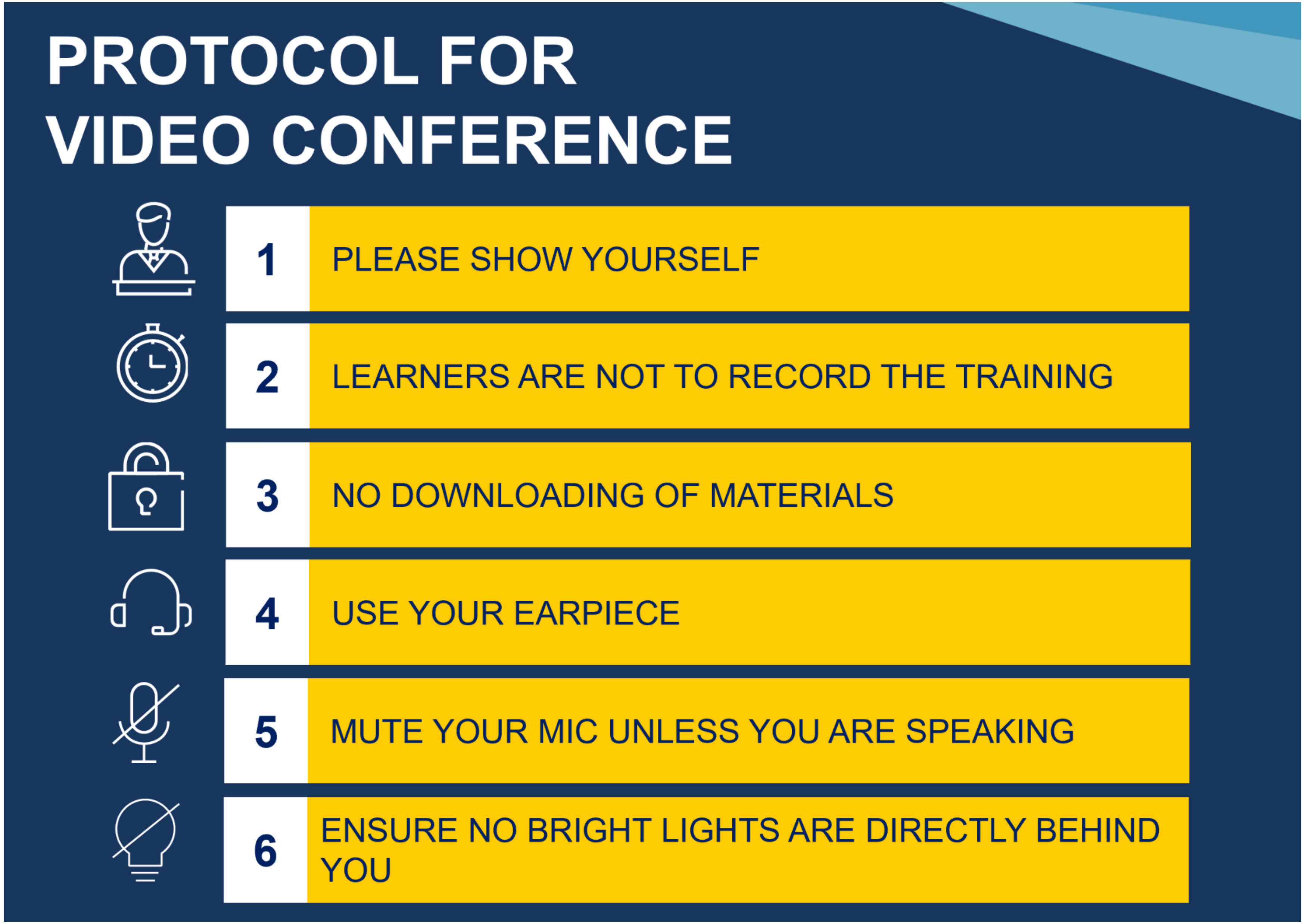
ntuc LearningHub

Machine Learning and Advanced Analytics using Python (Day-1)



PROTOCOL FOR VIDEO CONFERENCE

1 PLEASE SHOW YOURSELF

2 LEARNERS ARE NOT TO RECORD THE TRAINING

3 NO DOWNLOADING OF MATERIALS

4 USE YOUR EARPIECE

5 MUTE YOUR MIC UNLESS YOU ARE SPEAKING

6 ENSURE NO BRIGHT LIGHTS ARE DIRECTLY BEHIND YOU

Today’s Schedule

Too Much Stress??

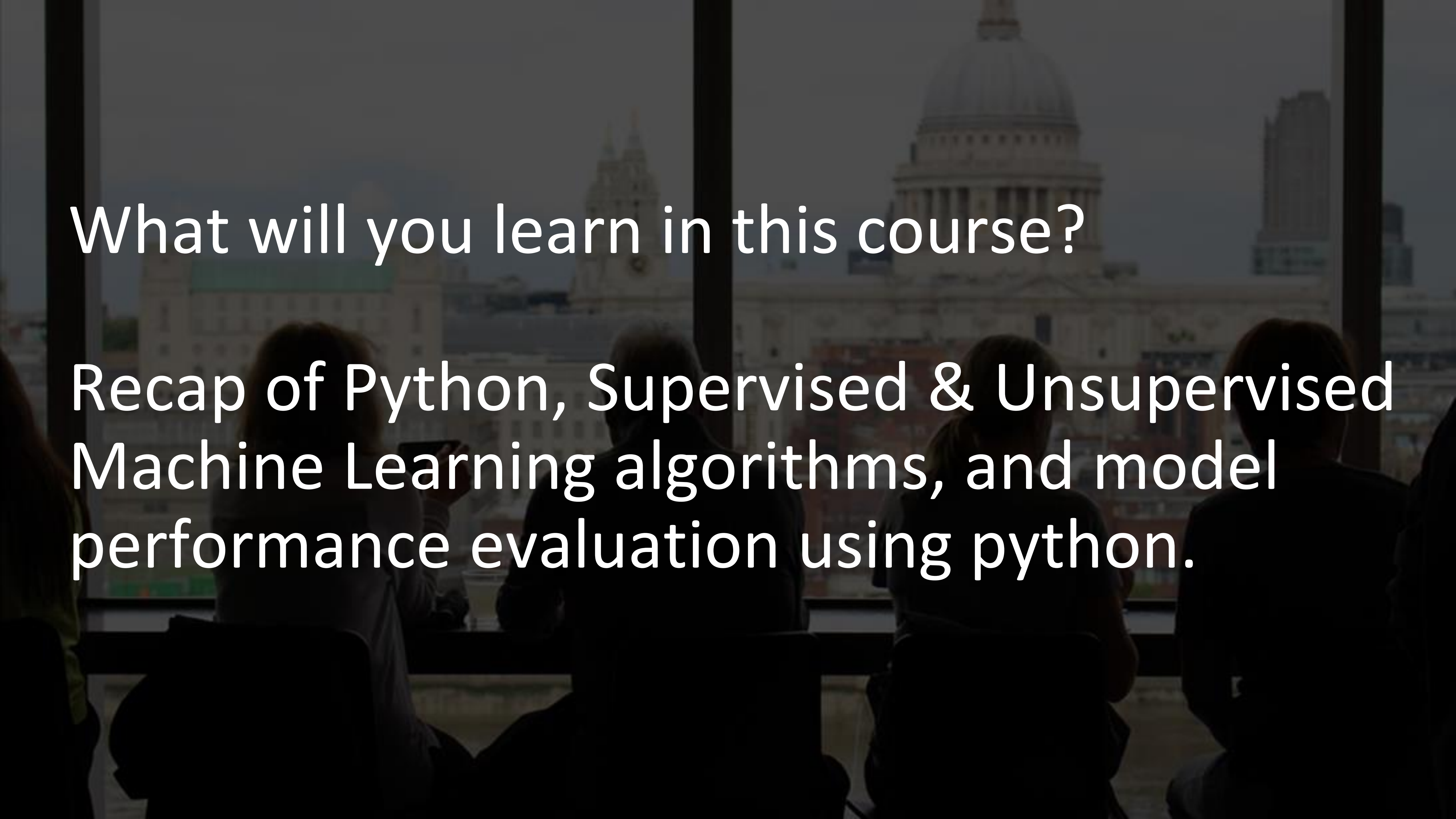
9am: Session Start

10-30am: 15min break

12-30pm: Lunch break

3-30pm: 15 min break

6pm: Session End



What will you learn in this course?

Recap of Python, Supervised & Unsupervised Machine Learning algorithms, and model performance evaluation using python.

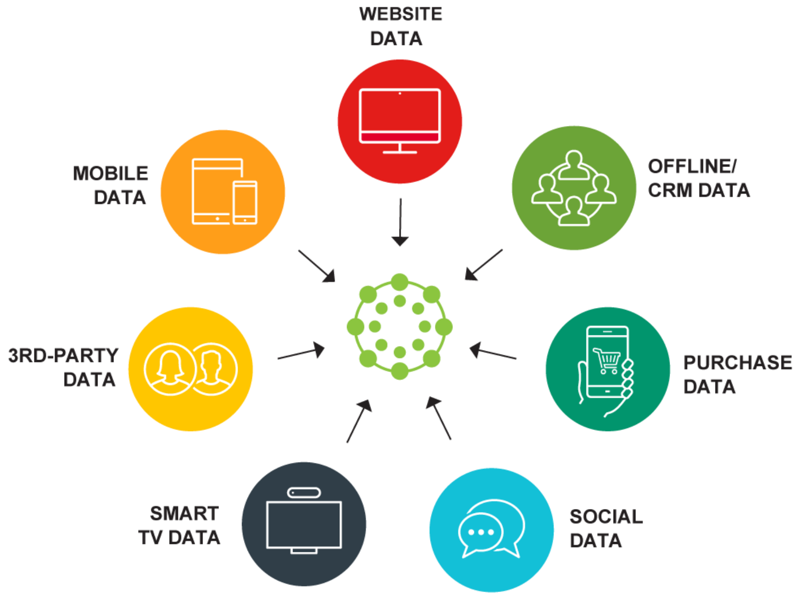
How much data do we create every single day?

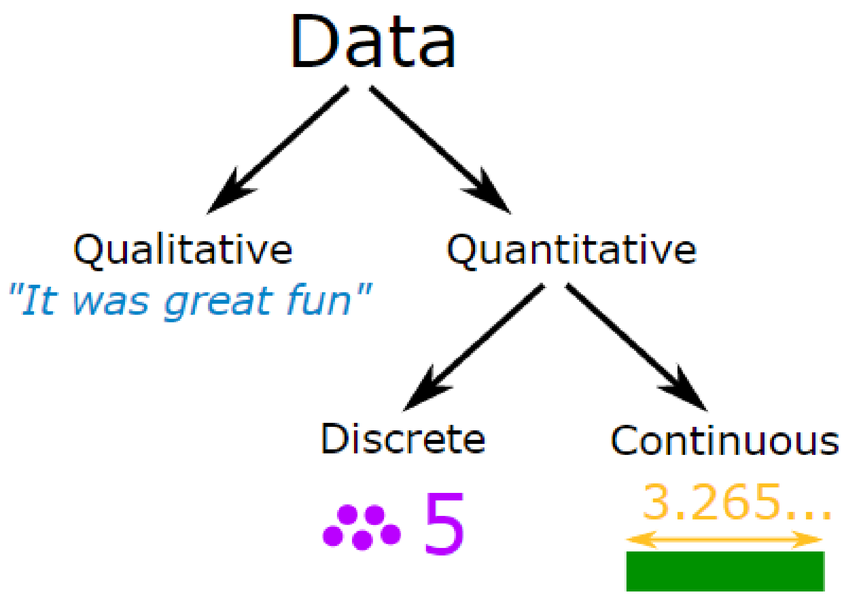
2,500,000,000,000,000,000

(Two and half quintillion)

What is data?

Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.





A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

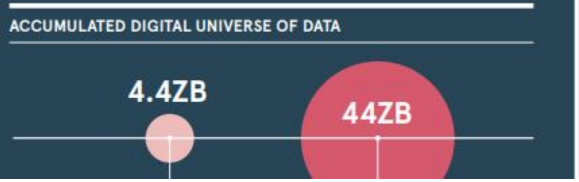
people use emails

4TB

of data produced by a connected car

Intel

ACCUMULATED DIGITAL UNIVERSE OF DATA



DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

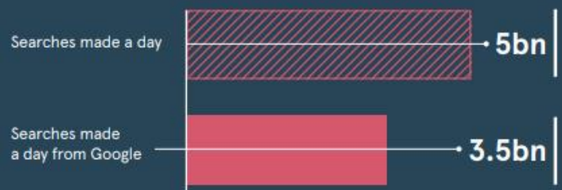
Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



Smart insights

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

28PB

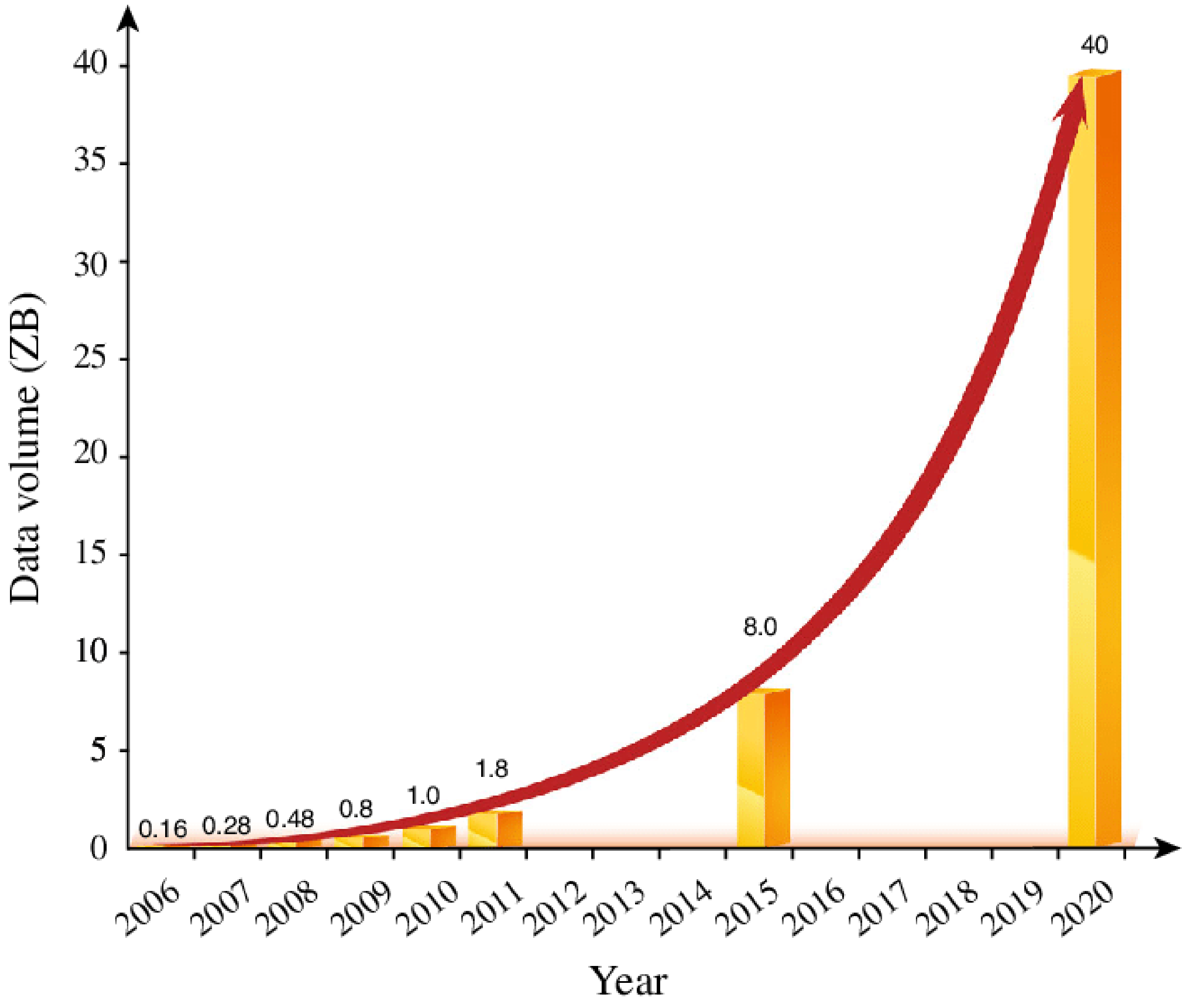
to be generated from wearable devices by 2020

Statista

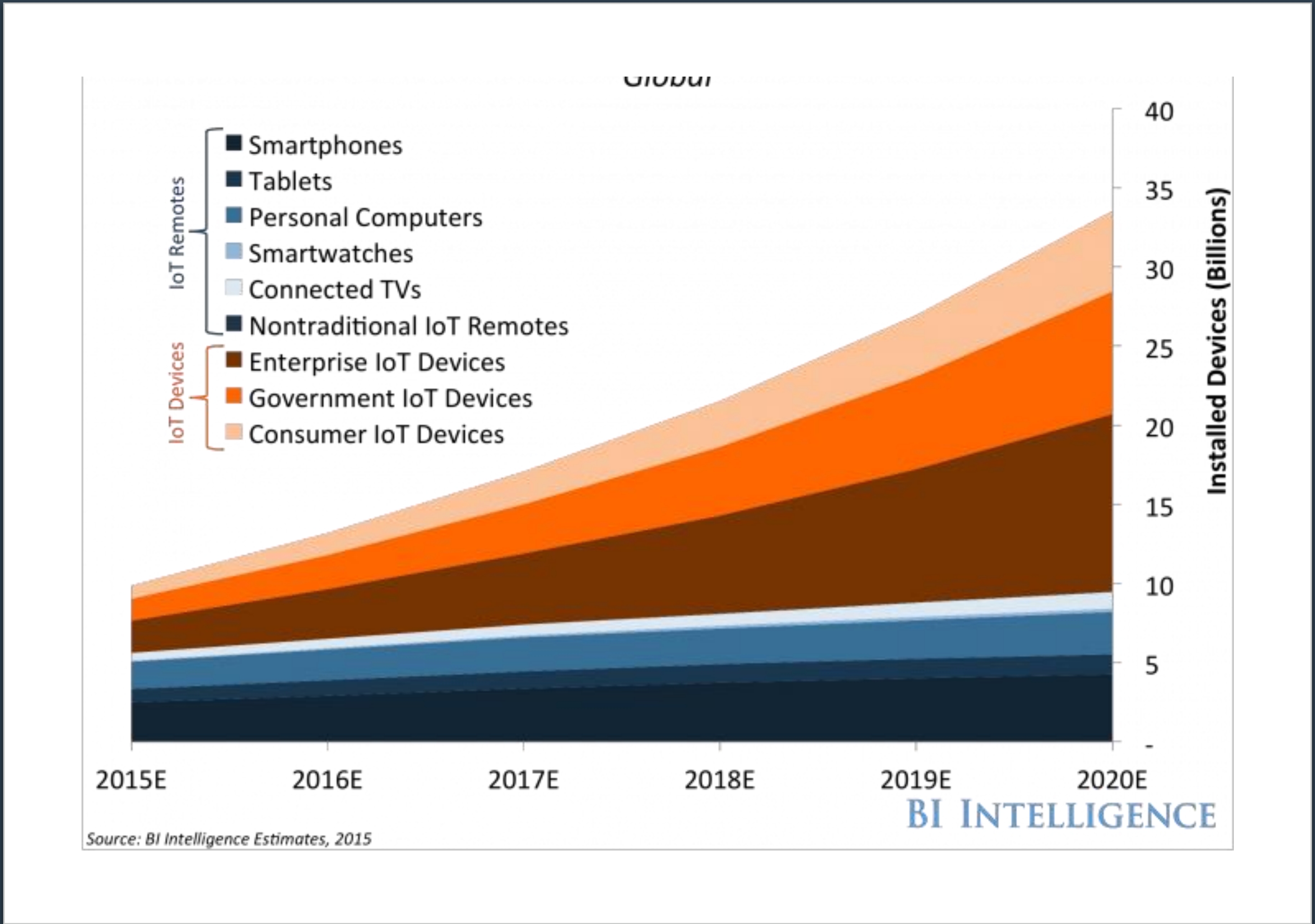
Source: Visual Capitalist

Significant growth in Data Science & Analytics

Explosion of data volume



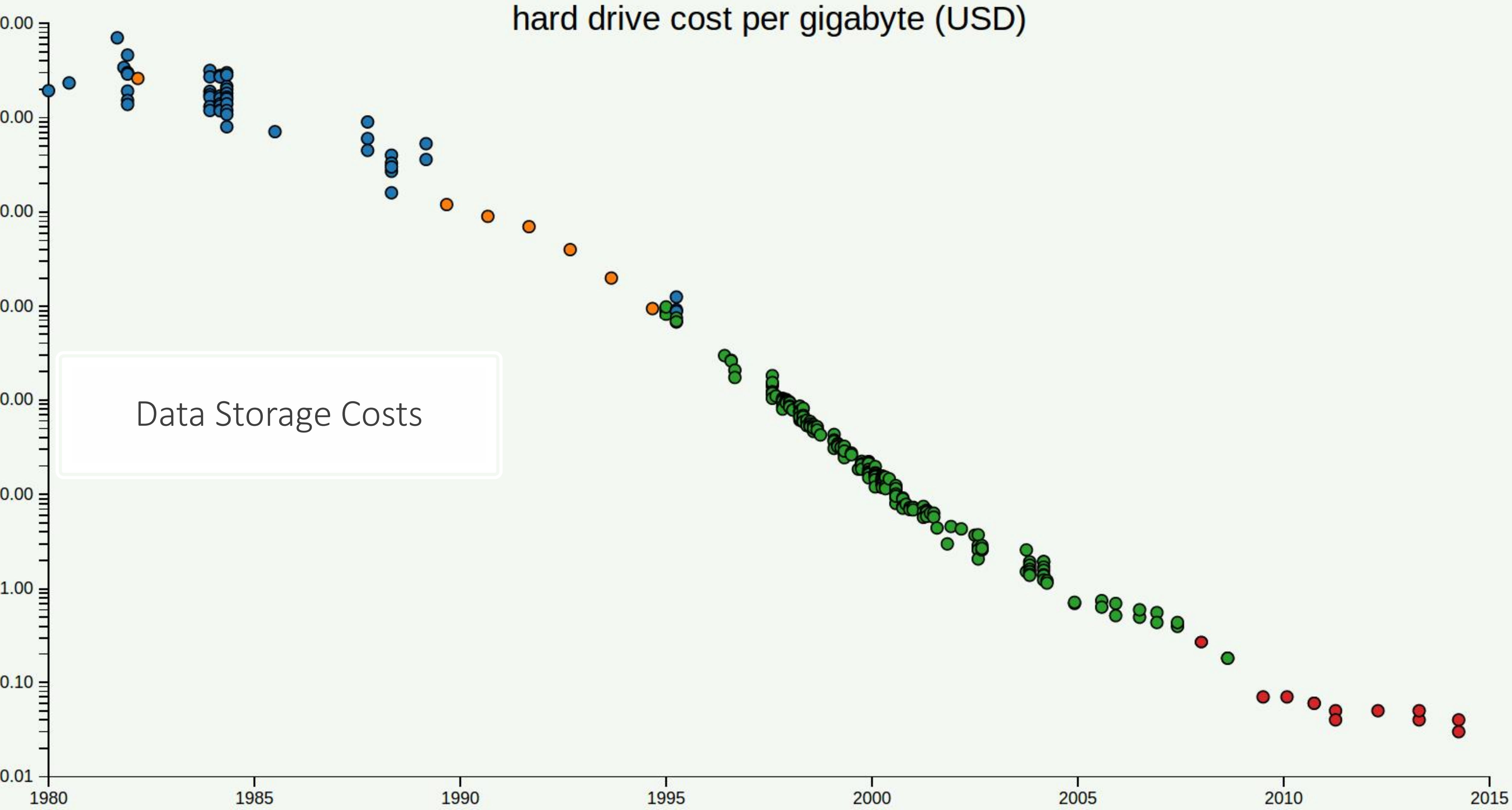
Source: ResearchGate



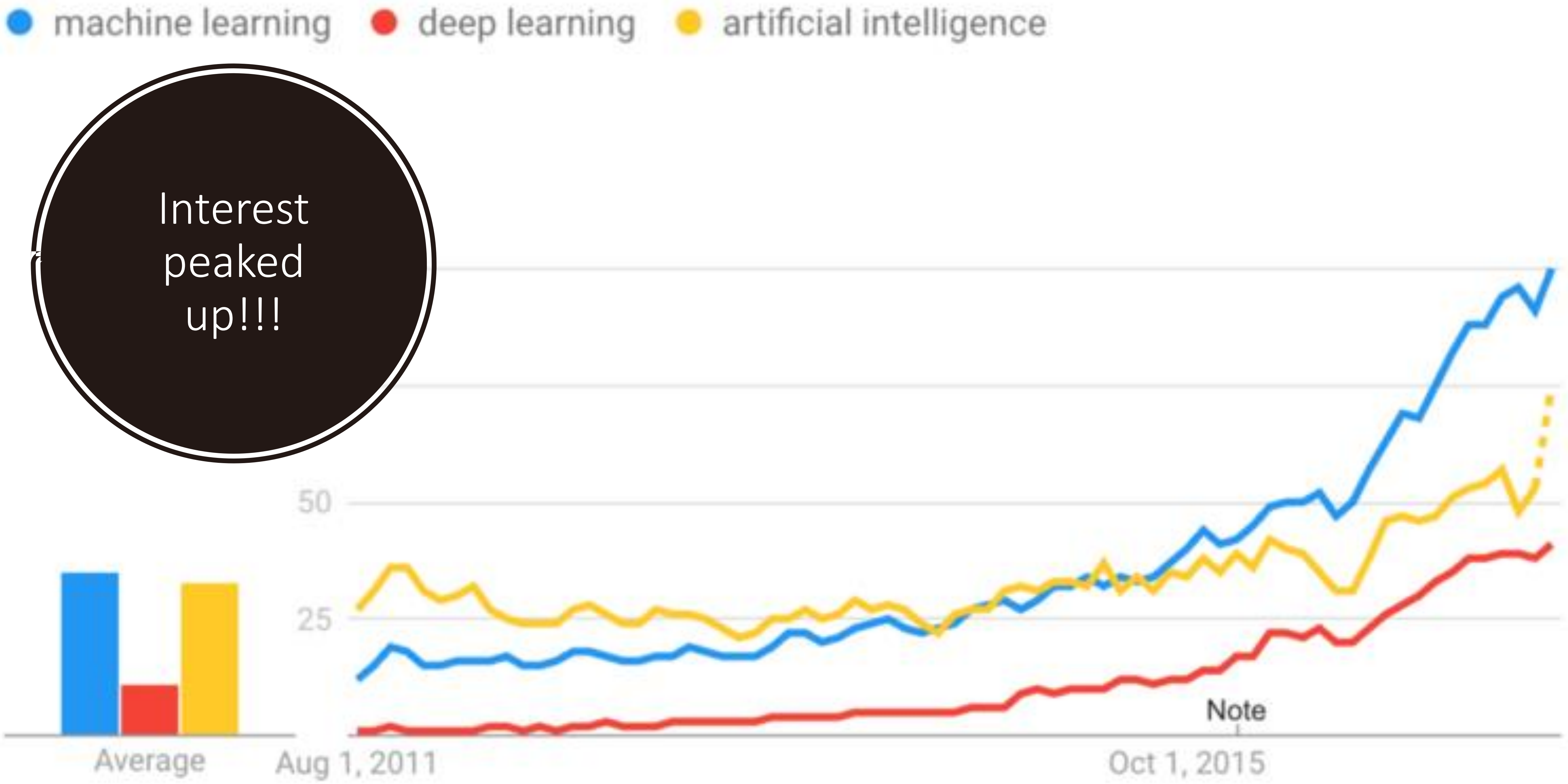
Source: BI Intelligence Estimates, 2015

Devices connected to the internet

Data Storage Costs



source: mkomo.com



Interest peaked up!!!

Worldwide. 7/9/11 - 8/9/17.

“DATA IS THE NEW OIL.”

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

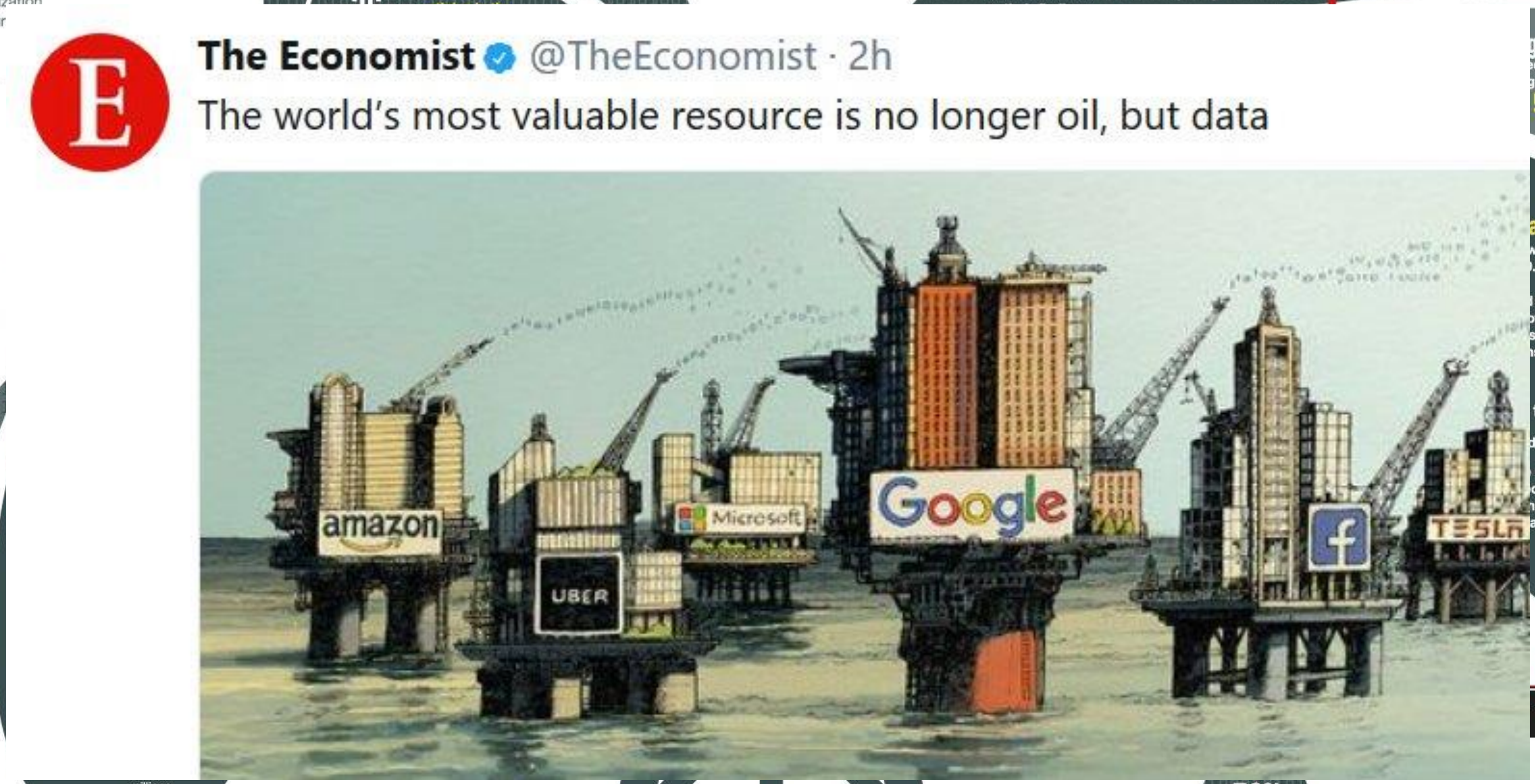
Coined in 2006 by Clive Humby, a British data commercialization entrepreneur this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data to be an economic asset,** like oil.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are **133 million BLOGS** on the web.

As of August 2012, there were just over

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly



The Economist @TheEconomist · 2h

The world's most valuable resource is no longer oil, but data

gorithms
ecisions to buy or sell a commodity.
laid under the Atlantic will shave
5 milliseconds
from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

able,
ween New York
.6 milliseconds.

saving is worth
of dollars to the trading
se the cable (and who will
s to do so).

they save 5 milliseconds

depth of the Atlantic Ocean varies.

new cable will lie on areas of the ocean
or that are up to 1,000 feet shallower
an the current fastest cable. By taking
a different route, the new cable is
shorter, meaning that the time it
takes for messages to travel
along it is shortened.

The new cable takes a shallower, therefore shorter route.

UK

Every hour, we create enough Internet traffic to fill **7 billion DVDs.** Side by side, that's that's seven times the height of Everest.

millions of users

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Information Graphics by Nigel Holmes

Agenda – Part 1

1

Recap of Python

2

Introduction to Machine Learning

3

Types of Machine Learning

4

Unsupervised Learning:

- K-Means Clustering

What is Python?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast.

Main elements of Python

DATA TYPES - BASIC AND ADVANCED

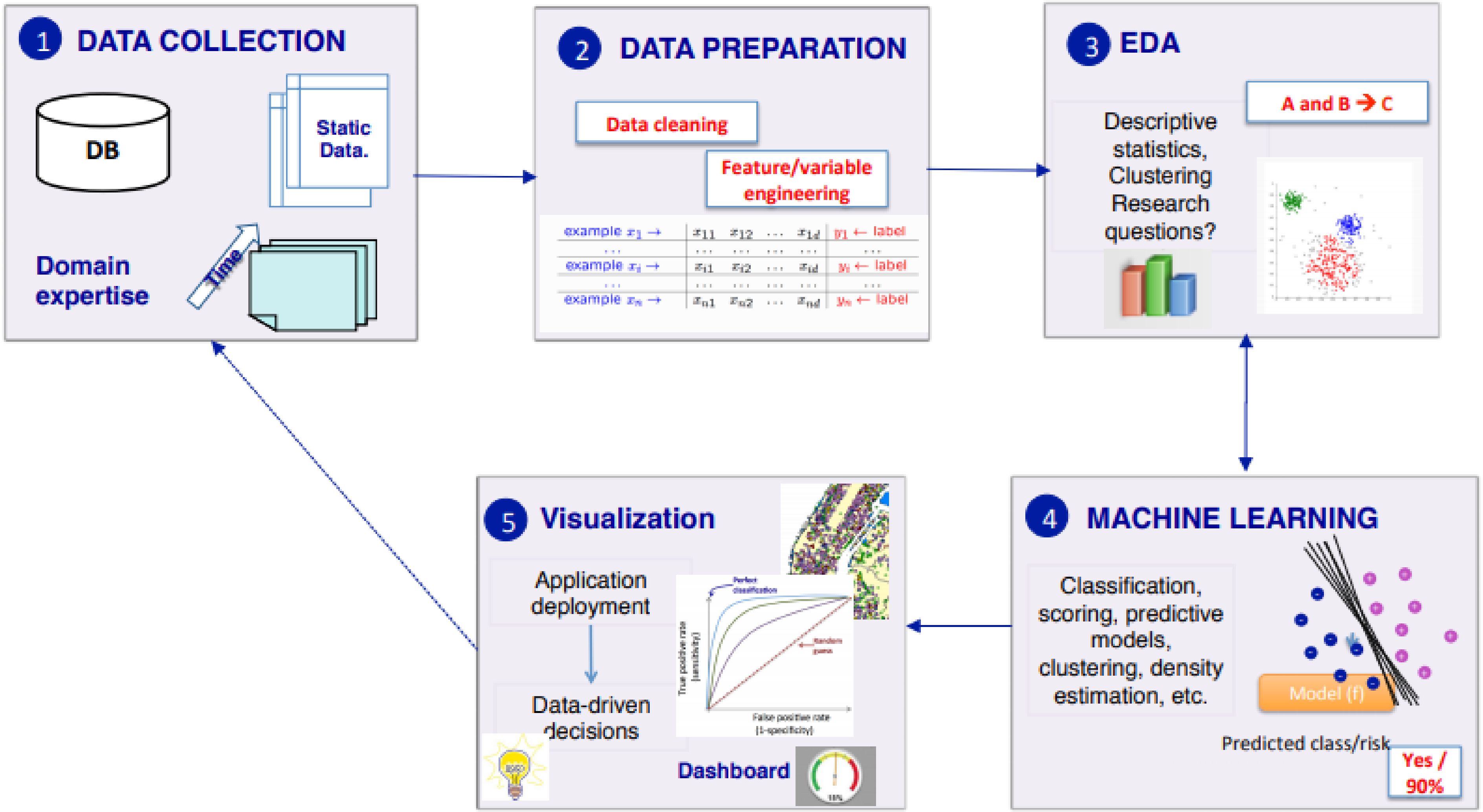
LIBRARIES

FUNCTIONS

FLOW CONTROL

BASIC VISUALIZATIONS

STATISTICAL ANALYSIS



1 DATA COLLECTION

DB

Static Data.

Domain expertise

Time

2 DATA PREPARATION

Data cleaning

Feature/variable engineering

example x_1 →	x_{11}	x_{12}	⋯	x_{1d}	y_1 ← label
⋯	⋯	⋯	⋯	⋯	⋯
example x_i →	x_{i1}	x_{i2}	⋯	x_{id}	y_i ← label
⋯	⋯	⋯	⋯	⋯	⋯
example x_n →	x_{n1}	x_{n2}	⋯	x_{nd}	y_n ← label

3 EDA

A and B → C

Descriptive statistics, Clustering Research questions?

4 MACHINE LEARNING

Classification, scoring, predictive models, clustering, density estimation, etc.

Model (f)

Predicted class/risk

Yes / 90%

5 Visualization

Application deployment

Data-driven decisions

Dashboard

Perfect classification
Random guess
True positive rate (sensitivity)
False positive rate (1-specificity)

Let’s discuss
What is **machine learning?**

What is Learning?

1

“Learning denotes changes in a system that ... enable a system to do the same task ... more efficiently the next time.” - Herbert Simon

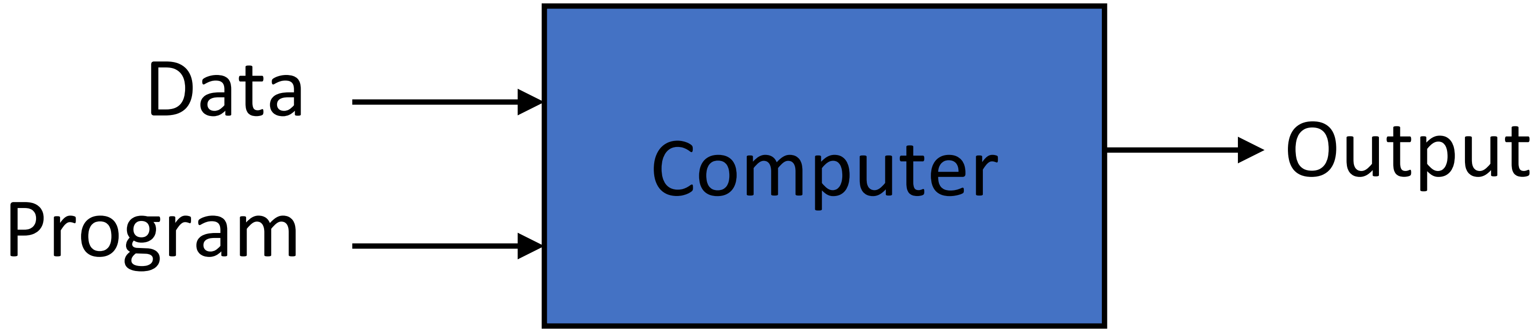
2

“Learning is constructing or modifying representations of what is being experienced.” - Ryszard Michalski

3

“Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge.”

Traditional Programming



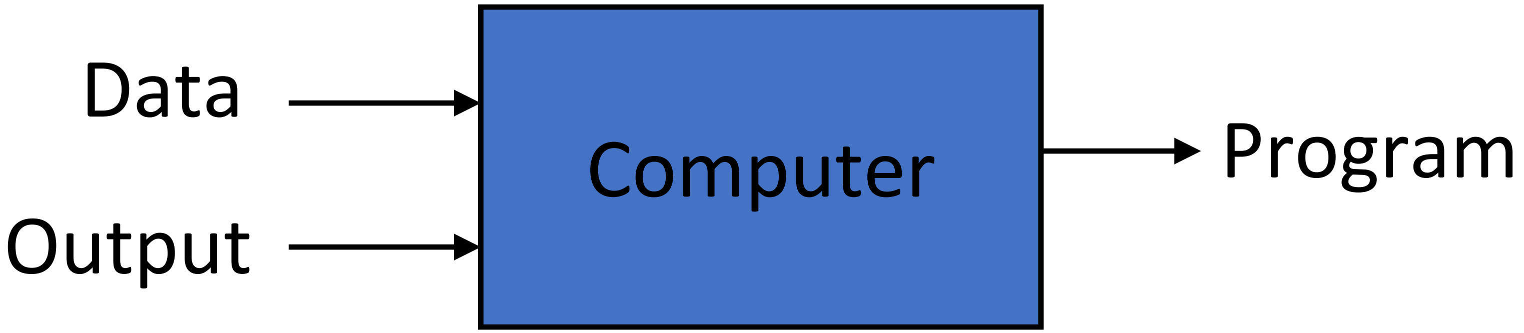
Data

Program

Computer

Output

Machine Learning



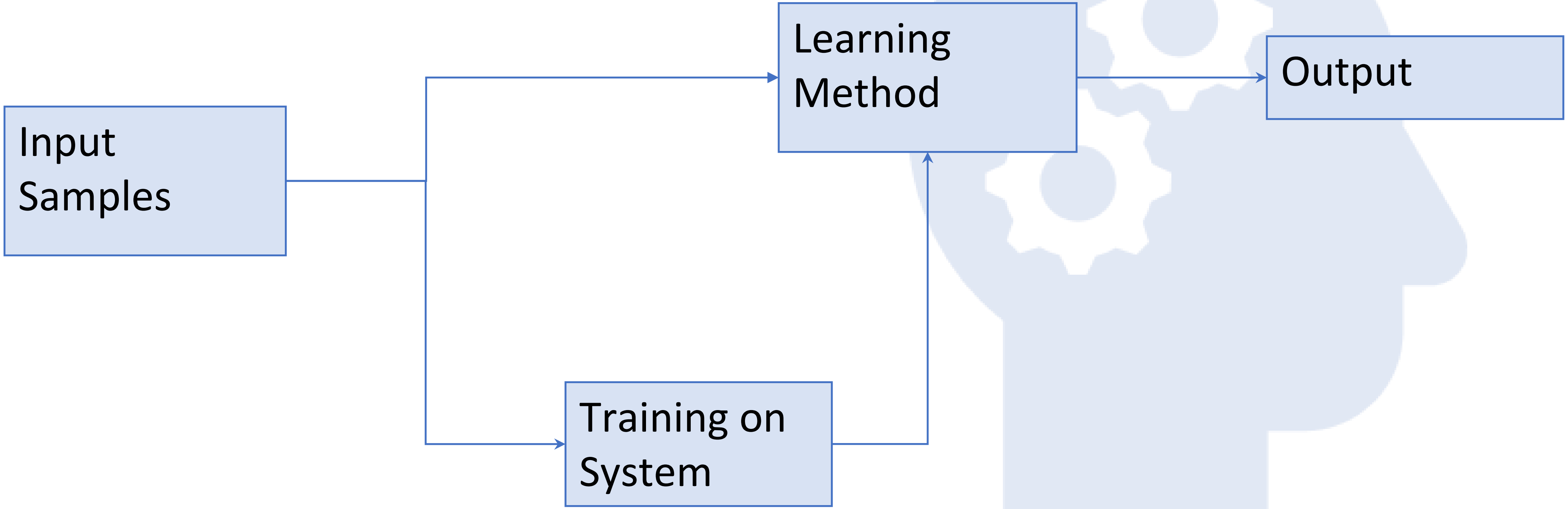
Data

Output

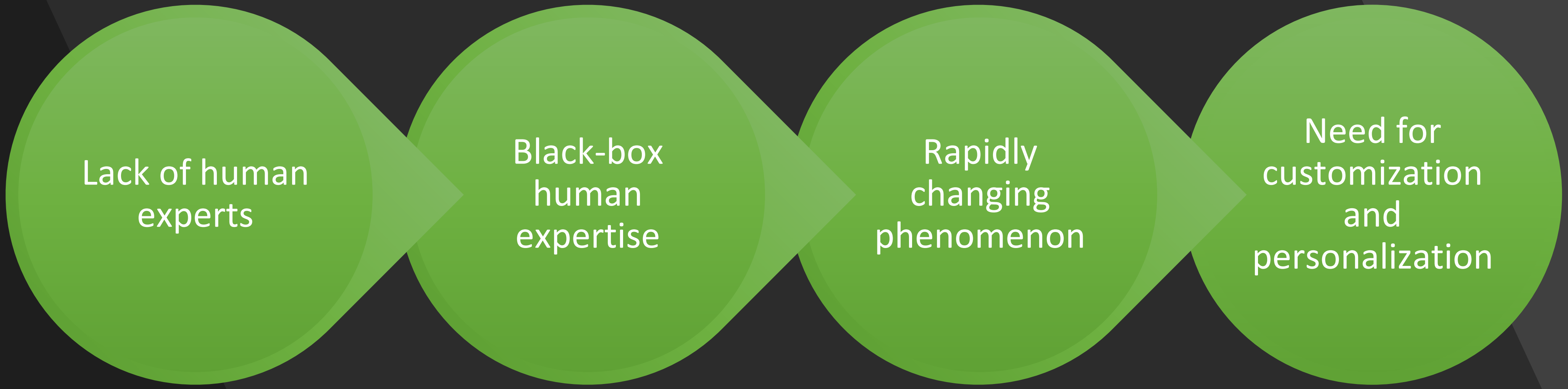
Computer

Program

Learning System Model



Why is machine learning required?



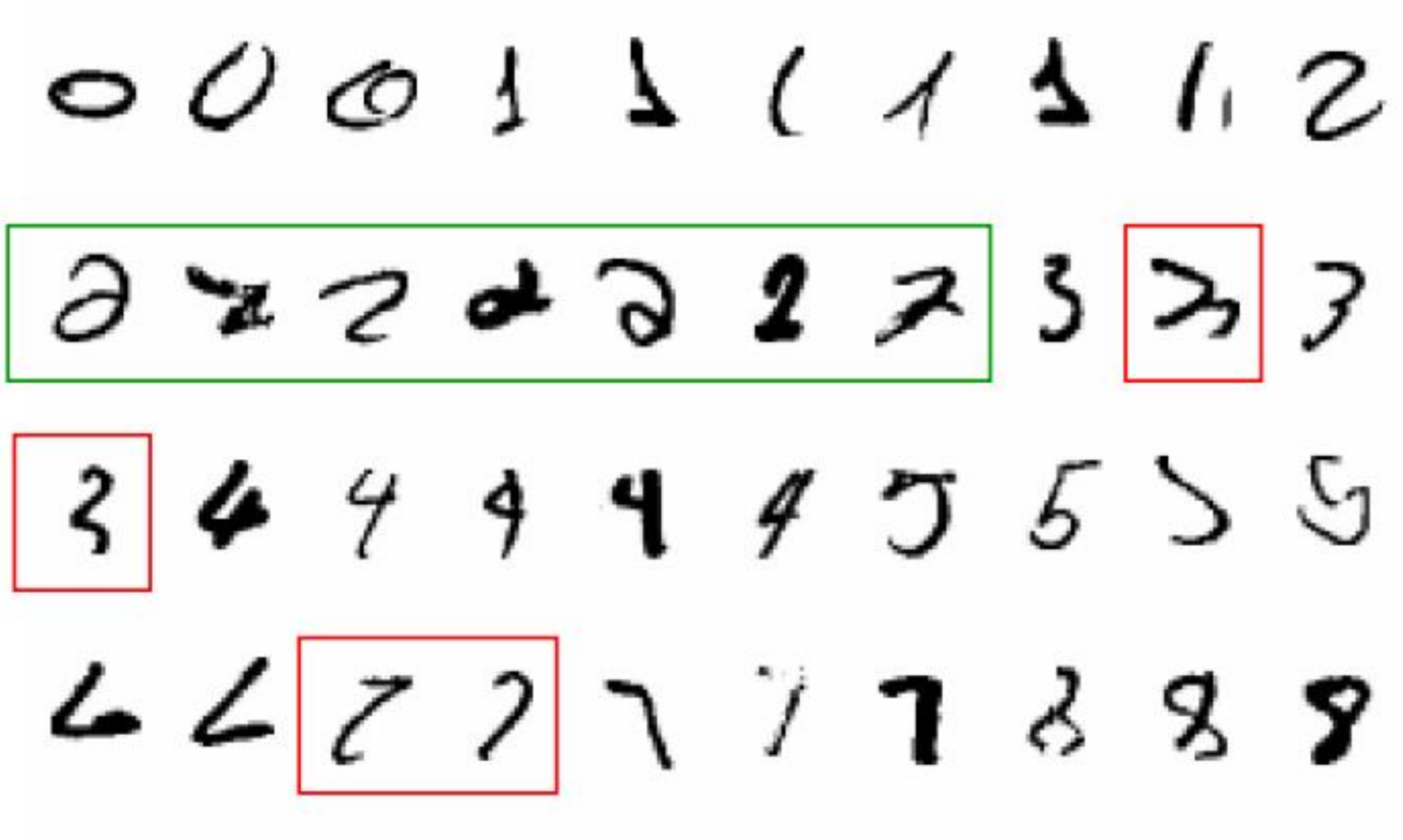
Lack of human experts

Black-box human expertise

Rapidly changing phenomenon

Need for customization and personalization

A classic example of a task that requires machine learning: It is very hard to say what makes a 2



Some examples that machine learning solves

Recognizing patterns:

- Facial identities or facial expressions
- Handwritten or spoken words
- Medical images

Generating patterns:

- Generating images or motion sequences

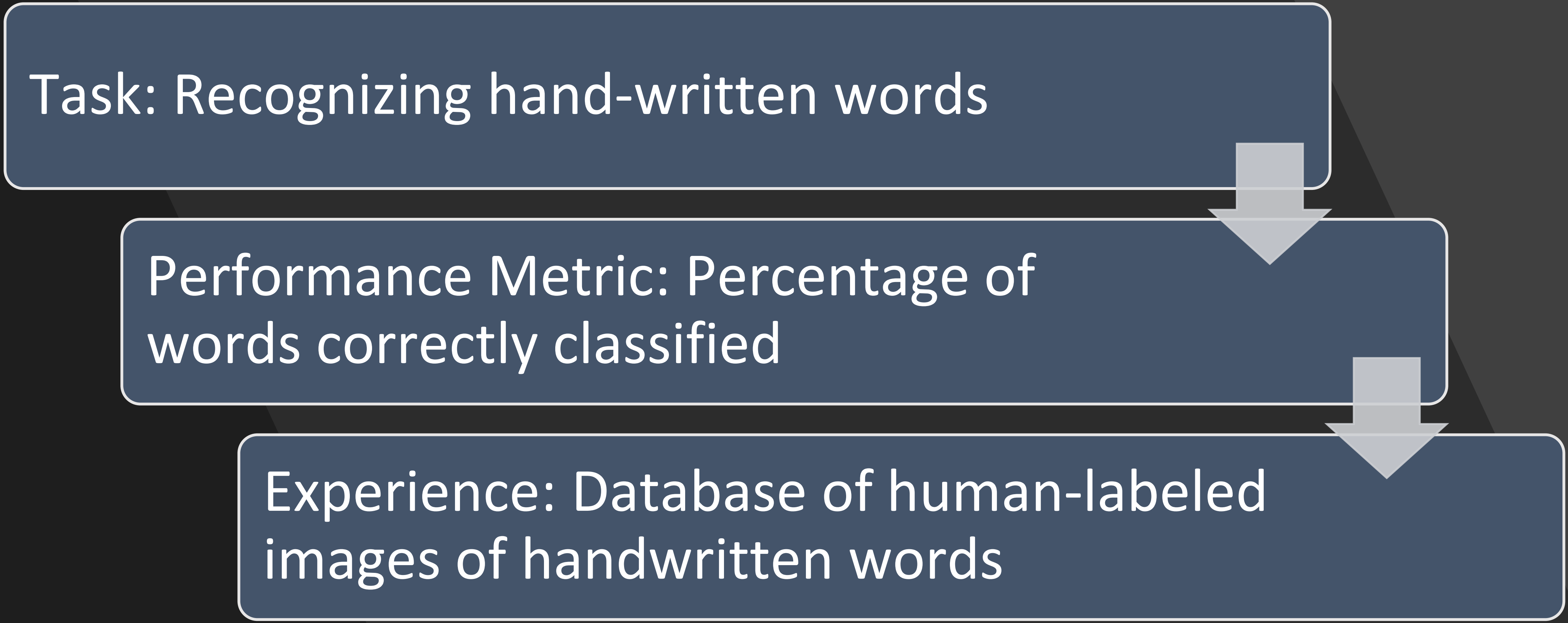
Recognizing anomalies:

- Unusual credit card transactions
- Unusual patterns of sensor readings in a nuclear power plant

Prediction:

- Future stock prices or currency exchange rates

3 vital things to define



Task: Recognizing hand-written words

Performance Metric: Percentage of words correctly classified

Experience: Database of human-labeled images of handwritten words

Types of Learning

Supervised (inductive) learning –

- Given: training data + desired outputs (labels)

Unsupervised learning –

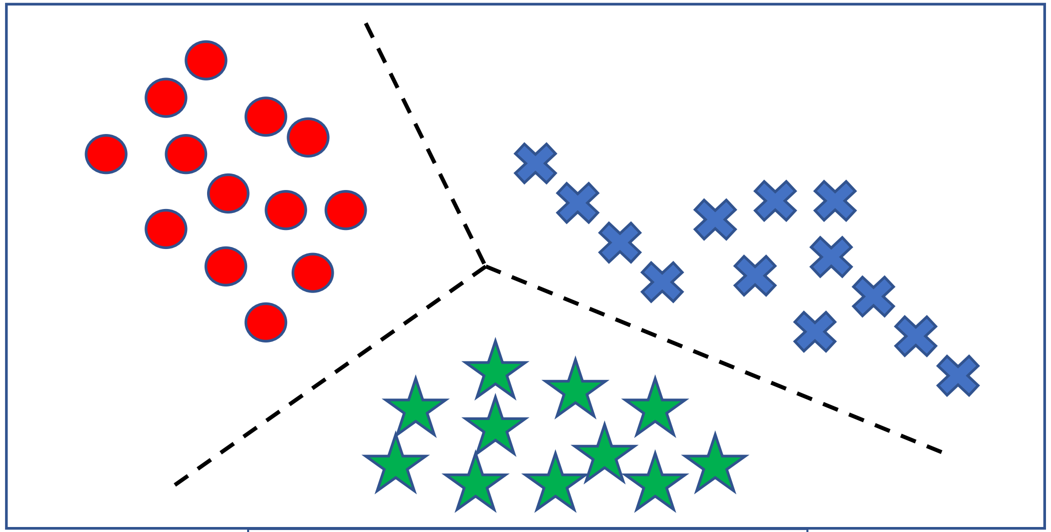
- Given: training data (without desired outputs)

Semi-supervised learning –

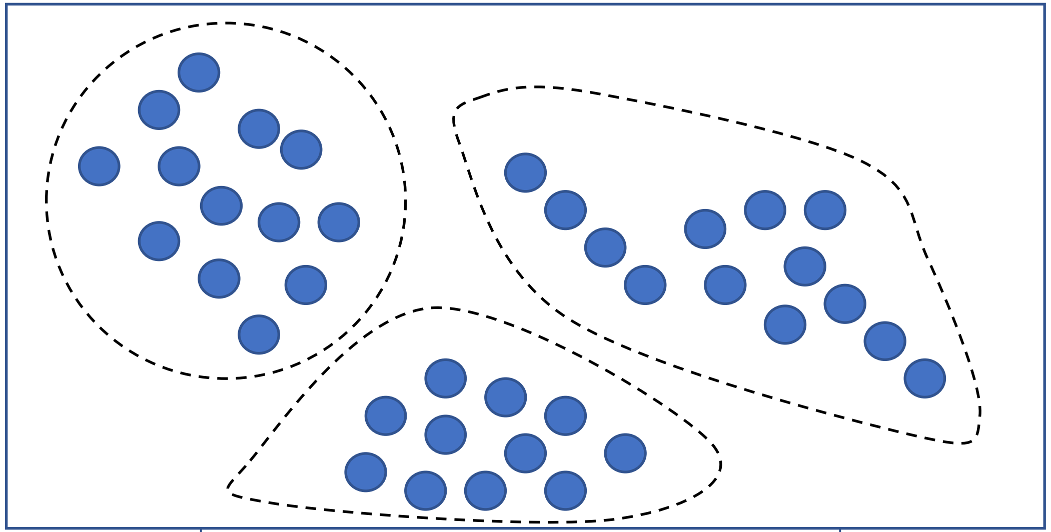
- Given: training data + a few desired outputs

Reinforcement learning –

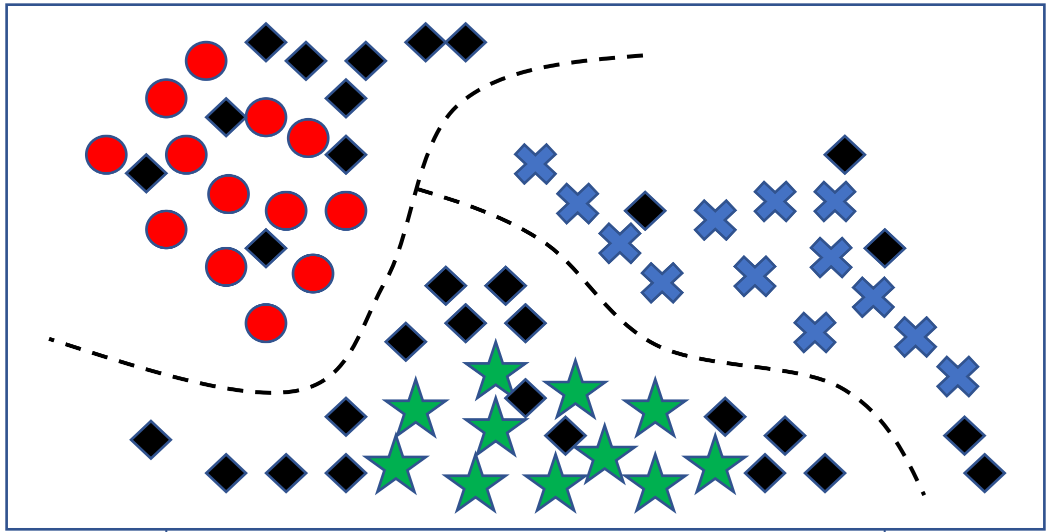
- Rewards from sequence of actions



Supervised learning

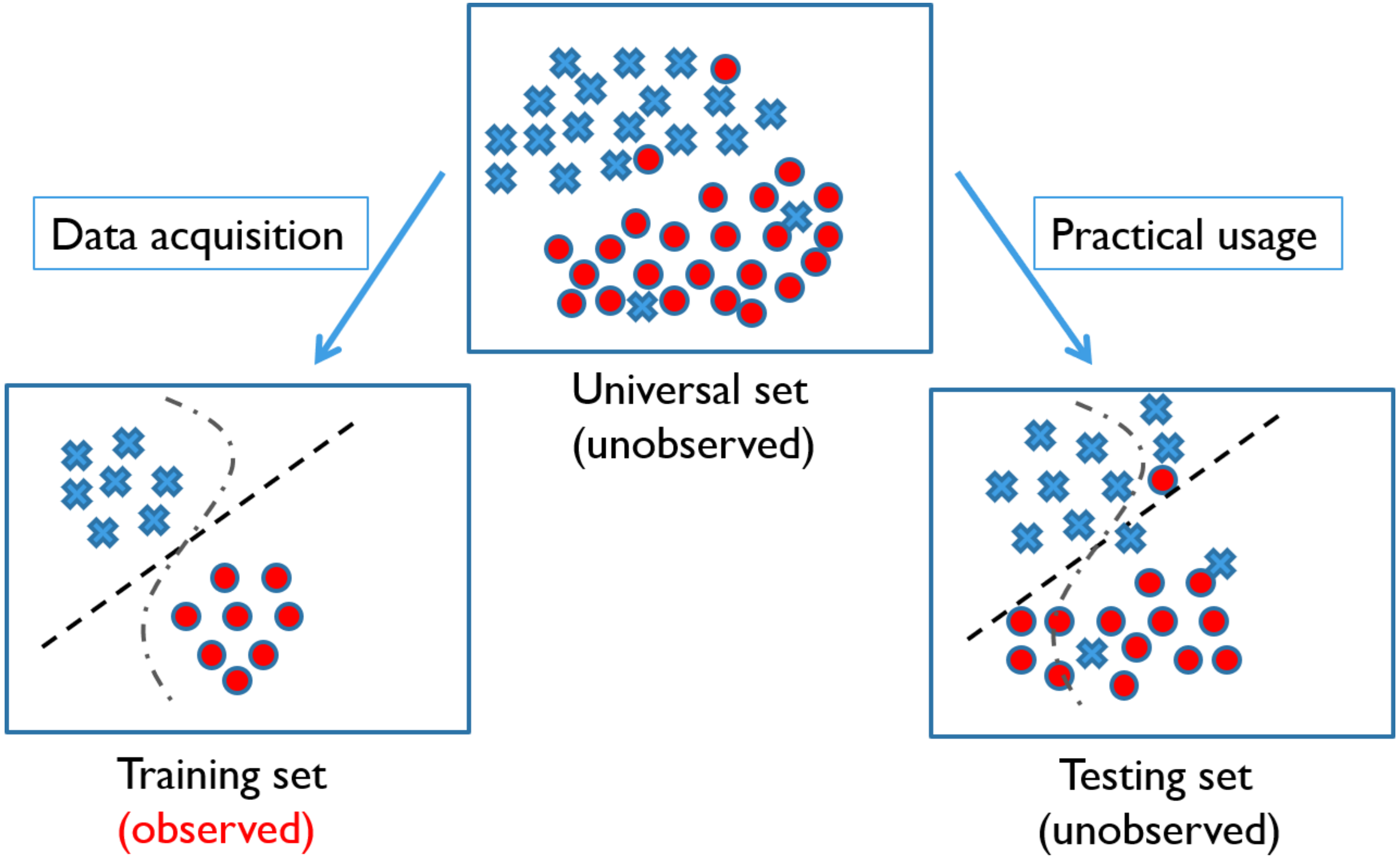


Unsupervised learning



Semi-supervised learning

Training and Test Sets



Unsupervised Learning

- The data has no target attribute.
- We want to explore the data to find some intrinsic structures in them.

What is Clustering?

Clustering

Clustering is a technique for finding similarity groups in data, called **clusters**. I.e.,

- It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.

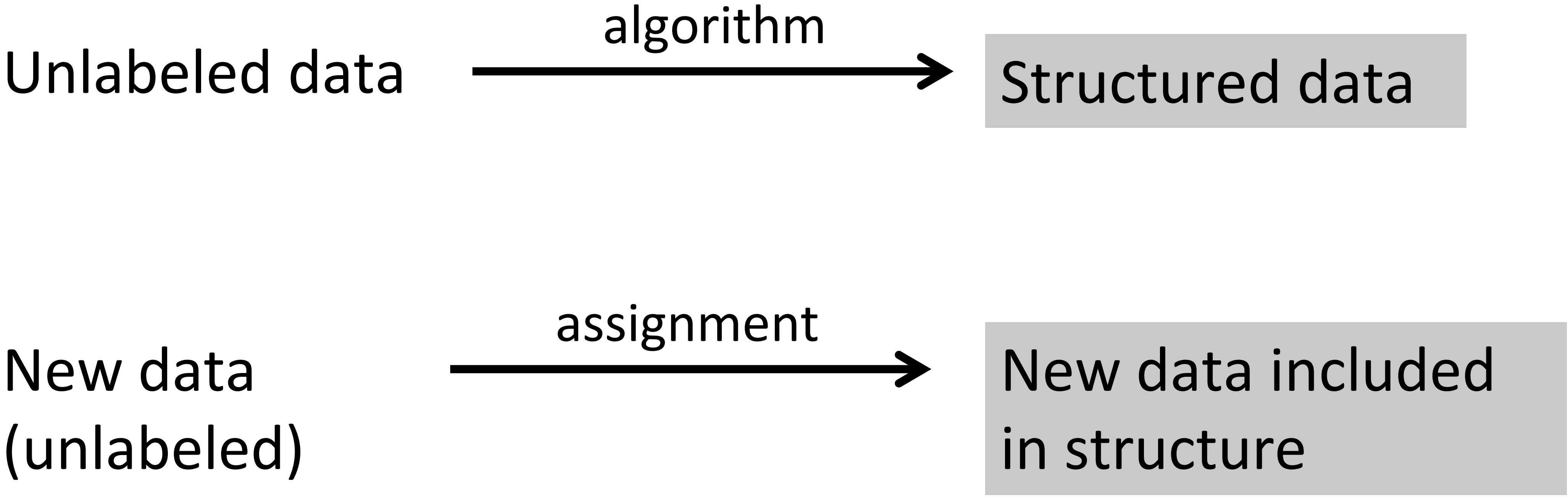
What’s a cluster?

Intuitive definition: Grouping of data points that are close to each other

To make this computer friendly, need a mathematical definition of “close.”

Closeness (most common definitions): based on distance or density

Clustering as unsupervised learning



Unlabeled data

algorithm

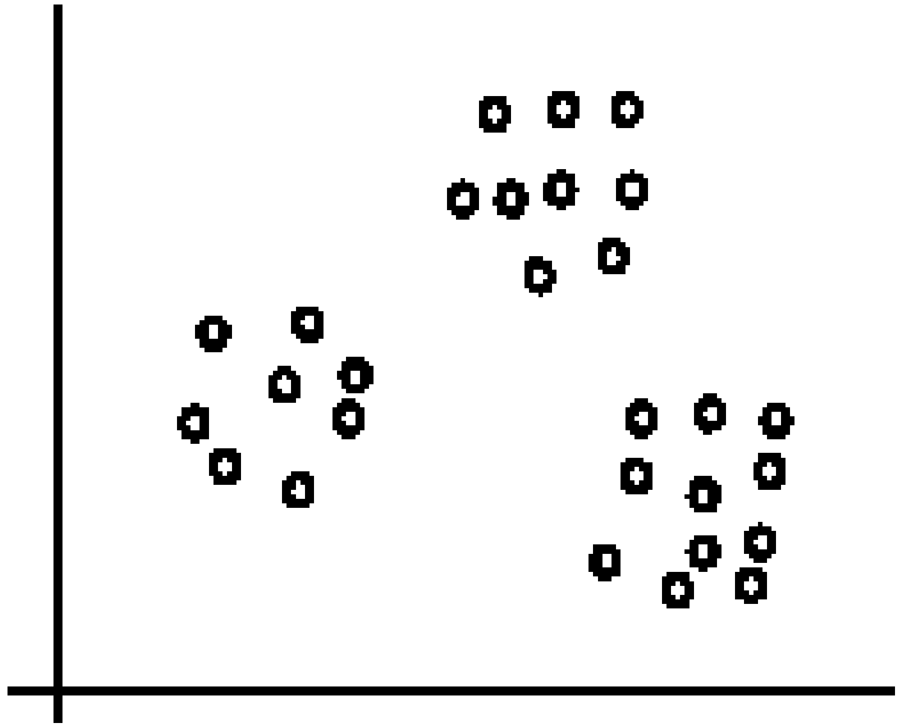
Structured data

New data (unlabeled)

assignment

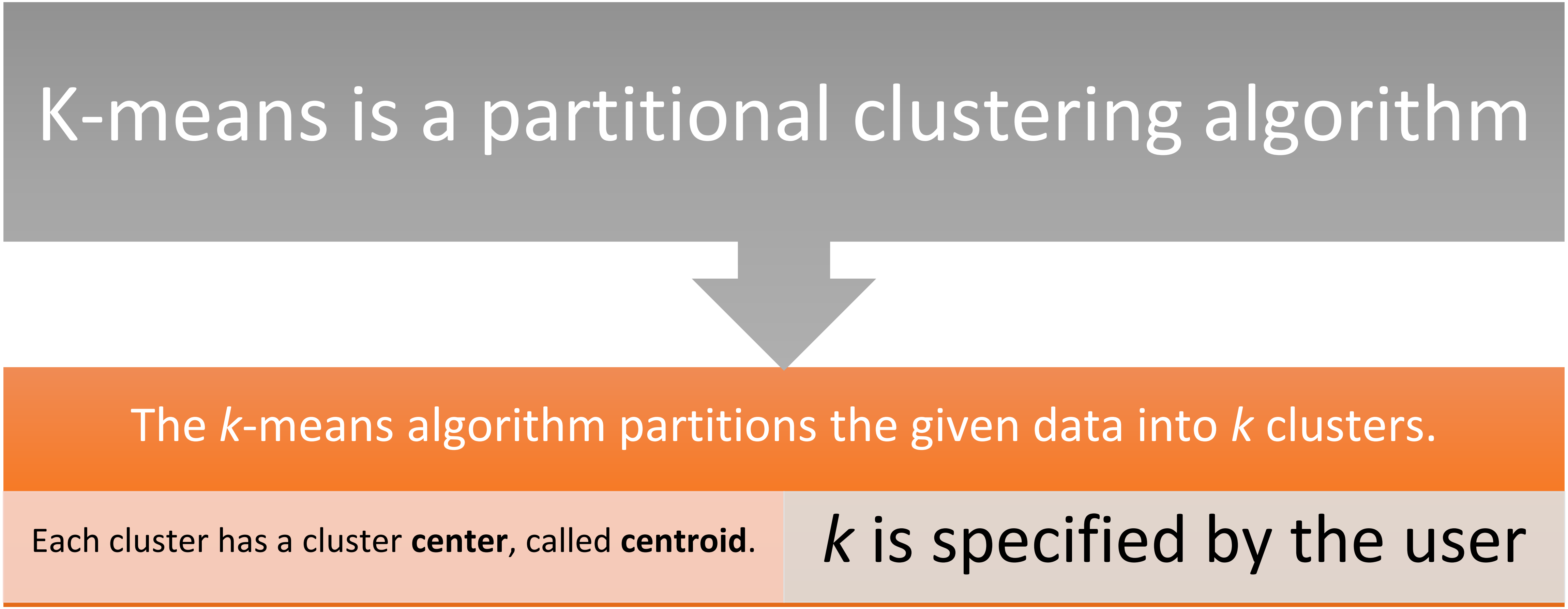
New data included in structure

Think of it like this – In layman figures



A Clustering Technique

K-Means Algorithm



K-means is a partitional clustering algorithm

The *k*-means algorithm partitions the given data into *k* clusters.

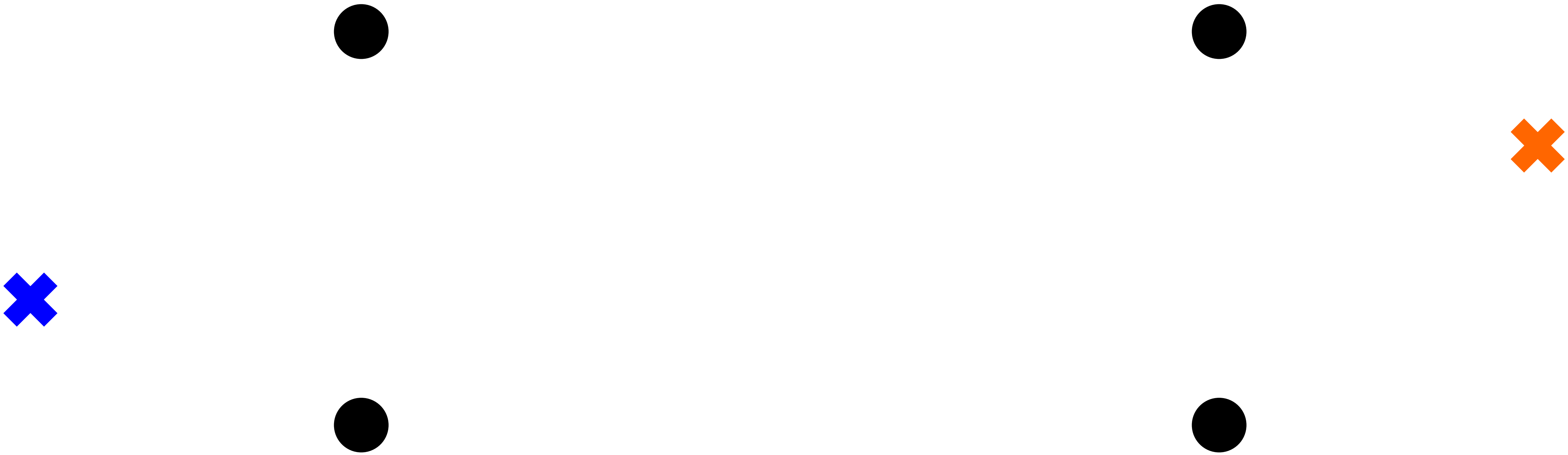
Each cluster has a cluster **center**, called **centroid**.

k is specified by the user

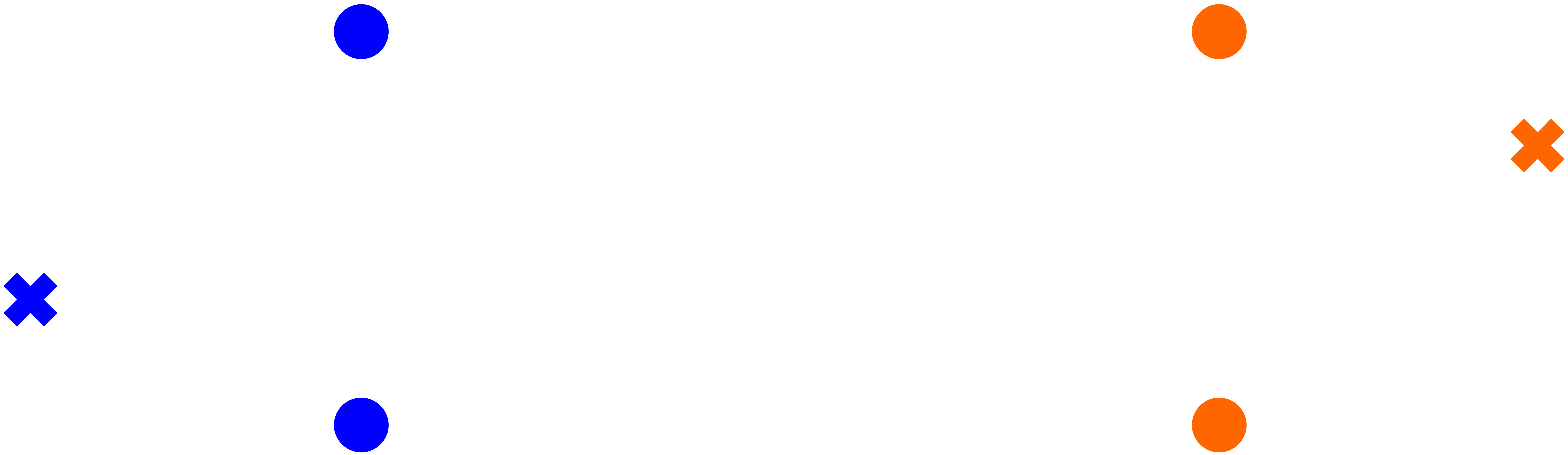
k-means clustering: the **algorithm**

- Choose *k* centroids
- Assign points to cluster based on nearest centroid
- Recompute centroids
- Repeat steps (2) and (3) until there is no more change to the centroids

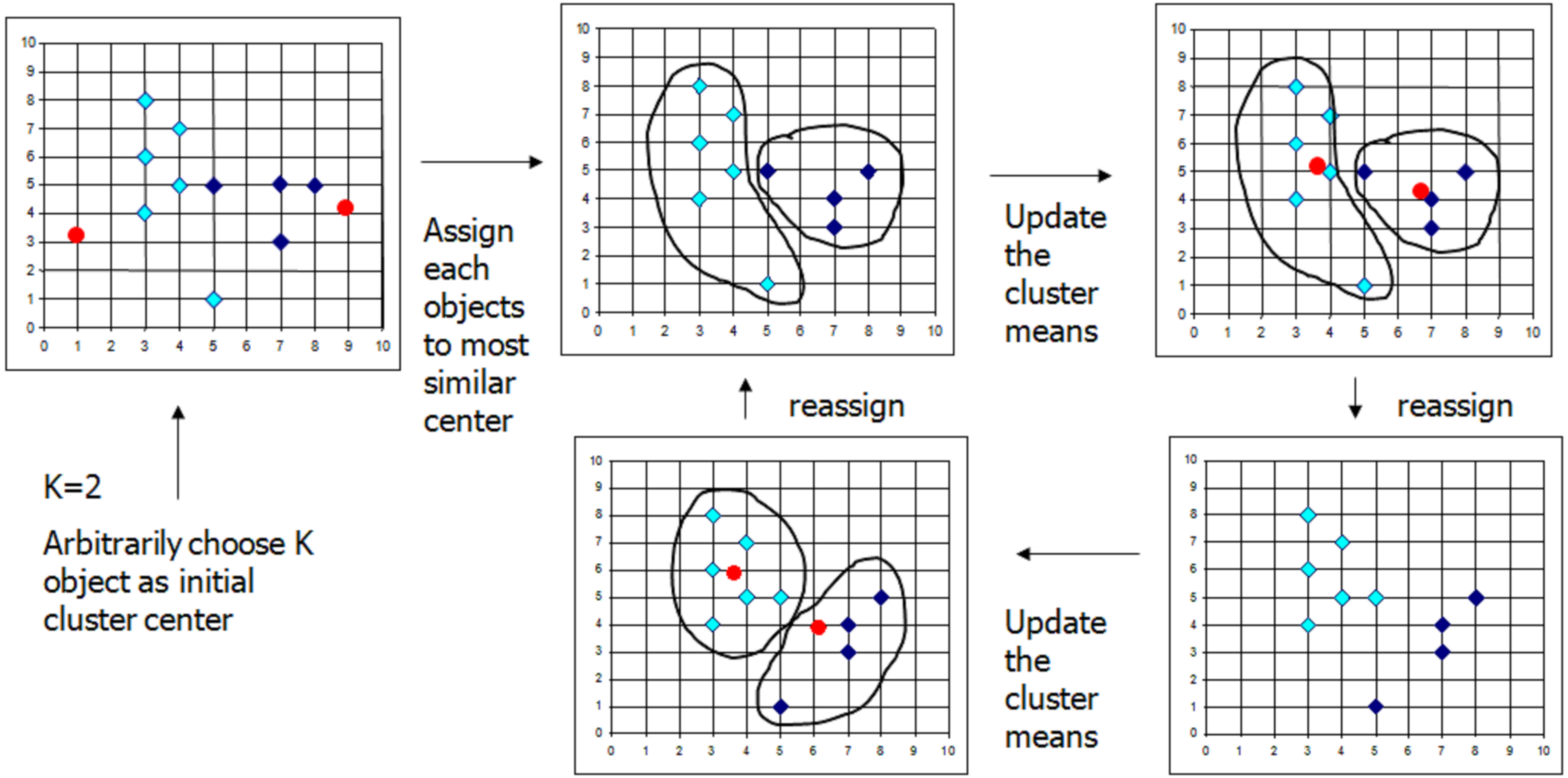
k-means: simple example



k-means: simple example



k-means: simple example



K=2

Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

Update the cluster means

reassign

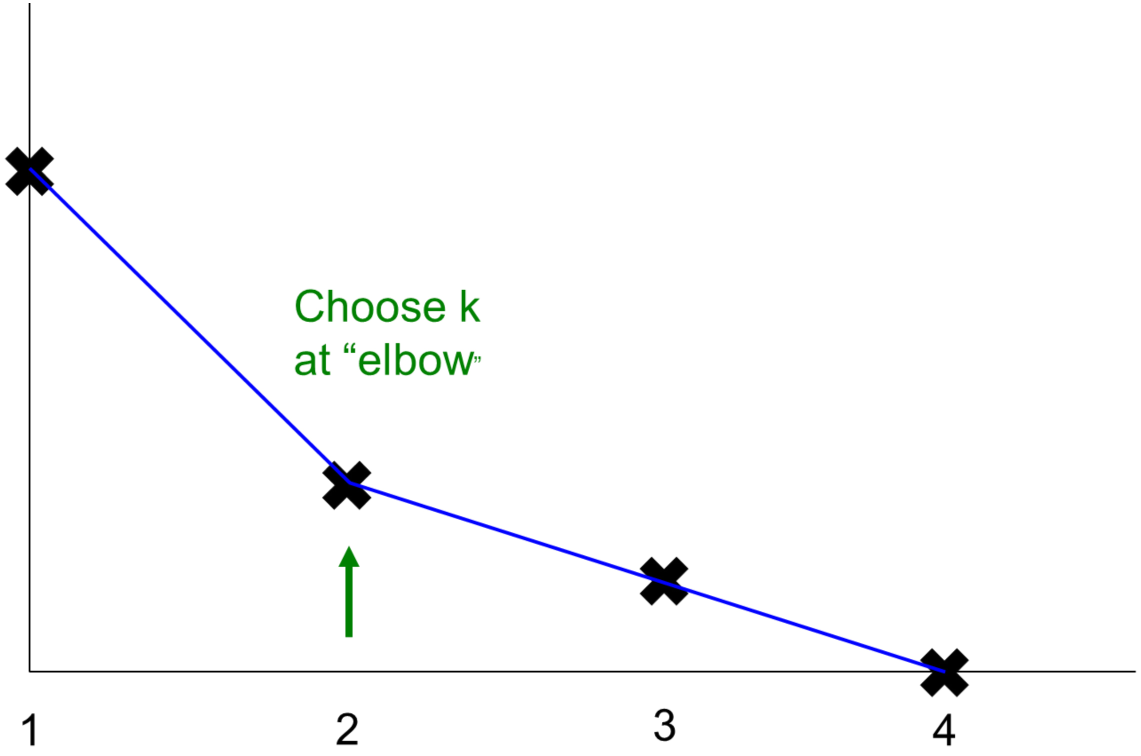
reassign

Update the cluster means

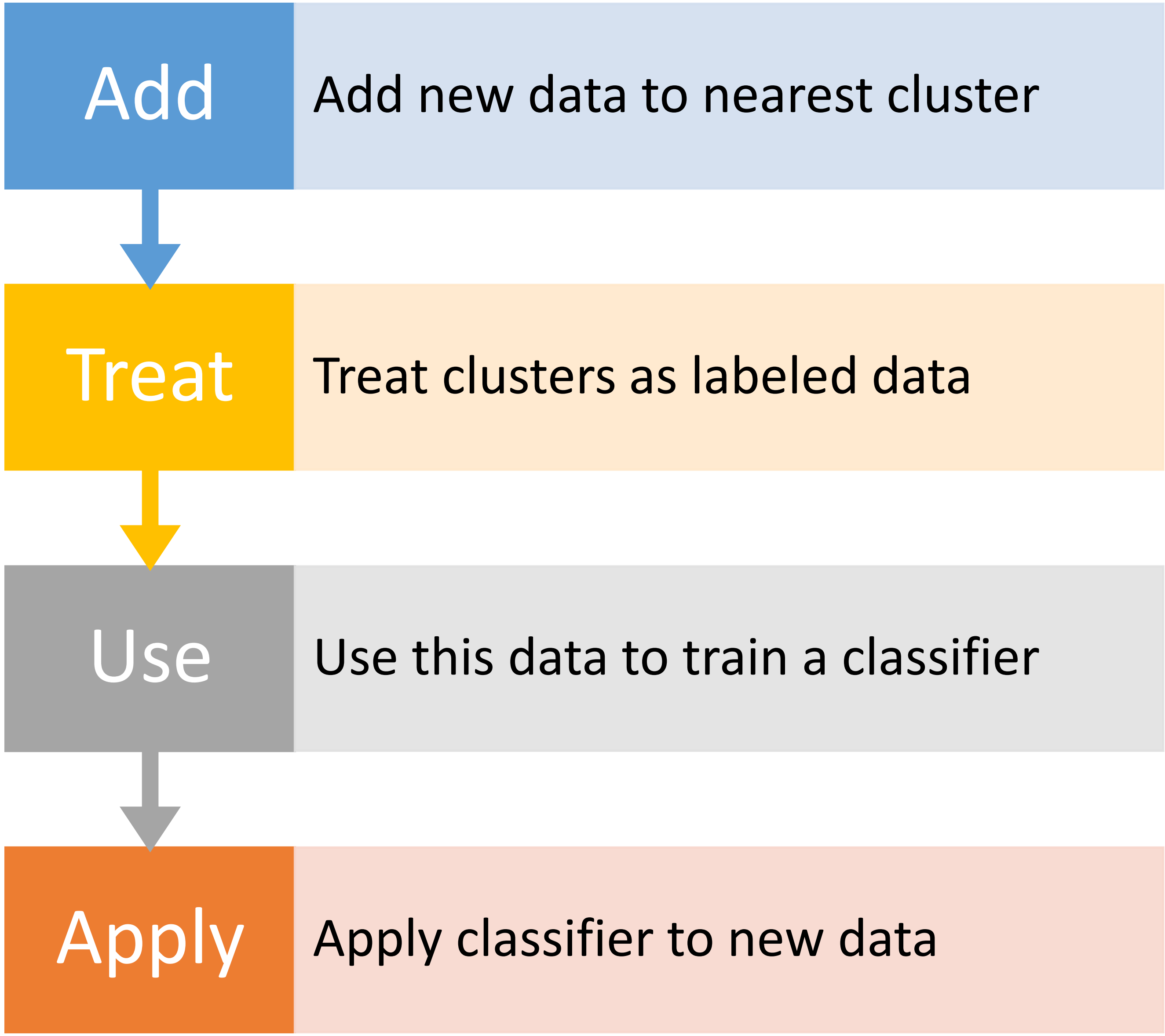
k-means performance

good clustering → points close to cluster centroids

k-means performance



k-means: adding new data



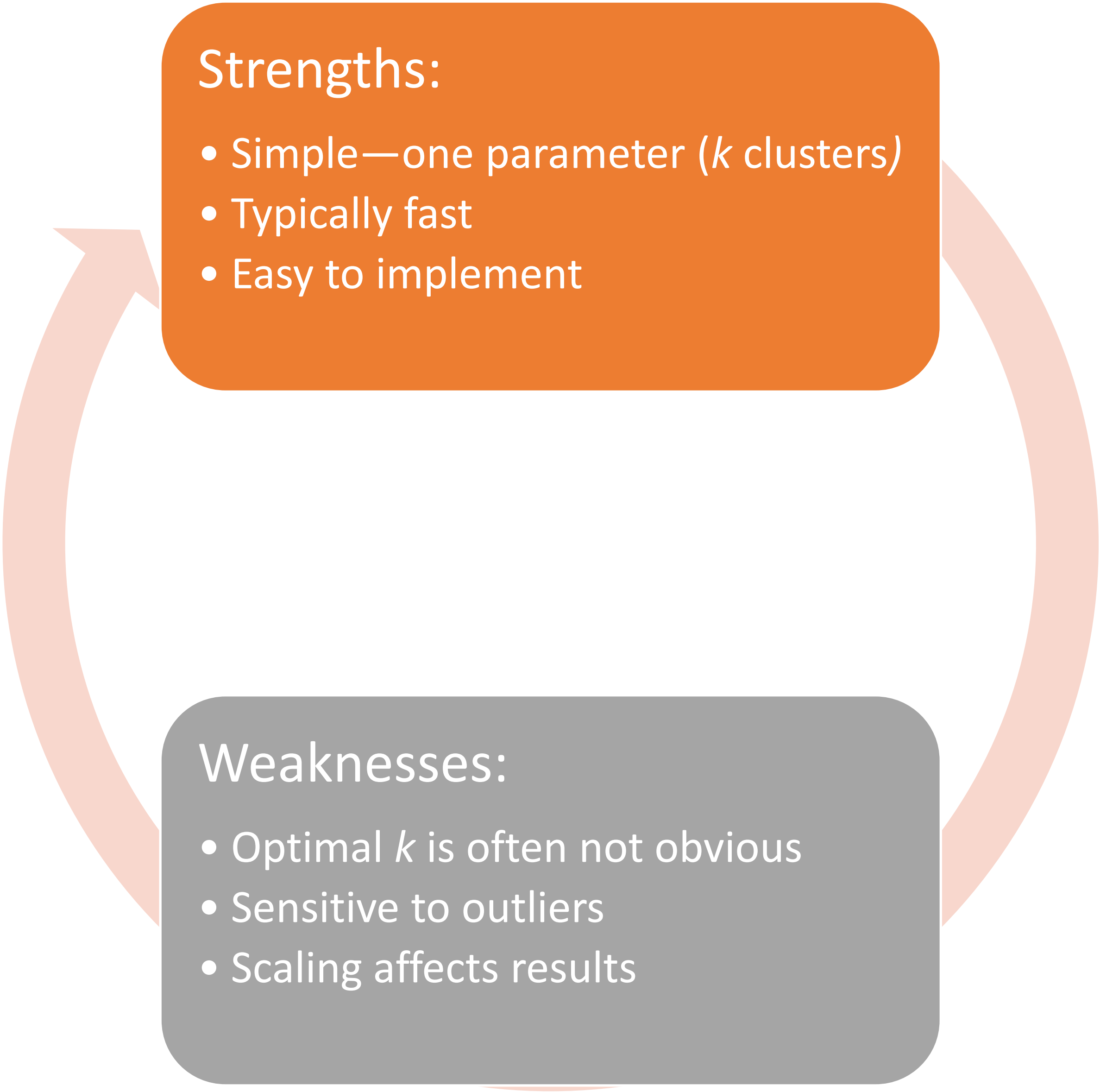
Add — Add new data to nearest cluster

Treat — Treat clusters as labeled data

Use — Use this data to train a classifier

Apply — Apply classifier to new data

k-means: strengths and weaknesses



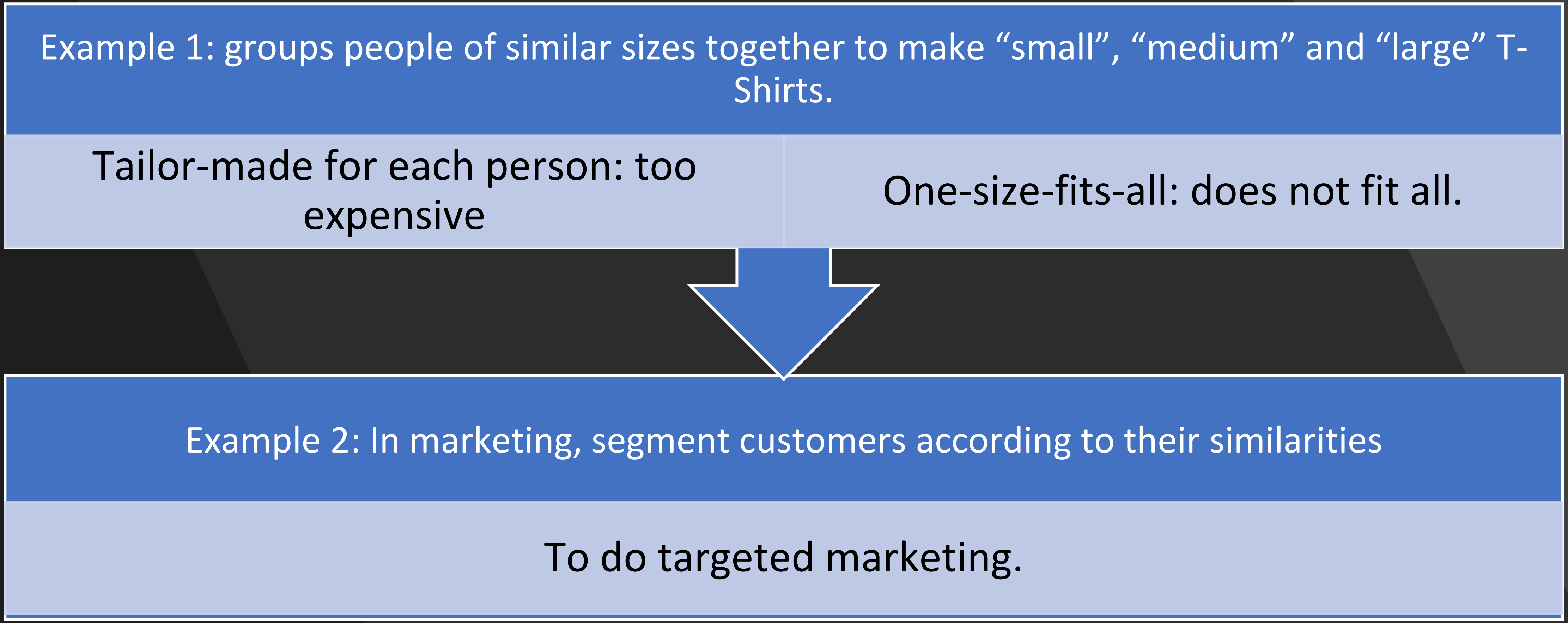
Strengths:

- Simple—one parameter (*k* clusters)
- Typically fast
- Easy to implement

Weaknesses:

- Optimal *k* is often not obvious
- Sensitive to outliers
- Scaling affects results

Clustering - Real life Examples



Example 1: groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

Tailor-made for each person: too expensive

One-size-fits-all: does not fit all.

Example 2: In marketing, segment customers according to their similarities

To do targeted marketing.

Let’s dive straight to the Hands-on using Jupyter notebooks

Other clustering algorithms

Self Organizing Maps (SOM)

Agglomerative Hierarchical Clustering

Agenda – Part 2

1

Supervised Machine Learning

- Linear Regression

2

Feature Engineering

Supervised Learning

Data includes both the input and the desired results.

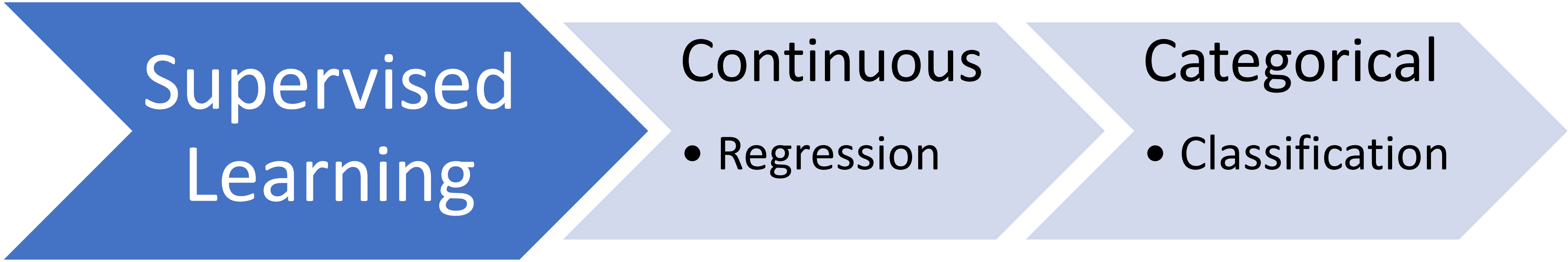
Think of the following examples.

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- **A decision is needed: whether to put a new patient in an intensive-care unit.**
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- **Problem**: to predict high-risk patients and discriminate them from low-risk patients.

Another example..

- A credit card company receives lots of applications for new cards. Each application contains information about the applicant for the card,
  - age
  - Marital status
  - annual salary
  - location
  - outstanding debts
  - credit rating
  - Family information etc
- **Problem**: to decide whether an application should be approved or not approved.

Types of Data vs Algorithm



Supervised Learning

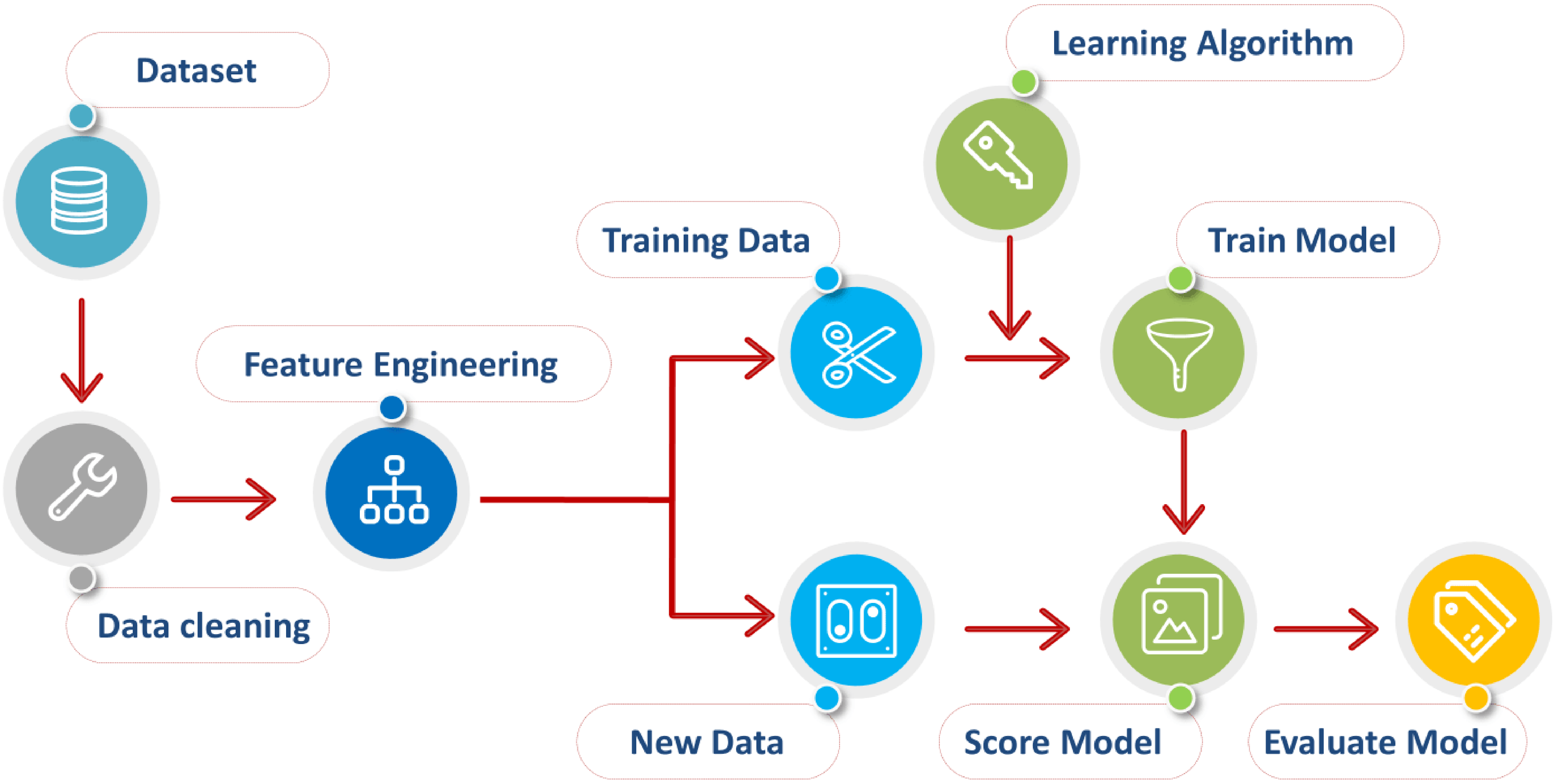
Continuous

- Regression

Categorical

- Classification

General Machine Learning Process



Dataset

Data cleaning

Feature Engineering

Training Data

New Data

Learning Algorithm

Train Model

Score Model

Evaluate Model

Jargons to be aware of!

Model Inputs: Features, Attributes, Predictors, Inputs, Independent Variables, Dimensions, probably more.

Model Outputs (what we’re trying to predict): Target, Response, Output, Dependent Variable, Labels

Row of Data (Inputs + Outputs): Observation, Datapoint, Record, Row

Labels: The values on the target variables in Supervised Learning

Feature Engineering

What is it all about?

Feature engineering

- The first thing we need to do when creating a machine learning model is to decide what to use as features.
- **Features** are key to a model, like a person's name or favorite color. pieces of information that we take from the text and give to the algorithm so it can work its magic.
- E.g, if we do classification on health, some features could be a person's height, weight, gender, and so on.
    - We would exclude things that maybe are known but aren't useful

Benefits of Feature Engineering

- **Reduces Overfitting** : Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy :** Less misleading data means modeling accuracy improves.
- **Reduces Training Time :** Fewer data points reduce algorithm complexity and algorithms train faster.

Techniques of Feature Engineering

- Introducing polynomial terms
- Introducing interaction terms

Linear Regression

Getting our line straight!

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

- Dependent variable:

The variable we wish to predict or explain

- Independent variable:

The variable used to explain the dependent variable

Simple Linear Regression Model

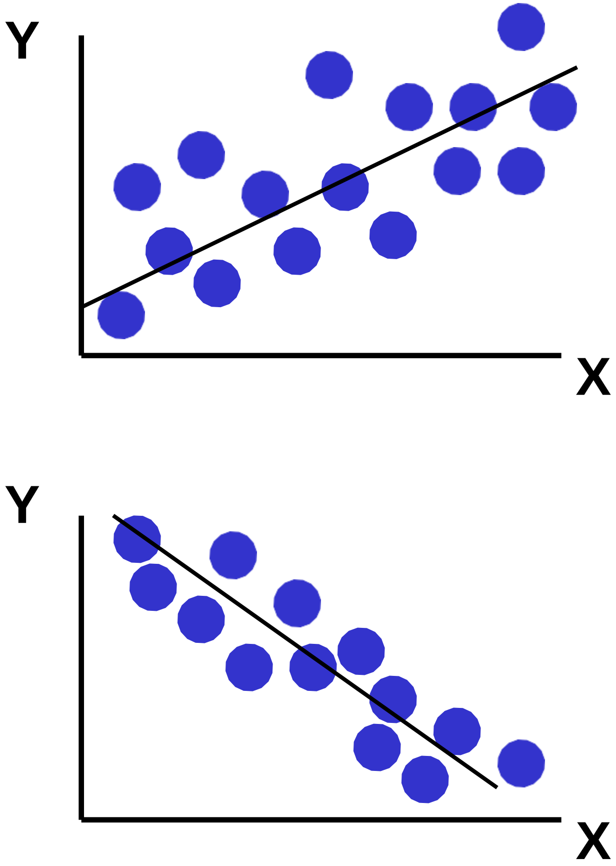
Only **one** independent variable, X

Relationship between X and Y is described by a linear function.

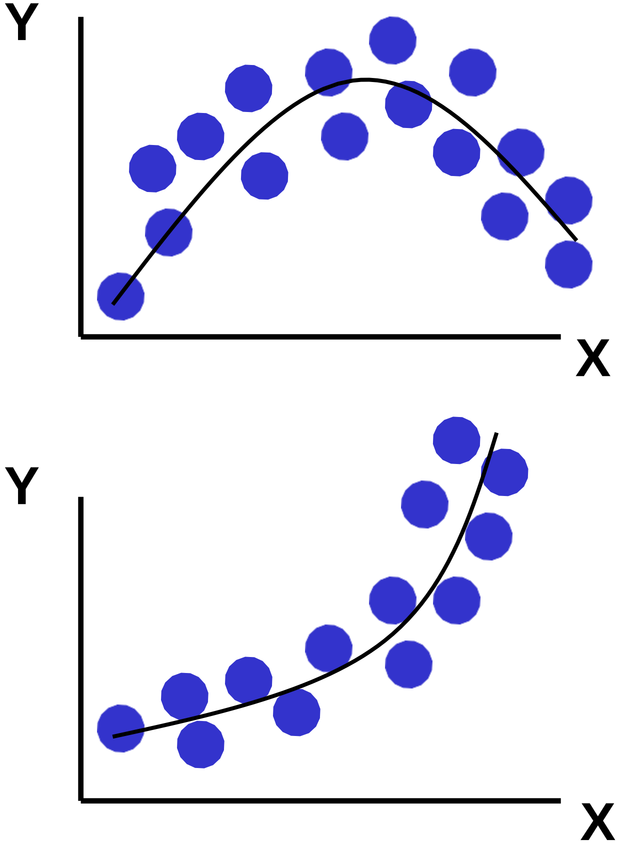
Changes in Y are assumed to be caused by changes in X

Types of Relationships

Linear relationships

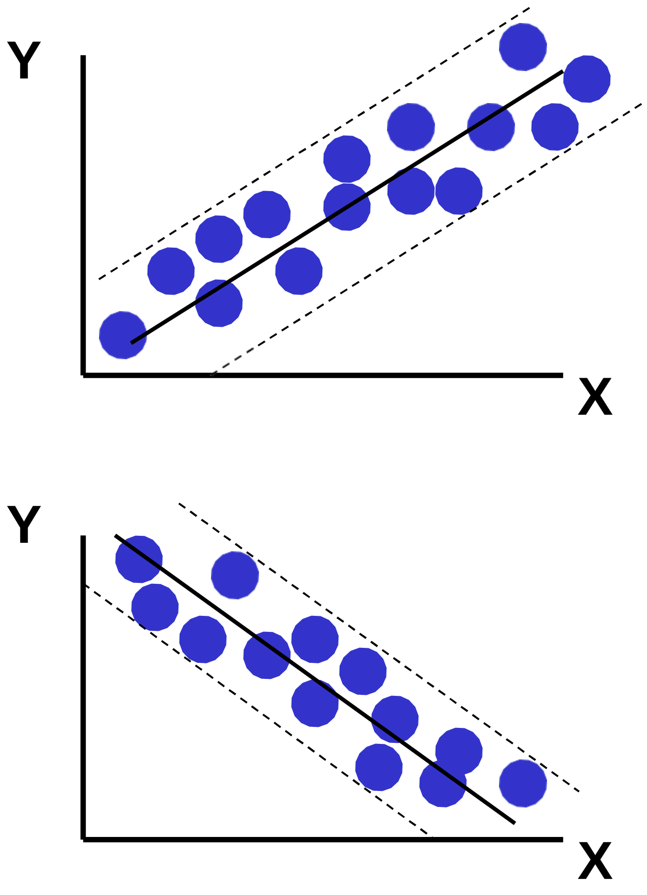


Curvilinear relationships

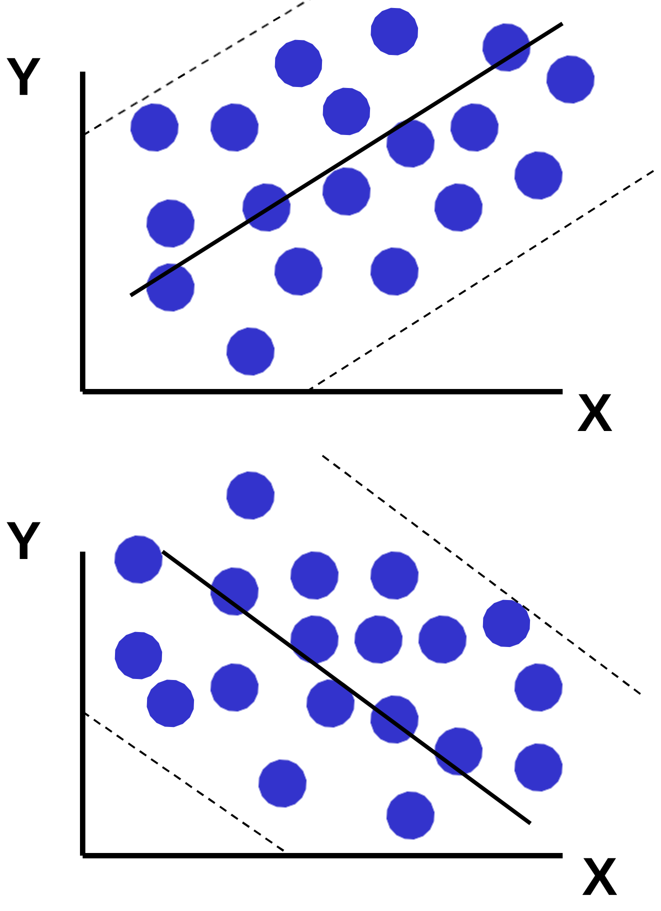


Types of Relationships

Strong relationships

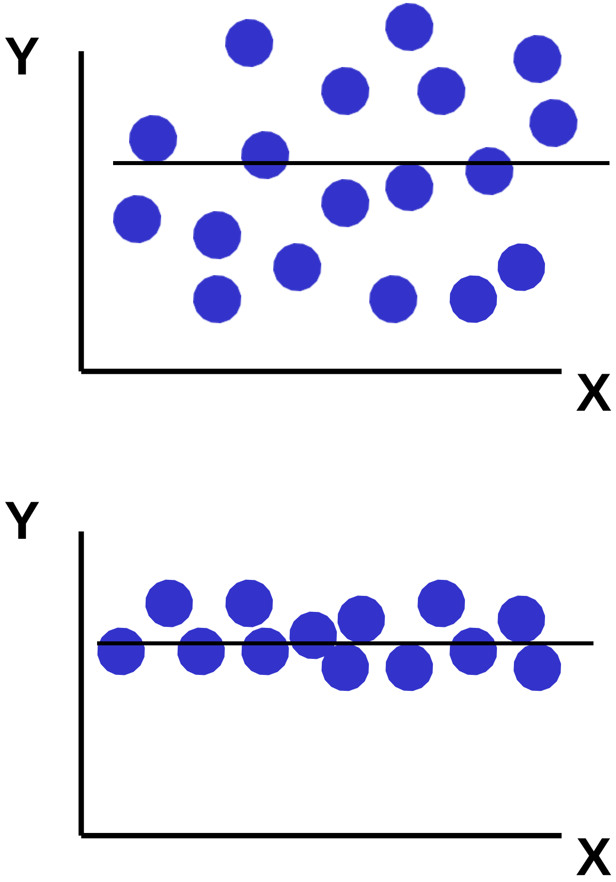


Weak relationships



Types of Relationships

No relationship



Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{b + MX_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

How do we determine if our Regression model is doing well or not?

Performance Metrics (Regression)

Mean Absolute Error - Sum of the absolute differences between predictions and actual values.

Mean Squared Error - Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

Let’s dive straight to the Hands-on using Jupyter notebooks

Agenda – Part 3

1

Logistic Regression

2

Model Evaluation

3

Support Vector Machines

Logistic Regression

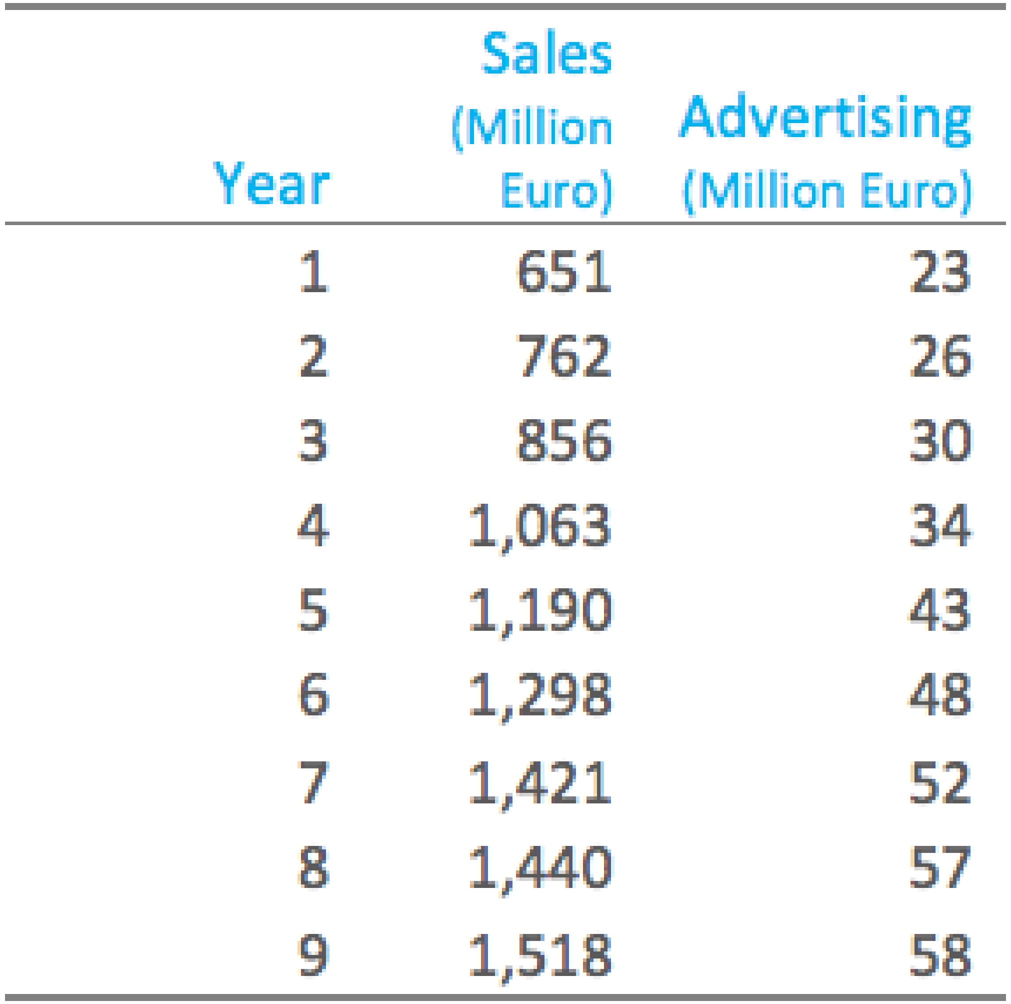
What is it and what is the algorithm?

What is the difference between Linear Regression & Logistic Regression?

Recap: What is linear regression?

- ***Linear regression*** quantifies the relationship between one or more *predictor variables* and one *outcome variable.*
- For example, linear regression can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Recap: Example



Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Sales = 168 + 23 Advertising

What is logistic regression?

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is **binary**.
- Like all regression analyses, the logistic regression is a predictive analysis.
- Logistic regression is used to describe data and to explain the relationship between one dependent **binary variable** and **one or more nominal, ordinal, interval or ratio-level independent variables.**

Good to know!

Nominal

- Nominal scales are used for labeling variables, without any quantitative value. “Nominal” scales could simply be called “labels.”
  - E.g Male/Female, Red/Green/Yellow

Ordinal

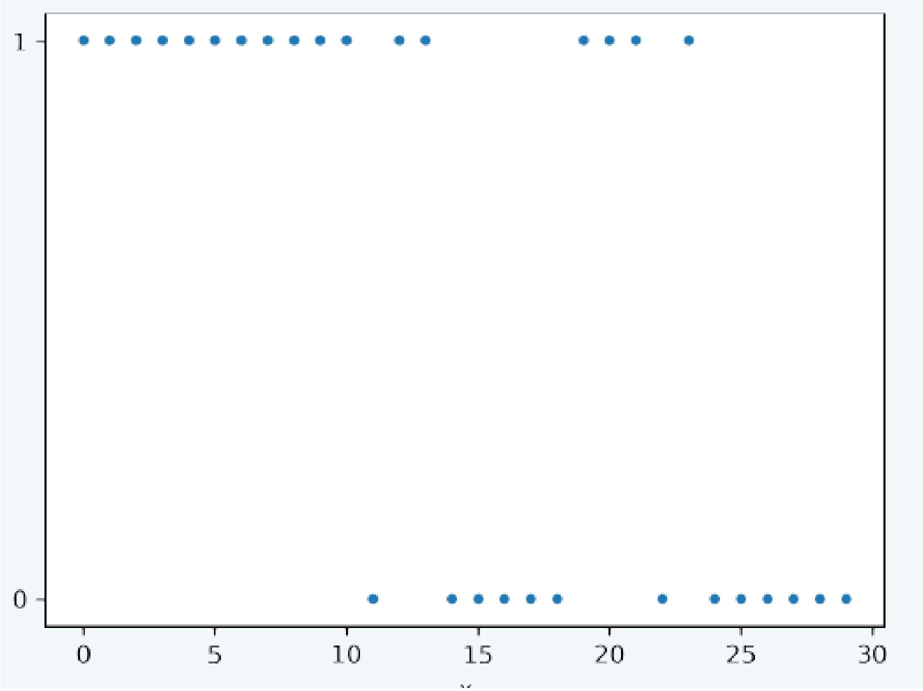
- With ordinal scales, the order of the values is what’s important and significant, but the differences between each one is not really known.
  - E.g Good, Very good, Excellent, Fantastic – 1#, 2#, 3#, 4#

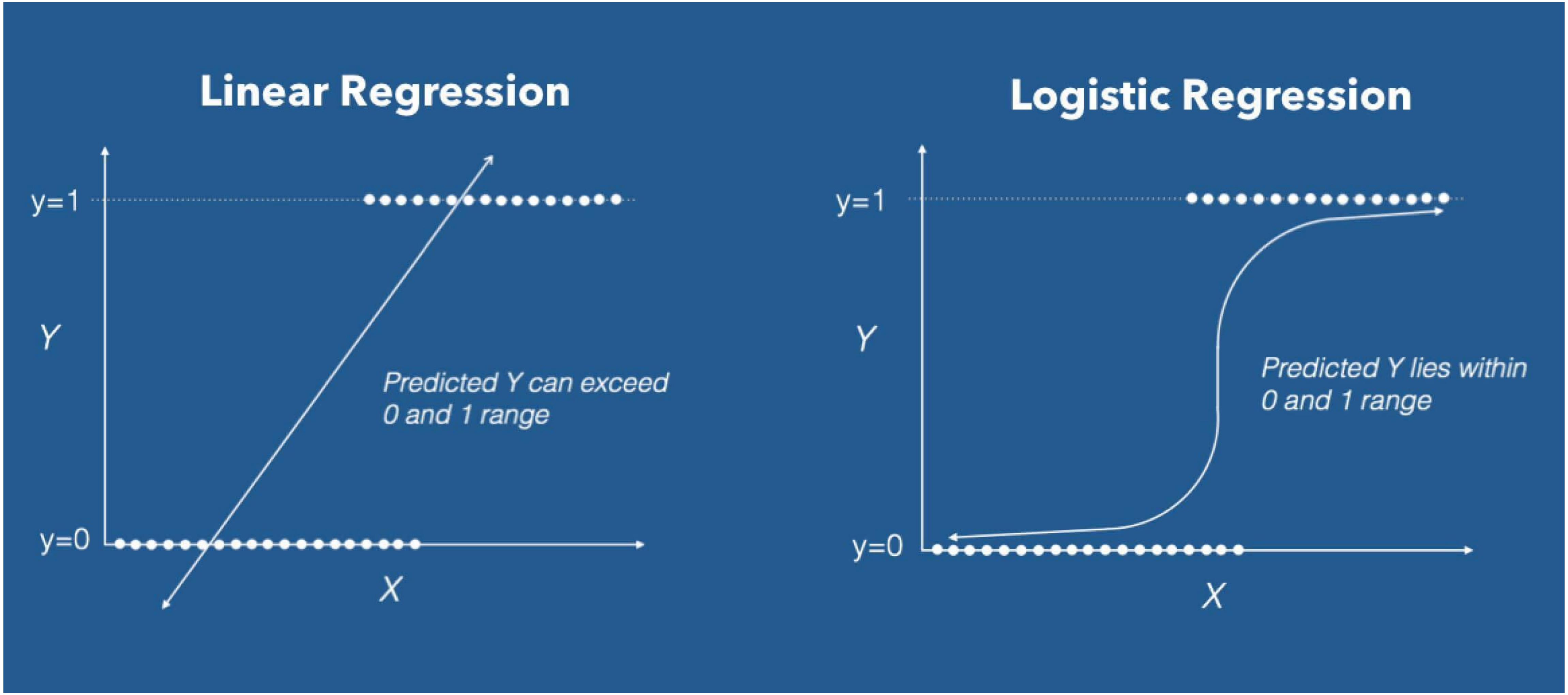
Interval

- Interval scales are numeric scales in which we know both the order and the exact differences between the values.
  - E.g Temp Celsius - because the difference between each value is the same.

Example – Log Reg – Scoring Goals!

- If we are kicking our soccer ball from a variety of distances.
- The results are going to be only Goal or no Goal.
- Our Standard Linear Regression will not work in this scenario!





Linear Regression

Predicted Y can exceed 0 and 1 range

Logistic Regression

Predicted Y lies within 0 and 1 range

Model Evaluation

Model Evaluation is an integral part of the model development process.

It helps to find the best model that represents our data and how well the chosen model will work in the future.

Performance Metrics (Classification)

Confusion Matrix

Accuracy

Precision and Recall

How do you evaluate classifiers?

Accuracy!

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

Confusion Matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes.

It is a table with 4 different combinations of predicted and actual values.

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted

		Actual Class	
		+	-
Predicted Class	+		
	-		

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- The actual classes are as seen.
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+		
		10	**90**
	-		

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	10	3
	-		83

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	8	3
	-	2	83

Let’s take an example of Confusion matrix

- Assuming there are 100 people which are to be predicted
- Now we get our predictions from our model.

		Actual Class	
		+	-
Predicted Class	+	**True +**	**False +**
	-	**False -**	**True -**

So how can we use the metrics?

Say we have 2 confusion matrix from 2 models

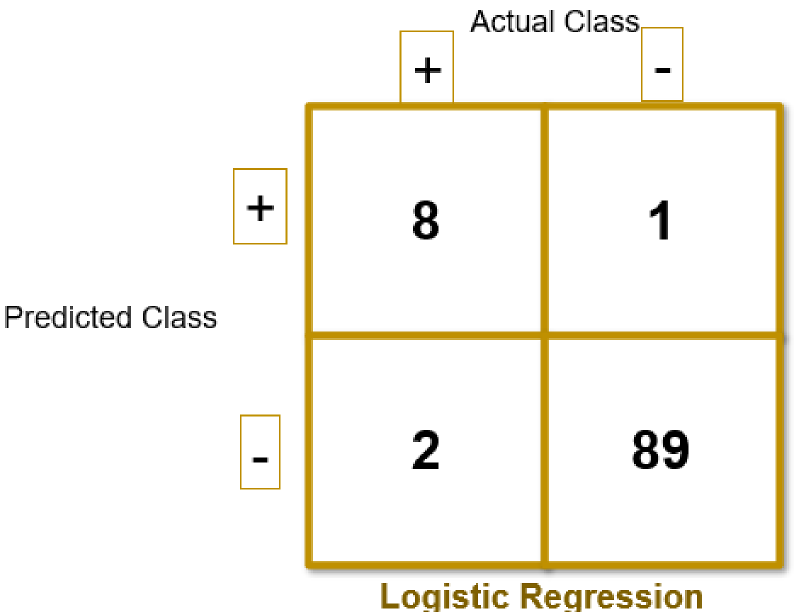
		Actual Class	
		+	-
Predicted Class	+	8	1
	-	2	89

Logistic Regression

		Actual Class	
		+	-
Predicted Class	+	10	10
	-	0	80

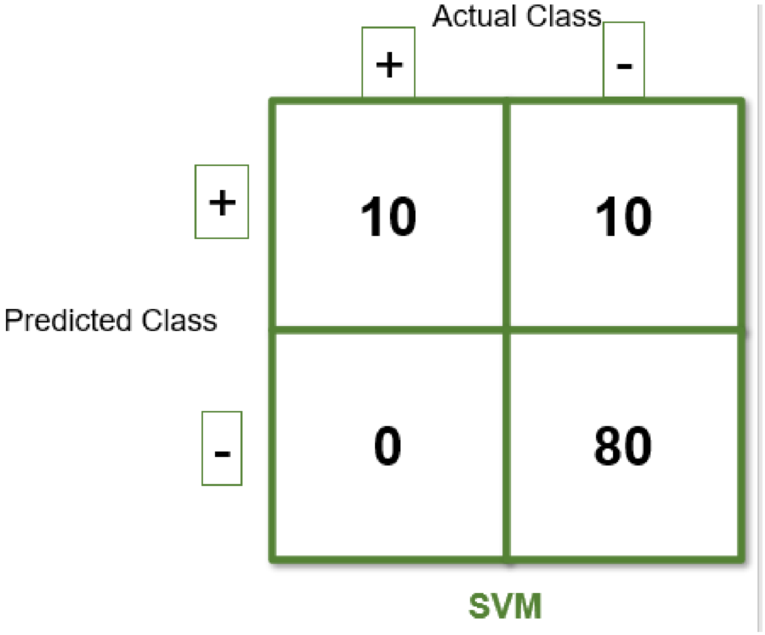
SVM

We can compare them!



Logistic Regression

		Actual Class	
		+	-
Predicted Class	+	8	1
	-	2	89



SVM

		Actual Class	
		+	-
Predicted Class	+	10	10
	-	0	80

Accuracy: (TP+TN)/(TP+TN+FP+FN)	97%	90%
Precision: TP/(TP+FP)	89%	50%
Recall: TP/(TP+FN)	80%	100%

	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 📧)	
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉ ✉ 320	FALSE NEGATIVE (FN) ✉ 📧 43	*Recall* $= \frac{TP}{TP + FN}$ $= \frac{320}{320 + 43} = 0.882$
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉ ✉ 20	TRUE NEGATIVE (TN) ✉ 📧 538	
	Precision $= \frac{TP}{TP + FP}$ $= \frac{320}{320 + 20} = 0.941$		

Precision and Recall

Precision attempts to answer the following question:
What proportion of positive identifications was correct?

Recall attempts to answer the following question:
What proportion of actual positives was identified correctly?

Let’s dive straight to the Hands-on using Jupyter notebooks

Machine Learning and Advanced Analytics using Python (Day-2)

Some Recap
What all did we learn yesterday?

Let’s try to answer the following questions!

What of the following is not a type of machine learning process?

- Unsupervised Learning
- Semi-supervised Learning
- Supervised Learning
- Pro-supervised Learning

A Self Organizing Map (SOM) is an example of which type of learning algorithm?

- Unsupervised Learning
- Supervised Learning

Imagine, you are solving a classification problems with highly imbalanced class.

The majority class is observed 99% of times in the training data. Which of the following is a suitable metric to look at?

- Accuracy
- Precision
- Mean Absolute Error
- None of the above

A feature F can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

Which of the following statement is true?

- Feature F is an example of nominal variable
- Feature F is an example of ordinal variable
- Both the above
- None of the ab

Back to last week!

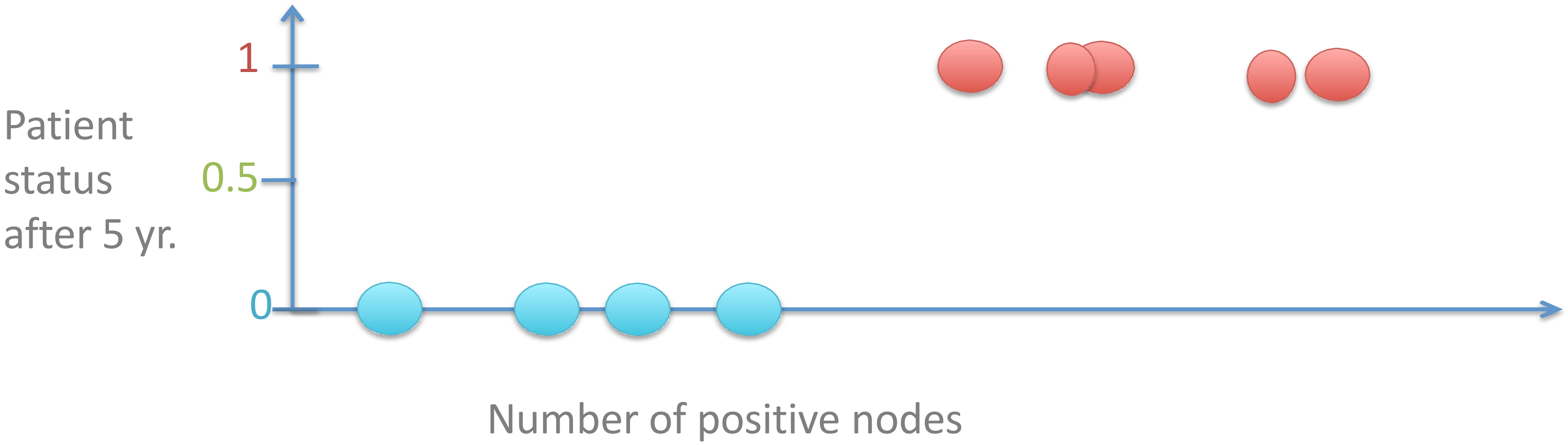
Logistic Regression!

Support Vector Machines

What are SVMs?

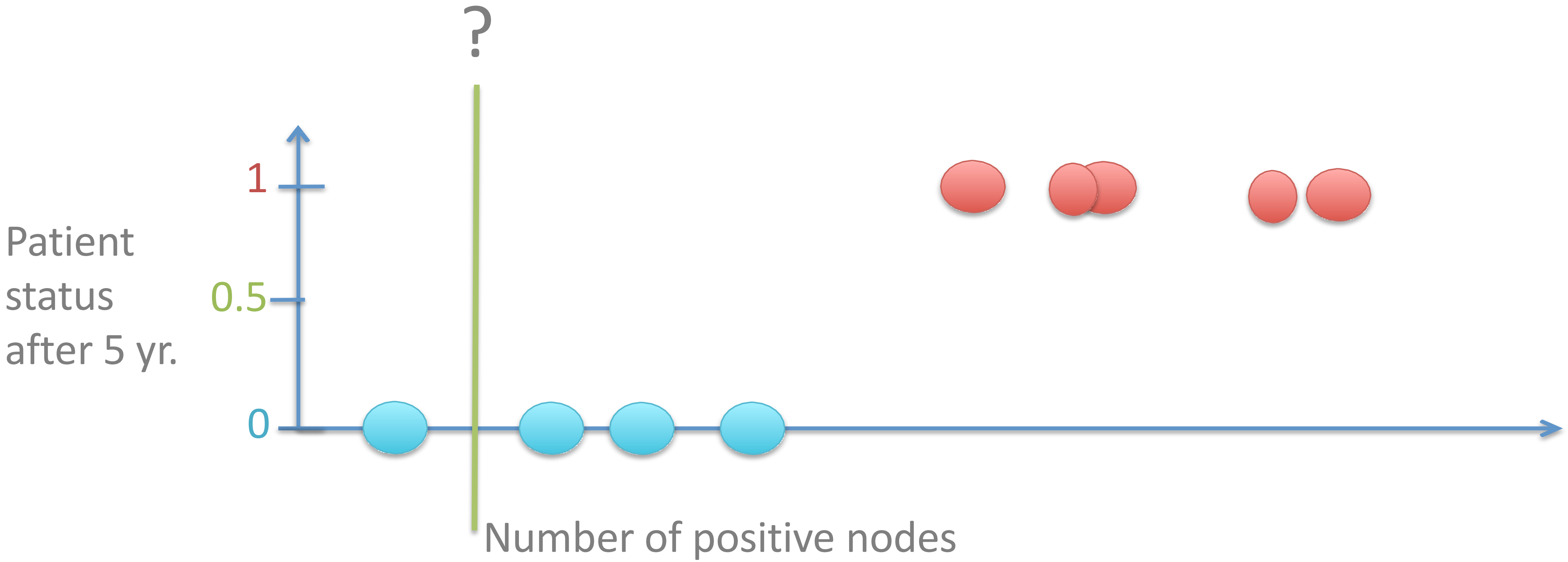
- SVMs are linear or non-linear classifiers that find a hyperplane to separate two class of data, positive and negative.
- SVM not only has a rigorous theoretical foundation, but also performs classification more accurately than most other methods in applications, especially for high dimensional data

Support Vector Machine (SVM)



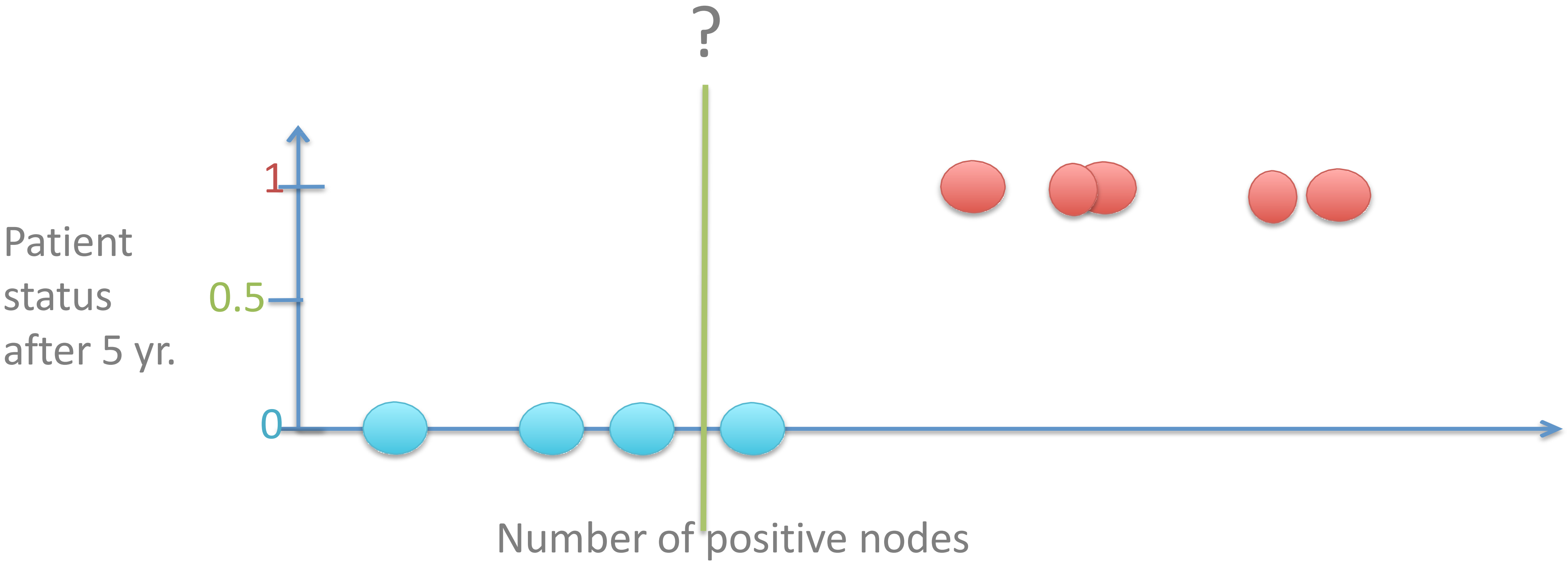
Find the best boundary that separates two classes

Support Vector Machine (SVM)



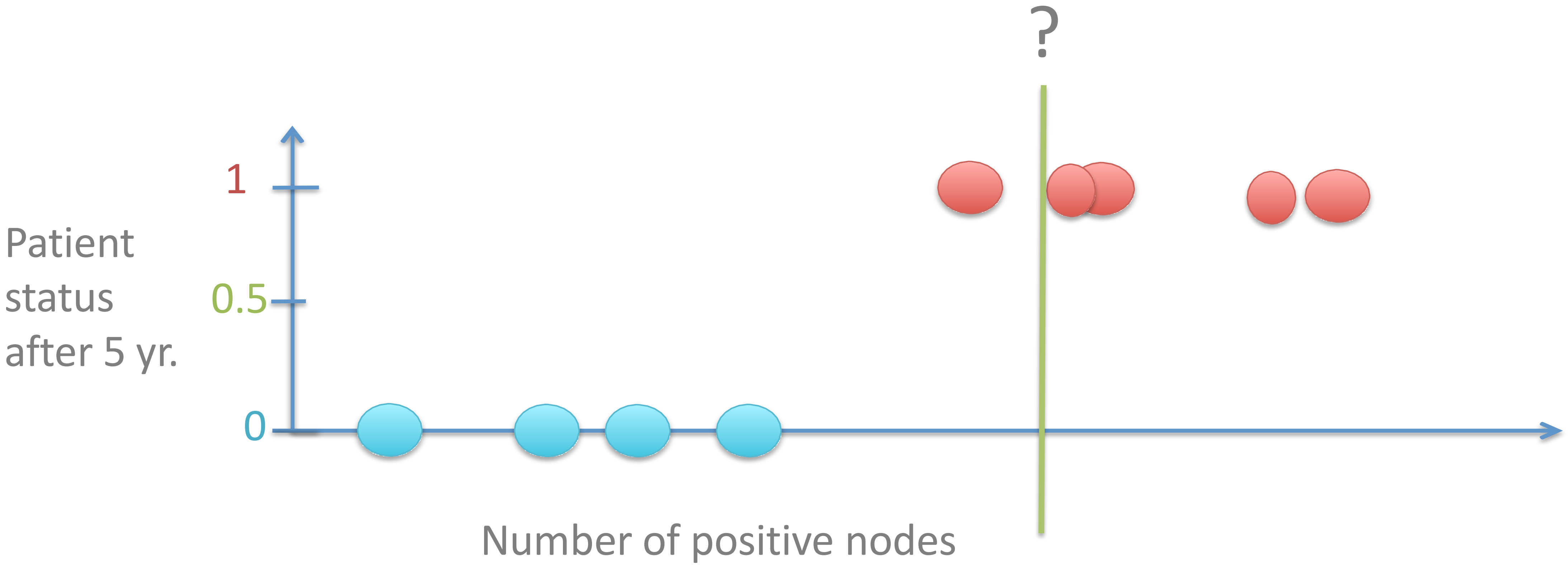
Bad: 3 misclassifications, accuracy 67%

Support Vector Machine (SVM)



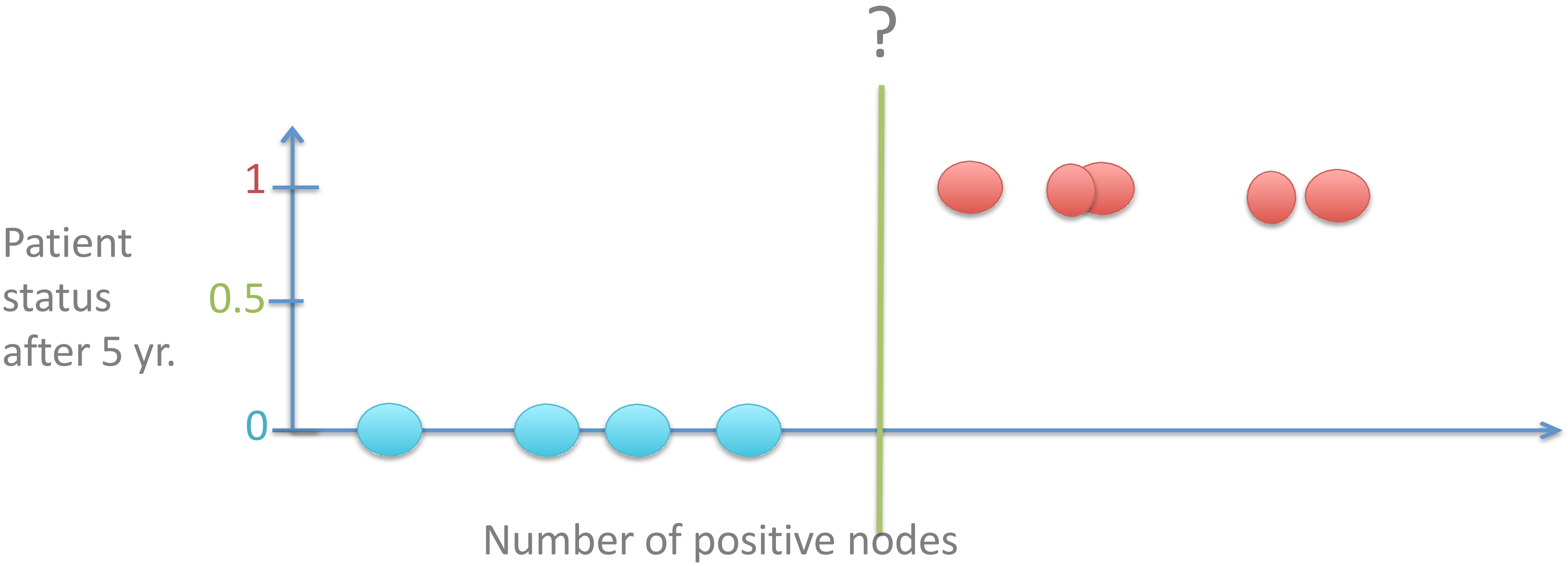
One misclassification, accuracy 89%

Support Vector Machine (SVM)



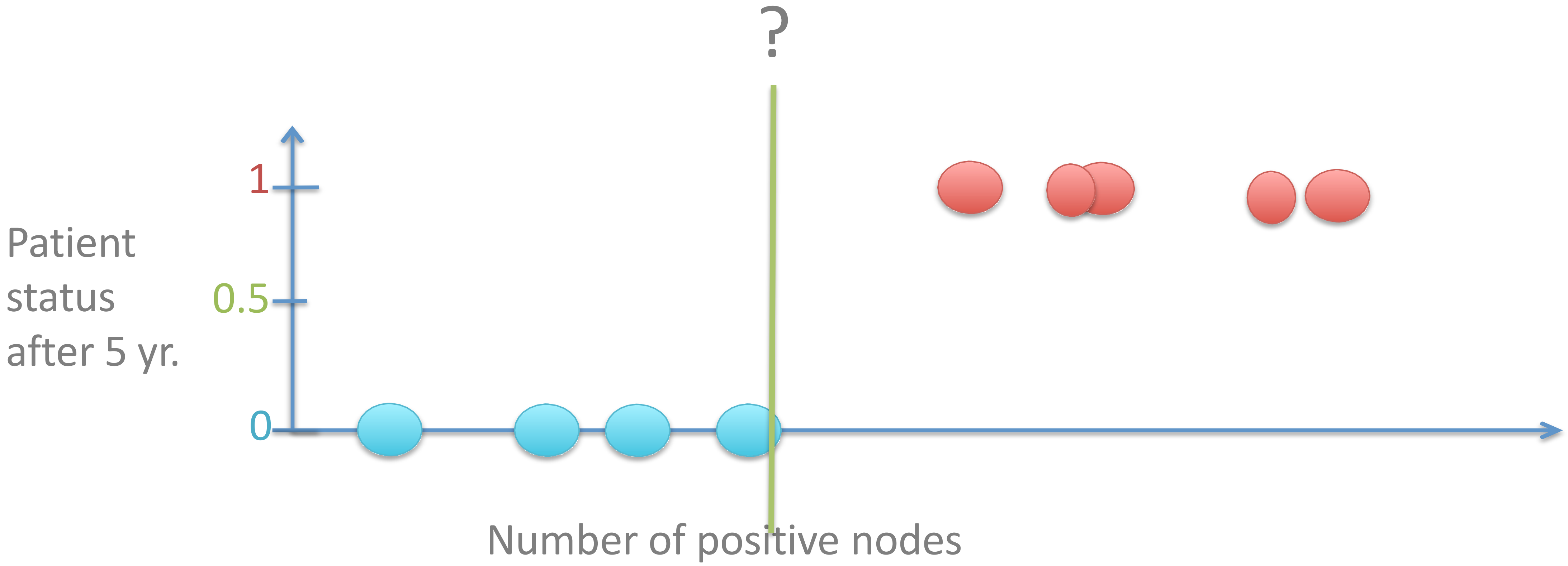
Accuracy: 78%

Support Vector Machine (SVM)



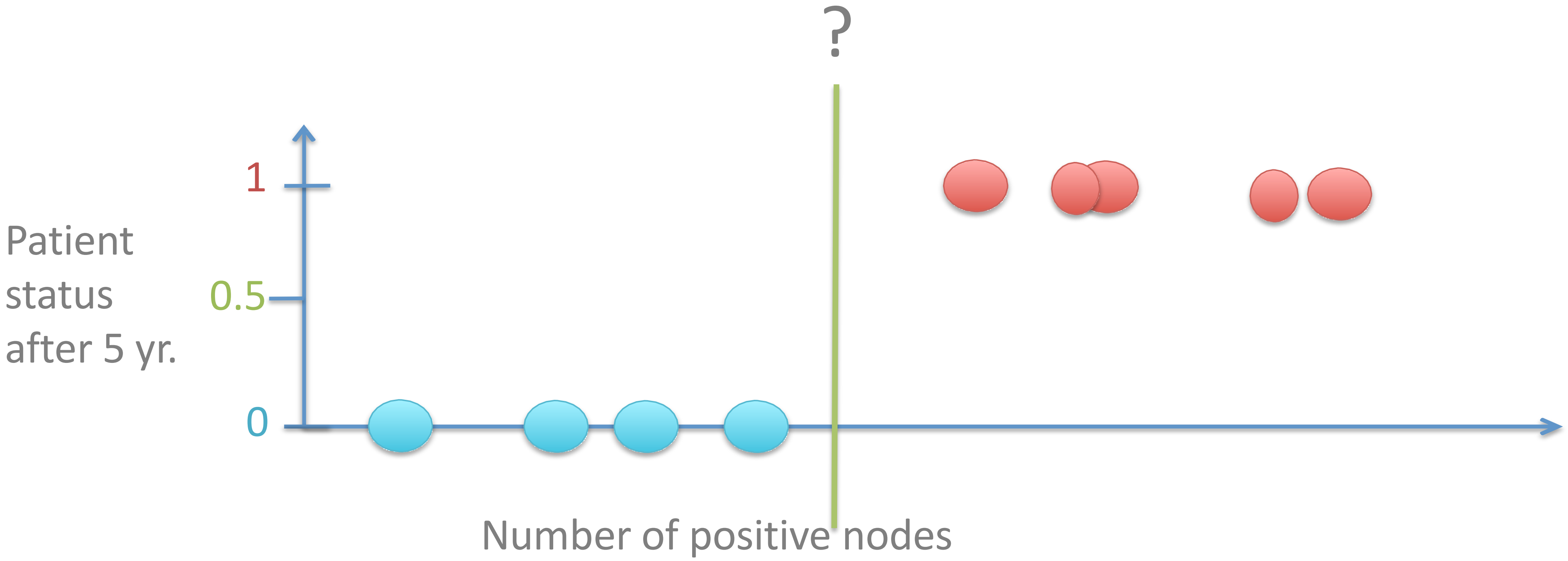
Accuracy: 100%

Support Vector Machine (SVM)



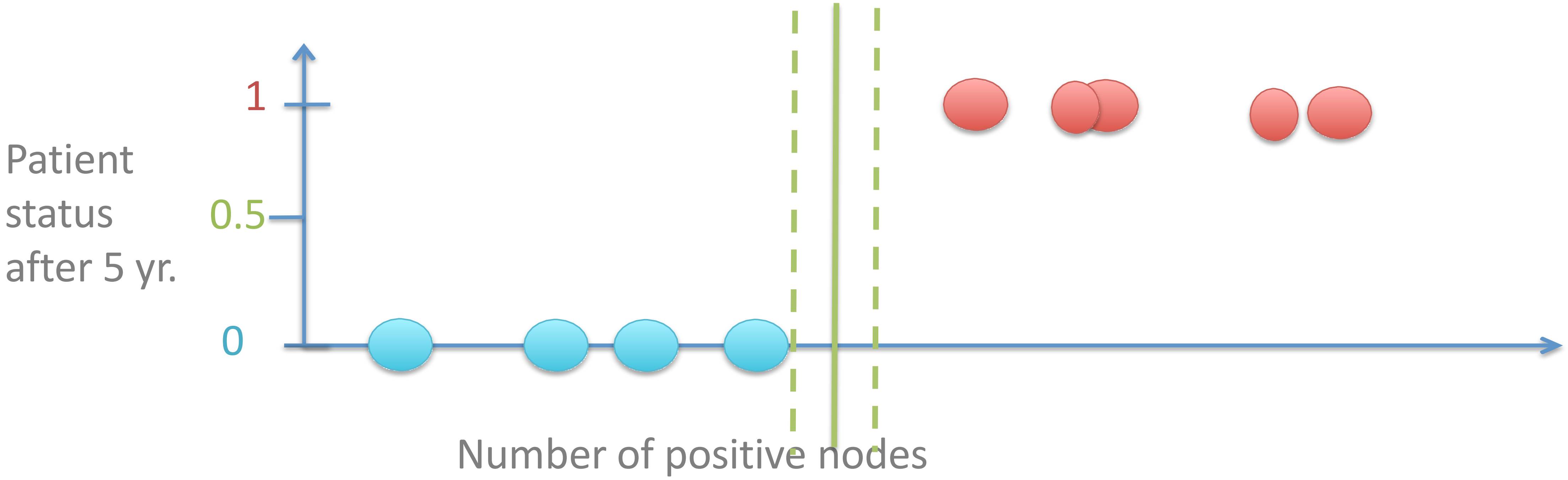
Accuracy: 100%

Support Vector Machine (SVM)



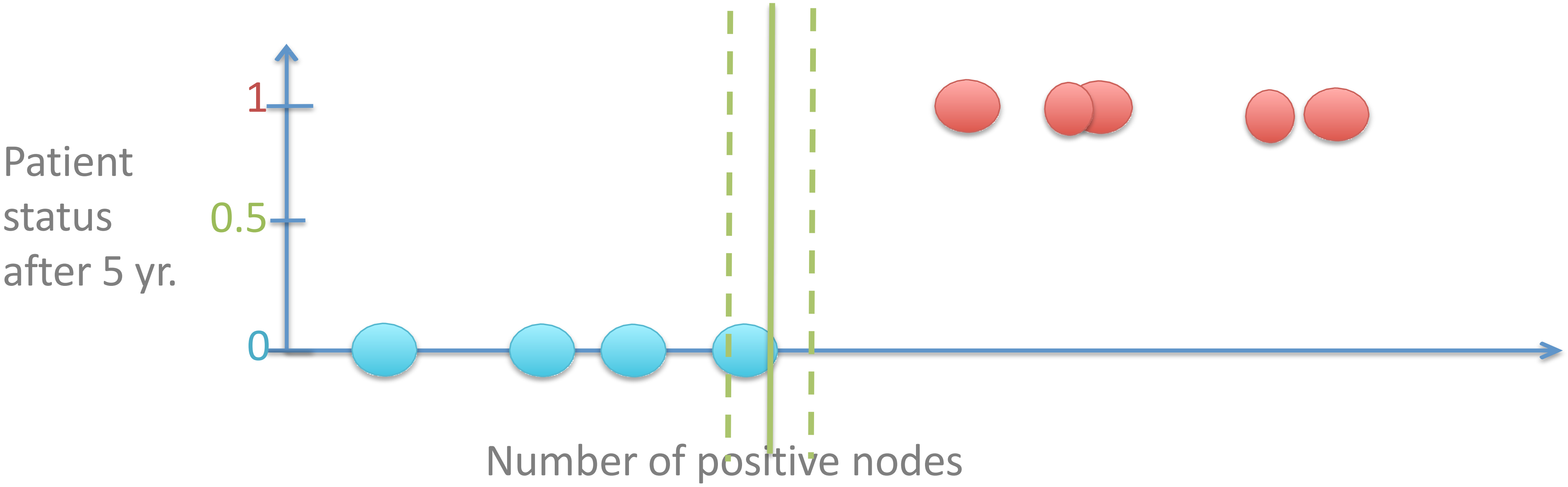
Accuracy: 100%

Support Vector Machine (SVM)



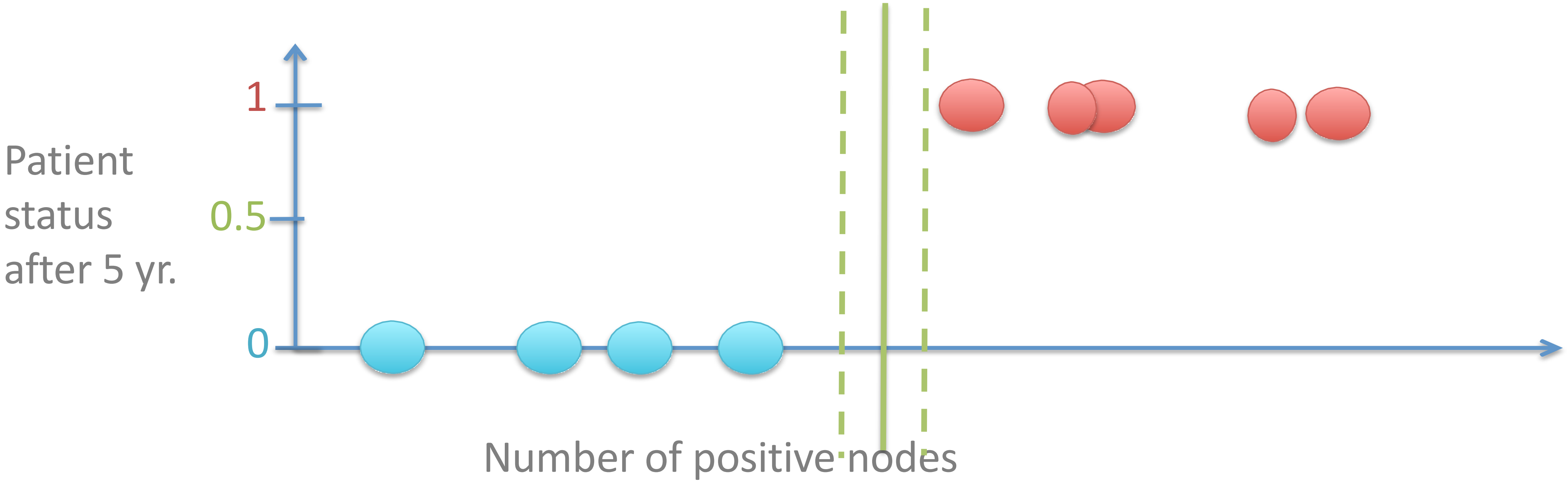
The margin: No man’s land

Support Vector Machine (SVM)



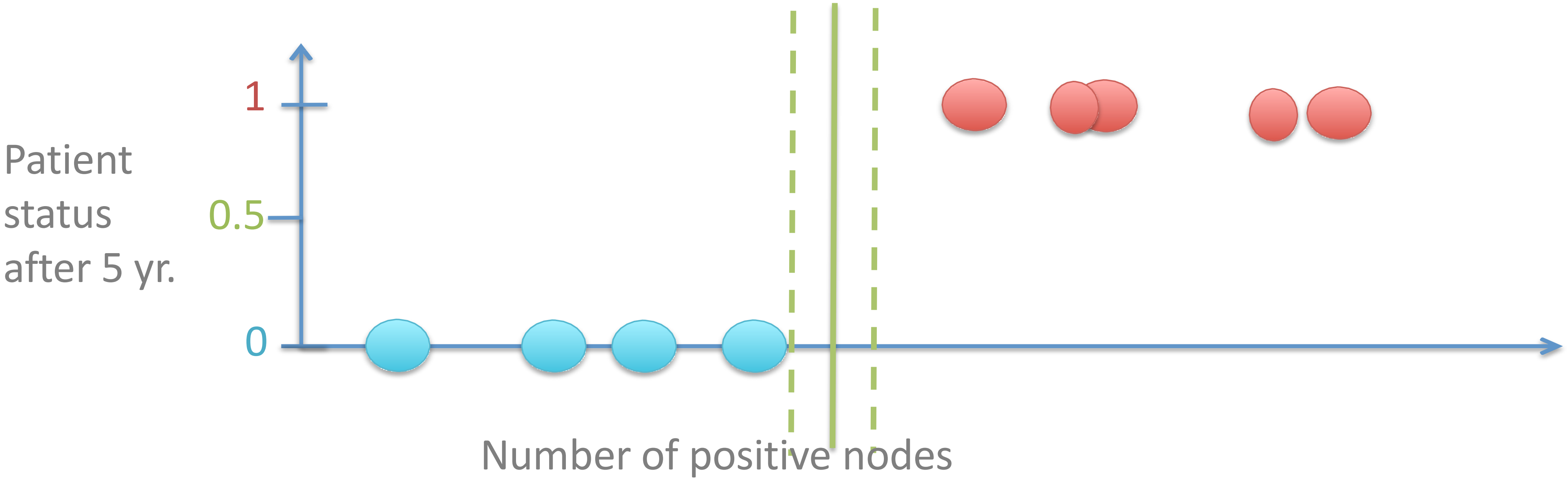
Extra error due to point in margin

Support Vector Machine (SVM)



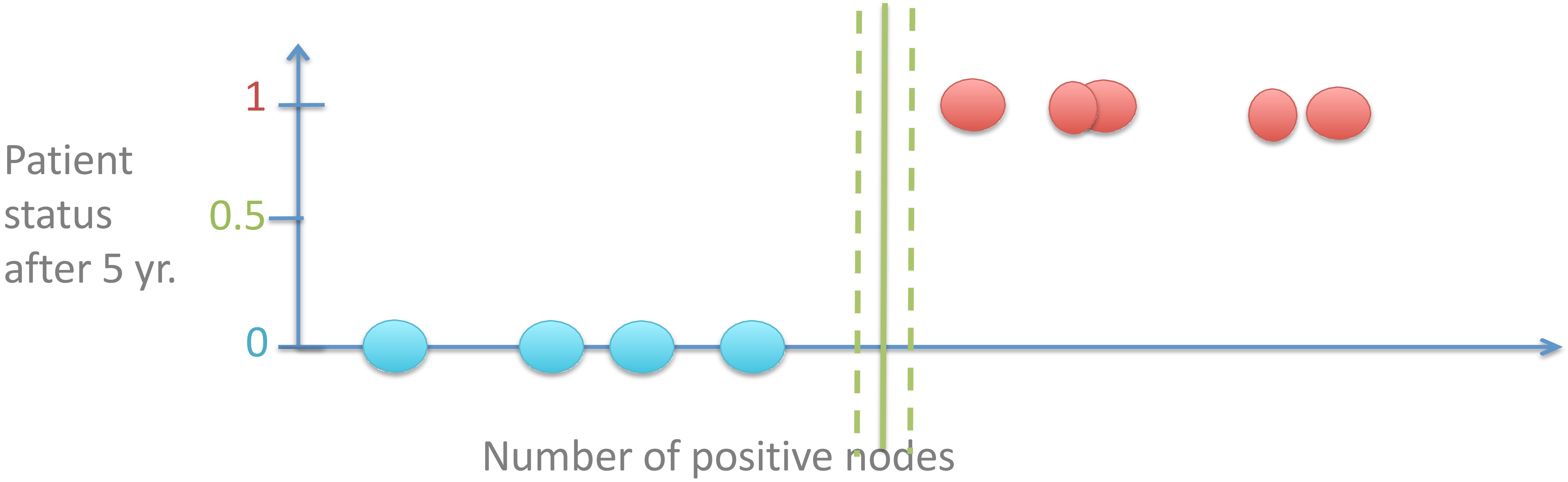
Extra error due to point in margin

Support Vector Machine (SVM)



Best boundary

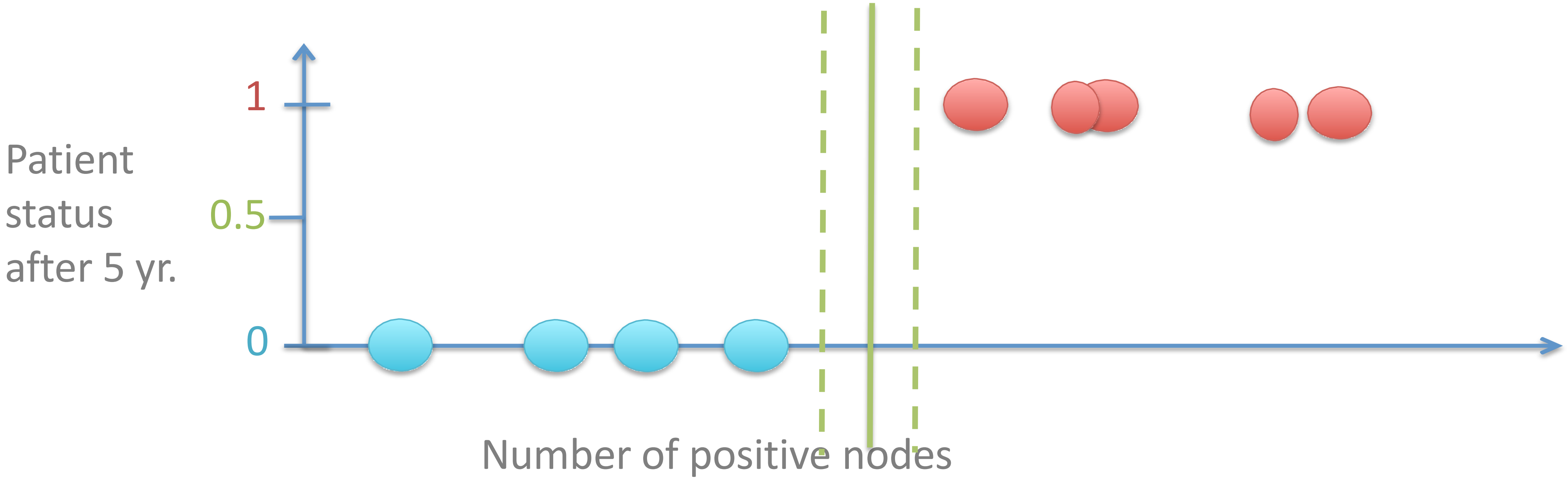
Support Vector Machine (SVM)



An even better way of doing this:

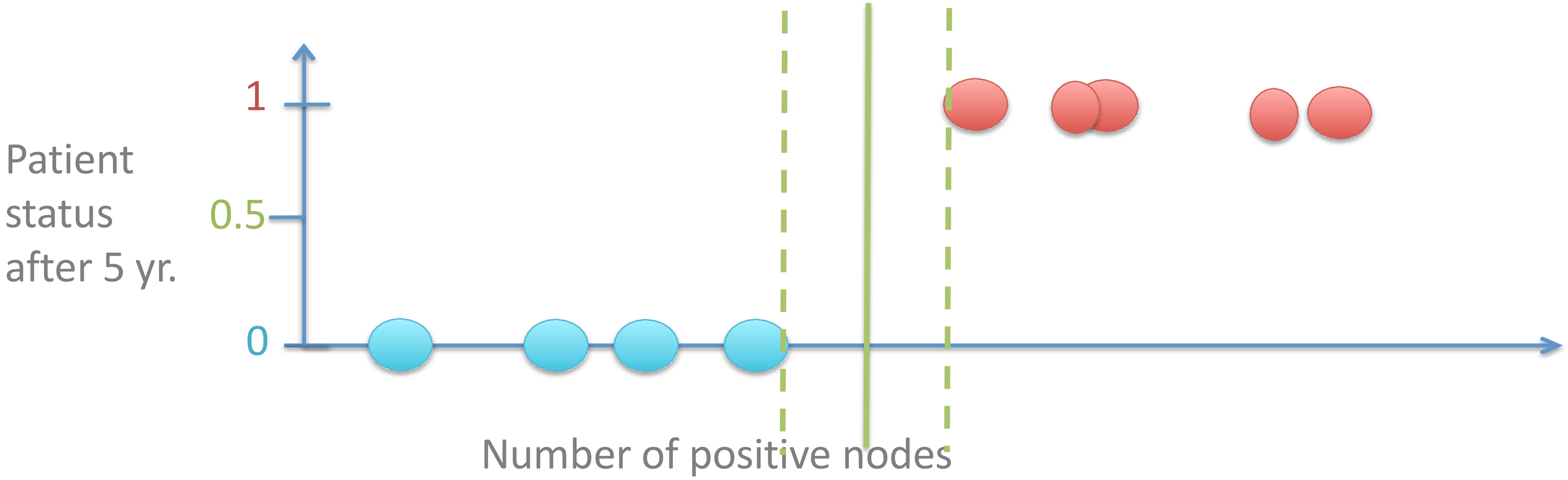
Find the boundary with the largest margin

Support Vector Machine (SVM)



Extra error due to point in margin

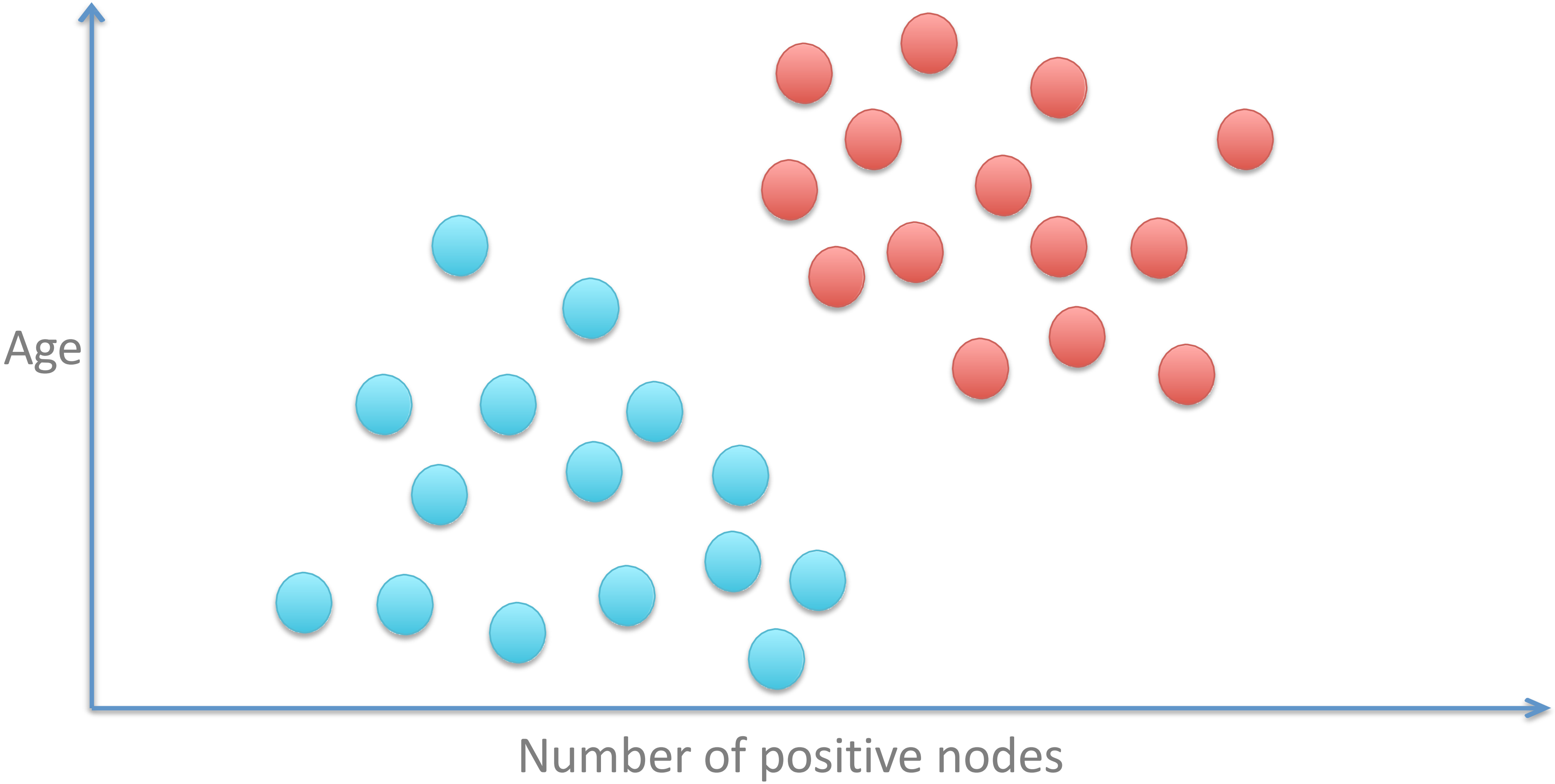
Support Vector Machine (SVM)



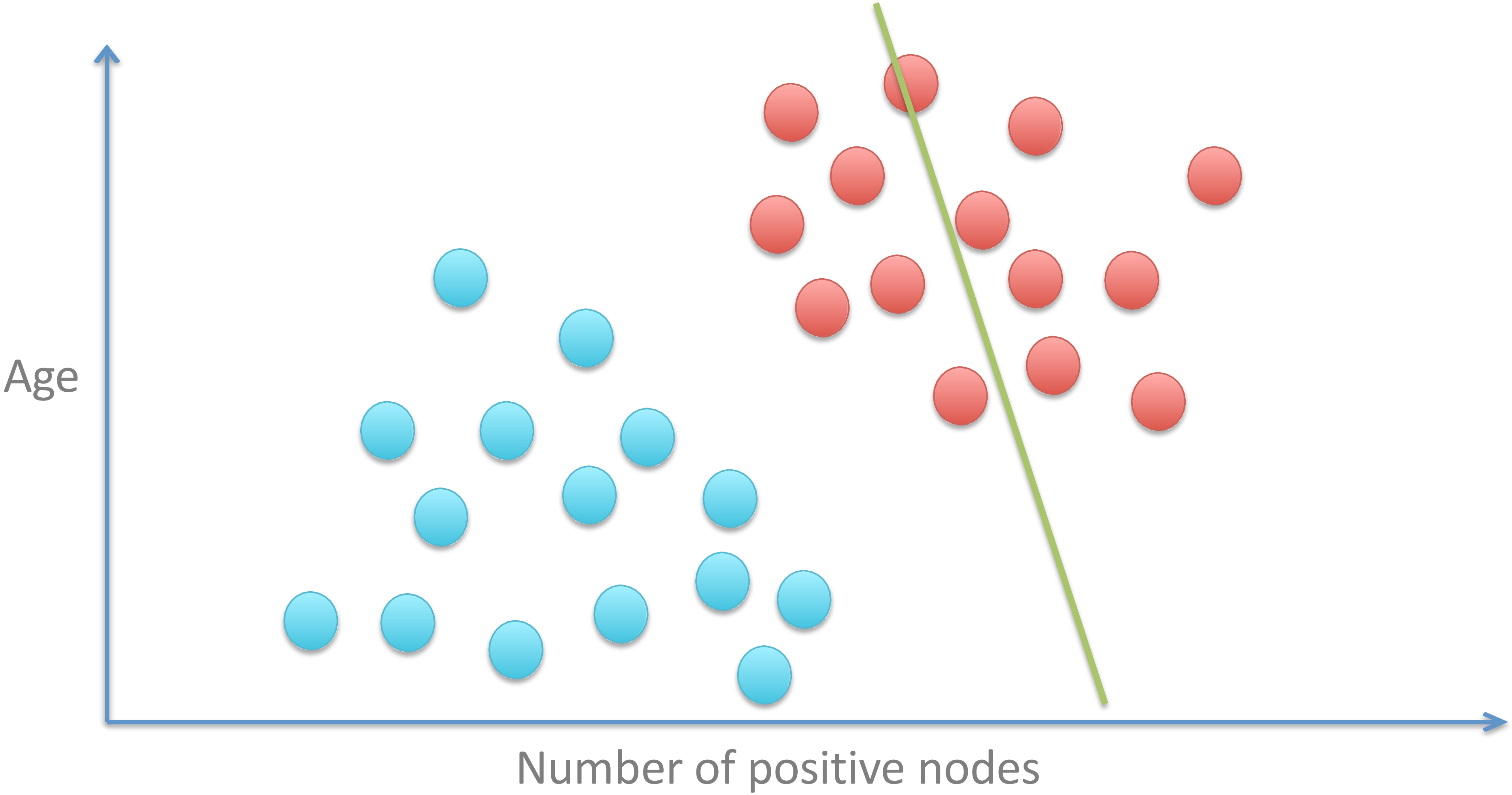
Best boundary

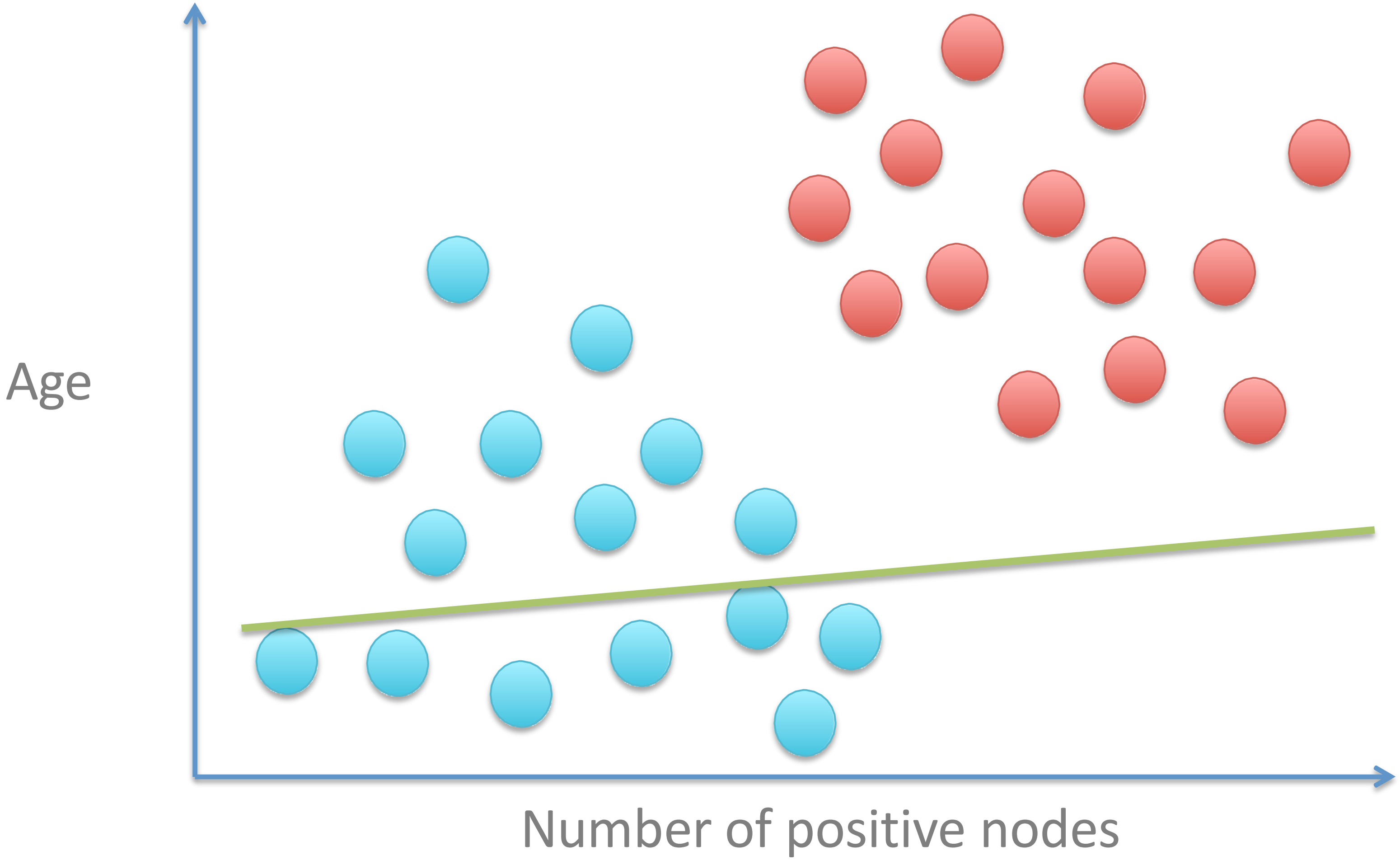
2 Features: Number of + nodes, Age

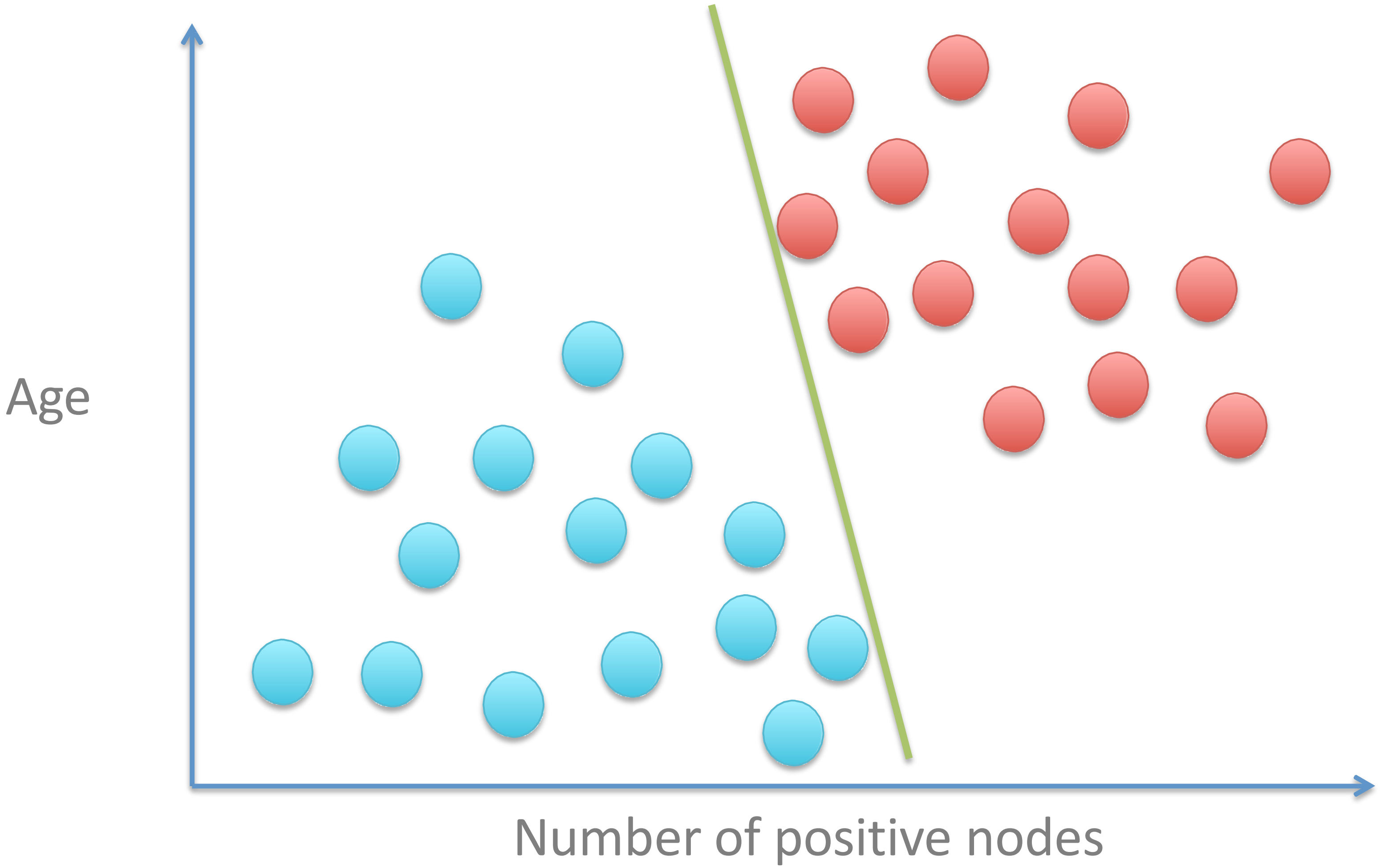
2 Labels: Survived / Lost

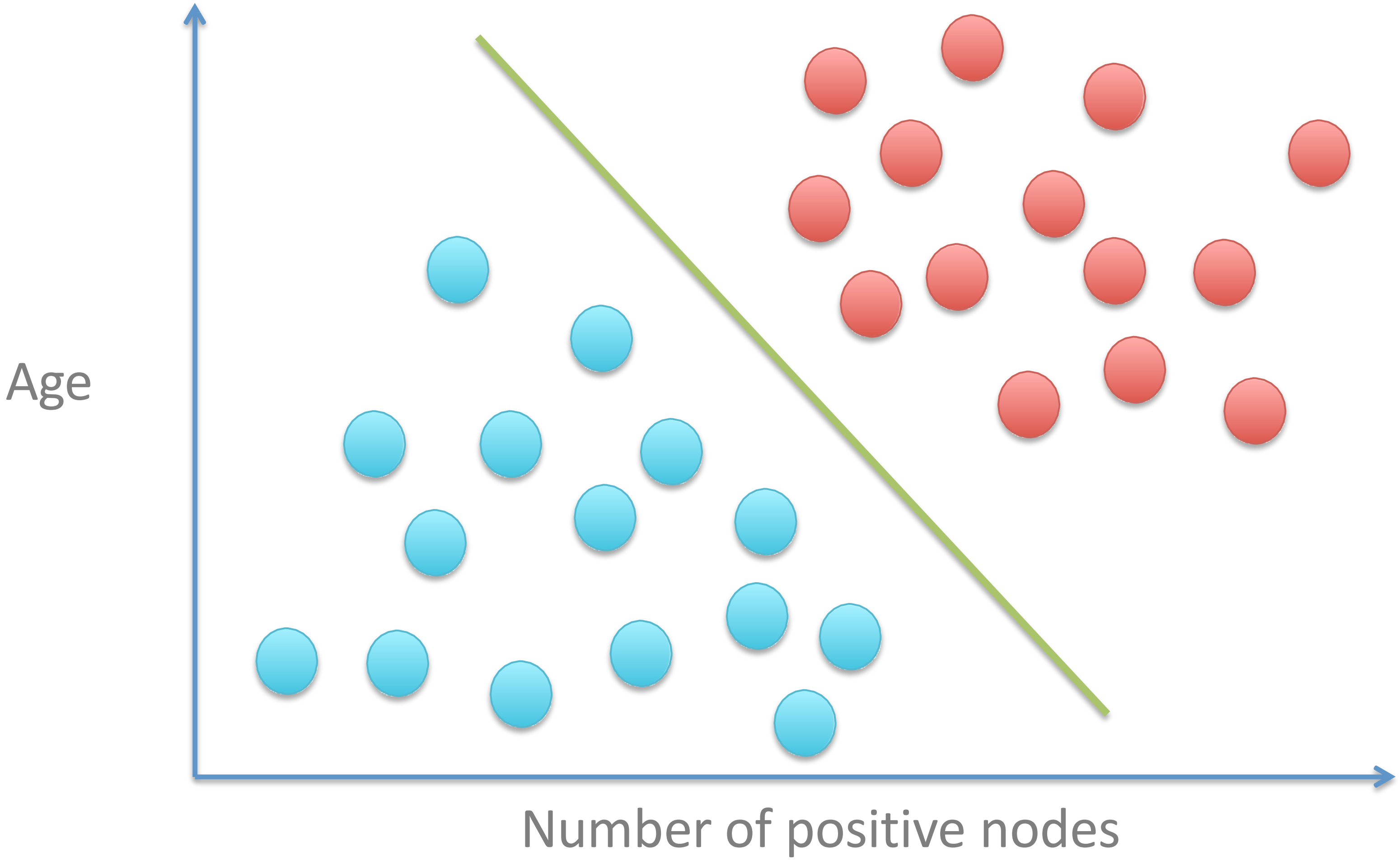


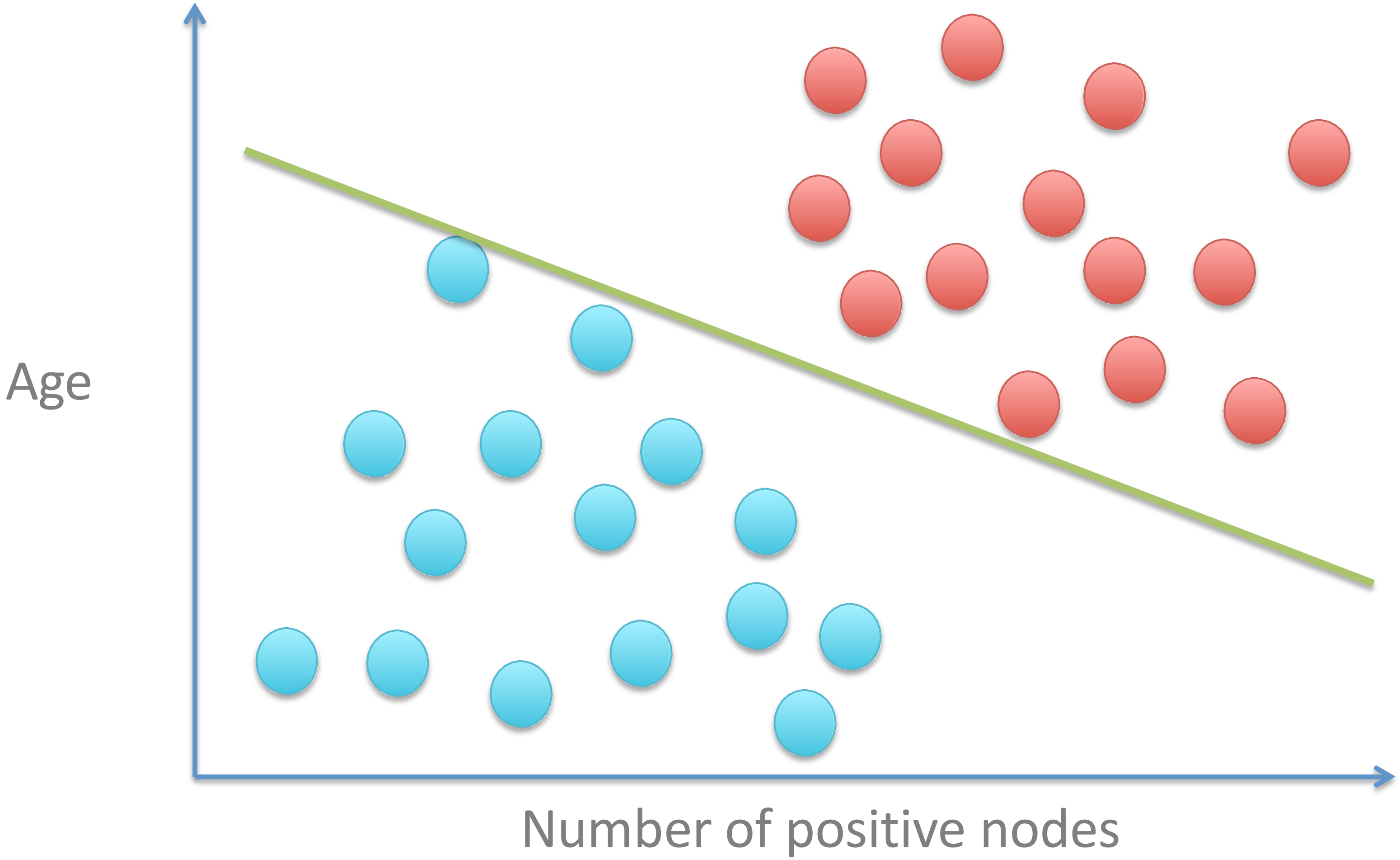
Find the line that separates the classes best

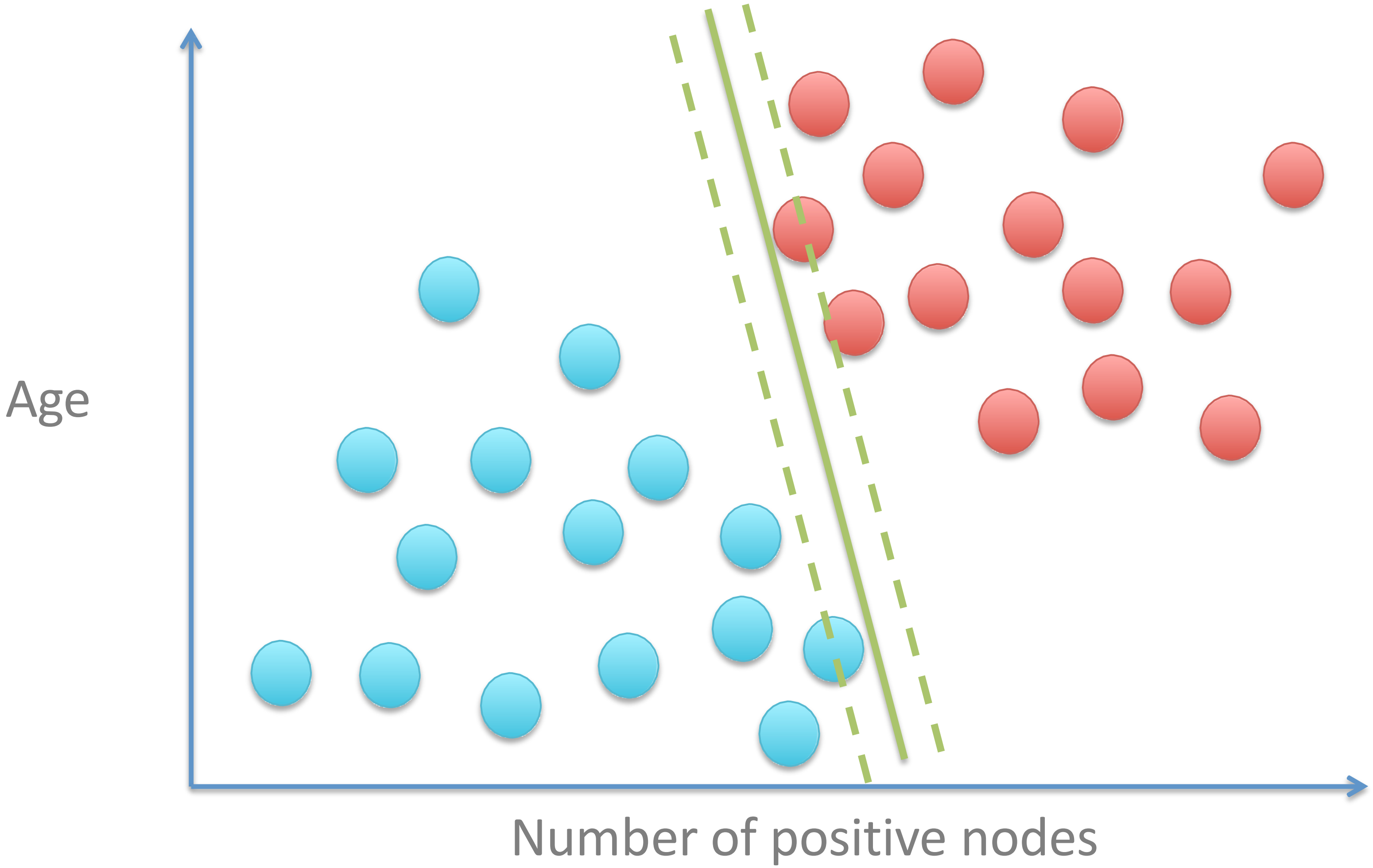


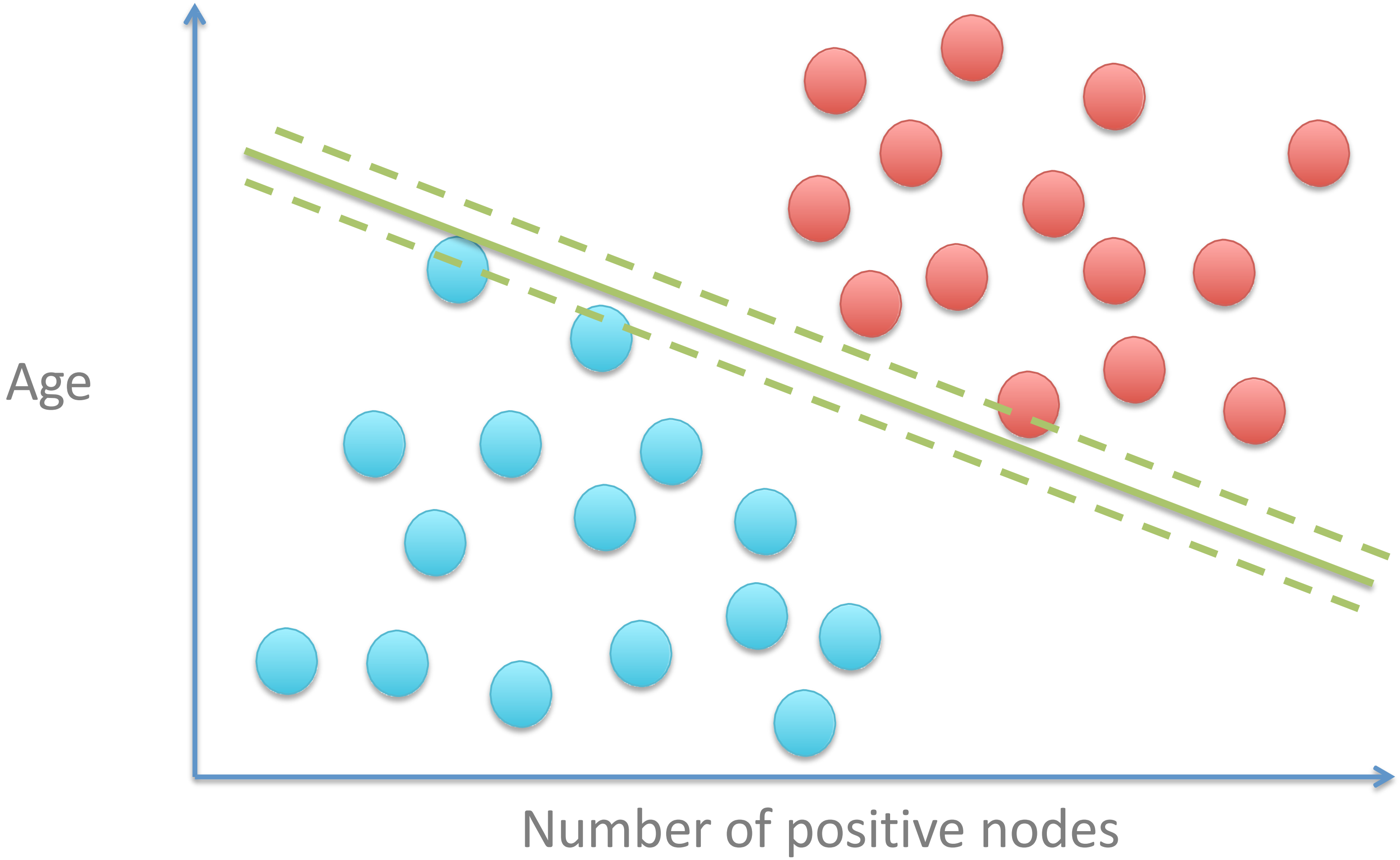

Age
Number of positive nodes


Age
Number of positive nodes

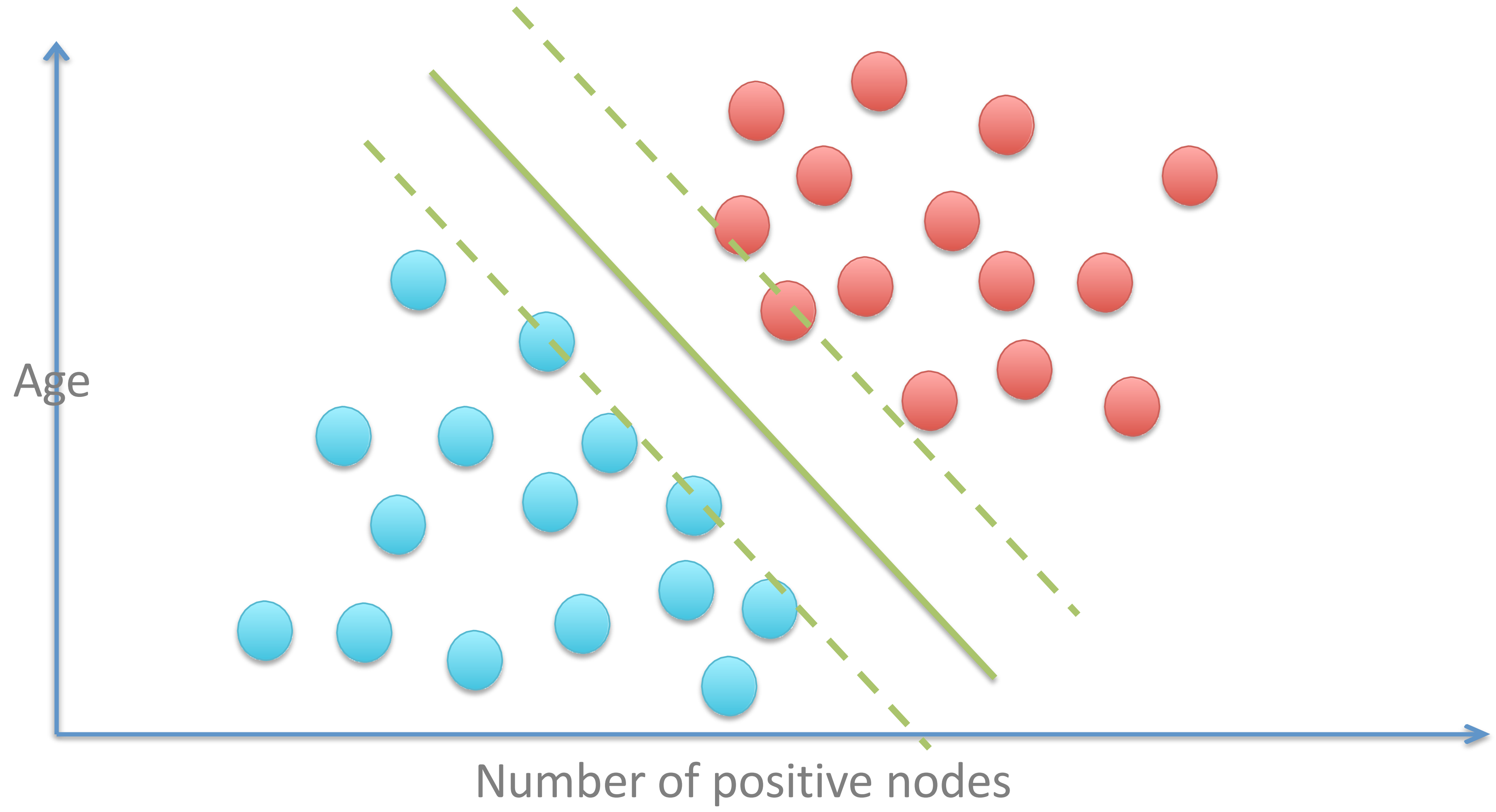

Age
Number of positive nodes


Age
Number of positive nodes


Age
Number of positive nodes

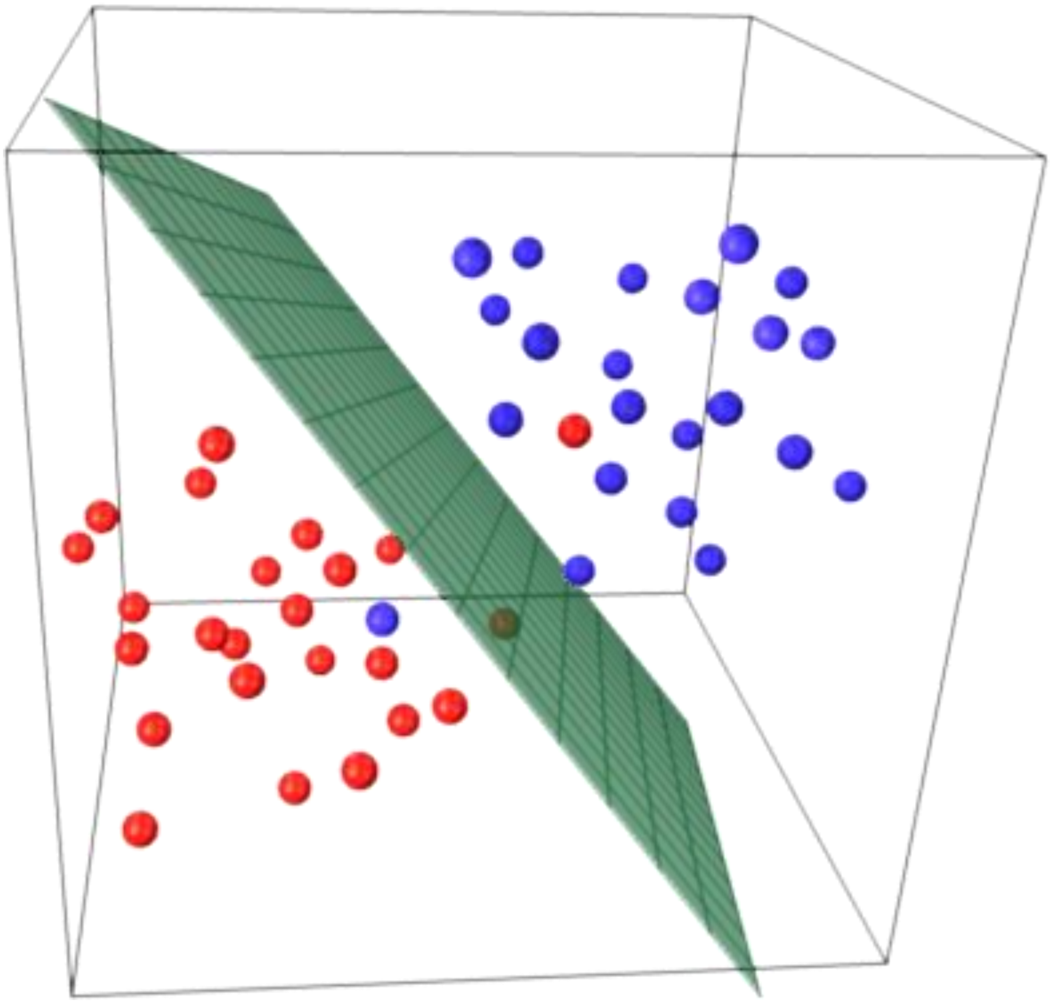

Age
Number of positive nodes

Best boundary



3 features: Find the best boundary plane

(More features: hyperplane)

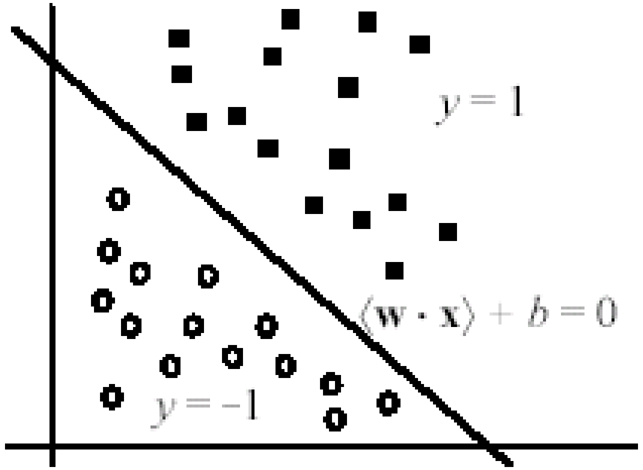


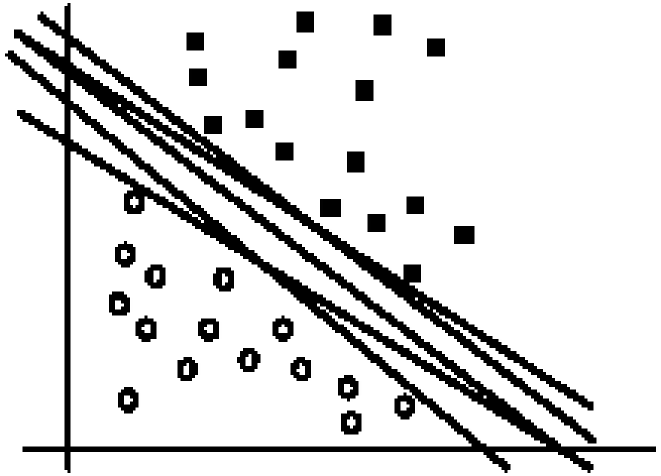
What is a hyperplane?

- The hyperplane that separates positive and negative training data is

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$$

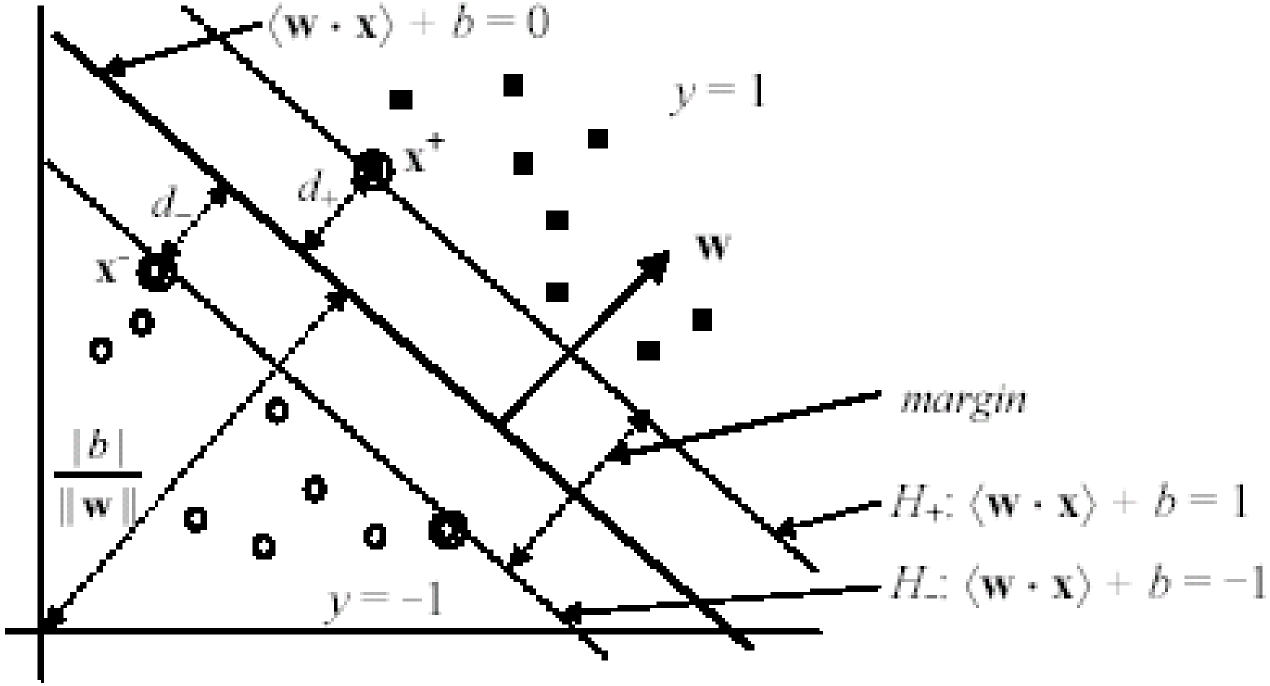
- It is also called the decision boundary (surface).





How to choose the best hyperplane?

- SVM looks for the separating hyperplane with the largest margin.
- Machine learning theory says this hyperplane minimizes the error bound



$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$
$y = 1$
$\mathbf{x}^+$
d_+
d_-
$\mathbf{x}^-$
$\mathbf{w}$
$margin$
$\frac{|b|}{\|\mathbf{w}\|}$
H_+: $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 1$
H_-: $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = -1$
$y = -1$

Pros

- Accuracy
- Works well on smaller cleaner datasets
- It can be more efficient because it uses a subset of training points

Cons

- Isn’t suited to larger datasets as the training time with SVMs can be high
- Less effective on noisier datasets with overlapping classes

Let’s dive straight to the Hands-on using Jupyter notebooks

Agenda – Part 4

1

Decision Trees

2

ML in real use cases

Quick Recap!!

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- **A decision is needed: whether to put a new patient in an intensive-care unit.**
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- **Problem**: to predict high-risk patients and discriminate them from low-risk patients.

Another example..

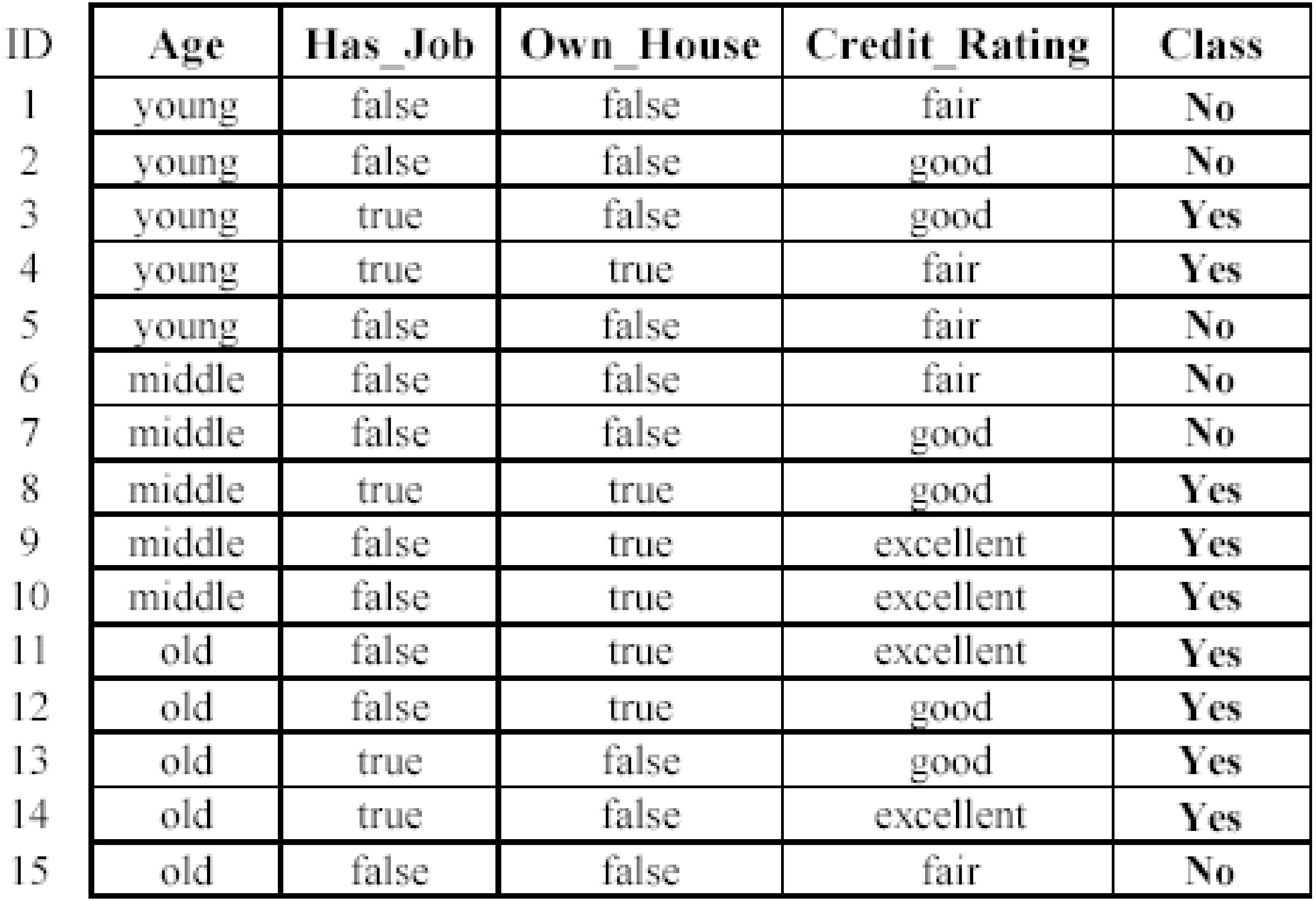
- A credit card company receives lots of applications for new cards. Each application contains information about the applicant for the card,
    - age
    - Marital status
    - annual salary
    - location
    - outstanding debts
    - credit rating
    - Family information etc
- <u>**Problem**</u>: to decide whether an application should be approved or not approved.

Decision Trees

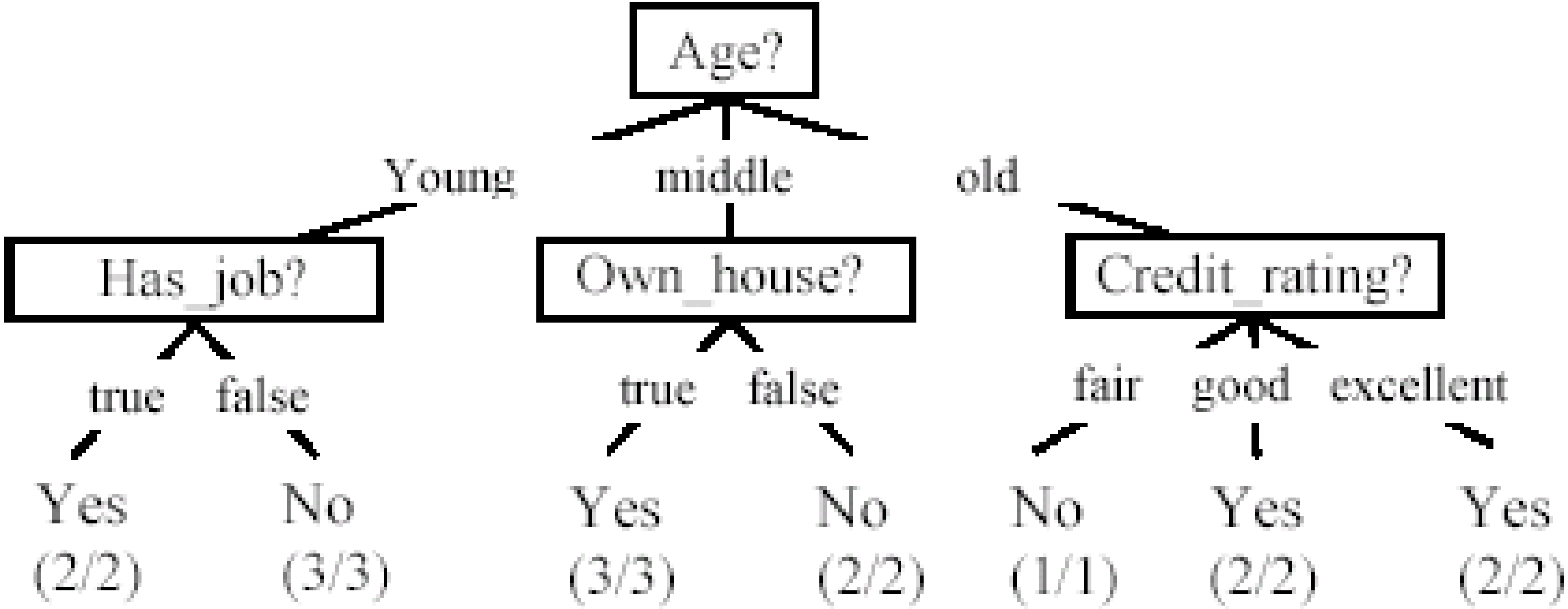
Introduction

Decision tree learning is one of the most widely used techniques for classification.

The classification model is a tree, called decision tree.



ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

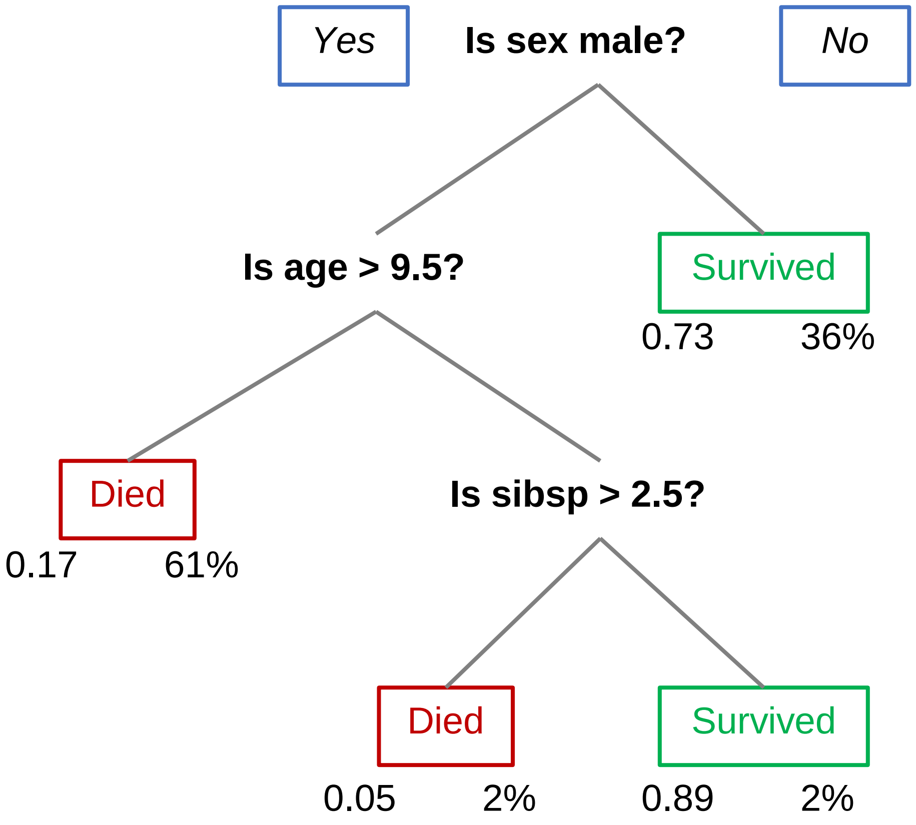


A decision tree can be converted to a set of rules

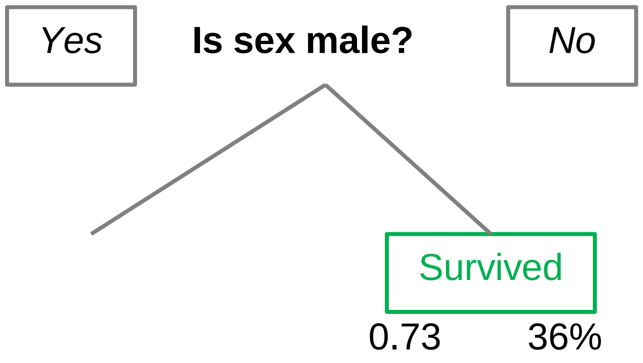
Each inner node is a decision based on a feature

Each leaf node is a **class label**

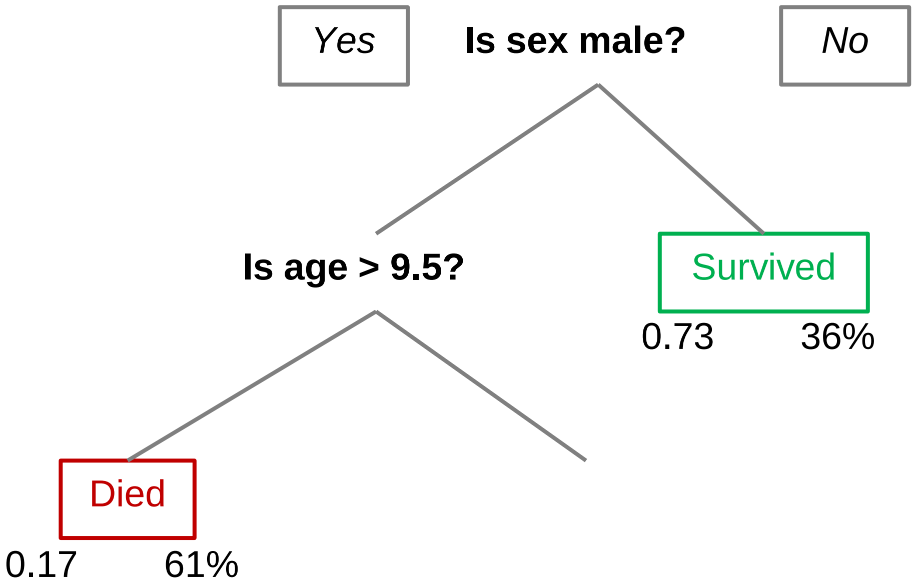
Predicting Titanic Survivors



Build tree split by split,
Find the best split you can at each step



Build tree split by split,
Find the best split you can at each step



Yes

Is sex male?

No

Is age > 9.5?

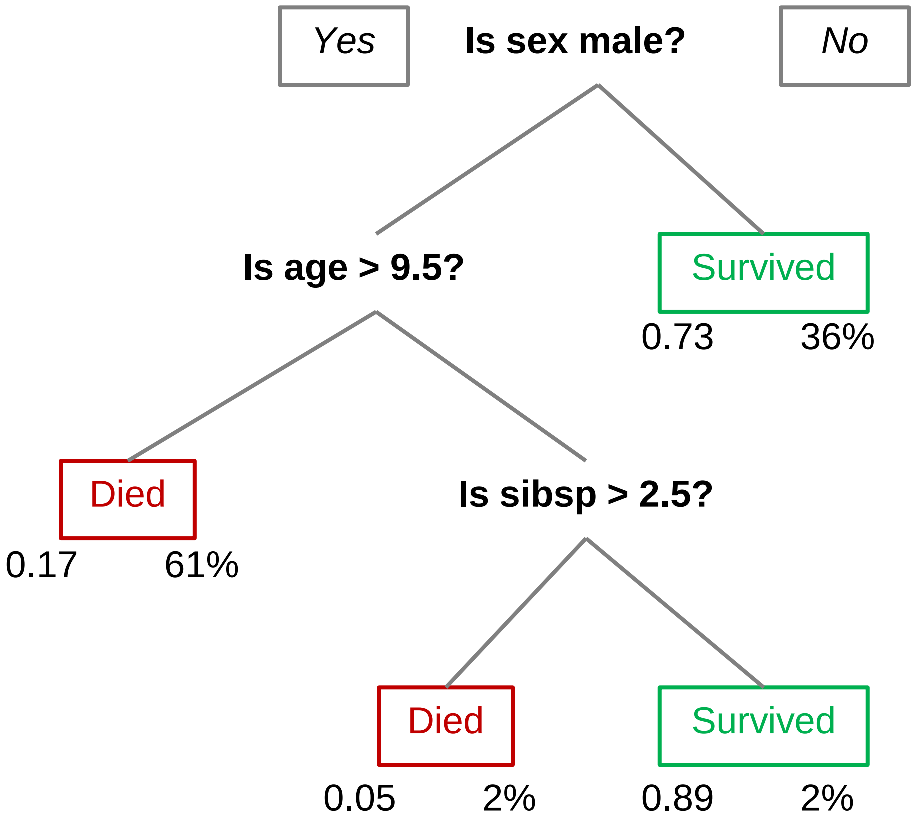
Survived

0.73 36%

Died

0.17 61%

Build tree split by split,
Find the best split you can at each step



Strengths of decision tree methods

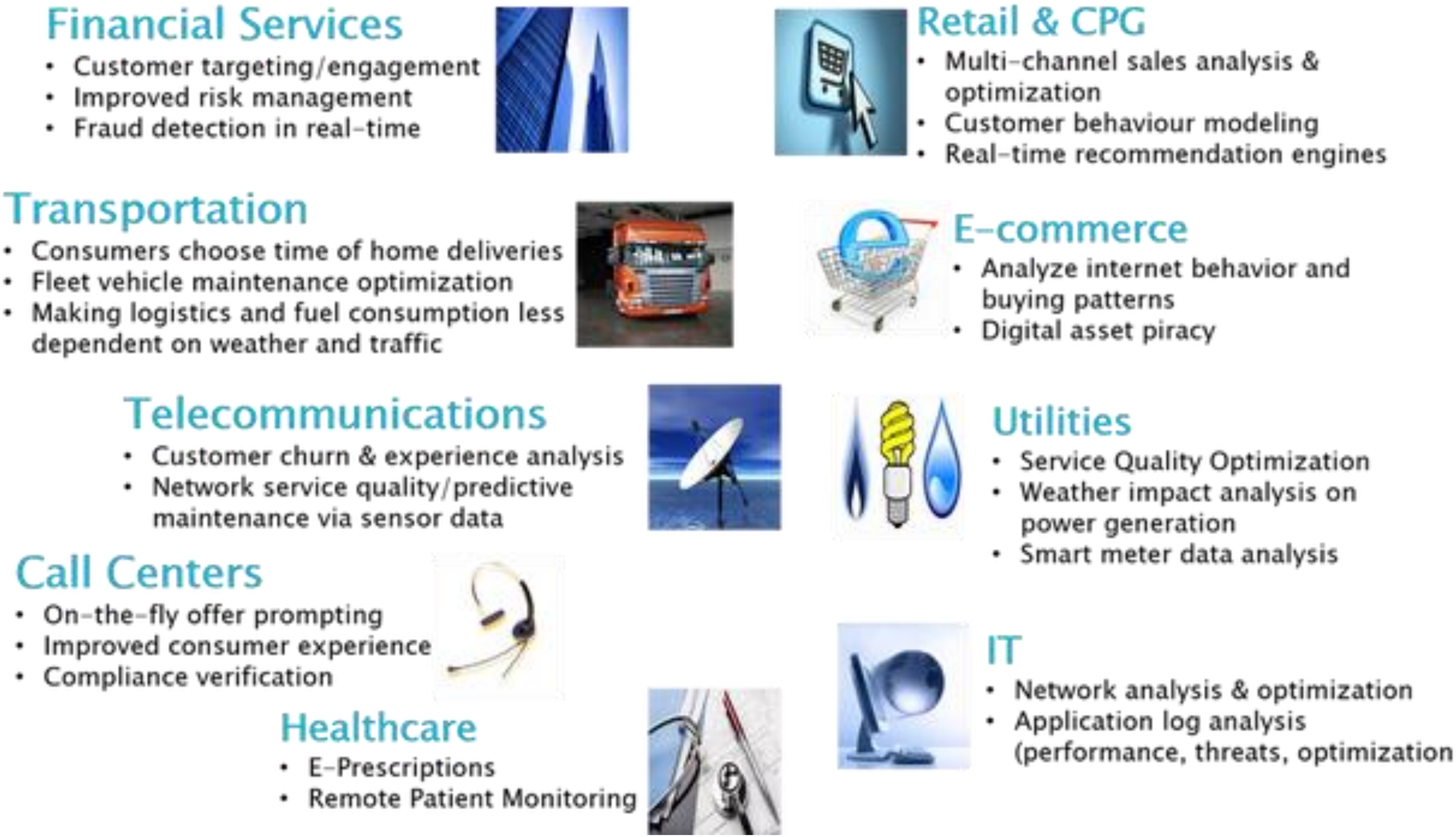
- Generates understandable rules.
- Perform classification without requiring much computation.
- able to handle both continuous and categorical variables.
- Provides a clear indication of which fields are most important for prediction or classification.
- Natural multiclass classifier.

Weaknesses of decision tree

- It is less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Prone to errors in classification problems with many class and relatively small number of training examples.
- Computationally expensive to train.
  - Growing a decision tree is computationally expensive.
  - At each node, each candidate splitting field must be sorted before its best split can be found.
- Small changes in input data can result in totally different trees.
- Can make mistakes with unbalanced classes.

Let’s dive straight to the Hands-on using Jupyter notebooks

What are the industry use cases of Machine Learning?



Financial Services

- Customer targeting/engagement
- Improved risk management
- Fraud detection in real-time

Transportation

- Consumers choose time of home deliveries
- Fleet vehicle maintenance optimization
- Making logistics and fuel consumption less dependent on weather and traffic

Telecommunications

- Customer churn & experience analysis
- Network service quality/predictive maintenance via sensor data

Call Centers

- On-the-fly offer prompting
- Improved consumer experience
- Compliance verification

Healthcare

- E-Prescriptions
- Remote Patient Monitoring

Retail & CPG

- Multi-channel sales analysis & optimization
- Customer behaviour modeling
- Real-time recommendation engines

E-commerce

- Analyze internet behavior and buying patterns
- Digital asset piracy

Utilities

- Service Quality Optimization
- Weather impact analysis on power generation
- Smart meter data analysis

IT

- Network analysis & optimization
- Application log analysis (performance, threats, optimization

Source: bigdata4analytics

Quick Recap

At least eighty percent of the time spent on a Web-based data mining project is devoted to this

- interpretation of results
- data mining
- goal identification
- data preparation

Which statement is true about the K-Means algorithm?

- All attribute values must be categorical
- The output attribute must be cateogrical
- Attribute values may be either categorical or numeric
- All attributes must be numeric

The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75.

What can be said about employee salary and years worked?

- Individuals that have worked for the company the longest have higher salaries
- There is no relationship between salary and years worked
- Individuals that have worked for the company the longest have lower salaries.
- The majority of employees have been with the company a long time.

Simple regression assumes a ___________ relationship between the input attribute and output attribute

- quadratic
- reciprocal
- inverse
- linear

A correlation coefficient enables you to:

- establish whether the data is telling you what you think it should tell you.
- quantify the strength of the linear relationship between two ranked or quantifiable variables.
- assess whether two variables measure the same phenomenon.
- measure the difference between two variables.

Exploratory Data Analysis (EDA) is:

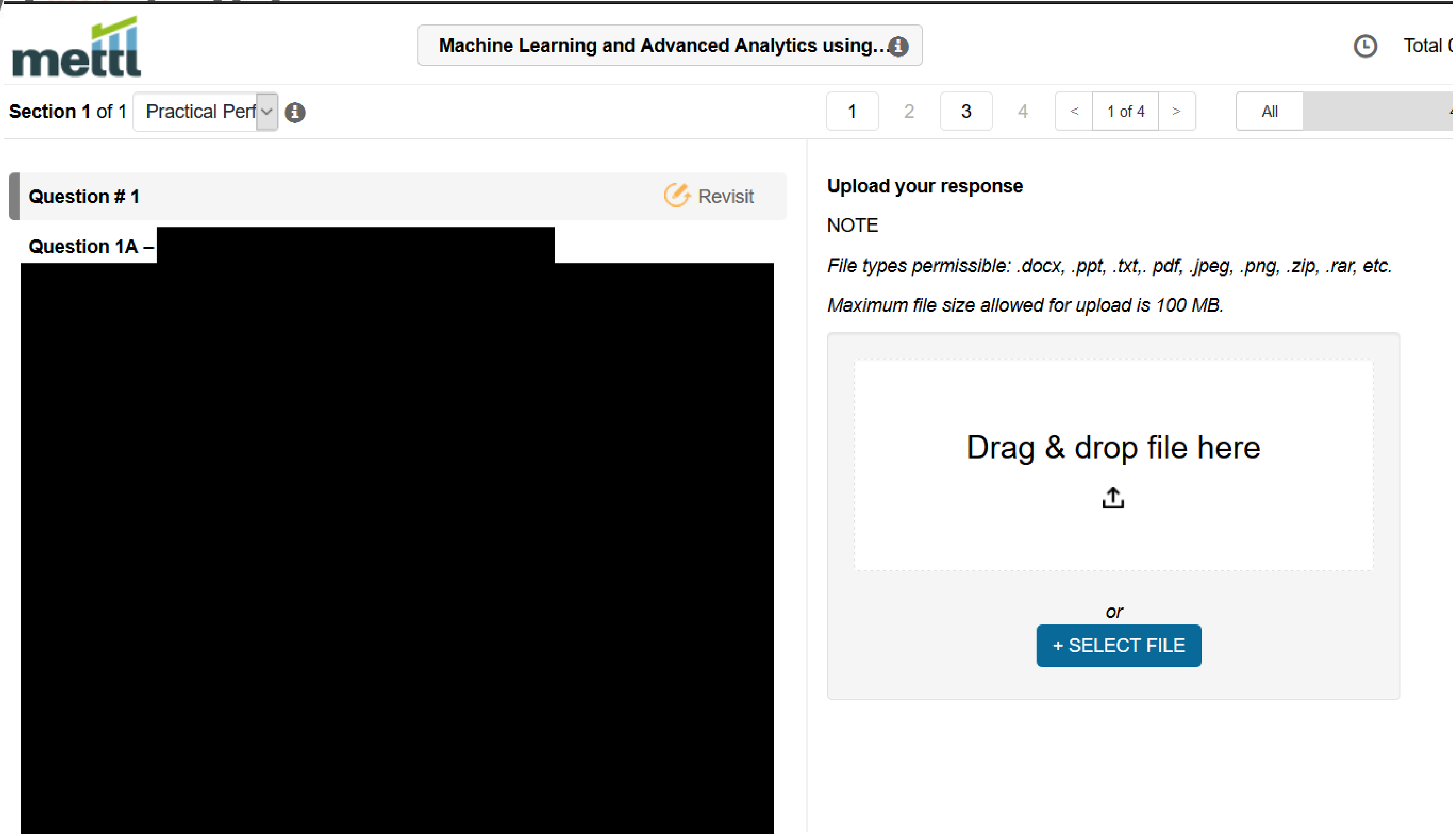
- A set of statistical methods specially designed for exploring a small, unruly data set and identifying any abnormalities in distribution or highly unusual scores
- The stage at which the data are described by the traditional measures of central tendency, spread and distribution shape
- Especially appropriate for nominal data
- Of limited value because no formal statistical tests are made

The average squared difference between classifier predicted output and actual output

- mean squared error
- root mean squared error
- mean absolute error
- mean relative error

Test Instructions (Mettl)

- 2 Parts to the test
  - MCQ 30 mins – Mettl Link will be provided
    - 16 questions
  - Practical 90 mins - Mettl Link will be provided
    - 4 questions
    - The question is for you to submit your juypter notebook files! Submit 2 files only.
      - **Q1A and Q1B – 1 jupyter notebook**
      - **Q2A and Q2B – 1 jupyter notebook**
- After Done with test – Come back to zoom
- Breakout rooms
- Recovery
- Evaluation
- End



Question # 1

Question 1A – [redacted]

Upload your response

NOTE

File types permissible: .docx, .ppt, .txt,. pdf, .jpeg, .png, .zip, .rar, etc.

Maximum file size allowed for upload is 100 MB.

Drag & drop file here

or

+ SELECT FILE