KØBENHAVNS UNIVERSITET

Kurser

AØKK08216U Summerschool Social Data Science Volume 2016/2017

Course information

Language	English
Credit	7,5 ECTS
Level	Full Degree Master Bachelor
Duration	1 semester
Placement	Summer And Autumn
Schedule	Summer 2016: August 8 to 26. Teaching: August 8 to 12 and 15 to 19, 9 AM - 4 PM. Writing assignment Agust 19 to 26 Autumn: The course will not be offered in Autumn 2016.
Continuing and further education	
Price	320 DKK per ECTS
Study board	Department of Economics, Study Council

Contracting department

- Department of Economics

Course responsibles

- Sebastian Barfort (sebastian.barfort@econ.ku.dk)
- David Dreyer Lassen (david.dreyer.lassen@econ.ku.dk)

Saved on the 18-03-2016

Education

Recommended elective from 3. year at BSc in Economics

Elective at MSc in Economics

Content

The objective of this course is to learn how to analyze, gather and work with modern quantitative social science data. Increasingly, social data--data that capture how people behave and interact with each other--is available online in new, challenging forms and formats. This opens up the possibility of gathering large amounts of interesting data, to investigate existing theories and new phenomena, provided that the analyst has sufficient computer literacy while at the same time being aware of the promises and pitfalls of working with various types of data.

Learning Outcome

1. We will introduce students to the state of the art social science literature using computational methods and social data.

2. We will present students with an overview of key benefits and challenges of working with different kinds of social data. We will show how various kinds of data (survey, web-based, experimental, administrative, etc.) can be used to answer different questions within the social sciences. Furthermore, we will discuss ethical challenges related to the use of different types of data.

3. We will introduce students to statistical techniques for predicting and classification, known as statistical learning, and we will discuss how these methods relate to existing empirical tools within economics such as causal inference and regression.

4. We will present modern data science methods needed for working with computational social science and social data *in practice*. Being an effective economist and data scientist means spending large fractions of our time writing and debugging code. In this section you will learn how to write code to clean, transform, scrape, merge, visualize and analyze social data.

In addition to core computational concepts, the class exercises will focus on the following topics

1. Generating new data: We will learn how to collect and scrape data from websites as well as working with APIs.

2. Data manipulation tools: Participants will learn how to import, transform, munge and merge data from various sources.

3. Visualization tools: We will learn best practices for visualizing data in different steps of a data analysis. Participants will learn how to visualize raw data as well as effective tools for communicating results from statistical models for broader audiences.

4. Reproducability tools: Participants will learn how to use version control and social coding using Github and how to effectively communicate the insights of an analysis using markdown.

6. Prediction tools: We will cover key implementations of statistical learning algorithms and participants will learn how to apply and interpret these models in practice.

1.

After the course the student should:

- Have strong knowledge of the state of the art social science literature using computational methods and social data.
- Have strong knowledge of advantages and challenges in using different kinds of data to answer various questions in the social sciences
- Strong practical data science skills such as the ability to scrape web pages, import and export data from numerous sources, basic knowledge of functional programming and effective data visualization skills.
- Have knowledge of widely used statistical prediction algorithms as well as the ability to estimate these models in practice.

- Strong working knowledge of the R programming language for statistical computing.

Literature

A comprehensive reading list as well as detailed information about the course is available on the course website at

http://sebastianbarfort.github.io/sds_summer/

Teaching and learning methods

The course will consist of 3 hours of lectures and 2 hours of exercises and problem solving per day. The lectures will focus on the broad topics covered in the course (part 1-3 listed above). In the exercise classes we will get our hands dirty and present data science methods needed for collecting and analyzing real-world data.

2 hours of exercises a day is not a large amount of time for learning how to code. We will use some of this time like development meetings: going over assignments, having detailed code reviews of various forms, and discussing blocking issues and potential solutions.

Schedule (tentative):
3 hours of lectures: August 8 to 12, 15, 16 and 19.th, 9.00-12.00 hrs
2 hours of exercise classes: August 8 to 12, 15 and 16.th 13.00-15.00 hrs and 17.th and 18.th 9-12 hrs.

For enrolled students please find more information of courses, schedule, rules etc at
https://intranet.ku.dk/economics_ma/courses/Pages/default.aspx

Timetable and classroom:
For time and classroom please press the link under "Se skema" (See schedule) at the right side of this page.

Academic qualifications

Because the course builds on a wide range of techniques, we do not have any hard requirements, but students are expected to have an interest in some subset of: statistics, econometrics, linear algebra, and a scripting language (we will use R in this course).

Sign up

Self Service at KUnet

Exam (Written take-home)

Credit	7,5 ECTS
Type of assessment	Written assignment, 7 days The exam is an 7-days assignment written in groups, but where the assignment must be handed in individually before deadline.
Exam registration requirements	Students are expected to complete at least 2 out of 3 mandatory assignments. Full participation at the summerschool is mandatory and the student must actively participate in all activities.
Aid	All aids allowed
Marking scale	7-point grading scale
Censorship form	External censorship 100% censorship
Exam period	Summer 2016: The assignment will be given the 19.th of August and has to be handed in not later than 26 August at noon. For enrolled students more information about examination, exam/re-sit, rules etc. is available at the *student intranet for Examination (English),student intranet for Examination (KA-Danish)* and *student intranet for Examination (BA-Danish).*
Re-exam	Same as the ordinary exam.

Criteria for exam assesment
The student must in a satisfactory way demonstrate that he/she has mastered the learning outcome of the course.

Workload

Category	Hours
Lectures	42
Preparation	124
Project work	40
Total	206

Saved on the 18-03-2016

Readings

Text books

We will use the two following books throughout the course

- Grolemund, Garrett and Hadley Wickham. 2016. *R for Data Science*.
- Imai, Kosuke. 2016. *A First Course in Quantitative Social Science*.

None of the books are available for purchase yet. The Grolemund and Wickham book is freely available online. The Imai book is forthcoming at Princeton University Press. Professor Imai has kindly given us permission to use the textbook free of charge in advance of its official release. I will make a PDF of the book available when the course begins. Please do not circulate it!

R & Programming Resources

Here are some books you may find of use throughout the course. None is required to purchase, and readings will be provided as PDFs as needed. But they're good. Note that many of these are available online (e.g. at Springer's SpringerLink website) in their entirety.

- Winston S. Chang. 2013. *The R Graphics Cookbook*. O'Reilly.
- Peter Dalgaard. 2008. *Introductory Statistics with R.* 2nd. Ed. Springer.
- Norman Matloff. 2011. *The Art of R Programming.* No Starch Press.
- Paul Murrell. 2006. *R Graphics*. Chapman & Hall/CRC.
- W.N. Venables and B.D. Ripley. 2002 *Modern Applied Statistics with S*. 4th Ed.
- Hadley Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
- Spraul, V. Anton. 2012. *Think like a Programmer*. San Francisco: No Starch Press.
- Shotts Jr., William E. 2012. *The Linux Command Line: A complete introduction*. No Starch Press, San Fransisco.

Data Visualization

Mandatory

- Schwabish, Jonathan A. 2014. "An Economist's Guide to Visualizing Data". *Journal of Economic Perspectives*, 28(1): 209-34.
- Healy, Kieran and James Moody. 2014. "Data Visualization in Sociology". *Annual Review of Sociology*, 40:105–128.
- Edward R. Tufte. 1983. *The Visual Display of Quantitative Information*. Graphics Press.
- Cox, Amanda. "Data Visualizations at the New York Times".
- Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapters 3 and 21.
- Kahle, David and Hadley Wickham. 2013.''ggmap: Spatial Visualization with ggplot2'', *The R Journal*, 5(1).

Inspiration

- Makela, Susanna. Yajuan Si and. Andrew Gelman. 2015. "Graphical visualization of polling results".
- Gelman, Andrew and Antony Unwin. 2012. ''Infovis and Statistical Graphics: Different Goals, Different Looks''.
- Dodhia, Rahul. Andrew Gelman and Cristian Pasarica. 2002. "Let's practice what we preach: turning tables into graphs". American Statistician, 56: 121–130.
- Wickham, Hadley. 2010. "A Layered Grammar of Graphics". *Journal of Computational and Graphical Statistics*, Volume 19, Number 1, Pages 3–28.

Data Manipulation

- Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis". Journal of Statistical Software 40(1).
- Wickham, Hadley. 2014. "Tidy Data". Journal of Statistical Software 59(10). *The R Journal*. 2(2): 38-40.
- Wickham, Hadley. 2016. "Making Data Analysis Easier". Workshop presentation organised by the Monash Business Analytics Team.

- Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapters 4, 9, 14 and 18.

- Gentzkow, Matthew and Jesse M. Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide". University of Chicago mimeo.

Inspiration

- Lovelace, Robin and James Cheshire. 2013. "Introduction to Spatial Data and ggplot2".

- Brey, Steven. 2014. "Working with Geospatial Data".

- Yollin, Bethany. 2014. "Working with Geospatial Data (and ggplot2)".

Data Import & Web Scraping

Mandatory

- Edelman, Benjamin. 2012. "Using internet data for economic research." *The Journal of Economic Perspectives*, 26.2: 189-206.

- Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapter 8.

- Shiab, Nael. 2015. "Web Scraping: A Journalist's Guide". Global Investigative Journalism Network.

- Shiab, Nael. 2015. "On the Ethics of Web Scraping and Data Journalism". Global Investigative Journalism Network.

- Wickham, Hadley. 2014. "rvest: easy web scraping with R". RStudio Blog.

- Peng, Roger. 2012. "Reading/Writing Data in R". Coursera course: Getting and Cleaning Data.

Inspiration

- Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics*, 118: 26-40.

- Stephens-Davidowitz, Seth, Hal Varian, and Michael D. Smith. 2016. "Super Returns to Super Bowl Ads?". R & R, *Journal of Political Economy*.

- Stephens-Davidowitz, Seth, and Hal Varian. 2015 "A Hands-on Guide to Google Data." Google working paper.

- Barberá, Pablo. 2015. “Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data.” *Political Analysis*, 23.1: 76-91.

- Cavallo, Alberto. “Scraped data and sticky prices”. No. w21490. National Bureau of Economic Research, 2015.

- DiGrazia, Joseph, et al. 2013. “More tweets, more votes: Social media as a quantitative indicator of political behavior.” *PloS one*, 8.11: e79449.

- Diaz, Fernando, et al. 2014. “Online and social media data as a flawed continuous panel survey”. Microsoft Working Paper.

Version Control and Reproducible Research

- Jones, Zachery. 2015. “Git & Github tutorial”.

- Rainey, Carlisle. 2015. “Git for Political Science”.

- Wickham, Hadley. 2015. “Git and GitHub”.

- Bryan, Jennifer. 2016. “Happy Git and GitHub for the useR”

Big Data

- Einav, Liran, and Jonathan Levin. 2014. “Economics in the age of big data.” *Science*, 346.6210: 1243089.

- Einav, Liran, and Jonathan D. Levin. “The data revolution and economic analysis”. National Bureau of Economic Research, No. w19035.

- Grimmer, Justin. 2015. “We are all social scientists now: how big data, machine learning, and causal inference work together.” *PS: Political Science & Politics*, 48.01: 80-83.

- Deutsche Bank Markets Research. 2016. “Big Data in Investment Management”.

- Toole, Jameson L., et al. 2015. “Tracking employment shocks using mobile phone data.” *Journal of The Royal Society Interface*, 12.107: 20150185.

- Gayo-Avello, Daniel. 2013. “A meta-analysis of state-of-the-art electoral prediction from Twitter data.” *Social Science Computer Review*, 0894439313493979.

- Bond, Robert M., et al. 2012. “A 61-million-person experiment in social influence and political mobilization.” *Nature*, 489.7415: 295-298.

- Yougov UK. 2015. “Memories of Iraq: did we ever support the war?”.

- Pew Research Centre. 2015. “From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls”.

- Blackwell, Matthew, and Maya Sen. 2012. ''Large Datasets and You: A Field Guide'', *The Political Methodologist* 20(1): 2-5.

- Mann, Adam. 2016. “Core Concepts: Computational social science.” *Proceedings of the National Academy of Sciences*, 113.3: 468-470.

Causal Inference vs. Statistical Learning

- Varian. Hal. 2014. “Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28.2: 3-27.

- Angrist, Joshua D., and Jörn-Steffen Pischke. 2014. “Mastering’metrics: The path from cause to effect”. Princeton University Press. (pages: XI-XV, 1-14)

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. “The elements of statistical learning”. Vol. 1. Springer, Berlin: Springer series in statistics. (pages: 15-42, 175-184, 214-227)

- Kleinberg, Jon, et al. “Prediction policy problems.” *American Economic Review*, 105.5 (2015): 491-495.

- Breiman, Leo. 2001. “Statistical modeling: The two cultures (with comments and a rejoinder by the author).” *Statistical Science*, 16.3: 199-231.

Inspiration

- Anderson, Chris. 2008. “The end of theory: The data deluge makes the scientific method obsolete.” *Wired*, 16-07.

- Ginsberg, Jeremy, et al. 2009. “Detecting influenza epidemics using search engine query data.” *Nature*, 457.7232: 1012-1014.

- Lazer, David, et al. 2014. “The parable of Google Flu: traps in big data analysis.” *Science*, 343.14.

- Broniatowski, David Andre, Michael J. Paul, and Mark Dredze. 2014. “Twitter: big data opportunities.” *Inform*, 49: 255.

- Lampos, Vasileios, and Nello Cristianini. 2012. “Nowcasting events from the social web with statistical learning.” *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3.4: 72.

- Askitas, Nikolaos, and Klaus F. Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly*, 55.2: 107-120.

- Choi, Hyunyoung, and Hal Varian. "Predicting initial claims for unemployment benefits." Google working paper.

Text as Data

- Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis*, 21.3: 267-297.

- Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapter 11.

Inspiration

- Fariss, Christopher J., et al. 2015. "Human Rights Texts: Converting Human Rights Primary Source Documents into Data." *PloS one*, 10.9: e0138935.

- Jonas, Zachery and Fridolin Linder. 2016. "Exploratory Data Analysis using Random Forests".

Privacy & Ethics

- Heffetz, Ori, and Katrina Ligett. 2014. "Privacy and Data-Based Research." *The Journal of Economic Perspectives*, 28.2: 75-98.

- Acquisti, Alessandro, Curtis Taylor and Liad Wagman. 2015. "The economics of privacy".

- Neuhaus, Fabian, and Timothy Webmoor. 2012. "Agile ethics for massified research and visualization." *Information, Communication & Society*, 15.1: 43-65.

Inspiration

- Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences*, 111.24: 8788-8790.

- Brykman, Steven. 2014. "Facebook's Creepy Case Of Emotional Contagion".

- Meyer, Michelle. 2014. "Misjudgements will drive social trials underground." *Nature* 511.7509: 265-265.

- Tufekci, Zeynep. 2014. “Facebook and Engineering the Public”.
- O’Neil, Cathy. 2016. “The Ethical Data Scientist”. Slate.

Social Data Science

Summer School 2016
Department of Economics
University of Copenhagen

sds_summer

Dette dokument indeholder et overslag over sideantallet på fagets foreløbige pensumliste. Det samlede estimerede sidetantal er 1097.

Derudover indeholder pensum en række videoer samt en tekstbog, som jeg ikke har kunnet finde sidetal på.

Titel på tekst	Sideantal
Total for estimerede sider:	**1097**
Text books	
Grolemund, Garrett and Hadley Wickham. 2016. R for Data Science.	250
Imai, Kosuke. 2016. A First Course in Quantitative Social Science.	Har ikke fundet sidetal
Data Visualization	
Schwabish, Jonathan A. 2014. "An Economist's Guide to Visualizing Data". Journal of Economic Perspectives, 28(1): 209-34.	26
Healy, Kieran and James Moody. 2014. "Data Visualization in Sociology". Annual Review of Sociology, 40:105–128.	23
Edward R. Tufte. 1983. The Visual Display of Quantitative Information. Graphics Press.	200
Cox, Amanda. "Data Visualizations at the New York Times".	Video
Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapters 3 and 21.	Text book
Kahle, David and Hadley Wickham. 2013.''ggmap: Spatial Visualization with ggplot2'', The R Journal, 5(1).	18
Data Manipulation	
Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis". Journal of Statistical Software 40(1).	29
Wickham, Hadley. 2014. "Tidy Data". Journal of Statistical Software 59(10). The R Journal. 2(2): 38-40.	2
Wickham, Hadley. 2016. "Making Data Analysis Easier". Workshop presentation organised by the Monash Business Analytics Team.	Video
Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapters 4, 9, 14 and 18.	Text book
Gentzkow, Matthew and Jesse M. Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide". University of Chicago mimeo.	45
Data Import & Web Scraping	
Edelman, Benjamin. 2012. "Using internet data for economic research." The Journal of Economic Perspectives, 26.2: 189-206.	17
Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapter 8.	Text book
Shiab, Nael. 2015. "Web Scraping: A Journalist's Guide". Global Investigative Journalism Network.	3

Dette dokument indeholder et overslag over sideantallet på fagets foreløbige pensumliste. Det samlede estimerede sidetantal er 1097.

Derudover indeholder pensum en række videoer samt en tekstbog, som jeg ikke har kunnet finde sidetal på.

Shiab, Nael. 2015. "On the Ethics of Web Scraping and Data Journalism". Global Investigative Journalism Network.	5
Wickham, Hadley. 2014. "rvest: easy web scraping with R". RStudio Blog.	3
Peng, Roger. 2012. "Reading/Writing Data in R". Coursera course: Getting and Cleaning Data.	Video
Version Control and Reproducible Research	
Jones, Zachery. 2015. "Git & Github tutorial".	13
Rainey, Carlisle. 2015. "Git for Political Science".	1
Wickham, Hadley. 2015. "Git and GitHub".	23
Bryan, Jennifer. 2016. "Happy Git and GitHub for the useR"	Video
Big Data	
Einav, Liran, and Jonathan Levin. 2014. "Economics in the age of big data." Science, 346.6210: 1243089.	8
Einav, Liran, and Jonathan D. Levin. "The data revolution and economic analysis". National Bureau of Economic Research, No. w19035.	24
Grimmer, Justin. 2015. "We are all social scientists now: how big data, machine learning, and causal inference work together." PS: Political Science & Politics, 48.01: 80-83.	4
Deutsche Bank Markets Research. 2016. "Big Data in Investment Management".	47
Toole, Jameson L., et al. 2015. "Tracking employment shocks using mobile phone data." Journal of The Royal Society Interface, 12.107: 20150185.	36
GayoAvello, Daniel. 2013. "A metaanalysis of state-of-the-art electoral prediction from Twitter data." Social Science Computer Review, 0894439313493979.	19
Bond, Robert M., et al. 2012. "A 61-million-person experiment in social influence nd political mobilization." Nature, 489.7415: 295298.	9
Yougov UK. 2015. "Memories of Iraq: did we ever support the war?".	3
Pew Research Centre. 2015. "From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls".	25
Blackwell, Matthew, and Maya Sen. 2012. ''Large Datasets and You: A Field Guide'', The Political Methodologist 20(1): 2-5.	4
Mann, Adam. 2016. "Core Concepts: Computational social science." Proceedings of the National Academy of Sciences, 113.3: 468-470.	3
Causal Inference vs. Statistical Learning	
Varian. Hal. 2014. "Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28.2: 3-27.	24

Dette dokument indeholder et overslag over sideantallet på fagets foreløbige pensumliste. Det samlede estimerede sidetantal er 1097.

Derudover indeholder pensum en række videoer samt en tekstbog, som jeg ikke har kunnet finde sidetal på.

Angrist, Joshua D., and JörnSteffen Pischke. 2014. "Mastering'metrics: The path from cause to effect". Princeton University Press. (pages: XIXV, 1-14)	14
Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. "The elements of statistical learning". Vol. 1. Springer, Berlin: Springer series in statistics. (pages: 15-42, 175-184,214-227)	49
Kleinberg, Jon, et al. "Prediction policy problems." American Economic Review, 105.5 (2015): 491-495.	5
Breiman, Leo. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical Science, 16.3: 199-231.	32
Text as Data	
Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political Analysis, 21.3:267-297	30
Grolemund, Garrett and Hadley Wickham. 2016. "R for Data Science". Chapter 11.	Text book
Privacy & Ethics	
Heffetz, Ori, and Katrina Ligett. 2014. "Privacy and Data-Based Research." The Journal of Economic Perspectives, 28.2: 75-98.	23
Acquisti, Alessandro, Curtis Taylor and Liad Wagman. 2015. "The economics of privacy".	58
Neuhaus, Fabian, and Timothy Webmoor. 2012. "Agile ethics for massified research and visualization." Information, Communication & Society, 15.1: 43-65.	22
Total for estimerede sider:	**1097**