Lecture 1: Medical Databases

CSCI6410/EPAH6410/CSCI4148

Finlay Maguire (finlay.maguire@dal.ca)

Learning Objectives

- Overview of the types of medical database
- Ways of maintaining data privacy with medical databases and some of their trade-offs
- How and why ontologies and survey weights are used in medical databases

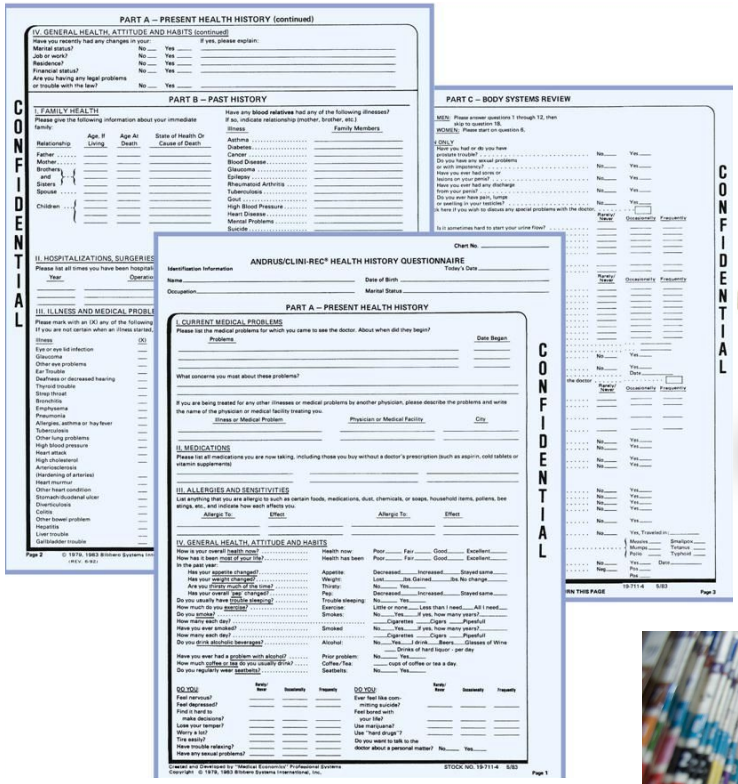
- Key strategies/approaches for exploratory data analysis
- Different types of dimensionality reduction
- Basics of supervised learning
- Accessing feature importances
- Aggregating simple/weak models to improve performance: boosting and bagging

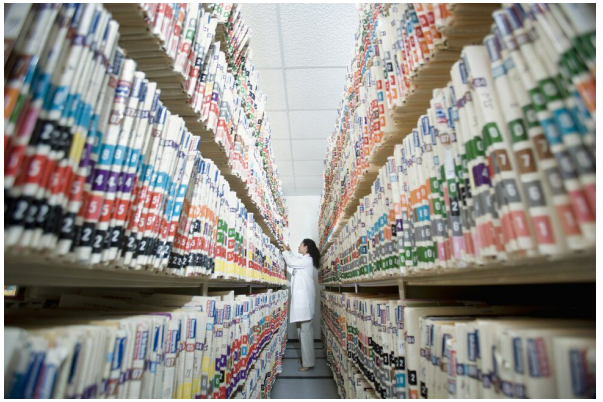
What is a database?

Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts

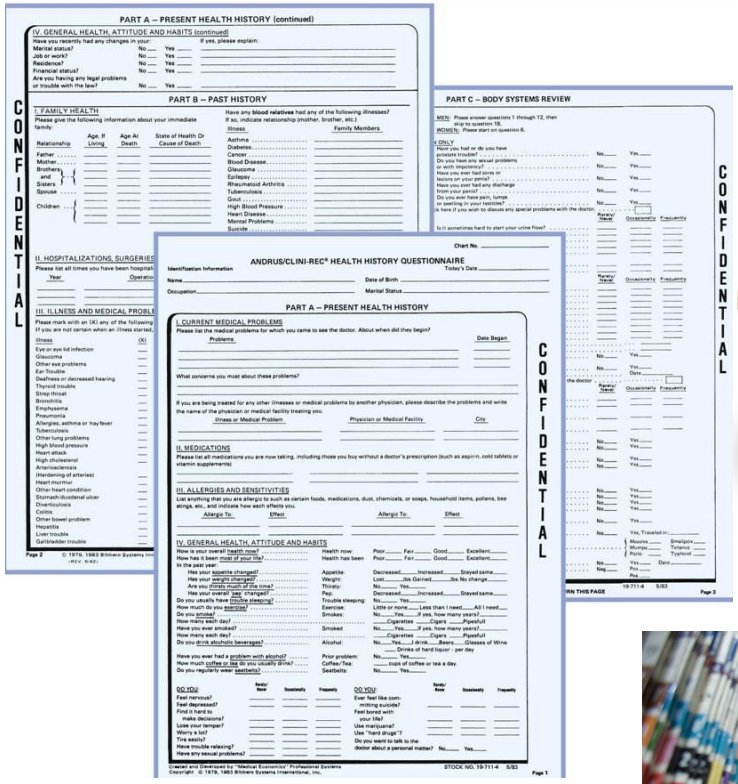


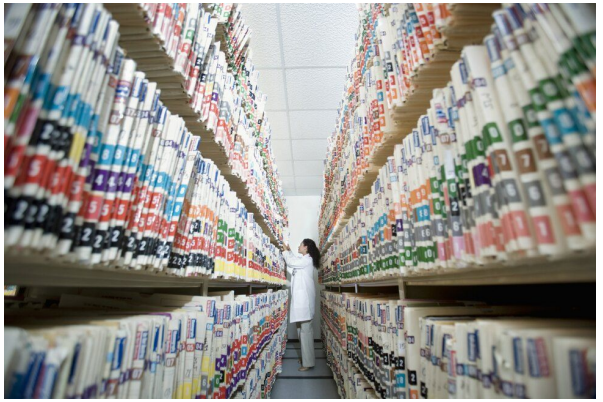


Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet





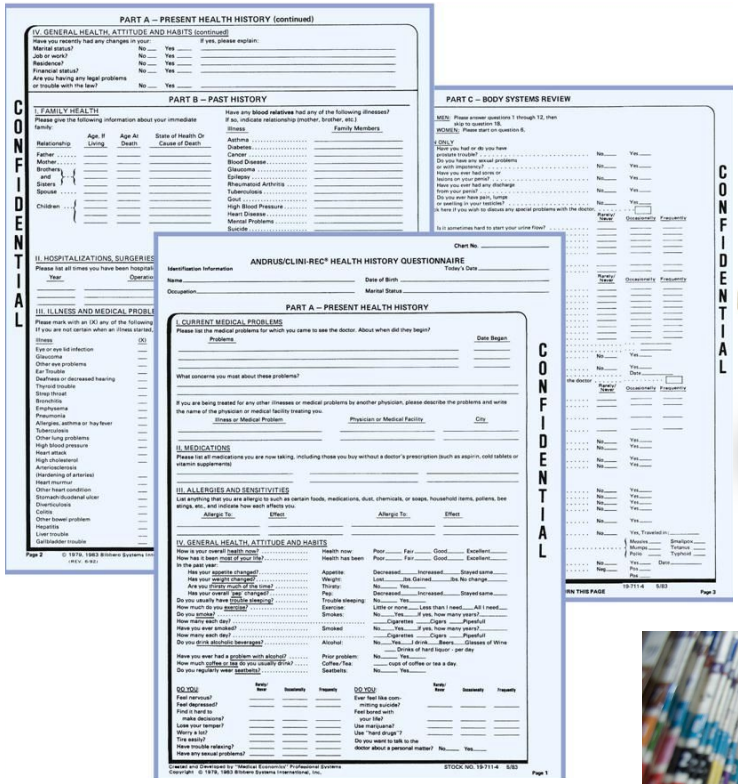
Databases (broadly) are ordered collections of data

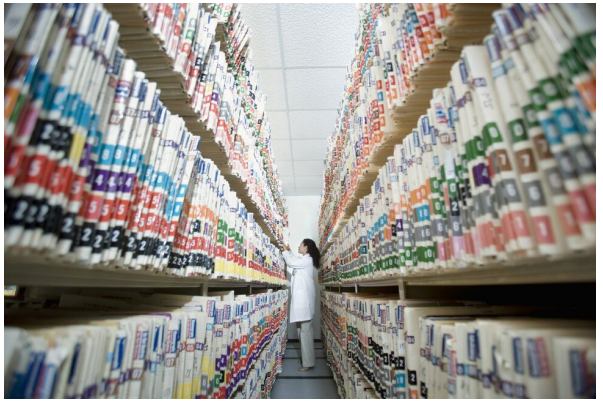
Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet

Ordering:

- Index
- Defined fields
- Standardisation

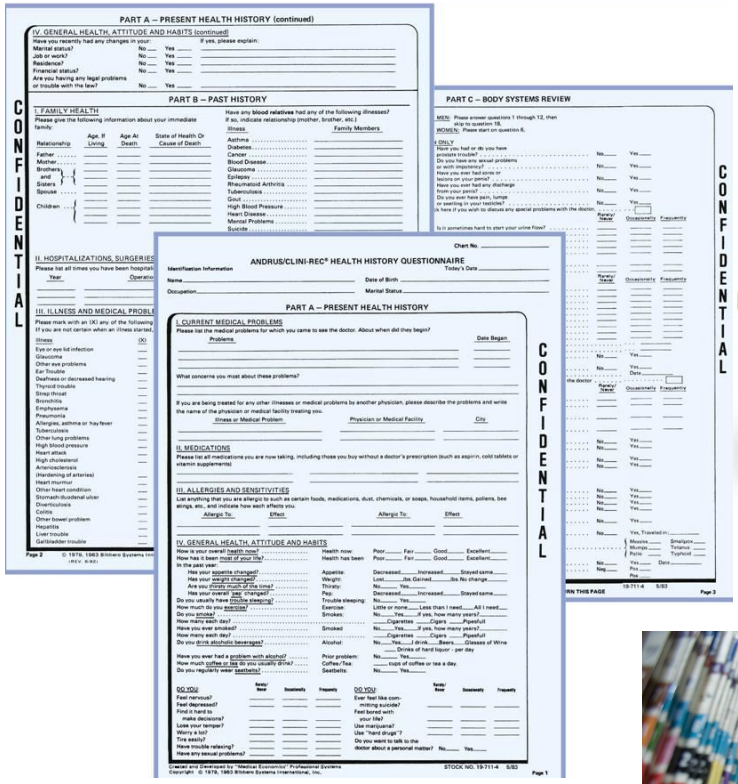




Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet

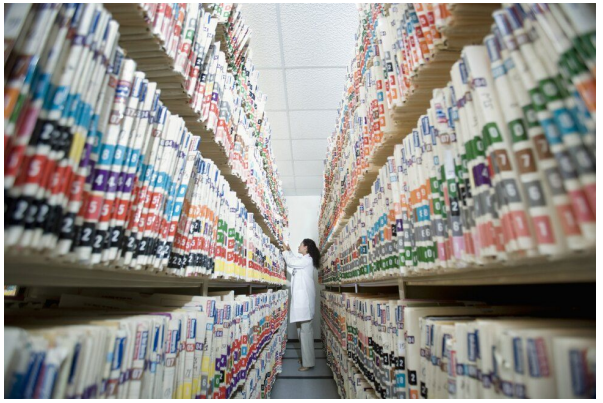


Ordering:

- Index
- Defined fields
- Standardisation

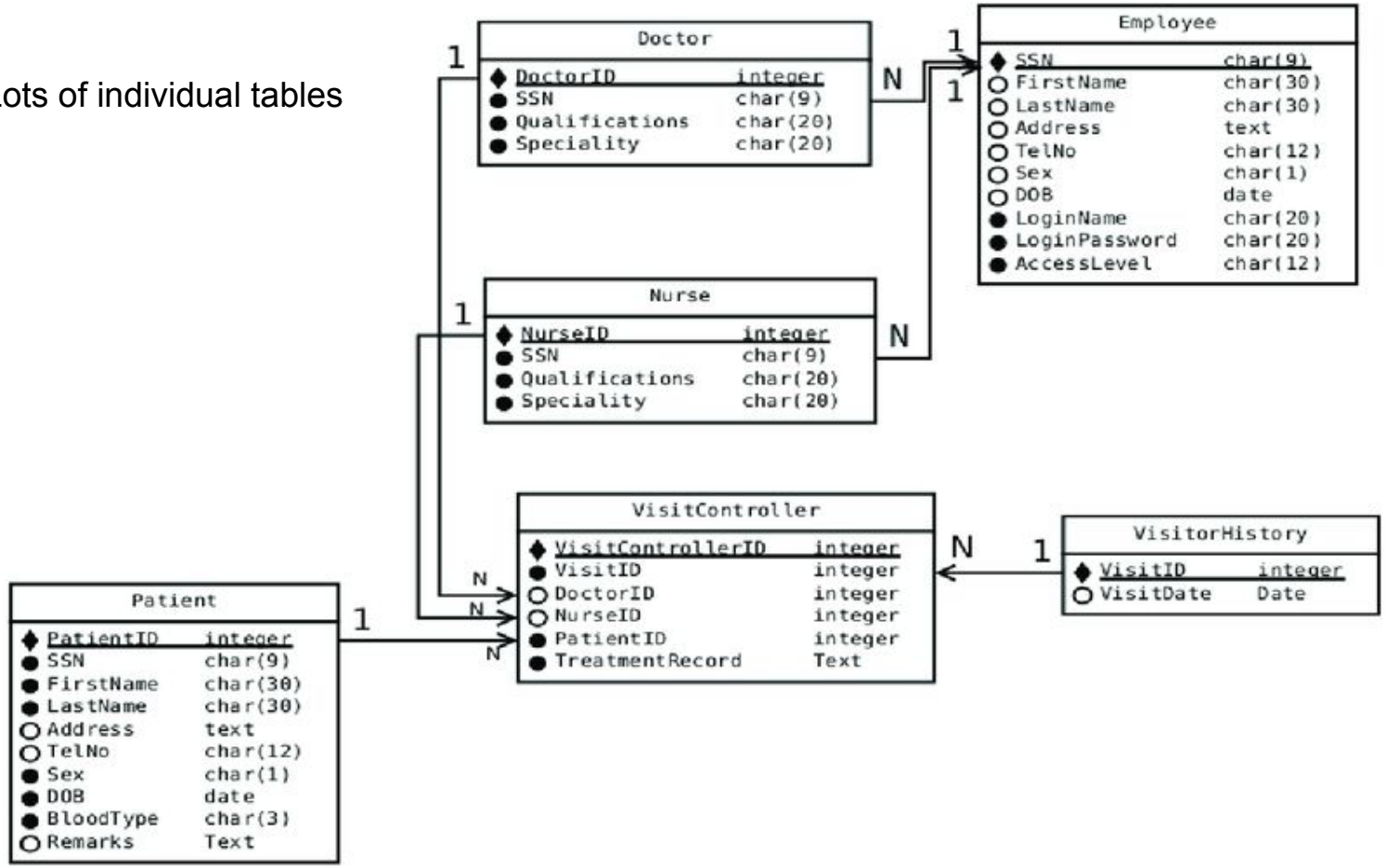
Organisation make some tasks easier/harder:

- Find all patients with the same condition
- Find the longest word in a dictionary
- Find an a number from an address in a phonebook



Most Common Type: Relational Databases

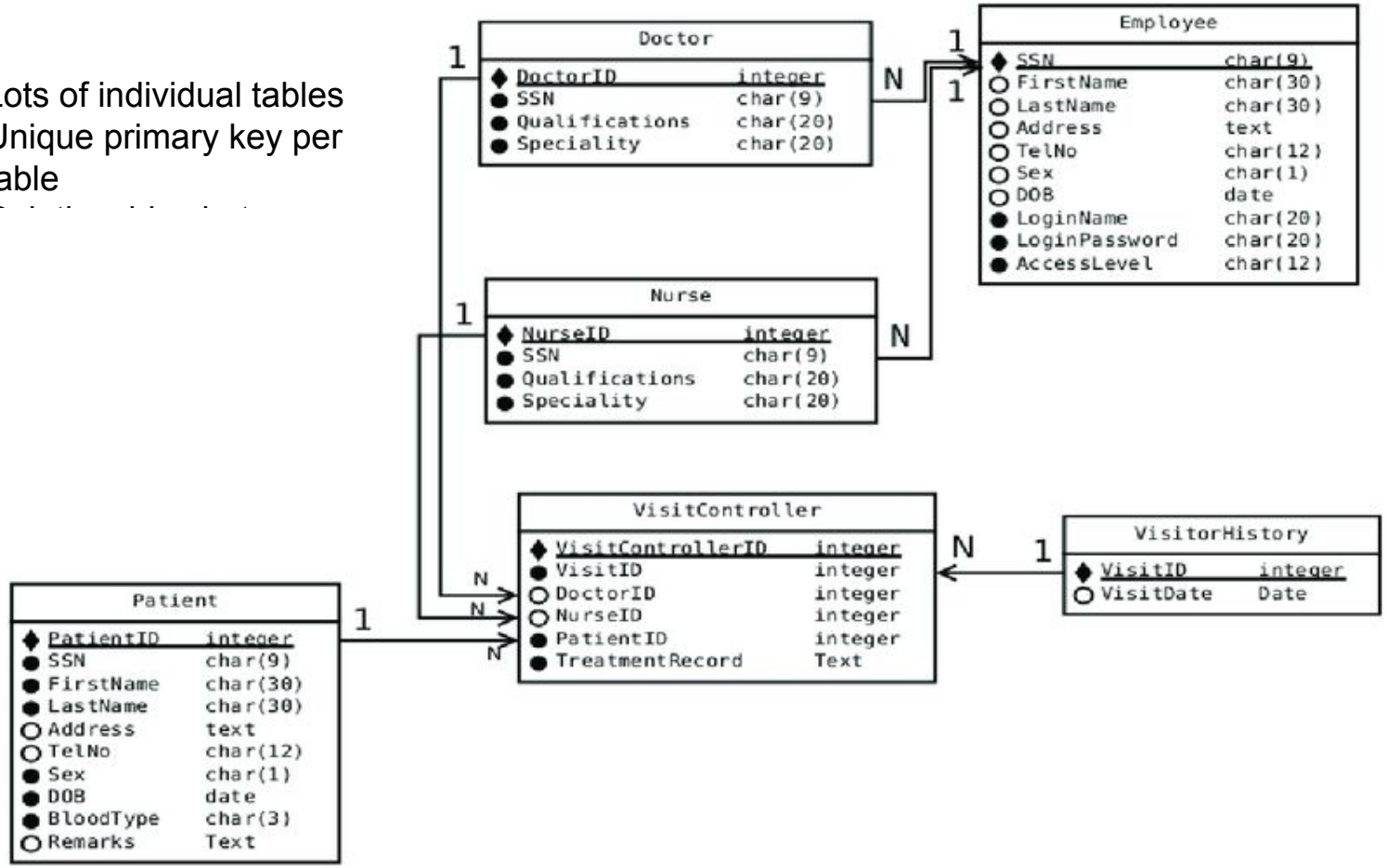
- Lots of individual tables



http://dx.doi.org/10.1016/j.procs.2015.08.441

Most Common Type: Relational Databases

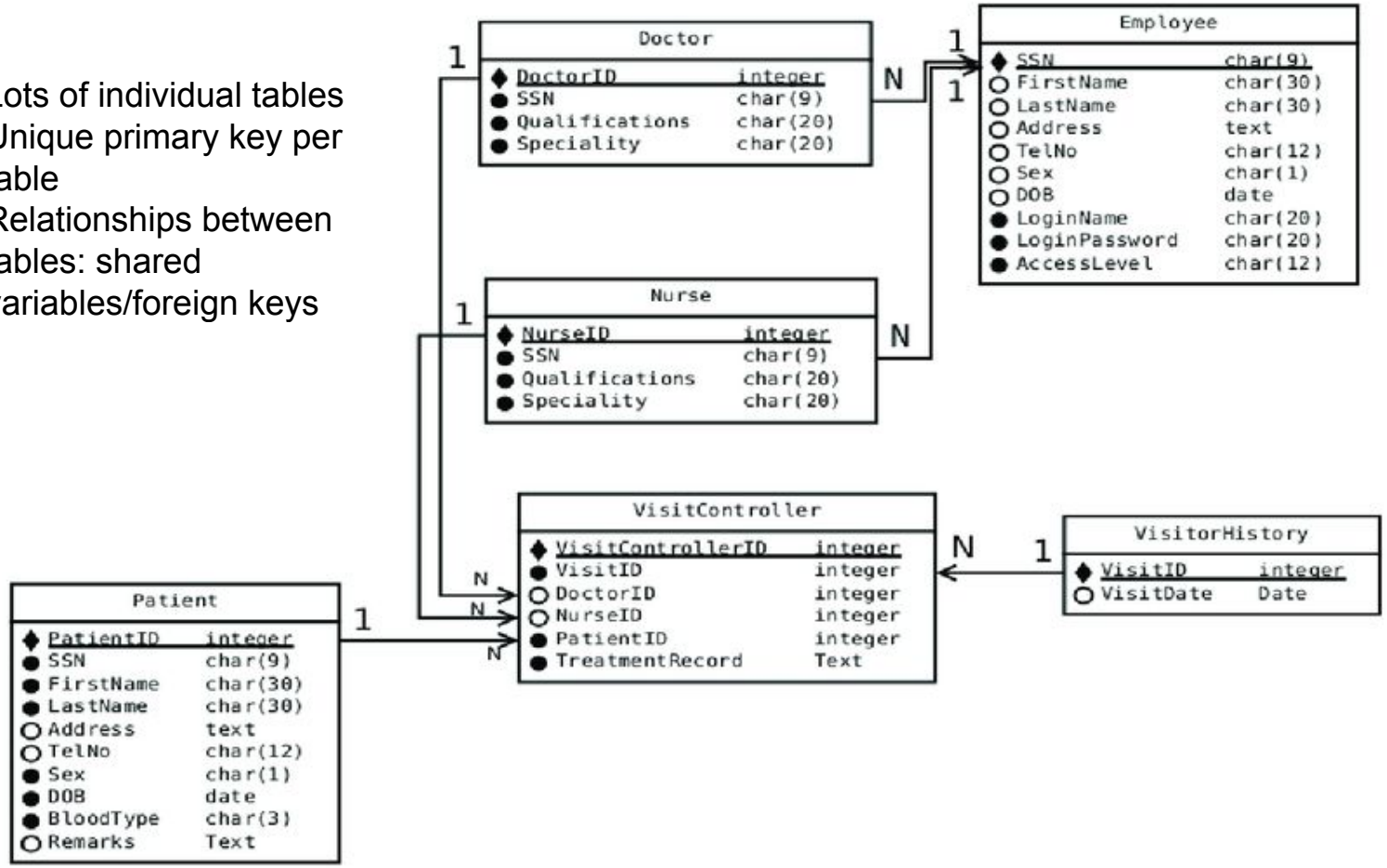
- Lots of individual tables
- Unique primary key per table



http://dx.doi.org/10.1016/j.procs.2015.08.441

Most Common Type: Relational Databases

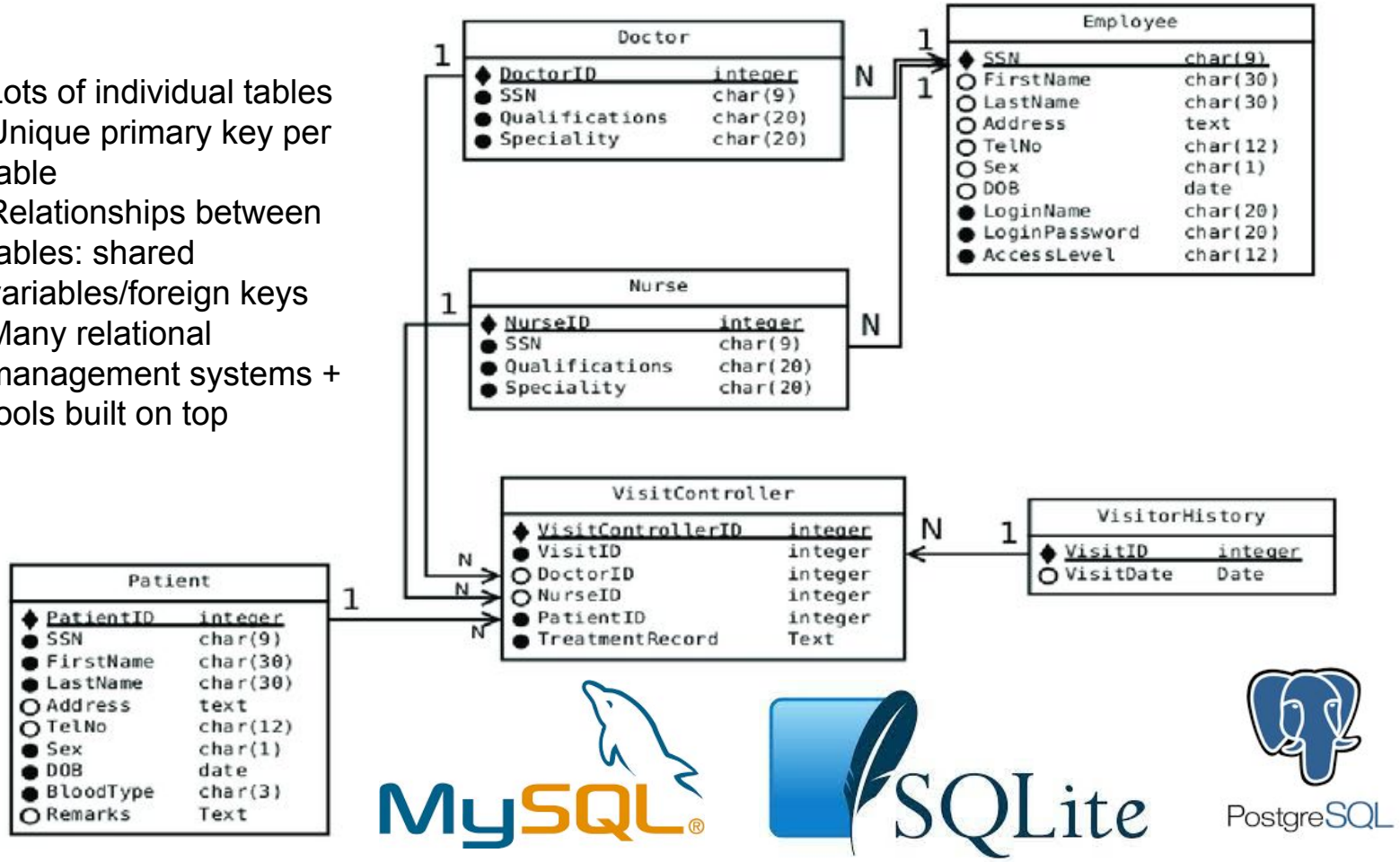
- Lots of individual tables
- Unique primary key per table
- Relationships between tables: shared variables/foreign keys



http://dx.doi.org/10.1016/j.procs.2015.08.441

Most Common Type: Relational Databases

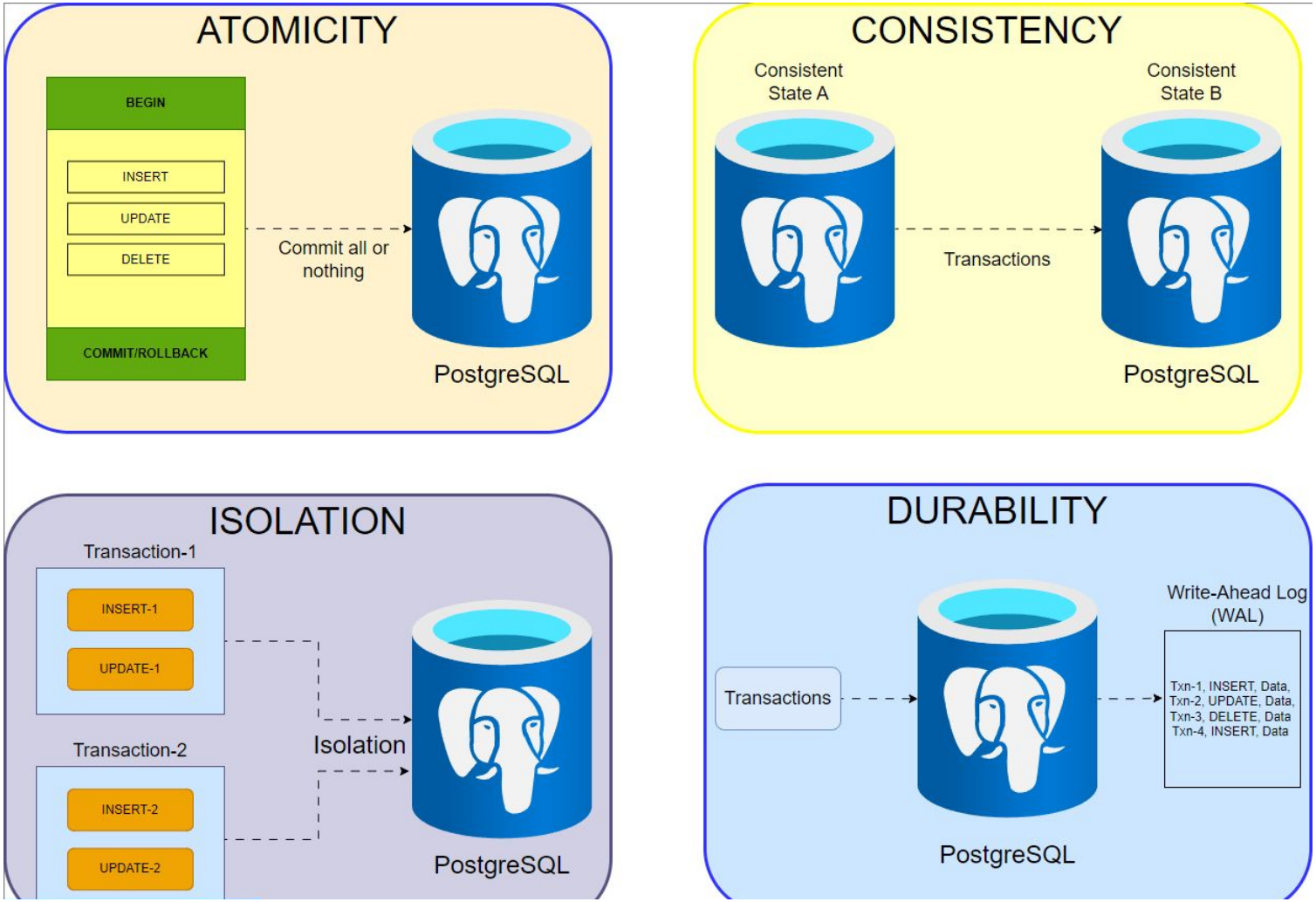
- Lots of individual tables
- Unique primary key per table
- Relationships between tables: shared variables/foreign keys
- Many relational management systems + tools built on top



http://dx.doi.org/10.1016/j.procs.2015.08.441

REDCap
Research Electronic Data Capture

Why bother with databases? ACID properties



Queried using Structured Query Language (SQL)

- Non-procedural Language
- Standardised/powerful/flexible

https://dbplyr.tidyverse.org/articles/sql-translation.html

Queried using Structured Query Language (SQL)

- Non-procedural Language
- Standardised/powerful/flexible
- Basis of many data tools
- Well-supported by dbplyr

https://dbplyr.tidyverse.org/articles/sql-translation.html

Queried using Structured Query Language (SQL)

- Non-procedural Language
- Standardised/powerful/flexible
- Basis of many data tools
- Well-supported by dbplyr

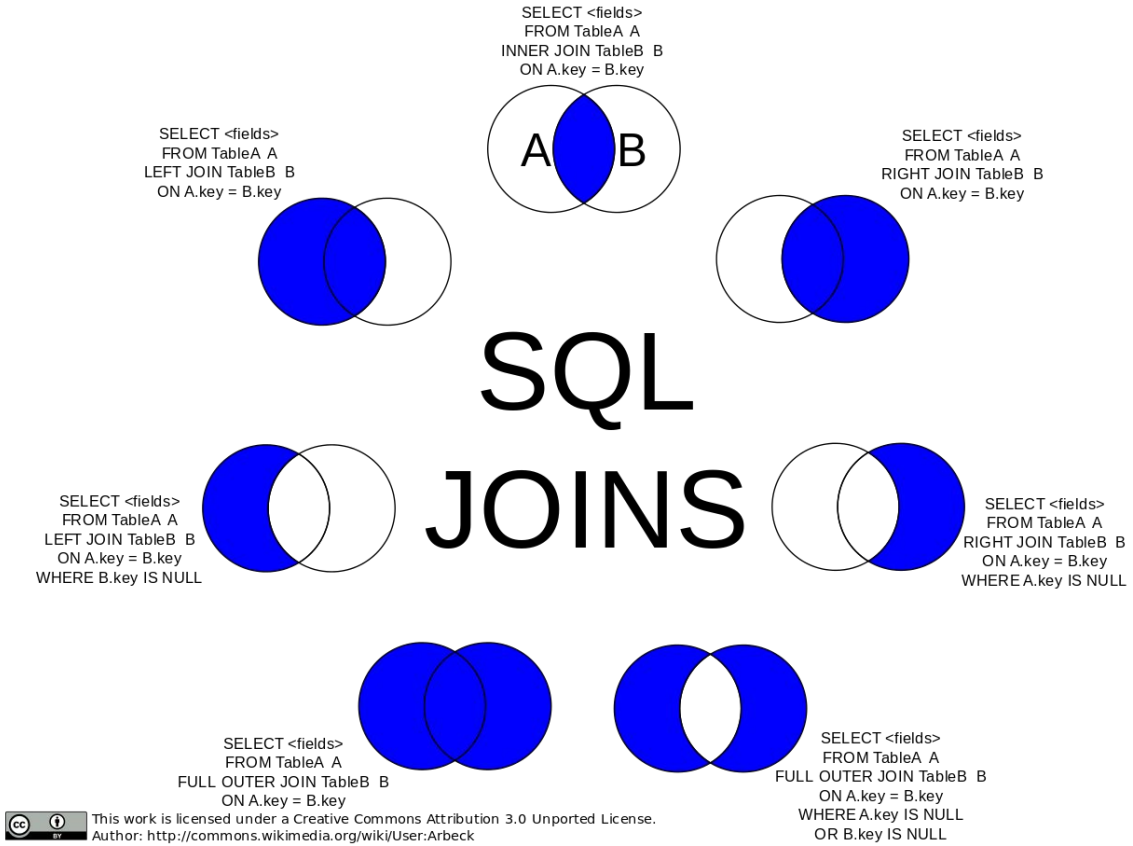
```
flights %>%
  select(contains("delay")) %>%
  show_query()
#> <SQL>
#> SELECT `dep_delay`, `arr_delay`
#> FROM `nycflights13::flights`
```

```
flights %>%
  group_by(month, day) %>%
  summarise(delay = mean(dep_delay)) %>%
  show_query()
#> Warning: Missing values are always removed in SQL.
#> Use `AVG(x, na.rm = TRUE)` to silence this warning
#> <SQL>
#> SELECT `month`, `day`, AVG(`dep_delay`) AS `delay`
#> FROM `nycflights13::flights`
#> GROUP BY `month`, `day`
```

```
flights %>%
  select(distance, air_time) %>%
  mutate(speed = distance / (air_time / 60)) %>%
  show_query()
#> <SQL>
#> SELECT `distance`, `air_time`, `distance` / (`air_time` / 60.0) AS `speed`
#> FROM (SELECT `distance`, `air_time`
#> FROM `nycflights13::flights`)
```

https://dbplyr.tidyverse.org/articles/sql-translation.html

SQL enables complex joins/queries



INNER JOIN - only assigned patients and theirs doctors

LEFT JOIN - all doctors including those without patients

RIGHT JOIN - all patients including those without doctors

OUTER JOIN - all patients and doctors

doctors

doctor_id	*name*	*specialty*
10	Dr. Smith	Cardiology
20	Dr. Jones	Neurology
30	Dr. Lee	Pediatrics
40	Dr. Patel	Oncology

patients

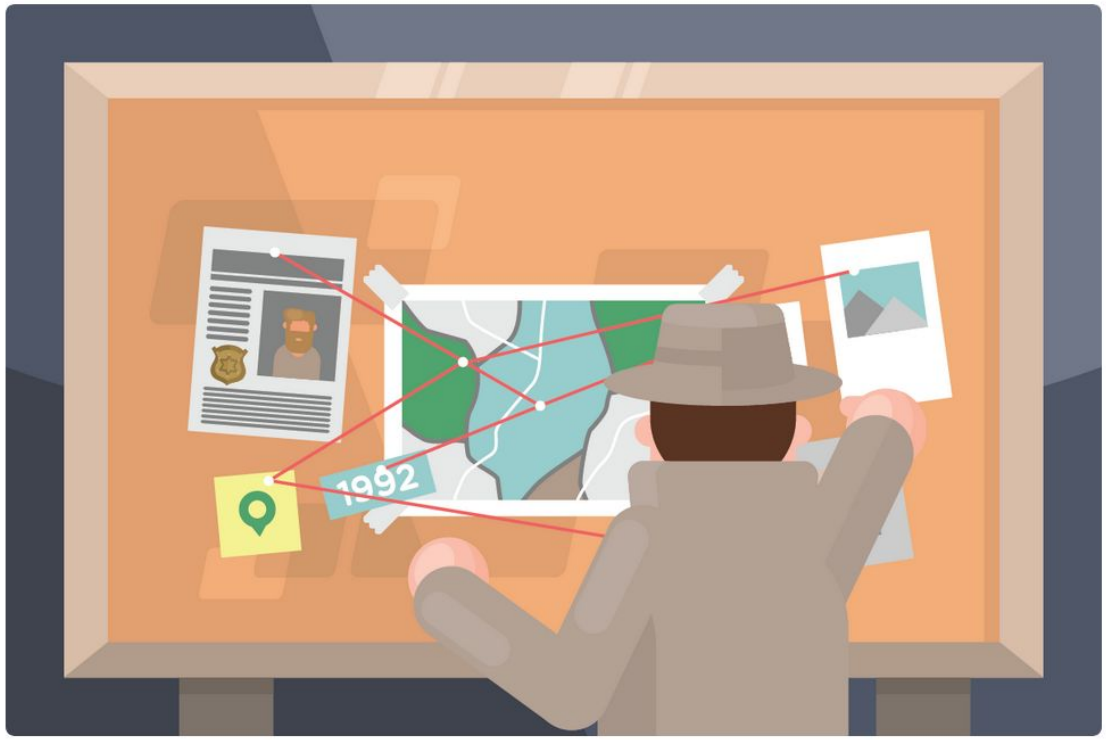
patient_id	*name*	*primary_doctor_id*
1	Alice	10
2	Bob	20
3	Carol	NULL
4	Dave	10

Fun way to learn basic SQL

https://mystery.knightlab.com/

SQL Murder Mystery

Can you find out whodunnit?

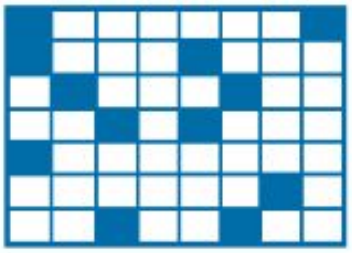


Are all databases relational?

Non-Relational Databases AKA NoSQL

- Less common than relational in medicine
- General focus on flexibility & performance
- Mostly for very large/unusual datasets or high demand:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies
- Or both:
  - Social media data

https://phoenixnap.com/kb/database-types



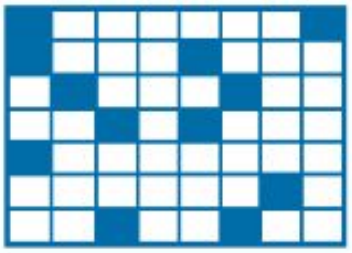
Column based

Google Big Query

Non-Relational Databases AKA NoSQL

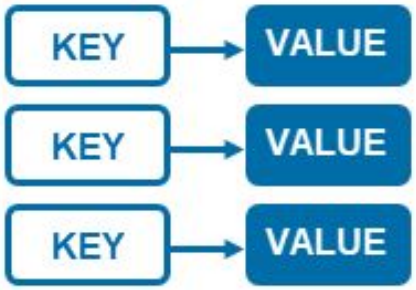
- Less common than relational in medicine
- General focus on flexibility & performance
- Mostly for very large/unusual datasets or high demand:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies
- Or both:
  - Social media data

https://phoenixnap.com/kb/database-types



Column based

Google Big Query



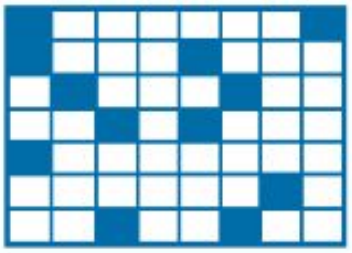
Key-value

redis

Non-Relational Databases AKA NoSQL

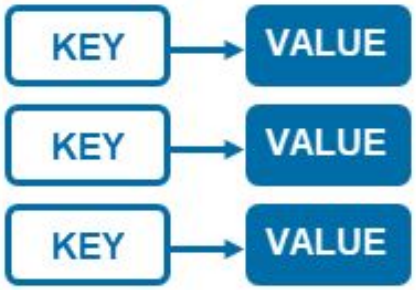
- Less common than relational in medicine
- General focus on flexibility & performance
- Mostly for very large/unusual datasets or high demand:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies
- Or both:
  - Social media data

https://phoenixnap.com/kb/database-types



Column based

Google Big Query



Key-value

redis

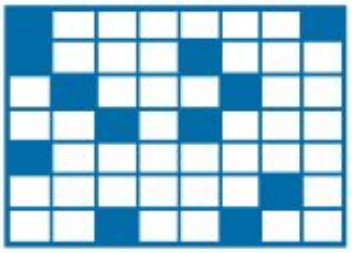
Graph

neo4j

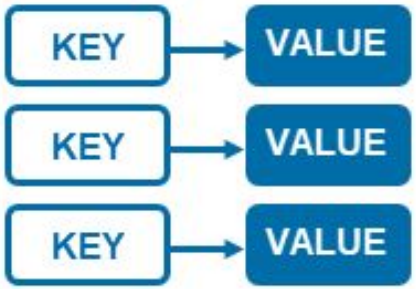
Non-Relational Databases AKA NoSQL

- Less common than relational in medicine
- General focus on flexibility & performance
- Mostly for very large/unusual datasets or high demand:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies
- Or both:
  - Social media data

https://phoenixnap.com/kb/database-types



Column based



Key-value

Graph

Document

Google Big Query

redis

neo4j

mongoDB.

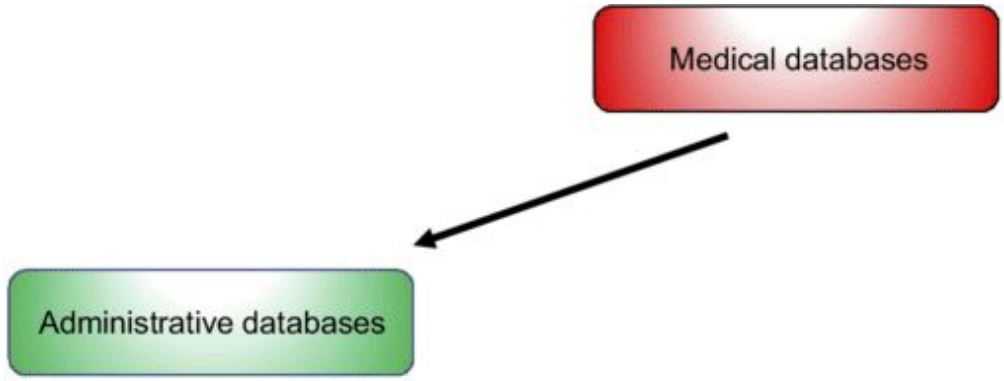
What are medical databases?

Many types of database

All types of registries and databases that contain health-related data

https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database

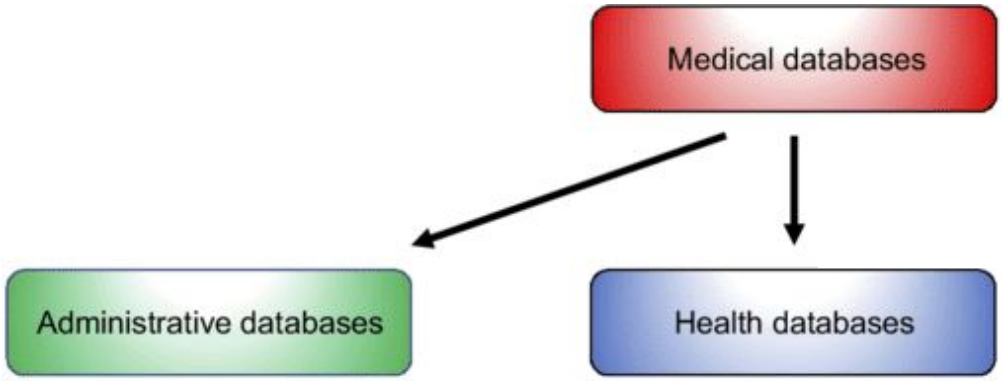


All types of registries and databases that contain health-related data

Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic

https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



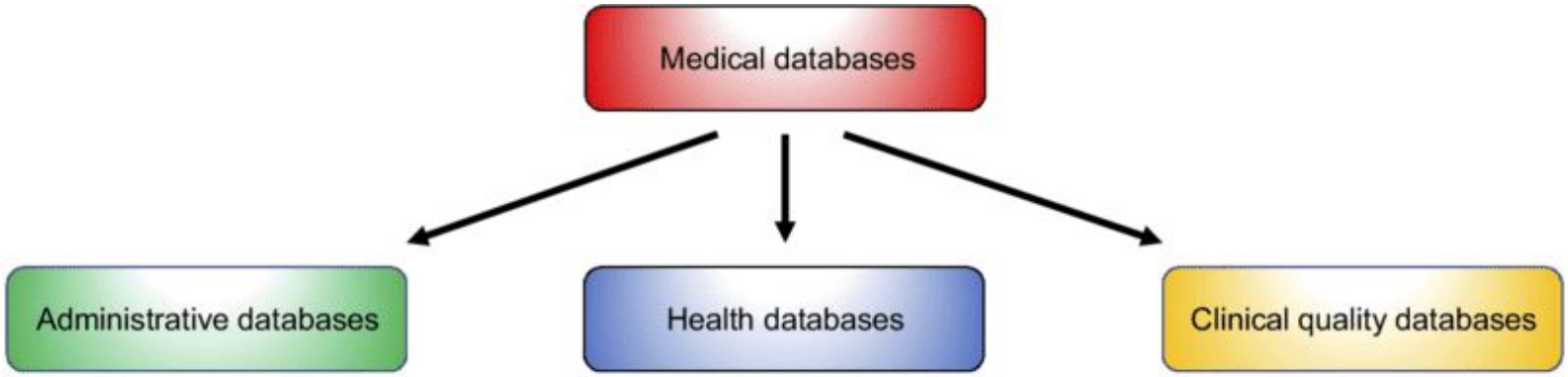
All types of registries and databases that contain health-related data

Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic

Register health data for the purpose of surveillance and research

https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

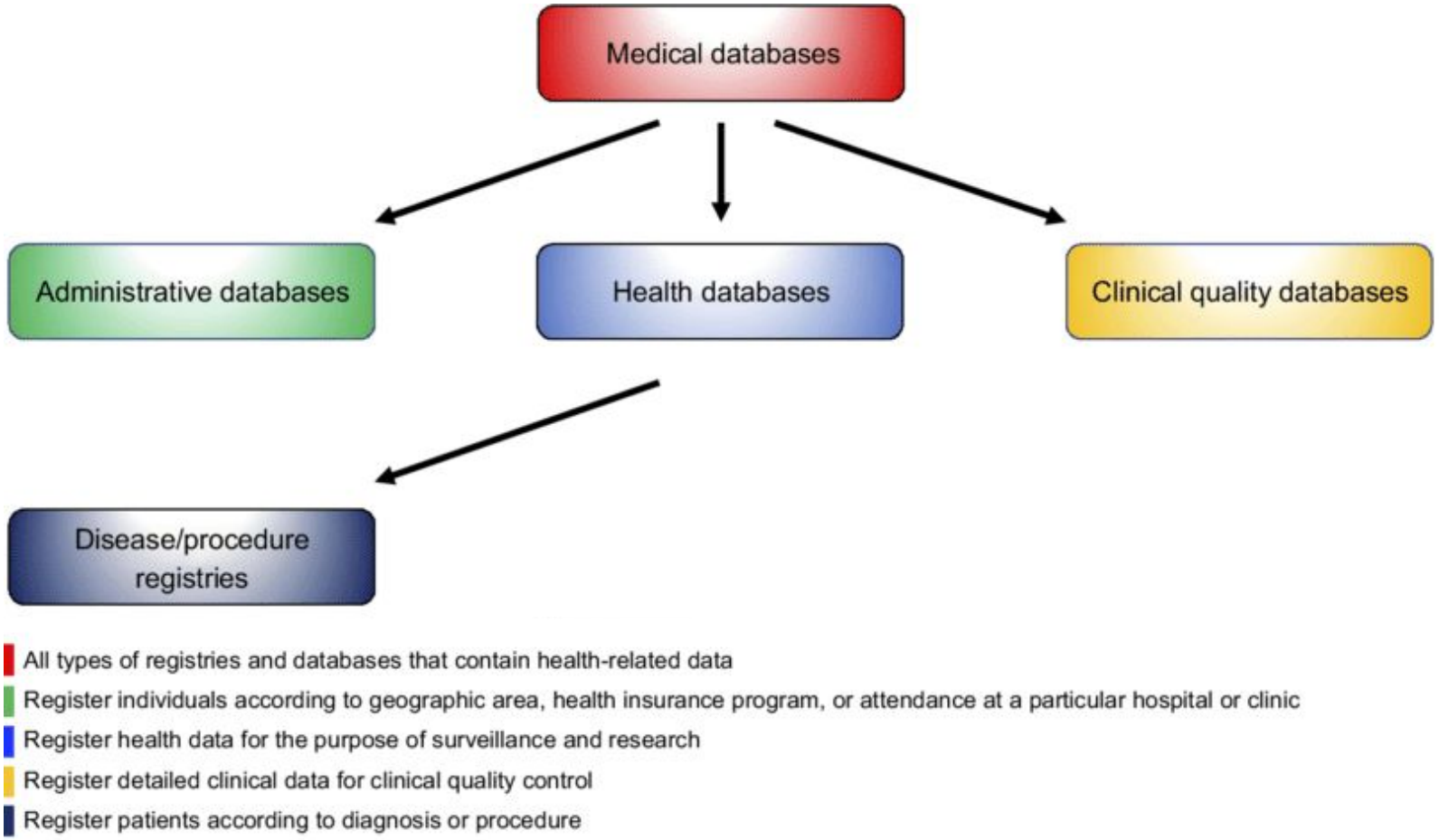
Many types of database



All types of registries and databases that contain health-related data
Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic
Register health data for the purpose of surveillance and research
Register detailed clinical data for clinical quality control

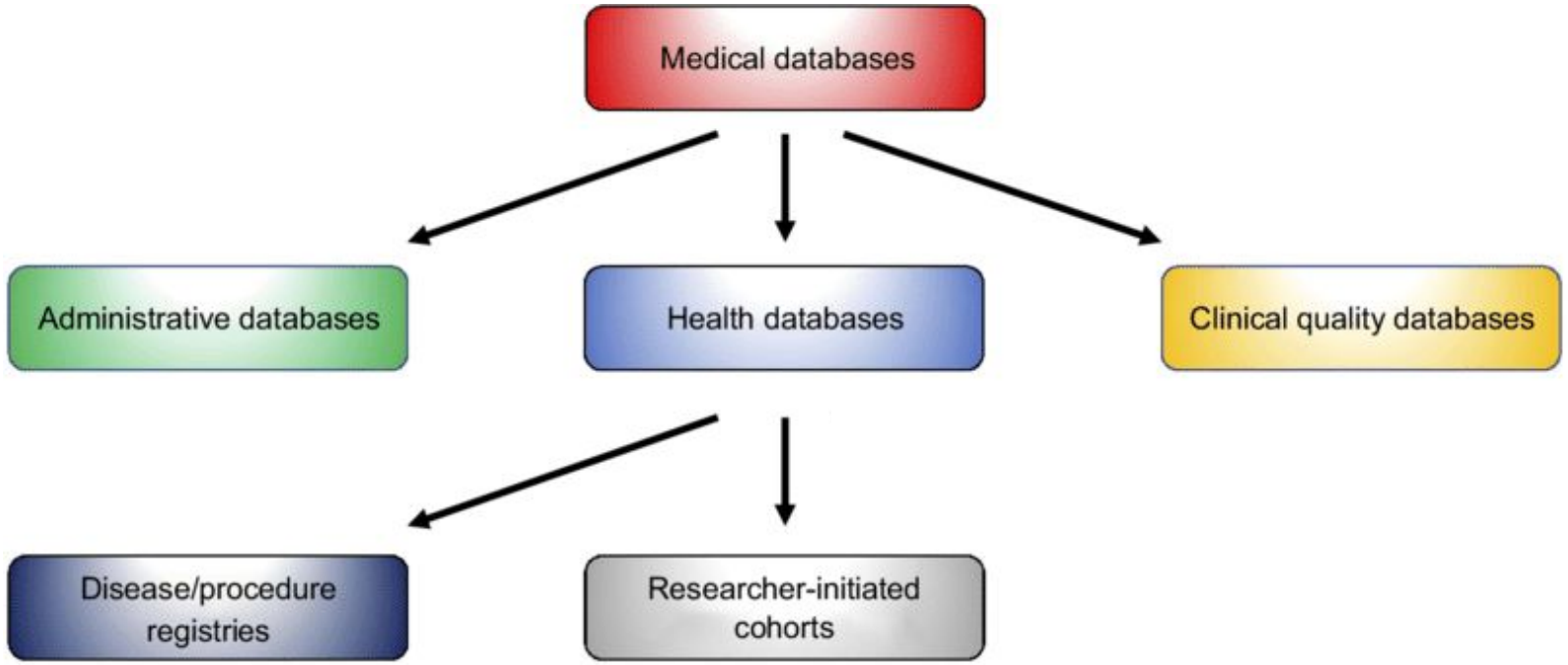
https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

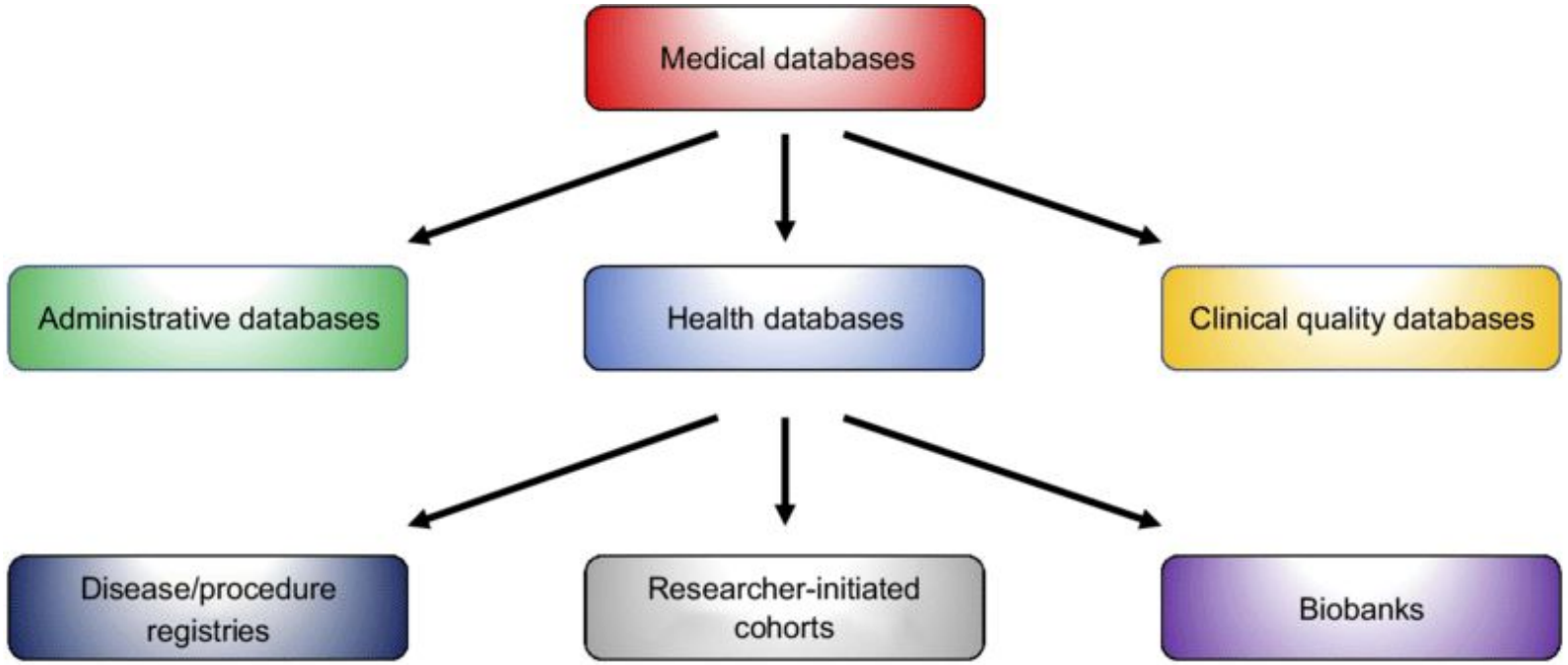
Many types of database



All types of registries and databases that contain health-related data
Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic
Register health data for the purpose of surveillance and research
Register detailed clinical data for clinical quality control
Register patients according to diagnosis or procedure
Register individuals according to prespecified criteria (eg, area of residency, age, sex, conscription, adoption, pregnancy, or survey participation)

https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



All types of registries and databases that contain health-related data

Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic

Register health data for the purpose of surveillance and research

Register detailed clinical data for clinical quality control

Register patients according to diagnosis or procedure

Register individuals according to prespecified criteria (eg, area of residency, age, sex, conscription, adoption, pregnancy, or survey participation)

Store biological samples (eg, blood and tissue)

https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Analysis based on primary record type & sampling scope

Primary Record Type:

- Individual procedures e.g., arthroplasty
- Prescriptions e.g., colistin
- Disease/Illness e.g., ovarian cancer
- Hospital Admission/Discharge
- Individual health interactions
- Patient
- Person
- Population

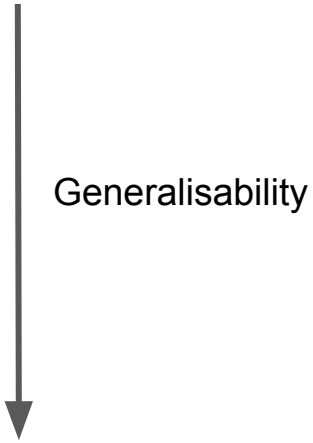
Analysis based on primary record type & sampling scope

Primary Record Type:

- Individual procedures e.g., arthroplasty
- Prescriptions e.g., colistin
- Disease/Illness e.g., ovarian cancer
- Hospital Admission/Discharge
- Individual health interactions
- Patient
- Person
- Population

Sampling Scope:

- Single physician
- Group of physicians
- Hospital
- Health Authority
- Province
- National
- International



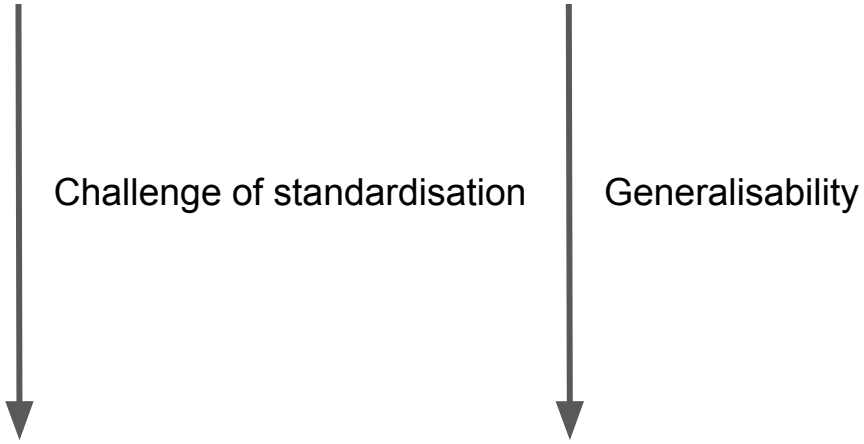
Analysis based on primary record type & sampling scope

Primary Record Type:

- Individual procedures e.g., arthroplasty
- Prescriptions e.g., colistin
- Disease/Illness e.g., ovarian cancer
- Hospital Admission/Discharge
- Individual health interactions
- Patient
- Person
- Population

Sampling Scope:

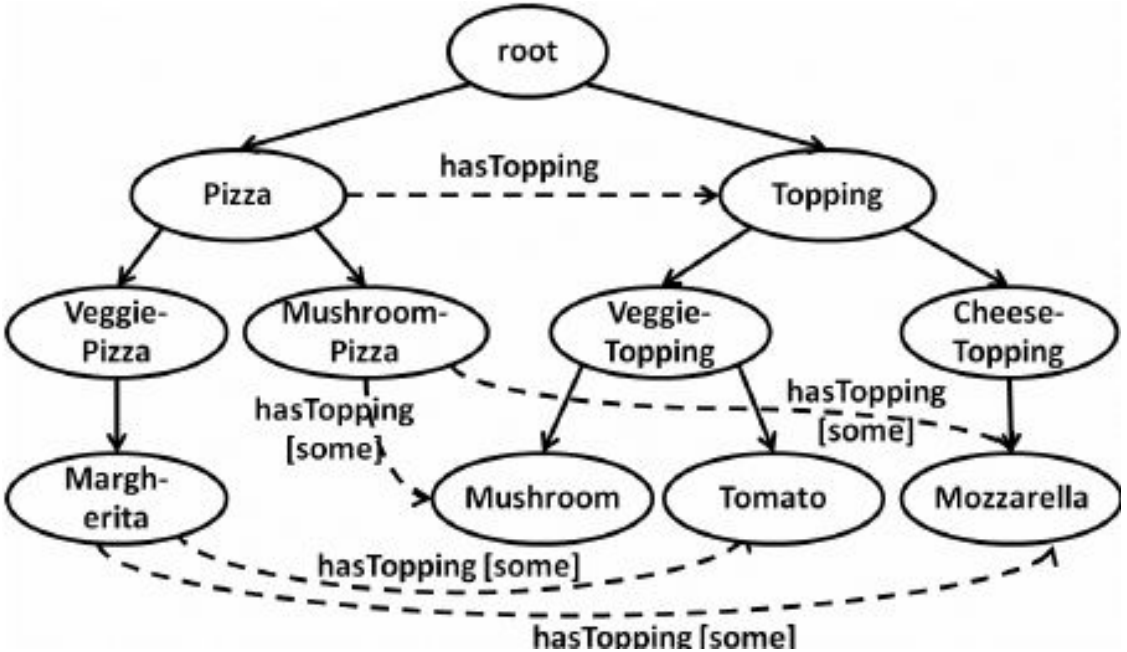
- Single physician
- Group of physicians
- Hospital
- Health Authority
- Province
- National
- International



How do medical databases try to handle standardisation?

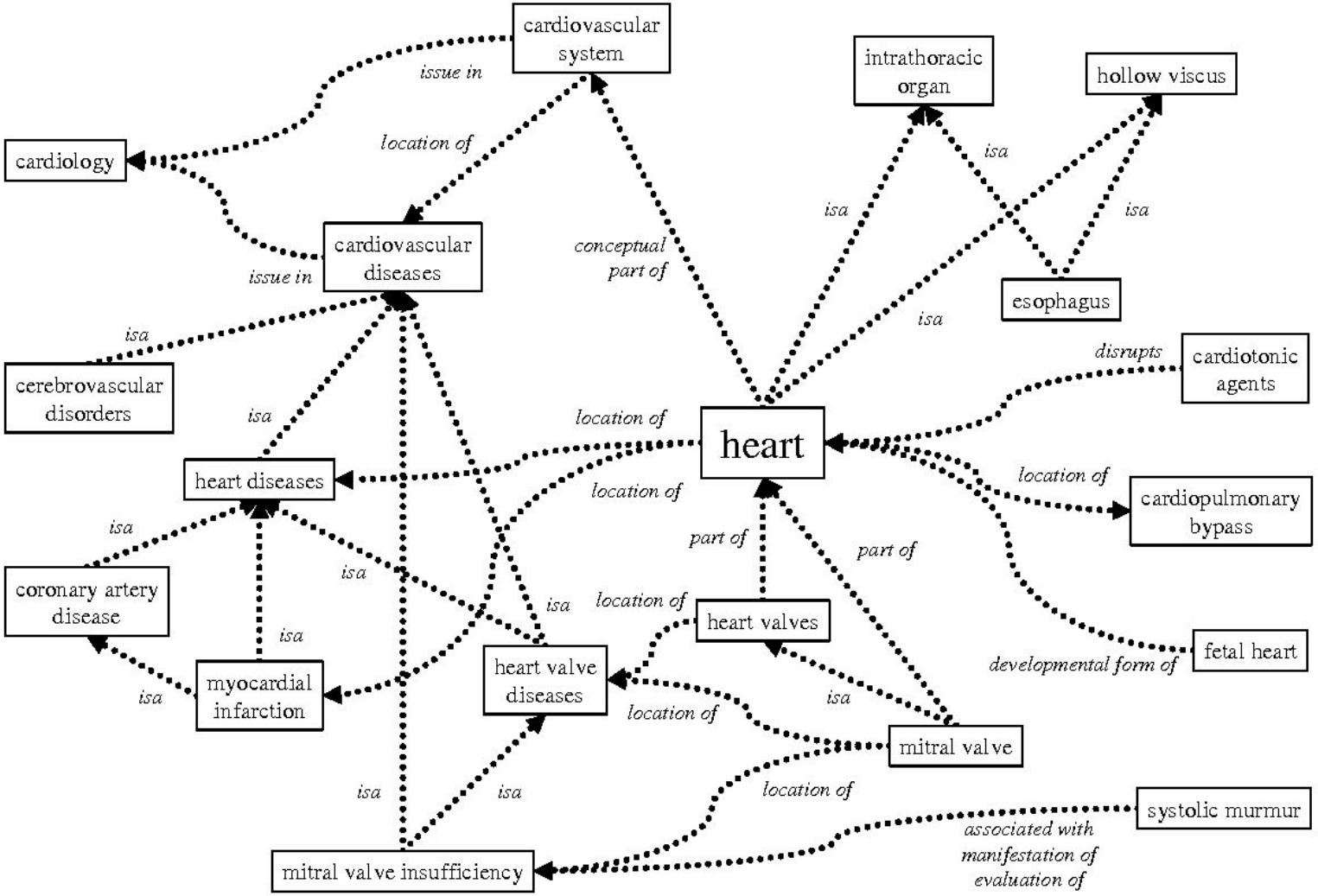
Ontologies for standardisation

- Standardised terms e.g., Pizza, Tomato, Mozzarella
- Standardised types of relationships between terms
- Acyclic links between terms
- Manual curation
- Automated curation



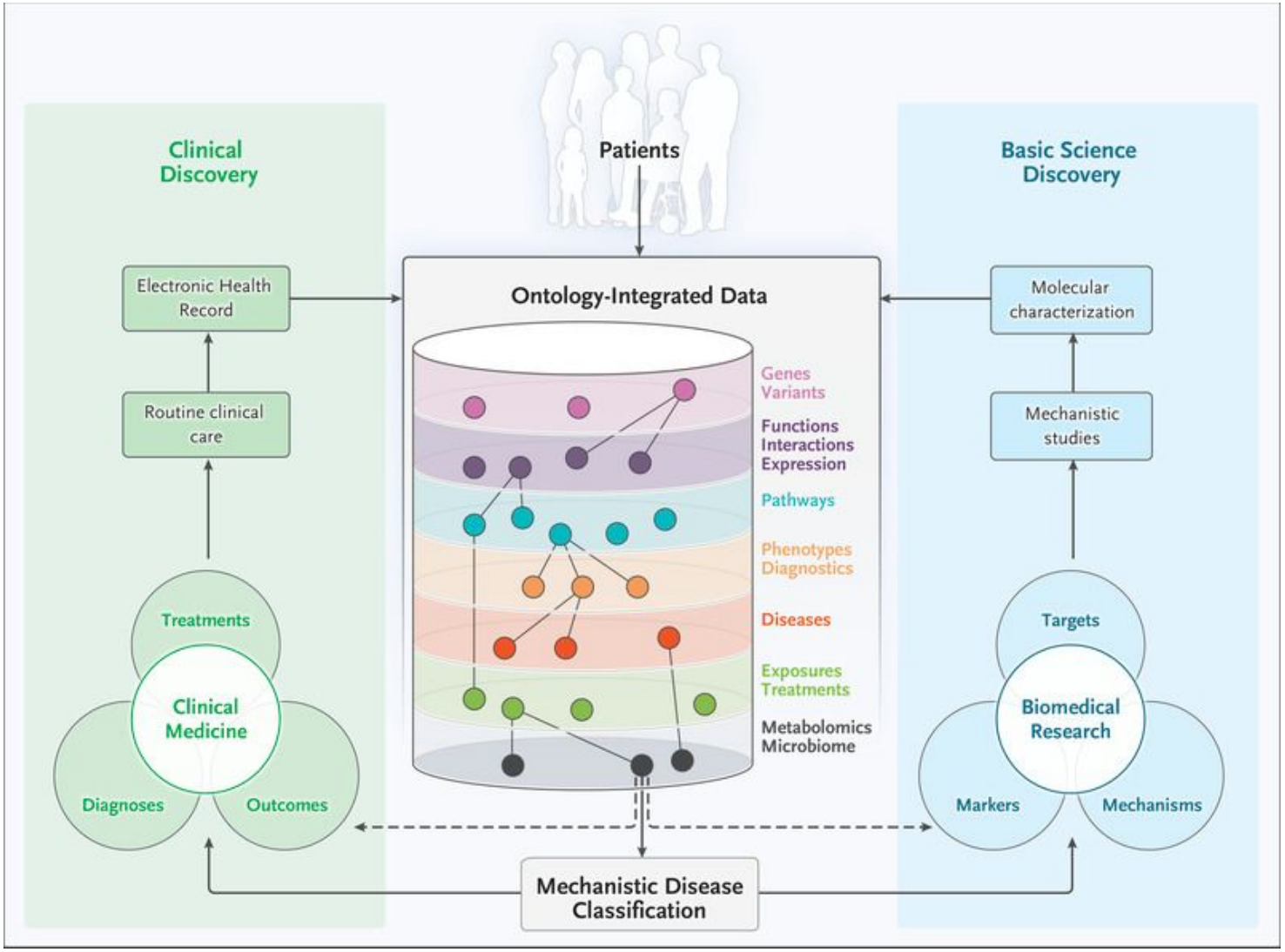
https://www.researchgate.net/figure/Example-pizza-ontology-represented-as-a-graph-G-a-and-a-changed-version-of-the-pizza_fig1_236842047

Medical Ontologies



https://www.semanticscholar.org/paper/Medical-Ontology-Research-Bodenreider/b27ef1466b5d0ca6b4b17d8d9e02406b996e0ba2

Ontologies for linking diverse types of data



https://www.nejm.org/doi/full/10.1056/NEJMra1615014

International Statistical Classification of Diseases and Related Health Problems (ICD-9, ICD-10)

- 2 ontologies
  - ICD-X-CM (medical diagnoses)
  - ICD-X-PCS (procedure coding)

International Statistical Classification of Diseases and Related Health Problems (ICD-9, ICD-10)

- 2 ontologies
  - ICD-X-CM (medical diagnoses)
  - ICD-X-PCS (procedure coding)
- ICD-9 -> ICD-10 (2015)

Differences Between ICD-9-CM and ICD-10 Code Sets		
	ICD-9-CM	**ICD-10 code sets**
Procedure	3,824 codes	71,924 codes
Diagnosis	14,025 codes	69, 823 codes

ICD-10 Code Structure Changes (selected details)		
	Old	**New**
Diagnosis Structure	ICD-9-CM • 3 -5 characters • First character is numeric or alpha • Characters 2-5 are numeric	ICD-10-CM • 3 -7 characters • Character 1 is alpha • Character 2 is numeric • Characters 3 – 7 can be alpha or numeric
Procedure Structure	ICD-9-CM • 3-4 characters • All characters are numeric • All codes have at least 3 characters	ICD-10-PCS • ICD-10-PCS has 7 characters • Each can be either alpha or numeric • Numbers 0-9; letters A-H, J-N, P-Z

https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm

International Statistical Classification of Diseases and Related Health Problems (ICD-9, ICD-10)

- 2 ontologies
  - ICD-X-CM (medical diagnoses)
  - ICD-X-PCS (procedure coding)
- ICD-9 -> ICD-10 (2015)

- “V97.33XD: Sucked into jet engine, subsequent encounter.”
- “Y93.D: V91.07XD: Burn due to water-skis on fire, subsequent encounter.”
- “Z63.1: Problems in relationship with in-laws.”
- “W22.02XD: V95.43XS: Spacecraft collision injuring occupant, sequela.”

Differences Between ICD-9-CM and ICD-10 Code Sets		
	ICD-9-CM	ICD-10 code sets
Procedure	3,824 codes	71,924 codes
Diagnosis	14,025 codes	69, 823 codes

ICD-10 Code Structure Changes (selected details)		
	Old	New
Diagnosis Structure	ICD-9-CM • 3 -5 characters • First character is numeric or alpha • Characters 2-5 are numeric	ICD-10-CM • 3 -7 characters • Character 1 is alpha • Character 2 is numeric • Characters 3 – 7 can be alpha or numeric
Procedure Structure	ICD-9-CM • 3-4 characters • All characters are numeric • All codes have at least 3 characters	ICD-10-PCS • ICD-10-PCS has 7 characters • Each can be either alpha or numeric • Numbers 0-9; letters A-H, J-N, P-Z

https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm

How do we sample from medical databases?

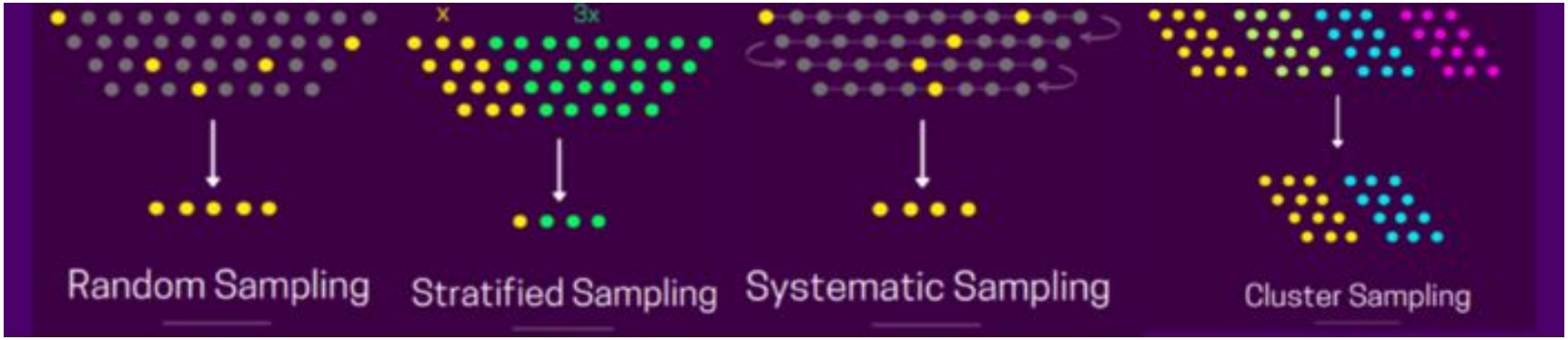
Sampling strategy

- Exhaustive in a database isn’t usually exhaustive in true population
- Numerous and often quite complex!
- Major source of bias so always carefully explore

https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124

Sampling strategy

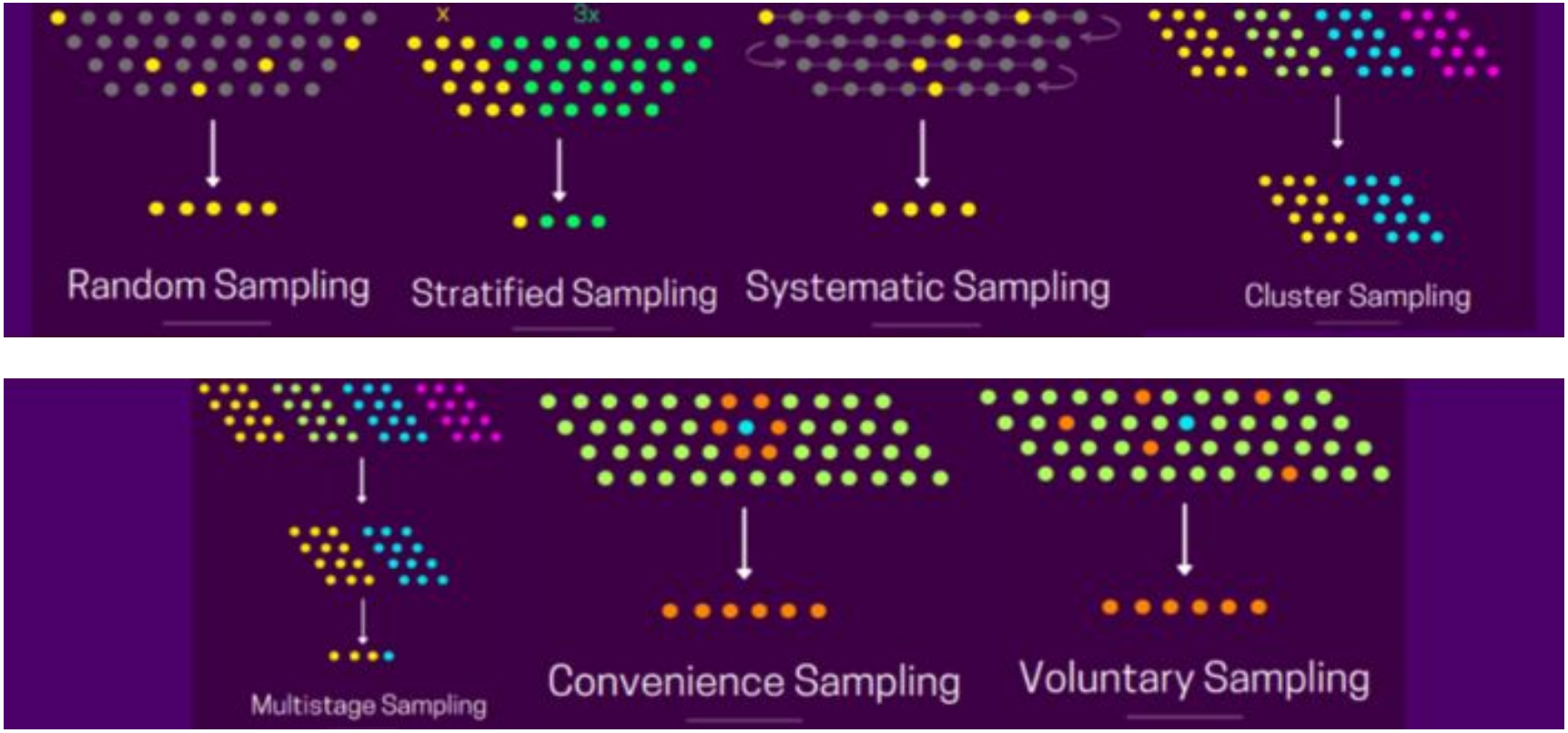
- Exhaustive in a database isn't usually exhaustive in true population
- Numerous and often quite complex!
- Major source of bias so always carefully explore



https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124

Sampling strategy

- Exhaustive in a database isn't usually exhaustive in true population
- Numerous and often quite complex!
- Major source of bias so always carefully explore



https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124

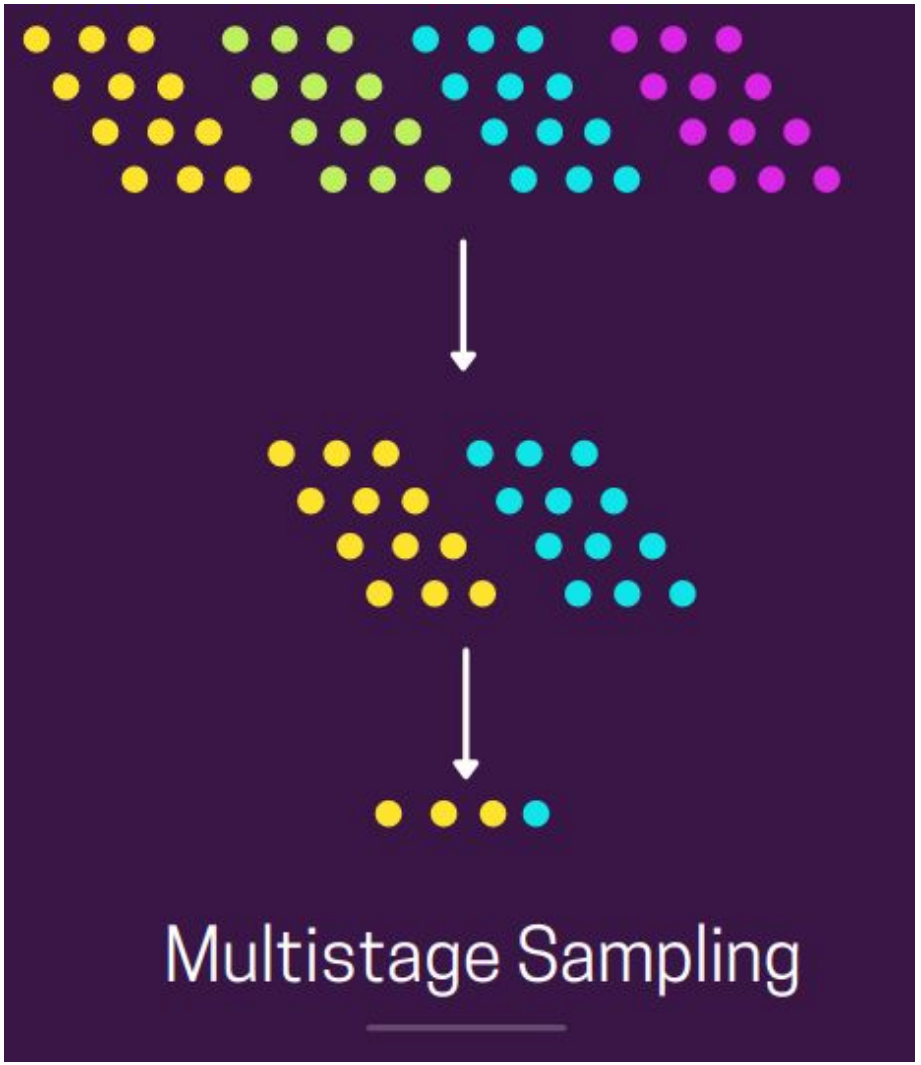
Survey/Sample weights

- Value/weight assigned to each record
- Make statistics calculated from database more representative of population
  - Weight=0.5 underweight this case
  - Weight=1
  - Weight=2 overweight the contribution of this case

Survey/Sample weights

- Value/weight assigned to each record
- Make statistics calculated from database more representative of population
  - Weight=0.5 underweight this case
  - Weight=1
  - Weight=2 overweight the contribution of this case

- Complex sampling strategies (e.g., deliberate oversampling of some populations, biasing recruitment) mean weights **MUST** be used.
- Not directly supported in all machine learning libraries (`sample_weights` implemented for some models)



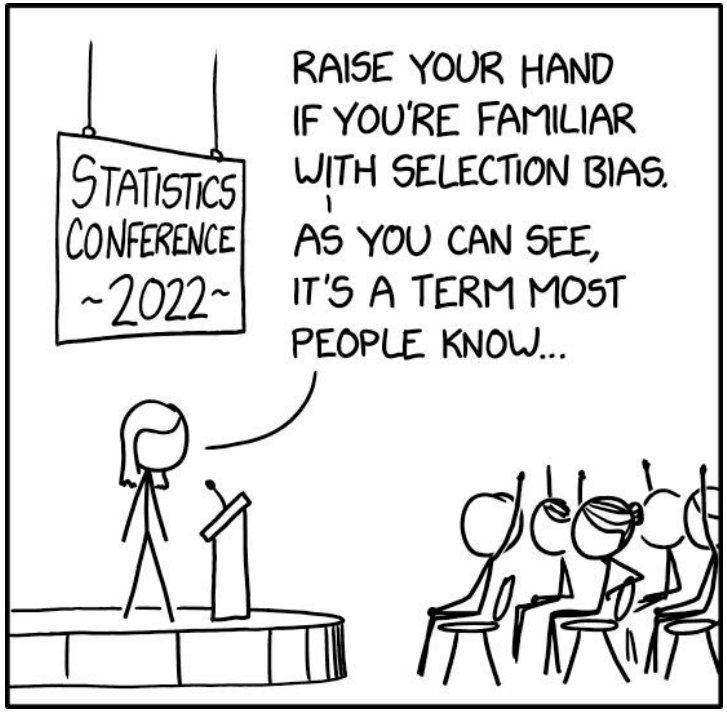
Types of weights

- Design Weights
  - Based on sampling strategy i.e., “design” of survey/database/data collection
  - Common to over-sample under-represented or rare groups
  - Need to correct for this or will overestimate statistics e.g., lower weight of over-sampled groups

Types of weights

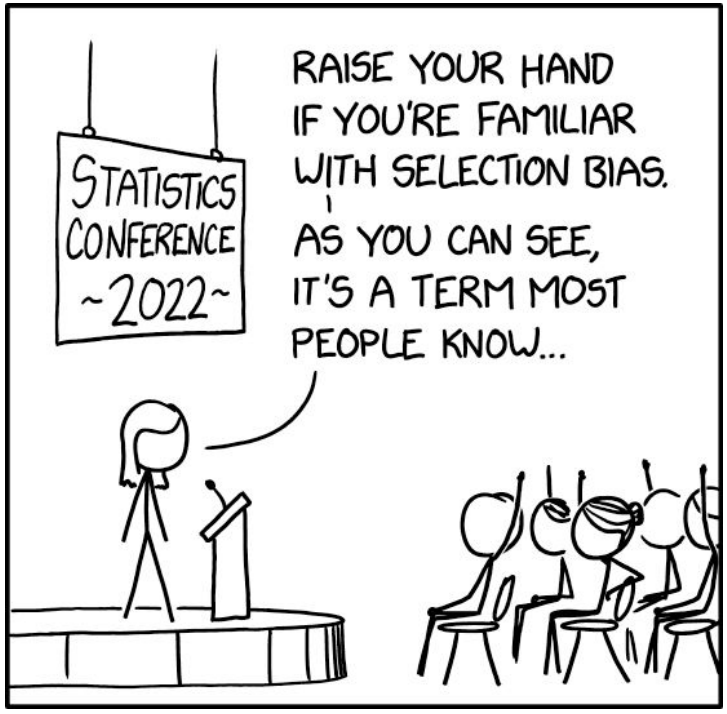
- Design Weights
  - Based on sampling strategy i.e., “design” of survey/database/data collection
  - Common to over-sample under-represented or rare groups
  - Need to correct for this or will overestimate statistics e.g., lower weight of over-sampled groups

- Post-stratification / Non-response weights
  - Based on collected data
  - Typically biases in whose data is collected
  - Over-represented groups need to be under-weighted



Types of weights

- Design Weights
  - Based on sampling strategy i.e., “design” of survey/database/data collection
  - Common to over-sample under-represented or rare groups
  - Need to correct for this or will overestimate statistics e.g., lower weight of over-sampled groups

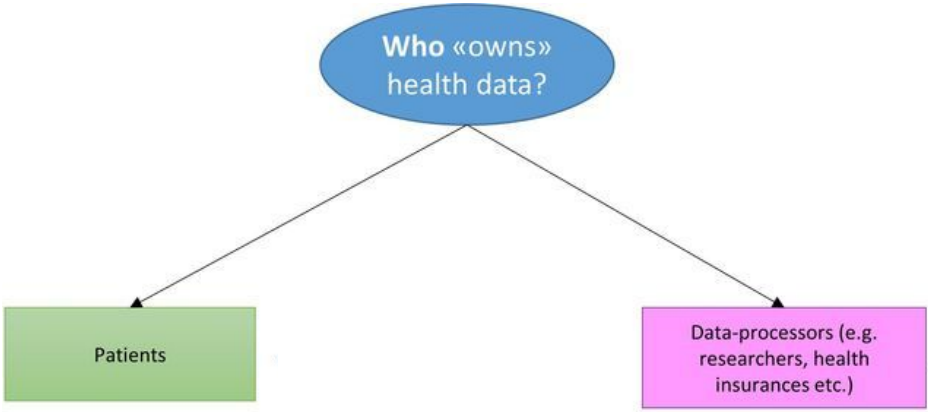


- Post-stratification / Non-response weights
  - Based on collected data
  - Typically biases in whose data is collected
  - Over-represented groups need to be under-weighted

- Often many different weights are combined

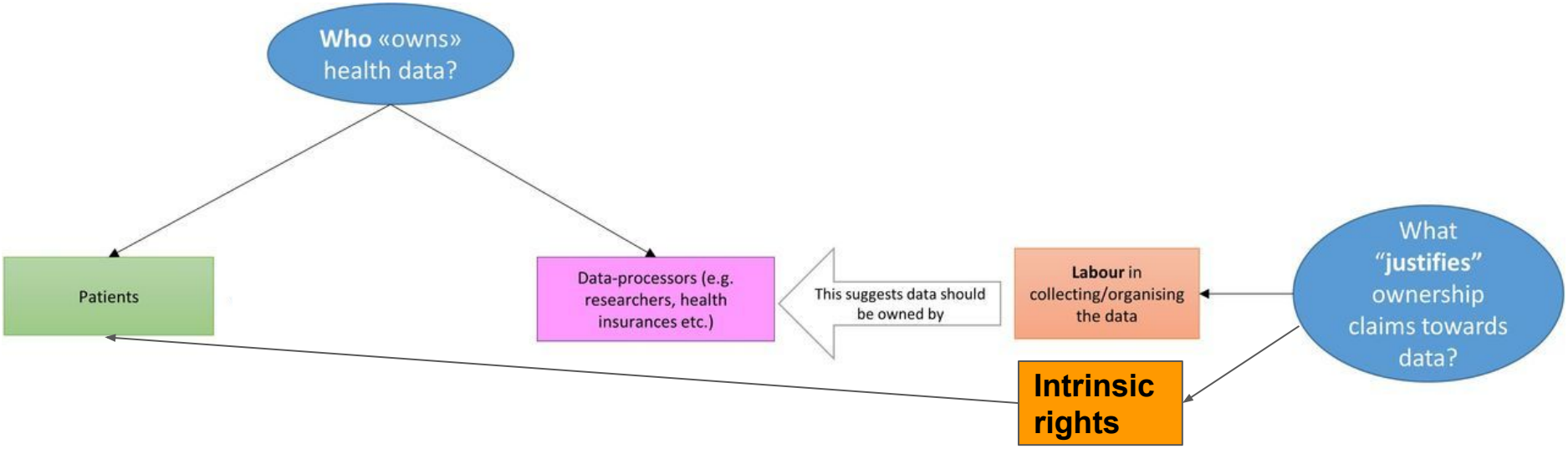
Who actually owns this data?

Data ownership is complicated



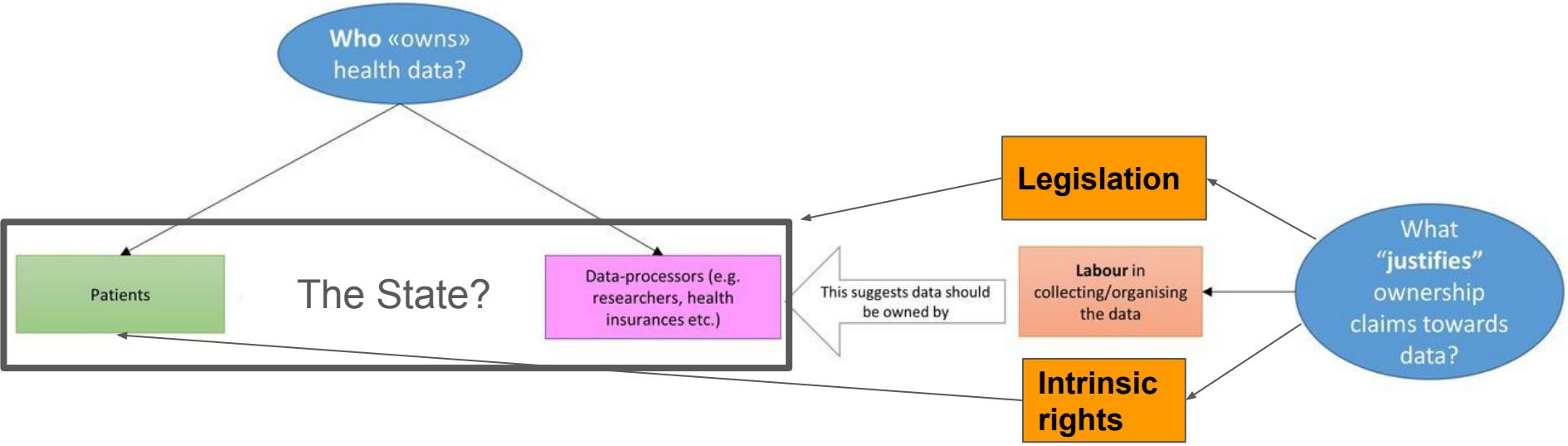
https://bmjopen.bmj.com/content/11/4/e045717

Data ownership is complicated



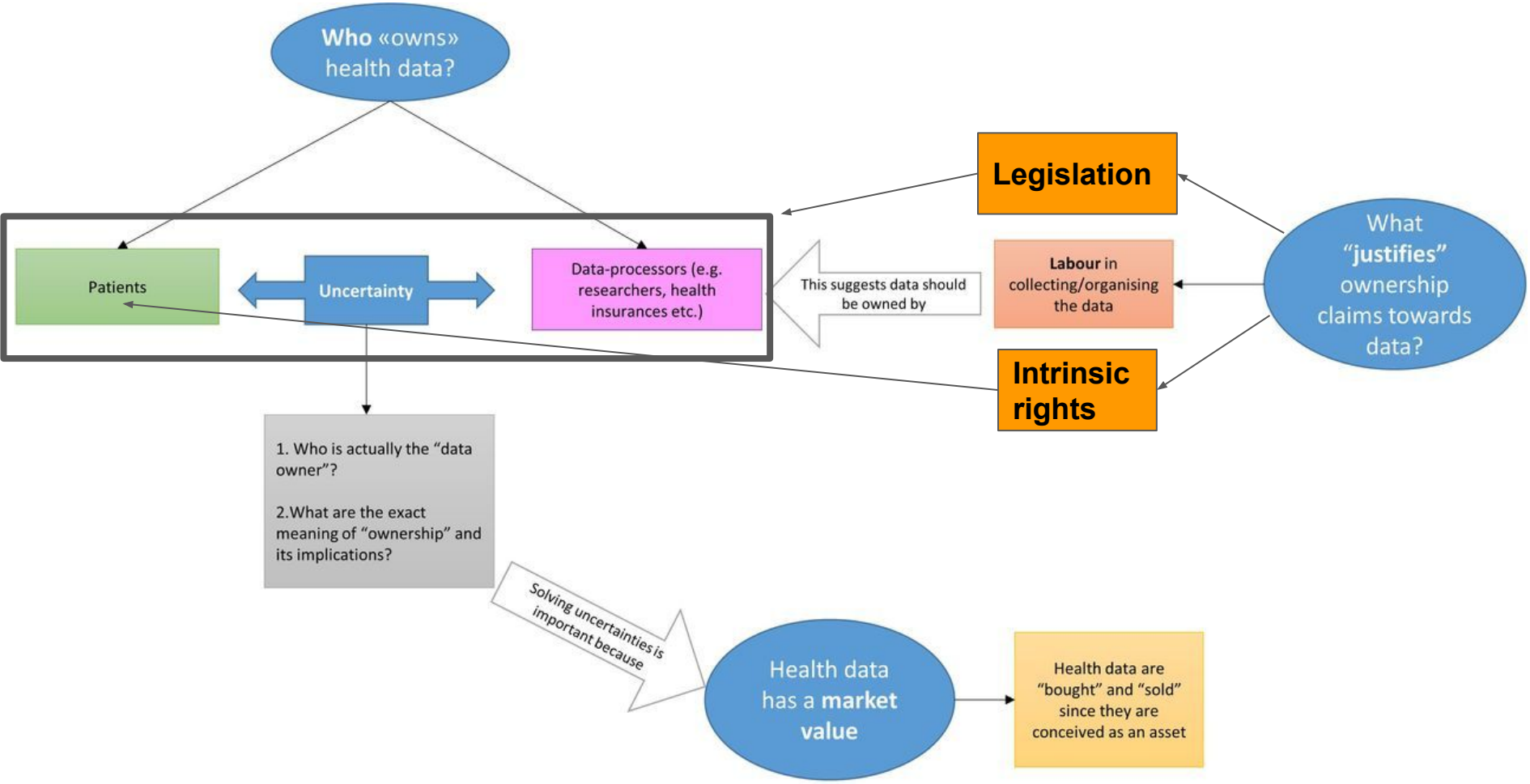
https://bmjopen.bmj.com/content/11/4/e045717

Data ownership is complicated



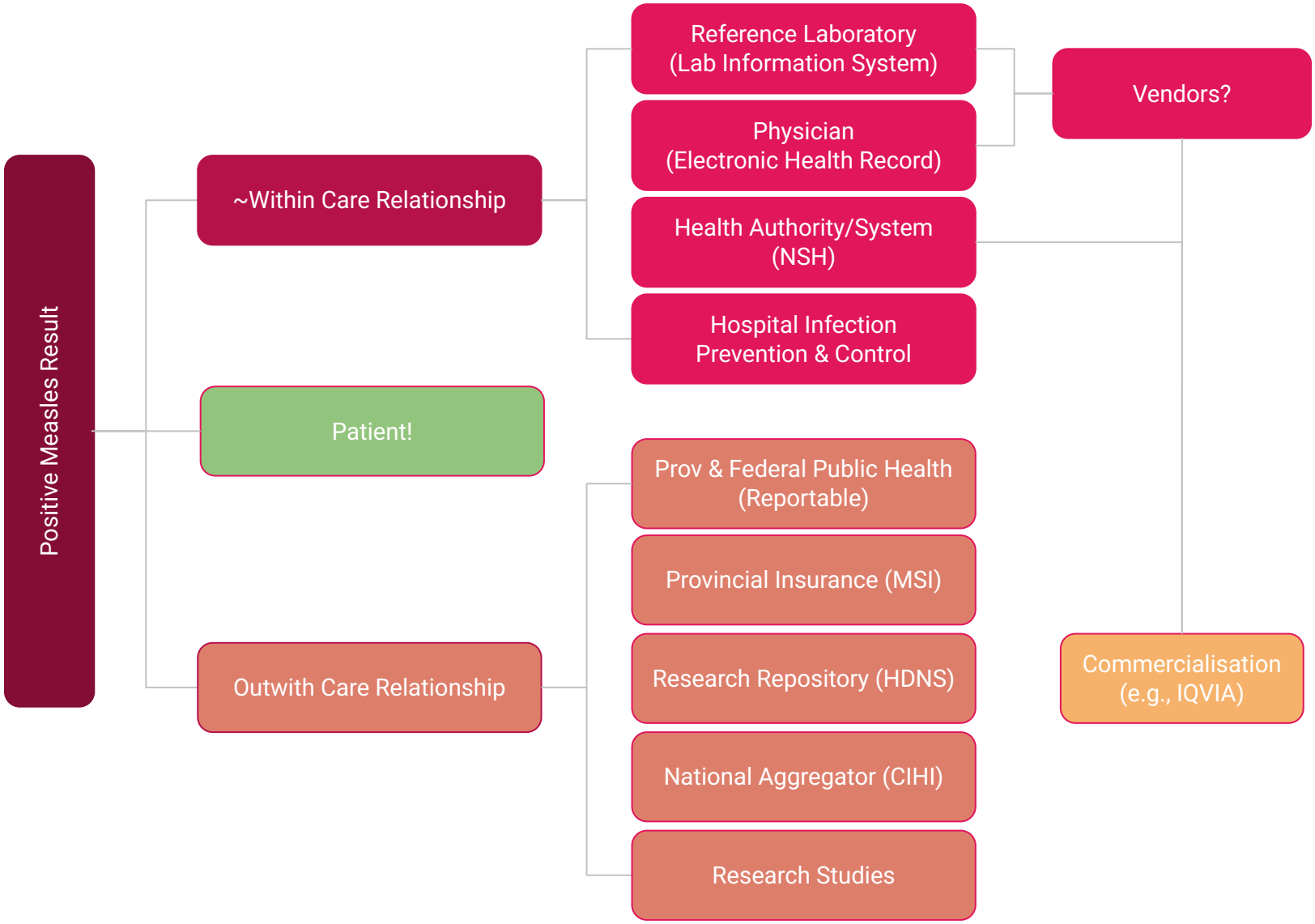
https://bmjopen.bmj.com/content/11/4/e045717

Data ownership is complicated



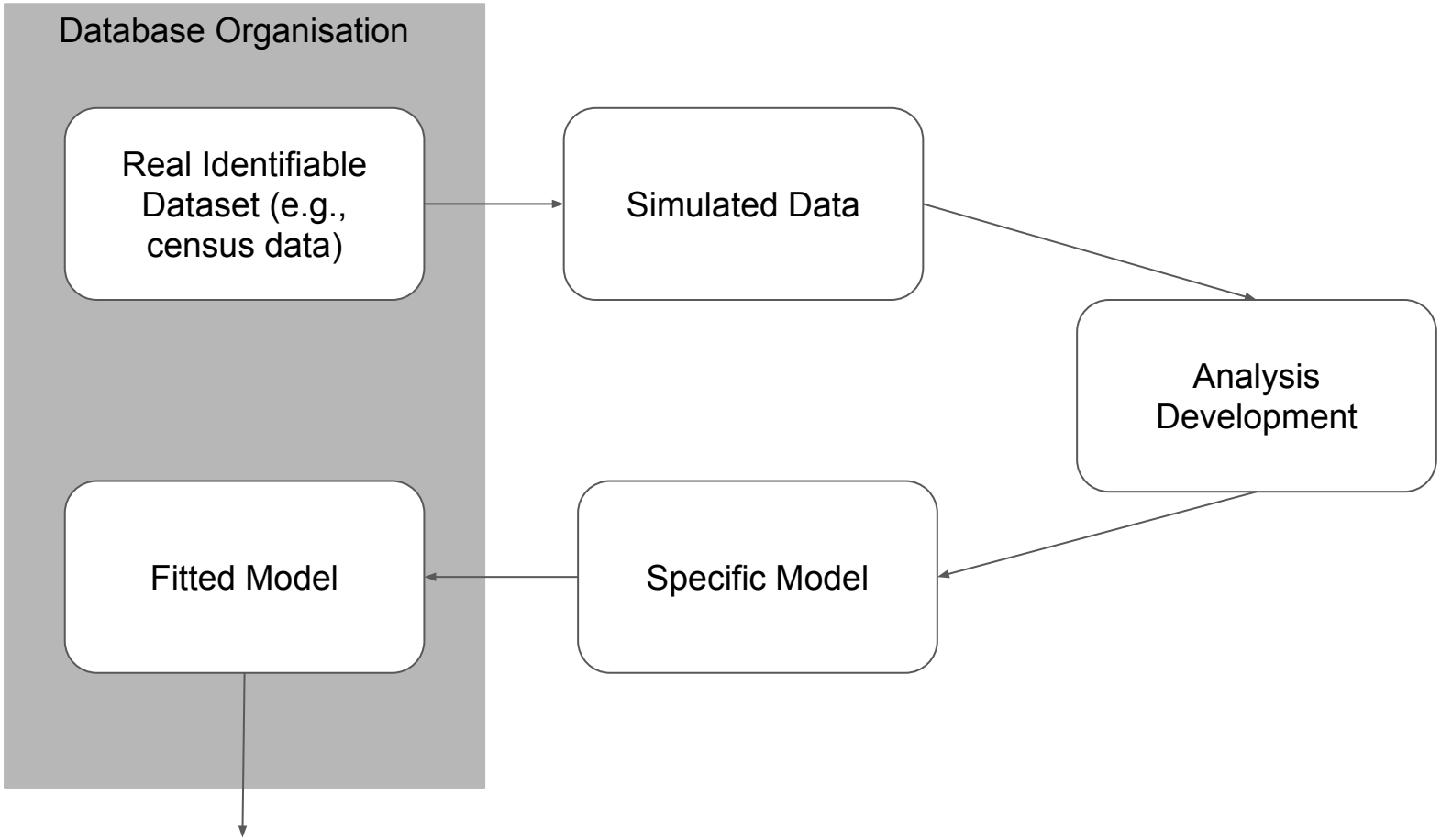
https://bmjopen.bmj.com/content/11/4/e045717

Ownership over which copy of same data?

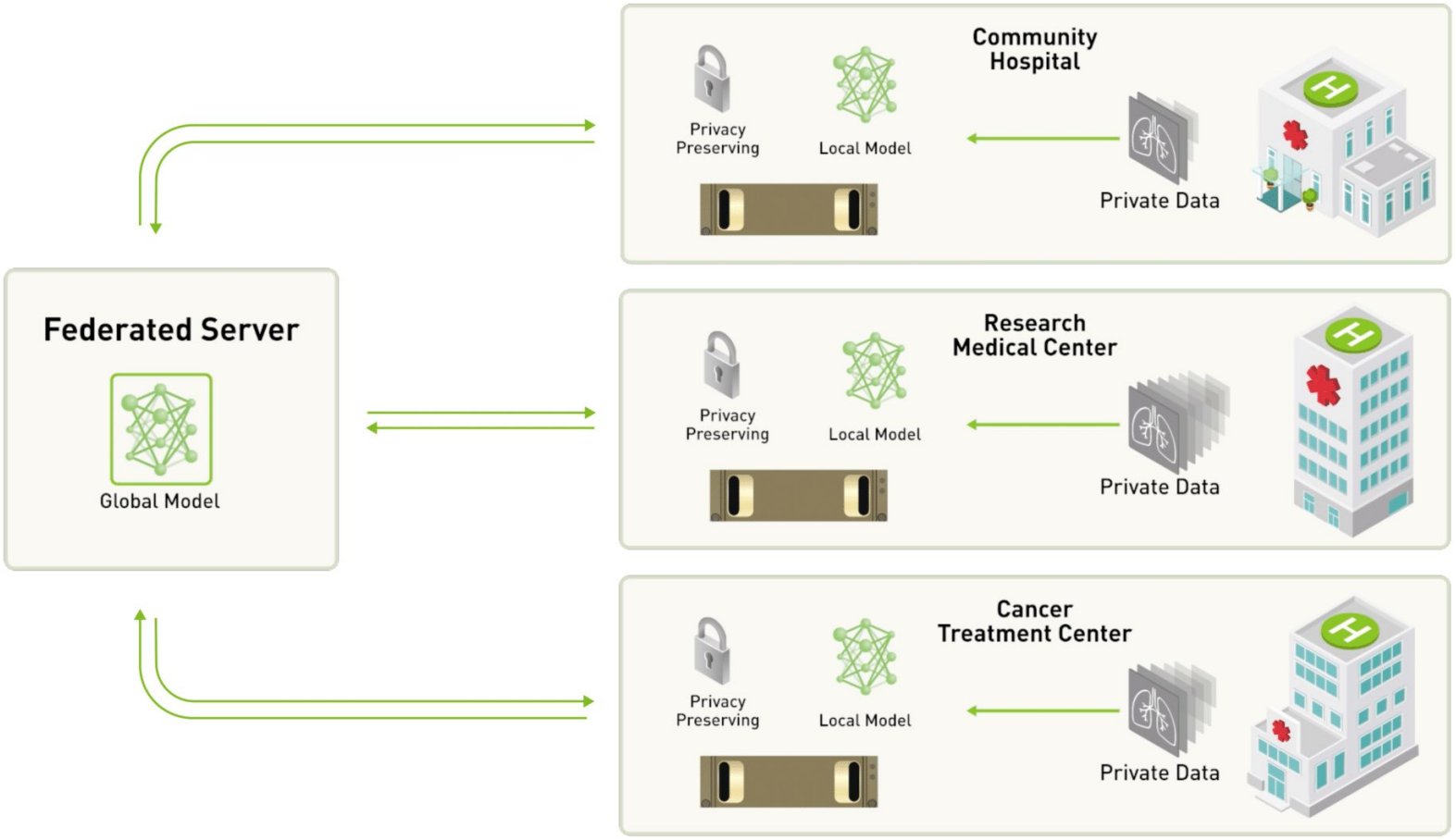


How do you protect privacy in these databases?

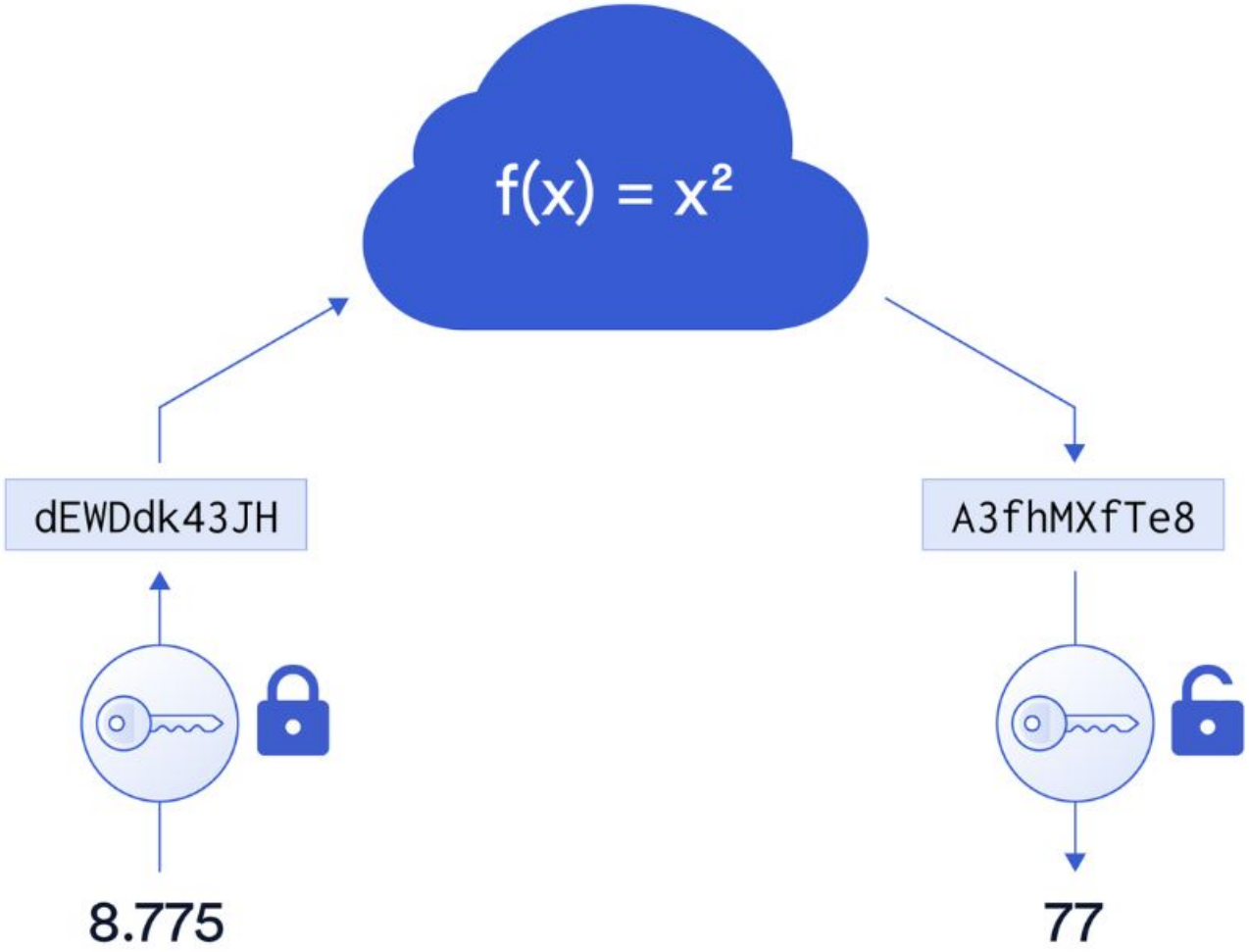
No direct data access



No external data access: federated training



Shared data but encrypted: homomorphic encryption



These are difficult and limited… so how can we share data directly but safely?

Data privacy is a continuum

Direct Identifiers:

- Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SIN, MSI Number)

Indirect Identifiers:

- Data that identifies an individual indirectly but can so in aggregate (e.g., DoB, gender, IP address, license plate)

Pseudonymous Data:

- Direct identifiers eliminated or transformed but indirect intact (e.g., names replace with unique identifier)

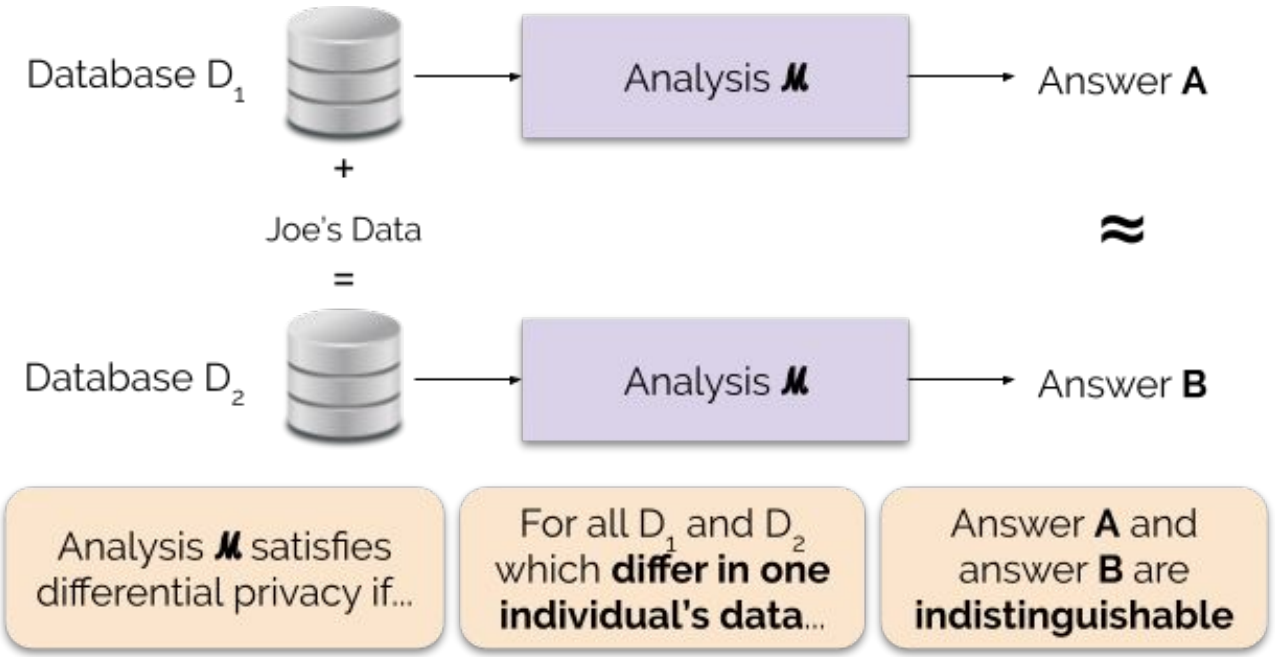
De-identified Data:

- Direct and indirect identifiers removed or modified to break linkage to real-world identities (e.g., data suppressed, generalized, perturbed, swapped - ranges applied)

Anonymous Data:

- Direct and indirect identifiers removed or modified to ensure mathematical guarantees against re-identification (e.g., aggregated census data, noise added to dataset)

Differential privacy: no singling out individuals



https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our

Differential privacy: no singling out individuals

Analysis $\mathcal{M}$ satisfies differential privacy if...

For all D_1 and D_2 which **differ in one individual's data**...

Answer **A** and answer **B** are **indistinguishable**

Probability of seeing output O on input D_1

Probability of seeing output O on input D_2

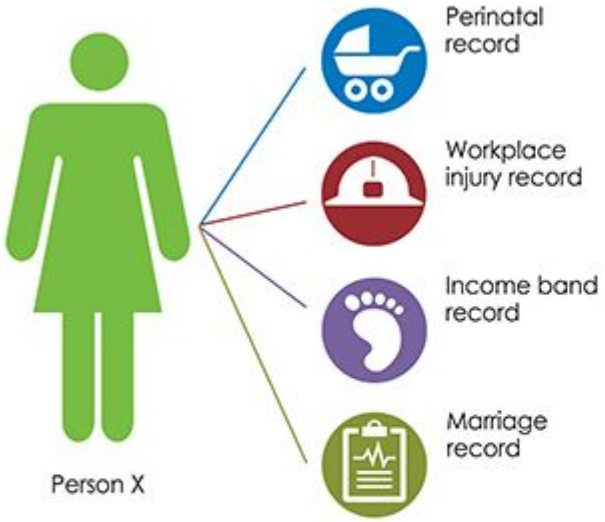
$$\frac{\Pr[\mathcal{M}(D_1) \in O]}{\Pr[\mathcal{M}(D_2) \in O]} \leq e^{\varepsilon}$$

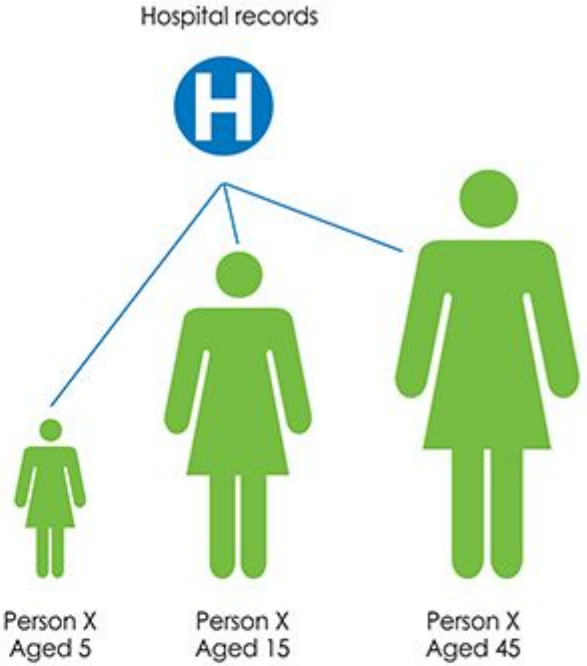
Indistinguishability: bounded ratio of probabilities

https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our

Data linkage is powerful but dangerous

- Linking between databases and resources -> identifiability
- Can be done probabilistically
- Often needs additional ethics/applications
- Can break a lot of data privacy operations





So, you’ve got access to a database, what now?

Data Cleaning: even “simple” fields can be a nightmare

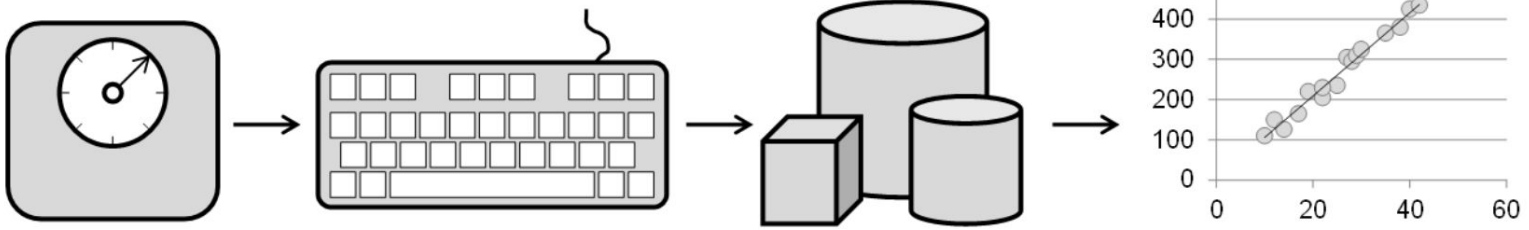
Data Quality

Actual value:
200.6 lbs.

Recorded value:
200 lbs.

Data warehouse value:
200 kg

Analytic value:
100 kg (mean 200 & 0)



Measured (same day)
• Validity challenge
198.9 | 198.9 | 198.9 lbs.
• Reliability challenge
200.6 | 198.9 | 202.2 lbs.

Measured (diff. days)

User Typed (one entry)
• Typos
200.6 lbs. → 20.06, 2006
• Mismatching units
200.6 lbs. → 200.6 kg
• Assumptions/truncations
200.6 lbs. → 200 lbs.
NULL → 0
• Free-text additions
200.6 lbs. → 200.6 pounds

DB Operations (one entry)
• Truncations/Rounding
200.6 → 200.0
• Error conversions
200.6 pounds → NULL
200.6 lbs. → 200.6 kg
• Cleaning
200+ lbs. → 200.0

Analytics (data points)
• Aggregation of data points
200 | 0 → mean of 100
• Selecting a representative
190 | 200 | 210 → 210 (first)
190 | 200 | 210 → 200 (mean)
190 | 200 | 210 → 210 (last)
• Removing outliers
200 | 200 | 350 → 200 | 200 | NULL

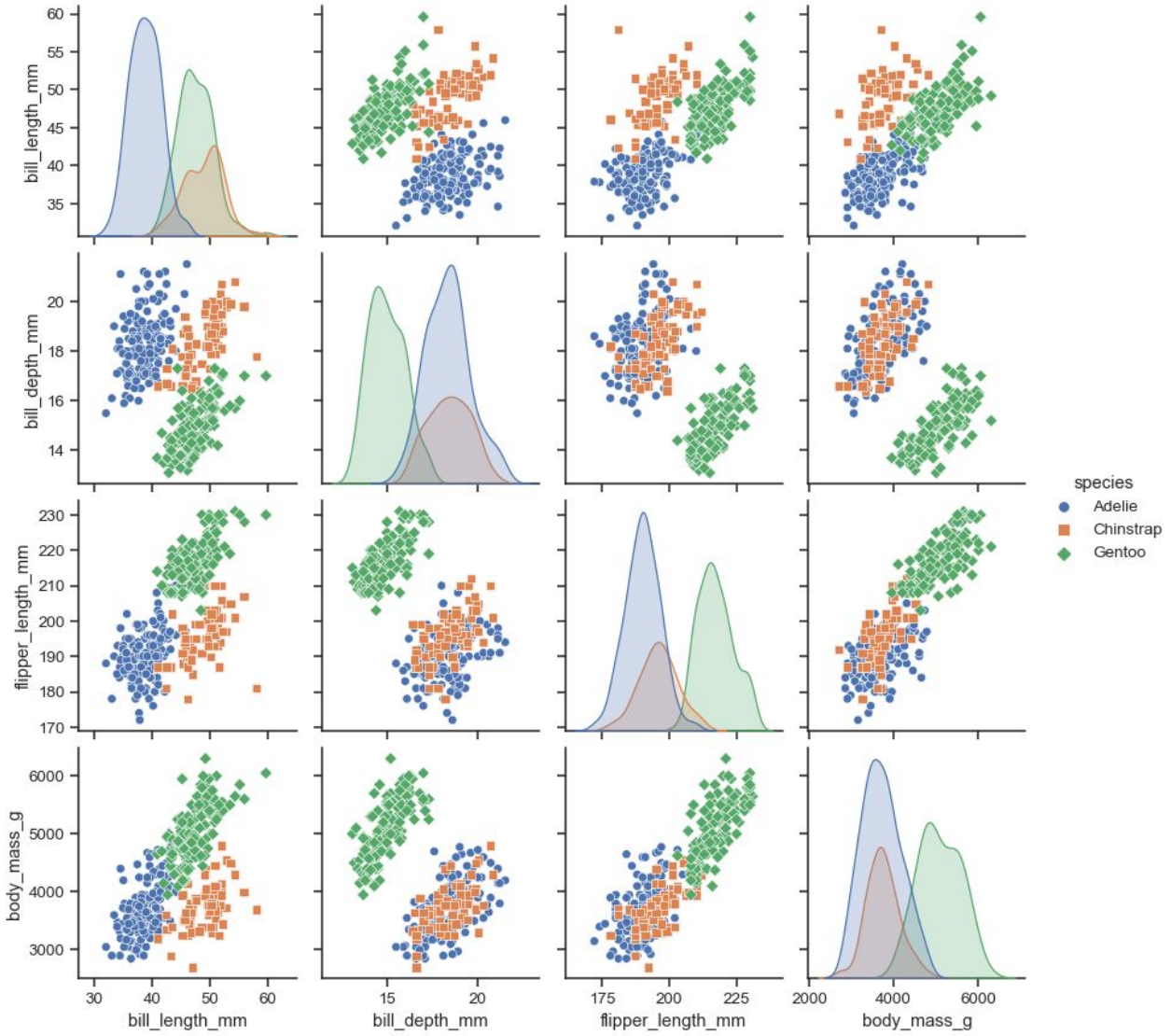
Under review

Slide from Dr. Hadi Kharrazi

9 months & >25 rules to clean weight

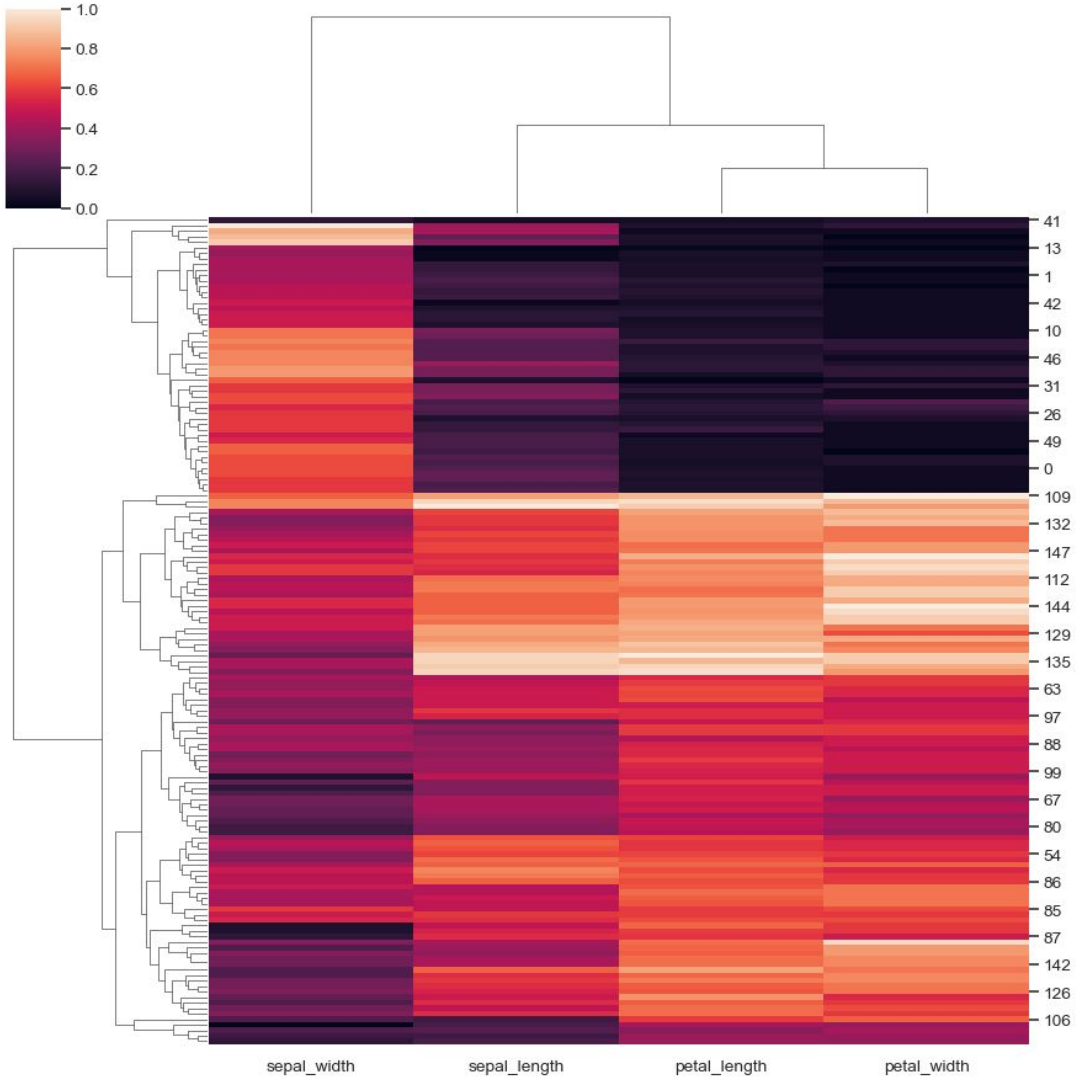
Exploratory Data Analysis

- Individual variable distributions
- Pairwise variable distributions
- Distributions relative to variable(s) of interest
- Point analysis of extreme values



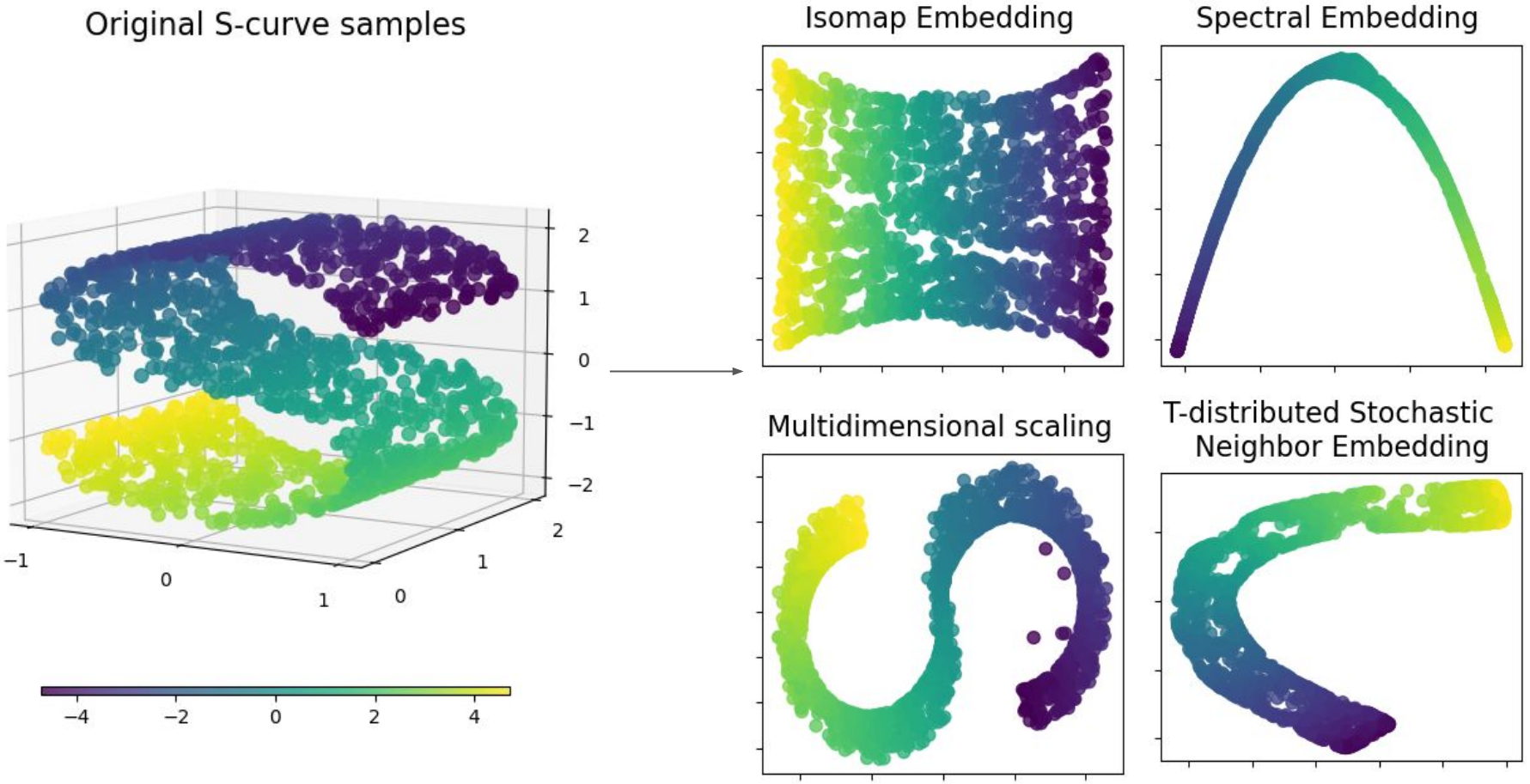
Exploratory Data Analysis

- Individual variable distributions
- Pairwise variable distributions
- Distributions relative to variable(s) of interest
- Hierarchical clustering of variables
- Point analysis of extreme values



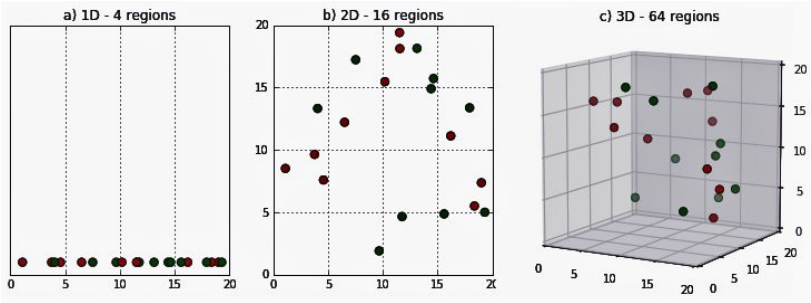
How do I look at all the data together?

Many dimensions to few: Manifold learning, Ordination, Decomposition, Dimensionality reduction



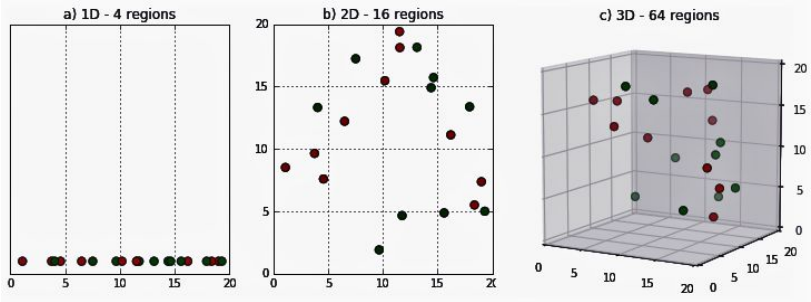
Why is this hard? High dimensions are counter-intuitive!

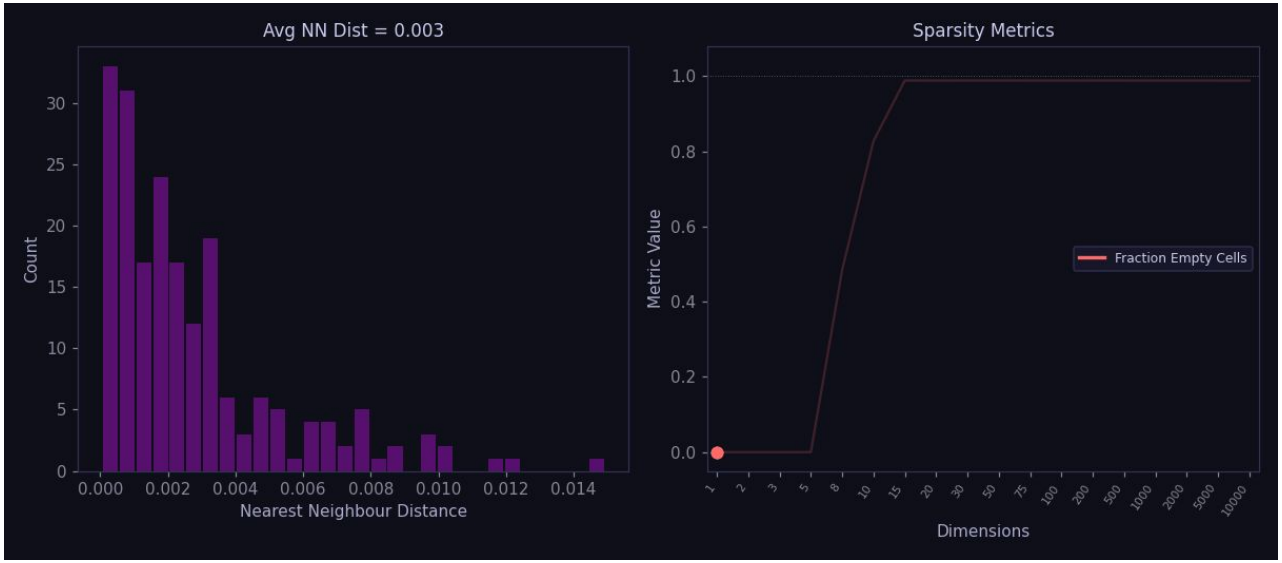
Data becomes increasingly sparse in high dimensions



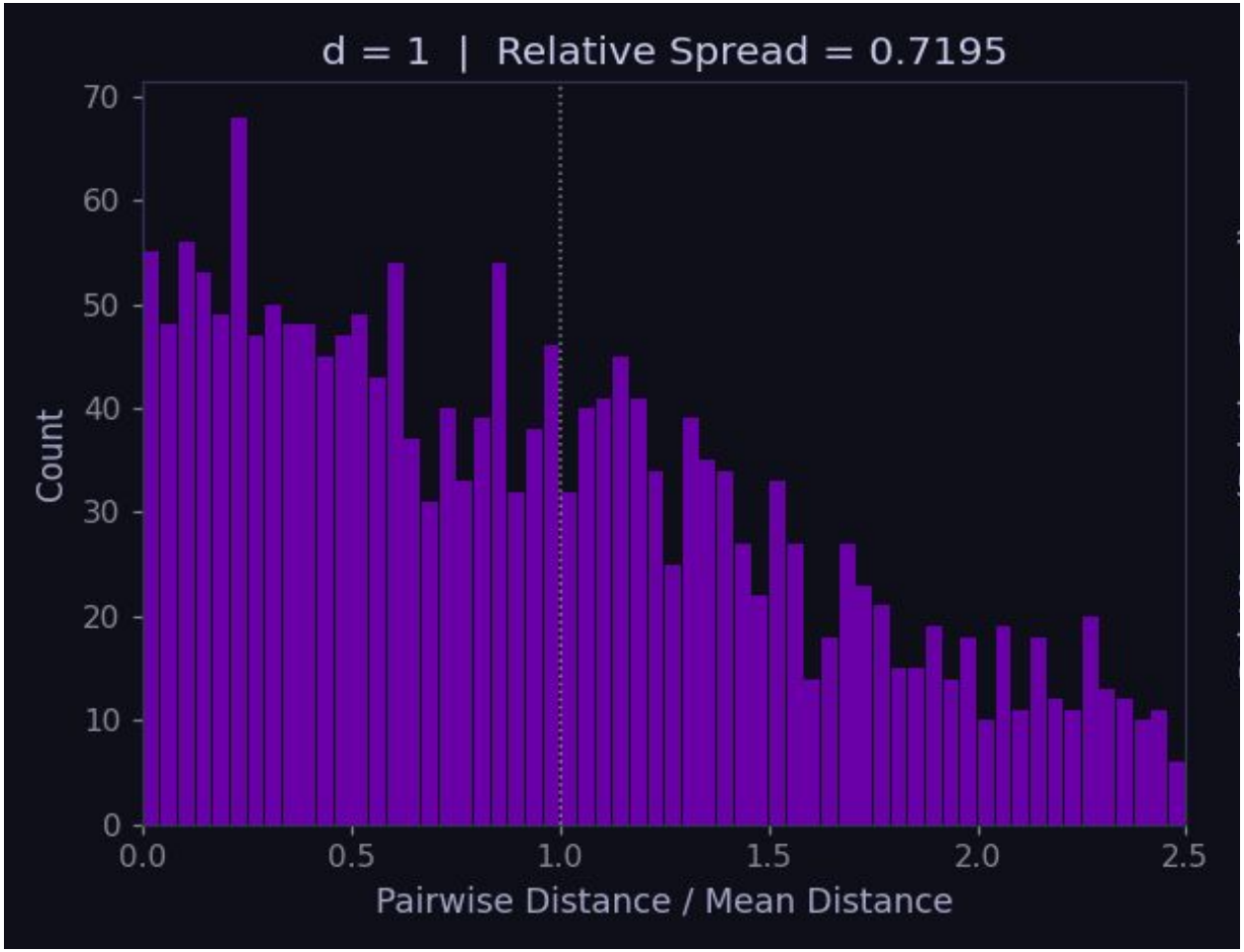
https://medium.com/analytics-vidhya/the-curse-of-dimensionality-and-its-cure-f9891ab72e5c

Data becomes increasingly sparse in high dimensions

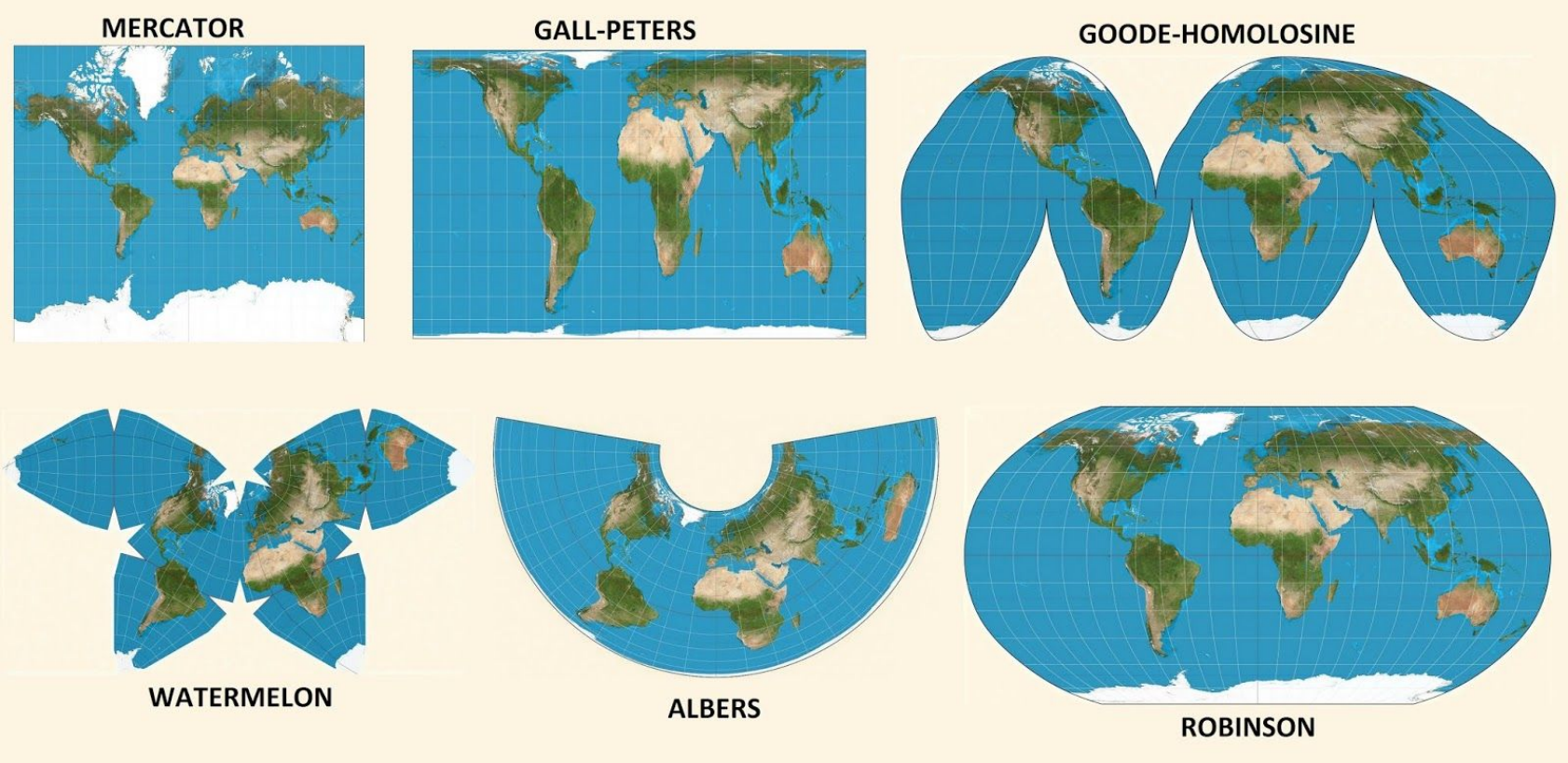




All points start to look equally far apart



No lower dimension representation will ever be perfect

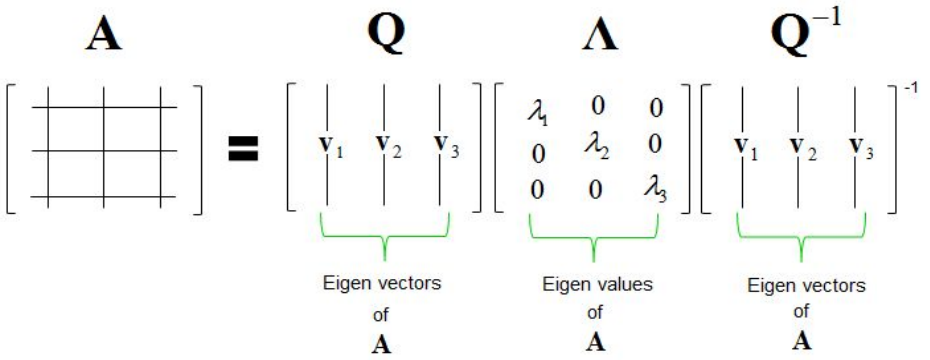


So, how can we do it?

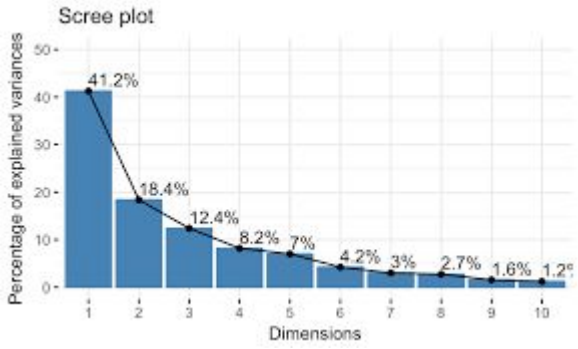
Principal Component Analysis - Simplest Method

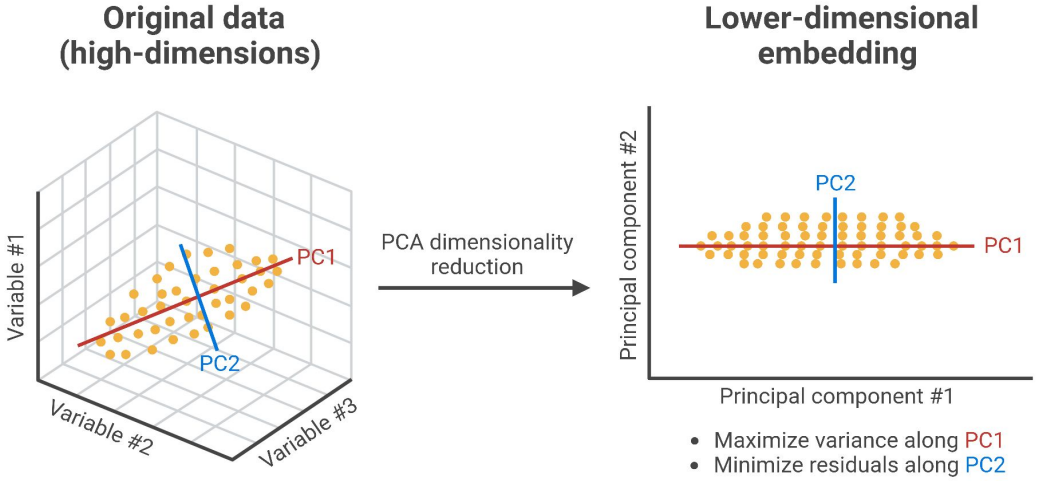
Reorient the data in the direction of maximal variance

1. Center the data
2. Calculate the covariance matrix
3. Perform eigendecomposition
4. Sort and select n principal components
5. Project the data onto the reduced space



How many components? Scree/elbow plot





https://www.biorender.com/template/principal-component-analysis-pca-transformation

MultiDimensional Scaling (MDS): Distances

$$Stress_D(x_1, x_2, \ldots, x_N) = \sqrt{\sum_{i \neq j = 1,\ldots,N} (d_{ij} - ||x_i - x_j||)^2}$$

The goal of the algorithm is to minimize the value of stress.

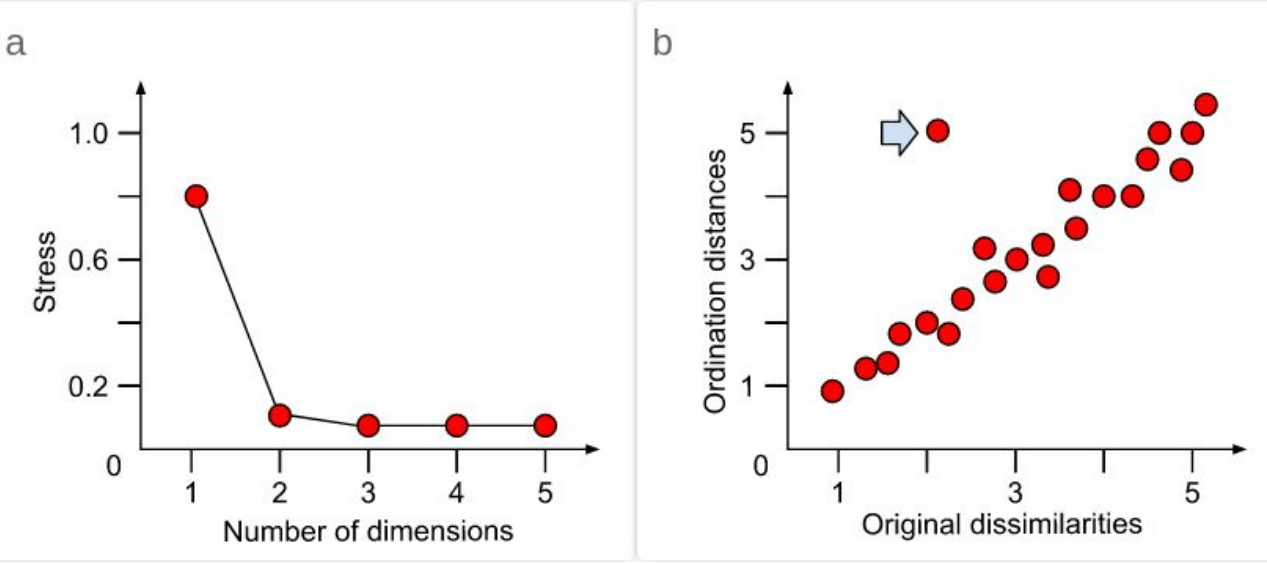
Where $x_1, \ldots, x_N$ are data points with their new set of coordinates in lower dimensional space.

d_{ij} is the actual distance we have calculated between the two corresponding data points in their original dimensional space.

$||x_i - x_j||$ is the distance between the two corresponding data points in their lower dimensional space.

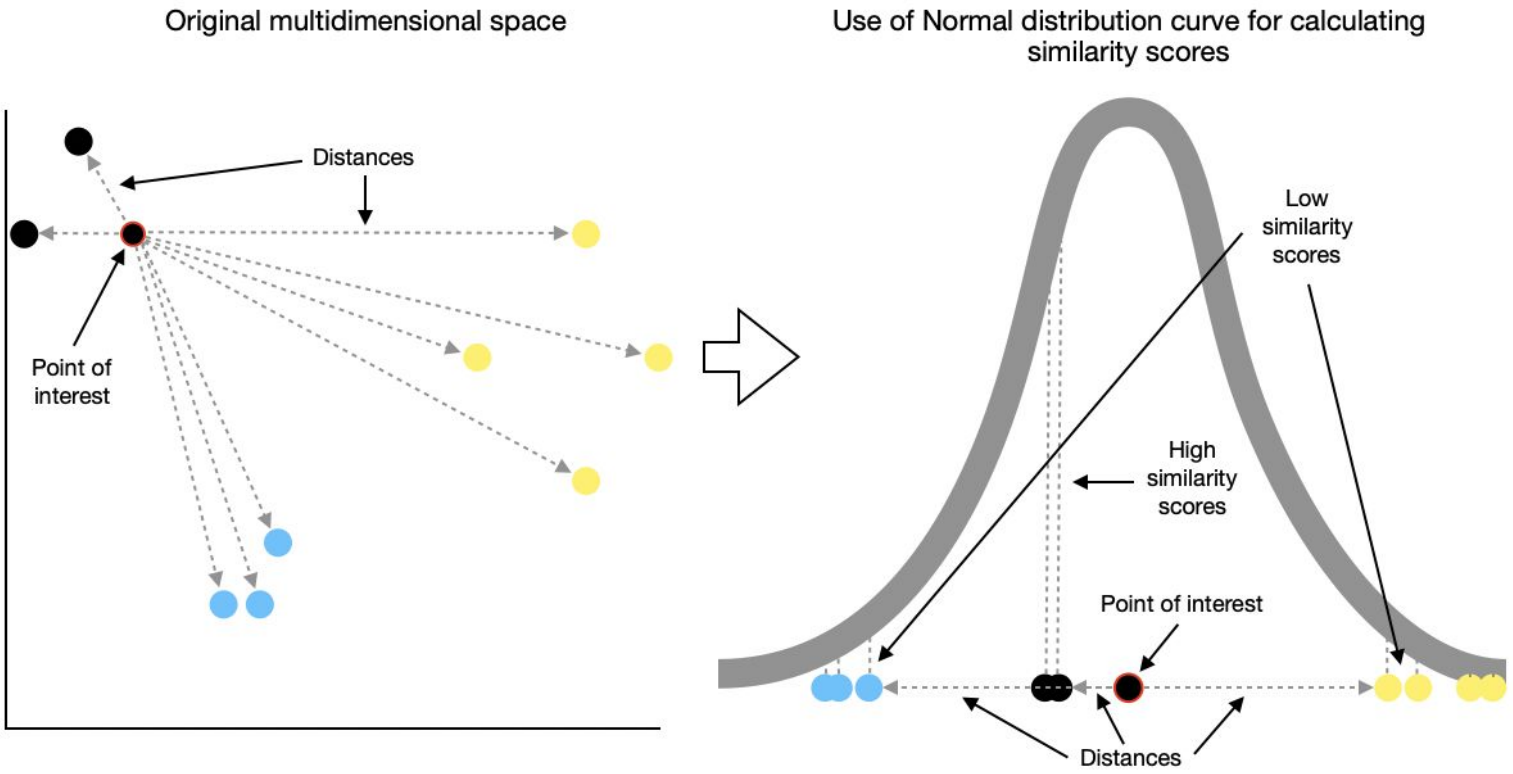
The closer the value of $||x_i - x_j||$ is to d_{ij} the smaller will be the value of stress.

Non-Metric: Ranks



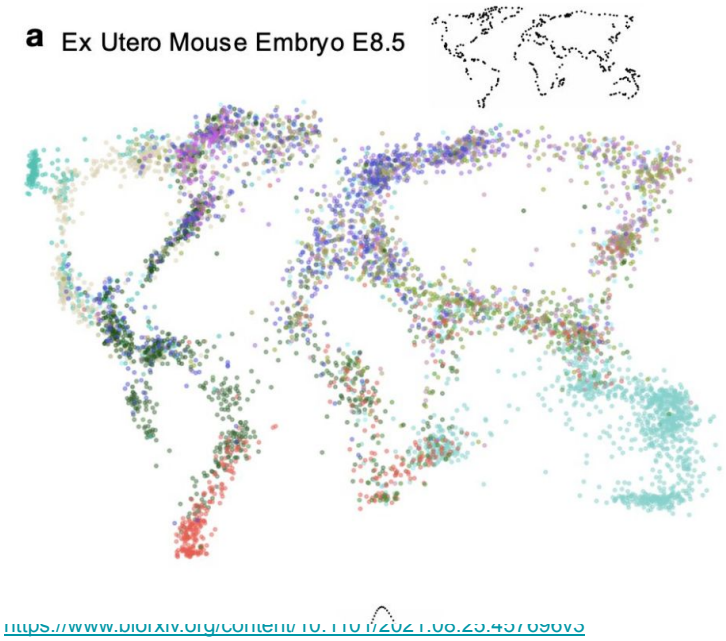
t-SNE/UMAP: Probabilities

- Pairwise probability distribution in all dimensions
- Pairwise probability distribution in few dimensions
- Stochastic minimisation of KL divergence between distributions

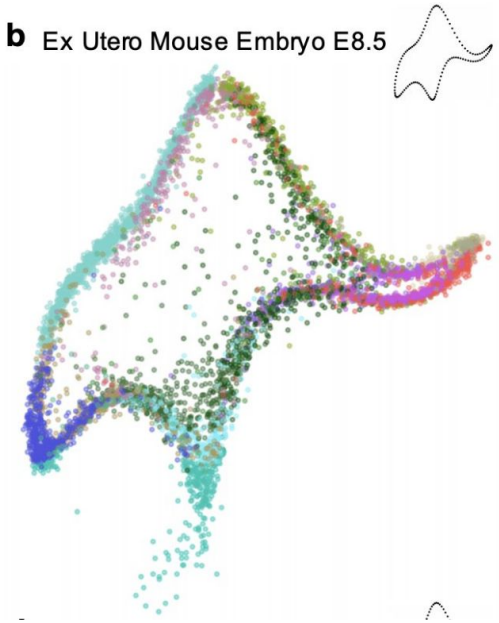


Avoid over-interpreting single plots

- Sensitive to hyperparameters
- Beware analysing these non-linear projections
- Can contribute to confirmation bias



https://www.biorxiv.org/content/10.1101/2021.08.25.457696v3

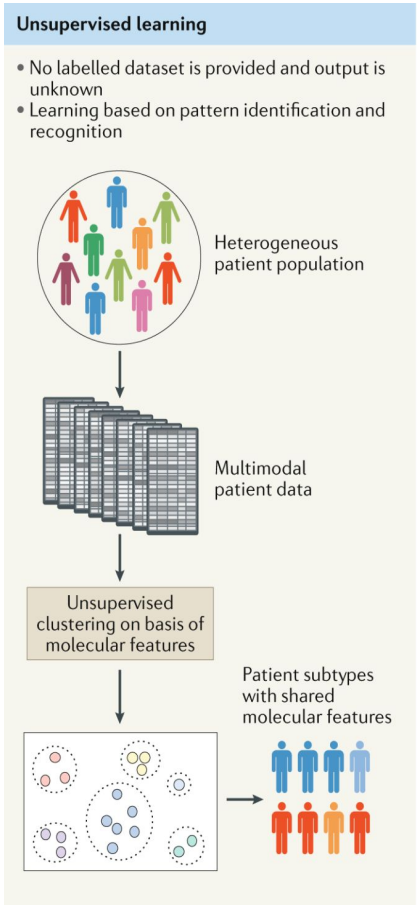


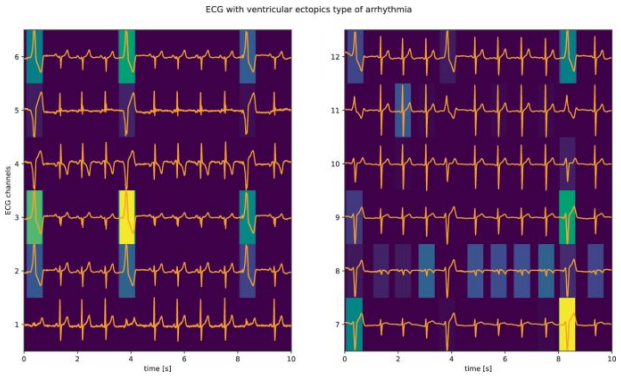
"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." - Von Neumann

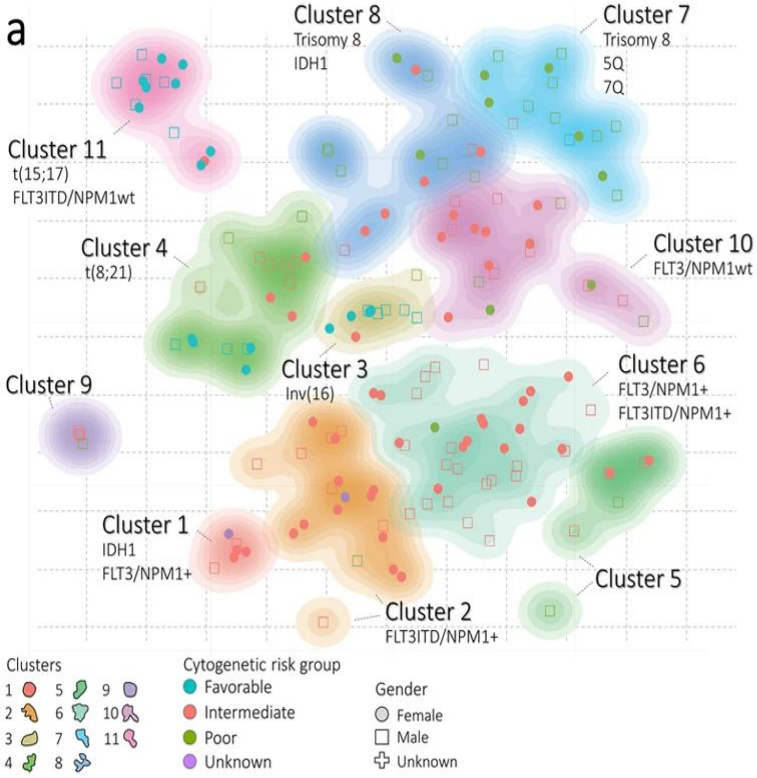
Finding structure in data: unsupervised learning

Anomaly Detection - ECG

Clustering - Acute Myeloid Leukemia Subtypes



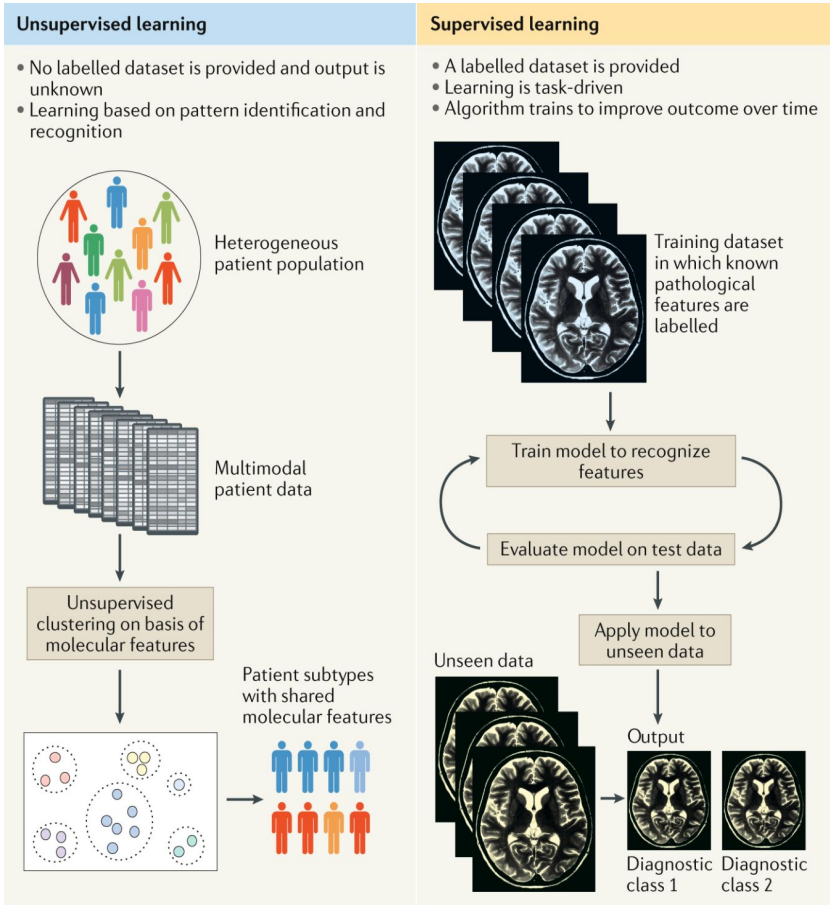




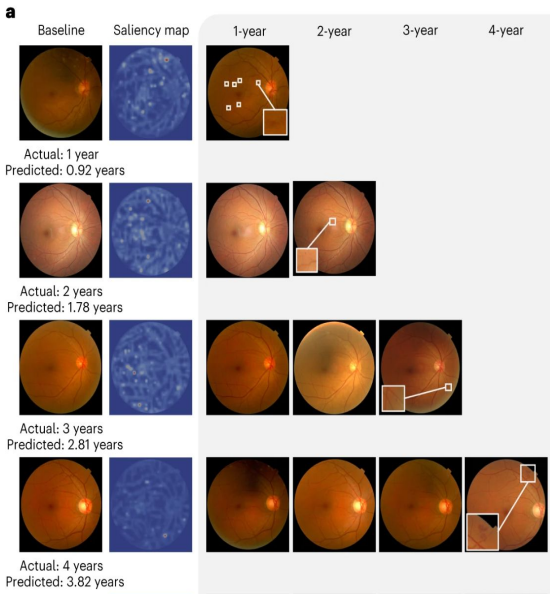
*Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol **16**, 440–456 (2020). https://doi.org/10.1038/s41582-020-0377-8*

*Taskesen, E., Huisman, S., Mahfouz, A. et al. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. Sci Rep **6**, 24949 (2016). https://doi.org/10.1038/srep24949*

Predicting outcomes from data: supervised learning

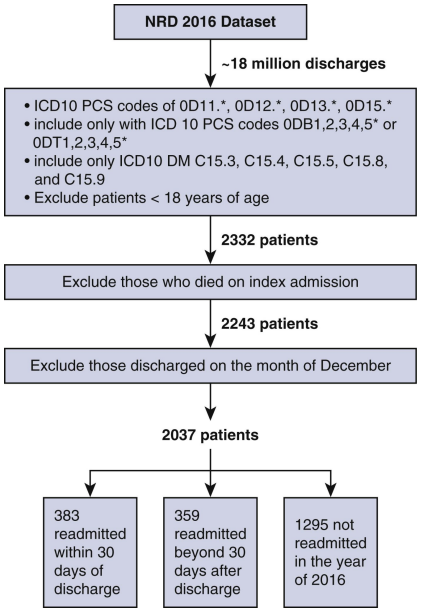


Prediction of Diabetic Retinopathy Progression



*Dai, L., Sheng, B., Chen, T. et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nat Med **30**, 584–594 (2024). https://doi.org/10.1038/s41591-023-02702-z*

Prediction of 30-Day Readmission



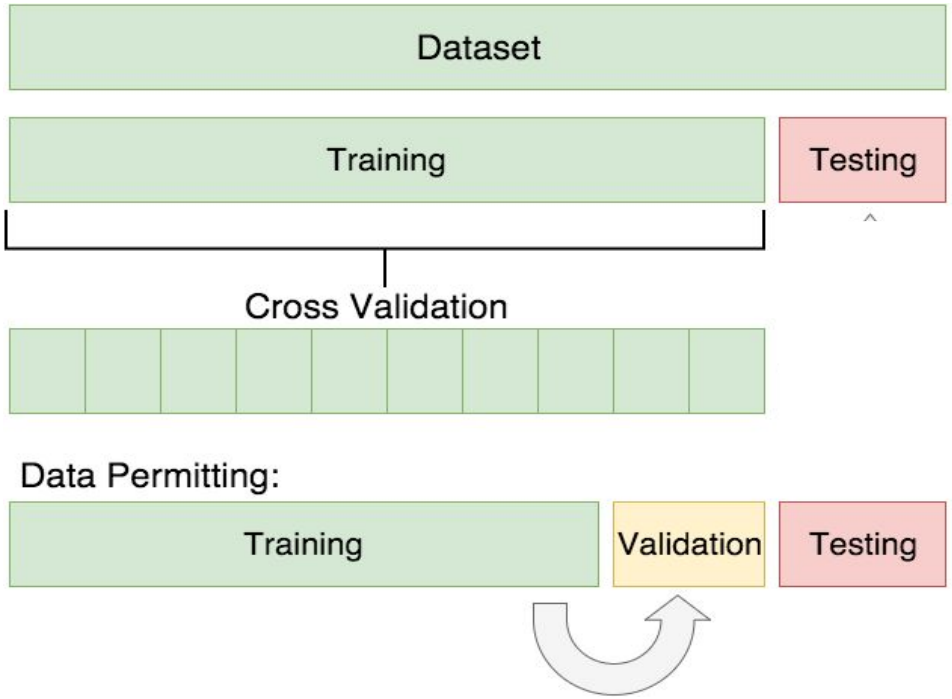
Bolourani, Siavash, et al. "Using machine learning to predict early readmission following esophagectomy." The Journal of Thoracic and Cardiovascular Surgery 161.6 (2021): 1926-1939.

Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P002	41	F	22.1	118	75	4.9	5.2	4.7	No	No	Healthy	4.1%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P005	58	M	29.9	145	91	6.9	6.5	5.8	No	No	Diabetic	14.7%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%

Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P002	41	F	22.1	118	75	4.9	5.2	4.7	No	No	Healthy	4.1%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P005	58	M	29.9	145	91	6.9	6.5	5.8	No	No	Diabetic	14.7%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	28.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%



What does it mean to fit a supervised model?

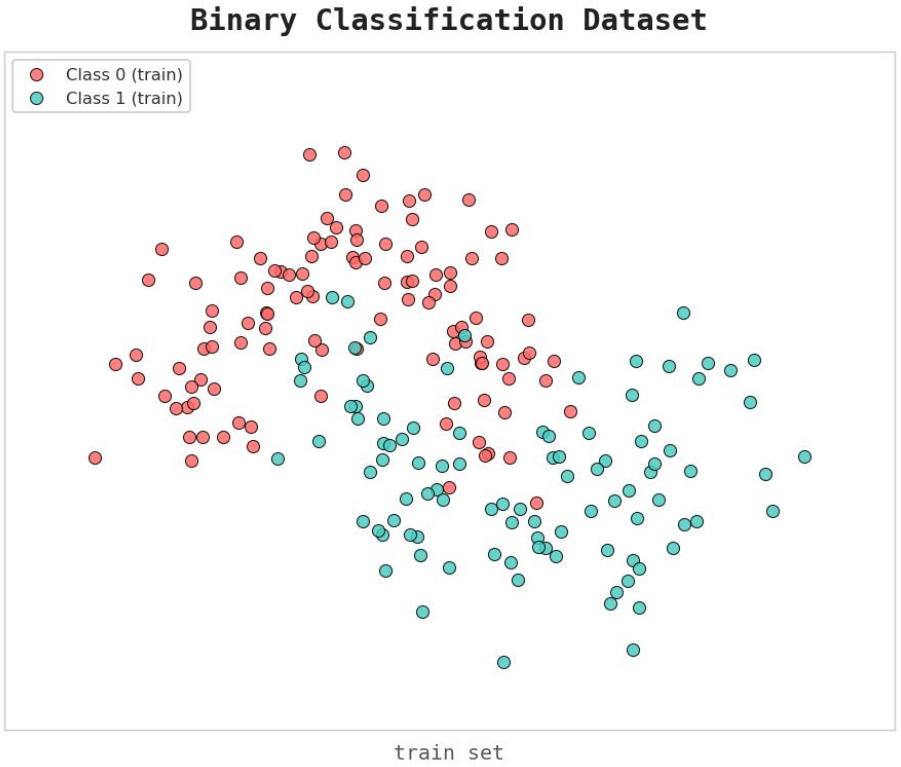
Decision boundaries and loss functions

Find β so our function maps X to y

$$\hat{y} = f_\beta(x)$$

In other words: find β that defines boundary between labels

Minimising some loss function like binary cross-entropy (log-loss)



$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

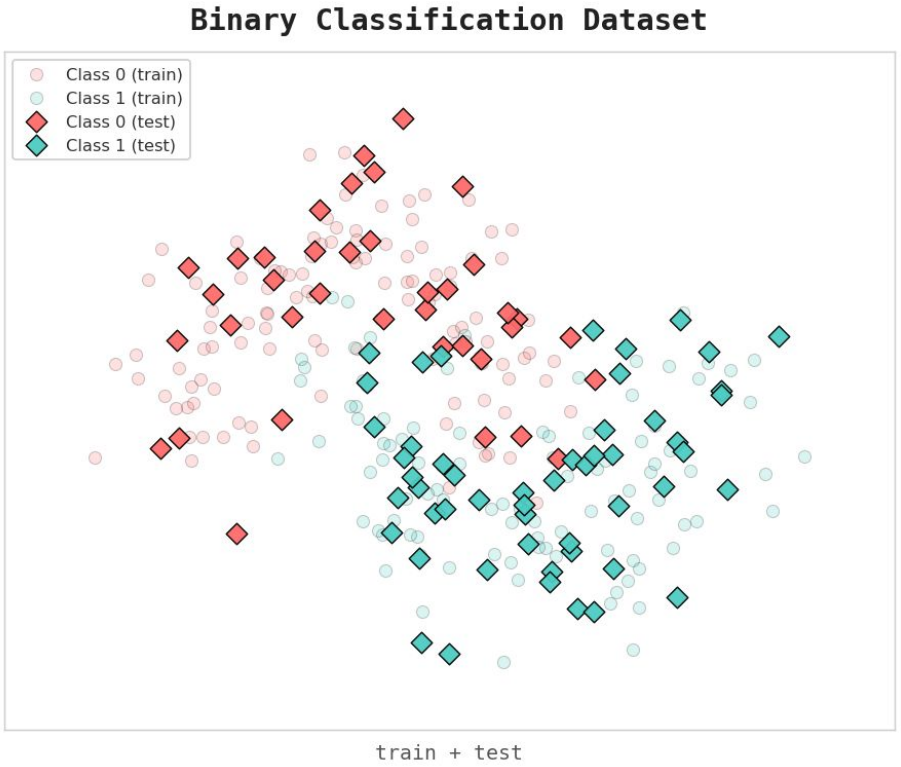
Decision boundaries and loss functions

Find β so our function maps X to y

$$\hat{y} = f_\beta(x)$$

In other words: find β that defines boundary between labels

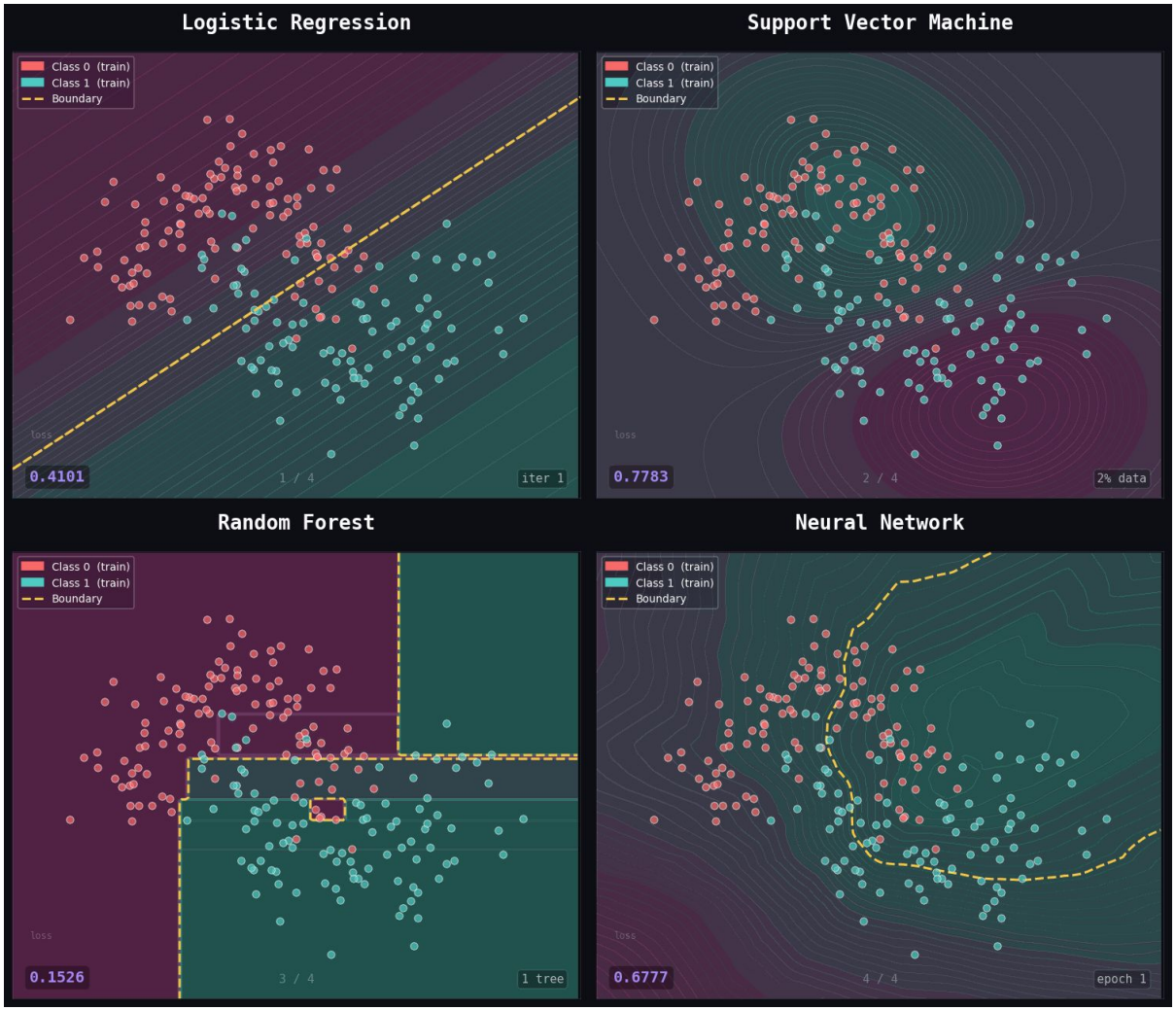
Minimising some loss function like binary cross-entropy (log-loss)



$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

Ideally in a way that generalises to new data!

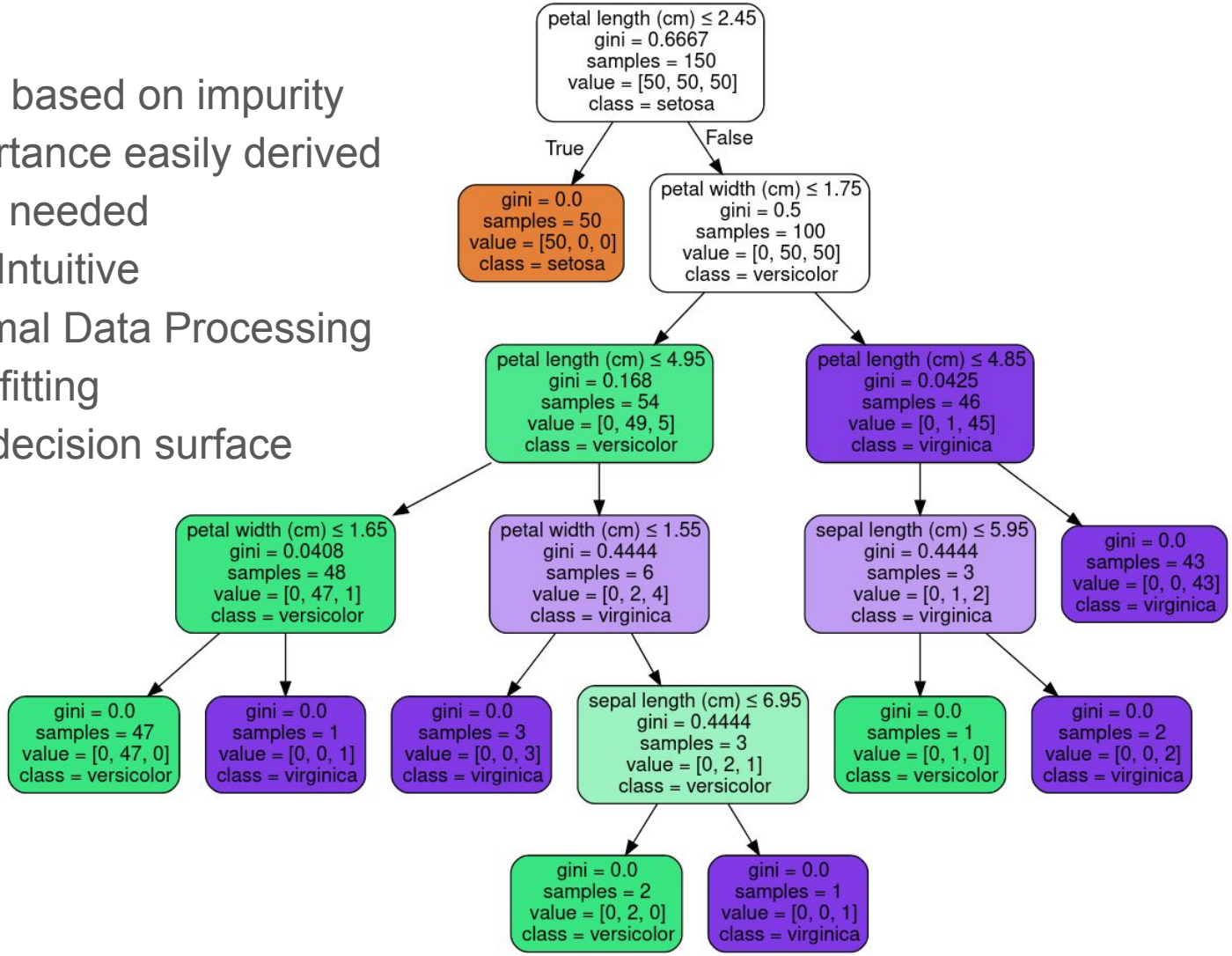
Decision boundaries and loss functions



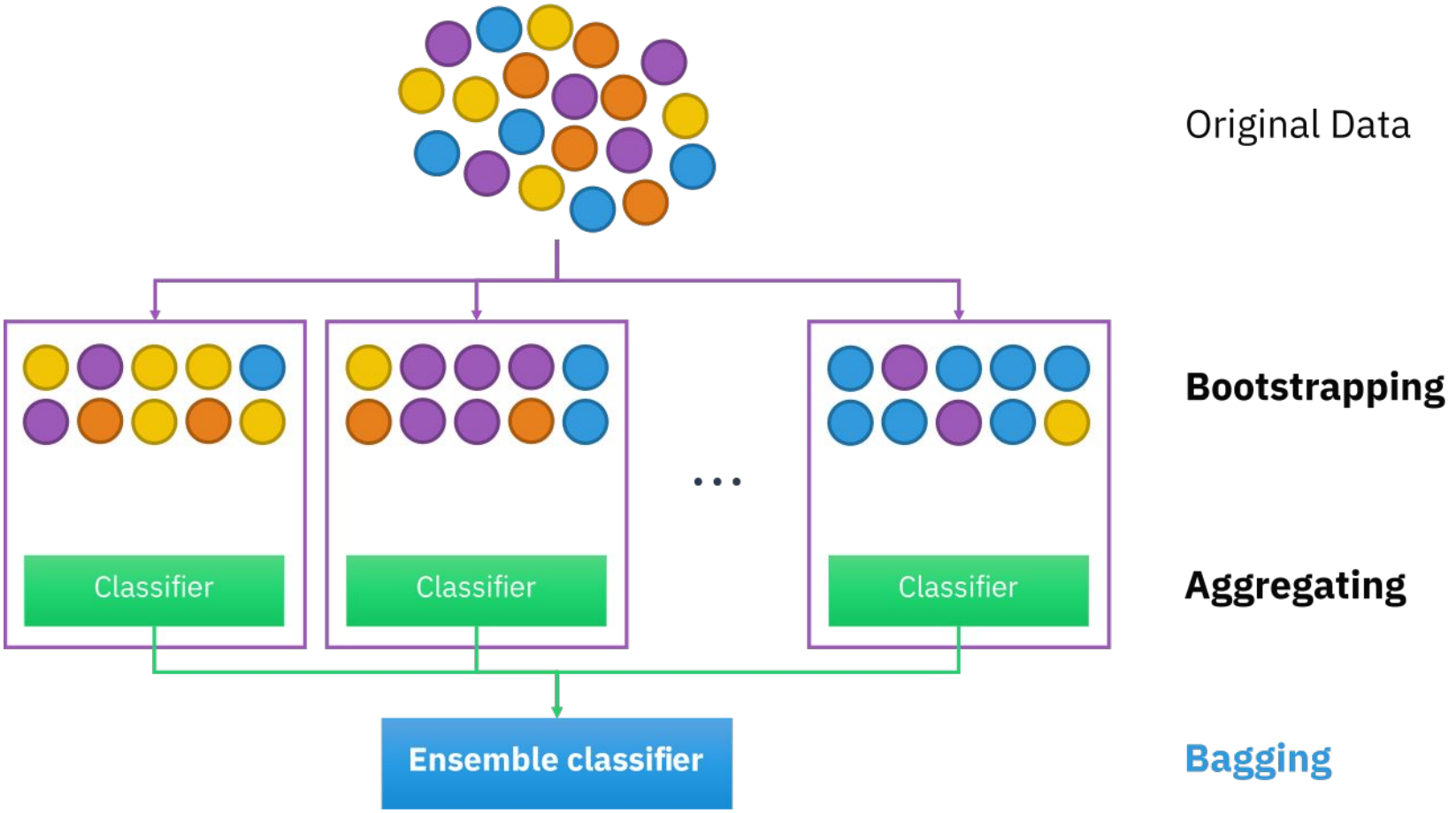
Can’t get a well-performing model? Just combine lots of weak ones!

Decision Trees

- Dataset splits based on impurity
- Feature importance easily derived
- Pruning often needed
- Interpretable/Intuitive
- Require Minimal Data Processing
- Prone to overfitting
- Non-smooth decision surface

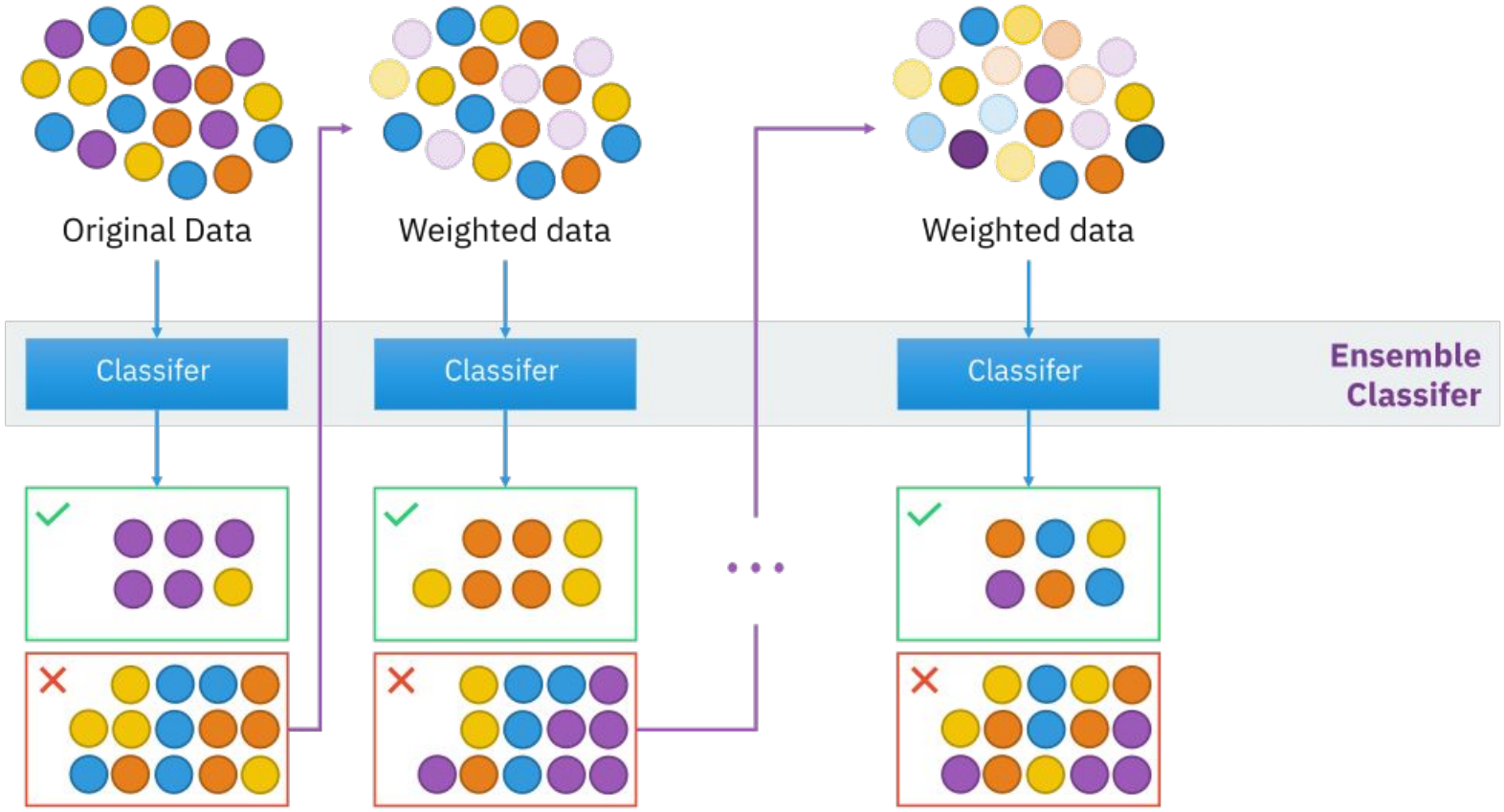


Many Decision Trees: Bagging



Random Forest: Bagging + Random Subset Per Split
Feature Importance: Average impurity decrease

Boosting: AdaBoost, Gradient Boosting, XGBoost



AdaBoost, Gradient Boosting, XGBoost

Decision Trees methods regularly outperform deep learning on tabular data

Tree-based methods deal well with common features of tabular data (even compared to well-tuned neural networks):

- Heterogeneous data
- Ignoring uninformative data
- Non-smooth decision boundaries
- Moderate size & dimensionality
- Skewed or heavy-tailed feature distributions and other forms of dataset
- Rotational invariance (column/row order is not informative)

But: difference is often negligible (except in computational efficiency!)

Why do tree-based models still outperform deep learning on typical tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
MLIA, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

When Do Neural Nets Outperform Boosted Trees on Tabular Data?

Duncan McElfresh[*1,2], **Sujay Khandagale**[3], **Jonathan Valverde**[4], **Vishak Prasad C**[5], **Ganesh Ramakrishnan**[5], **Micah Goldblum**[6], **Colin White**[1,7]
[1] Abacus.AI, [2] Stanford, [3] Pinterest, [4] University of Maryland,
[5] IIT Bombay, [6] New York University, [7] Caltech

Overview

- Medical databases are usually relational and are defined by their origin, primary record type, scope, and sampling strategy
- Standardisation is important and ontologies support that in medical databases
- Survey weights are key to compensate for complex sampling
- There is a continuum of approaches to retain data privacy (and data ownership is a complex issue)
- Individual and joint distributions are key EDA tools
- Dimensionality reduction (PCA, MDS, t-SNE) is very useful but can be challenging/misleading
- Start with simple classifiers e.g., logistic regression/decision tree
- Combine weak classifiers via bagging (bootstrapping data: Random Forest special form) or boosting (sequential training model on errors: AdaBoost/XGBoost) to improve performance.
- XGBoost gold-standard but requires more tuning than AdaBoost