

# 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?



[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

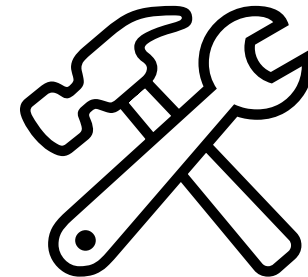
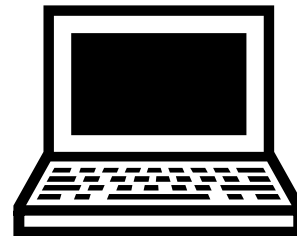
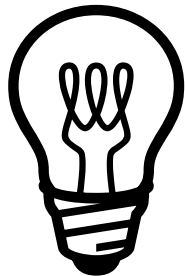


# What do we typically work on at the chair?

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

**SocioPatterns**

**business networks**

**Combined behavioral data**

**Amsterdam**

**Rome**

**blockchain networks**

UNIVERSITY OF MANNHEIM Business School

politics

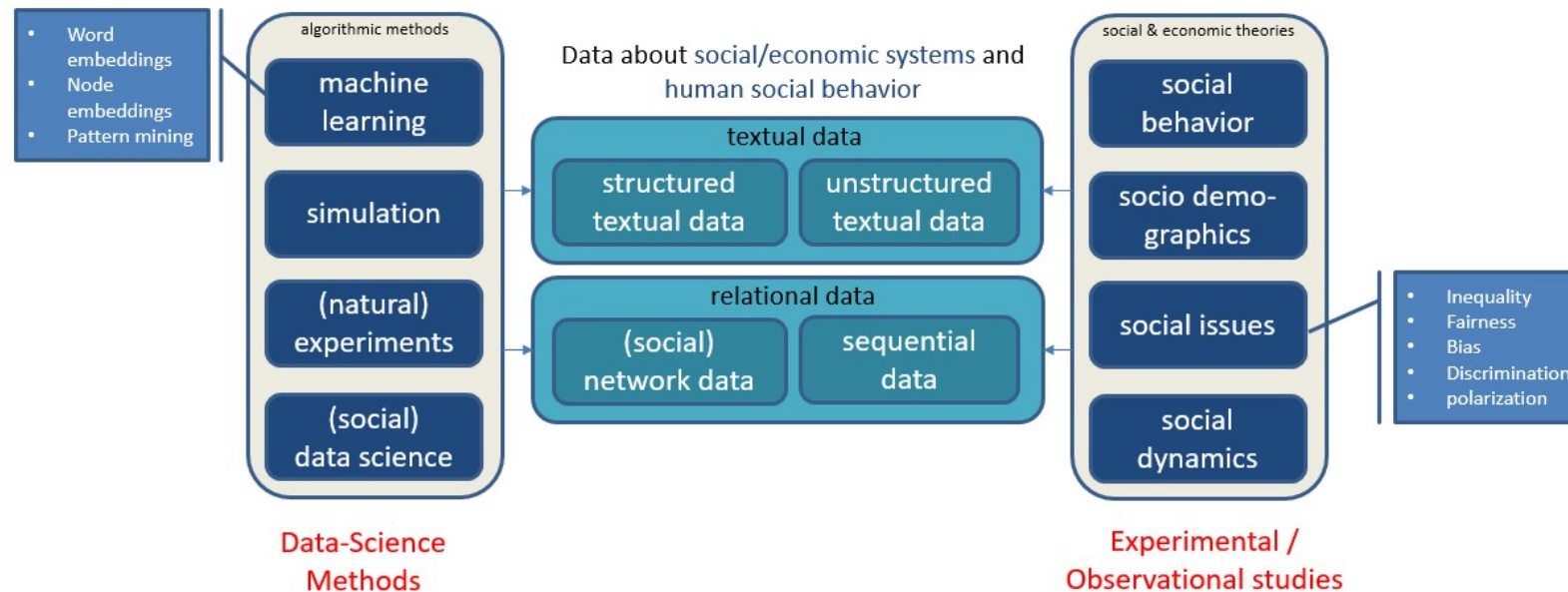
LinkedIn profile: Carsten, Chief Sales and Marketing Officer, Member of the Executive Board at E.M.P. Mercha...  
Oliver, Founder & Chief Executive Officer at Rocket Internet SE - Managing Partner at Glob...

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF AMBA ACCREDITED

Oxford Internet Institute @oioxford - 4m  
THIS WEEK: Hear from @sedyst @mikav @mikitaggerwal, Dr Aaron M...  
@TibaultLan & @C\_CS on tech governance during #COVID19 in our latest #OWWednesdayWebinar

require new computational methods and techniques!

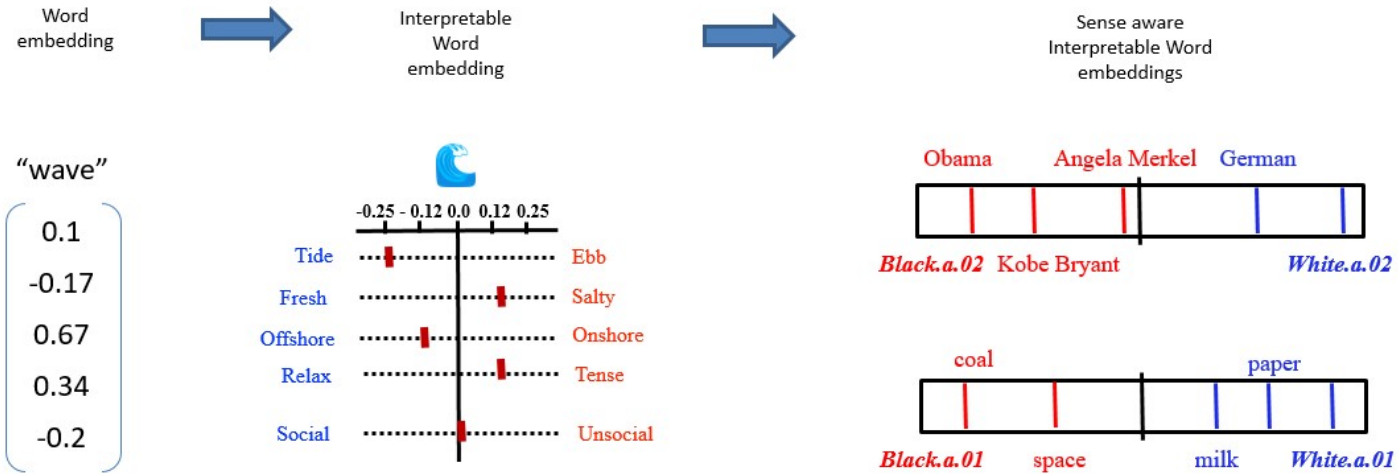
# Understanding social systems and modeling human social behavior via computational methods and new kinds of data.





# Example 1

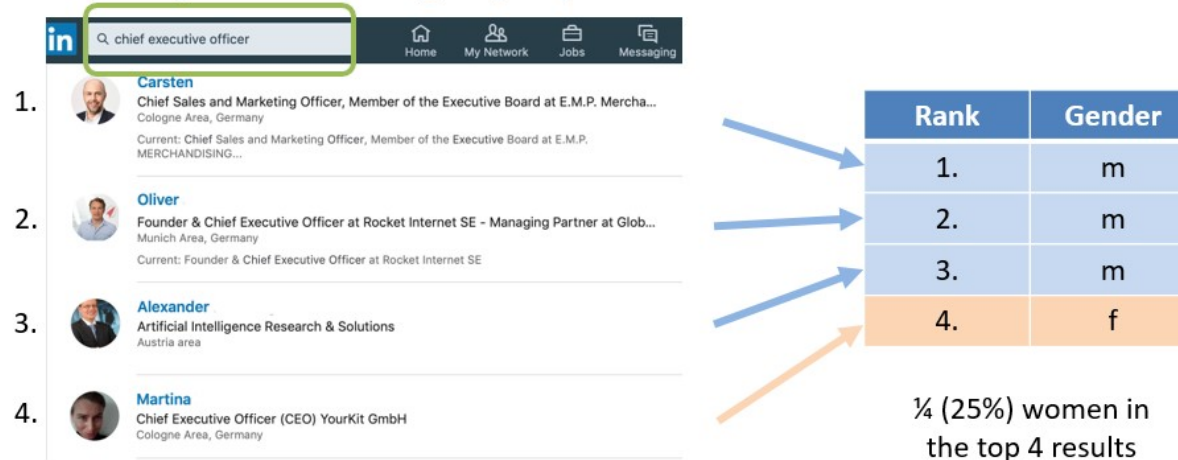
## Data-Science methods for textual data



# Example 2

## Data-Science methods for relational data

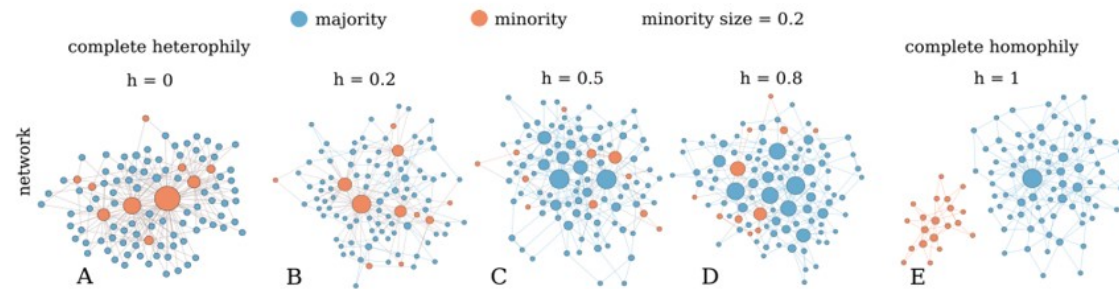
### Inequality in the ranking of people in online social networks



# Example 3

## Data-Science methods for relational data

### Inequality in the ranking of people in online social networks



10

# 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

Max Pellert

About

Materials

Talks



## Max Pellert

Junior Faculty (Assistant Professor)  
University of Mannheim

PhD  
Complexity Science Hub Vienna  
Medical University of Vienna



### > Bio

Max Pellert has a background in cognitive science and economics (University of Vienna, Austria and University of Ljubljana, Slovenia). He was a doctoral researcher affiliated to Complexity Science Hub Vienna and Medical University of Vienna in the WWTF research group “Emotional Well-Being in the Digital Society” lead by David Garcia (now University of Konstanz). After receiving his PhD, he gained industry experience as Assistant Researcher at Sony CSL Rome. Currently, he works at the Chair for Data Science in the Economic and Social Sciences at University of Mannheim (Markus Strohmaier) as junior faculty (assistant professor). His research focuses on analyzing the digital traces of individual and collective emotional behavior and affective expression on social media. He is broadly interested in the social sciences and uses traditional and novel computational methods from domains such as Natural Language Processing to study emotion dynamics, belief updating, collective emotions and other interesting phenomena.

<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

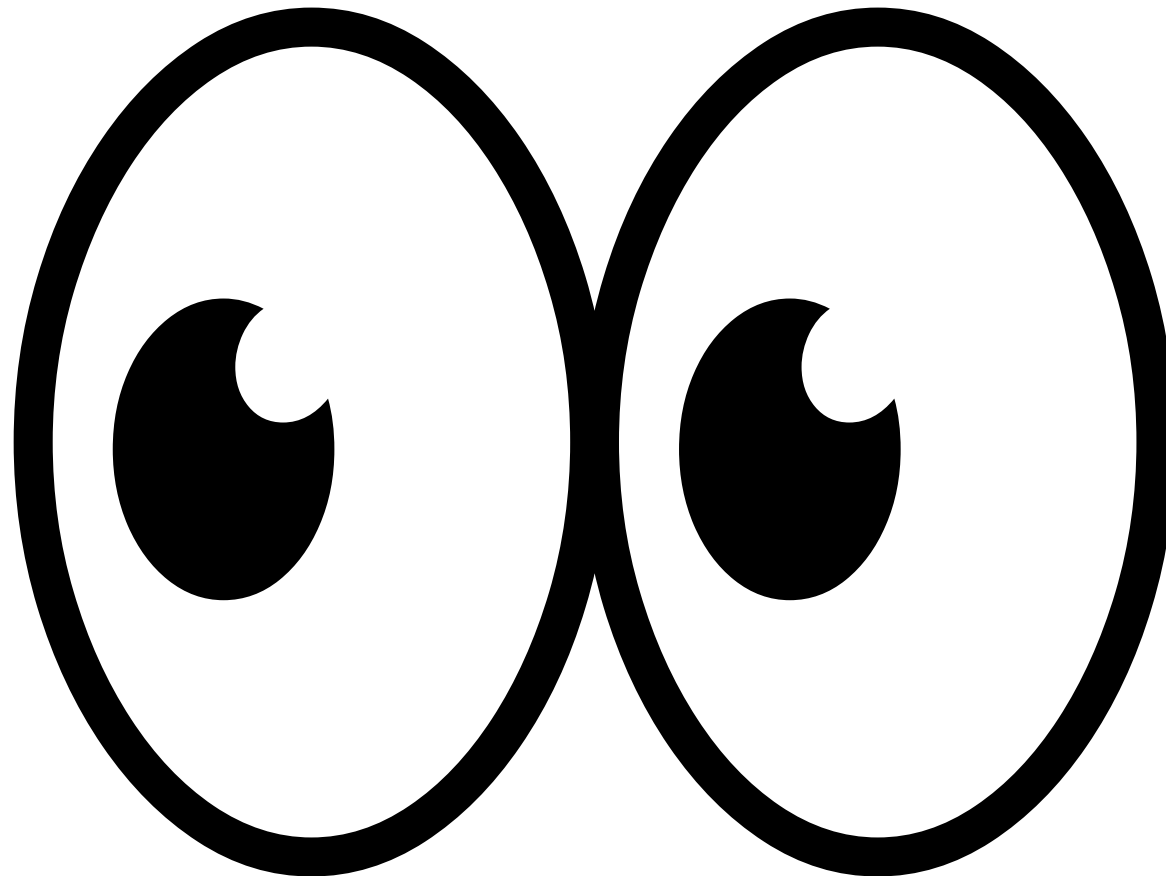
Before coming to Mannheim, I worked at Sony Computer Science Laboratories in Rome, Italy

# Max Pellert

## Research interests

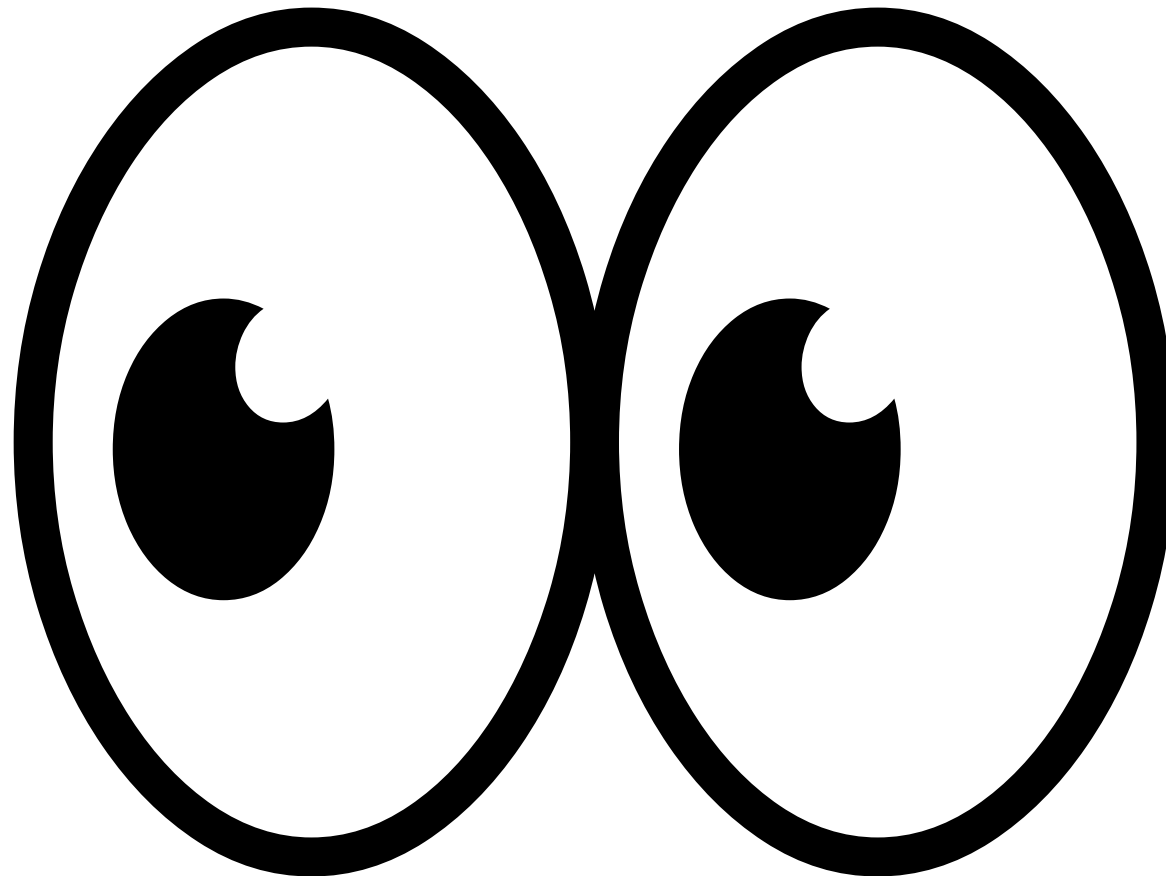
- Computational Social Science
- Digital traces
- Affective expression in text
- Natural Language Processing
- Collective emotions
- Belief updating
- **Psychometrics of AI**

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!**

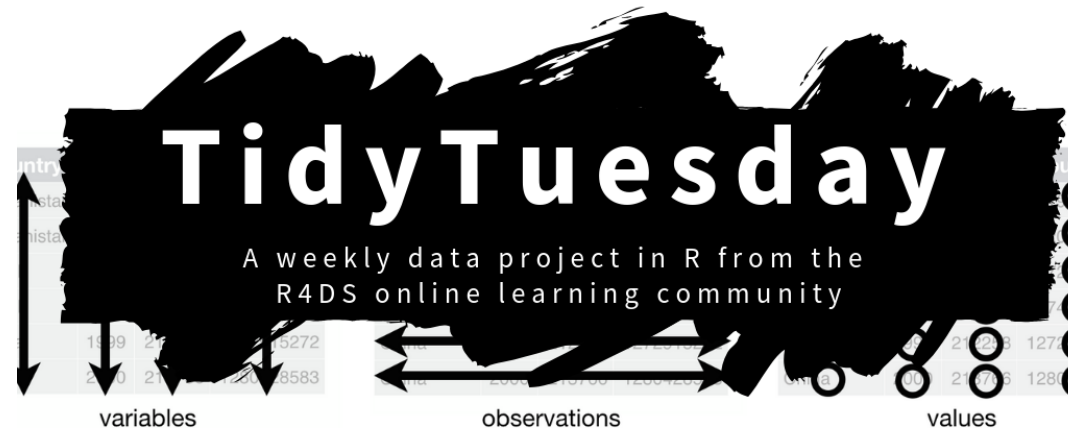
# Participation, Output

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

## 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.<sup>[12]</sup>

He uses the term "data-ink ratio" to argue against using excessive decoration in visual displays of quantitative information.<sup>[13]</sup> 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."<sup>[14]</sup>

## Pronunciation of "Tufte"?

How do you pronounce your last name?

-- Heather Strong

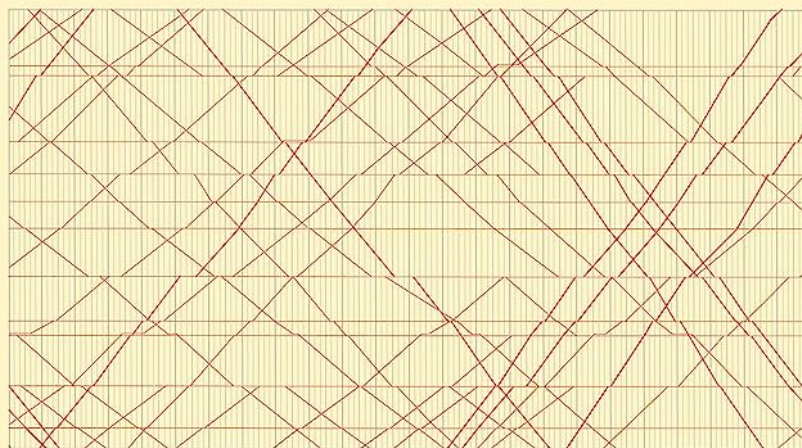
---

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,<sup>5</sup> 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.

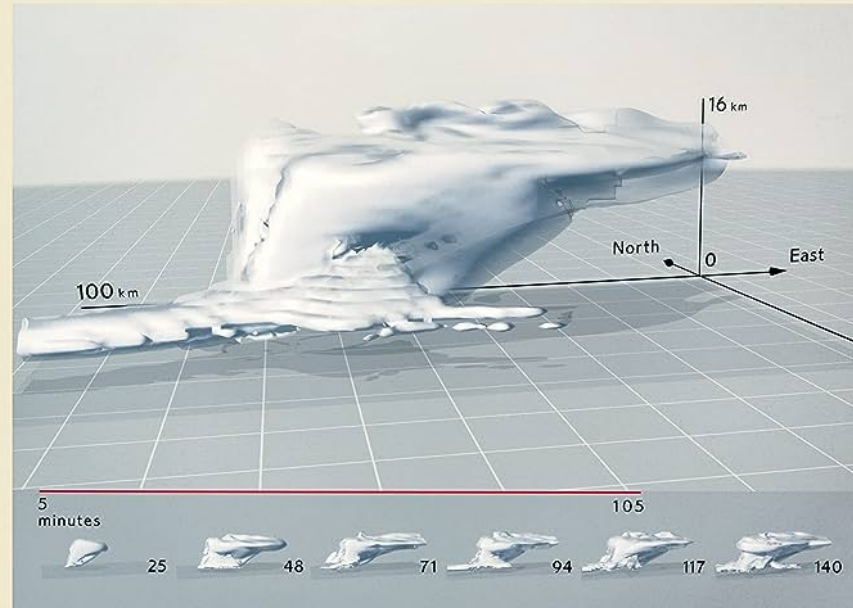


FIGURE 6.1: The basic layout of the Tufte style.

<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”

AMONG THE NUMEROUS AT TENDENCY TO BREAK AND C FOR THEIR CHARACTER AND DUE VALUE ON ANY PLAN V JUSTICE, AND TMENTS HA DERIVE THE BOTH AND D THAT THI ROM OUR TY, THAT OI ES ARE TOC ITERESTED E, OF KNOW SITUATION IMENTS; BL ND, PARTI OED FROM E WITH WI

NTAGES PROMISED BY F FROM THE VIOLENCE OF E, AS WHEN HE CONTEA H, WITHOUT VIOLATING I INTRODUCED HERE PERISI CIOUS DEC ODERN, GARRISON'S EFFECTUALLY OBIATED DERATE AND VIRTUOU ABENTS ARE TOO UNSTA IDED, NOT ACCORDING ARING MA ILL NOT PE OF THE DI: FOUND, AT THE SAME TI R THAT PREVAILING ANI : THE CONTINENT TO TH OUS SPIRIT HAS TAINTI

CONSTRUCTI N, THE FRIEN THEIR PROPL NCIPLES TO I DUNGLES, HA E, IN TRUTH TO BE THE FAVORITE AM E IMPROVEMENTS M UCH ADMIRERD F THIS SIDE, AS ILLY THE FRIEN JIBLES GOOD IS B OF JUSTICE AND TH OUSLY WE MAY WISH IHEY ARE IN SOME DEGRD H WE LABO VEE BEEN ER, I OTHER CAI ILL NOT ALO, EASING DIST PUBLIC END HER, THESE M UR PUBLIC A NS, BY A FAC, TEN AND E

NONE DESEF PULAR GOV TO THIS DA CH HE IS ATT TO BE THE FAVORITE AM E IMPROVEMENTS M UCH ADMIRERD F THIS SIDE, AS ILLY THE FRIEN JIBLES GOOD IS B OF JUSTICE AND TH OUSLY WE MAY WISH IHEY ARE IN SOME DEGRD H WE LABO VEE BEEN ER, I OTHER CAI ILL NOT ALO, EASING DIST PUBLIC END HER, THESE M UR PUBLIC A NS, BY A FAC, TEN AND E

IS TO BE MORE ACCURATELY MENTS NEVER FINDS HIMSEL DUS VICE, HE WILL NOT FAIL PROVIDES A PROPER CURE AL DISEASE FROM WI AN CONST JNWARAI ECTED, COI VATE FAITH ONFLICTS NOR PARTI AINTS HAD FOUND I ARGED ON FOR MANY ND ALARM ECTS OF TH DERSTAND Y CASE OF

TO EVERY ST REMEDY, I RES, BUT IT CO LD BE TO WISH TH EXPEDIENT IS AS II XERCISE IT, DIFFERE NS AND M'S PASSION, H THEMSE ES, THE DI: GLE TO A U ORMITY O RENT AND I DUAL FACI AND PRO I INFLUENCE FFERENT IN S AND PART HIT INTO D DEGREES C ERNING RELIGION, CONCERNIN RENT LEADERS AMBITIOUSLY COM I R, DIVIDED M INTERESTING TO THE HUMAN PASS IVE BEEN I EX AND OPPRESS EACH OTI FUAL ANIMOSITIES, THAT W ICIENT TO KINDLE THEI NS HAS BEEN TI

IONS, THE SAME HAN THE DISEASE, LIB OLLY TO ABOLISH LIBEF A OF AIR, WHICH IS LESS AS THE FIRST D B WILL BE FO A RECIPRO HE FACULTIES O THE PROTECTION O TURNING PROPERTY TH THE SENTIMENTS AND CAUSES OF FACTION TO THE DIF ANY OTHER ENCE AND HAVE B

UT CONCERNING THE O THE CAUSES WHICH THE ON ONE SIDE AND THE DEBTO, ELVES THE JUDGES, AND THE M DOMESTIC MANUFACTURES BE EN, WOULD BE I Y DECIDED B ND THE PL THE APPOR OST EXAC TY, YET THEB TRAMPLE C REDOMIN, CKETS, IT IS VIENT TO T : MADE AT # HE IMMEDI WHICH ON ATIS, THAT I ONTROLLI ECTS, IF A FAC MAJORITY, U DEFEAT ITS SINSTEP INABLE TO EXECUTE AND MASK OF POPULAR GOVERNMENT, C IGHTS OF OTHER CITIZEN RESERVE THE SPIRIT

CITIZENST PROPOSED CL JUSTICE OUGHT I ARTY, OR, IN OT ID IN WHAT DEGR J AND THE MANUFA OF TAXES ON THE VAK APS, NO LE ISLATIVE A JLES OF JU E, EVERY S, TO SAY TH UIGHTENE LIC GOOD, ITENED ST, HOUT TAKI O VIEW INL Y MAY FIN REGARDING USES OF FACTION CANNOT BE RE CONSISTS OF LESS THAN A MAJOR S BY REGULAR VOTE, IT MAY CLOG TI JENCE UNDER THE CO OF THE CO, /OTHER HANT SACRIFICE I VATE RIGHT THEN THE I

E THE DIFFERENT CLASSES O PRIVATE DEBTS? IT IS A QUE : BALANCE BETWEEN THEM, ST, ARE QUESTION WITH A SOLE REGAK WHICH SEEMS TO REC EMPTATIO ARE GIVEN /HICH THE NUMBER, I ESE CLASH INTERESTS, HELM, NOR ANY CASES, S, WHICH V RELY PREVA OOD OF TH THE INFE TO BE SOUGHT IN THE MEANS O PUBLICAN PRINCIPLE, WHICH ENA JNVULSE THE SOCIETY; BUT IT WILL B ITY IS INCL IN A F FACTION, THE FL INTEREST BC SUCH A F OUR INOLY LE SAME TIM LET ME ADL

# 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 Resources 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?

