Data Handling: Import, Cleaning and Visualisation

Lecture 1 :

Introduction

Prof. Dr. Ulrich Matter

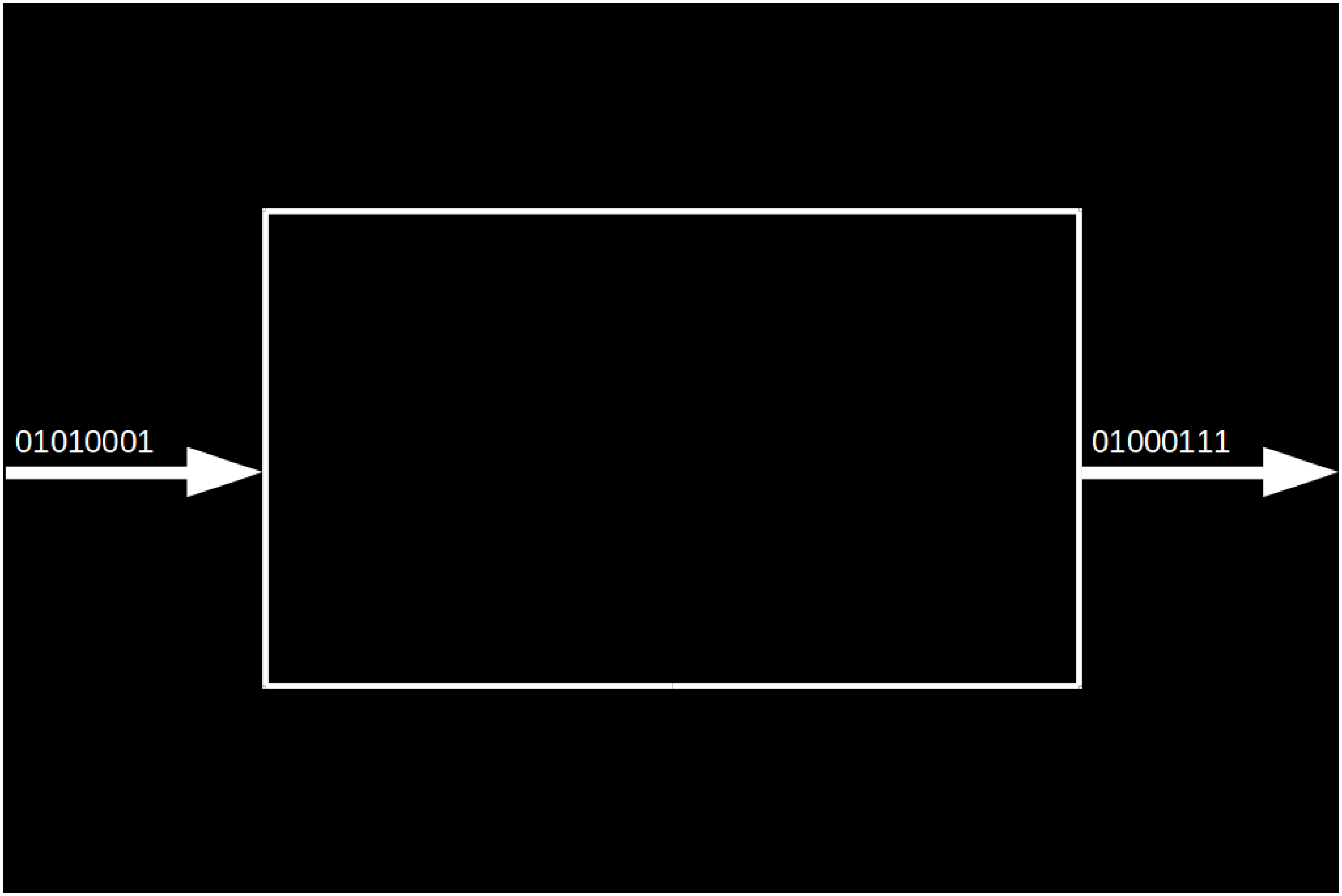
17/09/2020

Welcome to Data Handling: I.C.V. 2020!

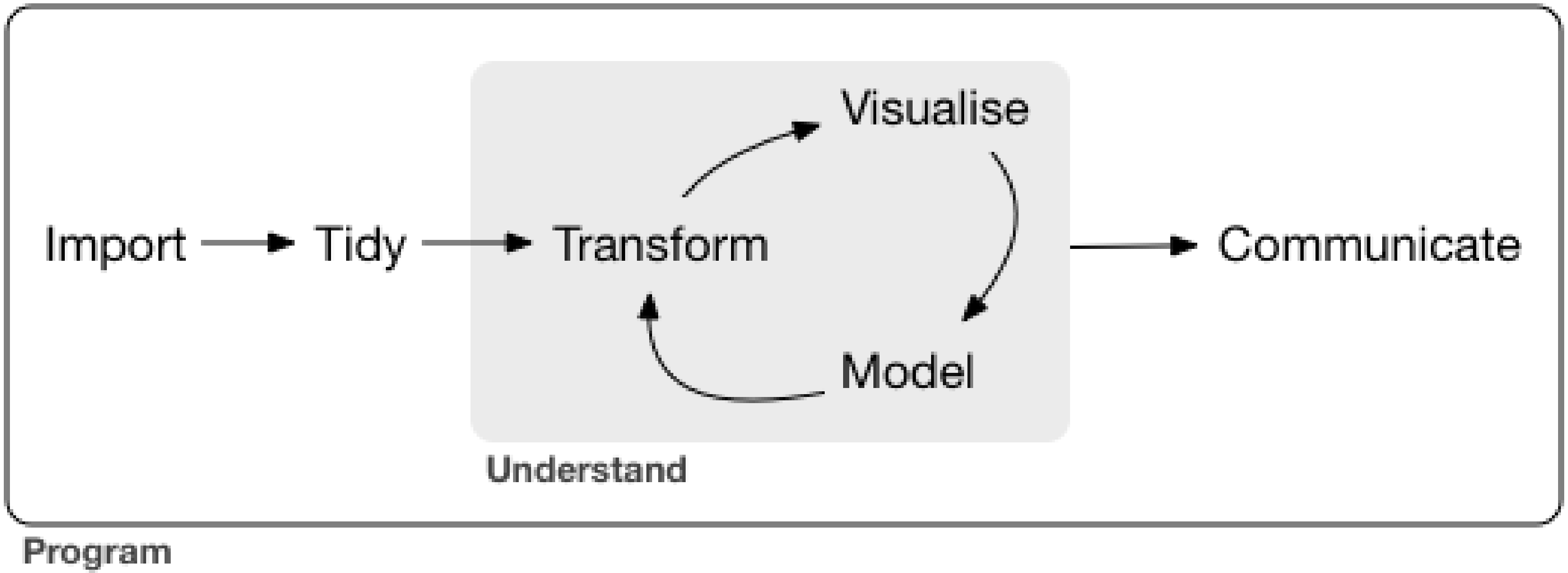
- Fire up your notebooks!
- Go to this page: http://bit.ly/datahandling-2020
- Use one row to respond to the questions in the column headers (see the first two rows for examples).

Introductory Example

Data input, processing, output

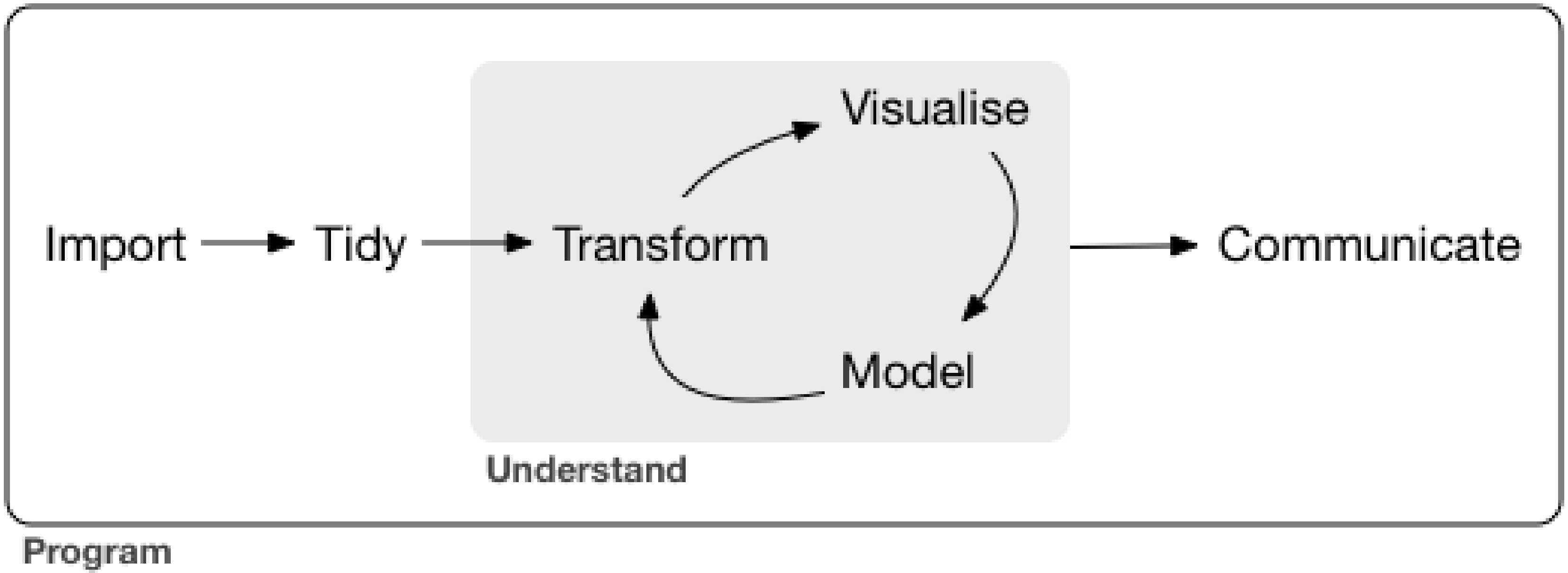


The Data Pipeline



Data Science workflow. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

The Data Pipeline



Data Science workflow. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

What could be the **output** of all this?

The Data Pipeline

- Research report/paper (e.g., BA Thesis)
- Presentation/Slides
- Website
- Web application (interactive; alas the introductory example)
- Dashboard for management
- Recommender system (i.e., a trained machine learning algorithm)
- …

‘Data Science’?

‘Data Science’?

“This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and inter-disciplinary applications.”

University of Michigan ‘Data Science Initiative’, 2015

But, what about statistics?!

“Seemingly, statistics is being marginalized here; the implicit message is that statistics is a part of what goes on in data science but not a very big part. At the same time, many of the concrete descriptions of what the DSI will actually do will seem to statisticians to be bread-and-butter statistics. Statistics is apparently the word that dare not speak its name in connection with such an initiative!”

David Donoho (2015). 50 years of Data Science

Background

What’s new about all this?

“All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: …”

What’s new about all this?

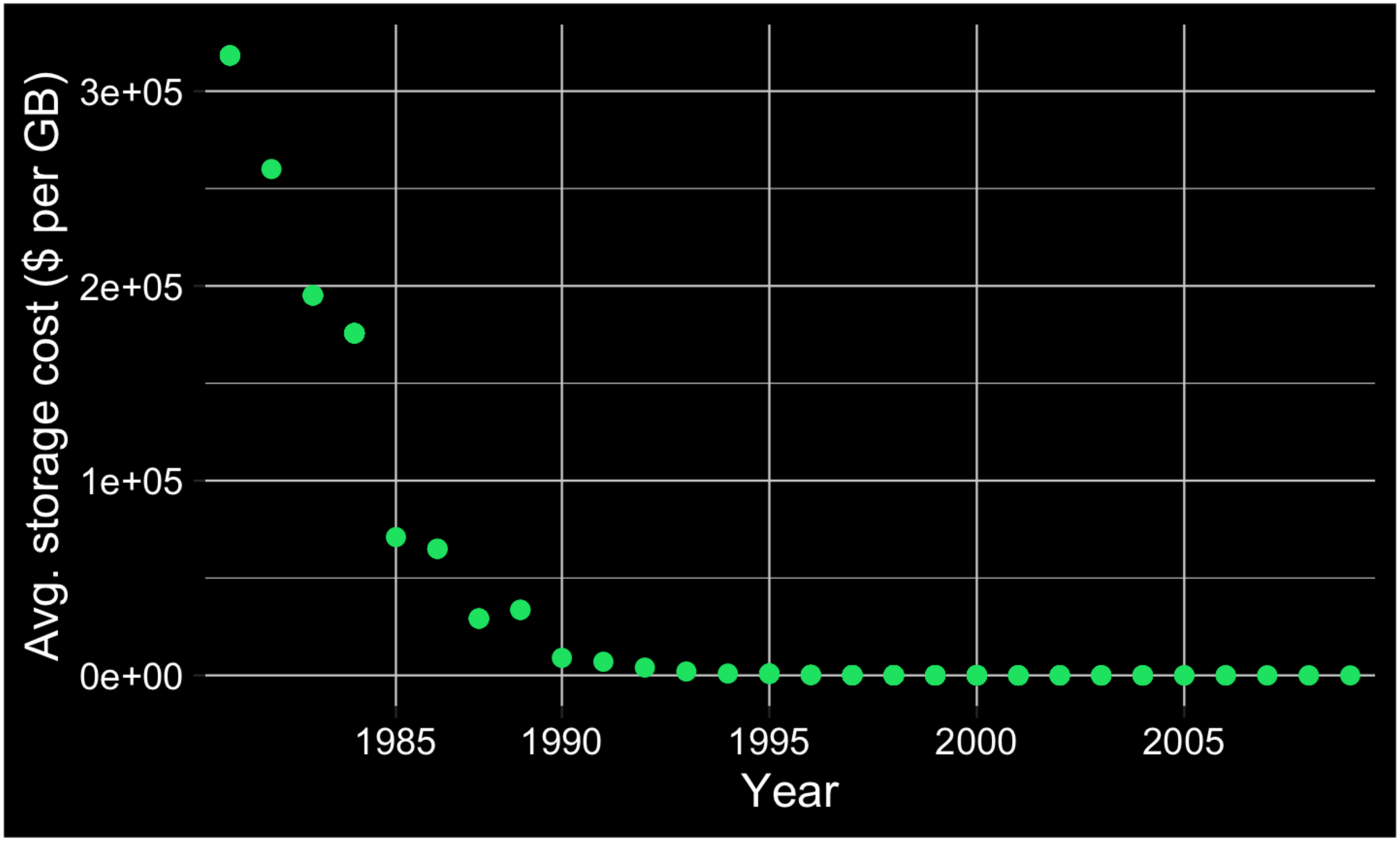
“All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

What's new about all this?

John Tukey (The Future of Data Analysis, 1962!)

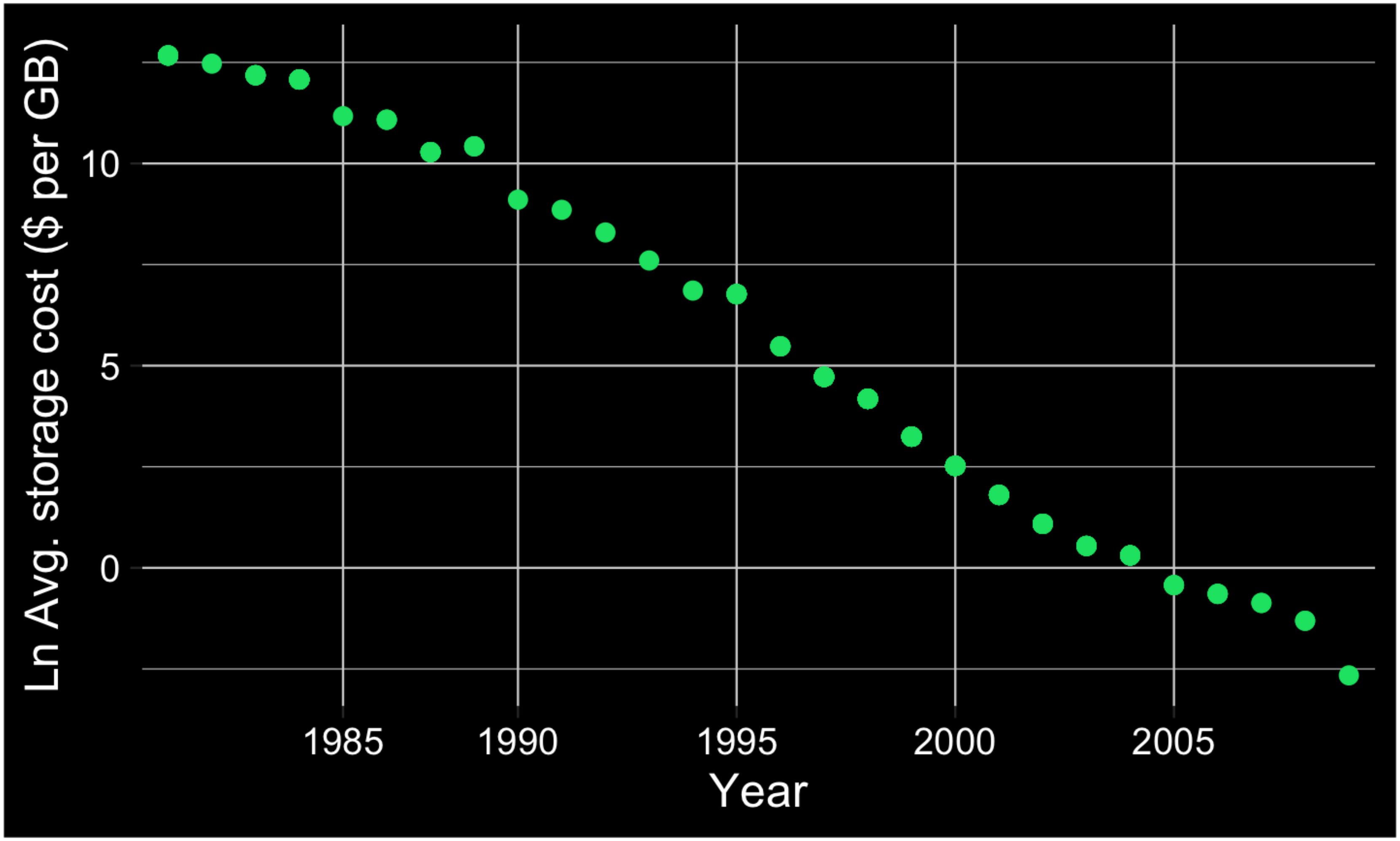
Technological change

Technological change

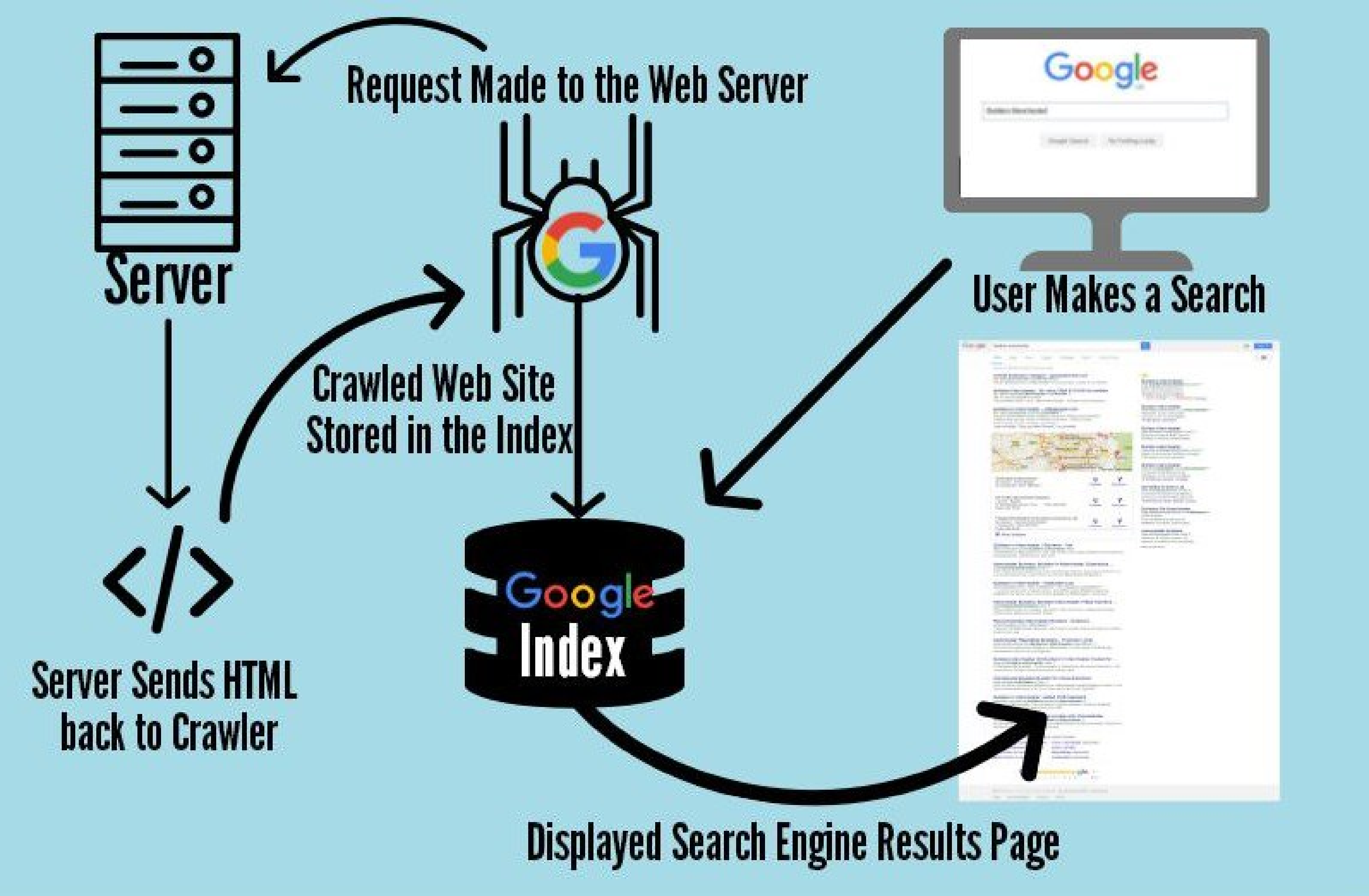


Data source: http://www.mkomo.com/cost-per-gigabyte

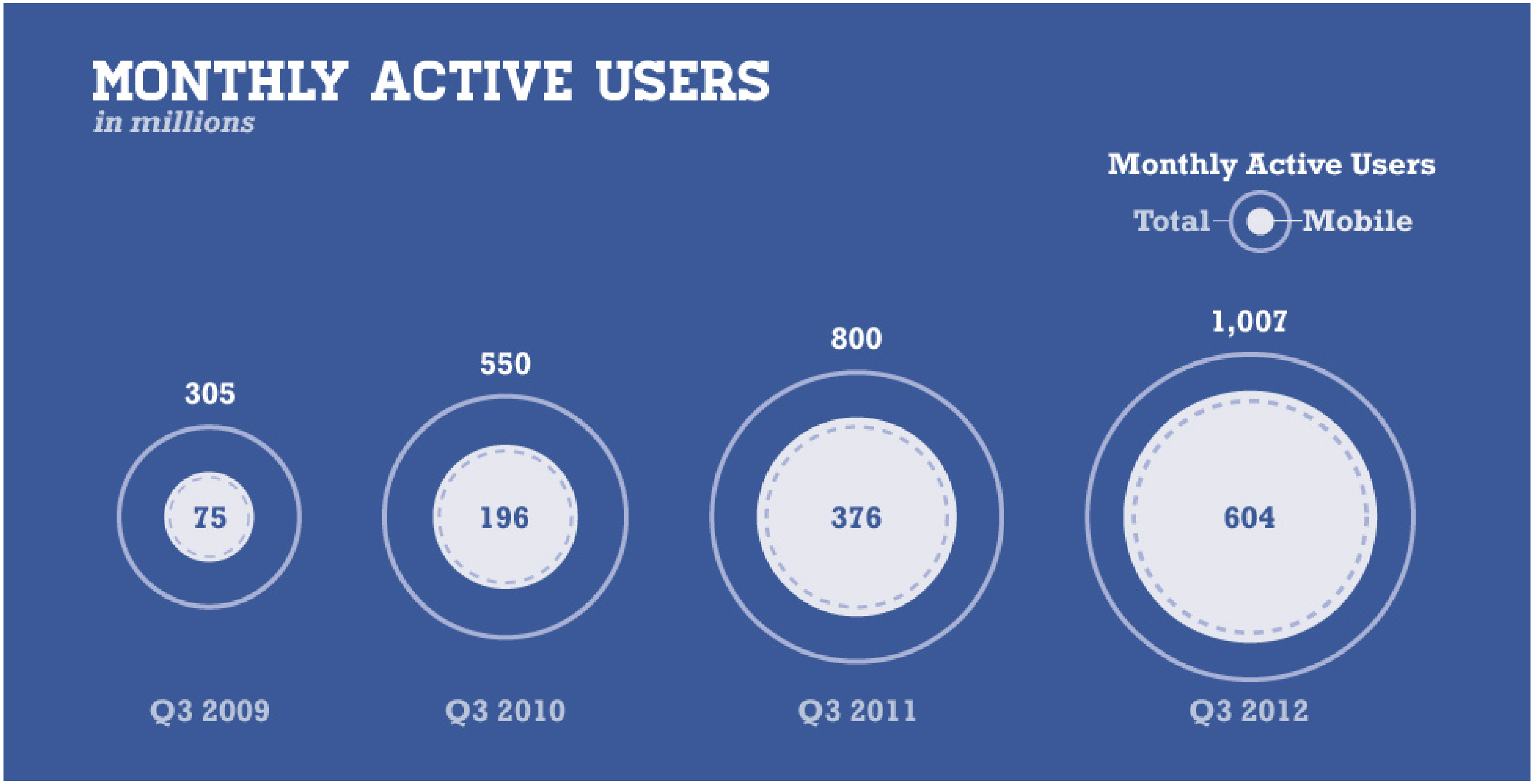
Technological change



Data source: http://www.mkomo.com/cost-per-gigabyte



Source: https://techxerl.net.



Source: statista.com.

The Economist

Theresa May v Brussels

Ten years on: banking after the crisis

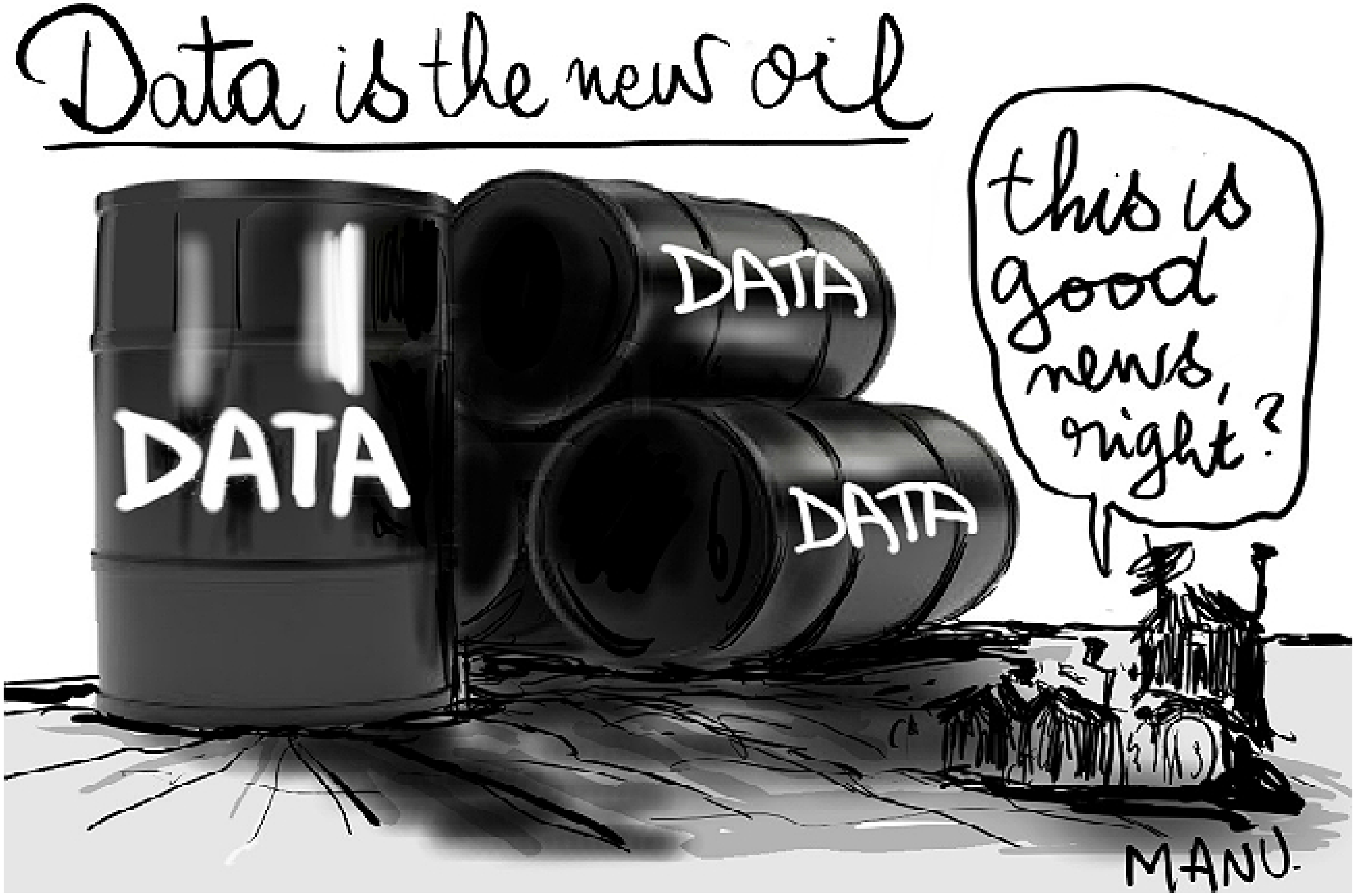
South Korea's unfinished revolution

Biology, but without the cells

MAY 6TH–12TH 2017

The world's most valuable resource

Data and the new rules of competition


Data is the new oil
DATA
DATA
DATA
this is good news, right?
MANU.

The AI Revolution Is Remaking Every Business in Every Industry

There is no typecast for savvy AI businesses. They come in all sizes and represent an ever broadening swath of industry. Simply put, the era of artificial intelligence is remaking business as we know it.

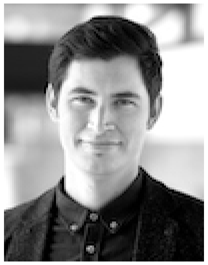
Businesses see AI as a long-term strategic priority. In a recent survey from Infosys, three-quarters of the respondents from large, multinational corporations cited AI as fundamental to the success of their organization's strategy. Sixty-four percent believe that their organization's growth is dependent on large-scale AI adoption.

The main challenge is in figuring out how best to put AI to work. There is no universal answer. That was clear from the hundreds of companies that participated at our GPU Technology Conference last month. And it's evident again at the O'Reilly AI conference this week in New York. Much like GTC, the conference draws thousands of participants in every industry, from startups to massive enterprises.

Organization of the Course

Our Team - At Your Service

Philine Widmer



Ulrich Matter

HELP
WANTED

Course Structure

Course concept

- Lectures (Thursday morning)
    - Background/Concepts
    - Live demonstrations of concepts
    - Illustration of ‘hands-on’ approaches

Course concept

- Lectures (Thursday morning)
    - Background/Concepts
    - Live demonstrations of concepts
    - Illustration of ‘hands-on’ approaches
- Workshops/Exercises (bi-weekly evening sessions)
    - Guided tutorials
    - Discussion of homework exercises
    - Recap of theoretical concepts

Course concept

- Lectures (every Thursday morning)
  - Background/Concepts
  - Live demonstrations of concepts
  - Illustration of ‘hands-on’ approaches
- Workshops/Exercises (bi-weekly evening sessions)
  - Guided tutorials
  - Discussion of homework exercises
  - Recap of theoretical concepts
  - **First Exercises (set up R/RStudio) is available on StudyNet/Canvas today**

Course concept

- Lectures (every Thursday morning)
  - Background/Concepts
  - Live demonstrations of concepts
  - Illustration of ‘hands-on’ approaches
- Workshops/Exercises (bi-weekly evening sessions)
  - Guided tutorials
  - Discussion of homework exercises
  - Recap of theoretical concepts
  - **First Exercises (set up R/RStudio) is available on StudyNet/Canvas today**
- Guest lecture and research insights

Course concept

- Strongly encouraged: (virtual) learning groups!
    - Biweekly exercises provide opportunity.
    - Tackle the tricky exercises together!

Part I: Data (Science) fundamentals

Date	Topic
17.09.20	Introduction: Big Data/Data Science, course overview
24.09.20	An introduction to data and data processing
24.09.20	Exercises/Workshop 1: Tools, working with text files
01.10.20	Data storage and data structures
08.10.20	’Big Data‘ from the Web
08.10.20	Exercises/Workshop 2: Computer code and data storage
15.10.20	Programming with data

Part II: Data gathering and preparation

Date	Topic
22.10.20	Research Insights
22.10.20	Exercises/Workshop 3: Programming with Data
29.10.20	Semester Break
05.11.20	Semester Break
12.11.20	Data sources, data gathering, data import
19.11.20	Data preparation and manipulation
19.11.20	Exercises/Workshop 4: Data import and data preparation/manipulation

Part III: Analysis, visualisation, output

Date	Topic
26.11.20	Guest Lecture
03.12.20	Basic statistics and data analysis with R
03.12.20	Exercises/Workshop 5: Applied data analysis with R
10.12.20	Visualisation, dynamic documents
17.12.20	Summary, Wrap-Up, Q&A, Feedback
17.12.20	Exercises/Workshop 6: Visualization, dynamic documents
18.12.20	Exam for Exchange Students

Core course resources

- All information and materials (notes, slides, course sheet, syllabus, etc.) available on StudyNet/Canvas.
- Exercises will be uploaded to Assignments in StudyNet/Canvas!
- This course is **open souce**: all raw materials (code, source code for slides, notes, etc.) are freely available on GitHub

Main textbooks

Murrell, Paul (2009). **Introduction to Data Technologies**, London: Chapman & Hall/CRC.

Wickham, Hadley and Garred Grolemund (2017). **R for Data Science**, 1st Edition. Sebastopol, CA: O'Reilly.

Further resources

- Stackoverflow
- Get inspired in the R blogsphere

Exam information

- Central, written examination.
- Multiple choice questions.
- A few open questions.
- Theoretical concepts and practical applications in R (questions based on code examples).

Exam information II

- Exercises towards the end of the term will contain sample questions.
  - Get familiar with the style/format of questions.
- Exchange students who need to take the exam before the central exam block:
  - Notify the course TA until the end of September: **philine.widmer@unisg.ch**!
  - Decentral exam for exchange students: **18 December 2020**.

Q&A

References

Wickham, Hadley, and Garrett Grolemund. 2017. Sebastopol, CA: O’Reilly. http://r4ds.had.co.nz/.