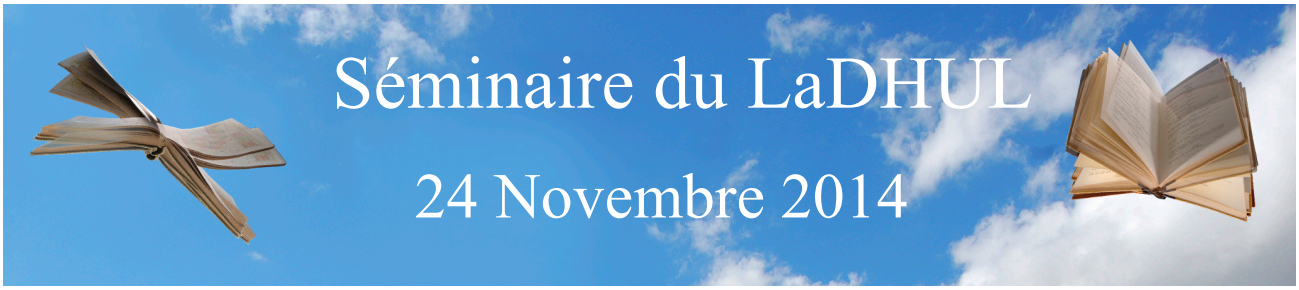

Periklis Andritsos, « Big Data Challenges, Opportunities and Avenues of Research, or "How did I grow up talking to data" »

Recently appointed full professor in information systems by UNIL's HEC faculty, Periklis Andritsos presents us some of his research interests as well as broader reflections on what is nowadays called "Big Data." As a computer scientist with deeper training in databases, he practiced his *data science* skills in different countries (Canada, Italy) for academic as well as industrial purposes. Besides his research activities, he is now responsible for a graduate course entitled "Web Scales Analytics."
The (quite problematized) summary of this talk will be organized as follows: we will first deal with broad characteristics of "Big Data" and mention some difficulties engendered by its promises. We will then present some of P. Andritsos's research interests. Finally we will use some of the previous elements to give some insight into what "Big Data" may *provoke*.

What, why and how is "Big Data"?

In this first part, we will try to briefly present the origin of the notion "Big Data" as well as some of its encapsulated promises that are not always easy to keep.

2011 : the McKinsey Report

In 2011, multinational consulting firm McKinsey published an influential 156 pages report entitled "Big Data: The next frontier for innovation, competition, and productivity." More than just defining "Big Data" as huge and always increasing datasets that require new forms of analysis, the report assigns it important marketable values: between 140 000 and 190 000 new job positions for people with deep analytical skills by 2018; data mining as a key production factor; managing big and heterogeneous datasets as key basis for competition and growth, etc. Even though skills and practices that organize, process, mine and interpret huge amount of heterogeneous data existed before 2011 – P. Andritsos met for example the concept in 1997 in a conference on "Very Large DataBase" (VLDB) – the *term* "Big Data" that associates the exploration of huge quantities of data with economic growth and innovation really started to become trendy after the publication of this report. Here, Google Trends provides us a nice visualization of McKinsey report's (almost) creation of the term "Big Data":

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

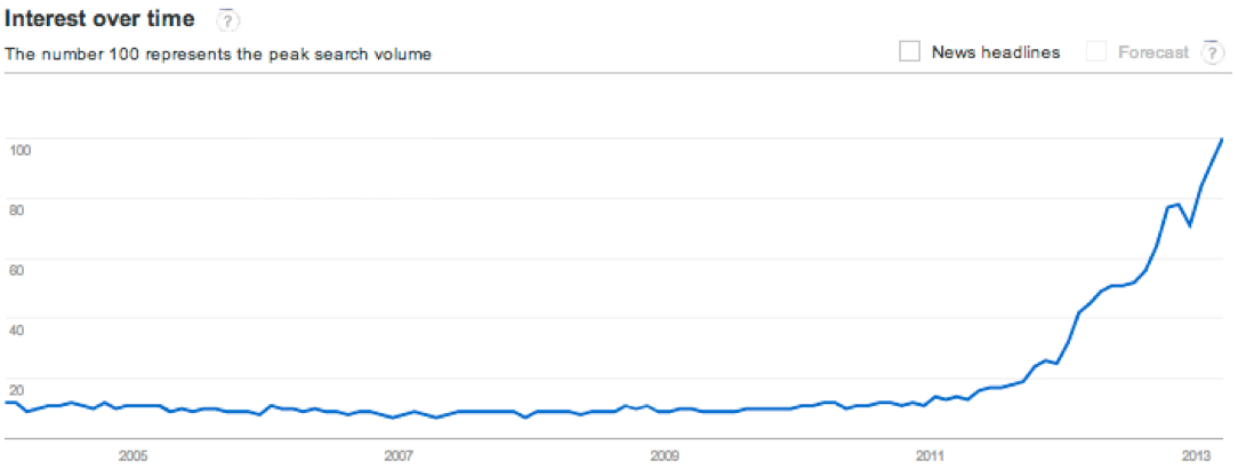

Figure 1: evolution of "googlization" of "Big Data" according to Google Trends. Note that the curve really starts after 2011. Difficult to see here but also interesting: the small bumps in December 2012 and 2013 which indicated that "Big Data" is less attractive during Christmas vacations.

3 then 4, then 5, then 6 V's

By predicting new opportunities *via* Big Data explorative capabilities, McKinsey report also created promises that are not always easy to keep. Indeed, exploring huge amount of data is hard, tricky and may quickly create *inaccurate* results. One example provided by P. Andritsos is a 2011 collaboration with biologists who wanted to probe academic publications about specific proteins called NOK, the abbreviation for "nagie oko protein." Was it worth the try for those biologists to invest in this NOK area of research? Since at that time PubMed already gathered about 20 millions articles, they needed some help to test the waters by means of text analysis. But text analysis of huge amount of data *by itself* may not prevent from semantic problems such as the fact that "NOK" is highly equivocal: it can refer to "nagie oko protein" but also to Norwegian Couronnes and the acronym of "Next of Kin." So instead of only extracting and analysing papers that speak about a specific protein, P. Andritsos and his text mining tool would have include – for example – papers about investments of the Norwegian health system in medical research or papers about family health histories. It is only *retroactively* and after quite a huge amount of post-processing that he could translate the different NOKs into vectors and thus evaluate their relative distances. Data that was finally sent to biologists for further analysis was truthful because of his efforts of adapted geometric translations.

This example and its emphasis on the *qualitative importance* of data mining indicate – perhaps – a kind of shift in data science. In 2001 – before the emergence of the term "big data" – former Meta Group (another consulting firm now part of Gartner Inc.) characterized the analysis of very large databases as dealing with *Volume* (massive amount of data), *Velocity* (streaming, quickly evolving data) and *Variety* (heterogeneous data with different formats and modes of production). After the Big Data "deluge," IBM proposed the inclusion of another "V" for *Veracity*: because nowadays people (and more importantly business leaders) know that volume of data is huge, quickly evolving and highly heterogeneous, they know that is hard to analyse them in an adaptive ways. Data analysts must then emphasize on the qualitative aspects of their mining strategies such as semantics or entity resolution problems. Another "V" labelled *Value* was recently included: because data is huge, swift, heterogeneous and tricky, Big Data projects must be adapted to specific situations with understanding of costs and benefits in order to produce marketable *value*. The NOK example is explicit in that way: without an emphasis on data *veracity*, P. Andritsos's work would have had less *value* for biologists.

In the continuity of this movement, a sixth "V" was recently proposed for "Visualization." Indeed, data analytics that can provide good visualizations of their processes of production tend to be more trustful and provide more value.

Current Work

After those introductory elements to better feel what Big Data may refer to, P. Andritsos presents us some of his current research interests, deliberately omitting the very technical details.

Extracting meaningful indicators to report, analyse (and predict?) events

Is it possible to compete with Google News? The answer of a start-up created in Toronto, Canada, by P. Andritsos and several other data scientists was (and still is) "yes, at least if you process streaming textual data that are produced *about* new events."

The idea here is quite subtle. Indeed, social media such as Twitter, Facebook or blogs provide sometimes (not always) huge amount of textual data *about* events that are not yet "captured" by professional journalistic organizations. And by means of clustering, it is actually possible to identify those not-yet-captured-by-professional-journalists events from the huge amount of textual data that is being shared, reTweeted, commented by witnesses, followers, Facebook friends, etc. In other words, by detecting *vectorially* close textual vibrations on social media, it is possible to provide exclusive outlines of events that have just appeared. Moreover, this kind of system could also be adaptive: since it relies on streaming data, it can adjust and update its aggregated outputs automatically.

But speed and adaptability of this *system of insights* (who's name was THOORA) is not the only advantages on Google News. Indeed, by operating from this unusual direction (from "comments" to "actual news"), it becomes computationally easier to provide statistical information about the evolution of one given piece of news. Is this piece of news "hot" (many times the topic of Tweets, comments on Facebook, etc.)? Since when does this topic provoke reactions? Those kinds of information could be provided by this system relying on big textual data analytics.

"Win a few, lose a few" as the saying goes. Because if this kind of Big Data system can quickly identify an event and statistically analyse it, it might be extremely hard for it not to propagate *rumours*. But once again, if the purpose of such a system is to *probe* and analyse geometrically close amateur textual productions, not fact-checked information such as rumours might be considered as interesting and thus deliberately be part of the outputs.

Finally, is it possible to *predict* the trajectory of a piece of news? Will a given piece of information continue to be "hot" or will it soon sink into oblivion? If a system is well designed enough to identify shared characteristics between trajectories of events, it might be able to assign statistical patterns. Yet, those computational ways to predict events still need research investments.

Exploring ways to resolve entity problems

Let's imagine that Migros and Coop have just merged. As a loyal customer of both shops, P. Andritsos has both fidelity cards. But those cards in his wallet imply that he had previously filled in both Migros and Coop's forms with his name, email, address, etc., which itself implies that at least two different employees put these two piece of information in two different client systems. But what if some mistakes occurred? Or, since he travels a lot, what if P. Andritsos had put his Canadian address in one form and his very recent Swiss address in the other? As long as Migros and Coop are two different companies, it does not really matter. But if they turn to merge, the same merged client database might store two different P. Andritsos: one who lives in Switzerland, another who lives in Canada. This is the entity problem: how to automatically understand from a massive amount of data that two different recordings are actually linked to same entity?

Once again, solutions have to do with clustering operations that consist in translating the content of the databases into accurate multidimensional *vectors* with specific *values*. Once you acquire vectors, you can operate *algorithmically* on them in order to produce *relative distances*. According to a certain distance threshold, those distances will then draw *packages* of entities that are more or less close from each other. This laboriously constructed closeness allow then to consider – at least geometrically – that two entities are somewhat, quite, very or even *actually* similar; this constructed closeness – if well designed – would for example allow the merged information system of Coop-Migros to calculate that P. Andritsos-who-lives-in-Canada and P. Andritsos-who-lives-in-Switzerland are close *enough* to be considered as one single entity. The solution of entity resolution problem relies then on adequate choices along this workflow that aims to produce case-specific geometrical spaces from which distances can be calculated. But once again, the *Veracity* of those distances matters: if your model does not fit your data and your purpose, the distances you will acquire might be inadequate.

Automatically providing databases from text files

In order to be able to make queries such as "How many models did Canon build in 1998?" big companies like Microsoft or Google have created – for marketable purposes – huge databases about products such as cameras, smart phones, TVs, etc. But they don't really have access to the detailed databases of, let's say, Canon, Panasonic or Nokia. They actually pay the manufacturers (or sometimes third party companies) to get metadata about their products in order to build their own specific big databases.

For many reasons (which remain unclear), metadata are sent as text files that summarize characteristics of products, as exemplified below:

- DELL LATITUDE X300 PENTIUM III 2.4GHZ 14.1" TFT WITH DOCKING STATION
- IBM THINKPAD T42 2490 PENTIUM 320 III 3.0GHZ 512MB 4GB XPP PRO
- IBOOK G4 NOTEBOOK 1.33MHZ 512MB 2.0GB DVD+RW MAC OS X 15.1" TFT
- IBM THINKPAD T42 INTEL PENTIUM III PROCESSOR 2.0GHZ 512MB 40 GB

Since you cannot use only text files to feed a database, laborious manual work has to be done in order to label those text files and indicate: " Here lies the name of the manufacturer"; "Here lies the name of the model"; "Here lies the speed of the processor" etc., etc. Some algorithms are then trained on those manually labelled data to learn how to detect these pieces of information automatically.

Manual labelling is hard and time-consuming because it is not always easy for the worker to know what term refers to what attribute. Mistakes can easily be done which may drive to errors in the final database. But what if the manual labelling tasks could be avoided? What if, from those textual files, some program could automatically create a database that could answer to queries like "Tell me how many cameras were produced in 1998 by Canon"? What if Google or Microsoft could use systems that detect *regular expressions* within text files, produce *dictionaries* and finally propose potential *candidates* to specific queries? That's an idea P. Andritsos is currently working on.

Bottom-up redesigning of relational databases

Relational databases work with functional dependencies between tables, attributes and keys. But as we have already seen, they do evolve through time: they may become bigger, people in charge may change (especially in the academic world), misunderstandings of data original protocol might occur, etc. In the end, primer indicators can be lost, functional dependencies weakened and redundancies multiplied. That is why some huge relational databases need to be reshaped. But how?

One way to contribute to this problem is to proceed "bottom-up" instead of "top-down." Instead of designing new tables and then populate them with values, the idea is find new hidden associations between the values of the old relational databases in order to give hints to people in charge of redesigning the database. Extracting new associations and problematic redundancies to help normalization; *talking to the data* in order to make a diagnosis and propose an adequate treatment: that is the basic idea that, of course, is easier said than done…

Conclusion: What *does* Big Data?

Based on the elements presented above, one may propose four preliminary statements about the performativity of Big Data. First, it tends to establish an influential research space for *unsupervised techniques* working on unlabelled data. Outputs can be distorted relative to initial expectations but the volume of data is assumed to make those distortions visible and thus correctable *without the need to label data manually*. Second, "correlations", "proximity" or "distance" are recurring words in Big Data results. Its geometric universe tends then to provide *tendencies* instead of – for example – explanations. Third, maybe because of its tendency to provide tendencies, *visualization* of results become one of Big Data great challenges. Fourth, Big Data capacity *and need* to practice on huge amount of unlabelled data creates new kinds of collaborations with the Humanities. In that sense, *Digital humanities* – especially through textual data mining – may also be considered as a laboratory for data science.

Florian Jaton

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL