4 Data-Centric AI

The next big step

“The model and the code for many applications are basically a solved problem. Now that the models have advanced to a certain point, we got to make the data work as well.”

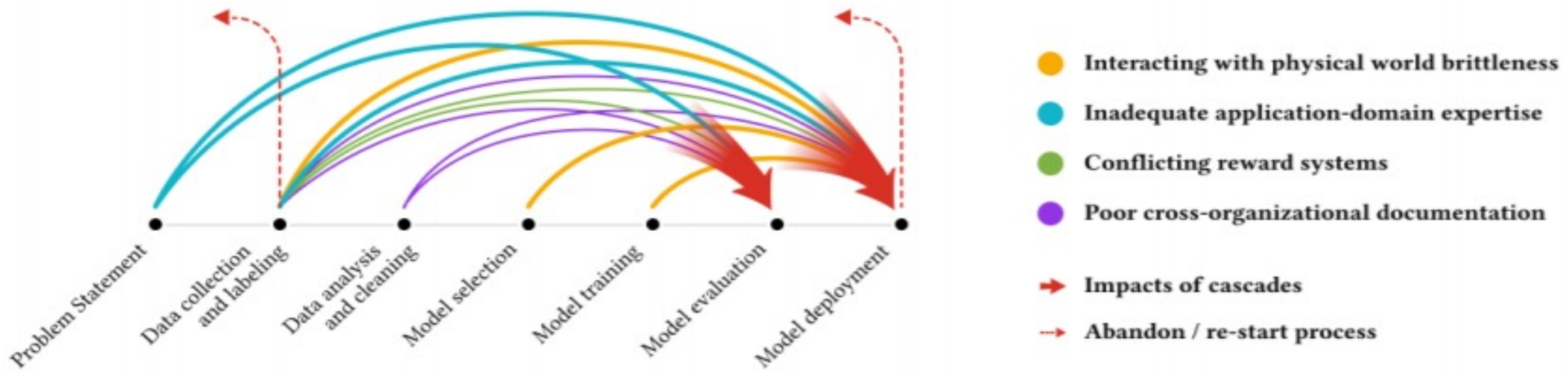
“Many data scientists have their own ways to clean data but what we don’t have is a systematic mental framework for doing it”

“Just like the rise of deep learning a decade ago spawned tons of new jobs, I hope that data-centric AI development will spawn tons of new jobs in many industries.”

-Andrew Ng

Data Cascades

Data Cascades are compounding events causing negative, downstream effects from data issues, that result in technical debt over time


Problem Statement
Data collection and labeling
Data analysis and cleaning
Model selection
Model training
Model evaluation
Model deployment
● Interacting with physical world brittleness
● Inadequate application-domain expertise
● Conflicting reward systems
● Poor cross-organizational documentation
Impacts of cascades
Abandon / re-start process

Sambasivan et al, 2021

From Big data to Good data

The rise of Big data allows companies to extract value from AI.

Models improved, but what if we focus on improving the data instead?

What is good data?

- Defined consistently
- Cover of important cases
- Has timely feedback from production data
- Sized appropriately

Model-Centric status quo

Since the Big Data revolution, Data Scientists focused on improving models and code as much as possible, rather than the data.

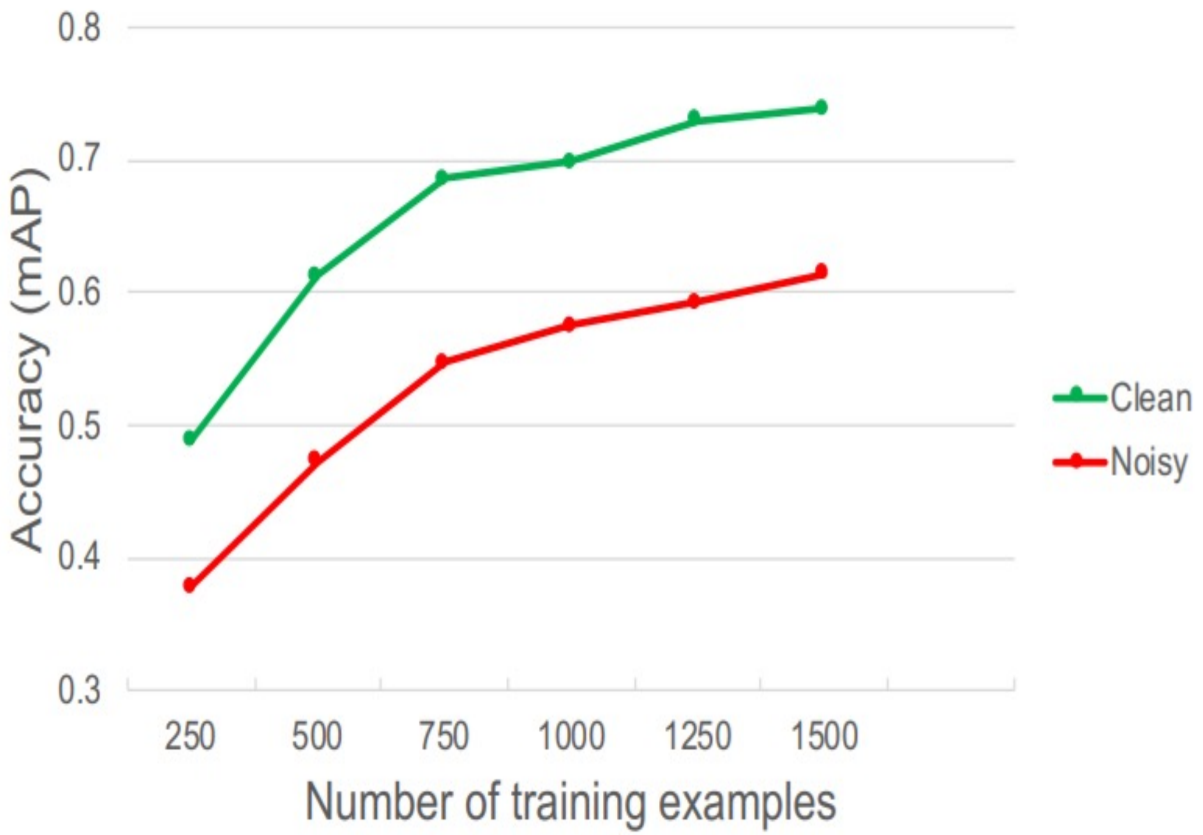
This is expected

- Industry follows academic development closely
- Benchmark data sets are used so development focus on models
- Open source culture has made model improvement accessible

This is not necessarily a bad thing

- Following this approach AI has made tremendous progress

Data-Centric approach



Aim of the data-centric approach is to make data quality a systematic issue

Why?

- Improving the data can greatly improv the quality
- Maintainability
- Ease of deployment

Model-Centric vs Data-Centric

AI = Code (and Model) + Data

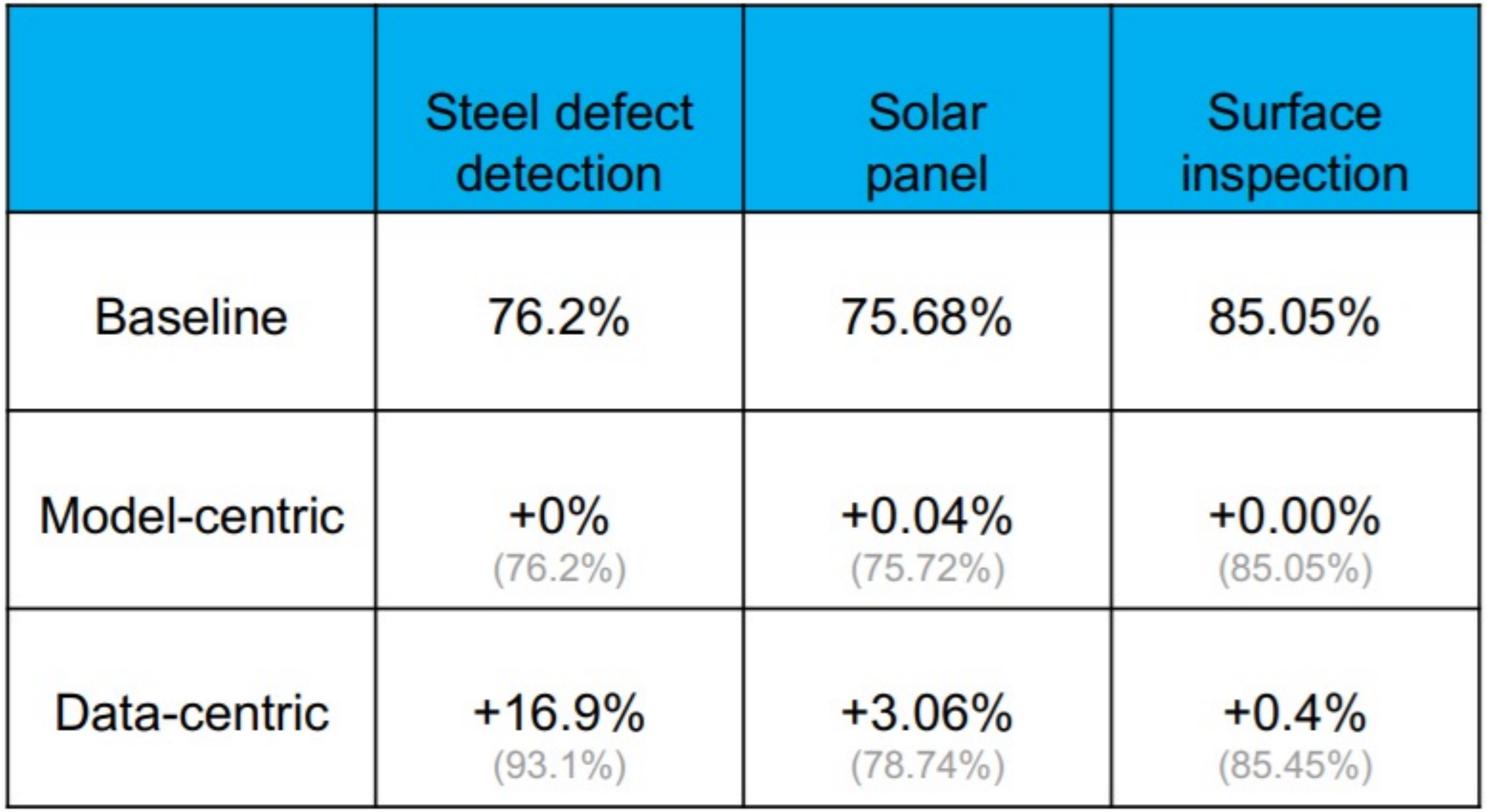
Model-Centric

- Collect all the data you can
- Develop a model good enough to deal with that data
- Hold data fixed and iteratively improve the model

Data-Centric

- Consistency and quality of the data is paramount.
- Allows multiple models to do well
- Hold the code fixed and improve the data

Model-Centric vs Data-Centric



	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

Takeaways

Systematic improvements

- For this approach to work, it should be and efficient and systematic process

High quality data

- Important to guarantee high quality data in the whole project life cycle

Is time to improve our approach

- The model-centric approach has taken us very far but we can still go further

Investing in data pays off

- Yet another reason why having a solid data strategy helps to improve results

A new systematic framework

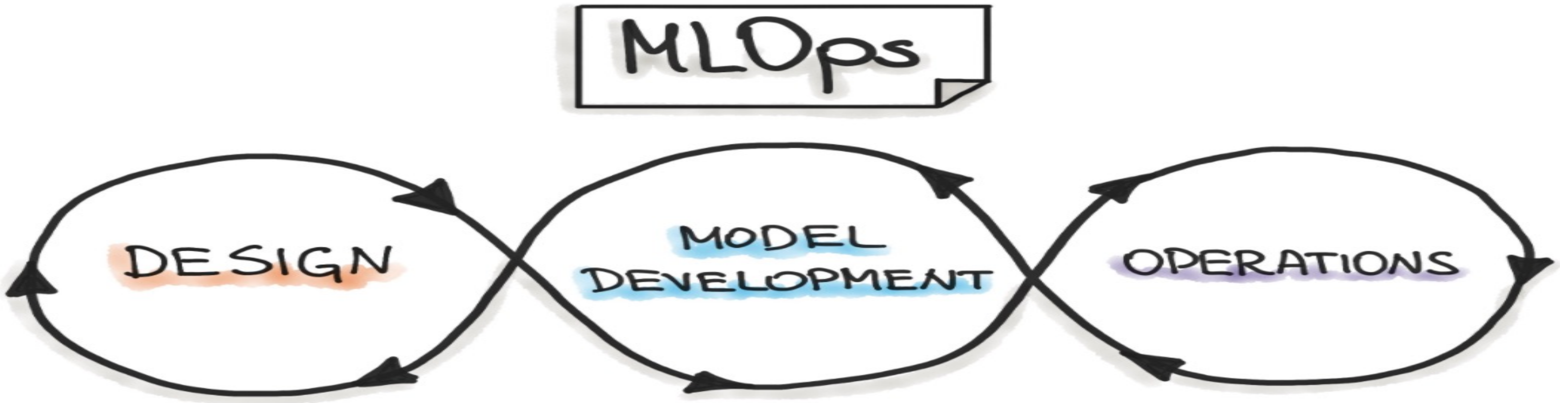
We know

Making data quality systematic improves general performance

But…What if?

We use the same principle for the whole project lifecycle

Enters…



PROPHECY LABS

ctg

nannyML

AI4Business

ESF
INVESTEERT IN
JOUW TOEKOMST

Europese Unie

Vlaanderen
is werk