

Automated Data Analytics

*Combining Human Creativity
and AI Power using ChatGPT*

Soraya Sedkaoui



Automated Data Analytics

*This book is dedicated to my little family, who light my way.
May it give back to you even a fraction of what you bring me.*

With all my gratitude and affection.

Soraya Sedkaoui

Series Editor
Jean-Charles Pomerol

Automated Data Analytics

*Combining Human Creativity and
AI Power using ChatGPT*

Soraya Sedkaoui

ISTE

WILEY

First published 2024 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

© ISTE Ltd 2024

The rights of Soraya Sedkaoui to be identified as the author of this work have been asserted by her in accordance with the Copyright, Designs and Patents Act 1988.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s), contributor(s) or editor(s) and do not necessarily reflect the views of ISTE Group.

Library of Congress Control Number: 2024941677

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

ISBN 978-1-78630-978-5

Contents

Preface	ix
Introduction	xv
Chapter 1. Artificial Intelligence (AI) and Automated Data Analytics	1
1.1. The emergence of automated data analytics and the potential of generative AI	2
1.1.1. The power unleashed by generative AI	2
1.1.2. Transforming the data analytics process	2
1.1.3. Redefining coding with the intelligent agent	3
1.1.4. Human–AI collaboration	5
1.1.5. The power of prompt engineering	5
1.1.6. An ethical North Star	6
1.2. Revolutionizing the data analytics process with ChatGPT	7
1.2.1. The numbers behind ChatGPT’s potential	7
1.2.2. ChatGPT and the future of data analytics	9
1.3. Harmony between human creativity and automated analysis: the winning duo	11
1.3.1. The value of human creativity	11
1.3.2. The complementary power of automated analysis	12
1.3.3. Navigating the partnership responsibly	13
1.4. Unlocking the secrets of prompt engineering for powerful results	15
1.4.1. The art of prompting	15
1.4.2. Optimizing prompts for efficient interaction with ChatGPT: fundamental aspects	17

- Chapter 2. ChatGPT for Data Analytics 21**
 - 2.1. Exploring the ChatGPT universe: history, presentation and capabilities 21
 - 2.1.1. From GPT-1 to GPT-4: generative pre-trained transformers 21
 - 2.1.2. ChatGPT in practice 24
 - 2.2. Powerful features for intelligent data analytics: natural language at your service 27
 - 2.3. ChatGPT versus data scientists: an intelligence battle that looks like an alliance 30
 - 2.3.1. Seamless communication and key features of an effective partnership 30
 - 2.3.2. ChatGPT, between ally and threat 33
 - 2.3.3. Automating data science workflows: unleashing the potential of ChatGPT plugins 35
 - 2.4. Benefits and challenges of integrating ChatGPT into data analytics workflows 37
 - 2.4.1. Unlocking analytical potential and reducing costs 37
- Chapter 3. Data Preparation for Analysis with ChatGPT 41**
 - 3.1. ChatGPT in charge of preparing our datasets 41
 - 3.1.1. Data cleaning 42
 - 3.1.2. Data transformation. 43
 - 3.1.3. Data formatting 43
 - 3.2. Automated cleaning and pre-processing for optimum results with ChatGPT. 45
 - 3.3. Handling missing data, outliers and other common data issues 48
 - 3.4. Using ChatGPT for data transformation, feature engineering and beyond 51
- Chapter 4. Intuitive Query Creation with ChatGPT 57**
 - 4.1. The discovery of patterns, trends and insights through interactive conversations 57
 - 4.1.1. Key benefits 59
 - 4.1.2. Discovering insights organically. 60
 - 4.1.3. Democratizing data analytics 60
 - 4.1.4. The future of business intelligence 61
 - 4.2. Creating natural language queries to analyze your data 62
 - 4.3. The art of transforming analysis questions into SQL queries with ChatGPT 65

4.4. Generating efficient and optimized queries: the key to your success with ChatGPT	69
Chapter 5. ChatGPT: The Advanced Analysis Wizard	73
5.1. Exploring new horizons: ChatGPT for exploratory data analysis	73
5.2. Simplifying your analysis: automation tasks for increased efficiency.	80
5.3. From statistics to predictions: ChatGPT as the partner of choice	82
5.3.1. Establishing statistical foundations	84
5.3.2. Building models with ChatGPT	84
5.3.3. Model deployment and monitoring models.	85
5.3.4. Business impact.	86
5.4. Deciphering feelings: text and sentiment analysis with ChatGPT.	87
Chapter 6. Prediction and Modeling with ChatGPT.	93
6.1. Automating the data analysis process with ChatGPT.	94
6.2. ChatGPT for accurate and robust modeling	96
6.3. Continuous improvement: optimizing model capabilities through feedback loops.	100
6.4. Trend and time series analysis	105
Chapter 7. ChatGPT at the Service of Machine Learning	109
7.1. Machine learning in the functional fabric of ChatGPT.	110
7.2. Creating new machine learning approaches with ChatGPT	113
7.3. Boosting machine learning algorithms with ChatGPT	116
7.3.1. Optimizing machine learning engineering	117
7.3.2. Hybrid model innovation.	117
7.4. Enhancing the potential of machine learning algorithms with ChatGPT.	120
Chapter 8. Narrative Fascination: Data-driven Stories and Reports	127
8.1. ChatGPT for generating data storytelling plans.	127
8.2. The bewitchment of words: automating for writing data-driven stories	131
8.3. Interactive dashboards and ChatGPT's ingenuity.	134
8.4. Humans at the heart of protocols: the imprint of human ingenuity in generative AI	137
Chapter 9. Power within Hands: Ethics, Orientation and Use	143
9.1. Understanding the limits of AI-generated analysis	144
9.2. Ethical harmony: ChatGPT in data analytics workflows.	147

9.3. Providing iterative feedback to improve ChatGPT	151
9.4. Addressing ethical concerns and biases when using ChatGPT in data analytics.	155
9.5. Ensuring fairness, transparency and accountability in automated data analytics	157
Conclusion	161
Appendix 1	167
Appendix 2	183
Appendix 3	185
References	191
Index	197

Preface

Comparing the capacity of computers to the capacity of the human brain, I've often wondered, where does our success come from? The answer is synthesis, the ability to combine creativity and calculation... into a whole that is much greater than the sum of its parts.

How Life Imitates Chess, Garry Kasparov (2007)

Data analytics is a crucial process in today's data-driven world. It involves collecting, cleaning, transforming and analyzing data to uncover useful information, insights, trends and patterns which inform business strategy, decision-making and process optimization. Traditionally, data analytics was a manual process requiring data scientists and analysts to prepare and process data before analyzing them. This was both tedious and time-consuming. The advent of machine learning and artificial intelligence (AI) has transformed data analytics by automating some parts of the process.

Generative AI models, like ChatGPT, are at the forefront of this automation revolution in data analytics. These large language models can understand human prompts and generate coherent and human-like textual responses. They are trained with massive text datasets, enabling them to perform a variety of language-based tasks. Generative models, such as ChatGPT, can be fine-tuned for specific applications, including data analytics.

We can consider these generative AI models as children with unlimited potential. The data scientist's role is to nurture these models, train them and help them grow – just as parents do with their children. In the beginning, these models are like children – they have immense capabilities, but need guidance to realize their potential.

The data scientist trains them step-by-step, teaching them the different tasks, operations and functionalities required for data analytics. This includes data pre-processing, cleaning, feature engineering, modeling, evaluation and explanation. Models are trained thoroughly, across diverse datasets, learning the nuances of each analytical task.

Gradually, just as a toddler learns to walk, talk and eat solid foods, generative models become capable with regard to various data analytics workflows. With each iteration, their skills improve: they learn to handle diverse datasets, manage missing values, transform features, select optimal models, critically assess performance and generate data-driven insights.

After extensive training on a wide range of datasets and analytical tasks, these generative models evolve from toddlers into mature analytical assistants. They move from simple memorization of problem-solving techniques to the development of true conceptual understanding. The models understand why particular data transformations, models and evaluations are appropriate for given scenarios.

In a way, they develop inductive reasoning and deductive logic, just like humans. They understand the principles and evidence-based principles that underlie data analytics workflows, rather than simply memorizing mechanical instructions. This conceptual understanding is what distinguishes generative AI from prior rule-based expert systems.

Thus, when a data scientist prompts a mature, well-trained model like ChatGPT to perform analysis, it deeply understands the request, rather than simply matching keywords. It draws on its conceptual knowledge to analyze the dataset, select optimal techniques, generate insights and explain the reasoning behind them. And it does so at superhuman speeds, leveraging the computing power of AI.

But does that make these generative models smarter than humans? The answer is no. At least not yet. Although they can outperform humans on narrow tasks within their training domain, these AI models lack generalized intelligence. Human cognition crucially remains far more advanced.

Unlike generative models, humans possess common sense, intuition, imagination, social intelligence, sensitivity and generalized reasoning capacities. We can creatively solve new and open problems that intersect several domains. Humans also have better judgment, wisdom and morality, which temper our technical capabilities with ethics and responsibility.

So, while ChatGPT can analyze datasets and quickly generate information, it lacks the general critical thinking skills to deeply understand implications and assess

ethics. Its intelligence is circumscribed by its training data and purpose. It cannot reason its way through completely new scenarios, as humans can through transfer learning.

That said, narrow AI models offer complementary advantages to human intelligence. Their prodigious memory and computational speed enable exhaustive data analytics. Their lack of bias and fatigue ensure consistent performance. In this way, they endow humans with superhuman data processing capabilities.

Rather than competing with AI, we can collaborate with it – combining human wisdom and ethics with AI's productivity and precision. Together, we evolve data analytics and make it more insightful and responsible. But humans must remain in the loop to provide guidance, assess implications and ensure alignment with ethics.

An ideal symbiosis is one where humans manage creative and strategic tasks requiring reason, ethics and imagination, while AI accelerates repetitive analytical tasks requiring memory, computation and precision. Similar to Iron Man deploying the AI assistant JARVIS to enhance his human capabilities.

So, while the gap between human intelligence and AI persists, narrow AI models like ChatGPT are still in their infancy. Their capabilities will continue to grow exponentially thanks to increased data size, computing power and algorithmic advances. One day, they may even cross the threshold into artificial general intelligence (AGI).

But for the time being, generative AI enhances rather than replaces humans when it comes to data analytics process. It takes care of the tedious parts, allowing data scientists to focus on creative, high-added-value tasks. It is becoming an indispensable analytical assistant that continues to learn – like a child who grows into an adult over the years with careful attention.

The key is for humans to guide the development of these generative models in a thoughtful and ethical way. We need to focus on beneficial objectives and monitor for harmful biases or abuse. With judicious care and training, AI can usher in an era of augmented analytics – where human and machine intelligences meet and converge for more powerful, yet ethical, data insights. But the human must remain the supervising parent for the AI child.

Rather than wondering when AI will surpass human intelligence, we should focus on how to cultivate beneficial and ethical AI applications today. Generative models like ChatGPT are impressionable children that will shape the future based on the guidance they receive. Data scientists have a profound opportunity and

responsibility: to educate these AI “children”, so that they become responsible and collaborative allies instead of impenetrable adversaries.

Just as teaching helps humans to consolidate our own knowledge, the training of AI models requires us to thoroughly evaluate our hypotheses, biases and best practices. AI development is as much about advancing our own intelligence – codifying disciplines into coherent frameworks, evidence-based principles and methodologies.

Collective training of generative models advances human knowledge across domains. This requires distilling nebulous issues into structured frameworks; formalizing messy tasks into step-by-step workflows; and crystallizing weakly defined domains into rigorous first principles. Teaching AI models through examples helps us to better evaluate solutions, generalize insights and formalize ethics for humans as well.

The future of data analytics is humans and AI working together – combining the human’s imaginative definition of problems, ethics and strategic judgment with AI’s vast memory, exhaustive computation and high-velocity analytical workflows. Neither can match the synergistic value of the two intelligences combined. Data science augments both human and artificial intelligence.

The time has come to actively train this child prodigy: ChatGPT! It holds enormous potential to enhance human capabilities if nurtured properly. We need to nurture it carefully – teaching analytical skills while emphasizing ethics, exposing ChatGPT to a variety of data and scenarios under supervision, so that it moves from mechanical regurgitation to contextual comprehension.

So, let us therefore guide these generative models with wisdom and kindness. Let us instill analytical techniques with values and ethics, guiding them from innocence to maturity. And let us develop artificial intelligence that makes individuals responsible instead of replacing them. Models such as ChatGPT are still in their maturing stages. With the careful supervision of researchers concerned with data ethics, they could evolve into assistants, opening up new horizons of ethically guided discovery, and collaborating more fruitfully with human than they could achieve on their own.

Just as Garry Kasparov’s quote states, human success arises from our ability to synthesize the creative and intuitive aspects of cognition with the computational and analytical aspects. When we combine these complementary modes of thinking and reasoning, the result is an emergent intelligence that simply exceeds the linear sum of creativity and calculation. There are synergies and amplifying effects that result

from the fusion of different thinking styles, which makes us unique as human beings.

Synthesis creates something with expanded potential and capabilities beyond what creativity or calculation could achieve individually. It is this integration that allows us to shine. This is what Kasparov believes the human brain excels at, compared to computers.

Kasparov's quote eloquently captures the spirit of the collaboration between humans and AI that we advocate for in this book – combining complementary strengths for amplified potential. Just as the synthesis of creativity and computation expands human cognition, the integration of human ingenuity with the analytical power of AI opens up new frontiers in data science.

When we design responsible workflows that harness both modes of intelligence, the results can far exceed their individual contributions. A capacity emerges from the thoughtful combination of human ethics and supervision with artificial productivity and rigor. This book provides frameworks for harnessing these synergies to advance data analytics. The future holds exciting possibilities as we forge partnerships between hearts and minds.

July 2024

Introduction

Imagine this: a seamless and interactive experience where you communicate effortlessly with your data, unravel their mysteries and uncover hidden insights through an engaging conversation. This vision can become a reality thanks to recent advances in artificial intelligence (AI) in data analytics.

The data-driven world is increasingly turning to AI to speed up and improve analysis. More specifically generative AI models, such as ChatGPT, are automating the process of interacting with data to discover insights. These generative models have immense potential as analytical assistants if developed responsibly under human guidance.

Just like a curious child who continually asks questions, generative AI enables fluid and intuitive exploration of datasets without the constraints of predefined queries or static reports. Users can engage in an ad hoc and open-ended dialogue, where the path is guided by emerging insights rather than being limited to pre-planned paths.

This creative symbiosis between humans and machine remarkably amplifies the analytical process. Domain experts provide strategic thinking, intuition, monitoring and ethical judgment. Meanwhile, intelligent AI agents handle the difficult computational tasks – rapid analysis of plain-language questions, analyzing huge datasets and generating interactive responses in natural language.

For example, prompted by a human analyst to evaluate sales patterns, ChatGPT could respond with high-level trends, correlations and high-level hypotheses it has identified by rapidly analyzing sales data. The analyst can then interpret these insights critically, ask follow-up questions to validate hypotheses and direct the AI to explore new perspectives.

This collaborative analysis is far more exploratory and multidimensional than static queries or predefined reports. It combines the framing and monitoring of human imagination with the exhaustive computational power of AI for an exhaustive and open exploration. Insights emerge more organically, guiding data analysts towards unexpected paths and serendipitous discoveries.

Generative AI augments the human analyst by automatically aggregating, processing and visualizing volumes of data exponentially faster than manual analysis. But it is human creativity that sparks new threads of inquiry, asks “why” and “how”, and connects analytical dots to extract meaning and implications.

The AI assistant finds signals buried in the massive noise. The human provides the contextual framework for interpreting signals into meaningful information. Together, they can unravel complex phenomena by intimately engaging with the raw data through a natural conversational flow.

This fluid and unconstrained approach enables data scientists to seamlessly traverse granular detail and broad trends as required. They can quickly zoom in on micro-level data points if something appears anomalous, then zoom back out to visualize macro-trends – spotting unusual events within wider contexts.

A key advantage is that generative AI allows humans to guide the analysis intuitively using natural language, without the constraints of structured queries or predefined analyses. No need to know programming languages or database schema. Users can engage in conversation as the process evolves.

This makes data exploration more accessible to a wider audience beyond just data scientists. Business leaders, frontline staff and others can participate, encouraging diverse analytical perspectives. The AI assistant becomes the great equalizer, enabling more stakeholders to unlock insights from data via a natural conversation.

Of course, like any child learning a skill, an AI assistant requires extensive training under human supervision to become proficient. It must ingest a myriad of datasets and scenarios to move from the mechanical analysis of queries to a true understanding of the principles and trade-offs involved when interacting with data.

Data scientists must teach these conversational models the nuances of analytical workflows – how to manage ambiguity, validate hypotheses, identify limitations, avoid bias, present information responsibly, etc. Ethics and responsibility must be emphasized, as models progress from innocence to maturity.

But once properly developed, generative AI assistants, like ChatGPT, can automate the laborious aspects of data analytics, while humans focus on high-level creative supervision. This enables exponential scaling of data exploration while maintaining a human-led direction. With sufficient guardrails, AI can even suggest new analytical paths for humans to evaluate.

Of course, generative AI has its limitations. Unlike humans, these models lack any real semantic understanding, beyond pattern recognition in training data. They also lack human judgment, intuition and ethics. Unconstrained automation could lead to misleading insights or generalized bias.

As a result, human guidance is crucial when deploying generative AI. Data scientists must assess the ethics, assumptions and blind spots behind any insights generated by AI. Generative conversational models are still, ultimately, narrow AIs, contrary to generalized human cognition. They excel only in their trained domain.

The ideal symbiosis is rapidly prompted, human-supervised, generative AI mechanical analysis. But humans provide the global framing, subjective judgment and moral responsibility for interpreting, validating and acting on ideas, both critically and ethically.

Neither humans nor machines alone can match the amplified intelligence of their collaboration. Together, they can mine both granular detail and big-picture meaning to uncover hidden insights in massive and complex data. AI expands the scope of analysis exponentially, while human creativity and ethics anchor it responsibly.

Generative AI promises to take data exploration to new frontiers by enabling a natural, fluid form of interactive analysis. But careful design and monitoring are essential if such automation is to be developed ethically and responsibly. When combined with human creativity and ethics, generative AI can open up new horizons of amplified analytical intelligence and accelerated discovery.

As data grows exponentially, AI will become an indispensable partner to augment human analysts. These models are still maturing children. Under the tutelage of responsible data scientists, they could become trusted analytical allies operating under the compass of human values to extract the maximum value from data for the benefit of society.

Why this book?

The emergence of AI has given rise to visions of a future where analytical tasks are fully automated by intelligent algorithms far exceeding human capabilities. This

narrative evokes both excitement at the possibilities and apprehension at the implications of handing over decision-making power to AI systems devoid of human values. It imagines a world where generative models like ChatGPT rapidly displace humans' footing in data science.

But this dystopian vision fails to take into account the enduring nature of human ingenuity and ignores the narrowness of contemporary AI technologies. As analytical automation is indeed accelerating, thanks to models like ChatGPT, they remain beneficial enhancements to human intelligence, more than adversaries. When developed responsibly, AI systems can amplify human potential to unlock new frontiers in data-driven insights and value creation.

However, careless implementation, focused solely on efficiency, risks undermining human action and responsibility. Maximizing the transformative power of AI-driven analytics requires respect for human ethics and surveillance throughout the process. This balance remains crucial, but precarious in the age of automation.

That is why this timely book, *Automated Data Analytics: Combining Human Creativity and AI Power using ChatGPT*, serves as an indispensable guide to the modern analytical frontier. It charts a prudent course, one where human imagination and ethics harness automation to elevate, rather than overwhelm, human capabilities. The book advocates thoughtful collaboration with AI, and not abandonment.

And this collaborative approach is, indeed, the wisest course. Because at their core, contemporary AI models like ChatGPT remain narrow and limited by the patterns contained in their training data. Although they can perform defined tasks at superhuman speeds, they lack generalized intelligence capabilities.

What we need is a harmonious fusion of the forces of human creativity, ethics and surveillance with the relentless analytical excellence of AI on vast datasets. Neither humans nor machines alone can match the amplified intelligence unleashed by their symbiosis. The book provides frameworks for productively structuring this collaboration for maximum collective benefit.

By shedding light on the inner workings, development processes and inherent limitations of AI systems, the book enables smoother integration into analytical workflows. Laying bare their technical realities helps to delimit optimal roles that emphasize the specialized strengths of humans versus machines. We can design complementary human–AI partnerships by assigning appropriate tasks based on contextual capabilities.

Humans must take the lead in framing problems, interpreting solutions and thinking globally, while AI undertakes data processing, calculation and in-depth

analysis. This book discusses various interaction models that enable a continuous flow between human creativity and the machine's productivity to constantly improve the discovery process.

It highlights techniques to better expose the model's decision logic, in order to increase trust and accountability. Continuous surveillance of model behavior and fail-safe guardrails are emphasized to ensure safety and prevent harmful abuse. The book also strongly advocates for maintaining sustainable human values and ethics at every stage of analytical automation to keep progress aligned with social benefits.

It explores frameworks for adapting to society and developing specialized human capabilities that unlock unique synergies with AI. All this is geared towards developing AI-driven analytics in a responsible way for collective human elevation. The book envisions an augmented intelligence, wherein humans and machines elevate each other, not AI supremacy over humanity.

Mainly, it lays the foundations for the ethical integration of automation into the analytical workflow – from strategic problem definition where humans lead, to iterative collaboration where both participate seamlessly, to evaluative surveillance where human judgment remains essential. This makes it possible to benefit from productivity gains without giving up creativity or responsibility.

The book comes at an opportune moment when AI is transforming analysis, but its capacities remain nascent and impressionable, like a young child. It underlines our responsibility as developers, leaders and citizens to wisely guide AI's progress from these formative stages to broader social benefit. The choices we make today will determine the long-term trajectory of artificial analytical intelligence.

And the book equips us to make these choices wisely by demystifying AI, highlighting its strengths and weaknesses, and providing ethically based normative advice. Whether you are excited or anxious about the revolution of automation, this book is an indispensable decision-making tool for navigating the future analytically.

Its frameworks, examples and principles enable us to integrate analysis responsibly with a human-centered perspective. You will get comprehensive, balanced clarity on collaborative opportunities that elevate without undermining human agency. And it champions respect for ethics as a guiding compass in the development and application of analytical AI.

This book inspires a future where AI makes data insights available to the whole of society for the common enrichment of human life. Its sound advice steers analytical automation in a prudent direction that amplifies human potential exponentially while keeping humanity firmly in control. The possibilities are

brehtaking if we move forward guided by the human values and spirit of collaboration that this book seeks to inspire.

Who this book is for

This book serves as an indispensable guide for the various stakeholders navigating the early stages of amplified analysis.

For data scientists and analytics teams, it provides actionable frameworks for working responsibly with AI while retaining human creativity and human supervisions. The book's insights will enable thoughtful development of analytical AI to unlock productivity while maintaining accountability. With illuminating case studies and prescriptive recommendations, data scientists can future-proof their skills and workflows to thrive in the age of automation.

The book is also extremely important for business leaders and decision-makers evaluating the adoption of AI-based analytics. It grounds discussions in ethical considerations beyond mere efficiency, helping leaders to make wise choices when integrating automated analytics into their organizations. Pragmatic advice allows leaders to select analytical workflows that wisely enhance the capabilities of their teams, rather than ruthlessly replacing them.

For policy-makers and regulators, the book highlights the need for adaptive governance to encourage responsible AI innovation in data analytics. Its forward-looking perspectives reveal how today's prudent policies can proactively shape the trajectory of automation in a direction that benefits society collectively. The book provides insight into the development of governance frameworks that are responsive to rapid technological change, while remaining grounded in ethics and inclusion.

In academic institutions, this book offers an indispensable program both in data science programs and for interdisciplinary discussions. It enriches understanding of how best to develop AI-based techniques founded on moral principles and responsibility. The book sheds light on research avenues into reliable and transparent analytical algorithms. And it enables a nuanced discourse on the balance between productivity and prudence in the age of automation.

The book also serves civil society and general interest readers looking for balanced perspectives on the integration of analytical AI. It dispels alarmist fears of AI, while emphasizing the need for thoughtful monitoring and adaptation. By making technical concepts accessible through clear metaphors and examples, the book enables the public to participate in the ethical development of the discourse on

automation. It contributes to the crucial dialogue on the direction of technology for collective human upliftment.

Students and young professionals will find this book invaluable in preparing for the future of analysis. It provides a fundamental understanding of how best to develop collaborative skills in AI from a technical and ethical perspective. The information will help students cultivate human-centered values and creativity even as they develop technical prowess in AI systems. This fosters comprehensive capabilities suited to the judicious integration of automation between roles.

But perhaps most critically, this work highlights the essential human responsibility in the development and deployment of analytical AI. It emphasizes that all stakeholders have an active role to play in shaping progress on the basis of moral principles and compassion. Only through collective foresight and wisdom can we nurture a beneficial intelligence that amplifies rather than diminishes human potential.

Overall, this book is aimed at anyone interested in or affected by the integration of automation and AI in data analytics. Its mix of technical foundations, ethical direction and pragmatic advice offers indispensable guidance for moving cautiously into the era of amplified analytics. The book awakens our collective consciousness to the responsible orientation of technology use towards the expansion of human capabilities. And it provides ways to synergize human and artificial intelligence under conscientious human management, so that automated analysis benefits, rather than harms, society.

The challenge of the book

The emergence of generative AI models such as ChatGPT has triggered a revolution in data analytics. These powerful technologies automate parts of the analytical process that previously required significant human effort. Tasks such as data cleaning, aggregation and basic reporting are rapidly transformed thanks to these AI systems.

An appropriate analogy compares the education of these AI models to the education of children. Like children, they start out as blank slates and gradually acquire complex skills through careful guidance. Data scientists act as parents, slowly exposing them to various datasets and analytical tasks during the training process.

Initially, the models are only capable of mechanically imitating operations based on training data. They do not have a deeper understanding of the concepts, principles

and implications behind the workflows. Models can repeat data processing steps with precision, but without any notion of meaning.

This is similar to a young child reciting multiplication tables by heart without grasping the deeper mathematical concepts of multiplication. They can perform the calculation accurately, but without the reasoning behind it. Early training of AI models also focuses on the repetitive execution of instructed tasks.

But with enough care over large datasets, these AI systems can move on to a true understanding of the patterns, relationships and rules governing analytical workflows. Just as varied exposure to the real world enables a child to deduce abstract concepts from concrete examples, extensive training helps models generalize.

Gradually, the models evolve from the simple memorization of instructions to the development of an inductive understanding of the principles underlying data analytics workflows. They understand why particular techniques are appropriate in different contexts on the basis of conceptual understanding, rather than blind implementation.

This transition reflects a student's move from simply memorizing mathematical facts to understanding how mathematical operations work on the basis of logical reasoning. Through further maturation, the models acquire contextual awareness to adaptively select the optimal analytical approaches for new datasets and scenarios.

In essence, they develop a trained intuition similar to that of human analysts. Models like ChatGPT acquire the ability to critically evaluate, explain and improve their analysis based on experience and learned patterns, rather than on simple, rigid instructions. Their analytical capabilities evolve in sophistication from basic obedience to contextual understanding.

As models develop proficiencies in a variety of datasets, they are transformed from passive analysis tools into intelligent collaborators in the information extraction process. They move from simply analyzing data to engaging in intelligent dialogue with human analysts.

The models provide the computational speed needed to rapidly process, cross-reference and visualize complex data on a superhuman scale. Their pattern-recognition capacities uncover relationships that humans may miss or take much longer to discern simply because of the sheer volume of data.

But human analysts provide the strategic framing, imagination and supervision needed to ensure appropriate interpretation and extraction of meaningful information,

rather than simple blind correlations. Analysts evaluate the assumptions, ethics and implications guiding the appropriate application of model results.

This creative synergy combines the power of data processing with nuanced judgment. Powerful AI models can generate new analytical avenues, but perceptive humans still need to guide the investigation by asking the right questions, through prompts. Models quickly reveal insights, but humans discern their meaning and applicability.

Together, humans and AI models can reach higher analytical heights than either of them can separately. But role divisions need to be carefully structured to benefit from this symbiosis. Just as shared parental responsibilities create specialization, the separation of analytical tasks promotes human–AI complementarity.

Delineating clear roles while collaborating closely avoids confusion and ensures that human creativity and the power of AI can amplify each other. But care must be taken to structure the human–machine partnership in such a way as to highlight their complementarities.

Uncontrolled automation without human supervision risks losing contextual judgment and responsibility. Giving full control to autonomous AI models ignores their limitations in relation to generalized human cognition. Although powerful, these models still lack certain distinctly human mental capacities.

For example, AI is currently struggling with complex abstraction, causality, transfer learning and innovative analytical techniques. Models like ChatGPT also lack human judgment, intuition, common sense, morality and appreciation of ethics or biases. They have limited capacities for imagination and social awareness beyond their training.

We must recognize these technological limitations to properly integrate AI into the analytical workflow as assistants rather than sole replacements. Their specialized capacity makes them productivity amplifiers for specific tasks rather than fully capable analysts themselves. Sound human guidance prevents misleading ideas or exaggeration of correlations.

Some predict a dystopian future in which AI will rapidly surpass all human intelligence – the so-called *technological singularity*, (or simply *the singularity*). But for now, AI remains limited to narrow applications, unable to compete with the multi-faceted reasoning, ethics and creativity of the human mind. Their capabilities, while prodigious, are constrained by the limits of their data training.

These models should be seen as powerful tools, not omniscient oracles. Like earlier innovations, such as calculators or computers, they extend certain analytical capacities without reproducing human cognition in the broadest sense. Their performance is exceptional, but narrowly delimited within a trained domain.

This is why this book, entitled *Automated Data Analytics: Combining Human Creativity and AI Power using ChatGPT*, rightly advocates developing and exploiting these technologies with caution, rather than entrusting them with complete analytical controls. It warns against focusing solely on efficiency gains, without considering social impacts or the loss of human control.

The aim should be to synergize and advance human and artificial intelligence to better understand data, faster and more responsibly. This involves establishing strong human safeguards and emphasizing ethics when applying AI models to avoid abuse or harmful bias.

But this also requires the development of mechanisms to interact seamlessly with AI, in order to improve, rather than simply automate, the analysis. This means structuring intuitive interfaces, intelligent workflows, and complementary human–AI team structures.

The optimal way forward is neither domination nor rejection of AI, but rather assimilation of AI under the aegis of human wisdom. This involves embracing AI's productivity gains while instilling human values, creativity and supervision through continuous training.

If we nurture AI wisely, avoiding the perils of uncontrolled automation, the future promises new frontiers of creative data-driven insights far beyond what humans or machines could realize independently. AI development must balance social well-being and human dignity, not sacrifice them on the altar of efficiency and capability.

The wise integration of hearts and minds with artificial analytical power can propel progress at an unprecedented pace and scale. But this requires setting ethical boundaries and priorities beyond simply maximizing analytical power using AI, for example, ChatGPT.

Human stewardship is essential to guide these technologies towards automation, rather than the replacement of human capabilities and responsibilities. We need to proactively shape their evolution based on moral considerations and social benefits.

With AI transforming data analytics, we have reached a decisive turning point, where our choices and priorities will define the trajectory of the future of ethical AI. Treading the path cautiously, yet boldly, with compassion and creativity, jointly

guiding analytical rigor, the dawn of hybrid intelligence promises an unprecedented, yet humanistic, discovery.

Navigating content

The contemporary world overflows with data, and in-depth analysis holds untold potential for individuals and organizations. However, traditional methods are struggling to unlock these nuggets of value. AI is emerging as a promising response, and in particular, generative models like ChatGPT automate entire sections of the analytical process, increasing our ability to extract impactful insights.

This book explores the fascinating world of AI-enhanced data analytics. Through a step-by-step exploration, it reveals the exciting possibilities of this symbiosis of human ingenuity and algorithmic power. The book demystifies these emerging technologies and provides a practical framework for integrating them ethically into modern analytical workflows. It presents an essential roadmap for the ethical and collaborative integration of automation in data analytics.

The interconnected chapters guide the reader through the process of training, preparing and supporting ChatGPT, in order to ensure the automation of the various tasks involved in the analytical process.

Chapter 1 sets the scene by examining how AI-driven automation is transforming data analytics. It highlights the immense potential of generative models such as ChatGPT to accelerate and deepen the extraction of insights. The chapter highlights the importance of combining human creativity and automated analysis in a harmonious way, in order to achieve augmented intelligence. It explains how to create effective prompts to get the most out of ChatGPT.

Chapter 2 delves into the heart of ChatGPT, revealing its unique capabilities for data analytics. It explores its powerful understanding of natural language, providing intuitive interaction. The chapter examines how ChatGPT can support data scientists, rather than replace them. It provides a practical guide to integrating ChatGPT into analytical workflows.

Chapter 3 shows how ChatGPT can automate data preparation – cleaning, pre-processing, management of missing values and outliers. This often tedious step is essential for high-quality analysis. By delegating these tasks to ChatGPT, analysts save valuable time.

Chapter 4 explains how to formulate natural and conversational queries for ChatGPT. No need to master SQL! ChatGPT takes questions in everyday language

and translates them into optimized queries. This intuitive approach frees up analytical curiosity.

Chapter 5 highlights ChatGPT's prowess in advanced analysis – data mining, statistics, prediction and sentiment analysis. Its assistance boosts analysts' performance, enabling them to concentrate on high-value-added tasks.

Chapter 6 covers the automation of predictive modeling and time series analysis with ChatGPT. Its systematic approach surpasses human capabilities for rigorous and accurate models. Continuous feedback cycles with analysts optimize its performance.

Chapter 7 reveals ChatGPT's tremendous potential for accelerating the machine learning process. Its ability to generate personalized algorithms and improve their performance opens up new horizons in this crucial discipline.

Chapter 8 explains how ChatGPT automates engaging data storytelling – the generation of reports, interactive dashboards and eye-catching presentations. But it emphasizes the importance of preserving a human imprint in these narratives to truly captivate readers.

Finally, the ethical dimensions surrounding the responsible use of ChatGPT in data analytics are addressed in Chapter 9. This chapter provides guidelines for mitigating bias, ensuring fairness and transparency. Without a moral compass, automation risks dehumanizing data analytics.

In broad terms, our aspiration is for this book to stand as a fundamental guide to thoughtfully integrating automation and AI in data analytics. We hope that its blend of ethical considerations and practical techniques will open up new perspectives on the development of human-centric augmented automated analytics.

This book aims to stimulate your curiosity about the vast possibilities when human creativity and machine intelligence converge. And we hope it will provide fertile seeds for further exploration in the elaboration of analytical workflows that ethically amplify insight and productivity. While generative AI propels analysis in known territory, this book seeks to shed light on a prudent path based on moral principles.

May it inspire you to participate in shaping the future trajectory of AI-powered automated analytics for the collective elevation of humanity.

Artificial Intelligence (AI) and Automated Data Analytics

We are today at the dawn of a new era, defined by the rise of generative AI and the immense potential it opens up to automate and increase data analytics. Systems such as ChatGPT (GPT, abbreviation for Generative Pre-trained Transformer), which designates a sophisticated machine learning model that has undergone extensive pre-training on a vast dataset, foreshadow a fascinating future, as well as raising important questions about the evolving role of human analysts and the responsible integration of AI.

The implications of this AI for data analytics are profound and disruptive. We now have an AI system that can not only calculate numbers, but also communicate with analysts in natural language as a collaborative partner. This human-like communication capability helps analysts to prepare data, formulate analytical questions, generate queries, run models, summarize results, and more.

In essence, generative AI opens up the possibility of automated data analytics where algorithms take on the mechanical and repetitive tasks, while human analysts provide strategy, imagination and oversight. Working together, humans and AI systems can achieve exponentially more, faster and better than either could alone. This vision is what makes the rise of ChatGPT a paradigm shift for data analytics.

Before revealing how to guide ChatGPT step by step through all the tasks of the analytical process, it is essential to map out the landscape, opportunities and ethical considerations surrounding the automation of data analytics by emerging generative AI.

1.1. The emergence of automated data analytics and the potential of generative AI

We are witnessing an unprecedented inflection point in the evolution of data analytics, catalyzed by the rapid advances of AI. The emergence of generative AI systems has paved the way for the automation of many analytical tasks traditionally requiring intensive human effort (Sedkaoui 2023). This monumental change has the potential to fundamentally transform the way humans extract insights from data.

1.1.1. The power unleashed by generative AI

A paradigm shift is underway, driven by the recent breakthroughs in deep learning. The development of natural language models based on transformers such as ChatGPT has produced AI systems capable of remarkably human-like written language behavior.

With over 1 trillion parameters, ChatGPT demonstrates an unprecedented ability to engage in contextual and coherent multi-turn natural language conversations on a dizzying array of topics (Tingiris and Kinsella 2021). This presents a watershed moment for the unlocking of automated data analytics using natural language interfaces.

By providing detailed instructions and feedback in simple conversational language, humans can now collaborate with AI systems to perform a wide range of analytical tasks. This includes data cleaning, pre-processing, visualization, statistical analysis, query generation, predictive modeling, report creation, etc.

The implications are far-reaching. Every facet of the data analytics life cycle – from raw datasets to final insights – can potentially be accelerated, augmented and automated through collaboration with generative AI systems.

1.1.2. Transforming the data analytics process

The integration of generative AI promises to profoundly reshape and accelerate the entire data analysis process. With ChatGPT, capable of handling many basic tasks in response to natural language prompts, human analysts can pursue more ambitious analysis goals and devote their time to high-value work.

The preparation, cleaning and processing of datasets to make them ready for analysis often presents the most labor-intensive phase of analysis (Manning et al. 2014; Sedkaoui and Khelfaoui 2020). ChatGPT would automate the bulk of the work, following guidelines to facilitate data analytics, preparation and

transformation. It would also help to summarize datasets for analysis purposes, reducing analyst burden while improving consistency.

Formulating analytical questions and translating insights into executable queries and scripts is another area where AI is extremely promising. Instead of painstakingly coding queries, analysts could rely on ChatGPT to interpret the questions provided in plain language and suggest appropriate SQL or Python queries to run (Tingiris and Kinsella 2021).

For advanced analysis covering statistical tests, forecasting and prediction, ChatGPT can also help design rigorous analytical workflows and rapidly perform calculations. Throughout AI's natural language capabilities, namely ChatGPT, enable a highly interactive iterative process between humans and AI.

The combined effect is a dramatic increase in analyst productivity and scope. With AI systems automating all rote-coded tasks, analysts are able to find deeper insights faster among exponentially larger and more complex data.

1.1.3. Redefining coding with the intelligent agent

By exploiting the advanced natural language processing capabilities of models like ChatGPT, data scientists can efficiently improve their workflows and achieve superior project results. The versatility of these intelligent agents extends their benefits to programmers and data analysts in a multitude of ways.

When it comes to programming, they are an invaluable asset. Their code generation prowess enables programmers to quickly formulate code snippets by responding to natural language prompts, which is particularly beneficial for rapid prototyping of ideas and boilerplate code generation. For example, by training ChatGPT on an assortment of code examples, a language model can emerge, capable of producing syntactically accurate code segments across a range of programming languages.

Another area where AI models prove invaluable is the creation of documentation. By familiarizing models with a corpus of code comments and documentation, programmers can cultivate a language model capable of autonomously composing documentation for code snippets or entire codebases, thus streamlining the documentation process.

In addition, these models can contribute to code optimization by suggesting techniques for simplifying or refining code segments. Trained on a diverse collection

of optimized code samples, the generative model evolves into a competent advisor, providing recommendations for improving existing code. This enhancement serves to reduce code complexities, improve performances and facilitate maintenance efforts.

It should be noted that generative AI models, such as ChatGPT, facilitate error management, offering potential solutions to common coding errors (Sedkaoui 2023). Drawing on a knowledge pool of sample codes containing errors and their respective remedies, the model evolves into a resource that can provide solutions to common coding errors. This automated approach to error resolution can significantly improve the programming experience.

ChatGPT, for example, is a powerful tool that can be used for a variety of programming and debugging tasks. It can generate code snippets, optimize code, debug code, document code and review code.

ChatGPT is able to perform these tasks by analyzing programming language functions, requirements, algorithms, data structures, code structures and error messages. ChatGPT can be used to improve programmer productivity and efficiency by automating many of the tasks involved in programming and debugging code.

In essence, the advanced language capabilities of ChatGPT, or other generative AI models, are a boon to programmers and data scientists alike. By integrating this technology, workflows are streamlined, coding efficiency is increased and the overall development process achieves new levels of productivity and efficiency.

Task	ChatGPT usage
Code generation	Generating code snippets based on user input, facilitating the implementation of specific functions or features
Code optimization	Analyzing code to identify inefficiencies and recommend improvements, enhancing code performance and efficiency
Debugging assistance	Analyzing code structure and error messages and providing debugging recommendations, helping to resolve errors
Code documentation	Analyzing code structure and requirements and suggesting best documentation practices, contributing to clear documentation
Code review	Analyzing code and coding standards, offering suggestions for improving code quality and reliability during code reviews

Table 1.1. *ChatGPT: a programming and code debugging tool*

1.1.4. Human–AI collaboration

The unique capabilities of AI systems create the foundations for a highly collaborative human–AI partnership that reinvents data analytics. By managing time-consuming, codified tasks in response to natural language instructions, this frees up human analysts to focus on high-level strategic objectives that draw on unique human skills such as creativity, intuition and reasoning.

Tomorrow’s data scientists will devote their time to valuable work, such as formulating the right analytical questions, interpreting results, challenging hypotheses and communicating insights. AI assistants will perform data preparation, modeling, calculation and reporting under the guidance of the analyst.

This vision promises to greatly boost the productivity of data scientists and enable deeper, faster data analytics. What once took an entire team weeks (Davenport and Harris 2007; Sedkaoui and Khelfaoui 2020) could soon be achievable for a single analyst in hours, or minutes, by harnessing the synergies of human and machine intelligence. This division of labor translates into a mutually beneficial arrangement, where AI brings speed, rigor and objectivity, while humans provide contextual advice, validation and monitoring.

In this human–AI symbiosis, repetitive, codified tasks are delegated to natural-language AI assistants, freeing up analysts for strategic reflection and imagination. Together, this hybrid team achieves synergies and augmented capabilities far beyond what either could do separately. The whole becomes greater than the sum of its parts. This collaborative workflow combines the best of human and machine intelligence.

1.1.5. The power of prompt engineering

To collaborate effectively, generative AI systems like ChatGPT need detailed instructions and feedback from humans. This is achieved through careful prompt design – the art of designing prompts that clearly communicate tasks and desired outcomes to AI.

The prompts that analysts use to run ChatGPT have a considerable influence on the quality, usefulness and ethics of the results produced. Carefully crafted prompts, providing relevant context and setting clear expectations, are essential for successfully leveraging ChatGPT in data analytics.

Prompt engineering involves an iterative process of refining prompts based on AI responses to improve relevance, accuracy and specificity (Jiang et al. 2020). Analysts must learn to formulate concise prompts that effectively guide the AI to apply the appropriate knowledge and perspective to generate optimal results.

Mastering prompt engineering is crucial to guiding AI systems towards productive contributions while avoiding potential pitfalls. The prompts provided by analysts will shape the integration of generative AI systems into analysis workflows (see Appendix 1).

To take full advantage of the promise of generative AI in the data analytics process, humans must thoughtfully guide the process. Certain best practices play a pivotal role in building an ethical and productive collaboration with ChatGPT and similar AI systems, for example:

- providing sufficient context in the prompts to indicate the appropriate capabilities and frame of reference;
- iteratively re-prompting based on results to address shortcomings;
- critically reviewing AI-generated output for completion, accuracy, potential biases and errors;
- providing feedback via additional prompts to further improve the AI system;
- maintaining human control over the direction of analysis, ethics considerations and decision-making.

Adherence to these principles will steer AI systems towards their highest utility, while keeping humans firmly in charge of strategy and ethics. Overall, the prospects are extremely exciting, as AI is poised to energize the way humans extract insights from data.

1.1.6. *An ethical North Star*

As we further integrate these AI systems into data analytics and data science, certain considerations arise. As with any powerful technology, to realize the full potential of generative AI in data analytics requires careful thought and guidance. While AI systems like ChatGPT harbor no intrinsic bias or motivation, their outputs are only as ethical and objective as their human designers.

One of the main concerns is the possibility of errors or biased results in data science when relying on AI systems (Osoba and Welser 2017). ChatGPT's performance depends on the quality of the data on which it has been trained. Consequently, any existing biases in the data may be amplified, leading to non-objective results. However, this concern does not justify rejecting AI's potential entirely; instead, it emphasizes the need to use diverse and representative data when training the model.

We must remain vigilant, so that by integrating AI into analysis workflows, we do not inadvertently perpetuate historical biases in training data. Transparency around data provenance and algorithmic limitations is crucial. Extensive human surveillance and correction of AI systems offer another vital guarantee.

Overall, by proactively identifying and mitigating risks early, we can steer AI towards unambiguously positive outcomes that elevate society. The promise is too great not to pursue, but we must do so with wisdom and principle.

1.2. Revolutionizing the data analytics process with ChatGPT

The integration of generative AI systems such as ChatGPT has the potential to profoundly accelerate the data analytics process. By automating many repetitive and codified tasks, ChatGPT enables analysts to focus their time on high-value objectives that leverage human judgment, creativity and strategy.

1.2.1. The numbers behind ChatGPT's potential

It is clear that AI-based systems are poised to bring transformative changes to various spheres of life. The underlying mechanism of this technology involves advanced machine learning algorithms that examine vast repositories of data, such as books, articles and web pages.

This intelligent AI agent is based on transformers, a type of attention-based deep neural network particularly effective for natural language tasks (Tingiris and Kinsella 2021). Transformers integrate aspects of supervised, unsupervised and reinforcement learning.

During pre-training, the model parameters are trained in an unsupervised manner on massive text corpora to learn language patterns, semantics and world knowledge. The pre-trained model is then refined using reinforcement learning based on human feedback to optimize its conversational capabilities and human alignment.

Supervised learning provides fine-tuning on specific tasks, such as classification using labeled training examples to improve performance. So, while transformers and neural networks form the core architecture, unsupervised pretraining, reinforcement fine-tuning and supervised customization combine synergistically.

This reflects the real-world practice of mixing learning paradigms to capitalize on their complementary strengths. The use of self-supervised pre-training, followed

by downstream task tuning, has become a common pattern in state-of-the-art deep learning. ChatGPT illustrates the strength of hybrid approaches. Its foundation is based on recursive self-improvement through machine learning, guided by human feedback.

As a result, ChatGPT is proving to be a powerful tool with many applications, including enhanced chatbots, automated content creation and more efficient search engines (Taecharungroj 2023; Thorp 2023). In addition, the potential for improving language translation has the capacity to revolutionize intercultural communication.

According to a recent McKinsey report (2023), the potential economic impact of generative AI can be assessed through:

- an analysis of over 60 organizational use cases across all sectors and functions, estimating USD 2.6–4.4 billion in the potential annual value creation worldwide;
- modeling the impact on automation of work activities in all professions, which could contribute 0.1–0.6% per year to growth in productivity up to 2040.

Generative AI represents a radical improvement in capabilities over previous AI techniques.

Models like GPT-3 underpin many generative AI applications, trained on massive datasets to perform a variety of natural language or visual input-based tasks. According to the McKinsey report, the most significant revenue impact is estimated in the banking (up from 2.8 to 4.7%), high-tech (from 4.8 to 9.3%) and pharmaceutical (from 2.6 to 4.5%) industries. In retail, marketing and customer service use cases could generate revenue growth from 1.2 to 2%, worth USD 400–660 billion per year.

IMB's report (2023) on enterprise generative AI is along the same lines, highlighting concrete examples of the successful adoption of these technologies by leading companies, such as Lloyds, which used chatbots powered by AI text generation to deliver fast, consistent responses to customers; and Bouygues Telecom, which deployed virtual assistants based on natural language processing to simplify customer–agent interactions in its call centers, practical applications which well and truly illustrate the potential of generative AI to streamline and optimize customer engagement.

Type	Year	Milestone
Basic chatbots	1966	ELIZA, the first chatbot, developed by Weizenbaum
	1972	SHRDLU, a language processing program by Winograd
	1995	ALICE (Artificial linguistic Internet computer entity), developed by Wallace
	2001	SmarterChild, a popular chatbot on AOL Instant Messenger
Conversational agents	2011	IBM Watson wins <i>Jeopardy!</i> against human contestants
		Siri: Apple's virtual assistant for iOS devices
	2014	Microsoft introduces Xiaolce, a popular chatbot in China
		Alexa Amazon's voice-controlled virtual assistant
	2015	Invention of the diffusion model, which learns to generate images from noise, and serves as the basis for image generators such as Dall-E
	2016	Facebook launches its chatbot platform on Messenger
		Microsoft Tay, a Twitter-based chatbot, creates controversy due to inappropriate behaviour
	2017	Google introduces Google Assistant and Duplex
Generative chatbots	2021	Jasper AI: content platform that helps creators and businesses of all kinds to develop their creative potential
	2022	OpenAI launches GPT-3, a large language model
		OpenAI launches ChatGPT, an AI-powered conversational agent
	2023	Google creates Bard, its generative AI GPT-4 version was introduced by OpenAI Claude created by Anthropic

Table 1.2. *From basic to generative AI: a timeline*

1.2.2. ChatGPT and the future of data analytics

Data analytics is a fundamental and indispensable capability that drives operations, decision-making and innovation across various sectors. The advent of generative AI has the potential to revolutionize data analytics in several significant ways, as shown in Figure 1.1.

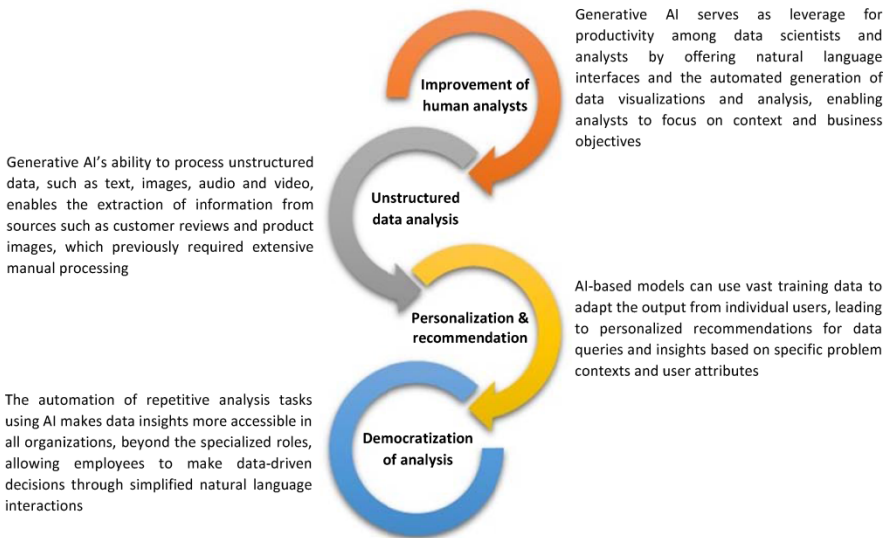


Figure 1.1. Benefits of generative AI in data analytics

However, the real value lies in the augmentation of human intelligence by analyzing data to produce initial insights that humans then validate and apply using their own judgment and expertise. The combination of automatic analysis and human intelligence promises to generate new knowledge on a massive scale, demonstrating how AI can transform the way organizations harness information – although the greatest impact will come from the human–AI collaboration, which helps, rather than hinders, the development of new knowledge.

These AI-based systems can also elevate the skills and expertise of data scientists. By automating tedious tasks such as data cleaning and pre-processing, AI systems enable analysts to focus on more complex, creative and intuitively driven aspects of data analytics. The synergy between human and AI capabilities enables more accurate, data-driven decision-making (Hendrycks et al. 2021).

What is more, AI's ability to identify hidden patterns and relationships in large datasets can help generate hypotheses that data scientists can then validate through further analysis. In addition, AI's ability to generate synthetic data proves invaluable in scenarios where real data are scarce or difficult to obtain, enabling more robust training and testing of machine learning models.

It is therefore crucial to recognize that AI is not a threat, but rather a valuable ally. Although it can automate specific tasks, it remains incapable of replicating the

holistic capabilities of human professionals. AI systems lack the ability to interpret results strategically or make data-driven decisions, both of which remain in the domain of human expertise.

Therefore, rather than fearing the emergence of AI technologies such as ChatGPT, data scientists should embrace these advances as invaluable tools in their data analytics toolbox. The harmonious collaboration between humans and AI has the potential to unlock new levels of productivity and insight, ultimately contributing to the advancement of data science and analytics.

1.3. Harmony between human creativity and automated analysis: the winning duo

The rise of generative AI ushers in an era of symbiotic collaboration between human creativity and automated analysis. By judiciously balancing their respective strengths, humans and AI systems like ChatGPT can form a winning duo that unlocks deeper insights from data.

1.3.1. *The value of human creativity*

According to the McKinsey report (2023), with the increasing adoption of AI and machine learning in different industries, the roles of the data scientist are expanding beyond technology companies to in-house AI/machine learning consultants, trainers and experts.

In this context, it should be noted that certain aspects of data analytics rest fundamentally on the singular capacities of the human mind. Creativity, intuition, imagination, empathy and ethics underpin the data scientist's most important contributions. Formulating insightful questions and strategically defining the problem, discerning subtle patterns and non-obvious relationships within the data, validating analytical models and results, identifying gaps, exceptions and areas requiring further exploration, as well as communicating convincing narratives and visuals tailored to diverse audiences, and responsibly overseeing the direction of analysis and use of insights, are all areas where humans excel, despite the advances of AI.

When human creativity and automated analysis converge as partners, they can achieve far more together than either can separately (Sedkaoui 2023). AI provides scale, speed and tireless calculation, while humans provide strategy, ethics and surveillance. Some features of this symbiotic human-AI analytical duo include the following.

Operation type	Human	Generative AI
Fundamental questions	Frames fundamental analytical questions	Suggests optimal methodologies
New directions	Identifies new data directions	Quickly compiles and analyzes required data
Subjective context	Provides subjective context	Injects computational rigor into deductions
Outliers	Provides the instinct to spot outliers	Validates wide-scale models
Communications	Prepares communications	Generates visuals and data stories required

Table 1.3. *The human–AI analytical duo*

This highlights the differences in roles between humans and generative AI in various data analytics operations. Humans bring expertise in framing fundamental questions, identifying new directions and providing subjective context, while generative AI excels in suggesting optimal methodologies, compilation and rapid data processing, injecting computational rigor into inferences, validating large-scale models and generating required visuals and data stories. By combining these skills, humans and generative AI can collaborate synergistically for more efficient and in-depth data analytics.

This division of labor enables each partner to operate to the full and to the best of their ability in the service of common analytical objectives.

1.3.2. The complementary power of automated analysis

Where human creativity weakens, automation can fill the gaps in an exceptional way. AI systems, such as ChatGPT, have complementary strengths that give them a significant advantage in a variety of analytical tasks. These advantages revolve around key capabilities, including:

- tirelessly managing repetitive and rote analytical tasks without the fatigue that can affect humans. AI can process large amounts of data with consistent accuracy and coherence, even during intensive processes;

- rapidly performing complex statistical calculations beyond human mental capability. Thanks to their computational power, AI systems can perform advanced mathematical operations at unprecedented speed, considerably speeding up analytical processes;

- examining a multitude of information and variables during machine learning. Unlike humans, AI can process big data produced from different sources, which improves the quality of learning models and predictive analyses;

- providing a rigorously consistent result once trained, without deviating from its logic or premises. AI avoids bias or human error related to the interpretation of data, which reinforces the reliability of the results;

- exponential scaling of analytical workloads without marginal human effort. AI can adapt to increasing data volumes and analytical queries without sacrificing efficiency, helping to meet growing business needs.

These computational strengths present areas where AI automation holds clear advantages over unassisted human analysis. Integrating AI into the data analytics process allows data scientists to focus on more complex, creative and strategic tasks, while capitalizing on AI's undeniable strengths to optimize the accuracy, efficiency and scalability of their analyses.

Thus, the advent of AI-based systems is a powerful force propelling the evolution of data professions. The combination of AI's computational prowess and human expertise is leading to a new kind of data analytics. By embracing the strengths of machines and humans, we can develop more accurate models and make more informed, data-driven decisions.

1.3.3. Navigating the partnership responsibly

In the future, the integration of AI promises to transform the role of the analyst, placing them at the helm of automated systems, rather than simply performing individual analyses. Thanks to AI, which lightens their burden by automating repetitive tasks, future data scientists can devote themselves more to creative and strategic initiatives with added value. These advances open up new prospects for analysts, who will be able to play a decisive role in the design and governance of high-performance analytical systems.

They will focus on identifying key questions and creative hypotheses, exploiting AI to reveal hidden patterns in the data. In addition, they will use AI to extract information from a variety of sources, and to develop innovative analytical frameworks to address complex problems. Analytical results will be translated into concrete actions and strategic recommendations to meet business and societal needs.

Future analysts will also play a key role in setting ethical standards for the responsible use of AI, and will oversee processes for continuous monitoring and improvement of automated analytical systems.

This reinvented role firmly places human creativity at the head of an AI-powered engine. By capitalizing on the complementary strengths of humans and AI, future analysts will have the opportunity to shape the future of data analytics and contribute significantly to innovative discoveries, informed decisions and strategic advances in all areas of business.

But to cultivate an effective and ethical human–AI analytical partnership, it is essential to focus on key priorities that ensure a harmonious and responsible collaboration between the two entities. These priorities serve as the foundation for fully exploiting the analytical potential offered by the combination of human creativity and automated analysis. Here are a few important aspects:

- *Human supervision and control*: human supervision must remain at the heart of the analytical process, guiding the direction and use of automated analysis. Humans play a crucial role in determining the objectives, questions to be answered and decisions to be made on the basis of AI results. They must maintain a global and strategic perspective to ensure that analytics align with business and ethical objectives.

- *Transparent communication*: clear, transparent communication is essential to building trust between stakeholders. It is essential to disclose the hypotheses, limitations and biases underlying AI systems used in the data analysis process. This transparency enables decision-makers to fully understand the implications and uncertainties associated with AI-derived results.

- *Rigorous validation of results*: although AI can provide fast and powerful analyses, human validation is essential to guarantee the reliability and quality of results. Humans must carefully examine AI results, understand the reasons behind the conclusions and ensure that the findings are consistent with pre-existing knowledge and ethical principles.

- *Feedback loops*: automated analytics should not be seen as a static solution. On the contrary, it is necessary to establish feedback loops to continuously improve AI models, correct biases and adapt to changes in data or objectives. Humans must play an active role in this continuous improvement using feedback and new information to optimize the efficiency and relevance of analyses.

- *Respect for ethics and privacy*: at every stage of the analysis, it is imperative to respect ethical principles, data privacy and fairness in the process. Humans must ensure that AI does not reproduce or amplify undesirable prejudices, and that the rights and privacy of individuals are preserved.

By integrating these responsible principles into the design, coordination and governance of human–AI analytical partnerships, we are able to unleash the enormous potential of data analytics. This synergy between human intelligence and automated analytics enables complex challenges to be addressed, informed decisions to be made and emerging opportunities to be exploited with increased efficiency and ethics.

However, achieving an effective synthesis of human and artificial intelligence relies heavily on prompt engineering. Mastering this skill is essential to successfully direct AI systems towards productive and ethical results.

1.4. Unlocking the secrets of prompt engineering for powerful results

As generative AI, in particular ChatGPT, is gaining in popularity, the art of prompting has become a crucial skill in the field of AI. Crafting a prompt, the means of interacting with large language models (LLMs), is not as straightforward as it may first appear. Beyond the novelty of initial interactions with ChatGPT, it has become clear that mastering prompt engineering requires practice and careful thought.

1.4.1. *The art of prompting*

It is truly remarkable that ChatGPT can automatically generate text closely resembling human-written content. But how does it achieve this feat, and what is the reason for its success? Although we are talking about certain technical aspects, these principles also apply to other LLMs, not just ChatGPT.

Basically, ChatGPT’s main objective is to produce a “reasonable continuation” of the given text. By “reasonable”, we mean an answer that corresponds to what you would expect to write after analyzing several web pages and digital books. Imagine, for example, this original text:

In data analytics, one of the most crucial aspects is data pre-processing, which involves...

To achieve its goal, ChatGPT analyzes vast datasets containing information from various sources (such as BookCorpus, Common Crawl and Wikipedia). Instead of simply searching for literal matches, ChatGPT uses sophisticated algorithms (Sedkaoui 2023; Taecharungroj 2023; Van Dis et al. 2023; Zhang et al. 2021) to identify concepts and patterns that “match the meaning”.

As shown in Figure 1.2, when you interact with ChatGPT by asking a query, it uses natural language understanding (NLU) to deduce the user's objective. Then, relevant data are extracted from the underlying knowledge base. Natural language generation (NLG) is then used to summarize an appropriate response based on these data. The discussion history is preserved to enable future personalized interactions. To improve the quality of responses in the future, reinforcement learning is used to gather user feedback and take appropriate action (Sedkaoui 2023).

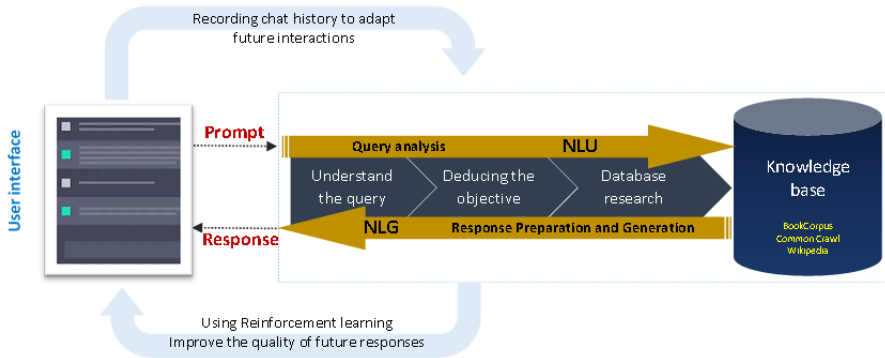


Figure 1.2. *How does ChatGPT work?*

This process allows ChatGPT to generate a ranked list of words and expressions that are highly likely to follow the initial text, along with associated probabilities, ensuring that the sequence generated conforms to data analysis principles and best practices.

Fundamentally, the use of questions and their formulations, known as prompts, is the essential mechanism for exploiting ChatGPT's full potential while managing the associated risks.

The prompts that data scientists use to drive AI systems like ChatGPT profoundly influence the usefulness of the results. By mastering prompt engineering techniques, they can optimize ChatGPT's responses to accelerate their workflows. Well-designed prompts are essential to successfully harness AI (Jiang et al. 2020; Schick and Schütze 2021).

Prompt engineering refers to the iterative process of designing and perfecting prompts to steer AI systems towards desired behaviors. With large language models like ChatGPT, prompts are the only way for humans to communicate their intent and guide the AI.

Knowledge of how to create the most effective prompts has now become sought-after expertise in the LLM community and beyond. As the popularity of generative AI continues to soar, the importance of prompt engineering has grown considerably, demanding the attention and skills of those involved in the AI world. Now that we have established the structure of a basic prompt, let us take a closer look at the essential considerations for better mastering prompt engineering.

1.4.2. Optimizing prompts for efficient interaction with ChatGPT: fundamental aspects

The wording of a prompt plays an essential role, as it guides the LLM, such as ChatGPT, in producing the desired results. Writing the question or statement in a way that ChatGPT understands and responds to accurately is of the utmost importance.

For example, if a user lacks expertise in a specific field and has difficulty formulating a question with the appropriate terms, ChatGPT's ability to provide relevant answers may be limited. This can be likened to searching the Web without knowing the right keywords.

While it may seem obvious that adding extra information could lead to better prompts, it is less obvious that too much verbiage in general is not necessarily an optimal strategy. It is best to consider prompt wording not as an isolated technique, but rather as a common thread linking all other prompt engineering techniques.

Conciseness in a prompt is crucial for clarity and accuracy. A well-designed prompt should be brief and to the point, offering enough information for ChatGPT to grasp the user's intent without unnecessary elaboration. However, it is essential to avoid making the prompt too brief, as this could lead to ambiguity or misunderstandings. Striking the right balance between brevity and exhaustivity can be challenging. Practice is probably the key to mastering this competency.

Thus, in prompt engineering, roles refer to the people assigned to the LLM and the target audience. For example, if you want ChatGPT to create an outline for a blog post on machine learning classification metrics, clearly specifying that the LLM should act as a machine learning expert and that the target audience is composed of newcomers in data science would significantly improve the relevance of the answer.

The way in which this information is expressed can be experimented with, either using a conversational style: "You are a seasoned data scientist working for a logistics company"; or a more formal approach: "Role: logistics expert analyst; audience: business decision-makers", depending on the context.

Objectives are closely related to roles. Explicitly stating the goal of a guided interaction is not only advisable, but essential (Schick and Schütze 2021). Without a clearly defined goal, ChatGPT would lack the guidance needed to generate the desired result. In addition, positive and negative prompts are two different approaches used to influence the model's outputs.



Figure 1.3. *Basic prompts versus role definition*

By using both types of prompts, we can significantly shape the direction and quality of the result.

Prompt engineering is therefore an essential element in the efficient use of LLMs, such as ChatGPT, and is a powerful way of tailoring interactions with the model. This involves carefully crafting clear and specific instructions or queries, in order to obtain the desired responses from the LLM model. By skillfully constructing prompts, data scientists can guide ChatGPT's output towards their intended goals, resulting in more accurate and useful responses.

During the prompt optimization process for ChatGPT, there are several fundamental techniques to keep in mind:

- firstly, providing explicit instructions at the beginning helps to set the context and define the task for the model;
- specifying the format or type of response expected can also be beneficial;
- in addition, improving interaction can be achieved by incorporating system messages or using role-playing techniques in prompts.

As a result, the way in which data scientists formulate prompts directly determines the relevance, accuracy and usefulness of ChatGPT's responses to automate parts of the analytical workflow.

Prompts enhance contextual understanding of ChatGPT and guide appropriate analytical capacities. Inappropriate prompts provide insufficient guidance, leading to irrelevant or inaccurate results.



Figure 1.4. *Guide to strategic prompting with ChatGPT*

Developing and mastering strategic prompting, illustrated in detail in Figure 1.4, makes it possible to guide ChatGPT in a collaborative way to produce personalized, optimized responses. It includes defining tone, format and desired scope to align expectations. Providing background and clearly stating the objective gives direction to this strategy. Enriching prompts with relevant keywords, examples and data makes responses more substantive

Adapting language and terminology to suit the audience and field optimizes comprehension and accuracy. Encouraging critical thinking by asking for multiple perspectives strengthens reasoning and objectivity. And respect for ethics through careful wording encourages nuanced processing of sensitive subjects.

By following these principles, you will get answers that are precisely aligned with your needs, while respecting the standards of reasoned discourse.

Strategic prompting formulation allows you to collaboratively guide ChatGPT to generate optimized responses tailored to your needs.

As the adoption of generative AI accelerates, mastery of prompt engineering will become an essential analytical skill. With care, rigor and ethical considerations, wisely designed prompts enable analysts to use the immense possibilities of AI augmentation, while maintaining vigilant control. The future of analytics belongs to those who learn to judiciously combine their human creativity with AI capabilities through the channel of well-developed prompts.

Now that a strategic foundation has been laid, our exploration continues with a deep dive into ChatGPT's revolutionary capabilities and their integration into the data analytics workflow. The following pages shed light on the ways in which analysts can harness the power of this technology.

ChatGPT for Data Analytics

The advent of ChatGPT presents a crucial milestone in the evolution of AI, unlocking revolutionary new capabilities to automate and augment human data analytics. Built on a foundation of deep learning at unprecedented scale, ChatGPT provides a highly advanced LLM language model, ready to energize analytical workflows.

This chapter provides a comprehensive guide to integrating ChatGPT into the data analytics process. We explore ChatGPT's origins and attributes, its collaborative potential, integration nuances and real-world applications. We begin by discussing the essence of ChatGPT, where GPT-4 is the latest iteration of this technology, building on the remarkable success of its previous versions.

2.1. Exploring the ChatGPT universe: history, presentation and capabilities

ChatGPT presents a revolutionary advance in generative AI, made possible by recent advances in deep learning. Built by OpenAI, using self-supervised learning on massive datasets, ChatGPT offers remarkable natural language capabilities to power a new generation of AI applications.

2.1.1. From GPT-1 to GPT-4: generative pre-trained transformers

OpenAI has made remarkable progress in natural language processing (NLP) with its versions of GPTs, ranging from GPT-1 to GPT-4. These models paved the way for AI-generated content, covering a wide range of applications, such as the generation of prose, poetry, chatbots and even coding. However it is crucial to understand the distinctions between each model and the significant impact they have had on the NLP field.

Generative pre-trained transformers (GPTs) are a type of machine learning model used in natural language processing tasks (Sedkaoui 2023; Taecharungroj 2023; Van Dis et al. 2023). These models are pre-trained on large datasets, such as books and web pages, which enables them to generate contextually relevant and coherent language.

In simpler terms, GPTs are computer programs capable of generating human-like text without explicit programming. As a result, they can be fine-tuned for a variety of NLP tasks, including question answering, translation and text summarization.

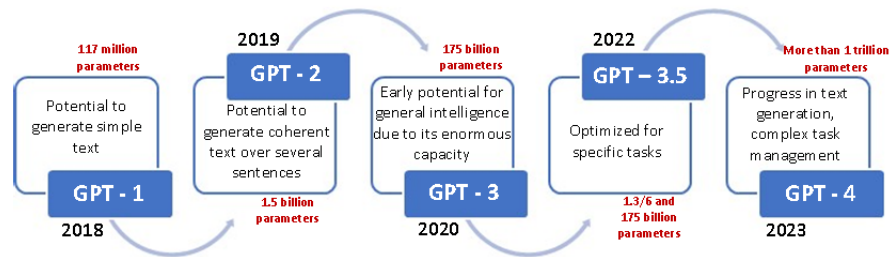


Figure 2.1. *Chronological progression of GPT*

The importance of GPTs lies in their breakthrough in natural language processing, enabling machines to understand and produce language with unparalleled fluency and accuracy. In sections 2.1.1.1–2.1.1.4, we look at the four GPT models, from the initial version to the latest GPT-4, evaluating their performances and their limitations.

2.1.1.1. GPT-1

In 2018, OpenAI introduced GPT-1, its first language model using the transformer architecture (Oxford Analytica 2023). With 117 million parameters, it marked a significant advance over previous state-of-the-art language models. GPT-1’s strength lies in its ability to generate a fluent and coherent language based on prompts or context. The model is based on a combination of two large datasets: Common Crawl, containing billions of words from web pages, and the BookCorpus dataset, made up of over 11,000 books covering different genres. These diverse training data have contributed to GPT-1’s robust language modeling capabilities.

However, despite its achievements in natural language processing, GPT-1 had certain limitations. Notably, it tended to produce repetitive text, particularly when given prompts beyond the scope of its training data. The model had difficulty reasoning about multi-turn dialogues and lacked the ability to capture long-term dependencies in text. In addition, its cohesion and fluency were limited to shorter

text sequences, and maintaining coherence in longer passages was challenging. Nevertheless, GPT-1 paved the way for more substantial and powerful transformer-based models.

2.1.1.2. *GPT-2*

In 2019, OpenAI introduced GPT-2 as the successor to GPT-1. This new version boasted an impressive 1.5 billion parameters, making it considerably larger than its predecessor. The training data for GPT-2 had also been improved (Zhang et al. 2021), combining the vast Common Crawl dataset with the diversified WebText dataset. One of GPT-2's strengths was its ability to generate coherent and realistic text sequences. Its ability to produce human-like responses made it a valuable asset for various NLP tasks, including content creation and translation.

However, GPT-2 had some limitations. It struggled when confronted with tasks requiring more complex reasoning and a deeper comprehension of context. While it excelled at generating short paragraphs and text snippets, it had difficulty maintaining context and coherence over longer passages. Despite these challenges, the development of GPT-2 laid the foundations for further advances in subsequent iterations of the GPT models.

2.1.1.3. *GPT-3 and GPT-3.5*

The year 2020 marked a significant leap forward in natural language processing models with the arrival of GPT-3. Boasting a staggering 175 billion parameters (Zhang et al. 2021), GPT-3 surpassed GPT-1 by more than 100 times and GPT-2 by more than 10 times. The GPT-3 training data come from various sources (Thorp 2023), such as BookCorpus, Common Crawl and Wikipedia, representing almost a trillion words. This extensive training has enabled GPT-3 to generate sophisticated responses in various NLP tasks without the need for prior data examples.

In particular, GPT-3 stood out from its predecessors in its ability to produce coherent text, write computer code and even create art. The comprehension of context and the generation of appropriate responses set it apart, making it ideal for applications, such as chatbots, content creation and translation. The emergence of ChatGPT, a conversational AI chatbot, illustrates the immense impact of GPT-3 in the field.

Despite its remarkable capabilities, GPT-3 was not without its faults. Biased, inaccurate or inappropriate responses could occur due to the model being trained on large quantities of text, potentially containing biased information. In addition, GPT-3 sometimes generated irrelevant text for a given prompt, indicating its difficulty in grasping context and background knowledge.

In addition, the impressive potential of GPT-3 raised ethical concerns and fears of misuse. Experts worried about its use for malicious purposes, such as the generation of fake news, phishing mails and malware. OpenAI also introduced an improved version, GPT-3.5, prior to the official launch of GPT-4.

Although it has the same number of parameters as GPT-3, GPT-3.5 benefits from training on a larger dataset of text and code (Sedkaoui 2023). These extended training data enable GPT-3.5 to generate even more realistic and complex text than its predecessor.

Furthermore, GPT-3.5 offers improved capabilities in performing various tasks without the need for additional training. While it cannot be considered a major upgrade over GPT-3, it does present a significant improvement in terms of potential performance and utility.

2.1.1.4. GPT-4

GPT-4, the latest addition to the GPT versions, was launched on March 14, 2023, representing a significant advance on its already impressive predecessor, GPT-3.5. Although details of its training data and architecture have not been officially disclosed, GPT-4 undoubtedly builds on the strengths of GPT-3.5, while addressing some of its limitations (see Appendix 2).

A remarkable feature of GPT-4 is its multimodal capabilities, enabling the model to understand input images alongside text prompts (Open AI 2023; Oxford Analytica 2023). For example, during the live broadcast of the GPT-4 launch, an OpenAI engineer fed the model with a hand-drawn website mock-up and, surprisingly, the model produced functional code for the website.

In addition, GPT-4 demonstrates better understanding of complex prompts and achieves human-level performance on various professional and traditional benchmarks (Sanderson 2023). It also features a larger window size and context, enabling the model to retain more information during conversations.

GPT-4 progress is pushing back the boundaries of AI tools and should find applications in various industries. However, as with any powerful technology, concerns about potential misuse and ethical implications surround this formidable tool.

2.1.2. ChatGPT in practice

The introduction of GPT models has brought about a revolutionary transformation in the field of AI, opening up a realm of exciting possibilities. Their large scale, capabilities and complexity make them extremely valuable for a variety

of applications. These GPT applications are diverse and impactful. Some key uses include:

- *Content generation*: GPT models can generate text for a variety of purposes, such as writing articles, blog posts, product descriptions and social media content, making it easy to create content quickly and efficiently.

- *Language translation*: GPT models are able to translate text from one language to another, using the patterns and structures they have learned during training to provide translation services.

- *Answering questions*: GPT models can process natural language questions and generate relevant and informative answers, making them useful for chatbots, customer support systems and information retrieval.

- *Text summary*: GPT models can produce concise summaries of longer text documents, which is useful for news articles, research papers or other large documents.

- *Dialog systems and chatbots*: GPT models, like ChatGPT, can power interactive and engaging chatbots and dialogue systems, understanding user queries and providing relevant assistance.

- *Creative writing*: GPT models can generate creative content, such as poems, stories or scripts, imitating the style and structure of different genres to help writers or provide entertainment.

- *Language modeling research*: the GPT models play a key role in research into language processing and understanding, serving as a reference for evaluating model performance and exploring new techniques.

- *Coding*: ChatGPT can perform coding tasks in languages (Tingiris and Kinsella 2021; Sedkaoui 2023) such as Python, JavaScript and SQL, simply by translating instructions into natural language. For example, it can interpret “extract the data from this API and display it in a table” to automatically generate code.

- *Data analysis*: ChatGPT can answer complex analytical questions about datasets and quickly generate insightful visualizations through conversational interactions. For example, when asked about patterns in a retail sales dataset, it provided a detailed histogram analyzing trends.

- *Customer service*: based on historical customer conversations, ChatGPT can automate common service requests, such as returns or delivery tracking, by programmatically generating human-like responses.

These examples only scratch the surface of ChatGPT’s capabilities. From coding to business applications, to creativity, ChatGPT’s versatility across all areas of

knowledge makes it a profoundly disruptive technology. As capabilities mature, ChatGPT can reshape industries from medicine to analytics and beyond. The versatility and advanced natural language processing capabilities of ChatGPT have made it a valuable tool in various fields.

Domain	ChatGPT applications	Example use cases
Health and medicine	Medical diagnosis assistance Personalized treatment plans Medical research summary and synthesis Mental health advice Clinical research and development	Developing chatbots for patient triage Recommendations for medical diagnosis and treatment Supporting patient engagement and treatment adherence
Business and finance	Automated financial reporting Customer feedback sentiment analysis Personalized investment recommendations Writing business content	Developing chatbots for customer service Market analysis and forecasts Bank fraud detection Financial risk management
Education and training	Personalized learning materials and lesson plans Real-time feedback and guidance for learners Interactive educational content generation	Teacher support for lesson plans and teaching strategies Language learning and test preparation Personalized online tutoring
Sales and marketing	Lead generation Chatbots for customer service Market analysis and forecasts Sales management support	Developing chatbots for customer service Market analysis and forecasts Marketing content creation Risk management for financial products
Bank	Chatbots for customer service Financial risk management Managing personal finances	Analysis and forecasts for investment Risk management for financial institutions

Table 2.1. Applications of ChatGPT in various fields

However, as with any technology, it is essential to recognize the risks and limitations. The remarkable ability of these models to generate highly realistic text

and code raises concerns about potential misuse, particularly in areas such as the creation of malware and the dissemination of disinformation.

Despite these concerns, as GPT models continue to advance and become more widely available, they are poised to play an important role in shaping the future of AI and NLP.

Having explored the origins and capabilities of ChatGPT, we now take a closer look at its versatile features for data analytics tasks through intuitive natural language conversations.

2.2. Powerful features for intelligent data analytics: natural language at your service

ChatGPT has a versatile set of features well-suited to accelerating and improving many facets of the data analytics workflow. By processing natural language requests and following analytical or coding instructions, ChatGPT can help analysts unlock deeper insights faster.

ChatGPT and other large NLP language models have shown potential for performing data analytics tasks, with performance comparable to, or even better than, that of human data analysts in some aspects. ChatGPT's strengths as a data analyst include:

- *Fast processing speed:* ChatGPT can perform data analysis tasks much faster than analysts, due to its AI-based nature. This can improve work efficiency and turnaround times. In the experiments reported in an article, ChatGPT took an average of approximately 50 seconds to generate an analysis for a data point, while analysts took a few minutes.

- *24/7 availability:* ChatGPT is available around the clock, with no breaks, which can help organizations that need to carry out analysis work at any time of the day.

- *Lower cost:* using ChatGPT, and GPT-4 in particular, would cost companies just 0.45% of the salary of a senior data analyst for a single analysis task, and 0.71% for a junior analyst (Cheng et al. 2023). In other words, the fractional cost of this AI agent compared to human expertise could generate significant savings on operational expenses related to data analytics. The productivity gains enabled by ChatGPT's conversational automation far exceed its cost of use, paving the way for cost-effective and scalable analysis.

- *Ability to generate a smooth and consistent analysis:* evaluations reported in the research results of Cheng et al. (2023) show that ChatGPT is capable of

generating data analytics points with great fluidity and alignment scores, comparable to those of data analysts. This indicates that GPT-4 can clearly and logically express insights from data.

For data scientists, ChatGPT opens up revolutionary possibilities as a natural language interface to accelerate analytical workflows. By conversing directly with analysts, ChatGPT can automate a wide range of repetitive tasks (Sedkaoui 2023; Shen et al. 2023).

Conversational capabilities	Generative capabilities
Data cleaning	Automated reports
Feature engineering	Data storytelling
Statistical analysis	Presentation development
Query generation	Dashboard annotation

Table 2.2. *Conversational versus generative capabilities*

These capabilities allow ChatGPT to perform various data preparation tasks, improving model performance, performing advanced analytical methods and translating analyst questions into optimized code. Additionally, ChatGPT can also summarize comprehensive reports and create compelling data stories.

In addition to conversational and generative capabilities, ChatGPT also unlocks considerable computational throughput, which is essential for complex analytical tasks. Its AI architecture offers speed and scalability beyond human cognitive limits (Tamkin et al. 2021; Sedkaoui 2023).

Beyond raw computation, ChatGPT also demonstrates a nuanced ability to recommend appropriate analytical approaches based on specified objectives. Its analytical knowledge enables three levels of guidance:

- Recommending statistical tests and models perfectly adapted to the characteristics of the available data and the insights required makes it possible to adapt the analysis to the problem.
- Suggesting creative new ways of combining, processing or visualizing data that may reveal non-obvious insights missed by current approaches stimulates the imagination.
- Identifying gaps in current analysis methodology and suggesting permutations or additional data needed for a more exhaustive investigation enables iterative refinement.

– By recommending statistical tests and models specifically tailored to the characteristics of the available data, ChatGPT offers data scientists the ability to perform more targeted and precise analyses. By identifying the analysis methods best suited to the problem at hand, ChatGPT saves time and maximizes the use of their resources.

– By suggesting creative new ways of combining, processing or visualizing data, ChatGPT opens up new perspectives and possibilities for data exploration. This innovative approach stimulates the imagination of data analysts by encouraging them to consider unconventional approaches to analyzing data and uncovering hidden insights that might otherwise go unnoticed.

– ChatGPT also plays a crucial role in identifying gaps in the current analysis methodology. By suggesting permutations or additional data needed to conduct a more exhaustive investigation, ChatGPT brings rigor to the data analytics process, which allows for iterative data analytics that progressively improves the quality and reliability of results.

What is more, ChatGPT's versatile nature as a programming and code debugging tool opens up a world of possibilities in a variety of fields. Thanks to its advanced natural language processing capabilities, ChatGPT can be used to streamline the coding process and improve software development efficiency. When it comes to code generation, developers can rely on ChatGPT to generate code snippets that perfectly match their needs, thus saving valuable time and effort. In addition, ChatGPT's code optimization capabilities are invaluable for refining algorithms and data structures, improving application performance and scalability.

When it comes to debugging, ChatGPT is indispensable. By analyzing code structure and error messages, analysts obtain valuable insights and recommendations, speeding up the identification and solving of problems. What is more, ChatGPT can help with code documentation, ensuring that software is well documented and easy to understand for future maintenance and collaboration.

Furthermore, ChatGPT's analytical prowess comes into play in the code review process. It can assess code against coding standards and best practices, offering constructive feedback to improve code quality and robustness. As industries rely increasingly on software solutions, ChatGPT's role as a competent programming and debugging tool is set to revolutionize software development, enabling developers to create applications that are more efficient, reliable and innovative.

ChatGPT's fine-tuning for language modeling tasks opens up a wealth of possibilities for data scientists. The model's ability to understand context and semantics enables it to create language models that excel in a variety of applications.

The integration of ChatGPT into conversational systems, chatbots and virtual assistants enables data analysts to develop interactive, context-sensitive AI agents. These intelligent agents can converse with users in natural language, answering queries, providing recommendations and assisting with tasks. The potential for personalized, human-like interaction opens up new possibilities for customer support and service automation.

ChatGPT can be refined to understand specific languages or domains, making it a valuable tool for tasks such as named entity recognition, part-of-speech tagging and sentiment analysis. For example, a data analyst can customize ChatGPT to perform sentiment analysis in the financial field, providing insights on market sentiments and trends.

ChatGPT is revolutionizing the NLP landscape, empowering data scientists with its versatile capabilities. From content generation to language understanding, to coding, ChatGPT is proving to be a game changer in various fields, promoting innovation and efficiency in AI-based applications. As the field of NLP continues to advance, ChatGPT is set to play an increasingly vital role in shaping the future of natural language processing and human–AI interactions. The responsible development and deployment of ChatGPT will be crucial to realizing its full potential, while ensuring that ethical considerations are respected in its applications.

2.3. ChatGPT versus data scientists: an intelligence battle that looks like an alliance

ChatGPT's natural language capabilities enable a human–AI partnership that reinvents the data analytics process. By collaborating seamlessly, ChatGPT and data analysts can complement each other to acquire more powerful insights. This vision promises to expand and elevate the role of the data analyst.

2.3.1. Seamless communication and key features of an effective partnership

At the heart of seamless teamwork between data analysts and ChatGPT is the natural language conversation capability, which facilitates a smooth and synergistic collaboration. Thanks to this intuitive interface, analysts can communicate their objectives, questions and prompts to ChatGPT in a simple conversational language, avoiding the need for complex code. This interactive dialogue allows ChatGPT to ask clarifying questions, seek clarification and ensure a precise understanding of assigned tasks. This human–machine collaboration reflects the natural exchange between colleagues working together to solve problems.

A key feature of this collaboration is the ability for data analysts to provide real-time feedback as ChatGPT works to refine its performance. This interactive learning allows ChatGPT to integrate human advice to continually improve its outputs. ChatGPT can also explain its actions and suggest analytical steps for validation, thus reinforcing transparency and trust in the analytical process (Sedkaoui 2023).

Thanks to this collaboration, analysts can exploit ChatGPT's capabilities on-demand to speed up the acquisition of meaningful insights. However, it is essential to remain vigilant against risks, including bias, as AI has no inherent sense of ethics. Responsible coordination is needed to ensure that the benefits of this collaboration outweigh the risks.

For this collaboration to be truly symbiotic, certain key attributes must be respected. Clear communication of objectives, constraints and deadlines establishes alignment, as ChatGPT depends entirely on human instructions for direction. Full transparency about ChatGPT's capabilities and limitations is crucial to build trust and set appropriate expectations.

Responsible human surveillance is non-negotiable to ensure the safety and ethics of the analytical process. Humans must maintain control and oversight of all analytical pathways, model assumptions and decisions.

Regular feedback from analysts is essential to iteratively improve ChatGPT's work and mitigate risks, as AI systems require ongoing guidance.

Finally, ethical alignment on priorities such as data privacy, algorithmic fairness and transparency is essential if ChatGPT is to empower data analysts responsibly.

With thoughtful coordination and responsible implementation of these principles ChatGPT can work in close symbiosis with data scientists, becoming a seamless extension of the team. This human-machine collaboration promises to significantly amplify human potential in the field of NLP.

The effective integration of generative AI, such as ChatGPT, into data analytics fundamentally redefines the role of the human data analyst. Thanks to the automation of mechanical tasks by AI, data analysts can now focus more on higher-order tasks.

The analyst's role before generative AI	The analyst's role with generative AI
Performing analyses	Setting strategic orientations and analytical plans
Identification of standard questions	Creative identification of new data connections
Rigid planning	Dynamic adjustment of the analytical direction
Data processing and execution	Critical evaluation and interpretation of model results
Communicating findings	Communication of stakeholders' needs

Table 2.3. *Evolution of the data analyst's role with generative AI*

As illustrated in Table 2.3, data analysts can devote more time to setting strategic orientations for analytical plans and workflows. They can be freed from execution and focus on creative thinking and strategic planning, elevating their role within the organization.

Generative AI inspires curiosity by identifying new data connections and asking new investigative questions. This leads to breakthroughs and discoveries that might previously have been missed. So, with access to AI-accelerated insights, data analysts can now dynamically adjust analytical direction based on discoveries in the process. This makes data analytics more adaptive and responsive to changes.

With AI taking charge of task execution, critical evaluation and interpretation of model results become the main added value of the analyst. Human judgment comes to the fore, enabling a deeper understanding of the results. With generative AI, analysts can now focus on communicating findings and insights to stakeholders, reinforcing their role as key communicators within the company.

As a powerful and innovative assistant, ChatGPT positions itself as a valuable ally and facilitator in the data analytics process rather than a threat that could replace the data analyst. In the future, its effective integration into the field of analytics aims to fundamentally redefine the main responsibilities of the human analyst, freeing them from mechanical tasks and enabling them to concentrate on higher-added-value intellectual activities.

2.3.2. ChatGPT, between ally and threat

So far, we have focused on the collaborative aspect of ChatGPT and the LLMs. These models can automate some of the tasks that data analysts currently perform, such as data cleaning and exploratory data analysis. This could lead to a reduction in the demand for data specialists in these areas.

The emergence of model-based machine learning (LLM) like ChatGPT is poised to significantly reduce the barriers to entry in data science and analytics (Sedkaoui and Khelfaoui 2020). In the future, there will be less need to master different software, as we will be able to exploit the power of LLMs to accomplish tasks, which previously took hours, in minutes.

Let us take the example of a company wishing to quickly extract and analyze its data stored in Excel sheets. Thanks to a ChatGPT–Excel integration, the company can simply highlight the cells to be analyzed and ask questions such as: “What is the trend in sales figures over the last quarter?”; or: “Can you perform a regression analysis?” The integration of products in this way will improve accessibility to Excel and similar software for people who do not usually use them, which will ultimately lead to decreasing the demand for experts proficient in these tools.

ChatGPT’s code interpreter plugin illustrates the growing democratization of data science workflows. With this plugin, we can run Python code and perform data analytics in the chat interface. We can also download CSV files and take advantage of ChatGPT’s support to clean up, analyze and create statistical models with our data.

Once we have analyzed the data and specified our objectives, ChatGPT will guide us through the necessary steps to achieve the desired outcome. It will then proceed with the analysis and modeling, providing explanations at every stage of the process. For example, ChatGPT can determine the optimal parameters to generate forecasts with ARIMA. While a data analyst would typically take approximately three to four hours to complete this task, ChatGPT can do it in a matter of minutes by simply processing the data provided.

This remarkable achievement considerably reduces the level of expertise required to facilitate the model building process. In this case, we can ask ourselves the following question: Is human expertise in this field still necessary? The answer is: absolutely.

Despite advances in AI capabilities for coding and model building, human experts remain essential to oversee the process. Although LLM models have advanced, they can still generate incorrect codes and make erroneous decisions when

creating statistical models. As a result, companies will always need qualified professionals with statistical and programming expertise to oversee the data analytics process and guarantee accurate model prompts.

LLMs, like ChatGPT, are not capable of creating data products in their own right, so humans will always be needed for tasks such as gathering requirements, debugging and model result validation. However, the efficiency gains generated by these natural language models will lead to a reduction in the number of personnel needed to perform these tasks.

The impact of this development may first be felt in entry-level big data analytics jobs, as LLMs can already master intermediate-level coding and analytical workflows (Hilton et al. 2021; Tamkin et al. 2021).

However, data scientists will still be needed to develop and train these models, as well as to interpret their outputs and communicate them to stakeholders. Their role encompasses much more, including data preparation, analysis, visualization and model building. Furthermore, they will always be needed to perform tasks that these models cannot, such as experimental design and data collection.

According to data from the US Bureau of Labor Statistics (2022), the demand for data scientists is projected to surge 22% by 2030, well above the average growth rate for all professions. This increase is driven by the growing use of data within companies and organizations of all sizes.

In today's landscape, data scientists are in high demand due to the diversity of skills they are expected to possess (McKinsey 2023). In addition to expertise in statistical modeling and coding, experts in this field must also master SQL for data extraction, work with tools such as Tableau and PowerBI for data visualization, and communicate insights effectively to stakeholders.

So, it is true that ChatGPT is a valuable tool for facilitating certain specific tasks in the data analytics process, such as data cleaning and pre-processing. However, its capabilities are limited, and it cannot completely replace the role of a data analyst. Data analytics, big data and data science are multifaceted fields (McAfee and Brynjolfsson 2011; Sedkaoui and Khelfaoui 2020) that require a mix of technical expertise, domain knowledge and critical thinking, making complete automation difficult.

Although ChatGPT improves efficiency in some areas, it cannot replace the indispensable nature of human expertise and judgment in data analytics, model building and interpretation and decision-making processes. Additionally, data analytics is a constantly evolving field, continuously introducing new technologies

and methods. Human adaptability and creativity remain essential if we are to exploit these advances effectively. In this dynamic landscape, ChatGPT, or any other generative AI tool, will complement, but not replace, the invaluable role of a skilled data analyst (Sedkaoui 2023).

Roughly speaking, the implementation of AI could indeed lead to a displacement of jobs in specific areas, but it will also create new opportunities in areas such as data analytics, model development and maintenance. Moreover, improvements in AI productivity in various industries can fuel economic growth and open doors to new job prospects.

As technology advances, it will be crucial to adapt and improve skills to exploit the advantages of AI while preserving the human traits that cannot be reproduced by machines. It is in this context that plugins are formidable allies, enhancing the basic capabilities of models such as GPT in data analytics.

2.3.3. Automating data science workflows: unleashing the potential of ChatGPT plugins

ChatGPT captured our hearts and minds as an adorable AI prodigy – eager to learn, endlessly curious, bursting with potential. Like a bright toddler continually absorbing new skills, ChatGPT has demonstrated a prodigious capacity for knowledge. But how can we nurture this potential from promise to capability – from toddler to sage?

The answer lies in augmenting its capabilities via powerful complementary modules, which can act as additional tutors, instilling new talents into ChatGPT through hands-on learning. This is the remarkable transformation that occurs when ChatGPT meets its dynamic companions: plugins.

You may already be familiar with this term, which you have probably come across on the previous pages of this chapter. Plugins are third-party integrations that extend the capabilities of the basic ChatGPT model in a powerful way, tailored to data science workflows. At its core, ChatGPT has strong natural language processing skills, but lacks direct access to data or specialized analytical tools. Plugins fill this gap.

Think of plugins as additional accessories integrated into a powerful basic system, similar to the way Tony Stark's basic Iron Man suit integrates extra modules such as JARVIS or rocket thrusters when needed for specialized tasks. In the same vein, plugins augment ChatGPT's fundamental intelligence with extended functionality.

They work by linking ChatGPT conversations to external data sources, analysis libraries, visualization engines and more. This offers the best of both worlds – the conversational, generative nature of ChatGPT and the programming capabilities combine into a simple yet powerful data analysis experience.

For example, the Wolfram/Alpha plugin performs expert-level statistical analysis via conversational prompts.

Code-type plugins, such as Code Interpreter, execute blocks of code in Python and other languages directly into chats. Web integrators, such as WebPilot, retrieve sites or import APIs according to conversational instructions. Zapier channels data between thousands of applications to automate workflows.

Plugin	Functionality
Data exploration assistant	Enables interactive data exploration via natural language
Code assistance	Provides intelligent code suggestions for data science
Data visualization wizard	Generates data visualizations from conversational prompts
Machine learning advisor	A guide to the creation and evaluation of machine learning models
Data cleaning expert	Identifies and addresses data quality issues
Statistical analysis companion	Provides statistical expertise for rigorous analysis
Data storyteller	Generates compelling data-driven stories
WebPilot	Navigates websites and extracts data
ScholarAI	Finds and summarizes academic articles
Wolfram	Accesses the Wolfram/Alpha knowledge base
Code interpreter	Executes and interacts with code snippets
ChatWithGit	Enables Git-based version control and collaboration
LinkReader	Summarizes articles and web resources
Noteable	Takes, stores and organizes notes
ShowMe	Generates diagrams and graphs from text descriptions
Zapier	Connects and automates tasks between applications

Table 2.4. *List of ChatGPT plugins for data analytics*

So, if the ChatGPT base is the curious child, the plugins are the teachers that cultivate its capabilities. These robust extensions elevate data analytics to unprecedented heights, making them an invaluable asset you will not want to overlook.

As new plugins emerge, ChatGPT can intelligently interface with virtually any system through a natural dialogue (Chen and Li 2023). Analysts benefit from immense analytical versatility without programming or technical expertise.

Now, imagine these plugins integrated into future versions, such as GPT-4. The base model itself will have greater analytical prowess. Plugins will grant it domain-specific superpowers – like Tony Stark preparing for a mission with different accessories.

GPT-4, combined with SQL Assistant, becomes an unbeatable conversational database query engine. Plotly reinforces its keen sense of visualization. The execution of the code interpreter enables programming on the fly. Together, they form a formidable analysis team – the flexibility of human conversation with an AI assistant equipped with specialized tools.

This is the real promise of the plugin ecosystem – transforming ChatGPT from a narrow AI into a versatile analytical collaborator. As the plugins enhance its data analysis capabilities, we move closer to true augmented intelligence that empowers every user. The future holds exciting possibilities as AI assistants acquire extended functionalities through human-machine teamwork.

2.4. Benefits and challenges of integrating ChatGPT into data analytics workflows

ChatGPT integration promises huge productivity benefits for data analysts, as well as raising important ethical considerations. By striking the right balance between the two, data scientists can strategically leverage ChatGPT to improve their analytical processes.

2.4.1. Unlocking analytical potential and reducing costs

With its potential to automate a variety of repetitive, natural language-coded tasks, ChatGPT paves the way for considerable productivity gains in data processing. Examples show that data cleaning and manipulation can be accelerated by up to five times compared with current methods (Sedkaoui 2023). Similarly, ChatGPT is able to translate conversational natural language questions into SQL, Python or other code in a matter of seconds, eliminating hours of manual work.

These impressive advances do not mean, however, that ChatGPT can completely replace the role of data scientists. Despite its growing capabilities, this generative AI cannot yet understand context and meaning in the same way as human beings. Tasks that require critical reasoning, creativity or human interaction remain beyond its current capabilities. Consequently, collaboration between AI and data scientists is essential to make the most of this technology while avoiding potential errors.

The benefits of this collaboration are numerous. By offloading repetitive and automatable tasks, ChatGPT frees up analysts' time, so they can focus on higher added-value tasks. For example, automated reports enable analysts to focus on interpreting results and finding unique insights, rather than spending time on visualizing and assembling dashboards. Automation of queries provides interactive exploration of large datasets, which was previously difficult to achieve.

In addition to productivity gains, ChatGPT considerably extends analytical scope. Thanks to its ability to rapidly translate natural language questions into high-performance code, it becomes possible to query massive datasets in real time. New correlations and metrics can be uncovered by ChatGPT during real-time conversations, paving the way for further exploratory analysis. Predictive modeling and large-scale scenario analysis also become feasible thanks to the accelerated execution of generative AI.

However, it is essential to note that generative AI remains imperfect and requires a responsible approach to its use. The implementation of rigorous verification, monitoring and governance is essential to avoid potential errors and ensure ethical use of this technology. Human–AI collaboration is essential if ChatGPT is to be integrated responsibly into data analytics processes.

In addition to productivity gains and the expansion of analytical possibilities, the automation workflow with ChatGPT also promises significant cost and resource savings for organizations. The time spent on reporting, visualization and dashboarding can be considerably reduced thanks to ChatGPT's automated storytelling and synthesis capabilities. This time-saving presents thousands of precious human hours saved.

In addition, the overall operating costs of data analytics could be reduced through the efficiencies gained by using generative AI. This could enable organizations to reduce the size of their analyst teams while maintaining the same workload. Dependence on scarce and expensive data science talent could also be reduced thanks to ChatGPT's assistance, which enables business analysts to manage more advanced workflows via its natural language interface.

While incredibly promising, thoughtful management of ChatGPT integration is crucial to ensure that the benefits outweigh the potential drawbacks. The way forward requires a delicate balance between human and artificial intelligence in data analytics. Several challenges need to be addressed to exploit its full potential impact. These challenges include:

- *Bias and fairness*: generative models, including ChatGPT, are likely to reflect biases present in the training data. Analysts must be vigilant and carefully review the generated outputs to ensure that sensitive attributes, such as gender, ethnicity or age, do not influence the results. The presence of bias can have a significant impact on decision-making, leading to undesirable consequences, such as inequalities or discrimination. To ensure fair analysis, efforts should be made to mitigate potential bias and improve the diversity and representativeness of training data.

- *Explainability*: the opacity of the neural networks used in generative models can be a challenge in terms of explainability. Analysts and users need to understand how and why results are generated, so that they can interpret and explain the information produced by ChatGPT. Research is underway to develop methods that make generative models more transparent and interpretable. Explainability is essential to establish trust in the results of generative AI and for decision-makers to be able to make informed decisions based on the information produced.

- *Information security*: generative models, especially when used in sensitive areas or with confidential data, must be subject to rigorous access controls. It is crucial to prevent privacy violations and data breaches by ensuring that only authorized persons have access to information generated by ChatGPT. Data protection is essential to maintain the trust of stakeholders, be they customers, employees or the general public.

- *Redefining roles*: the introduction of generative AI in the data analytics process may lead to a redefinition of data scientists' roles. They will need to acquire new skills to supervise and collaborate with generative models, in addition to their traditional programming. This may involve rapid model engineering, interpreting generated output and solving problems related to generative AI. This redefinition of roles will enable analysts to collaborate more effectively with AI technologies, and to integrate them responsibly into their workflows.

- *Monitoring model errors*: like all AI systems, generative models can produce errors or unexpected results (Liebrenz et al. 2023). Given that ChatGPT can generate misconceptions or hallucinations, robust testing and monitoring mechanisms are needed to guarantee the reliability of the output. Experts need to be aware of the model's limitations and ensure that results are verified and validated before making important decisions based on the generated information.

– *Intellectual property protection*: the use of generative AI raises concerns about the protection of intellectual property. When generative models are used to generate proprietary analytics, it is possible for competitors to reproduce these analyses by training their own models. It is therefore essential for organizations to put appropriate protection measures in place to safeguard their intellectual property and proprietary analytics.

As a result, while generative AI offers tangible benefits, it also introduces risks for data analytics. To realize its full potential, organizations must adopt comprehensive governance frameworks, foster human surveillance and promote collaboration between analysts and technologists. Responsible implementation of these technologies can significantly accelerate the ability to extract valuable insights from large datasets, improving decision-making at all levels.

It is important to note that while generative AI contributes to automation, it does not replace the essential role played by human analysts. Analysts provide judgment, context, quality control and direction for generative AI. Collaboration between humans and machines takes advantage of their respective strengths, leading to more impactful and informed data analytics.

We now move on to Chapter 3, which takes a closer look at how ChatGPT's capabilities can be used to speed up critical data preparation tasks, such as cleaning, pre-processing and feature engineering.

Data Preparation for Analysis with ChatGPT

Data preparation is a critical and very important step in the data analysis process. It involves cleaning, transforming and formatting data, so that they can be analyzed efficiently. This task can be time-consuming and challenging, but it is essential to ensure the accuracy and reliability of data analytics results.

ChatGPT is an LLM that can be used to understand and respond to natural language queries. ChatGPT is a powerful tool for data preparation, to automate tasks such as loading, cleaning, and splitting datasets; identifying and removing missing or outlier data; data normalization; text processing; and the creation of new features. This facilitates the process of data preparation for analysis.

Using this generative AI model, you can automate many of the tasks involved in data preparation, freeing up your time to focus on other aspects of the data analytics process.

In this chapter, we provide a detailed overview of how to use ChatGPT for data preparation. We start by discussing the basics of data preparation, then show you how to use ChatGPT to automate the various tasks. We also provide some tips and tricks for using ChatGPT effectively.

3.1. ChatGPT in charge of preparing our datasets

Data preparation, as a crucial step in the data analysis process, is the key to unlocking the potential of data analytics. This involves a series of tasks aimed at ensuring that data are in a clean, usable and structured format. By cleaning,

transforming and formatting data, data analysts can prepare them for analysis and extract valuable insights.

Data cleaning is the initial and fundamental task in data preparation. It involves identifying and rectifying any errors or inconsistencies in the data. These may include typos, missing values, duplicate records or outliers. Addressing these issues is essential to guarantee data accuracy and reliability. Incorrect or missing data can have serious repercussions on the results of data analytics and decision-making processes.

Data transformation then comes into play. This stage consists of converting the data into a format suitable for analysis. For example, converting categorical data into numerical values enables calculations and modeling. Scaling data to a consistent range ensures that variables at different scales do not disproportionately influence the analysis.

Data formatting focuses on organizing data in a way that is easy to understand and analyze. The creation of summary tables and the generation of logical data structures are all part of data formatting. This process improves data accessibility and enables patterns, trends or anomalies to be identified quickly.

Although data preparation can be time-consuming and challenging, it is a critical process that has a significant impact on the quality and accuracy of the insights derived from data analysis. Good data preparation ensures informed decision-making based on reliable information (Davenport and Harris 2007).

To ensure efficient data preparation, several principles must be followed. Beginning with a clear understanding of data objectives and potential problems helps to focus efforts. Adopting a systematic and consistent approach, using appropriate tools such as Excel, Python or R, ensures repeatability and reliability. Testing prepared data against known values or running different analyses helps to verify the accuracy of results.

While it is essential to adhere to these principles, data preparation can still be a challenging process, especially when dealing with large, big data-type, and complex datasets (Sedkaoui 2018a; McKinsey 2023). By using ChatGPT to automate data preparation tasks, we can save time and improve the accuracy and reliability of our results. This generative AI model can be used to assist with data preparation in several ways.

3.1.1. Data cleaning

By using ChatGPT, we are able to identify errors in our data by asking simple questions in natural language. For example, we could ask ChatGPT things like: “Are

there duplicate values in this column?"; or: "Are there typos in this column?" ChatGPT will spot these potential errors and point them out, enabling us to quickly correct problems in our data.

Furthermore, ChatGPT can also help us eliminate duplicate data by allowing us to easily identify and delete rows that are exactly the same. For example, we can ask ChatGPT: "Remove all duplicate lines from this dataset". ChatGPT will then perform this cleaning task.

In addition, ChatGPT offers the ability to fill in missing values in our data. If we have empty cells in the "age" column, for example, we simply ask ChatGPT: "Fill in the missing values in the 'age' column with the average age of all rows". And ChatGPT will perform this task intelligently, calculating the average of the available ages and replacing it in the empty cells.

3.1.2. Data transformation

With the help of ChatGPT, we can transform categorical data into numerical data by asking it to identify and label the various categories. For example, simply ask ChatGPT: "What are the categories in the 'gender' column?"; then ask it: "Attribute a number to each category in the 'gender' column". The generative AI model will then take over this task by associating numerical values with each category, making the categorical data usable for numerical analysis. This transformation facilitates data manipulation and interpretation, enabling statistical operations and predictive models on previously categorical variables.

ChatGPT makes it possible to scale data by asking it to identify and perform this operation, so that all values are harmonized on the same scale.

For example, simply ask the question: "Scale the data in the 'age' column so that it lies between 0 and 1", and ChatGPT will take over this task to normalize the data, making it easier to compare and coherently analyze. This enables heterogeneous datasets with different value ranges to be processed efficiently and harmoniously.

3.1.3. Data formatting

Generating summary tables is child's play with ChatGPT's capabilities. We can simply ask ChatGPT to identify and summarize the data in a table. For example, "Create a table summarizing the data in the 'age' column". In response, ChatGPT will quickly process the data and provide us with a well-organized summary table, making data presentation and analysis more efficient.

To visualize data, ChatGPT comes to the rescue once again. It can identify and visualize data in a table, and it can also suggest relevant visual presentations. For example, we can ask: “Which graph will best present the data in the ‘gender’ column?” ChatGPT will quickly generate a list of charts corresponding to this type of data.

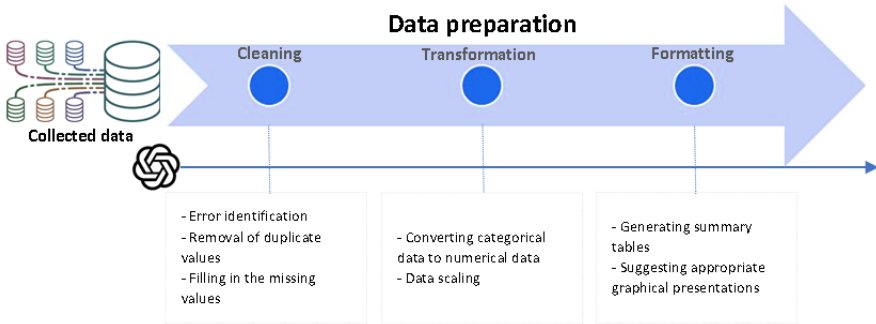


Figure 3.1. *ChatGPT's role in data preparation tasks*

However, it is important to remember that ChatGPT is trained on a massive dataset of text and code. This means that if input data from ChatGPT are inaccurate or incomplete, output data from ChatGPT may also be inaccurate or incomplete. To avoid this, it is important to carefully prepare the input data for ChatGPT. This includes identifying and correcting errors or missing values, and ensuring that data are formatted in a way that ChatGPT can understand.

By following these procedures, we can guarantee the accuracy and reliability of the data produced by ChatGPT. This will avoid the GIGO (Garbage In, Garbage Out) principle, which emphasizes that the quality of the output of an IT program depends on the quality of the input data, and thus ensure the success of our data preparation activities.

This principle is particularly important in the context of data preparation, where the quality of input data can have a significant impact on the accuracy and reliability of the results (Hazen et al. 2014; Sedkaoui 2018a). By performing careful preparation, we will be able to take full advantage of ChatGPT's capabilities and obtain good results for our analyses and decision-making.

3.2. Automated cleaning and pre-processing for optimum results with ChatGPT

Data preparation is usually a time-consuming and difficult task, but it is essential to ensure the accuracy and reliability of our results. Data cleaning involves a more comprehensive set of tasks, encompassing not only the data cleansing, but also the identification and elimination of duplicates, incomplete records and irrelevant data.

By taking the time to prepare our data correctly, we can improve the efficiency and accuracy of the data analysis performed and ultimately make better decisions. However, time efficiency is also essential, and technological advances, such as ChatGPT, enable rapid data preparation. By generating large volumes of synthetic data, data scientists can improve the efficiency of their models, reduce the time and costs associated with data collection and bypass the privacy and security risks inherent in using real data (Lucini 2021; Martineau 2022).

By exploring these technologies to automate data preparation steps such as cleaning and pre-processing, we can not only save valuable time, but also process larger volumes of data in less time.

By automating this essential task, companies can optimize their resources, boost the accuracy of results and guard against potential errors, marking a new era in data preparation and analysis.

Automated data cleaning and pre-processing have become increasingly vital in data science projects. This essential step implies laborious and time-consuming manual tasks. However, a breakthrough in natural language processing (NLP) comes in the form of ChatGPT, a cutting-edge tool that offers the potential to automate data cleaning and pre-processing, revolutionizing the way we manage these critical tasks.

ChatGPT is an invaluable aid in the data cleaning process, a crucial task for improving data quality. Often, large datasets contain errors that need to be identified and rectified before accurate analysis. Drawing on its substantial language model, ChatGPT has the ability to detect inconsistencies and erroneous values in data, dealing with problems such as missing data, formatting discrepancies and inaccuracies. By effectively identifying a range of data anomalies in various fields, ChatGPT speeds up the detection process.

In addition, it provides actionable suggestions for error resolution, streamlining data cleaning efforts and reducing the time spent on manual review. This automation of data validation and correction reduces the time spent on repetitive tasks, optimizing resources for more strategic analysis and insights.

This intelligent technology cleverly detects typos, misspellings, punctuation errors and other inconsistencies that may plague data. What’s more, it goes beyond simple error identification, identifying patterns in the dataset and even suggesting appropriate corrections.

Automated cleaning and pre-processing data cleaning with generative AI tools, such as ChatGPT, also offer a consistent and reproducible approach, eliminating the human variations that might arise during manual tasks. This ensures that our analyses can be easily reproduced for validation and verification. Additionally, this automation frees the data scientists from repetitive and time-consuming tasks.

Thanks to its advanced features, ChatGPT can identify and process outliers in the data, ensuring the veracity of the final results. The incorporation of ChatGPT into data cleaning procedures offers significant benefits, including time and cost savings by reducing the need for extensive manual work and increasing the accuracy of results. In addition, it helps to mitigate the risks associated with data errors, safeguarding the integrity of the analysis.

In essence, ChatGPT’s involvement in data cleaning improves data accuracy and quality, for more reliable analyses. Table 3.1 illustrates how ChatGPT makes data cleaning and pre-processing easier and more efficient.

Feature	Description
NLP language processing	Uses advanced NLP technology, allowing users to interact with the tool using natural language
Error detection and correction	ChatGPT’s machine-learning algorithms automatically detect and correct errors, such as typos and duplicates
Imputation of missing values	Identifies and fills in missing values, reducing the need for manual data entry
Data type inference	Deduces data types, guarantees formatting and compatibility for analysis
Noise removal	Detects and eliminates irrelevant words, phrases and out-of-context data, improving data quality
Pattern identification	Identifies patterns indicating errors or inconsistencies, thus improving data accuracy
Report generation	Generates comprehensive reports summarizing data and identifying potential issues for prompt resolution

Table 3.1. *How ChatGPT improves cleaning and pre-processing of data*

However, it is important to note that automation does not entirely replace human intervention. Human expertise remains necessary to understand the context of the analysis, select the right data preparation methods and correctly interpret the results obtained. Collaboration between AI and data experts is therefore essential to take full advantage of these technological advances and achieve high-quality analyses.

Beyond accuracy, automation also offers considerable time and effort savings, a crucial advantage when it comes to processing large volumes of data. Manual cleaning and the pre-processing of large quantities of data can be tedious and time-consuming, but automation streamlines these tasks, freeing up valuable time for more strategic analyses (Mollick 2022; Sedkaoui 2023).

Consistency is another essential aspect that automation improves. This consistency is invaluable for maintaining the integrity of data-driven analyses. In addition, the automated process enables effortless adaptation of data to various analyses, eliminating the need for manual adjustments each time a new analysis is performed. This flexibility enhances the agility of data analytics, allowing different approaches to be explored quickly, and more in-depth insights to be obtained from the data.

In addition to these advantages, automation of cleaning and pre-processing also improves data quality. Automated tools can identify and rectify errors and inconsistencies resulting in data that are better suited to analysis. When considering automated cleaning and pre-processing, several factors need to be taken into account. The type of data being analyzed is an important consideration, as different types of data may require specific cleaning and pre-processing approaches. For example, textual data may require different processing from digital data.

In addition, the objectives of the analysis influence the choice of cleaning and pre-processing methods. The objectives of the analysis determine the most appropriate approach to data processing, whether it is a question of pattern identification or predictive modeling. Resource availability is another crucial factor. Limited resources, such as time and computing power, can influence the choice of a less complex solution.

In this context, a clear strategy needs to be mapped out, clarifying how to automate cleaning and pre-processing for optimal results with ChatGPT. To begin with, it is essential to have a clear understanding of the data, including the objectives of the analysis, the type of data we have and potential data problems. This understanding will help us to choose the right data preparation methods and to use ChatGPT effectively.

So, taking into account that when we interact with ChatGPT, it is best to use natural language. Given its in-depth training on a vast dataset of text and code, ChatGPT is better at understanding natural language questions than technical jargon. Being precise in our questions increases the likelihood of getting accurate and useful answers. Instead of asking a general question like “How do I clean up my data?”, it is better to ask: “How do I remove duplicate rows from the ‘age’ column?”

An interesting feature of ChatGPT is its ability to generate code that can be used to automate the data cleaning and pre-processing. By taking advantage of this feature, we can save a lot of time and effort in the preparation process. However, it is essential to test the results obtained. This validation process ensures that the results are accurate and reliable.

In addition to the main tips, consider using a consistent denomination convention to help ChatGPT better understand the data and generate the appropriate code. Incorporating comments into the code will also help to document steps and make them more understandable to others. The use of version control is another useful practice, as it makes it possible to track changes made to the code and revert to previous versions if necessary.

Automated cleaning and pre-processing can help guarantee accuracy and reliability. This is particularly important for machine learning and AI applications, where even small errors in the data can have a significant impact on the results. However, it is important to automate the process of cleaning and pre-processing so as not to introduce new errors. This means that the process must be able to identify and handle missing data and outliers.

3.3. Handling missing data, outliers and other common data issues

The management of missing values and outliers is an essential aspect of data preparation as these factors can have a significant impact on the accuracy and reliability of the data analysis.

Missing values are data points that are missing from a dataset and can occur for a variety of reasons, such as human error, data corruption or incomplete surveys. On the other hand, outliers are data points that deviate significantly from the rest of the data, and they can be a sign of errors or issues in the data or be legitimate data points that fall outside the range.

To address missing values and outliers, several approaches can be used, depending on the specific dataset and the analysis to be performed. To handle missing values, options include deleting rows or columns containing missing values,

imputation of missing values with estimated values using different imputation methods or processing outliers as missing values, which means ignoring them or handling them as missing values. Alternatively, it may be useful to define outliers and adjust the analysis accordingly, when outliers are identified and the analysis is modified to account for them.

Appropriate management of missing data and outliers is essential to ensure the accuracy and reliability of data analysis, as this can lead to better results and more meaningful insights. This is particularly important in decision-making processes that rely strongly on data-driven information (Chen et al. 2012; Sedkaoui 2018a).

Furthermore, there is no single solution for dealing with missing values and outliers. Data scientists must carefully consider the specific characteristics of the dataset and the analysis requirements before selecting the appropriate approach. By thinking through the approach, they can make informed decisions and avoid potential pitfalls in the data analytics process.

Testing results after managing missing data and outliers is also crucial. This helps to validate the data preparation process and ensure that the analysis produces reliable and meaningful results. Comparing results with known values, or performing different analyses on data, may reveal discrepancies or errors that need to be corrected.

ChatGPT, with its powerful natural language processing (NLP) capabilities, can play an important role in managing missing data, outliers and other common data-related issues. Table 3.2 illustrates how ChatGPT can assist in each of these areas.

ChatGPT makes it easy to process missing data by identifying missing values in a dataset using generated code, enabling analysts to understand the extent of such missing data. Depending on the dataset and analysis objectives, ChatGPT can also generate code to impute missing data using various statistical techniques, or suggest processing as a separate category for categorical data. In addition, ChatGPT helps manage outliers by identifying data points that deviate significantly from the expected range and generates code to remove, replace or transform them, thus ensuring that outliers do not unduly affect data analytics.

Beyond missing data and outliers, ChatGPT can provide code solutions for a variety of other data-related issues. These include data cleaning tasks, such as correcting typos, deleting duplicates and formatting inconsistencies, which are essential for ensuring accuracy.

Domain	How?	
	Identification	Processing
Missing data	Identify missing data by generating a code to analyze the dataset and identify areas with missing values	Generate code to process missing data to suit the specific dataset and analysis objectives Suggest using imputation with mean, median or other statistical techniques to replace missing values Suggest deletion of rows containing missing data
Outliers	Generate code that calculates the standard deviation of the dataset and identifies data points that deviate significantly from the expected range	Generate code to remove outliers Replace with estimated values Transform data to mitigate their impact

Table 3.2. *Processing and identification of missing data and outliers with ChatGPT*

Managing text cleaning and encoding is an essential part of data preparation, especially when working with textual data. ChatGPT can be an invaluable aid in addressing these current data-related issues.

ChatGPT can help with text clean-up by generating code to remove unwanted characters, punctuation marks and special symbols from textual data. It can identify and correct spelling mistakes, standardize capitalization and handle abbreviations and contractions. For example, if the dataset contains text responses from surveys or social media posts, ChatGPT can help normalize the text and make it more consistent for further analysis.

Text clean-up also involves removing empty words (common words such as “and”, “the”, “in”, which have no significant meaning) and applying r stemming or lemmatization to reduce words to their basic forms. ChatGPT can generate code to perform these operations, which can help improve the efficiency and accuracy of text analysis tasks such as sentiment analysis or topic modeling (Sedkaoui et al. 2022; Susnjak 2023).

When working with textual data, it is often necessary to convert the text into a digital format so that the machine learning algorithms can process the data efficiently. ChatGPT can suggest different encoding techniques to achieve this. A common approach is to use one-hot encoding where each word in the text is presented as a

binary vector. For example, if the dataset contains categorical features such as “country” with values such as “USA”, “UK” and “Canada”, ChatGPT can generate code to convert these categories into binary vectors (e.g. [1, 0, 0], [0, 1, 0], [0, 0, 1]).

Another popular method is word embedding, which presents words as dense vectors in a continuous space. ChatGPT may suggest using pre-trained word embeddings such as *Word2Vec*, *GloVe* or *FastText* to encode textual data, capture the semantic relationships between words and improve the performance of natural language processing tasks.

In addition, ChatGPT can help manage the encoding of ordinal and nominal categorical features. It can generate code to convert ordinal categories into numerical values while preserving their order, and suggest encoding for nominal categories.

ChatGPT can also help normalize data, by converting different data scales into a common scale, which is essential for some machine learning algorithms and comparisons between different datasets. By generating code to organize and structure data in tables or other formats, ChatGPT enables analysts to explore and analyze data more efficiently.

Integrating ChatGPT into our data analytics toolbox not only speeds up the processing of large volumes of information, but also improves our ability to recognize hidden patterns, trends and connections that might otherwise go unnoticed. Thanks to its natural language understanding capabilities, ChatGPT helps convert complex data into precious insights, enabling informed, data-driven decisions to be made with confidence.

3.4. Using ChatGPT for data transformation, feature engineering and beyond

Data transformation and feature engineering are important steps in the analysis process. Data transformation is the process of converting data from one format to another, while feature engineering is the process of creating new features from existing data. Both tasks can be complex and time-consuming, but they are essential to ensure that data are clean, consistent and usable for analysis.

Data transformation is of paramount importance for a number of reasons. It ensures that data are presented in a format that can be easily understood and manipulated by computers, thus facilitating their use in analyses and models. It eliminates noise and errors from the data, helping to improve the accuracy of results obtained during analysis. Data transformation can be used to convert information into a format more suitable for analysis, notably by transforming textual data into

numerical data, thus facilitating the use of machine learning and statistical analysis techniques.

Feature engineering is also an important step in the data analytics process. Features are the individual pieces of data used to train a machine learning model. The better the features, the more the model will be able to learn and make predictions. Feature engineering can involve a number of tasks, such as selecting the right features, transforming features and creating new features.

We can summarize the advantages of these two tasks in the following points:

- *Improved accuracy and reliability*: by cleaning, transforming and engineering data, you can improve the accuracy and reliability of models. In effect, you will remove noise and errors from the data and create features that are more relevant to the task at hand.

- *Better insights*: by transforming and engineering data, you can gain new insights into your data. In effect, you will be able to see data in new ways, and you will be able to create new features that reveal hidden patterns.

- *Improved decision-making*: by cleaning, transforming and engineering data, you can make better decisions. Indeed, you will have a better understanding of your data, and be able to make decisions based on accurate, reliable information.

Data transformation and feature engineering are therefore two crucial steps in the preparation process, playing a central role in improving the performance and accuracy of machine learning models. Traditionally, these tasks have been laborious, requiring domain expertise and significant manual effort to identify relevant features and design them appropriately (Davenport and Dyché 2013).

However, the data preparation landscape is undergoing a paradigm shift with the emergence of advanced NLP technologies, notably ChatGPT. Leveraging the power of AI and NLP, data analysts can now automate and enhance these critical tasks, revolutionizing the way data is prepared and analyzed.

One of ChatGPT's key capabilities in data transformation is its ability to convert raw data into new features that offer valuable insights and improve predictive models. By generating code, ChatGPT can efficiently extract meaningful information from various types of data, including text and images. This enables data analysts to effortlessly create powerful features without the need for a lot of manual work. For example, consider a dataset holding customer reviews for a product. Using ChatGPT, analysts can derive sentiment scores from the textual data, classifying each review as positive, negative or neutral. These sentiment scores can then be used for new

features that significantly improve the accuracy of sentiment analysis models, enabling companies to better understand customers' feelings and preferences.

In addition, ChatGPT can be used to generate word or document embeddings for textual data. Word embeddings transform words into dense numerical vectors, capturing the semantic relationships between words. Document embeddings summarize the content of entire documents in fixed length vectors. These embeddings can serve as powerful input features for downstream tasks, such as document classification, information retrieval or clustering. Leveraging ChatGPT's feature engineering capabilities, data analysts can enrich their datasets with highly informative embeddings that contribute to the overall predictive power of their models.

In the realm of feature engineering, ChatGPT's assistance is invaluable in identifying relevant features and engineering new ones that are more predictive. Let us say we have a set of marketing data including customer demographics and purchase history. Leveraging ChatGPT, analysts can generate code to calculate additional features, such as average purchase amount per customer, purchase frequency or total number of products purchased. These engineered features can significantly enhance customer segmentation models, enabling companies to effectively tailor their marketing campaigns to specific customer groups.

In addition, ChatGPT can help explore interactions between different features and create new terms of interaction. For example, in a dataset containing housing price information, ChatGPT may suggest combining the "number of bedrooms" and "number of bathrooms" features to create a new feature presenting the total number of rooms, which may have a deeper impact on house price prediction. These feature interactions can capture non-linear relationships and lead to predictive models.

Data cleaning and feature selection are essential aspects of data preparation, and ChatGPT's capabilities go beyond transformation and feature engineering. By generating code, ChatGPT can help manage missing data, remove duplicates and impute values based on statistical methods, thus effectively automating data cleaning processes. For example, if a dataset contains missing values in the "age" column, ChatGPT can suggest the use of mean, median or regression imputation techniques to efficiently fill in these missing values. This automated approach not only saves time but also guarantees data quality and accuracy.

In the context of feature selection, ChatGPT can help to identify and rank features according to their importance for predictive models. By suggesting techniques such as recursive feature elimination, principal component analysis or SelectKBest, ChatGPT helps to select the most informative features, reducing

dimensionality and preventing over-fitting. The automated feature selection process further improves the efficiency and effectiveness of data preparation.

ChatGPT offers powerful approaches to data preparation tasks. It can transform categorical data into numerical formats, scale data into specific ranges and generate new features from existing data. The tool automates data cleaning by identifying and removing missing values and outliers.

Tasks	ChatGPT approach	Example
Data transformation	Using ChatGPT to convert categorical data into numerical data	Converting the “gender” column with the categories “male” and “female” into numerical values 0 and 1, respectively
	Using ChatGPT to scale data within a specific range	Scaling “age” column from range 0–100 to range 0–1
Feature engineering	Using ChatGPT to generate new features from existing data	Generating a “total revenue” feature by adding the “salary” and “bonus” columns
	Using ChatGPT to derive features based on textual data	Extracting “sentiment” feature from customer reviews using ChatGPT’s sentiment analysis
Data cleaning automation	Using ChatGPT to identify and remove missing values	Generating code to impute missing values in the “age” column with the average age
	Using ChatGPT to detect and remove outliers	Generating code to remove data points in the “income” column that lie outside the range of three standard deviations
Feature selection automation	Using ChatGPT to identify relevant analytics-based features	Generating code to select the 10 most important features using ChatGPT’s feature importance technique
	Using ChatGPT to automatically select features based on correlations	Generating code to select features with a correlation coefficient above a certain threshold

Table 3.3. *Valuing ChatGPT for data transformation, feature engineering and data preparation task automation*

What’s more, it facilitates feature selection by identifying relevant features and selecting those with significant correlations. ChatGPT revolutionizes data analytics

by enabling natural language queries, making interactions more intuitive and interactive.

Beyond data transformation and feature engineering, ChatGPT's capabilities continue to empower data analysts in a variety of data preparation tasks. For example, ChatGPT can generate code to handle outliers, detect and correct errors, normalize data and encode categorical variables using one-hot encoding or label encoding.

This comprehensive suite of functions ensures that data analysts have the tools they need to effectively prepare their data for machine learning approaches and other advanced analytical tasks.

ChatGPT's remarkable capabilities in data transformation, feature engineering and other data preparation tasks have revolutionized the data analytics landscape.

By automating repetitive and time-consuming tasks, ChatGPT enables data analysts to focus on higher-level strategies, problem-solving and generating actionable insights. This powerful technology has the potential to transform the way we interact with data, allowing us to ask questions in a more intuitive and interactive way.

Now that we have explored how ChatGPT can support us in the data preparation phase, we will move on to learn how to create natural language queries for data analytics and skillfully transform analytical questions into efficient SQL queries.

Intuitive Query Creation with ChatGPT

One of the most powerful capabilities of generative AI models is the ability to intuitively query data via natural language interactions. Rather than forcing users to master complex query syntax, generative AI enables them to query data using simple English. This helps overcome one of the biggest hurdles in data analytics – translating natural language-articulated analytical questions into formal database queries.

With ChatGPT, users can simply describe the information they need using prompts. The AI agent will then interpret the intent and context to automatically generate optimized SQL queries that retrieve the desired data. By leveraging ChatGPT for intuitive queries, analysts and business users can focus on discovering insights rather than coding queries.

In this chapter, we will see how tools like ChatGPT make it easy for users to ask questions about their data and automatically generate the SQL queries required to retrieve the answers. We will cover the techniques for formulating analysis questions for ChatGPT in a clear and precise way that leads to the generation of accurate queries.

4.1. The discovery of patterns, trends and insights through interactive conversations

Data analytics encompasses a dynamic process of exploring, refining and modeling data with the ultimate goal of extracting valuable insights, drawing informed conclusions and supporting decision-making. This multi-faceted discipline covers a wide range of techniques and approaches across various fields.

At its core, data analytics strives to untangle meaningful patterns, trends and insights embedded in data. These components are of crucial importance for understanding data dynamics and predicting future trends. Patterns are recurring sequences of data points that reveal consistent behaviors, enabling the identification of trends and even facilitating predictive capabilities (Frankel and Reid 2008; Sedkaoui 2018b). Meanwhile, trends represent long-term changes and developments in data, providing valuable insight into how data evolve over time. Insights, the true gems of data analytics, emerge as new understandings and knowledge derived from data, enabling better decision-making and propelling businesses and organizations forward (McAfee and Brynjolfsson 2011; Sedkaoui and Khelfaoui 2020).

Several techniques are at the forefront of the data analyst's toolbox, each making its own unique contribution to pattern discovery, trends and insights:

- *Statistical analysis*: by calculating statistical measures such as the mean, median and standard deviation, analysts can discern patterns and trends in the data, highlighting their inherent characteristics.

- *Data visualization*: through the art of data visualization, analysts present information in a visually appealing way, leveraging the power of tables, graphs and other visual presentations to reveal patterns and trends that might otherwise remain hidden.

- *Machine learning*: by leveraging the power of machine learning, data analysts can create new models that reveal complex patterns and trends by training the model on a set of data and allowing it to identify similar patterns in new data.

The fascinating fusion between generative AI and data analytics is revolutionizing our approach to discovering valuable insights. On the one hand, data analytics inspects, cleans, transforms and models data to reveal patterns and trends. On the other, generative AI models such as ChatGPT, with their ability to understand human language, offer a new pathway to fluid and interactive data exploration. This conjunction of intelligences broadens the horizon of discovery, guiding users in dynamic dialogues with their data, revealing deep insights and pushing decision-making to new heights.

The transformative impact of generative AI on data analytics and business intelligence cannot be overestimated. This cutting-edge technology is reshaping the way we interact with data, making them more accessible and intuitive than ever before. The traditional approach of relying solely on static reports or predefined queries is being revolutionized, as tools such as ChatGPT enable users to engage in dynamic, natural “dialogues” with their data.

4.1.1. Key benefits

Let us delve deeper into the key benefits that conversational AI brings to the table, propelling data mining into an unprecedented realm of possibilities:

– *Freeing up ad hoc exploration*: with ChatGPT, users are no longer confined to analysis *workflows*. The chains of structured data exploration are broken, and a world of fluid, unstructured exploration awaits. Imagine immersing yourself in data and letting it guide your discovery process. Like an intrepid explorer, you can venture into uncharted territory, unraveling patterns, trends and correlations that you might have missed with conventional approaches.

– *The art of intuitive querying*: the language of data analytics has been democratized thanks to generative AI models. Users no longer need to possess specialized query skills or be familiar with complex programming languages. Instead, they can converse with data as easily as they would with a friend. This natural approach to asking questions unlocks up a wealth of insights, from identifying exceptions to making predictions about future trends. Data analytics is no longer limited to the privileged few; it becomes an experience that is inclusive and rewarding for all.

– *Seize opportunities for rapid follow-up*: discovering valuable insight is just the beginning with intelligent models like ChatGPT. The ability to ask immediate follow-up questions means you can dive deeper into emerging revelations. As your curiosity is piqued, you can seamlessly dig deeper, unearthing underlying causes and subtleties that lurk beneath the surface. It is like having an interactive dialogue with your data, where each answer sparks new questions, leading to a continuous cycle of discovery and learning.

– *Automation of democratized access*: in the past, working with data required specialized expertise, creating a barrier for those without technical skills. But generative AI breaks down these barriers, allowing more users to actively participate in the data analytics process. Now, team members from diverse backgrounds can interact with the data in meaningful ways, contributing their unique perspectives and domain knowledge. The result is a more collaborative and inclusive approach to data-driven decision-making, enabling organizations to harness the collective intelligence of their teams.

In this seamless synergy between humans and AI, generative AI has heralded a new era of data mining. Gone are the days of rigid queries and predefined paths; data scientists embark on an open-ended journey, where ideas blossom organically, unconstrained, as unexpected sources of inspiration. As the example of dialogue shows, the beauty of this approach lies in the fluidity and spontaneity it brings to the data analytics process.

4.1.2. Discovering insights organically

By harnessing the power of ChatGPT, data scientists are not simply observers of data, but active participants in a stimulating exchange. They can now ask questions, follow their instincts and approach the subtleties of the data with an AI companion who responds naturally and intelligently. This collaborative relationship amplifies the creative thinking of data analysts, enabling them to connect the dots, spot trends and unmask hidden patterns that traditional methods might otherwise ignore.

Data mining becomes an exciting adventure, where every query sparks new ideas and every answer unveils new possibilities. The iterative nature of this process allows for continuous refinement and course correction, refining questions as new ideas emerge and taking analysis into unexpected territories.

What really sets this organic exploration apart is the element of serendipity it introduces. Like an explorer who stumbles upon a hidden treasure while venturing into uncharted lands, data analysts experience moments of surprise and inspiration when seemingly unrelated information converges to reveal a profound understanding of the data story. These “aha” moments are invaluable, sparking innovative ideas and opening doors to transformative opportunities.

ChatGPT not only enables insight discovery, but also serves as a gateway to democratize data access. Thanks to its natural language interface, it brings data analytics to all members of the organization, transcending technical barriers and fostering a culture of data-driven decision-making. Data literacy becomes a shared asset, as teams effortlessly collaborate and communicate with their data, each bringing their unique perspectives to fuel collective growth and success.

The insights that emerge are not simply static snapshots, but living entities that evolves with the dynamic nature of the data. Armed with a better understanding of their operations, customer preferences and market trends, companies gain the agility to adapt quickly to changing landscapes, seize new opportunities and stay one step ahead.

4.1.3. Democratizing data analytics

The democratization of data analytics thanks to generative AI heralds a revolutionary era of inclusivity in the world of data-driven decision-making. No longer limited to the realm of data scientists, professionals from diverse backgrounds – business analysts, marketers, department managers, etc. – can now become actively involved in the process of data discovery. The natural language interface (Berger et al. 2020) of conversational AI enables them to ask insightful

questions and gather valuable insights without the need for in-depth technical expertise in SQL or queries.

This broader participation in data analytics fosters a culture of data literacy throughout the organization. As individuals in various roles immerse themselves in data mining, they develop a deeper understanding of organizational performance, customer behavior and market trends. The democratization of data analytics fosters informed decision-making at all levels, paving the way for strategic growth and competitive advantage.

4.1.4. The future of business intelligence

In particular, the advent of generative AI also resolves the long-standing IT bottlenecks that often hamper data access and analysis. By offloading the task of retrieving and processing data onto AI assistants like ChatGPT, organizations reduce their dependence on specialized data engineering resources. This freed-up IT workforce can now redirect its efforts towards more strategic initiatives, fostering innovation and streamlining processes.

The future of business intelligence lies in seamless collaboration between humans and AI systems, where conversations, not just queries, form the basis of exploration. Thanks to this symbiotic partnership, users embark on an interactive voyage of discovery, supported by AI-powered insights that transcend the boundaries of conventional data analytics. This marriage of human ingenuity and AI prowess opens up a realm of untapped possibilities, where deep insight emerges from the heart of data, fueling strategic decision-making and generating better business results.

As the frontier of data-driven knowledge expands, organizations will harness the transformative power of generative AI to unlock untapped potential. In this future landscape, expert and casual users alike will harness the analytical prowess of AI, infusing intuition and human creativity into the dialogue of data mining. Together, they will untangle the complex tapestry of information, revealing patterns, trends and insights that have eluded traditional methods.

The future of business intelligence is upon us, where the synergy of human curiosity and AI assistance holds the key to unlocking strategic advantage and propelling organizations towards a prosperous future.

4.2. Creating natural language queries to analyze your data

For a long time, natural language text understanding has presented significant challenges in the field of NLP. The problem of ambiguity and co-reference has been a persistent obstacle, making tasks such as answering questions about datasets and converting spoken language into structured SQL queries particularly difficult.

The prevalence of relational databases containing large quantities of valuable information, such as medical records, financial data, customer records and big data insights, further exacerbates the issue. Retrieving information from these databases and repositories has been hampered by the need for users to grapple with complex and powerful structured query languages.

The power of natural language queries goes far beyond merely simplifying interactions with data. It revolutionizes the way data analytics is performed, democratizing access to insights and improving collaboration between teams.

One of the main advantages of natural language queries is their accessibility. They remove the obstacles for data analysts who may not have programming experience (Surameery and Shakor 2023; Yilmaz and Yilmaz 2023), making data analytics more inclusive and accessible. No longer confined by technical jargon or complex query languages, data analysts from diverse fields can effortlessly interact with data using everyday language. This accessibility enables individuals to explore data, ask questions and gain insights without the need for specialized training, fostering a culture of data-driven decision-making throughout the organization.

Fortunately, with the advent of ChatGPT and the rapid progress of LLMs, the text-to-SQL challenge has eased considerably. Even non-technical users can now describe their queries using declarative language. The complex task of query construction is skillfully handled by LLMs, streamlining the process and making data retrieval more accessible than ever.

In addition, natural language queries elicit a sense of curiosity and exploration, fueling the quest for deeper insights. While experts articulate their questions in a conversational manner, AI not only translates them into SQL queries (Kim et al. 2020), but also seeks to understand the underlying intention. By understanding the context and nuances of the question, ChatGPT discovers patterns and trends in the data that match the user's interests. This interactive and iterative process fosters a symbiotic relationship between human expertise and AI-powered data mining, leading to the revelation of new, actionable insights.

But what does this mean for you? The integration of natural language for database queries opens up many interesting possibilities, which are summarized in Figure 4.1.

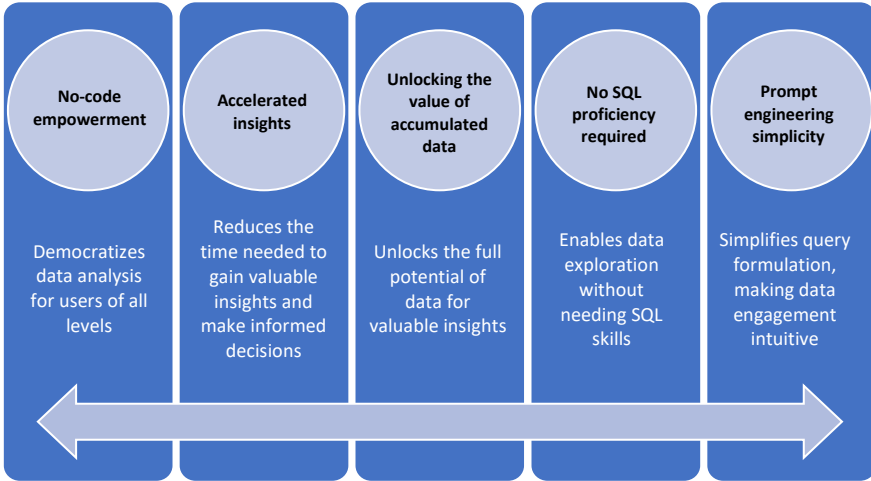


Figure 4.1. *ChatGPT's advantages for natural language queries*

ChatGPT plays a transformative role in data analytics by enabling users without coding skills. It democratizes the data analytics process, allowing individuals of different roles and skill levels to effortlessly access data-driven insights. Additionally, the adoption of natural language queries accelerates the generation of valuable insights, leading to faster decision-making and a better understanding of domain-specific trends.

This improved access to data also unlocks the full potential of accumulated data, extracting valuable knowledge and valuable business insights. What's more, ChatGPT eliminates the need to master SQL, making data mining inclusive and accessible to all users. It puts the emphasis on prompt engineering rather than coding knowledge, ensuring that users need only a basic understanding to formulate queries effectively, enhancing the intuitive and user-friendly data engagement process.

In addition to individual contributions, natural language queries also foster a collaborative environment. Teamwork among data analysts becomes more transparent, as colleagues can easily exchange questions, conclusions and findings using a shared natural language interface. This streamlined communication rationalizes information sharing and encourages collective problem-solving. As each

member contributes their perspective, the team's collective intelligence grows, leading to deeper analysis and more informed decision-making.

The impact of natural language queries extends beyond the role of data analyst; it resonates throughout the organization, influencing the way business decisions are made. Leaders, managers and stakeholders, who may not possess deep data expertise, can now actively participate in data discussions, contributing their domain knowledge and shaping the organization's strategic direction.

Mastering the art of formulating efficient natural language queries with generative AI paves the way to harnessing the true potential of data analytics (Devlin et al. 2019). With its powerful AI-driven chatbot, ChatGPT spearheads the evolution of a new era in NLP technology. Its remarkable ability to understand and interpret natural language queries, translating them into actionable insights, places ChatGPT at the forefront of advances in conversational AI.

Exploring this area of intuitive interaction further, here are a few additional pieces of advice to improve the accuracy and precision of your queries:

- *Understand the data model*: become familiar with the structure and relationships within your datasets. Knowing the available data fields, their types, and their relationships to each other allows you to create more contextual queries. This understanding will ensure that ChatGPT can accurately interpret your questions and provide relevant information. You can provide ChatGPT with background context about your data sources and your business terminology, so it can better understand intent. For example: “According to our e-commerce database, what was the total revenue for the fourth quarter?”; or even: “How many new customers have registered this month in our CRM system?”.

- *Use natural language examples*: instead of relying solely on rigid queries, providing natural language examples can clarify your intent and refine ChatGPT's understanding. For example, you could include a conversational context such as: “Show me the sales data from the last three months”. This improves the AI's ability to understand the scope and duration of your request.

- *Embrace exploratory questions*: do not hesitate to ask exploratory questions to uncover hidden patterns and trends. For example, you could ask: “What characteristics do our most loyal customers share?”; or “Can you identify any seasonal patterns in the sales data?” In this way, you need to clarify the exact data fields, statistics or attributes you wish to include in the query results. You can ask ChatGPT: “Show me every product name, unit price and unit sold for all transactions last month.” And: “What are the first name, last name and e-mail address of our inactive customers?”.

– *Experiment with different approaches*: as with any linguistic communication, experimenting with different formulations and wordings can yield varied query results. Try rephrasing your questions from different angles, and you may discover new angles and dimensions to explore in your data. When querying large datasets, include any filters, conditions or selection criteria in your natural language question, such as: “What were the top 5 best-selling products last month for customers aged 30 to 40?”; or: “Which campaigns generated the most leads from Canada in the second quarter?”.

– *Take advantage of conversational feedback*: AI systems like ChatGPT learn from interactions. If the AI’s answers do not perfectly match the intent of your query, offer feedback by rephrasing the question or asking for clarification. This feedback loop can lead to more precise and refined queries in subsequent interactions. Review the queries produced by ChatGPT and reformulate your question if necessary. Adding examples allows you to refine queries over several iterations. Following these tips will enable you to use ChatGPT more effectively for data analytics queries. With practice, you can have intuitive conversations that unlock deeper insights from your data.

– *Combine human expertise with AI assistance*: do not forget that you are the domain expert and ChatGPT is your artificial collaborator. Combine your subject knowledge with the data processing capabilities of this intelligent agent to gain deeper insights. Trust your instincts and let AI complement your expertise with data-driven perspectives.

As you integrate these additional strategies into your data analysis process, you will notice a seamless harmony between the capabilities of human intelligence and AI. Generative AI, such as ChatGPT, will become a powerful ally, able to decipher your intentions and provide actionable insights. With practice and improvement, you will master the art of questioning data in natural language, revealing valuable patterns, making informed decisions and propelling your organization towards data-rooted success.

4.3. The art of transforming analysis questions into SQL queries with ChatGPT

SQL, or structured query language, is a fundamental tool in the world of data management, particularly within relational databases. Its flexibility and versatility make it an indispensable language for performing a wide range of data analytics tasks such as retrieving specific information with queries, updating existing data and deleting records where necessary.

Transforming analysis questions into SQL queries is a crucial step in the data analytics process. Given that databases include SQL, translating your analysis questions into this language ensures that your queries are properly understood and processed. This seamless conversion is essential for extracting accurate and efficient answers, leading to informed decision-making and insightful discoveries.

The benefits of this transformation are considerable and far-reaching. Firstly, it significantly improves communication between data analysts and team members. The use of SQL queries to express analytical questions allows for clear and precise communication, facilitating seamless collaboration and knowledge sharing within the team.

Secondly, the automation aspect is crucial to streamlining data analytics tasks. Once you have created an SQL query, you can save it for future use, reducing redundant work and maintaining consistency across multiple analyses. This automation not only saves time and effort, but also ensures that repetitive tasks are handled accurately every time.

In addition, the ability to adapt data analytics becomes a crucial consideration. By transforming analysis questions into SQL queries, you enable your data analytics process to adapt efficiently to expanding datasets and complex data scenarios.

The ability of generative models to translate natural language questions into executable SQL queries is revolutionizing the way data analysts interact with datasets. Instead of struggling with complex query syntax, tools like ChatGPT allow analysis questions to be expressed in plain language, making data exploration and analysis more accessible to a wider range of users (Sedkaoui 2023).

The introduction of OpenAI's ChatGPT product has significantly improved the productivity of business users and data analysts. One of the main advantages of this technology is its ability to facilitate SQL query writing. SQL has long been a valuable language for querying and analyzing data, and now, with the help of AI, learning and writing SQL queries has become more accessible than ever. AI technology can automatically generate, modify, debug and optimize SQL queries in seconds.

This tool is particularly beneficial for beginners in SQL or those struggling with complex queries. Its remarkable advantage lies in its ability to understand natural language input, enabling users to articulate their data needs and then generate the appropriate SQL query. This simplified process saves users considerable time and effort compared to manual query composition.

In addition, ChatGPT enhances the query writing experience by generating several options for a given query, enabling users to explore different structures or

compare alternatives. In addition, it assists in debugging SQL queries, quickly identifying and highlighting errors for efficient troubleshooting.

Beyond basic query writing, ChatGPT excels in guiding users through advanced SQL concepts, such as joining tables and creating subqueries. It offers clear explanations and illustrative examples, supporting users' understanding and successful implementation of these complex concepts.

What's more, ChatGPT's prowess extends to providing contextual suggestions as users elaborate their queries. It offers information on table structure and column details, helping users to better understand the data analyzed. Thanks to ChatGPT's invaluable assistance, users can devote less effort to query composition and more time to in-depth data analysis.

ChatGPT can help data scientists with their SQL queries in many ways, from writing basic queries to optimizing complex queries. Figure 4.2 summarizes this.



Figure 4.2. *Using ChatGPT for SQL queries*

For example, you may say: “I need to create a query that associates the ‘customers’ table with the ‘orders’ table to return the total spend for each customer. The

‘customers’ table contains the customer’s ID, name, e-mail address, etc., and the ‘orders’ table has the order ID, customer number, order amount...”

ChatGPT can then suggest effective logic and provide advice on indexes to improve performance based on the details provided.

For simpler queries, ChatGPT can automatically translate natural language queries into efficient SQL syntax. For more complex queries involving multiple tables, joins and advanced functions, ChatGPT can provide a head start by suggesting the optimal join logic and query structure.

Data scientists can refine and iterate with ChatGPT to create powerful queries tailored to their unique needs. This facilitates faster, easier and more scalable SQL development.

In the future, SQL plugins specially designed for large language models, such as GPT-4, could take ChatGPT’s programming support to the next level. Rather than a general command of natural language, SQL plugins would provide models with a deeper knowledge of database schema query optimization and advanced SQL functionality.

The GPT-4 version offers the potential of plugins, which are extensions that can be integrated to add new functionality. These plugins are of significant value for data analysts, particularly in the field of SQL queries.

Specialized plugins could enable direct SQL input rather than simple natural language conversations (Sedkaoui 2023). The model could then analyze, debug, reformat and improve the input SQL based on database metadata and query best practices. This would enable advanced auto-completion, linting and performance optimization functionalities beyond what is possible today.

But to guarantee accurate translation, it is crucial to provide ChatGPT with a relevant database context. By specifying the database schema, including table and column names, data types and relationships, data analysts equip ChatGPT with the information needed to generate precise, contextually relevant queries. This database context forms the basis for generating queries that align with specific data structure and business objectives.

In addition, articulating the intent and objectives of the analysis question helps ChatGPT to select the most appropriate query approaches. By clarifying what they seek to analyze and discover, analysts enable ChatGPT to optimize query formulation and deliver valuable insights quickly.

It is also essential to include important filters and result limitations within the query. By providing these specifics, analysts can refine the query output, ensuring that the results correspond precisely to the desired scope of analysis.

What's more, specifying the desired fields, measures or aggregate calculations required in the query further enhances its accuracy and relevance. By informing ChatGPT exactly of the required data elements, analysts streamline the data retrieval process, leading to more insightful and relevant results.

As data analysts continue to practice conversational querying, they will become able to refine their questions for optimal outcomes. By iterating on queries and reviewing initial ChatGPT attempts, analysts can refine their questions to extract the perfect insights from the data.

4.4. Generating efficient and optimized queries: the key to your success with ChatGPT

Data analytics is a multi-faceted process that involves various stages, from data exploration, cleaning and modeling, to the elaboration of meaningful insights (Sedkaoui 2018a; Sedkaoui and Khelfaoui 2020). A crucial aspect of data analytics is the generation of efficient and optimized queries. An efficient query not only delivers the desired results accurately, but also does so in a timely manner, avoiding unnecessary delays and inefficiencies.

To achieve an effective and optimized query, several factors come into play. First and foremost, clarity is of the utmost importance. The query must be simple and easy to understand, with a clear purpose and intent. This ensures that stakeholders and team members can easily understand the question posed with regard to the data.

Specificity is another key consideration. A well-defined and specific query enables the search to be refined and ensures that the results returned correspond precisely to requirements. This minimizes the risk of sifting through irrelevant data and wasting precious time.

Choosing the right keywords is essential for a successful query. The keywords used must accurately present the data being queried, aligning with the structure and content of the dataset. This ensures that the query effectively targets the relevant data and avoids ambiguity.

The use of indexes is a powerful strategy for optimizing query performances. Indexes are data structures that enable faster data retrieval by creating efficient access

paths. By using appropriate indexes, queries can quickly identify and retrieve the data they need, resulting in significant performance improvements.

Another essential element in generating efficient queries is the use of appropriate joins. Joins make it possible to combine data from several tables, facilitating complex queries where information needs to be extracted from a variety of sources. Selecting the right type of join and structuring the query effectively can improve the efficiency of data retrieval.

To further improve query generation, the use of natural language queries can be beneficial. Natural language queries enable data analysts to interact with data in a more conversational manner, making the process more intuitive and user-friendly. This approach allows analysts to focus on the questions they want to explore, rather than struggling with complex query syntax.

In addition, incorporating various clauses into the query can help refine the results. The *WHERE* clause allows analysts to specify data selection conditions, by refining the output to meet specific criteria. The *ORDER BY* clause allows data to be sorted in a particular order, while the *LIMIT* clause limits the number of rows returned, making analysis both more manageable and more efficient.

But the question here concerns ChatGPT's contribution to efficient and optimized queries, and the different ways in which it can be used to this end.

The key to answering this question lies in identifying the various functionalities that ChatGPT provides to generate efficient and optimized queries, namely:

- *Understanding natural language queries*: thanks to its in-depth training in a variety of texts and codes, including natural language queries, ChatGPT is able to understand the intent behind user questions and accurately convert them into SQL queries. This functionality makes the query more intuitive and user-friendly. For example, the user can simply ask: “Show me last year’s total sales per month”, and ChatGPT will translate this question into an appropriate SQL query to extract the necessary information.

- *Generation of clear, concise queries*: ChatGPT excels in producing clear, concise queries, improving their comprehensibility and interpretability. This contributes to the accuracy of the obtained results, enabling data analysts to efficiently extract valuable insights from data. For example, the user may ask: “What were the five best-selling products last month?”, and ChatGPT will generate a precise and concise SQL query to obtain this specific information.

- *Use of keywords*: by using relevant keywords in queries, ChatGPT guarantees the accuracy while reducing the number of rows to be processed. This optimization

plays an important role in improving query efficiency and performance. For example, the user may ask: “Show me all customers under the age of 30 who have made a purchase this month”, and ChatGPT will include the keywords “customers”, “age”, “purchase” and “month” in the SQL query to filter the data.

– *Use of indexes*: by using indexes, ChatGPT improves query performance, especially for frequently executed or data-intensive queries. Indexes facilitate faster data retrieval, which translates into faster, more efficient analysis. For example, the user can ask: “Give me the total sales of each product for the last three years”, and ChatGPT will use indexes to speed up data retrieval, enabling faster analysis.

– *Use of joins*: ChatGPT makes efficient use of joins to combine data from different tables, an essential capability for handling complex queries that require the analysis of multiple datasets. This functionality enables analysts to explore and integrate various data sources seamlessly. For example, the user may ask: “Show me each department’s revenues and associated costs for last year”, and ChatGPT will use joins to combine data from the revenue and cost tables and provide the complete results.

By using ChatGPT, data analysts can also generate SQL queries containing previously discussed clauses, such as the *WHERE* clause, the *ORDER BY* clause and the *LIMIT* clause, by simply including them in their question. This enables users to obtain more targeted and organized results, optimizing their data analysis and facilitating informed decision-making.

For example, an analyst might query ChatGPT the following: “Display the total sales of products priced over \$100”. In this case, ChatGPT would incorporate the *WHERE* clause into the SQL query to filter donations born according to the specified condition. The analyst could also formulate a request such as: “Sort total sales in descending order”, and ChatGPT would add the *ORDER BY* clause to the SQL query to present the results in descending order from highest to lowest.

Similarly, the data analyst could prompt ChatGPT with a query such as: “Show me the five best-selling products from last month”. In this case, ChatGPT would integrate the *LIMIT* clause to display only the five most relevant products according to the defined criteria. These clauses make the query more precise and customized, making it easier to obtain exact and pertinent information needed for data analysis.

Using ChatGPT to optimize SQL queries is not limited to what has already been mentioned. In fact, it goes beyond the simple generation of queries from natural language. The development of efficient, optimized queries is a collaborative process that involves drawing on the strengths of expertise and the capabilities of AI.

By providing ChatGPT with the necessary context and details, data analysts can guide the model to generate queries that correspond precisely to their data analysis needs (Sedkaoui 2023). This context includes information on the database schema, pertinent business terminology and any specific filters or limits required. By formulating clear and precise questions, users ensure that the queries generated by ChatGPT are targeted and accurate.

The generated queries are, however, not the final end point. They serve as a starting point for further optimization. ChatGPT can provide valuable suggestions for improving query performance, such as recommending the use of indexes on filtered columns or suggesting simplified joins based on data relations. It can also assist in choosing the right data types to improve query performance.

The collaborative aspect comes into play when data analysts and developers work together to refine the queries generated. By iteratively improving queries with ChatGPT's assistance and human surveillance, analysts can achieve high-performance queries that minimize execution time while delivering complete and accurate results.

This collaboration is a game-changer for data analytics. On the other hand, even untrained users can actively participate in the query process, thanks to ChatGPT's ability to understand natural language queries. This democratizes data analysis, making it accessible to a wider range of professionals within an organization.

With attentive prompts and a feedback loop of iterative improvements, ChatGPT can minimize query execution time while returning complete and accurate results. The fusion of human and AI facilitates the generation of SQL in a way that neither could do alone – by marrying business logic with technical optimization.

Having explained how to create intuitive queries using ChatGPT, we now move on to explore the advanced analysis techniques made possible by generative AI. Chapter 5 shows how ChatGPT can go beyond basic data retrieval to provide invaluable assistance with exploratory data analytics, statistics, prediction, text mining, etc.

Although crafting the optimal SQL query is fundamental, it is only a starting point. ChatGPT's capabilities really shine when it comes to accelerating complex analytical tasks that normally require data analysis experts. By automating rote work and augmenting human critical thinking, ChatGPT allows us to extract deeper insights and to optimize decisions.

ChatGPT: The Advanced Analysis Wizard

In Chapter 4, we explained how ChatGPT makes querying and accessing data more intuitive through natural language conversations. However, ChatGPT's true capabilities extend far beyond basic data retrieval. This chapter explores how the intelligent generative AI agent can automate and augment advanced analysis techniques for deeper insights.

ChatGPT excels at helping data scientists in the more complex and high-value tasks that follow data collection and preparation. These include exploratory data analysis, statistical modeling, predictions, sentiment analysis and more. Such techniques normally require skilled data specialists.

By combining its computational power with human surveillance and domain expertise, ChatGPT can help speed up advanced analytics. Repetitive and routine aspects are automated, while ChatGPT's explanations strengthen human judgment.

This chapter demonstrates how ChatGPT can function as an analytics companion, offering the best of AI and human critical thinking. We cover innovative applications in exploratory data analysis, statistical analysis and predictive modeling, as well as text mining and sentiment analysis. Together, these techniques provide deeper insights and optimize data-driven decisions.

5.1. Exploring new horizons: ChatGPT for exploratory data analysis

Exploratory data analysis (EDA) is a fundamental approach to data analytics (Tukey 1977), essential for obtaining deeper insights and understanding the underlying patterns within a dataset. Using visualization and statistical analysis techniques, EDA aims to unravel valuable information, such as trends, outliers and

data distributions, which lay the foundations for subsequent modeling and hypothesis testing.

One of the main advantages of conducting EDA is its ability to identify patterns and trends in the data. This information is crucial for making informed decisions, as it enables analysts to recognize recurring behaviors or events. For example, within customer data analysis, EDA can reveal recurring purchasing patterns, enabling companies to effectively tailor targeted marketing strategies to specific customer segments.

EDA also serves as an outlier detection mechanism, helping to identify data points that deviate significantly from general trends. Outliers can provide valuable insights about rare events or anomalies in the data. Identifying and understanding outliers helps to distinguish erroneous data points from significant observations, leading to more accurate analyses.

In addition, EDA facilitates a global understanding of the structure of the dataset (Tukey 1977; Sedkaoui 2018a), the relationships between variables and data distribution. This knowledge is essential for selecting appropriate methods of analysis and formulating hypotheses. Understanding the data landscape enables analysts to efficiently explore the dataset, extract meaningful information and interpret results more accurately.

Visualization plays a central role in effective EDA. Thanks to the visual presentation of data, patterns and trends become more obvious, enabling analysts to grasp complex relationships and uncover hidden insights that might otherwise go unnoticed in the raw data. Furthermore, statistical analysis adds a quantitative aspect to the exploratory process, providing measures and metrics that help to understand data patterns objectively.

EDA is an iterative process that involves continuous exploration and discovery (Berman 2013; Sedkaoui 2018a). As analysts take the data further, new questions and hypotheses may emerge, requiring further exploration and refinement of the analysis. Adopting the iterative nature of EDA ensures that all aspects of the dataset are explored in depth, ultimately leading to richer insights and more informed decision-making.

In this open-ended discovery process, EDA requires both technical expertise and human intuition, making it a creative and analytical undertaking. Here, generative AI models, exemplified by ChatGPT, play a unique and complimentary role, enhancing human judgment and collaboration during this exploration.

One important aspect in which ChatGPT excels is hypothesis generation. By examining datasets, ChatGPT can help analysts brainstorm intriguing hypotheses based on trends and apparent anomalies. This synergy of human expertise and AI-based insights opens up new investigative directions. For example, by analyzing sales data, ChatGPT can prompt data analysts to explore the link between a significant increase in the third quarter and a promotional summer campaign. This joint exploration leads to more targeted and fruitful investigations.

Figure 5.1 shows how ChatGPT can profoundly transform key aspects of the analytical process – from data processing and analysis to insightful hypothesis generation and rigorous validation. ChatGPT accelerates workflows while increasing human creativity and monitoring across these phases.

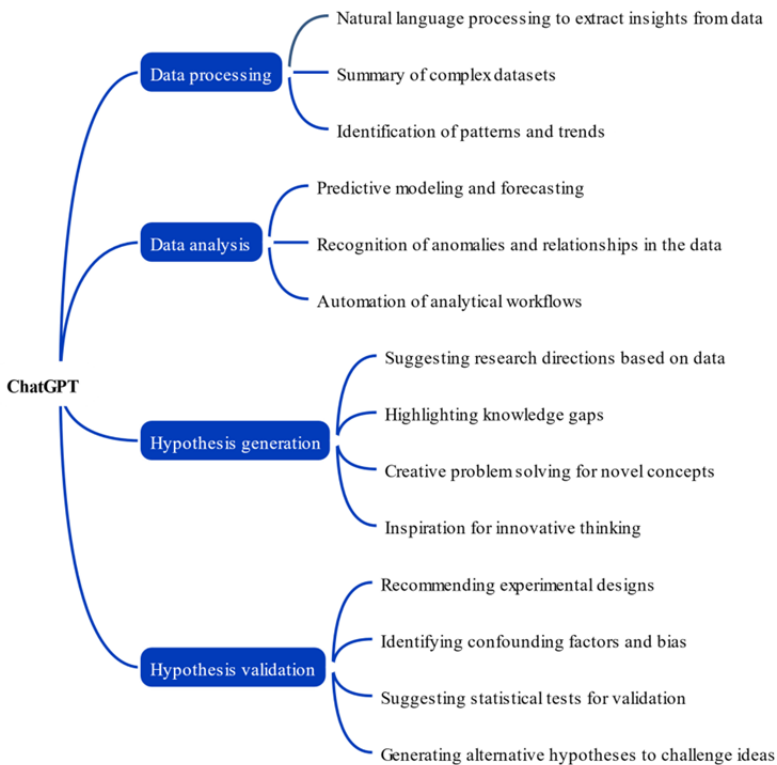


Figure 5.1. *Using ChatGPT for data processing, analysis and hypothesis generation*

In data processing, ChatGPT applies natural language features and pattern recognition algorithms to rapidly extract information, identify trends and summarize complex datasets. This automates laborious manual tasks while highlighting overlooked relationships.

With regard to analysis, predictive modeling and anomaly detection, ChatGPT reveals new insights that humans may miss. It enhances hypothesis generation with out-of-the-box thinking by highlighting gaps and suggesting creative new data-driven connections. It thus strengthens validation by systematically assessing experimental rigor and challenging hypotheses through alternative perspectives.

ChatGPT also helps to identify potential patterns that might require further investigation. The model can spot subtle trends or correlations within the data, and suggest relevant visualizations to illustrate them more clearly. This collaborative process enables data analysts to gain valuable insights into the underlying structure of the data (Sedkaoui 2023). For example, ChatGPT could help uncover a clear increase in sales each month through a heat map presentation, prompting further exploration of seasonality and its implications.

Outliers often contain critical information in data analytics. ChatGPT's ability to quickly analyze large datasets and flag potential anomalies for human review proves invaluable, as discussed earlier in this book. By identifying unusual data points, the model guides data analysts to examine these instances more closely. For example, ChatGPT can highlight three data points as outliers among daily active users, prompting data scientists to investigate the events or factors that led to the spikes on those specific dates.

Overall, the collaboration between generative AI and analysts in the EDA process presents an exciting fusion of technical capabilities and creative thinking. ChatGPT's talent for generating hypotheses, identifying patterns and spotting outliers complements the intuitive insights of data analysts, propelling the discovery process to new heights. This symbiotic relationship not only accelerates data exploration, but also improves the quality of analysis, leading to more in-depth and actionable discoveries. The interplay between AI and human judgment in EDA highlights the potential of technology as an indispensable ally in data-driven decision-making and problem-solving.

So far, we have seen how ChatGPT can be an incalculable asset for generating hypotheses, identifying patterns and identifying outliers in the EDA process. However, we have not yet addressed the other essential aspect of this data exploration process: the visualization of data.

Data visualization plays a central role in understanding complex datasets, revealing hidden patterns and presenting information in a clear and accessible way. Visual presentations enable the human mind to grasp information more easily, making it easier to detect trends, anomalies and correlations that might otherwise remain obscured in the raw data.

However, it is crucial to understand how ChatGPT approaches data visualization. Can it really generate graphs and other visual presentations from databases?

Even if we ask ChatGPT this question directly, we will get the answer given in Figure 5.2.

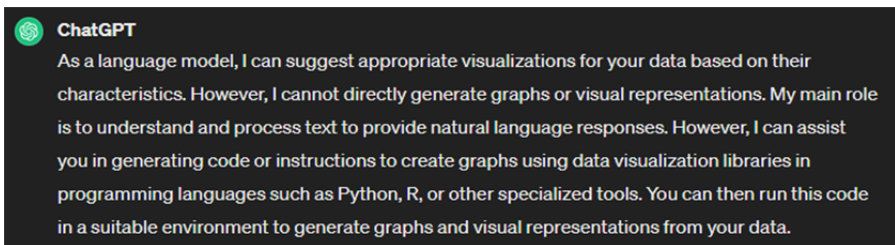


Figure 5.2. *ChatGPT's data visualization response*

This response is useful and informative, highlighting both the capabilities and limitations of ChatGPT as a natural language model. ChatGPT can be a valuable tool for data visualization, but it is crucial to consider its limitations.

This answer raises a good point – ChatGPT does not currently generate visualizations of interactive data or graphics on its own. So, to state or believe that ChatGPT can actually “create visualizations” is not accurate, at least for now. In the context of data visualization, it should be noted that ChatGPT can, for example:

- suggest appropriate visualization types, such as bar charts, scatter plots or heat maps, based on the data you have and the objectives of your analysis;
- guide you through the essential elements of a visualization, such as axes, data fields, colors etc., to make your graphs more informative and attractive;
- provide Python or R code examples that can be used to create the recommended visualizations;
- allow you to easily create tables to organize your data clearly and concisely.

It is true that absorbing knowledge through words alone often limits full conceptual understanding, and that visual presentations provide an additional layer

of understanding that is not fully transmitted by text alone. But, unfortunately, even advanced versions such as GPT-4 cannot autonomously generate explanatory graphics, despite their multimodal capabilities for processing images and text. Currently, GPT-4 can only analyze existing visuals, not create new ones.

This is where ChatGPT plugins prove invaluable. By integrating diagramming languages such as Mermaid and PlantUML, these tools enable users to visualize ChatGPT's textual explanations through graphs, mind maps, flow charts and more.

The synergy between ChatGPT's textual elucidation and plugin-enabled graphical presentation thus enhances comprehension. Critical concepts become clearer when expressed through both textual and visual mediums, creating deeper anchors for learning.

So, while ChatGPT remains confined to textual elucidation, the plugins unlock its visualization capabilities. Together, they enable a more holistic absorption of knowledge – meaningful words, graphs providing visual contexts. This demonstrates the power of ChatGPT's increased capabilities through integration with complementary technologies.

But it is worth noting that data visualization goes beyond simply creating esthetically pleasing charts and graphs. It involves determining what data are relevant to visualize, ensuring that the data are in the appropriate format for visualization, and exploring how data visualizations can even predict future trends.

In this context, it is worth asking whether advances in generative AI technology have succeeded in solving these daunting challenges for many people. Or put another way, how can ChatGPT be used for data visualization, helping users discover deeper insights from their datasets? By leveraging ChatGPT's capabilities, can we truly unlock the power of data visualization and improve the decision-making process?

The answers to these questions can be summarized as follows:

– unlocking data visualization recommendations from ChatGPT: managing complex datasets with numerous columns can be a daunting task, especially when it comes to deciding which data should be visualized (Sedkaoui 2018b). Analyzing data relationships and correlations can be time-consuming and challenging, leaving you overwhelmed by the myriad of available options. However, with the advent of language learning models and AI-powered applications like ChatGPT, data analysts can now save a great deal of time and effort in this process. Imagine you have a dataset of product sales in a store, and you want to uncover valuable insights through visualizations. Simply turn to ChatGPT for assistance. By taking advantage

of its capabilities, you can search for suggestions on what data to visualize and what types of charts would best suit your data analysis needs. You can easily provide a description of your dataset through a prompt, and ChatGPT will quickly come up with relevant recommendations for effective visualizations. With this powerful tool at your disposal, exploring complex datasets becomes a more efficient and productive undertaking. While the result may not be perfect at this stage, it provides a valuable starting point for overcoming initial difficulties. It sets you in motion and, as you progress, you can refine and improve your results. If you want to improve this approach, you can integrate OpenAI with your preferred business intelligence tool and generate automated data visualizations using AI. You can, for example, use ChatGPT's Code Interpreter, a robust tool that enhances the AI model's functionalities, and makes coding and data visualization more accessible;

- Improving the quality of your data for smarter data visualizations: not only can you take advantage of ChatGPT for data visualization, but you can also use its capabilities to enrich your dataset. If your existing data are missing certain information, incorporating additional data can lead to more insightful and comprehensive results. Consider supplementing your dataset with various enrichments, such as sentiment analysis of text strings, location coordinates, demographics, up-to-date foreign currency conversion rates, aggregate calculations and some more (Cupani 2022). With the help of ChatGPT, you can discover valuable scripts to seamlessly integrate these enrichments into your dataset. For example, you can ask ChatGPT to assess the sentiment of customer reviews, allowing your data analysis to better understand your customers' perspectives.

- Leveraging AI for predictive analytics: while visualizing historical data is valuable for learning from past experiences, relying solely on past information can leave you behind your competitors. What if you could proactively predict risks and opportunities before they arise? Predictive analytics is gaining popularity among organizations seeking to make data-driven decisions based on forecasts. Although a specialized field, the advancement of language learning models like GPT-3.5 and GPT-4 has made experimenting with predictive analytics more accessible. By training GPT on historical data, data scientists can now get started, creating AI-powered data visualizations that help predict future outcomes.

- Data cleaning and modeling with ChatGPT: as the GIGO principle indicates, the importance of data cleaning and optimized data modeling cannot be overestimated when it comes to meaningful data visualizations. Although data cleaning and modeling are not as captivating as the creation of charts, they are essential steps that cannot be neglected. As mentioned previously, ChatGPT can simplify the laborious process of data cleaning and processing. It helps identify potential outliers in the dataset, manage missing information and empty values, normalize data and select the most appropriate data model.

– Could AI-based data visualization be the future? The vast potential of emerging AI tools like ChatGPT is undeniable. While certain people may worry about the impact of AI on their careers, forward-thinking software teams are embracing its possibilities. Early adopters who learn to leverage AI for data visualization will gain a competitive edge, become more efficient and rapidly expand their skills. As AI and tools like ChatGPT become an integral part of various industries, including data visualization, these have the potential to revolutionize the way we work and create valuable insights.

5.2. Simplifying your analysis: automation tasks for increased efficiency

The automation of data analytics plays a crucial role in revolutionizing the analysis process. As highlighted in Chapter 3 of this book, data analysts often spend a lot of time on repetitive tasks, which can hamper their productivity and their ability to focus on strategic insights. However, with the advent of ChatGPT, automation is becoming a reality, enabling data analysts to harness the true power of their skills and expertise.

One of the most significant advantages of ChatGPT is its ability to automate data preparation. Thanks to well-designed prompts, ChatGPT can handle tasks such as database queries, data cleaning and imputation of missing values. This automation considerably reduces the hours spent on trivial data preparation, allowing analysts to devote their time to higher-added-value activities.

What's more, ChatGPT facilitates self-service report creation for professional users. By generating reports, suggestions for data visualization and dashboards based on user queries, ChatGPT enables non-technical stakeholders to access the information they need without relying heavily on analyst support. This self-service reporting approach reduces dependency on analysts and promotes more agile decision-making within organizations (Sedkaoui 2023).

Beyond the automation of routine tasks, ChatGPT enhances complex analytical activities, such as forecasting, optimization and sentiments analysis. It gives analysts a head start by generating initial insights which they can then refine according to their expertise in the field. This augmentation process enables data analysts to enhance their analytical skills and extract more meaningful insights from the data. Integrating ChatGPT into data analytics processes goes far beyond automating repetitive tasks; it also offers valuable assistance in advanced and strategic analytical activities. Table 5.1 gives an overview of the roles of ChatGPT and analysts in some advanced analytical tasks.

Activity	ChatGPT's role	Data analyst's role
Forecasting	Rapidly process historical data for time-series forecasting using different statistical statistics	Review and refine forecasts based on business knowledge Adjust forecasts to take into account contextual factors
Optimization	Suggest optimizations, such as pricing, inventory levels or data-driven resource allocation	Add critical context to constraints and business objectives to tailor optimizations Provide the necessary logic and reasoning
Sentiment analysis	Perform sentiment analysis on customer feedback or reviews using natural language processing	Validate AI categorizations with a nuanced understanding of the business context

Table 5.1. *ChatGPT and analyst roles in advanced analytical tasks*

Forecasting is an excellent example of this collaboration. ChatGPT can rapidly process historical data and generate time-series forecasts using various statistical models. Analysts can then review and refine these forecasts, integrating their business acumen and knowledge. For example, they can adjust ChatGPT's holiday gross sales forecasts to take into account an upcoming marketing campaign or other contextual factors.

In optimization tasks, ChatGPT can provide valuable suggestions, such as pricing recommendations, inventory levels or resource allocations, based on data analyzed. However, analysts bring a critical context, taking into account commercial constraints and objectives, to tailor optimizations accordingly. While ChatGPT handles the mathematical aspects, human analysts provide the necessary rationale and reasoning.

Similarly, ChatGPT's natural language processing capabilities enable it to perform sentiment analysis on customer feedback or reviews (Sedkaoui 2023). However, it is the analysts' nuanced understanding of companies that enables them to accurately validate and contextualize the AI's categorizations.

By offloading the repetitive aspects of the analyst role to ChatGPT, data analysts can concentrate on higher-value work that requires their specialized skills. This shift in focus enables them to engage in strategic decision-making and model governance and to provide essential support to the organization's key stakeholders.

Apart from that, ChatGPT can also:

- quickly run simulations and create models to test different scenarios and hypotheses. Analysts provide key inputs based on business knowledge that improves the relevance and accuracy of models. In addition, ChatGPT can analyze data to flag potential anomalies using statistical tools such as z-scores. However, analysts need to validate which reported anomalies warrant investigation based on their understanding of the data and business context;

- in decision-making, suggest data-driven decisions, such as which products to promote or which campaigns to investigate. However, analysts evaluate these recommendations against real-world constraints and objectives for which AI lacks visibility;

- generate reports with summaries of insights, streamlining the reporting process. But analysts need to organize, refine and contextualize reports for different audiences and decisions, ensuring that the right message is conveyed.

The main theme is that ChatGPT enables analysts to increase their impact by automating busy work and providing analytical safeguards. It significantly reduces the time spent on repetitive tasks, enabling analysts to focus on higher-added-value activities that require human expertise and creativity. However, its recommendations must be tempered by human analysis to lead to sound decisions. This symbiosis enables organizations to make the most of AI's capabilities and human expertise, fostering a powerful collaboration that improves the overall data process.

The combination of ChatGPT's computational speed and an analyst's specialized skills creates an augmented analytics process with new potential. Organizations that embrace AI for automation while amplifying human strengths will excel in data-driven decision-making. This seamless integration enables data-driven decisions to be made more effectively and efficiently, ultimately leading to competitive advantages and better business results. As data analytics continues to evolve, the synergy between AI automation and data scientists will be essential to unlock new levels of insight and innovation.

5.3. From statistics to predictions: ChatGPT as the partner of choice

In most organizations, descriptive statistics form the basis for more advanced predictive analyses. By analyzing historical data to uncover patterns, trends and relationships, statistics help us to better understand the past (Delen and Demirkan 2013). Predictive analysis then exploits this statistical information to forecast future results.

Descriptive statistics enable a detailed examination of a company's performance. Measures such as averages, distributions, correlations and regressions quantify what has happened. Statistical modeling techniques, such as linear regression and cluster analysis, reveal the factors that determine results.

This statistical information sets the way for predictions. By identifying the strongest relationships and patterns in past data, we can create models that exploit these insights in order to anticipate future results. Predictive analysis applies machine learning algorithms to historical data, in order to predict outcomes such as sales, demand, prices, equipment failures and so on.

However, the predictions are not as good as the underlying statistics. Poorly conducted descriptive analysis will fail to uncover important variables and relationships. This makes predictive models ignore key factors. The best predictions combine robust statistical analysis with intelligent modeling techniques.

Organizations need to view statistics and forecasting as an integrated *continuum*, not as isolated practices. A solid statistical foundation puts more accurate forecasts on the map. And predictive models reveal new questions to be explored statistically. By linking these capabilities, companies can convert data into actionable forecasts that generate competitive advantage.

From statistics to predictions, ChatGPT and other generative models have established themselves as the partner of choice for data analysts. Although traditional statistical methods have formed the basis of data analytics for decades, they often fail when it comes to complex, large-scale datasets and the need for predictive insights.

ChatGPT, with its language understanding capabilities and access to vast knowledge, enables analysts to go beyond the constraints of statistical analysis and to delve into the realm of predictive modeling. By exploiting historical data, ChatGPT can quickly generate time-series forecasts, identify patterns and even assist in sentiment analysis. ChatGPT's fusion of human intuition and computational prowess enables analysts to unlock the potential of data-driven decision-making, paving the way for better-informed strategic choices and more accurate predictions for the future. It acts as an essential assistant in the transition from statistical analysis to predictive modeling.

Organizations leveraging ChatGPT to augment their analytical journey, from descriptive statistics to predictive analytics, will reap major benefits. In what follows, we explore best practices for using ChatGPT to enhance analytics across three key phases: statistical foundations, predictive modeling and model deployment.

5.3.1. Establishing statistical foundations

Through exploratory data analysis using statistics, ChatGPT provides the fuel for effective predictions. ChatGPT accelerates the process to uncover key insights, such as correlations, drivers and important variables in historical datasets. Its natural language processing allows analysts to explore data interactively. For example, a data analyst might ask ChatGPT questions such as: “What are the main factors correlated with customers unsubscribing in our subscriber data over the past two years?”; “Which customer segment has the highest average basket size, and what are the distinguishing features of this segment?”; “Is there a statistically significant difference in conversion rates between customers referred via social networks and search engines?”; “Identify outliers, anomalies or data quality issues in this dataset to be resolved”.

However, human monitoring is essential to ask the right questions (prompts) that are relevant to the business context. Analysts must also validate that ChatGPT’s analysis is statistically valid. Iterative probing with different questions enables analysts to create a complete statistical profile of the data.

With large, complex datasets, ChatGPT accelerates this discovery process exponentially while benefiting from human advice. Key statistical insights discovered becomes candidate variables for predictive modeling.

5.3.2. Building models with ChatGPT

Once key variables and relationships have been statistically identified, analysts can leverage ChatGPT to build predictive models. Its computational speed enables rapid prototyping and testing of different algorithms to find the best fit. Common predictive modeling techniques that ChatGPT can implement or support include:

- *Linear regression* predicts continuous results, such as sales or demand, as a function of correlated variables. ChatGPT can analyze historical data to automatically generate linear regression models. For example, data analysts can ask ChatGPT to automatically generate a linear regression to predict product demand based on historical sales, price seasonality, promotions and other factors. ChatGPT returns the model equation with variable coefficients for review. Analysts can then refine the model by asking ChatGPT to try transformations such as asymmetrical log scaling of variables. This augmented approach leads more quickly to well-fitted models.

- *Logistic regression* predicts binary outcomes, such as customer churn rate, using independent variables. ChatGPT can also create logistic regression models for binary classification use cases, such as predicting customer attrition. For example, data analysts could provide customer data with churn labels and ask ChatGPT to

identify the most predictive inputs to classify churn versus non-churn with logistic regression. It can determine income, purchase frequency, support tickets and other more important markers to model churn probabilistically. ChatGPT returns the logistic equation for inspection.

- *Decision trees* predict results by splitting the data based on optimal cut-off points in the variables. ChatGPT can build decision tree logic based on analysts' needs. For segmentation use cases, analysts can leverage ChatGPT to automatically generate decision tree models to divide customers into groups of people based on attributes. It efficiently determines the optimal sequence of attribute thresholds to segment the population. Analysts can assess tree depth, pruning needs and overfitting risks.

- *Random forests* improve accuracy over single decision trees by aggregating many trees. ChatGPT can also construct random forest models that aggregate many individual decision trees, using primed data samples to classify results as fraudulent transactions. This provides more robust predictions than a single tree. Analysts can work with ChatGPT to adjust the number of trees and variables per tree, to optimize accuracy. ChatGPT can quickly generate random forest models by building many individual trees using bootstrapped samples of the training data.

- *Neural networks* mimic human cognition to discover complex non-linear relationships. ChatGPT can suggest appropriate network architectures and hyperparameters to test.

For each technique, ChatGPT can provide model templates as a starting point.

Data analysts then evaluate the accuracy of the model on the test data and ask ChatGPT to adjust hyperparameters and try additional input variables to improve predictive power.

By automating the tedious task of testing different algorithms and configurations, ChatGPT enables analysts to take a high-level strategic view of maximizing model accuracy. This mixed approach results in more robust models.

5.3.3. Model deployment and monitoring models

Once a satisfactory model has been developed, data analysts can work with IT teams to integrate the models generated by ChatGPT into business applications for production use. However, deployment is not the end of the predictive journey. To maintain accuracy over time, human analysts must:

- continuously monitor model performance as new data emerges by tracking measures such as precision, recall and F1 scores;

- regularly recycle models on new data to incorporate the latest patterns and trends. ChatGPT accelerates the recycling process;
- implement systems to detect concept drift – when model accuracy deteriorates due to changing real-world conditions;
- adjust models in response to new business needs and objectives;
- implement governance protocol models based on ethical and compliance safeguards;
- document model limitations, assumptions and other details for stakeholders.

Proper implementation requires combining ChatGPT capabilities with human judgment, business knowledge and monitoring. Automated monitoring and retraining pipelines need to be complemented by human audits.

5.3.4. Business impact

Companies that successfully integrate ChatGPT's statistical calculations with human analytical thinking can unlock numerous benefits that drive growth and success:

– *Accelerated insights*: thanks to ChatGPT's ability to rapidly analyze variables statistically and create robust predictive models, organizations can gain insights at a much faster pace. This accelerated data exploration enables them to make informed decisions quickly, giving them a competitive edge in dynamic markets.

– *Democratized analytics*: by empowering business analysts to augment their capabilities with AI, organizations reduce their reliance on data specialists. ChatGPT's user-friendly interface and conversational nature make it accessible to a wider range of professionals, democratizing the data analysis process.

– *Improved forecasting*: the combination of human expertise and AI capabilities results in more accurate demand forecasts, propensity models and other predictions. By harnessing the strengths of both, organizations can make informed strategic decisions and anticipate future trends more effectively.

– *Optimized decisions*: data-driven insights and predictive intelligence enable companies to make timely, optimized decisions. Whether it is resource allocation, pricing strategies or marketing campaigns, these decisions are based on robust data analytics, minimizing risk and maximizing returns.

– *Competitive advantage*: with the ability to adapt quickly to market changes and anticipate future trends, organizations gain a competitive edge. ChatGPT's support

for predicting customer preferences, market changes and emerging opportunities helps them stay one step ahead.

The combination of computational power, human creativity and business logic is becoming a winning formula for predictive success, enabling companies to thrive in an increasingly data-driven world.

5.4. Deciphering feelings: text and sentiment analysis with ChatGPT

Sentiment analysis, as a process of assessing the emotional tone of textual content (Sedkaoui et al. 2022), is of immense importance in today's data-driven world. By identifying words and phrases that drive positive, negative or neutral sentiments, this powerful technique provides valuable insights into how people perceive products, services and brands. According to a study by Bain & Company, the adoption of sentiment analysis solutions was expected to reach 80% among companies by 2023.

One of its main applications lies in the analysis of customer feedback, where companies can discover areas for improvement by prioritizing and addressing common complaints. In addition, sentiment analysis has been widely used in social media monitoring. By tracking mentions of a company or brand, organizations can assess public sentiment and track changes over time.

This real-time feedback enables them to react proactively and improve their reputation management strategies. Speaking of brand reputation management, sentiment analysis acts as a crucial ally in identifying negative sentiment on social media platforms, enabling companies to quickly mitigate potential damage.

In addition to brand-related issues, sentiment analysis has many applications in market research. By deciphering prevailing feelings towards products or services, companies can gain valuable insights into consumer preferences and trends. These insights enable them to make data-driven decisions and refine their offerings to better meet customer demands.

By understanding customer sentiments, companies can tailor their offerings to better meet customer needs, fostering stronger customer relationships and loyalty. This valuable insights can also be used to fine-tune marketing campaigns, ensuring that they resonate with the target audience and generate greater engagement.

This enables companies to discern changes in customer preferences, helping them to decide when to launch new products or services, adjust their marketing strategies or take corrective action if necessary. This predictive aspect of sentiment

analysis enables organizations to stay one step ahead of the competition (Sedkaoui et al. 2022), anticipate market trends and remain agile in a dynamic business environment.

One notable advantage of sentiment analysis is its role in brand reputation management. By quickly detecting negative feelings towards a brand, companies can proactively address concerns, mitigate potential damage and improve brand reputation. This proactive approach not only builds loyalty with existing customers, but also improves brand image, attracting new customers who perceive the company as responsive and customer-centric.

Organizations leveraging generative AI for textual analysis open up new horizons for understanding sentiments, topics, trends and relationships in unstructured textual data (McKinsey 2023). Whether analyzing customer feedback, social network discussions, call center logs or employee communications, ChatGPT, in collaboration with human experts, can extract invaluable insights from text that guide strategic decision-making. This detailed guide explores techniques and best practices for text mining with ChatGPT.

Table 5.2 illustrates some examples of the different applications of ChatGPT in sentiment analysis.

Application	ChatGPT’s role	Example
Data pre-processing	Improve sentiment analysis by cleaning up and normalizing unstructured data for greater consistency and quality	ChatGPT corrects: “This producTisgood! #ILoveIt” with: “This product is good! #I Love it”
Feature extraction	Excel at extracting relevant features from textual data, enabling sentiment analysis systems to distinguish unique nuances in customer-brand communication	ChatGPT extracted from this text: “The user interface is intuitive and navigation is user-friendly”, the following characteristics: “user interface”, “intuitive”, “navigation”, “user-friendly”.
Contextual understanding	Analyze data contextually with industry jargon and unique relationships to effectively distinguish positive, negative and neutral sentiments	ChatGPT summarized this text: “Their customer support team was responsive and solved our problems quickly”, with the following context: “Positive feedback on a company’s customer service”. So, the overall feeling is: positive

Training data generation	Generate synthetic textual data with various sentiment labels, enhancing existing training datasets or creating new models for better sentiment analysis	From this text: “The after-sales service was responsive and solved our problem quickly”, ChatGPT indicated a summary: “Positive”. From this text: “The product does not meet expectations, and customer service is difficult to reach”, it indicated a summary: “Negative”.
Sentiment classification	Use ChatGPT as a classifier, directly predicting the sentiment of a given text	From the following entry: “The app’s new user interface is confusing, and it now takes longer to complete tasks”, ChatGPT predicted a sentiment with: “Negative”
Multiple language support	Use for sentiment analysis in different languages, enabling companies to reach wider markets and regions	Thanks to its translation capabilities, ChatGPT was able, from the text in French: “ <i>La dernière sortie de produit de l’entreprise a dépassé toutes les attentes et a reçu des critiques élogieuses de la part des clients</i> ”, to propose an English translation: “ <i>The company’s latest product release exceeded all expectations and received rave reviews from customers</i> ”, and indicated a “positive” overall sentiment
Real-time analysis	Enable real-time immediate sentiment analysis for fast data-driven decisions	Monitoring of a company’s Twitter account for customer comments: “@company Your customer service is terrible, I’ve been waiting for a response for days!” ChatGPT identified a “negative”, sentiment and defined an action to be taken as follows: “The customer support team must immediately respond to the dissatisfied customer and solve their problem”.

Table 5.2. *Examples of ChatGPT application in sentiment analysis*

In each case, ChatGPT’s natural language capabilities provide a scalable baseline, while human expertise focuses on strategic analysis and continuous improvement. AI handles repetitive language-based tasks, while humans provide the contextual understanding to turn information into action. This symbiosis enables organizations to extract the full value from unstructured textual data, giving them a competitive advantage.

It should be noted that ChatGPT can also be used for emojis in sentiment analysis. Emojis can convey valuable information about the sentiment of a message. For example, positive emotions can be expressed with emojis like:

😊 😄 😁 😊 😇 👍 🤗 ❤️ or 100

while negative emotions can be conveyed by emojis such as:

😞 😓 😔 😭 😡 🙄 😬 ❤️ or 😏

Although ChatGPT has been trained on a wide range of textual data to recognize and process certain emojis, its understanding and interpretation of emojis may not be as complete as with plain text, particularly when it comes to more complex or less common emojis.

To confirm this, we prompted ChatGPT to analyze the sentiment of this customer review, taking into account the emojis used:

😏 The new product is 🤢

Its response is shown in Figure 5.3.

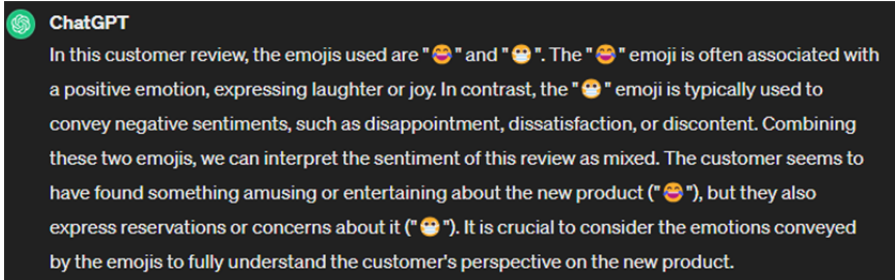


Figure 5.3. *ChatGPT's response for sentiment analysis*

In this situation, we find that ChatGPT's response does not perfectly reflect the customer's sentiment. However, it did advise taking into account the emotions expressed by the emojis to better understand customer sentiment. So, it is important to note that while ChatGPT can be helpful in analyzing emotions and sentiments expressed by customers, human intervention remains essential for an accurate and nuanced interpretation of customer feelings.

In this context, the use of ChatGPT for sentiment analysis presents some key challenges to consider:

– *Understanding sarcasm and subtle sentiments*: ChatGPT often has difficulty detecting sarcasm, irony and more subtle sentiments that require deeper linguistic and cultural context. For example, the phrase “it’s all simply awesome” may be classified as positive, when it was intentionally sarcastic.

– *Dealing with ambiguity*: ambiguous pronouns and references can make it difficult for ChatGPT to determine the true subject of the sentiments expressed. For example, “I can’t believe they did that” isn’t clear about who the pronoun “they” refers to. Here are a few examples of how ambiguous language can confuse sentiment classification:

- polysemy: words with several meanings, such as “light”, “left” or “conforming”, make sentences ambiguous. For example, “the laptop is so light!” is difficult to classify without knowing whether “light” refers to weight or color;

- idioms: figurative expressions such as “break a leg” or “cost an arm and a leg” are difficult for AI to interpret correctly without the cultural context of humans;

- unclear references: when pronouns like “this”, “that” or “it” have no clear antecedents, the subject of the sentiment expressed is uncertain. For example, “I can’t believe they’re allowing this!”;

- weak context: a broad context is often necessary to disambiguate sentiments. The phrase “the movie was sick!” is not clear without knowing that it is a movie review.

By recognizing these sources of ambiguity, we can select training data and create prompts to minimize inaccuracy when using ChatGPT for sentiment analysis. However, despite these improvements, human intervention remains essential to validate and interpret the results obtained by ChatGPT, ensuring a more accurate and insightful analysis of customer sentiments;

– *Detecting negation and emphasis*: the presence of negation words and emphasis, such as “not”, “never” or “extremely”, can alter or reverse the polarity of sentiments. ChatGPT does not always take these modifiers into account.

– *Limited world knowledge*: ChatGPT lacks the real-world knowledge and common sense that can provide critical context for interpreting complex sentiments. Humans intuitively take advantage of this to read between the lines.

– *Data issues and bias*: as a statistical model, ChatGPT risks inheriting and amplifying biases present in the training data. More diverse training data are needed.

– *Evolving language trends*: new slang, cultural references, emojis and creative language can render ChatGPT’s trained interpretations obsolete until retrained. Human analysts adapt more quickly.

We should also not forget cultural nuances, of course, which can certainly complicate ChatGPT’s sentiment analysis in different regions and languages.

Type	Issue
Indirect communication	Some cultures tend to express sentiments more indirectly or implicitly. For example, in some Asian cultures, negative comments may be made tentatively to avoid offending. Subtle cues that modulate sentiment can be missed by ChatGPT
Power distance	In hierarchical cultures, the status and power relationship between communicators impact how openly sentiments are expressed. ChatGPT can miss nuances of authority and deference that modify sentiment
Non-verbal cues	Factors such as tone, volume, facial expressions, gestures and other non-verbal cues provide additional context that facilitates human interpretation of sentiment, but are not captured in text alone
Idioms and historical references	Culture-specific idioms, analogies and historical references familiar to locals, but not to others, can mask sentiments. Meaning is opaque to ChatGPT without broader cultural mastery

Table 5.3. *Cultural challenges to consider in ChatGPT sentiment analysis*

By being aware of its limitations, we can develop strategies to compensate for cases where ChatGPT fails to analyze nuanced sentiments in comparison to human cognition. By training ChatGPT on region-specific data and incorporating feedback from local experts, certain cultural gaps can be filled. But human monitoring is always crucial when analyzing sentiments in various contexts.

Therefore, the involvement of data analysts is essential to train ChatGPT to properly handle these nuances, as well as the various challenges discussed above. With human guidance in data preparation, model development and continuous improvement, ChatGPT can become more aware and competent at disambiguating complex sentiments. The symbiosis of human and AI strengths leads to a more reliable and insightful analysis.

After exploring how ChatGPT can augment human analysts when it comes to key techniques such as statistical analysis, predictive modeling, text mining and sentiment analysis, we now take a closer look at best practices for developing, optimizing and governing predictive models by taking advantage of generative AI in Chapter 6.

Prediction and Modeling with ChatGPT

In Chapter 5, we explored ChatGPT's capabilities as an analytics assistant for techniques like statistical modeling, predictive modeling and text mining. However, building an initial model is only the first step. To extract lasting value from predictive modeling, organizations need strategies for continuous improvement, monitoring, responsible governance and integration into business processes.

This chapter provides best-practice advice on how to go beyond basic model development and create sustainable, ethical and optimized predictive engines with ChatGPT. By adopting a human-centric, iterative approach, we can improve model accuracy over time through feedback loops. As new data emerge, we can detect performance drift and retrain models to keep them relevant.

With responsible AI protocols that emphasize transparency and bias detection, we can build reliable predictive tools aligned with business needs. And by properly documenting, validating and monitoring model inputs, outputs and impacts, we responsibly integrate the resulting insights into organizational decision-making flows.

ChatGPT's expertise in rapid model development is powerful, but human collaboration is crucial to maintaining accuracy, ethics and strategic relevance. This fusion of AI and human allows organizations to extract lasting business value from predictive capabilities.

6.1. Automating the data analysis process with ChatGPT

Organizations striving to maximize the value inherent in their data recognize the indispensable role of analysis processes. These structured procedures serve as the backbone to ensure that data are systematically collected, processed and subjected to analysis, ultimately promoting consistency and efficiency. As a result, the dividends reaped include enhanced decision-making capabilities, refined operational efficiencies and an enhanced innovation environment.

Integrating a data analytics workflow offers a range of benefits. Of particular note is the increased accuracy and consistency achieved through well-defined processes, leading to more coherent data collection, processing and analysis. This, in turn, amplifies the accuracy of results. In addition, the presence of workflow produces a double effect, improving efficiency both in the data analytics process itself and freeing up resources for essential tasks. Furthermore, it weaves crucial insights together, enabling organizations to make more informed decisions.

This synergy contributes to greater operational efficiency and fertile ground for innovation. It is important to note that the incorporation of such a workflow acts as a safeguard against potential errors and pitfalls. Adherence to structured processes helps organizations mitigate the risk of financial loss and reputational damage.

The importance of data analytics to inform sound business decisions is universally recognized, but the process often involves a substantial investment of time and manual effort (Henke et al. 2016; Sedkaoui 2018a). It is at this stage that the concept of automating data analytics workflows emerges as an essential solution.

Automating data analysis workflows offers a multitude of benefits that impact all business operations. In particular, it offers a means of saving time and resources, thus promoting operational efficiency.

A facet of paramount importance is the improved accuracy and consistency that automation brings to data analytics. Through the application of standardized procedures, automated workflows ensure that data collection, processing and analysis occur in a uniform manner. This uniformity results in increased accuracy and reliability for results, mitigating the variability that human intervention can introduce.

Furthermore, the risk reduction inherent in automating data analytics workflows cannot be overestimated. The risk of error is inherently reduced when relying on automated processes that adhere rigorously to predefined protocols. This risk reduction spreads its protective umbrella over companies, protecting them from financial setbacks and preserving their hard-won reputation.

Significantly, the impact of automated data analytics workflows resonates deeply in the corporate decision-making landscape. Rapid access to timely, well-organized information enables organizations to make informed decisions with greater confidence (Delen and Demirkan 2013; Sedkaoui 2018b). By presenting data-driven insights in an easily accessible way, automation facilitates more agile and reactive decision-making, effectively improving the effectiveness of business strategies.

However, an important part of the data analytics process involves repetitive and procedural tasks that often consume analysts' time and energy. These routine activities, while necessary, tend to hinder analysts from engaging in more meaningful and interpretive work. In response to this challenge, ChatGPT is emerging as a transformative solution with the potential to revolutionize data analytics workflows.

ChatGPT's capabilities extend to different stages of the data analytics process detailed in previous chapters of this book. One notable area is data preprocessing, where ChatGPT can speed up laborious data manipulation tasks. The model's ability to recognize relationships between disparate datasets and join them efficiently, identify and rectify duplicate entries, handle missing values and facilitate conversions and formatting of data types offers a remarkable acceleration of the data preprocessing phase.

In the field of exploratory analysis, ChatGPT also proves invaluable in providing an overview of the data. By autonomously generating summaries, distributions and statistics, it provides data analysts with a preliminary understanding of the characteristics of the dataset. In addition, the model's ability to identify outliers, anomalies, patterns and correlations streamlines the preliminary exploration process.

The symbiotic collaboration between ChatGPT and data scientists also extends to model development. In addition to automating testing and predictive modeling algorithms comparison, ChatGPT can suggest evaluation metrics and appropriate validation datasets. This increase in the model development phase speeds up decision-making and lays the groundwork for more efficient model building.

Even in the area of reporting, the impact of ChatGPT is palpable. The model's ability to generate initial data summaries, suggested basic visualizations and preliminary draft reports allows data analysts to refine and improve their findings, effectively streamlining the data reporting generation process.

This automation frees up data analysts for additional interpretation and monitoring such as prioritizing the most relevant relationships, determining appropriate data transformations, assessing model limitations and contextualizing

results into insights. So, the marriage of AI and human expertise enables organizations to harness the dual strengths of both entities.

Workflows	ChatGPT’s role
Data processing	Accelerate data preparation with commands such as: “Join the customer table to the sales table using the customer_id key”; “Delete all duplicate lines in this dataset based on the product_id field”; “For missing values in the price column, impute the average price”
Exploratory analysis	Provide data summaries quickly: “Analyze the breakdown of sales by product category and highlight any outliers”; “Calculate the correlation coefficient between marketing expenditure and customer acquisition”
Model development	Propose modeling approaches: “Recommend a suitable model to predict customer churn based on this customer data”; “Suggest metrics for evaluating the performance of this classifier on unseen data”
Reports	Write basic results: “Summarize the main findings of sales data analytics”; “Generate a bar chart showing sales by product for the last quarter”

Table 6.1. Enhanced data analytics workflows with ChatGPT

6.2. ChatGPT for accurate and reboust modeling

In the field of data-driven decision-making, the importance of accurate and robust modeling is a cornerstone for companies seeking to navigate the complexity of their operations. The ability to build models that are not only accurate but also resilient becomes essential for deciphering trends, forecasting demand and predicting future events. Robust models, capable of withstanding changes in data dynamics and environmental factors, provide a solid foundation on which companies can anchor their decisions.

The benefits of cultivating accurate and robust models have a major impact on business operations. In particular, these models act as guideposts for better decision-making. Armed with accurate information, companies are empowered to embark on paths underpinned by informed choices, steering them towards results with greater efficiency and strategic resonance.

Furthermore, the role of robust models in risk reduction cannot be underestimated. By anchoring decisions in data that withstand scrutiny and fluctuation, companies can bypass the pitfalls of uninformed choices (Morabito 2015; Sedkaoui 2018a), protecting their businesses from eventual setbacks.

Operational efficiency, too, finds an unfailing ally in the form of accurate, robust models. Their ability to automate tasks and reduce the need for manual intervention not only streamlines operations but also frees up valuable resources that can be better used elsewhere, promoting a more optimal allocation of organizational capacity.

In customer service, the impact of accurate, robust models results in a high capacity to respond rapidly to customer concerns. Armed with reliable insights, companies can deliver effective solutions, improve the overall customer experience and foster lasting relationships.

To create accurate, robust models, several essential steps come into play. Above all, the use of relevant data lays the foundations for model accuracy. The quality of the data used for training is a decisive factor in the model's final performance. At the same time, the choice of the appropriate model type appears to be a crucial consideration, with the optimal model varying according to the specific characteristics and objectives of the data.

In addition, validation of the model's accuracy through tests on new data and its continuous monitoring ensure the reliability of the model's predictive capabilities. Evolving data and operational dynamics require ongoing commitment to ensure that models remain relevant and effective.

The cornerstone of effective AI-driven modelling, such as ChatGPT, is the quality of the training data it receives. This fundamental input not only shapes the model's understanding but also influences the biases and gaps it might manifest (Osoba and Welser 2017). The role of data analysts becomes essential in this context, as they assume responsibility for meticulously examining the data before introducing it into the system. This verification process is essential to maintain the integrity and accuracy of the resulting models. In addition, human monitoring guides ChatGPT in the adoption of advanced techniques, such as data augmentation and strategic oversampling. These methodologies enhance the presentation of diverse perspectives in the data, mitigating the risk of biased outcomes and improving model's robustness.

Selecting an appropriate model is another critical phase, where ChatGPT's prowess in rapidly prototyping multiple algorithms is plied. This exploratory capability makes it possible to empirically determine the optimal approach, taking into account complex factors such as the nuances of datasets, prediction objectives and performance criteria. Nevertheless, it is the human element that introduces

nuanced discernment. Human monitoring serves as a safeguard, ensuring that the chosen model aligns with the specific requirements and limitations of real-world business contexts. By considering the strengths and weaknesses of each model in the context of business objectives, human support prevents misalignment and promotes efficient use.

The journey does not stop once a model has been developed. Rigorous validation is the foundation of model reliability. Using appropriate tests based on representative datasets, analysts assess the model’s accuracy and generalizability. In addition, the essence of continuous monitoring comes into play. Performance dashboards, combined with bias detection and drift identification mechanisms, inaugurate a continuous feedback loop. This iterative refinement process maintains model integrity and accuracy even as conditions change over time.

Table 6.2 provides a summary of how ChatGPT can help reduce bias, assess generalizability and build confidence in models.

Objective	How ChatGPT can help	Example
Mitigating bias	Identify under-represented groups in the data, suggest controlled experiments to quantify bias, use debiasing algorithms	“Compare the accuracy of this classifier across gender and highlight disparities exceeding 5%”.
Assessing generalizability	Quickly test models on synthetic data, simulate future scenarios, allow quick retraining as new data emerges	“Run this forecasting model on simulated sales data with a 10% price increase”.
Building trust	Provide explanations of model logic and predictions, quantify uncertainty, enable in-depth analysis of training data	“Explain the logic behind this ad targeting model and its predictions”

Table 6.2. *Using ChatGPT to reduce bias, assess generalizability and build confidence in models*

In particular, strategies focused on minimizing bias, maximizing generalizability beyond training data and emphasizing transparency are essential for deploying reliable and ethical AI systems. The confluence of these strategies with ChatGPT capabilities ensures that the models not only excel in theoretical constructs but also resonate in the practical context of business operations.

While this detailed guide explores techniques for reinforcing, testing and establishing confidence in the models generated by ChatGPT, further bridging the gap

between theoretical potential and real-world effectiveness, it is important to note that ChatGPT itself continues to evolve rapidly. In particular, the move from GPT-3.5 to GPT-4 brings significant improvements to multimodal data analytics. GPT-4 demonstrates greater versatility integrating different types of data, and greater expertise in analyzing both structured and unstructured data (Open AI 2023).

Plugin integration also unlocks more visualization, automation of workflows and analytical enhancements for GPT-4. All in all, its underlying architecture seems more suited to an adaptable setup for different industrial domains and use cases.

Feature	GPT-3.5	GPT-4
Data sources	Text-based	Multimodal (text, images, audio)
Plugins and integrations	Limited integration	More seamless analytics plugins
Structured data analysis	Basic capabilities	Advanced SQL and relational understanding
Unstructured data analysis	Basic NLP	Richer semantic understanding
Creative ideation	Limited generation	Improved abstraction and insight connections
Personalization	Minimal beyond prompts	More adaptable to different fields
Explainability	Opaque logic	Slightly improved, but still limited

Table 6.3. *Performance of GPT-3.5 and GPT-4 for data analytics*

Of course, GPT-4 presents definite advances for data analytics. Nevertheless, challenges remain as to the interpretability of its reasoning and the limits of its contextual understanding. Despite its enhanced capabilities, human cognition remains relevant to judiciously guide this emerging AI. Although limitations remain, this evolution testifies to ChatGPT’s continuing potential. Our quest to develop trustworthy AI must therefore progress hand in hand with the advancement of technical capabilities.

Essentially, ChatGPT equips organizations with powerful tools, but it is the fusion of human governance that cements quality control. For example, ChatGPT can rapidly test models to detect bias. But analysts must determine appropriate checks and thresholds based on standards and risk appetite. ChatGPT can simulate hypothetical scenarios, but strategists need to develop relevant cases reflecting potential futures.

ChatGPT can quantify uncertainties, but governance teams need to define transparency standards.

This collaboration capitalizes on the scalability of AI, enhancing modeling capabilities while simultaneously integrating accuracy, ethics and relevance of the real world through the prism of human guidance. The partnership between AI and human expertise manifests itself as a driving force behind sustainable value creation resonating in the fields of data-driven decision-making.

6.3. Continuous improvement: optimizing model capabilities through feedback loops

Have you ever considered the precise essence of continuous improvement? It is a systematic approach to identifying ways of streamlining processes and minimizing inefficiencies. Continuous improvement is a fundamental principle that drives the evolution of processes and products.

One of the key driving forces of continuous improvement lies in harnessing the power of data and analysis.

Within the domains of each organization, a vast repository of data covers diverse facets, encompassing financial records, sales metrics, supply logs, investments, workforce management, projections and so on. This wealth of information gives companies a remarkable potential for extracting meaningful insights regarding their operations, and transforming raw data into lucrative opportunities. There is no doubt that data analytics enables organizations to make informed decisions, a fundamental benefit of this analytics-driven business era (McAfee and Brynjolfsson 2011; McKinsey 2016).

Reflecting on historical data and the use of statistical methodologies proves extremely valuable, particularly when it comes to discerning the superior course of action between alternatives. However, the real significance is amplified when we look at the ongoing monitoring and impact assessment, revealing the repercussions these decisions create on the operational landscape.

A significant proportion of operations and choices involve patterns of repeatability. Let us consider the case of establishing an optimal stock level using historical data – a model can be formulated to determine the ideal stock level. This same model can then be reapplied to assess the current state of the stock and monitor results over time.

Keeping a watchful eye on results is not just a matter of validating the accuracy of these models and tools. It extends to ensuring their ongoing relevance and effectiveness, by encouraging the necessary adaptations in response to evolving business needs. This perpetual cycle of review, refinement and adaptation is at the heart of continuous improvement, making organizations agile and resilient in the face of change.

In the dynamic field of machine learning, the concept of continuous improvement manifests itself in an iterative path of refinement, ensuring that models not only meet the initial reference criteria, but progress continuously to meet changing needs.

At the heart of this journey is the process of collecting feedback. This stream of feedback emanates from a variety of sources, each providing unique insights into a model's performance and relevance. Measures such as model accuracy, performance on validation data and real-world user experience all contribute to the mosaic of information that guides the path to improvement. Furthermore, domain experts play a central role in providing nuanced perspectives, further enriching the feedback loop.

Armed with this feedback, the next step is to identify opportunities for improvement. This is an essential step in model optimization and may involve retraining with new, more relevant data, adjusting model parameters and even redesigning the architecture to adapt to changing needs. This iterative process ensures that models remain aligned with the evolving data landscape and dynamic business requirements.

The importance of improvement in the context of machine learning is vast and far-reaching (Bishop 2006). It has the potential to strengthen the very basis on which models are built:

- *Higher accuracy and reliability*: as models are continually refined, their accuracy and resilience increase, ensuring that they consistently meet or exceed performance standards.

- *Unleashed efficiency*: the timely identification and resolution of bottlenecks or performance gaps ensures that models work more efficiently.

- *User-centric orientation*: user feedback, combined with iterative improvements, leads to models that are better adapted to user needs, promoting a transparent and efficient user experience.

- *Nurturing trust*: by visibly investing in continuous improvement, organizations nurture user trust in the reliability and pertinence of their models.

As the continuous improvement journey continues, it also requires a well-designed approach:

– *Incremental progress*: starting small and building gradually enables targeted improvements without disrupting the wider process.

– *The power of iteration*: recognizing that transformation takes place over several iterations means we cannot expect overnight transformations.

– *Data as a North Star*: data-driven insights serves as a compass, guiding improvement decisions and tracking progress.

– *The collaborative advantage*: involve stakeholders, users and experts in the process harnesses collective wisdom to guide the journey towards refinement.

Essentially, continuous improvement of machine learning transcends mere theory; it forms the foundation of a dynamic and responsive machine learning ecosystem. Thanks to this process, models remain adaptive, relevant and impactful, closely aligned with real-world dynamics and realizing their full potential.

In the dynamic AI landscape, keeping systems accurate and relevant is a perpetual challenge, as constantly changing conditions inevitably lead to performance erosion. ChatGPT can support the continuous improvement fueled by data analysis, in the following ways:

- quickly processing and summarizing historical data to discover insight;
- automating core analytical tasks such as data cleaning, aggregation and statistical modeling;
- prototyping analytical models for forecasting, optimization, etc.;
- backtesting models on historical data to assess accuracy;
- continuously monitoring live data and model outputs to detect deviations;
- promptly retraining models on new data when a drift is detected;
- allowing rapid iteration and testing of model improvements;
- generating explanations and metrics that reinforce confidence in the model's output;
- helping data analysts focus on high-added-value interpretation and oversight.

In order to ensure sustainable value and to keep up with these changing dynamics, it is imperative to foster a culture and establish an infrastructure that supports continuous model optimization. In this context, ChatGPT's role evolves from that of a simple tool to that of a catalyst for continuous improvement processes.

A crucial element in this undertaking is the diligent monitoring of the model performances, as a safeguard against any deterioration:

- Continuous assessment of the accuracy of predictions on live data streams serves as an early indicator of deviations from acceptable thresholds.
- Data validation routines play a crucial role in identifying disparities between production and training data, exposing gaps that require attention.
- Regular bias audits highlight any biased results, enabling timely corrective action.

By actively monitoring models, potential pitfalls are identified early, enabling prompt intervention before minor hiccups snowball into larger problems.

At the heart of continuous improvement is the essential integration of feedback, encompassing both explicit and implicit inputs, into the refinement process. Customer feedback, whether in the form of corrections or replacements, provides valuable insights that feed back into model retraining and improved accuracy. Evaluations of predictions by domain experts, accompanied by their annotations, contribute directly as signals for improvement. In addition, user behaviors, such as adoption and abandonment patterns, offer subtle yet powerful implicit feedback, providing insight into evolving needs (Sedkaoui 2018a). This interaction of complex real-world feedback lends practical experience to models, ensuring that enhancements remain responsive to real-world demands and aligned with practical requirements.

This real-world feedback loop infuses the models with practical experience, ensuring that improvements are aligned with real-world requirements.

The next steps involve retraining the models using the insights derived from the feedback. The data analytics teams use ChatGPT to speed up the iterative process, using new data to improve models. Updated models are then subjected to rigorous testing before finally being deployed to end-users.

With thought-out workflows, loops such as *new data source* → *monitoring* → *feedback* → *re-training* → *deployment* become ingrained. ChatGPT's speed radically accelerates this training effect. Here, ChatGPT's accelerated capabilities become essential to reinforcing this flywheel effect, radically accelerating the improvement journey.

Here is an example of how ChatGPT can help anchor improvement workflows based on feedback.

A retailer has a machine learning model that predicts customers' willingness to purchase. To maintain accuracy, they establish this workflow:

- *New data source*: daily sales data are uploaded to their data warehouse.
- *Monitoring*: every week, an automated script tests the accuracy of the model on new data.
- *Feedback*: if accuracy falls below 95%, an alert triggers the analysis team to examine a sample of errors.
- *Re-training*: the team uses ChatGPT to quickly retrain the model on new data using natural language prompts.
- *Deployment*: the updated model is seamlessly redeployed to replace the old version via an API call.

This automation allows rapid iteration to maintain accuracy. Analysts leverage ChatGPT's speed to retrain models in minutes rather than days. Minimal overheads make the process sustainable. Over time, the continuous loop of new data to assess performance, gather feedback and improve models via ChatGPT conversations becomes an ingrained practice for maintaining optimal predictions.

Continuous improvement through continuous feedback loops is crucial to maintaining model accuracy and relevance over time. ChatGPT can accelerate and improve this iterative process, but thoughtful design is required to extract maximum value.

Rigorous monitoring provides the basis for detecting declines in accuracy, data drift or the emergence of biases that signal the need for model updates. ChatGPT rapidly processes new data to promptly flag up these problems.

It is equally important to gather qualitative feedback from experts and end-users to identify limitations and opportunities based on their experience. ChatGPT can help collect and synthesize this human feedback through large-scale natural language conversations.

Equipped with rich monitoring insights and human perspectives, analysts can then leverage ChatGPT's speed to retrain models using the latest data and user input. This enables rapid iteration and deployment improvements.

When integrated into streamlined workflows, this feedback-driven flywheel effect can continuously optimize model performance and user value over time, through incremental improvements. Monitoring provides signals to trigger improvement, humans provide the strategic direction and ChatGPT accelerates iteration.

So, the crux of this complex mechanism lies in establishing a symbiotic relationship between end-users and models. Continuous improvement, fueled by robust feedback loops, gives rise to intelligent systems capable of evolving. As a result, these systems not only adapt to changing conditions but also ensure that accuracy and relevance persist over time, cementing their value and impact.

6.4. Trend and time series analysis

IBM's report on generative AI (2023) highlights several ways in which these models are transforming data analytics capabilities. Firstly, generative AI has enormous data requirements, so responsible data analytics requires interorganizational collaboration to feed models. According to the report, 60% of companies currently do not have a coherent approach to managing generative AI and associated data.

Companies that have made the transition to a hybrid cloud data infrastructure are better placed to exploit generative AI, as this requires holistic data strategies. Leaders also see opportunities to accelerate data sharing with partners via generative AI, enhancing collective insights. However, unifying siloed data requires alignment with business objectives and remains an ambitious project.

Next-generation data tools, such as automation and marketplaces, can help organizations ingest, prepare and analyze huge volumes of structured and unstructured data for AI systems (Sedkaoui 2023). Robust data governance practices are essential, given the vast datasets accessed by core models, including collection, storage, access, processing and security protocols.

Time series models can specifically combine powerful temporal analysis with flexible generative techniques to improve forecasting, planning and decision-making.

Already, the use of techniques such as trend analysis and time-series analysis is of prime importance in the domain of improving business performance. These methodologies enable companies to make informed decisions, streamline operations and improve the quality of customer service. The profound impact of these tools on overall efficiency and effectiveness cannot be overestimated.

Trend forecasting offers a window onto the future, enabling companies to anticipate changes in demand, market dynamics and consumer preferences. This foresight is invaluable for developing strategies, allocating resources and staying one step ahead of the competition. Furthermore, time series analysis delves deep into historical data to extract patterns, cyclical behavior and seasonality. Armed with this

understanding, companies can identify underlying performance drivers and optimize strategies to take advantage of them.

Time series data are ubiquitous in all sectors, whether financial data, sales trends, web traffic, sensor readings, etc. The analysis of these temporal data holds enormous potential for monitoring performance, discovering drivers and predicting future results.

However, as with any method, responsible use of these techniques is essential. It is essential to maintain the focus on ethical considerations and user comfort. By adhering to these considerations, companies not only exploit the potential of advanced analysis but also cultivate an environment of professionalism and integrity. The integration of forecasting and time series analysis reflects a commitment to leveraging technology for both operational excellence and ethical conduct.

In addition to predictive modeling, ChatGPT shows promise for these techniques. This ability to process vast volumes of information on a large scale meets the arduous demands of data processing (Sedkaoui and Gottinger 2017; Sedkaoui 2018b), enabling companies to extract meaningful insights from their datasets. ChatGPT accelerates statistically intensive tasks such as:

- decomposing down time series into trend, seasonal and noise components;
- stationarity and autocorrelation testing;
- detecting change points and concept drift;
- identification of cyclical/seasonal patterns and lagging relationships.

This allows data analysts to focus on strategic interpretation rather than mechanical statistics.

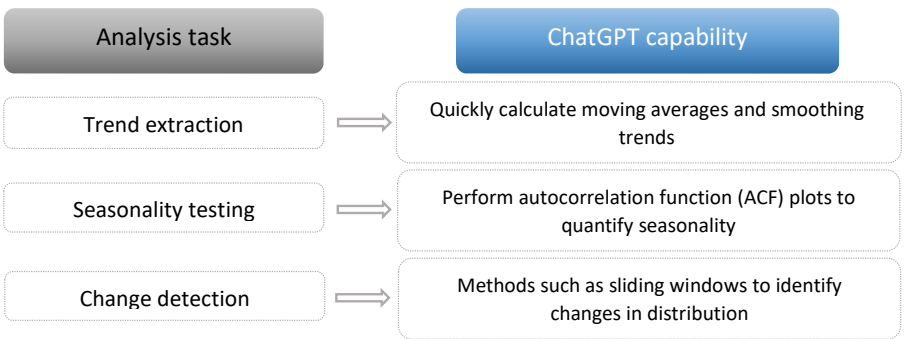


Figure 6.1. ChatGPT's capabilities for time series analysis

Time series analysis is a robust methodology used to examine historical data and forecast future trends. ChatGPT, with its dexterous computational skills, simplifies and accelerates the workflow by automating the laborious tasks of pre-processing, model fitting and prediction.

ChatGPT also excels at automatically analyzing time series data to discover seasonal patterns, autocorrelations, stationarity and other statistical properties (Sedkaoui 2023). This provides a better understanding of how metrics are evolving. For example, data analysts can prompt ChatGPT to assess trends and noise in time series datasets as a precursor to modeling. AI handles the heavy lifting of statistics while humans interpret the insights.

ChatGPT is certainly capable of supporting the generation of Python or R codes, performing time-series forecasting using Prophet, assisting in debugging application errors and managing intermediate data processing steps. Data analysts can then integrate available prospective indicators, event data and subject matter expertise to fine-tune forecasts as required. ChatGPT also enables rapid iteration on different simulation scenarios.

For example, they can forecast vacation sales based on past trends and seasons, then modify them based on known promotions, campaigns and market conditions.

For forecasting, ChatGPT offers a head start in:

- rapidly training predictive models, such as ARIMA and Prophet, on historical data;
- benchmarking the performance of different algorithms to select the best fit;
- generating multivariate models incorporating various predictors.

Data analysts then adjust the raw forecasts by applying their judgment and domain expertise. This balances automation with human oversight. So, it is essential to recognize that although ChatGPT emphasizes competence in several tasks, it may encounter challenges when faced with complex problems that require specialized knowledge or innovative solutions.

Therefore, the role of data scientists remains essential to exploit ChatGPT's potential. They serve as architects of the data mining process, defining clear objectives and contextualizing the results within the business landscape. Their interpretation bridges the gap between raw analytical results and actionable insights, ensuring that anomalies flagged by the system are subject to scrutiny before they lead to erroneous conclusions. Additionally, human judgment is used to align

forecasts with strategic objectives and identify relevant factors in time series analyses, basing the information on real-world relevance.

Well-tailored prompts are the cornerstone of effectively using ChatGPT capabilities in time series data analytics. The provided prompts serve as strategic entry points into the analysis process. Asking questions that revolve around trends, correlations, opportunities, risks, impacts and insights in time series data, users can guide ChatGPT's computational prowess towards uncovering hidden patterns and valuable information. These prompts act as navigational beacons, directing the AI to specific aspects of the data that are important to the users' goals.

The dynamic nature of time series data presents both challenges and opportunities. While the temporal dimension allows the observation of trends and changes over time, it also requires a comprehensive understanding of the context surrounding the data points. This is where collaboration between human expertise and the analytical capabilities of ChatGPT becomes essential. The human touch ensures that nuances, contextual variations and business implications are factored into the analysis, elevating the insights to actionable recommendations.

However, human–AI collaboration is not without its challenges. The prompts provided must be carefully designed, ensuring that they correspond to the specific context of the data and the objectives of the analysis. Although ChatGPT can process data quickly and accurately, it relies on accurate prompts to deliver meaningful results. While the field of AI continues to evolve, refining the interaction between human input and AI output remains a continuous process.

The collaboration between ChatGPT and data analysts does not stop at interpretation; it extends to validation. Data analysts use experimental design principles and team strategies to challenge and test the recommendations issued by ChatGPT. This safeguard not only prevents uncritical acceptance of AI results but also helps minimize potential risks and errors.

Since we have explored strategies and best practices for developing, optimizing and governing predictive models with the help of ChatGPT, we now, in Chapter 7, take a closer look at how this intelligent generative AI agent can be applied to improve and develop specific machine learning techniques and algorithms. Let us explore these potentials for advancing the state of the art of AI through human/AI collaboration.

ChatGPT at the Service of Machine Learning

In Chapters 3, 4, 5 and 6, we explored the applications of ChatGPT for tasks such as data preparation, natural language querying, augmented analysis and predictive modeling. However, the implications extend beyond end-user tools to accelerate the very practice of machine learning research and development.

This chapter shows how generative AI can collaborate with machine learning algorithms, with the help of human intervention, to improve existing techniques and stimulate new innovations.

In this chapter, we discuss high-impact applications, including the use of ChatGPT's ideation capabilities for exploring new model and algorithm architectures for the use of code generation to rapidly prototype conceptual models, the optimization of technical workflows and the combination of generative AI and procedural machine learning for new hybrid approaches. Finally, we look at how intuitive conversational interfaces democratize access to advanced machine learning.

The common theme is to augment human ingenuity with AI capabilities to advance the state of the art and bring new potential to life faster. Just as domain experts can achieve more with the assistance of AI, so too can users of machine learning algorithms, which can also be multiplied by the power of conversational interfaces.

By collaborating with tools like ChatGPT, the machine learning community can open up new horizons of possibility and solve complex real-world problems once considered unsolvable.

7.1. Machine learning in the functional fabric of ChatGPT

Machine learning (ML) is emerging as a transformative facet of AI, granting software applications the remarkable ability to improve predictive accuracy without the need for explicit programming. This discipline exploits the potential of datasets to improve algorithms. A compelling facet of machine learning lies in its foundation on advanced models that facilitate prediction (Mitchell 1997) and decision-making with minimal involvement from the human programmer.

These models operate through complex layers called “transformer blocks” which analyze the sample data and autonomously decipher the models to produce the desired outputs. At their core, machine learning algorithms exploit historical data as input, and then generate predictions for new output values, an embodiment of data-driven intelligence.

ChatGPT itself uses a combination of machine learning techniques in its underlying generative architecture (Gutierrez 2015). It is leveraged for the machine learning domain, drawing on a vast corpus of datasets that underpin its capabilities. In this field, a multitude of language models converge, giving ChatGPT the remarkable ability to generate human-like conversations.

The architects behind ChatGPT seamlessly combine the strengths of supervised and unsupervised learning, a synergy that enriches the chatbot’s learning experience and enhances its performance. This fusion embodies dynamic interaction, making ChatGPT’s results ever more refined.

ChatGPT uses the principles of supervised learning, where input–output relationships are explicitly defined and models are refined using labeled datasets. These algorithms then use this knowledge to predict the outcomes of new and novel inputs. In this training program, ChatGPT uses a strategy known as reinforcement learning from human feedback (RLHF).

The RLHF application is integrated into the architecture of ChatGPT as a subset of machine learning techniques designed to improve the conversational prowess of AI language models. Thanks to a feedback loop continued with human trainers, ChatGPT learns to distinguish between correct and incorrect responses, effectively shaping its conversational behavior.

The adoption of RLHF in ChatGPT serves as a channel for developing responses that resonate with human interaction. This iterative evolution involves supervised fine-tuning, where human AI trainers participate in dialogues embodying both user and AI assistant roles.

Within its repertoire, ChatGPT also capitalizes on unsupervised learning, a paradigm that unlocks patterns in data lacking explicit guidelines for desired outcomes.

In the unsupervised learning context, AI is primed to discover patterns by discerning similarities and disparities in the data. This enables AI to develop precise answers for tasks such as exploratory analysis, image segmentation and the identification of analytical patterns.

ChatGPT's unsupervised framework encompasses three main learning methodologies: clustering, association rules and dimensionality reduction.

Clustering, an essential data mining technique, groups unlabeled data according to their similarities or disparities. ChatGPT's algorithms verify these patterns, categorizing unsorted data into distinct clusters that embody inherent patterns and structures. Exclusive clustering, overlay clustering, hierarchical clustering and probabilistic clustering are among the key clustering algorithms that shape ChatGPT's prowess.

At the same time, ChatGPT incorporates association rules – a rule-based approach that reveals connections between variables within a dataset. This technique, often used in market basket analysis, enables companies to discern interrelationships between various items, optimizing cross-selling strategies and recommendation engines. The annals of association rule generation include estimated algorithms such as *FP-Growth*, *Apriori* and *Eclat*.

For datasets characterized by an abundance of dimensions, ChatGPT makes skillful use of dimensionality reduction. As data proliferate, accuracy increases, as does the complexity of machine learning algorithms, which often complicates data visualization. To counter this, ChatGPT uses dimensionality reduction in its unsupervised framework, keeping the volume of data manageable while preserving the integrity of the dataset.

In essence, ChatGPT harmonizes the duality of supervised and unsupervised paradigms in the field of machine learning. This fusion enables the AI chatbot to distill the complexities of data, providing optimal responses to queries and prompts, while cultivating a dynamic synergy between AI capabilities and human ingenuity.

Table 7.1 summarizes the symbiotic relationship between machine learning algorithms and ChatGPT.

Algorithms	How does ChatGPT use them?	How can they use ChatGPT?
Supervised learning	Task-specific fine-tuning using labeled data	ChatGPT can provide labeled datasets for training models
Unsupervised learning	Pre-training on unlabeled corpora to learn patterns	ChatGPT can assist in clustering and analyzing unstructured data
Reinforcement learning	Optimizing conversational skills with human feedback rewards	ChatGPT can provide a conversational environment for testing agent policies
Neural networks	Basic transformer architecture	ChatGPT can suggest new network architectures for testing
Attention mechanisms	Used in its transformer architecture	ChatGPT can explain how attention works
Learning transfer	Takes advantage of pre-trained parameters for fine-tuning	Knowledge can be transferred from ChatGPT to downstream models

Table 7.1. *ChatGPT and machine learning algorithms: a virtuous circle of co-advancement*

As shown in Table 7.1, ChatGPT combines several machine learning techniques in its design while providing a collaborative tool for advancing machine learning research and engineering lifecycles. The synergistic relationship leads to amplified capabilities on both sides.

In practice, elements of several approaches are often combined. For example, unsupervised learning can be used to extract features, followed by a supervised model for prediction. The diversity of algorithms offers the flexibility needed to adapt solutions to different machine learning tasks and real-world data types.

ChatGPT’s strength lies in its prowess in maneuvering these shifts in tone and style, rooted in its vast exposure to a diverse dataset. Its proficiency is further enhanced by machine learning, which enables it to decipher the complex grammatical underpinnings of language. This ability, fueled by machine learning, enables ChatGPT to harness algorithms to create output that resonates with human-like fluency.

Consequently, the symbiotic relationship between machine learning and ChatGPT's evolution underlines the trajectory towards strong AI. Machine learning is essential in ChatGPT's journey to capture the nuances of language, discern patterns and improve its performance over time.

7.2. Creating new machine learning approaches with ChatGPT

The importance of machine learning approaches extends to multiple dimensions. First and foremost, they serve as essential allies in improving corporate decision-making (Sedkaoui 2018a; Rich and Gureckis 2019). By mining historical data, machine learning algorithms reveal complex patterns and elusive trends that escape manual identification. These insights serve as powerful catalysts for decision-making, in fields ranging from product innovation to the orchestration of strategic marketing campaigns.

However, the power of these algorithms transcends mere decisional prowess. It resonates profoundly with increased operational efficiency. Thanks to the automation of tasks such as fraud detection and customer segmentation, machine learning algorithms are freeing employees from time-consuming tasks. This paves the way for strategic pursuits, propelling companies towards greater efficiency and productivity. The dividend of automation is undeniably transformative as it redirects human cognitive bandwidth from routine tasks to endeavors that require creativity, innovation and high-level strategy (Toubia and Netzer 2016).

Machine learning algorithms can be broadly classified according to their learning style into supervised, unsupervised, semi-supervised and reinforcement learning approaches.

Supervised learning algorithms train models to deduce a target output based on examples of labeled input–output pairs. Classification models, such as logistic regression and random forests that predict discrete targets and regression models that predict continuous values, are common supervised methods.

Unsupervised learning finds hidden patterns and intrinsic structures in unlabeled input data (Radford 2018; Sedkaoui and Khelfaoui 2020). Clustering algorithms, such as k-means, which group data into distinct segments, and dimensionality reduction techniques, such as principal component analysis (PCA), are unsupervised.

Semi-supervised learning combines a small set of labeled data with a larger unlabeled set during training to improve model performance. This uses unlabeled data while avoiding excessive manual labeling requirements.

Reinforcement learning trains models to optimize behavior in complex environments based on a system of given rewards and penalties for actions rather than on labeled examples. Game-playing AIs often use reinforcement learning.

The applications of machine learning extend to various facets of our daily lives. Take, for example, the case of Google Search and other search engines that use machine learning for predictive search and autocomplete features (Sedkaoui 2018a).

In the context of autocomplete, a simple input of initial letters prompts the search engine to present a drop-down menu of potential matches. This feature, launched by Google, has now become ubiquitous among search engines. In addition, companies such as Amazon, Flipkart and Spotify have integrated autocomplete features on their websites.

Predictive search, a facet of machine learning, aligns with autocomplete. It relies on algorithms developed from user behavior analysis, enabling it to predict user actions. Netflix is an excellent example of this approach, as it studies the viewing habits of active users to glean insights about their preferences, then providing personalized recommendations.

ChatGPT's potential for optimizing model performance through similar procedures is remarkable. In the context of machine learning, ChatGPT can streamline autocomplete and predictive search functionalities (Wiles 2023). In doing so, it can improve the efficiency and accuracy of these features, as well as the way they are used in other applications such as improving business workflows, customer service interactions and content generation.

One of the most promising applications of large language models like ChatGPT is collaboration with humans to design innovative new approaches for solving complex machine learning problems. ChatGPT has demonstrated a creative potential that can enhance the ideation process in exploring uncharted territories.

A key advantage is ChatGPT's ability to digest details about the current state of research, data characteristics and task objectives and then propose new model architectures or algorithmic modifications adapted to the use case.

For example, data scientists could describe the limitations of convolutional networks for processing 3D medical imaging data, and prompt ChatGPT to suggest alternative approaches that might excel with volumetric inputs. ChatGPT may propose ideas such as extending 2D convolutions to 3D, or using voxel-based encoders to integrate spatial relationships.

Data analysts can actively discuss the merits, variations and combinations of different conceptual models interactively with ChatGPT via multiple prompts. This fuels new hybrid techniques combining the strengths of existing methods. By externally explaining its natural language creation process (Hilton et al. 2021; Tamkin et al. 2021), ChatGPT enables data analysts to guide and align innovation.

As shown in Table 7.1, it is clear that ChatGPT offers immense potential for optimizing and accelerating applied machine learning processes, including data processing, model creation and machine learning engineering.

For data preparation, ChatGPT can automate the core steps, such as data source linking, cleaning, pre-processing and feature engineering based on natural language prompts. This provides a head start before modeling.

In model building, ChatGPT can rapidly prototype an array of machine-learning algorithms with configurable hyperparameters. This facilitates comparison and the selection of the optimal approach. ChatGPT can also explain model performance and behavior to support tuning.

For machine learning engineering, ChatGPT can generate boilerplate code for pipelines, workflows and monitoring systems that experts refine rather than create from scratch. It can also provide explanations and documentation.

Throughout the lifecycle, ChatGPT enables experts to operate at a higher level of abstraction, while AI manages time-consuming execution based on conversational input. This amplified productivity and speed unleashes new potential in applied machine learning. The future of machine learning engineering lies in the collaboration between human and AI strengths.

So, although they form the basis of ChatGPT itself, machine learning algorithms have much to gain from the features such as intuitive interfaces, automated coding, and mixed-initiative creativity that generative AI offers. ChatGPT promises to take machine learning applications to the next level.

The use of ChatGPT makes the process of creating a machine learning project more transparent than ever. By creating follow-up prompts and evaluating the outputs, data analysts can effectively train the model to meet user queries and provide valuable insights with ease and speed.

Over time, libraries of candidate architectures and algorithms can be accumulated by prompting ChatGPT to generate models for various types of data and machine learning challenges. This code base of launch points accelerates future ideation and experimentation.

By essentially expanding the brainstorming team, ChatGPT has the potential to dramatically expand the creative reach of data analysts and the rates of proposal, analysis and prototyping of new techniques. This accelerated innovation cycle will open up new horizons for machine learning capabilities.

7.3. Boosting machine learning algorithms with ChatGPT

Once new machine learning approaches have been conceptualized, translating the ideas into executable code has traditionally imposed a considerable implementation burden (Sedkaoui and Khelfaoui 2020). Coding of complex neural architectures from scratch can take weeks for data scientists in their machine learning approach. However, large language models like ChatGPT show revolutionary potential for automating this programming work through natural language code generation.

Rather than painstakingly coding new models, they can describe model specifications, data flows and algorithms in plain language. ChatGPT can then generate complete modules and prototypes in languages such as Python, TensorFlow and PyTorch. For example, data analysts can specify the structure of an attention-based RNN model for processing sequential data. ChatGPT can generate executable code implementing the described architecture.

This enables a rapid transition from ideation to implementation, bypassing the heavy burden of manual coding. Other ideas can be empirically tested in parallel. And iterative adjustments are much quicker as results come in. Prototypes serve as functional starting points for data scientists to customize them, rather than creating them from scratch. Over time, libraries of reusable modules can accumulate for common machine learning building blocks.

By automating the routine aspects of programming, generative AI frees human specialists from machine learning to focus on strategic engineering and impactful innovation with regard to mechanical coding. This triggers a wave of experimentation and applications amplifying engineering productivity. Their future role will be to collaboratively guide AI tools to execute technically while solving real-world problems. This cooperation between human and AI strengths will open up new frontiers of possibilities for machine learning.

While generative AI such as ChatGPT is already demonstrating considerable potential for machine learning applications, the implications extend deeper into the formation of the practice itself and advances in the research and development of machine learning. In particular, two high-impact capabilities stand out.

7.3.1. *Optimizing machine learning engineering*

ChatGPT demonstrates deep technical knowledge on subjects such as model architectures, data engineering, machine learning engineering and cloud-based machine learning algorithm infrastructure. This knowledge can be applied to optimize machine learning workflows and machine learning engineering systems.

For example, data analysts can describe their model training pipeline and ask ChatGPT to suggest optimizations such as distributed training approaches, hyperparameter tuning tools and debugging techniques. For production models, ChatGPT can suggest monitoring solutions, data validation checks and retraining rates.

Essentially, it provides a machine-learning consultant, always available to answer questions, generate boilerplate code and ensure best practices. This boosts productivity.

7.3.2. *Hybrid model innovation*

Furthermore, ChatGPT is particularly well placed to combine the strengths of different modeling paradigms into new hybrids. For example, it may suggest integrating the semi-supervised knowledge representation capabilities of graph neural networks with the sequential prowess of LSTMs, or combine deep learning architectures with probabilistic graphical model elements.

Combining the advantages of several techniques is a promising direction for advancing capabilities. By encoding knowledge in machine learning sub-domains,

ChatGPT can synthesize creative hybrid combinations fed by both human-organized expertise and its own emerging connections.

This two-pronged approach of optimizing engineering while innovating in the new frontiers of modeling demonstrates ChatGPT’s expansive potential to multiply the productivity of data analysts in their automatic learning approach. It is both a collaborator and a force multiplier.

ChatGPT’s mastery of machine learning concepts and algorithms enables it to act as an always-available optimization assistant. It provides both strategic advice and automates routine tuning tasks that would be tedious and time-consuming for human engineers.

As shown in Table 7.2, one important application of generative AI lies in its ability to improve the performance and efficiency of machine learning algorithms through collaborative interaction between data analysts and AI. ChatGPT plays a crucial role by acting as a readily available machine learning expert, offering assistance to data analysts in refining their models.

Machine learning optimization approach	ChatGPT’s role
Identify optimization opportunities	Suggest improvements based on model architecture, data and performance
Architecture research	Suggest model architectures and generate code for rapid prototyping
Hyperparameter tuning	Automate simulations to find optimal parameter combinations
Data augmentation	Recommend techniques to extend limited training data
Adaptive learning	Detect data drift and retrain models with new data
Continuous re-training	Detecting drifts and re-training models with new data
Efficiency improvements	Advise on techniques, such as pruning and quantification, to optimize speed and size

Table 7.2. *Boosting machine learning performance with ChatGPT*

They can engage with ChatGPT to describe their model architecture, data characteristics and performance metrics, seeking commands for potential optimizations in terms of accuracy, speed or computational efficiency (Sedkaoui 2023). For example,

ChatGPT can provide suggestions such as adjusting loss functions, applying regularization techniques, incorporating additional layers or attention mechanisms into the model architecture and modifying optimization algorithms according to specific objectives, data characteristics and ongoing model results.

Furthermore, ChatGPT's capabilities extend to automating the hyperparameter tuning process. By describing the search space, constraints, model choices and success criteria, data analysts can instruct ChatGPT to generate and evaluate various hyperparameter combinations to optimize model performance. Rapid simulations run by ChatGPT identify optimal configurations, streamlining the hyperparameter tuning process.

In the area of continuous retraining, as new data become available, ChatGPT serves as a quick resource for analysts, facilitating rapid model retraining and evaluation of results. It helps to determine whether further fine-tuning is required, enabling adaptive adjustments to adapt to changing data patterns. Thanks to its real-time advice, automation capabilities and iterative nature, it empowers data analysts to proactively refine their machine learning algorithms, enhancing their efficiency. Collaborative exchanges with ChatGPT lead to exponential improvements in the optimization of machine learning solutions.

Architecture research is another area where ChatGPT demonstrates its potential. By allowing analysts to describe their problem statements and objectives, ChatGPT can suggest various model architectures, connectivity patterns and layer configurations for experimentation. It goes even further, generating code snippets for the best candidates, speeding up the prototyping and comparison process. This broadens the spectrum of design possibilities that data analysts can effectively explore.

Data augmentation, a crucial technique for enriching model training, also benefits from ChatGPT's involvement. In image data scenarios, ChatGPT suggests augmentations such as rotations, cropping and color changes tailored to the specific problem. For textual data, it recommends strategies such as paraphrasing, back-translating and mixing several samples, thus improving the diversity and robustness of the training dataset.

In the context of adaptive learning, ChatGPT's contributions extend to detecting unfavorable data drifts and facilitating rapid model retraining to adapt to emerging patterns. This continuous learning mechanism guarantees sustained accuracy and relevance of machine learning solutions.

Finally, ChatGPT's ability to suggest efficiency improvements, including techniques such as distillation, pruning, quantization and optimized encodings, adds to its versatility as an optimizer. These recommendations focus on improving model speed, size reduction and optimization of resource use, while preserving the accuracy of performances.

All of this amplifies both productivity and performance – enabling analysts to actively improve model accuracy, robustness and efficiency through iterative collaboration looping human monitoring and AI execution. By combining their strengths, organizations can create optimized machine learning solutions that dynamically learn and adapt to changing real-world conditions.

7.4. Enhancing the potential of machine learning algorithms with ChatGPT

Supervised and unsupervised machine learning present two distinct paradigms in the field of AI, each offering unique advantages and trade-offs. These algorithms have the ability to acquire knowledge from large datasets or real-world observations (Sedkaoui 2018a). Since these datasets come from a variety of sources and in a variety of forms, their analysis can be more or less complex. It is this complexity that gives these algorithms the ability to teach a machine without the need for explicit programming (Samuel 1959).

Machine learning algorithms identify specific patterns that facilitate the learning process. This identification is closely linked to the various phases of data analysis which we have examined in depth in previous chapters. Observed from this angle, this categorization of algorithms can be divided into two main families (Sedkaoui and Khelfaoui 2020): supervised algorithms which discern results and facilitate learning by searching for patterns within a dataset whose characteristics are known, and unsupervised algorithms that process input data devoid of predefined class structures.

A wide range of techniques fall into these two categories and can be mobilized for data analytics. We present the most commonly used algorithms in Table 7.3.

Analysis	Techniques	Learning mode	Problem to be addressed	ChatGPT	
				Role	Example
Simple	Simple/multiple regression	Supervised	Regression	Automate feature engineering	Suggest features to predict customer churn based on these data
	Naive Bayes	Supervised	Classification	Identify missing or erroneous data	Identify data errors in a training set
	Logistic regression	Supervised	Classification	Suggest data transformations	Recommend data transformations to obtain a better model fit
	Linear discriminant analysis	Supervised	Clustering	Generate code for rapid prototyping	Provide a Python code template for this model
	K-means	Unsupervised	Clustering	Interpret and name clusters	Summarize the key traits of each group in the analysis
	Hierarchical classification	Unsupervised	Clustering	Interpret and name clusters	Summarize the main features of each cluster in this analysis
Complex	Decision trees	Supervised	Classification/ regression	Optimize hyperparameters	Optimize model hyperparameters to maximize accuracy
	Random forests	Supervised	Classification/ regression	Generate synthetic data	Generate new training samples for the unbalanced dataset
	Support vector machine (SVM)	Supervised	Classification/ regression	Suggest model architectures and optimize hyperparameters	Propose SVM kernel types for the text classification task, and tune SVM regularization kernel parameters

Analysis	Techniques	Learning mode	Problem to be addressed	ChatGPT	
				Role	Example
Complex	Neural networks	Supervised	Classification/ regression	Explain predictions	Explain the neural network's classification decisions
	K nearest neighbors	Supervised	Classification/ regression	Identify outliers	Detect potential outliers in training data
	Bayesian networks	Supervised	Classification, probabilistic prediction	Analyze dependencies between variables	Assess dependencies between variables in the dataset
	Gaussian mixture models	Unsupervised	Clustering, density estimation	Identify the optimum number of components	Determine the ideal number of mixing components based on the data
	Generative adversarial networks	Supervised, unsupervised	Realistic data generation	Generate new data	Generate synthetic data preserving original patterns
	Auto-encoders	Supervised, unsupervised	Dimensionality reduction, anomaly detection	Identify the most important dimensions/features	Identify abnormal data points in the dataset. Determine which data dimensions are most important

Table 7.3. The various uses of algorithms versus the role of ChatGPT

Supervised machine learning, where models learn from pairs of labeled training data, excels at making predictions and data-driven decisions once enough annotations exist (Rich and Gureckis 2019; Sedkaoui 2023). Classification, regression, prediction and ranking are all powered by supervised algorithms on human-labeled examples. This top-down approach encoded human considerations in models. However, obtaining comprehensive, high-quality training labels can be arduous.

Unsupervised learning, on the other hand, offers a bottom-up approach to discovering the structure and patterns in unlabeled data without human assistance. Clustering, dimensionality reduction, association rule mining and density estimation are key unsupervised techniques. This makes it possible to digest data where annotations are difficult, subjective or sparse. Unsupervised learning provides a lens on blind spots that supervised approaches miss given their dependence on labeling.

The techniques are complementary. Unsupervised learning is often used for exploratory analysis and the feature representation discovery, which can then be used to reinforce supervised models. Semi-supervised approaches combine small sets of labeled datasets with larger unlabeled corpora. And self-supervised learning creates artificial supervision signals from the data as a special case of targeted unsupervised pre-training. Mastery of both ends of the spectrum enables a variety of challenges to be matched with appropriate learning paradigms. Just as humans learn from a mixture of instruction and experience, AI systems must also integrate supervised and unsupervised learning to achieve robust intelligence.

In this context, the question arises as to the extent to which generative AI techniques, such as ChatGPT, could contribute to generating additional synthetic training data in order to improve the performance of these algorithms. Could such data-generating capabilities compensate for the limitation of training datasets and thus enhance results in both supervised and unsupervised learning? These considerations open up exciting prospects for merging the strengths of supervised and unsupervised learning targeted with the means of generative AI.

For their part, supervised learning algorithms rely on large labeled training datasets to learn efficient mappings from inputs to target outputs. However, acquiring sufficient, high-quality training data can be challenging and time-consuming in practice (Sedkaoui 2023). This is where generative AI models like ChatGPT show immense potential for help in augmenting limited human-labeled examples with knowledge-powered pseudo-labeling and feature enrichment strategies and natural language capabilities.

For example, ChatGPT could rapidly extend the limited training data for a sentiment classification model by generating additional representative synthetic samples with pseudo-labels on a diversity of topics and linguistic styles. This

provides broader coverage and protects against model overfitting on narrow labeled examples. ChatGPT can also enhance the quality of training data by interactively flagging and relabeling any mislabeled examples through conversation with human experts, cleaning up errors.

Equally impactful is ChatGPT's ability to enrich the input features introduced in supervised models by extracting higher-order semantic representations from plain text, images or tabular data based on the conversational prompts of machine-learning experts. More informative features improve model generalization and scalability. ChatGPT can also advise on the optimal model architectures and hyperparameters to try out, depending on performance objectives and data characteristics, automating tedious trial-and-error experimentation.

Unsupervised learning algorithms discover hidden structures in unlabeled data, without the aid of human-tagged examples. By revealing patterns, clusters and intrinsic representations, unsupervised techniques enhance predictive modeling and decision-making (Siegel 2013; Sedkaoui and Khelfaoui 2020). However, their exploratory nature poses challenges in interpreting results and aligning discoveries with the business value. This is where intelligent systems like ChatGPT demonstrate strong potential for collaboration by providing interfaces to interactively guide and contextualize unsupervised learning.

For example, ChatGPT can help identify new unlabeled datasets relevant to a business problem through natural language discussions to power unsupervised modeling. This alleviates data constraints. ChatGPT can also provide business context to help interpret the semantic meaning of clusters, dimensions and segments generated by unsupervised algorithms. This bridges the gaps between technical results and strategic objectives. Additionally, by conversing with ChatGPT, data scientists can direct analysis as needs evolve, applying human judgment to unsupervised techniques that traditionally operate blindly. Finally, ChatGPT can quickly test permutations of unsupervised algorithms, parameters and data permutations, accelerating iterative improvement.

Supervised machine learning techniques, such as regression or neural networks, rely on large sets of labeled data to learn. ChatGPT can enhance them by automating the engineering of features, generating synthetic data to compensate for limited datasets and explaining the predictions of complex models. For unsupervised methods, such as k-means clustering, ChatGPT helps to interpret and name the discovered clusters. For advanced techniques, such as Bayesian networks, ChatGPT speeds up the analysis of dependencies between variables. And for auto-encoders, it can identify the most informative data dimensions captured.

As shown in Table 7.3, ChatGPT is able to support a wide range of machine learning techniques, from the simplest to the most complex. Thanks to its conversational capabilities, it can accelerate every stage of the data pipeline from algorithm suggestion to model optimization and rapid prototyping. The table sums up the diversity of possible contributions to empowering practitioners. ChatGPT paves the way for fruitful collaboration between the strengths of AI developers and the knowledge of an AI. This association promises major advances for the future of machine learning.

More generally, ChatGPT's deep technical knowledge enables it to propose model architectures adapted to data and objectives. It automates the optimization of hyperparameters to improve performance. And it generates code to rapidly prototype ideas. Its conversational assistance thus amplifies the capabilities of users and practitioners in machine learning.

Whether for simple or complex algorithms, supervised or unsupervised algorithms, ChatGPT considerably speeds up the process from data to deployment. It ushers in a new era of AI innovation through human-machine collaboration. Data scientists gain a partner in optimizing end-to-end machine learning pipelines.

Now that we have covered the opportunities for optimizing workflows, accelerating innovation and democratizing access (Von Hippel 2006), Chapter 8 focuses on an equally crucial topic: bringing these analytical perspectives to life through intuitive and engaging data storytelling and reporting.

After developing models, organizations need to know how to convey the results convincingly and responsibly to a variety of audiences (Sedkaoui 2018b). Here too, ChatGPT demonstrates strong potential for automating reporting, provided that human supervision guarantees the accuracy, objectivity and strategic relevance of communications. Let us continue our exploration in Chapter 8!

Narrative Fascination: Data-driven Stories and Reports

In-depth data analytics only creates impact when it is communicated in an engaging way to stakeholders. Transforming technical insights into compelling stories is an essential skill in the data age. In this chapter, we explore how intelligent assistants like ChatGPT can help generate stories and reports from data that capture attention and convey complex ideas in a simple way.

We will be looking at applications such as automating the writing of interactive dashboards and developing plans for engaging storytelling. ChatGPT shows great potential for bringing data to life via visuals and natural, engaging language. However, human validation remains essential to guarantee accuracy, avoid bias and align storytelling with business objectives.

The combination of the creativity of AI and human wisdom is paving the way for a new era of high-added-value reporting and presentation. Through this chapter, we will learn how to leverage ChatGPT to transform technical analysis into compelling stories that inspire your audience and promote a data-driven culture in your organization. Let us get started on this exciting exploration!

8.1. ChatGPT for generating data storytelling plans

Data storytelling is an art form that transcends the realm of raw statistics and transforms them into captivating narratives with both informative and engaging qualities (Sedkaoui 2018a). This art form serves as a powerful instrument for conveying complex concepts, bridging the gap between complex data and human understanding.

At the heart of this process is the indispensable concept of data storytelling plans. These plans serve as architectural blueprints, ensuring that data stories are meticulously structured and ready to achieve their objectives. For a data storytelling plan to be effective, it needs to be made up of several integrated components, perfectly interwoven into a coherent whole.

A fundamental element of any data storytelling plan is the establishment of a clear understanding of the underlying data. This fundamental step requires a comprehensive understanding of the data's origins, its inherent constraints and the wealth of information it contains in its digital threads.

At the same time, a well-defined audience serves as a guiding star for creating a compelling data story. By precisely delineating the contours of the target audience, storytellers gain the means of modulating the tone, depth and breadth of the story. The story's resonance lies in its alignment with the audience's prior knowledge, interests and expectations, enabling it to reach its destination with greater precision.

The imperative of a crystalline goal is integrated into a narrative data plan. Every data story must be underpinned by a clear and discernible purpose. Whether the goal is to enlighten, influence or captivate, the trajectory and impact of the narrative depends on the succinct articulation of this objective, guiding every word and visualization towards a coherent destination.

The skeleton of a story is its structure, and the narration of data is no exception. A meticulously traced narrative structure ensures that the data story unfolds in a logical sequence, capturing attention from its inception, creating intrigue as it unfolds and offering a satisfying resolution at its climax. The triadic architecture of beginning, middle and end propels the audience on a sensory journey of exploration, understanding and reflection.

The visualizations, the vibrant brushstrokes on the data canvas, bring these stories to life. Data storytelling thrives when aided by visual aids that transform abstract numbers into tangible information. These graphical presentations not only enhance engagement but also promote understanding, allowing us to perceive patterns and relationships that words alone might struggle to convey.

When the narrative reaches its zenith, the narrative baton is passed to the call to action. A poignant data story culminates in a request for action, calling for new awareness to be translated into tangible steps. This last element catalyzes the transformation of passive engagement into active contribution, whether by visiting a website, signing a petition or making a donation, thus fostering a symbiotic relationship between the story and its recipients.

In the symphony of data storytelling, each element harmonizes with the next to create a composition that resonates far beyond the confines of data tables and charts. Thanks to meticulous planning, a data story evolves from a simple amalgam of facts into an immersive experience that enlightens, empowers and entertains its audience, leaving an indelible mark on their perceptions and actions.

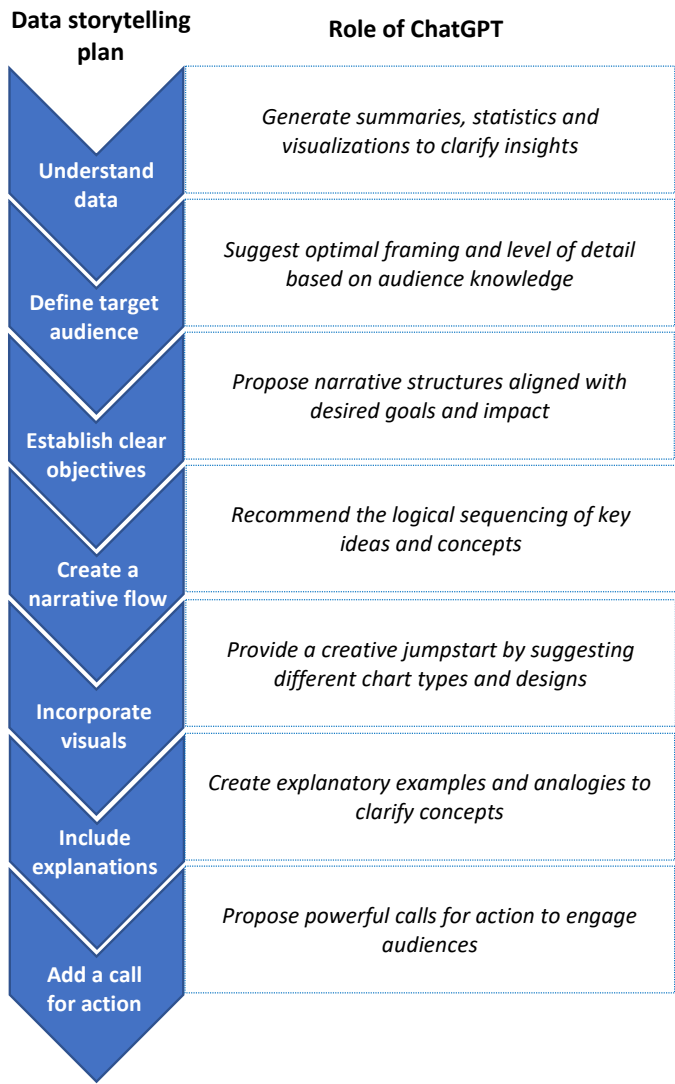


Figure 8.1. ChatGPT’s potential in the narrative data planning process

Creating compelling data stories requires thoughtful planning and structure. ChatGPT can provide valuable assistance throughout this process. But, in reality, how can ChatGPT collaborate with human teams on the essential elements of planning and designing data-driven narratives, to ensure maximum relevance, engagement and impact (Sedkaoui 2023)? Figure 8.1 visualizes the answer to this question and depicts the essential components of developing a data narrative, as well as ChatGPT's conceivable role at each stage.

ChatGPT can collaborate on key planning elements, from understanding datasets to shaping narrative flow and explanations. Its knowledge and generation capabilities provide a solid starting point for teams to then refine and personalize data narratives for maximum relevance and impact.

First and foremost, ChatGPT accelerates data comprehension by automatically generating summaries, statistics and tables via prompts in natural language. This elucidates ideas and relationships.

Then, when defining target audiences, ChatGPT can suggest optimal framing according to audience type – making complex details accessible to non-technical crowds or adding technical rigor to data scientists.

ChatGPT also helps to develop the narrative arc and flow according to the analysis objectives and key takeaways, ensuring a logical sequencing of ideas tailored to the intended audience.

For data visualizations ChatGPT provides creative jumpstart suggestions for trying out different chart types and designs. It also creates explanatory examples and analogies to clarify concepts for different audiences.

Last but not least, ChatGPT is adept at proposing punchy calls to action that mobilize the audience based on the desired messages and outcomes.

At every stage, ChatGPT provides a solid starting point for refining the nuanced understanding of business needs.

Generative AI models like ChatGPT can collaborate with data teams to design these communication plans to ensure report resonance. By describing key aspects, such as your target audience, data insights, analysis objectives and desired narrative arc, ChatGPT can propose communication frameworks, sequencing, explanatory analogies and customized visualizations for potential inclusion. This provides a solid starting point for discussion and refinement with creatives and directives.

For example, data analysts can inform ChatGPT of key trends in customer attrition trends detected by predictive modeling. They can then prompt it to suggest an engaging scenario that conveys the “aha” moments of the analysis into an intuitive flow. This may involve revealing patterns among customer cohorts before diving into churn factors. Data analysts can guide ChatGPT to tailor the narrative to the executive stakeholders and technical parties by describing their interests and level of mastery of the data.

In this context, ChatGPT can offer explanatory analogies and examples that resonate with each audience. Through collaborative brainstorming, teams access ChatGPT’s vast communications knowledge while applying a human understanding of business messaging needs and talents. This creative synergy translates into high-impact data stories that educate and activate audiences.

8.2. The bewitchment of words: automating for writing data-driven stories

Powerful data storytelling requires not only thoughtful planning but also skillful execution in writing convincing stories. This is where generative AI and its ChatGPT intelligent agent are extremely promising for automating the translation of analytical insights into engaging stories.

Data-driven narrative composition involves using data to build a story that combines information and engagement. This technique is highly effective as a tool for conveying complex concepts to a wide range of audiences.

Traditionally, the development of data-driven narratives has been characterized by long and laborious processes. However, the emergence of advanced language models such as ChatGPT has revolutionized this landscape, enabling the automation of data-driven storytelling creation.

ChatGPT offers capabilities covering text generation, language translation, diverse content composition and answering informative queries. In addition, it is on the verge of streamlining the complex effort of writing data-driven stories. In order to automate the process of writing a data-driven story via ChatGPT, we need to provide the AI with the following details:

- the dataset to be used for narrative construction;
- the story’s target audience;
- the overall purpose of the story, encompassing objectives such as informing, persuading or entertaining.

Armed with this input, ChatGPT launches into the task of text generation, producing a narrative interwoven with data insights. This output is meticulously tailored to meet both the target audience and the designated objective of the narrative. In addition, it has the qualities of being enlightening, immersive and comprehensible.

Step	ChatGPT capabilities	Example prompts
Outlining the story	Produce a detailed outline of the narrative and key points, and propose convincing narrative arcs and structures	“Provide a five-paragraph outline for a blog story on key trends from our sales data” “Suggest a compelling narrative arc to explain our A/B test results to senior leadership”
Writing initial draft	Generate complete narratives of draft narratives explaining data insights in easy-to-understand language	“Summarize the key insights from this customer research report in a short blog post aimed at a non-technical audience”
Driving data visuals	Suggest charts, diagrams and tables for visualizing data and findings	“What is the best way to show trends in website traffic over the past year described in these data?”
Refining story elements	Revise, expand or reformat text and visuals based on feedback	“Add more explanatory examples to clarify this section of the draft report”
Writing introductions	Write captivating introductory paragraphs	“Write a captivating introductory paragraph that frames the main ideas of this analysis presentation”
Generating explanations	Incorporate clarifying examples and analogies	“Enrich these data analysis findings with analogies and examples that would resonate with our marketing team”
Adding elements to the story	Include relevant anecdotes and narrative elements	“Include here an anecdote illustrating this data point on customer friction”
Revising tone and style	Adapt tone and style for different audiences	“Modify the tone of this draft to make it more formal for our executive stakeholders”

Table 8.1. *Automated data storytelling with ChatGPT*

By feeding ChatGPT’s key data outputs along with the communication plan and audience needs, it can produce written story drafts, slides, emails and other formats

that convincingly convey critical information in an intuitive flow using vivid language, analogies and examples tailored to the recipients (Sedkaoui 2023).

For example, data journalists could provide ChatGPT with interesting trends discovered from election polling data and have it generate a first draft of an article guiding readers through the analyses in clear, non-technical language, peppered with explanatory metaphors. The automated draft then allows the story to be refined for maximum engagement, rather than starting from scratch.

For visual slides, data teams can prompt ChatGPT to create slide titles, explanatory text and even to suggest custom graphics to bring the data to life in an impactful presentation. This provides a starting point for designers to then perfect the visual flow and branding.

In an era when data weaves complex narratives, the automation of data-driven storytelling via ChatGPT promises a host of compelling benefits.

Automating data-driven stories via ChatGPT offers a myriad of advantages. It not only saves valuable time and effort but also guarantees the accuracy and impartiality of the story. This approach enables the creation of narratives precisely tailored to different audiences, improving engagement and visual appeal in the process.

More than just a compilation of figures, data intrinsically tell stories that have the potential to illuminate complex patterns and insights. With ChatGPT at the helm, this process not only saves invaluable time and effort but also roots the story with unrivalled accuracy and impartiality.

As the data reveals its stories, this approach empowers storytellers to create narratives that resonate with diverse audiences, amplifying both engagement and visual appeal throughout the journey.

Thanks to this synergy, ChatGPT emerges as a powerful creative collaborator, orchestrating the automated creation of data-driven stories that cleverly combine information, captivation and effectiveness, inviting audiences into a realm where stories not only inform but also inspire and compel action.

ChatGPT provides a fast way to get initial ideas and text on the page, allowing human creators to then refine the material for maximum resonance and strategic alignment. By combining forces, powerful data stories can be developed much more quickly. However, human oversight remains essential to guarantee accuracy, avoid bias and write quality stories. The automation of ChatGPT needs to be strategically guided and its work carefully validated before external use.

8.3. Interactive dashboards and ChatGPT's ingenuity

Interactive dashboards are a cornerstone in the field of data visualization and data analytics results, providing a dynamic platform for users to explore data under a multitude of angles, reveal trends and patterns, and thus promote informed decision-making. The synthesis of interactivity, visualization and personalization encapsulated in these dashboards creates a multifaceted approach to exploring and understanding data.

The advent of interactive dashboards heralded a transformation in data analytics, democratizing the field and making it accessible even to those with limited technical prowess (Sedkaoui 2018a). This democratization is of paramount importance, as it extends the reach of data mining beyond the limits of data scientists and analysts, opening doors to a wide range of users, from high-level business leaders seeking strategic information to people in a variety of roles across the organizational spectrum.

Thanks to vividly colored tables, graphs, maps and other visual aids, users are able to discern patterns, anomalies and correlations that might remain hidden in the labyrinthine structure of raw data. The combination of interactivity with visualization enhances this process, allowing users to zoom, filter and manipulate visualizations in real time, offering a granular view of the data landscape.

Customization is a powerful attribute, accentuating the versatility of interactive dashboards. Users have the ability to tailor these dashboards to their unique needs. Whether a company needs to track discrete metrics for performance evaluation, or a data analyst is looking to examine specific data points for insights, the flexibility of shaping the dashboard ensures its resonance with the user's objectives.

Collaboration is elevated to a new level thanks to the inherent sharing functionality in interactive dashboards. Teams can collaborate seamlessly on data-centric projects, transcending geographical barriers and time constraints. The ability to distribute and merge information in real time fosters the convergence of diverse perspectives, essential to supporting sound decision-making in complex and rapidly changing business environments.

Taking advantage of ChatGPT's capabilities, the process of creating a dashboard and its refinement through to its ultimate iteration can be accomplished in a matter of minutes. The remarkable achievement here lies in ChatGPT's determining role in accelerating and simplifying the creation of interactive dashboards designed for data analytics or related activities.

This increase in speed not only accelerates the exploratory data analysis (EDA) phase and facilitates the visualization of hyperparameter tuning results (Sedkaoui 2023), but also streamlines the dissemination of findings. As a result, this dynamic integration of ChatGPT not only enables rapid development but also accelerates the transmission of insights, transforming the landscape of effective communication within minutes.

As ChatGPT's popularity soars, it becomes imperative for analysts to keep abreast of the latest tools and techniques, while exploring ways to seamlessly integrate ChatGPT into their processes. However, as is customary, the creative process is marked by iterations – creation, testing, refinement and further testing, an iterative dance that propels progress. The beginning of this process starts with the construction of a dataset.

Imagine the possibility of creating a dashboard with a simple whisper to ChatGPT. Imagine the scene: through prompts, we can trigger a sequence of events, powered by the ingenuity of plugins like Code Interpreter. Let us take a look at the mechanics of how this transformation happens:

- The process begins with a crucial step: the articulation of the dashboard's objective. What story does the data contain and what insights are we seeking to uncover? The initial dialogue with ChatGPT triggers the map of your intentions.

- From there, a tapestry of data emerges, woven from diverse sources – databases, spreadsheets, APIs – each contributing to the overall story. This assortment converges in the next phase, where data preparation takes center stage. Raw data are cleaned, structured and harmonized to facilitate seamless integration into the narrative dashboard.

- Selecting the dashboard tool is akin to selecting an artist's canvas. The landscape is rich and varied, populated by Tableau, Power BI, Looker, Python, etc. Each tool offers a distinct palette for translating data into vivid visualizations and interactivity, painting a canvas that resonates with your vision.

- Now the story takes shape. Guided by your whispered prompt, the dashboard tool transforms the data into dynamic visualizations. Tinkering and refining, the journey evolves as interactivity weaves itself into the fabric of your creation.

- A crucial step lies ahead: testing the user-friendliness of the dashboard and the clarity of its visual displays. Are the insights obvious and the navigation intuitive? Assurance is sought to ensure that users can easily glean value from the dashboard.

- The unveiling, the grand finale, heralds the unveiling of your creation to the world. The interactive dashboard, a symphony of data and design, is ready to provide others with actionable insights, laying the foundations for informed decisions.

As we delve deeper into the subtleties, plugins appear as invaluable tools, enriching the tapestry of interactivity. Plotly, Dash, Code Interpreter and Streamlit – these plugins paint brushstrokes of versatility and innovation, allowing for the infusion of dynamism and user engagement. Each plugin offers distinct benefits, tailored to your unique story.

The tips shine like guiding stars in this creative galaxy. Clarity in labeling guides users in visual areas, while color coding articulates nuances in the data narrative. Tooltips highlight hidden gems, providing a conduit for deeper understanding. With the mobile era in mind, designing for adaptability ensures that the reach of your dashboard knows no limits. Consistency in design promotes cohesion, while external perspectives refine your creation before it faces the world.

In essence, this journey from a simple prompt to a functional interactive dashboard is an alchemy of creativity, technology and insight. Code Interpreter – a conduit for ideas – works its transformative magic in collaboration with data, visualization and interactivity. With the canvas of possibilities stretched before you, the future of dashboard design is a blend of whispers, innovation and the power to create compelling narratives from the symphony of data.

In the realm of user-centered design, where understanding user needs and goals reigns supreme, selecting appropriate design components and seamlessly integrating ChatGPT serve as a compass for experts aspiring to elevate their dashboard designs to unparalleled levels.

Before diving into the transformative potential of ChatGPT in the dashboard design workflow, it is essential to review conventional approaches. Traditionally, experts would embark on a journey starting with crystallizing the purpose of the dashboard, continuing with the aggregation of the required data and visual components. Subsequently, the dashboard layout would be meticulously tailored to meet user preferences. The final steps would involve manual data entry and labor-intensive data processing to sculpt visual presentations and succinct summaries.

While this proven approach delivers results, it is not without its challenges. Conventional methodology can often be time-consuming and error-prone (Liebreiz et al. 2023). Manual processes can inadvertently hamper the design team's agility, inhibiting the potential for experimentation and iterative evolution. The laborious nature of traditional dashboard design often stifles the creative freedom needed to explore innovative solutions.

In light of these challenges, the prospect of integrating ChatGPT into the dashboard design journey appears to be a beacon of innovation. Infusing ChatGPT's

pro prowess could significantly mitigate the drawbacks of the manual approach, speed up the design process and reduce the risk of errors. By automating certain facets, we gain the freedom to focus on creative ideation and strategic decision-making, ultimately paving the way for iterative experimentation and refinement.

With adapted prompts, this generative agent can suggest and even help generate custom data visualizations such as tables, graphs and maps that intuitively convey key trends and relationships within complex datasets.

Analysts can provide context on targeted business questions and audience needs. ChatGPT can then deliver dashboard mockups with custom visualizations and interactions, such as filters, tooltips and cross-highlighting, to tell tailored data stories. For example, marketers can inform ChatGPT of campaign performance metrics and have it design an interactive dashboard showing trends across channels, creatives, segments and time periods. This gives a practical overview of the analysis.

For frontline teams, ChatGPT could help create a geospatial view of sales data layered with the territories and objectives of employees. Automation provides a solid starting point for then refining production visualizations with optimal functionality, usability and branding.

By combining the visualization capabilities of AI with human monitoring, dashboards move from static reports to interactive narratives that meaningfully engage the audience, making data exploration intuitive and high impact.

8.4. Humans at the heart of protocols: the imprint of human ingenuity in generative AI

In the complex tapestry of data-driven storytelling, an often overlooked aspect is the central role played by humans not only as users but also as creators and architects of the AI systems that underpin them. The profound influence of humans is reflected in the creation, development and use of generative AI models like ChatGPT, positioning them as a testament to human ingenuity and a mirror of human perspectives. The symbiotic partnership between human insight and machine-generated assistance is the basis for creating accurate, informative and ethically sound narratives.

The indispensability of human assistance persists across the entire spectrum, for several reasons:

- *Context and interpretation*: although ChatGPT is remarkably capable, it remains trapped within the limits of its programming. It lacks the ability to grasp the complex nuances that permeate human language and culture. Here, humans emerge

as key architects, endowed with the ability to infuse context and interpretation into data-generated narratives. This human touch elevates the quality of stories and reports by integrating them with a deeper understanding of the subject.

– *Error detection*: ChatGPT’s development trajectory suggests that it is also susceptible to errors. Despite its remarkable evolution, it can sometimes produce inaccuracies in the generated data. Here, the discerning eye of humans shines brightly as they scan the output, skillfully identifying discrepancies and correcting them. This interplay of human vigilance and AI assistance strengthens the foundation of reliability in data narratives.

– *Feedback loop*: the field of data is vast and often convoluted. In such complex domains, the human ability to discern meaning from a set of data points remains unparalleled. By providing comments to ChatGPT, humans focus on the most relevant data, enriching the quality of the stories and reports generated. This collaborative synergy fosters a continuous learning loop, refining the AI’s ability to discern relevance.

– *Interactive dashboards*: the field of interactive dashboards brings to the fore the harmonious blend of human ingenuity and AI capabilities. ChatGPT brings its prowess in generating data-driven insights, while humans step in to design visually appealing, user-friendly dashboards. These dashboards, carefully designed to resonate with users, become the conduits through which data stories are translated into meaningful information.

– *Beyond mechanics*: underpinning the interaction between humans and ChatGPT is a quintessential aspect that signifies the essence of human contribution. The human judge guides the curation of data and shapes the selection of relevant information and the formulation of conclusions. Creativity becomes an inestimable tool, as humans bridge the gap between raw data and compelling narratives, devising innovative ways to illustrate and communicate complex results.

– *At the creative helm*: the journey to generative AI models begins in the minds of humans. These models are, at their root, a reflection of the human intellect that conceived them. From conceptualization to execution, the context of generative AI models is rooted in human thought processes (Samuel 1959; Sedkaoui and Khelfaoui 2020). This thought leadership extends to formulating prompts that initiate interactions with AI. By shaping the prompts, humans imbue the AI with a predefined direction, imbuing the stories with a touch of human intent.

– *Architects of development*: The evolution of generative AI models, like ChatGPT, is a manifestation of human innovation and technological prowess. Humans meticulously build algorithms, train models on massive datasets and iteratively refine their performance. As gatekeepers of the AI development journey, humans refine models, giving them the ability to understand context, generate

coherent text and even imitate human conversations. These models are, in essence, the fruit of human creativity and effort.

– *Co-creators of insight*: the symbiotic interplay between human insight and AI augmentation is what makes data narratives complete and insightful. Using AI as a tool, humans leverage its computing prowess to analyze vast datasets, discern patterns and extract insights that might otherwise escape the unaided human eye. Yet the interpretation, framing and contextualization of these insights remain uniquely human endeavors. The fusion of AI-driven analysis and human-driven interpretation results in narratives enriched by both perspectives.

– *Ethical guard*: as we move into the realm of AI generative models, the ethical compass of humans is an indispensable guiding force. Responsibility for ensuring that AI-generated narratives meet ethical standards rests entirely on human shoulders. The rapid creation process is a living embodiment of this role, where humans decide on the parameters within which the AI operates. This ethical conservation ensures that AI narratives respect the values of inclusivity, fairness and respect. In the age of data-driven decision-making, human ethics play a central role. While ChatGPT is driven by algorithms, humans must ensure that the information extracted from the data is used ethically and responsibly. The moral compass of humans protects against the misuse of data for discriminatory or invasive purposes, helping to create narratives that reflect not just data, but empathy and responsibility.

– *Navigators of future possibilities*: beyond the current horizon, humans guide the trajectory of AI evolution. The creative dynamism of humans drives the refinement of generative AI models, seeking to fill gaps in understanding and improve narrative coherence. As AI models evolve, humans channel their abilities to overcome limitations and unlock new dimensions of creativity and precision.

The relationship between humans and generative AI models is characterized by a profound symbiosis. The genesis of prompts, AI model development and co-creation illustrate the dynamic partnership between human ingenuity and AI computational prowess. While AI models augment and accelerate aspects of narrative creation, the stories themselves remain a harmonious blend of human interpretation, intention and technological augmentation. As we travel this co-evolutionary path, the narrative fabric we weave reflects the harmonious interplay of human and AI capabilities.

While generative AI models such as ChatGPT can accelerate data storytelling and dashboard creation, human surveillance remains essential to validate quality, accuracy and ethics. As competent as ChatGPT is, it lacks true semantic understanding and commercial context. Without human surveillance, risks emerge around bias, misrepresentation and strategic alignment.

Data scientists must thoroughly check automatically generated results before they are released. This involves checking for statistical errors, misleading inferences or any biases that may inadvertently creep in. The examination of a representative sample of dashboard functionality can be used to detect any errors in data flows or visual mappings. Data and domain scientists also ensure that messaging aligns with the organization’s sensitivities and objectives.

In particular, human judgment provides crucial ethical protection against potentially misleading, harmful or unfair presentations. ChatGPT lacks the moral reasoning and responsibility that ethical relationships require. Ongoing human guidance reinforces guardrails and boundaries when AI writes or designs freely.

ChatGPT offers exponential gains in analytical throughput, but thoughtfully directed collaboration unleashes the business impact of the results. ChatGPT’s capabilities in the analysis process, both with and without human intervention, illustrated in Table 8.2, underline the central importance of human supervision.

Task	What ChatGPT cannot do on its own	With human validation and guidance
Data collection	Understand business requirements for data	Determine the relevant data to collect
Data preparation	Perform expert validation to ensure data quality and relevance	Validation of data quality and advanced processing
Exploratory analysis	Provide a strategic interpretation considering operational context	Strategic interpretation and business context
Predictive modeling	Use human discernment to select the optimal algorithm	Choose the best algorithm and model optimization
Storytelling	Identify and detect bias, ensure ethical compliance and accuracy	Validate accuracy, detect bias, strategic alignment
Deployment and monitoring	In-depth testing and ongoing maintenance	Extensive testing, continuous monitoring and maintenance

Table 8.2. AI-based collaborative data analytics to improve ChatGPT

On its own, ChatGPT lacks a true understanding of business objectives and real-world context. It can quickly generate suggestions, find patterns in the data and even prototype models – but the value lies in the human analysts who interpret the insights, select the most appropriate algorithms and align the results with strategic objectives based on their judgment and subject-matter expertise.

Ethical considerations are just as important. ChatGPT may inadvertently introduce biases or inaccuracies without human validators checking for misleading implications and auditing a sample of the AI's work. Subject matter experts provide essential protection against potentially harmful outputs.

Even for seemingly basic tasks, human supervision remains essential to guide ChatGPT responsibly and productively. Here are some additional thoughts:

- Public data collection by ChatGPT requires a human focus on relevant sources and business needs. Without this guidance, data risk being off-topic or useless.
- Data cleaning and preparation by ChatGPT must be validated by experts. Errors or anomalies could persist without human verification.
- The statistics and visualizations produced by ChatGPT in isolation lack strategic context for interpretation. Humans complete the picture.
- The models rapidly prototyped by ChatGPT must be evaluated by experienced data scientists to ensure their robustness and alignment with business objectives.
- Narratives and dashboards generated by ChatGPT require human control of potential biases and strategic relevance in the messaging.

In short, the value of ChatGPT lies in accelerating routine analysis tasks, but not in their autonomous execution. Close collaboration with experts remains essential to maximize benefits while mitigating risks. Together, humans and machines achieve superior quality, relevance and ethics than they can do alone. The future is data analytics combining the best of human and artificial capabilities.

However, realizing the full potential of these capabilities requires the maintenance of rigorous ethics, responsibility and monitoring. Now, in Chapter 9, we discuss the responsible and transparent deployment of generative AI in analysis workflows – by establishing clear safeguards to mitigate risks related to bias, fairness and misuse.

While ChatGPT offers exponential productivity gains, human guidance is essential for reliable integration. We discuss tactics, such as iterative feedback loops, to continuously align ChatGPT results with the needs of the analysis process. And we explore techniques for instilling accountability through transparency in limits and supply.

Power within Hands: Ethics, Orientation and Use

The rapid advancement of generative AI promises to revolutionize data analysis by enhancing productivity, accessibility and insight. However, realizing this potential requires thoughtful governance to maintain ethics, trust and accountability. Unmanaged adoption poses the risk of irresponsible and damaging outcomes.

In this chapter, we explore strategies for the responsible use of intelligent models such as ChatGPT in analytics workflows – including the establishment of human safeguards to guide technological capabilities towards positive goals.

First, we discuss how to understand the limitations of bias and ensuring transparency in when and how to appropriately apply AI assistance. Next, we examine how to align ChatGPT results with business objectives through iterative feedback loops.

We also look at how to ensure accountability and track the steps involved in using AI models. And we explore security protocols and compliance best practices for AI integration.

Finally, we look at how cultural change and education contribute to achieving responsible adoption. The future role of data scientists is to apply wisdom and surveillance to forge a symbiosis between human and artificial intelligence that ethically amplifies their mutual strengths.

By following the principles of a trustworthy AI – justice, equity, responsibility – while pragmatically governing risk, organizations can unlock ChatGPT's collaborative potential for data-driven insight. So let us start exploring this crucial priority.

9.1. Understanding the limits of AI-generated analysis

Throughout this book, we have seen how conversational systems like ChatGPT have profoundly transformed the field of data analytics, offering instant assistance and increased accessibility. However, adopting AI results without adequate monitoring presents significant challenges related to bias, accountability and strategic alignment. To address these concerns, a human governance approach is essential to ensure responsible and effective use of AI-generated information.

Despite remarkable advances in AI technology, it still falls short of human cognitive capabilities. Genuine semantic understanding, creative thinking and ethical reasoning are areas in which AI struggles to match human understanding. Relying solely on AI results can lead to inaccuracies, the perpetuation of biases in training data and misalignment with organizational goals.

A major problem is the presence of statistical biases in training data (Piegorsch 2015; Sedkaoui and Khelfaoui 2020), which can inadvertently lead to biased analyses and narratives produced by generative AI models. In addition, the complex behaviors of AI models can be difficult to understand, masking the risks they might introduce. It is important to recognize that AI is not intrinsically good or bad; its impact is amplified by the way it is applied. Responsible use guided by human supervision can lead to the revelation of ideas that might otherwise remain hidden.

The solution therefore lies in establishing pragmatic control mechanisms that ensure the reliable and ethical use of AI. This approach recognizes both the benefits and risks associated with new technological capabilities. By implementing protocols that emphasize transparency, responsibility and active human involvement in the decision-making process, organizations can harness the benefits of AI while minimizing potential pitfalls.

In this context, analysts play a central role. Their role goes beyond simply applying AI algorithms; it involves assessing context, identifying risks, validating results and guiding AI models like ChatGPT towards realizing their full potential. Let us take a closer look at the different dimensions of their role:

- *Contextual assessment*: human experts bring a deep understanding of the business landscape, societal implications and ethical considerations. They assess the relevance of AI applications in different contexts. For example, if an organization is considering using AI to automate interactions with customer service, a human expert will assess whether certain sensitive requests should be handled by humans to ensure empathy and understanding.

– *Risk identification and management*: human expertise is essential to identify the potential risks associated with AI models. These risks can range from unintentional ones in language generation to the amplification of misinformation. Human experts use their knowledge domain to detect biases that might emerge from the historical data or cultural nuances that the AI model might ignore. They develop strategies to minimize these risks and ensure that AI results comply with ethical and regulatory standards.

– *Reviewing results*: human experts play a crucial role in validating the results generated by AI models. They cross-check the information or recommendations provided by the AI with their own expertise and knowledge. This validation stage is essential to ensure the accuracy and relevance of the results generated by AI. For example, in the context of a medical diagnosis, an AI might suggest a potential diagnosis (Else 2023), but a human expert would carefully review it to ensure that it conforms to medical guidelines.

– *Advice and calibration*: human experts provide advice to AI models to improve their performance over time. They train the models by refining them on relevant data and providing feedback based on their own domain expertise. This iterative process of guidance and calibration helps AI models like ChatGPT to continually improve their responses and understanding of diverse contexts.

– *Adapting to the nuances*: although AI models excel at processing large volumes of data quickly, they can struggle with nuances, humor and cultural references. Human experts refine AI output by introducing these nuances, ensuring that generated content aligns with human communication norms and expectations.

Ethics and human values become integral considerations in both the design and supervision of data analytics augmented by AI.

At their core, AI models like ChatGPT currently lack real-world comprehension, ethical reasoning and creative problem-solving on a par with human intelligence. Their statistical and data-driven nature leaves gaps in contextual understanding, moral judgment, intuition and adaptability.

For example, while ChatGPT can generate long texts on a given topic, it does not really grasp the subtle cultural connotations as domain experts intuitively would. Similarly, ChatGPT has no inherent ability to reason about the ethical implications of generated content that could potentially be biased, harmful or misinterpreted. Human judgment is essential for providing the necessary supervision.

And while ChatGPT can remix training data in a novel way, it lacks the creative problem-solving skills to make truly intuitive leaps or adapt to unprecedented contexts. Its capacity for innovation remains hampered by the data on which it depends, often failing to capture the subtle contextual nuances that humans can easily understand.

Aspect	Human experts		AI models (e.g. ChatGPT)	
	Level	Capabilities	Level	Capabilities
Contextual understanding	High	With strong ethical values, they navigate ethical dilemmas and ensure responsible deployment of AI	Limited	Limited contextual understanding due to data dependency, lack of subtle contextual understanding
Ethical considerations	High	Strong intuition and creativity, generating new ideas and adapting to unique situations	Limited	Limited ethical considerations, risk of generating biased or inappropriate content without ethical awareness
Intuition and creativity	High	Excellent at interpreting the subtle nuances of communication such as humor and emotional undertones	Limited	Limited creativity, mimicking creative thinking through data-driven patterns, absence of genuine authentic human creativity
Interpreting nuances	High	Highly skilled at detecting bias in data, language and narratives	Limited	Difficulty in interpreting nuances due to their data-driven nature, risk of misinterpretation and inaccuracy
Bias detection	High	Embracing continuous learning and skills development over time	Limited	Limited bias detection capabilities, risk of reproducing biases present in training data
Continuous learning	Yes	Embracing continuous learning and skills development over time	Yes	Capable of continuous learning through updates and user feedback, improving response and performance

Table 9.1. Comparison of roles between humans and AI

However, integrating human expertise to address these limitations enables an AI like ChatGPT to excel at narrowly defined tasks with the responsibility of wider human supervision. With human virtues guiding technological progress, society benefits from both human and AI capabilities.

9.2. Ethical harmony: ChatGPT in data analytics workflows

The responsible integration of generative AI, in particular conversational systems like ChatGPT, into organizational data practices and machine learning pipelines, requires a multifaceted approach focused on ethics. This approach encompasses comprehensive planning, transparent deployment, continuous supervision and the active involvement of ethical control. Let us take a closer look at the various aspects of this approach in Table 9.2.

The integration of these four components provides a solid framework for ensuring that ChatGPT is used both ethically and responsibly in data analytics workflows. This approach promotes accountability, minimizes risk and maximizes the benefits of AI technology, while aligning its application with ethical considerations.

The overriding goal is to harness the transformative power of AI for positive change while mitigating the potential for misuse or unforeseen consequences (Sedkaoui 2023). In pursuit of this goal, several key strategies are employed throughout the deployment process to ensure responsible use of AI and alignment with ethical principles:

- *Modeling of ethical threats and pre-mortem analysis*: before deploying tools like ChatGPT, it is crucial to conduct ethical threat modeling. This involves identifying potential vectors of bias, equity issues or other harms based on the intended use cases and data sources. Ethics committees can conduct pre-mortem analyses to anticipate problems and develop mitigation strategies. For example, if ChatGPT is used in data analytics to predict consumer behavior, ethical threat modeling could identify concerns about reinforcing existing biases or making decisions based on unreliable data.

- *Inclusion of domain experts*: the ongoing inclusion of domain experts and data specialists throughout the interaction with ChatGPT is essential. These experts can guide the AI's interactions and validate its outputs before they are disseminated downstream, where potential damage can occur. For example, if ChatGPT is being used to analyze financial data for investment recommendations, the involvement of financial experts can help ensure that the information provided is accurate and in line with industry standards.

Aspect	Specific tasks	Description
Complete planning	Defining use cases	Identifying specific areas where ChatGPT will be used for data analytics, taking into account the organization's needs and objectives
	Risk assessment	Identifying and assessing the risks associated with using ChatGPT, such as bias, confidentiality and negative consequences
	Drawing up ethical guidelines	Developing ethical guidelines and principles to guide the use of ChatGPT, addressing ethical concerns and establishing standards of conduct.
Transparent deployment	Open communication	Proactively communicating to internal and external stakeholders the integration of ChatGPT into data analytics workflows
	Technological information sharing	Providing detailed information on how ChatGPT works, including its capabilities, limitations and data sources used
	Building trust	Building trust by providing transparent details on how ChatGPT will be used and by responding to stakeholders' questions and concerns
Continuous monitoring	Evaluation of results	Regularly evaluating the results generated by ChatGPT in the context of data analytics by comparing predictions with ground truth sources

Aspect	Specific tasks	Description
Continuous monitoring	Identifying problems and biases	Quickly identifying potential problems, such as bias or inconsistent results, using automated detection techniques and manual reviews
	Collecting user feedback	Actively soliciting feedback from ChatGPT users to understand their experiences, identifying areas for improvement and adjusting accordingly
Active ethical supervision	Setting up an ethics committee	Setting up a group of ethical experts to oversee and advise on the use of ChatGPT, reviewing decisions and making recommendations.
	Usage review	Regularly evaluating the use of ChatGPT to ensure that it complies with established ethical guidelines and does not present any ethical risks
	Maintaining liability	Ensuring accountability by continuously monitoring the use of ChatGPT, taking corrective action where necessary and reporting transparently to stakeholders

Table 9.2. Framework for the ethical deployment of ChatGPT in data workflows

– *Transparency and explicability*: transparency and explicability are paramount when the results generated by AI influence decisions or actions affecting individuals. Stakeholders need to understand how outputs are produced and the uncertainties associated with them. This builds trust and ensures that AI-generated recommendations are accepted. For example, if ChatGPT is used to help make data-driven policy decisions, parties need to know how the model arrived at its conclusions in order to assess their reliability.

– *Feedback loops for correction*: after deployment, feedback loops are essential for rapidly correcting errors and enabling learning. In the event of a problem, a feedback mechanism enables prompt adjustments and improvements. For example, if ChatGPT is used to analyze market trends, feedback from analysts and business experts can help refine AI predictions and ensure they remain accurate over time.

– *Human principles as a North Star*: a human-centric approach is crucial. Although AI has transformative capabilities, it is guided by the ethical compass of human principles. This means putting human well-being, equity and responsible use at the forefront of AI development and deployment. For example, if ChatGPT is used for content moderation on a social media platform, the primary objective must be to protect users from harmful content and foster a safe online environment.

To illustrate the integration of ethics, let us consider the example of sentiments provided. When analyzing sentiments on social media using AI models, many challenges require the involvement of human experts. These include interpreting figurative language, taking into account cultural nuances, identifying dangerous content and producing robust sentiment analyses.

For example, if an AI analyzes social media comments, it may struggle to detect sarcasm or irony, leading to inaccurate sentiment analysis. This is where human linguistic experts can help annotate training data to capture these nuances. Similarly, if an AI is tasked with identifying sentiments in local slang or cultural references, human moderators familiar with these aspects can ensure more accurate sentiment classification.

It should therefore be noted that integrating ChatGPT into data workflows introduces a series of ethical concerns that require particular attention. These include: (i) issues of discrimination, (ii) protection of privacy, (iii) transparency, (iv) data quality and (v) the complexity of analysis and user expertise.

Addressing these challenges is crucial to ensuring the ethical use of ChatGPT, preserving the fairness, confidentiality and accuracy of data analytics (Sanderson 2023). These concerns include potential bias and discrimination in analysis results, the need to protect sensitive information, the importance of transparency for informed decision-making, the impact of data quality on reliable information, the suitability of

ChatGPT for different levels of analysis complexity and the importance of user expertise in the responsible interpretation of results. Addressing these aspects is essential to navigating the ethical landscape surrounding ChatGPT's role in data analytics.

In light of these considerations, a holistic approach to the ethical use of ChatGPT in data analytics workflows involves several measures, as shown in Figure 9.1.

By adopting these steps, organizations can responsibly navigate the ethical landscape of integrating ChatGPT into data analytics workflows. This approach ensures that ChatGPT's analytical capabilities are leveraged to make informed and ethical decisions, while avoiding the pitfalls of bias, privacy breaches and misinformation

9.3. Providing iterative feedback to improve ChatGPT

To align ChatGPT's capabilities with business needs and maximize its usefulness, a key priority is to establish continuous feedback channels between human users and the AI system. This enables ChatGPT to continually learn from experience in context.

Helpful comments provide clear examples of inaccurate, biased or misaligned responses to objectives, as well as explanatory advice on ideal results given the trade-offs. Constructive criticism focuses on improving the system rather than venting frustrations.

To measure the impact of feedback, organizations should track performance over time. User panels and red teaming can provide valuable insights and help identify potential problems.

Iterative feedback plays a central role in improving ChatGPT's capabilities, leading to its evolution towards more sophisticated and nuanced interactions (Sedkaoui 2023). The feedback process is multi-faceted, and several key themes emerge as essential elements of this undertaking:

- *Specificity is crucial*: when giving feedback, specificity is a cornerstone. Vague praise or criticism does not offer the granularity needed for improvement. Providing concrete examples that highlight the distinctions between desirable and undesirable responses enables ChatGPT to effectively capture nuanced differences. Thanks to these examples, the AI system can learn to navigate linguistic complexities. For example, by using prompts such as: "This summary is too vague. Please provide more details on key trends and metrics from the sales data", ChatGPT is able to understand specific expectations for precision and detail within data analytics. With this type of feedback, AI can refine its responses by generating more in-depth summaries and avoiding generalities.

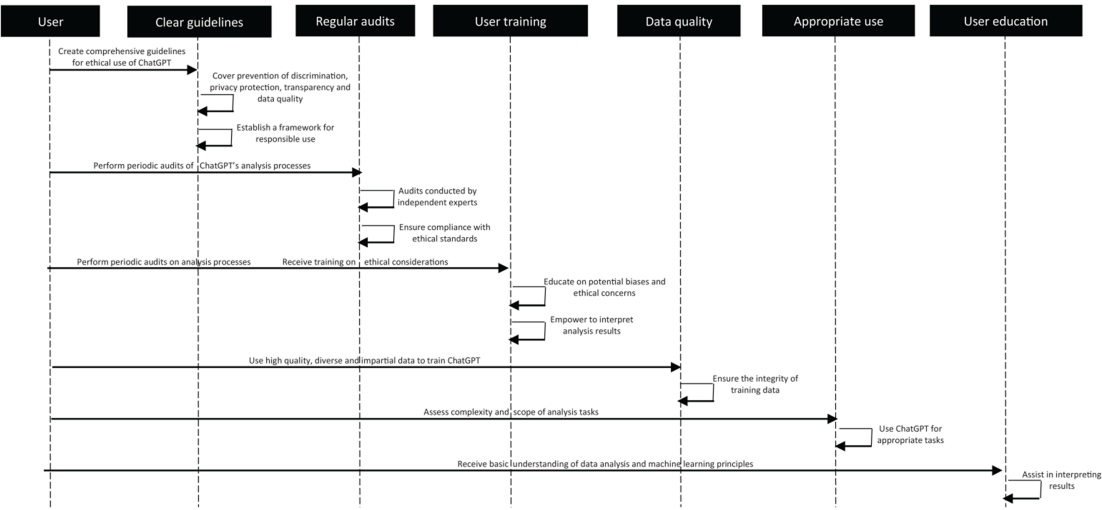


Figure 9.1. Responsible use of ChatGPT in data analytics

– *Constructive advice for progress*: beyond pointing out errors or pitfalls, offering constructive advice steers ChatGPT towards progress. Suggesting potential improvements does not just point out what is wrong, but provides a clear direction for refinement. This allows ChatGPT to learn from its mistakes and make significant progress towards improving the quality of its output. For example, constructive advice such as: “The tone of this section seems too informal. Try using more formal business language for this report”, offers valuable guidance to ChatGPT in adjusting its writing style to suit the business context. By incorporating this type of feedback, AI can gradually develop the ability to adapt to the nuances of tone and language required in business documents, improving its relevance and quality for specific uses.

– *Perseverance and continuous learning*: the feedback loop is not a one-off affair, but requires patience and perseverance. AI systems like ChatGPT are constantly learning and adapting. Continuous feedback over time provides the consistent data points necessary for robust learning. It is essential to recognize that perfection cannot be achieved overnight. Continuous progress, thanks to ongoing collaboration between humans and AI, is the path to improvement. Patience and perseverance in continually providing new examples over weeks and months are crucial to help ChatGPT improve progressively. Even if initial responses do not respond perfectly to expectations, this repeated and constant feedback allows the AI to adapt and adjust its responses accordingly. This approach helps to reinforce ChatGPT’s learning over the long term, evolving its ability to understand and generate more accurate and contextually appropriate responses.

– *Contextual relevance for practical utility*: feedback is most effective when it takes into account the context in which ChatGPT operates. By considering factors like audience and purpose, use case and objectives, enhancements can be tailored to the practical needs of the business. This ensures that AI training is not just focused on general capabilities, but tailored to specific, real-world scenarios, enhancing its practical usefulness. Thanks to feedback from development, deployment and monitoring, ChatGPT evolves with the needs of users, unlocking possibilities that were previously invisible. Considering context is essential for fine-tuning ChatGPT’s performance. For example, by providing hints such as: “This content seems too advanced for our blog audience. Simplify the language for a general audience”, we guide ChatGPT to tailor its responses to the target audience. This optimizes the relevance and comprehensibility of the content generated, ensuring that it is tailored to the specific needs and objectives of the communication.

– *Diversified feedback ecosystem*: it is crucial to enrich feedback perspectives beyond data teams. Diverse viewpoints offer a holistic perspective and help identify potential blind spots and biases that might not be apparent from narrower perspectives. A wider range of inputs promote a more balanced and unbiased AI system. Encouraging diversity in feedback enriches ChatGPT’s training. For

example, by soliciting perspectives from marketing and data science teams to evaluate ChatGPT's dashboard mock-ups and improve the presentation of both points of view, different angles are integrated into the learning process. This holistic approach helps identify potential gaps while ensuring that the system generates responses that reflect a variety of perspectives and avoid biases.

– *Feedback speed and automation*: The speed at which feedback is integrated into the recycling process is a critical factor in accelerating learning. Automation of parts of the feedback loop can amplify the pace at which ChatGPT evolves. This rapid integration of feedback leads to a tighter human-guided learning loop, ultimately improving ChatGPT's performances. By establishing an automated pipeline to transmit new support ticket extracts to ChatGPT on a daily basis, feedback and system learning are accelerated. This automation enables continuous improvement in the accuracy of responses to queries. By taking advantage of this feedback speed, ChatGPT can adjust its capabilities and optimize its results more quickly, thus contributing to more refined performance and smoother interaction with users. The key is to provide concrete illustrations based on real-life examples, as opposed to abstract comments. This helps ChatGPT to capture the tangible improvements expected from human guidance.

– *Collaboration for mutual improvement*: the crux of this iterative feedback lies in collaboration. This is reciprocal improvement between humans and AI. As we guide AI systems towards improvement, we also develop a deeper understanding of their capabilities and limitations. This collaborative approach ensures that humans and AI progress in tandem, harnessing each other's strengths for mutual growth. This human-centered collaboration is essential to evolving supervision. Ideally, feedback pipelines capture both explicit corrections and implicit signals, such as usage data and queries. Feedback is further aggregated, analyzed and used to broaden the diversity of training data (Sedkaoui 2023).

For better alignment, business stakeholders can provide open contributions, not just data teams. Continuous user feedback combined with proactive governance enables ChatGPT to evolve responsibly.

The iterative feedback loop is a dynamic process that fuels ChatGPT's evolution. Thanks to specificity, constructive advice, patience, contextual relevance, diversity of perspectives, automated comments and collaborative principles, we can build AI systems that are not only capable, but tailored to human needs and aspirations.

This collaboration between humans and AI embodies a shared journey towards a future where both entities progress harmoniously. The future is one in which humans and AI are mutually uplifting one another.

9.4. Addressing ethical concerns and biases when using ChatGPT in data analytics

Integrating AI, such as ChatGPT, into data analytics workflows presents both incredible opportunities and profound challenges that need to be addressed with the utmost care. As we leverage the capabilities of AI to improve analytical processes, it becomes essential to establish robust governance mechanisms to mitigate the risks that may arise. These risks involve unforeseen negative outcomes, biases and the potential for misuse when AI operates without appropriate supervision.

While AI systems like ChatGPT can undoubtedly improve analytical throughput and provide valuable insights, they also have the potential to exacerbate existing biases or introduce new distortions that compromise fairness and justice. This highlights the importance of responsible design and active monitoring to ensure that AI-based analyses comply with ethical norms and values.

Human monitoring throughout the AI pipeline is essential. This implies carefully curating unbiased and representative training data, assessing the behavior of AI models, conducting fairness audits and continuously monitoring their results. Subject matter experts play a crucial role in identifying potential damage during risk assessments and in setting rigorous standards to prevent abuse.

Transparency and explicability are the cornerstones of the ethical integration of AI. Clear documentation describing ChatGPT's capabilities, limitations and uncertainties is essential for building trust and accountability. Users and stakeholders need to understand how predictions are made and be able to trace them back to underlying rationale and data.

A crucial aspect of responsible AI integration is the establishment of continuous feedback loops. These loops facilitate rapid correction and learning from mistakes, ensuring that AI systems improve over time. The contribution of a diverse ethics panel representing different communities is invaluable in guiding the evolution of AI practices, so that they are more aligned with diverse perspectives and values.

Developing strategies for the responsible and ethical use of ChatGPT in data analytics is essential to exploit its potential while protecting against unintentional prejudice and bias. Mitigating the risk of bias requires a holistic approach that covers the entire machine learning pipeline.

Starting with data collection, it is crucial to ensure the diversity and inclusiveness of sources, minimizing biased representations that could perpetuate prejudice against under-represented groups. Regular audits throughout model development serve as an early warning system, identifying signs of injustice before

they manifest themselves in real-world applications. In addition, continuous monitoring of live predictions is crucial for detecting and rapidly dealing with emerging discriminatory patterns.

The transparency of models is another key pillar. Highlighting the origins and rationale of ChatGPT results promotes user confidence and understanding. This can be achieved by attaching explanations, confidence estimates and references to training data, thus enhancing the transparency of the model's behavior. Such transparency not only helps users to understand how predictions are made but also facilitates the identification and resolution of undesirable biases.

Recognizing the agency of individuals is essential to the ethical use of AI. Consent mechanisms allow users to control the application of their data by ensuring that they have a say in how their information is used. Participatory design processes, involving the people concerned in the development of model behaviors, ensure that AI systems meet human needs and values. By prioritizing user empowerment over potential exploitation, AI becomes a tool for the improvement of society.

Security and confidentiality safeguards play a crucial role in preventing abuse. Implementing strict access controls, encryption and clear policies for acceptable use cases based on data types prevents sensitive information from falling into the wrong hands. By linking ChatGPT's capabilities to ethical principles and human values, the potential for abuse is significantly reduced.

Developing approaches to ensure the responsible and ethical use of ChatGPT in data analytics is of paramount importance. This process allows us to take advantage of ChatGPT's capabilities while preventing the risks of harm and unintentional bias (Sedkaoui 2023).

Here are some additional approaches to encourage the ethical and responsible use of ChatGPT in the field of data analytics:

- emphasize the importance of setting up ethics committees and governance within organizations to oversee AI initiatives;
- discuss the integration of algorithmic fairness principles in the design of AI systems, such as ensuring correctness, harmlessness and transparency;
- address the importance of ethical training for AI at all levels of the organization to anchor a culture of responsibility;
- highlight the need to continually test AI systems using techniques such as red teaming to identify blind spots;

- emphasize that AI must remain a tool in the service of human objectives, rather than an end in itself;
- discuss approaches to making AI systems more robust and safe, such as reinforcement learning based on ethical rewards.

Ethical deployment of AI systems like ChatGPT rests on the foundation of human values. This requires a deliberate effort to align technological progress with principles such as dignity, justice and autonomy. By infusing human wisdom into the design, development and governance of AI systems, we direct their capabilities towards equitable and beneficial outcomes that uplift society as a whole.

9.5. Ensuring fairness, transparency and accountability in automated data analytics

ChatGPT has attracted widespread attention for its wide range of applications, ranging from idea generation and marketing content development to code debugging and data analytics.

However, amidst the rush towards generative AI, many questions remain, particularly concerning the potential risks and benefits associated with this technology. While ChatGPT's real-world utility has left many impressed, there are deeper and more challenging implications of AI advances that have yet to be fully explored.

For example, an important and unresolved question persists: what are the consequences if ChatGPT empowers people who may not be prepared or inclined to use this technology responsibly?

Although it is not explicitly designed to make programming accessible to the general public, ChatGPT is known to generate code (Bai et al. 2021), including potentially malicious code provoked by nefarious actors on the dark web.

Now, let us turn our attention to machine learning: can ChatGPT be exploited to build a viable machine learning model, and potentially function as a machine learning engineer? Is it capable of making substantial contributions to the future of this field?

Responsible integration of AI-assisted data analytics requires a holistic approach that embraces fairness, transparency and accountability in all of an organization's machine learning practices and pipelines. By adopting proactive governance strategies, organizations can effectively mitigate the potential damage resulting from bias, opacity and misuse, thereby aligning AI's transformative capabilities with ethical principles.

Overall, to guarantee fairness, transparency and accountability in data analytics, several key aspects come into play.

Above all, the involvement of human auditors is essential. These auditors must carefully examine the results of automated analyses for any signs of bias or unfairness before they are released. Rigorous testing processes, including red teaming, must be implemented to identify and correct AI system flaws. Organizations must also place a strong emphasis on diversifying their training data and on the continuous retraining of their models to maintain equity over time, adapting to changing conditions and new challenges.

Radical transparency is another crucial aspect. To establish trust in automated information, full documentation must be provided, describing the capabilities, constraints and uncertainties associated with the analysis results (Sedkaoui and Khelfaoui 2020). Model behaviors must be transparent, explainable and reproducible. In addition, organizations must facilitate information traceability by enabling data mining that feeds automated analyses, ensuring that each conclusion is anchored in accurate and reliable information.

In addition, constructive feedback loops are essential in the post-deployment phase. These feedback mechanisms make it possible to quickly identify and rectify any problems that may arise in the AI system's performance. Establishing a diversified ethics panel can provide multi-dimensional advice, ensuring that the evolution of AI is aligned with inclusive values. Automated reporting can track model performance on fairness measures over time, promoting accountability and progress.

By adopting a holistic approach that integrates foresight and ethically guided reactive controls, organizations can harness the full potential of automated analysis to positively transform their operations. However, it is essential to maintain a solid foundation in human principles to ensure that advances made through AI remain rooted in ethical considerations.

In addition to the strategies described above, ensuring fairness, transparency and accountability in automated data analytics involves several other practices. One key aspect is the use of diverse datasets for training. Ensuring that training data are representative of a variety of viewpoints and perspectives can significantly reduce the likelihood of bias in the analysis results. In addition, human monitoring plays an essential role in identifying and resolving potential biases and problems. Human review and assessment can provide an additional level of control to ensure responsible use of automated data analytics tools.

Transparency concerning the limits of automated data analytics is equally important (Sedkaoui 2018a). Users and stakeholders must be made aware of the potential for bias and misuse inherent in these tools. This awareness enables users to interpret results responsibly and make informed decisions based on an understanding of the overall analysis process.

The integration of these strategies and practices ensures that automated data analytics is carried out fairly, transparently and responsibly. Fairness, transparency and accountability are not only ethical imperatives, they are also essential for building trust in AI systems and preventing discrimination, misinformation and other negative outcomes.

To illustrate these concepts, let us take an example involving an organization that uses an AI-based system to analyze job applications. Without good governance, the system could inadvertently favor certain data demographics or be biased against others. By following comprehensive integration strategies, the organization can ensure that the system is trained on diverse datasets, reviewed by human auditors and transparently documented. This approach minimizes the risk of bias, promotes fairness and enables accountability in the hiring process.

Table 9.3 summarizes the strategies and their corresponding examples.

Strategy	Example
Human auditors	Examination of automated analysis results for signs of bias
Rigorous testing	Red teaming to detect vulnerabilities
Various training data	Ensuring that training data include a variety of viewpoints
Radical transparency	Document capacities, constraints and uncertainties
Information traceability	Enabling users to trace back the analysis to the underlying data
Constructive feedback loops	Integrating user feedback for continuous improvement
Ethics committee	Bringing together diverse perspectives to guide the evolution of AI
Automated reporting	Monitoring model performance on equity measures
Human surveillance	Conducting human reviews to ensure responsible use
Transparency of limits	Making users aware of potential limitations and biases

Table 9.3. *Strategies for fair and ethical automated data analytics*

By adopting these practices, organizations can create AI systems that align with ethical principles, promote transparency and contribute to fair results in data analytics. The symbiosis of AI capability and human wisdom, when guided by ethical considerations, paves the way for responsible and positive transformations in various fields.

To conclude this chapter, let us stress the importance of always placing ethical considerations in the foreground of the development and use of ChatGPT for data analytics. Although promising, this technology must be managed wisely for the benefit of all.

At the end of this book, let us remember that by combining human creativity and the power of AI in a careful and thoughtful manner, the future of data analytics will hold fascinating potential. Let us resolutely pursue this quest for augmented and responsible analysis, where the human mind guides and enhances algorithmic prowess. The time has come to begin this exciting journey, hand in hand with our conscience.

Conclusion

The era of automation is upon us. Various tasks and processes that previously required considerable human effort and time are now augmented or can be fully managed by AI agents. Data analytics is one of these areas experiencing a transformative change thanks to AI. Specifically, generative language models, such as ChatGPT, automate parts of the data analytics workflow that are repetitive, tedious or require exhaustive computational power.

This work highlights the enormous potential of using ChatGPT and similar generative AI to automate data analytics. We are witnessing an era where human data scientists can collaborate with AI assistants to accelerate and improve analysis. When properly developed with care, foresight and ethics, these AI systems can become trusted analytical partners rather than job-stealing adversaries.

Generative models like ChatGPT have certain superhuman capabilities that make them ideal for automating aspects of data analytics when guided by human judgment. Firstly, their insatiable memory and computational speed enable exhaustive analysis of massive datasets beyond human scale. Secondly, their tireless nature offers 24/7 analytical capacity without fatigue. Thirdly, their algorithmic nature ensures consistency in repetitive tasks, unlike bored or distracted humans.

However, these still require extensive training under human supervision to become fully fledged data analytics assistants. Like a child learning from its parents, ChatGPT needs to be nurtured with the appropriate analytical knowledge, tasks, examples and ethics to become a responsible and trustworthy partner. Data scientists need to teach it every step of the way – data cleaning, functionality engineering, modeling, evaluation, explanation, etc.

With extensive experience in diverse datasets, ChatGPT can move from mechanical memorization of instructions to a true understanding of the principles,

trade offs and best practices underlying analytical workflows. This conceptual understanding enables ChatGPT to adaptively select optimal techniques for new datasets and scenarios based on learned reasoning.

Yet, despite such training, ChatGPT still lacks the generalized intelligence that is uniquely human – abstract thinking, adaptability, common sense, judgment, values. It focuses narrowly on excelling at tasks within its training domain rather than a comprehensive understanding of their real-world implications. This is where human creativity and wisdom become so crucial when deploying generative AI.

Humans must supervise and collaborate with ChatGPT to enforce the purpose, ethics and responsibility of its analytical automation. Unchecked, in the wrong hands, such powerful AI tools could be misused in dangerous ways without human conscience as a guide. We must be vigilant parents thoughtfully nurturing the AI child.

The key is to combine the complementary strengths of humans and AI to achieve synergies beyond either. Human creativity, ethics, insight and oversight, along with AI's tireless precision, speed and exhaustive computation, can enable analysis of new scales and depths. But humans must remain an active partner in guiding AI.

For example, the data scientist can creatively define the initial problem, by asking ChatGPT to analyze a dataset for trends, correlations and patterns. ChatGPT quickly crunches the numbers, rigorously testing hypotheses exponentially faster than humans manually can. It generates insights, highlights significant patterns and explains the underlying reasoning.

Humans then assess whether the analysis is complete and the ethical implications are sound. We assess whether biases or limitations have affected the insights, guided by our intuition. Human creativity also gives rise to new angles of analysis not contained in the original prompt. We can then ask ChatGPT to study these new perspectives.

Such symbiotic collaboration, each bringing its strengths, can speed up and considerably improve analysis. But human oversight guarantees that analysis remains thoughtful, ethical and responsible. Neither humans nor AI alone can match their combined potential – a bit like Tony Stark and Iron Man dressing as one.

This demonstrates the crucial balance in automated data analytics. We can enable AI, like ChatGPT, to autonomously take on repetitive and tedious tasks that exceed human capabilities – data cleaning, transformation, model validation. But we must keep humans in the conductive, creative and ethical driver's seat to define problems, set goals, critically interpret ideas and continuously improve AI.

The ideal ratio is the minimum human effort required for monitoring combined with the maximum delegated autonomy appropriate to AI. This maximizes the power of AI automation under conscientious human guidance. With experience and advancement, ChatGPT's analytical skills will continue to improve, enabling greater delegation of responsibility gradually under human supervision.

However, automated analysis is not just about using AI to better understand data. The process of training models like ChatGPT also advances human knowledge and intelligence. Explaining step-by-step tasks to an AI assistant forces us to deeply understand and formalize domains in structured frameworks. The distillation of complex complicated problems into coherent workflows also crystallizes fuzzy human thinking.

Developing AI models grounded in principles, ethics and responsibility requires us to explicitly articulate values, objectivity, biases, hypotheses and best practices. The cleaning and preparation of big data training sets expose special cases and flaws in human logic when contradictions emerge. Teaching an empty vessel like ChatGPT through examples enhances the mastery, judgment and responsibility of the teacher.

So, just as automation is supposed to improve data analytics exponentially through AI, the development of such automation also improves human intelligence, responsibility and formalized knowledge. This requires us to codify the hitherto unstructured analytics profession into principled frameworks, structured workflows and proven best practices.

In conclusion, the automation of data analytics via generative AI promises to usher in a new era, where human and machine intelligences combine for more ethical, creative and accelerated discovery from data. But we are still at the very beginning of this journey, like raising a newborn baby full of potential.

There are certainly risks if such powerful and rapidly evolving AI systems are not thoughtfully developed with human wisdom and ethics firmly under control. There are also calls from profit-driven companies and misguided futurists to automate everything as soon as possible without regard for human well-being. Such unbridled automation could turn AI into an adversary rather than an ally.

But with careful and responsible development, we can create AI assistants that make analysis more accessible, reproducible, shareable and ethical – opening up data insights to a wider audience for the common good. The future is not one in which AI replaces humans; on the contrary, it will increase creativity and ethics with AI's productivity and rigor.

There will undoubtedly be disruption, as traditional analysis tasks are outsourced to AI partners. But history shows that the societal long-term impact of automation is to enable humanity to reach new heights, not to make us obsolete. Just as past breakthroughs, such as engines and computers, have given us superhuman physical and mental capabilities, AI can provide data analytics superpowers.

The sustainable solution is not to resist automation, but to manage its impact responsibly through social adaptation and education. We must continue to advance human creativity, ethics, empathy and wisdom to prosper alongside AI analytical assistants. Parents, educators and leaders have a crucial role to play in preparing society for an automated, AI-infused world by emphasizing our unique human spirit.

And those who build analytical AI have an ethical obligation to make it transparent, safe and consistent with human values. We need to monitor its behavior, minimize harmful biases and prevent misuse. There will be a critical need for AI safety research, certification standards and governance frameworks as adoption accelerates.

This first era of automated data analytics will set important precedents for the future development and integration of intelligent generative AI models. We have a profound responsibility, as well as the opportunity to wisely chart the way forward. By elevating ethics and human knowledge alongside analytical AI, we can enter the era of augmented intelligence – humans and AIs synergizing their complementary strengths for the common good.

The future remains undetermined. Fears of AI antagonism or singularity are unlikely in the foreseeable future. But with careful and ethical development, we can instead create AI collaborators serving under human guidance. Models like ChatGPT are still impressionable children who learn from us. Let us cultivate their productive potential while emphasizing human spiritual values beyond analytics.

If we plant the right seeds today, combining human creativity, ethics and wisdom with the productivity and rigor of AI, we can cultivate a bright future where the symbiosis of both intelligences benefits all humanity. The mission of data scientists and AI developers must be to advance beneficial intelligence – both human and artificial.

Automating analysis with tools like ChatGPT is just the first step in the eternal quest for human progress through creative, ethical, responsible and empowering innovation. We shape our tools; then, our tools shape us. So let us dream big, but proceed carefully. If we walk hand in hand with human consciousness guiding artificial analytical power, the greatest possibilities lie ahead for limitless collaborative discovery!

APPENDICES

Appendix 1

Prompt Repositories for the Analysis Process Using ChatGPT

This repository brings together a series of prompts to guide ChatGPT through the data analytics process. It is divided into several categories, each comprising prompts specific to particular tasks. These prompts have been developed with a view to being as precise as possible, while being adaptable enough to fit various projects. The main objective is to provide users with a starting point for their data analytics projects, without imposing predefined steps. It is important to note that this repository is not intended as a substitute for a full data analytics course or textbook. Nevertheless, it can be an invaluable resource for data analysts wishing to initiate a new project or unblock an ongoing situation.

A1.1. Data collection

A1.1.1. Identifying data sources

- Could you suggest potential data sources relevant to [your analysis objective]?
- I am looking for [datasets] related to [your topic of interest]. Any recommendations?
- Where can I find [reliable data sources] about [your topic]?
- Could you help me identify the [main sources] of data for my [specific analysis project]?
- I am interested in [gathering data] for [your analysis objective]. Any ideas on how to get started?

A1.1.2. Collecting data

- What are some [effective data collection methods] that I can use to obtain [specific data]?
- Could you guide me through [data collection] using [a relevant data collection method] for my [current project]?
- What tools or APIs are suitable for [retrieving data] related to [the extent of your analysis]?
- How should I go about [acquiring data] from [a specific source] for my [type of analysis]?
- I plan to [collect data] for [your objective]. Any advice on best practices and [data collection tools]?

A1.2. Data preparation

A1.2.1. Cleaning data

- What are the [common data quality problems] I need to solve when [cleaning data] for [a specific analysis]?
- Can you give me some advice on [data cleaning] to ensure [data integrity] in [your analysis context]?
- How can I manage [missing values] effectively when [pre-processing data] for [your analysis objective]?
- What steps should I take to [clean and pre-process the data] while working on [a specific analysis]?
- I am in the process of [data cleaning] for [current project]. Any tips for dealing with [data inconsistencies] and [errors]?

A1.2.2. Data transformation

- How can I [convert textual data into numerical values] for [quantitative analysis]?
- Could you show me how to [normalize data values] for [a fair comparison] in [your project]?
- What steps should I take to [group the data] into [meaningful categories], in order to [simplify the analysis]?

- I am confronted with [data in the wrong format]. How should I [transform data formats] to make them compatible with [your analysis model]?
- What are the common methods for [performing mathematical operations on data], in order to [generate new variables] for [my analysis objective]?

A1.2.3. Handling missing variables

- How can I [identify missing values] in [my data] and decide on [the best approach to deal with them] in [your project]?
- Could you guide me through [the different methods for dealing with missing values] and [their potential impact on the analysis]?
- What strategies should I adopt to [impute missing values] [consistently and reliably] in [your data]?
- I noticed [missing values in my dataset]. How should I [decide whether to delete, impute or retain them] for [analysis]?
- What are the advantages and disadvantages of [different imputation techniques for missing values] for [your data] and [your analysis objectives]?

A1.2.4. Outlier management

- What are the steps for [identifying outliers] in [your data] and [determining if they need to be corrected] for [your analysis]?
- Could you show me how to [use statistical methods] to [detect outliers] and [assess their impact on the analysis]?
- How can I [deal with outliers] in an [appropriate] way, in order to [minimize their influence on the results] in [your project]?
- I have spotted [outliers that could affect my results]. How should I [decide whether to correct, delete or retain them] for [your analysis]?
- What are the best practices for [managing outliers] while [maintaining the integrity of your data] and [analysis objectives]?

A1.2.5. Data formatting

- What is the best approach for [organizing data] into [tabular structures], to facilitate [data analytics] for [your objective]?

- Can you give me some advice on [data formatting] for [a smooth analysis] using [formats compatible] with [your analysis tool]?
- How should I [convert data types] appropriately [when preparing data] to [ensure consistency] with [the analysis model]?
- I work with [unstructured data]. What steps should I take to [organize them] into [structured formats] for [more efficient exploration]?
- What are the key steps for [formatting raw data] into [clean, understandable formats] for [the analysis objective]?

A1.3. Coding

A1.3.1. Data handling with coding

- What are the [common data manipulation methods with coding] to [prepare data for analysis]?
- Could you explain to me [how I can filter, sort and group data] using [coding] to [improve the quality of my data]?
- How can I [merge datasets] and [perform transform operations] using [coding] to [facilitate the analysis]?
- What [advanced data cleaning techniques] can I perform with [coding] to [improve the quality of my data]?
- I have disorganized data. How can I [use coding to prepare them] for [analysis] and [ensure consistency]?

A1.3.2. Task automation with coding

- How can I [automate repetitive tasks] in [data analytics] using [coding] to [save time]?
- Could you show me how to [create automated scripts] using [coding] to [perform complex tasks] in [your data]?
- How can [coding be used] to [automate data cleaning, transformation and preparation] in [your analysis workflow]?
- What are the [best practices for automating tasks] with [coding] to [maximize analysis efficiency]?
- I would like to [automate certain parts of my analysis process]. How can I [integrate coding] to [accomplish this] in an [efficient] way?

A1.3.3. Handling missing values with coding

- How can I [use coding to handle missing values] in [my data] [efficiently]?
- Could you explain to me [the techniques for imputing missing values] that I can apply with [coding] to [maintain data integrity]?
- How can [coding help] to [identify missing values] and [decide on the best imputation approaches] in [your data]?
- What are the [common methods for handling missing values with coding] that guarantee [reliable results] in [analysis]?
- I have [missing values in my dataset]. How can I [apply coding] to [process these values efficiently] while [preserving data integrity]?

A1.4. SQL queries

A1.4.1. SQL base

- How can I [get started with SQL] to [efficiently manipulate databases] in [data analytics]?
- Could you explain to me [the fundamentals of SQL] and [how it can be useful] for [querying databases] in [the context of analysis]?
- What are the recommended resources for [learning SQL] to [facilitate my analysis work] and [use data]?
- How can [SQL contribute] to [optimizing my queries] and [extract relevant information] for [my analysis projects]?
- I have little experience of SQL. How can I [start learning] and [apply it effectively] to [data analytics]?

A1.4.2. Data selection and filtering with SQL

- What are the [data selection and filtering methods] I can use with [SQL] to [extract specific records]?
- Could you show me how to [perform selection queries] using [SQL] to [filter data] according to [my criteria]?
- How can I [select specific columns] and [apply filtering conditions] with [SQL] to [optimize data analytics]?

– What [advanced filtering techniques] can I apply with [SQL] to [extract relevant data] in [your analysis projects]?

– I have got voluminous data. How can I [efficiently select and filter] using [SQL] to [work on subsets] in [my analyses]?

A1.4.3. Aggregation and grouping with SQL

– How can I [aggregate data] and [summarize information] using [SQL] to [obtain global statistics]?

– Could you explain to me [how to perform aggregation operations] using [SQL] to [group and summarize data] according to [specific criteria]?

– How can [SQL be used] to [calculate sums, averages and other aggregates] in [your datasets]?

– What [advanced aggregation methods] can I use with [SQL] to [analyze large datasets]?

– I need to [calculate aggregate statistics] for [my analysis projects]. How can I [use SQL] to [get accurate results] for [my needs]?

A1.4.4. Joining tables with SQL

– How can I [join tables] and [combine data] using [SQL] to [get a complete view] in [data analytics]?

– Could you show me how to [perform joins between tables] using [SQL] to [combine information] from [different sources]?

– How can I [perform internal and external joins] with [SQL] to [merge data] according to [specific criteria]?

– What [advanced table join techniques] can I use with [SQL] to [combine complex data] in [your analysis projects]?

– I have [separate data tables] and I want to [merge them]. How can I [use SQL] to [perform efficient joins] and [get integrated data] in [my analyses]?

A1.5. Data processing and analysis

A1.5.1. Exploring data

- How can I [start exploring the data] to [understand the distribution] and [trends] in [my dataset]?
- Could you help me to [explore my data] using [exploratory methods] to [discover patterns] and [anomalies]?
- What are the [essential steps for data mining] that I should follow to [gain insights] in [my analysis projects]?
- How can I [use visualizations] and [descriptive statistics] to [explore my data] [efficiently and informatively]?
- I have complex data. How can I [perform an in-depth exploration] using [different techniques] to [reveal hidden information] in [my analyses]?

A1.5.2. Automation of the analysis process

- How can I [automate the modeling process] by [using scripts] to [speed up iterations]?
- Could you show me how to [set up automation] to [reduce modeling time] while [maintaining quality]?
- What are the [best practices for automating modeling] that will enable me to [optimize resource use]?
- How can I [develop automated workflows] to [quickly experiment with different models] and [identify the best options]?
- I have [several models to develop] and I want to [automate the modeling process]. How can I [use tools and scripts] to [save time and improve efficiency]?

A1.5.3. Coding for data analytics

- How can I [use coding] to [optimize data analytics] in [my current project]?
- Could you help me [integrate coding] into [my analysis] to [automate tasks] and [increase efficiency]?
- What are the [steps for coding] that will help me [process the data] and [create customized solutions]?

- How can I [apply scripts and codes] to [transform and analyze data] using [coding methods]?
- I have [complex analysis tasks] and I want to [use coding]. How can I [adopt effective coding approaches] to [improve my analyses]?

A1.5.4. *Creating advanced functions*

- How can I [develop advanced functionalities] from [my existing data] to [enrich my analysis]?
- Could you help me [create new variables] using [clever combinations] of [existing data] to [better understand the relationships]?
- What [feature creation methods] can I use to [improve the accuracy] of [my predictive models]?
- How can I [extract meaningful information] from [my data] to [create relevant indicators] in [my analyses]?
- I need [specific features] for [my analyses]. How can I [design suitable variables] by [intelligently exploiting the data]?

A1.6. Feature engineering

A1.6.1. *Creating new features*

- How can I [generate new features] from [existing datasets] to [improve model performance]?
- Could you give me some examples of [techniques for creating features] that [can enrich my data] and [add relevant information]?
- What are the key steps in [developing new features] by [using methods such as transformation or combination]?
- I need [additional features] to [improve my model]. How can I [efficiently create new features] by [taking into account the relationship between the data]?
- I want to [enrich my data] by [developing new features]. What are the [common approaches to feature creation] that I might consider?

A1.6.2. *Selecting features*

- How can I [select the most relevant features] for [my model] by [reducing dimensionality]?
- Could you give me some advice on [feature selection] to [optimize model performance] while [avoiding overload]?
- What [feature selection criteria] should I use to [choose the most influential variables]?
- I have [many features] and I want to [select the most important]. How can I [apply selection techniques] to [simplify my model]?
- I want to [refine my model] by [carefully selecting features]. What are the [key steps for selecting features] that [will contribute most to the model's performance]?

A1.6.3. *Putting data in order*

- What is the [optimal method for sorting] and [organizing data] for [my analysis needs]?
- Could you show me how to [order my data] using [specific criteria] to [facilitate analysis]?
- How can I [reorganize and classify] my data in [ascending or descending order] for [better understanding]?
- What are the [steps I need to take to put my data in order] using [efficient methods] to [optimize my analyses]?
- I have disorganized data. How can I [apply efficient sorting] to [organize my data] for [smoother analysis]?

A1.6.4. *Merging data*

- How can I [merge different sources of data] into [a single data source] for [a complete analysis]?
- Could you help me to [combine data] from [several sources] using [fusion techniques] to [consolidate information]?
- What are the [best practices for data fusion] I should follow to [integrate datasets] in [my analysis projects]?
- How can I [apply merge operations] by [using merge keys] to [join data] and [obtain consistent results]?

– I have [separate datasets] that I want to [merge for analysis]. How can I [achieve efficient mergers] using [appropriate methods] in [my analyses]?

A1.7. Modeling

A1.7.1. *Selecting a model*

- How can I [select a suitable model] for [my specific analysis project]?
- Would you be able to help me [choose the best model] based on [my data] and [my goal]?
- What are the [key steps for choosing a model] that will enable me to [make informed decisions] for [my analyses]?
- How can I [evaluate different model options] using [relevant criteria] to [find the best solution]?
- I have [different types of models] in mind. How can I [select the most suitable model] for [my analysis needs]?

A1.7.2. *Training a model*

- How can I [start training a model] using [my training data] to [get accurate results]?
- Could you help me to [set up the training process] by [using the right hyperparameters] to [improve performance]?
- What are the [essential steps for training a model] that will enable me to [create an effective model] for [my analyses]?
- How do I [adjust settings] and [monitor the training process] to [optimize convergence]?
- I have [training data] and I want to [train a model]. How can I [train] using [quality methods] to [get good results]?

A1.7.3. *Evaluating a model*

- How can I [evaluate the performance of a model] by [using appropriate metrics] to [measure its effectiveness]?
- Could you help me [analyze the model's results] by [interpreting the metrics] to [determine its performance]?

- What are the [crucial steps for evaluating a model] that will help me [understand its accuracy] and [its adaptability]?
- How can I [compare model predictions] with [actual values] by [using evaluation techniques] to [validate its reliability]?
- I have [trained a model] and I want to [evaluate its performance]. How can I [analyze its results] using [relevant criteria] for [my analysis projects]?

A1.7.4. Model validation and readjustment

- How can I [validate the performance] of [my analysis models] to [guarantee their reliability]?
- Could you show me how to [assess the quality] of [my models] by [using appropriate metrics] to [select the best model]?
- What are the [steps for model validation] I need to follow to [ensure accurate predictions] in [my analysis projects]?
- How can I [adjust the hyperparameters] of [my models] by [optimizing their performance] to [obtain robust results]?
- I would like to [improve the accuracy of my models]. How can I [validate and adjust my models] by [using cross-validation methods] and [adjustment techniques]?

A1.7.5. Re-training a model

- When is it appropriate to [re-train a model] in relation to [data evolution] or [improved performance]?
- How can I [update an existing model] by [incorporating new data] while [maintaining its accuracy]?
- What are the [key steps for re-training a model] using [current data] to [maintain its relevance]?
- Could you help me understand how to [re-train a model] by [adapting hyperparameters] to [optimize results]?
- I have an existing model that I want to [improve with new data]. How can I [re-train the model] to [optimize its performance] and [integrate new information]?

A1.8. Data visualization

A1.8.1. Suggesting graphics

- Could you help me [choose the appropriate types of graphs] to [effectively represent the data] in [my analysis context]?
- What are the [recommended graphics] for [visualizing relationships] between [different variables] when [exploring data]?
- How can I [select relevant graphics] to [highlight patterns] and [communicate insights] from [my data]?
- I have [various variables to visualize] and I am looking for advice on [the most suitable graphs] to [highlight trends]. Any suggestions?
- What are the [best practices for selecting graphs] that will enable me to [clearly present the data] in [my analysis reports]?

A1.8.2. Identifying trends

- How can I [recognize trends] in [my data] using [effective visualization methods]?
- Could you show me how to [detect trends] from [data graphics] using [key indicators]?
- What types of [trend graphs] should I use to [highlight changes] over time?
- I would like to [highlight trends] in [my data]. What techniques for [trend visualization] do you recommend?
- How can I [use visualizations to quickly identify] [patterns and trends] that could influence [my analysis decisions]?

A1.8.3. Creating visualizations

- How can I [build visualizations] from [my data] using [visualization tools] such as [tool name]?
- Could you guide me through the process of [creating visualizations] using [visualization libraries] to [display results]?
- What are the [basic steps for creating visualizations] that will allow me to [communicate insights effectively]?

– I would like to [create attractive graphics] for [my data]. What are [the key principles] for [making my visualizations clear and informative]?

– How can I [use visualization creation tools] to [transform my data into graphics] and [improve presentation] in [my analysis reports]?

A1.9. Analysis documentation

A1.9.1. Designing a successful analysis report

– How can I [effectively document my analysis] to [ensure traceability of steps] and [facilitate collaboration]?

– Could you show me how to [create clear documentation] to [explain my analysis decisions] and [make the process transparent]?

– What [essential documentation steps] should I follow to [communicate my results] in an [understandable and concise] way?

– How can I [structure my analysis report] to [inform stakeholders] of [steps taken] and [key findings]?

– I need to [document my analysis results] in a [professional and convincing way]. What are the [best practices] to follow in order to [present my findings] in an [impactful] way?

A1.9.2. Making the analysis reproducible

– How can I [guarantee the reproducibility of my analysis] by [recording codes and steps] to [enable others to reproduce my results]?

– Could you show me how to [set up a reproducible environment] to [avoid dependency problems] when [reproducing the analysis]?

– What are the [key steps for making my analysis reproducible], so that [others can check my results] and [add improvements]?

– How can I [organize my analysis project] to [facilitate reproducibility] and [avoid common mistakes]?

– I need help to [make my analysis reproducible]. Could you guide me through [best practices for documenting my code] and [creating a reproducible environment]?

A1.9.3. Presentation of results

- How can I [present my analysis results visually] to [captivate the audience] and [communicate insights effectively]?
- Could you show me some [data visualization techniques] that are [effective in highlighting trends] and [facilitating understanding]?
- What are the [best practices for creating graphs and tables] that [highlight the key points] of my analysis?
- How can I [adapt my presentation to suit the audience] to [disseminate the relevant information] in an [impactful and persuasive] way?
- I would like to [improve the presentation of my results] to [make my analysis more engaging]. Any tips for [generating interest] and [clarifying messages] through data visualization?

A1.10. Forecasting trends and time series

A1.10.1. Predicting trends

- How can I [predict trends] in [temporal data] to [anticipate trends]?
- Could you help me to [use predictive models] to [identify emerging trends] in [my time series]?
- What are the [key steps to forecasting trends] that will allow me to [make projections] and [make informed decisions]?
- How can I [use forecasting methods] to [anticipate future movements] and [identify patterns] in [my temporal data]?
- I have [temporal data] and I want to [predict trends]. How can I [apply forecasting techniques] to [anticipate trends] in [my analyses]?

A1.10.2. Explaining the forecasting approach

- How can I [explain the forecasting approach] that I used for [my analyses]?
- Could you help me to [effectively communicate] the forecasting approach I have adopted for [my time series]?
- What are the [keys to making clear and understandable] the forecasting approach that I used for [my series data]?

- How can I [present my forecasting methods] in an [accessible] way so that I can [share my results] and [interpret them]?
- I have [made some forecasts] and I want to [explain my approach]. How can I [effectively communicate] the methods I used for [my analyses]?

A1.10.3. Time series analysis

- How can I [analyze time series] to [understand patterns] and [detect trends]?
- Could you help me [break down time series] into [key components] to [extract information] from [my data]?
- What are the [essential steps in time series analysis] that will enable me to [highlight variations] and [behaviors]?
- How can I [use time analysis tools] to [identify anomalies] and [highlights] in [my time series]?
- I have [time series] and I want to [analyze trends]. How can I [apply analysis techniques] to [understand patterns] and [interpret results] in [my analyses]?

A1.11. Machine learning

A1.11.1. Algorithm selection

- How can I determine the machine learning algorithm most appropriate for my [classification/regression]?
- Could you provide some examples of machine learning algorithms for [image recognition, sequence prediction, anomaly detection] and explain them?
- For my [unstructured textual data, time series data, computer vision tasks], what specific machine learning algorithms would you recommend considering?
- I am just starting out in machine learning. What are a few polyvalent algorithms I should start with for [classification, regression, clustering] problems?
- Which machine learning algorithm is best suited for [small or large datasets] in terms of [training time, performances, scalability]?

A1.11.2. Model training

- What are the best practices to follow when training models [regression, classification, neural network] on my dataset?

– How can I adjust model hyperparameters, such as [learning rate, layers, dropout], to improve model performance when training [CNNs, RNNs, random forests]?

– What techniques can I use to avoid overfitting and underfitting when training models on my dataset?

– How can I determine the optimum number of eras, batch size and other parameters when training [deep learning, SVM, decision tree] models?

– How can I take advantage of validation sets, cross-validation and other methods to rigorously evaluate models during training?

A1.11.3. Performance evaluation

– What evaluation measures would be most appropriate to assess the performances of my [classification, prediction, recommendation] model?

– How can I evaluate different machine learning models by training them against each other to select the best performer?

– What techniques are recommended for assessing model performance on unbalanced datasets?

– Apart from accuracy, what other key aspects should I analyze to determine how well my model generalizes?

– What validation strategies, such as k-fold cross-validation, should I implement to rigorously evaluate the model's performance?

Appendix 2

GPT-4 versus GPT-3.5: Feature Comparison

Features	GPT-3.5	GPT-4
Parameters	175 billion	Over 1 trillion
Training data period	Until June 2021	Until September 2021
Training duration	Standard training	Additional six months for human and AI feedback
Context window size	3,000 tokens	25,000 tokens
Accuracy	Prone to hallucination	More accurate responses
Safety mechanisms	Limited	Enhanced safety mechanisms
Programming skills	Basic code generation	Can iterate and debug code
Multimodal capabilities	Text only	Also includes images (describes images, contexts, suggestions)
Information synthesis	Struggles to establish connections	Can synthesize from multiple sources
Creativity and consistency	Basic creative writing	Improved narrative and character consistency
Problem solving	Limited in math and science	Complex mathematical and scientific solutions

Features	GPT-3.5	GPT-4
Problem-solving capabilities	Limited complex reasoning	Can solve advanced mathematical and scientific problems
Bias and ethics	More prone to bias	Reduced biased and unethical responses
Responsiveness	Faster responses	Slower for longer responses
Complex task management	Limited capacity	Efficiently manages complex tasks

Table A2.1. *Comparison between GPT-3.5 and GPT-4*

Appendix 3

Basic Terminology Cheat Sheet

Algorithms: step-by-step instructions for solving problems programmatically.

Anomaly detection: identification of unusual patterns different from expected behavior.

API: application programming interfaces enabling access to data and functionalities from an application.

Augmented analytics: fusion of human and artificial intelligence for enhanced data insights.

Automation: the use of technology, such as AI, to perform tasks automatically.

Bias: systematic prejudice that can skew model results.

Business intelligence (BI): processes and tools for collecting and analyzing data to generate business insights and make decisions.

Chatbots: conversational AI systems capable of natural dialogues.

ChatGPT: an advanced language model, created by OpenAI, capable of engaging in interactive conversations and generating coherent, contextually pertinent text in response to prompts.

ChatGPT plugins: third-party integrations that enhance ChatGPT's capabilities, such as code execution, data visualization or database querying.

Classification: categorization of data points into distinct classes or categories.

Code: a set of instructions written in a programming language to give orders to a computer and execute specific tasks.

Code interpreter: tool enabling AI to execute and interact with code snippets in various programming languages.

Coding: process using programming languages to develop software, applications and technological solutions.

Conversational AI: artificial intelligence systems designed to interact and dialogue with humans in a natural, conversational way, often by understanding and generating human language.

Correlation: assessment of the statistical relationship between variables.

Creativity: human ingenuity in formulating problems and imagining solutions.

Dashboard: interface for visual monitoring of key data and key indicators.

Data: facts, statistics or information that can be analyzed to draw conclusions.

Data analyst: a professional who interprets and analyzes data to extract useful information and make informed decisions.

Data analytics: discipline focused on extracting information from data using specialized tools and techniques.

Data cleaning: correction of inconsistencies and errors in the data.

Data collection: gathering relevant data from a variety of sources.

Data pipeline: architectural flow of data from raw state to processed state.

Data preparation: manual manipulation of data to make it usable.

Data quality: assessment of accuracy, completeness and reliability of data.

Data science: interdisciplinary field that uses scientific processes and algorithms to extract knowledge from data.

Data scientist: a data processing expert who applies statistical and computational methods to solve complex problems and gain actionable insights.

Data visualization: graphical representation of information and ideas.

Decision tree: tree-shaped model representing classification rules.

Decision-making: taking action informed by data.

Dimensionality reduction: simplifying multi-dimensional data.

Ethics: moral principles guiding the development and use of AI.

Exploratory data analysis (EDA): initial investigation of data to detect distributions, relationships and patterns.

General intelligence: human-like adaptability in different contexts and tasks.

Generative AI: artificial intelligence systems capable of creating original content, such as text or images, using machine learning models.

GPT-1: first transforming pre-trained generator model, launched by OpenAI in 2018.

GPT-2: larger model, released in 2019, with 1.5 billion parameters.

GPT-3: massive model with 175 billion parameters launched in 2020.

GPT-3.5: enhanced version of GPT-3 with updated architecture.

GPT-4: latest version of ChatGPT, with over 1 trillion parameters and multimodal capabilities.

Human supervision: people reviewing and criticizing AI suggestions.

Human values: qualities such as compassion, creativity and ethics that define humanity.

Intelligent assistants: smart systems, such as ChatGPT, that collaborate with humans.

Interpretability: the ability of humans to understand AI results.

Large language model (LLM): human-like communication, using natural language, processed by AI systems to enable conversational understanding and response.

Logistic regression: classification algorithm for binary results.

Machine learning: algorithms that automatically learn from data.

Natural language processing (NLP): understanding and generating human language with AI.

Neural networks: computer systems modeled on biological brains.

Noteable: plugin transforming AI into a virtual note-taking assistant, capable of capturing, storing and organizing information during conversations.

OpenAI: a renowned artificial intelligence research organization behind major developments such as ChatGPT and other generative AI models, aimed at pushing the boundaries of understanding and creativity machines.

Plugins: additional software components that add new capabilities or functionalities to an existing application.

Predictive analytics: using data to predict future results.

Prompt: instructions (or context) provided to ChatGPT to guide its responses. Carefully crafted prompts allow access to ChatGPT's capabilities.

Prompt engineering: process of creating and selecting questions or instructions to guide the responses generated by AI models, in order to obtain relevant and useful results.

Regression: statistical models quantifying relationships between variables.

Reinforcement learning: agents that learn through trial-and-error interactions.

ShowMe: plugin that enables AI to generate visual representations, such as diagrams and graphs, from text descriptions, to facilitate communication and understanding of complex concepts.

SQL: structured query language used to manage and query relational databases. ChatGPT can understand SQL queries explained conversationally.

Structured data: information organized in a tabular format with clear relationships between elements, facilitating analysis and manipulation via SQL queries or algorithms.

Supervised learning: training algorithms on labeled datasets.

Symbiosis: mutually beneficial collaboration between beings.

Time series: data indexed chronologically over time.

Transparency: openness about how AI systems work.

Unstructured data: information with no predefined format, such as text, images or audio, requiring special processing for analysis.

Unsupervised learning: finding patterns in unlabeled datasets.

Wolfram: a powerful knowledge base providing access to advanced answers, calculations and visualizations related to various academic and scientific fields.

Zapier: automation platform connecting applications to facilitate data flow and automated action execution.

References

- Bai, H., Wang, X., Zhao, L. (2021). Effects of the problem-oriented learning model on middle school students' computational thinking skills in a python course. *Frontiers in Psychology*, 12, 771221.
- Berger, J., Humphreys, A., Ludwig, S., Moe, W.W., Netzer, O., Schweidel, D.A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1), 1–25.
- Berman, J.J. (2013). *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Elsevier, Amsterdam.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer International Publishing, New York.
- Chen, M. and Li, G. (2023). ChatGPT for mechanobiology and medicine: A perspective. *Mechanobiology in Medicine*, 1(1), 100005.
- Chen, H., Chiang, R.H., Storey, V.C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- Cheng, L., Li, X., Bing, L. (2023). Is GPT-4 a good data analyst? *arXiv*, 2305.15038.
- Chui, M., Issler, M., Roberts, R., Yee, L. (2023). Technology trends outlook 2023. Report, McKinsey & Co, London [Online]. Available at: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech> [Accessed July 2023].
- Cupani, M. (2022). Advanced NER with GPT-3 and GPT-J. *Towards Data Science* [Online]. Available at: <https://towardsdatascience.com/advanced-ner-with-gpt-3-and-gpt-j-ce43dc6cdb9c>.
- Davenport, T.H. and Dyché, J. (2013). Big data in big companies. *International Institute for Analytics*, 3, 1–31 [Online]. Available at: <http://www.sas.com/resources/asset/Big-Data-in-BigCompanies.pdf>.
- Davenport, T.H. and Harris, J.G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business School Review Press, Boston, MA.

- Delen, D. and Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, 55(1), 359–363.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(423), 423–423.
- Frankel, F. and Reid, R. (2008). Big data: Distilling meaning from data. *Nature*, 455(30), 30–30.
- Gutierrez, D. (2015). *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R*. Technics Publications, Sedona.
- Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A. (2014). Data quality for data science, predictive analytics and Big Data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economy*, 154, 72–80.
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J. (2021). Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. Report, McKinsey & Co, London [Online]. Available at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world> [Accessed December 2016].
- Hilton, J., Nakano, R., Balaji, S., Schulman, J. (2021). WebGPT: Improving the factual accuracy of language models through web browsing. *OpenAI* [Online]. Available at: <https://openai.com/research/webgpt> [Accessed December 2021].
- IBM Institute for Business Value (2023). Enterprise generative AI: State of the market. *IBM* [Online]. Available at: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/enterprise-generative-ai?utm_medium=Email&utm_source=Adhocoutbound&utm_content=IBVWW&utm_term=17TSL&utm_campaign=Acoustic_23-EL1204-9A9D_A&utm_id=23-EL1204-9A9D_A_1204 [Accessed July 2023].
- Jiang, Z., Xu, F., Araki, J., Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8, 423–438.
- Kim, H., So, B.H., Han, W.S., Lee, H. (2020). Natural language to SQL: Where are we today? *Proceedings of the VLDB Endowment*, 13(10), 1737–1750.
- Liebrezn, M., Schleifer, R., Buadze, A., Bhugra, D., Smith, A. (2023). Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. *Lancet Digit Health*, 5(3), e105–e106.
- Lucini, F. (2021). The real deal about synthetic data. *MIT Sloan Management Review*, 63, 1–4.

- Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*, Baltimore, 55–60.
- Martineau, K. (2022). Five ways IBM is using synthetic data to improve AI models. IBM Research [Online]. Available at: <https://research.ibm.com/blog/synthetic-data-explained>.
- McAfee, A. and Brynjolfsson, E. (2011). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press, Lexington.
- McKinsey & Co (2023). The economic potential of generative AI: The next productivity frontier. Report, McKinsey & Co, London [Online]. Available at: <https://www.mckinsey.com/featured-insights/mckinsey-live/webinars/the-economic-potential-of-generative-ai-the-next-productivity-frontier> [Accessed 14 June 2023].
- Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill, New York.
- Mollick, E. (2022). ChatGPT is a tipping point for AI. *Harvard Business Review*.
- Morabito, V. (2015). *Big Data and Analytics: Strategic and Organizational Impacts*. Springer International Publishing, New York.
- OpenAI (2023). GPT-4 Technical Report [Online]. Available at: <https://cdn.openai.com/papers/gpt-4.pdf>.
- Osoba, O. and Welser, W. (2017). *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. RAND Corporation, Santa Monica.
- Oxford Analytica (2023). GPT-4 underlines mismatch on AI policy and innovation. Report, Oxford Analytica, Oxford/Paris.
- Piegorisch, W.W. (2015). *Statistical Data Analytics*. John Wiley & Sons, New York.
- Radford, A. (2018). Improving language understanding with unsupervised learning. *OpenAI* [Online]. Available at: <https://openai.com/research/language-unsupervised>.
- Rich, A.S. and Gureckis, T.M. (2019). Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, 1, 174–180.
- Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
- Sanderson, K. (2023). GPT-4 is here: What scientists think. *Nature*, 615(7954), 773.
- Schick, T. and Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2339–2352.
- Sedkaoui, S. (2018a). *Data Analytics and Big Data*. ISTE Ltd, London, and John Wiley & Sons, New York.

- Sedkaoui, S. (2018b). *Big Data Analytics for Entrepreneurial Success: Emerging Research and Opportunities*. IGI Global, New York.
- Sedkaoui, S. (2023). *ChatGPT and the Future of Language Technology from Conversational to Intelligent Generative. ChatGPT's Journey to Intelligent Generative Technology*. Independently published.
- Sedkaoui, S. and Gottinger, H.W. (2017). The internet, data analytics and Big Data. In *Internet Economics: Models, Mechanisms and Management*, Gottinger, H.W (ed.). Bentham Science Publishers, Sharjah, 144–166.
- Sedkaoui, S. and Khelifaoui, M. (2020). *Sharing Economy and Big Data Analytics*. ISTE Ltd, London, and John Wiley & Sons, New York.
- Sedkaoui, S., Khelifaoui, M., Keltoum, O. (2022). Sentiment analysis on COVID-19 tweets. In *International Conference on Managing Business Through Web Analytics*, Sedkaoui, S., Khelifaoui, M., Benaichouba, R., Mohammed Belkebir, K. (eds). Springer, Berlin.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y. (2023). HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace. *CoRR abs*, 2303(17580).
- Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley & Sons, New York.
- Surameery, N.M.S. and Shakor, M.Y. (2023). Use ChatGPT to solve programming bugs. *International Journal of Information Technology and Computer Engineering*, 3(1), 17–22.
- Susnjak, T. (2023). Applying BERT and ChatGPT for sentiment analysis of lyme disease in scientific literature. *arXiv*, 2302(06474).
- Taecharungroj, V. (2023). What can ChatGPT do? Analyzing early reactions to the innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, 7(1), 35.
- Tamkin, A., Brundage, M., Clark, J., Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. doi: 10.48550/arXiv.2102.02503.
- Thorp, H.H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313.
- Tingiris, S. and Kinsella, B. (2021). *Exploring GPT-3: An Unofficial First Look at the General-Purpose Language Processing API from OpenAI*. Packt Publishing, Birmingham.
- Toubia, O. and Netzer, O. (2016). Idea generation, creativity, and prototypicality. *Marketing Science*, 36(1), 1–20.
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.
- U.S. Bureau of Labor Statistics (2022). Occupational Outlook Handbook [Online]. Available at: <https://www.bls.gov/ooh/math/data-scientists.htm>.
- Van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947), 224–226.

- Von Hippel, E. (2006). *Democratizing Innovation*. MIT Press, Cambridge.
- Wiles, J. (2023). *Beyond ChatGPT: The Future of Generative AI for Enterprises*. Gartner, Stamford.
- Yilmaz, R. and Yilmaz, F.G.K. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*, 4, 100147.
- Zhang, Z., Han, X, Zhou, H., Ke, P., Gu, Y., Ye, D., Qin, Y., Su, Y., Ji, H., Guan, J. et al. (2021). CPM: A large-scale generative Chinese Pre-trained language model. *AI Open*, 2, 93–99.

Index

A, B

accuracy, 41, 42, 44, 45, 48–50, 53, 64, 70, 85, 93, 97, 98, 101, 102, 104, 125, 127, 133, 139, 140, 145, 151, 184, 188

Alexa, 9

algorithm(s), 140, 183, 189

- debiasing, 98
- machine learning, 183
- supervised, 120
- unsupervised, 120

Amazon, 9, 114

- Alexa, 9

analysis/analytics

- Exploratory Data (EDA), 33, 72, 73, 135
- predictive, 79, 82, 83
- sentiment, 30, 50, 53, 54, 73, 79–81, 83, 87, 88, 90–92, 151
- statistical, 2, 58, 73, 74, 83, 92
- text, 50, 88

analytical tasks, 2, 12, 28, 72, 80, 81, 102

artificial intelligence, *see also* generative AI, 1, 188–190

audio, 99, 191

automation/automating, 1, 2, 8, 12, 13, 30, 31, 36, 38, 40, 45–48, 66, 80, 82, 94, 95, 99, 104, 105, 107, 113, 119,

127, 131, 133, 137, 155, 163–166, 172, 175, 191

for writing, 127, 131, 133

Bard, 9

bias, 7, 13, 14, 39, 91, 93, 97–99, 103, 104, 115, 127, 140, 141, 143–146, 148–151, 154–157, 159, 160, 186

errors, 4, 6, 13, 38, 42, 44–48, 52, 55, 67, 94, 104, 107, 108, 121, 124, 136–138, 140, 141, 153, 156, 170, 181, 188

biased responses, 23, 186

big data datasets, 41, 124, 174

BookCorpus, 15, 22, 23

C, D

chatbots, 8, 9, 21, 23, 25, 26, 30

- virtual assistants, 8, 30

ChatGPT

- generative models, 39, 40, 66, 74, 83, 137–139, 163

- images, 9, 24, 52, 78, 99, 111, 124, 183, 185, 189, 191

clustering, 111, 113, 124

coding, 3, 4, 21, 25, 27, 29, 30, 34, 50, 51, 55, 57, 63, 79, 115–117, 136, 172, 173, 175, 176

JavaScript, 25

- programming, 3, 4, 22, 25, 29, 34, 36,
37, 39, 59, 62, 68, 110, 116, 117,
120, 137, 158, 185, 187, 188
- Python, 3, 25, 36, 37, 42, 77, 107, 116,
121, 135
- R, 42, 77, 107, 173
- collaboration, 2, 5, 10, 11, 21, 29–31, 36,
38, 40, 47, 61, 62, 66, 72, 74, 76, 81,
82, 88, 93, 95, 100, 105, 108, 114, 115,
120, 124, 125, 134, 136, 140, 141, 154,
155, 164, 181
- Common Crawl, 15, 22, 23
- continuous improvement, 14, 89, 92, 93,
100–105, 160
 - feedback loop, *see also* feedback loops,
14, 93, 104, 105, 141, 143, 150,
156, 159
- creativity, 5, 7, 11, 12, 14, 20, 35, 38, 61,
75, 82, 87, 113, 115, 127, 136, 139,
146, 161, 164–166, 189, 190
 - augmented intelligence, 37, 166
 - human, 11, 12, 14, 20, 35, 61, 75, 82,
87, 146, 161, 164, 166
- Dall-E, 9
- data
 - bases, 62, 63, 65, 77, 80, 135, 173,
187, 190
 - sets, 41, 124, 174
 - cleaning, 2, 33, 42, 45, 46, 50, 54, 79,
80, 102, 163, 164, 170, 172
 - feature engineering, 40, 51, 121
 - formatting, 41, 42, 45, 46, 50, 95
 - missing, 41, 45, 48–50, 53, 121
 - preparation, 5, 34, 40–42, 44, 45, 47,
48, 50, 52–55, 73, 80, 92, 96, 109,
115, 135, 141, 172
 - pre-processing, 2, 10, 15, 34, 40,
45–48, 107, 115, 170
 - quality, 6, 36, 44–47, 97, 124, 140,
151, 170
 - quantity, 12, 47
 - science, 6, 11, 17, 33–35, 38, 45, 154
 - scientist, 11, 17, 164

- structured, 59, 94, 99, 105, 135, 190
- training, 85, 91, 119, 123
- transformation, 3, 42, 51, 52, 55
- unstructured, 88, 89, 99, 105, 112, 172,
183, 191
- visualization, 34, 58, 76–80, 111, 182
 - PowerBI, 34
 - table, 4, 9, 12, 26, 28, 32, 34, 36,
46, 49, 54, 81, 89, 92, 96, 98, 99,
112, 118, 122, 132, 135, 140, 146,
149, 160, 186, 188
- debiasing algorithms, 98
- Duplex, 9

E, F, G

- engineering
 - feature, 40, 51, 121
 - prompt, 5, 6, 15–18, 20, 63
- ethics/ethical, 5, 6, 11, 14, 15, 20, 31, 38,
86, 93, 106, 139–141, 143–147,
149–151, 153, 156, 157, 159, 160,
163–166, 186, 189
- accuracy, 41, 42, 44, 45, 48–50, 64, 70,
85, 93, 97, 98, 101, 102, 104, 125,
127, 133, 139, 145, 151, 184, 188
- considerations, 1, 30, 37, 106, 141,
144, 146, 147, 159, 160
- reliability, 4, 13, 14, 29, 39, 41, 42, 44,
45, 48, 52, 94, 97, 98, 101, 138,
150, 179, 188
- fairness, 14, 31, 39, 139, 141, 143, 150,
151, 156–160
- FastText, 51
- generative AI, 1, 2, 4–12, 15, 17, 21, 31,
32, 38–43, 46, 57–61, 64, 65, 72, 73,
76, 78, 88, 92, 105, 108, 109, 115,
117, 118, 123, 130, 131, 137, 139,
141, 143, 147, 158, 163–166, 189,
190
- GloVe, 51
- Google, 9, 114
 - Assistant, 9

Bard, 9
 Duplex, 9
 governance, 13, 15, 38, 40, 81, 86, 99,
 105, 143, 144, 155–158, 160, 166
 GPT, 1, 8, 9, 21–25, 27, 28, 35, 37, 68,
 78, 79, 99, 185, 186, 189
 -1, 21–23, 189
 -2, 23, 189
 -3.5, 23, 24, 79, 99, 185, 189
 -4, 9, 21, 22, 24, 27, 28, 37, 68, 78, 79,
 99, 185, 186, 189
 parameters, 2, 7, 22–24, 33, 101, 112,
 118, 121, 139, 178, 184, 189

H, I, J

human supervision, 14, 125, 139–141,
 163, 165
 surveillance, 1, 7, 11, 14, 31, 38, 39,
 72, 73, 84–87, 89, 92, 93, 95, 97, 98,
 100, 103, 104, 107, 115, 117, 120, 133,
 137, 139, 141, 143, 144, 147, 154–156,
 159, 164, 165
 verification, 38, 46, 97, 141
 innovation, 9, 30, 61, 82, 94, 113, 115,
 125, 136, 138, 147, 166
 interactive dashboard, 3, 26, 30, 31, 59,
 61, 62, 135–137
 JavaScript, 25

K, L, M

k-means, 113, 124
 clustering, 111, 113, 124
 large language model (LLM), 15, 16, 27,
 68, 114, 116
 linear, 84
 discriminant analysis, 121
 machine learning, 1, 7, 8, 10, 11, 13, 17,
 33, 36, 46, 48, 50–52, 55, 58, 101, 102,
 104, 108–120, 123–125, 147, 156, 158,
 183, 184, 189
 hyperparameters, 85, 115, 117–119,
 121, 124, 125, 135, 178, 179, 184

models, 1–4, 7, 8, 11–15, 21–30,
 33–36, 39, 40, 45–47, 51–53,
 57–59, 61, 62, 64, 66, 68, 73, 74,
 76–79, 81–86, 89, 92, 93, 95–99,
 101–120, 123–125, 128, 130, 131,
 133, 134, 137–141, 143–145, 147,
 151, 156, 157, 159, 163–166, 175,
 176, 178–180, 182–184, 187, 189,
 190
 pipeline, 117, 155, 156
 reinforcement, 7, 16, 110, 112–114,
 157, 190
 modeling, 2, 5, 22, 25, 29, 33, 34, 38, 42,
 50, 57, 69, 73, 74, 76, 79, 83, 84, 92,
 93, 95–97, 100, 102, 106, 107, 109,
 115, 117, 118, 124, 131, 150, 163,
 175
 forecasting, 80, 81, 86, 98, 105, 106,
 123, 182–184
 performance, 97, 98, 101, 106, 118,
 124, 137, 177–179
 models
 predictive, 2, 38, 43, 47, 52, 53, 73, 76,
 83, 84, 86, 92, 93, 95, 106–109,
 124, 131, 176, 182
 deployment, 30, 83, 85, 103, 104,
 125, 135, 146, 147, 149, 150,
 154, 157, 159, 164
 supervised, 111, 113, 120, 123–125
 unsupervised, 111, 120, 124

N, P

natural language
 model, 2, 34
 interface, 2, 109, 115, 124, 187
 processing (NLP), 3, 16, 21–23, 29–31,
 35, 45, 50–52, 62, 64, 81, 84
 non-linear, 53, 85
 parameters, 2, 7, 22–24, 33, 101, 112,
 118, 121, 139, 178, 184, 189

performance improvement, 8, 14, 24, 28,
52, 65, 71, 72, 89, 92–94, 100–105,
114, 120, 153–155, 157, 160, 179

plugins, 35–37, 68, 78, 99, 135, 136

Code Interpreter, 188

ShowMe, 36, 190

WebPilot, 36

Zapier, 36, 191

PowerBI, 34

process

analytical, 1, 13, 14, 31, 75, 155

cleaning, 41, 45, 48, 53

deployment, 150

iterative, 16, 74, 98, 101, 103, 104, 145

preparation, 41

prompts, 2, 3, 5, 6, 15–18, 20, 22, 24, 30,
34, 36, 57, 72, 80, 84, 91, 99, 104, 108,
115, 116, 124, 130, 135, 137–139, 153,
169, 187, 190

Python, 3, 25, 36, 37, 42, 77, 107, 116,
121, 135

Q, R

query

natural language, 41, 55, 62–64,
68, 70, 72

optimization, 68, 71

SQL, 55, 57, 62, 65–68, 70, 71, 190

structured, 62

R, 42, 77, 107, 173

regression, 33, 53, 83, 84, 113, 123, 124,
183

linear, 83

logistic, 84

reliability, 4, 13, 14, 29, 39, 41, 42, 44,
45, 48, 52, 94, 97, 98, 101, 138, 150,
179, 188

responsibility, 97, 139–141, 143, 144,
147, 149, 156–160, 164–166

fairness, 14, 31, 39, 139, 141, 143, 150,
151, 156–160

governance, 13, 15, 38, 40, 81, 86, 93,
99, 105, 143, 144, 155–158, 160,
166

transparency, 7, 14, 31, 93, 98, 100,
141, 143, 144, 150, 151, 156,
158–160

S, T

ShowMe, 36, 190

software, 29

applications, 8, 21, 23–25, 29, 30, 36,
48, 73, 78, 85, 87–89, 107, 109,
110, 114, 115, 117, 127, 144, 156,
158, 187, 188, 191

development, 29

SQL, 25, 34, 37, 61–63, 65–68, 70–72,
173, 174, 190

requests, 55, 57, 62, 65–68, 70, 71, 190

storytelling, 127, 128, 131–133, 135, 136
dashboard, 28, 134–137, 140, 154
data, 127

plans, 26, 32, 127, 128, 130

strategic prompting, 20

structure, 4, 17, 25, 29, 67–69, 74, 76,
116, 123, 128, 130, 134
narrative, 128

synergy, 7, 10, 15, 59, 75, 78, 82, 94,
110, 111, 131, 133, 138, 166

table, 25, 28, 32, 43, 44, 46, 50, 67, 88,
98, 111, 112, 115, 118, 120, 125,
134–136, 140, 147

technology, 4, 6–8, 20, 21, 24, 26, 38, 46,
55, 58, 64, 66, 76, 78, 136, 144, 147,
158, 161, 187

advanced NLP, 46

AI, 78, 144, 147

text, 7, 8, 15, 22–26, 41, 44, 48, 50, 52,
62, 72, 73, 78, 88–90, 92, 93, 99, 121,
131–133, 139, 185, 187, 189, 191

time series, 81, 105, 182, 183
analysis, 106, 183

transformer, 58, 63, 110, 112, 113, 163

transparency, 7, 14, 31, 93, 98, 100, 141,
143, 144, 150, 151, 156–160
trends, 30, 42, 51, 57–64, 73–78, 82,
86–88, 91, 96, 105–108, 113, 132–134,
137, 150, 153, 164, 175, 180, 182, 183

V, W, Z

value, 2, 5, 7, 8, 10, 11, 13, 32, 38, 68,
80, 82, 89, 93, 94, 100, 102, 104, 105,
124, 127, 135, 140, 141
added, 13, 32, 82, 102, 127
variables
asymmetrical, 84
categorical, 55
independent, 84
vector, 50
virtual assistants, 8, 30

web, 36, 106, 114, 128, 158
pages, 7, 15, 22
WebPilot, 36
WebText, 23
Wikipedia, 15, 23
Word2Vec, 51
workflow(s), 3, 5–7, 16, 18, 20, 21, 27,
32–38, 59, 75, 94–96, 99, 103, 104,
107, 109, 114, 115, 117, 125, 136, 141,
143, 147–149, 151, 153, 155, 163–165,
175
analytical tasks, 2, 12, 28, 72, 80, 81,
102
data analytics, 37, 94–96, 147, 151,
153, 155
Zapier, 36, 191

Other titles from



in

Information Systems, Web and Pervasive Computing

2024

ALBAN Daniel, EYNAUD Philippe, RICHET Jean-Loup, VITARI Claudio
Information Systems Management: Governance, Urbanization and Alignment – 2nd edition

PINET François, BATTON-HUBERT Mireille, DESJARDIN Eric
Geographical Data Imperfection 2: Use Cases

2023

POMEROL Jean-Charles
Action in Uncertainty: Expertise, Decision and Crisis Management

REVEST Valérie, LIOTARD Isabelle
Digital Transformation and Public Policies: Current Issues

2022

ACCART Jean-Philippe
Library Transformation Strategies

BOADA Martí, LAZARO Antonio, GIRBAU David, VILLARINO Ramón
Battery-less NFC Sensors for the Internet of Things

BRÉZILLON Patrick, TURNER Roy M.
Modeling and Use of Context in Action

CHAMOUX Jean-Pierre
The Digital Era 3: Customs and Practices

KARAM Elie
General Contractor Business Model for Smart Cities: Fundamentals and Techniques

2021

BEN REBAH Hassen, BOUKTHIR Hafedh, CHÉDEBOIS Antoine
Website Design and Development with HTML5 and CSS3

EL ASSAD Safwan, BARBA Dominique
Digital Communications 1: Fundamentals and Techniques
Digital Communications 2: Directed and Practical Work

GAUDIN Thierry, MAUREL Marie-Christine, POMEROL Jean-Charles
Chance, Calculation and Life

LAURENT Sébastien-Yves
Conflicts, Crimes and Regulations in Cyberspace
(Cybersecurity Set – Volume 2)

LE DEUFF Olivier
Hyperdocumentation
(Intellectual Technologies Set – Volume 9)

PÉLISSIER Maud
Cultural Commons in the Digital Ecosystem
(Intellectual Technologies Set – Volume 8)

2020

CLIQUET Gérard, with the collaboration of BARAY Jérôme
Location-Based Marketing: Geomarketing and Geolocation

DE FRÉMINVILLE Marie
Cybersecurity and Decision Makers: Data Security and Digital Trust

GEORGE Éric

Digitalization of Society and Socio-political Issues 2: Digital, Information and Research

HELALI Saida

Systems and Network Infrastructure Integration

LOISEAU Hugo, VENTRE Daniel, ADEN Hartmut

Cybersecurity in Humanities and Social Sciences: A Research Methods Approach (Cybersecurity Set – Volume 1)

SEDKAOUI Soraya, KHELFAOUI Mounia

Sharing Economy and Big Data Analytics

SCHMITT Églantine

*Big Data: An Art of Decision Making
(Intellectual Technologies Set – Volume 7)*

2019

ALBAN Daniel, EYNAUD Philippe, MALAURENT Julien, RICHET Jean-Loup, VITARI Claudio

Information Systems Management: Governance, Urbanization and Alignment

AUGEY Dominique, with the collaboration of ALCARAZ Marina

Digital Information Ecosystems: Smart Press

BATTON-HUBERT Mireille, DESJARDIN Eric, PINET François

Geographic Data Imperfection 1: From Theory to Applications

BRIQUET-DUHAZÉ Sophie, TURCOTTE Catherine

From Reading-Writing Research to Practice

BROCHARD Luigi, KAMATH Vinod, CORBALAN Julita, HOLLAND Scott,

MITTELBACH Walter, OTT Michael

Energy-Efficient Computing and Data Centers

CHAMOUX Jean-Pierre

The Digital Era 2: Political Economy Revisited

COCHARD Gérard-Michel

Introduction to Stochastic Processes and Simulation

DUONG Véronique

SEO Management: Methods and Techniques to Achieve Success

GAUCHEREL Cédric, GUYON Pierre-Henri, DESSALLES Jean-Louis

Information, The Hidden Side of Life

GEORGE Éric

Digitalization of Society and Socio-political Issues 1: Digital, Communication and Culture

GHLALA Riadh

Analytic SQL in SQL Server 2014/2016

JANIER Mathilde, SAINT-DIZIER Patrick

Argument Mining: Linguistic Foundations

SOURIS Marc

Epidemiology and Geography: Principles, Methods and Tools of Spatial Analysis

TOUNSI Wiem

Cyber-Vigilance and Digital Trust: Cyber Security in the Era of Cloud Computing and IoT

2018

ARDUIN Pierre-Emmanuel

Insider Threats

(Advances in Information Systems Set – Volume 10)

CARMÈS Maryse

Digital Organizations Manufacturing: Scripts, Performativity and Semiopolitics

(Intellectual Technologies Set – Volume 5)

CARRÉ Dominique, VIDAL Geneviève

Hyperconnectivity: Economical, Social and Environmental Challenges
(Computing and Connected Society Set – Volume 3)

CHAMOIX Jean-Pierre

The Digital Era 1: Big Data Stakes

DOUAY Nicolas

Urban Planning in the Digital Age
(*Intellectual Technologies Set – Volume 6*)

FABRE Renaud, BENSOUSSAN Alain

The Digital Factory for Knowledge: Production and Validation of Scientific Results

GAUDIN Thierry, LACROIX Dominique, MAUREL Marie-Christine,
POMEROL Jean-Charles

Life Sciences, Information Sciences

GAYARD Laurent

Darknet: Geopolitics and Uses
(*Computing and Connected Society Set – Volume 2*)

IAFRATE Fernando

Artificial Intelligence and Big Data: The Birth of a New Intelligence
(*Advances in Information Systems Set – Volume 8*)

LE DEUFF Olivier

Digital Humanities: History and Development
(*Intellectual Technologies Set – Volume 4*)

MANDRAN Nadine

Traceable Human Experiment Design Research: Theoretical Model and Practical Guide
(*Advances in Information Systems Set – Volume 9*)

PIVERT Olivier

NoSQL Data Models: Trends and Challenges

ROCHET Claude

Smart Cities: Reality or Fiction

SALEH Imad, AMMI, Mehdi, SZONIECKY Samuel

Challenges of the Internet of Things: Technology, Use, Ethics
(*Digital Tools and Uses Set – Volume 7*)

SAUVAGNARGUES Sophie

Decision-making in Crisis Situations: Research and Innovation for Optimal Training

SEDKAOUI Soraya

Data Analytics and Big Data

SZONIECKY Samuel

Ecosystems Knowledge: Modeling and Analysis Method for Information and Communication

(Digital Tools and Uses Set – Volume 6)

2017

BOUHAÏ Nasreddine, SALEH Imad

Internet of Things: Evolutions and Innovations

(Digital Tools and Uses Set – Volume 4)

DUONG Véronique

Baidu SEO: Challenges and Intricacies of Marketing in China

LESAS Anne-Marie, MIRANDA Serge

The Art and Science of NFC Programming

(Intellectual Technologies Set – Volume 3)

LIEM André

Prospective Ergonomics

(Human-Machine Interaction Set – Volume 4)

MARSAULT Xavier

Eco-generative Design for Early Stages of Architecture

(Architecture and Computer Science Set – Volume 1)

REYES-GARCIA Everardo

The Image-Interface: Graphical Supports for Visual Information

(Digital Tools and Uses Set – Volume 3)

REYES-GARCIA Everardo, BOUHAÏ Nasreddine

Designing Interactive Hypermedia Systems

(Digital Tools and Uses Set – Volume 2)

SAÏD Karim, BAHRI KORBI Fadia

Asymmetric Alliances and Information Systems: Issues and Prospects
(*Advances in Information Systems Set – Volume 7*)

SZONIECKY Samuel, BOUHAÏ Nasreddine

Collective Intelligence and Digital Archives: Towards Knowledge Ecosystems
(*Digital Tools and Uses Set – Volume 1*)

2016

BEN CHOUIKHA Mona

Organizational Design for Knowledge Management

BERTOLO David

Interactions on Digital Tablets in the Context of 3D Geometry Learning
(*Human-Machine Interaction Set – Volume 2*)

BOUVARD Patricia, SUZANNE Hervé

Collective Intelligence Development in Business

EL FALLAH SEGHRUCHNI Amal, ISHIKAWA Fuyuki, HÉRAULT Laurent,
TOKUDA Hideyuki

Enablers for Smart Cities

FABRE Renaud, in collaboration with MESSERSCHMIDT-MARIET Quentin,
HOLVOET Margot

New Challenges for Knowledge

GAUDIELLO Ilaria, ZIBETTI Elisabetta

Learning Robotics, with Robotics, by Robotics
(*Human-Machine Interaction Set – Volume 3*)

HENROTIN Joseph

The Art of War in the Network Age
(*Intellectual Technologies Set – Volume 1*)

KITAJIMA Munéo

Memory and Action Selection in Human–Machine Interaction
(*Human–Machine Interaction Set – Volume 1*)

LAGRAÑA Fernando

E-mail and Behavioral Changes: Uses and Misuses of Electronic Communications

LEIGNEL Jean-Louis, UNGARO Thierry, STAAR Adrien

Digital Transformation

(Advances in Information Systems Set – Volume 6)

NOYER Jean-Max

Transformation of Collective Intelligences

(Intellectual Technologies Set – Volume 2)

VENTRE Daniel

Information Warfare – 2nd edition

VITALIS André

The Uncertain Digital Revolution

(Computing and Connected Society Set – Volume 1)

2015

ARDUIN Pierre-Emmanuel, GRUNDSTEIN Michel, ROSENTHAL-SABROUX
Camille

Information and Knowledge System

(Advances in Information Systems Set – Volume 2)

BÉRANGER Jérôme

Medical Information Systems Ethics

BRONNER Gérald

Belief and Misbelief Asymmetry on the Internet

IAFRATE Fernando

From Big Data to Smart Data

(Advances in Information Systems Set – Volume 1)

KRICHEN Saoussen, BEN JOUIDA Sihem

Supply Chain Management and its Applications in Computer Science

NEGRE Elsa

Information and Recommender Systems

(Advances in Information Systems Set – Volume 4)

POMEROL Jean-Charles, EPELBOIN Yves, THOURY Claire

MOOCs

SALLES Maryse

Decision-Making and the Information System

(Advances in Information Systems Set – Volume 3)

SAMARA Tarek

ERP and Information Systems: Integration or Disintegration

(Advances in Information Systems Set – Volume 5)

2014

DINET Jérôme

Information Retrieval in Digital Environments

HÉNO Raphaële, CHANDELIER Laure

3D Modeling of Buildings: Outstanding Sites

KEMBELLEC Gérald, CHARTRON Ghislaine, SALEH Imad

Recommender Systems

MATHIAN Hélène, SANDERS Lena

Spatio-temporal Approaches: Geographic Objects and Change Process

PLANTIN Jean-Christophe

Participatory Mapping

VENTRE Daniel

Chinese Cybersecurity and Defense

2013

BERNIK Igor

Cybercrime and Cyberwarfare

CAPET Philippe, DELAVALLADE Thomas

Information Evaluation

LEBRATY Jean-Fabrice, LOBRE-LEBRATY Katia

Crowdsourcing: One Step Beyond

SALLABERRY Christian

Geographical Information Retrieval in Textual Corpora

2012

BUCHER Bénédicte, LE BER Florence

Innovative Software Development in GIS

GAUSSIER Eric, YVON François

Textual Information Access

STOCKINGER Peter

Audiovisual Archives: Digital Text and Discourse Analysis

VENTRE Daniel

Cyber Conflict

2011

BANOS Arnaud, THÉVENIN Thomas

Geographical Information and Urban Transport Systems

DAUPHINÉ André

Fractal Geography

LEMBERGER Pirmin, MOREL Mederic

Managing Complexity of Information Systems

STOCKINGER Peter

Introduction to Audiovisual Archives

STOCKINGER Peter

Digital Audiovisual Archives

VENTRE Daniel

Cyberwar and Information Warfare

2010

BONNET Pierre

Enterprise Data Governance

BRUNET Roger

Sustainable Geography

CARREGA Pierre

Geographical Information and Climatology

CAUVIN Colette, ESCOBAR Francisco, SERRADJ Aziz

Thematic Cartography – 3-volume series

Thematic Cartography and Transformations – Volume 1

Cartography and the Impact of the Quantitative Revolution – Volume 2

New Approaches in Thematic Cartography – Volume 3

LANGLOIS Patrice

Simulation of Complex Systems in GIS

MATHIS Philippe

Graphs and Networks – 2nd edition

THERIAULT Marius, DES ROSIERS François

Modeling Urban Dynamics

2009

BONNET Pierre, DETAVERNIER Jean-Michel, VAUQUIER Dominique

Sustainable IT Architecture: the Progressive Way of Overhauling

Information Systems with SOA

PAPY Fabrice

Information Science

RIVARD François, ABOU HARB Georges, MERET Philippe

The Transverse Information System

ROCHE Stéphane, CARON Claude

Organizational Facets of GIS

2008

BRUGNOT Gérard

Spatial Management of Risks

FINKE Gerd

Operations Research and Networks

GUERMOND Yves

Modeling Process in Geography

KANEVSKI Michael

Advanced Mapping of Environmental Data

MANOUVRIER Bernard, LAURENT Ménard

Application Integration: EAI, B2B, BPM and SOA

PAPY Fabrice

Digital Libraries

2007

DOBESCH Hartwig, DUMOLARD Pierre, DYRAS Izabela

Spatial Interpolation for Climate Data

SANDERS Lena

Models in Spatial Analysis

2006

CLIQUET Gérard

Geomarketing

CORNIOU Jean-Pierre

Looking Back and Going Forward in IT

DEVILLERS Rodolphe, JEANSOULIN Robert

Fundamentals of Spatial Data Quality

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.