Introduction to Social Data Science

Department of Economics
Faculty of Social Sciences
University of Copenhagen

Summer 2024

Lectures and classes:
Andreas & Hjalte

Teaching assistants:
Johan, Jonathan & Mikkel

Welcome!
Good to see you

(always bring computer & headphones!)

https://isdsucph.github.io/isds2024/
+ Absalon homepage
+ GitHub repo

This class

1. Who are we? Who are you?

2. (Relatively) new course: Why and (so) What?

3. Course logistics:
   - plan, reading list, teach format + beyond the course
   - Python, Absalon vs. Github,
   - groups, assignments, exam project, course evaluation, Q&As

4. Course culture and ethics

5. Learning to code

6. Reproducibility tools: Git and markdown

Who are we?

- We are:
  - Andreas: Assoc. Prof. @ Econ Dept. & SODAS
  - Hjalte: PhD Stud. @ Econ Dept. & CEBI
    - Three amazing Teaching Assistants (TAs)
      - Johan, Jonathan, Mikkel

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

Who are you?

8 question survey NOW!

URL: bit.ly/4ftKur8

(estimated time to complete: < 1min)

Social data science:

What and why

ISDS 1

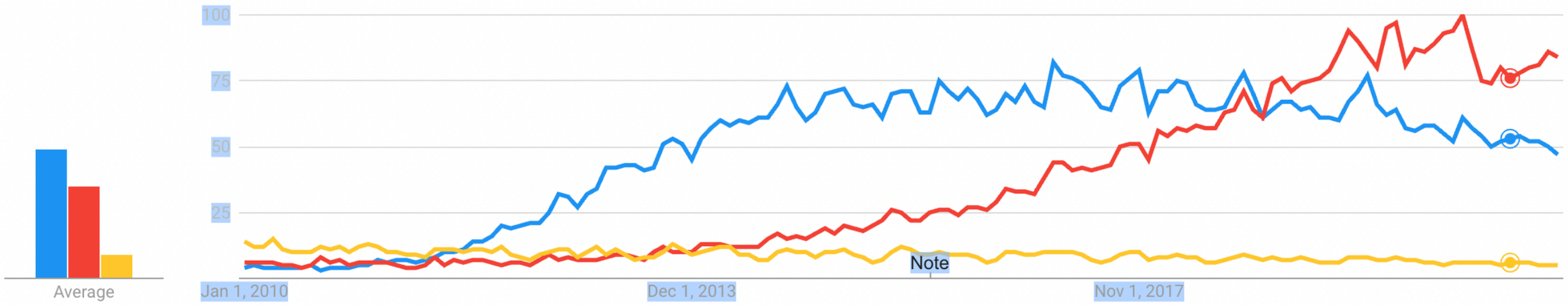
- Background: Why Social Data Science
    - Big Data / Deep Data / New Data (Lazer and Radford, 2017): Dramatic increase in availability of digital or digitalized data
    - Taking Data Science Back - from computer science, engineering, physics

Google Trends

big data	data science	econometrics	+ Add comparison
Search term	Search term	Search term	

Worldwide ▾ 1/1/10 - 8/1/21 ▾ All categories ▾ Web Search ▾

Interest over time

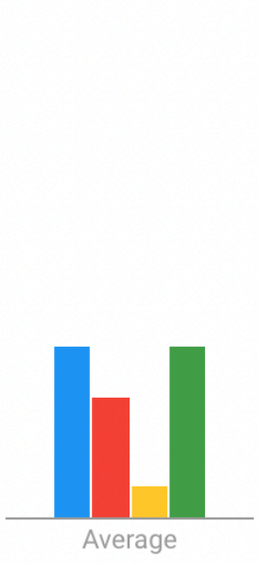


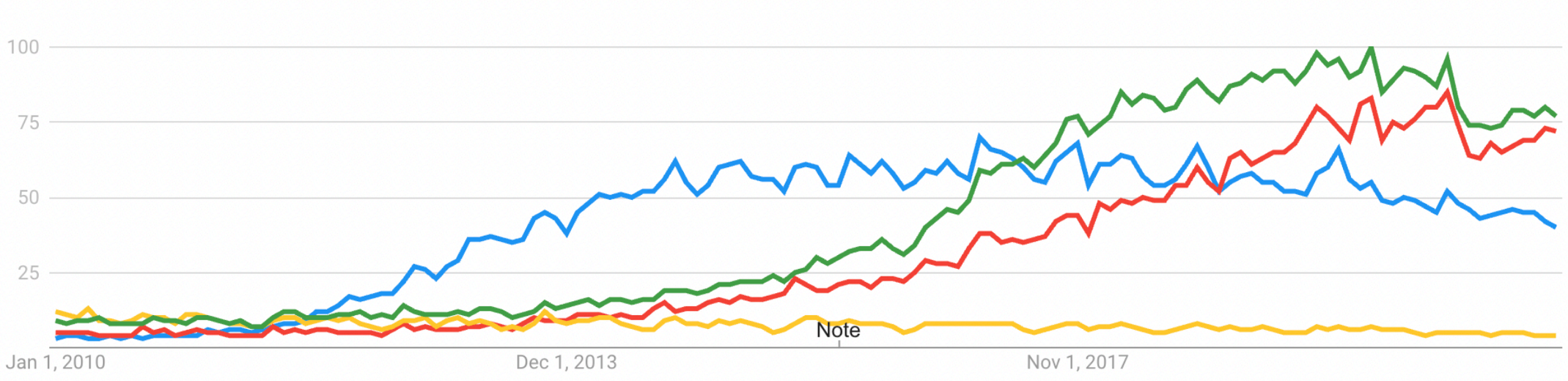
Google Trends

big data	data science	econometrics	machine learning	+
Search term	Search term	Search term	Search term	

Worldwide ▾ 1/1/10 - 8/1/21 ▾ All categories ▾ Web Search ▾

Interest over time



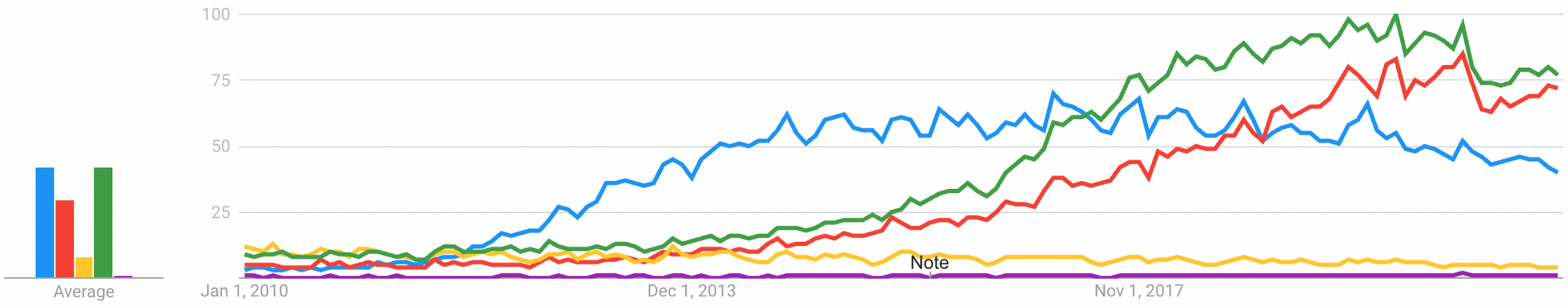


Google Trends

big data	data science	econometrics	machine learning	data ethics
Search term	Search term	Search term	Search term	Search term

Worldwide ▾ 1/1/10 - 8/1/21 ▾ All categories ▾ Web Search ▾

Interest over time



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 digitalized data
    - Taking Data Science Back - from computer science, engineering, physics
- 1) Social science methods: collection and structuring of human data
    - Anthropology, Economics, Political Science, Psychology, Sociology
    - Why important: research/substantive decisions taken along the way - “informed data cleaning”
- 2) New tools for predictive modelling

ISDS 2

- Important for
  - Research - new measures, new methods, new questions, checks on Big Tech and private/public sectors
  - Private sector - lots of new data, but what to do with them? E.g. algorithmic pricing, ad-tech?
  - 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 up to 75% of time and space

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

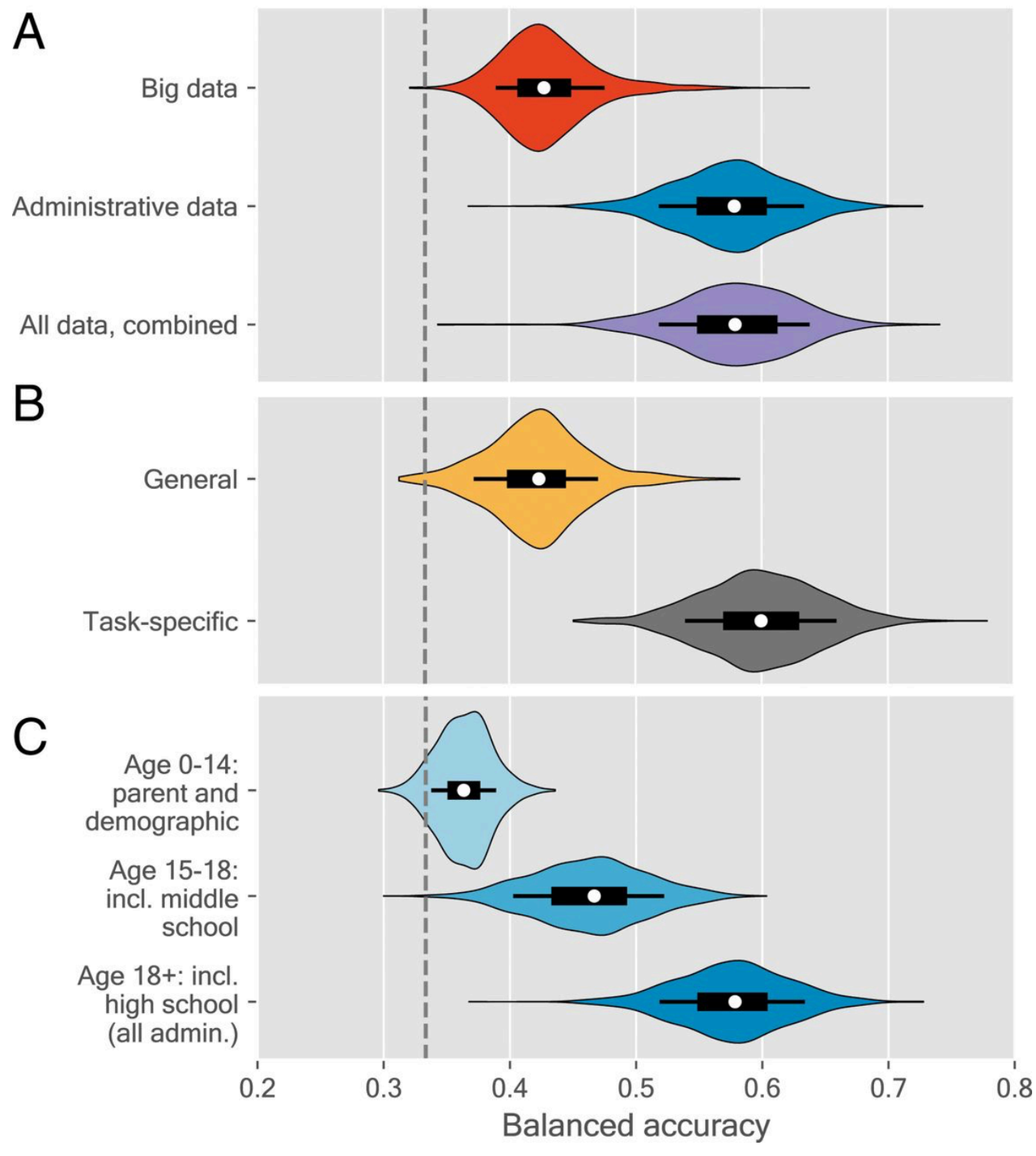
- Challenge: Big (social science) data not the product of scientific design, but scraps from admin (business, government) and life itself (e.g. mobile phones) - sometimes hard to get, sometimes hard to make meaning of.

Predicting academic success?

- People leave traces everywhere

- Example: this paper investigates whether using surveillance technology can enhance predictions of academic success

- Finding: they do not enhance already strong models



Measuring culture using Facebook data

- Facebook user data can be used to measure similar between users expressed interests and behaviours

- This paper leverages Facebook data to generate new measure of cultural similarity

FACEBOOK Data for Good

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

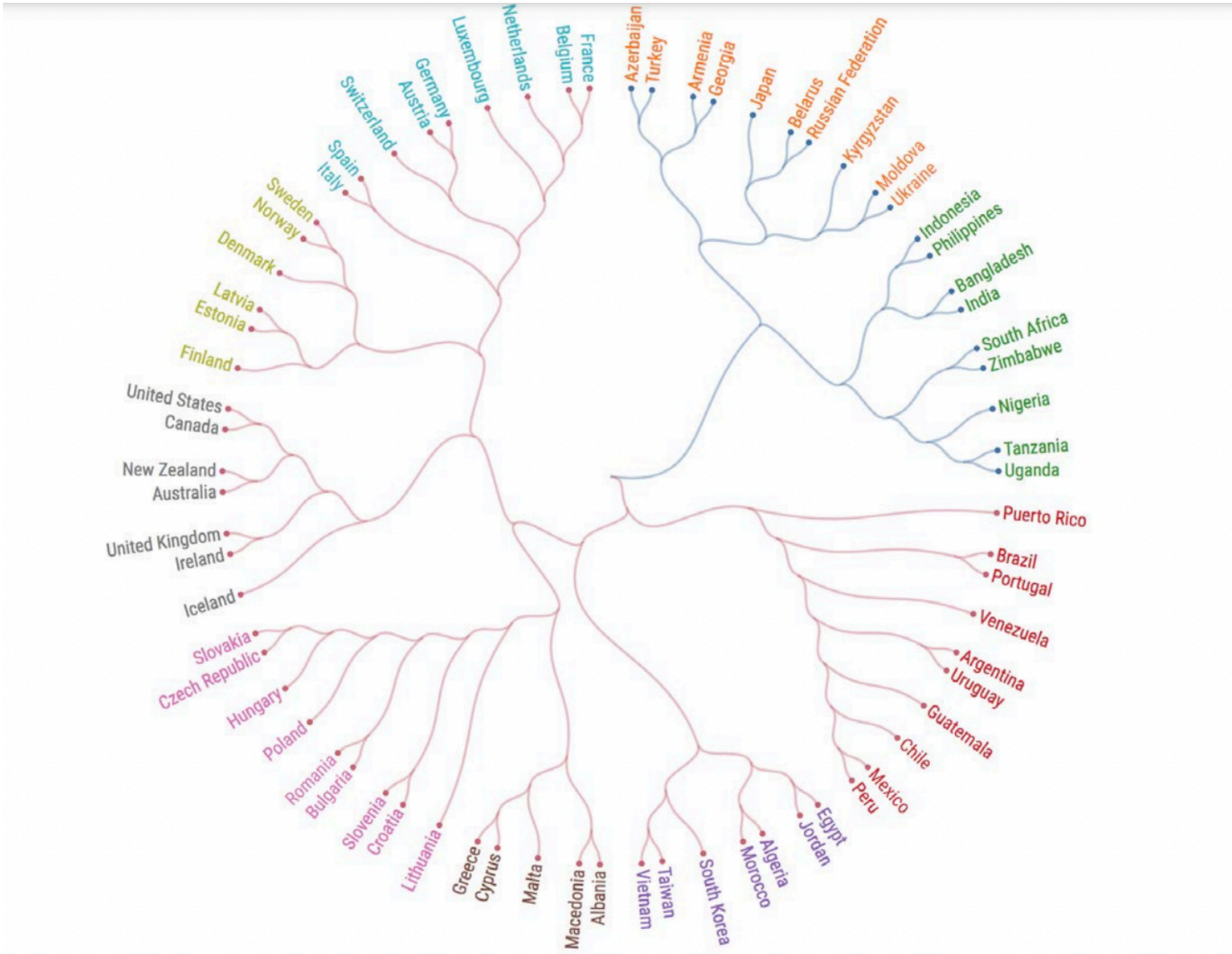


Figure 3: ***Bottom-up measurement of culture returns sensible clusters of countries.*** *Dendrogram is generated using cosine distance and the Ward linkage method (see Methods). Countries and territories with at least 300,000 persons, a Facebook penetration rate of more than 5%, and representation in both the Facebook and WVS data were selected. The color of a country's link represents its membership to a main cluster, while the color of its name represents its membership to a sub-cluster. Two countries of the same color are closer to each other than to a country of a different color.*

Some topics

We will present a social science view on data science methods needed for **collecting** and **analyzing real-world data**. Focus points tools:

- **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),

- 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

Course logistics

Reading list / Lecture plan

- Reading list at Github

  - New and fast moving topic - brand new excellent textbooks:
    - Bit by Bit (Sagalnik)
    - Python Machine Learning 3rd 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

  - Research papers, blogs

- Required vs. inspiration vs. background

  - What to actually read?

Logistics I

- We “meet” every day - some people online only
- 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 - GitHub issues

Plan this week

Aug 05	9-11	1a. Course welcome + intro to SDS	ABN	
	11-12	1b. Meet group	TAs	
	13-16	2. Data Structuring 1	ABN & TAs	
Aug 06	9-12	3. Data Structuring 2	ABN & TAs	
	12-13	Office Hour	ABN	
	13-16	4. Plotting	ABN & TAs	
	23:59	Assignment 1 posted	–	
Aug 07	9-12	5. Strings, Queries and APIs	ABN & TAs	
	13-16	6. Scraping 1	HFB & TAs	
Aug 08	9-12	7. Scraping 2	HFB & TAs	
	12:59	Fill in supervision sheet	–	
	13-15	TA help + Supervision *	TAs	
Aug 09	9-11	9. Data Ethics	ABN	
	13-14	Exam talk	HFB	
	14-17	TA help + Supervision *	TAs	
Aug 09	23:59	Assignment 1 hand-in	-	abs

Plan next week

Aug 12	9-12	8. Scraping 3	HFB & TAs	
	12-13	Office Hour	HFB	
Aug 13	9-12	10. ML Introduction	ABN & TAs	
	13-16	11. Regression and Regularization	ABN & TAs	
	23:59	Assignment 2 posted	–	
Aug 14	9-12	12. Model Selection and Cross-validation	ABN & TAs	
	13-16	13. Performance Metrics, Non-linear ML, and Perspectives	ABN & TAs	
Aug 15	9-12	14. Text as Data	HFB & TAs	
	13-15	TA help *	TAs	
	15-16	Office Hour	HFB	
Aug 16	9-12	Exam talk + TA help + Supervision	HFB & TAs	
	13-15	TA help + supervision *	TAs	
	23:59	Assignment 2 hand-in		abs

What we don't cover

- Social science theory (not much, anyway)
- Standard statistical methods / econometrics
- Social Data Science vs. Computational Social Science
- Lots and lots of advanced material (data structures, machine learning algorithms)

Where to - and who else?

- More knowledge
  - Degree M.Sc. in Social Data Science @ UCPH with exciting courses. Two electives are open:
    - (i) Data Governance: Law, Ethics and Politics
    - (ii) Advanced Social Data Science II (on text as data)
    - Machine learning and Econometrics: advanced course on machine learning and ties to econometrics.
    - More advanced courses in statistical learning, machine learning, data science: Computer science at KU (DIKU), DTU Compute, possibly ITU.
- 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.
- Professional
  - Several large DK corporations (Danske Bank, Mærsk, etc) upgrading significantly on Data Science; key focus area for Statistics Denmark, government at all levels. Obviously, Amazon, Facebook, Google etc. Also obviously, consulting

Course culture and ethics

Course culture and ethics

- Philosophy: Open source, everyone contributes
- Help each other: within groups, across groups
    - Discussion forum on Github
- 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 Monday
- So far
  - don’t be an (unduly) burden
  - Identify yourself (as students from UCPH)
  - First 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.

Learning to code:

Get ready to get frustrated

No free lunch..

- This is course is *not* easy
- Learning without supervision:
  - Data structuring as experimentation
  - Struggle with simple stuff
  - How I felt..

Some encouragement

- **Hadley Wickham:** *"The bad news is that when ever you learn a new skill you're going to suck. It's going to be frustrating. The good news is that is typical and happens to everyone and it is only temporary. You can't go from knowing nothing to becoming an expert without going through a period of great frustration and great suckiness."*

- **Kosuke Imai***: "One can learn data analysis only by doing, not by reading."*

Coding mindset

- Maintain healthy curiosity - how could we do things better?
- Practice and try as much as possible
- Type the code in yourself - then you see what is going on.

Coding advice

- Be careful: Think before you code - what you are trying to make it do?
- Be lazy: reuse code and write reusable code (functions)
- Make understandable: think about audience
    - Future you? May not recall this at all.
    - Group members or world? May need some background explanation/documentation.

Contact points

- Whenever you have a question you do as follows:
- 1: You ask other people in your group.
- 2: You ask the neighbouring groups.
- 3: You search on online (more advice will follow).
- 4: You raise an issue in our Github repo or you ask us

- In very rare cases: email
- Don’t call us (and we won’t call you)

Coding tutor?

Powered up learning with AI

This year: We allow AI tools as (exam) aids, e.g.

- ChatGPT
- GitHub Copilot (great with VS Code as code editor)

Many use cases: Debugging, Q&A, copy editing, brainstorming

- But: ChatGPT often suggests overly complex, inefficient, and bloated code.

- Econ take: It is first a substitute, then a complement for own skill. But without coding yourself, you will not get better – a dilemma.

.

Debugging with ChatGPT

- *My prompt: "Please debug*

```
print(1-A)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/var/folders/02/sx0gjqb54z16jm_hb9w7g81r0000gn/T/
ipykernel_58378/1515943728.py in <module>
----> 1 print(1-A)

NameError: name 'A' is not defined:"
```

- Example response:

"The error you're encountering, `NameError: name 'A' is not defined`, indicates that the variable `A` has not been defined before you tried to use it in the `print` statement. To resolve this issue, you need to define `A` before you use it.

Here's an example of how you can define `A` and then use it in the `print` statement:

```
A = 0.5  # Define the variable A with a value, for example, 0.5
print(1 - A)  # Now you can use A in the print statement:
```

Reproducibility tools

What is git?

- Enables you to work with several people on the same code or text, at the same time.
- Send/receive changes to/from shared folder
  - (folder is called repository).
  - Changes are explicit - not automatic

Why git?

- Enables you to work with several people on the same code or text, at the same time.
- Your closest collaborator is yourself 6 month ago - and you are not answering e-mails.
- Why not dropbox/Google Drive?
  - Detailed log of all changes: When, who, what
    - Easy to revert back to previous versions (remembers forever)
    - Clear attribution of work (who contributed what)
  - Some files are shared, some not (need to be explicit)

Without git

Coordinates_rawdata.csv	18/05/2015 19:07
coords_2015-09-09.csv	23/09/2015 17:18
coords_2015-09-09_modif.csv	05/11/2015 15:20
coords_2015-10-11_modif_YC.csv	17/11/2015 13:49
coords_2015-10-18_modif_YC.csv	18/11/2015 17:26
coords_2015-12-26_modif_YC.csv	28/12/2015 13:33
coords_2015-12-26_modif_YC_years.csv	30/03/2016 19:38
Pulido et al_SM1_Data.csv	20/10/2015 11:55
Pulido et al_SM1_Data_modif_YC_2015-12-26.csv	28/12/2015 13:30
qualitative_data.csv	04/07/2016 15:50
cleandata.xlsx	25/06/2015 01:14
cleandata_YC.xlsx	30/06/2015 16:22
COORDENADAS PACO_20-05-2016 CON REVIEWS.xlsx	20/05/2016 16:23
COORDENADAS PACO_20-05-2016 CON REVIEWS_FRS.xlsx	27/05/2016 19:41
COORDENADAS_paper195(Girella_elevata).xlsx	08/06/2016 13:09
coordenadas_raw_2016-06-08.xlsx	09/06/2016 15:53
coordenadas_raw_2016-06-08_old.xlsx	08/06/2016 16:00
coordenadas_raw_2016-06-21.xlsx	21/06/2016 16:12
coords_2015-09-09_modif.xlsx	05/11/2015 15:23
coords_2015-10-11_modif_YC.xlsx	17/11/2015 13:37
coords_2015-10-11_modif_YC_PACO.xlsx	17/11/2015 17:06
coords_2015-10-18_modif_YC.xlsx	18/11/2015 17:24
coords_2015-12-26_modif_YC.xlsx	30/03/2016 19:38
coords_2016-04-02.xlsx	06/04/2016 17:46
coords_2016-04-02_YC.xlsx	06/04/2016 18:03
coords_2016-04-08_YC.xlsx	11/04/2016 13:51
dataset_y_coords_09_09_15.xlsx	23/09/2015 17:18
Datos metaanalisis_18-04-2016.xlsx	19/04/2016 16:24
FINAL METAANALISYS_14-6-2016_WITH REVIEWS.xlsx	21/06/2016 16:15
FINAL METAANALISYS_16-6-2016_WITH REVIEWS.xlsx	21/06/2016 16:13
FINAL METAANALISYS_2016-04-27_WITH REVIEWS.xlsx	25/05/2016 18:05
FINAL METAANALISYS_2016-04-27_WITH REVIEWS_FRS.xlsx	27/05/2016 18:44
FINAL METAANALISYS_2016-04-29_EXCLUDING REVIEWS.xlsx	08/06/2016 13:06
FINAL VOTECOUNTING_1-7-2016.xlsx	04/07/2016 15:46
fitnessdata_2016-06-22.xlsx	22/06/2016 21:00
IFs for Bastien_19-3-2016_YC.xlsx	28/03/2016 19:26
Metaanalysis final_01-05-2015 with coordinates.xlsx	18/05/2015 19:20
Metaanalysis final_22-05-2015 coords.xlsx	24/06/2015 15:50
Metaanalysis final_25-06-2015.xlsx	30/06/2015 16:55
Metaanalysis y coords revisadas_06-08-2015_AH_JE.xlsx	23/09/2015 12:57
Pulido et al_SM1_Data_2016-05-27.xlsx	27/05/2016 18:48

With git

exclosure_damage_raw.csv	04/07/2016 21:21
exclosures_cover_raw.csv	04/07/2016 20:49
sitenames.csv	04/07/2016 20:42
sites_info_raw.csv	30/06/2016 20:03
species_info_raw.csv	05/07/2016 15:53

Markdown

- Markdown is an easy to use text editor
  - Alternative to latex - simpler
    - Almost WYSIWYG (what you see is what you get)
- Use
  - Typeset text in Jupyter Notebook / Lab
  - Also useful for making homepages (e.g. ISDS)

Markdown - font formatting

```
# Largest heading

## Second largest heading

### And so on
```

Largest heading

Second largest heading

And so on

- **Text in bold** -> **Text in bold**
- *Text in italics* -> *Text in italics*

Markdown - lists

```
- fruits
    - apples
        - macintosh
            - red delicious
    - pears
    - peaches
- vegetables
    - broccoli
    - chard
```

- fruits
  - apples
    - macintosh
      - red delicious
  - pears
  - peaches
- vegetables
  - broccoli
  - chard

The end

Next up: meet your groups