12: 'Big data' og maskinlæring

Videregående kvantitative metoder i studiet af politisk adfærd

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

Institut for Statskundskab
Københavns Universitet

28. november 2016

1. Formalia
2. Opsamling fra sidst
3. Big data I: hype
4. Big data II: skepsis
5. Maskinlæring
   - Regression/classification trees
   - LASSO
6. Implementering i R
7. Kig fremad

Uge	Dato	Tema	Litteratur	Case
1	5/9	Introduktion til R	Imai kap 1	
2	12/9	Regression I: OLS	GH kap 3, MM kap 2	Gilens & Page (2014
3	26/9	Regression II: Paneldata	GH kap 11	Larsen et al. (2016)
4	29/9	Regression III: Multileveldata, interaktioner	GH kap 12	Berkman & Plutzer
5	3/10	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	
6	10/10	Matching	Justesen & Klemmensen (2014)	Ladd & Lenz (2009)
	17/10	*Efterårsferie*		

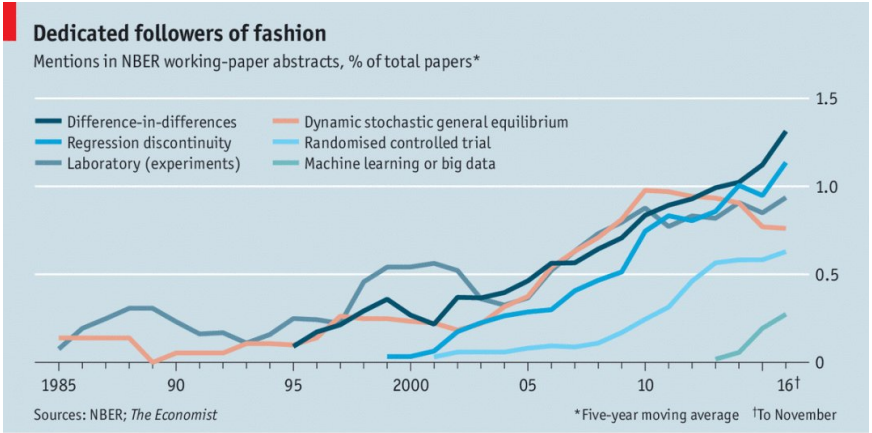
Uge	Dato	Tema	Litteratur	Case
	17/10	*Efterårsferie*		
7	24/10	Eksperimenter I	MM kap 1, GG kap 1+2	Gerber et al. (2008)
8	31/10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	14/11	Instrumentvariable	MM kap 3	Arunachalam & Watso
10	14/11	Regressionsdiskontinuitetsdesigns	MM kap 4	Eggers & Hainmueller
11	21/11	Difference-in-difference designs	MM kap 5	Enos (2016)
12	28/11	'Big data' og maskinlæring	Grimmer (2015), Varian (2014)	
13	5/12	Scraping af data fra online-kilder	MRMN kap 9	
14	12/12	Tekst som data	Grimmer & Stewart (2013), Imai kap 5	

frivillig workshop mandag d. 5/12 kl. 13-15 i `Digital Social Science Lab`
$\rightarrow$ send spm. hertil senest torsdag d. 1/12

- DiD-estimatoren
- parallel trends assumption
- DiD i regressionsform
- lange ctr. korte data
- 'piping'
- case I: flygtningekrisen
- case II: housing demolition

Spørgsmål?

Er big data/ML 'the next big thing'?



Hvad er big data/ML?

»Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value« (De Mauro et al., 2016)

→ defineres ofte med afsæt i 'de 3 V'er'

- Volume: doesn't sample; it just observes and tracks what happens
- Velocity: often available in real-time
- Variety: draws from text, images, audio, video

Hvad er big data/ML?

> the subfield of computer science that »gives computers the ability to learn without being explicitly programmed« (Samuel, 1959)

- machine learning + statistik kaldes nogle gange *data science*
- centralt: fokus på *klassifikation* ctr. kausalitet
- kanoniske eksempler: Google Self-Driving Car Project, Netflix Prize

vigtig, hyppig sondring inden for ML:

- superviserede metoder
  - out-of-sample klassifikationer bygger på kendte værdier i et 'training set'
  - eks.: logit-model
- usuperviserede metoder
  - klassifikationer bygger på in-sample-fit
  - eks.: cluster- eller faktoranalyse

$\rightarrow$ sondringen genoptages i u. 14!

Typisk samfundsvidenskabeligt datagrundlag de sidste $\sim$50 år:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

h/t: Gary King

1. Unstructured text: emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
2. Commerce: credit cards, sales, real estate transactions, RFIDs
3. Geographic location: cell phones, Fastlane, garage cameras
4. Health information: digital medical records, hospital admittances, accelerometers & other devices in cell phones
5. Biological sciences: genomics, proteomics, metabolomics, imaging producing numerous person-level variables
6. Satellite imagery: increasing in scope & resolution
7. Electoral activity: ballot images, precinct-level results, individual-level registration, primary participation, campaign contributions
8. Web surfing artifacts: clicks, searches, and advertising clickthroughs, multiplayer games, virtual worlds

h/t: Gary King

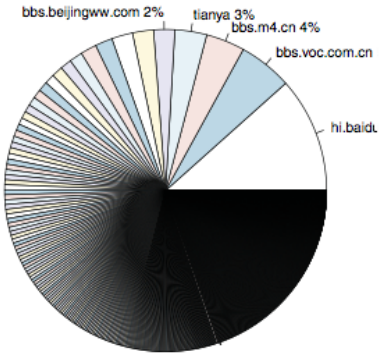
- Opinions of activists: A few thousand interviews → billions of political opinions in social media posts (1B every 2 Days)
- Exercise: A survey: "How many times did you exercise last week? → 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" → continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics → satellite images of human-generated light at night, road networks, other infrastructure

h/t: Gary King

King et al. (2013): indsamling af 3.7M posts, analyse af 127k

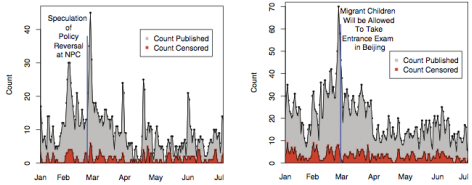
Figure 1. The Fractured Structure of the Chinese Social Media Landscape

(a) Sample of Sites



(b) All Sites excluding Sina

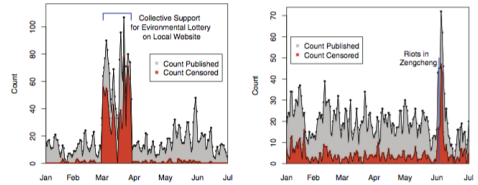
Figure 6. Low Censorship on News and Policy Events (in 2011)



(a) One Child Policy (b) Education Policy

(c) Corruption Policy (Bo Xilai) (d) News on Power Prices

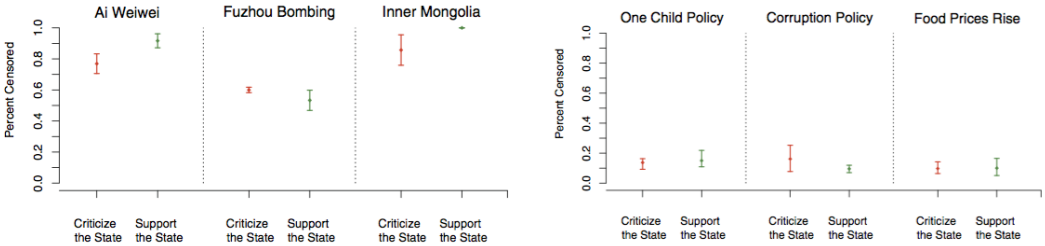
Figure 5. High Censorship During Collective Action Events (in 2011)



(a) Chen Fei's Environmental Lottery (b) Riots in Zengcheng

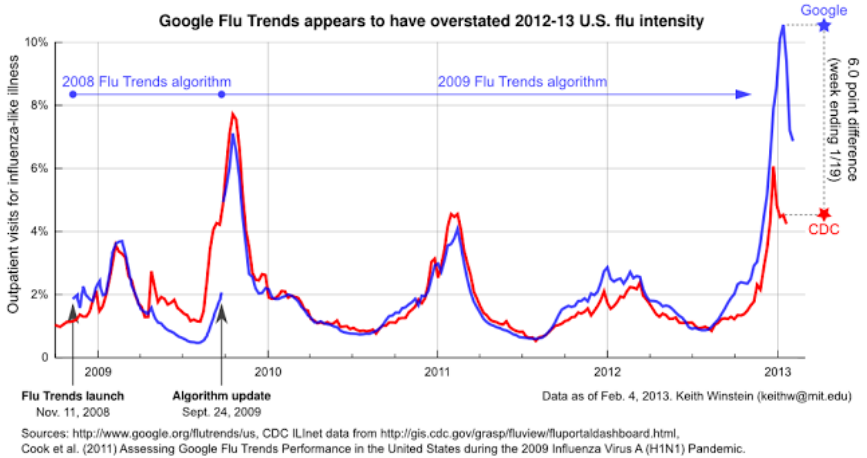
(c) Dissident Ai Weiwei (d) Inner Mongolia Protests

Figure 8. Content of Censored Posts by Topic Area

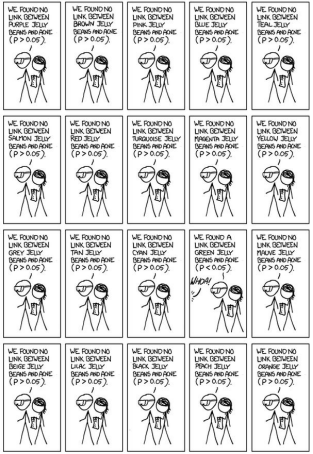


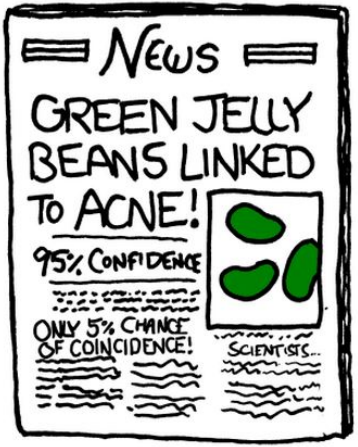
Big data $\approx$ The Literary Digest Poll

Paradigmatisk anekdote: Google Flu Trends



Multiple hypothesis testing

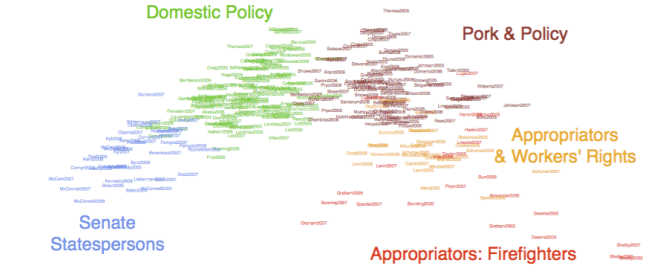




Grimmer:

- mange nødvendige social science-korrektiver til maskinlæring (systematisk målefejl, selektionsbias etc.)
- men: maskinlærings prediktionsmetoder anvendelige til *deskriptiv inferens*
- og: big data kan addressere klassiske sparsity-problemer
- fx: matching, RDD

Eks. fra Grimmers egen forskning: klassifikation af 127k pressemeddelelser



Senate Position Taker	Domestic Policy	Pork & Policy	Appropriators
- Iraq War	- Environment	- WRDA grants	- Fire Grants
- Intelligence	- Gas prices	- Farming	- Airport Grants
- Intl. Relations	- DHS	- Health Care	- University Money
- Budget	- Consumer Safety	- Education Policy	- Police Grants

Eksempel i Varian (2016): overlevelse i Titanic-forliset

Figure 1
A Classification Tree for Survivors of the *Titanic*

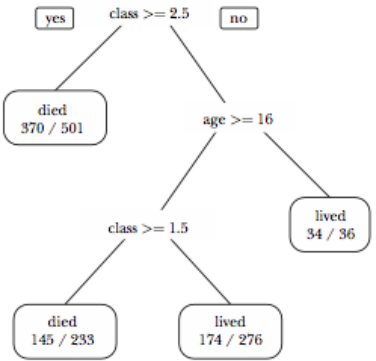


Table 3
Logistic Regression of Survival versus Age

Coefficient	*Estimate*	*Standard error*	*t value*	*p value*
Intercept	0.465	0.0350	13.291	0.000
Age	−0.002	0.001	−1.796	0.072

Note: Logistic regression relating survival (0 or 1) to age in years.

Logikken i regressionstræer:

1. antag fx. to kovariater X_{i1}, X_{i2}
2. SSE uden kovariater: $\sum_{i=1}^{N}(Y_i - \hat{Y})^2$
3. split X_{i1} eller X_{i2} ved c sådan at c minimerer SSE
4. gentag (3) i hvert af de to nye subset ('blade')
5. fortsæt sålænge kriterium for forbedring i fit er opfyldt

Udgangspunkt: least squares-estimatoren for n observationer og p variable:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2.$$

i *penalized regression* estimeres i stedet:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} [(1-\alpha)|\beta_p| + \alpha|\beta_p|^2]$$

$\rightarrow$ den ekstra sum er en *regulariseringsterm*

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} [(1-\alpha)|\beta_p| + \alpha|\beta_p|^2]$$

- denne generelle form: *elastic net regression*
- hvis $\lambda = 0$: reducerer til OLS
- hvis $\alpha = 1$: *ridge regression*
- hvis $\alpha = 0$: *least absolute shrinkage and selection operator* (LASSO)

$\rightarrow \lambda$ fungerer som *tuning-parameter*

Regularisering reducerer risiko for over-fitting $\rightarrow \uparrow$ out-of-sample fit:

»one might consider why the penalty term is needed at all outside the case where there are more covariates than observations. (...) Ordinary least squares is unbiased; it also minimizes the sum of squared residuals for a given sample of data. That is, it focuses on in-sample goodness- of-fit. One can think of the term involving the penalty as taking into account the 'over-fitting' error, which corresponds to the expected difference between in-sample goodness of fit and out-of-sample goodness of fit.« (Athey & Imbens 2016, 47)

LASSO illustrerer dermed også spændingen ml. maskinlæring og kausal inferens:

»LASSO penalizes the inclusion of covariates, and some will be omitted in general; LASSO will favor a more parsimonious functional form, where if two covariates are correlated, only one will be included, and its parameter estimate will reflect the effects of both the included and omitted variables. Thus, in general LASSO coefficients should not be given a causal interpretation.« (Athey & Imbens 2016, 53)

Spørgsmål?

- regressionstræer: `rpart()` i `rpart`-pakken + plots med `rpart.plot`
- LASSO: `glmnet()` i `glmnet`-pakken

Næste gang:

- scraping af data fra online-kilder
- pensum: MRMN kap. 9+14
- vigtigst i kap. 9: afsnit 9.1.10+
- forarbejde: `lav en twitter API key`

Tak for i dag!