Social Data Science

David Dreyer Lassen

UCPH ECON

Nov 4, 2015

In God we trust,
all others must bring data

W. Edwards Dewing

Today:
Big Data in Economics

But first: 3 slides on strategic data management and production

Strategic data management and production

- People / firms / governments do not always provide truthful and/or complete data
- Example: No penalty for lying in surveys – but no reason not to either
- Political reasons for obscuring or inventing data: Greece in EU, Chinese economy
- Firms: Proprietary info, competition reasons, fooling customers and regulators (VW)

Strategic data management and production

- Individual demand for privacy (We return to this)
  - Could be instrumental:
    - lack of privacy decreases consumer surplus by better estimate of reservation price (e.g. Steering: Mac vs PC when ordering online)
    - Concerns about political issues
  - Or an objective in itself: Privacy as a political goal

Social desirability bias I

- Key concern in surveys, but more general problem:
  What if people answer so as to conform with general notions of what's desirable?
  - Examples: Won't admit to not voting or having sexually transmitted diseases, exaggerates income
  - Reports buying healthy food vs unhealthy food
  - Important for asking/assessing sensitive questions

Social desirability bias II

- Why?
- Distinguish
  a) self-deception
  b) impression management
- Example: Scrape data from dating websites and link (hypothetically) to income data
  – Is there a correlation between beauty and income? (Yes, but not from such data)
  – Bias could be both (a) and (b)

Today:
Big Data in Economics

David Dreyer Lassen

UCPH ECON

September 24, 2015

No agreed upon definition what Big Data is

- Large N?
- High frequency / much detail?
- Many different measurements?
- Based on what people do (‘honest signals’)
  - ctr surveys
  - Not always honest
- Different to different people/traditions
- To Americans, Danish register data is big data

Administrative data

- Denmark, Norway, Sweden
  - Population-wide
  - Ex: Know population ‘by pressing Enter’
    - Most other countries: census (counting people), surveys, rough approximations
  - In DK, built on Central Person Registry number
  - System constructed for source taxation in 1960s, now used as ubiquitous identifier
- Why do some countries have CPR-like systems and some not?

Administrative data

- Pros
  - Often full population
  - In DK: third party reported -> no reporting bias, no survey bias
  - Very detailed, no survey fatigue
  - Often very precise, since used for admin purposes
- Cons
  - No soft data (attitudes, expectations); can be linked to surveys
  - Privacy concerns
  - Restricted to what is collected for admin reasons, both type and frequency (e.g. annual)

Administrative data

- Lots of work in Danish econ utilizes register data
  - Taxation
  - Education
  - Health
  - Financial decisions
  - Labor market

- Combined with
  - Personality measures
  - Attitudes/political prefs from surveys
  - Expectations from surveys
  - Biological data (neuro-measures, genetics)
  - Data from experiments

‘Big data’

- Pros
  - Often based on real decisions (as admin data), but more detail, e.g. auctions
  - High frequency (e.g. wifi), high granularity -> almost large N ethnographic data
  - Sometimes cheap/free

- Cons
  - No established protocol for collection
  - Sometimes dubious quality, selection issues (both known/unknown)
  - Start-up costs
  - Even more privacy concerns
  - Corporate gatekeepers -> bias in access

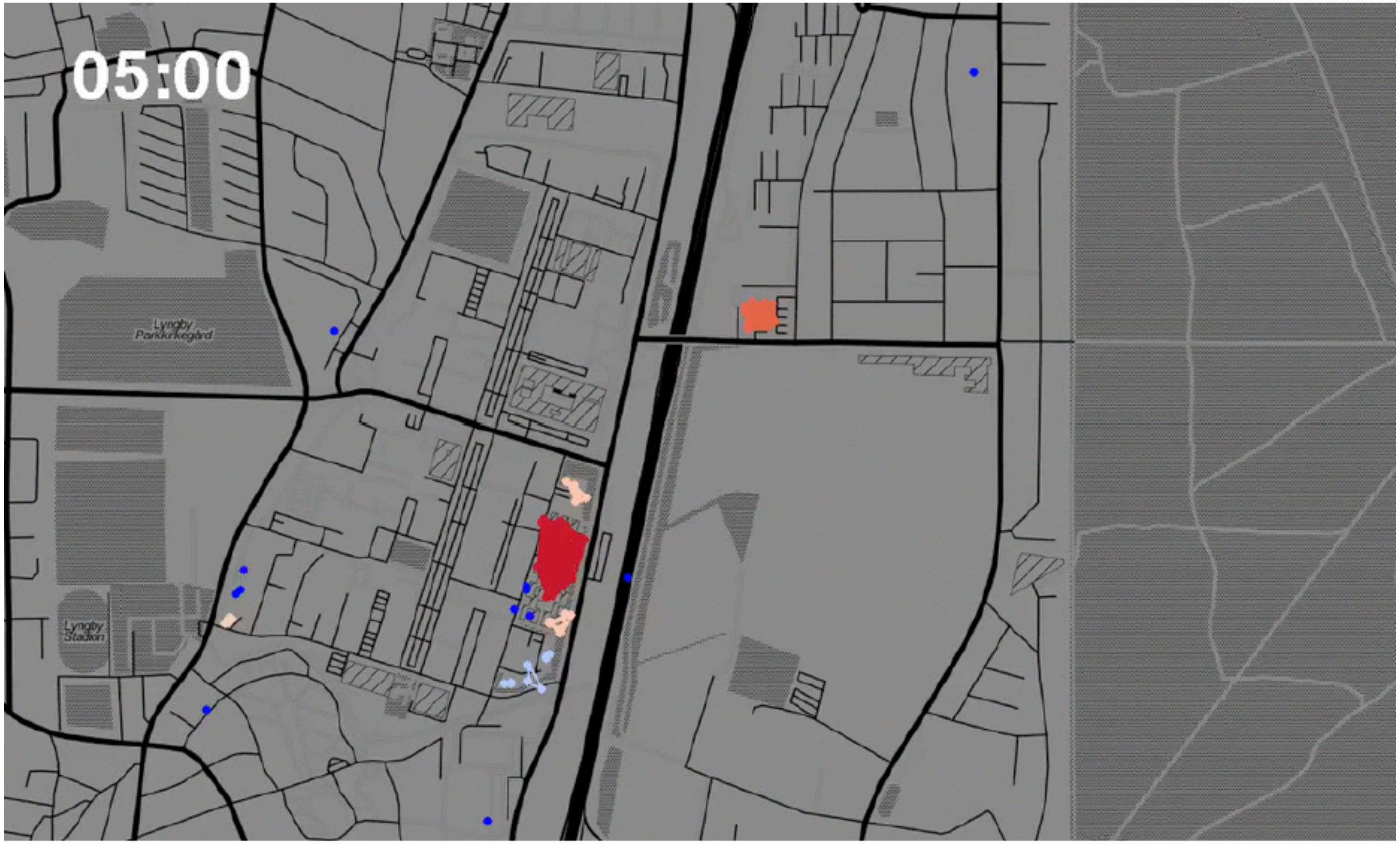
Characteristics of ‘big data’

- Structured (row/column-style) vs. unstructured (images/sound)
- Temporally referenced (date, time, frequency)
- Geographically referenced (wifi, bluetooth, Google)
- Person identifiable (identify vs. distinguish individuals vs. not distinguish individuals)
  - Separate medium (e.g. phone) from owner

Example: Social Fabric

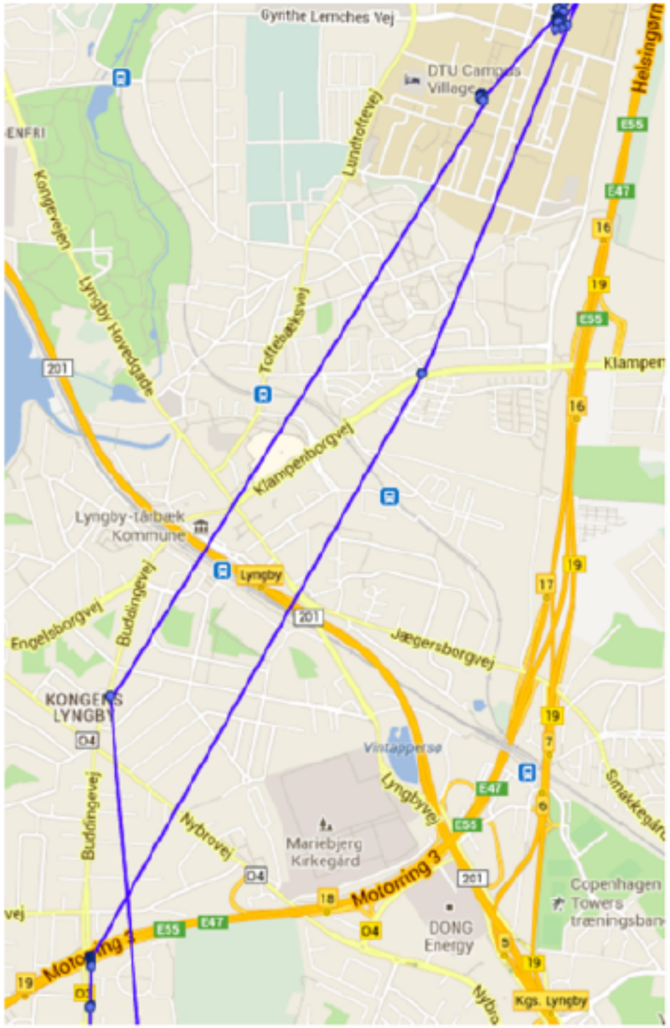
- Large-scale (N=1000) big data project
- Handed out smart phones to DTU freshmen
- Collected phone, SMS/text/email (not content), GPS, wifi, bluetooth data
- -> Where, when, with whom
- -> social networks

Example: Social Fabric

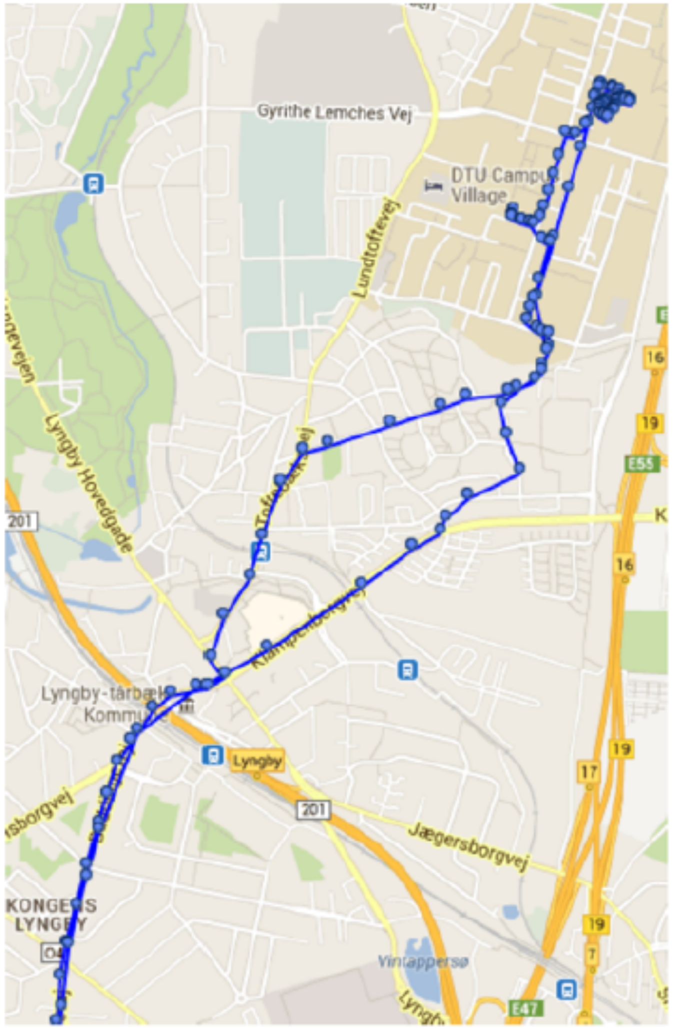


Phone locations 0500h Monday morning -> can predict where people at given time with 85% accuracy

Example: Social Fabric

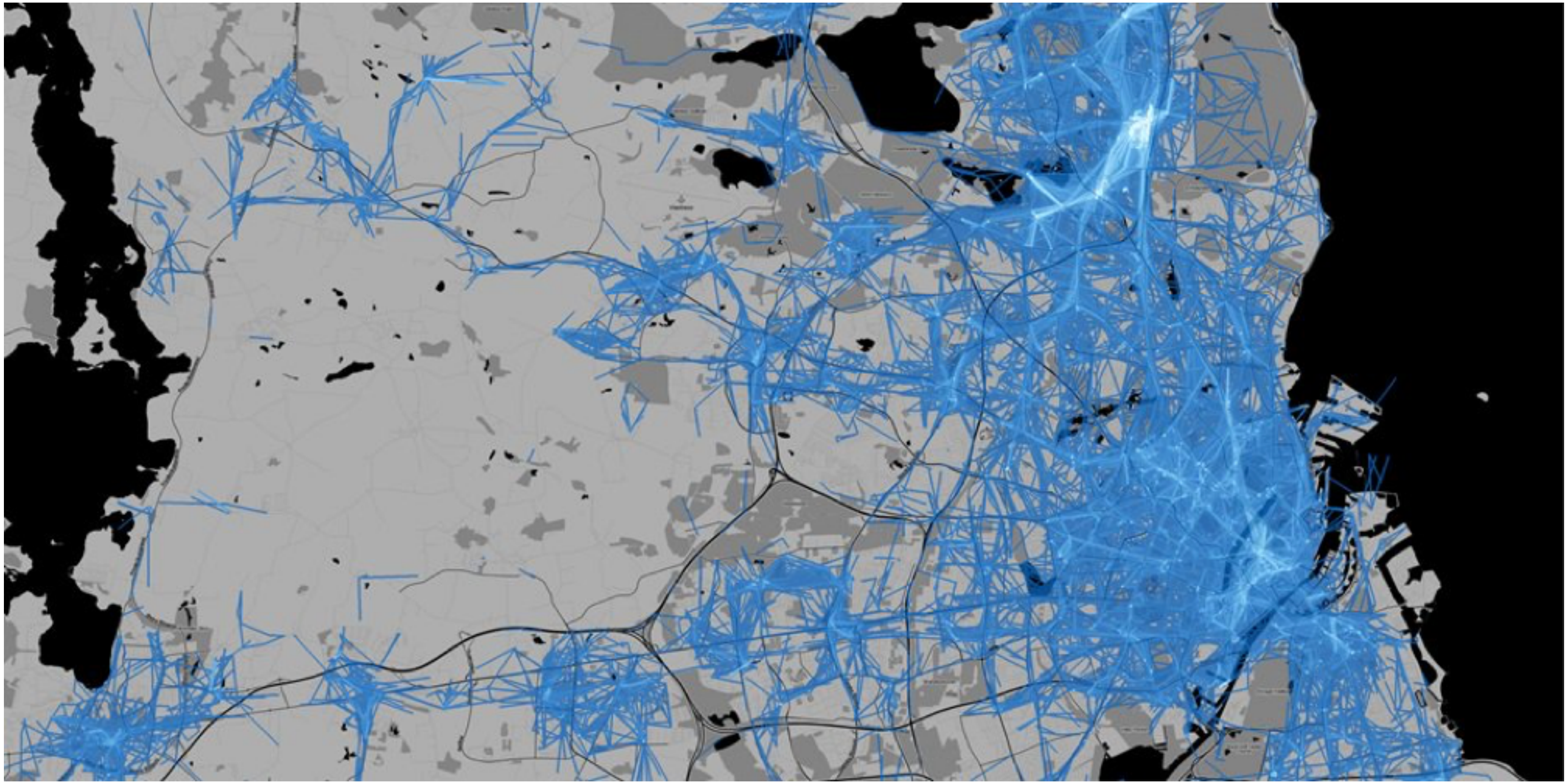


10 min GPS



wifi

Example: Social Fabric



Example: peer effects in education economics

- Students allocated to study and social groups, called vector groups (randomly)
- Are there peer effects, i.e. are students' grades/health behavior/study behavior affected by the group?
- Literature: sometimes yes, sometimes no; very heterogeneous
- Why? Perhaps being allocated to group is not = to actually meeting / using group

Example: peer effects

- Think of allocation to group as intention to treat (similar to offering treatment)
- Interesting example: Carrell et al, ECMA 2013. Small groups, yes peer effects; large groups: no peer effects – WHY?
- Use phone to measure frequency of group members being together physically, measured by bluetooth
- Three parts: (i) yes they are more together; (ii) more together => work better together; (iii) peer effects?

Broader issue: Who meets, and how close are they?

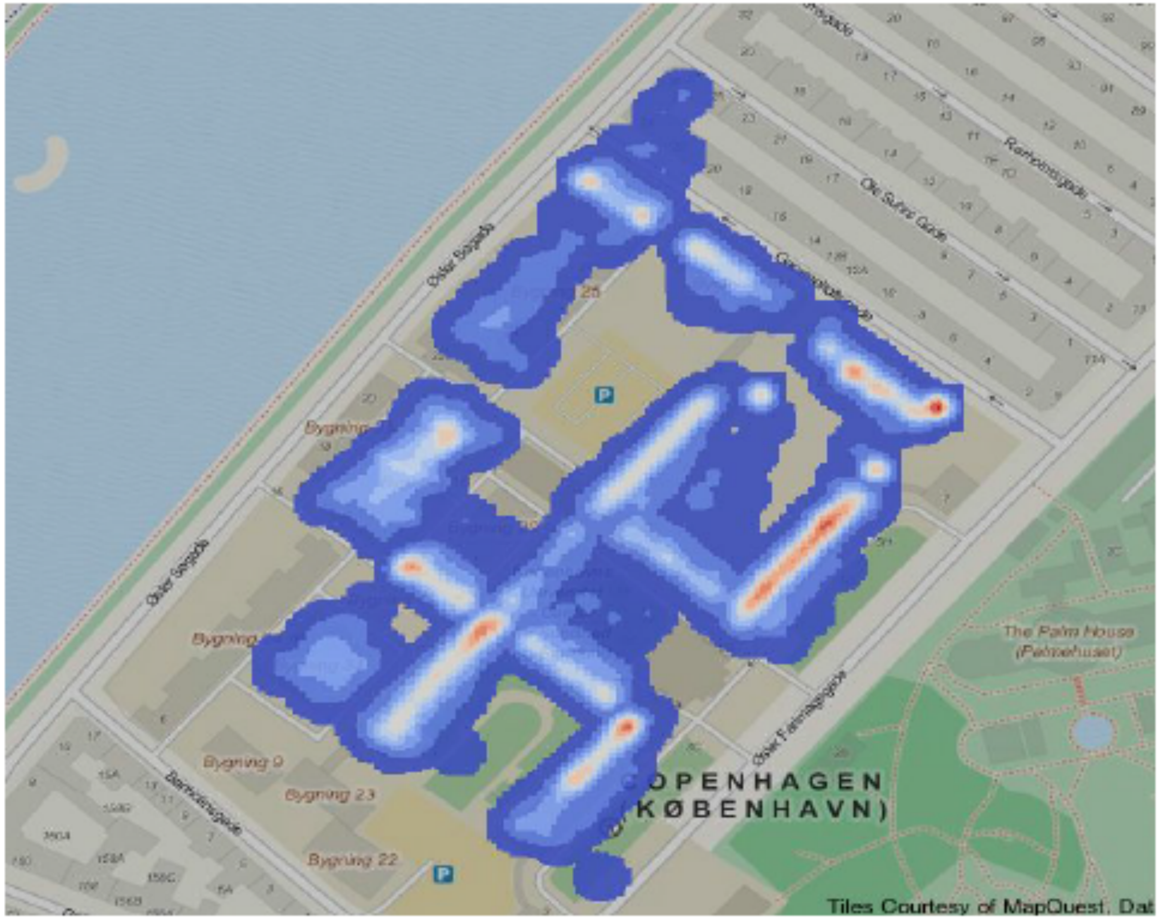
- (This is Kristoffer’s Master’s thesis)
- Again: use bluetooth signals to measure meetings (duration, participants)
- Analyzes 3.1 mio meetings over two months
- Some results:
  - Women/women pairs -> closer
  - Facebook friends -> closer
  - Same study -> closer
  - Difference in beauty -> further apart
  - One overweight, one not -> further apart
- People who stand very close have fewer friends

Example: why phone data

- Phones as **sociometers**
- Many/most people carry phone with them all the time
- Would be IMPOSSIBLE to have people report in detail for every 10 min every day for a year

- For this project: tailored software, but realized that many apps collect detailed wifi-data without telling

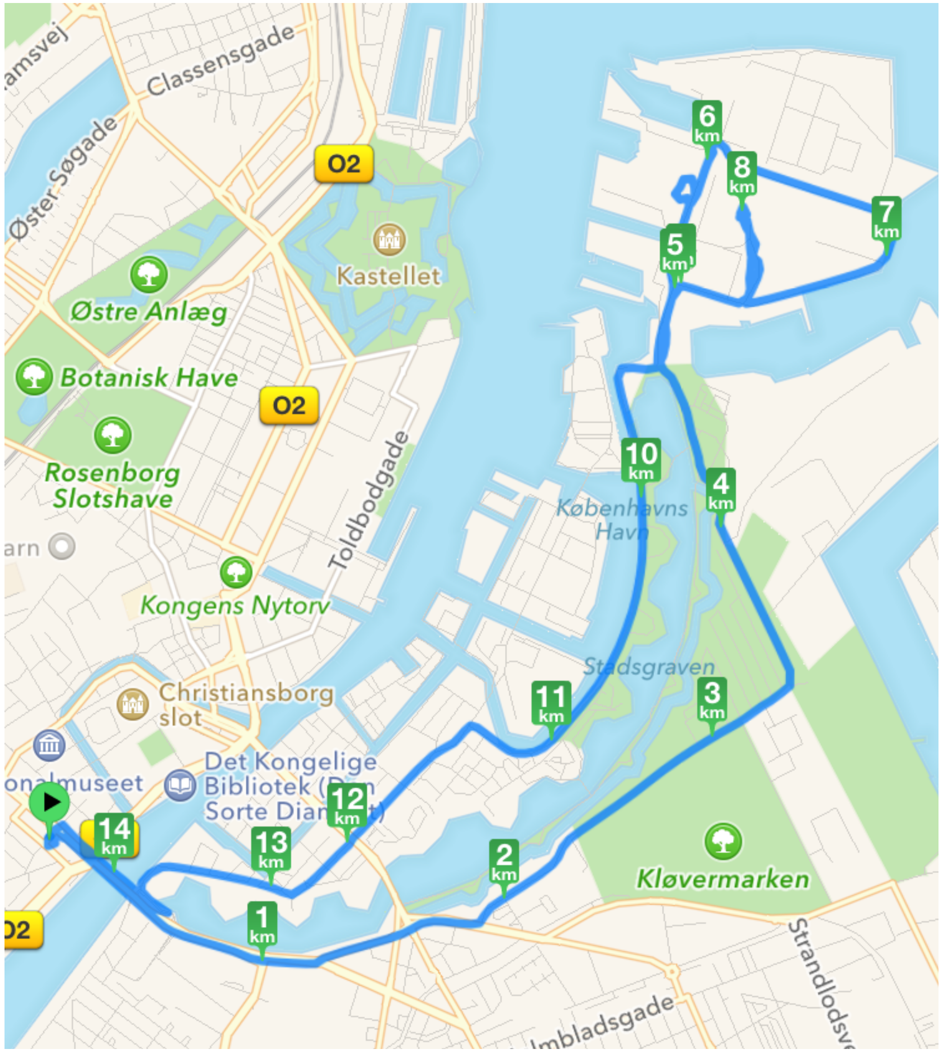
Example: CSS

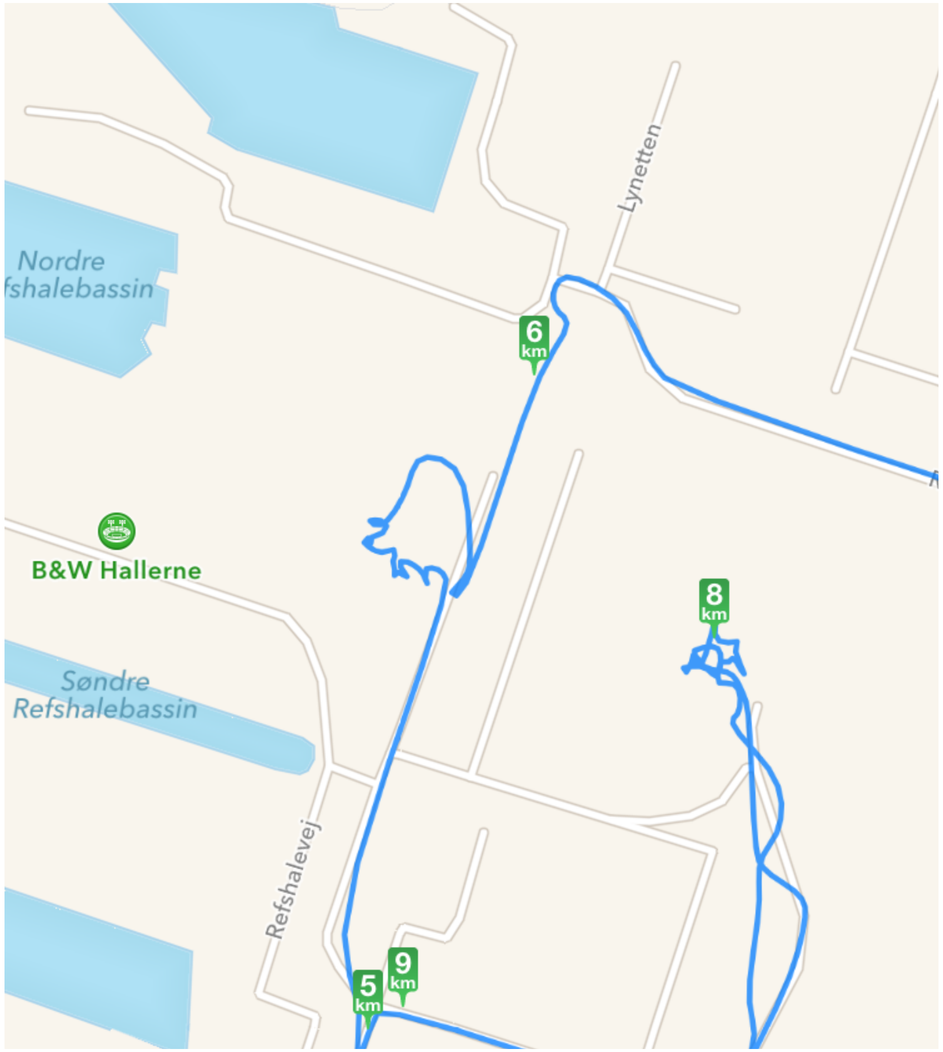


Tiles Courtesy of MapQuest, Dat

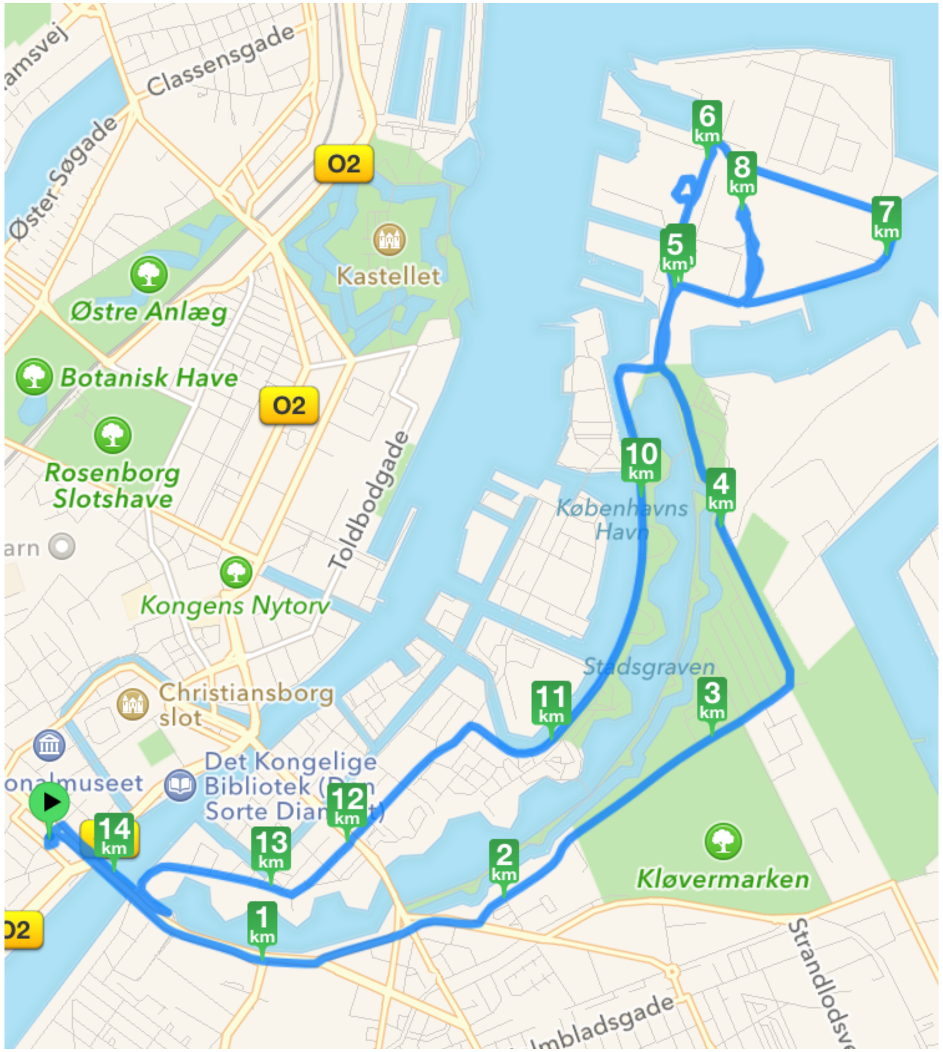
Heatmap of people with mobile devices on CSS (anonymous)

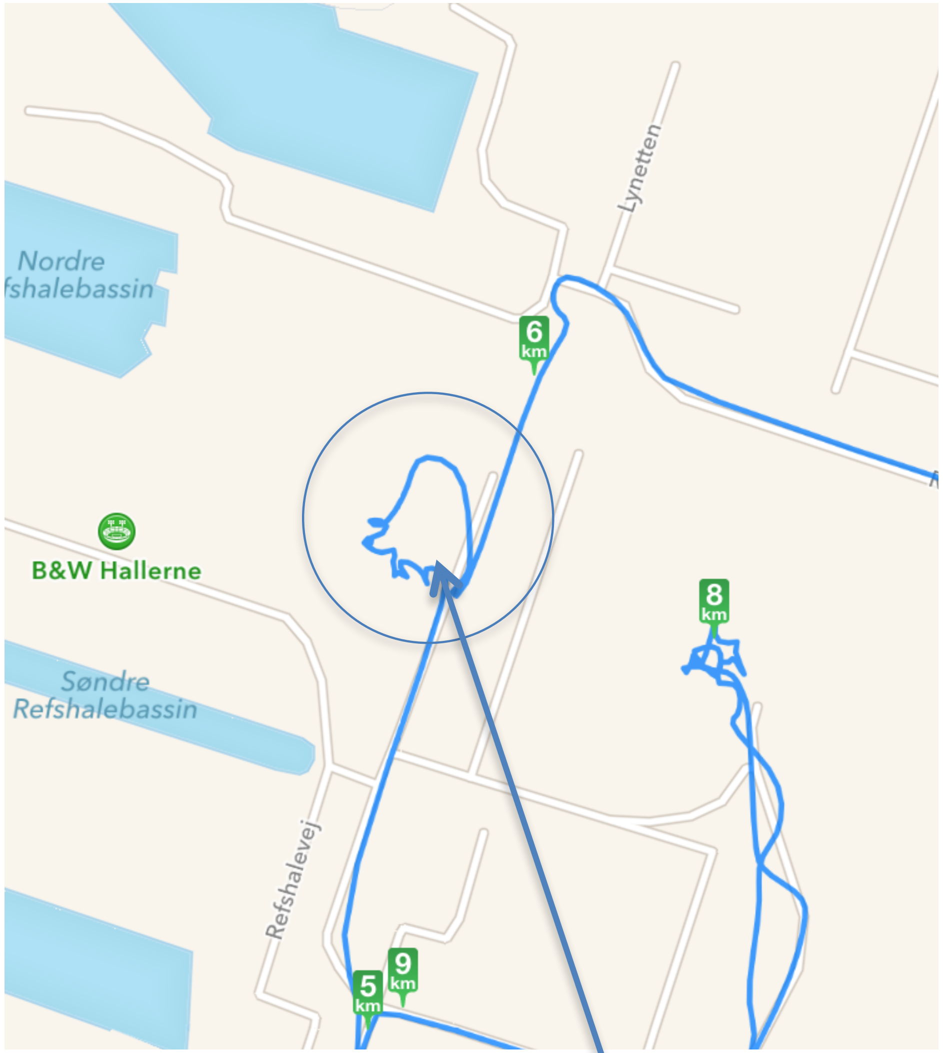
Example: David on Saturday





Example: David on Saturday





Flea market

Example: how to measure consumer spending

- Economically important:
  - Indicator of health of economy
  - Important for understanding individual responses to policy
  - d.o. to economic shocks
  - Important for consumer prices -> inflation -> adjustments of wages and transfers
  - In developing countries: important for estimates of poverty, inequality

Example: consumer spending

- Traditional methods:
  - Consumer expenditure surveys (DK: forbrugsundersøgelsen)
  - Diary or scanner
  - Errors, selection
- Economists wanted access to individual spending data from Dankort for a long time
  - No luck
- Recently, Statistics Denmark got access to COOP-card data to measure inflation
  - To be made public soon, pretty good fit with existing measures (and much faster)
  - Nice idea, incentive compatible
  - Indep of payment type
  - But selection

Example: consumer spending

- Attempts in developing economics
  - Use smart phones as scanner or means of payment
  - what can we infer about individuals from smart phone use (dedicated users)
  - Selection into who has smart phones
  - But should be seen against other ways of collecting data

- Qs:
  - How can we use smart phones to infer spending better?
  - What kinds of economically interesting data can we collect via smartphones?

Statistical analysis of Big Data

- Many observations: what does statistical significance mean?
  - And what is practical relevance? Size effects
- Multiple testing problems? If big data generates many variables, why not run through them all to see what is significant?
  - Correct standard errors
- In some cases, ‘eyeball econometrics’ can be difficult
  - Need systematic approach

Statistical/machine learning

- Suppose you have no or very little theory to guide you
- OLS is not only linear, but also presumes some idea of what actually goes in there and how
- Varian’s Titanic example: who survived the Titanic
  - Two variables: Class and age
  - Researcher decide / guess vs. data analysis yield most likely (decision tree, but lots more complicated -> Sebastian, later)
  - Einav, Levin: Econ should consider machine learning

Statistical analysis of Big Data

- But what if you have theory (or think you have) – e.g.combine econometrics and machine learning
- Goes back to old debate in economics
  - Milton Friedman (1953): judge a model by its predictions, not its assumptions
  - Machine learning made for prediction not for hypothesis testing and theory (in)validation