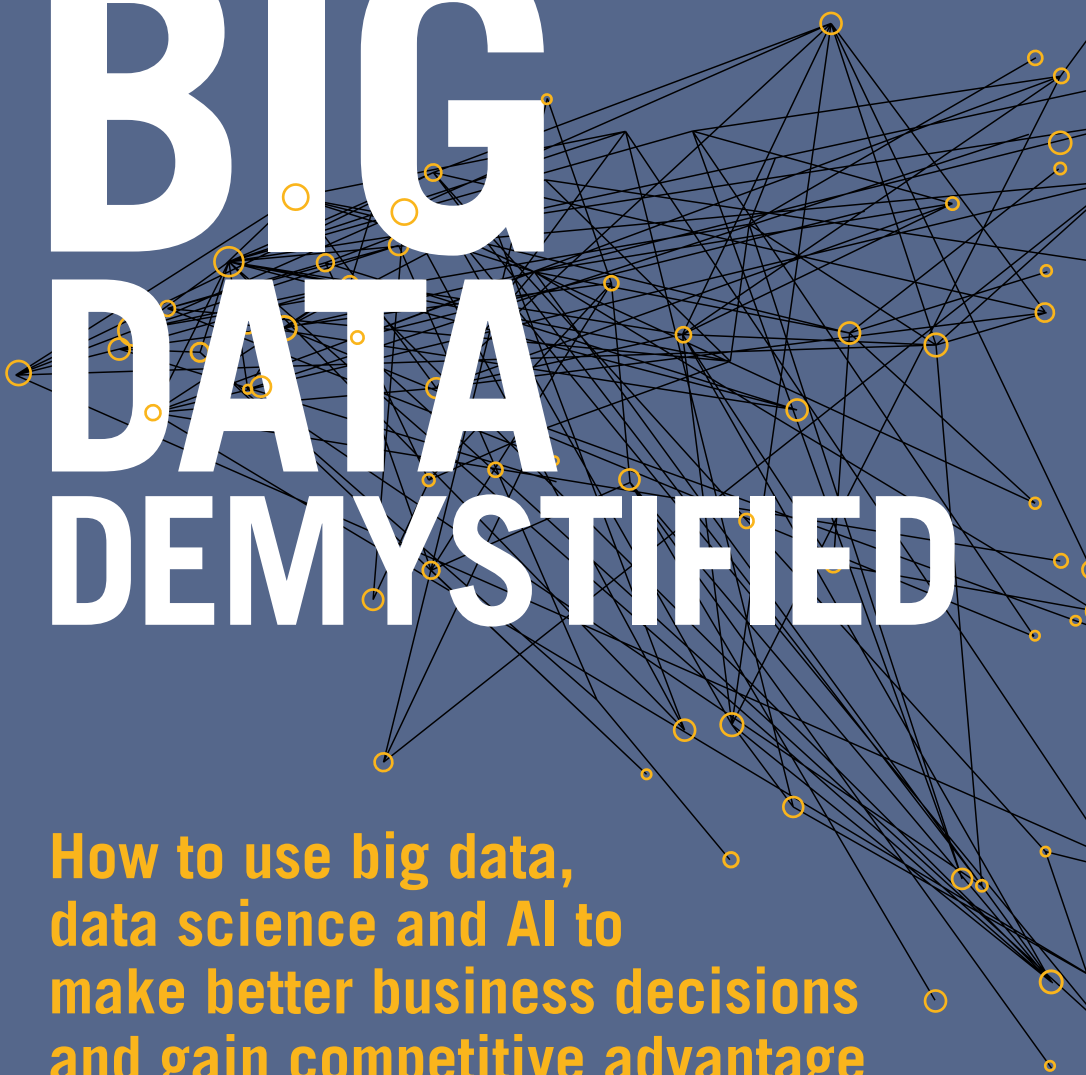
'Read this book! It is an essential guide to using data in a practical way that drives results.'
IAN MCHENRY, CEO, BEYOND PRICING

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

Praise for *Big Data Demystified*

'Before you embark on any kind of big data initiative at your organisation, read this book! It is an essential guide to using data in a practical way that drives results.'

Ian McHenry, CEO, Beyond Pricing

'This is the book we've been missing: big data explained without the complexity! And it will help you to look for big data opportunities in your day-to-day work.'

Marc Salomon, Professor in Decision Sciences and Dean, University of Amsterdam Business School

'Big data for the rest of us! I have never come across a book that is so full of practical advice, actionable examples and helpful explanations. Read this one book and start executing big data at your workplace tomorrow!'

Tobias Wann, CEO, @Leisure Group

'Dr Stephenson provides an excellent overview of the opportunities and tools that a modern business can exploit in data, while also going deep into the technical, organisational and procedural solutions. This book can be used as a best-practice education for both data analytics n00bs and seasoned professionals looking to identify gaps in data strategy.'

Clancy Childs, Chief Product and Technology Officer, Dow Jones DNA; Former Product Manager, Google Analytics

Marketing

Marketing is one of the first places you should look for applying big data. In Dell's 2015 survey,[1] the top three big data use cases among respondents were all related to marketing. These three were:

1. Better targeting of marketing efforts.
2. Optimization of ad spending.
3. Optimization of social media marketing.

This highlights how important big data is for marketing. Consider the number of potential ad positions in the digital space. It's enormous, as is the number of ways that you can compose (via keyword selection), purchase (typically through some bidding process) and place your digital advertisements. Once your advertisements are placed, you'll collect details of the ad placements and the click responses (often by placing invisible pixels on the web pages, collectively sending millions of messages back to a central repository).

Once customers are engaged with your product, typically by visiting your website or interacting with your mobile application, they start to leave digital trails, which you can digest with traditional web analytics tools or analyse in full detail with a big data tool.

Marketing professionals are traditionally some of the heaviest users of web analytics, which in turn is one of the first points of entry for online companies that choose to store and analyse full customer journey data rather than summarized or sampled web analytics data. Marketing professionals are dependent on the online data to understand the behaviour of customer cohorts brought from various marketing campaigns or keyword searches, to allocate revenue back to various acquisition sources, and to identify the points of the online journey at which customers are prone to drop out of the funnel and abandon the purchase process.

Social media

Social media channels can play an important role in helping you understand customers, particularly in real time. Consider a recent

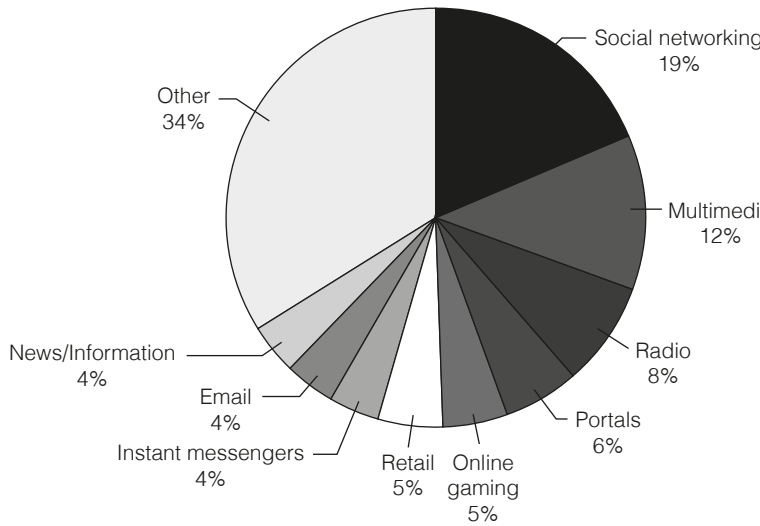


Figure 4.1 Share of the total digital time spent by content category.
Source: comScore Media Metrix Multi-Platform, US, Total Audience, December 2015.[33]

comScore report showing that social networking accounts for nearly one out of five minutes spent online in the US (see Figure 4.1)

Social media gives insight into customer sentiment, keyword usage and campaign effectiveness, and can flag a PR crisis you need to address immediately. Social media data is huge and it moves fast. Consider Twitter, where 6000 tweets are created each second, totalling 200 billion tweets per year.[34] You'll want to consider a range of social channels, as each may play an important role in understanding your customer base, and each has its own mixture of images, links, tags and free text, appealing to slightly different customer segments and enabling different uses.

Pricing

You may be using one or more standard pricing methods in your organization. These methods are specialized to fit specific sectors and applications.

Financial instruments are priced to prevent arbitrage, using formulas or simulations constructed from an underlying mathematical model of market rate movements. Insurance companies use

risk- and cost-based models, which may also involve simulations to estimate the impact of unusual events. If you are employing such a simulation-based pricing method, the big data ecosystem provides you with a scalable infrastructure for fast **Monte Carlo simulations** (albeit with issues related to capturing correlations).

If you are in commerce or travel, you may be using methods of dynamic pricing that involve modelling both the supply and the demand curves and then using experimental methods to model price elasticity over those two curves. In this case, big data provides you with the forecasting tools and methods mentioned earlier in this chapter, and you can use the micro-conversions in your customer journey data as additional input for understanding price elasticity.

Customer retention/customer loyalty

Use big data technologies to build customer loyalty in two ways.

First, play defence by monitoring and responding to signals in social media and detecting warning signals based on multiple touch points in the omni-channel experience. I'll illustrate such an omni-channel signal in the coming section on customer churn. In Chapter 6, I'll also discuss an example of customer service initiated by video analysis, which is a specific technique for applying non-traditional data and AI to retain customers and build loyalty.

Second, play offense by optimizing and personalizing the customer experience you provide. Improve your product using A/B testing; build a recommendation engine to enable successful shopping experiences; and deliver customized content for each customer visit (constructed first using offline big data analytics and then implemented using streaming processing for real-time customization).

Cart abandonment (real time)

Roughly 75 per cent of online shopping carts are abandoned.[35] Deploy an AI program that analyses customer behaviour leading up to the point of adding items to shopping carts. When the AI

predicts that the customer is likely to not complete the purchase, it should initiate appropriate action to improve the likelihood of purchase.

Conversion rate optimization

Conversion rate optimization (CRO) is the process of presenting your product in a way that maximizes the number of conversions. CRO is a very broad topic and requires a multi-disciplinary approach. It is a mixture of art and science, of psychology and technology. From the technology side, CRO is aided by A/B testing, by relevant recommendations and pricing, by real-time product customization, by cart abandonment technologies, etc.

Product customization (real time)

Adjust the content and format of your website in real time based on what you've learned about the visitor and on the visitor's most recent actions. You'll know general properties of the visitor from past interactions, but you'll know what they are looking for today based on the past few minutes or seconds. You'll need an unsampled customer journey to build your customization algorithms and you'll need streaming data technologies to implement the solution in real time.

Retargeting (real time)

Deploy an AI program to analyse the customer behaviour on your website in real time and estimate the probability the customer will convert during their next visit. Use this information to bid on retargeting slots on other sites that the customer subsequently visits. You should adjust your bidding prices immediately (a fraction of a second) rather than in nightly batches.

Fraud detection (real time)

In addition to your standard approach to fraud detection using manual screening or automated rules-based methods, explore alternative machine learning methods trained on large data sets.[36]

The ability to store massive quantities of time series data provides both a richer training set as well as additional possibilities for features and scalable, real-time deployment using **fast data** methods (Chapter 5).

Churn reduction

You should be actively identifying customers at high risk of becoming disengaged from your product and then work to keep them with you. If you have a paid usage model, you'll focus on customers at risk of cancelling a subscription or disengaging from paid usage. Since the cost of acquiring new customers can be quite high, the return on investment (ROI) on churn reduction can be significant.

There are several analytic models typically used for churn analysis. Some models will estimate the survival rate (longevity) of your customer, while others are designed to produce an estimated likelihood of churn over a period (e.g. the next two months). Churn is typically a rare event, which makes it more difficult for you to calibrate the accuracy of your model and balance between false positives and false negatives. Carefully consider your tolerance for error in either direction, balancing the cost of labelling a customer as a churn potential and wasting money on mitigation efforts vs the cost of not flagging a customer truly at risk of churning and eventually losing the customer.

These traditional churn models take as input all relevant and available features, including subscription data, billing history, and usage patterns. As you increase your data supply, adding customer journey data such as viewings of the Terms and Conditions webpage, online chats with customer support, records of phone calls to customer support, and email exchanges, you can construct a more complete picture of the state of the customer, particularly when you view these events as a sequence (e.g. receipt of a high bill, followed by contact with customer support, followed by viewing cancellation policy online).

In addition to utilizing the additional data and data sources to improve the execution of the traditional models, consider using

artificial intelligence models, particularly deep learning, to reduce churn. With deep learning models, you can work from unstructured data sources rather than focusing on pre-selecting features for the churn model.

Predictive maintenance

If your organization spends significant resources monitoring and repairing machinery, you'll want to utilize big data technologies to help with predictive maintenance, both to minimize wear and to avoid unexpected breakdowns. This is an important area for many industries, including logistics, utilities, manufacturing and agriculture, and, for many of them, accurately predicting upcoming machine failures can bring enormous savings. In some airlines, for example, maintenance issues have been estimated to cause approximately half of all technical flight delays. In such cases, gains from predictive maintenance can save tens of millions annually, while providing a strong boost to customer satisfaction.

The Internet of Things (IoT) typically plays a strong role in such applications. As you deploy more sensors and feedback mechanisms within machine parts and systems, you gain access to a richer stream of real-time operational data. Use this not only to ensure reliability but also for tuning system parameters to improve productivity and extend component life.

This streaming big data moves you from model-driven predictive maintenance to data-driven predictive maintenance, in which you continuously respond to real-time data. Whereas previously we may have predicted, detected and diagnosed failures according to a standard schedule, supplemented with whatever data was periodically collected, you should increasingly monitor systems in real time and adjust any task or parameter that might improve the overall efficiency of the system.

Supply chain management

If you're managing a supply chain, you've probably seen the amount of relevant data growing enormously over the past few years. Over half of respondents in a recent survey of supply chain industry