HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 2 — Data acquisition and open data. Data science ethics

Jorge N. Tendeiro

Hiroshima University

Today

- Data acquisition and open data.
- Data science ethics.

Data acquisition and open data

Data

Data is the foundation of data science:

> *Data is the material on which all the analyses are based.*

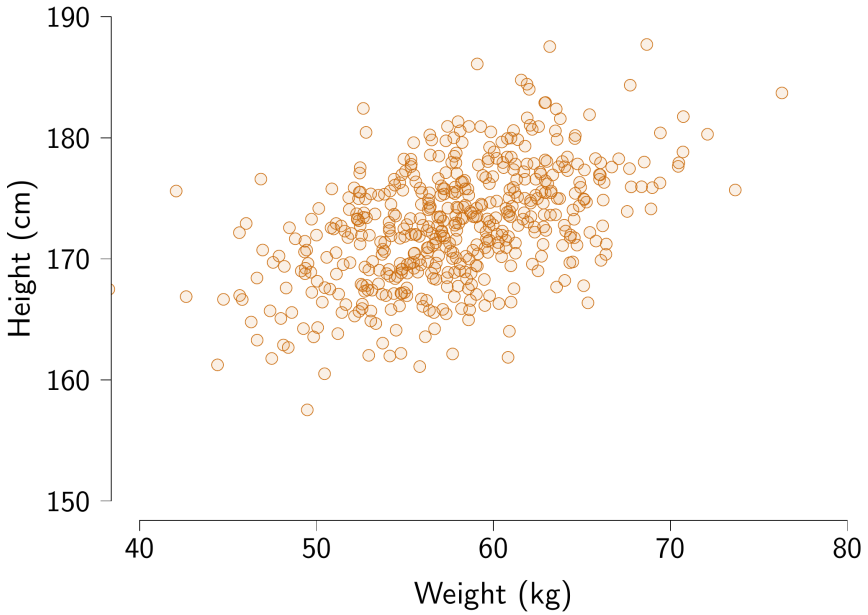
(Source)

Data are supposed to capture relevant features of what we are studying.
Some features are easier to observe and measure (e.g., age), whereas others are quite difficult (e.g., emotions).

Data typically must be digitalized so that they can be analyzed with the help of computers.

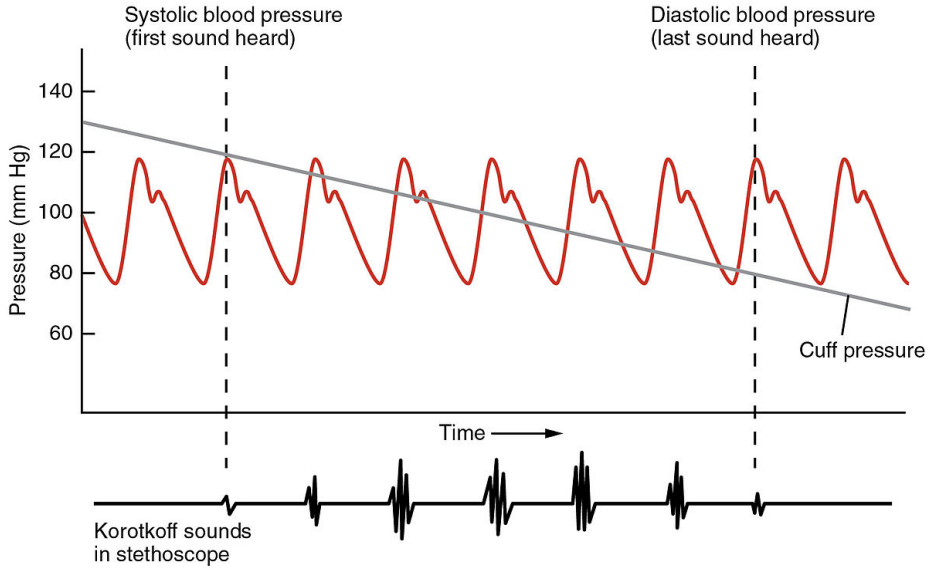
Examples of data

Height, weight:



Source: Kaggle.

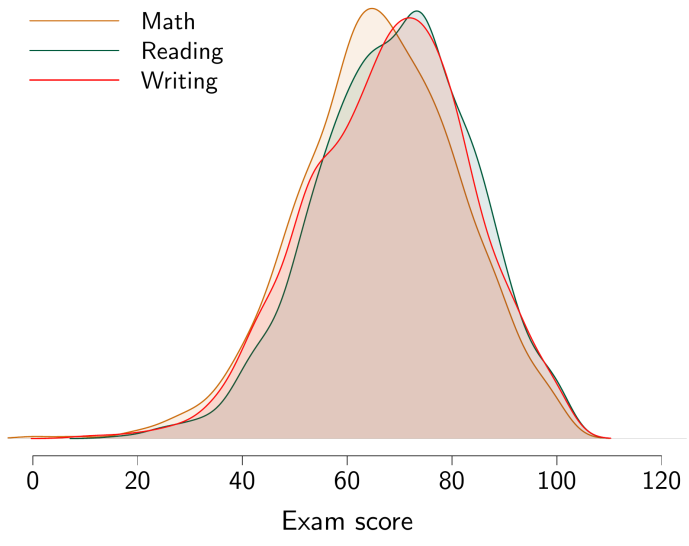
Blood pressure:



Source: Wikipedia.

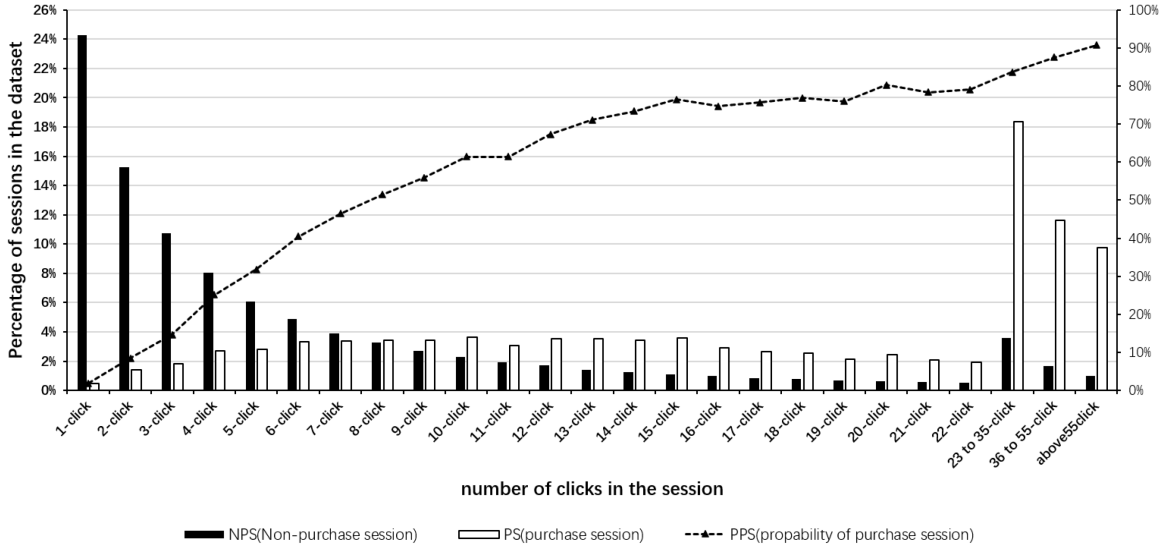
Examples of data

Exam scores:



Source: Kaggle.

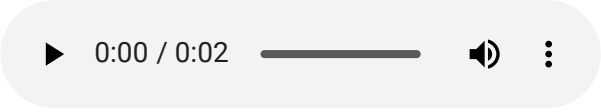
Marketing:

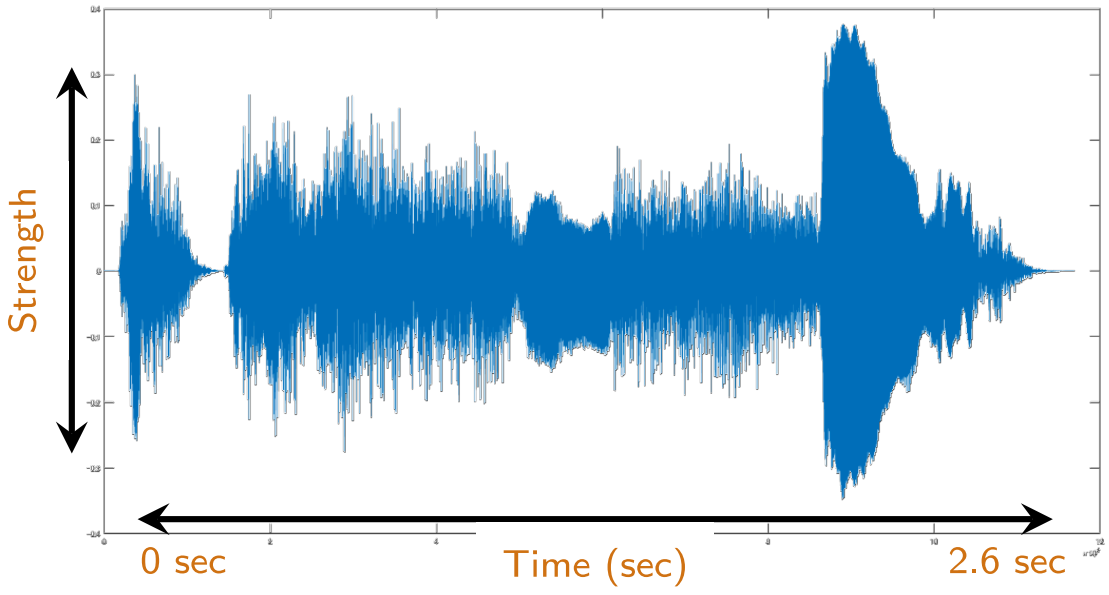


Source: Wen et al. (2023).

Examples of data (audio)

Audio data expressed as a wave:



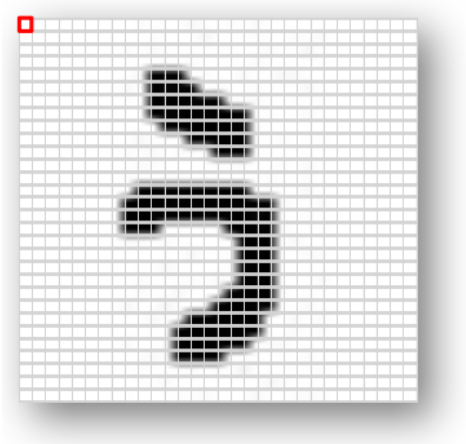


0.054262
0.037141
0.033052
0.033387
-0.0034181
-0.031831
-0.036073
-0.036653
-0.059999
-0.09238
-0.071169
-0.06357
-0.085177
-0.011414
0.04416
0.0053713
0.056154
0.12592

Numeric data of strength (about 120,000 numeric values in 2.6 sec).

Examples of data (image)

Hand-written letter:



Split an image into pixels (small squares), and quantify the shading of the color in each pixel.

Numerical data (0: black $\sim$ 255: white):

255	255	255	255	255	255	255	255	255	251	255	254	255	248	255	255	252	255	250	255	254	255	255	254	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	253	255	255	255	246	255	254	253	252	255	255	251	255	251	255	255	255	255	255	255
255	255	255	255	255	255	255	255	252	255	250	254	255	254	254	255	253	255	255	252	254	255	247	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255	255	251	255	255	254	255	250	255	255	254	255	253	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	247	0	0	0	255	252	255	255	249	255	255	248	249	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	249	255	2	7	3	0	255	250	255	255	252	255	255	255	248	254	255	255	255	255	255	255
255	255	255	255	255	255	255	255	254	255	0	0	0	5	4	1	251	253	250	254	255	253	255	253	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	0	5	5	0	4	0	0	5	255	253	255	254	251	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	0	0	0	0	0	0	5	250	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	247	255	253	255	0	5	0	3	0	254	255	251	255	254	253	255	255	255	255	255	255
255	255	255	255	255	255	255	255	251	255	249	255	251	251	254	1	2	0	255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	251	252	255	255	255	255	252	253	250	255	255	252	255	250	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	252	253	255	254	252	255	253	255	255	255	254	252	254	255	253	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	4	0	0	2	0	0	0	0	0	255	252	255	255	249	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	0	0	0	0	6	2	0	2	1	2	0	8	249	253	255	251	255	255	255	255	255	255
255	255	255	255	255	255	255	255	1	0	5	1	0	0	6	0	1	0	8	0	255	255	252	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	1	1	0	254	253	255	254	253	2	0	2	1	251	255	255	253	255	255	255	255	255	255
255	255	255	255	255	255	255	255	250	250	255	251	255	255	255	255	255	0	0	1	255	248	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255	255	247	254	254	253	0	4	0	253	255	251	252	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	255	249	254	255	255	255	253	7	0	0	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	252	254	255	255	252	255	250	253	0	4	0	255	255	255	250	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	251	255	255	247	255	255	255	5	1	0	0	252	255	254	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	248	255	255	244	255	255	252	0	0	1	5	0	255	255	250	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	249	255	251	0	1	1	1	0	0	255	255	249	255	251	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	254	255	0	1	1	0	0	3	0	255	255	251	255	252	255	255	255	255	255	255
255	255	255	255	255	255	255	255	250	255	255	251	4	0	0	5	2	0	254	252	250	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	253	249	255	0	5	1	0	254	255	253	255	254	255	251	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	253	255	255	253	250	255	252	255	242	255	255	255	254	255	252	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	254	250	255	255	255	255	255	255	255	248	251	254	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	254	255	255	255	253	254	251	255	255	255	248	255	255	253	251	253	255	255	255	255	255	255

Obtaining data

There are various methods available to collect data.

- Collect the data by yourself.

> *E.g.: Distribute questionnaires, perform experiment/observation.*

✓ You collect the data that you want.

✗ It can take a lot of time and effort.

- Use data collected by someone else.

> *E.g.: Request materials, download publicly available data online.*

✓ Easier to obtain.

✗ The data is not always available or shared.

✗ The data need not exactly match what you want.

Open data

> *Open data is data that is free to access, use, modify, and share by anyone — subject, at most, to measures that preserve provenance and openness.*

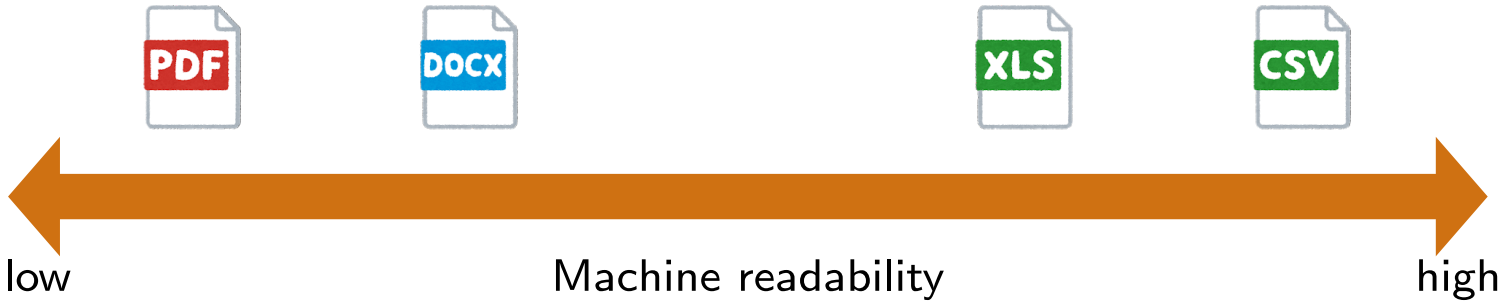
Open Definition

Requirements that open data must meet (source):

- *Open license*
  Data must be available in the public domain or provided under an open license allowing secondary use.
- *Access free of charge*
  Data must be provided online and free of charge
  (or at most at no more than a reasonable one-time reproduction cost).
- *Machine readability*
  Data must be provided in a form readily processable by a computer (see next page for details).
- *Open format*
  Data must be provided in an open format, placing no restrictions, monetary or otherwise, upon its use.
  Data can be fully processed with at least one open-source software tool.

Machine readability

> *Machine readability: Ease with which files can be imported or edited by computer programs.*



In data analysis, tabular files (e.g., XLS, TXT, CSV) are easy to work with.
CSV files are often used in R.

Important

> *Even with CSV files, you need to first clean data before analyzing them!*

It is a part of data science task to convert data into a format which is machine readable.

Benefits to open data

1. Solving various issues and revitalizing the economy through the promotion of public participation and public-private collaboration.

$\longrightarrow$ *Creation of new businesses and services.*

2. Sophistication and efficiency of administration.

$\longrightarrow$ *Enhancement of public services.*

3. Improve transparency and reliability.

Reference: Guidelines for open data https://cio.go.jp/sites/default/files/uploads/documents/data_shishin.pdf.

Open data — Also good to keep in mind

- Reproducibility
  Results can be reinspected after publication. Mistakes are easier to find.
- Accountability
  We do not need to "take someone's word for it" (i.e., to believe without proof).
- Trust
  Open data is a means to show that our work is trustworthy and honest.
- Contribution
  Open data is a way to give back to the society.
- Preserve the scientific record
  Open data is a means to not lose valuable information.
- Save money
  Open data facilitate future data collections, minimizing waste.

Sources: here, here.

Open data — Examples

There is a boom of projects which rely on access to open data.
Below are a couple of interesting sources to read about this:

- https://citizens-guide-open-data.github.io/guide/3-open-datas-use.
- https://www.luzmo.com/blog/examples-of-open-data-in-government.

I will highlight three examples.

Open data — Example 1: Public health and environment

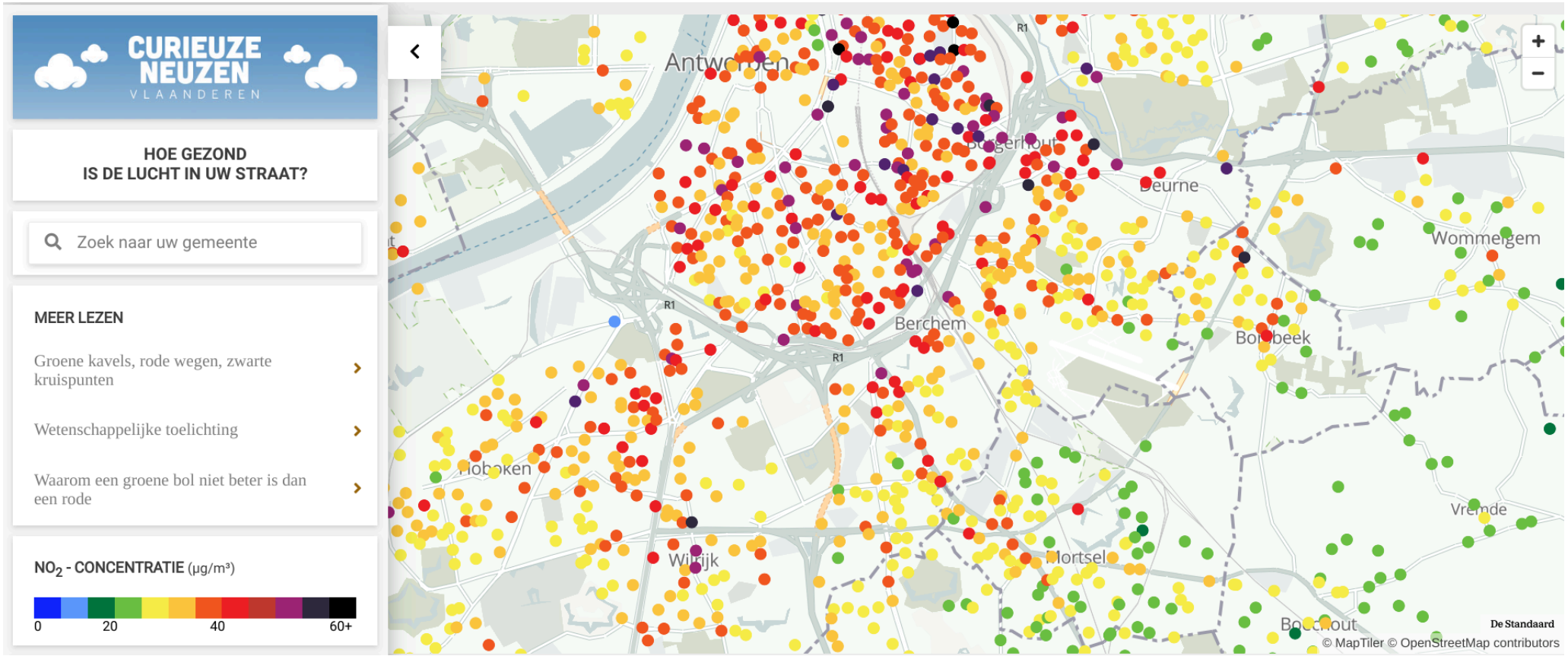
CurieuzeNeuzen is a citizen participation project in Belgium.

Goal
Participate in a large-scale research of air quality in Flanders.

Outcome
They developed an interactive app that monitors the air quality in Flanders.

Data are contributed by each citizen!

Open data — Example 1: Public health and environment



Source: https://www.standaard.be/curieuzeneuzen/map/#11.99/51.19321/4.40892

Open data — Example 2: Helping the police

See the full blog post here:
Parking Data, Ben Wellington.

Ben Wellington used the parking ticket data set from the New York Police Department's open data portal to identify a problem:

> *The police had been giving parking tickets for cars that were **legally** parked.*

Mr. Wellington used the open parking ticket data to prove that the police officers were not following a new traffic rule.

This saved the community from paying millions of dollars of unfair fines!

Open data — Example 2: Helping the police

Follow iquantny

I Quant NY

MAILING LIST RSS ARCHIVE

Quantitative Analysis of NYC Open Data: Every data set that the city releases tells a story. This blog is all about telling those stories, one data set at a time.

About Me About You Interviews Press Topics Subscribe

MAY 11, 2016

The NYPD Was Systematically Ticketing Legally Parked Cars for Millions of Dollars a Year- Open Data Just Put an End to It

New York City is a complex place to drive. And when it comes to parking, there are plenty of rules and regulations to follow. It's no wonder that sometimes people get confused and end up getting their cars ticketed or towed.

But in all of these rules, there is one thing that very few drivers seem to know. As of late 2008, **in NYC you can park in front of a sidewalk pedestrian ramp, as long as it's not connected to a crosswalk.** It's all written up in the NYC Traffic Rules, and for more detail, take a look at this article. The local legislation making these parking spots legal was proposed by Council Member Gentile, and adopted by the Department of Transportation before it ever made it for a vote. Though few people seem to know about the change.

Open data — Example 3: Spotify and MusicBrainz

Spotify is a popular digital music, podcast, and video streaming service.

Spotify relies on MusicBrainz to collect music metadata (like the name of the artists, album, year, music titles, etc.).

MusicBrainz is an open music encyclopedia that makes all of its data available to the public.

Also **you** can use MusicBrainz, for example, when setting up your own home media center (using for instance Jellyfin, Kodi, or Plex)!

Where to get open data?

There are many repositories.

In Japan:

- Websites to obtain open data:
  - e-Stat 政府統計の総合窓口
    https://www.e-stat.go.jp
  - DATA GO.JP データカタログサイト
    https://www.data.go.jp
  - e-GOV
    https://data.e-gov.go.jp/info/ja

- Example of analyzing open data:
  - Open data 100
    (Government CIO's portal Japan).
    https://cio.go.jp/opendata100

In the World:

- Websites to obtain open data:
  - 538 abc NEWS
    https://abcnews.go.com/538
  - kaggle
    https://www.kaggle.com/datasets
  - Open Science Framework
    https://osf.io/
  - ...

Data science ethics

Collecting data

I want to broadcast about the approval rate of the governor of Hiroshima among Hiroshima citizens during tomorrow's news.

Let's ask people at Hiroshima station. We tell upfront that the survey is to broadcast the approval rating of the governor at my news.

(After broadcast) To prove that I didn't fake it, let's publish the collected data on our website!

ID	Name	Sex	Approve (0=no, 1=yes)
1	Mike Brown	male	1
2	Sue Allen	female	0
3	John Watson	male	1
4	Tom Hanks	male	1
5	Norah Jones	female	0
6	Scarlett Johansson	female	0
7	Anne Hathaway	female	0
8	Meryl Streep	female	1

Looking at the published data, do you see any problems?

There are **many problems**!!!

Ethics in data collection and utilization

Ethical considerations in data collection and utilization:

- It is important to obtain **informed consent** after explaining the purpose of data collection and how the data will be utilized.

  Informed consent (medical):

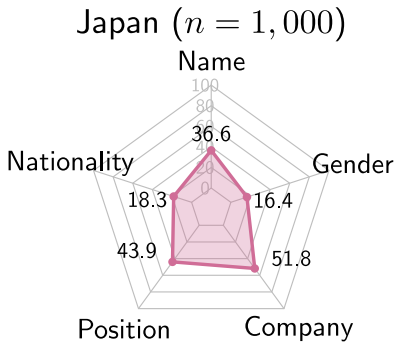
  > *A process in which patients are given important information, including possible risks and benefits, about a medical procedure, so that patients can decide if they want to take part in the trial.*
  >
  > *(National Cancer Institute, NIH)*

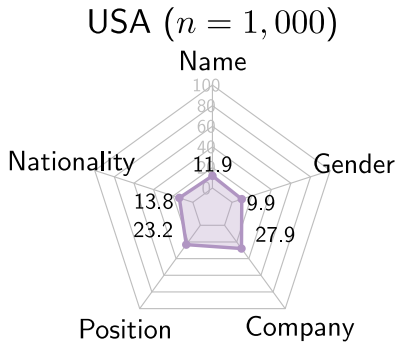
- Build a system that allows participants to withdraw their data easily, if they are willing to.
- Note that each participant has a different background and views on which data to be kept personal.
  $\longrightarrow$ When collecting data, we need to consider the *social and cultural background* of the data subjects.

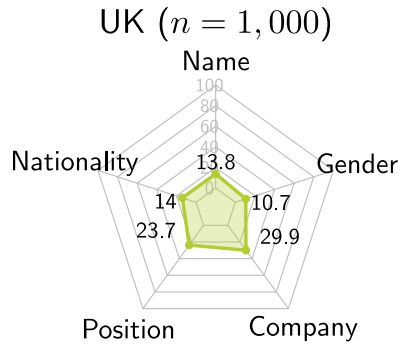
Also, keep in mind that the information that people don't want to disclose to others varies across regions or countries.

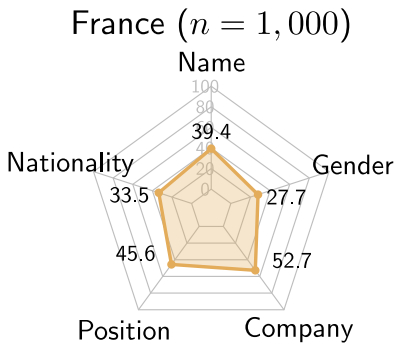
Survey: "Data I don't want to provide or disclose under any circumstances"

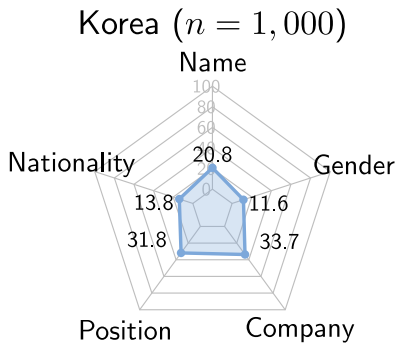
General personal data (not highly sensitive personal data; $n = 1,000$):

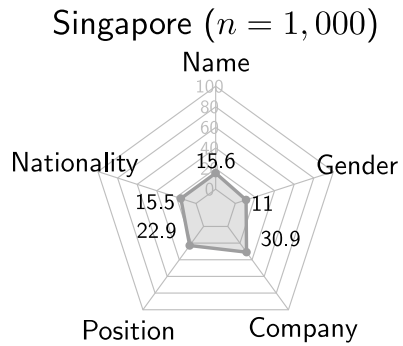








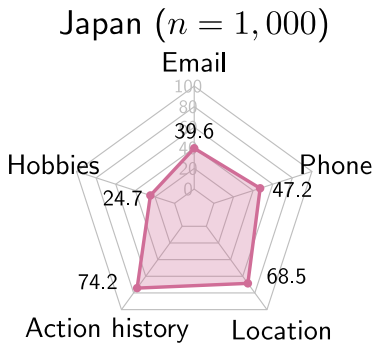


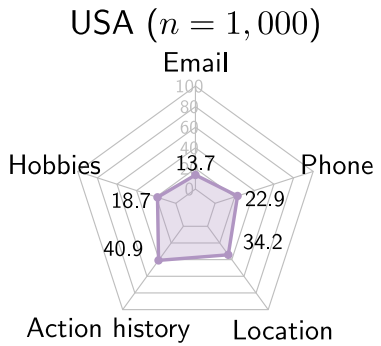


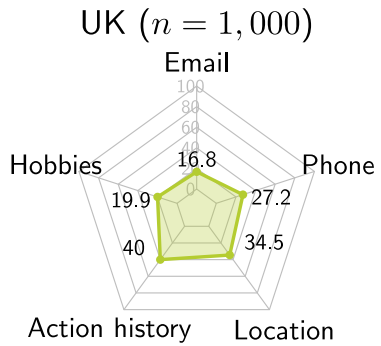
Source: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

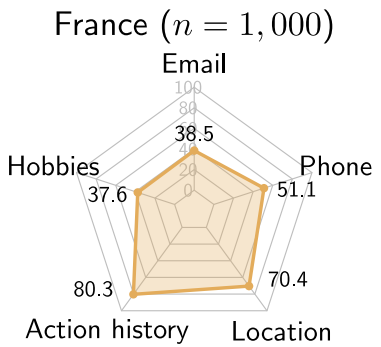
Survey: "Data I don't want to provide or disclose under any circumstances"

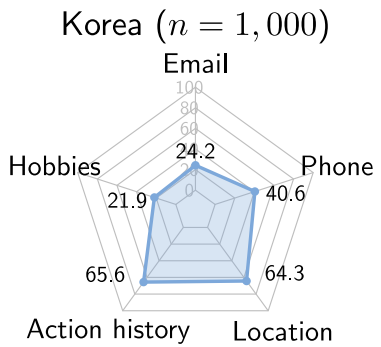
Sensitive personal data (data requiring careful handling; $n = 1,000$):

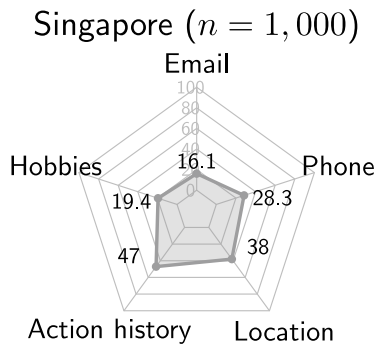








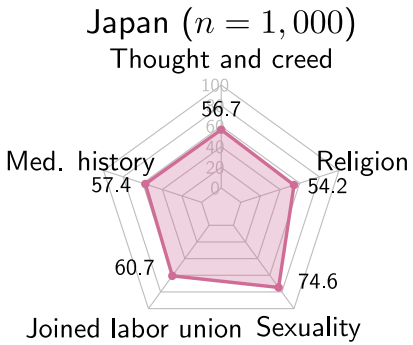


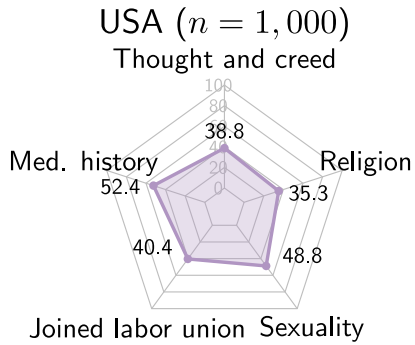


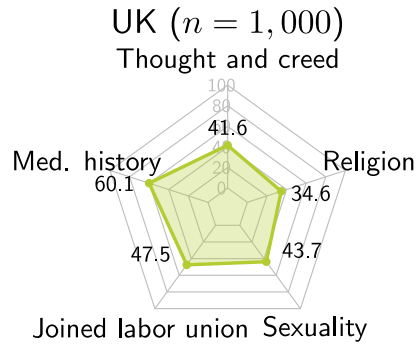
Source: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

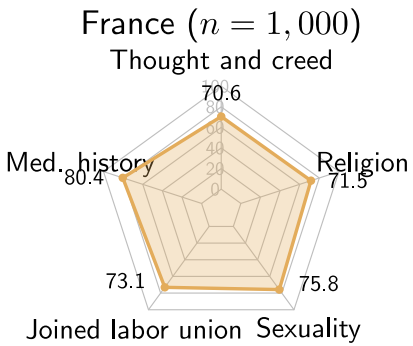
Survey: "Data I don't want to provide or disclose under any circumstances"

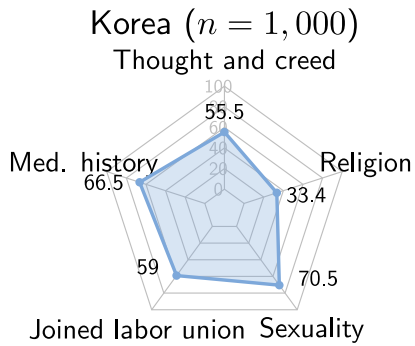
Highly sensitive personal data ($n = 1,000$):

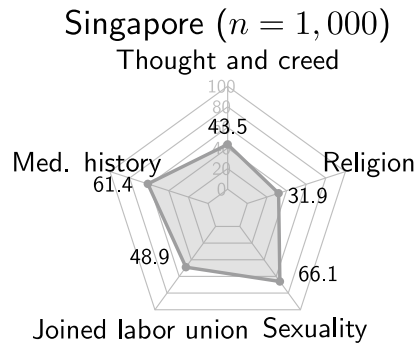












Source: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

Anonymization and pseudonymization of data

Anonymizing data:

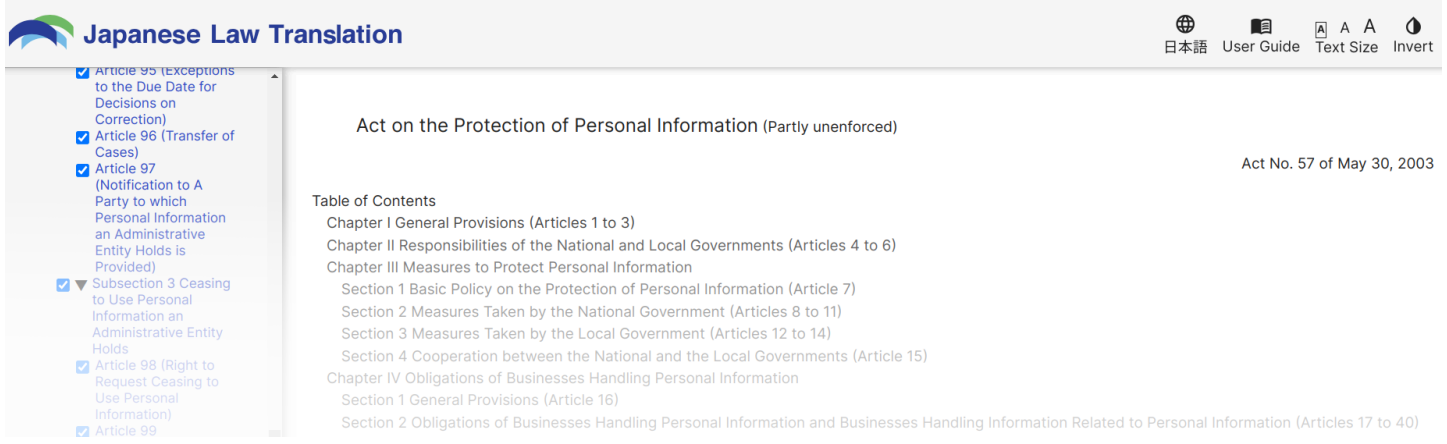
> *Making it impossible to identify an individual.*

Example:
Delete personal identifiable elements (name, address, etc.).

Pseudonymizing data:
Make sure that an individual is not identifiable *without additional information*.

Example:
Encrypt person-identifiable information (e.g., name) by ID.

Anonymization and pseudonymization of data



From the Act on the Protection of Personal Information (APPI):

- Anonymized data:
  In principle, it *can be provided* to a third party without consent.
- Pseudonymized data:
  In principle, it is *impossible to provide* to a third party without consent.

Anonymization and pseudonymization of data: Example

ID	Name	Age	Disease
235125	Mike Brown	103	diabetes
552412	Nina Brown	83	cancer
100485	Nicole Kidman	23	pneumonia
232039	Frank Jones	26	pneumonia

Age	Disease
103	diabetes
83	cancer
23	pneumonia
26	pneumonia

Impossible to identify an individual

ID	Age	Disease
235125	103	diabetes
552412	83	cancer
100485	23	pneumonia
232039	26	pneumonia

By connecting IDs and names, possible to identify an individual

Anonymization and pseudonymization of data: Example

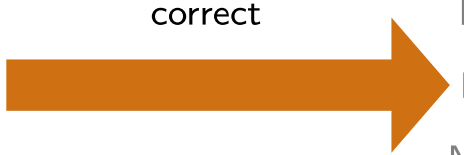
Even if you think that you have anonymized or pseudonymized data sufficiently, it may still be possible to identify an individual.

Example:
When only one person has age over 100 in the data.

	Age	Disease
Mike	103	diabetes
Nina	83	cancer
Nicole	23	pneumonia
Frank	26	pneumonia

Possible to identify Mike

correct



	Age	Disease
Mike	Over 80	diabetes
Nina	Over 80	cancer
Nicole	20~30	pneumonia
Frank	20~30	pneumonia

Impossible to identify Mike

> ***Always*** *pay close attention to personal identification!!*

So, do you now see any problems?

I want to broadcast about the approval rate of the governor of Hiroshima among Hiroshima citizens during tomorrow's news.

Let's ask people at Hiroshima station. We tell upfront that the survey is to broadcast the approval rating of the governor at my news.

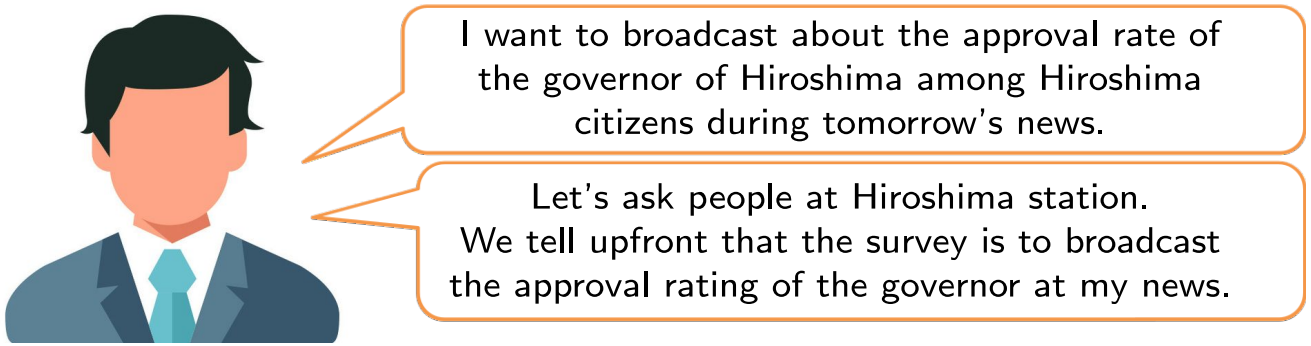
(After broadcast) To prove that I didn't fake it, let's publish the collected data on our website!

ID	Name	Sex	Approve (0=no, 1=yes)
1	Mike Brown	male	1
2	Sue Allen	female	0
3	John Watson	male	1
4	Tom Hanks	male	1
5	Norah Jones	female	0
6	Scarlett Johansson	female	0
7	Anne Hathaway	female	0
8	Meryl Streep	female	1

Problems:

- Individuals are clearly identifiable.
  Anonymization/pseudonymization is needed.
- Insufficient informed consent.
  Persons must be carefully informed about, and informed consent must be obtained afterwards.

Other sorts of problems



Although you want to know about the approval rate of all Hiroshima citizens, data were only collected at Hiroshima station.

The estimated approval rate is **biased**:

> *The results do not reflect **all** residents of Hiroshima.*

Selection bias

Selection bias:

> *Bias caused by an inappropriate way of collecting data.*
> *As a result, the collected data do not reflect the population at large.*

Example:
Study the *prevalence rate* of a certain disease among Japanese people.

Data collection plan:
Collect data from patients who visited hospitals.

Q: What is the problem?

A: Many people who go to the hospital are already sick.
The prevalence rate of such a sample will be **biased** upwards (inflated, that is, larger than it actually is in the population).

Information bias

Information bias:

> *Problems related to confusing information or improper measurement methods.*
> *As a result, the observed measurements may be biased.*

Example:
Ask people *"Did you catch a cold last month?"*

✗ Only people who recognized their symptoms as a "cold" would answer "yes" to this question.

Example:
Ask people *"How much savings do you have?"*

✗ People may answer more than their actual savings (e.g., to conceal financial problems).

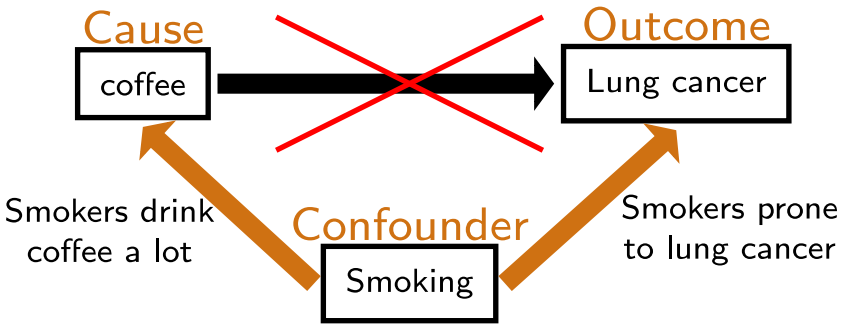
Confounding bias

Confounding bias:

> *Bias caused by factors (**confounders**) affecting both a 'cause' and an 'outcome'.*
> *This could lead to a **wrong** research outcome.*

Example:
Does drinking coffee increase the risk of lung cancer?
✗ Well, *not necessarily*.

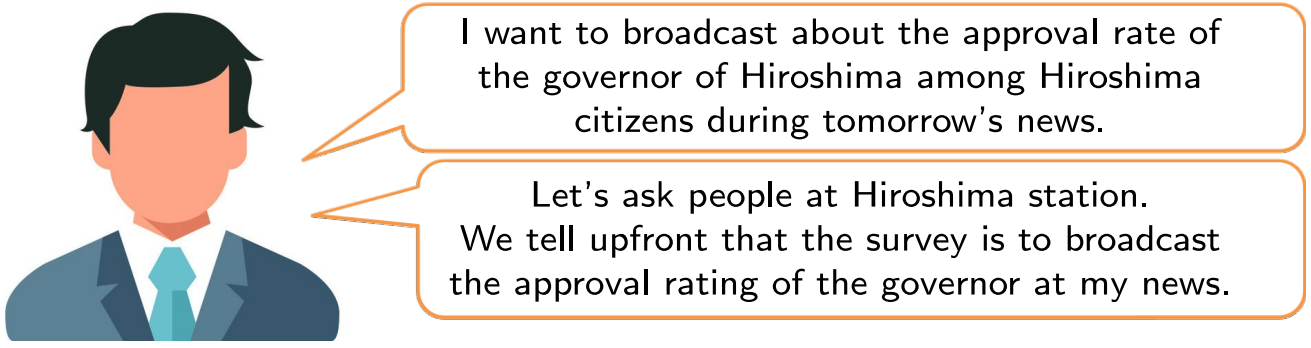
Here is an alternative *plausible* explanation:



The relation between "coffee" and "lung cancer" no longer seems causal once "smoking" is taken into account:

> *Smoking is a **confounder**.*

Recap: Which bias is taking place here?



Although you want to know about the approval rate of all Hiroshima citizens, data were only collected at Hiroshima station.

The estimated approval rate is **biased**:

> *The results do not reflect **all** residents of Hiroshima.*

Selection bias:
It is important to sample more randomly from the entire Hiroshima Prefecture.

There are many other biases. To reduce biases, we need to think **carefully** about how to collect data.

Summary

Data acquisition and open data:

- Data is crucial for data science!
- Open data is convenient, responsible, and is easy to use.

Data science ethics:

When collecting data by yourself and utilizing them, be aware of the following:

- Ethical considerations.
- Anonymization and pseudonymization of data.
- Various biases.