Chapter 6

Applying the Analysis Rubric

In this chapter, we pivot towards taking the view of a team building new data science applications. Their work begins when someone creates a concept for a worthwhile and plausibly achievable technique, product, or service. Goals may range from scientific pursuit to commercial gain. They may be motivated by the need to solve an existing problem or by a novel way of extracting information out of an existing data source.

Following conceptualization, the design process typically continues with further analysis and refinement of the initial idea, often with its decomposition into more solvable subcomponents. All of this has the ultimate goal of creating an implementation that delivers value. As we mentioned in the beginning of Chapter 5, most design approaches mix bottom-up and top-down thinking. There are many methodologies for designing new products and services (or even experiments), but they are a topic for a separate book. We recommend the classic *The Design of Everyday Things* by Don Norman (Norman, 2013), who we mentioned earlier as founder of the field of visualization. He can also be credited with moving computer science towards the empirical approaches that truly made computers much easier to use.

Whatever design approach is chosen, teams can realize value from the initial concept by applying the Analysis Rubric to ensure sufficient attention is paid getting the data science right. We now illustrate its use in 26 applications in six different domains, as listed in Table 6.1 through Table 6.6. More naturally solvable problems are towards the top of each table, and more difficult ones towards the bottom.

In the tables:

- A tick does not indicate that the Analysis Rubric is *easily* met, just that there is a path towards satisfying it. For example, web search has working technical solutions, but implementing them requires enormous creativity and labor. (We admit to there being considerable nuance in our assignments of ticks, so there is room for educational dialectic and disagreement.)

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

Table 6.1 *Transport and mapping applications of data science.*

Transport and mapping applications	Tractable data	Feasible technical approach	Depend-ability	Understand-ability	Clear objectives	Toleration of failures	ELSI
Traffic speed	✓	✓	Feasible, but risks	✓	Subtle challenges	Individual but not system-wide	✓
Route finding	✓	✓	Feasible, but risks	✓	Nuances and potential externalities	No egregious errors and not system-wide	A few ELSI issues
Level-5 (fully autonomous) cars	✓	Feasibility unproven	Resilience challenge	Explanation likely needed	Difficult challenges	Great safety required	All difficult

- A table element's few words cannot fully address an application's difficulties in meeting an Analysis Rubric element.

Some applications have considerable prose explanations, but in the grand tradition of "The proof is left to the reader," some descriptions are sparse. The more cursory ones, however, provide a greater opportunity to think through how we filled in the table rows with ticks and other annotations.

6.1 Transport and Mapping Applications of Data Science

Traffic speed estimation (Table 6.1) is a relatively straightforward data science application of great interest to drivers. Starting with implementation-oriented rubrics, data comes from cell phones that know their own location and are networked to data-processing cloud-based servers. The technical approach is based on a system knowing a cell phone's location at the beginning and end of some time interval. With that information, the system can compute the average speed by dividing the distance between those locations by the length of the interval. If captured from enough cell phones, the average data accurately reflects the speed of traffic on a stretch of road.

Privacy and security risks are minimal since this application need not use any identifiable cell phone or owner data. Anything identifiable is extraneous and can be discarded. The implementation resists abuse because spoofing requires physically manipulating many cell phones' locations. However, even this application did prove hackable when a Berlin artist put 99 cell phones in a slow-moving wagon in February 2020 and generated fake traffic jams on Google Maps (Bonifacic,

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

2020). This undoubtedly reminded the product team to protect against very close proximity, nearly identical location signals. Finally, the algorithm is extremely simple and functional, although it cannot differentiate between a road nobody happens to be driving on and a closed road. This is a **corner condition**, and most computer applications have them. To deal with this, the system designers could add additional data such as road-closing data from municipal or state websites or simply not report traffic speed on roads with insufficient data.

The application is also straightforward in the sense that it has no real need to provide explanation, show causal relationships, or provide reproducible data to others. Failures are tolerable, because the estimates are so much better than nothing. Also, drivers are aware of the likely limitations, since traffic can unpredictably and rapidly change due to an accident, inclement weather, or some unusual event.

The objectives are reasonably clear, but there are nuances to consider. Should the system show the current traffic speed, or should it show the traffic at a predicted arrival time? Should it show traffic speed relative to a road's speed limit (green might mean traveling at the speed limit) or perhaps versus the expected traffic speed (green is pretty good for rush hour)? These are relatively minor differences, but such details need precise definition, the data needed for implementation must be available, and the comparisons must be added to the software.

As for the ELSI review, there seem to be few problems. However, some might be concerned that providing traffic speed estimates tends to make driving easier, consequently reducing the appeal of mass transit. There is potentially some significant risk if the application fails completely, particularly as the driving population becomes more dependent on it.

Overall, this application seems quite straightforward and it is labeled with mostly ticks in the table.

Route finding is a related problem, listed to illustrate its much greater technical challenges and the unexpected complexity of objectives. Addressing the approach first, a classic operations research method would model roads as a network and algorithmically compute a shortest path based on predefined parameters such as speed limits and distances. A data-driven approach might only look at what others have done and use their successful paths (that is, paths that take the least amount of time or fuel for given start and end points). Actual approaches combine analytical techniques from the operations research's graph models with dynamically changing vehicle experience data.

There are many possible enhancements a design team can consider in this combined approach. With real-time data, how should the model (or routing algorithm) adapt when a traffic accident occurs? Should a model be calibrated with

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

historical delay data as of when a route was requested or at a vehicle's predicted arrival time in a particular area? Should the directions account for feedback, that is, the impact on future traffic of current drivers following new traffic directions? How can a direction-finding system include creative driver paths taken from the history of travel – combining these driver-found solutions with those proposed by a model? Route-finding systems are increasingly using these hybrid approaches.

While an *obvious* goal is to minimize delays, property owners on quiet side streets might not want them used as overflow capacity for major highways. While not addressing this particular issue, Google (in the spring of 2021) added new possible objectives to its navigation system so users can prioritize safer or more fuel-efficient routes.

There is also the question of priority. Is a system first-come, first-served, or should buses, trucks, multiple-occupant, or lower-emissions vehicles, or even higher-paying vehicles, have priority?

From a failure perspective, small errors are fine, but it is not fine to direct a vehicle onto a one-way road against traffic or to send a vehicle onto a road that is closed. Furthermore, the complete breakdown of vehicle routing systems is growing problematic as drivers become increasingly dependent on them. Breakdowns could occur due to application failures, cloud infrastructure failures, attacks on the GPS system, etc., though there is enormous redundancy built into each component.

Level-5 (fully autonomous) cars are a complex application that incorporates many data science components built from many kinds of datasets and technical approaches.

Self-driving cars need to process and act on many forms of data representing road networks, lanes, interchanges, danger areas, traffic bottlenecks, etc. Bespoke sensing activities (such as Google or Bing street mapping cars) and car-mounted sensors can provide much of what is needed. Important data can also be gained from the many vehicles that have accurate location data. As we have observed, vehicular location data is extremely useful in showing traffic speed, but geotraces can be of even greater use for training autonomous driving models.

Self-driving cars can clearly incorporate machine learning techniques, which have proven very successful in object recognition. Examples include detecting curbs, stop signs, bicycles, or turning vehicles. Autonomous vehicles also need to learn how human drivers react to the sight of such objects and many other driving situations.

This application has significant dependability requirements. While privacy issues are not particularly different from other applications that know user location, there are significant security issues. If bad actors gain control, individually or

collectively, cars (which some liken to two-ton projectiles) can do great damage. Some people might even try to bait (or, perhaps, abuse) a self-driving car into behaving improperly. Algorithms must be resilient in the face of many unanticipated conditions. They must respond appropriately to human drivers and their often iffy driving habits.

The objectives are difficult to get right. As two examples, consider first balancing arrival time versus risk tolerance, then determining right of way in complex situations. Autonomous vehicles may need to provide explanations for their actions, especially if they get into an accident with property loss or injury. In particular, they may need to justify their action as the best possible under the circumstances. While no systems (humans) are perfect, self-driving cars are very intolerant of failure. Many errors have extreme legal, ethical, and financial risks.

We don't know if present approaches are sufficient to allow for Level-5 Automation (fully attention-free self-driving cars in all conditions) (US Department of Transportation, n.d.). However, they almost certainly will allow autonomous cars to operate under specific conditions (Level 4) with better safety than cars with human drivers. However this plays out, data-science-based techniques will continue to make driving safer for all.

6.2 Web and Entertainment Applications of Data Science

There are many data science applications on the Web, in part because they are so natural given the large bodies of data stored and accumulated from users. Because of our familiarity with them and the previous web-related applications of spelling correction, speech recognition, and recommendation engines, our explanations of the examples in Table 6.2 are more succinct. In particular, news and video recommendations are similar to music recommendations, although they have more profound ELSI considerations.

Identifying copyrighted material became important when video sharing on the Web became prevalent in the 2000s. If sites could do this, they could then offer copyright holders the opportunity to take down or perhaps monetize their videos. When Google purchased YouTube, it quickly confronted the copyright problem and developed ContentID, a matching system. It uses machine learning to match uploaded content to previously registered (and provided) copyrighted material. A match triggers a notification to the copyright holder and makes the alleged infringer ineligible to receive advertising revenue. The objective is clear: to reduce copyright infringement.

Abuse is the biggest implementation challenge, as copyright infringers can attempt to camouflage material or bad actors can claim copyright they don't

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

Table 6.2 *Web and entertainment applications of data science.*

Web and entertainment applications	Tractable data	Feasible technical approach	Depend-ability	Understand-ability	Clear objectives	Toleration of failures	ELSI
Identifying copy-righted videos	✓	✓, but not foolproof	Abuse	✓	✓	✓	✓
In-session video game personalization	✓	✓	Abuse	✓	Balance tricky	✓	Ethics, financial
Targeted or personalized ads	✓	✓	Privacy, security, abuse	✓	Difficult	✓	Legal, risk, ethics
Video recommendations	✓	✓	✓	✓	Ambiguity	✓	Complex
Web search	✓, but voluminous	✓, but very many techniques	Privacy, security, abuse	✓	Significant nuance	Certain failures serious	Legal, risk, ethics
News feed recommendations	Fake news	Diverse challenges	Resilience, abuse	Increasingly important	Significant nuance	Certain failures serious	Legal, risk, ethics

possess. To deal with abuse, YouTube has a dispute resolution system with human oversight. As a truly unanticipated result of ContentID, police officers sometimes play copyrighted music to prevent recordings by bystanders during confrontations from being posted to YouTube (Schiffer & Robertson, 2021). The inclusion of that music makes it very likely that copyright holders will ask for the video be removed.

Otherwise, implementation is straightforward. The application is reasonably tolerant of failure, as uploaders or copyright owners can dispute the automated system's answers and request human arbiters make the final decision. The system appears to be in conformance with legal structures and has few ethical concerns.

In-session video game personalization can utilize game and player data to make video games more compelling, and perhaps more addictive as well. This is most feasible with games that have probabilistically occurring events (a certain card being dealt) or when the computer is a player. The data and technical approaches exist. However, there are dependability (in particular, abuse) challenges, hard-to-define objectives, and ELSI issues, e.g., relating to how addictive a game should be. There will be more on this, particularly in Chapter 12.

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

Targeted or personalized ads is one of data science's most prevalent web applications. Targeted ads may be shown when someone searches the Web, shops online, or views entertainment or news sites. This application has some commonality with recommendation systems, as its goal is to place (or recommend) an ad that meets some goals. Ads have been a primary revenue source for many internet companies, which try to make *personalized ads* beneficial to these four different constituencies, which are also illustrated in Figure 6.1:

- The **consumer** viewing the ad, who wants to see pleasing, relevant, and useful ads.
- The **publisher**, such as a periodical, a video site, or a blogger, who receives revenue for providing ad space and viewership. They want ads that maximize revenue but don't detract from their site's value.
- The **advertisers** who place ads to enhance their image or to sell products. An advertiser often desires a particular target audience so it can customize its ads, making them more effective.
- The **advertising platform** that coordinates the matching: the "right" ad for the "right" user on the "right" site in the "right" context. The advertising system makes a commission based on the value of the ads placed or clicked-on, how many people see or click on an ad, or perhaps even how much they buy. In some

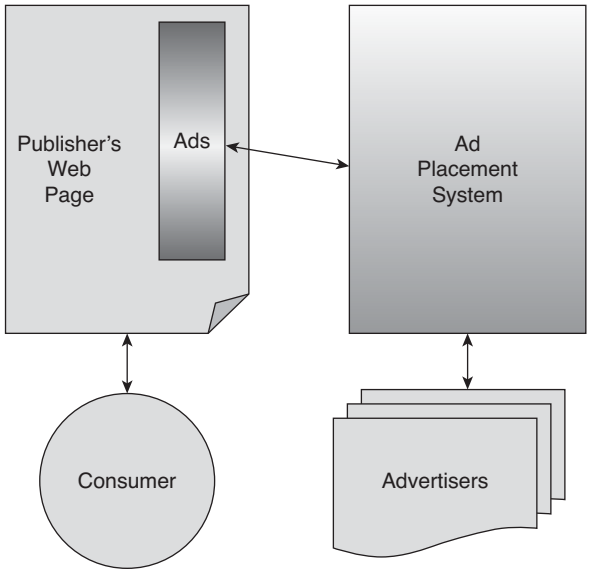

Figure 6.1 Advertisers bid to have their ads inserted into publishers' web pages; an ad placement platform chooses the best ads to be shown to each consumer.

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

instances, such as search and social networking advertising, the ad placement system and the publisher may be the same.

Vast amounts of data are available for ads personalization, with its type depending on the particular situation. Data may include geographical location, recently viewed websites, purchases, and (in the case of closed systems such as search engines or Facebook) detailed information on interests or searches. Advertisers have data that powers systems that provide automated recommendations on advertising budgets, keyword selection, and ad placement. Recently, there has been considerable evolution in how to gather data, due to public and regulatory concerns, and it seems likely that the role of web cookies will decline.

There are many technical approaches for generating a recommended ad from that data, often based on machine learning based using deep neural networks. Requirements must inform the technical approach, because users, advertisers, and publishers will tolerate only some mistargeted advertisements.

The engineering complexity of a personalized ad system is very high. Once appropriate ads have been identified, online auctions direct ads to the buyers who find the space most valuable. For any given query or page view, the potential ads must be identified and the auctions run so quickly that the user does not notice any delay in page load. Advertising systems may need to handle thousands of queries per second.

Perhaps targeted advertisements' greatest implementation challenge relates to dependability concerns. In part, this is due to the significant privacy and security issues when vast amounts of personal data are applied to a problem. As further discussed in Section 10.1, not only does the data need to be protected from disclosure, it cannot be used in ways a user would feel are spooky. However, increased privacy protections could also decrease competition, as discussed in Section 10.1.5. In addition to privacy issues, the vast amounts of money associated with advertising systems invite financial and other forms of abuse.

Turning to the requirements rubrics, advertising can be opaque, with little need for explanation. But there is the complex question of how to balance the needs of the consumer, publisher, advertiser, and advertising system. For example, should near-term site revenue or long-term customer satisfaction be optimized? How is revenue allocated between the publisher of the site showing the ad and the system brokering the advertisements?

Ethical, legal, and societal considerations also come into play in setting objectives, for example:

- If the objective were solely to maximize clicks on an ad (in the near term), deceitful ads that practiced "bait and switch" would be truly overwhelming.

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

- If the objective were only to sell a product or maximize economic activity, there would be no limits on advertising. But protecting consumers is also a concern: advertising of medicines is carefully controlled, and much of the world has banned cigarette ads, optimizing health over economic activity.
- To reduce foreign influence on political systems, there may be laws that restrict political ads paid for by foreign entities.
- Even more generally, some argue that ads elicit consumer behavior not in consumers' true self-interest. Ads encourage consumers to overindulge in alcohol, tobacco, and unhealthy food, or to spend beyond their means. Nobel Prize-winning economists Akerlof and Shiller in their book *Phishing for Phools* suggest regulation (Akerlof & Shiller, 2015).
- Advertising payments may support the creation and display of harmful content, or the ads themselves may be harmful even if not illegal.
- There are financial risks if the ads are not seen by actual potential customers due to insufficient system-wide abuse prevention.

Chapter 12 addresses the challenges in setting objectives, and Part III, in aggregate, addresses many other challenges. Despite them all, data-science-driven advertising has been a very powerful addition to commerce and benefited the public by paying for much of the growth of highly popular and useful web services.

Web search makes the Web's vast agglomeration of data easily accessible to users. We expect search to *locate* websites possessing desired information, to *navigate* to a specific website, to *return* an answer, or perhaps even to *perform* a transaction, such as finding a route from city A to city B (Broder, 2002). Web search engines are critically important because they are a frequent gateway to the Web.

Web search, like music recommendation, relies on three broad sources of data:

- **content** – the words on a page and the concepts behind them;
- **metadata** – facts about the page, the reputation of the website hosting the page, and facts about the links to this page; and
- **user reactions** – popularity of a page in general, or as an answer to a specific query.

Abuse is a significant problem. All webmasters want their websites to be ranked highly, and some achieve high rankings by providing quality content. But some "web spammers" try to abuse the system by creating fake web links and manipulating the system in other ways. Search engines need to stay on top of each new type of manipulation and reward good content, not manipulation. Some complex issues arise in setting objectives:

- Search objectives may be subjective and conflicting: Designers need to consider how to arrive at answers for political queries and what balance of results should

be presented. Should the website's response speed influence the ranking? Should a search engine try to return search results from multiple sites? Should search engines provide answers when they can, or restrict themselves to returning page links?
- Many ethical questions that impact the objectives also arise, such as: What focus should there be on the source's likely veracity? Should the searcher's pre-existing views be taken into account?

Failure tolerance is complex. When search engines return links to web pages, the searcher knows that some links will be excellent, but others may not be. Thus, search engines are tolerant of some forms of failures. For example, most users will understand if the query "cataract" returns information on eye diseases rather than waterfalls, and refine their search accordingly. However, users may not tolerate other types of errors. For example, a search engine's reputation would decline greatly if even a small percentage of its results were truly atrocious.

For example, in 2004 a Google search on "miserable failure" returned "George Bush" due to Google's then-inability to prevent a kind of abuse, termed a "Google Bomb." For this reason, many search engines use algorithms, tuned with considerable human labor, to prevent very poor results. This reminds us that data science approaches should consider rare downsides and mitigate the effect of bad actors.

Video recommendations and **news feed recommendations** are important data science applications that share some characteristics with music recommendations as described in Section 4.3 and Section 5.3. Both build off of similar techniques, but there are two key differences:

- The scope of their corpora varies. As presented, music recommendations focused on a relatively constrained corpus as defined by the music publishing industry. However, video and news feed recommendation applications have far larger and more irregular corpora, particularly if there is user-supplied content. Publishers are highly motivated to have their content viewed, and they sometimes go to extreme lengths to game recommendation systems.
- The individual and societal impact of video and news content is far greater. While all content creators want their content to be seen, many who post video or news stories have significant political goals and go to considerable efforts to achieve them. Recommendation engines have their own complex goals (e.g., perhaps to moderate content or suppress fake news) that are both challenging to define and to meet.

Because of these two differences, all of the Analysis Rubric elements are more complex to satisfy. We will discuss these applications' many challenges

in Part III and summarize some ethical issues relating to news recommendations in Chapter 7.

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.

Table 6.3 *Medicine and public health applications of data science.*

Medicine and public health applications	Tractable data	Feasible technical approach	Depend-ability	Understand-ability	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

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

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 (Google, n.d.-a). Google engineers felt these reports would show society's acceptance of government recommendations, and perhaps catalyze safer behaviors or better governmental responses (Aktay et al., 2020). 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 (Nuño et al., 2020).

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. (2021). Modeling approaches for reducing vaccine hesitancy would seem to be particularly difficult.

Identifying disease outbreaks using crowdsourced 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 1800s, when Gram staining started using stains to

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

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, MRI (magnetic resonance imaging) scans, and multispectral cameras, can then be classified to carry out or aid diagnoses (Shen et al., 2017). In fact, there are now published reports indicating that some techniques are approaching human capabilities (Esteva et al., 2017; Nabulsi et al., 2021).

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 (GWAS), 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 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" (Collins, n.d.; Eisenstadt, 2017). 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 (Claussnitzer et al., 2015).

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 not only use these data as the basis for a scientific

result, but also 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 also 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 Analysis 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, and 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 Analysis 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 the 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 both are informed of risks and have to consent to participation. There is significant risk to researchers, to 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.

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 (Institute of Medicine of the National Academies, 2009) 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 Analysis 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 10-fold by some metrics (Chiang et al., 2020; Williams, 2020). 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 (measuring the speed at which celestial objects are moving) references from the SDSS 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 reproducibility because the astrophysical data was publicly available and the scientists could publish their models.

Weather or earthquake prediction are data science applications based on different physical principles, though, like the astrophysics example, they center around forecasting. Weather predictions are now useful enough that we rely on

Table 6.4 *Science-oriented applications of data science.*

Science-oriented applications	Tractable data	Feasible technical approach	Depend-ability	Understand-ability	Clear objectives	Toleration of failures	ELSI
Determining the historical temperature of the universe	✓	✓	✓	Reproducibility	✓	✓	✓
Weather or earthquake prediction	Insufficient sensor coverage	Very complex problem	✓	Explanation, reproducibility, causality	✓	Some harmful	Risk

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

them daily, but earthquake prediction has achieved only limited success. For somewhat distant earthquakes, systems can provide seconds of advanced warning since electronic broadcasts travel faster than seismic waves.

For weather forecasting, it has been demonstrated that more sensor data, such as temperature, pressure, and/or wind data, at more locations, improves forecasting. Seismologists believe this may also be true in their domain. Weather prediction models are hugely complex, and small errors in the measurements or the models can cause large changes in the predictions. Furthermore, small differences in an event's location (e.g., the exact path of a tornado) can result in very different effects.

For earthquakes, modeling is at a very early stage, though hopefully machine learning approaches will prove useful (The Economist, 2022b). For new data science approaches, scientists will want reproducibility to verify the results. If data science leads to new scientific results, providing explanation or demonstrating causality may also be important. If society were to become dependent on earthquake predictions, ELSI Analysis Rubric elements could be of considerable importance. The risk from mistakes could be considerable, in both economic and human costs. However, at least these scientific examples have no privacy risks.

6.5 Financial Services Applications of Data Science

The financial services sector is a huge part of the economy. In the US, it contributes about 10% to GDP and includes sectors such as banking, investing, insurance, lending, and more. Prediction problems abound because of the enormous value in knowing a future interest rate, an equity or bond price, the likelihood of a claim or a default, or even a customer's real identity. Good predictions strongly reward those who can make them, encourage safe practices because of insurance pricing incentives, and may benefit society at large by guiding capital to areas of higher returns.

To support prediction models, companies and governments are capturing ever-growing amounts of data, including detailed information on customers, businesses, markets, and financial transactions. Some countries, such as India and China, are pushing hard for almost all transactions to be digital to enable easy capturing of almost all financial data.

Very large datasets coupled with statistical and machine learning techniques are already used to evaluate individual investments, create portfolios of those investments, and analyze their risk under varying assumptions. Data science applications may provide insight to analysts and portfolio managers who then apply their own discretionary judgment. Alternatively, algorithmic investors might also use them to draw conclusions and execute investment choices, as described by the book *Inside the Black Box* (Narang, 2013). Recently, consumer finance has been affected by

automation with the advent of "Robo Advisors" providing individual investors automated investment advice.

Data science is also used to detect fraud and to ensure compliance with anti-fraud regulations, such as "know-your-customer" identification rules. Its tools predict health, mortality, and property/casualty risks, and thus contribute to pricing insurance policies. Data science helps predict customer wants and needs, thereby better tailoring marketing campaigns and recommendations.

However, even with voluminous data, many of the financial services' data science problems are hard to solve. Much of the data requires immense processing to make it comparable. Some data, such as stock prices, must be available nearly instantly. Furthermore, in some finance applications, market dynamics can rapidly change and invalidate previously useful predictive models.

Almost all financial services problems use confidential customer data, and so have significant privacy and data security risks. There are also security risks beyond data loss, because bad actors or nation states have motivation to spy on or attack individual institutions or the financial sector at large. Below, we'll say a few more words on each element of Table 6.5.

Table 6.5 *Financial services and economic applications of data science.*

Financial services and economic applications	Tractable data	Feasible technical approach	Dependability	Understandability	Clear objectives	Toleration of failures	ELSI
Stock market investment selection and trading	Depends on approach	Complex, but there are successes	Depends on approach	✓ opacity may be acceptable	✓	Certain failures intolerable	Legal, risk, ethics
Underwriting/ pricing of property/ casualty insurance	✓	✓	Privacy, security, abuse, resilience	Explanation	Competing objectives	✓	Legal, risk, ethics
"Know-your-customer" warnings for financial entities	✓	✓	Tricky privacy, security	Explanation	✓	Some, not all	Legal, risk, ethics
Country-wide economic prediction	Insufficient data	Feasibility unproven	Privacy, security, abuse, resilience	Explanation, reproducibility, causality	✓	Probably	Legal, risk, ethics

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

In **stock market investment selection and trading**, data scientists with specialized finance-related knowledge, called quants, modelers, or researchers, choose datasets and create models that recommend trading financial instruments to construct profitable portfolios. These activities are often referred to as **quant trading** or *algorithmic approaches to investment*.

Much of the vast and diverse quantity of data has low predictive value. Market-oriented prediction problems are game-theoretic in that other profit-seeking players may change the financial situation before one can profit from a prediction. In high-speed trading domains, predictions are made and acted on in microseconds.

In addition to the most obvious task of **forecasting** a tradeable entity's price at some future time, quants consider numerous other factors, including the following:

1. The market impact of the trade itself. If a firm is buying or selling a large amount of something, buy orders tend to raise the price and sell orders have the opposite effect. This **slippage** results in a lower trade value.
2. The way the trade should be executed. The proper setting of the size and timing of orders can benefit the prices paid or received.
3. The portfolio optimization so that its individual components' aggregated value has a higher likelihood of achieving investor goals.

There are many approaches to creating prediction and optimization algorithms. Earlier algorithms were based on statistical regression, but today they increasingly involve machine learning. Models are validated by back-testing them on historical data and forward-testing them on simulated future scenarios.

But it is difficult to know how well the simulations will correspond to the actual future. Model development is challenging because investors want to (a) maximize expected returns, (b) avoid big negative swings, and (c) have some resilience to unforeseen circumstances, such as significant changes in investor sentiment. Certain quant challenges, such as price forecasting, are particularly prone to such changes in sentiment, while others, such as trade execution, are less so. When sentiment changes are rapid and broad, they are called **regime change**, a topic that Section 9.1 addresses in more depth.

Algorithms do not have to provide insight as to why they work, though investors would certainly prefer to know. Objectives are usually quite clear. Poor results have some degree of acceptability, since investment returns are known to be probabilistic. However, certain types of errors require regulatory disclosure that cause both reputational and financial risk. For example, firms cannot exceed certain ownership limits on securities or commodities, and they must disclose errors they make. Most investment activities are highly regulated, meaning many seemingly predictive models may be off-limits, with serious legal risks for violations.

There are ethical risks, for instance, in achieving justice. Specifically, data science makes it possible to create products that lure naive investors while giving professional investors opportunities to achieve high returns from mining their data and mistakes. This can result in potentially undesirable wealth transfers.

Underwriting/pricing of property/casualty insurance is a data science problem with a long history dating back to antiquity. The 17th century saw increasing use of math and data in evaluating risks, with the actuarial profession being formalized in the mid-18th century. Today, vastly more data is available for predicting risks and pricing insurance policies, but it is often hard to assemble. Additionally, regulations may prohibit using certain data, such as zip codes for property/casualty insurance or gender for automotive insurance (because using them could lead to unfair bias).

Technical approaches abound, given large amounts of historical data and since, for insurance, the past is usually a predictor of the future.[12] The heavy dependence on individual client data may cause even greater security and privacy challenges than it does in investment management applications, and approaches must be resilient in the face of unexpected behavior. Abuse typically relates to guarding against false claims or fraudulent representations during application processes.

Objectives are quite complex and must balance at least all of the following:

- pricing decisions that will win customers;
- expected profitability margins;
- overall risk to an insurer of specific portfolios (given an insurer may not want too many eggs in one basket); and
- equitable treatment of subpopulations.

Insurance typically cannot use opaque approaches due to needed regulatory reviews. As with investing, losses are inevitable, but certain failures are unacceptable. Legal and ethical issues abound, certainly when considering the pros and cons of different regulatory regimes. As an example, European Union (EU) regulations require auto insurance products not to consider gender, though young men and young women drivers presumably have differential risk.

Know-your-customer (KYC) compliance regulations are part of anti-money-laundering laws to prevent criminal money management use of the financial system. KYC obviously begins with the ability to accurately identify a criminal. This may be more accurately determined by the pattern analysis of activities than from what an account application states. With vast amounts of transactional data

[12] While the past is *usually* a predictor, climate change could increase property claims, and increased human lifespan could be increasing costs for insurance such as long-term care insurance.

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

available, both hard-coded algorithms and machine learning approaches to clustering and prediction can be applied to warn about suspicious behavior.

We need transparency since regulators want to know financial institutions are applying proper due diligence. Guarding against abuse is what this is all about, and this anti-abuse system is itself subject to abuse! Inevitably there are major legal, risk, and ethics challenges. For example, KYC systems will inevitably have occasional false positives that point a finger at innocents. Thus, it is ethically crucial for different subpopulations to be treated fairly. Additionally, automated systems require human appeals channels to resolve problematic results.

Country-wide economic predictions might be better made using the torrent of data and techniques used by financial services firms. While this is in the tradition of econometric forecasting, predictions might be more timely and accurate if guided by far more real-time data.

In 2009, Varian and Choi wrote about using aggregate information on Google Search traffic to better predict sector activities (e.g., retail or home sales) that are material to an economy as a whole (Choi & Varian, 2012). The trend towards using more data has continued (Einav & Levin, 2014), and in 2021 *The Economist* summarized its growing importance in "The Real-Time Revolution" (The Economist, 2021b).

Perhaps, as economies are increasingly digitized, individual transaction data could be utilized for more timely and specific economic prediction and, perhaps, more accurate and effective governmental interventions. One can almost, in a science fiction sense, envision a world where policy makers have a large real-time economic dashboard with economic controls and predicted impacts of all changes. It is not our goal to invent such a system, but rather to map such an application against the Analysis Rubric.

Trying to use all economic transactions would result in a truly huge amount of data. The needed models would be very complex and hard to test, in part because an economy has so many different possible configurations and is affected by so many different stimuli. As with investment optimization, changes in consumer or business sentiment may cause regime change and render existing models unpredictive. Dependability issues are extreme; there are the risks of exposing 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).

Table 6.6 *Government service and political applications of data science.*

Government service and 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 and societal welfare	Legal, risk, ethics

https://doi.org/10.1017/9781009272230.008 Published online by Cambridge University Press

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 in those which are hopeless for them. In systems where candidates need to fund-raise, 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 (Nickerson & Rogers, 2014).

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 that 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.
- Fund-raising may try to either maximize total funds raised, or perhaps demonstrate a broad-based groundswell of appeal by receiving many small donations.

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 (Eriksson et al., 2008). 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 1980s, 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 (Bloom, 1984). 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 Principle of beneficence, 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 (Plonsky et al., 2021). Ideally, individuals with similar criminal histories who commit the same crime in similar circumstances would be treated similarly, which is called **algorithmic fairness** (Dwork et al., 2012).

Today, US courts are using such tools, though some researchers have shown that the risk assessment tools are statistically biased (Eckhouse et al., 2019). However, other researchers have shown that using data-driven decision aids can reduce bias and increase accuracy of pre-trial decisions (Kleinberg et al., 2018). 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.