

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 *SD*

How to make sure, the new system will be reliable (will not cause new problems)

Visual feed – how to register?

What's the new sensor cost?

The current control requires manual input. How to adjust?

What's the risk associated with human factor?

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

2

Data availability and practical biases *SD*

Decision system

Either way – we have to get a **representative dataset** (covering entire range of possible situations)

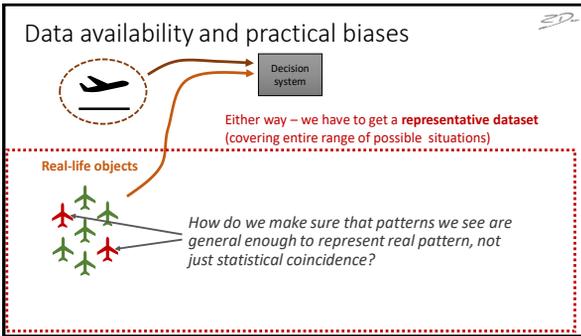
Real-life objects

Simulated digital twins

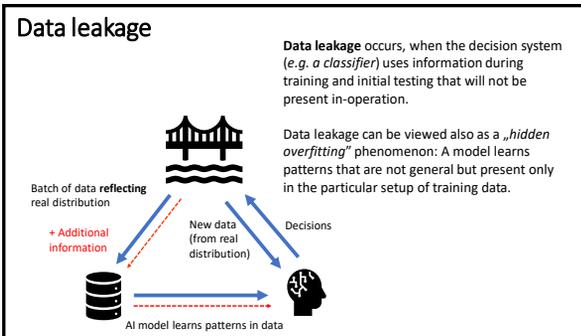
We need consistent sources Or good inter-object mapping – which is challenging and costly

We need good physically-informed models which are challenging and time-consuming

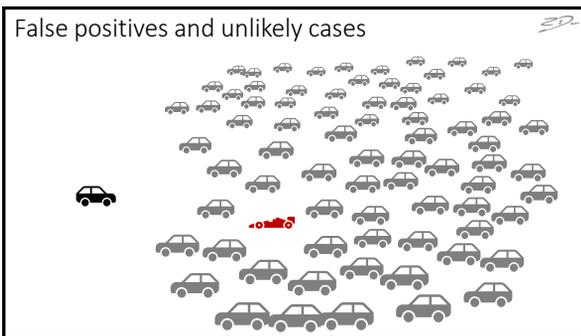
3



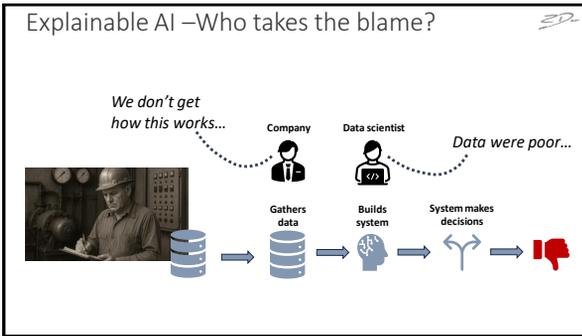
4



5



6



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.

Worker: I'll need a lot of well-labeled data

Worker: I'll need a general machine picture, data for independent subsets, 10k - 20k diverse examples should be OK

Data Scientist: Sure, no problem

Data Scientist: We have 3 machines, we'll gather data for 5 different states, 10 hours, I'll describe everything I know, should be enough

9

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

OK, I can "train" a neural network on this data, but it won't be general and I won't optimize metaparameters

Even if I have better data, optimization will take a month at least

If you can train it, that is enough. We will worry about less important stuff next week

Why? Training took only one day!

❖ 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 SD

I'll need much more and better data than you provide me...

AI surely can do anything now, perhaps this DS guy is just not that good...

11

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

We've given you data from three machines already. Why can't you make your system work for the fourth machine! They are similar!

I know we have different standards of data acquisition on all of them - just, you know, generalize, or something?

12

❖ Majority of industry practitioners misjudge importance of overfitting and data leakage risks SD

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 SD

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 SD

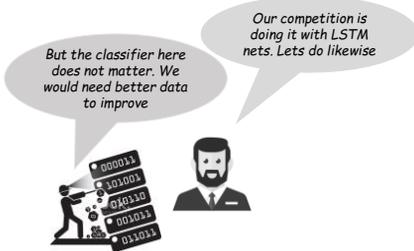
But the classifier here does not matter. We would need better data to improve

Our competition is doing it with LSTM nets. Lets do likewise

No, they are doing it so we have to as well!

15

❖ Majority of industry practitioners overestimate importance of a classification method used ED

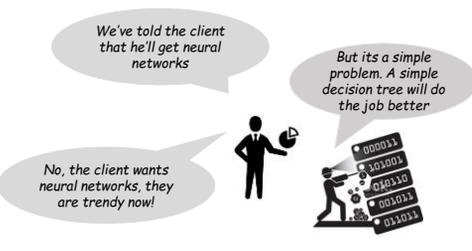


Our competition is doing it with LSTM nets. Lets do likewise

But the classifier here does not matter. We would need better data to improve

16

❖ In industry there is a high pressure towards „marketing effect“ and „selling dreams“ instead of real quality ED



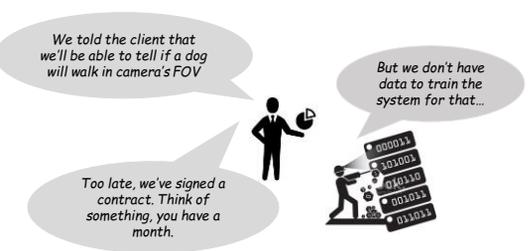
We've told the client that he'll get neural networks

But its a simple problem. A simple decision tree will do the job better

No, the client wants neural networks, they are trendy now!

17

❖ In industry there is a high pressure towards „marketing effect“ and „selling dreams“ instead of real quality ED



We told the client that we'll be able to tell if a dog will walk in camera's FOV

But we don't have data to train the system for that...

Too late, we've signed a contract. Think of something, you have a month.

18

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

I'll need 100 000 data samples

But we work an hour for every sample for you!

Tough luck, I'll go for a walk and wait

19

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

OK, your data does not look like a proper dataset. You have outliers, you have labels missing, clean it for me

But I don't know what all of these words mean...

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**

Dataset was too small. Gather 10x more next time

21

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

OK, we noticed several interesting facts about our machines. For starters, their operational parameters tend to correlate with outside temperature. Also, we...

OK, whatever. I'll just feed the data to my fancy DCNN

22

□ Some ML experts poorly estimate importance of false positive indications ED

I have 99% accuracy. Surely this is enough

We take measurements every second. We want you to pass them through your system and raise an alarm if something is wrong

Hey, we have an alarm every two minutes. Are you sure it works properly?

23

□ Some ML experts poorly identify factors contributing to the final result ED

OK, I'm predicting power output with 87.64 accuracy with simple approach but I can raise it by 0.02 if I demand another experiment and use ANN. I'll first optimize ANN and then if this won't help, I'll maybe try better features

24

Required mindset SD

<p> Be prepared to work with poorly prepared data <i>(including mislabeled samples, outliers, lack of description etc.)</i></p>	<p> Be prepared to compromise and think non-standard <i>(Some problems can't be solved using typical approaches and good practices)</i></p>
<p> You need a credit of trust <i>(can be gained based on experience and communication skills)</i></p>	<p> Protect salesmen from themselves <i>(„Would you like a merit check on your leaflet?“)</i></p>
<p> Think about the actual goal of your work <i>(It is rarely „to build a classifier“)</i></p>	

25

Required skillset SD

<p> Basic engineering knowledge <i>to understand context and meaning of the task at hand</i></p>	<p> Good understanding of data leakage and overfitting problems <i>these are usually the most important challenges that need to be solved in the first place. We've got a course for that!</i></p>
<p> Ability to learn and find sources <i>(We usually work with unique applications every time)</i></p>	<p> Basics of AI <i>General overview rather than deep understanding of methods' details</i></p>
<p> Experience from several projects with emphasis on design and optimization of the solution</p>	<p> Python (+ its AI libraries)</p>

26

Gaps in YOUR knowledge SD

<p> Only a handful of models <i>We covered only examples in a few categories...</i></p>	<p> Big data <i>When there are thousands of features and millions of examples... but there's a course for that in MSc program...</i></p>
<p> Lack of engineering statistics <i>Its not the point of this course, but you need to learn it, seriously!</i></p>	<p> Deep AI was only mentioned here... <i>And also not in full. No LSTMs, no advanced deep models for labeling, etc.</i></p>
<p> AI control and structural modeling <i>(There's a course for that in MSc program)</i></p>	<p> Python <i>(yes, its a gap, obviously... But there's a course for that in 7th semester...)</i></p>

27

For repetition...

SD

- 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
