Social Data Science

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

Summer 2019

Lectures and classes:
Andreas Bjerre-Nielsen
Snorre Ralund

Guest appearance: David Dreyer Lassen
Teaching assistants: Anne, Asger, Edith, Jakob, Kristian, Kristoffer, Lykke

Welcome!

always bring computer!

https://abjer.github.io/sds2019/
+ Absalon homepage

Today

1. Who are we? Who are you?
2. (Relatively) 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, assistant professor of Econ & SDS @ SODAS
  David: Professor econ, Director of SODAS
  Snorre: M.Sc. sociology, PhD student @ SODAS

  Anne: polisci student, research assistant @ SODAS
  Asger: statistics student, soon-to-be PhD @ SODAS
  Edith: PhD student, econ
  Jakob: BA Econ, employed @ Datamaga
  Kristian: econ student, research assistant @ CEBI
  Kristoffer: M.Sc. Econ, PhD student @ SODAS
  Lykke: econ student, research assistant @ SODAS

- What is sodas.ku.dk: Copenhagen Centre for Social Data Science

Who are you?

8 Q survey NOW!

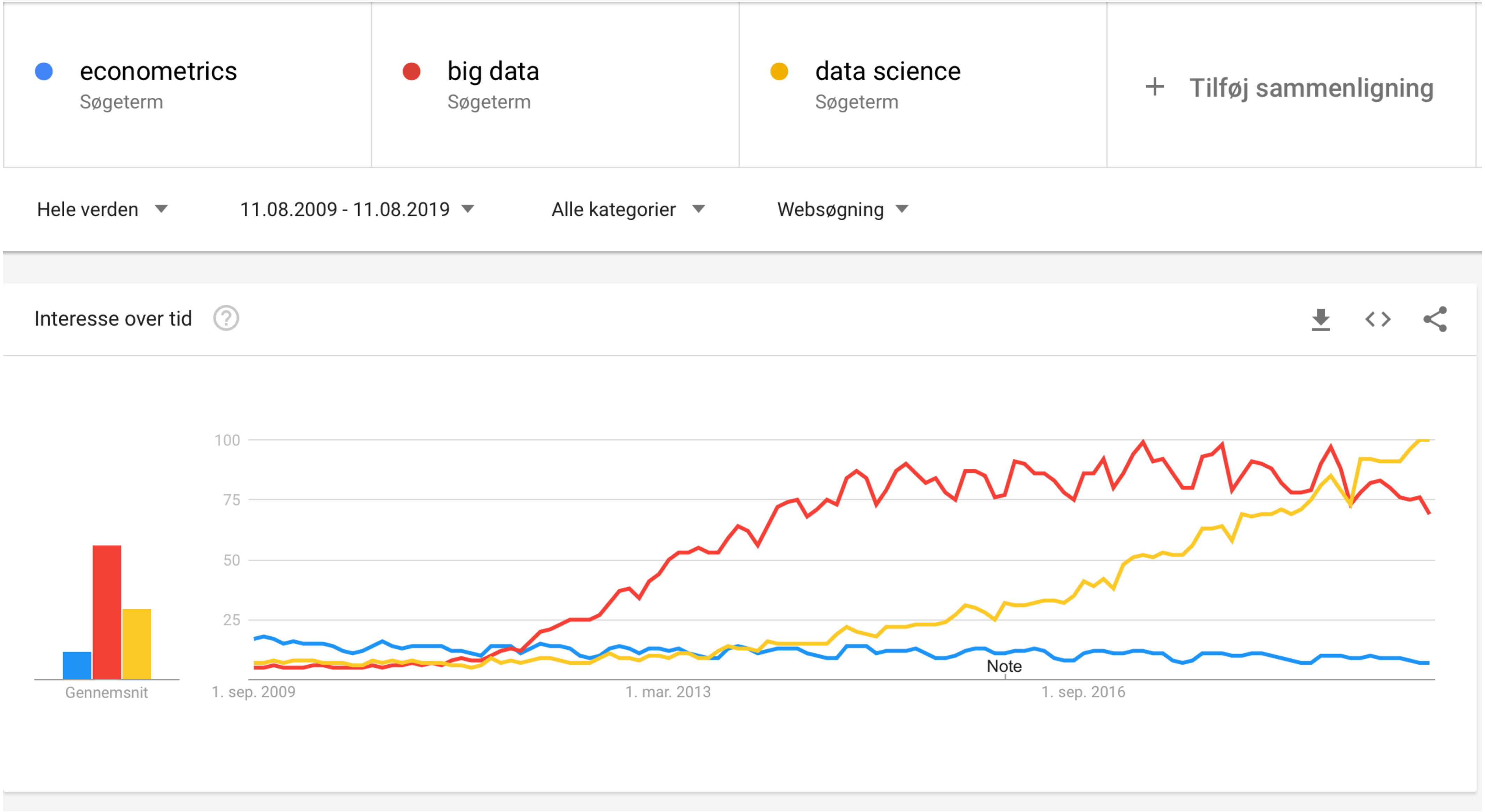
https://daviddreyerlassen.typeform.com/to/NcguZk

(estimated to to complete: 41 seconds)

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



● econometrics
Søgeterm

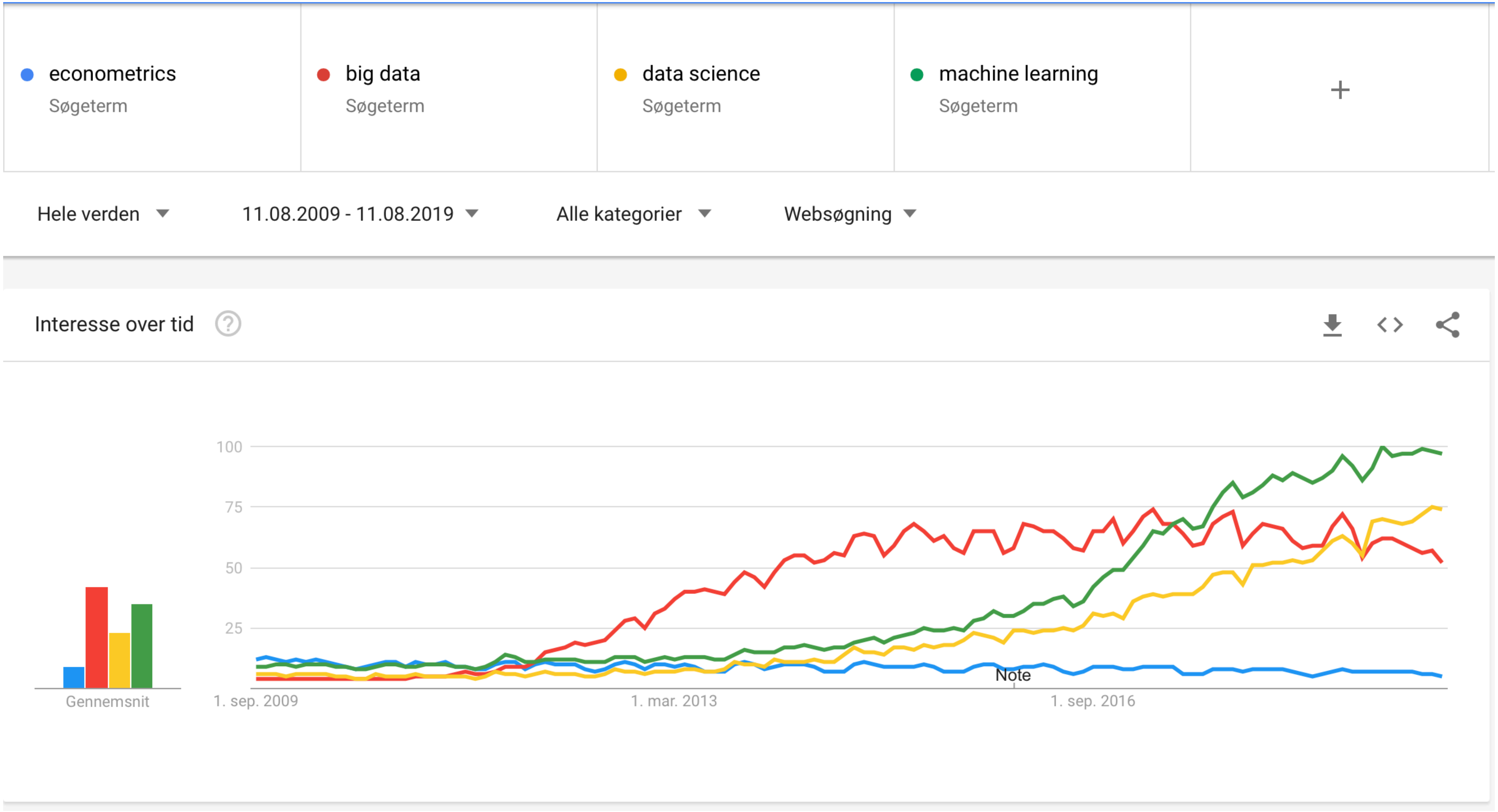
● big data
Søgeterm

● data science
Søgeterm

+ Tilføj sammenligning

Hele verden ▾ 11.08.2009 - 11.08.2019 ▾ Alle kategorier ▾ Websøgning ▾

Interesse over tid



● econometrics
Søgeterm

● big data
Søgeterm

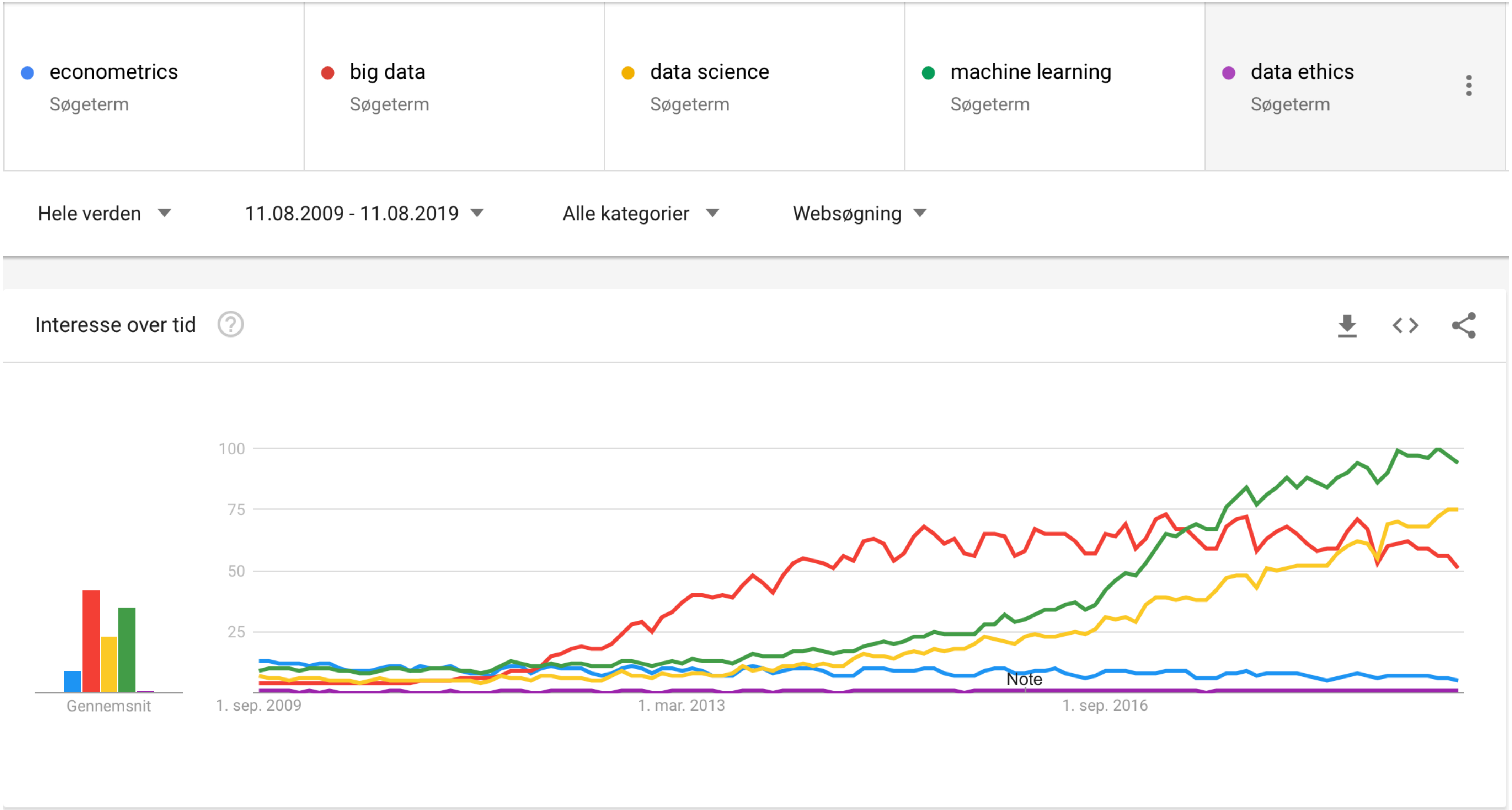
● data science
Søgeterm

● machine learning
Søgeterm

+

Hele verden ▾ 11.08.2009 - 11.08.2019 ▾ Alle kategorier ▾ Websøgning ▾

Interesse over tid



● econometrics
Søgeterm

● big data
Søgeterm

● data science
Søgeterm

● machine learning
Søgeterm

● data ethics
Søgeterm

Hele verden ▾ 11.08.2009 - 11.08.2019 ▾ Alle kategorier ▾ Websøgning ▾

Interesse over tid

What does ‘big data’ really mean?

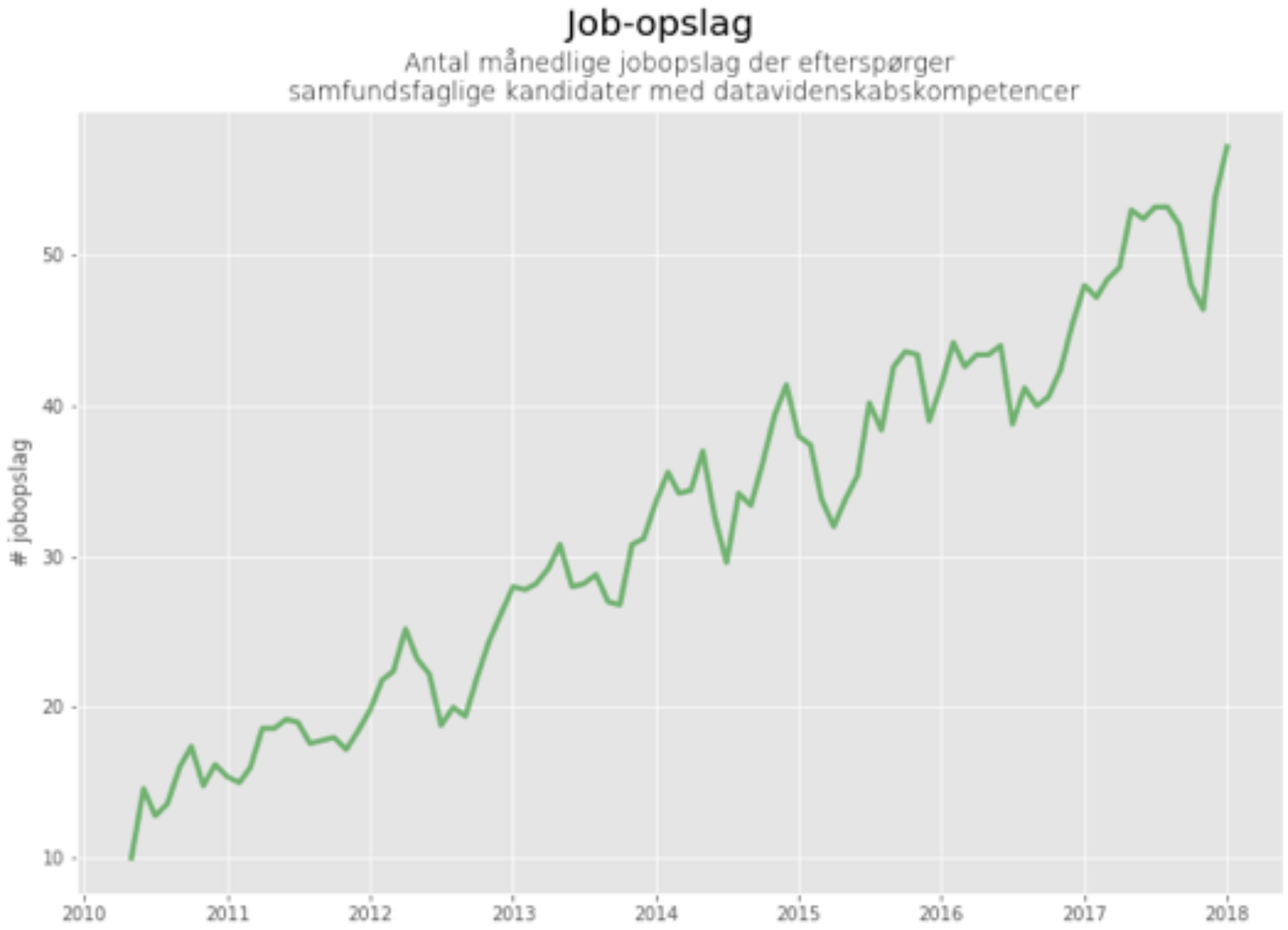
- Originally: outside the scope of trad software processing
- focus on
  - Volume (size: no. of obs, Gigabytes)
  - Variety/complexity (incl. text, pictures, sound etc)
  - Velocity (often high frequency)
  - Veracity (‘honest signals’, behavior)

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

New course II

- Important for
  - Research - new measures, new questions, checks on Big Tech and private/public sectors
  - Private sector - lots of new data, but what to do with them?
  - Public sector - lots of new data, more efficient and/or equitable public sector?



Job ads on Danish labor market combining some version of social science and some version of data skills. 1/3 public sector, 2/3 private sector

Data: Scraping Jobindex, 2.9 mio job ads 2007-18. Method: word2vec (data driven similarity of latent constructs). Ask Snorre.

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

4. Construction of variables of interest, sometime probabilistic

5. Validation

6. Analysis

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
4. Construction of variables of interest, sometime probabilistic
5. Validation
6. Analysis

Note: In some Social Data Science theses, 🟩 takes approx 75% of time and space, maybe even more

New course III

- 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 (Social) 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 IV

- 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, little market data

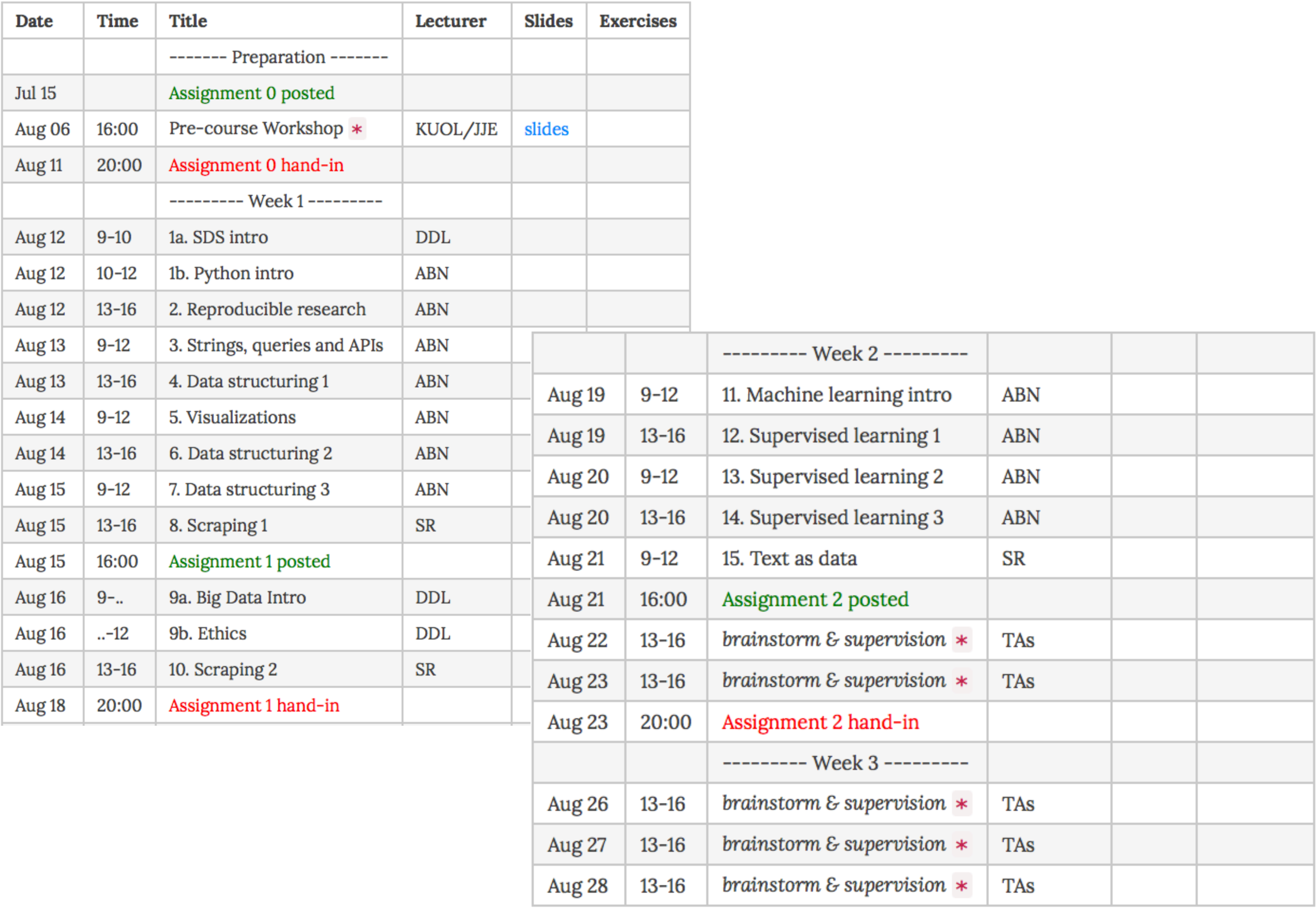
- Less or worse data: more theoretical assumptions

- Availability of data inevitably dictates what questions we ask - risk looking for the keys under the lamplight

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), **reproducibility 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



Date	Time	Title	Lecturer	Slides	Exercises
		------- Preparation -------			
Jul 15		Assignment 0 posted			
Aug 06	16:00	Pre-course Workshop ∗	KUOL/JJE	slides	
Aug 11	20:00	Assignment 0 hand-in			
		--------- Week 1 ---------			
Aug 12	9-10	1a. SDS intro	DDL		
Aug 12	10-12	1b. Python intro	ABN		
Aug 12	13-16	2. Reproducible research	ABN		
Aug 13	9-12	3. Strings, queries and APIs	ABN		
Aug 13	13-16	4. Data structuring 1	ABN		
Aug 14	9-12	5. Visualizations	ABN		
Aug 14	13-16	6. Data structuring 2	ABN		
Aug 15	9-12	7. Data structuring 3	ABN		
Aug 15	13-16	8. Scraping 1	SR		
Aug 15	16:00	Assignment 1 posted			
Aug 16	9-..	9a. Big Data Intro	DDL		
Aug 16	..-12	9b. Ethics	DDL		
Aug 16	13-16	10. Scraping 2	SR		
Aug 18	20:00	Assignment 1 hand-in			
		--------- Week 2 ---------			
Aug 19	9-12	11. Machine learning intro	ABN		
Aug 19	13-16	12. Supervised learning 1	ABN		
Aug 20	9-12	13. Supervised learning 2	ABN		
Aug 20	13-16	14. Supervised learning 3	ABN		
Aug 21	9-12	15. Text as data	SR		
Aug 21	16:00	Assignment 2 posted			
Aug 22	13-16	*brainstorm & supervision* ∗	TAs		
Aug 23	13-16	*brainstorm & supervision* ∗	TAs		
Aug 23	20:00	Assignment 2 hand-in			
		--------- Week 3 ---------			
Aug 26	13-16	*brainstorm & supervision* ∗	TAs		
Aug 27	13-16	*brainstorm & supervision* ∗	TAs		
Aug 28	13-16	*brainstorm & supervision* ∗	TAs		

∗ : optional participation

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

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 finance, freight supply, media usage, customer churn, firm bankruptcy etc.
- More advanced courses in statistical learning, machine learning, data science: Computer science at KU (DIKU), DTU Compute, possibly ITU.
- Topics in Social Data Science advanced course on machine learning, networks, text etc held in Spring 2018, 2019. To be repeated in modified, and enlarged, form in the Spring 2020.
- Several large DK corporations (Danske Bank, Mærsk, etc) upgrading significantly on Data Science; key focus area for DST, municipalities. Obviously, Facebook, Google etc. Also obviously, consulting

Logistics and Plumbing I

- We meet every day
- Two teaching sessions a day – one in the morning, one in the afternoon – mix of lectures and exercises
- 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

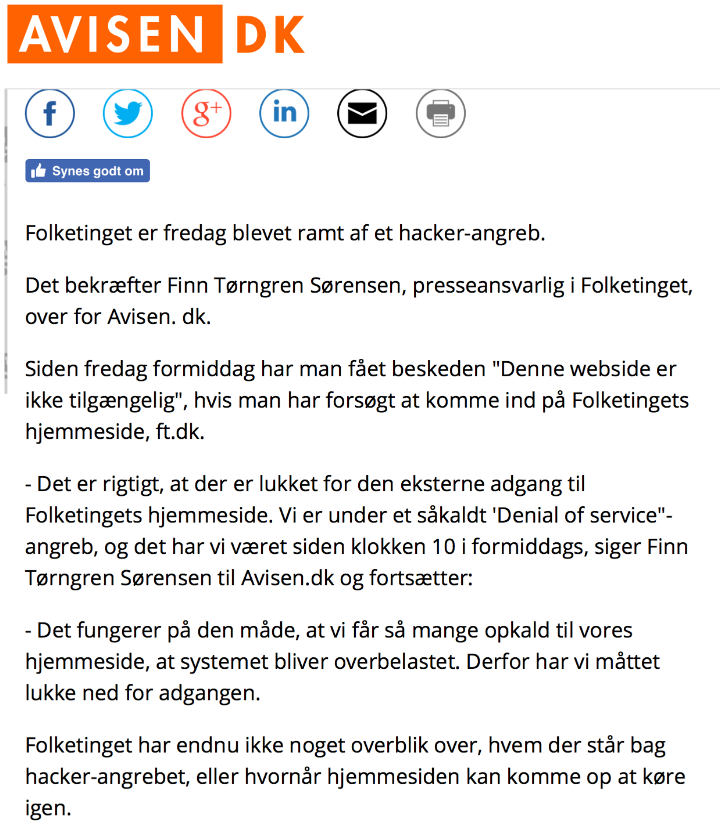
Course culture and ethics

- 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. Everyone in the group should contribute!
- Share, but don’t copy (really, don’t)

Course culture and ethics

- Ethics of data collection: will cover this at some length on Friday

- So far: don't be an (unduly) burden



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 excellent textbooks:
    Bit by Bit (Sagalnik) + Python Machine Learning 2nd ed (Raschka) + Python for Data Analysis (McKinney)
  - Some alternatives:
    Big Data and Social Science: A Practical Guide to Methods and Tools
    Kosuke Imai's Quantitative Social Science - good for R-users
    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?

Contact points

- For most questions, try
  - your group
  - other groups / Absalon (or github?) discussion forum
  - in-class consultations
  - Stack overflow and others
- 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 (a few people are not here this morning due to exams)
- If <whatever>, come see us.