

## 3: Data fra online-kilder

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth

[fh@ifs.ku.dk](mailto:fh@ifs.ku.dk)

[fghjorth.github.io](https://fghjorth.github.io)

@fghjorth

Institut for Statskundskab  
Københavns Universitet

20. september 2018

- 1 Opsamling fra sidst
- 2 Screen scraping
- 3 Case I: Taletider fra Folketinget
- 4 Etik i scraping
- 5 API'er
- 6 Case II: skalering af danske twitter-brugere
- 7 Kig fremad

## Opsamling fra sidst

- kriterier for tidy data
- tidy data verber
- tidyverse
- øvelse:
  - ① hent et af datasættene fra  
`https://github.com/fivethirtyeight/russian-troll-tweets`
  - ② hvad er de hyppigste og næsthyppigste sprog?
  - ③ i hvilken region har trolde flest følgere?
  - ④ hvilket sprog retweetes oftest?
  - ⑤ hvor ofte nævnes hhv. Trump og Clinton i tweetsne?
- løsning(er) under 'solutions'

Flytning af undervisningsgang 5 → se Doodle-link på Github

## 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



Gorm den gamle d. n. 940  
Gorm the Old dead appx. 940



Harald Blåtand o. 940-o. 987  
Harald Bluetooth appx. 940-appx. 987



Svend Tveskæg o. 986-1014  
Svend Forkbeard appx. 986-1014



Knud den Store 1018-1035  
Canute the Great



Svend Estridsen  
1047-1074



Erik Ejegod 1095-1103  
Erik the Good



Niels den Gamle 1104-1134  
Niels the Old



Valdemar den Store 1157-1182  
Valdemar the Great



Knud VI 1182-1202  
Canute VI



Valdemar Sej 1202-1241  
Valdemar the Victorious



Erik Plovpenning 1241-1250  
Erik Plovpenning

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

Der er for så vidt muligt anvendt samtidige portrætter, men de fleste portrætter i afsnittet er ikke samtidige (normalt fra 1600-tallet) og kan ikke forventes at have nogen lighed med regenten.

Navn	Billede	Født	Tiltrådte	Fratrådte/Død
Gorm den Gamle		Slutningen af 800-tallet evt. mellem 908 og 918. <sup>[5]</sup>  Søn af Hardeknud	Ukendt (Nævnes første gang i 936)	Antagelig død 958 mellem 40 og 50 år gammel
Harald Blåtand		Ca. 935 Søn af Gorm den Gamle og Thyra Dannebød	Konge 958 ved Gorm den Gamles død	Død 985, 986 eller 987 <sup>[6]</sup> 50-52 år gammel
Svend Tveskæg		Ca. 960 Søn af Harald Blåtand og Gunhild	Konge 986 eller 987 <sup>[8]</sup> ved Harald	Død 3. februar 1014 i Lincoln i England

Elements Console Sources Network Timeline Profiles >> ⚠ 1

```
<p>Der er flere muligheder for, hvem Knud den 1. var:</p>
<p>...</p>
<h2>...</h2>
<p>Kongerækken gælder disse områder:</p>
<ul>...</ul>
<h3>...</h3>
<div class="detail">...</div>
<p>...</p>
<p>...</p>
<p>...</p>
```

▼ table style="text-align:center; width:99%" class="wikitable" == \$0

▼ <tbody>

html body #content #bodyContent #mw-content-text table.wikitable tbody tr th

Styles Event Listeners DOM Breakpoints Properties

Filter :hov .cls +

element.style {  
text-align: center;  
width: 99%;  
}

table.wikitable { load.php?debug=&skin=vector:1  
margin: 1em 1em 1em 0;  
background-color: #f9f9f9;  
border: 1px solid #aaa solid;  
border-collapse: collapse;  
}

table.wikitable { load.php?debug=t.gadget.Rev:1  
margin: 1em 0;  
background-color: #f9f9f9;  
border: 1px solid #aaa;  
border-collapse: collapse;  
color: #000;  
}

margin 14  
border 1  
padding -  
649 × 6110  
1  
14

Filter Show all

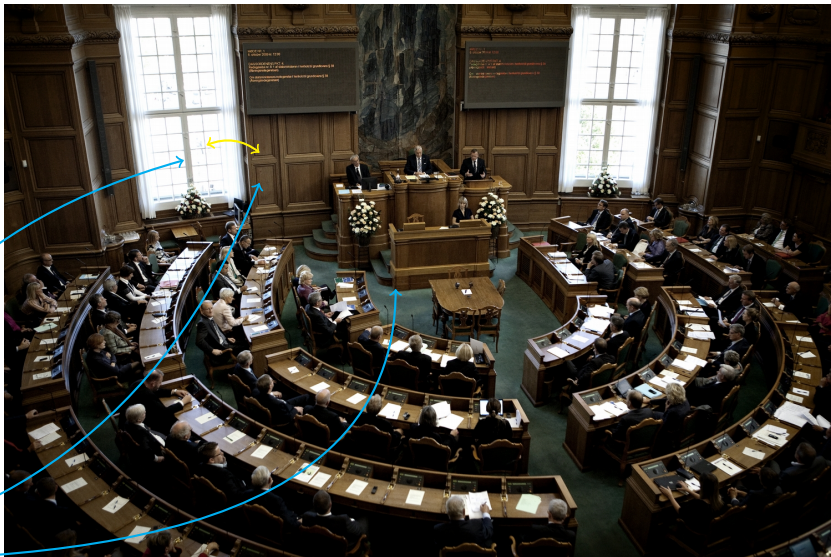
- background-at... scroll
- background-cl... border-...
- background-co... rgb(2...
- background-im... none
- background-or... padding...

## Case I: Taletider fra Folketinget

### Nøgleroller:

- Formand
- Taler
- andre MF'ere

Nogle gange er taler og formand 'copartisans'





## Online resuméer af forhandlinger:

3. møde (1. udgave - med f... ✕ +)

www.ft.dk/forhandlinger/20101/20101M003\_2010-10-07

Frederik Hjorth Pocket : My List Save to Mendeley START nabolags... fghjorth (Frederik ... Google Scholar Grimmer slides

Det næste punkt på dagsordenen er:  
**2) Forhandling om redegørelse nr. R 1:**  
**Redegørelse af statsministeren i henhold til grundlovens § 38.**  
(Redegørelsen givet 05.10.2010. Meddelelse om forhandling 05.10.2010). Kl. 10:04

**Forhandling**

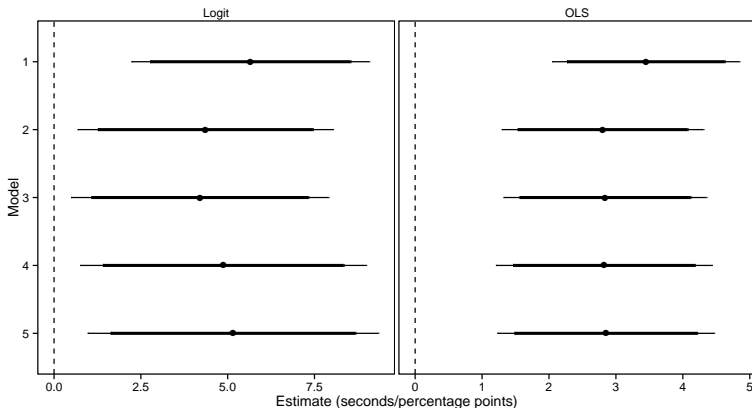
**Formanden :**  
Forhandlingen er åbnet. Den første, der får ordet som ordfører, er hr. Peter Christensen fra Venstre. Kl. 10:04

(Ordfører)  
**Peter Christensen (V):**  
Ligesom med afstemningssystemet lysner det i Danmark efter nogle hårde år. Danmark kan regne med vækst igen. Siden sommeren 2009, dengang krisen kradsede mest, er dansk økonomi kommet i gang igen. Bruttonationalproduktet steg med 1,7 pct. i andet kvartal, og i forhold til samme periode sidste år er BNP steget med 3,7 pct. Også de seneste arbejdsløshedstal giver anledning til håb. Siden udgangen af 2009 er arbejdsløsheden stagneret og holder sig fortsat på et lavt niveau på godt 4 pct. Vi har grund til optimisme.  
Danmark blev hårdt ramt af finanskrisen. Lykkeligvis var vi godt polstret takket være en regering, som netop ikke købte hele verden, men som nedbragte gælden; takket være en regering, som viste mådehold, da oppositionen skreg på endnu større offentlige udgifter; takket være en regering, som har vist stilfærdig reformvilje; takket være en regering, som ikke udskyder regningen for krisen, men som påtager sig ansvaret for at få den betalt.  
Nu spirer det så småt i dansk økonomi. Regeringen stimulerede økonomien, da det var nødvendigt, men det var også enormt dyrt. Derom vidner

```
<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

## Estimator:



## 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



**Emil OW Kirkegaard** @KirkegaardEmil · May 8  
The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) [openpsych.net/forum/showthre...](https://openpsych.net/forum/showthread.php?p=10000)

 26  36 ...

**Ethan Jewett** @esjewett · May 11  
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?



 3  9 ...

**Emil OW Kirkegaard**  
@KirkegaardEmil




 

@esjewett No. Data is already public.

LIKE  
1



1:30 PM - 11 May 2016


 ...





**Ethan Jewett** @esjewett · May 11  
@KirkegaardEmil Differing degrees of "public". Also different ethical guidelines. IMO, you should speak with a research ethicist/IRB ASAP.

**SAM** @samuelmehr · May 11  
@BrianNosek @KirkegaardEmil super cool. interesting that the data are scraped and not provided by @okcupid, are they OK with this?


 1  ...

**Emil OW Kirkegaard**  
@KirkegaardEmil

@samuelmehr @BrianNosek Don't know, don't ask. :)

RETWEET  
1



11:08 AM - 11 May 2016

## Vigtige principper for etisk web scraping:

- brug altid en API hvis muligt
- få tilsagn før større projekter uden om API'er
- hent data med passende pauser
- offentliggør ikke data uden tilsagn





## 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 TTransfer

- 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 `twitterR`
- 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`

**George G. Vega Yon** @gvegayon · Nov 30  
Hey @kearneymw what is/are the main difference(s) btwn rtweet and twitterR?  
[#rstats](#)

2 3 3

**Mike Kearney** @kearneymw  **Following**

Replying to @gvegayon

Unlike `twitterR`, `rtweet` is (a) not deprecated, (b) up-to-date with recent API changes, (c) only R pkg to interact with both REST and stream APIs, (d) more data frame-centric  
[#rstats](#)

7:59 AM - 30 Nov 2017

3 Retweets 31 Likes

1 3 31

## Muligheder og begrænsninger i REST API'en

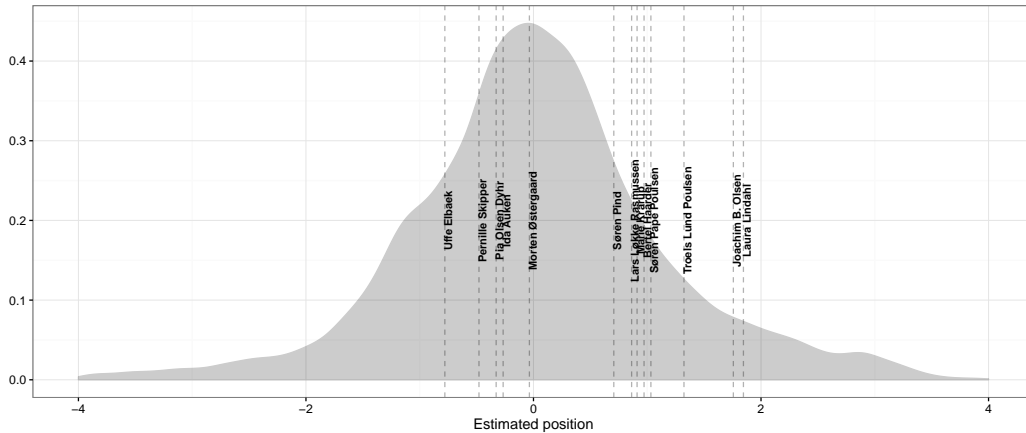
- kan tilgå information om relationer ml. konti
- kan tilgå alle tweets for en given bruger
- kan kun returnere emneordssøgninger 7 dage bagud
- alternativ: gem løbende tweets fra streaming API'er vha. `streamR`

Mål: estimér politiske positioner hos twitter-brugere (→ 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 estimér positioner for MF'ere og followers vha. multidimensionel skalering

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

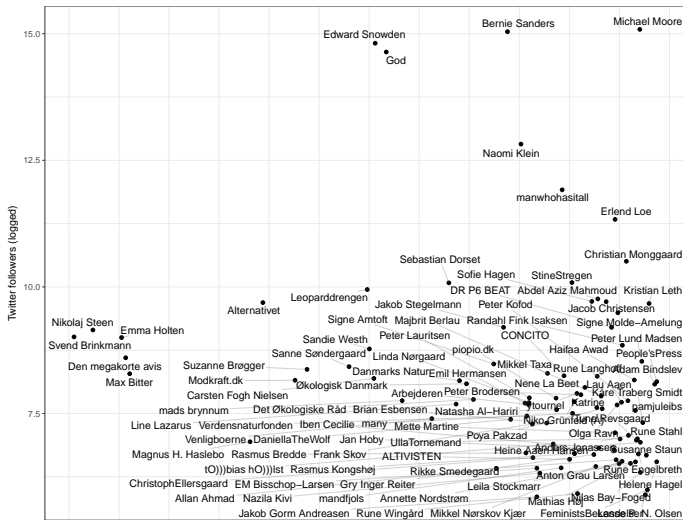


## Trin 2:

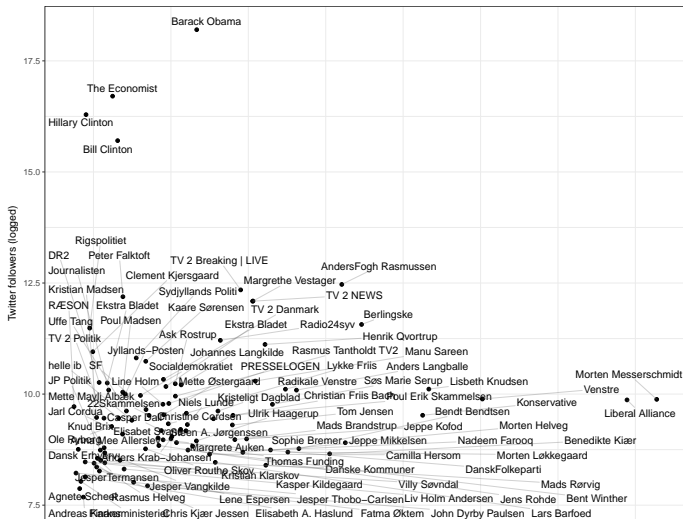
- ① sample  $N_L \simeq 500$  konti fra mest venstreorienterede decil og  $N_R \simeq 500$  fra mest højreorienterede decil
- ② for hver af  $\sim 410k$  konti fulgt af en af fløjene, beregn 'exclusivity score' som antal followers fra højre fløj – antal followers fra venstre fløj
- ③ fjern MF'ere, reskalér 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: tekst som data

- Grimmer & Stewart, Benoit & Nulty
- case: Baturu & Mikhaylov (læs s. 144-148)

## Øvelse til næste gang:

- besøg Conservapedia og hent data om hhv. fremragende konservative film ([https://www.conservapedia.com/Essay:Greatest\\_Conservative\\_Movies](https://www.conservapedia.com/Essay:Greatest_Conservative_Movies)) og elendige venstreorienterede film ([https://www.conservapedia.com/Essay:Worst\\_Liberal\\_Movies](https://www.conservapedia.com/Essay:Worst_Liberal_Movies))
- konstruér (til husbehov) en variabel der måler hver films indtjening i \$
- besvar m. `t.test()`: er fremragende konservative film eller elendige venstreorienterede film mest indbringende?

Opsamling  
○○

Screen scraping  
○○○

Case I  
○○○○○

Etik i scraping  
○○○○

API'er  
○○○○○

Case II  
○○○○○

Kig fremad  
○●

Tak for i dag!