ARTIFICIAL INTELLIGENCE BASICS

A NON-TECHNICAL INTRODUCTION

Tom Taulli

Apress®

CHAPTER

2

Data

The Fuel for AI

Pinterest is one of the hottest startups in Silicon Valley, allowing users to pin their favorite items to create engaging boards. The site has 250 million MAUs (monthly active users) and posted $756 million in revenue in 2018.[1]

A popular activity for Pinterest is to plan for weddings. The bride-to-be will have pins for gowns, venues, honeymoon spots, cakes, invitations, and so on.

This also means that Pinterest has the advantage of collecting huge amounts of valuable data. Part of this helps provide for targeted ads. Yet there are also opportunities for email campaigns. In one case, Pinterest sent one that said:

> You're getting married! And because we love wedding planning—especially all the lovely stationery—we invite you to browse our best boards curated by graphic designers, photographers and fellow brides-to-be, all Pinners with a keen eye and marriage on the mind.[2]

The problem: Plenty of the recipients of the email were already married or not expecting to marry anytime soon.

[1] www.cnbc.com/2019/03/22/pinterest-releases-s-1-for-ipo.html
[2] www.businessinsider.com/pinterest-accidental-marriage-emails-2014-9

Pinterest did act quickly and put out this apology:

> Every week, we email collections of category-specific pins and boards to pinners we hope will be interested in them. Unfortunately, one of these recent emails suggested that pinners were actually getting married, rather than just potentially interested in wedding-related content. We're sorry we came off like an overbearing mother who is always asking when you'll find a nice boy or girl.

It's an important lesson. Even some of the most tech-savvy companies blow it.

For example, there are some cases where the data may be spot-on but the outcome could still be an epic failure. Consider the case with Target. The company leveraged its massive data to send personalized offers to expectant mothers. This was based on those customers who made certain types of purchases, such as for unscented lotions. Target's system would create a pregnancy score that even provided estimates of due dates.

Well, the father of one of the customers saw the email and was furious, saying his daughter was not pregnant.[3]

But she was—and yes, she had been hiding this fact from her father.

There's no doubt that data is extremely powerful and critical for AI. But you need to be thoughtful and understand the risks. In this chapter, we'll take a look at some of the things you need to know.

Data Basics

It's good to have an understanding of the jargon of data.

First of all, a bit (which is short for "binary digit") is the smallest form of data in a computer. Think of it as the atom. A bit can either be 0 or 1, which is binary. It is also generally used to measure the amount of data that is being transferred (say within a network or the Internet).

A byte, on the other hand, is mostly for storage. Of course, the numbers of bytes can get large very fast. Let's see how in Table 2-1.

Table 2-1. Types of data levels

Unit	Value	Use Case
Megabyte	1,000 kilobytes	A small book
Gigabyte	1,000 megabytes	About 230 songs
Terabyte	1,000 gigabytes	500 hours of movies
Petabyte	1,000 terabytes	Five years of the Earth Observing System (EOS)
Exabyte	1,000 petabytes	The entire Library of Congress 3,000 times over
Zettabyte	1,000 exabytes	36,000 years of HD-TV video
Yottabytes	1,000 zettabytes	This would require a data center the size of Delaware and Rhode Island combined

Data can also come from many different sources. Here is just a sampling:

- Web/social (Facebook, Twitter, Instagram, YouTube)
- Biometric data (fitness trackers, genetics tests)
- Point of sale systems (from brick-and-mortar stores and e-commerce sites)
- Internet of Things or IoT (ID tags and smart devices)
- Cloud systems (business applications like Salesforce.com)
- Corporate databases and spreadsheets

Types of Data

There are four ways to organize data. First, there is structured data, which is usually stored in a relational database or spreadsheet. Some examples include the following:

- Financial information
- Social Security numbers
- Addresses
- Product information
- Point of sale data
- Phone numbers

For the most part, structured data is easier to work with. This data often comes from CRM (Customer Relationship Management) and ERP (Enterprise Resource Planning) systems—and usually has lower volumes. It also tends to

be more straightforward, say in terms of analysis. There are various BI (Business Intelligence) programs that can help derive insights from structured data. However, this type of data accounts for about 20% of an AI project.

The majority will instead come from unstructured data, which is information that has no predefined formatting. You'll have to do this yourself, which can be tedious and time consuming. But there are tools like next-generation databases—such as those based on NoSQL—that can help with the process. AI systems are also effective in terms of managing and structuring the data, as the algorithms can recognize patterns.

Here are examples of unstructured data:

- Images
- Videos
- Audio files
- Text files
- Social network information like tweets and posts
- Satellite images

Now there is some data that is a hybrid of structured and unstructured sources—called semi-structured data. The information has some internal tags that help with categorization.

Examples of semi-structured data include XML (Extensible Markup Language), which is based on various rules to identify elements of a document, and JSON (JavaScript Object Notation), which is a way to transfer information on the Web through APIs (Application Programming Interfaces).

But semi-structured data represents only about 5% to 10% of all data.

Finally, there is time-series data, which can be both for structured, unstructured, and semi-structured data. This type of information is for interactions, say for tracking the "customer journey." This would be collecting information when a user goes to the web site, uses an app, or even walks into a store.

Yet this kind of data is often messy and difficult to understand. Part of this is due to understanding the intent of the users, which can vary widely. There is also huge volumes of interactional data, which can involve trillions of data points. Oh, and the metrics for success may not be clear. Why is a user doing something on the site?

But AI is likely to be critical for such issues. Although, for the most part, the analysis of time-series data is still in the early stages.

Big Data

With the ubiquity of Internet access, mobile devices, and wearables, there has been the unleashing of a torrent of data. Every second, Google processes over 40,000 searches or 3.5 billion a day. On a minute-by-minute basis, Snapchat users share 527,760 photos, and YouTube users watch more than 4.1 million videos. Then there are the old-fashioned systems, like emails, that continue to see significant growth. Every minute, there are 156 million messages sent.[4]

But there is something else to consider: Companies and machines also generate huge sums of data. According to research from Statista, the number of sensors will reach 12.86 billion by 2020.[5]

In light of all this, it seems like a good bet that the volumes of data will continue to increase at a rapid clip. In a report from International Data Corporation (IDC) called "Data Age 2025," the amount of data created is expected to hit a staggering 163 zettabytes by 2025.[6] This is about ten times the amount in 2017.

To deal with all this, there has emerged a category of technology called Big Data. This is how Oracle explains the importance of this trend:

> Today, big data has become capital. Think of some of the world's biggest tech companies. A large part of the value they offer comes from their data, which they're constantly analyzing to produce more efficiency and develop new products.[7]

So yes, Big Data will remain a critical part of many AI projects.

Then what exactly is Big Data? What's a good definition? Actually, there isn't one, even though there are many companies that focus on this market! But Big Data does have the following characteristics, which are called the three Vs (Gartner analyst Doug Laney came up with this structure back in 2001[8]): volume, variety, and velocity.

---

[4] www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#788c13c660ba

[5] www.forbes.com/sites/louiscolumbus/2018/06/06/10-charts-that-will-challenge-your-perspective-of-iots-growth/#4e9fac23ecce

[6] https://blog.seagate.com/business/enormous-growth-in-data-is-coming-how-to-prepare-for-it-and-prosper-from-it/

[7] www.oracle.com/big-data/guide/what-is-big-data.html

Volume

This is the scale of the data, which is often unstructured. There is no hard-and-fast rule on a threshold, but it is usually tens of terabytes.

Volume is often a major challenge when it comes to Big Data. But cloud computing and next-generation databases have been a big help—in terms of capacity and lower costs.

Variety

This describes the diversity of the data, say a combination of structured, semi-structured, and unstructured data (explained above). It also shows the different sources of the data and uses. No doubt, the high growth in unstructured data has been a key to the variety of Big Data.

Managing this can quickly become a major challenge. Yet machine learning is often something that can help streamline the process.

Velocity

This shows the speed at which data is being created. As seen earlier in this chapter, services like YouTube and Snapchat have extreme levels of velocity (this is often referred to as a "firehouse" of data). This requires heavy investments in next-generation technologies and data centers. The data is also often processed in memory not with disk-based systems.

Because of these issues, velocity is often considered the most difficult when it comes to the three Vs. Let's face it, in today's digital world, people want their data as fast as possible. If it is too slow, people will get frustrated and go somewhere else.

Over the years, though, as Big Data has evolved, there have been more Vs added. Currently, there are over ten.

But here are some of the common ones:

- *Veracity*: This is about data that is deemed accurate. In this chapter, we'll look at some of the techniques to evaluate veracity.
- *Value*: This shows the usefulness of the data. Often this is about having a trusted source.

- *Variability*: This means that data will usually change over time. For example, this is the case with social media content that can morph based on overall sentiment regarding new developments and breaking news.
- *Visualization*: This is using visuals—like graphs—to better understand the data.

As you can see, managing Big Data has many moving parts, which leads to complexity. This helps to explain why many companies still use only a tiny fraction of their data.

Databases and Other Tools

There are a myriad of tools that help with data. At the core of this is the database. As should be no surprise, there has been an evolution of this critical technology over the decades. But even older technologies like relational databases are still very much in use today. When it comes to mission-critical data, companies are reluctant to make changes—even if there are clear benefits.

To understand this market, let's rewind back to 1970, when IBM computer scientist Edgar Codd published "A Relational Model of Data for Large Shared Data Banks." It was pathbreaking as it introduced the structure of relational databases. Up until this point, databases were fairly complex and rigid—structured as hierarchies. This made it time consuming to search and find relationships in the data.

As for Codd's relational database approach, it was built for more modern machines. The SQL script language was easy to use allowing for CRUD (Create, Read, Update, and Delete) operations. Tables also had connections with primary and foreign keys, which made important connections like the following:

- *One-to-One*: One row in a table is linked to only one row in another table. Example: A driver's license number, which is unique, is associated with one employee.
- *One-to-Many*: This is where one row in a table is linked to other tables. Example: A customer has multiple purchase orders.
- *Many-to-Many*: Rows from one table are associated with rows of another. Example: Various reports have various authors.

With these types of structures, a relational database could streamline the

But despite the advantages, IBM was not interested in the technology and continued to focus on its proprietary systems. The company thought that the relational databases were too slow and brittle for enterprise customers.

But there was someone who had a different opinion on the matter: Larry Ellison. He read Codd's paper and knew it would be a game changer. To prove this, he would go on to co-found Oracle in 1977 with a focus on building relational databases—which would quickly become a massive market. Codd's paper was essentially a product roadmap for his entrepreneurial efforts.

It was not until 1993 that IBM came out with its own relational database, DB2. But it was too late. By this time, Oracle was the leader in the database market.

Through the 1980s and 1990s, the relational database was the standard for mainframe and client-server systems. But when Big Data became a factor, the technology had serious flaws like the following:

- *Data Sprawl*: Over time, different databases would spread across an organization. The result was that it got tougher to centralize the data.
- *New Environments*: Relational database technology was not built for cloud computing, high-velocity data, or unstructured data.
- *High Costs*: Relational databases can be expensive. This means that it can be prohibitive to use the technology for AI projects.
- *Development Challenges*: Modern software development relies heavily on iterating. But relational databases have proven challenging for this process.

In the late 1990s, there were open source projects developed to help create next-generation database systems. Perhaps the most critical one came from Doug Cutting who developed Lucene, which was for text searching. The technology was based on a sophisticated index system that allowed for low-latency performance. Lucene was an instant hit, and it started to evolve, such as with Apache Nutch that efficiently crawled the Web and stored the data in an index.

But there was a big problem: To crawl the Web, there needed to be an infrastructure that could hyperscale. So in late 2003, Cutting began development on a new kind of infrastructure platform that could solve the problem. He got the idea from a paper published from Google, which described its massive file system. A year later, Cutting had built his new platform, which allowed for sophisticated storage without the complexity. At the core of this was MapReduce that allowed processing across multiple

Eventually, Cutting's system morphed into a platform called Hadoop—and it would be essential for managing Big Data, such as making it possible to create sophisticated data warehouses. Initially, Yahoo! used it, and then it quickly spread, as companies like Facebook and Twitter adopted the technology. These companies were now able to get a 360 view of their data, not just subsets. This meant there could be more effective data experiments.

But as an open source project, Hadoop still lacked the sophisticated systems for enterprise customers. To deal with this, a startup called Hortonworks built new technologies like YARN on top of the Hadoop platform. It had features like in-memory analytic processing, online data processing, and interactive SQL processing. Such capabilities supported adoption of Hadoop across many corporations.

But of course, there emerged other open source data warehouse projects. The well-known ones, like Storm and Spark, focused on streaming data. Hadoop, on the other hand, was optimized for batch processing.

Besides data warehouses, there was also innovation of the traditional database business. Often these were known as NoSQL systems. Take MongoDB. It started as an open source project and has turned into a highly successful company, which went public in October 2017. The MongoDB database, which has over 40 million downloads, is built to handle cloud, on-premise, and hybrid environments.[9] There is also much flexibility structuring the data, which is based on a document model. MongoDB can even manage structured and unstructured data at high petabyte scale.

Even though startups have been a source of innovation in database systems and storage, it's important to note that the mega tech operators have also been critical. Then again, companies like Amazon.com and Google have had to find ways to deal with the huge scale of data because of the need for managing their massive platforms.

One of the innovations has been the data lake, which allows for seamless storage of structured and unstructured data. Note that there is no need to reformat the data. A data lake will handle this and allow you to quickly perform AI functions. According to a study from Aberdeen, companies who use this technology have an average of 9% organic growth compared to those who do not.[10]

Now this does not mean you have to get rid of your data warehouses. Rather, both serve particular functions and use cases. A data warehouse is generally good for structured data, whereas a data lake is better for diverse environments. What's more, it's likely that a large portion of the data will never be used.

[9] www.mongodb.com/what-is-mongodb

For the most part, there are a myriad of tools. And expect more to be developed as data environments get more complex.

But this does not mean you should chose the latest technology. Again, even older relational databases can be quite effective with AI projects. The key is understanding the pros/cons of each and then putting together a clear strategy.

Data Process

The amount of money shelled out on data is enormous. According to IDC, the spending on Big Data and analytics solutions is forecasted to go from $166 billion in 2018 to $260 billion by 2022.[11] This represents an 11.9% compound annual growth rate. The biggest spenders include banks, discrete manufacturers, process manufacturers, professional service firms, and the federal government. They account for close to half the overall amount.

Here's what IDC's Jessica Goepfert—the program vice president (VP) of Customer Insights and Analysis—said:

> *At a high level, organizations are turning to Big Data and analytics solutions to navigate the convergence of their physical and digital worlds. This transformation takes a different shape depending on the industry. For instance, within banking and retail—two of the fastest growth areas for Big Data and analytics—investments are all about managing and reinvigorating the customer experience. Whereas in manufacturing, firms are reinventing themselves to essentially be high tech companies, using their products as a platform to enable and deliver digital services.*[12]

But a high level of spending does not necessarily translate into good results. A Gartner study estimates that roughly 85% of Big Data projects are abandoned before they get to the pilot stage.[13] Some of the reasons include the following:

- Lack of a clear focus
- Dirty data
- Investment in the wrong IT tools
- Problems with data collection
- Lack of buy-in from key stakeholders and champions in the organization

[11] www.idc.com/getdoc.jsp?containerId=prUS44215218
[12] www.idc.com/getdoc.jsp?containerId=prUS44215218

In light of this, it is critical to have a data process. Notwithstanding there are many approaches—often extoled by software vendors—there is one that has widespread acceptance. A group of experts, software developers, consultants, and academics created the CRISP-DM Process in the late 1990s. Take a look at Figure 2-1 for a visual.

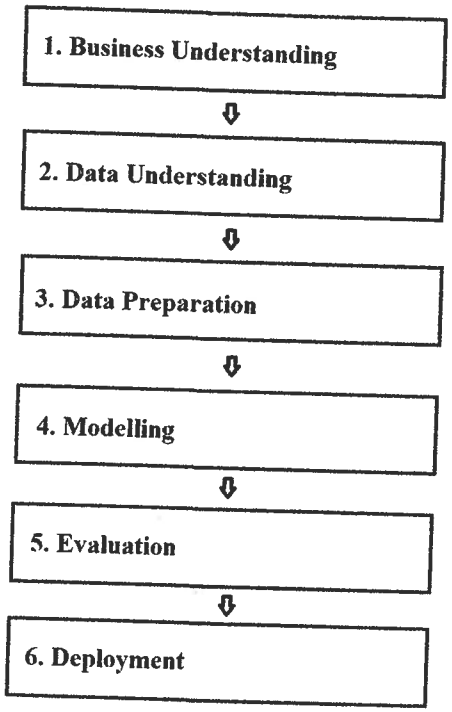

Figure 2-1. The CRISP-DM Process

In this chapter, we'll take a look at steps #1 through #3. Then in the rest of the book, we'll cover the remaining ones (that is, we will look at Modelling and Evaluation in Chapter 3 and Deployment in Chapter 8).

Note that steps #1–#3 can account for 80% of the time of the data process, which is based on the experience of Atif Kureishy, who is the global VP of Emerging Practices at Teradata.[14] This is due to factors like:

The data is not well organized and comes from different sources (whether from different vendors or silos in the organization), there is not enough focus on automation tools, and the initial planning was insufficient for the scope of the project.

It's also worth keeping in mind that the CRISP-DM Process is not a strict linear process. When dealing with data, there can be much iteration. For example, there may be multiple attempts at coming up with the right data and testing it.

Step #1—Business Understanding

You should come up with a clear view of the business problem to be solved. Some examples:

- How might a price adjustment impact your sales?
- Will a change in copy lead to improved conversion of digital ads?
- Does a fall in engagement mean there will be an increase in churn?

Then, you must establish how you will measure success. Might it be that sales should increase by at least 1% or that conversions should rise by 5%?

Here's a case from Prasad Vuyyuru, who is a partner at the Enterprise Insights Practice of Infosys Consulting:

> *Identifying which business problem to solve using AI and assessing what value will be created are critical for the success of all AI projects. Without such diligent focus on business value, AI projects risk not getting adopted in the organization. AB Inbev's experience in using AI to identify packaging line motors that are likely to fail is a great example of how AI is creating practical value. ABInbev installed 20 wireless sensors to measure vibrations at packaging lines motors. They compared sounds with normally functioning motors to identify anomalies which predicted eventual failure of the motors.*[15]

Regardless of the goal, it's essential that the process be free of any prejudgments or bias. The focus is to find the best results. No doubt, in some cases, there will not be a satisfactory result.

Or, in other situations, there may be big surprises. A famous example of this comes from the book *Moneyball* by Michael Lewis, which was also made into a movie in 2011 that starred Brad Pitt. It's a true story of how the Oakland A's used data science techniques to recruit players. The tradition in baseball was to rely on metrics like batting averages. But when using sophisticated data analytics techniques, there were some startling results. The Oakland A's realized that the focus should be on slugging and on-base percentages. With this information, the team was able to recruit high-performing players at lower compensation levels.

The upshot is that you need to be open minded and willing to experiment.

In step #1, you should also assemble the right team for the project. Now unless you work at a company like Facebook or Google, you will not have the luxury of selecting a group of PhDs in machine learning and data science. Such talent is quite rare—and expensive.

But you also do not need an army of top-notch engineers for an AI project either. It is actually getting easier to apply machine learning and deep learning models, because of open source systems like TensorFlow and cloud-based platforms from Google, Amazon.com, and Microsoft. In other words, you may only need a couple people with a background in data science.

Next, you should find people—likely from your organization—who have the right domain expertise for the AI project. They will need to think through the workflows, models, and the training data—with a particular understanding of the industry and customer requirements.

Finally, you will need to evaluate the technical needs. What infrastructure and software tools will be used? Will there be a need to increase capacity or purchase new solutions?

Step #2—Data Understanding

In this step, you will look at the data sources for the project. Consider that there are three main ones, which include the following:

- *In-House Data*: This data may come from a web site, beacons in a store location, IoT sensors, mobile apps, and so on. A major advantage of this data is that it is free and customized to your business. But then again, there are some risks. There can be problems if there has not been enough attention on the data formatting or what data should be selected.
- *Open Source Data*: This is usually freely available, which is certainly a nice benefit. Some examples of open source data include government and scientific information. The data is often accessed through an API, which makes the process fairly straightforward. Open source data is also usually well formatted. However, some of the variables

- *Third-Party Data*: This is data from a commercial vendor. But the fees can be high. In fact, the data quality, in some cases, may be lacking.

According to Teradata—based on the firm's own AI engagements—about 70% of data sources are in-house, 20% from open source, and the rest from commercial vendors.[16] But despite the source, all data must be trusted. If not, there will likely be the problem of "garbage in, garbage out."

To evaluate the data, you need to answer questions like the following:

- Is the data complete? What might be missing?
- Where did the data come from?
- What were the collection points?
- Who touched the data and processed it?
- What have been the changes in the data?
- What are the quality issues?

If you are working with structured data, then this stage should be easier. However, when it comes to unstructured and semi-structured data, you will need to label the data—which can be a protracted process. But there are some tools emerging in the market that can help automate this process.

Step #3—Data Preparation

The first step in the data preparation process is to decide what datasets to use.

Let's take a look at a scenario: Suppose you work for a publishing company and want to put together a strategy to improve customer retention. Some of the data that should help would include demographic information on the customer base like age, sex, income, and education. To provide more color, you can also look at browser information. What type of content interests customers? What's the frequency and duration? Any other interesting patterns—say accessing information during weekends? By combining the sources of information, you can put together a powerful model. For example, if there is a drop-off in activity in certain areas, it could pose a risk of cancellation. This would alert sales people to reach out to the customers.

While this is a smart process, there are still landmines. Including or excluding even one variable can have a significant negative impact on an AI model. To see why, look back at the financial crisis. The models for underwriting mortgages were sophisticated and based on huge amounts of data. During normal economic times, they worked quite well as major financial institutions like Goldman Sachs, JP Morgan, and AIG relied on them heavily.

But there was a problem: The models did not account for falling housing prices! The main reason was that—for decades—there had never been a national drop. The assumption was that housing was mostly a local phenomenon.

Of course, housing prices more than just fell—they plunged. The models then proved to be far off the mark, and billions of dollars in losses nearly took down the US financial system. The federal government had little choice but to lend $700 billion for a bailout of Wall Street.

Granted, this is an extreme case. But it does highlight the importance of data selection. This is where having a solid team of domain experts and data scientists can be essential.

Next, when in the data preparation stage, there will need to be data cleansing. The fact is that all data has issues. Even companies like Facebook have gaps, ambiguities, and outliers in their datasets. It's inevitable.

So here are some actions you can take to cleanse the data:

- *De-duplication*: Set tests to identify any duplications and delete the extraneous data.
- *Outliers*: This is data that is well beyond the range of most of the rest of the data. This may indicate that the information is not helpful. But of course, there are situations where the reverse is true. This would be for fraud deduction.
- *Consistency*: Make sure you have clear definitions for the variables. Even terms like "revenue" or "customer" can have multiple meanings.
- *Validation Rules*: As you look at the data, try to find the inherent limitations. For example, you can have a flag for the age column. If it is over 120 in many cases, then the data has some serious issues.
- *Binning*: Certain data may not need to be specific. Does it really matter if someone is 35 or 37? Probably not. But comparing those from 30–40 to 41–50 probably would.

- *Staleness*: Is the data timely and relevant?
- *Merging*: In some cases, the columns of data may have very similar information. Perhaps one has height in inches and another in feet. If your model does not require a more detailed number, you can just use the one for feet.
- *One-Hot Encoding*: This is a way to replace categorical data as numbers. Example: Let's say we have a database with a column that has three possible values: Apple, Pineapple, and Orange. You could represent Apple as 1, Pineapple as 2, and Orange as 3. Sounds reasonable, right? Perhaps not. The problem is that an AI algorithm may think that Orange is greater than Apple. But with one-hot encoding, you can avoid this problem. You will create three new columns: is_Apple, is_Pineapple, and is_Orange. For each row in the data, you'll put 1 for where the fruit exists and 0 for the rest.
- *Conversion Tables*: You can use this when translating data from one standard to another. This would be the case if you have data in the decimal system and want to move over to the metric system.

These steps will go a long way in improving the quality of the data. There are also automation tools that can help out, such as from companies like SAS, Oracle, IBM, Lavastorm Analytics, and Talend. Then there are open source projects, such as OpenRefine, plyr, and reshape2.

Regardless, the data will not be perfect. No data source is. There will likely still be gaps and inaccuracies.

This is why you need to be creative. Look at what Eyal Lifshitz did, who is the CEO of BlueVine. His company leverages AI to provide financing to small businesses. "One of our data sources is credit information of our customers," he said. "But we've found that small business owners incorrectly identify their type of business. This could mean bad results for our underwriting. To deal with this, we scrape data from the customer website with AI algorithms, which helps identify the industry."[17]

Data cleansing approaches will also depend on the use cases for the AI project. For example, if you are building a system for predictive maintenance in manufacturing, the challenge will be to handle the wide variations from different sensors. The result is that a large amount of the data may have little value and be mostly noise.

Ethics and Governance

You need to be mindful of any restrictions on the data. Might the vendor prohibit you from using the information for certain purposes? Perhaps your company will be on the hook if something goes wrong? To deal with these issues, it is advisable to have the legal department brought in.

For the most part, data must be treated with care. After all, there are many high-profile cases where companies have violated privacy. A prominent example of this is Facebook. One of the company's partners, Cambridge Analytica, accessed millions of data points from profiles without the permission of users. When a whistleblower uncovered this, Facebook stock plunged—losing more than $100 billion in value. The company also came under pressure from the US and European governments.[18]

Something else to be wary of is scraping data from public sources. True, this is often an efficient way to create large datasets. There are also many tools that can automate the process. But scraping could expose your company to legal liability as the data may be subject to copyrights or privacy laws.

There are also some precautions that may ironically have inherent flaws. For example, a recent study from MIT shows that anonymized data may not be very anonymized. The researchers found that it was actually quite easy to reconstruct this type of data and identify the individuals—such as by merging two datasets. This was done by using data in Singapore from a mobile network (GPS tracking) and a local transportation system. After about 11 weeks of analysis, the researchers were able to identify 95% of the individuals.[19]

Finally, make sure you take steps to secure the data. The instances of cyberattacks and threats continue to increase at an alarming rate. In 2018, there were 53,000+ incidents and about 2,200 breaches, according to Verizon.[20] The report also noted the following:

- 76% of the breaches were financially motivated.
- 73% were from those outside the company.
- About half came from organized criminal groups and 12% from nation-state or state-affiliated actors.

The increasing use of cloud and on-premise data can subject a company to gaps in security as well. Then there is the mobile workforce, which can mean access to data that could expose it to breaches.

[18] https://venturebeat.com/2018/07/02/u-s-agencies-widen-investigation-into-what-facebook-knew-about-cambridge-analytica/

The attacks are also getting much more damaging. The result is that a company can easily suffer penalties, lawsuits, and reputational damage.

Basically, when putting together an AI project, make sure there is a security plan and that it is followed.

How Much Data Do You Need for AI?

The more data, the better, right? This is usually the case. Look at something called Hughes Phenomenon. This posits that as you add features to a model, the performance generally increases.

But quantity is not the end-all, be-all. There may come a point where the data starts to degrade. Keep in mind that you may run into something called the curse of dimensionality. According to Charles Isbell, who is the professor and senior associate dean of the School of Interactive Computing at Georgia Tech, "As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially."[21]

What is the practical impact? It could make it impossible to have a good model since there may not be enough data. This is why that when it comes to applications like vision recognition, the curse of dimensionality can be quite problematic. Even when analyzing RGB images, the number of dimensions is roughly 7,500. Just imagine how intensive the process would be using real-time, high-definition video.

More Data Terms and Concepts

When engaging in data analysis, you should know the basic terms. Here are some that you'll often hear:

Categorical Data: This is data that does not have a numerical meaning. Rather, it has a textual meaning like a description of a group (race and gender). Although, you can assign numbers to each of the elements.

Data Type: This is the kind of information a variable represents, such as a Boolean, integer, string, or floating point number.

Descriptive Analytics: This is analyzing data to get a better understanding of the current status of a business. Some examples of this include measuring what products are selling better or determining risks in customer support. There are many traditional software tools for descriptive analytics, such as BI applications.

Diagnostic Analytics: This is querying data to see why something has happened. This type of analytics uses techniques like data mining, decision trees, and correlations.

ETL (Extraction, Transformation, and Load): This is a form of data integration and is typically used in a data warehouse.

Feature: This is a column of data.

Instance: This is a row of data.

Metadata: This is data about data—that is, descriptions. For example, a music file can have metadata like the size, length, date of upload, comments, genre, artist, and so on. This type of data can wind up being quite useful for an AI project.

Numerical Data: This is any data that can be represented by a number. But numerical data can have two forms. There is discrete data, which is an integer—that is, a number without a decimal point. Then there is continuous data that has a flow, say temperature or time.

OLAP (Online Analytical Processing): This is technology that allows you to analyze information from various databases.

Ordinal Data: This is a mix of numerical and categorical data. A common example of this is the five-star rating on Amazon.com. It has both a star and a number associated with it.

Predictive Analytics: This involves using data to make forecasts. The models for this are usually sophisticated and rely on AI approaches like machine learning. To be effective, it is important to update the underlying model with new data. Some of the tools for predictive analytics include machine learning approaches like regressions.

Prescriptive Analytics: This is about leveraging Big Data to make better decisions. This is not only focued on predicting outcomes—but understanding the rationales. And this is where AI plays a big part.

Scalar Variables: These are variables that hold single values like name or credit card number.

Transactional Data: This is data that is recorded on financial, business, and logistical actions. Examples include payments, invoices, and insurance claims.

Conclusion

Being successful with AI means having a data-driven culture. This is what has been critical for companies like Amazon.com, Google, and Facebook. When making decisions, they look to the data first. There should also be wide availability of data across the organization.

Without this approach, success with AI will be fleeting, regardless of your planning. Perhaps this helps explain that—according to a study from NewVantage Partners—about 77% of respondents say that "business adoption" of Big Data and AI remain challenges.[22]

Key Takeaways

- Structured data is labeled and formatted—and is often stored in a relational database or spreadsheet.
- Unstructured data is information that has no predefined formatting.
- Semi-structured data has some internal tags that help with categorization.
- Big Data describes a way to handle huge amounts of volumes of information.
- A relational database is based on relationships of data. But this structure can prove difficult for modern-day applications, such as AI.
- A NoSQL database is more free-form, being based on a document model. This has made it better able to deal with unstructured and semi-structured data.
- The CRISP-DM Process provides a way to manage data for a project, with steps that include business understanding, data understanding, data preparation, modelling, evaluation, and deployment.
- Quantity of data is certainly important, but there also needs to be much work on the quality. Even small errors can have a huge impact on the results of an AI model.

CHAPTER

3

Machine Learning

Mining Insights from Data

A breakthrough in machine learning would be worth ten Microsofts.

—Bill Gates[1]

While Katrina Lake liked to shop online, she knew the experience could be much better. The main problem: It was tough to find fashions that were personalized.

So began the inspiration for Stitch Fix, which Katrina launched in her Cambridge apartment while attending Harvard Business School in 2011 (by the way, the original name for the company was the less catchy "Rack Habit"). The site had a Q&A for its users—asking about size and fashion styles, just to name a few factors—and expert stylists would then put together curated boxes of clothing and accessories that were sent out monthly.

The concept caught on quickly, and the growth was robust. But it was tough to raise capital as many venture capitalists did not see the potential in the business.

[1] Steve Lohr, "Microsoft, Amid Dwindling Interest, Talks Up Computing as a Career: Enrollment in Computing Is Dwindling," *New York Times*, March 1, 2004, start page C1, quote page C2, column 6.