Data types and sources

Sonja Aits

Lund University
2025-09-22

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

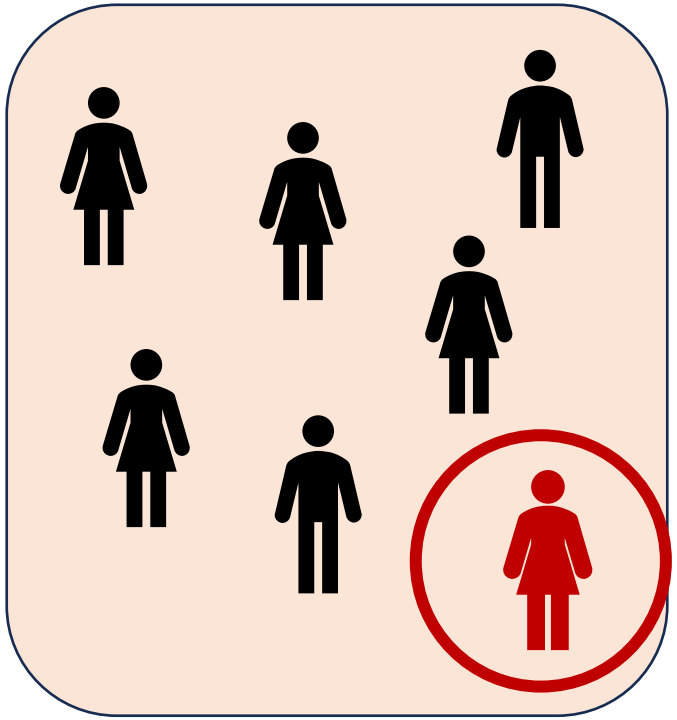
What data do you use?

Take a pen and write it down

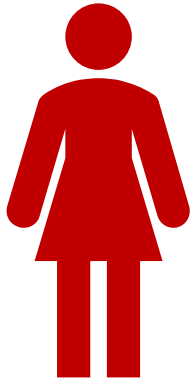
What is data?

What is data?

Dataset



Example



Feature	**Value**
Age	35
Sex	Female
Height	1.70 m
Weight	78 kg
Diabetic	Yes

Types of data

Big data ↔ Small data

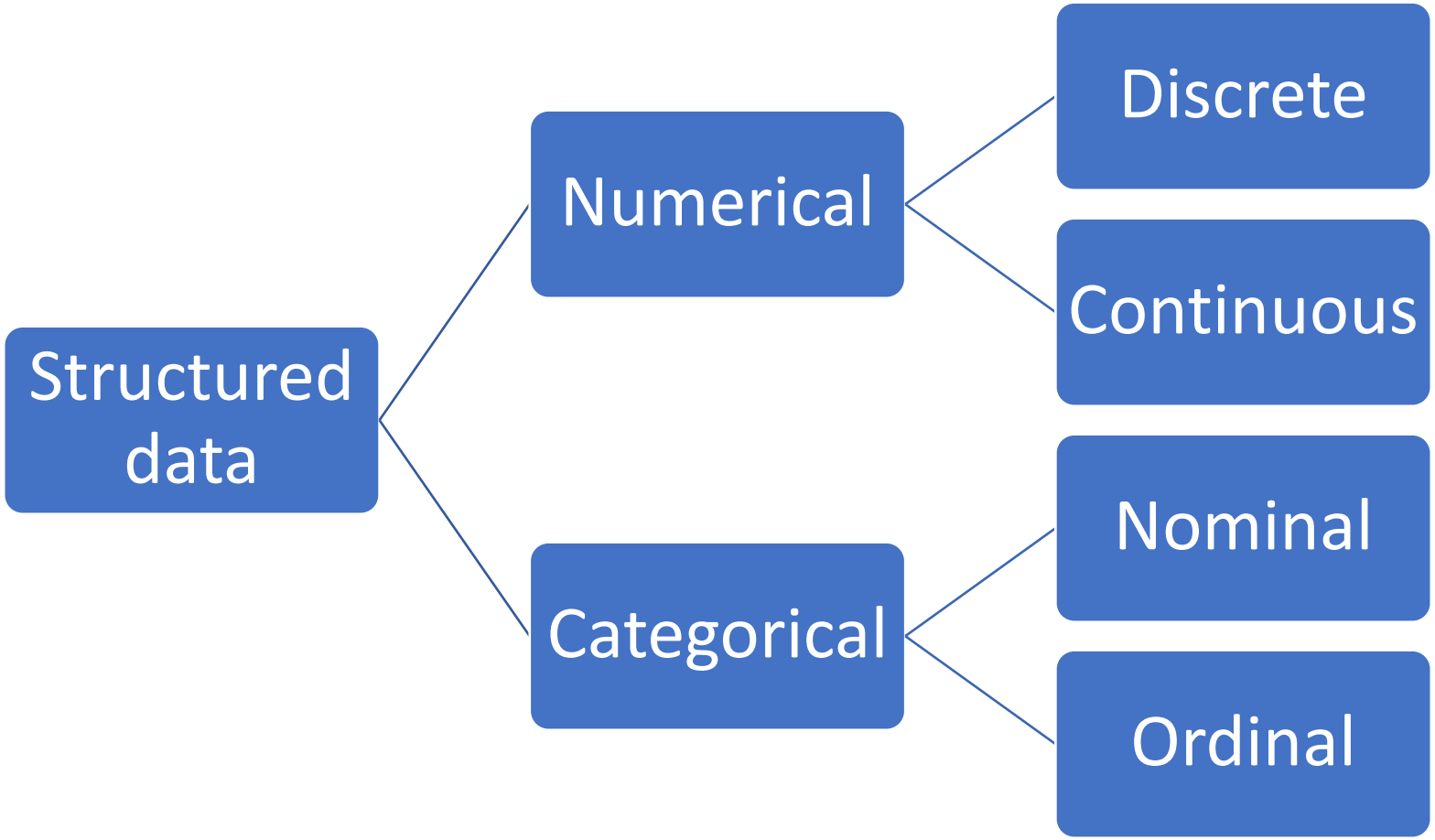
Single instance ↔ Sequence

Labelled (annotated) ↔ Unlabelled

Unstructured ↔ Structured

Sensitive ↔ 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
  - ttps://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

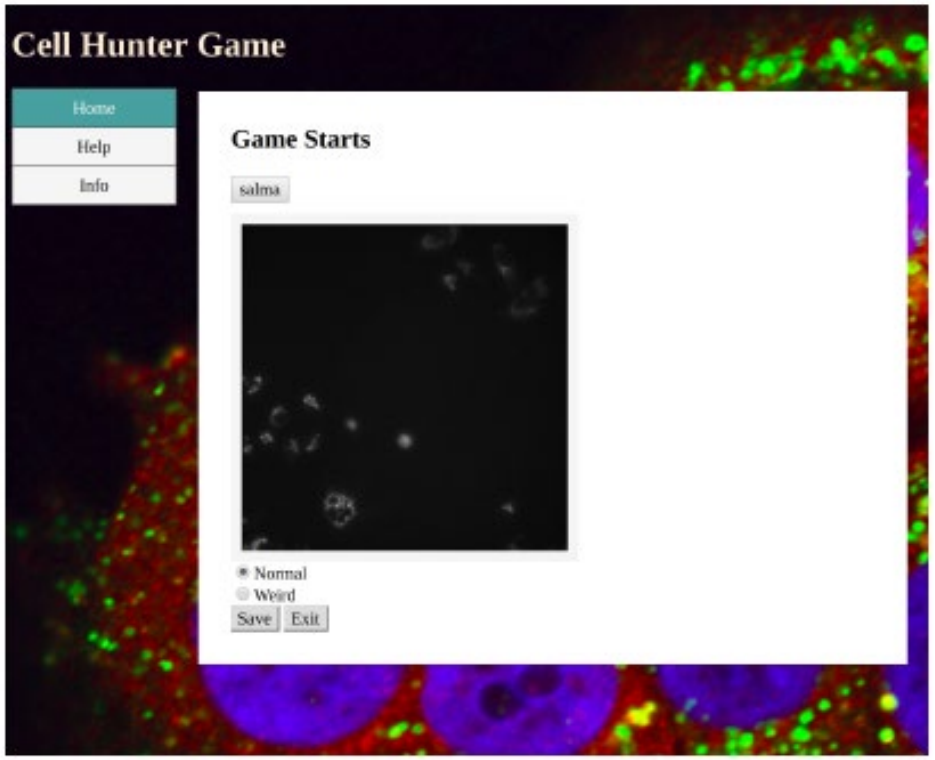
→ special requirements for handling – cannot be processed on most computational infrastructure

→ **Avoid if possible!**

Getting your own data

Survey tool

Citizen science annotations



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