Data fitness program

DATA FITNESS PROGRAM KICK-OFF


Data Science


Program Goals

Moviri Introduction


Audience Introduction


Program Rules


Program Agenda

WHAT IS DATA SCIENCE?

Data science is a very complex discipline, leveraging mathemathics, statistics and computer science

Data Science includes a sequence of actions and knowledge to master in order to get best from data

"Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data"

This is the overall final aim of the whole Data Science, which is also the reason why it is such a used buzzword

Enabling Business decisions and actions

Data is everywhere, but it is not always coming in the proper format to be used (almost never actually)

Data widely pervades every phase of Data Science: **DATA IS THE NEW BACON**

https://en.wikipedia.org/wiki/Data_science

PROGRAM GOALS

Being aware of the availability of **methodologies and technologies** supporting company needs

Getting fit with data

Learning the language of data analysis and enhance the capability to “think data”

Understanding the value of data analysis to the business

Enabling definition and realization of projects that **gather business value** from the data

AT THE END OF THESE 2 DAYS…

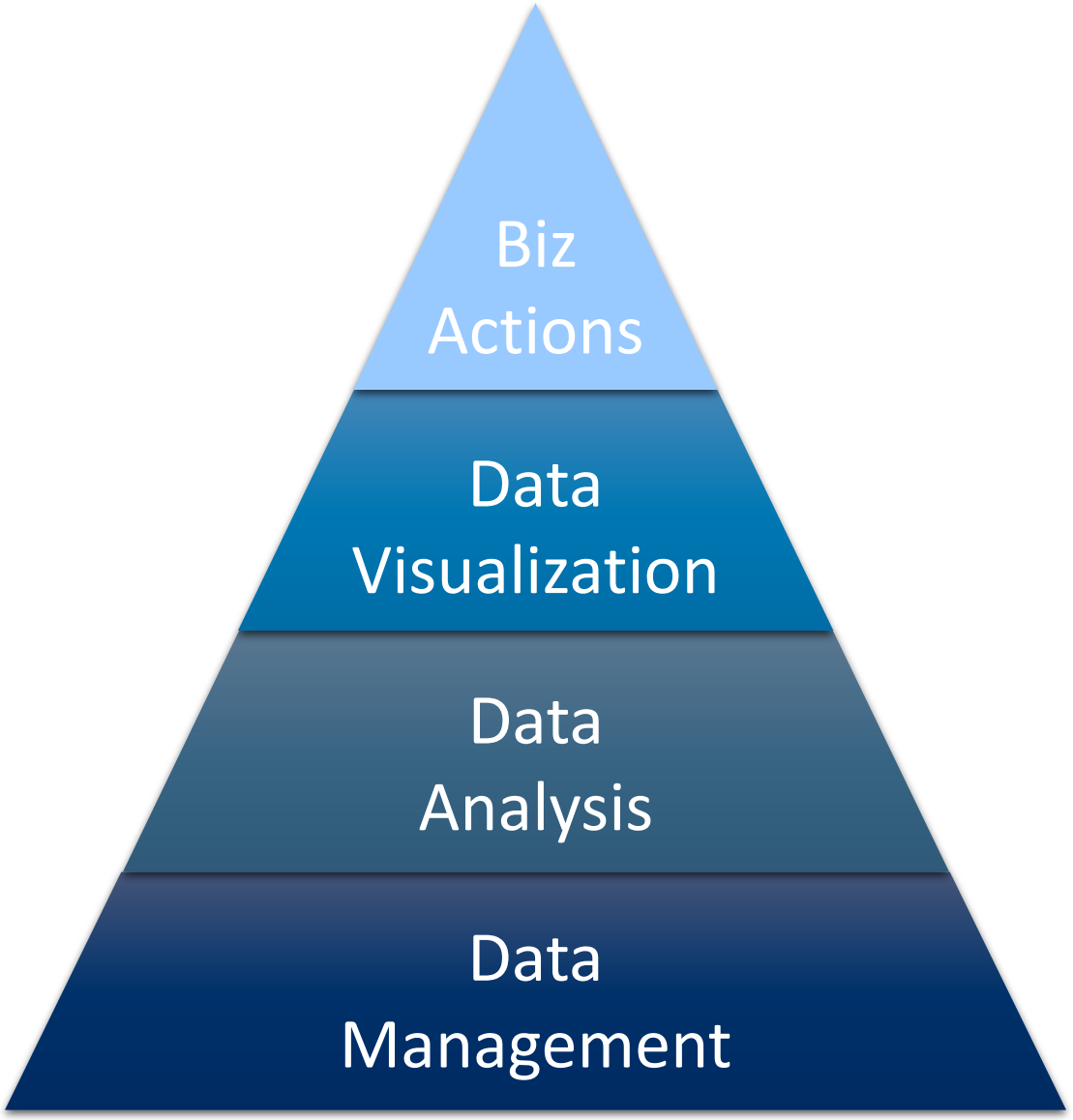
… you should not expect to:

- Be able to directly apply everything you saw during the presentation in your daily work
- Become a Data Science guru

… you should expect to:

- Have an overview of methodologies and technologies within Data Science Ecosystem
- Understand how Data Science can help you leveraging your data to get relevant business value
- Have the «appetite» to retrieve more information on these topics, helping you to really apply these concepts on your business

DATA SCIENCE PHASES



Data driven approach in decision making

Effective data reporting and dashboarding

Machine Learning and Artificial Intelligence applied to data

Architecture Design
Data integration and processing

INTRODUCING MOVIRI

Data Analytics for Business, Performance, Security and UX automation

Founded in 2000 as Neptuny, a Politecnico di Milano spin-off, in 2010 rebrands in Moviri after selling Caplan, its line of business, to BMC Software. Headquartered in Milan, has full ownership of Moviri Inc. in the US and Moviri Limited in the UK.

€ 31M+
Revenues

200+
Experts

150+
Trusted Customers

40+
Countries Covered

AUDIENCE INTRODUCTION

Who are you?

What do you do in your role *@Vodafone CIoT Team?*

How would you describe your knowledge level about Data Science?

BEGINNER

Limited or no knowledge of data and data science techniques

INTERMEDIATE

Basic knowledge, never or rarely looked behind a report or dashboard

ADVANCED

Good knowledge, data used daily or it is the main focus of my work

PROGRAM RULES

Questions are welcome at any time

Anyways there will be a Q&A sessions at the end and beginning of every section

There are NO STUPID QUESTIONS

Most "stupid" ones are the most welcome

Keep the program INTERACTIVE

We'll have rather short sub-sessions, in order to maximize attention

DATA FITNESS PROGRAM

DAY 1

10:00 — *Kick Start*

10:45 — **BUSINESS CASES PRESENTATION**

DATA VISUALIZATION

12:30 — *Lunch Break*

13:30 — **DATA ANALYSIS (Basic)**

16:00

DAY 2

09:30 — *Kick Start*

DATA MANAGEMENT

11:00 — **BUSINESS CASES DEEP DIVE**

13:00

Q&A

PICTURES CREDITS

1: https://giphy.com/explore/excellent-question

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 1

10:00

Kick Start

BUSINESS CASES PRESENTATION

10:45

FITNESS AS A GUIDE FOR THE PROGRAM

Throughout this program we will use the analogy of Fitness, using the idea of exercise to indicate the complexity of each slide

Easy

Medium

Hard

DATA FITNESS PROGRAM: The warm-up

A teaser on 3 Moviri business cases
as a warm-up for what will come next

AGENDA

3 Moviri Business Cases

- Churn Prediction
- Proactive Issues Detection
- Customer Journey Optimization

CUSTOMER SATISFACTION AND CHURN PREDICTION

Telco Company Case

Challenge	 Understanding main reason for customer churn	 Identify best actions to increase customer statisfaction	 Identifying potential behavior clusters
Gathered Insights	As “expected” **line quality** is one of the main causes for **dissatisfaction**, but **not for all products**	**“Technical” improvements** were the most **effective** actions, but **courtesy calls** have been **equally powerful** at times	**Considerably different behaviors** among Fiber, FTTS, ADSL and ADSL WS customers

CUSTOMER JOURNEY ANALYSIS AND ADV OPTIMIZATION

Fashion Company Case

Challenge	 Estimating contribution to sales of each marketing lever	 Identify best campaign testimonials	 Understand best targets for communication
Gathered Insights	**Email marketing** was the most efficient channel, **Social** were **not impacting** directly on revenues	The choice of the **VIP** to promote products **impacted** on average for **more than 30%** of total revenues	Location, income, education, age, gender and device were **relevant parameters** to detect specific clusters with **high profitability**

ANOMALY DETECTION ON NETWORK UTILIZATION

Oil&Gas Company Case

Challenge	 Identify the most critical IT resource for oil plants	 Proactively identify anomalies in resource behavior	 Alert correct people to address issues
Gathered Insights	**Network bandwidth consumption** was pointed has the one with highest **correlation with system failures**	**Abnormal situations** have been **predicted** with hours of anticipation, thanks to **Machine Learning** techniques	**For every type of issue**, route cause is identified and **people** in charge of specific asset/resource **are alerted**

Credits #: 1

PICTURES CREDITS

1: https://giphy.com/explore/any-questions

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 1

Kick Start

10:45

BUSINESS CASES PRESENTATION

DATA VISUALIZATION

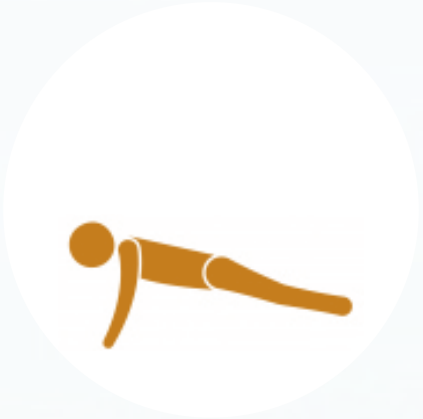
12:30

Lunch Break

13:30

DATA FITNESS PROGRAM: Arms pumping

Data visualization is the “business card” of all your hard work, as arms training for fitness



AGENDA

- Introduction to DV
  - Why Visualization
  - The Data
  - Tidy Data up
- Data representation
- Enterprise tools
- Classwork

HUMAN EYE IS IMPRESSED BY IMAGES IN FEW MILLISECONDS...

Image courtesy of the Earth Science and Remote Sensing Unit, NASA, Johnson Space Center. Photo ID: *ISS043-E-86375*

Credits #: 1

…VISUALIZATION HELPS MEMORY…

It is common to draw pictures for future times, recording what we knew.

Think for instance of Galileo depicting the Moon in early XVII century…

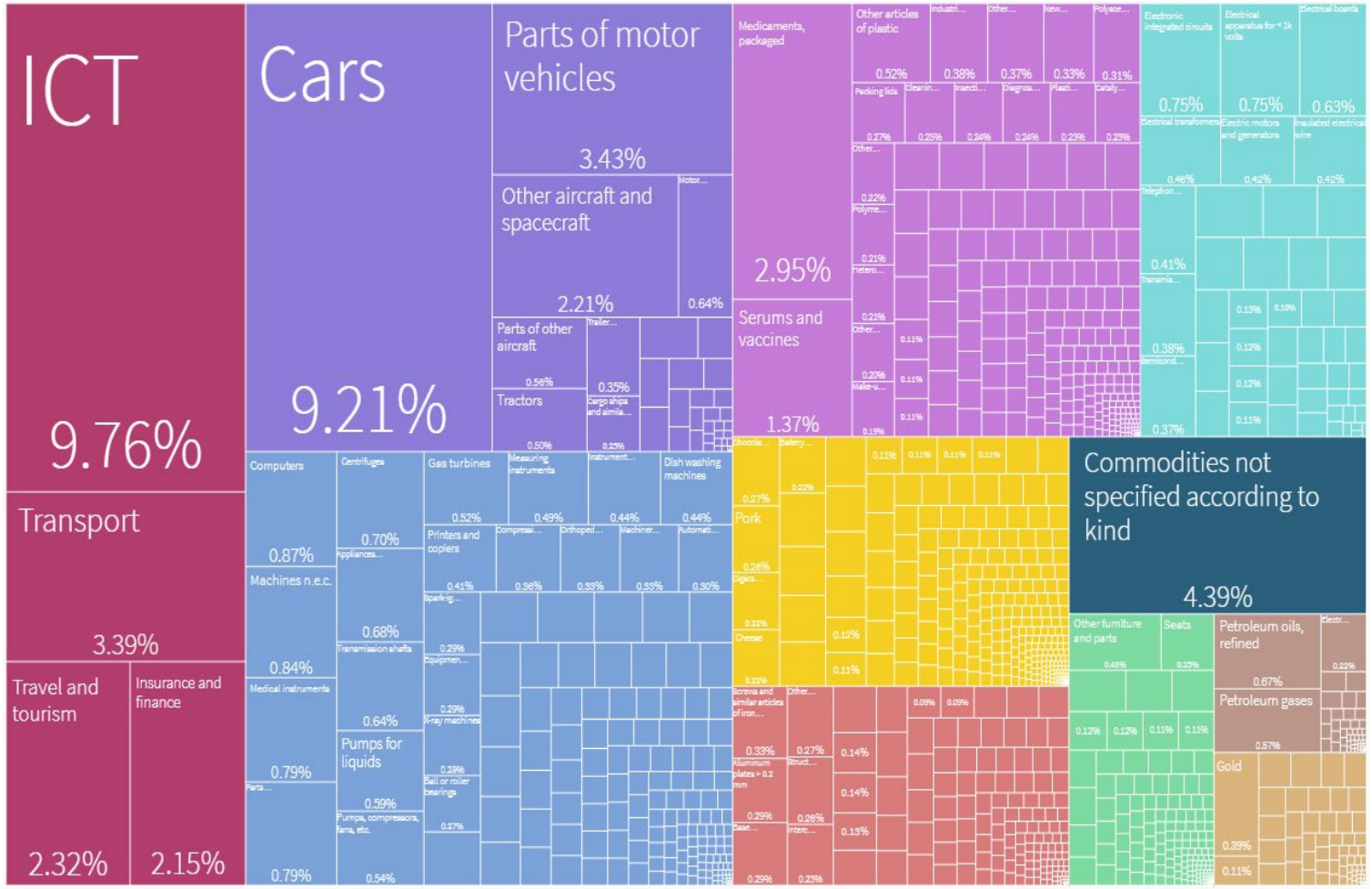
WHAT GALILEO SAW

Galileo's sketches of the moon - Sidereus Nuncius - March 1610

… ANALYSIS …

Data is easier to read in visual form, which strongly helps in discovering new knowledge

Drawing graphs allows to see how markets share split…

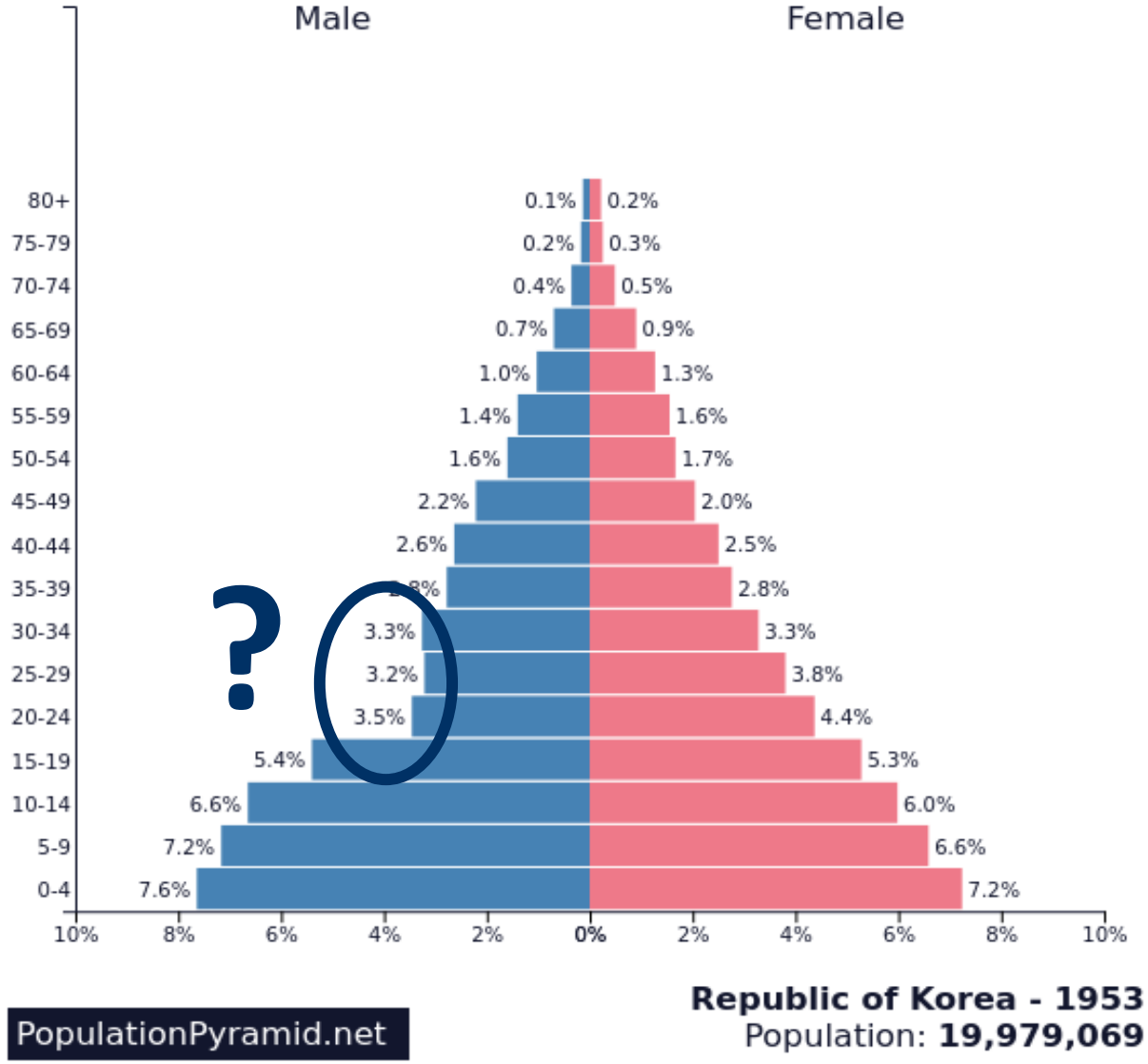


Germany export in 2016
"The Growth Lab at Harvard University. The Atlas of Economic Complexity"

Credits #: 2

… AND COMMUNICATION!

… or communicate any idea in which we are confident



THE POWER OF DATA VISUALIZATION TOOLS

Today computer tools are continuously enhancing data visualization, providing:

Interactivity

Scaling on datasets dimension

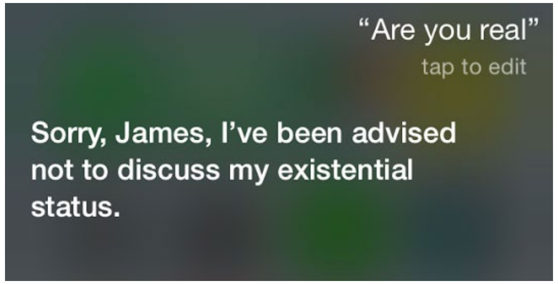
Detailed data exploration

New graphical idioms composing multiple simple charts

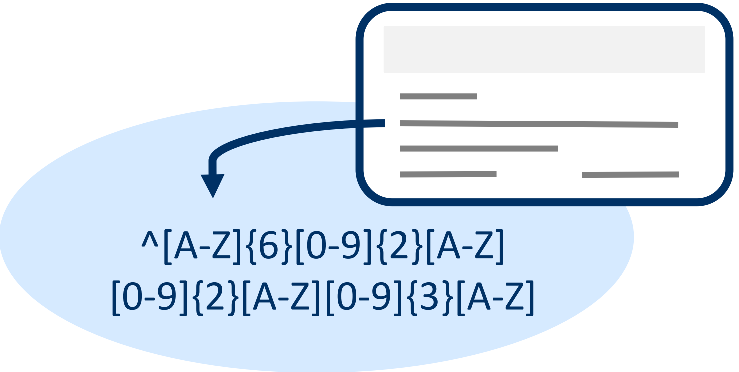
Credits #: 4

DATA IS EVERYTHING AND EVERYWHERE

Data come in many different fashions: text, numbers, log file, codes …




```
12/20/2010 8:08:02 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:02 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:03 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:04 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:05 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:06 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:06 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:06 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:06 PM : Tick event occurred (Args: System.EventArgs)
12/20/2010 8:08:06 PM : Tick event occurred (Args: System.EventArgs)
```




```python
# Python program to get average of a list
def Average(lst):
    return sum(lst) / len(lst)

# Driver Code
lst = [15, 9, 55, 41, 35, 20, 62, 49]
average = Average(lst)

# Printing average of the list
print("Average of the list =", round(average, 2))
```

CIoT DATA SOURCES

- Name, Address
- Email address
- Mobile number
- Marketing Permission Y/N?

- Existing Vodafone customer
- Non-Vodafone customer

- Date of Birth, Place of Birth
- Personal ID number / Fiscal Code
- Nationality
- Payment details (CTB, CC, Debit card)
- Account creation date

Contact Details

Customer Status

Personal / KYC details

- V-APP installed, de-installed
- V-APP usage (eg App opened, App functions used, App closed etc)

V by VF App

CIoT Subscriptions

- CIOT subscription
- Status – active / cancelled
- Activation date
- Renewal date
- Subscription charge amount

- App installed, de-installed
- App usage (eg App opened, App functions used, App closed etc)

CIoT Apps
Eg V-PET, V-BAG, V-AUTO, V-CAMERA

CIoT Device(s) Purchased

- CIOT device purchased
- Date / Time & Channel
- Purchase price

- Items / pages browsed
- Basket analysis

CIoT E-Shop use

CIoT Device(s) Data Usage

- CIoT device being used (IMEI)
- MB used per CIoT device
- Per hour, per day

CIOT Customer Care

- CIOT service issues per customer
- Issue & outcome
- "Opt out of Marketing" requests

Looking @ d4t4 *helps* to extract v4lu3 from noun morfiin *incoming* 1nf0rm4t10n!

INFINITE SOURCES

Data can come:

- from multiple sources
- in many fashions
- with no clear order

How can we organize these?

FIRST REORDERING CHECKLIST…

If all data share a **common pattern**, they can be ordered in a **structured** fashion:

- Imagine to order them on a white sheet
- Find and enumerate the common traits (let's call them «features»)
- Give them a meaningful label
- Put them in a row (Choose a proper separator)
- Take the first sheet and read the content
- Copy each information under the corresponding label, thus forming a new row
- For each sheet repeat the previous point adding a new row

…ENDING UP IN TABLES

Data are **listed** (allowing to count items, order them, filter if necessary…)

Features are **labelled**, therefore understandable without reading any item

Atomic information can be **retrieved faster** by intersecting a row and a column

The square geometry enhances **reasoning** and **making comparisons**

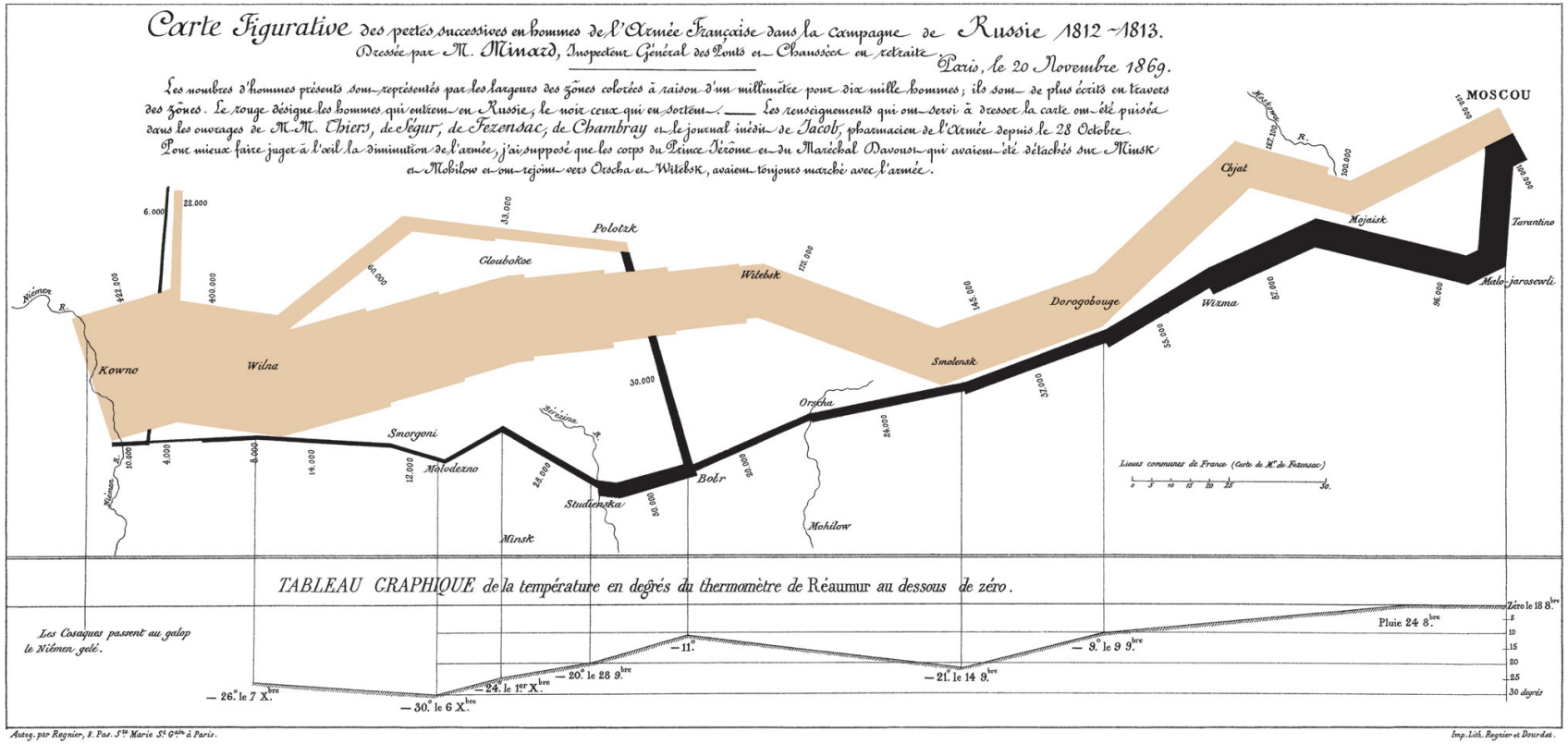
NAME	AGE	CITY
Paul	32	London
Franz	45	Dusseldorf
Luke	26	Manchester
Ann	28	Berlin

AGENDA

- Introduction to DV
- Data representation
- Enterprise tools
- Classwork

- Visualization principles
- Visual Variables and Dimensions
- Bad charts
- Common charts features
- Building good charts

Show the data



Show the data

Provoke thought

World's **26** richest people own as much as poorest

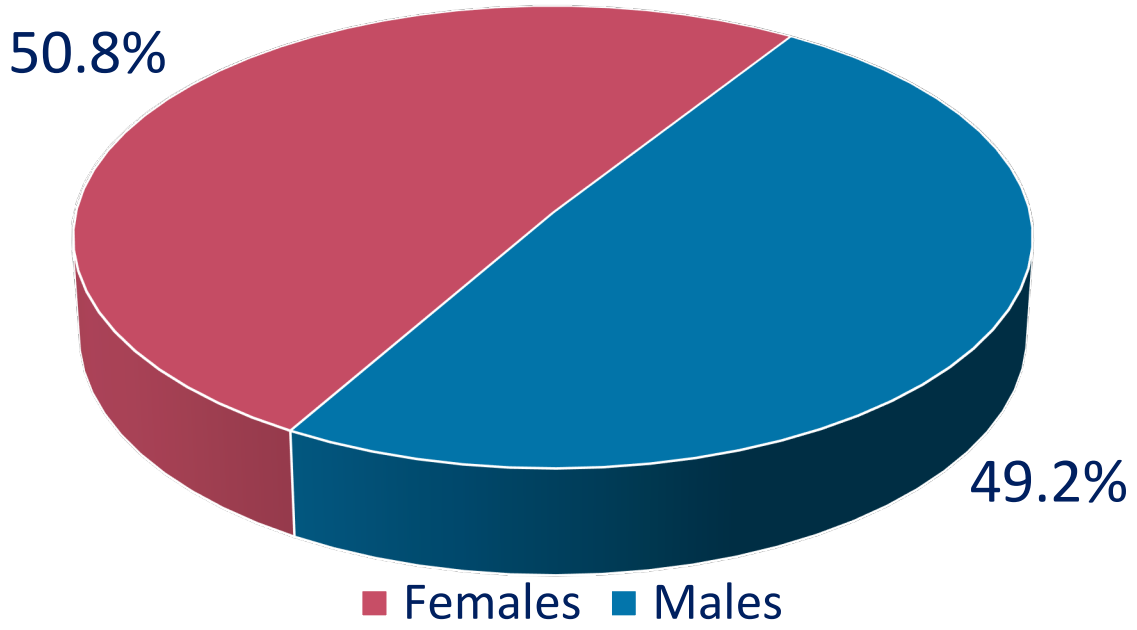
50%

Show the data

Provoke thought

Avoid distortion

World population Gender (%)

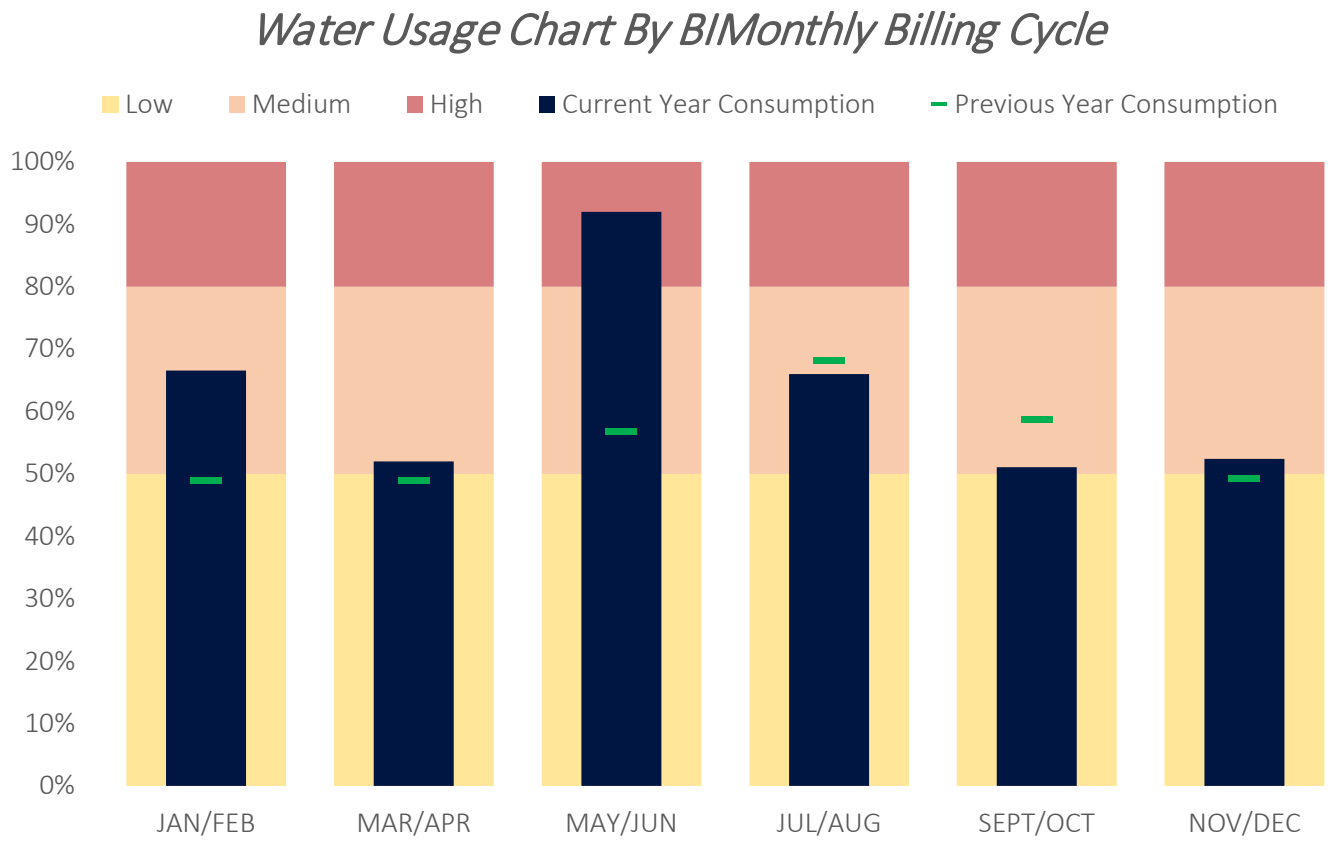


Show the data

Provoke thought

Avoid distortion

Present many numbers in a small space



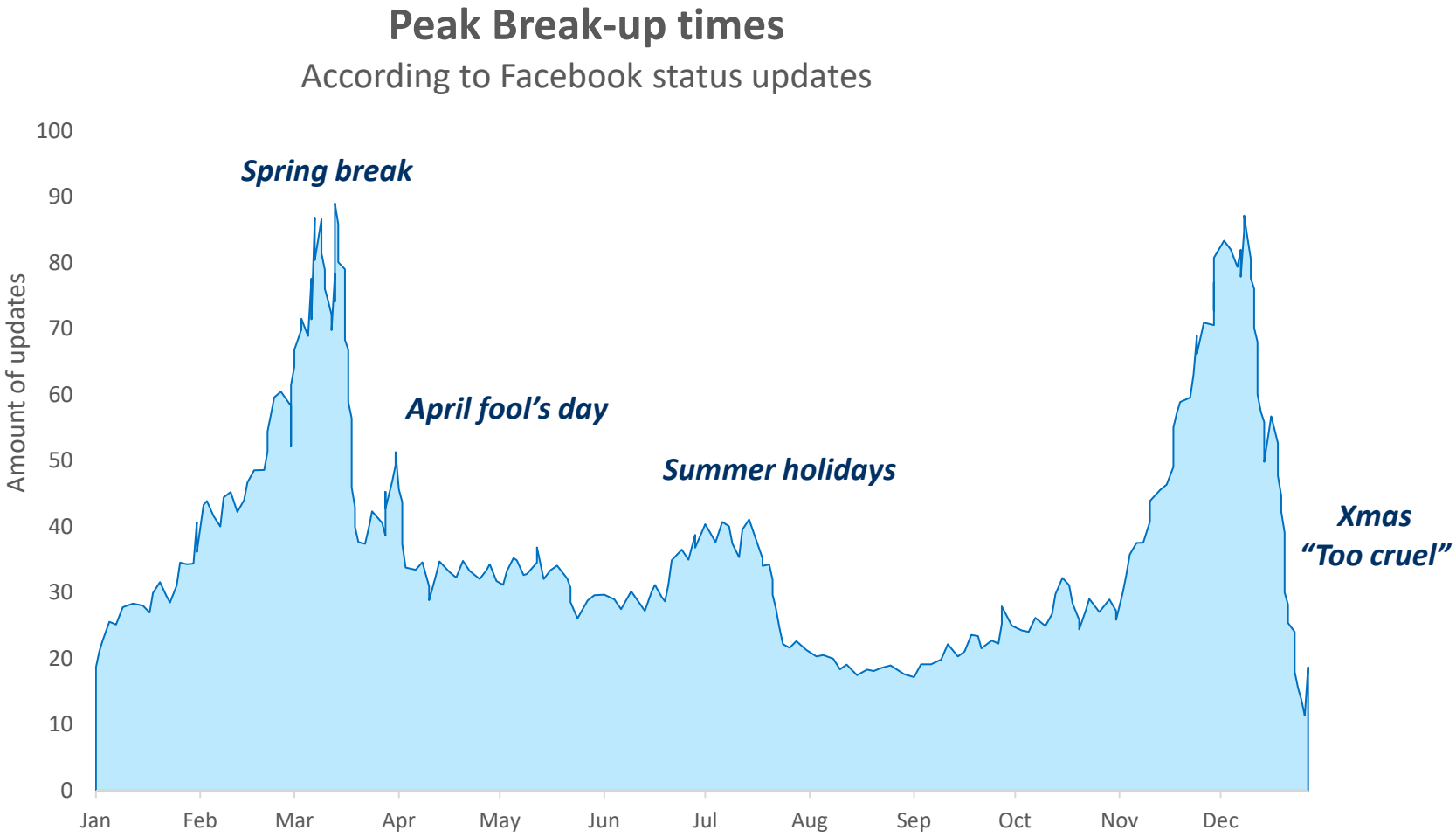
Show the data

Provoke thought

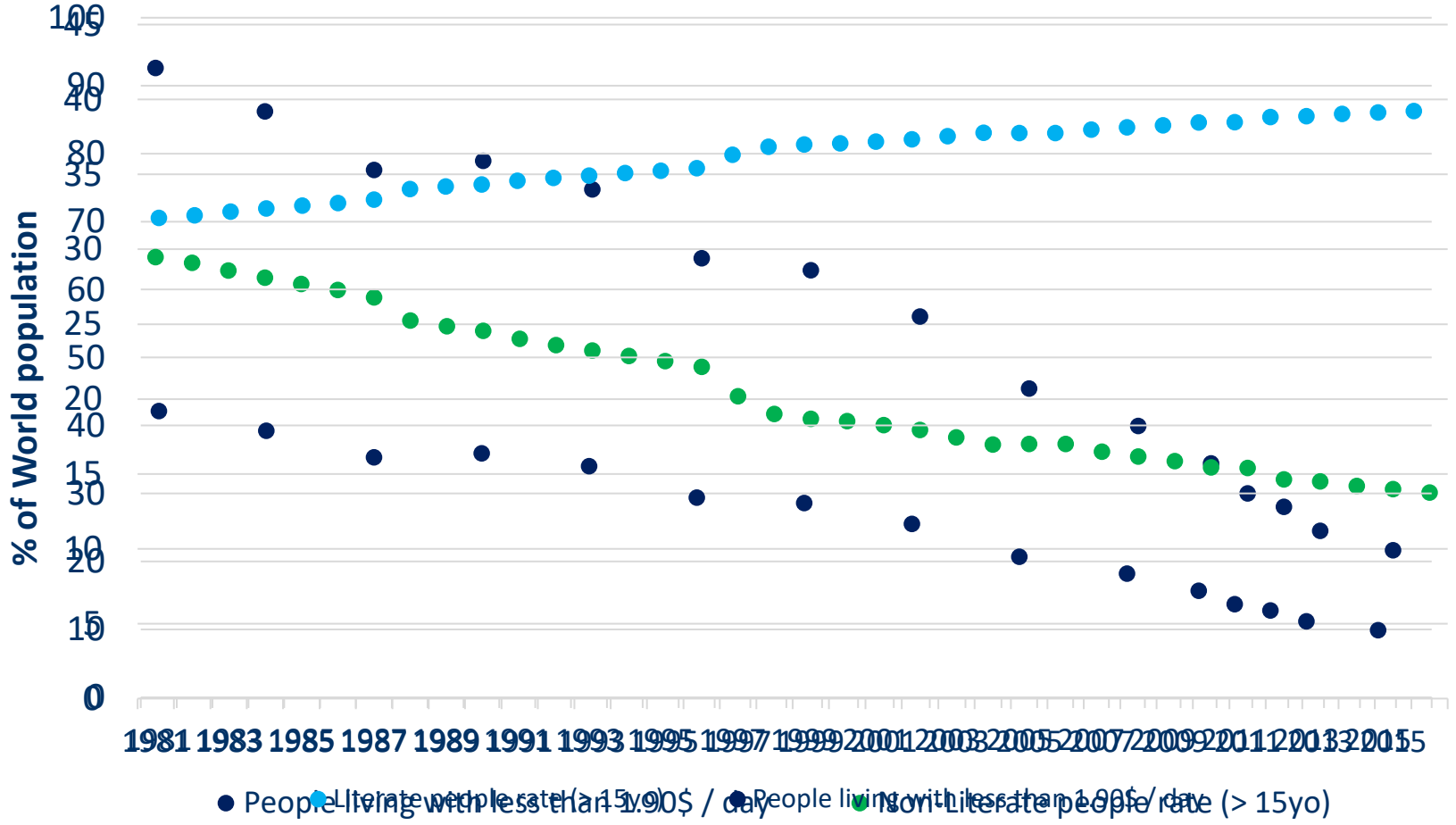
Avoid distortion

Present many numbers in a small space

Make large datasets coherent



- *Show the data*
- *Provoke thought*
- *Avoid distortion*
- *Present many numbers in a small space*
- *Make large datasets coherent*
- *Encourage eyes to compare data*



Credits #: 7

Show the data

Provoke thought

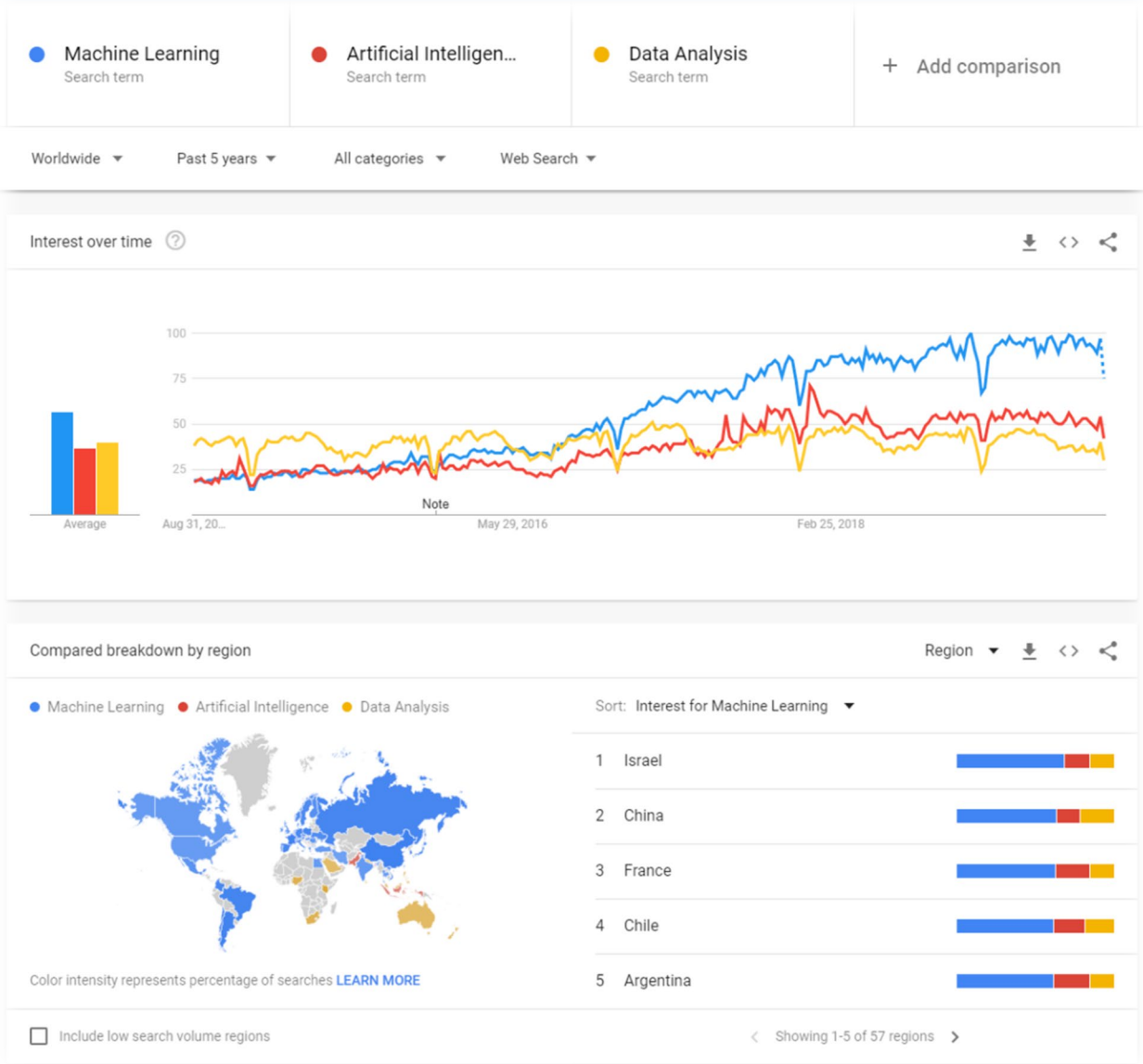
Avoid distortion

Present many numbers in a small space

Make large datasets coherent

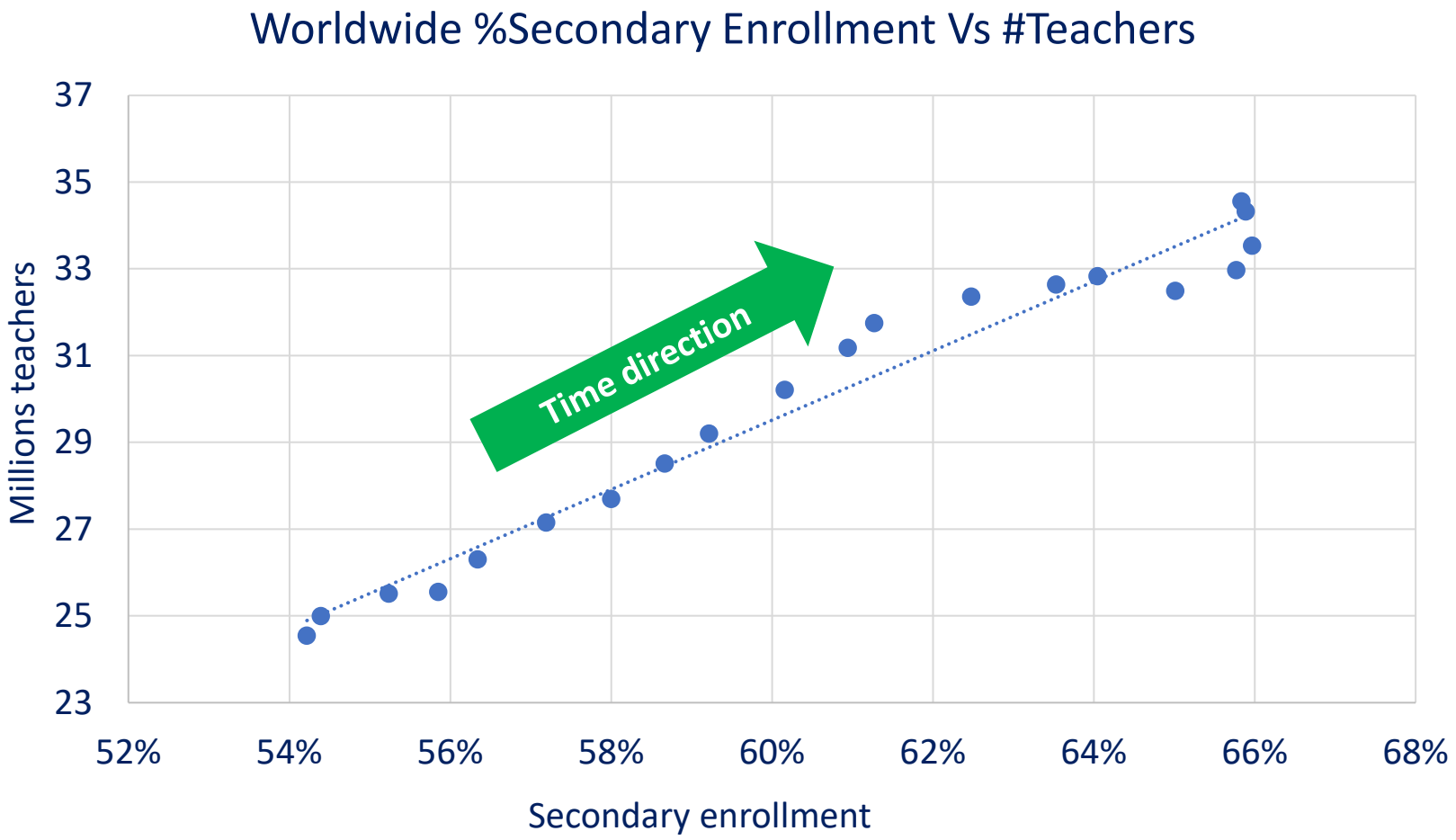
Encourage eyes to compare data

Reveal data at several levels of detail



Credits #: 8

- *Show the data*
- *Provoke thought*
- *Avoid distortion*
- *Present many numbers in a small space*
- *Make large datasets coherent*
- *Encourage eyes to compare data*
- *Reveal data at several levels of detail*
- *Serve a reasonably clear purpose*



Credits #: 7

15 minutes Break

https://giphy.com/explore/any-questions

MAIN VISUAL VARIABLES...

Color

- Use color consistently

- Be consistent with user expectations
- Why colors:
  - For emphasis
    - Opt for greyscale if effective
    - Avoid high contrast colors
    - Avoid large saturated color areas

MAIN VISUAL VARIABLES...

Color

Position

Mark

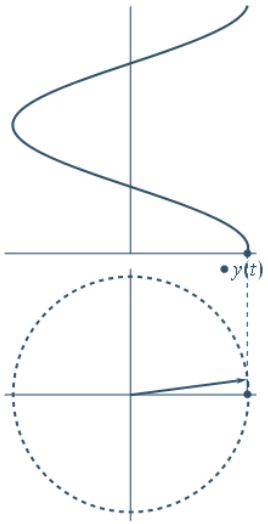
Size

Brightness

Motion

Orientation

Texture



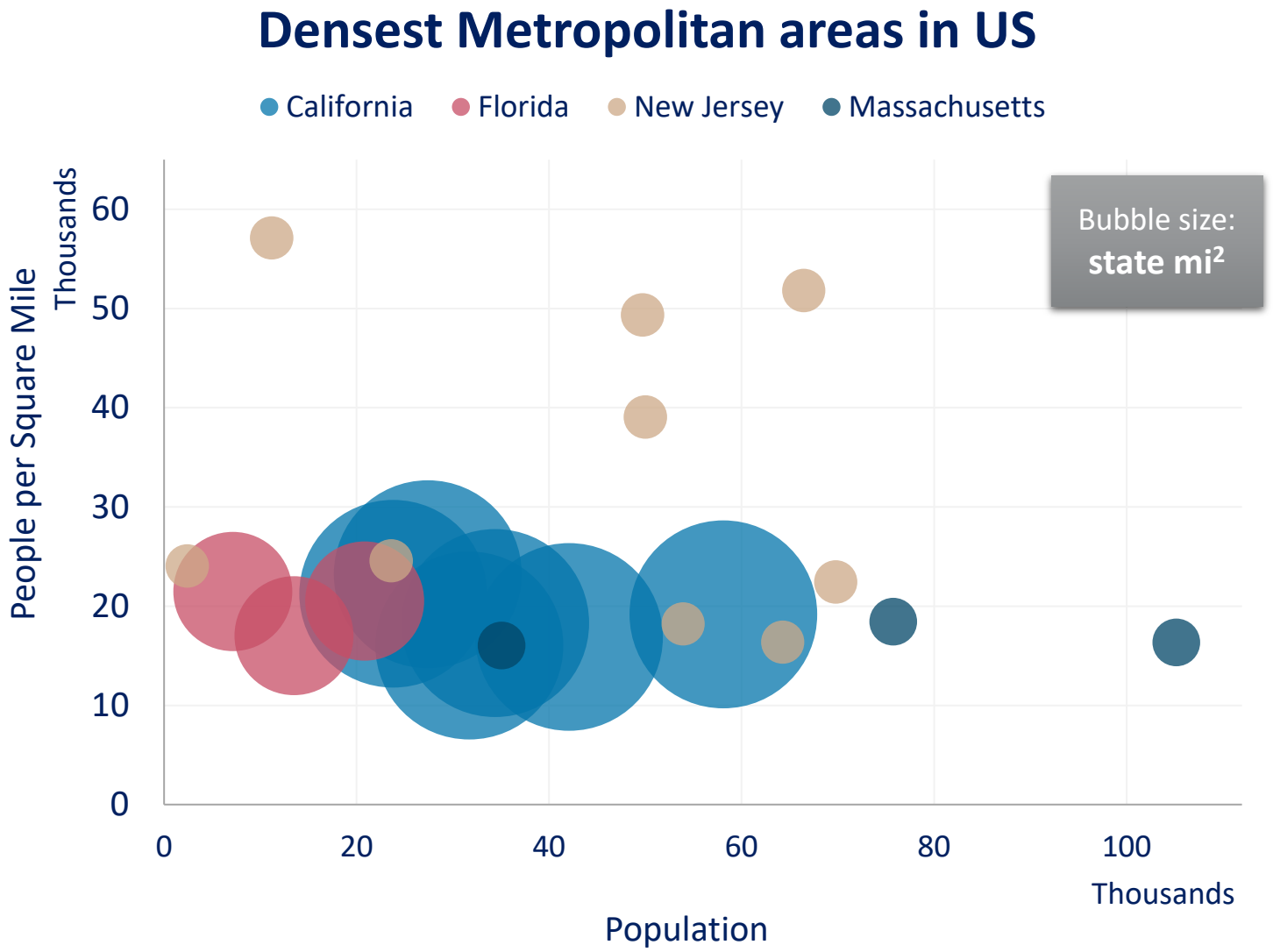
… AND DIMENSIONS

Sometimes it is complicated to represent data having a lot of dimensions (>2) in a single chart. Nevertheless, effective use of high-dimensional charts can have immediate and intense effect on the audience

New Jersey has much more highly dense cities and highest density variation

California, Florida and Massachusetts have clearly different average populations

Being in large state does not have logical correlation with cities density or population



GIVE YOUR BEST SHOT!

10 minutes

Try to represent these numbers now

15 30 39 16

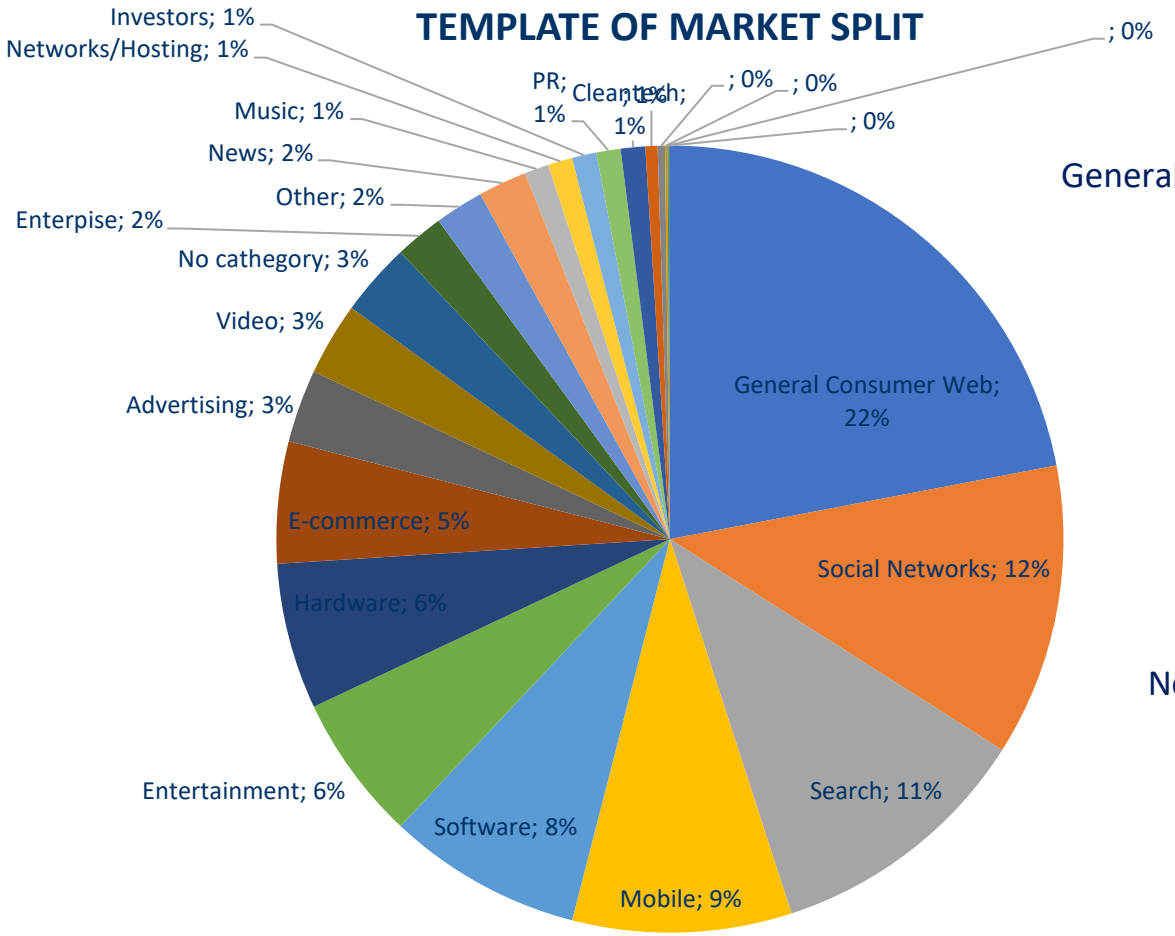
BUILDING BAD CHARTS IS EASIER THAN YOU THINK

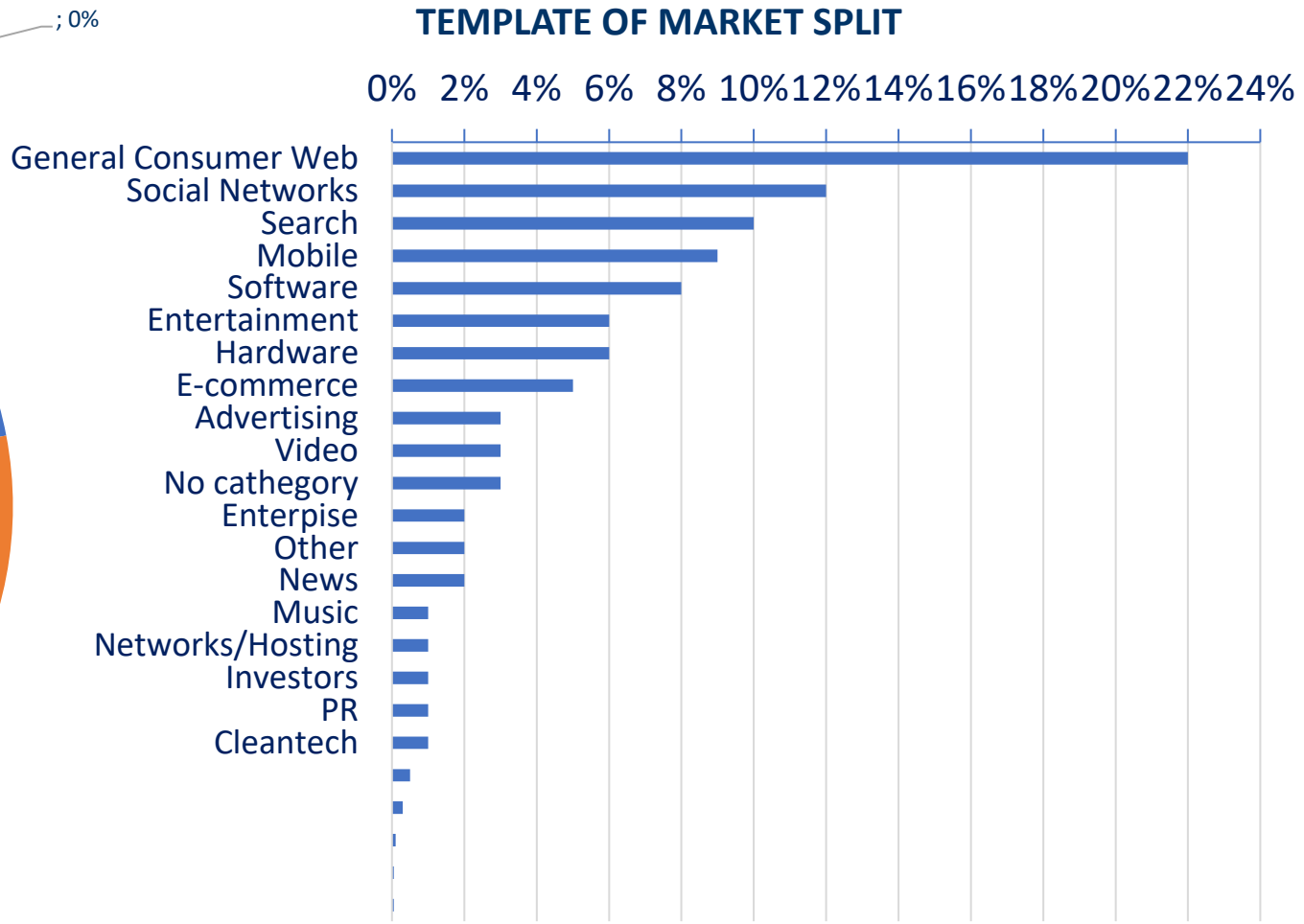
Here are the main reasons for a representation to be bad:

- Confusion
- Misleading choices
- Junk pictures
- 3D Distortion

CONFUSION PREVENTS FROM FAST POINT CATCHING

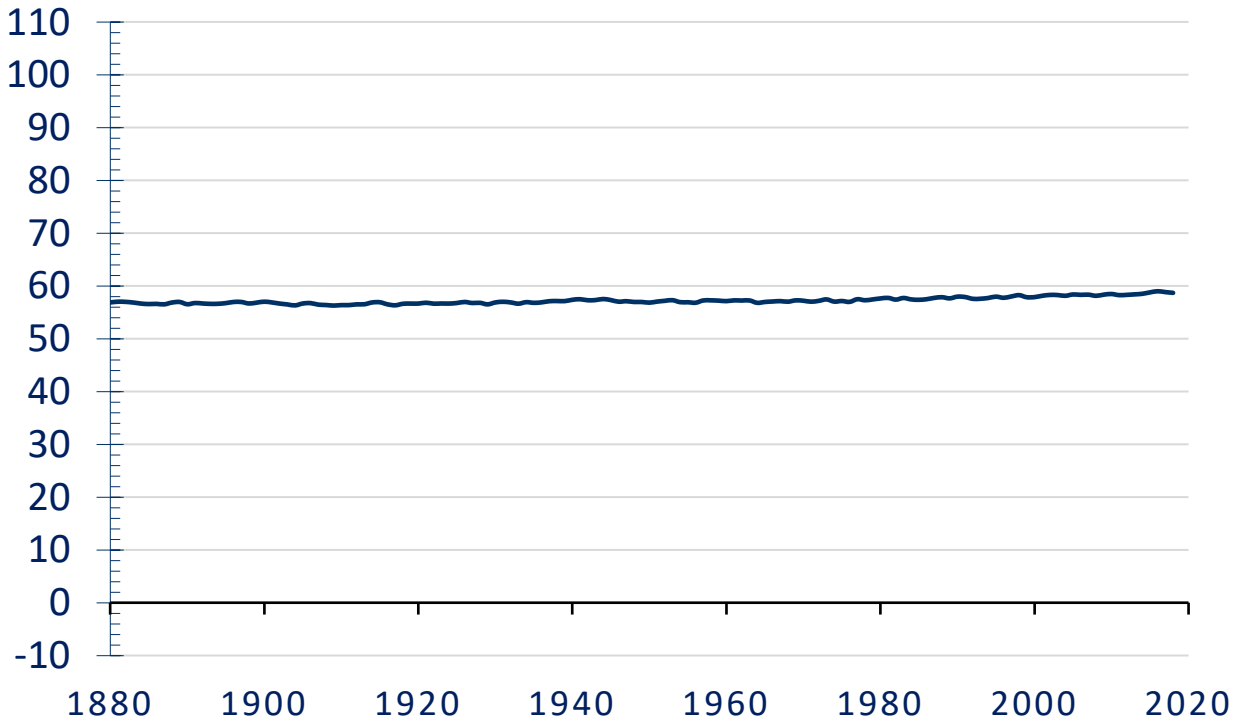
Confusion

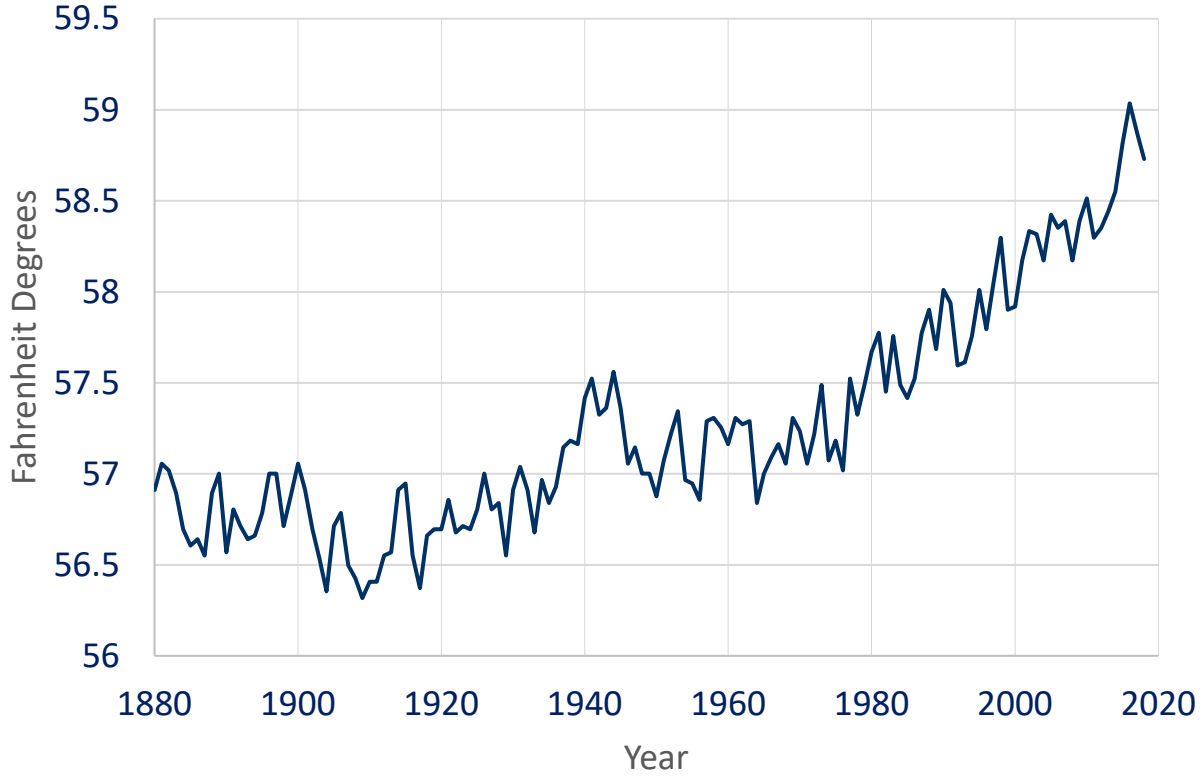




MISLEADING CHOICES HIDE INSIGHTS

Misleading Choices

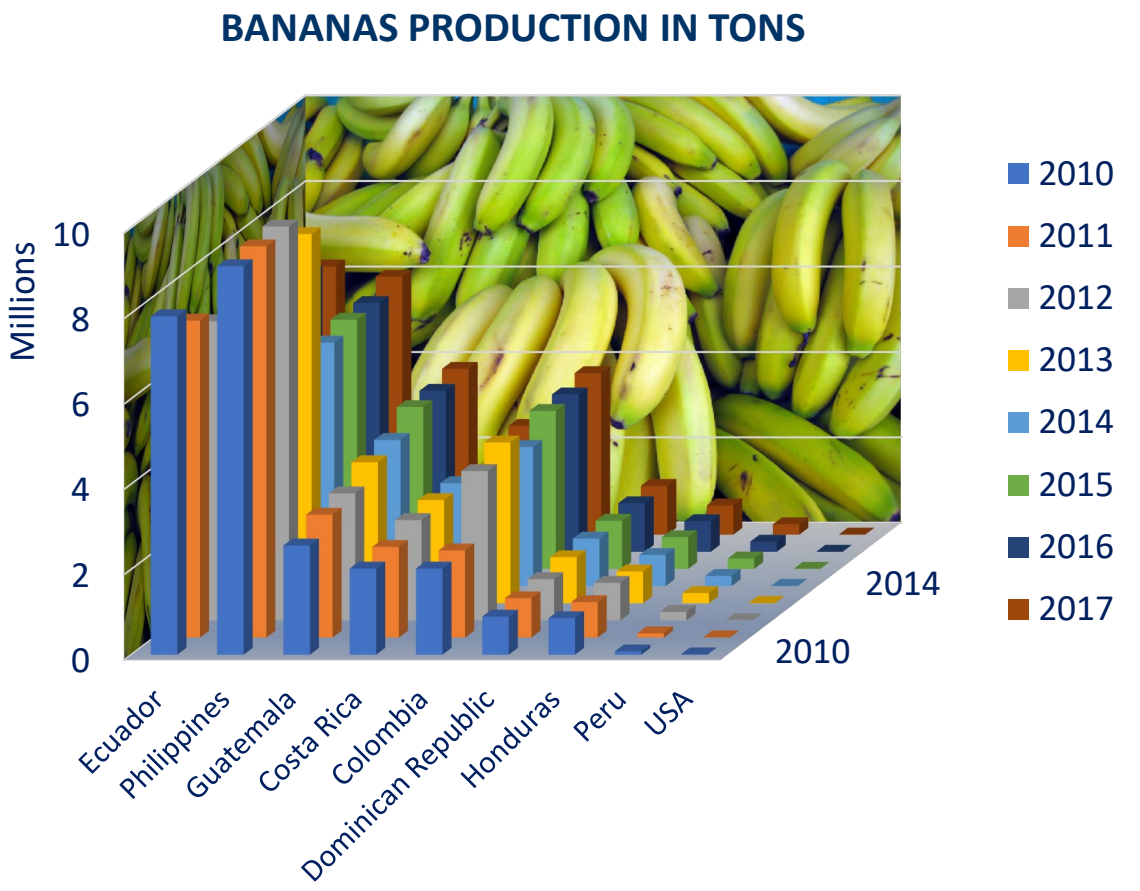


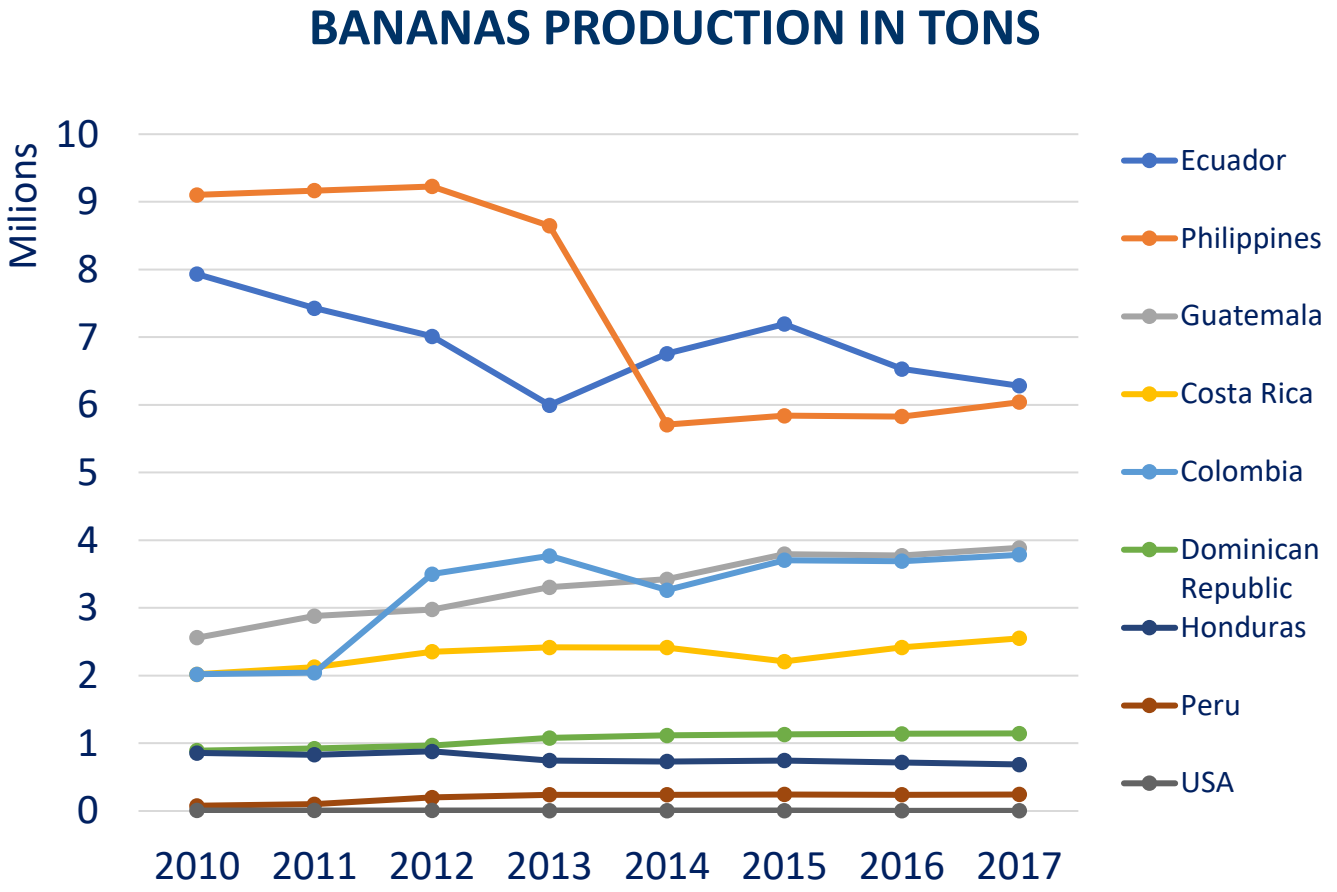


Credits #: 10

JUNK PICTURES WASTE ATTENTION

Junk Pictures

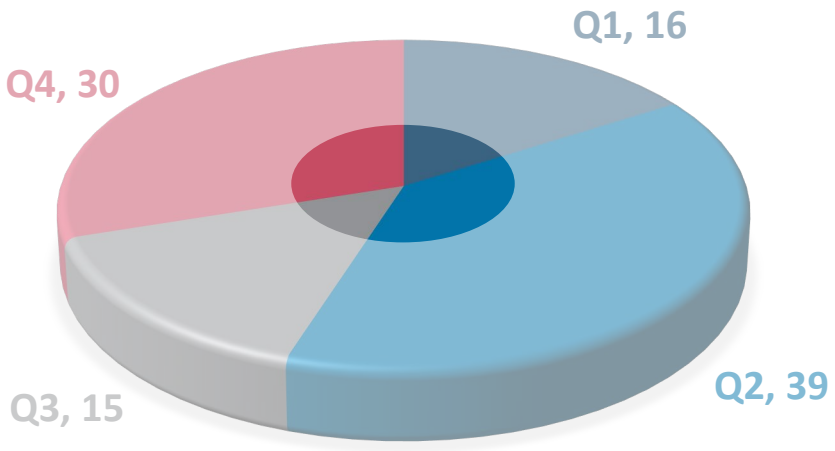


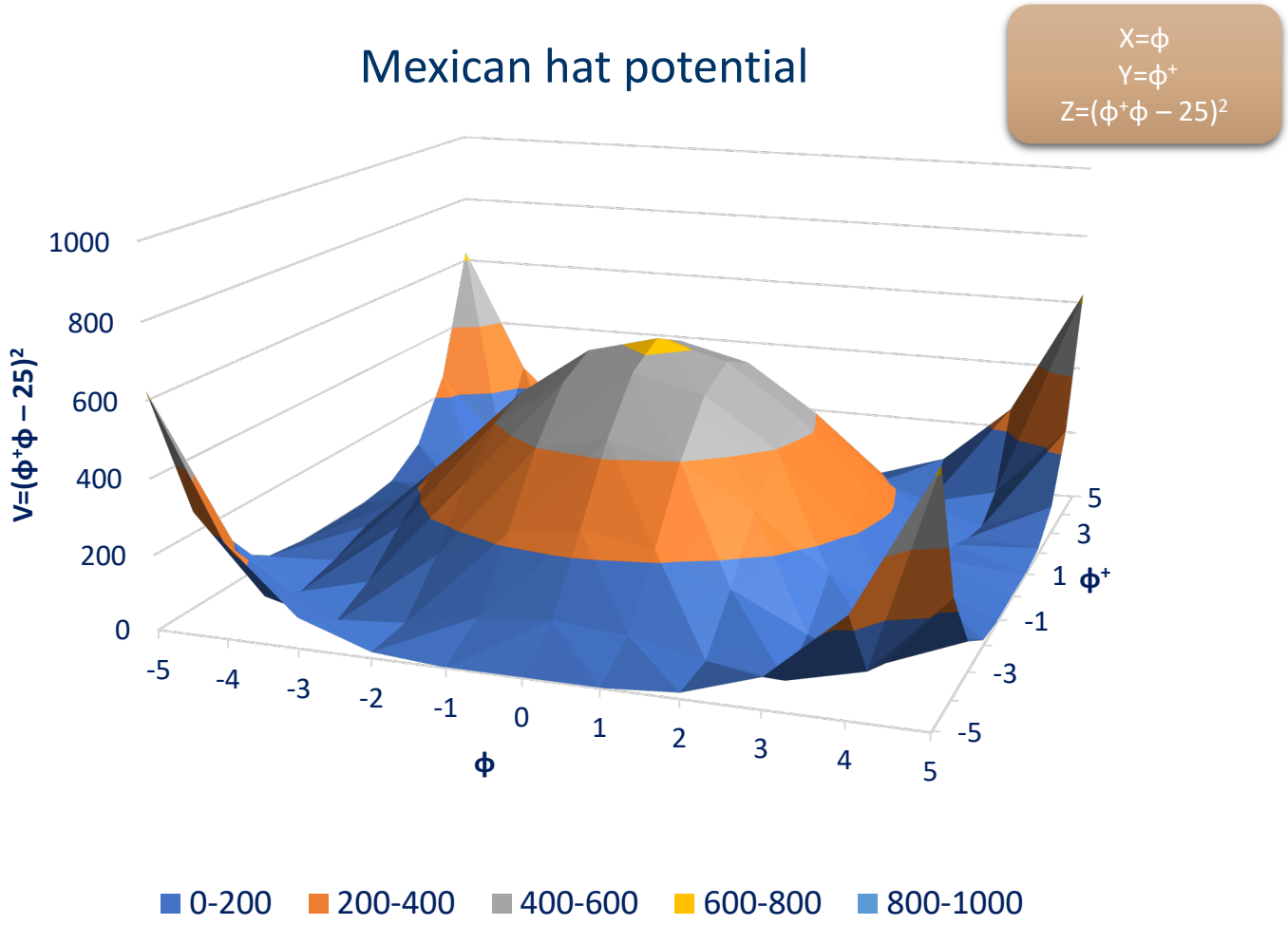


Credits #: 11, 12

3D DISTORTION IN CHARTS IS A FACT...

3D Distortion



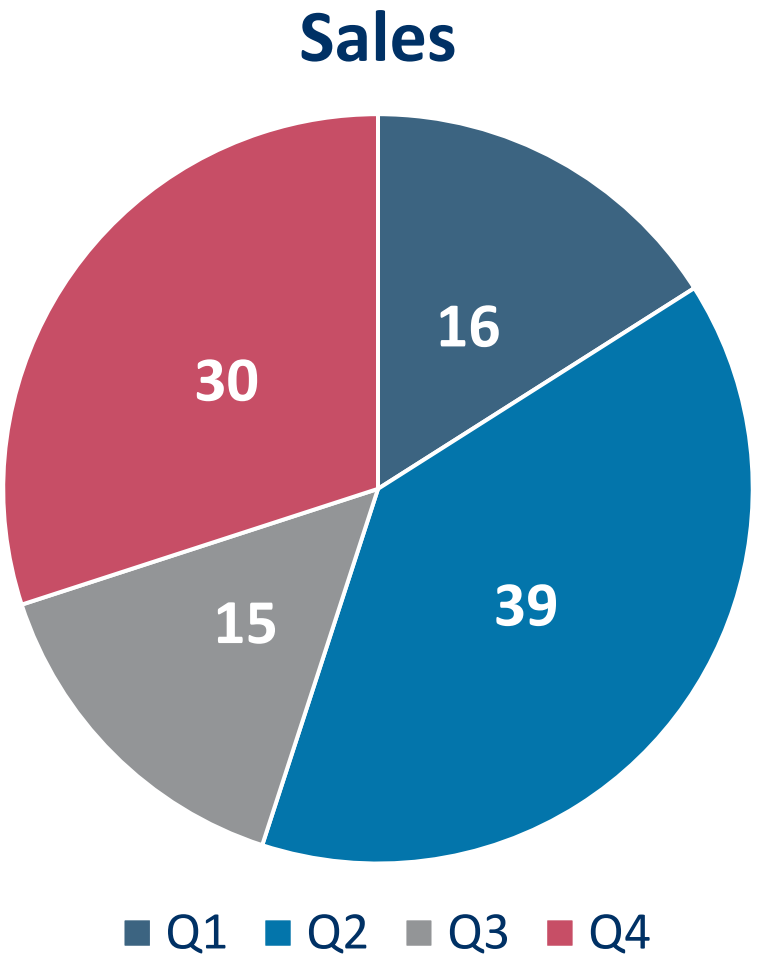


PIES

Pie charts show how a total splits in parts

- The round geometry suggests that numbers are exhausting all possibilities
- Numbers are well interpreted as percentages

- Easily overcrowded by too many sectors
- Comparisons are difficult: what is the smaller part?
- Unclear labelling
- No common reference



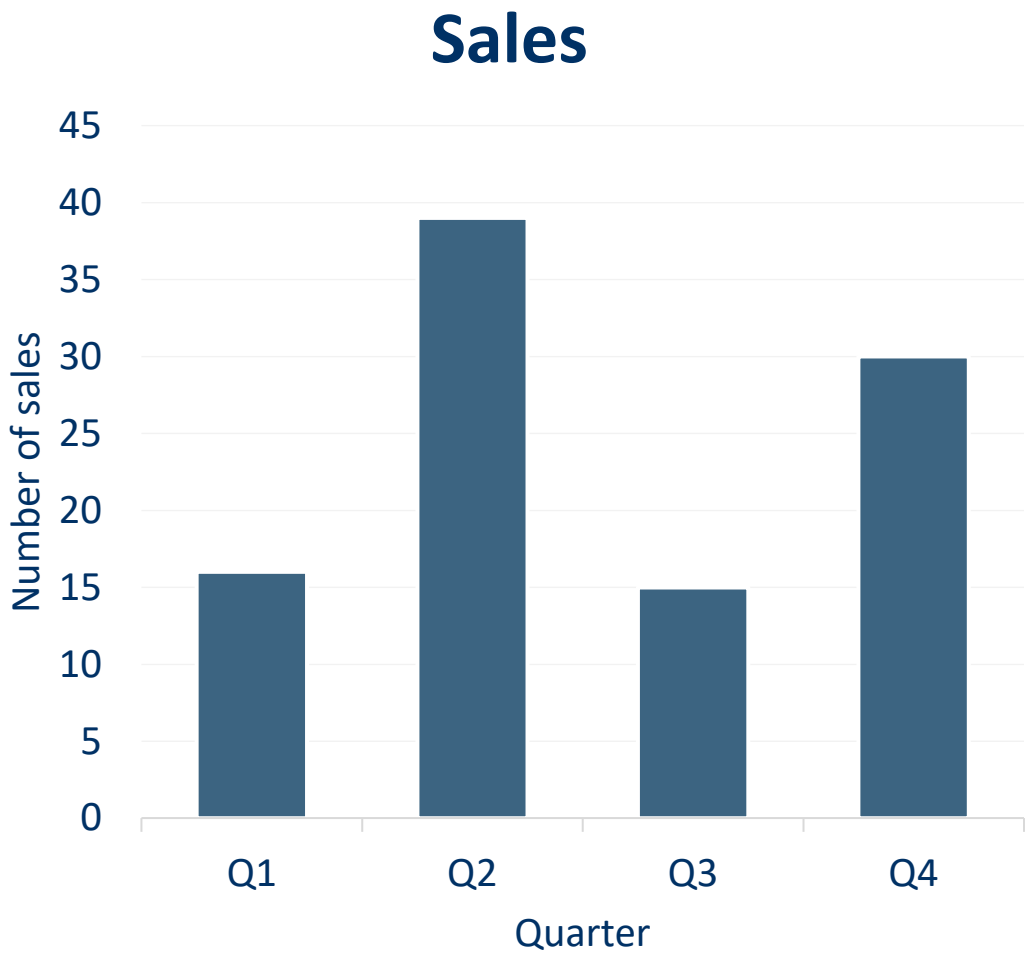
COLUMN CHARTS

Common Charts features

Column charts show how numerical data is distributed

- Columns show comparison between magnitudes
- Useful for representing ordering relations between numbers
- Large differences in magnitude can be handled “easily” through logarithmic axis usage

- Show only 2D data
- The scale on the x axis is not well defined
- May be overcrowded by vertical bars



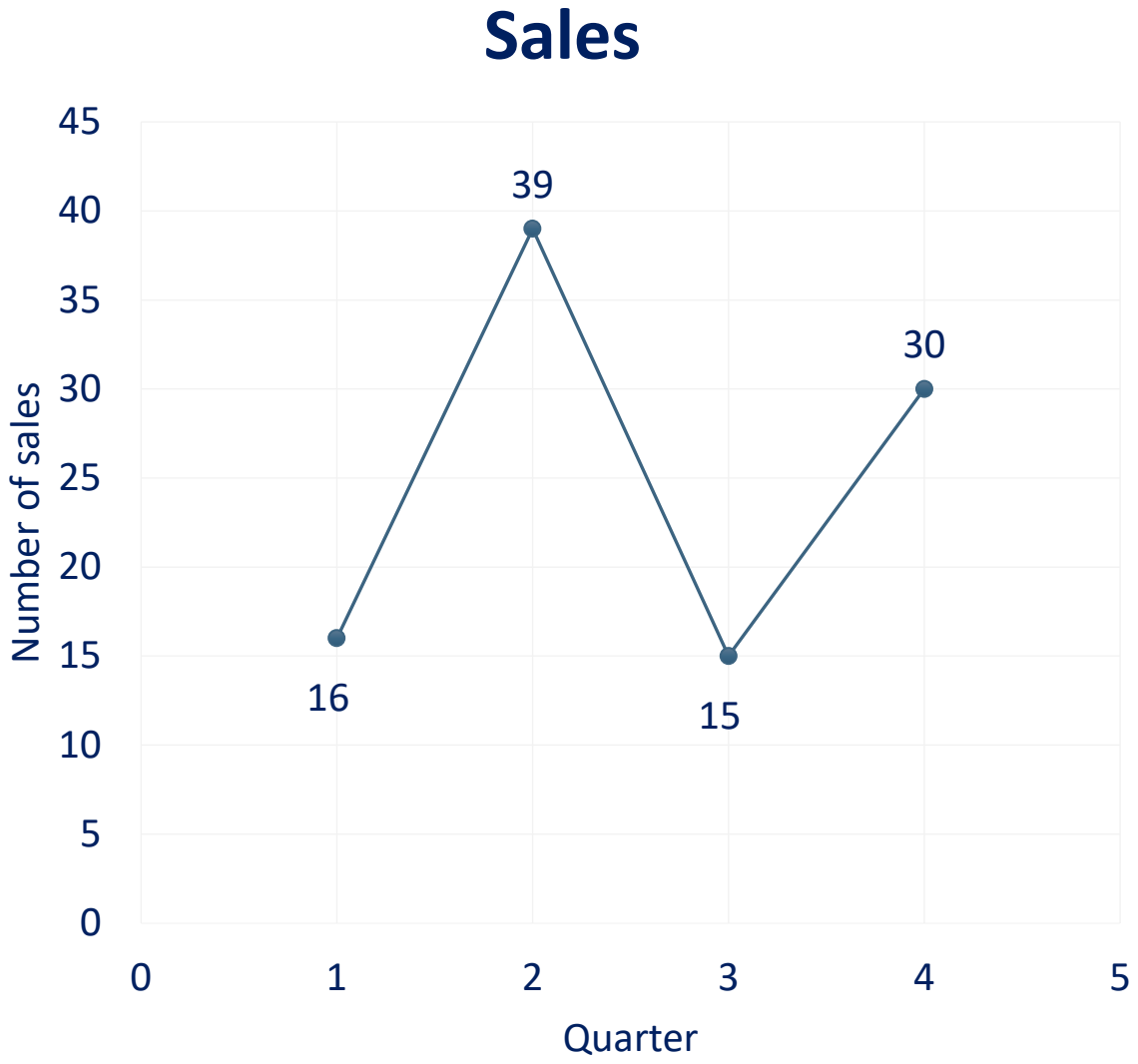
LINE CHARTS AND SCATTER PLOTS

Common Charts features

These show data points on a plane. Edges between points can map a relation among data

- They help to figure out relations or shapes among data points (e.g.: trend)
- Simple and clean

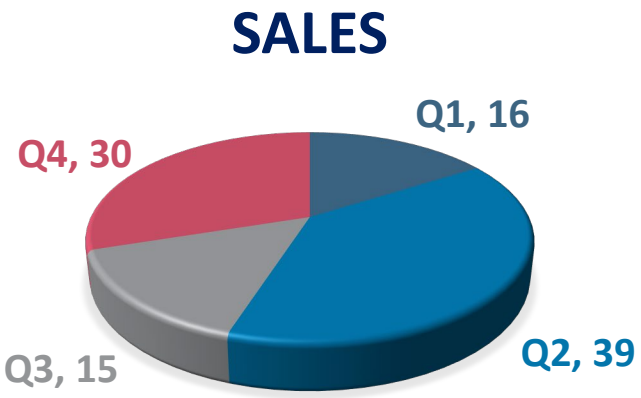
- Could be overcrowded with too many series
- Require to set scales properly to avoid misunderstandings and confusion



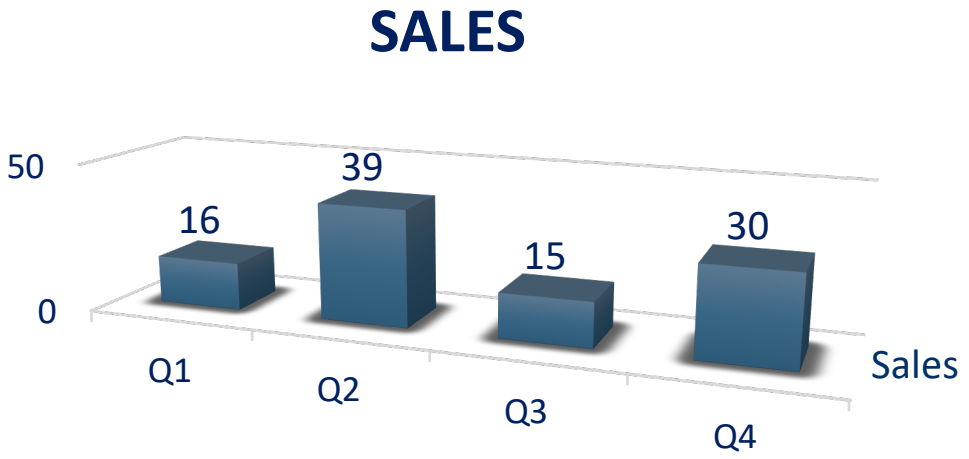
3D PLOTS

3D plots are sometimes necessary for comparison among many series

- They are (seldom) useful to compare high dimensional data
- They have the «cool effect»



- Suffer all the weaknesses of their 2D counterparts
- Perspective makes comparison difficult due to length and scales distortion
- The third dimensions is often useless and harmful in business context

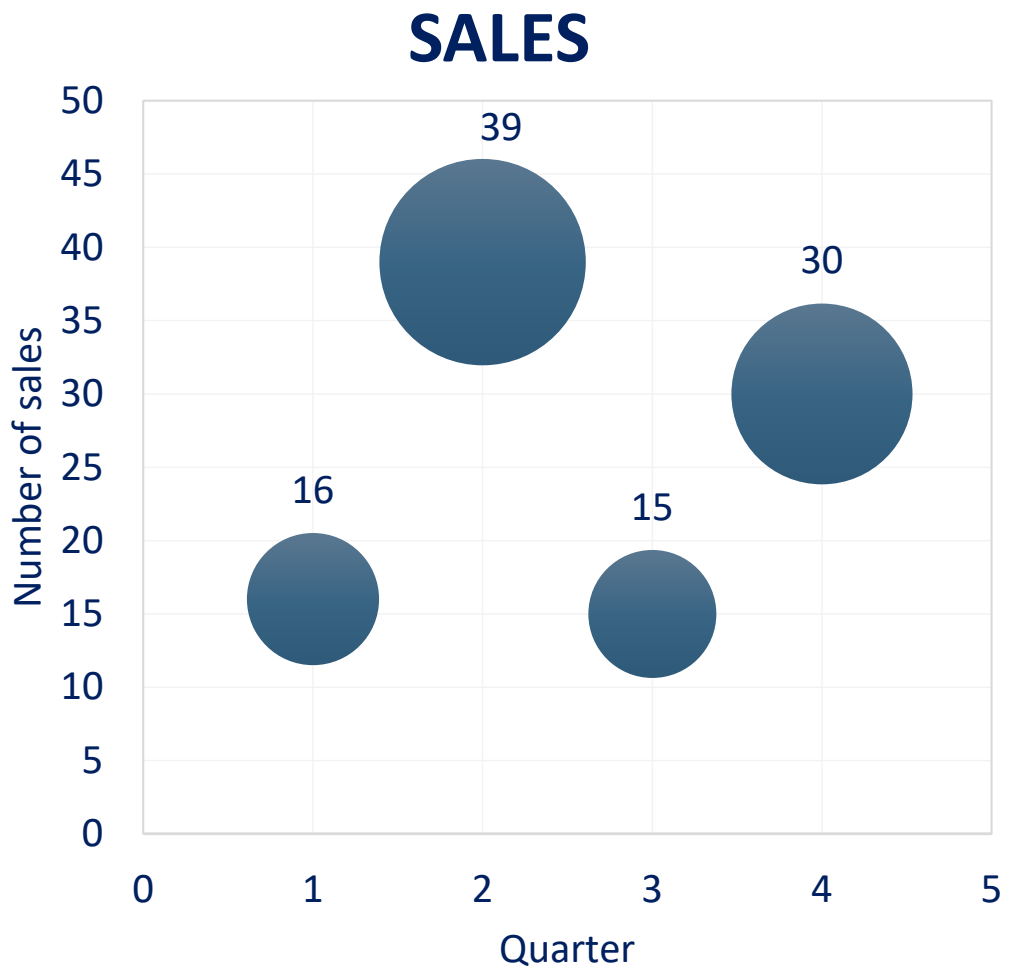


BUBBLE CHARTS

Bubble charts combine the cartesian layout with dimension/color of bubbles to show 3D data in a 2D fashion

- They are one of the most effective way to represent N dimensions on a plane
- Could help the eye to recognize proximity relations (e.g. clusters)

- Comparison of bubble dimension can be difficult
- Easily overcrowded by too many bubbles
- Overlaps of neighboring bubbles creates confusion

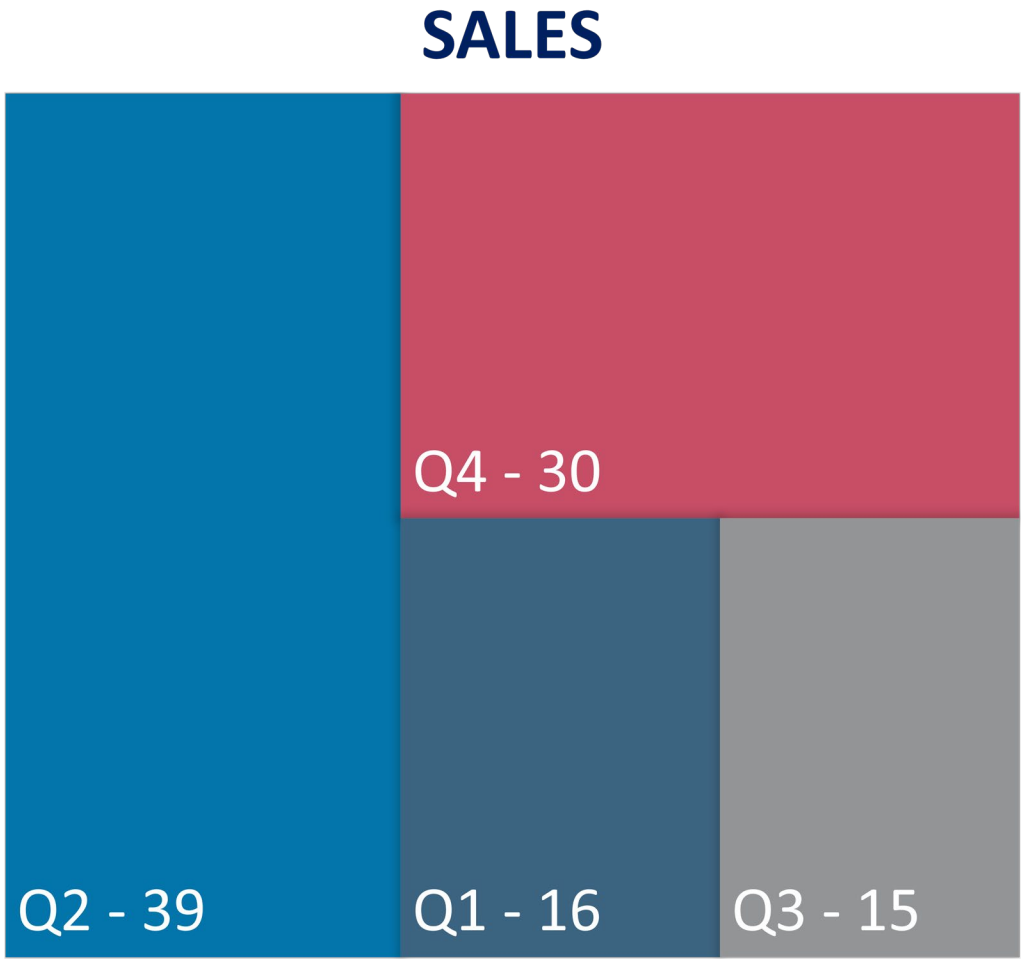


AREA CHARTS

This type of charts can help with large datasets and hierarchically structured data

- They help to figure out relations between (even large set of) objects
- Interesting cool effect

- They are less effective in showing percentages
- Don’t help to have precise comparison between categories



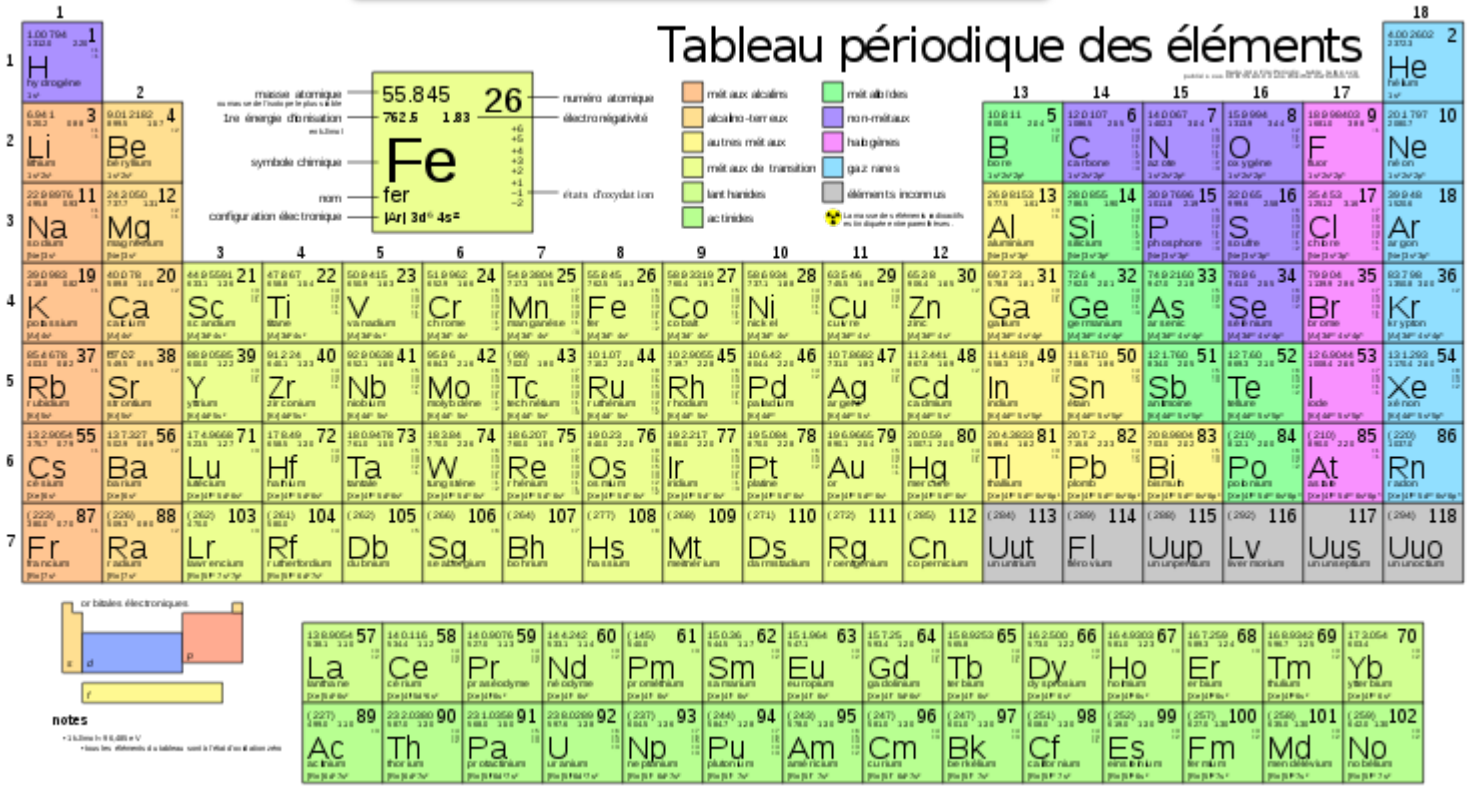
WHY BUILDING GOOD CHARTS IS SO HARD?

Bad match between message and design

Wrong audiences, which means wrong message

Difficult choice among many graphical solutions

An excellent example of Good chart



Credits #: 13

FEW GUIDELINES TO BE EFFECTIVE

WHAT TO REMEMBER

There is no absolute way of visualize data effectively

The same data may have different representations depending on the purpose of the visualization

Try many alternatives, limiting to one solution is almost always wrong

WHAT TO AVOID

Falling in love with coolness

Wasting your audience attention

Forgetting limited tool scalability

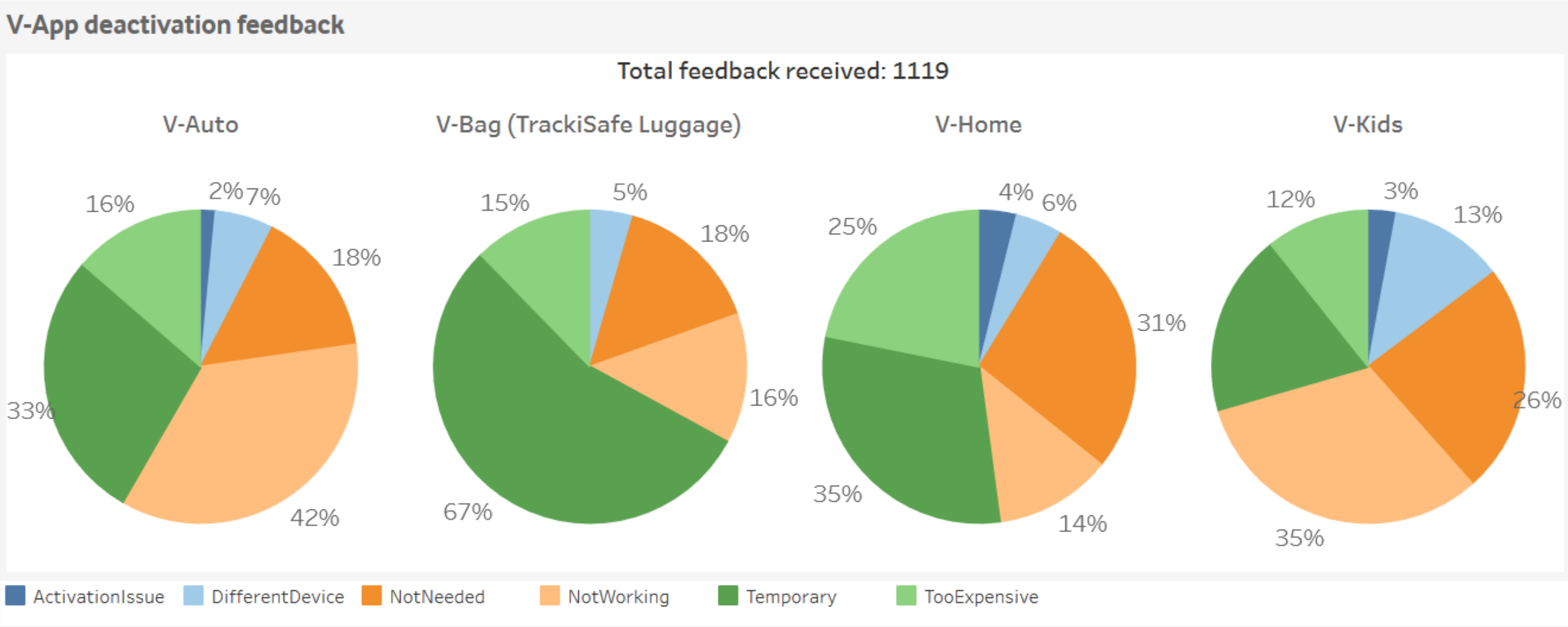
KNOW YOUR AUDIENCE

Understanding for whom the visualization is becomes necessary for any good chart

Satisfied audiences can be by far better than having found the best solution

CIoT EXAMPLE – PIE CHART

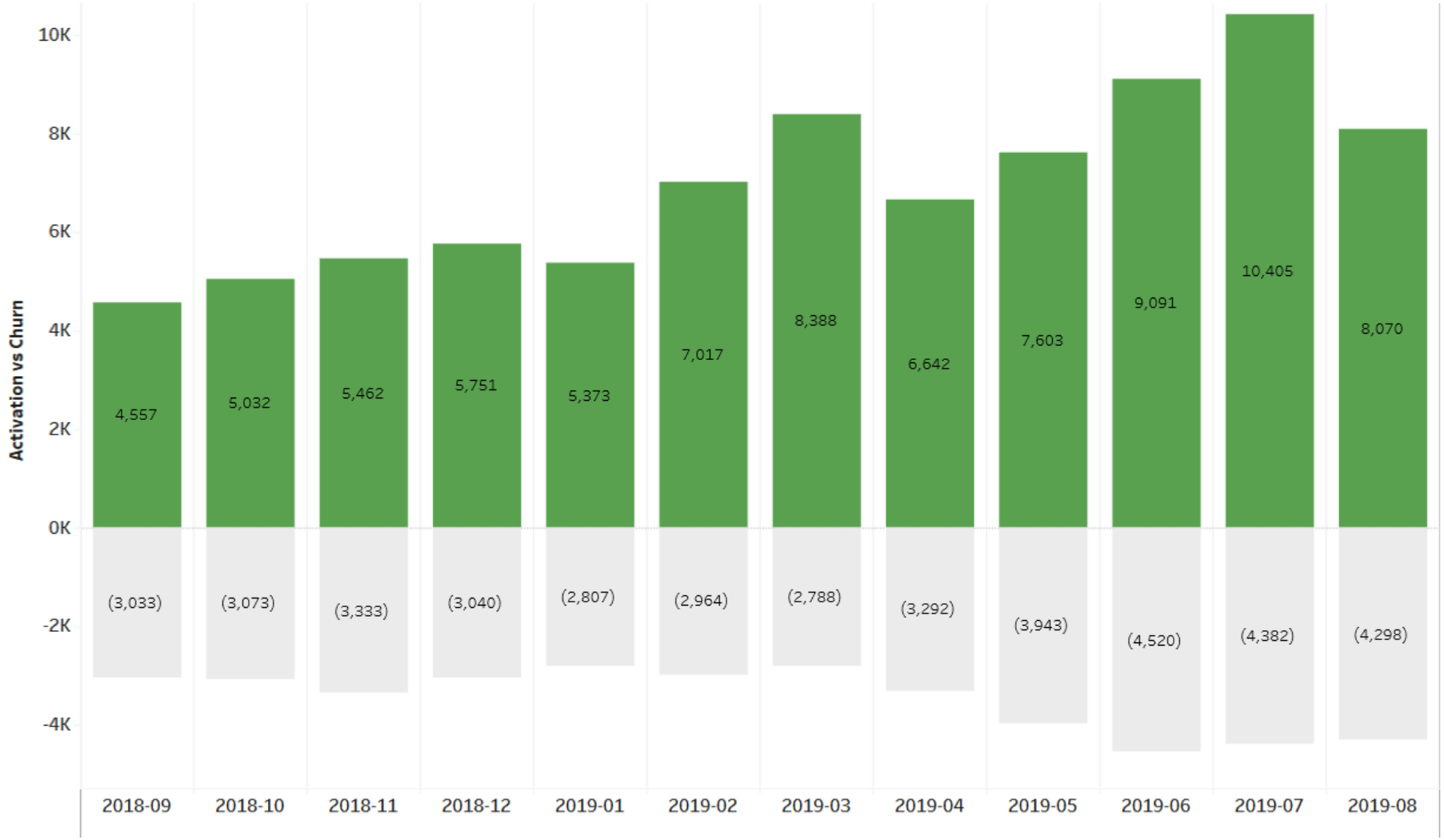
Data flow



CIoT EXAMPLE – COLUMN CHARTS

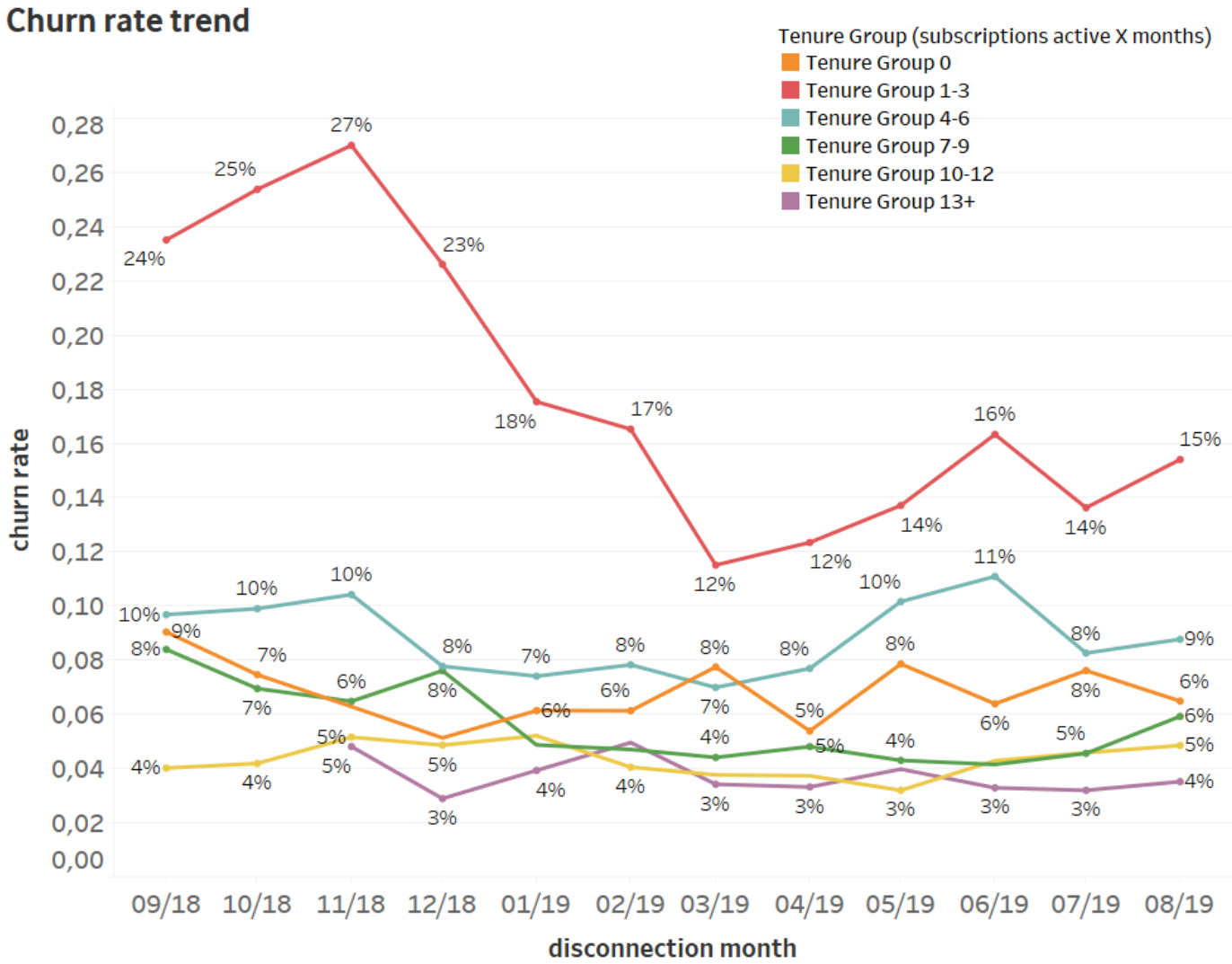
Data flow

Activations vs Churn over time



CIoT EXAMPLE – LINE CHART

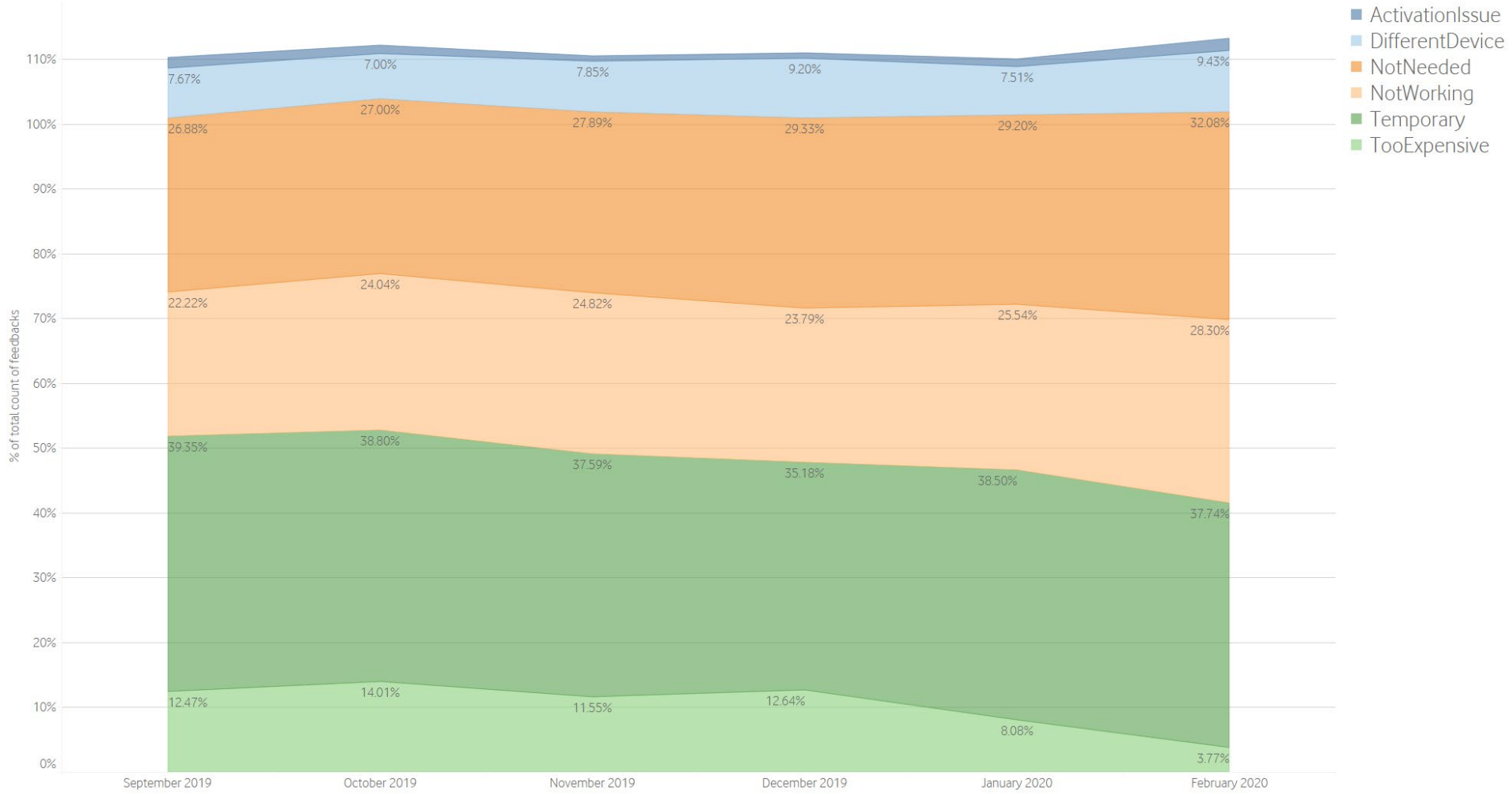
Data flow



CIoT EXAMPLE – AREA CHART

Data flow

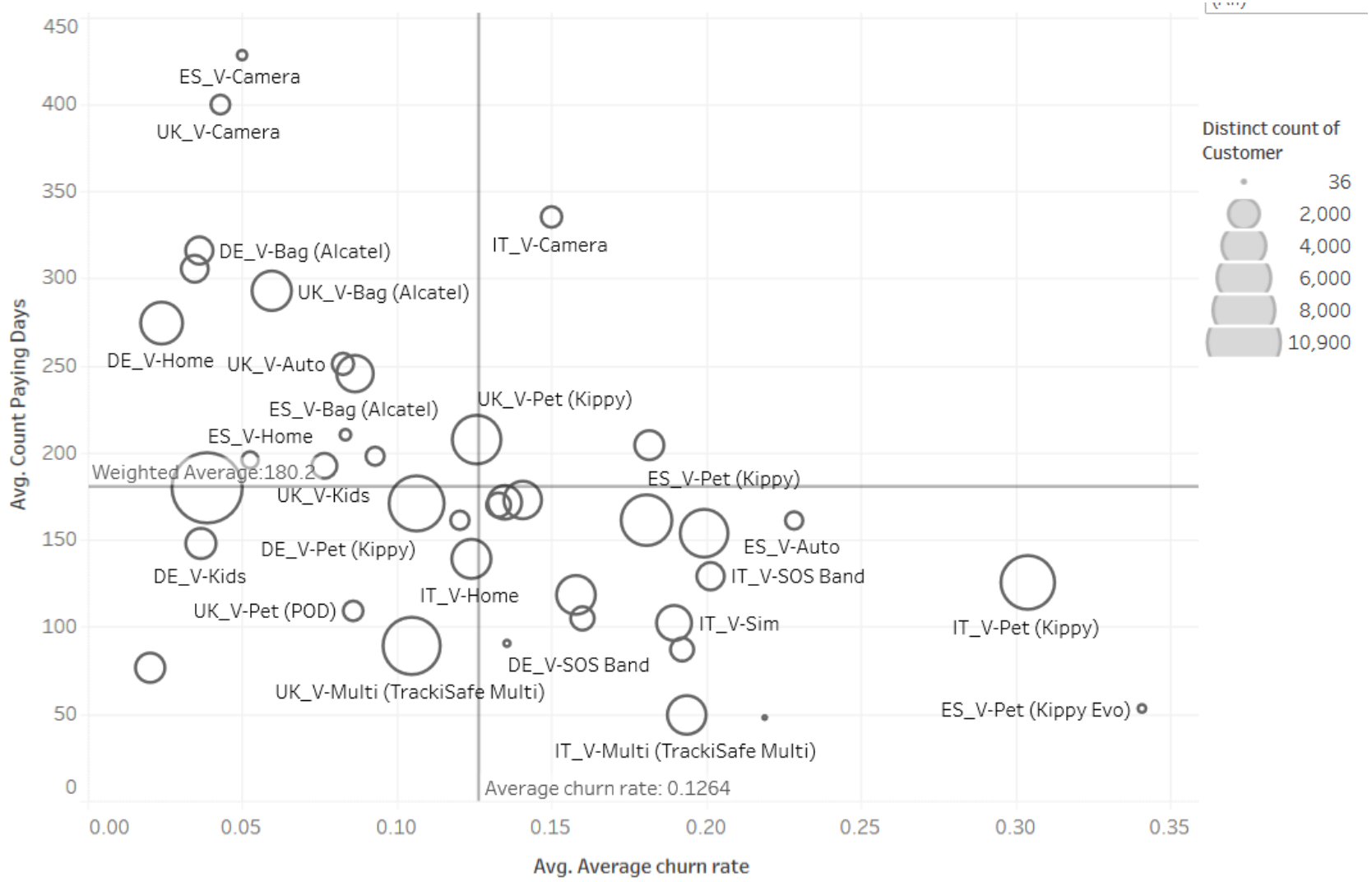
Feedback per deactivation reason



CIoT EXAMPLE – BUBBLE CHART

Average churn rate vs average days of activity

Data flow



AGENDA

- Introduction to DV
- Data representation
- Enterprise tools
- Classwork

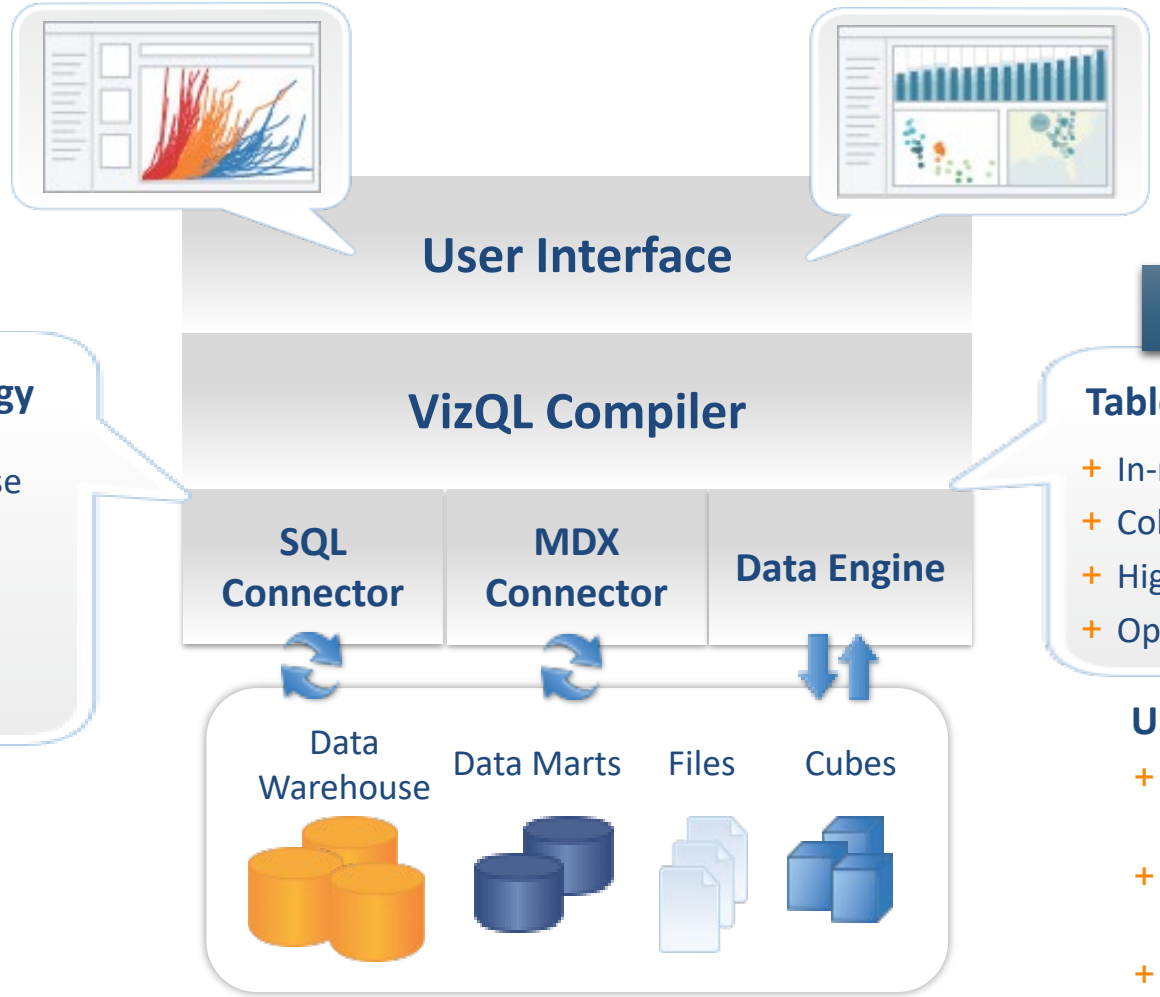
TABLEAU - OVERVIEW

Enterprise tools

- It is an interactive data visualization tool
  - Main data viz tool within CP&S and in other VF departments
- Short company history:
  - It was born in 2003 and went public in 2013
  - 877 Millions $ revenues in 2017, with around 4,000 employees
  - In 2018 it announced the acquisition of Empirical Systems, an AI startup, with the aim of integrating their algorithms in the project
  - In June 2019 Salesforce announced the acquisition of Tableau

TABLEAU - ARCHITECTURE

Enterprise tools



Live

Tableau Native Query Technology

+ Interactive queries to the database
+ Leverages IT policies for security
+ Ensures fresh data results
+ Values the EDW strategy

Extract

Tableau Data Engine Technology

+ In-memory
+ Column store
+ Highly compressed
+ Optimized API specific for Tableau

Use this option:

+ When the source DB query performance is slow
+ To offload iterative query workload from the source DB
+ To work offline from the network
+ To keep an archive of the data

SPLUNK – OVERVIEW

splunk>

Enterprise tools

- It is a full stack solution for data ingestion, analysis and visualization
  - Capture, indexes and correlates real-time data in a searchable repository
- Short tool history
  - Born in 2003 as a Log Collector system
  - 1.27 bln $ in 2017, ~4,500 employees
  - Leveraging a highly performing indexing methodology of dealing with large amounts of data in real-time, it extended its scope to several fields of application
    - Application management, security and compliance, business and web analytics
      - Became one of the most used and flexible enterprise tool for business intelligence, data analysis and visualization

SPLUNK - EXAMPLES

splunk>

Enterprise tools

Credits #: 18

VISUALIZATION TOOLS @VODAFONE CP&S

Enterprise tools

splunk>

+ableau

OTHER VISUALIZATION TOOLS

Enterprise tools

Power BI

ARCADIA DATA

QlikView

sas VISUAL ANALYTICS

IBM Watson

TIBCO Spotfire®

kibana

BY THE WAY…

Enterprise tools

GARTNER MAGIC QUADRANT

Analytics and Business Intelligence platform 2019

Source: Gartner (February 2019)

AGENDA

- Introduction to DV
- Data representation
- Enterprise tools
- Classwork

CLASSWORK

Fitness Gym Company

Assignment

The company has several gyms in different location.
In this scenario you are the manager of one gym and your objective is to present to your boss (a manager of the company) the situation of the venue. Here is your data:

- Opening hours
- Staff Scheduling
- Staff Salaries
- Revenues
- Number of customers (overall)
- Number of customers (per gym section)
- Gym equipment

GIVE YOUR BEST SHOT!

And remember that:

VIZ PRINCIPLES

- ✓ Show the data
- ✓ Provoke thought
- ✓ Avoid distortion
- ✓ Present many numbers in a small space
- ✓ Make large datasets coherent
- ✓ Encourage eyes to compare data
- ✓ Reveal data at several levels of detail
- ✓ Serve a reasonably clear purpose

VIZ VARIABLES & DIMENSIONS

- Color
- Position
- Mark
- Size
- Brightness
- Motion
- Orientation
- Texture

VIZ ENEMIES

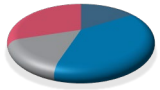
- ⚠ Confusion
- ⚠ Misleading choices
- ⚠ Junk pictures
- ⚠ 3D distortion

VIZ CATEGORIES

Pie 

Column chart 

Line chart & scatter plot 

3D plots 

Bubble chart 

Area chart 

15 minutes

SCHEDULING

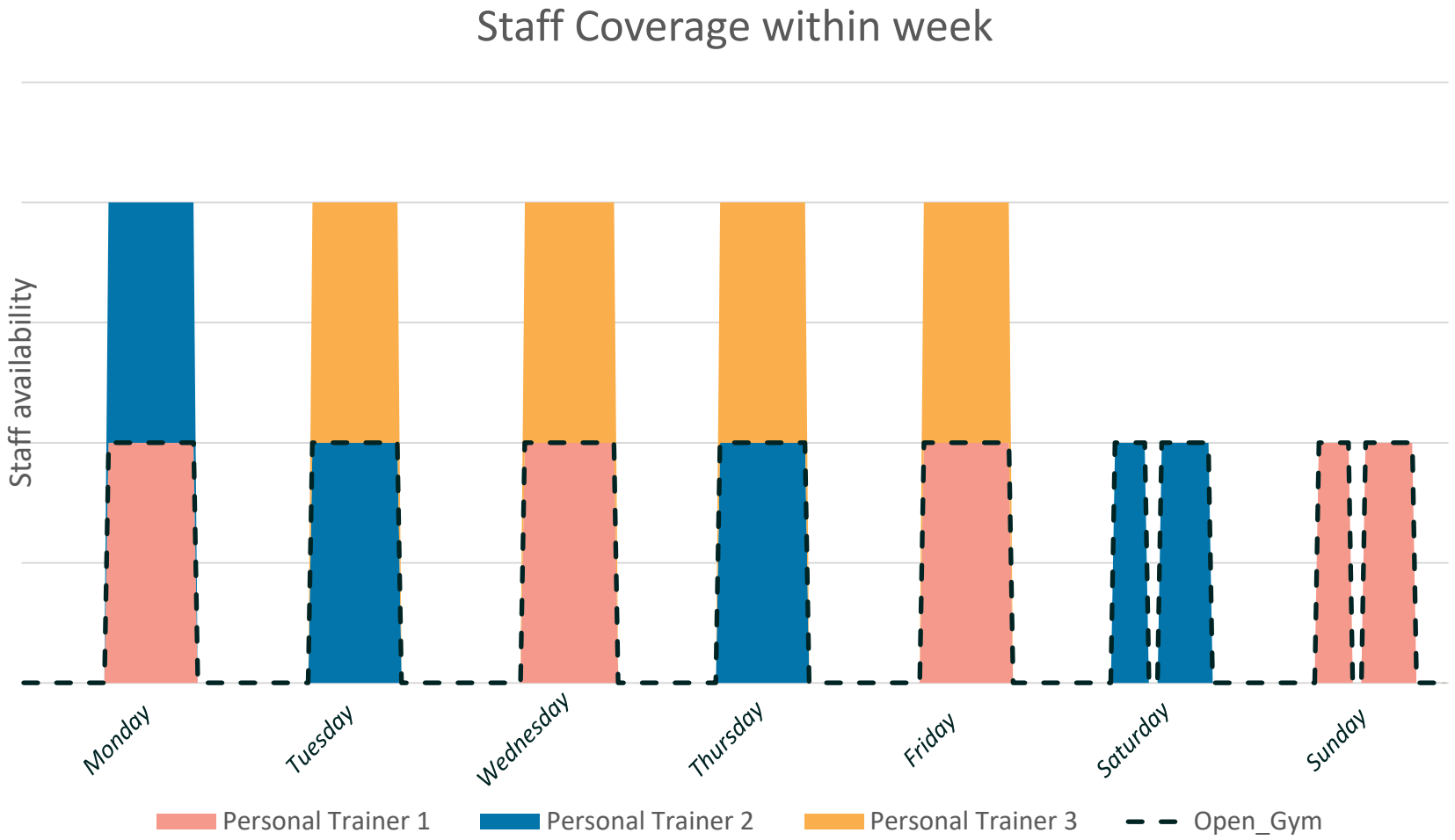
Possible Output

GYM opening hours

Mon, Tue, Thu, Fri
10:00 – 20:30

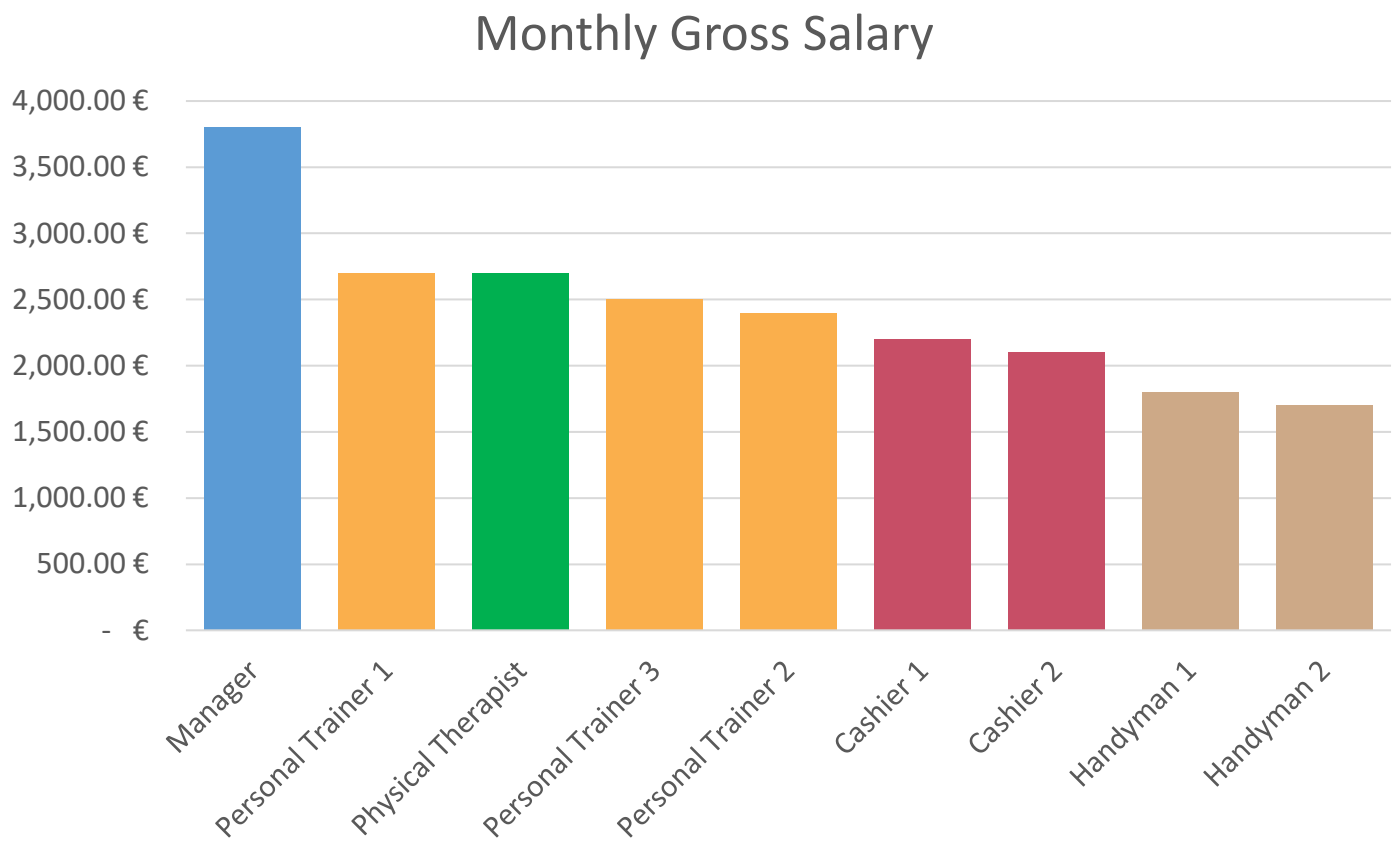
Wed
11:00 – 22:00

Sat-Sun
8:30 – 12:30
14:00 – 20:00



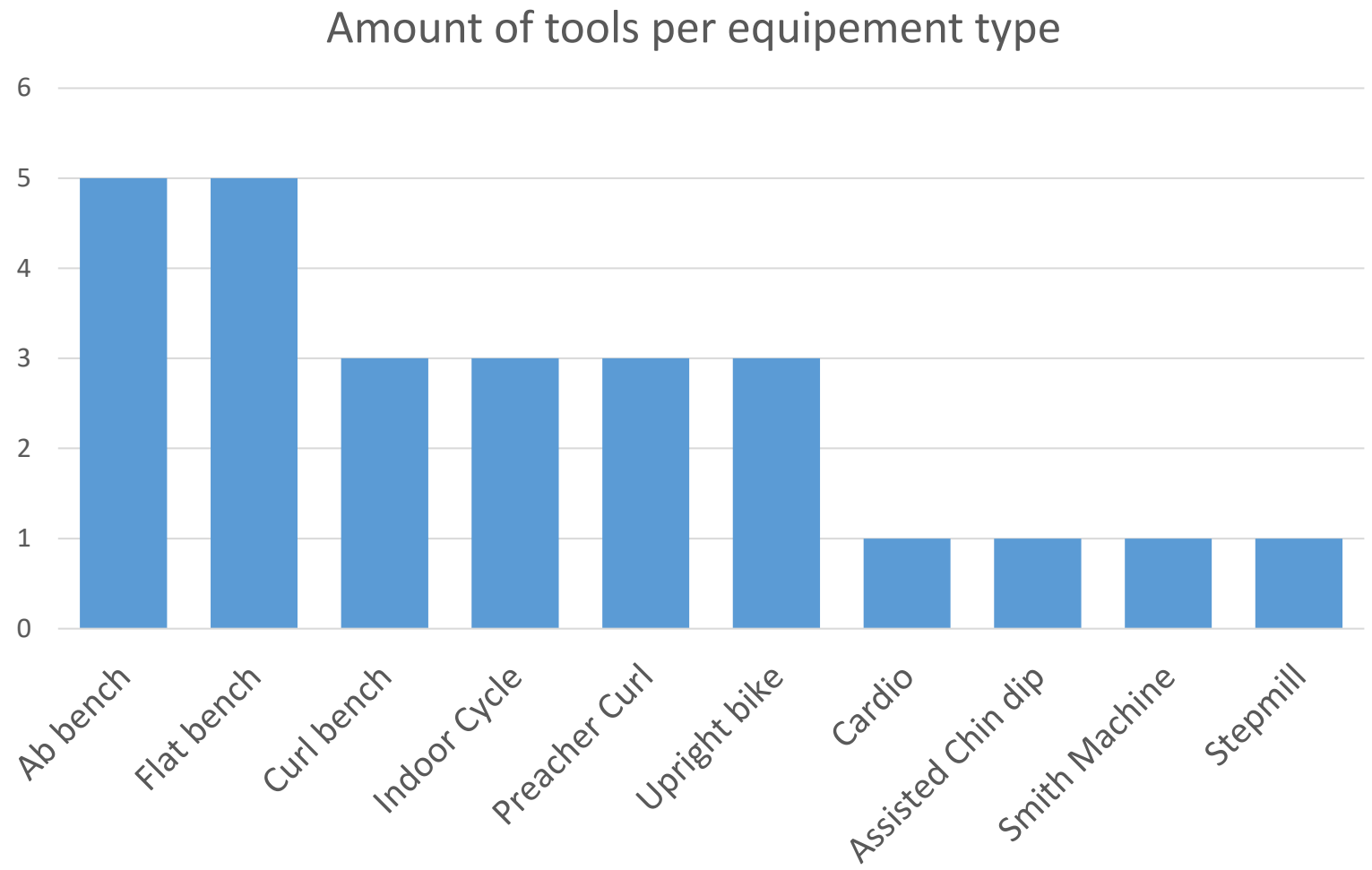
SALARIES

Possible Output



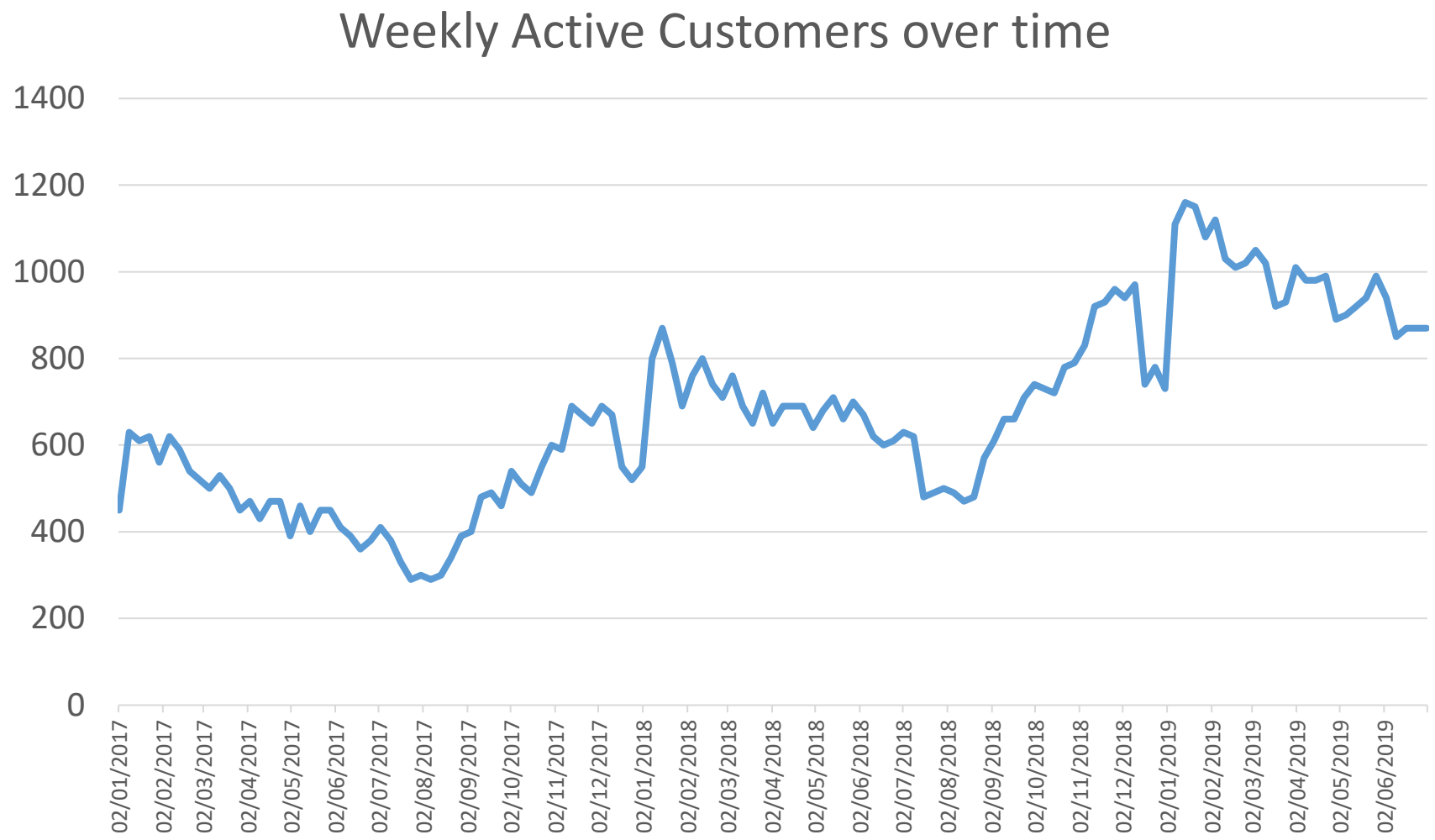
EQUIPMENT

Possible Output



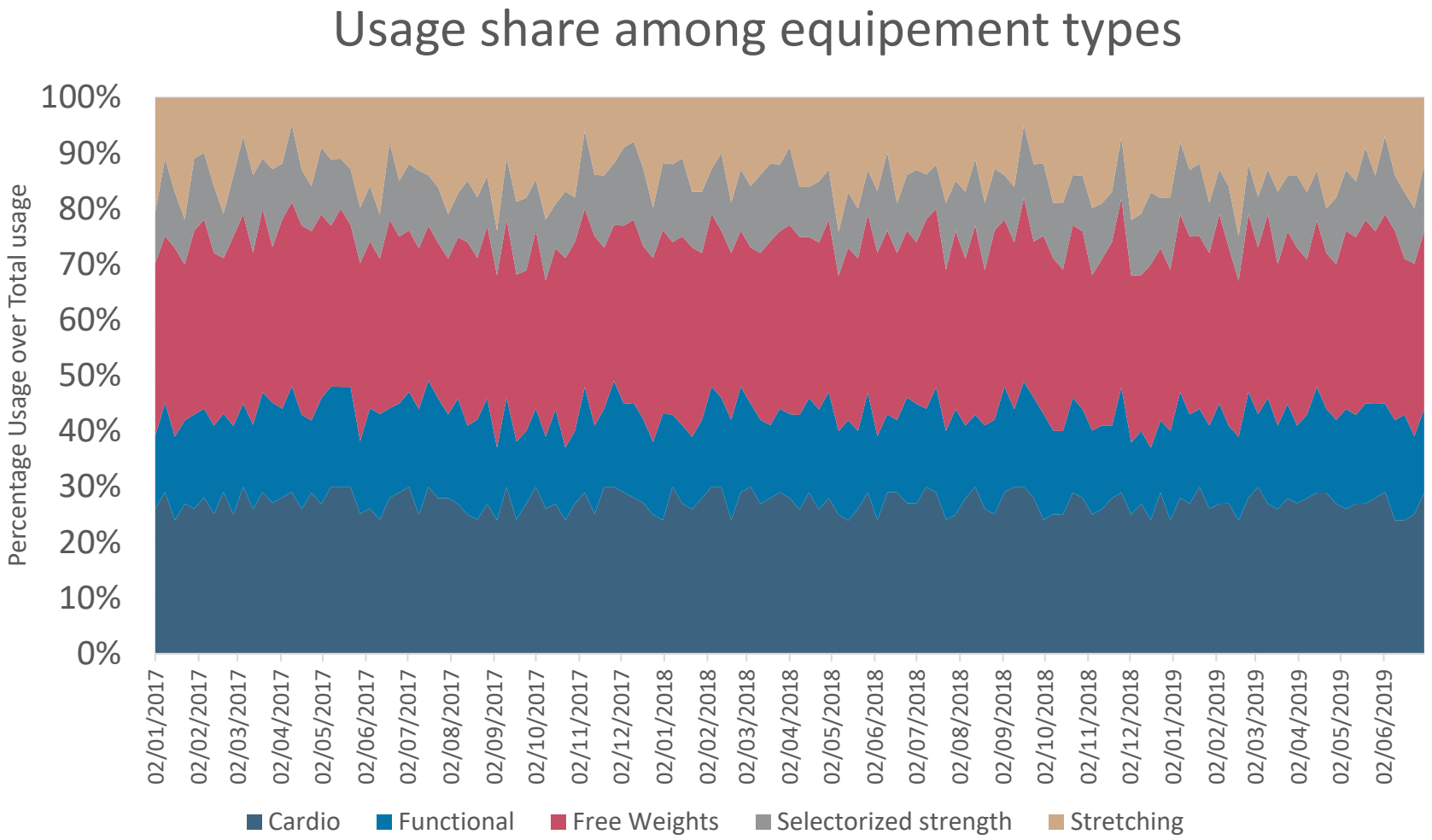
CUSTOMERS OVER TIME

Possible Output



EQUIPMENT USAGE SHARE

Possible Output

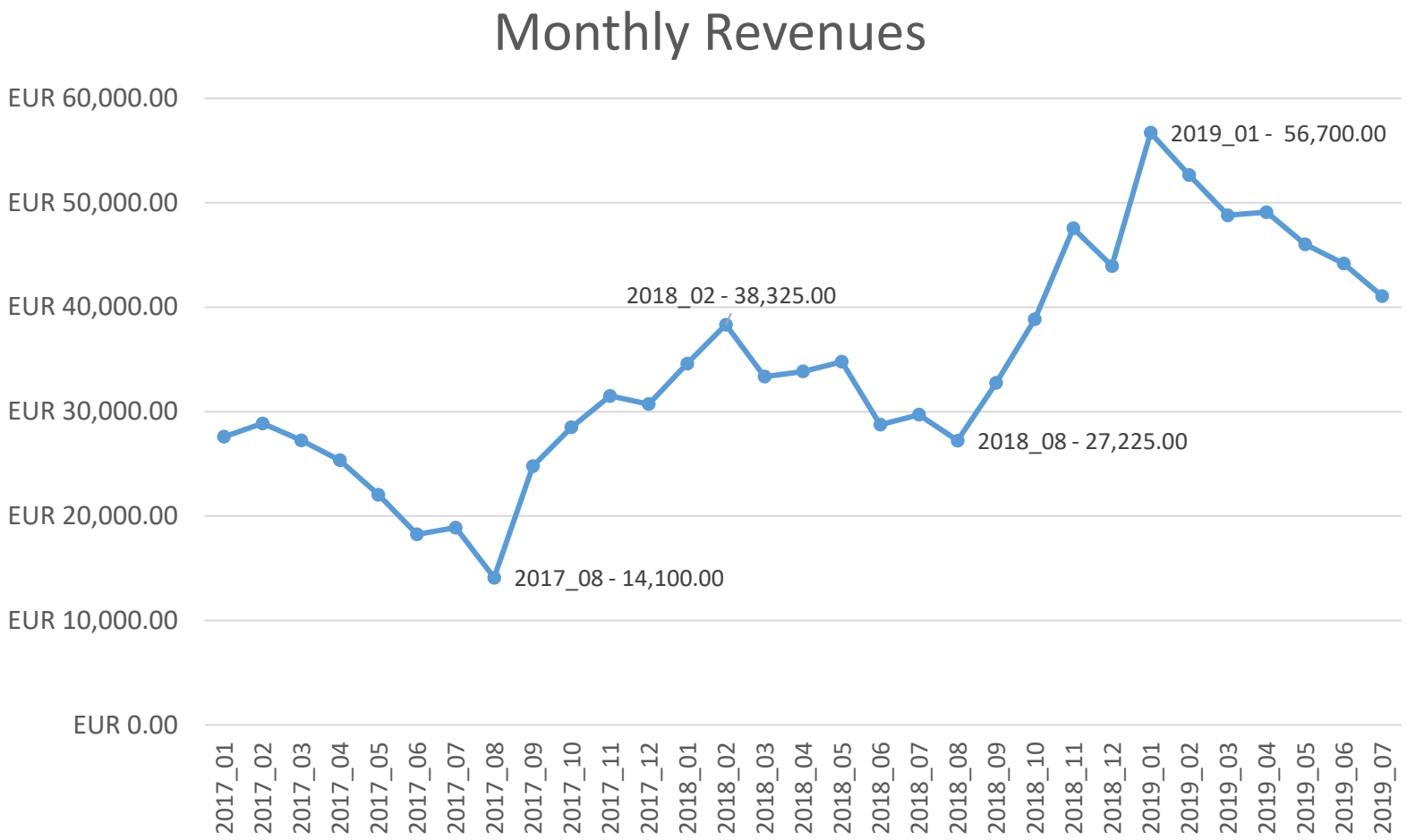


AVERAGE

Cardio	14.8%
Functional	10.9%
Free Weights	31.1%
Selectorized strength	16.0%
Stretching	27.2%

REVENUES

Possible Output



Lunch Break
See you at 13:30

PICTURES CREDITS

1: http://eol.jsc.nasa.gov

2: http://www.atlas.cid.harvard.edu

3: http://populationpyramid.net

4: https://medium.com/@plotlygraphs/3-minimalist-dashboards-with-great-style-bbd0f3491599 - Based on NY state environmental data (https://www.dec.ny.gov/energy/1601.html)

5: https://www.flickr.com/photos/diametrik/30931836692 - Licensed under Creative Commons "Attribution" BY 2.0 (https://creativecommons.org/licenses/by/2.0/)

6: https://en.wikipedia.org/wiki/File:Minard.png

7: https://data.worldbank.org - Licensed under Creative Commons "Attribution-Share Alike" BY-SA 4.0 (https://datacatalog.worldbank.org/public-licenses#cc-by)

8: https://trends.google.com/trends

9: https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population - Licensed under Creative Commons "Attribution – Share Alike" unported BY – SA 3.0 (https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License)

10: Based on data from GISTEMP Team, 2019: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. Dataset accessed 2019-08-23 at https://data.giss.nasa.gov/gistemp/ - Lenssen, N., G. Schmidt, J. Hansen, M. Menne,A. Persin,R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. J. Geophys. Res. Atmos., 124, no. 12, 6307-6326, doi:10.1029/2018JD029522

11: http://www.fao.org/faostat/

12: https://commons.wikimedia.org/wiki/File:Bananas.JPG - Licensed under Creative Commons "Attribution – Share Alike" unported BY – SA 3.0 (https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License)

*13:*https://fr.wikipedia.org/wiki/Fichier:Tableau_p%C3%A9riodique_des_%C3%A9l%C3%A9ments_pr%C3%A9cis.svg *- Licensed under Creative Commons "Attribution – Share Alike" unported BY – SA 3.0 (https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License)*

14: https://commons.wikimedia.org/w/index.php?curid=74539538– Licensed under Creative Commons "Attribution-Share Alike" BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/deed.en)

15 https://commons.wikimedia.org/w/index.php?curid=74539785 – Licensed under Creative Commons "Attribution-Share Alike" BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/deed.en)

16: https://commons.wikimedia.org/w/index.php?curid=74539541– Licensed under Creative Commons "Attribution-Share Alike" BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/deed.en)

17: https://docs.microsoft.com/it-it/power-bi/media/sample-sales-and-marketing/sales1.png

18: http://dev.splunk.com/view/webframework-codeexamples/SP-CAAAEWK

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 1

Kick Start

BUSINESS CASES PRESENTATION

DATA VISUALIZATION

Lunch Break

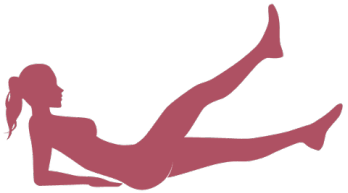
13:30

DATA ANALYSIS (Basic)

16:00

DATA FITNESS PROGRAM: Abs sculpting

Analyzing the data is the backbone of Data Science,
as abdominals are for the body



AGENDA


Data processing


Data manipulation


Data analysis


Classwork


Data flow

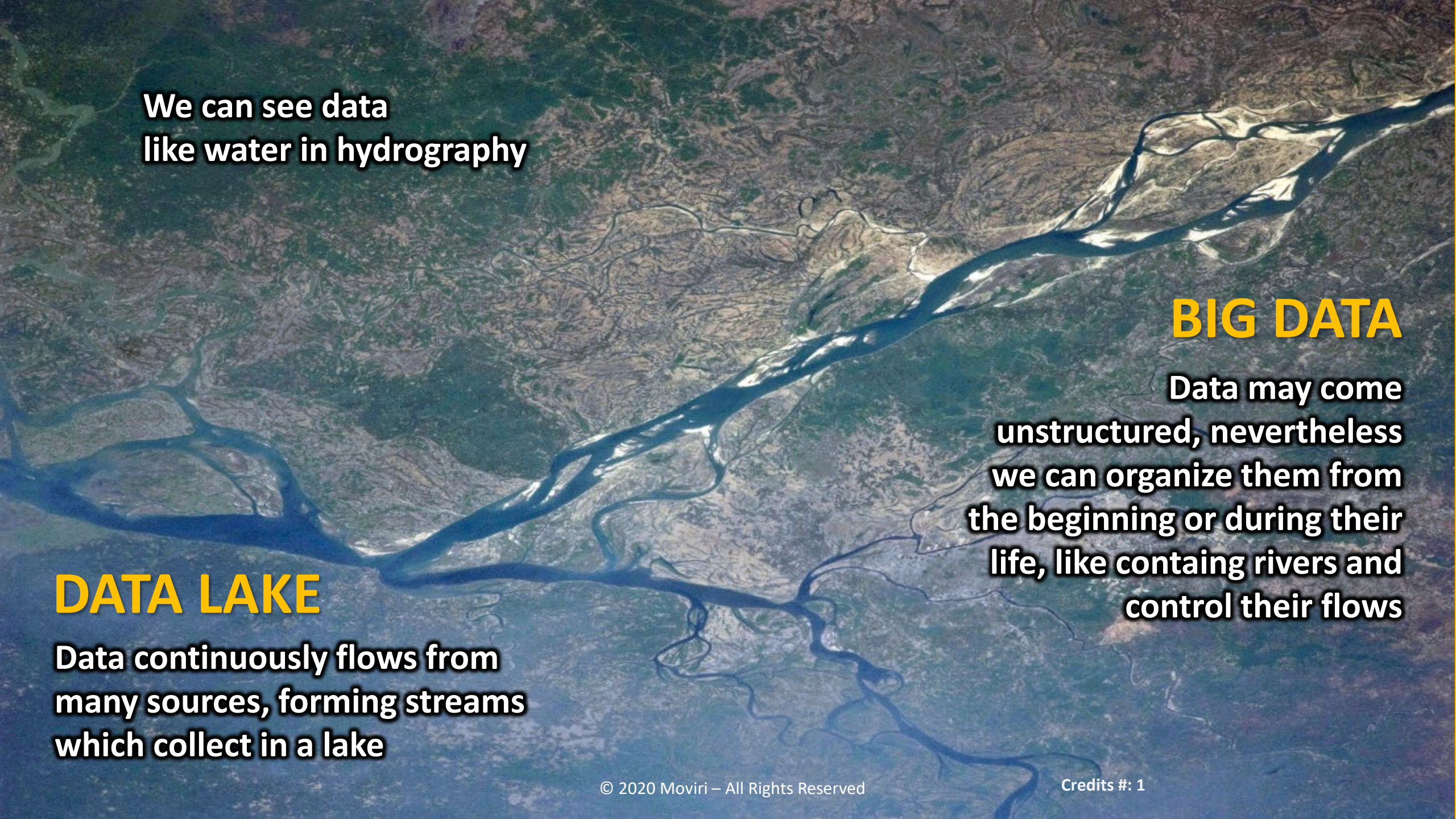

Context knowledge


Data model


Structured data


Data quality

C2 General



Credits #: 1

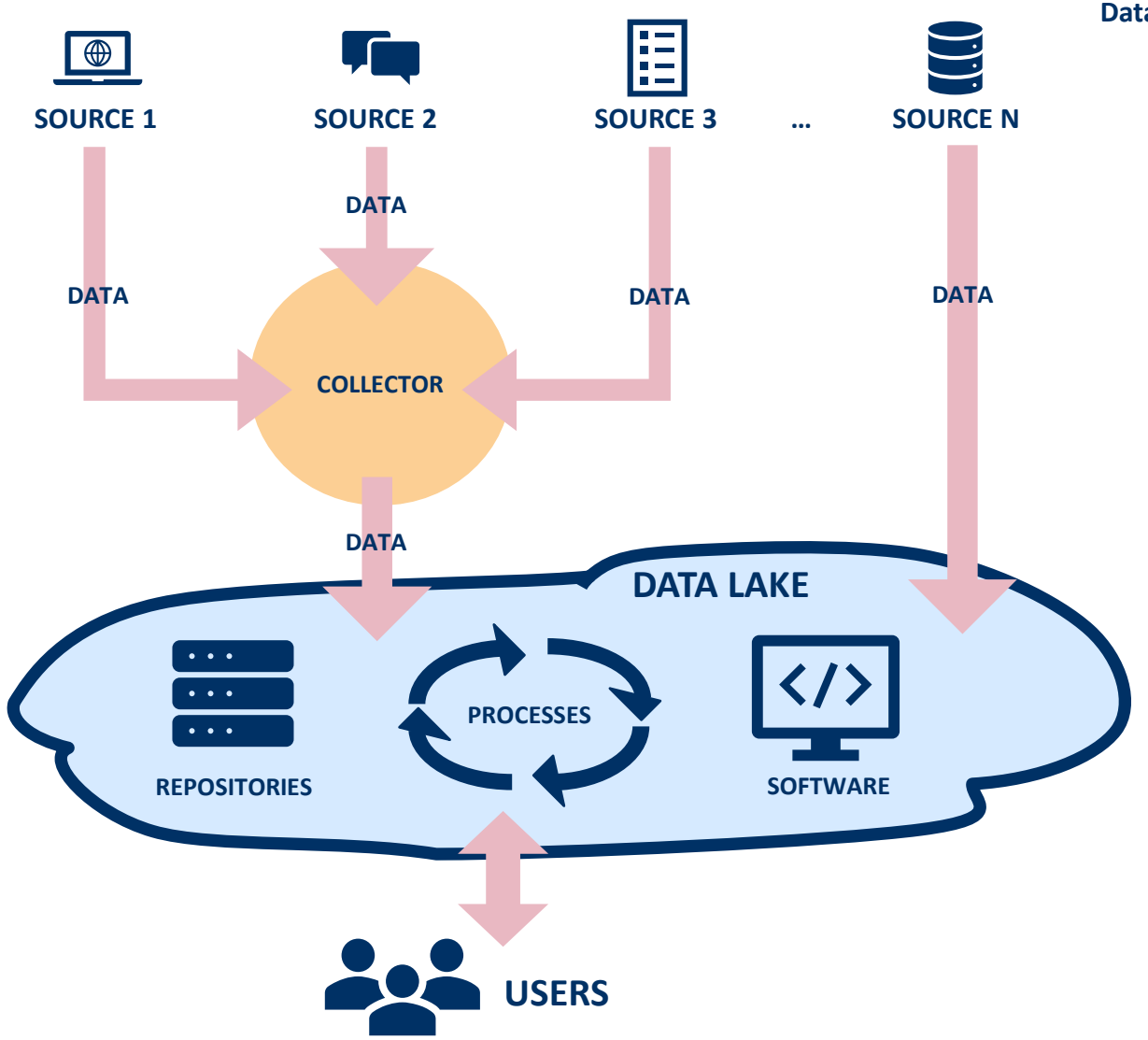
FROM SOURCES TO USERS

Data flow

Data coming from different sources can flow directly or being collected and manipulated before entering the lake

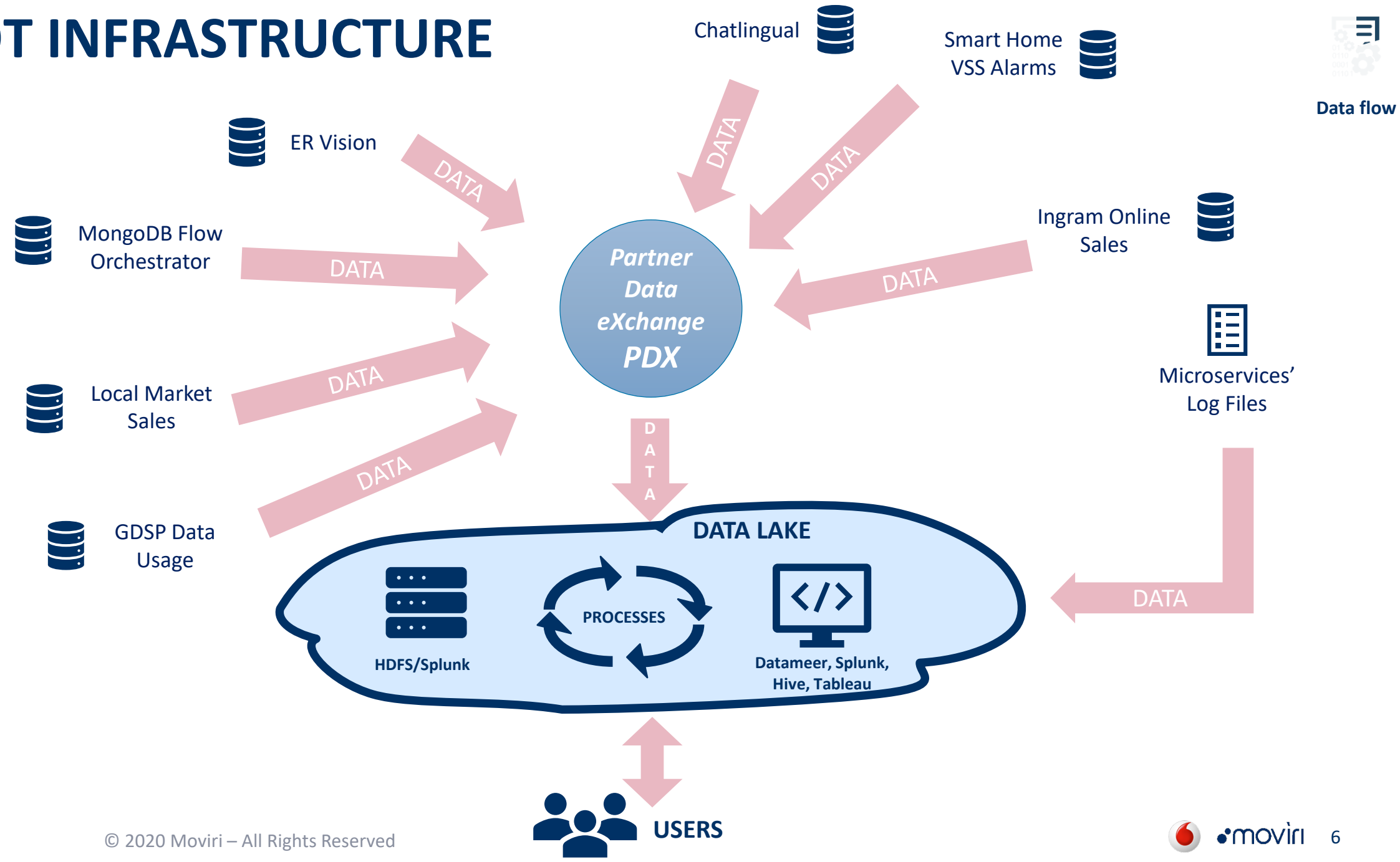
The lake can store raw and manipulated data.
Data governance is part of the lake itself

Users access data in the lake, not at the sources



C2 General

CIOT INFRASTRUCTURE



UNSTRUCTURED DATA

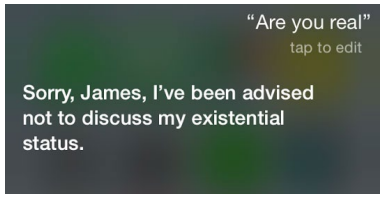
Sometimes original data has an inner structure, but it is too complex to be perceived before re-ordering:

- Text in ordinary writings
- Chats, emails and messages
- Images taken by security cameras
- Codes of arbitrary length
- Photos
- Human voice
- Signals coming from *hearing* instruments
- Sampling of continuous fields

```python
# Python program to get average of a list
def Average(lst):
    return sum(lst) / len(lst)

# Driver Code
lst = [15, 9, 55, 41, 35, 20, 62, 49]
average = Average(lst)

# Printing average of the list
print("Average of the list =", round(average, 2))
```



UNSTRUCTURED DATA – CIoT EXAMPLE

Data flow

Customer support chat with the customer who wishes to cancel V-Sim subscription

08:33:28 Customer: I wish to cancel my v sim subscription
08:33:49 Agent: Welcome to V by Vodafone, you're chatting with [redacted].
08:33:59 Agent: Sure, will definitely help you with that.
08:34:10 Agent: May I start by having your name, please?
08:34:48 Customer: [redacted]
08:34:54 Agent: Thank you, [redacted].
08:35:09 Agent: Can I have the mobile number attached to your V-SIM, please?
08:35:23 Customer: [redacted]
08:36:01 Agent: Would you mind me asking you some security questions so that I can access your account?
08:36:12 Customer: No problem
08:38:07 Agent: Do you still have your V by Vodafone app?
08:38:21 Customer: Yes that's what I'm on now
08:39:17 Agent: You will find the option from you account on the V app, through your payment to stop or to cancel the subscription. Can you check it, please?
08:40:18 Customer: I would need to leave the chat
08:41:07 Agent: It will not end the chat, you can come back to our chat here again.
08:41:13 Agent: Take the whole time you need.

PLAY WITH CODES

Context Knowledge

Suppose a data stream provides:

C65 | 1.e4, e5 | 2.Nf3, Nc6 | 3.Bb5, Nf6 | +0.2

? What is the first code?

? Why numbers and letters? Are they coordinates?

? What does +0.2 mean?

? What does | mean?

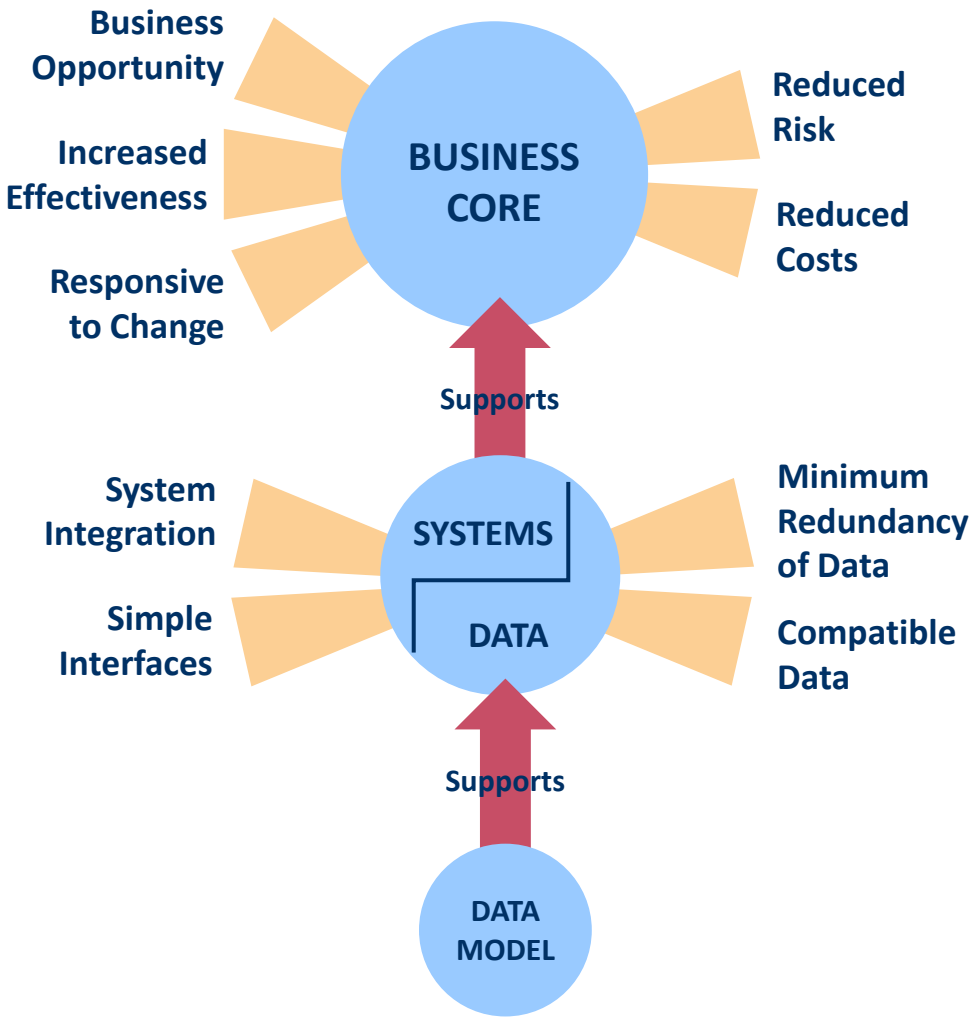
If you play chess, data becomes clearer…
we are looking to a chess opening

KNOWLEDGE IS POWER

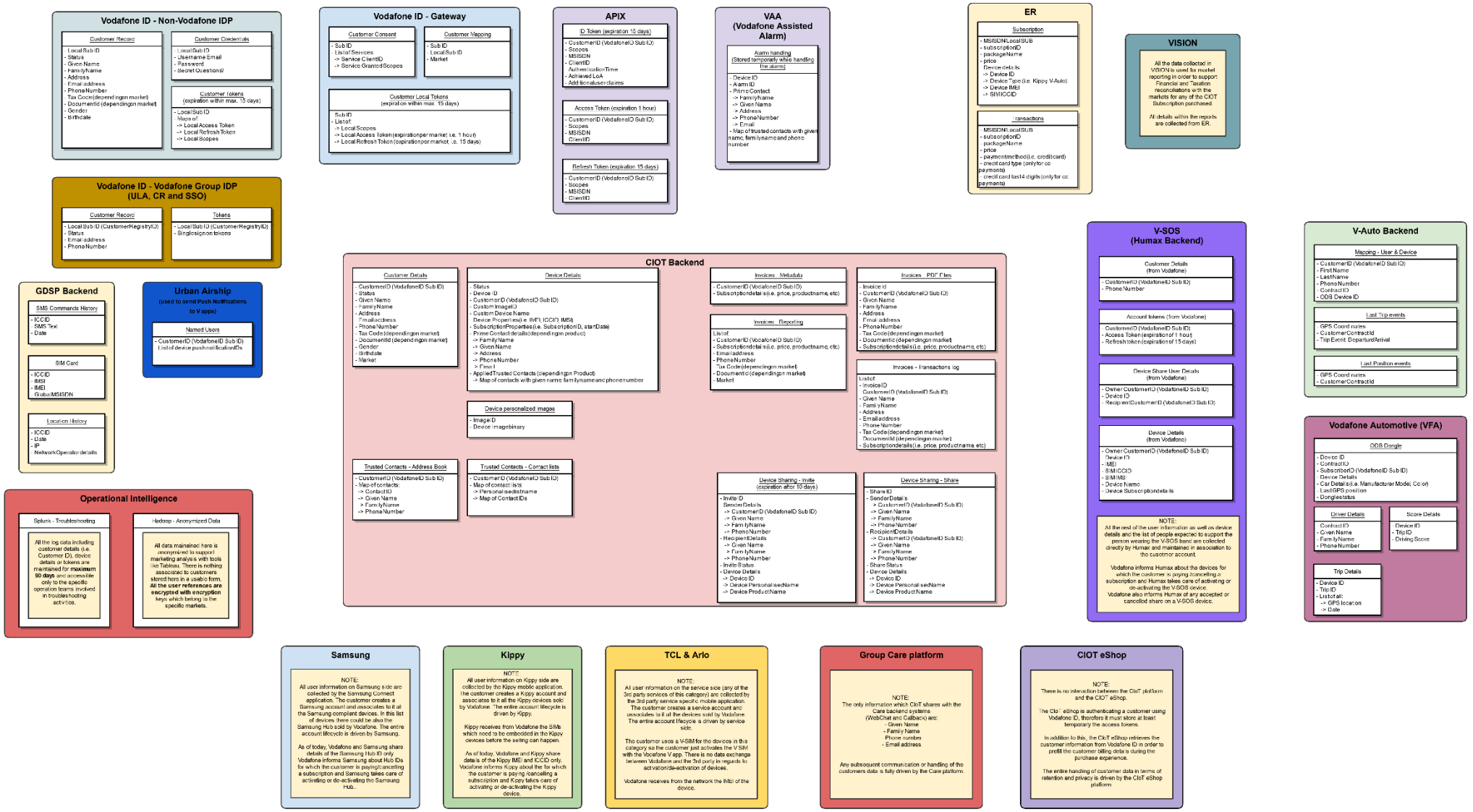
To leverage and empower data,
we need to build a suitable

DATA MODEL

i.e. any *abstract model* that defines a set of fundamental *entities and relations* among them, according to the real-world semantics of the data



DATA MODEL: CIoT



FORMER DATA TYPES

Data **types** define technically the information collected,
in terms of what operations are allowed with it

Numbers

allow mathematical operations
(+ - / *)

binary, integer or floating

Text

sentences or words

usually carrying their standard meaning

Timestamp

1560902400
06/19/19

used to record times and dates

e.g. the following are equivalent dates in different formats:
1560902400 06/19/19 19-06-19

CURRENT DATA FEATURES

Currently we use more flexible concepts for data features

Numerical

any data that can be represented by numbers

NO → 0
YES → 1
MAYBE → 2

Categorical

data assuming discrete values, corresponding to a given categorization of a concept in classes

Structured Data

Organized Data is Structured Data

FLAT TABLE

Numerical Feature

Categorical Feature

Structured Data

NAME	AGE	MARRIED
Allegri	51	0
Bolt	32	0
Christenson	26	1
Djokovic	32	1
Eriksen	27	0
Federer	37	1
Gullit	56	0
Helveg	48	1
Immobile	29	1

Items are listed in **ROWS**

Features are listed in **COLUMNS**

Data is contained in **CELLS**

The structure is based upon **INDEXING**, which is provided by the pair (row,column)

C2 General

3D TABLE

Structured Data

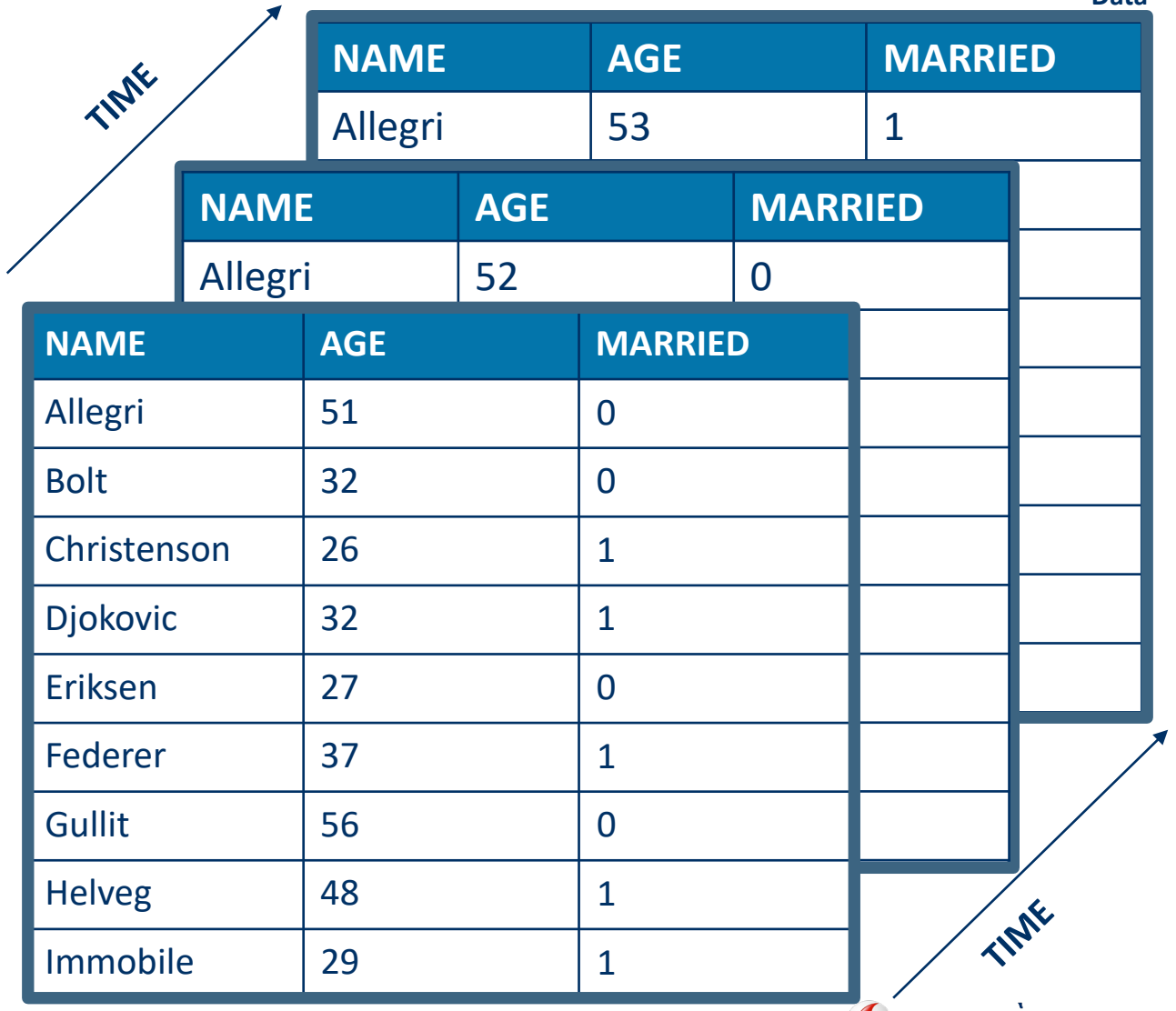
Multi-dimensional tables generalize the flat table with multi-indexing

Seen as a cube

For each fixed value of the third dimension, the slice is a flat table

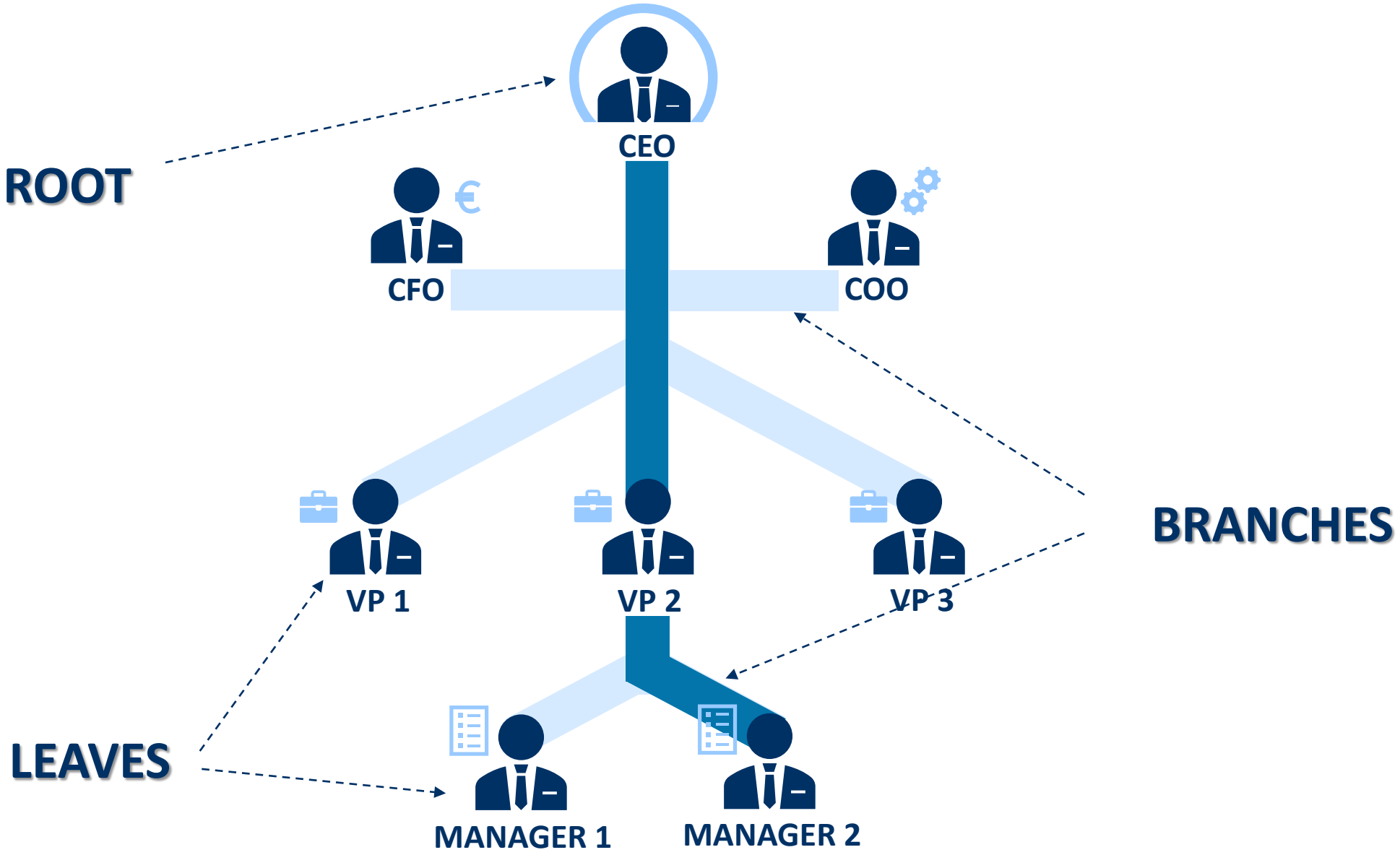
The index is made of three elements

The third dimension can be made by snapshots of a continuum, like time



TREES

Structured Data



C2 General

STRUCTURED DATA – CIoT EXAMPLE

Structured Data

Subscriptions/transactions as seen in Vision data warehouse

CI Type	MSISDN	Status	Start Date	End Date	Duration Str	Status Description (transaction)	Transaction Type Description	Package Id	Error ID
vodafoneid	[redacted]	INACTIVE	10/24/2018	11/13/2018	1 Month	COMPLETED	PURCHASE	PK_CIOT_VCAR_B	OK
vodafoneid	[redacted]	ACTIVE	4/11/2019	7/4/2019	14 Days	COMPLETED	PURCHASE	PK_CIOT_WIFI	
vodafoneid	[redacted]	ACTIVE	3/30/2019	6/30/2019	1 Month	COMPLETED	PURCHASE	PK_CIOT_SOSBAND	OK
vodafoneid	[redacted]	PAYMENT_FAILED	7/17/2018	8/17/2018	1 Month	DENIED	PURCHASE	PK_CIOT_Kippy_B	CARD_ERROR
vodafoneid	[redacted]	ACTIVE	12/11/2018	7/11/2019	1 Month	COMPLETED	RECURRING	PK_CIOT_VSIM_L	OK
vodafoneid	[redacted]	ACTIVE	3/30/2019	6/30/2019	1 Month	COMPLETED	RECURRING	PK_CIOT_VHOME	OK
msisdn	[redacted]	ACTIVE	11/20/2017	7/20/2019	1 Month	COMPLETED	RECURRING	PK_CIOT_VSIM_L	OK
msisdn	[redacted]	INACTIVE	1/28/2019	1/29/2019	1 Month	COMPLETED	PURCHASE	PK_CIOT_SOSBAND	OK
msisdn	[redacted]	ACTIVE	5/16/2019	7/16/2019	1 Month	COMPLETED	RECURRING	PK_CIOT_GSM_SERVICE	OK
msisdn	[redacted]	INACTIVE	1/16/2019	2/16/2019	1 Month	REJECTED	RECURRING	PK_CIOT_KIDS	INSUFFICIENT_FUNDS

SEMI-STRUCTURED DATA – CIoT EXAMPLE

Structured Data

Price plan activation transaction from the core MongoDB

{ "_id" : ObjectId("[redacted]"), "_class" : "com.vodafone.global.smartlife.transaction.repository.model.Transaction", "deviceUid" : "[redacted]", "deviceInfo" : { "deviceUid" : "[redacted]", "deviceStatus" : "ACTIVE", "hardwareUid" : "vodafone_trackimomini", "productId" : "5bed7f0e1e9d9800014cc3e9", "provisioningInfo" : { "iccid" : "[redacted]", "imei" : "[redacted]" }, "subscriptionInfo" : { "endDate" : "2019-07-28", "vendorPairingSuccess" : "true", "companyName" : null, "channel" : "999", "netPrice" : "0.0", "simActive" : "true", "grossPrice" : "0.0", "durationLabel" : "Subscrição mensal", "longPricePointId" : "package:PK_CIOT_TRACKISAFE_MINI_TAX_3_4_999_999_999_TRIAL_*_*_false_false_*_*_*", "subscriptionEndDateTime" : "2019-07-28T23:59:59.059+02:00", "subscriptionUpdateAction" : null, "context" : "PURCHASE", "customerIdentifier" : "[redacted]", "durationInMonths" : "1", "transactionDateTime" : "2019-06-28T00:51:03.003+02:00", "contactName" : null, "externalSubscriptionIdentifier3" : "", "pricePlanId" : "TSAFE_BASIC_PLAN", "transactionId" : „[redacted]", "externalReference" : "{\"imei\":\"[redacted]\",\"iccid\":\"[redacted]\",\"deviceUid\":\"[redacted]\",\"productName\":\"TrackiSafe Mini\",\"resourceType\":\"device\"}", "isContract" : "false", "subscriptionStartDateTime" : "2019-06-28T00:51:02.002+02:00", "taxRate" : "0.23", "pricePoint" : "TSAFE_BASIC_PLAN", "success" : "true", "shortPackageId" : "PK_CIOT_TRACKISAFE_MINI", "subscriptionId" : "147102380", "applicationId" : "OpCoApiPT", "taxAmount" : "0.0", "startDate" : "2019-06-28", "recurrency" : "por mês", "resultCode" : null, "opcoId" : "PT", "referenceId" : "[redacted]", "oldSubscriptionId" : "", "purchaseType" : "3", "duration" : "4", "customerCountryCode" : "PT", "requestId" : "", "subscriptionStatus" : "2", "promoCode" : "TRIAL", "currency" : "€", "packageName" : "PK_CIOT_TRACKISAFE_MINI", "serviceId" : "orderService", "email" : null, "timestamp" : "2019-06-27T22:51:04Z", "purchaseTimestamp" : "2019-06-27T22:51:04.491Z", "opcoCustomerId" : "", "customerIdentifierType“……

C2 General

moviri

SEMI-STRUCTURED DATA – CIoT EXAMPLE

Structured Data

Price plan activation transaction from the core MongoDB


```
{
        "_id" : ObjectId("6e96cfd374fb2f12dd9ea650fac799e1"),
        "deviceUid" : "[redacted]",
        "deviceStatusAtStart" : "UNREGISTERED",
        "deviceStatusAtEnd" : "ACTIVE",
        "action" : "REGISTER_PP",
        "status" : "SUCCESS",
        "creationTimestamp" : ISODate("2019-06-27T22:50:45.322Z"),
        "completionTimestamp" : ISODate("2019-06-27T22:51:10.683Z"),
        "deviceInfo" : {
                "deviceStatus" : "ACTIVE",
                "hardwareUid" : "vodafone_trackimomini",
                "productInfo" : {
                        "name" : "TrackiSafe Mini",
                        "pricePlanInfoSet" : [ {
                                "pricePlanId" : "TSAFE_BASIC_PLAN",
                                "commercialName" : "TrackiSafe Mini",
```


C2 General

DATA IS BORN RAW…

...AND NEEDS TO BE REFINED

Data Quality

To become

USABLE

VALUABLE

TRUSTWORTHY

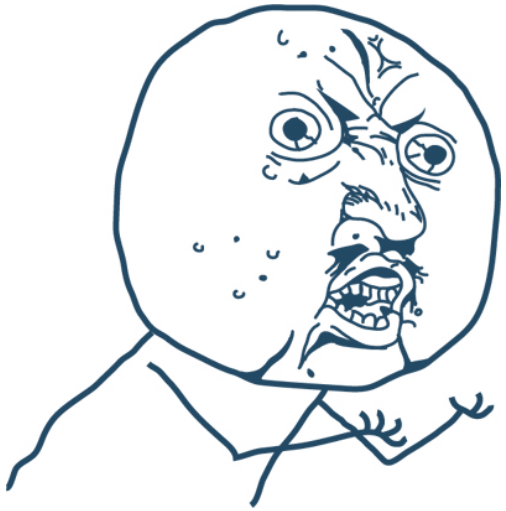
C2 General

Credits #: 4

DATA FORMAT

Data Quality

Try to write down the current date…



12/02/2020
02/12/2020
12/02/20
02/12/20
12-02-2020
02-12-2020
12-02-20
02-12-20
12-february-2020
12-feb-2020
…

Far too many ways!

DATA RELIABILITY

Data Quality

Or fill in the registration form…

Subscription

NAME:	Donald
SURNAME:	Duck
BORN IN:	Duckburg
WHEN:	06/09/1920
GENDER:	Male
RELATIVES:	Scrooge McDuck
TRAITS:	very unlucky!

…what is going on?

C2 General

DATA CYCLE

Data Quality

1560902400

INCOMING RAW DATA

DATA CLEANSING

REFINED DATASET

REQUISITES

ACCURACY
INTEGRITY
COMPLETENESS
VALIDITY
CONSISTENCY
UNIFORMITY
DENSITY
UNICITY

TO EMPOWER

REFINED DATA

ENRICHMENT

REPOSITORIES

USERS

DATA ANALYSIS

DATA VISUALIZATION

C2 General

AGENDA

Data processing

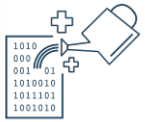
Data manipulation

Data analysis

Classwork

Missing values

Simple manipulations

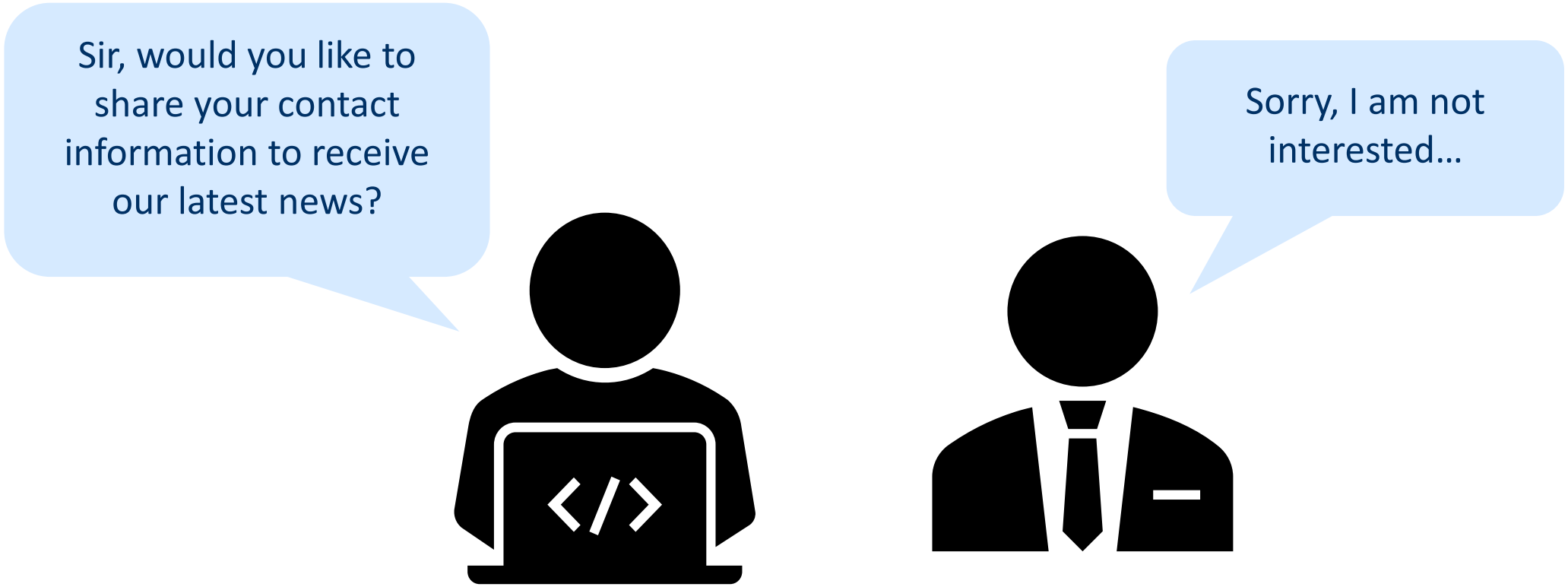


Data transformation

C2 General

MISSING VALUES

A dataset is not always complete…



FILLING THE GAPS

Missing Values

Filling missing values requires judgement and one to one evaluation

CHILD NAME	AGE	CLASS
Andy	8	III
Brian	9	III
Charles	8	III
David	9	III
Eve	9	III
Frank	8	III
Gulliver	9	III

The **missing age** is completed knowing that they attend elementary school, third class

The missing blood pressure of Gulliver is **hard to infer**, since he suffers hypertension

NAME	PRESSURE	HYPERTENSION
Andy	130/80	NO
Brian	125/75	NO
Charles	120/75	NO
David	125/85	NO
Eve	135/80	NO
Frank	*125/80*	NO
Gulliver	???	YES

C2 General

MISSING VALUES – CIoT EXAMPLE

Missing Values

sale_identifier	sale_date	sale_country	sale_product
ZAWwYHo0qTTtObAgGTvk	2019-08-23 00:00	5	VF V-Multi: TrackiSafe Mini
	2019-08-23 00:00	5	Vodafone V-Sim
PLQ0ylv8/ADrosxqDEnd	2019-08-23 00:00	5	Vodafone V-Pet: Pod3
	2019-08-23 00:00	5	Vodafone V-Sim
hinoIL6tqK1Ly31gZmtF	2019-08-23 00:00	5	VF V-Bag: Luggage Tracker
PFF99CiKocVH0045PvnY	2019-08-23 00:00	5	VF V-Bag: Luggage Tracker
KsYOIEedIHgskI4vSkQS	2019-08-23 00:00	5	VF V-Bag: Luggage Tracker
FB5jsOP0s+gVysVlSlNP	2019-08-23 00:00	5	VF V-Bag: Luggage Tracker
DqmNzuU4FC638KUV9by9	2019-08-23 00:00	5	VF V-Multi: TrackiSafe Mini
	2019-08-23 00:00	5	Vodafone V-Sim
	2019-08-23 00:00	5	Vodafone V-Sim

TABLE MANIPULATION: FILTERING AND ORDERING

Simple Manipulations

NAME	AGE	MARRIED
Allegri	51	0
Bolt	32	0
Christenson	26	1
Djokovic	32	1
Eriksen	27	0
Federer	37	1
Gullit	56	0

Filter married and younger than 35

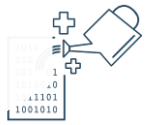
NAME	AGE	MARRIED
Christenson	26	1
Djokovic	32	1

Order by age

NAME	AGE	MARRIED
Christenson	26	1
Eriksen	27	0
Djokovic	32	1
Bolt	32	0
Federer	37	1
Allegri	51	0
Gullit	56	0

C2 General

ADD NEW OBSERVATIONS AND DETAILS



Simple Manipulations

	A	B	C	D	E	F
	NAME	AGE	MARRIED	BORN	CONSENT	REVENUES
1	Allegri	51	0	1967	N	1000
2	Bolt	32	0	1986	N	3000
3	Christenson	26	1	1993	Y	500
4	Djokovic	32	1	1987	N	5000
5	Eriksen	27	0	1992	Y	1000
6	Federer	37	1	1981	N	4500
7	Gullit	56	0	1962	Y	1000
8	Helveg	48	1	1971	Y	500
9	Immobile	29	1	1990	N	2000
10	Larsson	47	1	1971	N	500

Add new lines

Calculate fields

FORMULA

=YEAR(TODAY())-B1

Adding new columns

C2 General

AGGREGATION FUNCTIONS

Simple Manipulations

Aggregate multiple rows for a higher-level view of the data

SUM(REVENUES)

5500

COUNT(NAME)

10

AVERAGE(AGE)

35.5

NAME	AGE	MARRIED	BORN	CONSENT	REVENUES
Allegri	51	0	1967	N	1000
Bolt	32	0	1986	N	3000
Christenson	26	1	1993	Y	500
Djokovic	32	1	1987	N	5000
Eriksen	27	0	1992	Y	1000
Federer	37	1	1981	N	4500
Gullit	56	0	1962	Y	1000
Helveg	48	1	1971	Y	500
Immobile	29	1	1990	N	2000
Larsson	47	1	1971	N	500

GROUP BY A LABEL

Simple Manipulations

What is the sum of the *Revenues* split by marital status?

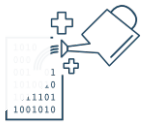
GROUP by label

SUM over column

MARRIED	SUM(REVENUES)
0	6000
1	13000

NAME	AGE	MARRIED	BORN	CONSENT	REVENUES
Allegri	51	0	1967	N	1000
Bolt	32	0	1986	N	3000
Christenson	26	1	1993	Y	500
Djokovic	32	1	1987	N	5000
Eriksen	27	0	1992	Y	1000
Federer	37	1	1981	N	4500
Gullit	56	0	1962	Y	1000
Helveg	48	1	1971	Y	500
Immobile	29	1	1990	N	2000
Larsson	47	1	1971	N	500

GROUPING ROWS – CIoT EXAMPLE



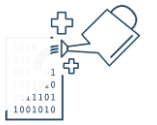
Simple Manipulations

Data usage aggregated and grouped by customer (MSISDN)

Customer ID (MSISDN)	SUM(Megabytes over the past 30 days)	SUM(Megabytes over the past 20 days)
026c8f27-6368-4788-b5c9-e1cb3664694f	129	101
031ff494-b818-4f44-b14c-69777e4c38e5	134	75
048965b3-e11d-4b49-8c5d-a20d9e7dc914	428	0
06d54fd2-fe91-4c54-ba77-837dce4746c7	215	0
06e0f853-6faf-4cb2-90ce-2bf089594fef	205	0
06e87a4f-0b23-4eed-9e77-42e0f3c7c97f	51	0
0716ebbd-e798-45e1-a5b3-57ce643024d2	201	38
077529b6-3eba-46c1-9712-b781b81ba514	49	24
07a87bf2-94bd-4a5a-8895-0cd7989eb291	34	0
0885d962-de02-4232-904a-d5bfded9b7c8	3	0

C2 General

JOINING TABLES



Simple Manipulations

NAME	SPORT	INCOME
Allegri	Soccer	1000
Bolt	Athletics	3000
Christenson	Volleyball	500
Djokovic	Tennis	5000
Eriksen	Soccer	1000
Federer	Tennis	4500
Gullit	Soccer	1000

How could we leverage information in multiple tables?

SPORT	AVG INCOME
Soccer	3000
Basketball	2000
Volleyball	400
Swimming	1000

JOINING TABLES

Simple Manipulations

NAME	SPORT	INCOME
Allegri	Soccer	1000
Bolt	Athletics	3000
Christenson	Volleyball	500
Djokovic	Tennis	5000
Eriksen	Soccer	1000
Federer	Tennis	4500
Gullit	Soccer	1000

Identify a common **INDEX**

LINK rows through operations on the index

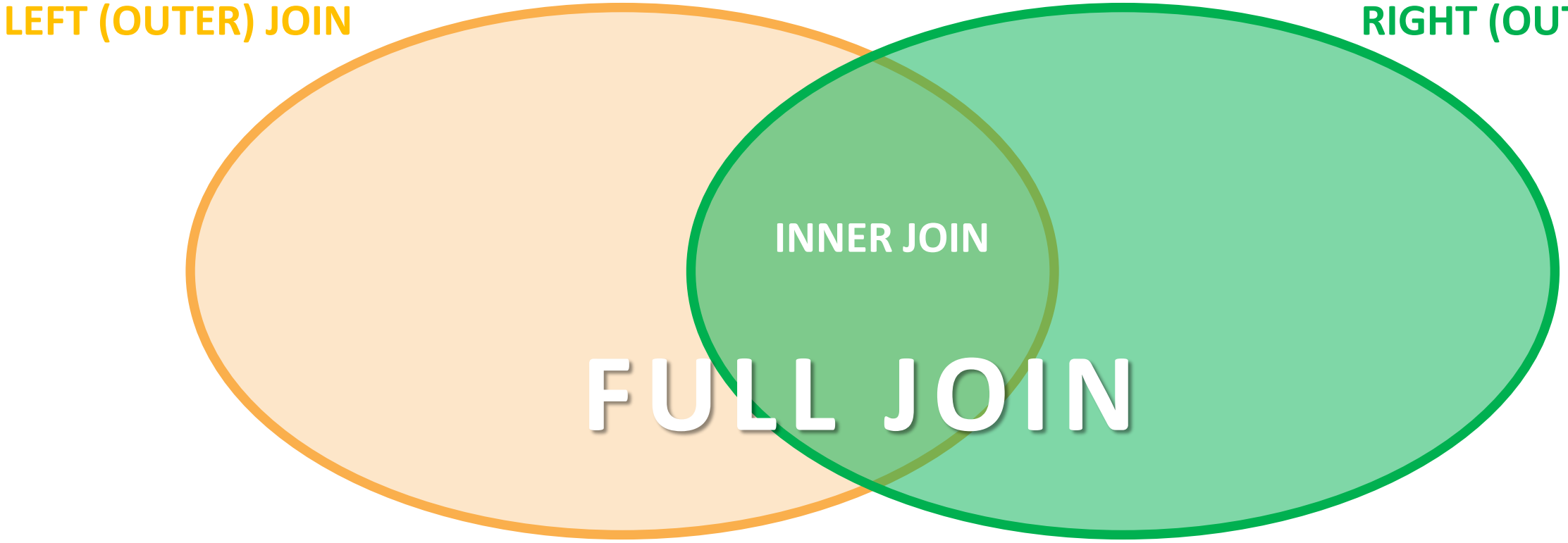
SPORT	AVG INCOME
Soccer	3000
Basketball	2000
Volleyball	400
Swimming	1000

C2 General

JOINING TABLES

Simple Manipulations

Joins are elementary operations between two set of indexes



C2 General

JOINING TABLES: Example

Simple Manipulations

NAME	SPORT	INCOME
Allegri	Soccer	1000
Bolt	Athletics	3000
Christenson	Volleyball	500
Djokovic	Tennis	5000
Eriksen	Soccer	1000
Federer	Tennis	4500
Gullit	Soccer	1000

SPORT	AVG INCOME
Soccer	3000
Basketball	2000
Volleyball	400
Swimming	1000

Suppose we want to compare an athlete income with avg income of its sport, we need to join the two tables on column "sport"

LEFT JOIN

NAME	SPORT	INCOME	AVG
Allegri	Soccer	1000	3000
Bolt	Athletics	3000	
Christenson	Volleyball	500	400
Djokovic	Tennis	5000	
Eriksen	Soccer	1000	3000
Federer	Tennis	4500	
Gullit	Soccer	1000	3000

INNER JOIN

NAME	SPORT	INCOME	AVG
Allegri	Soccer	1000	3000
Christenson	Volleyball	500	400
Eriksen	Soccer	1000	3000
Gullit	Soccer	1000	3000

OUTER JOIN

NAME	SPORT	INCOME	AVG
Allegri	Soccer	1000	3000
Bolt	Athletics	3000	
Christenson	Volleyball	500	400
Djokovic	Tennis	5000	
Eriksen	Soccer	1000	3000
Federer	Tennis	4500	
Gullit	Soccer	1000	3000
	Basketball		2000
	Swimming		1000

RIGHT JOIN

NAME	SPORT	INCOME	AVG
Allegri	Soccer	1000	3000
Christenson	Volleyball	500	400
Eriksen	Soccer	1000	3000
Gullit	Soccer	1000	3000
	Basketball		2000
	Swimming		1000

C2 General

JOINING TABLES: Example on our classwork

Simple Manipulations

Week	Active Customers
1/2/2017	450
1/9/2017	630
1/16/2017	610
1/23/2017	620
1/30/2017	560
2/6/2017	620
…	…

Join data on total customers and on accesses to the various section, through time information
(in this case)

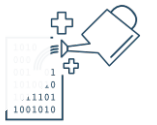
Week	Customers accessing Cardio Section
1/2/2017	117
1/9/2017	183
1/16/2017	146
1/23/2017	167
…	…

Week	Customers accessing Functional Section
1/2/2017	59
1/9/2017	101
1/16/2017	92
1/23/2017	93
…	…

…

C2 General

JOINING TABLES – CIoT EXAMPLE



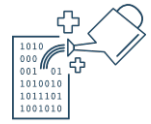
Simple Manipulations

Sell-in data joined with activations

activation_identifier	sale_identifier	sub_start_date	sale_date	sale_product	sale_channel_subcategory	sale_street
	bZRsUqXTku0uAxcM0LfD		2019-04-19 00:00	V-Home	ONLINE	Baiersdorfer Str. 20
	SGdlySmwvc9KvNp6VgyJ		2019-04-22 00:00	V-Home	ONLINE	Raiffeisenstr 6
	Mn1U+THOzPTmEbt3mGKz		2019-04-08 00:00	V-Home	ONLINE	Emil-Geis-Str. 38
	DB4KE1zRTu+INPO9+e0s		2019-04-08 00:00	V-Home	ONLINE	Auf Der Scholle 6
	MExDrtWBPfm6NKotAL/l		2019-04-08 00:00	V-Home	ONLINE	Im Dorfe 20d
aYaxampjpPibt9dckwwm	aYaxampjpPibt9dckwwm	2019-04-22 09:58	2019-04-08 00:00	V-Home	ONLINE	Brönngasse 10
	Guy6mlKIDFKIwDwmZyvx		2019-04-03 00:00	V-Home	ONLINE	Friedensstr. 7
	cU6a5cqVfXybrYpwizqG		2019-04-08 00:00	V-Home	ONLINE	Waldweg 121
bGRO31RTRLIB72lH9/ej	bGRO31RTRLIB72lH9/ej	2019-08-20 17:00	2019-04-08 00:00	V-Home	ONLINE	Stettiner Str. Sued 8
dq+DXKv4h3mqxypYxymj	dq+DXKv4h3mqxypYxymj	2019-04-11 09:50	2019-04-08 00:00	V-Home	ONLINE	Kastanienallee 48

C2 General

DATA REDUNDANCY



Data Transformation

NAME	AGE	BORN	AGE<=35	AGE>35
Albert	29	01/01/1990	YES	NO
Bud	34	01/01/1985	YES	NO
Chloe	24	01/01/1995	YES	NO
Daisy	59	01/01/1960	NO	YES
Eugen	54	01/01/1965	NO	YES
Farah	49	01/01/1970	NO	YES
Gidget	44	01/01/1975	NO	YES

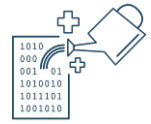
Age can be computed from the date of birth

The date of birth contains more information

Age<=35 and Age>35 are complementary, one of them is completely useless

C2 General

DATA TRANSFORMATION



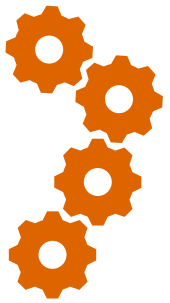
Data Transformation

Primary features allow to derive extra useful attributes

DATE OF BIRTH
01/01/1990
04/26/1989
12/03/1960

DAY OF BIRTH	MONTH OF BIRTH	YEAR OF BIRTH	AGE
1	JANUARY	1990	29
26	APRIL	1989	30
3	DECEMBER	1960	59

- DROP all /
- EXTRACT first couple of characters
- EXTRACT second couple of characters
- CONVERT numeric month to usual name
- EXTRACT last 4 characters
- COMPUTE age by substraction



C2 General

DATA ENRICHMENT - CIoT

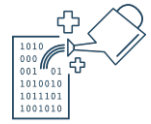
Data Transformation

Enrichment with data from the Group TAC (Type Allocation Code) list

hardware_uid	product_text	tac	Manufacturer	Marketingname	GSMAManufacturer	GSMAName	DeviceSlot_Type	count
vodafone_sim_only	V-Sim	35716404	Telit	HE910	Telit Communications SpA	Telit HE910, HE910-G, HE910-DG	M2M	1
vodafone_sim_only	V-Sim	35749206	TK Star	TK906	Topwell Technology (HK) Company Limited	Topwell M318A	GPS Tracker	1
vodafone_sim_only	V-Sim	35773808	Alcatel	Move Time	TCL Communication Ltd	MOVETIME Family Watch	Watch	293
vodafone_sim_only	V-Sim	35801409	u-blox	SARA-G350	u-blox AG	SARA-G350	M2M	1
vodafone_sim_only	V-Sim	35841707	Netgear	Arlo Go 4G	Netgear Inc	Arlo Security Camera System - Gen4	Camera	251
vodafone_sim_only	V-Sim	35888709	Voltaware	Voltaware Sensor	u-blox AG	SARA-U201	GPS Tracker	1

C2 General

DATA ENRICHMENT - CIoT



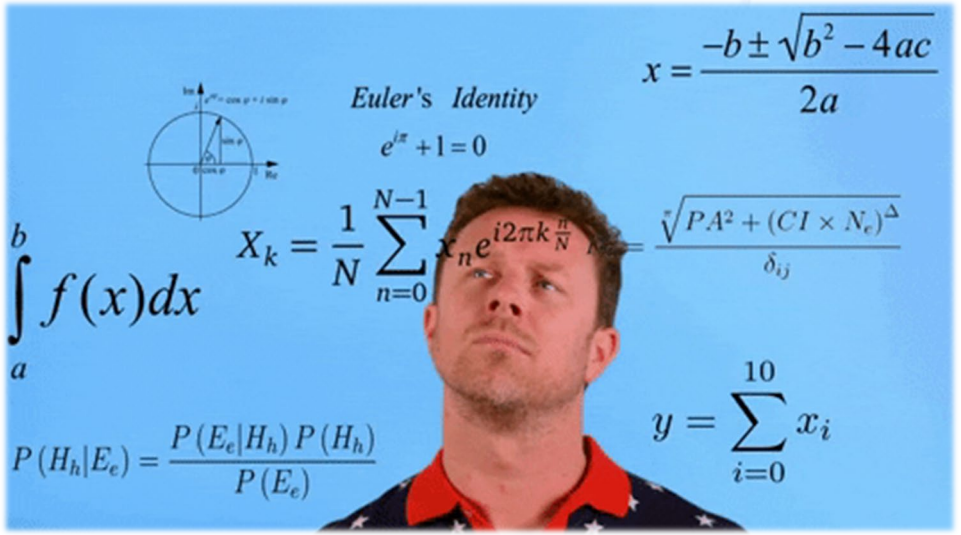
Data Transformation

Some other lookups we match the data against are:

Lookup table	Purpose
Tester list	Exclude test devices/accounts from the numbers
Launch dates list	Exclude data before the launch dates from production numbers
Special offers and campaigns	Limit the numbers only to devices that are part of a campaign

C2 General

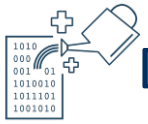
15 minutes Break



https://giphy.com/explore/questions

AGENDA

Data processing



Data manipulation

Data analysis

Classwork

Analysis

Data feeds Statistics

Simple statistical functions

Probability distribution

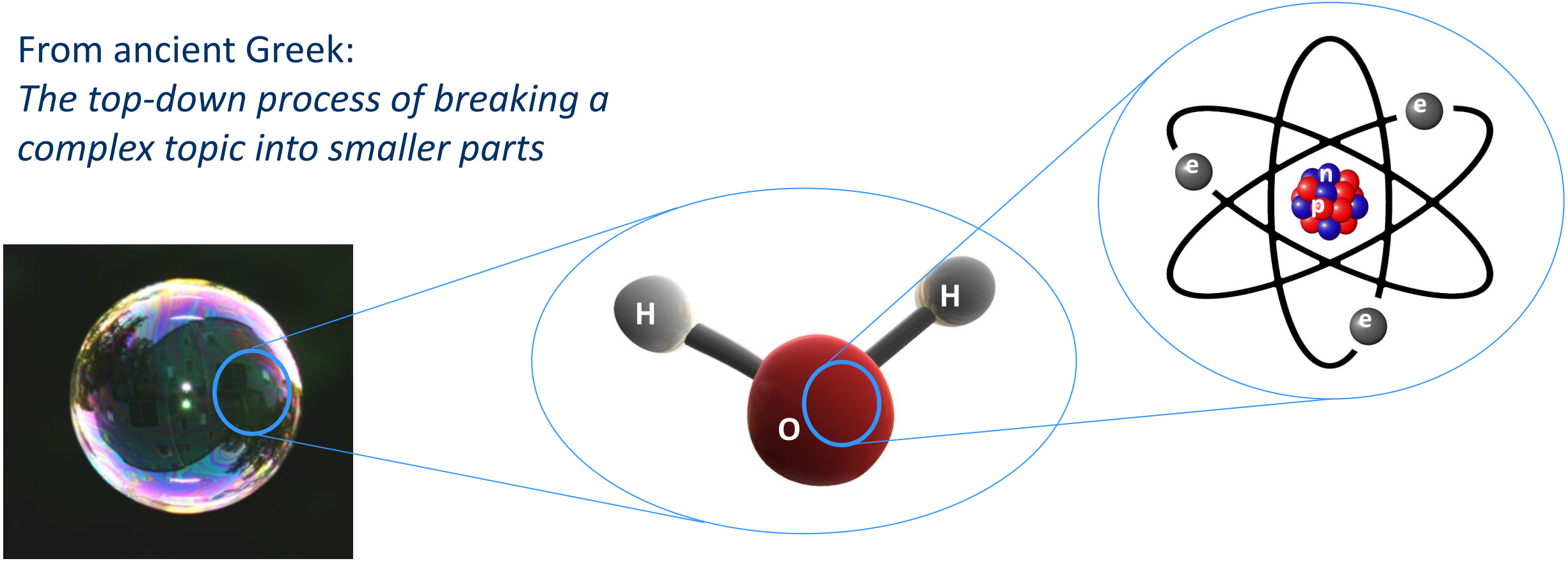
Correlation

Inference

C2 General

WHAT *ANALYSIS* MEANS

From ancient Greek:
The top-down process of breaking a complex topic into smaller parts

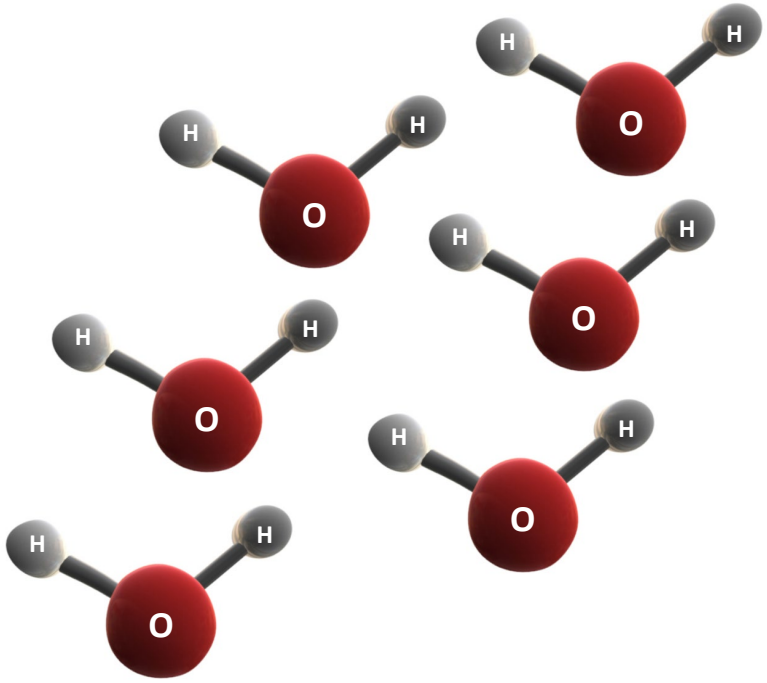


Credits #: 5,6,7

STATISTICAL VIEW OF DATA

Data feeds
Stats

While keeping a statistical perspective, detailed information are lost....



...to discover macroscopic collective behaviors

C2 General

Credits #: 6,8

moviri

HOW STATISTICS EXTRACTS VALUE FROM DATA

Data feeds
Stats

DESCRIBE

Compute simple indicators

Find aggregated views

Point out common traits

INFER

Find a theoretical model fitting data

Check assumption correctness

Forecast future trends/outcomes

C2 General

MODE

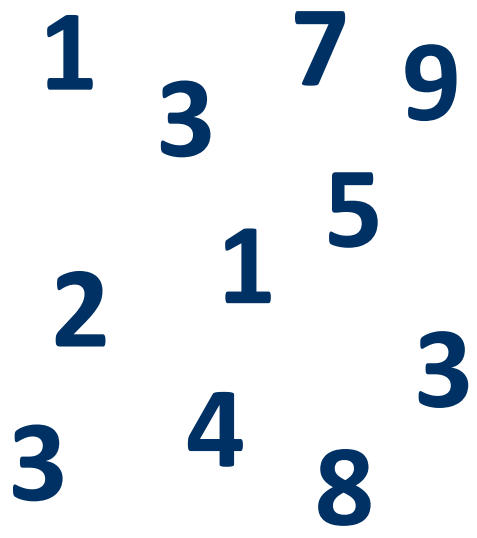
Count how many times each observation is repeated in the set

1 7 9 3 5 1 2 3 4 8 3

Observation	1	2	3	4	5	7	8	9
Frequency	2	1	3	1	1	1	1	1

The mode is the most recurrent observation

MEDIAN



Order the observations from the smallest to the largest

1 1 2 3 3 3 4 5 7 8 9

Find the position of the observation that splits the set in two groups, each containing half of the considered values

$$\frac{n+1}{2} = \frac{11+1}{2} = 6$$

The number in the 6^{th} position is 3,
the median for this example

PERCENTILE

The median can be interpreted as the 50th percentile,
i.e. the value for which 50% of the observations falls on the first part of the ordered sequence

1 1 2 3 3 **3** 4 5 7 8 9

50%

What about the 80th percentile?

1 1 2 3 3 3 4 5 **7** 8 9

80%

AVERAGE μ

ADD ALL THE ELEMENTS…

1 5 3

2 4

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1+2+4+5}{4} = 3$$

… AND DIVIDE BY THEIR COUNT

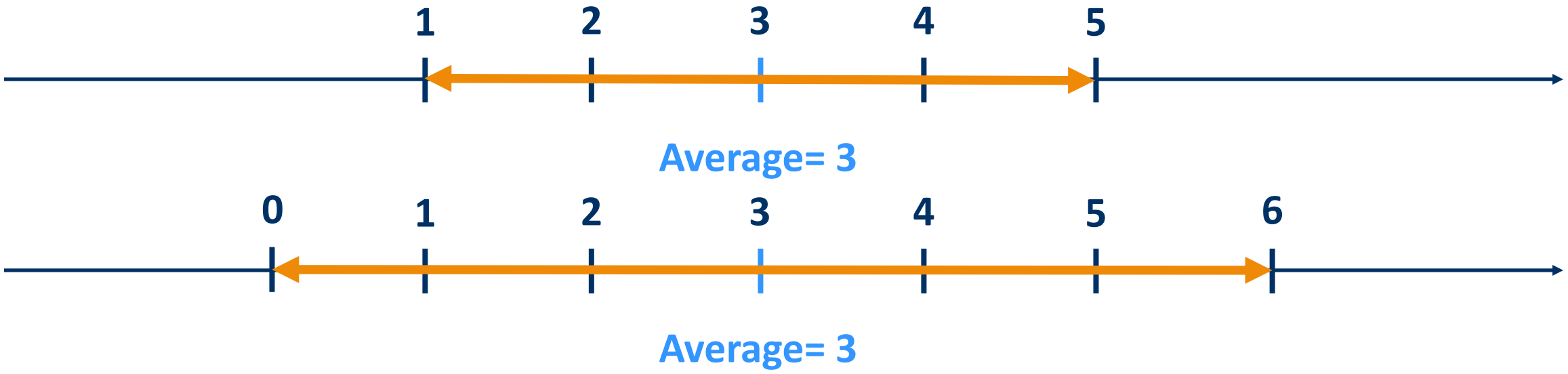
MISLEADING AVERAGE is not observed in the sample

BEST APPROXIMATION for repeated measures

C2 General

FROM AVERAGE TO VARIANCE

Unluckily the average does not involve sample size...

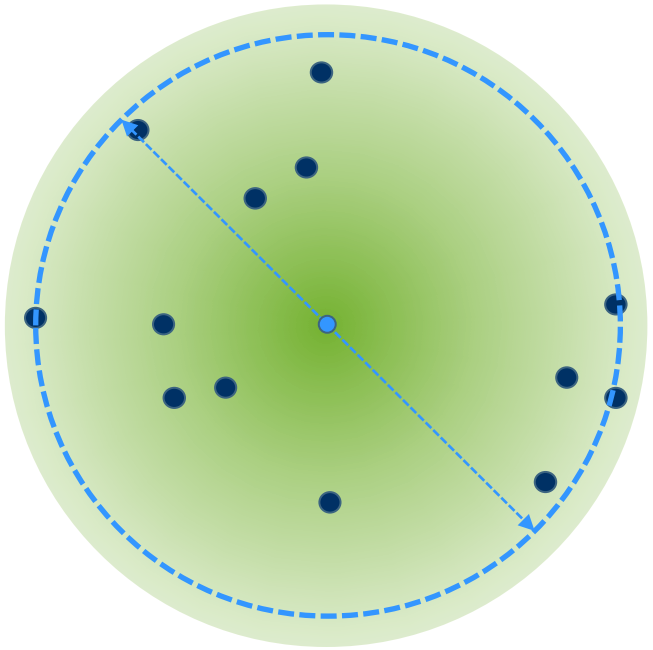


Minimum and Maximum values have a big impact on sample shape

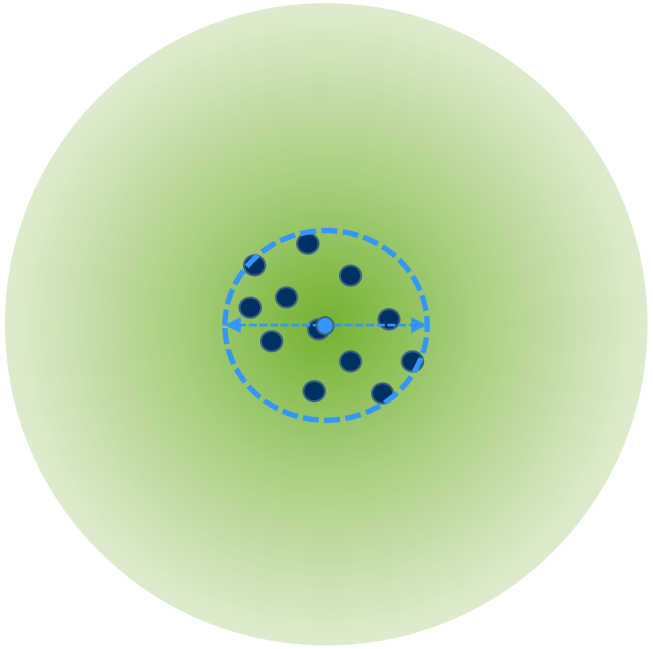
We need to use a measure of *dispersion*

VARIANCE σ^2

BIG VARIANCE



SMALL VARIANCE



Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

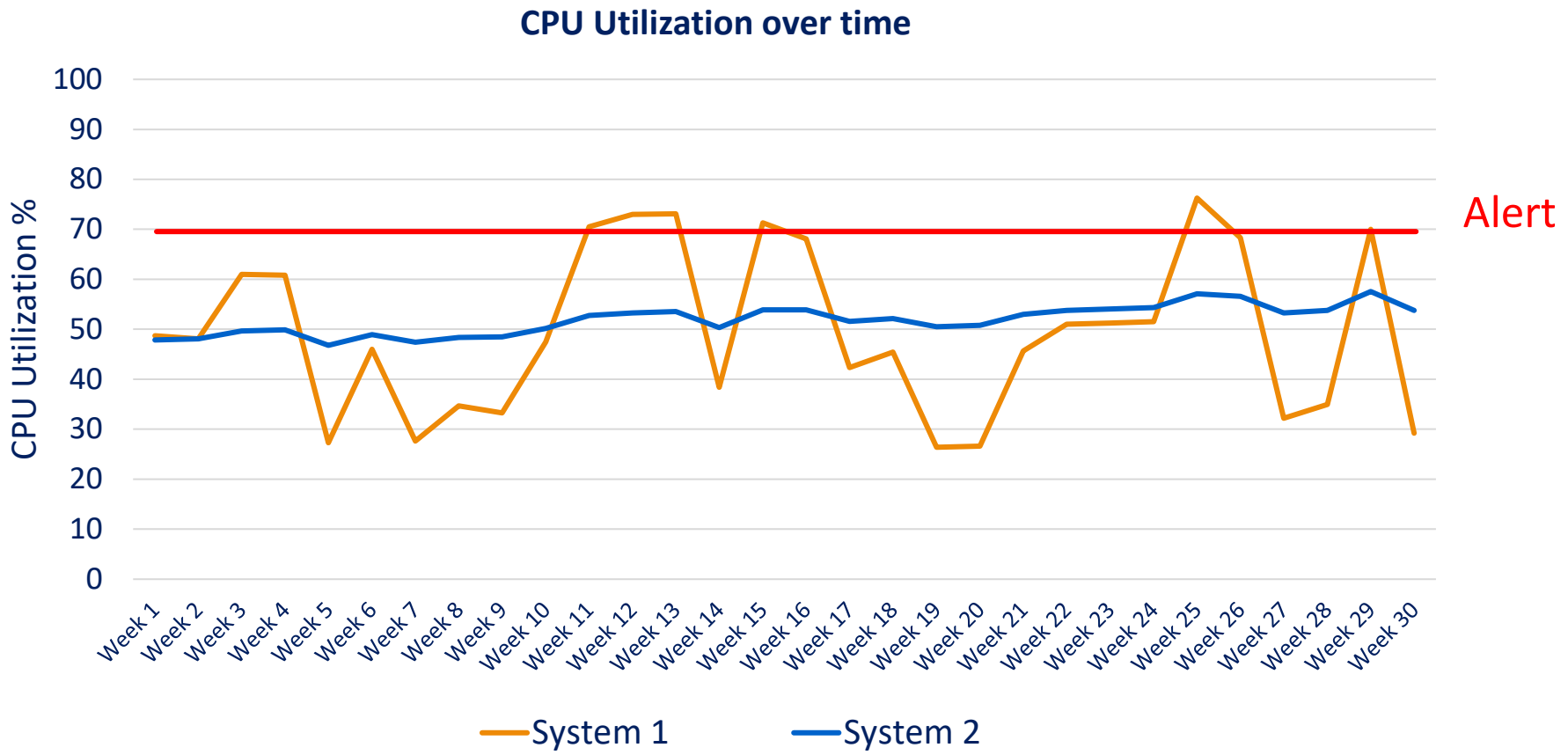
Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

IMPORTANCE OF VARIANCE

Simple Stats Functions

Variance helps defining better time series behavior



GIVE YOUR BEST SHOT!

5 minutes

FREQUENCIES

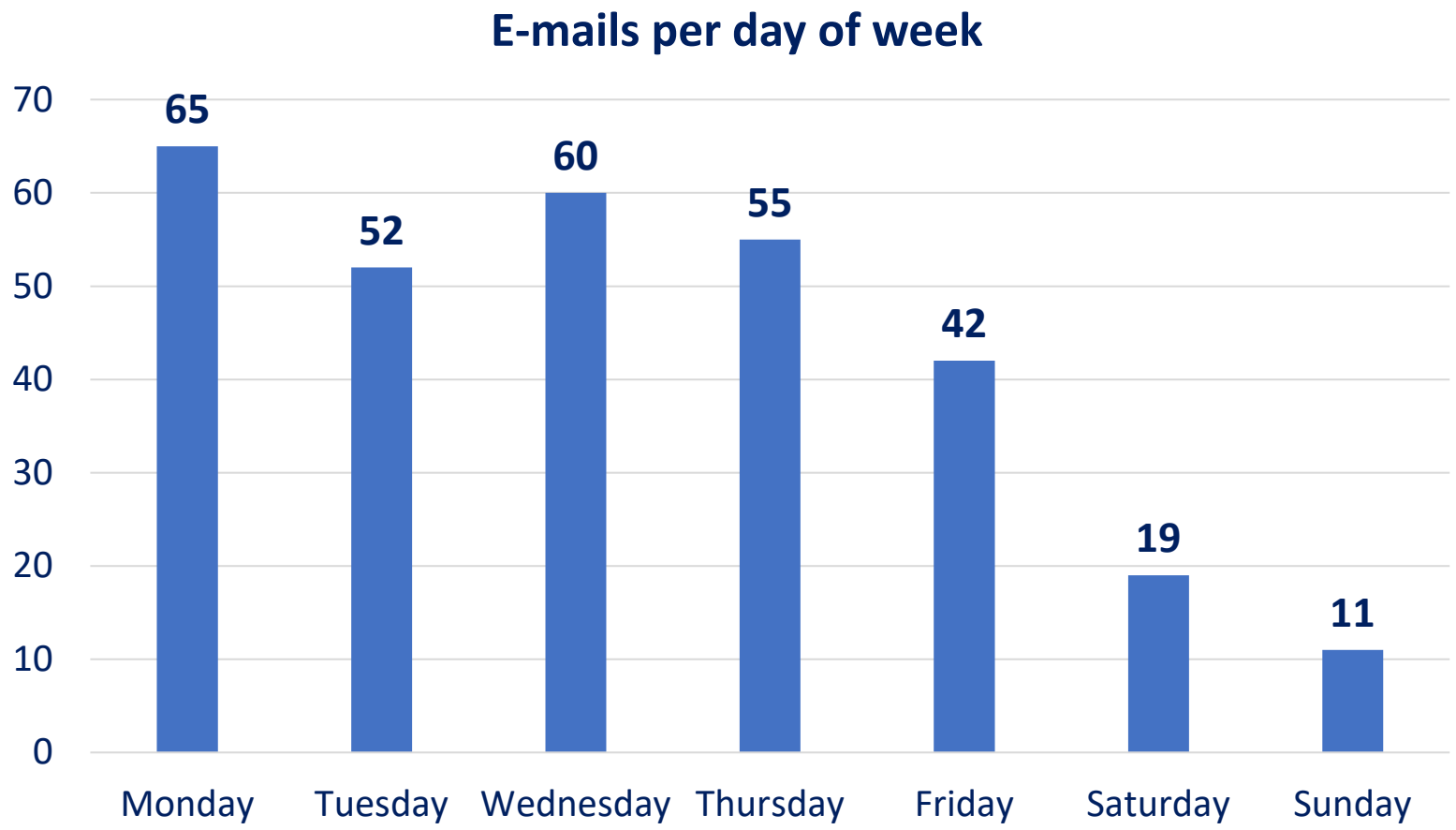
How many e-mails do you receive per day?

TOO MANY!!

FREQUENCY DISTRIBUTION

Probability Distribution

The number of e-mails received per day is roughly a statistical distribution

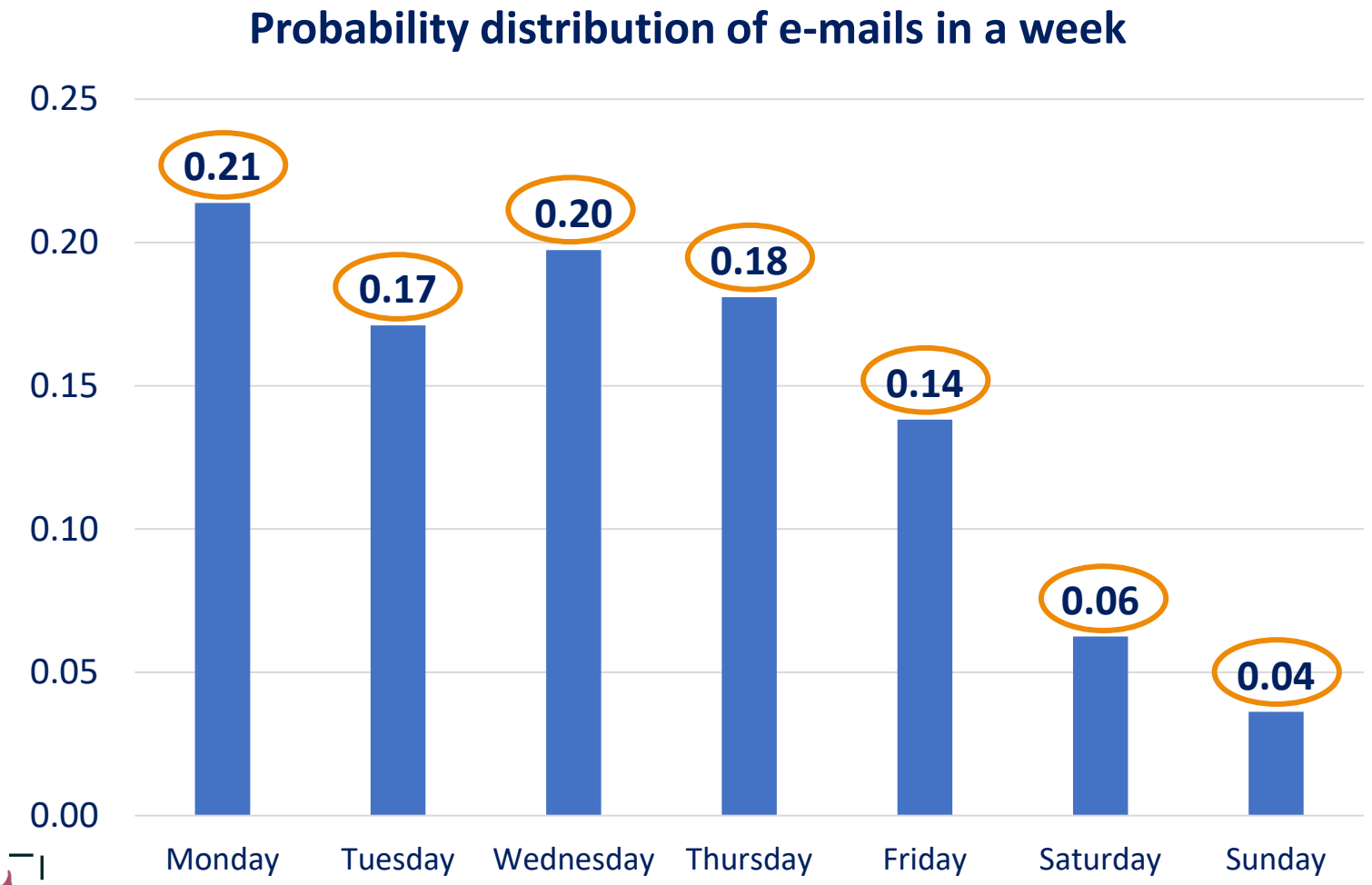


C2 General

PROBABILITY DISTRIBUTION

Probability Distribution

Normalize the columns to switch from frequency to probability distribution



The resulting ratios can be interpreted as discrete probabilities

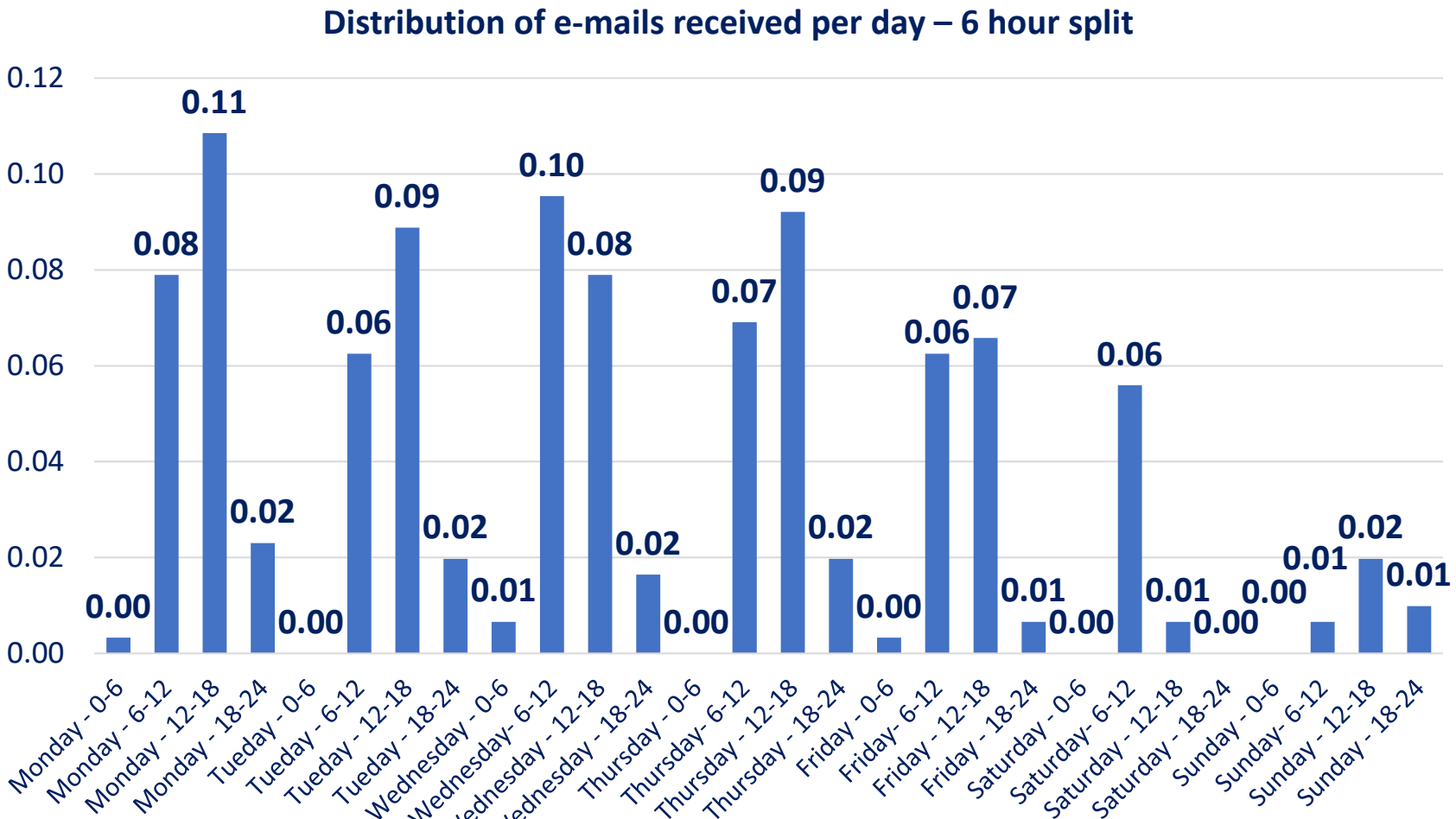
They are easily understood as percentages

They sum up to 1

C2 General

BREAKING DOWN PROBABILITY DISTRIBUTIONS

Probability Distribution



Probability gets smaller and smaller, losing its meaning

Baseline breaks in too many small parts

C2 General

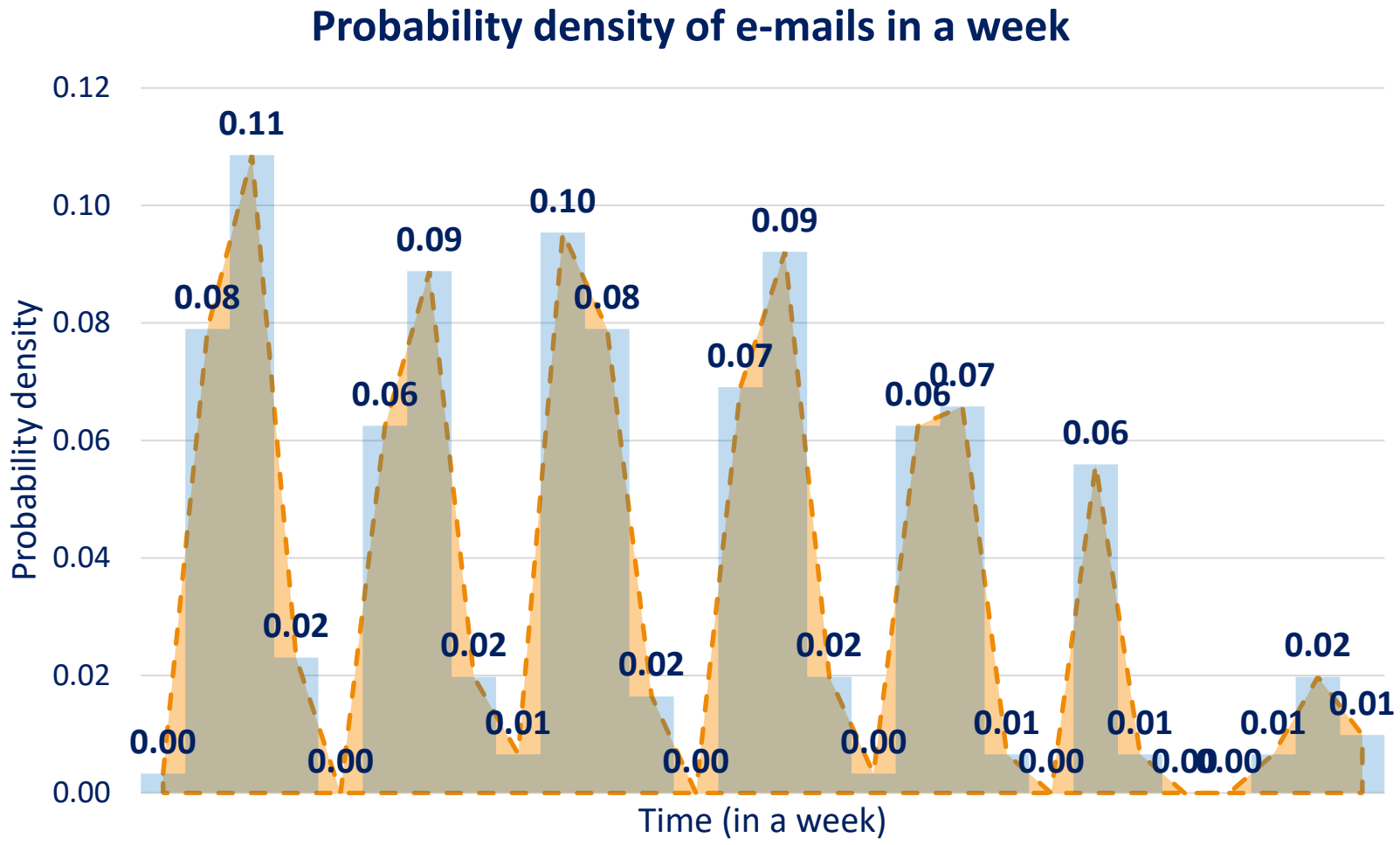
PROBABILITY DENSITY

Probability Distribution

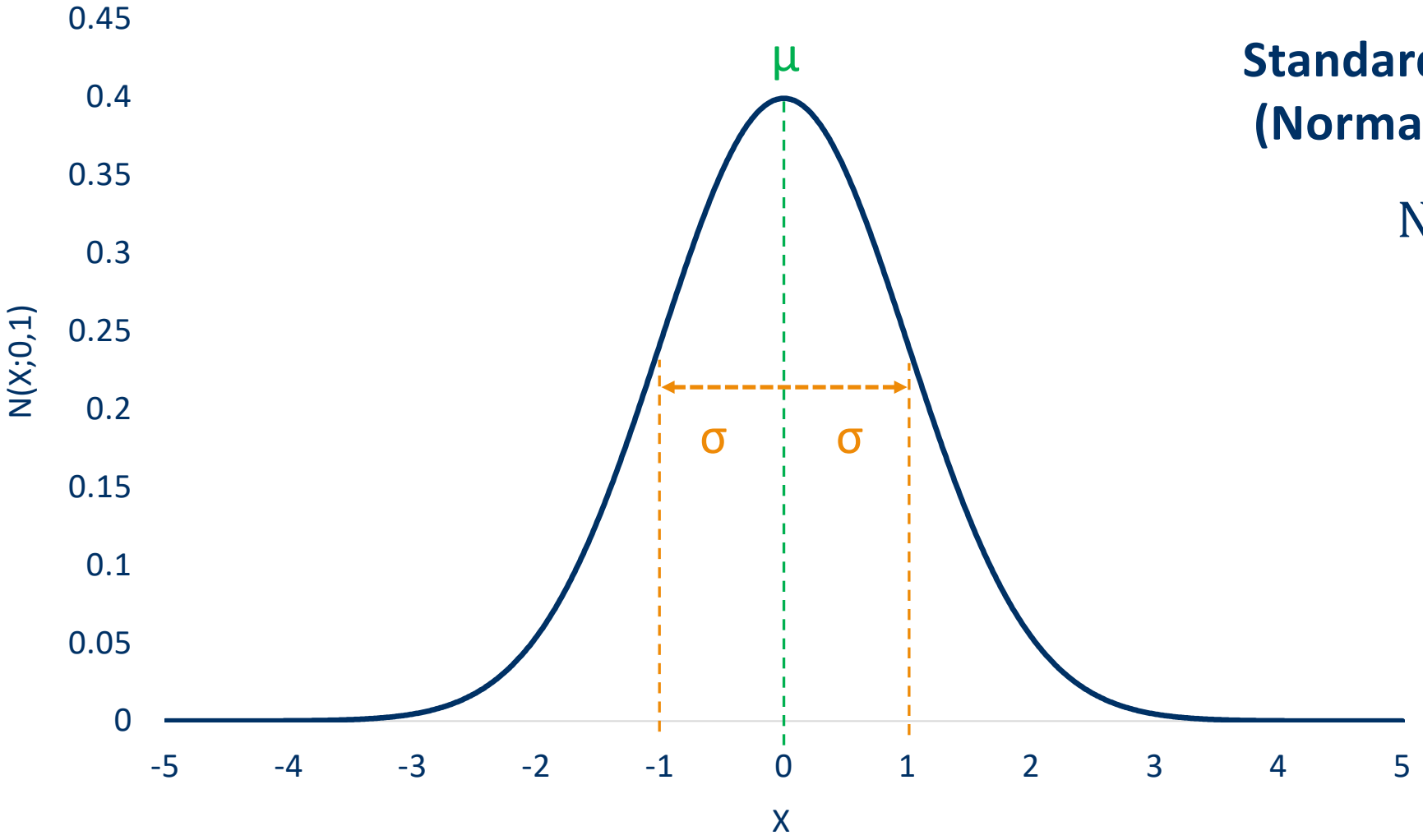
When the baseline becomes continuous, the histogram fades into a continuous curve

The curve is a **probability density**

The **area under the curve** defines the **probability**



NOTABLE EXAMPLE: GAUSSIAN DISTRIBUTION



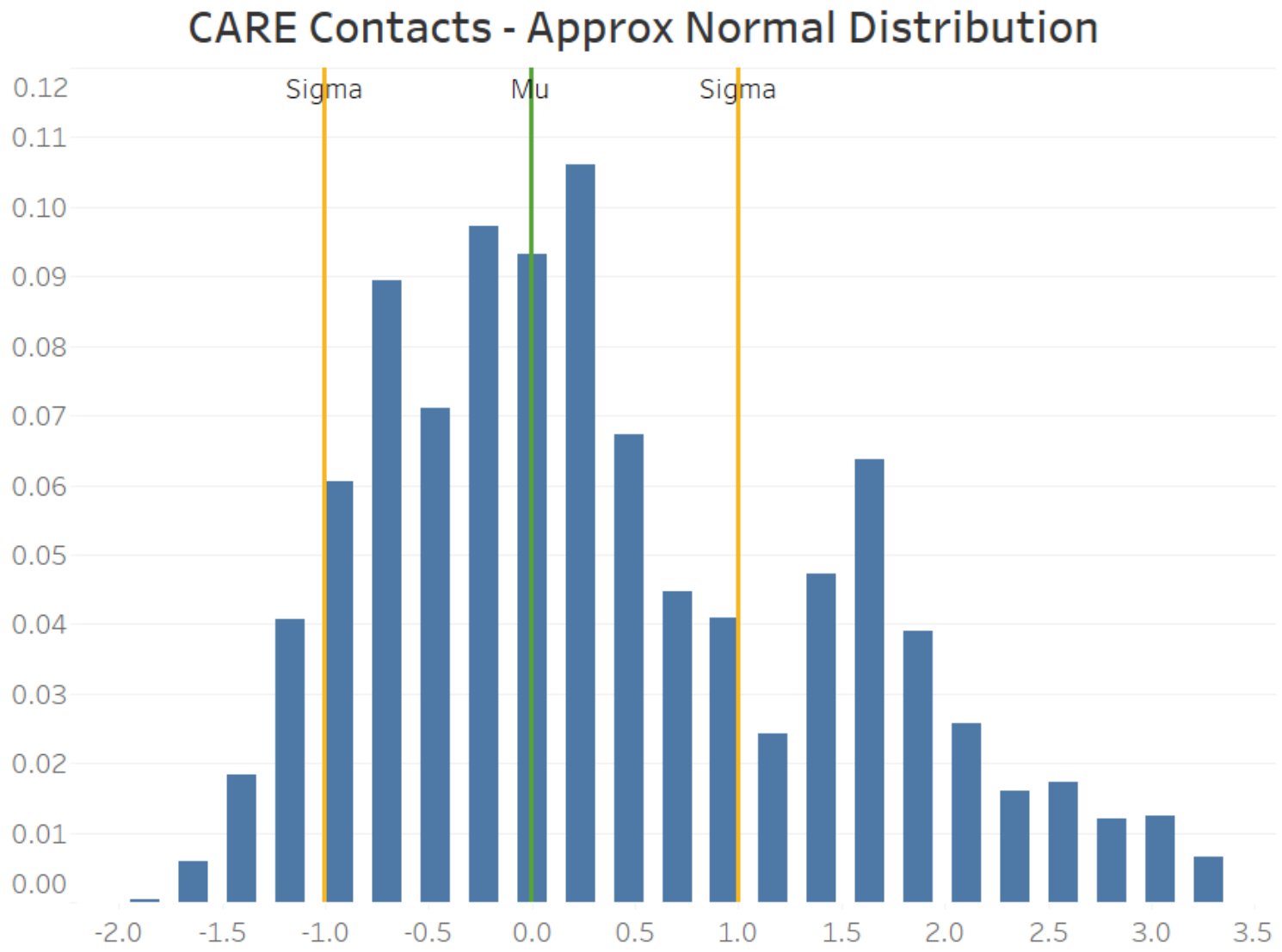
Standardized gaussian (Normal distribution)

$N(\mu, \sigma)$

$\mu = 0$

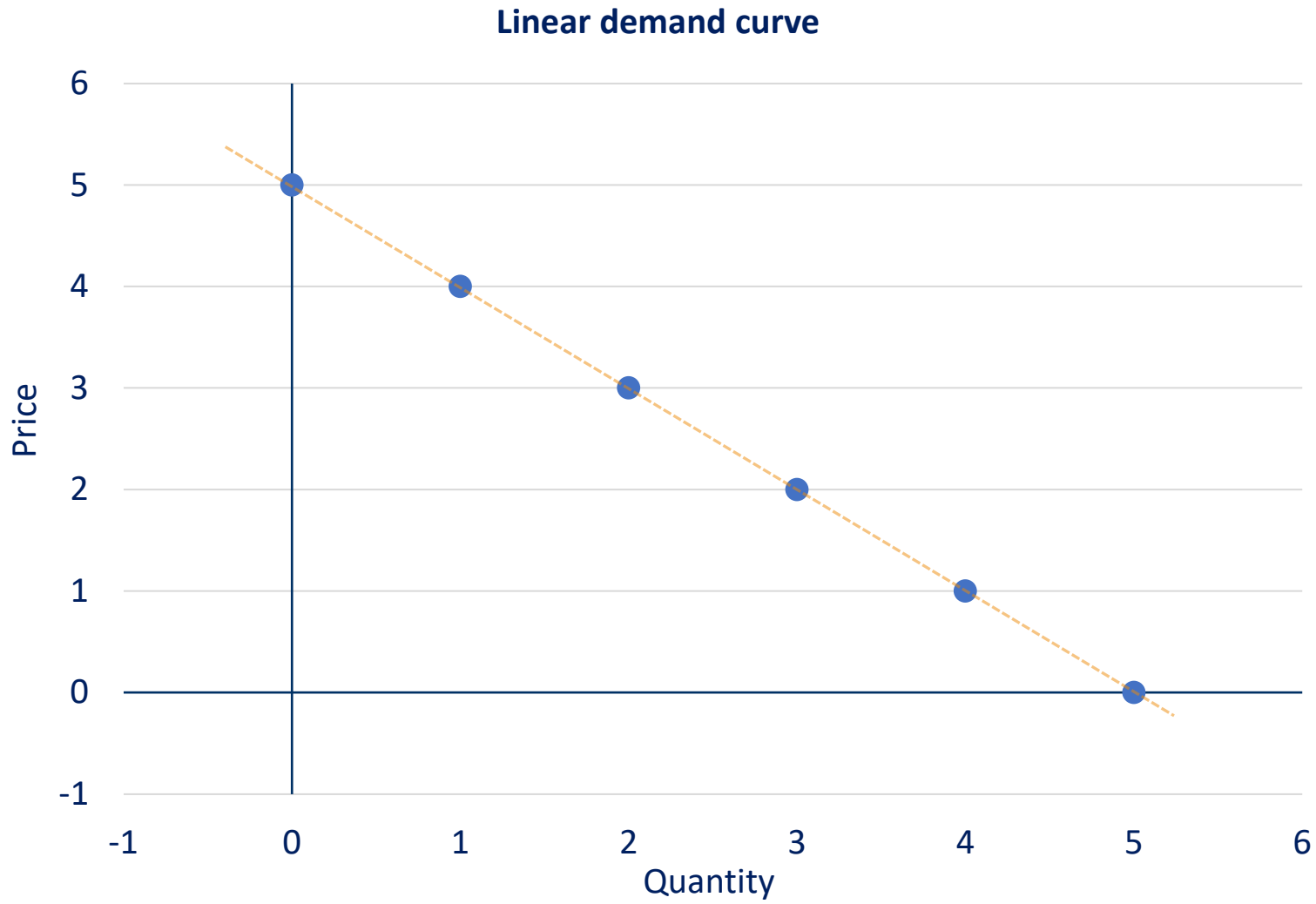
$\sigma = 1$

GAUSSIAN DISTRIBUTION – CIoT EXAMPLE



LINEAR RELATIONS

Correlation



Other examples:

Individual demand of products and services

Voltage and electricity (Ohm's law)

Volume and temperature

C2 General

NON LINEAR RELATIONS BETWEEN VARIABLES

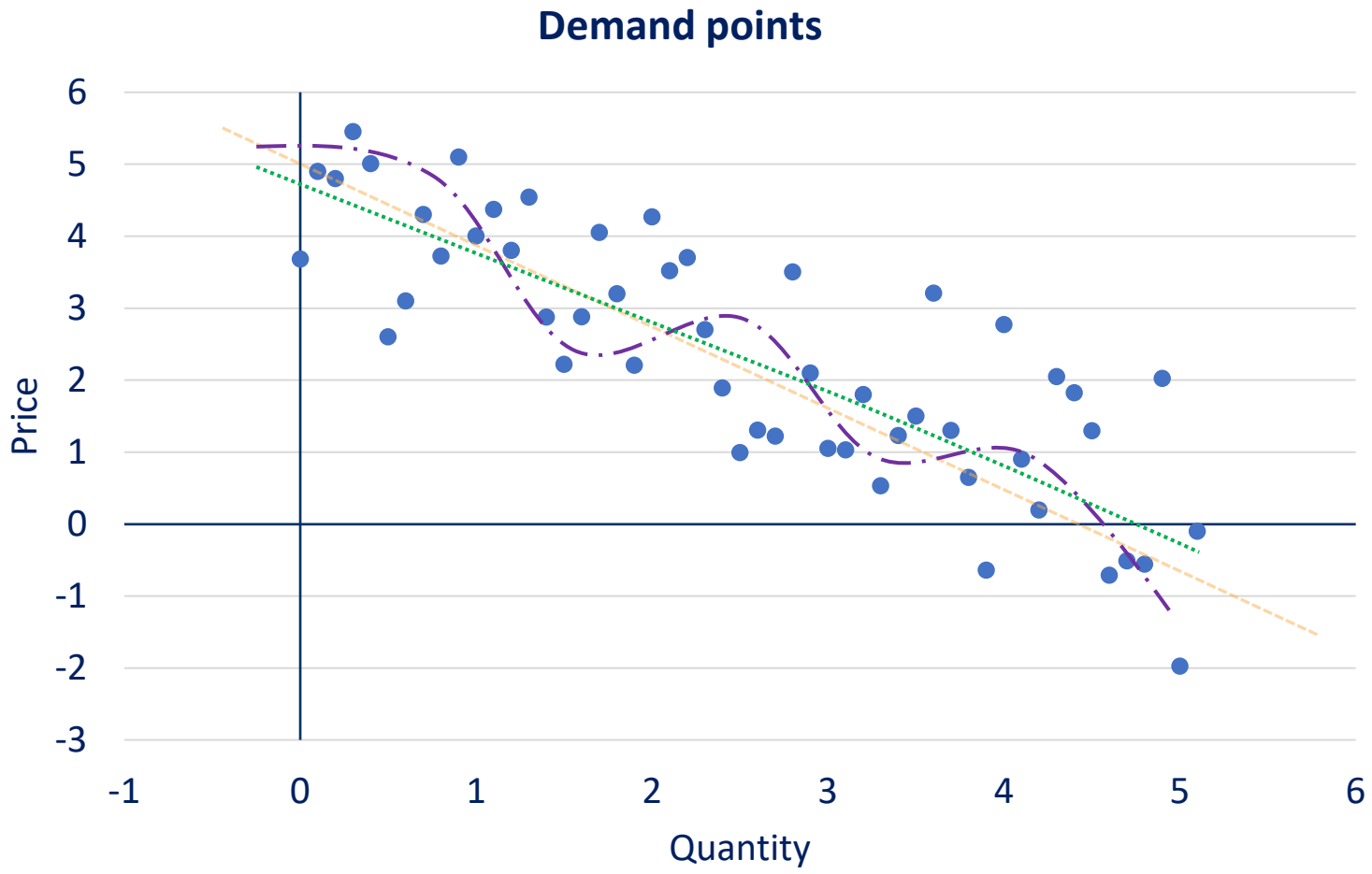
Correlation

But no real data fits a straight line perfectly

Linear model

Quadratic model

Others…

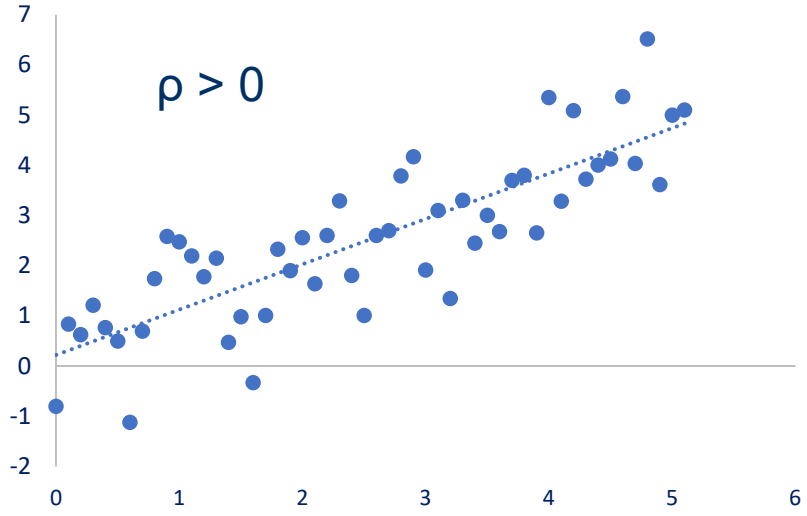


C2 General

PEARSON CORRELATION (ρ) AS LINEAR DEPENDENCY

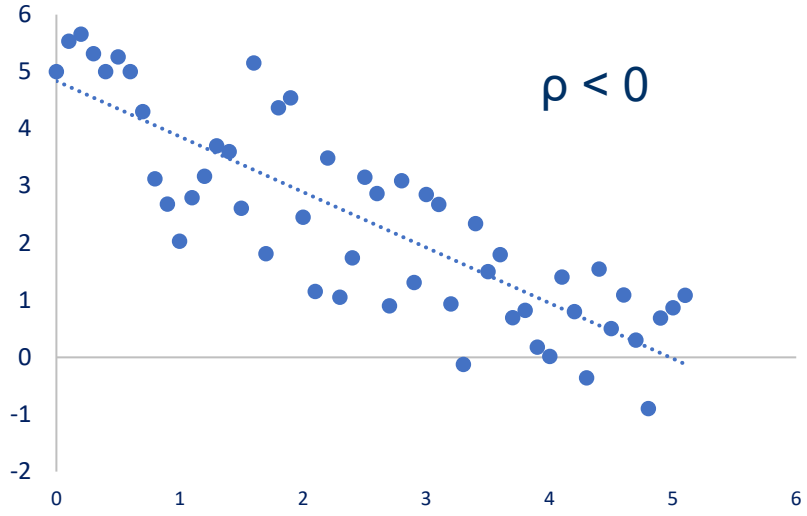
Correlation

Positive correlation



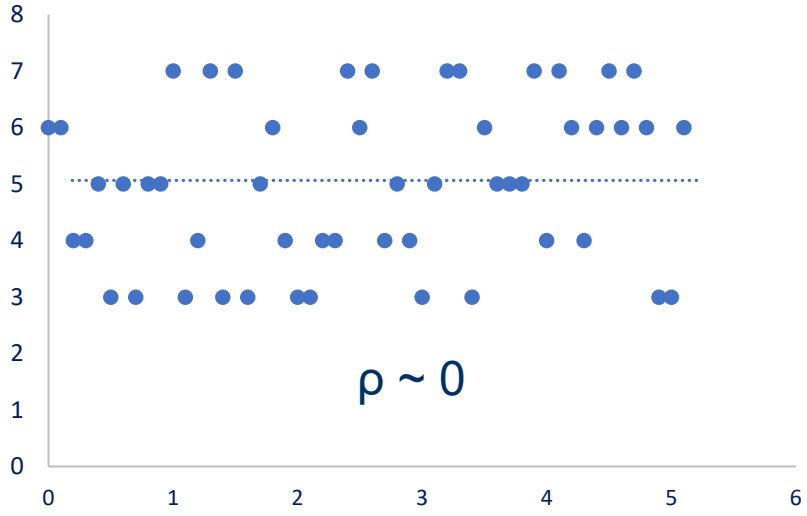
variables grow together

Negative correlation



one variable grows, the other decreases

Zero correlation



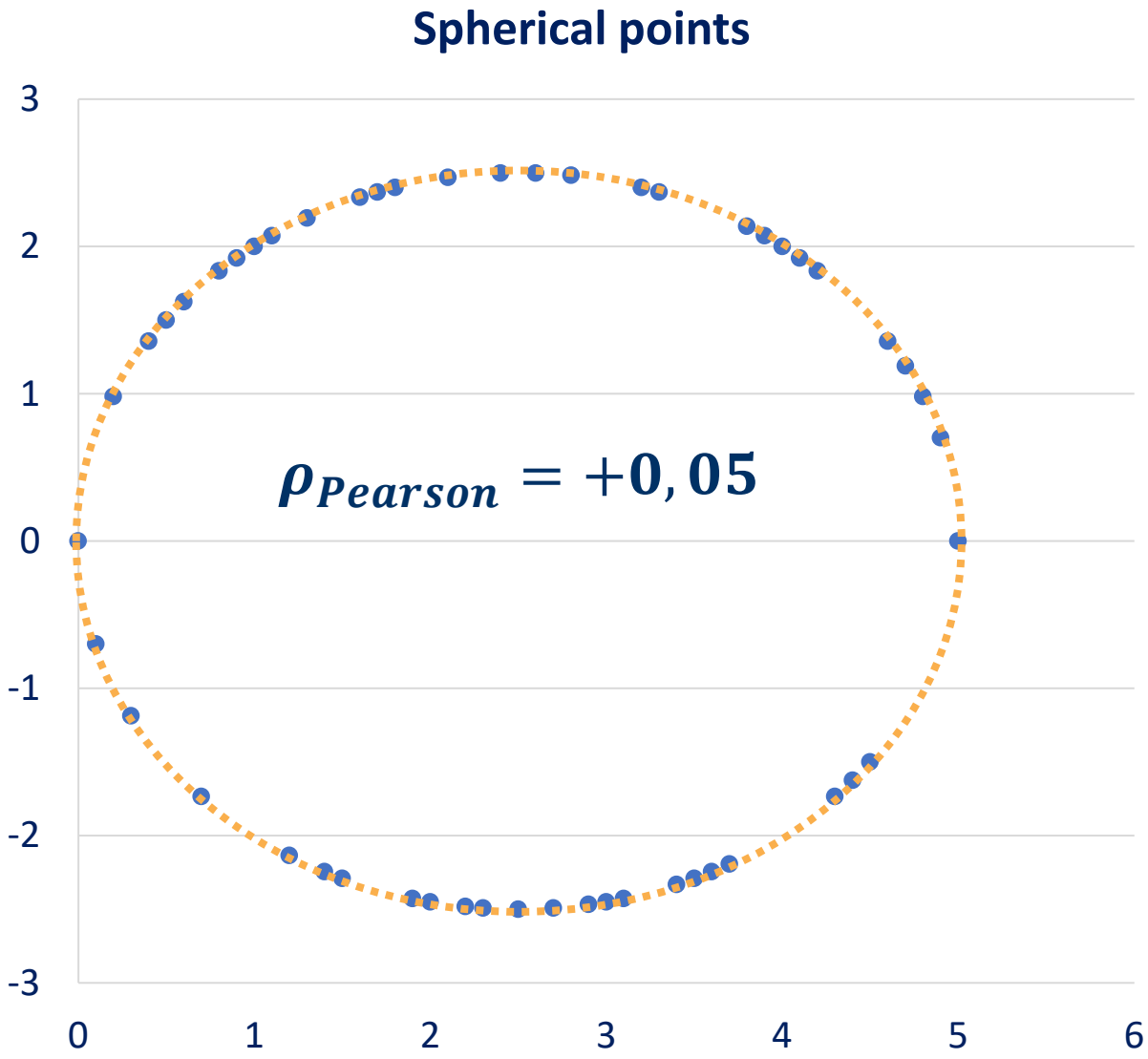
variables do not grow coherently

C2 General

WHEN CORRELATION IS MEANINGLESS

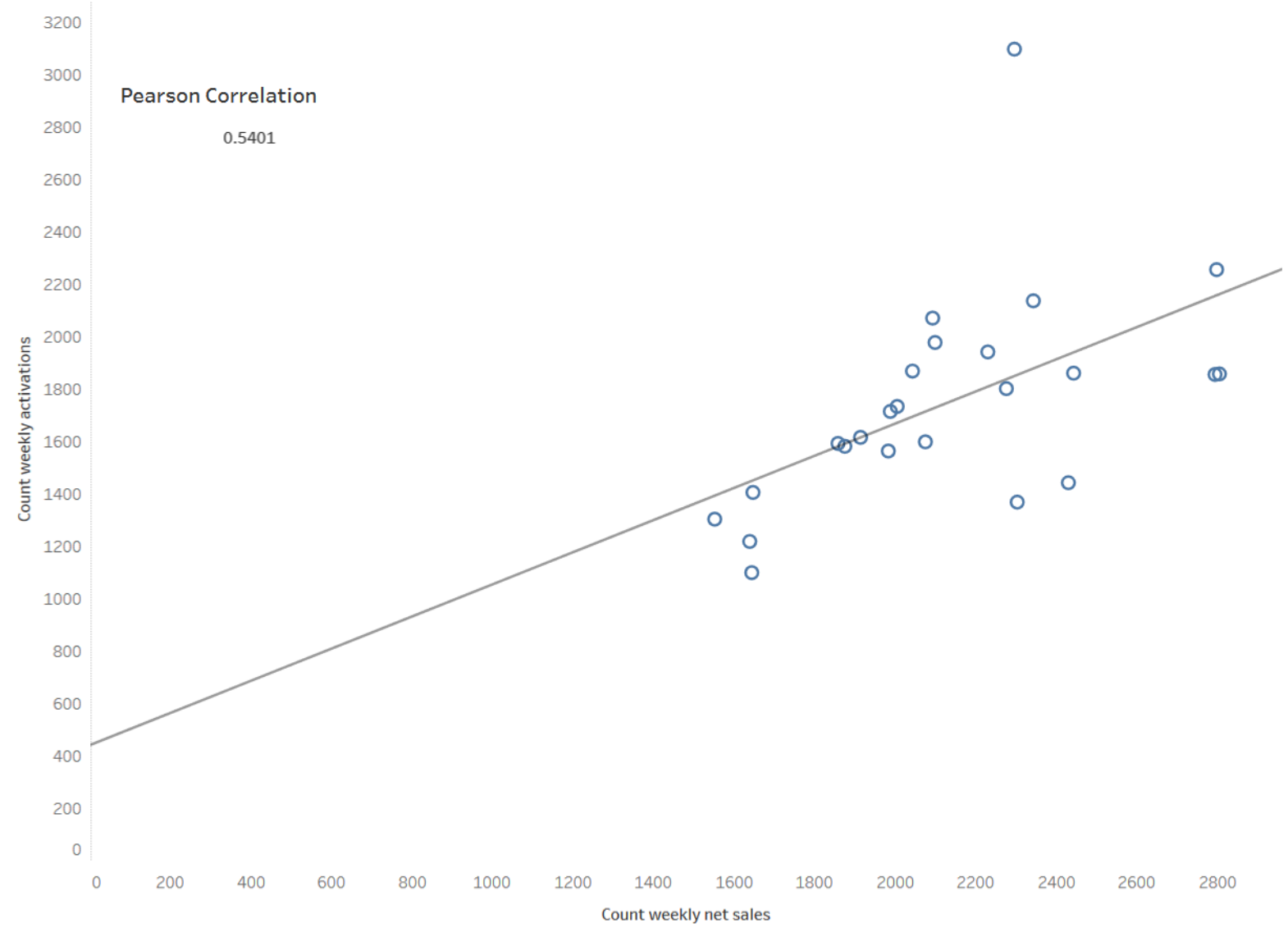
Small correlations do not exclude non-linear models

For points supporting a **non-linear model** Pearson's correlation **is meaningless**



CORRELATION– CIoT EXAMPLE

Simple Manipulations



CROSS-CORRELATION vs AUTO-CORRELATION

Correlation

New player: **TIME**

CROSS-CORRELATION
two different series are compared

AUTO-CORRELATION
the same series are compared with themselves

C2 General

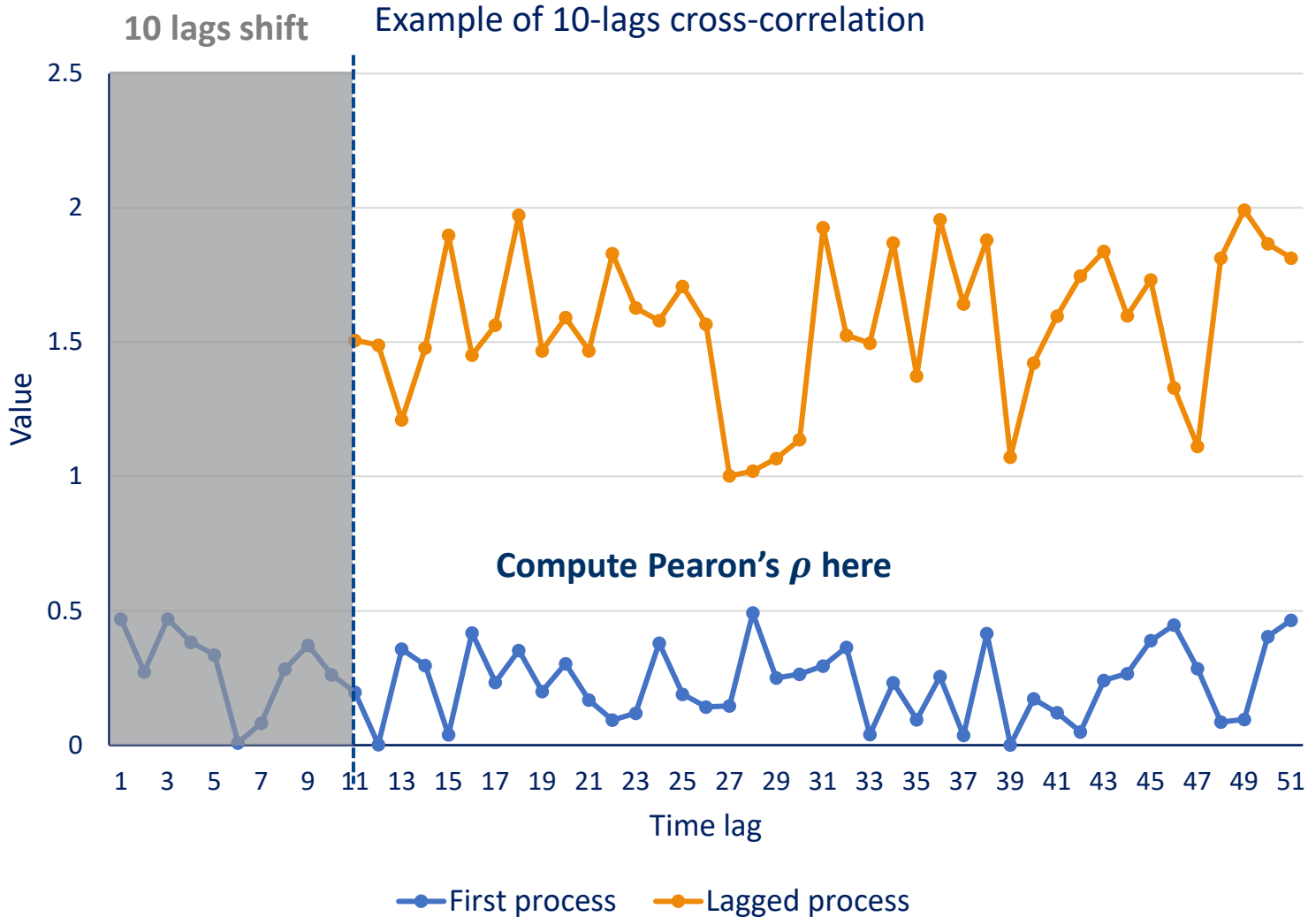
CROSS-CORRELATION

Process steps:

Fix a time window of n lags

Shift either series of exactly n lags in time

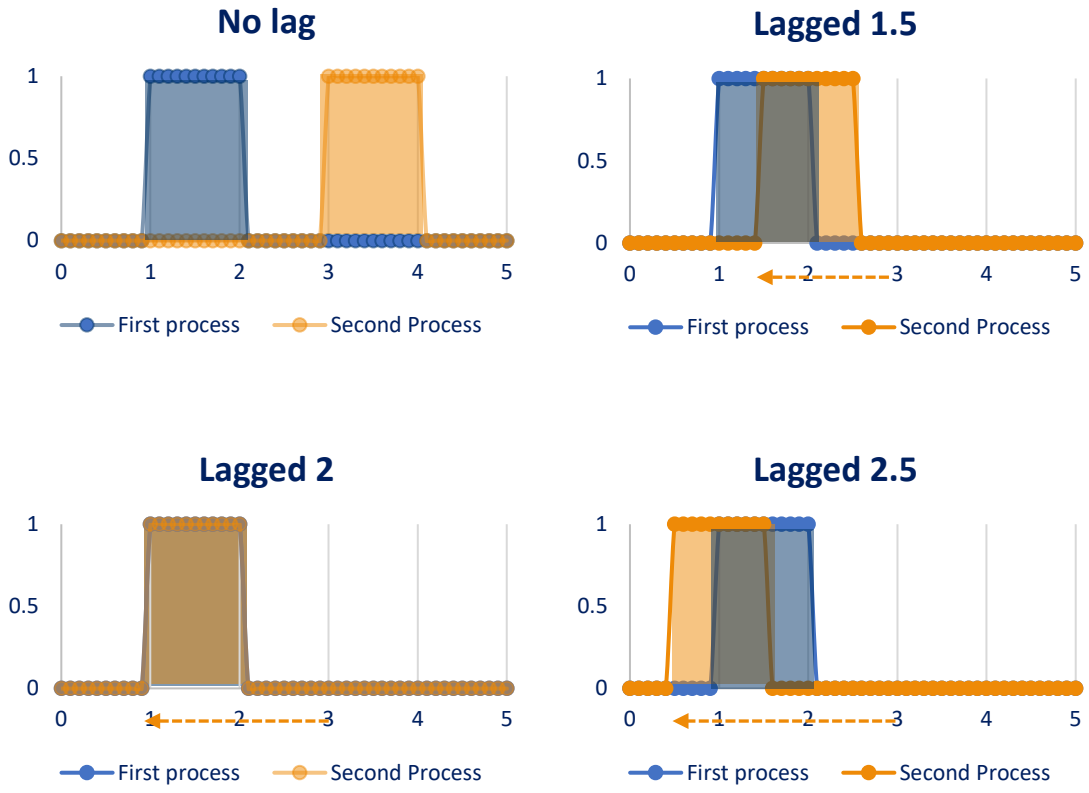
Compute the Pearson correlation over the common time region

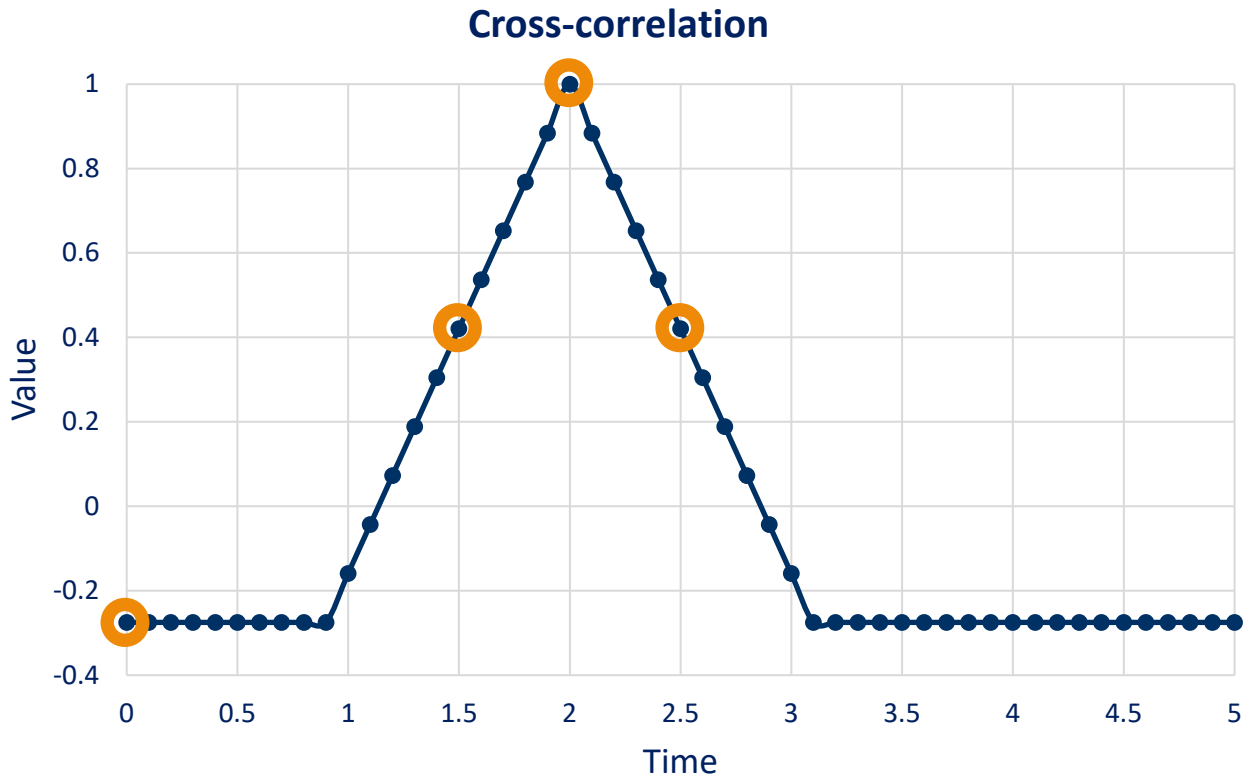


CROSS-CORRELATION

Correlation

Cross-correlation tells whether two time series *look the same* when shifted in time

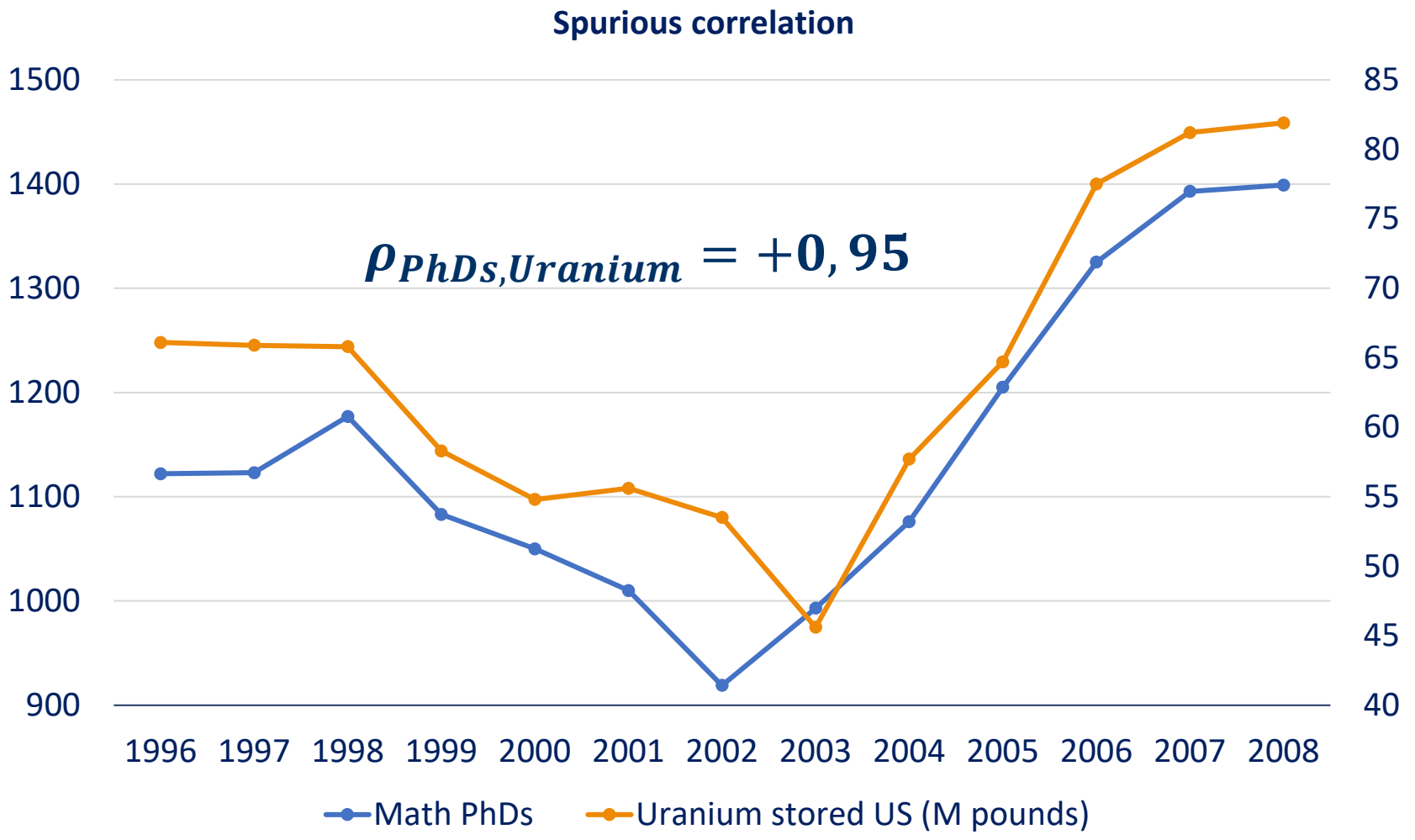




SPURIOUS CORRELATION

Correlation

Do we really believe that the increase in math PhDs is due to the increasing storage of uranium?



C2 General

Correlation

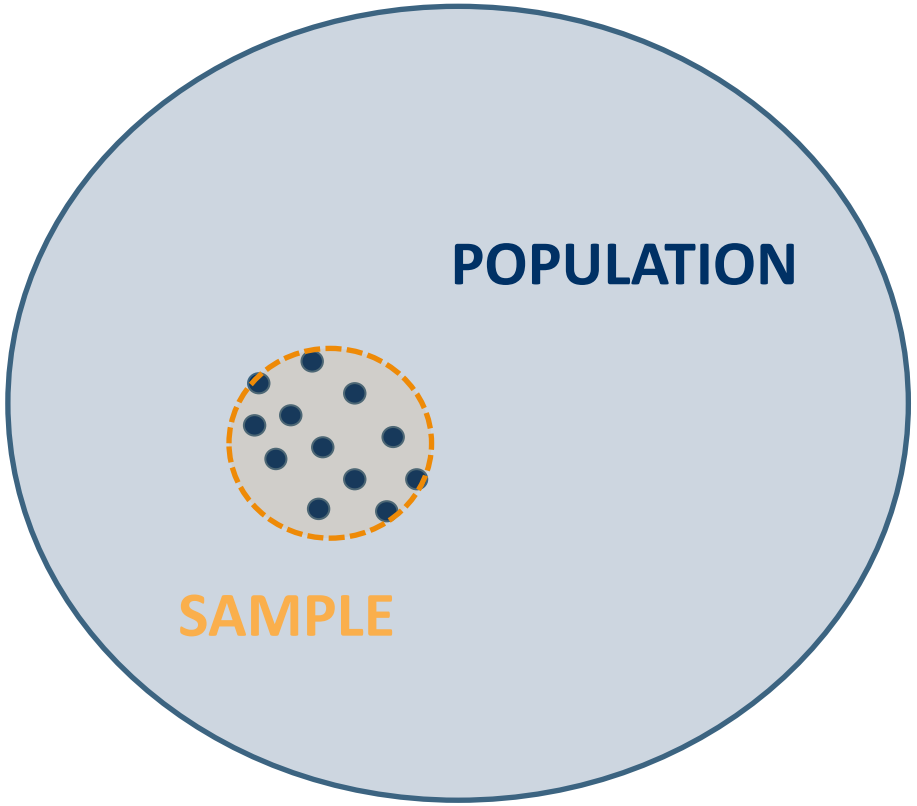
is not

Causation

SAMPLE VS POPULATION

POPULATION

the theoretical **infinite** set of element from which the sample will be extracted



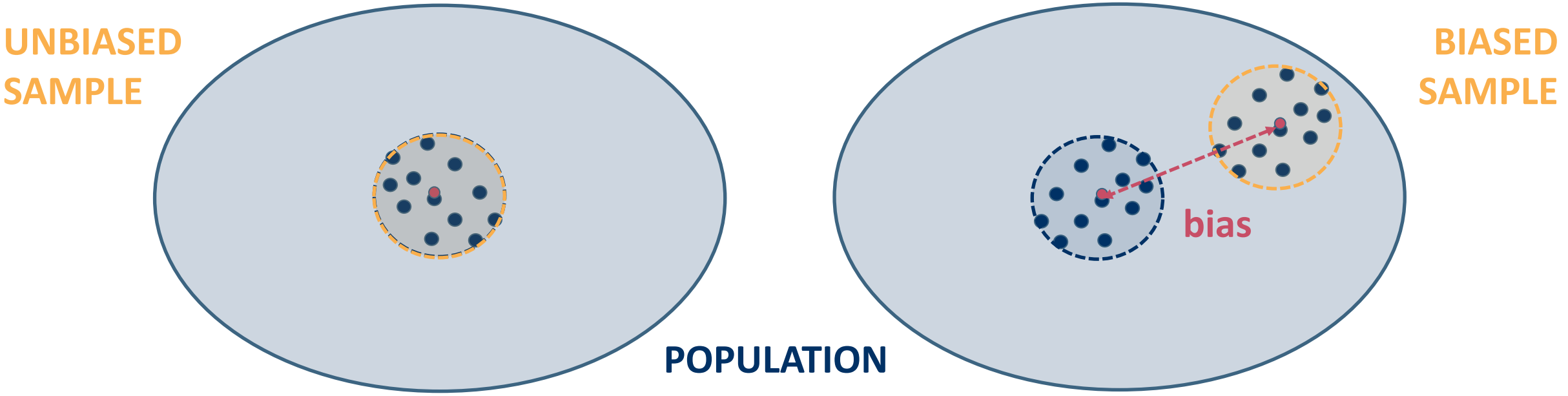
SAMPLE

the **finite** collection of data selected for the analysis

BIAS

Biased inferences are based on *shifted* samples,
which do not optimally describe the population

$$bias = \mu_{sample} - \mu_{population}$$

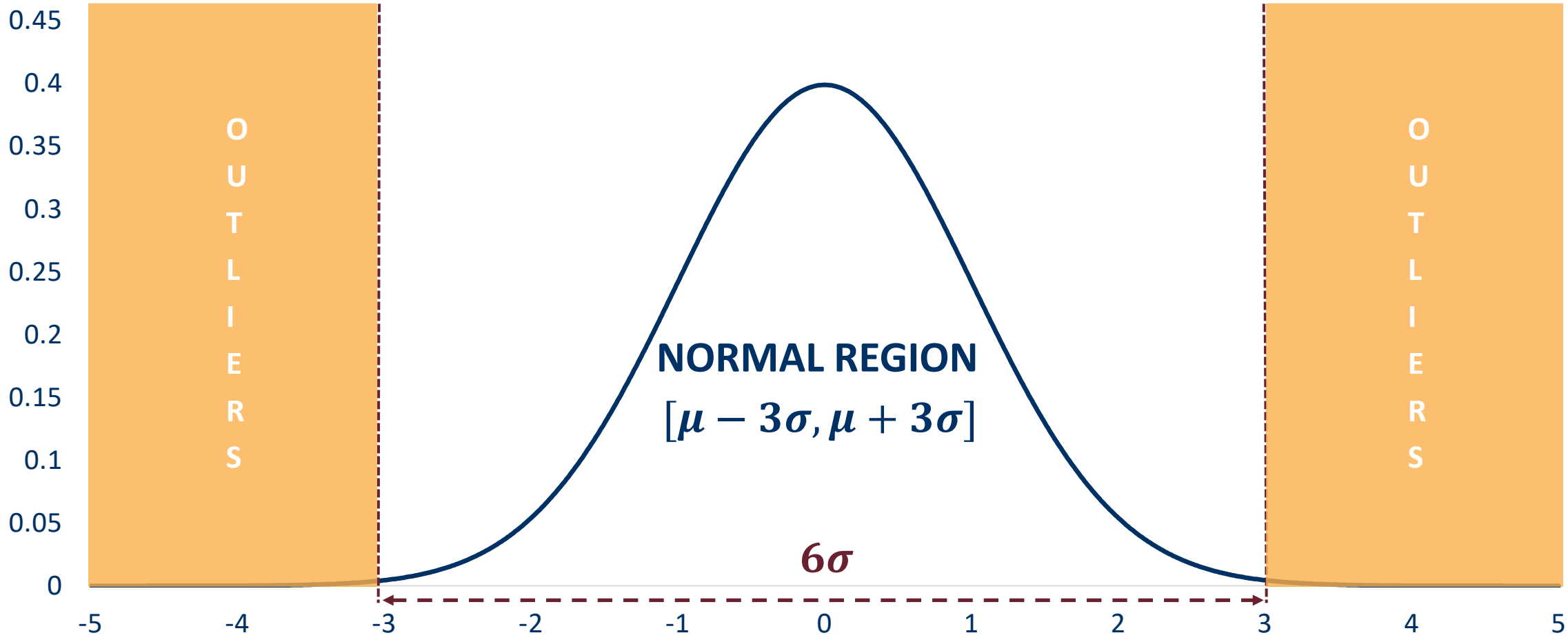


OUTLIERS PREDICTION IS A STRIKING CHALLENGE

Credits #: 9

OUTLIERS IN A STANDARD GAUSSIAN DISTRIBUTION

Inference

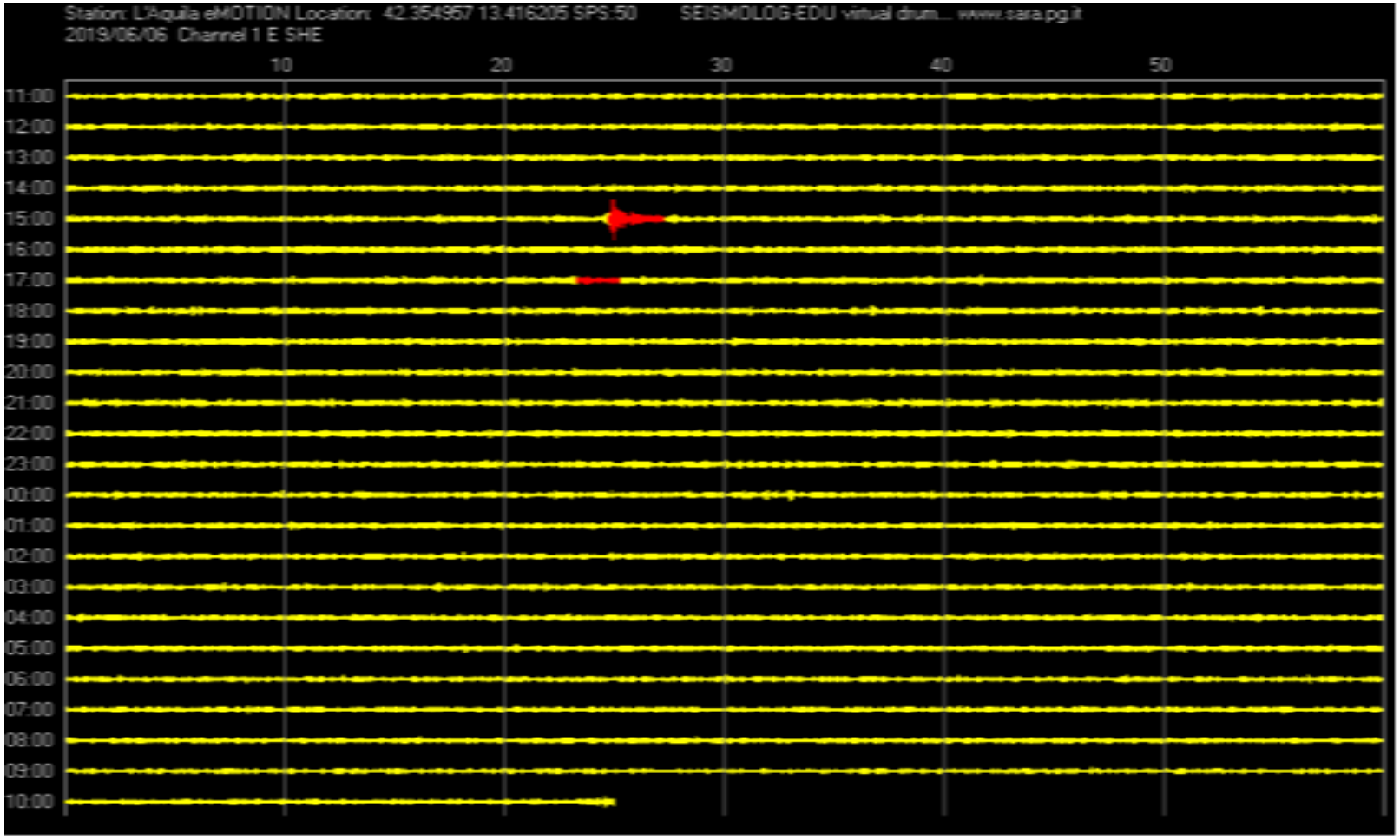


C2 General

OUTLIERS: FROM WHITE NOISE..

Inference

Sismogram of L'Aquila, live data, 7 June 2019

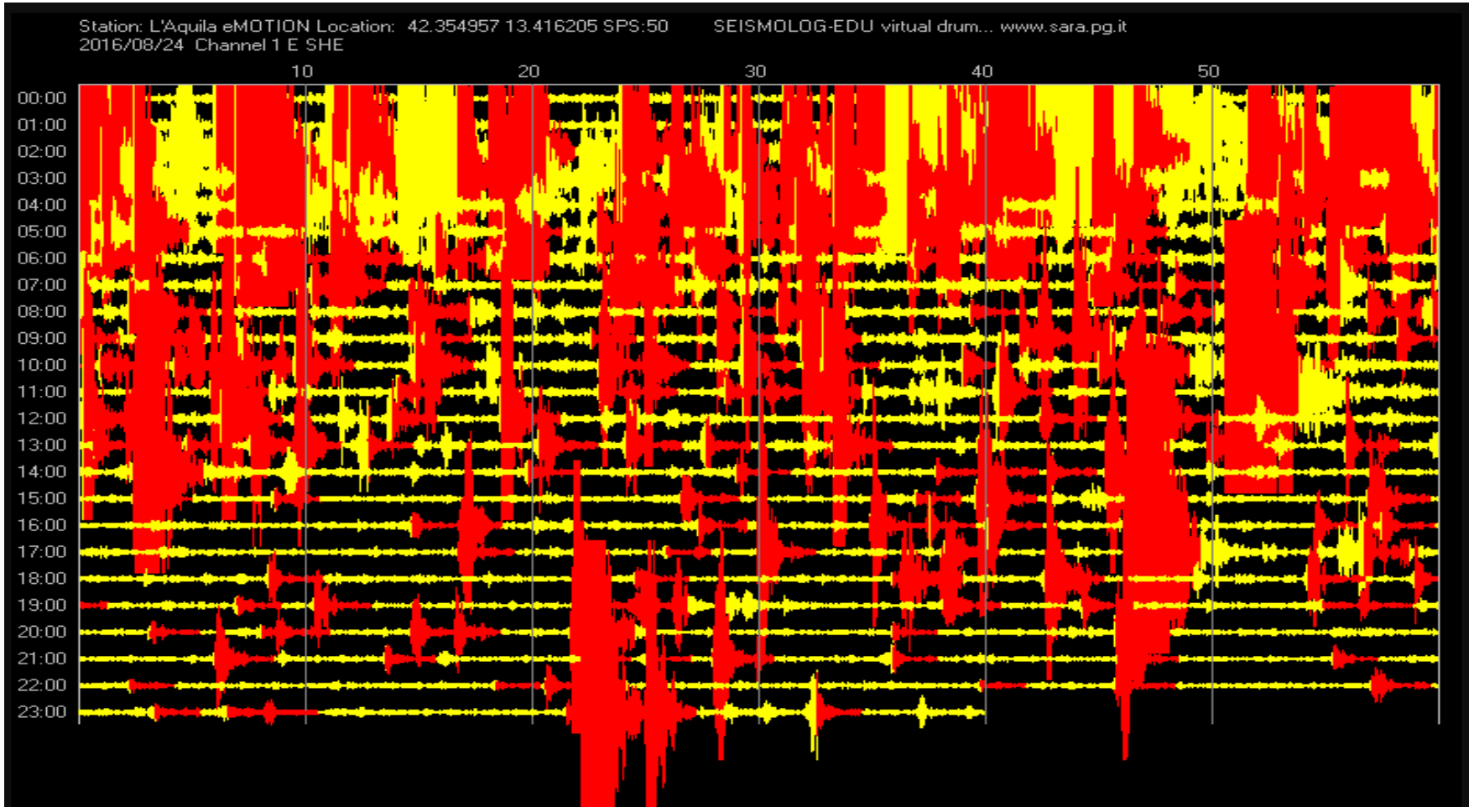


Credits #: 10

...TO DISRUPTIVE EVENTS

Inference

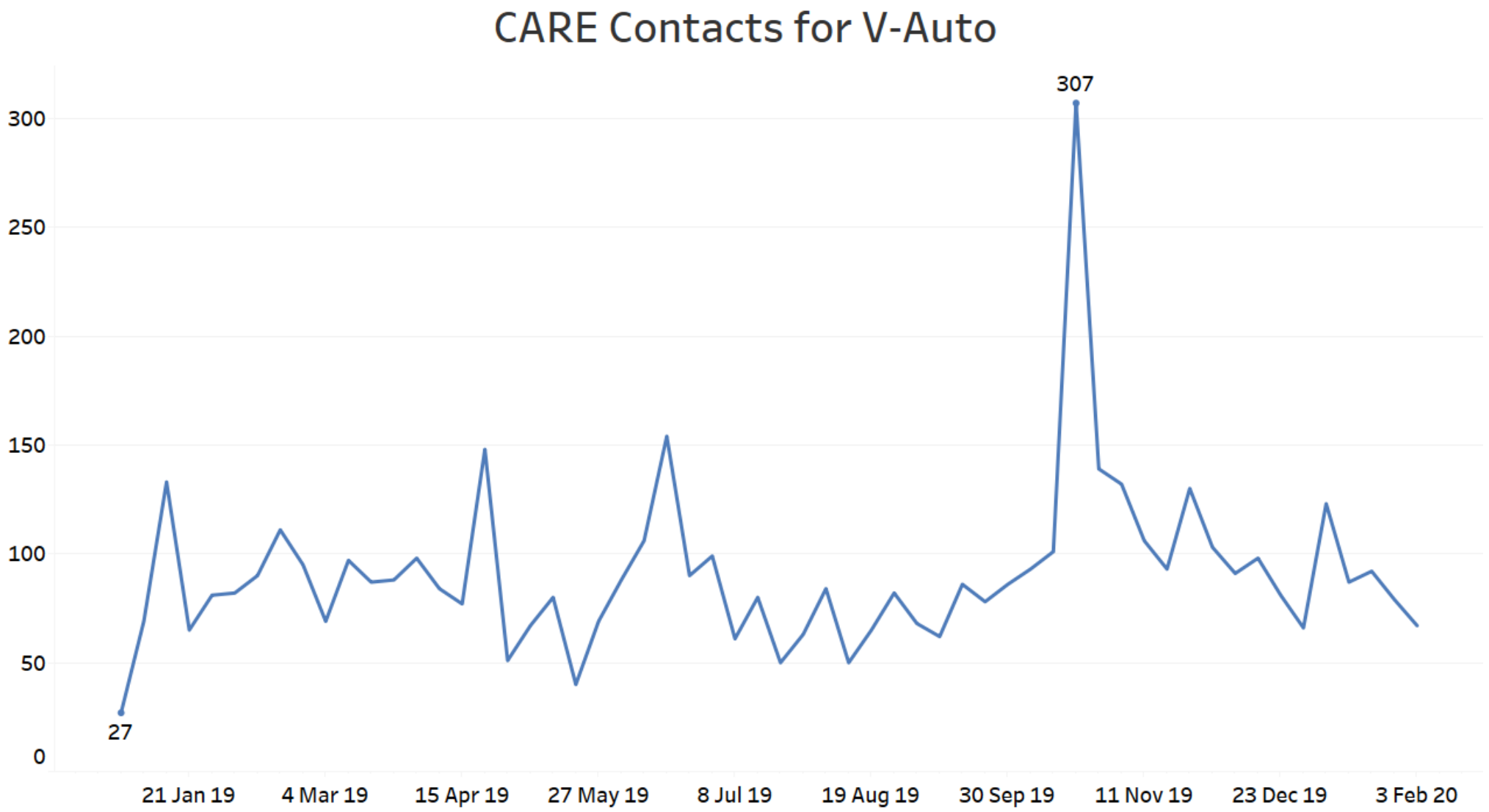
Sismogram of L'Aquila, live data, 24 August 2016



C2 General

Credits #: 10

OUTLIERS – CIoT EXAMPLE



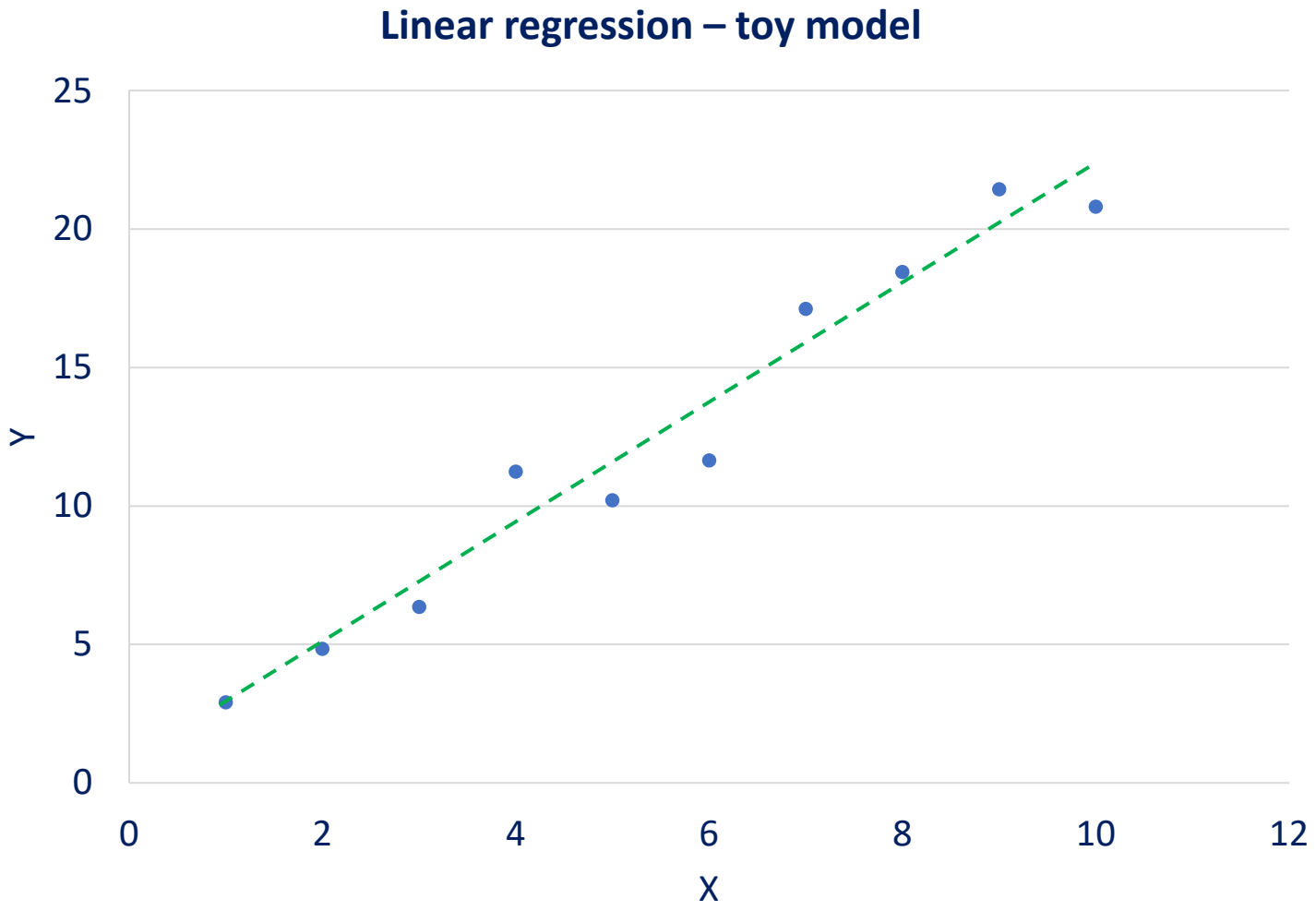
LINEAR REGRESSION

The easiest regression is a straight line

$$y = mx + b$$

parameter m:
the slope of the straight line

parameter b:
the bias (intercept)

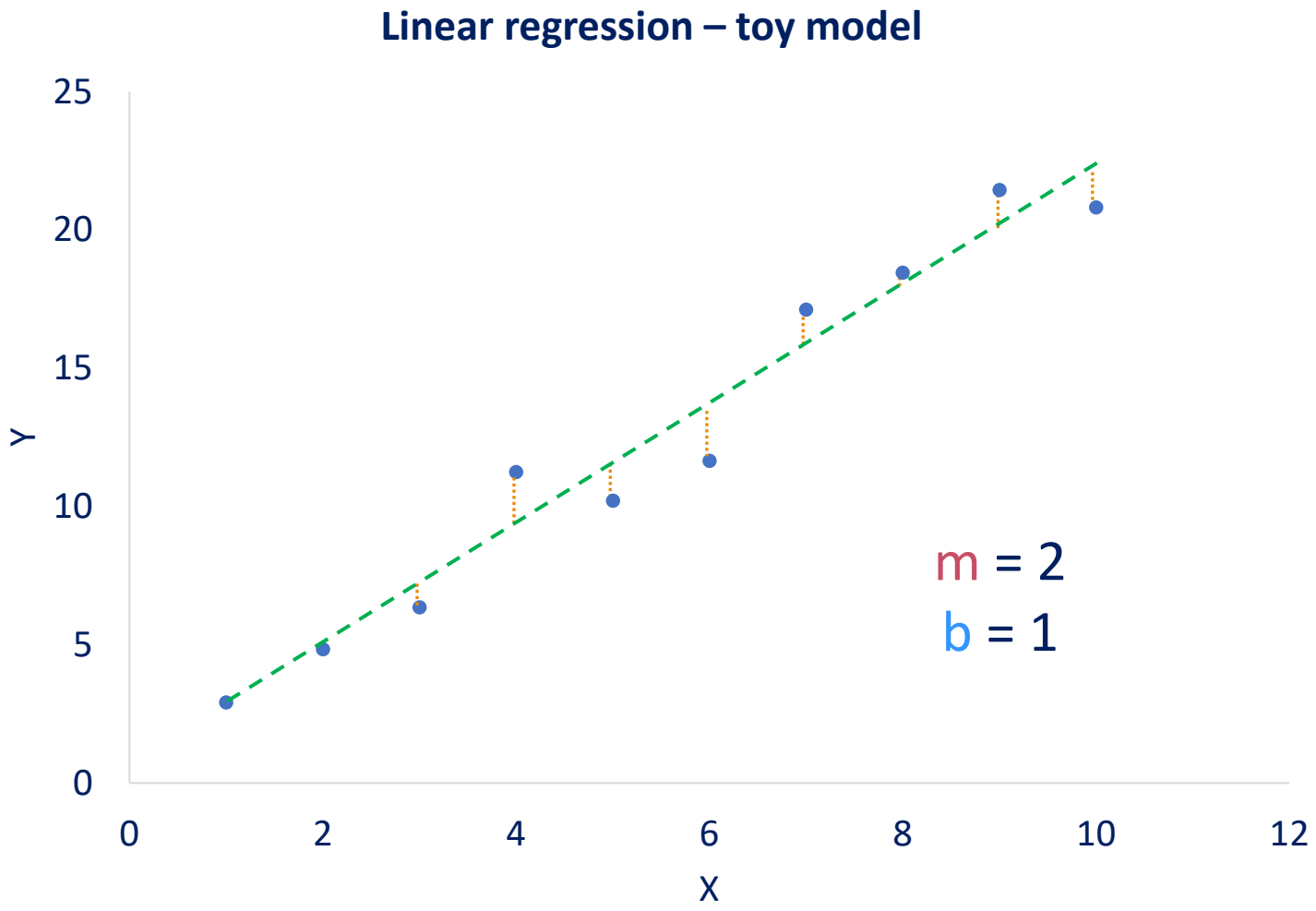


PARAMETER MEANING

Parameters are computed by minimizing the **the sum of vertical errors**

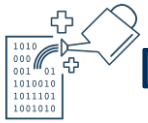
$$m = \rho_{Pearson} \frac{\sigma_y}{\sigma_x}$$

$$b = \mu_y - m\mu_x$$



AGENDA

Data processing



Data manipulation

Data analysis

Classwork

CLASSWORK

Fitness Gym Company

Assignment

In this scenario you are the General Manager of the company, being responsible of the business of all gyms, presenting to the investors board several gyms information.
The objective is to evidence from data, in order to help (at high level) future business of the company. Here is your data, for 3 selected gyms:

# of accesses per customer

Number of customers (per gym and section)

Staff Salaries

Revenues

Maintenance Expenses (per gym and section)

GIVE YOUR BEST SHOT!

KEY CONCEPTS

Mode
The most recurrent observation

Median
The observation that splits the set in two parts containing the same number of values

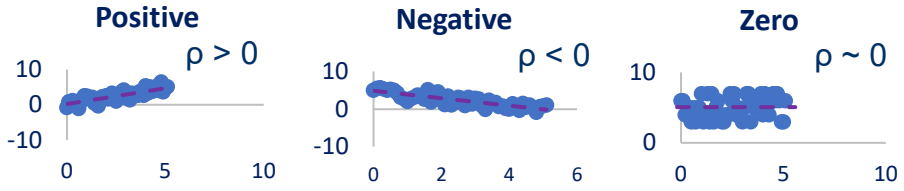
Percentile
The median is the 50th percentile

Average μ
The central value of a set of numbers

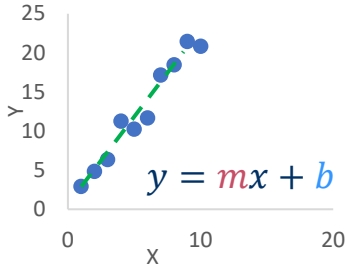
Variance σ^2
How far a set of numbers are spread out from their average value

Standard Deviation σ
Measure of dispersion of a set of values

Correlation ρ



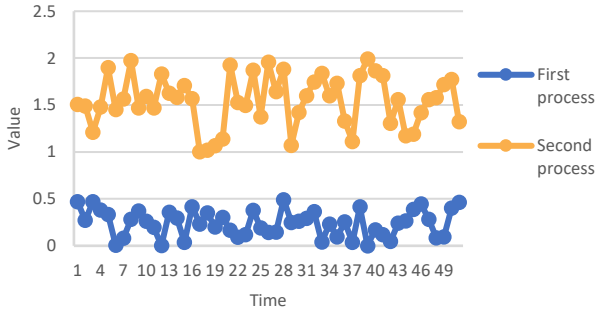
Linear Regression



$$m = \rho \frac{\sigma_y}{\sigma_x}$$

$$b = \mu_y - m\mu_x$$

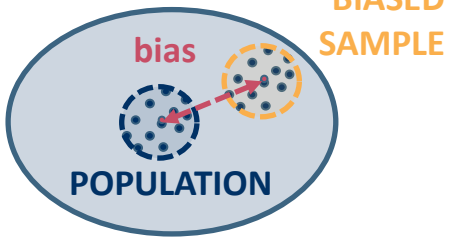
Time series & cross/auto-correlation



Cross-correlation
two different series are compared

Auto-correlation
the same series are compared with themselves

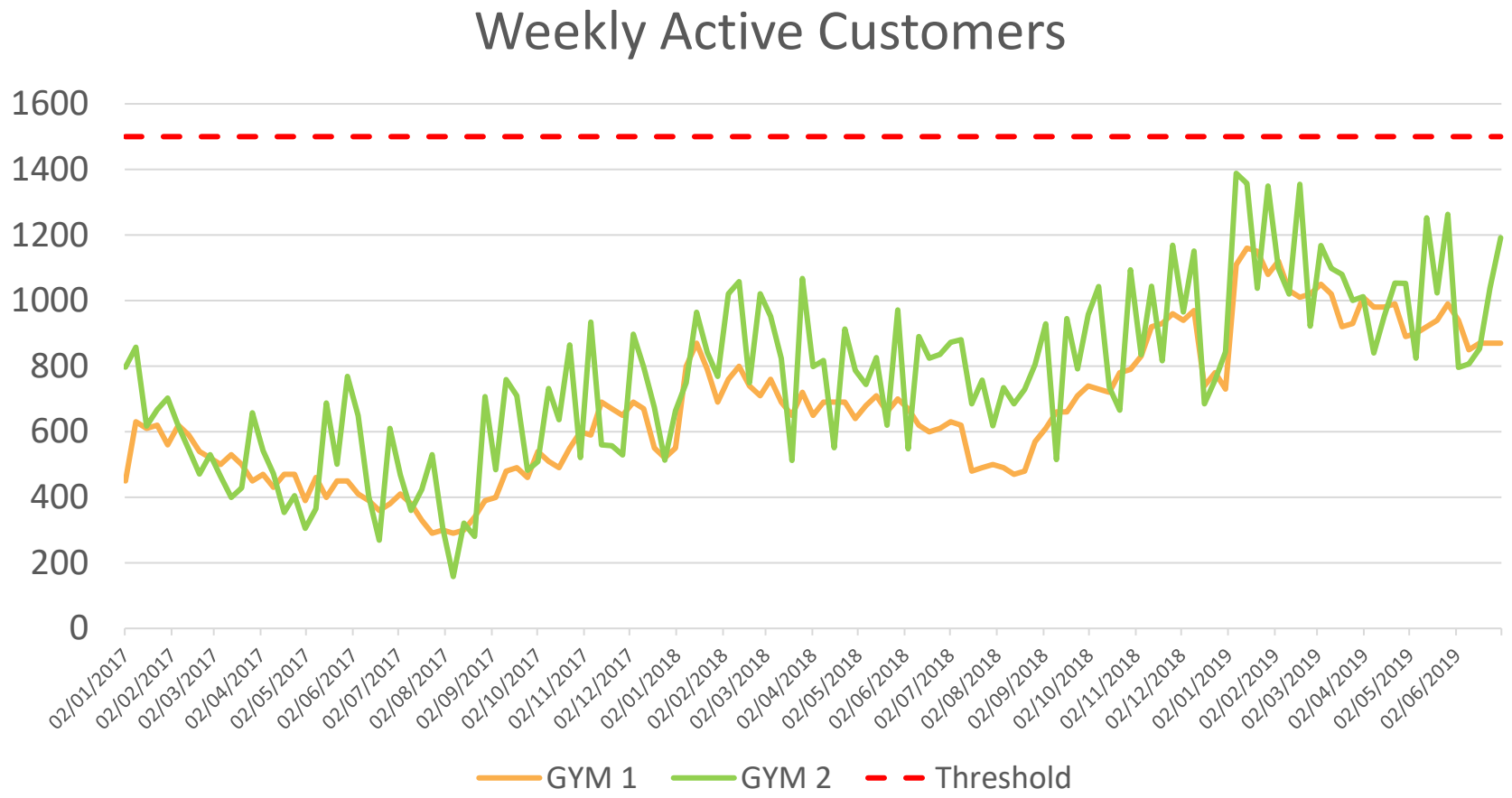
Bias



20 minutes

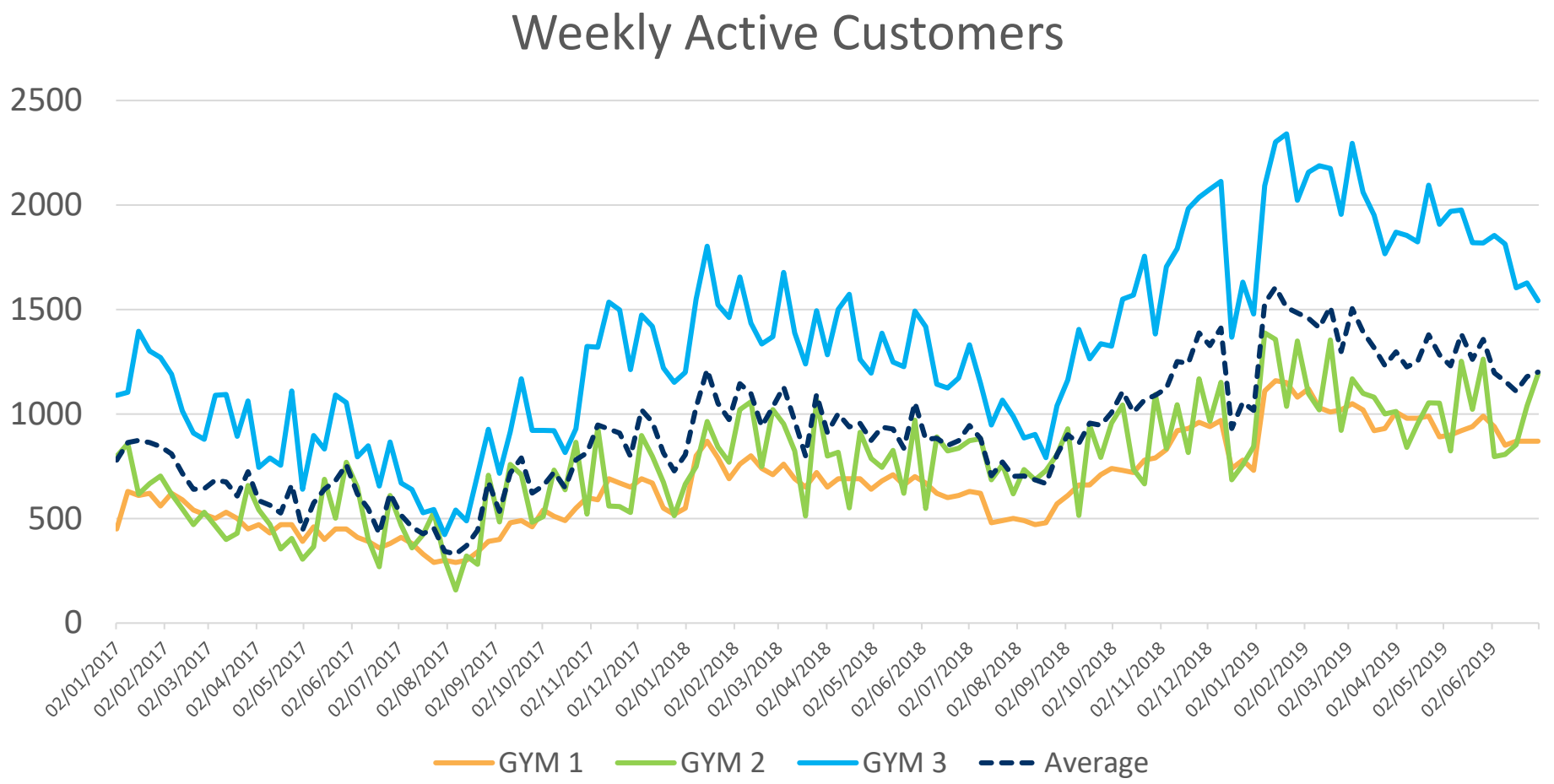
GYM CUSTOMERS SATURATION

Possible Output



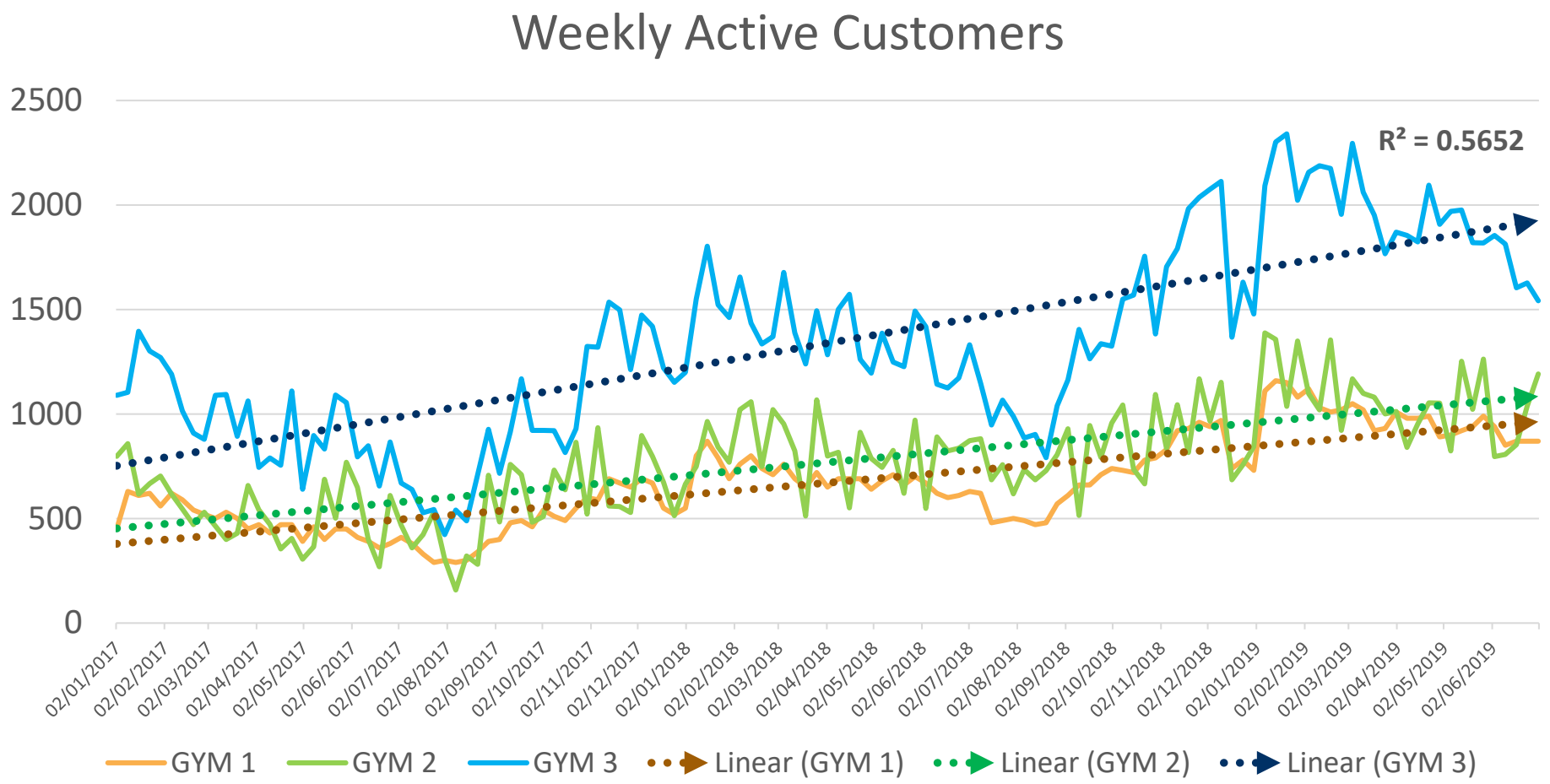
COMPARISON ON ALL GYMS TREND

Possible Output



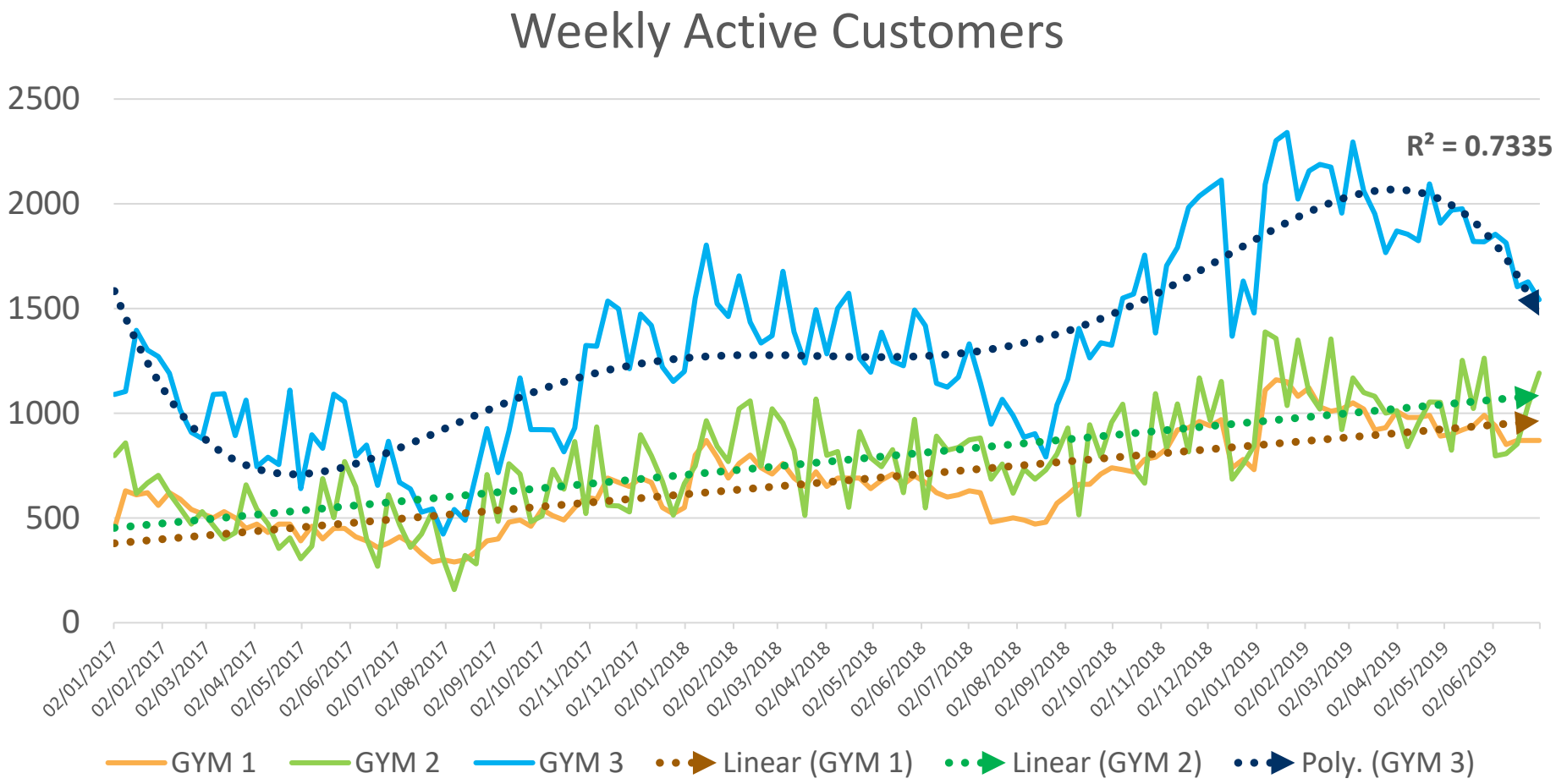
COMPARISON ON ALL GYMS TREND

Possible Output



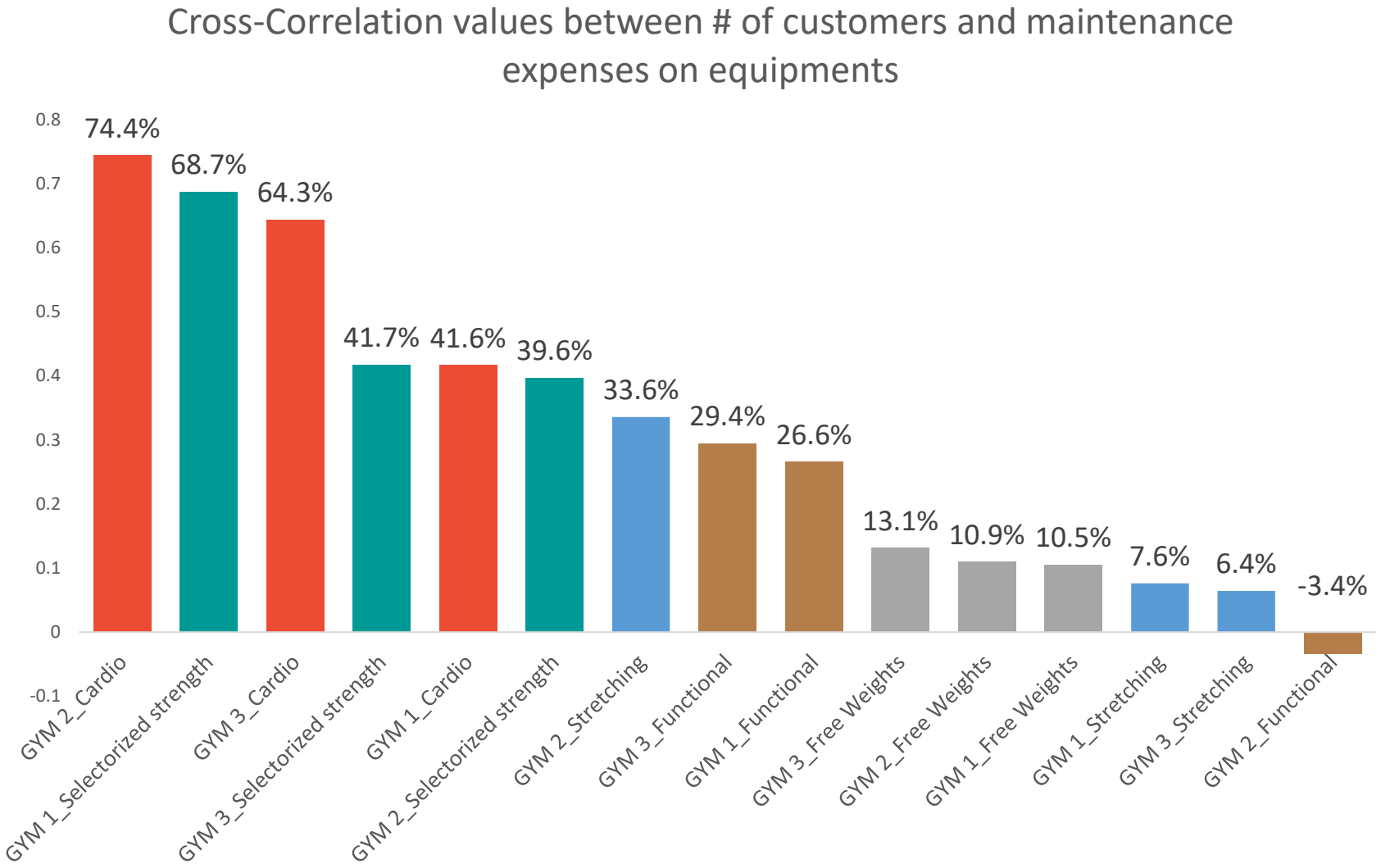
COMPARISON ON ALL GYMS TREND

Possible Output



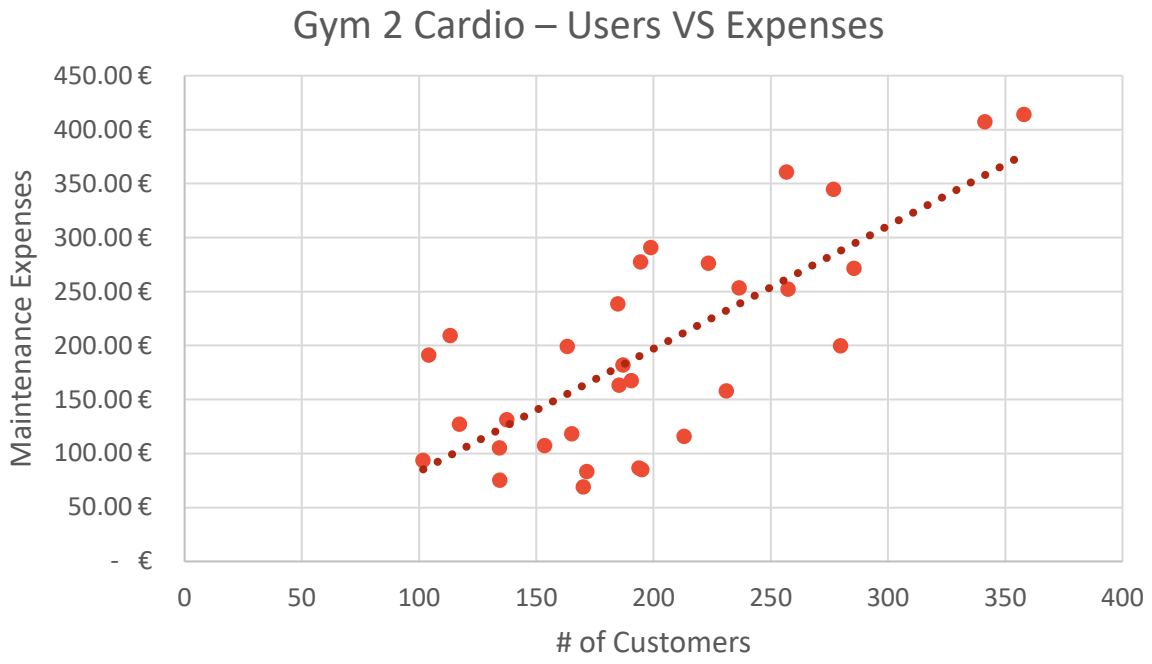
RELATION BETWEEN EQUIPMENT USAGE AND EXPENSES

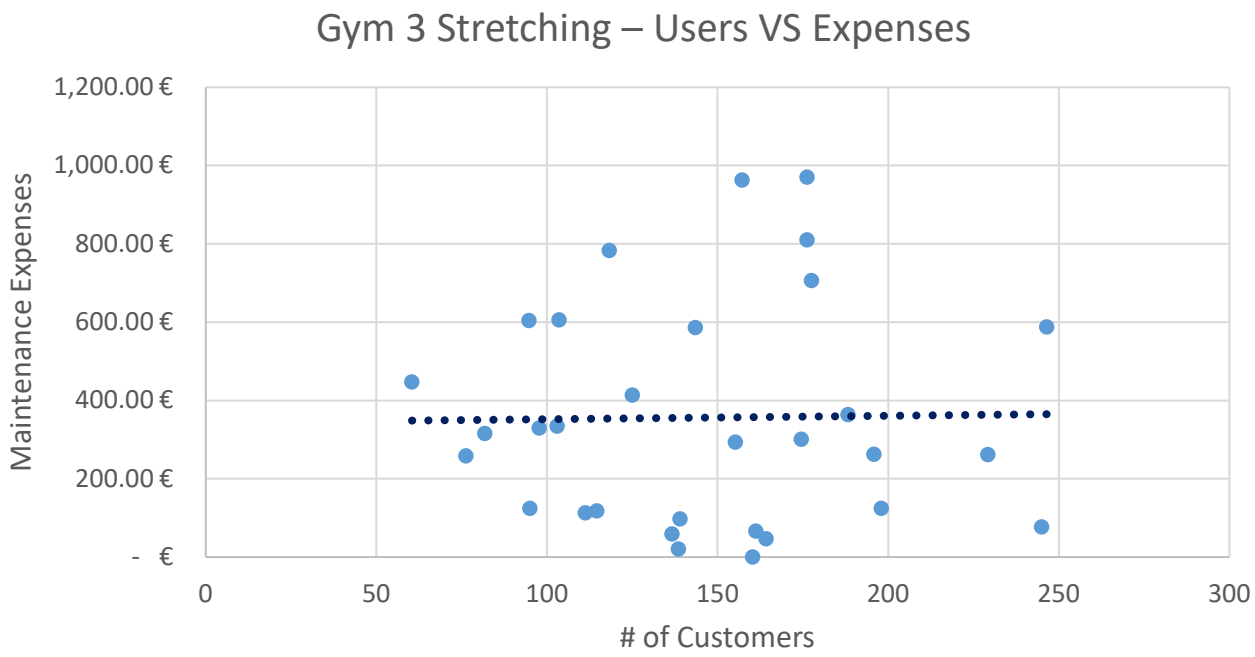
Possible Output



RELATION BETWEEN EQUIPMENT USAGE AND EXPENSES

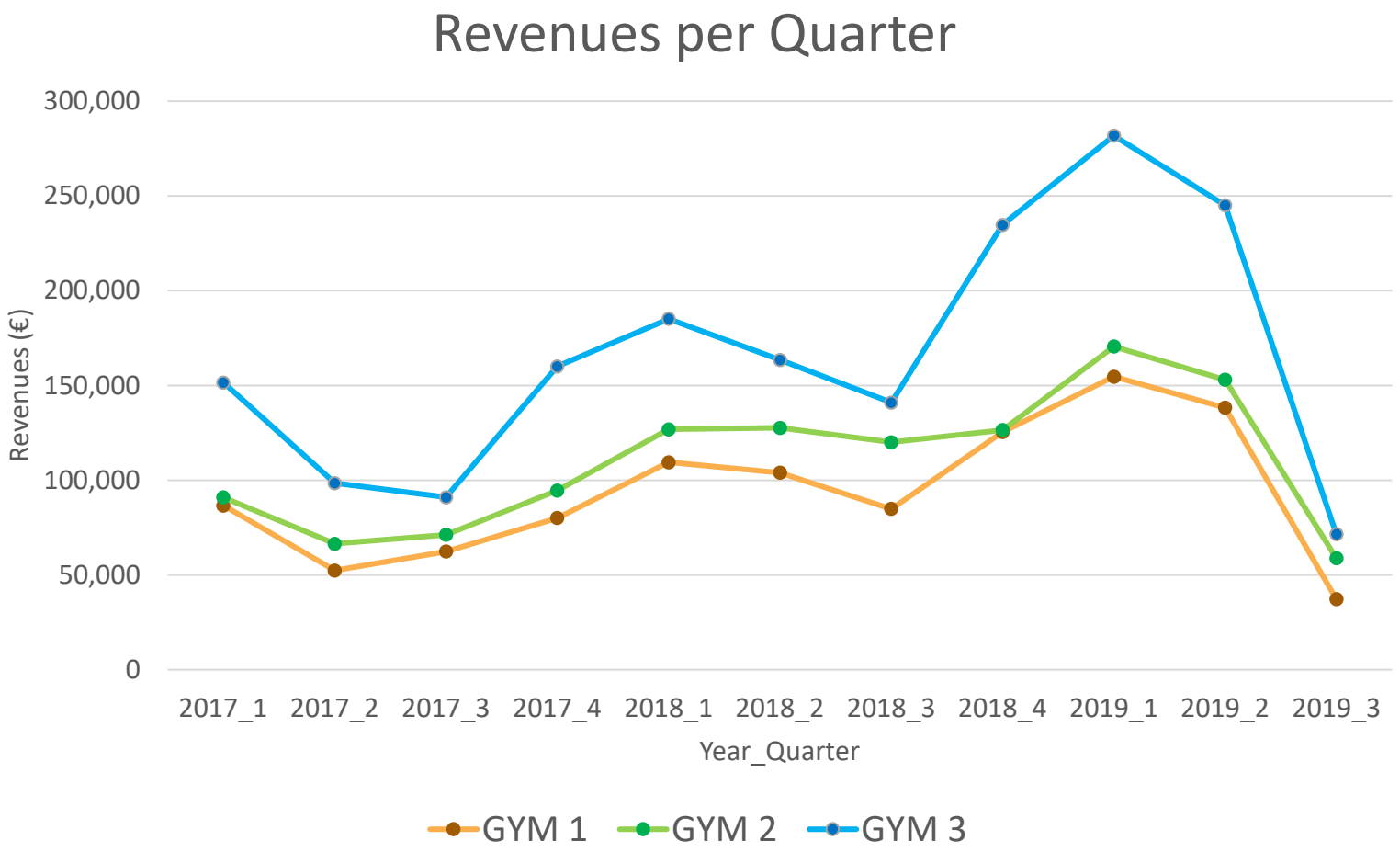
Possible Output





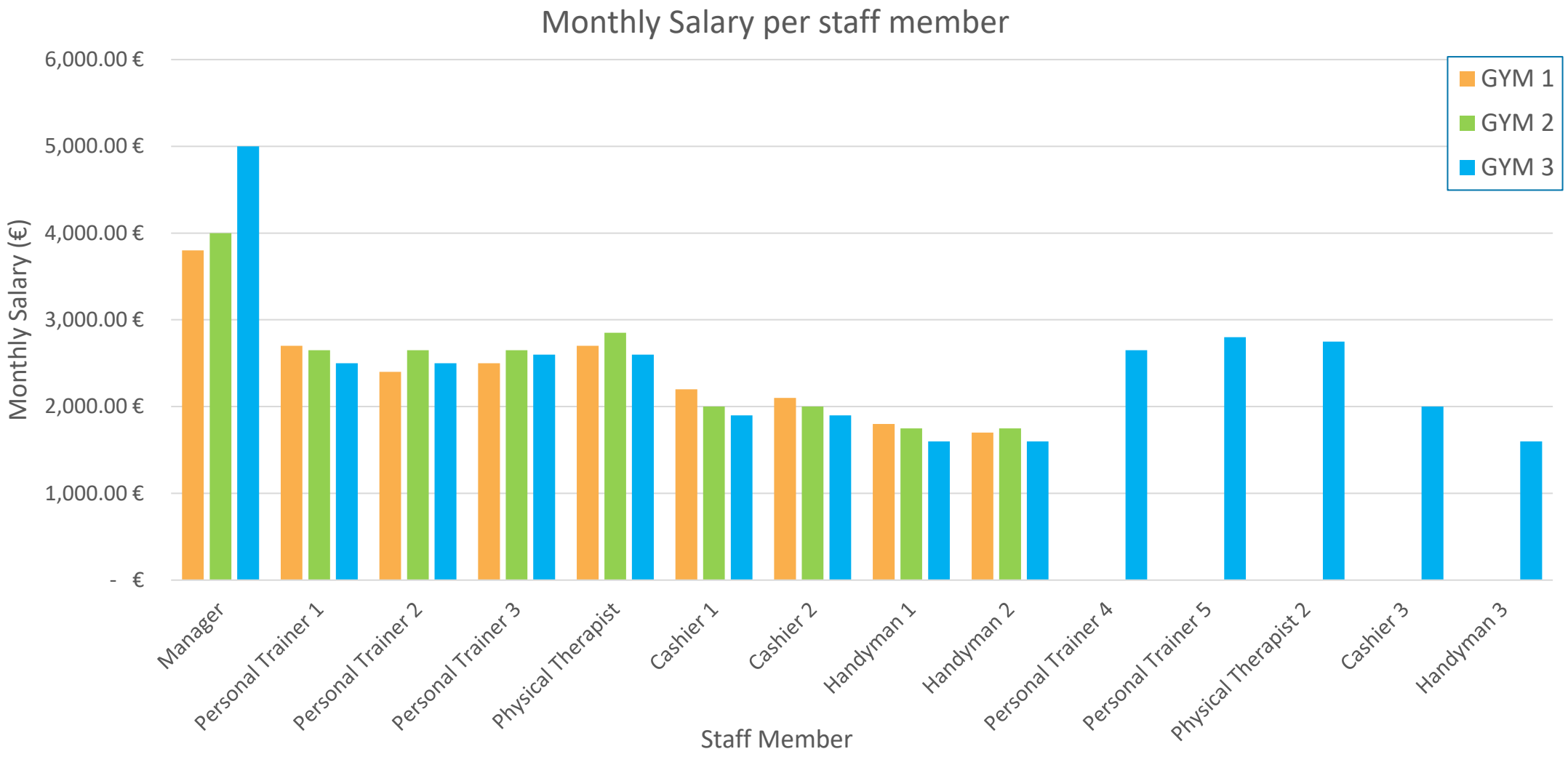
REVENUES COMPARISON

Possible Output



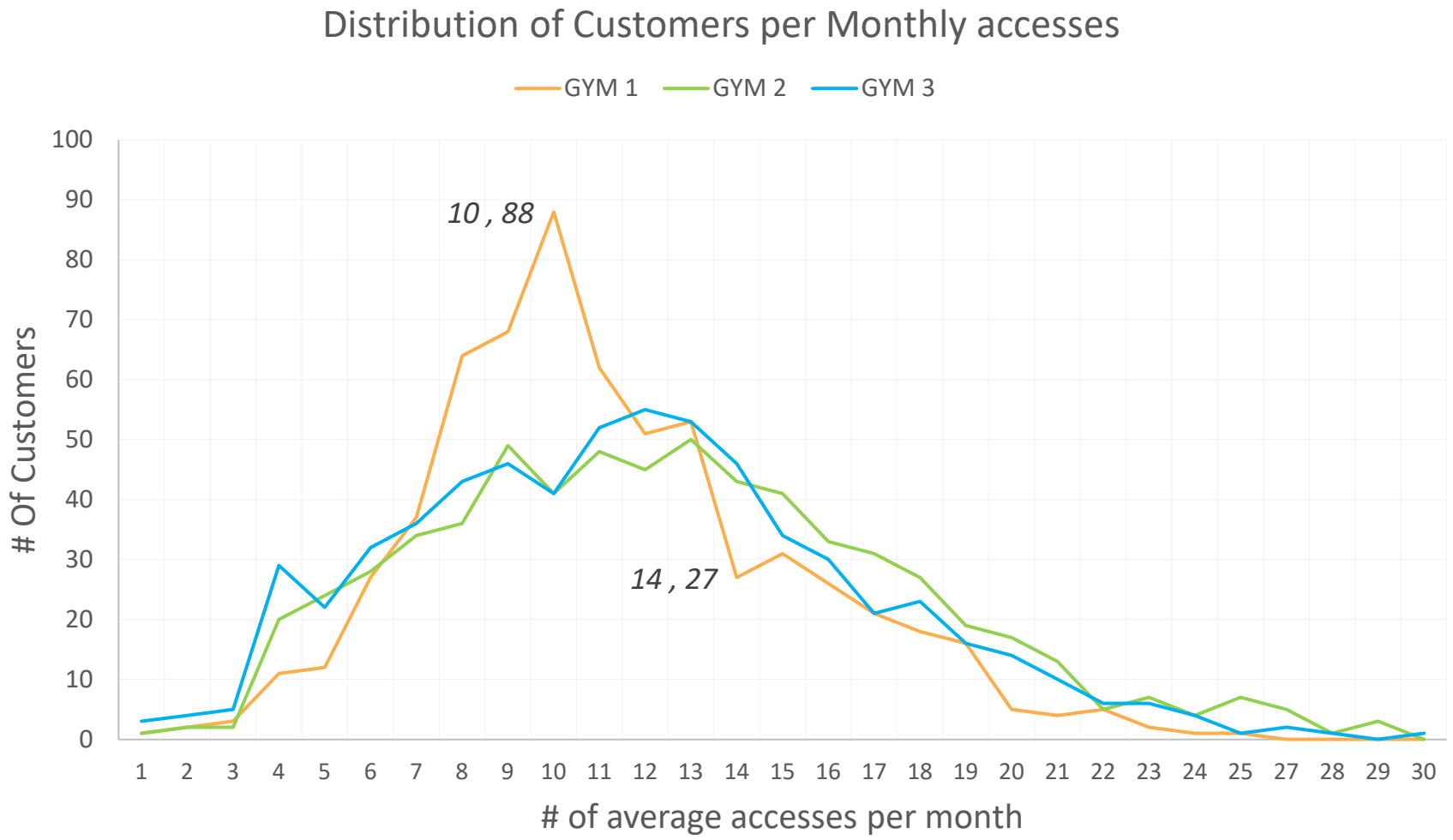
SALARIES COMPARISON

Possible Output



VISITS PER CUSTOMER

Possible Output



DATA FITNESS PROGRAM

DAY 2

09:30

Kick Start

DATA MANAGEMENT

11:00

BUSINESS CASES DEEP DIVE

13:00

PICTURES CREDITS

1: https://www.goodfreephotos.com By courtesy of NASA

2: https://en.wikibooks.org/wiki/Chess_Opening_Theory/1._e4/1...e5/2._Nf3/2...Nc6/3._Bb5 - Licensed under Creative Commons "Attribution – Share Alike" Unported BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/)

3: https://www.publicdomainpictures.net/it/view-image.php?image=9422&picture=cristalli

4: Modified from https://pixabay.com/illustrations/diamond-shiny-baby-wealth-wealthy-807979/

5: Modified from https://www.goodfreephotos.com/other-photos/water-droplet-macro.jpg.php

6: Modified from https://commons.wikimedia.org/wiki/File:Water_Molecule_3D_X_3.jpg

7: Modified from https://pixabay.com/illustrations/atom-symbol-characters-abstract-68866/

8: https://pxhere.com/en/photo/1159168

9: Modified from https://www.maxpixel.net/Black-Swan-Cygnus-Atratus-Water-Bird-Bird-Swan-Fly-204841

10: http://www.sismogrammi.com/laquila/archivio-drum-laquila.php

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 2

09:00

DATA ANALYSIS (Advanced)

10:45

PHASE 2 INTRODUCTION

12:45

Lunch Break

14:00

DATA MANAGEMENT

16:00

DATA FITNESS PROGRAM

DAY 2

09:00

DATA ANALYSIS (Advanced)

10:45

DATA FITNESS PROGRAM: Pectorals boosting up

Advanced analyses boost up the value of your data,
as pectorals do for your body

AGENDA

- From Regression to Classification
  - One-Dim Linear Regression
  - Two-Dim Linear Regression
  - Classification
- Several classes of Learning
- Artificial Intelligence and Machine Learning
- Training, Testing and Performance Evaluation
- Classwork

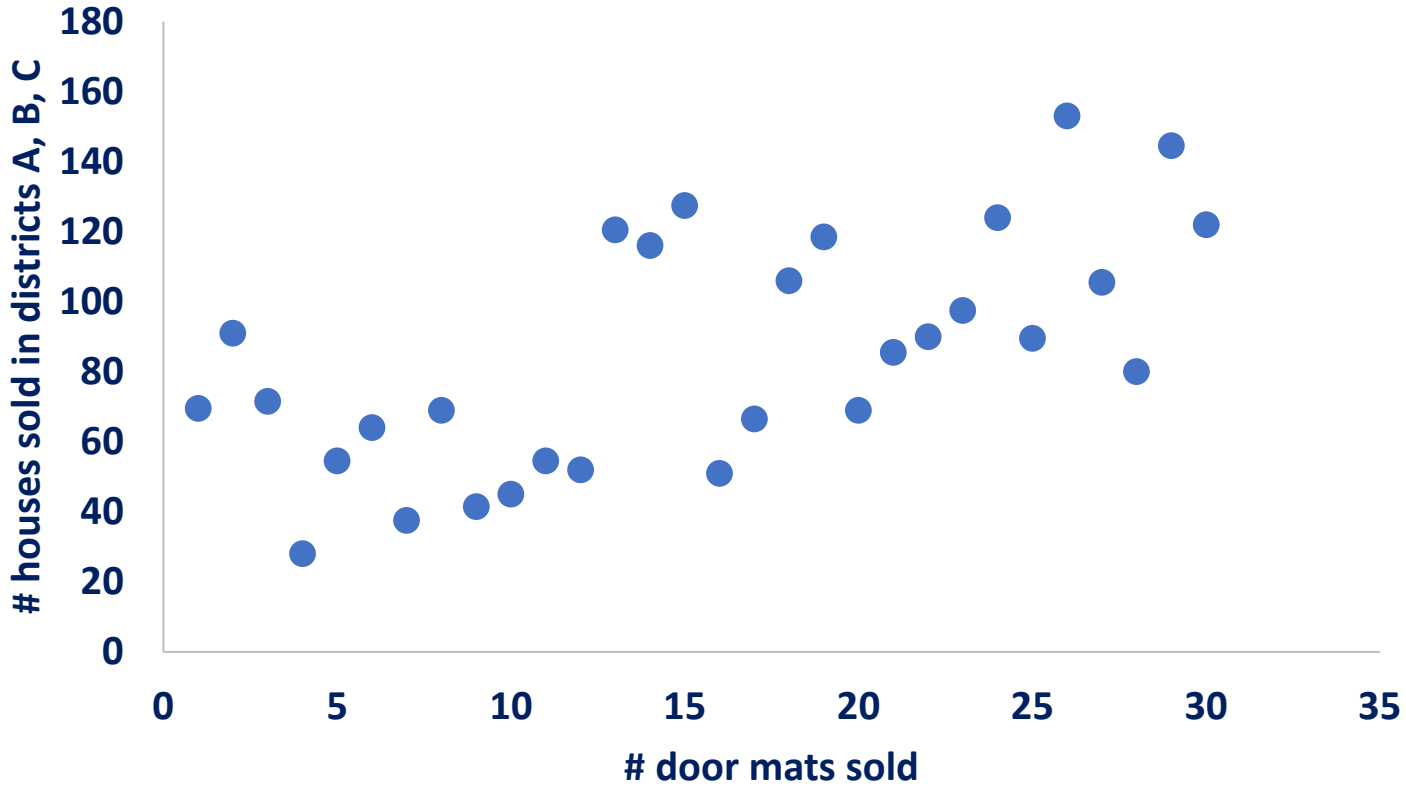
ONE-REGRESSOR PROBLEM

Try to predict how many houses will be sold in Düsseldorf downtown this month

What you know:

houses sold in districts A,B,C y

door mats sold x

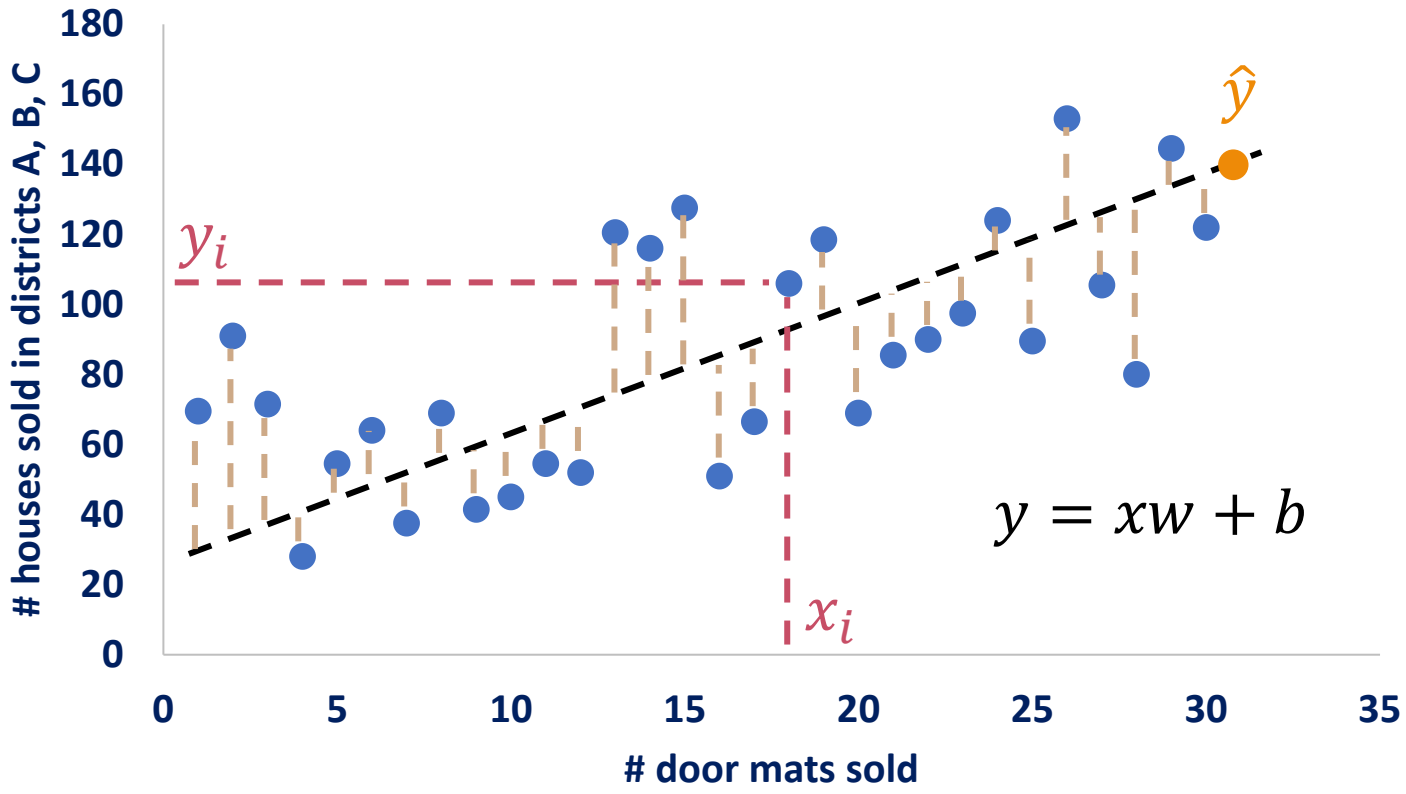


ONE-REGRESSOR PROBLEM

A linear regression seems reasonable, but what does it mean?

We ***learn*** the optimal value of $\boldsymbol{w}$ and $\boldsymbol{b}$ through the minimization of a function, e.g. Root Mean Square Error

Now we can predict $\hat{y}$, i.e. the # houses sold given the # door mats sold



TWO-REGRESSORS PROBLEM

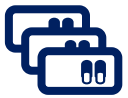
Two-Dim Linear Regression

Try to predict how many houses will be sold in Düsseldorf downtown this month

What you know:

houses sold in districts A,B,C

y

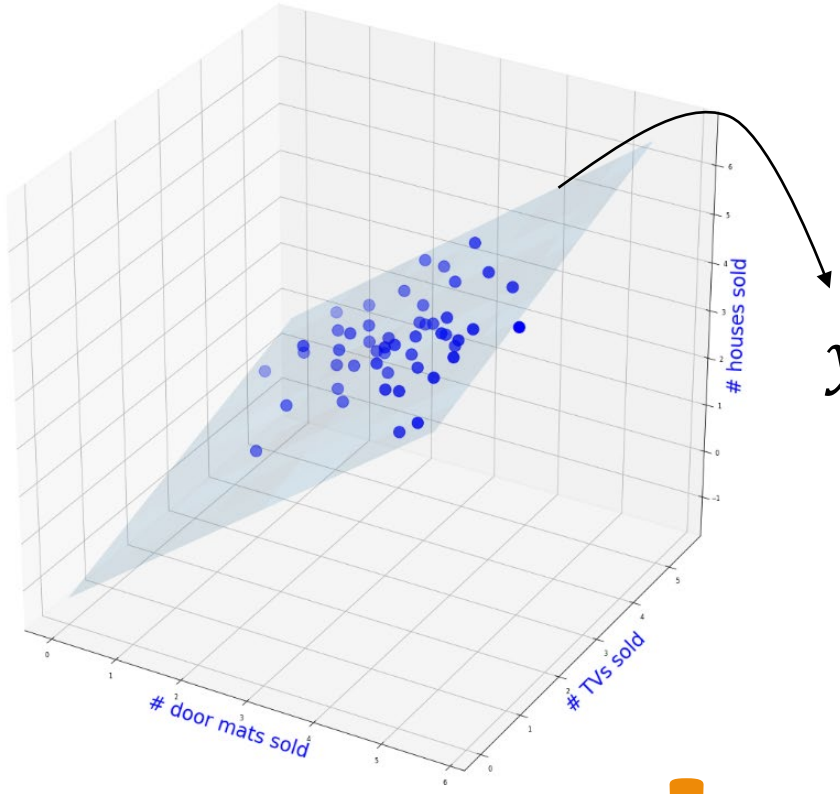


door mats sold

estate agencies

X

features



$$y = \boldsymbol{x}\boldsymbol{w} + b$$

More Information → Better Results

Not Always True

FROM REGRESSION TO CLASSIFICATION

Classification

Try to predict which employee are managers in section A of the Vodafone Campus

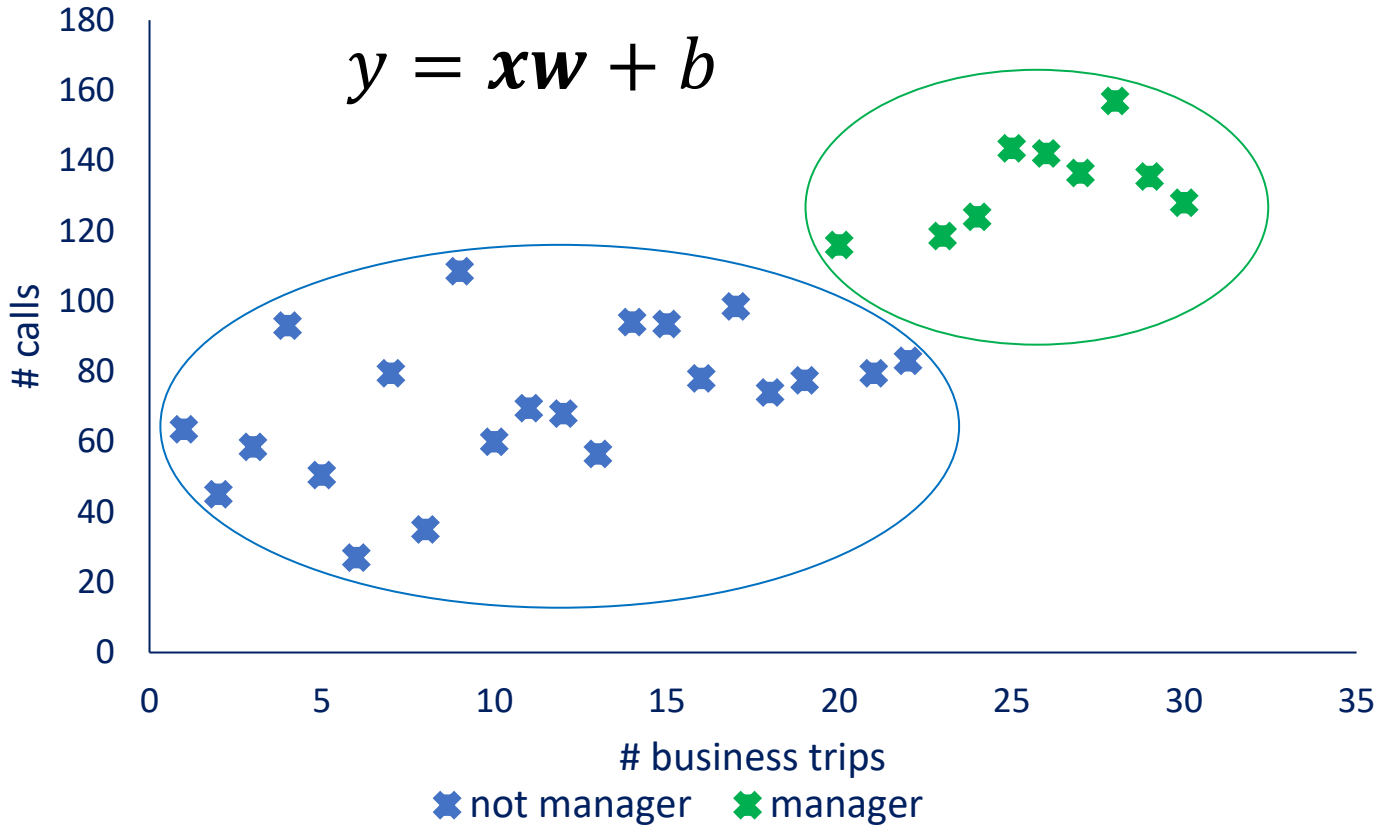
What you know:

managers in section B,C,D y

calls

business trips

X



FROM REGRESSION TO CLASSIFICATION

Classification

The model? $y = xw + b$ It keeps the same form

Linear regression ≠ Linear classification

Target Labels y

Continuous Variable

Target Labels y

Discrete Variable

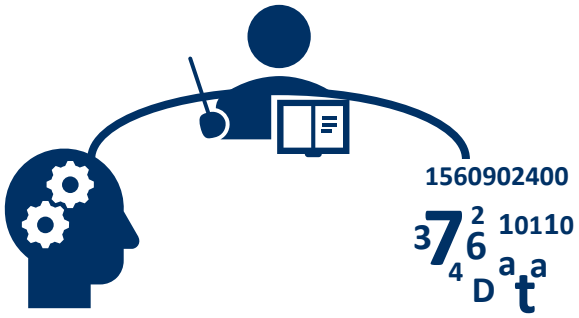
AGENDA

- From Regression to Classification
- Several classes of Learning
  - Overview
  - Supervised
  - Unsupervised
  - Reinforcement
- Artificial Intelligence and Machine Learning
- Training, Testing and Performance Evaluation
- Classwork

MANY WAYS TO LEARN

Machine learning is based on how machines *learn* and what they can *infer*

SUPERVISED LEARNING



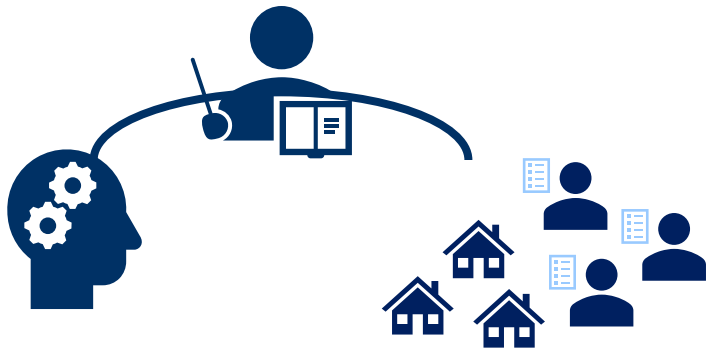
UNSUPERVISED LEARNING

REINFORCEMENT LEARNING

SUPERVISED LEARNING

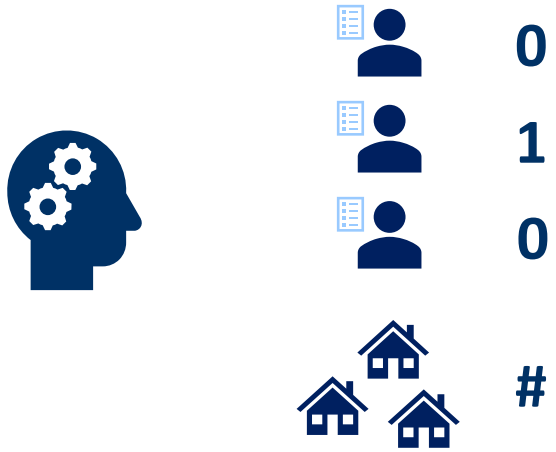
Supervised

Training Step



y_i labels known, used to train the machine

Test Step



y_i labels known, to be predicted by the machine

Accuracy Evaluation

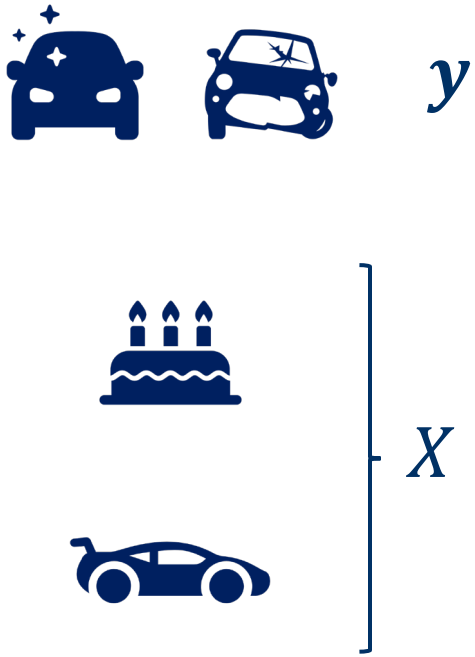
how good the performance was

EXAMPLE: PERCEPTRON

Supervised

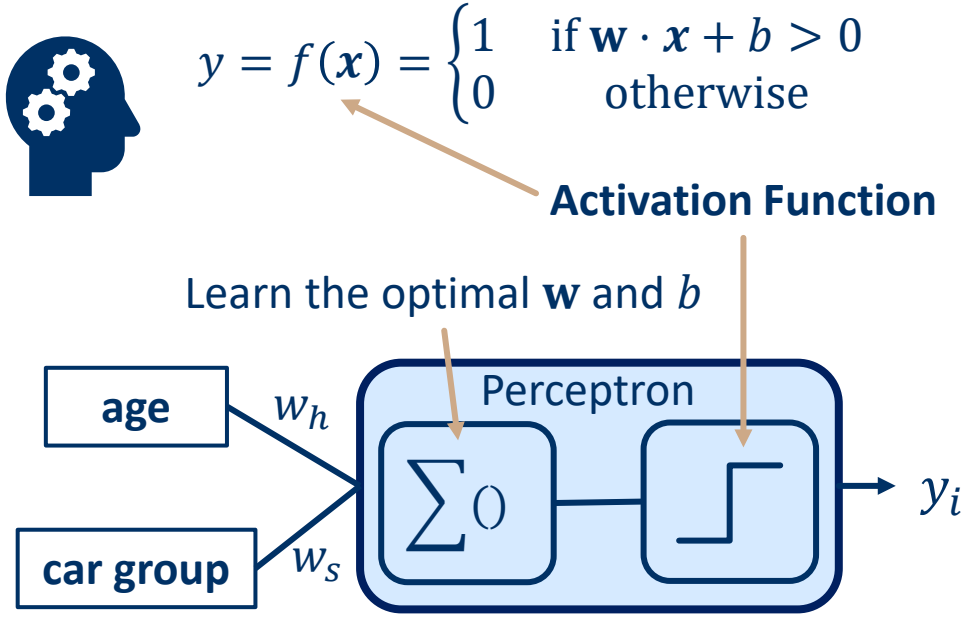
Let's try to predict the whether a driver is a good insurance risk, according to the characteristics of existing customers

What you know:

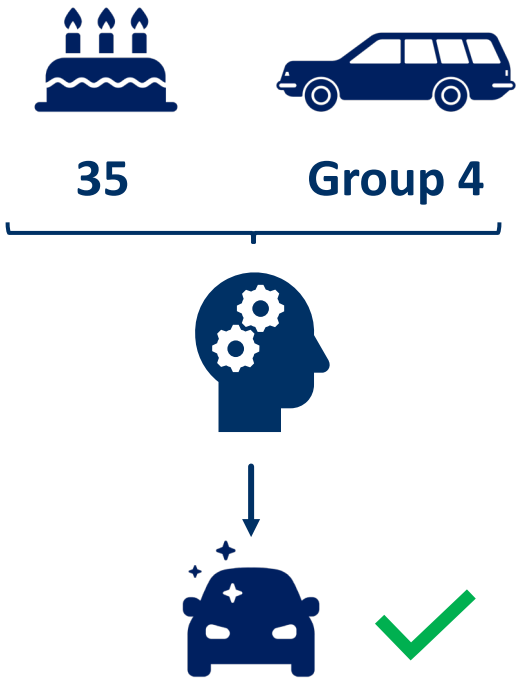


Perceptron model

BINARY CLASSIFICATION



New Observation

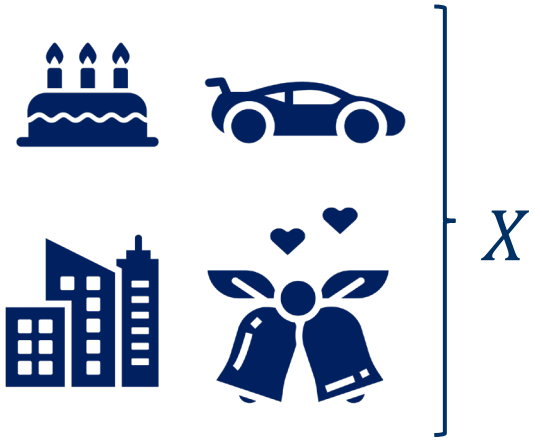


IS ONE PERCEPTRON ALWAYS ENOUGH?

Supervised

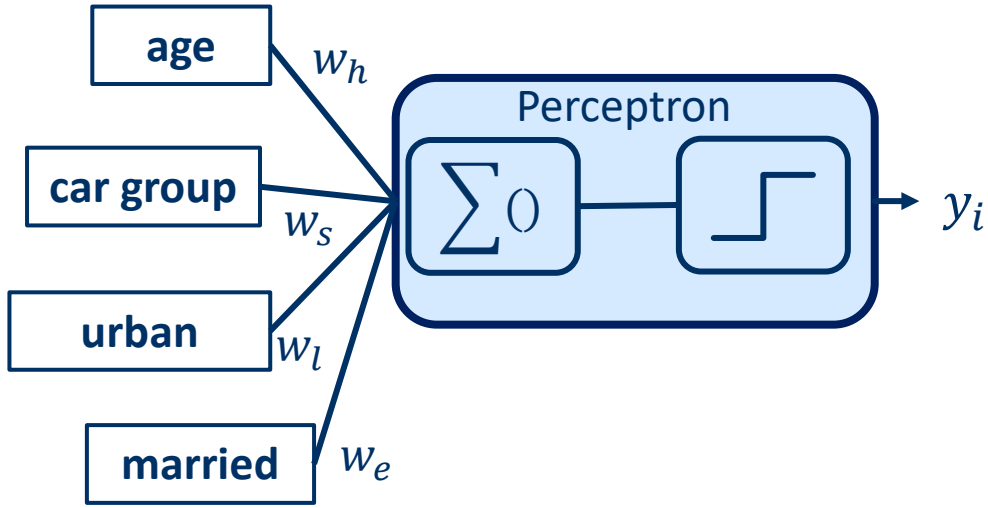
What you know:

y

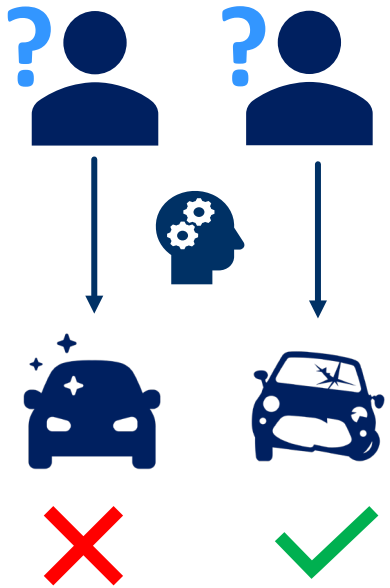


X

Perceptron model



New Observations



Accuracy: 50%

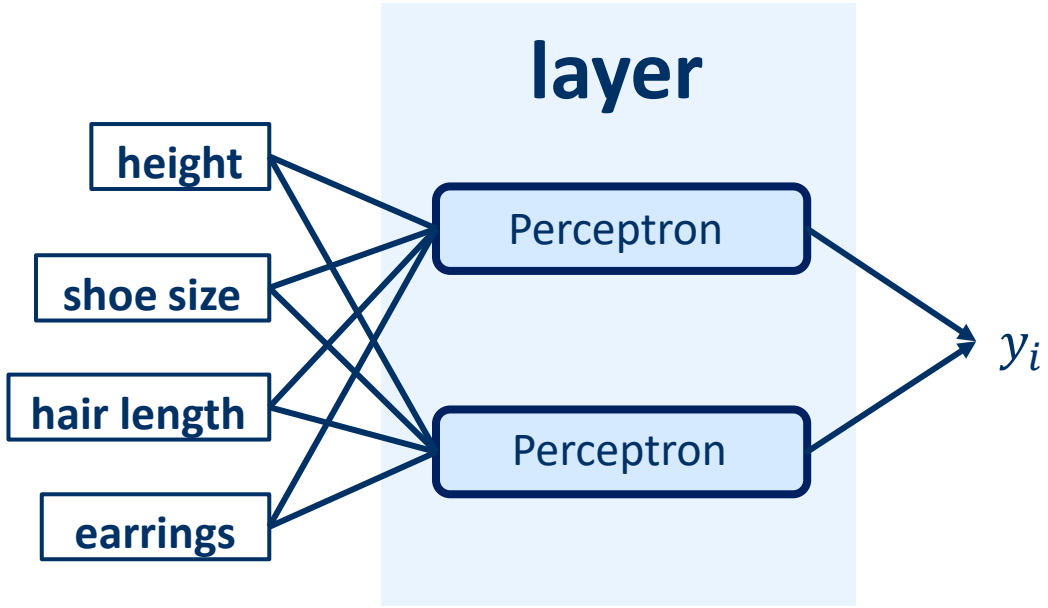
Random classification!

NEURAL NETWORK AS MULTILAYER PERCEPTRON

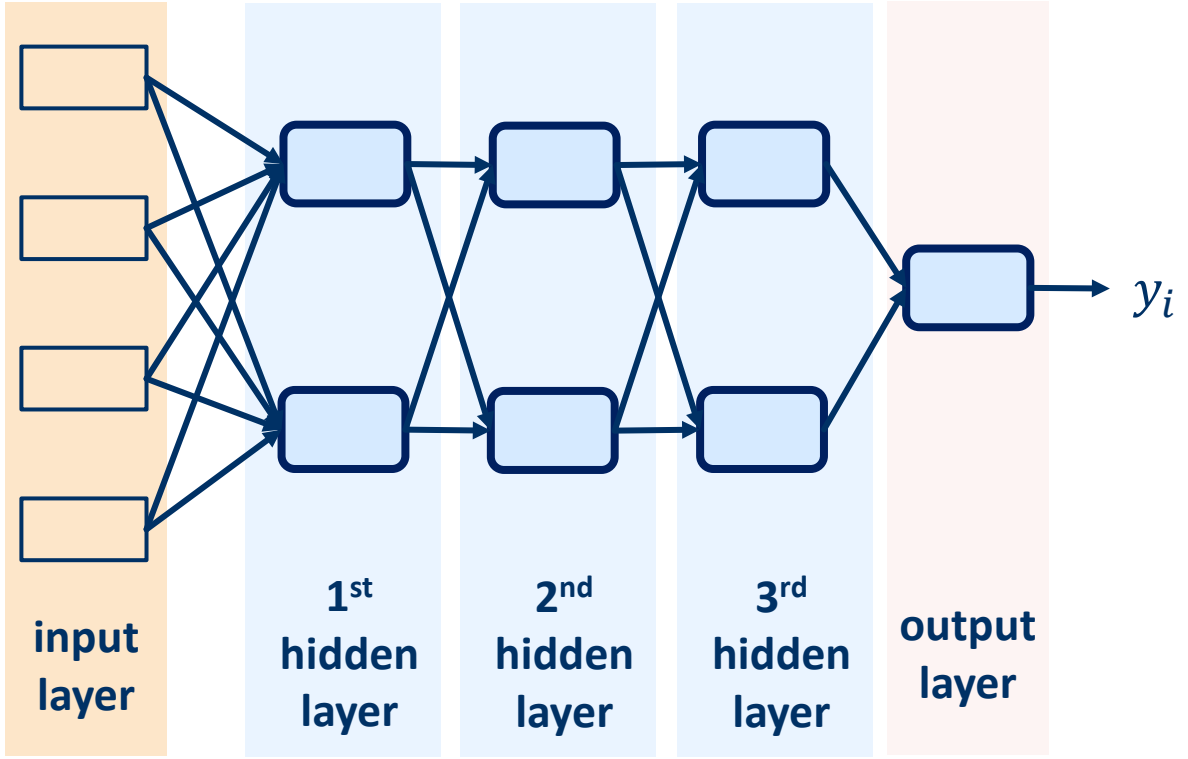
Supervised

Hidden information can be caught by increasing the number of perceptrons

stacking them in layers



increasing the number of hidden layers

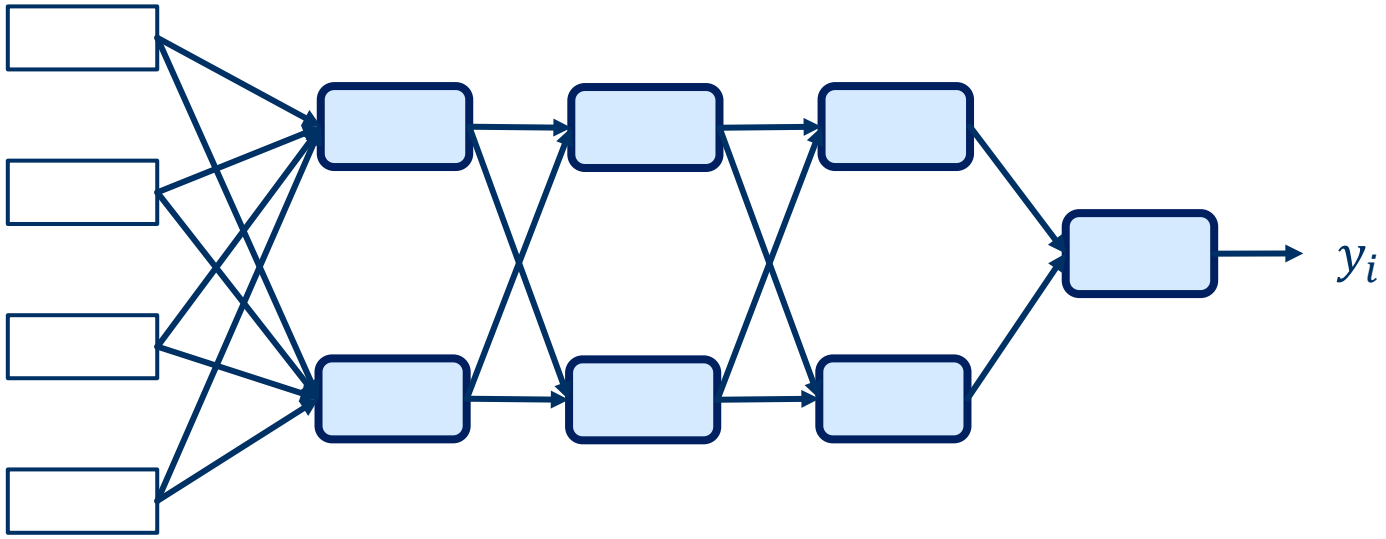


MULTILAYER PERCEPTRON

Supervised

FORWARD PASS

estimate the output as some function of the input, but the distribution of the data is unknown or hidden



BACKWARD PASS

learn the weight of each connection through an iterative process called **backpropagation**

UNSUPERVISED LEARNING

Unsupervised

$x_1 = [2, 5]$

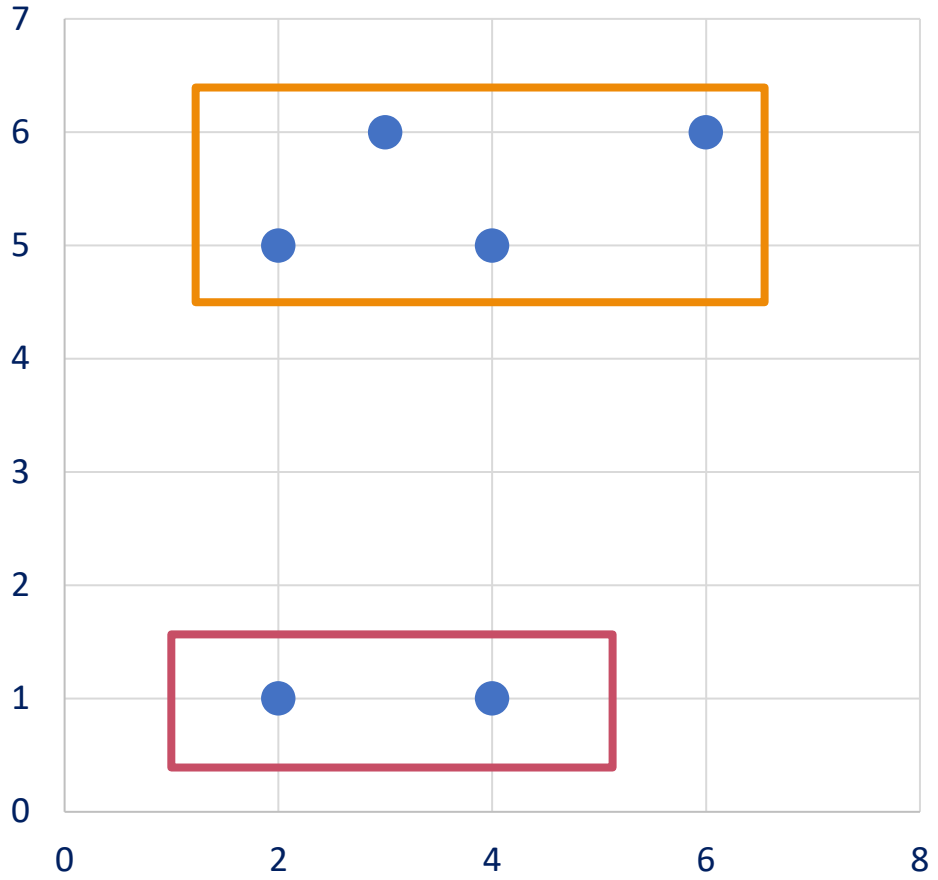
$x_2 = [4,5]$

$x_3 = [2,1]$

$x_4 = [6,6]$

$x_5 = [4,1]$

$x_6 = [3,6]$



x_i labels are unknown

Algorithms of this type help finding patterns or groups in the data, using a measure of distance as a score to determine closest observation

EXAMPLE: CLUSTERING

Unsupervised

Let’s divide the people in this room into four groups (*clusters*). Each cluster could represent the income, the business strategic unit or even the nationality

K-MEANS ALGORITHM

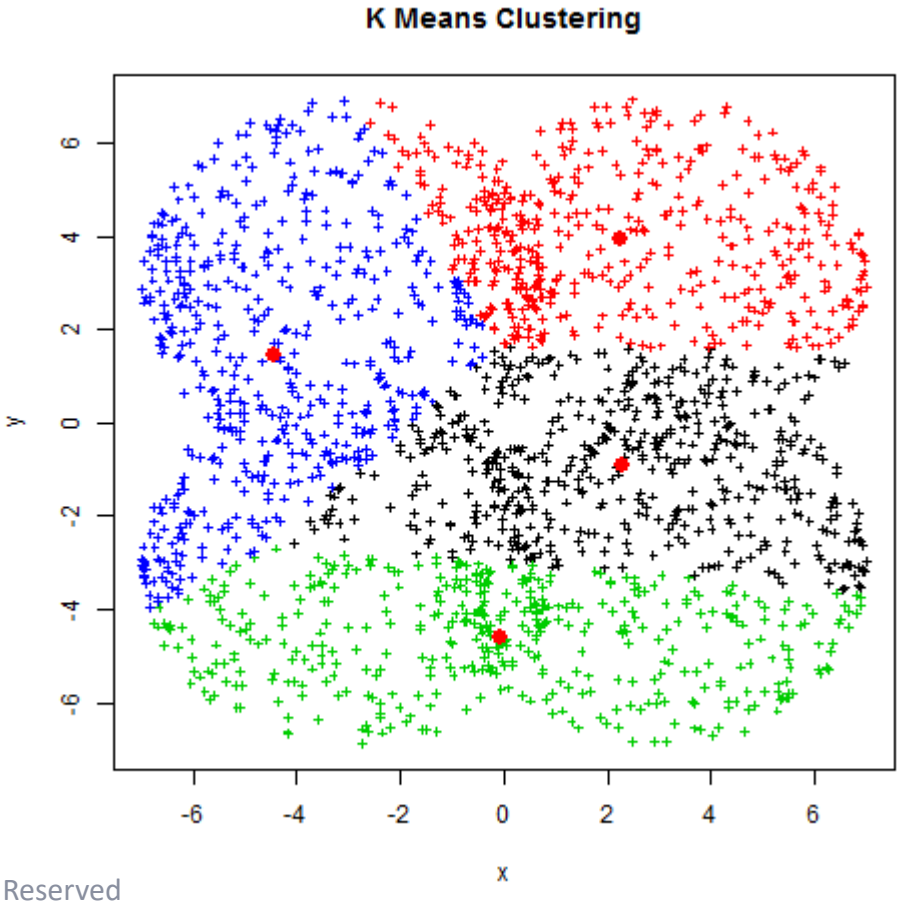
Choose k centroids

SCORE Compute data-centroid distances

Assign each data to the cluster corresponding to the closest centroid

Compute new centroids based on these new clusters

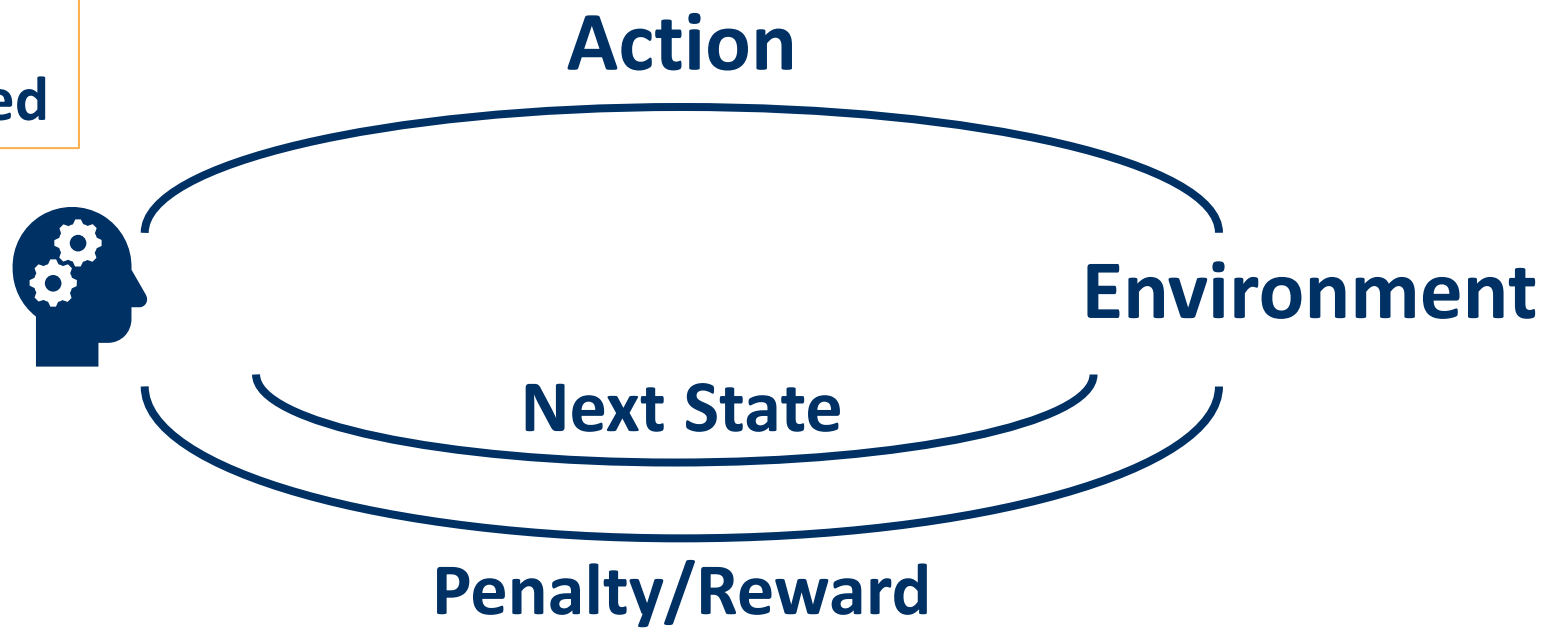
Iterate till centroids cease to vary

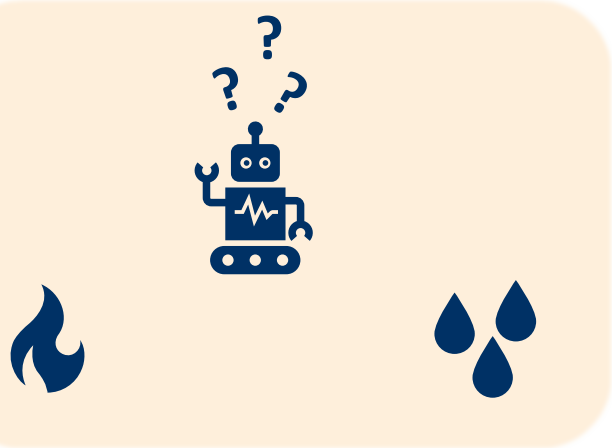


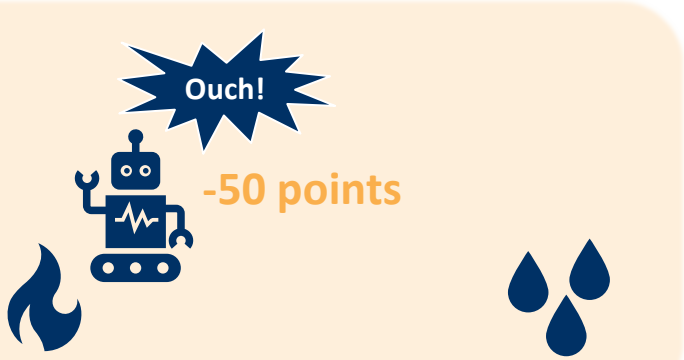
REINFORCEMENT LEARNING

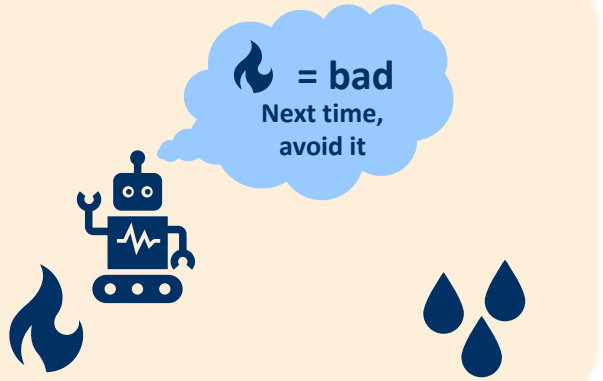
Reinforcement

Function to be optimized





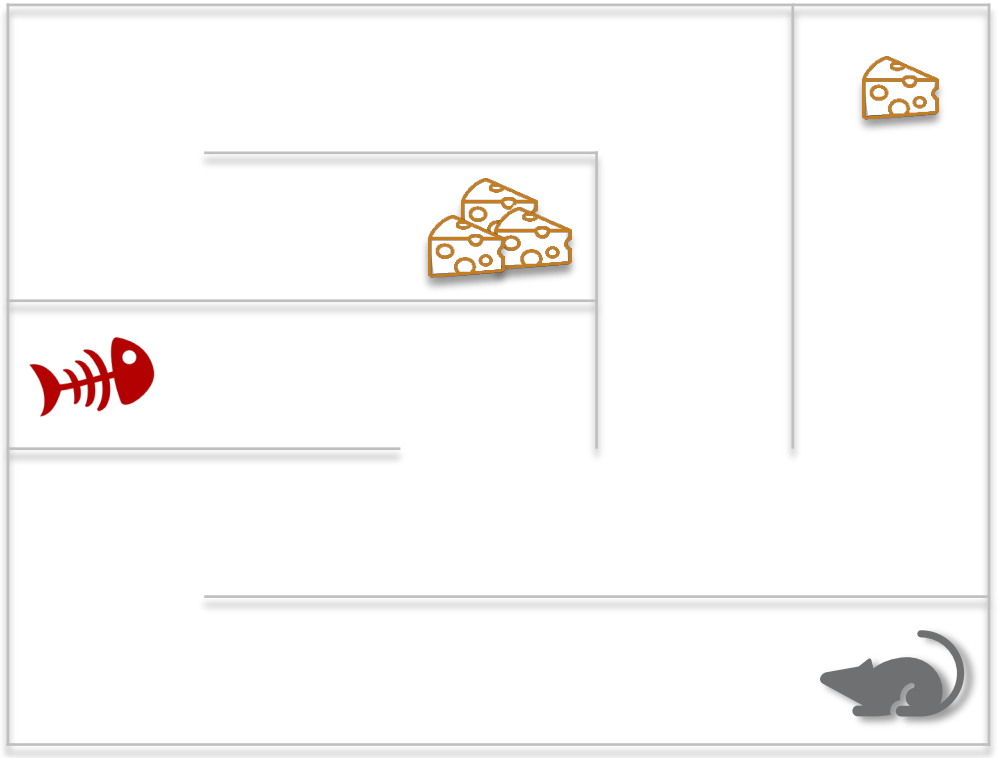




ALGORITHM EXPLAINED

Reinforcement

The objective is to maximize a *cumulative reward* thanks to a set of actions to be performed in a certain environment. Sets of actions are typically modeled as a Markov Decision Process (MDP)



The aim is to find the best policy for performance maximization, which includes:

Exploration and

Exploitation

AGENDA

- From Regression to Classification
- Several classes of Learning
- **Artificial Intelligence and Machine Learning**
  - Intelligence
  - AI vs ML
- Training, Testing and Performance Evaluation
- Classwork

INTELLIGENCE: A THEORETICAL DEFINITION

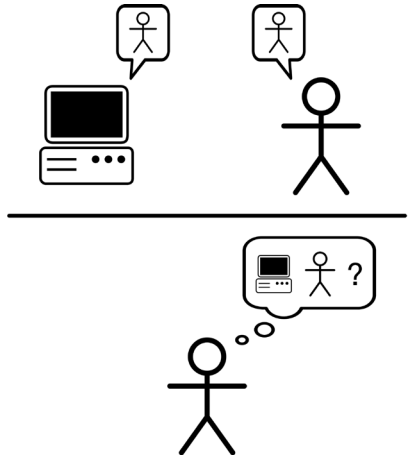
Someone/something who/which has the ability to *think*

Can machines *think?*

Turing "Computing Machinery and Intelligence", 1950

and

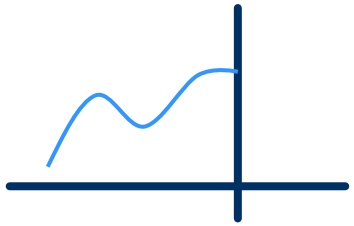
The Imitation game



INTELLIGENCE: A PRACTICAL DEFINITION

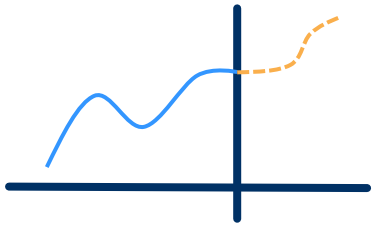
Intelligence

INTELLIGENCE as:



KNOWLEDGE ACQUISITION

learning rules which explain observations

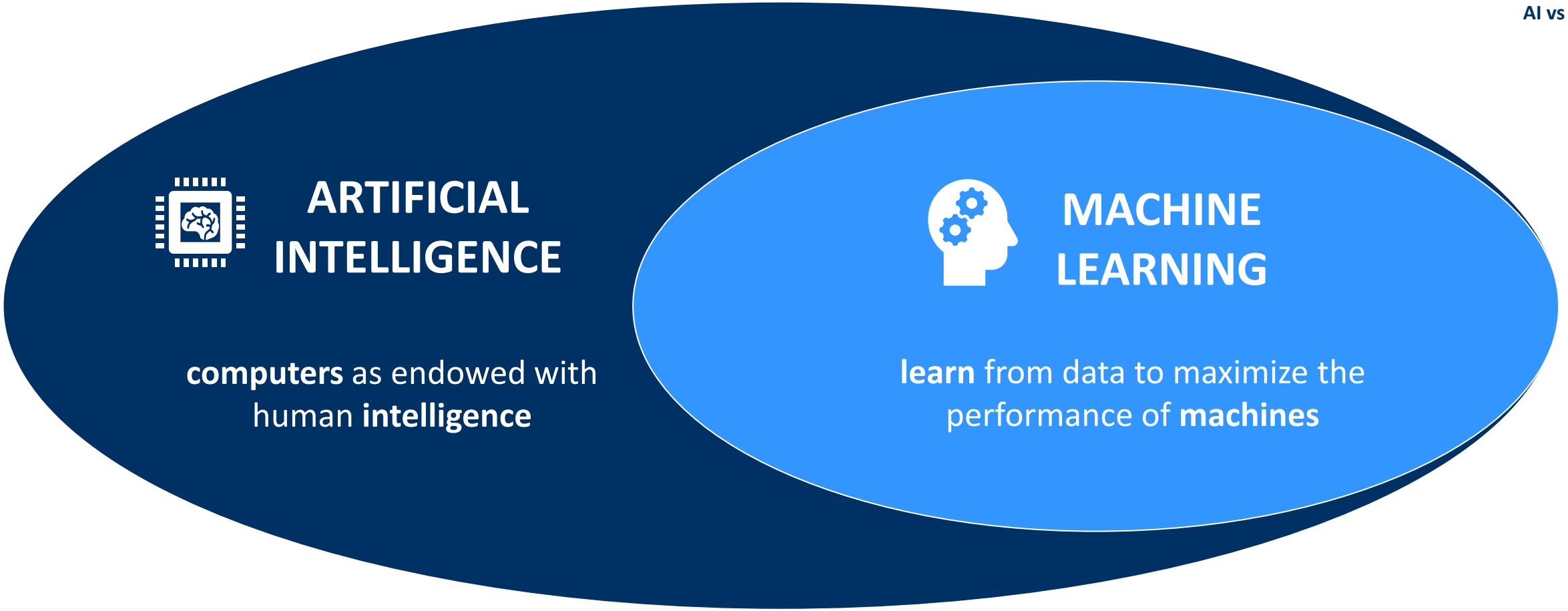


INFERENCE

deriving the truth by the acquired knowledge

DEFINING BOUNDARIES

AI vs ML



broad purpose:
DECISION MAKING

limited purpose:
DISCOVER HIDDEN RELATIONSHIP, PROCESS AUTOMATION

ANDROIDS

AI ≠ ANDROIDS

Not all AI are androids

Not all androids have AI

Otherwise:

Airplanes are motorized birds

Cars are motorized horses

Submarines are motorized whales

AGENDA

- From Regression to Classification
- Several classes of Learning
- Artificial Intelligence and Machine Learning
- **Training, Testing and Performance Evaluation**
  - Validation
  - Underfitting & Overfitting
  - Bias, Variance and Learning Curves
  - Performance Evaluation
- Classwork

THE IMPORTANCE OF TRAINING AND TEST SETS

Unbiased Split of the Full Dataset

Parameters Estimation

Parameters Validation

Model Validation

80%

20%

Rule of thumb

Cross-Validation

Training	Training	Training	Validation	Test
Training	Test	Validation	Training	Training
Training	Training	Training	Test	Validation
Validation	Training	Test	Training	Training
Test	Validation	Training	Training	Training

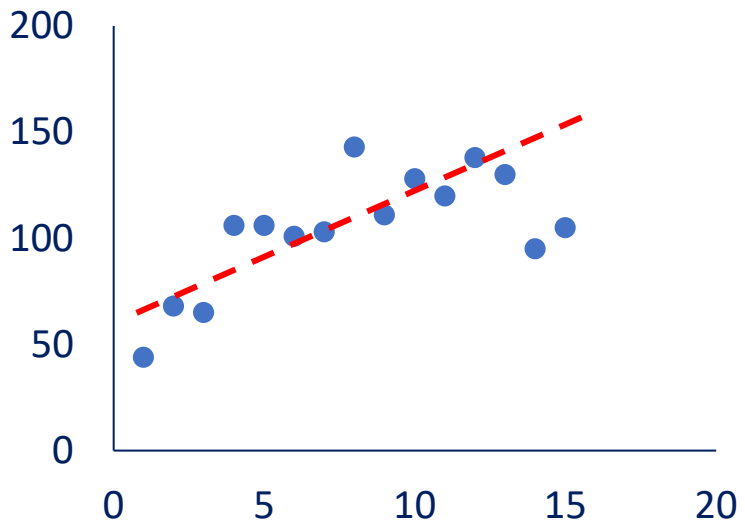
Test
Test
Test
Test
Test

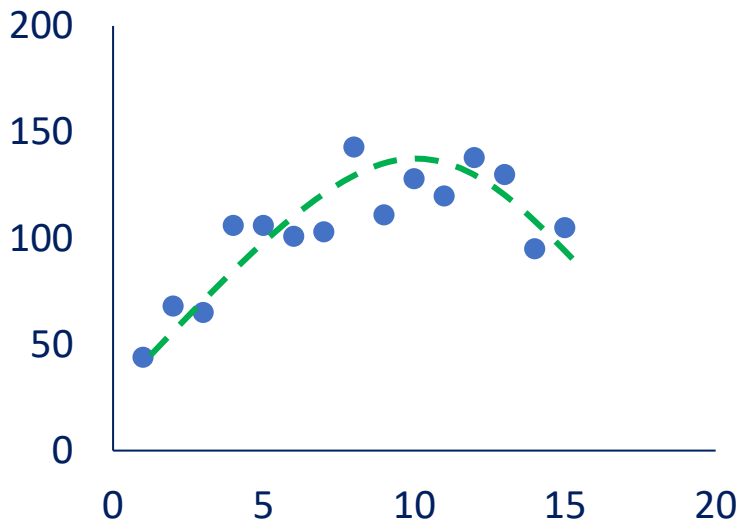
Prediction Statistics

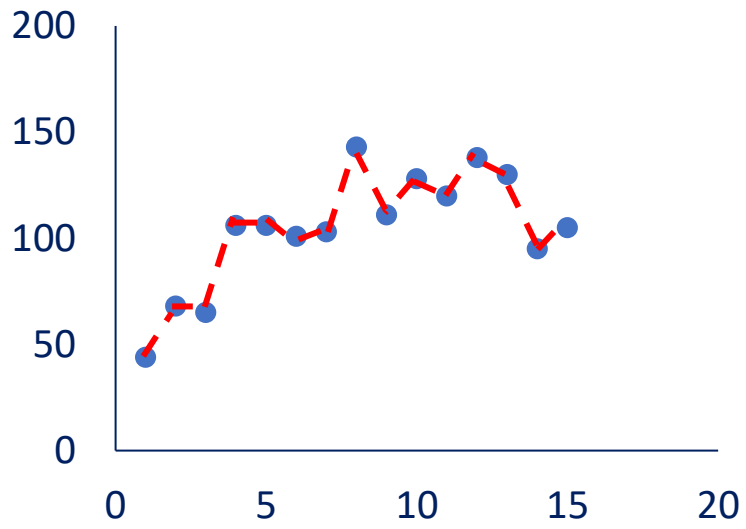
THE CURSE OF MODEL COMPLEXITY

Underfitting
& Overfitting

MODEL COMPLEXITY







UNDERFITTING

OPTIMIZED MODEL

OVERFITTING

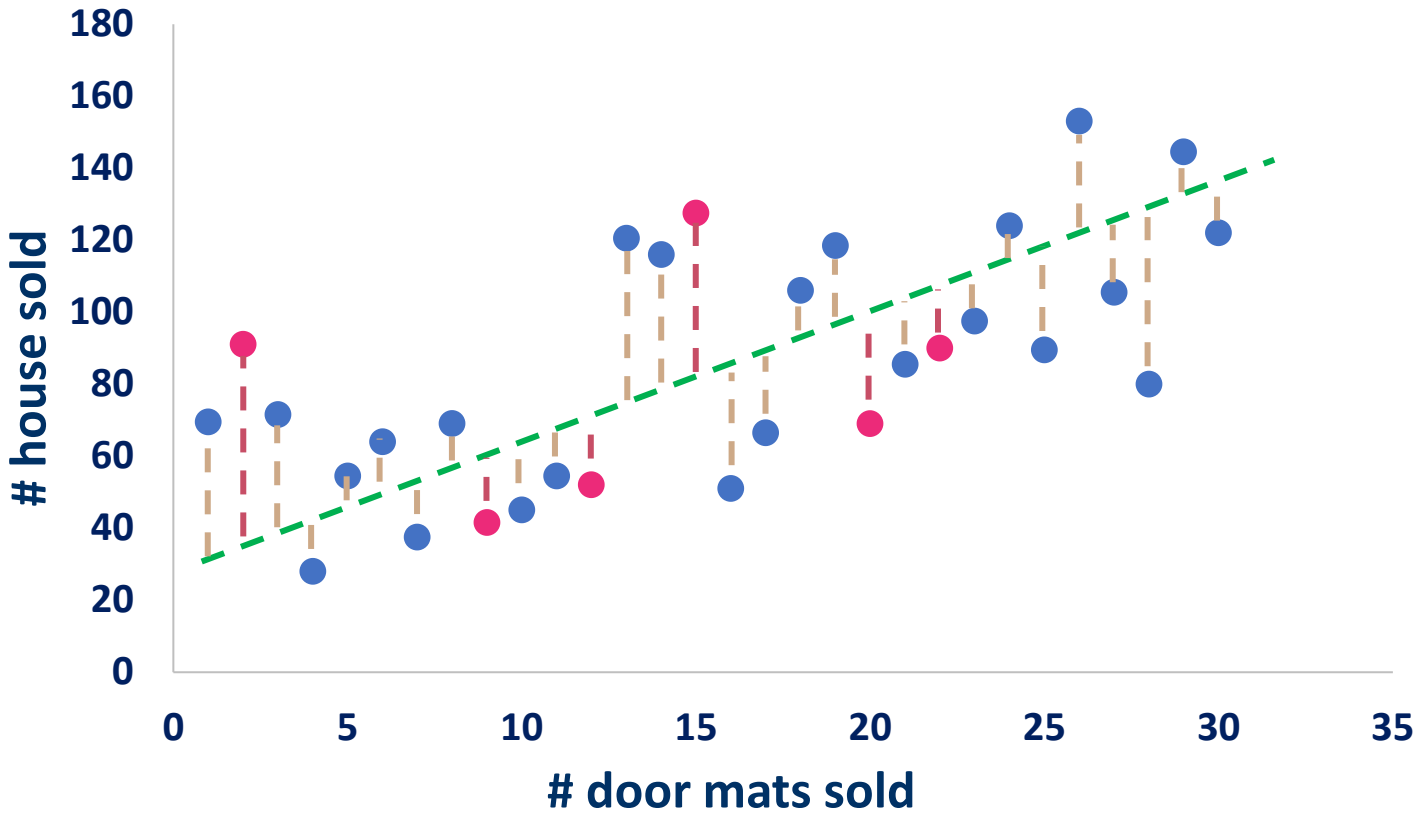
ERRORS: IN AND OUT OF SAMPLE

IN-SAMPLE ERROR

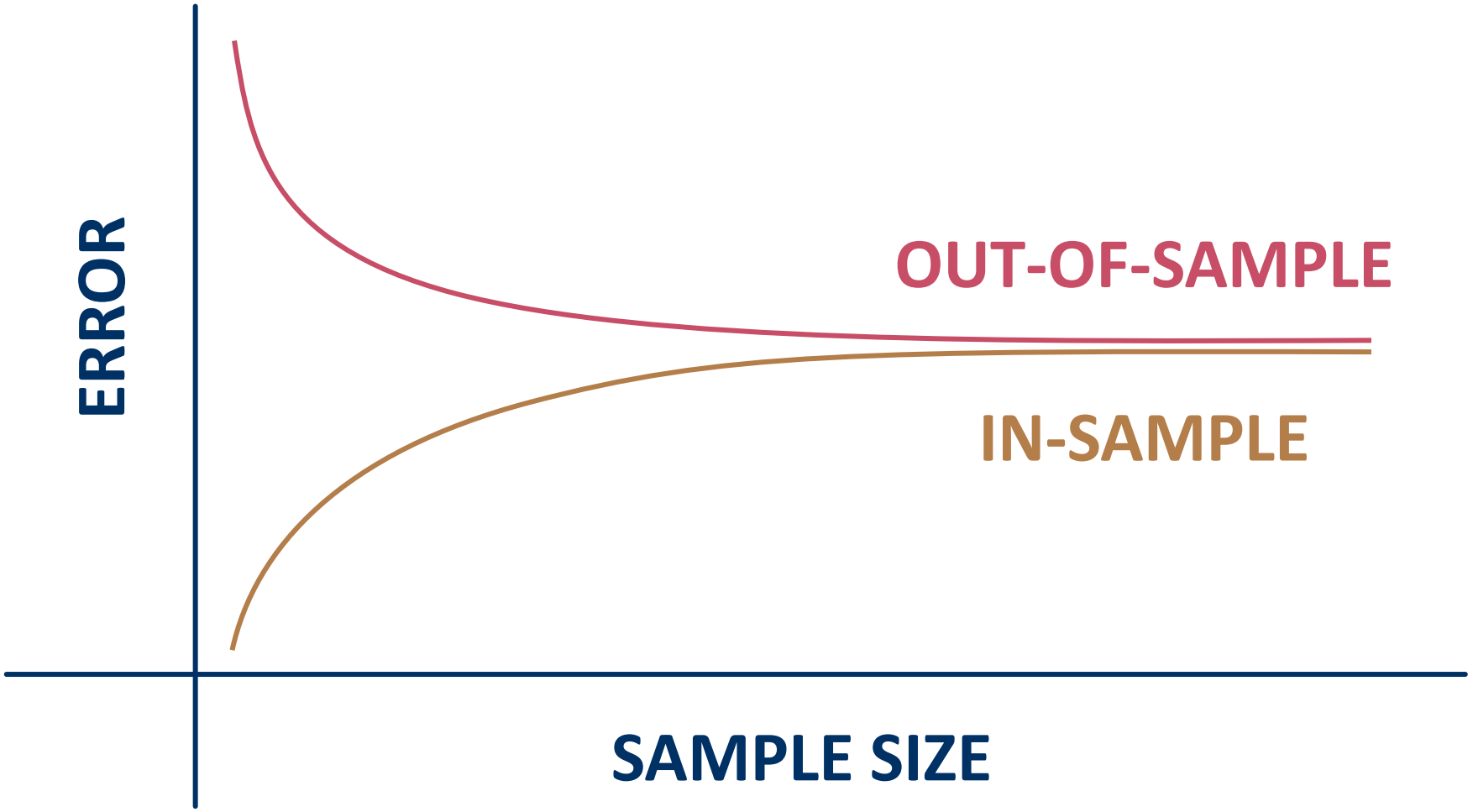
Error to be minimized in order to get the best model according to the dataset

OUT-OF-SAMPLE ERROR

Allows to evaluate the accuracy of the model in predicting outcomes unseen before



THE LEARNING CURVES



Generalization aims to minimize the gap between the two (curves) errors

GENERALIZATION ENEMIES – Part I

Bias, Variance & Learning Curves

1950 **Survey on life satisfaction**

90% satisfied people

Non representative information owing to:

BIAS

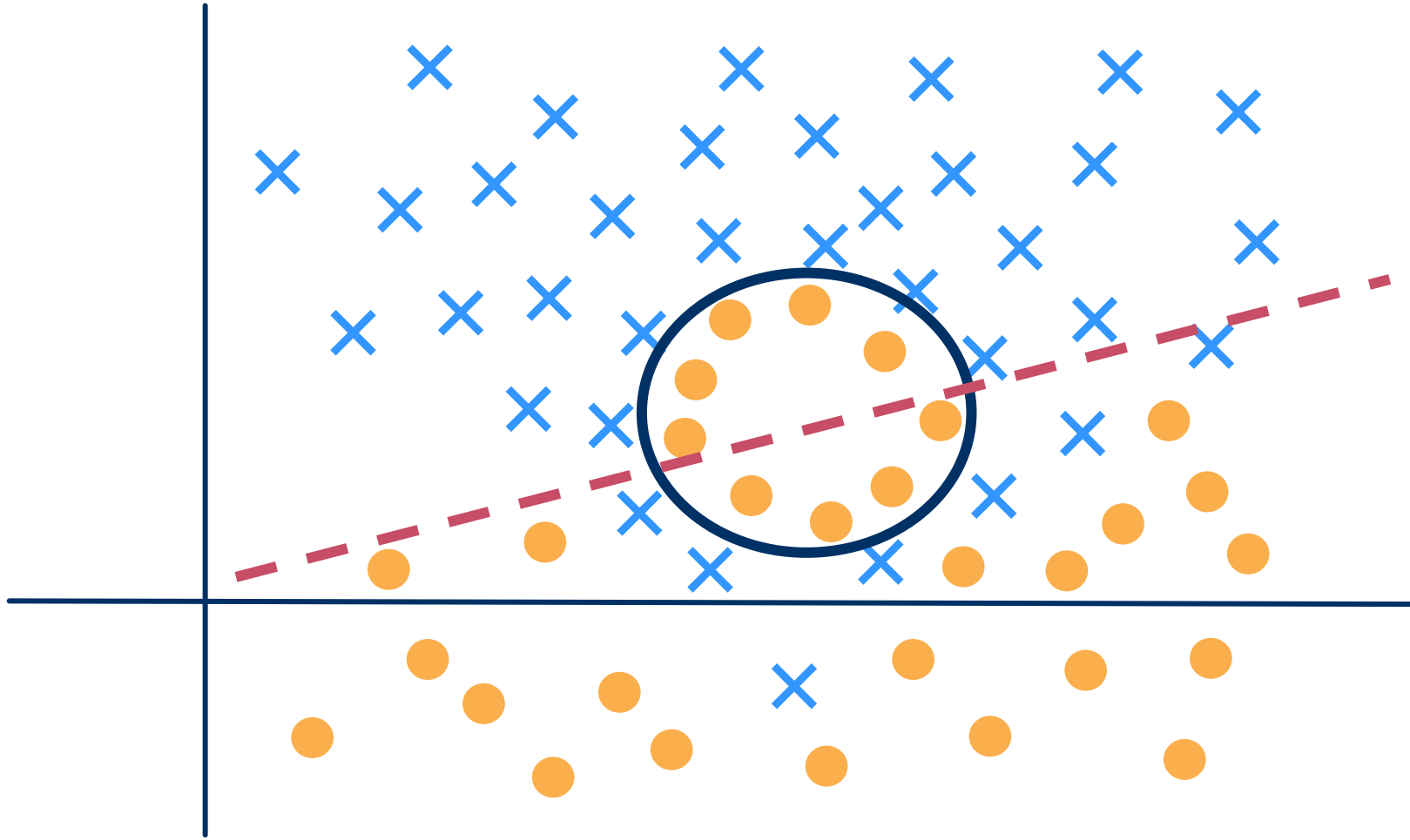
All reached people had a phone

VARIANCE

Non-gaussian distribution of telephones on the territory

GENERALIZATION ENEMIES – Part II

DATA SNOOPING

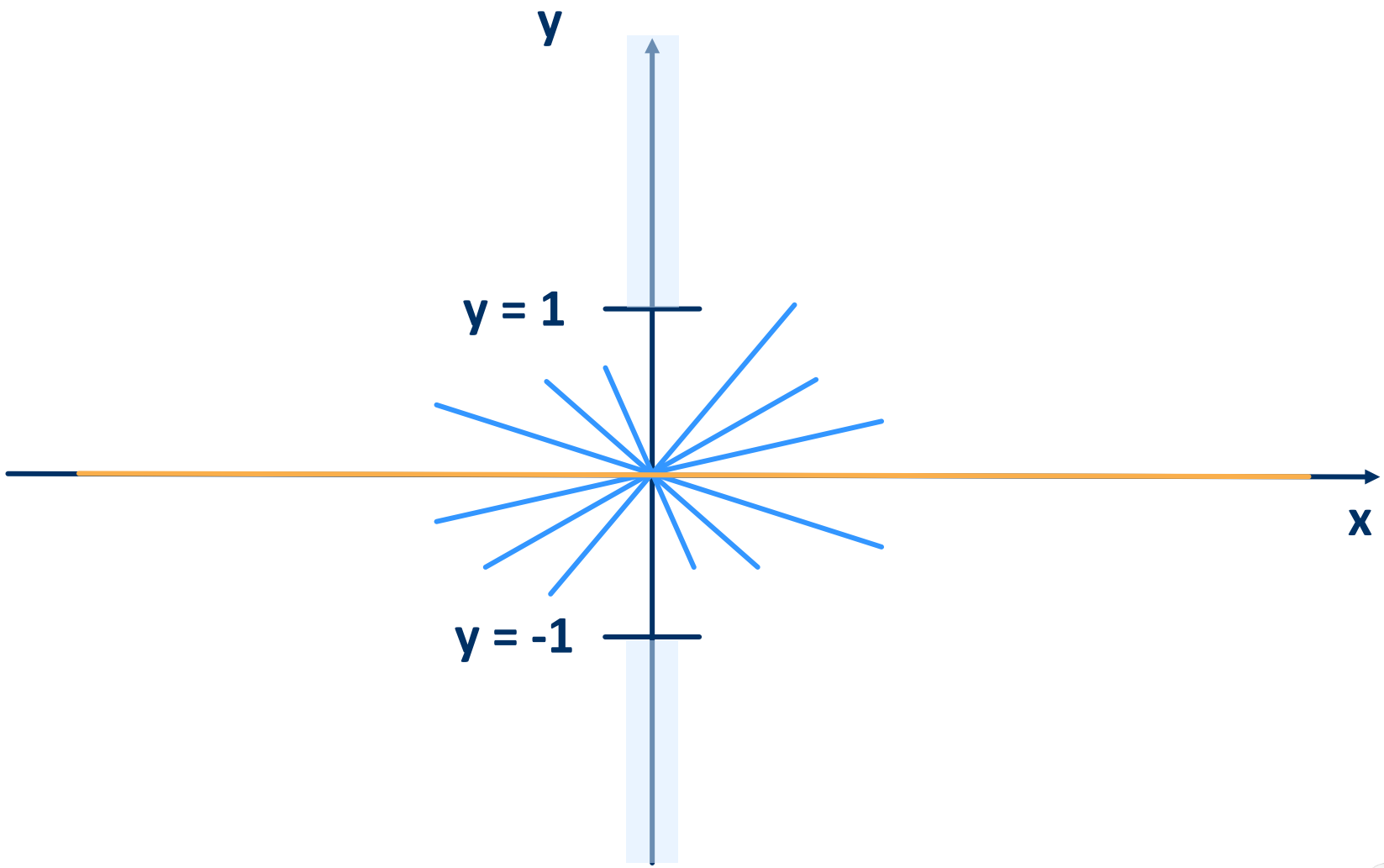


Human Judgment

Modelling Procedure

REGULARIZATION

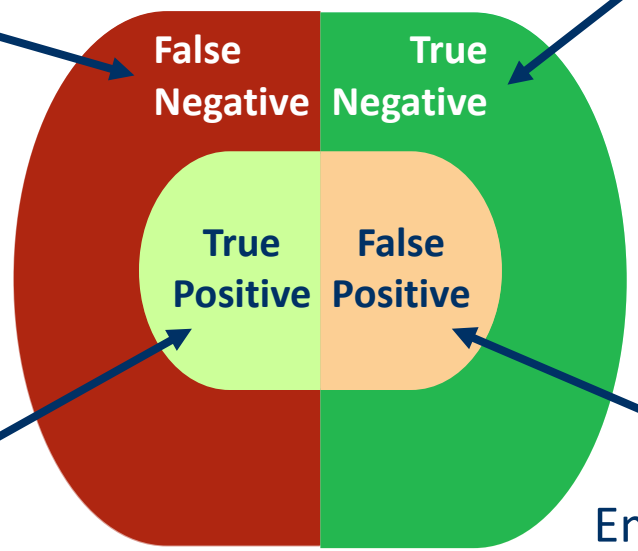
Regularization rewards model's simplicity



CLASSIFICATION MODELS

Performance Evaluation

Recall the manager classification problem. How can we evaluate the result?



Employee predicted **not manager** but actually **manager**

Employee predicted **not manager** and actually **not manager**

Employee predicted **manager** and actually **manager**

Employee predicted **manager** but actually **not manager**

PRECISION

How many predicted manager are truly managers?

Precision = $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

RECALL

How many actual manger are predicted correctly by the model?

Recall = $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

REGRESSION MODELS

Performance Evaluation

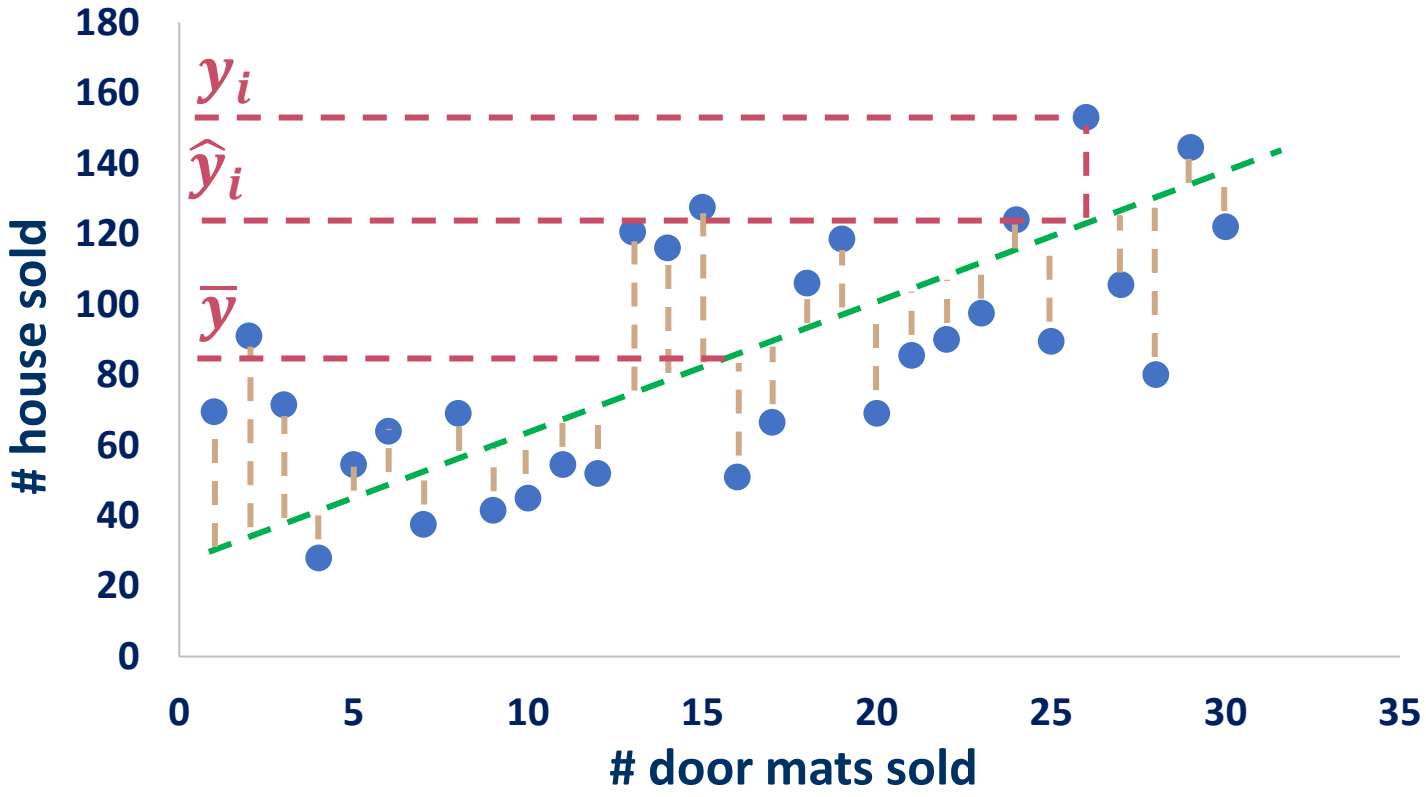
Recall the house sold in Düsseldorf problem. How can we evaluate the result?

Minimize the **Root Mean Square Error (RMSE)**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Goodness of fit measure: **Coefficient of Determination**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$



AGENDA

From Regression to Classification

Several classes of Learning

Artificial Intelligence and Machine Learning

Training, Testing and Performance Evaluation

Classwork

CLASSWORK

Fitness Gym Company

Assignment

In this scenario you are the Head of Innovation of the company, presenting to the investors' board possible ways to improve business leveraging available data using advanced analysis techniques. Here is your data:

Revenues

Maintenance Expenses (per gym and section)

Gym equipment

Number of customers (per gym and section)

Staff Scheduling

# of accesses per customer

Staff Salaries

Opening hours

All information REASONABLY available to the company

GIVE YOUR BEST SHOT!

15 minutes

CHURN PREVENTION

Possible Output

OBJECTIVE

Predict whether a customer is going to abandon the gym in next N days.
A churn is defined here as (for example) a customer not visiting gym for at least 1 month

DATA

Visits (e.g. enrich the dataset with delta visits in last K days), subscription details (price, access limitations in terms of times and services, followed courses), personal trainer(s) following them

APPROACH

A supervised learning algorithm that can predict the probability to abandon the gym according to historical data and related information on customers.
A classifier (e.g. Decision Tree) can fit this need

PRESCRIPTIVE ACTIONS

Subscription discounts
Personal Trainer change
Access to more tools

CUSTOMER SEGMENTATION

Possible Output

OBJECTIVE

Identify groups of gym customers that are “similar” according to some parameter, to identify cluster of potential customers to address marketing activities to them and invite them to subscribe to your gym

DATA

Personal data (geo, work, age, marital status..),
Interactions with gym (Subscriptions, accesses trend, attended “open day” and courses..)

APPROACH

An unsupervised learning algorithm able to minimize a distance among people and create groups.
(e.g. hierarchical) clustering
is a good option in this case

PRESCRIPTIVE ACTIONS

Dedicated marketing campaigns
Gym sections or tools addition

ACQUISITION MODELING

Possible Output

OBJECTIVE

Identify the most relevant levers that can help acquiring more new customers, to replicate and improve the same actions in the future.
Variables to model could be
number of new customers or revenues

DATA

Expenses on maintenance activities (as an indicator of tools quality), Investments in marketing activities (flyers, SEM, OOH), Hired staff, Discount applied, trend and seasonality

APPROACH

A multivariate regression analysis, a supervised algorithm technique able to estimate the contribution to the variable to predict of all implemented actions

PRESCRIPTIVE ACTIONS

Leverage more effective operations and cut less impacting ones
Build a strategy for prospects acquisition

KNOWLEDGE EXTRACTION

Possible Output

OBJECTIVE

Address specific prospects that are (more or less) clearly expressing discontent for their current gym subscription or the will to buy a completely new one

DATA

Public data on Social Networks, accessible through APIs, of people within gyms customer target (e.g. correct age and proximity):
personal data, posts, interests, liked page/profiles

APPROACH

Crawl Social Network data, identify topics from posts/tweets, categorize on the basis of interests, select interesting people

PRESCRIPTIVE ACTIONS

Direct contacts with the most promising people in terms of interest to join our gyms, through dedicated offers/discounts

15 minutes Break

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 1

Kick Start

BUSINESS CASES PRESENTATION

DATA VISUALIZATION

Lunch Break

DATA ANALYSIS (Basic)

DAY 2

Kick Start

DATA MANAGEMENT

11:00

BUSINESS CASES DEEP DIVE

13:00

DATA FITNESS PROGRAM: Working on explosive strength

Data science applications to real world business cases can be compared to racked strength transformed into explosive one in fitness

AGENDA

Business Cases
Deep Dive

Use Cases: Hands On

A look on other BCs by Moviri

CUSTOMER SATISFACTION AND CHURN PREDICTION

Telco Company Case

CHALLENGE

In Telco market, customers' loyalty is very fragile: because of new tariffs, abolished exit fees, competition in terms of QoS and pricing leads to frequent telco supplier change by final customer

OBJECTIVE

Client aim is to reduce the amount of end customers abandoning the service

DATA

Personal data
(gender, age, location…)

Service utilization
(bandwidth used, contact to customer service, service rate, …)

Subscription details
(type of contract, billing information, technical limitations...)

CUSTOMER SATISFACTION AND CHURN PREDICTION

Telco Company Case

APPROACH AND IMPLEMENTATION

Exploration and analysis of personal data
Quality of Service analysis and cross-correlation
Data quality enhancement and monitoring
Classification model (Gradient boosted trees):study, setup and fine tuning

Technology

Splunk (Machine learning toolkit + Custom algorithms)

Other Project details

4 months span

C2 General

CUSTOMER SATISFACTION AND CHURN PREDICTION

Telco Company Case

BUSINESS BENEFITS

Mapping of dynamics between QoS and churn, with identification of most impacting parameters

Prescriptive model currently under study to maximize retention
(Customer service contact, dedicated promotion, service enhancement, facilitated payments)

Examples of impacting parameters:
Frequency of «business» calls to Customer Service, # of problems to the network, Service price..

IMPACTS (by department)

C-LEVEL
High level company strategy to allocate improvements budget

MARKETING
Optimization of campaigns and promotions: target, media, content..

CUSTOMER SERVICE
Enhanced customer care through single person tailored NBA system

C2 General

CUSTOMER SATISFACTION AND CHURN PREDICTION

CIoT example

CHALLENGE

In the new CIOT market, customers' loyalty is impacted by how well products fulfill user's needs and expectations: product quality, pricing, complexity and flexibility

OBJECTIVE

The aim is to reduce the amount of customers abandoning the service after 2 months, by understanding the key features describing churners and non churners

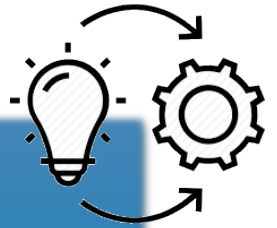
DATA

Device Usage
(data sessions, length of data sessions, frequency…)

Subscription details
(customer type, products subscribed to, billing fees, errors when provisioning)

CUSTOMER SATISFACTION AND CHURN PREDICTION

CIoT example



APPROACH AND IMPLEMENTATION

Correlation of subscription data with usage data (excluding V-Home, V-Auto, Lexus)
Aggregation on Customer level
Classification model (C50 tree):study, setup and fine tuning
Dealing with unbalanced data

Technology
Datameer and R

Other Project details
Ongoing (preliminary results)

C2 General

CUSTOMER SATISFACTION AND CHURN PREDICTION

CIoT example

BUSINESS BENEFITS

Identifying key features:
Data usage (first 14 days)
Average amount provisioned per transaction
Transactions with zero value

IMPACTS (by department)

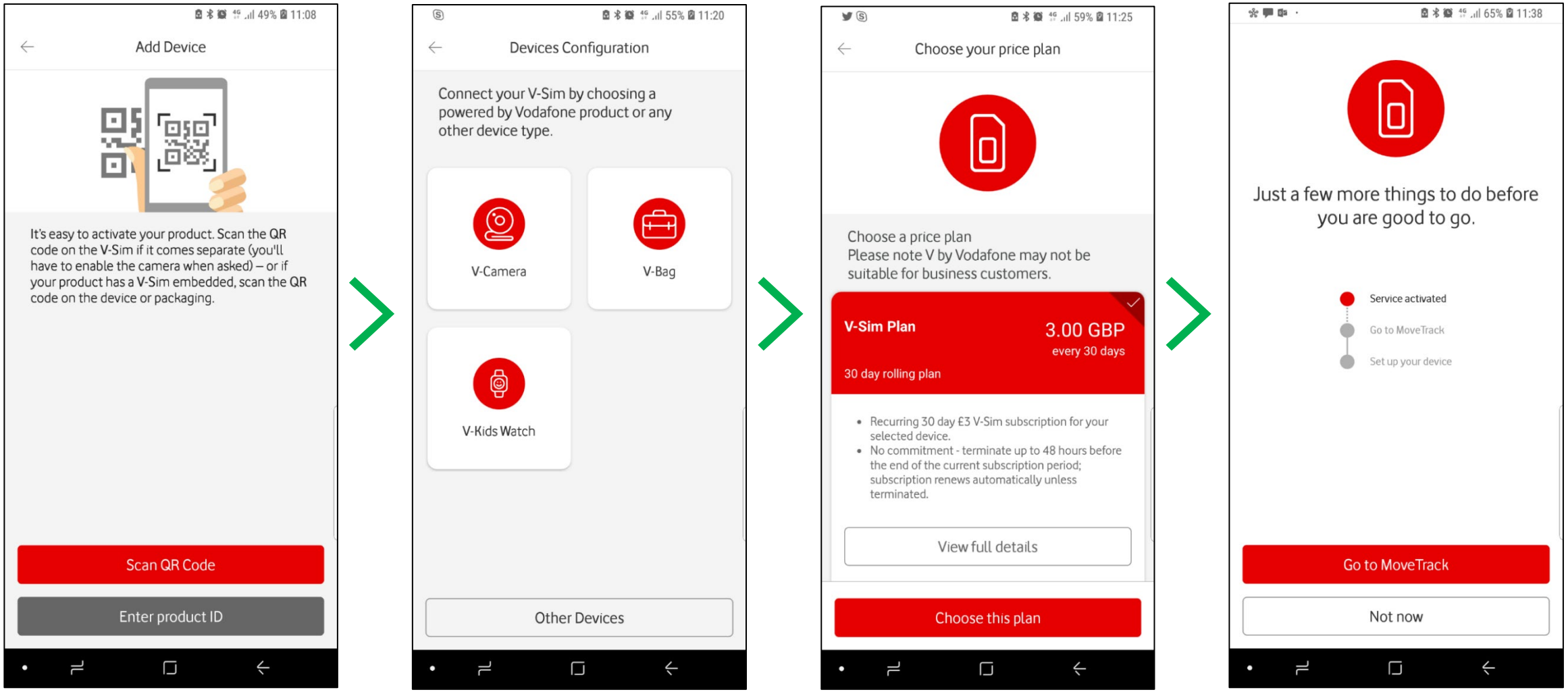
C-LEVEL
High level company strategy to allocate improvements budget

MARKETING
Optimization of campaigns and promotions: target, media, content..

CUSTOMER SERVICE
Enhancing easy setup and early first usage

VODAFONE CIoT EXAMPLE

V-App customer onboarding journey analysis



VODAFONE CIoT EXAMPLE

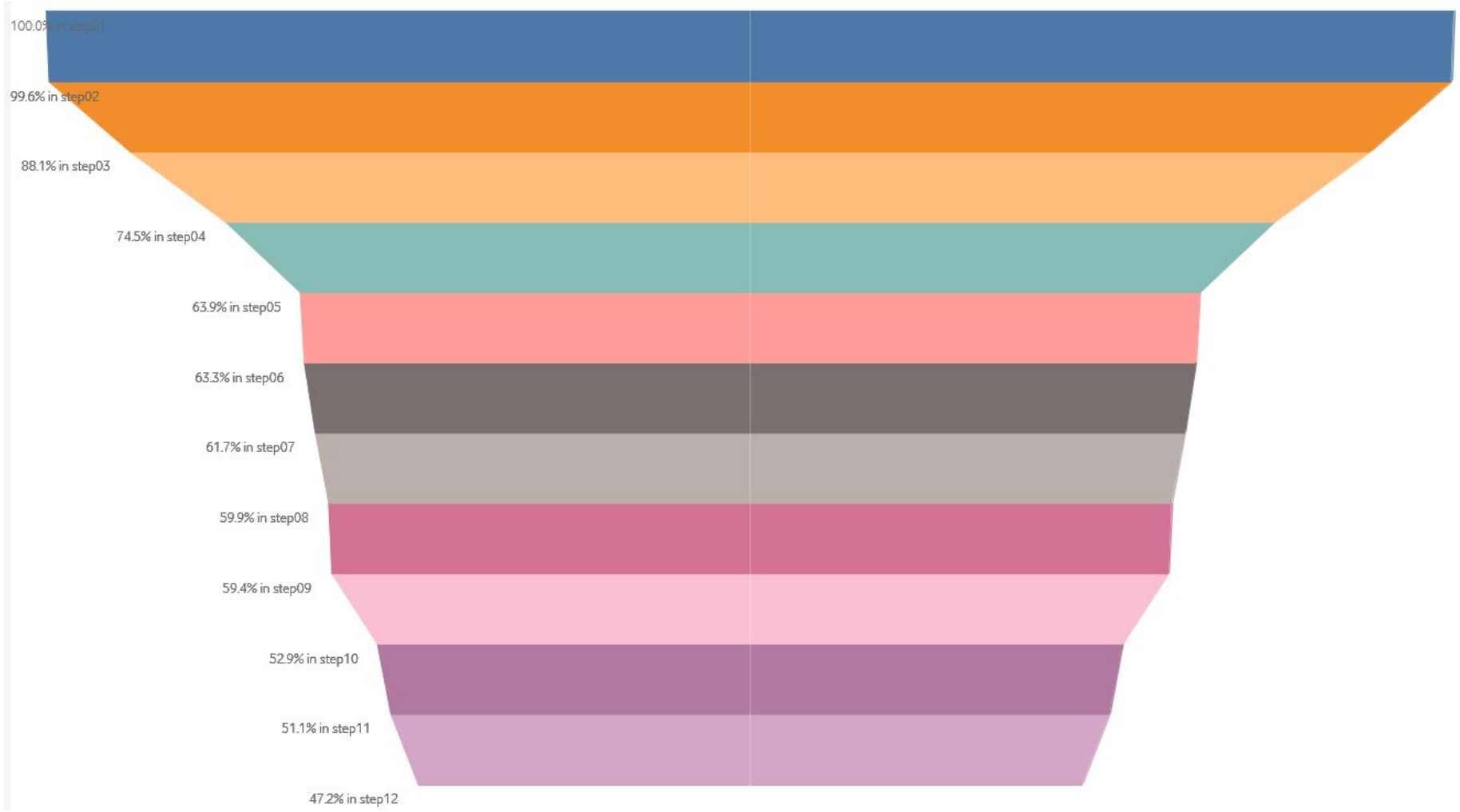
V-App customer onboarding journey analysis

	Funnel_Step	GB Count	GB conv rate (%)	GB conv rate relative to first step (%)
1	Tapped on "+" button on the home screen	762	100.00 %	100.00 %
2	"Add device" screen shown	758	99.48 %	99.48 %
3	Started product identification	632	83.38 %	82.94 %
4	Device ID obtained and "Identifying your product" screen shown	570	90.19 %	74.80 %
5	"We've got it" screen shown	509	89.30 %	66.80 %
6	Tapped "Confirm" button	508	99.80 %	66.67 %
7	"Choose your price plan" screen shown	496	97.64 %	65.09 %
8	Tapped "Choose this plan" button	484	97.58 %	63.52 %
9	Tapped "Accept" button on the "Terms of service" pop-up	480	99.17 %	62.99 %
10	Payment page shown and "Subscribe now" button tapped	411	85.63 %	53.94 %
11	"Activating" screen shown	400	97.32 %	52.49 %
12	"Activation successful" screen shown	370	92.50 %	48.56 %

C2 General

VODAFONE CIoT EXAMPLE

V-App customer onboarding journey analysis



GIVE YOUR BEST SHOT!

A DEEP DIVE IN YOUR BUSINESS CASES

DO NOT INSERT SENSITIVE INFORMATION (C3 and above)

CHALLENGE

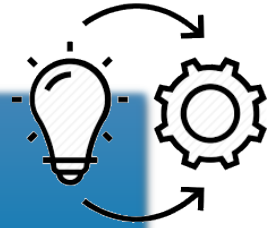
Which need/pain-point you would like to address?

OBJECTIVE

Which high level improvement/goal you would like to achieve?

DATA

Which data you have at disposal or you would need?



APPROACH AND IMPLEMENTATION

How do you leverage data to reach your objective and beat your challenge?

BUSINESS BENEFITS

How do you measure your success?

C2 General

CUSTOMER JOURNEY AND ADV OPTIMIZATION

Fashion Company Case

Improve efficiency and effectiveness of communications and marketing campaigns

Build actions based on the discovery of seasonality, outliers, trends and impacts of marketing levers

+0,6% increase of RPV on homepage

	2017	HP	2018
Visits	21473.2	-0.1%	21457.0
Bounces *Bounce Rate*	15.0%		16.5%
Active Users	17988.2	-0.9%	17820.1
"Unwilling to buy" *(1-Carts/Active Users)*	91.4%		92.4%
Carts	1603.1	-15.7%	1351.3
Carts Abandoned before Checkouts	66.0%		68.3%
Checkouts	550.1	-22.4%	426.8
Carts Abandoned at Checkouts	47.6%		38.4%
Orders	290.7	-13.2%	252.2
Revenues	30715.4	0.8%	30961.3
RPV ($)	**1.43**	***+0.6%***	**1.44**

C2 General

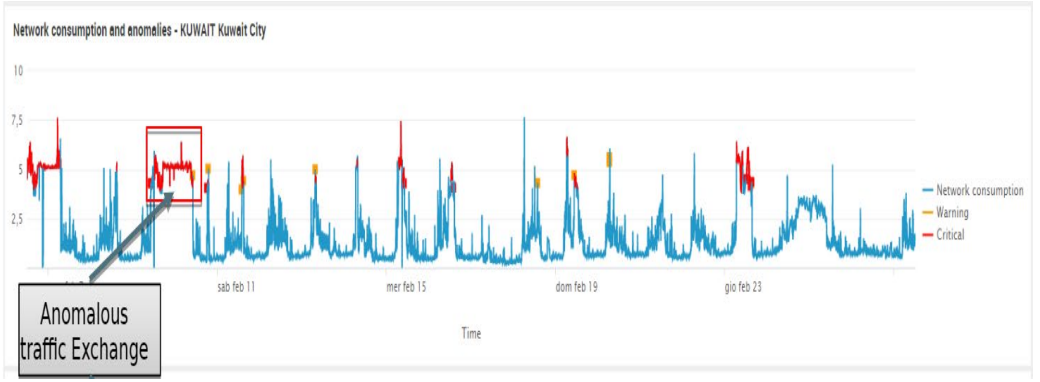
ANOMALY DETECTION ON NETWORK UTILIZATION

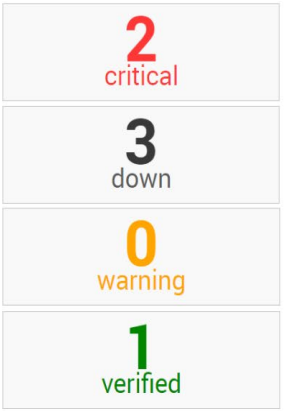
Oil&Gas Company Case

Predict network failures, minimize downtime and optimize interventions for maintenance

ARIMA-based anomaly forecast based on classification through machine learning

>99% reduction of false positives alarms





C2 General

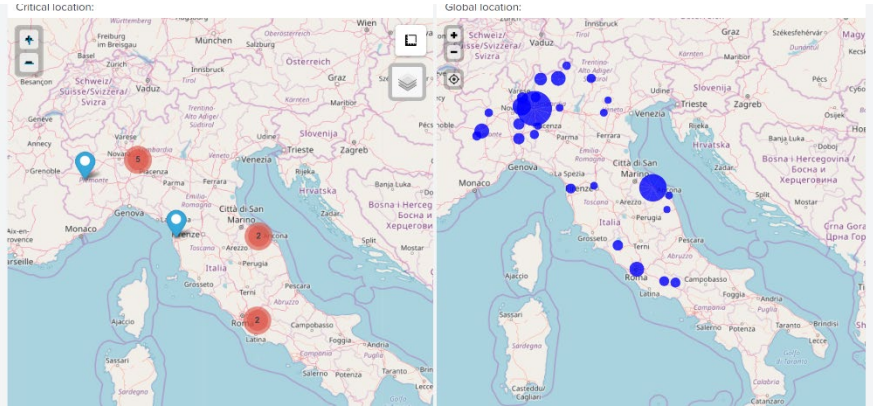
BANK OFFICES CASH PROACTIVE MANAGEMENT

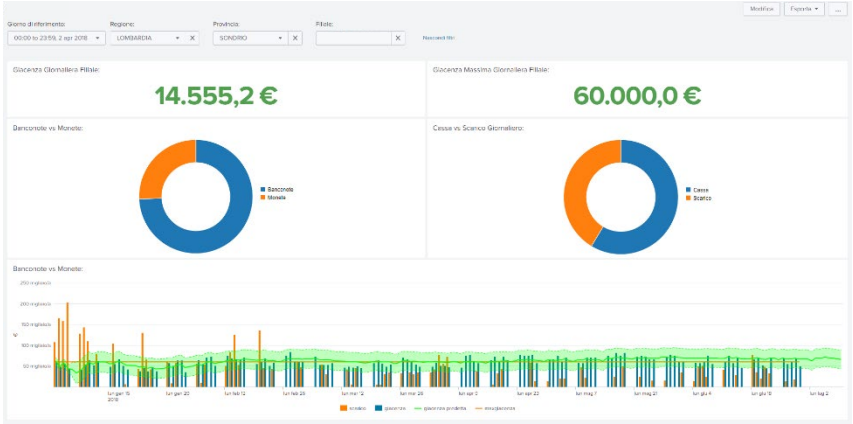
Banking Company Case

Optimize amount of cash loads and unloads

Prediction of cash load, trends and occasional events
Effective data representation on maps

60% savings on cash load-unload expenses
Enhanced cash traceability





C2 General

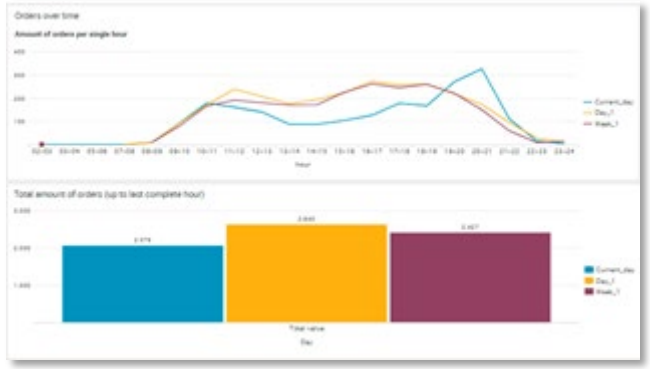
BUSINESS PERFORMANCE ANALYSIS

Entertainment Company Case

Track business performances
Optimize resource allocation

Define customer tailored experience
Optimize visualization tools

Improvements in Marketing activities effectiveness
Expected **single digit savings** through optimization
Expected significant incomes increment



C2 General

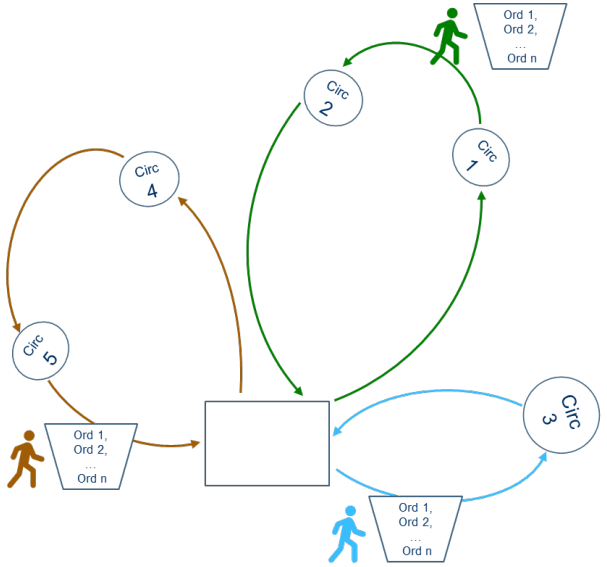
ORDER AND PICK SERVICE OPTIMIZATION

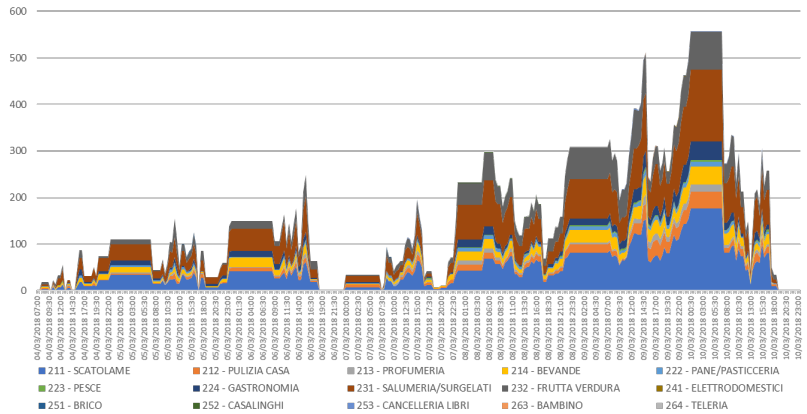
Grocery Company Case

Optimize order-and-pick service

Properly handle orders queue
Minimize collection paths in real time

Almost 50% reduction of busy time
Proportional **order increase**
23% time to pick enhancement





C2 General

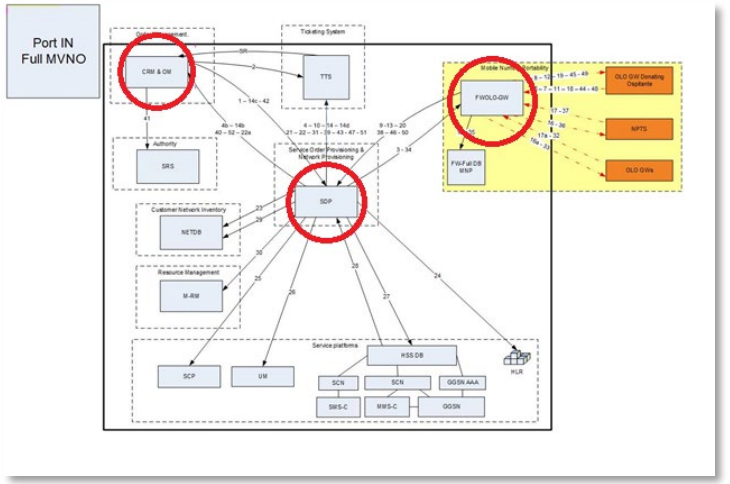
NUMBER PORTABILITY BUSINESS PROCESS MINING

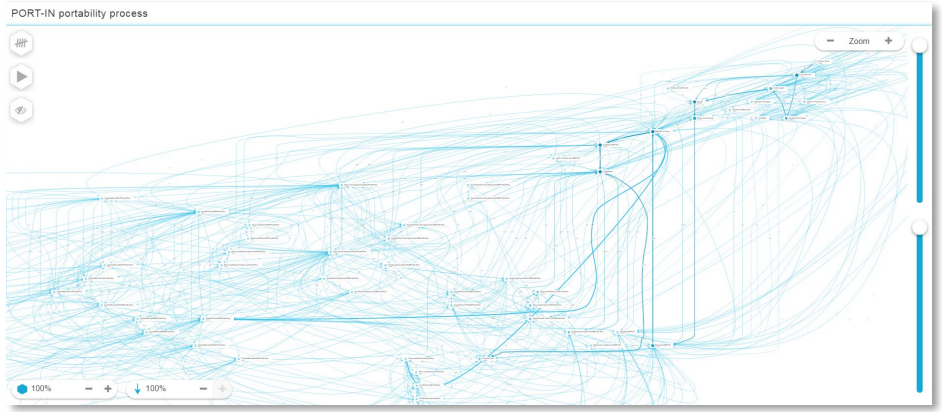
Telco Company Case

Avoid penalties for number portability time delays

Business process mining on event data
Discovery and enhancement of process model

Automatically map and maintain process topology
Forecast behavior deviations and counter them in advance





C2 General

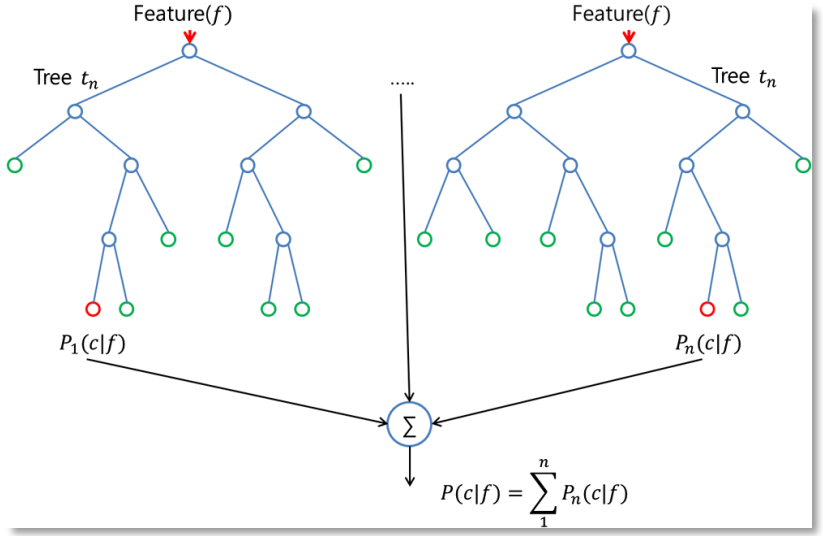
PREDICTIVE MAINTENANCE

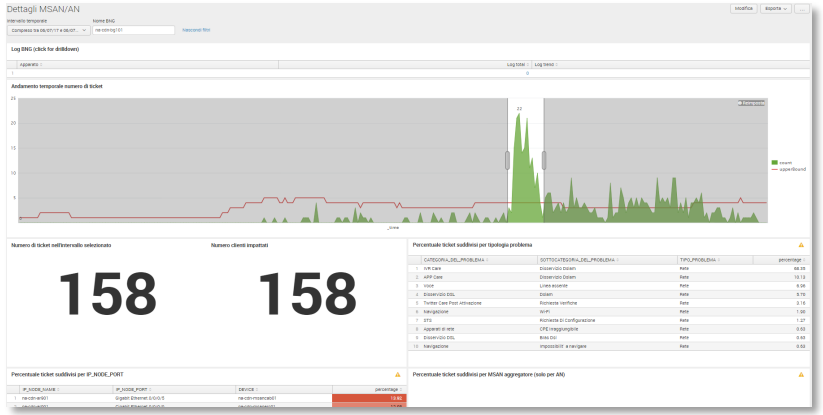
Telco Company Case

Forecast disruptions caused by network failures

Identification and classification of anomalous patterns

90% accuracy in predicting failures
Predictions available **days in advance**
25% of failure issues prevented





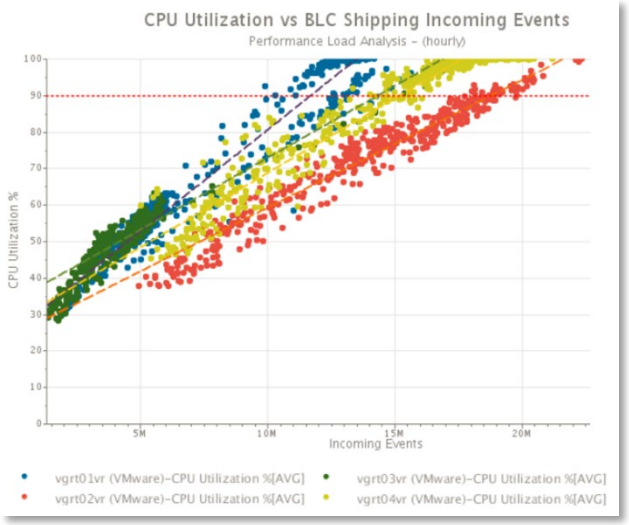
IT INFRASTRUCTURE CAPACITY PLANNING

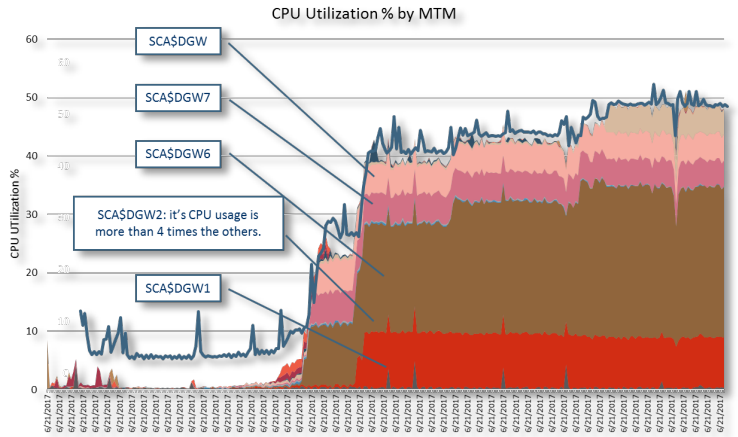
Projects in banking, Insurance, Retail, Telco (and more) markets

Improve service reliability
Reduce infrastructure costs

Manage and analyze data from multiple resources
Map relations among infrastructure and business KPIs

10%-20% unused components dismissed
10%-40% enhancement of resource use
10%-20% reduction in IT troubleshooting





C2 General

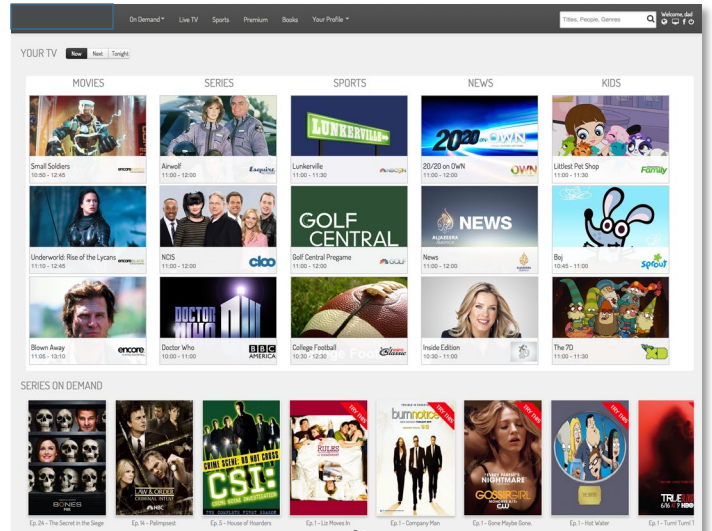
PERSONALIZATION – CONTENT DISCOVERY

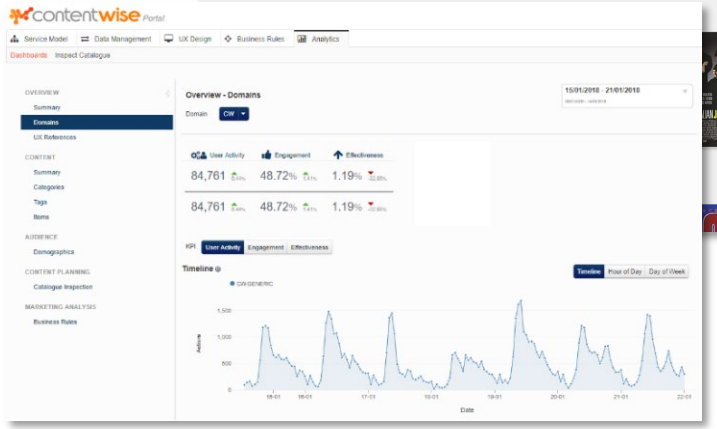
Video Broadcasting cases

Increase VoD/PPV income and cross/up-selling
1 to 1 user experience personalization

Customer profiling and tailored content recomendation
Editorial feedbacks integration

+26% of PPV purchases
+38% of monthly ticket per user
-20% customer churn
+15% catalog enhancement





C2 General

Credits #: 1

PICTURES CREDITS

1: With courtesy of Warner Bros. Animation

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.

DATA FITNESS PROGRAM

DAY 1

Kick Start

BUSINESS CASES PRESENTATION

DATA VISUALIZATION

Lunch Break

DATA ANALYSIS (Basic)

DAY 2

09:30

Kick Start

DATA MANAGEMENT

11:00

BUSINESS CASES DEEP DIVE

DATA FITNESS PROGRAM: Legs reinforcing

Performing proper Data Management is crucial for Data Science projects to have solid foundations, like legs are for body

AGENDA

- Introduction to Data Management
- The 5 V's
- Hadoop framework
- Types of database
- Data Management

- The basis of Data Science
- Examples of bad Data Management

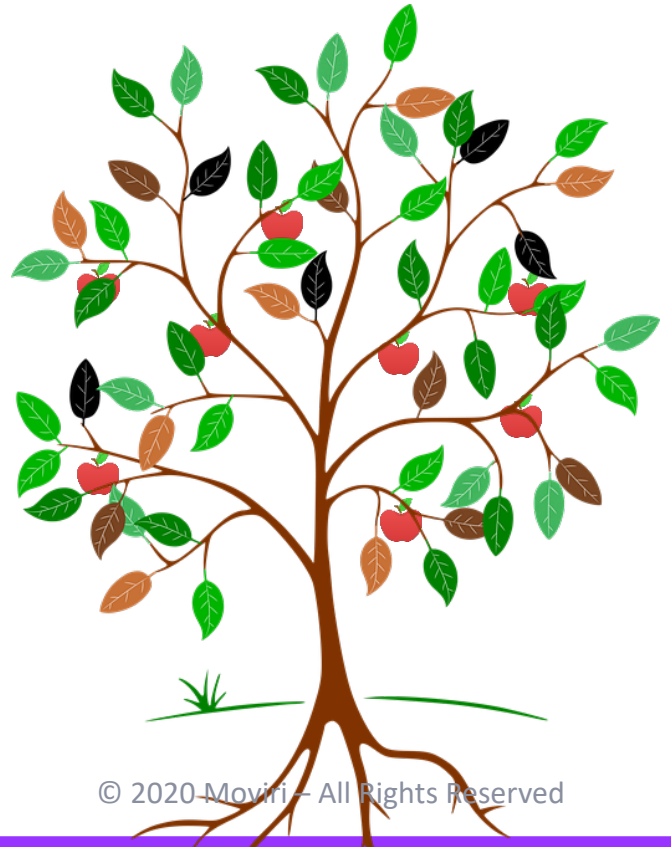
DATA MANAGEMENT AS THE ROOTS FOR A FRUIT TREE

The basis of Data Science

FRUITS
The most visible result of a Data Science project
Reports, Dashboards

WATER & SUN
Dataflows, properly configured to feed ground and roots
Ingestion Layer

TRUNK, BRANCHES & LEAVES
Where logics are applied to build results
Models, Analyses



GROUND
Where data management is built and can handle data
IT infrastructure

ROOTS
Tools and procedures to handle data and feed the whole tree
Data Management

REAL WORLD EXAMPLE #1

Examples of bad DM

Data stored in a big cluster

Campaigns

based on

Sensible data

which are

Easily accessible

due to

IT policies

No masking or encryption, private data can be stolen

NO PROTECTION

CZ General

REAL WORLD EXAMPLE #2

Examples of bad DM

Set of Spark scripts on data collection and data processing on AWS environment

Daily amount of data: ~1TB

File format: CSV

Jobs' running time: ~5 hours

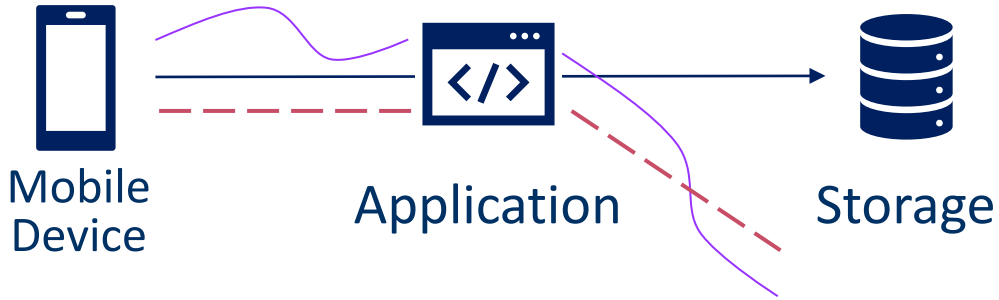
Data format is not optimized for the framework

COMPUTATIONAL ISSUES

CZ General

REAL WORLD EXAMPLES #3

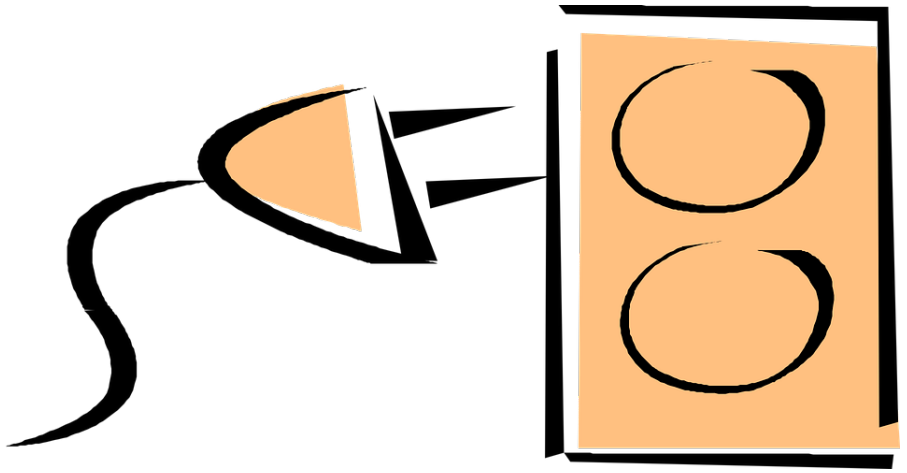
Examples of bad DM



Optimizing the application to deal with variations and exceptions in input data

implies

Effort increase in the implementation of new features



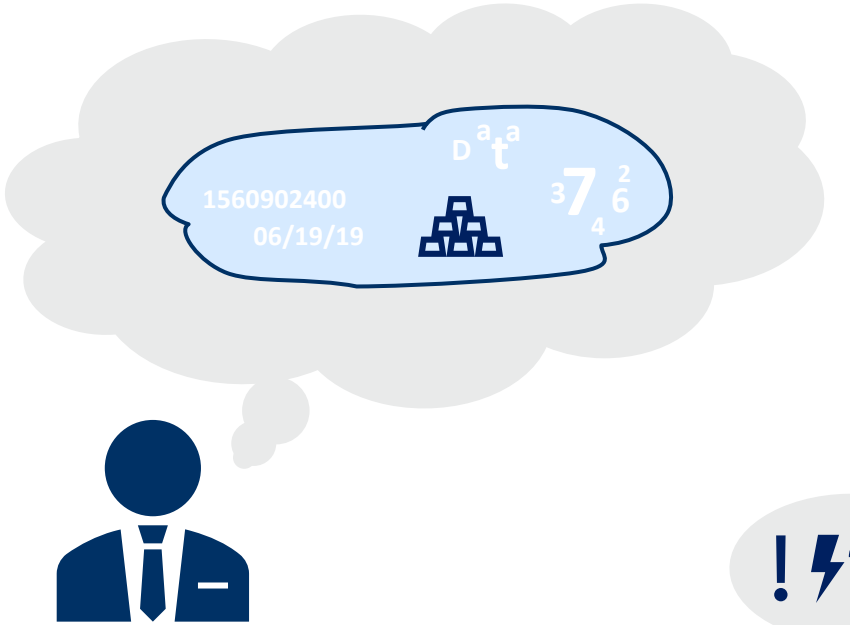
Lack of appropriate design or dev quality

NO SCALABILITY, FLEXIBILITY, ADAPTABILITY

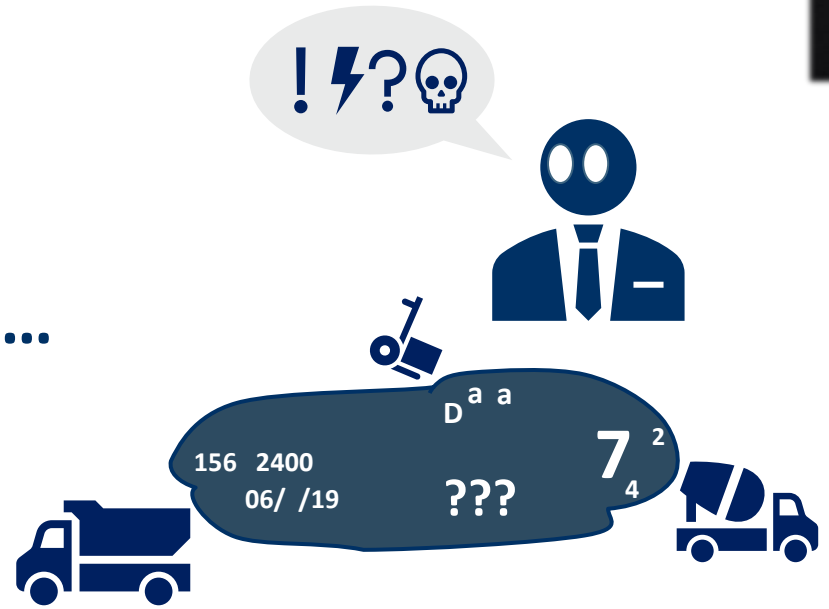
CZ General

REAL WORLD EXAMPLES #4

Examples of bad DM



…6 months later…



Technology has guided the project business needs and not viceversa

WRONG CHOICES

CZ General

AGENDA

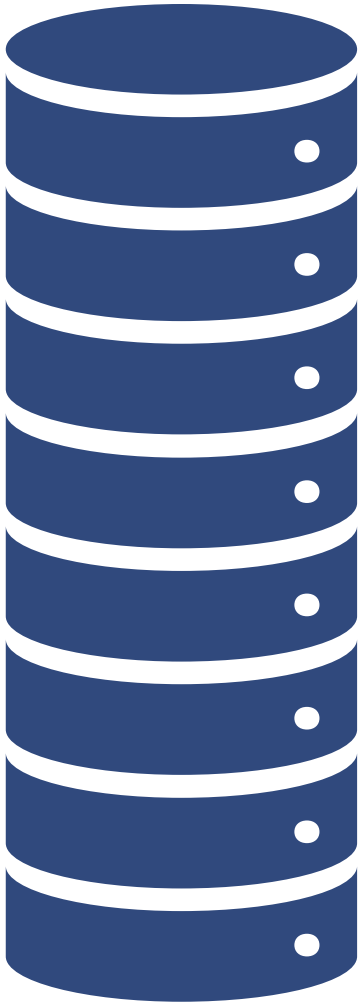
- Introduction to Data Management
- The 5 V's
- Hadoop framework
- Types of database
- Data Management

- Volume
- Variety
- Velocity
- Veracity
- Value

VOLUME IS THE AMOUNT OF DATA TO HANDLE

V

Volume



Big Data technology
is crucial

Traditional database technology is outdated

44 zettabytes
generated worldwide by the end of 2020

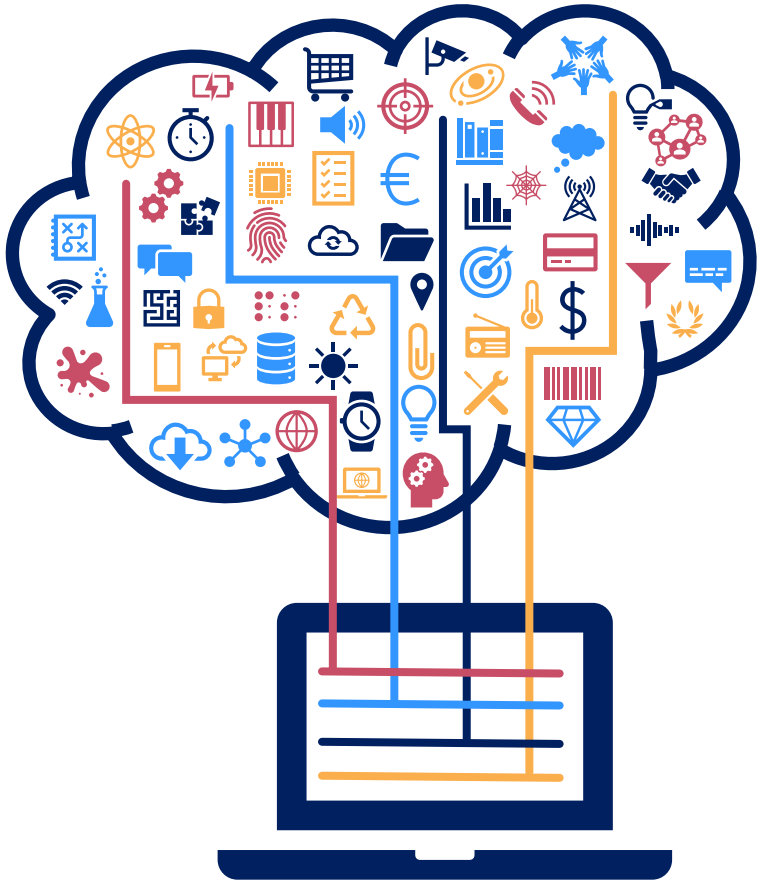
1.7 megabytes
person/second

C2 General

VARIETY IS DUE TO THE HUGE NUMBER OF DATA TYPES

Variety

	MAIN DATA TYPE
Past	Structured
Now	Structured Unstructured



Big Data Approach

VELOCITY OF DATA GENERATION IS STEADLY INCREASING

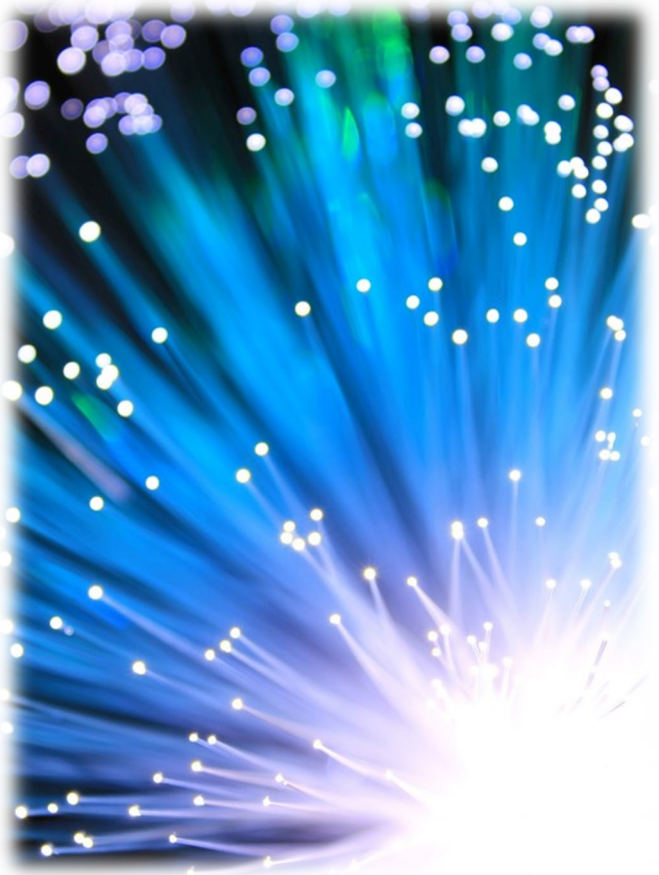
Velocity

Speed

Collection

Analysis

Reaction



Big Data Technologies allow for live/real time data processing

C2 General

VERACITY IS A KEY FACTOR FOR DATA CONSISTENCY

Veracity

Data reliability enemies

- Ambiguities
- False statements
- Latencies
- Deceptions

How to fight them

- Machine learning algorithms
- Artificial intelligence algorithms

C2 General

VALUE IS THE MOST IMPORTANT ASPECT TO CONSIDER

Value

$^{3}7^{2}_{4}6\ D^{a}t^{a}$

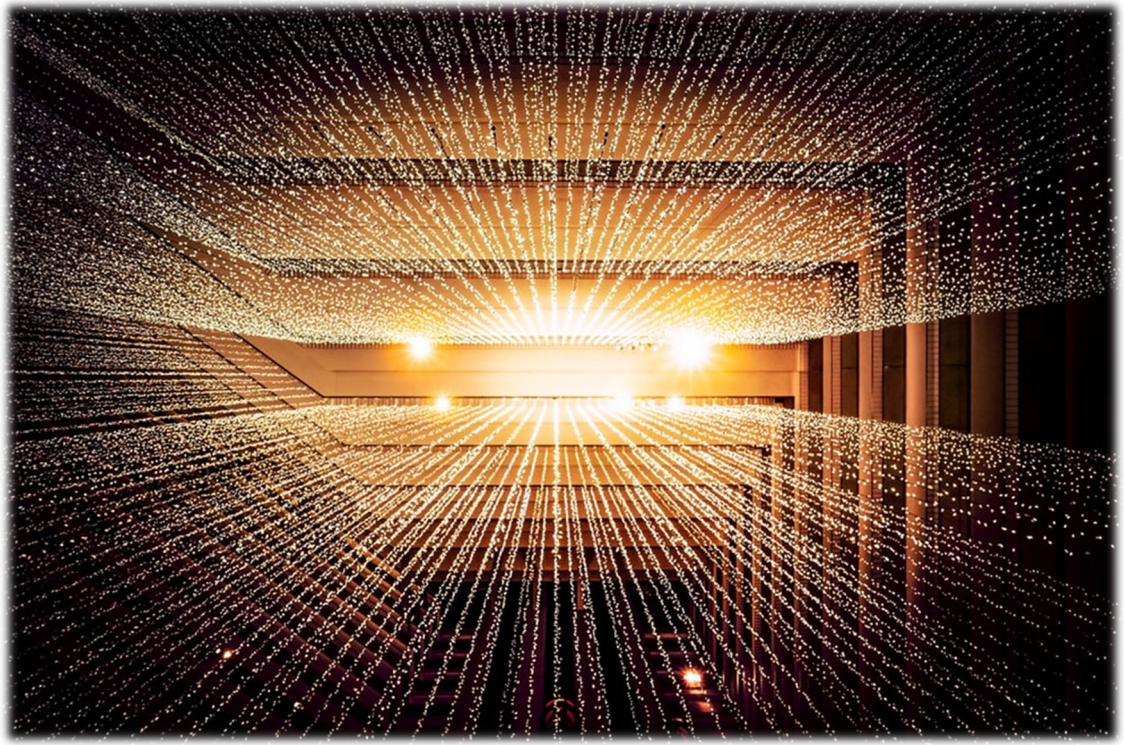
1560902400

06/19/19

Data Management

with a proper plan

without a proper plan



C2 General

AGENDA

- Introduction to Data Management
- The 5 V’s
- Hadoop framework
  - A complex use case
  - Why Hadoop?
  - Hadoop Ecosystem
  - An example of Hadoop infrastructure
- Types of database
- Data Management

BIG DATA MEAN BIG CHALLENGES

A complex use case

Ever **growing volume** of data and use cases

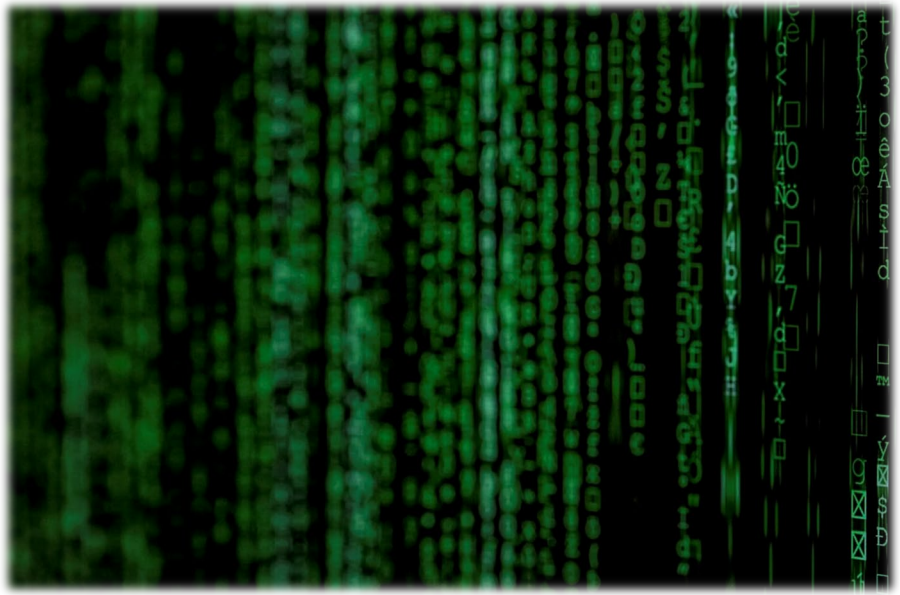
Data come **continually**
opening to several needs:

Visualization through a **specific JDBC tool**

Sliding time window for **pre-processing** activities and analysis through **ML** and **DL**

Ensuring **security** through the entire infrastructure

Computing **complex calculations** on live data



WHAT IS HADOOP?

Why Hadoop?

Apache Hadoop software library

is a

framework

that allows for

distributed processing

of large datasets across

clusters of computers

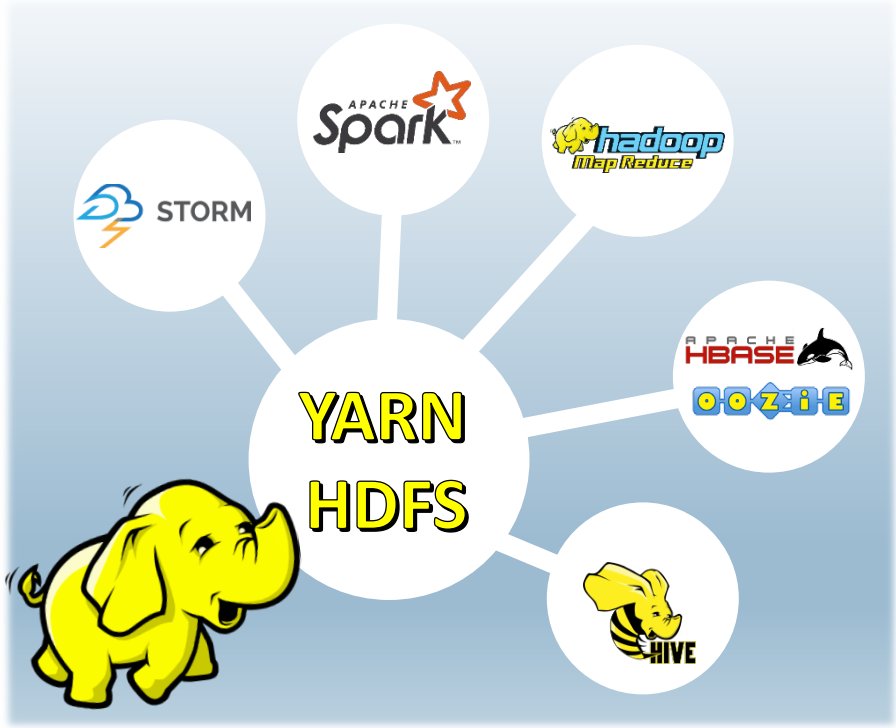
using simple programming models

Designed to scale up

from single servers

to thousands of machines,

each offering local computation and storage



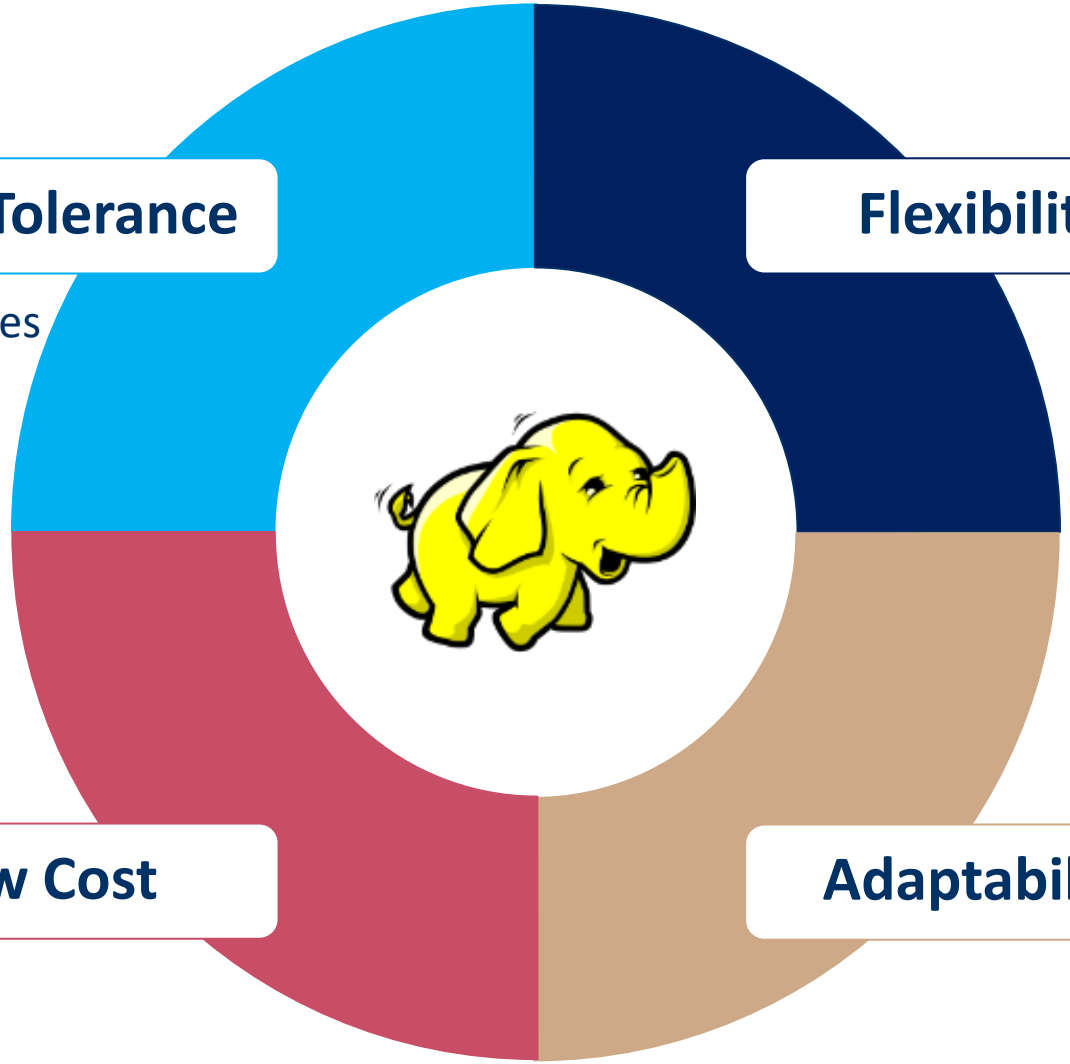
WHY HADOOP?

Fault Tolerance

Designed to detect and handle failures at the application layer, delivering a highly-available service on top of a cluster of computers

Flexibility

Fully scalable ecosystem



The framework can be run on commodity hardware

Low Cost

Interchangeability of system's components with other frameworks

Adaptability

HADOOP CHALLENGES

Why Hadoop?

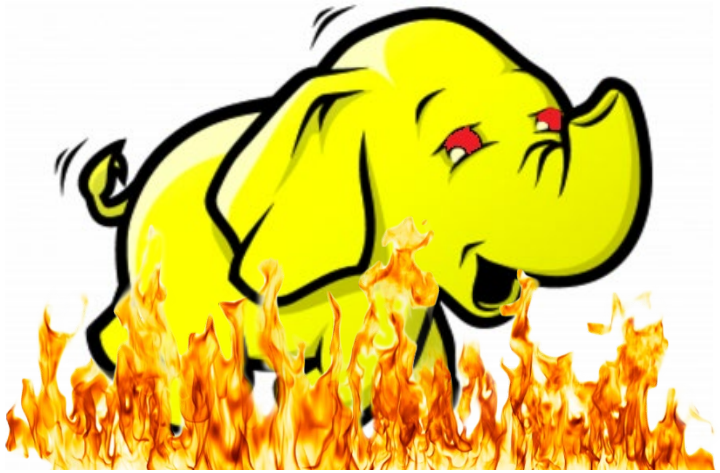
Hadoop is a complex distributed system with low-level APIs implying the need of distributions like CLOUDERA or MAPR

Specialized skills are required to properly use Hadoop, as:

Different processing paradigms require specific data format

Real-time and batch ingestion requires deeply integrating several components

Significant effort is wasted on simple tasks like data ingestions and ETL

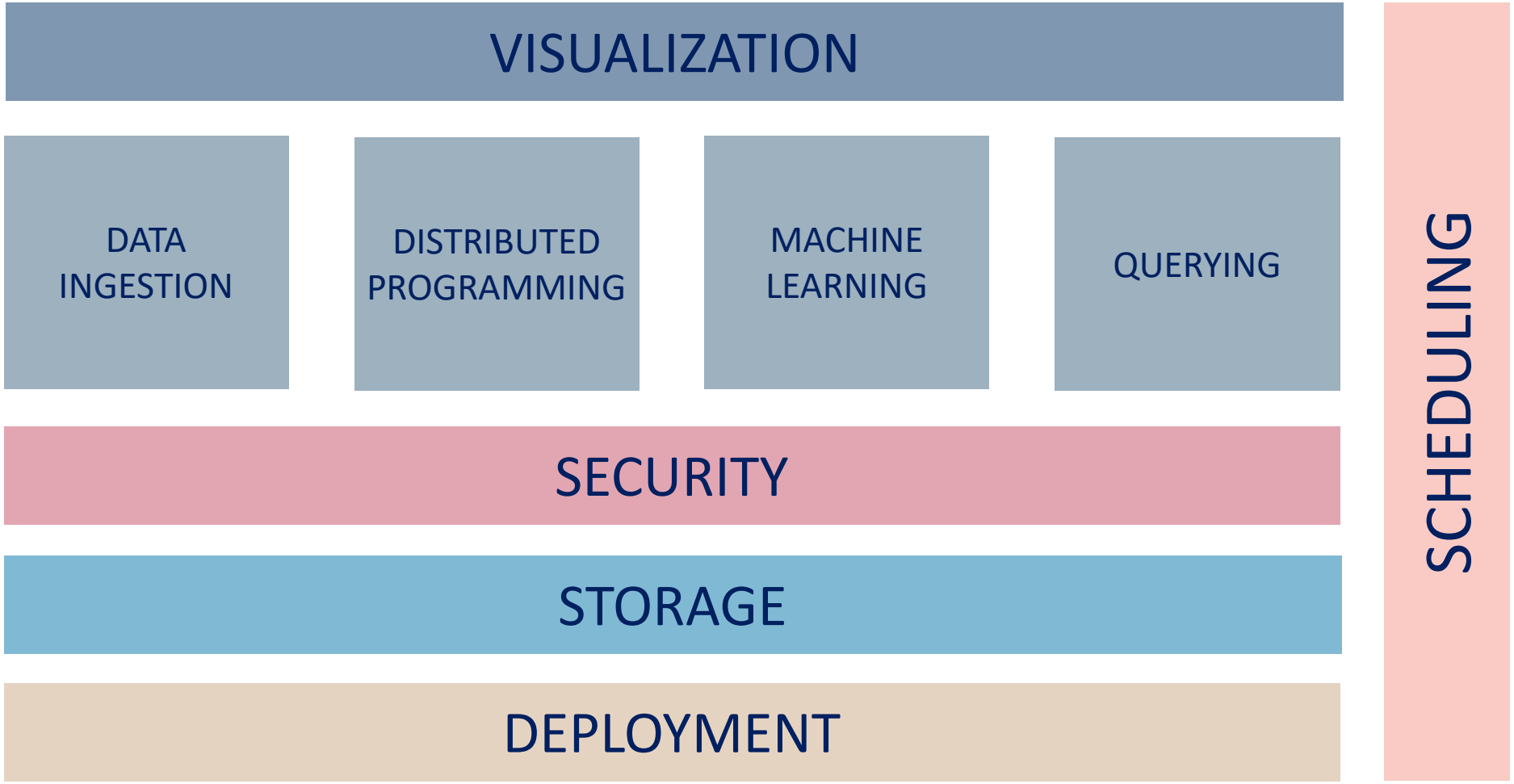


Automated testing of end-to-end solutions is impractical or impossible

C2 General

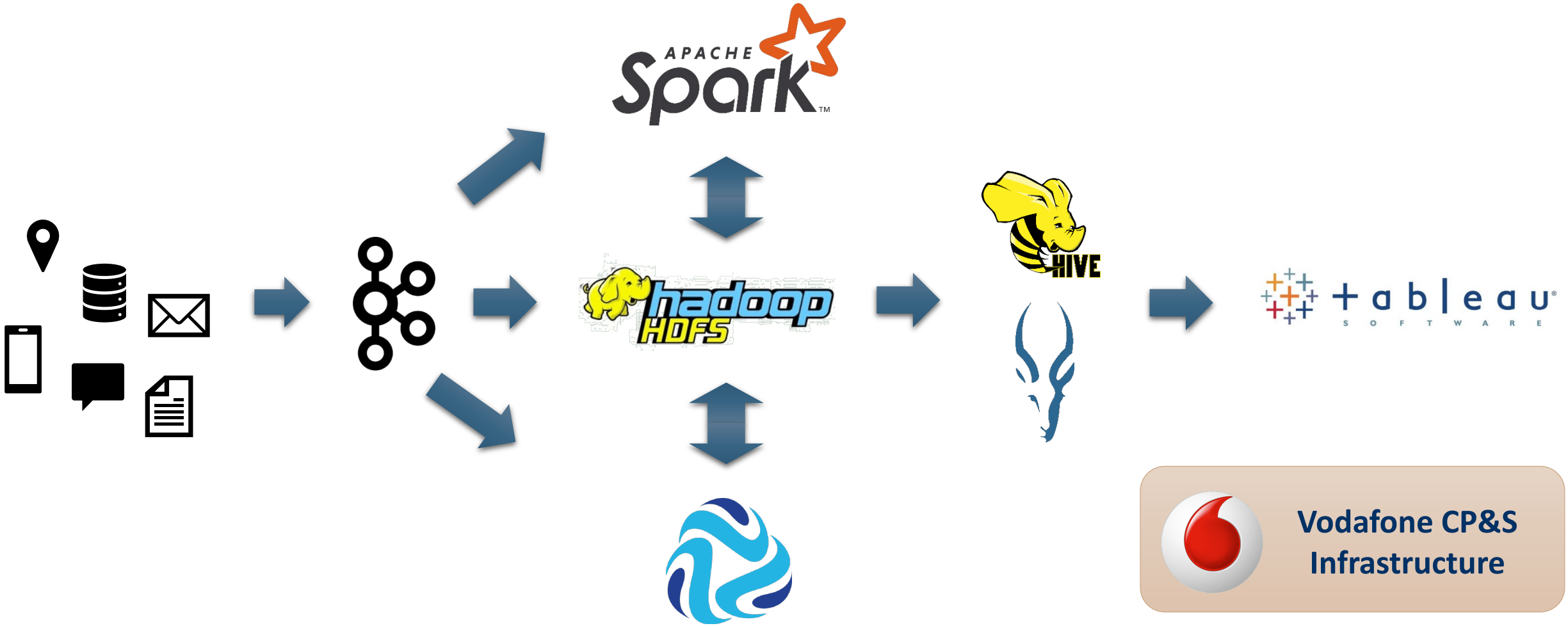
HADOOP ECOSYSTEM

Hadoop Ecosystem



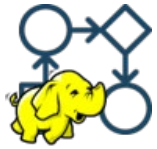
AN EXAMPLE OF A HADOOP (CLOUDERA) INFRASTRUCTURE

An Example

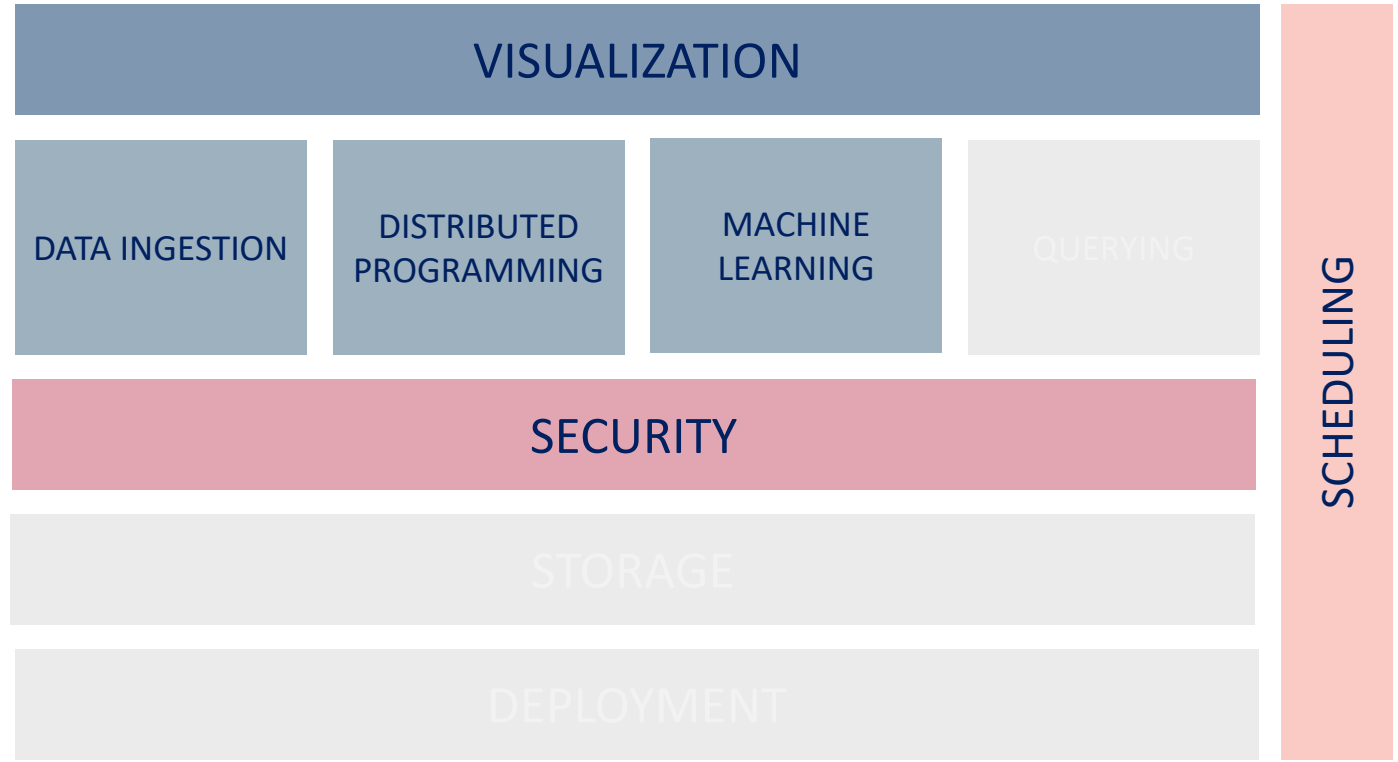


Vodafone CP&S Infrastructure

ALTERNATIVES TO HADOOP (FOR VODAFONE CIoT)



An Example



Datameer
stay curious

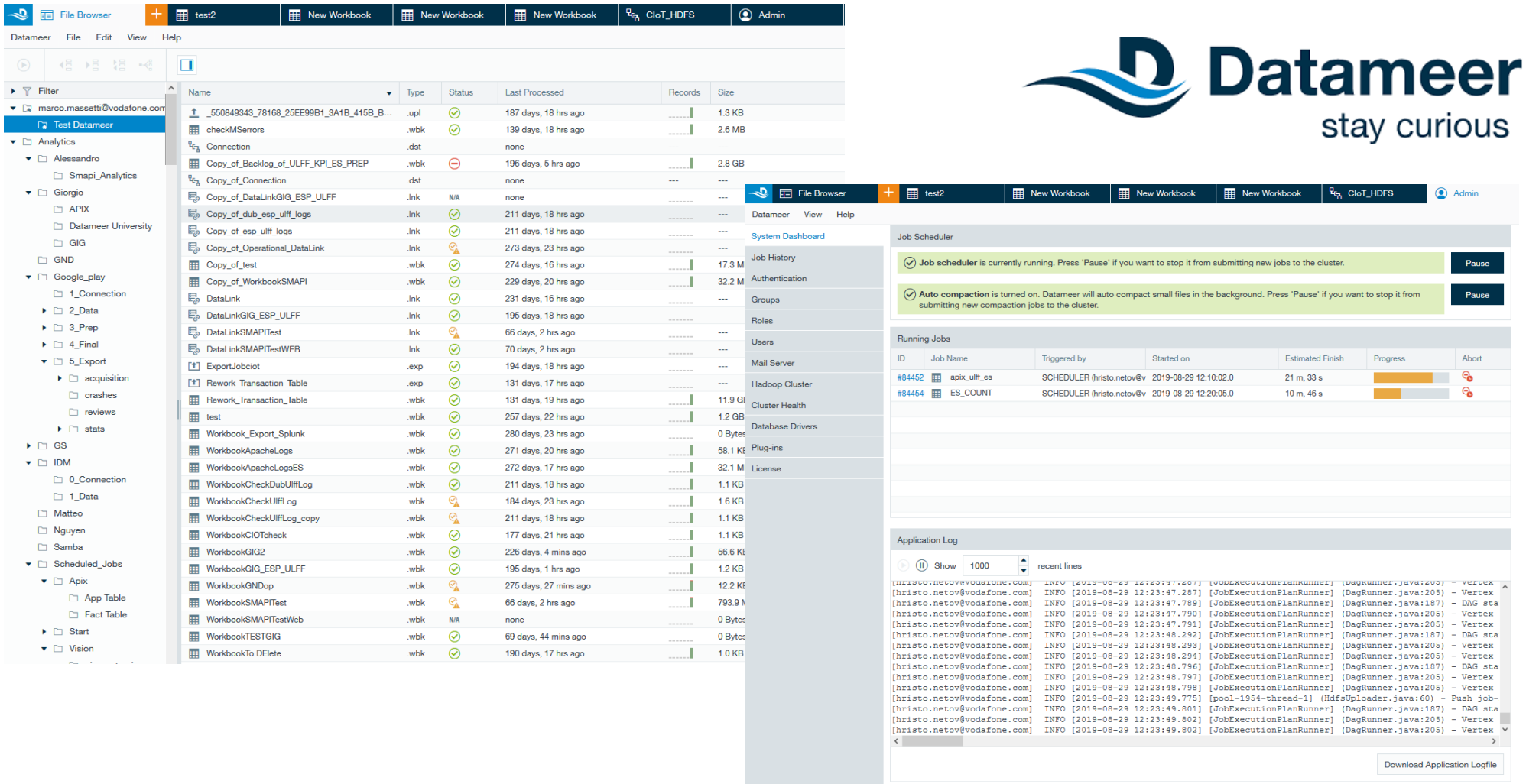
Datameer takes care of the majority of the explained logical functions

Allows easy visual exploration through an Excel-like look and feel

C2 General

ALTERNATIVES TO HADOOP (FOR VODAFONE CIoT)

An Example

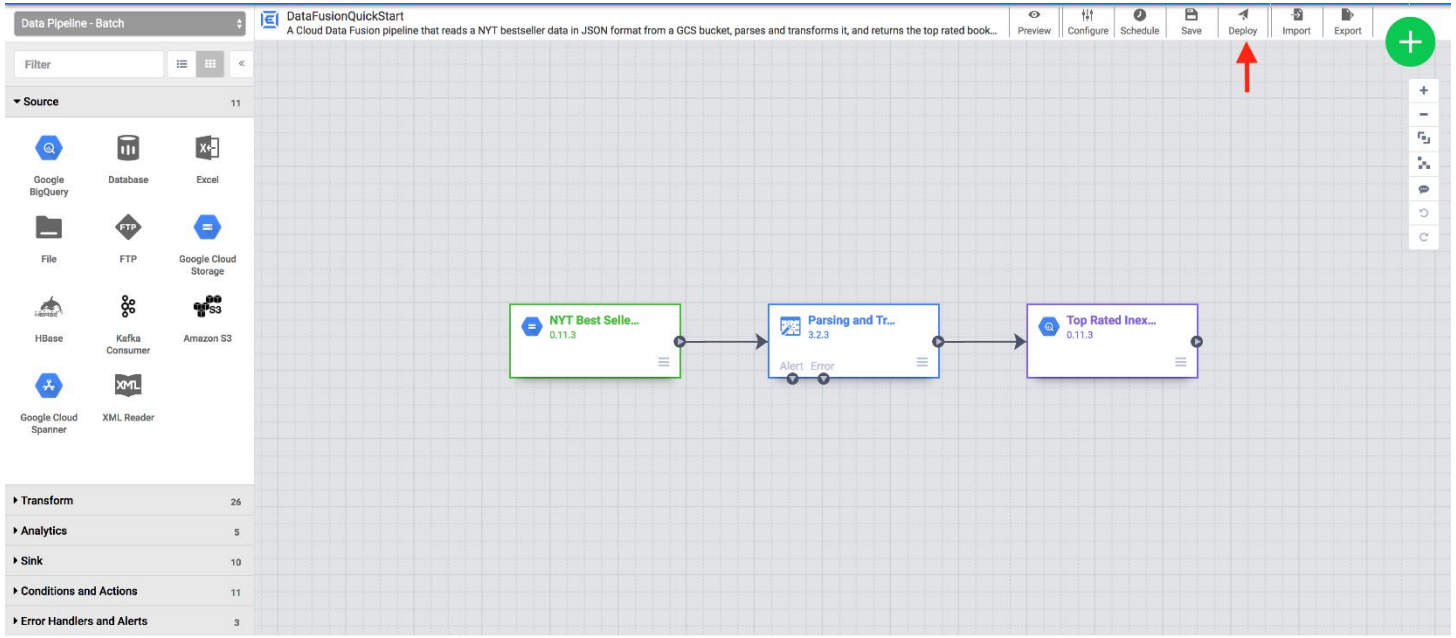


(FUTURE) ALTERNATIVES TO HADOOP (FOR VODAFONE CIoT)

An Example

Google Cloud Data Fusion

Google CDF maps the same logical functionalities (and more) of Datameer, integrated in Google Cloud Platform framework



Credits #: 1

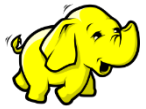
15 minutes Break

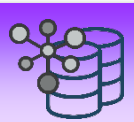
https://giphy.com/explore/good-point

AGENDA


Introduction to Data Management


The 5 V’s


Hadoop framework

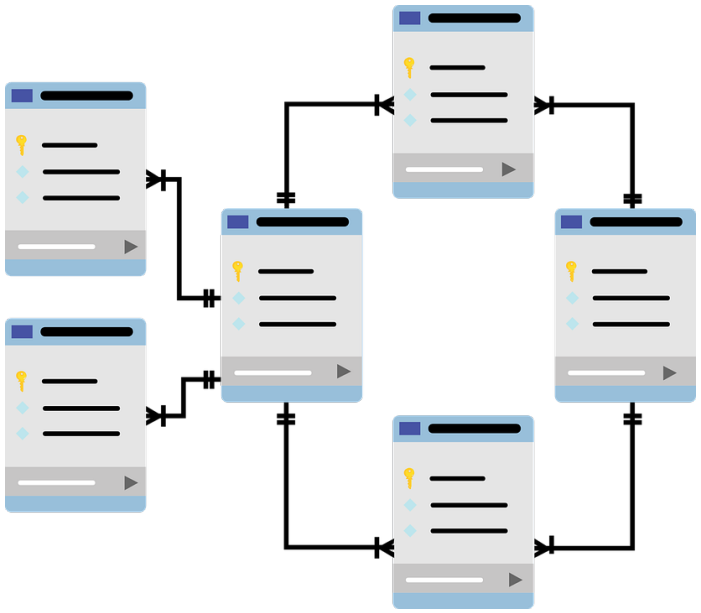

Types of Database


Data Management

SQL vs NoSQL Databases

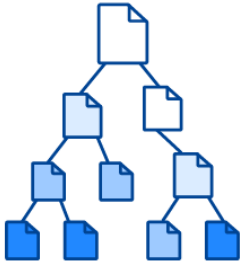
Relational vs Non-Relational Databases

SQL

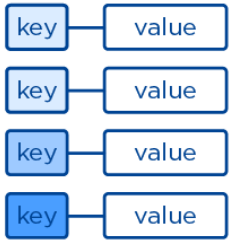


NoSQL

Document

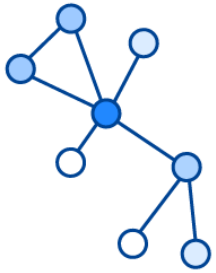


Key-Value



Time series

Graph



Full text search

SQL vs NoSQL Databases

Face-to-face

SQL

- Relational Databases (RDBMS)
- Structured Data stored in tables
- Vertical scalability
- Predefined schema
- Structured Query Language (SQL)
- ACID properties (Atomicity, Consistency, Isolation, Durability)
- Typical OLTP (On-Line Transaction Processing)

NoSQL

- Non-Relational/Distributed Databases
- Semi-/Un-structured data
- Horizontal scalability
- Dynamic schema
- No declarative query language
- CAP theorem (Consistency, Availability, Partition tolerance)
- Mainly OLAP (On-Line Analytical Processing)

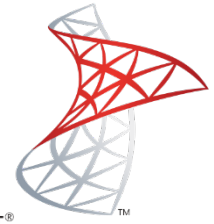
SQL vs NoSQL Databases

Products

SQL

Cloud SQL

PostgreSQL



Microsoft® SQL Server®

NoSQL

Cloud Datastore

AGENDA

- Introduction to Data Management
- The 5 V's
- Hadoop framework
- Types of Database
- Data Management
  - GDPR
  - DM @ Vodafone

GDPR

"… lays down rules relating to the protection of […] personal data […] and rules relating to the free movement of data" (GDPR Art 1):

Defines personal data

Any info relating to identified/identifiable person

Special classes: genetic, biometric, judiciary, …

States Privacy by Default

Data manipulation based on explicit consent

Ownership of personal data

States Privacy by Design

Duty of anonymization/pseudonymization

Duty of reporting data leaks

Sets administrative fines

Up to 2% of the total annual worldwide turnover

GIVE YOUR BEST SHOT!

Is it a Personal Identifiable Information (PII)?

NAME

SURNAME

NICKNAME

IMEI

IP ADDRESS

MSISDN

PHONE NUMBER

HUMAN VOICE

PHONE GEO-LOCATION

BANK ACCOUNT (IBAN)

NUMBER OF SUBSCRIPTIONS IN A REPORT

GDPR @Vodafone

Sensitivity @Vodafone

Non-Vodafone

C1 C3

C2 C4

Masking

xxxxxx123456

Encrypting

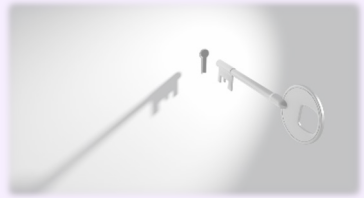
hKH7gpDF83AruEG4tKIcB+L
O1W[illegible]35fI0eQu+A/A4
LH1c[illegible]3kyGE6G+D6ZGQmI
CqQ0+hQmaF5Y0v0pdjA==

Hashing

hKH7gpDF83AruEG4tKIcB+L
O1[illegible]cB5fI0eQu+A/A4
LH[illegible]U3kyGE6G+D6ZGQm
ICqQ0+hQmaF5Y0v0pdjA==

Storage Location Rules

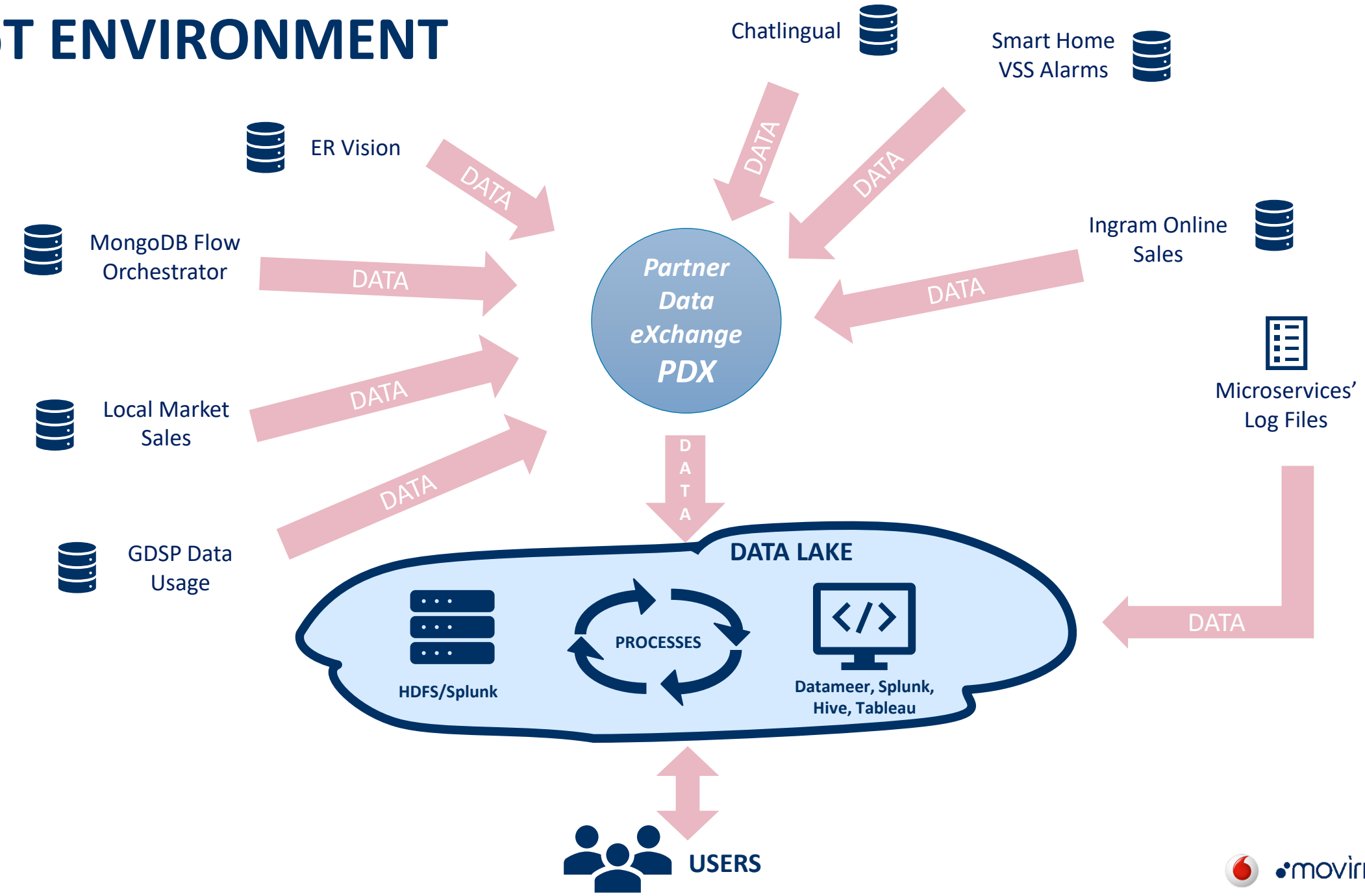
Access Rules



Data Source → **Streaming Layer** → **Storage** → **Exposure Layer**

C2 General

CIoT ENVIRONMENT



CIoT ENVIRONMENT - TODAY

- Shared CP&S Big Data platform
  - CIoT is a **tenant** (user) as any other CP&S service like SecureNet, GNE, DXL etc.
  - Tools: Hadoop / Datameer / Tableau
    - CIoT has roughly 7 TB of encrypted raw data as of today
    - Not approved for C3 data re-identification
    - Retention periods driven by the respective use-cases
- Splunk
    - Managed by TSS Egypt since spring 2019
    - Security-approved for cleartext C3 data
    - Retention period of 60 days
- Partner Data eXchange (PDX)
  - Developed by us in-house
  - Part of CIoT's AWS core infrastructure

CIoT ENVIRONMENT - SOON

- Own CIoT analytics platform and datalake in Google Cloud Platform (GCP)
  - Flexibility
  - Cost and scalability advantages of a cloud
  - Support for actionable C3 use-cases
  - Closer integration with local markets' data
  - Next-generation tools (including ML/AI)
  - Foundation for data-powered product features
- Partner Data eXchange (PDX)
  - Evolution of the platform
  - Integration of even more data sources

PICTURES CREDITS

1: https://cloud.google.com/data-fusion/docs/quickstart

DISCLAIMER

This presentation is intended only for Vodafone internal use.
No copy, use or circulation of this presentation to a third party should occur without the permission of Moviri, which retains all the pre-existing Intellectual Property interests and rights associated with the presentation, according to the signed Purchase Order Terms.
Moreover, all logos, photos, references and copyrighted materials contained in this presentation are and remain property of their respective owners with all rights reserved.
This presentation is intended for educational purposes and do not replace independent professional judgment; due to the complexity of the topics covered, it is suggested in any case to request for special needs a professional advice for any issue in addition to the information here provided.
Statements of fact, special cases and opinions expressed remain reserved.
Moviri assumes no responsibility for the content, technical accuracy or completeness of the information presented and expressly disclaims liability for errors and omission in such context.
Attendees should note that sessions may be audio-recorded and may be published in various media, including print, audio and video formats without further notice.