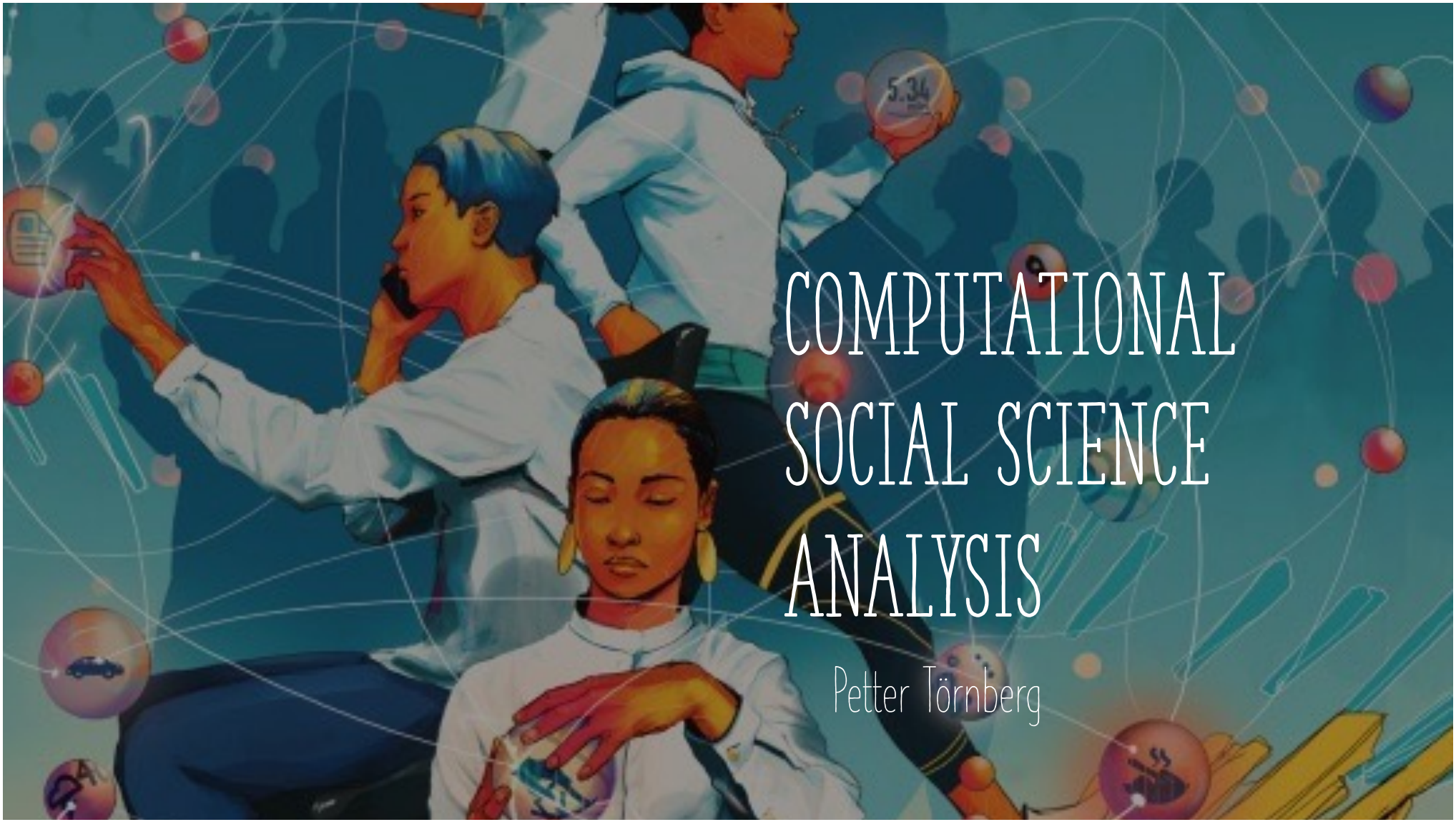


COMPUTATIONAL SOCIAL SCIENCE ANALYSIS

Petter Törnberg

THIS LECTURE

Part 0. Welcome!

Part 1. What is Computational Social Science?

Part 2. How to write a CSS research paper (in 6 weeks)

PART 0: WELCOME

WHO AM I?

Petter Törnberg (He/him)

Associate Professor in Computational Social Science at the Institute for Logic, Language and Computation.

Background in computer science and physics. Complexity science. But now social scientist.

My research: Social media, polarization and radicalization, platformization and mediatization. Large Language Models and AI. Social theory.

WHO ARE YOU?

- Pronouns?
- What did you study?
- Do you know programming?
- What are your research interests?

THIS COURSE

- Practical course to using CSS methods in the social sciences.
- Learn to write a research paper.
- Course mimics the research process:

  1. Research proposal. 2. Research paper draft. 3. Peer-review. 4. Revised final paper
- Each week will have a lecture, a workshop and an independent co-work session.
- Lecture gives theory. Workshop gives code.
- GitHub: https://github.com/cssmodels/ComputationalSocialScienceCourse

WHAT YOU CAN EXPECT OF THIS COURSE

- Freedom to pursue your own interest and focus on what you want
- You will learn powerful and useful digital research methods
- A *challenging* but exciting course

MY EXPECTATIONS OF YOU

- That you work independently and take your own responsibility for your learning.
- That you do your own readings and get the resources you need for your project.
- That you ask me for support whenever you need it
- That you submit work on time.
- That you show up: attendance is mandatory (you can miss a maximum of 2/12.)
- That you show up on time! (If you miss the attendance sheet, you missed the class.)

WHAT ARE YOUR EXPECTATIONS?

- Why did you choose to take this course?
- What are you hoping to get out of this course?

ASSESSMENT

The course is organized around the process of writing a paper.

1. Project proposal - team based (Pass/fail)
2. Research paper draft - team based (Pass/fail)
3. Peer-review on your classmates' papers - individual (20% of grade)
4. Final revised paper - team based (80% of grade)

DUO-GROUP WORK

Based on feedback from last year, the work will now be organized in groups of two.

It is *your* responsibility to find a collaboration partner!

Find someone with shared interest and complementary skills.

Register your group on Canvas.

ASSESSMENT 1: PROJECT PROPOSAL (FROM TODAY; DEADLINE APRIL 15; 11:00)

The primarily goal of this exercise is to make sure that you have an appropriate, feasible, and interesting project idea.

Write a concise one-page document outlining your research project within the realm of Computational Social Science.

What literature are you speaking to? What is your research question? How will you answer it? What are your expectations? What is your data? What is your method?

Write it as you expect to write the final paper, making a guess for what you will find! *Find your story!*

Length: max 1 page (A4, 11pt)

Grading: The grading will be pass/fail. If you fail, you will get the opportunity to revise your plan based on feedback from the instructor.

You will get feedback to guide your project.

See Canvas for more details!

ASSESSMENT 2: RESEARCH PAPER DRAFT (FROM APRIL 15-MAY 21)

The culmination of your project will be a research paper, written in the style of a scholarly article for *Journal of Computational Social Science*.

Present your findings from your project and how they contribute to the literature.

Length: 3000 ± 200 words in length (excl reference list)

Draft deadline: 2024-05-21 (before noon; 12:00)

See Canvas for more details!

ASSESSMENT 3: PEER REVIEW (FROM MAY 21 - MAY 23 17:00)

Individual task! You will provide peer-review on another group's papers. Read them carefully and make suggestions on how they can be improved.

Your task is to help the authors get the highest grade possible.

Look at the paper grading criteria and journal guidelines. Imagine that you're me.

Your feedback should be thorough, fair, empathetic, and constructive. The author should feel empowered and supported to improve their paper.

Length: around 1 page in length (A4, Times New Roman, 11pt).

Deadline: May 23, 17:00.

Grading: 1-10. Represents 20% of the final course grade

See Canvas for more details!

ASSESSMENT 4: REVISED RESEARCH PAPER (DEADLINE: MAY 28, 17:00)

- For the final submission, you will revise your paper based on the feedback from the reviewers.
- Final submission should include a 1-page response to each reviewer comment. How did you address the problems they raised? Be kind and grateful for their work helping you with your paper!

Length: 3000 ± 200 words paper + 500-600 words response to reviewers

Grading: 1-10. Represents 80% of the final course grade

DEADLINE OVERVIEW

	Deadline date
Research paper proposal	2025-04-15 (before class begins; 11:00)
Research paper draft (to reviewers)	2025-05-21 (before noon; 12:00)
Peer-reviews	2025-05-23 (before end of workday; 17:00)
Final research paper submission	2025-05-28 (before end of workday; 17:00)
[Paper Resit/Repair]	2025-06-30 (before end of workday; 17:00)

Read the course manual!

COURSE CONTENT

Week	Lecture	Workshop	Project work
1	What is Computational Social Science? + Introduction to Programming/Python	Introduction to Python continued	Programming homework and proposal writing
2	Acquiring Big Data: Scraping and APIs + Ethics & Law	Web Scraping and APIs in practice	Programming and proposal writing *continued*
3	Natural Language Processing for the Social Sciences	Natural Language Processing in practice	Individual project work
4	Social Network Analysis	Social Network Analysis in practice	Individual project work
5	Machine Learning in the social sciences	Machine Learning in practice	Individual project work
6	Agent-Based Modeling and Online Experiments	Agent-Based Modeling in practice, + How to write a research paper	Individual project work

IT'S YOUR COURSE!

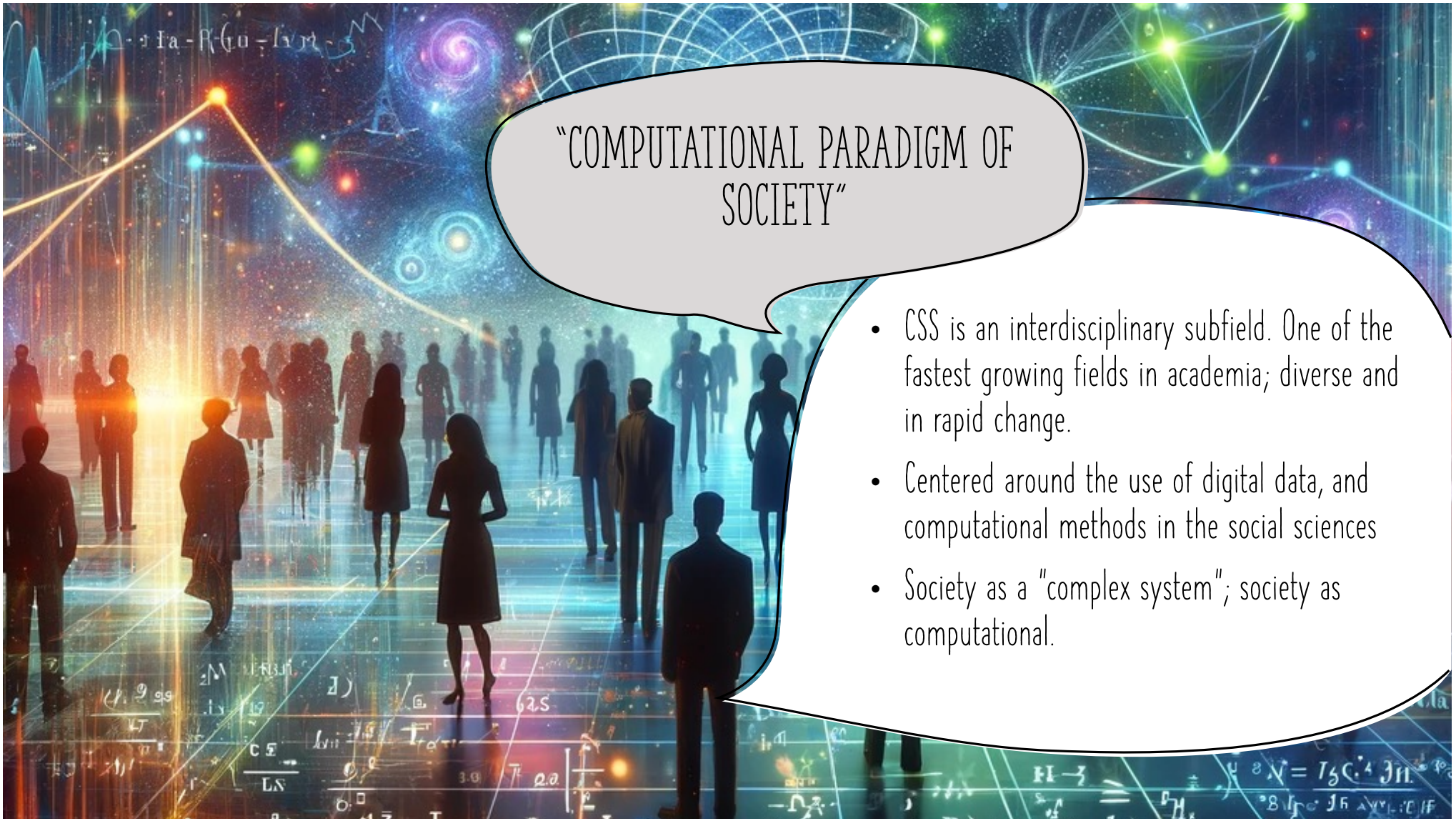
This course is new, and I will update it as we go along.

If you want to change anything, or you're unhappy with anything: just tell me!

For example: you want to move the deadlines? You want to change the order of the lectures? Want to learn more about something? You want me to talk less?

We can do it. Just let me know!

PART 1: WHAT IS COMPUTATIONAL SOCIAL SCIENCE?


"COMPUTATIONAL PARADIGM OF SOCIETY"
• CSS is an interdisciplinary subfield. One of the fastest growing fields in academia; diverse and in rapid change.
• Centered around the use of digital data, and computational methods in the social sciences
• Society as a "complex system"; society as computational.

'Digital is what gave culture the scale of physics, chemistry or neuroscience. Now we have enough data and fast enough computers to actually study the "physics" of culture'
Lev Manovich 2016

"just as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves ... we have finally found our telescope. Let the revolution begin"

Duncan Watts 2011: 266

QUANTITATIVE DATA

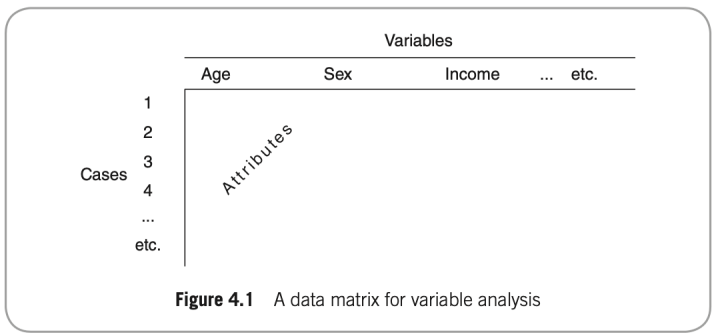
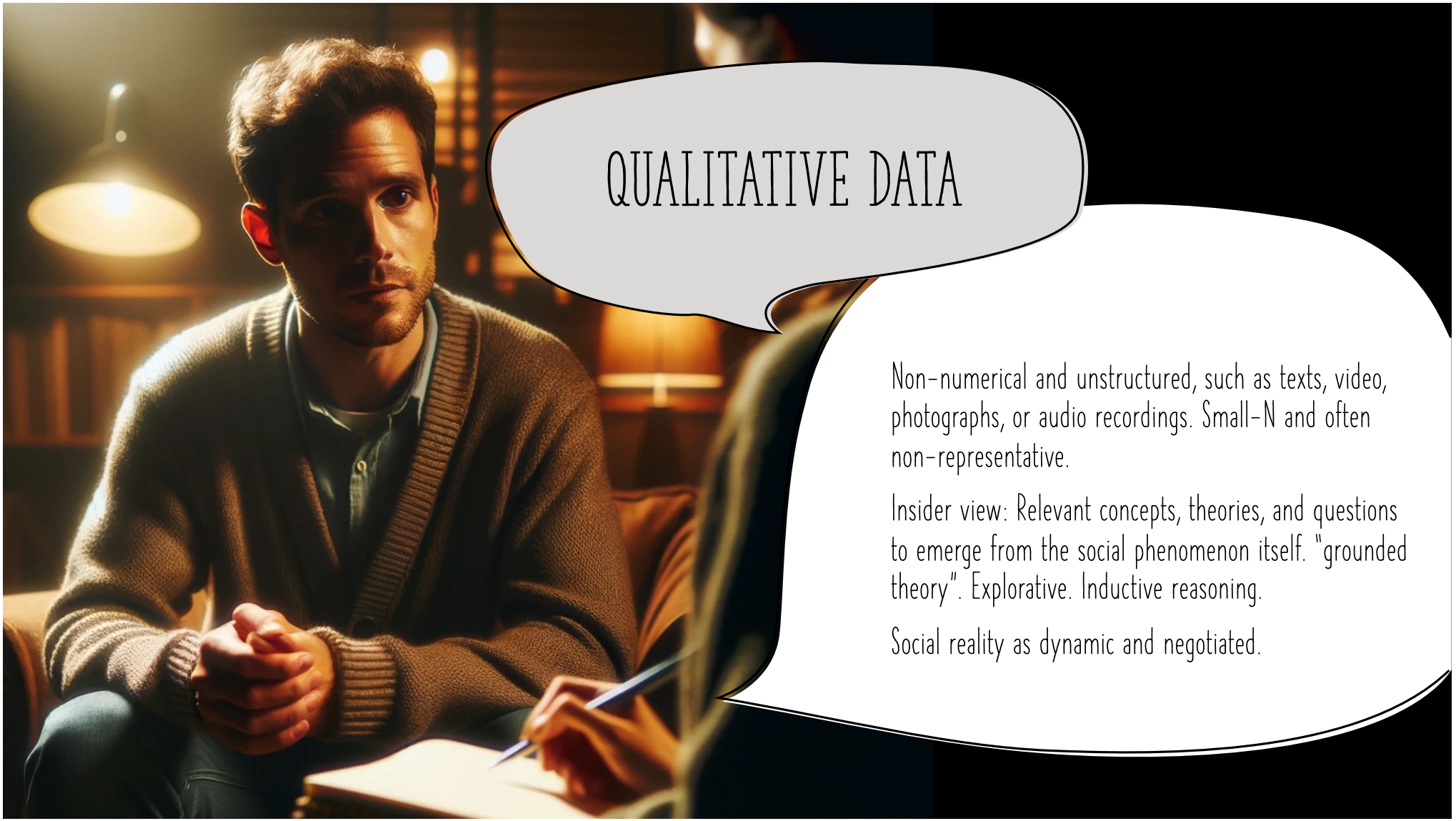


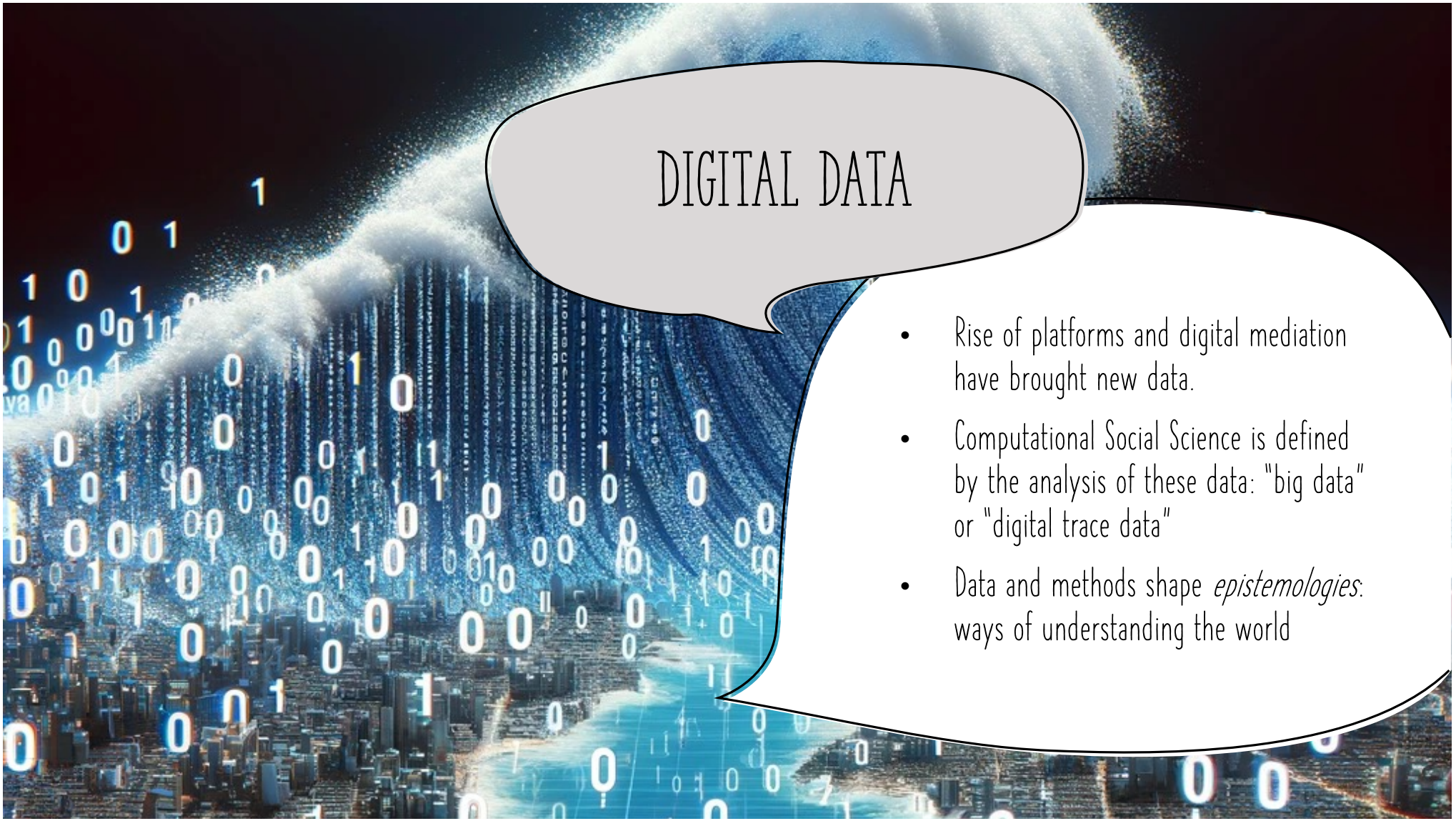
Figure 4.1 A data matrix for variable analysis

- Usually from surveys, distributed to random sample of individuals taken as representative of a certain population. List of rows with numerical or categorical answers, capturing attributes in columns.
- Goal of quant: test relationships between variables and make predictions about a wider population. *Confirmatory* and *deductive*. "Outside" view.

QUANTITATIVE DATA IS MADE FOR STATISTICS

- Quantitative methods squeeze the world into mold that fits the method.
- Variables are taken to represent some aspect of the world that the researcher believes is relevant for a phenomenon.
- *Designed* for statistical analysis: variable-variance analysis. "Gender". "Employment."


QUALITATIVE DATA
Non-numerical and unstructured, such as texts, video, photographs, or audio recordings. Small-N and often non-representative.
Insider view: Relevant concepts, theories, and questions to emerge from the social phenomenon itself. "grounded theory". Explorative. Inductive reasoning.
Social reality as dynamic and negotiated.


DIGITAL DATA
• Rise of platforms and digital mediation have brought new data.
• Computational Social Science is defined by the analysis of these data: "big data" or "digital trace data"
• Data and methods shape *epistemologies*: ways of understanding the world

THE THREE TYPES OF DATA IN THE SOCIAL SCIENCES

1. Attribute data:

E.g., survey, demographics

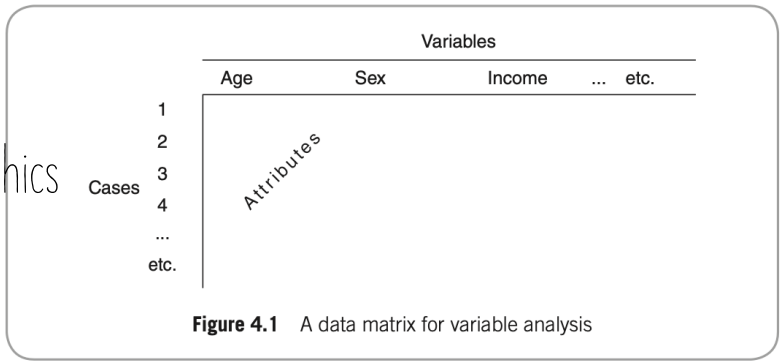


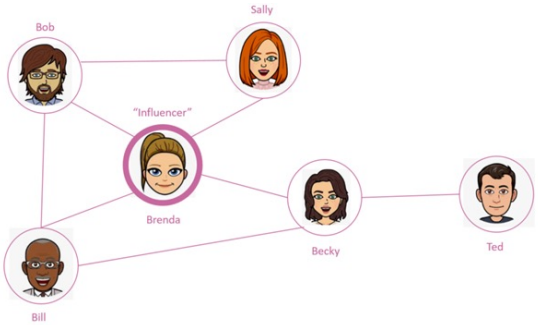
Figure 4.1 A data matrix for variable analysis

2. Ideational data:

E.g., text, speech, images, videos,...

3. Relational data:

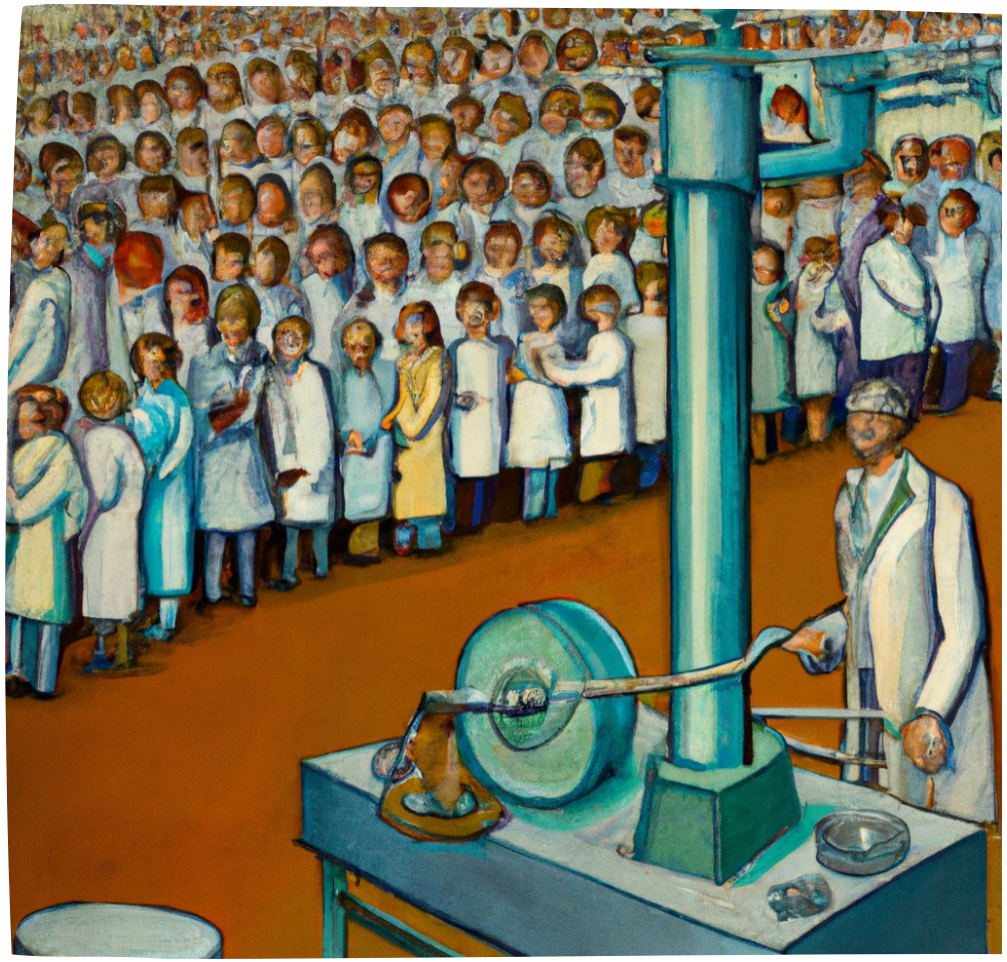
e.g. friendships, connections, etc.

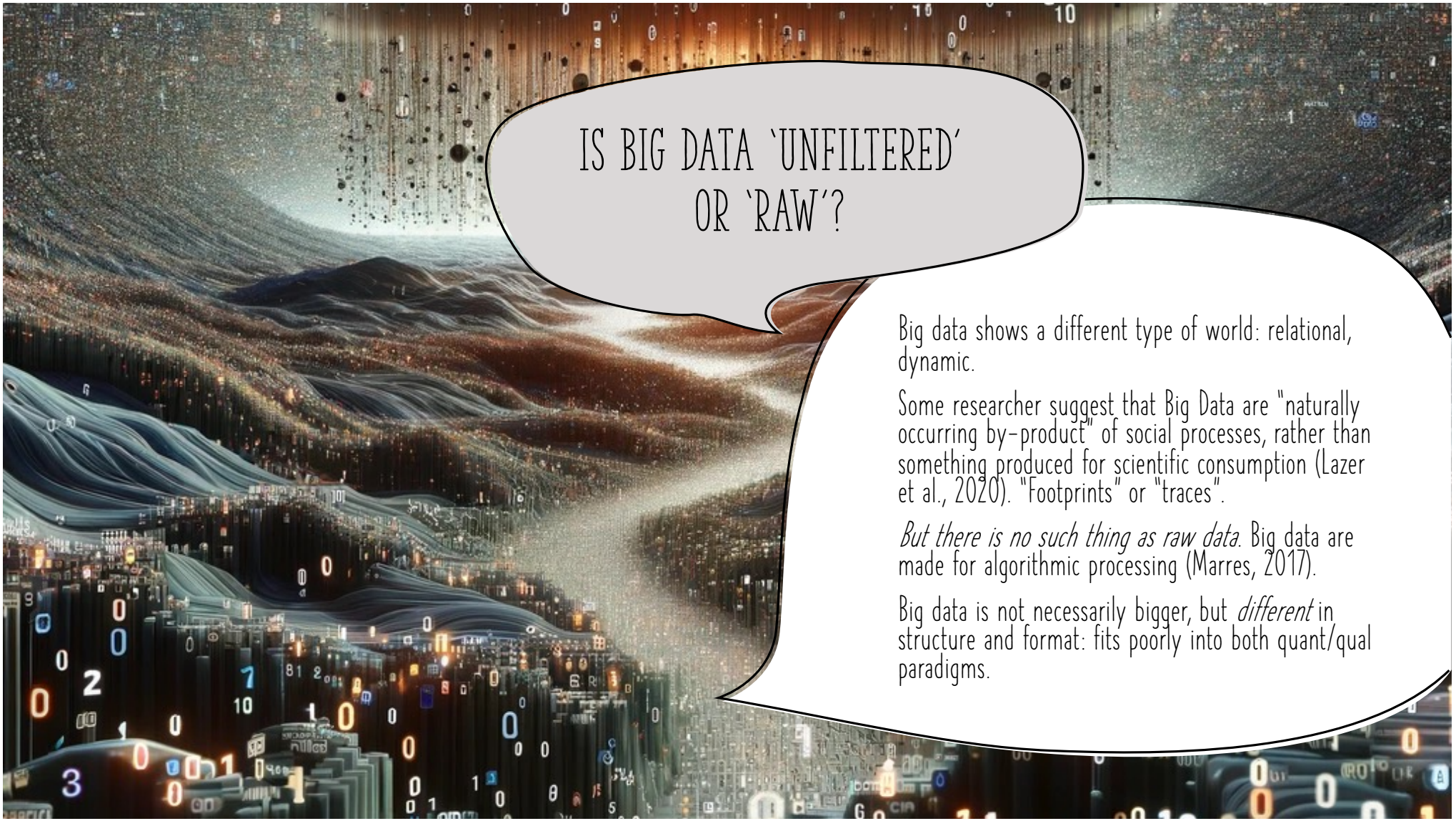


BIG DATA ANSWERS SOME LONG-STANDING CRITIQUE OF QUANTITATIVE PERSPECTIVE

"For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, ... the survey is a sociological meat grinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it."

Allen Barton, 1968




IS BIG DATA 'UNFILTERED' OR 'RAW'?
Big data shows a different type of world: relational, dynamic.
Some researcher suggest that Big Data are "naturally occurring by-product" of social processes, rather than something produced for scientific consumption (Lazer et al., 2020). "Footprints" or "traces".
But there is no such thing as raw data. Big data are made for algorithmic processing (Marres, 2017).
Big data is not necessarily bigger, but *different* in structure and format: fits poorly into both quant/qual paradigms.

"DATA CAPITALISM" OR "SURVEILLANCE CAPITALISM"

- Data are also a form of commodity; "the new oil".
- "Datafication": processes to extract information about individuals to market to data capitalist companies.
- "Dataism": the ideology that gives data its value.
- Computational Social Science can be understood as part of this mode of capitalism.

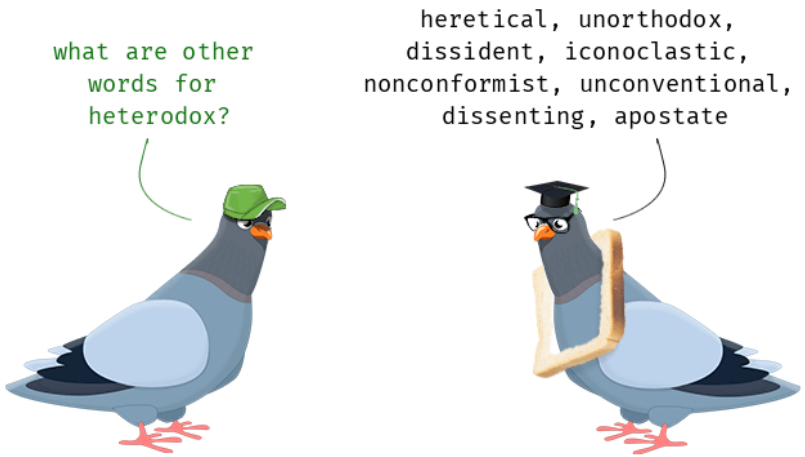
Törnberg, P., & Uitermark, J. (2021). For a heterodox computational social science. Big Data & Society, 8(2),.

We need to be aware of the context in which the data were produced.

Platforms are not neutral; they shape the structure of social life, and we are in part measuring the effects of these platforms.

FOR THE PURPOSES OF THIS COURSE

- We view CSS as a flexible label. Make it your own!
- CSS allows us to explore the boundaries outside the conventional methods: interpretation and exploration in large-scale data.
- "Heterodox CSS"
- You can combine CSS methods with a quant or a qual approach - or go beyond both!



IN SUMMARY: COMPUTATIONAL SOCIAL SCIENCE

- CSS is a new interdisciplinary field, in between computer science, social science and statistics.
- Characterized by use of digital data and computational methods to study social world
- Closely linked to big tech and Silicon Valley
- Neither conventional quantitative or qualitative, but shaped by the distinct epistemology of the digital: interactional, dynamic, fluid, relational, clusters...
- It's still being defined. You can be part of shaping it!

PART 2: HOW TO WRITE A CSS RESEARCH PROPOSAL

WHAT ARE APPROPRIATE TOPICS?

- Anything goes! Pick something you are interested in (and knowledgeable about.)
- Doesn't have to be a "computational" or "digital" topic.
- Are you part of some specific niche or community that is not broadly known? Might make for a good topic.
- Be unique and creative!

Be creative!

WHAT DATA CAN YOU USE?

We're looking for texts, relationships, or other forms of digital data.

Advice: use existing data, scraping new data will take too long!

Kaggle is a useful source of data.

I can offer: Historical political manifestos. Social media posts from parties across the world, Stormfront.org

- All of Reddit is easy to download - but hard to digest. TikTok, BluSky and YouTube are relatively easy to scrape.
- Airbnb is free and easy: Insideairbnb.com
- Parliamentary speeches
- Newspapers, headlines.
- Global Event database (GDELT), Global terrorism database
- Online experiments (e.g., MTurk)
- ... what else?

Be creative!

WHAT METHODS CAN YOU USE?

What we cover in this course:

- Natural Language Processing (Vectorization, Sentiment analysis, Topic modeling...)
- Large Language Models
- Social Network Analysis
- Machine Learning
- Online Experiments
- Agent-Based Modeling

WHAT CHARACTERIZES A GOOD RESEARCH QUESTION?

1. Clarity and specificity: It should be formulated in a way that leaves no ambiguity about what you are investigating.
2. Feasibility: Possible to answer the research question within the limited time and resources
3. Relevance: Should contribute to the existing body of knowledge or address a gap in the research.
4. Researchability: Should actually be possible to answer.
5. Originality: Should address a gap in the literature, propose a new idea, or explore understudied area.
6. Pragmatism: You need to be able to write a paper regardless of what you find.

TYPES OF RESEARCH QUESTIONS

1. Descriptive: These questions aim to describe the characteristics of a phenomenon..

   Example: What is the level of toxicity of political conversations on Twitter, BlueSky and Mastodon?

2. Causal (Explanatory) : Identify a cause-and-effect relationship.

   Example: Did Elon Musk's takeover of Twitter make the platform more toxic?

3. Exploratory: Open-ended questions aiming to learn about a little-known phenomenon.

   Example: What linguistic factors of a Twitter message influence how much engagement it receives?

EXAMPLES OF BAD RESEARCH QUESTIONS

1. "How do media affect society?" [Vague and too broad in scope; not feasible]
2. "How does changing the Facebook algorithm affect toxicity?" [Not feasible for the course]
3. "Which city has the best pizza?" [Limited academic value]
4. "Can sentiment analysis of Twitter data predict stock market trends?"

   [Fine, but hard to prove a negative. What do you write about if you fail?]

EXAMPLE IDEAS

"Is Cunningham's Law true? An online experiment on Reddit"

"How is AI discussed by different party families in their party manifestos in Europe?"

"Have mainstream news headlines have become more 'clickbaity' over the last 20 years?"

DO'S AND DON'TS OF WRITING A RESEARCH PAPER

Do tell a story

- Writing a paper is storytelling!
- Needs to be something surprising; something interesting.
- Your job is also to be a bit of a journalist. Tell a story!
- "Hourglass" structure: Start broadly, narrow down to specific research questions or hypotheses, then broaden out again to discuss the implications of your findings.

Schimel. "Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded"

Do use theory

- Use theory to *tell your story.*
- Theory is our tool, and our language
- Focus on using theory to engage with a topical and broadly interesting discussion.

Do work with middle-range theory

- We contribute not to grand theories, nor on the smallest theory, but on the middle-range
- Don't: "I'm going to test whether Goffman's theory of self-presentation is true"
- Do: "How do content recommendation algorithms on YouTube contribute to the formation of echo chambers?"

Do look at examples of similar papers

- Look at other papers doing something similar
- See how they structure their story
- Draw inspiration! (But don't plagiarize)

Do speak to your strengths

- The aim here is to produce a paper that tells a good story – so don't do something that will be unnecessarily hard and unrewarding.
- Draw on your strengths and what you know. For instance, you are younger than 99% of researchers, which gives you unique insights!
- What's some community or group that you have insider insights into?
- Do something interesting with limited time and resources!

Don't be too ambitious

- Papers are often really small: they nail down a small but broadly important claim
- Most common problem with student work is that it does too much!.

FOR REFERENCE: ARCHETYPICAL STORIES

1. Gap Identification Structure: "While significant progress has been made in understanding X, a gap remains in our knowledge of Y. In this study, we address this gap by investigating Z."

2. Problem-Solution Structure: "The major problem in field X is Y. This paper proposes a novel solution, Z, which overcomes limitations of existing approaches by..."

3. Contradiction Resolution Structure: "The dominant idea in the field is X, but recent evidence suggests Y, creating a contradiction. Our research resolves this contradiction by demonstrating that Z."

4. Sequential Development Structure: "To date, research in X has progressed from understanding basic elements (A) to more complex components (B). Building on this knowledge, we introduce a new dimension, C, which advances the field by..."

5. Comparative Analysis Structure: "Several theories exist about phenomenon X-Theory A suggests Y, while Theory B suggests Z. Through our comparative analysis, we identify the strengths and weaknesses of each and propose a unified theory that..."

6. Evolutionary Structure: "Historically, the understanding of X has evolved from A to B. We extend this evolution by introducing C, a concept that not only builds on previous work but also opens new research avenues by..."

7. Question-Driven Structure: "A key question in the study of X is 'how does Y affect Z?' This paper addresses this question through a series of experiments/analyses, revealing that..."

8. Debate Engagement Structure: "The field of X is characterized by a longstanding debate between proponents of Y and advocates of Z. We engage with this debate by providing new evidence that suggests..."

9. Innovation Introduction Structure: "Current methodologies for studying X are limited by Y. To overcome these limitations, we introduce an innovative approach, Z, which allows for more accurate/efficient investigation by..."

10. Theory Extension Structure: "Theory X provides a comprehensive framework for understanding Y. However, it falls short in explaining Z. Our work extends Theory X by incorporating W, thereby offering a more complete understanding of..."

INTRODUCTION TO PROGRAMMING!

First two weeks will be proposal writing + programming.

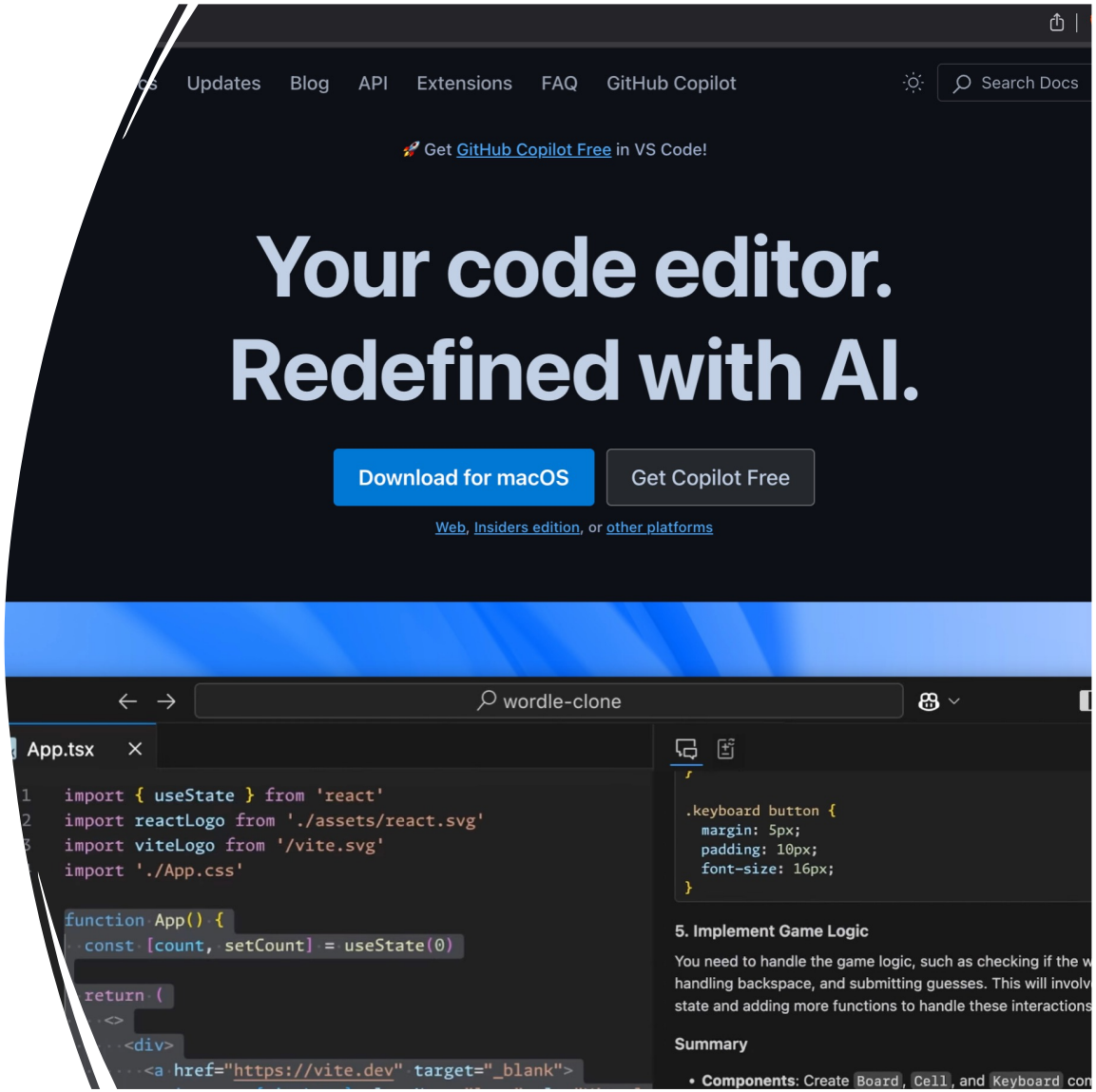
```python
from collections import Counter
import string

# Sample text
text = """Computational Social Science is fun! Data science and social science go hand in hand.
Analyzing data helps us find patterns and make decisions."""

# Function to clean and process text
def process_text(text):
    text = text.lower()  # Convert to lowercase
    text = text.translate(str.maketrans("", "", string.punctuation))  # Remove punctuation
    words = text.split()  # Split into words
    return words

# Function to count word frequencies
def count_words(words):
    word_counts = Counter()

    for word in words:  # Loop through words
        if word not in ["is", "and", "in", "us"]:  # Conditional to filter out common words
            word_counts[word] += 1

    return word_counts

# Main program
words = process_text(text)  # Clean and process text
word_counts = count_words(words)  # Count word occurrences

# Display the most common words
print("Most common words:")
for word, count in word_counts.most_common(5):  # Loop through the top words
    print(f"{word}: {count}")
```

```
Most common words:
science: 3
social: 2
data: 2
hand: 2
computational: 1
```

```python
from collections import Counter
import string

import nltk
from nltk.corpus import stopwords
nltk.download("stopwords")
stop_words = set(stopwords.words("english"))

with open('capital.txt', "r", encoding="utf-8") as file:
    text = file.read()

# Function to clean and process text
def process_text(text):
    text = text.lower()  # Convert to lowercase
    for char in string.punctuation:
        text = text.replace(char, "")  # Remove punctuation
    words = text.split()  # Split into words
    print(f"Number of words: {len(words)}")
    return words

# Function to count word frequencies
def count_words(words):
    word_counts = Counter()

    for word in words:  # Loop through words
        if word not in stop_words:  # Conditional to filter out common words
            word_counts[word] += 1

    return word_counts

# Main program
words = process_text(text)  # Clean and process text
word_counts = count_words(words)  # Count word occurrences


# Display the most common words
print("Most common words:")
for word, count in word_counts.most_common(5):  # Loop through the top words
    print(f"{word}: {count}")
```

Homework until Wednesday

1. Install Visual Studio Code
2. Get it running with Jupyter Notebook + Python
3. Find out which is the most common word in Marx's *Capital*



Homework 2

- Think up a project idea!
- Find a teammate
- Register your team on Canvas

Thursday independent work session

- You have space on Thursdays 13–17 to work in the JK building
- Collaborate, help each other, or just work in an inspiring environment

- Excellent opportunity to discuss project ideas and find a teammate!