Why Industrial Machine Learning is Hard

Markus M. Geipel

mgeipel.github.io linkedin.com/in/markus-geipel

*

November 11, 2026

1 Industrial Machine Learning is Hard

Machine learning in industry is hard. I do not mean hard in the sense that it is computationally demanding or algorithmically challenging. Both may be the case, though. The point is that industrial problems have characteristics which void many strengths of modern machine learning methods that have been successfully applied in the web and internet domains. I will point these out and give examples. I will also suggest some techniques that make machine learning work in an industrial context.

2 The Challenges

What makes industrial machine learning hard? I argue four characteristics:

2.1 Scarcity of Labels and Domain Knowledge

Most machine learning is supervised learning. This means regression and classification. For classification, you need labels. For regression, we call them regressand, endogenous variable, response variable, measured variable, criterion variable, or dependent variable. Whatever you call them, you need lots of them, especially if you want to use deep learning methods. In the case of web data, we are lucky. There are countless ways to extract labels: Images on webpages have captions and often there is an alternative text field in the HTML. Instagram posts and tweets come with hashtags. Text paragraphs often have titles, keywords, and categories which may serve as labels. And if you do not have

* This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. https://creativecommons.org/licenses/by-nc-nd/4.0/

any of these, you can crowdsource the labelling. After all, most internet users can tell a dog from a cat. With machine data, you are in a much tougher spot, though: they do not come embedded in context and crowdsourcing may prove difficult.

Imagine letting an arbitrary crowd label the characteristic pressure patterns caused by cavitation in a Francis turbine. In case cavitation is not a household name to you, it is quite a fascinating and relevant phenomenon. I hope I have convinced you that labelling industrial data can be prohibitively costly due to the specific domain knowledge needed.



Figure 1: A ship's propeller cavitating. U.S. Navy, public domain. Cavitation is the formation and violent collapse of vapour bubbles in a liquid caused by rapid pressure drops, and it produces distinctive noise and pressure signatures as it erodes turbine blades [1].

[2]: Wired (2015), *How Cats Took Over the Internet*

2.2 Adversely Skewed Data

If you are trying to find cat pictures online, you are lucky: as Wired.com claims, "Everyone knows cats rule the Internet"[2]. If we look at how data is created online, this is no surprise. On the web, we have what the futurist Alvin Toffler called prosumers. Users who often act as producers as well as consumers. And they produce what they are themselves interested in. As it turns out, mostly cat pictures. This is a lucky situation if you want to do machine learning for the online crowd: There is lots of data on the topics they are interested in. Industry is vastly different! In factories, machines produce data. If you want to offer a machine learning-based product, you will soon realise that the most relevant data, such as data on machine failures and anomalies, is often severely underrepresented. Systems are designed to run failure-free for a long time, and parts are often changed before actual damage occurs.

2.3 Cost of Exploration and Data Generation

Many forms of machine learning, such as reinforcement learning or recommender systems, rely on exploration. Social media platforms, online shops, video and music streaming services, and news sites constantly explore which content to deliver to maximise customer attention. While exploration in these cases is not without the risk of alienating customers with inappropriate content, the benefits of exploration consistently outweigh the costs or potential damage. The cost structure in industry, however, is often quite different. Suboptimal operation of machines can incur excessive costs resulting from wear and tear, energy, and raw materials consumption, as well as potentially severe damage. This does not mean that exploration is impossible, but exploration is often limited in two ways: firstly, the amount of data that can be collected is limited, and secondly, the scope of the exploration is constrained due to safety limits.



Figure 2: A steam turbine rotor being assembled at a Siemens factory. Machines like this are complex, expensive, and unforgiving of reckless exploration. Photo: Christian Kuhna/Siemens Pressebild, via Wikimedia Commons, CC BY-SA 3.0.

2.4 Observability

In the virtual world, everything is already digital by default. Industrial applications exist in physical reality, though, and not

all system aspects are measured by default. Sensors are installed based on a cost-benefit analysis. Retrofitting additional sensors is often not an option, and some system aspects are effectively unobservable (e.g., combustion dynamics inside gas turbines). When working on industrial machine learning, we need to take this into account and always ask: do we record all data necessary to solve the given problem? If not, does the benefit of a machine learning solution outweigh the extra data acquisition costs? This can be a tough call!

3 Some Strategies to Tackle the Challenges

So, what can we as industrial AI engineers do? There are a few strategies:

3.1 Data-Efficient Algorithms

Data-efficient algorithms make the most of the data you have! If you do not have much data, data-hungry methods such as classical deep learning from scratch may not be the right choice. Especially in image processing, we can benefit from data augmentation and transfer learning, in particular, the reuse of models pretrained on more easily available data. Websites such as huggingface.co/models offer a wide choice of ready-to-use pretrained models. In all other cases, there is nothing wrong with trying classical statistics. If linear regression does the job, do use it. It is simple, robust, and interpretable; all quite desirable properties for an industrial application.

You might also want to look at Bayesian methods such as Gaussian processes. These methods handle a lack of data gracefully and provide you with a measure of uncertainty which is also quite useful in industrial settings.

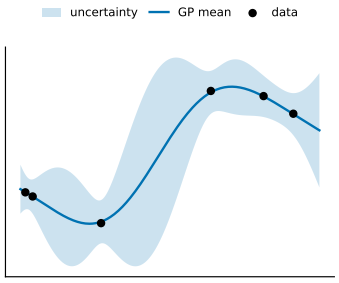


Figure 3: A Gaussian process fit to a handful of noisy measurements (dots). The shaded band is the model's uncertainty: it shrinks near data and grows in regions without observations.

3.2 Intelligence-Augmenting Algorithms

Industrial settings are often quite unforgiving because machine learning output feeds directly into the control of machines. In a web setting, it is often humans who receive the output of machine learning models. They then take the final decision: do I really want to buy the book my online bookstore recommends? Do I want to watch that video or listen to that recommended song? You can also leverage human intelligence in industrial setups! Think about AI systems that do not automate everything or directly control machines. Think about systems that help the human operators to do a better job. Build systems that augment human intelligence. Such a system could recommend a better operating mode to an engineer, warn an operator about anomalies, or make machine data easier to interpret. Keeping humans in the loop also leads to higher acceptance of AI solutions. Remember, the machine operators or engineers are key stakeholders and hold vital domain knowledge. They should always be your allies in improving the factory.

3.3 Symbolic AI Algorithms

In recent years, AI and deep learning have become almost synonymous. There is more to AI than machine learning, though! There are many symbolic approaches as opposed to purely data-driven ones. Symbolic approaches rely on explicitly formulated knowledge. This reduces the dependency on data and makes it easier to take existing domain knowledge into account. These approaches come in many forms and are sometimes so ubiquitous that we do not use the term AI anymore: heuristic search, for instance, is used to find the best route from A to B on a map. It can find much more and is a valuable tool to have in your toolbox. There is also the whole zoo of optimisation algorithms from discrete problems to continuous ones. Planning algorithms solve logistics problems and help robots to move and grasp without collisions. All of these can be applied in a much wider context and might just solve your industrial AI problem reliably without the need for data.

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

Figure 4: Some intellectual tasks, such as Sudoku, can be solved perfectly with classical AI algorithms and without any machine learning. The same is true for route planning and scheduling problems.

3.4 Simulations and Digital Twins

In some cases, a digital twin of the machine of interest exists. It can be a vital ally in the face of scarce data. Simulations can not only produce data for regression and classification tasks but also solve control optimisation problems via reinforcement learning. With a sufficiently accurate simulation at your disposal, deep learning is back in the game.

In some cases, you can even use machine learning to build a simulation model for particular aspects of interest, as it was done for gas turbines by Schäfer et al.[3]. Control policies trained on this model were later deployed successfully on real turbines.

[3]: Schaefer et al. (2007), *A Neural Reinforcement Learning Approach to Gas Turbine Control*

3.5 Interdisciplinary Teams

Besides algorithms and technology, there is an organisational aspect to industrial AI. Domain knowledge is vital, and domain experts are the keepers of this knowledge. For the duration of the project, they should be part of your data science or AI team. AI solutions are created in a build-measure-learn loop. The closer domain experts are to AI experts, the faster you can loop!

4 Final Thoughts

Machine learning in industry is hard, but there are many techniques that help you succeed. This extra challenge makes industrial AI interesting.

Bibliography

[1] Wikipedia contributors. *Cavitation*. 2026. URL: https://en.wikipedia.org/wiki/Cavitation (visited on 08/22/2026) (cited on page 2).

[2] Wired. *How Cats Took Over the Internet*. 2015. URL: https://www.wired.com/2015/08/how-cats-took-over-the-internet/ (visited on 08/22/2026) (cited on page 2).

[3] Anton Maximilian Schaefer et al. 'A Neural Reinforcement Learning Approach to Gas Turbine Control'. In: *2007 International Joint Conference on Neural Networks (IJCNN)*. 2007, pp. 1691–1696. DOI: 10.1109/IJCNN.2007.4371212 (cited on page 4).