CHAPTER 2

Data

The plural of anecdote is data.

—Raymond Wolfinger

Data is everywhere. Global disruption and international initiatives are driving datafication. Datafication refers to the modern-day trend of digitalizing (or datafying) every aspect of life [27]. This data creation is enabling the transformation of data into new and potentially valuable forms. Entire municipalities are being incentivized to become smarter. In the not too distant future, our towns and cities will collect thousands of variables in real time to optimize, maintain, and enhance the quality of life for entire populations. They may even be connected to your Amazon or Google accounts and aware of all significant events in our health and day-to-day living as human–computer interaction becomes even more embedded in smart care.

One would reasonably expect that as well as managing traffic, traffic lights may also collect other data such as air quality, visibility, and speed of traffic. One wonders whether a speeding fine may be contested based on the temperature from someone’s smartwatch or the average heart rate from a smart ring.

Data is everywhere. The possibilities are endless. Big data from connected devices, embedded sensors, and the IoT has driven the global need for the analysis, interpretation, and visualization of data. COVID-19 itself proved to be a great example of data sharing between countries and clinicians—particularly around the risk factors, comorbidities, and complications of the novel coronavirus.

What Is Data?

Data itself can take many forms—character, text, words, numbers, pictures, sound, or video. Each piece of data falls into two main types: structured and unstructured.
At its heart, data is a set of values of qualitative or quantitative variables. To become information, data requires interpretation. Information is organized or classified data, which has some meaningful value (or values) for the receiver. Information is the processed data on which decisions and actions should be based.

Healthcare is undergoing a data revolution, thanks to two huge shifts: the need to tame growing cost pressures, which is generating new incentives and reimbursement structures, and digital health, which is democratizing healthcare by empowering people through their digital devices and innovative technologies.

Clinical trends are changing. The advent of social media and immediate access to information through the Internet and health communities mean that patients are clued up about their health. Patients are generating data constantly and want to own and transfer their data between services. Concurrently, evidence-based medicine is striving for personalized care, led by an evidence-based approach to decision-making.

Patients and healthcare professionals alike are vocal in their appetite for all available clinical data to make more effective treatment decisions based on current evidence. A Diabetes.co.uk study in 2019 demonstrated that one in three patients would like to share their data with their healthcare professional, but only one in five patients can do so [9].

Aggregating individual datasets into more significant populations also provides more robust evidence because subtleties in subpopulations may be infrequent in that they are not readily apparent in small samples. For instance, aggregating patients with cardiovascular disease and high blood pressure may present health conditions such as high cholesterol with more confidence than separate datasets.

The incorporation of data in modern healthcare provides an opportunity for significant improvements within a plethora of areas. As medicine evolves to become evidence-based and personalized, data can be used to improve

- Patient and population wellness
- Patient education and engagement
- Prediction of disease and care risks
- Medication adherence
- Remote monitoring
- Disease management
- Disease reversal/remission
• Individualization and personalization of treatment and care

• Financial, transactional, and environmental forecasting, planning, and accuracy (of particular relevance to any operations and finance teams within healthcare organizations)

• User experience, which can be used to expedite all of the preceding items

Data facilitates information. Information powers insights, and insights lead to the making of better decisions.

**Types of Data**

Structured data typically refers to something stored in a database—in a structure that follows a model or schema. Almost every organization will be familiar with this form of data and may already be using it effectively. Most organizations will store at least some form of data in Excel spreadsheets, for example. Within a clinical setting, EHRs may also take a similar, structured form.

Readings from embedded sensors, smartphones, smartwatches, and IoT devices are typically forms of structured data—whether that be the provision of blood glucose readings, steps walked, calories burned, heart rate, or blood pressure.

Structured data is similar to a machine language. Highly organized in its format, structured data facilitates simple, straightforward search and information retrieval operations. Structured data would typically be stored in a relational database for this purpose.

Unstructured data refers to everything else. Unstructured data does not have a predefined model or schema. Data that is unstructured has no identifiable structure within it, and this presents problems for querying and information retrieval. Emails, text messages, Facebook posts, Twitter tweets, and other social media posts are good examples of unstructured data. Unstructured data can unleash a treasure trove of information and insight.

The Gartner report has indicated that data volume is set to grow 800% over the next 5 years, and 80% of this data will be in the form of unstructured data [29].
The problem that unstructured data presents is one of volume. The lack of structure makes compilation and interpretation a resource-intensive task in regard to time and energy. Any organization would benefit from a mechanism of unstructured data analysis to reduce the resource costs unstructured data adds. Unless the data is well understood, a subject matter expert may be required.

Structured data is traditionally easier to analyze than unstructured data due to unstructured data’s raw and unorganized form. However, analytics of unstructured, informal datasets is improving, with the use of data science and machine learning methods such as natural language processing (NLP) assisting in the understanding and classification of sentiment.

It may not always be possible to transform unstructured data into a structured model. For instance, the transmission of information through an email or notification holds data such as the time sent, subject, and sender as uniform fields. However, the contents of the message would not be easily dissected and categorized.

Semi-structured data fits the space between structured and unstructured data. Semi-structured data does not necessarily conform with the formal structures of data schemas associated with relational databases or data tables. However, semi-structured data may contain tags or markers to separate semantic elements and enforce grouping of records and fields within the data.

Languages such as JSON (JavaScript Object Notation) and XML (Extensible Markup Language) are all forms of semi-structured data. Both have been instrumental in facilitating data exchange.

Data is the fuel required to learn in an intelligent system, the most critical component for successful AI.

It is categorized into five sources:

- Web and social media data: Clicks, history, health forums
- Machine-to-machine data: Sensors, wearables
- Big transaction data: Health claim data, billing data
- Biometric data: Fingerprints, genetics, biomarkers driven from wearables
- Human-generated data: Email, paper documents, electronic medical records
When thinking about typical Excel spreadsheets, and in the domain of machine learning, the following definitions are also useful:

- **Instance**: A single row of data or observation.
- **Feature**: A single column of data. It is a component of the observation.
- **Data type**: This refers to the kind of data represented by the feature (e.g., Boolean, string, number).
- **Dataset**: A collection of instances used to train and test machine learning models.
- **Training dataset**: Dataset used to train the machine learning model.
- **Testing dataset**: Dataset used to determine accuracy/performance of the machine learning model.

### Big Data

The term big data is given to the collection of voluminous, traditional, and digital data that are sources for discovery and analysis.

Big data is a popular term that defines datasets that are too big to be stored and processed in a conventional relational database system. In this way, the term big data is vague—while size is undoubtedly a part of big data, scale alone doesn’t tell the whole story of what makes big data truly big.

Through the analytics of big data, we can uncover hidden patterns; unknown correlations, trends, and preferences; and other information that can help stakeholders make better and more informed decisions. Machine learning provides a toolbox of techniques that can be applied to datasets for this very purpose.

Big data was first described by Laney in 2001 as having the following characteristics, also known as the three Vs (Figure 2-1) [30]:

- **Volume**: The quantity of generated and stored data. Big data is typically high in volume. The sheer size of data brings with it its intricacies in the form of storage, indexing, and retrieval.
- **Variety**: Big data is varied in the types and nature of data, requiring efficient storage and analysis as well as systems for processing such data.
- **Velocity**: Big data is received at a speed that brings its demands and challenges.
Data science has been discussing the three Vs of big data for some time. However, there are two other aspects of data to consider that are perhaps more important than the three Vs of big data: the concepts of data veracity and data value. Mark van Rijmenam proposed four more Vs to further understand the incredibly complex nature of big data [31].

Further still, there are a total of ten Vs (Figure 2-2) [32]. In reality, differing subsets of Vs are important for organizations to keep in mind when developing a data strategy.

**Figure 2-1. The three Vs of data**

Data science has been discussing the three Vs of big data for some time. However, there are two other aspects of data to consider that are perhaps more important than the three Vs of big data: the concepts of data veracity and data value. Mark van Rijmenam proposed four more Vs to further understand the incredibly complex nature of big data [31].

Further still, there are a total of ten Vs (Figure 2-2) [32]. In reality, differing subsets of Vs are important for organizations to keep in mind when developing a data strategy.

**Figure 2-2. The ten Vs of data**
A subset of the many Vs now proposed have been included here.

A Gartner survey of 199 members demonstrated that investment in big data was increasing; but the investigation showed signs of slowing growth, with fewer companies having a future intent to invest in big data initiatives. Only 15% of businesses reported deploying their big data project, indicating that the adoption curve is still maturing and that talent and expertise is in high demand [33].

Patients and physicians are driving the adoption of big data. Patients drive data through clinical data, wearable devices, mobile health apps, and telehealth/remote healthcare services. Physicians meanwhile leave an exhaust of clinical data (such as written notes, imaging, and insurance data), patient records, and machine-generated data.

**Volume**

Laney defined a key characteristic, the first V, of big data in 2001—the proposition that there is a lot of it.

Key inertia in big data adoption has historically been the associated demand for storage and its respective costs. The movement of data centers from being locally housed to existing in the cloud has quashed this concern. Better still, it comes with benefits—primarily safety and scalability. This has significantly reduced storage costs as well as providing flexibility in data storage, collaboration, and disaster recovery. The cloud, or cloud computing, refers to distributed computing delivered through the Internet.

Moore’s Law has not just affected storage capacity; it has also benefited cost. Moore’s Law is a statement made by Gordon Moore, the Intel cofounder, in 1965. Moore states that the number of transistors that are able to fit onto an integrated circuit doubles approximately every 18 months. In 1981, the price of a gigabyte of storage was US $300,000. In 2004, this was $1.00; and it was $0.10 in 2010. Today, 1 GB can be rented in the cloud for $0.023 per month, with the first year of storage free [34].

Healthcare, whether that be delivered through public health services, governments, pharmaceutical organizations, or corporates, is no stranger to big data. Data is continuously collected both digitally and non-digitally and spans a variety of patient, behavioral, epidemiological, environmental, and medical information.

Data is consumed, managed, and generated for use in patient care, transactional records, research, compliance, processes, and regulatory requirements.
As of 2020, a patient’s EHR is not readily available digitally in a secure and transparent means. Many organizations are striving to place this information on the blockchain, which revolutionizes access to health data through creating decentralized network systems.

Patients and healthcare organizations are adopting smartphones, wearables such as smartwatches and fitness trackers, home sensors, intelligent personal assistants, and social networks within their environments. Data is plentiful, accessible, and determined to be reliable. However, it is currently not used by most healthcare providers, highlighting the following:

- Patient data generation happens outside of the healthcare system.
- Patient-generated data needs to be integrated and explored within a more extensive patient record or health timeline.
- Patient data extends beyond medical markers to more holistic markers including behavioral and environmental health.
- There is an enormous interest for healthcare organizations still to be realized from opportunity and cost benefits.

A recent initiative in the United Kingdom saw the prevalence of type 1 diabetes in children geographically mapped through the use of social media hashtags #facesofdiabetes and #letstalk. The sentiment of users was also profiled, with appropriate support provided to users through the use of machine learning tools to classify post types [35].

People-powered data and its use in near real time is the holy grail of personalized care. Datafication is well under way in the finance, energy, and retail sectors. With the prevalence of the health wearables and mobile healthcare, every individual and organization has the potential to tap into this growing data volume.
Coping with Data Volume

For big data to work for your organization, it is vital that storage is efficient and cost-effective. Choosing the most appropriate storage solution is dependent on several factors:

- **Approach**: Data storage costs increase alongside storage requirements. Hence, determining whether specific data is to be collected or whether a catchall approach to data is required depends on the project and regulatory requirements. Low-cost cloud storage providers make catchall approaches feasible. Beginners to data science and machine learning may wish to determine a set of key data metrics for evaluation for the sake of simplicity.

- **Types of data**: Consider whether structured or unstructured data is to be stored—and the source—for example, audio or video content, text, images, and so forth. Video, audio, or image storage will require more resources compared to text.

- **Deployment**: Determine how the solution will be deployed, which could include locally, internally, or cloud based.

- **Access**: Determine how the solution will be accessed—through an app, web interface, intranet, and so forth.

- **Operations**: This includes data structure, architecture, archiving, recovery, event logging, and legal requirements.

- **Future use**: This involves planning for further sources of data, extensions to the system, and potential future use cases.

The distributed nature of datasets has meant that inertia in the adoption of big data has previously referred to the fragmented nature of such systems.

Variety

The second V of data is variety. This refers not only to variations in the types of data but also sources and use cases. Twenty years ago, we used to store data in the form of spreadsheets and databases. Today, data may be in the form of photographs, sensor data, tweets, encrypted files, and so on. This variety of unstructured data creates problems
for storage, mining, and analyzing data. This is also an area that machine learning can significantly assist humans in.

Data is available in the form of structured data (in the form of clinical records and results), unstructured data (e.g., in the form of communication and interaction data), and also semi-structured data (e.g., an X-ray with written annotations).

The ever-increasing digitalization of services and the broad adoption of wearables and cyber-physical devices are continuously enabling new and exciting sources of data for discovery and analysis. This enables the creation of new risk models powered by sensor data, clinical records, communication and engagement, demographics, and billing records—which can make accurate predictions and save precious time and resources.

The Internet of Things

The Internet of Things refers to the rapidly increasing variety of smart, interconnected devices and sensors and the volumes of data they generate (Figure 2-3).

Figure 2-3. Evolution of the Internet of Things

Data is transferred between devices, services, and, ultimately, people. You may also hear similar terms such as the Industrial Internet of Things and the Healthcare Internet of Things, which refer to their respective vertical, or industry, within the Internet of Things.
Today, a variety of devices monitor every sort of patient behavior—from glucose monitors to fetal monitors to electrocardiograms to blood pressure monitors. Many of these measurements require a follow-up visit with a healthcare professional.

Smarter monitoring devices communicating with other devices, services, and systems could greatly refine this process, possibly lessening the need for direct clinical intervention and perhaps replacing it with a phone call from a nurse. Current innovations in smart devices include medication dispensers that can detect whether the medication has been taken—with data transmission occurring via Bluetooth. If it is determined that the user has not taken their medication, they are contacted via telephone to discuss and encourage adherence to their medication regime. There are apps used to detect skin cancer and provide urine analysis services. There are a wealth of opportunities enabled by the Internet of Things to not only improve patient care but to reduce healthcare costs at the same time.

A variety of big data includes traditional data sources as well as newer sources of both structured and unstructured data. As the variety of data grows, so too does the potential of what algorithms and machine learning tools deployed in healthcare can achieve.

As with a variety of sources, variety in big data can come in the form of the following:

- Data types: Text, numbers, audio, video, images, and so forth.
- Function: Use cases and user requirements.
- Value of data: Is the data fit for a purpose? This would be a focus on quality in the context of the data’s use. More data does not necessarily mean better data.

Just as there are a variety of data sources, values, types, and use cases, there are also a variety of applications for data. These include access points (i.e., Web, mobile, SaaS [software as a service] also sometimes known as MDaaS [medical device as a service], API [application programming interface]) and users (typically either humans or machines).

Data variety is increasing exponentially. As new healthcare IoT devices develop, it will become imperative to identify useful data features from noise.
Legacy Data

Legacy data, whether computer based or non-computer based, can also be used in your projects. An increasing number of services enable digitalization for natural language processing or classification purposes. By using legacy data, stakeholders can maximize the insights from underutilized data. Legacy, or traditional, data may also be referred to as little or small data, discussed later in this chapter.

Typically, legacy data is fragmented and incomplete, particularly for the criteria desired today. For example, many records from more than 10 years ago may not have an associated email address or correct phone number.

Velocity

The third V of big data is velocity, referring to the speed at which data is created, stored, and prepared for analysis and visualization. The velocity of data imposes unique demands on underlying computer hardware infrastructures.

A benefit of cloud computing has been the ability to quickly store and process the volume and variety of big data that would typically overwhelm a traditional server. Cloud computing is the preferred route for big data projects due to flexibility regarding storage and cost. Cloud providers can store petabytes of data and scale up thousands of servers in real time to requirements. Perhaps more valuable is that computational power is also inexpensive and distributable.

After the Haiti earthquake in 2010, Twitter data was a quicker way of detecting and tracking the deadly cholera outbreak compared to traditional methods. A subsequent research study determined that social media platforms outperformed official methods of monitoring disease prevalence in both speed and accuracy of detecting the progress of cholera [36].

In the big data era, data is created in real time or near real time. The ubiquity of cyber-physical devices, embedded sensors, and other devices means that data transfer can occur at the moment it is created, bar a MAC address and Internet access.

The speed at which data is created is unimaginable. It is widely understood that the data generated in the previous 2 years is greater than the data created from the start of time up until then [37]. The challenge organizations have is to cope with the enormous speed at which data is created and consumed in real time.
As the penetration of wearable devices and sensors continues, so too will their application within clinical healthcare (Figure 2-4). The key to maximum clinical value is in the integration of various and heterogeneous data sources.

| Data Type                     | Example                                                      | Characteristics                                                                                     |
|-------------------------------|--------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Patient behavior and sentiment| Social media, Smartphones Web forums                        | Most data is unstructured or semi-structured. Data is large and in real time. Opportunities for learning, engagement, and understanding. Typically open source. |
| Patient health data           | Sensors Blood glucose meters, Smartphones Fitness trackers Images | Owned by patient. Data is typically device reported. Storage requirements typically low. Image storage requirements vary depending on data required for storage. Sensor data is high velocity. Structured data enables ease of classification and pattern detection. Some items may require digitalization. |
| Pharmaceutical and R&D data  | Clinical trial data, Population data                        | Encompasses population and disease data owned by pharmaceutical companies, research, academia, and the government. |
| Health data on the Web        | Patient portals (Diabetes.co.uk) Digital interventions (e.g., Low Carb Program, Gro Health, Hypo Program) | Useful metadata. Typically device reported rather than clinical. |
| Clinical data                 | EHR Patient registries, Physician data Images (scans and others) | Structured in nature. Owned by providers.                                                          |
| Claims, cost, administrative data | Claims data Expenses data                                    | Structured in nature. Encompasses any datasets relevant for reimbursement.                         |
| Transactional data            | Health information exchanges                                 | May require preprocessing to become useful data.                                                    |
The traditional three Vs give an insight into the scale of data and the rapid speeds at which these vast datasets grow and multiply. However, data variety only begins to scratch the surface of the depth and challenges of big data.

The power of big data was best demonstrated by monolith Google in 2009 when it was able to track the spread of influenza with only a one-day delay across the United States, faster than the Centers for Disease Control and Prevention (CDC), through the analysis of associated search terms [38].

However, in 2013, Google Flu Trends got it wrong—which spurred questions on the concepts of value and validity of data [39].

In a clinical setting, data value and veracity would inherit an elevated position, as clinical decisions are made on data with lives at risk and go far beyond transactional implications.

**Value**

Value refers to the usefulness of data. Key to building trust in your data is ensuring its accuracy and usability. In healthcare, this would be evaluated through qualitative and quantitative analysis. Value can be circumscribed through a variety of factors, primarily clinical impact on patient outcomes or behavior modification, patient and stakeholder engagement, impact on processes and workflows, and monetization (cost saving, cost benefits, etc.).
McKinsey stated that the potential annual value of big data to US healthcare is $300 billion. They also mention that big data has a potential annual value of €250 billion to Europe’s public sector administration [40].

Data’s value comes from the analyses conducted on it. It is this analysis of data that enables it to be turned into information, with the aim of eventually turning it into knowledge. The value lies in how organizations use available data and become information centric and data driven for purposes of decision-making. Value is not found; it is made. The key is to make the data meaningful to your users and your organization by managing it well.

Google did a perfect job in tracking the searches for flu symptoms in 2013. However, it had no way of knowing that its results weren’t valid regarding predicting flu prevalence. Google Flu Trends, at least in this instance, wasn’t providing valuable results for this use case. To add insult to injury, Google was never to know that real-world cases of flu failed to match the search distribution and frequency—so it could never do anything about it. Perhaps if this project was to be extended, analyzing social media posts might have been more valuable in determining flu prevalence. Social media and wider unstructured data sources provide a wealth of often unexplored information.

Identifying a population at risk of type 2 diabetes, predicting atrial fibrillation (AFIB), assisting patients with recovery from health events, and recommending the latest evidence-based medicine therapy to a patient are tasks with a clear and valuable return on investments.

Veracity

Worthless data, no matter how plentiful, is always going to be worthless. Veracity refers to data truthfulness and whether data is of optimal quality and suitability in the context of its use relating to the biases, noise, and abnormality in data.

Various factors can affect the veracity of data, including the following:

- Data entry: Was data entered correctly? Have there been any errors or events? Is there an audit trail of data entry?
- Data management: What is the integrity of the data moving through the system?
- Integration quality: Is data appropriately referenced, commented on, and unique?
The six Cs of trusted data can help assess whether the data you are using for your projects has veracity. It is essential to ensure your datasets are clean, complete, current, consistent, and compliant.

- **Clean data**: This is the result of good data quality procedures such as deduplication, standardization, verification, matching, and processing. Clean data enables robust decisions to be made on quality,

- **non-tainted data. Clean data is the biggest challenge for user trust. Clean data is the product of thorough data preparation.**

- **Complete data**: This is the result of consolidated data infrastructures, techniques, and processes that enable robust decision-making.

- **Current data**: Fresh data is normally considered more trustworthy over stale data. The question then becomes, at what point is data not current? As real-time, current data grows, deciphering the signal from noise can become a more challenging task.

- **Consistent data**: Data must be consistent (thus enabling machine readability). This is required for cross-compatibility of systems and applies to metadata also.

- **Compliant data**: Compliance regulations can come from various sources such as stakeholders, clients, legal legislation, or as the result of a new policy. Although many data infrastructures and standards exist, data science as a category is still coming up with first-time problems for which governance and legislation are required. Compliance can mean different things to internal and external stakeholders. Internally, there will be standards to ensure data is compliant with quality, security, and privacy procedures. All stakeholders need to trust that data has been
accessed and distributed by internal and external regulations. Typically, organizations may have a governance board for data and information.

The sixth C is **collaborative data**, which refers to collaboration over data to ensure that data management and business management goals are aligned. The sixth C refers more to the approach to data than the data itself. While all of the Cs are required for trusted data, the purpose is to have clean data that represents the real world with ideally no to little bias.

While most advocate complete veracity in data for clinical care, the gold standard of clinical care is harder to achieve in real-world scenarios with such a varied input of data. One would question that if data isn’t being used effectively already, why there would be a need to add even more. Enforcing data quality may be considered too difficult to achieve cost-effectively. In this case, training an intelligent system to learn to estimate for parameters not seen may be of use. Veracity can refer to ensuring sizable training samples for rich model building and validation, which empowers whole-population analytics.

Veracity can be applied to the analyses performed on the data, ensuring they are correct. Not only does data need to be trustworthy but so too do the algorithms and systems interpreting it.

Data entry can be a risky point that goes unnoticed without good data science. This is typically a human-based problem. Even in small data settings, humans make mistakes. To be clinically relevant, data requires maximum veracity. As the veracity of data improves, machine learning on this data enables more veracious conclusions.

**Validity**

Similar to big data veracity, there is also an issue of data validity, referring to whether data is correct and accurate for the intended use. In clinical applications, validity may be considered to be the prioritized V—to ensure that only useful and relevant data is used.

Truthfulness or veracity of data is absolute, whereas validity is contextual. Valid data represents the real world without bias.
Variability

Big data is variable. Variability defines data where the meaning is regularly changing. Variability is very relevant in performing sentiment analysis. For example, in a series of tweets, a single word can have a completely different meaning.

Variability is often confused with variety. To illustrate, a florist may sell five types of roses. That is variety. Now, if you go to the florist for two weeks in a row and buy the same white rose every day, each day it will have a subtly different form and fragrance. That is variability.

To perform proper sentiment analysis, algorithms need to be able to understand the context of texts and be able to decipher the exact meaning of a word in that particular context. This is still very difficult, even with progress in natural language processing abilities.

Visualization

Visualization is the final of the eight Vs, referring to the appropriate analyses and visualizations required on big data to make it readable, understandable, and actionable.

Visualization may not sound complex; however, overcoming the challenge of visualizing complex datasets is crucial for stakeholder understanding and development. Quite often, it is data visualization that becomes the principal component in transferring knowledge learned from datasets to stakeholders.

Massive Data

The term massive data is applied to datasets that are enormous collections of records. EHRs would be considered to be massive data. Massive datasets may be simple two-dimensional relational tables. The resources required for massive data surpass simple spreadsheet analysis and are computationally intensive due to the enormous matrices involved in computation.
Small Data

Little data (or small data) is a direct contrast to big data. Big data is distributed, varied, and comes in real time, whereas small data is data that is accessible, informative, and actionable as a result of the format and volume.

Examples of small data include patient medical records, prescription data, biometric measurements, a scan, or even Internet search histories. In comparison to organizations and services such as Google and Amazon, the amount of data is far smaller.

Analytics has gone through a rapid evolution spurned by the adoption and demands of big data. Going back to traditional datasets and applying more modern techniques such as machine learning should not be overlooked and is a great place for a data scientist to start. It’s not the size of the data; it’s what you do with it that counts.

Metadata

Metadata is data about data—which is descriptive data about each asset, or individual piece, of data. Metadata provides granular information about a single file supporting the facility to discover patterns and trends from metadata as well as the data it is supporting.

Metadata gives information about a file’s origin, date, time, and format; and it may include notes or comments. It is key to investing resources, predominantly time, in strategic information management to make sure assets are correctly named, tagged, stored, and archived in a taxonomy consistent with other assets in the collection. This facilitates quicker dataset linkage and maintains consistent methodology for asset management to ensure files are easy to find, retrieve, and distribute.

Big data hype has overshadowed metadata. However, with big data comes big metadata, enabling organizations to generate knowledge and leverage value. For example, Google and Facebook use taxonomy languages such as Open Graph [41] that help create a more structured Web, enabling more robust and descriptive information to be provided to users. This, in turn, provides human-friendly results to users, optimizing click-through and conversion.

Metadata can be very useful in any data-led project. For instance, a machine learning algorithm could use the metadata belonging to a piece of music, rather than (or alongside) the actual music itself, to be able to suggest further relevant music. Features of the music such as genre, artist, song title, and year of release could be obtained from metadata for relevant results.
Healthcare Data: Little and Big Use Cases

Healthcare stakeholders understand they are surrounded by masses of data from patients, professionals, and transactions. It is key to know how to drive value and meet KPIs (key performance indicators). The following are a selection of exciting healthcare data use cases.

Predicting Waiting Times

In Paris, France, four of the hospitals that comprise the Assistance Publique-Hôpitaux de Paris (AP-HP) teamed up with Intel and used data from internal and external sources, including 10 years of hospital admissions records, to determine day- and hour-based predictions of the number of patients expected to enter their facility [42].

Time series analysis techniques were used to predict admission rates at different times. This data was made available to all surgeries and clinics, and demonstrates an immediate way data could be used to improve efficiency and empower stakeholders.

Most, if not all, clinics across the world have access to similar data, which demonstrates just how healthcare is only beginning to scratch the surface of data’s application.

Reducing Readmissions

The same approach to waiting times can be used to help manage hospital costs. Using data analytics, at-risk patient groups can be identified based on medical history, demographics, and behavioral data. This can be used to provide the necessary care to reduce readmission rates.

At the UT Southwestern hospital in the United States, EHR analytics led to a drop in the readmission rate of cardiac patients from 26.2% to 21.2% through successful identification of at-risk patients [23].

Predictive Analytics

The preceding examples could very well use static data (i.e., non–real-time small data) and be fairly accurate in predicting waiting times and readmission intervals. The same concept of analyzing data can be used at scale for prediction of disease and the democratization of care.
In the United States, Optum Labs has collected the EHRs for over 30 million patients, creating a database for predictive analytics tools to improve the delivery of care. The intention is to enable doctors to make data-driven, informed decisions with proximity and therefore improve patients’ treatment [44].

The robustness that 30 million health records provide allows models to be trained and validated to find people who fit predictive risk trends for certain diseases such as hypertension, cardiovascular disease, stroke, type 2 diabetes, heart disease, and metabolic syndrome.

By analyzing patient data including age, social and economic demographics, fitness, and other health biomarkers, providers can improve care at both an individual and population level through not only predicting risk but through the delivery of treatments for optimal patient outcomes.

**Electronic Health Records**

EHRs haven’t quite come to fruition as yet. The idea is theoretically simple: that every patient has a digital health record consisting of their details, demographics, medical history, allergies, clinical results, and so forth. Records can be shared, with patient consent, via secure computer systems and are available for healthcare providers from both public and private sectors. Each record comprises one modifiable file, which means that doctors can implement changes over time with no danger of data replication or inconsistencies.

EHRs make perfect sense; however, complete implementation across a nation is proving a task. In the United States, up to 94% of eligible hospitals use EHRs according to HITECH research. Europe is further behind. A European Commission directive has set the task of creating a centralized European health record system by 2020 [46].

In the United States, Kaiser Permanente has implemented a system that shares data across all their facilities and made it easier to use EHRs. A McKinsey report highlighted how the data sharing system achieved an estimated $1 billion in savings as the result of reduced office visits and lab tests. The data sharing system improved outcomes in cardiovascular disease [47].

The EHR is evolving into the blockchain, which seeks to decentralize and distribute access to data.
Value-Based Care/Engagement

No longer are patients considered passive recipients of care. Healthcare is now obliged to engage patients in their health, healthcare decision-making, care, and treatment. Engagement can be maintained through digital means. Note that patient engagement is not to be confused with patient experience, which is the pathway (journey) a patient may take.

Financial drivers have already seen healthcare practices becoming increasingly incentivized to best engage with each patient, to ensure services received are satisfactory and of quality.

One of the key drivers of data-driven solutions is the demand for patient engagement and the transition toward value-based care. Better patient engagement enhances trust between patients, treatment providers, and bill payers. Moreover, it leads to better health outcomes and cost savings (or some other benefit) to the provider.

Pioneering health insurance initiatives seek to engage patients with the goal of better health outcomes through entwining premiums to good health. Health innovators at Diabetes Digital Media in Warwick, England, are working with global insurance providers and health organizations to provide scalable digital health technologies for users to optimize their health and well-being [18].

Blue Shield of California is improving patient outcomes by developing an integrated system that connects doctors, hospitals, and health coverage to the patient’s broader health data to deliver evidence-based, personalized care [48]. The aim is to help improve performance in disease prevention and care coordination.

Healthcare IoT: Real-Time Notifications, Alerts, Automation

Millions of people use devices that datafy their lives toward the quantified self. Devices connected to the Internet currently include weighing scales; activity monitors (such as Fitbit, Apple Watch, Microsoft Band) that measure heart rate, movement, and sleep; and blood glucose meters, all of which send metrics in real time and track user behavior in up-to-the-second fashion. There is even a fetal heartbeat wearable [50].

The data recorded could be used to detect the risk of disease, alert doctors, or request emergency services depending on the biometrics received. Many integrated devices are going beyond pulse and movement to measure sweat, oxidation levels, blood glucose, nicotine consumption, and more.
With sophisticated devices come sophisticated solutions to novel problems. For instance, now that heart rate monitoring is cheaper and more pervasive, conditions such as atrial fibrillation can be detected far easier and far earlier than ever before. Fluttering of one’s heart rate above 300 bpm (rather than 60–80 bpm) could be a symptom of atrial fibrillation. Patients diagnosed with the condition are 33% more likely to develop dementia; and more than 70% of patients will die from a stroke [51, 52]. The treatment can often be simple: anticoagulants and blood thinners, which are up to 80% effective.

Healthcare providers are moving slowly toward sophisticated toolkits to utilize the massive data stream created by patients and to react every time the results appear disturbing. This adoption is being driven by established digital organizations and start-ups in conjunction with health bill payers.

In another use case, an innovative program from the University of California, Irvine, gave patients with heart disease the opportunity to return home with a wireless weighing scale and weigh themselves at regular intervals. Predictive analytics algorithms determined unsafe weight gain thresholds and alerted physicians to see the patient proactively before an emergency readmittance was necessary [31].

What is interesting to recognize is that these connected health devices do not necessarily motivate the audiences or populations most at risk of adverse health events. According to several randomized trials, Fitbit wearers do exercise more, but not enough to guarantee weight loss and improved fitness.

In fact, some studies have determined they can be demotivating [55]. Then there is the question of accuracy. A Cleveland Clinic study found that heart rate monitors from four brands on the market were reporting inaccurate readings 10–20% of the time, which demonstrates that there is some way to go in the precision of the technology [56].

While these devices show potential, there is still the problem of an average abandonment rate of more than 30% after a period of engaged usage. The software application is now the channel for engagement, used as an intelligent layer on top of devices and IoT, where they can also introduce behavior change psychology for sustainable behavior change and usage.

 Offering users of devices such as these tangible incentives like discounts on health or life insurance could become more mainstream and drive prevention of many chronic lifestyle-related diseases similar to incentivized car insurance with embedded black box sensors.
Real-time alerting can also be used to notify patients of adverse effects of medications they are prescribed. Currently, patients wouldn’t receive any notification unless they were registered with the treatment provider. Healthcare providers can use public feeds to notify their patients of potential adverse effects. Emails or text messages would suffice as a form of engagement and could provide as much instruction as required, reducing time with the clinician.

EHRs can also trigger warnings, alerts, and reminders for when a patient should get a new lab test or track prescriptions to see if a patient has been following treatment instructions.

**Movement Toward Evidence-Based Medicine**

Evidence-based medicine is a term denoting treatment given based on proven scientific methods in pursuit of the best possible outcomes. Clinical trials work on a small scale, testing new treatments in small groups with internal validity (i.e., no other conditions or concerns other than those specified) and looking at how well treatments work and establishing if there are any side effects.

With growing datafication, there is also increasing “real-world evidence” or data, which can be analyzed at an individual level to create a patient data model and aggregated across populations to derive larger insights around disease prevalence, treatment, engagement, and outcomes. This approach improves quality of care, transparency, outcomes, and value and, at its core, democratizes healthcare delivery.

Using data for similar groups of patients, we can understand the treatment plans patients with similar profiles have taken, allowing the best treatment plan to be recommended based on how others in the population responded to any given treatment. The recommended treatment pathway would be best for the patient and could be explained to both the patient and healthcare professional as to why it was the best pathway to follow. This has roots in predictive analytics and goes beyond identification of populations and treatments.

By mining real-world patient data, including clinical patient records and real-time, personal, patient health data—and with demographic and health data on disease prevalence, treatment pathways, and outcomes—we can learn about optimal treatment plans for individuals, facilitating precision medicine, which is the pinnacle of evidence-based medicine.
By connecting real-world patient and clinical data to the genome, it can be personalized down to the genetic makeup of patients and populations for the following:

- Prescribing of medications
- Adverse effects and reactions
- Prevention strategies
- Likelihood risk of future diseases

Public Health

Analysis of disease patterns and outbreaks allows public health to be substantially improved through an analytics-driven approach.

Big data can assist in determining needs, required services, or treatments; and it can predict and prevent future crises to benefit the population. By mapping patient location, it would be possible to predict outbreaks, such as influenza, that could spread within an area, making it easier to formulate plans for dealing with patients, vaccinations, and care delivery.

In West Africa, mobile phone location data proved invaluable in tracking the spread of the population, and as a result, helped to predict the Ebola virus’s expanse [57].

After the Haiti earthquake in 2010, a team from Karolinska Institute in Sweden and Columbia University in the United States analyzed calling data from 2 million mobile phones on the Digicel Haiti network [58]. Phone records were used to understand population movements and for the United Nations to allocate resources more efficiently. The data was also used to identify areas at risk of the subsequent cholera outbreak.

Evolution of Data and Its Analytics

Proximity is driving the future. As people, patients, or agents in a digital world, we expect a relationship regarding nearness in time between data that is given and feedback that is received. As the scale of data increases, so too does the demand that the time between data creation and insight and action be reduced. There is substantial economic value in reducing the time from detected event to automatic response. This could help spot fraudulent transactions before a transaction completes or even serve information matched to an interest or need.

Just as data has evolved, so too has the analytics of data (Figure 2-5).
As a concept, the analysis of data, or analytics, has gone through several iterations, driven by the demand for access to data to drive decision-making.

Analytics 1.0 took the form of traditional analytics on small, slower-velocity datasets, often in the form of spreadsheets and static documents. This is suitable for little data. Typical analytics were descriptive and report based.

Data would usually be structured in nature. This drove business intelligence in the 1990s with predefined queries and detailed or historic views—typically driven by web dashboards leveraging structured data like customer data, behavioral data, sales data, patient records, and so forth. Data about production processes, sales, interactions, and transactions were collected, aggregated, and analyzed within traditional relational databases. In the Analytics 1.0 era, data scientists spent much of their time preparing data for analysis rather than performing analytics themselves.

Analytics 2.0 marked the integration of big data with traditional analytics, with interfaces to support the querying of big datasets in real time. Big data analytics began in the 2000s, which facilitated competitive insight through complex queries combined with predictive views leveraging both structured and unstructured data such as social media, behavioral data, and the growing variety of user data. Big data that couldn’t fit or be analyzed fast enough on a centralized platform would be processed with frameworks such as Hadoop, an open source software framework for fast batch data processing across parallel servers both in the cloud or on-premises.

Figure 2-5. The evolution of big data and its analytics
Unstructured data would be supported through more than just SQL (Structured Query Language) databases, supporting key and value, document, graph, columnar, and geospatial data. Other big data technologies introduced during this period include “in-memory” analytics for fast analysis whereby data is managed and processed in memory rather than on disk.

Today, vast amounts of data are being created at the edge of the network, and the traditional ways of doing analytics are no longer viable. Organizations that engage with consumers—whether they make things, move things, consume things, or work with customers—have increasing amounts of data on those devices and activities. Each device, shipment, and person leave a trail referred to as a data exhaust.

Analytics 3.0 marked the stage of maturity in which organizations realized measurable business impact from the combination of traditional analytics and big data. The approach supports agile insight discovery and real-time analysis at the point of decision. Analytics 3.0 is mostly a combination of conventional business intelligence, big data, and the IoT distributed throughout the network. Organizations can analyze those datasets for the benefit of particular audiences to achieve their objectives—whether that’s improving health outcome or monetization. They also can embed advanced analytics and optimization in near real time into every product or service decision made on the front lines of their operations.

**Turning Data into Information: Using Big Data**

Data’s value is demonstrated in the ability to convert it into information that can drive actionable insight to drive behaviors and workflows. Data analysis/analytics is a subset of data science that seeks to draw insights from sources of raw data.

Today’s advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, and determine which gene is most likely to be responsible for specific diseases and, of course, which ads you are most likely to respond to on Facebook.

Analysis of big data can help organizations understand the information contained in their data as well as identifying data most important to business objectives, outcomes, and decisions. Stakeholders typically want the knowledge that comes from analyzing data, which is determined based on the type of results produced and the reasoning approach.
The majority of big data is unstructured. IBM estimates that 80% of the world’s big data is unstructured, which includes photos, audio, video, documents, and interactions [59].

Data needs to be processed to be of value. For this to take place, resources are required—in the form of time, talent, and money. Herein is one of the inertias in adopting big data approaches: is processing the data into something with usable insight worth the capital and risk of doing so?

There are four streams within the data science analytics taxonomy (Figure 2-6). The classifications are based on the types of results that are produced.

![Figure 2-6. Streams of analytics](image)

**Descriptive Analytics**

Descriptive analytics brings insight to the past, focusing on the question of what happened, through analyzing data from history. Descriptive analytics uses techniques such as data aggregation and data mining to provide historical understanding. There is much to be learned from descriptive analytics.

Common examples of descriptive analytics are reports that provide the answers to questions such as how many patients were admitted to a hospital last July, how many were readmitted within 30 days, and how many caught an infection or suffered from a mistake in care. Descriptive analytics serves as the first form of big data analysis.
This basic level of reporting is still lacking for many organizations. Proprietary data standards maintain unhelpful pits of information; a lack of available human expertise or proper organizational buy-in can leave data collecting and unused.

Data can become more valuable through dataset linking. For instance, linking physician appointments to hospital admissions, to education attendance, to medication adherence, and to access to services can provide substantial insight and understanding for health delivery optimization and cost savings.

Limitations of descriptive analytics are that it gives limited ability to guide decisions because it is based on a snapshot of the past. Although this is useful, it is not always indicative of the future.

### Diagnostic Analytics

Diagnostic analytics is a form of analytics that examines data to answer the question of why something happened. Diagnostic analytics comprises techniques such as decision trees, data discovery, data mining, and correlations.

The next two types of analytics, predictive and prescriptive analytics, require the use of learning algorithms and are covered in more depth in following chapters.

### Predictive Analytics

Predictive analytics allows us to understand the future and predict the likelihood of a future outcome. Predictive analytics uses the data that you have at your disposal and attempts to fill in missing data with best-guess estimates.

Predictive analytics is characterized by techniques such as regression analysis, multivariate statistics, data mining, pattern matching, predictive modeling, and machine learning. Predictive analytics uses historical and current data to forecast the likelihood of future events or actionable outcomes.

Predictive analytics skills are in demand as healthcare providers seek evidence-based ways to reduce costs, take advantage of value-based reimbursements, and avoid the penalties associated with the failure to control chronic diseases and adverse events that are within their power to prevent. Predictive analytics is in the midst of disruption. The last 5 years of innovation has made significant progress with measurable impacts to the lives of patients. Wearable technology and mobile apps now make it possible to spot conditions such as asthma, atrial fibrillation, and COPD [60].
Predictive analytics remains elusive, as it requires access to real-time data that allows near–real-time clinical decision-making. To support this, medical sensors and connected devices must be fully integrated to provide up-to-the-moment information on patient health.

Besides, clinicians need to be experienced in using such data. As well as using the individual patient’s data, more accurate diagnoses and treatments must draw from as much patient data as is available, including population-level data.

As data sources and technologies develop, advanced decision support from cognitive computing engines, natural language processing, and predictive analytics can help providers pinpoint diagnoses that may otherwise elude them. Population health management tools can highlight those most at risk of hospital readmission, at risk of developing costly chronic diseases, or responding adversely to medication.

Use Case: Realizing Personalized Care

Metabolic health has been demonstrated to influence the risk of a variety of health conditions, including type 2 diabetes, hypertension, some dementias, and some cancers. While acting as a digital intervention for type 2 diabetes, the Low Carb Program app used a variety of features in the form of health biomarkers and demographics including blood glucose, weight, gender, and ethnicity to score a patient’s metabolic health. As the app expanded to integrate with wearable health devices, increased weight and blood glucose data enabled the algorithm to extend to determine the likely risk of pancreatic cancer. This is communicated to patients with their healthcare team where relevant and is referenced through cohort data comparison [16].

Users receive supported notifications that their data does not seem to fit within the demographic that is expected of them and prompts the user to speak to their healthcare professional. As demonstrated in this example, as further biosensors and algorithms are developed, the ethical implications of predictive analytics are significant.

Use Case: Patient Monitoring in Real Time

A typical hospital ward would see ward nurses manually visit patients to monitor and confirm vital health signs. However, there is no guarantee that the patient’s health status won’t deviate, for better or worse, between the time of scheduled visits. As a result, caregivers often react to problems as the result of adverse events, whereas arriving earlier in the process can make a significant difference to patient well-being. Wireless
sensors can capture and transmit patient vitals far more frequently than human beings can visit the bedside.

Over time, the same data could be applied to predictive analytics techniques to determine the likelihood of an emergency before it can be detected with a bedside visit.

**Prescriptive Analytics**

Prescriptive analytics strives to make decisions for optimal outcomes, that is, to use all available data and analytics to inform and evolve a decision of what action to take—that is, smarter decisions. Prescriptive analytics attempts to quantify the effect of future decisions to advise on possible outcomes before decisions are made.

At its best, prescriptive analytics predicts not only what will happen but also why it will happen to provide recommendations regarding actions that will take advantage of the predictions.

Prescriptive analytics uses a combination of AI techniques and tools such as machine learning, data mining, and computational modeling procedures. These techniques are applied against input from many different datasets including historical and transactional data, real-time data feeds, and big datasets.

Prescriptive analytics performs an in-context prediction, taking into account the available evidence. For example, typical hospital readmission analytics displays a forecast of patients likely to return in the next 14–30 days. A more useful predictor may integrate the same dashboard with projected costs, real-time bed counts, available education material, and follow-up care. With additional information, clinicians can determine patients with a high risk of readmission and take actions to minimize this risk, resulting in improved patient outcomes and reduced consumption of hospital resources. Population management approaches could consider patients who are obese, adding filters for factors such as high cholesterol or heart disease risk to determine where to focus treatment and which treatments to target. Pharmaceutical companies can use prescriptive analytics to expedite drug development through identifying audiences that are most suitable for clinical trials. This includes patients presumed to be compliant and those not expected to have complications.

Prescriptive analytics enables the optimization of population health management through identification of appropriate intervention models for individuals from risk-stratified populations—combining clinical, patient, and wider available health data.
Analytics is undergoing a further transformation, from prescriptive analytics through to contextual analytics. Contextual analytics considers wider data—such as environmental, location, and situational data—when prescribing the best course of action.

**Use Case: From Digital to Pharmacology**

Different treatment pathways work differently for different people. Collected on a large scale, physicians can prescribe therapies based on the latest evidence-based and predictive and prescriptive analyses of patient populations.

Incorporating digital education as well as pharmacological therapies, delivery of treatments through such resources democratizes healthcare access by ensuring the most precise treatment is always prescribed, with trackable adherence and progress insight alongside ensuring resources are focused where required.

**Reasoning**

A system can determine conclusions with information available in a knowledge base using three main methods: deduction, induction, and abduction. Although a background in logic is not required, it is useful to understand the differences between the different reasoning approaches. Any data scientist should be familiar with these reasoning techniques.

**Deduction**

Deductive reasoning allows you to make statements that are necessitated by facts that you know. For example, you are given two facts:

1. It rains every Saturday.
2. Today is Saturday.

Deductive reasoning allows you to determine the true statement that it is going to rain today if today is Saturday.

In deductive reasoning, one infers a proposition q, which is logically sensical from a premise p. Most reporting systems and business intelligence software are deductive.
**Induction**

Inductive reasoning enables you to make statements based on the evidence you have accumulated until now. However, the key point is that evidence is not the same as fact. Substantial evidence for something only suggests, however strongly, that some “thing” is a fact. Consequently, statements that are determined are only very likely to be true in an inductive approach, rather than absolute.

With inductive reasoning, one attempts to infer a proposition q, which follows from p, but which is not necessarily a logical consequence of p.

For example, if it has rained in Warwick, England, during December for a recorded history of over 50 years, you would have strong evidence (data) for the inductive statement that it will rain in the coming December in Warwick. However, this does not mean it is going to happen; it is not a fact.

Statistical learning is about inductive reasoning: looking at some data, scientifically guessing at a general hypothesis, and making statements or predictions on test data based on this premise.

**Abduction**

Abduction is an adaptation of inductive reasoning. Abductive reasoning attempts to use a hypothesis p to explain a proposition q. In abduction, reasoning flows in the opposite direction to deductive reasoning. The best hypothesis, that is, the one that most effectively explains the data, is inferred to be the most probably correct one.

A classic example is the following:

- (q) Observation: When I woke up, the grass outside my window was wet.
- (p) Knowledge base including information: Rain can make the grass wet.
- Abductive inference: It probably rained at night.

There are often many up to infinite possible outcomes, so abductive reasoning is useful in determining how to prioritize explanations to investigate. Big datasets provide opportunities for induction and inference through machine learning.
How Much Data Do I Need for My Project?

Although big data is defined by the fact it is stored in distributed systems, when it comes to creating machine learning algorithms, it’s about quality and not quantity. For instance, if you have over 1,000 data points for a patient, how are you to understand the most important attributes from what is noise? Further still, is it likely that you would get better outcomes from 5,000, 10,000, or 100,000 data points? This problem is best addressed through data science and is entirely dependent on the domain and the quality of your data samples.

The future of evidence-based medicine is the realization of data-driven, personalized healthcare. Data-driven healthcare occurs when there is a synergetic partnership between healthcare providers, patients, and data obtained from all clinical documentation. This results in analytics that drives actionable insights in real time, allowing for an enhanced patient experience and streamlined clinical and administrative workflows.

Challenges of Big Data

There are several challenges in the development of big data initiatives.

Data Growth

By the very nature of big data, storage is a challenge. There is broad agreement that the size of the digital universe will double every 2 years, which equates to a fiftyfold growth from 2010 to 2020.

Most big data is unstructured, meaning that it doesn’t reside in a traditional database schema. Documents, photos, audio, videos, and other unstructured data can be difficult to search, analyze, and retrieve. Hence, management of unstructured data is a growing challenge. Big data projects can evolve as quickly as the data used within them.

Infrastructure

Big data consumes technical resources in the form of infrastructure, storage, bandwidth, databases, and so forth. The challenge is not technical so much as one of finding reliable vendors of services and support and of getting the correct economic model of remuneration.
A solution delivered from the cloud will have greater scalability, cost-effectiveness, and efficiency compared to an on-premises solution.

**Expertise**

Finding good-quality expertise (or a deep stack expertise) in the disciplines of data analysis and data science is a challenge faced by many organizations. Compared to the amount of data generated, there are very few data scientists in proportion.

**Data Sources**

A significant challenge with big data is the volume and velocity at which data is generated and delivered. Managing the plethora of incoming data sources is a task in itself.

**Quality of Data**

Data quality is certainly not a new concern, but the ability to store each piece of data generated in real time compounds the problem. Common causes of dirty data must be addressed and include user input errors, duplicate data, and incorrect data linking. As well as maintaining data, big data algorithms can also be used to help clean data.

**Security**

There are significant risks in the security and privacy of patient data. With sensitive health data, there is rightly an increased sensitivity to security and privacy concerns. The tools used for analysis, storage, management, and utilization of data are from a variety of sources, which expose the data to risk. Cybersecurity is a concern. Significant data breaches have occurred in the healthcare industry. In the United Kingdom, the Information Commissioner’s Office (ICO) fined the Brighton and Sussex University Hospitals NHS Trust after it was found that the sensitive data of several thousand people were resident on hard drives placed up for auction on eBay [61, 62].
To date, the most substantial healthcare breach recorded has been of American health insurance company Anthem [63]. The breach exposed the personal records, including home addresses and personal information, of over 70 million current and former members. Data loss, including data theft, is a genuine problem that is continuously tested to its limits.

Security challenges include authentication for every team and team member accessing datasets, restricting access based on user needs, recording data access histories, and meeting compliance regulations and ensuring data is encrypted to protect from snoopers and any malintent.

Security of medical devices poses a unique threat because of their technological diversity. Medical devices, from health applications on a smartphone to insulin pumps, are becoming increasingly networked, leaving unique openings for hackers [64]. This kind of activity is regarded as dangerous, although for those more technical, it is understood that this can be utilized for positive benefit. Insulin pumps and continuous glucose monitors have been “hacked” to function as an artificial pancreas by enthusiasts way before pharmaceutical companies or digital companies had. However, the key concern is that, if exploited, similar device vulnerabilities could lead not only to data breaches but fatalities in people relying on medical devices. Today, the balance between disruption and progress is one of real-world experience. Yesterday’s disruptors are today’s innovators.

Then there is the topic of third-party services and vendors. Most health devices and health apps offer API access. It’s useful to remember that external vendor services and APIs are only as good as the developers who have developed them. Evidenced trust and reputation is all that one truly relies on.

Resistance

Lack of understanding of value and outcomes from machine learning projects can leave internal stakeholder resistance. You can only manage a project efficiently if you can ascribe clear value to results and prioritize clear explanations of outputs to stakeholders.

Over and above inertia, once key patterns have been identified, organizations must be prepared to make necessary actions and implement changes to derive value from them. This is often a primary inertia, so precise and jargon-free documentation and governance is useful for understanding the project and governance purposes.
Policies and Governance

The vast amount of big data captured through wearable devices and sensors raises questions as to how data is collected, how data is treated, how data is analyzed, and how data should affect the policy-making process. Besides, data privacy and security are legal areas in a state of constant development and evolution.

Fragmentation

Most organizational data is highly fragmented. Hospital teams have their patient data, as do physicians and tertiary care teams. This creates challenges at several levels: syntactic (i.e., defining common formats across teams, sites, and organizations), semantic (agreeing on common definitions), and political (determining and confirming ownership and responsibilities). This is before the patient’s own data, collected through phones, health apps, wearables, and so forth.

Lack of Data Strategy

To benefit from data science investments, organizations must have a coherent data strategy. It’s key to establish the need for the data and identify the right type and sources of data for use. What is the desired goal with the data and what are the objectives? It is important to establish the data needs, to save wasting resources and efforts collecting data that are of no use.

Within a hospital emergency department, for instance, collecting the time of patient admission, triage, and consultation is required to understand how long it takes for patients to be seen, but it is of negligible use to determine quality of care or outcomes.

Visualization

To truly benefit from data science investments, stakeholders must understand the outputs of data analysis and exploration. Poor visualization of data is connected to lack of data strategy: if specific use cases and objectives are not determined, visualization of results from data can prove challenging due to a lack of clear direction.
Timeliness of Analysis

Linked to proximity, the value of data may decrease over time. Applications such as fraud detection in healthcare, just like with banking, require data to be delivered as close to real time as possible to be effective.

Ethics

Data ethics is emerging as a topic as datasets become big enough to raise practical rather than theoretical concerns about ethics. Patients are leaving a constant data trail through sensors, wearables, transactions, social media, and transportation. For instance, the genetic testing company 23andMe sells reports on genetic risk disposition for over ten diseases including Alzheimer’s disease and Parkinson’s disease. This spurns challenging conversations about what is and what is not appropriate, discussed more in Chapter 8.

Data and Information Governance

Data is a precious asset to any business or organization; but in healthcare, data governance is paramount to ensure consistency, safety, and progression of any organization toward becoming data driven and analytically driven.

The need for data governance is driven primarily by the requirement for accountability in a risk-averse industry. As the value of data is realized, data governance too is evolving to define the approach, management, and utilization of data in an organization. Processes and standards such as ISO (International Organization for Standardization) 10027 and GDPR (General Data Protection Regulation) have been used to bring the entire data industry in line with best practice.

Data governance and information governance are often used interchangeably but do differ.

Governance of data refers to the practice of managing data assets throughout their lifecycle to ensure that they meet standards of quality and integrity. The goal of data governance is to provide user trust in the data through optimal data validity and accountability. It sets the framework for creating high-quality data and using data in a secure, ethical, and consensual manner. Data governance in healthcare seeks to improve efficiency, create accountability, and establish a pool of data that providers and patients can use to make proper health decisions.
Often the best approach to data governance is to govern to the least extent necessary to achieve the most common good. In several cases, governance is being set for circumstances that don’t require governing yet.

Organizations can waste tremendous amounts of time through management and decision-making, including constraints on data, which can hinder project development and direction. Healthcare organizations should be motivated by the acquisition of more data to learn—to gather information to manage risks better and improve outcomes.

Data governance is required to uphold a variety of criteria to return quality information.

**Data Stewardship**

The steward is accountable for the data as defined and its appropriate use. The objective of this aspect of data governance is to assure quality regarding data accuracy; data accessibility; and data consistency, completeness, and updating. This is typically the organization running the project. Further still, most organizations would have an appointed officer(s), responsible for accessing the data, with elevated privileges.

**Data Quality**

Ensuring data quality is arguably the most important function of data governance besides security. Poor data quality has a detrimental impact on accuracy, particularly for learning projects. Governance should seek to enforce quality, accuracy, and timeliness of data.

**Data Security**

Security of data is essential. First and foremost, data governance sets out how the data is protected, typically how it is encrypted, who has access to it, processes for handling data—including how the data may be used—and what is done in case of a breach. This is motivated by several drivers: regulatory requirements that demand action (such as GDPR taking place in Europe), the growing risk of cybersecurity, and the fact that perceptions of digital weakness can harm reputation and trust. This is demonstrated in how the number of websites that operate with a secure https:// protocol has risen by 4,000% since 2012 [65]. Patients expect their data to be secure.
Having said that, the Garmin ransomware attack of 2020 showed that people will be patient with services that they rely on, with messages of support and jubilation as services were returned online [66]. The situation was similar with Babylon Health which shared the recordings from other GP consultations in public view and access to a minority of patient users [67].

People are aware data security is a constant concern for all involved, and the Garmin cyberattack showed people will be patient with services and brands they prefer or that give them use or value. It is noteworthy that even when founders and executives from technology companies are brought before governors, time has shown there is very little understanding of the technologies involved or intricacies of use from those responsible for questioning.

**Data Availability**

Fragmentation of data is one of the biggest problems healthcare organizations face. Ensuring stakeholders’, providers’, and patients’ data access is critical.

People “own” their data; and as a result, access needs to be transparent, quick, and efficient for all. Data access governance may set permissions and authentication levels for user roles to systems.

**Data Content**

The data and information governance for a project may state the types of data collected, including health data; metadata; and location, profile, and behavior data. Data governance may also detail what the data is used for.

**Master Data Management (MDM)**

As data transfer becomes a priority, the MDM acts as a master reference to ensure consistent use of data across organizations. The goal of MDM is to create a common point of reference that can be shared within an organization.
Use Cases

In almost all cases, data governance policies will include use cases for data access and “what would happen if” scenarios. It is useful to have all team members required in action to be involved in dummy runs to ensure that data access or access requests are not met with periods of silence.

Use cases would include the set of processes that would take place in light of a request for data from a patient, request for deletion, ensuring appropriate roles were able to access the data, ensuring appropriate transactional logs were kept, and so forth.

Data governance provides several benefits over and above those that are internal, including the following:

- Protecting the interests of data stakeholders—particularly the data “giver”
- Standardizing procedures and processes for streamlined repetition and minimization of error
- Reducing costs and improving effectiveness
- Greater transparency and accountability between data transfer parties

Information governance has a slightly different purpose. Information governance is the management and control of information. This information is formed through the use of data assets (Figure 2-7).

| EXAMPLES OF DATA                      | EXAMPLES OF INFORMATION                  |
|--------------------------------------|------------------------------------------|
| 150/80 BLOOD PRESSURE READING        | JOHN’S BLOOD PRESSURE READING ON 9/15/15 |
| DATA ON AN EMPLOYEE APPLICATION      | EMPLOYEE APPLICATION RECORD               |
| NUMBER OF HEART ATTACKS IN JUNE      | ABC HOSPITAL HEART ATTACK RATE FOR JUNE   |
| EMPLOYEE ADDRESS                     | EMPLOYEE RECORD                           |

*Figure 2-7. Data vs. information—what’s the difference?*
A single data point, such as a blood glucose reading, without any additional context or metadata, offers very little. However, a series of blood glucose readings over the past 3 months can show whether the patient has good or bad blood glucose control, for example, which can be a risk factor for comorbidities. It pays attention to personal and sensitive information linked to patients and employees.

In most organizations, data and information governance is consolidated; but it is useful to understand the distinction.

A resolute data governance program would involve a governing body or council, a defined set of procedures, and the way in which said procedures are upheld and verified.

As learning systems develop, information governance is becoming more critical. Digital health apps and integrated bio-human health devices will soon be able to predict and determine risk and time to disease as datasets grow infinitely large.

Regardless of the question of accuracy, legislation on what can and can’t happen with this information is required.

**Conducting a Big Data Project**

Getting started with any big data project involves three steps. Big data projects do not have to be part of a bigger machine learning project:

1. Understand how the project will impact your organization.
   
   This involves identifying the business case; objectives (what is it I want to achieve?); current and required infrastructures; data sources; and quality, tools, and KPIs regarding measuring success. Identifying quantitative objectives is key to measuring success and performance. Determining a measurable ROI to determine the success of the project is useful to demonstrate value to stakeholders. Identifying goals with members from technical, medical, and operational teams helps to maintain clarity of focus and direction. Identify potential use cases and create use case scenarios and personas if time allows and if desired.

2. Find the skills and technology you need.

   Identification of the use cases for the project can help define the technology, infrastructure, and capabilities required for
the project. Skills in Python and R (open source programming languages) are most likely going to be required, with analytics skills also desired for data science. This can be outsourced, contracted in, or developed into full-time roles.

3. Implement the project.

   Identify the data you wish to include and make a job of identifying data that will also be excluded from use cases. Identify the types of analyses required on the data to gather the desired output from the project. Explore the requirements necessary for collecting and preparing data into usable formats for algorithms; the governance required for holding and using the data; and how it will be presented to stakeholders and, most importantly, the end users. Architectural decisions such as database structures, platform providers, and type of models may be made at this point. It’s at this point you can consider the types of analytics applied to the data.

   You may wish to consider phasing of the project and the gaps between current and future capabilities that require addressing.

   Develop the project in a test environment and present results to users in a meaningful way.

**Big Data Tools**

There are various open source and paid-for tools that can be used for big data collection and machine learning projects. The fundamentals begin with Hadoop and NoSQL (non-relational) databases.

   Hadoop is synonymous with big data and is a Java-based, open source software framework, or software ecosystem, for distributed storage of large datasets on computer clusters. Hadoop can process big and complex datasets for insights and answers. Hadoop Distributed File System (HDFS) is the storage system used by Hadoop that separates big data and distributes it across nodes within a cluster. This can also replicate data in a cluster, providing high availability. For those not wanting to delve into Java themselves, Microsoft Azure, Cloudera, Google Cloud Platform, and Amazon Web Services all provide big data hosting services powered by Hadoop.
MongoDB is a modern approach to databases that is useful for managing data that changes frequently or data that is unstructured or semi-structured in nature. The architecture has better write performance, takes less storage space with compression, and reduces operational overhead. MongoDB is part of a growing movement known as NoSQL databases, which are used to store unstructured data with no particular schema. Rows can all have their own set of unique column values. NoSQL has been driven by the requirements to provide better performance in storing big data.

R is an open source software and programming language used for data mining and statistical analysis and graphical techniques.

Python is another popular open source programming language used in machine learning and data mining for data preprocessing, computing, and modeling.

**Conclusion**

Big data and healthcare mean significant progress toward realizing a plethora of opportunities. This includes remote patient monitoring, precision medicine, preventing readmissions, and advancing disease understanding.

From mundane tasks of data entry and record keeping to advanced and functional applications like observing and understanding people’s blood glucose patterns, data is becoming part of our day-to-day fabric.

The potential for positive impact that big data-driven solutions can bring seemingly outweighs any negative sentiment. The growing trend toward centralization of medical data, toward the custodianship of the patient, is causing concern; yet as long as privacy and security are maintained, it is likely to facilitate the development of new treatments and contribute toward the evolution of medical understanding. The healthcare industry remains well within its infancy of leveraging big data for clinical and business use. The explosion of wearables and health IoT has allowed the measurement of heart rate or steps to be taken during a day, which may seem like a relatively ordinary dataset; but the potential impact on preventative medicine and the general health of a population is immense.

Big data provides the opportunity for precision medicine through the delivery of data-driven, evidence-based medicine. The benefits are plentiful, first and foremost through the delivery of more precise treatments for patients.

With this comes cost savings through fewer mistakes, fewer hospital admissions, decreased mortality, and improved patient satisfaction. This also saves money by optimizing the use of the physician and wider healthcare team’s time, as more time is spent with patients in need.