Introduction to Social Data Science

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

Summer 2020

Lectures and classes:
Andreas Bjerre-Nielsen
Terne Thorn Jakobsen
Niklas Johansen
David Dreyer Lassen

Teaching assistants:
Emil, Esben, Kristoffer, Jakob, **Jonas**, Josefine, Lykke, Mads, Mathilde, Michael

Welcome!
Good to see you

always bring computer!

https://abjer.github.io/isds2020/
+ Absalon homepage
+ GitHub

Today

1. Who are we? Who are you?
2. (Relatively) new course: Why and (so) What?
3. Reading list and Lecture plan + not covered
4. Logistics
   Python, Absalon vs. Github, groups, assignments, exam project, course evaluation, Q&As
5. Course culture and ethics
6. Covid-19 issues

Who are we?

- We are:
  Andreas: PhD econ, assist. prof. of Econ & SDS @ SODAS
  Terne: PhD student, NLP/text as data @ SODAS
  David: Professor econ, Director of SODAS
  Nicklas: soon-to-be PhD student @ SODAS

  Lots of good Teaching Assistants (TAs)

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

Who are you?

10 Q survey NOW!

https://daviddreyerlassen.typeform.com/to/NBUL2IHn

(estimated time to complete: 41 seconds)

ISDS 1

- Background: Why Social Data Science
  - Big Data / Deep Data / New Data (Lazer and Radford, 2017): Dramatic increase in availability of digital or digitalized data
  - Taking Data Science Back - from computer science, engineering, physics

Google Trends

econometrics
Søgeterm

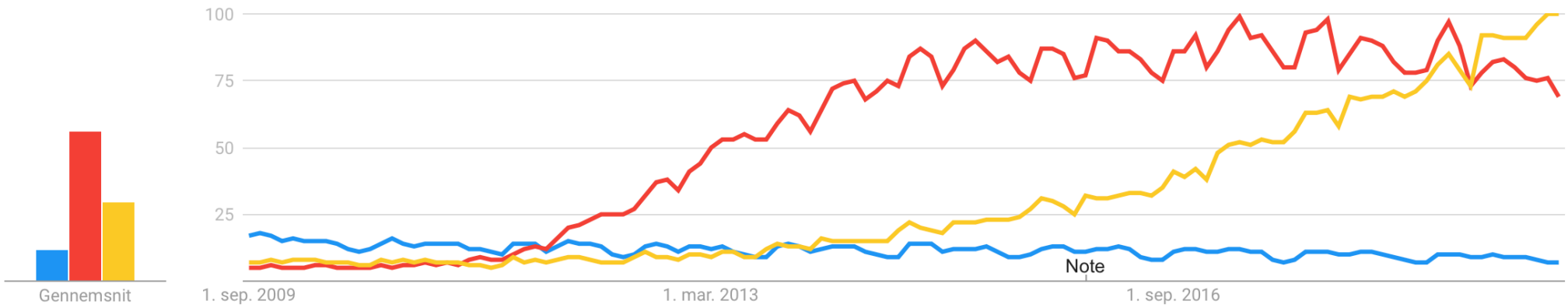
big data
Søgeterm

data science
Søgeterm

+ Tilføj sammenligning

Hele verden | 11.08.2009 - 11.08.2019 | Alle kategorier | Websøgning

Interesse over tid



econometrics
Søgeterm

big data
Søgeterm

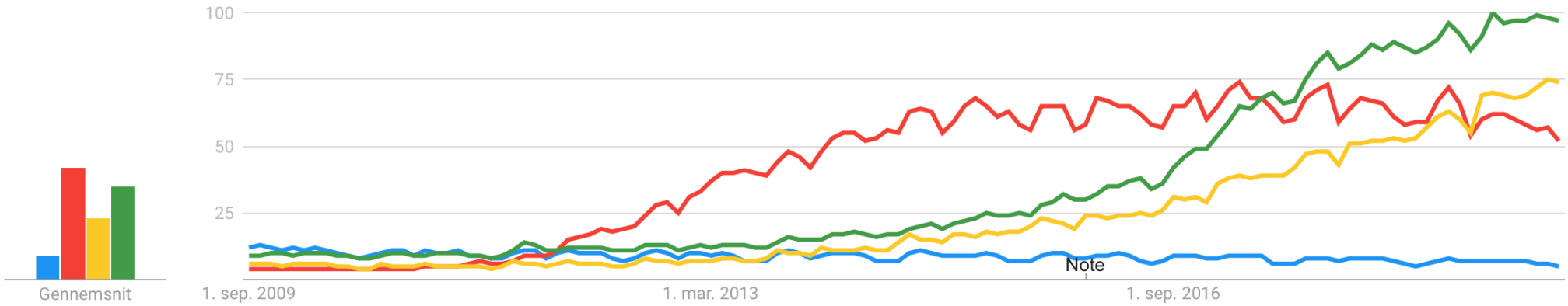
data science
Søgeterm

machine learning
Søgeterm

+

Hele verden | 11.08.2009 - 11.08.2019 | Alle kategorier | Websøgning

Interesse over tid



econometrics
Søgeterm

big data
Søgeterm

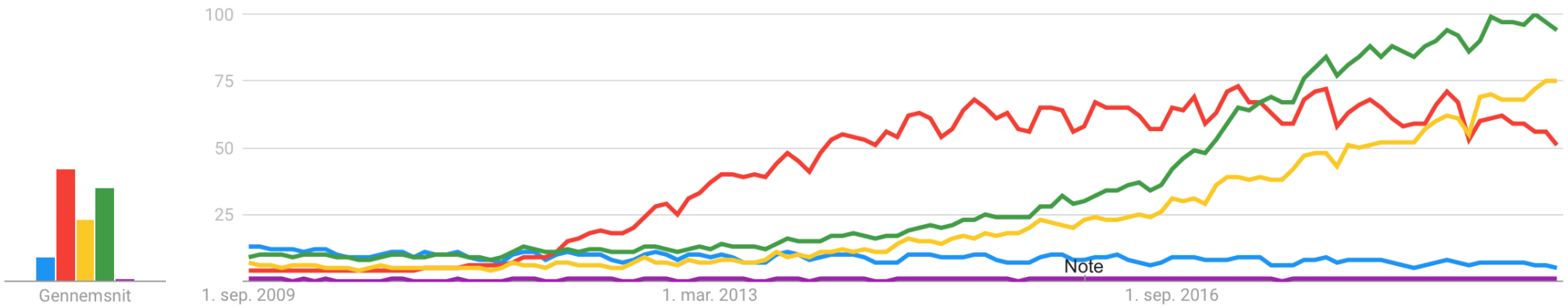
data science
Søgeterm

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?

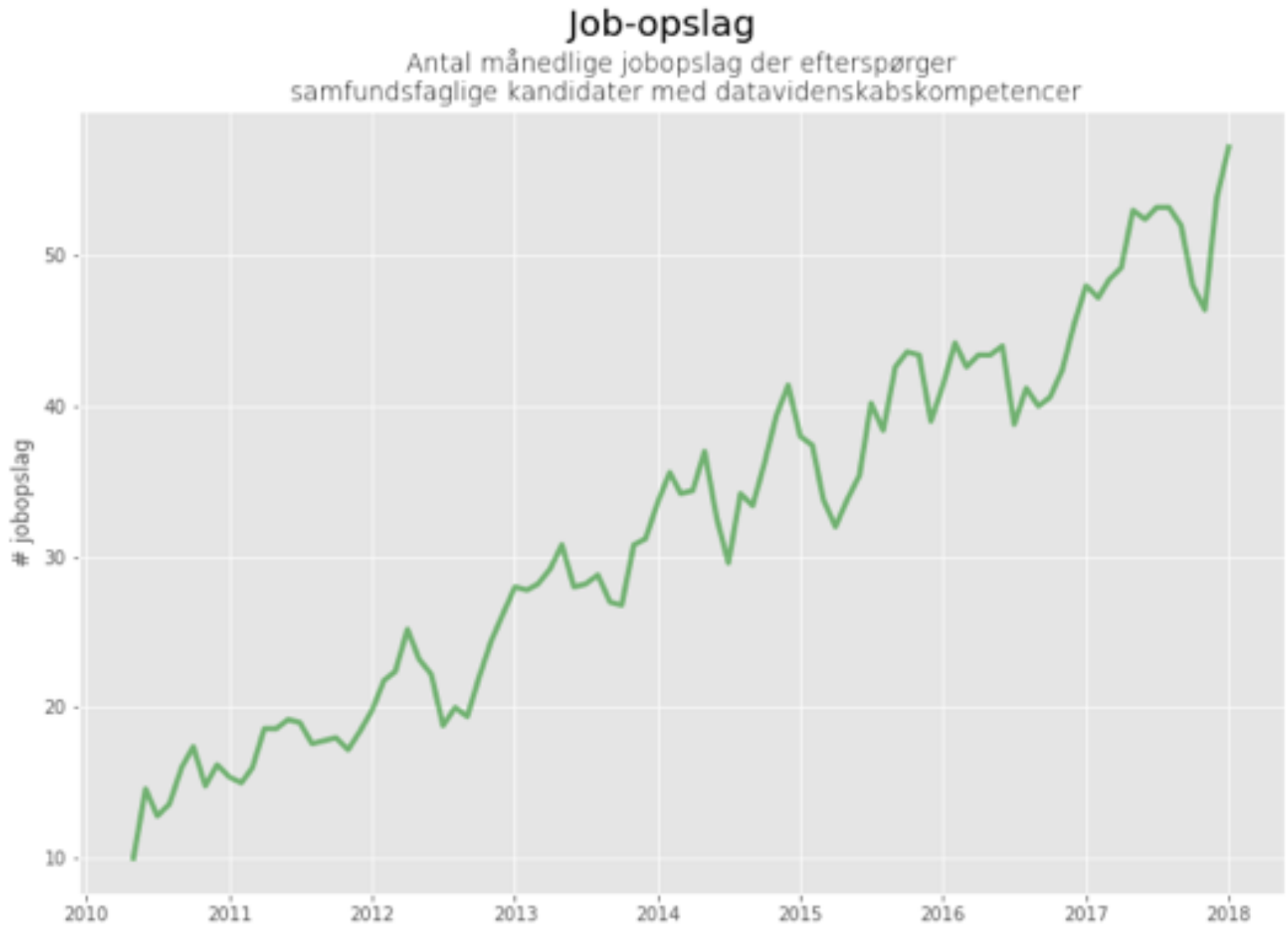
- Originally: outside the scope of traditional software processing
- focus on the 4 Vs
  - Volume (size: no. of obs, Gigabytes)
  - Variety/complexity (incl. text, pictures, sound etc)
  - Velocity (often high frequency: yearly vs. 5 min)
  - Veracity (‘honest signals’, behavior)

ISDS 1

- Background: Why Social Data Science
    - Big Data / Deep Data / New Data (Lazer and Radford, 2017): Dramatic increase in availability of digital or digilalized data
    - Taking Data Science Back - from computer science, engineering, physics
- 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 - “informed data cleaning”

ISDS 2

- 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 - talk to Terne about it)

The Construction of Data

1. Object(s) of interest
2. Data collection and structuring: 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 and structuring: 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

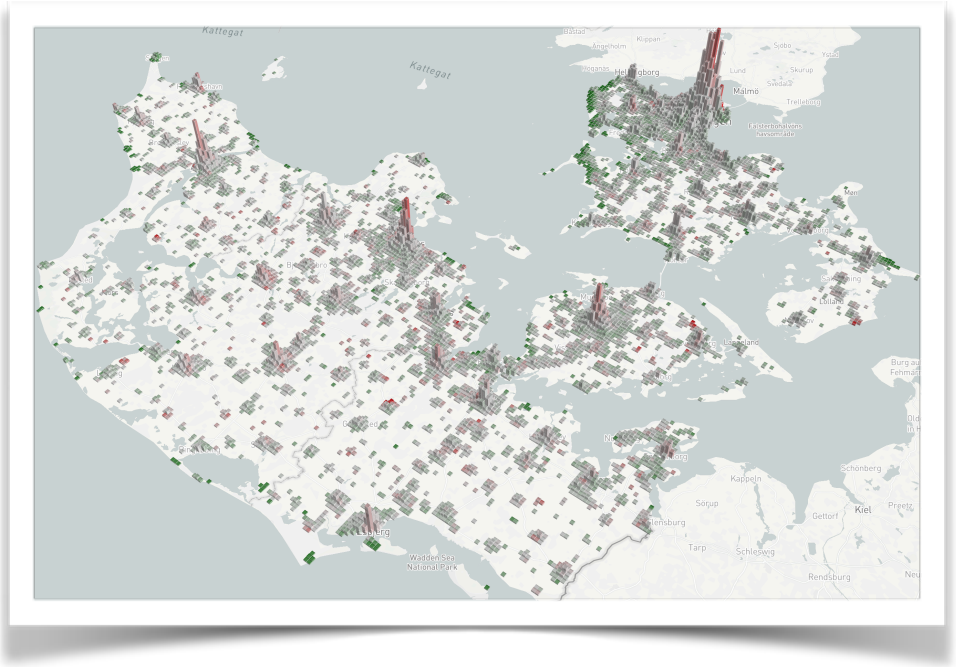
Note: In some Social Data Science theses, [redacted] takes approx 75% of time and space, maybe even more

ISDS 3

- Internet/digital data allows for more/new/realtime data: consumer prices, Uber, Facebook. Often requires scraping data, typically in forms not developed for analysis/research
- 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.

Example with Covid-19 data

- People leave traces everywhere
- Example: SODAS and DTU Compute work based on aggregated data from Facebook's "Data for Good"-platform
- Red = fewer than usual
  Green = more than usual



From: covid19.computer.dtu.dk

FACEBOOK Data for Good

We use data to address some of the world's greatest humanitarian issues.

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	Teachers	Material
		------- Preparation -------		
Jul 05		Assignment 0 posted		nb
Aug 04	15-18	Assignment 0 online workshop *	JSRP	pdf
Aug 07	12:00	Assignment 0 hand-in		
		--------- Week 1 ---------		
Aug 10	9-10	1a. Course welcome	DDL	
	10-11	1b. Python intro	ABN	
	11-12	1c. Meet group and intro to git	TAs	
	14-17	2. Strings, queries and APIs	ABN & TAs	nb
Aug 11	9-12	3. Visualizations	ABN & TAs	nb
	14-17	4. Data structuring 1	ABN & TAs	
Aug 12	9-12	5. Data structuring 2	ABN & TAs	
	14-17	6. Scraping 1	NJ & TAs	
	23:59	Assignment 1 posted	--	
Aug 13	9-12	7. Scraping 2	NJ & TAs	
	14-17	8. Scraping 3	NJ & TAs	
Aug 14	9-11	9a. Big Data and Ethics	DDL	
	11-12	9b. Assignment 1 workshop	TAs	
	14-16	Supervision 1 *	TAs	
Aug 14	9-11	9a. Big Data and Ethics	DDL	
	11-12	9b. Assignment 1 workshop	TAs	
	14-16	Supervision 1 *	TAs	
	23:59	Assignment 1 hand-in		
		--------- Week 2 ---------		
Aug 17	9-12	11. Machine learning intro	ABN & TAs	
Aug 17	14-17	12. Supervised learning 1	ABN & TAs	
Aug 18	9-12	13. Supervised learning 2	ABN & TAs	
Aug 18	14-17	14. Supervised learning 3	ABN & TAs	
Aug 18	20:00	Assignment 2 posted		
Aug 19	9-12	15. Text as data	TJ & TAs	
Aug 20	23:59	Assignment 2 hand-in		
Aug 21	9-11	*Supervision 2* *	TAs	
		--------- Week 3 ---------		
Aug 23	20:00	Exam project description due		
Aug 24	9-11	*Supervision 3* *	TAs	

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.
- Machine learning and Econometrics (Spring 2020): advanced course on machine learning and ties to econometrics. Will be offered at some point again.
- New M.Sc. in Social Data Science @ UCPH with interesting courses
- Several large DK corporations (Danske Bank, Mærsk, etc) upgrading significantly on Data Science; key focus area for DST, government at all levels. Obviously, Facebook, Google etc. Also obviously, consulting

Logistics I

- We “meet” every day - two online-only classes
- Typically 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 II

- Groups - some self-chosen, some allocated
- Assignments to help you through the material: everyone should work on these, don't be the one fetching the pizzas
- Week three: Group based exam project (see website 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)

Data collection ethics

- Ethics (and legalities) of data collection: will cover this at some length on Friday
- So far
  - don't be an (unduly) burden
  - Identify yourself (as students from UCPH)
  - Last year: "man in the middle" attack

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 excellent textbooks:
    Bit by Bit (Sagalnik) + Python Machine Learning 2nd ed (Raschka and Mirjalili) + Python for Data Analysis 2nd ed. (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 / TAs
  - stackoverflow and the www in general
- In rare cases: email
- Don’t call us (and we won’t call you)

Important! points

- Due to Covid-19, this is a special situation - we follow the rules set out by the Faculty of Social Sciences
- Observe the standard rules for hygiene
- Not everything will work as planned or intended - for us, for your group or for you - allow for adjustments, we will work it out
- Online group work has turned out surprisingly well, but it is not perfect. Be aware the group work may be harder or stranger than usual. Get in touch if problems.
- For offline students: If you are concerned about Covid-19-related issues during the course, get in touch and talk to your group about it