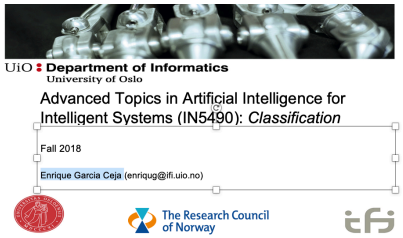
advanced issues in data preparation and modeling

Carlos Soares

(partly using materials kindly provided by José Luís Borges, from Han, Kamber & Pei, from Moreira, Carvalho & Horvath and from Ceja)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
- … and modeling
  - dealing with unbalanced classes
- final reflections on data quality

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
    - context
    - attribute aggregation
    - feature selection
- … and modeling
  - dealing with unbalanced classes
- final reflections on data quality

data reduction

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- obtain a reduced representation of the data set that is
  - **much smaller** in volume
- … producing the **same analytical results**
  - or almost the same
- … improved visualization of data
- … with more interpretable models
- … much faster

dimensionality reduction: data is cursed!

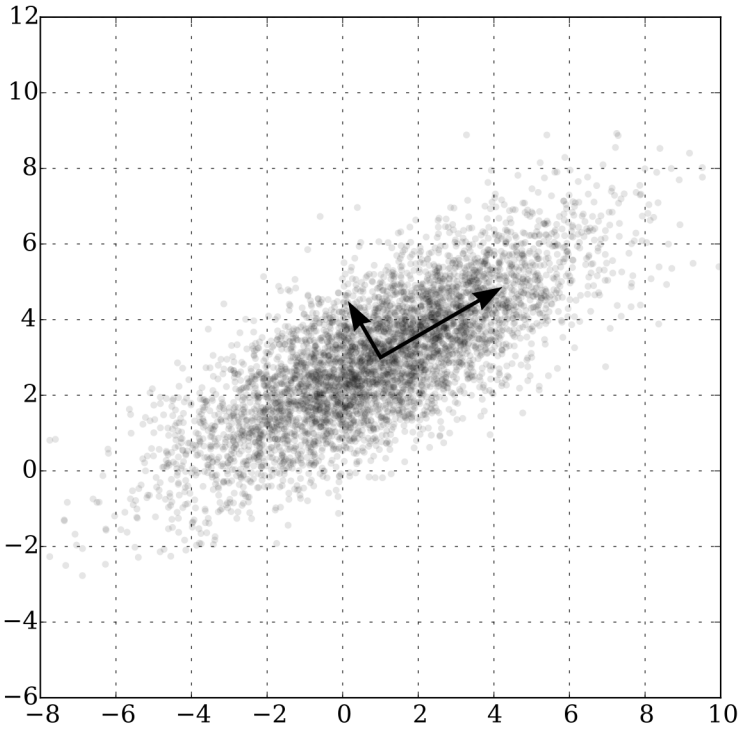
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- curse of dimensionality
  - when dimensionality increases, data becomes increasingly sparse
  - density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - the possible combinations of subspaces will grow exponentially
- … number of data points required for robust patterns grows exponentially with number of attributes

two approaches

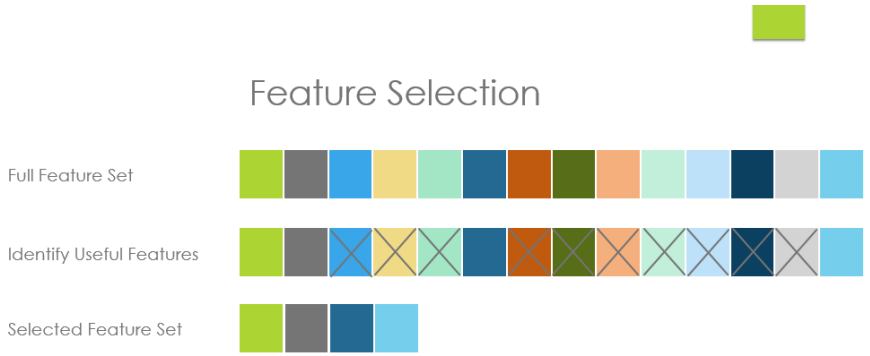
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

attribute aggregation



By Nicoguaro - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=46871195

feature selection

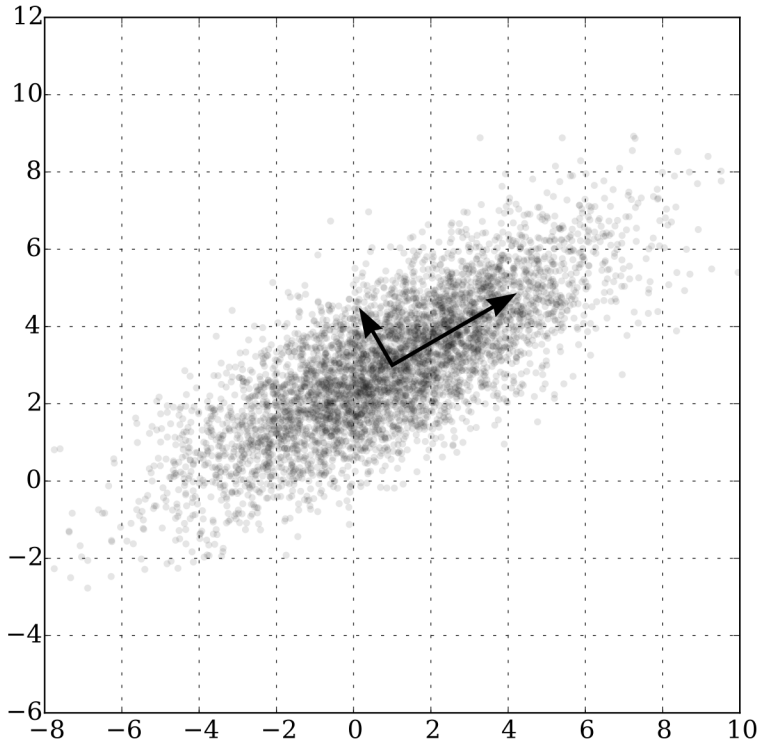


https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96

attribute aggregation: PCA

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Principal Component Analysis
  - *n* new features
  - linear combinations of existing *n* attributes
  - orthogonal to each other
  - *k* << *n* explain most of the variance in the data

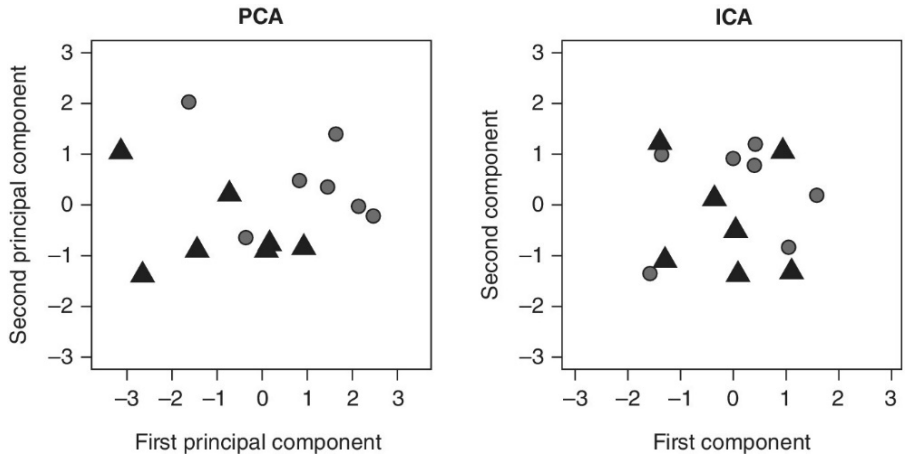


By Nicoguaro - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=46871195

attribute aggregation: ICA vs PCA

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

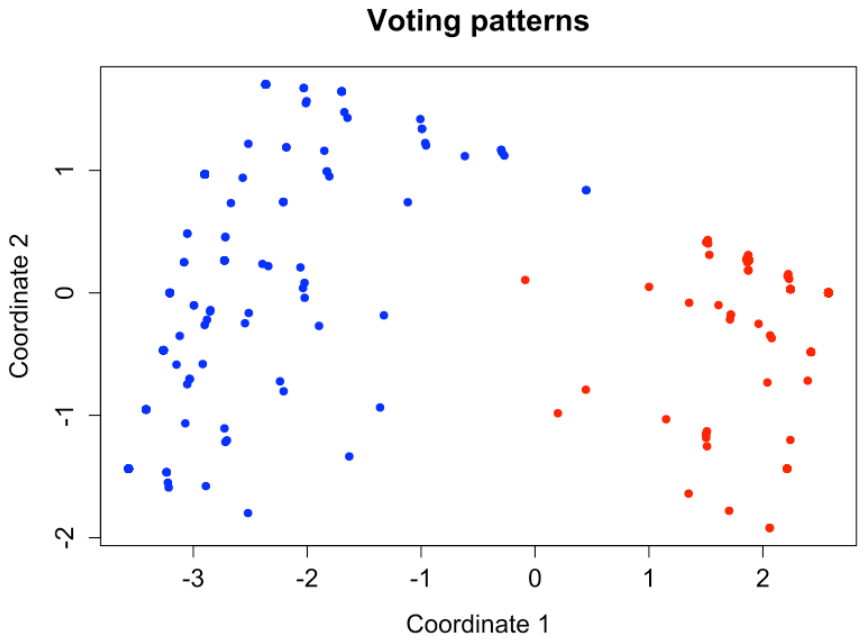
- both create linear combinations of the attributes
- ICA assumes the original attributes are statistically independent
- … reduces higher order statistics
  - e.g. kurtosis
- … does not rank components



attribute aggregation: multidimensional scaling

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- linear projection of a data set
- uses the distances between pairs of objects
  - not the values of the attributes of the objects
- particularly suitable when it is difficult to extract relevant features to represent the objects



https://en.wikipedia.org/wiki/Multidimensional_scaling

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
    - context
    - attribute aggregation
    - feature selection
- … and modeling
  - dealing with unbalanced classes
- final reflections on data quality

feature selection: eliminate...

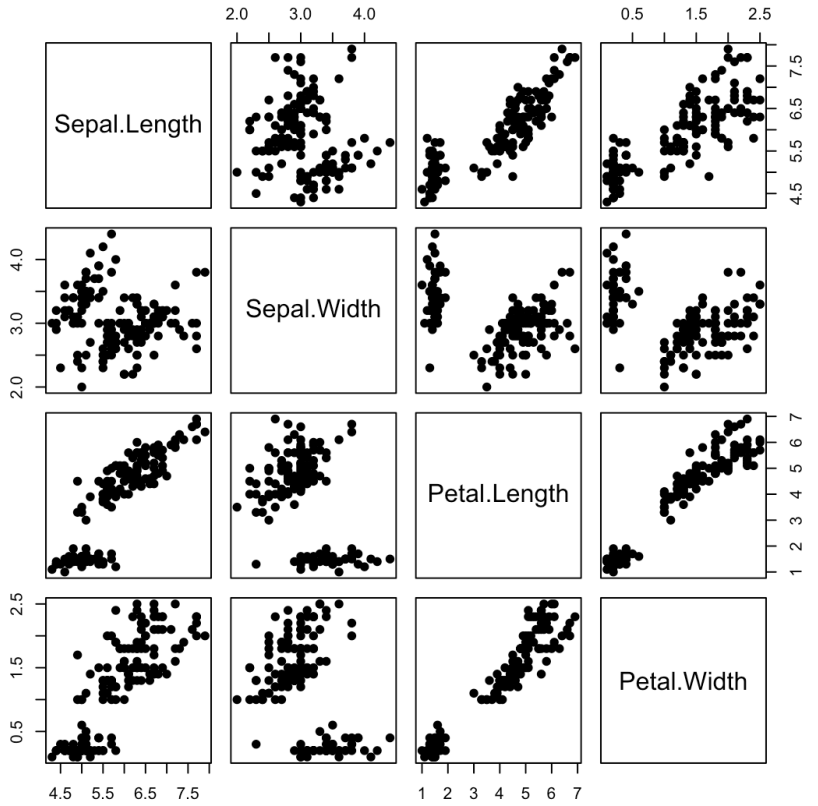
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- redundant attributes
  - duplicate much or all of the information contained in one or more other attributes
    - e.g. purchase price of a product and the amount of sales tax paid
- irrelevant attributes
  - contain no useful information
    - e.g. students' ID is often irrelevant to the task of predicting students' GPA

feature selection: filter methods

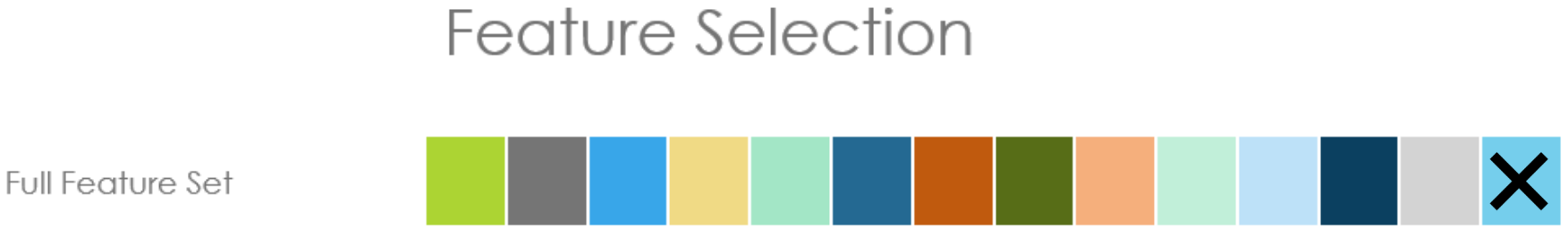
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- 2 attributes
  - remove redundant attributes
- 1 attribute vs target
  - identify relevant variables



feature selection: wrapper methods (1/4)

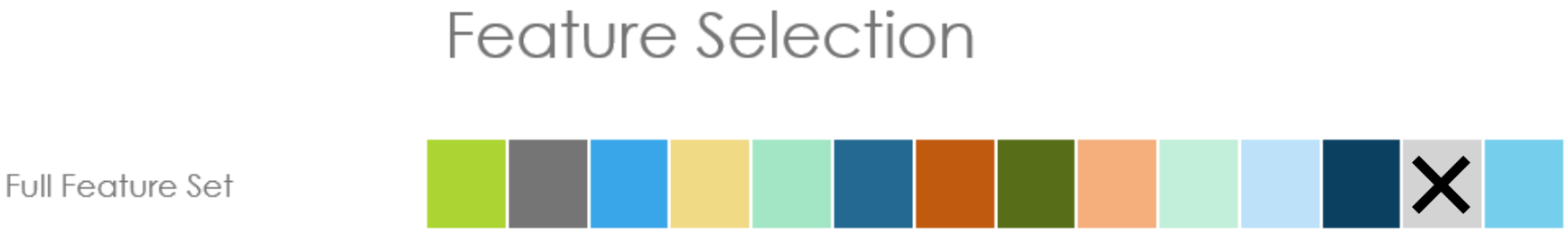
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96

feature selection: wrapper methods (2/4)

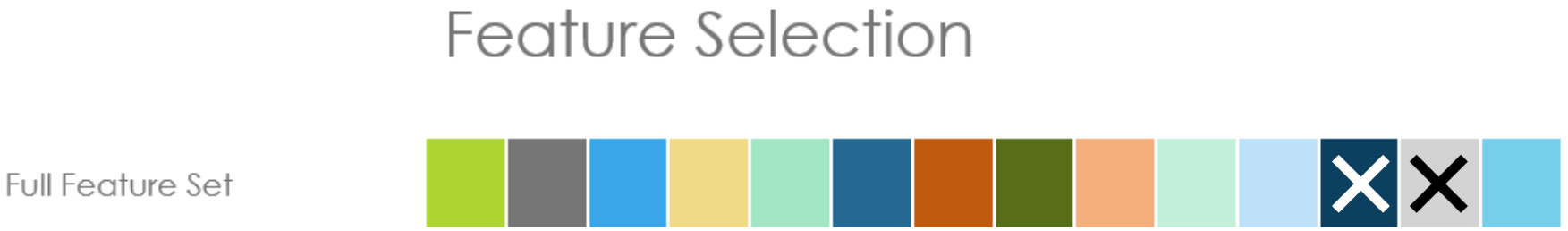
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96

feature selection: wrapper methods (3/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96

feature selection: wrapper methods (4/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96

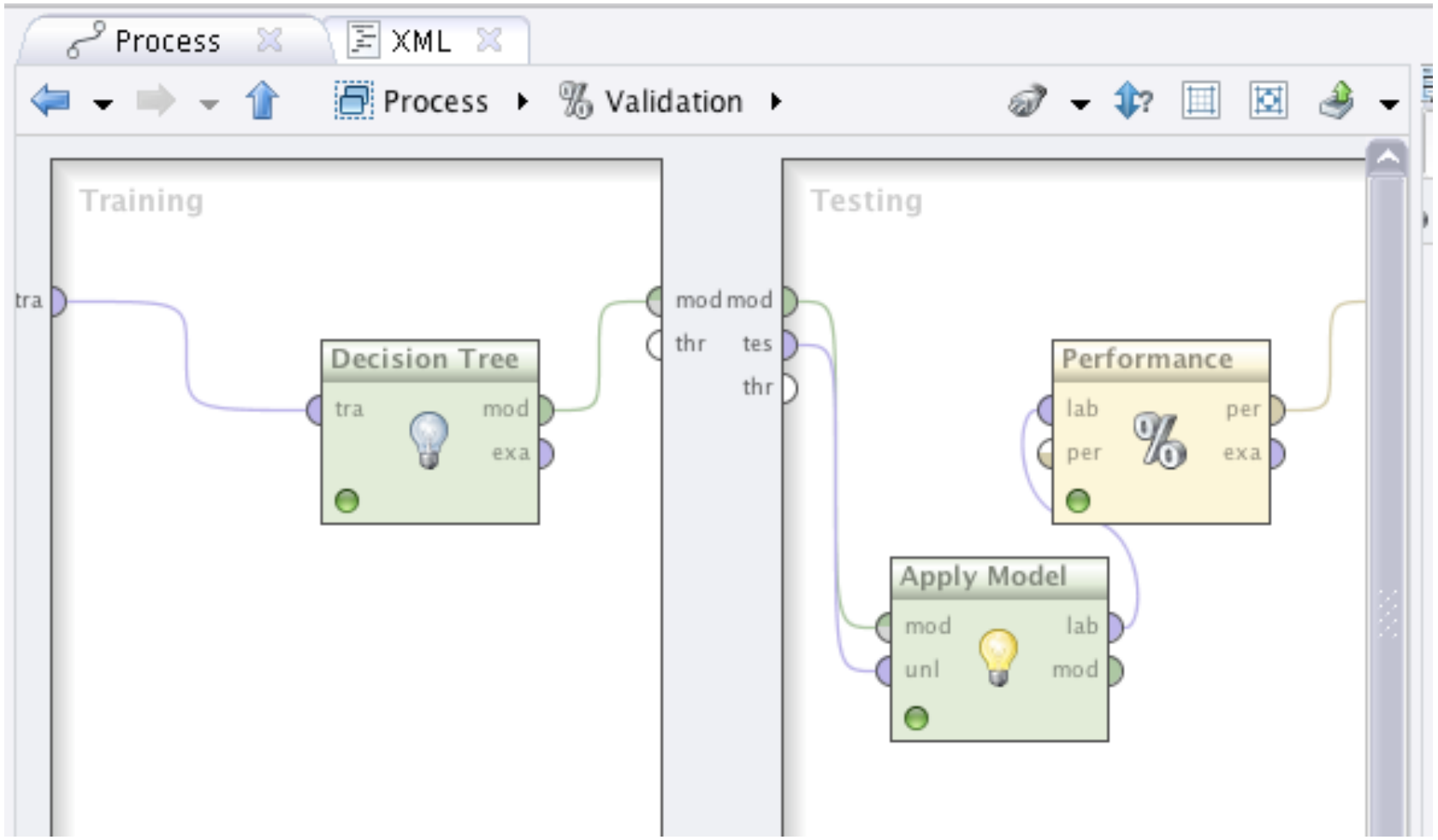
gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
- … and modeling
  - dealing with unbalanced classes
    - context
    - resampling
    - create synthetic data with SMOTE
    - cost sensitive learning
- final reflections on data quality

you developed a model for the competition…

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

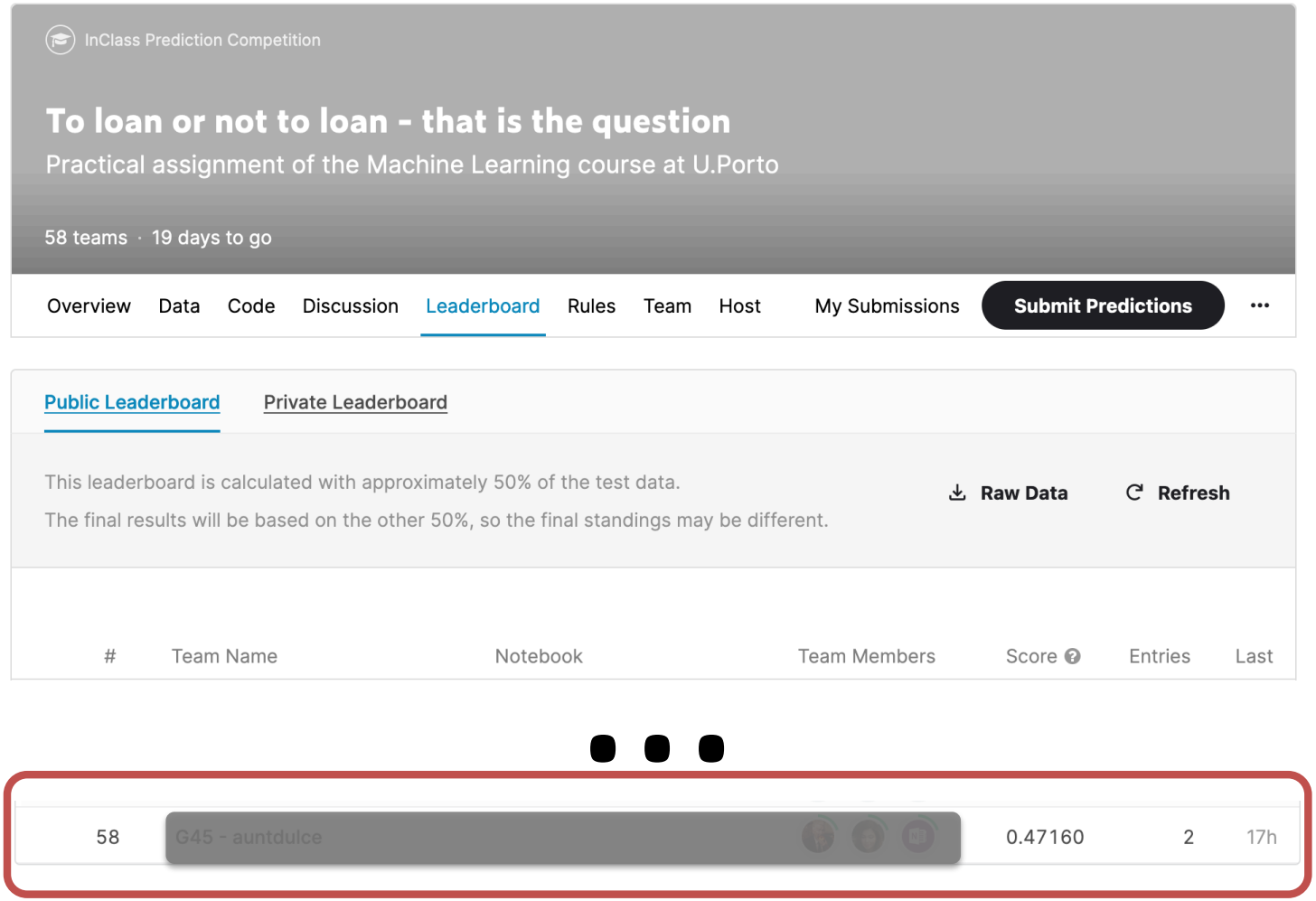


accuracy = 99.9%!!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

… wth?!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



all predictions = negative class!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

maybe I should have done the exploratory data analysis

(as advised by the lecturers…)

so what?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

		Predicted class	
		Yes	No
Actual class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

- ML methods usually minimize FP+FN
  - … or, at the very least give the same weight to both types of errorsa
- … but potentially FP >> FN
  - i.e. quality of the model more affected by FP
- … so algorithm effectively minimizes FP!
- … and there's an easy model for that
  - prediction = N

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

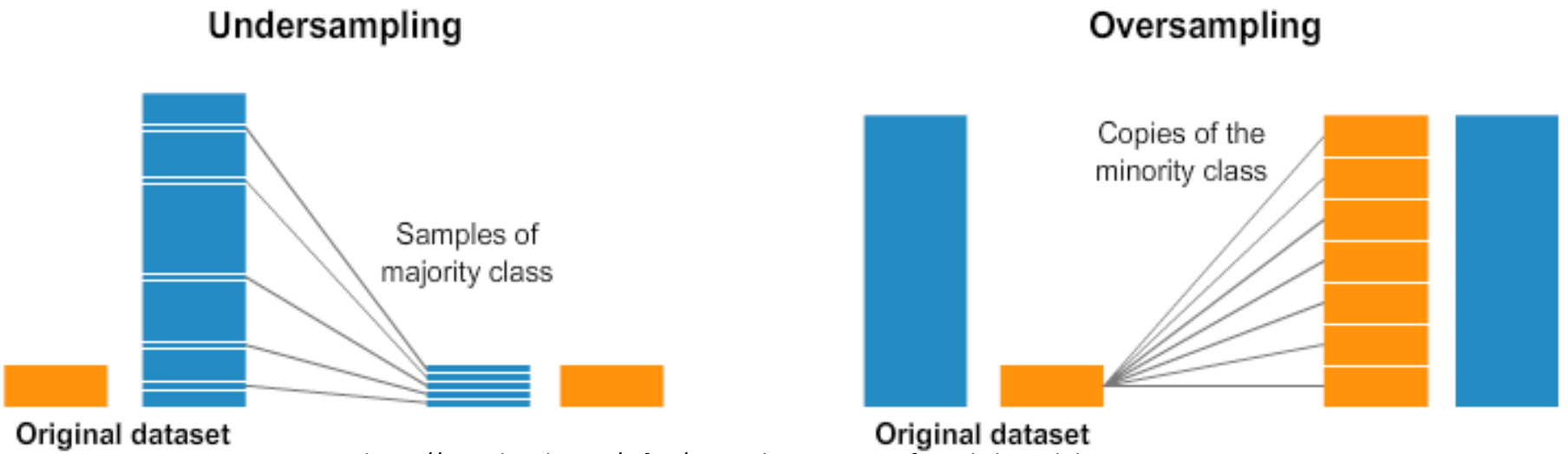
- advanced issues in data preparation
  - data reduction
- … and modeling
  - dealing with unbalanced classes
    - context
    - resampling
    - create synthetic data with SMOTE
    - cost sensitive learning
- final reflections on data quality

learning with class imbalance

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- collect more data
  - difficult in many domains
- resample existing data
  - delete data from the majority class
  - duplicate data from the minority class
- create synthetic data
  - e.g. SMOTE
- adapt your learning algorithm
  - e.g. cost sensitive learning

resampling



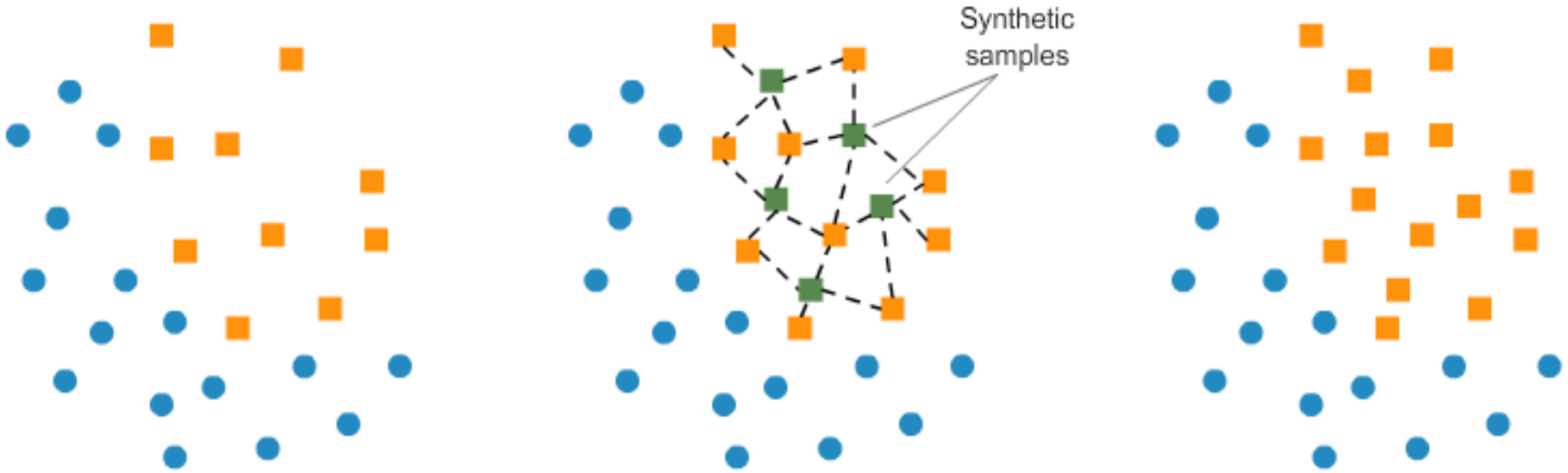
https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1

- … but possible loss of information
- … but fixed boundaries
- … and danger of overfitting

furthermore, what is the best ratio?

SMOTE
(Synthetic Minority Over-sampling Technique)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

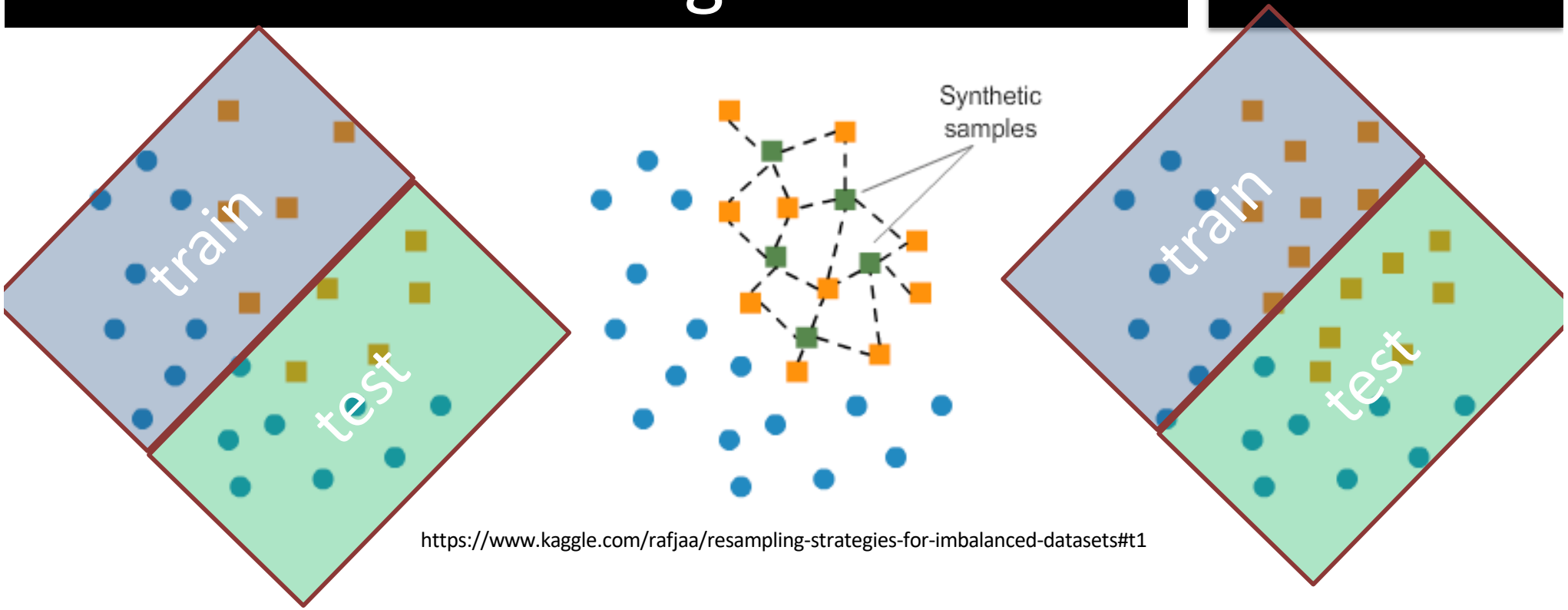


https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1

- … but possibility of inadequate boundaries
  - what happens if minority observations are too far apart?
- … and danger of overfitting

SMOTE + lack of basic statistical knowledge?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1

which one?

(or both?)

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
- … and modeling
  - dealing with unbalanced classes
    - context
    - resampling
    - create synthetic data with SMOTE
    - cost sensitive learning
- final reflections on data quality

the cost of errors

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

		Predicted class	
		Yes	No
Actual class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

- FP and FN errors often incur different costs
  - medical diagnostic
  - loan decisions
  - marketing campaigns
  - fraud detection in bank transactions
  - fault detection in machines
- … but ML methods still usually minimize FP+FN

error vs missclassification costs: medical diagnosis example

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

unnecessary suffering + more expensive procedures + eventually death
e.g. 100

		Predicted class	
		Yes	No
Actual class	Yes	10	5
	No	5	80

unnecessary exams & anxiety
e.g. 5

- error = 10 / 100 = 10%
- missclassification costs = 5 x 100 + 5 x 5 = 525
- … per pacient
  - 5.25

cost-sensitive learning

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- simple methods
  - resampling according to costs
  - weighting according to costs
    - basically, the same thing
- complex methods
  - e.g. metacost

metacost

1. create bootstap replicates of training data
2. learn model from each replicate
3. relabel examples

$$\text{argmin}_i \sum_j P(j|x)C(i,j)$$

   - *C(i | j)* = cost of mistaking *j* by *i*
   - *P(j | x)* = class probability of *x* by voting
4. learn model on relabelled data

- independent of algorithm

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- advanced issues in data preparation
  - data reduction
- … and modeling
  - dealing with unbalanced classes
- final reflections on data quality

coming next...

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- data
- cleaning
  - data quality: a data scientist's worst nightmare
  - quality issues
  - can we do better?
- integration
- reduction
- transformation and discretization
- challenges

data quality: multidimensional view

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- accuracy
  - correct or wrong, accurate or not
- completeness
  - not recorded, unavailable, …
- consistency
  - some modified but some not, dangling, …
- timeliness
  - timely update?
- believability
  - how trustable is the data and its sources?
- interpretability
  - how easily the data can be understood?

a data scientist's worst nightmare

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

no worries: our data is clean!

https://pt.wikipedia.org/wiki/RMS_Titanic#/media/File:St%C3%B6wer_Titanic.jpg

no worries: our data is clean! (1/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

we have a data warehouse

- DWs are built with a different purpose
  - descriptive
  - aggregated data
    - typically

https://pixabay.com/en/archive-bookcase-boxes-business-1850170/

no worries: our data is clean! (2/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

our IS was just revamped

- how far was analytics taken into account in the process?

https://pixabay.com/en/confused-muddled-illogical-880735/

no worries: our data is clean! (3/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

we had a major data cleanup

- how was project success measured?

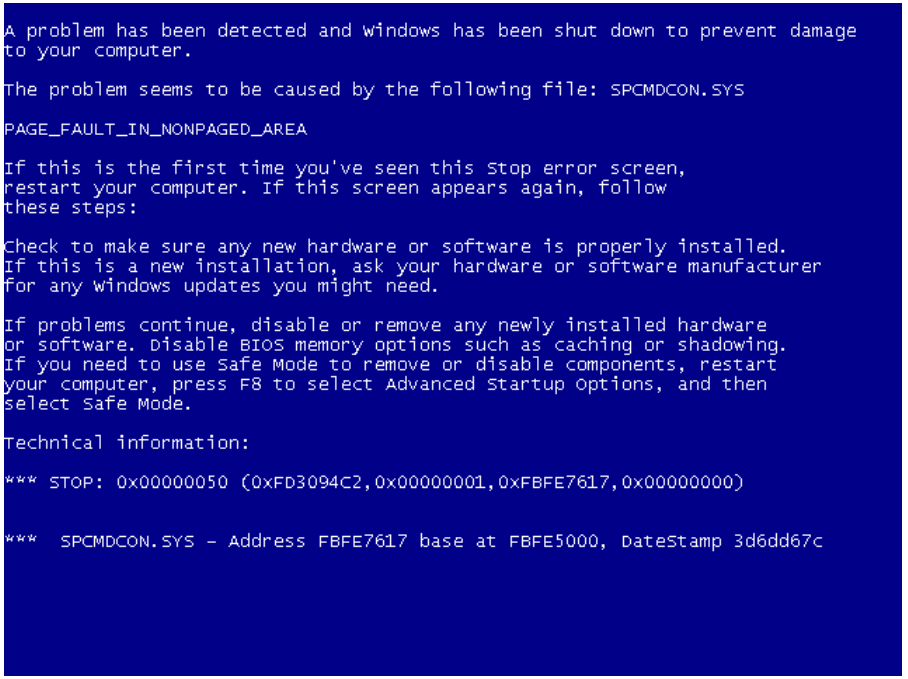
https://pixabay.com/en/clean-dare-cleaning-1706439/

no worries: our data is clean! (4/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

our data is collected automatically

- ... and we all know that machines never break



https://en.wikipedia.org/wiki/Blue_Screen_of_Death#/media/File:Windows_XP_BSOD.png

no worries: our data is clean! (5/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

our data collection is human-error proof

- “Data errors, uh, find a way”

http://knowyourmeme.com/memes/life-uh-finds-a-way

no worries: our data is clean! (6/6)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

tell us what you need: we have everything

- should be read as

"if something goes wrong, it's your fault"

- … often associated with

"you do magic, right?"

https://en.wikipedia.org/wiki/Marvin_(character)#/media/File:Marvin_(HHGG).jpg

can we do any better? (1/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

human resources

- remember when IT director was not a C-level job?

Fair use, https://en.wikipedia.org/w/index.php?curid=24782741

can we do any better? (2/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

analytics at the core of IS development

- requirements analysis includes analytics
- analytics components built in the same process as the rest of the functionalities

- we don't mind that systems still do old-fashioned tasks
  - sell stuff, pay salaries, etc.

https://pixabay.com/en/scaffolding-workers-construction-1617969/

can we do any better? (3/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

data quality is a continuous process

- data steward
- ... the sexiest job of the XXII century?

https://pixabay.com/en/lost-places-toilet-urinal-pforphoto-1610652/

data cleaning as a process

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- discrepancy detection
  - validate with metadata (e.g., domain, range, dependency, distribution)
  - check field overloading
  - check uniqueness rule, consecutive rule and null rule
  - commercial tools
    - scrubbing: use simple domain knowledge to detect errors and make corrections
      - e.g. postal code, spell-check
    - auditing: discover rules and relationship to detect violators
      - e.g. correlation and clustering to find outliers
- migration and integration
  - data migration tools
    - allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools
    - allow users to specify transformations through a graphical user interface

Summary

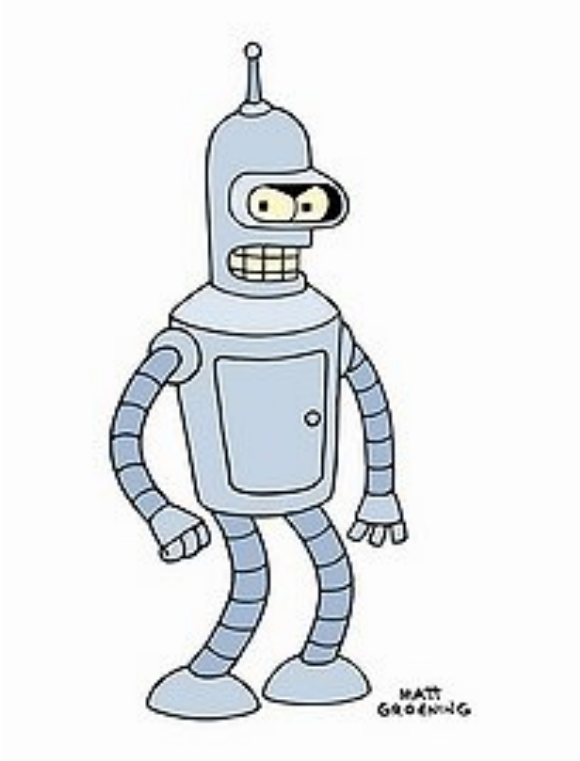
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- data
- ... quality
  - accuracy, completeness, consistency, timeliness, believability, interpretability
- … cleaning
  - e.g. missing/noisy values, outliers
- integration from multiple sources
  - entity identification problem is challenging
- reduction
  - curse of dimensionality and dimensionality reduction
  - numerosity reduction
- transformation and discretization

automation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- automl & metalearning
  - some progress on algorithm selection
  - ... early work on workflow
    - including data preparation
  - ... not really data cleaning

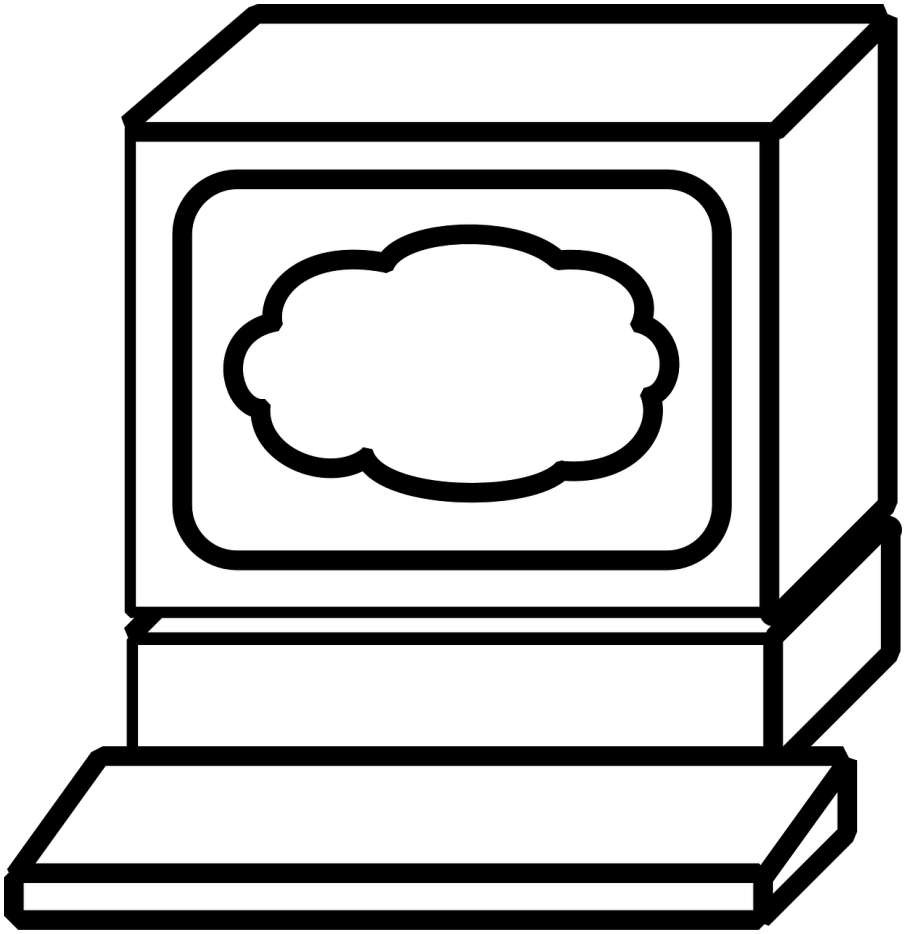


https://pt.wikipedia.org/wiki/Ficheiro:Bender_(personagem).jpg

DQaaS?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- if automation is possible
  - DQ can become a commodity?
- perhaps there is hope?
- ... many issues
  - confidentiality
  - computational costs



https://pixabay.com/en/computer-server-internet-network-294036/