DATA SCIENCE 2 - DATA & A.I. 3

V MACHINE LEARNING

1 WHAT IS MACHINE LEARNING

KdG Karel de Grote Hogeschool

PYTHON BASICS

Python for data science

I

WORKING WITH ARRAYS

Numpy

II

DATA ENGINEERING

pandas

III

DATA SCIENCE 2
DATA & A.I. 3

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

WHY

Why are we doing this

What do we want to achieve

What do we want as a result

DATA ANALYSIS (OBJECTIVES)	DATA ANALYSIS (TECHNIQUES)
DESCRIPTIVE ANALYSIS	SUMMERISATION
EXPLORATIVE ANALYSIS	VISUALISATION
CONFIRMATORY ANALYSIS	STATISTICS
PREDICTIVE ANALYSIS	MACHINE LEARNING
PRESCRIPTIVE ANALYSIS	EXPLAINABLE AI
OPERATIONS RESEARCH	OPTIMISATION

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

MACHIN 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-occurence / 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

UNSUPERVISED

- 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

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 colums/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 featurs (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

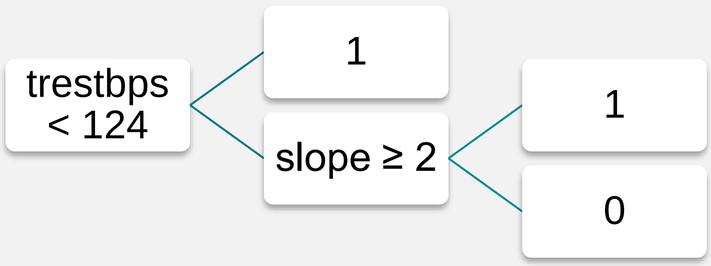
Methods: Decision tree; logistic regression; SVM; artificial neural network

(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)

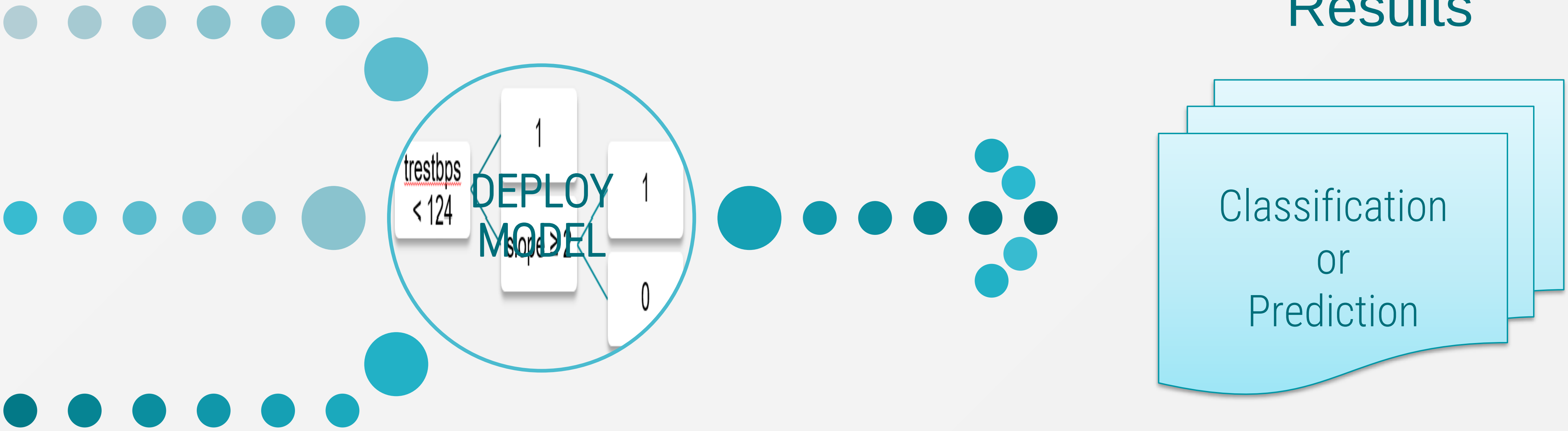
Model



SUPERVISED LEARNING : DEPLOYMENT (APPLY MODEL ON NEW DATA)

New data (new cases)

Results



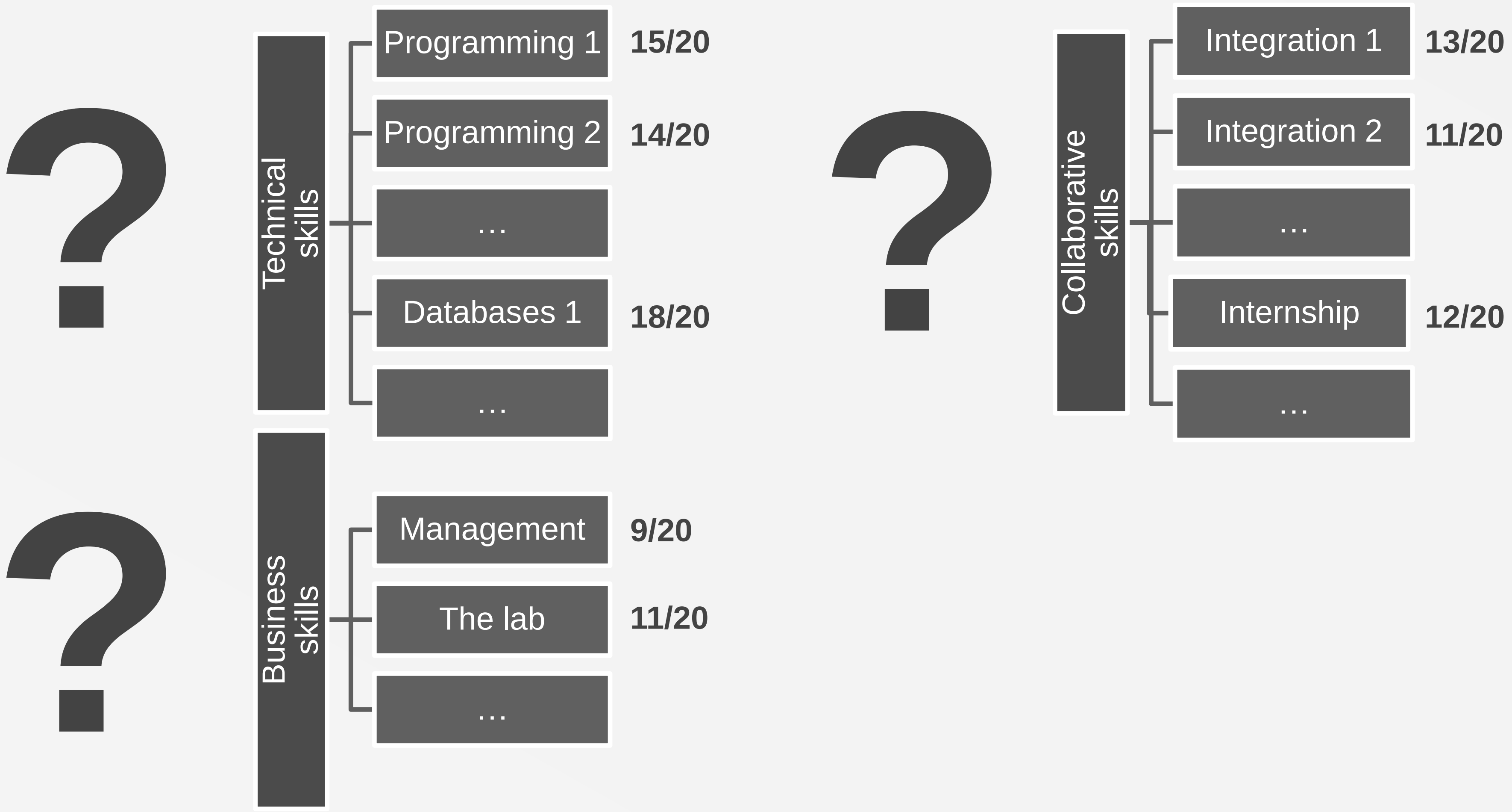
DIMENSIONALITY REDUCTION
(NOT COVERED IN DATA SCIENCE 1 / DATA & A.I. 2)

PRACTICAL EXAMPLE : IT STUDENT SKILL CHARACTERIZATION

You want to characterise IT students at the end of their studies to get an idea in what fields the student is strong.
You could use all the course grades of the students to do that, indicating for which courses the student was strong. But that might result in tens of numbers to charactirse the student (e.g. 30 grades after 3 years of bachalors study). That would work well for students performing very well on all courses, or students performing very poor on all courses. But for students scoring good at some courses and poor on others, 30 grades is just too much detail to get a quick grasp.

As there is some correlation between course grades, you do not need all the detailed grades, but you could group them in relevant classes: technical skills, business skills, collaboration skills, ..., so you only need a few numbers to quickly get an idea of the skills of a student.

DIMENSIONALITY REDUCTION
(NOT COVERED IN DATA SCIENCE 1 / DATA & A.I. 2)



DIMENSIONALITY REDUCTION
(NOT COVERED IN DATA SCIENCE 1 / DATA & A.I. 2)

PRACTICAL EXAMPLE : IT STUDENT SKILL CHARACTERIZATION

You could do this manually, carefully grouping courses into groups (e.g. group courses about programming and databases into ‘technical skills’) and calculate a mean score for every group to end up with a limited set of indictors (‘technical skills’, ‘business skills’, ...)

For big datasets it is not straightforward anymore to define those groups, although you could use clustering to help you with that, as clustering students on course grades will result in groups of students that score similar on a set of courses; this will probably reveal the technical students (all scoring high on technical courses), more business-oriented users (all scoring high on business courses), and so on.

DIMENSIONALITY REDUCTION
(NOT COVERED IN DATA SCIENCE 1 / DATA & A.I. 2)

PRACTICAL EXAMPLE : IT STUDENT SKILL CHARACTERIZATION

The clue of dimensionality reduction as a machine learning technique is to do this automatically, i.e. let the computer combine a large set of features into a limited set of new features based on the correlation structure in the data (by combining features with high correlation).

Mind the relation with clustering: clustering will combine observations (rows in a dataset) into groups based on similar features. Dimensionality reduction will combine features (colums in a dataset) into groups based on shared correlation. The difference is that clustering will define groups, while dimensionality reduction will replace a set of features by a limited set of new features.

DIMENSIONALITY REDUCTION
(NOT COVERED IN DATA SCIENCE 1 / DATA & A.I. 2)