13: Data fra online-kilder

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth
fh@ifs.ku.dk
fghjorth.github.io
@fghjorth

Institut for Statskundskab
Københavns Universitet

6. december 2017

Fagets opbygning

Blok 1

Gang	Tema	Litteratur	Case
1	Introduktion til R	Leeper (2016)	
2	R workshop + tidy data	Wickham (2014), Zhang (2017)	
3	Regression I: OLS brush-up	AP kap 3	Newman et al. (2015), Solt et al. (2017)
4	Regression II: Paneldata	AGS kap 4	Larsen et al. (2017)

Fagets opbygning

Blok 2

5	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	Eckles & Bakshy (2017)
6	Matching	Justesen & Klemmensen (2014)	Nall (2015)
Efterårsferie			
7	Eksperimenter I	AP kap 1+2, GG kap 1+2	Gerber, Green & Larimer (2008)
8	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	Instrumentvariable	AP kap 4	Lundborg et al. (2017)
10	Difference-in-differences	AP kap 5	
11	Regressionsdiskontinuitetsdesigns	AP kap 6	Eggers & Hainmueller (2009)

Fagets opbygning

Blok 3

12	Tekst som data	Grimmer & Stewart (2013), Benoit & Nulty (2016)	Baturo & Mikhaylov (2013)
13	Scraping af data fra online-kilder	MRMN kap 9+14	Hjorth (2016)
14	'Big data' og maskinlæring	Varian (2014), Montgomery & Olivella (2017)	Theocharis et al. (2016)

How Sudden Censorship Can Increase Access to Information*

William Hobbs† and Margaret E. Roberts‡

January 17, 2017

Abstract

Conventional wisdom assumes that increased censorship will strictly decrease access to information. We delineate circumstances when increases in censorship will expand access to information. When governments suddenly impose censorship on previously uncensored information, citizens accustomed to acquiring this information will be incentivized to learn methods of censorship evasion. These tools provide continued access to the newly blocked information and also extend users' ability to access information that has long been censored. We illustrate this phenomenon using millions of individual-level actions of social media users in China before and after the block of Instagram. We show that the block inspired millions of Chinese users to acquire virtual private networks (VPNs) and join censored websites like Twitter and Facebook. Despite initially being apolitical, these new users began browsing blocked political pages on Wikipedia, following Chinese political activists on Twitter, and discussing highly politicized topics such as opposition protests in Hong Kong.

Talk m. Molly Roberts **onsdag d. 6. december 10-11 i frokoststuen**

http://bit.ly/vkme17evaluering

Opsamling fra sidst

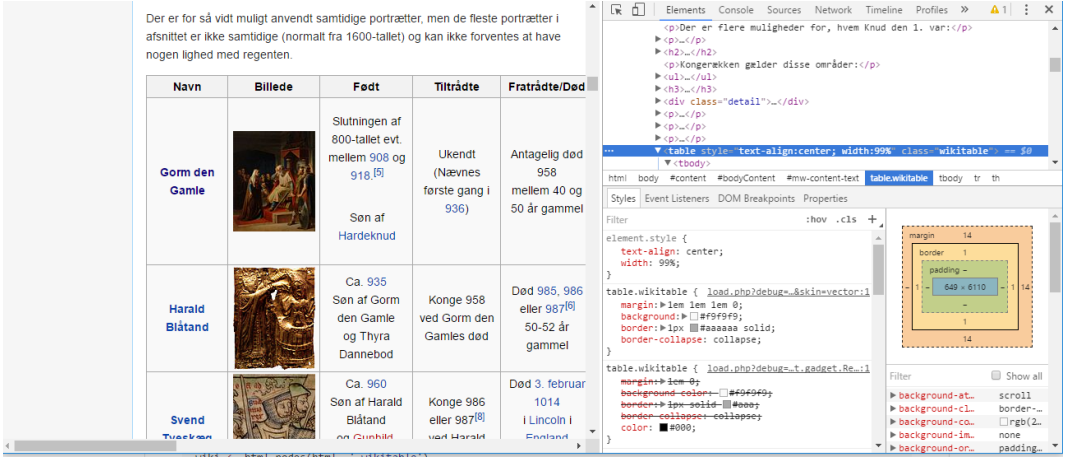
- intro til text as data
- klassifikation
- tf-idf
- skalering
- wordscores
- samler op i dag: Baturo & Mikhaylov

Nyttig pakke til screen scraping: `rvest`

- Typisk fremgangsmåde:
  1. indlæs html med `read_html()`
  2. find relevante elementer med `html_nodes()`
  3. konverter elementerne til data frames med `html_table()`

Eksempel på scraping m. `rvest`: den danske kongerække

Trick: brug 'Inspect...' til at finde HTML/CSS-tegn i kildekoden

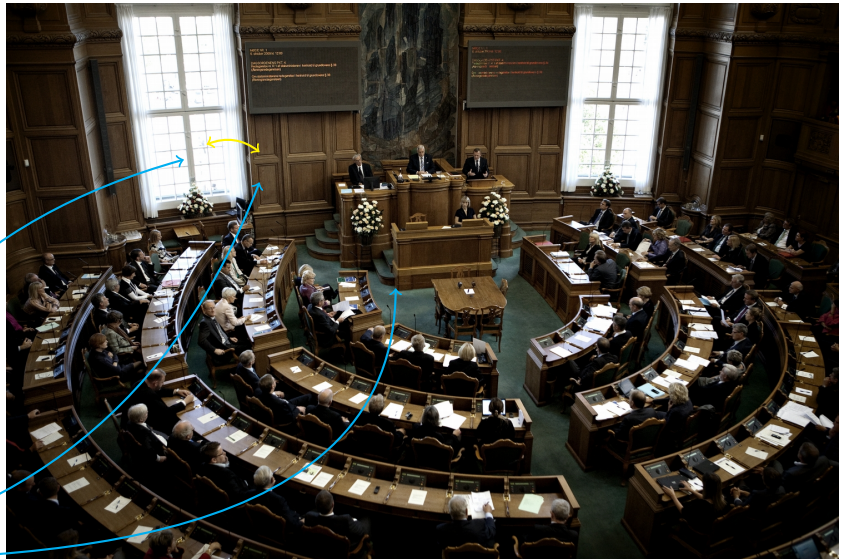


Case I: Taletider fra Folketinget

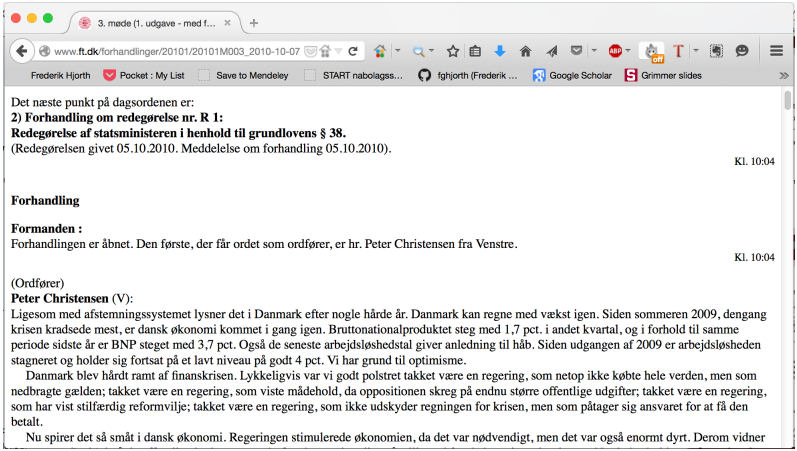
Nøgleroller:

- Formand
- Taler
- andre MF'ere

Nogle gange er taler og formand *'copartisans'*



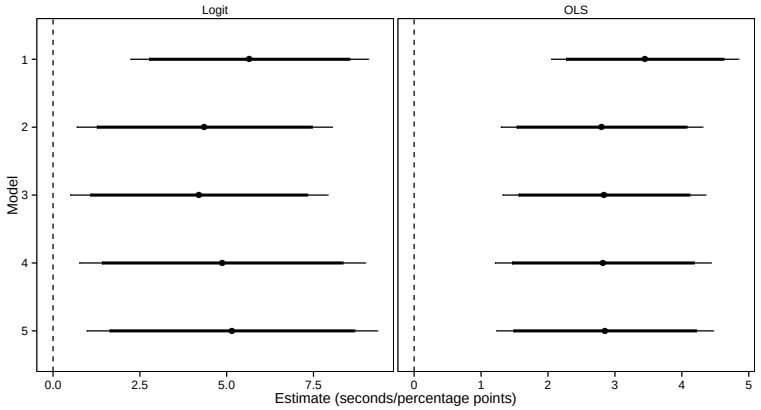
Online resuméer af forhandlinger:



```
 <meta name="OratorFirstName" content="Peter"> </meta> <meta name=
"OratorLastName" content="Christensen"> </meta> <meta name="Group
NameShort" content="V"> </meta> <meta name="OratorRole" content=" medlem">
</meta> <meta name="End MetaSpeakerMP"> </meta> <p class
="TaleType">(Ordfører)</p> <p class="TalerTitelMedTaleType"> <span
class="Bold">Peter Christensen</span> (V):</p> <meta name="Start
MetaSpeechSegment"> </meta> <meta name="LastModified" content="
2011-02-21T09:50:20"> </meta> <meta name="EdixiStatus" content= "Typeset">
</meta> <meta name="StartDateTime" content="2010-10- 07T10:04:20"> </meta>
<meta name="EndDateTime" content="2010-10 -07T10:09:37"> </meta> <meta
name="End MetaSpeechSegment"> </meta> <p class="Tekst">Ligesom med
afstemningssystemet lysner det i Danmark efter nogle hårde år.
```

	fullname	party	starttime	endtime	chair
1	Thor Pedersen	V	2010-10-07 10:04:03	2010-10-07 10:04:20	1
2	Peter Christensen	V	2010-10-07 10:04:20	2010-10-07 10:09:37	0
3	Thor Pedersen	V	2010-10-07 10:15:35	2010-10-07 10:16:24	1
4	Margrethe Vestager	RV	2010-10-07 10:16:24	2010-10-07 10:17:17	0
5	Thor Pedersen	V	2010-10-07 10:17:17	2010-10-07 10:17:19	1
6	Peter Christensen	V	2010-10-07 10:17:19	2010-10-07 10:18:21	0
7	Thor Pedersen	V	2010-10-07 10:18:21	2010-10-07 10:18:22	1
8	Margrethe Vestager	RV	2010-10-07 10:18:22	2010-10-07 10:18:48	0
9	Thor Pedersen	V	2010-10-07 10:18:48	2010-10-07 10:18:50	1
10	Peter Christensen	V	2010-10-07 10:18:50	2010-10-07 10:19:31	0

Estimater:


Logit
OLS
Model
1
2
3
4
5
0.0
2.5
5.0
7.5
0
1
2
3
4
5
Estimate (seconds/percentage points)

Cautionary tale I: Aaron Schwartz vs. JSTOR

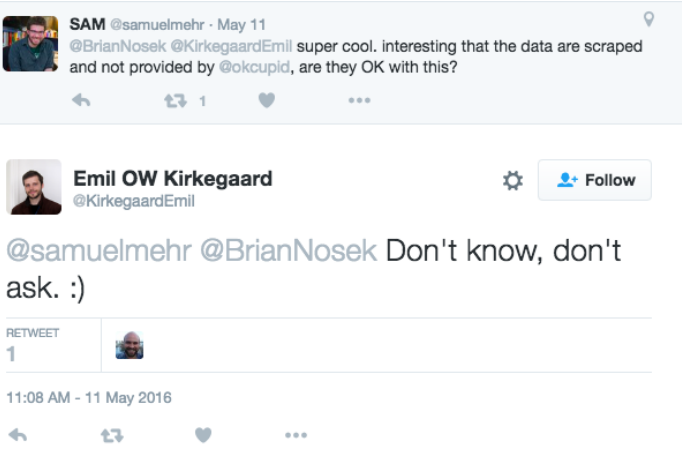
Cautionary tale II: OKCupid data dump

Researchers just released profile data on 70,000 OkCupid users without permission

Updated by Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

→ baggrund hos Vox.com

offentligt tilgængeligt ↛ åbne data



API: Application Programming Interface

»tools which enable programmers to connect their software with 'something else'« (p. 259)

- essentiel byggesten i 'Web 2.0.'
- vigtigt her: værktøj til at rekvirere data

Autorisering

- de færreste API'er er åbne
- typisk autoriseringsstandard: OAuth
- bruges fx. til at autorisere databrug i tredjepartsapps
- kræver *client credentials* for at tilgå data fra API
- typisk procedure:
  1. registrér app på API'ens side
  2. skaf *consumer key+secret*
  3. autorisér vha. key+secret i R

REST: REpresentational State Transfer

- API-standard udviklet til at hente statiske data
- bruger standard HTTP-metoder såsom GET, POST
- API'er der opfylder REST-standarder kaldes RESTful
- mange nyttige R-pakker er wrapper-funktioner for API'er

Twitter's REST API

- tidligere standard: pakken `twitteR`
- nyere, bedre pakke: `rtweet` af Mike Kearney
- consumer key/secret kan tilgås på `apps.twitter.com`
- rate limits: afhængig af type data 15 el. 180 requests pr. 15 min.
- mere info: `http://rtweet.info`

Muligheder og begrænsninger i REST API'en

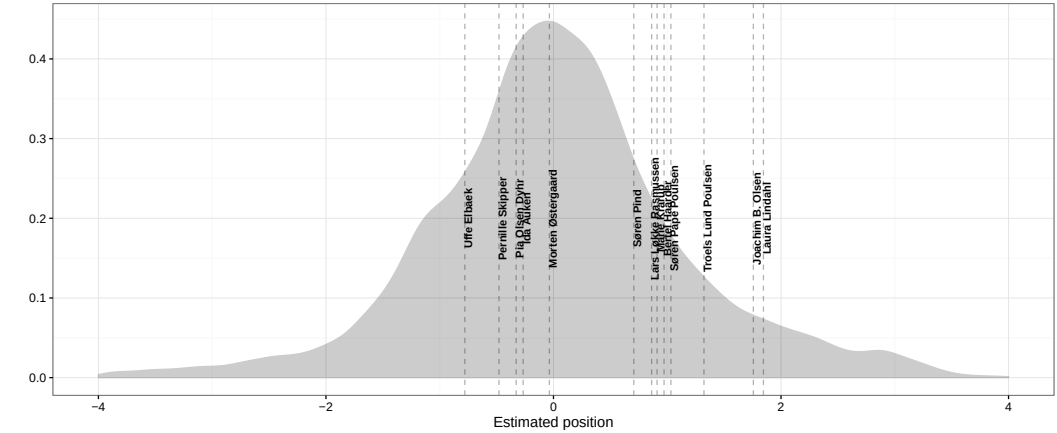
- kan tilgå information om relationer ml. konti
- kan tilgå alle tweets for en given bruger
- kan kun returnere emneordssøgninger 7 dage bagud
- alternativ: gem løbende tweets fra streaming API'er vha. `streamR`

Mål: estimér politiske positioner hos twitter-brugere ($\rightarrow$ sml. m. online adfærd)

Trin 1:

1. find liste med danske folketingsmedlemmer på Twitter ($N = 154$)
2. saml vha. REST API liste over alle unikke followers ($N = 164,041$)
3. konstruér *adjacency matrix* med relationer ml. alle MF'ere og followers
4. reducér adjacency matrix til followers af 10+ MF'ere ($N = 20,091$)
5. estimèr positioner for MF'ere og followers vha. multidimensionel skalering

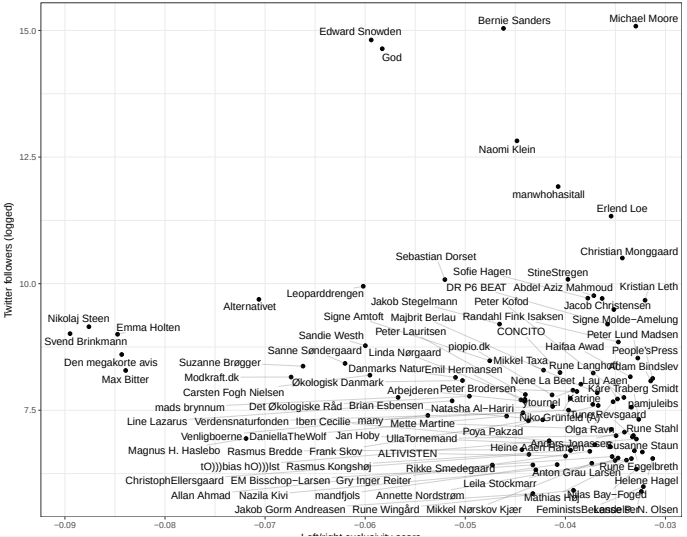
Est. positioner af udvalgte MP'er + 20k followers



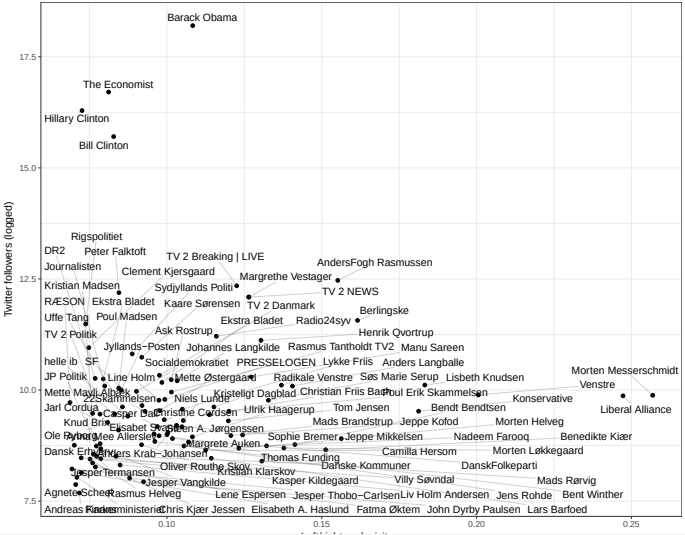
Trin 2:

1. sample $N_L \simeq 500$ konti fra mest venstreorienterede decil og $N_R \simeq 500$ fra mest højreorienterede decil
2. for hver af $\sim$ 410k konti fulgt af en af fløjene, beregn 'exclusivity score' som antal followers fra højre fløj $-$ antal followers fra venstre fløj
3. fjern MF'ere, reskalér til -1 til $+1$
   - -1 $\Rightarrow$ fulgt af alle N_L venstrefløjskonti, ingen N_R højrefløjskonti
   - $+1$ $\Rightarrow$ fulgt af alle N_R højrefløjskonti, ingen N_L venstrefløjskonti
4. fokus på 100 konti med mest ekstreme exclusivity scores

Konti m. exclusivity score < 0



Konti m. exclusivity score > 0 (i.e., right-wing)



Næste gang: 'big data' og maskinlæring

- Varian → god oversigtstekst om big data og ML
- Montgomery & Olivella → specifik om træmodeller i politologi
- Theocaris et al. → læs som case på anvendelse af alle fagets 'data science' moduler

Tak for i dag!