Social Data Science

(Department of Economics)
Faculty of Social Sciences
University of Copenhagen

Summer 2017

Lectures and classes:
Andreas Bjerre-Nielsen
David Dreyer Lassen
Snorre Ralund

Welcome!

always bring computer!

https://abjer.github.io/sds/
+ Absalon homepage

Today

1. Who are we? Who are you?
2. New course: Why and (so) What?
3. Logistics and Plumbing
   Python, Absalon vs. Github, groups, assignments??, exam project, course evaluation, Q&As
4. Course culture and ethics
5. Reading list and Lecture plan
6. Groups and details
7. Computer stuff

Who are we?

- We are:
  Andreas: PhD econ, postdoc @ SODAS
  David: Professor econ, Director of SODAS
  Snorre: Sociologist, PhD student @ SODAS

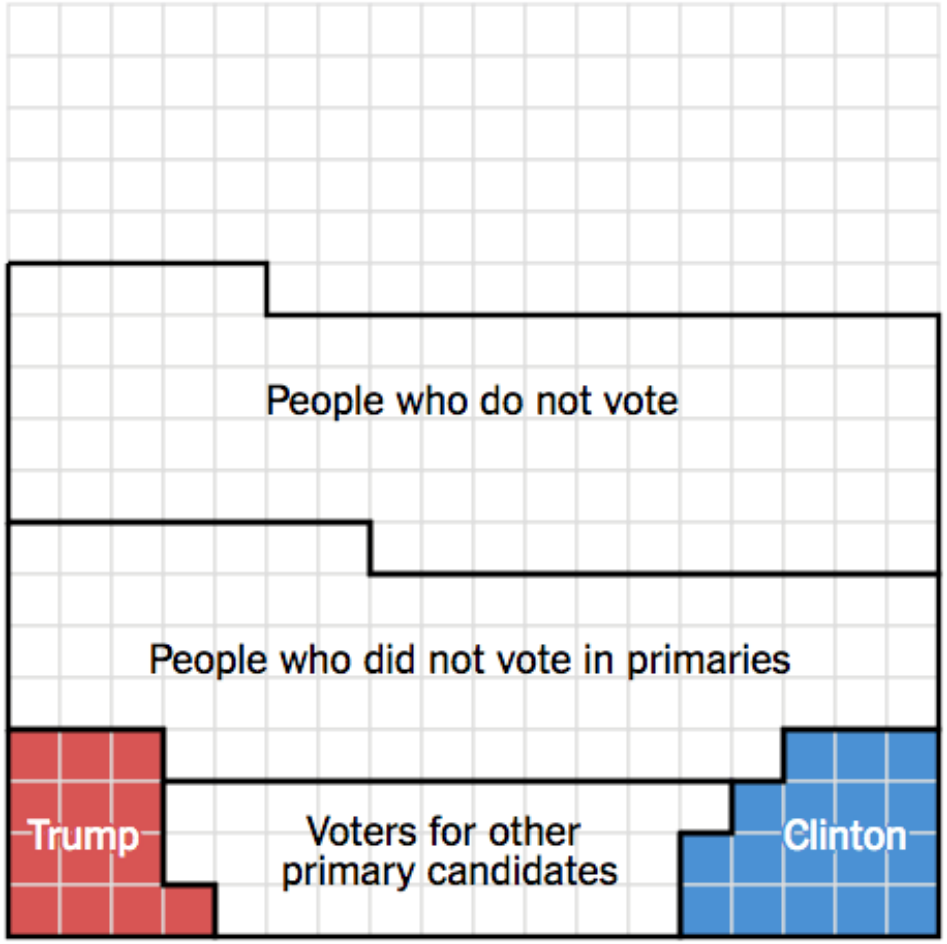
- All part of sodas.ku.dk
  Copenhagen Centre for Social Data Science

Who are you?

- 7 Q survey NOW!

https://daviddreyerlassen.typeform.com/to/gyshU0

Sometimes the best data is just about counting

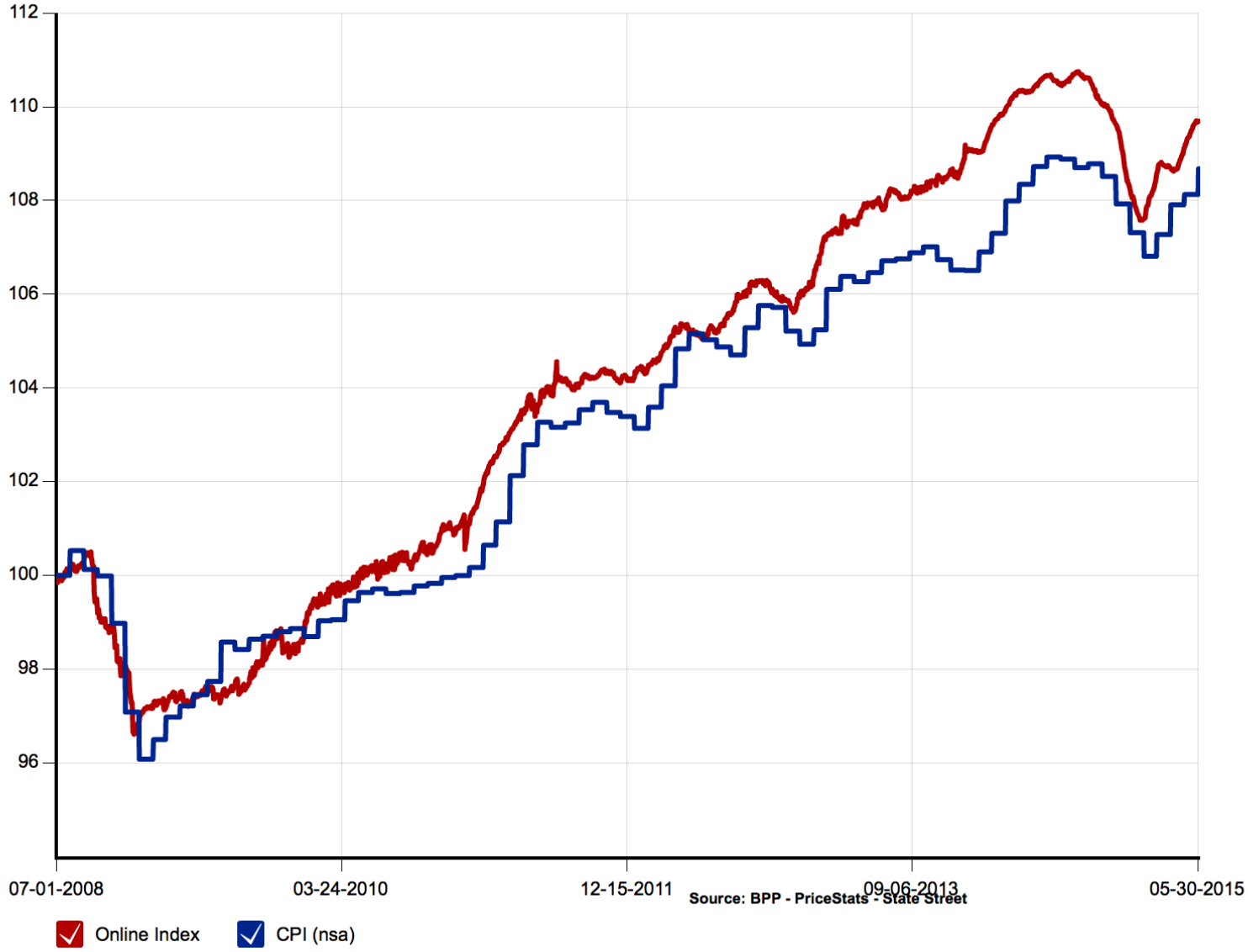


NY Times August 1,
http://www.nytimes.com/interactive/2016/08/01/us/elections/nine-percent-of-america-selected-trump-and-clinton.html

One key graph of the past 20 years' econ discourse

Figure 6-10
Share of Total U.S. Income Earned by Top 1 Percent, 1913–2010

Note: Total income includes wages and salaries (including bonuses and stock-option exercises), pensions, profits, farm income, dividends, interest, and rental income.
Source: Piketty and Saez (2003, 2010); authors provided an estimate for 2010 based on partial returns.



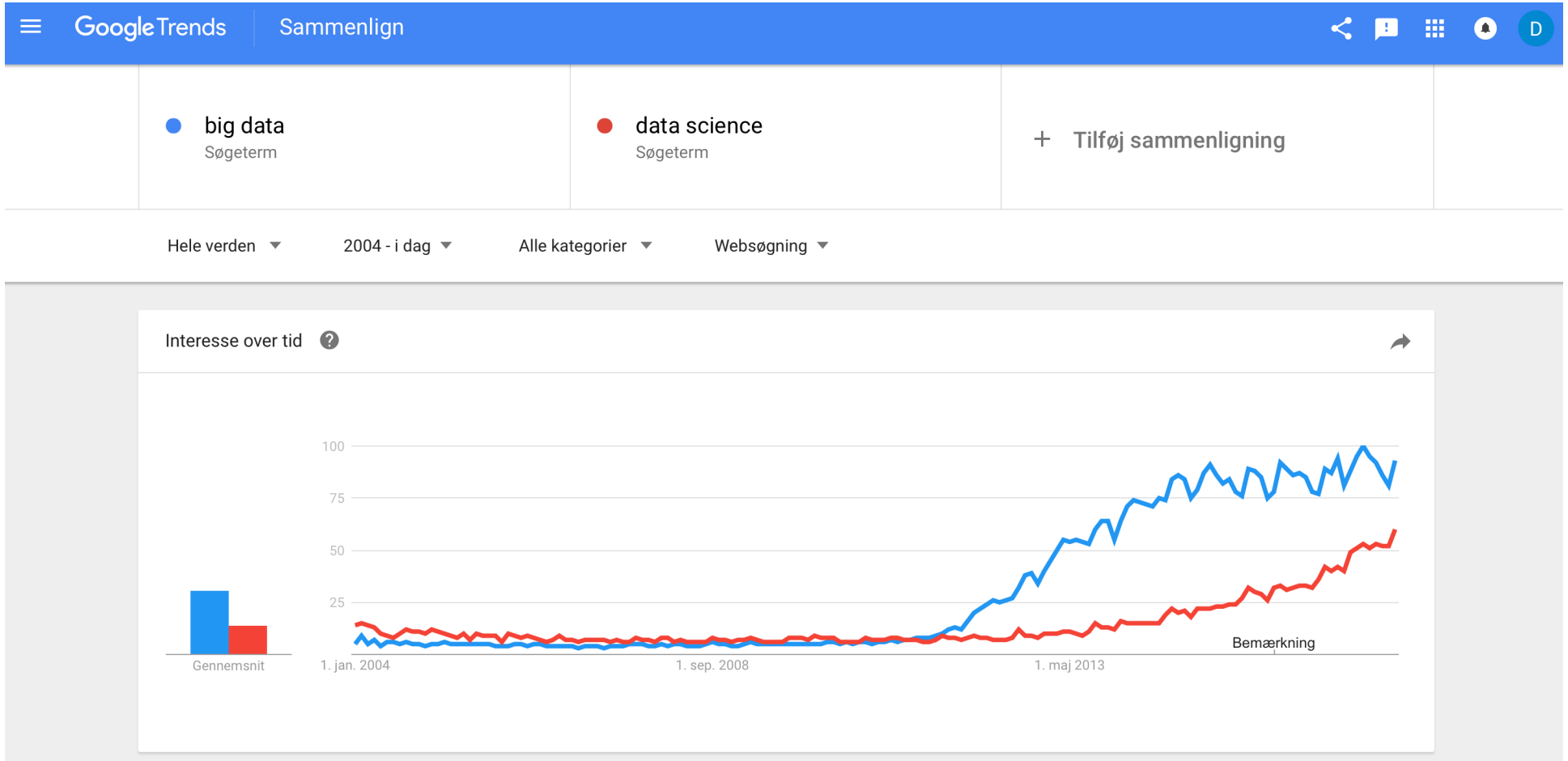
US Inflation

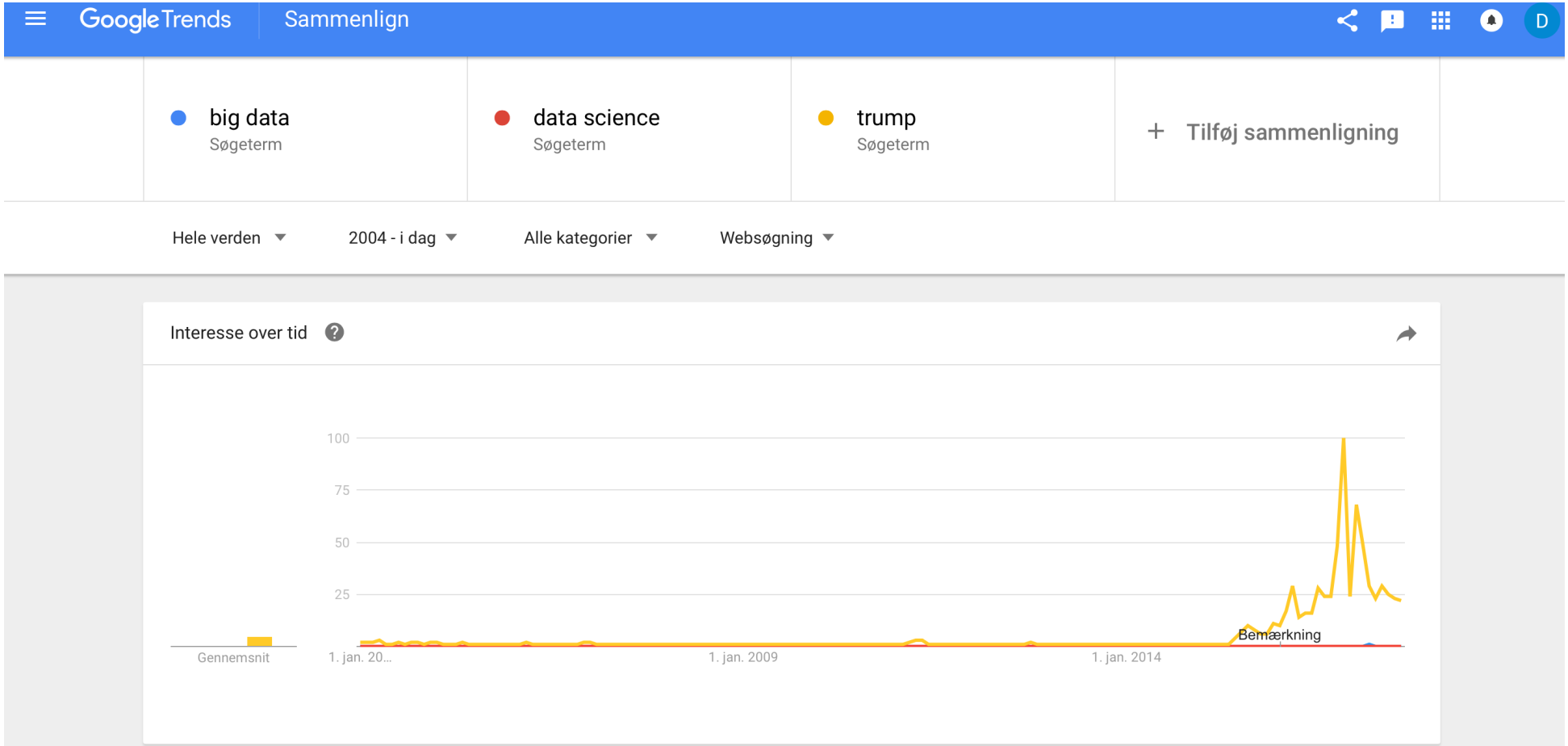
The Billion Prices Index: http://bpp.mit.edu


BIG DATA
IN
ACTION
FOR
DEVELOPMENT
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Latin America & the Caribbean
Opportunities for All
WORLD BANK GROUP
Macroeconomics & Fiscal Management
SECONDMUSE

New course I

- Background: Why Social Data Science
  - Big Data / Deep Data / New Data (Lazer and Radford, 2017)
  - Social Fabric / Taking Data Science Back


Google Trends
Sammenlign
D
big data
Søgeterm
data science
Søgeterm
+ Tilføj sammenligning
Hele verden
2004 - i dag
Alle kategorier
Websøgning
Interesse over tid
100
75
50
25
Gennemsnit
1. jan. 2004
1. sep. 2008
1. maj 2013
Bemærkning


Google Trends
Sammenlign
D
big data
Søgeterm
data science
Søgeterm
trump
Søgeterm
+ Tilføj sammenligning
Hele verden
2004 - i dag
Alle kategorier
Websøgning
Interesse over tid
100
75
50
25
Bemærkning
Gennemsnit
1. jan. 20...
1. jan. 2009
1. jan. 2014

What does ‘big data’ really mean?

- Originally: outside the scope of trad software processing
- focus on
  - Volume (size: no. of obs, Gigabytes)
  - Variety/complexity (incl. text, pictures, sound etc)
  - Velocity (often high frequency)
  - Veracity (‘honest signals’, behavior)

New course I

- Background: Why Social Data Science?
  - Big Data / Deep Data / New Data (Lazer and Radford, 2017)
  - Social Fabric / Copenhagen Network Study / Taking Data Science Back
- Economics: Not Econometrics, not standard Methods
- Social science methods: data collection, data construction
  - Sociology, political science, anthropology, psychology
  - Why important: research/substantive decisions taken along the way

The Construction of Data

1. Object(s) of interest
2. Data collection: feasibility (legal, ethics, (programming) skills, cooperation, time), costs
3. Data cleaning: what are objects of interest, what are outliers and errors (perspective: Latour, Pandora's box)
4. Construction of variables of interest, sometime probabilistic
5. Validation
6. Analysis

New course II

- Internet/digital data allows for more/new/realtime data: consumer prices, Uber, Facebook. Often requires scraping data.
- New methods allow for better extracting meaning from text (Text as Data, e.g. Facebook) and images
- Goals: ability to construct new data aimed at answering old and new social science questions. Make you informed consumers of Data Science literature
- Challenge: Big (social science) data not the product of scientific design, but scraps from admin (business, government) and life itself (e.g. mobile phones) - sometimes hard to get, sometimes hard to make meaning of

New course III

- Danish register data: admin data, full population, 1980-. Unique in the world, but: often very little data on actual behaviour, basically no data on social setting (network)

- Less or worse data: more theoretical assumptions (DK networks paper)

CONTAGIOUS POLITICAL CONCERNS: IDENTIFYING UNEMPLOYMENT SHOCK INFORMATION TRANSMISSION USING THE DANISH POPULATION NETWORK *

JAMES E. ALT †
AMALIE JENSEN ‡
HORACIO A. LARREGUY §
DAVID D. LASSEN ¶
JOHN MARSHALL ‖

JUNE 2017

It is widely believed that social pressure influences voters. However, there is little solid evidence that information transmitted through networks affects voter beliefs, policy preferences, and behavior. We investigate this function of networks with respect to unemployment shocks in post-financial crisis Denmark, where we link panel surveys to rich administrative data covering the entire population. Mapping each respondent's educational, familial, and vocational ties, we find that unemployment shocks afflicting "friends of friends"—individuals that a voter does not interact with directly—increase a voters' self-assessed risk of becoming unemployed, perception of the national unemployment rate, support for unemployment insurance, and probability of voting for left-wing political parties. Voters' own unemployment concerns and political preferences respond primarily to unemployment shocks afflicting second-degree connections in similar industries, whereas voters update about national aggregates from all shocks equally. This implies that political preferences driven by information transmitted through weak ties principally reflect self-interested—rather than sociotropic—motives.

New course III

- Danish register data: admin data, full population, 1980-. Unique in the world, but: often very little data on actual behaviour, basically no data on social setting (network)
- Less or worse data: more theoretical assumptions (DK networks paper)
- Prerequisites: Interest, willingness to program, will refer to regression models occasionally
- New course - hard to know time table

What we don’t cover

- Social science theory (not much, anyway)
- Standard statistical methods
- Social Data Science vs. Computational Social Science
- Networks
- Lots and lots of advanced material

Some topics

We will present a social science view on data science methods needed for **collecting** and **analyzing real-world data**. Focus points: **generating new data** (collecting, scraping, working with APIs), **data manipulation tools** (transforming, cleaning), **visualization tools** (visualizing raw data and model results), **reproducability tools** (git, github), an introduction to statistical techniques for predicting and classification, known as **statistical learning / machine learning (unsupervised / supervised)**

Meta and non-meta: What is data, types of data & types of questions, ethics, privacy, costs and benefits of data driven research / big data

Where to - and who else?

- Use insights from SDS in other courses / theses / workplace to generate new data for standard analysis
    - Recent theses: Friendships and group formation, GDP forecasting, predictive policing, machine learning approaches to predictive finance
- More advanced courses in statistical learning, machine learning, data science: Computer science at KU, DTU. SODAS advanced course on machine learning and networks in Spring 2018.
- Several large DK corporations (Danske Bank, Mærsk) upgrading significantly on Data Science; key focus area for DST. Obviously, Facebook, Google etc.

Logistics and Plumbing I

- We meet every day
- Typically, but now always: lectures in the morning, lab/exercises in the afternoon
- Always bring computer - Python!
- Absalon vs. Github

Logistics and Plumbing II

- Groups - we have allocated you, more shortly
- Assignments to help you through the material
- Week three: Group based exam project (see upcoming Github post)
- Course evaluation - formal and informal
- Discussion forum - Absalon

Course culture and ethics

- Philosophy: Open source, everyone contributes
- Help each other: within groups, across groups
  - Discussion forum
- But don’t free ride :-) Only fun if y’all pitch in
- Share, but don’t copy (really, don’t)

Course culture and ethics

- Ethics of data collection: will cover this at some length separately
- So far: don't be an (unduly) burden

AVISEN DK

Synes godt om

Folketinget er fredag blevet ramt af et hacker-angreb.

Det bekræfter Finn Tørngren Sørensen, presseansvarlig i Folketinget, over for Avisen. dk.

Siden fredag formiddag har man fået beskeden "Denne webside er ikke tilgængelig", hvis man har forsøgt at komme ind på Folketingets hjemmeside, ft.dk.

- Det er rigtigt, at der er lukket for den eksterne adgang til Folketingets hjemmeside. Vi er under et såkaldt 'Denial of service"-angreb, og det har vi været siden klokken 10 i formiddags, siger Finn Tørngren Sørensen til Avisen.dk og fortsætter:

- Det fungerer på den måde, at vi får så mange opkald til vores hjemmeside, at systemet bliver overbelastet. Derfor har vi måttet lukke ned for adgangen.

Folketinget har endnu ikke noget overblik over, hvem der står bag hacker-angrebet, eller hvornår hjemmesiden kan komme op at køre igen.

Reading list / Lecture plan

- Reading list at Github
  - New and fast moving topic - brand new textbook:
    Big Data and Social Science: A Practical Guide to Methods and Tools (BDSS) - good for Python
    Alternative: Kosuke Imai's **Quantitative Social Science** - good for R-users, good alternative
    Tons of bad and really bad books out there
  - Chapters (at Absalon), links to papers, blogs (UCPH domain)
- Required vs. inspiration vs. background
  - What to actually read?

Syllabus

A tentative syllabus, subject to change, is below.

Date	Time	Keywords	Slides/Exercises
Aug 7	9-12	Introduction to SDS	
Aug 7	13-15	Introduction to Python	An introduction to Python (notebook)
Aug 8	9-12	Data Visualization	Modern Python Plotting (notebook)
Aug 8	13-15	Data Manipulation (I)	Data structuring (notebook)
Aug 9	9-12	Data Manipulation (II)	
Aug 9	13-15	Exercises + Brainstorming	Ex 1: Hollywood's gender divide
Aug 10	9-12	Generating New Data	
Aug 10	13-15	Exercises + Brainstorming	
Aug 11	9-12	Reproducible Research	
Aug 11	13-15	Big Data in Social Science	
Aug 14	9-12	Causation & Prediction	
Aug 14	13-15	Exercise + Brainstorming	
Aug 15	9-12	Statistical Learning	
Aug 15	13-15	Exercise + Brainstorming	
Aug 16	9-12	Text as Data	
Aug 17	9-12	Privacy	
Aug 18	9-12	Wrap Up	

———— End of course ————

Contact points

- For most questions, try
  - your group
  - Other groups / Absalon (or github?) discussion forum
  - in-class
- In rare cases: email
- Don’t call us (and we won’t call you)

Groups

- Absalon has randomised you into groups of 4
- In Absalon: Go to People - Groups and use Search to find yourself.
- Right now: Find your group
- If you are not four in your group (if people didn't show up), come tell us. If <whatever>, come see us.