Data Handling: Import, Cleaning and Visualisation

Lecture 1 :

Introduction

Prof. Dr. Ulrich Matter

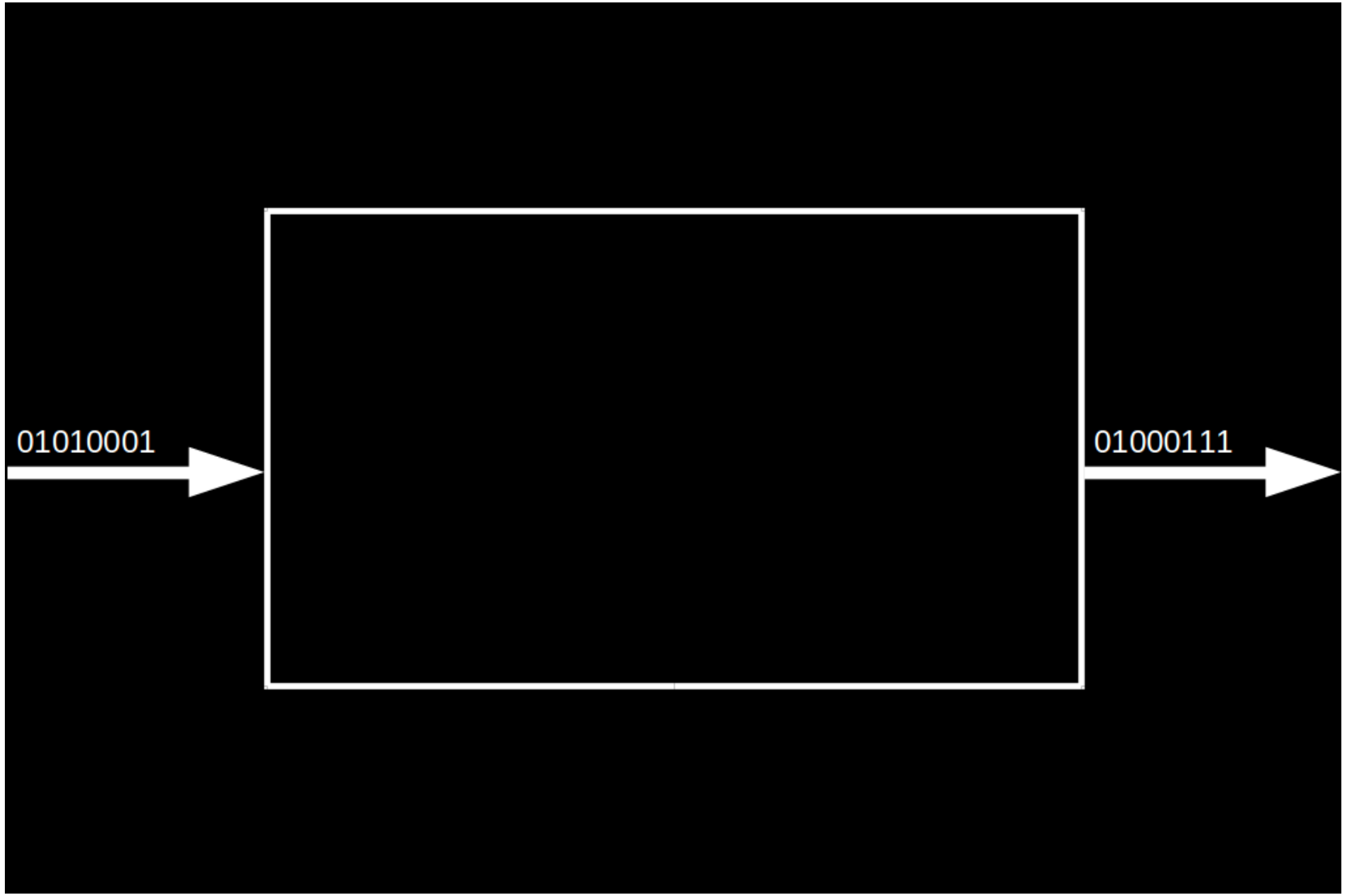
23/09/2021

Welcome to Data Handling: I.C.V. 2021!

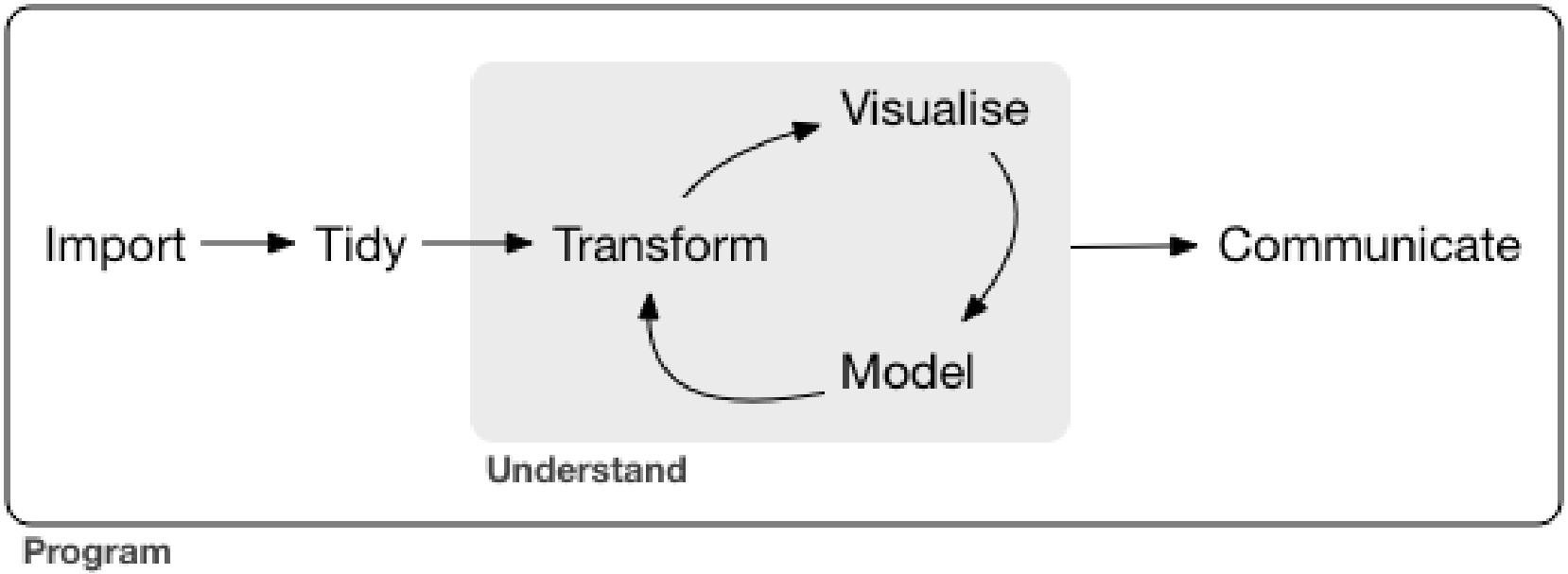
- Fire up your notebooks!
- Go to this page: http://bit.ly/datahandling-2021
- Use one row to respond to the questions in the column headers (see the first two rows for examples).

Introductory Example

Data input, processing, output

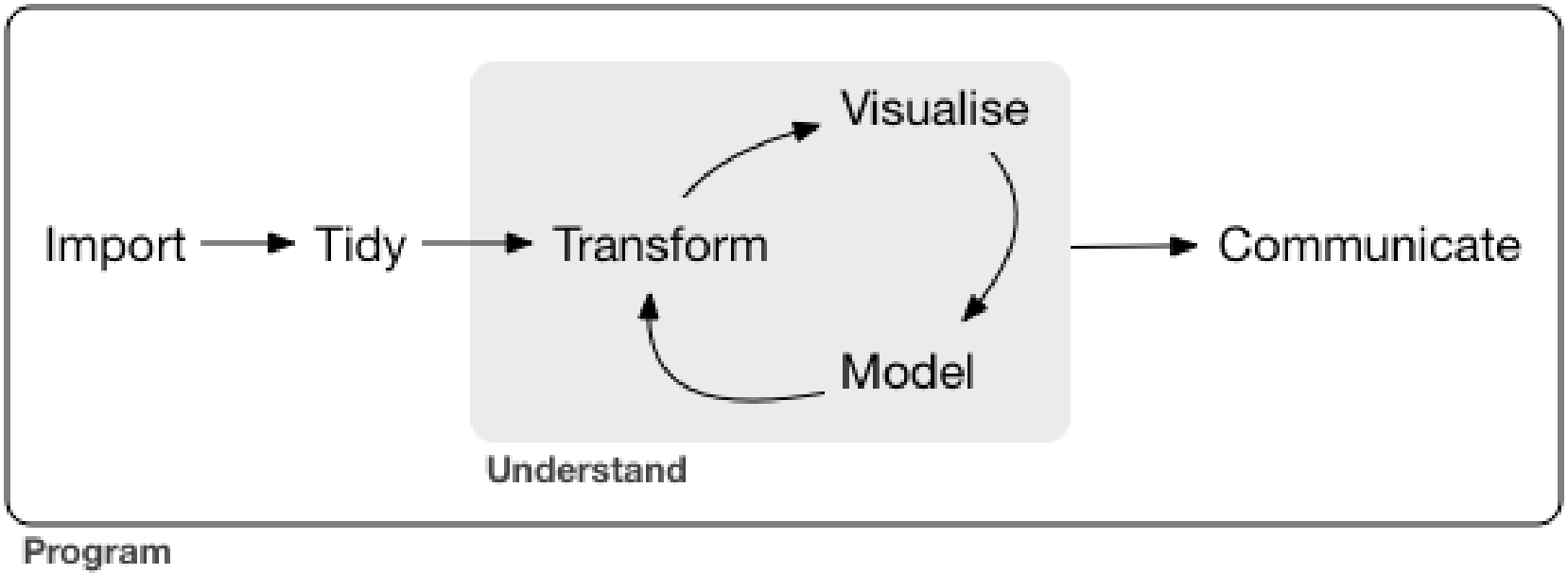


The Data Pipeline



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

Relevance for modern economic research

SOCIAL SCIENCE

Computational Social Science

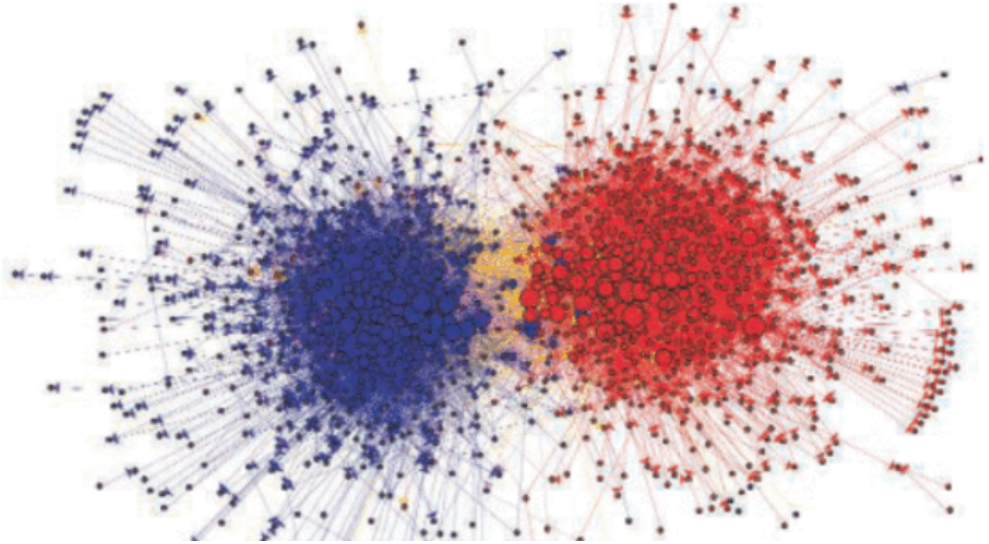
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the



Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 189–206

Using Internet Data for Economic Research

Benjamin Edelman

The data used by economists can be broadly divided into two categories. First, structured datasets arise when a government agency, trade association, or company can justify the expense of assembling records. The Internet has transformed how economists interact with these datasets by lowering the cost of storing, updating, distributing, finding, and retrieving this information. Second, some economic researchers affirmatively collect data of interest. Historically, assembling a dataset might involve delving through annual reports or archives that had not previously been organized into a format ready for research. In some cases researchers would survey stores, factories, consumers, or workers; or they could carry out an experiment. For researcher-collected data, the Internet opens exceptional

Relevance for modern economic research

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 3–28

Big Data: New Tricks for Econometrics†

Hal R. Varian

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools.

Relevance for modern economic research

Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

Text as Data†

Matthew Gentzkow, Bryan Kelly, and Matt Taddy*

An ever-increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications. (*JEL* C38, C55, L82, Z13)

Economy/Business

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.

Organisation of the Course

Our Team - At Your Service

Aurélien Sallin

Michael Tüting

Ulrich Matter

Course Structure

Course concept

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

Course components I

- Lectures (Thursday morning)
  - Background/Concepts
  - Live demonstrations of concepts
  - Illustration of ‘hands-on’ approaches
- Exercises (handed out every other week)
  - Some conceptual questions (as they appear in the exam)
  - Hands-on exercises/tutorials in R
  - Detailed solution videos
  - **First Exercises (set up R/RStudio) is available on StudyNet/Canvas today**

Course components II

- Workshops/Exercises (bi-weekly evening sessions)
  - Discussion of exercises and additional input
  - Recap of theoretical concepts
  - Q&A, support
- Guest lecture and research insights

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

- Learning mode in this course: Visit the lecture, recap key concepts in lecture notes (self-study), work on exercises, watch solution video, come to exercise session, repeat…
- Strongly encouraged: (virtual) learning groups!
  - Biweekly exercises provide opportunity.
  - Tackle the tricky exercises together!

Part I: Data (Science) fundamentals

Date	Topic
23.09.2021	Introduction: Big Data/Data Science, course overview
30.09.2021	An introduction to data and data processing
30.09.2021	Exercises/Workshop 1: Tools, working with text files
7.10.2021	Data storage and data structures
14.10.2021	Big Data from the Web
14.10.2021	Exercises/Workshop 2: Computer code and data storage
21.10.2021	Programming with data

Part II: Data gathering and preparation

Date	Topic
28.10.2021	Research insights
28.10.2021	Exercises/Workshop 3: Programming with Data
NA	Semester Break
NA	Semester Break
18.11.2021	Data sources, data gathering, data import
25.11.2021	Data preparation and manipulation
25.11.2021	Exercises/Workshop 4: Data import and data preparation/manipulation

Part III: Analysis, visualisation, output

Date	Topic
02.12.2021	Guest Lecture
09.12.2021	Basic statistics and data analysis with R
09.12.2021	Exercises/Workshop 5: Applied data analysis with R
16.12.2021	Visualisation, dynamic documents
23.12.2021	Summary, Wrap-Up, Q&A, Feedback
23.12.2021	Exercises/Workshop 6: Visualization, dynamic documents
24.12.2021	Exam for Exchange Students

Core course resources

- All information and materials (notes, slides, course sheet, syllabus, etc.) are available on StudyNet/Canvas.
- Exercises will be handed out via GitHub Classroom!
- Solutions to the exercises will be made available on 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:
  - Questions: **michael.tueting@unisg.ch**
  - Decentral exam for exchange students: **TBD**.

Q&A

References

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