

Data Science in Context:

Foundations, Challenges, Opportunities

Alfred Z. Spector, Peter Norvig, Chris Wiggins, Jeannette M. Wing
Pre-Publication Authors' Draft

October 2022, V.M1 (Author's Manuscript)

This material has been published by the Cambridge University Press in October 2022 as
Data Science in Context: Foundations, Challenges, Opportunities.

<http://www.cambridge.org/9781009272209>

This pre-publication draft is free to view and download for personal use only. **Not for redistribution, re-sale, or use in derivative works.**

© Spector, Norvig, Wiggins, Wing 2023

Introduction

The quantity of data that is collected, processed, and employed has exploded in this millennium. Many organizations now collect more data in a month than the total stored in the Library of Congress. With the goal of gaining insight and drawing conclusions from this vast sea of information, data science has fueled many of the vast benefits brought by the internet and provided the business models that pay many of its costs.

Beyond the marvels of present data science applications, there are even greater breakthroughs on the horizon: semi-autonomous cars and trucks and perhaps even fully autonomous ones; widespread precision medicine leading to longer and healthier lives; transformative improvements to education and the pursuit of science; new ways to pursue the humanities; and evolution in the workplace. There are new data science applications brewing in almost every field of human endeavor.

However, no new technology arrives without complications: Some of the complications are technological, based on challenges in both developing algorithms and then perfecting software and computer systems. For example, with so much data and processing capability, there are inevitably security, privacy, and reliability challenges. And, if applications of data science become as omnipresent as predicted, society needs the technologists to ensure they are rock solid.

Some complications are broader, relating to the very premise of using data to valuable effect. With mountains of data and correlations becoming available, we need to learn to cut through them to ascertain fundamental truths, not erroneous associations which may obfuscate the truth. Deeper risks arise when using data in decision systems; as new applications become available, and we can predict and optimize many outcomes, we must decide what we are really trying to achieve.

Some challenges are truly fundamental, as data science may change the operation of our society and impact our own humanness. We must come to grips with limitations on how much mechanistic advice and control we are willing to act on or even receive. As these systems alter our jobs and our socio-political systems, we will need to understand their effects and adjust in ways that we do not yet understand. Data science is affecting us already, and it may even challenge our notions of ourselves as the intelligent masters of our world.

Because of these very broad impacts, data science as a field has led to entirely new research agendas outside of its foundational fields of computer science, statistics, and operations research. Data science is also changing many other disciplines (e.g., how we think about and practice political science, but many more). There are also growing transdisciplinary relationships between data science and many of the humanities and social sciences.

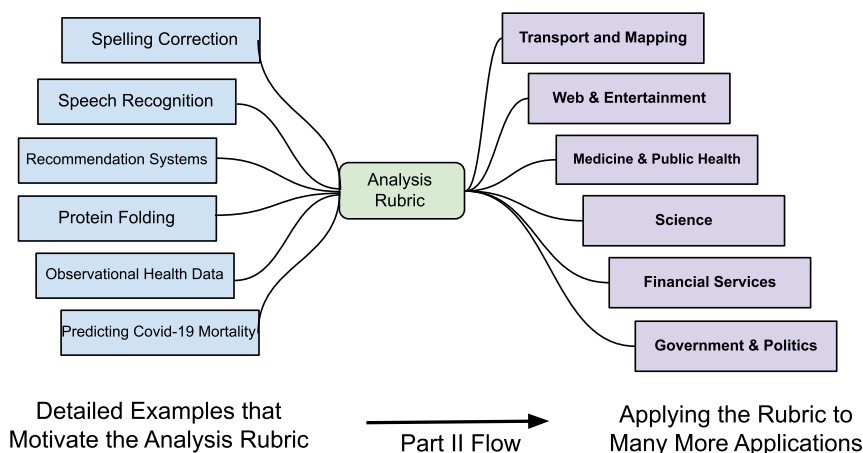
This book's holistic approach to data science leads us through these topics:

- [Part I](#), Data Science, provides a unifying definition of data science and sets forth the field's goals. It then provides a historical perspective on how data science arose from its foundational fields (statistics, operations research, and computing, metaphorically illustrated in [Figure I.1](#)) and describes its relationship to the sciences, social sciences, and humanities. The historical story is an exciting one due to exceedingly rapid progress that has changed the course of technology, many domains of applicability, and even our society writ large.



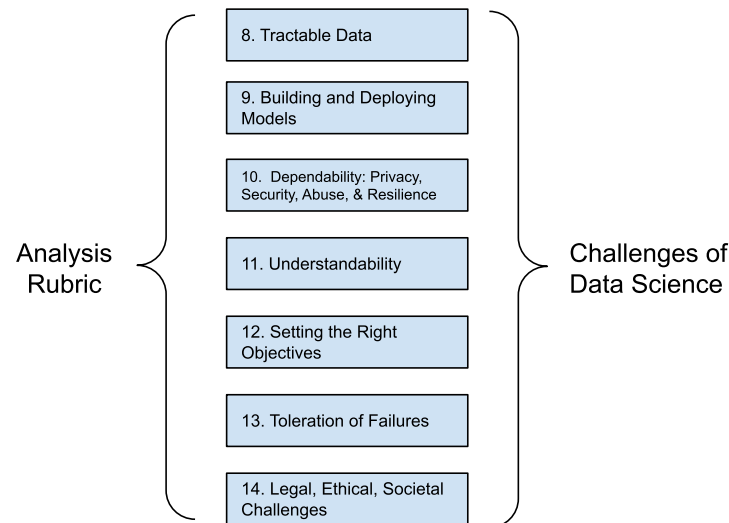
This metaphorical braid shows the integration of the foundational fields, labeled S, OR, and C.
 Figure I.1 Integration of Statistics, Operations Research, and Computing

- [Part II](#), Applying Data Science, presents examples of data science applications from the domains of technology, commerce, science, medicine, and more. Based on our detailed exposition of six applications, the chapter develops a seven-element Analysis Rubric to help us analyze the relative ease or difficulty of applying data science to other applications. We then review twenty-six more applications against the rubric. Some of these are straightforward; others gnarly but feasible; yet others nearly impossible. Almost all have unintended consequences that require care and thought. [Figure I.2](#) illustrates this part's flow.



Part II introduces six applications, uses them to induce a rubric, and then demonstrates its application.
 Figure I.2 Applying Data Science

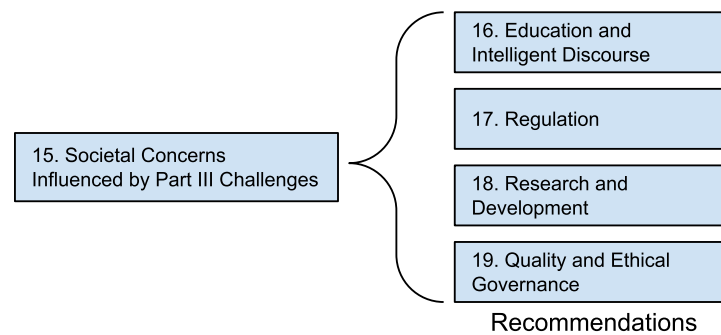
- [Part III](#), Challenges in Applying Data Science, builds off the seven elements in the Analysis Rubric to present the technical, contextual, and societal challenges in making data science work well. (This is illustrated in See [Figure I.3](#).) With care, users of data science can often navigate many of these challenges effectively. However, some are perilous and very difficult to resolve, implying that in some cases, data science is simply not the right tool for the job. [Part III](#) is quite clear about the risks of the misapplication of technology.



The analysis rubric's seven elements motivate the challenges in Chapters 8 to 14.

Figure I.3 Challenges in Applying Data Science

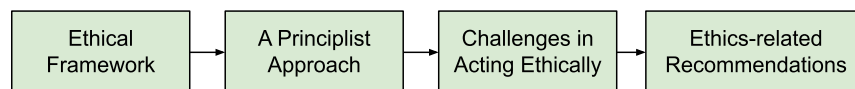
- [Part IV](#): Addressing Concerns, describes many societal concerns regarding data science and its applications – concerns which in turn are influenced by [Part III](#)'s challenges. It then discusses some approaches for mitigating these concerns while still allowing us to reap the rewards. In some cases, we make prescriptive proposals: For example, we recommend increasing data science education at the secondary school level and above, even if this means reprioritizing a little of the current mathematics curriculum and substituting more probability, statistics, and computing. In other areas, we only set forth some considerations that decision makers should take into account. Part IV's flow is described in [Figure I.4](#).



Part IV's summary of societal concerns motivate the recommendations in Chapters 16 to 19

Figure I.4 Societal Concerns and Recommendations

- An ethics thread flows through the book, with a focused section near the end of each part. Data science must consider ethical matters carefully because many data science applications have significant societal consequences and often rely on personal data to create computational models. As illustrated in [Figure I.5](#), the thread starts by defining ethical principles relevant to data science and then reviews some of the [Part II](#) applications in light of those principles. While most chapters of the book (and particularly those in [Part III](#)) present ethics-related issues, the ethics thread augments these discussions with the organizational challenges of balancing incentives and governance to achieve good outcomes. The ethics thread concludes in [Part IV](#), which ends with three recommendations.



This figure illustrates the flow of the ethics thread, which spans Parts I to IV.

Figure I.5 Ethics Flow Throughout the Book

While the chapters build on each other, data science courses will vary in what examples from [Part II](#) they emphasize. Some readers may choose to omit [Part IV](#) (which bridges from the *challenges* of [Part III](#) to societal *concerns*), while others may wish to omit some of the technical details in [Part II](#) and [Part III](#).

In all, this book's broad perspective on the field of data science aims to educate readers about the data science applications they regularly use, to apply that understanding to new applications, to more fully recognize the challenges inherent in data science, and to educate and catalyze thoughtful analysis, debate, and action to make data science ever more beneficial.

Chapter 1. Foundations of Data Science

This chapter first defines data science, its primary objectives, and several related terms. It continues by describing the evolution of data science from the fields of statistics, operations research, and computing. The chapter concludes with historical notes on the emergence of data science and related topics.

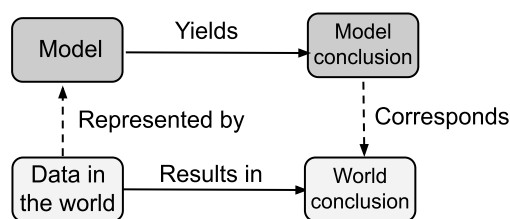
1.1 Definitions

Data science is the study of extracting value from data – value in the form of **insights** or **conclusions**.

- A data-derived insight could be:
 - A hypothesis, testable with more data;
 - An “aha!” that comes from a succinct statistic or an apt visual chart; or
 - A plausible relationship among variables of interest, uncovered by examining the data and the implications of different scenarios.

- A conclusion could be in an analyst’s head or in a computer program. To be *useful*, a conclusion should lead us to make good decisions about how to act in the world, with those actions either taken automatically by a program, or by a human who consults with the program. A conclusion may be in the form of a:
 - **Prediction** of a consequence;
 - **Recommendation** of a useful action;
 - **Clustering** that groups similar elements;
 - **Classification** that labels elements in groupings;
 - **Transformation** that converts data to a more useful form; or
 - **Optimization** that moves a system to a better state.

Insights and conclusions often arise from **models**, which are abstractions of the real world. A model can explain why or how something happens and can be tested against previously unseen inputs. This is shown schematically in [Figure 1.1](#).



From data in the world, we build a model of some aspects of it, reason about the model to draw conclusions, and check that these conclusions correspond to what happens in the world. The better the model, the better the correspondence between the model's conclusions and the real world. Dashed arrows denote the mapping between world and model, and solid arrows are within the world or model.

Figure 1.1 Models and the World

Of course, scientists and lay people have used data and models for centuries. Today's data science builds on this usage. But it differs from classical data use due to the scale it operates at and its use of new statistical and computational techniques.

There is still no consensus on the definition of data science. For example, the *Journal of Data Science* in its initial issue says “By ‘Data Science’ we mean almost everything that has something to do with data”; Mike Loukides, co-author of *Ethics and Data Science*, says “Data science enables the creation of data products”;² Cassie Kozyrkov, Google’s Chief Decision Scientist, says “Data science is the discipline of making data useful.”³ We believe our definition is consistent with other definitions and that it is usefully prescriptive.

If a retailer tracks a billion customer transactions, analyzes the data, and learns something that improves their sales, that’s a data science insight. If the retailer then automatically recommends to customers what to buy next, that’s a data science conclusion enabled by a model, perhaps one that uses machine learning.

Data Science touches all of society. We will highlight many applications in transportation, the web and entertainment, medicine and public health, science, financial services, and government. However, there are many others in the humanities, agriculture, energy systems, and virtually every field. In recognition of data science’s cross-disciplinary nature, this book presents data science issues from multiple points of view.

1.1.1 Data Science – Insights

Data science offers insights by permitting the exploration of data. The data may show a trend suggesting a hypothesis in the context of a model that leads to useful conclusions – which themselves can be tested with more data. A trend might indicate that two (or more) things are **correlated**, meaning the variables are related to each other, such as smoking and cancer. A potential correlation is an insight, and a hypothesis that can be tested. The data may even

suggest the possibility of an underlying **causal relationship**, which occurs when one thing causes another – smoking causes cancer, though cancer does not cause smoking. Or perhaps a conclusion is not obvious, but can be explored with many what-if analyses that also draw on more data.

Insights are facilitated by interactive tools that simplify this exploration and let us benefit from vast amounts of data without bogging down and missing the forest for the trees:

- Tools to help us gain insight start with data transformation, which converts units, merges names (such as “Ohio” and “OH”), combines data sources, and removes duplicates, errors, and outliers.
- Tools to automate experiments by providing integrated modeling capabilities that simplify creation, execution, exploration, and record keeping.
- Tools that offer interactive capabilities that guide us to non-obvious conclusions.

Pioneering data scientist John Tukey said, “The simple graph has brought more information to the data analyst’s mind than any other device,”⁴ but modern visualization offers many other beautiful and useful ways to gain insight. However, graphs must be scrutinized very carefully for meaning.

As an example of a graph that provides some insight but that also leads to many questions, the scatter plot in [Figure 1.2](#) shows the relationship between mortality and COVID-19 vaccination rates during the US delta variant wave. It shows four series of points representing different time periods ranging from delta’s beginning mid-2021 to its late 2021 end. Each point represents the vaccination rate and number of COVID-19 deaths in each of the fifty states and the District of Columbia. We show **regression lines** for each of the four series of data – each line represents the linear equation that best fits the data. Critical analysis would be served with error bars for each data point, but this information was unavailable.

The 6-Sep-21 and 27-Sep-21 series data were from the peak of the wave, and they tilt strongly down and to the right, meaning that higher state vaccination rates were strongly correlated with lower death rates. The 11-Jul-21 and 16-Dec-21 regressions (beginning and ending of the wave) showed small negative slopes, but reports of the CDC’s imprecision in vaccination reporting⁵ sufficiently concerned us that we provided a prominent warning on the graph, which also demonstrates a good visualization practice. Clearly, this data’s association of vaccination rate on mortality declined after the delta wave crested. During the five-month period, the chart also shows that vaccination rates increased by about 13% (absolute).

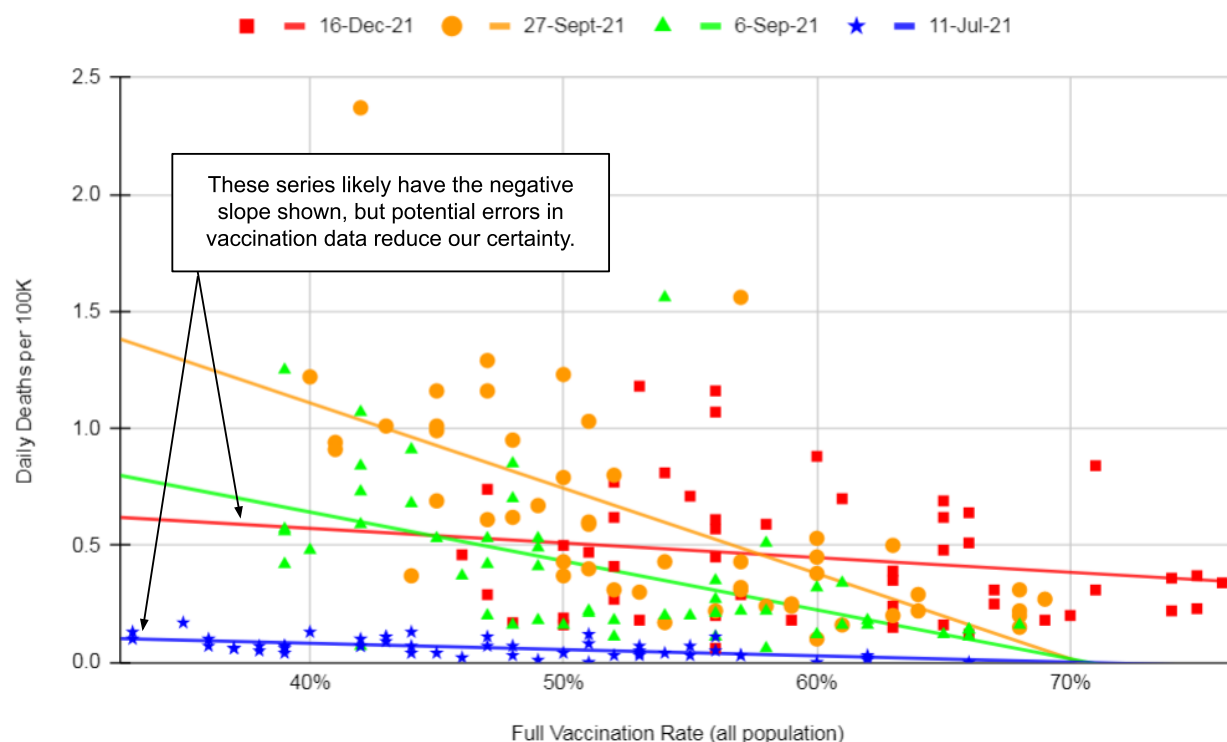
This data and our prior understanding of vaccine biochemistry lead us strongly to believe there is an underlying causal relationship – that vaccinations reduce the risk of deaths. (The CDC COVID Data Tracker provides even stronger evidence of a causal relationship.⁶) However, [Figure 1.2](#) does *not* provide conclusive insight, as there *could* be other explanations for some of the effects. States differ along many relevant variables other than vaccination rate, such as

population age, density, and prior disease exposure. This is not a randomized controlled experiment where each state was randomly assigned a vaccination rate. The reasons the curve flattened at the end of the wave may not be because of reduced vaccine efficacy against the delta variant but rather because of the impact of behavioral changes, changes in the locale of the wave as it spread across different states, increase in immunity from prior exposure, waning vaccine efficacy over time, and the very beginning of the follow-on Omicron wave.

A data scientist could gain further insight from the analysis of outliers. If not an artifact of the data, the twin 1.6 per 100K points that came from Florida, for example, may result from disease in the state's large at-risk elderly population. Data scientists could construct and evaluate many hypotheses from this graph using additional data and visualization techniques. But data scientists need to exercise caution about the quality of individual data points.

Daily Deaths/100K (7 day Trailing) vs Fully Vaccinated Percentage of Population

Series Representing Different Time Periods in Covid-19 Delta Wave



Each point shows the 7-day trailing average daily COVID-19 mortality of 50 US states and the District of Columbia plotted against their respective vaccination rates at the end of the time period. This data (though not this visual) was copied from the NYTimes Coronavirus in the US: Latest Map and Case Count during the period represented by this graph.⁷ The NY Times itself gathered this data from government authorities, and this limited data was likely to be comparable across regions and time periods. US CDC data (not shown) reported state totals that vary from NYTimes data, but the trend lines are very similar.

Figure 1.2 Deaths Versus Full Vaccination

The US omicron wave, which followed the delta wave, showed a different regression line. While Figure 1.2 does not illustrate this, state per capita mortality and vaccination rates became positively correlated for a brief period in mid-January 2022, though just slightly so. There are many possible explanations for this such as the specifics of the omicron mutation and the earlier arrival of the variant in vaccinated states. The reversal, and indeed this chart, reminds us to scrutinize data and visualizations carefully and to exercise due caution, recognizing the limitations of the data and its presentation. [Section 11.4](#) discusses this topic further.

1.1.2 Data Science – Conclusions

Let's look at some examples of our six types of conclusions from the beginning of [Section 1.1](#). Conclusions can be embedded in programs or serve to provide insight to a data analyst.

- **Prediction:**
 - Predict how a protein will fold, based on its structure.
 - Auto-complete user input, based on the characters typed so far.
- **Recommendation:**
 - Recommend a song, based on past listening.
 - Suggest possible medical therapies, based on laboratory results.
 - Show an ad to a user, based on their recent web searches.
- **Classification:**
 - Assign labels to photos (e.g., “cat” or “dog”).
 - Identify a bird's species, from its song.
 - Determine if a client is satisfied or unsatisfied, via sentiment analysis.
 - Label email as spam.
- **Optimization:**
 - Find the optimal location to build a new warehouse based on minimizing supplier/consumer transportation costs.
 - Schedule product manufacturing to maximize revenue based on predicted future demand.
- **Transformation:**
 - Translate a sentence from Chinese to English.
 - Convert astronomical images to entities.
- **Clustering:**
 - Cluster together similar images of cancerous growths to help doctors better understand the disease.
 - Cluster email messages into folders.

Models that generate these conclusions may be **clear box** or **opaque box**. A clear box model's logic is available for inspection by others, while an opaque box model's logic is not. The “opaque box” term can also apply to a model whose operation is not

comprehensible, perhaps because it relies on machine learning. Context usually clarifies whether opacity refers to unavailability, incomprehensibility, or both.

This book is filled with many examples of using data to reach conclusions. For example, [Chapter 4](#) leads off by discussing data-driven spelling correction systems, which may *classify* words into correct or mispelled variants (perhaps underlining the latter), *recommend* correct spellings (“did you mean, misspell?”) or automatically *transform* an error into a correct spelling. Returning to the mortality insight discussion that concluded the previous section, we also discuss COVID-19 mortality prediction in greater detail, but we will see this is hard to do even when there is much more data available.

1.1.3 Scale

Some data science success is due to new techniques for analysis, and new algorithms for drawing conclusions. But much is due to the sheer scale of data we can now collect and process.⁸

As examples of the size of data collections as of 2021: There are 500 billion web pages (and growing) stored in the Internet Archive. The investment company Two Sigma stores at least a petabyte of data per month. YouTube users upload five hundred hours of video per minute.⁹ The SkyMapper Southern Sky Survey is 500 terabytes of astronomical data; the Legacy Survey of Space and Time is scheduled to produce 200 petabytes in 2022.¹⁰ See [Table 1.1](#) below, which describes the scale of data with representative examples.

Table 1.1 Scale of Data and Representative Examples

Size			Example
10 ³	KB	Kilobyte	A half page of text, or a 32x32 pixel icon
10 ⁶	MB	Megabyte	The text of two complete books, or a medium-resolution photo
10 ⁹	GB	Gigabyte	An hour-long HD video, ten hours of music, or the Encyclopedia Britannica text
10 ¹²	TB	Terabyte	One month of images from Hubble Space Telescope or a university library's text
10 ¹⁵	PB	Petabyte	Five copies of the 170 million book Library of Congress print collection
10 ¹⁸	EB	Exabyte	Twenty copies of the 500 billion page Internet Archive, or two hours of data at the planned rate of the Square Kilometer Array telescope in 2025
10 ²¹	ZB	Zettabyte	World's total digital content in 2012, or total internet traffic in 2016

Chapter 2. Data Science is Transdisciplinary

Data science's primary progenitors are statistics, operations research, and computing, but the sciences, humanities, and social sciences are also all part of its story for these reasons:

1. **New Application Areas:** Data science adds valuable techniques to a large and growing number of application domains, building on their unique data and pre-existing capabilities. The combined capabilities of data science and domain-specific know-how can solve important scientific and social problems and have commercial value.
2. **Advancing Data Science:** Some domains possess advanced methods for dealing with their data. Data science benefits by incorporating these methods and then making them available for wider use. For example, petabyte-scale datasets generated by experiments in physics, astronomy, and biology led to inventing new data science techniques.
3. **Building Coalitions:** Data science operates in a societal context. Making sure we “get it right” requires partnerships which must include viewpoints from non-STEM domains such as sociology, law, economics, philosophy, and political policy. Good solutions can have great societal benefit, while poor ones can cause harm.

This chapter's title includes the term *transdisciplinary* to emphasize that data science has been able to achieve its theoretical, methodological, and practical results by combining the approaches of different disciplines to create a new field.^{F6} This combination is needed not only for data science's core disciplines, but also for its application areas and the fields that influence its proper use.

2.1 New Application Areas

We use two approaches to illustrate data science's broad applicability. First, we discuss data science's relevance to each economic sector. Second, we consider its relevance to academic research areas.

For the first approach, we divide the entire economy into major buckets, using the US Bureau of Economic Analysis GDP Report as a guide.⁷² Despite the data's US-centricity, the categories of economic activity are probably representative of most economies, and an analysis shows existing and growing data science roles in each one. [Table 2.1](#) lists the categories and a few example data science applications for each.

⁶ *Multidisciplinary* is when different fields separately contribute their approaches to a problem. *Interdisciplinary* is when the approaches interact. *Transdisciplinary* is when a new field emerges from that interaction.

Table 2.1 Components of the Economy and Data Science Applicability

Sector of US Economy	Areas of Data Science Applicability
Agriculture, forestry, fishing	Precision cultivation and harvesting, fishery management, quality control, risk reduction
Mining	Predicting resource location, risk management, pricing
Utilities	Fault detection, optimized energy sources, production automation, predictive maintenance
Construction	Scheduling, logistics, optimized design and materials use
Manufacturing - durable goods	Quality Control, production scheduling, automated design and manufacturing, customer support
Manufacturing - non-durable goods	Risk management, logistics, demand prediction, pricing
Wholesale trade	Inventory management, demand forecasting, logistics
Retail Trade	Merchandising, advertising, pricing, upsell, loyalty programs, inventory management
Transportation and warehousing	Optimized routing, storage, pricing, tracing, safety monitoring, semi- and fully automated vehicles, yield management
Information	Advertising, audience engagement, content moderation, translation services,
Finance and Insurance	Risk assessment, portfolio construction, security, regulatory monitoring
Real estate and rental and leasing	Construction, maintenance, & property mgmt., service automation
Professional, scientific, & tech. services	Mapping and surveying automation, new s/w development tools, data-driven marketing, data-driven science
Management of companies & enterprises	Decision support of all forms, improved communications
Other administrative services	Employee hiring and scheduling, automated transcription, security monitoring, credit scoring
Education, health and social assistance	Personalized education, remote health monitoring, disease diagnosis, social service delivery, fraud detection
Arts, entertain., recreation, food services	Personalization, pricing, upsell opportunities, automation, immersive experiences
Other highly diverse services	Fault diagnosis, dating services, locating civic needs, fund-raising
US national government defense	Logistics, guidance and targeting, decision support, maintenance, readiness, cybersecurity, wargaming
US national government non-defense	Tax audit, civic outreach, societal and economic monitoring
US state and local government	Maintenance operations, educ. programs, criminal justice system, monitoring service fairness

This first column divides the breadth of economic output into buckets. These come from the Bureau of Economic Research, which sources them from the North American Industry Classification System.⁷³ The second column is an eclectic list of data science-enabled applications, either existing or soon to be likely.

While this book provides detailed examples from many of the areas, we have inevitably omitted some. Among others, we do not devote much attention to data science's many uses in national

defense-related topics such as logistics, guidance and targeting, decision support, maintenance, or cybersecurity. We also don't discuss uses in precision agriculture, factory automation, building management, optimizing social service delivery, and so on.

For the second approach (relevance to academic research areas), we consider the role of data science in the various university education and research disciplines. Below, we discuss science, social science, engineering, and the humanities.

2.1.1 Sciences

The sciences have been a major source of data science use cases. As data becomes easier to use, scientific models have become more complex and highly tuned to real world inputs. In some cases, data science has automated the process of creating models.

As an example, efforts to combat the COVID-19 pandemic show data science's increasing role in biomedical and social applications. Vaccine development and deployment could not have happened as quickly without pre-existing infrastructure, tools, and data ready to be used in a new situation:

- Genetic and protein structure databases and other tools for simulating structure facilitated the rapid decoding and promulgation of the underlying SARS-CoV-2 genetic structure.
- Tools were in place to manipulate this genetic data.
- Large scale data management technologies were used to rapidly create, locate, manage, and monitor well-structured clinical trials.
- Logistics data and algorithms helped plan and control the complex supply chain for vaccinating large populations.

On the other hand, our COVID-19 experience showed some data science weaknesses: in particular, limitations in our ability to draw conclusions from public health monitoring and prediction. We give examples of these and other weaknesses throughout this book.

Aspirationally, Turing Award winner Jim Gray proposed a new model for scientific research, which he termed *the Fourth Paradigm*, in a talk at a 2007 National Academies meeting. In it, he talked about how science can increasingly benefit from new tools and techniques for data capture, analysis, and communication/publication. Gray was first and foremost a computer scientist specializing in databases, but he became involved in projects at the borders of astrophysics, mapping, and computer science. This led to his advising scientists in many disciplines on their growing data-related problems. As captured and edited by colleagues at Microsoft Research, Gray said:

The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline.

The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science.⁷⁴

Science is increasingly moving in this direction, particularly with the rapid growth in machine learning capabilities. As an example, scientists trained a neural network on thousands of molecules with known antibacterial properties. They then applied that network to a dataset of over six thousand compounds with potential antibiotic activity. This approach quickly uncovered a potential new drug, Halicin, for treating certain antibiotic-resistant bacteria.⁷⁵ While this data-centric approach is capable of screening far more than six thousand compounds and preliminary laboratory studies showed positive results, this particular drug may well encounter roadblocks on the path to approval.⁷⁶

2.1.2 Social Sciences

Much of the social sciences involves gathering and analyzing data. In economics, this was traditionally confined to the subfield of econometrics, which Samuelson explained as enabling one to “apply the tools of statistics ... to sift through mountains of data to extract simple relationships.”⁷⁷ However, today data science more broadly impacts all of economics. The related area of finance has been at the forefront of using large datasets and sophisticated predictive models in diverse applications. We present specific examples from economics and finance in [Section 6.5](#)

In broader society, governments have always needed information about their populations. For example, at the most basic level they need to count their people and collect taxes. They collect a great deal of other information as well; for instance, labor statistics, transportation statistics, and health data. From the other side, voters are interested in the results of pre-election polls. Polling agencies, and social media platforms also collect a great deal of political and sociological data. More and more, information gathering underlies societal systems. We discuss some examples of political and governmental data science in [Section 6.6](#).

In the future, perhaps all economic transactions will be digital, and physical currency will no longer be used. (China's early 2020's experimentation with a national digital currency may foreshadow this.) Economists may be able to measure and track all financial flows, and offer new mechanisms for economic governance. In a very ambitious proposal, data scientists and economists hypothesize that data science results could dynamically set tax policies to more efficiently balance revenue and equity objectives, as discussed in [Section 6.5](#).⁷⁸ A related, much simpler, present-day example is congestion pricing, in which highway tolls vary to reduce congestion and motivate use when there is extra capacity.

2.1.3 Engineering

Engineering disciplines abound with data science use cases, both in the design of new products and services and the efficient operation of complex systems. For the former, data science and

machine learning are the basis of GitHub Copilot, which assists programmers.⁷⁹ Google researchers have developed a system that learns to do the physical layout of devices on computer chips.⁸⁰ Civil engineers are likewise researching similar approaches to design far larger structures made of steel rather than silicon. Many engineering challenges, such as speech recognition as discussed in [Section 4.2](#) and [Section 5.2](#), were practically unsolvable until we applied data driven approaches to them.

In the domain of engineering operations, data, often directly provided by embedded instrumentation, can predict maintenance needs, provide early warnings of failures, and optimize system operations for applications including cars, power grids, rails, naval propulsion systems, jet engines, and many more. Neural networks can classify work orders and optimize environmental and power systems. Continual anomaly detection is at the heart of many computer security systems, whether they are looking for operational failures or intrusions.

2.1.4 Humanities

In the humanities, applications have been slower to come into common use. The earliest days of computers did see occasional projects to connect digitized data with literature, art, and history. These early, one-of-a-kind, efforts all focused on single works. As large web-based archives of books and images were built, interest grew substantially in applying data science tools and techniques. Circa 2010, academicians began to recognize the opportunities,⁸¹ and the US National Endowment for the Humanities launched its Digital Humanities Initiative. In 2009, the National Science Foundation (where co-author Jeannette was working at the time), in partnership with the National Endowment of Humanities, the Joint Information Systems Committee of the UK, and the Social Sciences Research Council of Canada, launched a “Digging into Data” challenge. It asked the question “What could you do with a million books?”.^{F7} It continues to this day, with 17 additional international partners.

Leveraging the millions of Google-scanned books, Google Research gave out awards in the digital humanities,⁸² initially focusing on text analysis. JB Michel et al. (including co-author Peter) used the Ngram viewer’s frequency count of words or phrases in millions of books to gather insight on many societal or linguistic changes, ranging from the impact of censorship on books published to the rate at which irregularly conjugated English verbs change to become regularly conjugated.⁸³

Matt Connelly, circa 2015, created the History Lab Project at Columbia, which maintains the world’s largest collection of declassified documents and lets researchers analyze them using data science techniques.⁸⁴ Legal scholars are analyzing the tens of millions of online court judgments released by China since 2014, to better understand the Chinese legal system and its rulings and consequences.⁸⁵ Many projects now geocode large numbers of records, placing

⁷ A million books may not sound like a lot, but reading a book a day from birth to age 100 only amounts to 36,525 books (give or take a day for leap years).

people, historical events, or statistical measures on a map to provide insight or bring history to life. The Smithsonian Collections Search Center is an online catalog with 17.2 million records, “records relating to areas for Art & Design, History & Culture, and Science & Technology with over 6.6 million images, videos, audio files, podcasts, blog posts and electronic journals.”⁸⁶

We could discuss many more applications in virtually every domain. We conclude this section by observing that data science's applications are broad and growing, while requiring the fusion of data science and discipline-specific capabilities to achieve their goals.

2.2 Advancing Data Science

As other fields address their data-related problems, they often end up making their own important contributions to data science. Necessity being the mother of invention, if a new problem requires a new capability, the field may be rapidly advanced.

A great example of a new capability coming out of physics is the World Wide Web (WWW). Sir Tim Berners-Lee created what became the WWW to help the greater CERN supercollider research community communicate and collaborate. It turned out this was such a good idea that it rapidly caught on elsewhere and became a fundamental pillar of both data science and our everyday lives. As another example, many important technologies for efficient pattern matching came from computational biologists' need to match nucleic acid sequences. Finally, social scientists' need for census data is forcing consideration about making aggregations of data broadly and accurately available while still preserving privacy. We have a new understanding of privacy-preserving data aggregation – regrettably, one showing its difficulty.

Soon after the World Wide Web became available to the general public, online advertising became a big business. Advertising fees paid for the growth of many well-known internet services and also brought huge attention and money to data science. Deciding what ad to show to a viewer is a data science problem; there is data on what the current user is searching for the pages they have been browsing, and there is past data on what similar users have done in the past. Economists contributed by introducing data scientists to algorithmic game theory. Among other things, this helps determine what auction style is best for selling ad space, balancing interests of consumers, sellers, and web publishers.

2.3 Building Coalitions

Data scientists need to partner with many other disciplines to ensure their resulting work will be maximally beneficial, societally acceptable, or perhaps even legal. Here are six examples:

- **Philosophers** can help frame ethical considerations about data science work, including issues of privacy, free will, and fairness. These will be considered throughout this book, starting in the next section, as we lay out an ethical framework.
- **Lawyers, politicians, and political scientists** can help with the legal and policy issues relating to data stewardship and the governance of data-intensive applications. Public

opinion can be strongly influenced by information publishing and recommendation systems

- **Designers** and **psychologists** can help data scientists present aesthetically pleasing and accessible results that users find easy to understand and use.
- **Economists** (in microeconomic, behavioral economics, and more) bring economic modes of analysis to data science problems that can help achieve efficiency or fairness.
- **Sociologists** provide insights for new ways to study human behavior and social relationships using data from large digitally connected social networks.
- **Journalists** can aid data science's goal of explanation so it emphasizes truthfulness and is valuable to users and society. Computing a number or displaying a bar graph is not enough; the results need to tell a coherent and truthful story.

Chapter 3. A Framework for Ethical Considerations

Data-empowered algorithms are reshaping our professional, personal, and political realities, and they are likely to have an even larger effect going forward. However, as with all developing technologies, increases in impact inevitably give rise to unanticipated consequences. These challenge our norms for how we use technology in ways consistent with our values. Many scholars, educators, and technology companies refer to these as **ethical challenges**, building on the applied ethics tradition from basic sciences.

Some challenges are best met by inventing improved or more nuanced technological approaches. However, many challenges will still arise based on how we deploy technology as products, or how statistical analysis interpretations guide law and policy.

While the word *ethics* may imply a branch of somewhat obscure philosophy, the applied ethical tradition is about both defining ethics and designing ethical processes clearly enough to help guide good choices. In the case of data science, it is also to develop programs that make good choices.

3.1 Professional Ethics

Companies and professional societies, including the American Statistical Association (ASA), the Institute for Operations Research and the Management Sciences (INFORMS), the IEEE, the Association for Computing Machinery (ACM), and Engineers Canada have long had important and useful ethical codes addressing matters of personal conduct and technical execution.^{87–90} These include principles such as honesty, impartiality, and integrity.

The introduction to the ASA's code observes, "The discipline of statistics links the capacity to observe with the ability to gather evidence and make decisions, providing a foundation for building a more informed society. Because society depends on informed judgments supported by statistical methods, all practitioners of statistics—regardless of training and occupation or job title—have an obligation to work in a professional, competent, respectful, and ethical manner."⁸⁷

As the impact of statistics, operations research, and computing (and analogously, data science) has grown, many of these codes are being generalized to include broader societal considerations. Gotterbarn and Wolf write in a preamble to the 2018 ACM Code of Ethics that, "we find ourselves in situations where our work can affect the lives and livelihoods of people in ways that may not be intended, or even be predictable. This brings a host of complex ethical considerations into play."⁹¹

3.2 The Belmont Commission

In the human subjects research community, the *Belmont Report* is the central document of applied ethics in biomedical and behavioral research. In it, ethics is defined in terms of general principles.⁹² The Belmont commission met monthly for four years in response to the 1932-72 US Public Health Service Syphilis Study at Tuskegee, a morally and scientifically flawed medical experiment. By including commissioners from a wide range of fields, including researchers, lawyers, administrators, and philosophers, the organizers hoped to protect human subjects while balancing societal norms, legal constraints, and society's need for innovation.

Despite its roots in human subjects research context, the report outlines principles that are sufficiently general to be a basis for a useful ethical framework for data science research and products. In Belmont, these principles are called “respect for persons, beneficence, and justice.” In more detail, they were then framed as:

- **Respect for persons.** This means ensuring the freedom of individuals to act autonomously based on their own considered deliberation and judgements. Often summarized as informed consent, this principle also includes having sufficient transparency to make judgements and also defending the autonomy of those with diminished consent, e.g., children or those who may be coerced into making a decision.
- **Beneficence.** Belmont encourages researchers *not* to limit their thinking to “do no harm,” but to maximize benefits and balance them against risks. Doing so requires careful consideration of the immediate risks and benefits as well as a commitment to monitor and mitigate harms as results occur.
- **Justice.** The consideration of how risks and benefits are distributed, including the notion of a fair distribution. Fair may not mean “equal” but rather that the risks are borne by the populations who stand to benefit (and are not born by populations who will not ultimately have access to the fruits of the research).

These principles are intended to be broad and therefore applicable to yet unenvisioned technology changes and their consequences. At the same time, they are intended to be sufficiently specific that communities can come to a shared, deliberative consensus as to their implied best actions. In other words, from general, common principles, a community derives more context-specific standards and instance-specific rules. For a technologist, these rules imply even more specific design choices in modeling or in data product development.

This principled approach to ethics does not offer a single all-encompassing checklist that one consults for an answer that is the same in all contexts. Instead, principles are, by design, in tension with each other. They provide a basis to ask specific questions, which often do not have a right or wrong answer, but illuminate the tension in a situation or between positions.

3.3 Belmont Application to Data Science

The breadth of data science's impact argues for applying Belmont-like principles to it. Numerous scholars,⁹³ researchers, and technologists have suggested how these principles can guide applied ethics even in the context of data-empowered algorithms. They sometimes also argue for extending the principles to emphasize the impact on society at large.^{94,95} However, as Belmont frames them, the principles provide a common vocabulary for researchers, data scientists, product developers, and regulators with which to reach consensus.

As co-author Jeannette writes in her 2020 essay, *Ten Research Challenge Areas in Data Science*: "The ethical principle of Respect for Persons suggests that people should always be informed when they are talking with a chatbot. The ethical principle of Beneficence requires a risk/benefit analysis on the decision a self-driving car makes on whom not to harm. The ethical principle of justice requires us to ensure the fairness of risk assessment tools in the court system and automated decision systems used in hiring."⁹⁶

Another voice in this area comes from the European Commission, the executive branch of the European Union. Their *Ethics guidelines for Trustworthy AI* shows that the need for ethical frameworks in technological areas is recognized by world governments, as well as researchers and ethicists.⁹⁷

We, of course, accept there are distinctions between data science-oriented implications of Belmont and human subject research, in particular medical research. In medical research, the principles motivate standards such as **informed consent** (a process of disclosing risks and benefits to an individual before gaining approval) and **fair selection of subjects**.

However, the digital domain can have different standards that are also consistent with the Belmont Principles, often because algorithms initiate automated actions. Here are some example considerations:

- Informed consent is hard to achieve in our current digital environment. To use a digital product, users must click "I agree," most often without comprehending the long and complex terms of service authorizing software actions over an extended period. Barocas and Nissenbaum identify "complex data flows" such as in digital services as possessing what they term a **transparency paradox**.⁹⁸ They argue that information disclosure provided to users is so simple as to be incomplete or deceptive, or it is so complex as to be incomprehensible. Data scientists adhere best to informed consent by respecting user norms at a level of transparency that avoids deception or unfairness, while allowing more detailed auditing and critique (e.g., via appropriate technical documentation or open source).
- For software, the risk-benefit balance of beneficence includes thinking through unintended consequences. It also requires the humility to recognize how hard it is to anticipate all the ways people will experience or use a product. That requires a commitment to monitor and mitigate harms as they are revealed.

- Justice, in the context of data-driven products, includes ongoing assessment of their fairness (technical and otherwise) as well as their training datasets. Justice includes fairness, with the understanding that defining “fair,” even in technical communities, can be subjective,⁹⁹ and is not as simple as giving it the same meaning as “equal.” Our norms of justice also include an understanding of addressing and redressing prior harms, where possible. We say much more on fairness in [Section 12.3](#).

We use the Belmont Principles to organize the analysis of several case studies in [Part II](#). We then address the challenges to aligning ethics with a university’s or technology company’s data scientists’ operational process in [Part III](#). [Part IV](#) contains recommendations on how to proceed in the future.

While the Belmont Principles are our ethical starting point, we recognize that applications of data science may also require the consideration of other ethical principles on which societies are based. For example, the principles of justice of war (**jus ad bellum**) and the conduct of war (**jus in bello**) are relevant to data science applications in military domains.¹⁰⁰ Furthermore, data science is a sufficiently new field that we may eventually need to identify new relevant principles for its ethics.

Chapter 5. The Analysis Rubric

This chapter defines the **Analysis Rubric**, which consists of seven major considerations for determining data science's applicability to a proposed application. While these considerations may not be fully understood at a project's inception, there needs to be a belief that answers will be forthcoming prior to completion. Three of these address requirements-oriented aspects ("For What or Why") of data science applications, and three address implementation-oriented aspects ("How To"). The seventh addresses legal, societal, and ethical implications (ELSI^{F8}). Collectively, these considerations, or rubric elements, cover the complex trade-offs needed to achieve practical, valuable, legal and ethical results.

Implementation-Oriented Elements

- **Tractable Data.** Consider whether data of sufficient integrity, size, quality, and manageability exists or could be obtained.
- **A Technical Approach.** Consider whether there is a technical approach grounded in data, such as an analysis, a model, or an interactive visualization, that can achieve the desired result.
- **Dependability.**^{F9} Dependability aggregates four considerations: Does the application meet needed privacy protections? Is its security sufficient to thwart attackers who try to break it? Does it resist the abuse of malevolent users? Does it have the resilience to operate correctly in the face of unforeseen circumstances or changes to the world?

Requirements-Oriented Elements

- **Understandability.** This means the approach must enable others to understand the application. Consider whether the application needs to only provide conclusions or if it will have to explain "why" it has rendered these conclusions. Will the application need to detail the causal chain underlying its conclusions? Or will it make its underlying data and associated models, software, and techniques transparent and provide **reproducibility** – that is, the ability for analysts or scientists to understand, validate, duplicate, or extend the results?
- **Clear Objectives.** Consider whether the application is trying to achieve well-specified objectives that align with what we truly want to happen.

⁸ The acronym, ELSI, stands for Ethical, Legal, & Social Implications. It was coined by James Watson in October 1988 as described in "[ELSI: Origins and Early History](#)."¹³⁰ We will typically address these issues in a more operationally focused order that begins with legal issues, followed thereafter by societal and ethical issues.

⁹ We devoted much effort before settling on *dependability* to aggregate privacy, security, abuse-resistance, and resilience. While dependability is often a generic term, this book will consistently use it as a placeholder for these four properties.

- **Toleration of Failures.** Consider both the possible unintended side effects if the objective is not quite right and the possible damage from failing to meet objectives. Many data science approaches only achieve good results probabilistically, so occasional poor results must be acceptable.

Ethical, Legal, and Societal Implications (ELSI) Element

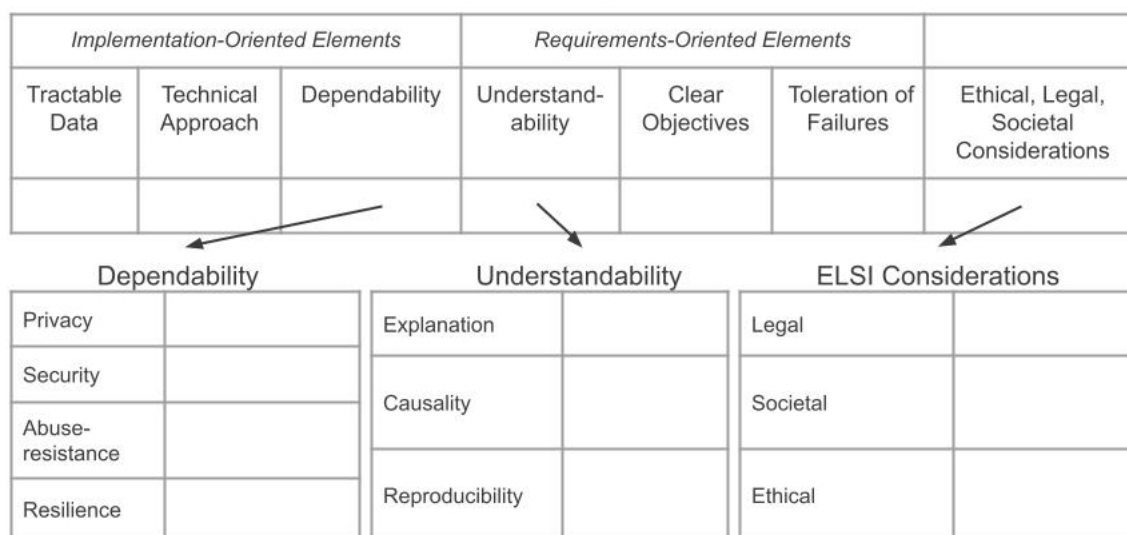
- **Ethical, Legal, and Social Issues.** Consider the application holistically with regard to legality, risk, and ethical considerations. Many of the topics under Dependability or Clear Objectives topics are relevant, but this holistic analysis is broader.

Many applications start with a **bottom-up approach**, focusing on implementation-related rubric elements relating to data availability, a technical approach providing the necessary results, and techniques to provide needed dependability. This analysis then informs the requirements definition and influences its refinement.

Others require a **top-down approach**, first focusing on the requirements-oriented rubric elements relating to understanding, clarity of objectives, and failure tolerance. This analysis then informs the implementation approach and influences its refinement.

Most commonly, the bottom-up and top-down are mixed, and there is iterative flitting back and forth between different considerations. No matter what design approach is used, the ethical, legal, and societal implications must be considered throughout the design and analysis. They cannot be bolted on at the last minute, and they must be carefully reviewed before any effort is declared complete.

Data Science Analysis Rubric



This graphic shows the seven top-level elements of the analysis rubric and the further breakdown of Dependability, Understandability, and Ethical, Legal, and Societal Considerations.

Figure 5-1 Graphical Summary of Analysis Rubric

6.3 Medicine and Public Health Applications of Data Science

[Table 6.3](#) lists several such health applications, supplementing the two presented in [Chapter 4](#) and [Chapter 5](#). We will discuss three briefly, but devote more attention to disease diagnosis, genome-wide association studies (GWAS), and understanding the cause of a disease.

Mobility reporting was introduced by Google in 2020 during the early COVID-19 quarantines and used individuals' location data to chart regional movement trends over time. Its reports were broken down not only by region but also by categories such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residences.¹⁴⁷ Google engineers felt these reports would show society's acceptance of government recommendations, and perhaps catalyze safer behaviors or better governmental responses.¹⁴⁸ These were referenced by over 2,000 scholarly papers as of September 2021.

The application uses similar data to the traffic speed application of [Section 6.1](#), though mobility reporting needs to solve much harder privacy issues. After all, its objective is to report on travel patterns, but to do so without divulging anything that could be used to infer private information about any individual or to exacerbate societal divides.

In addition to Google, other organizations introduced tools that showed changes in mobility. Apple introduced a tool based on counting the number of requests made to Apple Maps for directions, stratified by mode of transportation. Facebook provided mobility data based on the number of geographic tiles an individual moved to, relative to a baseline. Other datasets were also used as a proxy for mobility.¹⁴⁹

Vaccine distribution optimization involves balancing a truly wide variety of competing objectives against the likely operational success of achieving rapid vaccine uptake. Objectives might include minimizing mortality, supporting childhood education or economic activity, ensuring caregivers stay safe and willing to work, demonstrating fairness across multiple subgroups, being politically expedient, and many more.

Models must take into account the likelihood of supply or distribution constraints (e.g., refrigeration), the predilection of subpopulations to accept vaccines, the likelihood that vaccines prevent disease transmission, and even the effects of influencers – who might not themselves be a priority but might positively influence others. There are many papers evaluating different strategies, for example, this one by Bubar et al.¹⁵⁰ Modeling approaches for reducing vaccine hesitancy would seem to be particularly difficult.

Identifying Disease Outbreaks using crowd-sourced data has potential value. However, we defer this discussion to [Section 11.3](#) on Reproducibility Challenges, to allow us to focus on the problems created by this application's opaque nature.

Disease Diagnosis represents an opportunity to use large scale training data and machine learning to provide new diagnostics. While there have been specific tests for diseases since at least the late 1800's when Gram staining started using stains to classify bacteria, it is ever more possible to create new classifiers using neural networks on new forms of sensor data. Data diagnostic tests of many varieties, including X-rays, MRIs, and multispectral cameras, can then be classified to do or aid diagnoses.¹⁵¹ In fact, there are now published reports indicating that some techniques are approaching human capabilities.^{152,153}

Privacy and security issues can be minimized by anonymizing training data and protecting patient imagery and diagnoses the same way as healthcare data is protected. There is little likelihood of abuse, but resilience is very challenging, as errors are very problematic. False negatives (or underdiagnosis) result in untreated disease and false positives (or overdiagnosis) cause patient anxiety, financial costs, and potentially unnecessary treatment. See [Section 11.4.4](#) for more on false positives.

Reproducibility of results is certainly needed and seemingly feasible. Explanation and Causality would be very beneficial for acceptance by both medical professionals and patients. Unfortunately, achieving these is difficult, particularly if the primary technique is machine-learned image classification.

While the objective appears clear, its complexity relates to the toleration and distribution of errors. While human doctors are imperfect, data science approaches must nearly always make the right call. Society at large and its legal frameworks are likely to hold automated systems to a higher than human standard.

Medical regulations, as well as the liability and ethical considerations relating to errors and the associated financial risks, make the ELSI element rife with complexity.

Genome-wide Association Studies, according to Francis Collins, the former long-serving leader of the US National Institutes of Health, are defined in this way: “A **genome-wide association study (GWAS)** is an approach used in genetics research to *associate* specific genetic variations with particular diseases. The method involves scanning the genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease. Once such genetic markers are identified, they can be used to understand how genes contribute to the disease and develop better prevention and treatment strategies.”^{154,155} For example, GWAS has been used to show an association between certain variants located in the FTO gene and an increase in the energy storing white adipocytes (fat cells) that contribute to obesity.¹⁵⁶

Strictly speaking, GWAS refers to the gathering of genomic mutation data and associating that with a label of interest (e.g., disease state). However, a typical published GWAS study will use not only these data as the basis for a scientific result, but augment them with other qualitative

and quantitative research. This includes the stratification of the population and researchers' domain expertise in order to suggest not only correlations, but ideally mechanistic or causal associations. This is both to reduce the risk of time-wasting, expensive, spurious results and to speed the translation of positive results to treatments.

More generally, diseases may have many contributing causes (genetic predisposition and specific exposures and patient activity over a long period) making the underlying analysis even more challenging. Causal sequences may be very long, with some stimulus A influencing B1 and B2 in the same way; B2 influencing C; and C influencing D (say mortality). By just looking at correlations, it would be easy to conclude that if B1 were somehow controlled, D might also be controlled, but this would likely not be true. B1 is not in the causal chain, and also is a **confounder**, a non-causal correlate only associated with disease. [Section 9.1](#) and particularly [Section 11.2](#) have further discussions on causality.

Against the data rubric element, GWAS requires genomic and phenotypic data for sure. It may also need to contain other information about individuals such as age or race. They may also need historical information covering diet, exercise, environmental risks, stress levels, and communicable disease exposure, as these can trigger gene expression. There may be strong reasons for the data to represent a complete cross-section of the society being studied. Health data is often imprecise, inaccurate, incomparable across health centers or populations, and is subject to many regulatory protections. All of this makes its use difficult.

Even if all the data were available, a GWAS study might require an exceedingly complex model. This is due to the possibility of delayed impacts (e.g., hereditary, late-onset Parkinson or Tay-Sachs disease), complex causal pathways, and the previously mentioned risks related to confounders.

Relating to the dependability rubric element, health-related data science applications require laser focus on minimizing the risk of public exposure of private data. They must use the anonymization and encryption techniques described later in this book. In the case of genetic data, exposure not only affects individuals, but may also adversely affect their relatives.

GWAS results almost always trigger much additional work to pin down causality and find therapeutic agents, so great care in engineering and statistics must be taken to minimize the risks of errors. False positives are particularly prevalent. A positive association, if not carefully promulgated, can result in useless or even harmful effects. However, systems may not need to pay too much attention to abuse unless there is significant crowdsourcing of information.

At a minimum, systems need to show their evidence for associating particular genomes, and perhaps other factors, with disease. It is impossible for a system to decree “trust me.” The biomedical sciences strongly value peer review, so GWAS studies would be under great pressure to publish methods, the data, and the detailed semantic understanding needed for its

use. However, this is always difficult given the underlying data's ownership and privacy issues and the complexity of the analysis.

GWAS has reasonably clear high-level objectives, though there may be ambiguity in seeing the right threshold of association versus complete explanations to minimize wasted downstream efforts.

There are laws, some with significant financial and other penalties, governing how research data is used. Others govern how human research subjects are both informed of risks and have to consent to participation. There is significant risk to researchers, their institutions, and to human participants should problems occur. The Belmont Principles address many of the relevant ethical issues.

Understanding the cause of a disease represents a tremendous opportunity for data science. It has the ability to aggregate information on disease incidence and on a growing number of underlying, potentially causative factors including, though certainly not limited to, genetic information.

Table 6.3 Medicine and Public Health Applications of Data Science

Medicine and public health applications	Tractable data	Feasible technical approach	Dependability	Understandability	Clear objectives	Toleration of failures	ELSI
Mobility reporting by subregion during quarantine	✓	✓	Tricky privacy	✓	✓	✓	Perhaps, ethics
Vaccine distribution optimization - when limited supply	✓	Plausible ✓	✓	"Why" is needed	Numerous, conflicting	✓	Ethics
Identify disease outbreak from aggregated user inputs	✓	Plausible ✓	Abuse, resilience, privacy	Explanation, reproducibility	✓	✓	Perhaps, ethics
Disease diagnosis	Training data difficult to obtain	✓ for some diseases	resilience	Reproducibility, explanation, possibly causality	Agreeing on error rates	Wrong diagnoses very harmful	Legal, risk, ethics
Genome-wide association study	Difficult to obtain	Complicated by confounders, complexity	Privacy, security	Reproducibility, explanation, possibly causality	Agreeing on error rates	✓	Ethics
Understanding cause of a disease	Difficult to obtain	Complex	Privacy, security	Reproducibility, explanation, possibly causality	Agreeing on error rates	✓	Ethics

However, gathering the needed breadth of consistent and comparable data faces considerable challenges. Truly measuring and recording all the potential disease-causing factors would have to deal with extreme privacy and security issues. Measures already in place to protect such data add significant complexity to medical research data science applications,¹⁵⁷ and even more measures might be needed. Abuse is unlikely to be an issue, but resilience is important. The technical approach may be very difficult due to the breadth of the problem. Among other things, many factors (e.g., environmental ones) may take years or decades to cause disease.

Objectives are clear. Failures are acceptable if they are not too likely or costly and if results can be independently confirmed.

In the category of understandability, scientists need reproducibility to validate results. Beyond our previous medical examples' need for explanation, this application is (by definition) trying to show causality to enable development of good public health measures, prophylactics, or cures. For example, for many years, correlation between coffee drinkers and cancer implicated coffee as a carcinogen. But researchers eventually concluded that coffee consumption was correlated with cigarette smoking, and smoking turned out to be the smoking gun. See [Section 11.2](#) for more on causality. Applying data science to understand the causes of disease is challenging across all rubric elements.

6.4 Science-Oriented Applications of Data Science

As discussed in [Section 2.1](#) and [Section 4.4](#), data science can be of enormous value to science. At a minimum, it can provide the intuition for creating more and better hypotheses. It can also generate new knowledge and contribute to understanding causality. In this section, we discuss two more examples of using data science in the scientific realm. (See [Table 6.4](#).)

Determining the historical temperature of the universe is a scientific application that has confirmed the universe has warmed tenfold by some metrics.^{158,159} Scientists have determined this by amassing data from the Sloan Digital Sky Survey (SDSS) and the European Space Agency's Planck Infrared Astronomical Satellite. As background, every day the SDSS accumulates 200 gigabytes of data, all of which is eventually made public, so there are no privacy or security issues. More than 5,800 scientific papers using this data have been published.

In this case, scientists gathered two million spectroscopic redshift references from the SDSS (measuring the speed at which celestial objects are moving) and combined these with sky intensity maps (which indicate temperature). Since objects moving faster are further away, and their measurements are from longer ago, the scientists thus had a technical approach for measuring the change in temperature over time. In this instance, the scientists' deep theoretical understanding lets them apply big data and get their desired results. There was no problem with

all citizens' transaction data, security attacks that cause economic warfare, and the resilience problem of the “Oh no, we forgot to include that effect!” as well as many others.

Opaque systems that are neither reproducible nor comprehensible are probably unacceptable. For example, economic policy makers would find it very hard to act on economic predictions to change interest or tax rates without first understanding them. While it is easy to find correlates with economic growth, causality, especially over the long-term, is hard to show. Forecasting would seem to have clear objectives, but there would be difficulty in determining the requisite granularity and necessary accuracy. While forecasting will always be imprecise, some failures would have catastrophic effects affecting entire nations. There is no end to the legal and ethical risk.

We end this section on financial services by noting its data science applications are continually evolving with the growth in data, computational capability, and machine learning. The final example was more of a grand challenge research problem pushed to the limit. But there is no doubt that the increasing amount of data coming from the economy's digitization will lead to significant changes in economic forecasting.

6.6 Social, Political, and Governmental Applications of Data Science

Governments provide diverse and critical services to vast numbers of people. Operating at scale, there is great opportunity to sense opinions, needs, successes, and outcomes, and to optimize results. Possible uses range from political campaigns to operations of state agencies and include the domains of economics, health, education, and more. (See [Table 6.6.](#))

Targeting in political campaigns refers to the interest that political candidates have in knowing what positions appeal to voters, which communication channels to use, and even what exact words to use to disseminate their positions. Furthermore, candidates do not want to waste resources either in areas they are sure to win or which are hopeless for them. In systems where candidates need to fundraise, data science is critical for helping candidates focus their messages as well as the target audiences to raise the most money. For better or for worse, big data allows candidates to truly slice and dice populations and send out targeted messages to best appeal to fine-grained constituencies.¹⁶⁵

Significant amounts of data are already available. In the US, data begins with voter registries from which campaigns can get voting rolls (including party registration) and historical data on when individuals have voted, though NOT for whom they voted. Political parties and both not-for-profit and for-profit entities augment this data with additional individual and aggregate district data. For example, campaigns commission polls to learn voter positions and interests.

The application space is broad with many applicable clustering and prediction techniques. For example, campaigns predict the likelihood of sympathetic voters within a small region and then target voter registration drives to those regions with mostly sympathetic voters. There are the

usual privacy and security issues with some personal data, though campaigns can buy recommendations from others and possibly avoid directly holding too much data. Abuse is increasingly likely, even by nation state actors seeking only to create chaos.

Given western democracies' extreme focus on elections, election-related data science is a fertile area for seeing how objective functions vary:

- Candidates may have different goals at different times. During a primary, they need to maximize votes from members of their own party. During the general election, they need to maximize votes across a more politically diverse electorate. Data scientists on a campaign may suggest a candidate's approach and messaging vary accordingly.
- An individual vote's value may differ depending on the voting district. A vote in a contested district is far more important than one from a safe district. The fluidity in changing voter perceptions makes this challenging.
- Fundraising may try to either maximize total funds raised, or perhaps demonstrate a broad-based groundswell of appeal by receiving many small donations.

Table 6.6 Government Service and Political Applications of Data Science

Government service & political applications	Tractable data	Feasible technical approach	Dependability	Understandability	Clear objectives	Toleration of failures	ELSI
Targeting in political campaigns	✓	✓	Privacy, security, abuse	✓	Competing objectives	✓	Legal, ethics
Detect maintenance needs	Insufficient sensor coverage	✓	Security, resilience	✓	Complex due to prioritization	Certain failures intolerable	Legal, risk, ethics
Personalized reading tutor	✓	✓	Privacy, security, abuse, resilience	Explanation	✓	✓	Legal, risk, ethics
Criminal sentencing and parole decision-making	✓ but may be hard to assemble	✓	resilience	Explanation, reproducibility	Conflicting	Individual freedom & societal welfare	Legal, risk, ethics

Political campaigns may well accept opaque systems, and certain failures are both likely and acceptable, given the application's inherent uncertainty. There are legal regulations on campaign operations, but the biggest ELSI challenges are ethical. Candidates need to balance their own views on what is "right" with increasingly explicit recommendations on what positions the electorate wants them to take. Data science may also tell a candidate that one part of the electorate wants them to take position A, while another part wants the opposite position B. This leaves a candidate to decide whether to take no stand, to choose one stand, or possibly to take

different stands with different audiences. While candidates have always had to make such complex decisions, data science quantifies them and makes them explicit.

We briefly cover the next two topics:

Detecting maintenance needs is a considerably more mundane application than targeting in political campaigns. Data science can make it possible to provide early warnings of potential failures based on data from vibration, corrosion, and other failure precursor instrumentation or from crowdsourcing from cameras or vibration sensors on vehicles.¹⁶⁶ These warnings are important because it's both safer and more cost-effective to identify and fix problems prior to failure than after.

Depending on the specific application, there are a variety of models to use this data, taking into account structural, failure, and risk properties. Remember though, there is always the challenge of balancing false positives with false negatives. Also, bad actors might try to interfere with a systems operation to cause societal harm. Maintenance officials must understand this application's objectives and coverage to avoid complacency leading to undetected errors and catastrophic failures.

As our next example, we turn to the domain of **education**. While there are many possible examples, ranging from school budgeting to student/class scheduling to personalized learning, we focus on the latter.

For subjects taught to most students, such as reading and writing, there are vast amounts of pupil data to work with, and it might be possible to create customized education that better motivates students and is more effective. In the 1980's, researchers such as Benjamin Bloom showed that students learn best with an approach known as **mastery learning** – studying a subject at their own pace until mastery is reached.¹⁶⁷ Having an individual tutor to guide each student has been prohibitively expensive, but systems that gather individualized data may make it possible.

Personalized reading tutors are a good place to start. Already, there are online reading tutors for early childhood education that provide compelling material and immediate student feedback based on individual interests and level of mastery. Online reading education could be extended to additional populations, as data science techniques could categorize vast amounts of reading material. Systems could learn from a large student population's signals, such as engagement or comprehension. Their prediction abilities could reduce boredom from repeating known materials or the confusion caused by excessively fast-paced instruction.

Student data collection is a serious concern from a privacy and security perspective. However, resilience would seem the biggest dependability problem if optimization techniques can fail. As in healthcare, widespread adoption of educational innovations may require proof of success in small, controlled trials. This makes explanation and reproducibility of high importance.

Reading education's exact objectives are often unclear and vary by region and over time. There are also debates on how best to teach the subject. This makes it hard to create applications that can be deployed widely, which reduces both available funding and data. Failures are harmful, and education involves significant ethical issues. Applying the Belmont beneficence principle, we must carefully balance the benefit and risk to a student's educational progress when replacing a known approach with an automated tutor. Educational solutions must benefit many students, so balancing benefits is challenging.

Criminal sentencing and parole decision-making is our final example. Data science applications in this area might provide judges with decision aids for use during pre-trial detention, criminal sentencing, and parole assignment. These tools could enable judges to make decisions more consistently and lessen the variability of human judgment. They could better ensure consistency by a single judge over the course of each day or over an entire judicial tenure. Better yet, they could ensure some degree of consistency across judges in the same or different jurisdictions. For example, tools could mitigate “serial position effects,” the widely studied biases that may influence judicial decisions based on when a case is scheduled.¹⁶⁸ Ideally, individuals with similar criminal histories who commit the same crime in similar circumstances would be treated similarly, which is called **algorithmic fairness**.¹⁶⁹

Today, US courts are using such tools, though some researchers have shown that the risk assessment tools are statistically biased.¹⁷⁰ However, other researchers have shown that using data-driven decision aids can reduce bias and increase accuracy of pre-trial decisions.¹⁷¹ There is more detail on this in [Section 12.3](#) on Fairness.

In principle, the needed data is available. In practice though, different jurisdictions may collect different types of data and differently code/format what they have. Data can be incomplete and noisy, and data collected for the same individual can be inconsistent. Moreover, many criminal justice systems still use manual processes, so much data may still be only on paper. Data must be balanced in the sense that it will not lead to unfair treatments for any population. Once sufficient data is available and processed to be comparable, we can apply straightforward statistical models, from logistic regression to deep learning.

An algorithmic decision-making tool's failure can have disastrous and potentially long-term consequences. Choosing to develop and deploy such tools demands consideration of the ethics and societal risks, not just the statistical challenge. Denying bail or parole to a low-risk individual can have mental and economic consequences for the person and his/her family. Granting bail or parole to a high-risk individual could lead to another crime. We will refer back to this example in [Chapter 7](#), and also have more to say on it in the context of fairness in [Section 12.3](#).

Chapter 7. A Principlist Approach to Ethical Considerations

In this chapter, we describe how the ethical framework we introduced in [Chapter 3](#), based on the Belmont principles of respect for persons, beneficence, and justice, can be applied in the context of data science. This principlist approach to ethics attempts to provide a shared analytic framework and vocabulary to help communities and teams resolve difficult questions. Principles are most useful when broad enough to be comprehensive and capture, rather than ignore, the tensions that make questions of “right” and “wrong” so difficult.

As the Belmont report states: “the objective is to provide an analytical framework that will guide the resolution of ethical problems,” however, “these principles cannot always be applied so as to resolve beyond dispute particular ethical problems.” That is, our goal is not to provide a universal yes-or-no algorithm for ethics. Rather, it is to guide ethical decision-making so it provides practitioners and stakeholders a shared understanding of a decision-making process and logic.

To illustrate, we chose five of [Chapter 6](#)’s use cases to explain with respect to the Belmont principles: criminal sentencing, newsfeed recommendation, vaccine distribution, mobility reporting, and insurance underwriting. The three principles are not ranked in importance, and each example may not have concerns related to all three. As in the context of Belmont’s original deliberations, “beneficence” includes not only individual harms and benefits but also those of society at large.

The observation that a data science application may not satisfy each Belmont principle does not necessarily mean we advocate discontinuing the application. Vaccine mandates, for example, would prioritize public good over individual autonomy. This complexity of this balance is reflected in US law: “Since *Jacobson v Massachusetts* (1905), the judiciary has consistently upheld vaccination mandates,”¹⁷² while there are COVID-19 vaccination decisions that show nuance. The original Belmont Report, similarly notes of principles: “at times they come into conflict” and “cannot always be applied so as to resolve beyond dispute particular ethical problems.” Instead, they are meant to “provide an analytical framework that will guide the resolution of ethical problems.”

Here are applications of the Belmont framework’s three principles to some of [Chapter 6](#)’s examples:

Criminal Sentencing, and Parole Decision-making

As discussed in [Section 6.6](#), algorithms for criminal pre-trial, sentencing, and parole decisions are fraught with ethical challenges:

- **Respect for Persons:** All stakeholders’ autonomy would be challenged if such algorithms were opaque: defending attorneys lack understanding of how decisions are made now, and defendants who may become incarcerated lack understanding of how

their actions may be scored in such an automated decision system. The transparency paradox discussed in [Chapter 3](#) thwarts complete information. Instead of over-explaining technical subjects or under-explaining (which can lead to deceptive or unfair practices), proponents of increased use of automated decision systems can adopt a “tiered” approach.¹⁷³ This would involve explaining the basic concepts in plain language while providing extra detail to those who want increased transparency. Another concern is the likely event that algorithmic approaches do not provide interpretability. The answer to “why” a decision is rendered is both important and perhaps not answerable.

- **Beneficence:** The claimed benefits of using such algorithms, such as efficiency and uniformity, must be explicitly evaluated with respect to their impact on many parties and performance versus existing human approaches. These include defendants, potential parolees and other stakeholders, including the judiciary, the criminal justice system, and society writ large. For example, algorithms may be trained to minimize expected errors against a test set (e.g., prior decisions by human judges). However, this training may not minimize the total number of future crimes, the total expense of the justice system, or other more beneficial, societal goals.
- **Justice:** Biases resulting from deploying algorithms for criminal sentencing and parole decision support – e.g., different model outputs for demographically different defendants with similar criminal data – is the subject of ample research and journalistic inquiry. Many of these inquiries argue, via statistics as well as case studies and interviews, that some are benefiting relative to others. Also, perfectly accurate models trained on biased data perfectly reproduce these biases. As discussed in [Section 6.6](#), the use of data science in algorithmic sentencing and parole decisions catalyzed research discussions on the multiple ways to quantify fairness.¹⁷⁴ We discuss this in more detail in [Section 12.3](#).

In short, this example includes a variety of data science and societal challenges. The ethical considerations are clear, with reduced accountability, serious risk of harm to individuals and society, and opportunity for amplifying unfairness and injustice as applied to individuals. However, there could be enhanced judicial uniformity and possibly other benefits. Aside from ethical considerations, different stakeholders (defendants, lawmakers, members of the judiciary) might have extremely different notions of what objective an algorithm is to optimize.

News Feed Recommendations

As discussed briefly in [Section 4.3](#) and [Section 6.2](#), news feed recommendations are considerably more challenging than music recommendations. News corpuses vary in size and quality, as well as the motivations of those who add to them. News also has a much greater impact on individuals and society.

- **Respect for Persons:** Informed consent is challenged when news recommendation algorithms are opaque. The transparency paradox is exacerbated given their complexity, which prevents even their designers from fully understanding the resulting technical systems. A widely discussed example of informed consent was Facebook's 2012 “Emotional Contagion” automated news feed experiment, published in *PNAS* in 2014.¹⁷⁵

In this experiment, Facebook users' news feeds were experimentally manipulated to amplify posts of positive or negative sentiment. Subsequent posts by those users were then used to quantify whether the algorithmic changes caused subsequent increased or decreased happiness. [Section 12.4](#) has more detail on this.

- **Beneficence:** These algorithms can indeed inform and delight when delivering content that optimizes engagement through joy and surprise. However they can also create filter bubbles ([Section 12.4.2](#)) or amplify fear or hate, leading to a radicalizing “rabbit hole.”¹⁷⁶ The complexity of content ranking algorithms, and the unpredictability of their impact on users' well-being, is driving companies to spend more time investigating unanticipated ways content recommendation may lead to harm. Of course, they also have to develop ways to mitigate these effects.¹⁷⁷ We need long-term studies, where users' behavior is observed for several weeks or months, for developers to know the impact of a news feed's content recommendation algorithm.
- **Justice:** Without question, algorithms affect different societal groups in different ways, leading to many considerations of fairness. Some impacts are benign (perhaps, a propensity for a subgroup to get sports entertainment recommendations), while others reinforce societal problems.

We have chosen this example because there has been increasing societal reliance on algorithmic news feeds. This has led to increased public scrutiny. The complexity of the algorithms prevents designers and readers alike from understanding what content and worldview is being amplified. As to harms, the attribution of benefit and risk to these algorithms is debated daily in the press, by researchers, by companies doing news recommendations (including their own researchers), and by lawmakers and regulatory agencies.

Vaccine Distribution Optimization

Distributing a vaccine with supply, logistical, and vaccine hesitancy constraints is a truly complex problem, as briefly discussed in [Section 6.3](#).

- **Respect for Persons:** This is a concern as societies increasingly pressure and even mandate that vaccine skeptics be vaccinated. If vaccines were outright forced, vaccination would violate the principle of informed consent. Society-level rationing also reduces individual autonomy.
- **Beneficence:** Beneficence's role in vaccine distribution, particularly since both the supply and distribution channels may be limited, is extremely difficult to know in advance. The coupling of optimization and health policies requires complex trade-offs, e.g., between supporting opening schools or reducing disease in prisons or assisted living facilities. Here, a commitment to beneficence includes adjusting distribution policy as supplies and health policies change and as the effects of a distribution policy become evident.
- **Justice:** At the individual distributor level, whether state or private, appointment booking mechanisms have varied usability and interface complexity. The technology divide may contribute to a “vaccine divide” among those with and without the technology access to

secure vaccine appointments. The digital divide may correlate with demographic divides and could result in unfair outcomes.

We have included this example because of its widespread importance and serious ethical complexity. Even though data-driven models are needed to optimize vaccine distribution, ethics are of paramount importance in this example, and balancing objectives is particularly complex.

Mobility Reporting

The Google team that showed aggregate regional movement trends was cognizant of the need to preserve individual location privacy and effectively used differential privacy techniques to protect that data. (See [Section 10.1](#) for more on differential privacy.) However, even if individual privacy is preserved, using geographical data gives rise to ethical issues worthy of consideration.

- **Respect for Persons:** Neither any individual nor Google would have contemplated in advance that anonymized location data would be used for this purpose. However, Google's anonymization policy is explicitly written to allow Google wide latitude in the use of anonymized data.¹⁷⁸ On the one hand, users might be surprised, given how few users read the policy. On the other hand, opt-in would have greatly reduced the likely effectiveness of this application.
- **Beneficence:** Correlations between who is mobile and how disease progresses offer societal benefit for informing health policy. However, there may be implications to individuals, even if identities cannot be inferred from published mobility data. For example, mobility data could be correlated with widely available demographic data and reinforce stereotypes or create societal divides. This same effect could occur in many other applications that aggregate anonymized individual data.
- **Justice:** Since mobility data is often gathered via smart phones, it risks being skewed towards those users. This requires a careful analysis of the policy's efficacy and impact based on such data. Sampling bias is discussed more in multiple places in [Part III](#).

We used this example to show there are ethical concerns beyond the most obvious one, which is privacy. It shows the subtle issues a design team must navigate even when the primary goal is to produce an information system intended to benefit public health and policy.

Underwriting/Pricing of Property/Casualty Insurance

Ascertaining risk to enable better selection pricing of insurance policies is a traditional application of data, and it is significant given how important it is to people. As discussed in [Section 6.5](#), data can be applied to many aspects of the problem space.

- **Respect for Persons:** Availability and pricing of insurance should be based on the specific risks of an individual application, not exogenous factors which may not be related. Opaque algorithms which set loan policies and insurance rates for individuals

challenge the concept of informed consent, as neither applicants nor insurance regulators may be able to determine the rationale for an underwriting decision. Some who particularly need insurance (e.g., those in fragile economic circumstances) have diminished autonomy, possibly meriting increased protections against hard-to-understand or deceptive terms and conditions.

- **Beneficence:** Increased use of more personal data can itself affect the risk-taking behavior of individuals or groups. For example, high penetration of insurance and low reimbursements could drive practitioners out of a medical sub-field, causing societal harm. Insurance's increased use of personal health information could also motivate individuals to avoid useful health diagnostic tests, thus causing societal harm due to a lack of preventative testing or even increased disease transmission.
- **Justice:** Such algorithms can reinforce societal bias, e.g., if they are accurately trained to reproduce biased human insurance underwriting decisions, they would constitute a form of "digital redlining". Data could facilitate the creation of new, finer-grained risk pools (for example, assigning people with genetic predisposition to disease) thereby increasing differentials in insurance costs. This unequally distributed harm illustrates the Belmont Report's multiple meanings of justice, "in the sense of 'fairness in distribution' or 'what is deserved,'" to quote the original.

We include this example to illustrate how a mechanism that predates digital computation can, by including far more data and complex algorithms, risk amplification of already present harms and injustice.

To close, we refer back to Gottenbarn and Wolfe, who state, "...every decision requires us to identify a broader range of stakeholders and consider how to satisfy our obligations to them. A primary function of the Code is to help computing professionals identify potential impacts and promote positive outcomes in their systems."⁹¹ While they are talking about computing and the ACM Code, their quote is also consistent with our principlist approach to ethics. It underlies our view, as demonstrated with this section's five examples, that ethics must be considered as many types of decisions are made. We believe that doing such analyses against a set of principles, like the Belmont Principles, (1) reminds data scientists to think about difficult challenges, (2) acts as a check on significant errors, and (3) motivates practical improvements.

Recap of Part II – Transitioning from Examples and Learnings to Challenges

Reaching the end of [Part II](#), we hope our detailed discussions of the initial six examples and the somewhat more cursory discussions of twenty-six more have been enlightening:

- While we have not surveyed all possible applications, we hope we demonstrated data science's valuable potential in domains ranging from research to entertainment to medicine to commerce to finance to government, and more. In many cases, we also explained a bit of their “how-to” and demystified how many applications operate.
- We showed that data science is often applied in different ways to multiple sub-problems of an application. For example, in a video streaming application, we illustrated copyright identification, video recommendation, search, and advertising. Video streaming could also benefit from additional applications of data science, such as for closed captioning, language translation, summarization, and more.
- By our repeated referral to the Analysis Rubric, we illustrated seven important considerations in applying data science to a problem. We believe that examining an application in detail against the rubric elements teases out its design's hard parts.
- Whether contemplating a new data science application or evaluating an existing one, careful consideration of data science's unique and complex aspects is a necessity. While the rubric elements are of necessity listed in a particular order, we fully recognize that the application of the rubric may be done in a more bottom-up or top-down way, depending on the application.
- Our analysis shows a few applications to be straightforward, many to be challenging, and some to be very hard or perhaps presently impossible. When difficulties occur, they are usually because:
 - Data capture of sufficient quality and scale is impossible for some reason.
 - Existing technical approaches are insufficient.
 - The cost of achieving dependability is too high.
 - Opaque approaches in either sense of the term are insufficient; furthermore, there may be a particular need to prove causality.
 - Objectives are in dispute.
 - Failures are essentially intolerable.
 - Insurmountable difficulties arise from the ELSI criteria.

Finally, we resumed our ethics thread. We explored how the principles set forth in [Chapter 3](#) can guide us when applying data science to gnarly problems.

All of this sets the stage for [Part III](#), a considerably deeper discussion of data science challenges.

Alfred Z. Spector: Post-Modern Prometheus

I recently came upon an article arguing that data science had hit “Peak Metaphor.”^{F31} This is no surprise given the contemporary importance of the field, and the need for many of us to find an apt turn of phrase to summarize some point of view. On my mind is this message:

Data Science: Powerful Technology. Great and Increasing Value. Handle with Care.

I realize this is hardly new. It’s been repeated countless times, perhaps beginning with the Greek Myth of Prometheus, who delivered fire, a technology of unarguable value and lasting impact. However, Prometheus suffered acutely for pilfering the gods’ trade secret, and we are still dealing with fire’s disadvantages some twenty-eight hundred years after the Greek poet Hesiod’s writing. So:

- **Fire:** It’s easy to start, diversely useful, but it’s risky and has harmful side effects. The harms have been relatively evolutionary and despite repeated catastrophes, we’ve found ways to deal with them. Concerns over CO₂ will curtail bulk use, but otherwise fire will remain.
- **Data Science:** It’s ever easier to gather data and create great insights or conclusions. Like fire, it is astoundingly useful. It’s also a risky endeavor with subtle problems that have harmful effects. It may even be with us for twenty-eight hundred years more.

The really big question we don’t answer in this book is whether data science (and the overlapping field of artificial intelligence) will have a gradually increasing impact or whether it will catalyze fundamentally extreme and discontinuous change. Addressing this topic from the vantage point of AI, Bostrum wrote in 2014 that the time is near when we have “superintelligence,” which he defines as “any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.”⁴¹⁵ Kissinger et al. wrote in 2021 about AI causing “a new epoch” and an “alteration of human identity and the human experience of reality at levels not experienced since the dawn of the modern age.”³⁶

Nearer term, pragmatists like me will be focused on what to learn and do now. While I strongly endorse deep thinking about the longer-term issues, our work is mostly to solve the challenges we perceive today. I see these as dividing into ones that are primarily technical and others that relate to application or use:

Technical concerns: Research already underway will solve many of data science’s technical challenges though some breakthroughs are needed: In particular, we can’t yet replicate the type of transfer learning that enables humans to quickly learn from books. We also don’t have a good handle on how to combine common sense knowledge with machine-learned models. Modeling

³¹Sondregger used this phrase, perhaps sarcastically, to note the recent spate of metaphors for data science.⁴¹⁴

applications with concept drift seems almost impossible, particularly if there is sudden change. I also don't see how we will solve the interpretability problems discussed in [Section 11.1](#). Finally, resilience in the face of adversarial attacks seems like it will be a long-term challenge.

Application concerns: As we have said, when data science is used incorrectly, there are negative consequences.

- Data science exacerbates the problem of misaligned incentives. It is too easy to tune systems to meet a narrow goal (optimizing an overly simplistic objective) that is not to the long-term benefit of individuals, organizations, or society. The problems of incentives may be greater if they are created by a government or a small number of large organizations, potentially reducing a society's pluralistic voices.
- Data science methods coupled with the sheer quantity of fine-grained data can lead to compelling but false insights. In particular, our book warns against confusing correlation and causation. Creators and consumers of information should practice the greatest care in communicating and understanding data science results, and should pay particular heed to the list of cognitive biases co-author Peter has assembled in [Section 11.4](#).
- Data science makes it harder to agree on reasonable, but imperfect, solutions to difficult problems. Even though it may be possible to quantify mathematical trade-offs between different solutions, this analysis may only serve to highlight the inevitable limitations of each and prevent pragmatic action. Many systems are zero sum games, and data science can be used to highlight each loss. It's hard to remember "The Great is the Enemy of the Good," when confronted with quantified objections.
- Data science solutions are often insufficiently tolerant of errors or abuse. Some errors naturally occur because of the probabilistic nature of data science solutions or the innate difficulty of solving certain types of problems. Security vulnerabilities also play a big role, and they are extremely hard to prevent. Risks are heightened because many of us, myself included, were late in realizing that nation-states would engage in attacks on non-military applications. These challenges are not solely technical, because they often arise because of interactions between people and computers. We thus need to think carefully about where we are applying data science.

Many of these challenges will require the transdisciplinary efforts of the diverse coalitions we referred to in [Chapter 2](#). We surely need to apply an ethical lens as we make important decisions, and societal norms may also change. As with the control of fire, we will also require sensible laws and regulations, though we must take care to avoid negative regulatory consequences. Solutions will take time, and as with fire (and all good inventions), there will inevitably be residual risks that we learn to live with.

In her 1818 novel *Frankenstein*, Mary Shelley explored the consequences of a powerful and groundbreaking technology – in this case, one that created a living creature. In recognition of the parallel between Frankenstein's delivery of a synthetic life and Prometheus's delivery of fire, she subtitled her book *The Modern Prometheus*. We data scientists are perhaps, collectively, a

“Post-Modern Prometheus,” who can and should strive to minimize the risks of our own fire, for the well-being of ourselves and our societies.

However, we should receive encouragement to pursue our dreams from the words of Percy Shelley, her husband, who also wrote about Prometheus. He concluded his play, *Prometheus Unbound*, in an uplifting manner:

To defy Power, which seems omnipotent;
To love, and bear; to hope till Hope creates
From its own wreck the thing it contemplates;
Neither to change, nor falter, nor repent;
This, like thy glory, Titan, is to be
Good, great and joyous, beautiful and free;
This is alone Life, Joy, Empire, and Victory.⁴¹⁶

Final Thoughts

Data science practitioners have employed enormous effort and creativity to create applications of great value. They have addressed many difficult challenges to deliver results embraced by billions of people every day. Data science has brought increased understanding, economic growth, and new tools and entertainment.

We believe that data science will continue to thrive and extend its reach in important areas such as healthcare, education, climate, transportation and logistics, commerce, sports and games, and economic development, to name but a few. We should proactively encourage and engage in data science, while also addressing its pitfalls. The increasing international competition in data science means nation states will very likely reach the same conclusion.

There are indeed very hard foundational questions underlying data science: How do we deal with missing or differentially sampled data? What does it mean to be fair? How do we distinguish correlation and causation? How do we explain conclusions? Some real-world applications may continue to elude data science solutions, due to the sparsity of data, complexity of the problem, or cleverness of adversaries. We reiterate that an application of data science has not provided a complete solution if it does not meet the breadth of the Analysis Rubric considerations.

Some problems are particularly hard to set proper objectives for, as discussed in [Chapter 12](#). Simple metrics, such as maximizing clicks or counting near term revenue, are unlikely to suffice from either a business or ethical perspective. When data science is asked to provide solutions where people have not agreed on the preferred outcomes, the solutions will not please everyone. Gaining a consensus requires advice from ethicists, governments, economists, political scientists, other experts, and the general public.

Data science is being asked to provide solutions to very difficult problems. For example, it is plainly difficult to optimize complex systems that exhibit non-stationarity and which have adversarial responses, as we discussed in the country-wide economic prediction example of [Section 6.5](#). In recognition of this and the difficulty of establishing consensus objectives, these problems have been called **wicked**, and are acknowledged to be very difficult.⁴¹⁷

Finally, we admit the field's breadth and speed make it hard to keep up with everything. We ourselves are confronted with the rapid changes in application areas, technical approaches, and problems, though we find that this book's frameworks allow us to put these changes in perspective. We are less sure about all of the details, and we know we have probably made errors or provided overly shallow discussions of some topics so that we could cover the full scope of the challenges and opportunities we see. In recognition, we expect to put updates on our book's website, DataScienceInContext.com. We also acknowledge that some of our examples will become stale.

This book has not covered three topics that may have practical implications in the future:

- The application of quantum computing to solve currently intractable problems.
- The wide-spread deployment of capable robots throughout society.
- The development of artificial general intelligence, rather than AI for specific applications.

We close by stressing that data science is important to society – too important to be done poorly. We thus hope this book stimulates more of us – data scientists and humanists, ethicists, social scientists of all types, scientists, politicians, jurists, and more – to study data science's opportunities and challenges and work together to better our world.