Value in Data

Warith Harchaoui and Laurent Pantanacce

4e Révolution

June 2021

Version française

Introduction

Each industrial revolution is driven by a driving force: a raw material, an energy source, a creative technology that redefines the economy. Since the 19th century, we can list steam, coal, oil, electricity, radio, the transistor, computers, and today artificial intelligence (AI). This fourth industrial revolution is troubling because its commodity is abstract: data. At the scale of humanity, this seems a more significant milestone than a means of communication such as radio or the Internet. Beyond the industrial and economic consequences, artificial intelligence should be understood as an upheaval comparable to agriculture and sedentary life 10,000 years before Christ, the invention of writing 3,000 years before Christ or even printing in the 15th century. The very tangible impact of such a contribution profoundly changes our relationship to the world. A machine is now able to perceive, process and decide on the basis of received information and human beings are then forced to question their certainties. For companies, almost all areas of our contemporary world are now impacted by this artificial intelligence. Philosophically, it is the automated emergence of certain aspects of the mystery of intelligence thanks to mathematics applied with computers. Concretely, artificial intelligence is the science that allows a machine to make decisions without the scenarios having been exhaustively explained by a human intervention.

Taking a good look at data, giving it the structure it deserves, exploiting it, enriching it and transforming it are the real technological horsepower of what is much more than a contemporary industrial revolution. By way of comparison, there is no difference between coal and diamond other than the molecular structure: it is "only" the geometric arrangement of the same carbon atoms repeated for each space dimension that is different. However, the exploitation of coal has allowed tremendous industrial repercussions that cannot reasonably be compared to Marilyn Monroe's best friends, even if we can admire the grace of the artist, the beauty of the famous song and the brilliance of her diamonds. This same idea of changing structure, organization and arrangement is crucial for artificial intelligence and its raw material, data. Diamonds are not combustible, and coal is brittle,

but they are both only carbon. Likewise, data requires care, concern and purpose in its application.

In order to avoid sub-optimal exploitation of a raw data mine, the challenge is to address three central questions: Where is the data and in what state? What tools are available to refine the data and make it usable? How can we use this data economically, commercially and financially once it is deemed fit for consumption?

Data in all their moods

The topic of data access is not trivial. The recent news of the LinkedIn vs. HiQ[1] case is a good example. Indeed, the American justice system made the decision on September 9, 2019 (confirmed on June 4, 2021) to let HiQ suck up information made public by users on the professional social network LinkedIn. The accepted term for this practice is web scraping and the jurisprudence of this decision will offer new "data" opportunities for this flourishing field.

According to the Statista[2] portal, the total amount of computer data in the world has increased by a factor of 20 from 2010 to 2020 to represent 47 trillion Gigabytes. It is also expected to increase by almost 500 times between 2020 and 2035. This is so huge that we can naively think that there is probably some value in all this public and free data on the internet but unfortunately it is not that simple. Indeed, this data is not always useful and true but let's assume that it is. To exploit this data in the most relevant way, we encounter some technical obstacles that bring together several trades to get by. First, the data is not always publicly accessible and free.

For a company, the "Data" effort is an enduring one, otherwise it is unfortunately vain in what we observe in practice. In fact, it is a real voluntary policy encouraged by management, otherwise some work may be done in the right direction, but we would systematically fall back on the problem of data locked up in IT "silos" with internal political rivalries or even baronies. The risk is to pay for the effort of collecting data without benefiting from data that is de facto inaccessible: buried for human reasons, security, computer and/or graphic interface that is not very user-friendly, obsolescence, in short, poor management.

On the Internet, information can be present in the form of text but also images, photographs, sounds and videos. In these cases, the use of perceptive artificial intelligence is relevant. For handwritten or typed text contained in images and photographs, OCR (Optical Character Recognition) is used, which transforms the text present in the form of

[1] https://www.reuters.com/technology/us-supreme-court-revives-linkedin-bid-shield-personal-data-2021-06-14

[2] https://tinyurl.com/worldwide-data

pixels into editable text. For sound, the audio transcription into text popularized for doctors' dictaphones has found a new lease of life with the advent of podcasts (sound broadcasts on the web). We also have the recognition of objects and people in photographs with the spectacular progress of the Computer Vision field. Since the 2010s, these technologies have become really convincing for professional use, as a service, a kind of basic commodity like water, electricity and gas to the point that the expressions AI is the new electricity and AI as a Service have become standards. For the decade 2020, we expect huge progress well underway for text understanding: emotions, intentions and meanings to open up even more opportunities.

Extract oil from the data

At this point, the data is stored in your database system because the computational effort has been made and perhaps also the perceptual AI effort. From Machine Perception we can evolve to Machine Learning[3] which are obviously sister sciences without entering into the variations of terminology linked to the changing fashions according to the decades. At present, we can make exploratory steps called descriptive statistics which is a largely underestimated field whereas it is one of the simplest and initial of a well constructed AI process. It is a matter of questioning the data by calculating simple things like averages, the most rudimentary indicators simply to have the orders of magnitude of the data we are manipulating. This physicist intuition[4] is not only a question of scientists, we humbly recommend it to company managers as well. It is a matter of taking hold of the issues and making up one's mind according to one's intellectual curiosity.

In artificial intelligence, without seriously summarizing in a single paragraph such a huge scientific field, we can still distinguish :

- **supervised** learning where one has collected input and desired output data (often manually annotated) to build a model and hope to have the right output on an input that the system has never seen
- **unsupervised** learning where we only have the data and we try to decompose it: in characteristic dimensions (dimensionality reduction to see the data in 2D or 3D), in groups (clustering), by learning to imitate them (generation) for example
- **reinforcement** learning where we make decisions without always knowing if they are immediately good. For example, in chess, in Go, in robotics, it is very useful

[3] https://research.google/research-areas

[4] *The Pleasure of Finding Things Out*, Richard Feynman (posthumously published in 1999)

Driven by companies like GAFAM[5] and private foundations like OpenAI, we are gaining access to on-demand (for a fee) artificial intelligence models of very high quality because they are learned on gigantic annotated amounts of data. This gives smaller enterprise customers the opportunity to offer AI-enhanced functionalities to their customers, such as chatbots (conversational agents), machine translation, speech or image recognition, customer relationship management, etc. In concrete terms, we can say that 95% of the models are ready to use (this is just a way of speaking) and the remaining 5% have to be developed specifically if necessary. We can say that this is often the basis of a good first version before either specializing for a new need, or having the conscious ambition to redo everything at home with the means that this implies. It seems important to us to specify that a brand new scientific library with impressive results is not a product even when the temptation to download it for free is great without the real effort of adaptation on the internal data and those of the customer, not to mention the setting in production.

A definition of the Data Scientist job could be this famous tweet of Josh Wills (to whom we owe Spark) in 2012:

A data scientist is someone
who is better at Statistics than any software engineer
and better at Software Engineering than any statistician

This job consists of kneading data until it can be cooked into a format that has value. It's a computer science and mathematical experiment to cross-reference data to enrich it. It's also about understanding that the best toolkits in the world can't do anything on bad, poorly arranged or poorly verified data. It's better to keep things simple first because not everyone can afford to fund a Google-like research lab in-house (GAFAM researchers themselves first make baselines in their papers before going further, at least to compare performances). Putting into production a reliable method controlled by even rudimentary performance indicators (KPIs) is a sound approach for employees, their hierarchy and therefore the whole company.

As Ted Benson explains in his book *Automating Paperwork*[6], the reason why artificial intelligence of text (Natural Language Processing/Understanding) is having a strong impact on service businesses is that the vast majority of processes can be seen as parsing, extracting and converting text information. The inputs and outputs of these steps are all text (including spreadsheets and databases). According to Gartner (2020), there are more than one billion knowledge workers as of 2019, i.e., people whose work receives information as raw material and outputs information as product. We can therefore

[5] GAFAM: Google, Amazon, Facebook, Apple and Microsoft, and Tesla joining this group in terms of market capitalization

[6] https://edwardbenson.com/automating-paperwork

understand the magnitude of this strong trend on companies and our societies. With a little perspective, we can see that text still remains a great challenge for companies as a medium of information, probably because language is deeply linked to human intelligence itself.

Creating value with data

Now that the data has been aggregated, roasted, enriched and even synthesized, it's time to add value to it and exploit the threads of value it can provide.

Let's make the value reside where we don't expect it!

Data should never be used for reassurance. Too often, we try to make data say what we already believe; without even attempting to use it to counter our biases, preconceived notions or worse our lack of perspective as data scientists but also as business leaders. "Let the data show us what our minds don't see or refuse to see" is the watchword, it's about sometimes stopping ourselves from imagining and practicing to confront our intuition with the stubborn reality. The worst phrase that can be used in this case is "besides, the numbers don't lie!". These few words should sound like a warning signal; every time they are quoted directly or via synonyms, distrust must be born! As the saying goes, "Numbers are like people. If you torture them enough, you can make them say anything"[7]. Data should be there to change our way of seeing things rather than confirming what we already think: data destroys myths and even our mental barriers!

A few years ago, in the middle of launching a connected object project to manage gardens, we were desperately looking for statistics on gardens, pools, houses (main or secondary)... All the quality studies available on the market were worth much more than the capital of our newly created company. Forced to be frugal, we had to find a different way to get an idea of the size of the market. That's when I came up with the idea of counting pools; any satellite view suddenly becomes your best ally; first a count in the south of France and then more widely. The geographical distribution of the pools is astonishing: almost as numerous at the seaside as in the more remote areas; strangely numerous in Germany... All this was counter-intuitive. But above all, it forced us to ask ourselves real questions. The quasi-miracle was that once the pools were pointed out and counted, the reverse geocoding tools made it possible to obtain the address of the houses and often the telephone number of the occupants: a commercial manna that no PDF study could have offered! This anecdote shows how many free tools are available on the internet (Google Maps here but usually GitHub or even interactive websites) and that by persevering, a

[7] *Nombres en Folie - les divagations du mathématicien fou* (2013) by Didier Hallépée

company manager who plays at being a Data Scientist finally creates value. Shouldn't we finally unleash the Data Scientist in all of us?

Data analysis as a movie, not a photograph.

They say "it's the journey that matters not the destination" which perfectly applies to data. What creates value is not an analytical PDF produced for a single purpose. On the contrary, value is created by using living data via a creation/production process that can be reused over time to track changes. In other words, whether the analysis is done by an intern or by a large consulting firm, it will never create as much value as if it can be recreated on the fly with regard to updated news. Later, the satisfaction of re-running the same analysis but on the updated data is a creative inspiration amplified by the experience of what actually happened. In this process of analysis, a proactive mind creates irreplaceable value by going back and forth between data and reflection. The worst thing that could happen is an excellent (and possibly expensive) static analysis on the basis of data that will very quickly be forgotten, neither the source nor the methodology used to produce it.

From sharing comes even more value.

Once data has created value for you, your department, your company, your customers... ask yourself how much more value it could create? Often, Chief Data Officers sit on top of data like a treasure. They are right, it is often a part of the company's history! However, data is more like a sack of wheat. Unused, it rots. Used properly, it feeds new analyses and creates even more value.

Late Hans Rosling, Public[8] Health Professor, and above all a huge advocate of open, accessible and shared data, used to hammer home the following point in many of his lectures: "Some countries have agreed to make their data public, but what we really need is, of course, a search function. A search function that would allow us to copy the data into an accessible format and share it with the world."

Today, companies should have the same approach to sharing their data, they would be surprisingly pleased by treasures forced to unmask themselves behind the data; not only for the benefit of others but also for their own. Sharing one's beautiful data is getting human neurons to reveal the value that was hidden in one's own data. Finally, let's

[8] To open up your mind :

- Great Book by Hans Rosling *Factfulness : Pourquoi le monde va mieux que vous ne le pensez*, 2019
- Hans Rosling TED's video : https://tinyurl.com/hrosling-ted

celebrate the specificity of this *data* commodity: if I give knowledge to someone, he has become richer and I have not become poorer!

Conclusion

When gleaning data in the hope of extracting value, common sense reflexes are required: if someone has collected this data, who is it? Is the source reliable? can we say after examination that a curation work has been done? These are important checks because they are a source of bias, which is poisonous for statistical models. From simple averages to the latest generation neural networks to the best algorithms of the great specialists, statistical pollution is sneaky. Being at least aware that the data is only a statistical sample, (sometimes large), can avoid major errors.

In philosophy of science, we owe the statistician George Box a quote of such British phlegm:

All models are wrong, but some are useful

This powerful idea reminds us how crucial the goal of the scientific approach is. We set goals that are as tangible, objective and measurable as possible, otherwise the company loses its creative energy. Controlling each artificial intelligence module is essential: in input data for verification and in output data to measure performance. We argue that doing otherwise is like tilting at windmills of unfulfilled promises. Data is only relevant and models are only useful if a goal has been defined beforehand. For the prediction of a particular phenomenon, data will not be collected in the same way. Misguidedly ignoring goals in the name of stakhanovistically downloading algorithms and data is a scabrous source of bias and a money pit that nobody needs.

When dangers are avoided, when data is well ordered, its uses finally take on a decisive value for companies. By taming one's own will and preconceived ideas, with the rigor of a museum curator and the adventurous spirit of a discoverer, the treasure trove of data ultimately opens up to companies.