Social Data Science

David Dreyer Lassen
UCPH ECON
August 17, 2017

In God we trust,
all others must bring data

W. Edwards Dewing

Today:
Privacy and Ethics in Big Data Production and Use

Why privacy?

- Privacy for its own good – a principle of privacy
- Privacy to preserve informational rents
  - Consumers, firms
- Privacy and politics

Why privacy?

- Privacy for its own good – a principle of privacy
  - May simply value privacy in itself
  - But: public goods problems
    - Example: medical research. Share existing info on medical history, no cost to individuals. Some will not contribute, citing privacy concerns – but benefits of research accrue to everybody
    - DK: no consent necessary for register studies or re-use of data
    - Similar: Privacy for social science research, or monitoring in public places

Why privacy?

- Privacy to preserve informational rents
  - Consumers: willingness to pay (WTP), characteristics, and behavior often private information
    - Willingness to pay: 1st class vs. 2nd class
    - Characteristics: Taste, Genetics, Personality
    - Behavior: e.g. driving and insurance, physical activity
    - Value of time / search costs
    - Example: Internet steering
  - Firms: Intellectual Property Rights, strategy
    - Industrial espionage major problem
    - LinkedIn-story; Firms where data is only asset

Why privacy?

- Privacy and politics
  - Authorities may not register party identification
    - Originally for freedom of political expression but also: majority in city council could pay out cash assistance / kontanthjælp based on, say, union membership
  - These days: Privacy as a political platform

Danish law of personal data "Persondataloven"

- Lays down the rules on
  - All electronic use of personal data
  - Data in registers
- In general: strict rules governing authorities' and firms use of data / information
- Sensitive data singled out:
- Sensitive:
  - Political views
  - Philosophical views
  - Sexual preference
  - Ethnicity/Race
  - Union membership
  - Health
  - Serious social problems

Process example

- Combine survey data on economic expectations with administrative data on taxable income
- Combines two sets of individual data
- Requires permission from Danish Data Authority

- What about comments / likes from Facebook?
- What about usernamed rating on website?
- What about data on houseprices – or data on owners of houses?

From 2018: Replace Danish law with The EU data protection directive

- Link: http://ec.europa.eu/justice/data-protection/
- ”The objective of this new set of rules is to give citizens back control over of their personal data, and to simplify the regulatory environment for business.”

Process example

- Combine survey data on economic expectations with administrative data on taxable income
- Combines two sets of individual data
- Requires permission from Danish Data Authority

- What about comments / likes from Facebook?
- What about usernamed rating on website?
- What about data on houseprices – or data on owners of houses?

Example:What can we know from Facebook-likes?

- Quite a lot
- "Private traits and attributes are predictable from digital records of human behavior" Kosinski et al. PNAS 2013.
- 58,000 volunteers gave access to Facebook-likes, demographic info + took psychometric test
- Results: Facebook-likes -> stat learning model that correctly predicts
  - Sexual orientation 88%
  - Afri-Am vs Causcasian 95%
  - Dem vs. Rep 85 %
- As good as personality test for traits
- Implications for privacy and online behavior?

Apropos of Facebook

- EU DPD posits a right to data portability
- This means: easier to move personal data from one provider to another, incl social networks
- Compare: Phone companies
- Interesting regulatory consequences
- Old days: Phone companies owned phone number, large costs if switching. Now, individually owned
- Could one own one's social graph?

In DK: No need for informed consent on processed data

- § 5. Oplysninger skal behandles i overensstemmelse med god databehandlingsskik.
  Stk. 2. Indsamling af oplysninger skal ske til udtrykkeligt angivne og saglige formål, og senere behandling må ikke være uforenelig med disse formål. Senere behandling af oplysninger, der alene sker i historisk, statistisk eller videnskabeligt øjemed, anses ikke for uforenelig med de formål, hvortil oplysningerne er indsamlet.

- § 5.

  Stk. 2. Collection of information can happen only for learly specified and factual reasons. Subsequent processing must not be in disagreement with the reasons.
  Subsequent processing that happens only for historical, statistical or scientific reasons are not considered in disagreement with the purpose for which data is collected.

Individual data and privacy

- Stat Denmark: Data users cannot present data at the individual level
- Examples
  - Max of the income distribution
  - Median of income distribution
  - Max income in parish

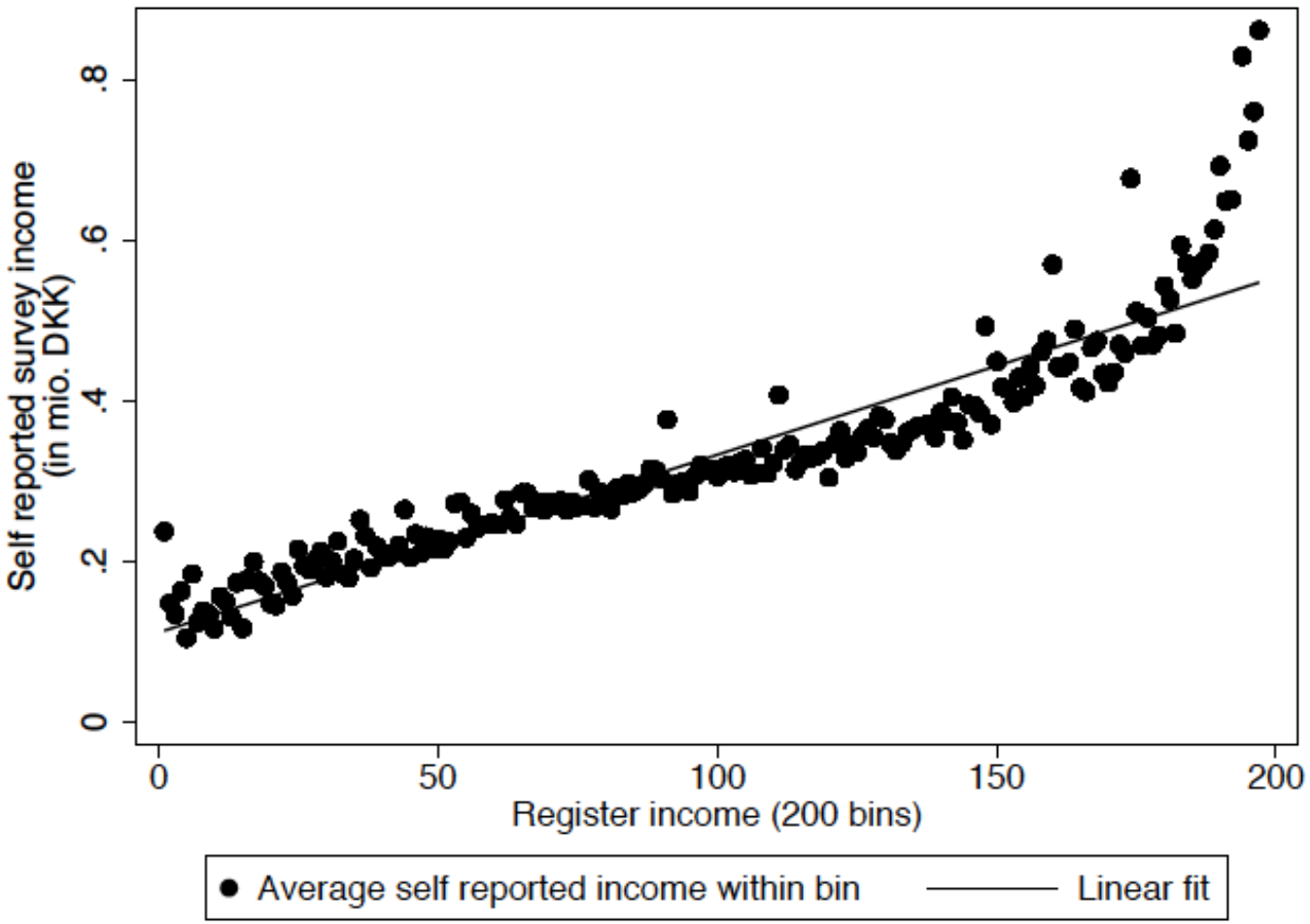
- Well-known examples of re-identification from public data
  - Often in combination with auxiliary data
  - An overview
  - An example based on credit card data

Trade-offs

- Sometimes: Sacrifice accuracy for privacy
- In some cases: no trade-off in analysis, only in presentation

Hariri and Lassen, 2016
Public Opinion Quarterly

Figure 1: Incomes: Register and Survey



Notes. The horisontal axis shows register based income in 200 equal-sized bins ranked from lowest income to highest income. The first bin is defined as the N/200=25 lowest ranked individuals and the figure plots the average register income for this group against their average survey reported income – and continues to do so for the 199 other income bins. The vertical axis has been censored at self reported survey income above 1 mio. DKK. Figure A.1 in the online appendix shows the full sample.

Trade-offs

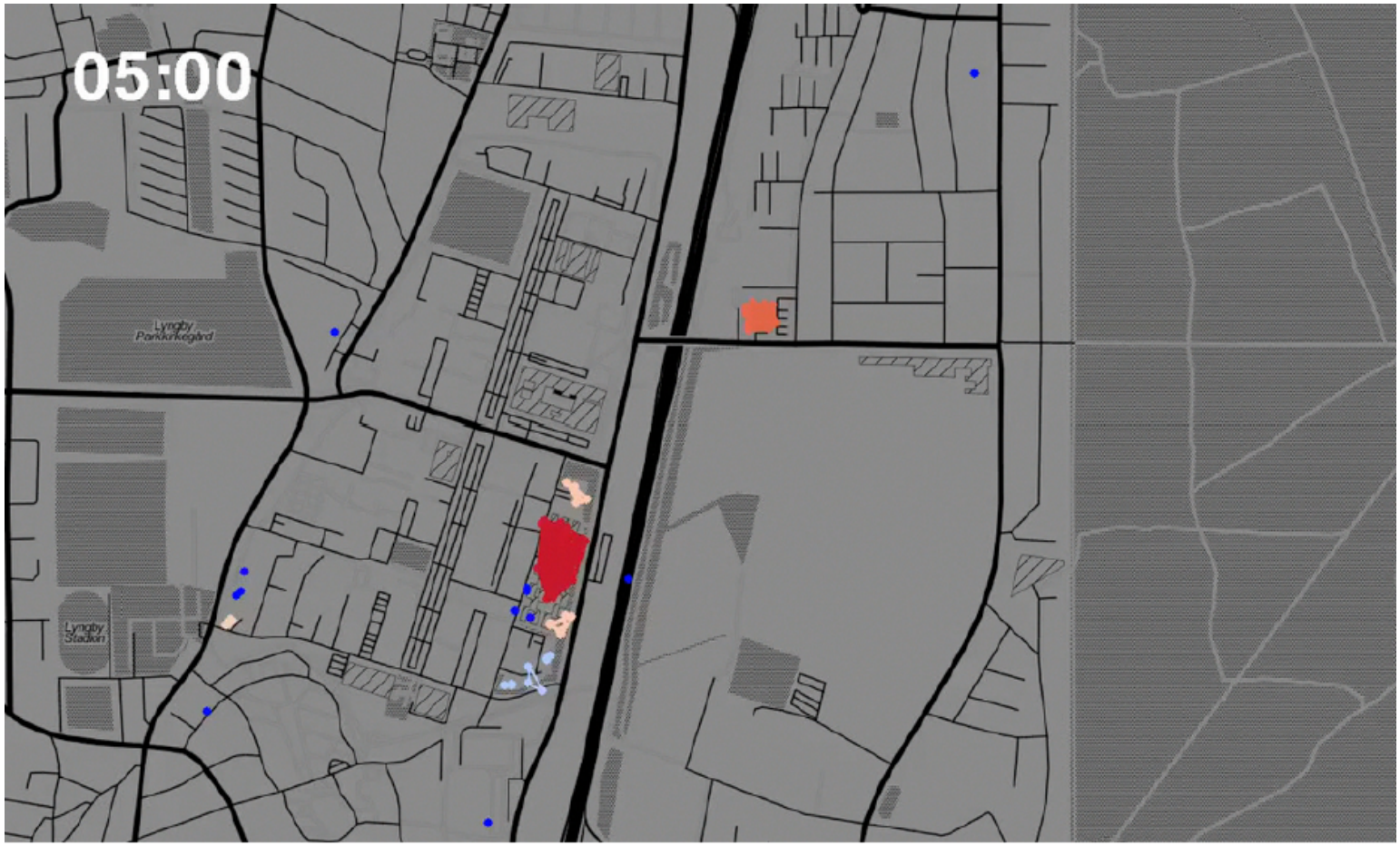
- Sometimes: Sacrifice accuracy for privacy
- In some cases: no trade-off in analysis, only in presentation
- Sometime: only have, say, interval data
- Danish firm data: Stat Denmark does not report figures for industries with very few firms

- New approaches: analysts don't see data, but can make calculations on it
  - May limit *feel* for data
- More general problem: how much info do we get from data under constraint of 'no identifiability'?
  Active research area in computer science

Economic analysis of privacy

- Heffetz and Ligett (Read): Principles for privacy preserving data handling
  - a bit complicated in places
- Active research area
  - Combine with mechanism design
  - Economic theory
  - Combine computer science and economics
- See Acquisti et al. for more on this (if interested)
- Also: behavioral economics aspects + genuine uncertainty: "Even ex post, only few of the consequences of privacy decisions are actually quantifiable; ex ante, fewer yet are."
  - from Acquisti&Grossklags, 2007 "What Can Behavioral Economics Teach Us About Privacy?"

From last time:



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

Ethics of Big data

- ”Web scraping: a journalist’s guide” + ”on the ethics of web scraping and data journalism”
- For journalists, but interesting for us as well
- Neuhaus and Webmoor 2012: ”Agile ethics for massified research and visualization”
- Do read (not econ)!
- Also (google): Zimmer (2010) ”But the data is already public”: on the ethics of research in Facebook.

What is Ethics?

- A systematic approach to moral judgments based on reason, analysis, synthesis and reflection
- Moral standards: Impartial, take precedence over self-interest, universal
- But not *one* set of standards

- Are student or researcher ethics different from personal ethics?

Ethics of Big data

- Ethics in universities often governed by
  - Institutional Reviews Boards (IRBs)
  - Personal ethics or feelings of right and wrong
- The law: the institutional embodiment of ethics

- Denmark: Only formal ethics board for bio-medical research
- no IRBs in economics
- DK-wide in polisci
- Some in psychology
- Sociology??

Key goal of ethical considerations

- Reduce potential risk of participants in research
  - In medicine: benefits vs. harms
  - In social science: typically identifiability/privacy, but could also be stigma or long term consequences in field experiments
- Is informed consent enough?
  - Is consent informed if shrouded in 80 pages of legal click-thru?
  - If photographing people in public places is ok, is noting what they say on Facebook also ok?

Challenges

- But: Not unethical to find correlation btw smoking and lung cancer, even if insurance companies use this to increase premiums for smokers
  - What about correlation between genetic markers and, say, chronic diseases, increased mortality risk?
- ethics is not about preventing stuff from being done – but reasonable balance between costs and benefits (ex: hidden camera/mike : not ok for mundane things, but maybe ok if benefits are huge; random drug screening of emplyees may violate privacy, but ok if job involves public safety)

ethical considerations for big data

- What about business ethics?
  - Example: Google Location. Show where friends/family are in real time – but requires consent
  - Are predictive location algorithms ethical?
- Algorithms as ”Weapons of Math Destruction”
  - Insurance based on where you live, your name/ethnicity
  - Entry into university based on prediction of completion?
  - Loan interest rates based on past behavior?

ethical considerations for big data

- Is it ethical to scrape competitors' *likes* on Facebook? Is it illegal?
  - ethics (and law) sometimes used as arguments to stiffle competition. See LinkedIn case
- Can you scrape data and resell? Or repackage?
- Does data collection cause significant costs (time or money) to firms and/or individuals?

Questions for proposed projects

- Do you respect privacy?
- Can single individuals be identified?

- Are there ethical considerations
  - With respect to individuals?
  - With respect to firms?
  - Should you report your research to the Danish Data Protection Agency? See here (in Danish) for exemptions wrt personal data and students' projects.