14: ‘Big data’ og maskinlæring

Videregående kvantitative metoder i studiet af politisk adfærd

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

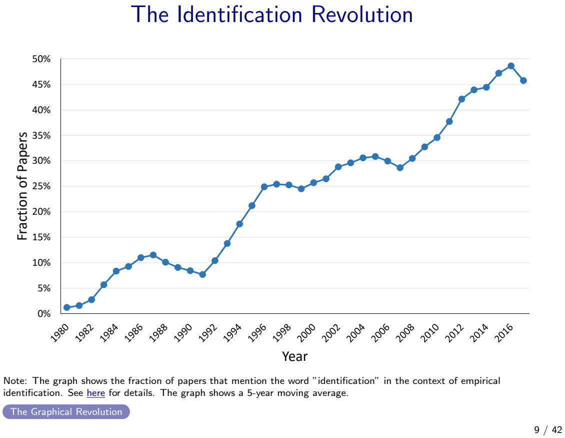
Institut for Statskundskab
Københavns Universitet

13. december 2018

Opsamling fra sidst

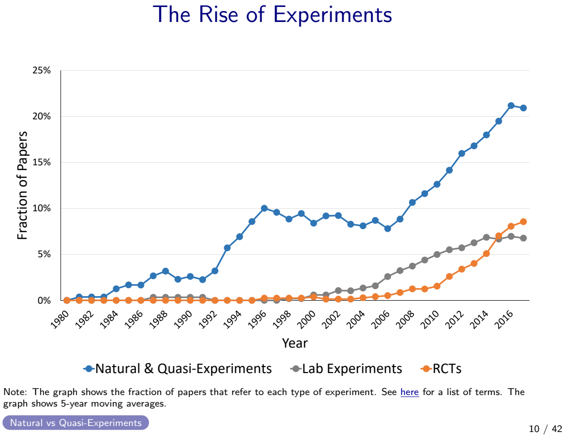
- intro til RD
- formel definition
- RD vs. diff-in-diff
- case: Eggers & Hainmueller

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



Kilde: Henrik Kleven, "Language Trends in Public Economics", July 2018

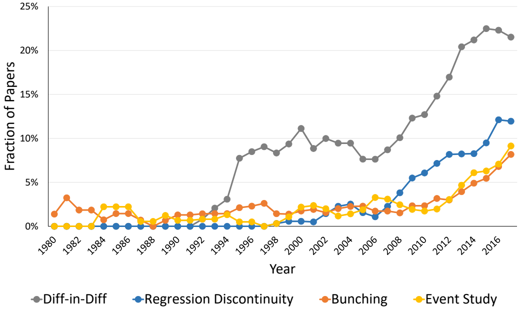
Er big data/ML 'the next big thing'?



Kilde: Henrik Kleven, "Language Trends in Public Economics", July 2018

Er big data/ML ‘the next big thing’?

The Rise Of Quasi-Experiments

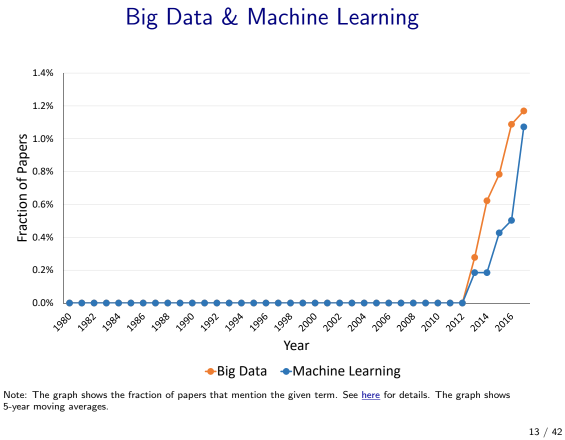


Note: The graph shows the fraction of papers that refer to each type of quasi-experiment. See here for a list of terms. The graph shows 3-year moving averages.

11 / 42

Kilde: Henrik Kleven, “Language Trends in Public Economics”, July 2018

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



Kilde: Henrik Kleven, "Language Trends in Public Economics", July 2018

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/prædiktion* ctr. kausalitet
- kanoniske eksempler: Google Self-Driving Car Project, Netflix Prize

$$\hat{y} \text{ ctr. } \hat{\beta}$$

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

Typisk samfundsvidenskabeligt datagrundlag de sidste ∼50 år:

- surveyforskning
- officiel statistik på aggregeret niveau
- single-case studier af steder, individer eller begivenheder

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

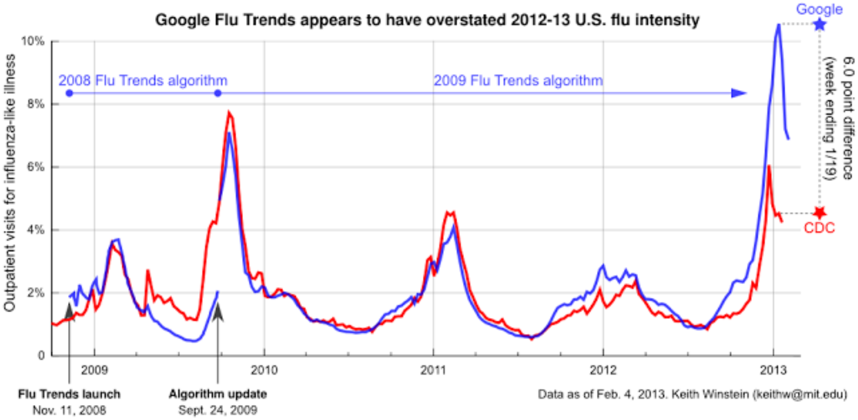
h/t: Gary King

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

h/t: Gary King

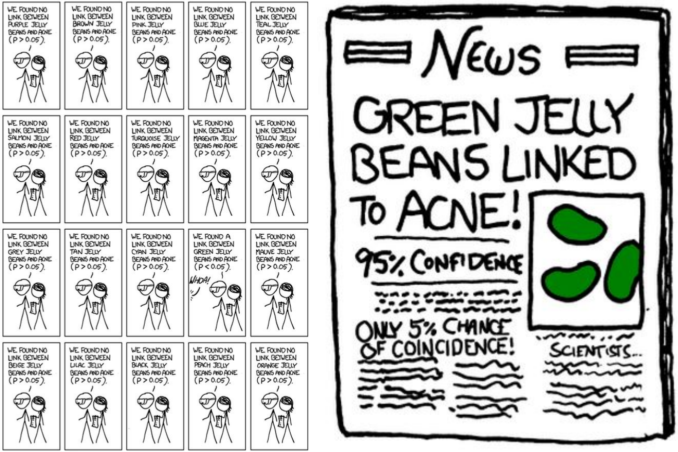
Big data $\approx$ The Literary Digest Poll

Paradigmatisk anekdote: Google Flu Trends



Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

Multiple hypothesis testing

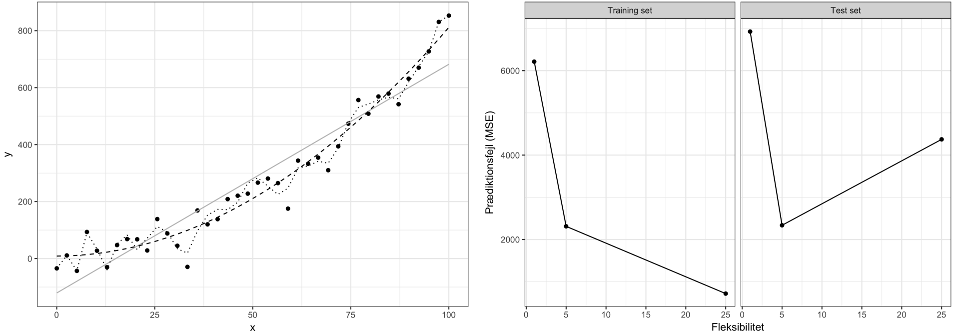


Maskinlæring kan reproducere sociale patologier:

- billedsøgning på 'CEO' returnerer kun hvide mænd
- Google Photo identificerer sorte mænd som 'gorillaer'
- YouTubes text-to-speech modul kan ikke genkende kvindestemmer
- HP Cameras' ansigtsgenkendelsesmodul kan ikke genkende asiatiske ansigter
- Amazon klassificerer LGBT-litteratur som porno
- søgninger efter afroamerikanske navne giver annoncer for baggrundstjeks for kriminalitet

Kilde: Kate Crawford, `https://twitter.com/math_rachel/status/938170475594649600`

Centralt analytisk mål i ML: prediktion uden overfitting



Kilde: Bach, Alexander, Jesper Svejgaard & Frederik Hjorth: “Maskinlæring som politologisk metode". Revise & resubmit, *Politica*.

Centralt analytisk mål i ML: prediktion uden overfitting

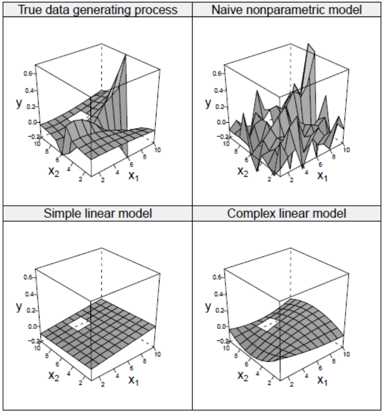


Figure 2. True and recovered relationship in simulated data. The true DGP is $y_i = \frac{1}{100} \times \sqrt{x_{i1}x_{i2}}\frac{(5.5-x_{i1})^2}{(5.5-x_{i1})(5.5-x_{i2})} + \epsilon_i$ where $\epsilon_i \sim N(0, 0.35)$. The simple linear model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, while the complicated model is $y = \beta_0 + \text{poly}(x_1, 2) \times \text{poly}(x_1, 2)$ (where $\text{poly}(x, d)$ is the sequential polynomial generating function, d is the highest degree generated, and the $\times$ operator generates all main effects and interactions).

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

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

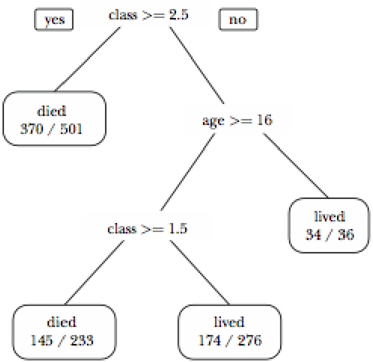


Table 3
Logistic Regression of Survival versus Age

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

Regressionstræ anvendt på Montgomerys simulerede data:

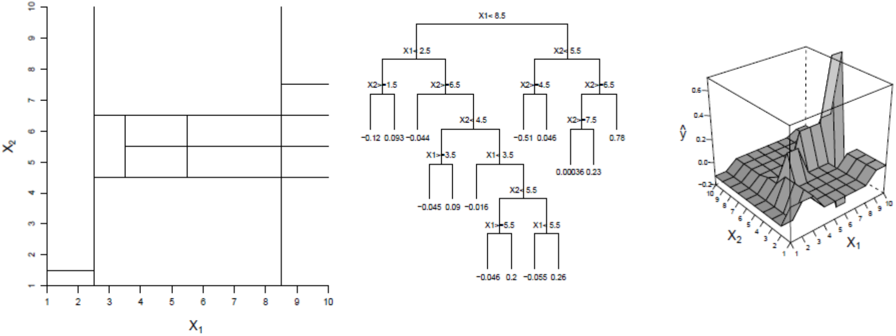


Figure 3. Left: example of a partition of a 2-covariate space into 14 rectangular prediction regions. Center: A binary tree that corresponds to the partition depicted on the left. Right: 3D plot of the prediction surface corresponding to regions defined in the left and center panels.

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

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

i *penalized regression* estimeres i stedet:

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

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

»The second (and contrary) need is to avoid overfitting the data, a goal sometimes labeled regularization. Overfitting occurs when the model is so complex that it makes predictions based on idiosyncratic features of the data unrelated to the true DGP.« (Montgomery & Olivella, 3)

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} [(1-\alpha)|\beta_p| + \alpha|\beta_p|^2] \right) \tag{3}$$

- 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)
- λ fungerer som *tuning-parameter*

vigtig egenskab ved regulariseret regression: mange koefficienter sættes til 0 $\rightarrow$ kan håndtere data hvor antal variable $>$ N

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)

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

Throwback til holdtime 04: document-feature matrice med ≈113k folketingstaler

```
> spdfm
Document-feature matrix of: 113,104 documents, 192,155 features (99.9% sparse).
> spdfm[1:10,1:10]
Document-feature matrix of: 10 documents, 10 features (39% sparse).
10 x 10 sparse Matrix of class "dfm"
                                                      features
docs                                                    : regeringens forslag om  , at alle skal betale 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-10.txt  2           1       1  4 18 11    3    2      2 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-100.txt 2           0       0  1 21  7    0    3      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-101.txt 2           0       0  1 22  5    0    0      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-102.txt 1           0       0  3 16  6    1    1      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-103.txt 2           0       1  2 15  7    0    3      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-104.txt 3           1       0  2 43 18    0    6      0 1
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-105.txt 3           0       0  1  9  7    0    1      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-106.txt 1           1       1  4 21 12    0    2      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-107.txt 3           0       0  2 25 14    0    1      0 0
  19971-1997-10-09-00322384/19971-1997-10-09-00322384-108.txt 1           1       1  2 19  6    0    0      0 0
> length(spdfm)
[1] 21733499120
```

Tak for denne gang og glædelig jul!