FACULTEIT INGENIEURSWETENSCHAPPEN
EN ARCHITECTUUR

VAKGROEP INFORMATIETECHNOLOGIE
IDLAB

Gastles - ENSEMBLE METHODS

dr. ir. Joeri Ruyssinck

Personal introduction

Nice to meet you

Educational background

Ghent University: Software engineer – 2011 | AI for robotics
Ghent University: PhD - 2018 | AI for Biological Network Inference

Data-efficient AI research group

Professional background

imec – Senior AI Engineer and Project lead– 2017-2024
Ghent University– Postdoc and lecturer– 2018-2024

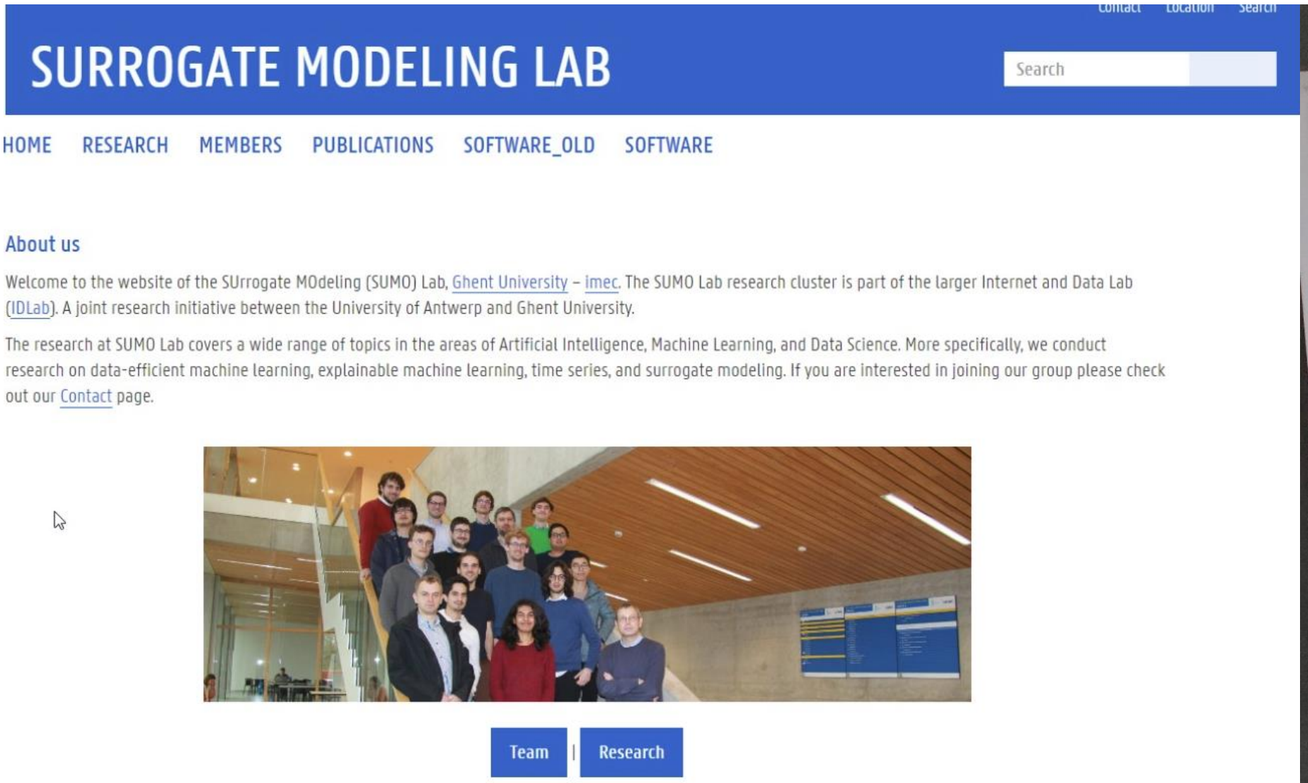
ML2Grow – June 2024

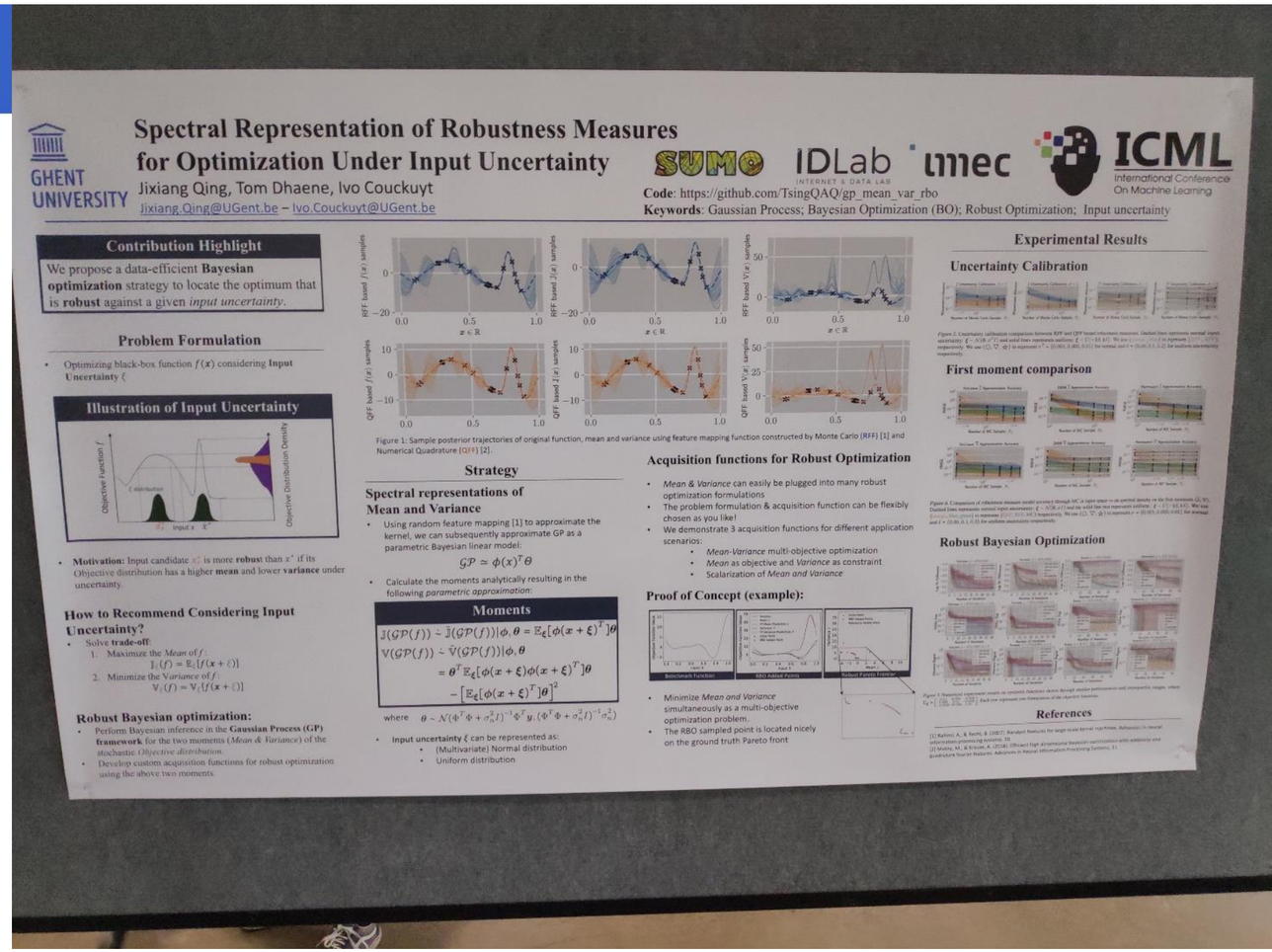
Entrepreneurial background

ML2Grow– ML engineer and CEO– 2017

Joeri Ruyssinck

Surrogate modelling lab (SUMO) – part of IDLab





ML²GROW

Data-driven yield optimization

Tomatomasters (Deinze) – large tomato greenhouse!

Horticulture is a very competitive sector with a trend for more manual and sensor data-gathering

PHYTOSTEM UNITS

De plant sensoren bestaan uit enerzijds een **sapstroom-sensor** en anderzijds een **diameter-variatie-sensor**.

PHYTOCLIP UNITS

Sinds kort hebben we temperatuur gecorrigeerde **micro sensoren** ter beschikking!

2GROW

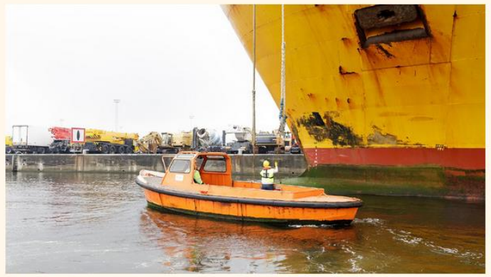
DE TIJD

Nieuws Markten LIVE Netto

NIEUWS > ONDERNEMEN > TECHNOLOGIE

Antwerpen lanceert 'Waze voor de scheepvaart'

TWITTER
FACEBOOK
WHATSAPP
LINKEDIN
E-MAIL
BEWAAR
SCHENK DIT ARTIKEL
REAGEER



Een bootman in de haven maakt een schip vast aan de kaai ©STUDIO CLAERHOUT

MARC DE ROO | 18 februari 2020 16:56

Elk schip dat te lang in de haven blijft, kost geld. Het havendienstenbedrijf Port+ en het loodsenbedrijf Brabo zetten algoritmes in om schepen zo vlug mogelijk in en uit de haven te krijgen.

Artificiële intelligentie

Binnenkort gaat Brabo nog een stap verder. Via artificiële intelligentie wil het voorspellingen maken van het werk dat op het bedrijf afkomt. En dat op basis van data uit het verleden die rekening houden met de tijdstippen waarop een schip wordt gesleept of aankomt in de sluis en aan de kaai'. De Groof: 'Heel belangrijk, als je weet dat we binnen het uur na een oproep een schip moeten kunnen beloodsen. Voor een werkleider is het onmogelijk acht uur vooruit te kijken. 26.000 scheepsbewegingen per jaar is gigantisch. De testen vorig jaar waren veelbelovend.'

'Elke periode dat er met een schip 'niets gebeurt', kost geld'

Deel op

JAN VAN DOOREN
CEO PORT+

Het systeem moet helpen om capaciteits- en personeelstekorten beter in te schatten, automatisch te plannen en if-scenario's in te stellen als iets erg voorvalt. 'Nu zetten we personeel in op basis van pieken', zegt De Groof. 'Het gaat permanent om 40 bootsmannen en 20 loodsen, 24 op 24 uur, zeven op zeven. Soms hebben die geen werk. Met AI kunnen we onze mensen accurater inzetten. In 2018 hebben we 8.000 keer mensen dringend thuis moeten oproepen, nu is ons streefdoel 5.000.'

'Brabo is een van onze beste voorbeelden van hoe je met kleine AI-ingrepen grotere veranderingen kunt teweegbrengen', zegt Joeri Ruyssinck van ML2Grow, dat de AI-toepassing bij Brabo bedacht. 'Met een minimum aan investeringen heb je een quick win.'

A port call can be compared to a F1 pitstop. Various stakeholders each must perform their own activity on a vessel, often in a specific sequence.

BRABO GROUP

port+

NIEUWS > ONDERNEMEN > TOERISME

Camera's aan de kust kunnen meer dan mensen tellen

De camera's worden opgehangen aan de zeedijk en in drukke winkelstraten. ©Dieter Telemans

WIM DE PRETER | 10 juni 2020 18:27

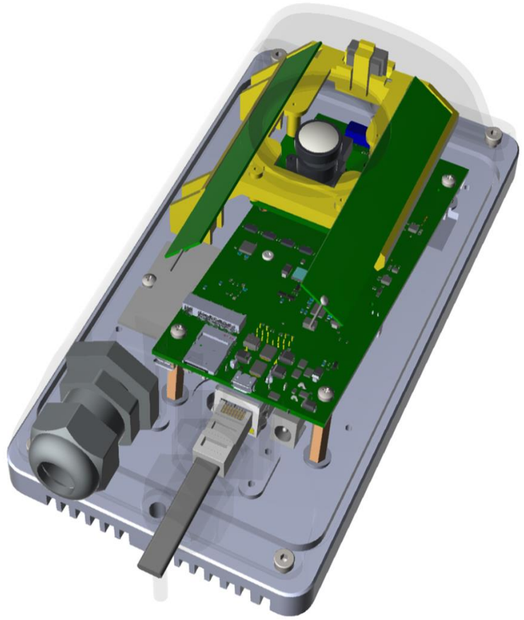
De 250 nieuwe camera's aan de kust zijn in eerste instantie alleen bedoeld om anoniem mensen te tellen. Toch zijn ze, met een laagje extra software, tot meer in staat.

De plannen van de kustgemeenten om deze zomer de drukte te meten met een netwerk van slimme camera's lokte woensdag een verontruste reactie uit van de Gegevensbeschermingsautoriteit (GBA), de opvolger van de vroegere Privacycommissie.

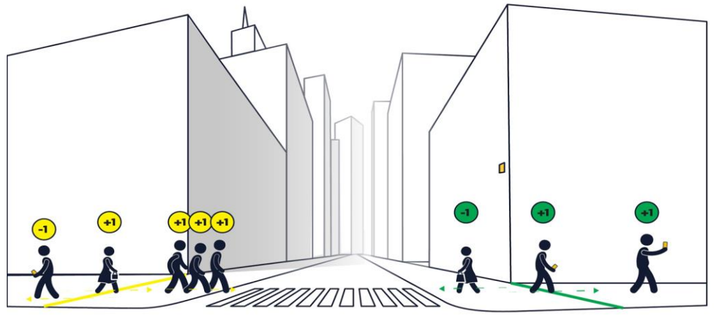
WHAT IS FOOTFALL AI?

CITYMESH

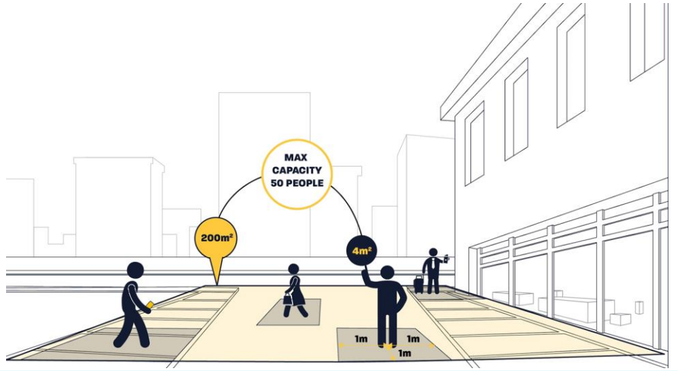
Footfall AI **accurately and automatically** maps your visitor flows. In many instances, counting visitors is still done manually. This is extremely time-consuming and often inaccurate. Qualitative footfall data is therefore **real white whale data**: rare and difficult to collect without the right methods. That is why Citymesh has developed Footfall AI. With this smart solution you get **accurate data that is displayed in a structured way in a dashboard**.



COUNTING TRIP LINES



COUNTING AREAS



Context

Social employment company, non-profit.
Focus on re-introducing goods into the second-hand market.Currently employing around 5000 people in Belgium

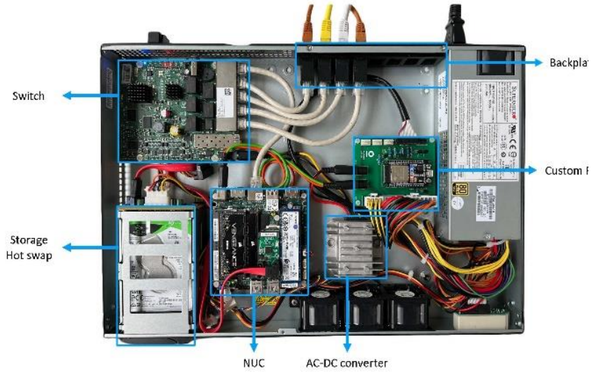
Challenge

Even with social employment and a non-profit viewpoint. It is impossible to give growing volumes of cheap textiles a second-life.

Innovative idea

Make use of AI computer vision systems

- To characterize clothing and textiles
- To spot defects or characterize wear



ML2Grow's Vision Hub

“A.I. is the new electricity”

Just as electricity **transformed industry after industry 100 years ago**, today I actually have a hard time thinking of an industry that I don’t think **AI will transform in the next several years**.

A clear mission and vision

At ML2Grow, we are convinced AI is a **breakthrough technology** that must be value enabling and accessible for **all organizations,** regardless of the organization’s size, sector or purpose.

Our expertise and platforms increase the adoption chance, raise the project success rate and optimize the return on investment.

Andrew NG,

CEO Deeplearning.ai, Landing AI, Partner AI fund , co-founder Coursera , Google Brain, ex-CSO Baidu, Stanford professor– 2017

Being the first to have electricity doesn’t matter

Without the

- ✓ infrastructure to deliver it
- ✓ the human capital to manage and improve it
- ✓ the standards to commercialize it

So are you ready for exponential AI adoption?

Benjamin Franklin and the kite experiment	*1752*	127y
Thomas Edison and the lightbulb patent	*1879*	41y
Half of American homes have electricity	*1920*	

AI is a general-purpose technology

The correct business goals, innovation appetite, change management and technical expertise come together to create impact

CONTENT

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

This lecture:

- What are ensemble methods? Why do they work?
- Different plan-of-attack of ensemble methods
  - Reducing variance error of base models: e.g. bagging methods (Random Forests)
  - Reducing bias error of base models: e.g. boosting methods (Adaboost)
  - Stacking
- Why use ensemble methods?
- Example cases

But first: A small experiment using people as machine learning models

WISDOM OF THE CLASS

HOW LONG IS THIS ROPE?

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

DIFFERENT PREDICTIONS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

?

One person

Average of the entire classroom

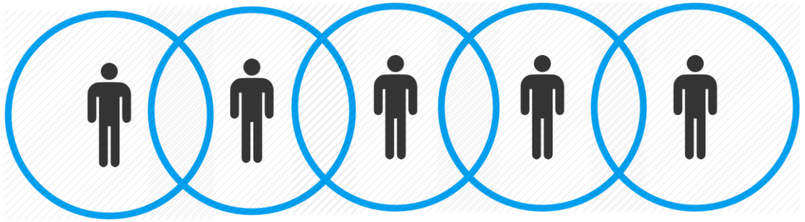
Chains with intermediate feedback

One person predicts based on the predictions of three persons of choice

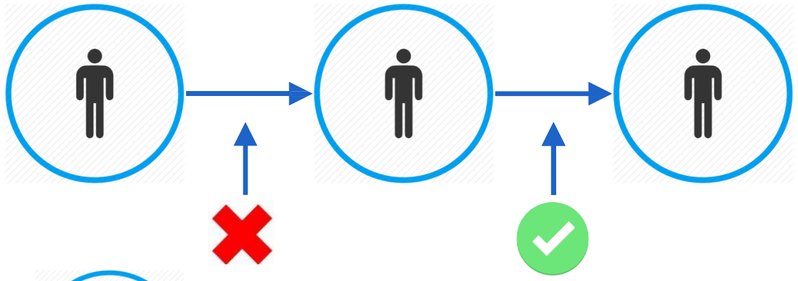
RESULTS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

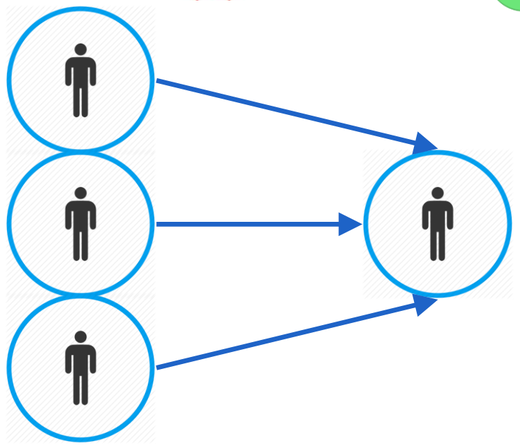
One person



Average of the entire classroom



Chains with intermediate feedback



One person predicts based on the predictions of three persons of choice

WHAT ARE ENSEMBLE METHODS?

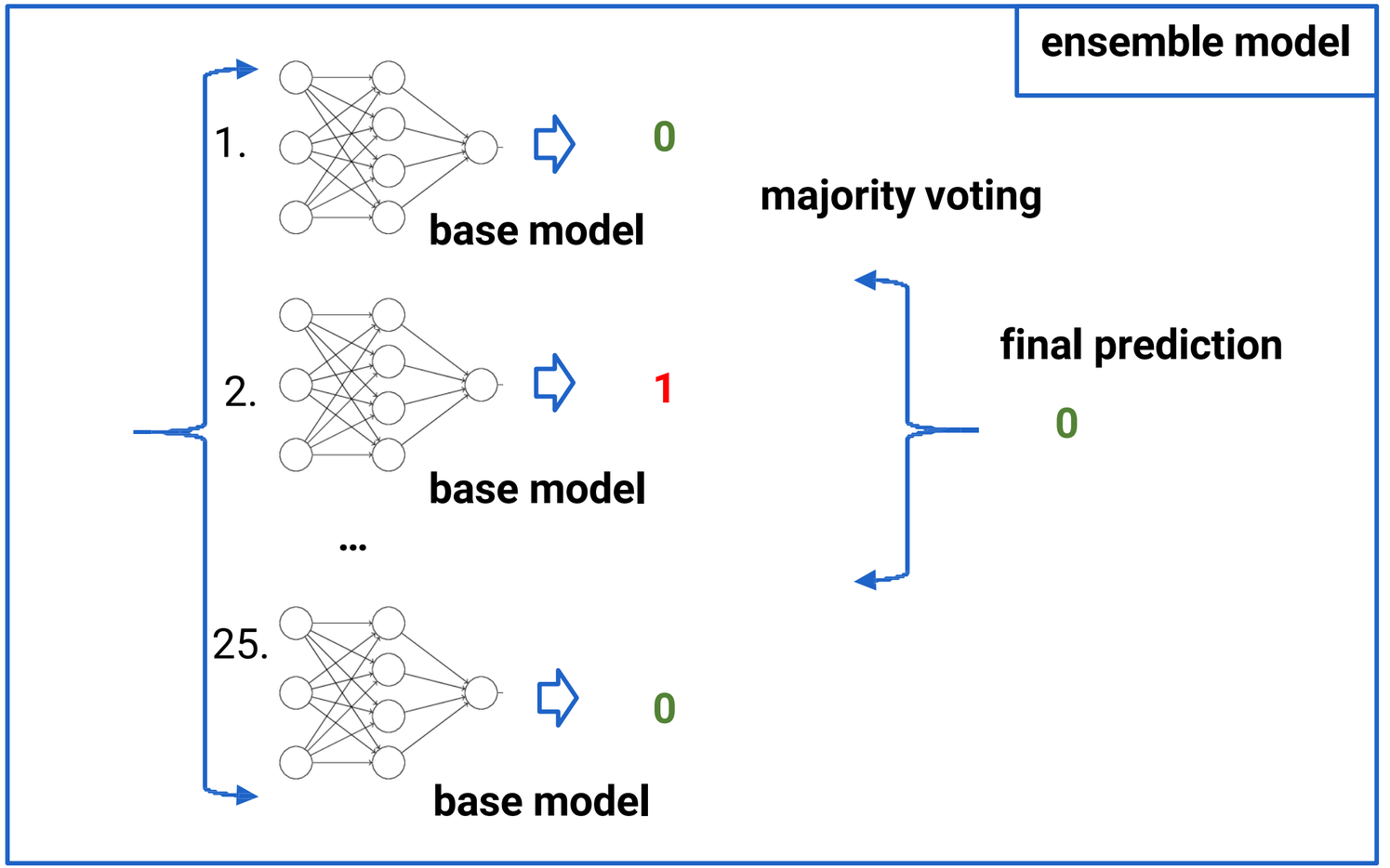
WHAT ARE ENSEMBLE METHODS?

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Basic idea: Use **multiple** classifiers or regression models to obtain a prediction which is in general **better than** the prediction of any of the **individual models**

Example:

Combining 25 classifiers for binary classification by voting



WHAT ARE ENSEMBLE METHODS?

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Basic idea: Use **multiple** classifiers or regression models to obtain a prediction which is in general **better than** the prediction of any of the **individual models**

Why would that work?

Consider a binary classification problem
Assume you have **25 classifiers** and each of them has a **35% chance to predict wrong**

Assume that the chance a classifier will make a mistake does not depend whether other classifiers made a mistake => each classifier is completely independent of the others

⇨ Final prediction is wrong if 13 or more classifiers make a mistake, chance of that happening is +- 6 %

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} \approx 0.06 \ll \varepsilon$$

WHAT ARE ENSEMBLE METHODS?

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Basic idea: Use **multiple** classifiers or regression models to obtain a prediction which is in general **better than** the prediction of any of the **individual models**

Why would that work? – Combining **independent** models with an error rate of **< 0.5** improves accuracy

Problem: **How to create independent models using the same training set**

Many different strategies have been developed to create **diverse base learners from a single training set.**

Diversity in the base models is they key to success for any ensemble method!

PLAN OF ATTACK OF ENSEMBLE METHODS

MACHINES MAKE MISTAKES

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Fact: **Perfect machine learning models do not exist**

If they would, they would:

Capture perfectly the properties and trends in the training set

Generalize perfectly when used to predict unseen data points

In practice: the error of machine learning models can always be split in three parts:

$$\mathrm{E}\left[\left(y-\hat{f}\left(x\right)\right)^2\right]=\mathrm{Bias}\left[\hat{f}\left(x\right)\right]^2+\mathrm{Var}\left[\hat{f}\left(x\right)\right]+\sigma^2$$

Prediction error BIAS VARIANCE UNAVOIDABLE ERROR

BIAS-VARIANCE TRADE-OFF

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

$$\mathrm{E}\left[\left(y-\hat{f}\left(x\right)\right)^2\right]=\mathrm{Bias}\left[\hat{f}\left(x\right)\right]^2+\mathrm{Var}\left[\hat{f}\left(x\right)\right]+\sigma^2$$

Prediction error — BIAS — VARIANCE — UNAVOIDABLE ERROR

A model with a **high error due to bias** and low variance:

Fails to capture important properties in the (training) data ✖

Has a similar behavior on unseen data points and training data ✔

Intuitively: model is not complex enough and underperforms (underfitting)

A model with a **high error due to variance** and a low bias:

Captures the behavior of the training data very well ✔

Fails to generalize on samples that deviate from the training set ✖

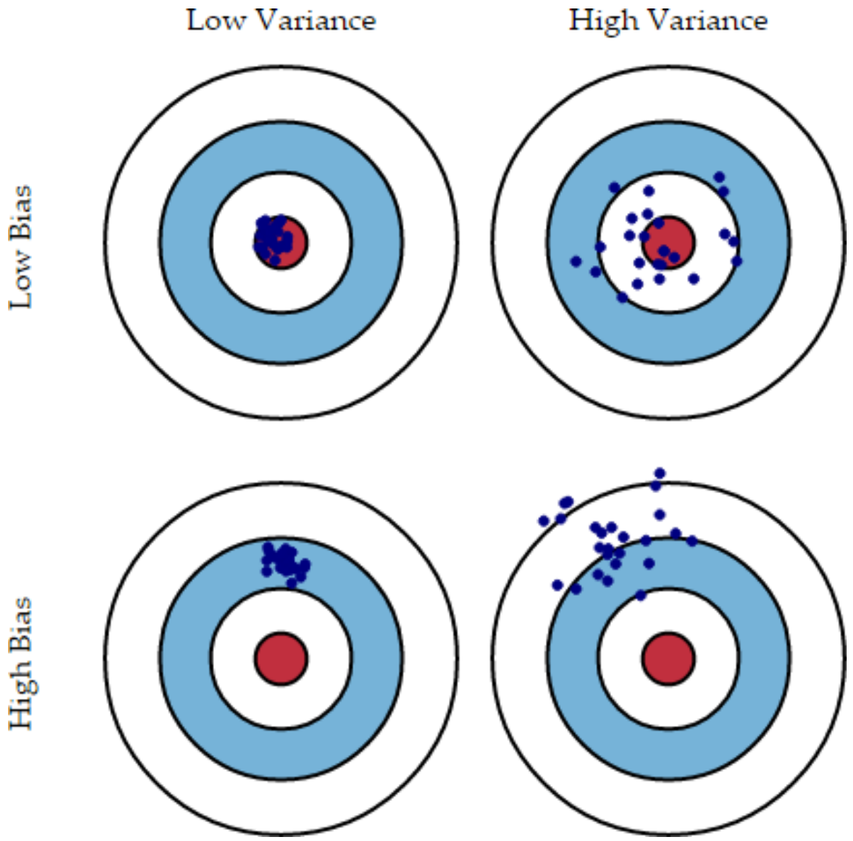
Intuitively: model is too complex and models the noise instead of general trends (overfitting)

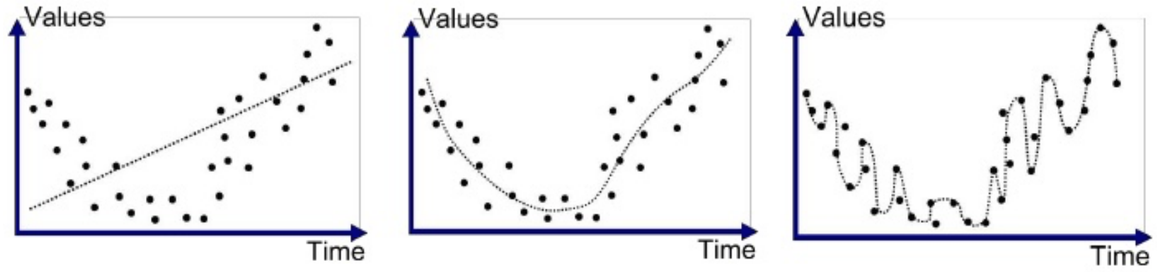
BIAS-VARIANCE TRADE-OFF

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

$$\mathrm{E}\left[\left(y-\hat{f}\left(x\right)\right)^2\right]=\mathrm{Bias}\left[\hat{f}\left(x\right)\right]^2+\mathrm{Var}\left[\hat{f}\left(x\right)\right]+\sigma^2$$

Prediction error BIAS VARIANCE UNAVOIDABLE ERROR

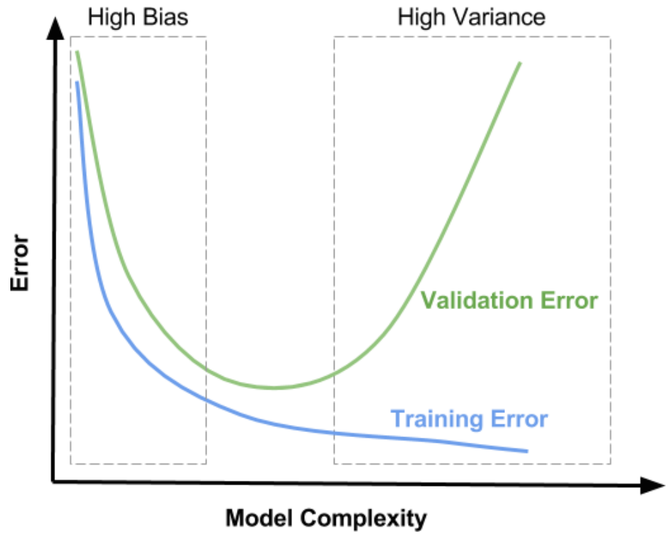




high bias
low variance

medium bias
medium variance

low bias
high variance



PLAN OF ATTACK OF ENSEMBLE METHODS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Different categories of ensemble methods exist:

Some ensemble methods attempt to improve:

1) Both the bias and variance error, with less theoretical assumptions
    E.g. **Stacking**

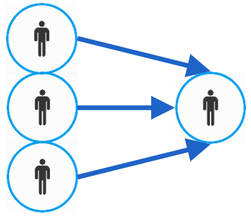
2) The variance error of base models with a low bias, high variance
    E.g. **Bagging** : decision trees -> Random Forest

3) The bias error of base models with a high bias, low variance
    E.g. **Boosting** : decision stumps -> AdaBoost

STACKING

STACKING – STEP 1

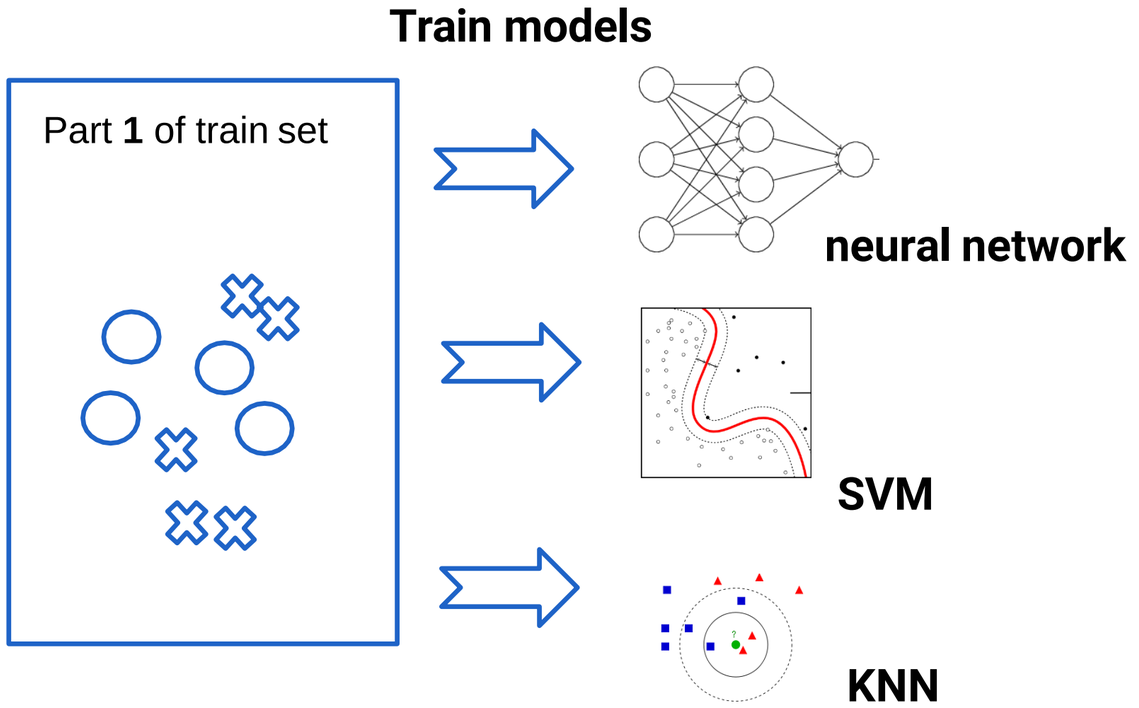
UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck



One person predicts based on the predictions of several persons of choice and the data sample

Outcome:

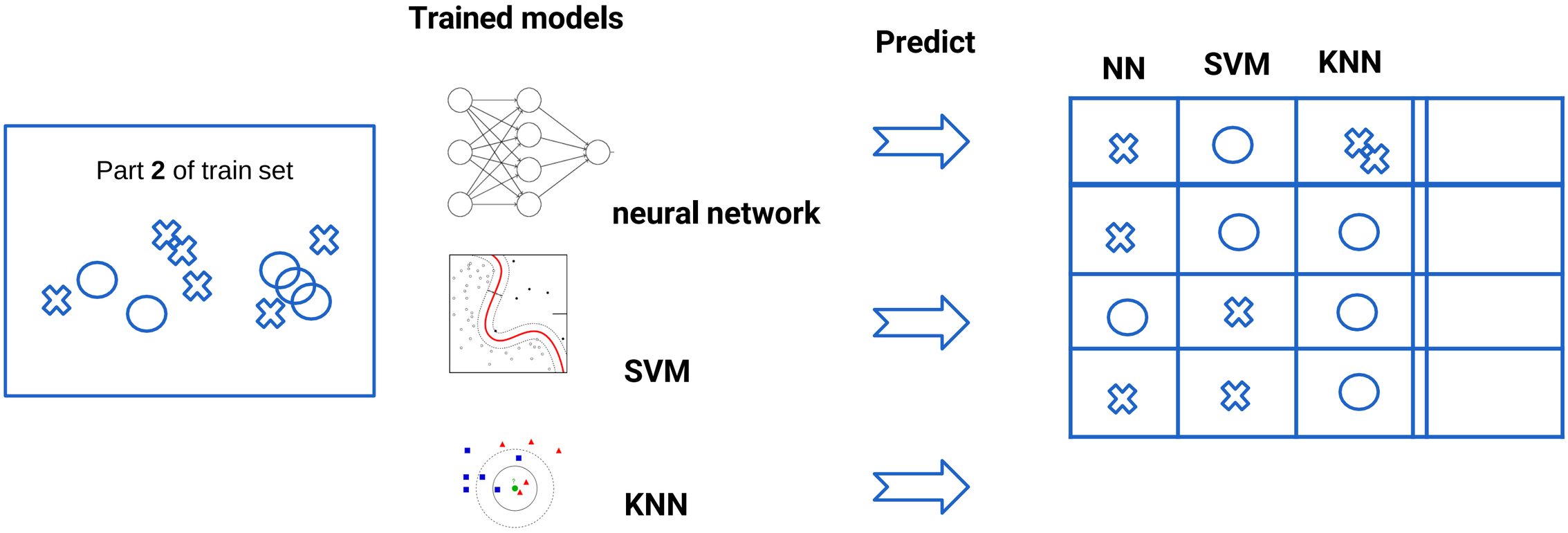
Step 1) Train several base models (strong predictors) using only a part of the training set



STACKING –STEP 2

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

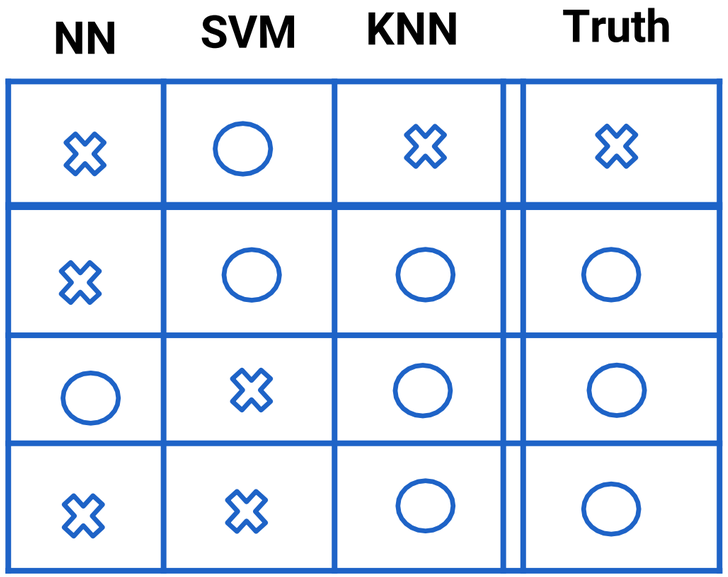
Step 2) Used the trained models to make predictions on part 2 of the training set and note prediction



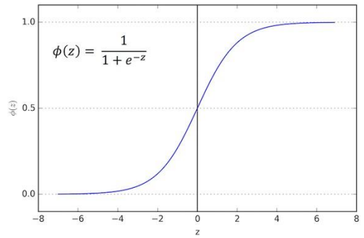
STACKING –STEP 3

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Step 3) Add the labels of part 2 of the train set to the table and learn another model using this newly created dataset. This is the final model used for testing.



Train



logistic regression

The model which uses the output of the previous models is typically called a ‘meta-learner’ in this context

STACKING –OVERVIEW

Stacking: Learn **L** classifiers on part of the training data. Use their predictions and a second part of the training data to train another classifier (meta-learner)

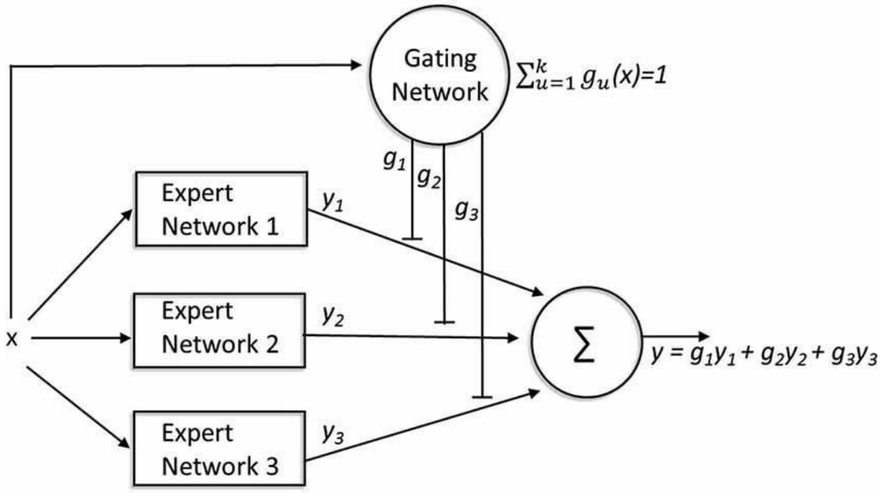
Many different variants exist: e.g. more than 2 levels, using cross-validation, different base learners. This was just a basic example

Why use stacking?

- Often performs better than just averaging different models because it can learn the connections between the different base models
- Different classifiers that are pre-trained might already exist and we wish to use them
- A single classifier might not be compatible with all of the features available to train the model. So, build different classifiers using different sets of features and combine using stacking

VARIANT OF STACKING

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck



Mixture of experts:

create ‘specialist’ models for certain parts of the data and use features of the data sample to determine which of the ‘experts’ can decide what the final prediction is

E.g. Sport quiz

Expert 1: Joos

Expert 2: De Cauwer

INTERMEZZO: THE NETFLIX PRICE

NETFLIX PRIZE (1 MILLION DOLLAR)

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Open online competition (pre-Kaggle)

Goal was to improve on Netflix own recommendation algorithm - Cinematch. The competition would end if a team would achieve a > 10% reduction in error rate (RMSE) on a certain validation set

Dataset contained 100,480,507 ratings that 480,189 users gave to 17,770 movies

Competition ended in 2009 with Team: 'BellKor's Pragmatic Chaos' winning 1 million dollars.

See documentary: https://youtu.be/ImpV70uLxyw?t=287

How many models were eventually combined?

Leaderboard

Display top 20 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelay Industries !	0.8579	9.83	2009-07-26 02:49:53
6	PragmaticTheory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dace	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8633	9.26	2009-07-21 02:04:40
14	Gravity	0.8634	9.25	2009-07-26 15:58:34
15	Ces	0.8642	9.17	2009-07-25 17:42:38
16	Invisible Ideas	0.8644	9.14	2009-07-20 03:26:12
17	Just a guy in a garage	0.8650	9.08	2009-07-22 14:10:42
18	Craig Carmichael	0.8656	9.02	2009-07-25 16:00:54
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	acmehill	0.8659	8.99	2009-04-16 06:29:35
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
Cinematch score on quiz subset - RMSE = 0.9514				

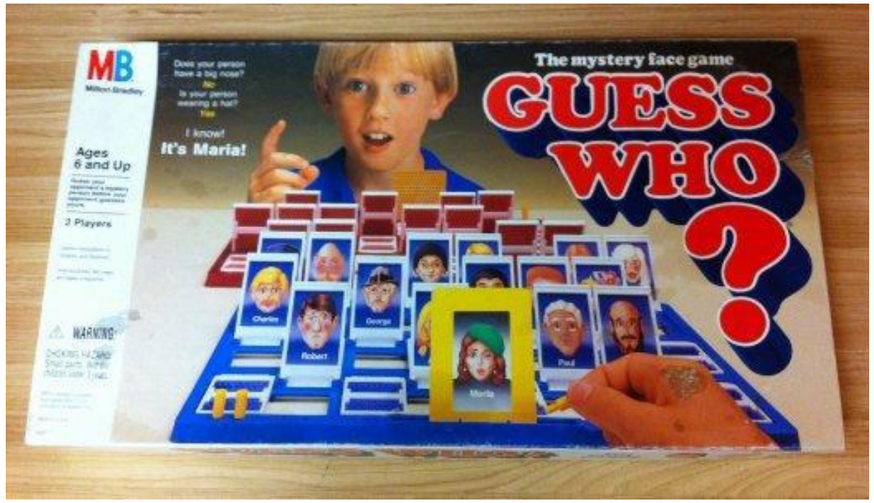
BAGGING

BASE MODEL: DECISION TREE

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

To explain bagging, we will need a base model which typically has a low bias error but a high variance error

A good example of this would be a fully grown **decision tree**



BASE MODEL: DECISION TREE

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

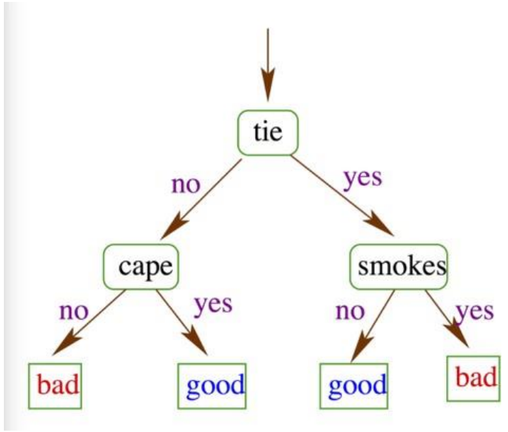
Decision trees are popular machine learning models because the model is fully interpretable.
Consider the following dataset

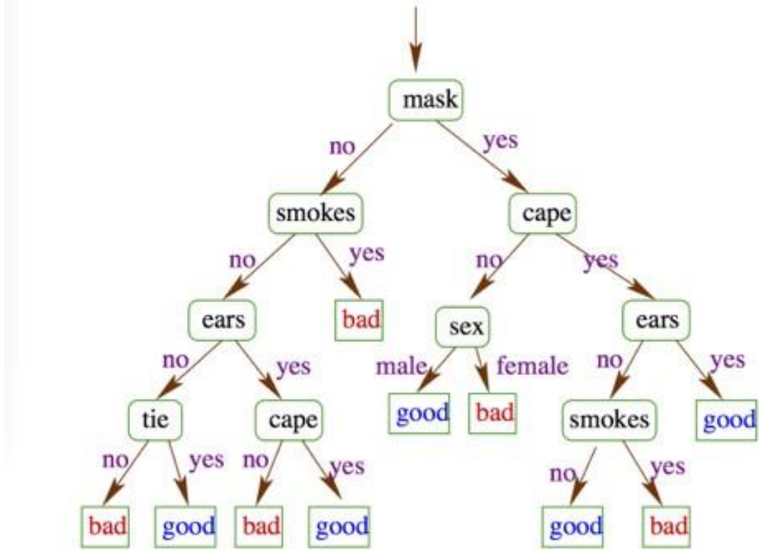
	Sex	Mask	Cape	Tie	Ears	Smokes	Target
Batman	male	yes	yes	no	yes	no	**Good**
Robin	male	yes	yes	no	no	no	**Good**
Alfred	male	no	no	yes	no	no	**Good**
Penguin	male	no	no	yes	no	yes	**Bad**
Catwoman	female	yes	no	no	yes	no	**Bad**
Joker	male	no	no	no	no	no	**Bad**

BASE MODEL: DECISION TREE

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Consider the following two decision trees:





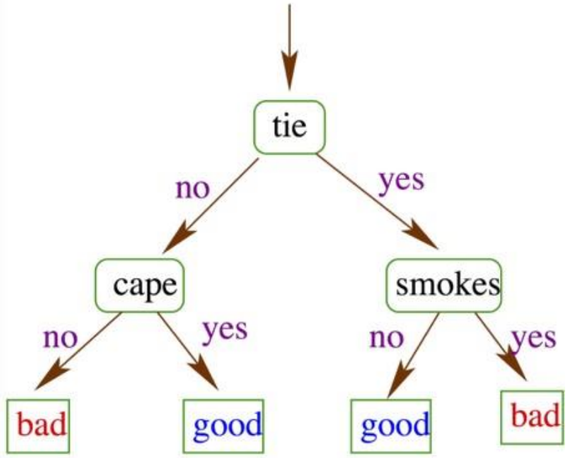
	Sex	Mask	Cape	Tie	Ears	Smokes	Target
Batman	male	yes	yes	no	yes	no	**Good**
Robin	male	yes	yes	no	no	no	**Good**
Alfred	male	no	no	yes	no	no	**Good**
Penguin	male	no	no	yes	no	yes	**Bad**
Catwoman	female	yes	no	no	yes	no	**Bad**
Joker	male	no	no	no	no	no	**Bad**

Both of the decision trees can perfectly classify our training set, but how do they perform on unseen data?

BASE MODEL: DECISION TREE

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Consider the following two decision trees:



Test set

	Sex	Mask	Cape	Tie	Ears	Smokes	Target	Model 1	Model 2
Batgirl	female	yes	yes	no	yes	no	**Good**	?	?
Riddler	male	yes	no	no	no	no	**Bad**	?	?

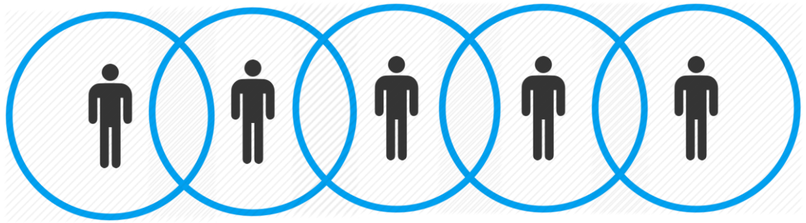
Decision trees with a lot of leaf nodes with only a few training examples per leaf **quickly tend to overfit**

Fully grown decision trees are therefore good examples of models which have a low bias error but a high variance.

BAGGING

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Idea: create diversity among the base models by training each model **using different training data.** Then average the predictions.



Average of the entire classroom

Outcome:

*Every one of you have seen different cereal collections in your life, so your training set is different.

Problem: there is never enough training data, so if we split the data in different parts, one part might not be sufficient to train a (low bias) model.

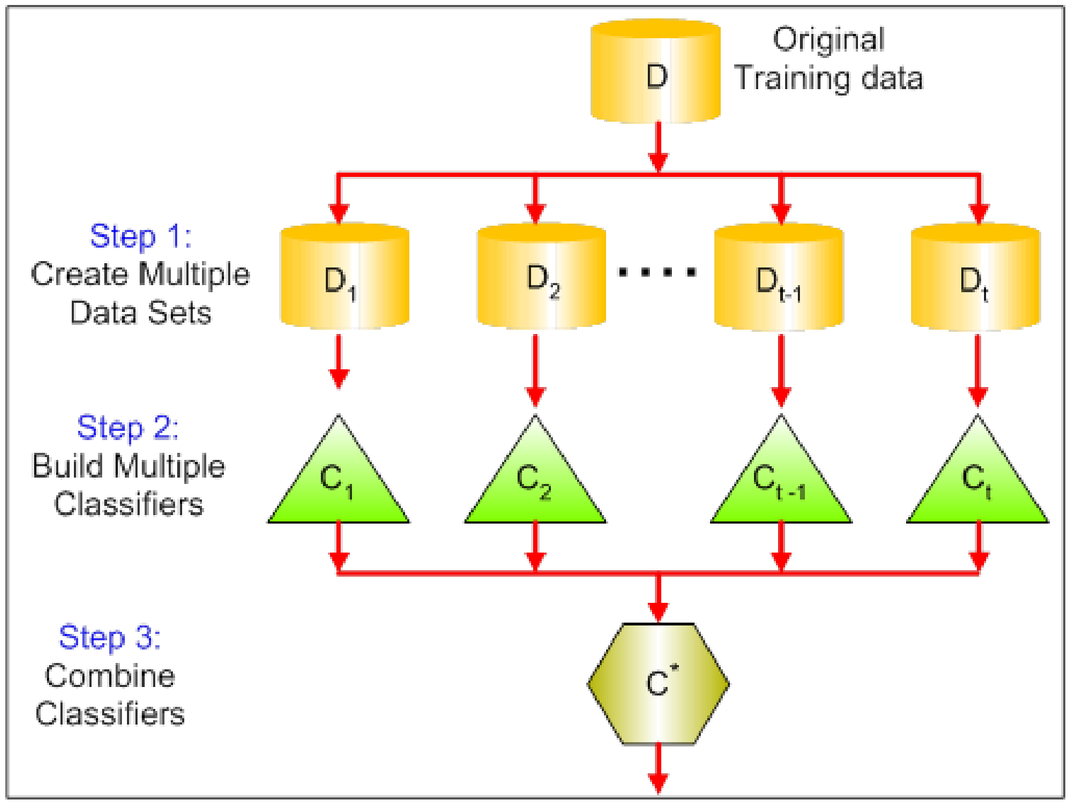
Bagging: "Bootstrap aggregating" – create new training sets by randomly selectin *M* samples (with replacement) from the full set of *N* data points.

BAGGING - SCHEMATIC

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Idea: create diversity among the base models by training each model **using different training data.** Then average the predictions.

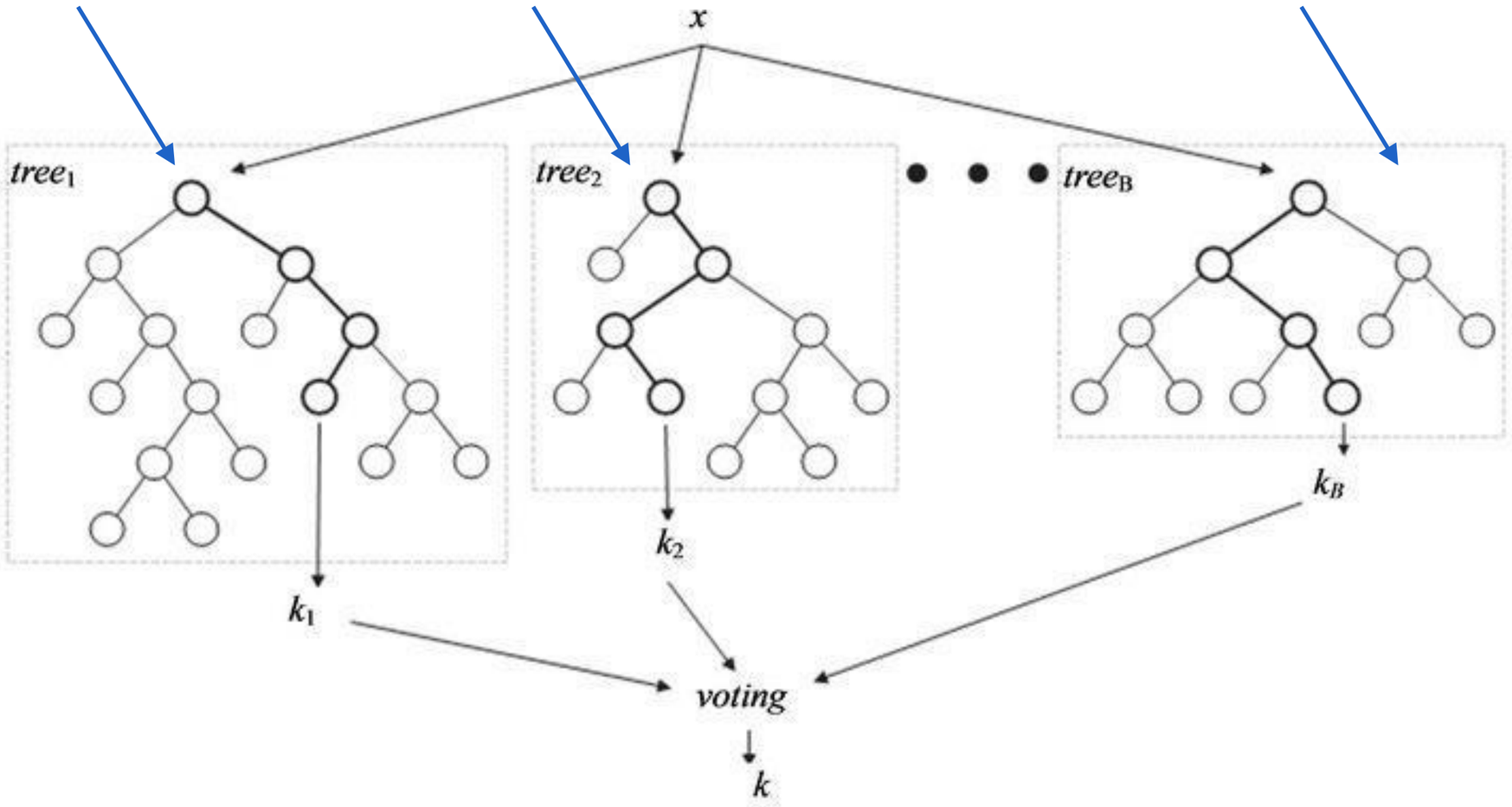
Bagging: "Bootstrap aggregating" – create new training (=bootstraps/bags) by randomly selecting *M* samples (with replacement) from the full set of *N* data points.



BAGGED TREES

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

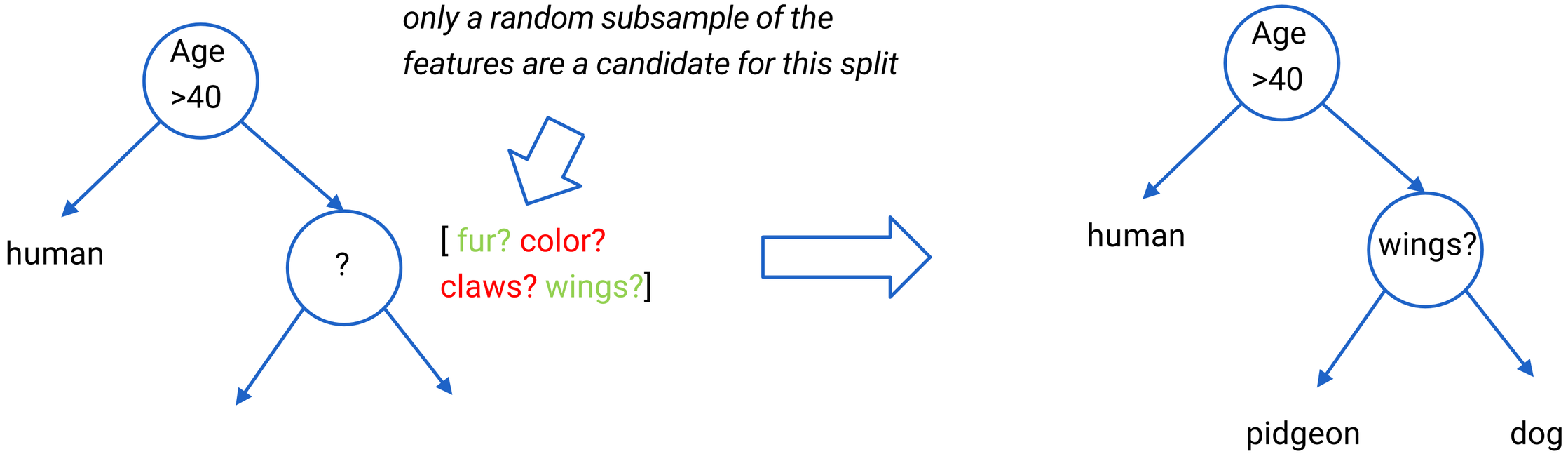
Trained using bag/bootstrap 1 **Trained using bag/bootstrap 2** **Trained using bag/bootstrap B**



BAGGING THE FEATURES

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

To create extra diversity in the base learners, sometimes not only the training set is bagged or subsampled but also the features/dimensions that can be used to create a split in the tree.



RANDOM FORESTS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Basically, bagged trees with random subsampling of the features at each split

One of the most popular machine learning methods in history

Why?

- Fast to train: each tree can be trained in parallel
- Though to beat and versatile: good performance on a large variety of tasks
- Difficult to go wrong : do not require careful tuning of parameters, hard to overfit

This resulted in the use of Random Forests by a lot of users that are not experts in machine learning

RECAP

Bagging to create ensemble methods works because it reduces the variance of the individual classifiers.

In practice, often it can both decrease the bias and variance error. Although, in theory, there are more guaranties that it lowers the latter.

Therefore, typically used in combination with base learners which have a low bias error but high variance such as decision trees.

Random Forests is a popular ensemble method which uses decision trees as base learners and employs bagging of both the training data and the features that can be picked from at each split.

Random Forests is often an excellent choice for classic machine learning problems because it's hard to create a bad Random Forest as the hyperparameters typically don't affect the model that much and it's hard to overfit.

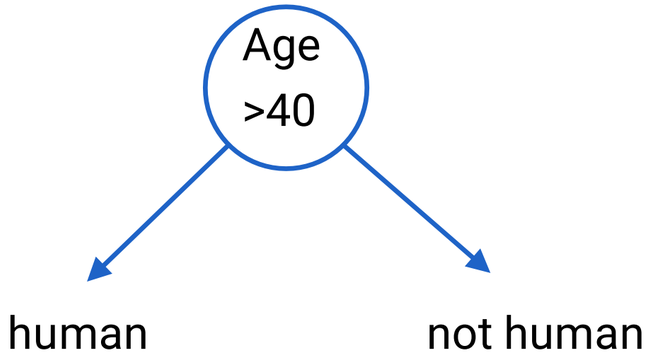
BOOSTING

BASE MODEL: DECISION STUMP

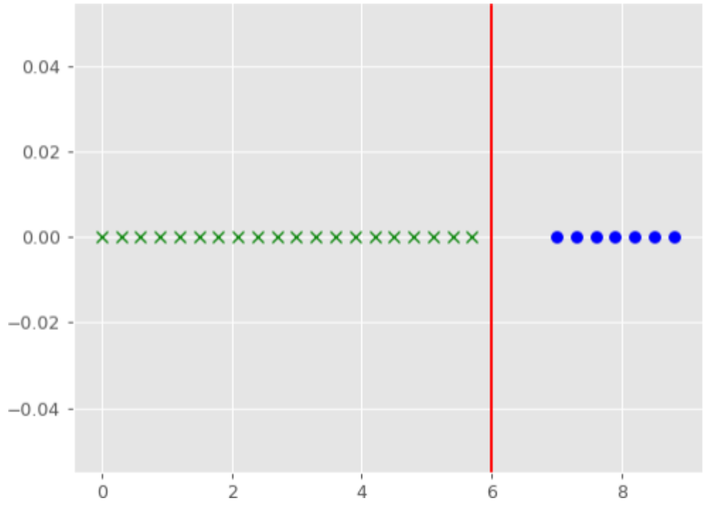
UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

To explain boosting, we will need a base model which typically has a high bias error but a low variance error

A good example of this would be a **decision stump**



A decision stump is a decision tree with only 1 split. Therefore it splits the entire dataset using only 1 feature/dimension and 1 split value.

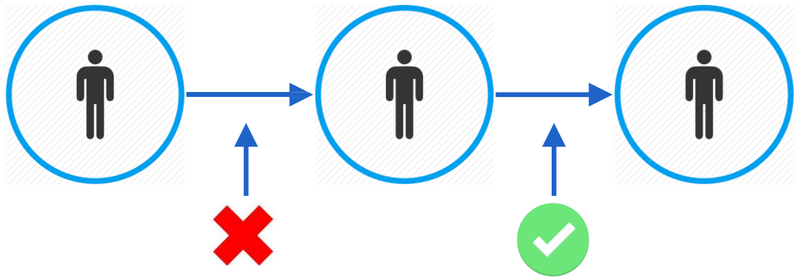


split value

BOOSTING

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Idea: create diversity among the base models by training each model **by learning from the mistakes from the previous model.**



Chains with intermediate feedback

Outcome:

Problem: How to create a new base learner which focusses on the mistakes of the previous classifier(s)?

Prestigious method: 'AdaBoost' [Adaptive Boosting]

Construct a final 'strong' ensemble classifier by taking a weighted sum of the predictions of 'weak' classifiers.
Weak classifiers = less than 50% training error

Each data point in the training set receives a weight and after each base learner is trained, the data points which are misclassified by that model receive a higher weight

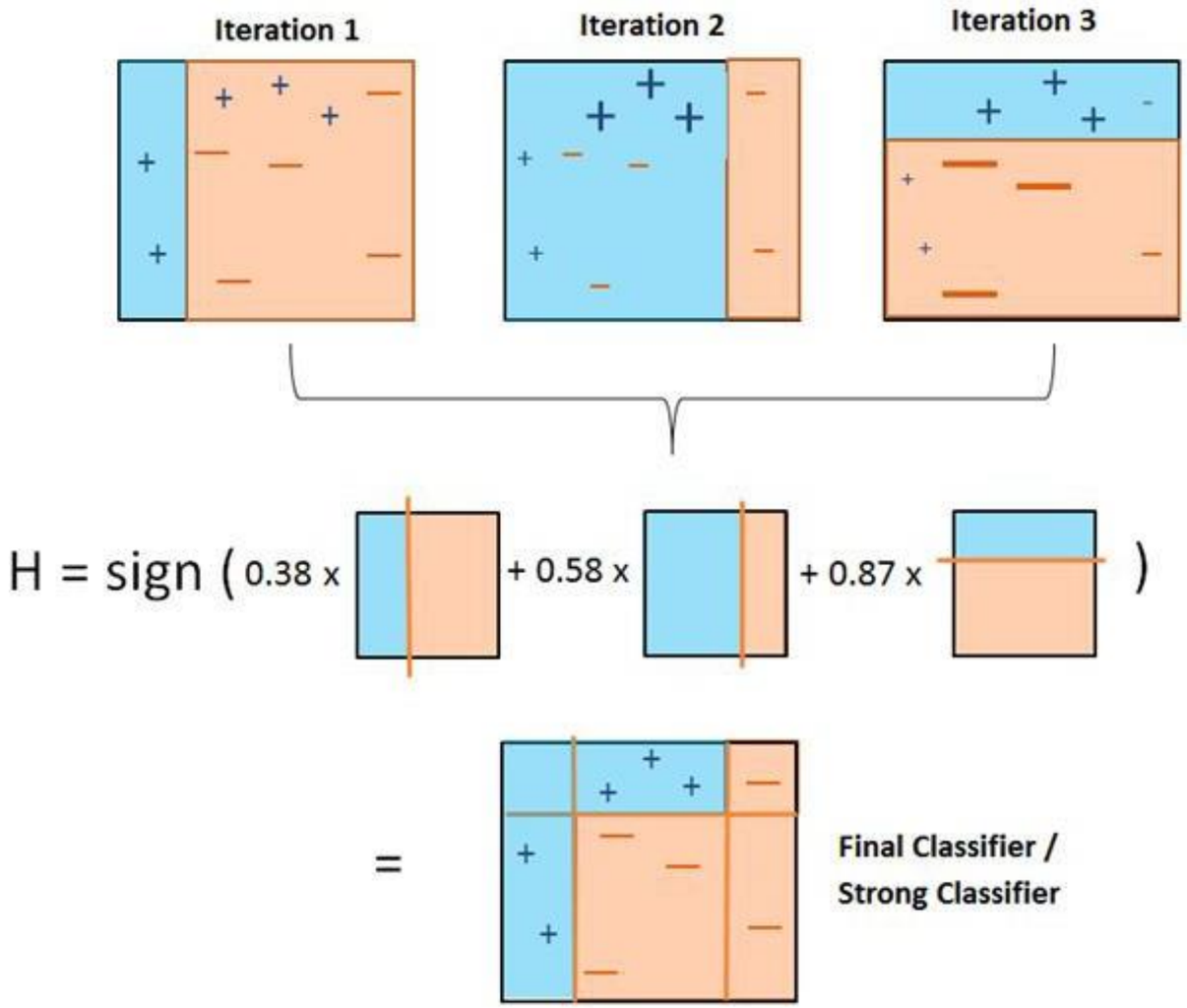
Base models with a higher accuracy are received more important in the final ensemble

ADABOOST- VISUALIZATION

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Idea: create diversity among the base models by training each model **by learning from the mistakes from the previous model.**

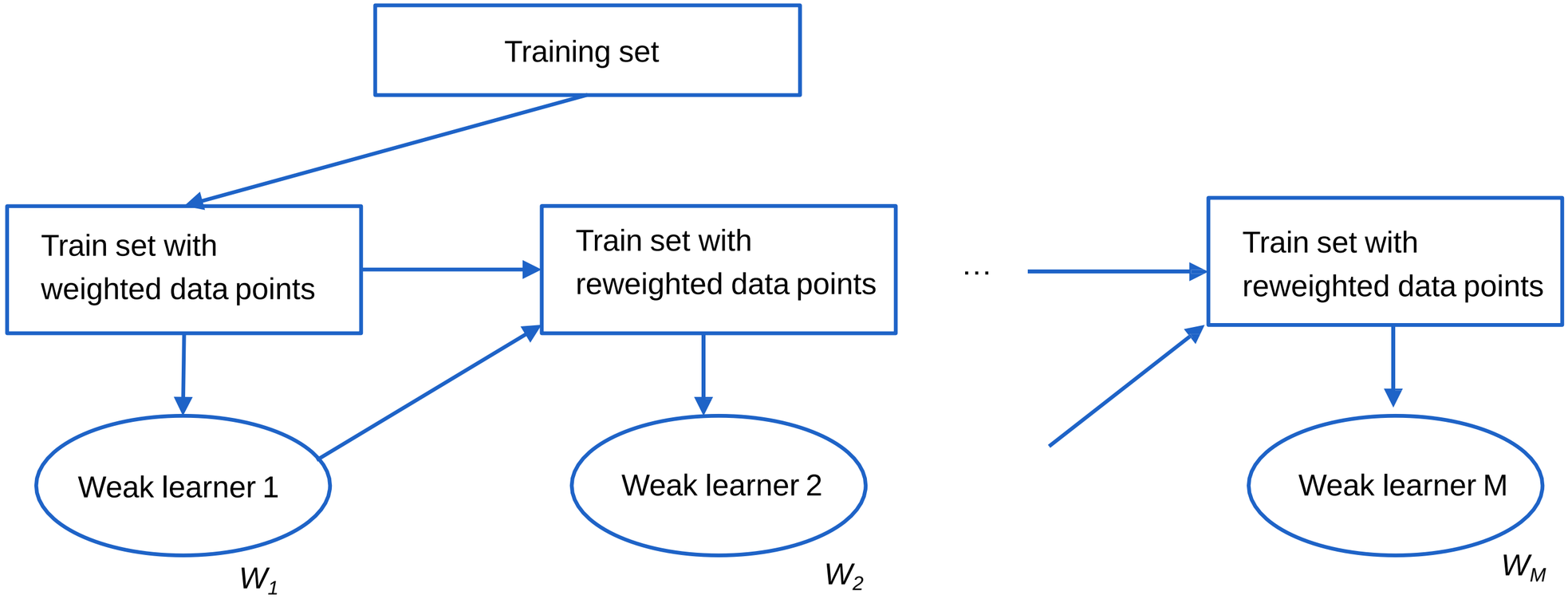
- Learns base models sequentially.
- Each data point in the training set receives a weight.
- After each base learner is trained, the data points which are misclassified by that model receive a higher weight
- Each base model also receives a weight related to the training error it achieves. Base models with a higher accuracy are received more important in the final ensemble.



ADABOOST- SCHEMATIC

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Idea: create diversity among the base models by training each model **by learning from the mistakes from the previous model.**



Base models

Weights learner

Ensemble models **Weighted sum of base model predictions:** W_1 Weak learner 1 + W_2 Weak learner 1+….

ADABOOST- PSEUDOCODE (SEE LAB)

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Let $(x_1, t_1), \ldots, (x_N, t_N)$ be the training data, where $x_n \in X, t_n \in Y = \{-1, 1\}$ and I is the indicator function

Initialize $w_n^1 = 1/N$ for $n = 1, \ldots, N$.

For $m = 1, \ldots, M$

1) Train the classifier $y_m(\cdot)$ by minimizing the weighted error function

$$\epsilon_m = \sum_{n=1}^{N} (w_n^m I(y_m(x_n) \neq t_n)).$$

← **single weak classifier tries to find best split**

where $I(y_m(x_n)) \neq t_n$ equals 1 when $y_m(x_n) \neq t_n$ and 0 otherwise.

2) Compute ϵ_m itself.

← **compute training error**

3) Compute the classifier weight alpha:

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right)$$

← **calculate importance of this weak learner based on training error**

4) Update the datapoints weights:

$$w_n^{m+1} = \frac{w_n^m \exp(\alpha_m I(y_m(x_n) \neq t_n))}{\sum_{k=1}^{N} w_k^m \exp(\alpha_m I(y_m(x_k) \neq t_k))}$$

← **give more importance to misclassified points**

← **forces the weights to sum to 1**

After finishing the loop, make the final prediction:

$$Y_M(\cdot) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m y_m(\cdot) \right)$$

← **final prediction is weighted sum of predictions of individual weak models**

ADABOOST

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Winner of Gödel Prize 2003

Basically, create diversity by (re)weighing data points based on the mistakes of previous model(s)

Very powerful machine learning technique

Why?

- Easy to implement and few parameters to set
- Resistant to overfitting (see lab)
- Can have very good performance

Disadvantage: sensitive to outliers and noisy data because the algorithm can give more importance to them over time

RECAP

AdaBoost to create ensemble methods works because it reduces the bias of the individual classifiers.

In practice, often at the start of the chain it mainly decreases bias but in the later stages it can also reduce variance. In theory, the focus is on the first.

Therefore, typically used in combination with base learners which have a high bias error but low variance such as decision stumps.

AdaBoost is often an excellent choice for classic machine learning problems because it’s easy of use and often better performance than bagging methods.

WHY USE ENSEMBLE METHODS

WINNER ALGORITHMS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

All types of ensemble methods have proven to be able to deliver top performance in numerous challenges.

Want to win a Kaggle competition? **Use ensemble methods.**

kaggle

More tricky to train, but extremely powerful in the right hands

Almost all competitive submissions to Kaggle leaders boards use some kind of ensemble method.

Dropout, an essential technique in training deep learning networks is essentially a form of extreme bagging.

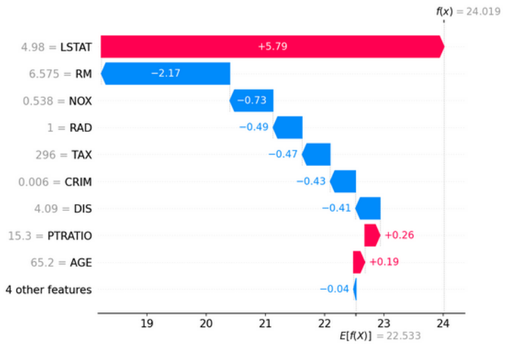
FEATURE IMPORTANCES

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

Tree ensemble example (XGBoost/LightGBM/CatBoost/scikit-learn/pyspark models)

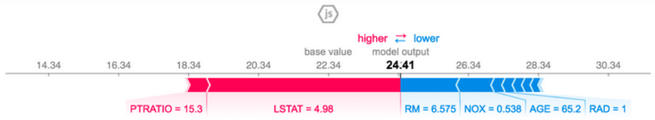
While SHAP can explain the output of any machine learning model, we have developed a high-speed exact algorithm for tree ensemble methods (see our Nature MI paper). Fast C++ implementations are supported for *XGBoost*, *LightGBM*, *CatBoost*, *scikit-learn* and *pyspark* tree models:

```
import xgboost
import shap

# train an XGBoost model
X, y = shap.datasets.boston()
model = xgboost.XGBRegressor().fit(X, y)

# explain the model's predictions using SHAP
# (same syntax works for LightGBM, CatBoost, scikit-learn, transformers, Spark, etc.)
explainer = shap.Explainer(model)
shap_values = explainer(X)

# visualize the first prediction's explanation
shap.plots.waterfall(shap_values[0])
```



The above explanation shows features each contributing to push the model output from the base value (the average model output over the training dataset we passed) to the model output. Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue. Another way to visualize the same explanation is to use a force plot (these are introduced in our Nature BME paper):

```
# visualize the first prediction's explanation with a force plot
shap.plots.force(shap_values[0])
```



If we take many force plot explanations such as the one shown above, rotate them 90 degrees, and then stack them horizontally, we can see explanations for an entire dataset (in the notebook this plot is interactive):

SHAP is a fairly recent and promising Python package to get (more advanced) insights on the model inner working.

It integrates with most tree ensemble models via a fast C++ link

ENSEMBLE METHODS FOR HUMANS

UNIVERSITEIT GENT
CAMPUS KORTRIJK
J. Ruyssinck

‘Wisdom of the crowd’

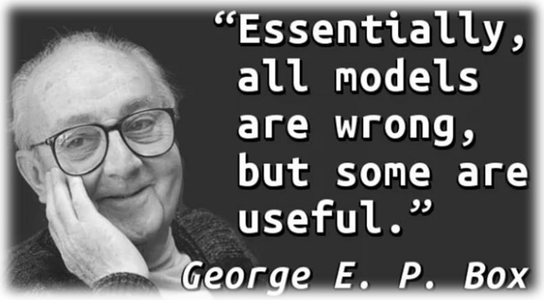
Idea was first suggested in Ancient Greece that a collective opinion of a group is as good as and often superior than any individual person.

More recent success stories include

And this classroom?

EXAMPLE CASES

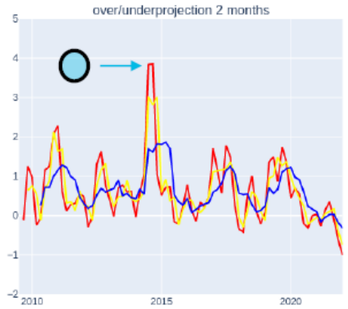
Stock management via demand forecasting



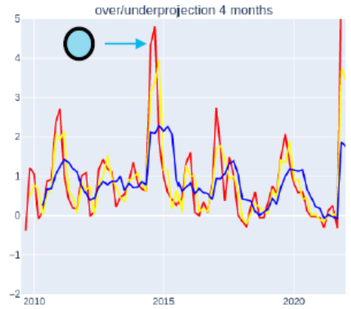
Creation of benchmark of relevant (business) error metrics and visualizations

woodkopmaat	totaal volume	volume 2 maand	overprojectie 2 maand	2 maand projectiefout	volume 4 maand	overprojectie 4 maand	4 maand projectiefout	volume 6 maand	overprojectie 6 maand	6 maand projectiefout
[redacted]	9523.95	9018.4	-0.05	0.74	9425.55	-0.01	0.83	9305.22	-0.02	0.89
[redacted]	7499.05	8174.14	0.09	0.8	8175.17	0.09	0.76	7423.65	-0.01	0.67
[redacted]	2721.16	2955.97	0.09	1.29	3235.55	0.19	1.21	3406.5	0.25	1.13
[redacted]	2407.66	2521.97	0.05	1.37	2483.08	0.03	1.33	2492.99	0.04	1.36
[redacted]	2331.39	2342.46	0	0.81	2286.22	-0.02	0.82	2334.74	0	0.85
[redacted]	1796.33	1601.47	-0.11	1.45	1731.26	-0.04	1.32	1765.95	-0.02	1.49
[redacted]	1719.27	1886.41	0.1	1.49	1932.96	0.12	1.7	1997.55	0.16	1.79
[redacted]	1314.9	1141.53	-0.13	0.98	1165.47	-0.11	1.05	1203.19	-0.08	1.02
[redacted]	1265.33	1353.99	0.07	1.75	1331.67	0.05	2.34	1324.53	0.05	2.37
[redacted]	1184.96	1249.99	0.05	1.38	1383.98	0.17	1.41	1386.98	0.17	1.33
[redacted]	1134.13	1587.15	0.4	2.23	1677.9	0.48	2.15	1570.36	0.38	2.05
[redacted]	769.06	782.97	0.02	1.89	905.23	0.18	1.65	872.48	0.13	1.53
[redacted]	612.67	557.59	-0.09	2.61	569.67	-0.07	3.01	598.94	-0.02	2.14
[redacted]	586.32	668.1	0.14	0.97	572.85	-0.02	1.17	517.27	-0.12	0.91
[redacted]	535.53	680.64	0.27	2.08	953.49	0.78	2.46	813.03	0.52	2.12
[redacted]	384.73	297.3	-0.23	1.68	293.22	-0.24	1.62	271.21	-0.3	1.51
[redacted]	211.44	308.07	0.46	1.54	318.07	0.5	1.35	216.63	0.02	1.22

Model 2 maand op voorhand



Model 4 maand op voorhand



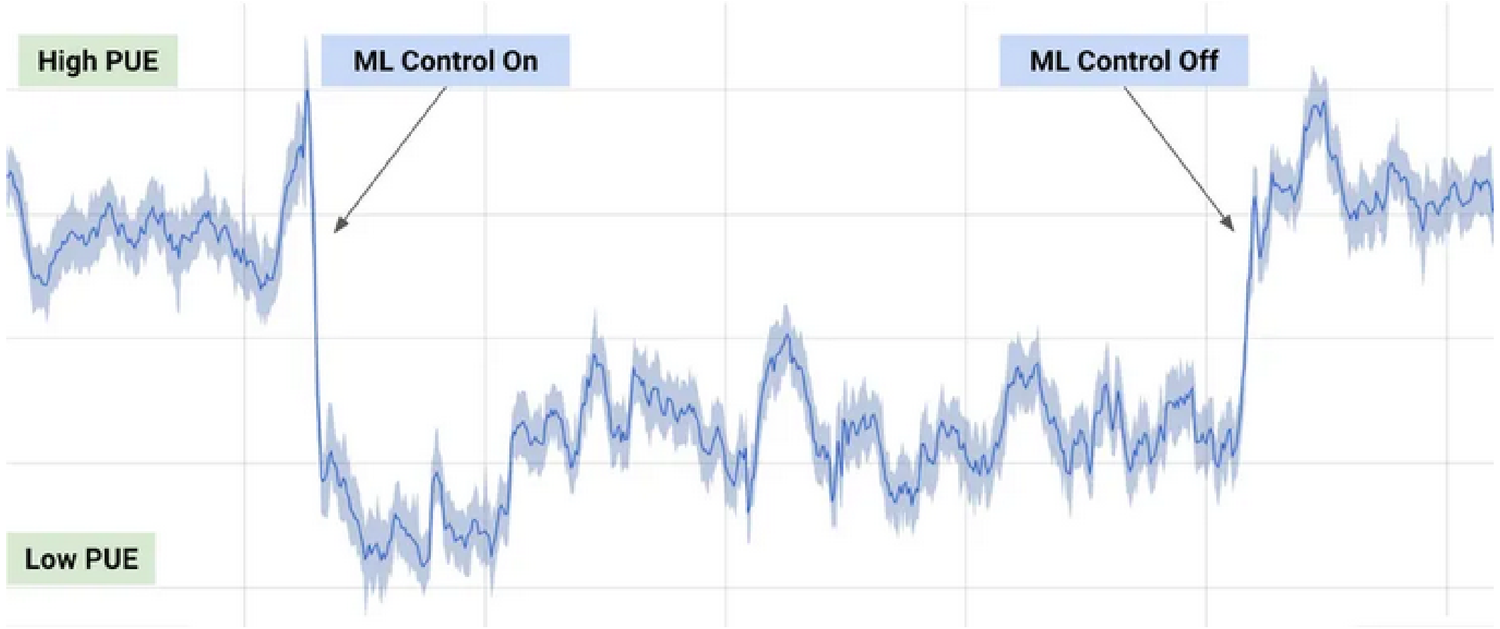
Creation of many custom performance measures of models (More than 1000 SKU modeled)

van Hoorebeke
We take timber seriously.

Client Use case: Demand forecasting models for a large wood importer

SAVINGS OF HUNDREDS OF MILLIONS OF DOLLARS

After accounting for "electrical losses and other non-cooling inefficiencies," this 40 percent reduction translated into a 15 percent reduction in overall power saving, says Google. Considering that the company used some 4,402,836 MWh of electricity in 2014 (equivalent to the amount of energy consumed by 366,903 US households), this 15 percent will translate into savings of hundreds of millions of dollars over the years. And when you remember that Google reportedly paid $600 million for UK-based DeepMind back in 2014, it seems the company's bet on AI will pay for itself before too long.



Polluting diesel generators being replaced by AI driven battery systems

Foodm

Smart fridges combatting food waste with AI

Authenticity analysis of oregano: development, validation and fitness for use of several food fingerprinting techniques

Jet Van De Steene [a,b,*], Joeri Ruyssinck [c], Juan-Antonio Fernandez-Pierna [d], Lore Vandermeersch [e], An Maes [f], Herman Van Langenhove [e], Christophe Walgraeve [e], Kristof Demeestere [e], Bruno De Meulenaer [f], Liesbeth Jacxsens [f], Bram Miserez [a]

[a] Ciboris, Technologiepark 90 zone A6b, 9052 Zwijnaarde, Belgium
[b] Primoris Belgium, Technologiepark 90 zone A6b, 9052 Zwijnaarde, Belgium
[c] ML2Grow, Reigerstraat 8, 9000 Ghent, Belgium
[d] Quality and authentication of products unit, knowledge and valorization of agricultural products Department, Walloon Agricultural Research Centre, Chaussée de Namur 24, 5030 Gembloux, Belgium
[e] Department of Green Chemistry and Technology, Research Group EnVOC, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Ghent, Belgium
[f] Department of Food Technology, Safety and Health, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

ML²GROW

ABSTRACT

Several analytical techniques, i.e. spectroscopic techniques as Near Infrared (NIR) and Mid-Infrared (MIR), Hyper Spectral Imaging (HSI), Gas Chromatography coupled to Mass Spectrometry (GC–MS) and Proton-transfer Reaction Time-of-Flight Mass spectrometry (PTR-TOF-MS), combined with chemometrics, are examined to evaluate their potential to solve different food authenticity questions on the case of oregano. In total, 102 oregano samples from one harvest season were analyzed for origin and variety assessment, 159 samples for adulteration-assessment and 72 samples for batch-to-batch control. The Gaussian Process Latent Variable Model (GP-LVM) was applied as technique to obtain a reduced two-dimensional space. A Random Forest Regression algorithm was used as regression model for the adulteration assessment. Prediction rates of more than 89% could be achieved for origin assessment. For variety assessment, prediction rates of more than 78% could be obtained. Batch-to-batch control could be successfully performed with NIR and PTR-TOF-MS. Detection of adulteration could be successfully performed from 10% on with HSI, NIR and PTR-TOF-MS.

5 different data analytics techniques for the purpose of:

Batch-2-Batch analysis: Can we distinguish between batches of products (harvests)

Bio versus convention: Can we distinguish between biological and conventional yields

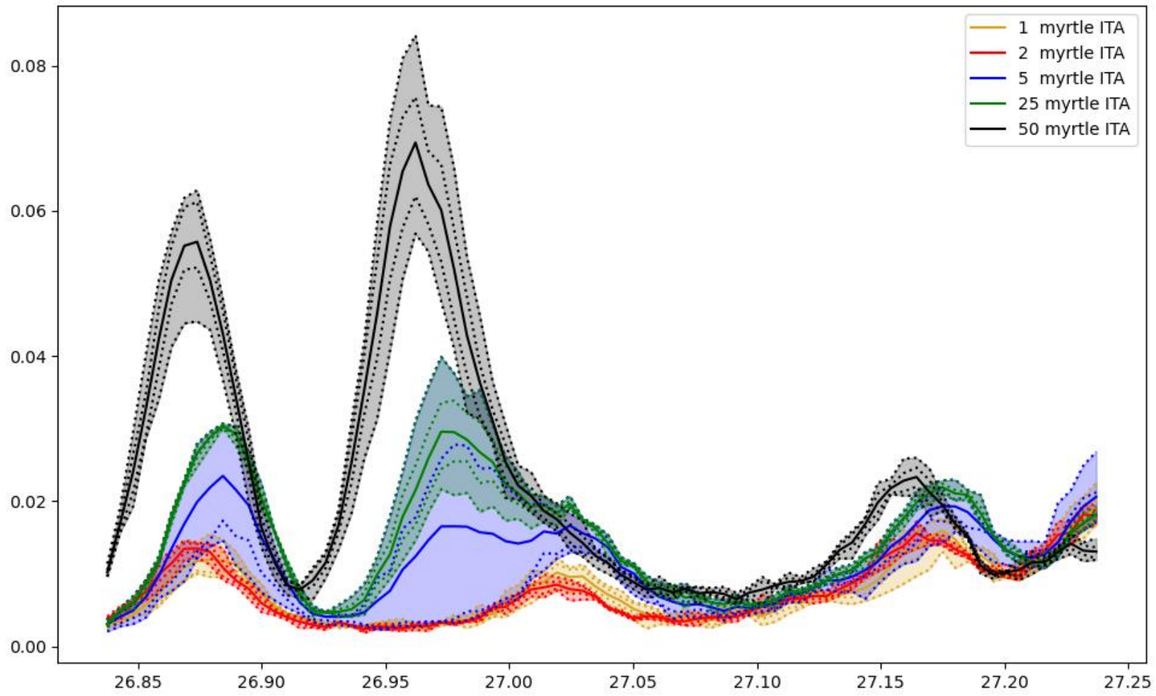
Vulgare versus Onites: Can we distinguish different varieties of Oregano?

Geographical variance: Can we distinguish or detect from which region the Oregano is from?

Adulteration: Can we detect tampering?

GC-MS

Gas chromatography–mass spectrometry (GC-MS) is a staple analytic method combining gas-chromatography and mass spectrometry. It results in a chromatogram with on the x-axis the retention time and on the y-axis the concentration of certain component. We will use standardization based on an internal standard.



(peak detection)

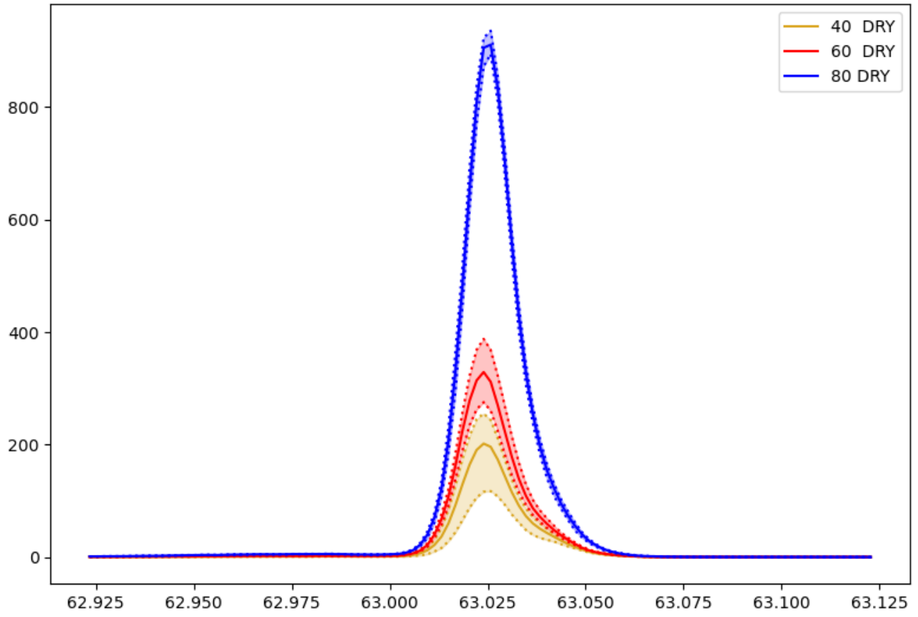
Data type: ‘2DSeries’ → Tabular data

Each row: 1 sample
Each column: the peak value or area at a certain retention time

PTRMS

Proton Transfer Reaction - Time of Flight – Mass Spectrometry (PTRMS) is a non-targeted method used in real-time detection of volatile organic compounds (VOCs) for different applications (environmental chemistry, food and beverages, etc.). It is an innovative technique based on soft chemical ionization where proton transfer from H_3O+ ions all compounds with a higher proton affinity (PA) than water are ionized. Common constituents of air, such as N_2, O_2, Ar, CO_2 etc. have lower PAs than H_2O and are therefore not detected, which enables very low detection limits for volatile trace compounds (≤ ppt_v).

Due to precisely controlled ion source and drift tube parameters, absolute quantification of VOC concentrations is possible. PTR-TOF-MS is a highly sensitive technique which does not involve chromatographic separation so it allows rapid screening of a large number of samples (analysis time 2 -5 min per sample).



(peak detection)

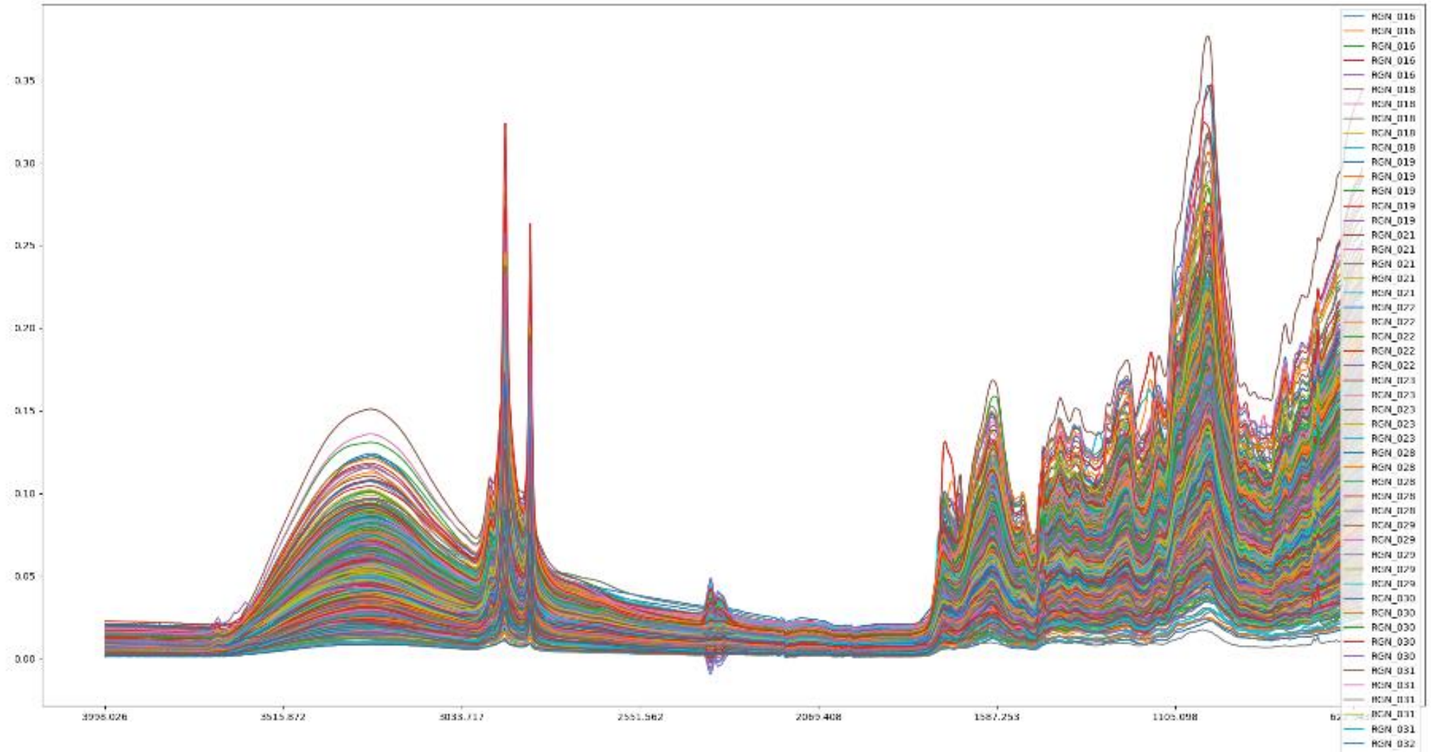
Data type: ‘2DSeries’ → Tabular data

Each row: 1 sample
Each column: the peak value or area at a m/z

MIR

Mid-infrared (MIR) spectroscopy is nondestructive, noninvasive and requires minimal or no sample preparation, and only need a few seconds for spectral collection. It relies on light absorption with a sample being targeted by incident IR light causing biochemical bonds to vibrate. Vibrational modes of molecule thereby cause specific amounts of energy from the incident IR beam to be absorbed, reducing the intensity of the subsequently detected IR beam. The difference in energy between incident and detected IR radiation produces a complex interferogram, which is deconvoluted using a FT operator. This separates the individual wavelengths of the measured IR range into component wavelengths, producing a wavenumber spectrum. MIR spectroscopy, in contrast to Raman spectroscopy, relies on a dipole moment present only in diatomic or more complex

liation can only penetrate the sample to a very shallow depth



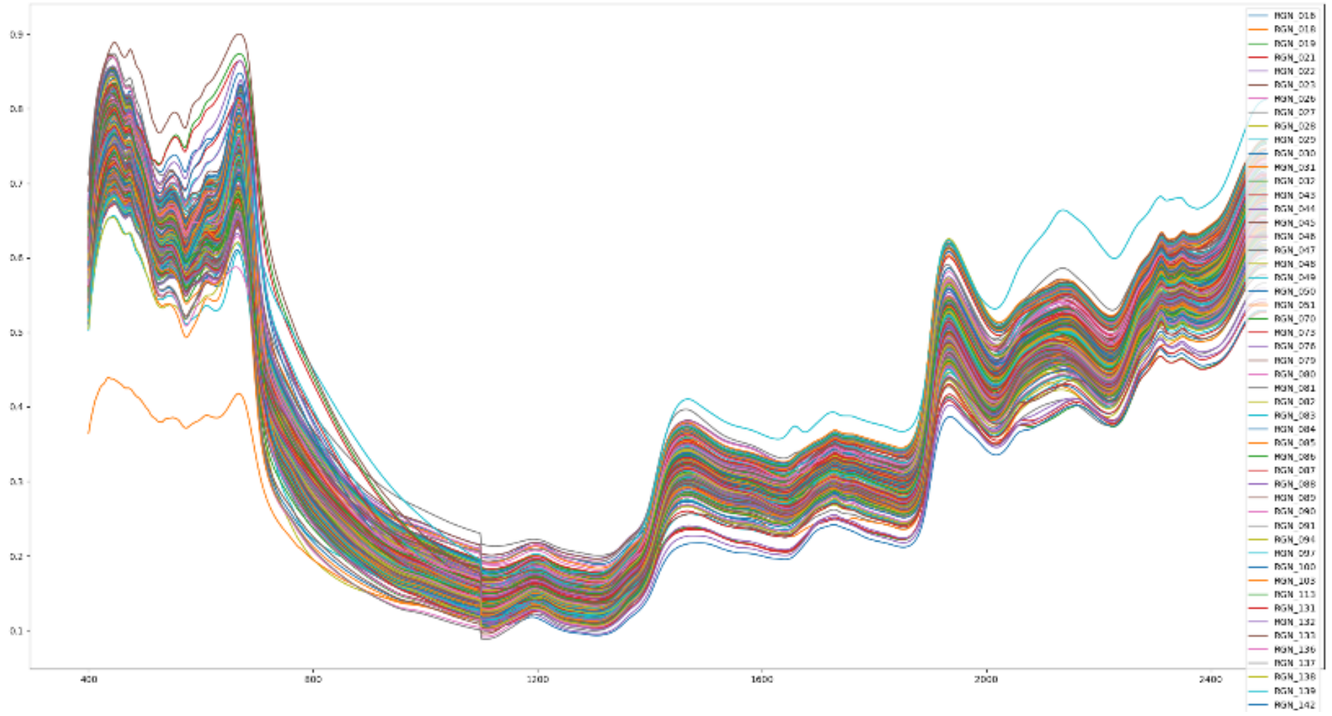
(pre-processing)

Data type: ‘2DSeries’ → Tabular data

Each row: 1 sample
Each column: certain derivative values in the series

NIR

Near-infrared (NIR) spectroscopy, just like MIR, is a nondestructive, noninvasive and requires minimal or no sample preparation, and only need a few seconds for spectral collection. The sample is typically illuminated by near infrared light in a range of (800-2500nm) which can be absorbed or reflected. Compared to MIR, a major advantage of NIR over MIR is that NIR light can penetrate much farther into a sample. Further information can be found in the MIR report.

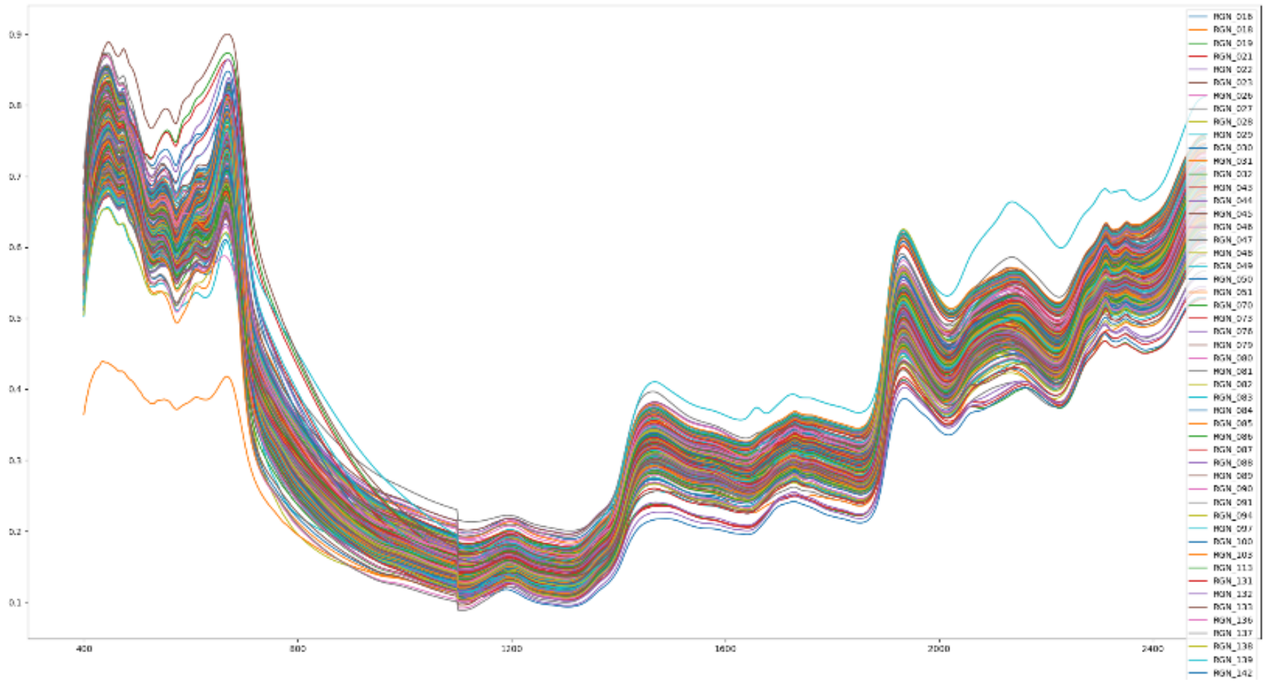


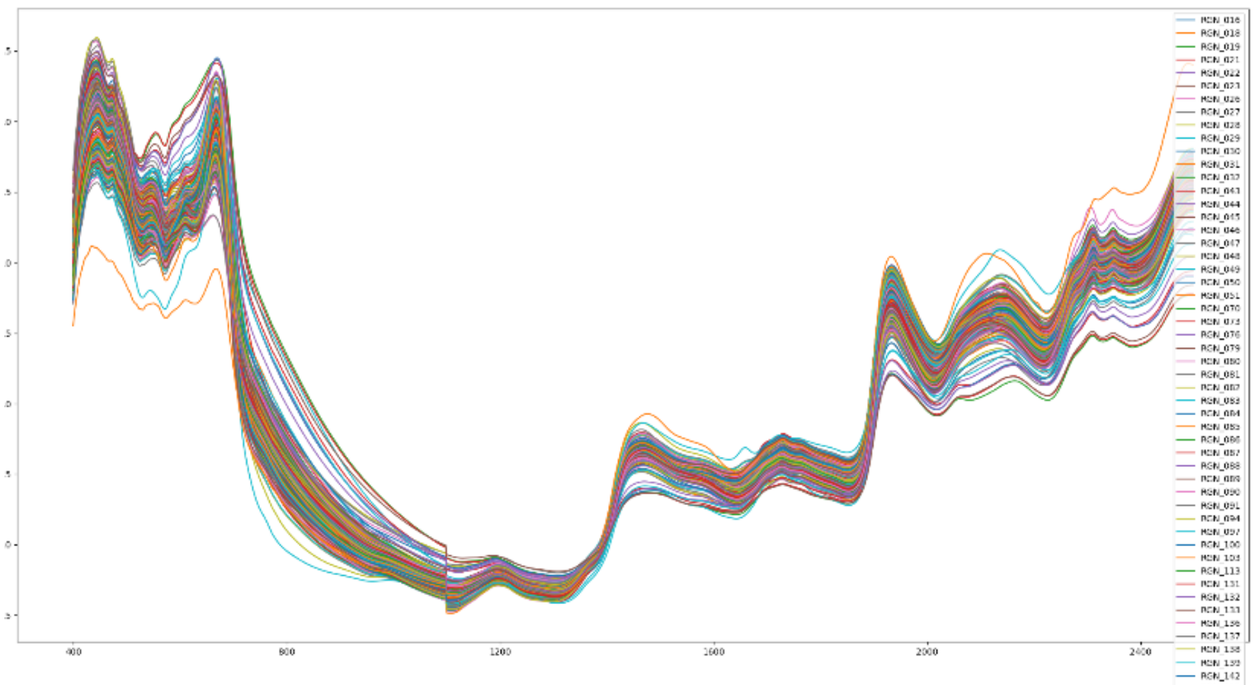
(pre-processing)

Data type: ‘2DSeries’ → Tabular data

Each row: 1 sample
Each column: certain derivative values in the series

SNV procedure





The obtained spectrum is normalized further by calculating the SNV (Standard Normal Variate). SNV is a weighted normalization procedure often used for NIR and MIR spectra which does not require a reference spectrum. Simply put, each spectrum is individually normalized by first mean-centering the spectrum by taking away it mean across the wavelengths. Then the mean centered spectrum is divided by its own standard deviation. Or for each spectrum X¬I , XiSNV =(Xi−mean(Xi))/ stdi .

First-derivative is taken, because the change is signal is more informative than the actual value (even after SNV)

HIT

In this setting, the hyperspectral camera using near-infrared wavelengths is used to create a 3-dimensional data file per sample, opposed to the processed NIR and MIR variants, where each sample is a single spectrum. The sample is measured with the result being a Height X Width X Wavelengths matrix.

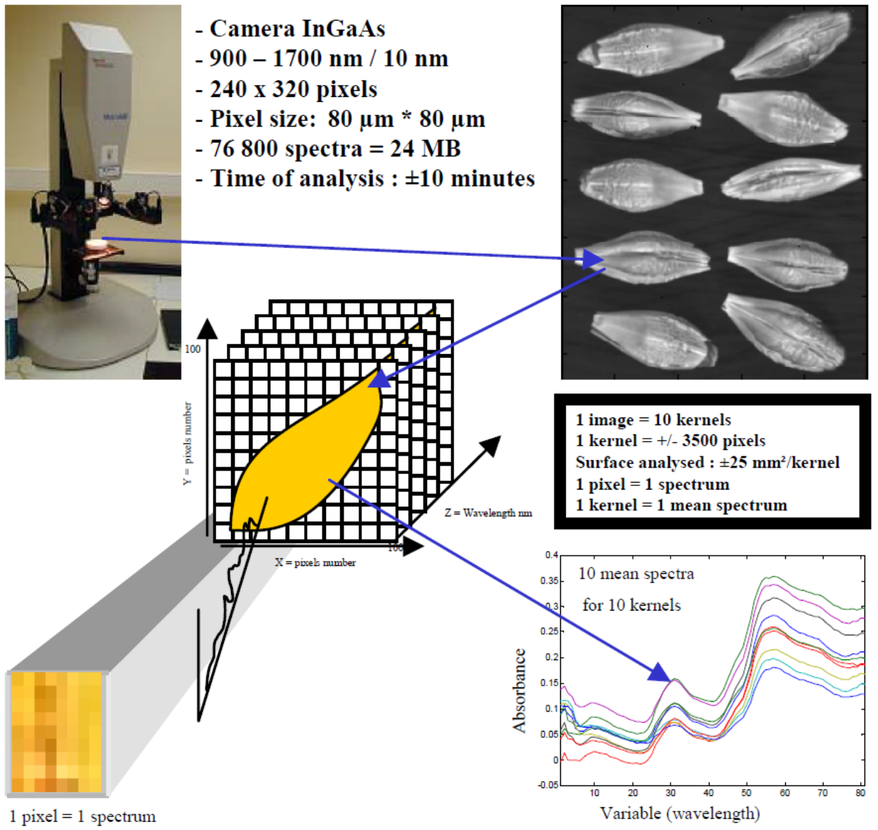
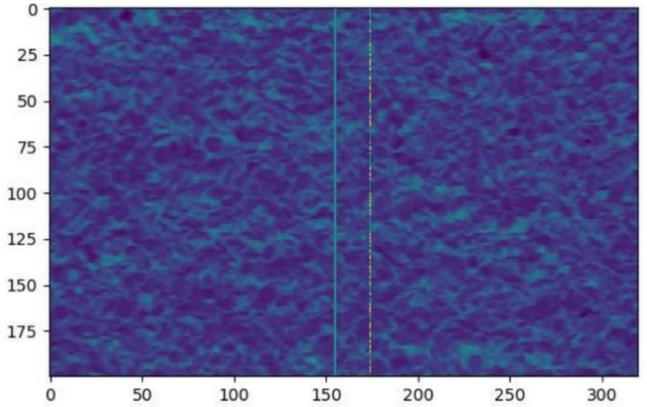


Figure 1. NIR camera and its characteristics.



(pre-processing,feature engineering)

Data type: ‘3DSeries’ → Tabular data

Each row: 1 sample
Each column: certain derivative values in the series

HIT feature engineering

We use the following procedure to process the data:

1. Only samples which were measured 3 times were used.
2. For each of the three measurements of a single sample in turn:
    a. We take each pixel (which results to a spectra) and SNV normalize it. Standard Normal Variate is a weighted normalization procedure often used for NIR and MIR spectra which does not require a reference spectrum. Simply put, each spectrum is individually normalized by first mean-centering the spectrum by taking away it mean across the wavelengths. Then the mean centered spectrum is divided by its own standard deviation. Or for each spectrum X_i, $X_i^{SNV} = (X_i - \text{mean}(X_i)) / \text{std}_i$. The effect of the SNV procedure can be seen in the MIR and NIR reports.
    b. All pixels spectra are then grouped into a single data matrix and several quantiles of the spectra are calculated. Namely the 0.03, 0.15, 0.25, 0.50, 0.75, 0.85, 0.97 quantiles.
3. We then average each of the quantiles over the three measurements, which can be done, as they are now spatially invariant.
4. We then write the extracted information to a .CSV file per RGN_XXX.

Next using the preprocessed .CSV files, we convert all files into a single data matrix by including the following features per sample. The data matrix contains rows which represent a single RGN_XXX sample and columns which indicate the features below.

1. The difference between the 0.03 quantile and the 0.97 quantile
2. The difference between the 0.15 quantile and the 0.85 quantile
3. The median spectrum
4. The first derivative of the median spectrum
5. The first derivative of the difference between the 0.03 quantile and the 0.97 quantile

Unsupervised analysis

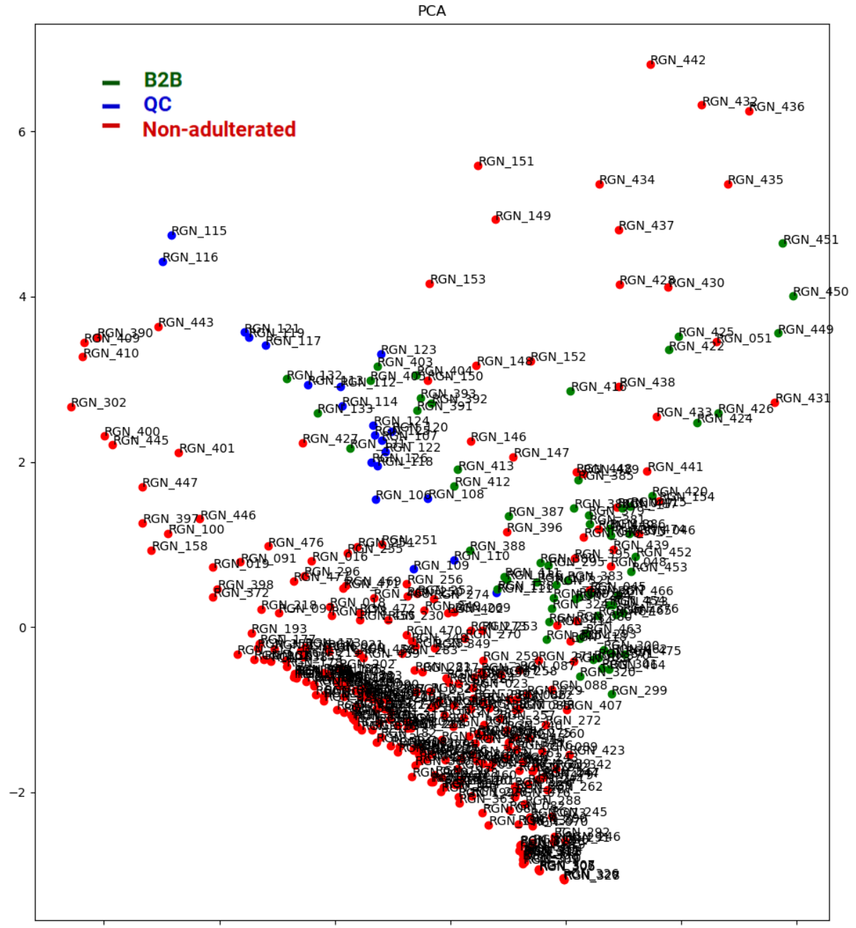
Why?

Data imbalance did not always allow for a supervised approach, as some classes were too small to split of a representative test-test

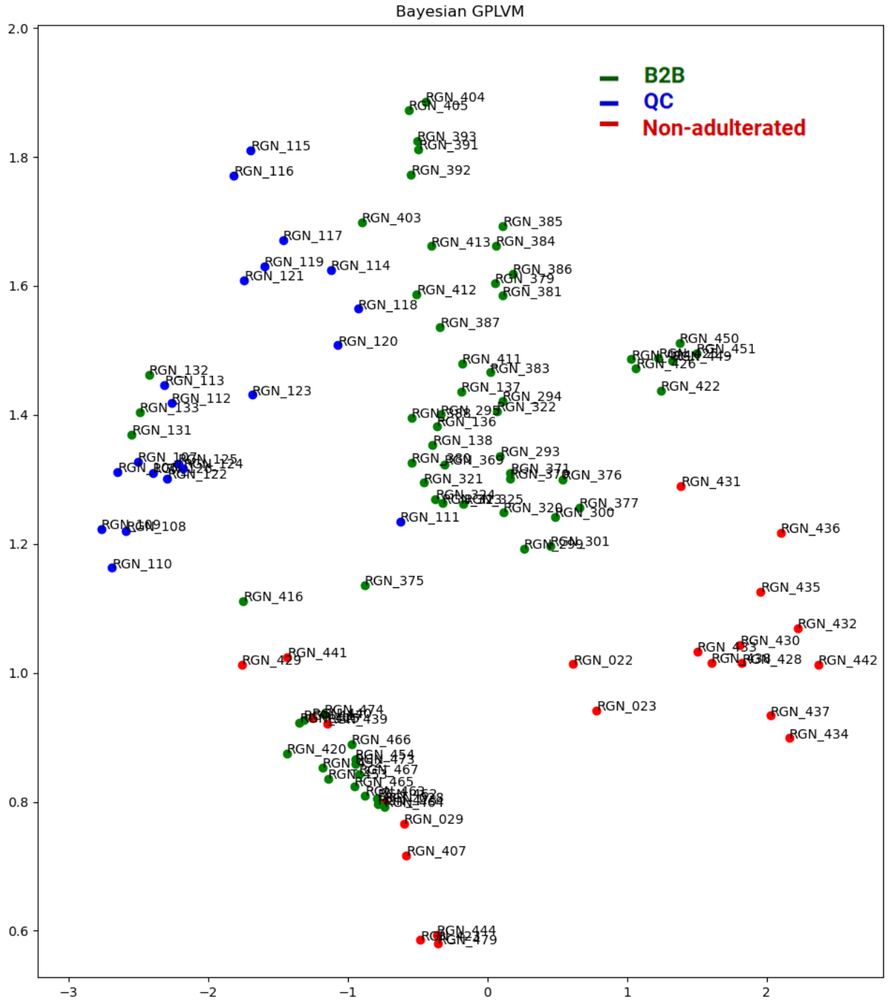
Unsupervised also allows us to see connections/correlations in the data, which we did not expect to see based on the labels

Exploratory analysis: allows to further interpret the data and look for biological conclusions, as opposed to black-box performance metrics

PCA versus GP-LVM

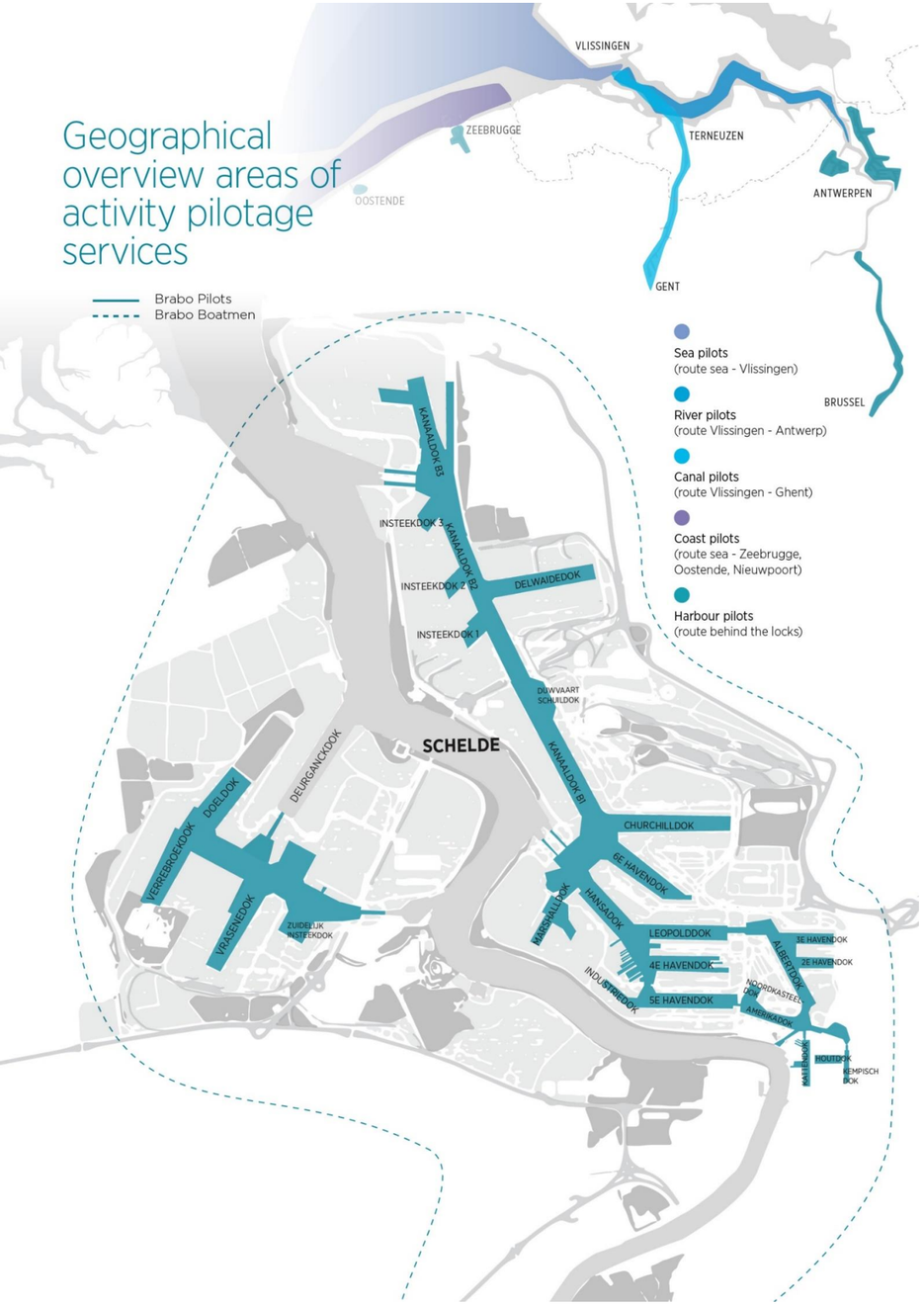


PCA



GPLVM

Geographical overview areas of activity pilotage services

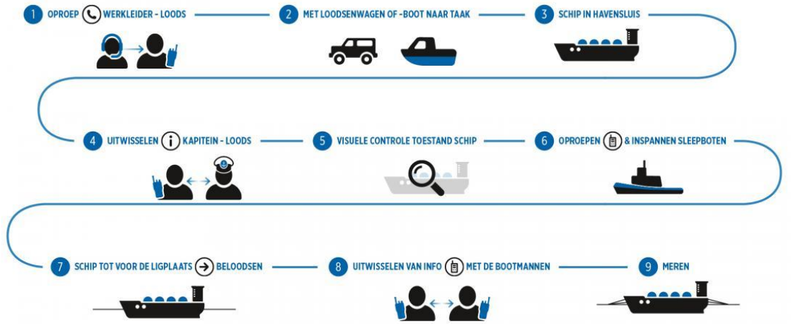


BRABO GROUP

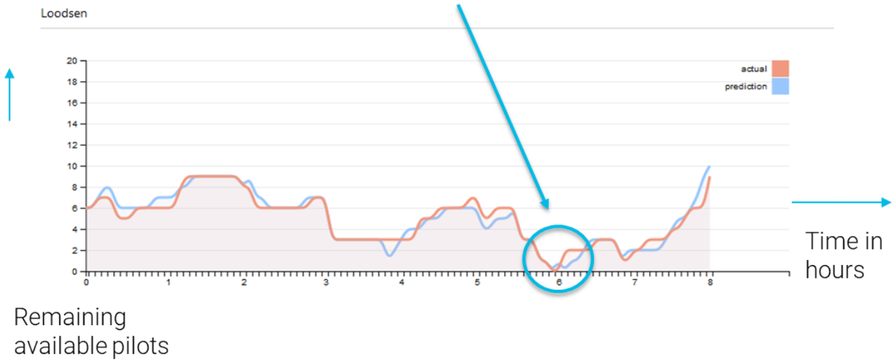
Boatsman: +- 200 personen
Pilots: +- 65 personen

2024: **26.000 ship movements**

Beloodsen van inkomende zeeschepen...



Pilot capacity prediction model demonstrated the ability to predict capacity shortages 6-8h ahead



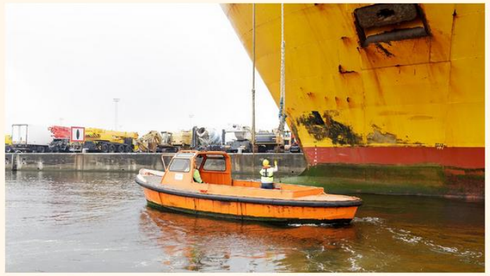
DE TIJD

Nieuws Markten LIVE Netto

NIEUWS > ONDERNEMEN > TECHNOLOGIE

Antwerpen lanceert 'Waze voor de scheepvaart'

TWITTER
FACEBOOK
WHATSAPP
LINKEDIN
E-MAIL
BEWAAR
SCHENK DIT ARTIKEL
REAGEER



Een bootman in de haven maakt een schip vast aan de kaai ©STUDIO CLAERHOUT

MARC DE ROO | 18 februari 2020 16:56

Elk schip dat te lang in de haven blijft, kost geld. Het havendienstenbedrijf Port+ en het loodsenbedrijf Brabo zetten algoritmes in om schepen zo vlug mogelijk in en uit de haven te krijgen.

Artificiële intelligentie

Binnenkort gaat Brabo nog een stap verder. Via artificiële intelligentie wil het voorspellingen maken van het werk dat op het bedrijf afkomt. En dat op basis van data uit het verleden die rekening houden met de tijdstippen waarop een schip wordt gesleept of aankomt in de sluis en aan de kaai'. De Groof: 'Heel belangrijk, als je weet dat we binnen het uur na een oproep een schip moeten kunnen beloodsen. Voor een werkleider is het onmogelijk acht uur vooruit te kijken. 26.000 scheepsbewegingen per jaar is gigantisch. De testen vorig jaar waren veelbelovend.'

'Elke periode dat er met een schip 'niets gebeurt', kost geld'

JAN VAN DOOREN
CEO PORT+

Deel op

Het systeem moet helpen om capaciteits- en personeelstekorten beter in te schatten, automatisch te plannen en if-scenario's in te stellen als iets erg voorvalt. 'Nu zetten we personeel in op basis van pieken', zegt De Groof. 'Het gaat permanent om 40 bootsmannen en 20 loodsen, 24 op 24 uur, zeven op zeven. Soms hebben die geen werk. Met AI kunnen we onze mensen accurater inzetten. In 2018 hebben we 8.000 keer mensen dringend thuis moeten oproepen, nu is ons streefdoel 5.000.'

'Brabo is een van onze beste voorbeelden van hoe je met kleine AI-ingrepen grotere veranderingen kunt teweegbrengen', zegt Joeri Ruyssinck van ML2Grow, dat de AI-toepassing bij Brabo bedacht. 'Met een minimum aan investeringen heb je een quick win.'