DATA SCIENCE COMPLEXITY AND SOLUTIONS IN REAL INDUSTRIAL PROJECTS

ROSEN

empowered by technology

WHO AM I

Artur Miller

🎓 Electrical Engineer   💼 Data Scientist

🐦 @arturmillerblog   ⚙ https://github.com/arturmiller

✉ amiller@rosen-group.com

Slides: https://github.com/rosen-group/conferences

INTRODUCING THE ROSEN GROUP

ROSEN develops and manufactures equipment, software, and methods for the inspection, diagnosis, and protection of industrial structures in a wide range of industries.

Because damage can cause serious impacts!

CHALLENGES

- **Classification** of installations and anomalies and accurate **estimation** of severity of defects.

- Our tools record **a huge amount data**, up to multiple terabytes per run.

- Severe defects threaten the integrity of the pipelines, therefore there is a **high risk** for environment and clients.

We address these challenges with **machine learning** and **Python**!
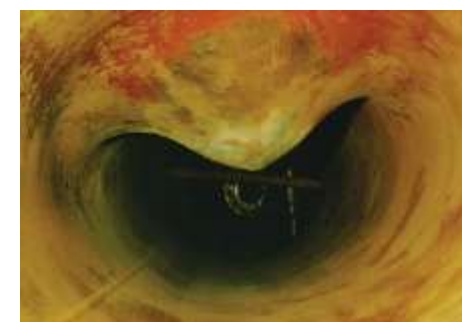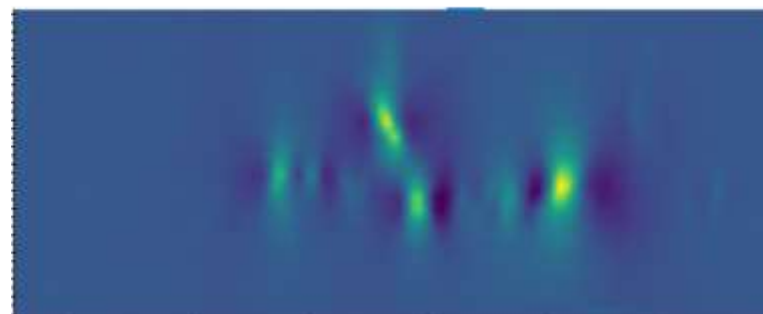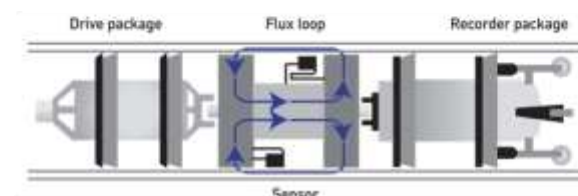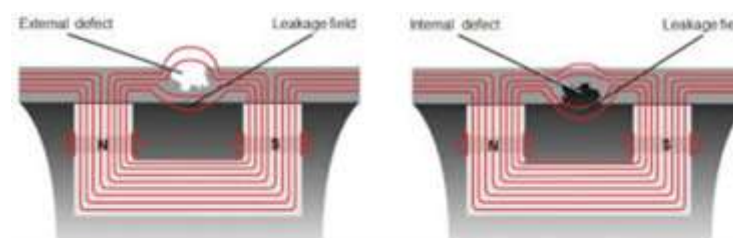
DATA SCIENCE MISCONCEPTION

- Misconception: Most of the time is spend tweaking machine learning models
- Real data is much more complex than toy datasets
- What is missing?
  - Data gathering
  - Data cleaning
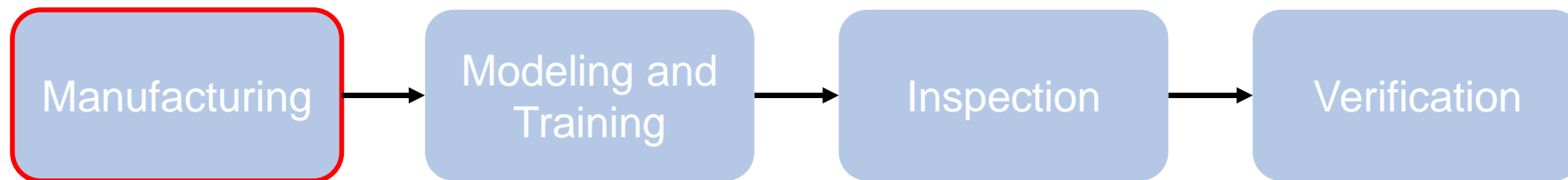 - Missing and imbalanced data
  - Scaling
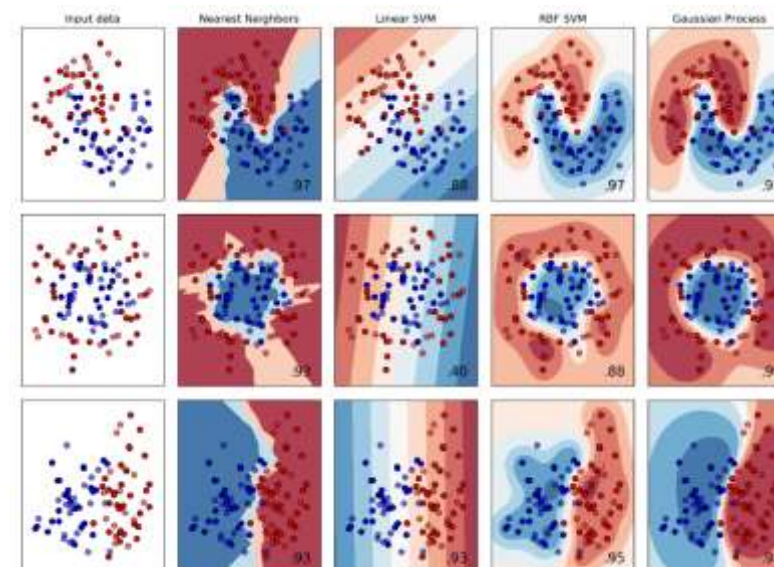
CLASSIC APPROACH

- Manufacturing and calibration of tools
- Models are based on pull-tests and laboratory measurements
- Verification results aren't fed back into the models

Manufacturing → Modeling and Training → Inspection → Verification
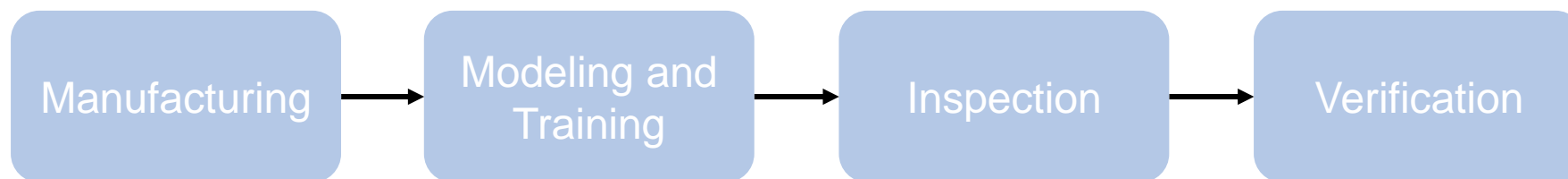
MANUFACTURING

- Tool is manufactured
- Laboratory measurements
- Pull-test or Pump-test
- Calibration
- Artificial defects


```
┌──────────────┐     ┌──────────────┐     ┌──────────────┐     ┌──────────────┐
│ Manufacturing│ ──> │ Modeling and │ ──> │  Inspection  │ ──> │ Verification │
│              │     │   Training   │     │              │     │              │
└──────────────┘     └──────────────┘     └──────────────┘     └──────────────┘
```

MODELING AND TRAINING

- Estimating the depth and shape of defects with regression models
- Classification of installations and anomalies
- Data from
  - Pull-tests and pump-tests
  - Laboratory measurements


```
Manufacturing → Modeling and Training → Inspection → Verification
```
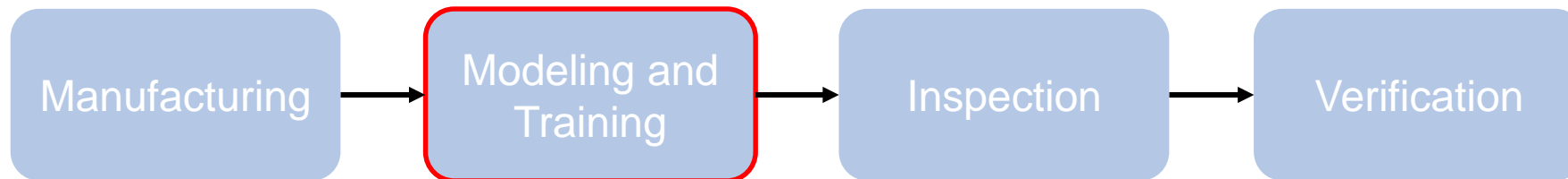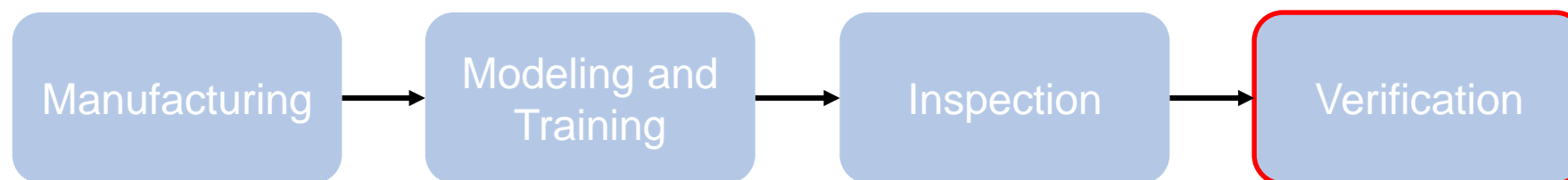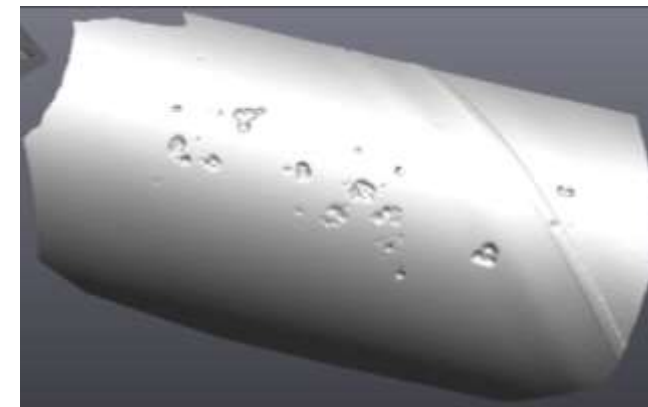
INSPECTION

- Preparing, transporting and launching the tool
- Tool takes measurements of the pipeline
- Processing and analyzing the data
- Reporting

Manufacturing → Modeling and Training → **Inspection** → Verification
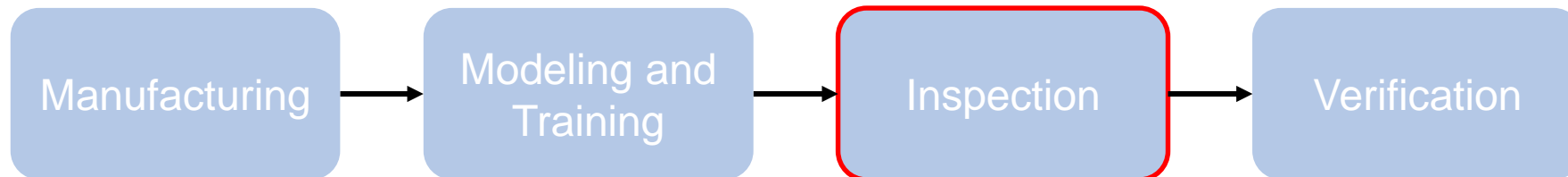
VERIFICATION

- Dig-ups
- Only severe defects
- Laser-scans
- X-ray computed tomography

Manufacturing → Modeling and Training → Inspection → **Verification**
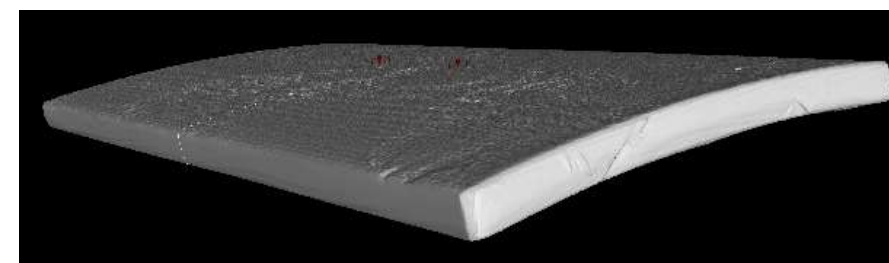
VERIFICATIONS

Laser-scans

- High resolution image of the outer pipeline wall
- Good image of corrosion
    - Depth and shape
- Better than artificial defects (e.g. ellipses)
- Non destructive
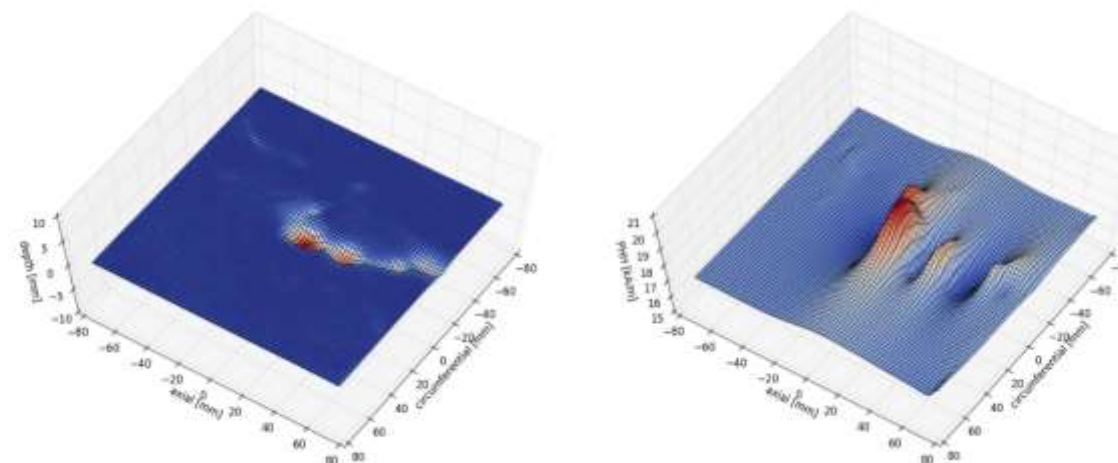
X-ray computed tomography

- High resolution 3D image of pipeline
- Good image of cracks (submillimeter resolution)
- Very expensive
- Specimen has to be cut out (destructive)

IMPROVED APPROACH

- Strong potential to improve the quality of our models
- Feeding back verification results into machine learning models
- Real defects are better than artificial defects
- A lot of distributed and non standardized data in-house
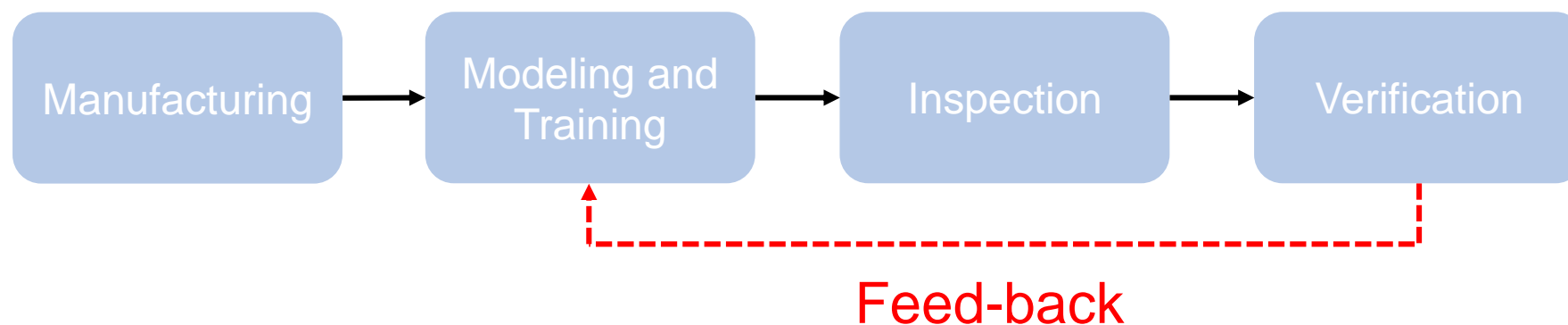
Manufacturing → Modeling and Training → Inspection → Verification

Feed-back

WHAT MAKES THIS FEEDBACK LOOP HARD TO ACHIEVE?

- How to get the verifications from the clients to the data scientists?
- We don't define what is verified
- Data is not clean
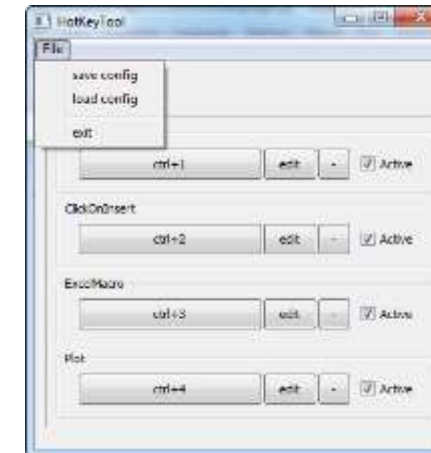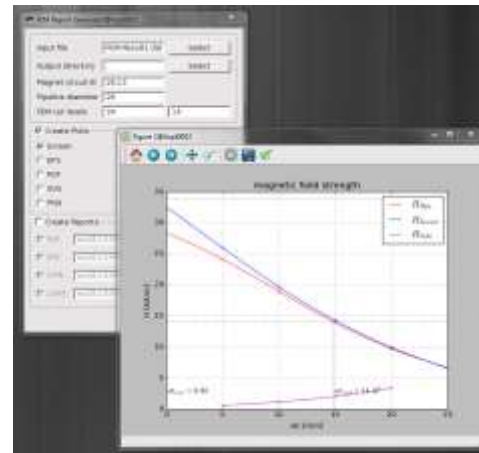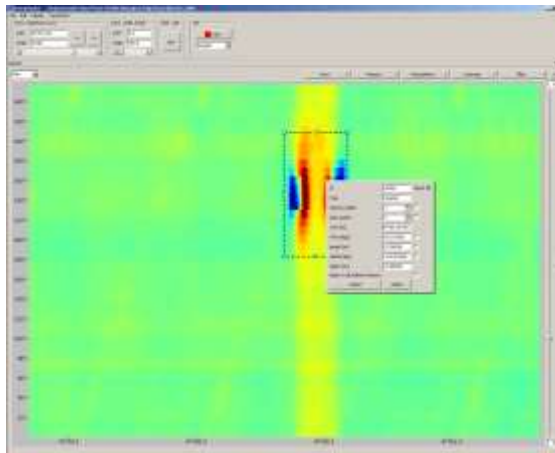- Data is not aligned

WHAT MAKES THIS FEEDBACK LOOP HARD TO ACHIEVE?

- How to get the verifications from the clients to the data scientists?
- We don't define what is verified
- Data is not clean
- Data is not aligned

HOW TO GET THE VERIFICATIONS FROM THE CLIENTS TO THE DATA SCIENTISTS?

- Know the delivery chain

  - Client ➡ Project Manager ➡ Data Engineer ➡ Annotator ➡ Data Scientist

- Define structures and processes

- Inform and train contact persons

- Help people who can help you

  - E.g. write small tools (in Python), which help them solve their problems
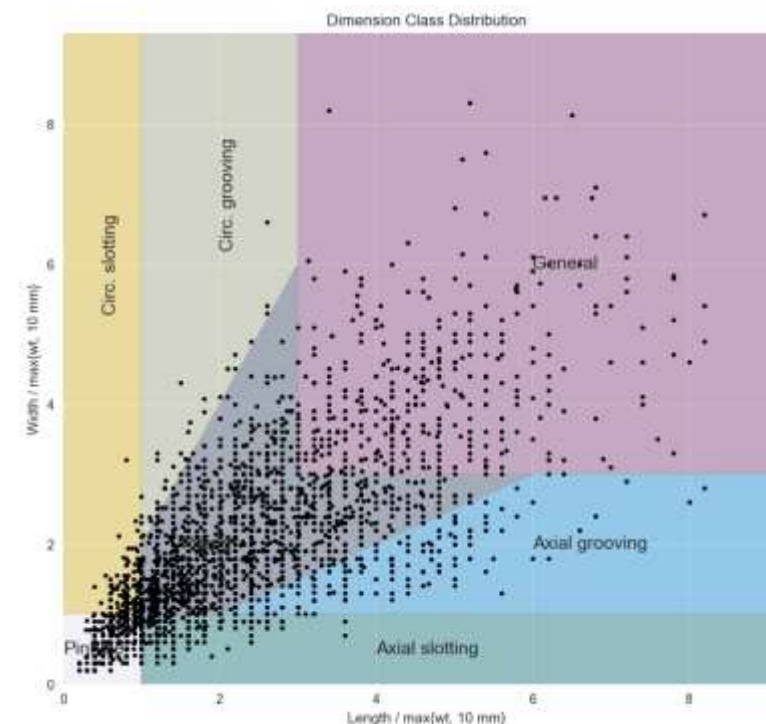
  - Automate the boring stuff

WHAT MAKES THIS FEEDBACK LOOP HARD TO ACHIEVE?

- How to get the verifications from the clients to the data scientists?
- We don't define what is verified
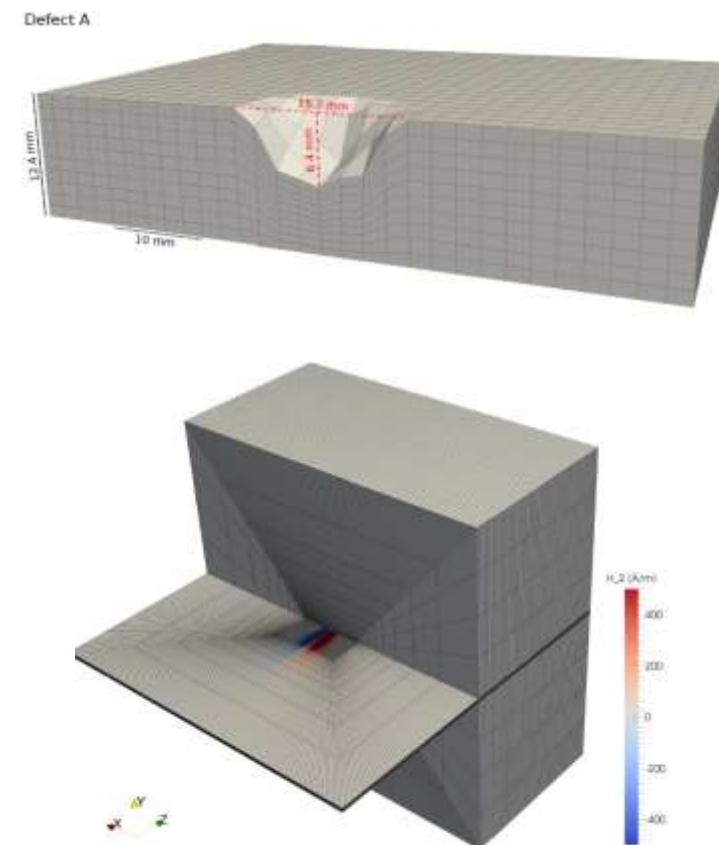- Data is not clean
- Data is not aligned

WE DON'T DEFINE WHAT IS VERIFIED

- Clients decide where excavations are done
- Very expensive!

- Fill the gaps:

  - Pull-tests and pump-tests

  - Laboratory measurements

  - Synthetic data (PyCon2017: Synthetic Data for Machine Learning Applications)

SMART WAYS TO FILL GAPS

- Synthetic defects

  - Distort laser-scans

  - Basic geometric shapes (e.g. ellipsoids)

  - Simulate corrosion growth (3D Cellular automata)
- FEM Simulations
- How to scale?

  - 15 min for each FEM simulation on one core

  - Distributed computation with a Docker cluster

WHAT MAKES THIS FEEDBACK LOOP HARD TO ACHIEVE?

- How to get the verifications from the clients to the data scientists?
- We don't define what is verified
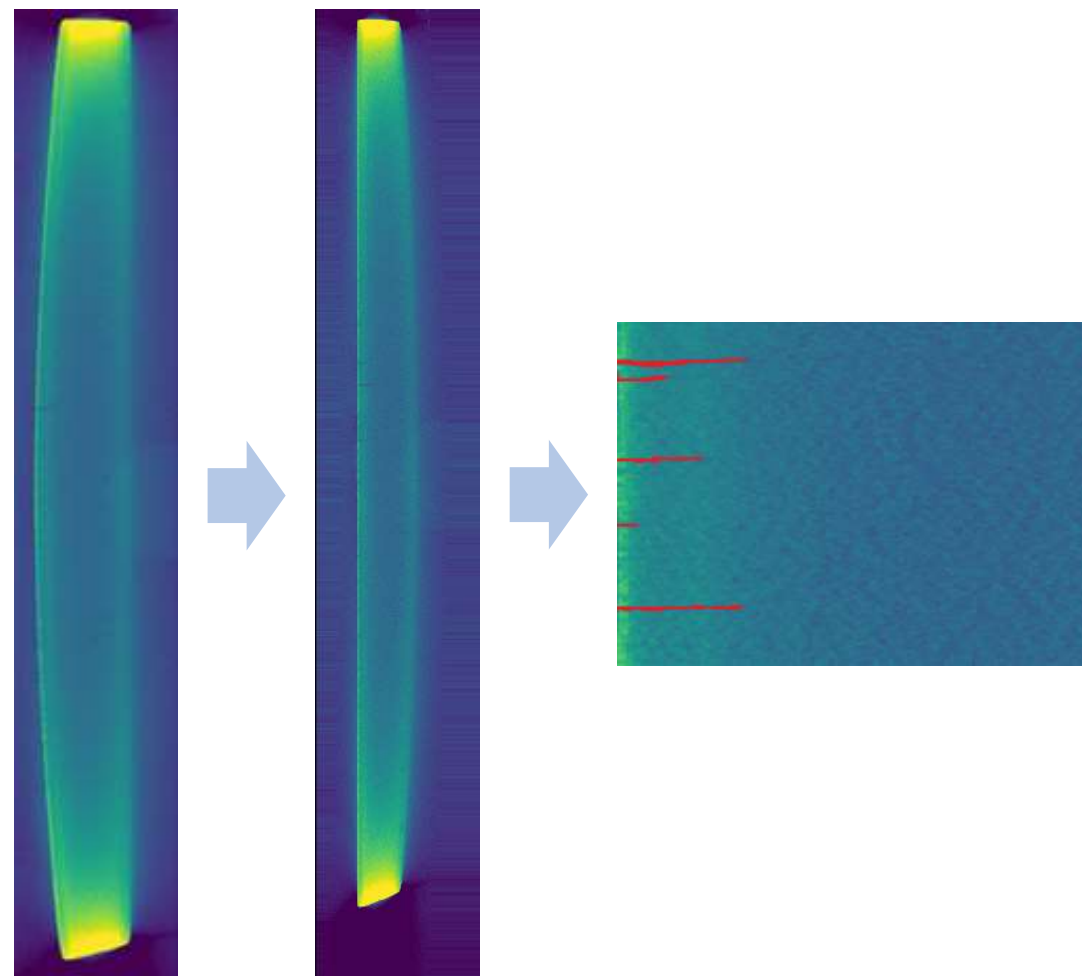- Data is not clean
- Data is not aligned

DATA STANDARDIZATION

- Clients use different data formats
- Flexible converter tools
- CSV or HDF5 as data container
- MongoDB for the meta data
- Proper interfaces for data access
- Data storage
- IT-support

DATA NORMALIZATION

- Data is not comparable to other data
  - Wall-thickness
  - Pipeline curvature
- Image processing
  - Filtering
  - Edge detection
  - Hough transformation
- Extraction of important information

WHAT MAKES THIS FEEDBACK LOOP
HARD TO ACHIEVE?

- How to get the verifications from the clients to the Data Scientists?
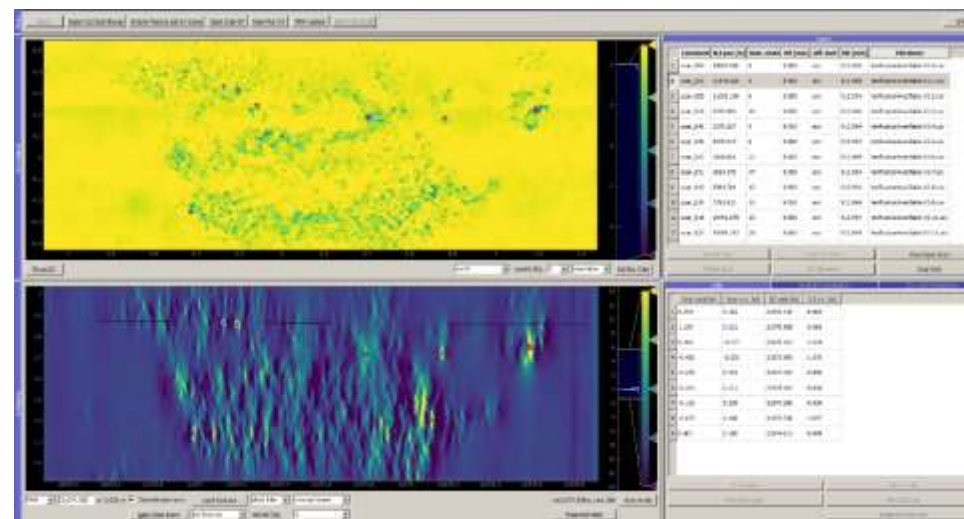- We don't define what is verified
- Data is not clean
- Data is not aligned

ALIGNMENT

Scan Alignment Tool

- Tool to align laser-scans and ILI-data
- Combines data from different sources
- Manual alignment

Automated alignment

- Template matching
- Direct comparison of laser-scans and ILI-data is not possible
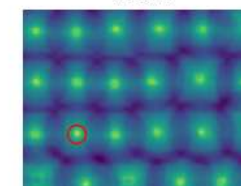- FEM simulations bridge this gap

template   image   `match_template` result

SUMMARY

- Data science challenges in ILI
- Misconception about data science
- Our classic and improved approaches
- Feeding back verification data is hard
- Various methods to tackle these challenges

Conclusion

- It is hard to feed-back validation data into our models
- But it is worth it!

  - Strong increase of classification and regression accuracies

  - Completely new approaches became possible
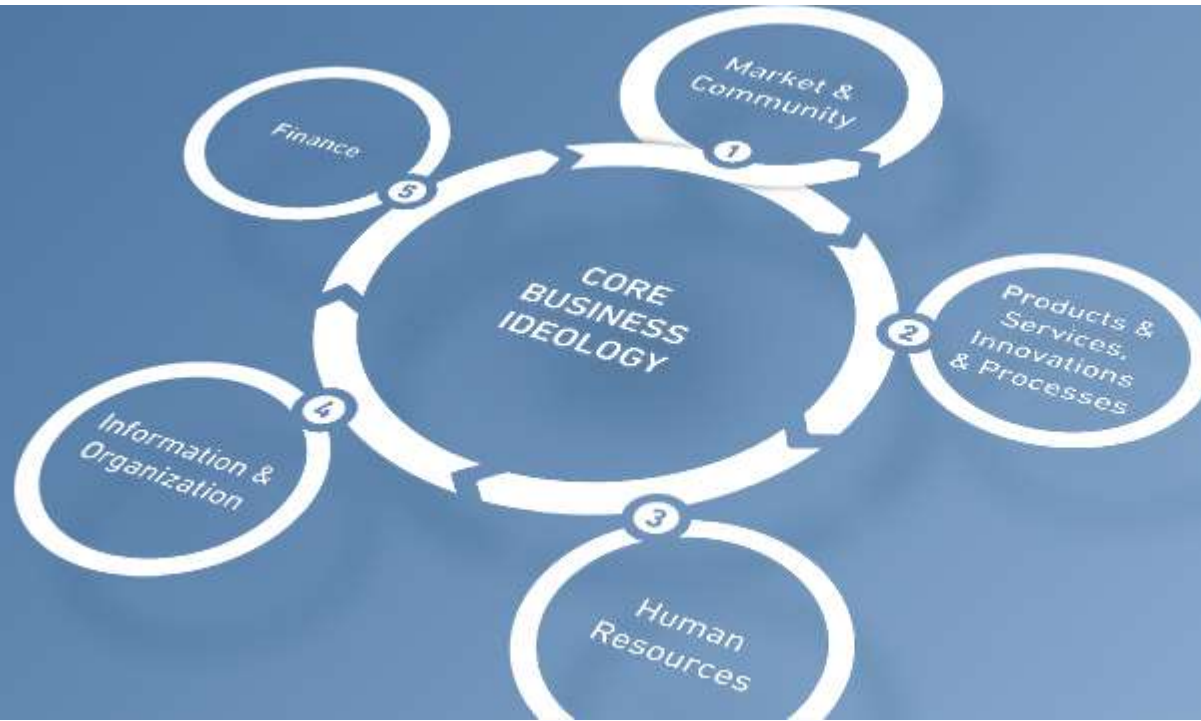
THANK YOU FOR JOINING
THIS PRESENTATION.

ROSEN

empowered by technology