Introduction to Computational Social Science

JProf. Dr. Claudia Wagner
University Koblenz Landau
GESIS – Leibniz Institute for the Social Sciences
About me

- Professor in Computer Science at Univ. of Koblenz-Landau
- Scientific director of the department „Computational Social Science“ at GESIS
Stefan Vujović
svujovic@uni-koblenz.de
Goal for Today

- What is computational social science?
- What to expect from this course?
- Logistics
Was ist Computational Social Science?

Computer Science + Social Science?
What is

- … Social Science?
- … Computer Science?

R. P. Feynman on social sciences

(1981)
https://vimeo.com/118188988

Richard Feynman on Computer Science — Talk at Bell Labs (19...

(1985)
http://www.youtube.com/watch?v=IL4wg6ZAFIM&list=FLg5a1LWfGRn4grwWIwCzcLA
Social Science Questions + Computational Methods and Systems + New Types of Data
Social Scientists are interested in how people

- think/feel/behave in social situations (social psychology),
- relate to each other (sociology),
- govern themselves (political science),
- handle wealth (socio-economics) and
- create culture (anthropology)
Unit of Analysis

Political Science, History, Sociology

Societal
Institutional
Group
Interpersonal
Intrapersonal

Organization science
Sociology
Psychology
Frequently Used Research Methods

- Laboratory experiments
- Surveys
- Case research
- Ethnographic research
Survey research is a research method involving the use of questionnaires and interviews to gather information from individuals.

Systematic collection of information using standardized procedures

Measurement instrument

Standardized Stimulus

Census taker visits a Romani family living in a caravan, Netherlands 1925

https://en.wikipedia.org/wiki/Census
Questions are designed to measure a theoretical constructs

- Scales are developed for this construct (e.g. political leaning, sexism)
- Multiple items (questions) are designed that should measure all facets of the theoretical construct

- Validity is assessed by experts and by reusing the same scale in different settings
- Reliability of the scale is measured by inter-item correlations (Cronbach’s Alpha), test-retest reliability
Some Limitations and Problems

- Researcher Bias

- Non-response Bias
  - Response rate of 10-20% are normal.

- Sample Bias
  - some subgroups (e.g. women) are more willing to respond
    → If you don’t know anything about your target population before, you cannot compensate the bias
  - Probabilistic (stratified) samples are not always possible.

- Social Desirability Bias, Memory issues
Types of Surveys

- Longitudinal Survey (Cohort study)
  - uses time as the main variable
  - observe how a small sample changes and fluctuates over time.

- Cross sectional survey
  - takes a snapshot of a population at a certain time
  - aims to measure effects of one or more independent variables upon a dependent variable of interest at a given point in time

- Panel: longitudinal and cross sectional
There are two general types of found data:

- **Accretion** - a build-up of physical traces
- **Erosion** - the wearing away of material

Unobtrusive research

• Unobtrusive or non-reactive

• Observing people without them knowing

• Methods of studying social behaviour without affecting it

• Such methods can be qualitative or quantitative.

• Surveys and Experiments are reactive (or obtrusive)
Found data is everywhere
What Happens in an Internet Minute?

- 639,800 GB of global IP data transferred
- 135 Botnet infections
- 6 New Wikipedia articles published
- 20 New victims of identity theft
- 204 million Emails sent
- 20 million Photo views
- 1,300 New mobile users
- 100+ New Linkedin accounts
- 320+ New Twitter accounts
- 100,000 New tweets
- 47,000 App downloads
- 83,000 In sales
- 277,000 Logins
- 6 million Facebook views
- 2+ million Search queries
- 30 Hours of video uploaded
- 1.3 million Video views

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second
“Between the dawn of civilization and 2003, we only created five exabytes of
information; now we’re creating that
amount every two days.”

Eric Schmidt, Google (and others)
Data is generated by human
Smarter Devices
Ubiquitous Connectivity
New kinds of data (macro-scale)

*Human mobility in societies*

Check-ins (Foursqaure, Gowalla, Twitter, …)

New kinds of data (meso-scale)

- *Urban movement analysis from GPS/phone data*

The Amsterdam Real Time Project

New kinds of data (micro-scale)

Social Sensing via RFID

Building connections: How scientists meet each other during a conference (Génois et al. 2017, Working Paper)
Global patterns of physical activity

Number of daily steps taken by ~700,000 people from 11 countries, collected via the smartphone Argus activity monitoring app.

https://www.nature.com/articles/nature23018
In more walkable cities, activity is greater throughout the day and throughout the week, across age, gender, and body mass index (BMI) groups.

Activity inequality is a better predictor of obesity prevalence in the population than average activity volume.

https://www.nature.com/articles/nature23018
The **digital world is tracking the social world more and more closely.**

This enables us to **use computation** to

- discover patterns,
- build models,
- validate social theories and
- learn about societies.
Potential Limitations

- not representative
  ➢ population biases
- poor in attributes
  ➢ unknown demographic attributes
- dominated by a few
  ➢ Power law phenomena
  ➢ Self-selection
- shaped by systems
  ➢ algorithmically mediated
- noisy
  ➢ users != people

Potential Advantages

- highly granular
  ➢ high temporal resolution
  ➢ High spatial resolution
- rich in structure
  ➢ multi-relational data
- rich in sources
  ➢ integration of different data types
- cheap and big
Computational Social Scientists need knowledge about

- computational methods (data cleaning, data mining, machine learning, statistics, computational models)
- new types of data
- social science theories and questions
What types of Research Questions do social scientists ask?

What types of Research Questions do computer scientists ask?
Research Questions

- Descriptive Research Questions
  - How often does X happen?
  - What are the properties of X?

- Comparative Questions
  - How does X compare to Y?

- Relational Questions
  - What is the relation between X and Y?

- Causal Questions?
  - Does X cause Y?
Status attainment: why do people obtain their position in society? Status attainment is affected by both achieved factors, such as educational attainment, and ascribed factors, such as family income.

Radicalization or Crime: why do some individuals radicalize / become criminal?

Voter Turnout: what impact will who vote?

Voting behavior: which factors explain political leaning?

Happiness: what makes people happy? Which factors explain happiness?

Opinion Formation: how do people form opinions? Which factors impact the opinion formation process most?

Mental Health: why does religion impact mental health? What mechanism explains that?
Examples

- **Macro-level**
  - **Inequality**: which factors and mechanisms explain the emergence of inequality?
  - **Polarization**: what leads to polarization in society?

- **Meso-level**
  - **Team performance/creativity**: what makes team successful? Which factors explain team success?
  - **Cooperation**: what impacts that people cooperate, i.e. contribute to a public good?
  - **Social Norms**: how do groups of people reach consensus?
CSS Research Questions

- Research about groups of users, institutions and societies
  - Use digital behavioral data and develop methods to address traditional social science research questions

- Research about socio-technical systems
  - Measure social issues (e.g. inequality, conflicts) in socio-technical systems → which mechanism explains how they arise?
  - Some of these systems impact society (e.g. Wikipedia)
    - Measure and explore their impact on offline world
    - Explain social phenomena in these systems (e.g. popularity of content and users; activity or inactivity of users)
Agenda

- What is computational social science?
- What to expect from this course?
- Logistics
What will you learn?

- Social Science Problems and Computational Methods
  - Read and evaluate research applying computational methods to social science problems

- Learn how to collect and manipulate quantitative data

- Develop skills necessary to analyze large and heterogeneous datasets
Tentative Syllabus

- Data and Measurements
- Text Mining
  - Supervised & unsupervised methods
- Networks
  - Static and dynamic networks
  - Dynamic processes in network
- Analytic Sociology (Guest Lecture)
- Agent Based Modelling (Guest Lecture)
- Social Issues: Inequality (maybe also polarization)
- Social issues in socio-technical systems
- Causal Inference & Experiments
Evaluation

- **50% Exam (25.7.)**
- **50% Research Project (Exercise)**
  - **Task 1 (20%):** individual → propose research project: come up with good questions, datasets and explain methods that you want to use to answer questions
  - **Task 2 (30%):** form teams and work on a research project together; write a small report about the project and present it at the end of the semester
    - Use python and create notebooks
    - Report should only contain 2-4 figures/tables that clearly answer your research question
    - Connect your work with existing research (related work)
Timeline

- **Task 1:** submission deadline 30.4.
- **2.5.** you present your ideas in class (~5min)
  - What question
  - What data (found data or will you generate new data)
  - Which number/plot will answer the question
  - Why is the question important
  - Has someone answered the question before? What related work exists?
- **9.5.** we select the best ideas and form teams
  - You work on the projects in May and June
- **6.6.** status report from all teams
- **11.7.** and **18.7.** Final Presentations
- **25.7.** Final Report & Notebook submission
Example 1

- **Dataset**
  - Politicians on Wikipedia
  - Register of politicians (e.g. Bundestag)

- **Research Question:**
  - Has Wikipedia become more gender equal over time?

- **Methods:**
  - Proportion of articles about male and female politicians over time → compare this with offline proportions
  - Article length of male and female politicians over time
Example 2

- Dataset
  - Politicians on Wikipedia
  - Register of politicians (e.g. Bundestag)

- Research Question:
  - Who is covered on Wikipedia?

- Methods:
  - Which features are predictive for being covered on Wikipedia?
Other pointers

- Google books ngrams
- Google correlate, Google trends
- Network data
- Eventregistry
  - [http://eventregistry.org/](http://eventregistry.org/)
- Official statistics about Germany
  - [https://www.destatis.de/DE/ZahlenFakten/ZahlenFakten.html](https://www.destatis.de/DE/ZahlenFakten/ZahlenFakten.html)
- Survey data
- Predict official statistics with yelp data
  - [http://www.hbs.edu/faculty/Publication%20Files/18-022_b618d193-9486-4de3-abc4-232e1baecbeb.pdf](http://www.hbs.edu/faculty/Publication%20Files/18-022_b618d193-9486-4de3-abc4-232e1baecbeb.pdf)
Coding Style and Good Computing Practices

Based on


Data Management

- Save raw data as originally generated
- Create the data you wish to see in the world:
  - Open, non-proprietary formats: e.g. .csv
  - Informative variable names
  - Recode missing values to NA
  - Informative filenames

- Record all steps used to process data and store intermediate data files if computationally intensive (easier to rerun parts of a data analysis pipeline)

- Separate data manipulation from data analysis
Data Management

- Prepare README with codebook of all variables
- Periodic backups (or Dropbox, Google Drive, etc.)
- Sanity checks: summary statistics after data manipulation
Code Management

- Number scripts/notebooks based on execution order: → e.g. 01-clean-data.py, 02-recode-variables.py
- Write an explanatory note at the start of each script/notebook
- Use version control (git)

- Rules of thumb for modular code:
  - Any task you run more than once should be a function (with a meaningful name!)
  - Functions should not be more than 20 lines long
  - Separate functions from execution
Organized Collaboration

- Create a README file with an overview of the project: title, brief description, contact information, structure of folder

- Shared to-do list with tasks and deadlines

- Choose one person as corresponding author

- Split code into multiple scripts/notebooks to avoid simultaneous edits

- ShareLaTeX or Overleaf to collaborate in writing of manuscript
Folder Structure

- Logical and consistent folder structure:
  - `src` for all scripts
  - `data` for raw data
  - `temp` for temporary data files
  - `output` or `results` for final data files and tables
  - `figures` or `plots` for figures produced by scripts
  - `manuscript` for text of paper
  - `docs` for any additional documentation
- Docker Image solves dependency hell problem
  - Binary image in which all the software has already been installed, configured and tested → also data files can be included
  - It’s like a VM but more lightweight
- Use Python Notebooks, Github and Binder
- Binder creates docker image from your git repository

Turn a GitHub repo into a collection of interactive notebooks

- Who has experience with python?
Next week 18th of April

- 10:15 Tutorial: Python, Docker, Binder
- Bring your laptop, ask questions

- 14:15 Lecture „Data and Measurements“
Any further questions?

Have a good start in the new semester!
See you next week