Inference Control

Ross Anderson

Cambridge

Haifa, September 7th 2020

3RD EDITION

SECURITY ENGINEERING

A GUIDE TO BUILDING DEPENDABLE DISTRIBUTED SYSTEMS

ROSS ANDERSON

WILEY

Forty years of inference control

- Early 1980s: early work on statistical disclosure control by Dorothy Denning, Tore Dalenius, …
- 1990s: we hit applications such as medical records where the data are too rich. Policy people in denial
- 2000s: search engines can identify people in large data sets such as movie preferences. Policy people call for PETs: along comes differential privacy
- 2010s: social media, location histories and genomics widen the gap between policy and reality
- Implications: from GDPR through opsec to ethics…

‘Anonymised data’ is one of those holy grails, like ‘healthy ice-cream’ or ‘selectively breakable crypto’

– Cory Doctorow

Statistical Disclosure Control

- Started about 1980 with US census
- Before then only totals & samples had been published, e.g. population and income per ward, plus one record out of 1000 with identifiers removed manually
- Move to an online database system changed the game
- Dorothy Denning bet her boss at the US census that she could work out his salary – and won!

Statistical Disclosure Control (2)

- A naïve approach is query set size control. E.g. in New Zealand a medical-records query must be answered from at least six records
- Problem: tracker attacks. E.g back when we had one female prof and six males:
  - 'Average salary professors'
  - 'Average salary male professors'
- Or even these figures for all 'non-professors'!
- On realistic assumptions, trackers exist for almost all sensitive statistics

Statistical Disclosure Control (3)

- A *characteristic formula* selects a *query set* (e.g. `all professors')
- The smallest query sets are cells
- If the set of *disclosed statistics* is D and the set of *sensitive statistics* is P, then we need $D \subseteq P'$ for privacy
- If $D = P'$ the privacy is *exact*
- Unfortunately if the minumum query set size $n < N/4$ where N is the total number of statistics, general trackers are easy to find

Statistical Disclosure Control (4)

- Cell suppression (Dalenius): suppose we can't reveal exam results for two or fewer students

Major:	Biology	Physics	Chemistry	Geology
Minor:				
Biology	-	16	17	11
Physics	7	-	32	18
Chemistry	33	41	-	2
Geology	9	13	6	-

Statistical Disclosure Control (5)

- But this is expensive! With n-dinemsional data, complementary cell suppression costs 2^n cells for each primary suppression

Major:	Biology	Physics	Chemistry	Geology
Minor:				
Biology	-	blanked	17	blanked
Physics	7	-	32	18
Chemistry	33	blanked	-	blanked
Geology	9	13	6	-

Statistical Disclosure Control (5)

- Query auditing – this is NP-complete, it ‘uses up’ your privacy budget, and users may collude
- Trimming – to remove outliers (e.g. the single HIV-positive patient in Chichester in the mid-1990s)
- Random sampling – answer each query with respect to a subset of records, maybe chosen by hashing the query with a secret key
- Swapping – exchange some records (e.g. census)
- Perturbation – add random noise

1995: UK HES Database Project

- The UK government wanted to start a research database of all hospital treatment in the UK
- Idea: dig out from records of hospital payments
- The BMA got me involved and we objected, pointing out the difficulties
- The government set up the Caldicott Committee which found many illegal data flows
- After the 1997 election, the new government just passed a law to legalize them
- Hospital Episode Statistics system started in 1998

Inference Control in Medicine

- Big problem in medical databases: context
- 'Show me all 34-yo women with 9-yo daughters where both have psoriasis'
- If you link episodes into longitudonal records, most patients can be reidentified
- Add demographic, family data: worse still
- Active attacks: worse still
- Social-network stuff such as friends, or disease contacts: worse still
- Only way to stay ethical: consent (via an opt-out)

Inference Control in Medicine (2)

- UK case law was established by the Source Informatics system for sanitised prescribing data. About as far as you can safely go – and even this was harder than it looks!

	Week 1	Week 2	Week 3	Week 4
Doctor 1	17	21	15	19
Doctor 2	20	14	3	25
Doctor 3	18	17	26	17

In Other Countries…

- In 1998 a startup (DeCODE) offered Iceland's health service free IT systems in return for access to records for research (by the Swiss drug company Roche)
- Records to be 'de-identified' by encrypting the social security number, but would be linked to genetic and family data, and run live (so active attacks possible)
- The Icelandic Medical Association persuaded 11% of citizens to opt out
- Eventually the Icelandic Supreme Court ruled the system should be opt-in, and the business collapsed

In Other Countries… (2)

- Germany: after 1989, they found they had valuable cancer registries from the former East Germany whose records were fully identifiable, thus illegal
- Netherlands, Austria: projects for central electronic heath records led to medical privacy activism
- USA: Latanya Sweeney identified the records of Massachussetts governor William Weld from the database of `anonymous’ VA records.
- Clinton government pushed through HIPAA to provide a (low) baseline of health privacy

Subsequent UK history

- Tony Blair ordered a “National Programme for IT” in the NHS in 2002
- Idea: replace all IT systems with standard ones, giving “a single electronic health record” with access for everyone with a “need to know”
- This became the biggest public-sector IT disaster in British history
- £11bn wasted, years of progress lost, lawsuits, and the flagship software didn’t work

European case law

- European law based on s8 ECHR right to privacy, clarified in the I v Finland case
- Ms I was a nurse in Helsinki, and was HIV+
- Her hospital's systems let all clinicians see all patients' records
- So her colleagues noticed her status – and hounded her out of her job
- The Finnish courts refused her compensation, but Strasbourg overruled them in 2010
- Now: we have the right to restrict our personal health information to the clinicians caring for us

Secondary Uses of Medical Data

- Cost control, clinical audit, research…
- Differing approaches:
  - USA: well-scrubbed incident data for open uses, lightly-scrubbed for controlled uses
  - Denmark, NZ: lightly scrubbed data kept centrally with strict usage control
  - Germany: no central collection
  - UK HES has summary data with postcode, date of birth
- UK approach appeared contrary to law, as people who tried to opt out were ignored

Limits of Medical Anonymisation

- Suppose you want Tony Blair's record
- A web search shows he was treated for an irregular heartbeat in Hammersmith hospital on 19 October 2003 and 1 October 2004
- Given a record like HES that links up successive hospital episodes, you've got him!
- If it doesn't, you can't do serious research with it
- So what's the solution?

The Political Track

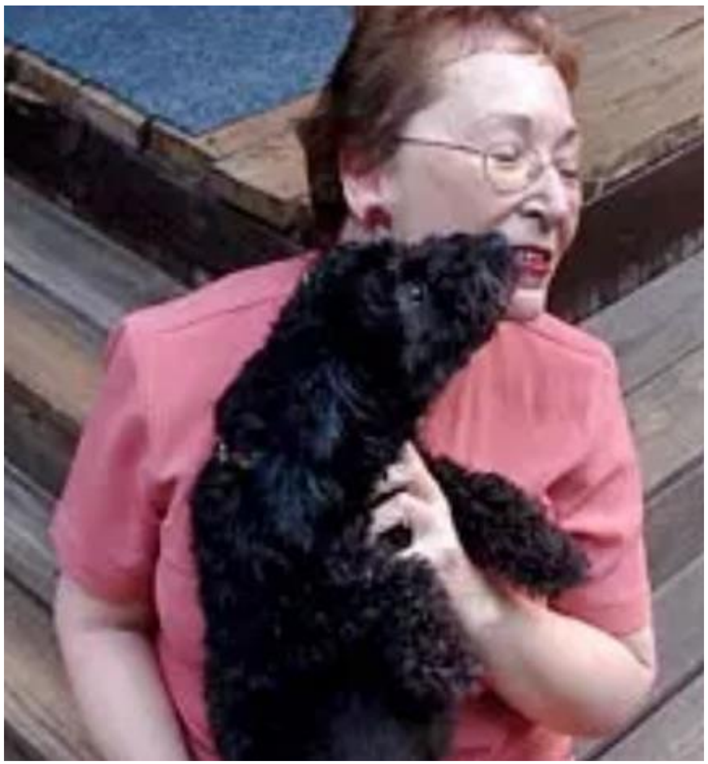
- 1980: Margaret Thatcher’s view of data protection
- David Waddington’s 1984 fix
- Tony Blair’s 1998 update
- The Information Commissioner’s conflict of interest
- The Caldicott Guardians’ conflict of interest
- The Thomas-Walport Review of 2007
- Paul Ohm’s ‘Broken Promises’ paper in 2009: computer scientists have known for 30 years that anonymization doesn’t work, but policy people stopped their ears

2010: ‘Transparency’

The care.data scandal

- Cameron policy announced January 2011: make ‘anonymised’ data available to researchers, both academic and commercial, but with opt-out
- In July 2013 the opt-out was removed (again) – NHS opt-outs have the wrong defaults and obscure mechanisms that get changed whenever too many people learn to use them (like Facebook’s)
- Apr 3 2014: we find that HES data were sold to 1200 universities, firms and others since 2013
- HES database is by now 22Gb, with 1 billion finished consultant episodes since 1998

The Third Wave



- AOL released 20m searches over three months by 657,000 people
- It was easy to see that user 4417749 was Thelma Arnold, 62, of Lilburn, Ga.
- AOL fired its CTO and the staff involved

The third wave (2)

- Netflix published `anonymized' ratings of 500,000 customers, offering $1m for a better recommender system
- Arvind Narayanan and Vitaly Shmatikov showed many subscribers could be reidentified against public preferences in the Internet Movie Database
- 'Long tail' insight: apart from the 100 most popular movies, people's preferences are pretty unique
- Policy response: try harder! Regulators call for research into Privacy Enhancing Technologies (PETs)

Differential privacy

- 2003: Kobbi Nissim and Irit Dinur considered reconstructing a database by linear algebra from random queries; if noise is small enough, you don’t need many of them. So the defender must add noise
- 2006: Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith showed how to analyse privacy systems that added noise to prevent disclosure
- Key insight: no individual's contribution to the results of queries should make too much of a difference, so you calibrate the standard deviation of the noise according to the sensitivity of the data

Differential privacy (2)

- A privacy mechanism is ε-indistinguishable if for all databases X and X' differing in a single row, the probability of getting any answer from X is within a factor of 1+ε of getting it from X’
- I.e., you bound the logarithm of the ratios
- Noise with a Laplace distribution gives indistinguishability with noisy sums; things compose, and become mathematically tractable
- I’ll leave the technical details for Kobi to discuss …

Differential privacy (3)

- DP gives us a dependable measure of privacy when we want to answer specific questions, not an anonymous database that will answer any question
- Now getting a full test in the 2020 US census!
- The 2010 census edited file (CEF) has 44 bits on each resident, 38% of which could be reconstructed using the Nissim-Dinur technique from the billions of bits in the published microdata summaries
- Only people who were swapped were protected; but the 2020 census will try to protect everybody

Differential privacy (4)

- But: adding noise means the totals don’t all add up
- As state totals need to add up to national totals, for Congressional districts, noise is added top down
- More noise in counties, more still in blocks, with special handling for edge cases (colleges, prisons…)
- Bu you no longer need to enumerate all the side information an attacker might use
- Extensive simulations suggest a value for ε of between 4 and 6

GDPR

- Germany, France were unhappy with the UK, Ireland implementing the Data Protection Directive with many deliberate loopholes
- So: General Data Protection Regulation 2016/679
- The most heavily-lobbied law ever in the European parliament with over 3000 amendments proposed
- Still no enforcement (so Max Schrems sues the Irish regulator, behind whom Google and Facebook hide)
- UK Information Commissioner hides behind the UK Anonymisation Network

The fourth wave

- The big changes since the second edition of my book are location, social and machine learning
- Universal smartphones and social networks both mean more data, while ML means better inference
- 2013: Yves-Alexandre de Montjoye, César Hidalgo, Michel Verleysen, and Vincent Blondel showed that four mobile-phone sightings are enough to identify
- Snowden tells us about ‘cotraveler’ and court cases since then tell about co-location analysis
- Private phone location data used by bounty hunters

The fourth wave (2)

- Example of ‘more data’: Stuart Thompson and Charlie Warzel bought a dataset of 50bn pings from 12m phones over several months in 2016–7
- Followed lots of different people:
  - both cops and demonstrators home from demos in DC
  - a singer at Trump’s inauguration, and secret service too
  - visitors to celebs and vice clubs
  - a Microsoft engineer who interviewed at Amazon, then shortly afterwards moved there
- See their “Twelve Million Phones, One Dataset, Zero Privacy”, New York Times Dec 19, 2019

The fourth wave (3)

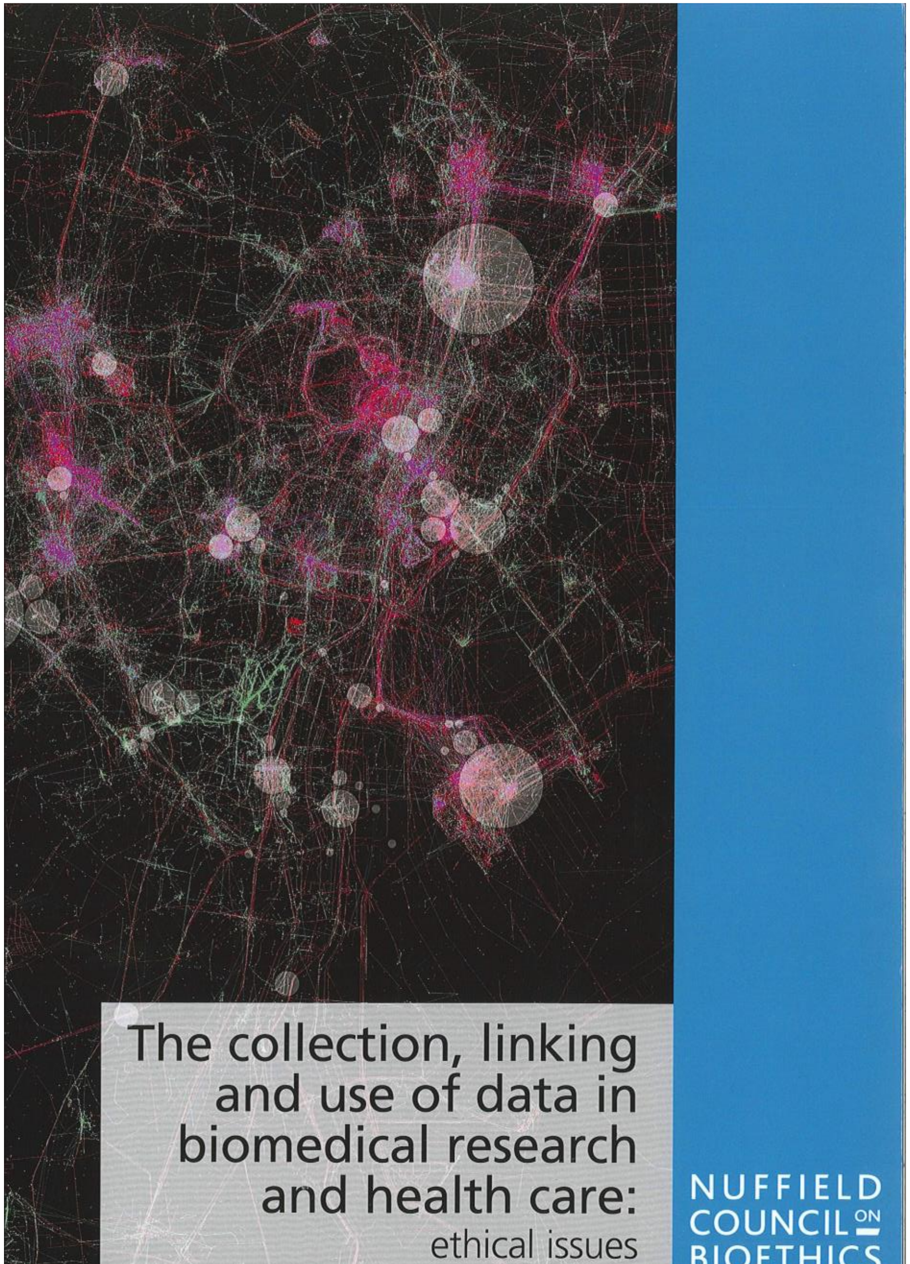
- Example of ‘better inference: Kumar Shahrad and George Danezis show you can use a random forest classifier to re-identify traffic data (CDRs identified by comparison with a social-network graph)
- Another example: the Cambridge Analytica scandal
- Starts when one of our postdocs figures out he can tell from 4 Facebook likes whether you’re gay
- A former colleague extends to personality traits, ethnicity, political preferences; 200k FB app users
- Analyses their many millions of ‘friends’ and sells this data to the Brexit and Trump campaigns

The fourth wave (4)

- Example of abuse: Google's AI subsidiary Deepmind persuaded the Royal Free Hospital, London, to give them patient records, saying they'd develop an app to diagnose acute kidney injury
- The hospital gave all 1.6m records, not those of the 60,000 relevant patients
- The ICO reprimanded the hospital but did not force Google to destroy the data
- The medical director of the hospital got promoted and is now a bigwig in the UK's Covid response

An Ethical Approach?

- It's long been accepted in medicine that the law's boundaries are way too wide
- If you do everything you can't be jailed or sued for, you'll quickly lose patients' trust
- So what is an ethical approach to medical practice, and medical research, in a world of cloud-based health records and genomics?
- Nuffield Bioethics Council set up a project …



The Nuffield Biodata report

- What happens to medical ethics in a world of cloud-based health records and pervasive genomics?
- 12 authors: from IT, medicine, ethics, insurance, pharma …

Problem Statement (1)

- Until 2003 all GP records were kept in PCs in the GP's surgery
- Government offered to pay for them
- Steadily everything moved to the cloud
- Hospital systems too, starting with radiology
- Now most clinical information is on a few big server farms
- Similar tech and policy trends elsewhere

Problem Statement (2)

- There's lots more data
  - Cloud-based primary and secondary care records
  - Genomics: from 100,000 patients to 50 million
  - Patient-generated stuff like fitbit
  - Comms data, lab data, all sorts of other stuff …
- And lots more capability to store & process it
- This led to all sorts of dumb initiatives from selling 10^9 records for £2000 to 1000+ users, through giving over 10^6 records to Google Deepmind

Problem Statement (3)

- In the old days, there was a clear distinction between operational and statistical uses
- The former had access controls, while the latter had inference controls
- Now the move to ‘personalised medicine’ is breaking down the barriers (is Deepmind direct care or research?)
- Anonymisation has turned out to be a ‘broken promise of privacy’ (in Paul Ohm’s words) or an `abomination’ (according to iPhone autocorrect)

Moral values and interests

- Distinction between public and private evolved over millennia – before history
- Norms of disclosure are important for formation and maintenance of identity and relationships
- Consent is how patient relationships work
- Public interests exist such as public health and research but these are not just in opposition to private interests in confidentiality

Law and governance

- Laws reflect emerging social consensus (albeit with a time lag and a big lobbying bias)
  - Data protection law
  - Human-rights law: s8 ECHR, I v Finland
- Usual take: ‘consent or anonymise’
- But anonymisation doesn’t work, and consent is becoming steadily harder!
- Regulators are captured and parliament doesn’t care
- What should an ethical researcher do?

Principle 1 – Respect for persons

- **The set of expectations about how data will be used in a data initiative should be grounded in the principle of respect for persons**
- This includes recognition of a person's profound moral interest in controlling others' access to, and disclosure of, information relating to them held in circumstances they regard as confidential

Principle 2 – Human rights

- **The set of expectations about how data will be used in a data initiative should be determined with regard to established human rights**
- This will include limitations on the power of states and others to interfere with the privacy of individual citizens in the public interest (including to protect the interests of others)

Principle 3 – Participation

- **The set of expectations about how data will be used (or re-used) in a data initiative, and the appropriate measures and procedures for ensuring that those expectations are met, should be determined with the participation of people with morally relevant interests**
- Where it is not feasible to engage all those with relevant interests, the full range of relevant interests and values should nevertheless be fairly represented

Principle 4 – Accounting for decisions

- **A data initiative should be subject to effective systems of governance and accountability that are themselves morally justified**
- This should include both structures of accountability that invoke legitimate judicial and political authority, and social accountability arising from engagement of people in a society
- Accountability must include effective measures for communicating expectations and failures of governance, execution and control to people affected and to society more widely

Application to security research?

- Started thinking about this following Facebook app that led to the Cambridge Analytica scandal
- Our Device Analyzer ran on 20k+ Androids
- For user: personal analytics (best phone plan)
- For us: understanding smartphone use, energy consumption, cybercrime and much else
- We then extended this to all our cybercrime work, much of which involves data that will never be ‘open data’ for variousreasons

The Cambridge Cybercrime Centre

- Until 2015, cybercrime research wasn't a science…
- To help fix this, the Cambridge Cybercrime Center now collects and curates masses of data on malware, spam, phish, botnet c&c traffic, crime forum posts, …
- These are licensed to 100+ researchers at 30+ universities in Europe & elsewhere
- If you have data, we can get it to academics who can use it
- If you want to do research on cybercrime, we have a lot of data you can use

Limitations of Ethics as an Approach

- Ethics committees fix the problems of mens rea in criminal law and the ‘standards of the industry’ in tort law
- In other words, they protect the researcher, not the data subject
- The dark side is the wicked security economics!
- Yet the reality of modern research is shown by Ben Goldacre’s work on Covid epidemiology. If you work directly with the data you can get the results

Future Directions?

Privacy is a transient notion. It started when people stopped believing that God could see everything and stopped when governments realised there was a vacancy to be filled.

– Roger Needham