Data Handling: Import, Cleaning and Visualisation

Lecture 1 :

Introduction

Prof. Dr. Ulrich Matter

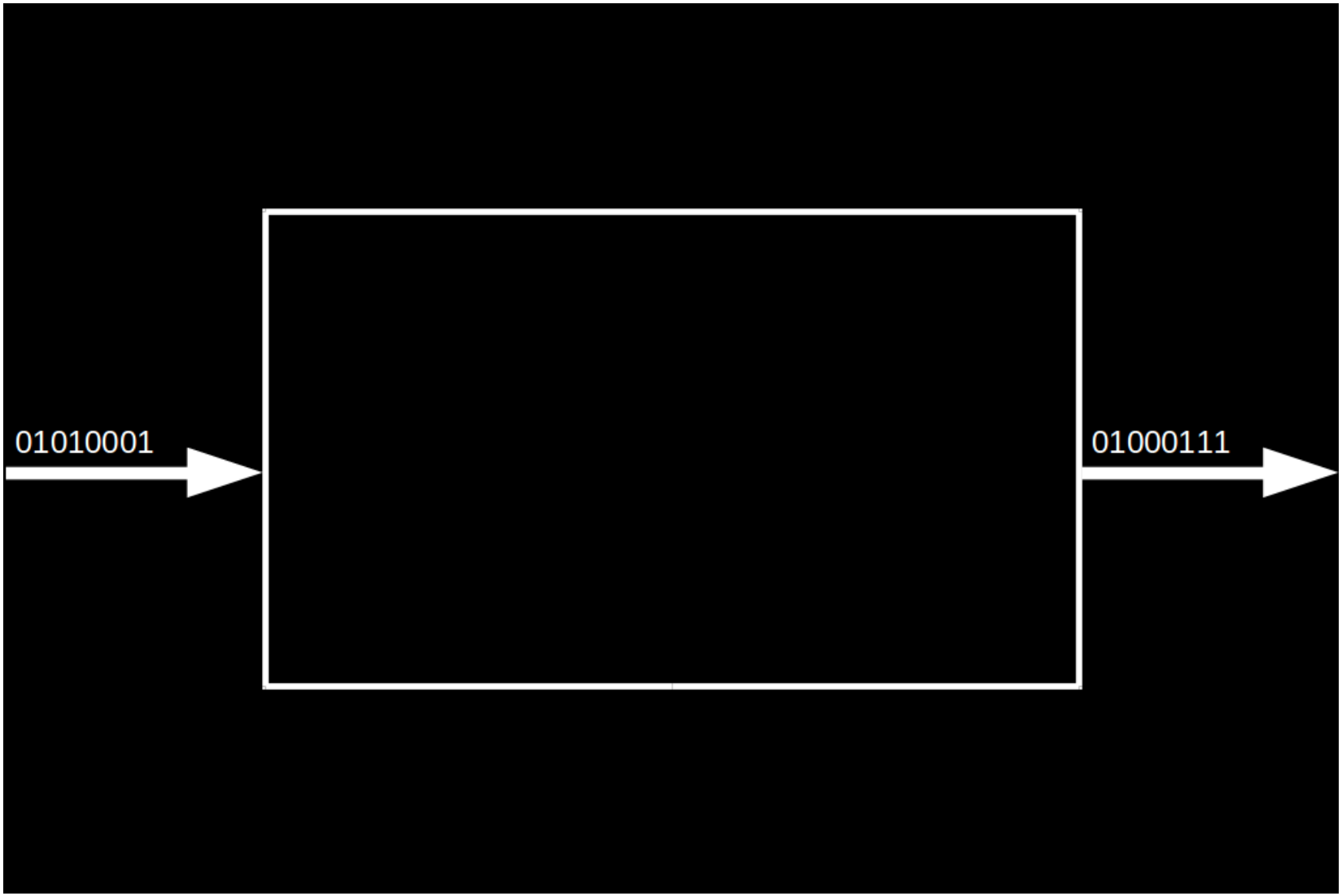
19/09/2019

Welcome to Data Handling: I.C.V. 2019!

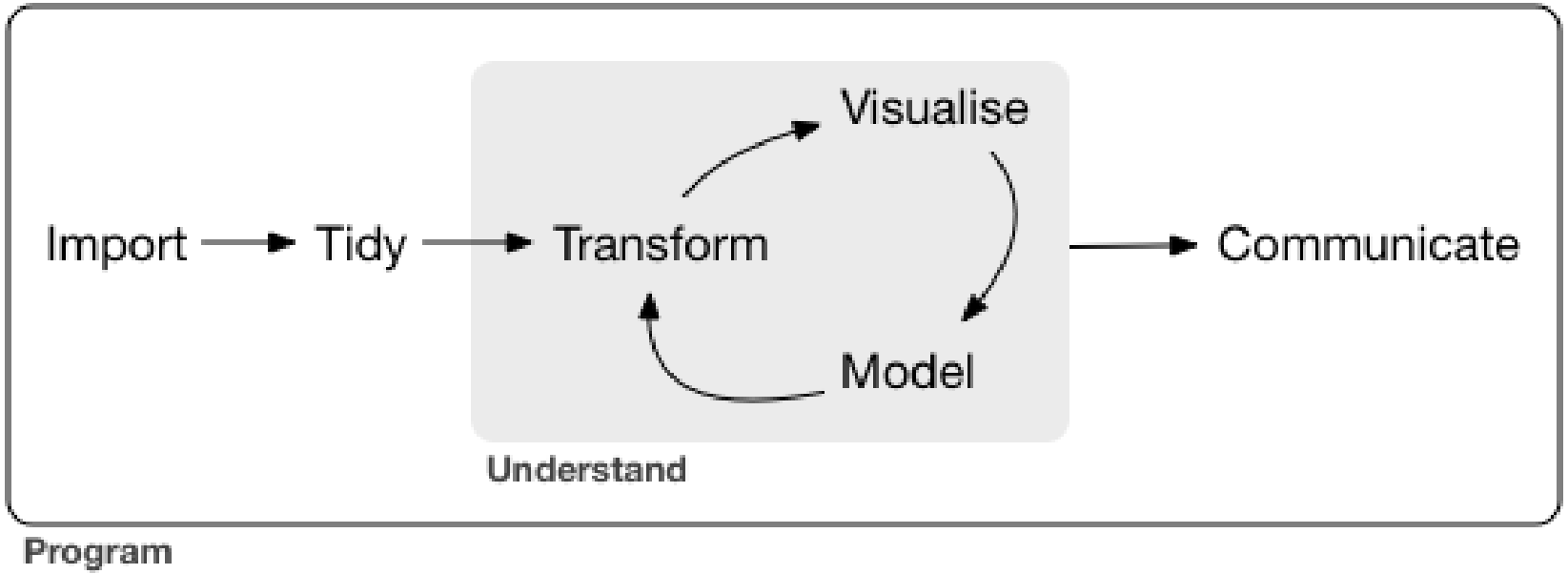
- Fire up your notebooks!
- Go to this page: http://bit.ly/datahandling-2019
- Use one row to respond to the questions in the column headers (see the first two rows for examples).

Introductory Example

Data input, processing, output

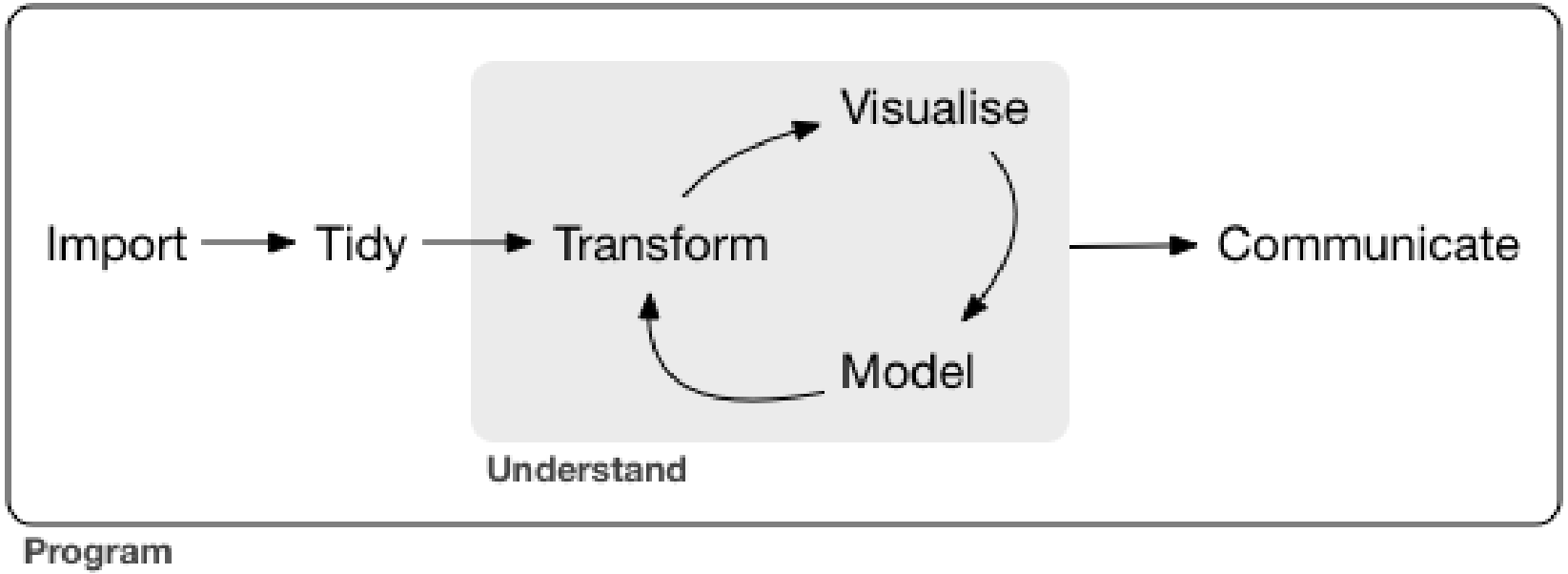


The Data Pipeline



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?

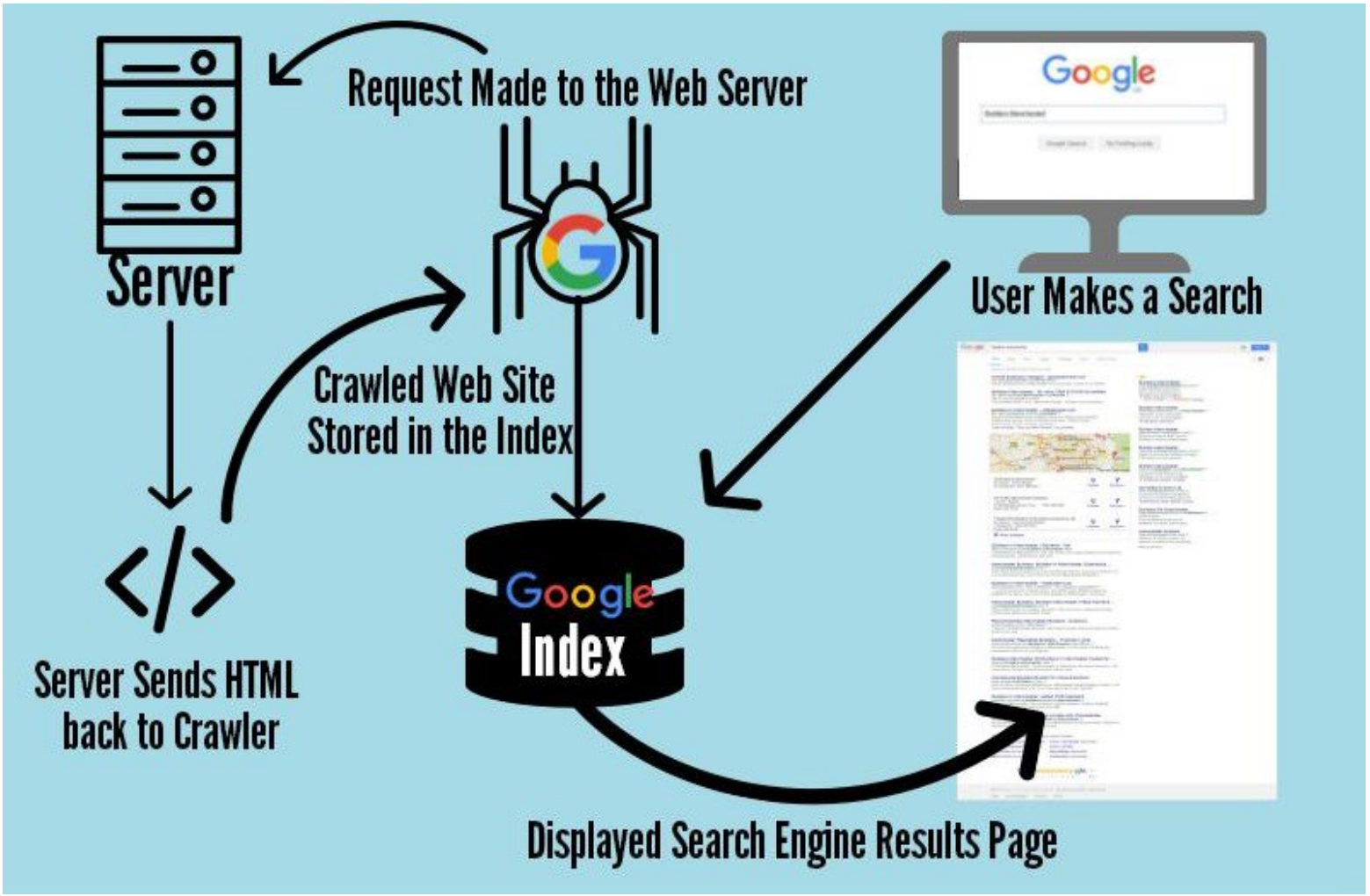
"All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more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

What's new about all this?

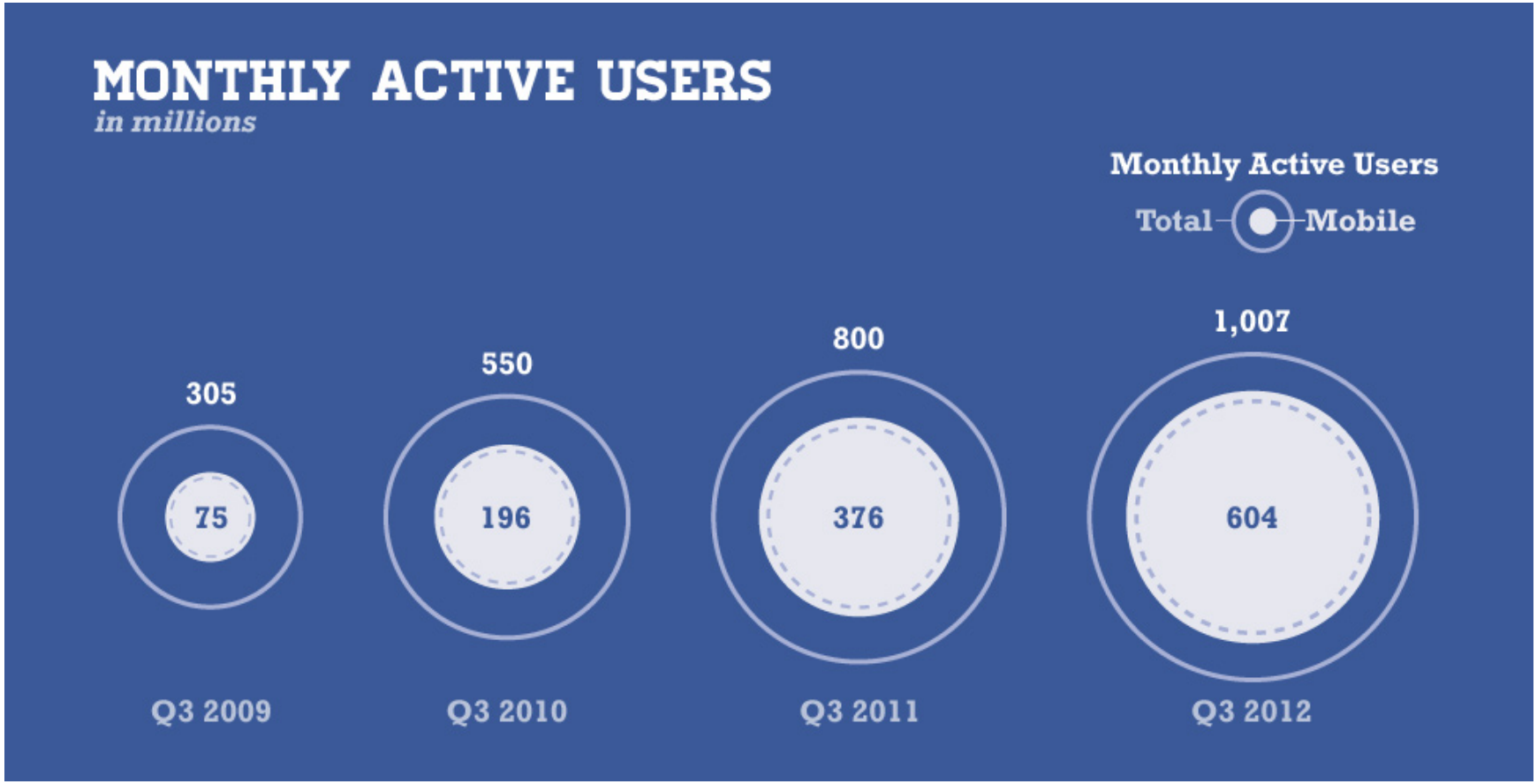
John Tukey (**The Future of Data Analysis**, 1962!)

Technological change

Technological change

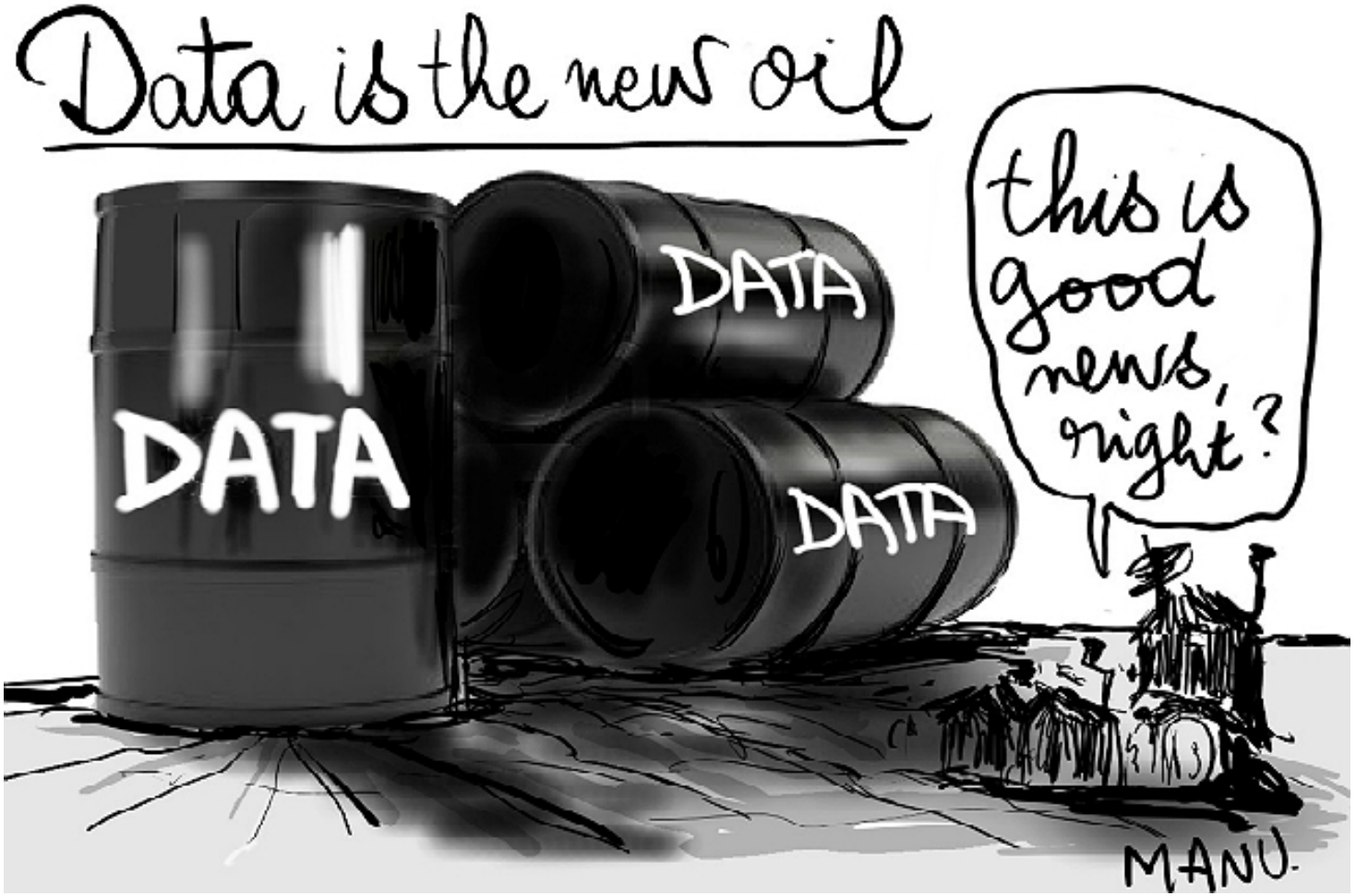


Source: https://techxerl.net.



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
amazon
Google
Data and the new rules of competition

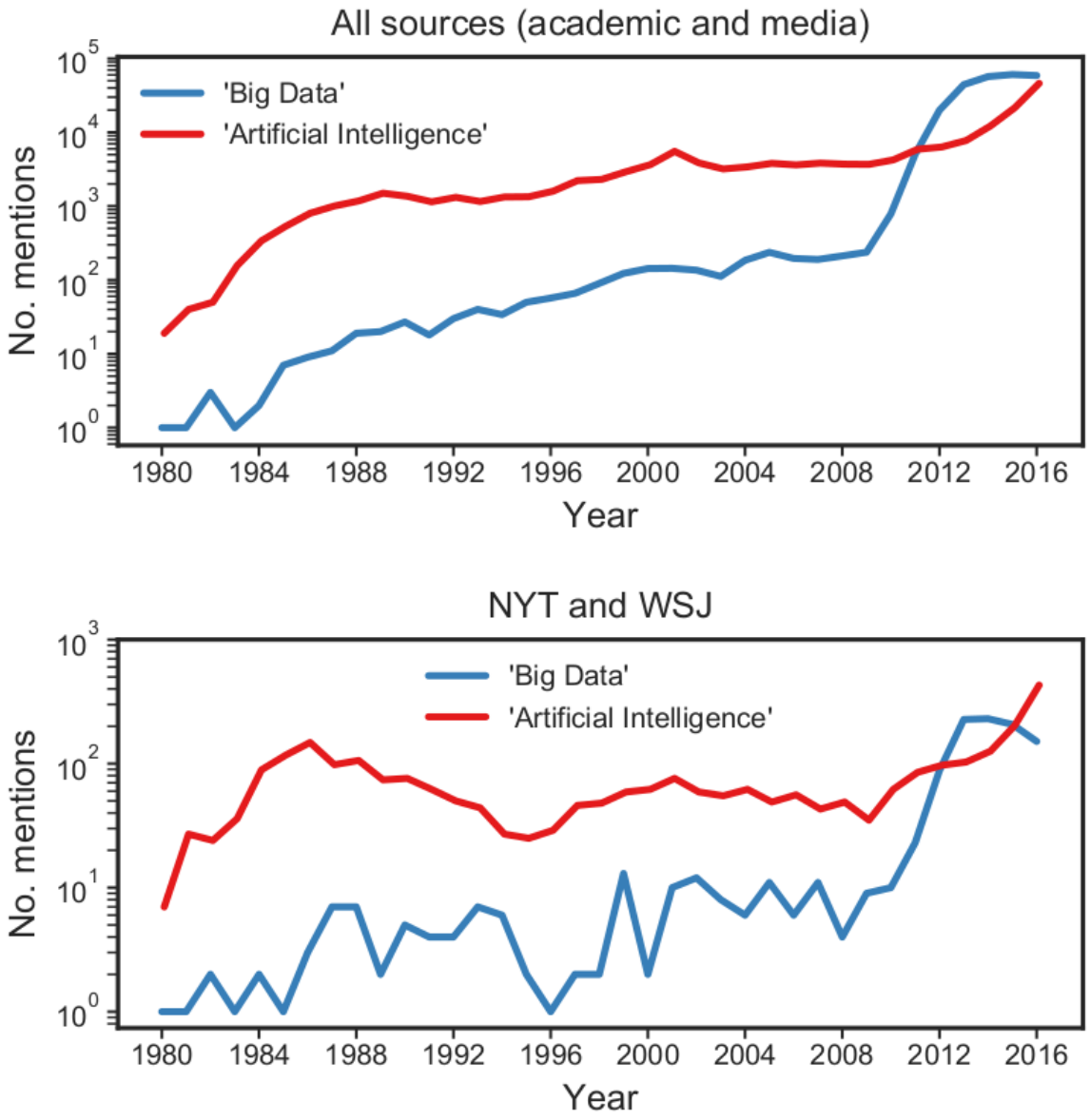

Data is the new oil
DATA
DATA
DATA
this is good news, right?
MANU.

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.


All sources (academic and media)
'Big Data'
'Artificial Intelligence'
No. mentions
Year
NYT and WSJ
'Big Data'
'Artificial Intelligence'
No. mentions
Year

Organization of the Course

Our Team - At Your Service

Mirjam Bächli

Philine Widmer

Ulrich Matter


HELP
WANTED

Help wanted

- Experienced R user?
- Assist fellow students during exercises in class
- Disclaimer: this is not an official TA position!

Help wanted

- Experienced R user?
- Assist fellow students during exercises in class
- Disclaimer: this is not an official TA position!
- **Interested?**
  - Approach us at the end of today's lecture!
  - Or send us an email: philine.widmer@unisg.ch

Course Structure

Course concept

- Lectures (every 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
  - (Guest Lecture)

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 Lectures

Course concept

- Strongly encouraged: Learning groups!
  - Workshops/Exercises-Sessions will provide opportunity.
  - Tackle the tricky exercises together!

Part I: Data (Science) fundamentals

Date	Topic
19.09.19	Introduction: Big Data/Data Science, course overview
26.09.19	An introduction to data and data processing
26.09.19	Exercises/Workshop 1: Tools, working with text files
03.10.19	Data storage and data structures
10.10.19	'Big Data' from the Web
10.10.19	Exercises/Workshop 2: Computer code and data storage
17.10.19	Programming with data

Part II: Data gathering and preparation

Date	Topic
24.10.19	No lecture in the morning
24.10.19	Exercises/Workshop 3: Programming with Data
31.10.19	Semester Break
07.11.19	Semester Break
14.11.19	Data sources, data gathering, data import
21.11.19	Data preparation and manipulation
21.11.19	Exercises/Workshop 4: Data import and data preparation/manipulation

Part III: Analysis, visualisation, output

Date	Topic
28.11.19	Guest Lecture or Research Insights
05.12.19	Basic statistics and data analysis with R
05.12.19	Exercises/Workshop 5: Applied data analysis with R
12.12.19	Visualisation, dynamic documents
19.12.19	Research Insights/Summary, Wrap-Up, Q&A
19.12.19	Exercises/Workshop 6: Visualization, dynamic documents;
20.12.19	Exam for Exchange Students

Core course resources

- Course materials: umatter.github.io/courses
- All information (course sheet, syllabus, etc.) always also available on StudyNet/Canvas.
- Exercises will be uploaded to Assignments in StudyNet/Canvas!

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's until the end of September: **philine.widmer@unisg.ch**, **mirjam.baechli@unisg.ch**!
  - Decentral exam for exchange students: **19 December 2019**.

Q&A

References

Katz, Yarden. 2017. “Manufacturing an Artificial Intelligence Revolution.” https://ssrn.com/abstract=3078224.

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