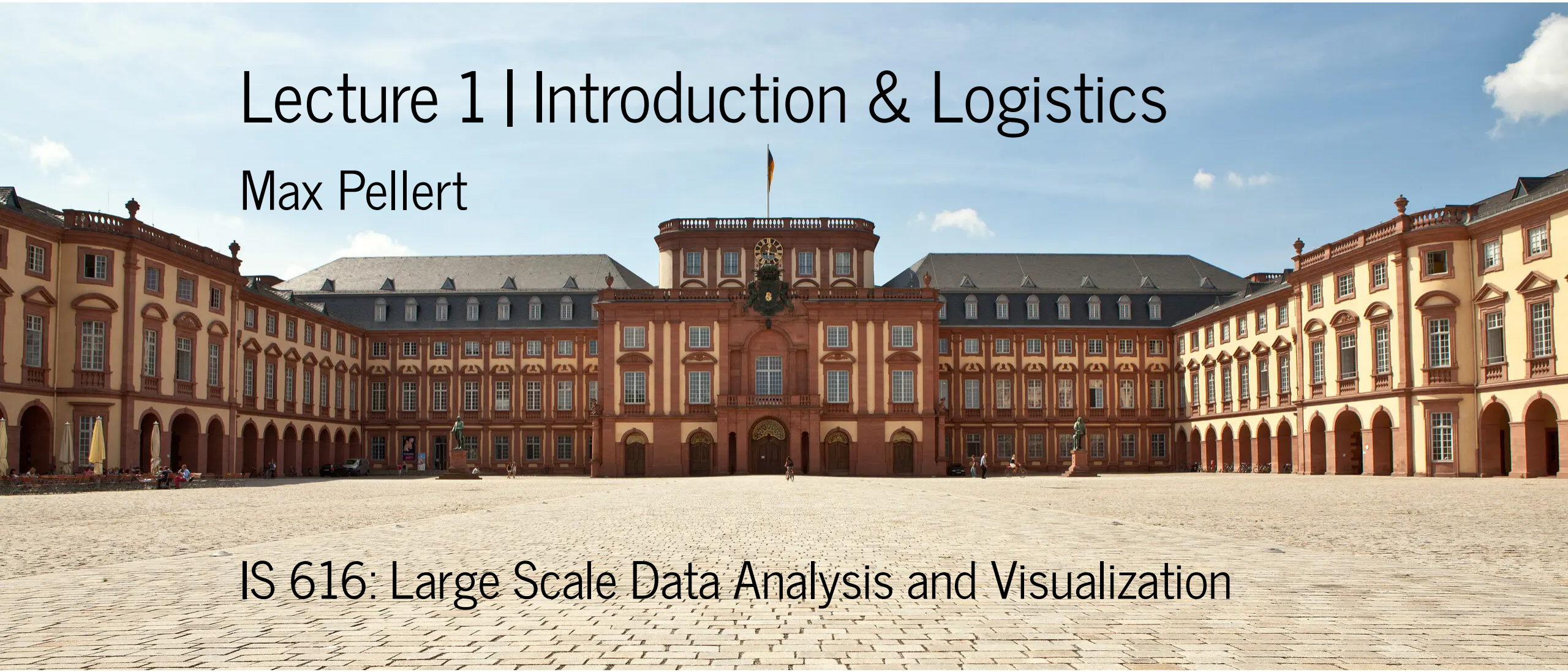
Lecture 1 | Introduction & Logistics

Max Pellert



IS 616: Large Scale Data Analysis and Visualization

The course

IS 616: Large Scale Data Analysis and Visualization

IS 616: Large Scale Data Analysis and Visualization

Contents

This course teaches students principles of scientific visualization of data using the R and Python programming languages. Starting from introductory large scale data handling and basics of visualization, more advanced methods for visualization will also be covered. Important libraries and frameworks that are essential for data analysis and visualization are introduced.

Learning outcomes

On completion of the course, students should be familiar with libraries in the R and Python programming languages that enable them to create professional scientific visualizations. This outcome includes the application of those scientific libraries, handling of large datasets and knowledge of many examples of how challenges in scientific visualization were overcome and in what ways creative solutions were found. *Skills:*

- Knowledge on how to include scientific visualization in research projects

Where are you?

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

Lehrstuhl Forschung Lehre Team Stellenangebote

Betriebswirtschaftslehre ■ Area Information Systems ■ Prof. Dr. Strohmaier

Lehrstuhl für Data Science in den Wirtschafts- und Sozialwissenschaften

Prof. Dr. Markus Strohmaier

Markus Strohmaier

Background:

- Applied Computer Science, TU Graz (2007-2013)
- Computational Social Science, GESIS & U. Koblenz-Landau (2013-2017)
- Computational Social Science, RWTH Aachen University (2017-2021)

Markus Strohmaier

Today:

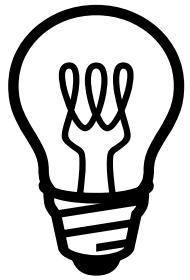
- Chair for Data-Science in the Economic and Social Sciences, Business School **University of Mannheim**, since January 1st 2022
  - Mannheim Center for Data-Science
- Scientific Coordinator at **GESIS** – Leibniz Institute for the Social Sciences since 2017

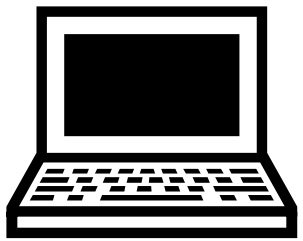
What do we typically work on at the chair?

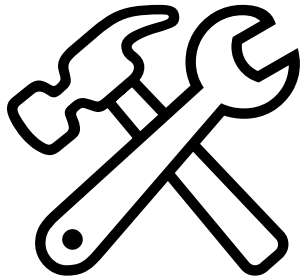
We want to give you now an overview on our research and our expertise

First conceptual

Followed by a few specific examples

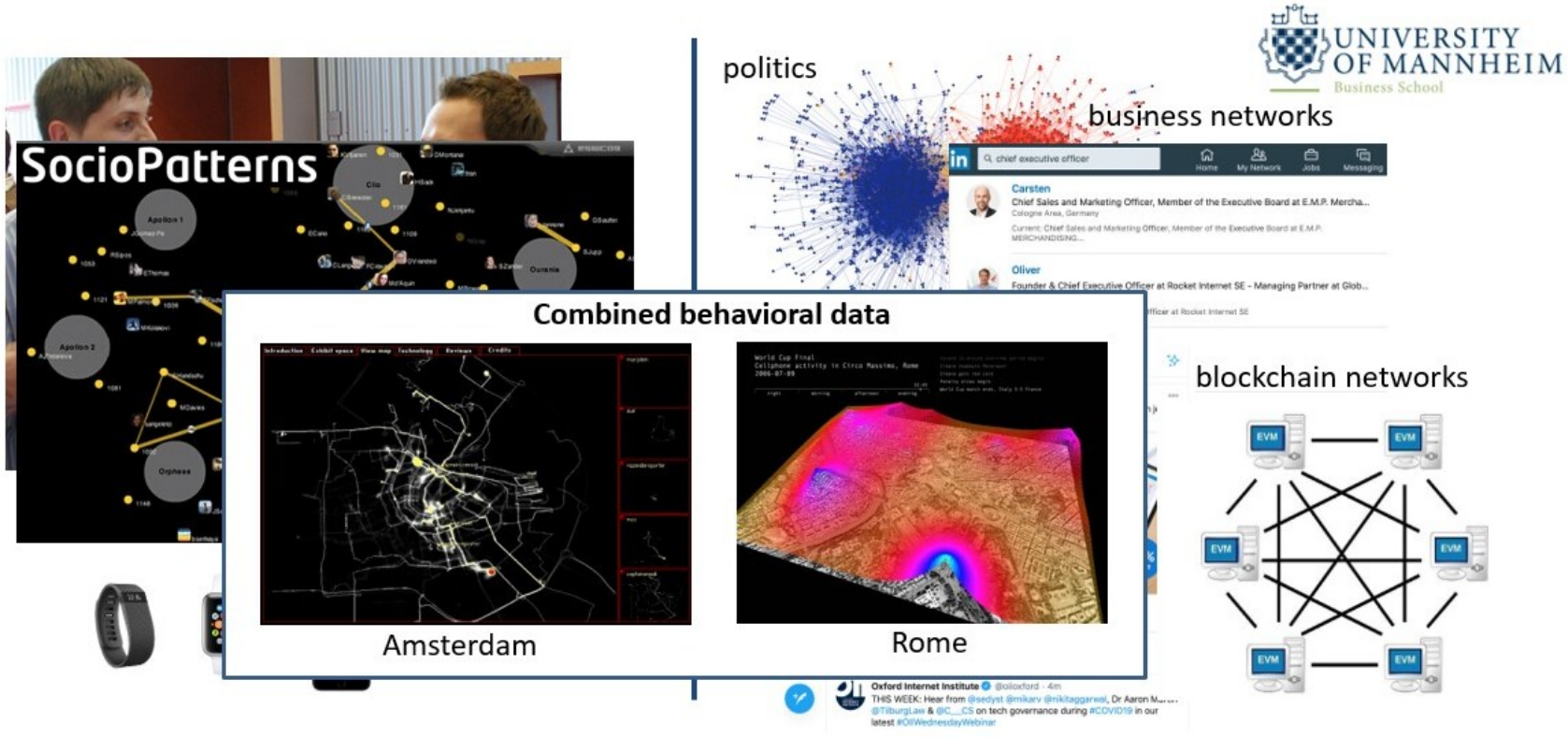






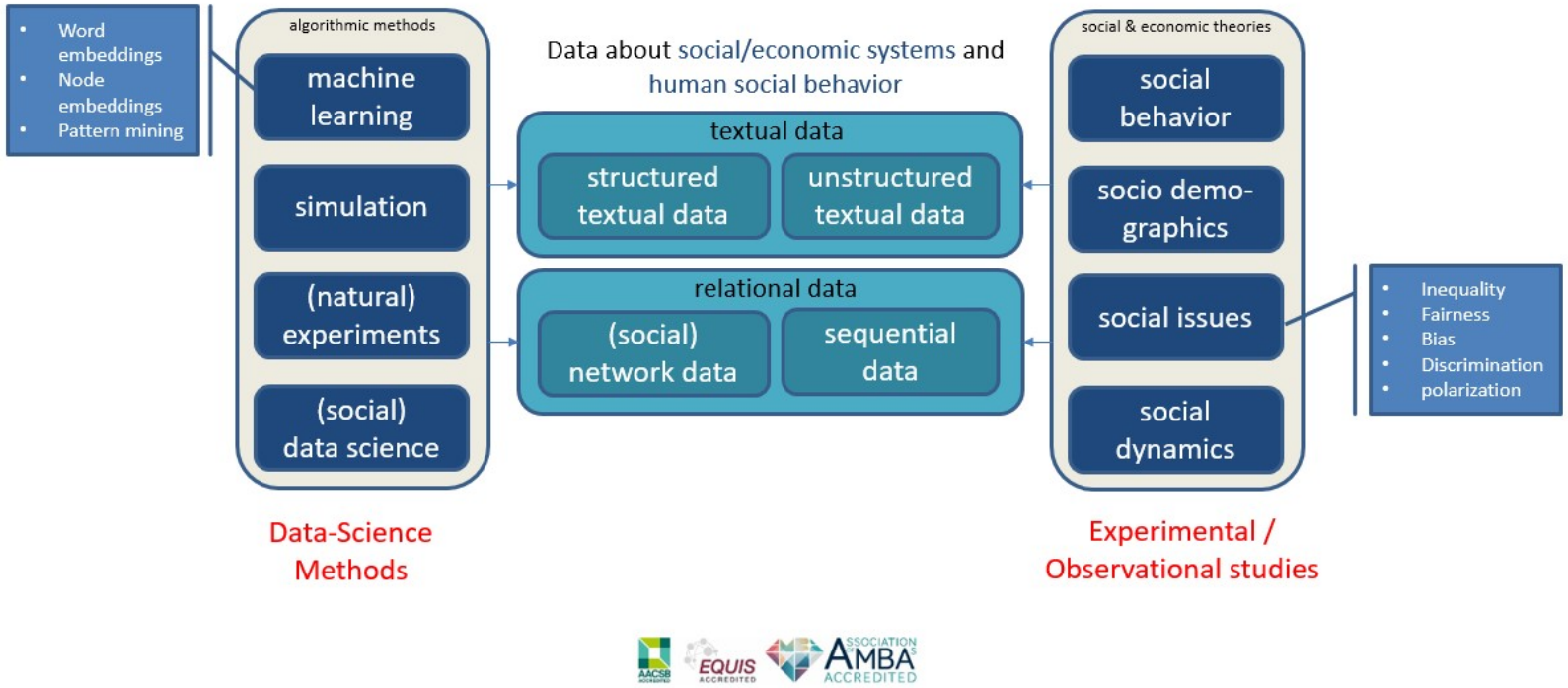
Physical behavioral data

Online behavioral data



require new computational methods and techniques!

Understanding social systems and modeling human social behavior via computational methods and new kinds of data.



Example 1

UNIVERSITY OF MANNHEIM
Business School

Data-Science methods for textual data

Word embedding → Interpretable Word embedding → Sense aware Interpretable Word embeddings

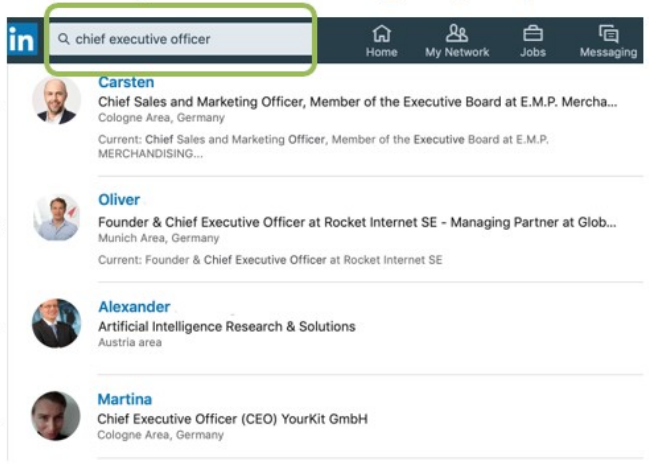
AACSB EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

Example 2

UNIVERSITY OF MANNHEIM
Business School

Data-Science methods for relational data

Inequality in the ranking of people in online social networks



Rank	Gender
1.	m
2.	m
3.	m
4.	f

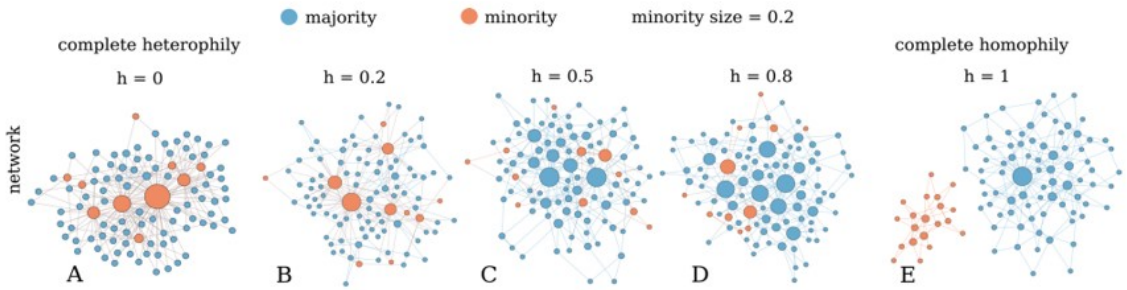
¼ (25%) women in the top 4 results

Example 3

UNIVERSITY OF MANNHEIM
Business School

Data-Science methods for relational data

Inequality in the ranking of people in online social networks



AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

What courses do we offer?

- Textual data analysis
  - IS 661 Text Analytics I (Master level)
  - IS 809 Advanced Data Science Lab II (GESS)
- Relational data analysis
  - IS 622 Network Science (Master level)
  - IS 808 Advanced Data Science Lab I (GESS)
- Seminars and master theses topics
  - CS 721 Methods of Data-Science
  - IS 723 Empirical Studies
  - IS 556 Public Blockchains
- Programming
  - IS 557 Scientific Programming with Python (Master level, for non CS-students)

Max Pellert

https://mpellert.at

Max Pellert

Interdisciplinary background: BSc Economics (and studies in Psychology), MSc Cognitive Science and a PhD in Complexity Science

All of the degrees are from Vienna (University of Vienna and Medical University of Vienna), semester abroad in Ljubljana, Slovenia

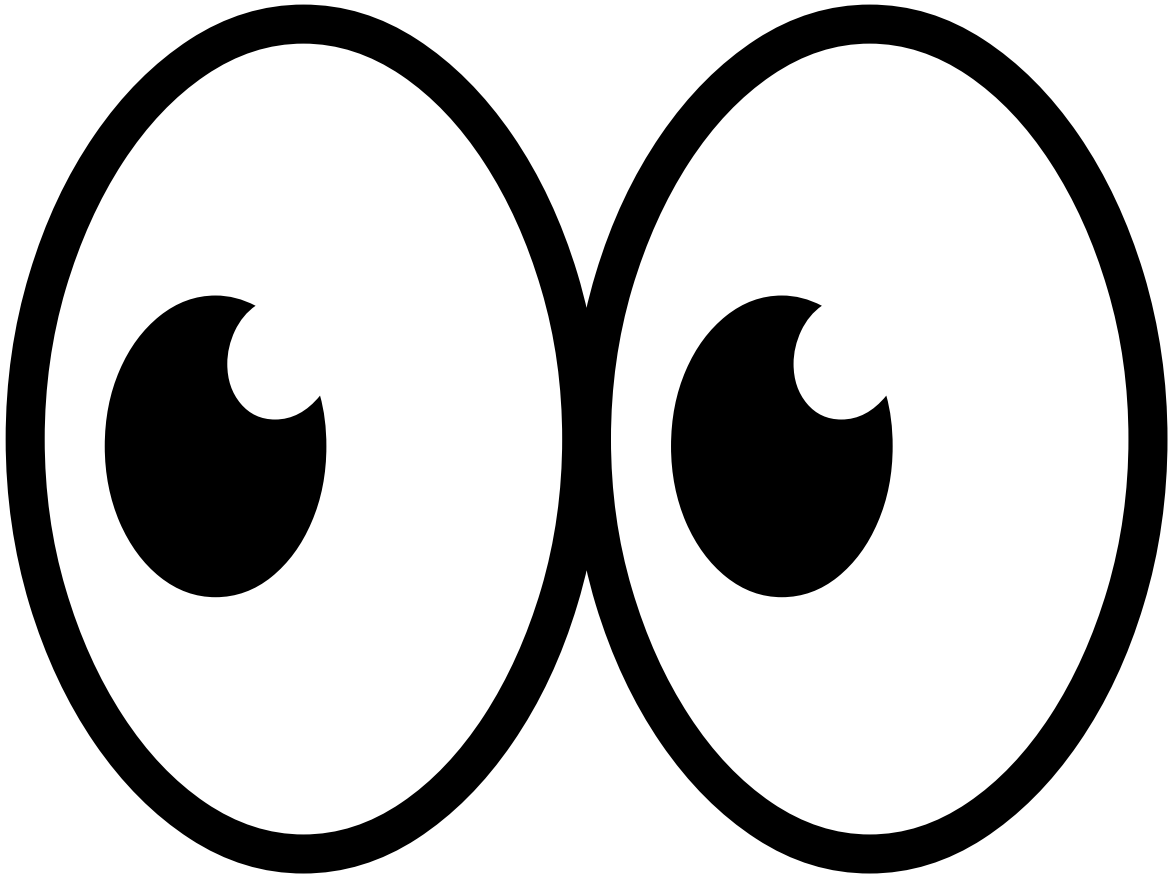
Before coming to Mannheim, I worked at Sony Computer Science Laboratories in Rome, Italy

Max Pellert

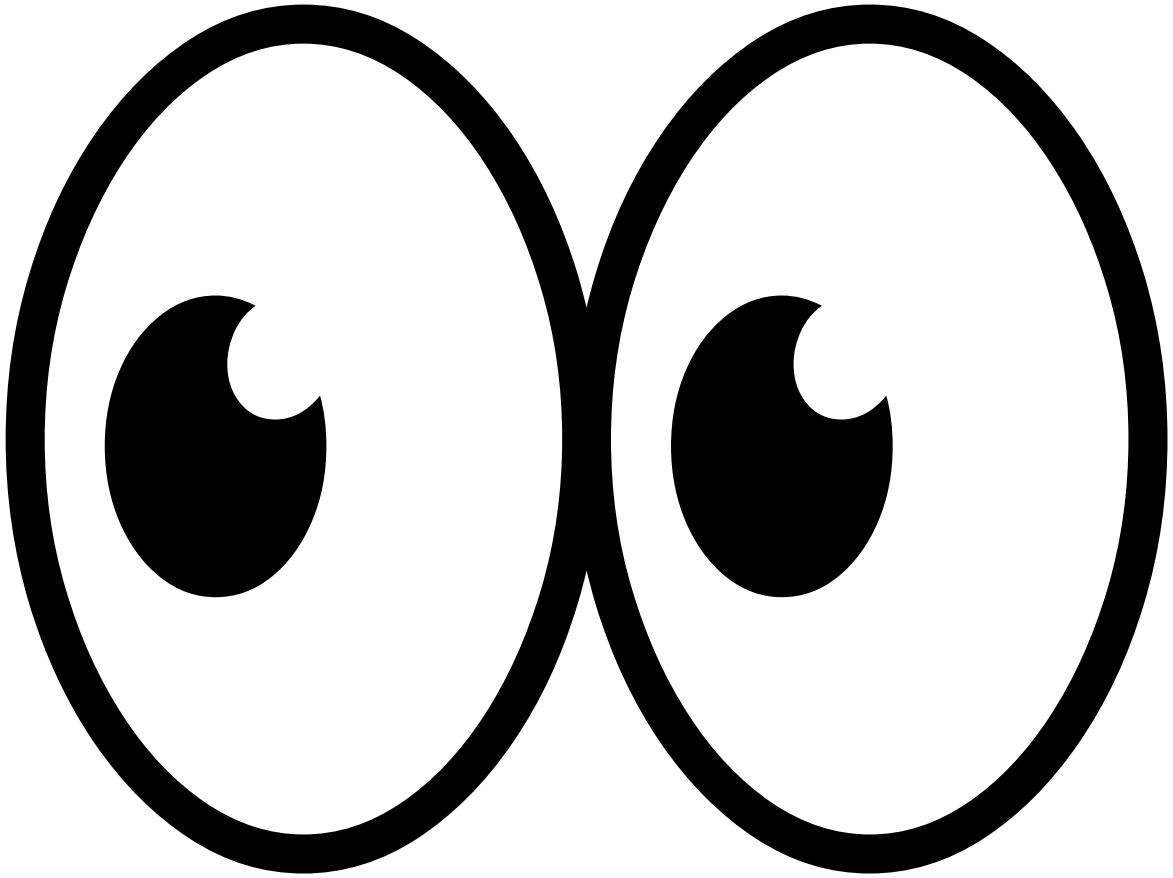
Research interests

- Computational Social Science
- Digital traces
- Affective expression in text
- Natural Language Processing
- Collective emotions
- Belief updating
- **Psychometrics of AI**

Who are you?



Your expectations?



Overall course format

13 Units (no class on German Unity Day, 3.10.2023):

- First half of each unit: lecture part
- Short break of 15 minutes
- Second half of each unit: hands-on exercise part
- Hand-In Exercises (2 planned)

These hand-in exercises have to be completed and submitted to be allowed to take part in the exam!

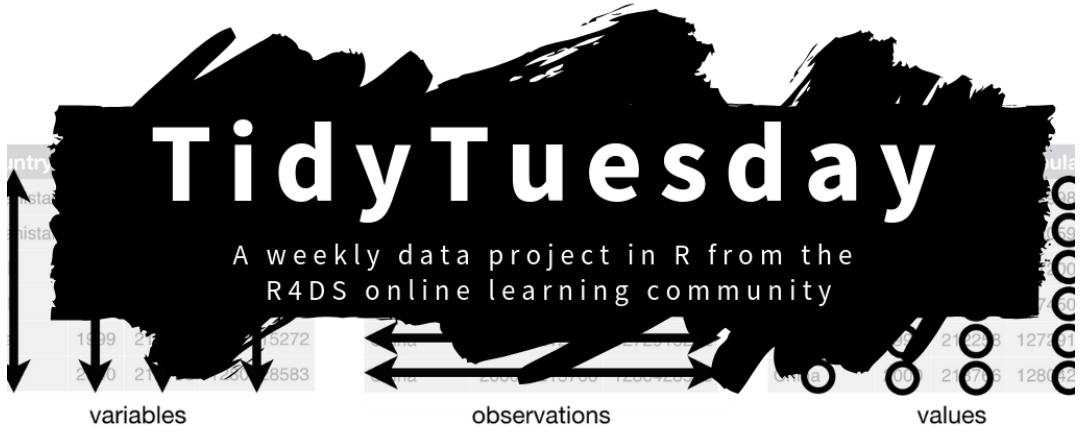
Participation, Output

Participation

- Students are expected to actively follow the lecture part
- Lecture part will provide the materials to show what *can* be done in data visualization and large-scale data processing
- Lecture part will discuss best-practice examples what *should* be done
- Exercise part will show and instruct *how* things can be done
- Students are expected to have their systems and programming environments set up to participate in the exercise part

Participation, Output

- Students need to hand-in two solutions to exercises (planned 17.10.2023 and 14.11.2023)
- You will have to complete and submit both hand-in exercises to be allowed to take part in the exam



https://github.com/rfordatascience/tidytuesday

(Preliminary) program for the course

Unit 1 | Introduction & Logistics
Unit 2 | Motivation
Unit 3 | Basics of Data Analysis I
Unit 4 | Basics of Data Analysis II
Unit 4 | History of Scientific Visualization
Unit 6 | Theory of Data Graphics I
Unit 7 | Theory of Data Graphics II

Tuesday, **13:45 - 15:15** (Lecture Part) & **15:30 - 17:00** (Exercise Part)

(Preliminary) program for the course

Unit 8 | Accessibility
Unit 9 | Grammar of Graphics I
Unit 10 | Grammar of Graphics II
Unit 11 | Advanced Visualization Techniques I
Unit 12 | Advanced Visualization Techniques II
Unit 13 | Wrap Up, Exam Preparation & Questions

Tuesday, **13:45 - 15:15** (Lecture Part) & **15:30 - 17:00** (Exercise Part)

If you are unsure if the course is right for you because …

… you have too much other obligations this semester

… you feel like you need to catch up on basics first (of programming for example)

… of many other other possible reasons

Consider deregistering now (in the beginning) to help people on the waiting list!

Books

The following books are sorted according to importance for the course

This course builds heavily on Edward Tufte's work

Edward Rolf Tufte (/ˈtʌfti/;[2] born March 14, 1942),[1] sometimes known as "**ET**",[3] is an American statistician and professor emeritus of political science, statistics, and computer science at Yale University.[4] He is noted for his writings on information design and as a pioneer in the field of data visualization.[5]

Information design [edit]

Tufte's writing is important in such fields as information design and visual literacy, which deal with the visual communication of information. He coined the word *chartjunk* to refer to useless, non-informative, or information-obscuring elements of quantitative information displays. Tufte's other key concepts include what he calls the *lie factor*, the *data-ink ratio*, and the *data density* of a graphic.[12]

He uses the term "data-ink ratio" to argue against using excessive decoration in visual displays of quantitative information.[13] In *Visual Display*, Tufte explains, "Sometimes decoration can help editorialize about the substance of the graphic. But it is wrong to distort the data measures—the ink locating values of numbers—in order to make an editorial comment or fit a decorative scheme."[14]

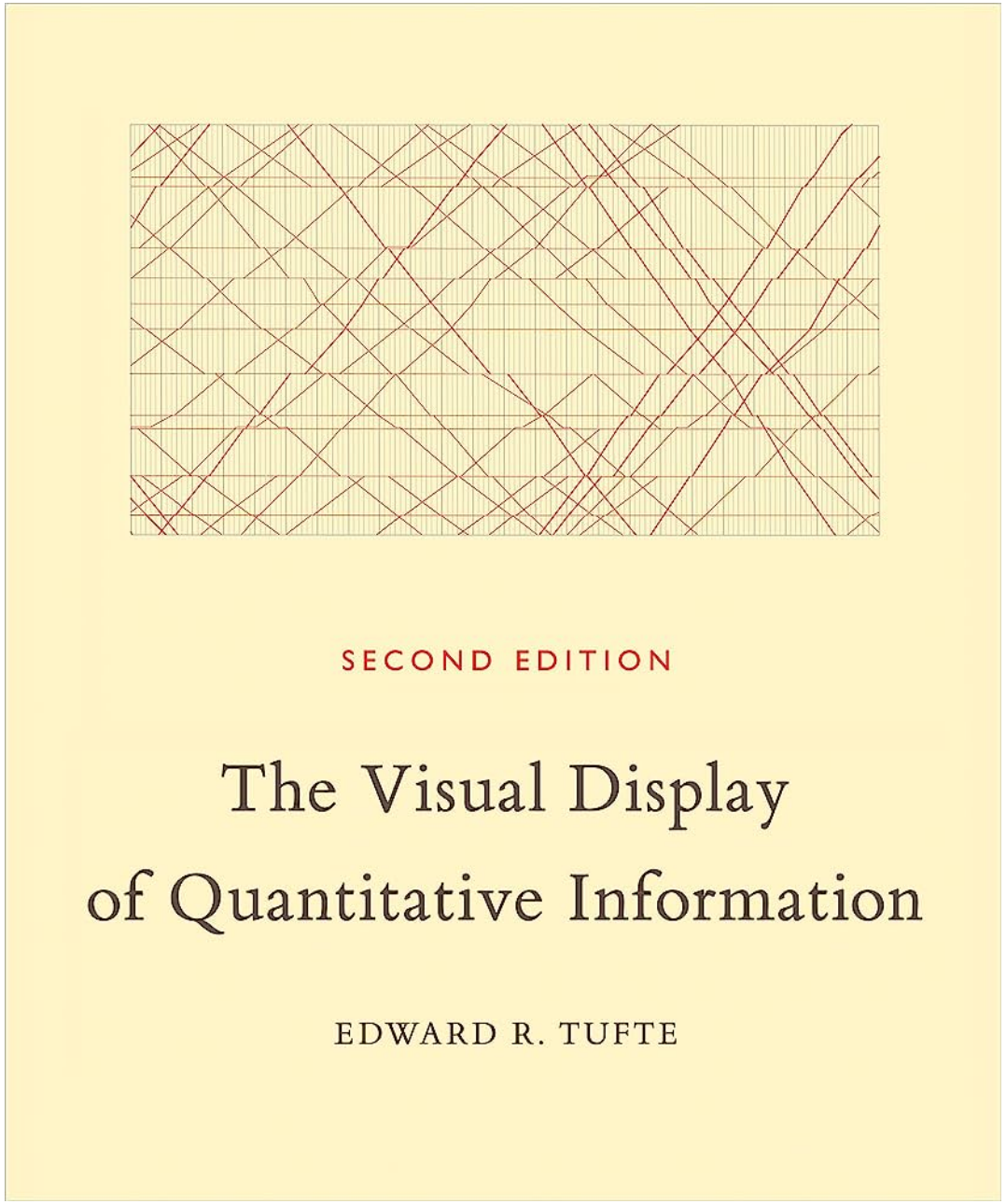
Pronunciation of "Tufte"?

How do you pronounce your last name?

-- Heather Strong

TUFF-tee

-- Edward Tufte


SECOND EDITION
The Visual Display
of Quantitative Information
EDWARD R. TUFTE

The Visual Display of Quantitative Information

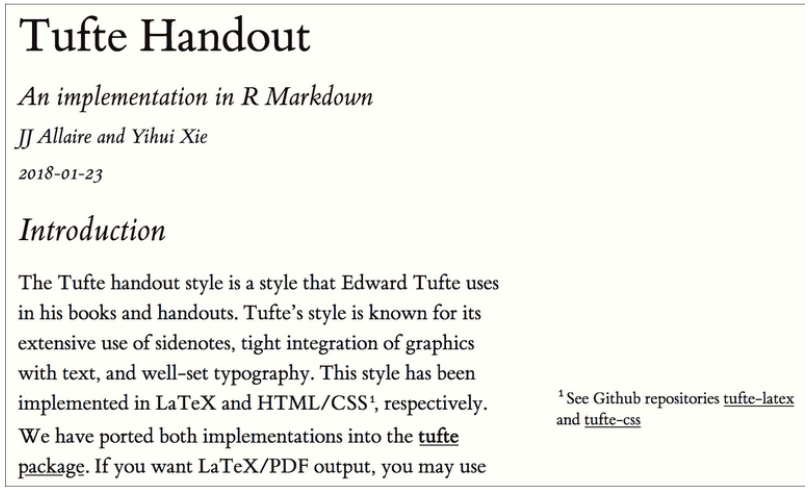
- A very influential book
- Entertaining read
- Provides a history of scientific visualization, good as well as bad examples from early to contemporary times, theoretical principles of good information design and many other things
- Also the use of “sidenotes, tight integration of graphics with text, and well-set typography” in the book itself was influential:

Chapter 6 Tufte Handouts

The Tufte handout style is a style that Edward Tufte uses in his books and handouts. Tufte's style is known for its extensive use of sidenotes, tight integration of graphics with text, and well-set typography. This style has been implemented in LaTeX and HTML/CSS,[5] respectively. Both implementations have been ported into the **tufte** package (Xie and Allaire 2022). If you want LaTeX/PDF output, you may use the `tufte_handout` format for handouts, and `tufte_book` for books. For HTML output, use `tufte_html`, e.g.,

```
---
title: "An Example Using the Tufte Style"
author: "John Smith"
output:
  tufte::tufte_handout: default
  tufte::tufte_html: default
---
```

Figure 6.1 shows the basic layout of the Tufte style, in which you can see a main column on the left that contains the body of the document, and a side column on the right to display sidenotes.



Tufte Handout

An implementation in R Markdown

JJ Allaire and Yihui Xie

2018-01-23

Introduction

The Tufte handout style is a style that Edward Tufte uses in his books and handouts. Tufte's style is known for its extensive use of sidenotes, tight integration of graphics with text, and well-set typography. This style has been implemented in LaTeX and HTML/CSS[1], respectively. We have ported both implementations into the tufte package. If you want LaTeX/PDF output, you may use

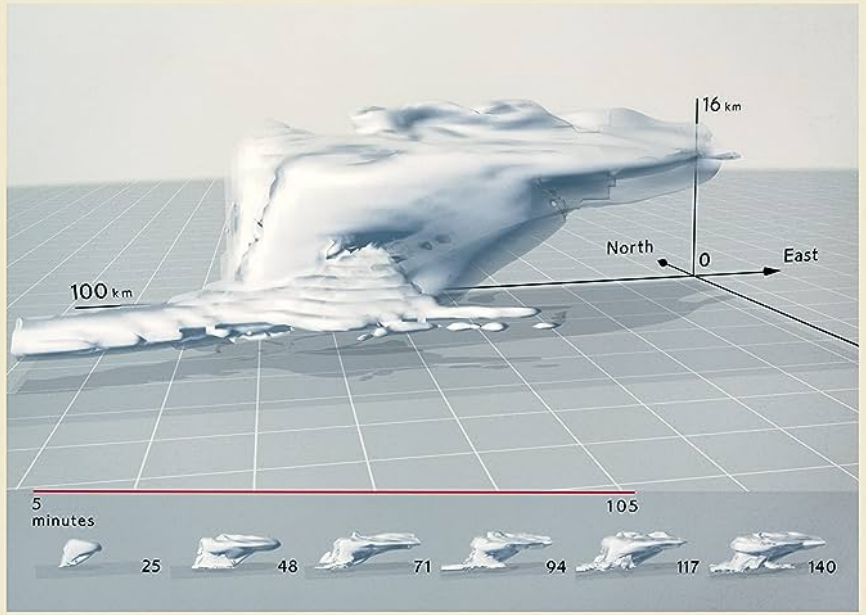
[1] See Github repositories tufte-latex and tufte-css

FIGURE 6.1: The basic layout of the Tufte style.

https://bookdown.org/yihui/rmarkdown/tufte-handouts.html

EDWARD R. TUFTE

VISUAL EXPLANATIONS



IMAGES AND QUANTITIES, EVIDENCE AND NARRATIVE

Visual Explanations

- Kind of a follow-up book to “The Visual Display of Quantitative Information”
- Outlines several interesting, classic case studies of the power of visualization in scientific analysis, including the story of John Snow and the cholera outbreak in London of 1854
- In total, Tufte wrote a series of 4 books on the topic: the two discussed and “Beautiful Evidence” and “Envisioning Information”

A New Framework for Machine Learning and the Social Sciences

Justin Grimmer | Margaret E. Roberts | Brandon M. Stewart

“Text as Data”

- The book can help to find out what methods are available and how they can (and have been) used to tackle research questions
- Starts with meta-theoretical considerations and gives you some kind of roadmap on how to use text as data to tackle scientific questions

“Text as Data”

- Builds up sophisticated machinery by going from simple to advanced in a very concise, efficient way (providing lots of pointers to additional materials)
- Can serve as a work of reference to look up certain methods that you might need and get inspiration on how to use them (for example different clustering techniques are covered in one of the chapters)

Other Ressources to warm up

You are expected to fill in the blanks in your skills on your own!

R Self-Assessment Test

R Crash Course by David Garcia + Materials

Python Crash Courses by University Libraries at the University of North Carolina at Chapel Hill

Python Data Science Handbook by Jake VanderPlas

Questions?