Pragmatic Big Data and smart manufacturing

Amine Lies Benhenni[1, *]

[1]Dataswati, R&D Department, 91400 Orsay, France

Abstract. Smart manufacturing is usually associated with fancy cobots and automated production lines. But the Big Data ecosystem has a part to play, and is probably faster and cheaper to start with. The numerical revolution has dramatically accelerated these last years, with many implications in different industries. We propose in this article to present these new tools, and discuss some real use cases. We will also present the best practices and some of the inevitable challenges related to data projects.

The upcoming Industry 4.0 movement is full of promises. However, many still think it is the vision of a distant future, implying heavy changes and new fancy but expensive machines and tools. But it is first and foremost a digital transformation, and as such, will generate its share of data. It will therefore be possible to leverage the Big Data technological ecosystem to:
- harness and bring value out of the data,
- improve quality and conformance,
- automate controls on the production line,
- have a tight control on industrial and manufacturing processes as well as on the machinery and tools.

1 Big Data: a technological revolution

Since the first personal computer, every aspect of modern life was or is on the verge of being digitized with many evolutions and revolutions.

The Big Data ecosystem emerged as a technological solution to the challenges brought by the ongoing numerical transformation leading to:
- powerful computation capabilities,
- wired and wireless networks with high bandwidth,
- availability of storage resources at a massive scale (physical as well as on the cloud),
- higher adoption of smartphones.

The last decade was fruitful in the diversity of developed solutions, mainly to solve technological issues related to the now infamous 3V (Volume, Velocity, Variety).

While computers, chips, networks and storage are still going cheaper, connected objects are becoming a common commodity. New communication technologies and protocols are emerging to enable every single object to be connected to the internet in some way, with a growing popularity to for Low-Power Wide-Area Networks (LPWAN), specifically designed for IoT.

Today the ecosystem covers most of the technological needs, and starting a data project is more about choosing the right set of tools.

2 Data Science: a scientific revolution

The availability of data and resources at scale played a significant role in the ongoing acceleration in the Machine Learning and Artificial Intelligence fields, enabling fast experimentations with cheap resources.

Internet companies like Google and Facebook, are leading the movement, both on business and on R&D. Google is a pioneer in some of the key Big Data technologies [1,2], and is now a reference in AI with more than 200 journal or conference papers published in 2016 on AI related topics.

Nowadays, a larger set of open-source predictive tools for Machine Learning [3–6] and Deep Learning [7–10], which lead to an exploding adoption among data practitioners. These tools are very versatile and can be used for different tasks, depending on the data available and the questions asked. They can be formalized as:
- Supervised learning, which is about learning the mapping between a set of features $(x_1, x_2, \ldots, x_n)$ and a set of output $(y_1, y_2, \ldots, y_n)$ known from the available data. The x_i can be scalars as well as vector, or even higher dimensions tensors. The output can be a continuous value (regression) or a discrete label (classification). This is by far the easiest and most popular approach, as it minimizes the volume of data necessary to have decent predictive results. Regression can be used to predict measurements and sensors values, while classification will learn optimal decision boundaries, like the ones needed for predictive maintenance and quality control automation.
- Unsupervised learning, where there are no outputs related to our inputs. Instead, algorithms look for patterns, for instance by grouping similar types of failures into distinct groups to help diagnose and fix them. It is possible of course to have a mixed approach, where some of the data are labelled, and some are not. Semi-supervised learning techniques will leverage the

---

* Corresponding author: albenhenni@dataswati.com

available information from a small set of labelled data to improve the overall learning accuracy.
- Finally, reinforcement learning is about learning a specific behaviour directly from the environment through actions and rewards.

2.1 Deep Learning for Signal and Image Processing

The first types of neural networks have been known for more than half a century. The multilayer perceptron was already in use in industrial applications like computer vision, automated translations and speech recognition for quite a time. But many problems, like vanishing gradients, would limit their capacity to learn when multiple layers are stacked, restricting them to very specific and expert questions. The first steps to overcome these limitations can be traced back to the 90s, and since then, slow but steady successes were accumulated until reaching a tipping point, facilitated by the Big Data.

Deep neural networks are now a reality, with spectacular results in many complex data analysis and pattern recognition tasks, particularly with unstructured data.

Text data are an interesting source of knowledge in industry, and deserve an article on their own. Images and sensors data value on the other hand is immediate to grasp by analogies to other fields. Initiatives in medical data analysis show really promising adoptions of deep neural networks. Images are used to diagnose precisely various pathologies [11]. EEG analysis are also a good example when it comes to learning patterns on complex combination of sensors/time-series. There are also interesting works in the speech and audio data, as well as on smartphone sensors (some recent examples [12,13]). An interesting initiative to notice is the approach that mimics how a doctor or a sound engineer use their eyes to analyse the results of an ECG or a spectrogram, by transforming sensors data to image data [14]. These approaches are also useful when it is necessary to fusion multiple sensors sources [15–17].

Deep Learning is usually supposed to require large volume of data to be able to capture the various levels of information from raw data to the final output. But this is mostly true when working with raw data (like raw pixel in pictures and images). But the algorithms can be applied on higher level of information. For instance, in the audio case, the spectrogram is fed as input instead of the raw audio file (amplitudes).

The volume being a major concern, many ongoing researches are focusing on new methods like transfer learning, to use the learned information from one case to another, sharing similar data, or few-shot/one-shot learning, that aims at learning from few instances, preferably one.

2.2 Machine Learning as a modelling tool

Using neural networks to simulate an industrial process isn't new, and one can easily find early examples in the

literature [18–20]. But the trend has clearly been accelerating lately, with many papers testing Machine Learning and Deep Learning models within a wide variety of fields (physics, chemistry, biology) [21–25]. The differences between statistical modelling and Machine Learning modelling were already extensively discussed, and a good reference can be found in [20].

One of the core feature of machine learning is the ability to capture complex behaviours and patterns, by learning them from within the data. The idea is to predict a value from the various available inputs. One can easily imagine complex combinations of signals from heterogeneous sources of data, like PLC, sensors and images to predict the target variable of any given industrial process, which is unimaginable with the classical approach.

This can prove useful when modelling complex processes in short times, or with limited resources, as the task of modelling from first principles at a sufficient precision level can be daunting, very long, and have a large cost. Even when a model is already available, it is usually built to capture the leading effects to understand the dynamic of the system, and there's always a trade-off between the time spent on building the model and the reached precision.

In the data-driven approach, the trade-off will be on the time and resources spent acquiring and storing the data, and the cost of the acquisition system is quickly paid back. If the system changes, there is no need to start the modelling process, as the learning algorithm will capture it from the new data.

3 The road to Smart Manufacturing

3.1 Data: a key resource

3.1.1 Data availability

Using data-driven methods requires data. As logical as it may sound, today this is still a major obstacle. To name a few:
- handwritten measurements are still very common on the shop floor, where many results are reported on paper cards or notebooks with a pen,
- complex proprietary machinery and protocols can be challenging to access, without the need of dedicated resources from the machines suppliers and manufacturers,
- low quality network can put some data beyond reach, or lead to packet loss. Wireless networks aren't always reliable in a factory. There are hopes that LPWAN will finally help bring stable wireless networking inside warehouses and factories. In the meantime, old-fashioned wired connections are still an interesting alternative to collect data from the production line.

3.1.2 Data quality

"Garbage in, garbage out" is a popular saying in the machine learning community. Having data is an

important milestone, but the quality of the data is crucial. Predictive algorithms will have difficulties distinguishing between signal and noise if the data is messy, and the measurement process isn't under control. Rigorous measurement procedures can only improve predictions quality. Especially if there's a human in the loop, with a process requiring a complex know-how. It goes beyond instruments conformance. Repeatability must be guaranteed, and reproducibility can also play a key role among different processes and/or factories, to ensure that the same models can be used across different sites.

The best one can expect is a successful predictive algorithm within the uncertainty range of the predicted variable. If the uncertainty is beyond the specifications tolerances, there is still a chance the algorithm can correctly identify the central tendency, but only if there are enough data, and no bias between the real value and the sample mean of the learning dataset (as expected from basic statistics).

3.2 Quick overviews of the possibilities

3.2.1 Control of complex industrial processes

Many industrial processes can be long and complex, with various steps leading to the final product. Each step builds on the previous ones, and having a clear and precise supervision on every combination of parameters can be tedious. Being able to predict at each step the expected output early can help to better control the whole process (see for instance [27,28]. All the following points are linked somehow to this one, and they all represent a complementary piece for a complete supervision of the production and the tools, with an expected increase in throughput and quality.

3.2.2 Virtual metrology for all

Virtual metrology is already used in the semiconductor industry [29,30] to speed up the manufacturing process, by predicting some properties that can be blocking if not known at the right time. There are even cases of classical neural network usage [31]. This can easily be generalized to any process and really deserve to be better known.

To make clear the difference with the previous paragraph, in this case, instead of predicting an output in the future, the idea is to capture the dependency of a value that is hard or long to measure (usually samples in a laboratory far from the production line) from other variables easier to access, on the same time horizon.

3.2.3 Speed-up quality control and inspection

Having fast algorithms that can validate quickly the conformance and/or safety of a product can greatly boost the production output, and dramatically reduce defects. Instead of randomly sampling, it is now possible to have a systematic algorithmic control, and predicting pieces

most likely to fail. A more thorough human control can then focus on the flagged cases, with an optimal allocation of human intelligence and know-how.

3.2.4 Easier image metrology

Numerous industrial processes rely on image analysis. Some quality controls in the previous paragraph can rely on microscopic imagery (metals, foams, etc). With new tools to analyse images at a level never known before, many scientific fields are experiencing a major shift in automated information and pattern extraction from images and videos. Once again, the medical field is interesting to look at, but many examples can be found for self-driving cars, spacecraft docking, etc [32–35].

The potential is nearly infinite when comparing to how many industrial processes require a visual human supervision.

3.2.5 Predictive maintenance

Many smart industry projects tend to focus on predictive maintenance. The advantage is that the value is clear, compared to a corrective maintenance that happens only once the failure has happened.

Predictive maintenance can be declined in diverse ways, but it isn't uncommon to have a misunderstanding of what it really is. For instance, having a planned schedule built from a global statistical analysis is hardly predictive. It doesn't account for daily operations impact per machine, and therefore doesn't optimize the maintenance operations.

Predictive maintenance should start with real-time monitoring. But this is not enough, and a common error is to condition maintenance operations based on fixed thresholds, calculated and fixed by hand. We are actually looking for a decision boundary, and machine learning based classification algorithms are built and designed specifically for this task. They are by far faster, and by definition, the decision boundaries they calculate are optimal for a given dataset.

Please note that as popular and straightforward this option seems, it is not necessarily the easiest one. Failures aren't so common in the data, and there's a risk that the algorithms will end up biased, potentially leading to many missed alarms.

3.3 From strategy to execution: methods and implementation

Having the tools of the trade is good, knowing how to use them is better. It is important to approach innovation pragmatically, otherwise the risk is to start with expectations that aren't aligned with the real potential of the tools, leading inevitably to failing projects and disillusion [36].

The examples above are presented in a generic manner, but each case can be implemented in numerous ways depending on the context, defined by the available data and their intrinsic quality, as well as operational and field constraints. Even small variations can lead to

substantial changes from one process implementation to another. And two factories from the same group, manufacturing the same product, can exhibit large variations.

Moreover, Big Data and AI have their own specificities that require an adequate handling. The first being the uncertainty due to the strong dependency on the available data and their quality (usable information within). By definition, an algorithm is supposed to have a deterministic behaviour, but predictive algorithms are different animals. Their behaviour is mainly dictated by the data used during the learning phase, meaning that the same algorithm can exhibit different behaviours for the same use-case, but in different contexts. Implementing a predictive approach requires to go through a knowledge funnel, from the initial question to the eventual answer, reducing uncertainty along the way. Methods exist to handle the associated risks.

3.3.1 Defining precisely the question

The first factor that plays a key role in the success or failure of a data project is the formal definition and setting of the question to be answered.

The design thinking framework [37] is a good start to help identify operational pain points that are good candidates, helping managers to remember that real problems come from the field, and favour creative thinking to find innovative solutions.

The basic principles are easy to understand, but somewhat subtle to put in place. The general workflow consists of alternating phases of creative and analytical thinking, twice. The problem space is explored first, then analysed to sort and filter the most interesting pain points, then the same is done for the solution space. Despite its simplicity, a well-conducted design thinking workshop can do wonders, and help generate novel ideas with a strong potential to grow as innovative solutions.

3.3.2 Validate rigorously the answer

Once a candidate use-case is identified, a rigorous validation process must be put in place. A project must start by considering an optimal trade-off between risk exposure and innovation, and demonstrate its value rapidly with a functional piece of software, referred to as the Minimal Viable Product (MVP) in the Lean Start-Up philosophy [39,40].

The latter draws inspiration from the Lean Manufacturing movement, in the sense that it recommends eliminating anything that doesn't contribute directly to the value demonstration.

Once validated, the MVP constitutes the basis upon which the developing team iterates and incrementally adds new features.

Iterative and incremental approaches were introduced to manage unpredictability in software engineering. Agile project management [38] is the typical archetype in a fluid business environment, where all the requirement can't be defined upfront. It usually consists of teams with transverse competencies

advancing in small steps, and validating each time a working piece of software with all the stakeholders (managers, field workers, final users, etc). This minimizes the risk of failures, by having regular feedback loops.

4 Synthesis

This paper provides a quick overview of the elements that should be considered when one wants to start experimenting with data, with a focus on industrial use-cases. The Big Data ecosystem comes as a technological layer exposing the raw data. Then the scientific tools, like Machine Learning and Deep Learning come into play to extract the value. This value is an actual answer to a real operational problem, and must validated using rigorous methods.

References

[1] J. Dean and S. Ghemawat, *Commun. ACM*, **51**, 107–113 (2008).

[2] S. Ghemawat, H. Gobioff, and S.-T. Leung, The Google file system, in: ACM SIGOPS Oper. Syst. Rev., ACM, (2003), pp. 29–43.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, *J. Mach. Learn. Res.*, **12**, 2825–2830 (2011).

[4] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in: Proc. 22nd Acm Sigkdd Int. Conf. Knowl. Discov. Data Min., ACM, (2016), pp. 785–794.

[5] C.-C. Chang and C.-J. Lin, *ACM Trans. Intell. Syst. Technol. TIST*, **2**, 27 (2011).

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, *ACM SIGKDD Explor. Newsl.*, **11**, 10–18 (2009).

[7] J. Bergstra, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, I. Goodfellow, A. Bergeron, Y. Bengio, and P. Kaelbling, (2011).

[8] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, and others, *ArXiv Prepr. ArXiv160304467*, (2016).

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proc. 22nd ACM Int. Conf. Multimed., ACM, (2014), pp. 675–678.

[10] F. Chollet and others, "Keras", (2015).

[11] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez, *ArXiv170205747 Cs*, (2017).

[12] C.A. Ronao and S.-B. Cho, *Expert Syst. Appl.*, **59**, 235–244 (2016).

[13] M.M.A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R.R. Yager, *Inf. Sci.*, **345**, 340–354 (2016).

[14] M.S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwicki, *ArXiv170101077 Cs*, (2017).

[15] A. Jain, H.S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, *ArXiv160100740 Cs*, (2016).

[16] F.J. Ordóñez and D. Roggen, *Sensors*, **16**, 115 (2016).

[17] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, *Front. Neurosci.*, **7**, (2013).

[18] V. Venkatasubramanian and K. Chan, *AIChE J.*, **35**, 1993–2002 (1989).

[19] S.A. Mandavgane, M.A. Siddiqui, A. Dubey, and S.L. Pandharipande, *Chem. Eng. World*, **39**, 75–77 (2004).

[20] K. Yetilmezsoy and S. Demirel, *J. Hazard. Mater.*, **153**, 1288–1300 (2008).

[21] M. Paganini, L. de Oliveira, and B. Nachman, *ArXiv170502355 Hep-Ex Physicshep-Ph Stat*, (2017).

[22] G.B. Goh, C. Siegel, A. Vishnu, N.O. Hodas, and N. Baker, *ArXiv170606689 Cs Stat*, (2017).

[23] Y. Li, W. Zhao, and J. Pan, *IEEE Trans. Autom. Sci. Eng.*, **14**, 1256–1264 (2017).

[24] M. Tehrani and M. Ahmadi, *SciRate*, (2017).

[25] Z. Li, S. Wang, Y. Yu, and J. Xu, *ArXiv170407207 Cs Q-Bio*, (2017).

[26] L. Breiman, *Stat. Sci.*, **16**, 199–231 (2001).

[27] N. Rana, Y. Zhang, D. Wall, and B. Dirahoui, Predictive data analytics and machine learning enabling metrology and process control for advanced node IC fabrication, in: 2015 26th Annu. SEMI Adv. Semicond. Manuf. Conf. ASMC, (2015), pp. 313–319.

[28] J. Moyne, J. Samantaray, and M. Armacost, Big data emergence in semiconductor manufacturing advanced process control, in: 2015 26th Annu. SEMI Adv. Semicond. Manuf. Conf. ASMC, (2015), pp. 130–135.

[29] Y.-J. Chang, Y. Kang, C.-L. Hsu, C.-T. Chang, and T.Y. Chan, Virtual Metrology Technique for Semiconductor Manufacturing, in: 2006 IEEE Int. Jt. Conf. Neural Netw. Proc., (2006), pp. 5289–5293.

[30] J.C. Yung-Cheng and F.-T. Cheng, Application development of virtual metrology in semiconductor industry, in: 31st Annu. Conf. IEEE Ind. Electron. Soc. 2005 IECON 2005, (2005), p. 6 pp.-.

[31] M.H. Hung, T.H. Lin, F.T. Cheng, and R.C. Lin, *IEEEASME Trans. Mechatron.*, **12**, 308–316 (2007).

[32] S. Singh and V. Kumar, *ArXiv160907597 Cs*, (2016).

[33] F. Liao, X. Chen, X. Hu, and S. Song, *ArXiv170203833 Cs*, (2017).

[34] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, *ArXiv170508260 Cs*, (2017).

[35] I.S. Fomin, A.V. Bakhshiev, and D.A. Gromoshinskii, *Procedia Comput. Sci.*, **103**, 59–66 (2017).

[36] J. Fenn and M. Raskino, "Mastering the hype cycle: how to choose the right innovation at the right time", Harvard Business Press, (2008).

[37] R.L. Martin, "The design of business: Why design thinking is the next competitive advantage", Harvard Business Press, (2009).

[38] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, and others, (2001).

[39] S. Blank, *Harv. Bus. Rev.*, **91**, 63–72 (2013).

[40] E. Ries, "The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses", Crown Books, (2011).