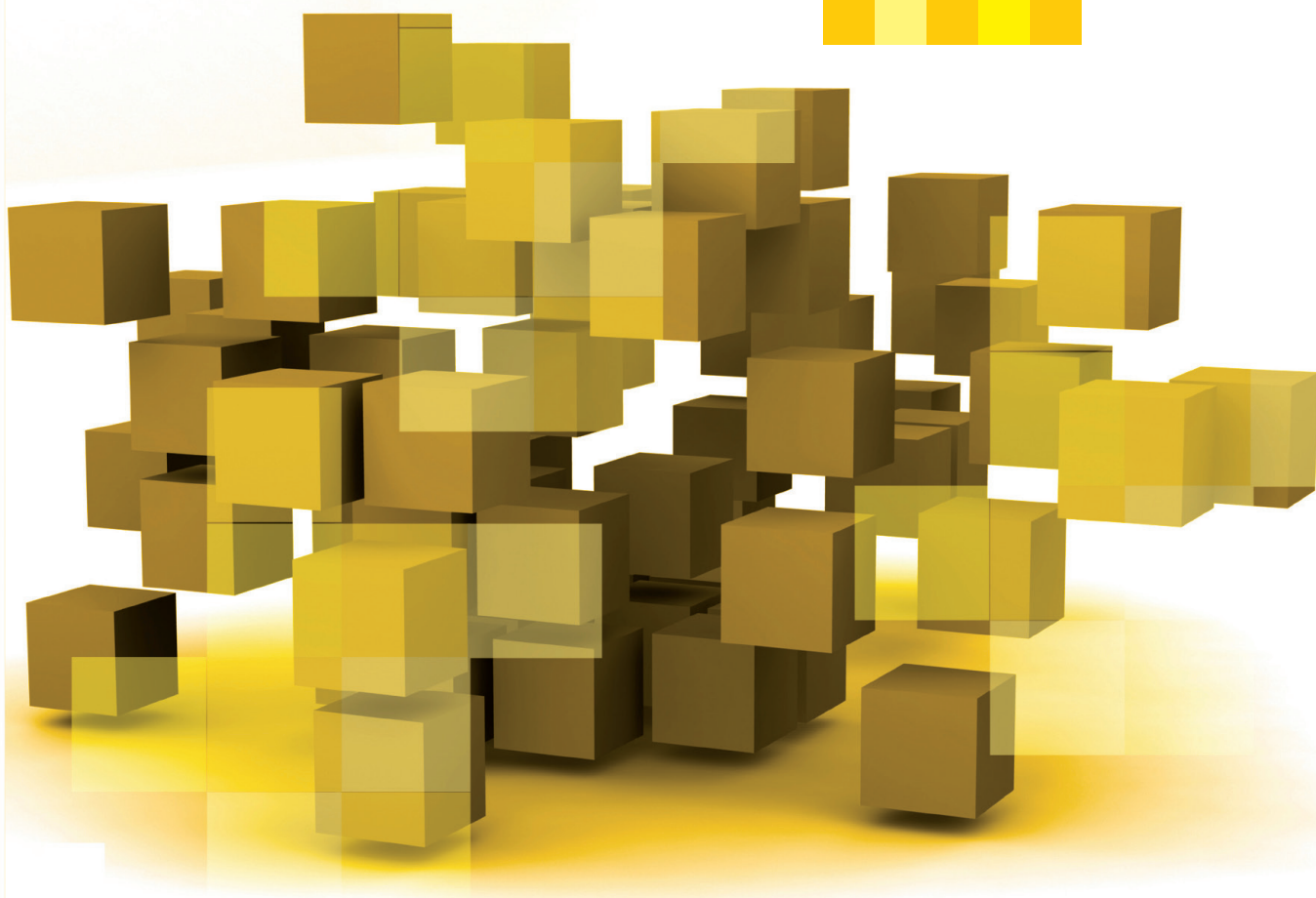


Managing data in a hyperconnected world

Les cahiers de veille de la Fondation Télécom



FONDATION
TELECOM



Editorial

May 2013

The digital transformation is not a linear phenomenon. It proceeds by waves of innovation, rich with their own issues and their own dynamics.

After the wave of computing and the wave of the Internet, we are probably entering a third sequence, the data wave: changes that we know for the most part, but which deserve to be examined as a whole.

The data revolution is, of course, the explosion of data available in organizations or accessible over the Internet. It is the falling cost of data production systems, which allows anyone to deploy information systems they would not have dreamed of a decade ago. It is the flow of radical innovations that hide behind words like «big data» or «cloud computing». It is also political and social change: the relationship between citizens and their own digital identity revealed by the Quantified Self movement, the privacy concerns, the growing demand for transparency in government agencies and large organizations. It is new government practices, such as open data, leading to new forms of open and collaborative government. And new risks too, especially in the cyber security domain...

In our organizations, the data revolution means new kinds of jobs that modify the ways we work and represent of the world: datascientists, data visualizers. It brings new concerns too, such as data governance, and new training needs that our schools must respond to.

It is, by definition, difficult to measure a wave of radical innovation. One always starts by comparing the new with the old, with what we already know: «Data is the new oil for our economies... A treasure to keep or distribute... A danger to privacy...» These analogies are not false, but they are incomplete and may cause us to miss the point.

The key point is that the flow of data exchanged on the networks now forms the backdrop of the economy, an important part of the social bond, a tool to forge new forms of public action, an essential component of our digital identities... and a treasure of risks and opportunities.

If I had to make an analogy, I would suggest rather: our economy is in the middle of the same revolution that the life sciences lived when biochemistry was discovered. With the ability to work effectively on large distributions at very granular levels, we are in a position to rebuild all our old learnings, to open new ones, and imagine radically new forms of intervention.

There are big challenges to come, indeed...

Henri Verdier

Director, Etalab

Contents

3	A deluge of data
3	Predicting the future with new amounts of data
5	Completely new horizons
6	Why is (big) data an opportunity now?
8	From Big Data to Small Data: giving the power back to the people.
9	The data landscape
9	What is data?
9	What is generating data?
10	Cleaning and contextualizing data
11	Processing data
11	Different types of data (structured and non structured)
12	Managing the petascale computing
14	The data market
16	Unlocking data value
16	Machine learning
18	Realtime analytics
18	Data analysis for everyone
19	Data visualization: telling stories with data
20	Rise of the data jobs
21	Privacy and trust concerns
22	Challenges in the air
22	Scientific challenges
	Towards a better understanding of the world? // Pour a maximum of data and solve big problems // Build and share data-oriented infrastructures // Learn to apply context to the numbers // Make the data qualitative and meaningful
24	Technical challenges
	Will "delete" become a forbidden word? // Anonymize for good // Do not neglect big data risks // Master data correlation // Master the big data cycle // Make the mobile phone your data assistant // Enable dataviz on new devices // Invent the future of shopping
26	Societal challenges
	Open new ways of thinking // Democratize data management // Teach the future datascientists
27	Working with the <i>Institut Mines-Télécom</i>
27	Glossary

A deluge of data

Big Data is a slippery concept, and big data-sets sizes are a constantly moving target, currently ranging from a few dozen terabytes to many petabytes. Big data is data that is too large, complex and dynamic **for any tools, procedures and processes available at the moment** to create, capture, store, manipulate, manage and analyze.

In this *cahier de veille* we write the term *Big Data* with capital letters sparingly. We use *data* when it refers to ordinary data, open data, fast data, big data... whereas *Big Data* with capital letters refers to the marketing phenomenon.

Data law #1: The faster you analyze your data, the greater its predictive value.

From David Feinleib's 8 laws of Big Data.
See references page 8.

In 2008, Google was able to spot trends in the Swine Flu epidemic two weeks before the U.S. Center for Disease Control by analyzing searches that people were making. Nowadays, Google Flu Trends (blue line) uses aggregated Google search data to estimate current flu activity around the world in near real-time. Figure on the right presents French results, compared with data collected from the public Sentinelles health network (orange lines).

Source: <http://goo.gl/SUqJd>

BIG DATA HAS BEEN DESIGNATED trend of the year 2012 and continues to make the headlines in 2013: reports, conferences and announcements follow one another at a furious pace, making it more and more a part of the data landscape. To the deluge of data associated with big data adds a flood of information on the phenomenon, and it is more necessary than ever to see more clearly. Over the past few years, big data has expanded to fast data and smart data, open data from public administrations, dark data that need to be revealed, and small data that is more specific. It is now time to learn how to manage all these flavors of data.

We choose in this *cahier de veille* to explore data from a particular angle: *managing* the data from the traces we leave in both the real world and the connected world, traces related to the digitalization of our lives. Everyday objects (phone, tablet, car, bathroom scales) all now have a software

component that becomes dominant in the design of these objects. These programs generate traces in large quantity and in real time, giving their location, their task, the identity of the user, their internal state, and even information about their environment, like the network status or the level of attacks to which they are subject. These traces can be easily exploited in the connected world to extract a user profile to best adapt the service provided by the object to the user needs, but also to the interests of brands (with a balance between privacy and economy), to get a state of the object in order to detect the need for maintenance, and to perform better inventory management to improve product availability.

Big data is not for big business only. Data is no longer an isolated IT discussion. We are now quantifying every aspect of our lives, and the data we generate, own and publish must remain under control, whether you are a small organization or an individual.

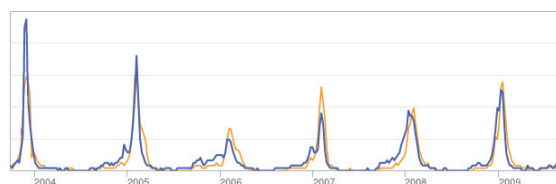
Predicting the future with new amounts of data

Managing large amounts of data is not new, and «big data» has been a concept known by IT manager for years. However, producing large amounts of data, accessing massive new data sources and having all of them at our fingertips marks a change in the way we work.

New insights come from sources that were not known before or were previously impossible to analyze. What is huge today is that businesses can leverage the entire Web as a data source, and at relatively low costs. One of the first uses of big data hitting the public attention was the Google flu trend service, launched in 2008. Google found at that time that certain search terms, among the millions of users queries around the world for health information, were good indicators of flu activity during flu season. These results were published in the prestigious journal Nature.

Google compared the flu-related query counts with traditional flu surveillance systems and found a *pattern* emerging between how many people search such queries and how many people actually have flu symptoms. The same close relationship could be obtained worldwide, and updated every hour or so, when traditional surveillance services could only update their results once a week, on a country basis. A new set of data (the Google queries, compared to counting people visiting a doctor) was delivering a totally new set of information that could be used as a complement to predict phenomenon, and this opened the path to the research on web-produced data.

While Google had somehow neglected to update its Google Trends service before 2007, it now updates information daily, and Hot Trends is updated hourly. It is worth noting that this is a publicly available service for everybody.



10^{24} Yottabyte

our digital universe

today

10^{21} Zettabyte

1.3 ZB of network traffic

by 2016

10^{18} Exabyte

250 million DVDs

Diameter
of the Sun

France largest
inner distance

10^{15} Petabyte

In May 2013, Microsoft migration
of Hotmail accounts was over
150PB of user data.

10^{12}

Niagara Falls
length

Terabyte

Boeing Jet generates 10 TB
of data per engine every 30
minutes.

10^9 Gigabyte

10^6

Megabyte



If 1 Megabyte
represents the
size of an ant...

3 V

However, flu prediction is mostly a way to make a monitoring process more accurate, and other sets of web collected data can give better insights on the future. This is particularly the case with data in the financial market context. Indeed, a recent scientific study published in Nature in early 2013 showed that it may be possible to quantify trading behavior in financial markets using the Google Trends service.

As the crises in financial markets affect humans worldwide, being able to better understand, and predict, stock market moves, will be surely of some help. These trading decisions reflect the complexity of human behavior. The authors of the study suggest that «*massive new data sources resulting from human interaction with the Internet may offer a new perspective on the behavior of market participants in periods of large market movements.*»

Their findings show that *Google Trends data can reflect the current state of the stock markets but may have also been able to anticipate certain future trends.* They found patterns that may be interpreted as **early warning signs** of stock market moves. It could be generalized to other complex systems that human being faces, the environmental questions for instance.

In fact, all is contained in that «may be interpreted». There is a short step from using data to interpret the world and using it to act on it,

so what's going on if decisions are wrong? On April 23, the very same week the stock market article was published in Nature, a bogus Twitter message on a false report of explosions at the White House was published via the Associated Press account and caused financial markets to drop sharply for a short period. This was evidence for some that automatic algorithms were using social network data and could take decisions without waiting to substantiate things.

Predicting moves as a function of past behavior can also be done with web collected data from its own customers. Lilligo.com, a real-time travel search engine, simplifies the user experience, by providing information about past price evolution and forecasts based on historical data. As it constantly records the result pages, the information collected allows the visualization of the price time series for every ticket, and can be used to gain insight into the effect of the yield management policy conducted by the travel companies. Based on vast streams of heterogeneous historical data collected through the Internet from more than 250 travel sites, agencies and tour operators, researchers at Institut Mines-Télécom proposed approaches to forecasting travel price changes at a given horizon, taking as input variables a list of descriptive characteristics of the flight, together with possible features of the past evolution of the related price series.

what is

When are we talking big?

Big Data is often defined first through the three «V's» – **Volume, Velocity, and Variety**, volume being the facet that is most discussed.

Nowadays, humanity produces in two days more data than throughout its history from its beginnings until 2003. In eight years, the **volume** of data will be fifty times greater than it is today. The big data market will reach €50 billion and 10 times more servers will be needed. Big data is less than ten years old, but the technology is already massively deployed by Facebook, Twitter, Google... As a comparison, every minute on average 350,000 tweets and

15 million SMS messages are sent globally. In 2011, the volume of data found on the Internet was 1.8 zettabytes and a volume of 2.9 zettabytes is expected for 2015.

Big data can come fast. The **velocity** is defined as the pace at which the data are to be captured and consumed (referring to both streams of data coming from Internet of Things and the speed of growth in large-scale data repositories). Just think of algorithmic trading engines that must detect trade buy/sell patterns in datastreams coming in at 5,000 orders per second.

Data processed by large web companies is not structured and neatly formatted. The users re-

Completely new horizons

“*The data streams coming from our devices are challenging the traditional approaches to data management.*”

Data is generated and collected all around us every second, and this opens completely new horizons when managing our everyday life in a hyperconnected world. First of all, the mobile phone in our pocket can track every movement you make and every sound you hear. It «knows» when you are at home or away. Its accelerometer sensors even «know» this is you wearing the phone, by how you walk. And with the advent of the Internet of Things (see our *cahier* #3), all sorts of sensors are monitoring you, in your car, your fridge, your scale and your toothbrush, billions of devices that can sense, communicate, compute and potentially actuate.

All the data streams coming from these devices are challenging the traditional approaches to data management, and can even infer usages not thought of before. Here is another example of the new kind of data being captured. The London transportation agency has deployed significant amount of sensor roadways in order to monitor traffic in real-time to optimum traffic management during the 2012 Olympics. This includes surveillance cameras in all the underground railway stations, parking areas, light rail stations and piers, at the bus stations and on the buses themselves. These cameras operate in real time, catch people speeding on the roads, and can record data in an analogue or digital format. They were deployed with the focus of collecting traffic information. However, other organizations would like to access this sensor data so they can use it for a different perspective, traffic updates or weather information for instance. The London transportation agency would provide a so called Sensor network as a Service, and these third party organizations would provide their customer with value added services. Indeed, the real value of the London video surveillance system would come from understanding the meaning of the images themselves, not just from the metadata associated with them.

The next example demonstrates the application of well known web insights to real world premises, along with the analysis of meaningful images. What if you could get analytics on your sales in the same way you get logs from your website, such as the number of visits, the bouncing rate (website immediate exit), the length of time on

each page/visited pages, the previously visited sites, the transformation from visitor to customer? Clirisgroup is a company once hosted in the Institut Mines-Télécom incubators, that measures every flow via video-analysis, inside, outside and at each key point of a retail outlet. These flow-measurements are strengthened with the data of the store information system itself: staff time-schedules, realized sales, etc. and are integrated with external data: weather, roadworks, marketing operation inside the retail outlet, etc. This allows the calculation of the store and universes attractiveness, measurement of the impact of communication campaigns, optimization of the advice and sales forces, measurement of the impact of externalities and unforeseen circumstances... When the customers are digitally identified (via mobile payment, QR code flashing...) these online data could be correlated with real world data and give specific insights never seen before.

In some use cases, concerns about big data big brother are arising. Beginning with companies, where the analysis of e-mails, instant messaging, phone calls, and mouse click (clickstreams) can now be employed in the quest of greater business efficiency. Indeed, the data produced by workers is becoming a valuable asset. Applying *datascience* to human resources is more and more appealing among academics and entrepreneurs. Gild, an 18-month-old start-up company, provides services to automatically discover talented programmers. In order to predict how well a programmer will perform in a job, Gild is tuning algorithms that evaluate the candidates through no fewer than 300 variables. At a time when recruiters do not have enough time to read all the resumes they receive (they spend an average of 6 seconds per CV), and where automated systems are struggling to sort the candidates, big data and machine learning systems could reduce human bias during the selection process to identify the best candidates.

Back to sensors. The Bank of America has equipped 90 of its employees with badges to study their movements and interactions. Data collection told an interesting story, and the bank decided to promote the taking of breaks in groups rather than alone. This increased the productivity by 10%!

big ?

quire the ability to store, query, and integrate results across a **variety** of information types, including text, image, audio, video... This is the third V in the Big Data definition. And more V are to come, as the reader will see later.

«Big» is a relative concept. EMC has defined big data as «*any attribute that challenges constraints of a system capability or business need*». For example, a 40MB Power Point presentation is big data, as is a 1TB medical image.

«Big» relates to the data itself, the size of the dataset, the velocity, the number of data and the types, or any combination of these.

Why is (big) data an opportunity now?

“ Four effects & five major trends. ”

The foundations for big data were laid down in February 2001 by Gartner analyst Doug Laney, following SGI Chief Scientist John Mashey's seminars in '97 who emphasized on the volume challenges. In an analyst report, he anticipated both the growth and impact that big data would have on technology, business and day to day life. He was the one who described big data as a 3-dimensional data challenge of increasing data volume, velocity and variety, and characterized big data as *«data that's an order of magnitude greater than data you're accustomed to.»*

Whereas big data has been around for a decade now, experts agree to say that 2013 is the year when it will find its ways out of the data centers and the tech rooms. The chief financial officer, chief marketing officer and chief sales officer, who are all in charge with growing revenue, are now aware what sort of opportunities big data are offering.

There are four effects and five major trends causing organizations today to rethink the way they approach data management.

1. **The Hardware effect:** a significant drop in the cost of hardware as a whole and its commoditization (servers, storage, network ...). This enable to distribute data through large clusters of nodes, at relatively low costs.

2. **The Tools / Framework / Software effect:** the development of open source load balancing systems on distributed architectures for very large volumes of data, among them the *Hadoop* and *MapReduce* projects. Many of these tools are available free under the open source licensing, helping to keep the cost of big data implementation and management under control.
3. **The Data effect:** the emergence of fast growing data volumes much larger and more heterogeneous than before. A lot of unstructured contents are produced everyday, for the most ill-suited to traditional relational databases, but full of key information for businesses and administrations. The multiplication of traces on the Internet and the proliferation of data from the real world through a wider network of sensors generate large amount of data that is more real-time oriented.
4. **The Market effect:** big data has a huge economic and scientific potential. The data accumulated by the services industry have within them significant value to the knowledge of customers. This will give the path to the revolution in real personalized services to best meet the needs of users.

These effects will be of no use if not accompanied by the trends which are changing minds both within companies and ordinary people.

quantified-self

From Flu Trends to Quantified Self: how the healthcare sector can benefit on Big data

Typical areas that produce large datasets are meteorology, space exploration, genomics, physics, simulations, biology medical research and environmental science. Some of the potential application areas of big data analytics are smart homes and smart grids, energy management, (cyber)security and automation, law enforcement, terrestrial, air and sea traffic control, transportation, location based systems, urbanism, telecommunications, search quality, manufacturing, retail, online marketing, customer service, billing, trade analysis, financial market and services, fraud and risk management. Defense applications, business transactional systems, embedded systems are some exam-

ples of existing applications producing high velocity data.

Among all, healthcare is flooded with a deluge of data, collected from multiple sources, from professionals – lab results, biometrics, medical claims, pharmacy claims, point-of-care – to individuals themselves through the internet and social media as with the Google Flu Trends.

At a larger scale – hospitals, regions, countries – the health information exchanges initiatives can *«be utilized for medical research, contributing greatly to evidence-based medicine, better assessment of incidence, prevalence and*

causative analysis on certain diseases», as reported by analysts at Wipro. \$300 billion was the estimated potential induced by Big Data for the healthcare sector in the US in 2012.

The healthcare sector regularly invites developers to work on their use cases. One of last year Health 2.0 hackathon in Boston rewarded a team that created a website «No Sleep Kills», through which people can see how poor sleeping patterns can lead to drowsy drivers and auto accidents. This data analytics are useful inputs for different users: vehicle industry, insurance, and customers. [<http://goo.gl/KomYh>] The Health 2.0's SXSW Code-a-thon, held in march 2013, was sponsored by BodyMedia, a pioneer in the development of wearable body monitors that

1. **Big data is no longer confined to the technical floors:** like virtualization and cloud computing in the recent past, and in some ways a continuation of this, it is becoming obvious that spending in the big data is worth the penny, especially as the cost for analytics is minimal compared to the cost for a traditional data warehouse. Moreover, big data cuts significantly data integration costs, and gives tools for data exploration that have never been seen before, offering new ways to use and understand data.
2. **Big Data is now mainstream:** both the open source community (for instance Hadoop) and long established IT companies (among them IBM, Microsoft, Oracle, SAP...) have joined forces to enable companies and administrations to invest in confidence in Big Data. This will not be a short hype cycle or the next bubble, but rather an underlying trend for the coming years, based on technical and social breakthrough innovations.
3. **Skeptics can be overcome:** big and small companies are beginning to see good ROI from the insights provided by the use of Big Data. It is especially true when they can better understand their businesses and the businesses of their competitors.
4. **The Chief Information Officer and the Chief Marketing Officer can better work together:** once these company executives experience the strengths of big data analytics and how they can better understand their customer, they will never go back to their previous practices.
5. **The power of analytics are now available to the masses:** following up the XaaS (X as a service) family, BDaaS (Big Data as a service) providers are emerging with an entire stack of services (acquisition, storage, analytics, visualization...) easily usable by companies of all sizes, even start-ups with no funds and individuals with no skills. This gives everybody a chance to play with the Big Data ocean.

However, there has been some warns on «Big Data» fatigue in the first quarter of 2013. People seem to get stuck on the 3 V's definition of Big Data, and other V's are proposed (we emphasize them in this *cahier de veille*) that are all pertinent. But enterprises do not want everlasting promises made years ago about the benefits of big data and analytics, they want solution, and that may be on not ever bigger data. It is now time to go beyond the marketing campaign, to use «Big Data» with parsimony in order not to be filtered out, if we do not want to completely miss the rise of the data-driven economy.

Dave Feinleb (see references page 8) adds 3 I's to the Big Data pot. [<http://goo.gl/NTrk0>]

- **Immediate** – in the sense that you need to do something about it now
- **Intimidating** – what if you don't?
- **Ill-defined** – what is it, anyway?

collect physiological data for use in improving health, wellness and fitness. Developer had to use the BodyMedia **API**. [<http://goo.gl/Tz789>]

Wipro analysts explain how it is possible to reshape the healthcare sector with new technologies: «As healthcare goes now far beyond hospitals, it is possible to have wearable, even internal, sensor-based devices to monitor vital signs and symptoms of patients. From clothes embedded with sensing devices to headsets that measure brainwaves, wearable devices can be seamlessly incorporated into the ensemble.»

According to IMS Research, the wearable technology market was worth \$2 billion in 2011 and will reach \$6 billion by 2016. The findings

reveal that 14 million wearable devices were shipped in 2011, and that number is likely to reach 171 million in 2016.

The French Withings body scale connects to various Health 2.0 services such as Google Health and Microsoft HealthVault as well as diet and exercise sites such as DailyBurn. Cityzen Sciences is another French company created in 2008 that specializes in smart textiles conception and development. Smart textiles are embedded with micro-sensors enabling them to monitor temperature, heart rate, speed and acceleration as well as to geolocate. Beyond Cityzen Sciences activities, the entity Cityzen Data, in the Telecom Bretagne incubator held in Brest, will be in charge of managing the storage

of data produced and extracted by smart textiles, then from other sensor networks, in an anonymous and secure environment, and in charge of developing a high-added-value service offer based on the collected data.

Monitoring him/herself and voluntarily sharing the data on web services is called **Quantified Self**, a «movement to incorporate technology into data acquisition on aspects of a person's daily life in terms of inputs (e.g. food consumed, quality of surrounding air), states (e.g. mood, arousal, blood oxygen levels), and performance (mental and physical). Such self-monitoring and self-sensing, which combines wearable sensors (EEG, ECG, video, etc.) and wearable computing, is also known as **lifelogging**.»

From Big Data to Small Data: giving the power back to the people.

“ Owners and users of the data can be individuals, not only large organizations. ”

And here is a possible miss. Big data is actually mostly customer data: data about customer behavior. From data collected on the web, the marketing team want first to quickly identify what seems to interest a customer, and then to organize the customer navigation (search results, navigation, ads shown...) on the site, then on the rest of the web.

Organizations, equipped with much higher means than those of ordinary people, become able to handle ever-increasing volumes of increasingly heterogeneous information to detect ever more subtle phenomena, to increasingly relevant decisions – and ultimately to strengthen their position or to occupy new. At no time do they try to talk to or listen to the visitor, but they prefer to guess its intention from its old moves. Big data could thus be the last means that companies have found to not to talk to their customers.

This approach has drawbacks, though. Not everybody is equal before the predictive algorithms, because not everybody is on the social networks, or even so, produces data with the same rate. And beyond these behavioral data on which are applied predictive algorithms, there is no room left for *serendipity* anymore, and everybody is gradually presented the same set of results or choices.

Let's change our point of view towards a customer-centric use. In the near future the challenge is not to help the organization to do things better – sell more and better – but to help people as individuals to do things better. This is what we can call *Small Data*. Big Data is dealing with statistics and trends, not specifics and immediate utility? Small Data is quite the opposite. Individuals, more equipped and connected, become active agents able to benefit from the same technology that companies, to analyze the past, to draw conclusions that make sense for them, to plan for the future, to be helped to make decisions, and implement these decisions towards their stakeholders. They share the right data about them, what they want to do right now, and want practical answers, not inferred from a motive or an intention emerging from a crowd pattern. This shift from company-centric to user-centric is the shift from *data-crunching* to data-sharing.

References and further readings

Google Flu Trends How-To <http://goo.gl/SUqJd>

Foundations for big data by Gartner analyst Doug Laney, 2001, PDF <http://goo.gl/C5iHn>

«Quantifying Trading Behavior in Financial Markets Using Google Trends», Scientific Reports in Nature, February 2013
<http://goo.gl/J8f2K>

David Feinleib slideshares, producer of The Big Data Landscape and bigdatalandscape.com website. <http://goo.gl/r4TLN>
He is the author of 8 Laws of Big Data that we disseminated in this *cahier* as a reference.

<http://whatsthebigdata.com/>

<http://humanfaceofbigdata.com/>

Gil Press, Forbes: «A Very Short History Of Big Data», May 2013 <http://goo.gl/jXn3i>
«Big Data News Roundup: Correlation vs. Causation», April 2013 <http://goo.gl/qmfDv>

Miller, H.E. [2013]. «Big-data in cloud computing: a taxonomy of risks» Information Research, 18[1] paper 571.
[Available at <http://goo.gl/oJUW0>]

«Emc And Big Data – A Fun Explanation», February 2013, 9' Video <http://goo.gl/09SSo>

«Big data, big dead end», January 2012: <http://goo.gl/YGvyy> ; «Big data, big illusion», April 2012: <http://goo.gl/x1kBc> ; «Big Data, Big Hype, Big Danger», April 2013: <http://goo.gl/hACvn>

«Le recrutement et la productivité à l'heure des Big Data», May 2013 <http://goo.gl/7RiHT>

Cléménçon et. al. Telecom ParisTech / Liligo [2012] «A Data-Mining Approach to Travel Price Forecasting» <http://goo.gl/xP44p>

The data landscape

Uses of data (1/3): Describe		
Types of data	Data elements	Patient A's blood pressure at 9 a.m. on Wednesday
	Aggregates	Histogram of current blood pressure readings for 45-54 year old females
	Clusters	Plot of average blood pressure readings by age group for males and females in the country

From Miller, H.E. (2013)

Miller, H.E. proposes first a taxonomy of types of data: *atomic data elements*, *aggregates of data* and *clusters of data*. Three uses of data are then listed for the healthcare sector: «describe» above, «analyze» and «act» on next pages.

What is data?

DATA IS A BRUTE FACT, which has not yet been interpreted. Data is not information, it is a value assigned to a thing. To create *information* – and then *knowledge* – out of data, we need to interpret that data. «19°C» is a data we can read on our thermostat. All the collected thermostat data of a building make a set of data. «Flats from this building seem to overheat» is an information, that can be derived from the comparison of this set of data with past data, or with data collected from other buildings in the area. Data is thus typically the result of measurements.

In today's digitally hyperconnected world **data is all around us**, in our phones, RFID's clothes, cars, food... Examples of data include a table of numbers representing blood pressure over a month, the characters on this page, the recording of sounds made by a person complaining about overheating at the phone. Almost everything we touch or use belongs to and feeds back to a larger data set.

Raw data is not useful as itself. It is unprocessed data that must be refined, interpreted to gain more value. In the data pipeline (see sidebar below), processed data is often the raw data of the next stage.

Data is never neutral. New combinations of data can create new knowledge and insights that were not thought at the beginning. For instance, monitoring the temperatures in a apartment building can give information about the habits of a resident, that can be correlated with, say, his tax return, and cast doubt on the actual composition of his family.

What is generating data?

75% of data are nowadays generated by individuals. This includes the digital footprints left by people living most of their lives online, the telemetry generated by their devices, and all sources of information about their behaviors. In 80% of cases, companies play a role in the life cycle of data: they store it, protect it, preserve it confidentiality or ensure good distribution.

Data law #5: Plan for exponential grow.

All these new datasets come mostly from:

- The web: website traffic logs, indexation, search queries, online transaction, friendships and social media relationships, document, images and video storage...
- Commercial data collected in the real world: transaction logs in a retail store
- Personal data: medical records...
- Public and open data...
- This human generated data is just the beginning. A lot more data is now generated by machines and the Internet of Things: sensors networks, RFID, GPS, phone traffic logs...
- Field data, which is data collected in an uncontrolled in situ environment
- Experimental data and scientific investigation by observation and recording: genomics, astronomy, meteorology, environmental science...



The data pipeline What do we do with data? Whatever the different types of data, almost all processing can be expressed as a set of incremental stages through the data pipeline above. With small projects, not each of these stages may be necessary. Ultimately, archived data can be reinserted in the pipeline for new insights.

“ *Leading organizations of the future will be distinguished by the quality of their predictive algorithms. This is the CIO challenge, and opportunity.* ”

Cleaning and contextualizing data

Data need to be cleaned and transformed first, to remove invalid records and to obtain a sane set of values. Cleaning and contextualizing is a first reinterpretation process and allows to have a new look on the original dataset.

Cleaning means to combine different datasets into a single table, remove duplicate entries and apply normalization processes. This can be the more time-intensive aspect of processing data, and still needs human intervention. Badly formatted numbers can indeed be corrected automatically, but without ultimate human control this could lead to big errors depending on the numbers nature. Inconsistencies in data and file corruptions are also corrected at this stage.

Data need context too, in order to be really useful. Context turns disparate data points into a story. Data without context tells a misleading story, as a former professor of the New York University reports in the NY Times. He wanted to determine with the help of sensors whether students used the elevators more than the stairs, and whether that changed throughout the day. The experiment went well and the data collected told a story: students seemed to use

the elevators in the morning and switch to the stairs at night. It was then interpreted as follows: perhaps students were tired from staying up late, and became enough energized during the day to use the stairs. But this appeared to be an entirely other story when the professor had a discussion with security guards who revealed that one of the elevators broke down a few evenings during the week, and the lazy students had no choice but to use the stairs.

Back to the Google flu trend. According to an article in the journal Nature, Google's algorithms were wrong this year: their results were double the actual estimates by the Centers for Disease Control and Prevention. Indeed Google Flu trends is only one source in addition to the flu surveillance methods, but it nevertheless raises questions. So what went wrong? «*Several researchers suggest that the problems may be due to widespread media coverage of this year's severe U.S. flu season*» the author's wrote in Nature. Add social media on top of this, and you understand how the news of the flu spread quicker than the virus itself.

«In other words, Google's algorithm [was] looking only at the numbers, not at the context of the search results.»

open data

Open data: when freeing data benefits to all people and organizations

True indeed, data is under the spotlights, whether big or small, acquired in real time or belonging to large archived sets. In this data storm, the open data phenomenon plays a particular role, because it opens the way to combine different public and private datasets together and thereby to develop more and better products and services.

According to the opendefinition.org website, open data is «*data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to **attribute and sharealike**.*» Saying «Open data» can emphasize three different meanings: the data that is open; doing the act of opening data; asking people to open their data.

Whereas big data focuses primarily on the benefits offered by exploiting ever-growing

massive set of data, the creation of value in the open data context depends more on the sharing and interoperability possibilities.

As anybody can now produce open data, with no assumptions on the «for what purpose?», it is worth knowing how these data need to be characterized:

- **Availability and Access:** *the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the Internet.* Open data is not raw data and must also be available in a convenient and modifiable form.
- **Reuse and Redistribution:** *the data must be provided under terms that permit reuse and redistribution including the in-*

termixing with other datasets. Open data are technically, legally and economically open.

- **Universal Participation:** *everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavor or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.*

When opening up data, it is important to focus on non-personal data, that is, data which does not contain information about specific individuals. Cleaning and anonymizing data can thus be a mandatory first step.

Processing data

Different types of data (structured and non structured)

The two major categories of data are qualitative and quantitative data.

- **Qualitative** data refers to the quality of something: colors, shape, texture of an object...
- **Quantitative** or numerical data refers to numbers: temperature, size, price, number of items...

Beyond qualitative and quantitative, data falls into more categories:

- **Discrete data** which is numerical data like a count. Shoes sizes are discrete data, even if they are not same amongst different countries.
- **Continuous data** which is numerical data within a continuous range. Feet sizes are natural and continuous data, with a minimum and a maximum.
- Data can describe an item **by categories** it belongs to: one's foot can be qualitatively describe as big, or belong to the «big» category; one shoe can be «brand new», «old fashioned» or «broken».

Data falls too in a continuum spectrum from structured to unstructured.

- **Structured data** refers to data that is identifiable because of a high degree of organization, such that inclusion in a traditional relational databases is seamless.
- **Unstructured data** is all sort of documents in all formats (word, PDF...), **XML** or **JSON** documents, emails, images, videos, slideshare presentations, and more recent social media posts and comments on social network walls, tweets, and all natural language logs such as customer service call and chat logs... Unstructured data is any data not in a relationship format, and not related to a predefined data model.

- There is always some sort of structure: unstructured data can indeed be structured at some level. This is the case with web logs, a commonly cited form of unstructured data. URI, which forms part of the log, or dates, are well structured data, but at a different level at which analysis is to be performed. One should better describes web logs as **mixed-type data**. A text document is inherently **semi- or poly-structured**. Some can look at it as a plain bag of words when others will query the words via stemming and synonyms, and others will study the document's grammatical structure.

V⁴ This is where comes a fourth V in the big data definition: **variability**, when the data format changes, or when one adds just a field in the initial data structure. Format can also change over time, for the user's purposes, or because similar data are added from another source that was different in structure and format.

Data law #3: Use more diverse data, not just more data.

Let's not confuse variability with variety. You meet variety when you go to the newsstand and have access to dozens of different newspapers. Let's choose one and come over the week to choose the same title: every day it is fresh news and stories, and new styles to tell the stories. This is variability.

Big data solutions allow data to be stored in its original form, either structured, unstructured or semi/poly-structured, in its entire variety and variability, and be available for analysis when a user queries the data.

Unstructured data are often estimated to account for 70-85% of the data in existence. «*Big data is about looking ahead, beyond what everybody else sees,*» said Peter Sondergaard, senior vice president at Gartner and global head of Research. «*You need to understand how to deal with hybrid data, meaning the combination of structured and unstructured data, and how you shine a light on 'dark data.'* Dark data is the data being collected, but going unused despite its value. Leading organizations of the future will be distinguished by the quality of their predictive algorithms.»

Uses of data (2/3): Analyze		
Types of data	Data elements	Time series of Patient A's blood pressure
	Aggregates	Correlation between 45-54 year old female blood pressure readings and daily calories consumed
	Clusters	Regression analysis of group health status vs. caloric consumption with sex and age as 'dummy' variables

Uses of data (3/3): Act		
Types of data	Data elements	Prescribe medication and dosage level to treat hypertension
	Aggregates	Change budget for proper eating habits information campaign
	Clusters	Run simulation model to predict change in group health status using various budgets as health intervention scenarios

From Miller, H.E. (2013)

Managing the petascale computing

Data law #2: maintain one copy of your data, not dozens.

“Big data is not new, but the tools are.”

limits

Specialized software and hardware are needed for the data challenge, either big, fast, open, secured or real-time: developments in parallel and distributed processing are necessary for working in a reasonable amount of time. Let's begin with two key enablers, not completely related to each others: the NoSQL database revolution, and the Hadoop computing stack and ecosystem.

The NoSQL revolution

NoSQL (mostly interpreted as « Not Only SQL ») offerings are closely associated with Web application providers, becoming key foundations of any web-scale computing stack whether for online Create, Read, Update, Delete (CRUD) applications or for off-line analytics. It is a broad class of database management systems identified by its non-adherence to the widely used relational database management system model, ACID [see below the limits of traditional database management systems]. NoSQL databases are not primarily built on tables, and as a result, generally do not use the widely used SQL language for data manipulation. Instead, they are accessed via plain get and put commands. Much of the structured data are eliminated in the queries, which are often reduced to straight key-value pairs. Fixed table schema are not mandatory, and join operations are avoided.

NoSQL products like Cassandra, HBase (also a part of the Hadoop ecosystem, see below), Riak, CouchDB and MongoDB (to mention a few)

eliminate relationships between data entities for the benefit of allowing for greater horizontal scalability, partitioning across several machines and replication. NoSQL also gives the developer a more flexible «on-read» schema model that has its benefits.

There are currently 4 types of NoSQL databases: Key-value [e.g. Memcached] ; Column oriented or clones BigTable [e.g. Cassandra] ; Document oriented [e.g. CouchDB, MongoDB] ; Graph [e.g. Neo4j]. See [<http://goo.gl/Z3Rh>] for a complete list of NoSQL solutions.

The Hadoop stack and ecosystem

The technology solution that is most associated with big data is Hadoop, an open source database, and is now 10 years old. It can be seen today as a catalyst for Big Data. Hadoop and others have democratized the data management, off-line batch computing and analytics, to make them accessible — both practicality and cost — to small companies and organizations.

One of the first big data challenge was to harvest, index and rank the billions of web page out there, and make better search engines. In 2004 Google issued a famous paper on the «MapReduce algorithm» [a computational approach that involves breaking large volumes of data down into smaller batches, and processing them separately] and their Google File System. This inspired people at Yahoo! who wrote an open source version in Java, as an Apache

The limits of the traditional database management systems

Relational database management systems (RDBMS) store data in relational tables (tables related to each other via one common element), structured in rows and columns, and use **Structured Query Language** (SQL) for accessing and manipulating data inside. Unstructured data cannot fit well in such tables, as their **schema** is not known in advance, and is typically stored in key-value pairs not accessible via SQL.

SQL and relational database have failed over the past decade to keep up with the scaling demands and requirements with regards to performance or agility, coming from large social networks datasets, and this paved the way to new different approaches and various database alternatives. In this earlier days, analytics were

performed on frozen data, after a so-called **ETL** (Extract, Transform and Load) approach. Data was extracted from **OLTP** (On-Line Transactional Processing) systems where they obeyed the **ACID** (Atomicity, Consistency, Isolation and Durability) rule, and loaded on data warehouse systems able to handle large volume of data. This is good for analyzing past performance, but fails to deal with real-time insights demands. Additionally, new data is coming at increasing speeds, 80% of it being unstructured and fail-

ID	Lastname	Firstname	Blood pressure
1	Smith	Mary	90
2	Jones	Lena	119
3	Martinez	Sara	104
4	Smith	Clara	123

Data organized in a row-oriented database

project: the Hadoop framework. It supports the running of applications on large clusters of **commodity hardware** (commodity servers that in turn use commodity disks), divided into many small fragments of work (the MapReduce idea), each of which may be executed or re-executed on any node in the cluster, meaning that clusters can grow as data volumes do. It is associated with its own distributed file system (HDFS) that stores data on the compute nodes, providing high bandwidth across the cluster. Nowadays, Hadoop refers to a larger ecosystem of software packages, including MapReduce, HDFS, and many software packages to support the import and export of data into and from HDFS, and other distributed file system like Amazon S3, or newer, more efficient DFS.

Hadoop can be used for any sort of work that is batch-oriented rather than real-time, that is very data-intensive (and data that does not have to be structured), and is able to work on pieces of data in parallel. However, Hadoop goes now far beyond than the mere MapReduce jobs. It can be used for other applications, many of which are under development at Apache: the HBase database, the Apache Mahout machine learning system, the Apache Hive Data Warehouse system (essentially providing a SQL abstraction over MapReduce), or Yarn, the nextgen Hadoop framework for job scheduling and cluster resource management, for naming a few.

ing to fit in the traditional rows and columns. Row-based systems are designed to efficiently return data (for instance: Smith Mary 90) for an entire row, or record, in as few operations as possible. In a column-oriented DBMS, data is stored as sections of columns, making efficient to «find all the people with the last name Smith» in one operation.

OLAP (Online analytical processing) is an approach to answer multi-dimensional analytical. Databases stored in so-called OLAP Cubes use a multidimensional data model, and provide fast access to knowledge through techniques that include pre-aggregated, pre-built analytics in the cube. Big data needs ad-hoc, data exploration and knowledge self-discovery, which is not possible in the OLAP cube based on requirements and assumptions.

Massively Parallel Processing

Big data is not all about MapReduce. Massively Parallel Processing (MPP) is another approach to distribute query processing, more corporate oriented than academic- or research-oriented MapReduce. In both approaches processing of data is distributed across a cluster of compute nodes, each separate nodes processing data in parallel, and the final result being assembled at the node-level output. However, MapReduce and MPP are used in rather different scenarios, and use different hardware. MPP is used on expensive, specialized hardware tuned for CPU, storage and network performance. MPP products are thus bound by the cost of these assets and the software, and by its finite hardware.

MPP columnar analytic databases have become a common choice for any real-time analytics on structured data. Whereas in traditional relational databases the data is organized in rows, MPP use a column-oriented organization, in a compressed form, yielding faster queries. Moreover, the nature of MPP facilitates the «scale out» by simply adding more commodity hardware.

Grid / Cache

The NoSQL solution still relies on data stored on disk, making it not practicable for real-time due to all the reads and writes. Additionally, the delay in replication could result in datasets out of date. Transactions that need speed and reliability are not good candidates for NoSQL.

Big data grids now can store data across many in-memory nodes, rather than data stores on disk. All the reads and writes from disk are eliminated, data can be queried up to 10x more rapidly, and the overall performance is more consistent. These databases are able to incor-

porate real time data with learned behavior, and react in real time.

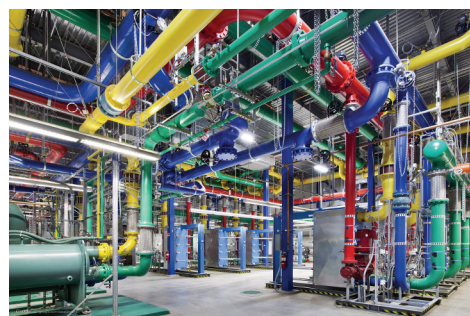
This was made possible thanks to the decrease of memory costs in the past few years, and its easy commoditization in the cloud. Real-time in-memory databases simplify the internal optimization algorithms and makes better use of the hardware.

NewSQL

NewSQL addresses the problem of datasets mixed with structured and unstructured data. The term was first used by 451 Group analyst Matthew Aslett in a 2011 research paper. NewSQL is a set of various *new* scalable/high-performance SQL database vendors (or databases), it is not a new query language.

NewSQL databases provide the same scalable performance of NoSQL systems for OLTP workloads while still maintaining the ACID guarantees of a traditional single-node database system. They are built on a scale-out, shared-nothing architecture, capable of running on a large number of nodes without suffering bottlenecks.

Three different approaches are adopted by vendors. New databases are proposed, designed from scratch to achieve scalability and performance. Some changes to the code may be required and data migration is still needed from older databases. Performance is allowed via non-disk (memory) or new kinds of disks (flash/SSD) data store. Some solutions can be software-only (VoltDB, Nuodb and Drizzle). Secondly, vendors offer new MySQL storage engines: MySQL is used extensively in OLTP and in web services. The third approach is to ensure scalability of the OLTP databases by providing a pluggable feature to cluster transparently.



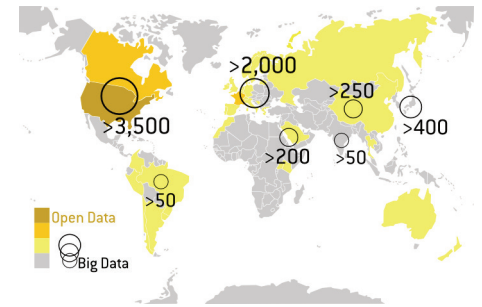
Huge storage databases need huge cooling. Here is a glimpse at Google's data centres.

The data market

Law #8: Big Data is transforming business the same way IT did.

The data landscape

This figure compiles several landscapes of the data market (see refs at the bottom right). «6» indicates the page number where details are provided.



Big data is creating a new layer in the economy at full speed. All is about turning data into information, knowledge and, ultimately, into revenue. This will accelerate growth in the global economy. Estimation from analysts in 2013 suggest that the global Big data market will grow to a staggering \$16.9 billion by 2015. Big data is not a business model, it is business. It is an enabler for new ways of conducting business.

New business need data-driven business models

Hardware, software and networks having been commoditized to the point that they are essentially free, Data is the only business (model) that is left: either monetize data, or provide the infrastructure to enable the monetization of data.

Big data enables companies to create new products and services, enhance existing ones, and invent or refine business models. Companies which don't take the turn of (big) data, or which do not take the time to analyze the dark data they have in hands, or how collecting data might change their business, will soon be wiped out

Traditional enterprise data is only growing at 20% year over year when the amount of new data being stored is growing in the order of 50%. There will thus be two key shifts within the storage industry:

- a move towards more commodity-based storage that can potentially take the place of traditional storage
- a new set of high-scale storage architectures to store all this new data.

These application development platforms are often platforms. They provide additional analytical capabilities natively provide. Example includes Infochimps, Acunu, Wibidata, Causata, LucidWorks, Cityzen Data...

Hortonworks, MapR, Vertica, Cloudera, Greenplum, EMC2, IBM, Kognitio, Datas-tax, Exasol, Actian ParAccel...

ANALYTICS INFRASTRUCTURE

Oracle, SAP, IBM, Microsoft, bime, chart.io, GoodData, Talend, Jaspersoft...

BUSINESS INTELLIGENCE

Tableau, metaLayer, Teradata, Datameer, visual.ly, panopticon, Treerank, Dataveyes, ¹⁸ kwypesoft, factory, Squid...

ANALYTICS AND

Talend, Informatica, Pentaho, Flume, Oqoop, Squid solutions...

DATA INTEGRATION

Splunk>, Loggly, Sumo Logic...

LOG DATA APPS

Some companies or solutions can be found in several classes, and this can change over time. French companies are underlined.

Factual, GNIP, DataSift, Inrix, LexisNexis, Kaggle, Windows Azure Marketplace, Space Curve, Loqate...

DATA AS A SERVICE

Data sources include: environmental sensors, social & commercial web, service providers, infrastructure providers, telecom networks, open data ¹⁰, quantified self data ⁶, geocoding services, datasets and datasets directories...

SOURCES

The large sensors clouds are still in their infancy. Open source initiatives like OpenIoT (Open Source blueprint for large scale self-organizing cloud environments for IoT applications) is aimed at developing an open source middleware platform to connect Internet-objects to the cloud. <http://openiot.eu/>

SENSOR AS A SERVICE

Cloudera, Intel, Hortonworks, EMC2, IBM, MapR, Hadapt, Ubeeko...

HADOOP

CouchDB, HBase, InfoGrid, Infinite Oracle Coherence,

NO SQL

AmazonRDS, FathomDB, Sa-Drizzle Handler-

NEW SQL

Amazon web services, Windows Azure, Infochimps, Google BigQuery...

INFRASTRUCTURE AS A SERVICE

Dell, Cisco, EMC2, NetApp, SGI, Fusion, IO, Withings...

HARDWARE PROVIDER

The amount of new data stored varies across geography: new data stored in Petabytes by geography in 2010 (source: Wipro); Open data initiatives by regions as of 2012 (source: Open Knowledge Foundation)

built on top of Hadoop and/or scale-out ties beyond what the underlying database Continuity,

HORIZONTAL PLATFORMS

Palantir, dataspota, Metamarkets, ClearStary, platfora, alteryx, Cinequant, Visibrain, Focusmatic, The Metrics solutions...

VISUALIZATION

Forbes, IBM, Deloitte, ThinkBig, Cetadata...

SERVICES

Rocketfuel MediaScience, Bluefin, Data collective, Recorded Future...

AD MEDIA APPS

VERTICAL PLATFORMS

Specialized applications for a specific industry vertical: Predictive Policing, Bloomreach, Myrrix, Kyruus, Splunk, Palantir, Explorys, Sumo Logic, Safetyline, Citizen Sciences...

A Do Tank for the French Big Data community

In France, Aproged and Cap Digital founded in March 2013 the Alliance Big Data, and have been joined by ADBS, APEC, GFII, Institut Mines-Télécom, and others.

<http://www.alliancebigdata.com/>

Interindividual Greenbureau...

PRIVACY

Gazzang virtru Dataguisse

SECURITY

MongoDB, Neo4J, Graph, Cassandra, db4o, ObjectStore, GemStone, Polar...

Teradat Vertica Netezza Oracle...

EDW / SQL

SQLAzure, scaleBase Continuent Socket, VoltDB...

GridGain, Teracotta, Infispan, memcached...

GRID / CACHE

Amazon web services, Eucalyptus, AppScale, GoGrid, Zinix, Intercloud...

CLOUD PROVIDER

by more agile company reinventing their business through a creative destruction of today's business models. At the beginning of this century, data offers a new Industrial Revolution, that was announced by the Internet and the mobiles. Smart companies first think data as an asset, upon which they experiment and build business models.

Data registers, API and mashups at the heart of new businesses

Two keys of understanding are needed to explain how this data ecosystem develops itself so quickly. All is happening in the interfaces, what is called **mashups and APIs**. As data grew to big data, companies used to store it in the cloud, and others developed tools in the cloud to process it. But data gains tremendous value when correlated with other data from various sources you don't own. Using the **directories of public or paid datasets** is the second key. In a matter of hours, clever developers during Startup Weekends can create new services by tapping a continual stream of information from internal and external sources, and by querying the APIs of big web services like twitter, Amazon or Ebay, or smaller one APIs discovered on directories

like <http://www.programmableweb.com/apis>. All these APIs are well documented, and some of the datasets they are connected to can even be queried in natural language. A billion records dataset such as Versium enables specific individual queries that cross traditional marketing boundaries: social-graphic, demographic, and psycho-graphic, in both the online and off-line worlds.

France's assets

France's entrepreneurs and scientists can play an essential role in the new Economy of data. First of all, the **French School of Mathematics** is renowned all around the world. In France 27% of college students earn a degree in math, science, technology or engineering, compared with only 17% in the U.S. Of the 52 winners of the Fields Medal, 11 have been French. Second point is its **Telecom industry**, used to provide detailed and quick reporting across tera or petabytes of data. Ultimately, the French big data startups and champions, the clusters dedicated to the digital creative industries, and the higher education institutes and research centers all are **working together** with the French government on major data projects.

References

The [big] data market is currently a very noisy market. This figure joins several landscapes made by the analysts and the industry:

the big data landscape by Dave Feinleb [2012 <http://goo.gl/MJu3M>]
the big data open source tools [jan'2013 <http://goo.gl/WUITe>]
the Hadoop ecosystem by Datameer [jan'2013 <http://goo.gl/se8wH>]
the database map by the 451 Group [feb'2013 <http://goo.gl/QgRI5>]
the big data ecosystem by Sqrrl [mar'2013 <http://goo.gl/y7Nc8>]

All these landscapes demonstrate the lack of agreement on what are the different segments of the [big] data market and what to call them (see «the big data landscape revisited» [april 2013 <http://goo.gl/hEsVG>]).

For an evolution of these references over the years, read «Getting a grip on the rapidly changing DBMS landscape» [<http://goo.gl/cgHwz>].

Unlocking data value

Scikit-learn integrates machine learning algorithms in the tightly-knit scientific Python world, building upon numpy, scipy, and matplotlib. As a machine-learning module, it provides versatile tools for data mining and analysis in any field of science and engineering. It strives to be simple and efficient, accessible to everybody, and reusable in various contexts. See also «contributions» page 27.

[<http://scikit-learn.org/>]
For a presentation of Scikit-learn, read
[PDF: <http://goo.gl/kiZ2i>]

Machine learning

The question is then: what is the meaning of all this data? Basic analytical methods used in Business Intelligence and Enterprise Reporting Tools delivered simple sums, counts, averages and results from SQL queries. These analytics were specified by humans who knew what should be calculated and how to do it. This is not possible anymore with datasets too large for comprehensive analysis. It is not possible for an analyst to test all hypotheses and unlock the value buried in the multiple data sources. New algorithms are needed to deal with big data: existing statistical algorithms do not scale, and using sampling for prediction may miss important facts or phenomena.

Machine learning systems automate decision making on subsequent data points, automatically producing outputs like classification, recommendations (as with Amazon's products) or groupings.

Technologies includes WEKA, Mahout, MOA, scikit-learn, SkyTree...

Value is the fifth V of the Big Data

IT departments have had to make decisions about which data to keep and how long to keep it, at a time where the processing power required to perform analysis was far beyond their capacities. Machine learning helps now to unlock the **Value** in the datasets, as soon as the data is produced.

ly handled by tensor-based computation, such as multi-linear subspace learning.

Additional technologies being applied to big data include massively parallel-processing databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud based infrastructure (applications, storage and computing resources) and the Internet.

Source: Wikipedia

Machine learning classes

Machine learning algorithms generally fall into two classes depending on whether the training data set includes the correct or desired answer or not. When the desired answer is known and drives the algorithm, this is called **supervised learning**. A training dataset is provided to the algorithm, with all the data attributes for each training instance, but also the correct class for the instance. At its opposite, **unsupervised learning** makes possible to find insights on data with no clues. It is provided with unlabeled data inputs and has to discover by itself any associations or relationships between the data instances. This algorithm is an autodidact. It proceeds by seeking and grouping similar data. It is very interesting to use this technique when we do not know what we are looking for. For example, speech recognition, especially the isolation of the voice with respect to a noisy environment, uses an unsupervised neural network algorithm.

Reinforcement learning is a class of algorithms which is somewhere between supervised and unsupervised learning. From supervised learning it borrows the knowledge of a desired outcome. A sequence of steps is needed to arrive to this known outcome, but it is not known if every step goes effectively towards the goal or not. The right answer is never given, and like unsupervised learning, reinforcement learning systems are trained with unlabeled data. However, some distance measure to the goal is done, and the internal mechanics of the algorithm are rewarded or punished, according to their positive or poor progress towards the desirable outcome.

Machine learning applications

According to SkyTree, a leader in advanced analytics services, Machine Learning can be applied wherever data is available to gain new insight and improve decision making:

- **E-Tailing:** product recommendation engines, cross channel analytics, events/activity behavior segmentation
- **Retail/Consumer:** merchandising and market basket analysis, campaign management and optimization, supply-chain

New algorithms are needed to deal with data

A 2011 McKinsey report suggests suitable technologies for Machine Learning to include A/B testing, association rule learning, classification, cluster analysis, crowd-sourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization.

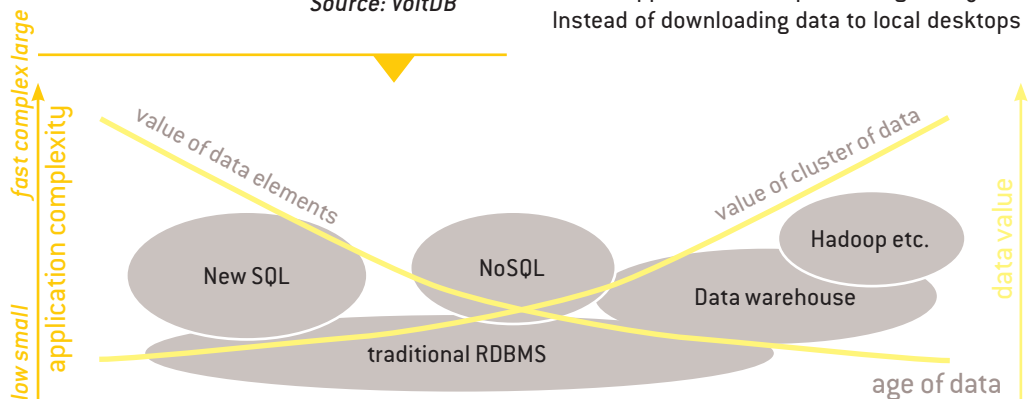
Multidimensional big data can also be represented as tensors, which can be more efficient-

“*Nowadays, organizations want to perform «deep analytics» on the massive datasets. This ranges from statistics (averages, correlations, regressions) to more complex functions such as graph analysis and predictive analytics by using advanced machine learning.*”

Value of data over time

The value of a data elements goes down with time and the value of a cluster of data goes up with time.

Source: VoltDB



Transactional

Interactive	Real-time	Record Lookup	Historical	Exploratory
milliseconds	10 milliseconds	second(s)	minutes	hours
<ul style="list-style-type: none"> place trade serve ad examine packets approve trans. 	<ul style="list-style-type: none"> calculate risks leaderboard aggregate count 	<ul style="list-style-type: none"> retrieve click streams show orders 	<ul style="list-style-type: none"> backtest algos business intelligence daily reports 	<ul style="list-style-type: none"> algo discovery log analysis fraud pattern match

Realtime analytics

Once the data is acquired and cleaned, it is time to extract insights from it. The three biggest analytics areas are customer interaction (for further recommendation and personalization), network and sensor monitoring, and game and mobile application back-ends. Add it algorithmic trading, anti-fraud, risk measurement, law enforcement/national security, healthcare and stakeholder-facing analytics.

Advanced analytics can also enable organizations to penetrate new markets, grow revenues, track competition, forecast demand, drive product and service differentiation, analyze detailed transactions to better understand customer patterns, provide advertisers with more granular targeted advertising, acquire and retain new customers, better predict customer churn and profitability, enhance visitor experiences, and respond to market dynamics and regulations.

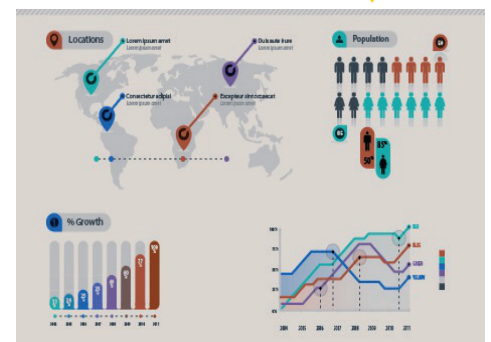
Real-time analytics enables users to get up-to-the-minute fast data by tapping directly what's happening on their ecosystem, as it happens. Many companies have thus rethought traditional approaches to performing analytics. Instead of downloading data to local desktops

or servers, they are running complex analytics in the database management system itself. In France, Squid Solutions offers analytics as a service on fast data, providing a low-latency high-availability worldwide server network to track their customer interactions with all of their web properties, affiliates, display advertising, etc...

New software are needed to process real-time information from highly dynamic sources, and web companies such as twitter have developed and released in open source these new technologies. Storm from Nathan Marz (now with Twitter) operates on streaming data that is expected to be continuous. These complex event-driven systems allow to identify meaningful events from a flood of events: one of the most interesting example provided by Twitter is the generation of trend information.

Data analysis for everyone

A lot of companies still don't have the technological sophistication to understand the whole data science and cannot hire data scientists to dig in their data. The hardest work is the so-called *data munging* and it gets harder with scale. Advanced analytics on unstructured data online services like Precog (see example of web analytics below), offer tools targeting specific use cases. The sentiment analysis and natural language processing are highlighted in the social media one, while the web analytics focuses on features such as behavioral clustering. And the visualization of the results have been particularly tuned.



Data visualization: telling stories with data



Same data, different visualization. In which company would you like to invest?

Data visualization tells a story on data, and this story could be misleading the reader. The figure on the left shows the same data with two different presentations. Say it is the percentage of production of 3 goods, A, B and C, over a period of three years. A is 20% of the whole production the first year, and B and C 30 and 50% respectively. Depending on how you classify the goods and the colors you choose, you will not give the same impression on the company situation.

Data visualization helps to find new meanings to data that were misinterpreted before, and to share that meaning with others. Through her work as a nurse in the Crimean War, Florence Nightingale was a pioneer in applied statistics and data visualization, back in 1855. She gathered data on relating death tolls in hospitals to cleanliness and communicated her results through a new sort of pie chart, called afterwards roses, polar area charts or coxcombs. This particular visualization emphasized the real causes of death among soldiers and showed that more soldiers were dying from preventable illnesses than from their wounds.

As these data could not have been explainable with a simple pie chart, it is worth nowadays to be aware of all the different visual representations we have in the toolbox. More than one hundred of visualization methods exist, and Ralph Lengler & Martin J. Eppler listed them as a periodic table in 2007, available as an interactive visualization online. This applies to data, information, concept, metaphor, strategy and **compound visualization**.

Data visualization can tell stories to better understand complex data and / or large amount of multidimensional data. In a famous video lecture published on YouTube, Hans Rosling, the Swedish statistician and medical doctor, runs through 200 years' worth of augmented-reality data visualization telling the story of economic development and health in 200 countries over 200 years using 120,000 numbers in a mere four minutes. Plotting life expectancy against income for every country since 1810, Hans shows how the world we live in is radically different from the world most of us imagine.

[<http://goo.gl/Aat8Y>]

Visualization tools are now available for the masses. Free data visualization tool can help create an interactive viz in minutes and embed it in on website or share it. Lots of visualization javascript libraries and frameworks are available to allow scientists, journalists and companies to make their data more visual and thus, sharable. The major social networks provides tools to make their user view, and understand the relationships between you and your connection, and how to grow and use it better. These tools can make sense out of the social network noise.

Beware a new deluge of data visualization! It is above all about giving attention to data that is too valuable to be remained on the shelf, and communicating an idea that will drive action. Three requirements and three legitimate reasons must be satisfied to make data visualization valuable and worth the effort:

- **Information must be interpretable.** With so much unstructured data used today, interpretation must come from the meta-data associated: what is the data, where, when and how it was collected...
- **Information must be relevant** to the people looking for insights, and to the targeted purposes.
- **Information must be original** or, as Nightingale proved in 1855, shed new light on a phenomenon.
- **The dashboard reason:** visualizations can help to check assumptions about how a system we are interested in operates, to see how it can deviate from a predefined model, and to make relevant decisions.
- **The gameification reason:** playing with data, develop intuition and new insights on the behavior of a known system, replay long time series data in a shorter experimental time frame.
- **The exploration reason:** when data is too complex to be understandable with mere statistics, visualization tools can help to build a model that make possible to ask the good questions to the system.



Rise of the data jobs

Data law #7: Put data and humans together to get the most insight.

The data science Venn diagram by Drew Conway [<http://goo.gl/gKkJP>]

talents

In the search for data talented people ?

Here are more than 75 – and still growing – job interview questions maintained by Dr. Vincent Granville <http://goo.gl/Y48ZU>

See also «7 new types of jobs created by Big Data», September 2012, <http://goo.gl/oE9KM>

New skills are required to deal with data, and each of these skills has a unique function in big data analytics. Moreover, the data specialists are operating at different levels:

- **data architects** put together disparate types of data in new ways to create fresh insights
- **data engineers / operators** develop the architecture that helps analyze and supply data
- **data visualizers** translates analytics into comprehensive and sharable information

Then, there are the **data change agents** who have an evangelist informal role, **data stewards** that ensure that data sources are properly accounted for and maintained, and **data virtualization / cloud specialists** who build and support the Database as a Service functions.

«But there is a challenge» said Peter Sondergaard, senior vice president at Gartner and global head of Research. «There is not enough talent in the industry. Our public and private education systems are failing us. Therefore, only one-third of the IT jobs will be filled. Data experts will be a scarce, valuable commodity. IT leaders will need immediate focus on how their organization develops and attracts the skills required. These jobs will be needed to grow your business. These jobs are the future of the new information economy.»

McKinsey projects that by 2018, the U. S. will need 140,000 to 190,000 people with expertise in statistical methods and data analysis, the «deep analytical talent», and 1.5 million more data-literate managers, people capable of analyzing data in ways that enable business decisions.

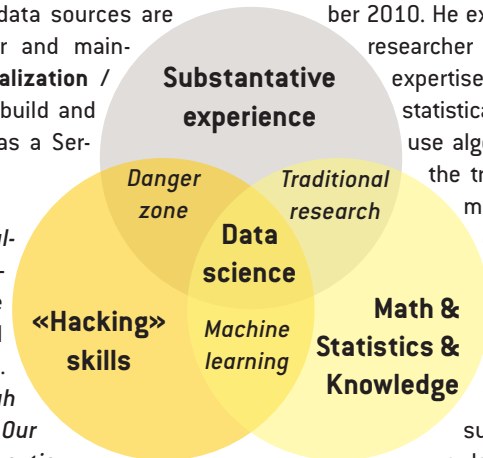
«By 2015, 4.4 million IT jobs globally will be created to support Big Data, generating 1.9 million

IT jobs in the United States,» adds Peter Sondergaard. «In addition, every big data-related role in the U.S. will create employment for three people outside of IT, so over the next four years a total of 6 million jobs in the U.S. will be generated by the information economy.»

However, companies attempting to handle the data challenges with silo-ed statisticians, computer scientists or MBAs will certainly fail. What is needed are professionals with a convergence of skills, somewhat called «**data scientists**», with a strong background in artificial intelligence, natural language processing and data management.

This convergence of skills view of the data scientist is attributed to Drew Conway in September 2010. He explains that the traditional researcher may have substantive expertise and learning, as well as statistical skills and the ability to use algorithms, but he is lacking the training and experience to manipulate raw data in a clever and skillful way. To him, data plus math and statistics only gets you machine learning. Such a specialist knows hacking and math/stat, but is substance-free. In other words, most of the analytics in machine learning are theoretical and model-free. Finally, Conway places a danger zone at the convergence of hacking skills and substantive experience. These are people who «know enough to be dangerous». They may have discovered an interesting fact, but as they do not master statistics, they cannot distinguish between this random event and a systematic pattern. «Either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created» says Conway.

Thus the Data Scientist is a rare bird indeed and the three complementary skills – mathematics and statistics, substantive experience, hacking skills – combined in one person are highly valuable.



Privacy and trust concerns

“What is legal?

What is ethical?

What will the public find acceptable?

In an ideal world, these three considerations would lead to the same result...”

Bill Franks, Chief Analytics Officer for Teradata's global alliance program

Google statement on privacy is highlighted on the Google Flu Trend page [op. cit.] as follows: *«Google Flu Trends can never be used to identify individual users because we rely on anonymized, aggregated counts of how often certain search queries occur each week. We rely on millions of search queries issued to Google over time, and the patterns we observe in the data are only meaningful across large populations of Google search users.»*

Only a few years ago, we were cautioned not to put online our name or birth date, but societal norms are shifting, and we now check online when traveling, telling everyone we are away. But even if we stay cautious, aggregation of data from innocuous datasets can be easy, and relatively low cost computing can be done by governments, criminals or our neighbors to predict our buying behavior.

The data-gathering technology raises questions about the limits of people surveillance. *«You don't know what data is being collected and how it is used»* says Marc Rotenberg, executive director of the Electronic Privacy Information Center about data collected in workplaces (see page 5). The possibility of a startup scavenging the social web to build complete profiles of people – with email, name, location, interests and such – raises the question of a better international regulation, balancing the rights to privacy vs security vs commodity.

And this is urgent matters, when one considers the impact of the drones, a question debated for two years in the U.S., and yet largely unknown by citizens in Europe. Surveillance drones are collecting amounts of data, which can then be combined with facial recognition technology and big datasets. Airspace will soon be open for private drones, and technology makes them stealth and small, recalls the Electronic Frontier Foundation.

One last question, and another V in our bag. Now that we have all this data, here is the pivotal question: can it be trusted? Data collected from a drone tells I have a terrorist behavior: can it be trusted? Data collected from flu trends launches a large production of vaccines: what if this causes a financial mess, or panic, or other unexpected phenomenon?

V? This is the essence of **Veracity**. That is «conformity with truth or fact» or in short, Accuracy or Certainty. And this can be caused by any of inconsistencies, model approximations, ambiguities, deception, fraud, duplication, spam and latency.

Ensuring that data is full of veracity at any time of its life cycle might be one of the biggest challenges we will face.

References and further readings

«When Google got flu wrong», Article in Nature, February 2013, <http://goo.gl/Cgbt9>

V. N. Vapnik, «The Nature of Statistical Learning Theory», Springer, 2000

«The history of Hadoop: From 4 nodes to the future of data», 2013, <http://goo.gl/mh30f>

On open data, see a brief history in the Paris-Tech Review, 2013, <http://goo.gl/EqzPK>

«Pioneering the dataviz: Nightingale's 'Coxcombs'», in the *Understanding Uncertainty* blog, 2008, <http://goo.gl/OAhBo>

A thoughtfully curated selection of tools that will make your life easier creating meaningful data visualizations.

<http://selection.datavisualization.ch/>

«Towards A Periodic Table of Visualization Methods for Management», PDF & interactive: <http://goo.gl/YMCpJ> & <http://goo.gl/Abk8>


Data Visualization Tools by the visual.ly website: <http://goo.gl/IBEHJ>

«Process real-time big data with Twitter Storm», April 2013, IBM, <http://goo.gl/uSNKQ>

«Big Data, Big Value», Les Entretiens de Télécom ParisTech, experts' opinions and 10 french startup interviews, Decembre 2012, <http://goo.gl/hLmGb>

Challenges in the air

Scientific challenges



Data law #6: Solve a real pain point.

WE NOW PROPOSE A SERIES OF 17 scientific, technical and/or societal challenges which address major issues in data management. We arranged them so as to show their connections. Other challenges do exist, but this list provides a preliminary current view, ranging from solving the big problems of the world to everyday data management for everyone of us. Some challenges might be in several sections. We first classify

in the 'scientific section' the challenges which may need further research, even if in some cases applications are already on the market. Then come technical challenges which need a appropriate answer, and societal challenges which imply us more deeply. For most of them we propose an opening statement, and then clear objectives or still open questions, and possibly issues and blocking points.

#1

Towards a better understanding of the world?

Data exploration is the last scientific exploration paradigm.

Thousand years ago science was only **empirical**, describing natural phenomena. Then came the **theoretical** branch a few hundred years ago, using models and generalizations, with Kepler's Laws, Newton's Laws of Motion and so on. The **computational** branch appeared during the last century, simulating complex phenomena, when the theoretical models were too complicated to solve analytically and the computers were available to do the job. Nowadays science has changed, and here is how: scientists do not look anymore directly at their instruments. They are looking at data captured by instruments or generated by simulations, before being processed and analyzed.

«*This data exploration is the fourth paradigm for scientific exploration*», said Jim Gray, the eminent database researcher, in January 2007, and calling **eScience** this transformed scientific method where «*IT meets scientists*.» This data-intensive science consists of three basic ac-

tivities: **capture**, **curation**, and **analysis**. Much of the vast volume of scientific data captured by instruments on a daily basis, along with information generated via computational models, reside then forever in a live and curated state for the purposes of continued analysis.

Some say it is sufficient to understand the world. Chris Anderson, editor of Wired magazine, hypothesizes the end of hypotheses: «*We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot*.» Rely on data is already reshaping the explanation schemes: categories, that once were a key for describing the world, are no longer used (see sidebar below).

Question: a – somehow philosophical – question still remains. To what extent are we going to

The quest for raw data through the open and big data movements: will categories disappear?

Journalists, sociologists and statisticians use to make visible the social world from surrounding data with different techniques and methodologies. The former study the data to reveal hidden realities, to show the public things which it does not have access. For their part, sociologists make visible the effects of structures.

Statisticians, and especially those from government agencies, reveal correlations between

statistical categories. This is a «zoom out» function to understand the world. These categories propose a system of conventions for describing the social world.

However, two current crisis are changing the situation: a general crisis of representation of the world, especially through the media and politics, and a crisis with regard to categorical representations related to statistical conventions that describe the social world.

Indeed, said Dominique Cardon, sociologist at Orange Labs, the categories that allowed us to

describe and structure the society are falling. They do not seem to explain the world anymore.

Currently, the world of big data and open data suggest to predict user behavior through learning algorithms on raw data so that everyone can project its own interpretation, its own agenda, its own objectives. And these two phenomena will now produce new interpretations of our societies.

«*From statistics to big data: what changes in our understanding of the world?*» Article in French, December 2012: <http://goo.gl/97rWa>

change our understanding of the world, in a way we could make more errors? Over-reliance on historical data for instance can create mischaracterization of data and solving the wrong problems. It also risks repeating the error of treating correlation as causation.

Further reading: «The Fourth Paradigm: Data-Intensive Scientific Discovery», October 2009, Microsoft Research, <http://goo.gl/qqr8B>

#2 Pour a maximum of data and solve big problems

Is this the new way for solving the world?

Objective: find quickly innovative solutions to solve big problems

Methodology: take a problem, let's say the environmental question, check all the useful disciplines to see what data are necessary to tackle the problem. Break large problems up into smaller problems, and collect these data.

There are numerous good reasons to choose the ecological question for such a challenge. One of them is the overgrowing source and heterogeneity of data, from low-cost sensors networks to satellite observations, produced by organizations, government agencies and people. The Internet connectivity enables then data sharing across organizations and disciplines. This «ecological science driven by data» presents new computing infrastructure needs and opportunities, which are the next challenge.

Further reading: «Redefining Ecological Science Using Data» in «The Fourth Paradigm: Data-Intensive Scientific Discovery», op.cit.

Is datascience a misnomer?

Fundamentally, Science, in the hypothetico-deductive model, proceeds by formalizing a hypothesis given a set of observations and assumptions, designing an experiment around that hypothesis, testing it and analyzing the data generated through that process to either corroborate or falsify the hypothesis. Science also refers to a body of knowledge itself.

#3 Build and share data-oriented infrastructures

Build the Vannevar Bush's memex.

We need not only new hardware and software architecture, but also new infrastructures dedicated to multidisciplinary research on large datasets and new insights on eScience. While Google has just announced its collaboration with NASA to use the recently quantum computer acquired by the Universities Space Research Association from the Canadian-based company D-Wave, the race is started over between organizations to solve the big problems with big data.

Objectives: to continue the development of storage and computing dedicated to big data, both for the scientific teams, and the industrials, including SMEs. This is already the case in France within the competitive clusters and/or the *Instituts de Recherche Technologique*, and must be done in a European perspective. For instance, the Institut Mines-Télécom coordinates the BADAP project (Big (a) Data Academic Platform), a big data as a service platform for researchers and SMEs. It is based on SQL-MPP solutions, and will cope fast data thanks to the Storm (see page 18) open source framework. A second objective is to unlock the scientific data, buried in books or in small labs. «Long-term data provenance as well as community access to distributed data are just some of the challenges.»

Industrials must provide real use cases with large datasets. As in the U.S., the healthcare sector can be one of the first provider of such use cases: it must be educated to big data.

Further reading: «As We May Think.» Bush, Vannevar. The Atlantic. July 1945. Starting point: http://en.wikipedia.org/wiki/As_We_May_Think

The term «datascience» must be used with caution, as it lacks the need to theory if it overrelies on data. Without theory there is no real learning and no real questions to ask. If we process massive amounts of data without human intervention and underlying theory, we will create a data cycle of collect-analyze-act which may appear valid, but is in fact a short-term view.

#4 Learn to apply context to the numbers

From content to context: data without context tells a misleading story.

We showed on page 10 how the Google's flu trend algorithm was looking at the numbers, not at the context of the search results.

How a computer can sense the context of search results, and more generally of data? The equivalent of the five human senses for computers are: date and time, geographical location, physical environment (the weather information are just a website away), topic of interests inferred from websites open by the user, emails or created contents.

Objective: develop proofing methods to ensure that the numbers/raw data are always analyzed in a meaningful context, and remain so when additional data are collected, especially via data correlation (see challenge #10). Make the data scientists able to ask the right questions by providing them a general set of fundamental questions.

#5 From data to sentiments

Sense the mood in the room.

Numbers and raw data are cold assets. Data is not information, information is not knowledge, and what does matter is subjective information: sentiments, emotions, intentions and opinions.

A five-year-old child can say immediately her parents mood, making sense out of heterogeneous contexts, but it is still hard for a computer program to figure that out. Computers can perform automatic sentiment analytics from digital texts, using machine learning algorithms such as semantic analysis or support vector machines. It is harder for face recognition systems. And still a challenge for mood trends to be known from our digital behaviors split on several social networks and made of small piece of contents. The so-called augmented pervasive intelligence will provide contextual applications, that learn from our daily behaviors through our mobile device, and propose help to facilitate our days.

Technical challenges

#7 Will “delete” become a forbidden word?

From storing important data to keeping all data.

The cost of storage has dropped dramatically, we end up keeping all the data, but can we keep everything forever? Beyond the technical and ecological questions, the main issue still remains that of privacy and that of the ‘right to be digitally forgotten’. This is a long debated regulatory question in Europe: the explanatory memorandum concerning the European legislative proposal for a General Data Protection Regulation can be found at <http://goo.gl/tU0r5>.

This may also be a cultural question between digital natives and other people, or even between countries. A technical answer and a growing trend is the so-called «short-term social networks» launched with the promise that the users set a time limit for how long their friends can view their contents, after which it will be hidden from the friend's device and deleted from the company's servers. However, it was reported recently that Snapchat, one of the more famous short-term network, didn't actually make contents disappear, and this raises the question of trust.

If we do not keep all data, then we cannot keep open all options for further interesting analytics, when datascientists develop new theories and models, and go back in time to understand these new models. But if we collectively choose to keep every data, we must ensure to never create data aggregates that compromise privacy.

#8 Anonymize for good

Data can either be useful or perfectly anonymous but never both. Really?

This is one of the most challenging problem, and a major blocking point for many collaboration between industrials and/or researchers. It's a scientific and societal challenge too, which still remains open.

Objective: ensure that whatever happens in the future, even with adding complementary data or crossing data not originally designed to, data that should not be associated with a person will never be.

Metaphor: how can we tell something to a child and be sure she will not remind the words in another context, embarrassing the whole family?

Question: is there a Heisenberg principle asserting a fundamental limit to the certainty with which certain pairs of properties of data, such as usefulness and anonymisation, can be established simultaneously?

#9 Do not neglect big data risks

Mischaracterizing data resulting in privacy violations are one of the risks pointed out by Miller, H.E. [op.cit.]. The fail to meet the user requirements due to the inability to process data, or the uncertainty regarding who owns customer information, are other risks affecting the end user.

Addressing the wrong problem, focusing on the near domain and ignoring the real problem or mischaracterizing data resulting in poor decisions are among the risks facing datascientists.

#6

Make the data qualitative and meaningful

Drop the DRIP syndrome.

The Data Rich/Information Poor syndrome refers to «the problem of an abundance of data that does nothing to inform practice because it is not presented in context through the use of relevant comparisons». It is still an important problem in a big data context. In a heterogeneous data context, it is easier to use data that we know the meaning of. Both scientists and organizations need new capabilities that rely on new semantic approaches, shared vocabularies and ontologies. This is a prerequisite to the ability to master data correlation (challenge #10) between disciplines or sources of data. And before being meaningful, data must be qualitative.

Dimensions of data quality

- **Relevance:** Do the data address their user's needs?
- **Accuracy:** Do the data reflect the underlying reality? Is the level of precision consistent with the user's demands?
- **Timeliness:** Are the data current, relative to user demands?
- **Completeness:** Does the level of completeness correspond with user de-

mands? Data can be incomplete or even too complete!

- **Coherence:** How well do the data hang together? Do irrelevant details, confusing measures or ambiguous format make them incoherent?
- **Format:** How are the data presented to the user? Is the context appropriate?
- **Compatibility:** Are the data compatible in format and definition with other data with which they are being used?

- **Accessibility:** Can the data be obtained when needed?
- **Security:** Are the data physically and logically secure?
- **Validity:** Do the data satisfy appropriate standards related to other dimensions such as accuracy, timeliness, completeness and security?

Source: Miller, H. [1996]. *The multiple dimensions of information quality*, *Information Systems Management*, 13 (2), 79-82

#10 Master data correlation

*How to cross data without a priori?
How to find relevant datasets?*

How do I use this data type, which I have never seen, with the data I use every day? One could give many more examples of questions arising in this «mashing up knowledge» that is being made possible. One possible answer to this challenge would be the deployment of a real worldwide network of open datasets and API directories, an artificial intelligence facilitating to cross any sort of data by simply testing, and even proposing new correlations by itself.

#11 Master the big data cycle

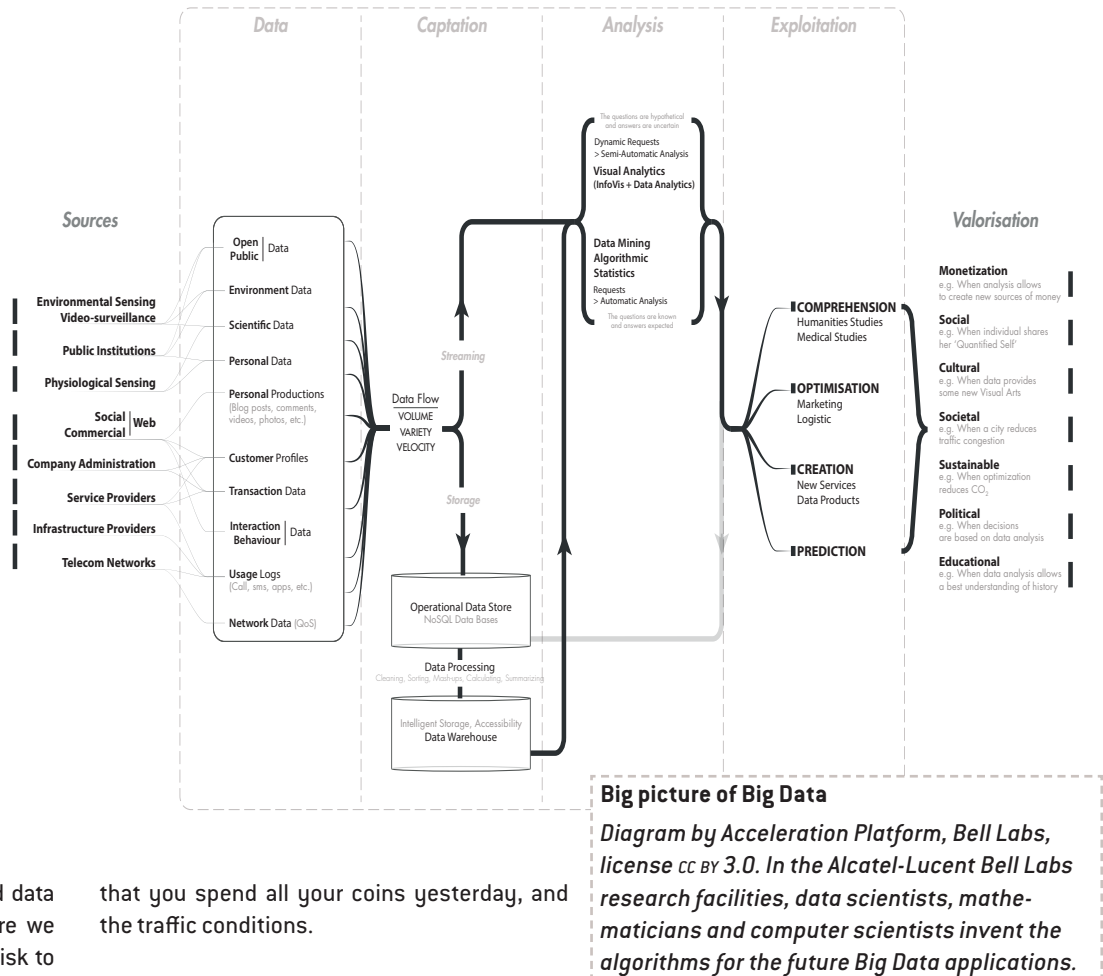
The more you use data, the more you produce data.

From the ocean of data to the storage cloud and then to the many actors processing data and producing new data that we saw in the data landscape on pages 14-15, data is continuously circulating as water in the water cycle. The more we consume data, the more we find data useful for our daily usages, and the more we try to produce even more data. With the risk to produce inaccurate or redundant data. With the risk to produce data which is useless or could be backfired on us in the future.

Issue: produce only relevant and not yet known data, and to be able to find relevant datasets before adding its own data. Produce neither too much nor too little data, in order to get results from our queries without preventing the serendipity, which is a central part of the quest to produce new knowledge.

#12 Make the mobile phone your data assistant

Use case: You are driving your car, in a hurry towards your lunch, you don't have money left to pay for it but you don't know it. Your mobile does know, and it will route you, *on its own initiative*, to the nearest cash machine. It does know your agenda, your habits, the fact that you are a bit absent-minded these days, the fact that this restaurant does only accept notes, the fact



Big picture of Big Data

Diagram by Acceleration Platform, Bell Labs, license cc BY 3.0. In the Alcatel-Lucent Bell Labs research facilities, data scientists, mathematicians and computer scientists invent the algorithms for the future Big Data applications.

that you spend all your coins yesterday, and the traffic conditions.

Mobile phones were once «personal data assistant». It is now time to make this real: a device that could even pass the *Turing tests*.

#13 Enable dataviz on new devices

People understand what they can see.

Information design pioneer Edward Tufte has one primary rule: **show the data**. A second rule was **show comparisons**. To support and encourage new powerful ways of thinking, the data must also be **manipulated** in their environment.

Use case: at a dinner with friends, talking about energy saving. Look at the wall and access in a comprehensive way to a public dataset of heat loss measures in your neighborhood.

Issues: data-oriented human-machine interfaces for mobile devices, for motion sensing input devices like the Kinect, for augmented reality devices like Google glasses, and for human body interfaces like Interaxon's Muse.

#14 Invent the future of shopping

From company-centric to user-centric data management.

With our big data oriented new devices, it is possible to establish the Small Data paradigm.

Scenario: forget all these hours spent on comparative engines to replace your just broken down espresso machine. Scan the QR code on the machine, and broadcast an «intentcast» to the marketplace, without revealing any personal information. You will get in return only relevant offers, without being polluted by irrelevant ads, and the fear to know your personal data are in the hands of third party vendors.

*Further reading: «The Customer as a God.», Essay, July 2012, Wall Street Journal.
<http://goo.gl/XkGyO>*

Societal challenges

#15 Open new ways of thinking

Both industrials and individuals must change their minds, and accept to open the data they own or produce. For industrials, in our world-wide competition, it is an illusion to protect their data assets anymore. In order to unlock its own data value, it is necessary to start thinking how to get more value from external interactions. Carefully tuned API can protect their data, while at the same time allowing their datasets to become an internationally-recognized reference. With regard to the citizens, the challenge is to accept to share more personal data in order to feed the humankind database, as long as privacy is respected. This new ways of thinking must be lead by visionary entrepreneurs.

#16 Democratize data management

As we saw in pages 16-17, there are thousand of machine learning methods, but textbooks are written for researchers, not practitioners. In this new arising economy of data, everyone should be confident with data and data manipulation.

Issue 1: make people understand what data is.

Inspiration: School of Data provides courses for everyone, from data-newbie to pro looking for new ideas. It works to empower civil society organizations, journalists and citizens with the skills they need to use data effectively.

<http://schoolofdata.org/>

Issue 2: make people want to play with data, and able to make their own minds from raw data.

Inspiration: In his «Ten Brighter Ideas» essay, Bret Victor proposes a prototype of a reactive document, full of data collected on the web. The reader can play with the premise and assumptions of various claims about energy saving, and see the consequences update immediately.

<http://goo.gl/TyM8p>

#17 Teach the future datascientists

The sexiest job of the 21st century?

The leading training institutions in datascience are largely Anglo-Saxon. Without being exhaustive, we can cite the masters programs dedicated to «machine learning» proposed by Carnegie Mellon University, Berkeley, Stanford University and the Stanford Center for Profes-

sional Development to «Data Mining and Applications Graduate Certificate» training three years in partnership with Sony and Cisco, MIT, Chicago Northwestern University (Predictive Analytics), North Carolina State University (MSc in Analytics in partnership with SAS) or UC San Diego (certificate program in data mining).

Next Fall, Telecom ParisTech shall propose a novel Professional Master program, fully dedicated to big data. It aims at teaching concepts and techniques required to manage and exploit massive data, in a progressive and very complete manner. It includes technical courses related to the following topics: semi-structured databases, machine-learning, web technologies, decision support systems, distributed computing, computer security. Beyond the acquisition of general knowledge related to the different fields involved in big data (computer science and applied mathematics especially), the goal of this training is to develop effective skills. The theoretical content of the lessons shall be illustrated by a variety of case studies and practical applications (e.g. design of recommending systems, design of search engines in information retrieval) arising from a variety of fields, ranging from e-commerce to finance through defense/security. Societal aspects of the Big Data phenomenon, legal (privacy) and economic, shall also be investigated at length during the training.

In order to keep the program in line with the needs in the industry, a number of companies (from start-ups to big companies) in a variety of sectors (defense, Internet, finance, high-tech) shall be involved in the training. Partners include in particular Thales, BNP ParisBas, Safran group, EADS, Criteo, Liligo, SAS, Capgemini, and IBM.



«User-centric data management, data democratization, new ways of thinking, numbers in context and meaningful visualization...» Our future in the new Economy of Data is full of promises, if we can avoid a «Data Divide» between those who have access and the opportunity to make effective use of data and those who do not.

This could be the biggest challenge we face. ■

“Before 1786, authors invariably presented quantitative data as tables of numbers before the economist William Playfair published a book called *The Commercial and Political Atlas*, full of line graphs, bar graphs, and other pie charts he created specifically. Today, people take these graphical forms for granted; they seem as obvious and fundamental as written language.”

Bret Victor, designer,
worrydream.com

Working with the Institut Mines-Télécom

This *cahier de veille* was written with the help of several **contributors** from the schools of the Institut Mines-Télécom. **Stephan Cléménçon** is a teacher-researcher at **Telecom ParisTech**, a member of the department TSI (Image and Signal Processing) and works in the lab LTCI (Communication and Information Theory). His main research contributions are in the fields of Markov processes, nonparametric statistics and statistical machine-learning. He is responsible of the Industrial Chair «Machine Learning» at Telecom ParisTech. **Alexandre Gramfort** is an assistant professor at **Telecom ParisTech**. His research interests are on mathematical modeling and the computational aspects of brain imaging. He is more generally interested in biomedical

signal and image processing with a taste for scientific computing, numerical methods, data mining and machine learning. He is one of the contributors of the scikit-learn machine learning framework (see some of his applications at <http://martinos.org/mne>). **Claire Levallois-Barth** is a teacher on legal and privacy aspects at **Telecom ParisTech** and **Telecom SudParis**, among others, she is responsible of the Chair «Values and Policies of Personal Information» launched by the Institut Mines-Télécom in April 2013, and member of the expert network for Etalab. **Bruno Defude**'s team at **Telecom SudParis** is specialized in the field of High Performance Computing. **Cécile Bothorel**, is a researcher at **Telecom Bretagne** in the department LUSSI,

and a member of the Social Networks Chair for eMarketing at the Institut Mines-Télécom. She conducts work on social networks analysis, focusing more particularly on the detection of implicit communities of interest over the social web. She is also involved in social network analysis-based recommendation problematics. **Claude Berrou**'s Neucod research program at **Telecom Bretagne** aims to identify and exploit the strong analogies observed between the structure and properties of the cerebral cortex and those of modern error correcting decoders. Awarded a grant of 1.9 million euro by the European Research Council, this project based on a multidisciplinary approach will give new fresh insights on machine learning.

Additional documents are available on the partner area of the site of the Fondation Télécom.

Ant by Jacob Eckert, from The Noun Project.

Glossary

anonymisation: the process of treating data such that it cannot be used for the identification of individuals.

API: Application Programming Interface. A way computer programs talk to one another. Can be understood in terms of how a programmer sends instructions between programs.

attribute & share alike: a Creative Commons license that requires attributing the original source of the licensed material, allowing derivative works under the same or a similar license.

commodity hardware: computer hardware that is affordable and easy to obtain.

compound visualization: the visualization resulting from two (or more) spatially distinct different data representations, each of which operating independently, with the possibility to be used together to correlate information in one representation with that in another.

dark data: the value of dark data is locked up in a way so that it isn't readily available for use by analytics. Also, data accumulated and still unused.

data crunching: a marketing term. The process of collecting and cleaning the data.

data munging: the process of converting or mapping data from one raw form into another format that allows for more convenient consumption of the data.

Data Science: a recent term (see sidebar page 23) with multiple definitions. Drew Conway's Data Science Venn Diagram is a good definition as a start (see page 20).

early warning signs: patterns one wishes to see long before their impact on an observed phenomenon.

eScience: a recent term and a new research methodology. Computationally intensive science that is carried out in highly distributed network environments.

Etalab: a service under the French Prime Minister, in charge of the French Open Data initiative.

www.etalab.gouv.fr

Hadoop: an open source software project administered by the Apache Software Foundation that enables the distributed processing of large data sets across clusters of commodity servers. See details page 10.

information: A piece of information. Value added from the process of collecting and organizing data. Information needs to be converted into knowledge before it can be used, by relating it to oneself, one's experiences, environment and other contextual information.

JSON: JavaScript Object Notation. A common format to exchange data.

knowledge: based on the value added from the process of organizing information, plus expert opinion, skills and experience, plus a far more complex mix of creativity, serendipity, and social and cultural binds.

lifelogging: the process of wearing specific devices to capture continuous physiological data.

MapReduce: a framework for processing parallelizable problems across huge datasets using a large number of computers (nodes), collectively referred to as a cluster. See page 12.

open data: data that can be used, reused and redistributed freely by anyone for any purpose. Details on page 10.

pattern: in machine learning, a non-null finite sequence of constant and variable symbols.

quantified self: a movement to incorporate technology into data acquisition on aspects of a person's daily life in terms of inputs, states, and performance. Also: self-monitoring and self-sensing. See page 7.

raw data: primary data that has not been subjected to processing or any other manipulation.

schema: the structure that defines the organization of data in a database system.

sentiment analysis: the application of statistical functions on comments people make online and through social networks to determine their mood.

serendipity: the ability of finding something good or useful while not specifically searching for it.

Small Data: a recent term with multiple definitions. Emphasizes the need to decentralized, more localized and ultimately user-centric, data management.

SQL: Structured Query Language. Initially developed at IBM in the early 1970s, it is a popular programming language designed for managing data held in a relational database management system (RDBMS).

transactional data: data that changes unpredictably.

Turing tests: a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of an actual human.

XML: a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

See also:

<http://schoolofdata.org/handbook/appendix/glossary/>

<http://data-informed.com/glossary-of-big-data-terms/>

Les cahiers de veille de la Fondation Télécom

The *cahier de veille de la Fondation Télécom* is the result of studies conducted jointly by Institut Mines-Télécom professors and industry experts. Each *cahier*, which deals with a specific topic, is given to researchers at the Institute who gather around them recognized experts. All at once comprehensive and concise, the *cahier de veille* offers a state of the art of the technology and an analysis of both the market and the economic and legal aspects, focusing on the most critical points. It concludes with perspectives that are all possible ways of joint working between partners of the Fondation Télécom and the Institut Mines-Télécom teams.



With the support from:

**Alcatel-Lucent, BNP Paribas, Google, Orange and SFR,
founding partners of the Fondation Télécom**

And with Accenture, Astrium Services, Cassidian
Cybersecurity, CDC, Sopra Group and Streamwide

Fondation Télécom

46, rue Barrault - 75634 Paris CEDEX 13 - France

Tel.: + 33 (0) 1 45 81 77 77

Fax: + 33 (0) 1 45 81 74 42

info@fondation-telecom.org

www.fondation-telecom.org

