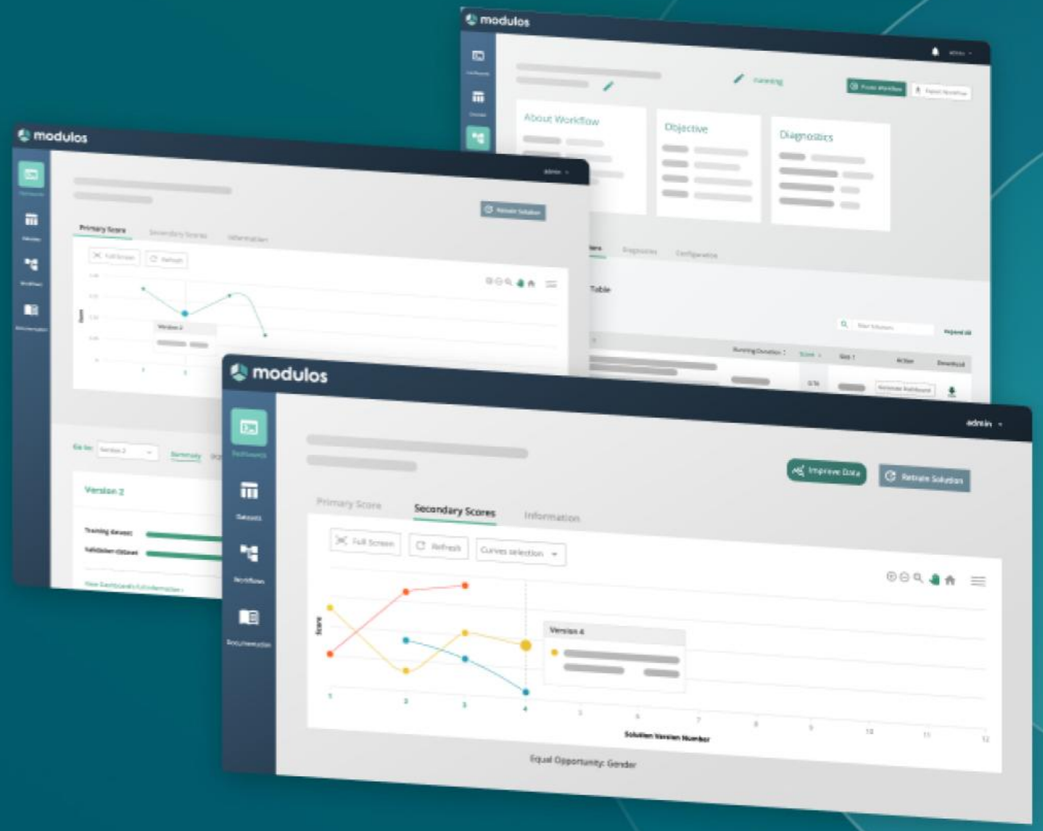
modulos

The Power of Data-Centric AI

"It is not about big data, but good data"



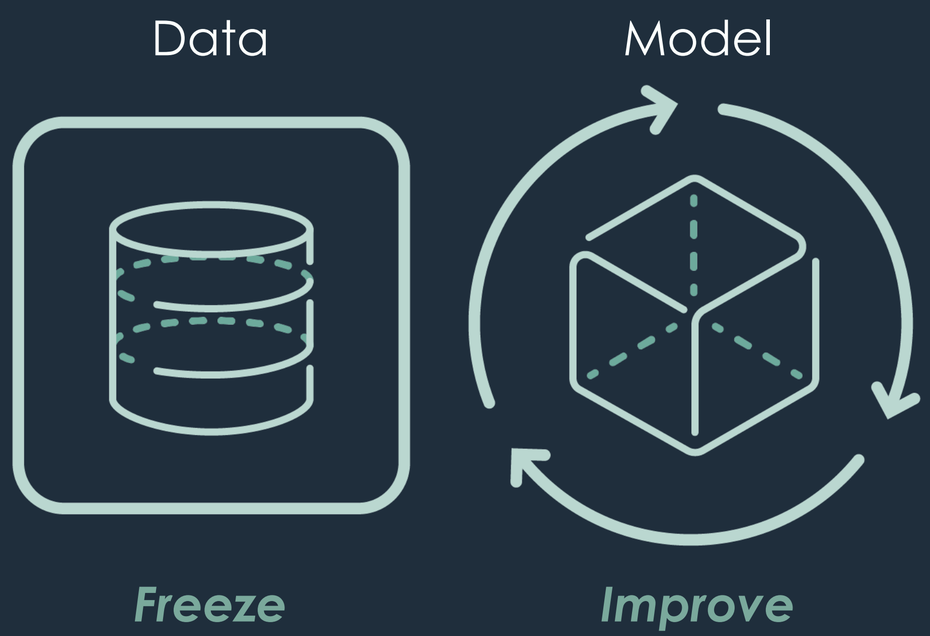
Anna Weigel, PhD
CTO

#datacentricai

swissnex AI Infrastructure Alliance swiss made software Spin-off | ETH zürich

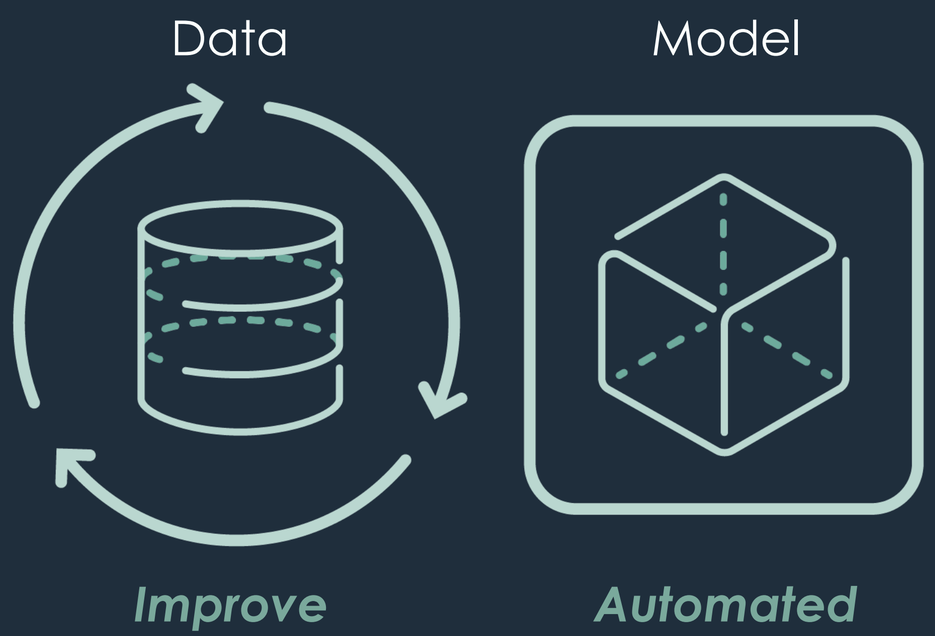
Paradigm Shift Towards Data-Centric Models

Model-centric ML Development



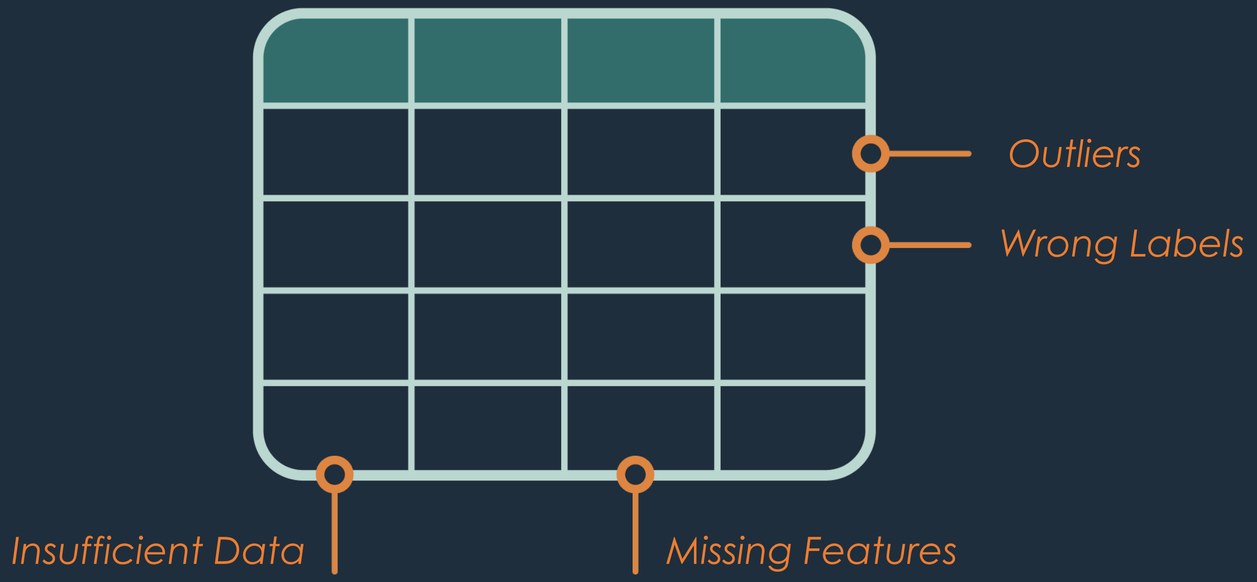
Traditionally, the main focus was on model improvement

Data-centric ML Development



However, it is important to focus on data as well

Model Training & Dataset Quality



Data

Which samples should you focus on to improve application performance?

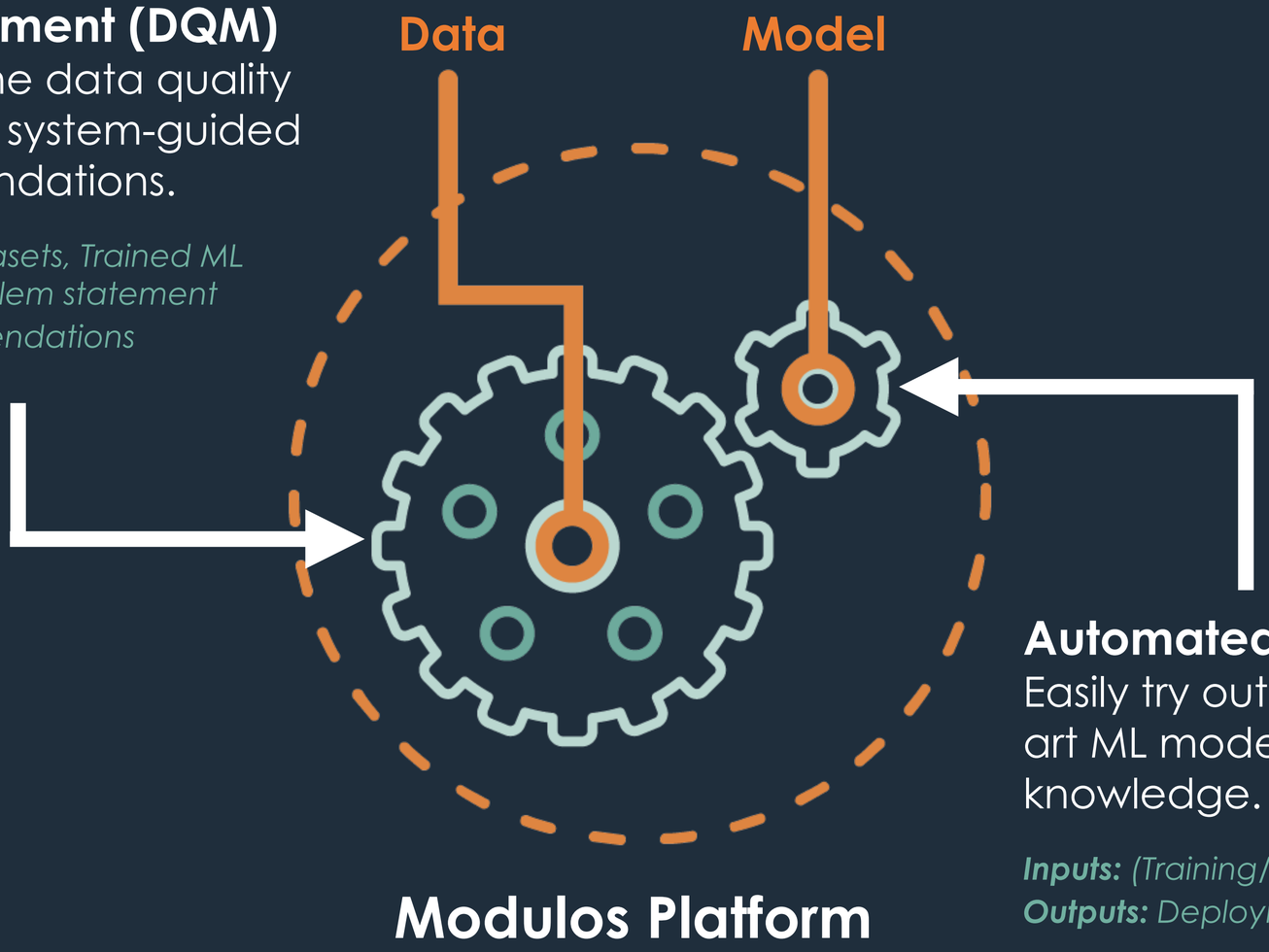
Model

How do you find the best model for the given dataset?

Main Modulos Platform Components

Data Quality Management (DQM)
Manage and improve the data quality of your ML datasets with system-guided Improvement recommendations.

Inputs: *(Training/Validation) Datasets, Trained ML Solutions, DQ improvement problem statement*
Outputs: *Improvement recommendations yielding cleaner datasets*



Automated Machine Learning (AutoML)
Easily try out and train different state-of-the art ML models without requiring ML expert knowledge.

Inputs: *(Training/Validation) Datasets, ML problem statement*
Outputs: *Deployment-ready ML Solutions*

7 Steps to Improve your Data Quality

Upload Data

Start by uploading your dirty datasets

Find and Train ML Solutions

Use Modulos AutoML to train the best Solutions

Select ML Solution

Pick the best solution you wish to improve iteratively

Assess Performance

Evaluate scores like fairness and benchmark your Solution

Improve Data Quality

Get prioritized recommendations on which samples to clean for reaching your objective

Retrain ML Solution

Retrain your Solution using the improved data

Deploy ML Solution

Deploy the output Solution to your private or public cloud servers to integrate it in your services

Use Case: Image Dataset

Experiment:

- Dataset: MNIST
- Training Dataset (10’000 images)
  - Pollute 1% (100) labels
- Validation Dataset (10’000 images)
- *Question: Can we detect the wrongly labeled training data?*

Use Case: Image Dataset

Experiment:

- Dataset: MNIST
- Training Dataset (10’000 images)
  - Pollute 1% (100) labels
- Validation Dataset (10’000 images)
- *Question: Can we detect the wrongly labeled training data?*

Random Cleaning:

Use Case: Image Dataset

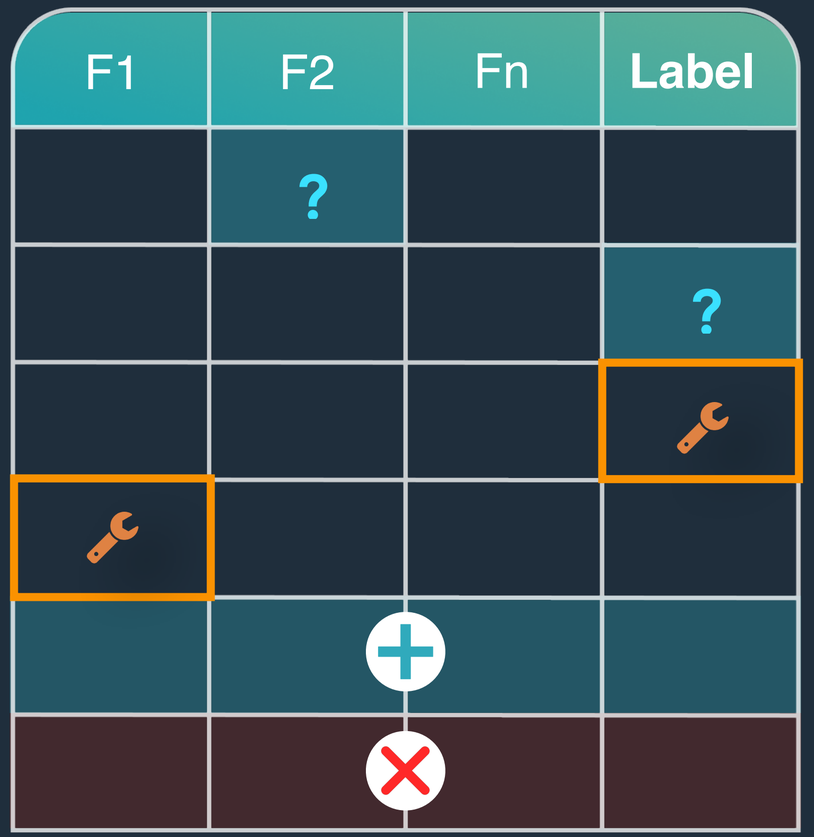
Experiment:

- Dataset: MNIST
- Training Dataset (10’000 images)
  - Pollute 1% (100) labels
- Validation Dataset (10’000 images)
- *Question: Can we detect the wrongly labeled training data?*

Targeted Cleaning:

Faster and Smarter data cleaning in ML application:

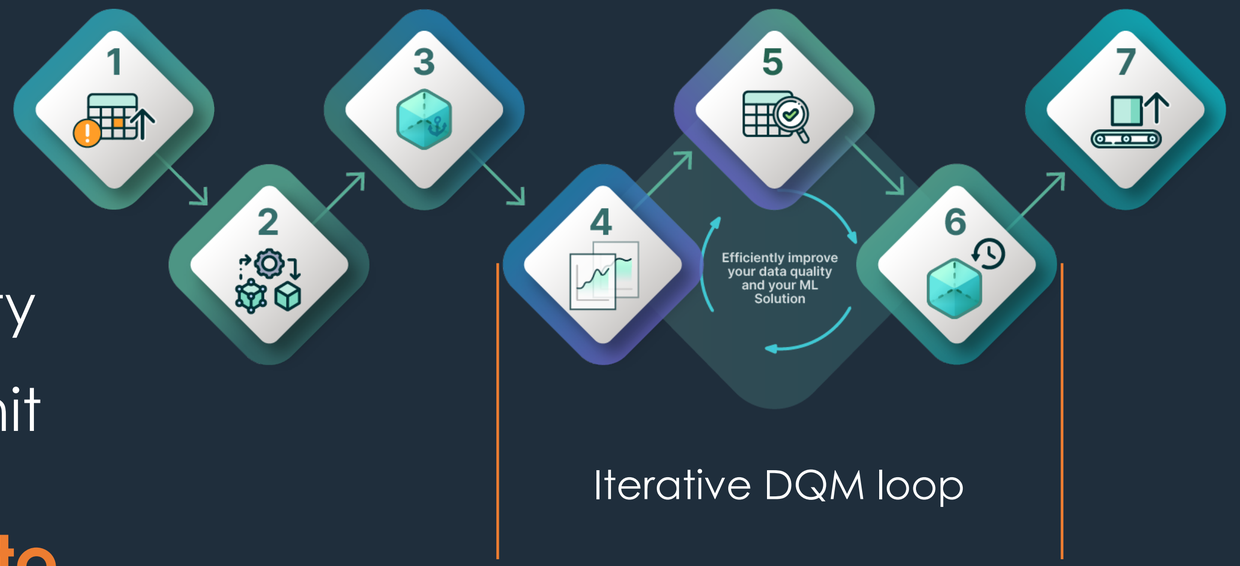
- Identify flaws and “bugs” in your training data
- Fast, cost-efficient, objective-oriented data cleaning
- Traceability and accountability
- Build more accurate and fairer models



Key points

Data-Centric AI

- Model performance = reflection of data quality
- Values data quality over data quantity
- Implies treating data & model as a unit



Modulos Platform empowers you to

Easily build fair and accurate ML models!

- Reduce time to market
- Reduce data acquisition costs
- Be a step closer to compliance

Thank you!

modulos.ai | contact@modulos.ai | @modulos_ai