Prototipazione Rapida con Open Data e Machine Learning

Dr.Ir. Jacopo De Stefani

Lecturer @ TPM-ESS-ICT

Cremona, 16/12/2022

TU Delft

About Me

Academic Background

- BSc in Computer Engineering, Politecnico di Milano, Italy (2011)
- MSc in Computer Science and Engineering, ULB, Belgium (2013)
- MSc in Computer Engineering, Politecnico di Milano, Italy (2015)
- PhD in Machine Learning and Time Series Analysis, ULB, Belgium (2022)

Scientific activity

- 4 international peer-reviewed journal publications
- 6 international peer-reviewed conference proceedings
- **1 international patent**
- Reviewer for International Journal of Forecasting, IEEE Access, Technology and Economics of Smart Grid and Sustainable Energy

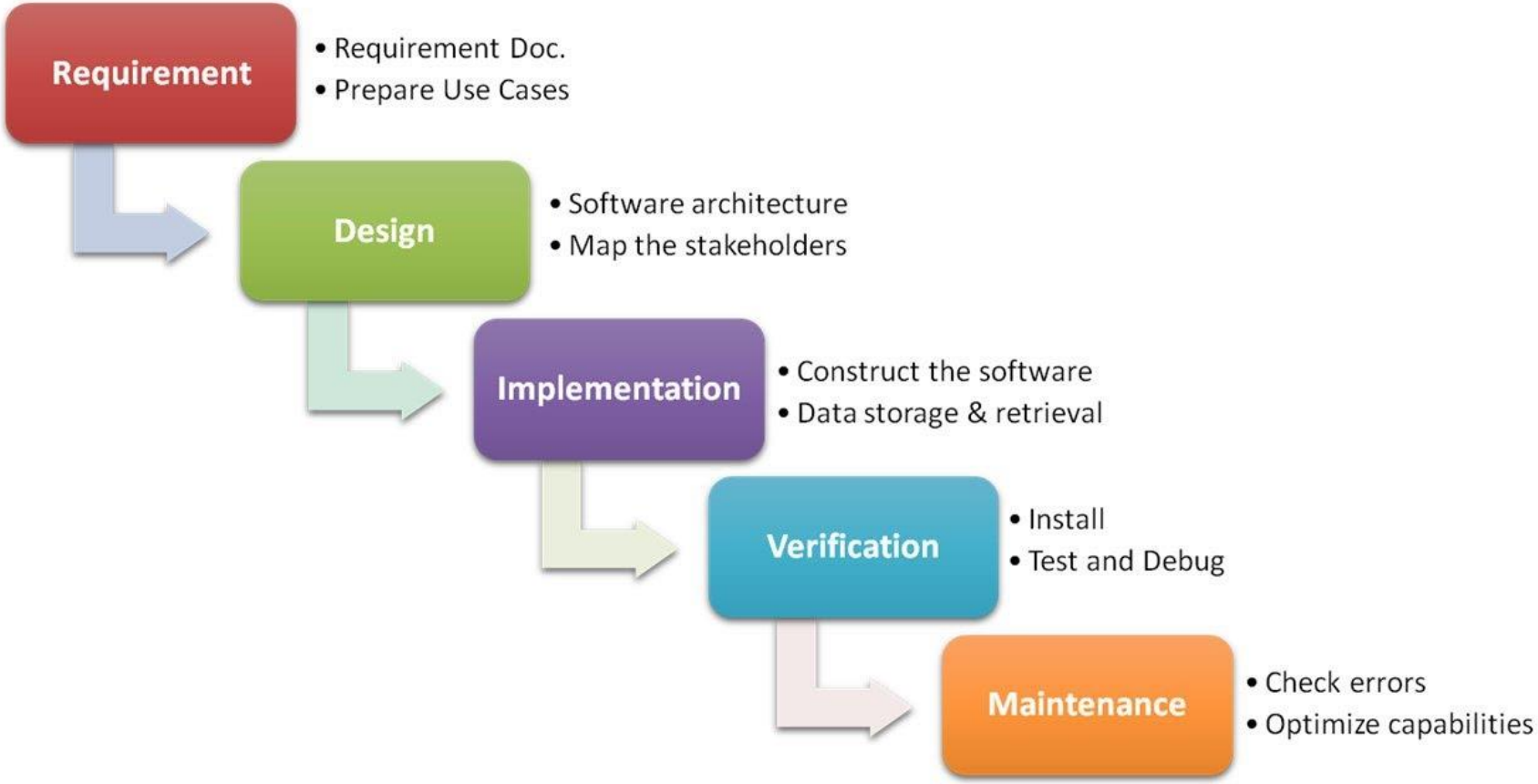
TU Delft

Outline of the workshop

- **Introduction to Open Data**
  - 10-20 min theoretical introduction
  - 30-40 min hands on exercises exploring/gathering open data
- **Introduction to Data Analytics**
  - 10-20 min theoretical introduction
  - 30-40 min hands on exercises on data preprocessing and visualization
- **Break – 15 minutes**
- **Introduction to ML**
  - 10-20 min theoretical introduction
  - 30-40 min hands on exercises implementing ML models using Sklearn/Keras
- **Wrap-up and publication of the final work**
  - 30-40 min Wrapping up and finalizing

TU Delft

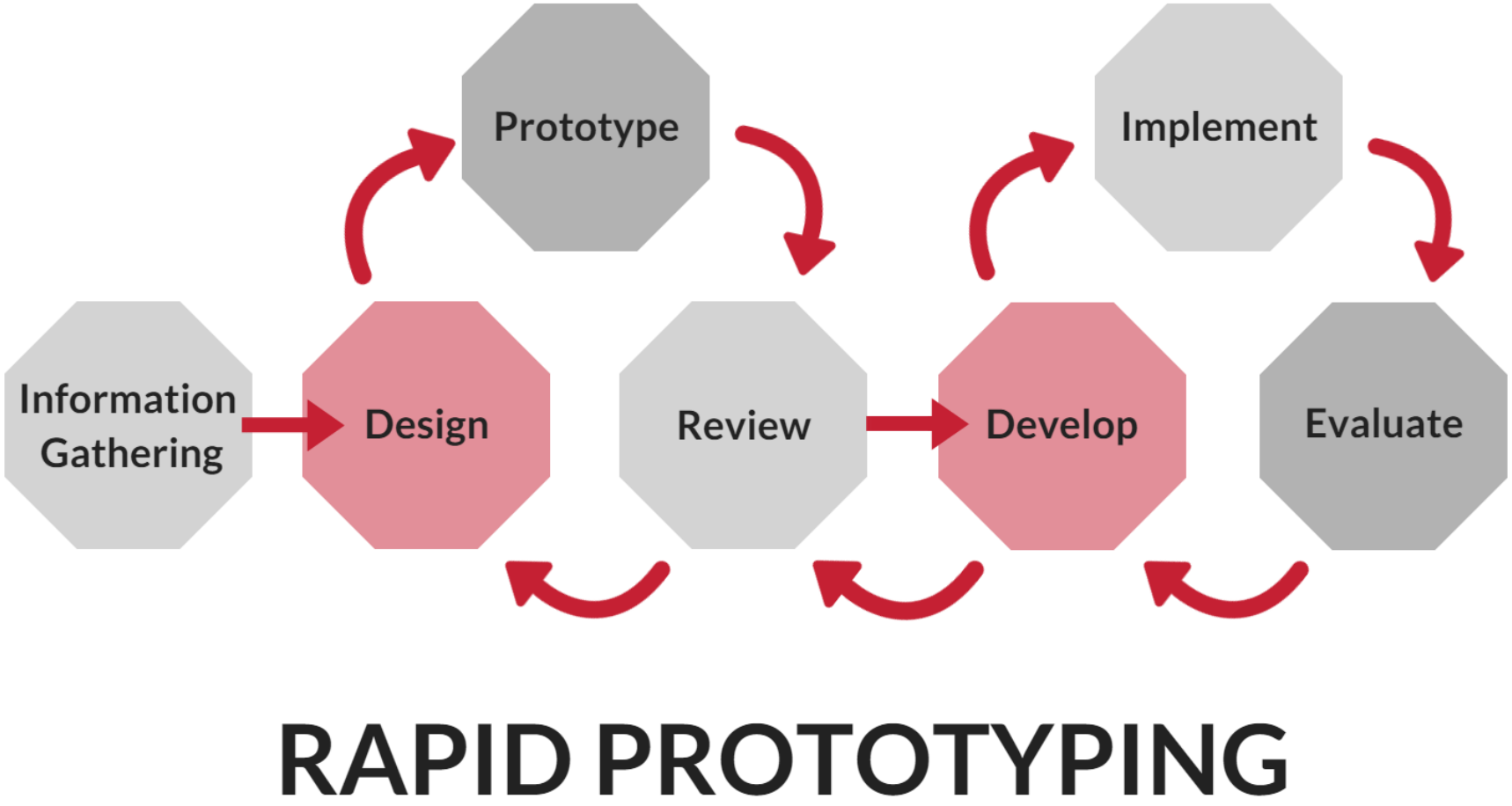
Traditional Software Development



Source: https://4.bp.blogspot.com/-FJ4rlfMcDfc/V8mDTJfcDQI/AAAAAAAAA4s/f1dd7JC9-QsBh-67gdIDpq5ThubrTcVywCK4B/s1600/maxresdefault.jpg

TU Delft

Rapid Prototyping



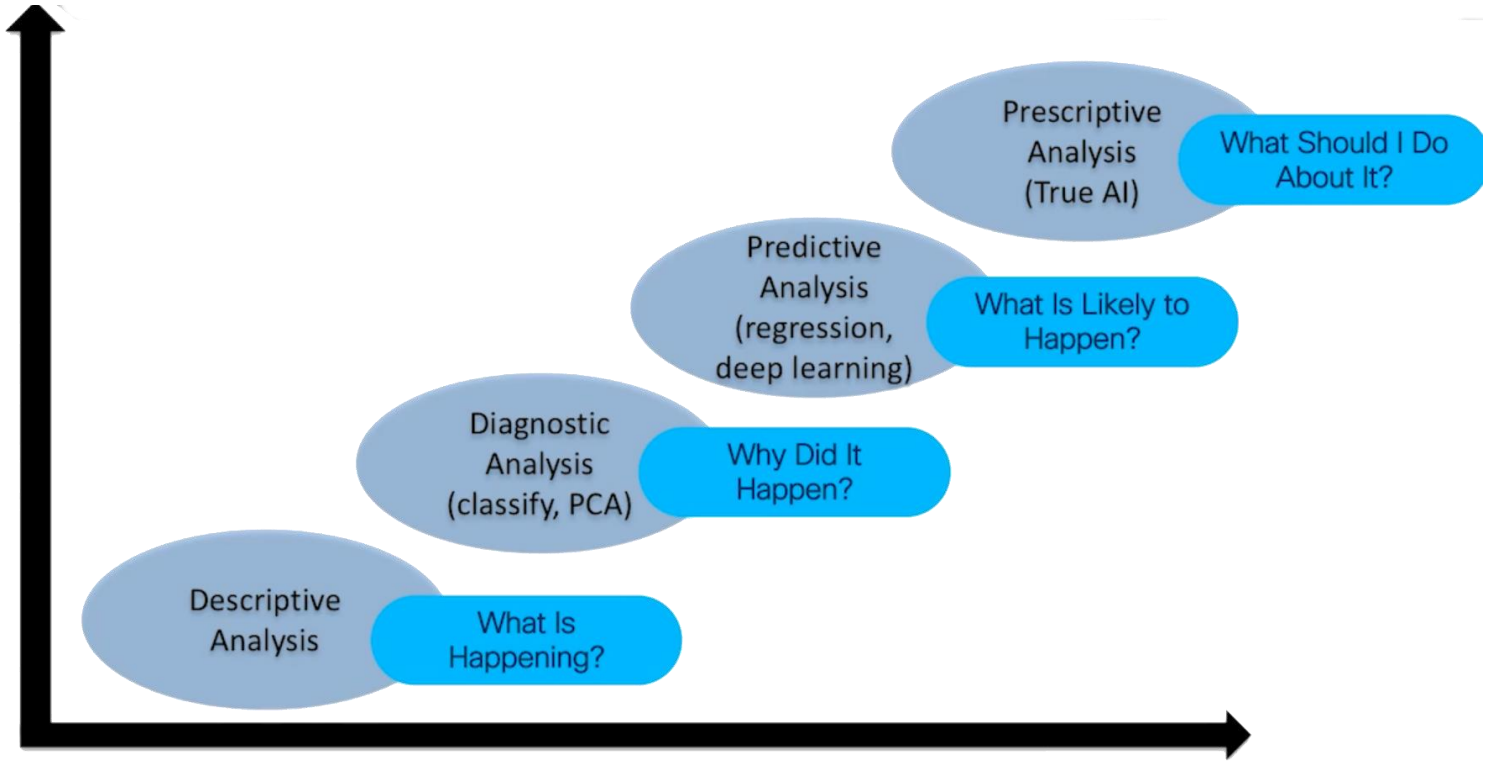
Source: https://www.dashe.com/hs-fs/hubfs/Rapid_Prototyping-1.png?width=1800&name=Rapid_Prototyping-1.png

TU Delft

How should be AI applied in practice?

From the perspective of the data:

1. Descriptive Analysis
2. Diagnostic Analysis
3. Predictive Analysis
4. Prescriptive Analysis



Source picture : Screenshot from Data Analytics and Machine Learning Fundamentals LiveLessons Video Training by Jerome Henry

TU Delft

01

Introduction to Open Data

“Il Machine Learning e’ nulla senza dati…”

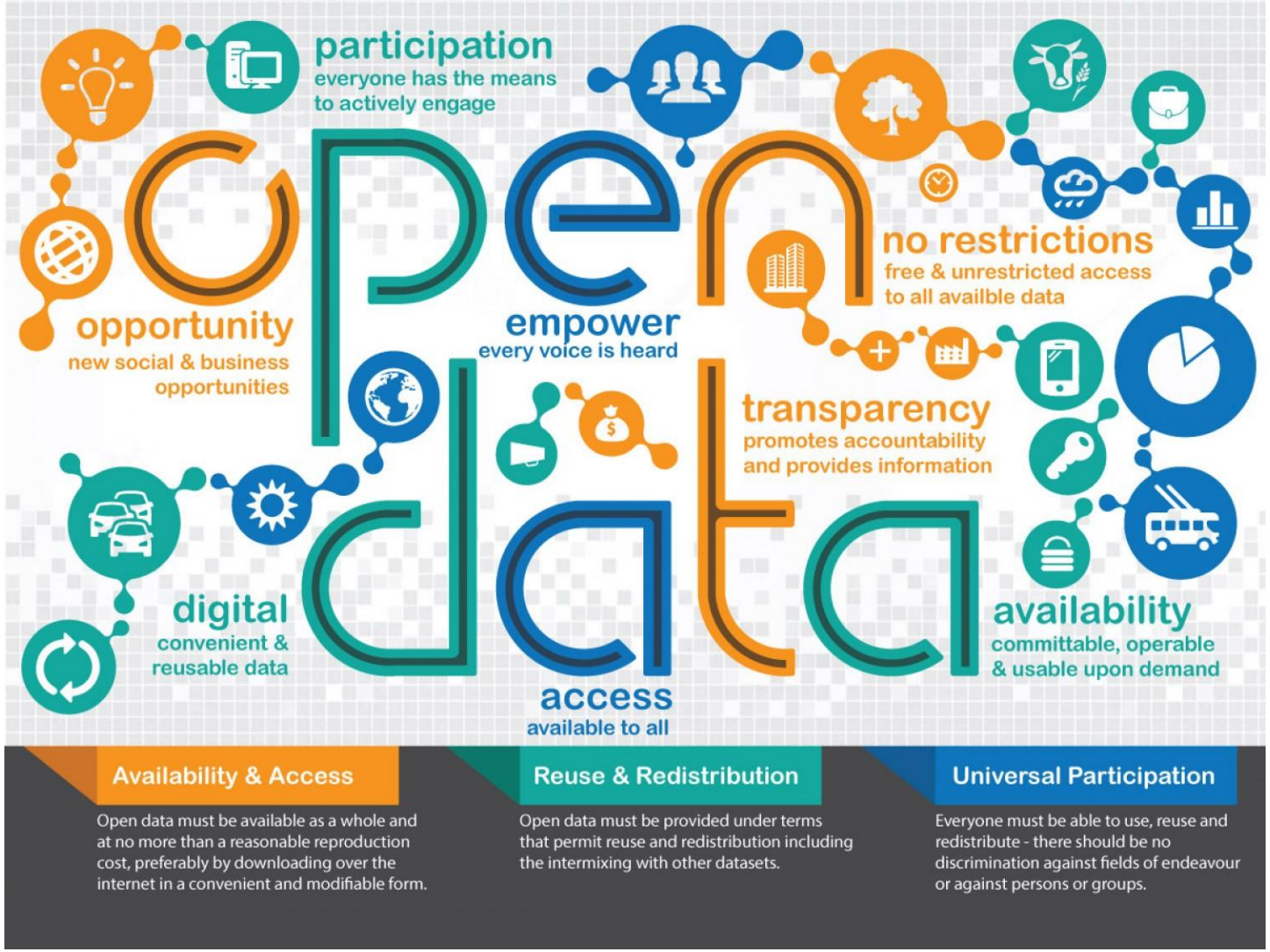
TU Delft

How would you define Open Data? Have you ever heard of it?

TU Delft

What is Open Data?

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.
- If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: **interoperability.**
- **Reference:** : https://opendatahandbook.org/guide/en/what-is-open-data/



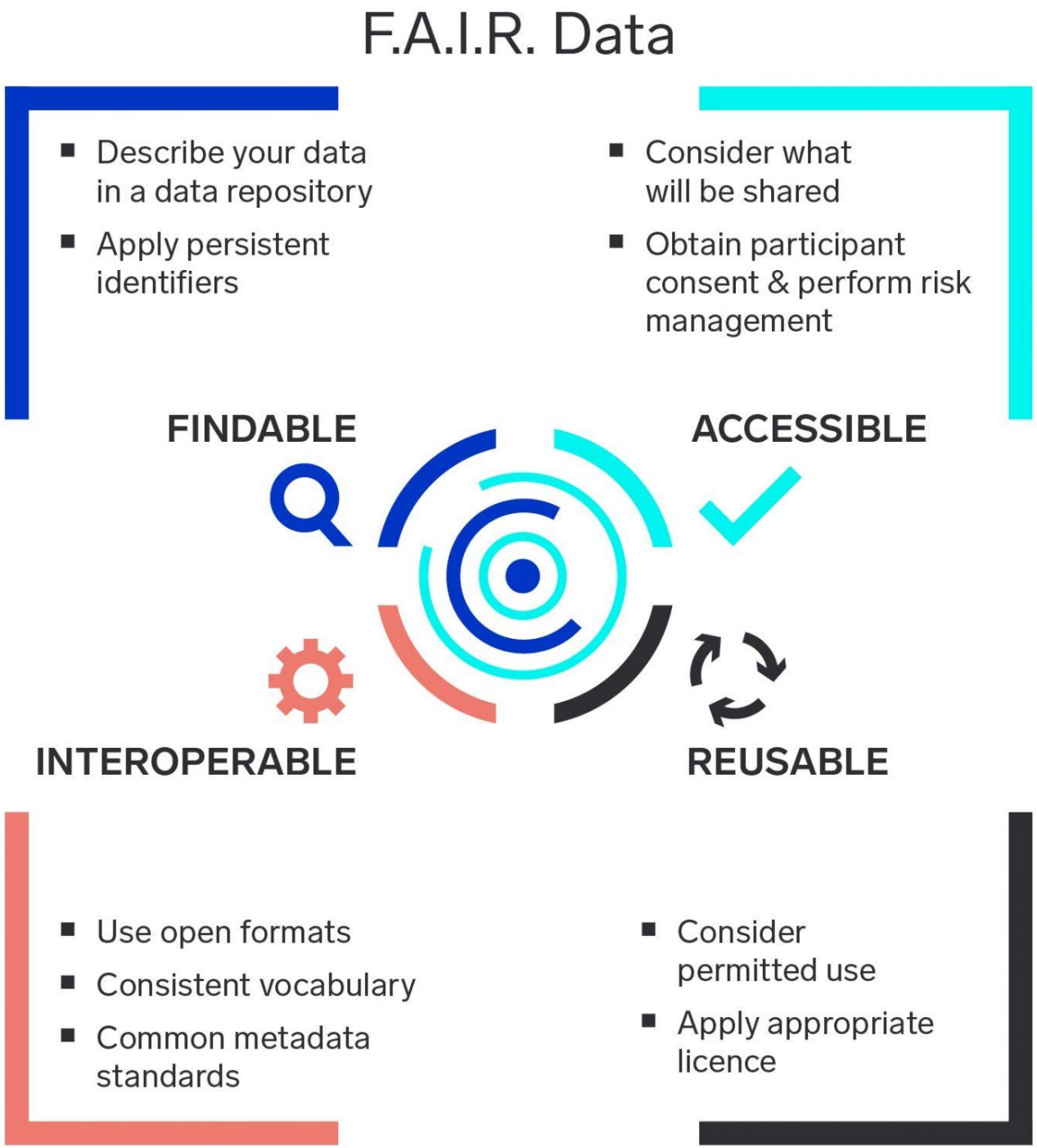
Source: https://id4communication.fr/wp-content/uploads/2018/10/what-is-open-data_54b39b466d9c9_w1500.jpg

TU Delft

FAIR

Reference: https://www-nature-com.tudelft.idm.oclc.org/articles/sdata201618

Source: https://www.go-fair.org/fair-principles/



TU Delft

Global platforms to find open data

- **Proprietary sources**
  - Kaggle
  - Google Datasets
  - …
- **Governmental sources**
  - **EU:** https://data.europa.eu/en
  - **US:** https://www.data.gov/
  - …
- **Academic sources**
  - **UCI Machine Learning Repository**
  - **…**

TU Delft

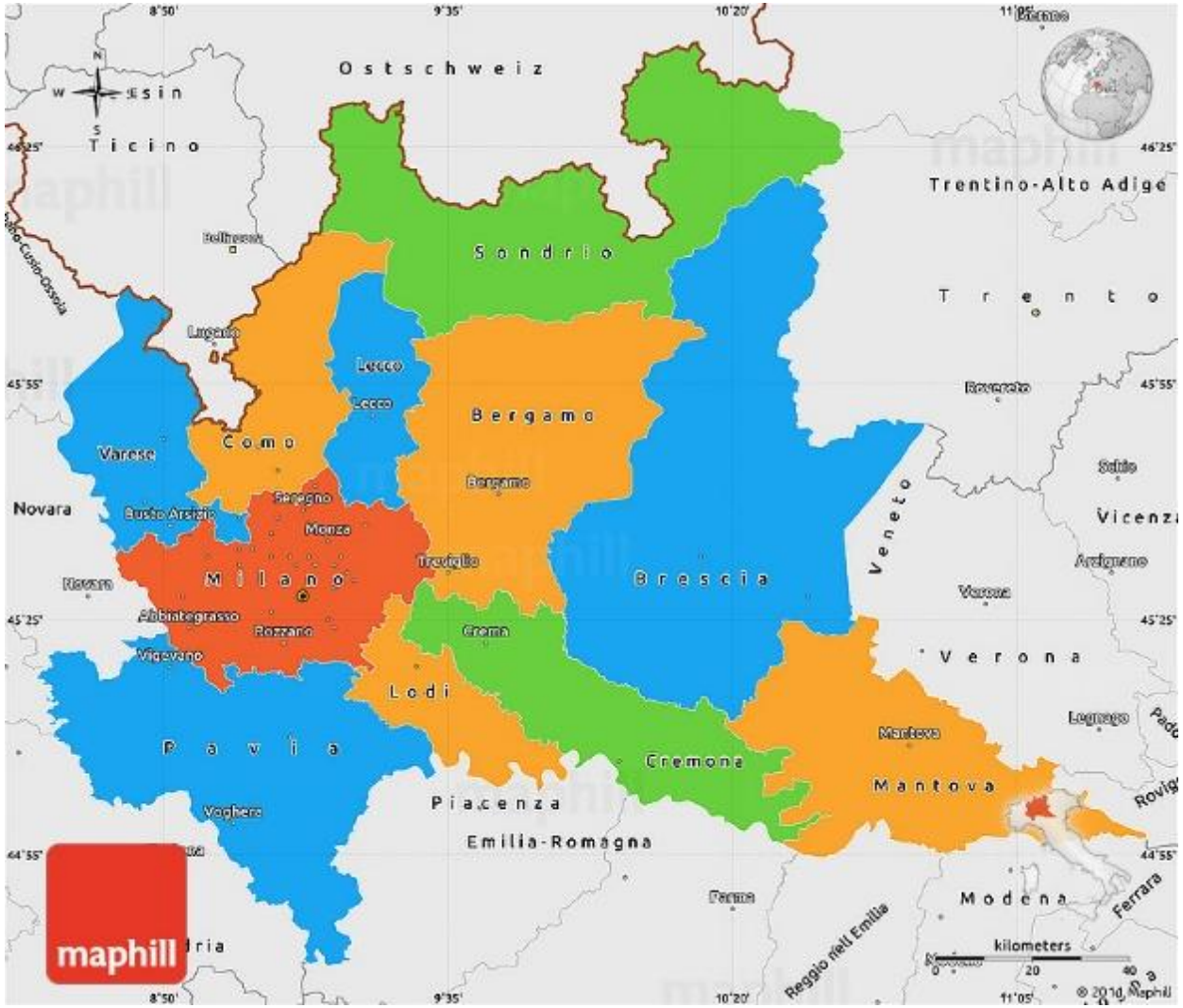
National platforms to find open data

- **Governmental sources**
- https://github.com/italia/awesome-italian-public-datasets
- https://www.kaggle.com/general/27278

TU Delft

Regional platforms to find open data

- **Governmental sources**
- https://www.dati.lombardia.it/



TU Delft

Examples of most common data formats

- CSV
- XML
- XLSX
- JSON

TU Delft

02

Introduction to Data Analytics

Quality data requires quality time

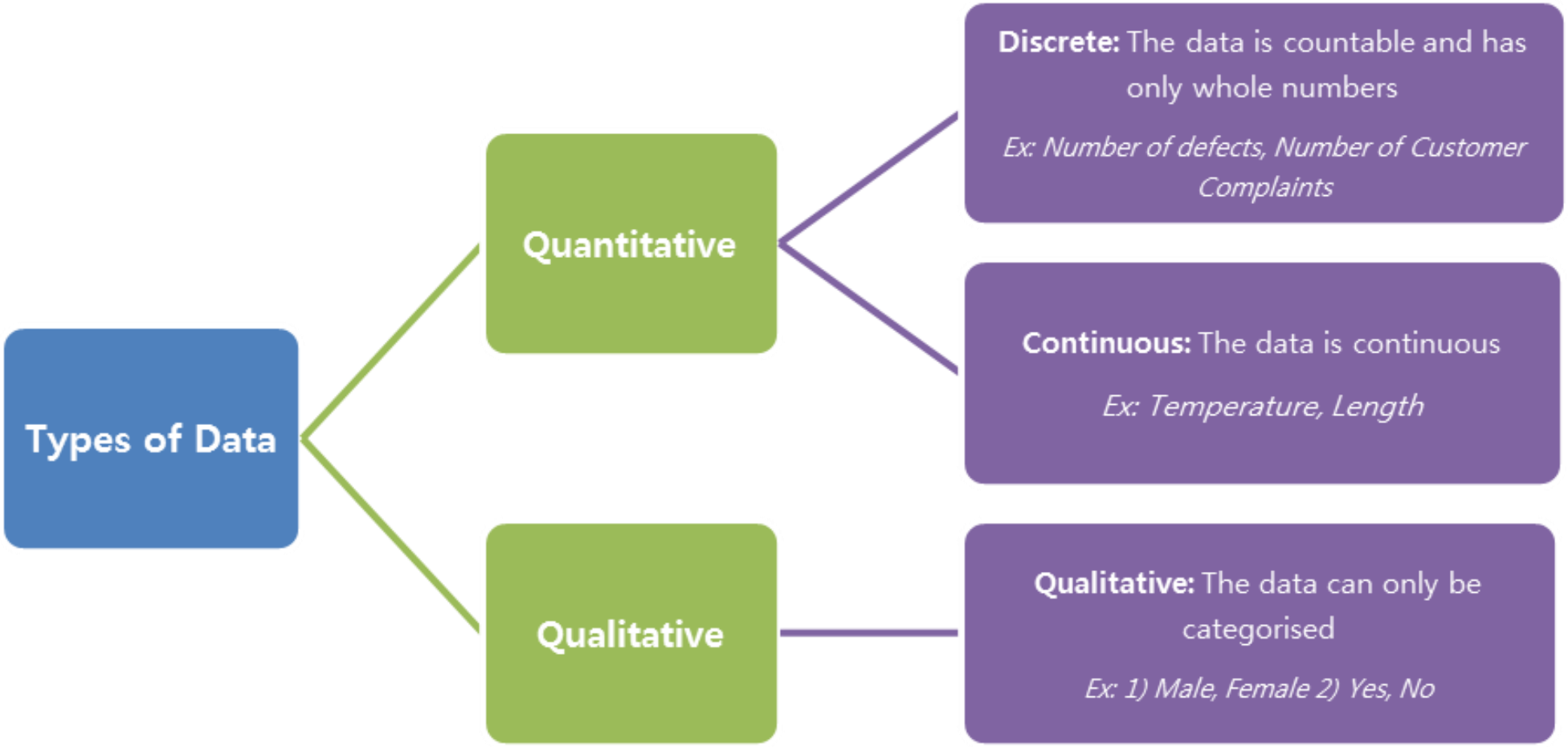
TU Delft

Outline

Describe data → Plotting → Missing values handling

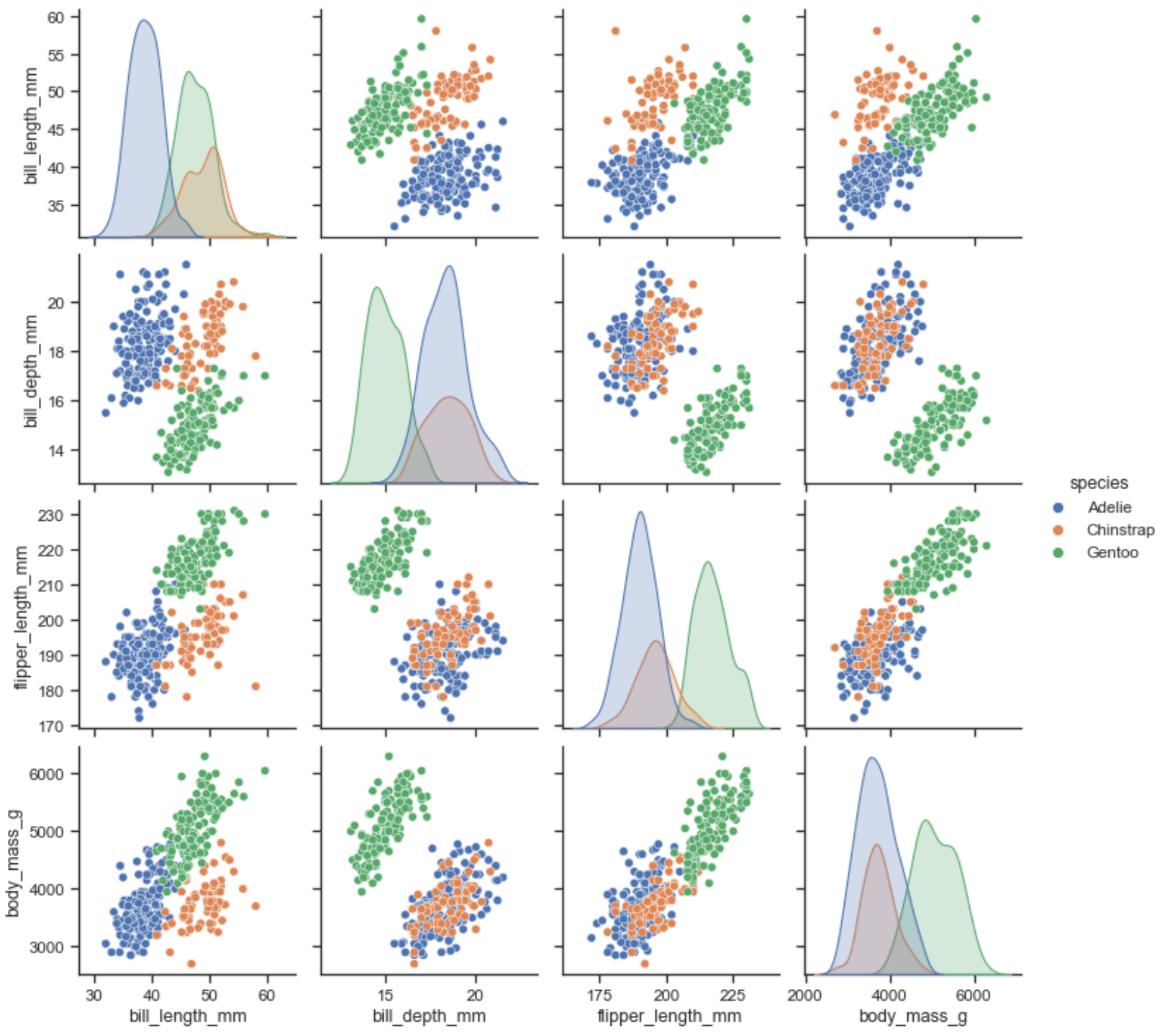
TU Delft

Data description



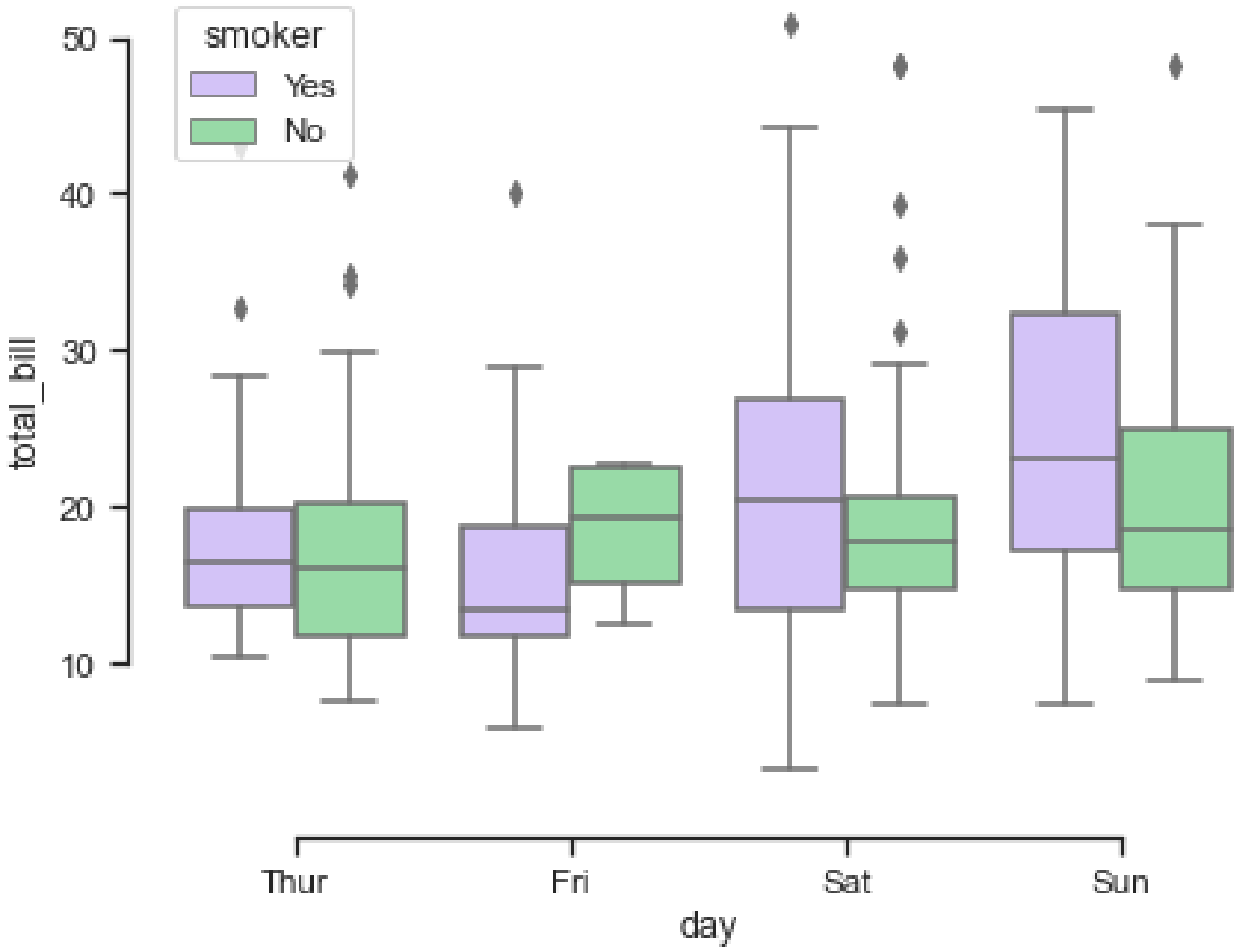
TU Delft

Plotting



TU Delft

Plotting



TU Delft

Missing data handling

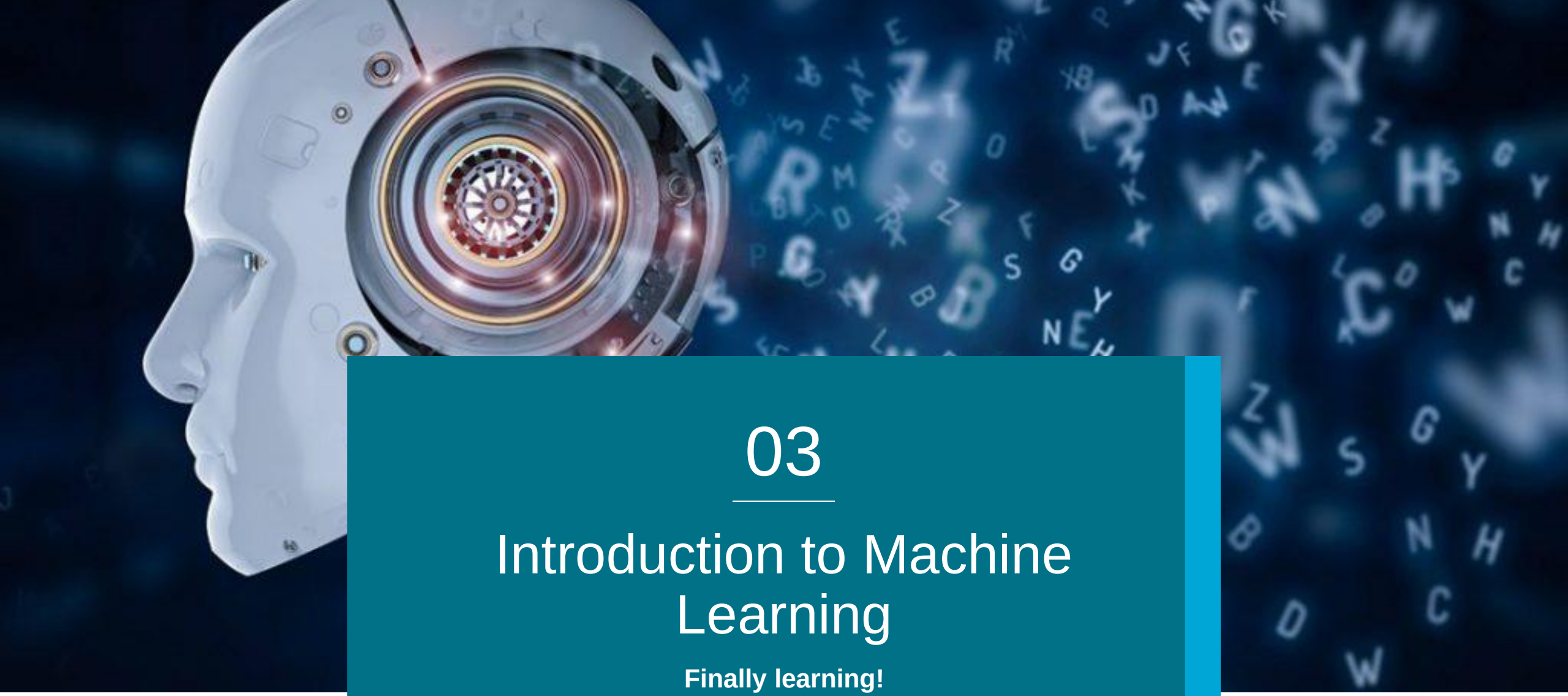
Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

Bed	Bath	Bed_was_missing
1.0	1.0	FALSE
2.0	1.0	FALSE
3.0	2.0	FALSE
2.0	2.0	TRUE

Source: https://medium.com/analytics-vidhya/how-to-handle-missing-values-byaryan-cb76b9dbaae2

TU Delft

TU Delft



03

Introduction to Machine Learning

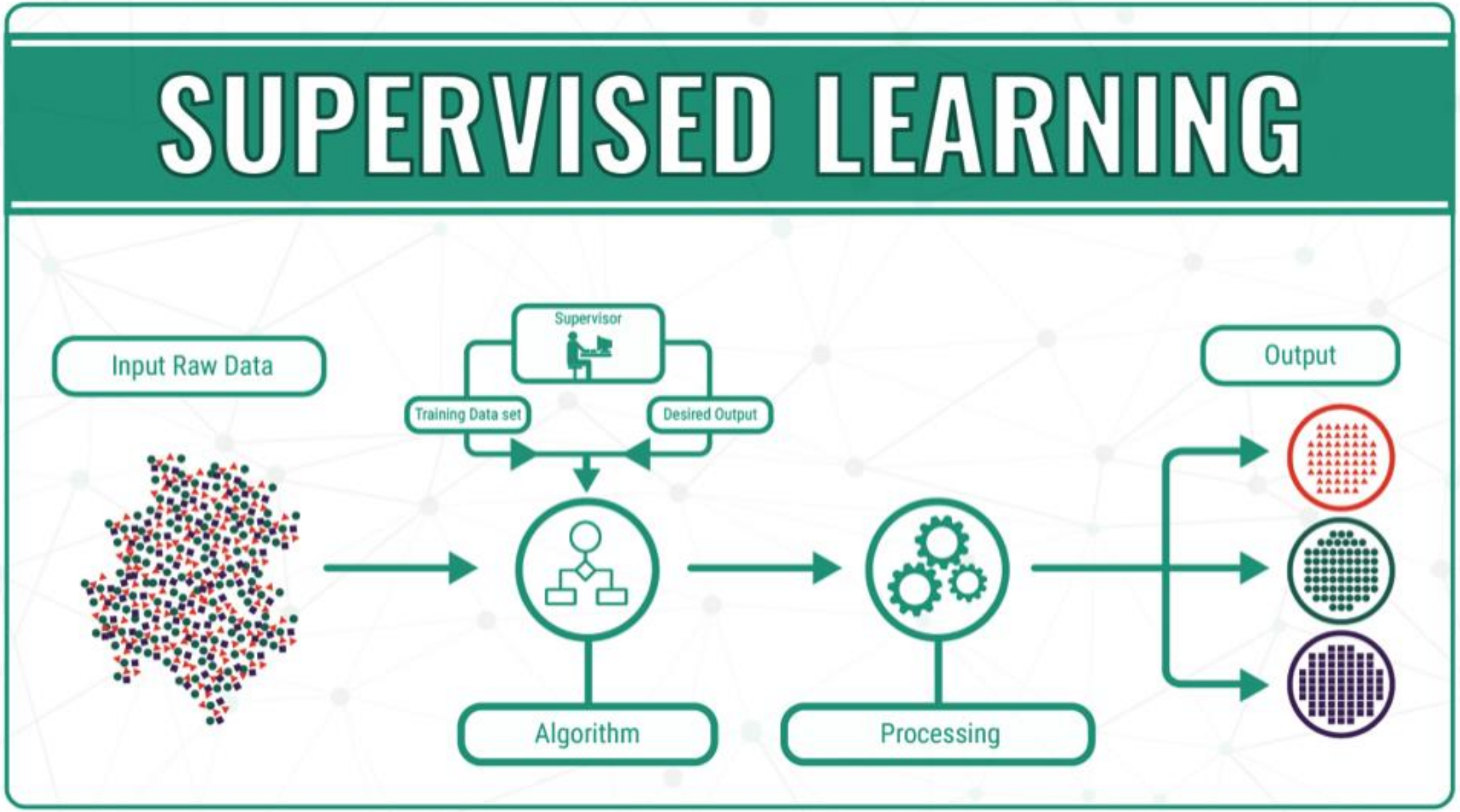
Finally learning!

TU Delft

Outline

Train/Test split → Categorical data → Fit benchmark model → Fit other models → Parameter optimization

TU Delft



Source: https://www.crayondata.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforcement-learning/

$$
\begin{array}{c|ccccc|c}
 & \multicolumn{5}{c}{\text{Features/Variables}} & \text{Target} \\
 & \mathbf{X}_1 & \cdots & \mathbf{X}_i & \cdots & \mathbf{X}_{n'} & \mathbf{y} \\
\hline
\text{Sample} & x_{11} & \cdots & x_{1i} & \cdots & x_{1n'} & y_1 \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & x_{k1} & \cdots & x_{ki} & \cdots & x_{kn'} & y_k \\
\hline
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & x_{N1} & \cdots & x_{Ni} & \cdots & x_{Nn'} & y_N
\end{array}
$$

Training set $\mathcal{D}_{train}$

Testing set $\mathcal{D}_{test}$

Figure 2.2: Example of generic dataset D_N of N samples, $n' = n \cdot c_{FE}$ variables, resulting of feature engineering of the original n variables, split into a training set $\mathcal{D}_{train}$ of k samples and a testing set $\mathcal{D}_{test}$ of $N - k$ samples.

TU Delft

Source: My Own PhD Thesis ☺

Models

Overview Features Models

Search Results - All predictions done on Evaluation Split

	fit_time	accuracy_score	balanced_accuracy_score	f1_score	roc_auc_score
PassiveAggressiveClassifier	0.00199	0.82960	0.81670	0.77907	0.90550
RidgeClassifier	0.00598	0.83408	0.82043	0.78363	0.90357
SGDClassifier	0.00200	0.82511	0.81109	0.77193	0.89351
DummyClassifier	NaN	0.51121	0.48575	0.36994	0.48575

Confusion Matrices

PassiveAggressiveClassifier

	Predicted Negative	Predicted Positive
Actual Negative	118	16
Actual Positive	22	67

RidgeClassifier

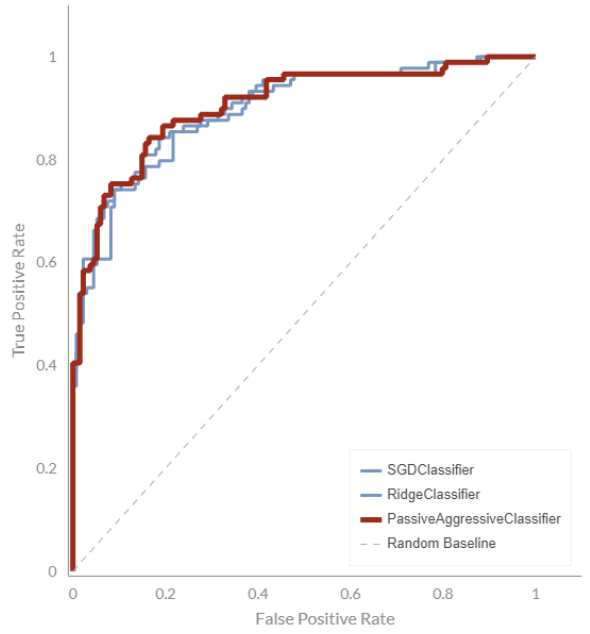
	Predicted Negative	Predicted Positive
Actual Negative	119	15
Actual Positive	22	67

SGDClassifier

	Predicted Negative	Predicted Positive
Actual Negative	118	16
Actual Positive	23	66

Result Curves Comparison

ROC Curve | Precision Recall Plot | Detection Error Tradeoff



Predictions Table

#	Survived	PassiveAggressiveCla	RidgeClassifier	SGDClassifier	Embarked	Parch	Cabin	Sex	SibSp	Ticket	Fare	PassengerId	Pclass	Name	Age
0	1	0	0	0	C	1	NaN	male	1	2661	15.2458	710	3	Moubarek, Master. Ha	NaN
1	0	0	0	0	S	0	NaN	male	0	C.A. 18723	10.5	440	2	Kvillner, Mr. Johan Her	31
2	0	0	0	0	S	0	NaN	male	0	SOTON/O2 3101287	7.925	841	3	Alhomaki, Mr. Ilmari Ri	20
3	1	1	1	1	S	1	NaN	female	0	248727	33	721	2	Harper, Miss. Annie Je	6

TU Delft

Source: https://discourse.bokeh.org/t/data-dashboard-library-for-explaining-the-data-and-creating-baseline-ml-models-bokeh-driven/7729

Repository: https://github.com/maciek3000/data_dashboard

Performance Evaluation – Confusion Matrix

Actual \ Predicted	Positive	Negative
Positive	True Positive (TP) *Correct objects*	False Negative (FN) *Missed objects*
Negative	False Positive (FP) *Extra objects*	True Negative (TN) *No objects*

Recall

$$\frac{TP}{(TP+FN)}$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Specificity

$$\frac{TN}{(TN+FP)}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Precision

$$\frac{TP}{(TP+FP)}$$

Negative predictive value

$$\frac{TN}{(TN+TP)}$$

TU Delft

04

Wrap-up and publishing!

Finally learning!

TU Delft

Outline

Select publication "venue" → Clean deliverable elements → Publication

TU Delft

Publishing data on GitHub

1. Structuring the content locally to separate code and data
2. Create an empty public repository on GitHub with a README
3. Cloning the empty repository on your PC
4. Adding your content
5. Committing your content
6. Pushing your content on GitHub

TU Delft

Thank you for your attention!
Any questions?

Dr. Ir. Jacopo De Stefani – J.deStefani@tudelft.nl

TU Delft