Social Data Science

David Dreyer Lassen

UCPH ECON

September 24, 2015

In God we trust,
all others must bring data

W. Edwards Dewing

Today:
Empirical design
data generating process
modes of collection
strategic data provision

David Dreyer Lassen

UCPH ECON

September 24, 2015

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data production

Different data for different questions
or
Different questions for different data

Sometimes possible to separate **data collection process** from underlying **data generating process** – and sometimes not

Fundamental difference between what people do and what they say they do

‘cheap talk’ / ‘put your money where your mouth is’ / honest/costly signaling

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data production

What is your question, again?

1. Research question from theory
2. Ideal empirical design
3. Feasible empirical design / collection
4. Results
5. Adjustment of theory/ question/design
6. New results
7. …

A. What data do we have

B. What question can they answer

C. Research question

D. Results

All models are wrong – but some are useful

George Box

Two key goals

1. Forecasting: individual behavior, policy consequences, voting, Champions League, … Data science / machine learning (but also macroeconomics)
2. Hypothesis testing, derived from theory ´Traditional' social science

1. Forecasting

- Example: Bank wants to forecast non-payment on loans (P_d: probability of default)
- Couldn't care less about theory
- Rough "Data Science": try to predict from all available data
- Suppose we find that birth weight predicts default
  - Bank is happy, better fit (defer ethics etc)
  - Policy: does investing in pre-natal care reduce defaults?
- In practice: set of predictors taken from (some) theory, even if casual
- Complications: if customers know that P_d depends on birth weight, would/should they disclose it? What if loans only to disclosers? Would they tell the truth?

2. Hypothesis testing

- Theory (rational choice, sociology, biology, common sense, …) posits effect of X on Y
  A. Selection/type theory: People who are impatient cannot defer immediate pleasures -> smoke and drink while pregnant -> gives birth sooner. If impatient parents -> impatient children (whether by nature or nurture), we have an explanation.
  B. Biological theory: low birth weight affects brain development and neurological wiring for patience.
- If (A), little role for policy; also, both can be true at same time
- How to distinguish: exogenous shock to birthweight, but ethically tricky ...

Goodhart’s law

- Most popular: “When a measure becomes a target, it ceases to be a good measure.”
- What he wrote: “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.”

Case of Google Flu

- Google Flu: web searches for Flu symptoms predicted actual flu cases
- By-product of Google’s main service
- But from 2010, not so well: overestimated actual flu cases, partly as result of autosuggest feature, partly because model was overfitted (we’ll return to that)
- Best predictor: number of cases past week

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data production

Effects of causes
vs.
Causes of effects

Different questions

- Effects of causes: intervention, what is effect of policy X on outcome Y
- Causes of effects: Why does Z occur?

Effects of causes (forward causal questions)

- Narrow questions, sometimes (but not always) policy interventions
  - Effect of tax change on behavior
  - Effect of regulation on risk taking
  - Effect of schooling on earnings
  - Effect of smoking on lung cancer propensity
  - Effect of public health on schooling in Africa
  - …
- Often, but not always, amenable to treatments/ randomization/experimentation

Causes of effects
(reverse causal inference)

- Much harder, but often more interesting
  - Why do some people smoke?
  - What are the causes of democratization?
  - Why do some people pursue a PhD why others drop out after primary school?
  - Why did Greece (almost) go bankrupt?
- Tensions with ”effects of causes” – search for causes sometimes derided as ‘party chatter’

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data production

Data generating process

What is the **data generating process**?

Observational: endogenous decisions, researcher passive collector of data

Randomization: treatment-control

(Some) exogeneity: policy interventions, sometimes with comparisons, researchers sometimes involved

Important: more data does not give better result/ more precision if estimator is biased

Randomized experiments

- Distinguish
  - Lab experiments: traditionally computer-based in econ, but also eye tracking/brain images (fMRI)/physiological
  - Survey experiments: assign survey respondents to different frames/treatments/primings, e.g. have SocDems and Liberals say same thing and look at support
  - Field experiments: experimental control in the real world, e.g. banks charging different rates to learn about mobility of customers; interventions against teacher absenteeism in India; …)

Randomized experiments

- Distinguish
  - Natural experiments (weather induced: effects of poverty on violence, randomization of names on election ballots, …)
  - Quasi-experiments (effects of change in policy; effect of tax reform on tax planning; effect of immigrant allocation on crime)
- Throughout: exogenous (outside of the individual) change

Randomized experiments

- Large, important current debate in (development) economics
- CofE: what are effects of penalties on teachers' absence in Indian village schools – evidence from randomized experiments
- **Randomly** selected teachers get harsh penalty for no-shows -> difference in absenteeism **causal effect** of penalty
- (Broader EofC Q: why is education sector in rural India so inefficient?)

Randomized experiments

- Strong on internal validity: from randomization **any** effect on absenteeism is from harsher penalties; good for testing theory
- Weak(er) on external validity – would effect be similar in Africa? Would effect from lab work outside lab? Why, why not?
- (compare: medicine works in similar ways across locations)

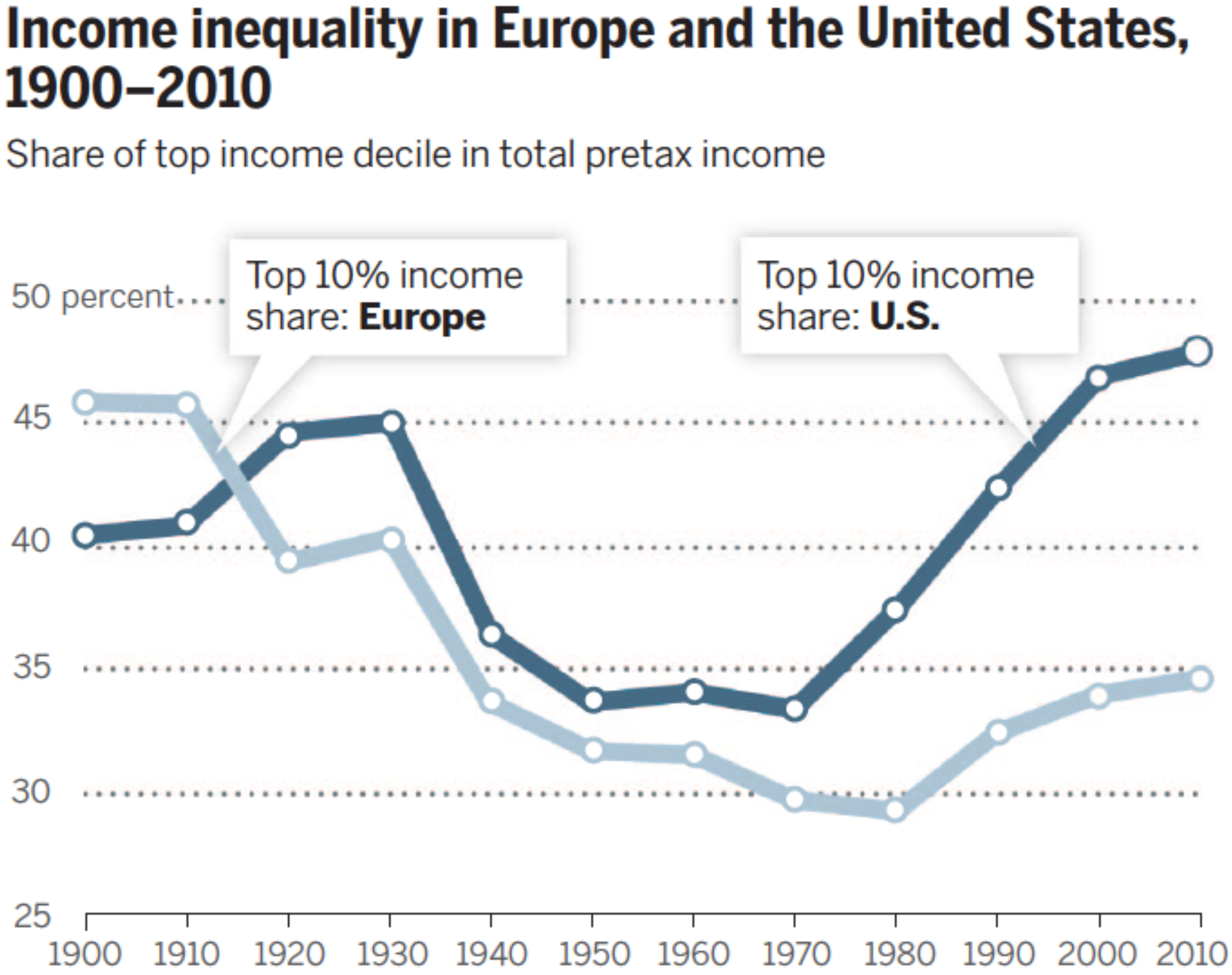
Randomized experiments

- Challenges
  - Limits to what can be studied by experimentation (ethics; law; feasibility)
  - Funding (field experiments expensive, survey exp less so)
  - Often **participation constraint** – voluntary participants' gain >= 0 or no incentive
  - Subjects leave for various (systematic) reasons
  - Large-scale randomization can be hard in field experiments

Observational data

- Generated without experimental or exogenous intervention
- Typically reveals correlations or descriptive patterns that can be interesting in themselves

Example: Inequality



Income inequality in Europe and the United States, 1900–2010

Share of top income decile in total pretax income

Top 10% income share: **Europe**

Top 10% income share: **U.S.**

Source: Piketty and Saez, Science 2014, tax return data

Observational data

- Generated without experimental or exogenous intervention
- Typically reveals correlations or descriptive patterns that can be interesting in themselves
  - Are in themselves silent about causality
  - Theory may be provide structure to learn about causal mechanism under strong assumptions
  - May conflate correlation and causality

Observational data

- Exple: Does being in private schools affect grades
  - Classic: Catholic schools and grades in US
  - Collect attendance and grades -> run regression
- But: suppose some parents are more focused on schooling than others
  - Send kids to private school more
  - More involved in school + homework
- What do higher grades measure?
  - Effect of private school OR effect of involved parents?

Observational data

- What to do?
  - Assign kids/parents randomly to private schools?
- More complicated
  - Waiting-list experiment design: people who sign up reveal themselves as school interested, compare grades between those in program and on waiting list -> much narrower design
  - Modeling (US case): use fact that Catholics are much more likely to choose Catholic schools

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data production

Modes of data collection

- (Ethnographic / participant observer)
- Survey
  - Interview survey (in person), phone survey, internet survey, …
- Administrative data
  - Used for administrative purposes
  - Some countries: census, tax return
  - DK: CPR-registry based
- (Primary collection: texts, counting)
- "Big data": in social sciences typically a by-product of digital information

Modes of data collection

- Note: survey, admin data, big data can all have randomized / exogenous elements or be purely observational
- Often in Lab/field experiments: ask about income, education etc – but may be biased
- Sometimes: combine experimental data with admin or big data (but rare)

Ethnographic

- Pros
  - Attempt to understand situations from participants' perspective
  - Very detailed observations (e.g. dynamics at a meeting: who speaks when, who listens, who nods off and flirts etc)

- Cons
  - Very difficult to generalize (if even the goal)
  - Typically very small n, not for stats
  - Hard to reproduce / replicate

Surveys

- Pros
  - Can be cheap
  - Elicit info on attitudes, beliefs, expectations
  - Necessary when no other means exist
  - Combine with open-ended info
  - Easily anonymized (firms; China)

- Cons
  - Can be expensive
  - Non-random samples, sometimes very much so (paid surveys)
  - Cheap talk
  - Diverse interpretations (e.g. 1-10 scales, Maasai example)
  - Very different quality: interview vs. internet
  - Not full researcher control: Interviewer completions

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)
  - Privacy concerns
  - Restricted to what is collected for admin reasons, both type and frequency

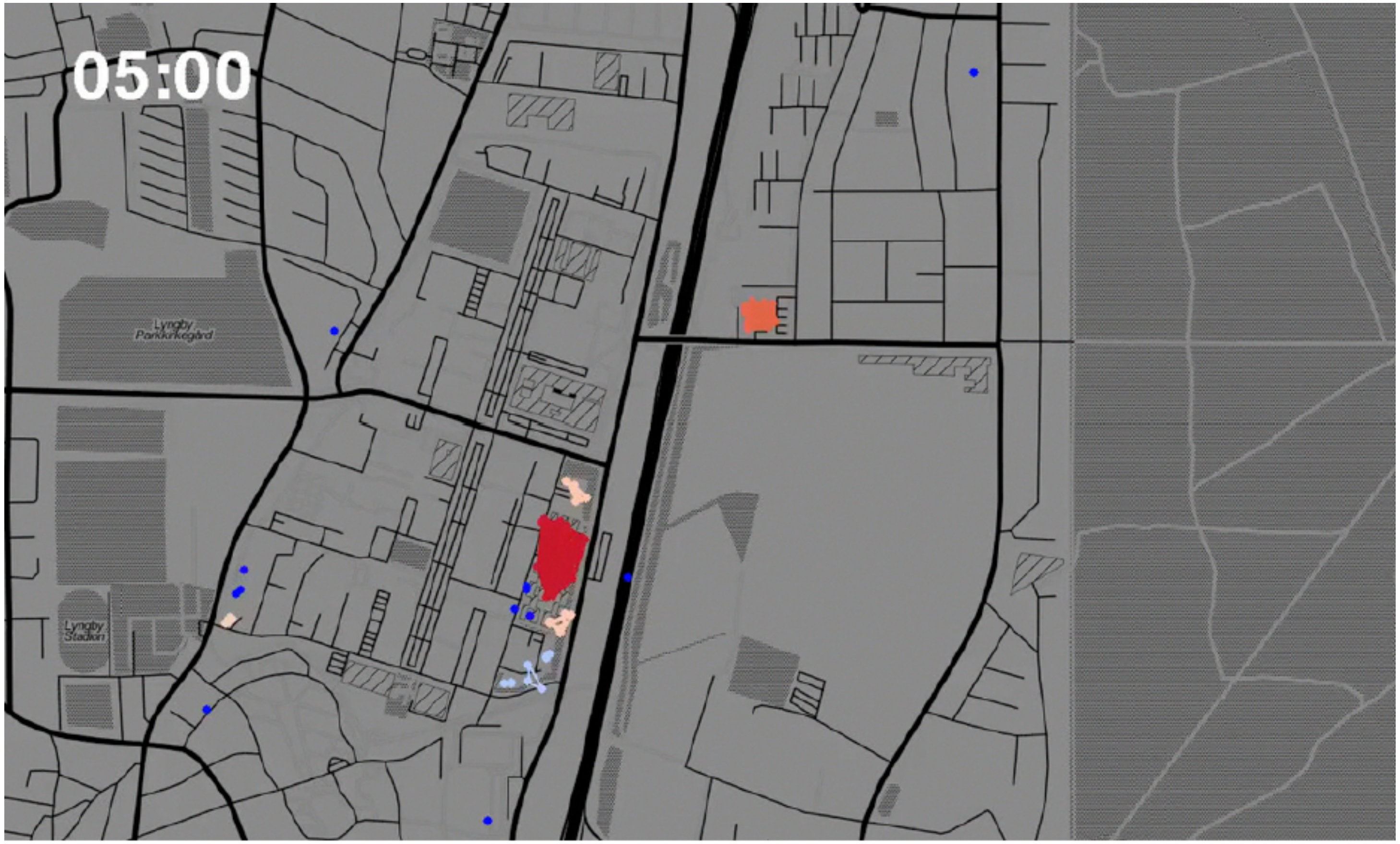
Big data

- 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
  - Cheap/free
- Cons
  - No established protocol for collection
  - Start-up costs
  - Even more privacy concerns
  - Corporate gatekeepers -> bias in access

Example: Social Fabric

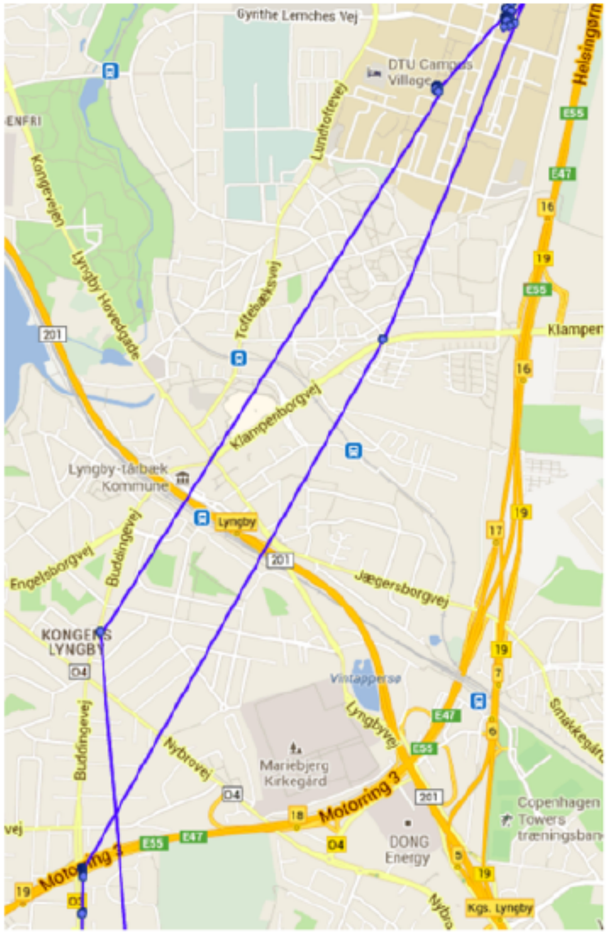
- Large-scale (N=1000) big data project
- Handed out smart phones to DTU freshmen
- Collected phone, SMS/text, GPS, wifi, bluetooth data
- -> Where, when, with whom
- -> social networks

Example: Social Fabric

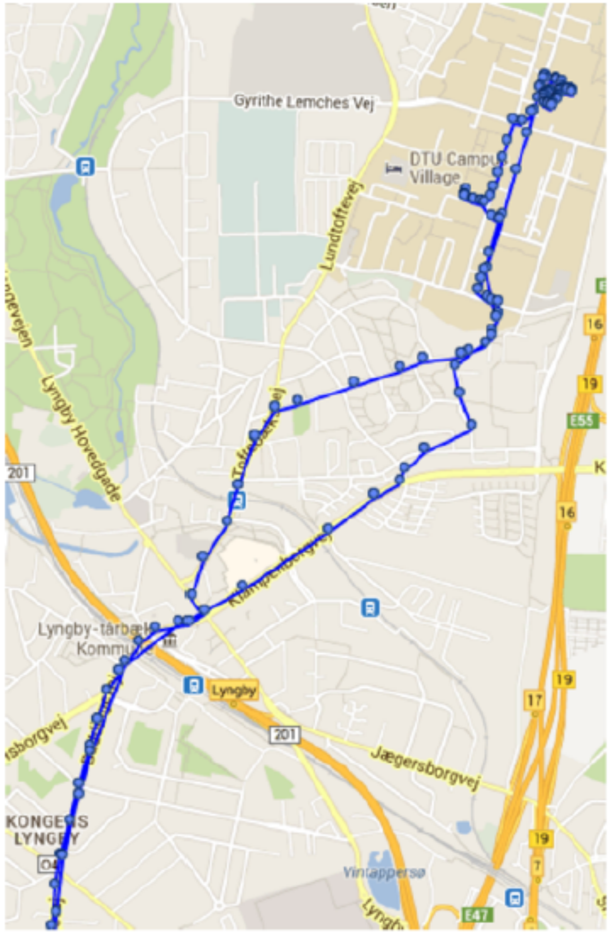


Phone locations 0500h Monday morning

Example: Social Fabric

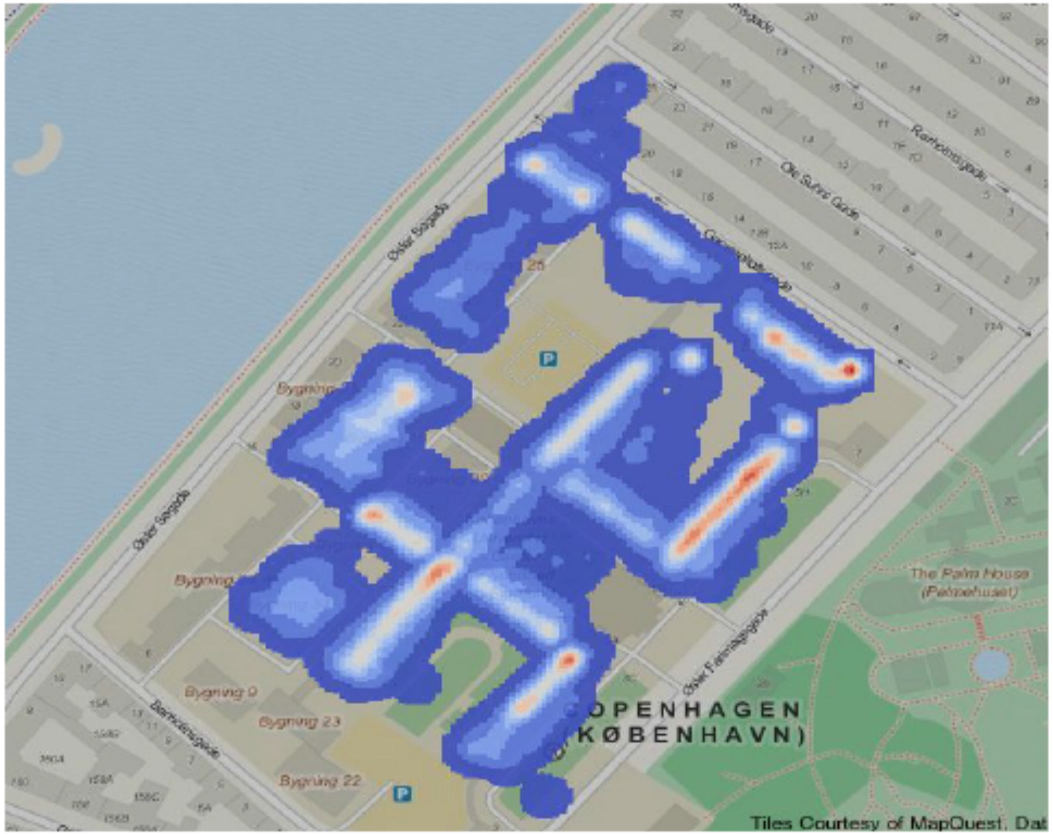


10 min GPS



wifi

Example: CSS



Heatmap of people with mobile devices on CSS (anonymous)

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

roadmap

- Different data for different questions
- Theory and empirics, forecasting and hypothesis testing
- Effects of causes vs. Causes of effects
- Data generating process
- Modes of data collection – pros and cons
- Strategic data management and data 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 to either
- Political reasons for obscuring or inventing data: Greece in EU, Chinese economy
- Firms: Proprietary info, competition reasons, fooling customers and regulators (VW)

Social desirability bias

- 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
  - Important for asking/assessing sensitive questions

Social desirability bias

- 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)