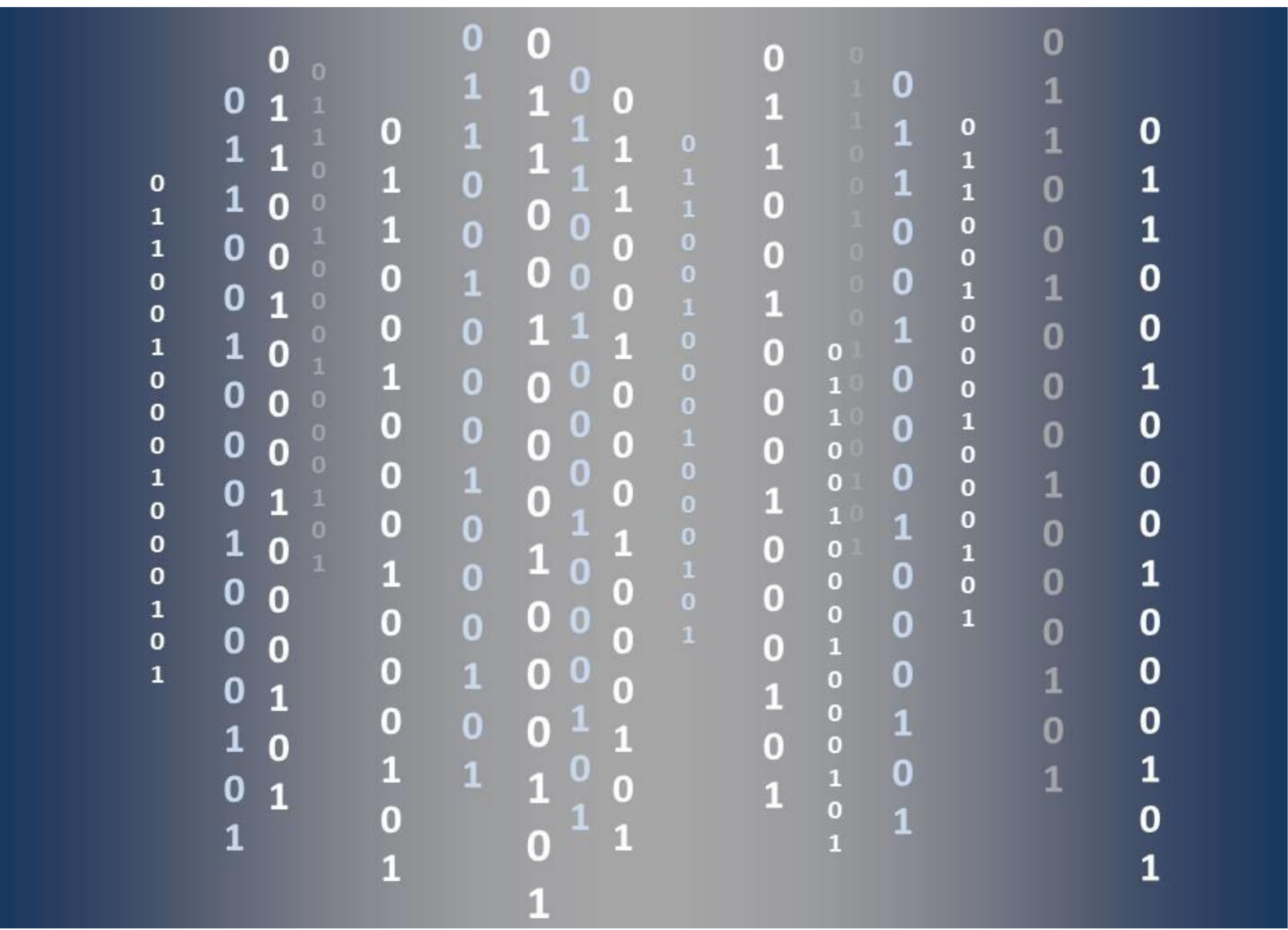
DataBerries

LARGE-SCALE DATA-DRIVEN DECISION-MAKING: THE NEXT REVOLUTION FOR TRADITIONAL INDUSTRIES

How new knowledge-extraction processes and mindsets derived from the Internet Giants' technologies will disrupt and transform traditional industries



Internet Giants such as Google or Amazon have long understood that data exploitation can provide key strategic advantages. They have suffused a "data mindset" throughout their entire organizations, and have developed the most advanced infrastructures and processing technologies to extract knowledge from data. This knowledge in turn fuels and enriches, in real time, the companies' decision-making processes. Yet, the Internet is not the only place where data is generated in large quantities. Various traditional industries also generate considerable amounts of data. Those companies that will be able to adopt a data-driven decision-making mindset will be able not only to significantly optimize their operations, but also to design new Value Propositions.

The Internet industry has become the undisputed leader in extracting knowledge from data.

The process that occurs when you load the sports page of your favorite website illustrates how far the Internet industry has gone with data. In less than 0.1 seconds, your website informs an online marketplace (called an Ad Exchange) that the front banner of its sports page is for sale. The marketplace then asks hundreds of buyers how much they are willing to pay to display their ad on the banner. Each of the buyers estimates the probability of your clicking on their ad, from which they deduce a business case and determine the maximum price to pay to display their ad on the banner. The marketplace then selects the highest offer, and the corresponding ad is displayed on the sports page of your website.

But what does this elaborate bidding procedure have to do with the extraction of knowledge from data? The estimate of the probability that you will click on the ad is made by big data algorithms that take into account a quasi-infinite number of parameters, including the weather, time of day, device you are using, your location, your entire available browsing history (from which they can deduce your gender, age, tastes, social class, etc.), as well as the webpage of the banner, etc.

And there is more. Even the ad that you are shown is designed in real time by algorithms that choose the color, offer, and message that have performed best among users who have similar profiles to your own. And there is even more. All of the data generated by this single transaction is then stored and re-used to perfect the algorithms. Each time a webpage is loading, this entire process starts up again in order to determine which ad will next be displayed on your screen. With more than 5 billion transactions occurring every single day in France, *a lot* of data is generated, stored, and exploited.

The Internet industry thus exploits data through a continuous cycle, which consists of three stages:

- Testing (for instance, different colors of a same banner)
- Measuring (the performance of each color)
- Learning (which color performs best with a given profile, based on past interactions)

Tremendous value can be created by adapting these digital techniques to the needs of traditional brick and mortar industries.

Although tests may be harder to implement, and data more difficult and/or costly to collect, in traditional industries as opposed to online, some brick and mortar pioneers have successfully adapted the aforementioned digital techniques to their businesses, thereby creating significant value.

UPS has fitted its truck fleet with sensors as well as GPS and wireless modules. By analyzing the real-time data thus captured, and comparing it with historical data, UPS is now able to optimize its delivery operations in numerous ways. For example, it can anticipate engine troubles and thereby reduce maintenance costs. It can also optimize travel routes in order to lower energy consumption, avoid traffic, or even reduce the likelihood of accidents. In 2011, UPS shortened its trucks' total travel distance by 30 million miles thanks to its big data analytics program.

In the USA, some insurance companies offer to place a wireless module in their clients' cars. Drivers can then subscribe to an insurance contract that is priced according to their driving habits. Following a test phase, the algorithms predict which driving behaviors are more likely to lead to accidents. The algorithms then continue to develop greater accuracy once the service is deployed, thereby allowing for further price fine-tuning going forward. This is a highly disruptive shift for the insurance industry: a new model, based on individual behavior, is replacing the traditional "pooled risk" paradigm.

Rolls Royce is a jet engine manufacturer and has invested in fleet monitoring technologies for more than 3700 of its engines. With the large amounts of data thus generated, the company has learned how to identify future problems before they even manifest themselves. Quantity of data is key here: an airline monitoring 200 engines will gain far less knowledge than Rolls Royce does when it monitors 3700 engines. Consequently, Rolls Royce is in a far better position to maintain and repair its fleet, and can achieve a lower total ownership cost for its engines. Engine servicing now accounts for more than 70% of the jet engine division's revenues.

On a more general level, a 2011 MIT study by Erik Brynjolfsson has shown that firms that promote data-driven decision-making achieve productivity rates that are 5 to 6% higher than those of their peers. Importantly, the relationship between data use and productivity seems to be a causal one.

New competencies are needed to extract value from data; competition to attract the most talented profiles will be intense.

Capturing value from data requires specialized competencies in the field of data science, big data, and machine learning. Building an efficient and reliable data science team is not an easy task: first, because this field is usually unknown or obscure to current management, and second, because the requisite skills often lie outside of the core competencies that are most prevalent in traditional industries. A data science team needs to combine:

- Deep understanding of the company's business issues
- Full proficiency in data-mining techniques and algorithms
- Solid knowledge of big data infrastructures, database technologies, and programming languages

To hire such qualified individuals, traditional industries will have to compete with very attractive Internet Giants such as Google, Facebook, Amazon, or Netflix, as well as Silicon Valley's start-ups and other tech players. As Google's Chief Economist Hal Varia acknowledges, *"data is so widely available and so strategically important that the scarce thing is the knowledge to extract wisdom from it. That is why statisticians, database managers, and machine learning people are really going to be in a fantastic position."*

Ultimately, the ability to use data to enhance decision-making processes will become a key strategic competency across many industries. Not only data itself, but also, and especially, the *ability to extract knowledge* from data, will be increasingly regarded as a strategic asset. In this context, it is no surprise that McKinsey's Global Institute is predicting a shortage of data scientists by 2018.

The extraction of value from data relies on a well-defined process.

Although there is an underlying process, extracting value from data remains an uncertain endeavor. It is not unusual that the value extracted is different from what one was looking for initially. A data science team's competencies are put to use throughout an iterative five-stage process.

1. **Define the business problem.** This might seem obvious, but all stakeholders must be aligned and have a common understanding of the ultimate business objectives of their data mining project. The data science team can then design a plan to reach the agreed-upon objectives.

2. **Understand the data** from which the knowledge will be extracted: where does the data come from? What are the biases? What information could be missing? Sometimes, it will be necessary to invest in creating data, for instance by sending promotional offers at random and monitoring responses. Usually there will be more than one data source, so it is important to understand how different sources can be linked together.

3. **Prepare the data** that will be used to build the model. The raw data will need to be cleaned and formatted before being utilized in the construction of the model. For instance, you may have retrieved all of a retailer's cash outflows for a given year, but the data you will then inject will only be (for each customer) the share of promotion in their purchases, the frequency of visits, and the share of spending in each retail category.

4. **Build the model.** This is the stage where the data-mining techniques and algorithms are applied to the data. What does a big data algorithm do, and why is it so powerful? The algorithm tries to determine the common characteristics of entities that belong to a same group. Basically it does what a human would do, but without cognitive limitations in terms of numbers of parameters it can handle and amount of information it can digest and retain. Fueled by historical data, the model can then analyze the characteristics of an entity that it has never seen before, and assign that entity to a group.

5. **Test the model.** Once the model is built, one must assess how it will perform with historical data that has not been used to build it. Such data is called a test set. The test set provides an evaluation of how well the model will perform in real business conditions, without taking any operational risks. In a classic development cycle, the lessons learned from the test are used to launch a new iteration of the five-stage process.

The Big Data value chain rests upon three fundamental elements: data, technical skills, and ideas.

Through their logistics and delivery operations, companies such as UPS and FedEx are amassing reams of information on merchandise flows from all around the globe. An American hedge fund, which wishes to remain anonymous, had the intuition that these delivery records would correlate with numerous macroeconomic indicators, and could thereby provide valuable insights on entire industries. Assisted by technical experts, the hedge fund mined the companies' data and used it to design an algorithm that can predict the evolution of macroeconomic indicators before they are published. Today the hedge fund uses those predictions to fine-tune its investment strategy.

The above experience illustrates how the Big Data value chain is structured. To capture the full potential of Big Data projects, three fundamental elements must come together: data, technical skills, and ideas. One can also think of this chain as: the raw material, the transformation process, and the design.

Data is commonly generated from business operations and customer interactions of large, well-established companies. Such companies collect information primarily for reporting, consolidation, or basic monitoring purposes; they do not necessarily have the skills to extract value from their data. However, some emerging companies now propose to handle data monetization. They gather numerous data sources and manage licensing strategy.

Technical skills generally lie with IT consultants, technology vendors, or analytics providers who have developed special expertise in managing big data. These technical experts master the infrastructure and database technologies, and can extract knowledge from data. Most of them focus on the technological and technical aspects of big data; they do not primarily focus on generating disruptive business ideas.

Ideas can flourish from anywhere, but they usually come from small players who have a big data mindset. Based on their more advanced understanding of the potentialities offered by the data, such players imagine innovative services or products. They are not afraid to explore their most creative ideas.

Usually, traditional industries own large amounts of data. However, with most of their attention devoted to core business operations, brick and mortar companies currently lack the resources necessary to develop both innovative ideas and technical knowhow to extract valuable knowledge from their vast troves of data. As a result, they may struggle to implement large-scale data-driven decision-making processes. To overcome these barriers, traditional brick and mortar companies have two options. They can start building data science teams and thus participate in the talent war against Internet Giants. Alternatively, or additionally, they can partner with smaller and more innovative players who will bring both fresh ideas and technical knowhow. By working with such players on specific business ideas, traditional industries will gain more knowledge about big data techniques, and will gradually modernize their mindset and culture. Ultimately, they will capture the full potential of large-scale data-driven decision-making.

Author

François Wyss, Co-founder, DataBerries
francois@databerries.com

About DataBerries

DataBerries masters the latest predictive big data technologies and leverages the power of mobile digital advertising to drive additional traffic in points of sales. Amongst others, our products embed the following technologies:

- Hadoop Spark predictive modeling
- Geolocation based targeting
- Mobile RTB Ad Exchange platforms

DataBerries was founded by François Wyss, former Sales Manager at Google and Consultant at A.T. Kearney; Benoît Grouchko, former Product Manager at Criteo; and Guillaume Charhon, former CEO at Arise.io.

Contact:
contact@databerries.com
(+33) 6 75 85 94 36