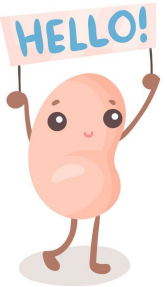
L1: Data Science Foundations Intro

Aneta Havlínová, Justina Ivanauskaite,
Thomas Browne

4th of April, 2023

WELCOME TO THE COURSE!



Tuesdays 18:00-21:00:

4.4. Data Science Introduction (Aneta, Justina, Thomas)

18.4. Data Pre-processing (Thomas, Josef, Andrea)

25.4. Hypotheses Testing with Linear Regression (Michal, Aneta) + HW assignment

9.5. Panel Data Analysis (Pavel, Michal)

16.5. Regression Tasks (Justina, Martin, Josef)

30.5. Classification Tasks part I (Justina, Pavel, Martin, Anna) + HW assignment

6.6. Classification Tasks part II (Anna, Josef, Martin)

20.6. Time Series Analysis (Michal, Aneta)

27.6. Unsupervised Learning (Thomas, Michal) + **CLOSING EVENT** in MSD

czechitas

ORGANISATION

Lectures: theoretical part introduced + demo of code examples

Coding: very basics of Python expected

Homeworks:

- two assignments, at least one to be submitted
- simple task similar to ones showed during the lectures, connected to specific topic: you should write a simple code and describe the results
- after 3rd and 6th lecture
- three weeks to complete

LECTURERS INTRO

Aneta Havlínová
Workday

Main lecturer to contact
aneta.havlinova@protonmail.com

Justina Ivanauskaite
MSD

Pavel Fišer
MSD

Andrea Štefancová
MSD

Michal Hakala
MSD

Anna Štrobová
MSD

Martin Koryťák
Workday

Josef Švec
Workday

Thomas Browne
Kiwi.com

PARTICIPANTS INTRO - we want to get to know you a bit :)

QUESTION 1: What is your **age**?

a) up to 20 years
b) 21-26 years
c) 27-35 years
d) >35 years

czechitas

PARTICIPANTS INTRO - we want to get to know you a bit :)

QUESTION 2: What is your experience with **statistics/data science**?

a) none
b) university
c) courses/projects
d) some working experience
e) self-study
f) multiple of the above

czechitas

PARTICIPANTS INTRO - we want to get to know you a bit :)

QUESTION 3: What is your experience with **IT/coding**?

a) none
b) university
c) courses/projects
d) some working experience
e) self-study
f) multiple of the above

czechitas

PARTICIPANTS INTRO - we want to get to know you a bit :)

QUESTION 4: Starting this course, you **feel**:

a) excited
b) scared
c) neutral
d) mixed feelings :)

czechitas

LET’S GET STARTED :)

TODAY'S LECTURE

1. Introduction & business aspects
2. Use case examples
3. Data Structures and Terminology
4. Data Science Areas

    4.1 Hypotheses Testing
    4.2 Unsupervised Learning
    4.3 Supervised Learning

5. Statistics vs. Machine Learning
6. Models Evaluation

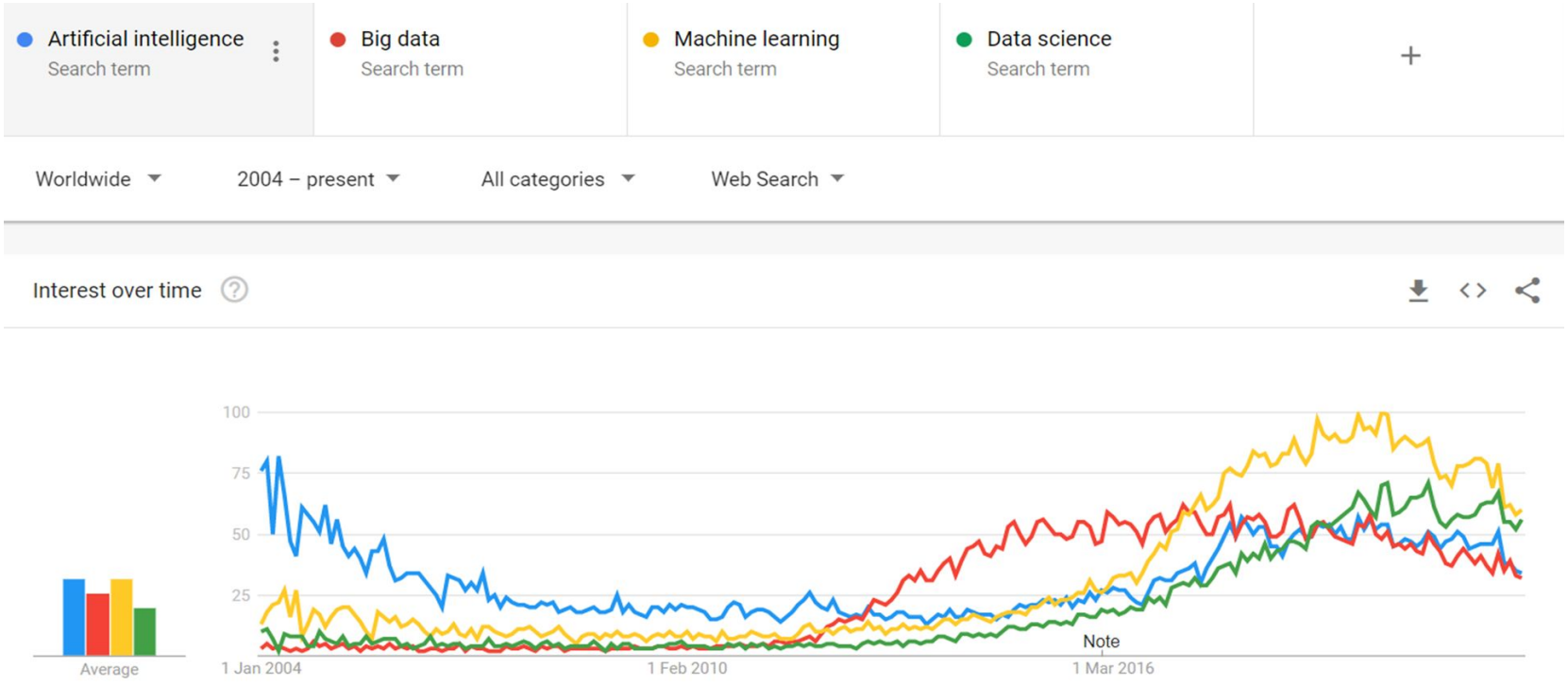
Anet

Justina

Thomas

czechitas

Buzzwords might change but basics stay the same



czechitas

What is data science?

- Nowadays, people often think it is about big data, machine learning, neural networks
- But those are only inputs, methods
- **The main aim is to deliver value through data**

Delivering business value depends on your ability to answer clients' questions with data

01 Decisions

+ Successful analytical projects are those that support decision making or automation

02 Questions

+ Key is to understand what questions need to be answered to make a decision

03 Approach

+ We choose analytical approach that is most suitable to answer these questions

04 Technology

+ It is important to pay attention to technology allowing implementation into the decision process

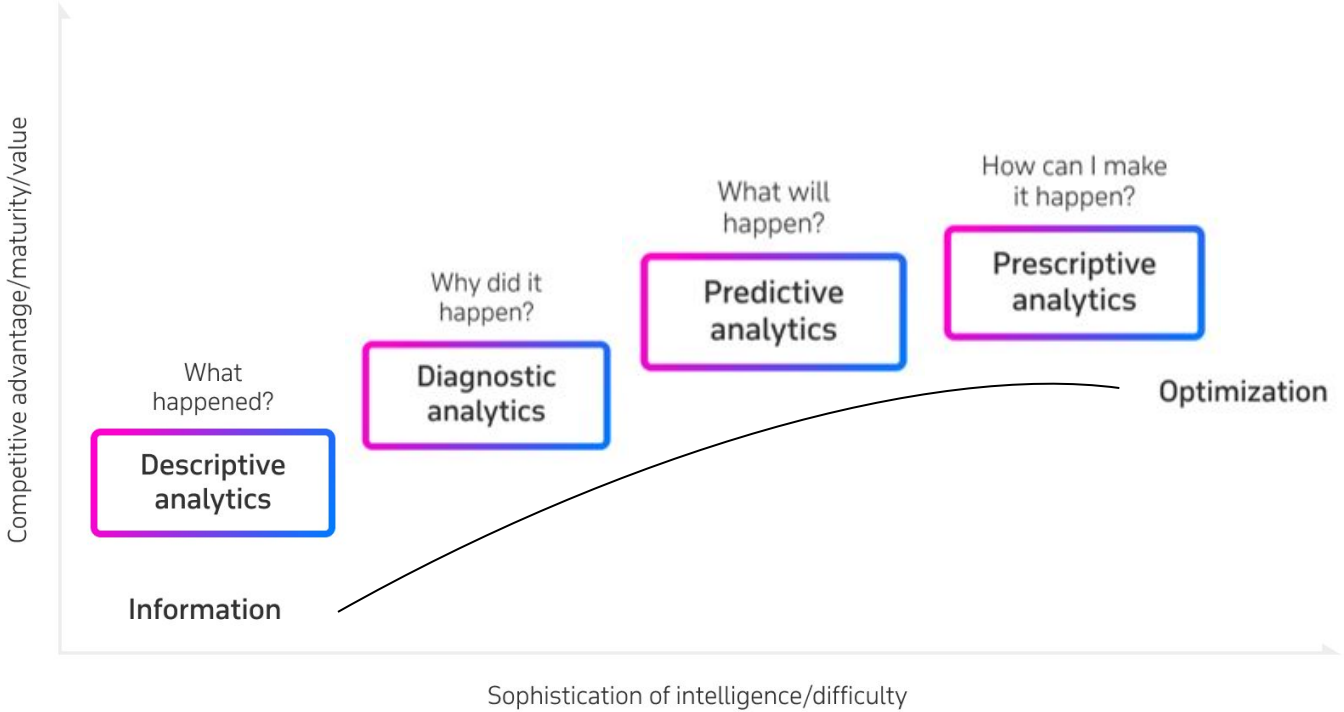
05 Implementation

+ We ensure, in cooperation with IT, implementation of the solution and take care of the maintenance process

Focus of the course

czechitas

Different types of analytics solve different business needs



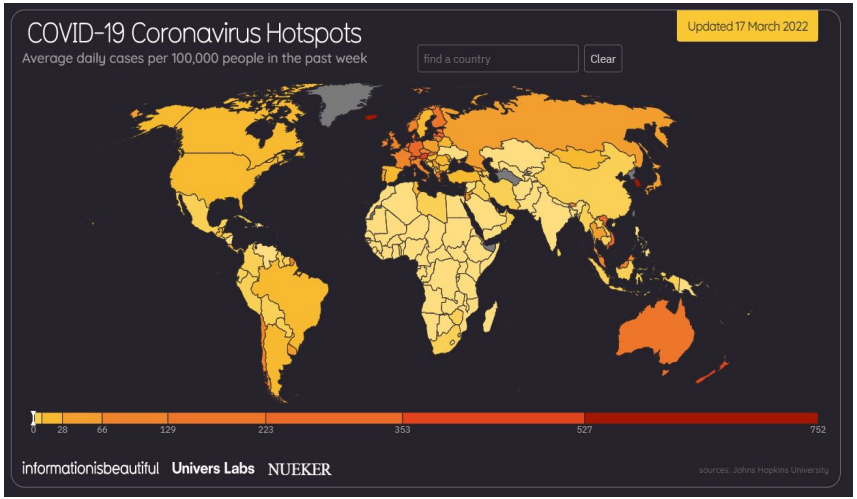
czechitas

Interactive visualization are used commonly to present complex data and allow for certain level of user self-service

Interactive reporting can serve to:

- Present complex data in user appealing form
- Let user decide what to have in the visualization
- Uncover hidden pattern
- Bring new questions, new ideas

- **Interactive visualizations are usually accessible through web interface which simplifies user experience**

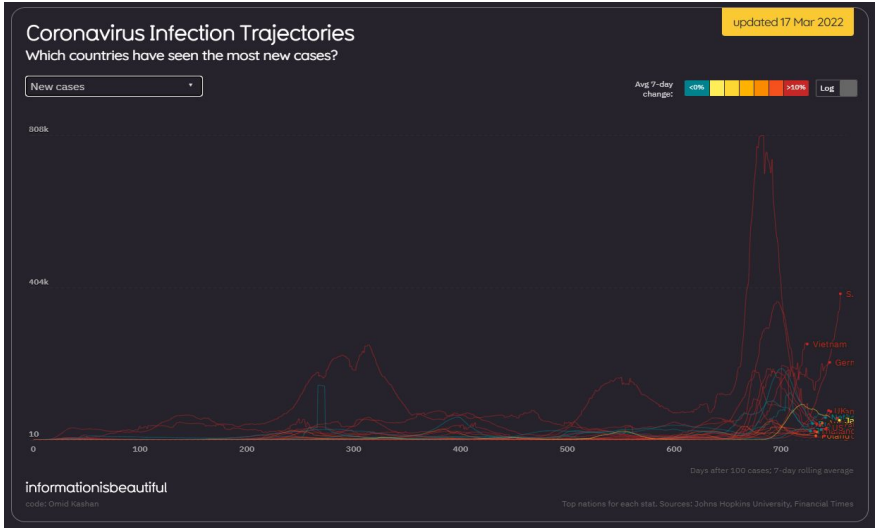


In diagnostic analytics we want to uncover hidden factors that drive certain behavior – for example different countries death toll

There are questions answered by diagnostic analytics:

- What are factors driving observed development?
- Which factors have strongest impact?
- Are there clusters of countries that are similar?
- What is the impact of vaccination?

- **There are several statistical methods that can help us uncover hidden relationships and we will cover them in this course in detail**

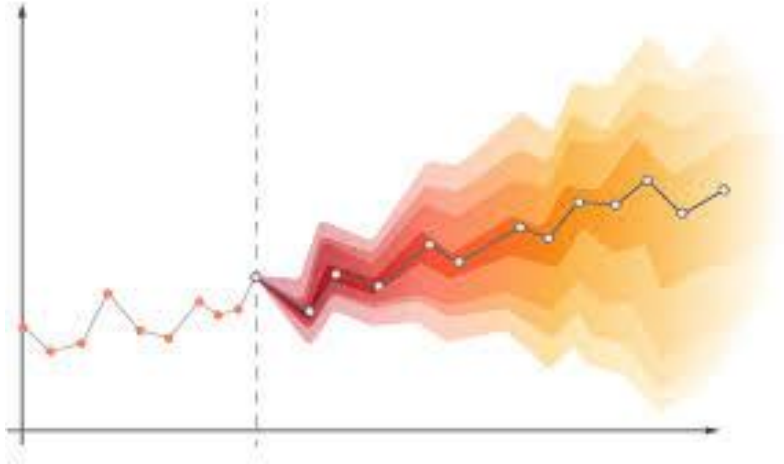


Forecasting techniques allow to predict chosen KPI(s) outlook or even probability of customer actions

Staying with Covid use-case, uncovered dependences allow us to forecast future development:

- What can we expect to happen in the Fall 22?
- What would happen if vaccination increases by 10%?
- If new variant is discovered with twice the infection and death rate, what could we expect?

- **Again, there are multiple analytical approaches, that can be used in forecasting tasks, both from the time-series as well as machine learning catalogue.**

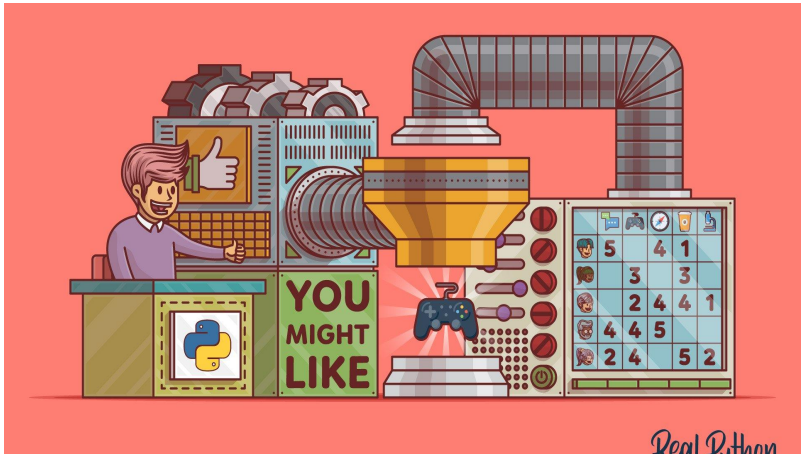


Prescriptive analytics accounts for constraints when making decisions

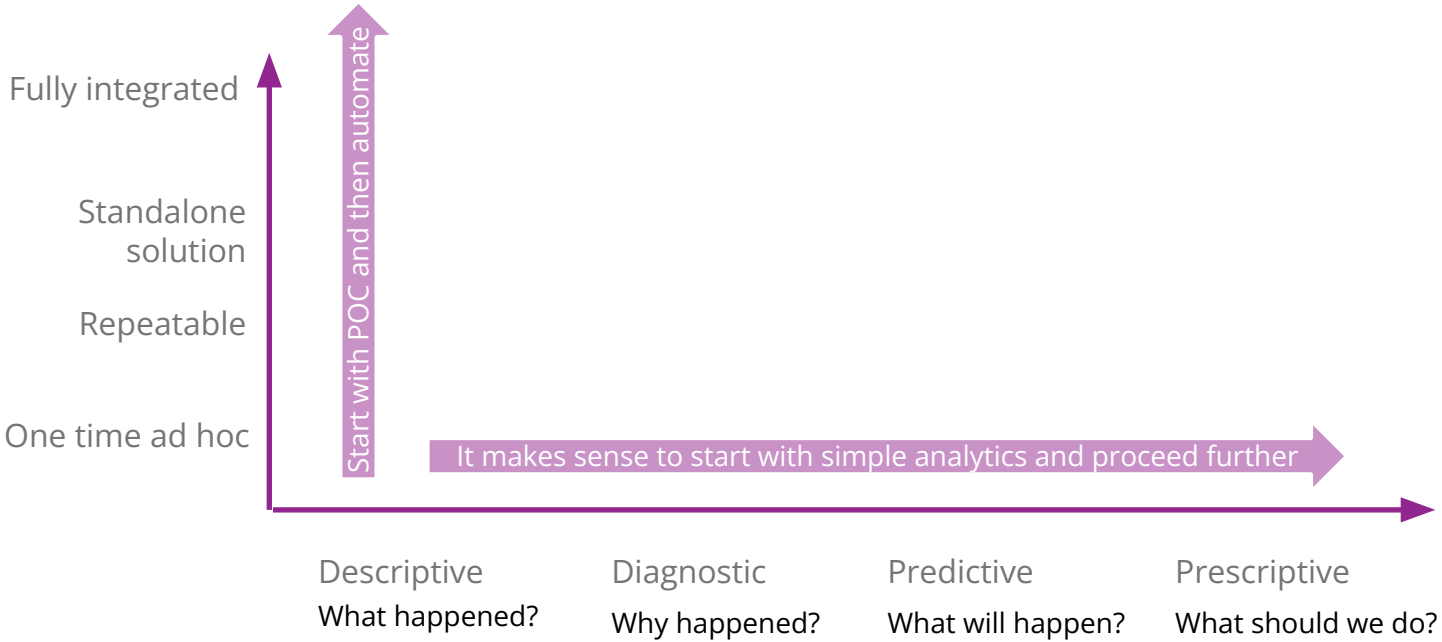
Still related to Covid use-case, countries optimize their response to Covid pandemic:

- With constrained initial supplies of vaccines, who should get it first?
- Should people get 3rd dose or rather focus on new vaccinations
- Who should get treatment first if resources are limited?

- **There are some optimization techniques, but usually these use-cases are so specialized, that they require specific tailored made approach.**

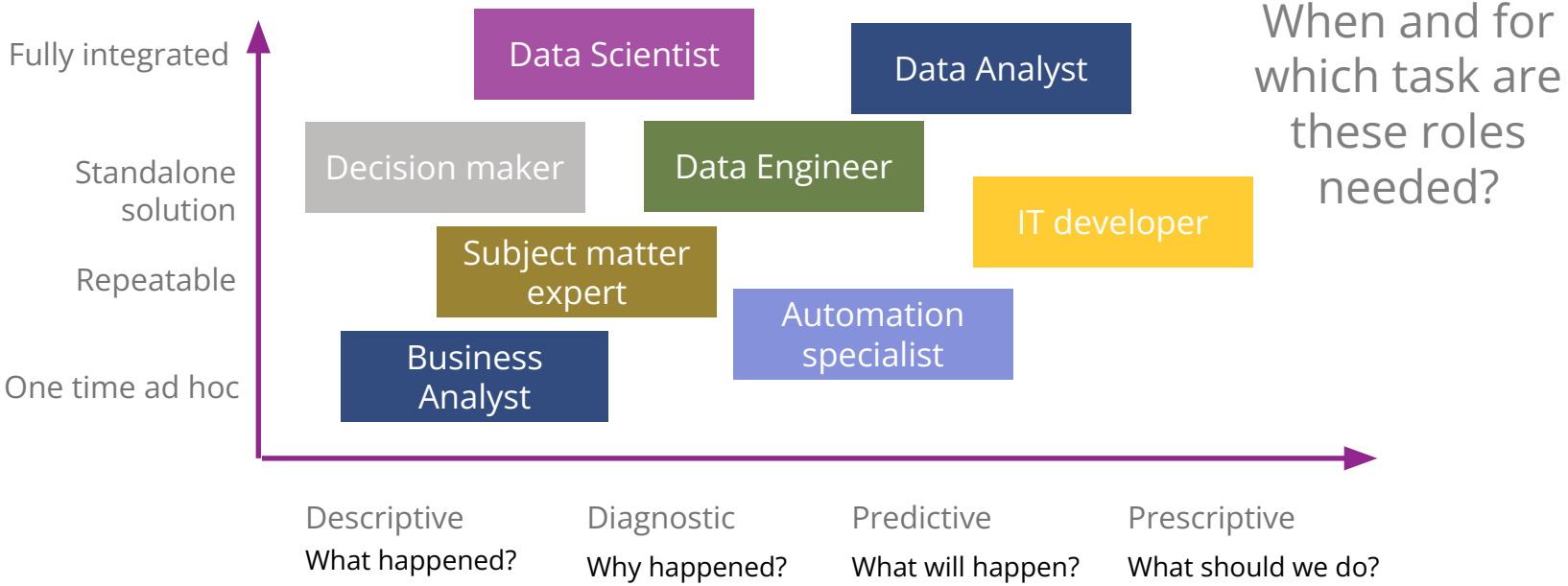


Different types of analytics



czechitas

Important roles for a data science project



czechitas

DATA SCIENCE USE CASES EXAMPLES

Situation: You are an academic who is doing research on impact of gender on salary. How would you approach this problem? What data do you need?

Type of Analytics	
Required data	
POC vs Automation	

czechitas

Situation: You are working for an airline company. You are tasked to predict flight delays. How would you approach this tasks?

Type of Analytics	
Required data	
POC vs Automation	

Situation: You are working for a major FMCG retailer. You want to understand who is your customer to better approach him through commercial or promo offers.

Type of Analytics	
Required data	
POC vs Automation	

czechitas

3. DATA STRUCTURES AND TERMINOLOGY

DATA = TABLE

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
…	…	…	…	…

Data on Employees' Income

DATA = TABLE

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Data Table:

→ Columns

DATA = TABLE

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

VALUES = DATA

Data Table:

→ Columns

czechitas

DATA = TABLE

HEADER = METADATA

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Data Table:

→ Columns

DATA = TABLE

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

ROW = OBSERVATION

Data Table:

→ Columns
→ Observations

czechitas

DATA STRUCTURE - WIDE

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
…	…	…	…	…

Typically: One column = one variable

→ Wide data format, **1 OBS = All Employee Information**

DATA STRUCTURE - LONG

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

PersonID	Variable	Value
2343	Gender	F
2343	Income	50000
2343	Education	17
2343	Age	35
1213	Gender	M
1213	Income	35000
1213	Education	15
...	...	...

Sometimes: Variable names in one column, values in another column

→ Long data format, **1 OBS = PersonID + Variable**

Usage: data processing in bulk (e.g., conversion to numeric), visualizations of multiple variables in one plot

czechitas

DATA STRUCTURE - WIDE vs LONG

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

PersonID	Variable	Value
2343	Gender	F
2343	Income	50000
2343	Education	17
2343	Age	35
1213	Gender	M
1213	Income	35000
1213	Education	15
...	...	...

→ Wide data format, **1 OBS = All Employee Information**

→ Long data format, **1 OBS = PersonID + Variable**

Contain **similar DATA**, Choice WRT GOAL

czechitas

DATA TYPES

1) CROSS-SECTIONAL DATA

PersonID	Income	Years Education
2343	50 000	17
1213	35 000	15
4533	40 000	15
4563	100 000	19
7554	50 000	18
6465	27 500	13
7453	34 000	13
...	...	...

- Characterized by individual units - people, companies, countries, ...

DATA TYPES

1) CROSS-SECTIONAL DATA

PRIMARY KEY

PersonID	Income	Years Education
2343	50 000	17
1213	35 000	15
4533	40 000	15
4563	100 000	19
7554	50 000	18
6465	27 500	13
7453	34 000	13
...	...	...

- Characterized by individual units - people, companies, countries, ...

DATA TYPES

2) TIME SERIES

Day	Microsoft share price (USD)
01/01/2020	301
02/01/2020	303.2
03/01/2020	302
04/01/2020	311.5
05/01/2020	312.6
06/01/2020	309.1
07/01/2020	311.4
...	...

- Data collected at several time points
- Stock prices, interest rates, exchange rates, GDP,...
- Many different frequencies (hourly, daily, monthly, quarterly,..)

DATA TYPES

2) TIME SERIES

PRIMARY KEY

Day	Microsoft share price (USD)
01/01/2020	301
02/01/2020	303.2
03/01/2020	302
04/01/2020	311.5
05/01/2020	312.6
06/01/2020	309.1
07/01/2020	311.4
...	...

- Data collected at several time points
- Stock prices, interest rates, exchange rates, GDP,...
- Many different frequencies (hourly, daily, monthly, quarterly,..)

DATA TYPES

3) PANEL DATA (LONGITUDINAL)

Company	Date	Net Income After Taxes
Apple	10/24/2020	28,755,000,000
Apple	7/25/2020	12,673,000,000
Apple	4/25/2020	11,253,000,000
Apple	1/25/2020	11,249,000,000
Cisco	10/24/2020	2,174,000,000
Cisco	7/25/2020	2,636,000,000
Cisco	4/25/2020	2,774,000,000
Cisco	1/25/2020	2,878,000,000

- Combines cross-sectional and time series data
- The same individuals (persons, firms, cities, etc.) are observed at several points in time (days, years, ..)

DATA TYPES

3) PANEL DATA (LONGITUDINAL)

PRIMARY KEY

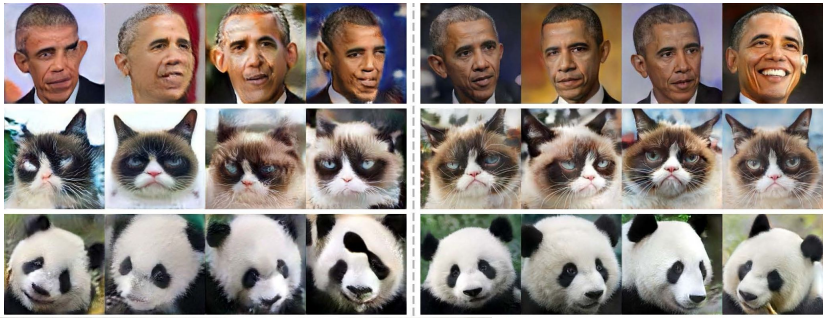
Company	Date	Net Income After Taxes
Apple	10/24/2020	28,755,000,000
Apple	7/25/2020	12,673,000,000
Apple	4/25/2020	11,253,000,000
Apple	1/25/2020	11,249,000,000
Cisco	10/24/2020	2,174,000,000
Cisco	7/25/2020	2,636,000,000
Cisco	4/25/2020	2,774,000,000
Cisco	1/25/2020	2,878,000,000

- Combines cross-sectional and time series data
- The same individuals (persons, firms, cities, etc.) are observed at several points in time (days, years, ..)

DATA TYPES

4) OTHER DATA TYPES

- Images
- Videos
- Text (news articles etc.)



Merkel should tell Obama to wipe her NSA files, opposition urges

Critics say credibility of western alliance is at risk

Chancellor's office lowers expectations of US visit

Philip Oltermann Berlin

Ukip likely to come out ahead in Europe poll

StyleGAN2 + DiffAugment

czechitas

DATA TYPES

4) OTHER DATA TYPES: Images & videos to numbers

=

Red, Green, Blue (RGB) array

(232, 190, 172) (232, 190, 169) (232, 191, 170) ...

(230, 183, 172) (232, 190, 172) (230, 189, 172)

...

..

...

Video = a sequence of images = a sequence of RGB arrays

czechitas

DATA TYPES

4) OTHER DATA TYPES:

=

Turned into Table!

Red, Green, Blue (RGB) array

(232, 190, 172)	(232, 190, 169)	(232, 191, 170)	…
(230, 183, 172)	(232, 190, 172)	(230, 189, 172)	
…			
..			
…			

Video = a sequence of images = a sequence of RGB arrays

czechitas

DATA TERMINOLOGY

Example: predict income based on person's age, education, and gender

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
…	…	…	…	…

Income is :

- Predicted variable
- **Dependent variable**
- **Target**
- Response variable
- Output variable
- Label

czechitas

DATA TERMINOLOGY

Example: predict income based on person's age, education, and gender

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
…	…	…	…	…

Income is :

- Predicted variable
- **Dependent variable**
- **Target**
- Response variable
- Output variable
- Label

Age, Education, and Gender are:

- Predictor variables
- **Independent variables**
- **Explanatory variables**
- **Features**
- Input variables
- Regressors
- Control variables

czechitas

DATA TERMINOLOGY

Example: predict income based on person's age, education, and gender

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
…	…	…	…	…

Income is :

- Predicted variable
- **Dependent variable**
- **Target**
- Response variable
- Output variable
- Label

Age, Education, and Gender are:

- Predictor variables
- **Independent variables**
- **Explanatory variables**
- **Features**
- Input variables
- Regressors
- Control variables

czechitas

DATA TERMINOLOGY: MODEL

Example: predict income based on person's age, education, and gender

PersonID	Gender	Income	Years Education	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
…	…	…	…	…

Income is :

- Predicted variable
- **Dependent variable**
- **Target**
- Response variable
- Output variable
- Label

Age, Education, and Gender are:

- Predictor variables
- **Independent variables**
- **Explanatory variables**
- **Features**
- Input variables
- Regressors
- Control variables

COLUMN TYPES

PersonID	Gender	Income	Years Education	Age	Birth Place
2343	F	50 000	17	35	SK
1213	M	35 000	15	32	LTU
4533	M	40 000	15	53	FR
4563	M	100 000	19	51	CZ
7554	F	50 000	18	28	CZ
6465	F	27 500	13	25	UK
7453	M	34 000	13	32	US
...	...	...	...	...	

COLUMN TYPES

PersonID	Gender	Income	Years Education	Age	Birth Place
2343	F	50 000	17	35	SK
1213	M	35 000	15	32	LTU
4533	M	40 000	15	53	FR
4563	M	100 000	19	51	CZ
7554	F	50 000	18	28	CZ
6465	F	27 500	13	25	UK
7453	M	34 000	13	32	US
...	...	...	...	...	

NUMERICAL

COLUMN TYPES

PersonID	Gender	Income	Years Education	Age	Birth Place
2343	F	50 000	17	35	SK
1213	M	35 000	15	32	LTU
4533	M	40 000	15	53	FR
4563	M	100 000	19	51	CZ
7554	F	50 000	18	28	CZ
6465	F	27 500	13	25	UK
7453	M	34 000	13	32	US
...	...	...	...	...	

STRING/FACTOR **STRING/FACTOR**

COLUMN TYPES

PersonID	Gender	Income	Years Education	Age	Birth Place
2343	F	50 000	17	35	SK
1213	M	35 000	15	32	LTU
4533	M	40 000	15	53	FR
4563	M	100 000	19	51	CZ
7554	F	50 000	18	28	CZ
6465	F	27 500	13	25	UK
7453	M	34 000	13	32	US
...	...	...	...	...	

NUMERICAL?
STRING?

ONE-HOT ENCODING OF NON-NUMERIC DATA

PersonID	Gender
2343	F
1213	M
4533	M
4563	M
7554	F
6465	F
7453	M
...	...

PersonID	GenderMale
2343	0
1213	1
4533	1
4563	1
7554	0
6465	0
7453	1
...	...

Categorical variables need to be transformed into numeric ones.

Usually, one category is dropped (use only male gender here) to avoid multicollinearity.

4. DATA SCIENCE AREAS

DATA SCIENCE PROBLEM TYPES OVERVIEW

HYPOTHESES TESTING

- With cross sectional data
- With panel data

SUPERVISED LEARNING

- Classification
- Regression

DATA SCIENCE

OTHER

- Image, Video processing
- Anomaly detection
- Optimizations
- Simulations

UNSUPERVISED LEARNING

- Dimensionality reduction
- Clustering

DATA SCIENCE PROBLEM TYPES OVERVIEW

STATISTICS

HYPOTHESES TESTING

- With cross sectional data
- With panel data

SUPERVISED LEARNING

- Classification
- Regression

OTHER

- Image, Video processing
- Anomaly detection
- Optimizations
- Simulations

UNSUPERVISED LEARNING

- Dimensionality reduction
- Clustering

czechitas

DATA SCIENCE PROBLEM TYPES OVERVIEW

STATISTICS

HYPOTHESES TESTING

- With cross sectional data
- With panel data

MACHINE LEARNING

SUPERVISED LEARNING

- Classification
- Regression

UNSUPERVISED LEARNING

- Dimensionality reduction
- Clustering

OTHER

- Image, Video processing
- Anomaly detection
- Optimizations
- Simulations

4.1 HYPOTHESES TESTING

HYPOTHESES - examples

1. Lack of feedback has a negative impact on employees' retention.

2. Women are less likely to get a promotion.

3. Job profile has an significant impact on employees' attrition.

HYPOTHESES TESTING

Personal ID	Gender	Income	Education (years)	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Assumption to test: Number of years of education have a statistically significant impact on income.

HYPOTHESES TESTING

Personal ID	Gender	Income	Education (years)	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Assumption to test: Women have smaller income.

HYPOTHESES TESTING

Personal ID	Gender	Income	Education (years)	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Assumption to test: Women have smaller income.

***mean*(Income_F) < *mean*(Income_M) ?**

czechitas

HYPOTHESES TESTING

Personal ID	Gender	Income	Education (years)	Age
2343	F	50 000	17	35
1213	M	35 000	15	32
4533	M	40 000	15	53
4563	M	100 000	19	51
7554	F	50 000	18	28
6465	F	27 500	13	25
7453	M	34 000	13	32
...	...	...	...	...

Methodology:

- Calculate ***mean*(Income_F)** and ***mean*(Income_M)** in table
- Assess Probability for ***mean*(Income_F) = *mean*(Income_M)**
- **P-value,** if too small conclude:

***mean*(Income_F) < *mean*(Income_M)**

czechitas

HYPOTHESES TESTING

***mean*(Income_F) = 42,500**
***mean*(Income_H) = 51,800**

Personal ID	Gender	Income	Education (years)	Age
2343	**F**	**50 000**	17	35
1213	**M**	**35 000**	15	32
4533	**M**	**40 000**	15	53
4563	**M**	**100 000**	19	51
7554	**F**	**50 000**	18	28
6465	**F**	**27 500**	13	25
7453	**M**	**34 000**	13	32
…	**…**	**…**	…	…

Methodology:

- Assume: ***mean*(Income_F) = *mean*(Income_H)**
- Calculate ***mean*(Income_F)** and ***mean*(Income_M)** in table
- Assess Probability for ***mean*(Income_F) = *mean*(Income_M)**
- **P-value,** if too small conclude:

***mean*(Income_F) < *mean*(Income_M)**

czechitas

HYPOTHESES TESTING

Personal ID	Gender	Income	Education (years)	Age
2343	**F**	**50 000**	17	35
1213	**M**	**35 000**	15	32
4533	**M**	**40 000**	15	53
4563	**M**	**100 000**	19	51
7554	**F**	**50 000**	18	28
6465	**F**	**27 500**	13	25
7453	**M**	**34 000**	13	32
…	**…**	**…**	…	…

Methodology:

- Assume: ***mean*(Income_F) = *mean*(Income_H)**
- Calculate ***mean*(Income_F)** and ***mean*(Income_M)** in table
- Assess Probability for ***mean*(Income_F) = *mean*(Income_M)**
- **P-value,** if too small conclude: ***mean*(Income_F) < *mean*(Income_M)**

- ***mean*(Income_F)** = 42,500, ***mean*(Income_H)** = 51,800
- *p-value* ≅ *Probability* (Difference = 9,300)
- *p-value* = 0,01 ⇒ ***mean*(Income_F) < *mean*(Income_H)**

czechitas

4.2 UNSUPERVISED LEARNING

Unsupervised Learning: Clustering

Customer ID	Income	Education	Age	Gender	Last Purchase Amount	Customer Segment
2343	50 000	17	35	M	2500	?
1213	35 000	15	32	F	34000	?
4533	40 000	15	53	F	12000	?
4563	100 000	19	51	M	2100	?
7554	50 000	18	28	M	760	?
6465	27 500	13	25	F	21000	?
7453	34 000	13	32	M	42000	?
6775	72 000	18	43	M	18000	?
4643	50 000	19	47	M	5600	?

Unlabeled Data

Goal:

Identify **groups** of Observation with **Similar Patterns**

Clustering: 4 Clusters

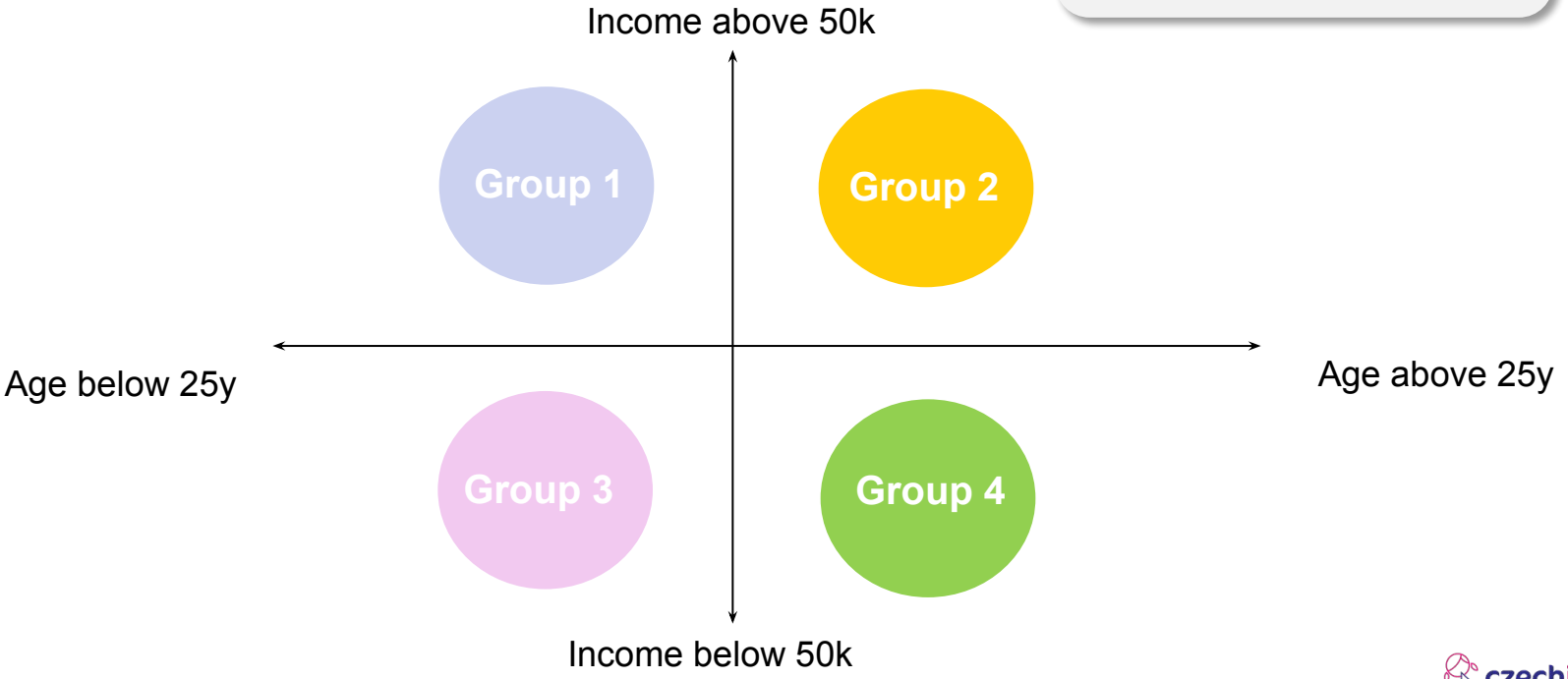
Customer ID	Income	Education	Age	Gender	Last Purchase Amount	Customer Segment
2343	50 000	17	35	M	2500	?
1213	35 000	15	32	F	34000	?
4533	40 000	15	53	F	12000	?
4563	100 000	19	51	M	2100	?
7554	50 000	18	28	M	760	?
6465	27 500	13	25	F	21000	?
7453	34 000	13	32	M	42000	?
6775	72 000	18	43	M	18000	?
4643	50 000	19	47	M	5600	?

Unlabeled Data

czechitas

Clustering: Labelling Unlabelled Data

If you don't have specified classes, you can do the grouping yourself



czechitas

Clustering: Automatic Labelling

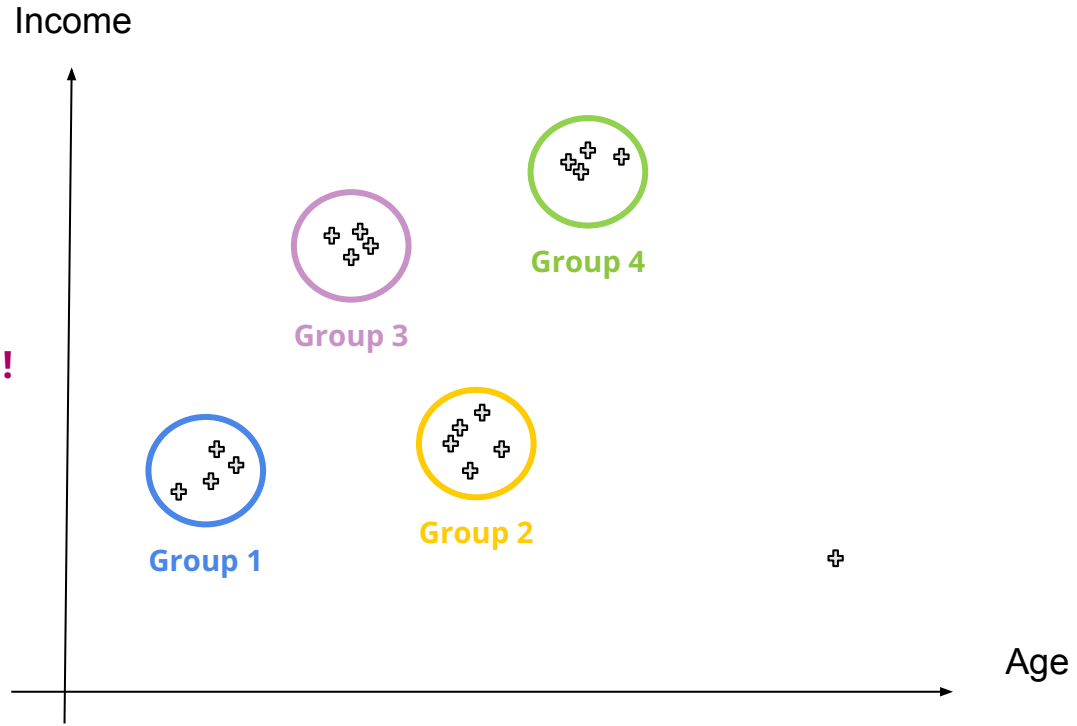
Each + is an **observation**

czechitas

Clustering: Automatic Labelling

Each + is an **observation**

AUTOMATIC CLUSTERING!



czechitas

Clustering: Automatic Labelling

Customer ID	Income	Education	Age	Gender	Last Purchase Amount	Customer Segment
2343	50 000	17	35	M	2500	?
1213	35 000	15	32	F	34000	?
4533	40 000	15	53	F	12000	?
4563	100 000	19	51	M	2100	?
7554	50 000	18	28	M	760	?
6465	27 500	13	25	F	21000	?
7453	34 000	13	32	M	42000	?
6775	72 000	18	43	M	18000	?
4643	50 000	19	47	M	5600	?

Unlabeled Data

czechitas

Anomaly Detection

Each + is an **observation**

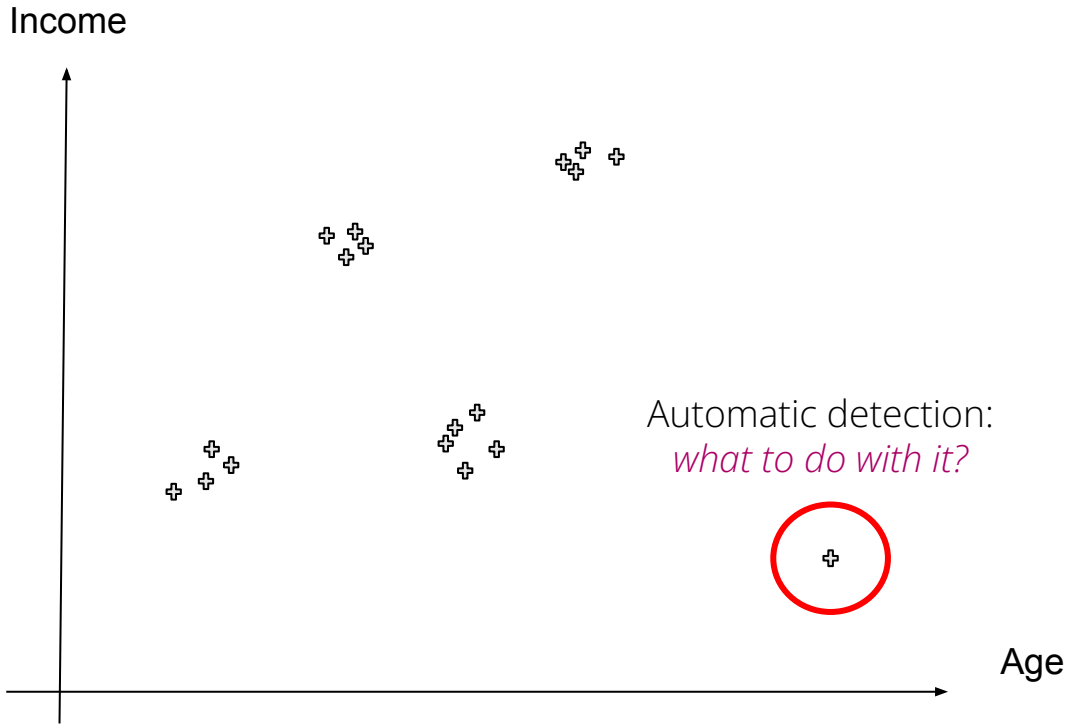
We want to be able to recognise **abnormal points**

czechitas

Anomaly Detection

Each + is an **observation**

We want to be able to recognise **abnormal points**



czechitas

Clustering: Automatic Labelling

Customer ID	Income	Education	Age	Gender	Last Purchase Amount	Customer Segment
2343	50 000	17	35	M	2500	?
1213	35 000	15	32	F	34000	?
4533	40 000	15	53	F	12000	?
4563	100 000	19	51	M	2100	?
7554	50 000	18	28	M	760	?
6465	27 500	13	25	F	21000	?
7453	14 000	65	112	M	42000	?
6775	72 000	18	43	M	18000	?
4643	50 000	19	47	M	5600	?

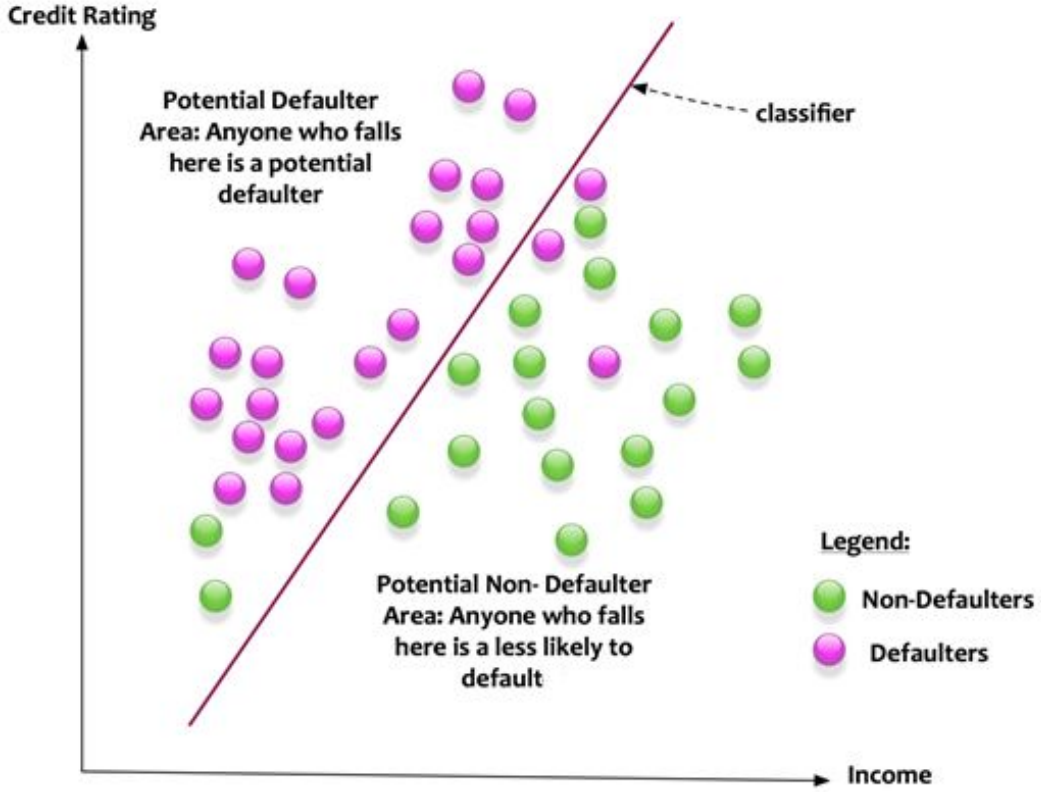
Need to **detect** such observations...

czechitas

4.3 SUPERVISED LEARNING

Classification & Regression

CLASSIFICATION



Credit Default = a binary variable!

czechitas

SUPERVISED LEARNING

CustomerID	Income	Years Education	Age	Default
2343	50 000	17	35	No
1213	35 000	15	32	Yes
4533	40 000	15	53	No
4563	100 000	19	51	No
7554	50 000	18	28	No
6465	27 500	13	25	Yes
7453	34 000	13	32	No
6775	72 000	18	43	No
4643	50 000	19	47	No
6886	48 000	19	37	?
8668	62 500	21	39	?
8765	78 000	23	46	?
9797	23 000	12	29	?

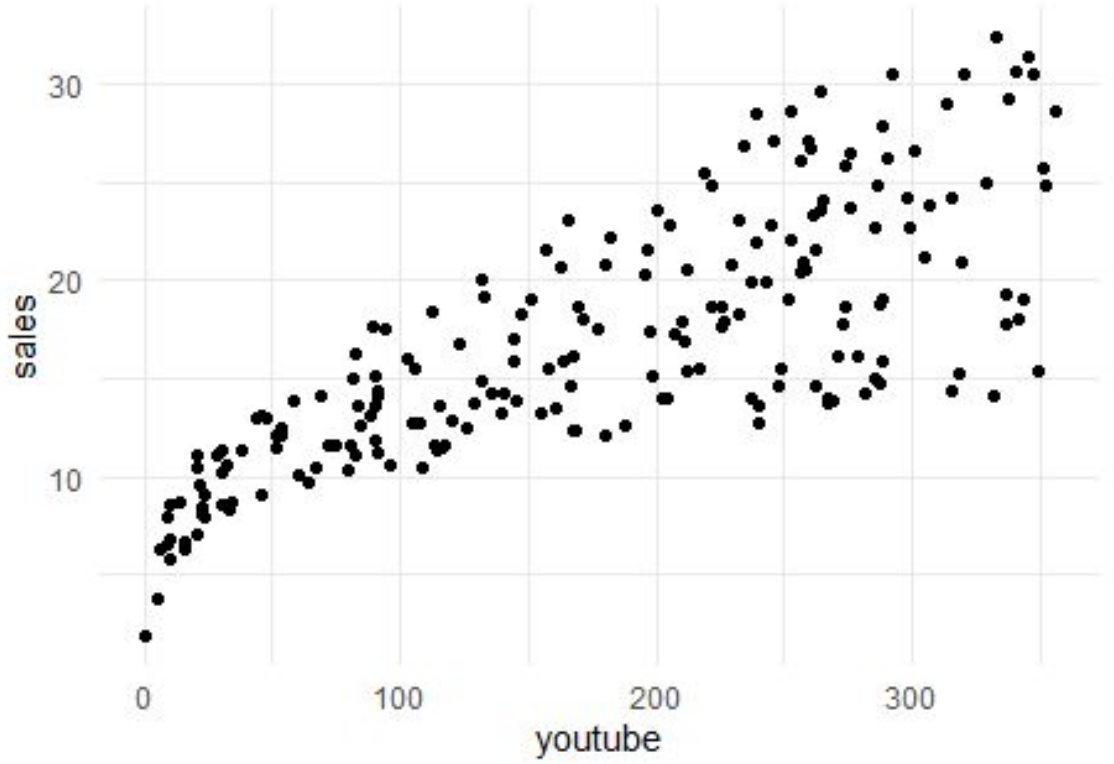
Labeled Data

Unlabeled Data

czechitas

REGRESSION

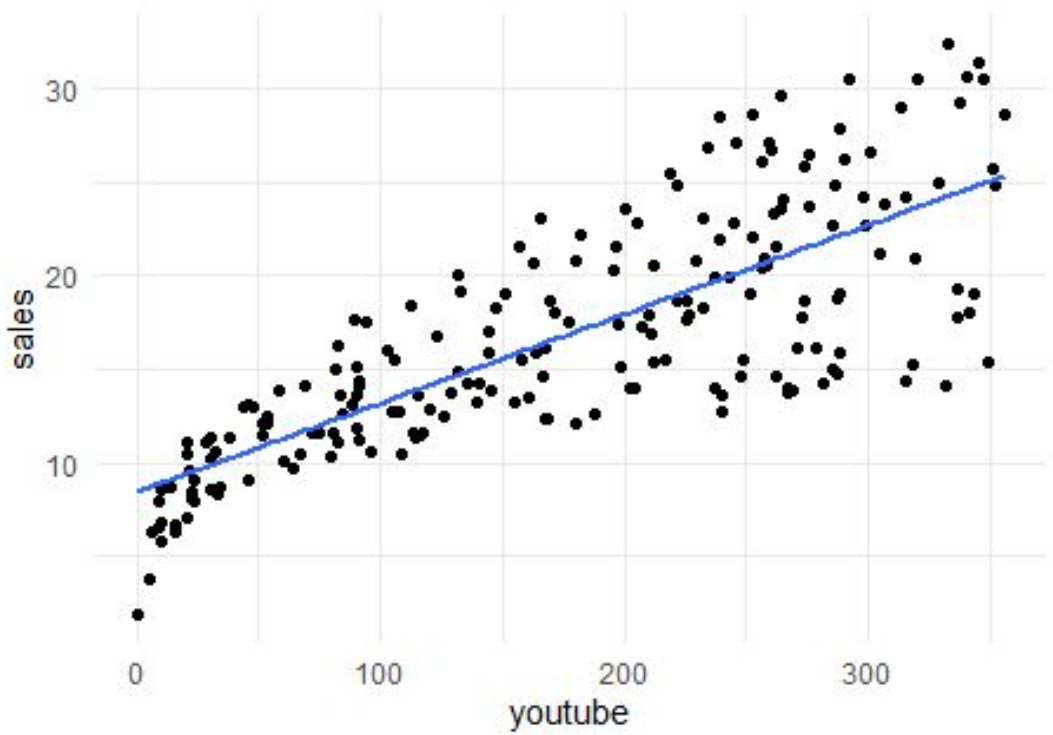
youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76



czechitas

REGRESSION

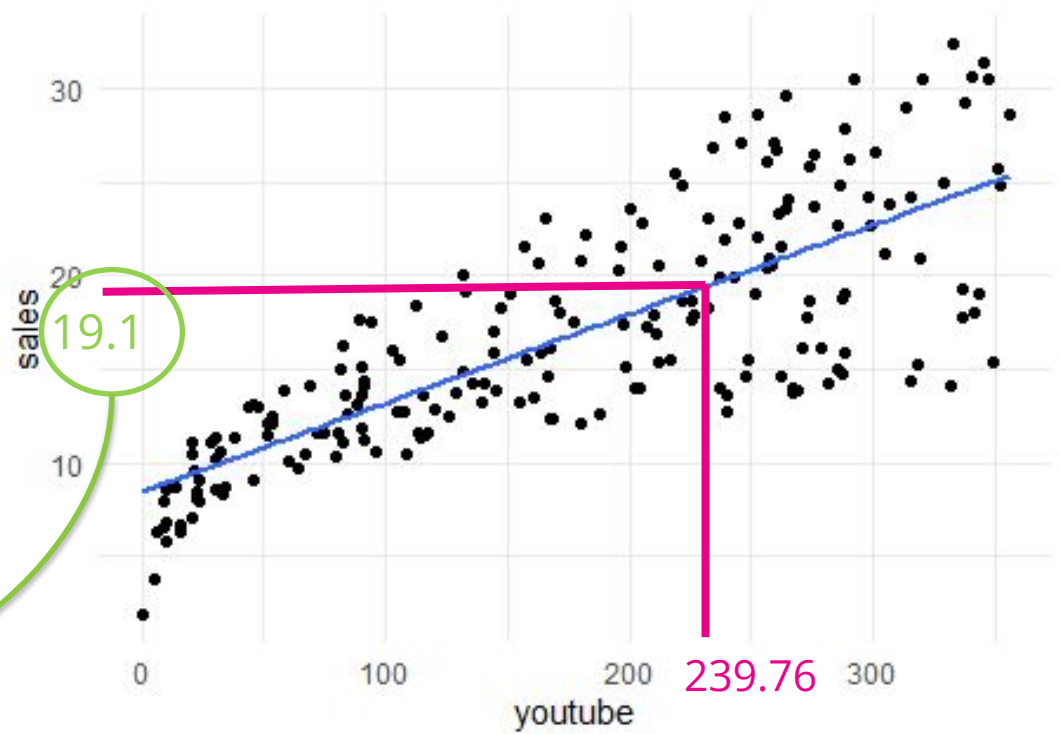
youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76
239.76	?
79.32	?
257.64	?



czechitas

REGRESSION

youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76
239.76	?
79.32	?
257.64	?



czechitas

REGRESSION

youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76
239.76	?
79.32	?
257.64	?

Labeled Data

Unlabeled Data

Sales = a continuous variable!

czechitas

5. STATISTICS VS. MACHINE LEARNING

STATISTICS VS. MACHINE LEARNING

STATISTICS

HYPOTHESES TESTING

- Cross sectional data
- Panel data

MACHINE LEARNING

SUPERVISED LEARNING

- Classification
- Regression

UNSUPERVISED LEARNING

- Dimensionality reduction
- Clustering

OTHER

- Image, Video processing
- Anomaly detection
- Optimizations
- Simulations

HYPOTHESES TESTING vs. PREDICTING WITH FITTED MODEL

Statistical models find a relationship between explanatory variables and dependent variable:

$$Income = 15000 + 1500\ MaleGender + 1800\ Education + 560\ Age + \varepsilon$$

We can **use it to predict** the income of a new person:

PersonID	Gender	Years Education	Age	Income
8112	F	17	35	???

$$Income = 15000 + 1500 * 0 + 2100 * 17 + 560 * 35 = \mathbf{65\ 200}$$

HYPOTHESES TESTING vs. PREDICTING WITH FITTED MODEL

Predict the income of a new person:

PersonID	Gender	Years Education	Age	Income
8112	F	17	35	???

$$Income = 15000 + 1500 * 0 + 2100 * 17 + 560 * 35 = \mathbf{65\,200}$$

We are interested in **how precise the prediction is.**

We are not so much interested in how exactly education or gender influence the income.

→ We want the model as a whole to do a good job when predicting for a future.

STATISTICS VS. MACHINE LEARNING

Statistics

- Look at **past trends**
- Support/Reject your **hypothesis**
- Examine **effects of individual factors**
- **Assumptions** are important

Machine Learning

- Main goal: **predict future** using past data
- Usually **no strict assumptions**
- Often **no examination of impacts** of individual factors
- **Maximize accuracy/ precision** of prediction (classification)
- **Minimize errors** of prediction (regression)
- Use train/test split

STATISTICS VS. MACHINE LEARNING

Statistics

- Linear regression (OLS)
- Logistic regression
- Panel regressions
  (Fixed effects, Random effects)
- Time Series models

Machine Learning

- **Classification:**
  - Logistic regression
  - Decision Trees
  - Support Vector Machines
    etc.

- **Regression:**
  - Linear regression (OLS)
  - Regression Trees
  - Support Vector Regression
  - etc.

6. MODEL PERFORMANCE EVALUATION

6.1 MODEL PERFORMANCE EVALUATION: HYPOTHESES TESTING

HYPOTHESES TESTING – MODEL VALIDATION

R^2:

- Indicates the percentage of the variance in the dependent variable that the independent variables explain collectively

- 0-100% scale (the higher the better)

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

HYPOTHESES TESTING – MODEL VALIDATION

R^2:

- Indicates the percentage of the variance in the dependent variable that the independent variables explain collectively

- 0-100% scale (the higher the better)

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

p-values:

- Show if we have some good variables in the model that explain well the dependent variable

czechitas

6.2 MODEL PERFORMANCE EVALUATION: PREDICTING

PREDICTING - CROSS SECTIONAL DATA

Statistical **models find a relationship between features and target** variable – for simplicity, imagine an equation such as:

$$Income = 15000 + 1500\ MaleGender + 1800\ Education + 560\ Age + \varepsilon$$

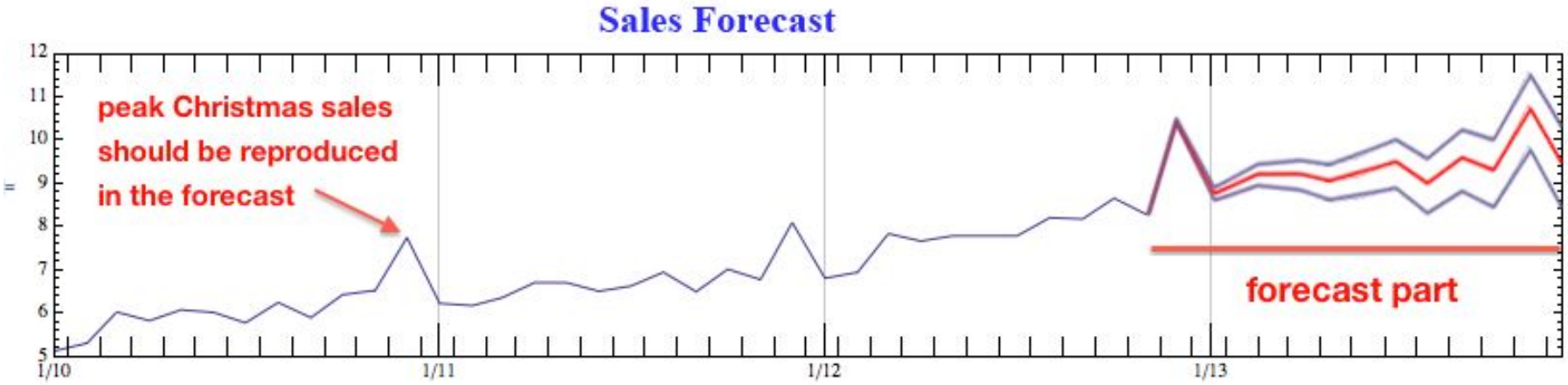
We can **use it to predict** the income of a new person:

PersonID	Gender	Years Education	Age	Income
8112	F	17	35	???

$$Income = 15000 + 1500 * 0 + 2100 * 17 + 560 * 35 = \mathbf{65\ 200}$$

czechitas

PREDICTING - TIME SERIES DATA



czechitas

6.2.1 PREDICTING: TRAIN/TEST DATA SPLIT

TRAIN/TEST DATA SPLIT – CROSS SECTIONAL DATA

CustomerID	Income	Education	Age	Default
2343	50 000	17	35	No
1213	35 000	15	32	Yes
4533	40 000	15	53	No
4563	100 000	19	51	No
7554	50 000	18	28	No
6465	27 500	13	25	Yes
7453	34 000	13	32	No
6775	72 000	18	43	No
4643	50 000	19	47	No
6886	48 000	19	37	?
8668	62 500	21	39	?
8765	78 000	23	46	?
9797	23 000	12	29	?

Before we use the model to predict new values, we need to **use the already labeled data to find the relationships** between variables.

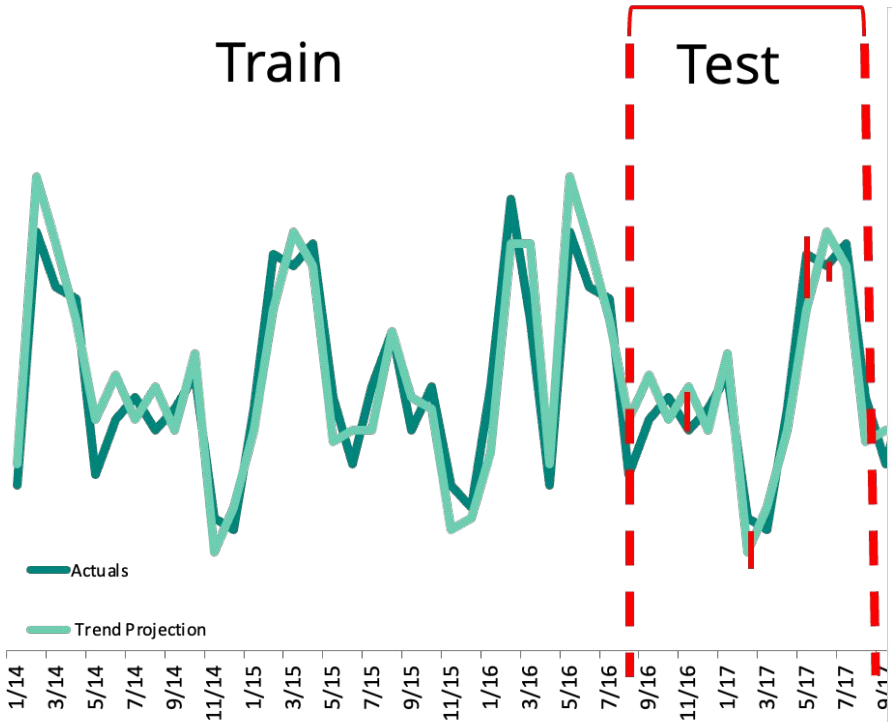
TRAIN/TEST DATA SPLIT – CROSS SECTIONAL DATA

CustomerID	Income	Education	Age	Default	
2343	50 000	17	35	No	TRAIN
1213	35 000	15	32	Yes	
4533	40 000	15	53	No	
4563	100 000	19	51	No	
7554	50 000	18	28	No	
6465	27 500	13	25	Yes	
7453	34 000	13	32	No	TEST
6775	72 000	18	43	No	
4643	50 000	19	47	No	
6886	48 000	19	37	?	
8668	62 500	21	39	?	
8765	78 000	23	46	?	
9797	23 000	12	29	?	

Use only a part of the data for finding those relationships, i.e., for model **training (~60-80%).**

Keep another part of the data for **testing**.

TRAIN/TEST DATA SPLIT – TIME SERIES



- Train sample
  - Long enough to catch seasonality
  - Some models might need more data to train than others
- Test sample
  - the size of your forecasting horizon

TRAIN/TEST DATA SPLIT RATIONALE

1. **Find the model using the training data**
2. Calculate model performance
3. **Use the same model to predict with testing data**
4. Calculate model performance

→ We want both training and testing model performance to be similarly good.

6.2.2 PREDICTING: METRICS

PREDICTIVE MODELS PERFORMANCE METRICS

Validation depends on:

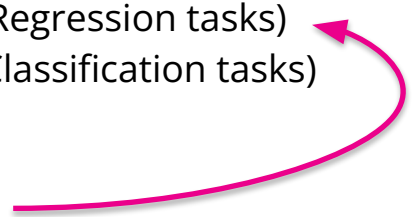
Data structure:

A. **Cross-sectional, Panel**

Also depends if target variable is:

- Continuous (Regression tasks)
- Categorical (Classification tasks)

B. **Time-Series**



Time series have the same performance metrics as regression tasks with cross-sectional/panel data.

REGRESSION PERFORMANCE

1) **Estimate the model** with historical data – the model is generalized, does not hold perfectly for every individual!

$$Income = 15000 + 1500\ MaleGender + 1800\ Education + 560\ Age + \varepsilon$$

2) Use the estimated model to **calculate the predicted target** variable
3) See how similar are the real, **observed values vs. predicted values**

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

REGRESSION PERFORMANCE: RMSE

How to best compare observed values vs. predicted values?

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

Dataset: 750 individuals

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Predicted_i - Observed_i)^2}$$

n – number of observations

$$RMSE = \sqrt{\frac{(65\,200 - 63\,000)^2 + (37\,300 - 35\,000)^2 + \ldots}{750}}$$

czechitas

REGRESSION PERFORMANCE: MAPE

How to best compare observed values vs. predicted values?

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

Dataset: 750 individuals

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Observed_i - Predicted_i}{Observed_i}\right|$$

n – number of observations

$$MAPE = \frac{1}{750}\left(\left|\frac{63\,000 - 65\,200}{63\,000}\right| + \left|\frac{35\,000 - 37\,300}{35\,000}\right| + \cdots\right)$$

czechitas

CLASSIFICATION PERFORMANCE

CustomerID	Income	Education	Default - Observed	Default - Predicted
2343	50 000	17	No	No
1213	35 000	15	**Yes**	**No**
4533	40 000	15	**No**	**Yes**
4563	100 000	19	No	No
7554	50 000	18	No	No
6465	27 500	13	Yes	Yes
7453	34 000	13	**No**	**Yes**
6775	72 000	18	No	No
4643	50 000	19	No	No

CLASSIFICATION PERFORMANCE

CustomerID	Income	Education	Default - Observed	Default - Predicted
2343	50 000	17	No	No
1213	35 000	15	**Yes**	**No**
4533	40 000	15	**No**	**Yes**
4563	100 000	19	No	No
7554	50 000	18	No	No
6465	27 500	13	Yes	Yes
7453	34 000	13	**No**	**Yes**
6775	72 000	18	No	No
4643	50 000	19	No	No

Confusion Matrix

	Predicted Yes	Predicted No
Observed Yes	1	1
Observed No	2	5

czechitas

CLASSIFICATION PERFORMANCE

Confusion Matrix

TRUE POSITIVES

	Predicted Yes	Predicted No
Observed Yes	1	1
Observed No	2	5

TRUE NEGATIVES

CLASSIFICATION PERFORMANCE

Confusion Matrix

TRUE POSITIVES

FALSE NEGATIVES

	Predicted Yes	Predicted No
Observed Yes	1	1
Observed No	2	5

FALSE POSITIVES

TRUE NEGATIVES

CLASSIFICATION PERFORMANCE: Accuracy

	Predicted Yes	Predicted No
Observed Yes	TRUE POSITIVES	FALSE NEGATIVES
Observed No	FALSE POSITIVES	TRUE NEGATIVES

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

CLASSIFICATION PERFORMANCE: Precision

Intuition: book recommendation analogy

- I do not want to waste my time reading books that I do not like
- I am ok if some good ones are not recommended

CLASSIFICATION PERFORMANCE: Precision

	Predicted Yes	Predicted No
Observed Yes	TRUE POSITIVES	FALSE NEGATIVES
Observed No	FALSE POSITIVES	TRUE NEGATIVES

$$Precision = \frac{TP}{TP + FP}$$

How many observations predicted as positive are really positive?

CLASSIFICATION PERFORMANCE: Recall (Sensitivity)

Intuition: burglar's analogy

- When stealing, I will not be checking whether all jewelry is real
- It is important to get as much as possible and to check things later

czechitas

CLASSIFICATION PERFORMANCE: Recall (Sensitivity)

	Predicted Yes	Predicted No
Observed Yes	TRUE POSITIVES	FALSE NEGATIVES
Observed No	FALSE POSITIVES	TRUE NEGATIVES

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

How many observations out of all positive observations have we classified as positive?

CLASSIFICATION PERFORMANCE

Other classification performance metrics:

- ROC curve and Area Under Curve (AUC)
- F1 Score

Quiz: what performance we want to achieve on training vs. testing dataset?

a) train performance should be ideally higher than test performance

b) they should be as similar as possible

c) test performance should be ideally higher than train performance

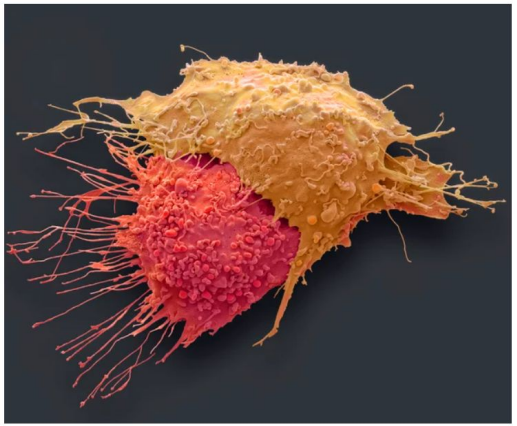
Quiz:
You have a drug that can save lives of women from deadly cancer but potentially has strong side effects.

Drug helps only some women and you want to build a model to identify to which women it will help. Which measure you should optimize?

a) accuracy

b) precision

c) recall



Ovarian cancer cells. Credit: Steve Gschmeissner *Getty Images*

czechitas

SUMMARY

Data science and problems it can solve:
- descriptive
- diagnostic
- predictive
- prescriptive

Data structures:
- wide vs. long
- cross-sectional, time-series, panel

Hypotheses testing vs. Supervised vs. Unsupervised Learning

Statistics vs. Machine Learning

Model Validation principles and metrics

Thank you for your attention!

Next time:
Data Pre-processing with Thomas, Josef, and Andrea