Inference Control

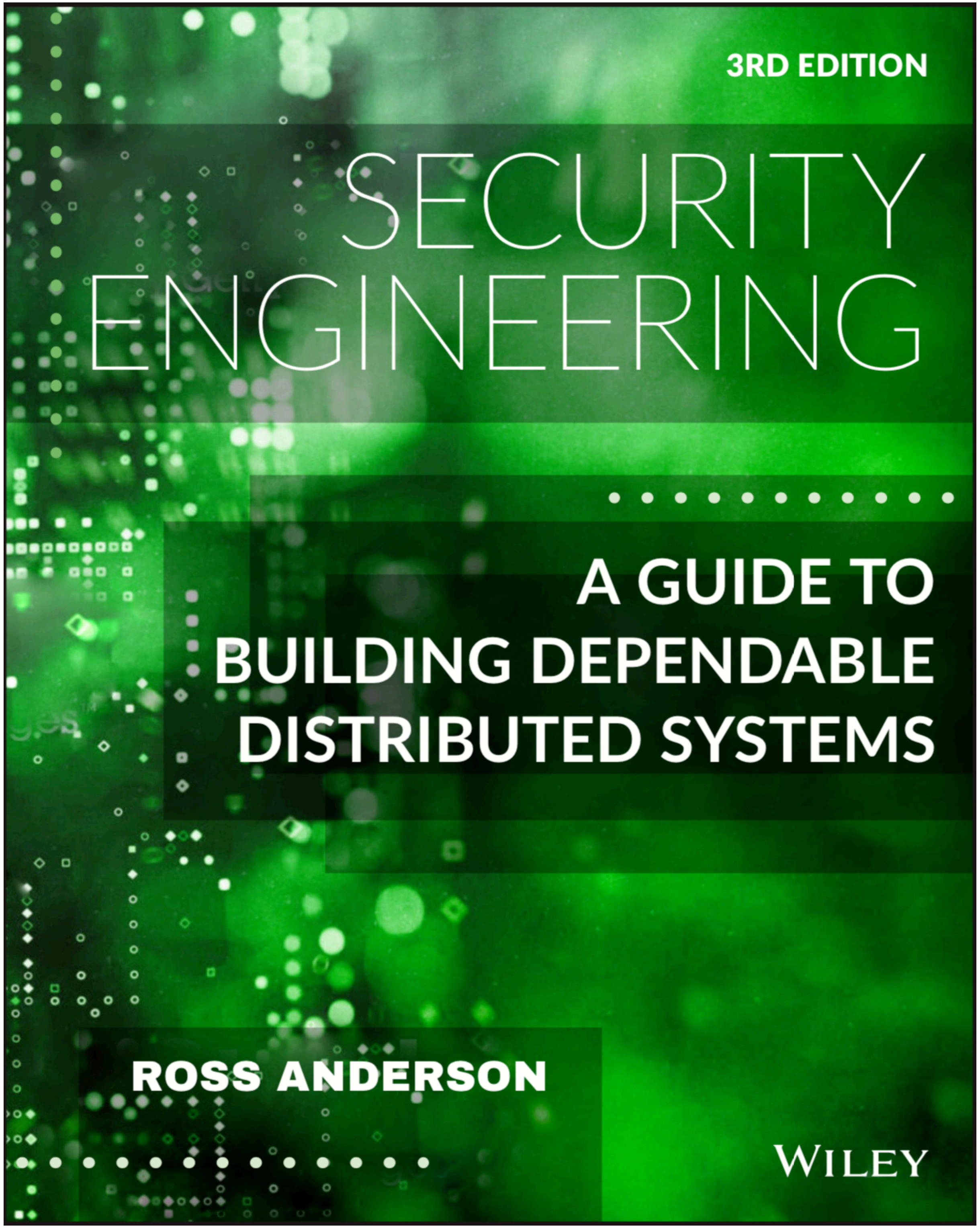
Ross Anderson

Cambridge

Haifa, September 7th 2020

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

– Cory Doctorow



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 still in denail: 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…

Statistical Disclosure Control

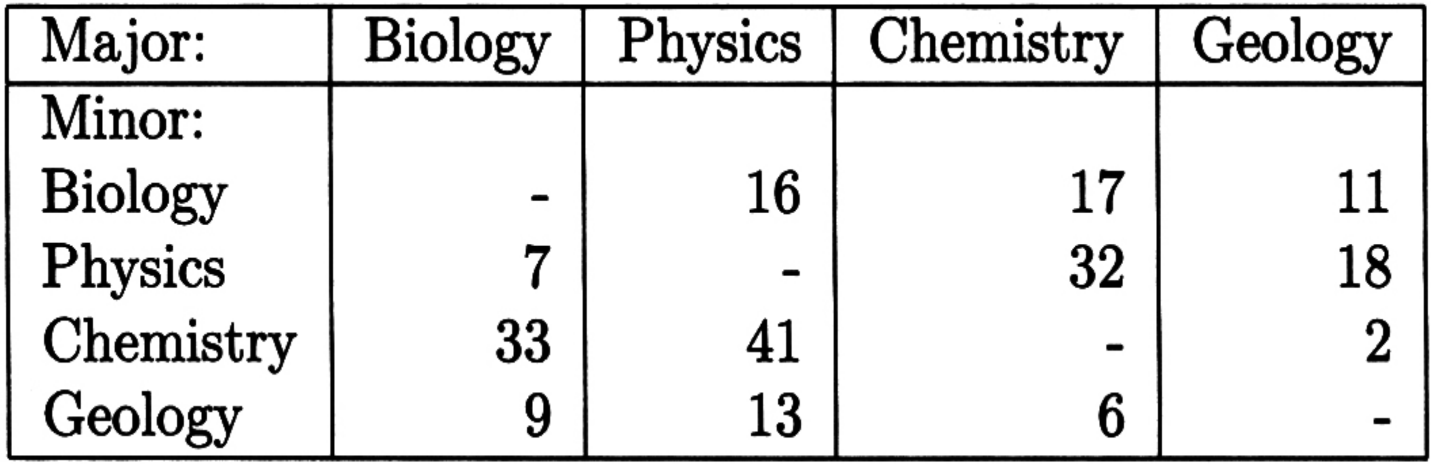
- Started about 1980 with US census
- Before then only totals & samples had been published, e.g. population and income per ward, plus one record out of 1000 with identifiers removed manually
- Move to an online database system changed the game
- Dorothy Denning bet her boss at the US census that she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

Statistical Disclosure Control (2)

- A naïve approach is to control query-set size. 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)

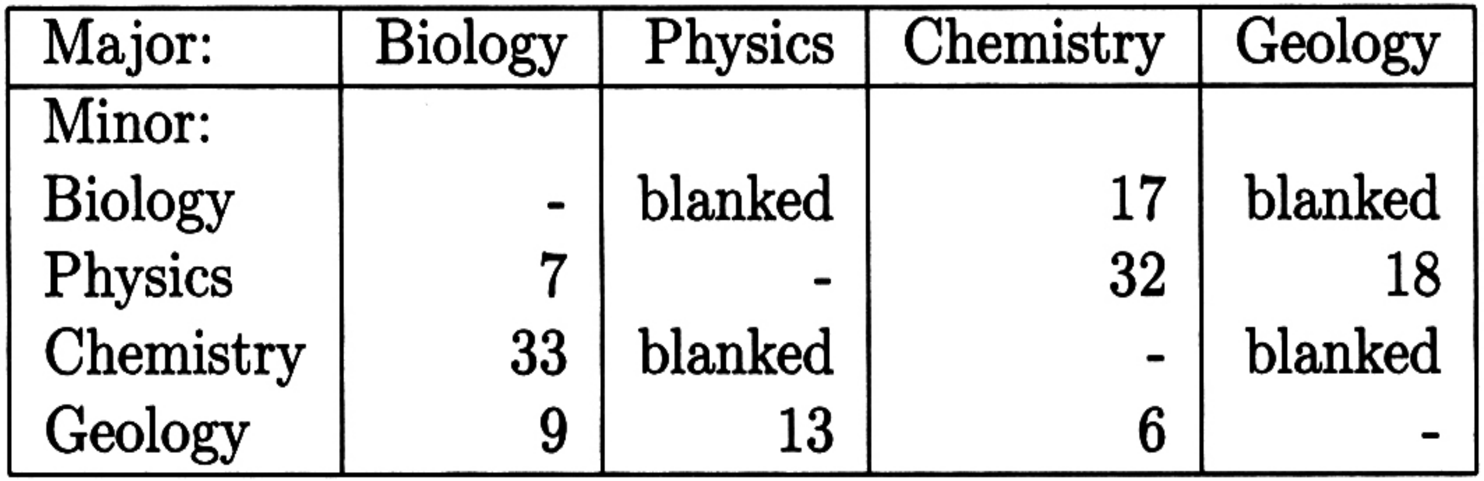
- Cell suppression (Dalenius): suppose we can't reveal exam results for two or fewer students



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 (4)

- 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 (UK etc census)
- Perturbation – add random noise (more later)

Secondary Uses of Medical Data

- Cost control, clinical audit, research…
- Differing approaches:
  - Germany: no central collection. But cancer after 1989
  - Denmark, NZ: lightly scrubbed data kept centrally with strict usage control (Germany followed for registries)
  - USA: lightly-scrubbed for controlled uses, slightly better scrubbed data for open uses. But Latanya Sweeney 1996
  - UK has summary data with postcode, date of birth
- UK had medical privacy issues from late 1990s, as people who tried to opt out were ignored

Limits of Medical Anonymisation

- A web search shows Tony Blair was treated for an irregular heartbeat in Hammersmith hospital on 19 October 2003 and 1 October 2004
- If a database links up successive hospital episodes, you’ve got him!
- If it doesn’t, you can’t do serious research with it
- 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)

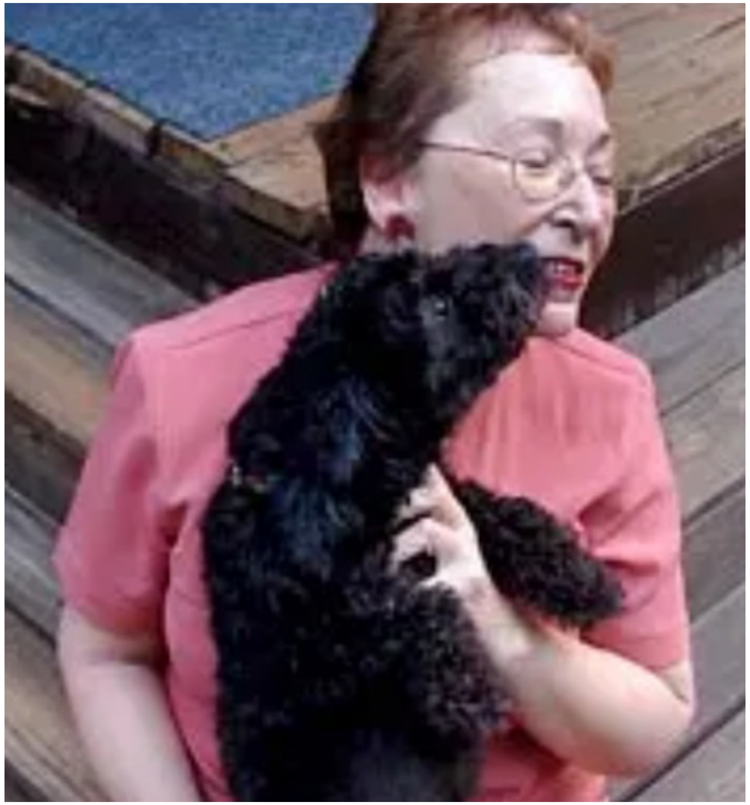
European case law

- 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

Britain’s care.data scandal

- Cameron policy from January 2011: make ‘anonymised’ data available to researchers, both academic and commercial, but with opt-out
- Opt-outs had the wrong defaults, difficult access and obscure mechanisms that got changed whenever too many people learned to use them (like Facebook’s)
- Apr 3 2014: we found that hospital data were sold to 1200 universities, firms and others since 2013
- The HES database they sold was 22Gb, with 1 billion finished consultant episodes 1998–2013

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’ data on 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
- Paul Ohm’s 2009 paper “Broken promises of privacy” – policymakers don’t want to know

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

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
- Much Big Tech behaviour is now illegal, but there’s no enforcement!
- So Max Schrems sues the Irish regulator, behind whom Google and Facebook hide, along with many other US firms

The fourth wave

- The big changes during the 2010s 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

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)

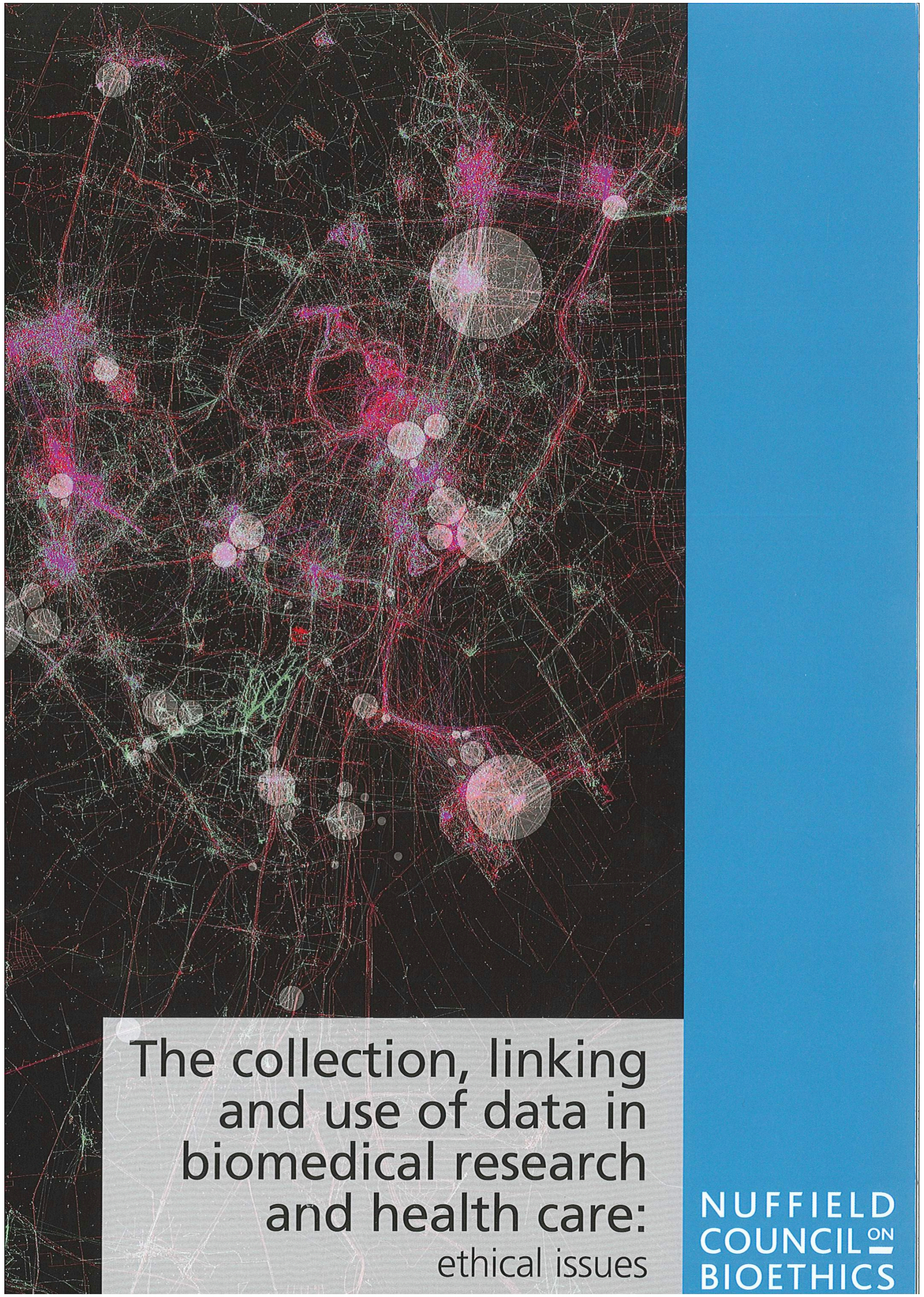
- Example of ‘better inference’: Kumar Sharad and George Danezis show you can use a random forest classifier to re-identify records of who phoned whom, by comparison with a social-network graph
- Another example: the Cambridge Analytica scandal
  - Started when a Cambridge postdoc figured out he could tell from 4 Facebook likes whether you’re gay
  - A (former) colleague extended to personality traits, ethnicity, political preferences; 200k FB app users
  - Analysed their many millions of ‘friends’ and sells this data to the Brexit and Trump campaigns
  - Clear breach of privacy law, election law, research ethics

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 kidney 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 collection, linking and use of data in biomedical research and health care: ethical issues

NUFFIELD COUNCIL ON BIOETHICS

The Nuffield Biodata report

- What happens to medical ethics once we have cloud-based health records and pervasive genomics?
- When ‘consent or anonymise’ fails, 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

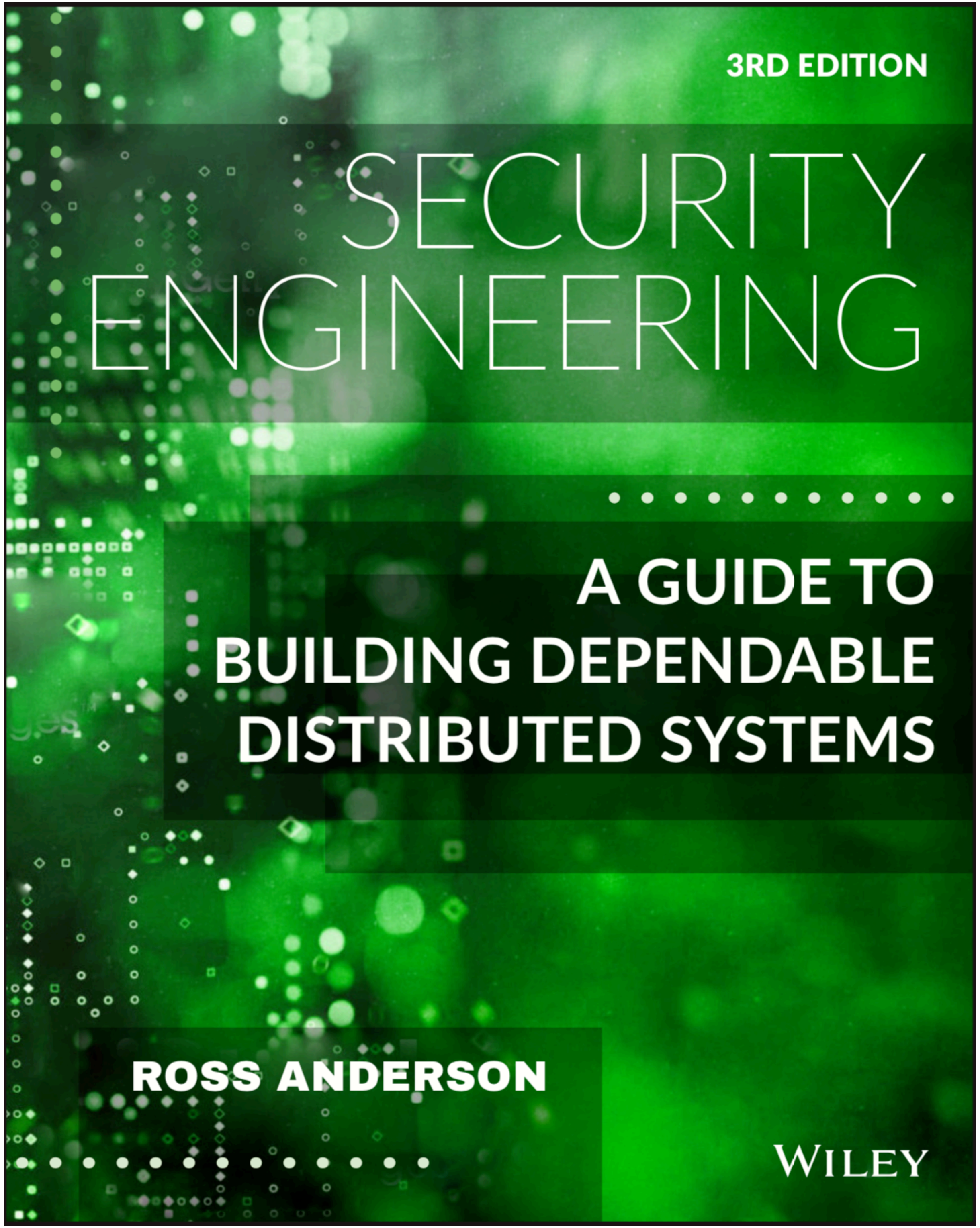
Limitations of Ethics as an Approach

- The reality of modern research is shown by Ben Goldacre’s ‘OpenSAFELY’ work on Covid. Work directly with the data in place to get the results
- Ethical approval essential
- We use the Nuffield approach in our work on cybercrime data
- Ethics committees only go so far though
- They protect the researcher, not the data subject
- Given wicked security economics, you need law too

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



3RD EDITION

SECURITY ENGINEERING

A GUIDE TO BUILDING DEPENDABLE DISTRIBUTED SYSTEMS

ROSS ANDERSON

WILEY