

The Data-Centric Development Process for AI in Industry

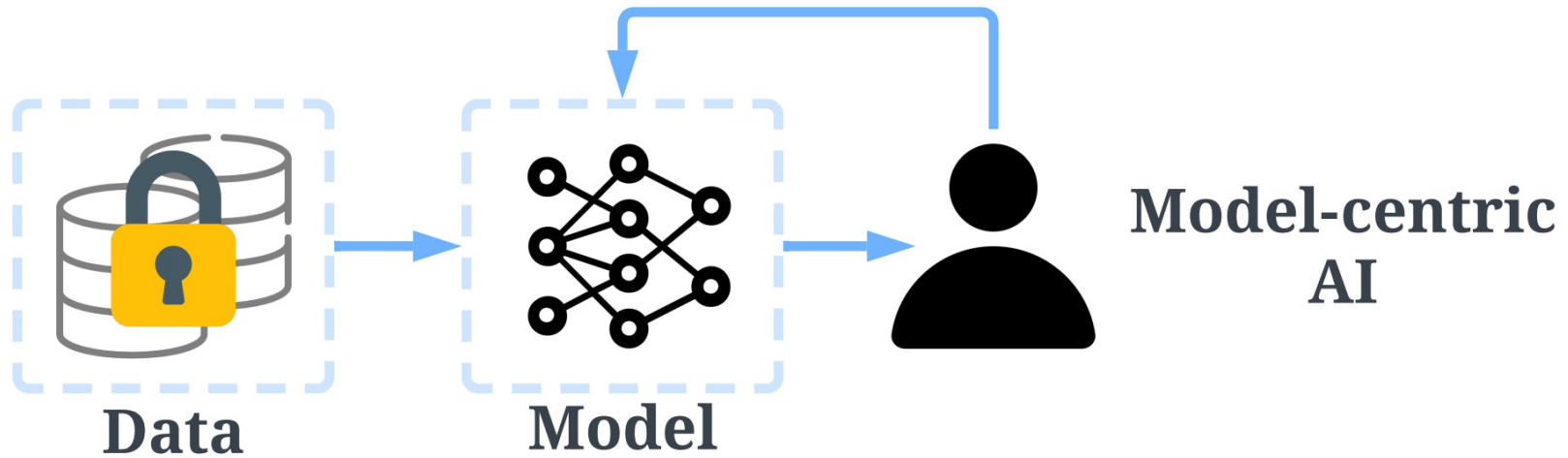
SDS2023, June 23, 2023

Paul-Philipp Luley
Gerrit A. Schatte



Source: <https://xkcd.com/1838/>

The Prevalent Approach: Model-Centric AI

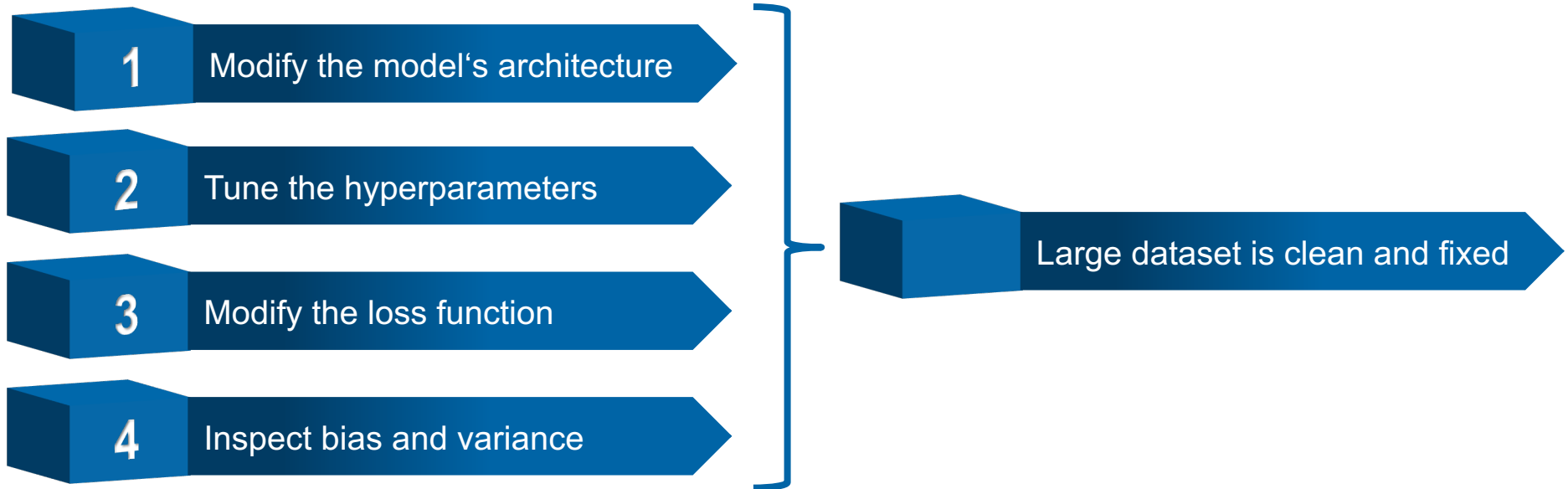


Source: Zha et al., «Data-centric AI: Perspectives and Challenges», 2023





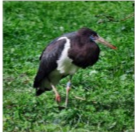
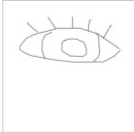







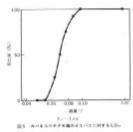
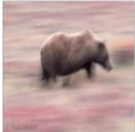





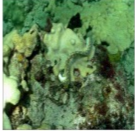

The Prevalent Approach: Model-Centric AI

- 1 Led to important advances
- 2 Scientific competition
- 3 We are taught this way

Model-Centric Development

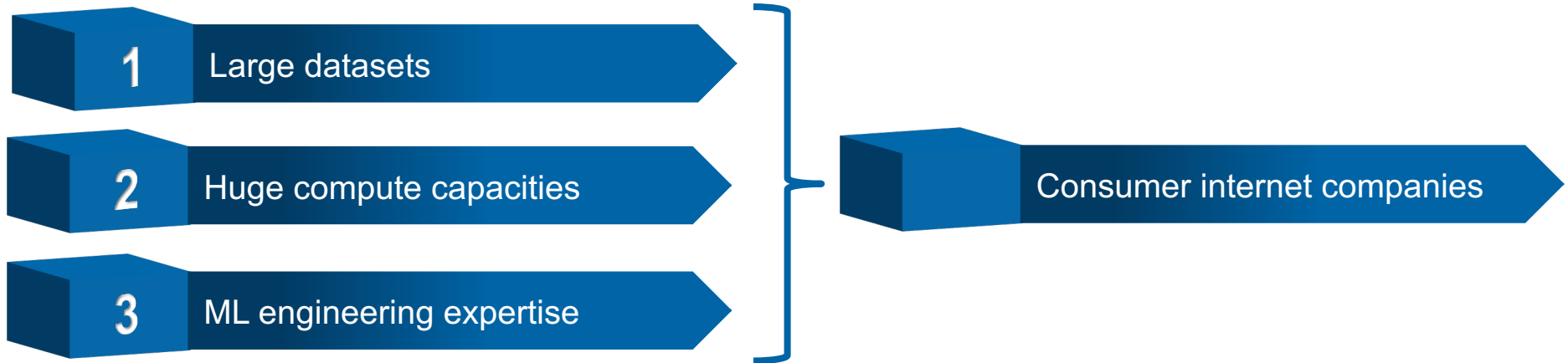


Model-Centric AI - Label Errors

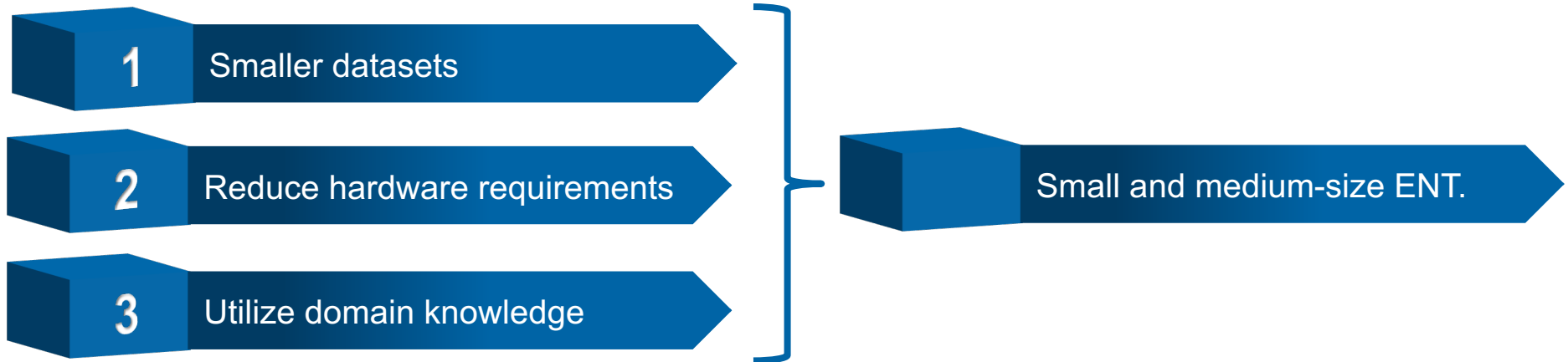
| | MNIST | CIFAR-10 | CIFAR-100 | Caltech-256 | ImageNet | QuickDraw |
|---------------|---|---|--|---|--|--|
| correctable |  given: 8 corrected: 9 |  given: cat corrected: frog |  given: lobster corrected: crab |  given: dolphin corrected: kayak |  given: white stork corrected: black stork |  given: tiger corrected: eye |
| multi-label | (N/A) | (N/A) |  given: hamster also: cup |  given: laptop also: people |  given: mantis also: fence |  given: wristwatch also: hand |
| neither |  given: 6 alt: 1 |  given: deer alt: bird |  given: rose alt: apple |  given: house-fly alt: ladder |  given: polar bear alt: elephant |  given: pineapple alt: raccoon |
| non-agreement |  given: 4 alt: 9 |  given: automobile alt: airplane |  given: dolphin alt: ray |  given: yo-yo alt: frisbee |  given: eel alt: flatworm |  given: bandage alt: roller coaster |

Source: Northcutt et al., «Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks», 2021

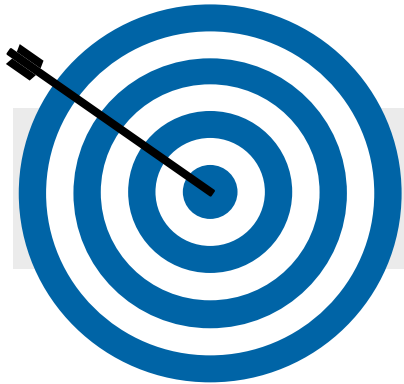
Model-Centric Development



Widespread AI Adoption



From Concept to Implementation: Data-Centric AI for Industry



Data-Centric AI.



Definition



Properties

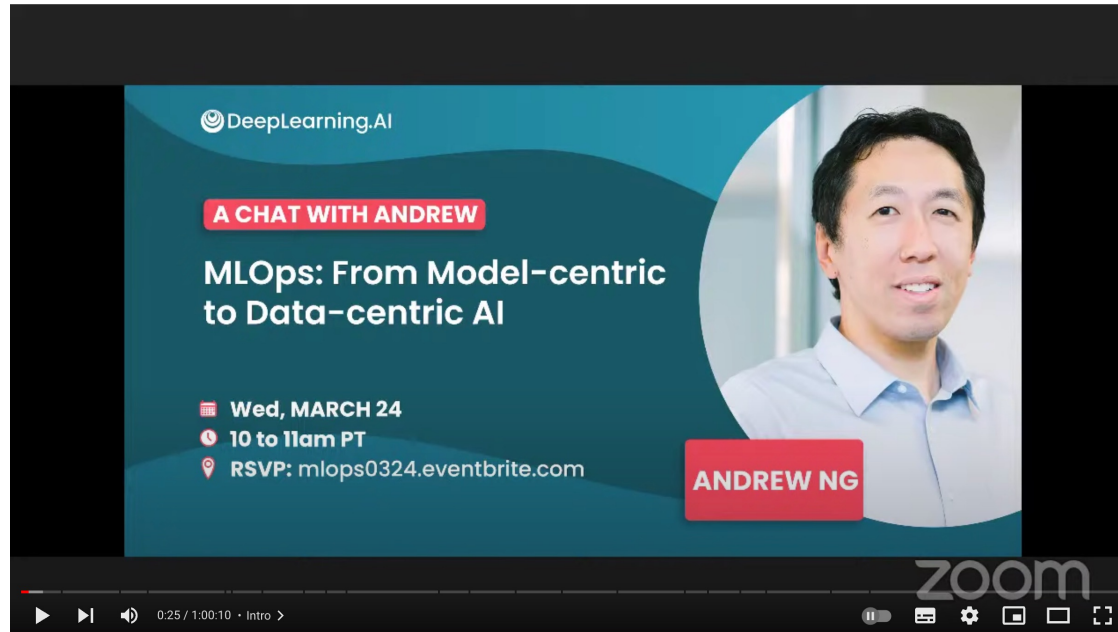


Implementation

Data-Centric AI - Origin

YouTube

Suchen

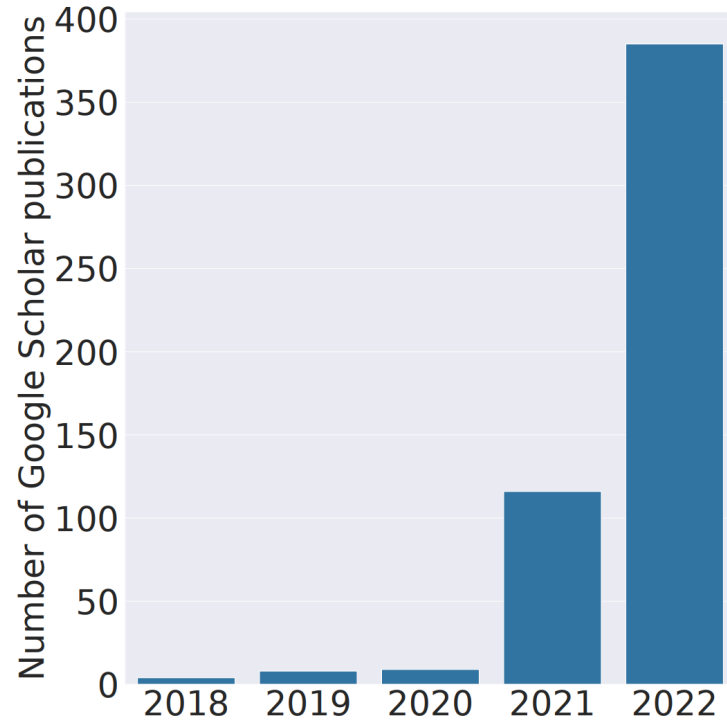


A Chat with Andrew on MLOps: From Model-centric to Data-centric AI



Source: <https://www.youtube.com/watch?v=06-AZXmWtHjo>

Data-Centric AI - Evolution



Source: Zha et al., Data-centric AI: Perspectives and Challenges, 2023

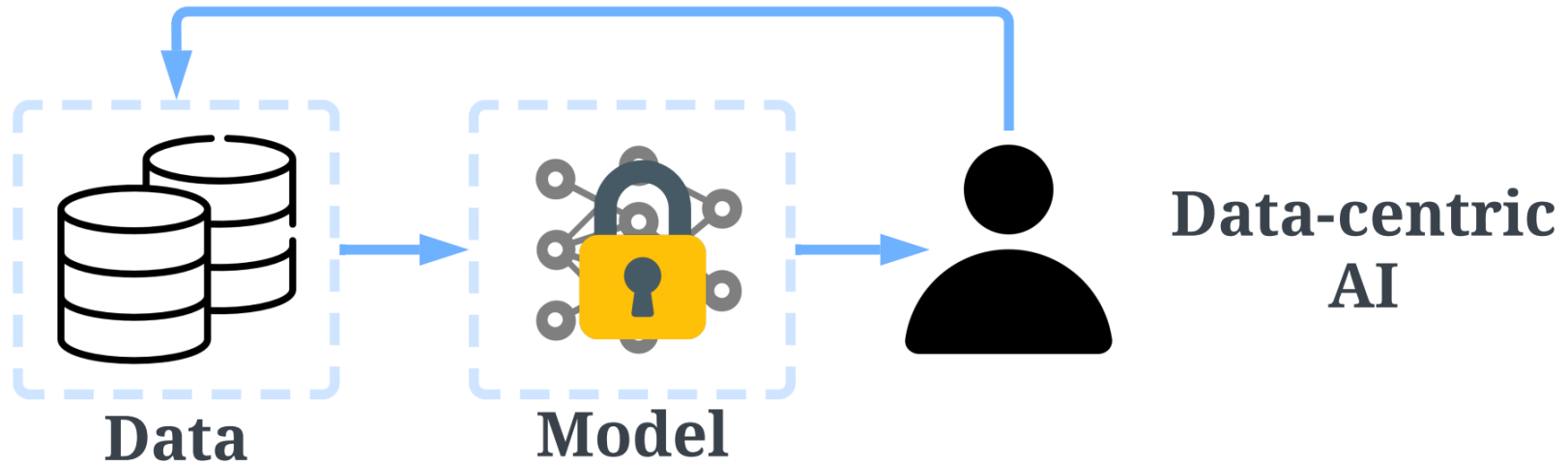
- 1** DCAI research accelerates
Examples: ZHAW, ETH, Fraunhofer
- 2** DCAI education starts
Examples: NeurIPS workshop, MIT
- 3** DCAI companies grow [3] [4]
Examples: Kistler, Landing AI, Snorkel,

What is Data-centric AI?

Data-centric AI is the discipline of systematically engineering the data used to build an AI system.

Source: <https://datacentricai.org/>

Data-Centric AI - Definition

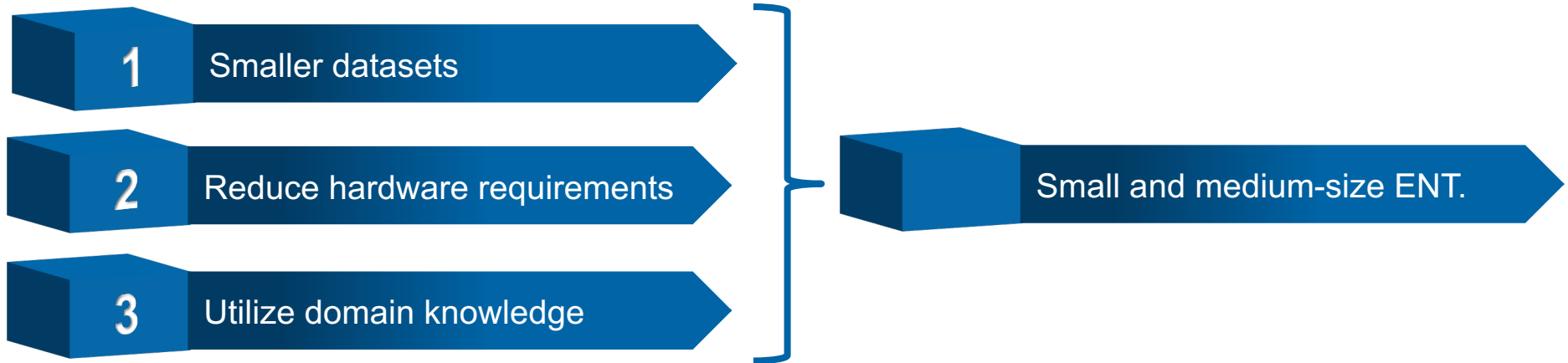


Source: Zha et al., «Data-centric AI: Perspectives and Challenges», 2023

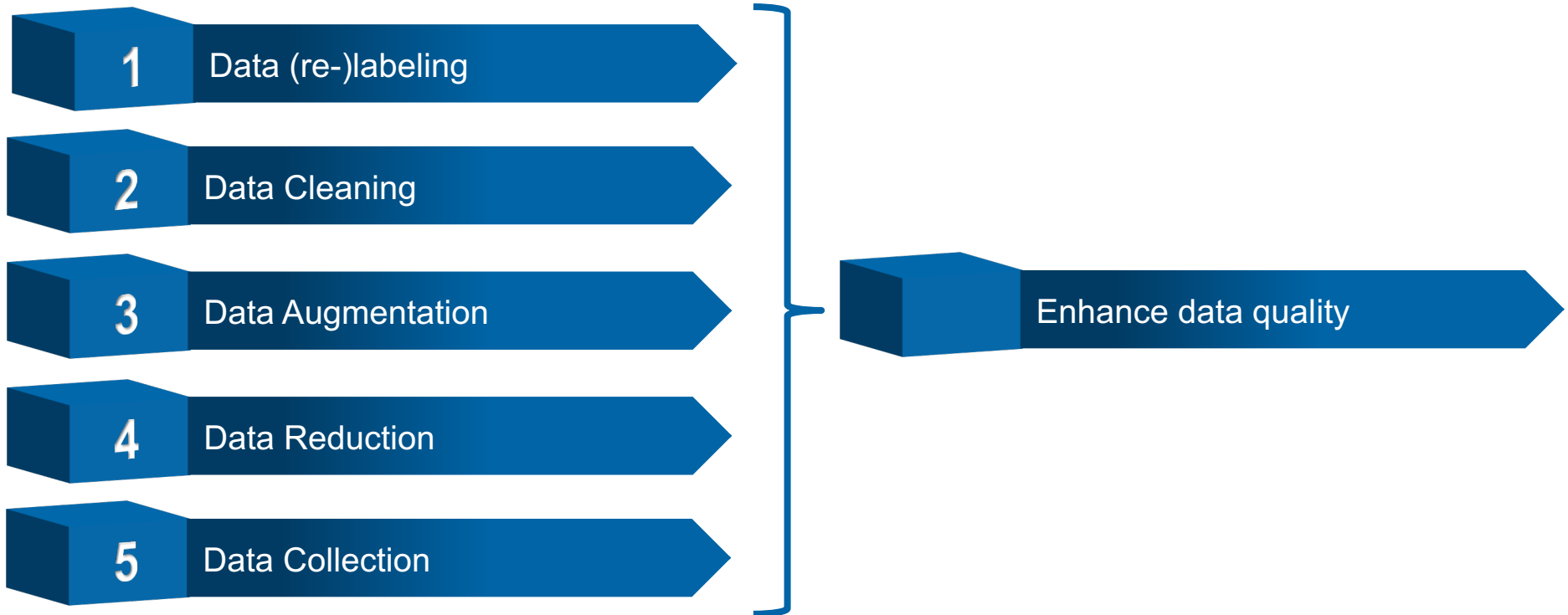
Data-Centric AI - Properties

- 1 Consistent labels
- 2 Representative, high-quality data
- 3 Detect concept & data drift

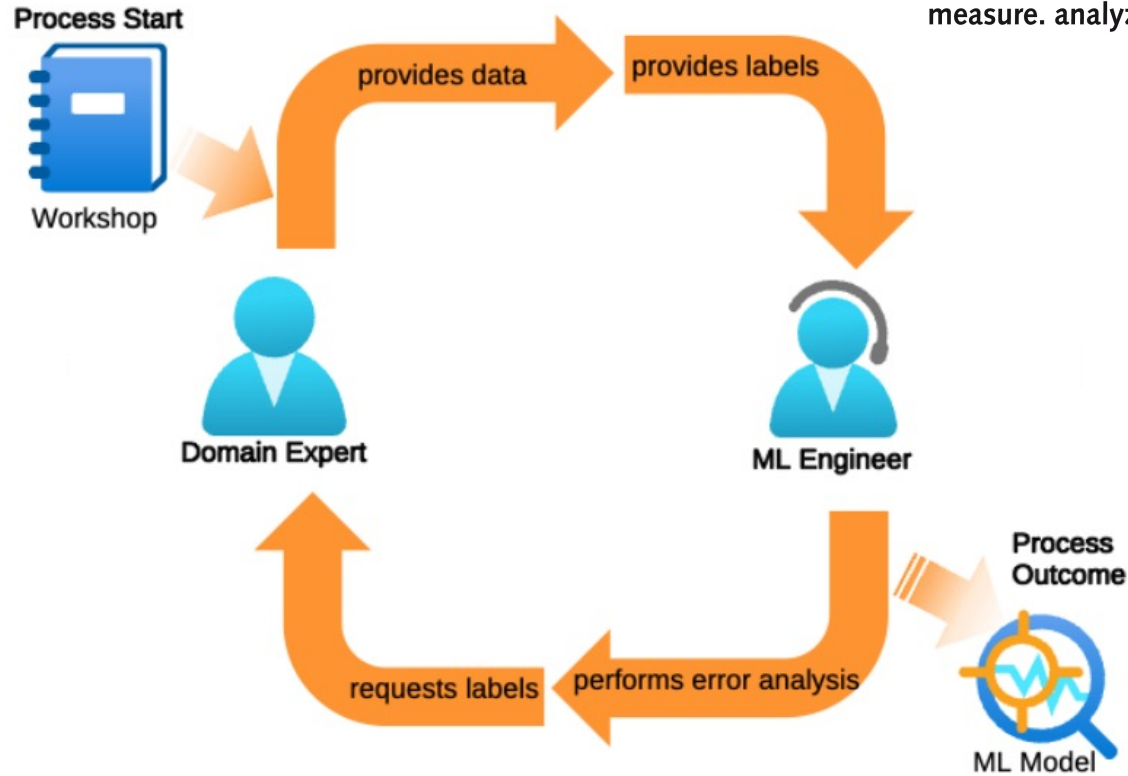
Widespread AI Adoption



Data-Centric AI – Engineering the Data



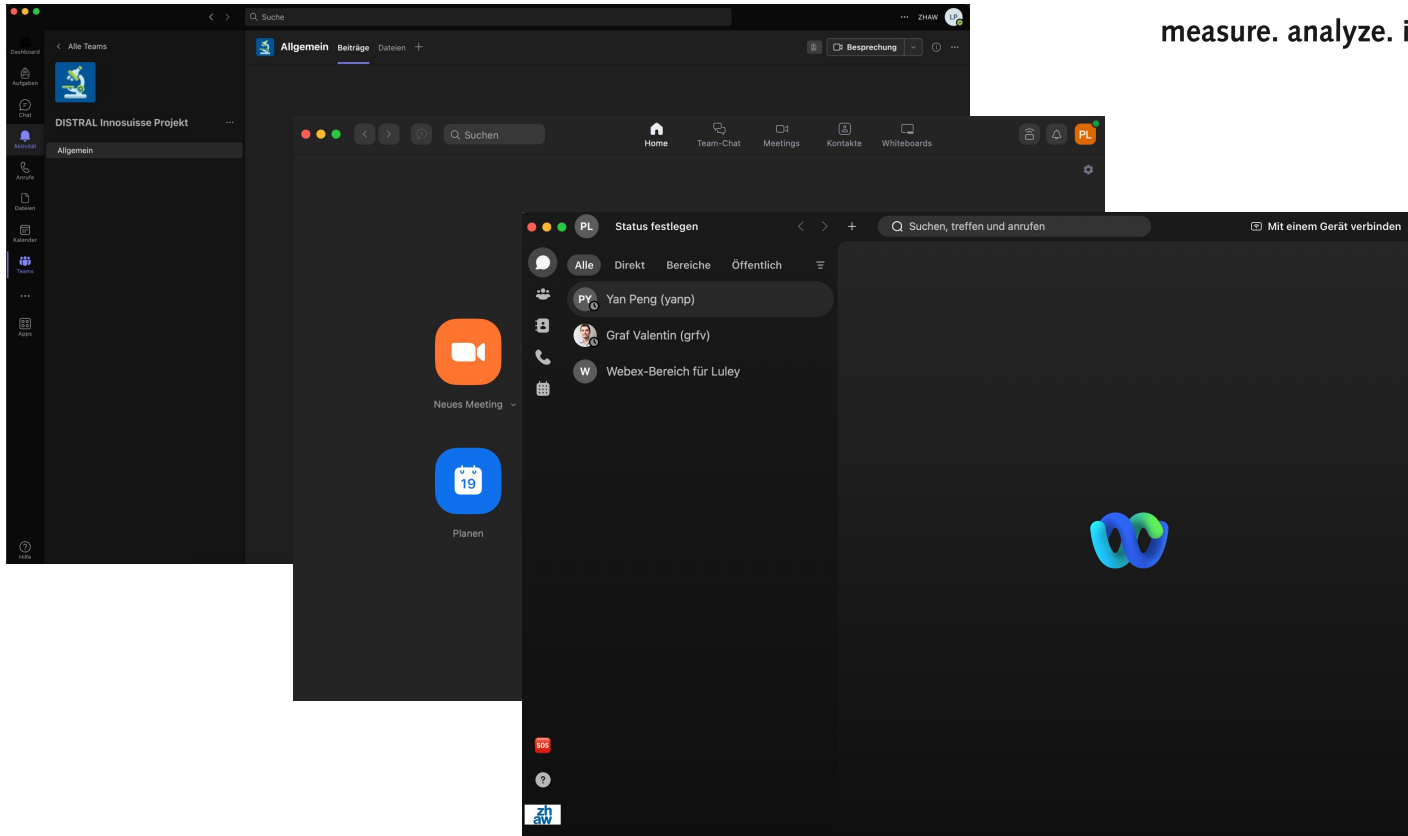
The Data-Centric Development Process



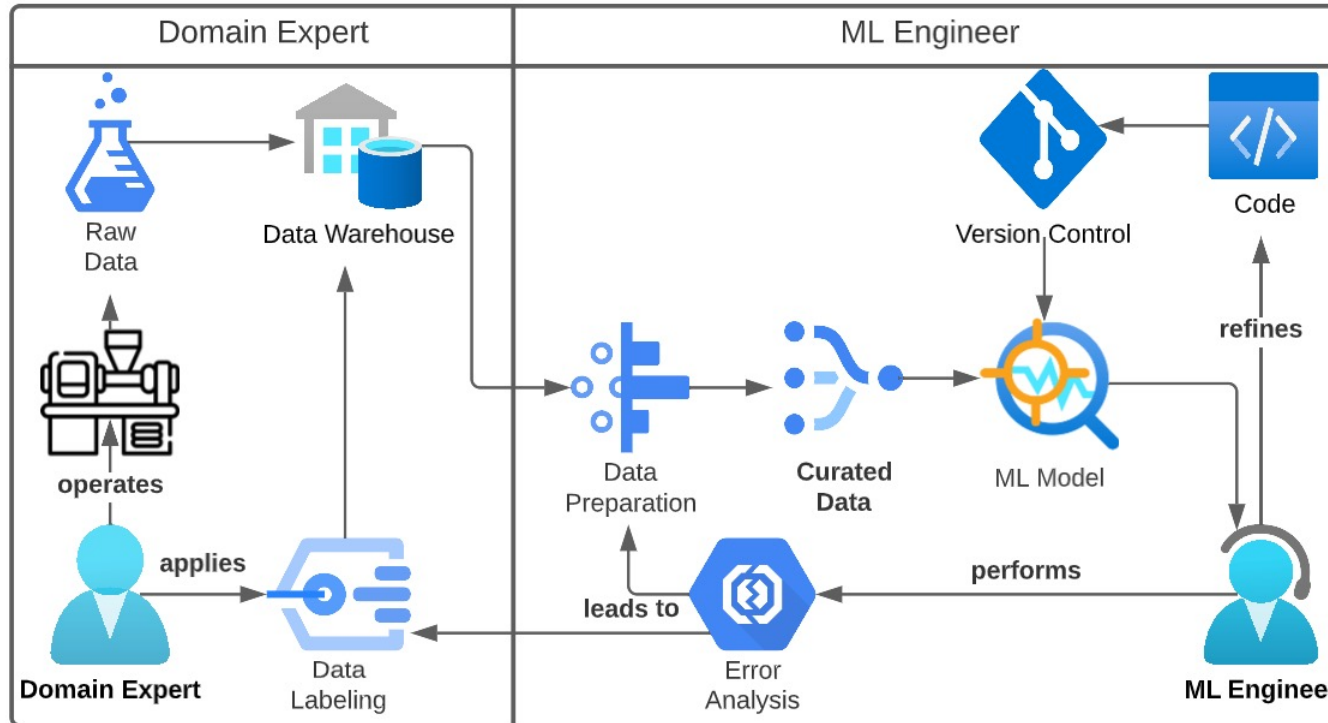
Source: Luley et al., «From Concept to Implementation: The Data-Centric Development Process for AI in Industry», 2023

The Data-Centric Development Process

KISTLER
measure. analyze. innovate.

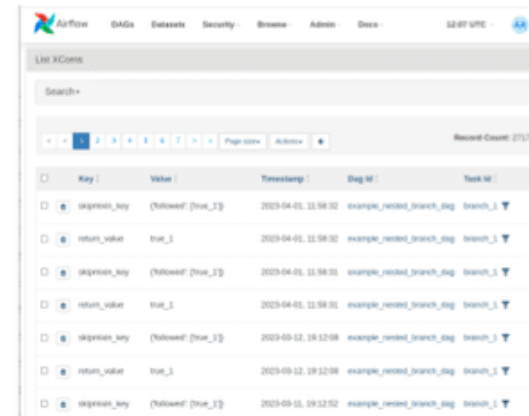
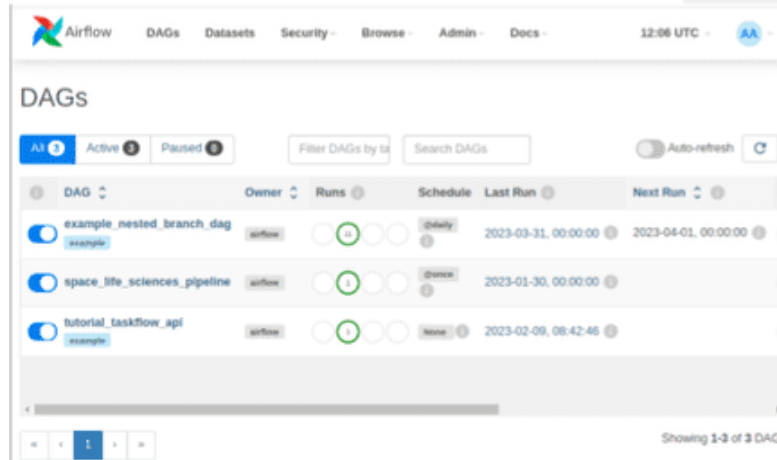
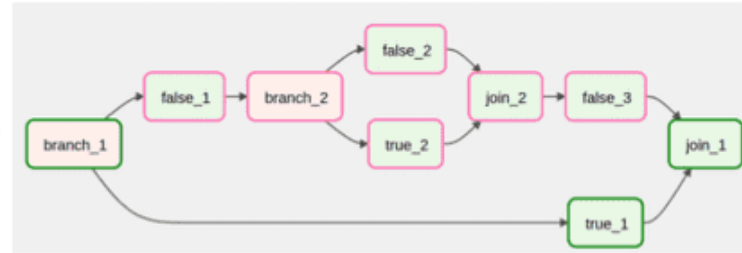


The Data-Centric Development Process



Source: Luley et al., «From Concept to Implementation: The Data-Centric Development Process for AI in Industry», 2023

The Data-Centric Development Process Orchestration



Source: <https://theaisummer.com/apache-airflow-tutorial/>

The Data-Centric Development Process

Model Lifecycle

mlflow Github Docs

Listing Price Prediction

Experiment ID: 0 Artifact Location: /Users/matei/mlflow/demo/mlruns/0

Search Runs: Search

Filter Params: Filter Metrics: Clear

4 matching runs Compare Selected Download CSV

| | | | | | Parameters | | Metrics | | |
|--------------------------|-------|-------|-----------|---------|------------|----------|---------|-------|-------|
| | Time | User | Source | Version | alpha | l1_ratio | MAE | R2 | RMSE |
| <input type="checkbox"/> | 17:37 | matei | linear.py | 3a1995 | 0.5 | 0.2 | 84.27 | 0.277 | 158.1 |
| <input type="checkbox"/> | 17:37 | matei | linear.py | 3a1995 | 0.2 | 0.5 | 84.08 | 0.264 | 159.6 |
| <input type="checkbox"/> | 17:37 | matei | linear.py | 3a1995 | 0.5 | 0.5 | 84.12 | 0.272 | 158.6 |
| <input type="checkbox"/> | 17:37 | matei | linear.py | 3a1995 | 0 | 0 | 84.49 | 0.249 | 161.2 |

Source: <https://datasolut.com/mlflow-machine-learning-plattform/>

The Data-Centric Development Process

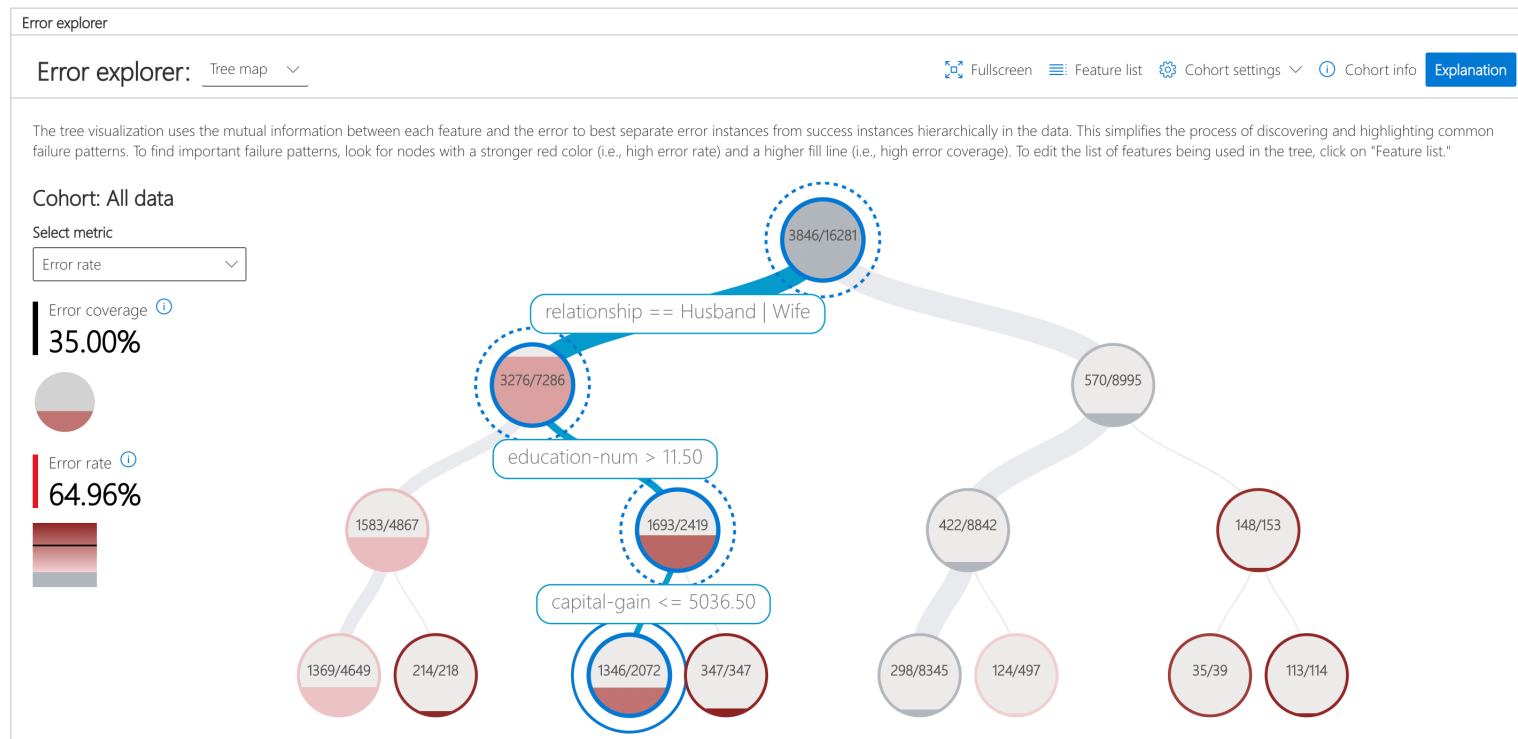
Version Control



Source: <https://www.codementor.io/@oussamalouati4/versioning-data-and-models-for-machine-learning-projects-with-dvc-1k3glsgqyf>

The Data-Centric Development Process

Error Analysis

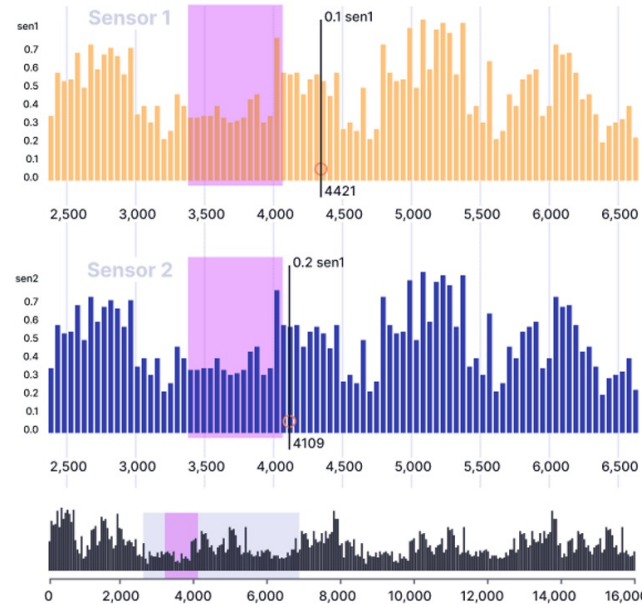


Source: <https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/erroranalysis-dashboard-README.md>

The Data-Centric Development Process

Label Studio

Defect^[1] | Diffusion^[2] | Incident^[3] | Failure^[4]



Skip

✓ Submit

Task ID: 1

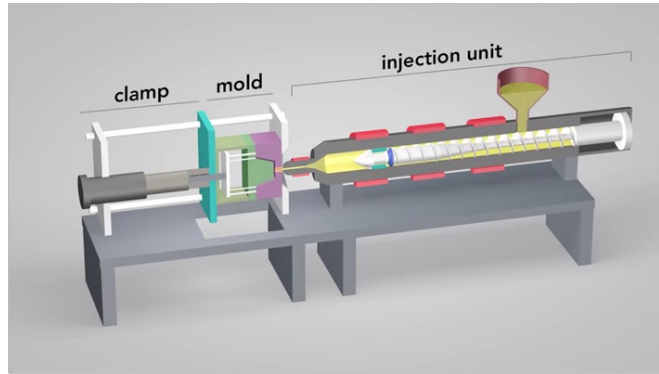
Source: <https://labelstud.io/>

Injection Molding Process

1. Injection molding machine and tool
2. Filling of the cavity
3. Pressure sensor in the cavity is picking up a signal
4. Pressure sensor is transferred to the process monitoring unit ComoNeo



State of the art quality control

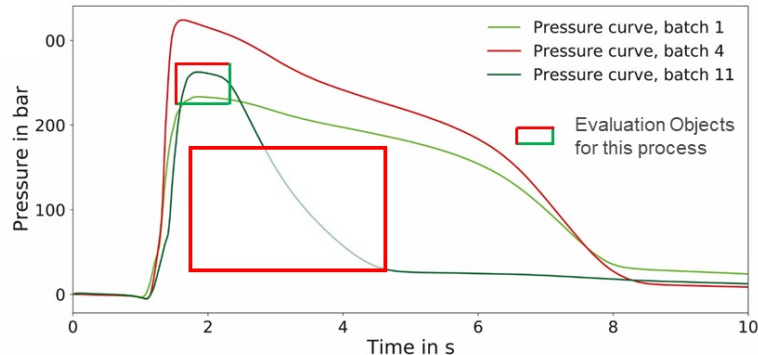


-Optical inspection of every produced part is possible but extremely expensive

-The good news / state of the art: Process parameters and quality “fingerprint” are encoded in the cavity **pressure curve**

-Experts can then define Evaluation Objects (EOs) reflecting an anomaly-free operation

-In case of errors: experts are needed to find causes and remove them



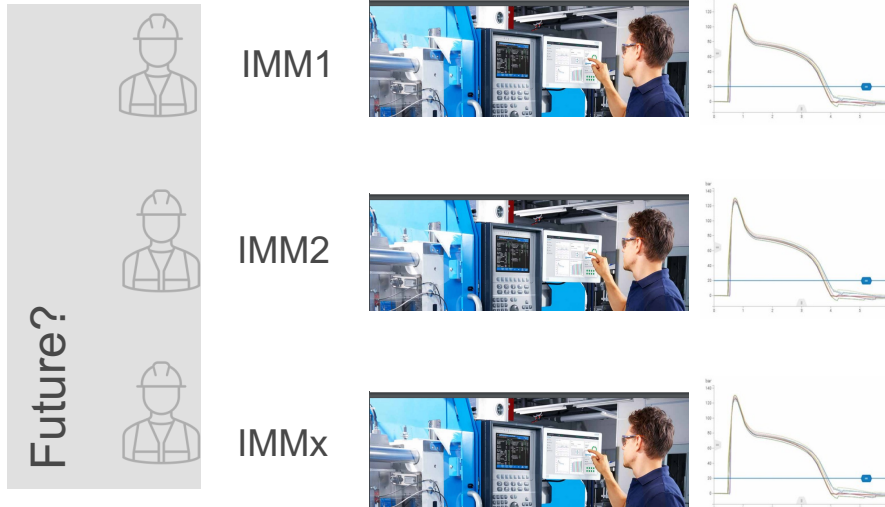
Challenges for Injection Molders in the Future



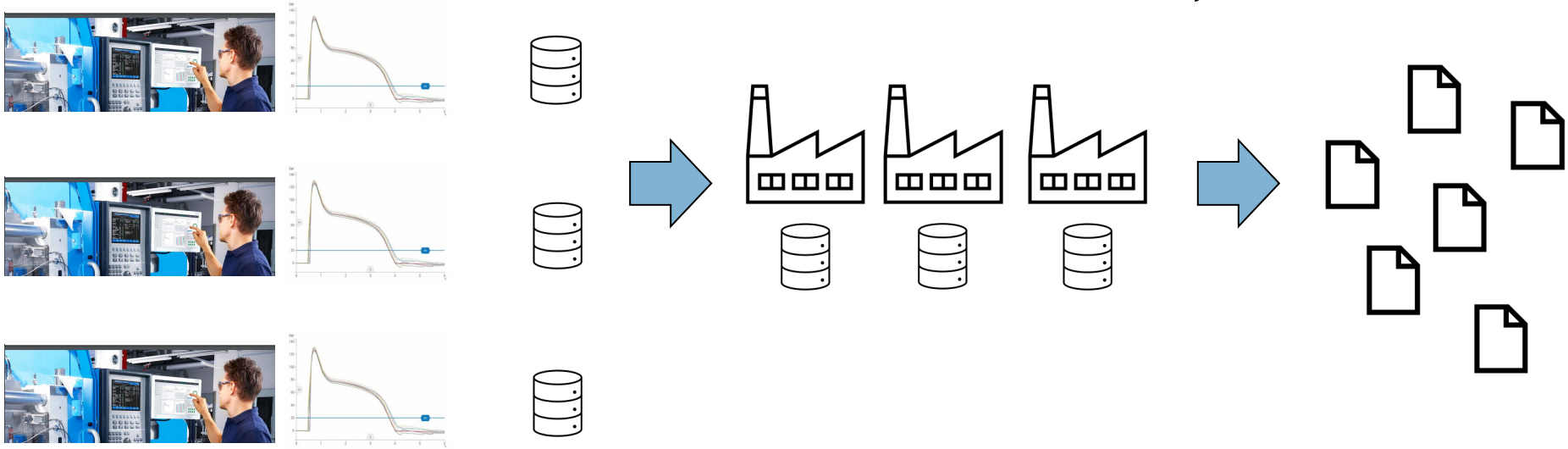
Shopfloor with multiple machines

-Usually have very large shopfloor with machines producing batches of 100'000s parts each

-In the future, the availability of experts will decrease (# experts / machine)



What Stands in the Way of AI to Encode Expert Knowledge?



- Historically: data handling of process monitoring not designed for machine learning
- Challenge: setup process for data centric development while using data supplied from the installed base. → Need for data centric development process

Take Away Messages



- 1 Data dev. should be integrated
- 2 DCAI is a shift not a turnaround
- 3 ...it can complement MCAI nicely
- 4 Choose your tools appropriately