EVALUATION OF PIPELINE INSPECTION DATA AS A SHOWCASE FOR INDUSTRIAL DATA SCIENCE

ROSEN
empowered by technology

WHOAMI

Hendrik Niemeyer

Theoretical Physics | Data Scientist

@hniemeye | https://github.com/hniemeyer

hniemeyer@rosen-group.com

Stephan Eule

Theoretical Physics | Data Scientist

seule@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!

INDUSTRIAL VS COMMERCIAL DATA SCIENCE

- Data Science practices and strategies can differ significantly between commercial and industrial Data Science applications
- Frequency of data input
- Cost of experiments and models
- Interpretability of data
- Demand on the accuracy of predictions

COMMERCIAL DATA SCIENCE

- Examples: online shop, click through rates, product recommendations, churn prediction
- E.g. identify and enforce preferable behavior, improve customer experience, increase revenue
- Usually large amount and high frequency of data input
- Data in general hard to misinterpret
- Allows real time experimentation (e.g. A/B testing)
- Predictions do not need to be very accurate due to high number of decisions
- Decisions of algorithm do need necessarily to be interpretable (although often desirable)

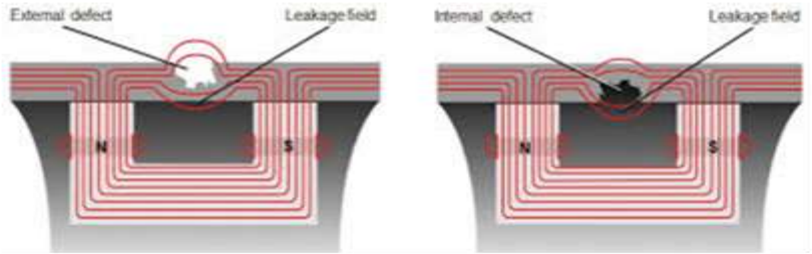
INDUSTRIAL DATA SCIENCE

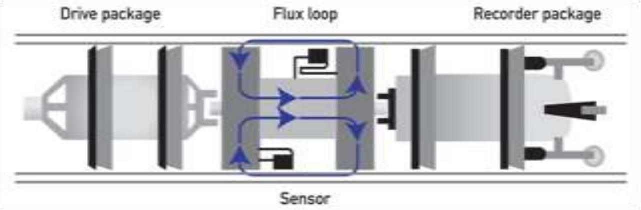
- Examples: Testing of devices, predictive maintenance of machines, planes, trains etc., (IoT applications)
- Data is often messy -> requires much more hands on management and analysis
- Create physical models using data sets
- Often time series data (sensors, successive measurements) -> restricts real time experimentation
- Automated testing is hard
- Predictions need to be very accurate, single prediction might be worth thousands of Euros
- -> a lot more money is allocated to single predictions, either by expensive algorithms or data collection/curation
- Models need to be interpretable
- (sometimes predictions need to be very fast, ‘on edge’)

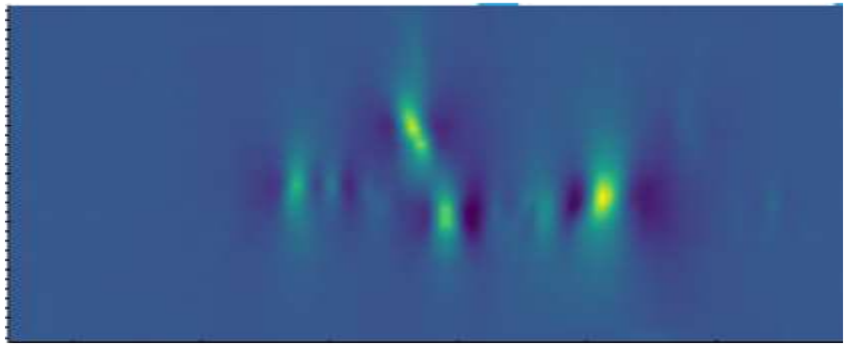
MAGNETIC FLUX LEAKAGE

- Measure volume loss in pipeline wall
- Indirect measurement principle
- Image-like data (2d array of amplitudes)

- Tasks: Detect, classify and estimate defect geometry from measured data.



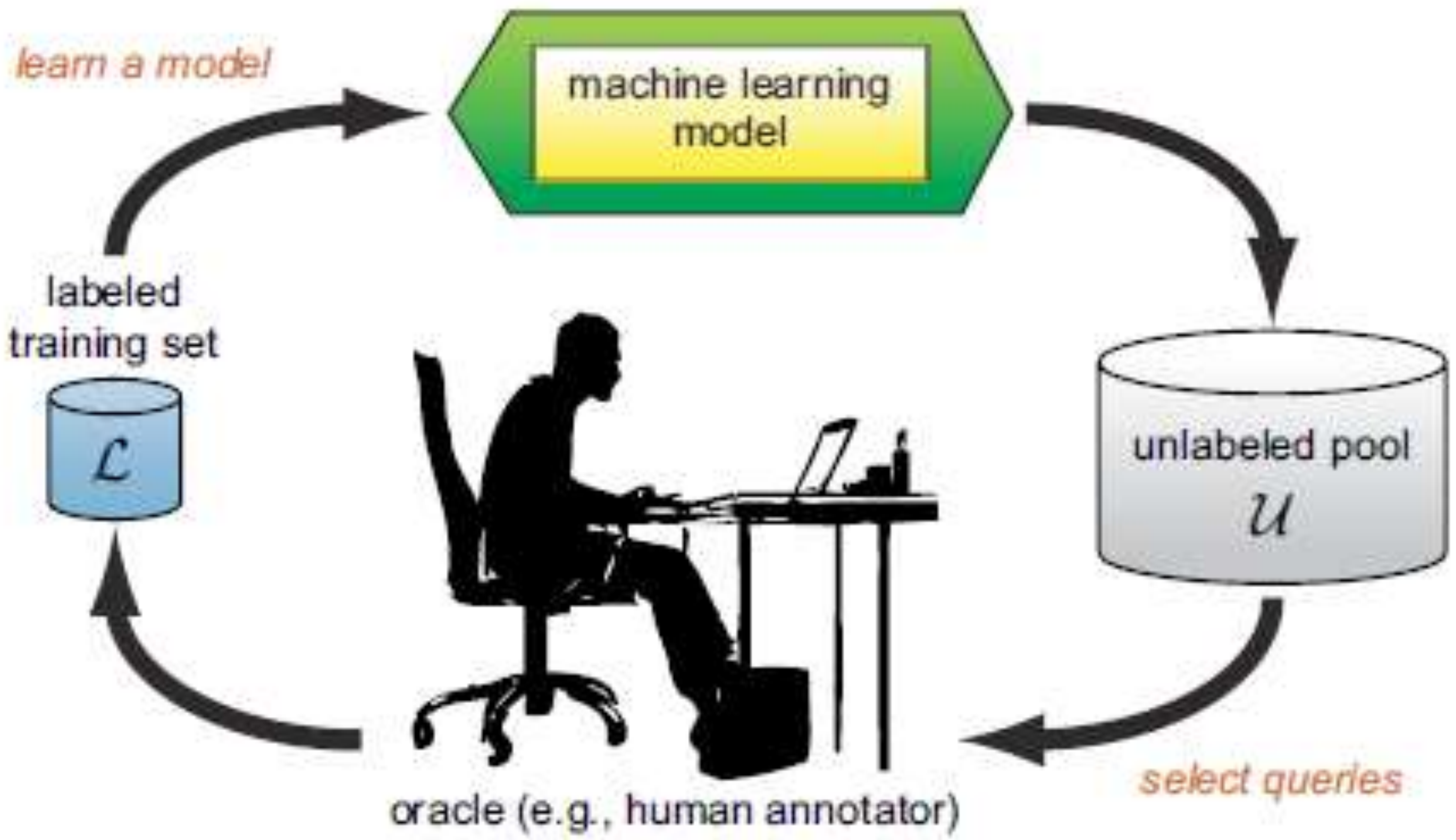




MOTIVATION FOR ACTIVE LEARNING

- Computer Vision problems
  - Large amounts of unlabeled data available
  - Human annotators can provide ground truth
  - Which instances to label?
  - Achieve high accuracy using as few labeled instances as possible

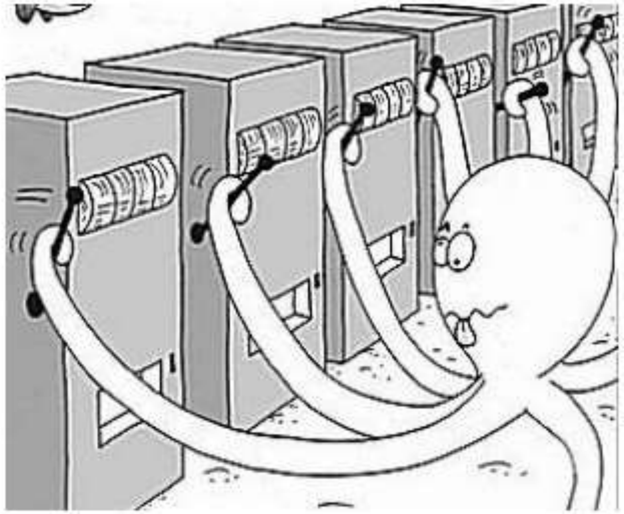
WHAT IS ACTIVE LEARNING?



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

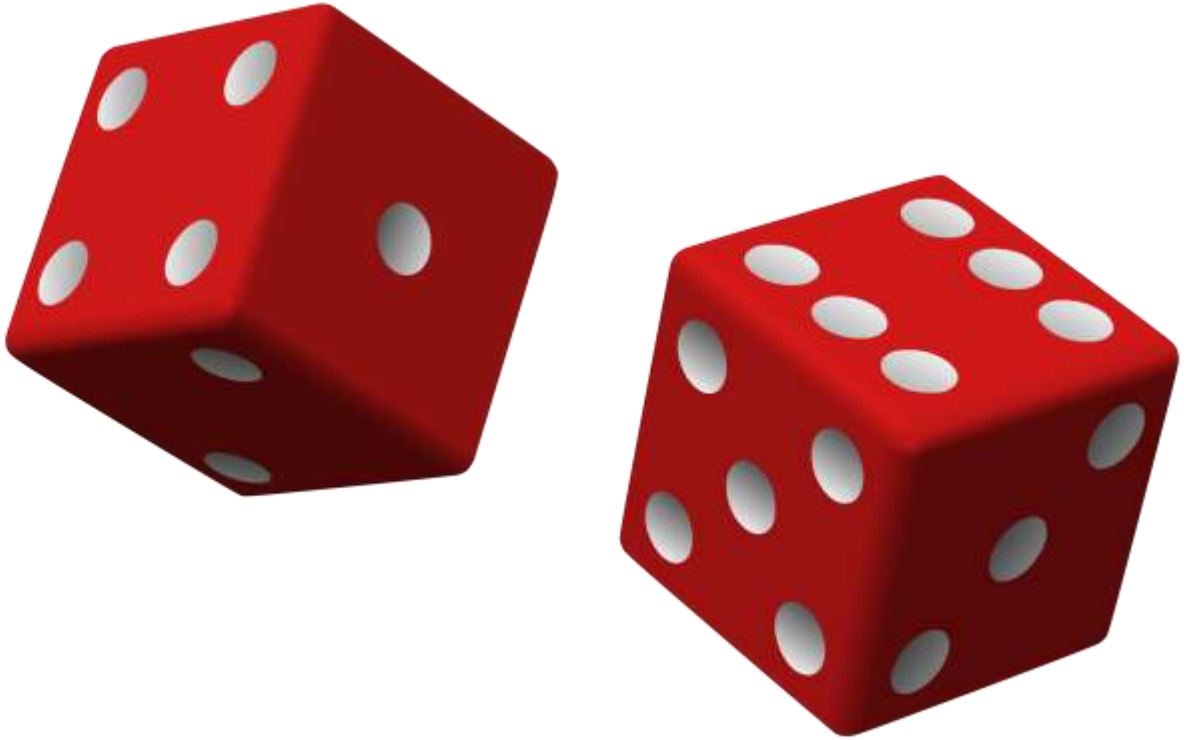
EXPLORATION VS EXPLOITATION DILEMMA

- How to design query algorithms?
  - Exploitation: make best decision based on currently available information
  - Exploration: gather more information



RANDOM QUERY

- Select instance to label randomly
- Good starting point
- Baseline algorithm : Compare other algorithms against random query



UNCERTAINTY SAMPLING

- Most popular algorithm
- Query the instance which the classifier is most uncertain how to label
- Least confident prediction in terms of prediction probability

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

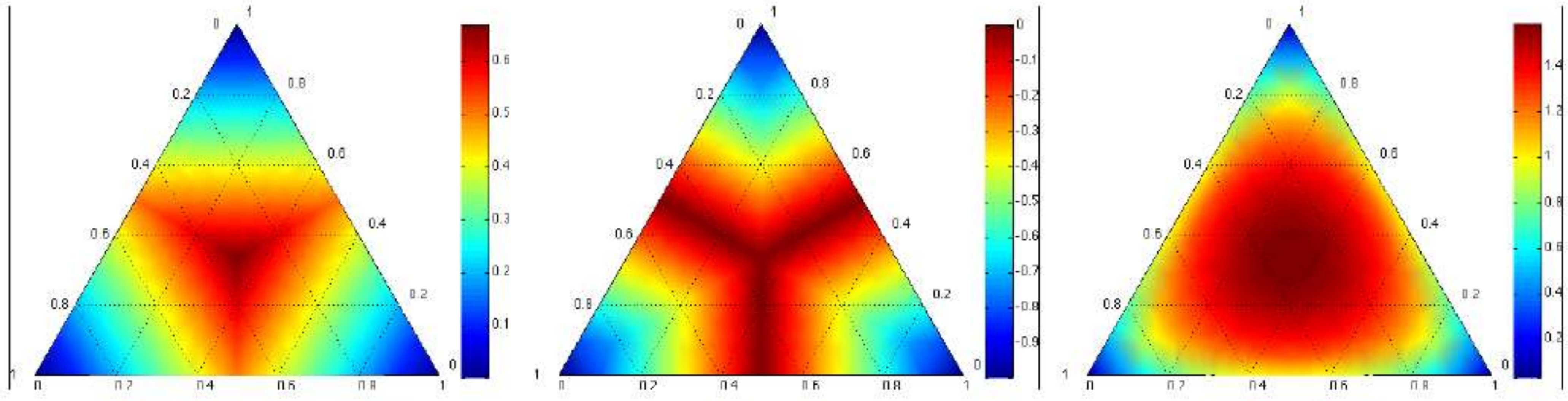
$$\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$$

- Make use of margin between most probable classes or entropy of class prediction probabilities to capture the whole distribution

$$x^*_M = \underset{x}{\operatorname{argmin}}\ P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

$$x^*_H = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

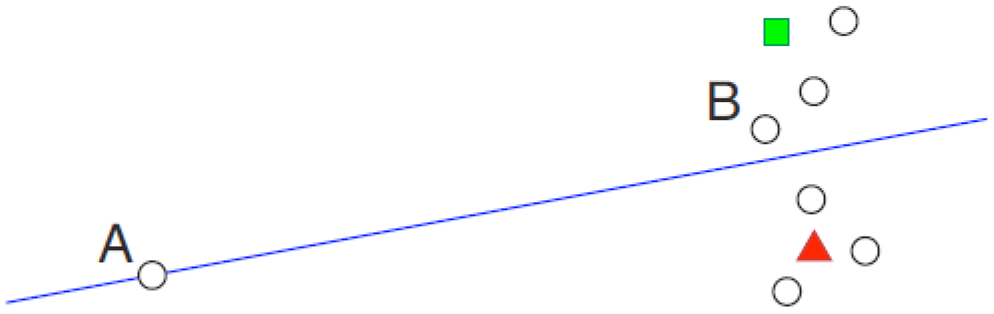
UNCERTAINTY SAMPLING



(a) least confident (b) margin (c) entropy

Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

SHORTCOMINGS OF UNCERTAINTY SAMPLING



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

OTHER ALGORITHMS

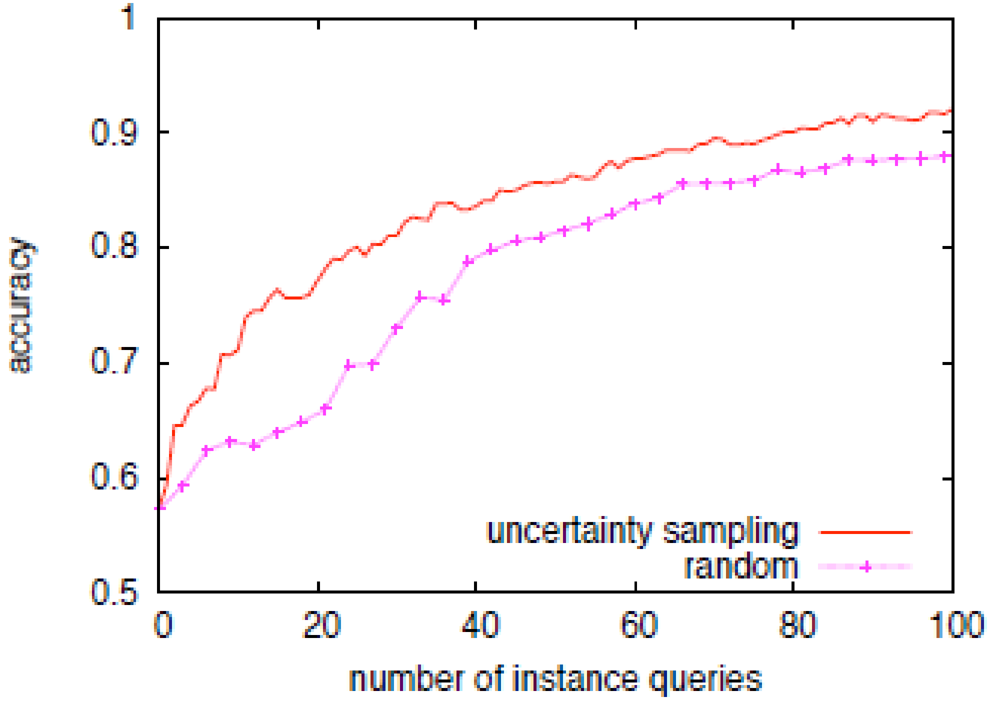
- Predict impact of model change when labeling a sample
- Predict expected error reduction induced by labeling a sample

WHICH ALGORITHM SHALL I USE?

- Algorithm for cold start when no classifier is available (random query, deterministic query)
- Balance exploration and exploitation
- Combine different selection strategies
- Test test test

TESTING QUERY ALGORITHM

- Active learning can be simulated on a fully labeled data set
- Human annotator is simulated by giving the correct label
- Compare performance over number of labeled instances against random query



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

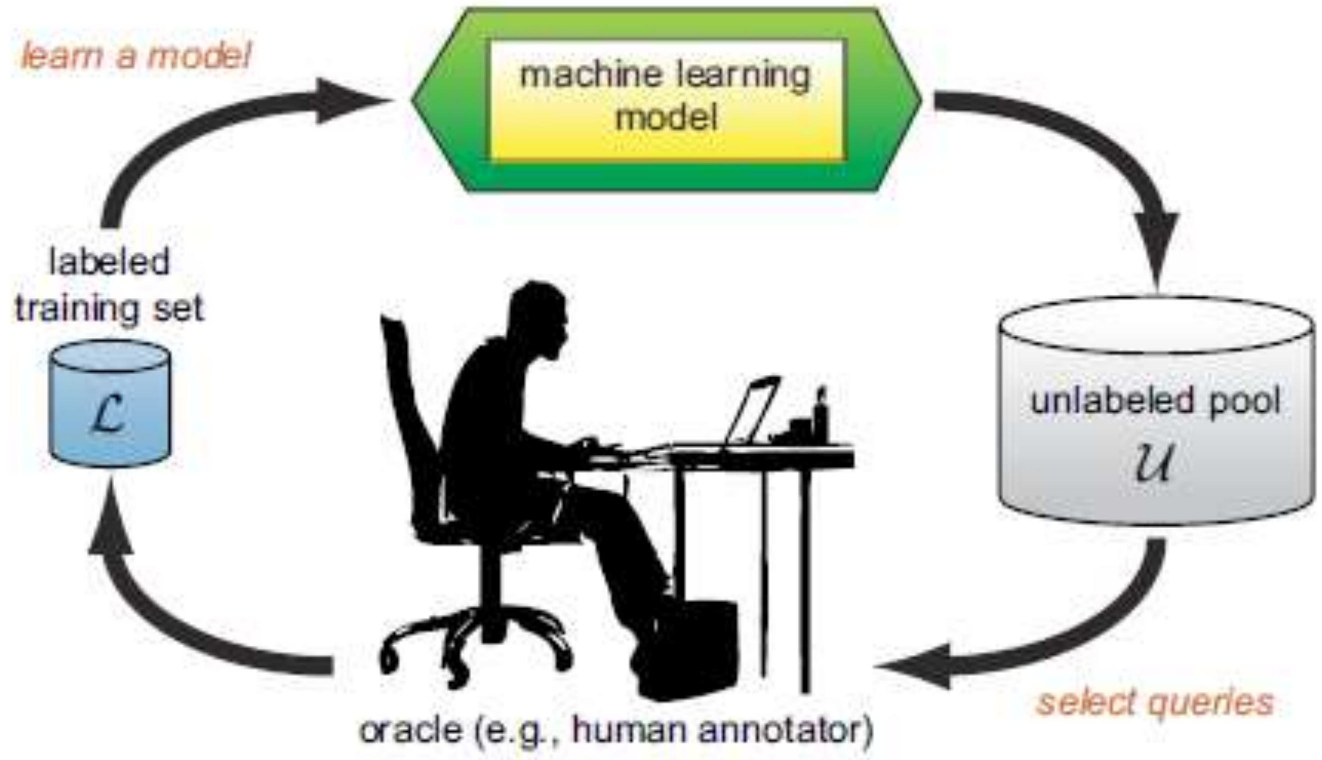
CONVERGENCE

- When to stop labeling?
  - Costs
  - Human annotator notices convergence
  - Stopping criterion

CONCLUSION

- Active learning can help you if you have a huge amount of unlabeled data and humans can provide ground truth
- There is no best query algorithm
- There is no best way to measure convergence

- Problems to consider:
  - Noisy human annotators
  - Class discovery
  - Training speed of ml algorithms



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

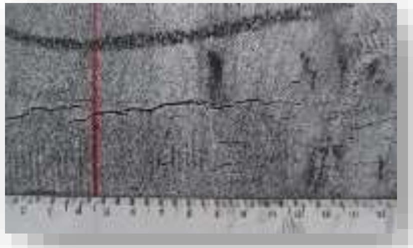
EMAT CRACK DETECTION AND COATING ASSESSMENT

22-24" EMAT ILI tool

Crack Detection

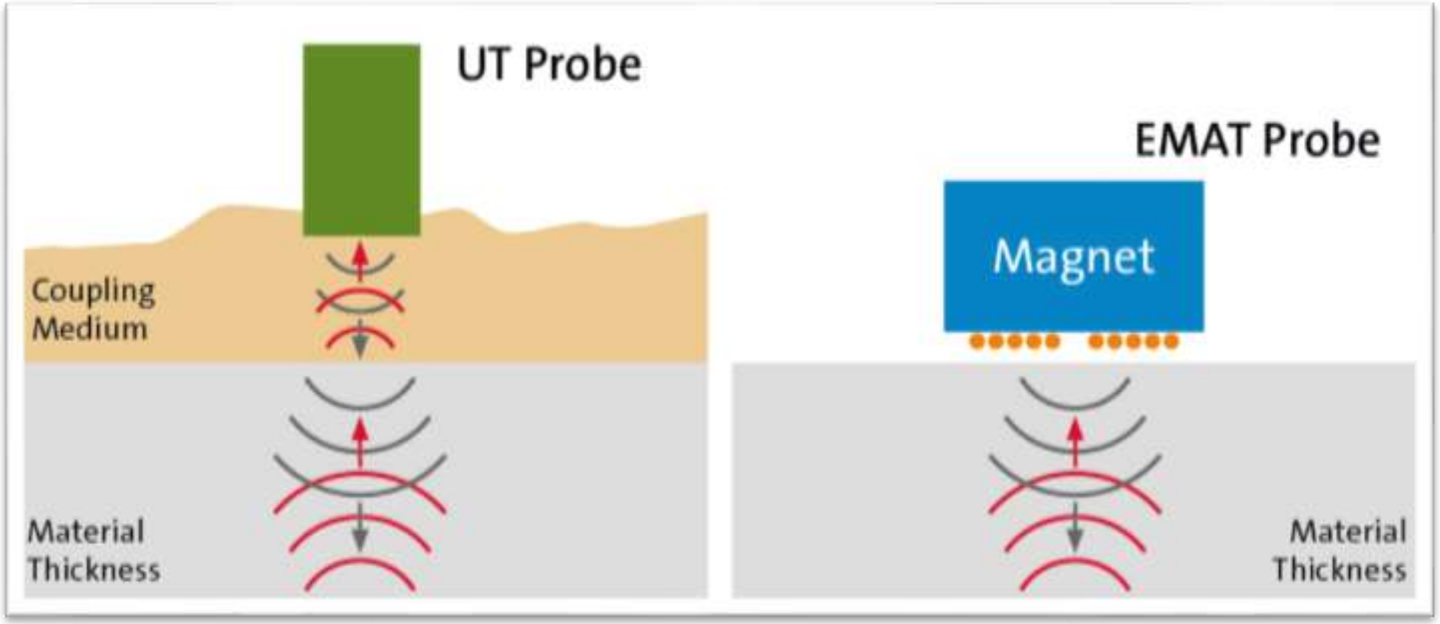
Coating Disbondment

Coating Identification

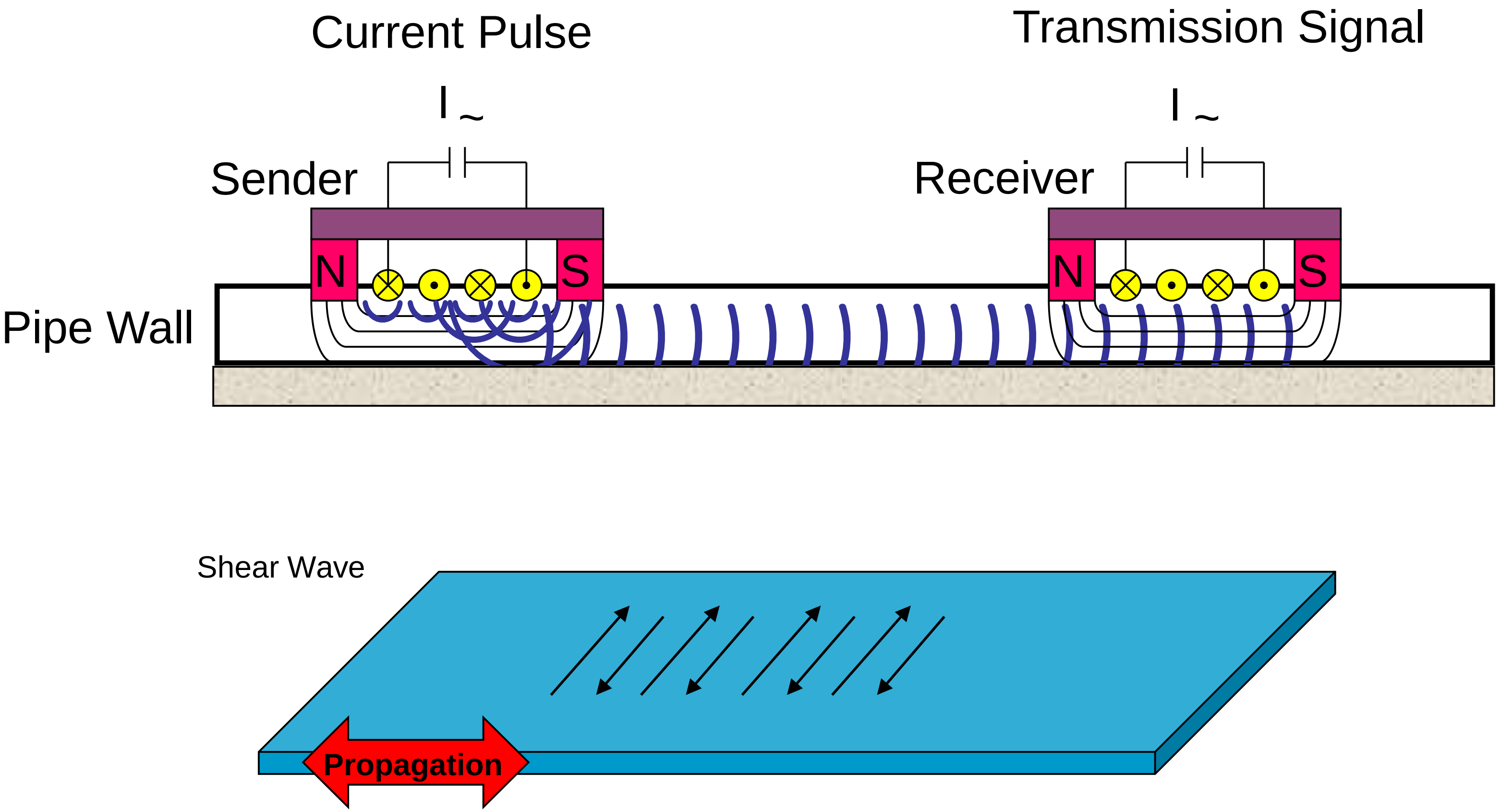


EMAT: COUPLANT-FREE ULTRASONIC INSPECTION

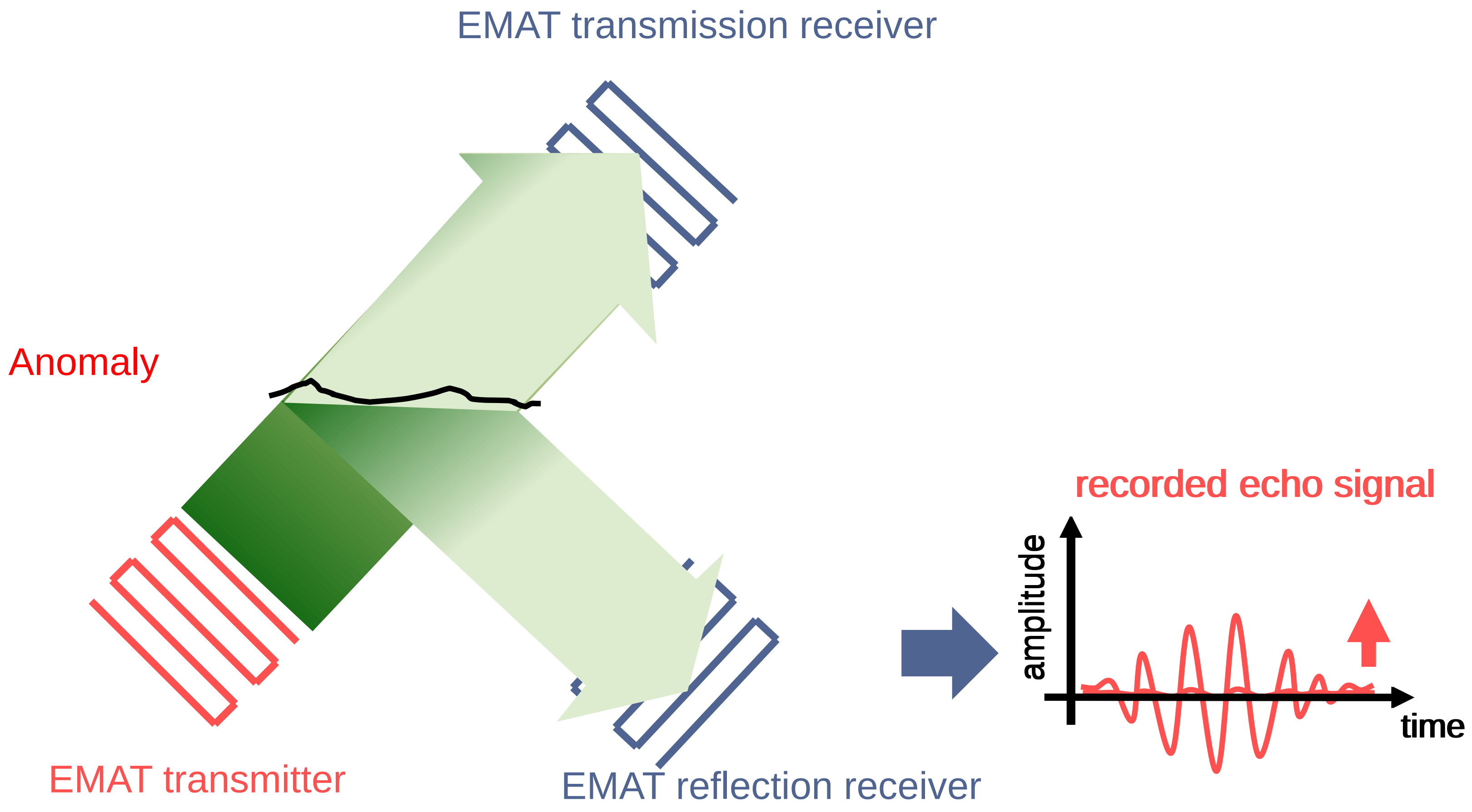
- EMAT: Electro Magnetic Acoustic Transducer
  - Applies forces to metal surface without direct contact
  - Detects movement in a metal surface without direct contact
  - Makes use of electromagnetic induction across air gap
  - Advantages: Contactless, Dry inspection, Usable at high temperatures and pressures, High volume coverage



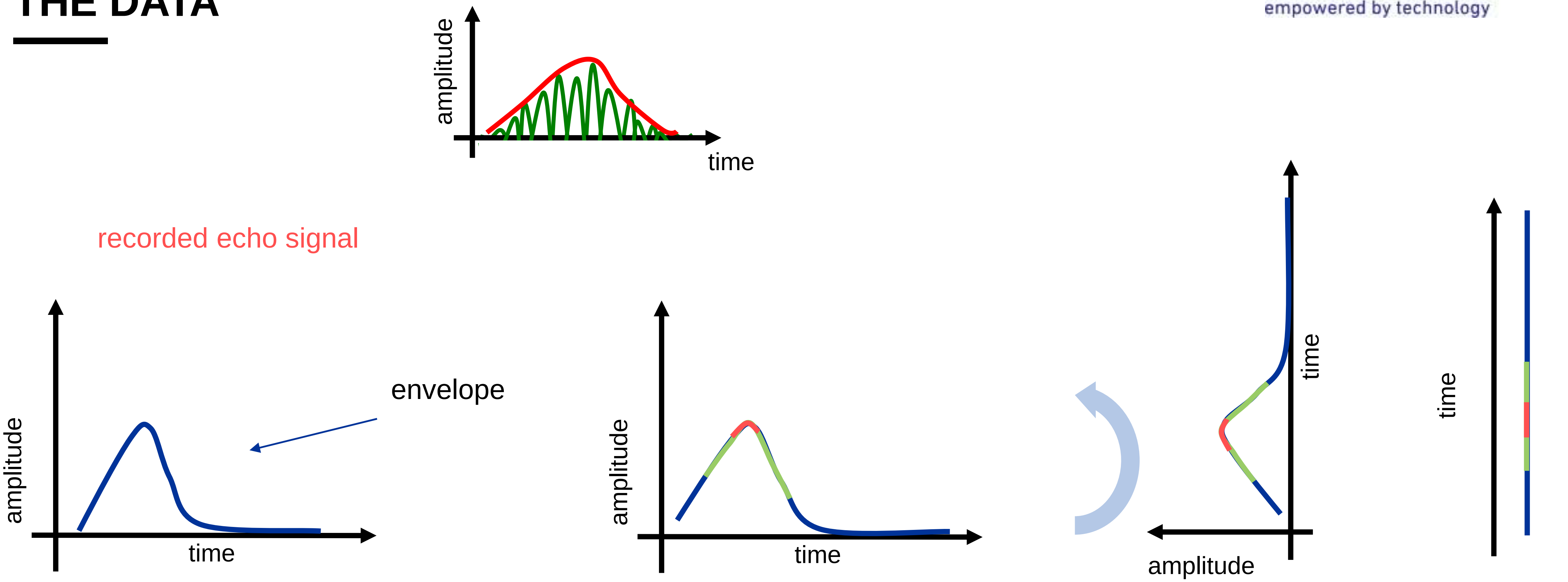
OVERALL SETUP



EMAT ANOMALY DETECTION



THE DATA


amplitude
time
recorded echo signal
envelope
amplitude
time
amplitude
time
time
amplitude
time

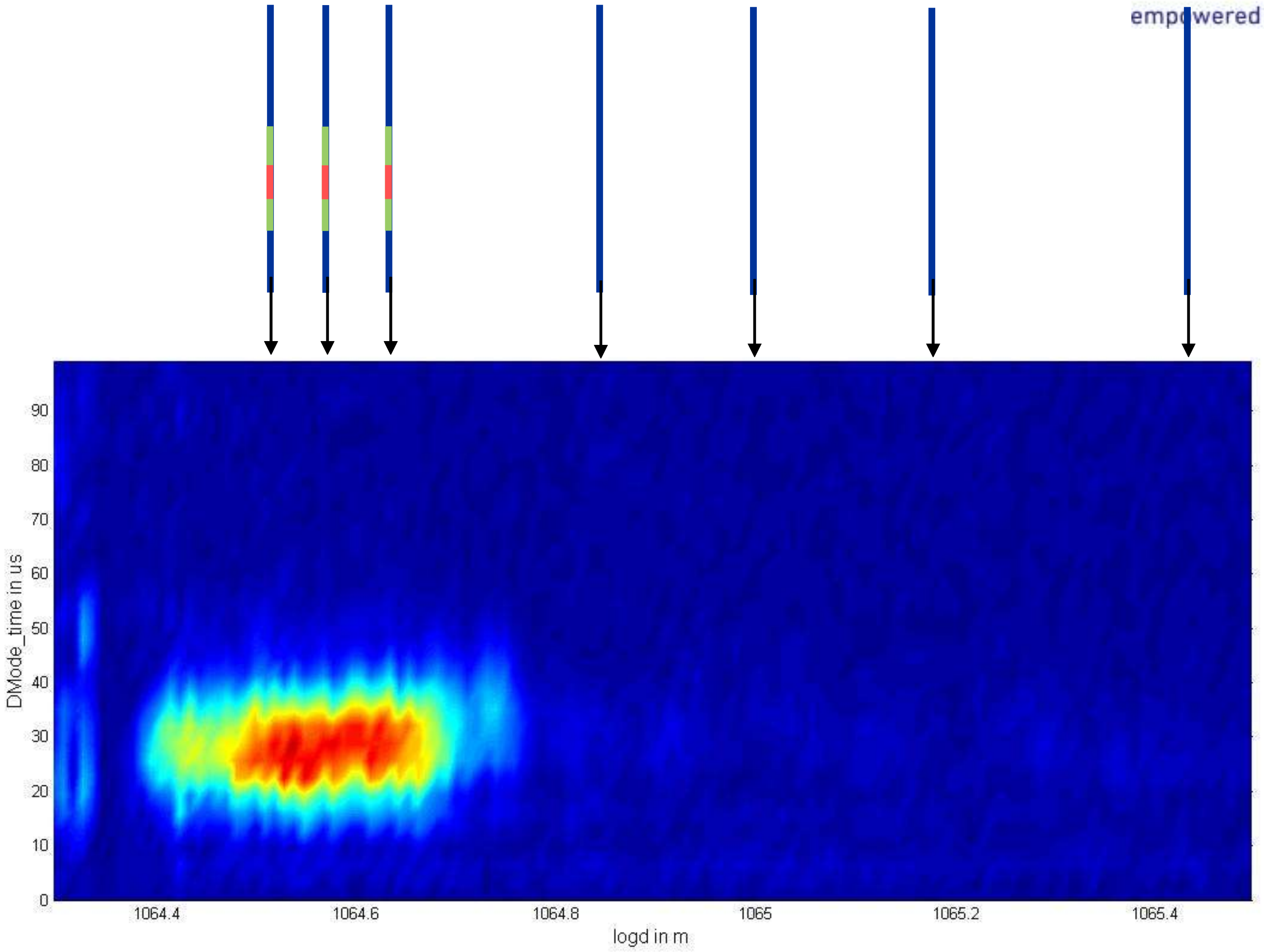
3D DATA

One channel:

2d image

Multiple channels

3D Data

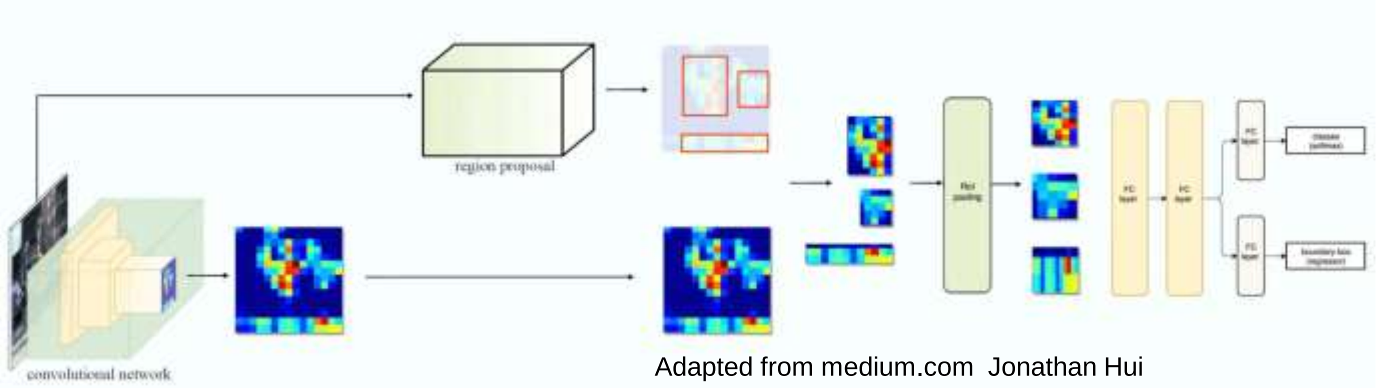


THE CHALLENGE

- Detect anomalies (objects, patterns) in 2D/3D data and classify anomalies
  - In principle a computer vision problem
  - Anomaly localization and classification
  - Object detection
  - Need labeled data
  - Provided by inhouse domain experts
  - Dig up campaigns
  - Recent breakthroughs in Deep learning

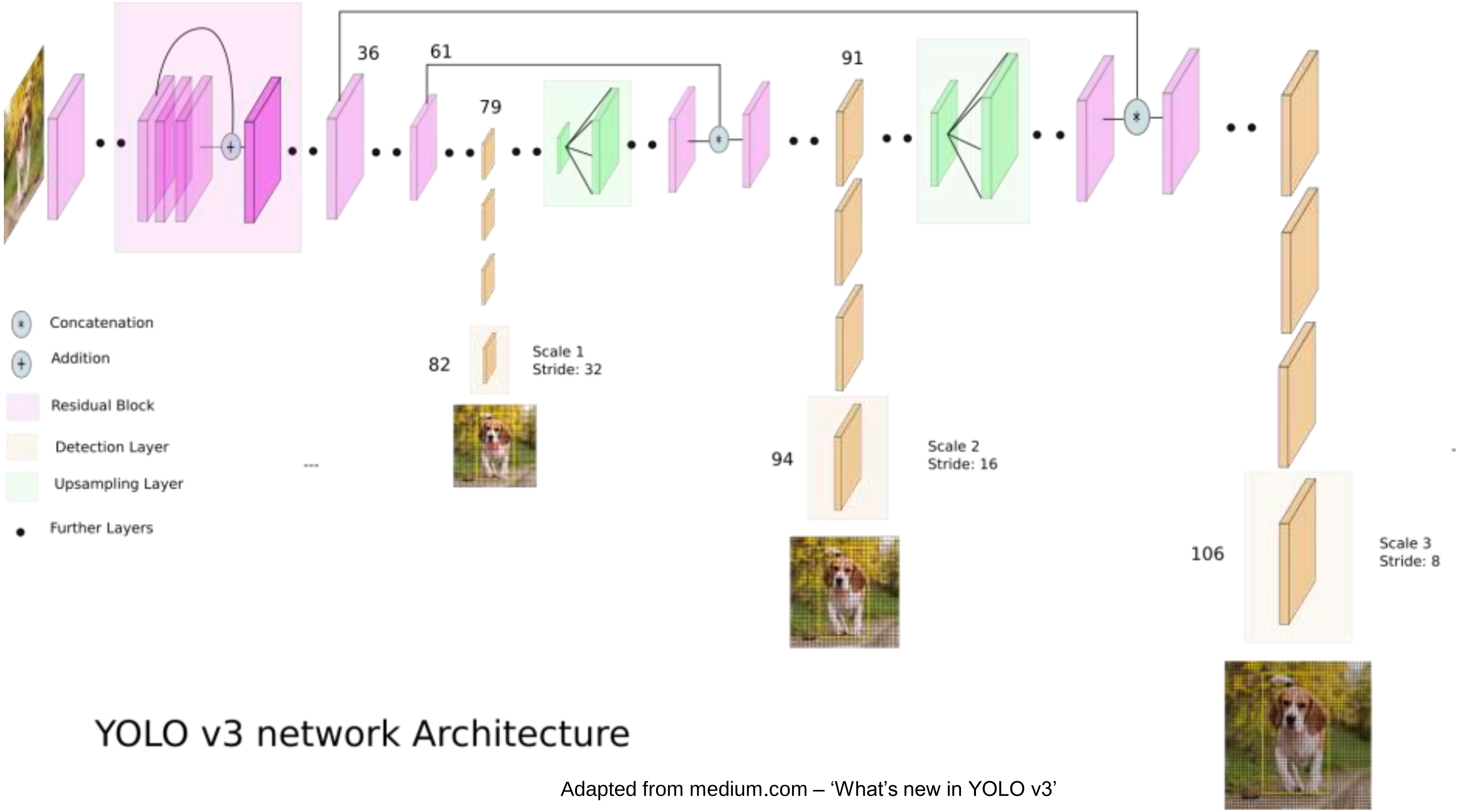
OBJECT DETECTION ALGORITHMS

- Modern Object detection algorithms are based almost exclusively on convolutional neural architectures
  - Easiest Problem: Image classification, e.g. cat/dog, t = (c1, c2, …, cN)
  - Object localization: where in the image is the object? t = (p, x, y, h, w, c1, c2, …, cN)
  - Object detection: Localize multiple objects in image
  - Image segmentation: classify each pixel according to class membership
  - Two types of networks: Single shot networks (e.g. YOLO) (fast!) and networks with region proposals (e.g. Faster RCNN) (precise!)
  - Method can be easily extended to include multiple information channels (c.f. multiple color channels)



Adapted from medium.com Jonathan Hui

EXAMPLE: YOLO



Adapted from medium.com – ‘What’s new in YOLO v3’

PERFORMANCE METRICS

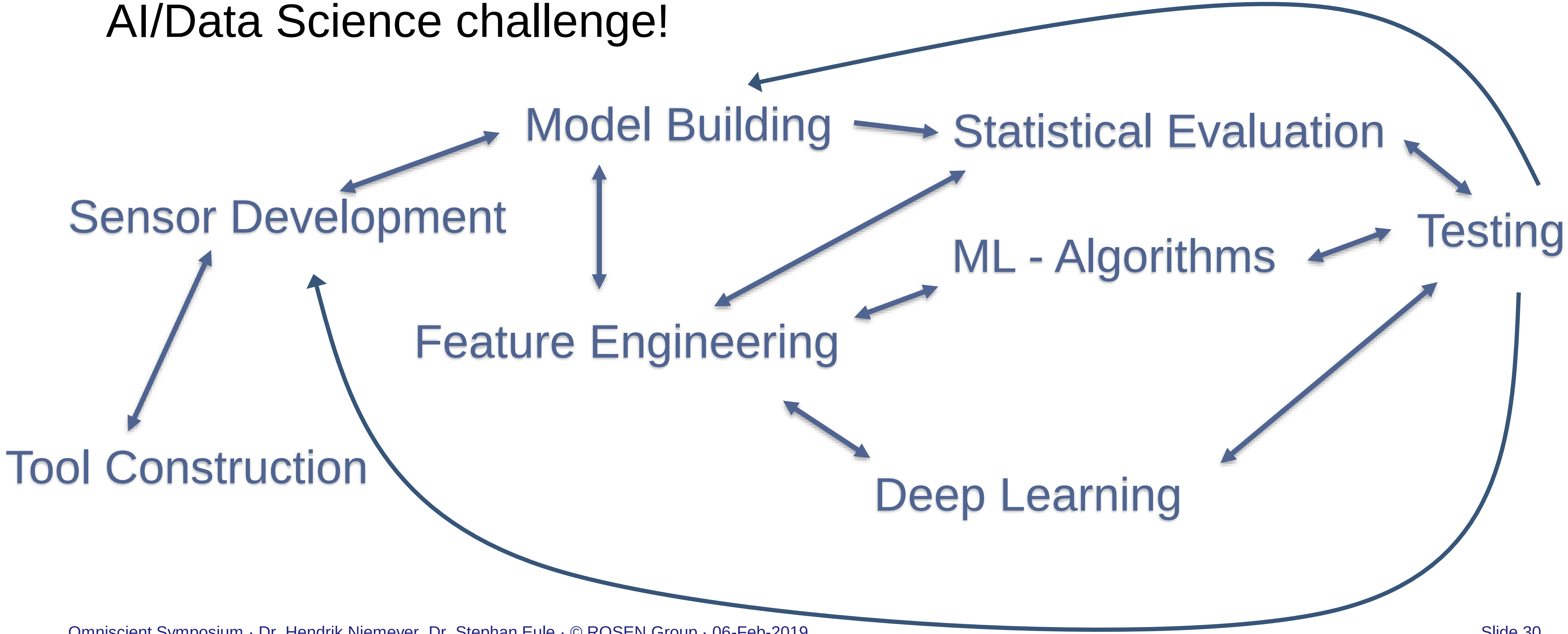
Maximize Precision: $\frac{TP}{TP+FP} = \frac{TP}{\text{all detections}}$

Constraint: Recall $\frac{TP}{TP+FN} = \frac{TP}{\text{all ground truth}} \cong 1$

Approach: First detect anomalies, second classify anomalies, last quantify anomalies

CONCLUSION

- Pipeline Inspection at ROSEN is an extremely interesting industrial AI/Data Science challenge!



THANK YOU FOR JOINING THIS PRESENTATION.

ROSEN
empowered by technology