

Data types and sources

Sonja Aits

Lund University
2025-09-22



What kind of data do we have in
medicine/life sciences?

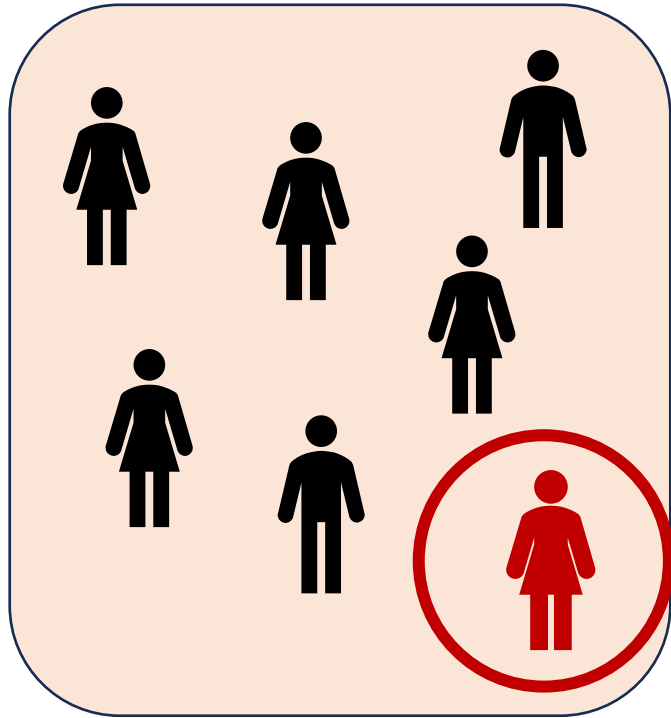
What data do you use?

Take a pen and write it down

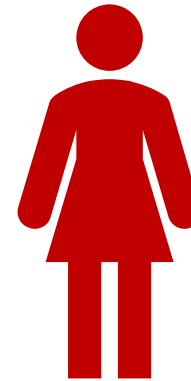
What is data?

What is data?

Dataset



Example



Feature
Age
Sex
Height
Weight
Diabetic

Value
35
Female
1.70 m
78 kg
Yes

Types of data

Big data \leftrightarrow Small data

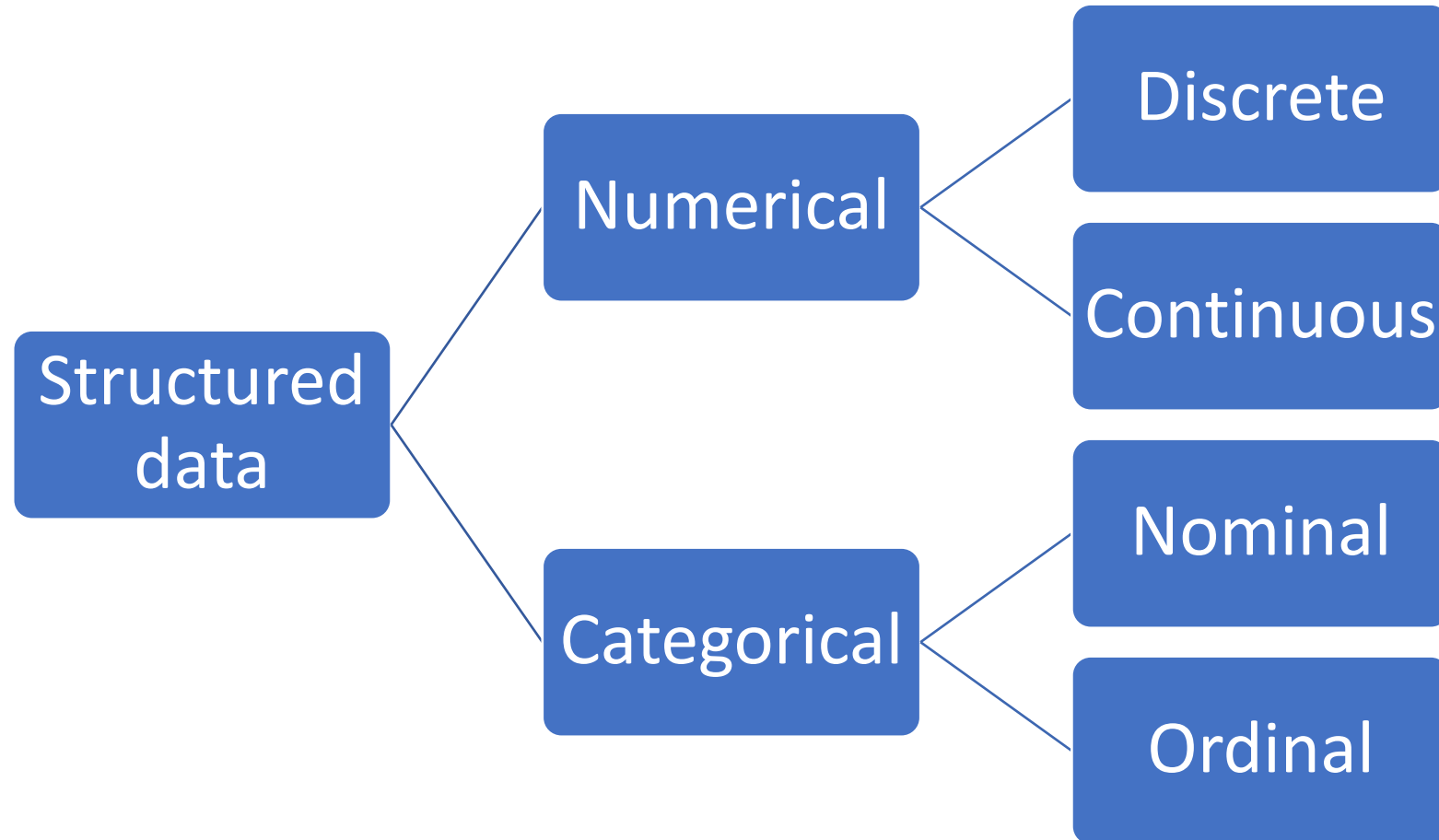
Single instance \leftrightarrow Sequence

Labelled (annotated) \leftrightarrow Unlabelled

Unstructured \leftrightarrow Structured

Sensitive \leftrightarrow Not sensitive

Types of data



What type of data do you have?

Write it down for one of the data types you use

How we perceive the world



How computers perceive the world

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

Finding data

- Search engines
 - <https://datasetsearch.research.google.com/>
 - <https://emedia.lub.lu.se/db>
- Collections of databases and repositories
 - <https://www.re3data.org/>
 - <https://fairsharing.org/databases/>
 - <https://snd.gu.se/en/find-data/international-data>
 - <https://www.oxfordjournals.org/nar/database/a/>
 - https://en.wikipedia.org/wiki/List_of_biological_databases
 - <https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories>
 - <https://www.ebi.ac.uk/>
 - <https://www.ncbi.nlm.nih.gov/>
- Data repositories
 - <https://zenodo.org/>
 - <https://figshare.com/>
 - <https://www.covid19dataportal.org/>
 - <https://idr.openmicroscopy.org/>

Finding data

- Machine learning data collections
 - <https://www.kaggle.com/datasets>
 - <https://huggingface.co/docs/datasets/>
 - https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
 - https://en.wikipedia.org/wiki/Biomedical_text_mining#Corpora
- Scientific databases
 - <https://www.gbif.org/species/> (species)
 - <https://www.rcsb.org/> (protein structures)
- Research infrastructures and consortia
 - <https://www.ukbiobank.ac.uk/>
 - <https://gdc.cancer.gov/>
- Governmental and non-governmental organizations
 - <https://www.who.int/data/collections>
- Data-centric journals
 - <https://www.sciencedirect.com/journal/data-in-brief>

Swedish data

- <https://www.lupop.lu.se/data-at-LU>
- [SoS](#) (Socialstyrelsen)
- [SCB](#) (Statistikmyndigheten)
- <https://rut.registerforskning.se/>
- <https://snd.gu.se/>
- <https://www.pathogens.se/>

Sensitive data must be handled with care

- Personally identifiable information (name, person number, email, phone, etc.)
- Sensitive personal data (health information, biometrical data, ethnicity, race, sexual orientation, etc.)
- Classified data
- Confidential data

→ special requirements for handling – cannot be processed on most computational infrastructure

→ **Avoid if possible!**

Getting your own data

Survey tool

[Employment](#) [Support and tools](#) [Research and Education](#) [Organisation and Governance](#)

Start › Research and Education › Education support › Quality assurance and enhancement › Sunet Survey – a survey tool

Research support

Education support

- ▶ Study administration
- ▼ Quality assurance and enhancement
 - Evaluation of higher education
 - Validation – quality assurance of new programmes and courses
 - National evaluations
 - Surveys
 - Course evaluations and course evaluation reports
 - Publications
 - Sunet Survey – a survey tool
 - Alumni Relations
- Training in teaching and learning in higher education
- ▶ Education Board
- ▶ Disciplinary matters
- Student rights
- ▶ Student support

[Se sidan på svenska](#)

Sunet Survey – a survey tool

Sunet Survey is a survey tool used at Lund University for online surveys within quality assurance and research.

Sunet Survey is a program for conducting online surveys. At Lund University, the programme is used for course evaluations and other types of surveys.

The programme facilitates creating questionnaires and distributing surveys through a link or by email. Sunet Survey is free of charge to staff and students of Lund University.

All employees have user authorisation in the system. Log in using your Lucat identity.

Students use their student account (StiL) to log in, but must initially contact the system administrator to obtain authorisation.

Researchers who wish to use Sunet Survey for collecting research data should contact the system administrator for further information.

Sunet Survey is a commercial IT system that has been procured in accordance with the Swedish Public Procurement Act.

[Log into Sunet Survey](#)

Page Manager: federica.savino@stu.lu.se | 21 March 2019

Citizen science annotations

Cell Hunter Game

Home
Help
Info

Game Starts

salma



☒ Normal
☐ Weird

Save Exit

Artificial data can often replace/complement real data

- Simulated measurements to train outlier detection
- Artificial images to train object detection
- Fake medical records to train summarization models
- ...

Benchmarking datasets

- High quality
- Often published for scientific competitions
- Leaderboards for tracking progress

Examples:

Computer vision: <https://www.image-net.org/>

BioNLP: <https://biocreative.bioinformatics.udel.edu/>

State-of-the-art (SOTA) performance

Websites tracking SOTA

- <http://nlpprogress.com> (text)
- https://github.com/syhw/wer_are_we (speech)
- <https://github.com/JunMa11/SOTA-MedSeg> (medical image segmentation)
- <https://paperswithcode.com/>

Community challenges (“shared tasks”)

- Kaggle
- Conferences

Recent research articles

Take home message

Strategic data acquisition is an ESSENTIAL part of most AI projects