Fallacy #1

BIG DATA SOLVES EVERYTHING

From Google to start-up analytics firms, many companies have successfully implemented business models around the opportunities offered by big data. The growing number of analytics use cases include media streaming, business-to-consumer (B2C) marketing, risk and compliance in financial services, surveillance and security in the private sector, social media monitoring, and preventive maintenance strategies (Figure I.1). However, throwing big data at every analytics use case isn't always the way to generate the best return on investment (ROI).

Before we explore the big data fallacy in detail, we need to define analytics use case, a term you'll encounter a lot in this book. Here is a proposed definition:

"An analytics use case is the **end-to-end analytics support solution** applied once or repeatedly to a **single business issue** faced by an **end user** or homogeneous group of end users who need to make decisions, take actions, or deliver a product or service **on time** based on the **insights** delivered."

What are the implications of this definition? First and foremost, use cases are really about the end users and their needs, not about data scientists, informaticians, or analytics vendors. Second, the definition does not specify the data as small or big, qualitative or quantitative, static or dynamic—the type, origin, and size of the data input sets are open. Whether humans or machines or a combination thereof deliver the solution is also not defined. However, it is specific on the need for timely insights and on the end-to-end character of the solution, which means the complete workflow from data creation to delivery of the insights to the decision maker.

Now, getting back to big data: the list of big data use cases has grown significantly over the past decade and will continue to grow. With the advent of social media and the Internet of Things, we are faced with a vast number of information sources, with more to come. Continuous data streams are becoming increasingly

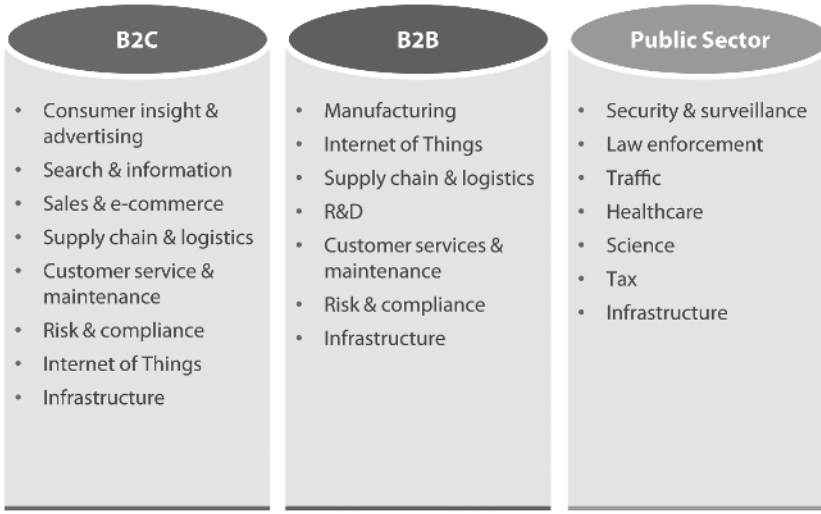


Figure I.1 Areas of Big Data Impact

prevalent. As companies offering big data tools spring up like mushrooms, people are dreaming up an increasing number of analytics possibilities.

One of the issues with talking about big data, or indeed small data, is the lack of a singular understanding of what the term means. It's good hype in action: an attractive name with a fuzzy definition. I found no less than 12 different definitions of big data while researching this book! I'm certainly not going to list all of them, but I can help you understand them by categorizing them into two buckets: the geek's concept and the anthropologist's view.

Broadly speaking, tech geeks define big data in terms of volumes; velocity (speed); variety (types include text, voice, and video); structure (which can mean structured, such as tables and charts, or unstructured, such as user comments from social media channels); variability over time; and veracity (i.e., the level of quality assurance). There are two fundamental problems with this definition. First, nobody has laid down any commonly accepted limits for what counts as big or small, obviously because this is a highly moving target, and second, there is no clear "so what?" from this definition. Why do all of these factors matter to the end user when they are all so variable?

That brings us to the anthropologist's view, which focuses on the objective. Wikipedia provides an elegant definition that expresses the ambiguity, associated activities, and ultimate objective:

> Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced

methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

High-ROI use cases for big data existed before the current hype. Examples are B2C marketing analytics and advertising, risk analytics, and fraud detection. They've been proven in the market and have consistently delivered value. There are also use cases for scientific research and for national security and surveillance, where ROI is hard to measure but there is a perceived gain in knowledge and security level (although this latter gain is often debated).

We've added a collection of use cases throughout this book to help give you insight into the real-world applications of what you're learning. They all follow the same format to help you quickly find the information of greatest interest to you.

ANALYTICS USE CASE FORMAT

Context: A brief outline of where the use case comes from: industry, business function, and geography

Business Challenge: What the solution needed to achieve for the client(s)

Solution: An illustration of the solution or processes used to create that solution

Approach: Details on the steps involved in creating the solutions along with the mind+machine intensity diagram, illustrating the change in the balance between human effort and automation at key stages during the implementation of the solution

Analytics Challenges: The key issues to be solved along with an illustration of the relative complexity of the mind+machine aspects applied in solving the case

Benefits: The positive impact on productivity, time to market, and quality, and the new capabilities stemming from the solution

Implementation: The key achievements and the investment and/or effort required to make the solution a reality (development, implementation, and maintenance, as applicable), illustrated where possible

I wanted to include some of the more exciting projects currently under development to show the possibilities of analytics. In these cases, some of the productivity gain and investment metrics are estimates and are labeled (E).

Innovation Analytics:
Nascent Industry Growth Index

Context

Organization
Corporate innovation departments, hedge funds, PE and VC firms

Function(s)
Companies investing in nascent industries

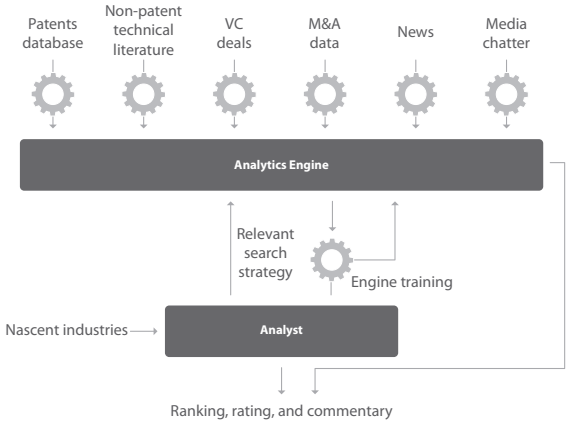
Industry
Corporates and financial services

Geography
Global

Business Challenge

- Build index to forecast probability of high near-future growth of specified nascent industries
- Read and interpret technical and business text from thousands of documents

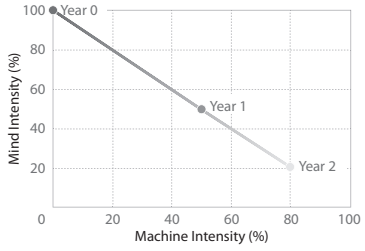
Solution



Nascent Industry Growth Index

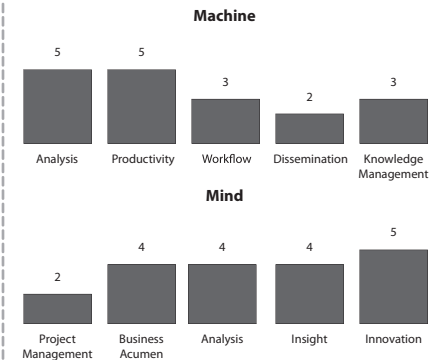
Approach

- Developed proof of concept to manually create industry-agnostic index
- Set up team of 4 FTEs (Full-Time Equivalent) (1 each from analytics and IP, FS, and corporates)
- Deployed text analytics tool (KMX) to automate innovation intensity ranking through free-text patent searching.
- Created production platform to aggregate disparate data sets and semiautomate growth forecast index creation



Analytics Challenges

- High variety of data sets to be integrated
- Iterative training of text analytics engine to accurately read large volumes of text
- Establishing judicious use of analyst involvement and time to screen relevant from irrelevant data
- Fine-tuning index over time to improve its accuracy in providing probability of high growth

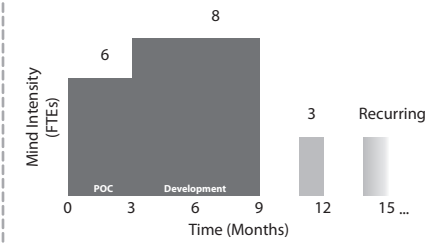


Benefits

Productivity	Time to Market	New Capabilities	Quality
• Reduces the need for an analyst by 50% to 90%	• Proof of concept built in 3 months • Product in 9 months	• Removed subjectivity in innovation prioritization criteria • Enabled faster prioritization of attractive markets and better judgment of next big industry wave	• Unmatched rigor in forecasting nascent industry inflection point • Initial accuracy is about 70%; improving with analyst intervention and iterative engine training

Implementation

- Proof of concept in 3 months with 6 FTEs
- Data warehousing, platform development, and testing for 6 months with 8 FTEs
- Thereafter, recurring engagement of 3 FTEs for 1 month per quarter
- Budget of USD 0.5 million in the first year
- Opex includes USD 0.1 million per year database cost and 2–3 FTEs



The big data hype has its origin in three factors: the appearance of new data types or sources, such as social media; the increasing availability of connected devices, from mobile phones to machine sensors; and the evolution of ways to analyze large data sets in short periods of time. The sense of possibility led to a proliferation of use cases. We cannot say how many of these untested use cases will survive. Ultimately, the question is not what *can* be done, but what *actually* delivers value to the end user.

Gartner predicts that 60 percent of big data initiatives will fail in 2017,[1] and Wikibon, an open-source research firm, maintains that the average ROI for big data projects is currently only about 55 cents on the dollar spent instead of the expected $3 to $4.[2] The latter assessment wasn't made by CFOs, but came directly from practitioners, who saw a "lack of compelling need" for big data in those use cases as a reason for the low returns. However, our experience is that CFOs are increasingly asking about the viability of such analytics.

For large companies, the investment in big data infrastructure and expertise can easily run into the tens of millions of dollars. It would seem obvious that prior to any such investment, the company would want to fully investigate the need, and yet in the 2012 BRITE/NYAMA "Marketing Measurement in Transition" study, 57 percent of companies self-reported that their marketing budgets were not based on ROI analysis.[3]

Measuring the ROI of analytics use cases is unfortunately not as easy as it sounds. This is especially true where companies have invested in infrastructure such as central data warehouses, software licenses, and data scientist teams. Properly calculating the desired impact at the use case level requires the corresponding governance and control, which is rare at this stage. In a series of initial interviews with companies that went on to become Evalueserve clients, seven areas were found to be lacking—in some cases, almost completely:

1. Governance structure for the data and use case ownership
2. Accountability for individual use cases, portfolio management, and associated economics
3. Clear definition of analytics use cases
4. Objectives and intended end user benefits for each use case
5. Tracking the actual results against the targets
6. Knowledge management allowing the efficient reuse of prior work
7. Audit trails for the people, timing, actions, and results regarding the code, data, and findings

That said, examples of excellent and highly focused big data use case management do exist. The use case *Cross-Sell Analytics: Opportunity Dashboard* shows solid accountability. The campaign management function of the bank

continually measures the ROI of campaigns end to end, and has built a focused factory for a portfolio of such analytics.

An example of a much weaker big data use case was recently proposed to me by a US start-up engaged in human resources (HR) analytics. The example illustrates some of the fundamental issues with the current hype. An ex-consultant and an ex-national security agent suggested using a derivative of software developed for the surveillance field for recruiting analytics. Based on the previous five to 10 years of job applications—the curriculum vitae (CV) or resume and cover letter—and the performance data of the corresponding employees, a black-box algorithm would build a performance prediction model for new job applicants. The software would deliver hire/no hire suggestions after receiving the data of the new applications.

We rejected the proposal for two reasons: the obvious issue of data privacy and the expected ROI. Having done thousands of interviews, I have a very simple view of resumes. They deliver basic information that's been heavily fine-tuned by more or less competent coaching, and they essentially hide the candidate's true personality. I would argue that the predictive value of CVs has decreased over the past 20 years. Cultural bias in CV massaging is another issue. Human contact—preferably eye contact—is still the only way to cut through these walls of disguise.

The black-box algorithm would therefore have a very severe information shortage, making it not just inefficient, but actually in danger of producing a negative ROI in the form of many wrong decisions. When challenged on this, the start-up's salesperson stated that a "human filter" would have to be applied to find the false positives. Since a black-box algorithm is involved, there is no way of knowing how the software's conclusion was reached, so the analysis would need to be redone 100 percent, reducing the ROI still further.

It was also interesting to see that this use case was being sold as big data. It's a classic example of riding the wave of popularity of a term. Even under the most aggressive scenarios, our human resources performance data is not more than 300 to 400 megabytes, which hardly constitutes big data. Always be wary of excessive marketing language and the corresponding promises!

These are just two isolated use cases, which is certainly not enough to convince anyone trained in statistics, including myself. Therefore, it is necessary to look at how relevant big data analytics is in the overall demographics of analytics. To the best of my knowledge, this is not something that has ever been attempted in a study.

At first, it's necessary to count the number of analytics use cases and put them into various buckets to create a demographic map of analytics (Figure I.2). One cautionary note: counting analytics use cases is tricky due to the variability of possible definitions, so there is a margin of error to the map, although I believe that the order of magnitude is not too far off.

Cross-Sell Analytics:
Opportunity Dashboard

Context

Organization
United States retail bank

Function(s)
Regional managers and financial advisers

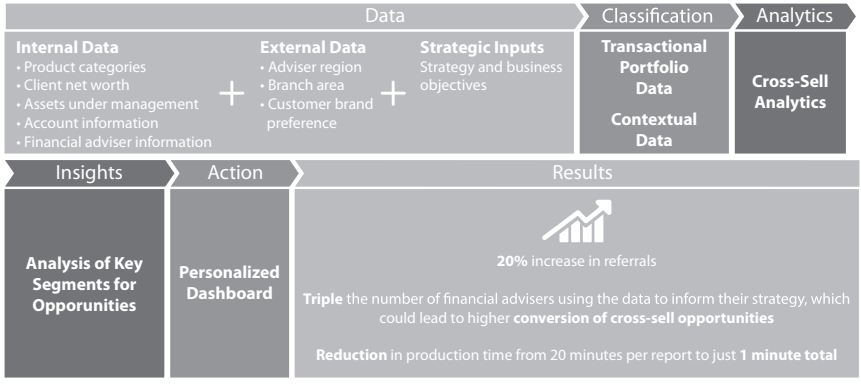
Industry
Retail banking

Geography
United States

Business Challenge

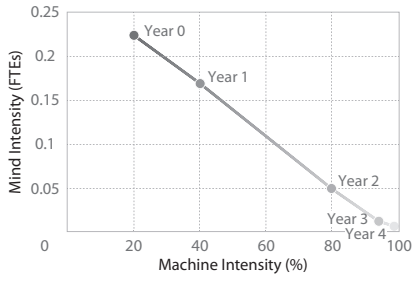
- Identify target customers and interesting products without centralized customer portfolios
- Efficiently and securely distribute customer data to support opportunity identification
- Give regional managers optimal oversight of their financial advisers

Solution



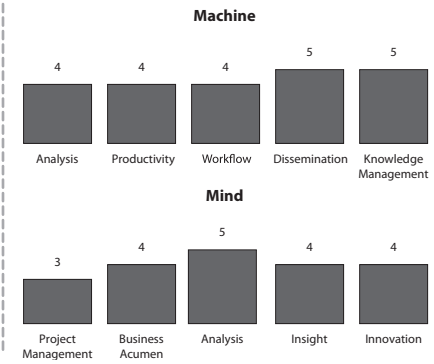
Approach

- Created a dashboard to allow financial advisers to identify the best cross-sell opportunities and generate individualized reports with suitable products for their customers
- Began auto-generation of weekly opportunity summaries for managers
- Implemented filtering so financial advisers only see their customers' information and managers only see information for their region



Analytics Challenges

- Decreasing data processing time to enable multiple iterations and replication across segments
- Working within financial advisory regulations regarding data security
- Ensuring only the necessary information reached the right person in the right format at the right time
- Providing appropriate monitoring for regional managers

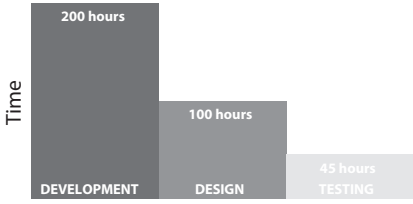


Benefits

Productivity	Time to Market	New Capabilities	Quality
• A 20% increase in referrals • Increased share of wallet for investment-only clients	• Report generation time down to 1 hour (from over 1 week)	• Stronger partnership for brokerage & retail banking divisions • Individualized reports for each financial adviser	• Increased efficiency and transparency for the financial adviser processes

Implementation

- Production time of 345 hours over 3 months (includes design, development, and testing)
- Rapid adoption: compared to their previous system, three times as many financial advisers use these reports to plan their cross-sell strategy



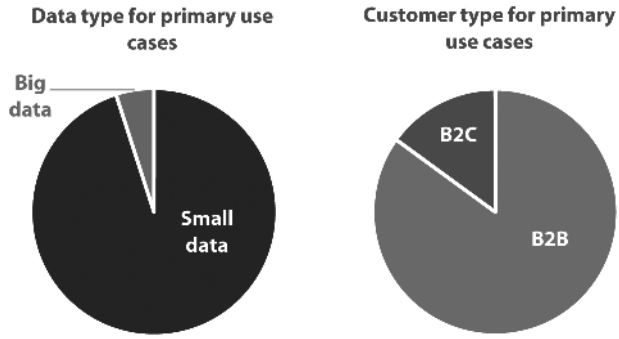


Figure I.2 Demographics of Use Cases

This map illustrates my first key point: big data is a relatively small part of the analytics world. Let's take a look at the main results of this assessment of the number of use cases.

1. Globally, there are a staggering estimated *one billion* implementations of **primary use cases**, of which about 85 percent are in B2B and about 15 percent in B2C companies. A primary use case is defined as a generic business issue that needs to be analyzed by a business function (e.g., marketing, R&D) of a company in a given industry and geography. An example could be the monthly analysis of the sales force performance for a specific oncology brand in the pharmaceutical industry in Germany. Similar analyses are performed in pretty much every pharmaceutical company selling oncology drugs in Germany.
2. Around 30 percent of companies require high analytics intensity and account for about 90 percent of the primary analytics use cases. International companies with multiple country organizations and global functions and domestic companies with higher complexity are the main players here.
3. The numbers increase to a staggering *50 to 60 billion* use cases globally when looking at **secondary implementations**, which are defined as micro-variations of primary use cases throughout the business year. For example, slightly different materials or sensor packages in different packaging machines might require variant analyses, but the underlying use case of "preventive maintenance for packaging machines" would still remain the same. While not a precise science, this primary versus secondary distinction will be very relevant for counting the number of analytics use cases in the domain of Internet of Things and Industry 4.0. A simple change in sensor configurations might lead to large

numbers of completely new secondary use cases. This in turn would cause a lot of additional analytics work, especially if not properly managed for reuse.

4. Only an estimated 5 to 6 percent of all primary use cases really require big data and the corresponding methodologies and technologies. This finding is completely contrary to the image of big data in the media and public perception. While the number of big data use cases is growing, it can be argued that the same holds true for small data use cases.

The conclusion is that data analytics is mainly a *logistical* challenge rather than just an analytical one. Managing the growing portfolios of use cases in sustainable and profitable ways is the true challenge and will remain so. In meetings, many executives tell us that they are not leveraging the small data sets their companies already have. We've seen that 94 percent of use cases are really about small data. But do they provide lower ROI because they are based on small data sets? The answer is no—and again, is totally contrary to the image portrayed in the media and the sales pitches of big data vendors.

Let me make a bold statement that is inevitably greeted by some chuckles during client meetings: "Small data is beautiful, too." In fact, I would argue that the average ROI of a small data use case is much higher due to the significantly lower investment. To illustrate my point, I'd like to present *Subscription Management: "The 800 Bits Use Case,"* which I absolutely love as it is such an extreme illustration of the point I'm making.

Using just 800 bits of HR information, an investment bank saved USD 1 million every year, generating an ROI of several thousand percent. How? Banking analysts use a lot of expensive data from databases paid through individual seat licenses. After bonus time in January, the musical chairs game starts and many analyst teams join competitor institutions, at which point the seat license should be canceled. In this case, this process step simply did not happen, as nobody thought about sending the corresponding instructions to the database companies in time. Therefore, the bank kept unnecessarily paying about USD 1 million annually. Why 800 bits? Clearly, whether someone is employed ("1") or not ("0") is a binary piece of information called a "bit." With 800 analysts, the bank had 800 bits of HR information. The analytics rule was almost embarrassingly simple: "If no longer employed, send email to terminate the seat license." All that needed to happen was a simple search for changes in employment status in the employment information from HR.

The amazing thing about this use case is it just required some solid thinking, linking a bit of employment information with the database licenses. Granted, not every use case is as profitable as this one, but years of experience suggest that good thinking combined with the right data can create a lot of value in many situations.

Subscription Management:
"The 800 Bits Use Case"

Context

Organization
Investment bank

Function(s)
Sell-side research

Industry
Financial services

Geography
Global

Business Challenge

- Collect and update subscription information by region, team, and analyst level
- Create a centralized information repository providing detailed subscription and license information
- Provide regular customized reports on usage and cost

Solution

Data sources

- Internal database
- Sourcing teams
- Business managers
- Subscription users

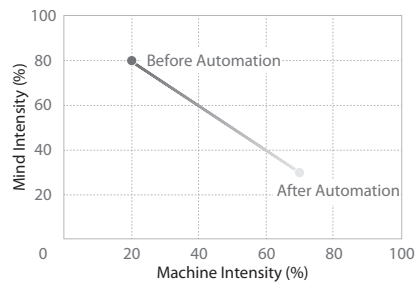
→ Define data and methodology to extract required data

→ Run codes to automate extraction process

→ Create dashboard with customization options

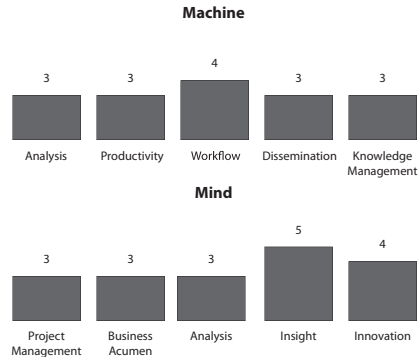
Approach

- Collated and analyzed the reporting requirements of client
- Consolidated the required data from multiple sources (internal database, sourcing teams, business managers, users)
- Automated the data extraction process
- Created dynamic dashboard to provide customized charts and tables on required parameters



Analytics Challenges

- Extraction of required data from multiple files from various sources
- Ensuring consolidation of data in a single file through appropriate coding
- Delivering consistent and easy-to-read visualizations

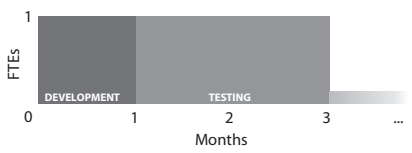


Benefits

Productivity	Time to Market	New Capabilities	Quality
• Centralized view of all subscriptions • Savings: over USD 1 million per year	• 70% faster time to delivery	• Ready availability of terms and conditions • Self-certification from subscribers to confirm requirement • Optimized usage & spend recommendations	• Zero error rate due to automation of data extraction and consolidation

Implementation

- 1 FTE developed customized automation tool in 3 weeks
- Sent first version of the dashboard for two monthly reporting periods for feedback
- Incorporated feedback and implemented the final version at the start of the third reporting month



This use case illustrates another important factor: the silo trap. Interesting use cases often remain unused because data sets are buried in two or more organizational silos, and nobody thinks about joining the dots. We will look at this effect again later.

Summing up the first fallacy: not everything needs to be big data. In fact, far more use cases are about small data, and the focus should be on managing portfolios of profitable analytics use cases regardless of what type of data they are based on.