Introduction to Social Data Science

Department of Economics
Faculty of Social Sciences
University of Copenhagen

Summer 2023

Lectures and classes:
Felix Chopra

Teaching assistants:
Marcus, **Mischa**, Jonathan

Welcome!
Good to see you

(always bring computer & headphones!)

https://isdsucph.github.io/isds2023/
+ 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, evaluation, Q&As

4. Course culture and ethics

5. Learning to code

6. Reproducibility tools: Git and markdown

Who are we?

- Felix: Asst. Prof. @ Econ Dept. & CEBI
- Three amazing Teaching Assistants (TAs)
  - Marcus, Mischa, Jonathan

Who are you?

12 question survey NOW!

https://forms.gle/moVR2SXwN1b2vA4XA

(estimated time to complete: 68 seconds)

Social data science:

What and why

ISDS 1

- Background: Why Social Data Science?

  1. Big Data / Deep Data / New Data (Lazer and Radford, 2017): Dramatic increase in availability of digital or digitalized data

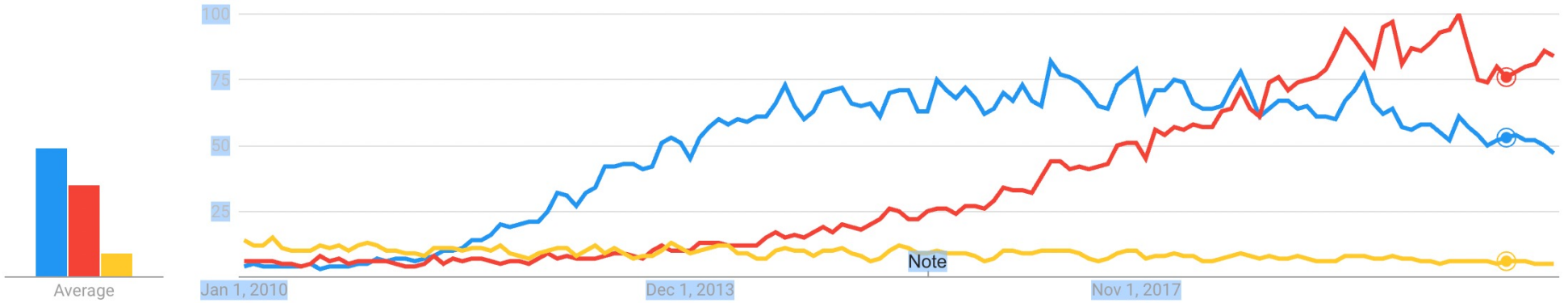
  2. Taking Data Science Back - from computer science, engineering, physics

Google Trends

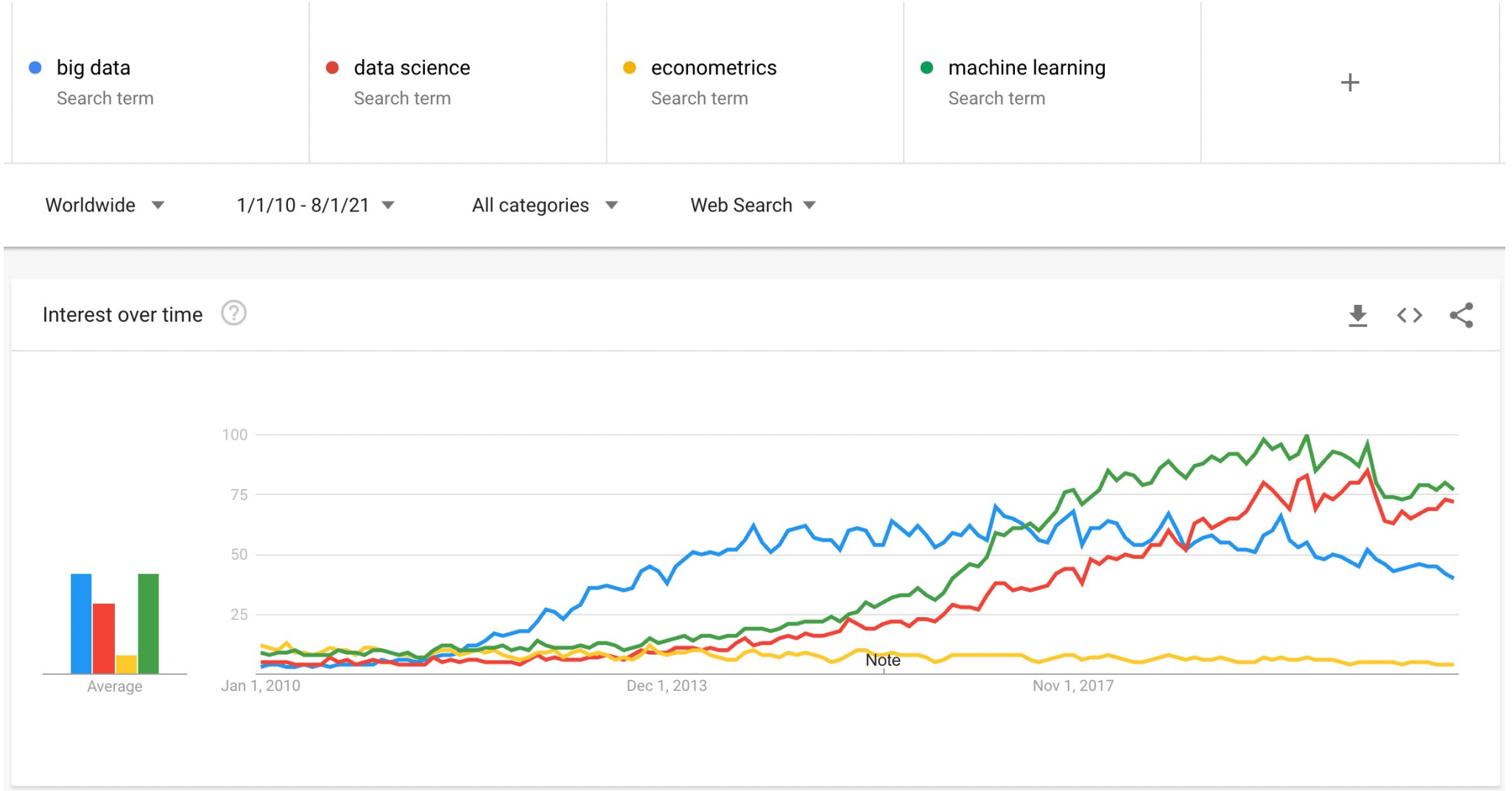
big data	data science	econometrics	+ Add comparison
Search term	Search term	Search term	

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

Interest over time



Google Trends

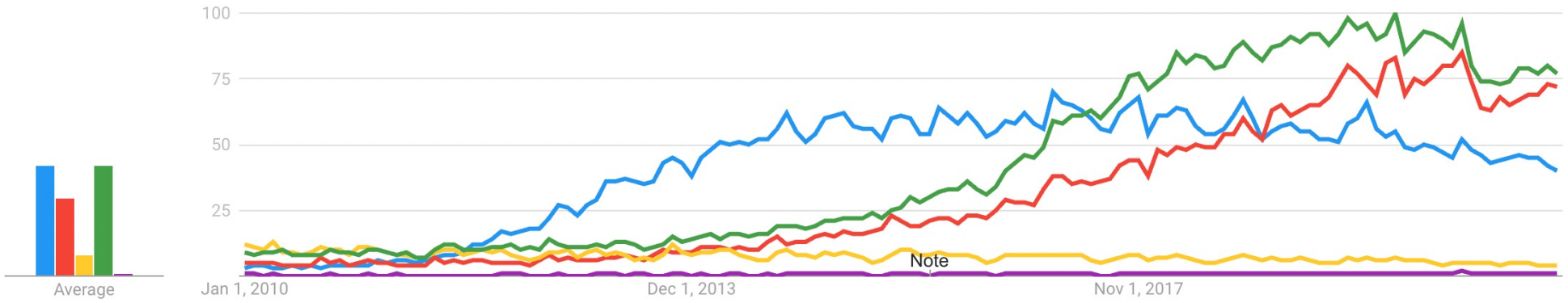


Google Trends

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: large N, 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

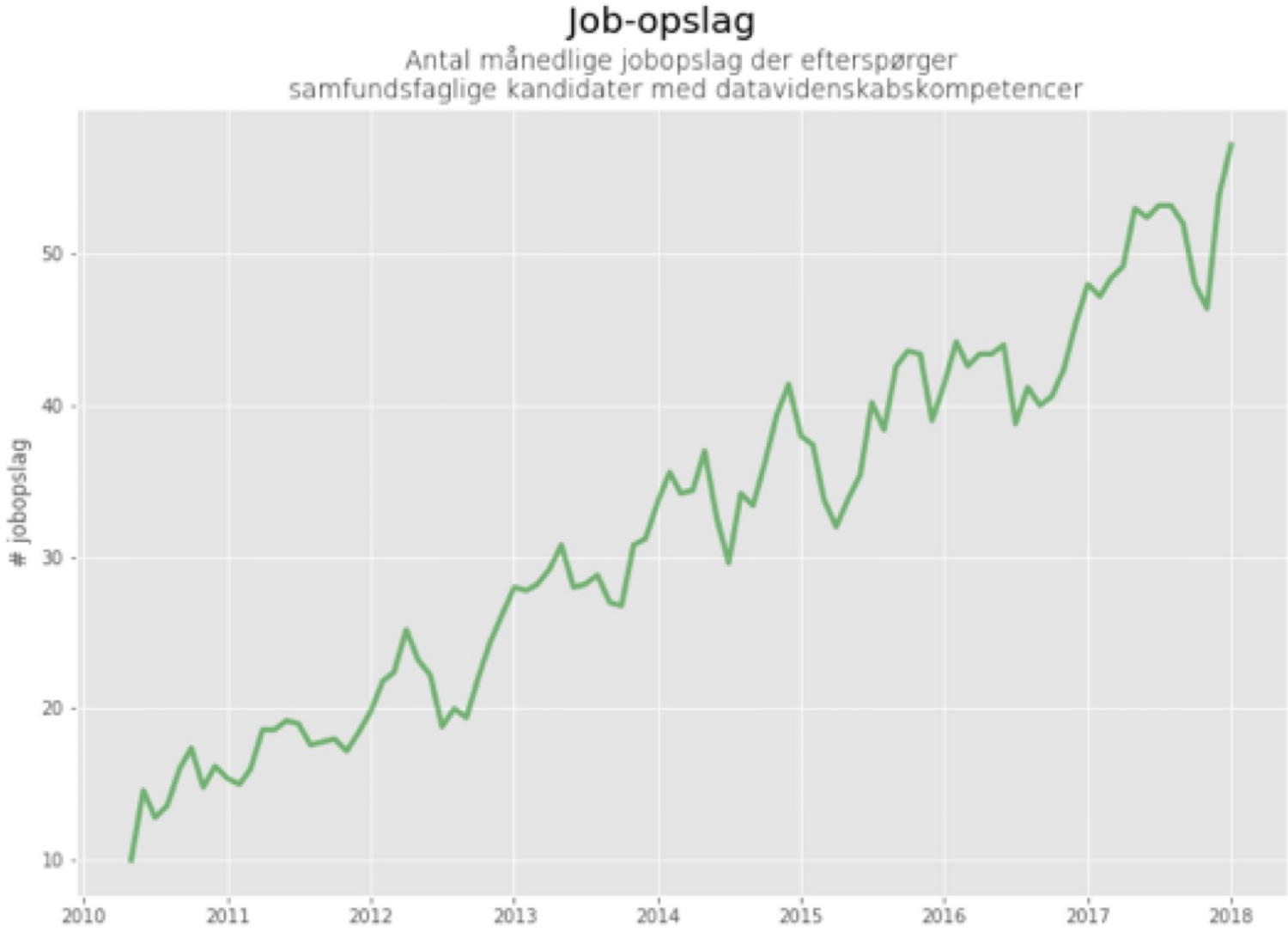
1) Social science methods: collection and structuring of human data

- 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)

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, steps 2-4 take up 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

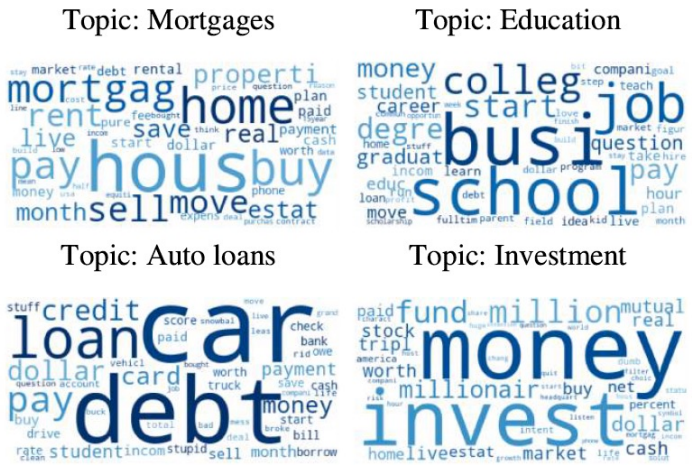
- New methods allow for better extracting meaning from text (Text as Data, e.g. Facebook) and images

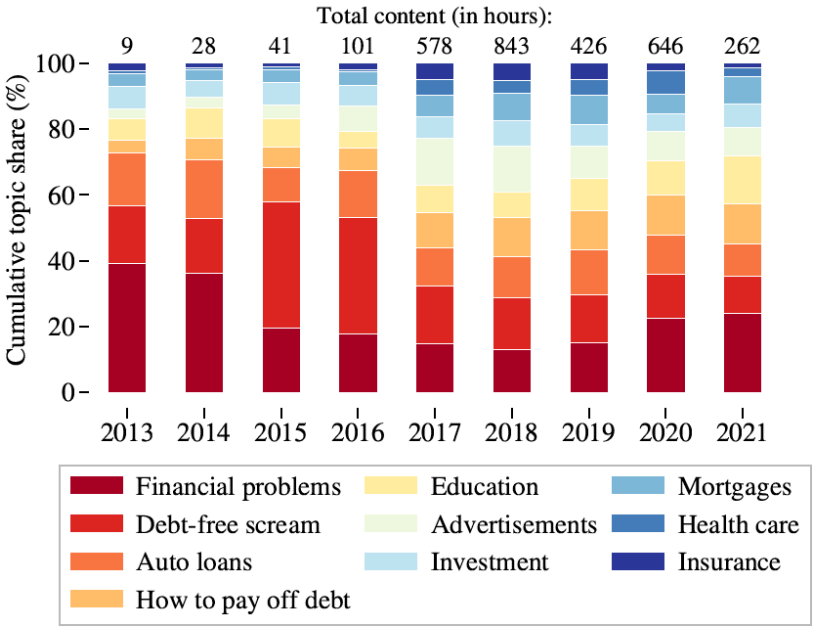
  - Goals: Ability to construct new data aimed at answering old and new social science questions. Make you informed consumers of (Social) Data Science literature.

  - 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.

New ways of enriching traditional research

- Rich data exists on social media.
  - **Example**: In this paper, I study the effect of popular financial advice on financial decision-making of households
  - **Strategy**: The *Dave Ramsey Show* – the most popular finance show in the US – slowly expanded across the US. Allows me to compare decisions before/after introduction.
  - **But:** What is said on the show? Can we answer this with data? Impossible without SDS methods.
  - **What I did**: Webscrape YouTube, download 3,000 hours of material, conduct text analysis.
  - You will be able to do the same after this course!





Peer effects in product adoption

- People leave traces everywhere
  - **Example**: This paper investigates whether peers influence our purchases using Facebook data
  - **Strategy**: Really every app knows your phone model; phones get lost/broken, you buy a new one. What happens to your friends?
  - **Finding**: they are more likely to buy a new phone, too.

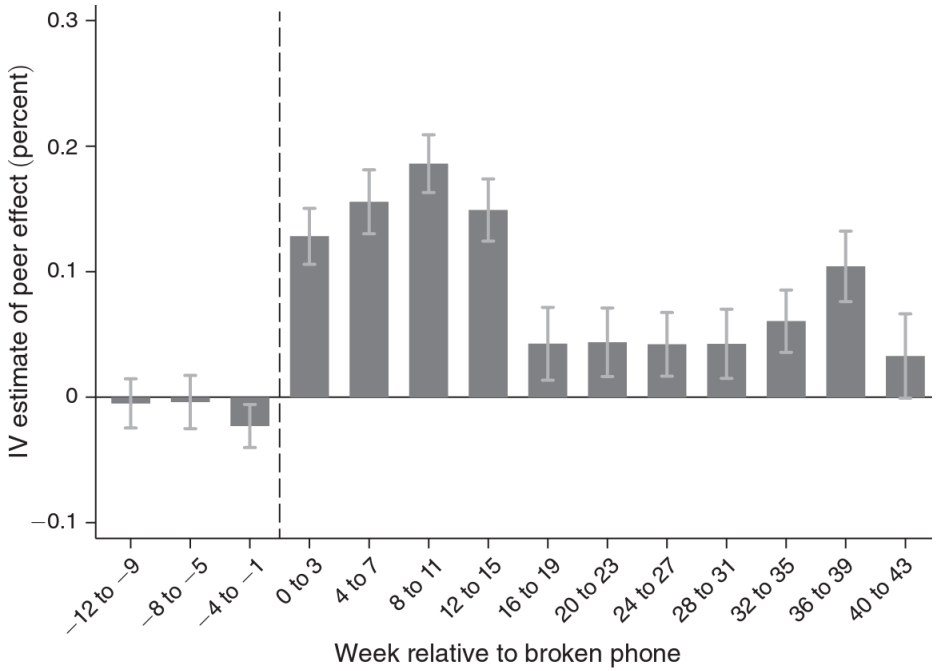


FIGURE 5. PEER EFFECTS AT ALTERNATIVE HORIZONS

Some topics

We will present a social science view on data science methods needed for **collecting** and **analyzing real-world data**. Focus points are on 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 discussions: 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 on our course website
  - New and fast moving topic - brand new excellent textbooks (some on Absalon):

    - 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
- Typically two teaching sessions a day – one in the morning, one in the afternoon – mix of lectures and exercises
- Always bring computer & headphones!
- Absalon for annoucements vs Github for content

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

		——— Week 1 ———		
Jul 31	9-11	1a. Course welcome + intro to git and markdown	FC	
	11-12	1b. Meet group	FC/TAs	
	13-16	2. Data Structuring 1	FC & TAs	
Aug 01	9-12	3. Data Structuring 2	FC & TAs	
	13-16	4. Plotting	FC & TAs	
	16-17	Office Hour	FC	gs
	23:59	Assignment 1 posted	–	
Aug 02	9-12	5. Strings, Queries and APIs	FC & TAs	
	13-16	6. Scraping 1	FC & TAs	
Aug 03	9-12	7. Scraping 2	FC & TAs	
	13:59	Fill in supervision sheet	–	gs, pdf
	14-16	TA help + Supervision *	TAs	gs, pdf
Aug 04	9-12	Exam talk + TA help + Supervision	FC & TAs	gs
	14-16	TA help + Supervision *	TAs	gs
Aug 04	23:59	Assignment 1 hand-in	-	abs

Plan next week

		——— Week 2 ———		
Aug 07	9-12	8. Scraping 3	FC & TAs	
	13-15	9. Data Ethics	FC	
	15-16	Office Hour	FC	gs
Aug 08	9-12	10. ML Introduction	FC & TAs	
	13-16	11. Regression and Regularization	FC & TAs	
	23:59	Assignment 2 posted	–	
Aug 09	9-12	12. Model Selection and Cross-validation	FC & TAs	
	13-16	13. Performance Metrics, Non-linear ML, and Perspectives	FC & TAs	
Aug 10	9-12	14. Text as Data	FC & TAs	
	13-14	Office Hour	FC	gs
	14-16	TA help *	TAs	All previous material
Aug 11	9-12	Exam talk + TA help + Supervision	FC & TAs	
	14-16	TA help + supervision *	TAs	gs
	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?

- More knowledge
  - Degree M.Sc. in Social Data Science @ UCPH with exciting courses, e.g. :
    - (i) Data Governance: Law, Ethics and Politics
    - (ii) Advanced Social Data Science II (on text as data)
  - Machine learning and Econometrics: advanced course on ML and ties to econometrics.
  - More advanced courses in statistical learning, machine learning, data science: Computer science at KU (DIKU), DTU Compute.
- Apply knowledge: 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, ML 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 issues
- 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
- Unil then, remember:
  - Don’t be an (unduly) burden on other servers
  - Identify yourself (as students from UCPH)
  - Don’t violate the terms of service (e.g. Twitter)

Learning to code:
Get ready to get frustrated

No free lunch..

- This 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, more efficiently?
- Practice and try as much as possible. It’s the only way to learn programming.
- Type the code 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, more on that later!)
- Make it understandable: think about audience
  - Future you? May not recall what you did at all.
  - Group members or even the 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 Google/StackOverflow (more advice will follow).

4: You raise an issue in our Github repo or you ask us

- In very rare cases: email (but no Absalon messages)
- Don’t call us (and we won’t call you)

ChatGPT

AI to the rescue?

- New 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

FE

Here is my Python code:

```
x = 10
def x_plus_one():
  x += 1
  print x
x_plus_one()
```

I want the function to add one to x.

This is the error message I get:

```
Traceback (most recent call last):
 File "<stdin>", line 1, in <module>
 File "<stdin>", line 2, in x_plus_one
UnboundLocalError: local variable 'x' referenced before assignment
```

What does the error message mean? Let's think step by step.

Certainly! Let's break down the error and what's happening step by step:

The Error Message

The error message you're getting is:

bash Copy code

```
UnboundLocalError: local variable 'x' referenced before assignment
```

What the Error Message Means

This error essentially tells you that you're trying to use a local variable before it's been assigned a value. But why is it treating `x` as a local variable? Let's delve into the code and how Python functions handle variables.

The Code: Step by Step

1. `x = 10`
   - Here, you're defining a global variable named `x` and assigning the value `10` to it.
2. `def x_plus_one():`
   - This is the definition of your function, `x_plus_one`.
3. Inside the function, you have: `x += 1`
   - This line tries to increment `x` by 1.
   - However, Python interprets this as an assignment operation. Since it's within a function, Python assumes you are referring to a local variable `x`.
   - Because of this assumption, and since you haven't defined any local variable `x` in the function prior to this, Python throws the error. It's basically saying, "Hey, you're trying to increment a local `x`, but there isn't one!"
4. Next, you have: `print x`
   - This is just meant to print the value of `x`, but we never get to this line because of the error in the previous step.

How to Fix the Error

If you want to modify the global `x` within the function, you need to explicitly tell Python that you're referring to the global variable:

Share Regenerate

Reproducibility tools

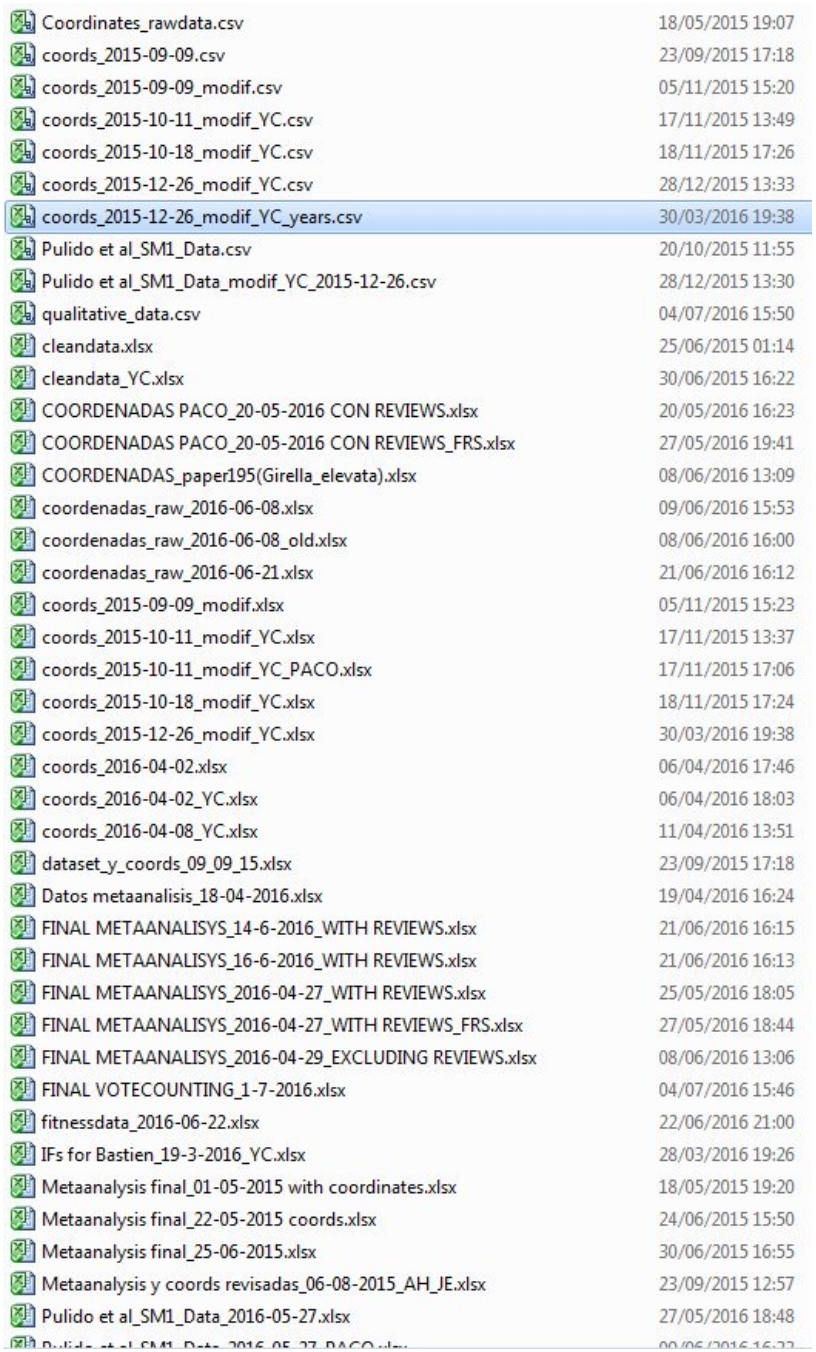
What is git?

- Enables you to work with several people on the same code or text, at the same time.
- Send/receive changes to/from shared folder
  - Git: folder is called a *repository*
  - Changes are explicit - not automatic
- Other KU econ coding/programming courses use it: learn it now and benefit from it throughout your studies

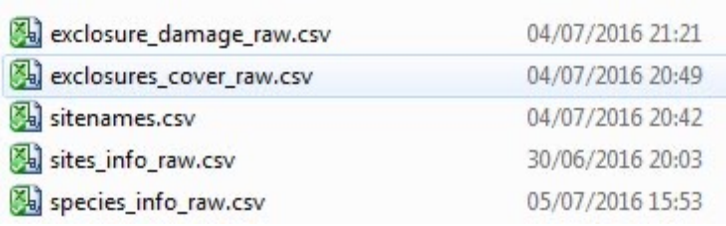
Why git?

- 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



With git



Markdown

- Markdown is an easy to use text editor
  - Alternative to latex - simpler
  - Almost WYSIWYG (what you see is what you get)
- Possible use cases
  - Typeset text in Jupyter Notebook / Lab (see website posts)
  - 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