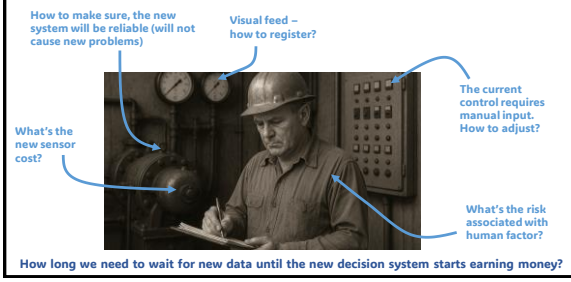
Mechatronic Engineering program

Basics of AI and Deep Learning:
13: Mechatronic Data Scientist: skillset and mindset, deployment of AI models

Ziemowit Dworakowski
AGH University of Krakow

1

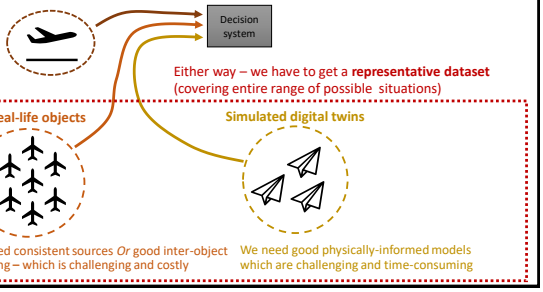
Data availability and practical biases



How long we need to wait for new data until the new decision system starts earning money?

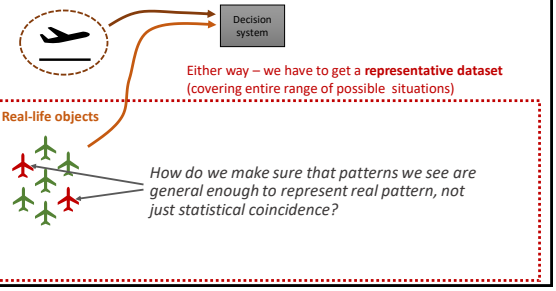
2

Data availability and practical biases



3

Data availability and practical biases

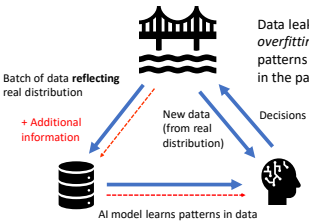


4

Data leakage

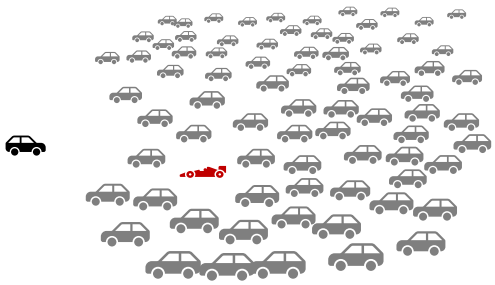
Data leakage occurs, when the decision system (*e.g. a classifier*) uses information during training and initial testing that will not be present in-operation.

Data leakage can be viewed also as a *„hidden overfitting"* phenomenon: A model learns patterns that are not general but present only in the particular setup of training data.



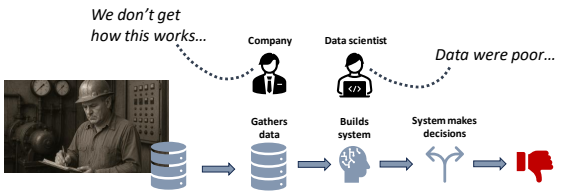
5

False positives and unlikely cases



6

Explainable AI –Who takes the blame?



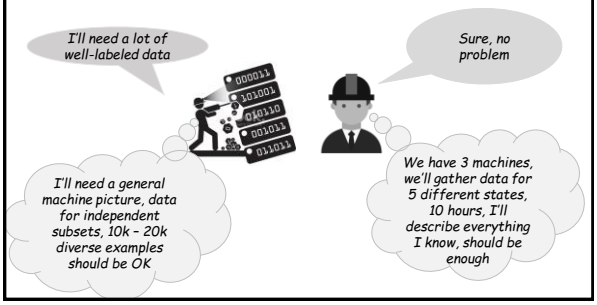
7

Why AI models tend to struggle in industrial environment?

- **They are often not reliable in extreme cases**
- **They are not intuitive, its hard to take responsibility**
- **Its hard to gather and label reliable data**
- **AI systems tend to weirdly malfunction**
  Usually because of data leakage
- **They are understood poorly**
  (Managers are rarely data scientists)

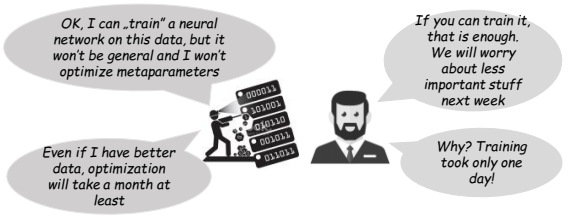
8

- Some terms are understood differently by industry and DS.



9

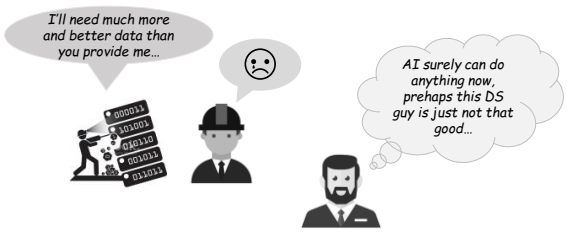
❖ Majority of industry practitioners **have no idea** which tasks are difficult



❖ Majority of industry practitioners cooperating with ML experts have poor understanding of **time required for particular tasks.** Typically they overestimate time for model setup and underestimate time for its configuration and fine-tuning.

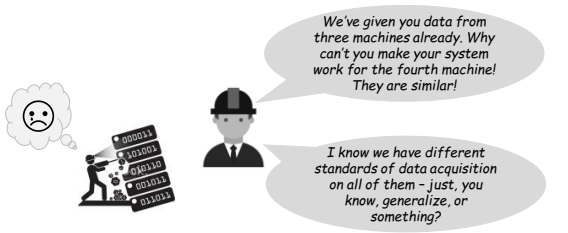
10

❖ Majority of people view ML systems as a „magic black box" that can do anything if the operator is good enough



11

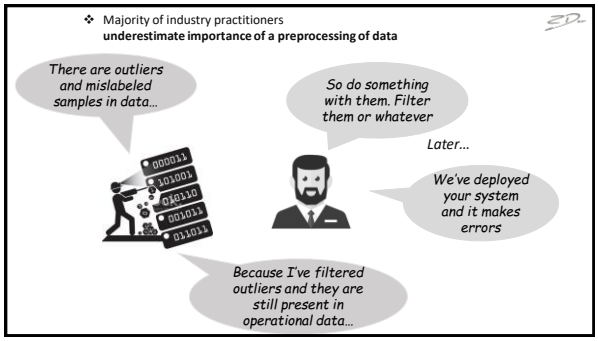
❖ Majority of industry practitioners **misjudge quality of available data**



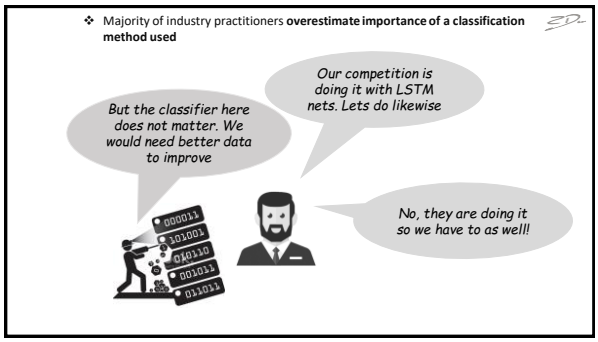
12


Majority of industry practitioners misjudge importance of overfitting and data leakage risks
We have one example of healthy machine and one example of damaged machine. Just check for differences and teach your net that way
You say our data are not independent. OK, whatever, train your net and lets worry about that later

13

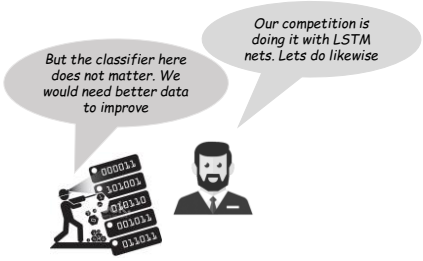

Majority of industry practitioners underestimate importance of a preprocessing of data
There are outliers and mislabeled samples in data…
So do something with them. Filter them or whatever
Later…
We've deployed your system and it makes errors
Because I've filtered outliers and they are still present in operational data…

14


Majority of industry practitioners overestimate importance of a classification method used
Our competition is doing it with LSTM nets. Lets do likewise
But the classifier here does not matter. We would need better data to improve
No, they are doing it so we have to as well!

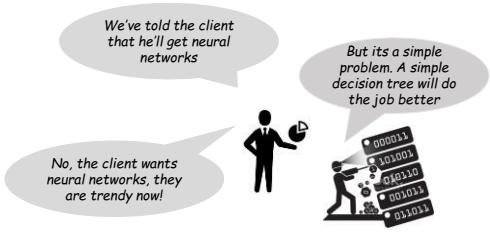
15

- Majority of industry practitioners **overestimate importance of a classification method used**



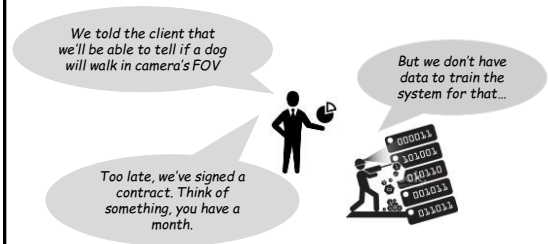
16

- In industry there is a high pressure towards „marketing effect" and „selling dreams" instead of real quality



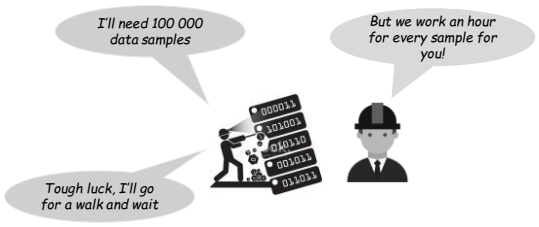
17

- In industry there is a high pressure towards „marketing effect" and „selling dreams" instead of real quality



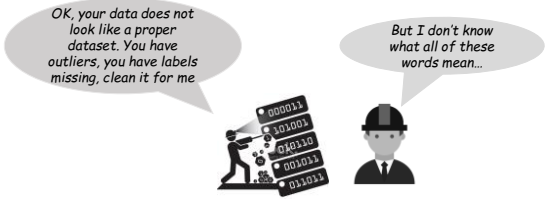
18

- Majority of ML experts do not understand how difficult data acquisition could be:



19

- Majority of ML experts are trained only on „easy" and „clean" data, with no outliers or mislabeled samples.

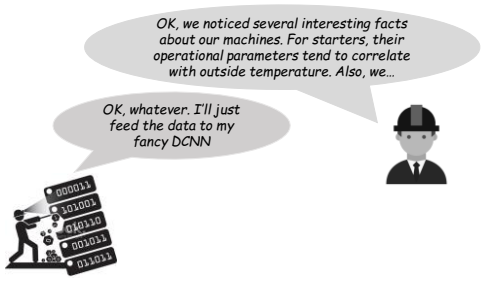


20

- ML experts like to take it the easy way, feeding vectors of data to a decision system and demand more data if it does not converge to a good result, instead of picking **better features** and use **context knowledge**

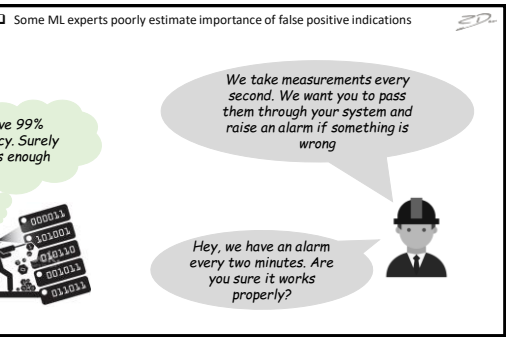
21

- Majority of ML experts underestimate context knowledge and overestimate the classifier capabilities



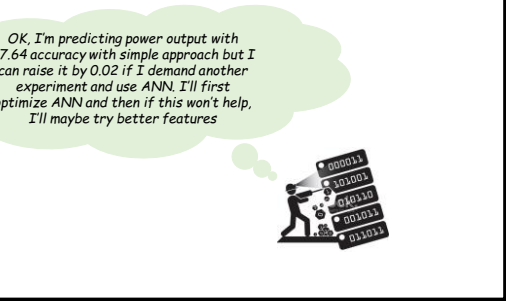
22

- Some ML experts poorly estimate importance of false positive indications



23

- Some ML experts poorly identify factors contributing to the final result



24

Required mindset

Be prepared to work with poorly prepared data
(*including mislabeled samples, outliers, lack of description etc.*)

Be prepared to compromise and think non-standard
(Some problems can't be solved using typical approaches and good practices)

You need a credit of trust
(can be gained based on experience and communication skills)

Protect salesmen from themselves
(*„Would you like a merit check on your leaflet?"*)

Think about the actual goal of your work
(It is rarely „to build a classifier")

25

Required skillset

Basic engineering knowledge
to understand context and meaning of the task at hand

Good understanding of **data leakage and overfitting** problems
these are usually the most important challenges that need to be solved in the first place. We've got a course for that!

Ability to learn and find sources
(*We usually work with unique applications every time*)

Basics of AI
General overview rather than deep understanding of methods' details

Experience from several projects with emphasis on **design** and **optimization** of the solution

Python (+ its AI libraries)

26

Gaps in YOUR knowledge

Only a handful of models
We covered only examples in a few categories…

Big data
When there are thousands of features and millions of examples… but there's a course for that in MSc program…

Lack of engineering statistics
Its not the point of this course, but you need to learn it, seriously!

Deep AI was only mentioned here…
And also not in full. No LSTMs, no advanced deep models for labeling, etc.

AI control and structural modeling
(There's a course for that in MSc program)

Python
(yes, its a gap, obviously… But there's a course for that in 7th semester…)

27

For repetition…

1) Explain what limits practical applicability of AI methods in industry (data leakage, unlikely cases management, explainability and representative datasets)

2) Explain basic mistakes in approach between data science and industry

28