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

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.

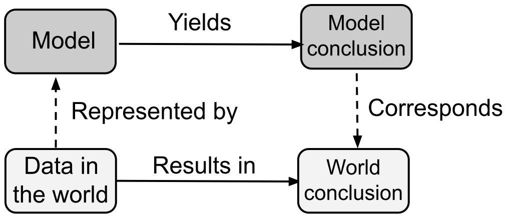
1.1 Definitions

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, Googles' 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.

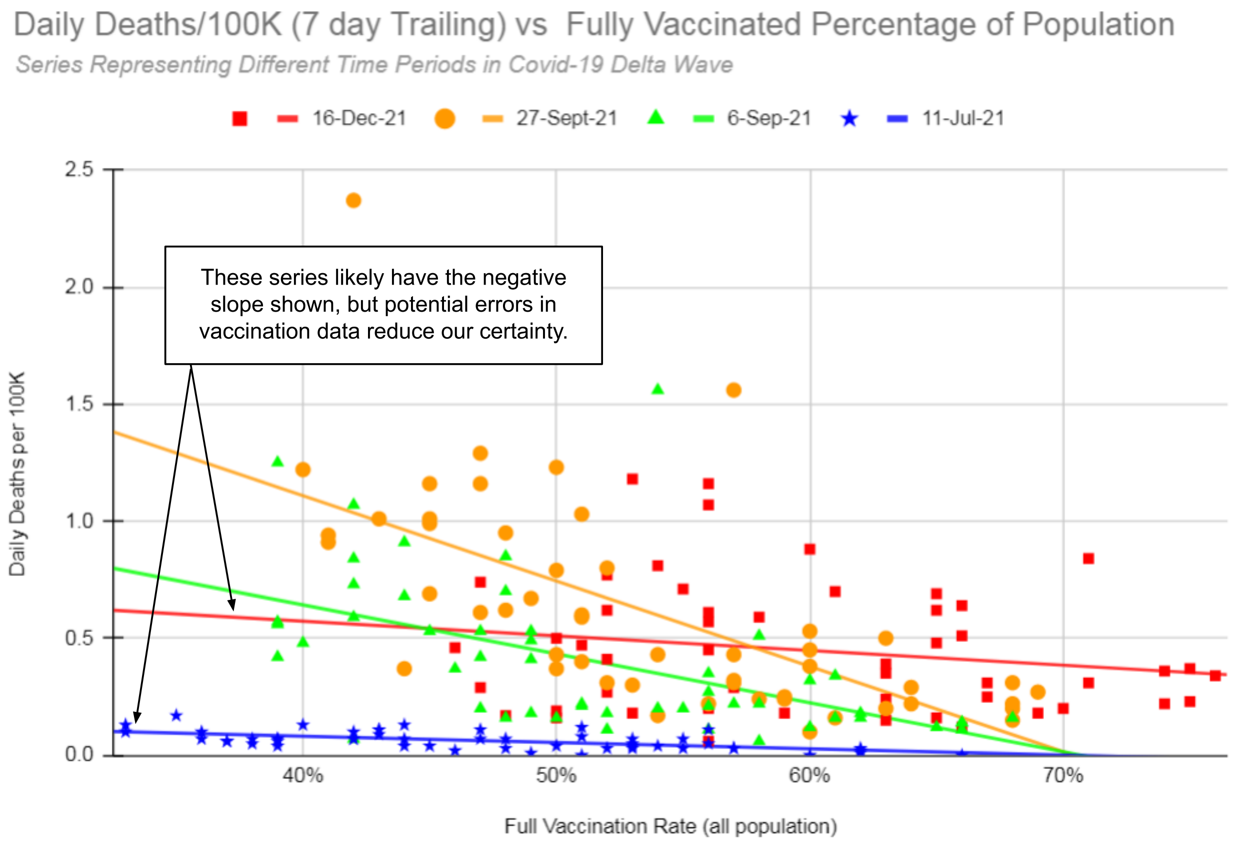
Pioneering data scientist John Tukey said, "The simple graph has brought more information to the data analyst's mind than any other device,"[4] but modern visualization offers many other beautiful and useful ways to gain insight. However, graphs must be scrutinized very carefully for meaning.

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 reporting[5] 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.



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 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.

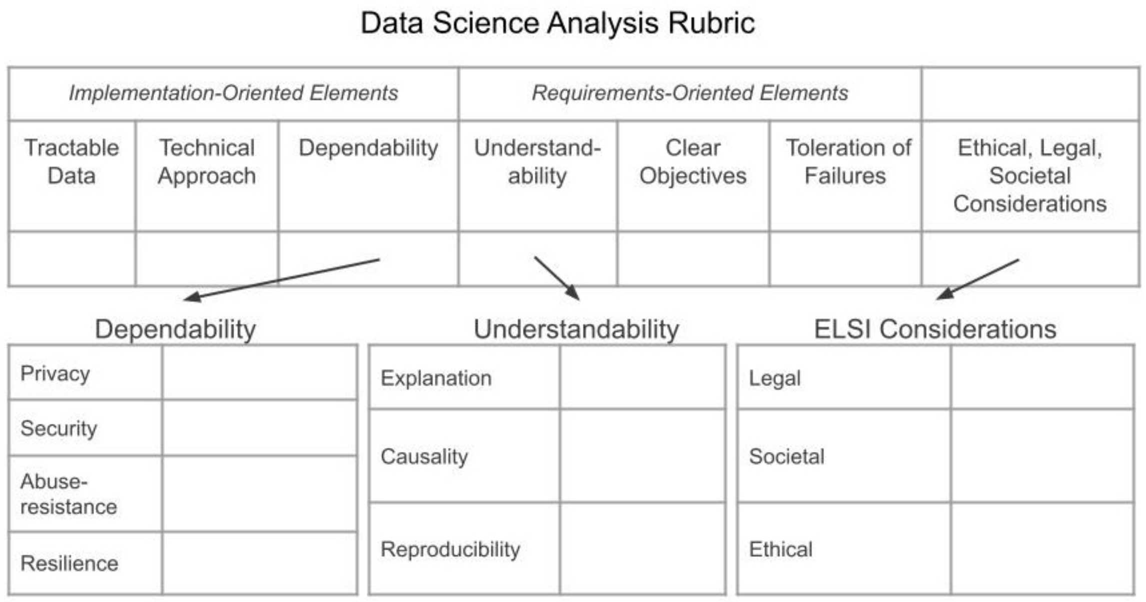
Ethical, Legal, and Societal Implications (ELSI) Element

- **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

Implementation-Oriented Elements			*Requirements-Oriented Elements*			
Tractable Data	Technical Approach	Dependability	Understand-ability	Clear Objectives	Toleration of Failures	Ethical, Legal, Societal Considerations

Dependability

Privacy	
Security	
Abuse-resistance	
Resilience	

Understandability

Explanation	
Causality	
Reproducibility	

ELSI Considerations

Legal	
Societal	
Ethical	

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

The Analysis Rubric is important to this book. It is illustrated in Figure 5-1, which summarizes its considerations in graphic. The next six sections will make the rubric more concrete by demonstrating its application to the six examples of Chapter 4.

5.1. Analyzing Spelling Correction

Spelling correction is a clear example of a really good data science application–as evaluation using the Analysis Rubric shows:

- **Tractable Data.** Anyone can easily collect an appropriate corpus of online text. A company already running a service can easily collect user feedback from spelling suggestions to verify which suggestions are good. There are “only” a few million distinct word tokens in any language, so individual word count data is relatively small. However, multi-word phrase data quickly grows in size–the Google Books Ngram project has a few hundred gigabytes of data for counts of phrases up to 5 words long.

- **A Technical Approach.** Section 4.1 outlined an approach to spelling correction in a search engine using word and phrase frequencies in the search corpus, together with user feedback from accepting or rejecting suggestions. The model is relatively simple, and a basic version takes just a few dozen lines of code.[131]

- **Dependability.** Spelling correction relies mostly on non-private data, so privacy and security are not major issues. However, privacy is always tricky, and a system that learns from an individual or institution should not expose confidential information (such as the spelling of code names) to outsiders. Erroneous corrections may occur, but the cost of a spelling error is low. Some care must be taken to prevent an attacker from spamming the spelling corrector with an incorrect spelling (perhaps to promote their brand name).

- **Understandability.** Users don't really care how spelling correctors work. Spelling correctors also don’t need to understand a spelling error's root cause. Finally, a spelling corrector’s internal operation can be opaque. Neither must its inner workings be understandable nor must its logic and data be published. This is good, because the specific words each individual user types must be kept secret.

- **Clear Objectives**. The clear goal is determining and providing the correct spelling. While a spelling corrector could correct “wheg” to many different words such as “when,” “where,” or “Whig,” the correct spelling is what the user *meant* to type.

- **Toleration of Failures**. While a spelling corrector should almost always do the right thing, almost all users are accepting if it does not correct a word’s spelling or even guesses a word incorrectly, as long as the failure is plausible. However, even a rare failure that “corrects” words to become profane or otherwise objectionable would be unacceptable.

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**	**Depend-ability**	**Under-standability**	**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**	**Depend-ability**	**Under-standability**	**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 20. Concluding Thoughts

We had several goals in writing *Data Science in Context*:

- We wanted to introduce data science as a *coherent field*, while illustrating the need to balance *its opportunities and challenges*.
- We wanted to advise our readers on how to both *apply data science* and to *critically understand* its uses in the world.
- We wanted to emphasize ethical considerations, through both the *ethics framework* and a large collection of ethics-related challenges.
- We wanted to summarize *societal concerns* about data science and make recommendations to address them.

This chapter briefly summarizes these points and concludes with a few lessons we learned while writing this book.

Data Science - A Coherent Field

Our explanation of this field began with a definition: “Data science is the study of extracting value from data – value in the form of **insights** or **conclusions**.” We then made more explicit what we mean by insights and specified six types of conclusions: **Prediction**, **Recommendation**, **Clustering**, **Classification**, **Transformation**, and **Optimization**.

As we described, data science's intellectual origins lie mostly in statistics, operations research, and computing. We find the story of the forces that combined to form data science over the decades prior to the term’s ~2010 breakout to be a compelling one, replete with visionaries, breakthroughs, the march of technology, and economic incentives. We illustrated data science’s broad and growing impact, complex challenges, and powerful future with many examples. We used the term “transdisciplinary” to emphasize its integration of many forms of knowledge, techniques, and modes of thought.

The **Analysis Rubric** and associated discussions of challenges completed this theme by showing data science’s breadth of problems and methods for addressing them.

Data Science - Opportunities and Challenges

One of our main aims has been to accurately and comprehensively cover both data science's positive benefits and its potential harms when misused.

- The domains where data science is proving applicable are already important and are growing rapidly. They affect almost everyone's day-to-day life. As co-author Jeannette

says, “Data science provides the 21st century methods to tackle 21st century problems,” meaning climate, public health, education, and more.
- Applying data science well is difficult. We discussed many challenges in Part III relating to data, modeling, dependability, supporting understandability, setting objectives, tolerating failures, and meeting ELSI objectives. They are mathematical, engineering, epistemological, societal, and political in nature. Because of them, society has developed concerns over data science’s actual and perceived harms, as we summarized in Chapter 15.

Understanding and Applying the Analysis Rubric

We have aimed to instruct students and practitioners on how to approach new data science problems by offering the Analysis Rubric with its seven elements and implied questions:

- Is there **data**?
- If the goal is to provide a conclusion, is there a **model** that will do so?
- Will the project be **dependable**?
- Can the project provide sufficient **understandability**?
- Are there clear and beneficial **objectives**?
- Can the application **tolerate failures?**
- Are the needed **ethical, legal, and societal** implications met?

A priori, the rubric helps determine if a proposed project is feasible. A posteriori, it can be used to see if it addressed needed issues. We do not advocate a particular top-down or bottom-up methodology, and we recognize that different project teams will use the rubric in different ways. We feel that the benefits of a rubric or checklist are well-documented,[407] and that our rubric is a good starting point for most teams.

This book presents many examples of applying the rubric. Some showed data science works naturally; others showed great challenges. The examples informed us not only as practitioners, but as people who interact with uses of data science on a daily basis. We acknowledge that some of our examples will become stale and that future readers will be surprised we omitted others of then-current contemporary importance.

As previously noted, we readily admit that aspiring data science practitioners need to augment our discussions with technical material from statistical, optimization, and computational texts.

Ethics

We believe a data science project is only a complete success when it satisfies an actual human need and doesn’t merely meet a statistical measure. To that end, a successful data scientist considers not only design constraints and statistical goals but also the context that defines success. Framed this way, and with a nod to our title, *Data Science in Context*, a data science project's success clearly depends on the human and societal context in which it exists.

By no means do we argue that it's easy to balance ethical and other objectives, but we do argue that the act of trying to do so results in better outcomes. Ethical consideration is not just for philosophers, it is a necessary and useful exercise that is the responsibility of all data science practitioners.

We recommend the Belmont Principles of **Respect for Persons**, **Beneficence**, and **Justice** (Chapter 3) as a concrete framework for thinking about ethics in data science. We also emphasized that ethical uses of technology necessitate scientists and engineers to successfully navigate all of Part III's challenges. We then discussed the organizational and governance challenges that make it hard to balance incentives and achieve good outcomes. Chapter 19 concluded the ethics discussion with recommendations on quality and organization.

Addressing Concerns

In Part IV, we divided societal concerns on data science into five categories. Summarized in Table 15.1, they are the *Data Science Implications on Economic and Fairness*, *Impacts on People and Institutions*, *Personal Implications to Data*, *Institutional and Societal Operation*, *the Environment*, and *Trust*. We then proposed some recommendations of varying specificity and complexity:

- Some are straightforward and relatively short-term. For example, some of our recommended technology improvements can occur quickly. As one example, we are seeing rapidly increased uses of federated learning to reduce privacy risks. Also, we could quickly define and use more precise vocabulary (e.g., for specific privacy concerns) and thus have clearer and more thoughtful policy debates.
- Some are clear to us but take time. A focus on education is of the utmost importance, as individuals with data science knowledge will gain leverage in their vocation and better understand their rapidly changing world. More practitioners will also speed progress. We want to emphasize that humanities and social science education provides data scientists with valuable perspectives.
- Others are complex. Regulation requires care due to negative, unintended consequences. Issues such as content moderation or the implications of scale are complex and require significant thought and consensus-building.

Reflections from Your Authors

We have each written a brief essay, representing our own individual interests and concerns.

Jeannette M. Wing: Where Does Data Science Fit in Academia?

"Will data science evolve as an academic field like computer science or like computational science?" This insightful and probing question asked by Ed Lazowska, renowned computer scientist at the University of Washington, at the inaugural Academic Data Science Leadership Summit in 2018, still has no answer–it is too early to tell. And maybe it doesn't matter.

Computer science as a field of study emerged from its roots of electrical engineering, mathematics, and business in the 1960s. Within two decades, one could major in computer science, get a Ph.D. in computer science, be a faculty member in a computer science department, be a dean of a computer science school, publish in computer science journals and conference proceedings, buy computer science textbooks, attend computer science conferences, get a job as a computer scientist, join computer science professional organizations, and win the equivalent of the Nobel Prize in computing (i.e., the Turing Award).

Funding agencies, such as the National Science Foundation and the Defense Research Projects Agency, had created directorates or offices dedicated to computer science. The information technology sector grew quickly on the shoulders of computer science giants. To date, industry demand for computer scientists continues to outstrip the supply. It took only a couple of decades, but computer science is now an established and accepted field of study worldwide. No question.

Computational science, in contrast, refers to the use of computational methods, tools, and thinking in the sciences. For the most part, it is not considered a single field of study. Rather, one can specialize or even major in computational astrophysics, computational biology, computational chemistry, computational materials science, computational neuroscience, computational physics, and more. But most universities do not have a computational science degree program or a computational science department.

Data science, like computer science, has its roots in other disciplines. Data science, also like computer science, has nearly universal applicability. So, will the foundations of data science solidify and evolve, much like they did for computer science, and lead to data science being its own discipline? Or will data science be so integral to each domain, where eventually each domain's repertoire of methods necessarily includes data science?

Here are two other suggestive analogies: On one hand, mathematics is the language of science, yet it remains an independent field of study. On the other hand, software engineering is typically studied as part of computer science, yet one of the first jobs a computer scientist might land in industry is titled "software engineer."

Universities today are embracing data science but in different ways. In some schools, it is a part of the computer science department or college (e.g., University of Southern California and University of Massachusetts, Amherst) or part of the statistics department (e.g., Carnegie Mellon University and Yale University). In some, data science is its own school (e.g., University of Virginia), alongside its computer science and statistics departments. At some schools, there is an independent data science institute (e.g., Columbia University, Georgia Tech, Harvard University, University of Chicago, University of Michigan, University of Washington), cutting across schools, and thus across disciplines; however, degree programs and joint faculty have homes in an academic department. And some schools have a hybrid approach: at MIT, the

Institute for Data, Systems, and Society, serves the entire university, cutting across all schools and disciplines, but organizationally, it is housed in the Schwarzman College of Computing; at New York University, the Center for Data Science serves the entire university but offers its own degree programs and hires joint faculty; and at UC Berkeley, the Division of Computing, Data Science, and Society, is a new academic entity, incorporating its computer science faculty, who are part of the School of Engineering, and Berkeley's School of Information.

Watching these multiple models emerge is not surprising, as data science builds on core strengths in computer science, statistics, and operations research. How a university embraces data science is related to its organization of these and other related disciplines. Universities understand the value of data science in the future of all academic pursuits, and thus to their own future, but today there is no one right answer to the question when the president asks "Where do I tuck data science in the org chart at my university?"

At the same time, interest in data science continues to skyrocket. The 2018 Academic Data Science Leadership Summit led to the creation of the Academic Data Science Alliance, a non-profit organization initially funded by the Gordon and Betty Moore Foundation, Alfred P. Sloan Foundation, and the National Science Foundation. As of 2021 it had 40 founding member institutions. It convenes annual meetings, already engaging over 100 organizations from academia, industry, and government to share best practices in education, research, and the ethics of data science.

And the next generation is voting with their feet. In late 2020, the NSF-funded Northeast Big Data Innovation Hub, headquartered at Columbia University, started an effort in the nine Northeastern states to engage directly with students interested in data science. This effort blossomed into the National Data Science Student Data Corps, which by January 2022 had 1922 student members (including high school students) from 348 colleges and universities, 40 states, and seven countries. Twenty-four percent of the members are from Minority-Serving Institutions. Students from over 40 academic institutions are asking to create their own NDSC chapters.

Regardless of how data science fits into an academic organizational structure, data science is here to stay. If your child asks you "Should I study data science?" reply "Yes!" because data science students learn techniques useful for any future profession–and useful for life.

Chris Wiggins: Rethinking Responsibility and Success

In May of 2017, I asked the scholar danah boyd how we engineering educators could convince students and practitioners that context was worth studying. Her suggestion was to push data scientists to think more deeply about what it means for data science research and data science products to be "successful": a success does not simply mean meeting a statistical goal (for example low generalization error) but rather that the research or product aims to actually improve lives.

Much of this book has been about the promise of data science. Certainly, in the last decade, it has become clear that computational advances for making sense of the world through data have vastly increased its impact. Arguably, the mindset of data science goes back to work by John Tukey, who split his career between industry and academia (Bell Labs and Princeton). A slightly earlier point of origin is the dawn of digital computation at Bletchley Park, where computing with data and the combined statistical and engineering mindset has been credited with shortening World War Two by two to four years. However, as Spider Man's uncle once warned him, “With great power comes great responsibility.”

For most of us raised as technologists, the idea that a technical subject can have “politics,” meaning it can change the dynamics of power, is unfamiliar and sometimes unbelievable. Like many earlier researchers in machine learning, my personal training was in physics, a field in which the potential politics of one's work has been inescapable since August of 1945. One of many significant differences from data science today is that the technical and financial barriers are lower than ever before to having a wide impact on a large number of people.

Part of our goal in this book, implicit in the title *Data Science in Context, is to* illustrate how data science as a technical field is built from and shares techniques with many adjacent fields of the last 50 to 100 years. A second meaning to “data science in context” is to remind practitioners that, particularly in industry, data science powers products – that is, things that real people use and which impact their lives. A similar sentiment guides our treatment of ethics. We hope that this book not only convinces you, our reader, that the context is worth thinking about, but also that it gives you the conceptual tools for thinking through this context and the difficult responsibilities data science practitioners now bear.

We hope that by introducing you not only to the fundamental technical concepts of data science, but also to fundamental concepts such as the Belmont Principles, we will help you expand and ground your conception of what constitutes a successful data science project and a successful career in data science.

Peter Norvig: From Algorithms to Data to Needs

When I started work in artificial intelligence in 1980, researchers were focused on inventing new algorithms to solve problems more effectively. By 1990 it became clear that the field of AI was changing, in three ways:

- The canonical approach shifted from an expert system (a program designed to mimic the thinking processes of human experts) to an intelligent assistant (designed not to imitate humans, but rather to optimize performance on some task–to do the right thing).
- Researchers (notably Pearl[408], along with Cheeseman[409], Heckerman, Horvitz, and others[410]) convincingly argued that reasoning with probabilities and decision theory was

superior to reasoning with logic for the types of problems AI faced–problems where uncertainty is a key component.

- Machine Learning grew from a subfield to the dominant approach within AI, and the emphasis of the field shifted from *algorithms* to *data*. No longer were knowledge bases carefully hand-crafted and curated by graduate students; instead we could appeal directly to the data. Researchers such as Banko and Brill[411] showed learning curves that continued to improve as the amount of data went from thousands to millions to billions of unlabeled examples. There was plenty of room at the top for more data, and the phrase "big data" came into vogue.

Stuart Russell and I were able to chronicle these changes in a textbook (first appearing in the mid-1990s),[412] and we had good luck in our timing; professors and students were eager to embrace this newly evolving picture of AI. Later, Alon Halevy, Fernando Pereira, and I were also able to put down some thoughts on the effectiveness of data.[8]

With the frontier of AI shifting from algorithms to data, I swapped my .edu address for .com to get the resources–computing power and teammates–necessary to harness big data. It was an exciting time and we created applications that were used by millions, and then billions, of people. Before anyone codified AI Principles, I learned to embrace the principles of the World Wide Web Consortium: "Put user needs first" and "The web should not cause harm to society." I'm proud of the dedication and hard work that my teammates put in towards achieving these goals.

One day in 2012 I was sitting by myself, contemplating what project to focus on next, when Geoff Hinton approached, very excited, and said "You've got to see this. It finally works!" He showed me the image classification network that was to win the ImageNet ILSVRC competition. I immediately realized that this would mark another significant change in the field, but I underestimated just how widespread the influence of deep neural networks would become in vision, speech recognition, natural language, robotics, and other fields.

By 2020, it looked like the field had changed again. This time it was a change in how we look at problems. We still had to answer "what's the right algorithm?" and "what data should we use?" but most often the hardest question to answer was "what is the goal?" or "what do we want to optimize?," and the related questions of "what is fair?" and "who is this for?"

Underlying all this is the deeper question "what context are we operating under?" I spent a lot of my time in college and grad school playing Ultimate Frisbee, and in 1982 I was called upon to serve on the committee to write the 8th edition of the rules. My experience with rule-based systems, both in AI and in sports, told me that when there is a specific set of rules, competitors look for loopholes in the rules. For example, in basketball, sometimes a player will intentionally foul an opponent, because doing so gives their team an advantage. To counter this, the rules are constantly updated with new penalties (e.g., the "clear path" rule and the "Hack-a-Shaq" rule).

I realized that it would never be possible to foresee all situations and codify all penalties, so instead the 8th edition rules state that “Ultimate has traditionally relied upon a spirit of sportsmanship which places the responsibility for fair play on the player” and “Such actions as taunting, dangerous aggression, intentional fouling, or other “win-at-all-costs” behavior are contrary to the *spirit of the game* and must be avoided by all players.” In effect, we told players that their first responsibility was not merely to optimize their chance of winning the game under a set of rules; their primary goal was to conscientiously contribute to the betterment of the community of players, and only secondarily to win the game.

I saw these lessons as both a challenge and an opportunity for machine learning systems. The challenge: any system that is described by a set of rules may have exploitable loopholes. The opportunity: it is easier to describe the boundaries of acceptable behavior with a set of examples than with rules, and machine learning systems are good at learning from examples. If we model things correctly, we can build machine learning systems that learn to act like conscientious members of a community, not like win-at-all-costs exploiters. We want to make it easier to create systems that are creative enough to, for example, come up with “move 37” in Go, yet are ethical enough to know that cheating is not the right way to win, and that turning the whole world into one big paper clip factory is not the right thing. I believe that a major area of research will be in finding better ways to communicate with machine learning systems, to have more effective ways of describing to them the bounds of what we want them to do, and to help us discover for ourselves what we really want.

To date, the computer industry does not have the best record of protecting the community from win-at-all-cost exploiters. The web is a global marketplace, for products, ideas, and attention; and we have made it all too easy to harvest user’s attention.[413]

The 20th century British philosophers Michael Philip Jagger and Keith Richards wrote that “you can’t always get what you want” but “you get what you need.” However, when it comes to the web, they got it exactly backwards. We have constructed a very efficient feedback mechanism to say what you *want*–a system that encourages you to consume the latest amusing game, meme, or video, and then uses collaborative filtering to make recommendations to others as well. But we don’t have a good system for saying what we really *need*–equality, justice, health, safety–and we don’t have good feedback systems to make sure everyone gets them.

The challenge for machine learning and data science is to build systems that align with society’s real needs, and work for everyone. I hope this book will inspire researchers to develop ideas that contribute to this; will enable developers to build systems that work for the betterment of all; and to empower consumers to know what they can ask for.

Alfred Z. Spector: Post-Modern Prometheus

I recently came upon an article arguing that data science had hit "Peak Metaphor."F[31] 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 CO_2 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

[31] Sondregger 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.