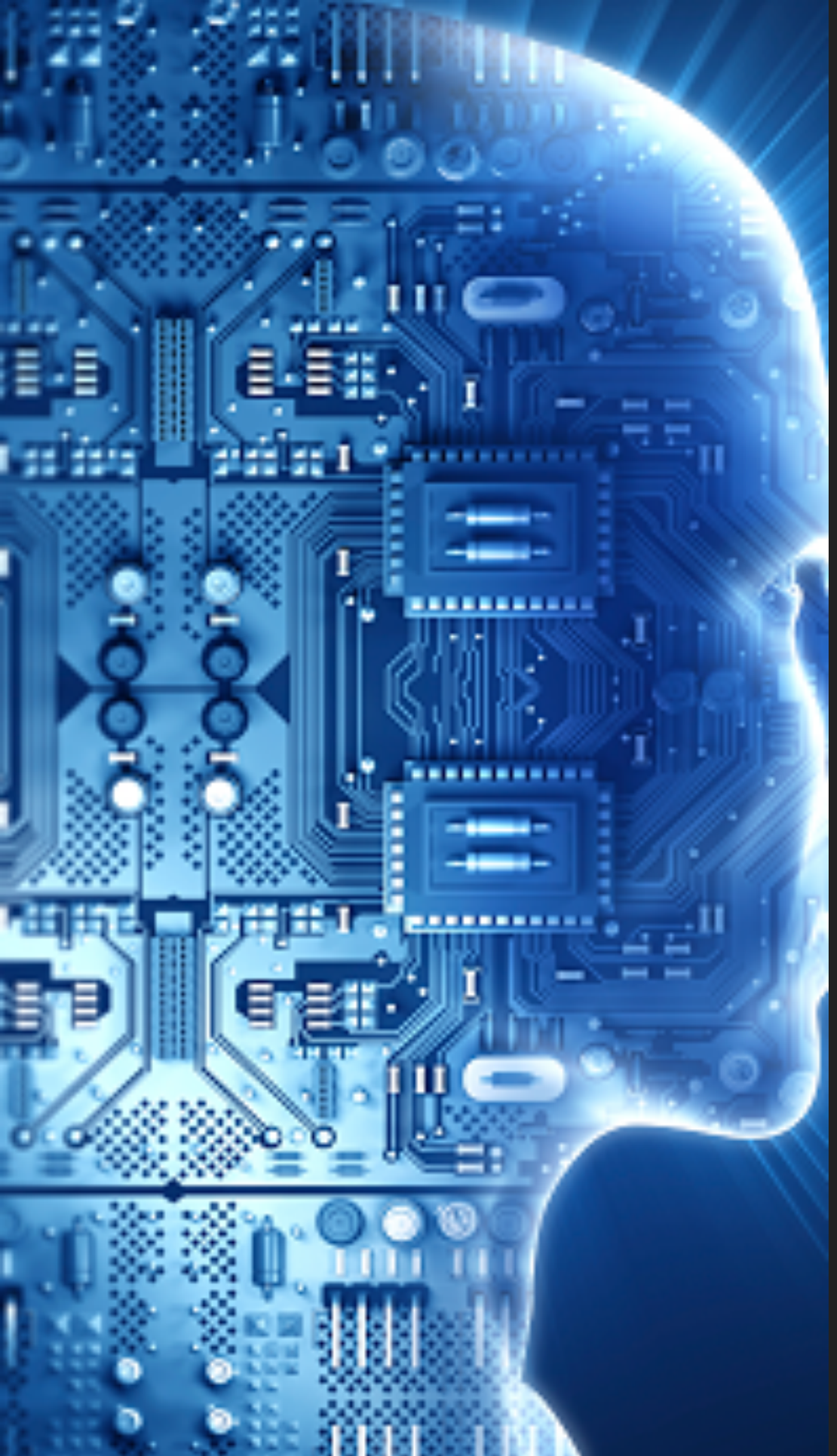


INTRODUCTION TO AI AND MACHINE LEARNING

SESSION #2

COURSE AGENDA

Session #1: Introduction to machine learning, concepts, basics, capabilities. Classification basics.

Session #2: Feature engineering, data wrangling. Regression basics.

Session #3: Working with textual data, text classification, NLP basics

Session #4: Introduction to neural networks, deep learning, image recognition

SESSION #2 AGENDA

SECTION 1

- Overview on classification algorithms
- Four level of data
- Feature engineering

SECTION 2

- Case Study: Predicting house prices on the King County House Sales dataset

SCIKIT LEARN CLASSIFIER OVERVIEW

Scikit Learn provides numerous classifiers to work with:

Simple algorithms:

- LogisticRegression
- DecisionTreeClassifier
- Support vector machines
- Naive-Bayes classifiers
- Nearest Neighbours classifier

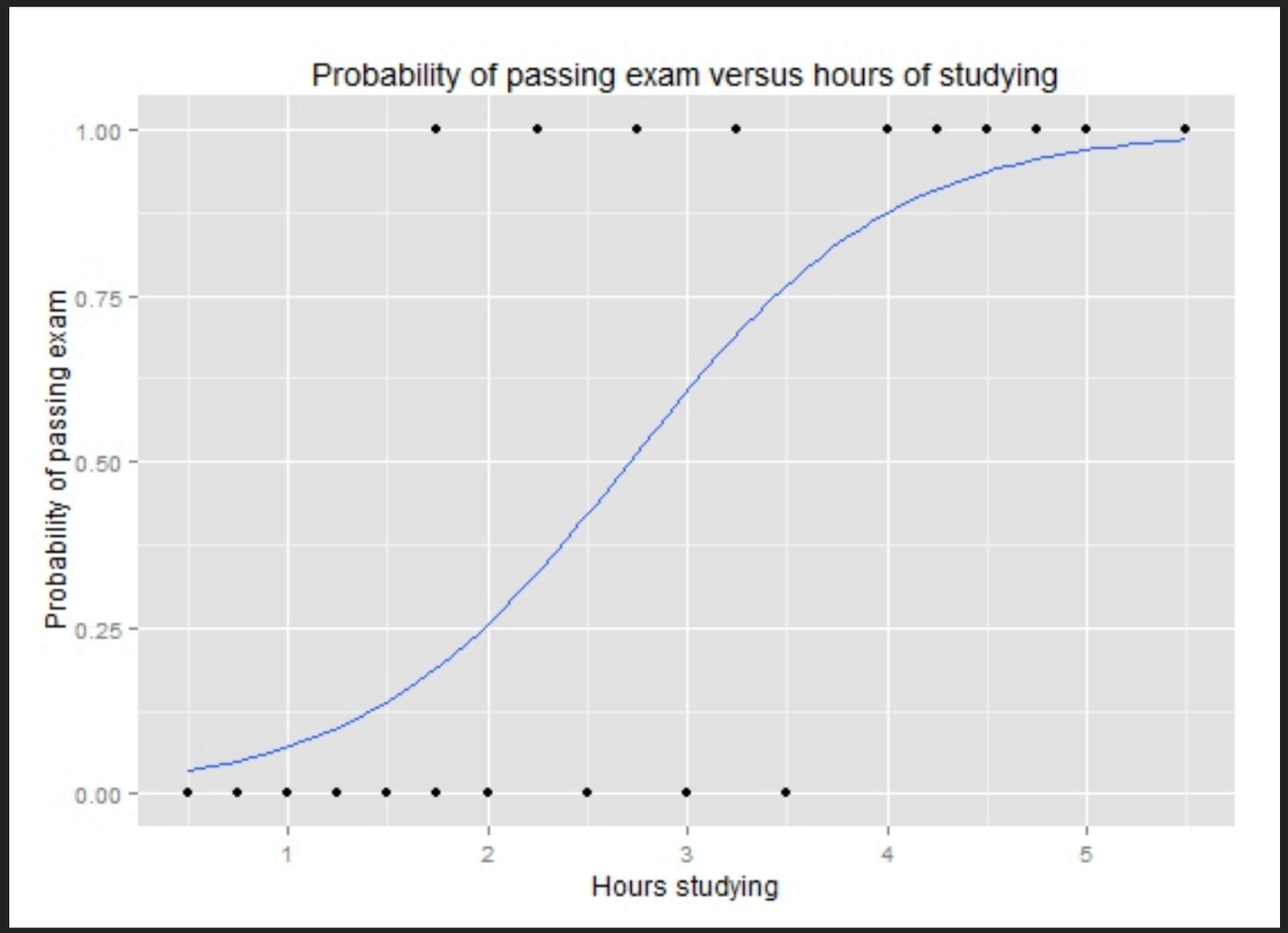
Ensemble methods:

- RandomForestClassifier
- BaggingClassifier
- GradientBoostingClassifiers
- VotingClassifier
- AdaBoostClassifier

LOGISTIC REGRESSION CLASSIFIER OVERVIEW

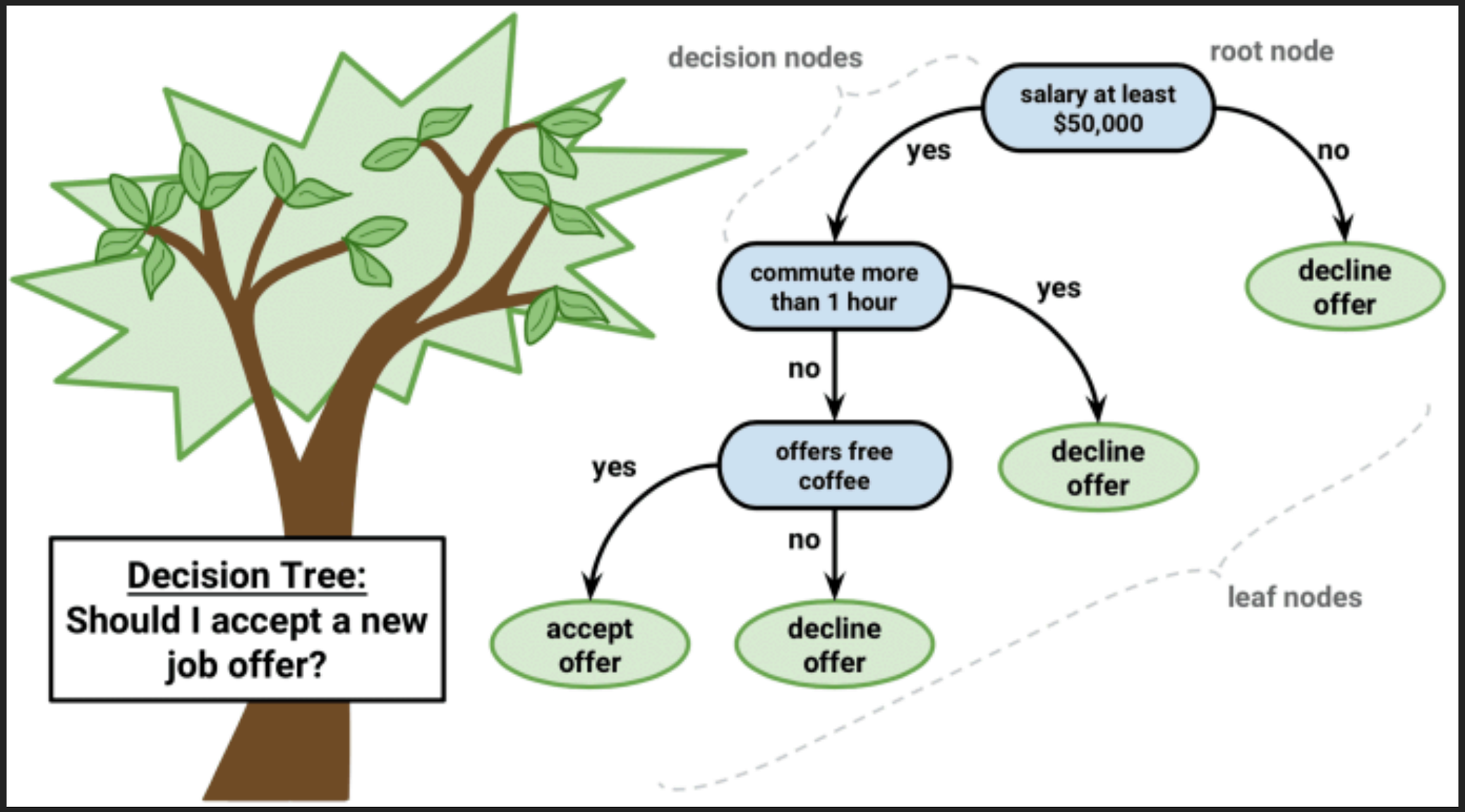
Table 1

Hours	Pass
0.50	0
0.75	0
1.00	0
1.25	0
1.50	0
1.75	0
1.75	1
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.50	1
4.75	1
5.00	1
5.50	1



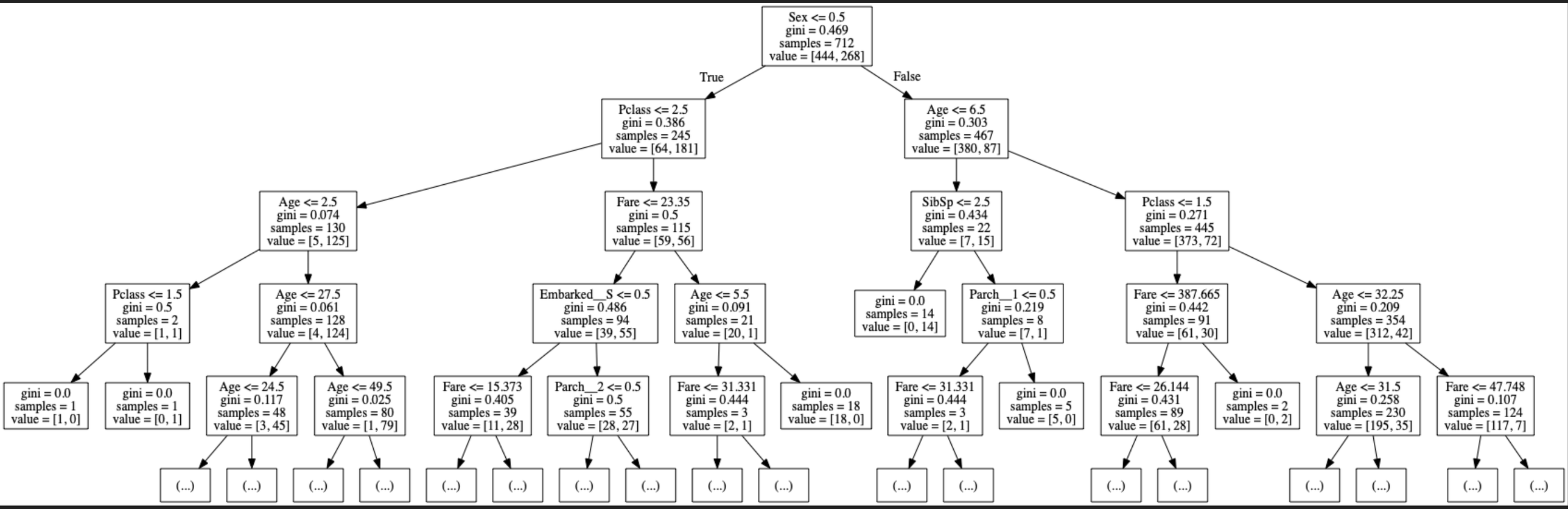
Source: https://en.wikipedia.org/wiki/Logistic_regression

DECISION TREE CLASSIFIER OVERVIEW



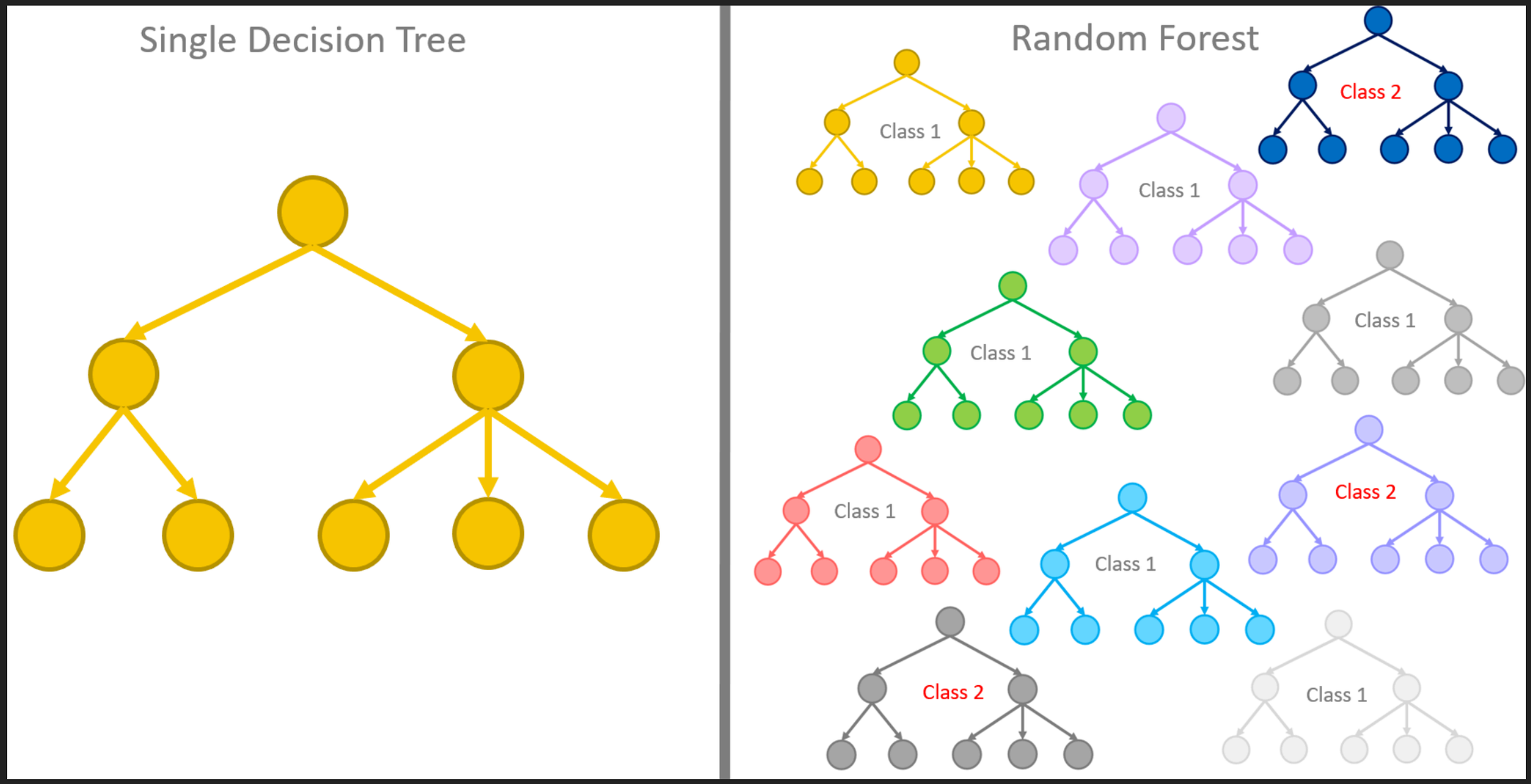
Source: https://mc.ai/mathematics-behind-decision-tree/

DECISION TREE FOR TITANIC



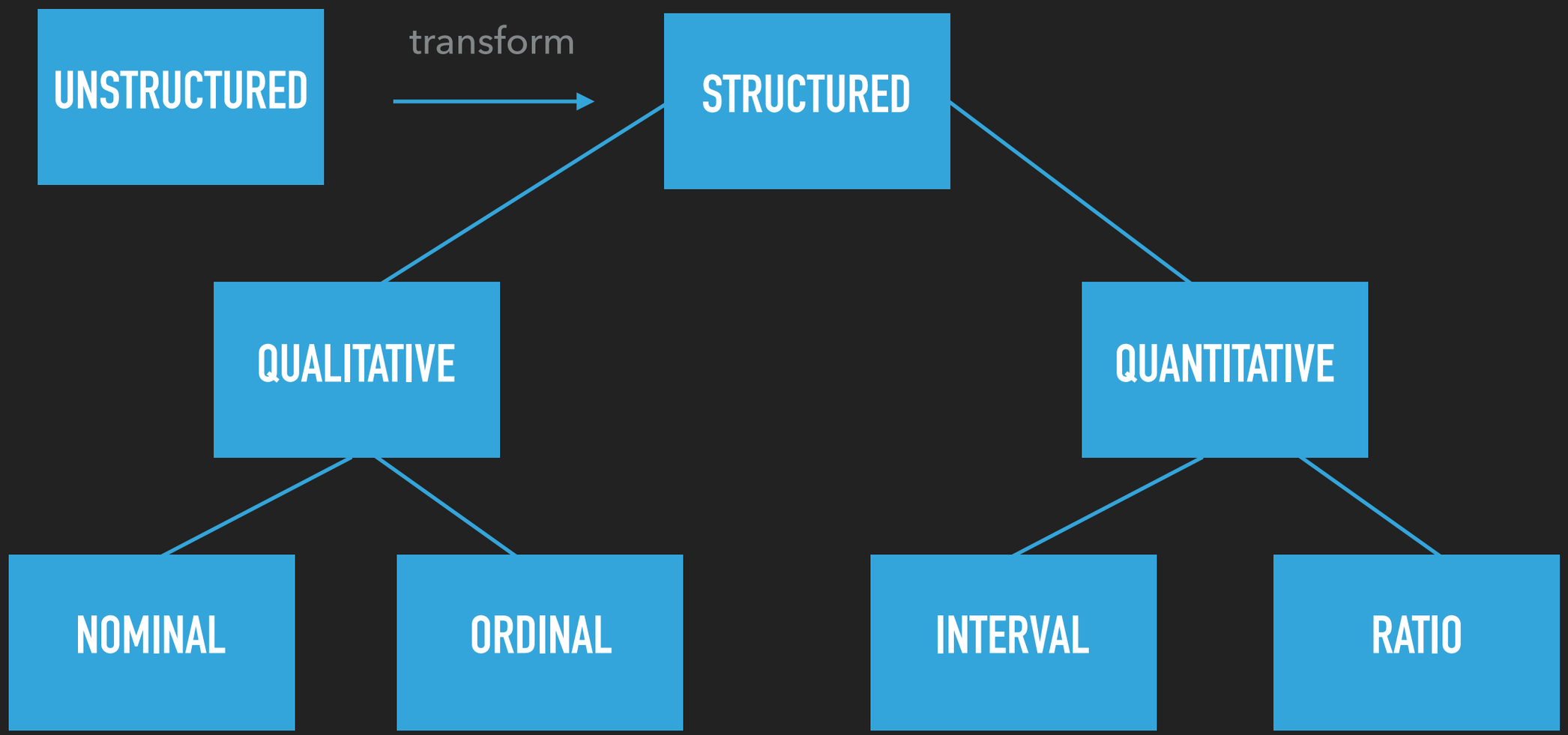
RANDOM FOREST CLASSIFIER OVERVIEW

Final decision is the predicted value voted by the most of the trees.



Source: https://www.dataversity.net/from-a-single-decision-tree-to-a-random-forest/#

DATA TYPES


UNSTRUCTURED
transform
STRUCTURED
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO

DATA TYPES: UNSTRUCTURED DATA

- text
- images
- sound, e.g.: phone recordings, music
- sensor data

Note: semi-structured data: e.g. email, XML, JSON

DATA TYPES: QUALITATIVE DATA

Nominal

- discrete values
- may be categorical
- often useless for ML
- no mathematical operations

Ordinal

- natural order exists
- ‘Good’, ‘Average’, ‘Poor’
- Likert scale
- numeric comparison possible

Qualitative data has to be converted numeric to be used for machine learning problems!

DATA TYPES: QUANTITATIVE DATA

Interval

- meaningful difference between values
- adding and subtraction possible
- no multiplication or division
- e.g temperature, degree

Ratio

- continuous
- introduces true zero (real absence)
- all mathematical operations possible
- e.g. money, weight

FEATURE ENGINEERING

Goals:

Preparing the proper input dataset for a given machine learning algorithm

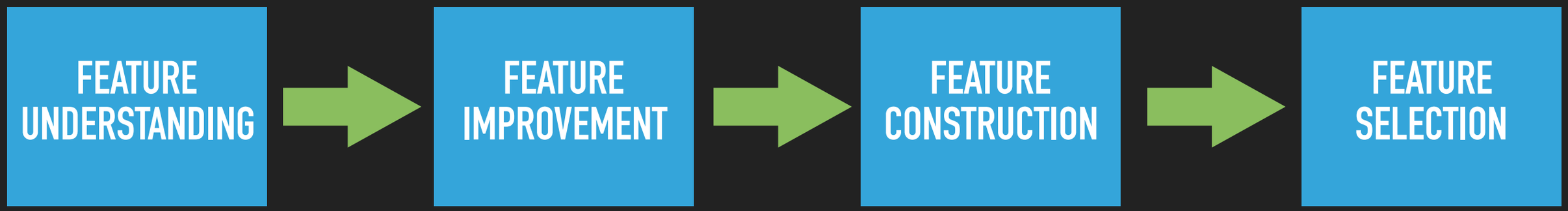
Improving the performance of a machine learning algorithm

Attribute: is or derived from an observable property

Feature: an attribute or an internal representation of data

Dimension: every feature creates a dimension in the feature space

FEATURE ENGINEERING



FEATURE UNDERSTANDING

Goals:

Understand the dataset

Including:

- structured vs. unstructured data, identifying data levels
- identifying missing values
- exploratory data analysis
- descriptive statistics
- data visualisations

FEATURE IMPROVEMENTS

Goals:

Clean the dataset and prepare for machine learning

Including:

- structuring unstructured data
- imputing missing data
- removing outliers
- data normalisation

FEATURE CONSTRUCTION

Goals:

Creating new features based on existing ones or other datasets

Including:

- encoding categorical variables
- deriving features from existing ones (e.g. date of birth -> age)
- creating features from feature interaction (e.g. weight / height -> BMI)
- bringing in features from additional data sources (e.g. GPS coordinates -> country)

FEATURE SELECTION

Goals:

Improve model performance (predictive power, speed)

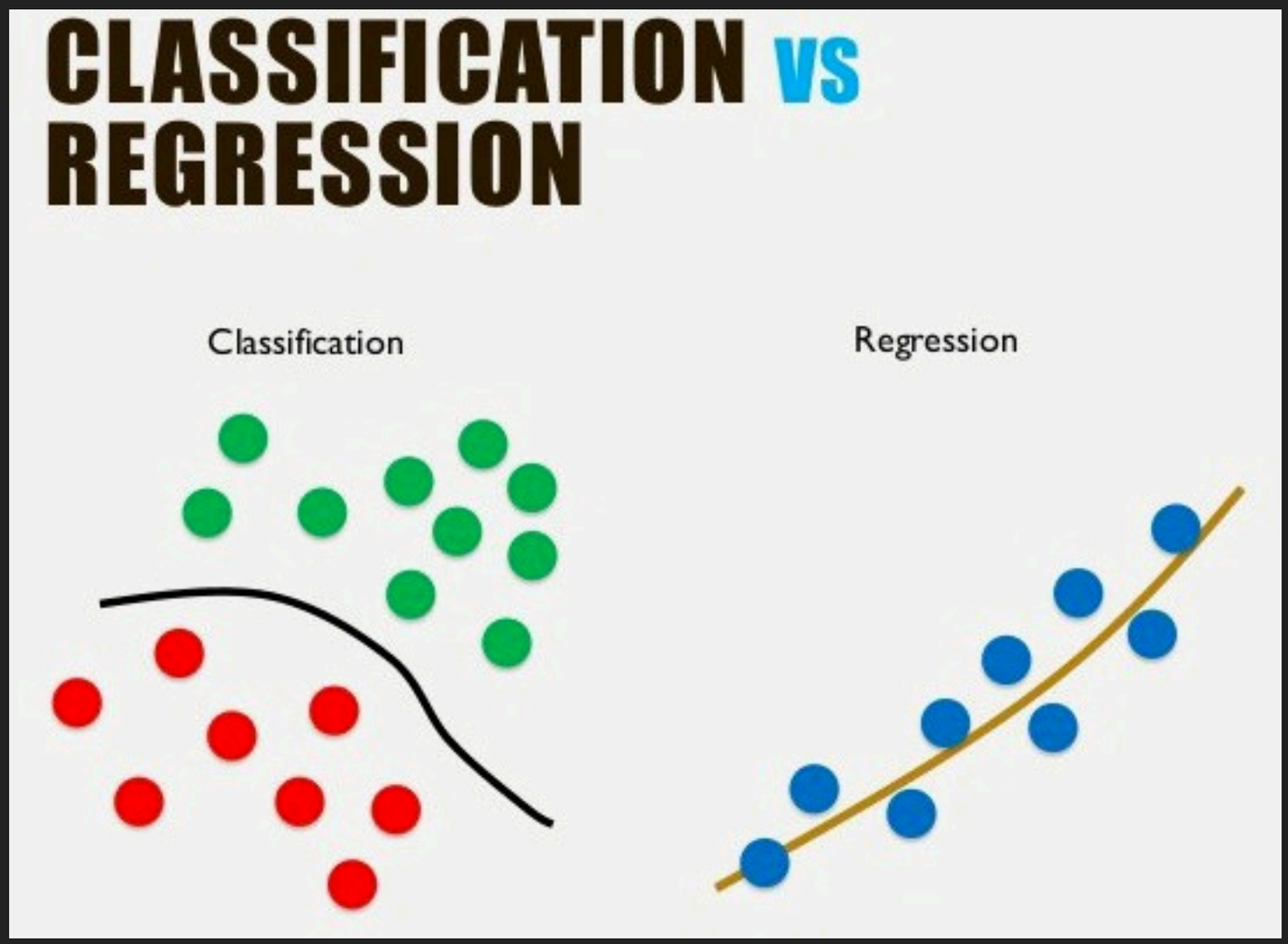
Including:

- selecting features based on statistical methods (e.g. correlation matrix)
- selecting features based on hypothesis testing (p-value)
- selecting features with model based or machine learning based methods
- manual feature selection based on domain knowledge

FEATURE ENGINEERING: EXERCISE

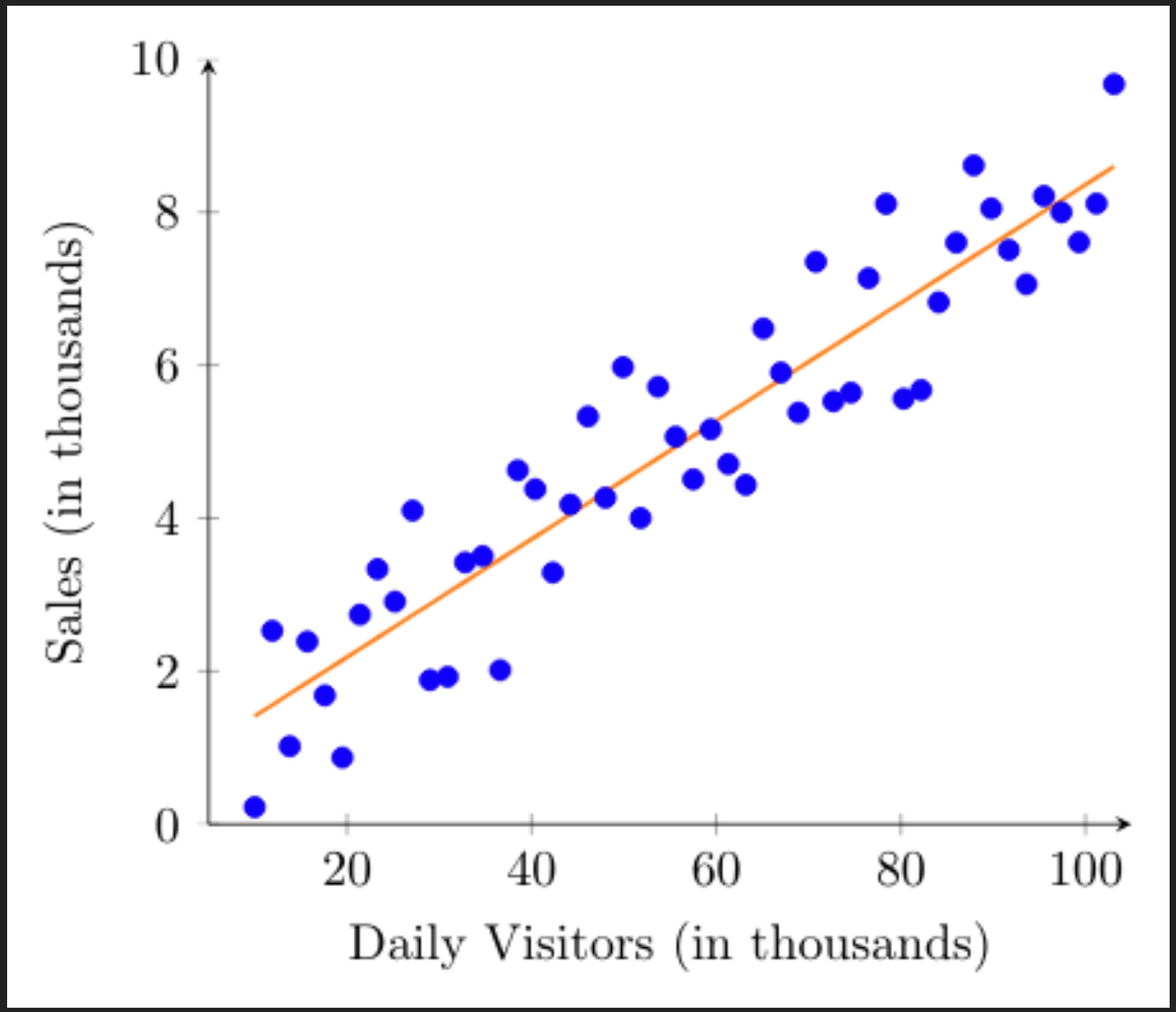
```
192.168.1.4 - - [26/Apr/2018:17:18:54 +0200] "GET /2017/05/30/how-to-retrieve-ii
s-http-logs-remotely/ HTTP/1.1" 301 671 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64;
 Trident/7.0; rv:11.0) like Gecko"
192.168.1.4 - - [26/Apr/2018:17:18:54 +0200] "GET /2017/05/30/how-to-retrieve-ii
s-http-logs-remotely/ HTTP/1.1" 200 21466 "-" "Mozilla/5.0 (Windows NT 6.1; WOW6
4; Trident/7.0; rv:11.0) like Gecko"
192.168.1.4 - - [26/Apr/2018:17:19:00 +0200] "GET /2017/05/ HTTP/1.1" 200 19215
"https://www.wptest.com/2017/05/30/how-to-retrieve-iis-http-logs-remotely/" "Moz
illa/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"
192.168.1.4 - - [26/Apr/2018:17:19:03 +0200] "GET /2017/05/30/hello-world/ HTTP/
1.1" 200 19742 "https://www.wptest.com/2017/05/" "Mozilla/5.0 (Windows NT 6.1; W
OW64; Trident/7.0; rv:11.0) like Gecko"
webmaster@USERVER:~$
```

RECAP: SUPERVISED LEARNING



Source: https://www.codeingschool.com/2019/06/regression-classification-supervised-machine-learning.html

LINEAR REGRESSION



DEMO

RECAP

Today we learnt:

- how some basic classification algorithms work on a high level
- what are the four levels of data
- the essentials of feature engineering
- we built a regression model to predict house prices

HOMEWORK

AMES HOUSE PRICING DATASET

Can you build a machine learning model to accurately predict house prices?

Data and description is also available at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data