**Chapter 11**

**Inference Control**

**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

**“Anonymized data” is one of those holy grails, like “healthy**

**ice-cream” or “selectively breakable crypto”**

– CORY DOCTOROW

**11.1** **Introduction**

Just as Big Tobacco spent decades denying that smoking causes lung cancer,  
 and Big Oil spent decades denying climate change, so also Big Data has spent  
 decades pretending that sensitive personal data can easily be ‘anonymised’ so  
 it can be used as an industrial raw material without infringing on the privacy  
 rights of the data subjects.

Anonymisation is an aspirational term that means stripping identifying in-

formation from data in such a way that useful statistical research can be done  
 without leaking information about identiﬁable data subjects. Its limitations

have been explored in four waves of research, each responding to the technology  
 of the day. The ﬁrst wave came in the late 1970s and early 1980s in the context  
 of the US census, which contained statistics that were sensitive of themselves but  
 where aggregate totals were required for legitimate reasons such as allocating  
 money to states; and in the context of other structured databases from college  
 marks through staff salaries to bank transactions. Statisticians started to study  
 how information could leak, and to develop measures for inference control.

The second wave came in the 1990s as medical records were computerised.

Both health service administrators and medical researchers saw this as a treasure  
 trove, and hoped that removing patients’ names and addresses would be enough  
 to make the data non-personal. This turned out to be insufficient because of  
 the richness of the data, which led to tussles in several countries including the

US, the UK, Germany and Iceland. There have since been multiple scandals  
 when inadequately anonymised data were leaked or even sold.

The third wave, in the mid-2000s, came when people realised they could

use search engines to identify people in large datasets of consumer preferences  
 such as movie ratings and search engine logs. An advance in theory came in  
 2006, when Cynthia Dwork and colleagues developed the theory of *differential*  
 *privacy*, which quantiﬁes the extent to which inferences can be prevented by  
 limiting queries and adding noise, enabling us to add noise where it’s needed.  
 This is now being used in the US census, whose experience teaches a lot about  
 its practical limits.

The fourth wave came upon us in the late 2010s with social media, pervasive

genomics and large databases of personal location histories collected by phone  
 apps and widely sold to marketers. Ever more companies who sell personal

information at scale pretend that it isn’t personal because names are somehow  
 tokenised. Ever more press articles show how bogus such claims usually are.  
 For example, in December 2019 the New York Times reported analysing the  
 mobile-phone location history of 12 million Americans over a few months, lo-  
 cating celebrities, rioters, police, Secret Service officers and even sex-industry  
 customers without difficulty [1885].

We face a yawning gap between what can be done using anonymisation and

related privacy technologies, and what stakeholders from medical researchers  
 through marketers to politicians would like to believe is possible. This gap has  
 been the subject of much discussion and, as with tobacco and carbon emissions,  
 political argument. As our knowledge of the re-identiﬁcation risks becomes ever  
 more detailed and certain, so the hopes of both governments and industry be-  
 come ever more unrealistic. Governments repeatedly call for proposals, and data  
 users call for contractors, to create services that cannot be created; all too of-  
 ten, contracts for privacy services are won by the more ignorant or unscrupulous  
 operators.

It must be said that not all governments have simply been ignorant. Both

the UK and Ireland, for example, annoyed other EU member states for years by  
 allowing ﬁrms to pretend that data were anonymous when they clearly weren’t,  
 and this was one of the factors that led the EU to pass the General Data  
 Protection Regulation (GDPR), as I will discuss later in section 26.6.1. Since it  
 came into force, the wriggle room for wishful thinking has become less – though  
 even the European institutions have sometimes had a rosy view of what can be  
 achieved by de-identiﬁcation.

**11.2** **The early history of inference control**

Inference control goes back to the 1920s when economic data were compiled in  
 ways that masked the contribution of individual ﬁrms, but it was ﬁrst studied  
 systematically in the context of census data. A census collects a lot of sensitive  
 information about individuals, including names, addresses, family relationships,  
 race, employment, educational attainment and income, and then makes statisti-  
 cal summaries available by geographical and governmental units such as states,

|  |  |  |
| --- | --- | --- |
|  |  |  |

counties, districts and wards. This information is used to determine electoral  
 districts, to set levels of government funding for public services, and as inputs  
 to all sorts of other policy decisions. Census data are a good simple case with  
 which to start as the data are in a standard format, and the allowable queries  
 are generally known in advance.

There are two broad approaches, depending on whether the data are sanitised

once and for all before publication, or whether the privacy mechanisms operate  
 one query at a time and work out whether it’s allowable. Mathematically, the  
 two types of processing are the same. For data of a particular type subject to  
 given privacy constraints, only a certain number of queries will be allowable; the  
 question is whether you determine these in advance, or dynamically in response  
 to user demand.

An example of the ﬁrst type comes from the US census data up till the 1960s.

One record in a thousand was made available on tape – minus names, exact  
 addresses and other sensitive data. There was also noise added to the data in  
 order to prevent people with some extra knowledge (such as of the salaries paid  
 by the employer in a company town) from tracing individuals. In addition to the  
 sample records, local averages were also given for various attributes. But records  
 with extreme values – such as very high incomes – were suppressed. Without  
 such suppression, a wealthy family living in a small village might increase the  
 average village income by enough for their own family income to be deduced.

In the second type of processing, identiﬁable data are stored in a database,

and privacy protection comes from restricting the queries that may be made. For  
 example, a simple rule might be that you answer no question unless the result  
 is computed using the data of three or more data subjects – the so-called *rule of*  
 *three*. Early attempts at this were not very successful, as people kept on coming  
 up with new attacks based on inference. A typical attack would construct a  
 number of queries about samples containing a target individual, and work back  
 to infer some conﬁdential fact. You might for example ask ‘tell me the number  
 of two-person households earning between $50,000 and $55,000’, ‘tell me the  
 proportion of households headed by a man aged 40–45 years earning between  
 $50,000 and $55,000’, ‘tell me the proportion of households headed by a man  
 earning between $50,000 and $55,000 whose children have grown up and left  
 home’, and so on, until you home in on the target individual. Queries to which  
 we successively add context to defeat query controls are known as *trackers*.

Related problems arise in many contexts. For example, a New Zealand

journalist deduced the identities of many officers in that country’s signals intel-  
 ligence service, GCSB, by scrutinising lists of military and diplomatic personnel  
 for patterns of postings over time [849]. Combining low-level sources to draw a  
 high-level conclusion is known as an *aggregation attack* in the national security  
 context.

**11.2.1** **The basic theory of inference control**

The basic theory of inference control was developed by Dorothy Denning and  
 others in late 1970s and early 1980s, largely in response to problems of the US  
 census [538]. This wave of research is summarised in a 1989 survey paper by

|  |  |  |
| --- | --- | --- |
|  |  |  |

Adam and Wortman [17]. The developers of many modern privacy systems are  
 often unaware of this work, and repeat many of the mistakes of the 1960s. The  
 following is an overview of the basic ideas.

A *characteristic formula* is the expression (in some database query language)

that selects a *query set* of records. An example might be ‘all female employees  
 of the Computer Laboratory at the grade of professor’. The smallest query sets,  
 obtained by the logical AND of all the attributes (or their negations) are known  
 as *elementary sets* or *cells*. The statistics corresponding to query sets may be  
 *sensitive statistics* if the set size is too small. The objective of inference control  
 is to prevent the disclosure of sensitive statistics.

|  |
| --- |
| If we let *D* be the set of statistics that are disclosed and *P* the set that are  sensitive and must be protected, then we need *D ✓ P 0* for privacy, where *P 0*  is the complement of *P*. If *D* = *P 0*, then the protection is said to be *precise*.  Protection that is not precise will usually carry some cost in terms of the range of |
| queries that the database can answer and may therefore degrade its usefulness. |

**11.2.1.1** **Query set size control**

The simplest protection mechanism is to specify a minimum query set size, so  
 that no question is answered if the number of records from which the answer is  
 calculated is less than some threshold *t*. But this is not enough. Say *t* = 6; then  
 an obvious tracker attack is to make an enquiry on six patients’ records, and  
 then on those records plus the target’s. And you must also prevent the attacker  
 from querying all but one of the records: if there are *N* records and a query set  
 size threshold of *t*, then between *t* and *N ffi t* records must be the subject of a  
 wrote the ﬁrst edition of this book, only one of the full professors in our lab was  
 female. So we could have found out her salary with just two queries: ‘Average  
 salary professors?’ and ‘Average salary male professors?’. So you have to avoid  
 successive queries of record sets *K* and *L* if *K ⇢ L* and *|L| ffi |K| < t*.

**11.2.1.2** **Trackers**

That is an example of an *individual tracker*, a custom formula that allows us  
 to calculate the answer to a forbidden query indirectly. There are also *general*  
 *trackers* – sets of formulae that will enable any sensitive statistic to be revealed.  
 A somewhat depressing discovery made in the late 1970s, due to Dorothy Den-  
 ning, Peter Denning and Mayer Schwartz, was that general trackers are usually  
 easy to ﬁnd. Provided the minimum query set size *n* is less than a quarter of  
 the total number of statistics *N*, and there are no further restrictions on the  
 type of queries that are allowed, then we can ﬁnd formulae that provide general  
 trackers [541]. So tracker attacks are easy, unless we restrict the query set size  
 or control the allowed queries in some other way. Such *query auditing* turns out  
 to be an NP-complete problem.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**11.2.1.3** **Cell suppression**

The next question is how to deal with the side-effects of suppressing sensitive  
 statistics. The UK rules for the 2010 census, for example, required that it be  
 ‘unlikely that any statistical unit, having identiﬁed themselves, could use that  
 knowledge, by deduction, to identify other statistical units in National Statistics  
 outputs’ [1416]. To take a simple concrete example, suppose that a university  
 wants to release average marks for various combinations of courses, so that  
 people can check that the marking is fair across courses. Suppose now that  
 the table in Figure 11.1 contains the number of students studying two science  
 subjects, one as their major subject and one as their minor subject.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Major: | Biology | Physics | Chemistry | Geology |
| Minor:  Biology  Physics  Chemistry  Geology | - | - | - | - |

Figure 11.1: Table containing data before cell suppression

The UK census rules imply a minimum query set size of 3, which makes

sense here too: if we set it at 2, then either of the two students who studied  
 ‘geology-with-chemistry’ could work out the other’s mark. So we cannot release  
 the average for ‘geology-with-chemistry’. But if the average mark for chemistry  
 is known, then it could be reconstructed from the averages for ‘biology-with-  
 chemistry’ and ‘physics-with-chemistry’. So we have to suppress at least one  
 other mark in the chemistry row, and for similar reasons we need to suppress one  
 in the geology column. But if we suppress ‘geology-with-biology’ and ‘physics-  
 with-chemistry’, then we’d also better suppress ‘physics-with-biology’ to prevent  
 these values being worked out in turn. Our table will now look like Figure 11.2,  
 where ‘D’ means ‘value suppressed for disclosure purposes’.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Major: | Biology | Physics | Chemistry | Geology |
| Minor:  Biology  Physics  Chemistry  Geology | - | D  -  D | - | D    D  - |

Figure 11.2: Table after cell suppression

This process, due to Tore Dalenius, is called *complementary cell suppression*.

|  |  |  |
| --- | --- | --- |
| If there are further attributes in the database schema – for example, if ﬁgures are  also broken down by race and sex, to show compliance with anti-discrimination | | |
| laws – then even more information may be lost. Where a database scheme  contains *m*-tuples, blanking a single cell generally means suppressing 2*m �* 1 | | |
| So even precise protection can rapidly make the database unusable. Sometimes  complementary cell suppression can be avoided, as when large incomes (or rare | | |
|  |  |  |

diseases) are tabulated nationally and excluded from local ﬁgures. But it is  
 often necessary when we are publishing microstatistics, as in the above tables  
 of exam marks. It may still not be sufficient, unless we can add noise to the  
 totals – as the possible values of the conﬁdential data are limited still further  
 by the information we disclose, and there may also be side information such as  
 the fact that no totals are negative.

**11.2.1.4** **Other statistical disclosure control mechanisms**

|  |
| --- |
| Another approach is *k-anonymity*, due to Pierangela Samarati and Latanya  Sweeney, which means that each individual whose data is used in calculating  a release of data cannot be distinguished from *k �* 1 others [1795]. Its limita-  mathematical deﬁnition of a privacy property; it’s not much help if *k* individuals  all possess the same sensitive attribute. Where the database is open for online  queries, we can use *implied queries control*: we allow a query on *m* attribute  values only if every one of the 2*m* implied query sets given by setting the *m*  attributes to true or false, has at least *k* records. An alternative is to limit the  type of inquiries. *Maximum order control* limits the number of attributes any  query can have. However, to be e↵ective, the limit may have to be severe. It  takes only 33 bits of information to identify a human, and most datasets are of  much smaller populations. A more thorough approach (where it is feasible) is  to reject queries that would partition the sample population into too many sets. |

We saw in the previous chapter how lattices can be used in compartmented

security to deﬁne a partial order of permitted information ﬂows between com-  
 partments with combinations of codewords. They can also be used in a slightly  
 different way to systematize query controls in some databases. If we have, for  
 example, three attributes *A*, *B* and *C* (say area of residence, birth year and  
 medical condition), we may ﬁnd that while enquiries on any one of these at-  
 tributes are non-sensitive, as are enquiries on *A* and *B* and on *B* and *C*, the  
 combination of *A* and *C* might be sensitive. It follows that an enquiry on all  
 three would not be permissible either. So the lattice divides naturally into a ‘top  
 half’ of prohibited queries and a ‘bottom half’ of allowable queries, as shown in  
 Figure 11.3.

**11.2.1.5** **More sophisticated query controls**

There are a number of alternatives to simple query control. During the late 20th  
 century, the US census used the ‘*n*-respondent, *k*%-dominance rule’: it would  
 not release a statistic of which *k*% or more was contributed by *n* values or less.  
 Other techniques included suppressing data with extreme values. A census may  
 include high-net-worth individuals in national statistics but not in the local  
 ﬁgures, while some medical databases do the same for less common diseases.  
 For example, a UK prescribing statistics system from that period suppressed  
 sales of AIDS drugs from local statistics [1249]; even during the AIDS crisis in  
 the early 1990s, there were counties with only one single patient receiving such  
 treatment.

Some systems try to get round the limits imposed by static query control

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |
| --- |
| (A, B, C)  Prohibited  (A, B) (B, C) (C, A)  A B C  Allowable  U |

Figure 11.3: – table lattice for a database with three attributes

by keeping track of who accessed what. Known as *query overlap control*, this  
 involves rejecting any query from a user that, combined with what the user  
 knows already, would disclose a sensitive statistic. This may sound like a good  
 idea, but in practice it suffers from two usually insurmountable drawbacks.  
 First, the complexity of the processing involved increases over time, and often  
 exponentially. Second, it’s extremely hard to be sure that your users don’t

collude, or that one user has registered under two different names. Even if your  
 users are all honest and distinct persons today, it’s always possible that one of  
 them will get taken over tomorrow.

**11.2.1.6** **Randomization**

By now it should be clear that if various kinds of query control are the only  
 protection mechanisms used in a statistical database, they will often impose  
 an unacceptable statistical performance penalty. So query control is often used  
 in conjunction with various kinds of randomization, designed to degrade the  
 signal-to-noise ratio from the attacker’s point of view while impairing that of  
 the legitimate user as little as possible.

Until 2006, all the methods used were rather ad hoc. They started with

*perturbation*, or adding noise with zero mean and a known variance to the data;  
 but this tends to damage the legitimate user’s results precisely when the sam-  
 ple set sizes are small, and leave them intact when the sample sets are large  
 enough to use simple query controls anyway. A later variant was *controlled tab-*  
 *ular adjustment* where you identify the sensitive cells and replace their values  
 with different ones, then adjust other values in the table to restore additive  
 relationships [490]. Then there are *random sample queries* where we make all  
 the query sets the same size, selecting them at random from the available rel-  
 evant statistics. Thus, all the released data are computed from small samples  
 rather than from the whole database, and we can use a pseudorandom number

|  |  |  |
| --- | --- | --- |
|  |  |  |

generator keyed to the input query to make the results repeatable. Random  
 sample queries are a natural protection mechanism where the correlations being  
 investigated are strong enough that a small sample is sufficient. Finally, there’s  
 *swapping*, another of Tore Dalenius’ innovations; many census bureaux swap  
 a proportion of records so that a family with two young teenage kids and an  
 income in the second quartile might be swapped for a similar family in a town  
 in the next county.

Since 2006, we have a solid theory of exactly how much protection we can

get from adding randomness: *differential privacy*. This is now being used for  
 the 2020 US census, and we’ll discuss it in more detail later in this chapter.

**11.2.2** **Limits of classical statistical security**

As with any protection technology, statistical security can only be evaluated in  
 a particular environment and against a particular threat model. Whether it is  
 adequate or not depends on the details of the application.

One example is a system developed in the mid-1990s by a company then

called Source Informatics for analysing trends in drug prescribing, which ﬁgured  
 in the key UK lawsuit about the privacy of anonymised data1. The system’s  
 goal is to tell drug companies how effective their sales staff are, by tracking  
 sales of different medicines by district. The privacy goal was to not leak any  
 information about identiﬁable patients or about the prescribing habits of indi-  
 vidual physicians2. So prescriptions were collected (minus patient names) from  
 pharmacies, and then a further stage of de-identiﬁcation removed the doctors’  
 identities too.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Week: |  |  |  |  |
| Doctor A  Doctor B  Doctor C  Doctor D |  |  |  |  |

Figure 11.4: Sample of de-identiﬁed drug prescribing data

The ﬁrst version of this system merely replaced the names of doctors in a cell

of four or ﬁve practices with ‘doctor A’, ‘doctor B’ and so on, as in Figure 11.4.  
 When evaluating it, we realised that an alert drug rep could identify doctors  
 from prescribing patterns: “Well, doctor B must be Susan Jones because she  
 went skiing in the third week in January and look at the fall-off in prescriptions  
 here. And doctor C is probably Mervyn Smith who was covering for her”. The  
 ﬁx was to replace absolute numbers of prescriptions with the percentage of each  
 doctor’s prescribing which went on each particular drug, to drop some doctors at  
 random, and to randomly perturb the timing by shifting the ﬁgures backwards  
 or forwards a few weeks [1249].

1Full disclosure: I was the evaluator, acting on behalf of the British Medical Association.  
 2Doctors are hounded all the time by drug sales reps and often say they’ll use some product

or other just to get them out of the surgery. It’s curious that such an important privacy case  
 had as its privacy objective a doctor’s ability to continue telling white lies.

|  |  |  |
| --- | --- | --- |
|  |  |  |

This is a good example of the sort of system where classical statistical secu-

rity techniques can give a robust solution. The application is well-deﬁned, the  
 database is not too rich, the allowable queries are fairly simple, and they remain  
 stable over time. Even so, the UK Department of Health sued the database oper-  
 ator, alleging that the database might compromise privacy. The Department’s  
 motive was to maintain a monopoly on the supply of such data to industry.  
 They lost, and this established the precedent that (in Britain at least) inference  
 security controls may, if they are robust, exempt statistical data from being  
 considered as ‘personal information’ for the purpose of privacy laws [1804].

In general, though, it’s not so easy. For a start, privacy mechanisms don’t

compose: it’s easy to have two separate applications, each of which provides the  
 same results via perturbed versions of the same data, but where an attacker with  
 access to both of them can easily identify individuals. This actually happened  
 in the Source Informatics case; by 2015, another competing system was avail-  
 able that used different mechanisms, and people realised that a drug company  
 with access to both systems could occasionally deduce some doctors’ prescribing  
 behaviour. If we were re-implementing such a system today, we’d prevent this  
 by using differential privacy, which I’ll describe later in this chapter.

**11.2.2.1** **Active attacks**

The Source Informatics system added a new tranche of records every week, but  
 it can sometimes happen that users have the ability to insert single identiﬁable  
 records into the database. In that case, *active attacks* can be particularly pow-  
 erful. A prominent case in the late 1990s was a medical research database in  
 Iceland. A Swiss drug company funded a local startup to offer the Reykjavik  
 government a deal: we’ll build you a modern health cards system if you’ll let us  
 mine it for research. The government signed up, but Iceland’s doctors mostly  
 opposed the deal, seeing it as a threat both to patient privacy and professional  
 autonomy.

Under their proposed design, every time a medical record was generated, it

would be sent to the Iceland privacy commissioner whose system would strip  
 out the patient’s name and address, replacing it with an encrypted version of  
 their Social Security number, and pass it to a research database. The privacy  
 commissioner controlled the encryption key. However, anyone in the system who  
 wanted to ﬁnd (say) the Prime Minister’s medical records would merely have to  
 enter some record or other – say a prescription for aspirin – and then watch it  
 pop up on the research system a second or two later. The Icelandic government  
 pressed ahead anyway, with a patient opt-out. Many doctors advised patients  
 to opt out, and 11% of the population did so. Eventually, the Icelandic Supreme  
 Court found that European privacy law required the database to be opt-in rather  
 than opt-out, which put paid to the project.

Iceland was particularly attractive to researchers as the population is very

homogeneous, being descended from a small number of settlers a thousand years  
 ago, and there are good genealogical records. This also made privacy problems  
 in the Icelandic database more acute. By linking medical records to genealogies,  
 which are public, patients can be identiﬁed by such factors as the number of  
 their uncles, aunts, great-uncles, great-aunts and so on – in effect by the shape

|  |  |  |
| --- | --- | --- |
|  |  |  |

of their family trees. There was much debate about whether the design could  
 even theoretically meet legal privacy requirements [66], and European privacy  
 officials expressed grave concern about the possible consequences for Europe’s  
 system of privacy laws [515]. This brings us to the broader question of rich  
 contextual data, which drove the second wave of work on inference control.

**11.2.3** **Inference control in rich medical data**

The second half of the 1990s saw the ‘dotcom boom’. The worldwide web was  
 new, and a torrent of money ﬂowed into tech as businesses (and governments)  
 tried to ﬁgure out how to move their operations online. Healthcare IT people  
 struggled with many questions around safety and privacy; records had already  
 been moving from paper to computers, but now all the computers started talking  
 to each other [63]. Could you use email to send test results from a hospital to  
 a doctor’s surgery, or would it be a web form? How would you encrypt it, and  
 who’d manage the keys? And could you make complete medical records safe  
 enough for use in research by removing names and addresses, as opposed to  
 just episode data such as individual prescriptions? Researchers had previously  
 done epidemiology by sitting in hospital libraries reading paper records, and  
 it would ‘obviously’ be better if you could do this at your desk. However, an  
 epidemiologist will usually want to be able to link up episodes over a patient’s  
 lifetime, so they can see long-term effects of treatments and lifestyle choices.  
 That is much harder to anonymise.

Health IT people faced this problem in many countries at once. New Zealand

set up a database with encrypted patient names plus a rule that no query may  
 answered with respect to fewer than six records, but realised that that was  
 not enough and restricted access to a small number of specially cleared med-  
 ical statisticians [1422]. The fall of the Berlin Wall caused an acute problem  
 for Germany, as the former East Germany had cancer registries with ﬁrst-class  
 data that were really useful for research but had patient names and rich con-  
 textual data, and these now fell under West Germany’s strict privacy laws. The  
 registry had to install protection mechanisms rapidly, which involve both de-  
 identiﬁcation and strict usage controls [266]. In Switzerland too, some research  
 systems were replaced at the insistence of privacy regulators [1681]. The British  
 Medical Association objected to a proposal for a centralised research database  
 in 1995–6 and a committee was set up under an eminent psychiatrist, Dame  
 Fiona Caldicott, to suggest a way forward.

The fact that the rich context of medical records had changed the statis-

tical security game was then brought into focus in 1997 by Latanya Sweeney  
 who tried, in her PhD thesis, to build a system that would anonymise medical  
 records properly, and discovered how hard it is. She showed that even the Health  
 Care Finance Administration’s ‘public-use’ ﬁles could often be re-identiﬁed by  
 cross-correlating them with commercial databases [1849]. She showed that 69%  
 of U.S. residents can be identiﬁed by date of birth and zip code, and discussed  
 the extreme difficulty of scrubbing medical records that contain all sorts of  
 contextual data, including free-form text [1849]. At the time, the Medicare

system considered *beneﬁciary-encrypted* records – with patients’ names and So-  
 cial Security numbers encrypted – to be personal data and thus only usable by

|  |  |  |
| --- | --- | --- |
|  |  |  |

trusted researchers. There were also *public-access* records, stripped of identiﬁers  
 down to the level where patients are only identiﬁed in general terms such as ‘a  
 white female aged 70–74 living in Vermont’. Nonetheless, researchers have found  
 that many patients can still be identiﬁed by cross-correlating the public access  
 records with commercial databases. Sweeney brought this to public attention  
 by identifying the records of Massachusetts governor William Weld. This got  
 the anonymity of medical research data on to the US political agenda.

As I describe in section 10.4, the Clinton administration issued a privacy

rule in 2000 under HIPAA that deﬁned a ‘Safe Harbor’ standard for the public  
 sharing of data, and then in 2002 the Bush administration adopted a more  
 relaxed rule. In 2017 Sweeney and colleagues examined a 2006 public-health  
 study of 50 homes in California, which had been cited hundreds of times in the  
 research literature, and showed they could identify 25% of the participants by  
 name and 28% by address [1850]. Even after redacting participants’ birth years  
 to 10-year ranges, they could still pinpoint 3% by name and 18% by address –  
 because of side information such as the type of housing.

The UK followed a similar trajectory. Dame Fiona Caldicott’s report identi-

ﬁed over sixty illegal information ﬂows within the health service [367]. Some re-  
 search datasets were de-identiﬁed very carelessly; others (including data on peo-  
 ple with HIV/AIDS) were re-identiﬁed deliberately afterwards, so that people  
 and HIV charities whose data had been collected under a promise of anonymity  
 were deceived. Parliament then passed a law giving ministers the power to

regulate secondary uses of medical data, but the broad direction was trusted  
 researchers; a committee vetted applications for data access. Patient consent  
 was obtained in some cases, but not for research involving the Hospital Episode  
 Statistics database, which contains records of over a billion hospital treatments  
 in England and Wales from 1998 to the present day. HES data are made avail-  
 able to researchers with the patient’s name and address removed and replaced  
 with an encrypted identiﬁer. (The encryption key is different for each research  
 organisation that licenses the data.)

But encrypting patient names isn’t enough. Suppose I want to look up the

record of former Prime Minister Tony Blair. A quick web search reveals that  
 he was treated in Hammersmith Hospital in London for an irregular heartbeat  
 on 19th October 2003 and 1st October 2004. That’s more than enough to pick  
 out his encrypted ID and look up everything else he’s had done. Such a leak  
 can be intrusive for anybody; for a celebrity, it can be newsworthy. What’s  
 more, in many systems there’s a cleartext postcode and date of birth; again,  
 this combination is enough to identify about 98% of UK residents3. Even if  
 the date of birth is replaced by a year of birth, I am still likely to be able to  
 compromise patient privacy if the records are detailed, or if records of different  
 individuals can be linked. For example, a query such as ‘show me the records  
 of all women aged 36 with daughters aged 14 and 16 such that the mother  
 and exactly one daughter have psoriasis’ can ﬁnd one individual out of millions.  
 Query set size control might stop this kind of tracker, but researchers do want  
 to make complex queries with lots of conditions to ﬁnd disease clusters with a

3UK postcodes have more resolution than US zip codes, with typically 30 buildings in

each postcode. The 1% or so of people for whom postcode plus date of birth is not unique are  
 mostly identical twins, or young people living in college halls of residence or military barracks.

|  |  |  |
| --- | --- | --- |
|  |  |  |

few hundreds or even a few dozens of patients. Such queries could be composed,  
 whether deliberately or by accident, in such a way as to identify individuals.

In 2006, UK privacy groups organised a campaign to alert people to the

|  |
| --- |
| risks and invite them to exercise their right to opt out of secondary data use.  In 2007, Parliament’s Health Select Committee conducted an inquiry into the  Electronic Patient Record, heard evidence from a wide range of viewpoints4 and  made many recommendations, including that patients should be permitted to  prevent the use of their data in research [925]. Privacy concerns are not the only  reason that a patient might reasonably request that their data not be used; for  example, a devout Catholic woman might demand that her data not be used to  develop pills for abortion or birth control. The Government rejected this. |

David Cameron’s government, elected in 2010, weakened privacy protection,

just as George Bush had done ten years earlier. Amidst talk of abolishing red  
 tape and making the UK the best place in the world for medical research, as I  
 discussed at greater length in section 10.4.4.3, he launched ‘care.data’, a central  
 research database that would add test results, prescriptions and GP data to the  
 existing HES database. In November 2013 it emerged that HES data were avail-  
 able via BT for sale online [948], and in February 2014, it emerged that copies  
 of the HES database had been sold to 1,200 organisations worldwide, including  
 not just academic researchers but commercial ﬁrms, from drug companies to  
 consultancies [774]. One of the big US consultancies had uploaded all 23GB of  
 data to the Google cloud ‘as it was too big for Excel’ and was making it available  
 to clients, despite laws that required the data to remain in the UK. The data  
 had been used for non-health purposes, speciﬁcally by actuaries to reﬁne insur-  
 ance premiums. A law was quickly passed stating that health and social data  
 could be shared and analyzed only when there was a ‘beneﬁt to healthcare’, and  
 never for other purposes. Another consultancy was hired to produce another  
 report, and people who’d opted out were told to opt out all over again. An aca-  
 demic case study tells the story, analyses the tensions between healthcare law  
 and data-protection law, and remarks that ‘this debate centers on the ability  
 to protect and maintain the anonymity of patient data, and there are no easy  
 answers’ [1548].

**11.2.4** **The third wave: preferences and search**

The next wave broke in 2006, by which time a signiﬁcant number of transac-  
 tions had moved online, recommender systems had emerged thanks to eBay and  
 Amazon, and search engines made it easy to ﬁnd needles in haystacks. Two  
 incidents that year brought this home to the public.

First, AOL released the supposedly anonymous records of 20 million search

queries made over three months by 657,000 people. Searchers’ names and IP  
 addresses were replaced with numbers, but that didn’t help. Investigative jour-  
 nalists looked through the searches and rapidly identiﬁed some of the searchers,  
 who were shocked at the privacy breach [167]. The data were released ‘for re-  
 search purposes’: the leak led to complaints being ﬁled with the FTC, following  
 which the company’s CTO resigned, and the ﬁrm ﬁred both the employee who

4Declaration of interest: I was a Special Adviser to the Committee.

|  |  |  |
| --- | --- | --- |
|  |  |  |

released the data and their supervisor. Search history, or equivalently your

clickstream, is highly sensitive as it reﬂects your thoughts and intentions.

Second, Netﬂix offered a $1m prize for a better recommender algorithm and

published the viewer ratings of 500,000 subscribers with their names removed.  
 At the time, it had only 6 million US customers and shipped them physical  
 DVDs, so this was a signiﬁcant minority of its customers. Arvind Narayanan  
 and Vitaly Shmatikov showed that many subscribers could be reidentiﬁed by  
 comparing the anonymous records with preferences publicly expressed in the  
 Internet Movie Database [1384]. This is partly due to the ‘long tail’ effect:

once you disregard the 100 or so movies everyone watches, people’s viewing  
 preferences are pretty unique. As US law protects movie rental privacy, the  
 attack was a serious embarrassment for Netﬂix.

The response of privacy regulators in Europe and Canada was to promote

*Privacy Enhancing Technologies* (PETs) – they hoped that if security researchers  
 were to work harder, we could come up with more effective ways of anonymising  
 rich data [649]. Researchers at Microsoft took them at their word, and developed  
 the theory of differential privacy, which I explain in 11.3. This does not get the  
 privacy regulators off the hook, as it clariﬁes the limitations of anonymisation.  
 Yet for years policy people talked about it as a solution without understand-  
 ing that it explains in more detail why we cannot resolve the tension between  
 researchers’ demand for detailed data, and the right of data subjects to privacy.

**11.2.5** **The fourth wave: location and social**

During the 2010s, the world was changed by smartphones and social networks.  
 Chapter 23 in the second edition of this book in 2008 describes the early social  
 network scene, as Facebook was just taking over from Myspace. I noted that  
 Robert Putman’s book ‘Bowling Alone’ had documented the decline of social  
 engagement through voluntary associations such as churches, clubs and societies  
 with the arrival of TV in the 1960s [1563], and the fact that the Internet’s  
 early Usenet newsgroups and mailing lists had managed to put some of that  
 back. The sweet spot the social networks hit was rolling this out to everybody.  
 However recondite your interests, you can connect with people who share them,  
 wherever in the world they are. We predicted that social networks would bring  
 all sorts of privacy problems directly, as social context makes it hard to hide. (Is  
 there anyone other than me who hangs out with cryptographers, with digital-  
 rights activists, and with people interested in the dance music of 200 years  
 ago?) Persistence adds further hazards, as when teens’ boasts about sex and  
 drugs come back to haunt them later in job interviews. Two things we missed  
 were the fact that masses of data have migrated to the cloud, and the sheer  
 amount of sensitive personal information that can be deduced from contextual  
 data about people. By 2011 Google was describing its core competence as ‘the  
 statistical data mining of crowdsourced data’; as the datasets got larger, and  
 basic statistical techniques were augmented with machine learning, the amount  
 we can learn has grown.

An example of ‘more data’ is location history. By 2012, Yves-Alexandre de

Montjoye and his colleagues had shown that four mobile-phone locations are  
 in general enough to identify someone, even when you only get their cell-tower

|  |  |  |
| --- | --- | --- |
|  |  |  |

location [1333]. Nowadays much more high-resolution data are widely available,  
 as many smartphone apps ask for access to your location – which can involve not  
 just GPS (with an average accuracy of perhaps 8m outdoors) but also which wiﬁ  
 hotspots are in range (which can tell where you are in a building). Most people  
 click to agree without a second thought, and there’s now a whole ecoystem of  
 companies buying and selling location trace data – which is now accurate to a  
 few metres rather than a few hundred. The data were sold not just to marketing  
 ﬁrms, but to private detectives, including bounty hunters who use it to track  
 down people who’ve jumped bail [489].

In December 2019 the New York Times got hold of the location traces of

12 million Americans over a few months and demonstrated graphically how  
 closely people can now be tracked. Your daily trace shows your home, when you  
 left, how you traveled to work, where you stopped for a coffee en route, where  
 your office is, where you went for lunch – everything. The journalists found  
 in their database a celebrity who had sung at a church service for President  
 Trump; hundreds of people working at the Pentagon and the CIA, as well as  
 the President’s Secret Service bodyguards, all of whom they could follow home;  
 and people visiting the sex industry. They found one man who’d worked at  
 Microsoft, then visited Amazon, then started working at Amazon the following  
 month. They looked at a riot, and found they could follow both rioters and  
 police officers home [1885]. There’s a stark contrast between the ease of buying  
 this data on the open market, and the hoops that law enforcement have to jump  
 through to get it by means of warrants. The location data companies all claim  
 that their data are anonymous; yet even though they might not actually use the  
 phone book or the voters’ roll to look up your name from your street address,  
 several sell your location data tied to an advertising identiﬁer based on one or  
 more cookies in your browser. With low-resolution location data, when you go  
 to Black Hat in Las Vegas, online gambling companies can put ads in front of  
 you. With high-resolution data, a foreign intelligence agency could locate people  
 who work at the Pentagon and also visit gay clubs or brothels. It can also follow  
 them home.

An example of ‘better inference’ comes from the behavioural analysis of

social-network data. The headline case here started when Michal Kosinski and  
 colleagues wrote a Facebook app that offered free psychometric testing and  
 persuaded tens of thousands of people to use it. They ﬁgured out that they  
 could tell whether someone was straight or gay from four Facebook likes; given  
 sixty likes, they could assess the user’s ‘Big Five’ personality traits: whether  
 you are open to experience or cautious, conscientious or easygoing, extravert or  
 introvert, agreeable or detached, and neurotic or conﬁdent [1086]. They can also  
 tell whether you’re white or black, conservative or liberal, Christian or Muslim,  
 whether you smoke, whether you drink, whether you’re in a relationship, and  
 whether you use drugs – with varying degrees of accuracy. This led some of  
 his colleagues to collect Facebook data on an industrial scale for marketing and  
 political campaigning, leading to the Cambridge Analytica scandal, which I’ll  
 discuss in Part 3. Later research showed that having behavioural data gives  
 publishers only an extra 4% of ad income compared with what they get over  
 contextual ads, so conceivably this practice might simply be banned [1239].  
 However, industry observers note that the platforms earn more than this, as  
 they get the lion’s share of ad income – so they can be expected to resist any

|  |  |  |
| --- | --- | --- |
|  |  |  |

such privacy law [1181].

In many cases, you can get both location data and social data, and get

them at scale. For example, the government of Victoria, Australia, made public  
 a database of transport ticket use covering a billion journeys by 15m tickets  
 from 2015–8. Although the card IDs had been anonymised, it usually took

only one or two journeys for a resident to identify their own card from the  
 touch on and touch off times; researchers found they could then identify their  
 co-travelers [502]. Next they identiﬁed people using Australian federal parlia-  
 mentary passes, who routinely get the train to their constituencies; hypotheses  
 could be conﬁrmed from the parliamentarians’ tweets. This dataset enabled the  
 researchers to analyse the sensitivity of travel time. They found that even if  
 travel times were truncated to the day, with hours and minutes thrown away,  
 four locations would identify over a third of travelers.

We now have many social side channels as well as location data. Location

history leaks so much data as it reveals who we live with, work with and party  
 with. Social networks are even richer with our contacts, preferences and selﬁes,  
 and can make these measurements more accurate. And social analysis can reach  
 right down into the lowest layers of the stack. For example, it turns out to be  
 fairly easy to match up two social graphs, even if they are not exact copies of  
 each other; so given a country’s anonymised mobile phone call data records, you  
 can re-identify them by comparing them with (say) the friend graph of a social  
 network [1719]. Mobile phone data already leak lots of information about our  
 personalities: extraverts make more calls, agreeable people get more calls, and  
 the variance of time between phone calls predicts conscientiousness [1334].

The combination of more data and better inference led to fresh controversy

in medical research too. Google’s AI subsidiary DeepMind announced a collab-  
 oration in 2016 with a London hospital to develop an app to diagnose kidney  
 injury. The following year, it turned out that the hospital had given DeepMind  
 not just the records of kidney injury sufferers, but all 1.6m fully-identiﬁable  
 records of all its patients, without getting their consent [1542]. The privacy  
 regulator reprimanded the hospital, as such access should be given only to ﬁrms  
 involved in direct patient care rather than for product research; however it did  
 not attempt to force DeepMind to delete the data. The company used VA data  
 from the US instead to develop diagnostic apps. It did set up an Ethics Board  
 that it claimed would control the technology, and did undertake not to give the  
 hospital data to its parent Google, but in 2017 an eminent member of the ethics  
 board resigned claiming it was window-dressing, and in 2018 it was announced  
 that Google was absorbing DeepMind’s health operation [909]. This slow train  
 wreck was followed by the news that Google was already under ﬁre for acquiring  
 the records of 50 million US patients [121].

So is it possible to do anonymisation properly? The answer is yes; in certain

circumstances, it is. Although it is not possible to create anonymous datasets  
 that can be used to answer any question, we can sometimes provide a dependable  
 measure of privacy when we set out to answer a speciﬁc set of research questions.  
 This brings us to the theory of differential privacy.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**11.3** **Differential privacy**

In 2006, Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith  
 published a seminal paper showing how you could systematically analyse pri-  
 vacy systems that added noise to prevent disclosure of sensitive statistics in a  
 database [595]. Their theory, *differential privacy*, enables the security engineer  
 to limit the probability of disclosure, even in the presence of an adversary with  
 unbounded computational power and copious side information, and can thus be  
 seen as the equivalent of the one-time pad and unconditionally secure authen-  
 tication codes in cryptography. Although it started as a paper on theoretical  
 cryptography, it has come to be seen as the gold standard for both statistical  
 database security and for anonymisation in general. The starting point was an  
 earlier paper by Kobbi Nissim and Irit Dinur, who had shown in 2003 that if  
 queries on a database each returned an approximation to a linear function of  
 private bits of information, then so long as the error was small enough the num-  
 ber of queries required to reconstruct the database would not grow too quickly;  
 such reconstruction attacks are, after all, based on linear algebra, so rather than  
 making carefully targeted tracker attacks, an attacker can just make a whole  
 lot of random queries, then do the algebra and get everything out [562]. So the  
 defender has to add noise if there will be more than a limited number of queries,  
 and the question is how much.

The key insight of differential privacy is that, to avoid inadvertent disclosure,

|  |
| --- |
| no individual’s contribution to the results of queries should make too much of a  di↵erence, so you calibrate the standard deviation of the noise according to the  sensitivity of the data. A privacy mechanism is called *✏*-indistinguishable if for  all databases *X* and *X0* di↵ering in a single row, the probability of getting any  answer from *X* is within a multiplicative factor of (1 + *✏*) of getting it from *X0*;  in other words, you bound the logarithm of the ratios. It follows that you can  use noise with a Laplace distribution to get indistinguishability with noisy sums,  and things compose, so it all becomes mathematically tractable. The value of *✏*,  which sets the trade-o↵ between accuracy and privacy, has to be set by policy.  Small values give strong privacy; but setting *✏* = 1000 is basically publishing  your raw data. |

There is now a growing research literature exploring how such mechanisms

can be extended for static to dynamic databases, to data streams, to mecha-  
 nism design and to machine learning. But can the promise of learning nothing  
 useful about individuals while learning useful information about a population,  
 be realised in practical applications?

**11.3.1** **Applying differential privacy to a census**

Differential privacy is now getting a full-scale test in the 2020 U.S. census.  
 The census is not allowed to publish anything that identiﬁes the data of any  
 individual or establishment; collected data must by law be kept conﬁdential for  
 72 years and used only for statistical purposes until then. First, the Census  
 Bureau reviewed the security of the 2010 census in the light of modern analysis  
 tools [752]. In 2010, the aggregated *census edited ﬁle* (CEF) of data collected  
 from US residents and then edited to get rid of duplicates and ﬁll in missing

|  |  |  |
| --- | --- | --- |
|  |  |  |

entries from data such as tax returns, had 44 bits of conﬁdential data on each  
 resident (a total of 1.7Gb). The problem is that the microdata summaries simply  
 contained a lot more data than this; writing everything out, you get several  
 billion simultaneous equations and can in theory solve for the conﬁdential data.

What about in practice? Census staff implemented ideas based on Kobbi

Nissim and Irit Dinur’s work, and found that they got all the variables right  
 about 38% of the time, covering a bit under 20% of the population. It took one  
 month on four servers, so it’s not entirely trivial. However, the lesson is that the  
 traditional approaches to statistical database security don’t really work. They  
 did provide some privacy, because the 2010 census swapped very identiﬁable  
 households with other blocks, so not everyone was compromised. If they’d

swapped all the households, it would have been OK, but the users wouldn’t  
 have put up with that; the fact that they gave exact population counts for a  
 block was a real vulnerability. Dealing with database reconstruction piecemeal  
 is hard; that’s the value of differential privacy.

The big policy question is where you set *ffl*. This is also an empirical question.

In 2018, census staff did an end-to-end test reporting four tables. In 2020 the  
 full system will process the CEF into a *microdata details ﬁle* (MDF) from which  
 the tabulations will be derived. Foreseeable issues include that numbers won’t  
 add up; so the number of members of the separate Native American tribes won’t  
 add up to the total of Native Americans, and that will have to be explained to  
 the public. The differential-privacy approach will protect everyone, while the  
 old system only protected people who were swapped, and it has to be done all  
 at once. Every record may be modiﬁed subject to an overall privacy budget, so  
 there’s no exact mapping between the CEF and the MDF.

The new top-down algorithm generates a national histogram without ge-

ographic identiﬁers, then sets out to build a geographic histogram top-down,  
 such that the state ﬁgures add up to the national ﬁgures (which is needed for  
 Congressional redistricting). The construction is then done recursively down  
 through state, county, tract, block group and block, after which they generate  
 the microdata. This can be done in parallel and enables sparsity discovery (e.g.  
 there are very few people over 100 belonging to 5 or more races). The top-down  
 approach turns out to be much more accurate than applying noise block-by-  
 block, in that county data have less error than blocks, and national data have  
 essentially no error. There are several edge cases needing special handling: a  
 prison won’t be turned into a college dorm, but if there are ﬁve dorms, you  
 might report four or six. Person-household joins are also hard; you can do the  
 number of men on a block, or the number of households, but the number of  
 children in households headed by a single man is more sensitive. But many  
 things that used to be suppressed no longer have to be; you no longer have to  
 enumerate all the sources of side information that might be used; and there will  
 at last be published error statistics.

Now that the outline design has been done, there’s a simulator you can use to

explore possible values of *ffl*. You can plug this into an economic analysis of the  
 tradeoff between the marginal social beneﬁt of better stats with the marginal  
 social costs of identity theft [928]; the outcome suggests a value of *ffl* between 4  
 and 6.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**11.4** **Mind the gap?**

On the political side, the use of lightly-deidentiﬁed data in research, whether  
 medical research or market research, has involved sporadic guerilla warfare be-  
 tween privacy advocates and data users for years, with regulators usually siding  
 with the data users except in the aftermath of a scandal. The regulators are  
 both overwhelmed and conﬂicted, as I’ll describe in section 26.6.1, and mostly  
 do not have the political support to take on big Internet service ﬁrms or govern-  
 ment departments. These ‘Big Data’ interests are generally adept at capturing  
 regulators anyway. For example, in 2008 Prime Minister Gordon Brown asked  
 the UK Information Commissioner and the head of Britain’s largest medical-  
 research charity to come up with guidelines on using data in research; they  
 ignored privacy rights, took an instrumental view of costs and beneﬁts, and  
 spun the secondary use of data as ‘data sharing’. As you might expect, neither  
 privacy lawyers nor security academics were pleased with the result [96].

In 2009 a highly inﬂuential paper, ‘Broken promises of privacy’, was written

by Paul Ohm, a distinguished US law professor [1465]. He noted that “scientists  
 have demonstrated they can often ‘reidentify’ or ‘deanonymize’ individuals hid-  
 den in anonymized data with astonishing ease” and confessed “we have made a  
 mistake, labored beneath a fundamental misunderstanding, which has assured  
 us much less privacy than we have assumed. This mistake pervades nearly ev-  
 ery information privacy law, regulation, and debate, yet regulators and legal  
 scholars have paid it scant attention.” For the previous thirty years, computer  
 scientists had known that anonymisation doesn’t really work, but law and pol-  
 icy people had stopped their ears. Here at last was an eminent lawyer spelling  
 out the facts, telling the story of AOL and Netﬂix, in a law journal and us-  
 ing lawyer-accessible language. Among other things he ridiculed Google’s claim  
 that IP addresses were not personal information (it argued that its search logs  
 should therefore fall outside the scope of data protection), denounced the binary  
 mindset of data as either personal or not, and called for a more realistic debate  
 on privacy and data protection. Might this change things?

In 2012, a report from the Royal Society called for scientists to publish their

data openly where possible but acknowledged the reality of re-identiﬁcation  
 risks: ‘However, a substantial body of work in computer science has now demon-  
 strated that the security of personal records in databases cannot be guaranteed  
 through anonymisation procedures where identities are actively sought’ [1627].  
 In that year, the UK Information Commissioner also developed a code of prac-  
 tice on anonymisation [80]; as the ICO is the privacy regulator, such a code  
 can shield ﬁrms from liability, and it was the target of vigorous lobbying. The  
 eventual code required data users to only describe their mechanisms in general  
 terms, and shifted the burden of proof on to anyone who objected [81]. This was  
 a less stringent burden than the ICO applies in freedom-of-information cases,  
 where a request for public data can be refused on the presumption that the  
 data subjects’ ‘friends, former colleagues, or acquaintances’ may know relevant  
 context. This tiptoes round a concept of some relevance to tactical anonymity –  
 the *privacy set*, or the set of people whom I might want to not know some fact  
 about me. For most people, this is your family, friends and work colleagues –  
 perhaps 100–200 people. For celebrities, it can be everybody; and problems can

|  |  |  |
| --- | --- | --- |
|  |  |  |

arise when someone suddenly becomes famous. Most of us can be anonymous  
 in a big city, but a celebrity can’t.

Another useful but quite different concept is the *anonymity set*, which is the

set of people with whom you might be confused. We’re all familiar with detective  
 ﬁlms or novels, where Poirot steadily reduces the number of people who might  
 have committed the murder from a dozen to one. Strategic mechanisms like  
 differential privacy focus on keeping the anonymity set large enough, while many  
 tactical mechanisms assess the risk that people with access to some application  
 will overlap your privacy set.

But you always have to think carefully about the threat model. While it may

be enough to worry about your privacy set when the concern is embarrassment,  
 when it’s scam artists you need to worry about the anonymity set. As we

noted in Chapter 3, phishing attacks often involve information leaks about the  
 victim that enable an attacker to impersonate the victim to some service, or  
 impersonate the service to the victim. In short, when it comes to phishing,  
 anyone who can tie your identity to some relevant context may be able to attack  
 you.

**11.4.1** **Tactical anonymity and its problems**

The ICO also set up the UK Anonymisation Network (UKAN), which is coor-  
 dinated by academics and by the Office of National Statistics. In 2016 UKAN  
 produced a book of guidance on how ﬁrms should make decisions on anonymi-  
 sation, duly signed off by the ICO [626]. Its authors see conﬁdentiality as being  
 about risk rather than duty; decisions have to be taken not just according to the  
 technical possibility of identifying data subjects but the institutional and social  
 context that determines whether this might be attempted. The threat model  
 should be based on plausible intruder scenarios. They talk of governance pro-  
 cesses rather than side channels; they dismiss differential privacy as ‘extreme’;  
 they see anonymisation as a process and advise against using ‘success terms’  
 like ‘anonymised’; and they deﬁne ‘de-identiﬁed’ as ‘can’t be re-identiﬁed from  
 the data directly’. Measures to manage re-identiﬁcation risk should be pro-

portional to risk and its likely impact; and anonymisation measures may have  
 a limited lifetime because of eventual triangulation from other datasets. Such  
 mechanisms therefore have to be seen as tactical anonymity, as opposed to the  
 strategic anonymity that is being carefully engineered into the US census. The  
 UKAN authors do not seem to have considered differential privacy seriously.

Despite its ﬂaws, the UKAN framework requires attention if you’re going

to rely on anonymisation, whether tactical or strategic, in the UK, as it’s the  
 yardstick by which the regulator will decide whether or not to take enforcement  
 action against you. It is likely to provide a shock absorber and liability shield for  
 both data users and regulators as anonymisation becomes steadily less effective.  
 It would have provided some protection for ﬁrms that based their EU operations  
 in the UK, but with Britain having left the EU this will no longer hold. It does  
 however contain a reasonable amount of practical advice on assessing the risks of  
 tactical anonymisation in applications where both the data and the environment  
 are reasonably well understood. As a result, there are now several ﬁrms whose  
 products and services aim at helping data users comply with it.

|  |  |  |
| --- | --- | --- |
|  |  |  |

An example of a ﬁrm operating openly under this framework is the mobile

network operator Vodafone, which sells ‘location insight’ products. The com-  
 pany aggregates the mobile phone locations of its customers into journeys with  
 implied origin, destination and mode of transport. The origin-destination ma-  
 trices are sold to local government and transport ﬁrms along with ﬂows along  
 main roads and railways. The privacy mechanisms consist of ﬁrst, allowing all  
 subscribers an opt-out and second, encrypting phone IMSIs to give a different  
 pseudonym per device, with a slowly changing key; the cell towers are easily re-  
 identiﬁable. One can indeed make an argument that the risk here is low; maybe  
 the analysts at the local council or bus company can identify you, especially if  
 you live in a small hamlet (as I do; four houses 200m from the nearest village).  
 So the anonymity set can be too small. Then you have to look at the privacy  
 set size. But suppose you work at a ﬁrm that becomes a target for activists. If  
 they recruit someone at the council, they could target company staff who live  
 in isolated houses in order to intimidate them or their families5.

The practical problems that have become evident have to do ﬁrst with scale

and second with the inherent conﬂicts of self-regulation. The scale is evident  
 not just in the number of data sources that might be matched externally to  
 identify people, but in the growing size and complexity of organisations’ internal  
 data warehouses too. A decisive factor has been Hadoop6: a ﬁrm can now

store everything, so it’s hard to keep track of what’s stored. As there are no  
 database schemas but the data are just piled up, you have no idea of linkage  
 risks, especially if your ﬁrm has a multitenant cluster with all sorts of stuff from  
 different subsidiaries. Such data warehouses are now used for fraud prevention,  
 customer analytics and targeted marketing. Firms want to be responsible, but  
 how do you give live data to your development and test teams? How can you  
 collaborate with academics and startups? How can you sell data products?

Anonymisation technology is all pretty rudimentary at this scale, and as you  
 just don’t know what’s going on, it’s beyond the scope of differential privacy or  
 anything else you can analyse cleanly. You can tokenise the data on ingest to  
 get rid of the obvious names, then control access and use special tricks for time  
 series and location streams, but noise addition doesn’t work on trajectories and  
 there are lots of creative ways to re-identify location data (e.g. photos of celebs  
 getting in and out of taxis). Things get even harder where people are partially  
 authorised and have partial access.

Future problems may come from AI and machine learning; that’s the fash-

ion now, following the ‘Big Data’ fashion of the mid-2010s that led ﬁrms to  
 set up large data warehouses. You’re now training up systems that generally  
 can’t explain what they do, on data you don’t really understand. We already  
 know of lots of things that can go wrong. Insurance systems jack up premiums  
 in minority neighbourhoods, breaking anti-discrimination laws. And machine  
 learning systems inhale existing social prejudices along with their training data;  
 as machine-translation systems read gigabytes of online text, they become much

5In 2003 I was an elected member of our university’s governing body, and we were targeted

by animal rights activists after the university proposed a new building for animals to be used  
 in medical research. Some colleagues had activists turning up at their homes to shout at them,  
 and a couple of activists were later convicted of terrorism offences after a similar campaign at  
 Oxford. Just about anyone can suddenly become a target.

6Open-source software originally developed by Yahoo to store data at petabyte scale on

clusters of servers and access it using NoSQL.

|  |  |  |
| --- | --- | --- |
|  |  |  |

better at translation but they also become racist, sexist and homophobic (we’ll  
 discuss this in more detail in section 25.3. Another problem is that if a neu-  
 ral network is trained on personal data, then it will often be able to identify  
 those persons if it comes across them again – so you can’t just train it and then  
 release it in the hope that its knowledge is somehow anonymous, as we might  
 hope for averages derived from large aggregates of data. Again, you just don’t  
 understand what the ML system is doing, so any claim you make to anonymity  
 should be treated with scepticism. And it’s not enough to say ‘We don’t sell  
 your data, we just target ads’: if you let the Iranian secret police target ads at  
 gay people who speak Farsi, they can simply pop up ads offering free pizza.

As the Information Commissioner’s Office doesn’t appear to have the capa-

bility or motivation to police anonymity services and applications, the industry  
 self-regulates; in effect, ﬁrms mark their own homework. This means adverse  
 selection, as the least conscientious provider will promise the most functional-  
 ity. As I already noted, there are many ﬁrms selling ﬁne-grained location data,  
 social data and the like who claim it’s anonymous even when it clearly isn’t.  
 Even where organisations are well-meaning, it’s rare for them to really under-  
 stand the issues until they hit trouble, and on more than one occasion we’ve had  
 providers approaching us for advice after they’d bitten off more than they could  
 chew. The data users often don’t want to talk to real experts once they hit a  
 problem as they realise that the more they know, the more expensive things will  
 be to ﬁx. As for beeﬁng up the regulator, the more a government did that, the  
 less competitive its information industries would become. One of the reasons  
 anonymisation is such a wicked problem is that its security economics are truly  
 dreadful.

**11.4.2** **Incentives**

Even imperfect de-identiﬁcation may protect data against casual browsing and  
 against some uses that are unsafe or even predatory. However, it may make  
 rascals feel empowered to do rascally things (especially since UKAN). So in  
 statistical security, the question of whether one should let the best be the enemy  
 of the good can require a ﬁner judgment call than elsewhere. As I discussed in  
 the chapter on economics, the most common cause of security failure in large  
 systems with many stakeholders is when the incentives are wrong – when Alice  
 guards a system and Bob pays the cost of failure. So what are the incentives  
 here?

The overall picture is not good. For example, medical privacy is conditioned

by how people pay for healthcare. If you see a psychoanalyst privately and pay  
 cash, then the incentives are aligned; the analyst will lock up your notes. But  
 in the US, healthcare is generally paid for by your employer; and in Britain, the  
 government pays for most of it. In both cases, attempts to centralise control  
 for management purposes have driven conﬂict with doctors and patients. While  
 such conﬂicts can be masked for a while by claims about anonymity, it is unlikely  
 that they can be resolved by any feasible privacy technology. Once people accept  
 this, a more realistic political conversation can begin.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**11.4.3** **Alternatives**

One approach is to combine weak anonymity with access control, whether re-  
 quiring the researcher to visit a secure site (as in New Zealand, and also for  
 research on tax data in the UK) or requiring licensing incorporating a non-  
 disclosure agreement plus access and use controls that forbid any attempt at  
 identifying subjects (as in Germany). This can be robust provided it is done:

1. competently, with decent security engineering;

2. honestly, without false claims that the data are no longer personal; and

3. within the law, which in the EU will involve giving data subjects a right

to opt out that is respected.

In medicine, the gold standard is doing research with explicit patient con-

sent. This not only allows full access to data, but provides motivated subjects  
 and much higher-quality clinical information than can be harvested simply as a  
 byproduct of normal clinical activities. For example, a network of researchers  
 into ALS (the motor-neurone disease from which Cambridge astronomer Stephen  
 Hawking suffered) shares fully-identiﬁable information between doctors and  
 other researchers in over a dozen countries with the full consent of the pa-  
 tients and their families. This network allows data sharing between Germany,  
 with very strong privacy laws, and Japan, with almost none; and data continued  
 to be shared between researchers in the USA and Serbia even when the USAF  
 was bombing Serbia. The consent model is spreading. A second example is  
 Biobank, a UK research project in which several hundred thousand volunteers  
 gave researchers not just full access to their records for the rest of their lives, but  
 answered an extensive questionnaire and gave blood samples so that those who  
 develop interesting diseases in later life can have their genetic and proteomic  
 makeup analysed. Needless to say, access with full consent also requires robust  
 security engineering as consent will be contingent on access being restricted to  
 researchers.

Whether you go the trusted-researcher route or the full-consent route, access

for research will also depend on ethical approval. In section 10.4.5.1 we discussed  
 the origins of medical ethics, in the Tuskegee experiments in the US and the  
 experiments performed by Nazi doctors in Germany, and the safeguards that  
 have now arisen: Institutional Review Boards (IRBs) in America and ethics  
 committees in Europe. If you’re a medical researcher with no realistic alternative  
 to using records collected from medical practice on a shaky legal basis and  
 protected using leaky de-identiﬁcation mechanisms, then you have no real choice  
 but to rely on your IRB or ethics committee. Although the exact processes differ  
 between (and within) institutions the key principle is that such research has to  
 be approved by someone independent of the researcher – typically one or more  
 anonymous colleagues, who assess both the aims of the investigation and the  
 proposed methods. There are, however, some serious moral hazards.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**11.4.4** **The dark side**

Ethics review processes provide researchers with a liability shield at two levels.  
 First, if something goes wrong and the researcher is sued for negligence, this  
 is assessed using ‘the standards of the industry’ as a yardstick. If you follow  
 the same processes as everybody else, and have each project approved by an  
 ethics committee that contains ‘independent’ members (which in practice means  
 professors from other universities, rather than representatives of the real data  
 subjects) then you can make a strong case that you followed those standards.  
 Second, if the worst happens and you face the possibility of criminal prosecution,  
 in common-law countries that involves a dual test: of ‘mens rea’ or wrongful  
 intent, as well as ‘actus reus’ or a prohibited act. Ethical approval processes are  
 designed to provide evidence that there was no mens rea. If you did what you  
 said you’d do, and for reasons that independent people approved, how can that  
 be wrongful intent? In short, ethics review processes are optimised to protect  
 the researcher and the institution, not the data subject.

This has not escaped the attention of Big Data. In section 11.2.5 I men-

tioned Google DeepMind’s ethics board and its failure to prevent the scandal;  
 Google managed to escape censure from the Information Commissioner (unlike  
 the hospital that handed over all its medical records). Unsurprisingly, ethics  
 boards are proliferating, especially as ﬁrms start throwing artiﬁcial intelligence  
 and machine learning techniques at large data warehouses with little clear idea  
 of what the outcome might be. AI ethics is a hot topic in academia and a

rapidly-growing source of jobs. The cynical operator will go through the mo-  
 tions of complying with some of the UKAN recommendations and then hire  
 some unemployed philosophers to talk about moral philosophy and the nature  
 of intelligence, while getting on with the business of selling your most intimate  
 personal information to the spammers. Ethics washing and data abuse now go  
 hand in hand.

What’s more, the existence of publicly-advertised privacy mechanisms may

deﬂect attention from abuse of the underlying personal data. In March 2007,  
 historians Margo Anderson and William Seltzer found that census conﬁden-  
 tiality was suspended in 1942, and microdata on Japanese Americans living in  
 Washington DC was given to the Secret Service in 1943 [1699]. Block-level data  
 were given to officials in California, where they rounded up Japanese-Americans  
 for internment. The single point of failure there appears to have been Census  
 Bureau director JC Capt, who released the data to the Secret Service following  
 a request from Treasury Secretary Henry Morgenthau. The Bureau has since  
 publicly apologised [1319]. But this was nothing new. The British government  
 used the 1911 census to target aliens for expulsion when WWI broke out in 1914;  
 the 1941 census was brought forward to 1939 to serve as a basis for conscription,  
 rationing and internment; and the security services continued to have a back  
 door into the census until the 1980s. Elsewhere, the Germans used census data  
 to round up Jews not only in Germany but in the Netherlands and other oc-  
 cupied territories. More recently, Cambridge Analytica and its parent company  
 SCL were granted covert access to full national census data by a number of  
 countries where they helped the incumbent government win re-election [2052].

There are many examples of publicly-advertised privacy mechanisms that are

|  |  |  |
| --- | --- | --- |
|  |  |  |

less effective than they seem. The UK is building a system of ‘smart meters’ that  
 report everyone’s gas and electricity consumption via a central clearinghouse,  
 from which it gets sent to your utility so they can bill you; other ﬁrms need an  
 approved privacy plan to get access to the data. However, when we look at a  
 typical privacy plan, we see a distribution network operator getting access to  
 half-hourly meter data for its distribution area, the Midlands, the South West  
 and Wales [2011]. The purpose is to predict when substation transformers will  
 have to be replaced. The distributor promises to aggregate this feed into half-  
 hourly totals for each feeder – these are the cables that leave the transformers  
 and supply a number of houses. But looking at the data, we see that 0.96% of  
 feeders serve only one house and 2.67% serve 3 or fewer. A more robust privacy  
 regulator would have told them to just install their own meters at their own  
 transformers. In fact, more sensible public policy would have been to not do  
 the smart meter project at all; I discuss this in Chapter 14.

As for medicine, the U.S. HIPAA system empowers the DHHS to regulate

health plans, healthcare clearinghouses, and healthcare providers, but leaves  
 many other organisations that process medical data such as lawyers, employers  
 and universities, outside its scope. Big tech companies may escape the regula-  
 tions depending on who they say they’re processing data for. In the UK, as we  
 already noted, neither the patient opt-outs nor the advertised de-identiﬁcation  
 mechanisms are effective. In many countries, more organisations than you might  
 think have access to fully-identiﬁable data.

**11.5** **Summary**

Lots of people want to believe that you can turn sensitive personal data into  
 an industrial raw material by stripping off overt identiﬁers such as names. This  
 only works in some well-deﬁned special cases, such as a national census – where  
 we have a solid theory in the form of differential privacy. In most cases, the  
 data are just too rich and re-identiﬁcation of data subjects is easy.

However policymakers, marketers, medical researchers and others want so

hard to believe that anonymity provides a magic solution to using personal data  
 that it’s difficult to disabuse them. The constant hype around big data and  
 machine learning makes the education task harder, just as these technologies  
 are making anonymity much harder still. We may expect serious trouble as  
 the scale and the scope of the privacy lawbreaking become ever more clear to  
 the public. It will probably take a scandal to bring real change, and when this  
 eventually happens, the disruption is likely to be non-trivial.

**Research problems**

At present there are several lively threads of research around anonymity and  
 privacy. First, there are practical researchers who look for new ways of de-

riving sensitive data from existing public data, or try to understand exploits  
 being carried out by marketers and cyber-criminals. Second, there are math-  
 ematicians looking at ways of doing differentially-private machine learning in

|  |  |  |
| --- | --- | --- |
|  |  |  |

various contexts, such as learning from data held by mutually mistrustful ﬁrms.  
 Third, there are privacy law scholars trying to work out how the gap between  
 law and practice could be closed. Fourth, there are practical campaigners (such  
 as EPIC, Privacy International and Max Schrems) who bring lawsuits to try to  
 stop practices that are becoming common yet which appear to violate the laws  
 we already have. This ecosystem of theory, practice, scholarship and campaign-  
 ing will no doubt continue to evolve as yet more of the stuff around us becomes  
 ‘smart’. Will ‘smart cities’ simply mean even more pervasive surveillance? In  
 the limit, will there be so much contextual information available that nothing  
 short of differential privacy will do? Or will society eventually say that enough  
 is enough, and impose radical limits on the collection, analysis and use of data  
 – and what limits might have some chance of working? Finally, the latest magic  
 potion is privacy-preserving federated machine learning. I’ve no doubt one can  
 ﬁnd edge cases in which something like that can be made to work, as with dif-  
 ferential privacy. But I suspect it will turn out to be just a variant of the snake  
 oil we’ve been fed about anonymisation over the past forty years. (Hey, if you  
 boil snake oil with sodium hydroxide, you should get snake soap.) What’s the  
 best way to debunk that?

**Further reading**

If you want to dive into the details of differential privacy, a good starting point  
 might be a long survey paper by Cynthia Dwork and Aaron Roth [594]. The  
 classic reference on inference control is Dorothy Denning’s 1982 book [538]; the  
 1989 survey paper by Adam and Wortman is a good summary of the state  
 of the art then [17]. An important reference for statisticians involved in U.S.  
 government work is the Federal Committee on Statistical Methodology’s *‘Report*  
 *on Statistical Disclosure Limitation Methodology’* which introduces the tools and  
 methods used in various US departments and agencies [667]; this dates back to  
 2005, so it’s somewhat out of date and is currently being rewritten. The UKAN  
 book is a must-read if you’re doing anonymisation for a client operating within  
 the UK’s jurisdiction [626]. As an example of a quite different application, Mark  
 Allman and Vern Paxson discuss the problems of anonymizing IP packet traces  
 for network systems research in [42]. Finally, Margo Anderson and William

Seltzer’s papers on the abuses of census data in the US, particularly during  
 World War 2, can be found at [52].

|  |  |  |
| --- | --- | --- |
|  |  |  |