Course Introduction and Background

Indira Sen

University of Konstanz
Research Projects in Computational Studies of Social Phenomena

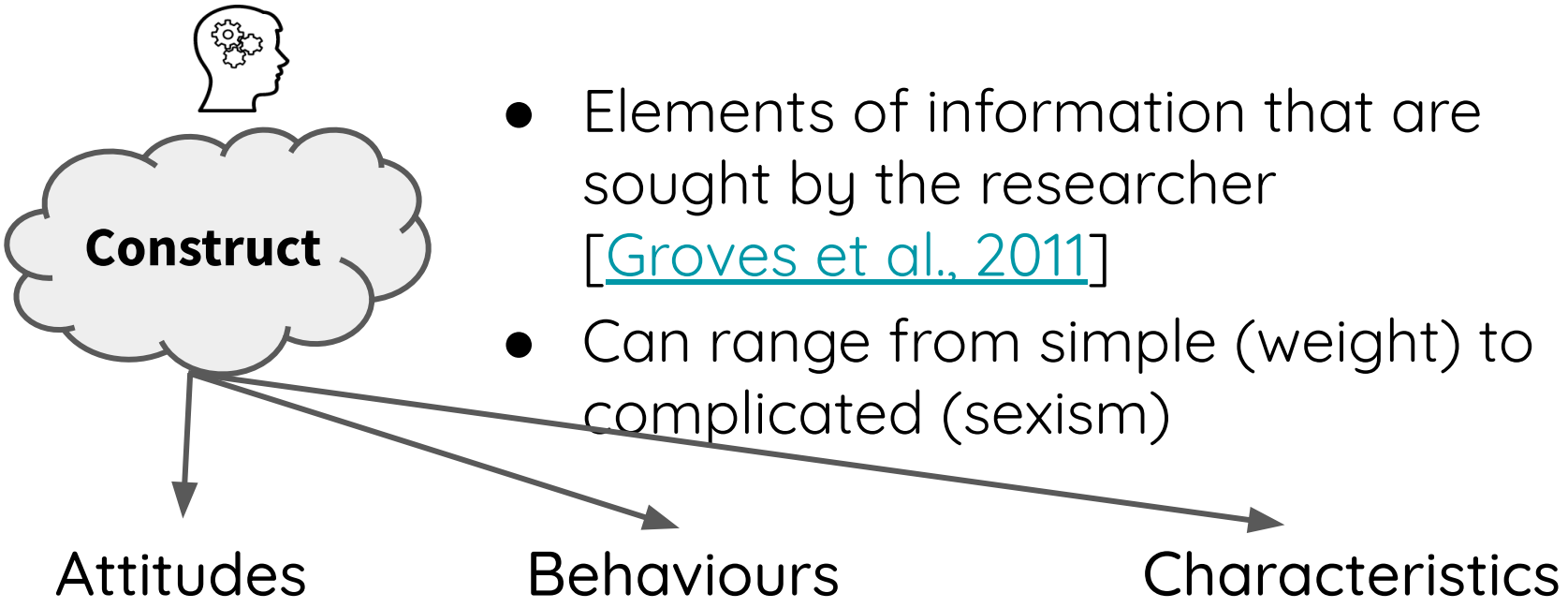
Agenda

- Background in Computational Social Science: reading pointers
- Course logistics and projects plan
- Ideas, resources, and discussion

About Computational Studies of Social Phenomena

Understanding Social Phenomena

Understanding Social Phenomena



- Elements of information that are sought by the researcher [Groves et al., 2011]
- Can range from simple (weight) to complicated (sexism)

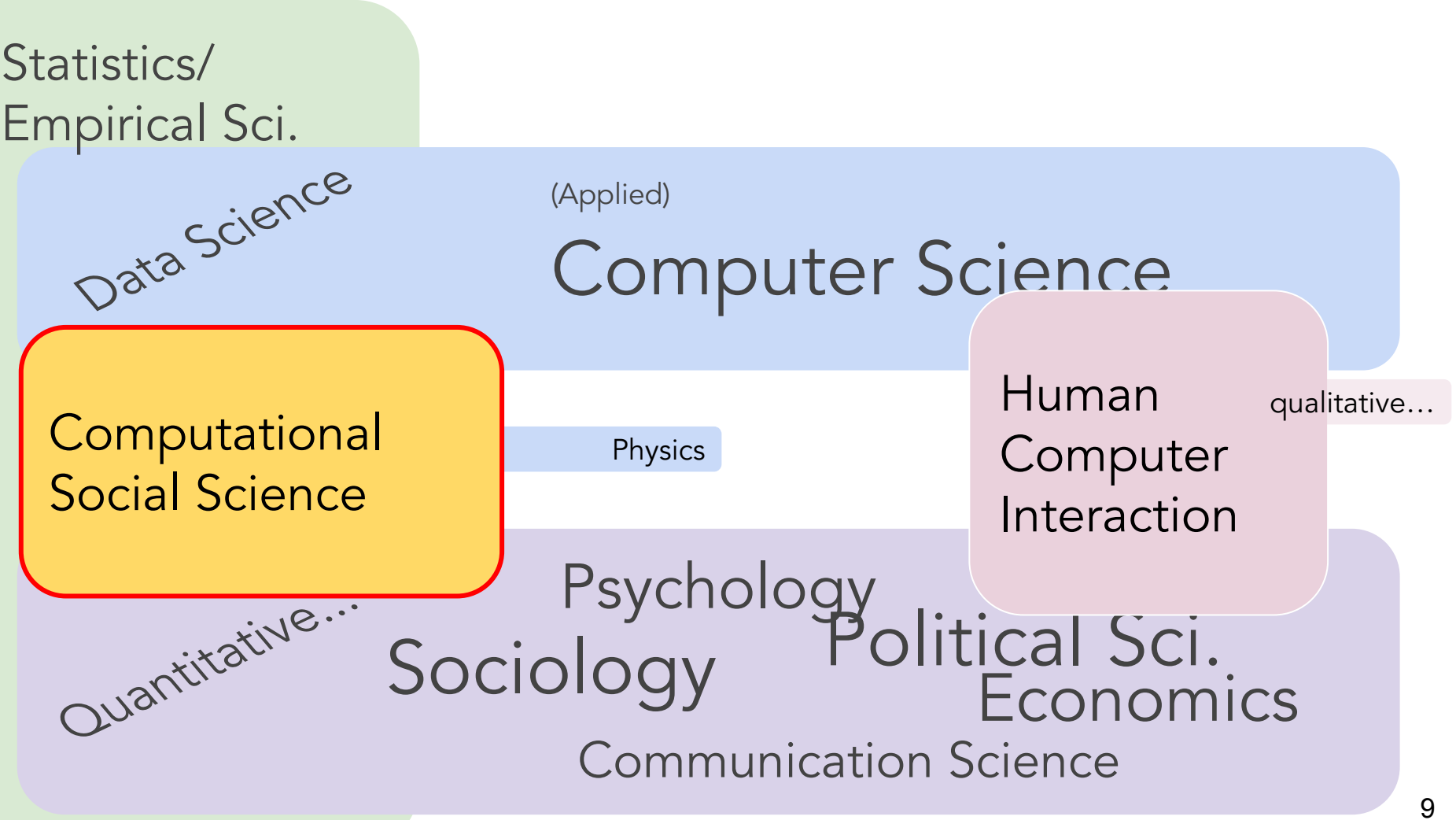
Understanding Social Phenomena

- The set of **people** to be studied [Groves et al., 2011]
- Can be easily defined (all the preschool students in a city) to more difficult (all refugees)
- Usually a national population, can also be any “system population”

Target Population

About Computational Studies of Social Phenomena

About Computational Studies of Social Phenomena
or just Computational Social Science?


Statistics/
Empirical Sci.
Data Science
(Applied)
Computer Science
Computational
Social Science
Physics
Human
Computer
Interaction
qualitative…
Quantitative…
Psychology
Sociology
Political Sci.
Economics
Communication Science

Highly Interdisciplinary research field

- multidisciplinary entry points
- diverse range of approaches (also mixed methods)
- many exploratory studies

CSS and Digital Traces

SOCIAL SCIENCE

Computational Social Science

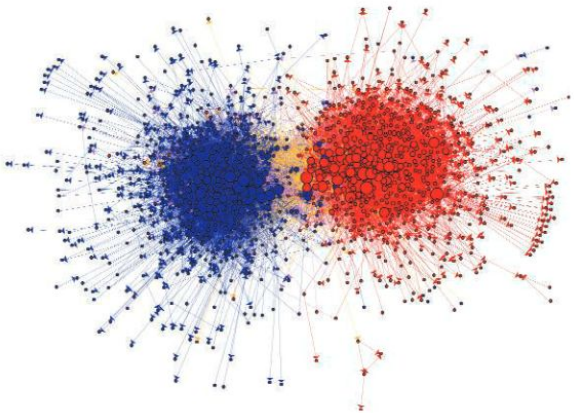
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

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.

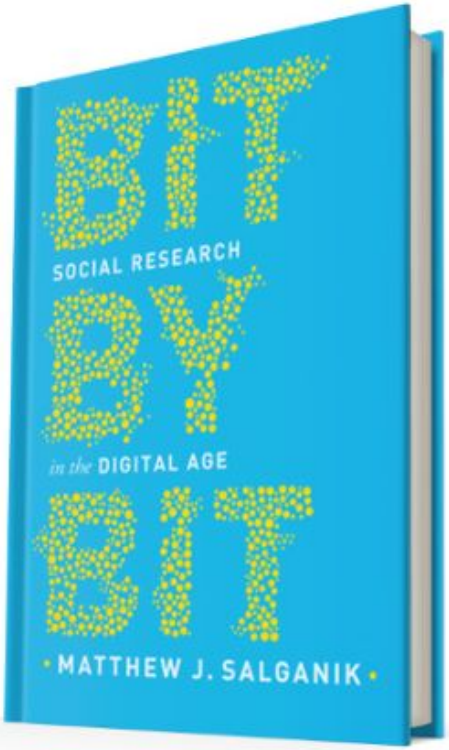
The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But 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



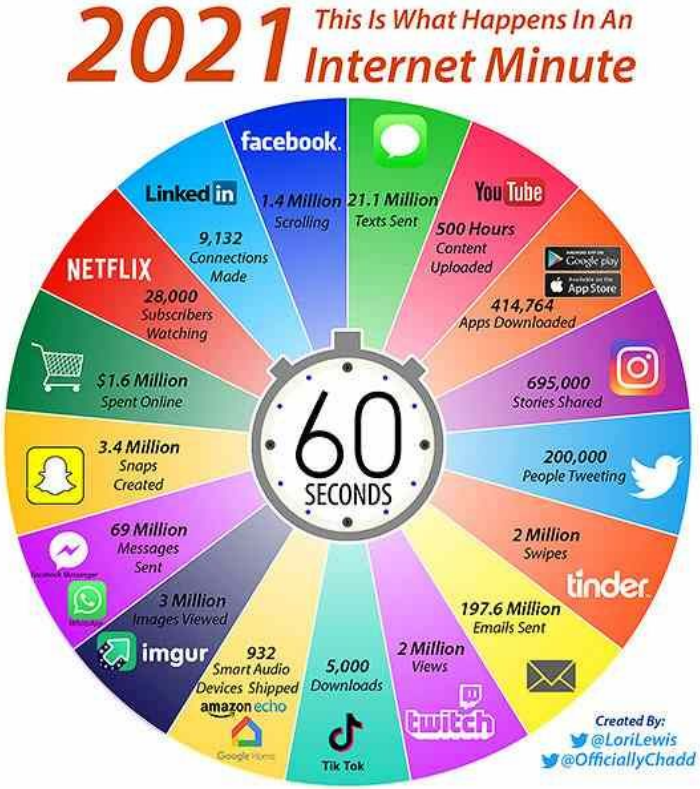
[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA.

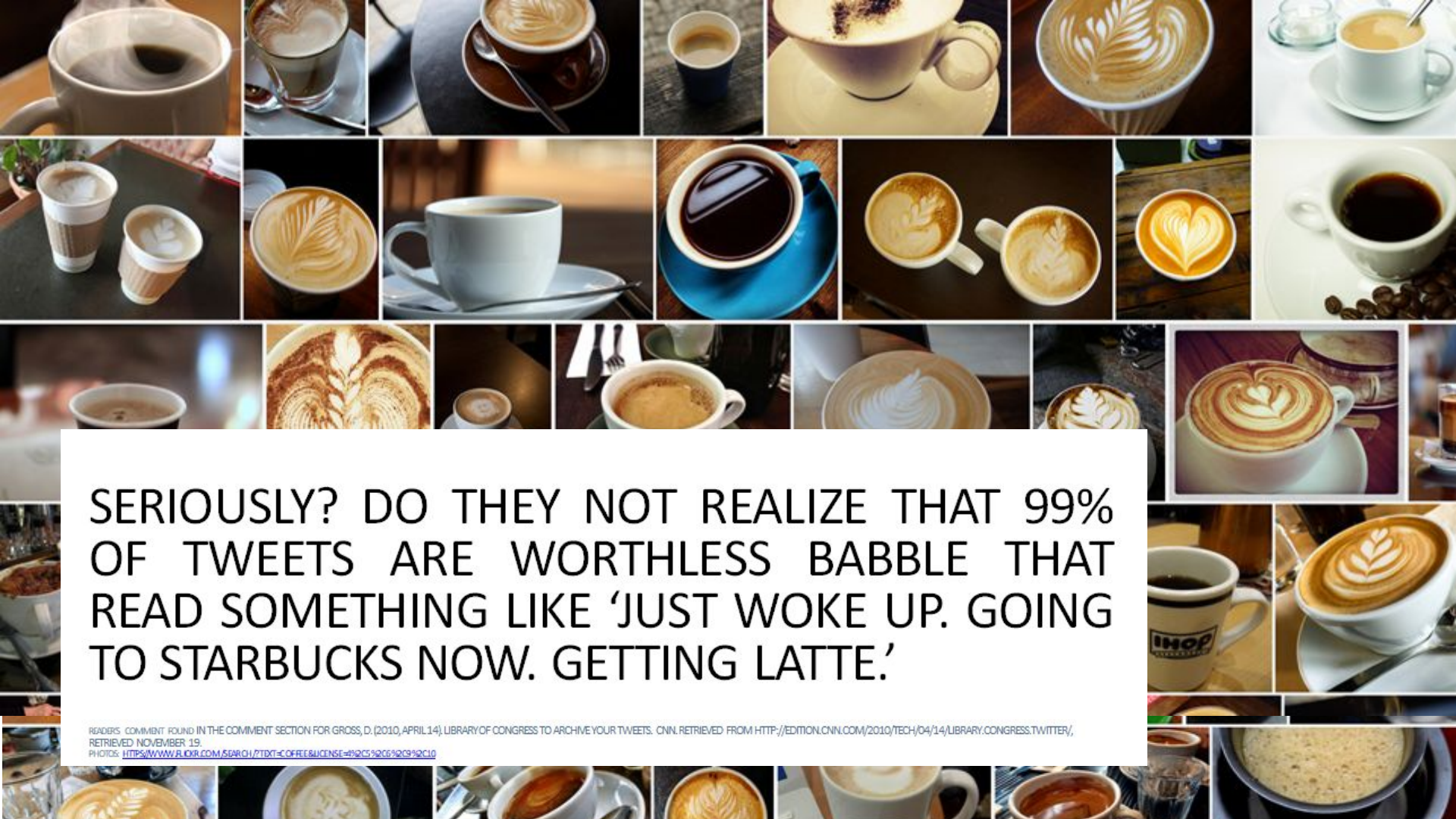
Lazer, David, et al. "Computational social science." *Science* 323.5915 (2009): 721-723.



Salganik, Matthew J. *Bit by bit: Social research in the digital age.* Princeton University Press, 2019.

Digital Traces: So much data!
(…about people's behavior and attitudes)





SERIOUSLY? DO THEY NOT REALIZE THAT 99% OF TWEETS ARE WORTHLESS BABBLE THAT READ SOMETHING LIKE 'JUST WOKE UP. GOING TO STARBUCKS NOW. GETTING LATTE.'

READERS COMMENT FOUND IN THE COMMENT SECTION FOR GROSS, D. (2010, APRIL 14). LIBRARY OF CONGRESS TO ARCHIVE YOUR TWEETS. CNN. RETRIEVED FROM HTTP://EDITION.CNN.COM/2010/TECH/04/14/LIBRARY.CONGRESS.TWITTER/, RETRIEVED NOVEMBER 19.
PHOTOS: HTTPS://WWW.FLICKR.COM/SEARCH/?TEXT=COFFEE&LICENSE=4%2C5%2C6%2C9%2C10

Studying digital traces as a new type of data

- Researchers value social media as a new type of data
- Previously „ephemeral data“ become visible
- Immediate – quick reaction to events
- Structured
- „natural“ data

“What I find really interesting is that structure becomes manifest in internet communication. So it’s the first time in history actually that we can, that social structures between people become manifest within a technology. (...) They become visible, they become crawlable, they become analyzable.”

Kinder-Kurlanda, Katharina, and Katrin Weller. "“I always feel it must be great to be a hacker!” the role of interdisciplinary work in social media research." *Proceedings of the 2014 ACM conference on Web science*. 2014.

Some typical studies

Computer Science

Social Sciences

2010 From tweets to polls [O'Connor et al.]

2014 Predicting tie strength with social media [Gilbert et al.]

2019 Investigating commentator bias in football broadcasts [Merullo et al.]

2013 Text as data [Grimmer et al.]

2013 Big Data in Survey Research: AAPOR Task Force Report [Japec et al.]

2019 Combining surveys and digital traces [Stier et al.]

Studying digital traces as a new type of data

Data on social software and platforms is not created for research purposes.

This leads to challenges in

- Accessibility
- Quality
- Interpretation
- Ethics

- Researchers value social media as a new type of data
- Previously „ephemeral data“ become visible
- Immediate – quick reaction to events
- Structured
- „natural“ data

“What I find really interesting is that structure becomes manifest in internet communication. So it’s the first time in history actually that we can, that social structures between people become manifest within a technology. (...) They become visible, they become crawlable, they become analyzable.”

Kinder-Kurlanda, Katharina, and Katrin Weller. ""I always feel it must be great to be a hacker!" the role of interdisciplinary work in social media research." *Proceedings of the 2014 ACM conference on Web science.* 2014.

Research ethics

- Research ethics practices are also still evolving
- Main focus on privacy
- Lack of informed consent
- “participants” not expecting research activities on their data [Fiesler & Proferes, 2018]
- Different assumptions for different types of user groups (e.g. vulnerable groups)
- Potential starting point: AoIR ethics guidelines

It's not a simple overlap or addition of disciplines [Wallach'2018]

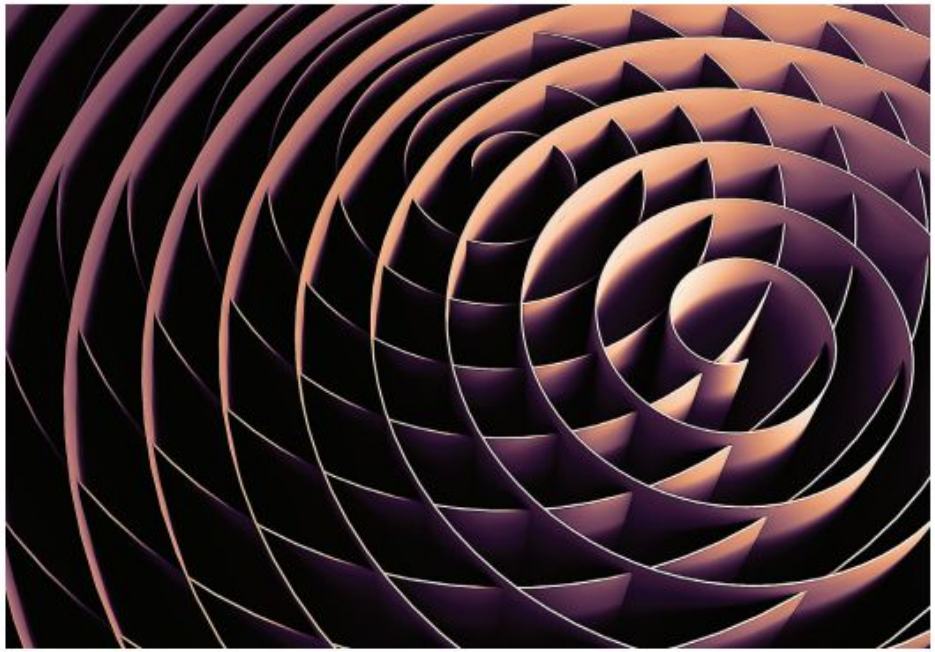
Computational Social Science ≠ Computer Science + Social Data

The important intersection of computer science and social science.

THIS VIEWPOINT IS about differences between computer science and social science, and their implications for *computational* social science. Spoiler alert: The punchline is simple. Despite all the hype, machine learning is not a be-all and end-all solution. We still need social scientists if we are going to use machine learning to study social phenomena in a responsible and ethical manner.

I am a machine learning researcher by training. That said, my recent work has been pretty far from traditional machine learning. Instead, my focus has been on computational social science—the study of social phenomena using digitized information and computational and statistical methods.

For example, imagine you want to know how much activity on websites



This Course: Objectives

- use digital trace data to answer social research questions
- develop computational models for large-scale analysis of digital trace data
- perform statistical analysis for empirical social research

A Project-based Seminar

- The course contains lectures on recent research in Computational Social Science and Social Data Science, and a large practical part to develop a data analysis project under my supervision.
- Projects are done in **groups of two-three students**, but grades are individual
- The purpose of the project is to learn by doing in a small data science project based on some suggested ideas*, focusing on data management, analysis, statistics, and interpretation.
- Projects have to start from a research question and have an empirical focus, but at the same time critically reflect on methods, conclusions, and limitations.

* you can also pitch your own project, but please discuss with me first

Course Logistics

[April-June 10th] initial overview of CSS research, description and discussion of project ideas, overview of methods:

exploratory data analysis

text-as-data

network analysis

reproducible research pipeline

[June 17th-July 29th] project discussion, guidance, and consulting ---> weekly office hours, but you can also email me or request ad-hoc appointments

If there are some methods you want to try and need help with them, we can also do a presentation on that during one of the office-hours (Please let me know one week in advance)

Course Schedule and Grade Allocation

date	time	type	title
Apr 8	1:30-3:00	lecture 0	course logistics, some ideas
Apr 15	1:30-3:00	lecture 1	examples of research with web and social media data
Apr 22	1:30-3:00	lecture 2	description of datasets
Apr 29	1:30-3:00	no class	
May 6	1:30-3:00	no class	
May 13	1:30-3:00	lecture 3	exploratory data analysis
May 20	1:30-3:00	lecture 4	text-as-data methods
May 27	1:30-3:00	no class (Corpus Christie)	
Jun 3	1:30-3:00	lecture 5	network science methods
Jun 10	1:30-3:00	lecture 6	reproducible research pipeline
Jun 17	1:30-3:00	lecture 7	project guidance and discussion
Jun 24	***1:30-3:00***	***lecture 8***	***project registration deadline (10%)***
Jul 1	1:30-3:00	lecture 9	project guidance and discussion
Jul 8	1:30-3:00	lecture 10	project guidance and discussion
Jul 15	1:30-3:00	lecture 11	project guidance and discussion
Jul 22	***1:30-3:00***		***midway presentation (15%)***
Jul 29	1:30-3:00	lecture 12	project guidance and discussion
mid-August	***TBD***		***Final presentation (25%)***
end-August	***TBD***		***Final Report (50%)***

Suggested Ideas (We'll discuss these in detail in Lecture 3)

1. **Reddit Posts and Comments about Politicians:** How do people discuss female politicians?
2. **Stance Detection Benchmark:** How well do current computational models perform at detecting stance towards different entities and topics?
3. **Annotator (Dis)agreement:** Characterizing differences in annotator perspective for subjective constructs
4. **X (Twitter) and Reddit discussion about Football:** How do people talk about non-white players?
5. **Lost in Simplification? English vs. Simple Wikipedia:** How does content and framing diverge in Simplified Wikipedia?
6. **One Day on Twitter**: What goes on in one day on Twitter?

Project steps

- **1. Form a group and choose a topic**
  - Start now! It's never too soon to think and start readings about a topic. There are suggested readings for each of the ideas mentioned earlier at the end of this slide deck.
  - You can get input from me via email or around lectures
- **2. Project registration**.
  - Special session to help with registrations on 17th June
  - More guidance sessions before presentations
- **3. Project presentations**.
  - Focus on the research design and hypotheses in the **mid-point presentation**
  - Focus on some of the interesting findings or results in the **final presentation**
  - There is more time to work and improve after the presentations. Results do not have to be final at the presentation, they are updates on your project state
- **4. Submit the final report**.

What you will be graded on

1. how 'good' your results are
    - good != positive direction, effect size, statistical significance, etc
    - but how robust, replicable, and reliable your results are
    - negative results are also results
    - how well your analysis supports your claims
2. research design [controlling for confounders, validation]
3. theoretical, statistical, and practical rigor

How you will be graded

Total credits: 4

initial abstract submission [200 words] [10%] ---> extended abstract describing the overall research questions, methods, and expected results

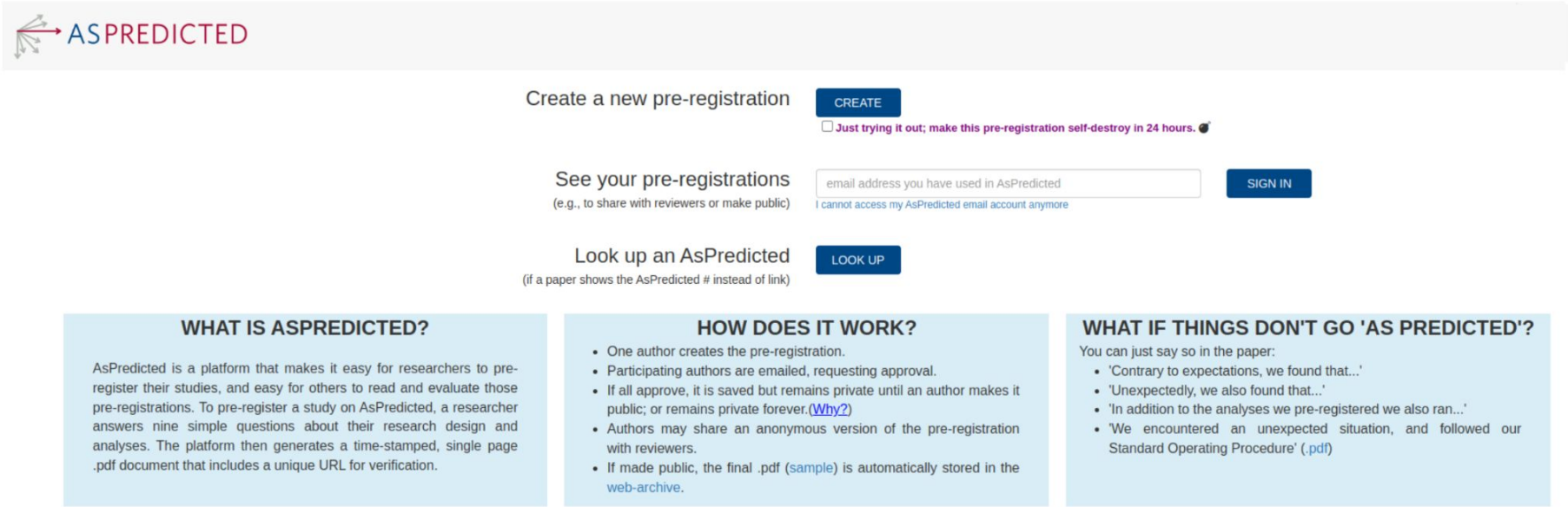
midway presentation [10 mins**] [15%] ---> initial progress and results (exploratory data analysis, research design overview)

final presentation [15 mins**] [25%] ---> short intro, pick the most interesting results, discuss implications

final report [50%] ---> incorporate feedback from the presentation

**we will be very strict about presentation timings in this course, so prepare accordingly

Initial Abstract Submission: Similar to Pre-registrations



https://aspredicted.org/

Initial Abstract Submission [10% of the grade]

Submit your initial abstract via email to indira.sen@uni-konstanz.de

Each abstract a short text that includes the following:

1. Project title
2. Names and email addresses of group members
3. Research question(s)
4. Planned data analysis to address the questions including:
    - precise data sources
    - exclusion criteria from retrieved data
    - methods used for all measurement of variables
    - statistical tests and outcomes
    - description of other visualizations or outputs (e.g. models)

Mid-Point [15%] and Final [25%] Presentations

- **10 and 15-minute** presentations about the state of results of their projects

- Students get questions from the audience (you and some other CSS researchers) regarding the topic and feedback on how to improve the project for the final report.

- Each presentation has to contain four parts:
  - Research question and motivation
  - Data and methods
  - Results
  - Conclusion and critique [only for the final presentation]

- All members of the group have to participate in their presentation and answer questions.

Final Report [50%]

Each report should consist of:

- **Abstract:** 250-word summary of your project including a formulation of its research questions and an overview of methods and results
- **Motivation:** What question(s) do you seek to answer and why? Elaborate on how this research matters in terms of science, technology and/or society
- **Research background:** What previous work is related to this project and what can we learn from it? Do not just cite papers but anything relevant and reflect about the cited contributions on your text
- **Data:** Describe your data sources including descriptive statistics and plots. Carefully document all references to data sources.

Final Report [50%] (contd….)

- **Methods:** Explain how you filtered data, normalized values, computed additional variables, etc. Detail the statistical analyses and other methods you will apply to assess your research question(s). Carefully document all references to methods and packages used.

- **Results:** Expose the results of your analysis including tables and figures that communicate and illustrate those results

- **Discussion:** Evaluate answers to the question and their reliability. Identify limitations and alternative explanations for your results.

- **Conclusion:** Short summary of the outcome of the project

- **Author contributions:** what each team member worked on [remember, all members need not get the same grade]

Submitting the Final Report

Send a final report as a PDF document (max. 10 pages, min. font size 11pt) via email to indira.sen@uni-konstanz.de.

References do not count towards the page limit.

Plots should be correctly shown (named axes, visible scales) and writing has to be understandable.

Projects can contain a link to a Github repository including the code to produce results, datasets if they can be shared, and additional figures or tables that can be referenced from the project report.

Extra points are given when projects are based on open science principles (e.g. data and code is available in a Github repository).

Reading related to project ideas

Reddit Posts and Comments about Politicians:

Hofmann, Valentin, Hinrich Schütze, and Janet B. Pierrehumbert. "The reddit politosphere: a large-scale text and network resource of online political discourse." ICWSM'2022.

Marjanovic, Sara, Karolina Stańczak, and Isabelle Augenstein. "Quantifying gender biases towards politicians on Reddit." PloS one 2022

Stance Detection Benchmark:

AlDayel, Abeer, and Walid Magdy. "Stance detection on social media: State of the art and trends." Information Processing & Management 2021

Schiller, Benjamin, Johannes Daxenberger, and Iryna Gurevych. "Stance detection benchmark: How robust is your stance detection?." 2021

Annotator (Dis)agreement:

Gordon, Mitchell L., et al. "Jury learning: Integrating dissenting voices into machine learning models." Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022.

Davani, Aida Mostafazadeh, Mark Díaz, and Vinodkumar Prabhakaran. "Dealing with disagreements: Looking beyond the majority vote in subjective annotations." TACL 2022

Reading related to project ideas

X (Twitter) and Reddit discussion about Football:

Vidgen, Bertie, et al. "Tracking abuse on Twitter against football players in the 2021–22 Premier League Season." 2022

Merullo, Jack, et al. "Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts." EMNLP 2019.

Lost in Simplification? English vs. Simple Wikipedia:

Yasseri, Taha, András Kornai, and János Kertész. "A practical approach to language complexity: a Wikipedia case study." PloS one 7.11 (2012): e48386.

Jatowt, Adam, and Katsumi Tanaka. "Is Wikipedia too difficult? comparative analysis of readability of Wikipedia, simple Wikipedia and Britannica." 2012.

One Day on Twitter:

Pfeffer, Juergen, et al. "Just another day on twitter: A complete 24 hours of twitter data." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 17. 2023.

What type of social phenomena are *you* interested in studying?