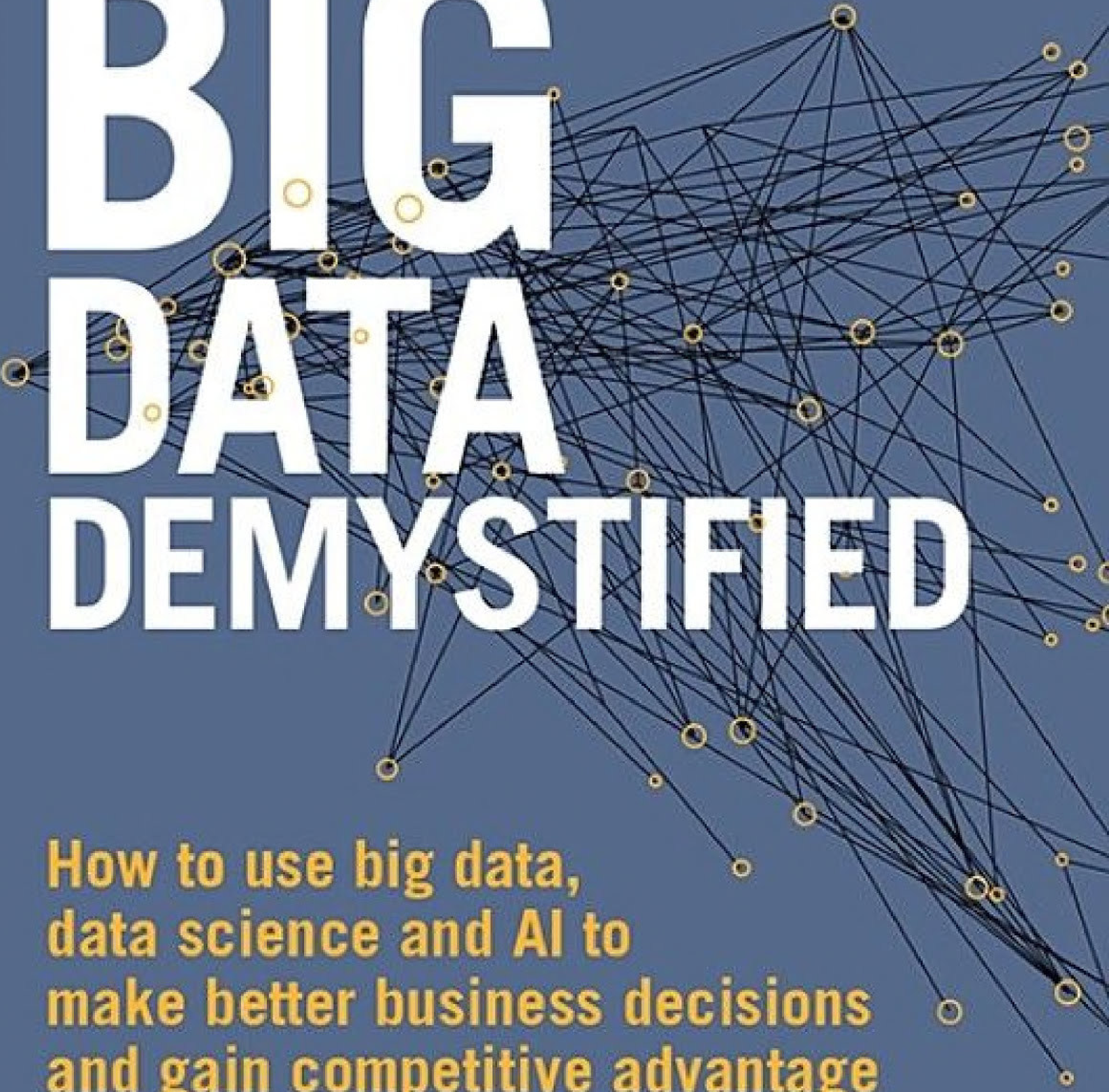
'Read this book! It is an essential guide to using data in a practical way that drives results.'
IAN MCHENRY, CEO, BEYOND PRICING

BIG DATA DEMYSTIFIED



How to use big data, data science and AI to make better business decisions and gain competitive advantage

DAVID STEPHENSON PhD

FT PUBLISHING
FINANCIAL TIMES

Contents

About the author

David Stephenson consults and speaks internationally in the fields of data science and big data analytics. He completed his PhD at Cornell University and was a professor at the University of Pennsylvania, designing and teaching courses for students in the engineering and Wharton business schools.

David has nearly 20 years of industry experience across Europe and the United States, delivering analytic insights and tools that have guided $10+ billion in business decisions and serving as an expert advisor to top-tier investment, private equity and management consulting firms. He has led global analytics programmes for companies spanning six continents.

David is from the USA but has been living in Amsterdam since 2006. More information and content are available on his company website at www.dsianalytics.com.

Acknowledgements

I would like to thank Eloise Cook at Pearson for her valuable editorial guidance, Uri Weis for comments on the first draft of the text and Matt Gardner for comments on the initial chapters. Thanks also to my father for proofing and providing helpful comments on the entire text. Despite their best efforts, any remaining errors are my own.

Introduction

You often hear the term ‘big data’, but do you really know what it is and why it’s important? Can it make a difference in your organization, improving results and bringing competitive advantage, and is it possible that not utilizing big data puts you at a significant competitive disadvantage?

The goal of this book is to demystify the term ‘big data’ and to give practical ways for you to leverage this data using data science and machine learning.

The term ‘big data’ refers to a new class of data: vast, rapidly accumulating quantities, which often do not fit a traditional structure. The term ‘big’ is an understatement that simply does not do justice to the complexity of the situation. The data we are dealing with is not only bigger than traditional data; it is fundamentally different, as a motorcycle is more than simply a bigger bicycle and an ocean is more than simply a deeper swimming pool. It brings new challenges, presents new opportunities, blurs traditional competitive boundaries and requires a paradigm shift related to how we draw tangible value from data. The ocean of data, combined with the technologies that have been developed to handle it, provide insights at enormous scale and have made possible a new wave of machine learning, enabling computers to drive cars, predict heart attacks better than physicians and master extremely complex games such as **Go** better than any human.

Why is big data a game-changer? As we will see, it allows us to draw much deeper insights from our data, understanding what motivates our customers and what slows down our production lines. In real time, it enables businesses to simultaneously deliver highly personalized experiences to millions of global customers, and it provides the computational power needed for scientific endeavours to analyse billions of data points in fields such as cancer research, astronomy and particle physics. Big data provides both the data and the computational resources that have enabled the recent resurgence in artificial intelligence, particularly with advances in **deep learning**, a methodology that has recently been making global headlines.

Beyond the data itself, researchers and engineers have worked over the past two decades to develop an entire ecosystem of hardware and software solutions for collecting, storing, processing and analysing this abundant data. I refer to these hardware and software tools together as the **big data ecosystem**. This ecosystem allows us to draw immense value from big data for applications in business, science and healthcare. But to use this data, you need to piece together the parts of the big data ecosystem that work best for your applications, and you need to apply appropriate analytic methods to the data – a practice that has come to be known as **data science.**

All in all, the story of big data is much more than simply a story about data and technology. It is about what is already being done in commerce, science and society and what difference it can make for your business. Your decisions must go further than purchasing a technology. In this book, I will outline tools, applications and processes and explain how to draw value from modern data in its many forms.

Most organizations see big data as an integral part of their digital transformation. Many of the most successful organizations are already well along their way in applying big data and data science techniques, including machine learning. Research has shown a strong correlation between big data usage and revenue growth (50 per cent higher revenue growth[1]), and it is not unusual for organizations applying data science techniques to see a 10–20 per cent improvement in **key performance indicators (KPIs)**.

For organizations that have not yet started down the path of leveraging big data and data science, the number one barrier is simply not knowing if the benefits are worth the cost and effort. I hope to make those benefits clear in this book, along the way providing case studies to illustrate the value and risks involved.

In the second half of this book, I'll describe practical steps for creating a data strategy and for getting data projects done within your organization. I'll talk about how to bring the right people together and create a plan for collecting and using data. I'll discuss specific areas in which data science and big data tools can be used within your organization to improve results, and I'll give advice on finding and hiring the right people to carry out these plans.

I'll also talk about additional considerations you'll need to address, such as data governance and privacy protection, with a view to protecting your organization against competitive, reputational and legal risks.

We'll end with additional practical advice for successfully carrying out data initiatives within your organization.

Overview of chapters

Part 1: Big data demystified

Chapter 1: The story of big data

How big data developed into a phenomenon, why big data has become such an important topic over the past few years, where the data is coming from, who is using it and why, and what has changed to make possible today what was not possible in the past.

Chapter 2: Artificial intelligence, machine learning and big data

A brief history of artificial intelligence (AI), how it relates to machine learning, an introduction to neural networks and deep learning, how AI is used today and how it relates to big data, and some words of caution in working with AI.

Chapter 3: Why is big data useful?

How our data paradigm is changing, how big data opens new opportunities and improves established analytic techniques, and what it means to be data-driven, including success stories and case studies.

Chapter 4: Use cases for (big) data analytics

An overview of 20 common business applications of (big) data, analytics and data science, with an emphasis on ways in which big data improves existing analytic methods.

Chapter 5: Understanding the big data ecosystem

Overview of key concepts related to big data, such as open-source code, distributed computing and cloud computing.

Part 2: Making the big data ecosystem work for your organization

Chapter 6: How big data can help guide your strategy

Using big data to guide strategy based on insights into your customers, your product performance, your competitors and additional external factors.

Chapter 7: Forming your strategy for big data and data science

Step-by-step instructions for scoping your data initiatives based on business goals and broad stakeholder input, assembling a project team, determining the most relevant analytics projects and carrying projects through to completion.

Chapter 8: Implementing data science – analytics, algorithms and machine learning

Overview of the primary types of analytics, how to select models and databases, and the importance of agile methods to realize business value.

Chapter 9: Choosing your technologies

Choosing technologies for your big data solution: which decisions you'll need to make, what to keep in mind, and what resources are available to help make these choices.

Chapter 10: Building your team

The key roles needed in big data and data science programmes, and considerations for hiring or outsourcing those roles.

Chapter 11: Governance and legal compliance

Principles in privacy, data protection, regulatory compliance and data governance, and their impact from legal, reputational and internal perspectives. Discussions of PII, linkage attacks and Europe's new privacy regulation (GDPR). Case studies of companies that have gotten into trouble from inappropriate use of data.

Chapter 12: Launching the ship – successful deployment in the organization

Case study of a high-profile project failure. Best practices for making data initiatives successful in your organization, including advice on making your organization more data-driven, positioning your analytics staff within your organization, consolidating data and using resources efficiently.

Part 1

Big data demystified

Chapter 1

The story of big data

We've always struggled with storing data. Not long ago, our holidays were remembered at a cost of $1 per photo. We saved only the very best TV shows and music recitals, overwriting older recordings. Our computers always ran out of memory.

Newer, cheaper technologies turned up the tap on that data flow. We bought digital cameras, and we linked our computers to networks. We saved more data on less expensive computers, but we still sorted and discarded information continuously. We were frugal with storage, but the data we stored was small enough to manage.

Data started flowing thicker and faster. Technology made it progressively easier for anyone to create data. Roll film cameras gave way to digital video cameras, even on our smartphones. We recorded videos we never replayed.

High-resolution sensors spread through scientific and industrial equipment. More documents were saved in digital format. More significantly, the internet began linking global data silos, creating challenges and opportunities we were ill-equipped to handle. The *coup de grâce* came with the development of crowd-sourced digital publishing, such as YouTube and Facebook, which opened the portal for anyone with a connected digital device to make nearly unlimited contributions to the world's data stores.But storage was only part of the challenge. While we were rationing our storage, computer scientists were rationing computer processing power. They were writing computer programs to solve problems in science and industry: helping to understand chemical reactions, predict stock market movements and minimize the cost of complicated resource scheduling problems.

Their programs could take days or weeks to finish, and only the most well-endowed organizations could purchase the powerful computers needed to solve the harder problems.

In the 1960s and again in the 1980s, computer scientists were building high hopes for advancements in the field of **machine learning (ML),** a type of **artificial intelligence (AI),** but their efforts stalled each time, largely due to limitations in data and technology.

In summary, our ability to draw value from data was severely limited by the technologies of the twentieth century.

What changed towards the start of the twenty-first century?

There were several key developments towards the start of the twenty-first century. One of the most significant originated in Google. Created to navigate the overwhelming data on the newly minted world wide web, Google was all about big data. Its researchers soon developed ways to make normal computers work together like supercomputers, and in 2003 they published these results in a paper which formed the basis for a **software framework** known as **Hadoop**. Hadoop became the bedrock on which much of the world's initial big data efforts would be built.

The concept of 'big data' incubated quietly in the technology sector for nearly a decade before becoming mainstream. The breakthrough into management circles seemed to happen around 2011, when McKinsey published their report, '*Big data: The next frontier for innovation, competition, and productivity.*'[2] The first public talk I gave on big data was at a designated 'big data' conference in London the next year (2012), produced by a media company seizing the opportunity to leverage a newly trending topic.

But even before the McKinsey paper, large data-driven companies such as eBay were already developing internal solutions for fundamental big data challenges. By the time of McKinsey's 2011 publication, Hadoop was already five years old and the University of California at Berkeley had open-sourced their **Spark** framework, the Hadoop successor that leveraged inexpensive **RAM** to process big data much more quickly than Hadoop.

Let's look at why data has grown so rapidly over the past few years and why the topic 'big data' has become so prominent.

Why so much data?

The volume of data we are committing to digital memory is undergoing explosive growth for two reasons:

1. The proliferation of devices that generate digital data: ubiquitous personal computers and mobile phones, scientific sensors, and the literally billions of sensors across the expanding **Internet of Things (IoT)** (see Figure 1.1).
2. The rapidly plummeting cost of digital storage.

The proliferation of devices that generate digital data

Technology that creates and collects data has become cheap, and it is everywhere. These computers, smartphones, cameras, RFID (radio-frequency identification), movement sensors, etc., have found their way into the hands of the mass consumer market as well as those of scientists, industries and governments. Sometimes we intentionally create data, such as when we take videos or post to websites, and sometimes we create data unintentionally, leaving a digital footprint on a webpage that we browse, or carrying smartphones that send geospatial information to network providers. Sometimes the data doesn't relate to us at all, but is a record of machine activity or scientific phenomena. Let's look at some of the main sources and uses of the data modern technology is generating.

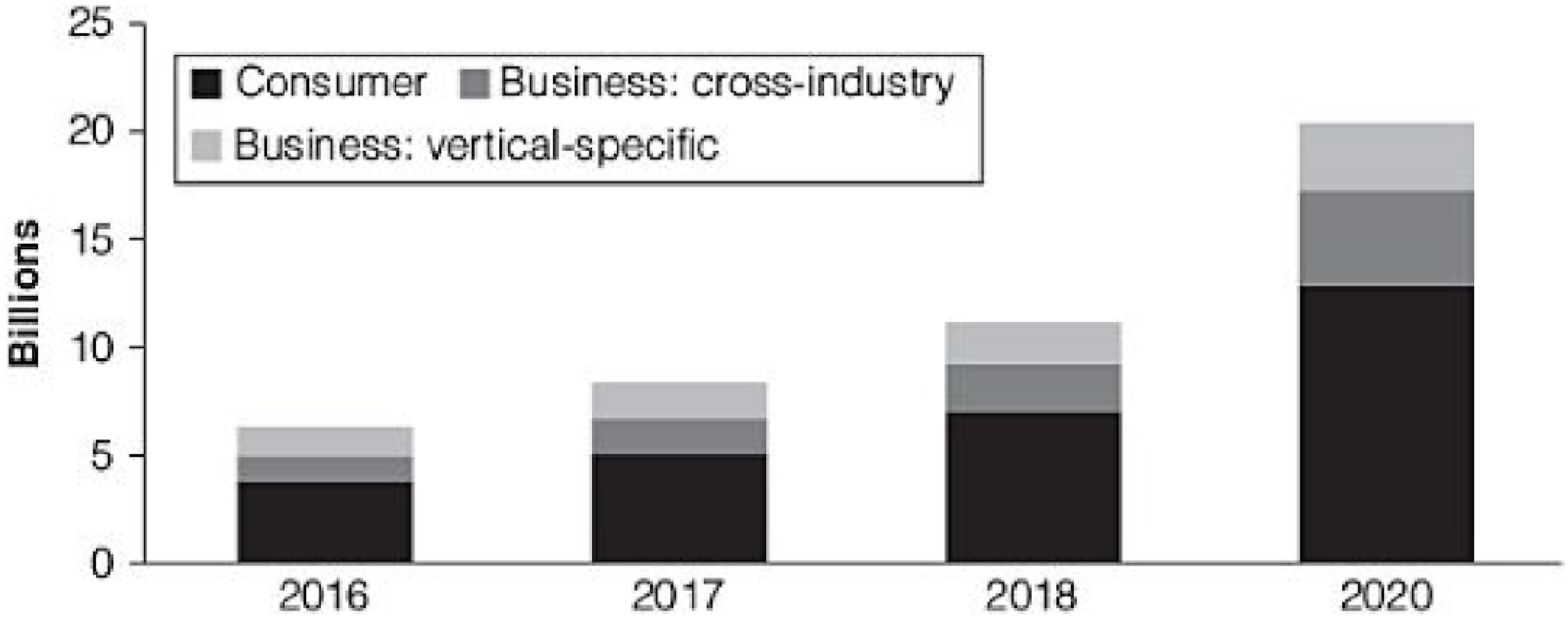


Figure 1.1 Number of IoT devices by category.[3]

Content generation and self-publishing

What does it take to get your writing published? A few years ago, it took a printing press and a network of booksellers. With the internet, you only needed the skills to create a web page. Today, anyone with a Facebook or Twitter account can instantly publish content with worldwide reach. A similar story has played out for films and videos. Modern technology, particularly the internet, has completely changed the nature of publishing and has facilitated a massive growth in human-generated content.

Self-publishing platforms for the masses, particularly Facebook, YouTube and Twitter, threw open the floodgates of mass-produced data. Anyone could easily post content online, and the proliferation of mobile devices, particularly those capable of recording and uploading video, further lowered the barriers. Since nearly everyone now has a personal device with a high-resolution video camera and continuous internet access, the data uploads are enormous. Even children easily upload limitless text or video to the public domain.

YouTube, one of the most successful self-publishing platforms, is possibly the single largest consumer of corporate data storage today. Based on previously published statistics, it is estimated that YouTube is adding approximately 100 **petabytes (PB)** of new data per year, generated from several hundred hours of video uploaded each minute. We are also watching a tremendous amount of video online, on YouTube, Netflix and similar streaming services. Cisco recently estimated that it would take more than 5 million years to watch the amount of video that will cross global IP (internet protocol) networks each month in 2020.

Consumer activity

When I visit a website, the owner of that site can see what information I request from the site (search words, filters selected, links clicked). The site can also use the **JavaScript** on my browser to record how I interact with the page: when I scroll down or hover my mouse over an item. Websites use these details to better understand visitors, and a site might record details for several hundred categories of online actions (searches, clicks, scrolls, hovers, etc.). Even if I never log in and the site doesn't know who I am, the insights are valuable. The more information the site gathers about its visitor base, the better it can optimize marketing efforts, landing pages and product mix.

Mobile devices produce even heavier digital trails. An application installed on my smartphone may have access to the device sensors, including GPS (global positioning system). Since many people always keep their smartphones near them, the phones maintain very accurate data logs of the location and activity cycles of their owner. Since the phones are typically in constant communication with cell towers and Wi-Fi routers, third parties may also see the owners' locations. Even companies with brick-and-mortar shops are increasingly using signals from smartphones to track the physical movement of customers within their stores.

Many companies put considerable effort into analysing these digital trails, particularly e-commerce companies wanting to better understand online visitors. In the past, these companies would discard most data, storing only the key events (e.g. completed sales), but many websites are now storing all data from each online visit, allowing them to look back and ask detailed questions. The scale of this customer journey data is typically several **gigabytes (GB)** per day for smaller websites and several **terabytes (TB)** per day for larger sites. We'll return to the benefits of analysing customer journey data in later chapters.

We are generating data even when we are offline, through our phone conversations or when moving past video cameras in shops, city streets, airports or roadways. Security companies and intelligence agencies rely heavily on such data. In fact, the largest consumer of data storage today is quite likely the United States' National Security Agency (NSA). In August 2014, the NSA completed construction of a massive data centre in Bluffdale, Utah, codenamed *Bumblehive*, at a cost somewhere between 1 and 2 billion dollars. Its actual storage capacity is classified, but the governor of Utah told reporters in 2012 that it would be, 'the first facility in the world expected to gather and house a **yottabyte**'.

Machine data and the Internet of Things (IoT)

Machines never tire of generating data, and the number of connected machines is growing at a rapid pace. One of the more mind-blowing things you can do in the next five minutes is to check out Cisco's Visual Networking Index™, which recently estimated that global IP traffic will reach over two **zettabytes** per year by 2020.

We may hit a limit in the number of mobile phones and personal computers we use, but we'll continue adding networked processors to devices around us. This huge network of connected sensors and processors is known as the **Internet of Things (IoT)**. It includes the smart energy meters appearing in our homes, the sensors in our cars that help us drive and sometimes communicate with our insurance companies, the sensors deployed to monitor soil, water, fauna or atmospheric conditions, the digital control systems used to monitor and optimize factory equipment, etc. The number of such devices stood at approximately 5 billion in 2015 and has been estimated to reach between 20 and 50 billion by 2020.

Scientific research

Scientists have been pushing the boundaries of data transport and data processing technologies. I'll start with an example from particle physics.

Case study – The large hadron collider (particle physics)

One of the most important recent events in physics was witnessed on 4 July 2012: the discovery of the Higgs boson particle, also known as 'the god particle'. After 40 years of searching, researchers finally identified the particle using the Large Hadron Collider (LHC), the world's largest machine[4] (see Figure 1.2). The massive LHC lies within a tunnel 17 miles (27 km) in circumference, stretching over the Swiss–French border. Its 150

million sensors deliver data from experiments 30 million times per second. This data is further filtered to a few hundred points of interest per second. The total annual data flow reaches 50 PB, roughly the equivalent of 500 years of full HD-quality movies. It is the poster child of big data research in physics.

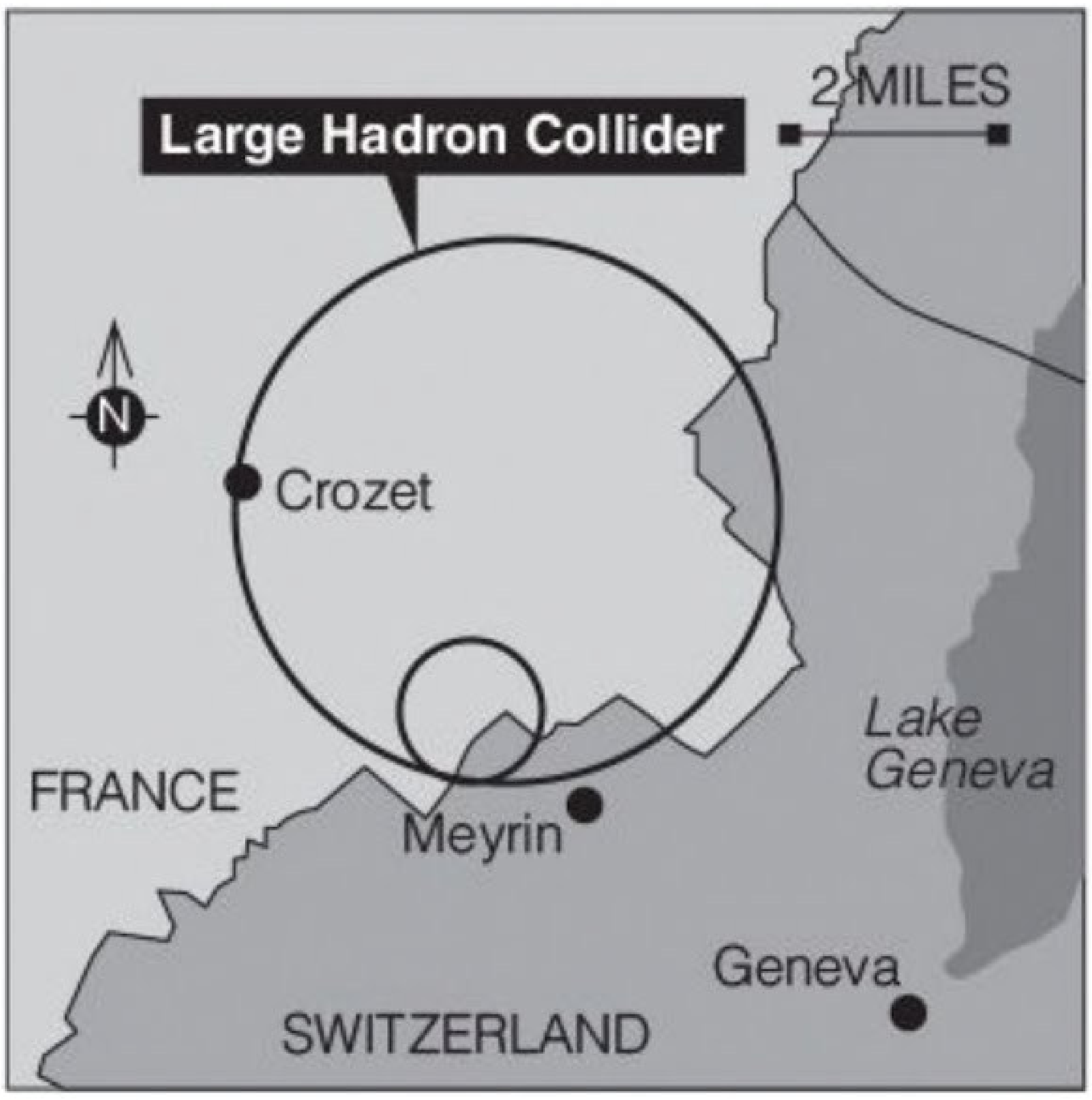


Figure 1.2 The world's largest machine.[5]

Case study – The square kilometre array (astronomy)

On the other side of the world lies the Australian Square Kilometre Array Pathfinder (ASKAP), a radio telescope array of 36 parabolic antennas, each 12 metres in diameter[6] and spanning 4000 square metres. Twelve of the 36 antennas were activated in October 2016[7], and the full 36, when commissioned, are expected to produce data at a rate of over 7.5 TB per second[8] (one month's worth of HD movies per second). Scientists are planning a larger Square Kilometre Array (SKA), which will be spread over several continents and be 100 times larger than the ASKAP. This may be the largest single data collection device ever conceived.

All of this new data presents abundant opportunities, but let's return now to our fundamental problem, the cost of processing and storing that data.

The plummeting cost of disk storage

There are two main types of computer storage: disk (e.g. hard drive) and random access memory (RAM). Disk storage is like a filing cabinet next to your desk. There may be a lot of space, but it takes time to store and retrieve the information. RAM is like the space on top of your desk. There is less space, but you can grab what's there very quickly. Both types of storage are important for handling big data.

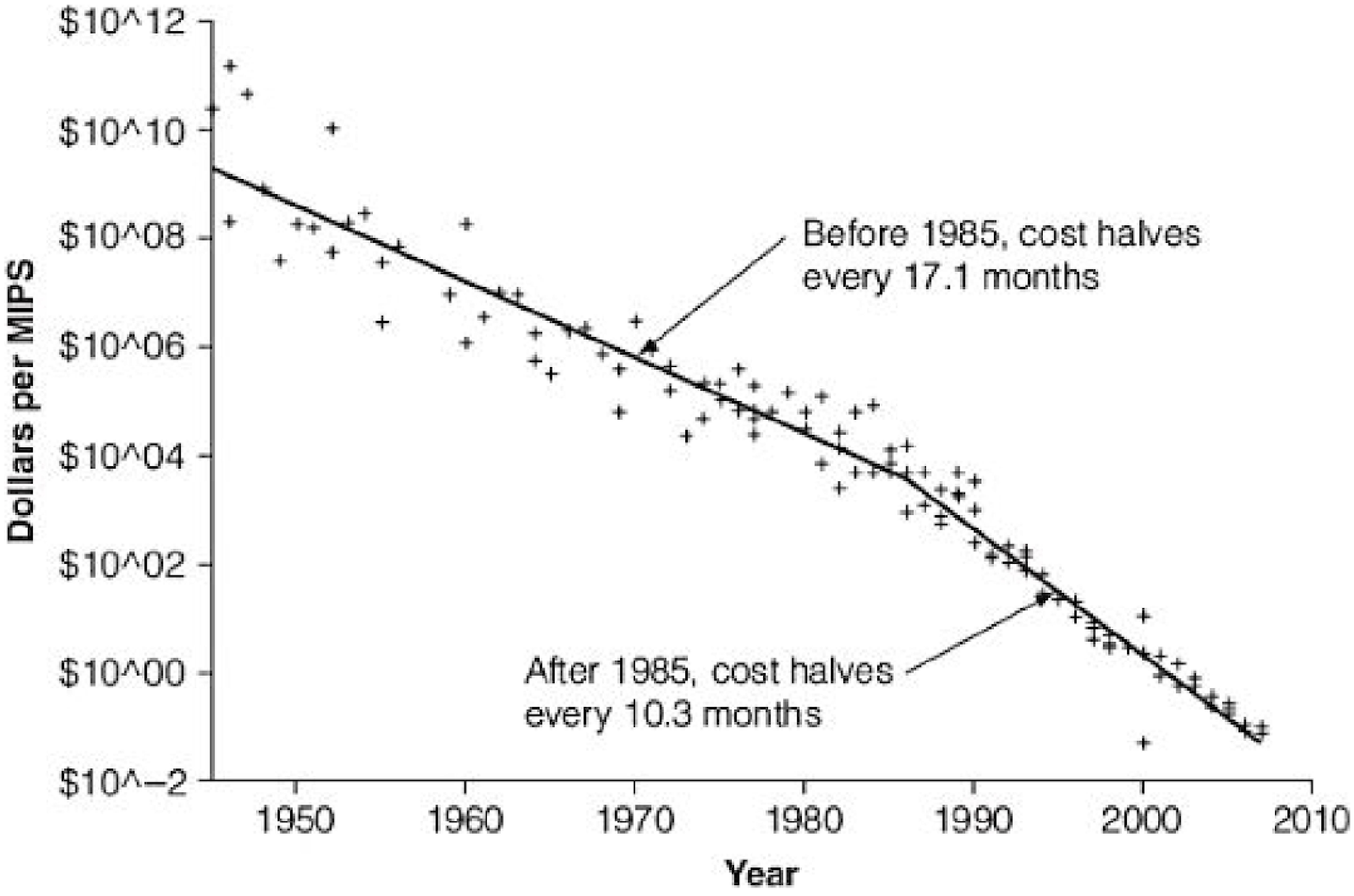


Figure 1.5 Historic cost of processing power (log scale).[10]

Why did big data become such a hot topic?

Over the last 15 years, we've come to realize that big data is an opportunity rather than a problem. McKinsey's 2011 report spoke directly to the CEOs, elaborating on the value of big data for five applications (healthcare, retail, manufacturing, the public sector and personal location data). The report predicted big data could raise KPIs by 60 per cent and estimated hundreds of billions of dollars of added value per sector. The term 'big data' became the buzzword heard around the world, drawn out of the corners of technology and cast into the executive spotlight.

With so many people talking so much about a topic they so little understood, many quickly grew jaded about the subject. But big data became such a foundational concept that Gartner, which had added big data to their **Gartner Hype Cycle** for Emerging Technologies in 2012, made the unusual decision to completely remove it from the Hype Cycle in 2015, thus acknowledging that big data had become so foundational as to warrant henceforth being referred to simply as 'data' (see Figure 1.6).

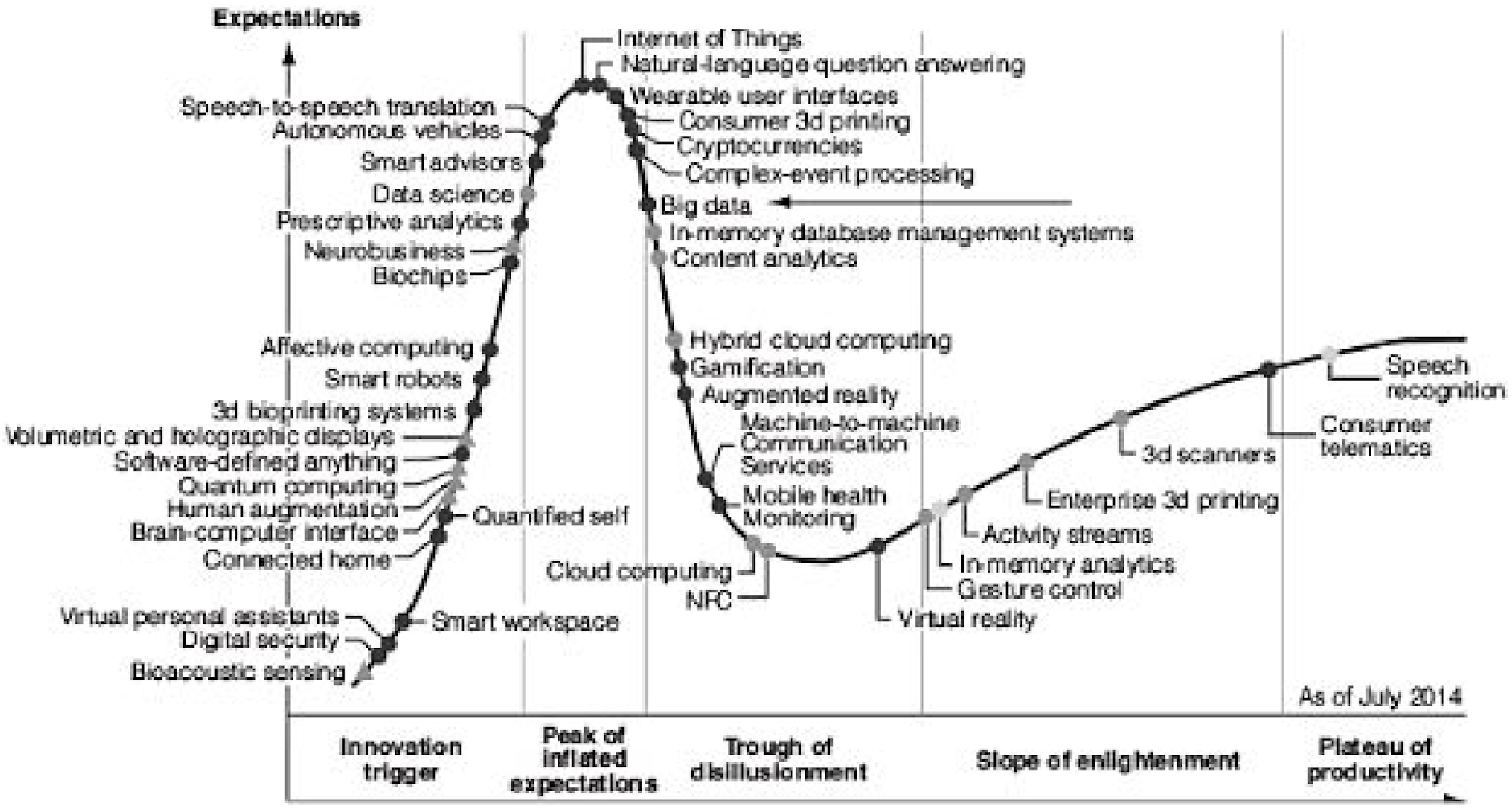


Figure 1.6 Gartner Hype Cycle for Emerging Technologies, 2014.

Organizations are now heavily dependent on big data. But why such widespread adoption?

- **Early adopters,** such as Google and Yahoo, risked significant investments in hardware and software development. These companies paved the way for others, demonstrating commercial success and sharing computer code.
- **The second wave** of adopters did much of the hardest work. They could benefit from the examples of the early adopters and leverage some shared code but still needed to make significant investments in hardware and develop substantial internal expertise.

Today, we have reached a point where we have the role models and the tools for nearly any organization to start leveraging big data.

Let’s start with looking at some role models who have inspired us in the journey.

Successful big data pioneers

Google’s first mission statement was ‘to organize the world’s information and make it universally accessible and useful.’ Its valuation of $23 million only eight years later demonstrated to the world the value of mastering big data.

It was Google that released the 2003 paper that formed the basis of Hadoop. In January 2006, Yahoo made the decision to implement Hadoop in their systems.[11] Yahoo was also doing quite well in those days, with a stock price that had slowly tripled over the previous five years.

Around the time that Yahoo was implementing Hadoop, eBay was working to rethink how it handled the volume and variety of its customer journey data. Since 2002, eBay had been utilizing a **massively parallel processing (MPP)** Teradata database for reporting and analytics. The system worked very well, but storing the entire web logs was prohibitively expensive on such a proprietary system.

eBay's infrastructure team worked to develop a solution combining several technologies and capable of storing and analysing tens of petabytes of data. This gave eBay significantly more detailed customer insights and played an important role in their platform development, translating directly into revenue gains.

Open-source software has levelled the playing field for software developers

Computers had become cheaper, but they still needed to be programmed to operate in unison if they were to handle big data (such as coordinating several small cars to move a piano, instead of one truck). Code needed to be written for basic functionality, and additional code needed to be written for more specialized tasks. This was a substantial barrier to any big data project, and it is where open-source software played such an important role.

Open-source software is software which is made freely available for anyone to use and modify (subject to some restrictions). Because big data software such as Hadoop was open-sourced, developers everywhere could share expertise and build off each other's code.

Hadoop is one of many big data tools that have been open-sourced. As of 2017, there are roughly 100 projects related to big data or Hadoop in the **Apache Software Foundation** alone (we'll discuss the Apache foundation later). Each of these projects solves a new challenge or solves an old challenge in a new way. For example, Apache **Hive** allows companies to use Hadoop as a large database, and Apache **Kafka** provides messaging between machines. New projects are continually being released to Apache, each one addressing a specific need and further lowering the barrier for subsequent entrants into the big data ecosystem.

Keep in mind

Most of the technology you'll need for extracting value from big data is already readily available. If you're just starting out with big data, leverage as much existing technology as possible.

Affordable hardware and open-sourced software were lowering the barrier for companies to start using big data. But the problem remained that buying and setting up computers for a big data system was an expensive, complicated and risky process, and companies were uncertain how much hardware to purchase. What they needed was access to computing resources without long-term commitment.

Cloud computing has made it easy to launch and scale initiatives

Cloud computing is essentially renting all or part of an offsite computer. Many companies are already using one or more **public cloud** services: AWS, Azure, Google Cloud, or a local provider. Some companies maintain **private clouds**, which are computing resources that are maintained centrally within the company and made available to business units on demand. Such private clouds allow efficient use of shared resources.

Cloud computing can provide hardware or software solutions. **Salesforce** began in 1999 as a Software as a Service (SaaS), a form of cloud computing. Amazon Web Services (AWS) launched its Infrastructure as a Service (IaaS) in 2006, first renting storage and a few months later renting entire servers. Microsoft launched its cloud computing platform, Azure, in 2010, and Google launched Google Cloud in 2011.

Cloud computing solved a pain point for companies uncertain of their computing and storage needs. It allowed companies to undertake big data initiatives without the need for large capital expenditures, and it allowed them to immediately scale existing initiatives up or down. In addition, companies could move the cost of big data infrastructure from **CapEx** to **OpEx**.

The costs of cloud computing are falling, and faster networks allow remote machines to integrate seamlessly. Overall, cloud computing has brought agility to big data, making it possible for companies to experiment and scale without the cost, commitment and wait-time of purchasing dedicated computers.

With scalable data storage and compute power in place, the stage was set for researchers to once again revisit a technology that had stalled in the 1960s and again in the 1980s: artificial intelligence.

Takeaways

- Modern technology has given us tools to produce much more digital information than ever before.
- The dramatic fall in the cost of digital storage allows us to keep virtually unlimited amounts of data.
- Technology pioneers have developed and shared software that enables us to create substantial business value from today’s data.

Ask yourself

- How are organizations in your sector already using big data technologies? Consider your competitors as well as companies in other sectors.
- What data would be useful to you if you could store and analyse it as you’d like? Think, for example, of traffic to your website(s), audio and video recordings, or sensor readings.
- What is the biggest barrier to your use of big data: technology, skill sets or use-cases?