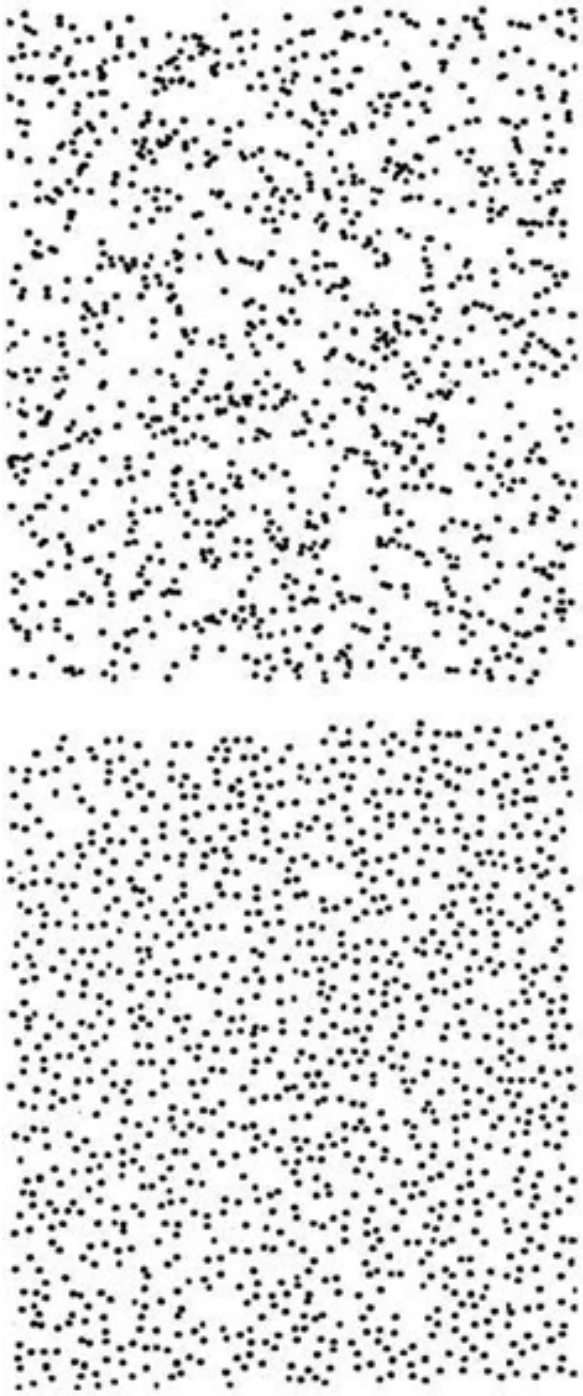


Data Science & Visualization

Exploratory Data Analysis

Gabriela Molina León
molina@uni-bremen.de

Institute for Information Management Bremen
Information Management Group (AGIM)

University of Bremen

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
08.04.24	Introduction to Data Science	Python Introduction
15.04.24	Basic Statistics & Supervised Learning	Practical Statistics + Supervised
22.04.24	Unsupervised Learning	Practical Unsupervised
29.04.24	Introduction to Data Visualization	Guest Lecture on NLP by Oxana Vitman
06.05.24	Exploratory Data Analysis	Data Science & Vis Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
13.05.24	Visual Encoding	Exposé Workshop
20.05.24	Pentecost	Pentecost
27.05.24	Design Studies	Exposé Presentation
03.06.24	Interaction Techniques	Practical Interaction
10.06.24	Data Science & Journalism	Progress Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
17.06.24	Machine Learning & Visualization	Project Workshop
24.06.24	Guest Lecture on Critical Data Studies by Paola Lopez	Observable (Plot)
01.07.24	Final Session: Recap	Final Project Presentation

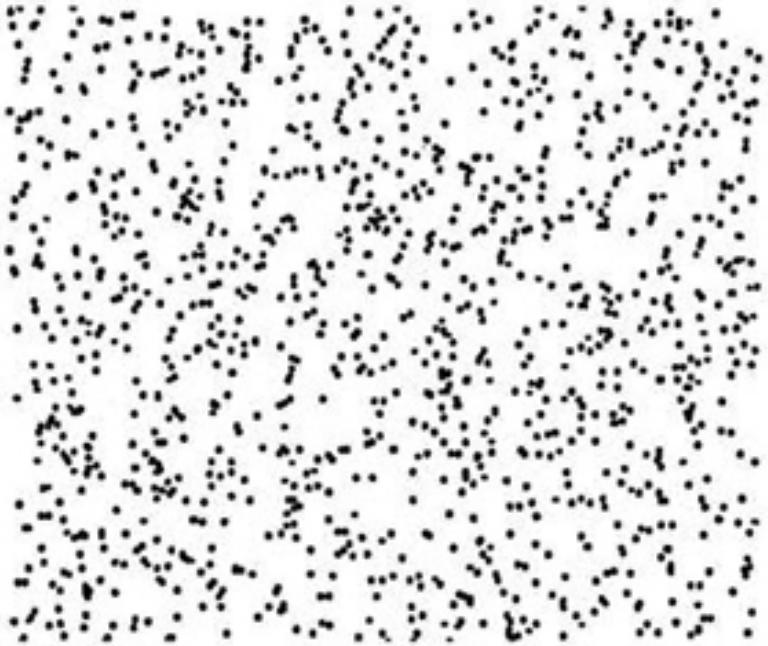
19.07.24 Deadline Final Report

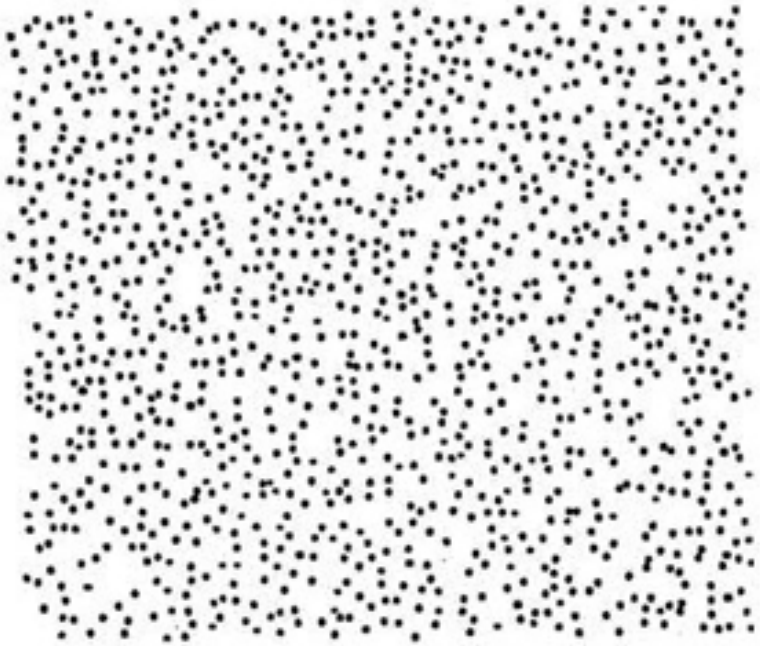
Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Exploratory Data Analysis: Goals

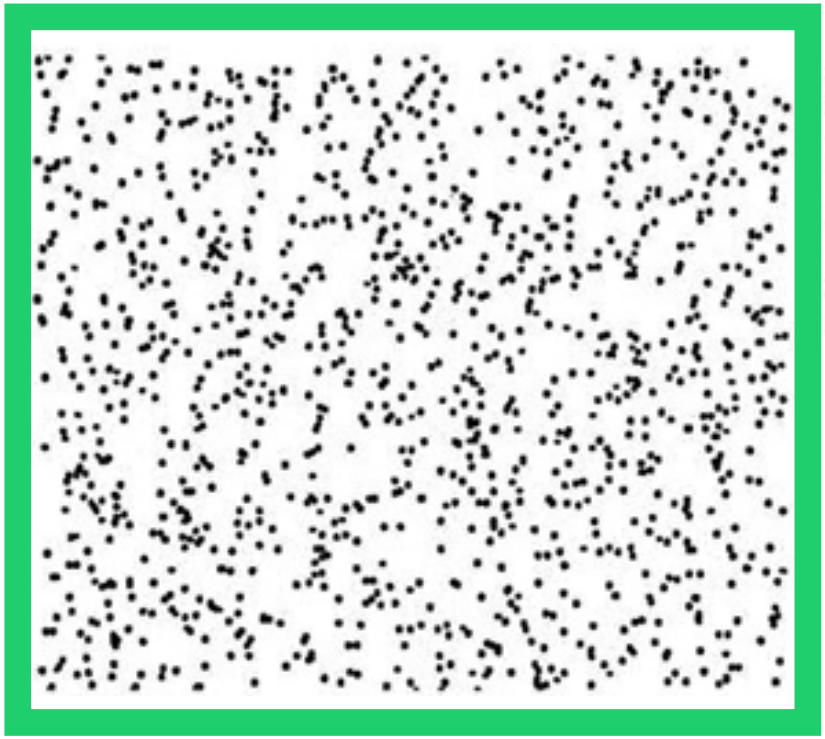
- Analyze the dataset to answer your questions using statistical methods and data visualization
- Build a machine learning model to make predictions of the future

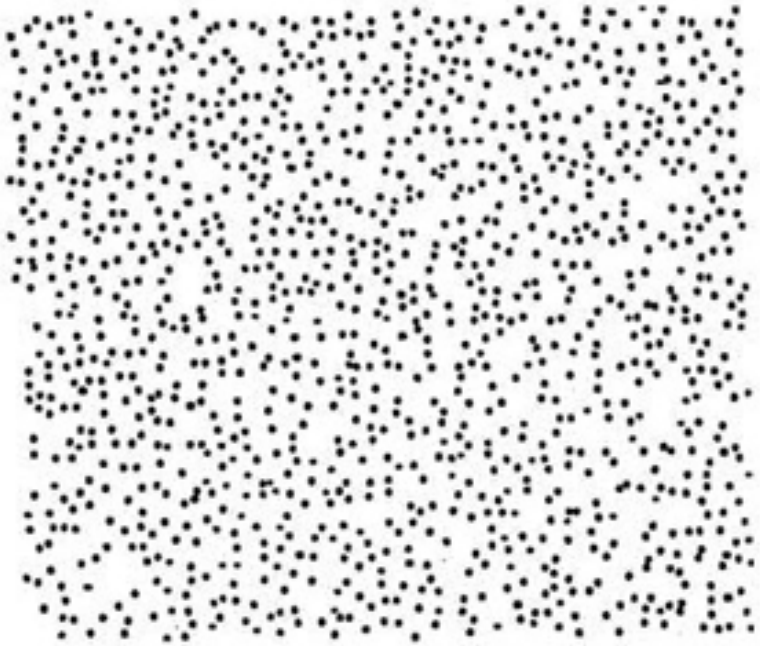
Which is more random?



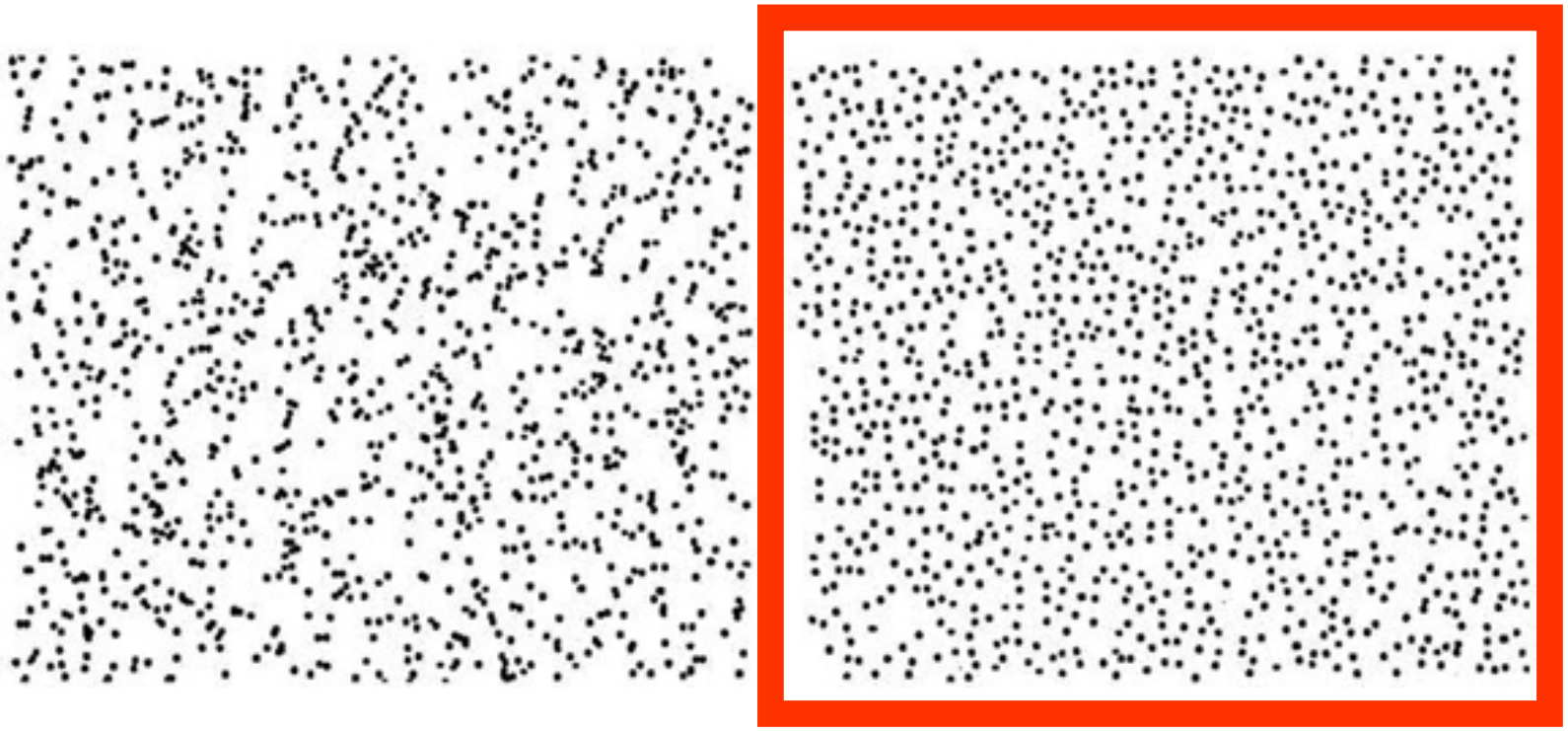


This is random





This is not random



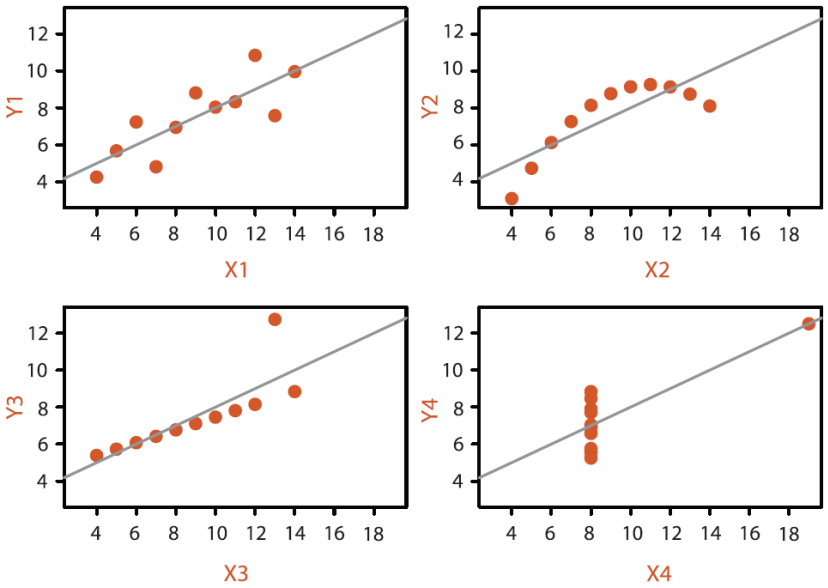
Five-number summary

- Mean: The sum of the values divided by the count of non-missing observations
  - Median: The number exactly in the middle of an ordered list of numerical values
- Minimum value
- Maximum value
- First quartile (q1)
- Third quartile (q3)

Visualize your data

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



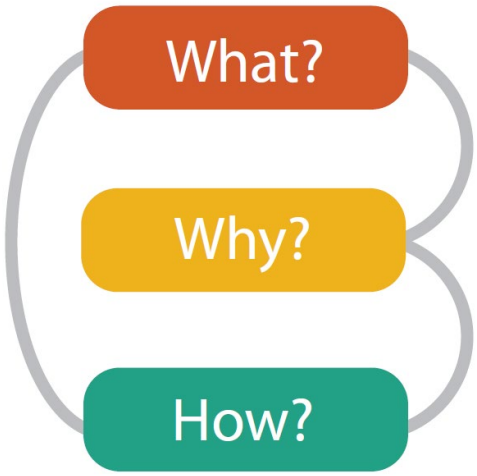
What to look at first?

- Descriptive statistics
- Five-number summary (mean/median, min, max, q1, q3)
- Histograms
- Box plots
- Scatterplots

What to look at first?

- Make use of the Python libraries available
  - Numpy
  - Pandas
  - Matplotlib
  - …

Munzner's framework



What data?

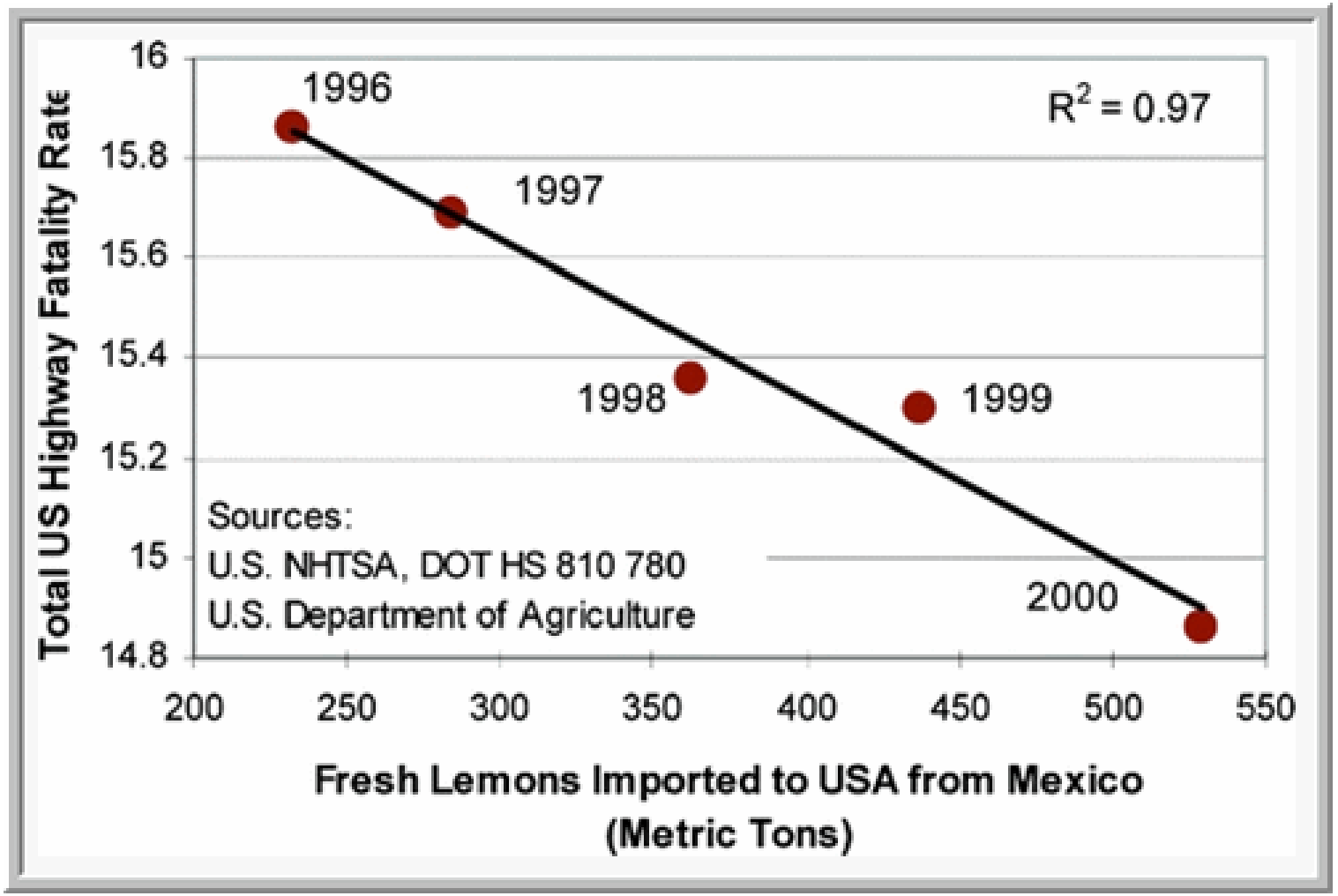
For whom?

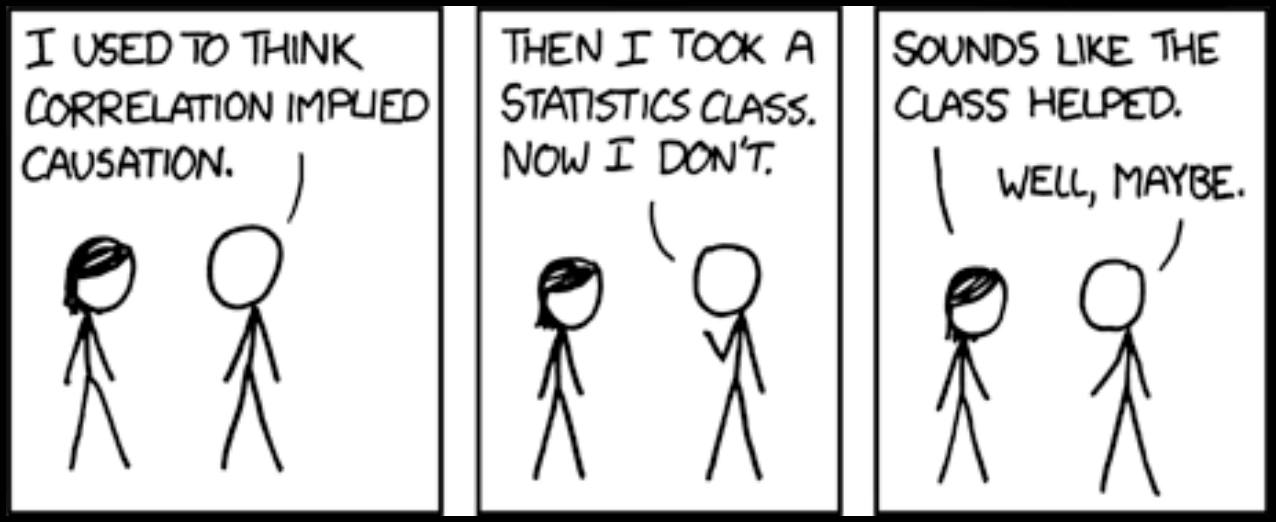
Which task will they perform?

How to encode the data visually?

How to interact with it?

Spurious correlations





Correlation **does not** imply causation

Exploratory Data Analysis: Goals

- Analyze the dataset to answer your questions using statistical methods and data visualization
- Build a machine learning model to make predictions of the future

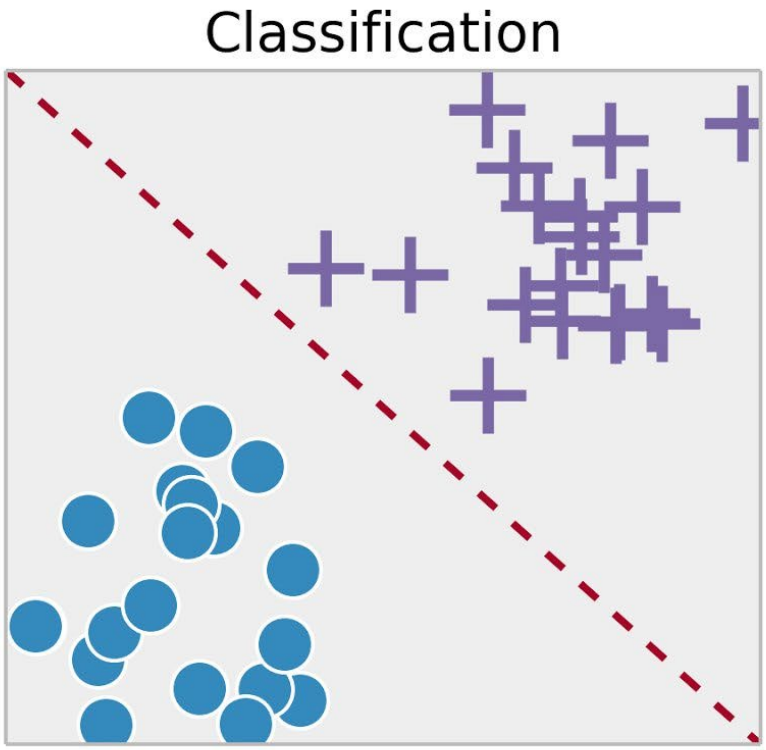
Supervised learning

- We have features (a X) and we have classes (a Y)
- The goal is prediction

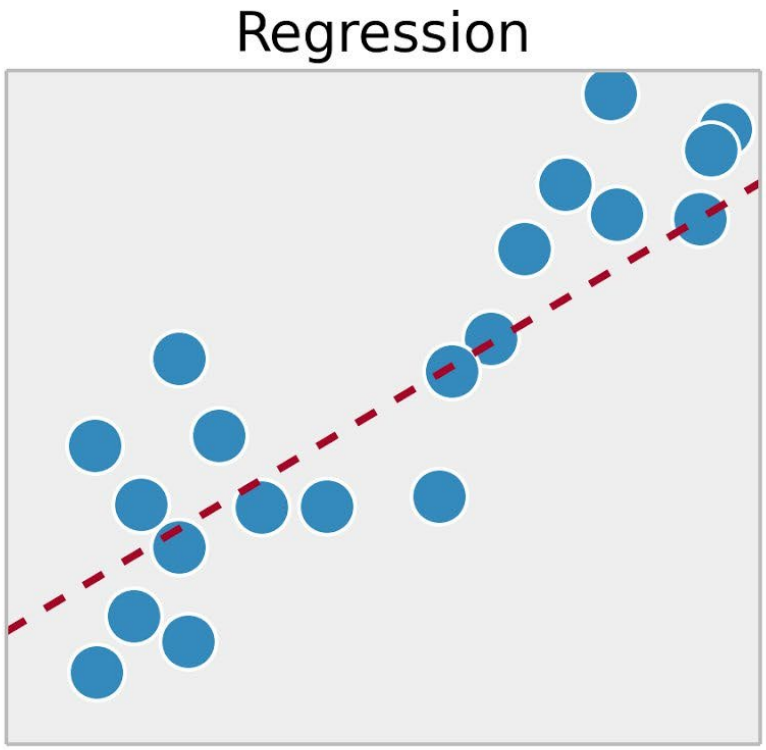
Unsupervised learning

- We have features (a X)
- The goal is exploration

Supervised learning



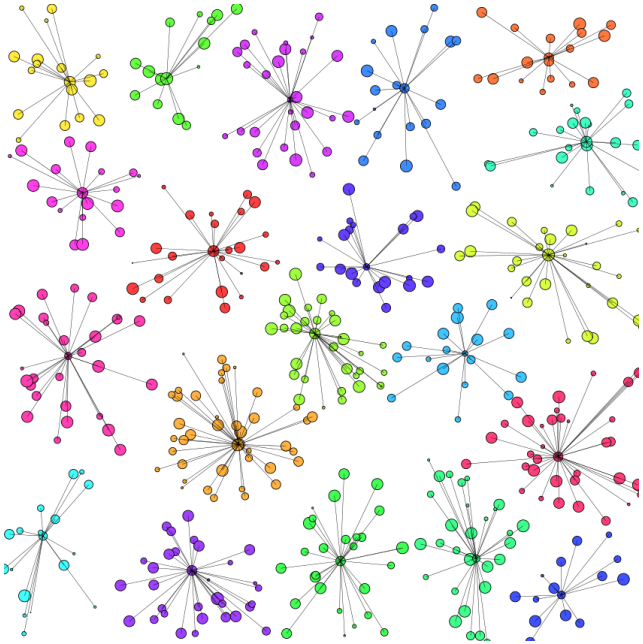
Predict the category of the items
(e.g., benign or malign)



Predict the value of the items
(e.g., house price)

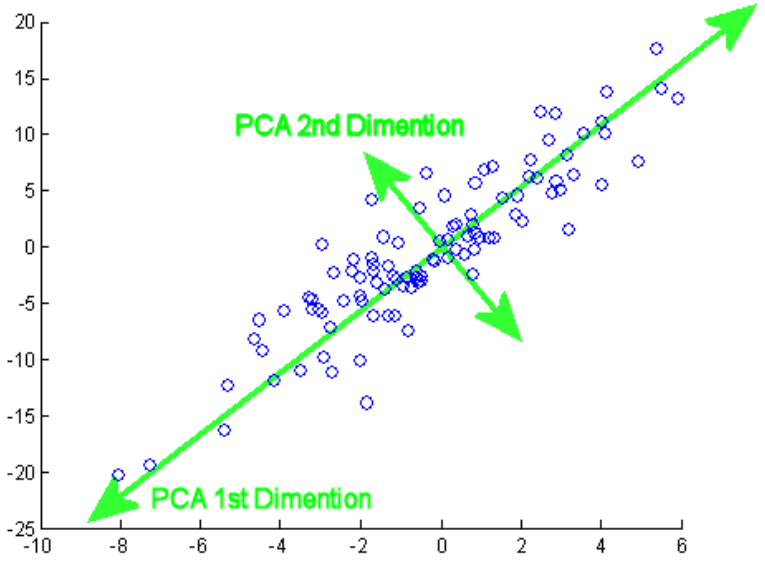
Unsupervised learning

Clustering

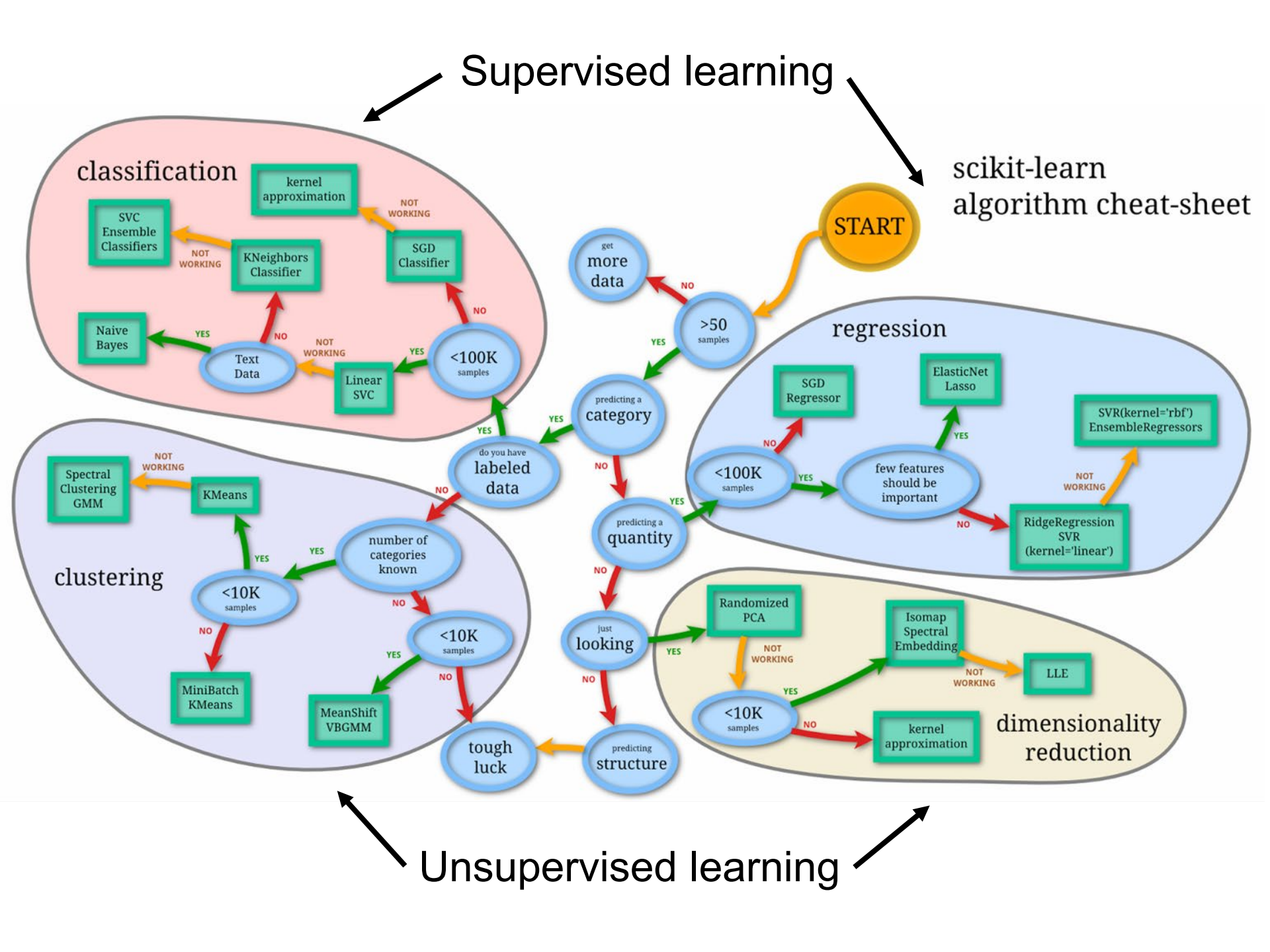


Group items based on similarity
(e.g., group customers based on what each one buys)

Dimensionality Reduction

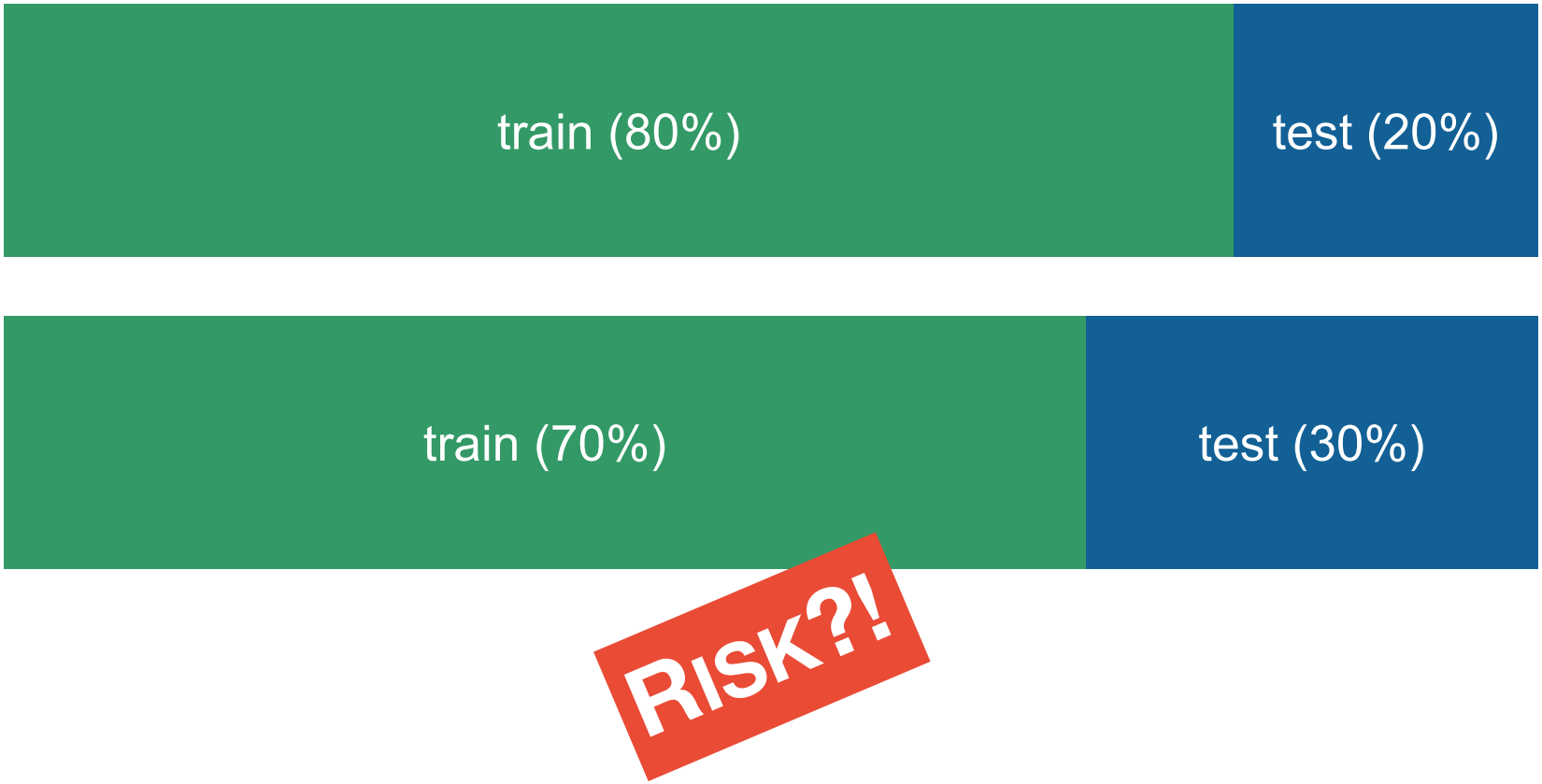


Reduce the features of the items
(e.g., biological features reduced to Body Mass Index)


Supervised learning
scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NO
SVC Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering GMM
NO
<10K samples
YES
MeanShift VBGMM
NO
MiniBatch KMeans
clustering
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Unsupervised learning

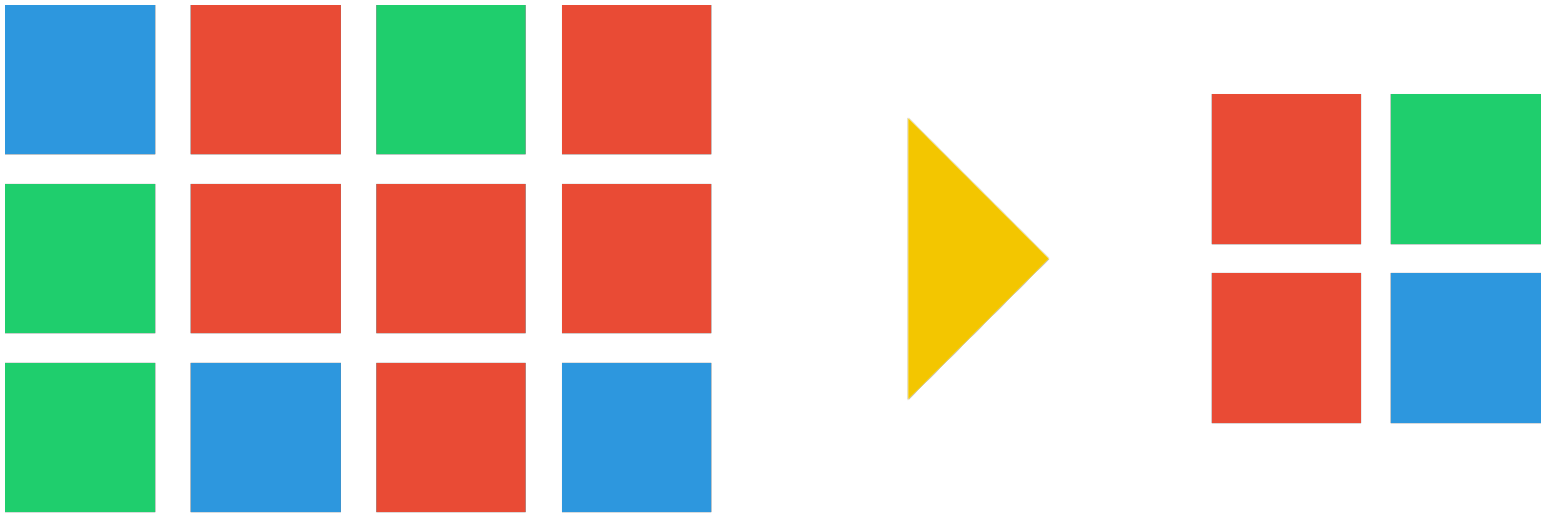
Train test split

You choose what portions of the dataset to use as training and test datasets



Stratification

- Strata: non-overlapping groups
- Stratified sampling:
  Proportional random sample per strata

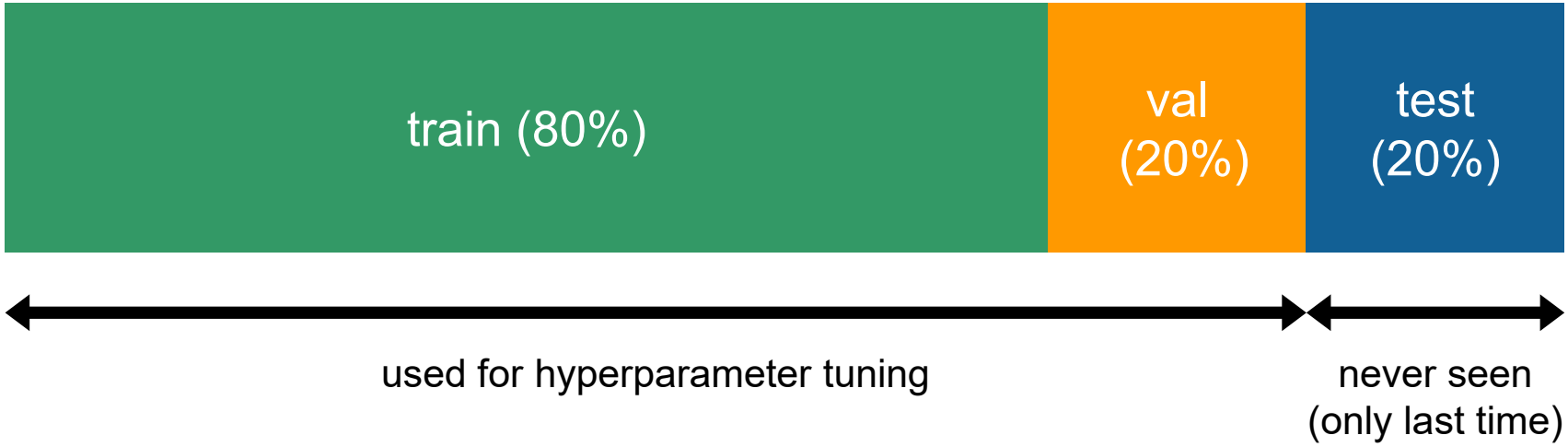


Population

Sample

Train val test split

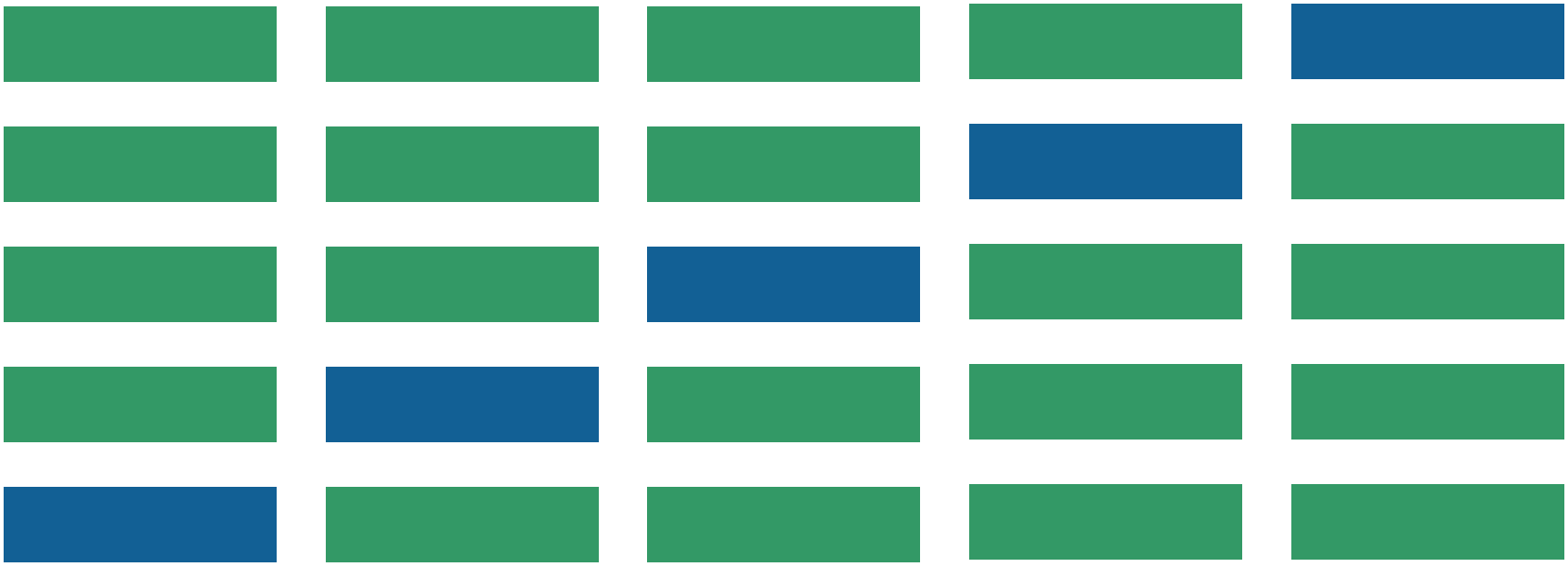
You choose what portions of the dataset to use as training and test datasets



K-Fold Cross validation

- Rather than one test-training split, use many
- Each split is then called fold

K = 5



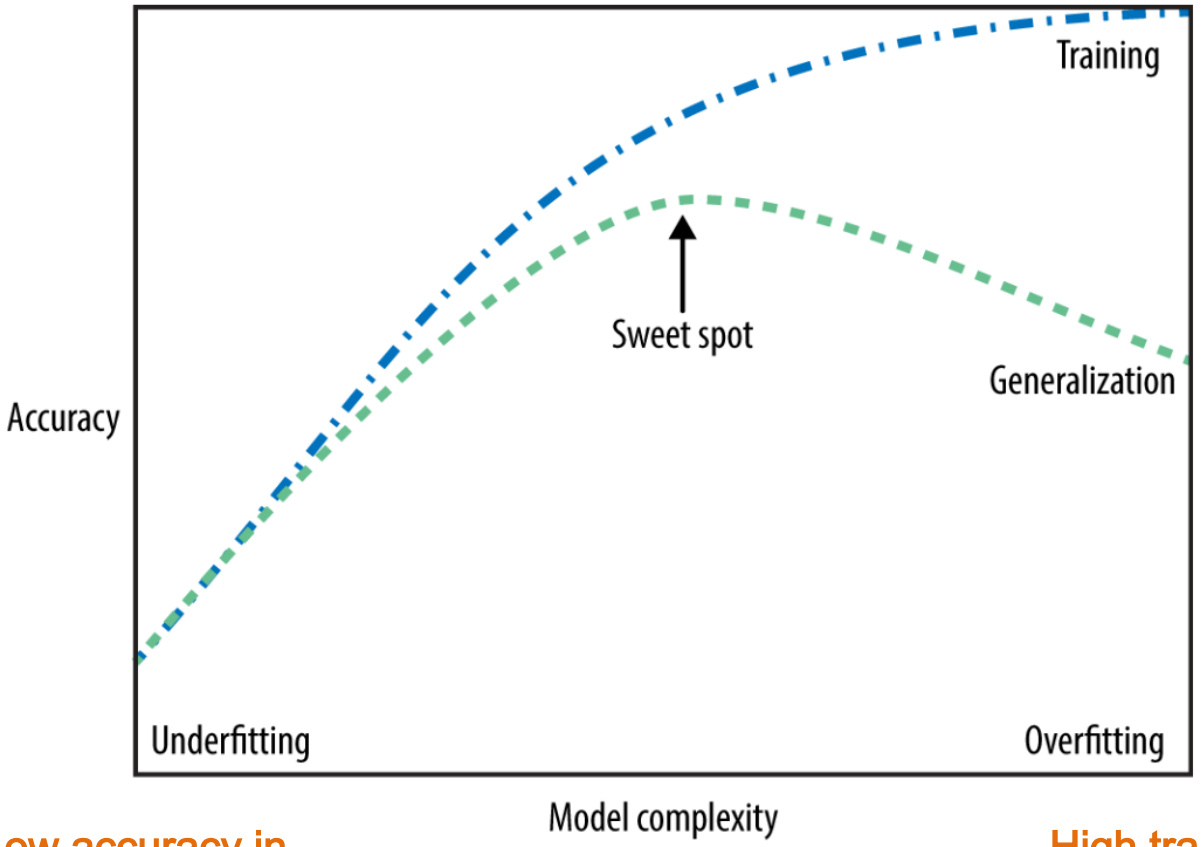
Report descriptive statistics of the K accuracy values

Shuffle -split cross validation

- Before each split, randomly shuffle the dataset
- ShuffleSplit
  test_size=0.5
  train_size=0.5
  n_splits=10

Evaluate your results

Accuracy

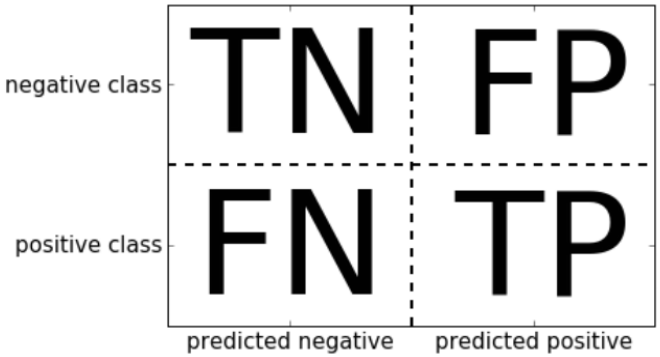


Low accuracy in training and testing

High training accuracy, low test accuracy

Evaluate your results

Confusion matrix

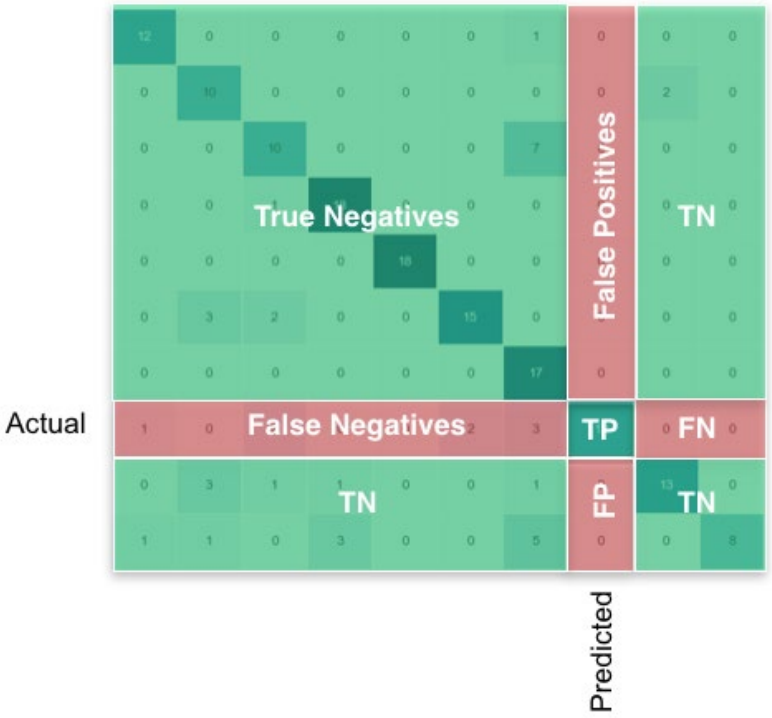


False Positive (Type I error)
The model predicts that a healthy patient has cancer

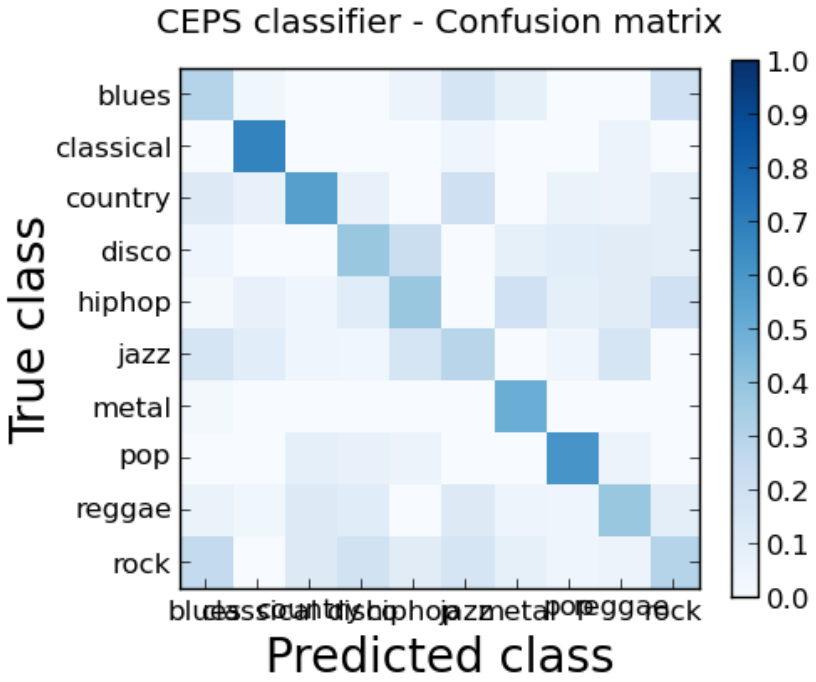
False Negative (Type II error)
The model predicts that an ill patient has no cancer

Evaluate your results

Confusion matrix



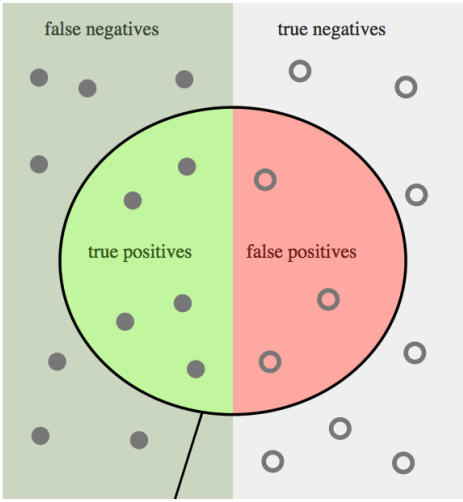
Example: Detection of music genre



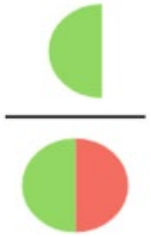
Source: https://github.com/jazdev/genreXpose

Evaluate your results

Precision and Recall



$$\text{Precision} = \frac{TP}{TP+FP}$$



How many of our positive predictions are correct?

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

How good is the model performance, even with imbalanced classes?

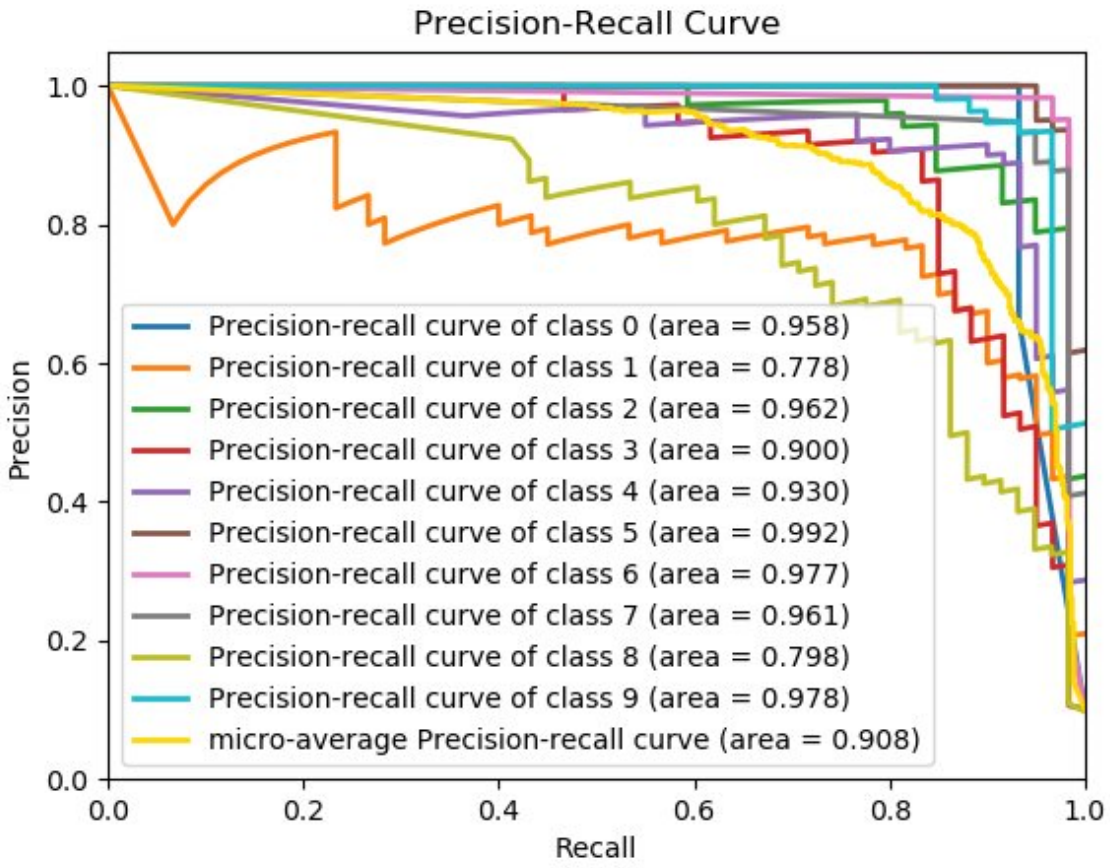
$$\text{Recall} = \frac{TP}{TP+FN}$$

How many correct predictions do we get among those we could have gotten?

Source: https://en.wikipedia.org/wiki/Precision_and_recall

Evaluate your results

Precision and Recall

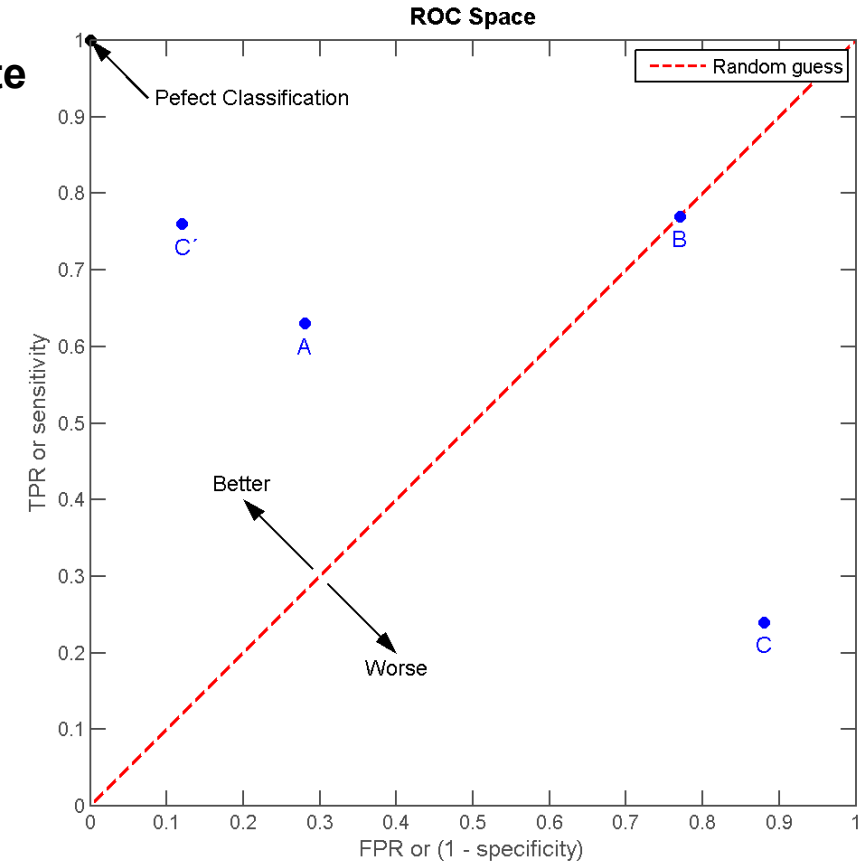


Source: Yellowbrick

Evaluate your results

Receiver Operating Characteristic curve

True positive rate (TPR) (== Recall)

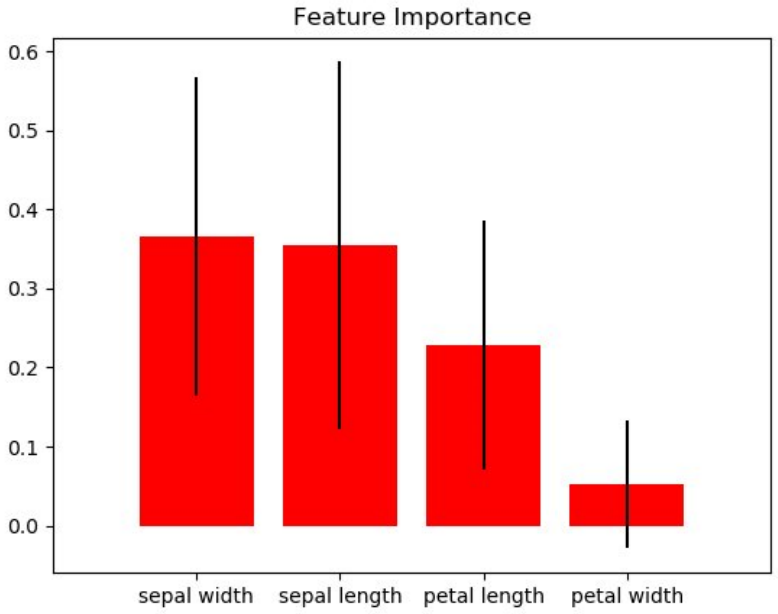


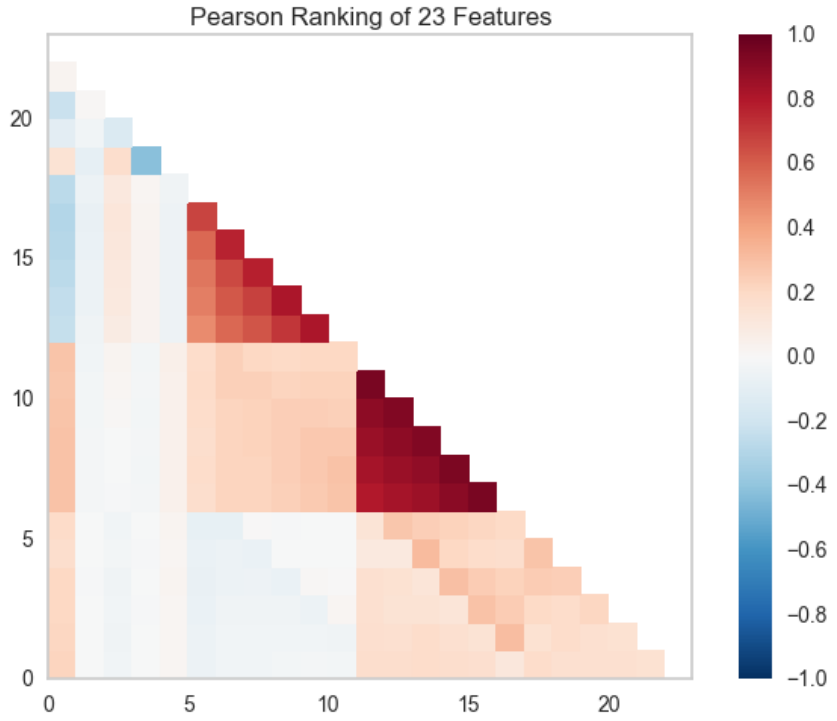
False positive rate (FPR)

Source: Wikipedia

Evaluate your results

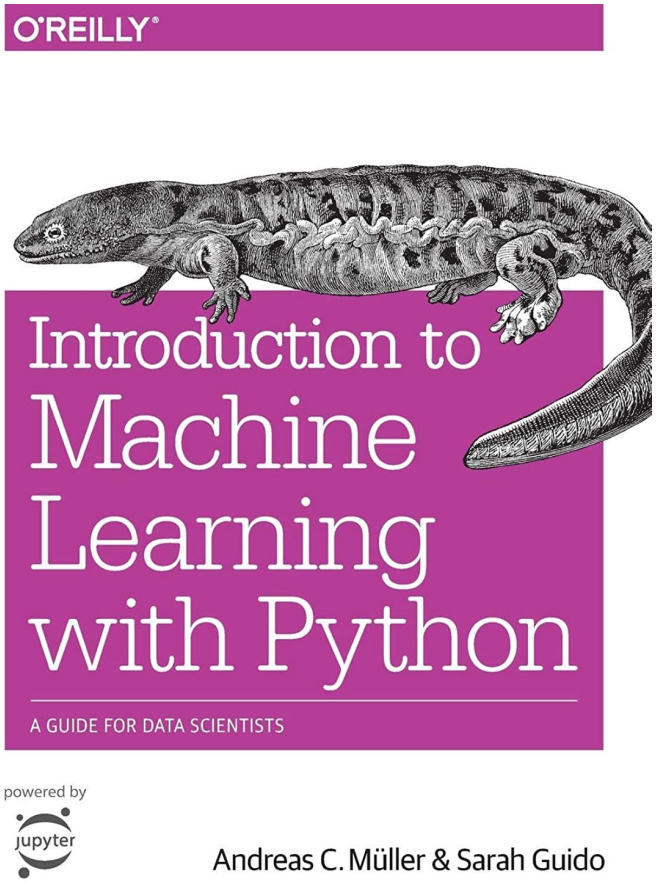
Feature Importance & Rank2D



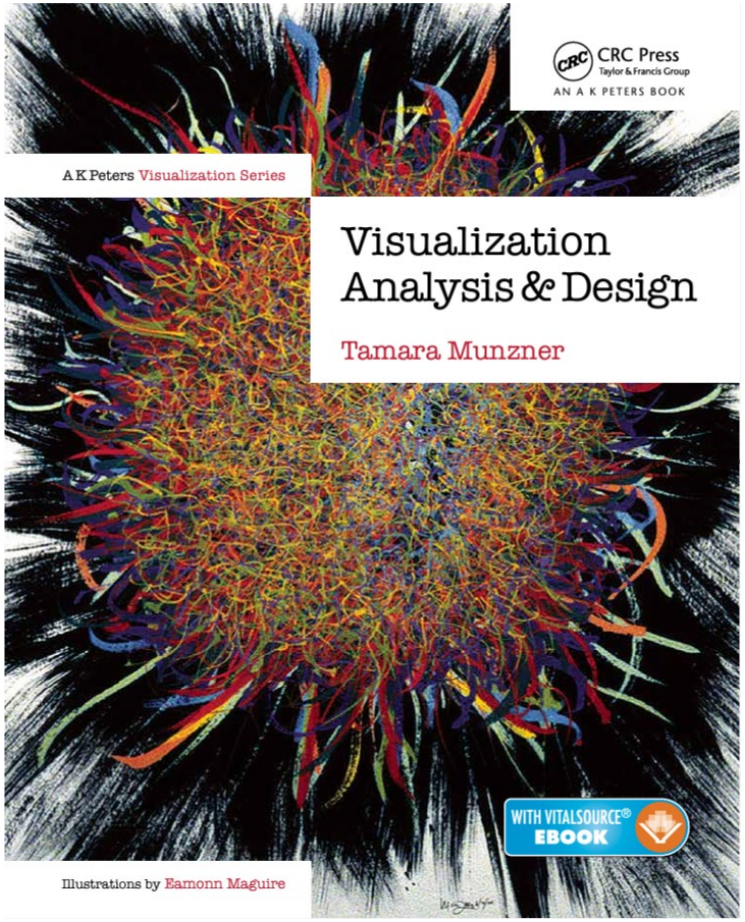


Source: [Yellowbrick]

Main textbooks



"Introduction to Machine Learning with Python" by Müller and Guido (e-book via SUUB)



"Visualization Analysis & Design" by Munzner (e-book via SUUB)

Data Science & Vis Presentation (20% of your grade)

- **Every group member needs to present.**

- If you can’t make it in person, talk to your team and make it possible in hybrid format.

Data Science & Vis Presentation

Goal: Get a shared understanding of data science & vis

- Why is it relevant to data science & vis?
- Why is it relevant to you?
- You have 5 minutes (4 minutes if you are group of 2)

Group:
Topic:

University of Bremen

	Excellent	Very good	Good	Satisfactory	Weak	Very Weak	Unacceptable	N/A
Clear Slides								
Clear Delivery								
Timing								
Clarity of topic / focus								
Command of content								
Critical analysis								
Conclusion								
Questions and Answers								
Overall Performance								

Comments:

Grading Criteria

- Data Science & Vis paper presentation (20%)
- Exposé & exposé presentation (required)
- Interim presentation (required)
- Final presentation (30%)
- Final report (50%)

Credits (ECTS)

6 ECTS == 180 hours

- Lectures: 24h
- Tutorials: 24h
- Learning Python: 6h
- Data Science & Vis Paper Presentation: 12h
- Exposé: 18h
- Progress Presentation: 12h
- Final Presentation: 12h
- Final Report: 24h
- Final Project: 48h

Group Project

- As groups of three people, you will work on a project throughout the course
- Pick a dataset
- Pick a research question
- Pick a suitable method
- Find the best* analysis and visualization techniques for your dataset, question, and method
- Write a report on your findings and motivate your choices

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
08.04.24	Introduction to Data Science	Python Introduction
15.04.24	Basic Statistics & Supervised Learning	Practical Statistics + Supervised
22.04.24	Unsupervised Learning	Practical Unsupervised
29.04.24	Introduction to Data Visualization	Guest Lecture on NLP by Oxana Vitman
06.05.24	Exploratory Data Analysis	Data Science & Vis Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
13.05.24	Visual Encoding	Exposé Workshop
20.05.24	Pentecost	Pentecost
27.05.24	Design Studies	Exposé Presentation
03.06.24	Interaction Techniques	Practical Interaction
10.06.24	Data Science & Journalism	Progress Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
17.06.24	Machine Learning & Visualization	Project Workshop
24.06.24	Guest Lecture on Critical Data Studies by Paola Lopez	Observable (Plot)
01.07.24	Final Session: Recap	Final Project Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom .

Exposé Presentation

Date:
27.05.2024

- Title — clear title that describes your project
- Abstract — short summary of approach and main findings
- Introduction — a text that introduces somebody new to the topic to what you did and why it is worth doing
- Background — correctly cite the tools and approaches you used
- Method — what was your method?
  - Data — what analysis/vis/ML techniques did you apply?
  - Collection — how did you get the data?
- Results — what are your answers to your research questions?
  - Discussion — what are highlights and lowlights?
what other interesting things did you observe?
what didn't work and why?
- Conclusion — quick summary of your work in 1-2 paragraphs

Project Report

Deadline:
19.07.2024

- Title — clear title that describes your project
- Abstract — short summary of approach and main findings
- Introduction — a text that introduces somebody new to the topic to what you did and why it is worth doing
- Background — correctly cite the tools and approaches you used
- Method — what was your method?
  - Data — what analysis/vis/ML techniques did you apply?
  - Collection — how did you get the data?
- Results — what are your answers to your research questions?
  - Discussion — what are highlights and lowlights?
    what other interesting things did you observe?
    what didn't work and why?
- Conclusion — quick summary of your work in 1-2 paragraphs