Social Data Science: big data and ethics

Andreas Bjerre-Nielsen
Econ & SODAS, UCPH
August 5, 2022

Plan today

1. Empirical design:
   - *data generating process* vs *modes of collection*
2. Data characteristics
   - Interventional vs. observational
   - Big data vs. administrative, survey, ethnographic
   - Strategic data provision
3. Ethics of data processing and use
   - Privacy and law
   - Ethics of using data for algorithms

Empirical design

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

What is your question, again?

From theory:

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

From data availability:

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, grades …
   Data science / machine learning (but also macroeconomics)
2. Hypothesis testing, derived from theory
   ´Traditional' social science

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 typically 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?
  - *> Strategic data provision*

Hypothesis testing

- Theory (rational choice, sociology, biology, common sense, …) posits effect of X on Y. Examples:
  A. *Selection/type theory*: People who are impatient cannot defer immediate pleasures -> smoke and drink while pregnant -> give 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 mechanism/theory (A) then little role for policy;
  - also, both can be true at same time
- How to distinguish: exogenous shock to birthweight, but ethically tricky ...

Interventional data

Data generating process

What is the **data generating process**?

Observational: endogenous decisions, researcher passive collector of data

Randomization/intervention: 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; Facebook experimenting with different algorithms; …)

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

- Narrow questions and narrow answers

- Large, important current debate in (development) economics
  - 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

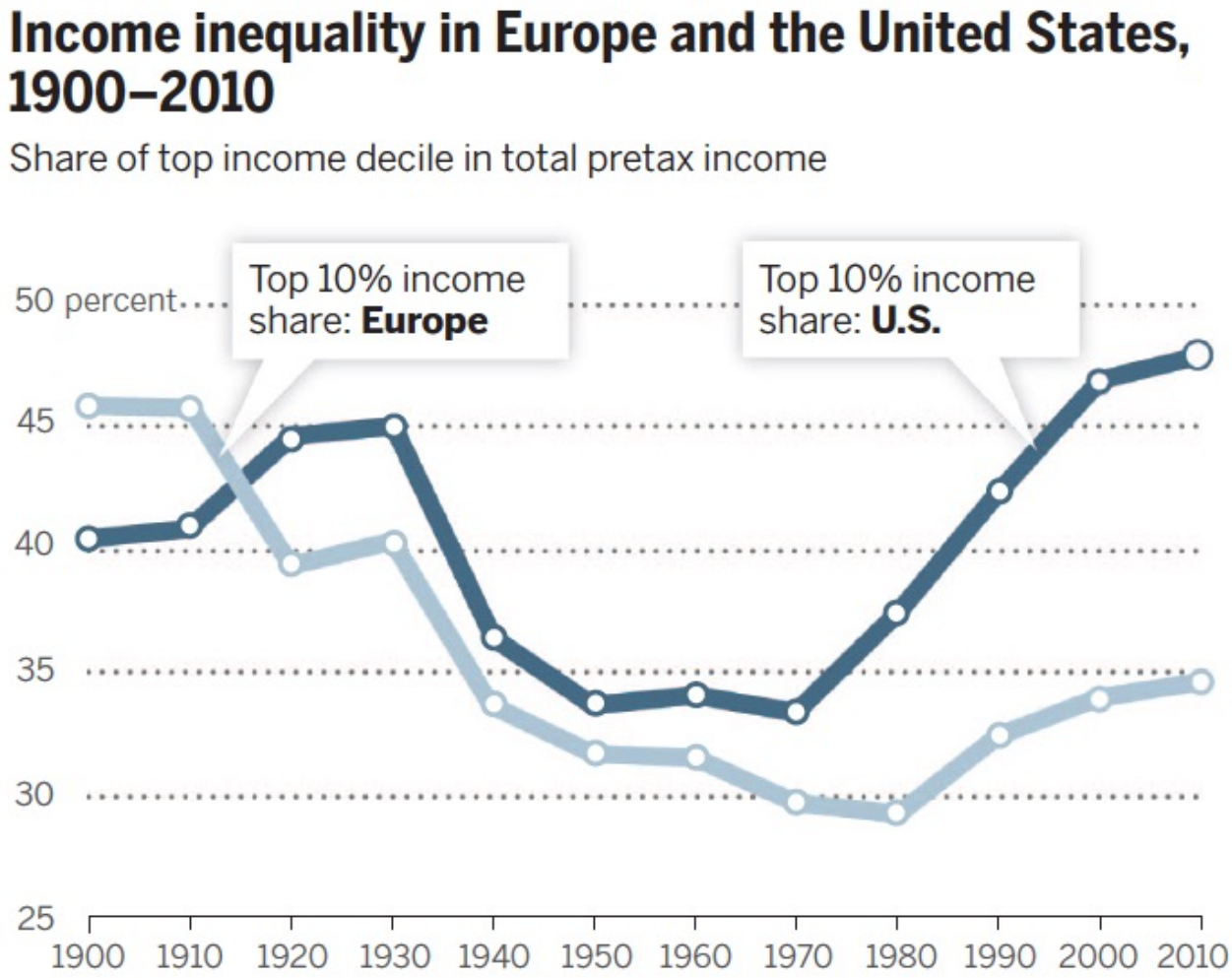
- Challenges
  - Limits to what can be studied by experimentation (**ethics**; law; feasibility)
  - Funding (field experiments **expensive**, survey experiment less so)
  - Often **participation constraint** – voluntary participants' gain >= 0 or no incentive
  - Subjects leave for various (systematic)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

Observational data

Observational data

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

Example: Inequality



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 provide structure to learn about causal mechanism under strong assumptions
  - May conflate correlation and causality

Observational data

- Example: 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

Big data is often observational

- Not always basis for causal claims
  - But interesting nonetheless: Description
- Can (potentially) be combined with natural/quasi-experiments.
  - Example: very detailed data on transportation/mobility and exogenous weather shocks-> effect of weather on mobility
  - Payday and consumption
- BUT: Google, FB, Amazon etc -> lots of field experimental data, in house

Modes of data collection

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 digitally stored transactions (in a broad sense)

Modes of data collection

- Note: survey, admin data, big data can all have randomized / exogenous elements or be purely observational (e.g. social media text and weather shocks)
- 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 *(strategic data provision)* and recall issues
  - Diverse interpretations (e.g. 1-10 scales)
  - Very different quality: interview vs. internet
  - Not full researcher control: Interviewer completions

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 and social science 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 on platforms etc.

No agreed upon definition what Big Data is

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

‘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’ = “deep data”
  - Sometimes cheap/free
- Cons
  - No established protocol for collection
  - Sometimes dubious quality, selection issues (both known/unknown), hard to validate
  - Start-up costs
  - Even more privacy concerns
  - Corporate gatekeepers -> bias in access (Facebook, Google)

Characteristics of ‘big data’

- Structured (row/column-style) vs. unstructured (images/text)
- 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: CSS

Heat map of people with mobile devices on CSS (anonymous)

Strategic data provision

Image concerns

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

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

Targets and Measures

- You cannot be told how your bank constructs a model for your likelihood of repayment. Why?
  - Goodhart's law: people will attempt to outmaneuver measure
  - (thought)example: spending on shoes good indicator of account overdraft -> shoe lovers will have others buy for them, ceases to be a good measure

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 (Joachim will return to that)
- Best predictor: number of cases past week

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.: Mac vs PC when ordering online)
    - Concerns about political issues
  - Or an **objective in itself**: Privacy as a political and commercial 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
    - COVID vaccination?
  - 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: What do you value most in a potential mate?
  - People say: "kind and understanding”
  - From dating data: physical attractiveness, status
  - Bias could be both (a) and (b)

Privacy

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
- For the weekend: visit https://teol.ku.dk/privacy/ Center for Privacy Studies, on how the concept originally evolved

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

Legal framework guiding personal data

- Before 2018:

  Persondataloven

- After 2018:

  GDPR +

  ”Lov om supplerende bestemmelser til forordning om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger (databeskyttelsesloven)”

https://www.retsinformation.dk/Forms/r0710.aspx?id=201319

From 2018: Danish law under EU data protection directive (GDPR)

- Link: https://gdpr-info.eu
- ”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.”
- **Individual consent** plays are much larger role (but special rules for DK)
- Some types of personal data are considered sensitive (health, political views, social problems)

GDPR

- Very different rules for
  - Research
  - Public administration
  - Private firms / organizations
- Potentially large penalties for non-compliance or misuse
- New job: DPO – Data Protection Officer
- Fair to say that interpretations of GDPR is work-in-progress, everywhere

Research and business don't mix

- 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 & FB uses this
- Implications for privacy and online behavior?
- When is a probability personal data?
- What if these data were passed on to private companies and political parties? (Cambridge Analytica, The Great Hack on Netflix)

Research exemption

- § 10. Oplysninger som nævnt i databeskyttelsesforordningens artikel 9, stk. 1, og artikel 10 må behandles, hvis dette alene sker med henblik på at udføre statistiske eller videnskabelige undersøgelser af væsentlig samfundsmæssig betydning, og hvis behandlingen er nødvendig af hensyn til udførelsen af undersøgelserne.
- Stk. 2. De oplysninger, der er omfattet af stk. 1, må ikke senere behandles i andet end videnskabeligt eller statistisk øjemed. Det samme gælder behandling af andre oplysninger, som alene foretages i statistisk eller videnskabeligt øjemed efter databeskyttelsesforordningens artikel 6.

Special for DK, also under GDPR: Can re-use data collected for other purposes for research or stats

- (50)Behandling af personoplysninger til andre formål end de formål, som personoplysningerne oprindelig blev indsamlet til, bør kun tillades, hvis behandlingen er forenelig med de formål, som personoplysningerne oprindelig blev indsamlet til. I dette tilfælde kræves der ikke andet retsgrundlag end det, der begrundede indsamlingen af personoplysningerne.

- Hvis behandling er nødvendig for at udføre en opgave i samfundets interesse eller henhører under offentlig myndighedsudøvelse, som den dataansvarlige har fået pålagt, kan EU-retten eller medlemsstaternes nationale ret fastsætte og præcisere de opgaver og formål, hvortil det bør være foreneligt og lovligt at foretage viderebehandling. Viderebehandling til arkivformål i samfundets interesse, til videnskabelige eller historiske forskningsformål eller til statistiske formål bør anses for at være forenelige lovlige behandlingsaktiviteter.

Individual data and privacy

- Statistics Denmark: Data users cannot present data at the individual level
- Examples that are no-go
  - Max of the income distribution
  - Median of income distribution
  - Max income in parish

- Things are different if you can anonymize data (e.g. a man in his 20s in Copenhagen)
- But! 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
- 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 (re)identifiability'? Active research area in computer science

Economic analysis of privacy

- Heffetz and Ligett: 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?"

Social Fabric data

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

Ethics in data use

Ethics of Big data

- Bit by Bit chapter 6
- Also:
  - ”Web scraping: a journalist’s guide” + ”on the ethics of web scraping and data journalism”
  - For journalists, but interesting for us as well
- Additional readings:
  - Neuhaus and Webmoor 2012: ”Agile ethics for massified research and visualization”
  - Also (google): Zimmer (2010) ”But the data is already public”: on the ethics of research in Facebook.

What is Ethics?

- In practice, so far:
  - social science: Ethics as a set of constraints; mixture of law and ”corporate social responsibiliy” / impression management
  - Data science: No ethics as absense of constraints; this is changing
- Compare: Medical science Are invasive procedures proportional to expected benefit?

- 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 and research often governed by
  - Institutional Reviews Boards (IRBs)
  - Personal ethics or feelings of right and wrong
  - Professional norms and codes of conduct (e.g. Econ vs Psychology); ”dual use” technology (research w military use)
- The law: (also) the institutional embodiment of ethics
- Denmark: Only formal ethics board for bio-medical research
  - Recently estbl. at Faculty of Social Sciences
- But what about firms?

Key goal of ethical considerations

- Reduce potential risk for 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?
  - Monopoly, mobility and informed consent

Is informed consent enough? Is it too much?

- 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?
  - Monopoly, mobility and informed consent
- Before GDPR: Firms often limited in what they could collect by law; now: jusk ask for informed consent
- But: Informed consent and public goods problems. Easy to say no and not give consent, but what if everyone does this?

Challenges

- Is it unethical 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 employees 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?
  - FAT ML: Fair, Accountable, and Transparent Machine Learning

Example – biased technology

- Buolamwini and Gebru found that face recognition was based against people of color and women
  - huge implications for consequences of other tech that depends on it: being found by the police, whether phone can unlock

- Potential biases in predictive algorithms:
  - recidivism risk (used in relation to criminal cases)
  - study completion (admisison, use of resources)
  - fraud detection (credit card, social benefits, tax)

The role of social science in tech

- What are the societal/economic consequences of adopting certain algorithms?
  - Can algorithms debias human biases, e.g. in police inspections? Preliminary answer – yes and simultaneously raise efficiency (e.g. jail likely re-offenders).
  - Can algorithms be used for inclusive policies, yes.

Ethical/legal 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 - now under review by SCOTUS, see here

25.764 views | Sep 10, 2019, 06:42am EDT

LinkedIn Data Scraping Ruled Legal

Emma Woollacott Senior Contributor
Cybersecurity

Photocredit: Getty GETTY

A court has ruled that it's legal to scrape publicly available data from LinkedIn, despite the company's claims that this violates user privacy.

San Francisco-based start-up hiQ Labs harvests user profiles from LinkedIn and uses them to analyze workforce data, for example by predicting when employees are likely to leave their jobs, or where skills shortages may emerge.

After LinkedIn took steps to block hiQ from doing this, hiQ won an injunction two years ago forcing the Microsoft-owned company to remove the block. That injunction has now been upheld by the 9th US Circuit Court of Appeals in a 3-0 decision.

Ethical/legal 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 - now under review by SCOTUS, see here
- Can you scrape data and resell? Or repackage?
- Does data collection cause significant costs (time or money) to firms and/or individuals?
- Typically more welcoming towards students, but be careful – and if in doubt, ask us!

Questions for proposed projects

- Do you respect privacy?
- Can single individuals be identified?
- What are potential consequences of (re)identification?
- What are terms and conditions for scraping and using data?

- Are there ethical considerations
  - With respect to individuals?
  - With respect to firms or organizations?