Introduction

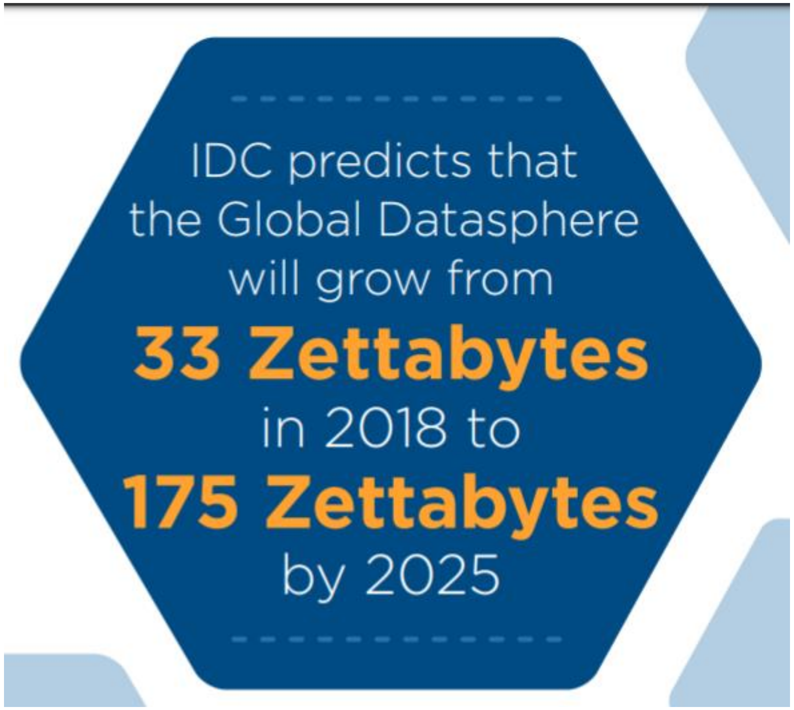
Text, Web and Social Media Analytics Lab

Prof. Dr. Diana Hristova

Introduction

Break-out rooms: what do you think you will learn during this course? Write your group ideas on a slide (or miro). Make a screenshot at the end. Please agree on one student who will represent the group. You have 10 minutes.

Motivation: Why do we need Text, Web and Social Media Analytics?



IDC predicts that the Global Datasphere will grow from **33 Zettabytes** in 2018 to **175 Zettabytes** by 2025

- 80%[1] of it is in unstructured form such as documents, news, e-mails, tweets, videos, pictures, audio streams

➔ Companies have more data than ever to support business decisions

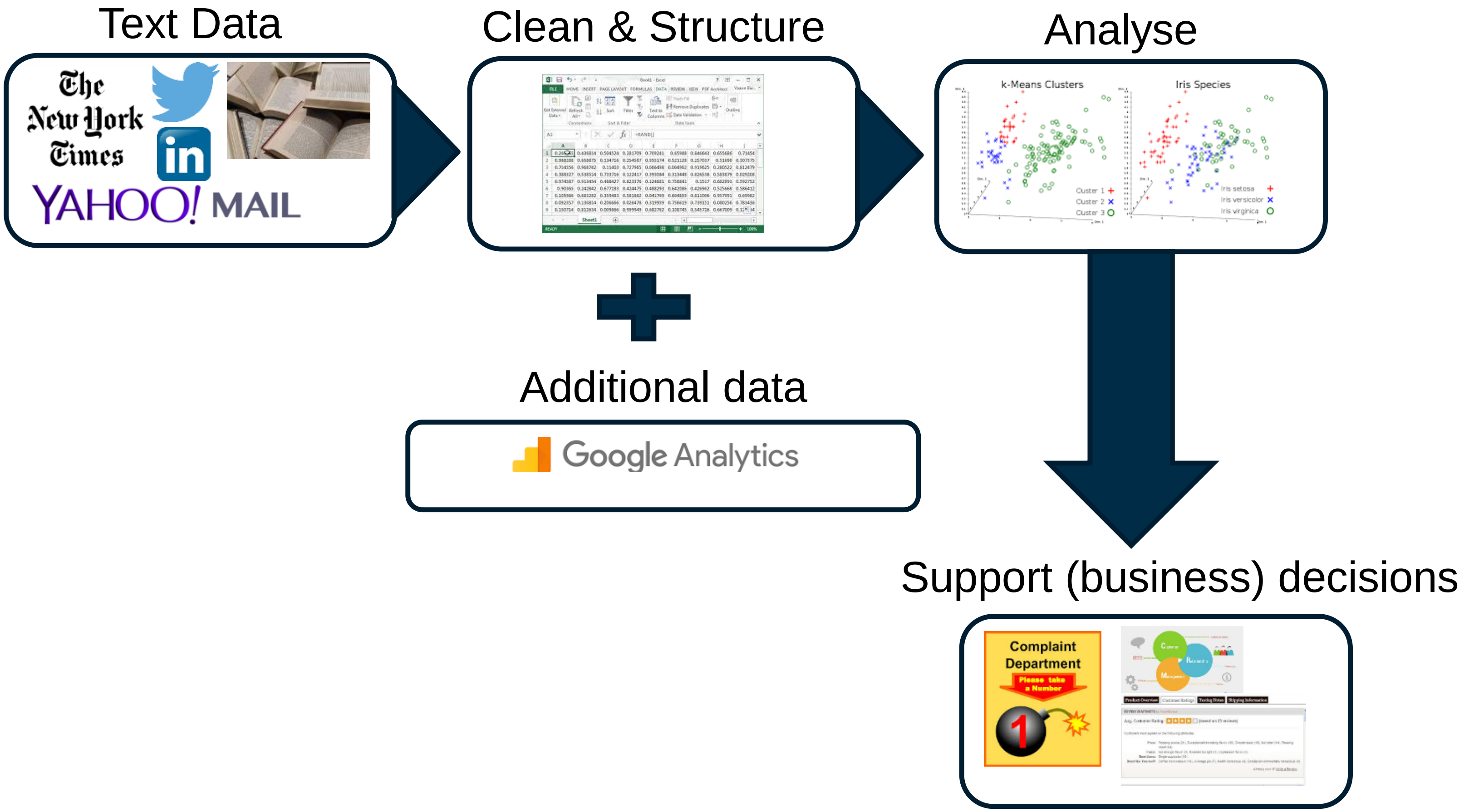
- BUT: Analysing unstructured data requires different capabilities than analysing structured data

Source: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[1] https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

? How can we (semi-)automatically analyse unstructured text data?

Motivation: How can we analyse unstructured text data?



Text Data

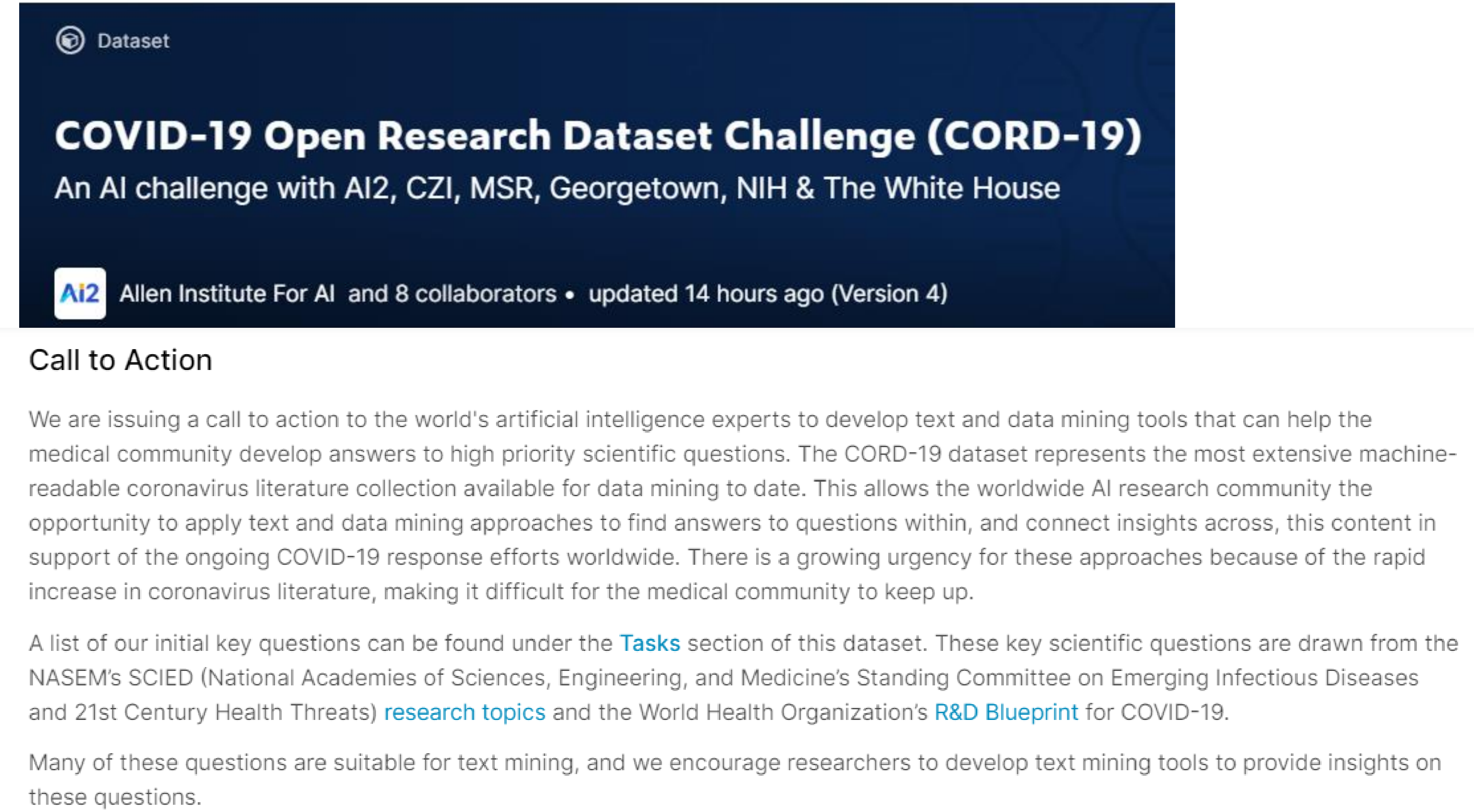
Clean & Structure

Analyse

Additional data

Support (business) decisions

Motivation: Is this really only about business decisions?



Dataset

COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

Allen Institute For AI and 8 collaborators • updated 14 hours ago (Version 4)

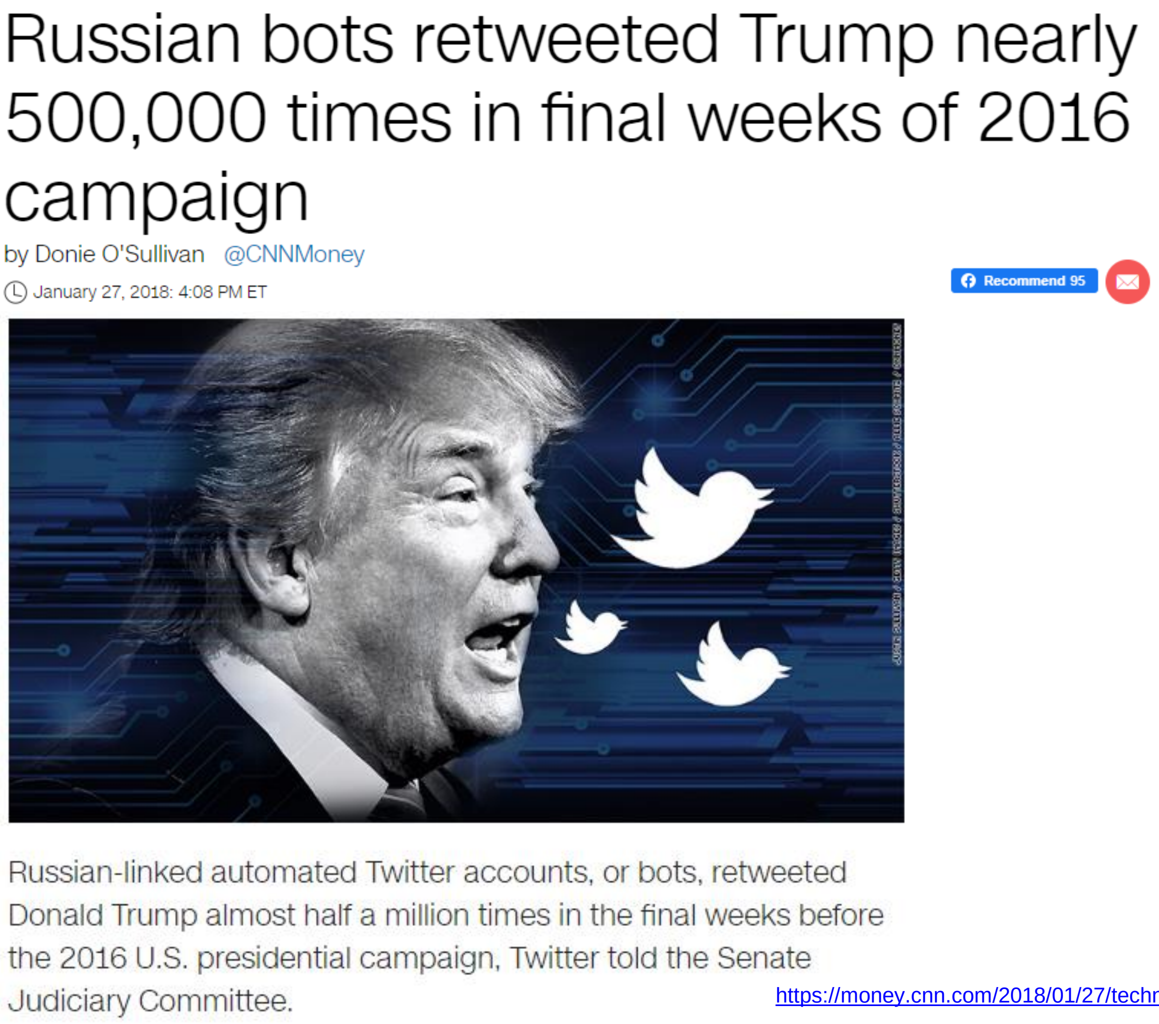
Call to Action

We are issuing a call to action to the world's artificial intelligence experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This allows the worldwide AI research community the opportunity to apply text and data mining approaches to find answers to questions within, and connect insights across, this content in support of the ongoing COVID-19 response efforts worldwide. There is a growing urgency for these approaches because of the rapid increase in coronavirus literature, making it difficult for the medical community to keep up.

A list of our initial key questions can be found under the Tasks section of this dataset. These key scientific questions are drawn from the NASEM's SCIED (National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) research topics and the World Health Organization's R&D Blueprint for COVID-19.

Many of these questions are suitable for text mining, and we encourage researchers to develop text mining tools to provide insights on these questions.

Motivation: Is this really only about business decisions (2)?



Russian bots retweeted Trump nearly 500,000 times in final weeks of 2016 campaign

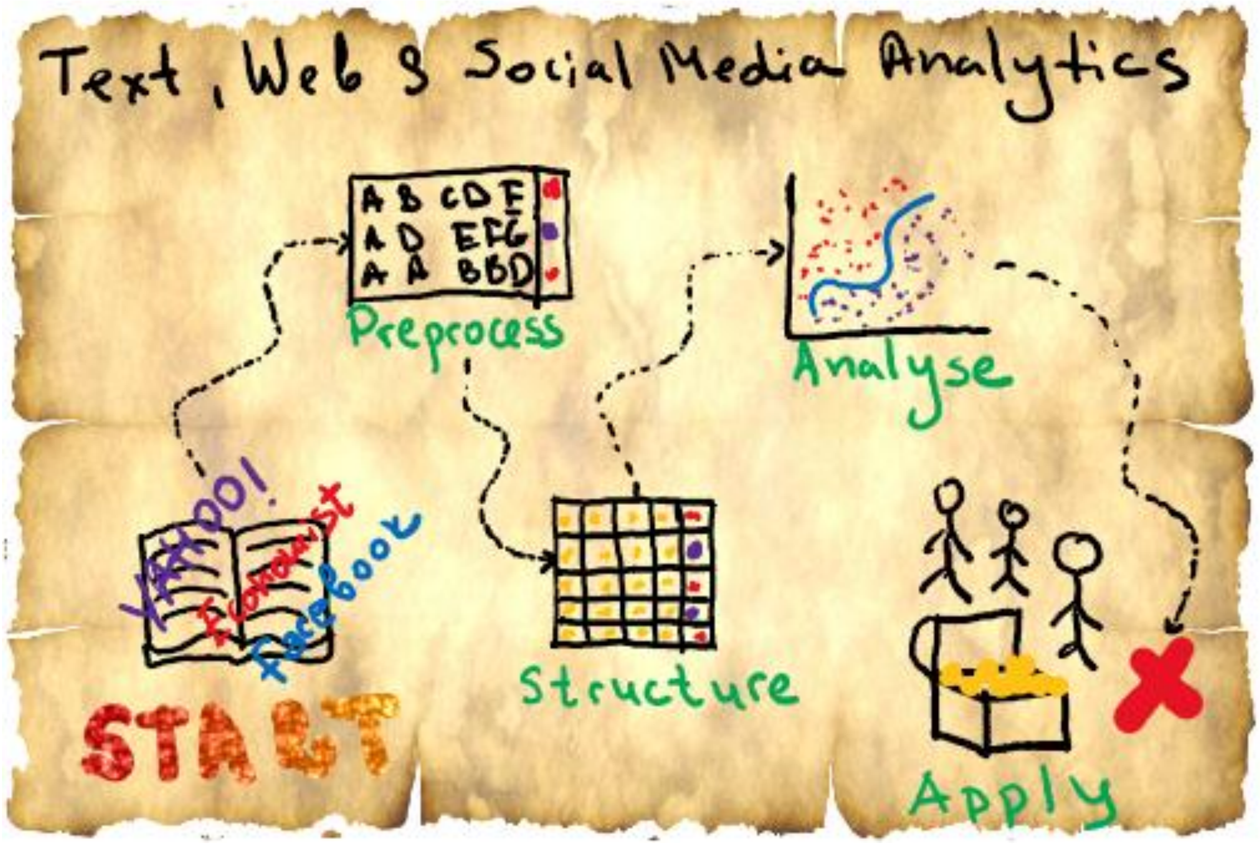
by Donie O'Sullivan @CNNMoney

January 27, 2018: 4:08 PM ET

Russian-linked automated Twitter accounts, or bots, retweeted Donald Trump almost half a million times in the final weeks before the 2016 U.S. presidential campaign, Twitter told the Senate Judiciary Committee.

https://money.cnn.com/2018/01/27/technology/business/russian-twitter-bots-election-2016/

Course structure: Treasury map



Course structure: Preliminary Timeline

Date	Lecture	Exercise
12.04.2021	Introduction	Technical Installation
19.04.2021	Text Preprocessing	Projects kick-off
26.04.2021	Text Representation	Preprocessing Newsgroups
03.05.2021	Text Representation (2)	Text Representation Newsgroups
10.05.2021	Text Classification	Text Representation Newsgroups (2)
17.05.2021	Text Clustering	Newsgroups Topic Classification
31.05.2021	Text Mining in Social Media	Newsgroups Topic Clustering
07.06.2021	Mining Social Graphs	Sentiment Analysis and Time Series in Twitter
14.06.2021	Projects Status Update	Projects Status Update
21.06.2021	Web Analytics	Mining Social Graphs in Twitter
28.06.2021	Mock Exam	Web Analytics in E-commerce
05.07.2021	Final Presentation	Final Presentation
19.07.2021	Submit Code & Written report	
t.b.a.	Exam	

	Legend
🟦	Preprocess
🟩	Structure
🟪	Analyse
🟦	Apply
⬜	Project

Final Grade

Date/ Deadline	Exam type	Weight
05.07.2021	Final Presentation	20%
19.07.2021	Group Written Report (1200 words)	40%
19.07.2021	Developed Software (Team)	
t. b. a.	Final Exam	40%

Organisation

- The course will take place Monday, 8 a.m. to approx. 11:30 a.m.
- Project groups have already been defined. Please contact your fellow students. Kick-off meetings will follow.
- Exercise sheets will be available one week in advance in Moodle and should be prepared in your group (same as project group).
- Every week, I will ask one group voluntarily to present their solution for the exercise. Please participate, even if you don’t have the perfect solution ➔ best way to learn for the exam and clarify questions.

Final wishes

- Active participation. This is a special situation for all of us, let’s do the best out of it!
- Please always feel free to provide constructive feedback. This will only improve the course.
- If you have questions or need help you could:
  - ✓ Ask the other students in your group
  - ✓ Ask your question in the Moodle-Forum
  - ✓ Write me an e-mail at diana.hristova@hwr-berlin.de
- Help your fellow students!

Exercise: Projects and Set-up

Text, Web and Social Media Analytics Lab

Prof. Dr. Diana Hrisova

Projects overview

Group	Topic	Supervisor
1	Extract relevant meta-information from unstructured & semi-structured process documents	Signavio
2	Extract process flows from unstructured & semi-structured process documents	Signavio
3	Web Analytics - Improving new Webshop	Strayz
4	Analysing user patterns in the Aam Digital application	Aam
5	Twitter analysis of mentions of talent	Bayer 04 Leverkusen
6	Social Media Analysis of discussions about Zalando	Zalando

Please contact your group members and organise a first get-to-know meeting until next session.

Data Science Projects: Tasks and Organisation

Project management:

- *Task:* makes the timeline with milestones, organizes regular meetings, writes minutes, makes sure the deadlines are obeyed, discusses pain points and tries to coordinate a solution
- *Who does that:* you should give this responsibility to one student of your group. This person should then be assigned less workload for other tasks.

Coding:

- *Tasks:* writes the code making sure that it follows standard coding principles e.g. clean, well-structured, documented, tested. We will use Python and Jupyter notebooks.
- *Who does that:* all team members, for instance by using a collaborative plattform such as github

Data Science Projects: Tasks and Organisation (2)

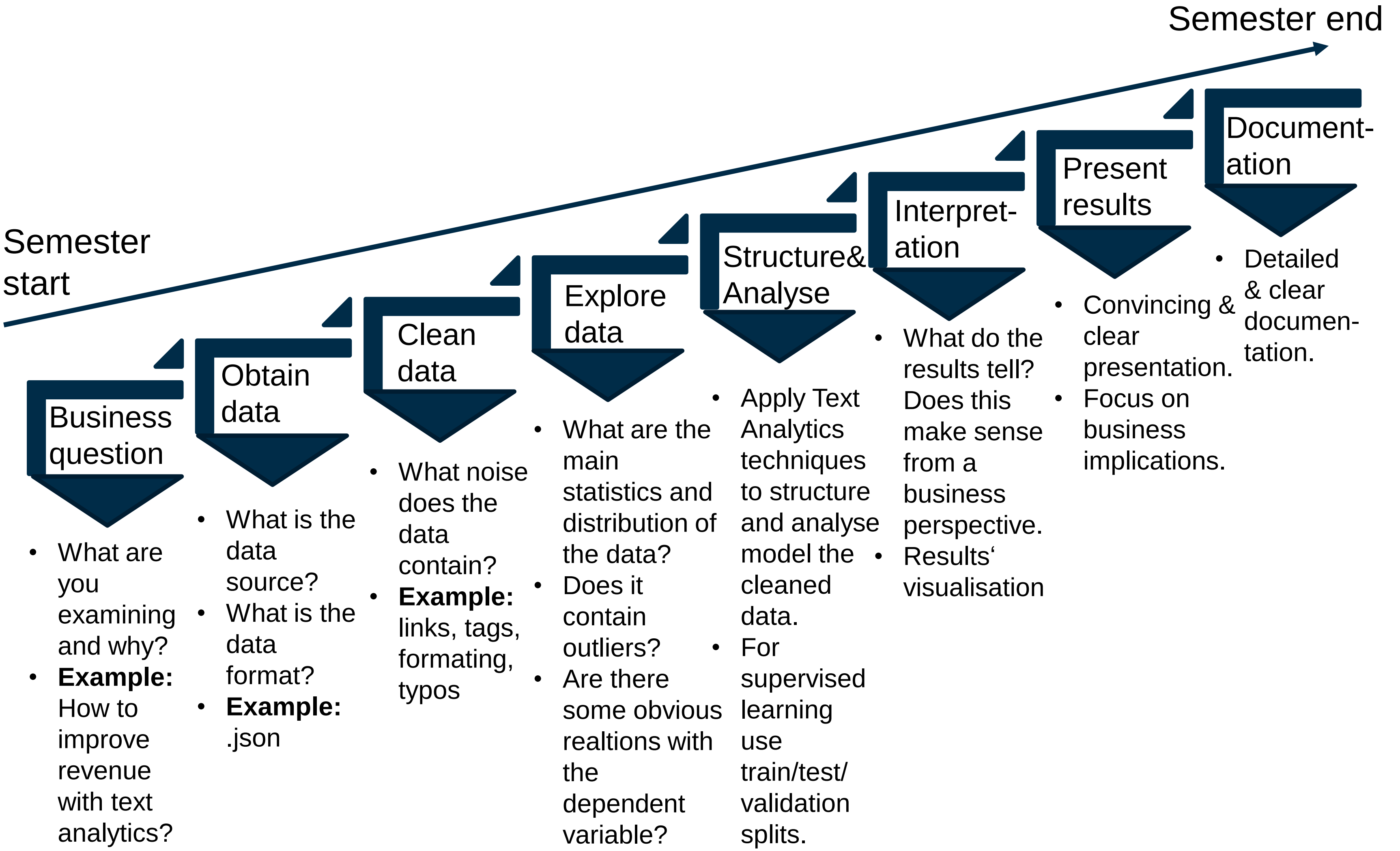
Documentation:

- *Tasks:* writes the documentation as an accompanying process to the development, NOT all at the end. The documentation should be precise and detailed.
- *Who does that:* all team members after finishing a task. One team member should be assigned the task of aggregating the separate documents.

Supervision:

- *Tasks:* meets with the team on a regular basis and discussed their progress and problems.
- *Who does that:* see the table on slide 13.

Data Science Projects: Milestones



Semester start

Semester end

Business question

- What are you examining and why?
- **Example:** How to improve revenue with text analytics?

Obtain data

- What is the data source?
- What is the data format?
- **Example:** .json

Clean data

- What noise does the data contain?
- **Example:** links, tags, formating, typos

Explore data

- What are the main statistics and distribution of the data?
- Does it contain outliers?
- Are there some obvious realtions with the dependent variable?

Structure& Analyse

- Apply Text Analytics techniques to structure and analyse model the cleaned data.
- For supervised learning use train/test/ validation splits.

Interpret-ation

- What do the results tell? Does this make sense from a business perspective.
- Results‘ visualisation

Present results

- Convincing & clear presentation.
- Focus on business implications.

Document-ation

- Detailed & clear documen-tation.

Survey

How far did you get with Exercise 1?

a. Finished Question 1 (Google Colab)

b. Finished Question 2 a. (Twitter Developer account)

c. Finished Question 2 b. (Twitter App)

d. Finished Question 2 d. (Notebook runs)

e. None

Exercise 1

Please choose the break-out room for the question you couldn‘t accomplish.

a. Room1= Question 1 (Google Colab)

b. Room2 = Question 2 a. (Twitter Developer account)

c. Room3 = Question 2 b. (Twitter App)

d. Room4 = Question 2 d. (Notebook runs)