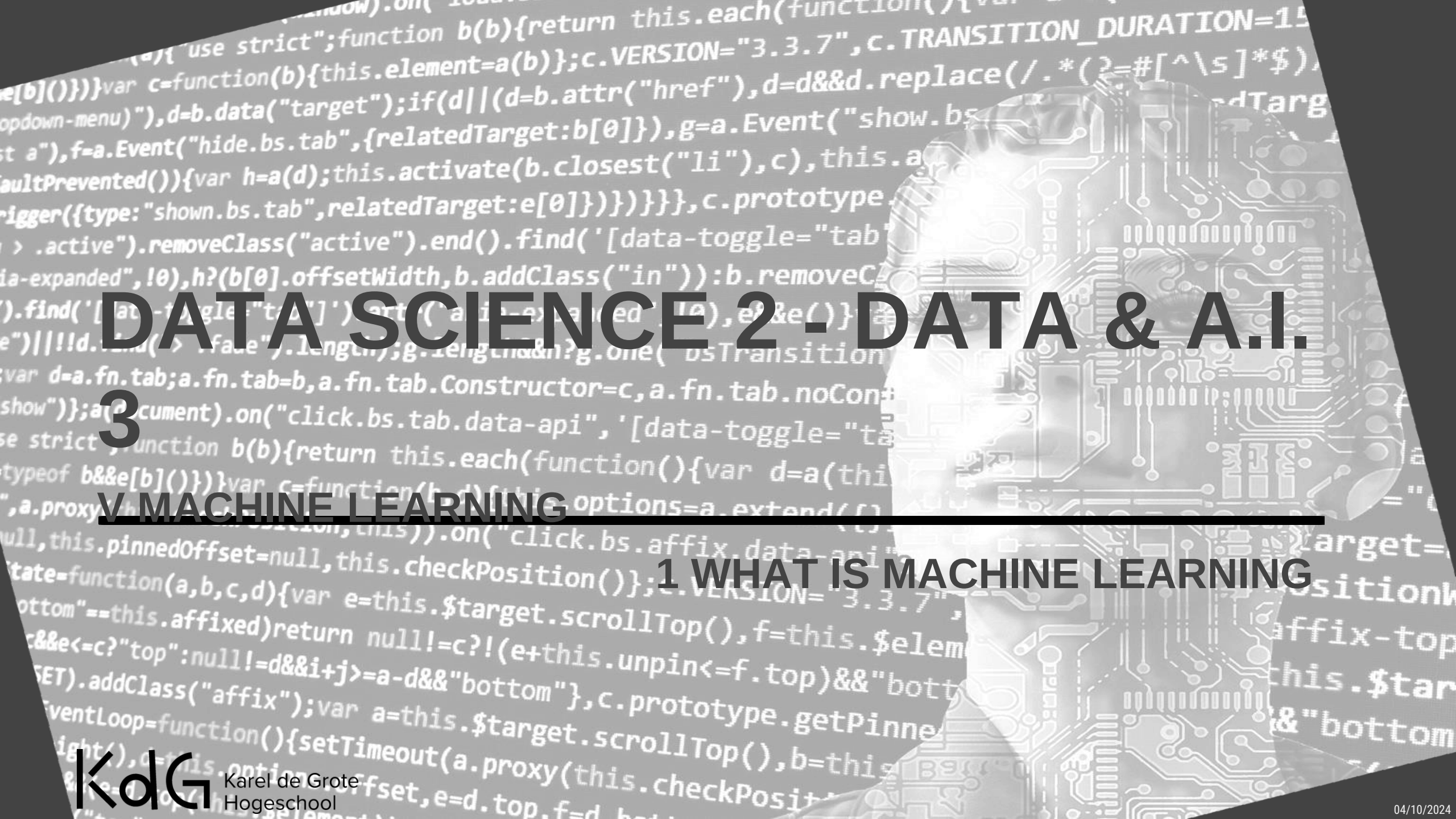
DATA SCIENCE 2 - DATA & A.I.

3 V MACHINE LEARNING

1 WHAT IS MACHINE LEARNING



Karel de Grote Hogeschool

DATA SCIENCE 2
DATA & A.I. 3

I

PYTHON BASICS

Python for data science

II

WORKING WITH ARRAYS

Numpy

III

DATA ENGINEERING

pandas

IV

DATA VISUALISATION

Matplotlib

V

MACHINE LEARNING

Automatically find patterns

V

MACHINE LEARNING

scikit-learn

MACHINE LEARNING

01 **WHAT IS MACHINE LEARNING**
Automatically find patterns

02 **INTRODUCING SCIKIT-LEARN**
Machine learning with Python

03 **HYPERPARAMETERS AND CROSS VALIDATION**
Holdout samples and cross-validation

04 **REGRESSION**
Best fitting line

05 **DECISION TREES**
Best separating lines

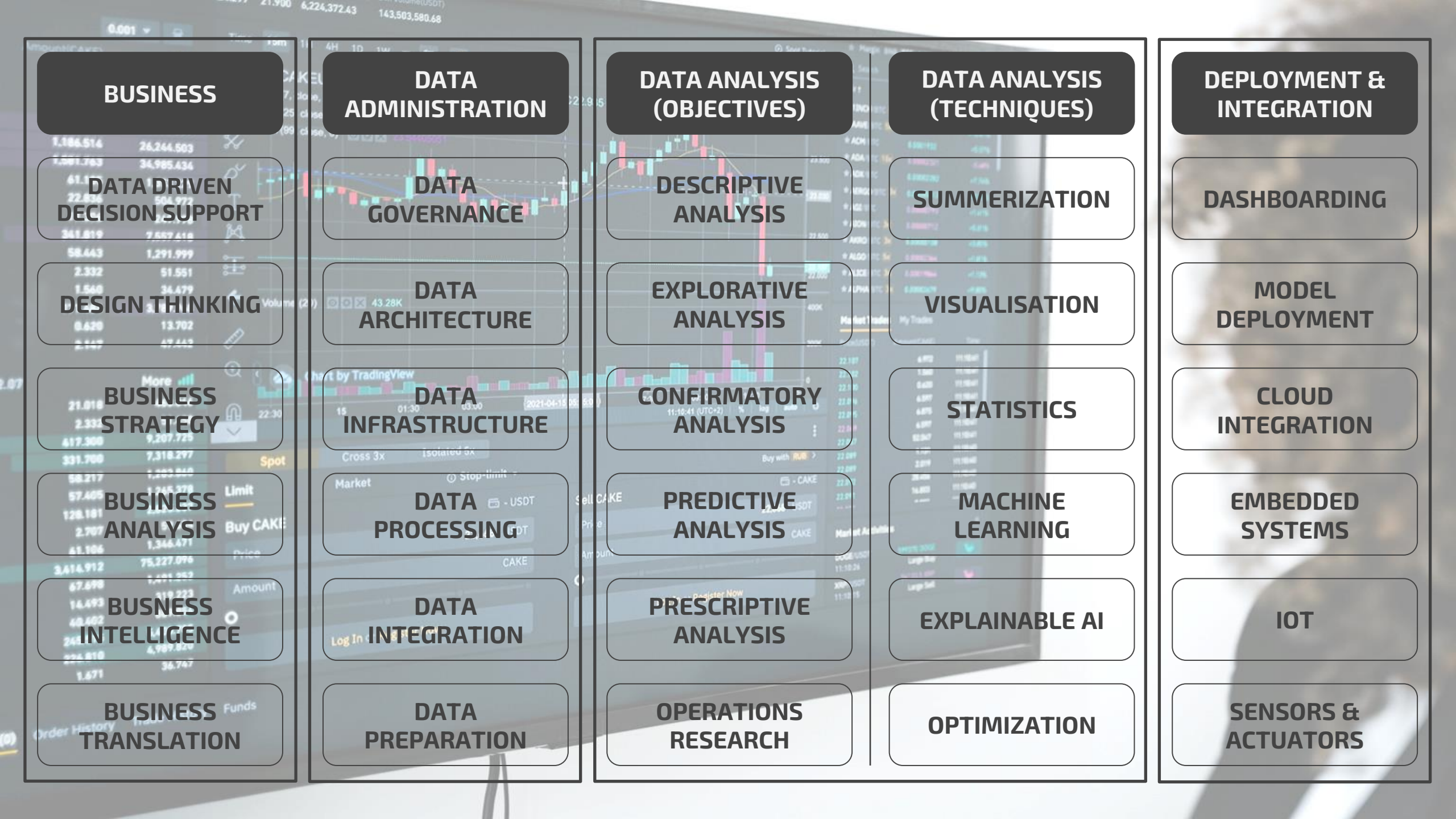
06 **K-MEANS CLUSTERING**
Object grouping

07 **ASSOCIATION RULES**
Frequent itemsets

08 **ARTIFICIAL NEURAL NETWORK**
Imitate the human brain

WHAT IS MACHINE LEARNING

Automatically find patterns

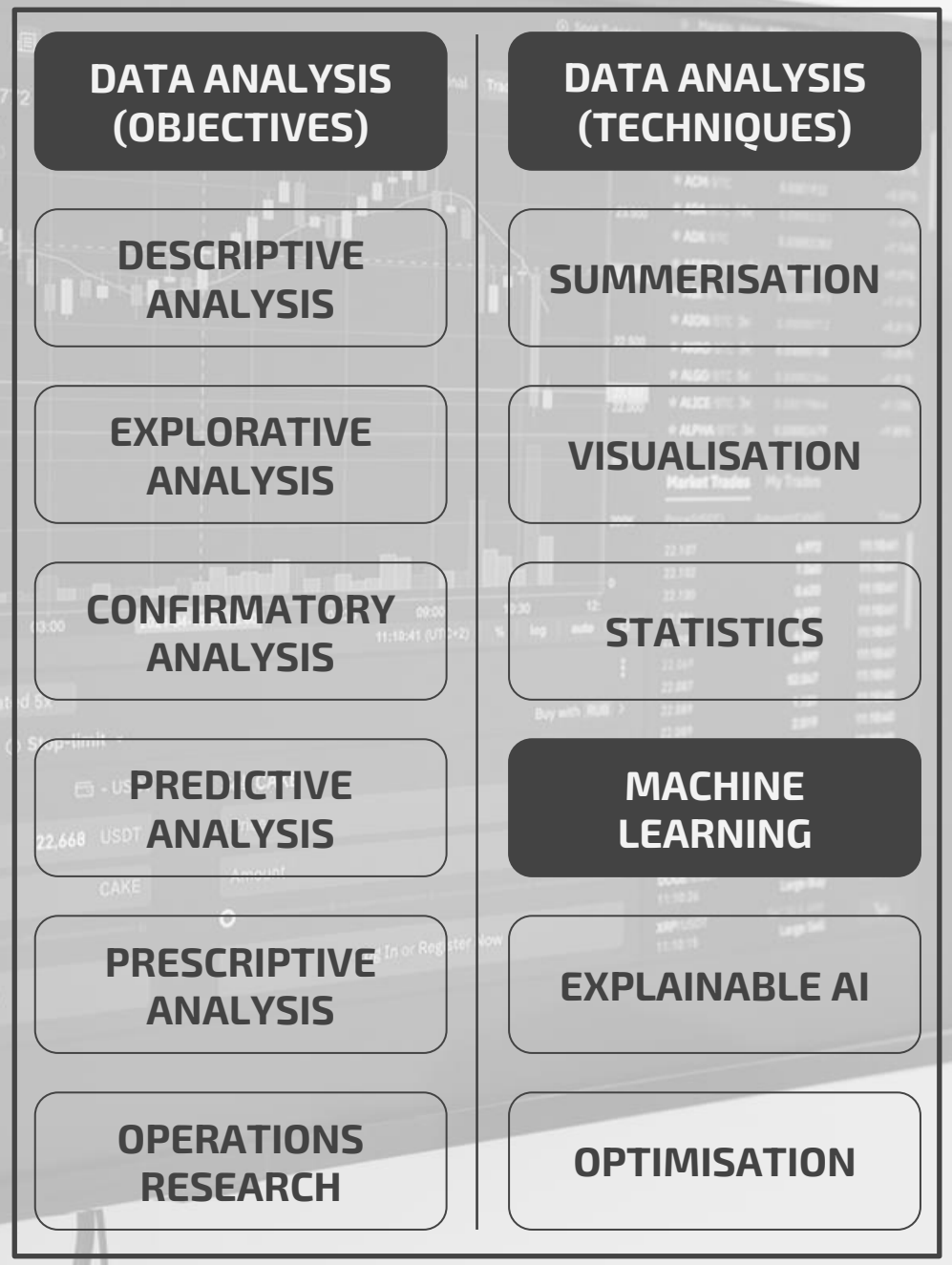

BUSINESS
DATA ADMINISTRATION
DATA ANALYSIS (OBJECTIVES)
DATA ANALYSIS (TECHNIQUES)
DEPLOYMENT & INTEGRATION
DATA DRIVEN DECISION SUPPORT
DESIGN THINKING
BUSINESS STRATEGY
BUSINESS ANALYSIS
BUSNESS INTELLIGENCE
BUSINESS TRANSLATION
DATA GOVERNANCE
DATA ARCHITECTURE
DATA INFRASTRUCTURE
DATA PROCESSING
DATA INTEGRATION
DATA PREPARATION
DESCRIPTIVE ANALYSIS
EXPLORATIVE ANALYSIS
CONFIRMATORY ANALYSIS
PREDICTIVE ANALYSIS
PRESCRIPTIVE ANALYSIS
OPERATIONS RESEARCH
SUMMERIZATION
VISUALISATION
STATISTICS
MACHINE LEARNING
EXPLAINABLE AI
OPTIMIZATION
DASHBOARDING
MODEL DEPLOYMENT
CLOUD INTEGRATION
EMBEDDED SYSTEMS
IOT
SENSORS & ACTUATORS

WHY

Why are we doing this

What do we want to achieve

What do we want as a result



How are we doing this

What steps will we take

What techniques will we use

HOW

DATA ANALYSIS

VISUAL EXPLORATION

Get insights from the data

- Difficult to visualize more than 3 variables
- Inconsistent interpretation

STATISTICAL ANALYSIS

Formal test of a model (confirmation)
Implicit significance testing
(likelihood pattern is observed by coincidence)

- Difficult to use when many variables are present and model unknown
- Assumptions on distribution characteristics of population

MACHINE LEARNING

Let the computer derive a pattern
No distribution characteristics needed
=> Far more methods available to find patterns

- No implicit significance testing
- Beware of overfitting

MACHINE LEARNING

“TRADITIONAL” DATA ANALYSIS

- Formulate hypothesis
- Collect data
- Explore data (and refine hypothesis if needed)
- Build model according to hypothesis and exploration
- Test for statistical significance (test for coincidence because of sample)
- Derive conclusions

=> SEARCH FOR PATTERNS YOURSELF

=> STARTS FROM HYPOTHESIS OR EXPLORATION

MACHINE LEARNING

- Collect data
- Train model
- Test model (test for coincidence because of sample
- Derive conclusions

=> LET THE COMPUTER SEARCH FOR PATTERNS

=> STARTS FROM DATA

WHAT KIND OF PROBLEMS CAN YOU SOLVE WITH MACHINE LEARNING

- Value estimation / regression: predict a continuous variable
  e.g. sales prediction; car use

- Classification: predict a categorical variable
  e.g. churn prediction; diagnosis

- Segmentation / clustering: split or group cases/observations
  e.g. customer segmentation; document topic search

- Co-occurrence / association rule discovery: events happening together
  e.g. market basket analysis ; recommendation system

WHAT KIND OF PROBLEMS CAN YOU SOLVE WITH MACHINE LEARNING

SUPERVISED

- Value estimation / regression: predict a continuous variable
  e.g. sales prediction; car use
- Classification: predict a categorical variable
  e.g. churn prediction; diagnosis

- Segmentation / clustering: split or group cases/observations
  e.g. customer segmentation; document topic search
- Co-occurence / association rule discovery: events happening together
  e.g. market basket analysis ; recommendation system

UNSUPERVISED

MAJOR DATA ANALYTICS SUBTASKS

CLUSTERING

I want to group/segment my cases, but

- I have no clue what kind of classification to use (how many classes and/or what kind of classes) => exploration
- I have a clue about the classification to use, but I'm unable to get example cases (labeled cases) => prediction

Make groups based on some kind of similarity in features/variables

Use case: Exploration, segmentation

Methods: Hierarchical clustering (Ward, Single linkage Complete linkage, …);

Non-hierarchical clustering (k-means)

MAJOR DATA ANALYTICS SUBTASKS

ASSOCIATION RULE DISCOVERY

I want to group events/features that happen together

Similar to clustering, but now we do not group cases (rows) but events/features (mostly columns/variables)

Make groups based on co-occurrence of events/features (frequent itemsets)

Use case: Recommender systems, market basket analysis

Methods: Association rule mining

(subtle difference with clustering: clustering will reveal which cases have something in common, association rule discovery will reveal what they have in common. You could get the same kind of information by interpreting your cluster solution, but better to do that by directly using association rule discovery)

MAJOR DATA ANALYTICS SUBTASKS

ESTIMATION

Predict the value of a numerical value based on a set of features (set of numerical or categorical variables)

Regress a model based on examples (labeled cases)

Use case: All kind of numerical predictions

Methods: (linear) regression

MAJOR DATA ANALYTICS SUBTASKS

CLASSIFICATION

Predict the value of a categorical value based on a set of features (set of numerical or categorical variables)
Related to clustering, but now we know the classes and we have example cases (labeled cases)

Make groups based by deriving rules or deriving hyperplanes based on example cases (labeled cases)

Use case: All kind of classifications

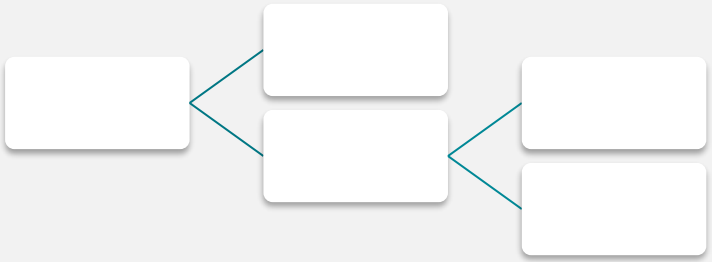
Methods: Decision tree; logistic regression; SVM; artificial neural network

(can also be used for estimation if you transform your numerical variable into a categorical variable using ranges)

SUPERVISED LEARNING : TRAINING (DERIVE A MODEL FROM LABELED DATA)

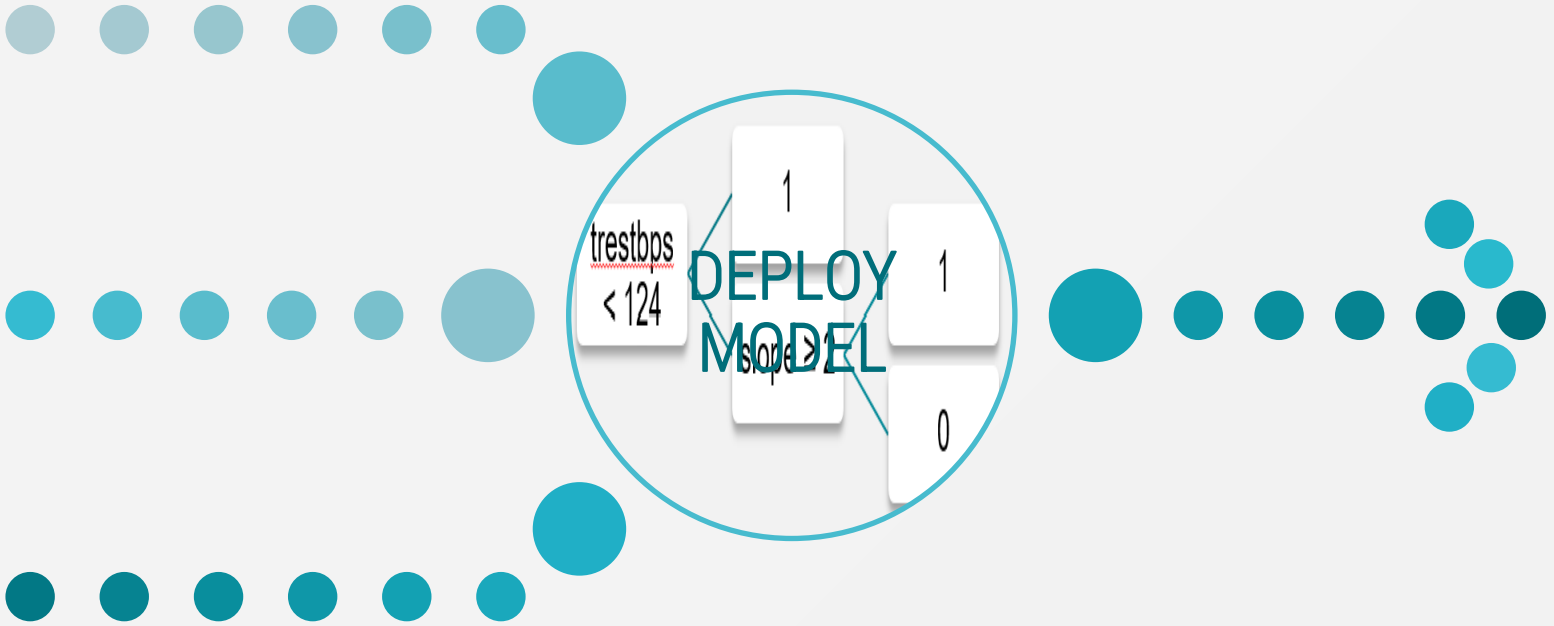
Examples (labeled cases)

Model



SUPERVISED LEARNING : DEPLOYMENT (APPLY MODEL ON NEW DATA)

New data (new cases)



Results

Classification
or
Prediction