Giver store data og data science nye (u)muligheder i samfundsvidenskaben?

David Dreyer Lassen*

Økonomisk Institut &
SODAS - Center for Social Data Science

SAMF
Københavns Universitet

VIVE
4. maj 2018

* med: Ulf Aslak, Sebastian Barfort, Andreas Bjerre-Nielsen, Kelton Minor, Sune Lehmann, Hjalmar Bang Carlsen, Snorre Ralund, Robert Klemmensen m.fl.

Google Trends | Sammenlign

"data science"
Søgeterm

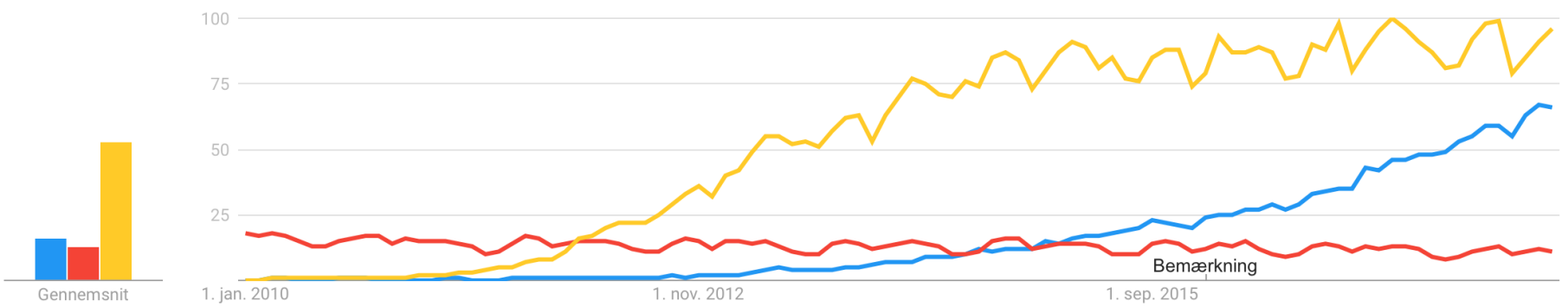
econometrics
Søgeterm

"big data"
Søgeterm

+ Tilføj sammenligning

Hele verden | 01/01/2010 - 18/03/2018 | Alle kategorier | Websøgning

Interesse over tid



Google Trends | Sammenlign

"data science" Søgeterm

econometrics Søgeterm

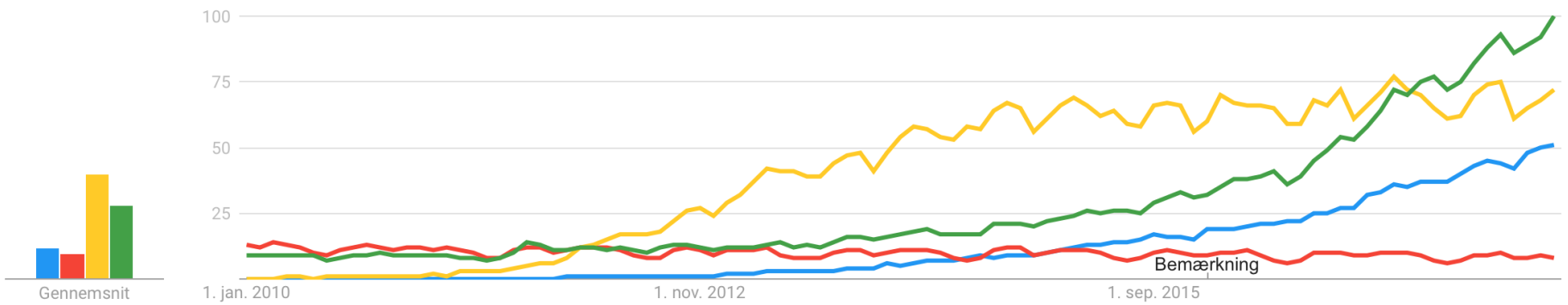
"big data" Søgeterm

"machine learning" Søgeterm

+

Hele verden | 01/01/2010 - 18/03/2018 | Alle kategorier | Websøgning

Interesse over tid



“By almost any market test, economics is the premier social science”

“The starting point in economic theory is that the individual or the firm is maximizing something […] The emphasis on maximization is important because it allows an analyst to make predictions in new situations. […] Other social sciences that are unwilling to assume maximization are in the position of being unable to predict in new situations.”

Lazear (2000, QJE): Economic Imperialism.

"All models are wrong, but some are useful" (Box, 1976)

"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete "
Chris Anderson, *Wired*, 2008

- Traditionel tilgang: Regelbaseret, bl.a. introspektion, teori - deduktiv
- Ny tilgang (machine learning): Lær regler fra træningsdata - induktiv

Datarevolutionen: (data)udbud skaber sin egen (metode)efterspørgsel

Datakilder:

Tidligere: survey, registerdata - analog -> digital administrative data, valideret og processeret centralt. Meget data i økonomi ‘andenhåndsdata’ - men valideret.

Nu: digitale data fra social medier, transaktioner, smartphones, web-scrapings. Førstehåndsdata - ofte ikke-valideret. Data i hænderne på dem, der frembringer dem.

Nogle gange handler data bare om at tælle: En af de vigtigste figurer i de seneste 10 års økonomiske debat

Figure 6-10
Share of Total U.S. Income Earned by Top 1 Percent, 1913–2010

Note: Total income includes wages and salaries (including bonuses and stock-option exercises), pensions, profits, farm income, dividends, interest, and rental income.
Source: Piketty and Saez (2003, 2010); authors provided an estimate for 2010 based on partial returns.

Noget gange skal man have noget at tælle først: Uber

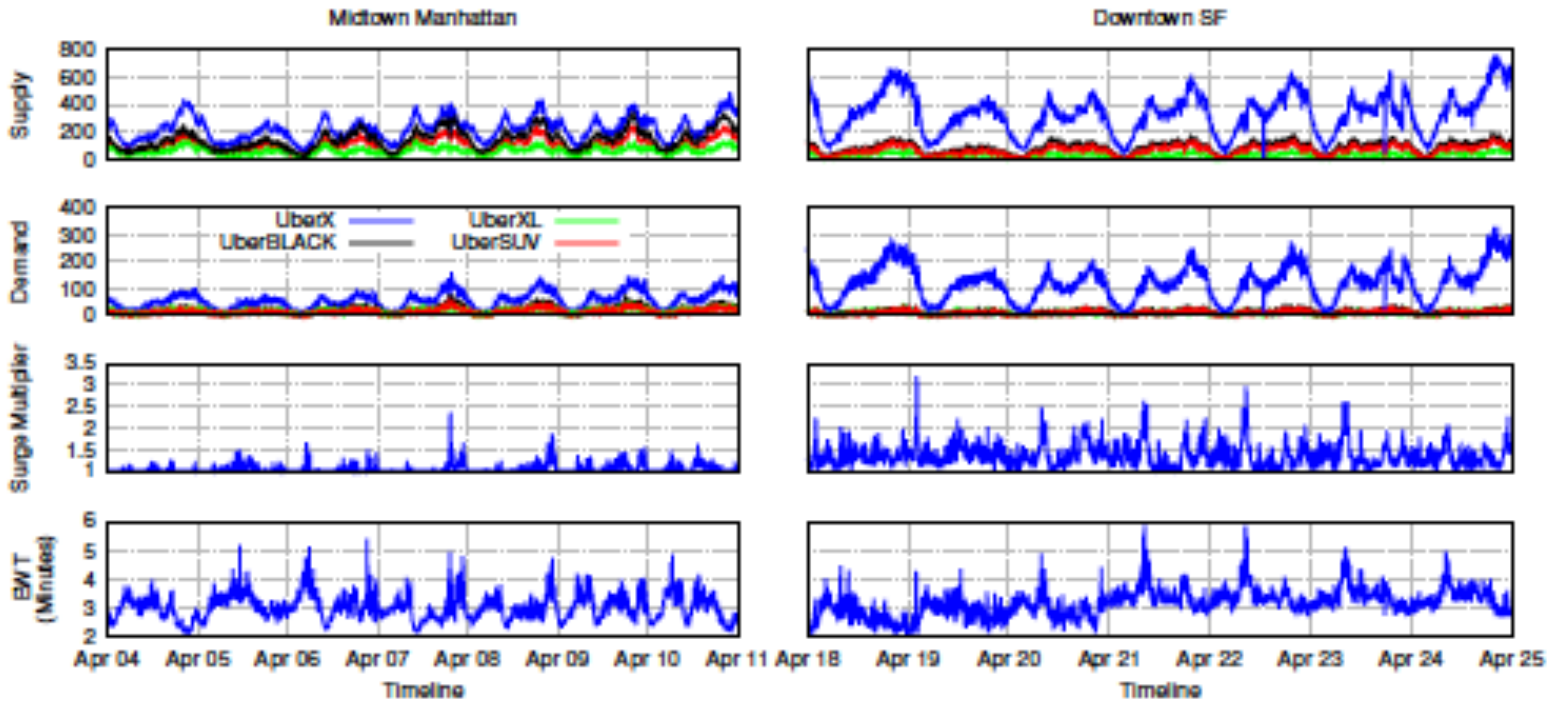


Figure 8: Supply, demand, surge multiplier, and EWT over time for midtown Manhattan and downtown SF. Surge multiplier and estimated wait time (EWT) are only shown for UberX. Diurnal patterns are observed in supply and demand, but the characteristics of the surge multiplier show less predictability.

Chen et al. 2015. “Peaking below the hood of Uber.” *ACM*.

Road map

- Hvad er ‘big data’ og ‘data science’?
- Hvad betyder data science for
  - måling og inferens - “økonometri”/“metode”
  - teori - nye aktører, nye ting der kan testes
  - politik
- AI / robotter / 4. industrielle revolution, arbejdsmarked
- Privacy / persondataforordningen etc.

Hvad betyder ‘big data’ egentlig?

- Oprindeligt: data som er for stort til at kunne håndteres i nuværende software
- fokus på
  - Volume (size: no. of obs, Gigabytes)
  - Variety/complexity (incl. text, pictures, sound etc)
  - Velocity (often high frequency)
  - Veracity (‘honest signals’, behavior)
- Ikke klar skillelinje: Registerdata benævnes ofte ‘big data’

Hvad betyder 'data science' egentlig?

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, uncertainty quantification, computational science, data mining, databases, and visualization.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

When Harvard Business Review called it "The Sexiest Job of the 21st Century"[6] the term became a buzzword, and is now often applied to business analytics,[7] or even arbitrary use of data, or used as a sexed-up term for statistics.[8] While many university programs now offer a data science degree, there exists no consensus on a definition or curriculum contents.[7] Because of the current popularity of this term, there are many "advocacy efforts" surrounding it.[9]

data science vs. social science!
eller
data science + social science?

- “Økonomi er for vigtigt til at overlade til økonomer”
  “Sociologi er for vigtigt til at overlade til sociologer”
  …

- Er data science for vigtigt til at overlade til ingeniører og dataloger?
  “Det er jo bare prediktion …”

Er prediktion vigtigt?

- hvem bliver udsatte børn?
- hvilke finansielle transaktioner er hvidvask?
- hvem reagerer på skatteændringer?
- hvem kan betale lån tilbage?
- hvilke iværksættere får succes?
- …

data science vs. social science!
eller
data science + social science?

- “Økonomi er for vigtigt til at overlade til økonomer”
  “Sociologi er for vigtigt til at overlade til sociologer”
  …

- Er data science for vigtigt til at overlade til ingeniører og dataloger?
  “Det er jo bare prediktion …”

- Sammenlign med statistik vs. sociologisk metode og økonometri eller economic man vs. behavioural economics

data science vs. social science!
eller
data science + social science?

Metoder

Videnskab

Dataficering

data science vs. social science!
eller
data science + social science?

Metoder:

- Machine learning, neurale netværk, deep learning, AI -> prediktionsmodeller,
- datareduktion, dataindhentning
- tekstanalyse, “kvantificering” af lyd, billeder; “kvalificering” af “kvantificering”

Machine learning

- Supervised machine learning
  - ≈ regression, logit -> kender y-variabel
  - Mange metoder: Lasso, random forests etc
  - Træner model - cross-validation, model averaging
- Unsupervised machine learning
  - kender ikke mønstre, bruges til kategorisering,
    ≈ faktoranalyse, typologier

$$y = \alpha + \beta x + \varepsilon$$

- Fokus i traditionel samf: $\hat{\beta}$

- Fokus i ML og prediktion mere generelt: $\hat{y}$

- Helt afgørende: metoder der minimerer bias i estimation af $\hat{\beta}$ vil typisk IKKE minimere varians i estimation af $\hat{y}$

- Trade-off mellem bias og varians

Hatte fra Roth: Introduction, Harvard U, Jan 24. 2018

data science vs. social science!
eller
data science + social science?

- Videnskab
  - Kausalitet / hypotesetest vs prediktion
  - Variabelkonstruktion
  - Aktører: AI og rationalitet
- Dataficering
  - Lovgivning, etik, politik

Eksempel: Selektion og prediktion

- Kleinberg et al. 2018: Beslutninger om varetægtsfængsling i USA.
- Dommer: skal prediktere om anklagede vil dukke op til retssag (og i øvrigt begå ny kriminalitet)
- Eks på problem: observerer kun outcome, hvis ikke varetægtsfængsling
- Her: naturligt eksperiment kombineret med prediktiv algoritme

Kausalitet eller prediktion?

- Simpelt eksempel: Deltagelse og karakterer på uni
- Konstruerer variable for tilstedeværelse baseret på smartphones (selvrapport vs lokation ifht skema vs lokation ifht gruppe)
- Kommer til timer hvis følelse af at få noget ud af det vs. kommer til timer og får faktisk noget ud af det
- Eksempel fra Datalogi: Håndholdt frafaldstjek
- Her: identificerer at-risk personer -> fokus årsager

Social Fabric / Sensible DTU Copenhagen Network Study

- Fulgte ca. 1000 DTU-studerende via smartphones over 1-1.5 år
- Højfrekvente målinger (5 s < < 5 min) af GPS, bluetooth, wifi, SMS, tlf, FB, skærmberøring
- Dynamiske netværk, peer effects (randomisering), sortering
- Her: kender skema, estimer hvorvidt faktisk undervisning

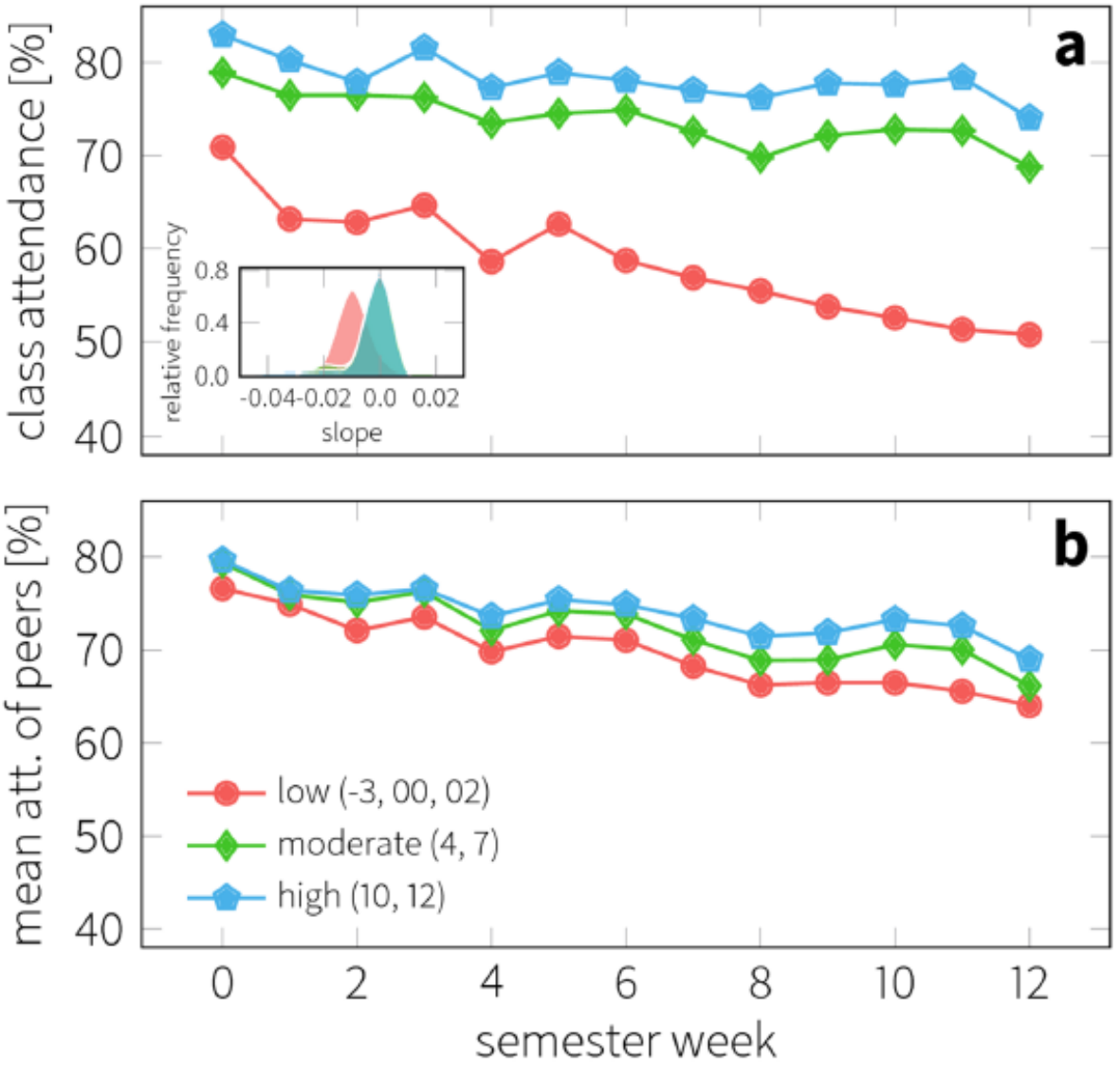


Fig 5. Change in class attendance over a single semester. a) Trends of attendance observed in the three performer groups: low (red circles), moderate (green diamonds) and high performers (blue pentagons), according to the Danish grading system. Inset shows the distribution of slopes measured for each pairs of data point in the trends. b) Mean attendance measured among the contacts of the students based on exchanged text messages.

https://doi.org/10.1371/journal.pone.0187078.g005

Kassarnig, Bjerre-Nielsen, Mones, Lehmann, Lassen. 2017. Class attendance, peer similarity, and academic performance in a large field study. PLOS One.

Eksempel: Peer effects

- Er der peer effects? hvordan virker de? hvorfor (ikke)? Kan man sammensætte grupper og hvad med compliance?
  Litteratur peger i alle retninger (Carrell et al; Angrist)

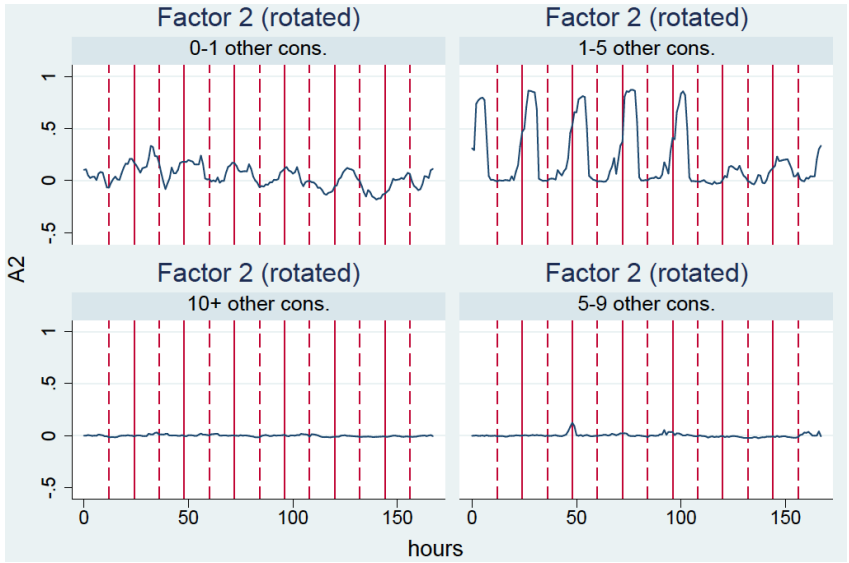
  - Hypoteser: Nogle grupper “virker”, andre “virker ikke” -> effekt = 0? Peer effects kan være forskellige ifht kontekst og outcomes (studievenner -> karakterer eller fredagsbar-venner-> sundhedsoutcomes)

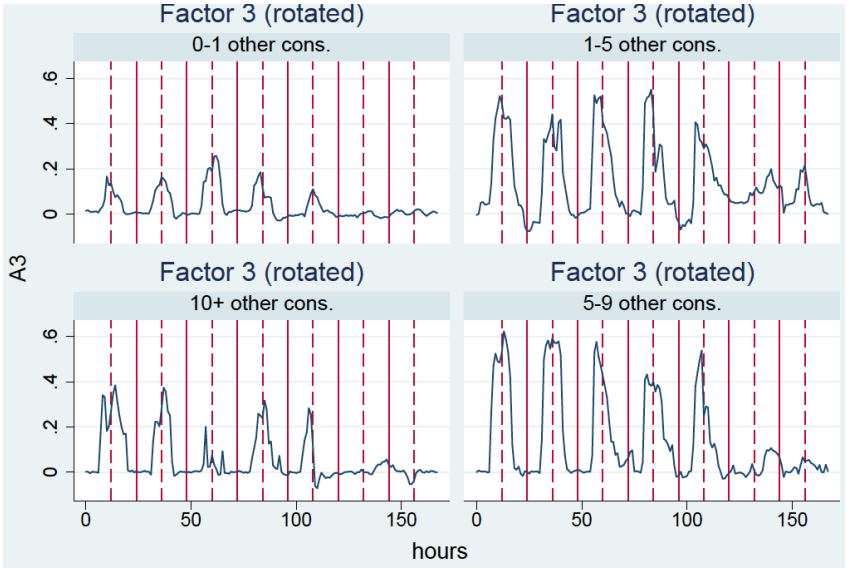
Bjerre-Nielsen, Harmon, Lassen. 2018. “Effective Peer Effects.”
Aslak, Bjerre-Nielsen, Lassen. 2018.

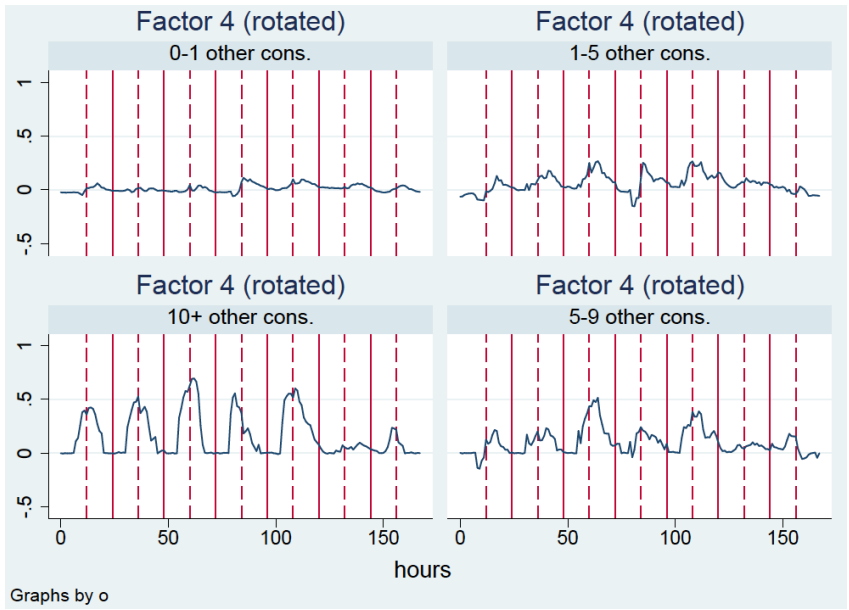
Eksempel: Peer effects

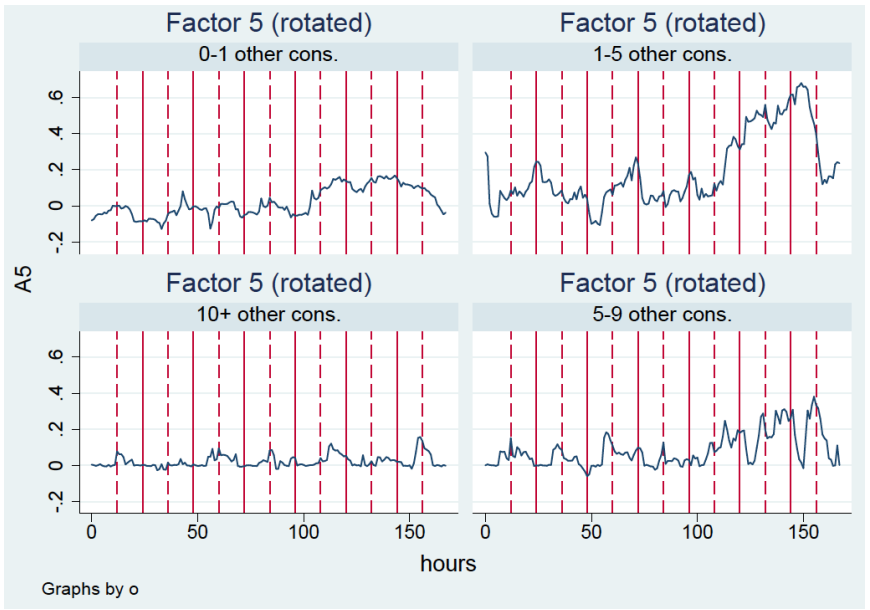
- Metode:
  - Trin 1: Randomiser studerende ind i ‘vektorgrupper’
  - Trin 2: Hvilke faktorer gør at grupper ‘virker’ - og hvordan måler man ‘virker’?
  - Trin 3: Estimer peer effects i forskellige dimensioner udfra ‘kontakt’
- Møder in {IRL, sms, tlf, FB}. Fokus her på 3 mio møder IRL over 3 måneder

Bjerre-Nielsen, Harmon, Lassen. 2018. “Effective Peer Effects.”
Aslak, Bjerre-Nielsen, Lassen. 2018.









Bjerre-Nielsen, Harmon, Lassen. 2018. “Effective Peer Effects.”
Aslak, Bjerre-Nielsen, Lassen. 2018.

Eksempel: Peer effects

- Foreløbige resultater: Gruppeinddeling (first stage) virker
- Tegn på dynamisk sortering på karakterer
- Peer effects?

Bjerre-Nielsen, Harmon, Lassen. 2018. “Effective Peer Effects.”
Aslak, Bjerre-Nielsen, Lassen. 2018.

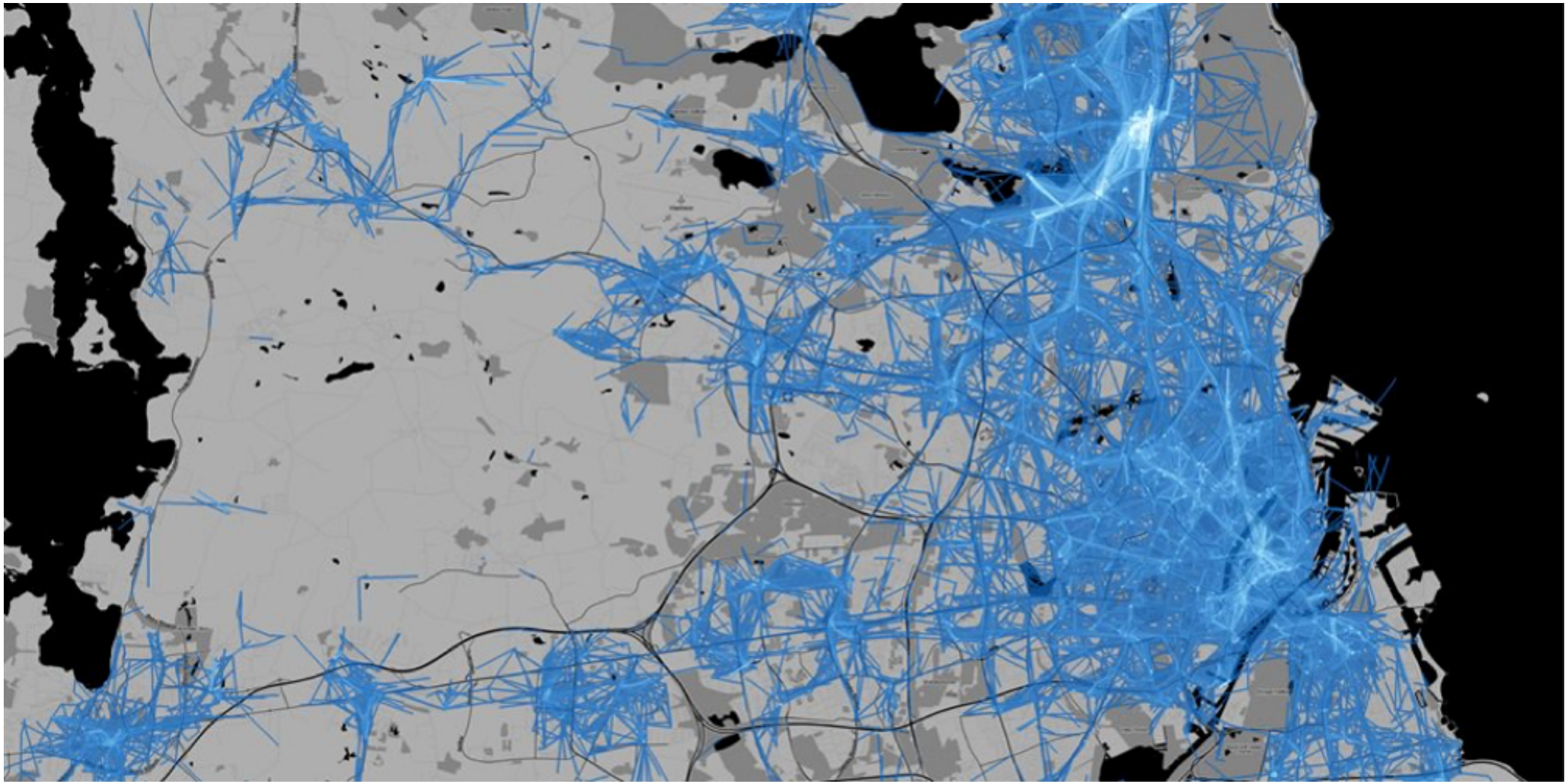
datakonstruktion

1. objekt (teori, politik)
2. Dataindsamling: feasibility (jura, etik, (programmerings-) evner, samarbejde, tid), omkostninger
3. Data-rensning: hvad er objekt, hvad er outliers og fejl (perspektiv: Latour, Pandora's Hope)
4. Variabelkonstruktion, undertiden probabilistisk
5. Validering
6. Analyse

Eksempel: Transportadfærd

- Hvordan transporterer folk sig?
  - Anonyme tællere - ingen individdata (incidens?)
  - Transportsurveys - upræcise, for små?
  - Registerdata om ejerskab af bil - men ikke om brug; potentielt rejsekortdata
  - Automatiseret via smartphones

Eksempel: Transportadfærd



Merged Location

PROJECT
1

By mapping all of the streetside Wi-Fi router locations across Copenhagen we improved the temporal resolution of the location data.

Merged Location

PROJECT
1

By mapping all of the streetside Wi-Fi router locations across Copenhagen we improved the temporal resolution of the location data.

Eksempel: Transportadfærd

- Mål: at inferere transport-type alene fra mobildata
- Hvordan infereres transport-type?
- Supervised ML kræver ‘labeled data’

  korrespondance mellem ‘ground truth’ og mobilsignal, detaljeret mobil-rejse-dagbog -> træningsdata

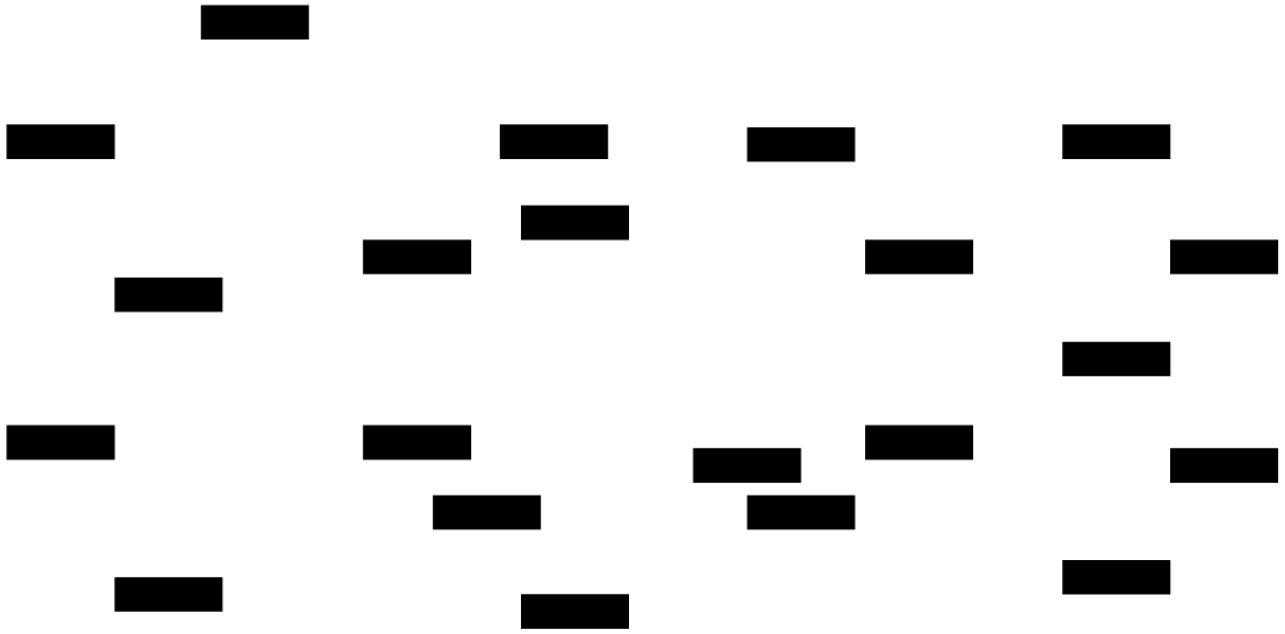
Bjerre-Nielsen, Minor, Sapiezynskic, Lassen, Lehmann. 2018. Wi-Finder: Urban Transportation Sensing Using Crowdsourced Wi-Fi and Contextual Clues. Manuskript, Kraks Fond for Byforskning

Machine Learning

PROJECT
1

Classification
Identify how many of these trips were taken on a bus

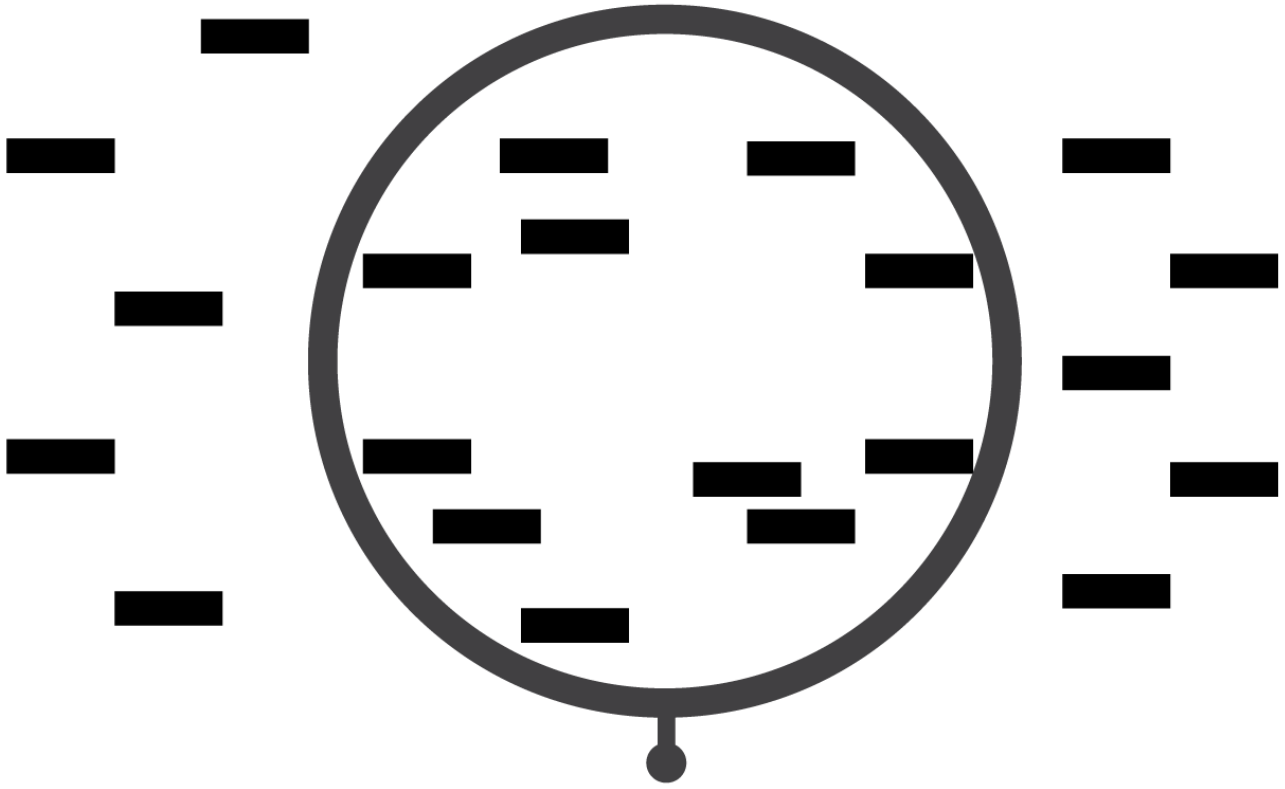
= 1 trip



Machine Learning

PROJECT
1

Classification
Identify how many of these trips were taken on a bus

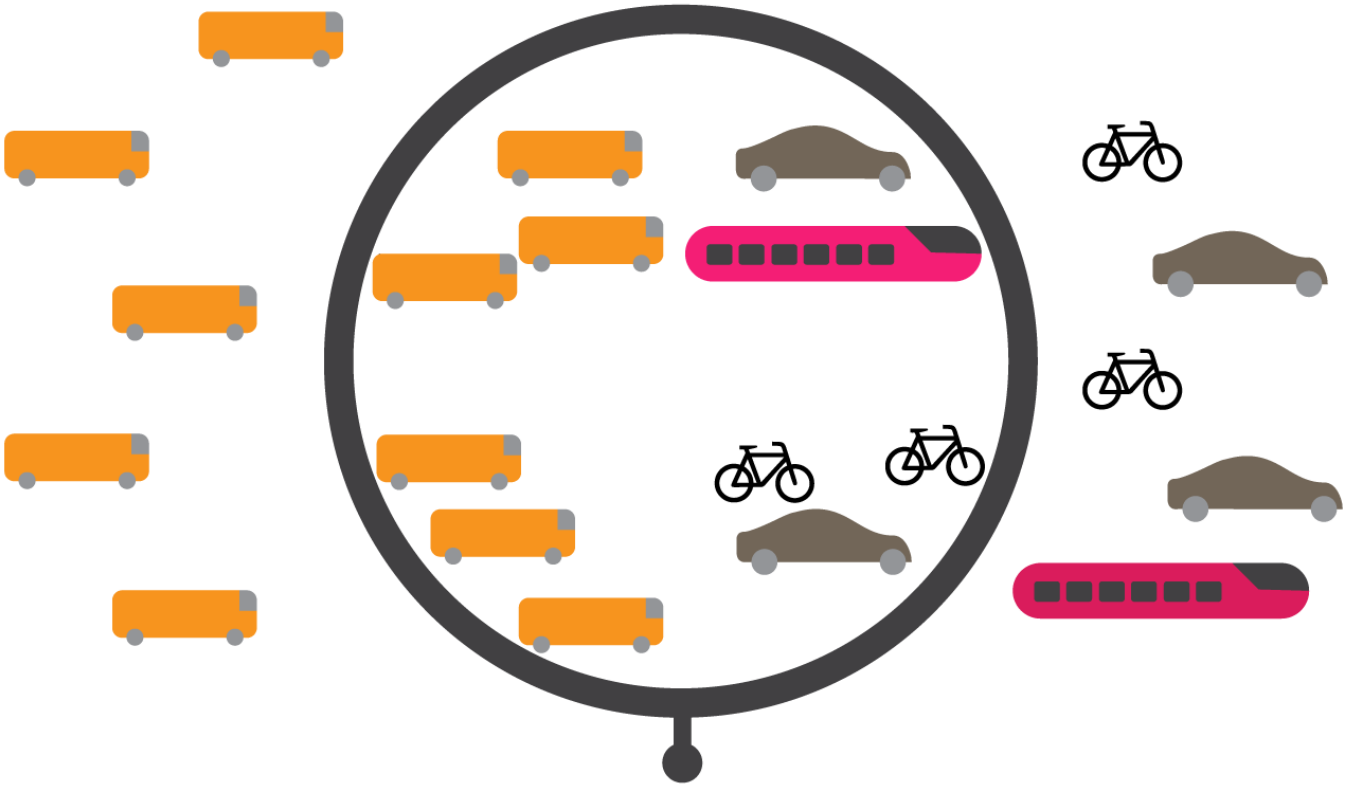


Machine Learning

PROJECT
1

Classification
Identify how many of these trips were taken on a bus

relevant items

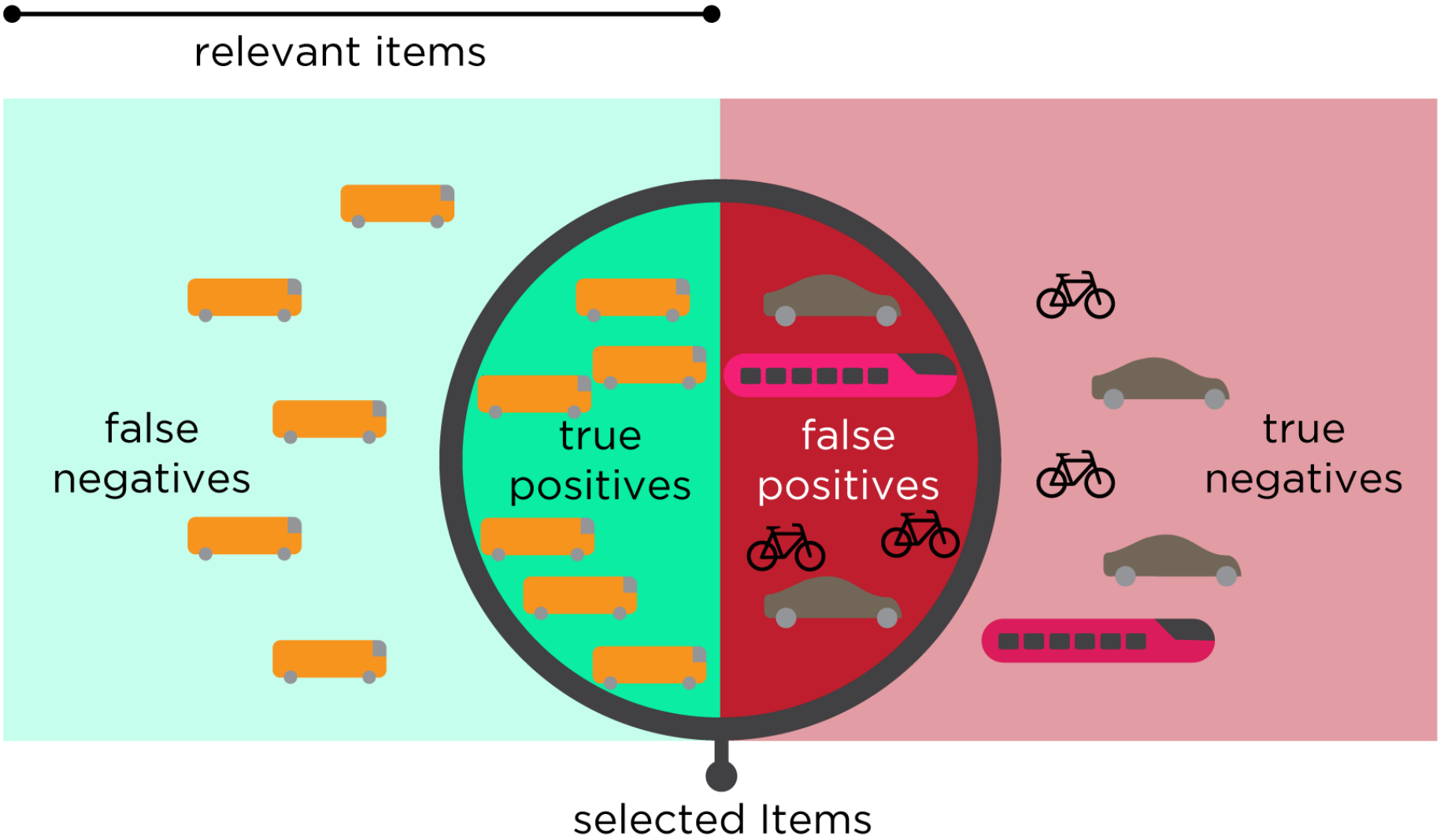


selected Items

Machine Learning

PROJECT
1

Classification
Identify how many of these trips were taken on a bus

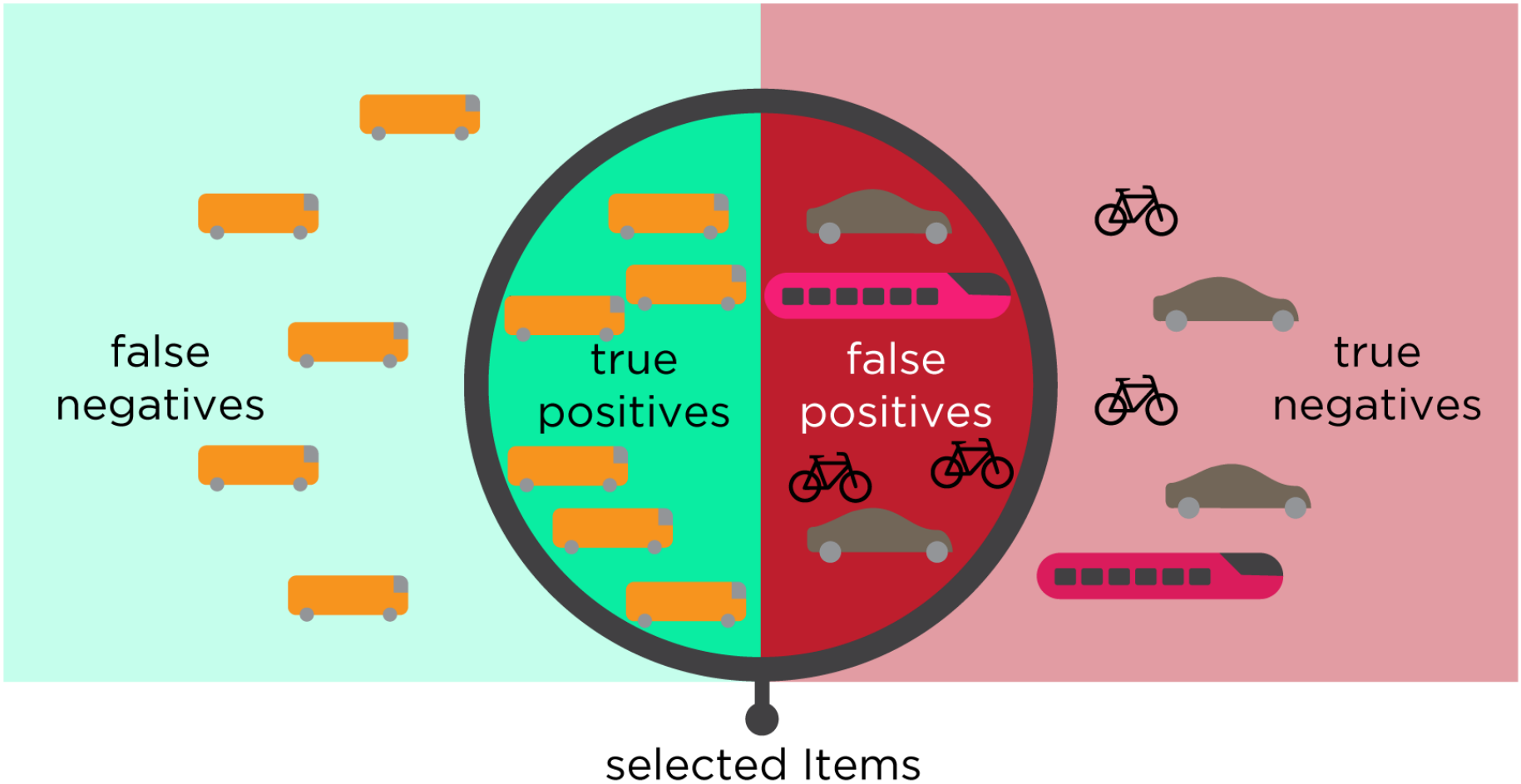


Machine Learning

PROJECT
1

Precision - Exactness
how many selected items are relevant?

relevant items

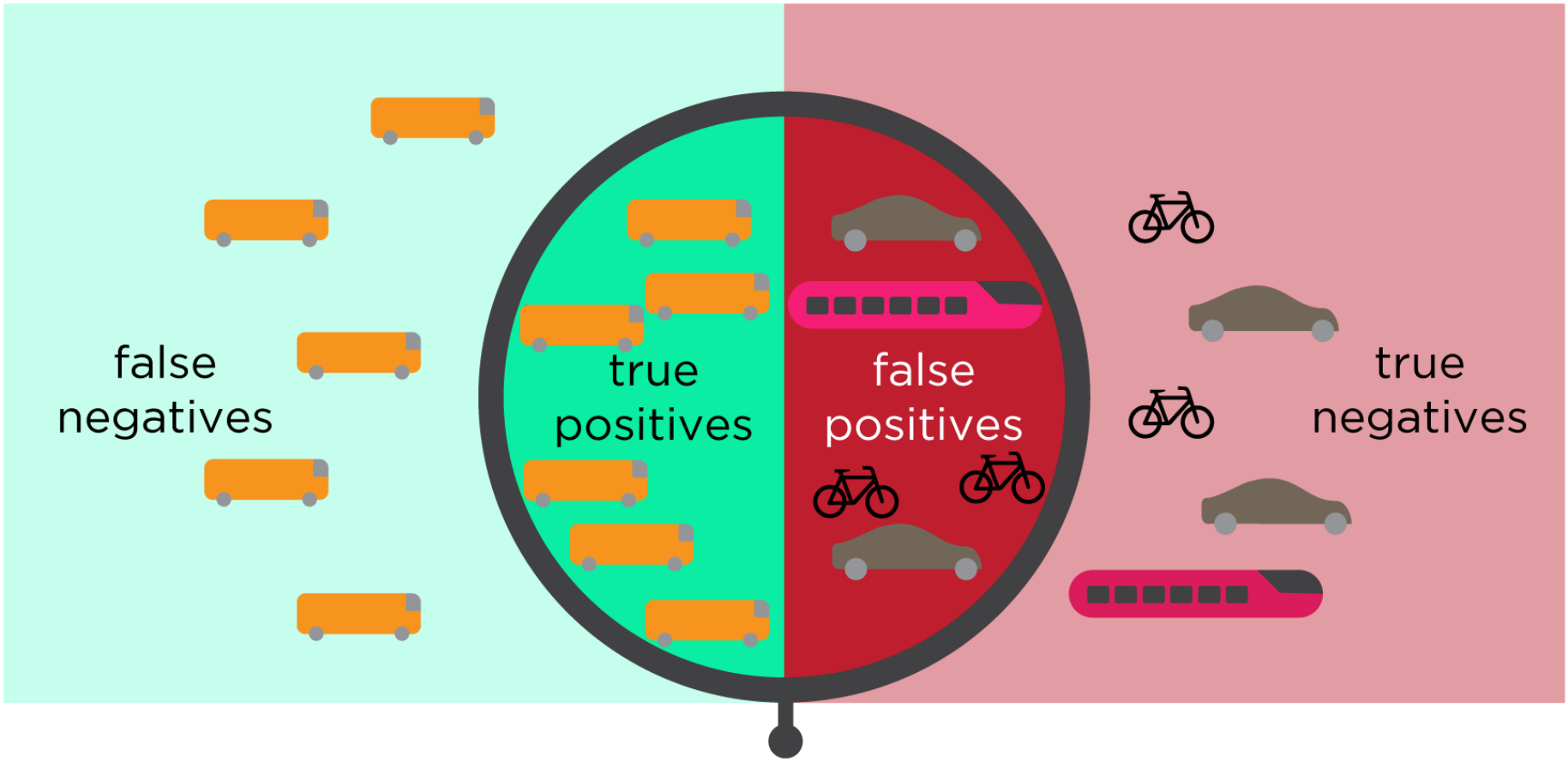


Machine Learning

PROJECT
1

Recall - Completeness
how many relevant items are selected?

relevant items



Eksempel: Transportadfærd

- F1 = “gennemsnit” af precision og recall
- Rugbrødsmotor vs. ikke-rugbrød: F1 = 0.89
- Lige nu:
  - Ekstremt vejr og transport - Klimaforandringer og adfærd
  - Real-tid app til cykling m Københavns Kommune

Bjerre-Nielsen, Minor, Sapiezynskic, Lassen, Lehmann. 2018. Wi-Finder: Urban Transportation Sensing Using Crowdsourced Wi-Fi and Contextual Clues. Manuskript, Kraks Fond for Byforskning

(mere kompliceret) eksempel: Den offentlige samtale

- Hvordan udvikler 'stemningen' i Danmark sig?
- Kan man måle sammenhængskraft over tid?
- Samarbejde: SODAS + Kraka
- Data: 45 mio. opslag fra 153K forskellige Facebook-sider med 300 mio kommentarer fra mere end 3,5 mio danskere 2008-17
- Idé: måle åben debat vs grøftegravning på SoMe
- Fx: Hvad betyder flygtningekrisen?

Bang Carlsen, Klemmensen, Lassen, Ralund. 2018. In Your Face: Ubiquitous Politics and the Polarization of Social Networks. Manuskript, SODAS og Kraka

(mere kompliceret) eksempel: Den offentlige samtale

- egentlig kvalitativ metode: forsker vurderer 'tone' (fx imødekommende, aggressiv, hånende, indifferent) i indlæg i FB-debat
  - men der er > 300 mio kommentarer …
- Koder fx 60,000 kommentarer i hånden -> træner model der relaterer kombinationer af ord til tone
- Kører model på hele datasættet …

Bang Carlsen, Klemmensen, Lassen, Ralund. 2018. Do Political Events Polarize Social Networks? Manuskript, SODAS og Kraka

Fagre nye verden I

- “Gladsaxe-modellen”
- Predikter udsatte børn vha “registersamkøring”, fx arbejdsmarkedstilknytning, tandlægebesøg etc.
  - Offentlig administration kræver tilladelse
  - Forskning: Kan lave model/algoritme t kommunalt plug-in
- Gladsaxe: “Målet helliger midlet.” Datadagsorden får praktikere til at tænke over hvad man kan gøre med data
- Problemer: Politik, Etik

Fagre nye verden II

- “Gladsaxe-modellen”
- Hvor meget i en sådan risikomodel foregår allerede?
  - indberetninger fra skolelærere, sociale myndigheder
- Algoritmer
  - Positivt: horisontal lighed, alle underlægges samme model
  - Negativt: Neurale netværk, AI uigennemskuelige
- DK: skøn vs. regel

Fagre nye verden III

- "Gladsaxe-modellen"
- Er algoritmer biased?
  - Hvis model trænes på biased data (eks: race i USA) kan prediktioner være biased imod særlige karakteristika
  - Men: Ex fra USA om beslutning om varetægtsfængsling finder at algoritmiske beslutninger reducerer kriminalitet, antal indsatte - *og bias mod minoriteter*
  - Kleinberg et al. QJE 2018: kombination af SAMF (selektion, counterfactuals) og ML nødvendig

Uddannelse

- *Social Data Science* (> 350 studerende)
    - 2015-6: Sebastian Barfort, David; 2017-8: Andreas, David, Snorre Ralund
- *Topics in Social Data Science*
    - 2018: Andreas, Snorre, Ulf Aslak
- ML som del af *Advanced Microeconometrics*
- Specialer: Mærsk, Danske Bank, Finansiel Stabilitet, Chr. Hansen, Zetland, Kbh Politi
- Andre steder (eksempler)
    - Harvard 2018: *The Econometrics of Machine Learning (and other `Big Data' Techniques)*
    - Coursera
    - MIT 2018- uddannelse i *Computer Science and Economics*

Uddannelse

- *Social Data Science* (
  - 2015-6: Sebastian e Ralund
- *Topics in Social Data*
  - 2018: Andreas, Sr
- ML som del af *Advan*
- Specialer: Mærsk, Da etland, Kbh Politi
- Andre steder (eksemp
  - Harvard 2018: *The r `Big Data' Techniques)*
  - Coursera
  - MIT 2018- uddanr

AVISEN DK

Synes godt om

Folketinget er fredag blevet ramt af et hacker-angreb.

Det bekræfter Finn Tørngren Sørensen, presseansvarlig i Folketinget, over for Avisen. dk.

Siden fredag formiddag har man fået beskeden "Denne webside er ikke tilgængelig", hvis man har forsøgt at komme ind på Folketingets hjemmeside, ft.dk.

- Det er rigtigt, at der er lukket for den eksterne adgang til Folketingets hjemmeside. Vi er under et såkaldt 'Denial of service"-angreb, og det har vi været siden klokken 10 i formiddags, siger Finn Tørngren Sørensen til Avisen.dk og fortsætter:

- Det fungerer på den måde, at vi får så mange opkald til vores hjemmeside, at systemet bliver overbelastet. Derfor har vi måttet lukke ned for adgangen.

Folketinget har endnu ikke noget overblik over, hvem der står bag hacker-angrebet, eller hvornår hjemmesiden kan komme op at køre igen.

Bottom line

- sociologi, økonomi, psykologi: veletablerede teorirammer til at forstå adfærd -> selektion, endogenitet, …
- Mere generelt: bud på mekanismer
- Data science: flere ting i værktøjskassen. Tillader
  - test af nye, endnu ikke udviklede sammenhænge
  - test af etablerede, men ikke empirisk undersøgte, sammenhænge
- Data science: gode/bedre bud på prediktion, vigtigt til nye/flere variable, tættere på hvad vi vil måle

Bottom line

- SAMF-forskere gode bud på folk som kan forstå og bruge data science
- Erfaringer fra samarbejder m fysikere, ingeniører:
  - Kræver at begge sider investerer i at forstå de andres metoder
  - Hvis SAMFere ikke er med kører de andre bare videre - uden os
- SODAS: hjælper SAMFere med at hjælpe sig selv …

Litteratur

- Big Data and Social Science: A Practical Guide to Methods and Tools
- Kleinberg, J., Lakkaraju, H., Leskovic, J., Ludwig, J., & Mullainathan, S. “Human Decisions and Machine Predictions.” NBER Working PaperAbstract w23180.pdf, forthcoming *QJE*.
- Mullainathan and Spiess. “Machine Learning: An Applied Econometric Approach”. *J. of Ec. Perspectives* 2017.
- Social Data Science, UCPH. Available at https://abjer.github.io/sds/

Tak!

Synspunkter og kommentarer til
ddl@econ.ku.dk

Slides på
https://daviddlassen.github.io

Mere om SODAS på
http://sodas.ku.dk