# Week 6: Data Ano

Charlotte Hadley

# Topics for today

1. Why is data anonymisation important?
2. What are specific risks of deanonymisation of heal
3. Anonymity measures: k-anonymity and l-dive
4. Case studies of deanonymisation

> … and why anonymity measures are often n

5. R packages for working with anonymouse da

# Why is data anon
# important?

# Data isn't always ca

Mostly during this course we've been talkin
explicitly being collected - and participan

But that's often not the case.

> Data is collected continuously about indi
> often without implicit consent.

# Data tracking across

Cookies [and similar t

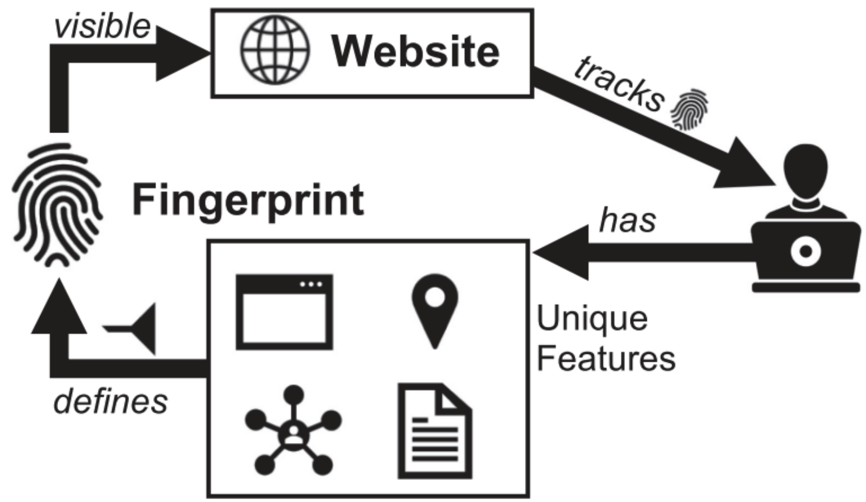
ever present in the mo

They allow websites to

websites

with websites.

> … but what about all those cookie
> popups?!




(b) A representation of how device fingerprinting enables the tracking of users on the web. The specific fingerprint may vary across different websites, since it can include different features unique to a particular device, which are requested by the website.

Source: Kretschmer[1] et

# GDPR and Cookies

The GDPR[2] Policy included
"cookie" ferris, which u
the cookie consent
everywhere.

For a common sense
the policy group / code[3]

The policy became in
the cookie avalanc

The policy was well m
necessary.

In 2016 it was demons
the top million websi
of tracking[4]

However, there's clear eviden
policy hasn't materia
privacy.[1]

# Cookies: Fingerprinting, ... patterns

These are the primary ways websites circumv...

- The biggest issue with ... m... tools like fingerprinting are harder to track.
- Users are nudged to accept cookies by auto select...
- Dark patterns employed to prevent users[5] from...

Overall, the policy has probably made thing...

> GDPR impacted smaller advertisement compa... brands, such as Google and Facebook, lead... these companies, which, in turn, may incr... decrease it Source:[1] Kretschmer et al 2021

# Why do websites want to t

There are simple un
this:

- Selling tracking d
  networks
- Using tracking dat
  users

But there are inter

- Anticipatory shipp
  purchases
  - Amazon first patented
    in 2013[6], thoroughly
    Eva-Maria Nyckel[7]
  - Used in the agro[8]

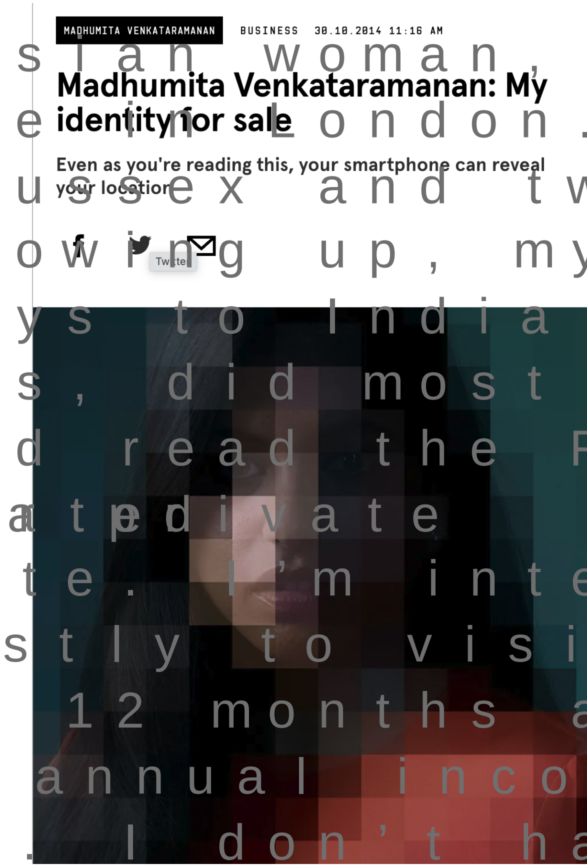
Can we get some exam
they track?

## Wired: Madhumita Venkataramanan: My iden

I strongly recommend reading wired.co.uk/article/

- The article is from 2014, but as we've dis made this situation worse.
- The article also provides a wealth of othe These would be good examples for your ass
- The quote below is from the article and fsu tracked without consent.

I'm a 26-year-old British Asian woman, wo
and living in an SW postcode in London. I
lived at two addresses in Sussex and two
east London. While I was growing up, my f
detached house, took holidays to India ev
donated to medical charities, did most of
shopping online at Ocado and read the Fin
Now, I rent a flat owned by a private
landlord and have a housemate. I'm intere
and startups. I have taken five holidays (mostly to visit
friends abroad) in the last 12 months and
bought flights within 14 days. My annual income
between £30,000 and £39,999. I don't have
watching scheduled television…



# Data tracking during

There's an A&E waiting s
sent to folks that

This survey is sent
into it.

> The Section 251 of
> 2006[10] provides for t
> confidential patient
> consent for purposes
> HRA or the Secreta
> Health and Social

That's an exception to
Act[11]

This means that the
at this data and de

This exception was
COVID-19 pandemic, u
surveying hospital

Note that the NHS m
opt out [once you'v

> However, as has alway
> case, patients/ser
> given the opportun

# Data anonymisation i
# because data is col
# all of the time

... how does that match up with GDPR and the

# Individual rights fr

The AD[11] P provides 8 rights for individuals:

- The right to be informed
- The right of access
- The right to rectification
- The right to erasure
- The right to restrict processing
- The right to data portability
- The right to object
- Rights in relation to automated decision making and profiling

If we don't know that … collected - or by whi… individual rights are

This opens up lots of … We'll discuss these i… about data ethics.

In this lecture we're … specific risks to deanony…

# What are the ri
# anonymisation

# Who can be at risk of anonymisation?

| Individuals | Organisations |
|---|---|
| It's the dangers to individuals which should primarily be considered. | However, organisations are also at risk if their anonymisation process is ... |
| There are significant risks to individuals' liberty, livelihood and life from deanonymisation. | • Organisations face reputational damage<br>• Organisations face legal difficulties, including fines |

> Let's focus on the individual. The ICO provides a useful data protection risk ... designed for organis...

# Specific risks of de-anonymi

First and foremost, th There is a risk of itself
but more specifically

- Information about someone's ending up in the publ
- Individual suffering distress, embarrassment, or anx sensitive information domain

There is a risk of sensitive information domain that:

- Individual suffering harassment attack and/or injury
- Individual suffering persecu

# Specific risks of deanonymis

Sensitive information third-party organisations change in service options loss.

- Private healthcare da might be sold to insurers resulting in increase terminate existing p costs
- Employers could poten information in employ

> Remember that (the GDPR and UK GDPR) protects (principle...) individuals' data'. If data is sold without knowl

# Specific risks of deanonymis

These are the “protected characteristics” defined in the Equality Act 2010[13]

- Age
- Disability
- Gender reassignment
- Marriage and Civil Partnership
- Pregnancy and Maternity
- Race
- Religion
- Sex
- Sexual Orientation.

Further, [illegible] prejudice against ind
groups.

This prejudice can be
actions, from hate gr
institutional policie

De-anonymisation of h
realistically [and of
individuals’ protected chara

All individual data s
private, but there is [illegible] risk
the de-anonymisation
healthcare data.

# What is health data

Recall how the Data Protection Act[11] of healt

- “biometric data” means personal data resulting from specific technical pro relating to the physical, physiological or which allows or confirms the unique identification of that individu or dactyloscopic data;
- “data concerning health” means personal data relating of an individual, including the provision about his or her health status;
- “genetic data” means personal data relating to th characteristics of an individual which giv the health of that individual and which result from an analysis of a biological sample from the individual in q

# A potted histo
# anonymisation

# Early evidence for d

In the late 90s there was a rapid conceptua
large, public datasets.

Latanya Sweeney[14] showed
using public data and
voting list (n=54,805)
easy to uniquely identify indi
only 2 pieces of infor

| | |
|---|---|
| birth date alone | 12% |
| birth date and gender | 29% |
| birth date and 5-digit ZIP code | 69% |
| birth date and full postal code | 97% |

**Table 3. Uniqueness of Demographic Fields in Cambridge, Massachusetts, Voter List.**

Source: Latanya[14] Sween

# Early evidence for d

In the late 90s there was a rapid conceptua
large, public datasets.

Three years later in 2
Sweeney[15] demonstrated

> 87% of the US population can be uniquely identified only ZIP, gender and date of birth.

form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

Source: Latanya[15] Sween

# k-anonymity: A Model for Privacy

# k-anonymity: Definition (I)

k-anonymity is a property of data that has been subject to anonymisation.

Unfortunately, a lot of data ... k-anonymity is ... people don't declare calculating k values.

k-anonymity integer ... guarantees uniqueness of individuals amongst k-1 individuals.

# k-anonymity: (II)

Lest 'consider a simple pre

| name | regi | age_r | disea |
|---|---|---|---|
| Saind | Engl | 20-30 | Heart |
| Enio | Engl | 20-30 | Heart |
| Daury | Engl | 20-30 | Heart |
| Alphu | Engl | 20-30 | Pancr |
| Balia | Engl | 20-30 | Pancr |
| Kenye | Engl | 20-30 | Pancr |
| Graci | Wale | 40-50 | Liver |
| Aliye | Wale | 40-50 | Liver |
| Kadin | Wale | 40-50 | Liver |
| Asaph | Wale | 40-50 | Liver |

We consider each column i... be an attribute.

These attributes can be c...nto two types of identifer:

- Unique identifiers
  - These are attributes t... identify individuals. These ... removed for anonymisation.
- Quasi-identifiers
  - These attributes could be used used to identify individuals, even after anonymisation.

# k-anonymity: Identification (III

Now we've thrown away the identifiers we need to decide wh attributes are sensitive

Sensitive attributes are m lthcare data that we need to prote the anonymisation process.

| Non-sens | Sensi |
|---|---|
| regi age_r | disea |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |

# k-anonymifitnyi :t iAonde(IV)

There arefenroewn t3 cdhiofices abo how we calculate k-anonymi data.

- Combining together all no attributes ecaocmhspeanrseidt itvoe attributes.
- Combining together all at
- For each individual attri

Lest'go through each of thes .

| Non-sens | Sensi |
|---|---|
| regi age_r | disea |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |

# k-anonymity: Definition (V)

> Combining together all non-sensitive attributes each compared to sensitive attributes.

In toy example like this … rough manually any count how ma… ls belong to each group.

Using this measure the da… anonymous as all individu… guaranteed anonymity amongst 6 others (ike-1).

| disea | regi | age_r | n_in_g |
|---|---|---|---|
| Heart | Engl | 20-30 | 6 |
| Heart | Engl | 20-30 | 6 |
| Heart | Engl | 20-30 | 6 |
| Pancr | Engl | 20-30 | 6 |
| Pancr | Engl | 20-30 | 6 |
| Pancr | Engl | 20-30 | 6 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |

# k-anonymity (VI)

> Combining together all attributes

When we measure all attributes the k-anonymity of the da... reduced.

Using this metric, the da... anonymity.

| disea | regi | age_r | n_in_g |
|---|---|---|---|
| Heart | Engl | 20-30 | 3 |
| Heart | Engl | 20-30 | 3 |
| Heart | Engl | 20-30 | 3 |
| Pancr | Engl | 20-30 | 3 |
| Pancr | Engl | 20-30 | 3 |
| Pancr | Engl | 20-30 | 3 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |
| Liver | Wale | 40-50 | 4 |

# k-anonymity: A Definition (VIII)

> For each individual attribute

When we measure the anonym... h individual variable the da... 2-anonymity

We always select the $k$ for our specific measure.

| disea | regi | age_r |
|---|---|---|
| Heart | Engl | 20-30 |
| Heart | Engl | 20-30 |
| Heart | Engl | 20-30 |
| Pancr | Engl | 20-30 |
| Pancr | Engl | 20-30 |
| Pancr | Engl | 20-30 |
| Liver | Wale | 40-50 |
| Liver | Wale | 40-50 |
| Liver | Wale | 40-50 |
| Liver | Wale | 40-50 |

# k-anonymity: Identifiability (VII

As we've seen, each of the gives a different measure of the of the data.

1. Combining together all no attributes each separately, non-sensitive attributes.
2. Combining together all at
3. For each individual attri

| Non-sens | Sensi |
|---|---|
| regi age_r | disea |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Heart |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Engl 20-30 | Pancr |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |
| Wale 40-50 | Liver |

Frustratingly it is quite rare for authors to explicitly state which combination of attributes they use.

The methods are listed roughly in terms of the frequency that I've seen them in the literature.

# k-anonymity: Definition (IX)

The definition I've given is not ... and precise.

But be aware that you'll see[16] more technical definitions that use set theory notation.

Ignore It: a dataset is having "k-anonymity 1( ... any row in the dataset there are at least 9 other rows identical to it.

**Example 3. Table adhering to *k*-anonymity**

Figure 2 provides an example of a table T that adheres to *k*-anonymity. The quasi-identifier for the table is $QI_T = \{Race, Birth, Gender, ZIP\}$ and $k=2$. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T. That is, for each sequence of values in $T[QI_T]$ there are at least 2 occurrences of those values in $T[QI_T]$. In particular, $t1[QI_T] = t2[QI_T]$, $t3[QI_T] = t4[QI_T]$, $t5[QI_T] = t6[QI_T]$, $t7[QI_T] = t8[QI_T] = t9[QI_T]$, and $t10[QI_T] = t11[QI_T]$.

**Lemma.**

Let $RT(A_1,...,A_n)$ be a table, $QI_{RT} = (A_i,\ldots,A_j)$ be the quasi-identifier associated with RT, $A_i,\ldots,A_j \subseteq A_1,\ldots,A_n$, and RT satisfy *k*-anonymity. Then, each sequence of values in $RT[A_x]$ appears with at least *k* occurrences in $RT[QI_{RT}]$ for $x=i,\ldots,j$.

**Example 4. *k* occurrences of each value under *k*-anonymity**

Table T in Figure 2 adheres to *k*-anonymity, where $QI_T = \{Race, Birth, Gender, ZIP\}$ and $k=2$. Therefore, each value that appears in a value associated with an attribute of QI in T appears at least *k* times. $|T[Race = \text{"black"}]| = 6$. $|T[Race = \text{"white"}]| = 5$. $|T[Birth = \text{"1964"}]| = 5$. $|T[Birth = \text{"1965"}]| = 4$. $|T[Birth = \text{"1967"}]| = 2$. $|T[Gender = \text{"m"}]| = 6$. $|T[Gender = \text{"f"}]| = 5$. $|T[ZIP = \text{"0213*"}]| = 9$. And, $|T[ZIP = \text{"0214*"}]| = 2$.

Source: Formal definition for k-anonymity from Sweeney 2002[16]

# k-anonymity: How is

We are responsible for manipulating our dat

> Even though a minimum k value of 3 is oft
> in practice is to en fivuer es itmhial ta rt hoebrsee[19] ravraet iao

We have two tools available to us:

# Task: Setup our pro

SLIDE 1 OF 1

1. Create a new project for week-6
2. Create a new RMarkdown document called data-anonymisation.Rmd
3. Install the {wakefield} and {aux} package

# {wakefield}

The {wakefield} package is very useful for creati[ng]
variables.

The package features 49 built-in variables with

| | | | | |
|---|---|---|---|---|
| age | dice | hair | mili | sex_in |
| animal | dna | height | mont | smokes |
| answer | dob | income | name | speed |
| area | dummy | internet | norm | state |
| car | educat | iq | poli | string |
| childr | employ | language | race | upper |
| coin | eye | level | relig | valid |
| color | grade | likert | sat | year |
| date_s | grade_ | lorem_ip | sent | zip_co |
| death | group | marital | sex | |

# wakefield::r_data_frame

We generate a data set with the r_data_frame() function:

```r
r_data_frame(10,
             id,
             name,
             dob,
             income,
             smokes,
             death)
```

```
# A tibble: 10 × 6
   ID    Name      DOB        Income Smokes Death
   <chr> <chr>     <date>      <dbl> <lgl>  <lgl>
 1 01    Dreniyah  2008-05-24 42462. TRUE   FALSE
 2 02    Alexiana  2008-11-06 60950. FALSE  TRUE
 3 03    Ashauria  2008-10-16 17069. FALSE  TRUE
 4 04    Azelea    2007-12-11 19028. FALSE  FALSE
 5 05    Krystof   2009-09-11 49067. TRUE   FALSE
 6 06    Loudes    2008-10-24 65993. FALSE  FALSE
 7 07    Rebeckah  2008-07-21 50860. FALSE  FALSE
 8 08    Jarek     2009-01-26 21266. FALSE  FALSE
 9 09    Gavrielle 2008-08-03 50405. FALSE  TRUE
10 10    Tarajah   2008-05-07 28798. FALSE  FALSE
```

> Can you explain why the data are not identical on your machine

# Pseudorandomness

When programming we use pseudorandom numbe[r…] numbers.

These are deterministic algorithms that give really random numbers wh[…]
We can therefore always get the same numbers by setting the seed of the algorithm.

```r
set.seed(1)
r_data_frame(10,
             id,
             name,
             income,
             dna,
             smokes,
             death)
```

```
# A tibble: 10 × 6
   ID    Name      Income DNA      Smokes Death
   <chr> <chr>      <dbl> <fct>    <lgl>  <lgl>
 1 01    Donaldeen 48108. Cytosine FALSE  FALSE
 2 02    Martiqua  36496. Guanine  FALSE  FALSE
 3 03    Juliaann  12130. Thymine  FALSE  TRUE
 4 04    Poyraz    53488. Cytosine FALSE  TRUE
 5 05    Boleslaus 46311. Adenine  FALSE  FALSE
 6 06    Duc       56733. Adenine  FALSE  FALSE
 7 07    Hadeer    52217. Thymine  TRUE   TRUE
 8 08    Camilya   31854. Thymine  FALSE  TRUE
 9 09    Ashlay    77153. Adenine  TRUE   TRUE
10 10    Dutch     62490. Adenine  FALSE  FALSE
```

# k-anonymity for our

Let's pretend our dataset reflects information about youth.[1]

When thinking about making anonymous...

- Are there any variables to suppress?
- How could we generalise variables?

| ID | Name | DOB | Inco | Smok | Dea |
|---|---|---|---|---|---|
| 01 | Donal | 2009-0 | 3185 | FALS | FALS |
| 02 | Marti | 2009-0 | 7715 | FALS | FALS |
| 03 | Julia | 2008-1 | 6249 | TRUE | TRU |
| 04 | Poyra | 2009-0 | 1546 | TRUE | FALS |
| 05 | Boles | 2008-0 | 1943 | FALS | TRU |
| 06 | Duc | 2008-0 | 2450 | FALS | FALS |
| 07 | Hadee | 2009-0 | 2753 | TRUE | TRU |
| 08 | Camil | 2008-1 | 1620 | FALS | FALS |
| 09 | Ashla | 2008-0 | 5750 | FALS | FALS |
| 10 | Dutch | 2008-0 | 5922 | FALS | TRU |

[1] I'm limited by the options given by {wakefield}. There is NO correlation implied here. You can tell because I don't have a citation.

# Task: k-anonymity c

SLIDE 1 OF 2

1. Use this code to create a dataset:

```
library(wakefield)
library(tidyverse)
library(lubridate)

set.seed(1)
smoke_data <- r_data_frame(
  50000,
  id,
  name,
  dob(start = ymd("1950-01-01"),
      k = abs(as.integer(days(ymd("1950-01-01") - Sys.Date()) - 365*18))),
  income,
  smokes,
  death)
```

2. Suppress inappropriate columns from the

# Task: k-anonymity c

SLIDE 1 OF 2

1. Generalise the remaining variables as f

- Extract year of birth
- Split income into 4 categories
  - "< £30,000"
  - "£30,000 - £70,000"
  - "£70,000 - £100,000"
  - "£100,000+"

2. Calculate the k-anonymity of the datase

# How to attack k-anon

There are 3 known attacks for attempt to de

- Unsorted matching attack
  - If an anonymised dataset is ordered in t recorded this is a. pDoetpeenntdieanlt aotnt atchke vaetctta either an addifieiro noarl (qwuoarnstt- icdaesnet)i a uniqu
  - Ist'a good practice to randomise the orde releases.

# How to attack k-anon
# (II)

There are 3 known attacks for attempt to de

- Unsorted ma

- Subsequent release attacks
  - Large healthcare datasets might be used multiple k-anonymised releases.
  - Temporal attacks are possible by [illegible] rows (r intervention) or the removal of data (eg
  - Protecting against these attacks requir consider the k-anonymity of combined rel

# How to attack k-anon (III)

There are 3 known attacks for attempt to de

- Unsorted mat
- Subsequent re

# l-diversity

# l-diversity (I)

l-diversity is a more sophisticated measure of the anonymi
variables in an anonym
introduced by Machanav
2006[21]

This method depends or
model of background kr

We are going to cover E
statistics. Read stats.github.io/book/ if
you're interested.

assignments $\psi$ compatible with the background knowledge such that $\psi(X) = s$ can be calculated as follows. $X$ is assigned the sensitive value $s$. Since $X[Q] = q$, out of the remaining $N_q - 1$ individuals having the nonsensitive value $q$, $N_{(q,s)} - 1$ of them are assigned $s$. For every other sensitive value $s'$, $N_{(q,s')}$ out of the $N_q - 1$ individuals are assigned $s'$. For every $q' \neq q$ and every $s'$, some $N_{(q',s')}$ out of the $N'_q$ individuals having the nonsensitive value $q'$ are assigned $s'$. The number of these assignments is

$$\frac{(N_q-1)!}{(N_{(q,s)}-1)!\prod_{s'\neq s} N_{(q,s')}!}\prod_{q'\neq q}\frac{N_{q'}!}{\prod_{s'\in S} N_{(q',s')}!} = \frac{N_{(q,s)}}{N_q}\prod_{q'\in Q}\frac{N_{q'}!}{\prod_{s'\in S} N_{(q',s')}!} \quad (2)$$

For each mapping $\psi$ such that $\psi(X) = s$, we count the number of $Z_n$'s such that $(\psi, Z_n) \vdash (T^\star, X)$ as follows. Let $q^\star$ be the generalized value of $q = X[Q]$. $X$'s record will appear as $t^\star_X = (q^\star, s)$ in the table $T^\star$. Apart from $t^\star_X$, $T^\star$ contains $n_{(q^\star,s)} - 1$ other tuples of the form $(q^\star, s)$. Hence, apart from $X$, $Z_n$ should contain $n_{(q^\star,s)} - 1$ other individuals $\omega$ with $\psi(\omega) = s$ and $\omega[Q] = q'$ where $q'$ generalizes to $q^\star$. For all other $(q^{\star\prime}, s')$ such that $q^{\star\prime} \neq q^\star$ or $s' \neq s$, $Z_n$ should contain $n_{(q^{\star\prime},s')}$ individuals $\omega'$ where $\psi(\omega') = s'$ and $q^{\star\prime}$ is the generalized value of $\omega[Q]$. The number of $Z_n$'s is given by

$$\binom{N_{(q^\star,s)}-1}{n_{(q^\star,s)}-1}\prod_{(q^{\star\prime},s')\in(Q^\star\times S)\setminus\{(q^\star,s)\}}\binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} = \frac{n_{q^\star,s}}{N_{(q^\star,s)}}\prod_{(q^{\star\prime},s')\in Q^\star\times S}\binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} \quad (3)$$

The cardinality of $\mathcal{T}^\star_{(X,s)}$ is therefore the product of Equations 2 and 3 and can be expressed as

$$\begin{aligned}|\mathcal{T}^\star_{(X,s)}| &= \frac{N_{(q,s)}}{N_q}\prod_{q'\in Q}\frac{N_{q'}!}{\prod_{s'\in S}N_{(q',s')}!}\times\frac{n_{q^\star,s}}{N_{(q^\star,s)}}\prod_{(q^{\star\prime},s')\in Q^\star\times S}\binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}}\\ &= n_{(q^\star,s)}\frac{N_{(q,s)}}{N_{(q^\star,s)}}\times\frac{1}{N_q}\prod_{q'\in Q}\frac{N_{q'}!}{\prod_{s'\in S}N_{(q',s')}!}\times\prod_{(q^{\star\prime},s')\in Q^\star\times S}\binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}}\\ &= n_{(q^\star,s)}\frac{N_{(q,s)}}{N_{(q^\star,s)}}\times\mathcal{E}\end{aligned}$$

The expression $\mathcal{E}$ is the same for all $s' \in S$. Hence, the expression for the observed belief is

$$\beta_{(q,s,T^\star)} = \frac{|\mathcal{T}^\star_{(X,s)}|}{\sum_{s'\in S}|\mathcal{T}^\star_{(X,s')}|} = \frac{n_{(q^\star,s)}\frac{N_{(q,s)}}{N_{(q^\star,s)}}}{\sum_{s'\in S}n_{(q^\star,s')}\frac{N_{(q,s')}}{N_{(q^\star,s')}}}$$

Using the substitutions $f(q,s) = N_{(q,s)}/N$ and $f(q^\star,s) = N_{(q^\star,s)}/N$, we get the required expression.

$$\beta_{(q,s,T^\star)} = \frac{n_{(q^\star,s)}\frac{f(q,s)}{f(q^\star,s)}}{\sum_{s'\in S}n_{(q^\star,s')}\frac{f(q,s')}{f(q^\star,s')}} = \frac{n_{(q^\star,s)}\frac{f(s|q)}{f(s|q^\star)}}{\sum_{s'\in S}n_{(q^\star,s')}\frac{f(s'|q)}{f(s'|q^\star)}}$$

Note that in the special case when $S$ and $Q$ are independent, The expression for the observed belief simplifies to

# l-diversity (II)

In order to estimate l-div … need to … once again consider the st … our dataset.

| Non-sensiti Keys | Sensi |
|---|---|
| regicage_ra | disea |
| Engla 20-30 | Heart |
| Engla 20-30 | Heart |
| Engla 20-30 | Heart |
| Engla 20-30 | Pancr |
| Engla 20-30 | Pancr |
| Engla 20-30 | Pancr |
| Wales 40-50 | Liver |
| Wales 40-50 | Liver |
| Wales 40-50 | Liver |
| Wales 40-50 | Liver |

We split our data into:

- Sensitive variables - the … variables.
- Keys - the non-sensitive …

A dataset is l-diverse if … nique combination of key attribu… are at leals"well-represented" values for each sensitive variable.

# l-diversity (III)

# Task: Calculate I-d

SLIDE 1 OF 2

1. Install {lsld ctMiec} package
2. Add this dataset to your .Rmd

```r
data_diseases <- tibble(
  name = c("Saindhavi", "Enio", "Daury", "Alphus", "Balian",
           "Kenyea", "Gracielynn", "Aliye", "Kadince", "Asaph"),
  region = c(rep("England", 6), rep("Wales", 4)),
  age_range = c(rep("20-30", 6), rep("40-50", 4)),
  disease = c(rep("Heart", 3), rep("Pancreatic", 3), rep("Liver", 4))
)
data_diseases
```

```
# A tibble: 10 × 4
   name       region  age_range disease
   <chr>      <chr>   <chr>     <chr>
 1 Saindhavi  England 20-30     Heart
 2 Enio       England 20-30     Heart
 3 Daury      England 20-30     Heart
 4 Alphus     England 20-30     Pancreatic
 5 Balian     England 20-30     Pancreatic
 6 Kenyea     England 20-30     Pancreatic
 7 Gracielynn Wales   40-50     Liver
 8 Aliye      Wales   40-50     Liver
 9 Kadince    Wales   40-50     Liver
10 Asaph      Wales   40-50     Liver
```

# Task: Calculate l-d

SLIDE 2 OF 2

1. Compute the l-diversity (d) of the disease attribute

```
ld_diseases <- data_diseases %>%
  mutate(disease = as_factor(disease)) %>%
  createSdcObj(keyVars = c("region", "age_range")) %>%
  ldiversity(ldiv_index = "disease")
```

1. Extract the l-diversity value

```
ld_diseases@risk$ldiversity
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     1.0     2.0     1.6     2.0     2.0
```

# Summarising k-ano
# diversity

# Summarising k-anonym… diversity

k-anonymity has well known vulnerabilities for high-dimensional …

identifiers [20]. Furthermore, *k*-anonymization completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases.

Source: Narayanan and …

However it is still worth …ng at least 5-anonymised datasets containing sensitive attributes.

20

l-diversity is a much more sophisticated tool that provides stronger privacy protections. It is NP-hard to re-identify individuals in l-diverse datasets

Even though a minimu… is often suggested, recommendation in pr… ensure that there ar… similar observations

19

However, l-diversity does not guarantee against re-identification. The coarsening of data might also degrade the usability of released data.

# Case Study: Netflix Prize

# Netflix Prize Dataset:

In October 2006, Netflix started a competition with the intention of improving their recommendation engine.[22]

Netflix sought an algorithm/methodology that would improve their recommendation algorithm.

Let's get into this a little more.

Netflix provided over 100 million ratings (and their dates) from over 480 thousand randomly-chosen, anonymous subscribers on nearly 18 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received by Netflix during this period. The ratings are on a scale from 1 to 5 (integral) stars.

| us | mov | date_of | gra |
|---|---|---|---|
| 13 | Ali | 2006-01 | 5 |
| 13 | Ali | 2006-01 | 4 |

It withheld over 3 million most-recent ratings from those same subscribers over the same set of movies as a competition qualifying set.

# Neflti x Prize Dataset:

Neflti x had an alCgionreimt which attempted's to oa ofilmX based on's tehxei sut movie ratings.

The intentCi ionne mwaatsoc hf provide more person recommendations tha average ratXnagc rfoosrs users.

Neflti x chose to use the squared error orC (RMSE) c and all other uscempraed usesr actual rating as accuracy

The Cinema adlgorithm was 10% more accurahteothea user rating

The competition chall further improve this an addition 10%.

# Netflix Prize Dataset:

The competition wa
and it wasn't won

In fact, quite a dram
winning entries su
one another in 200

I'd recommend thrilld
article by[23] Dan Jac

However,

16 days after Arvind Narayan
Vitaly Shmatikov demo
to re-identify users.[24]

This draft paper was finally published i
2008[20] and we'll look in
identification was possibl

# Netflix Prize Dataset:

With the announcement of a 2nd competition in 2009 a class-action suit was filed.

The suit described the Netflix Prize dataset as the biggest "voluntary privacy breach to date". The Federal Trade Commission (FTC) also got involved.

Unfortunately the case was settled privately so we don't know the damages. The best I can get is this quote from a deleted[25] blogpost

"To some, renting a m
Brokeback Mountain o
Passion of the Chris
personal issue that
want published to th
Jane Doe, a lesbian,
want her sexuality n
gay and lesbian films them
broadcast to the wor
anonymity in this ac

> In the past few months, the Federal Trade Commission (FTC) asked us how a Netflix Prize sequel might affect Netflix members' privacy, and a lawsuit was filed by KamberLaw LLC pertaining to the sequel. With both the FTC and the plaintiffs' lawyers, we've had very productive discussions centered on our commitment to protecting our members' privacy. We have reached an understanding with the FTC and

> have settled the lawsuit with plaintiffs. The resolution to both matters involves certain parameters for how we use Netflix data in any future research programs.
>
> In light of all this, we have decided to not pursue the Netflix Prize sequel that we announced on August 6, 2009." - Neil Hunt, Chief Product Officer @ Netflix

# Netflix Prize De-anonymisation (I)

The exact mechanics of the algorithm behind Arvind Narayanan and Vitaly Shmatikov's[20] de-anonymisation attack are beyond the scope of this course.

We're going to walk through the mechanics of the attack.

With 8 movie ratings ... allow 2 to be comple... dates that may have ... 96% of Netflix subscribers ... records have been re... uniquely identified in the dat...

Source: arXiv[24] pre-pr...

It's important to identify this paper introduced robust statistical de-anonymisation of large sparse datasets that is not unique to the Netflix dataset.

# Neflti x Prize Data fisc aett io (II)

The de-anonymisation aItMtDaBc kp riosv ipdoewse rae dl arge by two data sources: movie ratings.

- IMDB Ratings
- Date of m

An assumed similflaxritaym IMDB ratings prficoavtidoesn attack. vector

The authors hexatvertemrewmilyty ten an use fAuQf[26]. F

The paper provides a attackonulsyin5g0 ulsMeDrBs.

Despite this, tfidhenaudt positively cross-matc between the datasets.

# Netflix Prize Dataset (II)

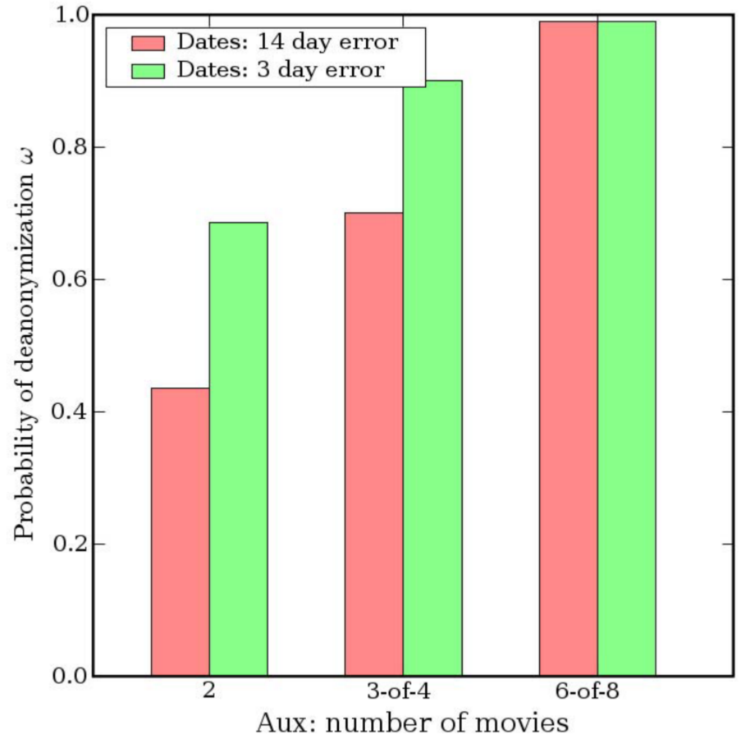
# Netflix Prize Dataset:
# (II)

The de-anonymisation aDtattaec ko fi sm opvoiwee rreadt i n g by two data sources: additional .attack vec

- IMDB R
- Date of movie review

Background knowledge account creation date vector for that user

The authors heaxvter ewmr eiltyt en an use fAuQl[26]. F



Selected other ca

# Selected other case

We're going to look at a few additional cas

Please note that for most of these examples identification of users from de-anonymised datas breach privacy -. sometimes publicly

All of these case studies can be used in th assessment. And please do share additional

Remember I mentioned https://www.wd.co.uk/articles/my-ide as being source of additional case studies.

# Case Study: Facebook

Facebook has a long a

privacy breaches and

activity

nBde aacwofnu lw ahsi sdteosriyg noefd t

qpuoesstt ipounracbhlaes es Facebo

feeds.

Facebook Beacon is one

examples, all the way

eT hoef ctohmep aonlyd eosrti g i n a l l

sbearcvki cfer owma s2 0“0o7p.t - i n ” ,

e v i d e n c e this was not

It's a rare example where Mark

Zuckerberg open talks

mistake[27]

Haobwoeuvyte ar ift t ears car i t i c i s m

provided an opt-out.

to turn the service.

> Discuss why Facebook Beacon
> breaches user privacy

A class action suit i

$9.5million.

# Case Study: Google B

Google Buzz was a v
networking tool:

- Launched: February[28]
- Discontinued: Dec[29]

It shut down explic
violations and Goog
$8.5 million[30] within one of mo
service launching.

Google automatically public
Google Profile pages usin
service.

The public pages disclo
user most frequentl
via email or chat w

This was as designe
Google designed thi

> When you first enter Google Buzz, to make the startup experience easier, we may automatically select people for you to follow based on the people you email and chat with most. Similarly, we may also suggest to others that they automatically follow you. You can review and edit the list of people you follow and block people from following you.
>
> . You may opt out of displaying the list of people following you and who you're following on your profile.

# Case Study: In-store

Do you know what these machines are for in Tesco?

# Case Study: In-store

These devices have emb... that can track custome... for the their attention.


The video feed from th... processed by software ...an estimate:

31

- Gender
- Age
- Mood

Their primary purpose makes for interesting reading.

"Quividi's software employs advanced facial detection, not facial recognition" Quividi marketing

# Anonymisation sof

# Anonymisation softwa

**Table 5.** Comparison of the off-the-shelf privacy model–based data anonymization tools in terms of available development options, anonymization functionality and risk metrics.

| Tool | Last release | Development support | | | | | Anonymiza-tion | Risk as-sessment |
|---|---|---|---|---|---|---|---|---|
| | | Open source | Public API[a] | Extensibility | Cross-plat-form | Programming language | | |
| ARX | November 2019 | ✓[b] | ✓ | ✓ | ✓ | Java | ✓ | ✓ |
| Amnesia | October 2019 | ✓ | ✓ | ✓ | ✓ | Java | ✓ | |
| μ-ANT[c] | August 2019 | ✓ | ✓ | ✓ | ✓ | Java | ✓ | |
| Anonimatron | August 2019 | ✓ | ✓ | ✓ | ✓ | Java | | |
| SECRETA[d] | June 2019 | | | | ✓ | C++ | ✓ | |
| sdcMicro | May 2019 | ✓ | ✓ | Poorly support-ed | ✓ | R | ✓ | ✓ |
| Aircloak Insights | April 2019 | | | | ✓ | Ruby | | |
| NLM[e] Scrubber | April 2019 | | | | ✓ | Perl | | |
| Anonymizer | March 2019 | ✓ | ✓ | ✓ | ✓ | Ruby | | |
| Shiny Anonymizer | February 2019 | ✓ | ✓ | ✓ | ✓ | R | ✓ | |
| μ-ARGUS | March 2018 | | | | | C++ | ✓ | ✓ |
| UTD[f] Toolbox | April 2010 | ✓ | | Poorly support-ed | ✓ | Java | ✓ | |
| OpenPseudonymiser | November 2011 | ✓ | | | ✓ | Java | | |
| TIAMAT[g] | 2009 | | | | ✓ | Java | ✓ | |
| Cornell Toolkit | 2009 | ✓ | | Poorly support-ed | ✓ | C++ | ✓ | Poorly supported |

In 2021 [32] Zuo performed a syste
de-anonymisation tools in

There we few that be ... built with R

- {sdcMicro}
  - This package contains ma
    measuring/exploring the
    via a {shiny} app.
- `{ShinyAnonymizer}
  - This package provides a
    anonymising healthcare d

# Anonymisation softwa

However. There isn't a one-size-fits-all methodology or guarantee of privacy through anonymisation.

k-anonymity and l-diversity are useful metrics and provide some assurance of privacy. But background knowledge attacks might undermine these.

We need to be vigilan

- Preparing data for re
- Designing services

# Simulating fake d

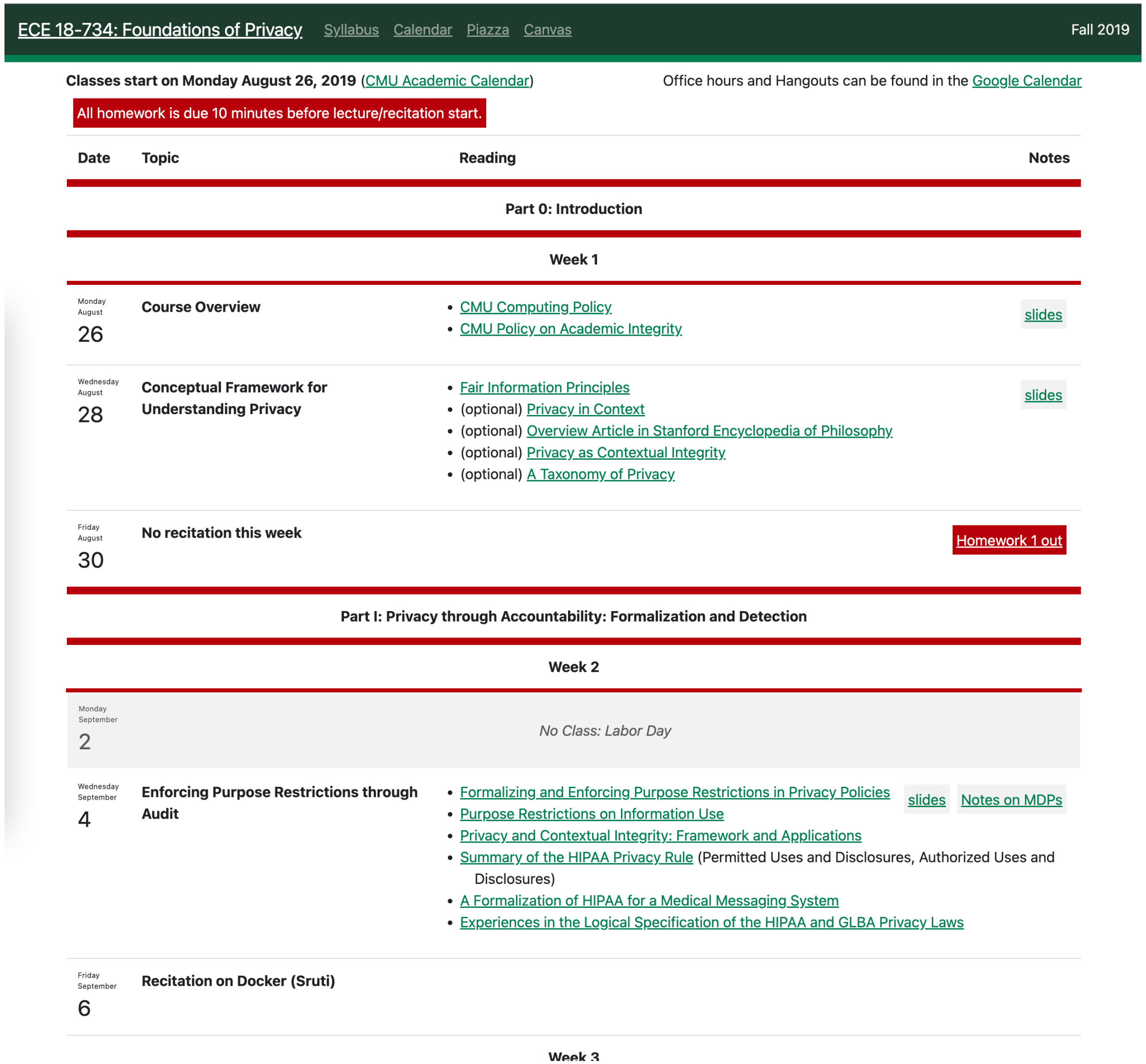
# Simulating fake data

We've used the {wakefield} package to
fake datasets.

The {faux} package is a more s
package for simulating data
Lisa DeBruine

# Useful resources

# Useful resources (I)



ECE 18-734: Foundations of Privacy | Syllabus | Calendar | Piazza | Canvas | Fall 2019

Classes start on Monday August 26, 2019 (CMU Academic Calendar) — Office hours and Hangouts can be found in the Google Calendar

All homework is due 10 minutes before lecture/recitation start.

| Date | Topic | Reading | Notes |
|---|---|---|---|
| | **Part 0: Introduction** | | |
| | **Week 1** | | |
| Monday August 26 | Course Overview | • CMU Computing Policy<br>• CMU Policy on Academic Integrity | slides |
| Wednesday August 28 | Conceptual Framework for Understanding Privacy | • Fair Information Principles<br>• (optional) Privacy in Context<br>• (optional) Overview Article in Stanford Encyclopedia of Philosophy<br>• (optional) Privacy as Contextual Integrity<br>• (optional) A Taxonomy of Privacy | slides |
| Friday August 30 | No recitation this week | | Homework 1 out |
| | **Part I: Privacy through Accountability: Formalization and Detection** | | |
| | **Week 2** | | |
| Monday September 2 | *No Class: Labor Day* | | |
| Wednesday September 4 | Enforcing Purpose Restrictions through Audit | • Formalizing and Enforcing Purpose Restrictions in Privacy Policies<br>• Purpose Restrictions on Information Use<br>• Privacy and Contextual Integrity: Framework and Applications<br>• Summary of the HIPAA Privacy Rule (Permitted Uses and Disclosures, Authorized Uses and Disclosures)<br>• A Formalization of HIPAA for a Medical Messaging System<br>• Experiences in the Logical Specification of the HIPAA and GLBA Privacy Laws | slides, Notes on MDPs |
| Friday September 6 | Recitation on Docker (Sruti) | | |
| | **Week 3** | | |

This is an excellent module o
privacy from Carne
University

You can find all lectur
and even exercises
https://course.ece

This quickly becom
technical.

# Useful resources (I)

Privacy in a Mobile-Social World

*CompSci 590.03*
*Instructor: Ashwin Machanavajjhala*



Ashwin Machanavajj has an
excellent course a
mobile world -
https://courses.cs

This quickly becom
technical.

# Assessment

# Assessment

In this section you must explain:

- What is an open dataset?
- What is "health data"?
- Why is it important to government, indust
  available?
- What are some benefits to individuals and groups
- What is data anonymisation and why is it
- What are the dangers to individuals and g
  deanonymised?
- What are some steps that can be taken to

> In answering these questions, you must in
> about data deanonymisation. Include as mu
> how the data was deanonymized.

# References


1. Kretschmer, M., Pennekamp, J. & Wehrle, K. Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. ACM Transactions on the Web 15, 1–42 (2021).
2. European Union. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. (2016).
3. Cookies, the GDPR, and the ePrivacy Directive. GDPR.eu (2019).
4. Englehardt, S. & Narayanan, A. Online Tracking: A 1-million-site Measurement and Analysis. in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security 1388–1401 (Association for Computing Machinery, 2016). doi:10.1145/2976749.2978313.
5. Habib, H., Li, M., Young, E. & Cranor, L. "Okay, whatever": An Evaluation of Cookie Consent Interfaces. in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems 1–27 (Association for Computing Machinery, 2022). doi:10.1145/3491102.3501985.
6. Spiegel, J. R., McKenna, M. T., Lakshman, G. S. & Nordstrom, P. G. Method and system for anticipatory package shipping. (2013).
7. Nyckel, E.-M. Ahead of Time: The Infrastructure of Amazon's Anticipatory Shipping Method. in Media Infrastructures and the Politics of Digital Time (eds. Volmar, A. & Stine, K.) 263–278 (Amsterdam University Press, 2021). doi:10.2307/j.ctv1xcxr3n.18.
8. Viet, N. Q., Behdani, B. & Bloemhof, J. Data-driven process redesign: Anticipatory shipping in agro-food supply chains. International Journal of Production Research 58, 1302–1318 (2020).
9. Venkataramanan, M. Madhumita Venkataramanan: My identity for sale. Wired UK (2014).
10. UK Government. National Health Service Act 2006. (2006).
11. UK Government. Data Protection Act 2018. (2018).
12. Information Commissioner's Office. Anonymisation: Managing data protection risk code of practice. (2012).
13. UK Government. Equality Act 2010. legislation.gov.uk (2010).
14. Sweeney, L. Weaving Technology and Policy Together to Maintain Confidentiality. The Journal of Law, Medicine & Ethics 25, 98–110 (1997).
15. Sweeney, L. Simple Demographics Often Identify People Uniquely. 0 Bytes (2000) doi:10.1184/R1/6625769.V1.
16. Sweeney, L. K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557–570 (2002).
17. Sweeney, L. Only You, Your Doctor, and Many Others May Know. Technology Science (2015).

18. Samarati, P. & Sweeney, L. Protecting Privacy when Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression. (1998).

19. El Emam, K. & Dankar, F. K. Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association : JAMIA 15, 627–637 (2008).

20. Narayanan, A. & Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. in 2008 IEEE Symposium on Security and Privacy (sp 2008) 111–125 (2008). doi:10.1109/SP.2008.33.

21. Machanavajjhala, A., Gehrke, J., Kifer, D. & Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. in 22nd International Conference on Data Engineering (ICDE'06) 24–24 (2006). doi:10.1109/ICDE.2006.1.

22. Bennett, J. & Lanning, S. The Netflix Prize. (2007).

23. Jackson, D. The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever. Thrillist (2017).

24. Narayanan, A. & Shmatikov, V. How To Break Anonymity of the Netflix Prize Dataset. (2006) doi:https://arxiv.org/abs/cs/0610105v1.

25. Waxman, S. Whoops! Netflix Gets Caught by FTC, Cancels Contest. (2010).

26. Narayanan, A. & Schmatikov, V. 'How to Break Anonymity of the Netflix Prize Dataset" - FAQ. (2007).

27. Zuckerberg, M. Our Commitment to the Facebook Community. Meta (2011).

28. Ho, E. Google Buzz in Gmail. Official Gmail Blog (2010).

29. Google Blog. A fall sweep. Official Google Blog (2011).

30. BuzzClassAction.com. Google Buzz User Privacy Litigation Class Action Settlement Website (2010).

31. Quividi. CONSUMER Privacy - Quividi - Insightful data without anonymous measurements. Quividi (2022).

32. Zuo, Z. et al. Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. JMIR Medical Informatics 9, e29871 (2021).