https://ai4mi.epotentia.com

AI FOR MATERIALS INDUSTRY

ARTIFICIAL INTELLIGENCE

A MASSIVE OPEN ONLINE COURSE

Hands-on session 1

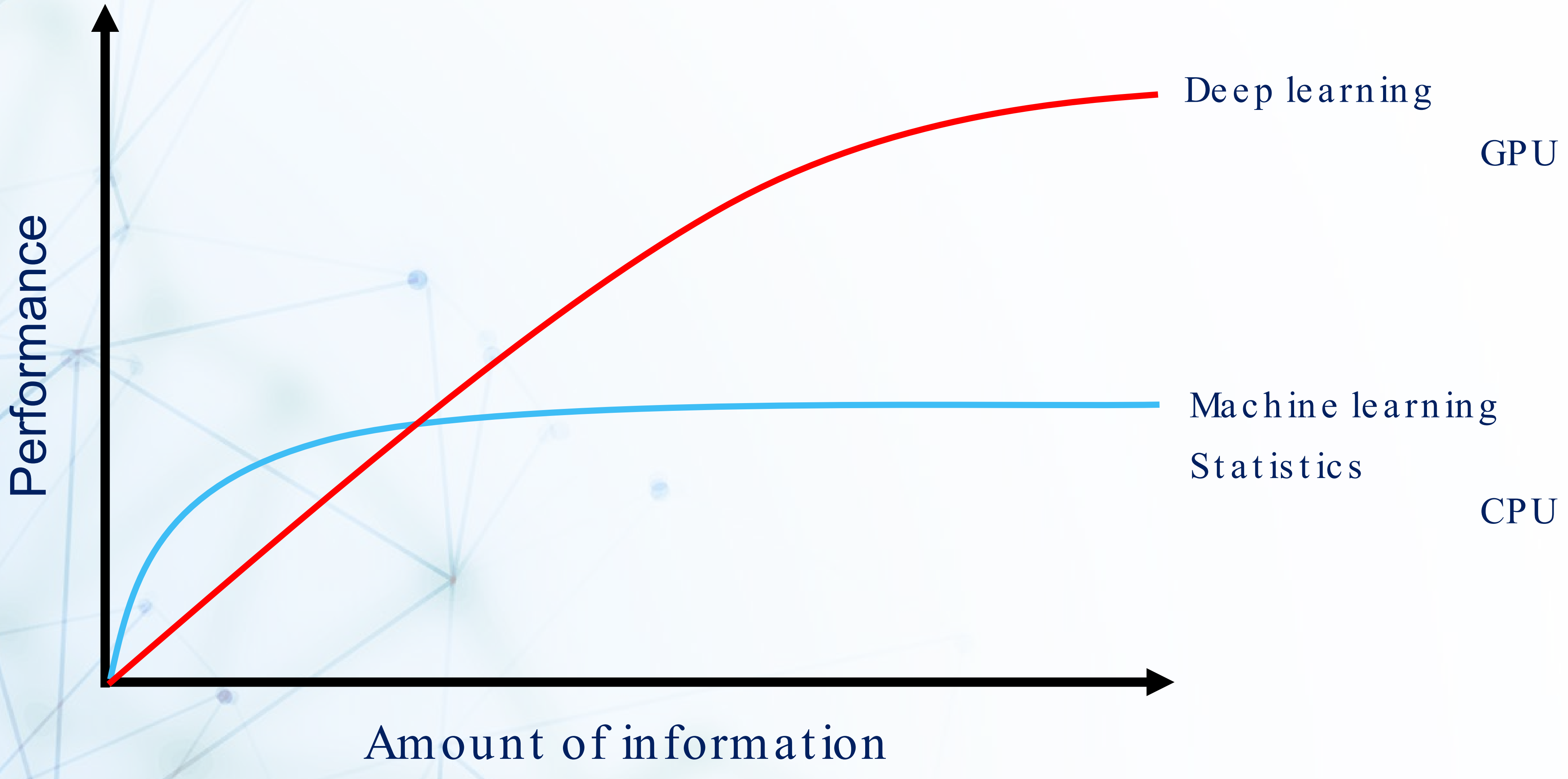
Steel defect classification with machine learning

AI FOR MATERIALS INDUSTRY

What is AI?

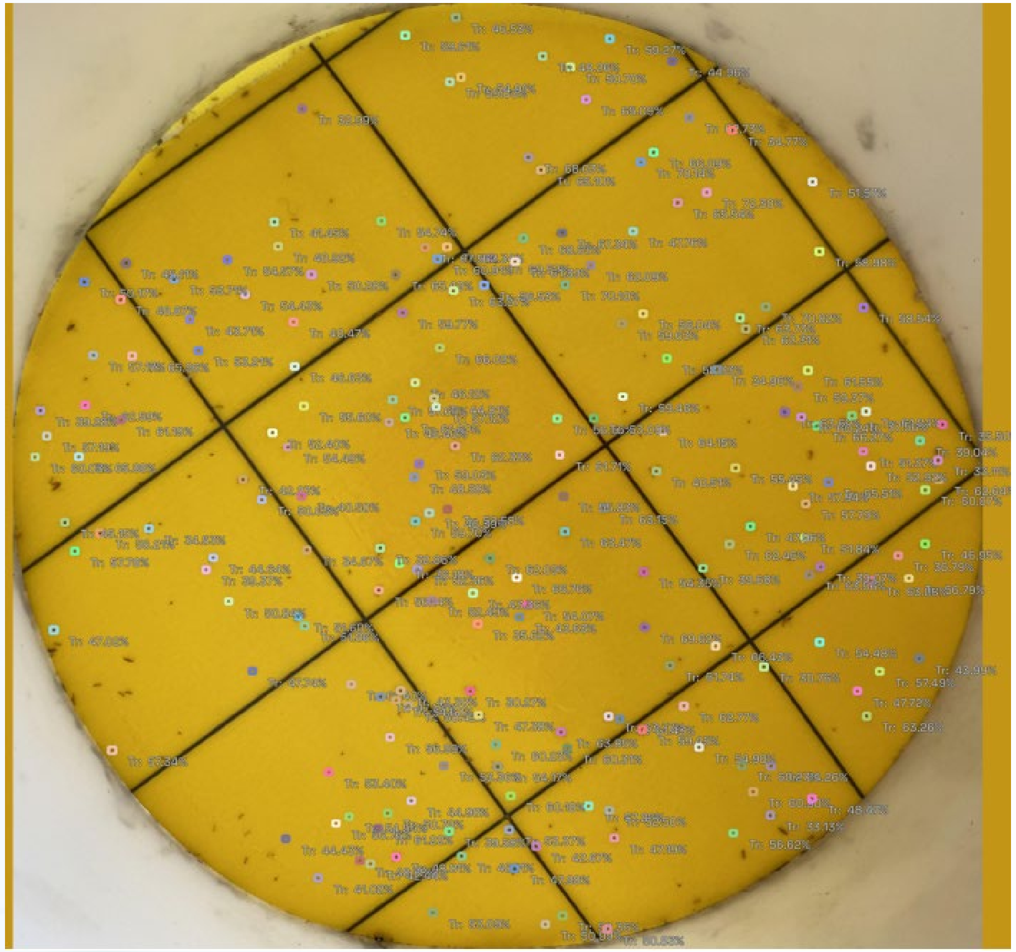
Series of techniques to extract **information** from data

Types of AI



The more data we have, the more we trade insight for predictive power

Example application: BioBest



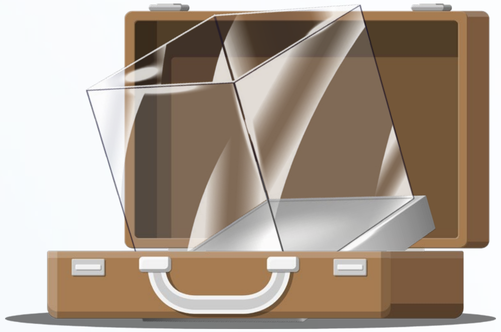
Case studies

Spreadsheet data

Steel

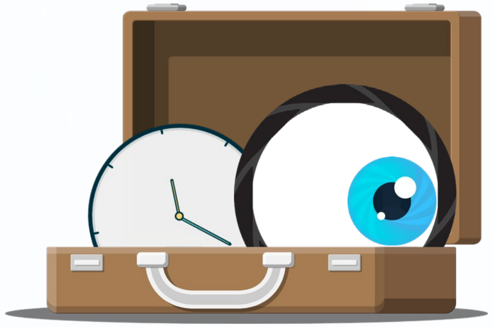
Computer vision

Glass



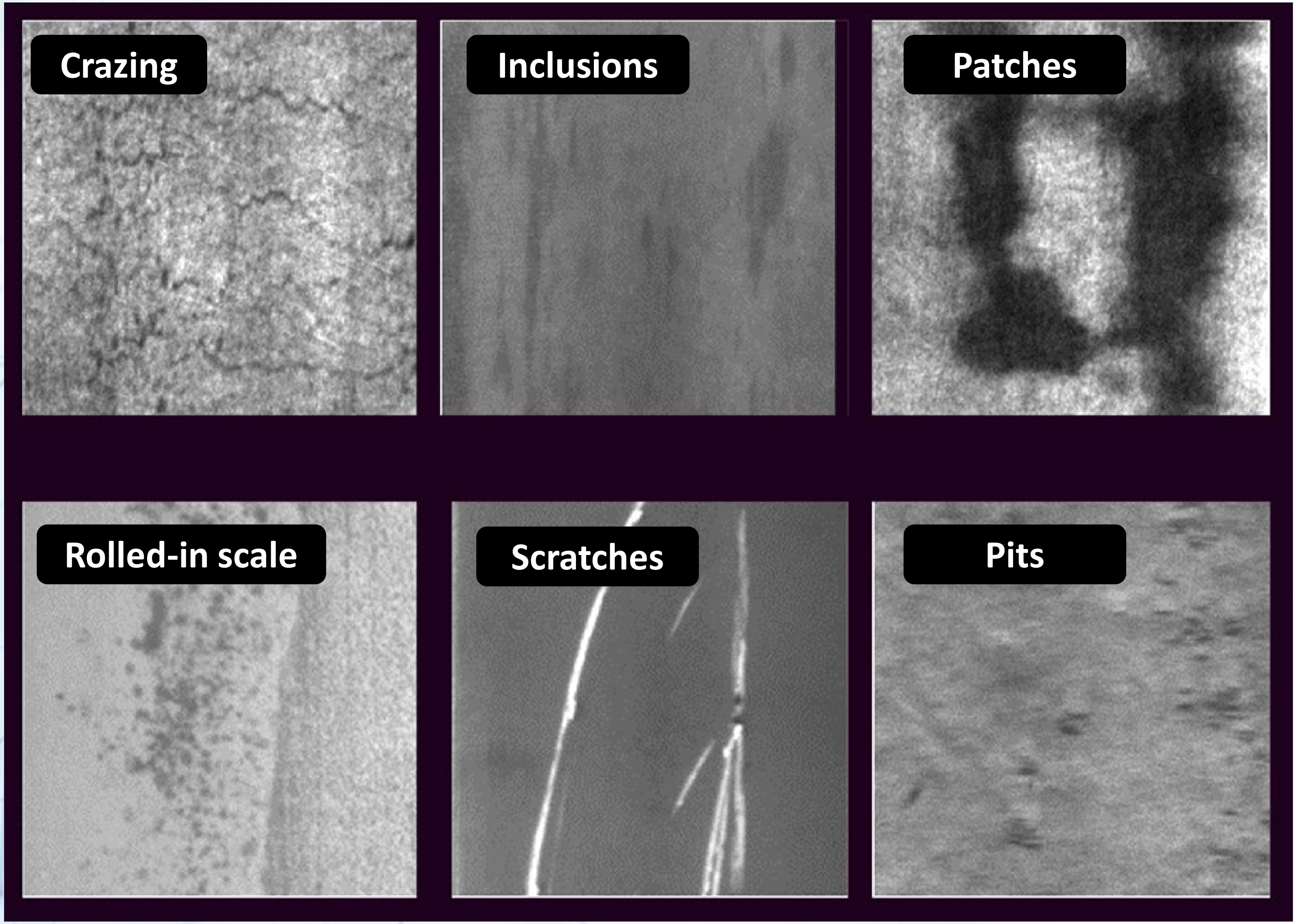
Materials discovery

Manufacturing

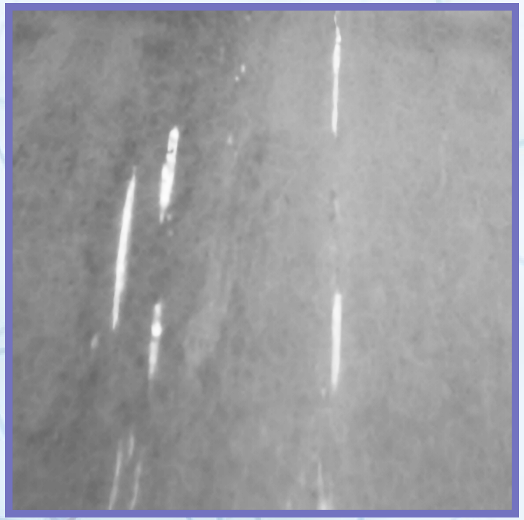


Sensor data

Steelplate defects



Multi-step pipeline



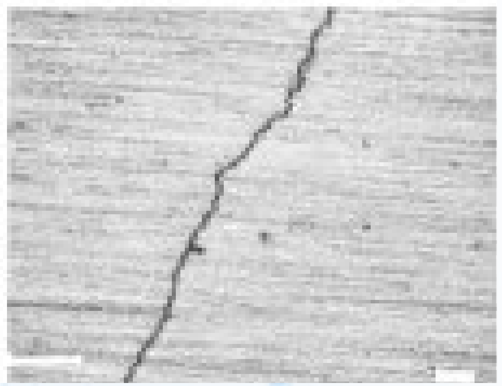
Detector → Featurizer → Classifier

Part of the pipeline needs to run in production

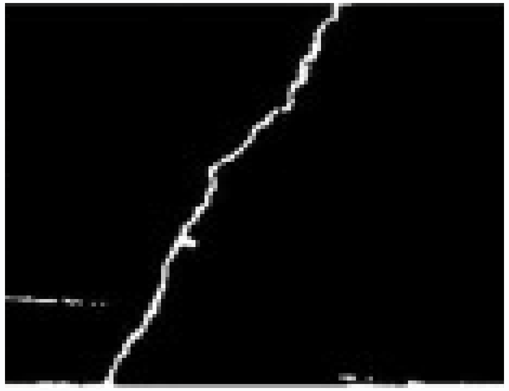
Detector

Original

Grayscale



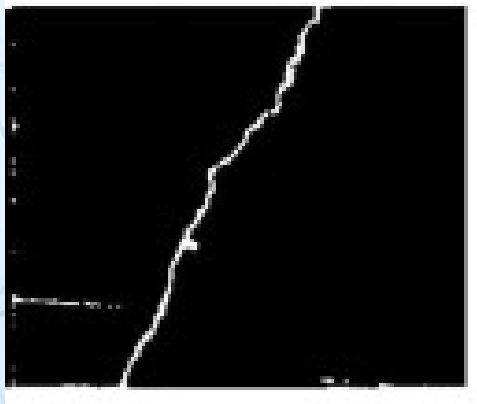
Threshold



Source: https://doi.org/10.1016/j.aej.2019.10.001

Simple and effective, but can be fickle

Featurizer



- Size and shape
- Brightness and contrast
- Materials properties
- Processing information

Features are numerical representation of the data created using our knowledge

This session

1. Analyze the features
2. (optional) Add more
3. Choose a model
4. Optimize the model
5. Evaluate the results
6. Interpret with explainable AI

Part of the pipeline needs to run in production

Requesting a job

Jupyter Lab

This app will launch a Jupyter Lab server on one or more nodes.

Cluster

dodrio cpu_rome

Time (hours)

12

Number of nodes

1

Number of cores per node

32 (quarter)

Mode

JupyterLab version

3.1.6 GCCcore 11.2.0

Custom code

```
source /dodrio/scratch/projects/explor_2022_008/init.sh
```

This code is executed before the JupyterLab is started. Primarily used for extra modules you want to load or swap.

Extra Jupyter Arguments

```
--notebook-dir=$PDIR/$USER
```

Project account

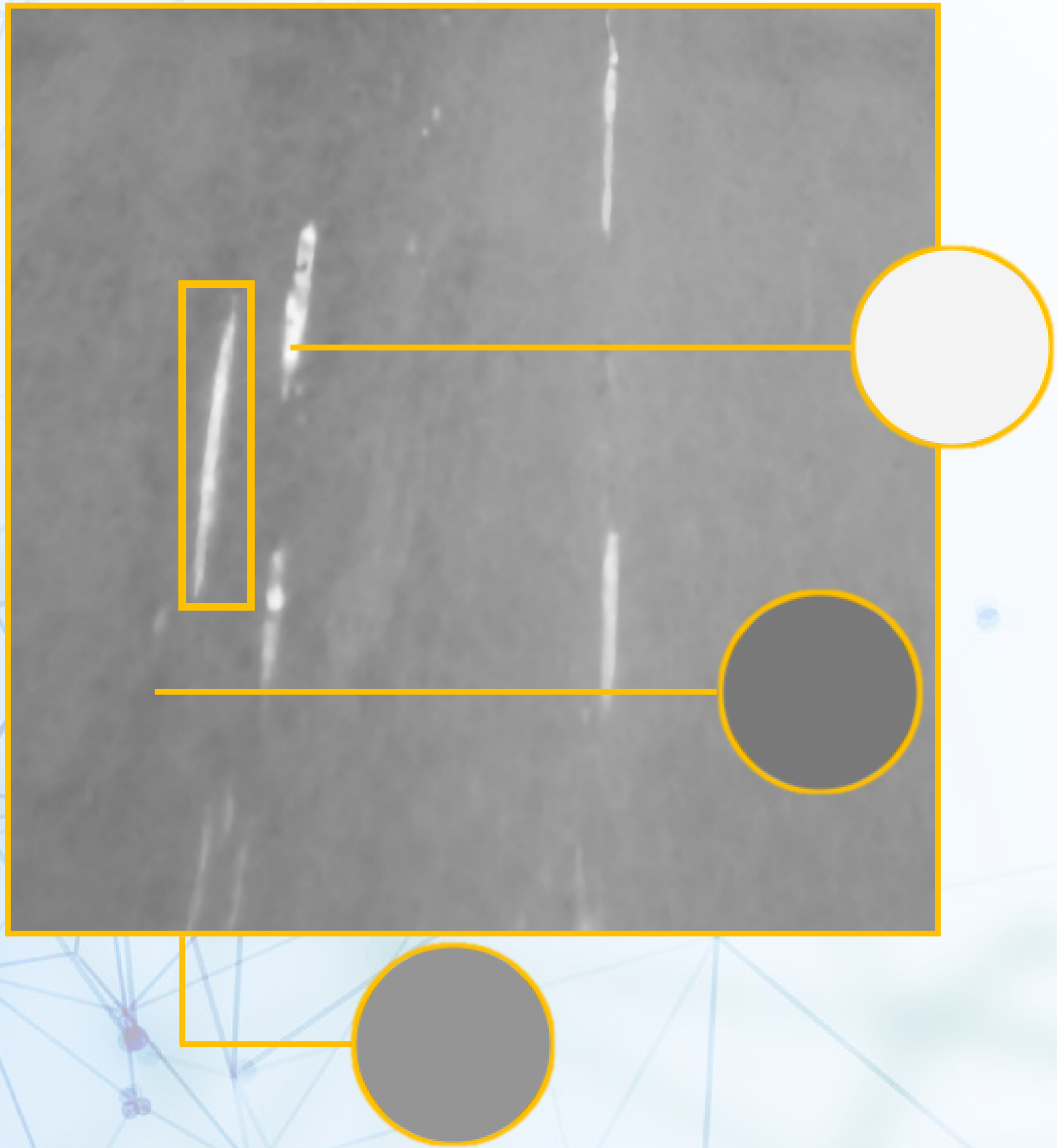
explor_2022_008

Extra sbatch arguments

```
--reservation=MaterialsAI-CPU
```

Log in at https://tier1.hpc.ugent.be

Dataset



- What is the dataset?
- Where did it come from?
- Who made it?
- What are the inputs?
- What are the outputs?

Tabular data

		Feature		
Datapoint		Value		

AI FOR MATERIALS INDUSTRY

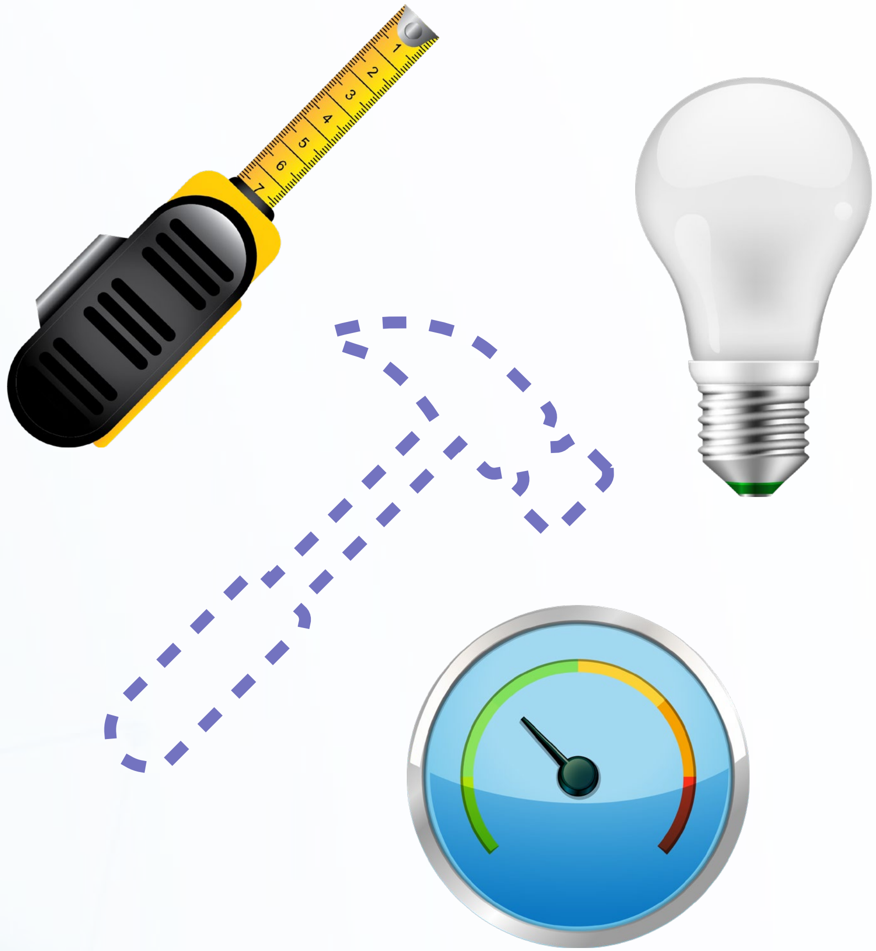
Python Pandas

pandas

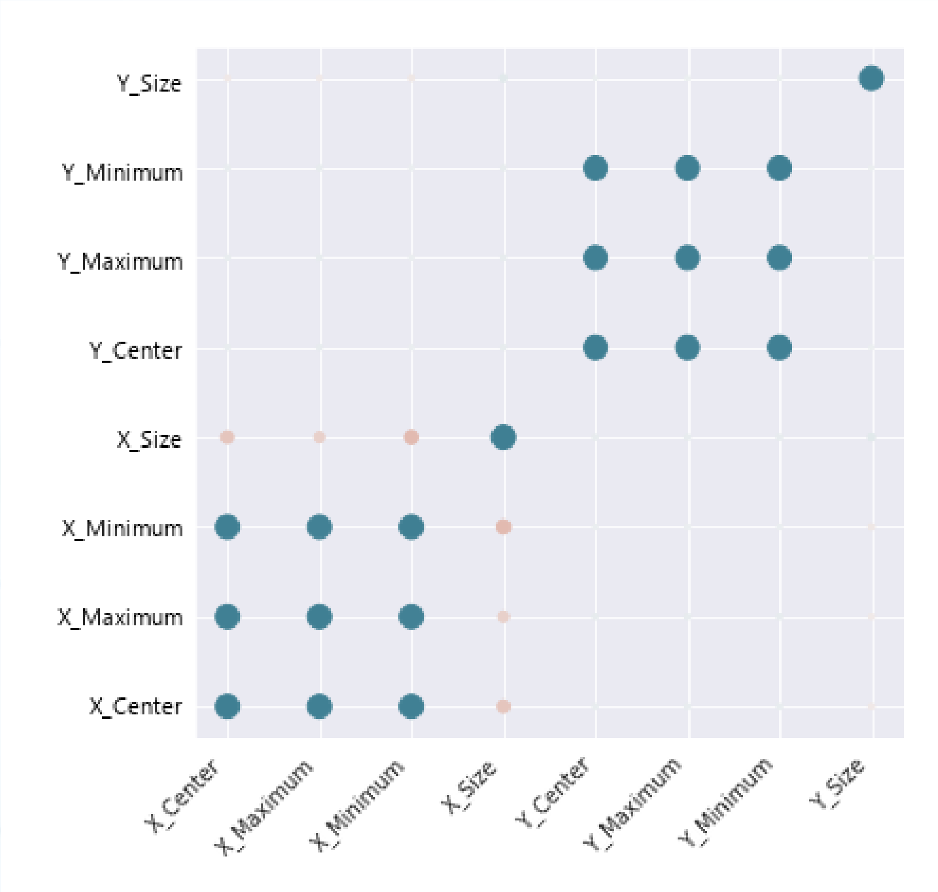
```
pip install pandas
```

Exploratory Data Analysis

- What kind of features are in the dataset?
- What is the range of the values?
- What do the features represent?
- Are they independent?
- Is there missing data?
- Are certain outputs rare or abundant?



Example: feature independence

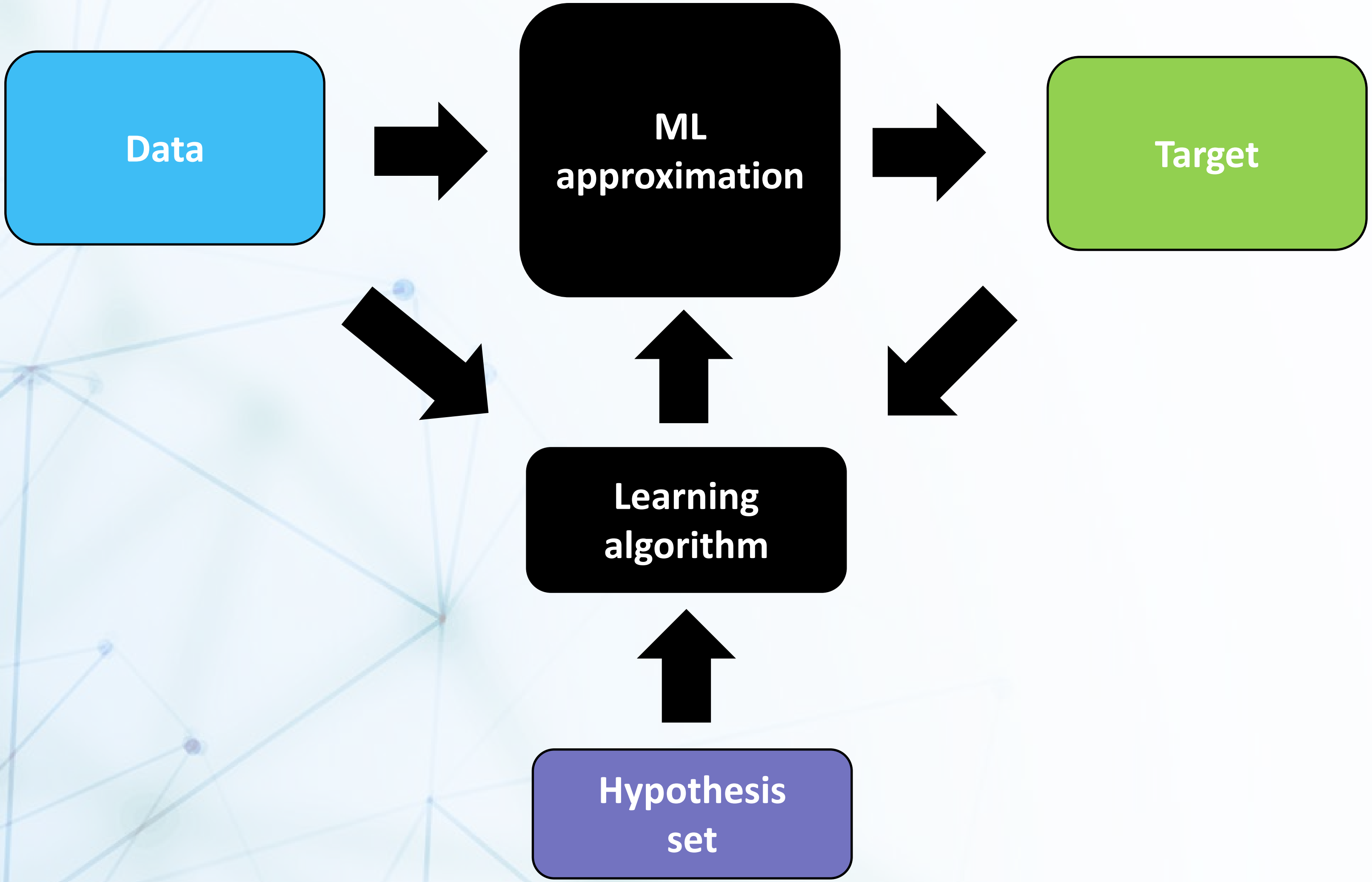


Strongly correlated features are opportunities for optimization

Machine learning ingredients

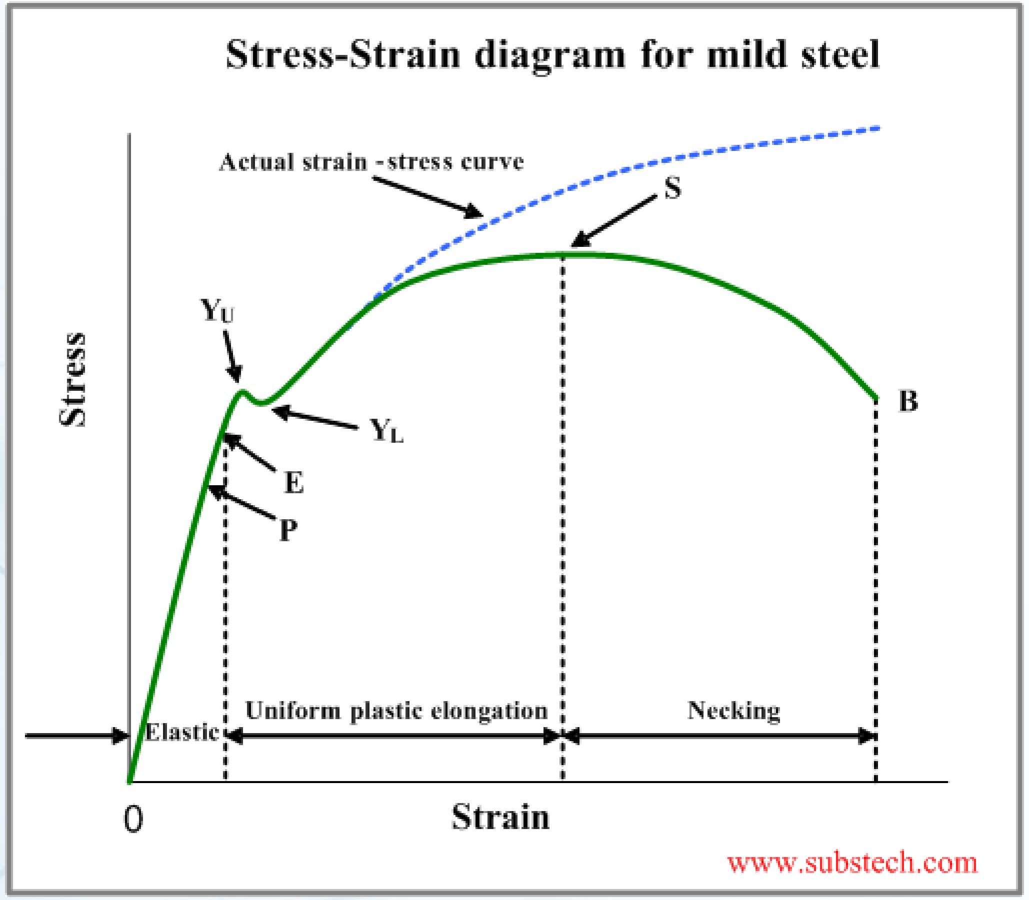
We must find the function that connects the inputs to the target

Machine learning ingredients



AI FOR MATERIALS INDUSTRY

Example: tensile curve



Type of representation can greatly affect the performance

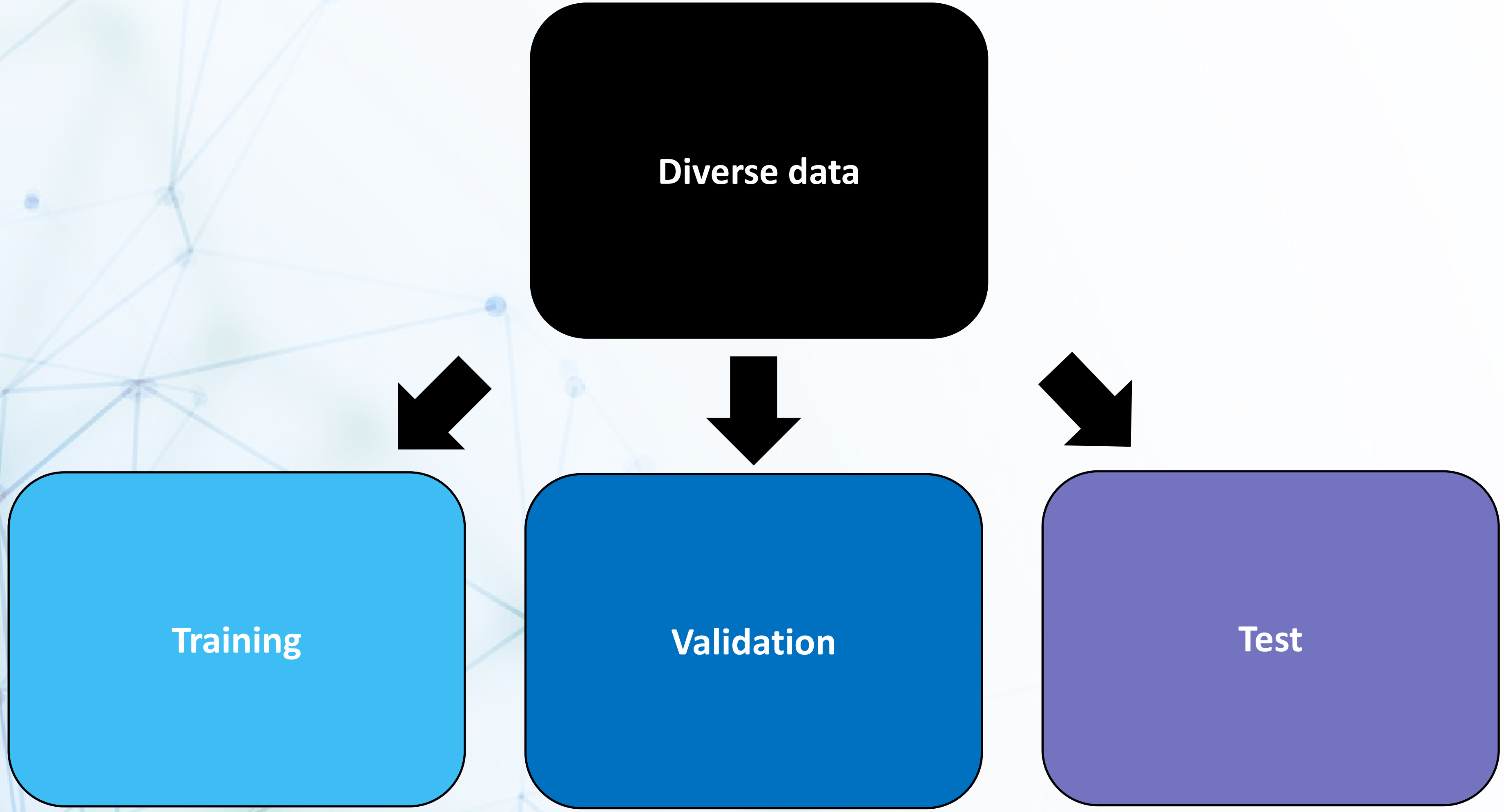
Constructing a dataset

Diverse dataset

- Materials
- Etchings
- Machines
- People

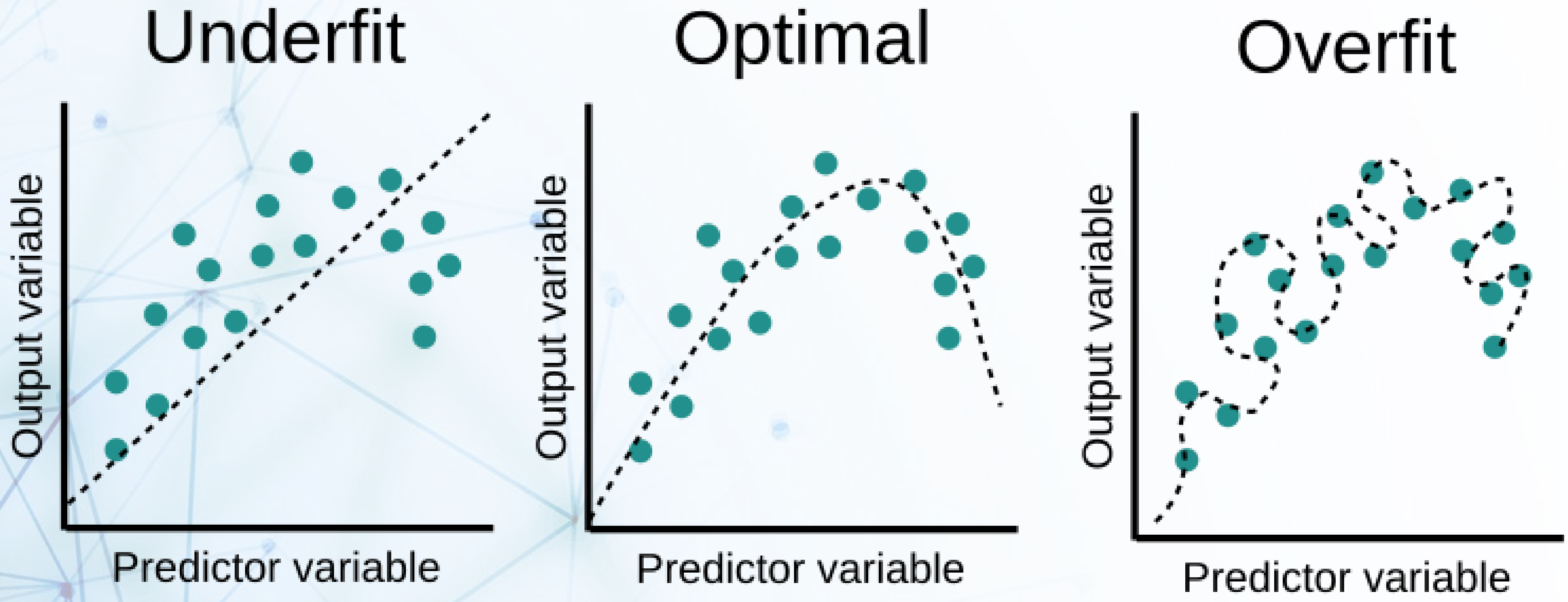
Collecting a good dataset takes time

Generalization



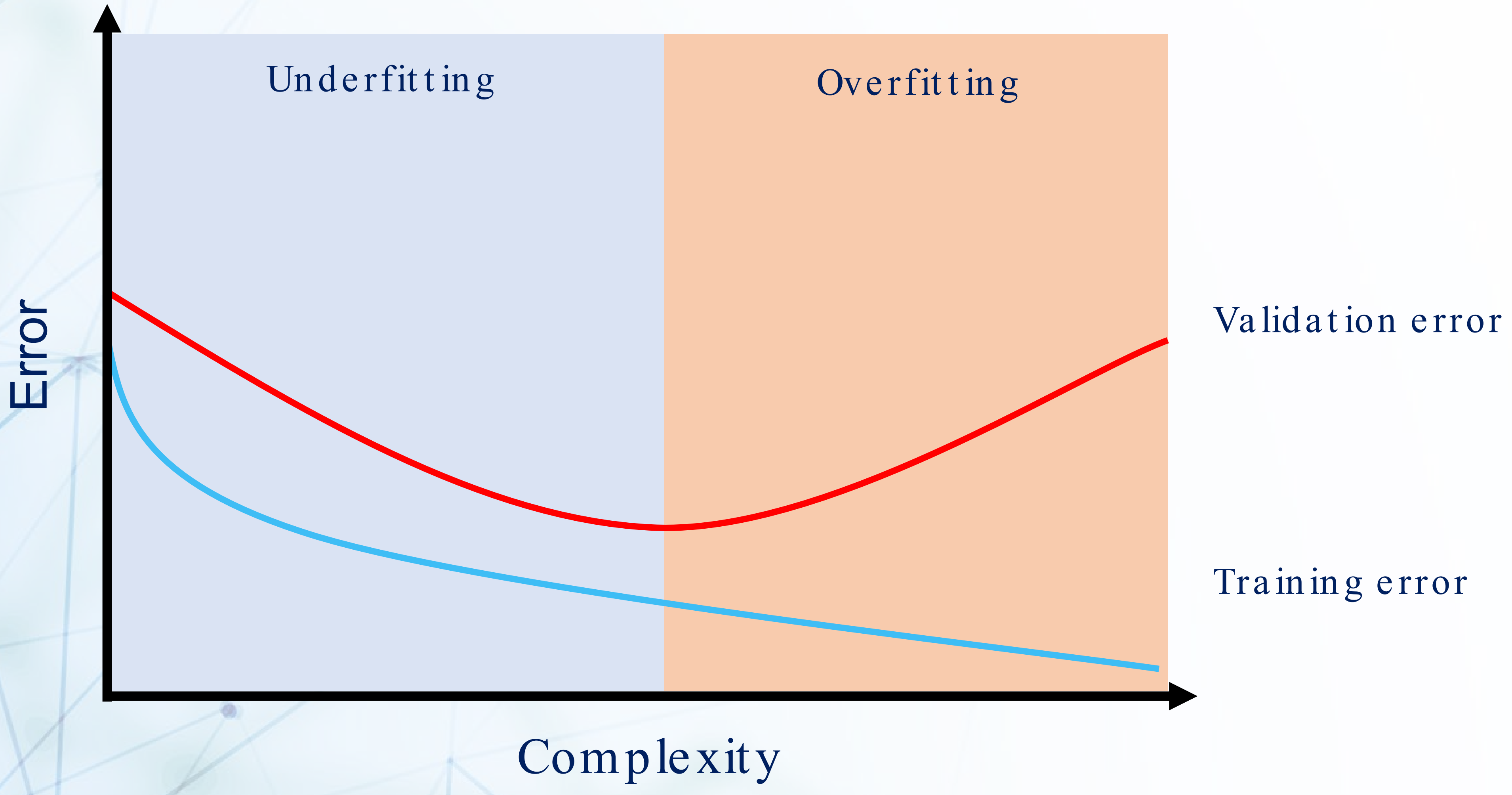
Be very careful for data leakage

Fitting your model



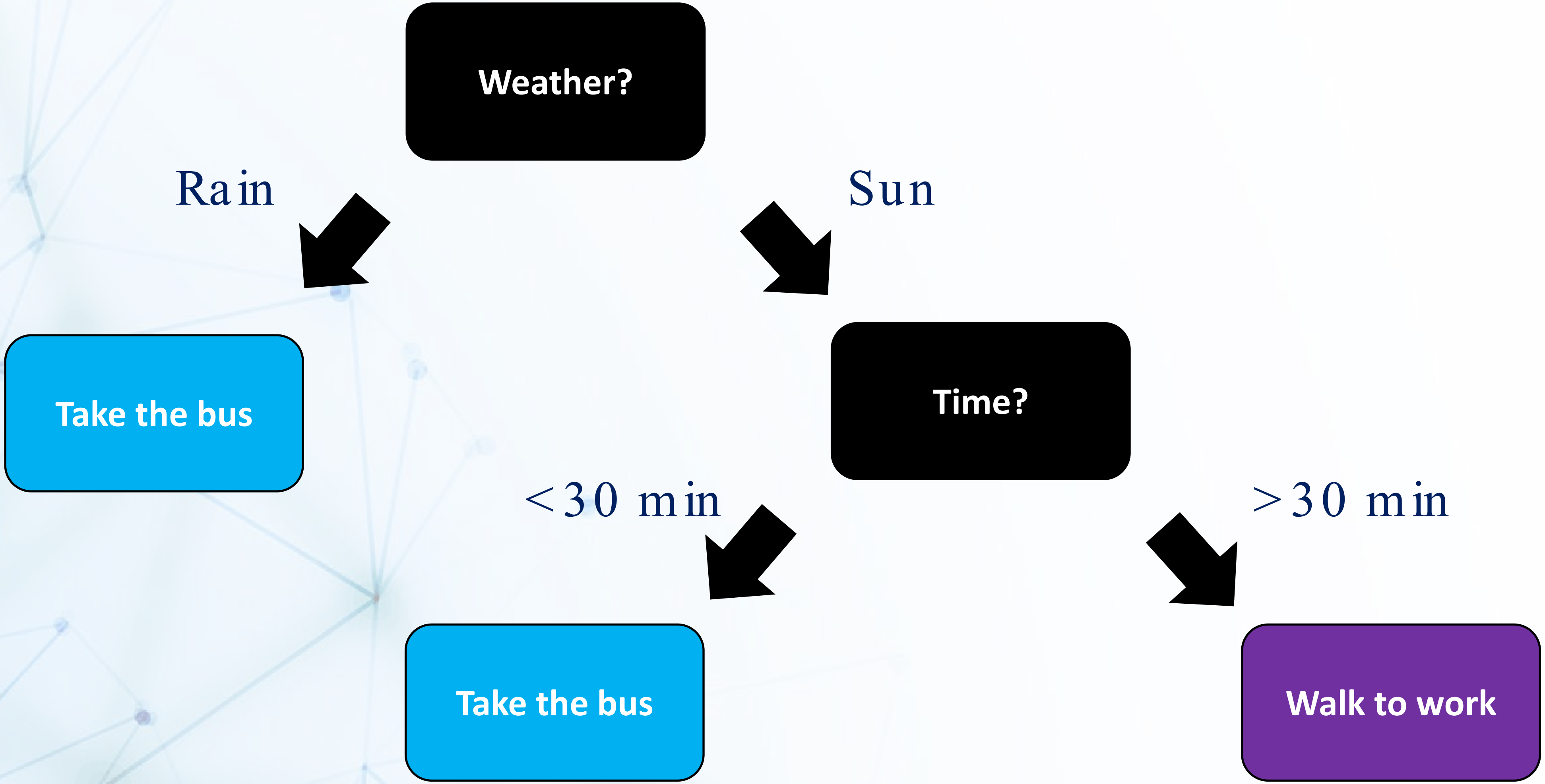
Choose a model of the right complexity and regularize where needed

Generalization



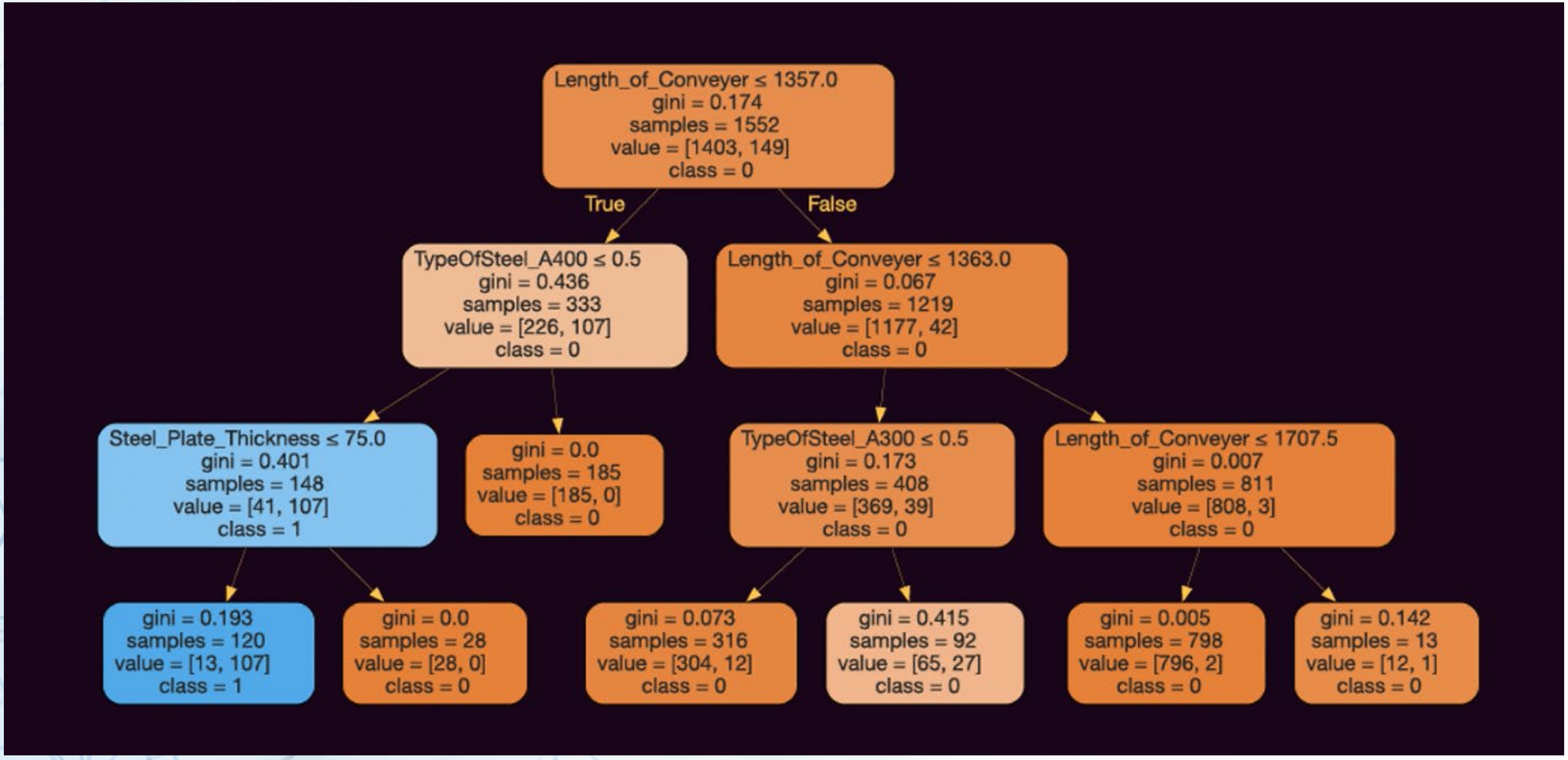
Regularization techniques can prevent overfitting

Decision trees



Robust, flexible in amount of data, controllable speed (CPU)

Bump decision tree



A single tree greatly risks overfitting

Random forests

By taking many simple trees, we average out the error and prevent overfitting

Out-of-bag score

N samples are chosen for each tree with repetition (bagging)

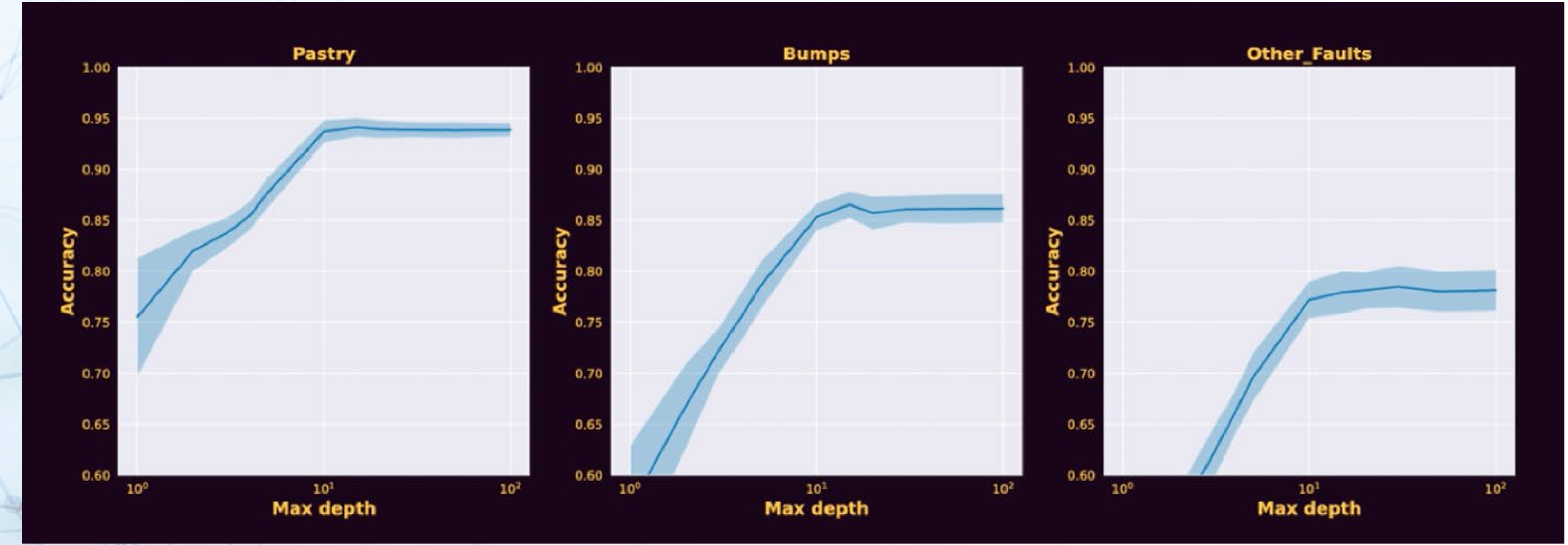
36,2% of the data is unseen by each tree which can be used for OOB validation

Hyperparameters

• Number of trees	`n_estimators`
• Depth of the trees	`max_depth`
• Amount of data per tree	`max_samples`
• Number of features per tree	`max_features`
• Samples per split	`min_samples_split`
• Samples per leaf	`min_samples_split`

Perfect for high-throughput screening

Example: tree depth



Deeper trees are better, but slower and risk overfitting

Accuracy

- **Decision trees** can split the dataset down to **individual samples**
- 100% accuracy on training is easy, almost trivial
- This doesn't mean good performance "in the wild", the model might have just learned to recognize distinct noise patterns that don't generalize
- This is why we need a validation and test set: a fair assessment

Metrics

- **Accuracy:** How often did we get it right? Gives no information about what type of mistakes are being made
- **Confusion matrix** : Very useful tool to understand what's going on, showing true positive (TP), false positive (FP), false negative (FN) and true negative (TN)

		Real data	
		+	-
Predicted data	+	TP	FP
	-	FN	TN

Metrics

		Real data	
		+	-
Predicted data	+	TP	FP
	-	FN	TN

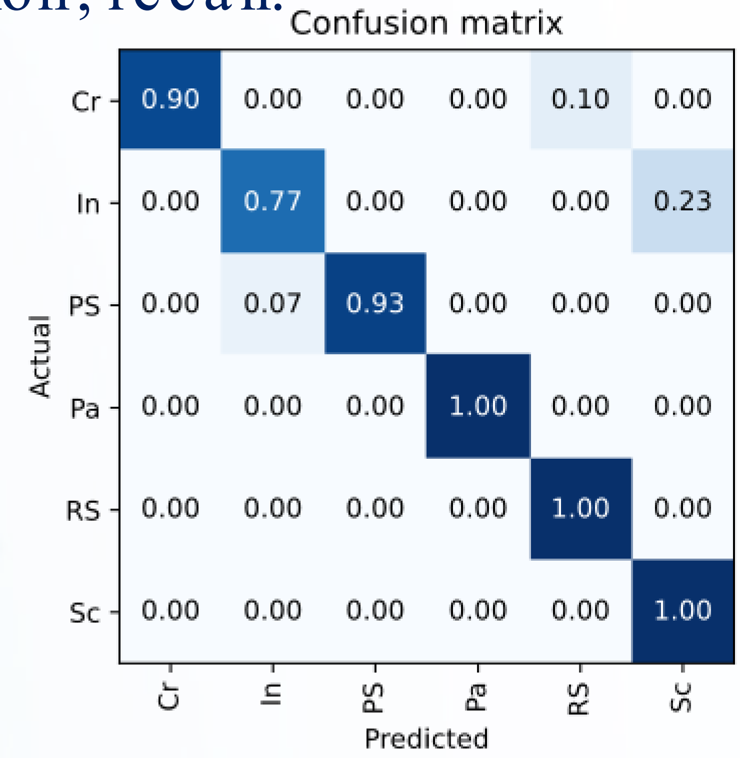
- **Accracy** : How often did we get it right? (TP and TN)
- **Precision** : Quality, how many of the positive samples are real? (FP)
- **Recall** : Quantity, How many of the positive samples did we find? (FN)

Choosing the right metric for your goal is important

Metrics

- The **confusion matrix** is more general than just true/false, can give insight in whether certain classes get confused for each other

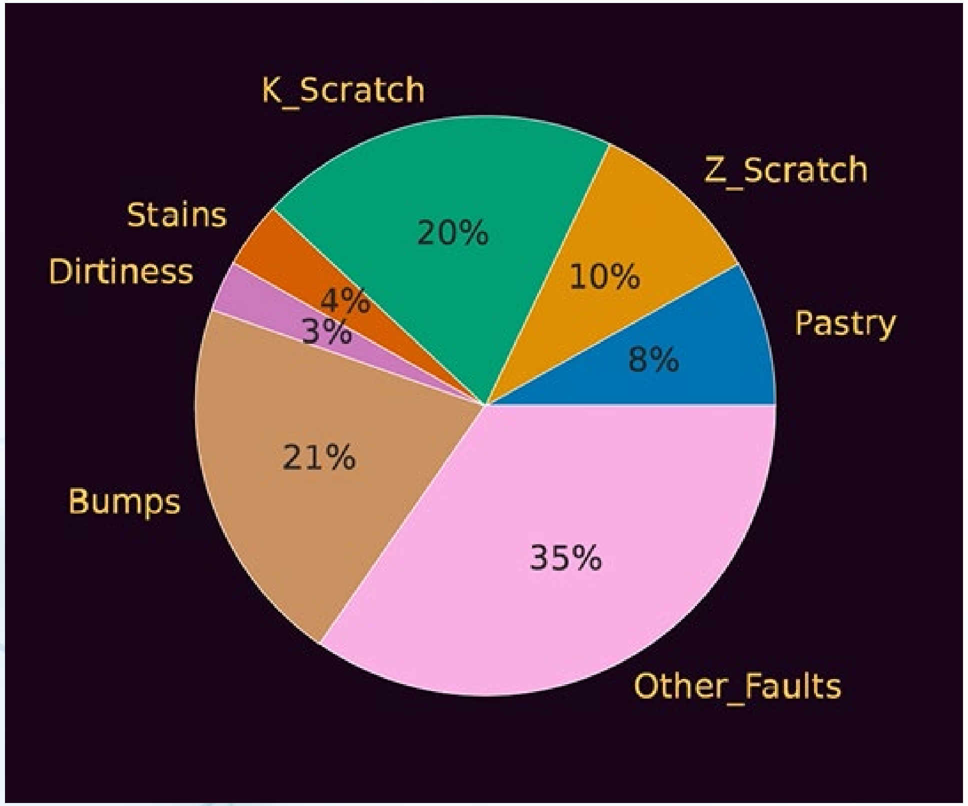
- **Other metrics** exist beyond accuracy, precision, recall.



- **F1-score:** geometric mean between precision and recall, good middle ground

- **Matthews Correlation Coefficient (MCC):** Cross-correlation between the real data and predicted data, ranges from -1 (anticorrelated) to +1 (correlated)

Results



Optimized	Pastry	Z Scratch	K Scratch	Stains	Dirtiness	Bumps	Other Faults
Accuracy	94%	98%	99%	100%	99%	88%	80%

Choosing the right metric for an application is important

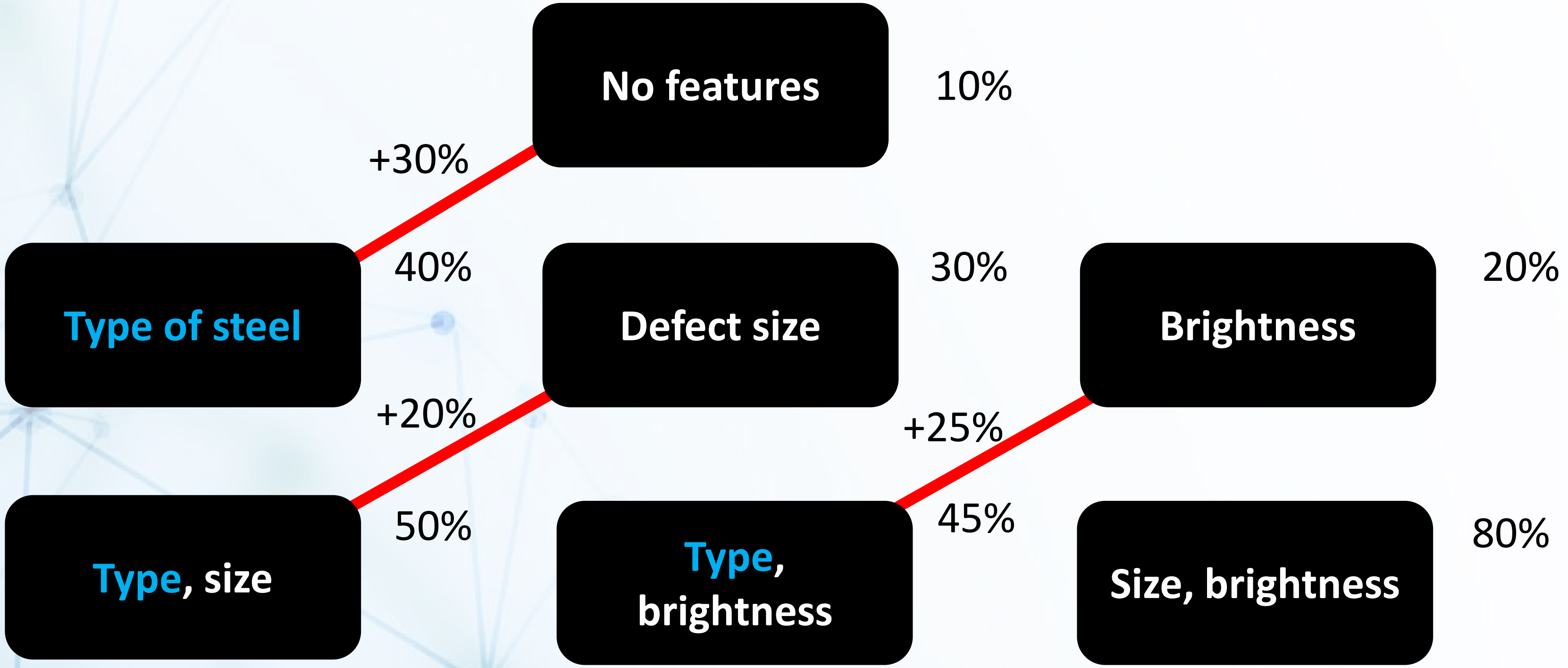
Explainable AI

- The model uses **features** to make predictions
- How **important** is a feature? Remove it and find out!
- **Drop-column feature importance:**
  1. Remove feature
  2. Retrain model
  3. Compare
- **Permutation feature importance:**
  1. Shuffle feature
  2. Make prediction with same model
  3. Compare

Explainable AI

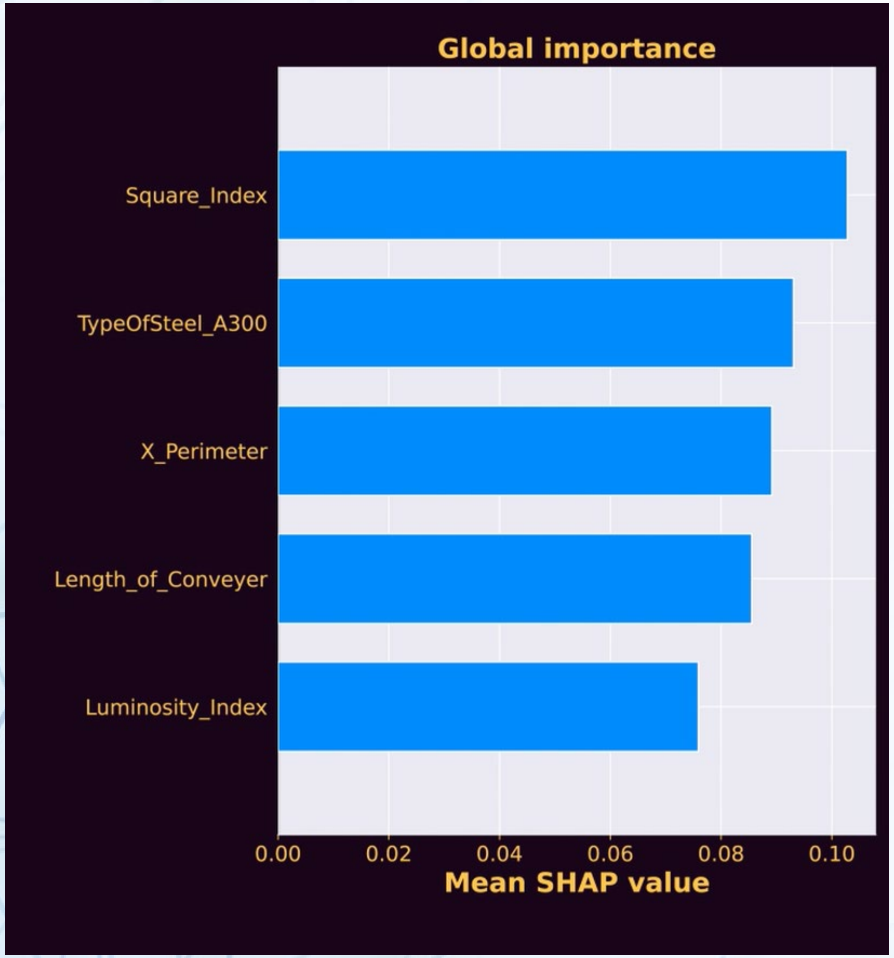
Features	Accuracy
No features	10%
Type of steel	40%
Defect size	30%
Brightness	20%
Type, size	50%
Type, brightness	45%
Size, brightness	80%

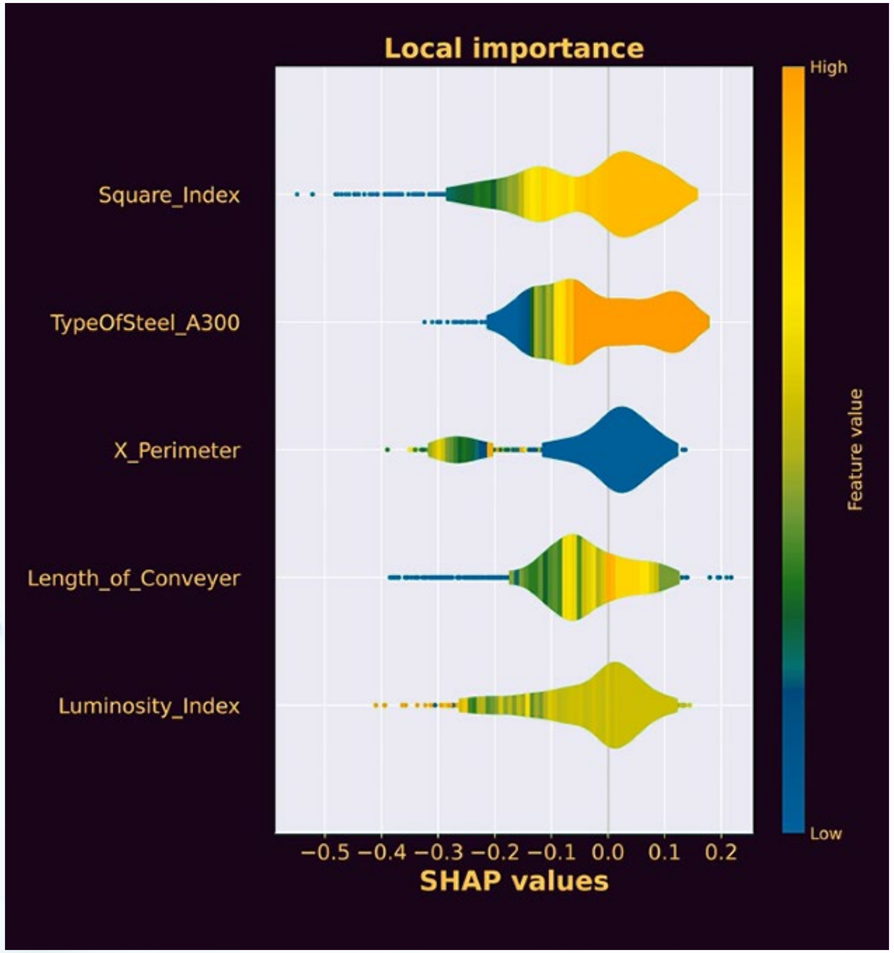
Explainable AI



If we could simulate all possible models we could evaluate the impact -> SHAP

Practical SHAP for bumps





https://github.com/slundberg/shap

SHAP helps us understand how a decision is reached

INVESTEERT IN
JOUW TOEKOMST

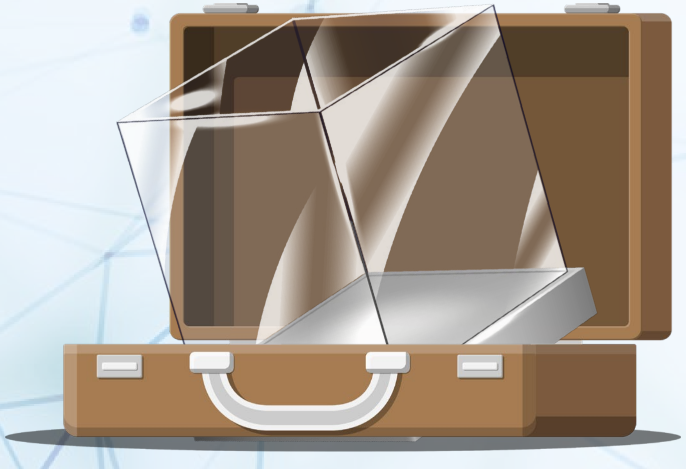
ESF Europese Unie Vlaanderen is werk

AI FOR MATERIALS INDUSTRY

View more online

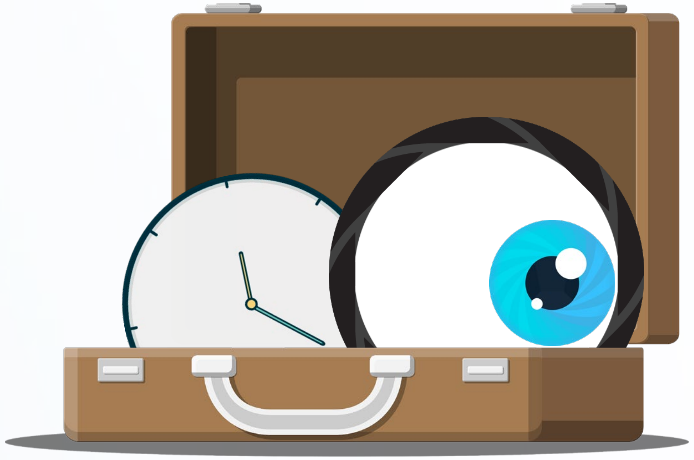
https://ai4mi.epotentia.com

Glass



Materials discovery

Manufacturing



Sensor data

GHENT UNIVERSITY

ePotentia
Data•Science•Hosting

ocas

umicore
materials for a better life

Sensor data

AGFA Agfa