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

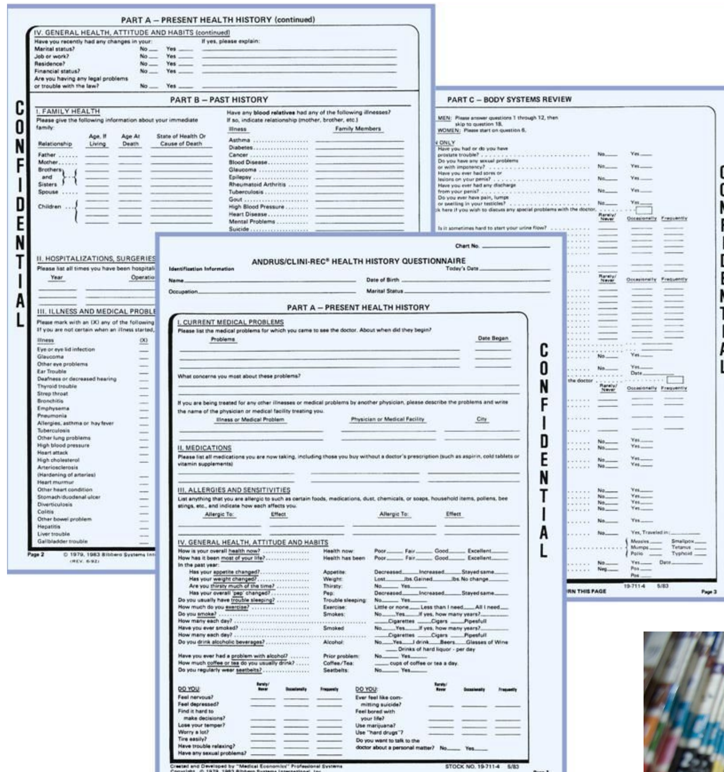
- Key strategies/approaches for exploratory data analysis
- Different types of dimensionality reduction
- Basics of supervised learning
- Accessing feature importances
- Aggregating simple/weak models to improve performance: boosting and bagging

What is a database?

Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts



PART A – PRESENT HEALTH HISTORY (continued)

IV. GENERAL HEALTH, ATTITUDE AND HABITS (continued)

PART B – PAST HISTORY

I. FAMILY HEALTH

PART C – BODY SYSTEMS REVIEW

ANDRUS/CLINI-REC® HEALTH HISTORY QUESTIONNAIRE

PART A – PRESENT HEALTH HISTORY

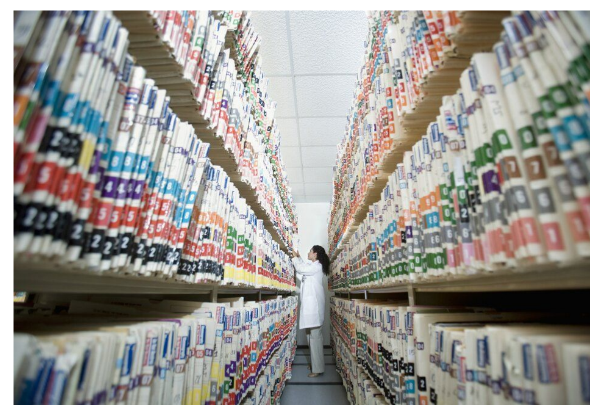
I. CURRENT MEDICAL PROBLEMS

II. MEDICATIONS

III. ALLERGIES AND SENSITIVITIES

IV. GENERAL HEALTH, ATTITUDE AND HABITS

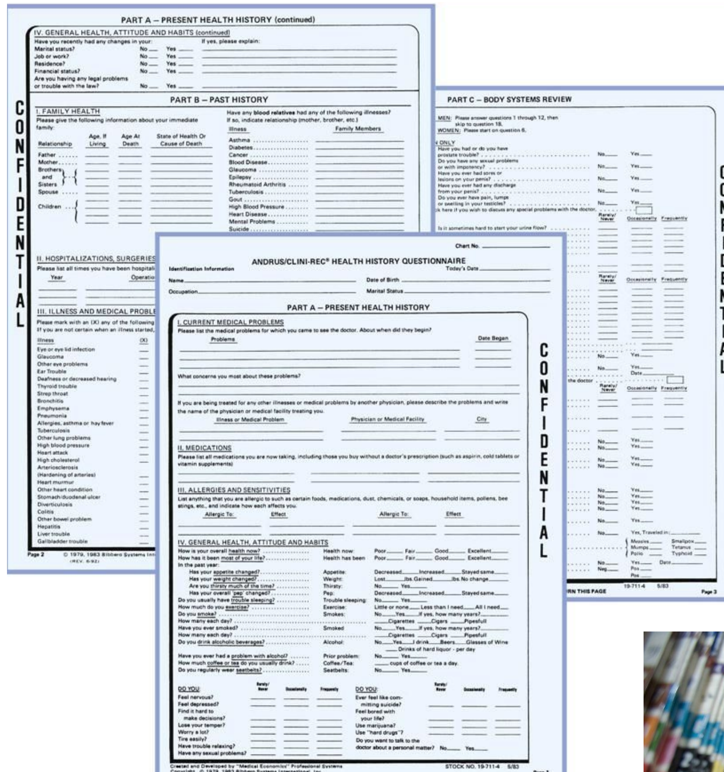
CONFIDENTIAL



Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet


PART A – PRESENT HEALTH HISTORY (continued)

IV. GENERAL HEALTH, ATTITUDE AND HABITS (continued)

PART B – PAST HISTORY

I. FAMILY HEALTH

PART C – BODY SYSTEMS REVIEW

CONFIDENTIAL

ANDRUS/CLINI-REC® HEALTH HISTORY QUESTIONNAIRE

PART A – PRESENT HEALTH HISTORY

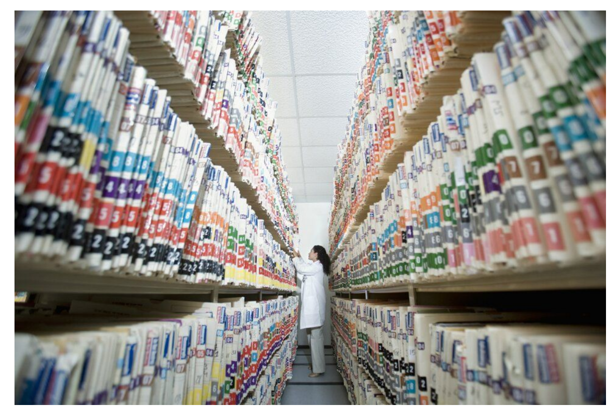
I. CURRENT MEDICAL PROBLEMS

II. MEDICATIONS

III. ALLERGIES AND SENSITIVITIES

IV. GENERAL HEALTH, ATTITUDE AND HABITS

CONFIDENTIAL

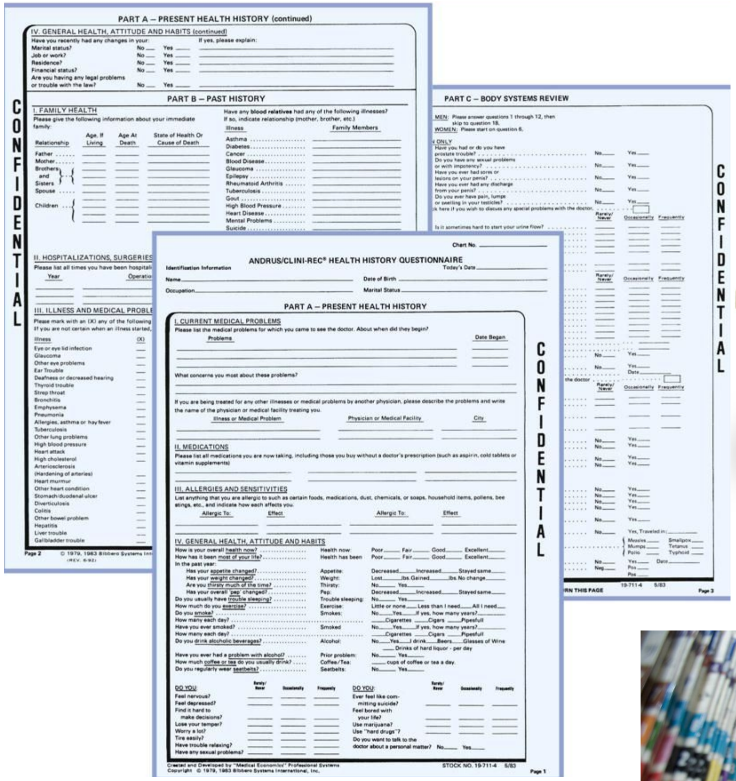
Databases (broadly) are ordered collections of data

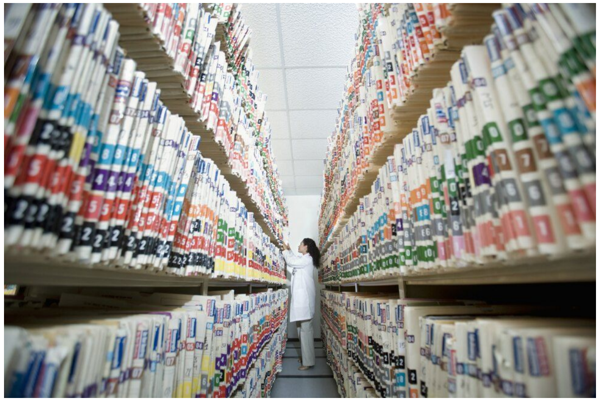
Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet

Ordering:

- Index
- Defined fields
- Standardisation





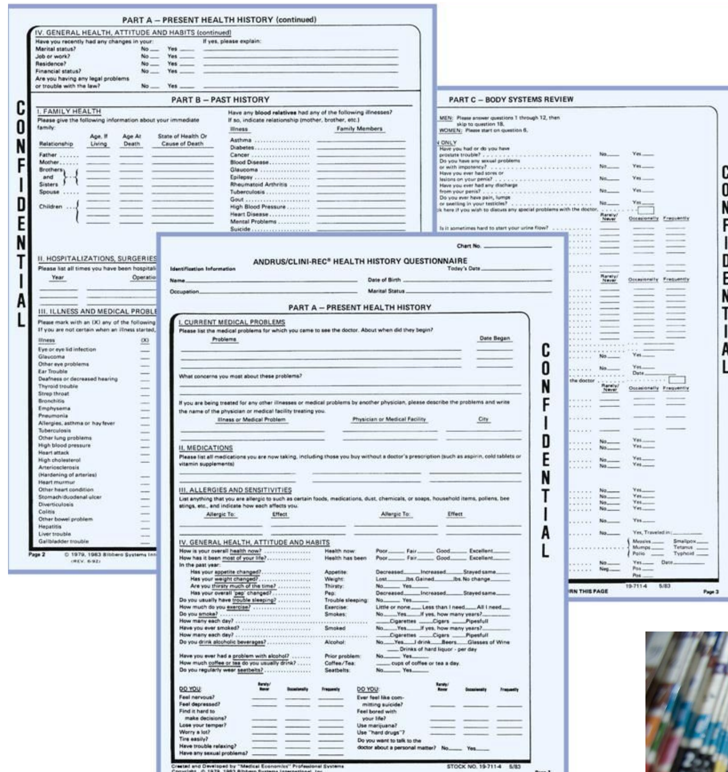
Databases (broadly) are ordered collections of data

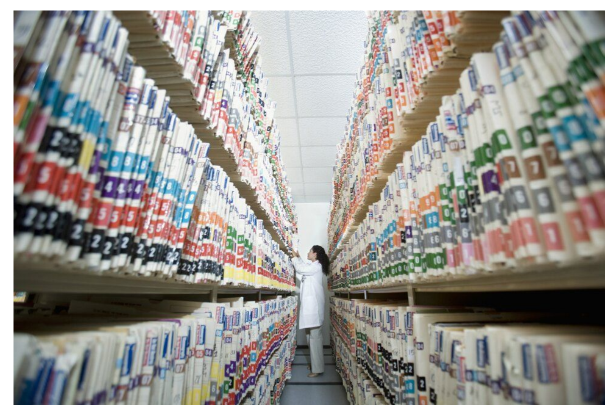
Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet

Ordering:

- Index
- Defined fields
- Standardisation



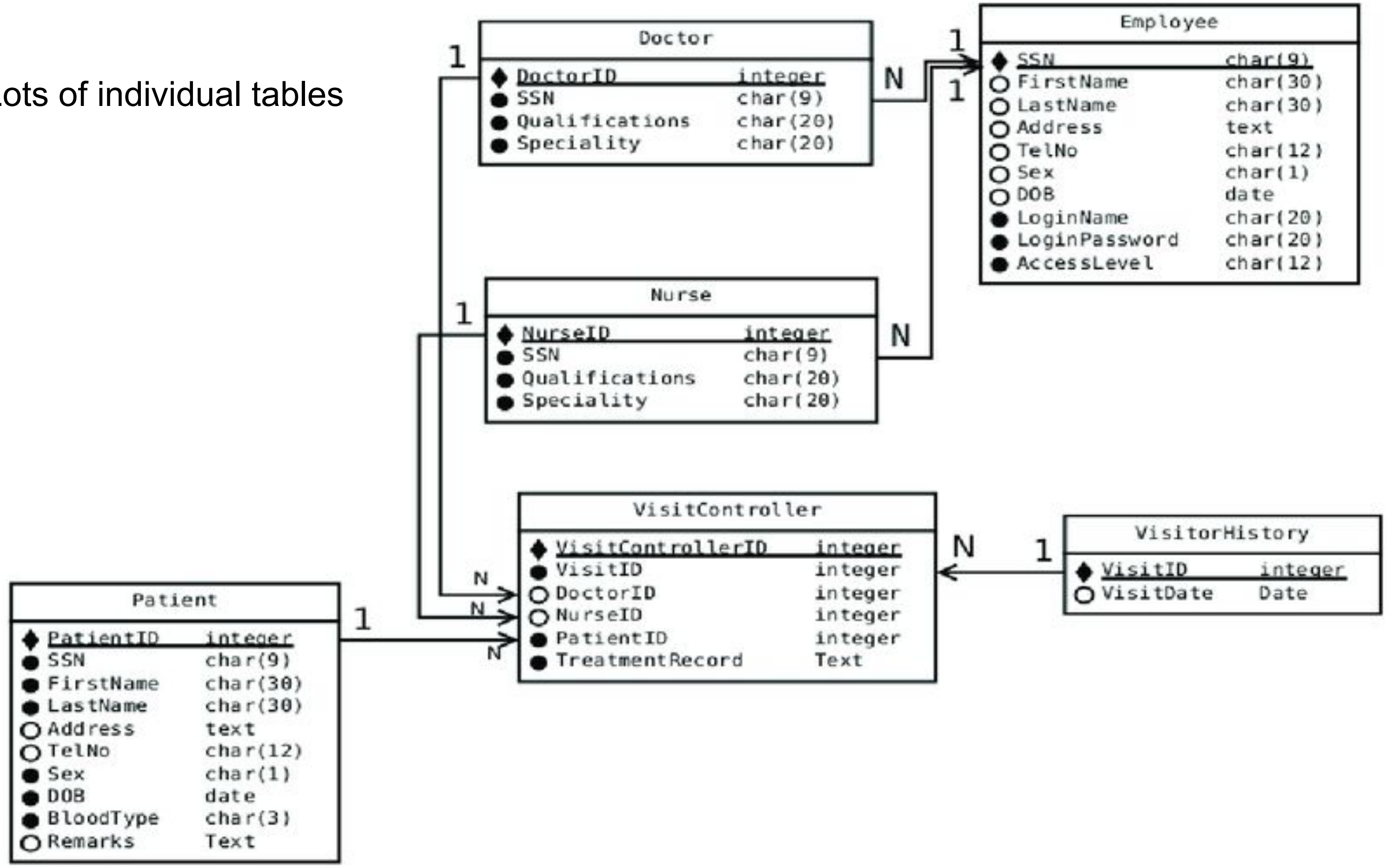


Organisation make some tasks easier/harder:

- 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

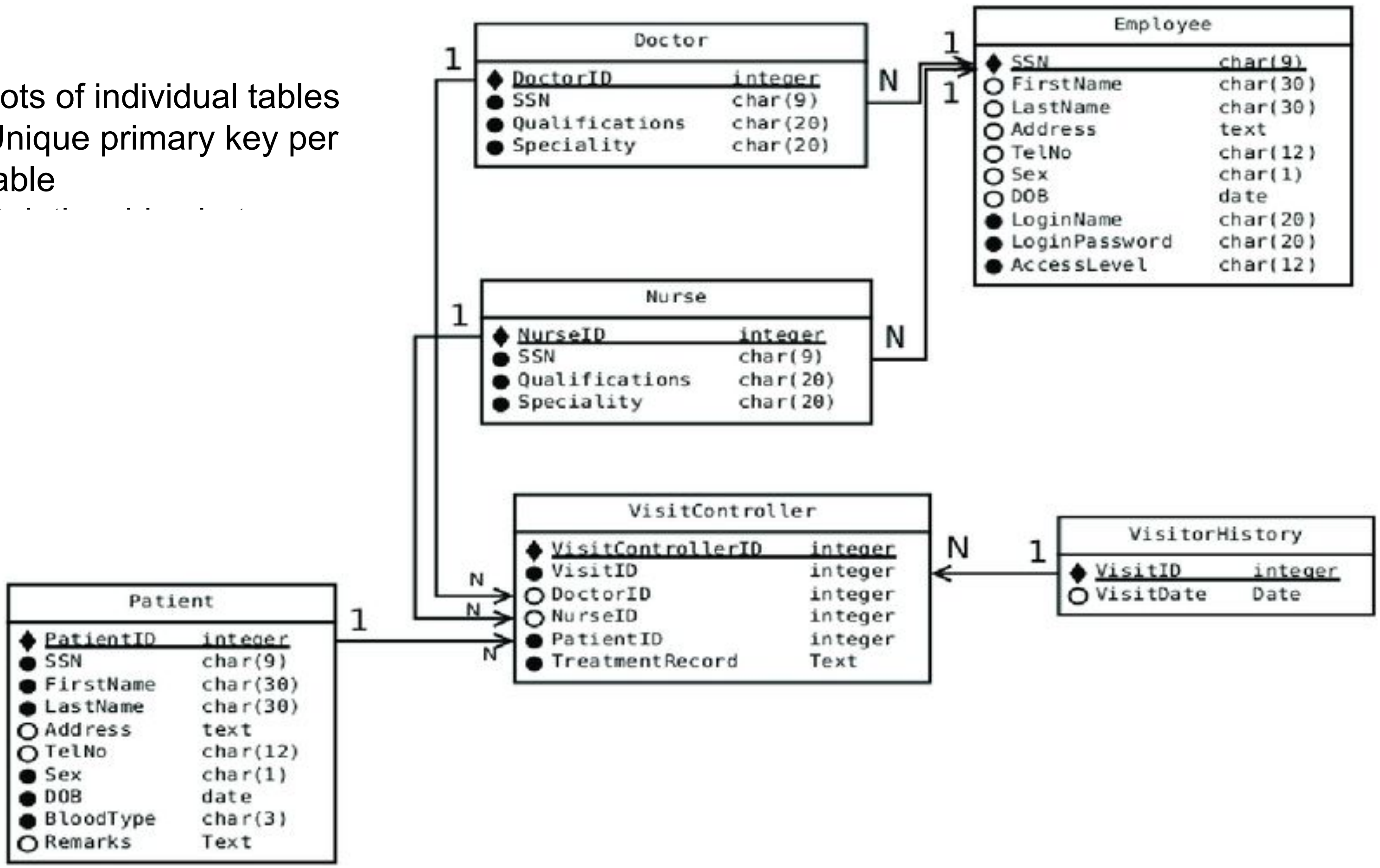
- Lots of individual tables



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

Most Common Type: Relational Databases

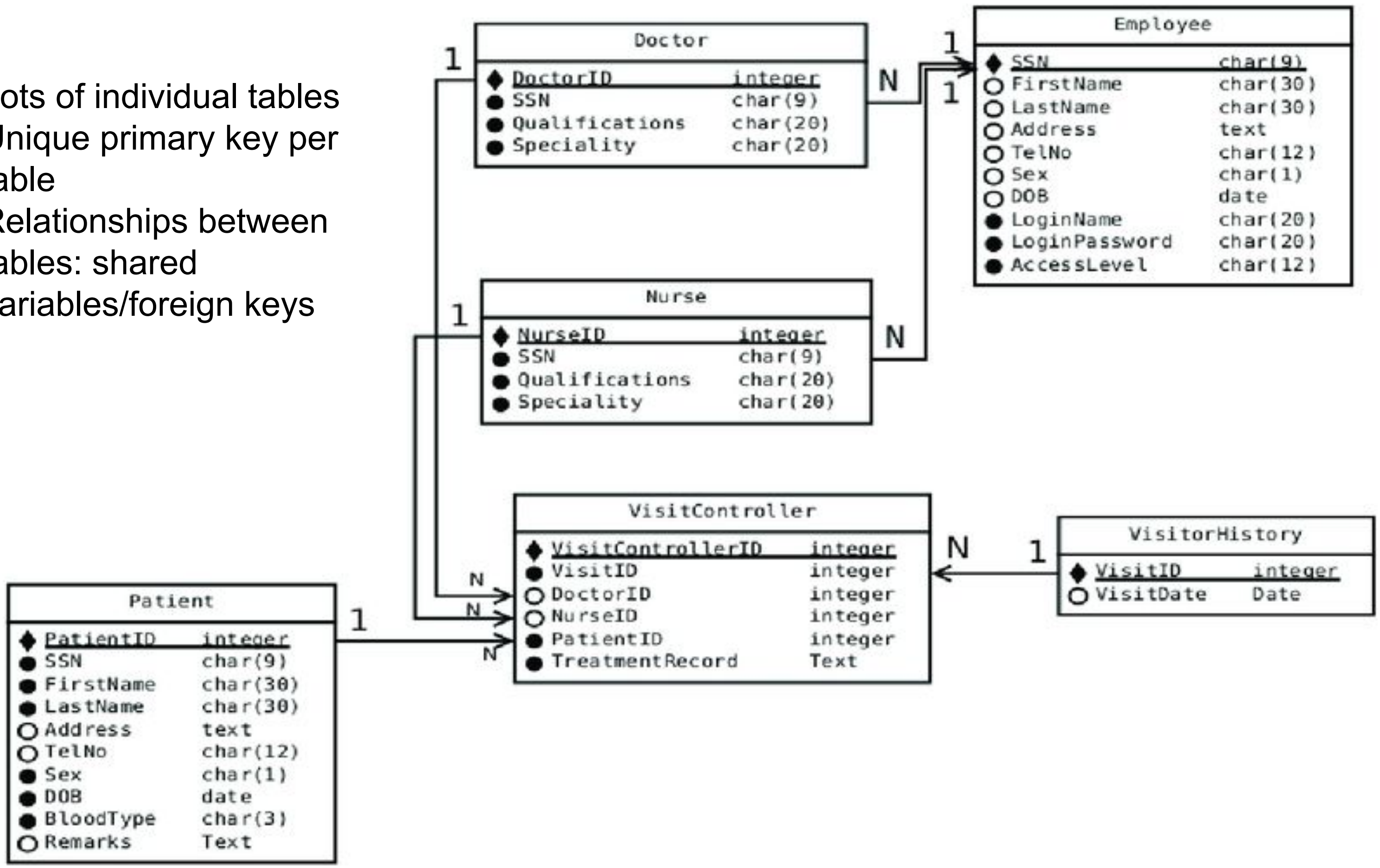
- Lots of individual tables
- Unique primary key per table



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

Most Common Type: Relational Databases

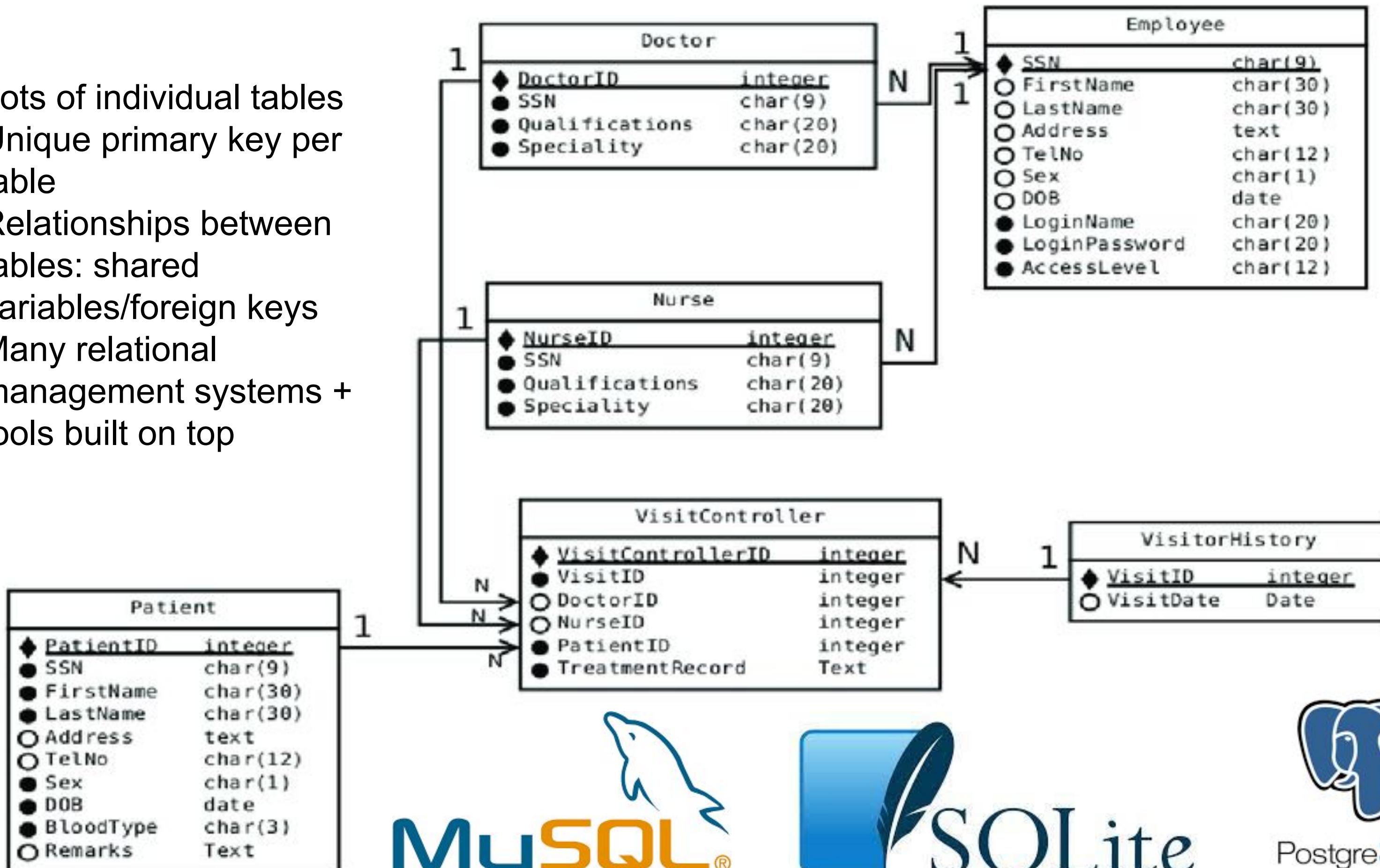
- Lots of individual tables
- Unique primary key per table
- Relationships between tables: shared variables/foreign keys



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

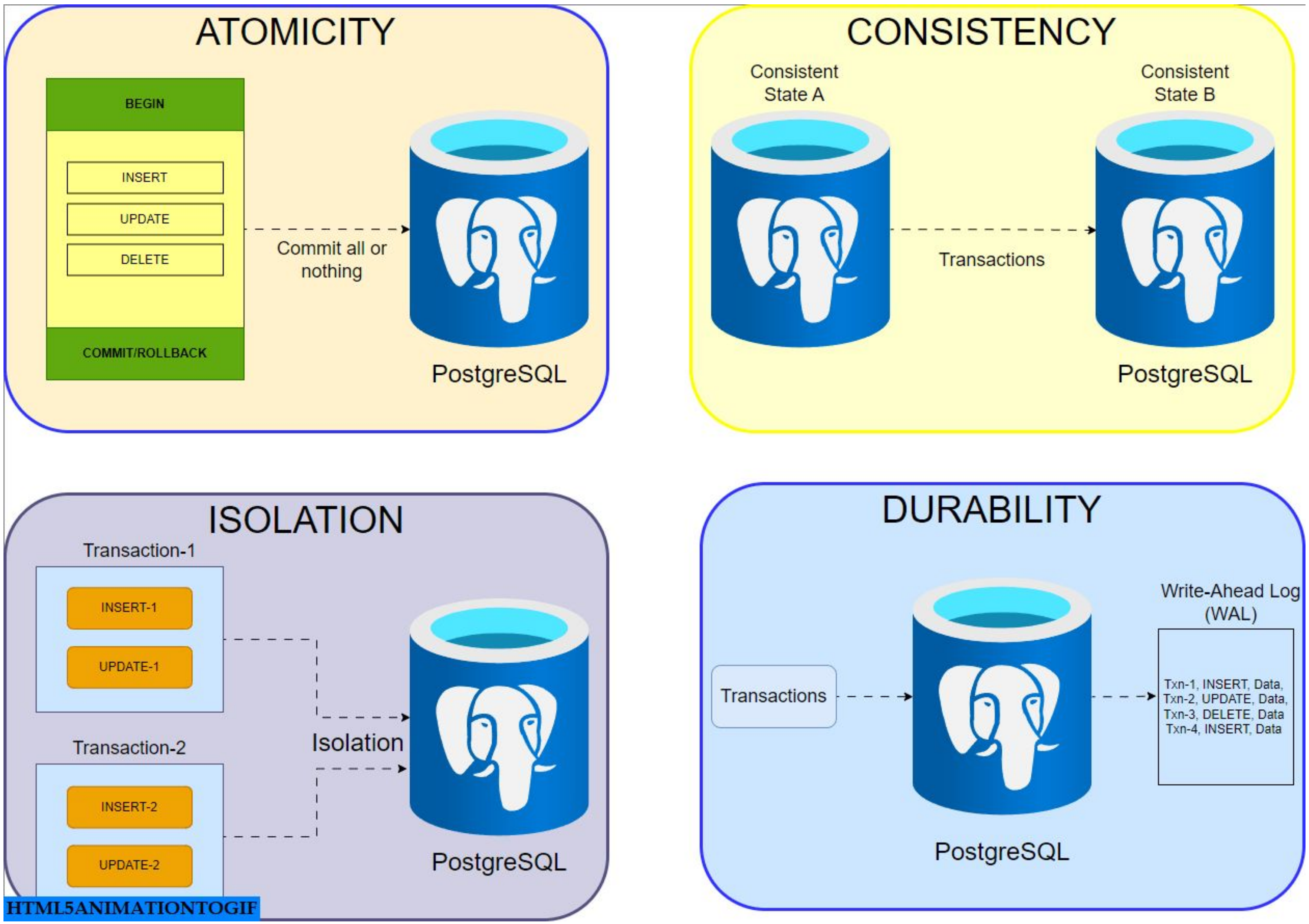
Most Common Type: Relational Databases

- 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

Most relational databases support 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

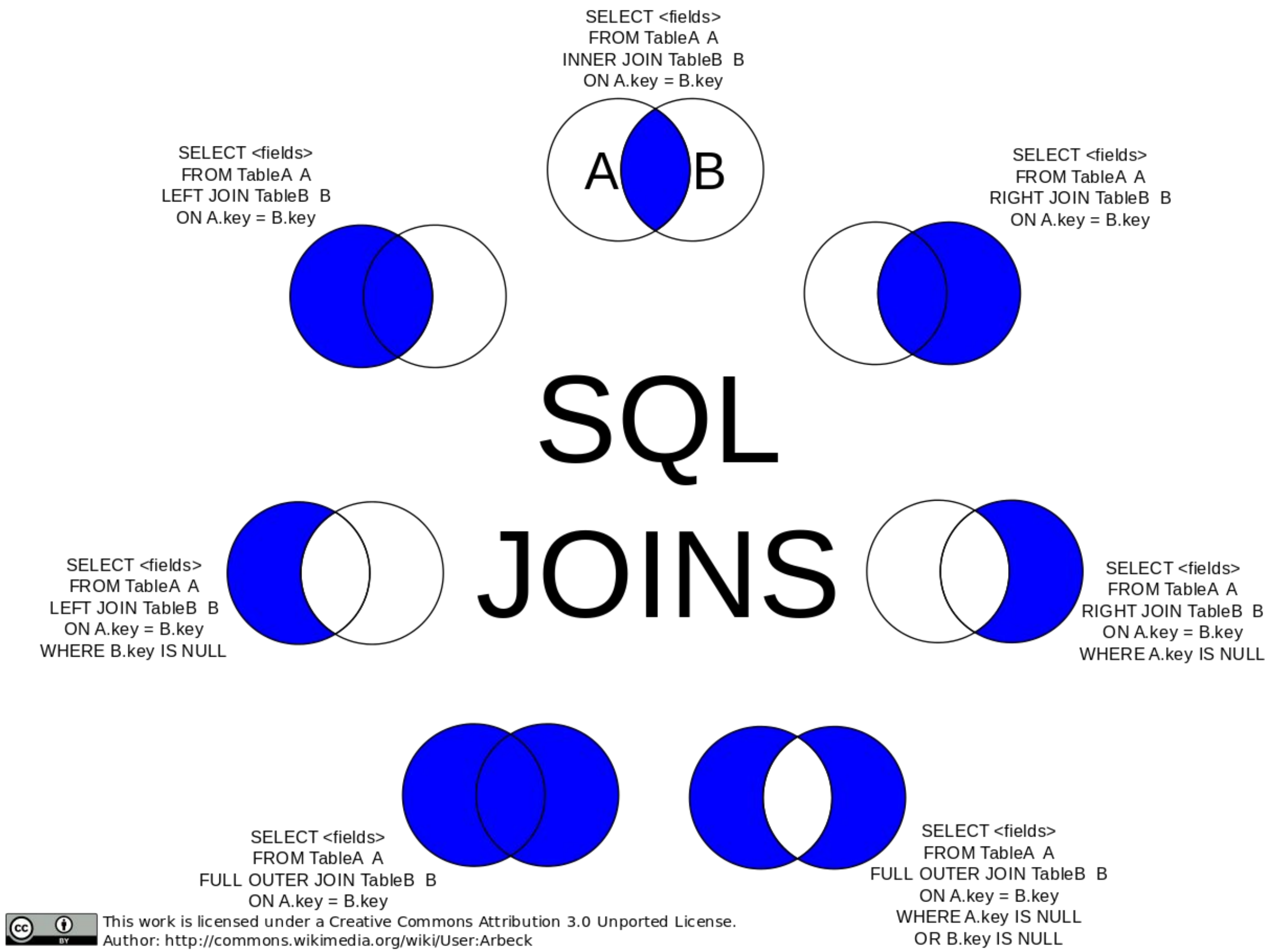
```
flights %>%
  group_by(month, day) %>%
  summarise(delay = mean(dep_delay)) %>%
  show_query()
#> Warning: Missing values are always removed in SQL.
#> Use `AVG(x, na.rm = TRUE)` to silence this warning
#> <SQL>
#> SELECT `month`, `day`, AVG(`dep_delay`) AS `delay`
#> FROM `nycflights13::flights`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

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

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

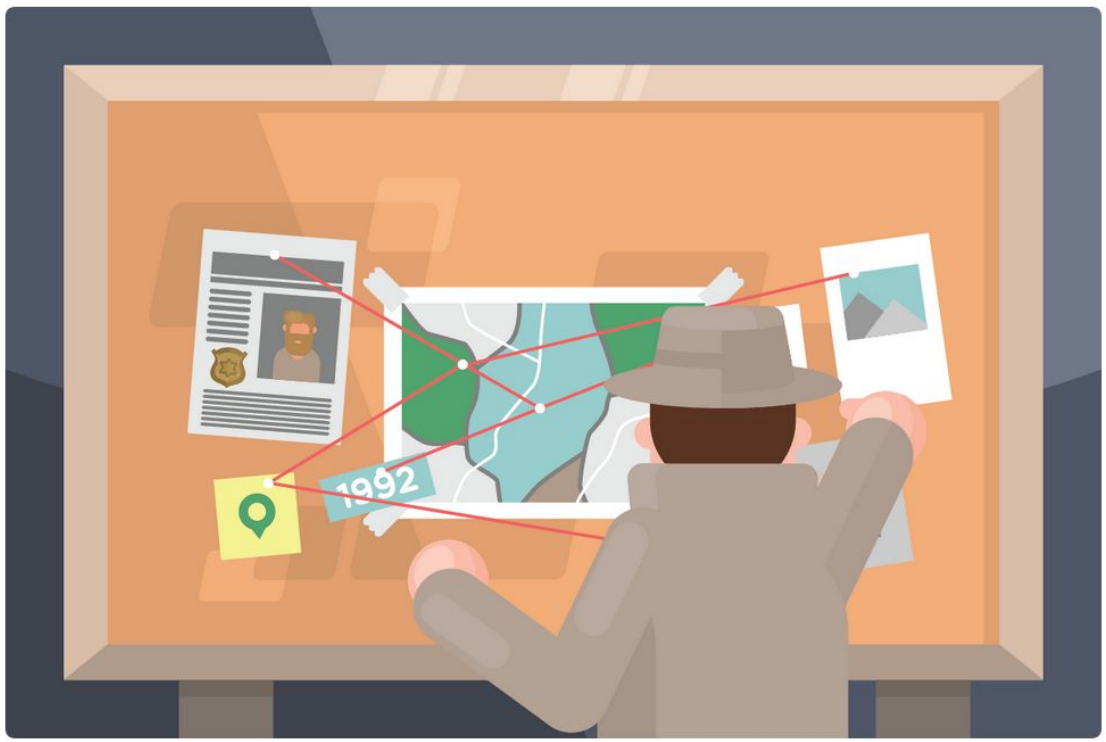


Fun way to learn basic SQL

https://mystery.knightlab.com/

SQL Murder Mystery

Can you find out whodunnit?

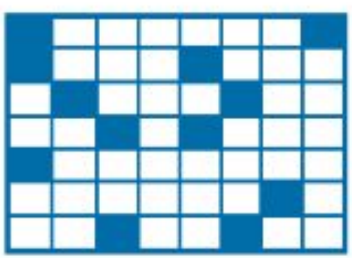


Are all databases relational?

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



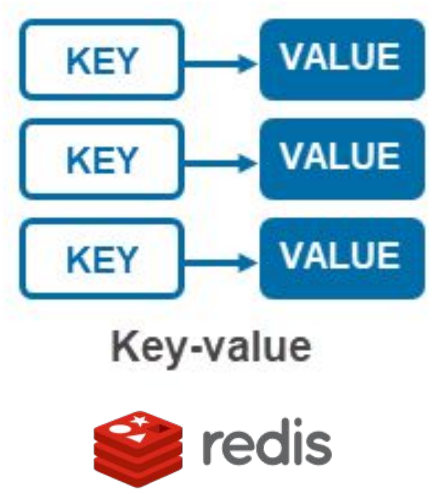
Column based

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

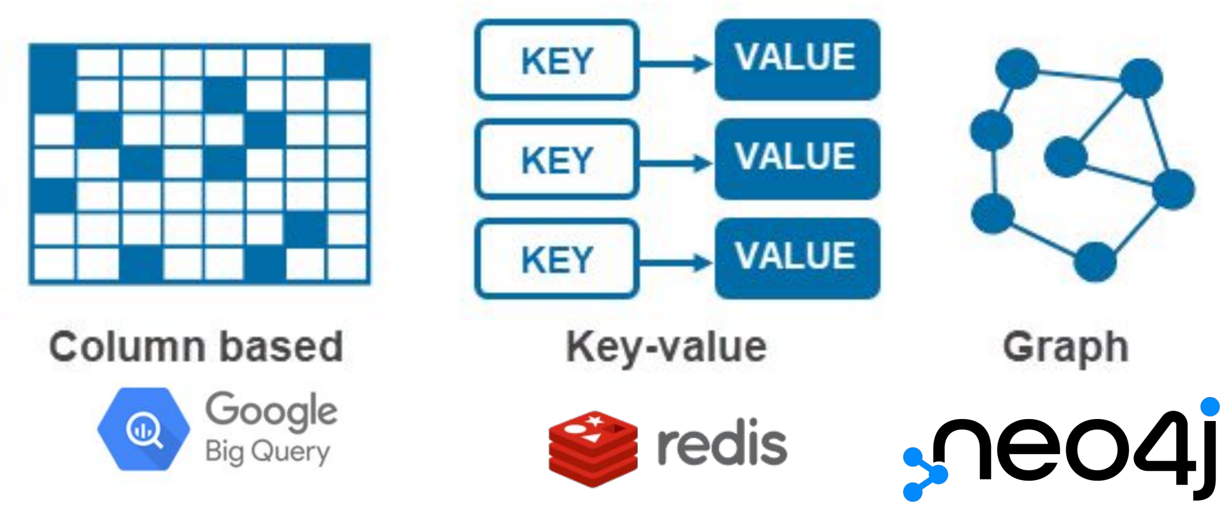
https://phoenixnap.com/kb/database-types



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

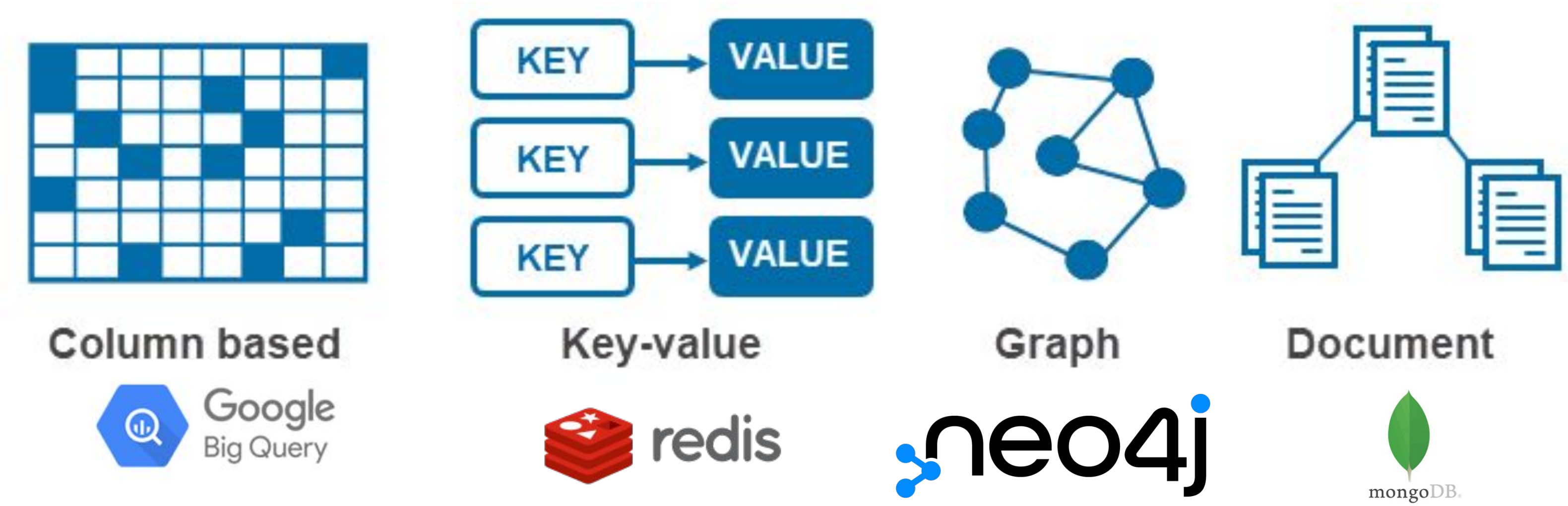
https://phoenixnap.com/kb/database-types



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

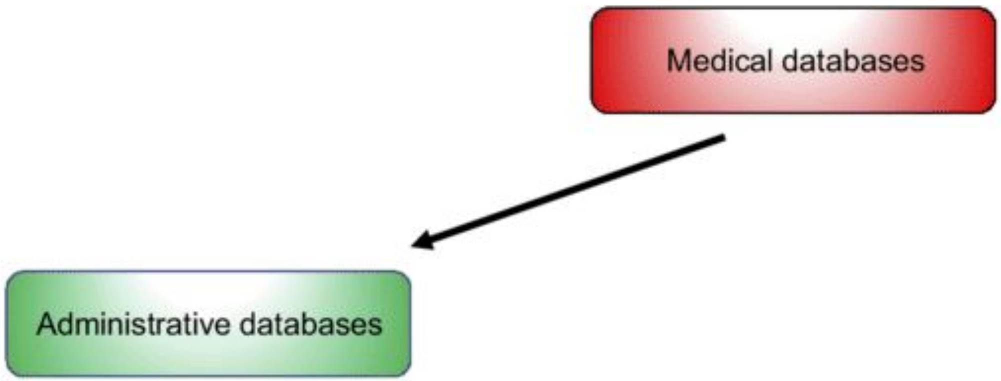


What are medical databases?

Many types of database

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

Many types of database

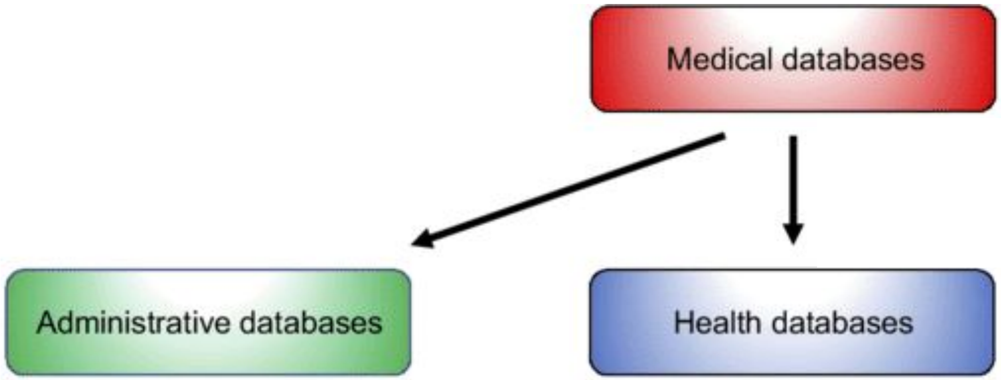


All types of registries and databases that contain health-related data

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

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

Many types of database



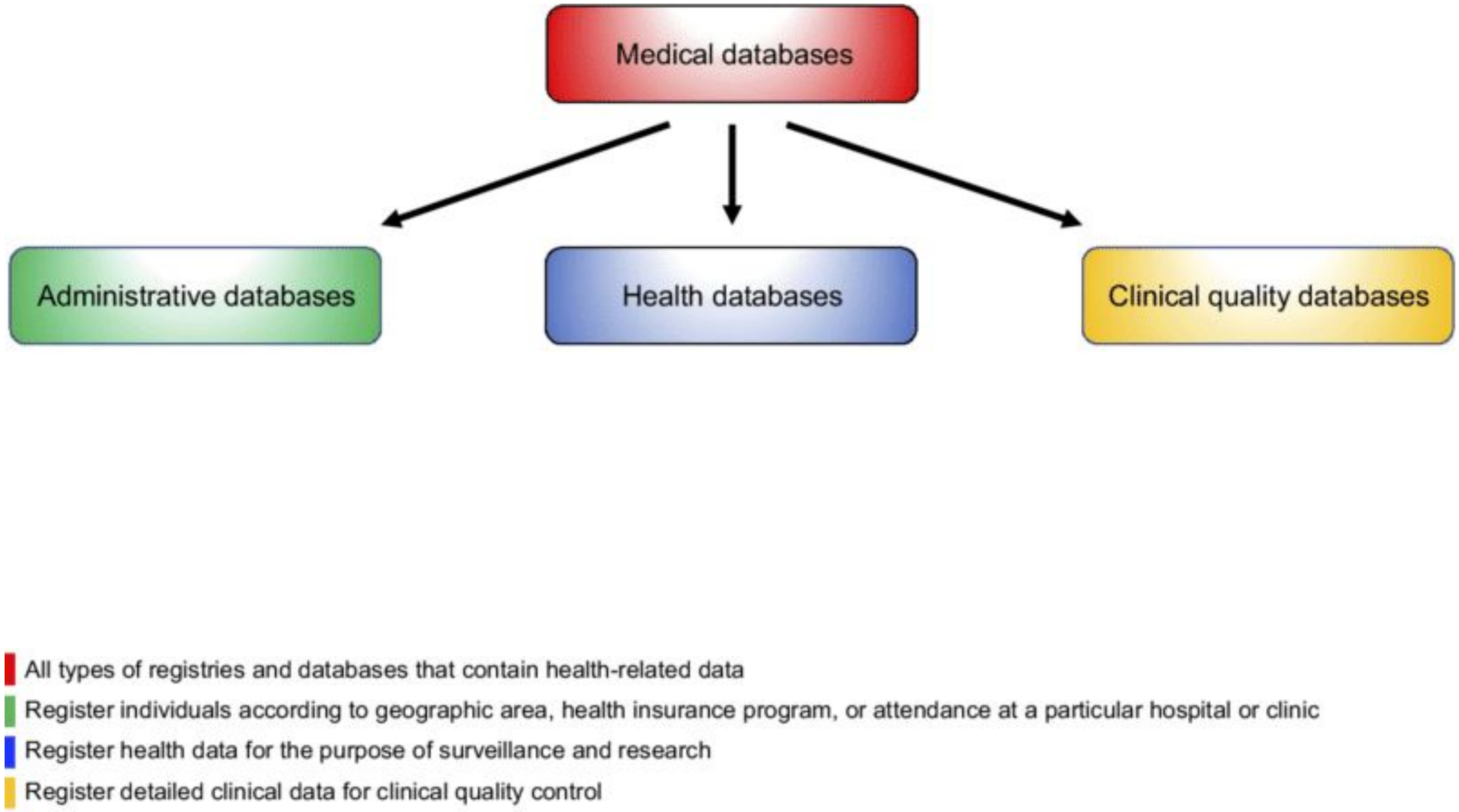
All types of registries and databases that contain health-related data

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

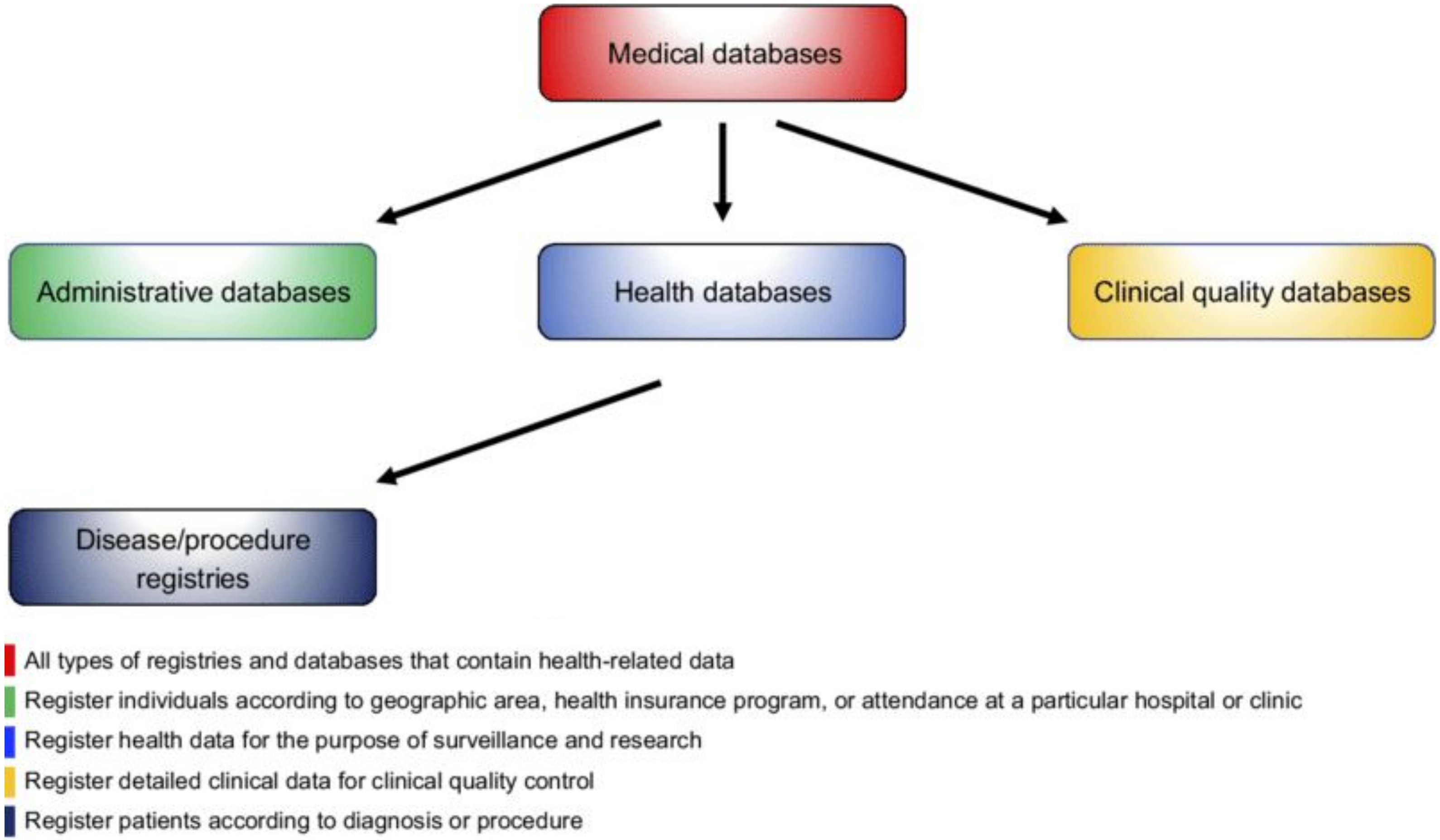
https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



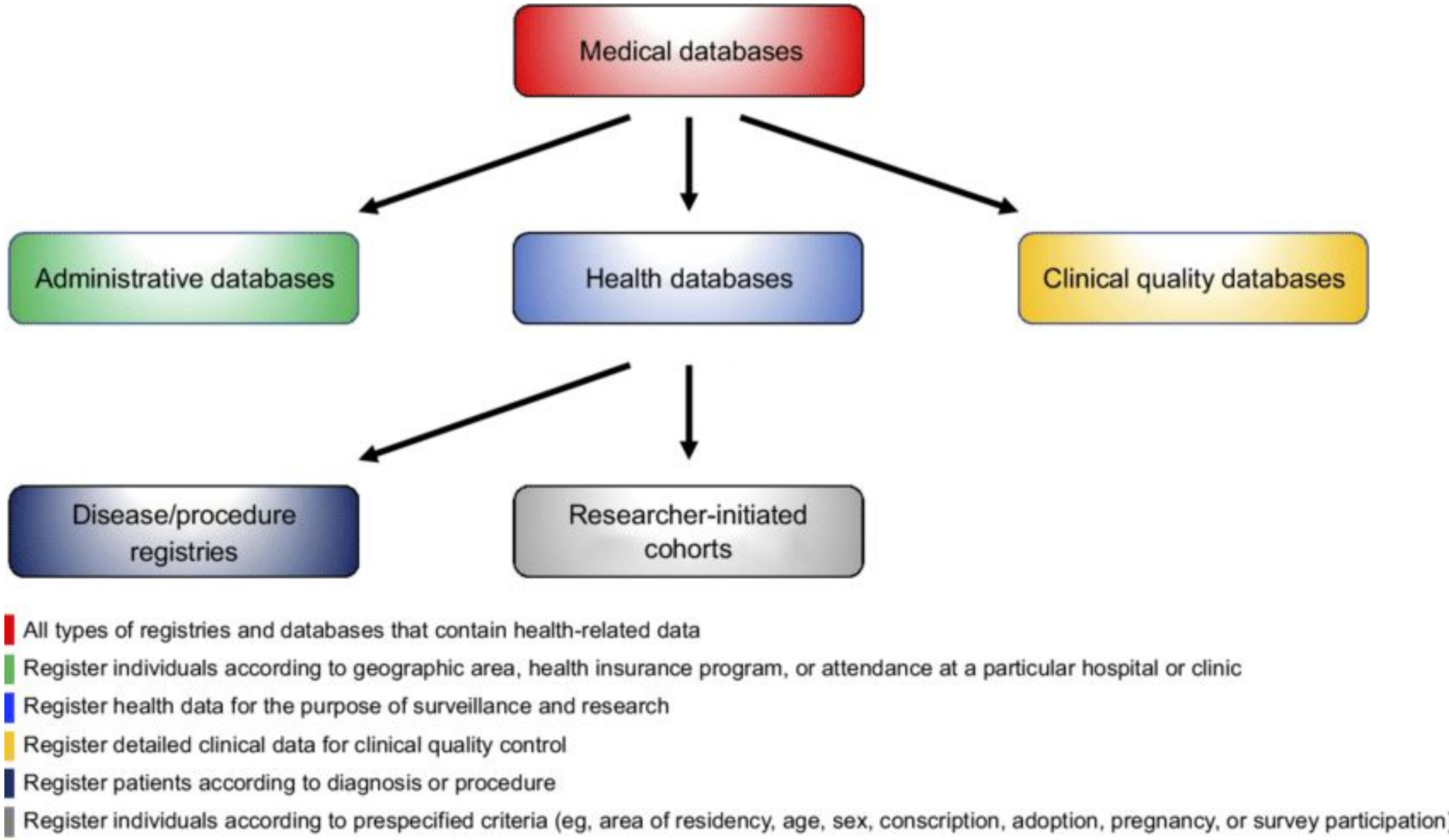
https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



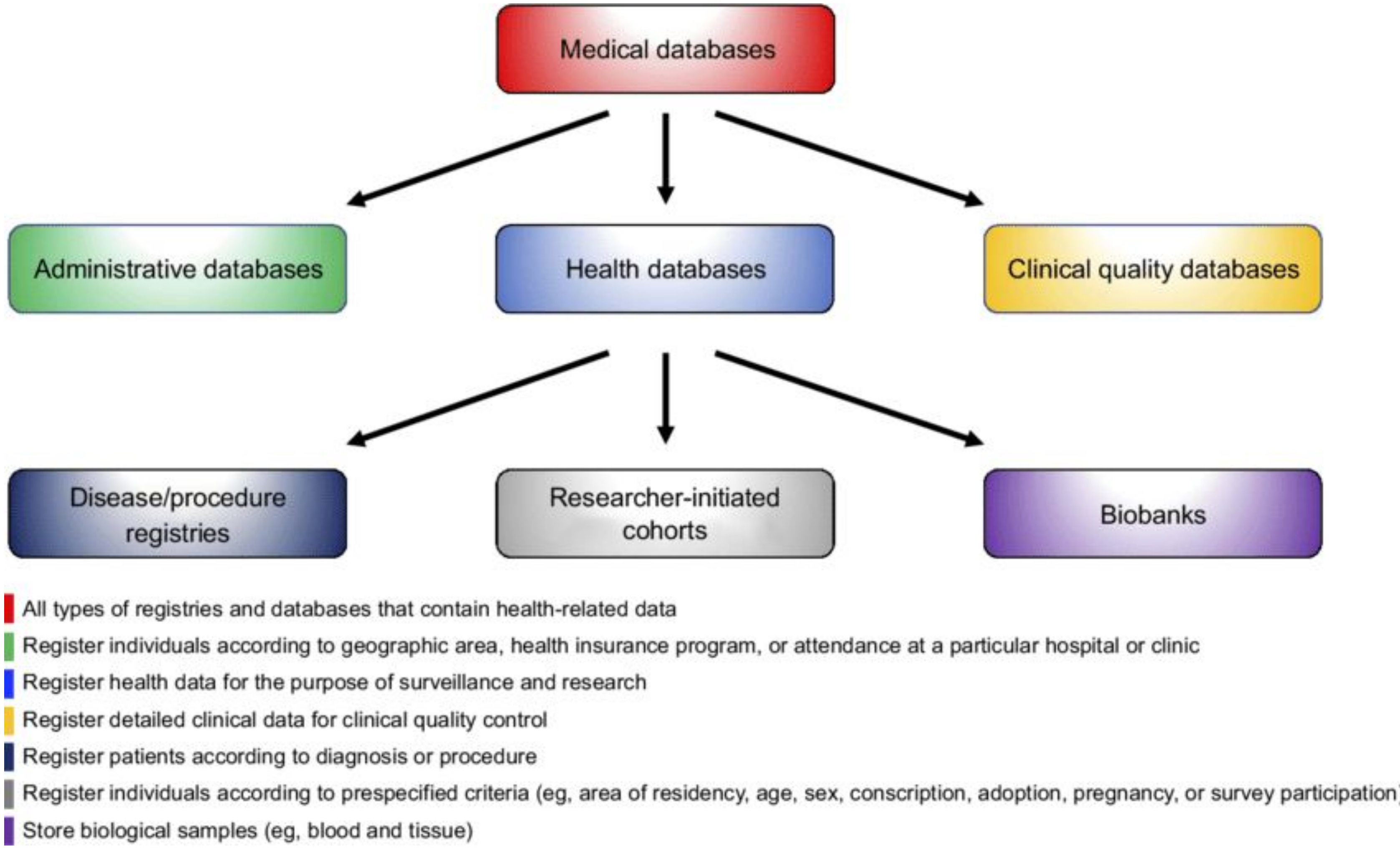
https://www.researchgate.net/figure/Classification-of-medical-databases_fig1_334421419

Many types of database



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

Consider 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

Generalisability

Sampling scope

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

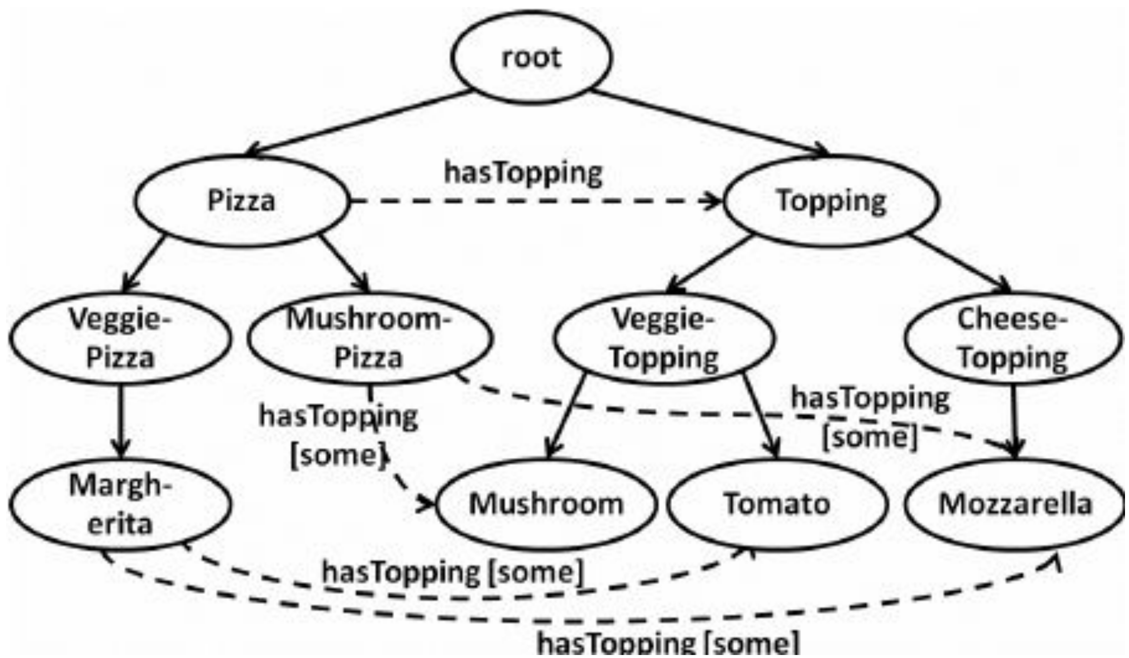
Challenge of standardisation

Generalisability

How do medical databases try to handle standardisation?

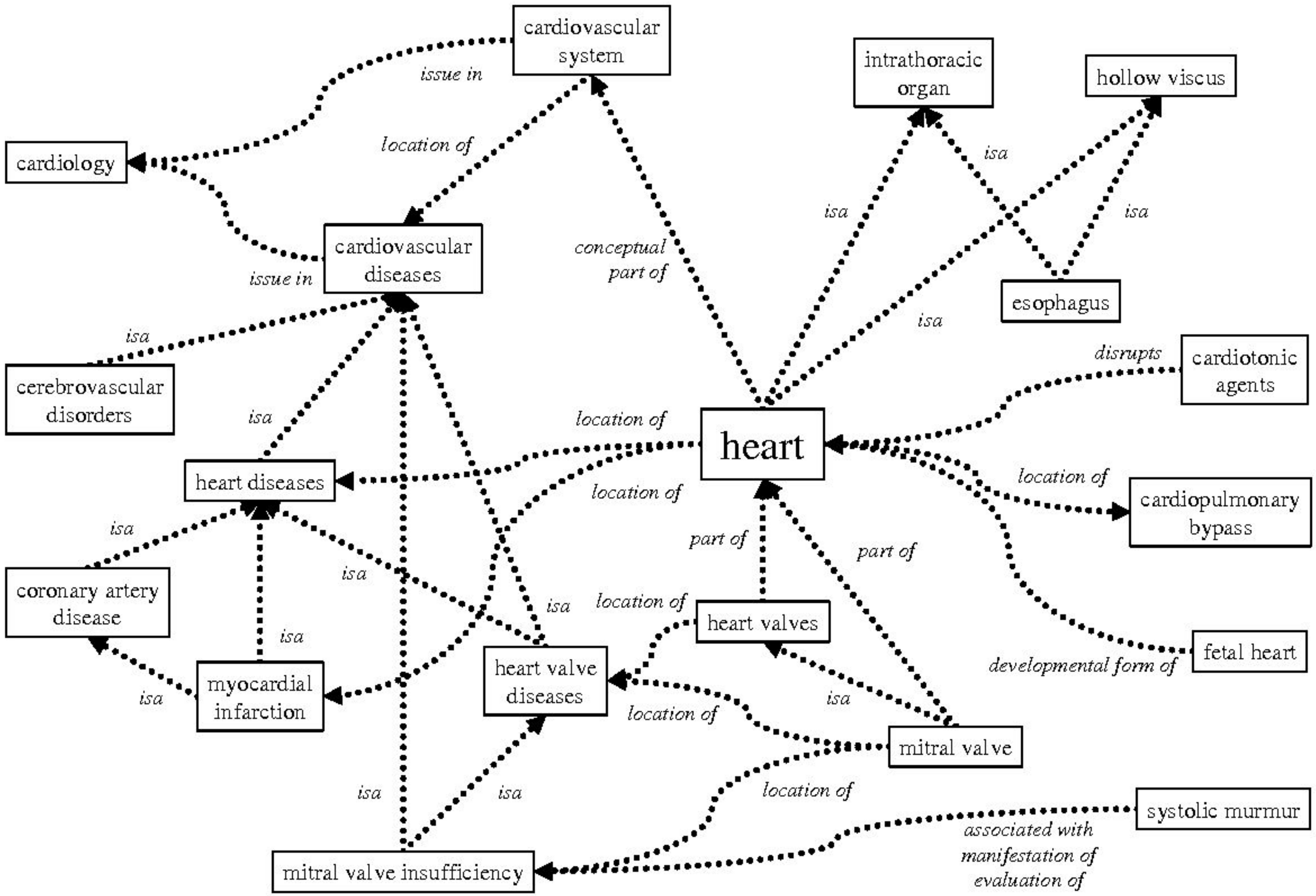
Ontologies for standardisation

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



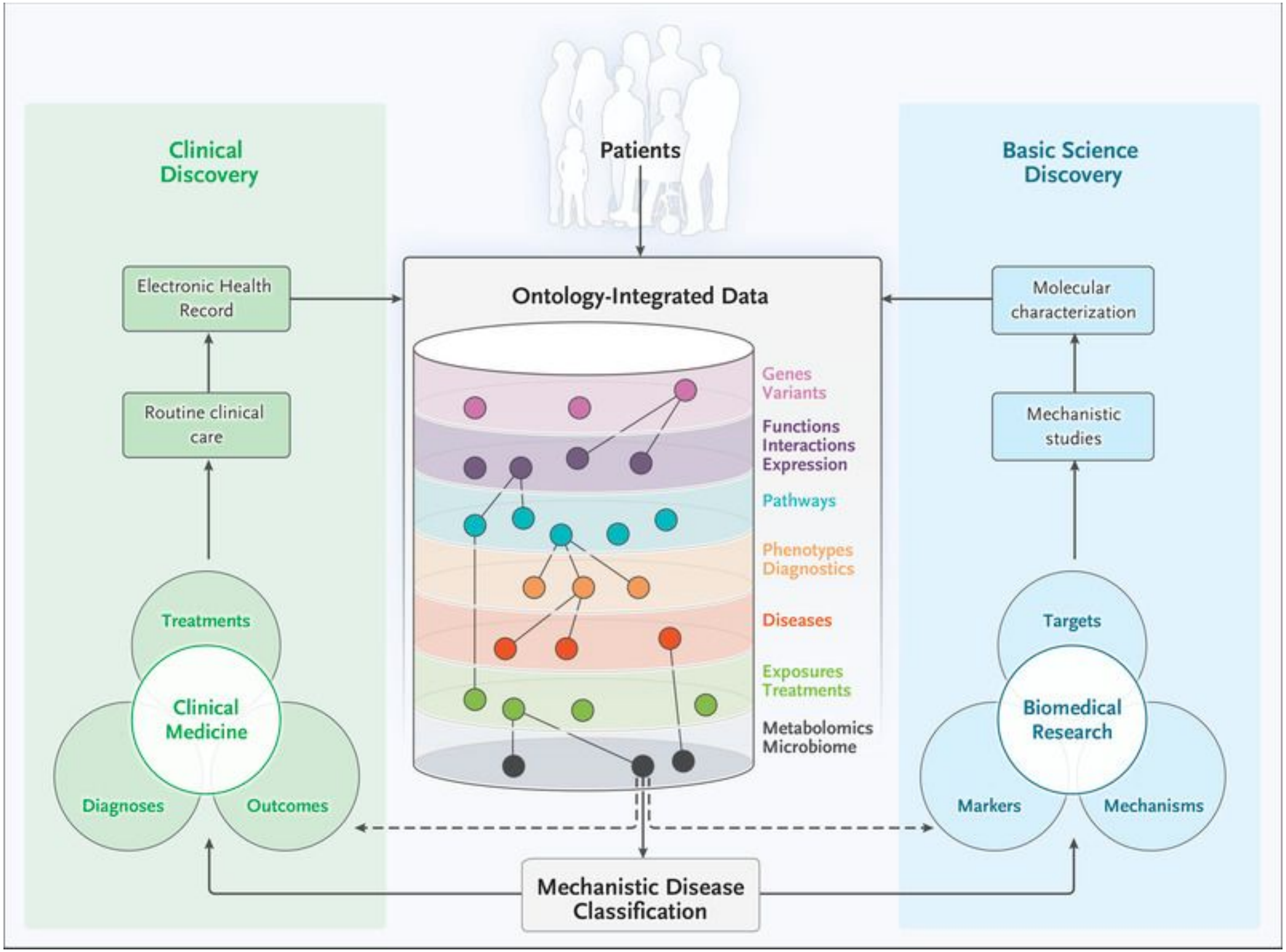
https://www.researchgate.net/figure/Example-pizza-ontology-represented-as-a-graph-G-a-and-a-changed-version-of-the-pizza_fig1_236842047

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?

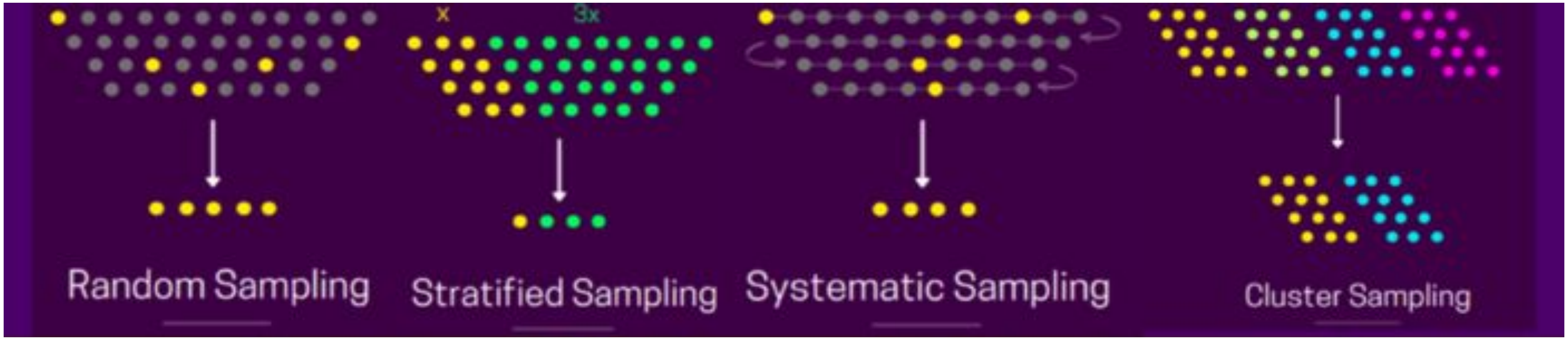
Sampling strategy

- Exhaustive in a database isn’t always 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

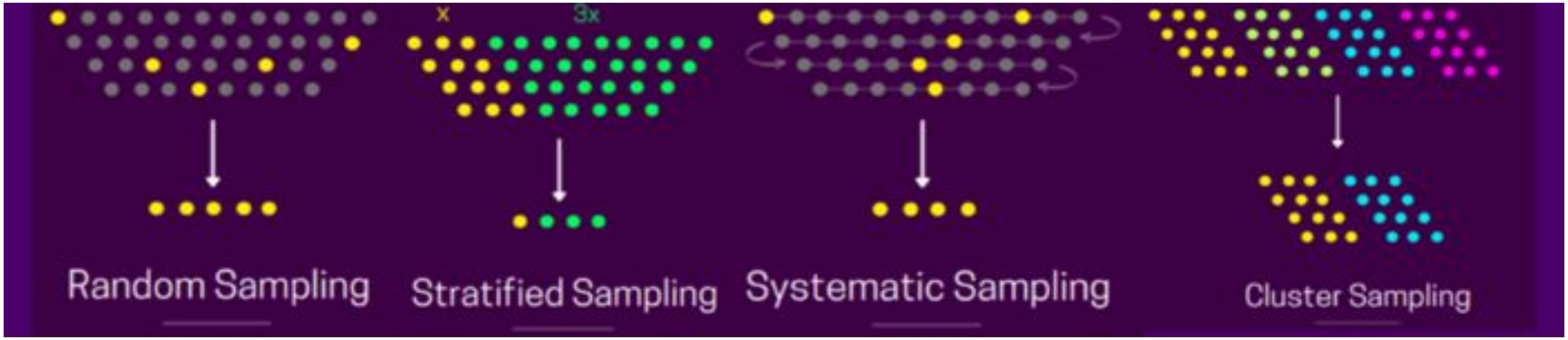
- Exhaustive in a database isn't always exhaustive in true population
- Numerous and often quite complex!
- Major source of bias so always carefully explore



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

Sampling strategy

- Exhaustive in a database isn’t always 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

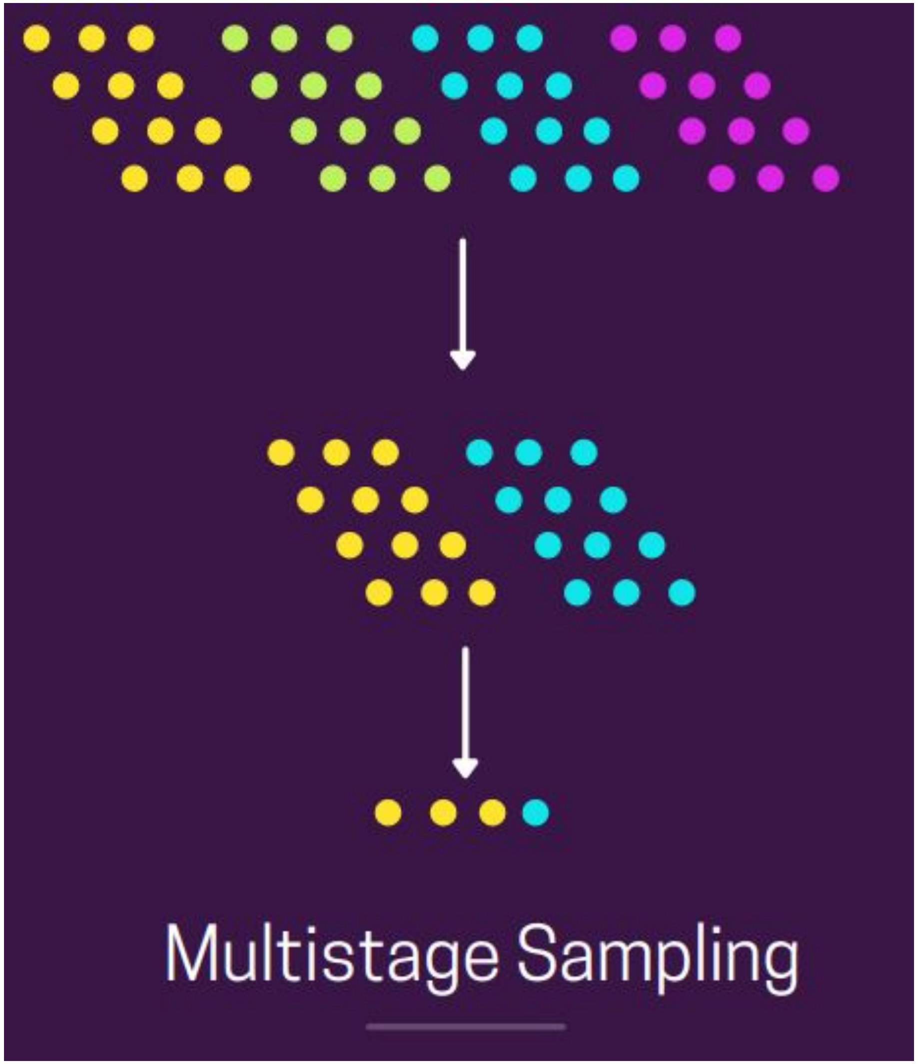
Survey/Sample weights

- 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)

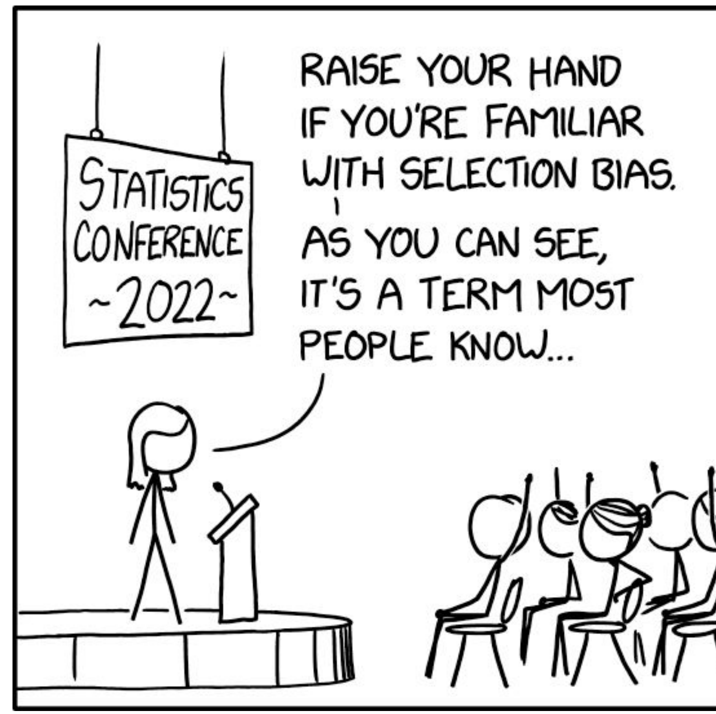


Types of weights

- 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

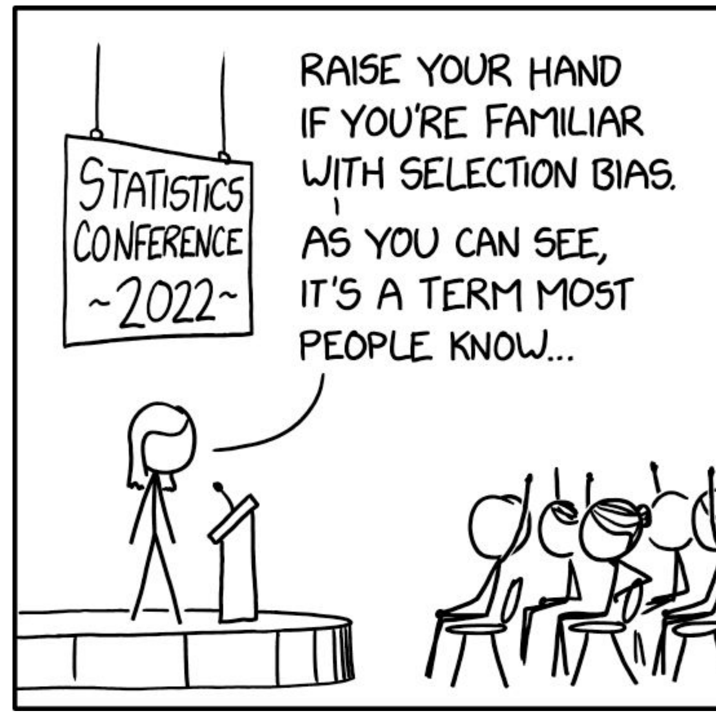


- Post-stratification / Non-response weights
  - Based on collected data
  - Typically biases in whose data is collected
  - Over-represented groups need to be under-weighted

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

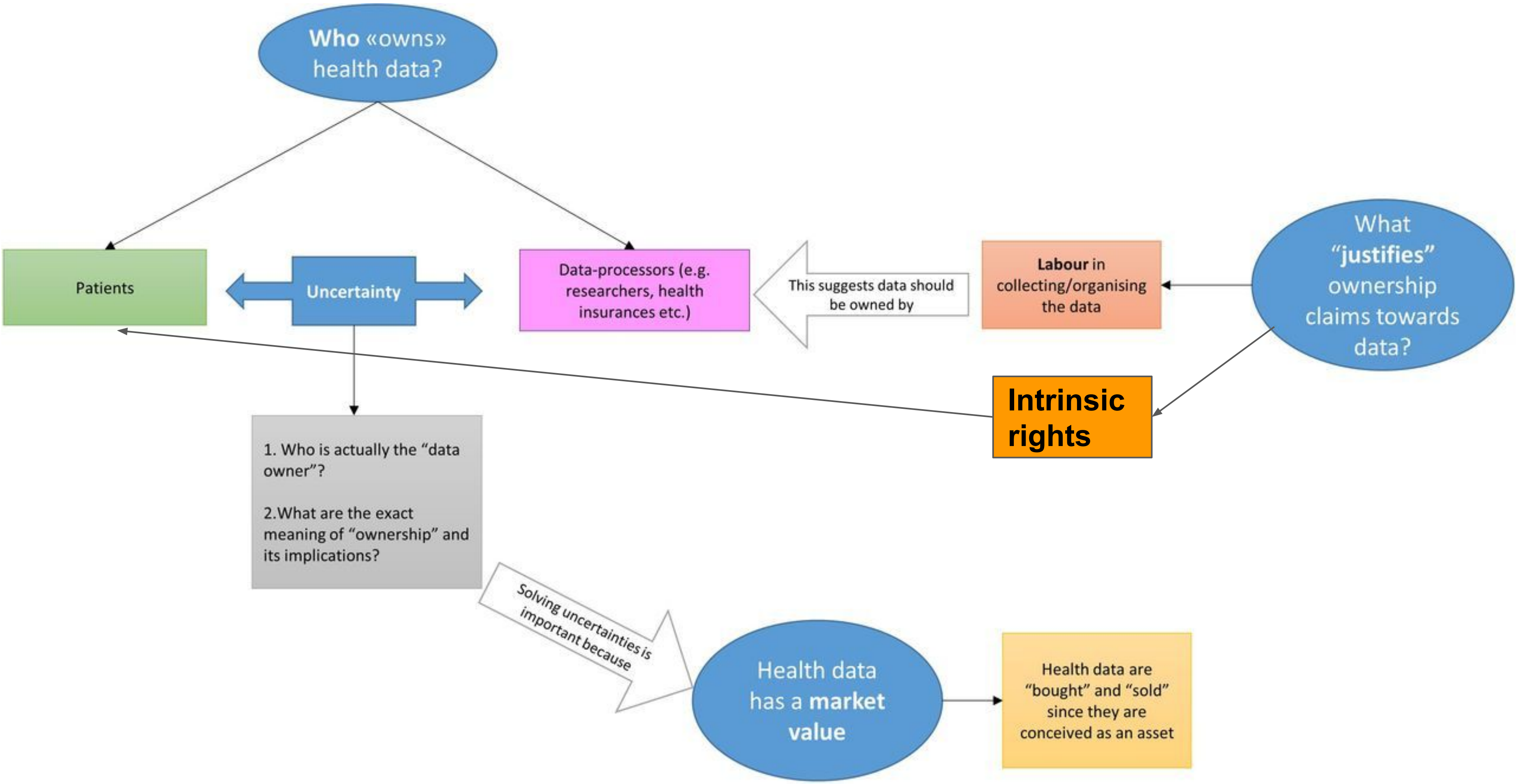
- Post-stratification / Non-response weights
  - Based on collected data
  - Typically biases in whose data is collected
  - Over-represented groups need to be under-weighted



- Often many different weights are combined:

Who actually owns this data?

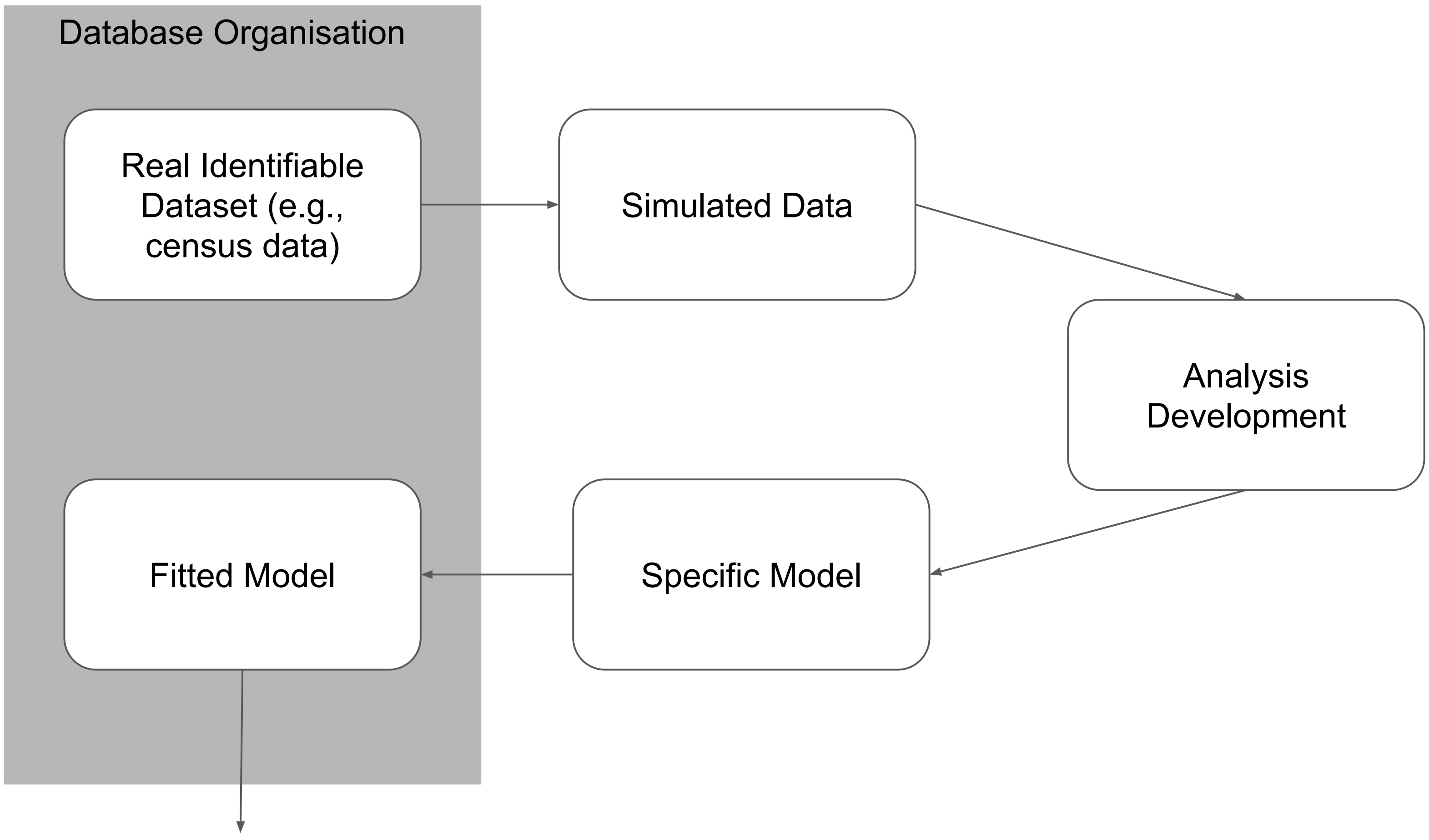
Data Ownership is Difficult



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

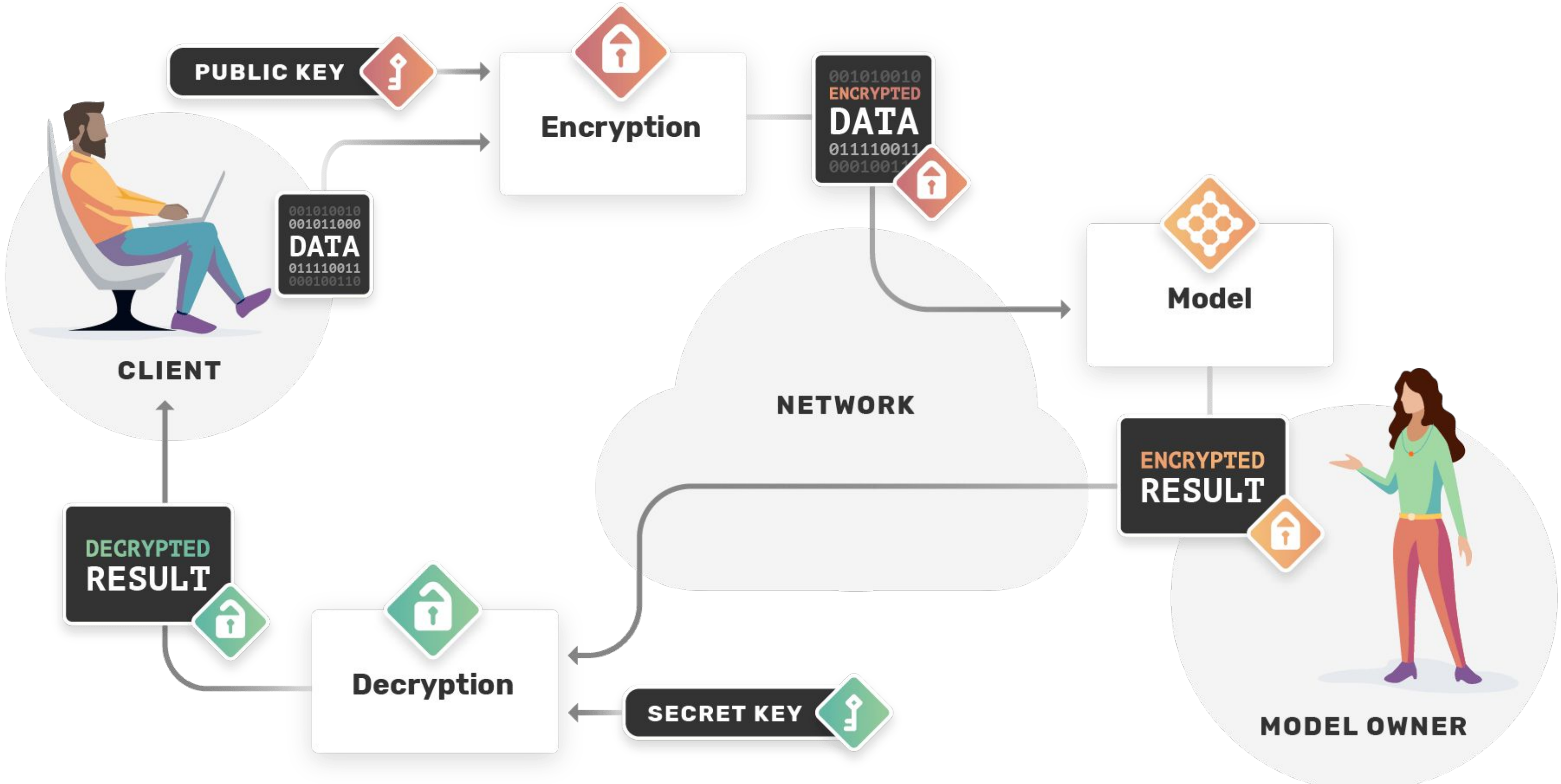
How do you protect privacy in these databases?

No direct data access



Shared data but encrypted: homomorphic encryption

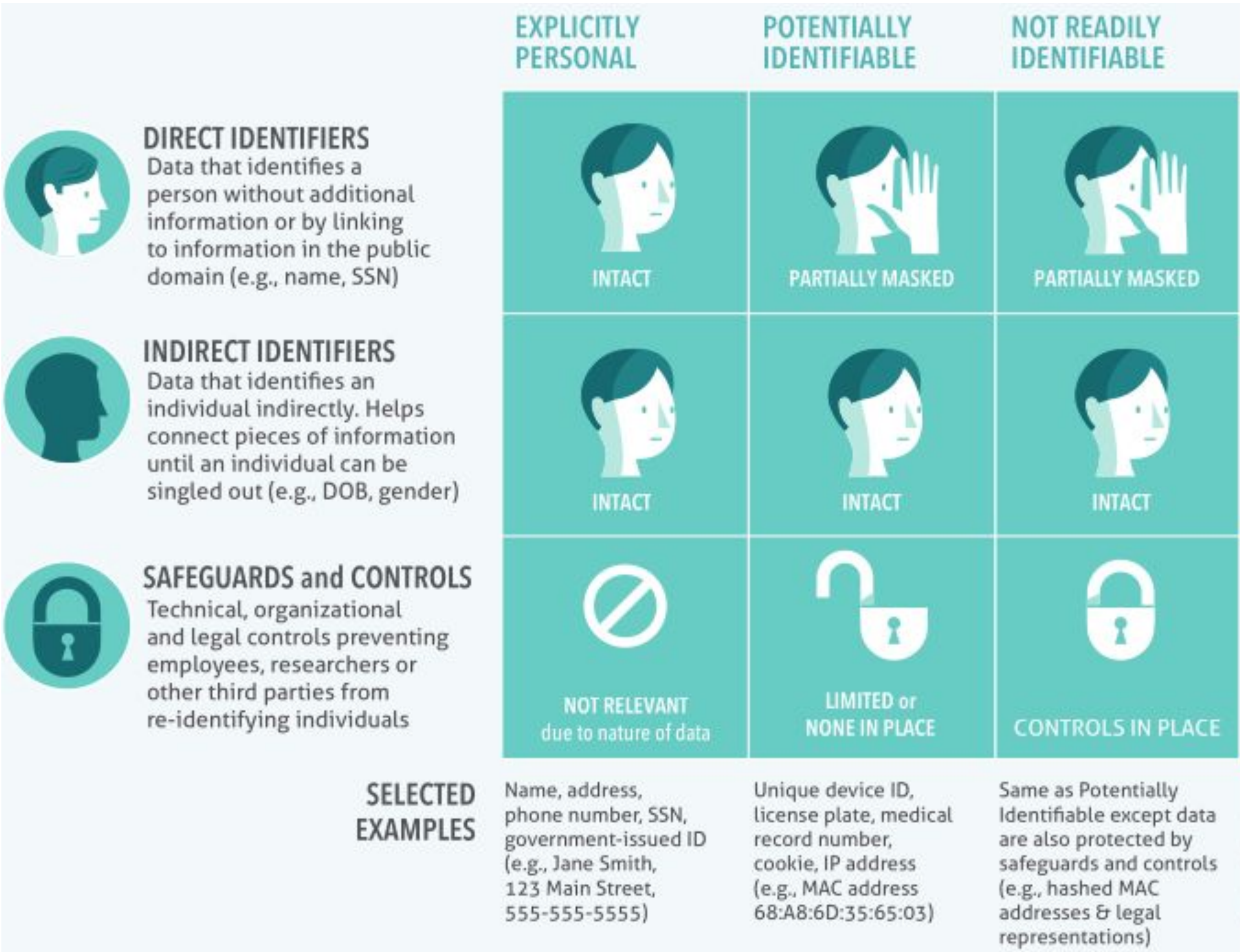
Partial to fully homomorphic encryption



https://research.aimultiple.com/homomorphic-encryption/

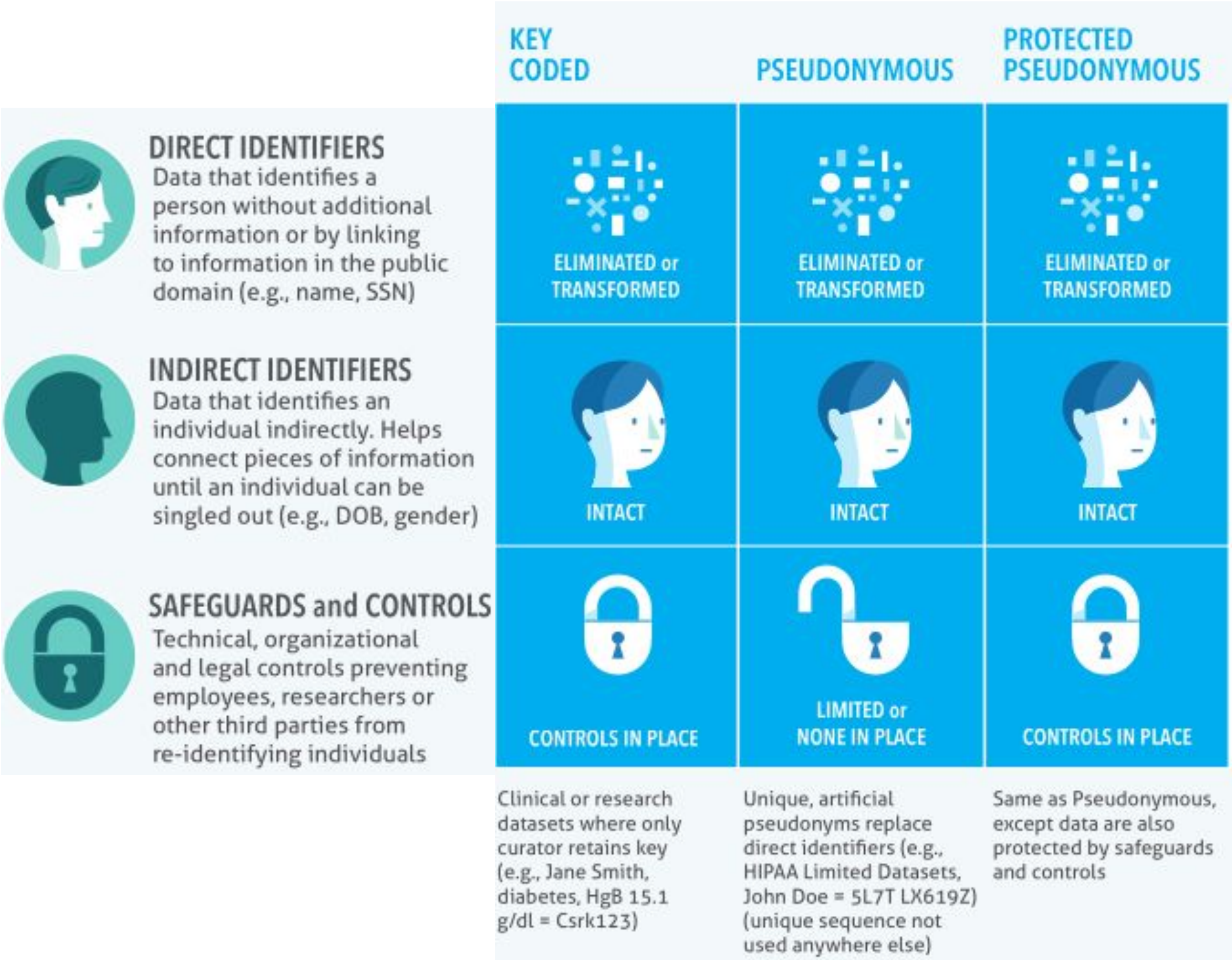
Both are difficult and limited… so how can we share data directly but safely?

Data privacy is a continuum



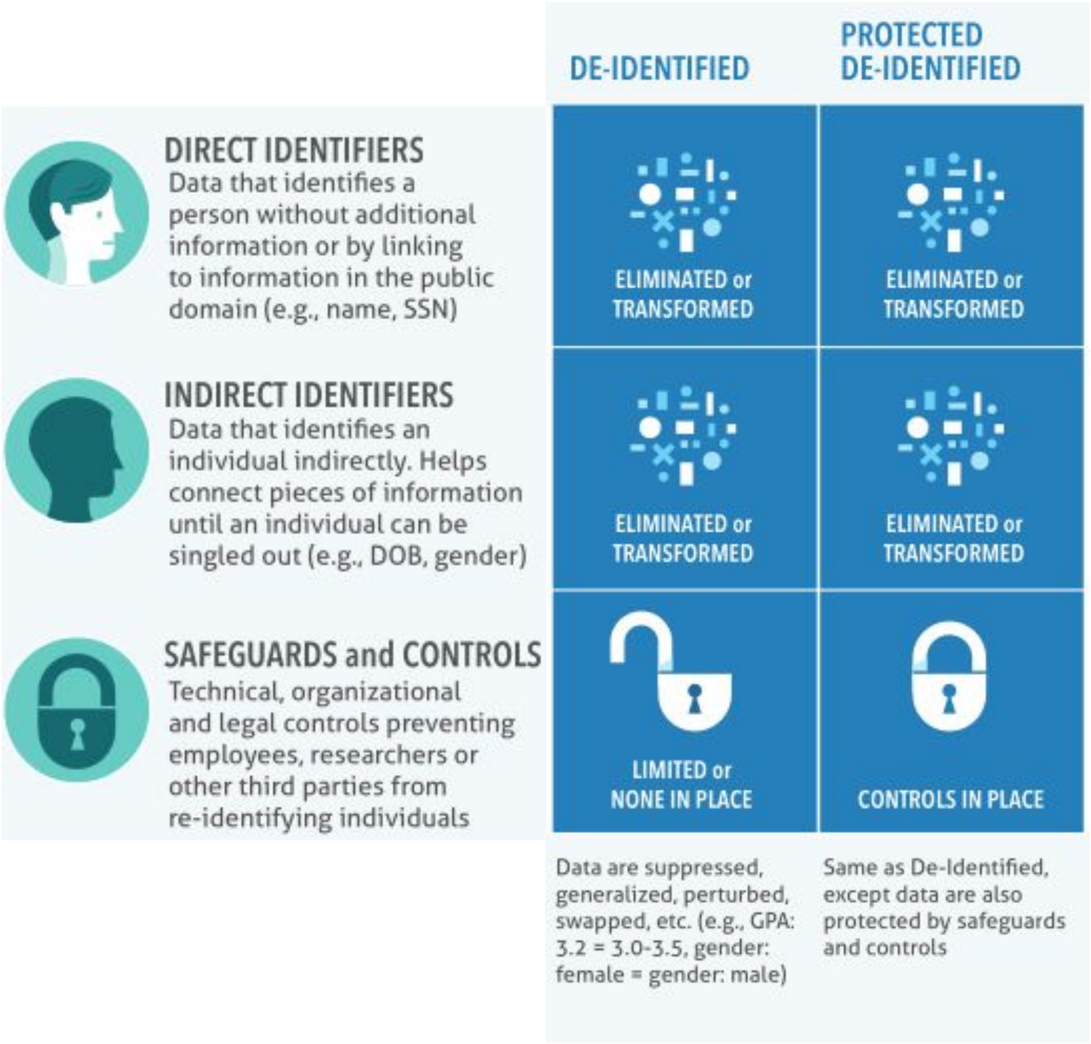
https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf

Indirectly identifiable: Pseudonymous Data



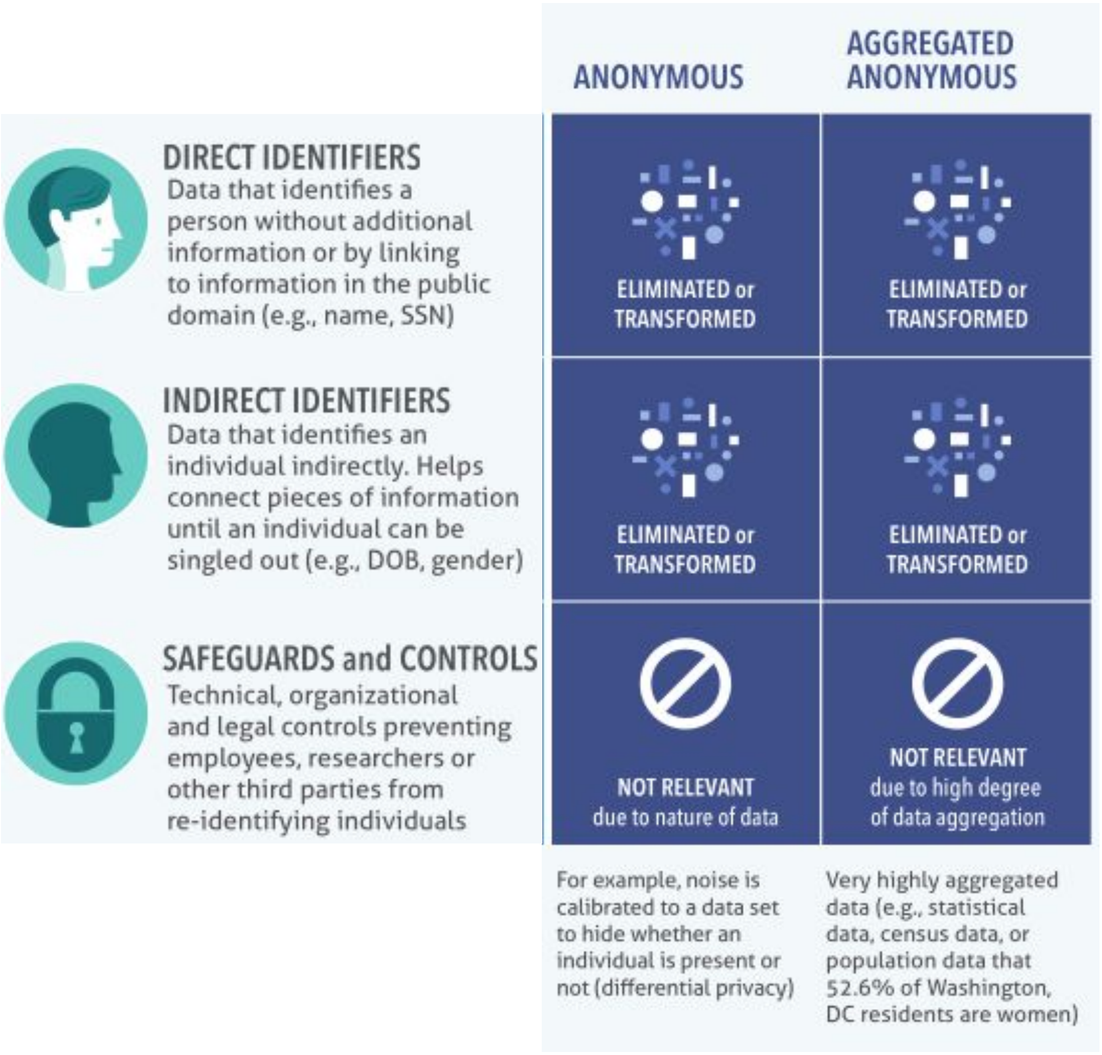
https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf

Identifiers removed/broken: De-Identified Data



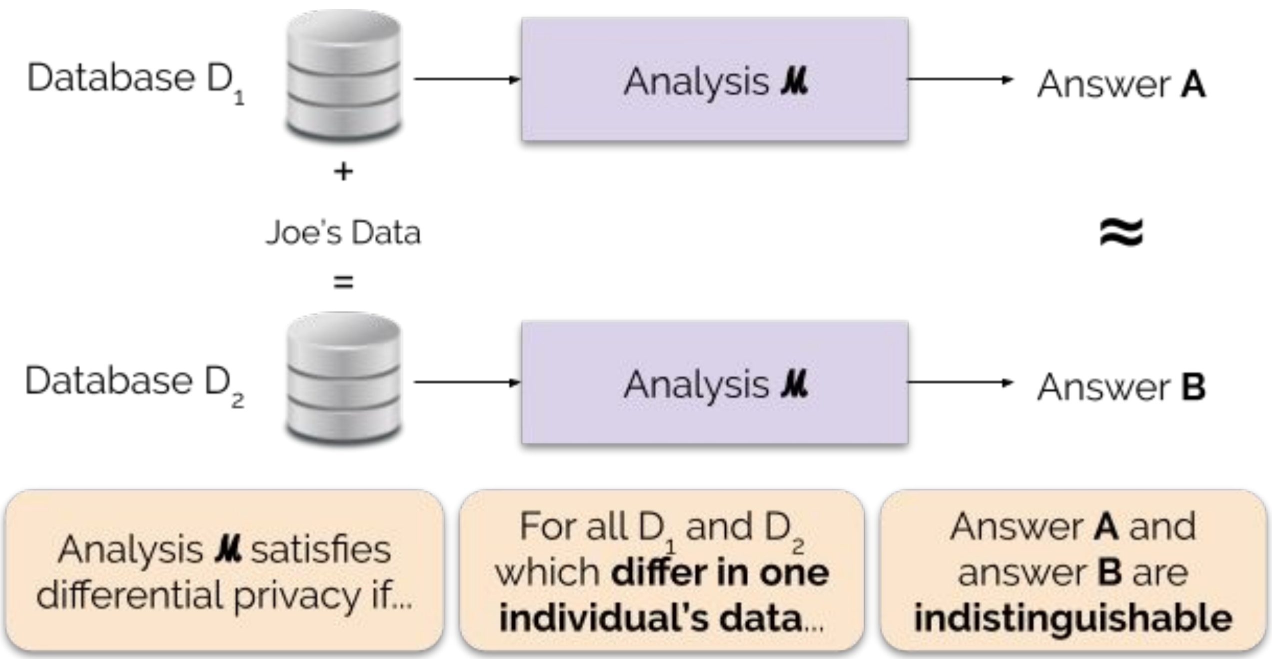
https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf

Non-identifiability Guarantee: Anonymous Data



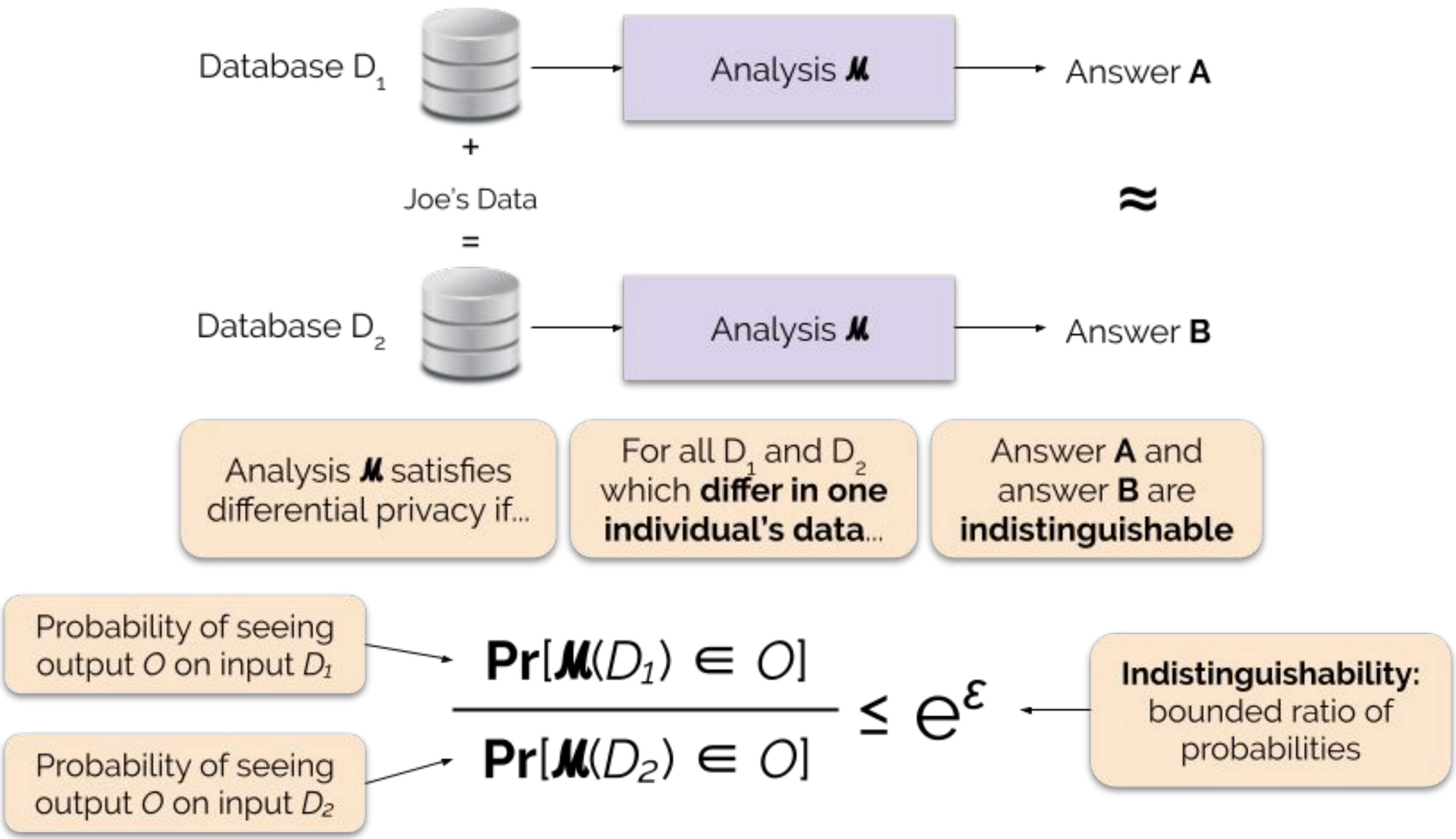
https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf

Differential privacy: no singling out individuals



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

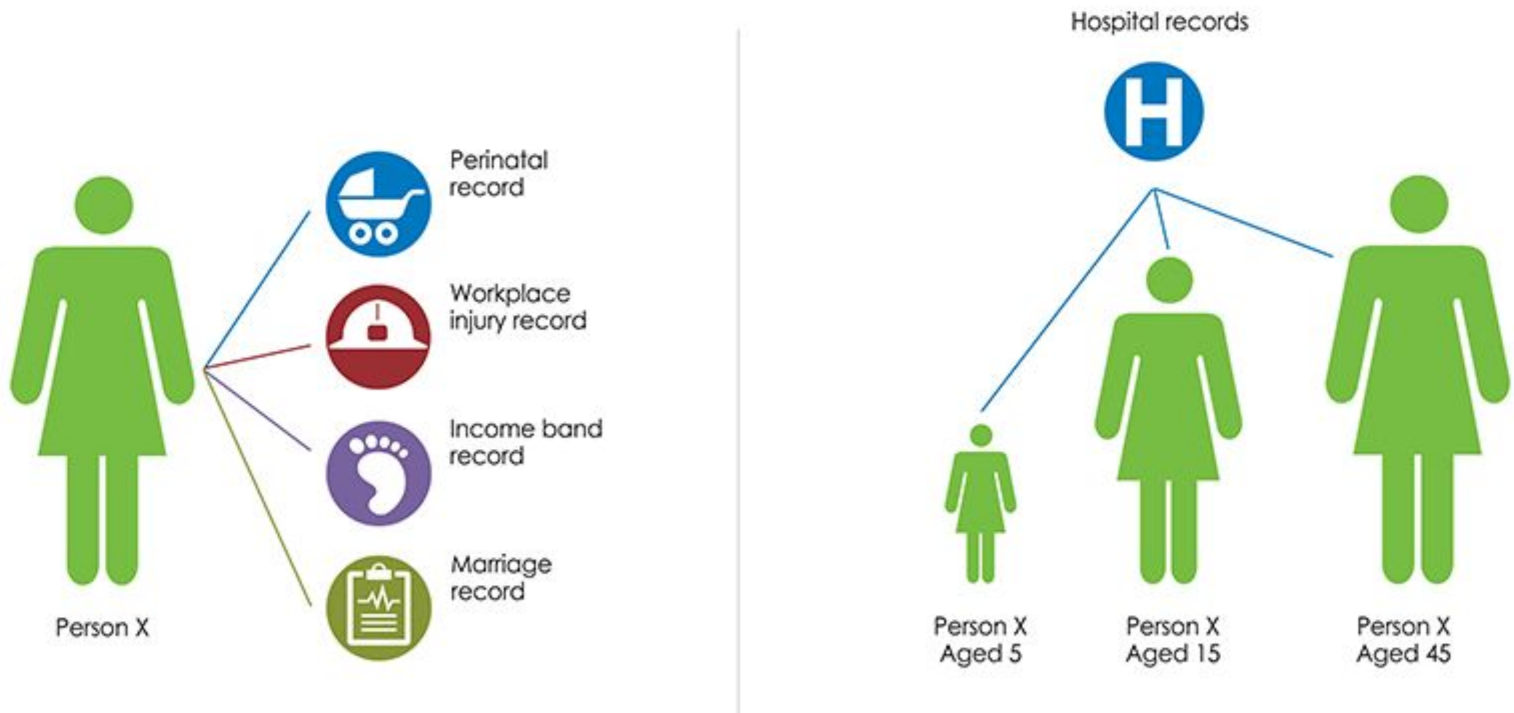
Differential privacy: no singling out individuals



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



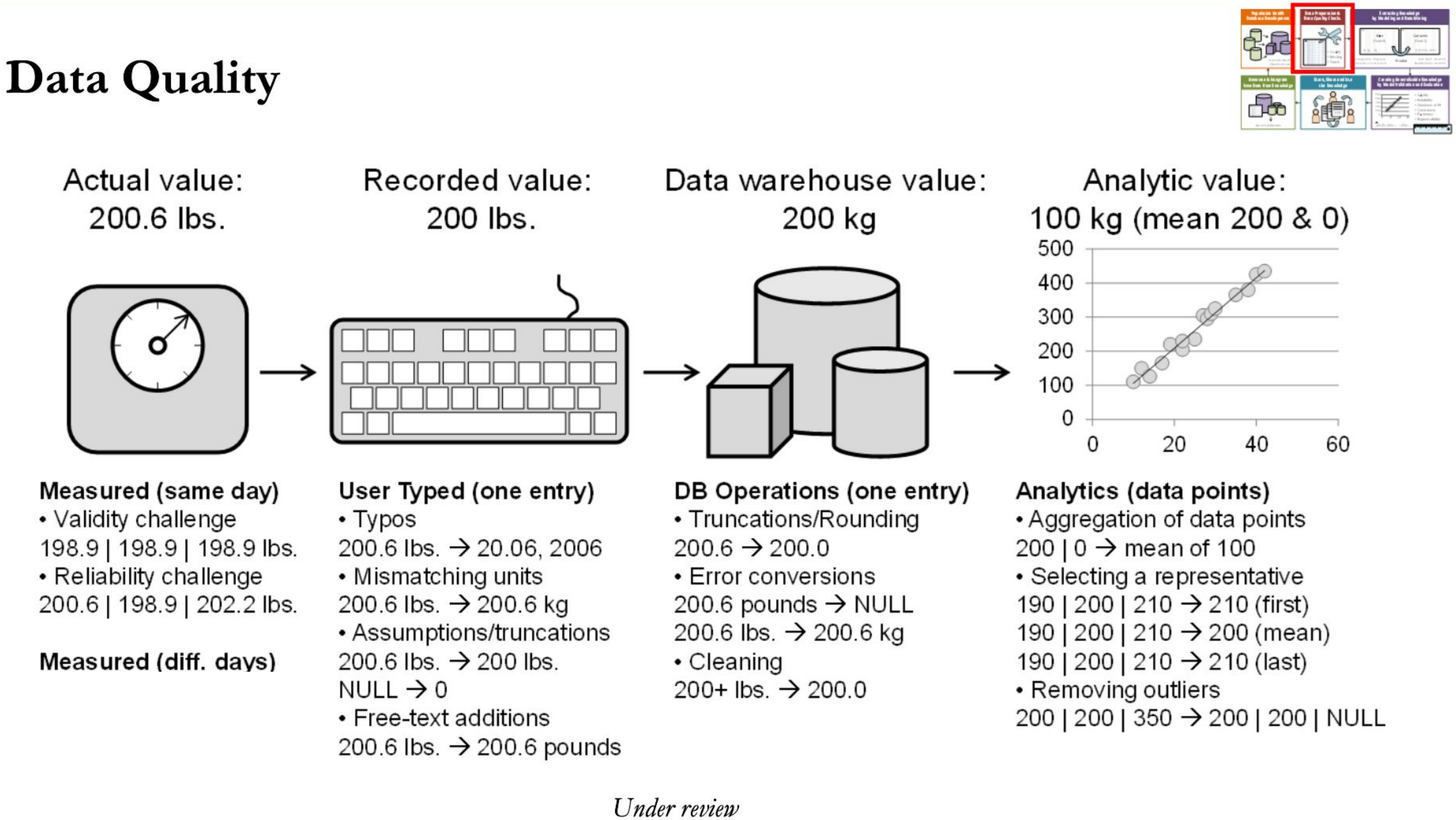
Many different data access processes

- Buy access and get processed data
- Apply for individual fields and justify why
- Full pre-registration of analysis

Let’s take a short break!

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

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

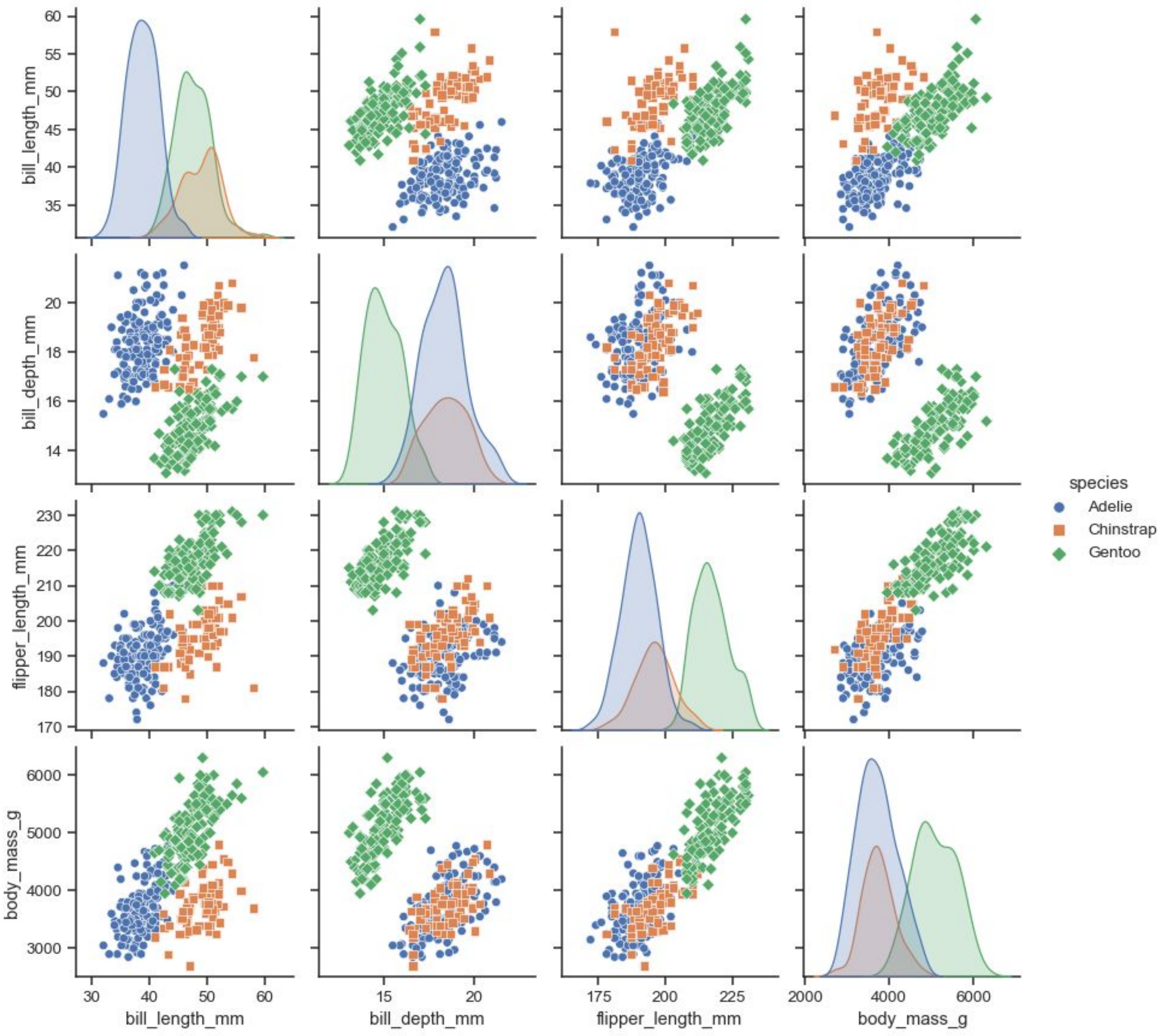


Slide from Dr. Hadi Kharrazi

9 months & >25 rules to clean weight

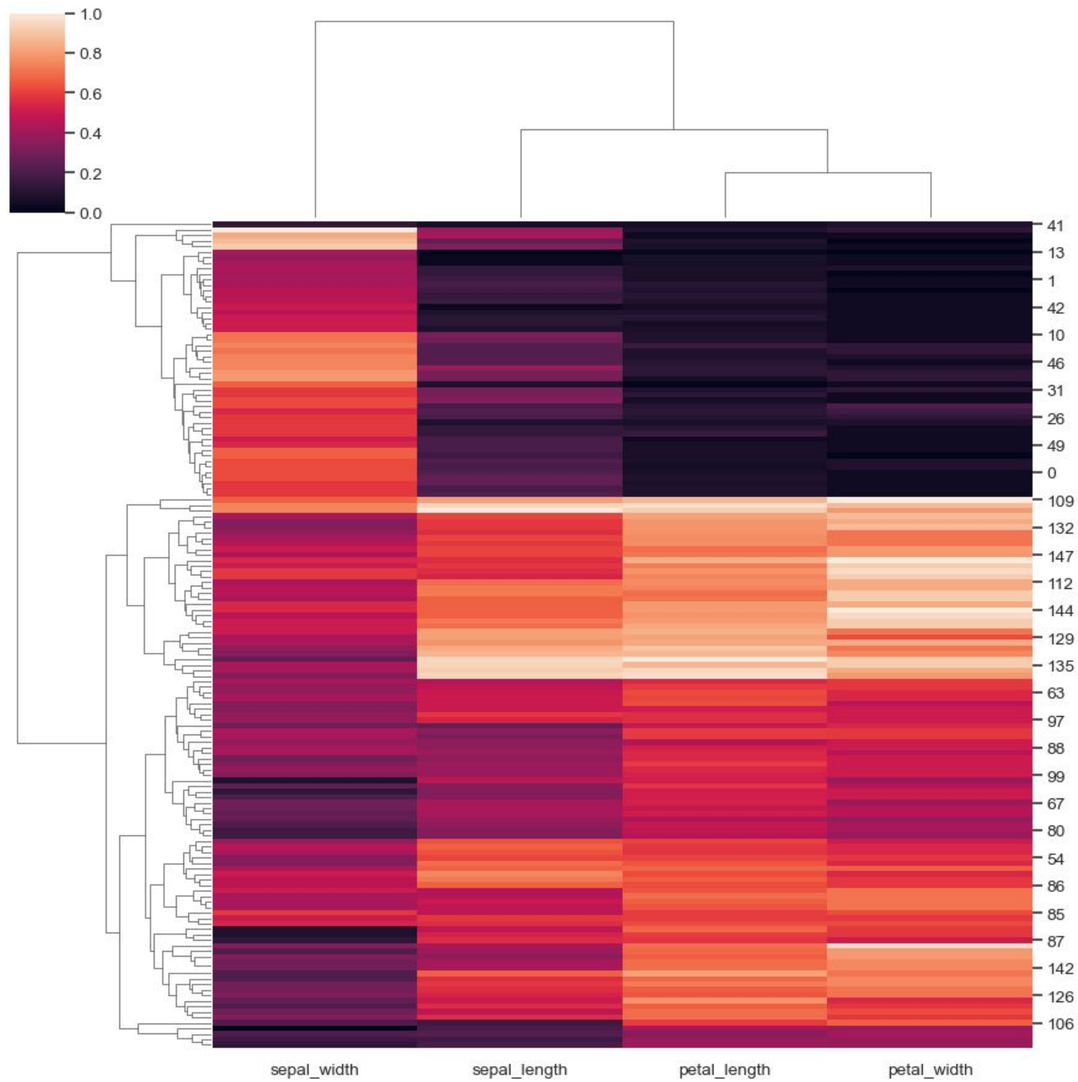
Exploratory Data Analysis

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



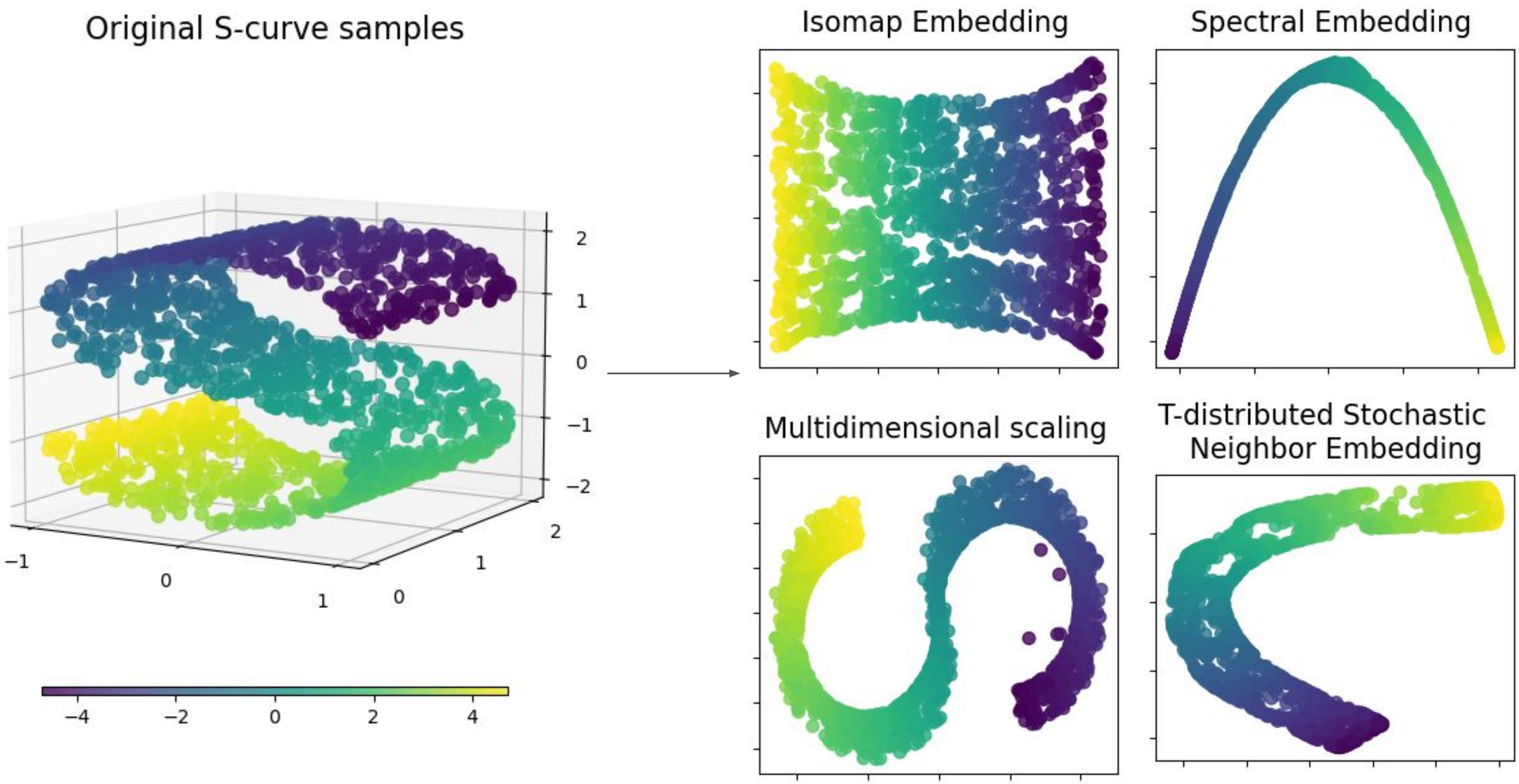
Exploratory Data Analysis

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



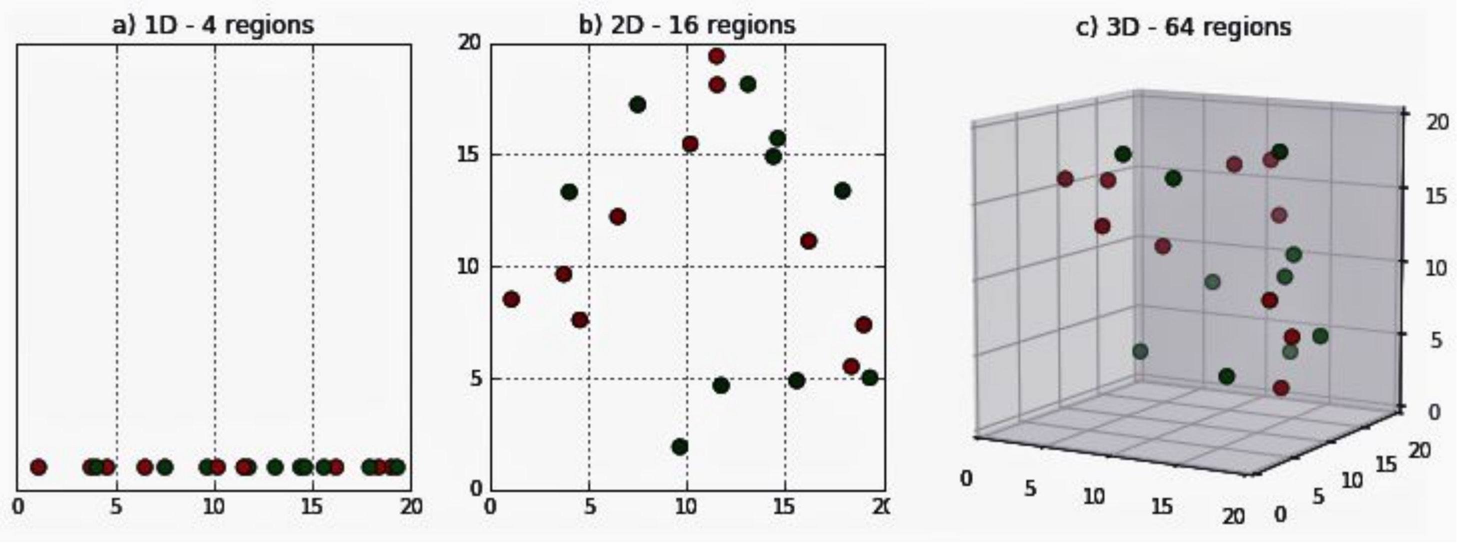
How do I look at all the data together?

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



Why is this hard?

High dimensional data is sparse

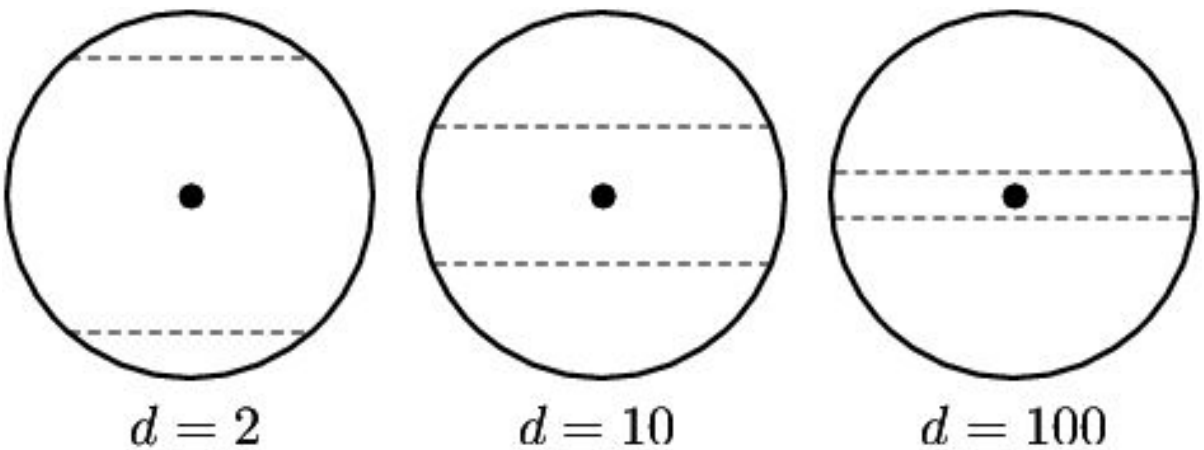


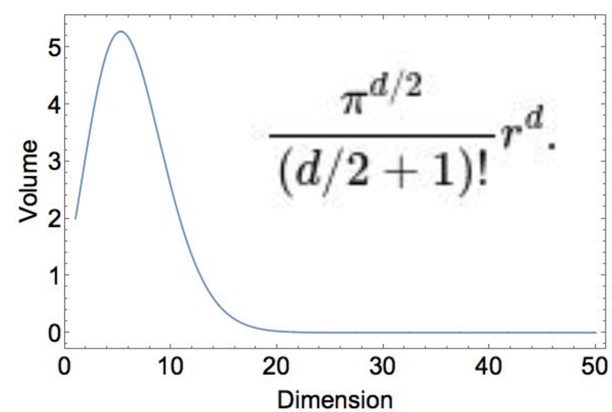
https://medium.com/analytics-vidhya/the-curse-of-dimensionality-and-its-cure-f9891ab72e5c

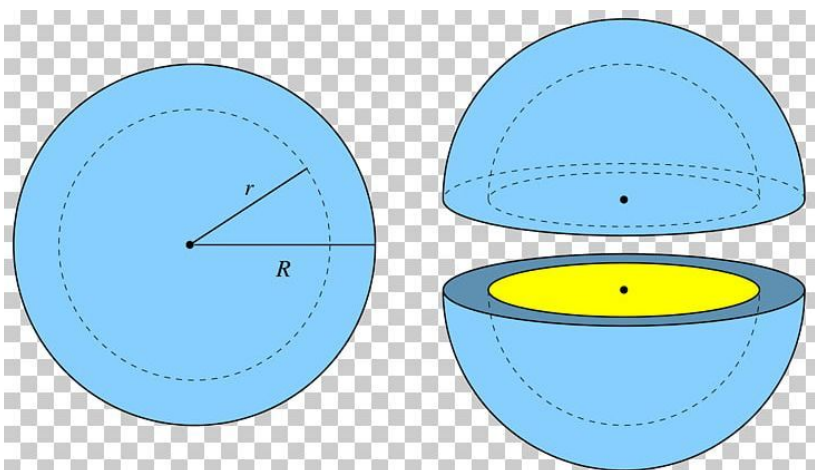
High dimensional space is counterintuitive

As dimensions increase volume enclosed by a d-sphere decreases ~ 0

Band-size to capture 99% of the volume of a sphere:



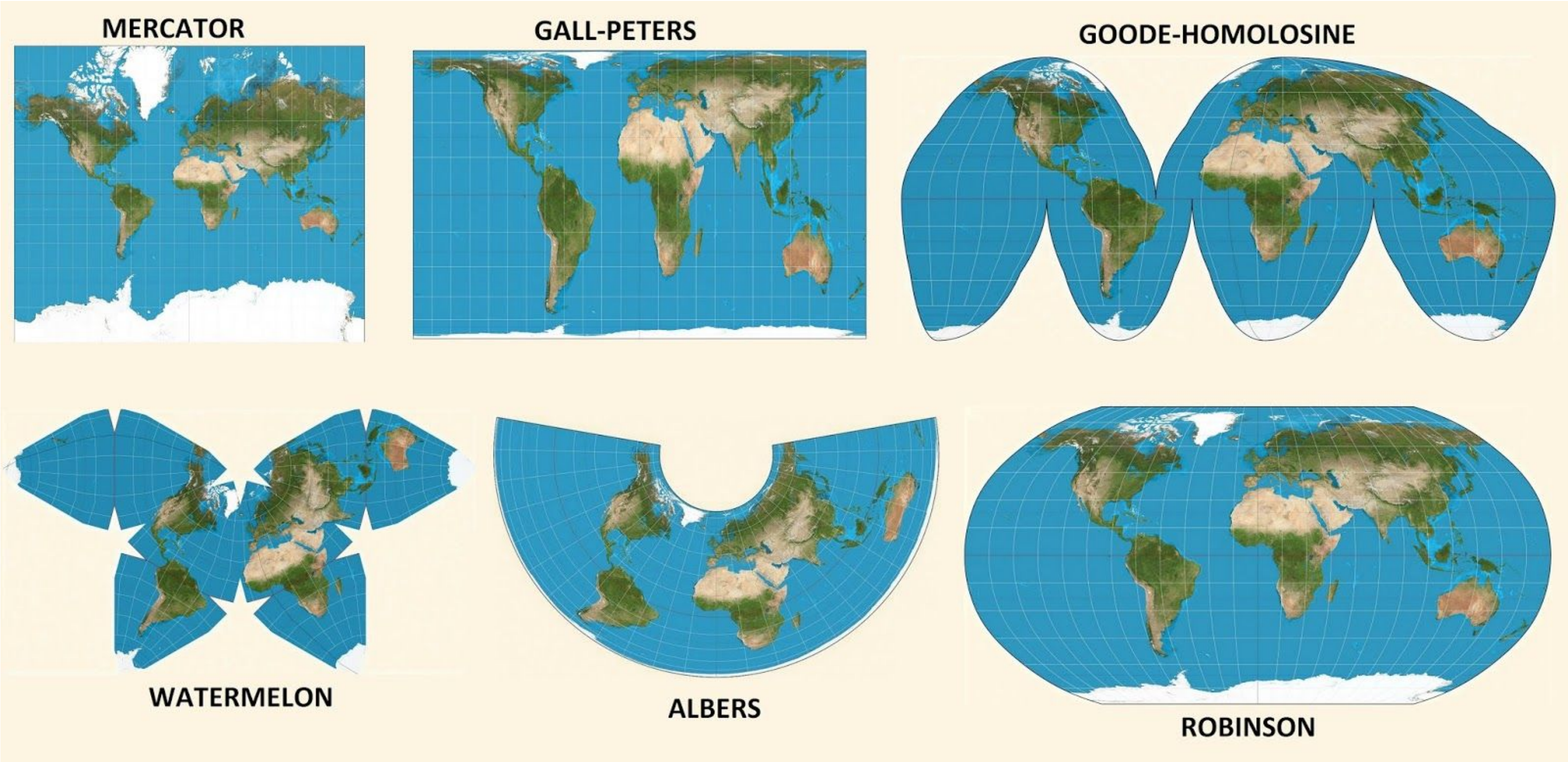




Mass becomes increasingly “shell-like”

HARD TO EFFICIENTLY SAMPLE

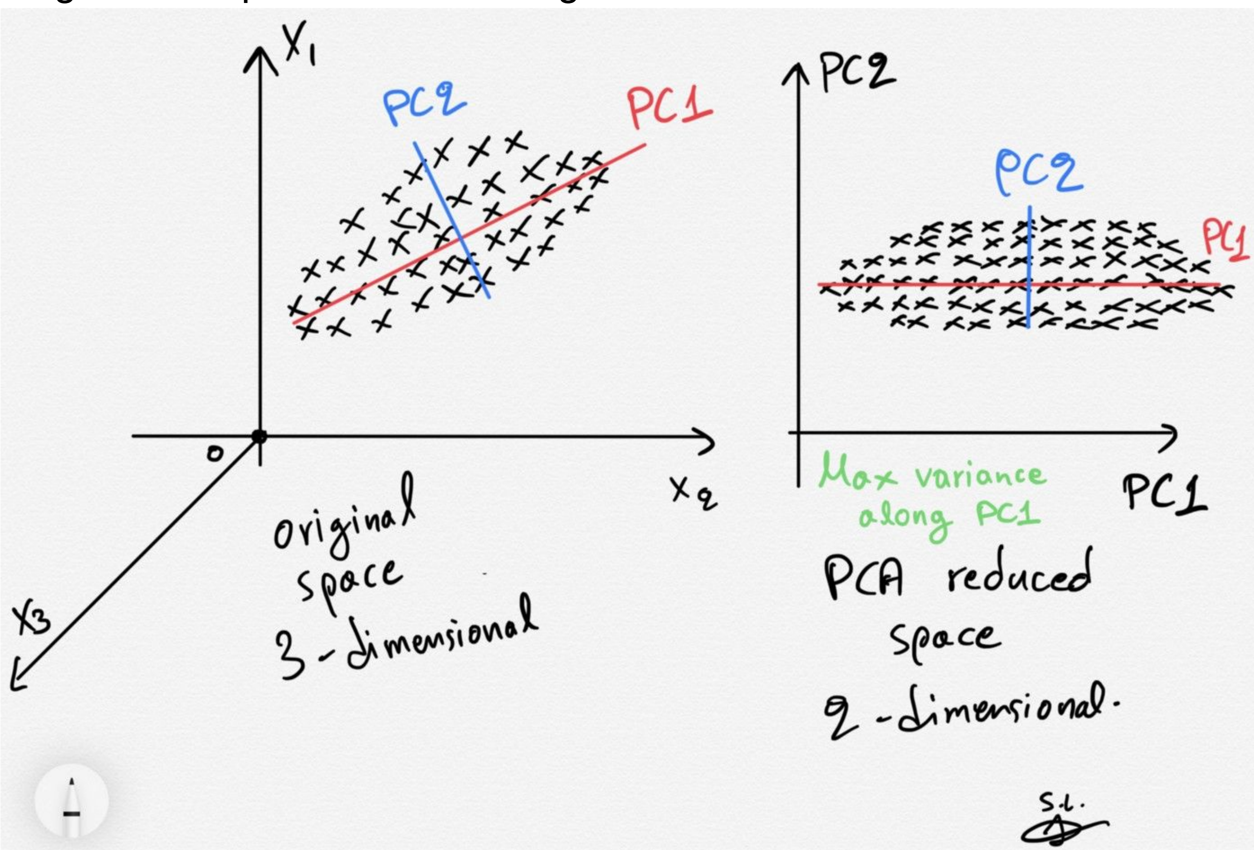
No representation is perfect



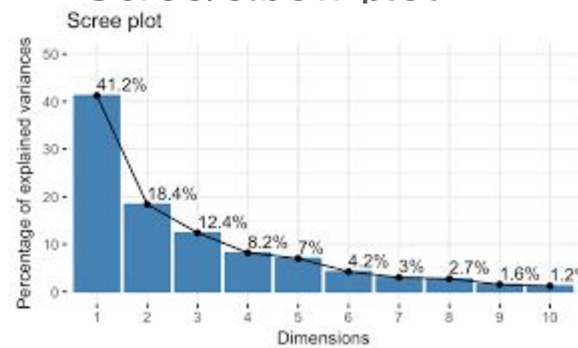
So, how can we do it?

Principal Component Analysis (PCA): Variance

Mean center data -> Generate Covariance Matrix -> Eigendecomposition -> Sort Eigenvalues



- How many components? Scree/elbow plot



- What variables contribute most to PCs? BiPlot

https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

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.

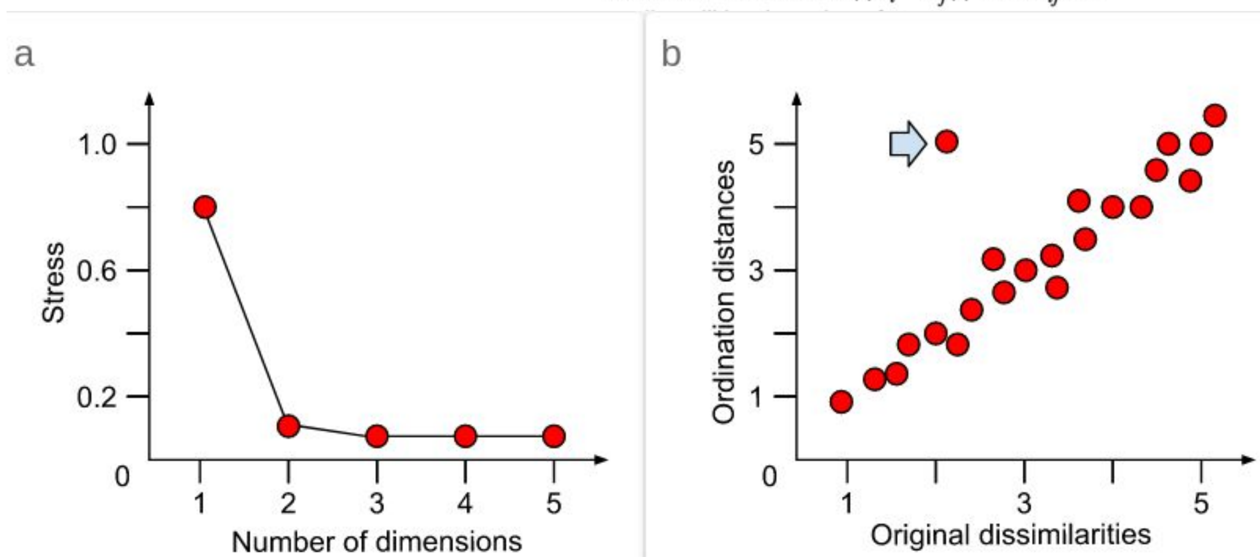
Where $x_1, \ldots, x_N$ are data points with their new set of coordinates in lower dimensional space.

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.

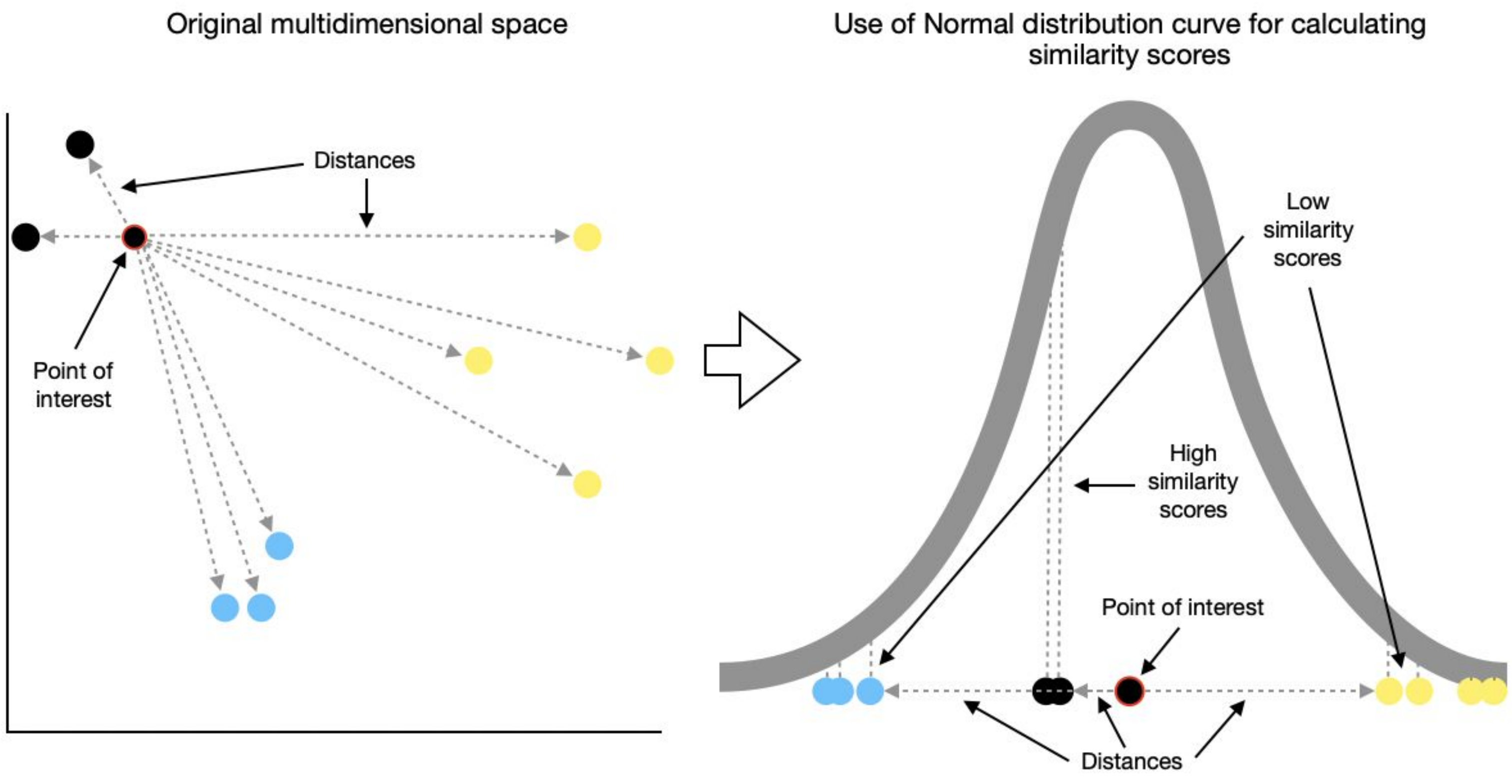
The closer the value of $||x_i - x_j||$ is to d_{ij} the

Non-Metric: Ranks

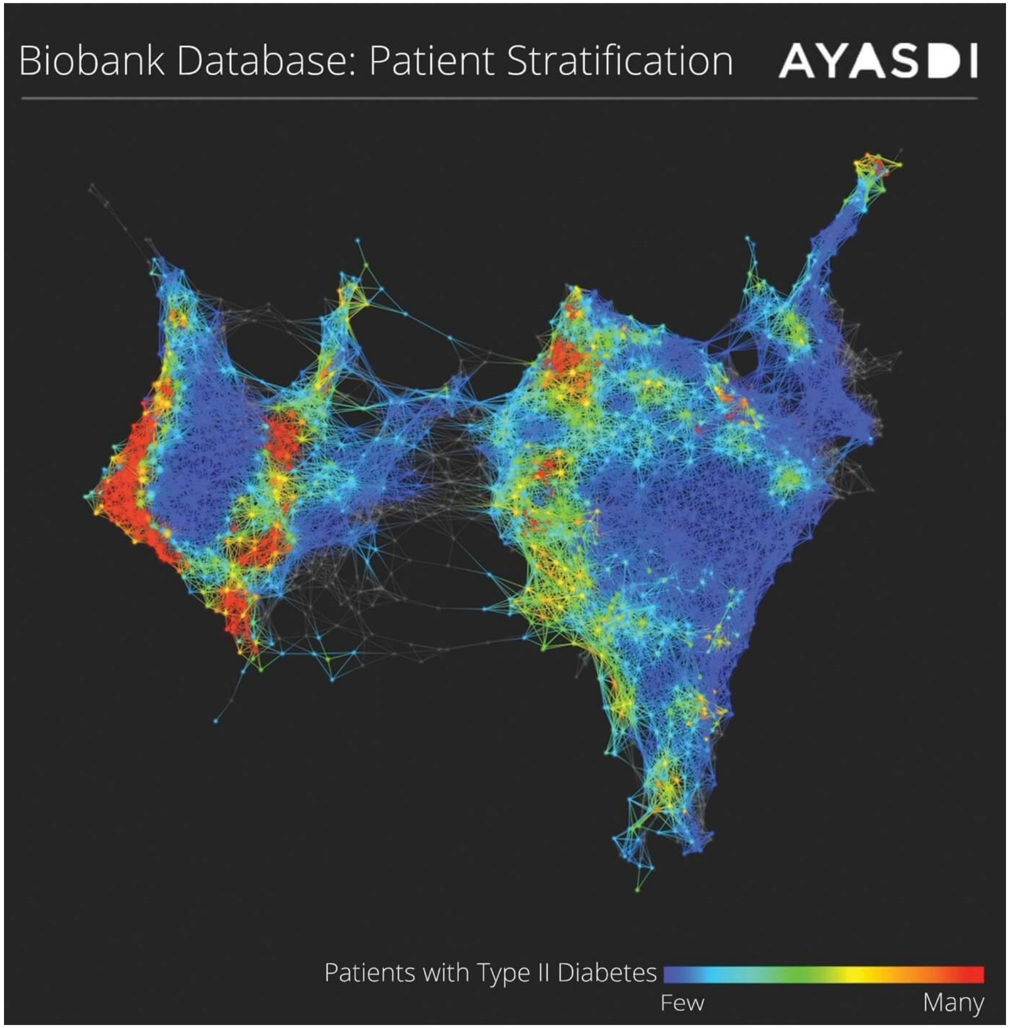


t-SNE/UMAP: Probabilities

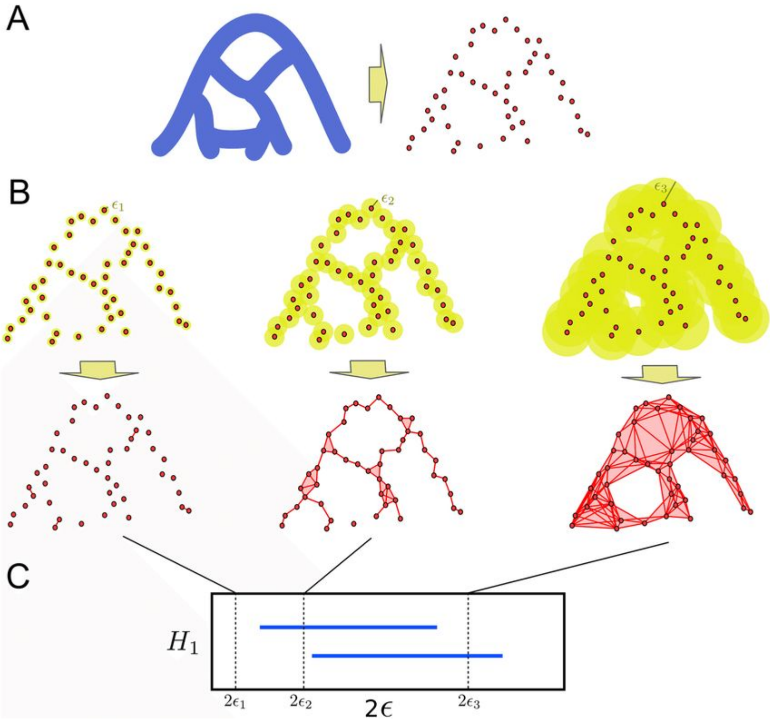
- Pairwise probability distribution in all dimensions
- Pairwise probability distribution in few dimensions
- Stochastic minimisation of KL divergence between distributions



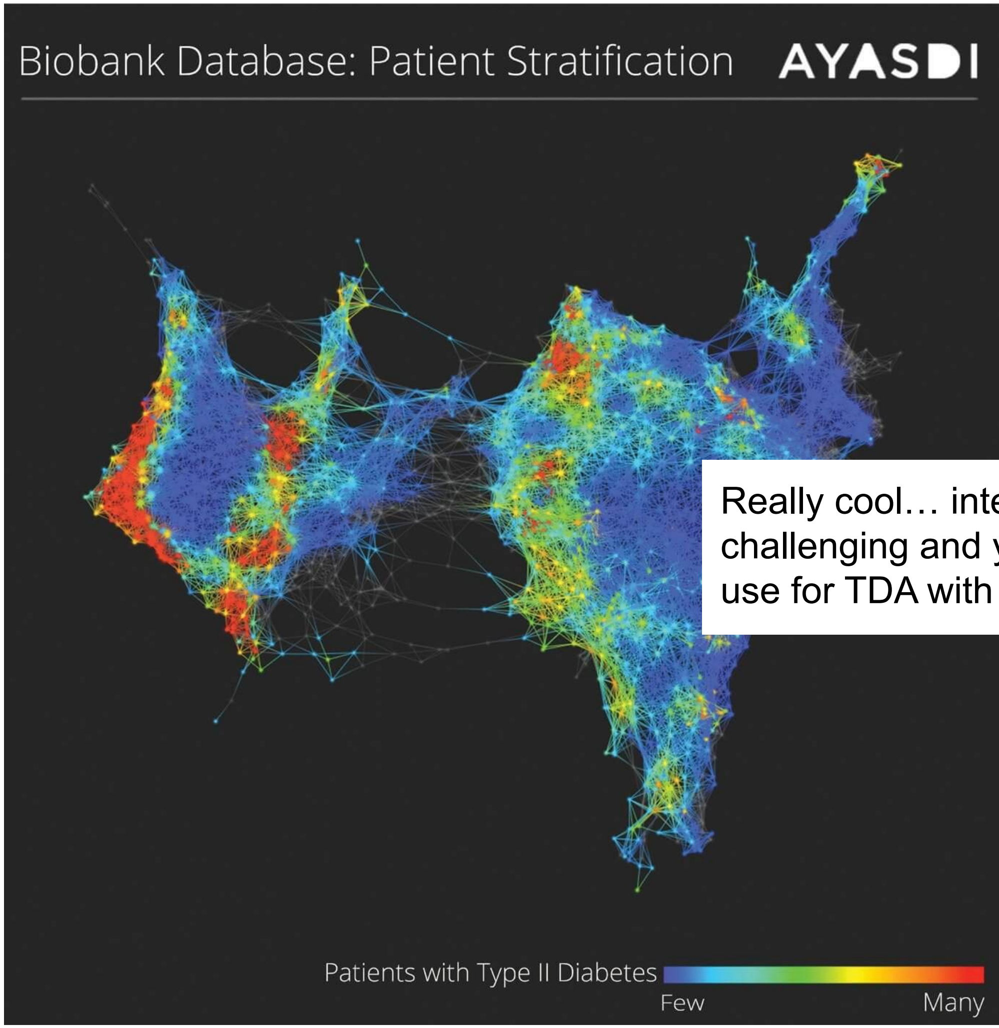
Topological Data Analysis



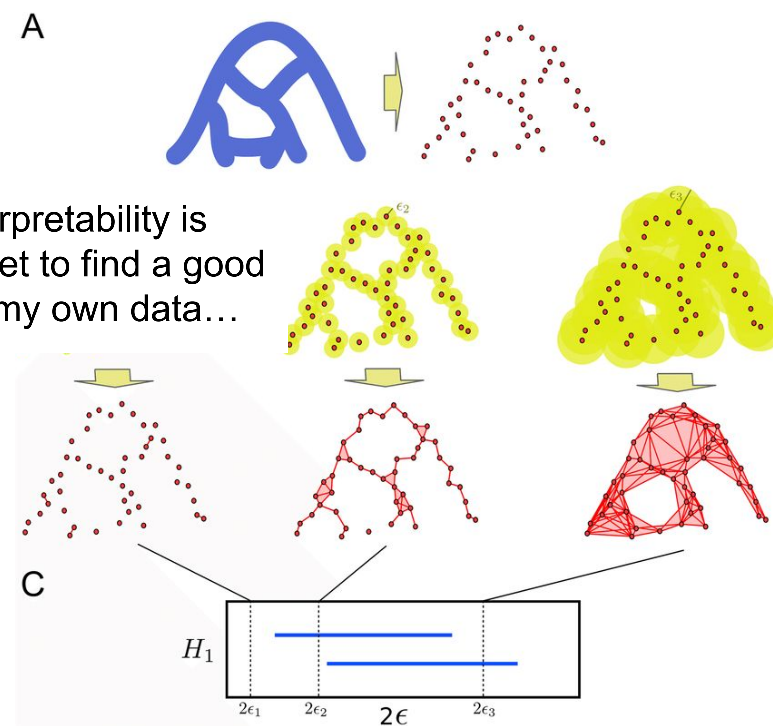
- Point clouds -> increase radius -> simplical complexes -> topological characteristics



Topological Data Analysis



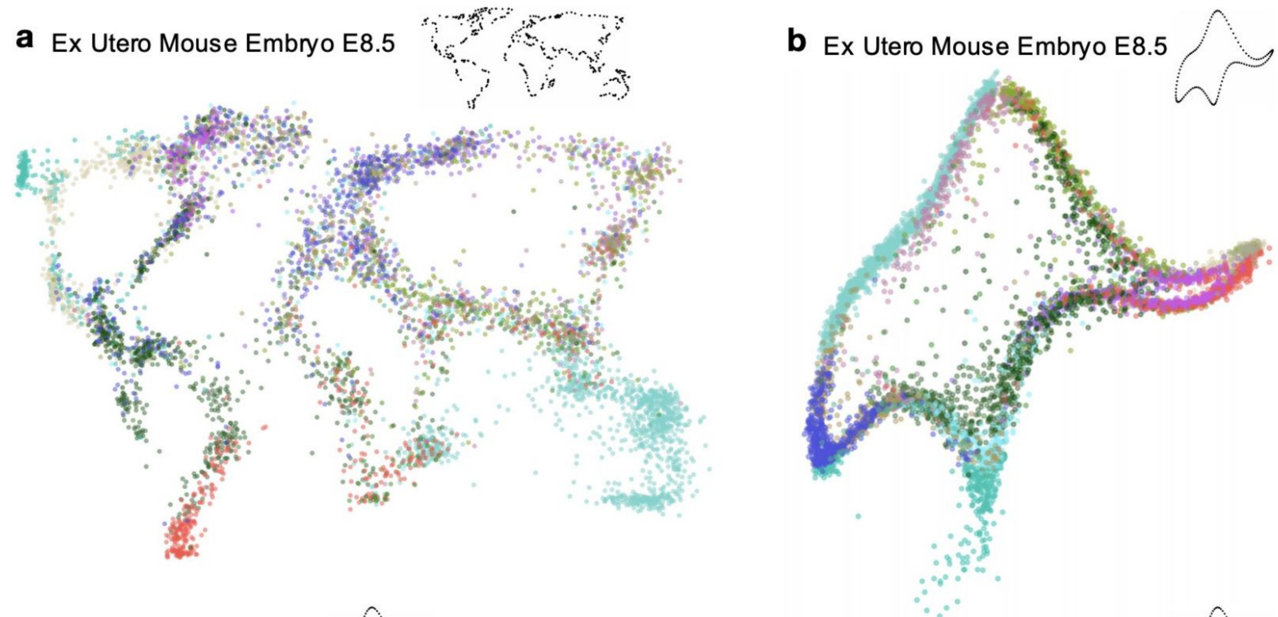
- Point clouds -> increase radius -> simplical complexes -> topological characteristics



Really cool… interpretability is challenging and yet to find a good use for TDA with my own data…

Avoid over-interpreting single plots

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



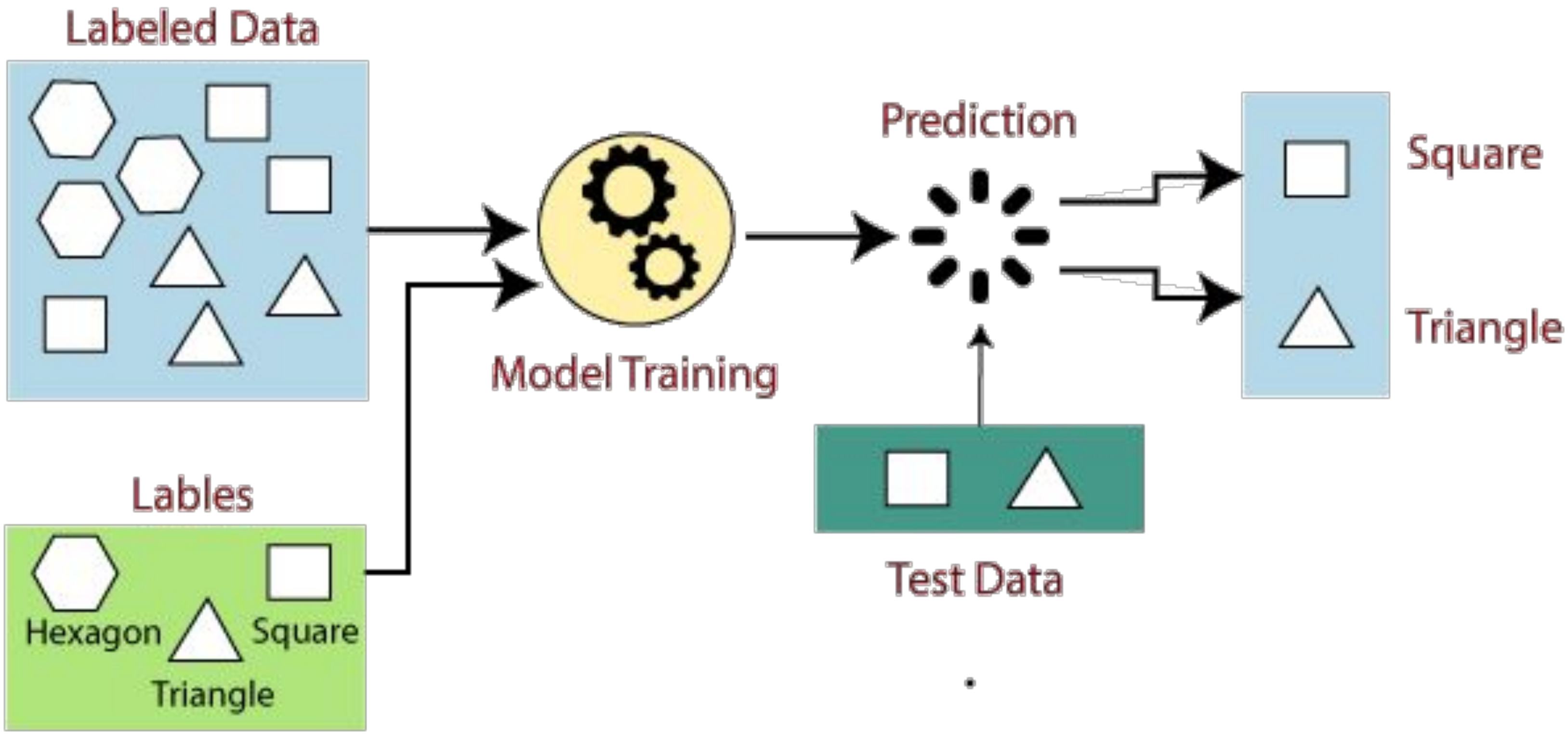
https://www.biorxiv.org/content/10.1101/2021.08.25.457696v3

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

Predicting using tabular data

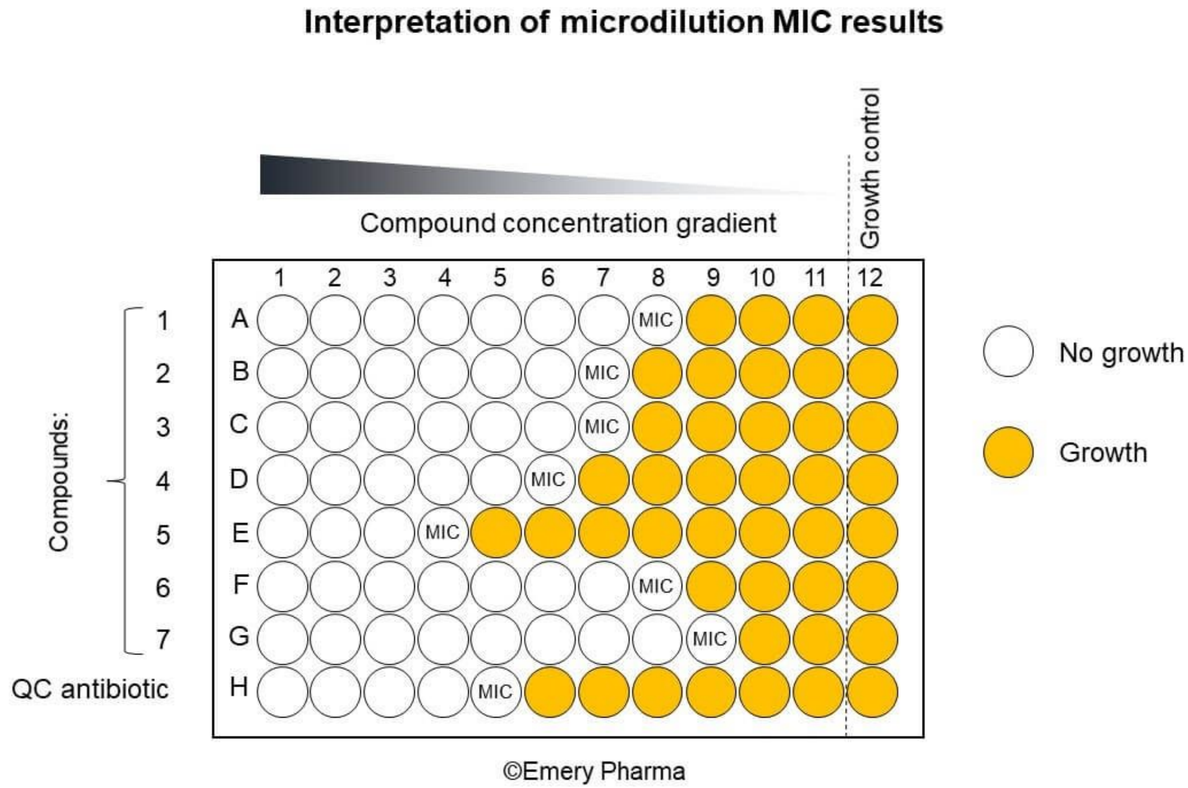
Predicting using tabular data

Predicting Labels or Values

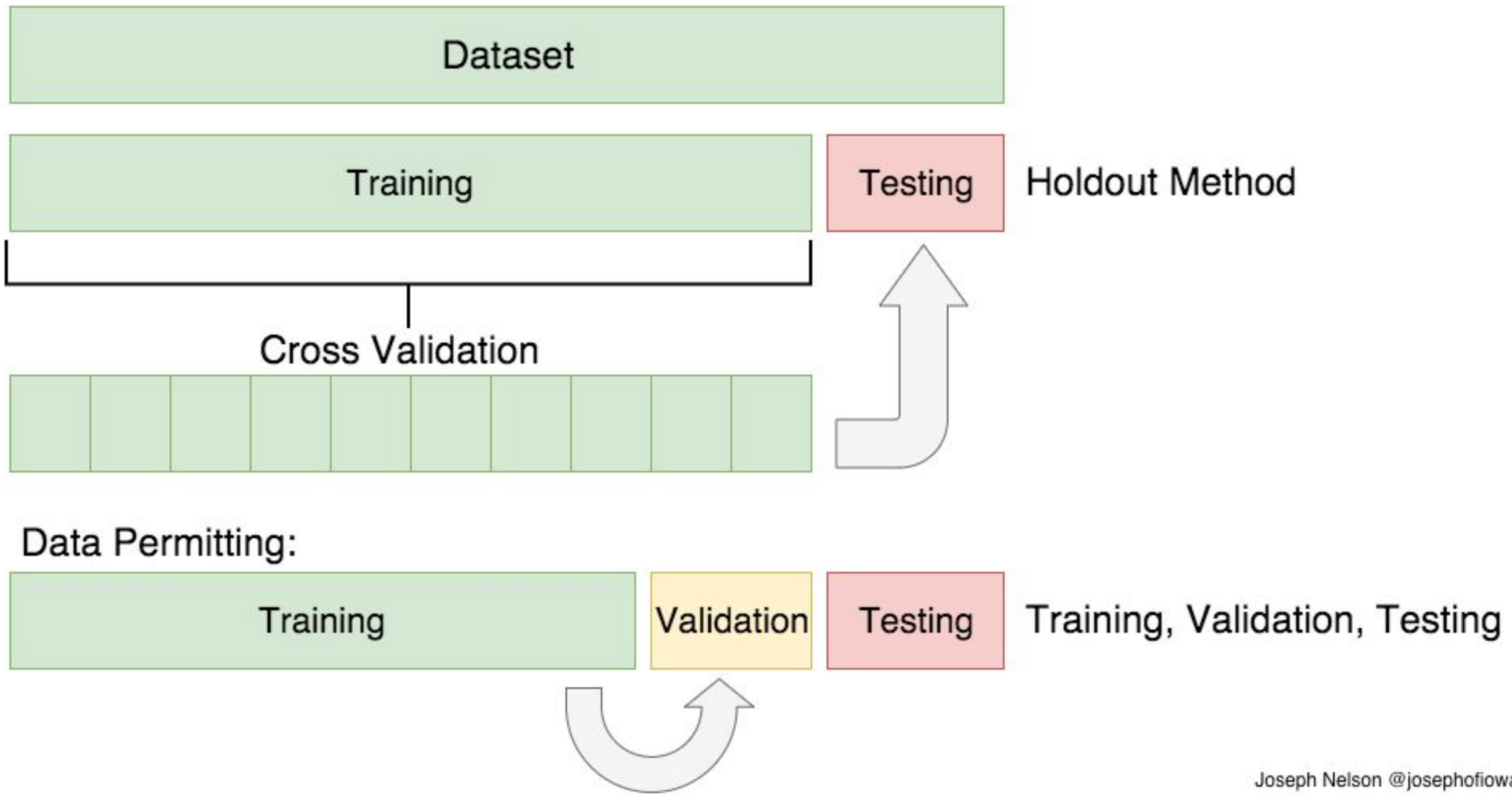


Values can be complex: interval prediction

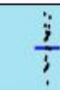
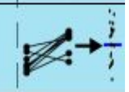
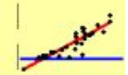
- MIC > highest concentration = *right-censored*
- MIC < lowest concentration = *left-censored*
- Serial Dilutions: MIC of x actually [x/2, 2x] = *unequal error*



Overfitting 101: Test-Train Split

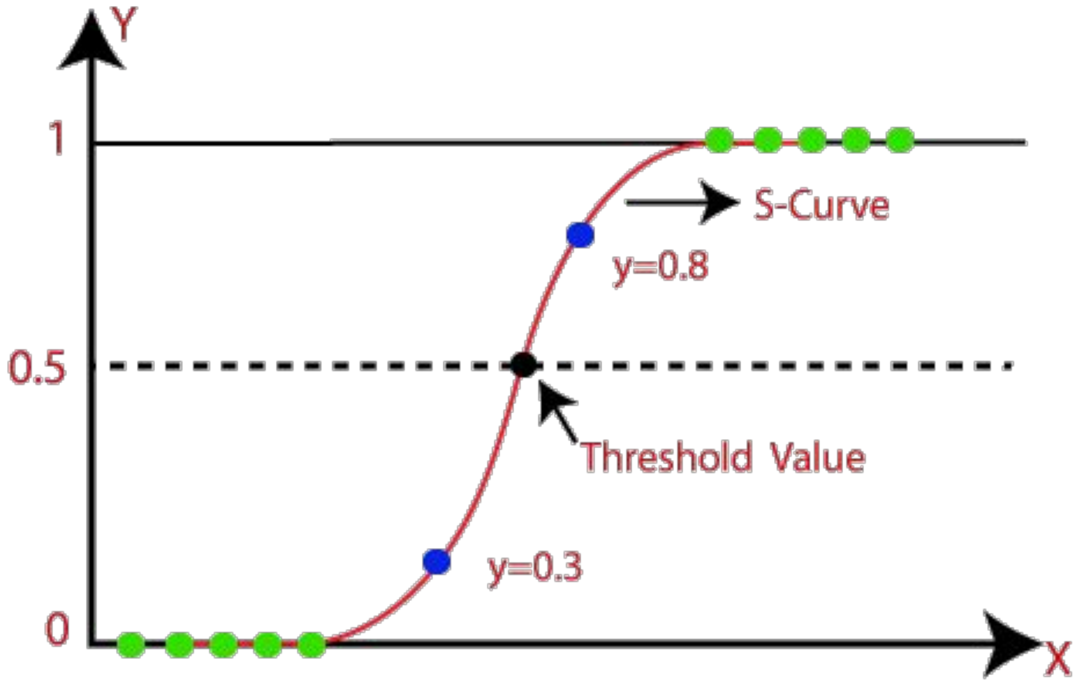


Start simple: linear regression

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=...[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +...)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + S_2 + S_3 +...+ S_K + G_2*S_2 + G_3*S_3 + ... + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $\mathbf{G_{2\ to\ N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\ to\ K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	[Coming]
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2 + G_3*S_3 +...+ G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

https://lindeloev.github.io/tests-as-linear/

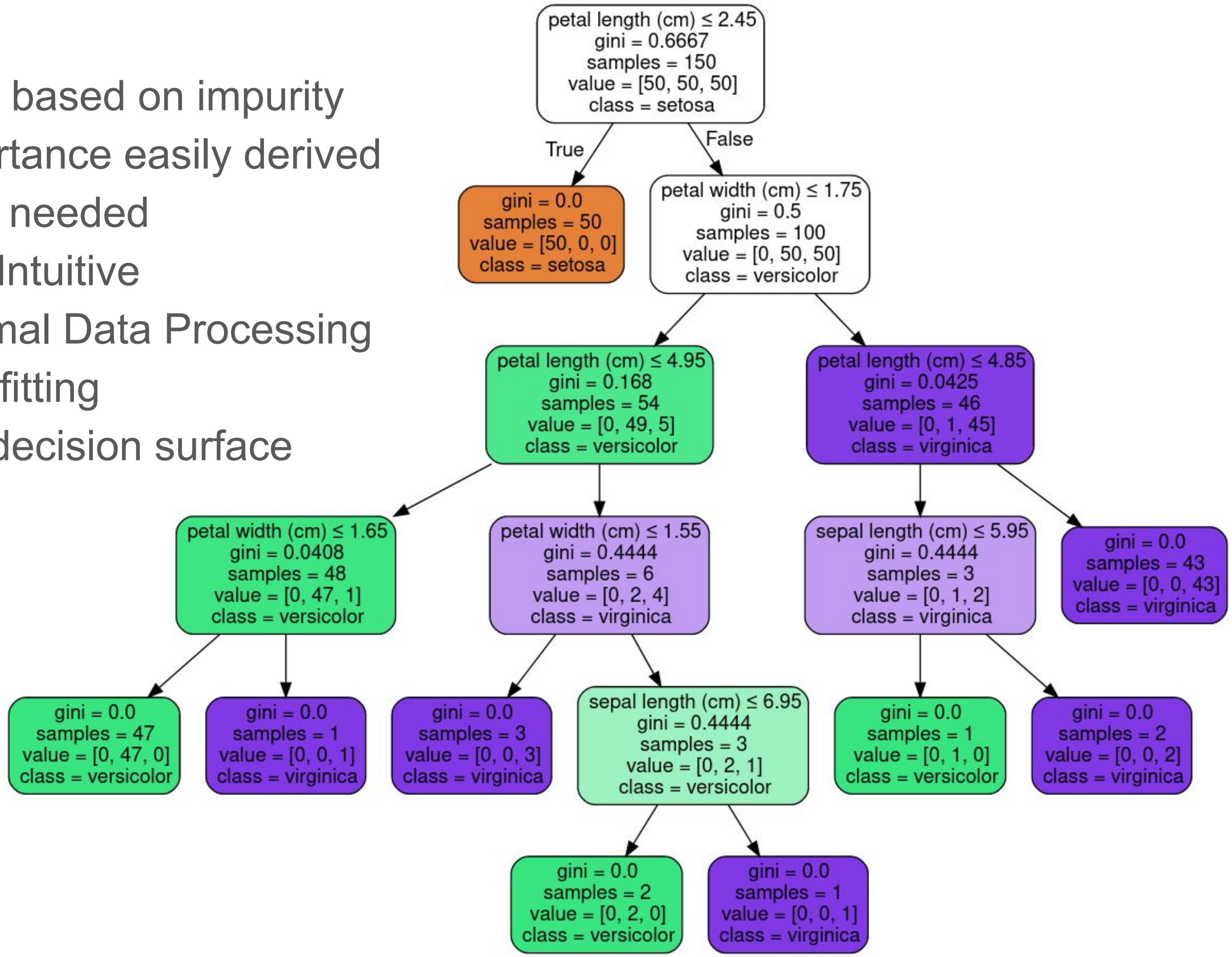
Add a sigmoid for classification



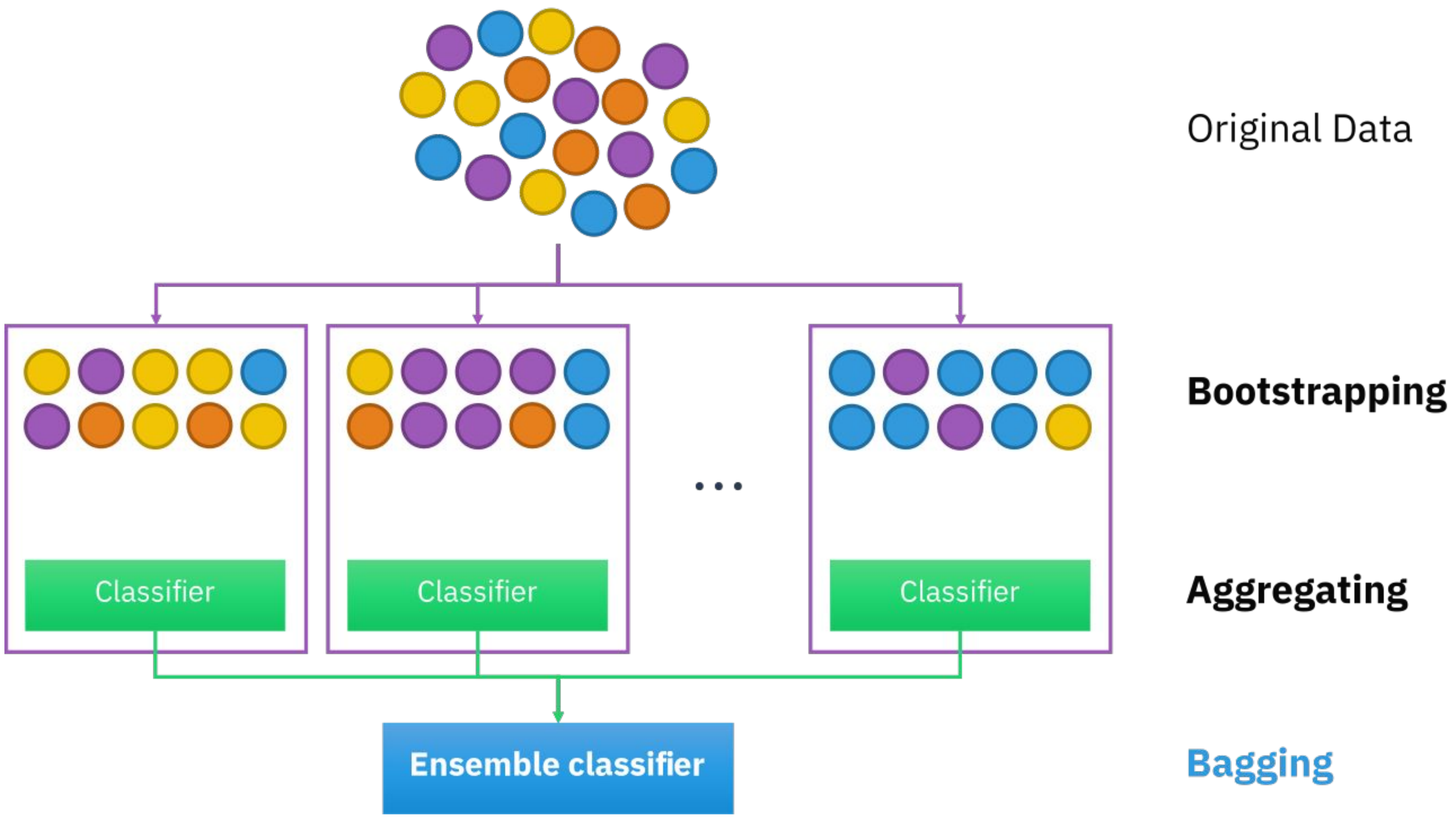
- Crude measure of feature importance (model coefficients)
- Specific feature selection can be a good idea
- Support for regularisation (Lasso/L1 -> sparsity vs Ridge/L2 -> minimal vs ElasticNet -> balance)
- Statistics has developed much better practices for treatment/interpretation of logistic regression

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

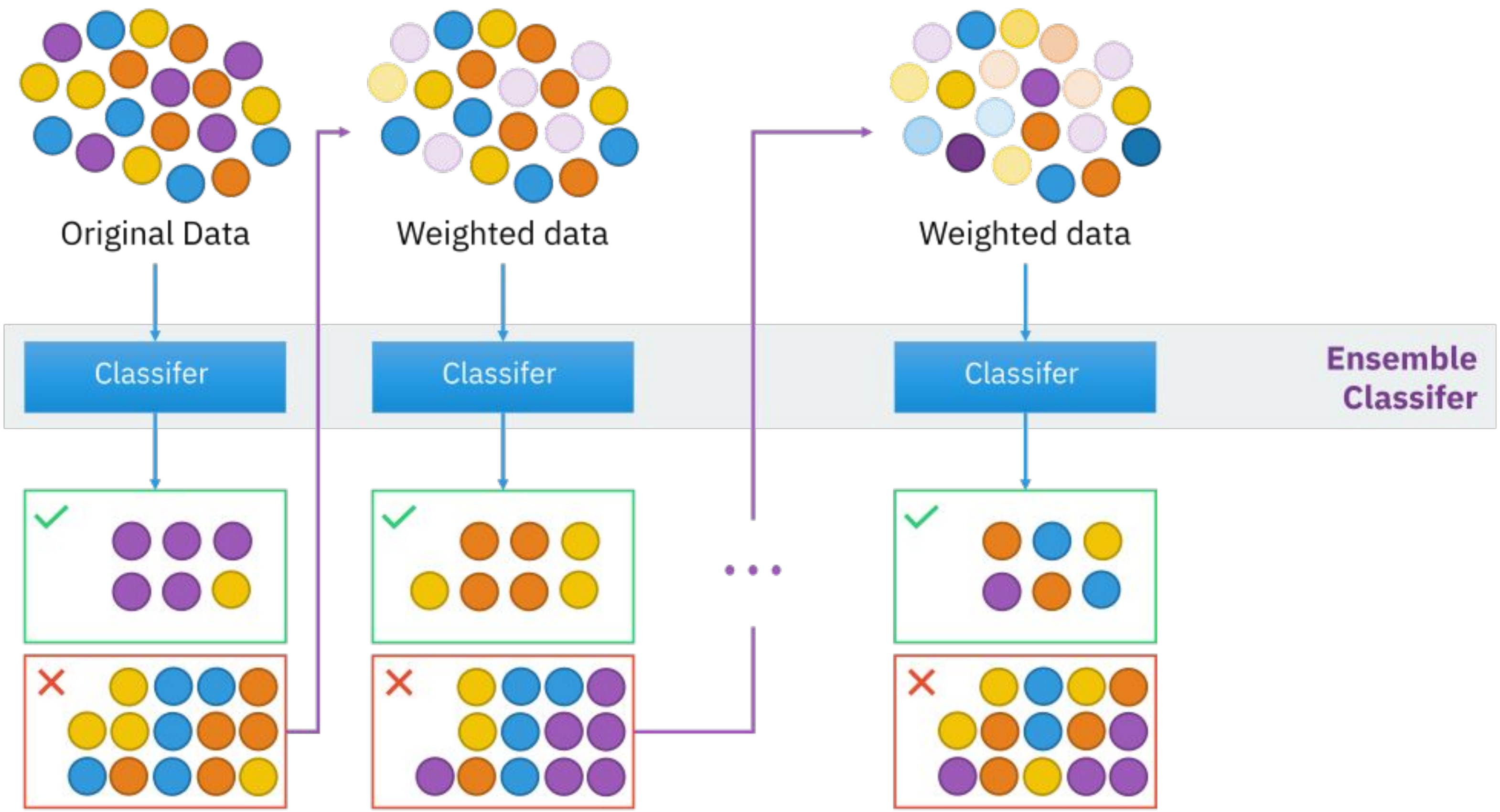


Many Decision Trees: Bagging



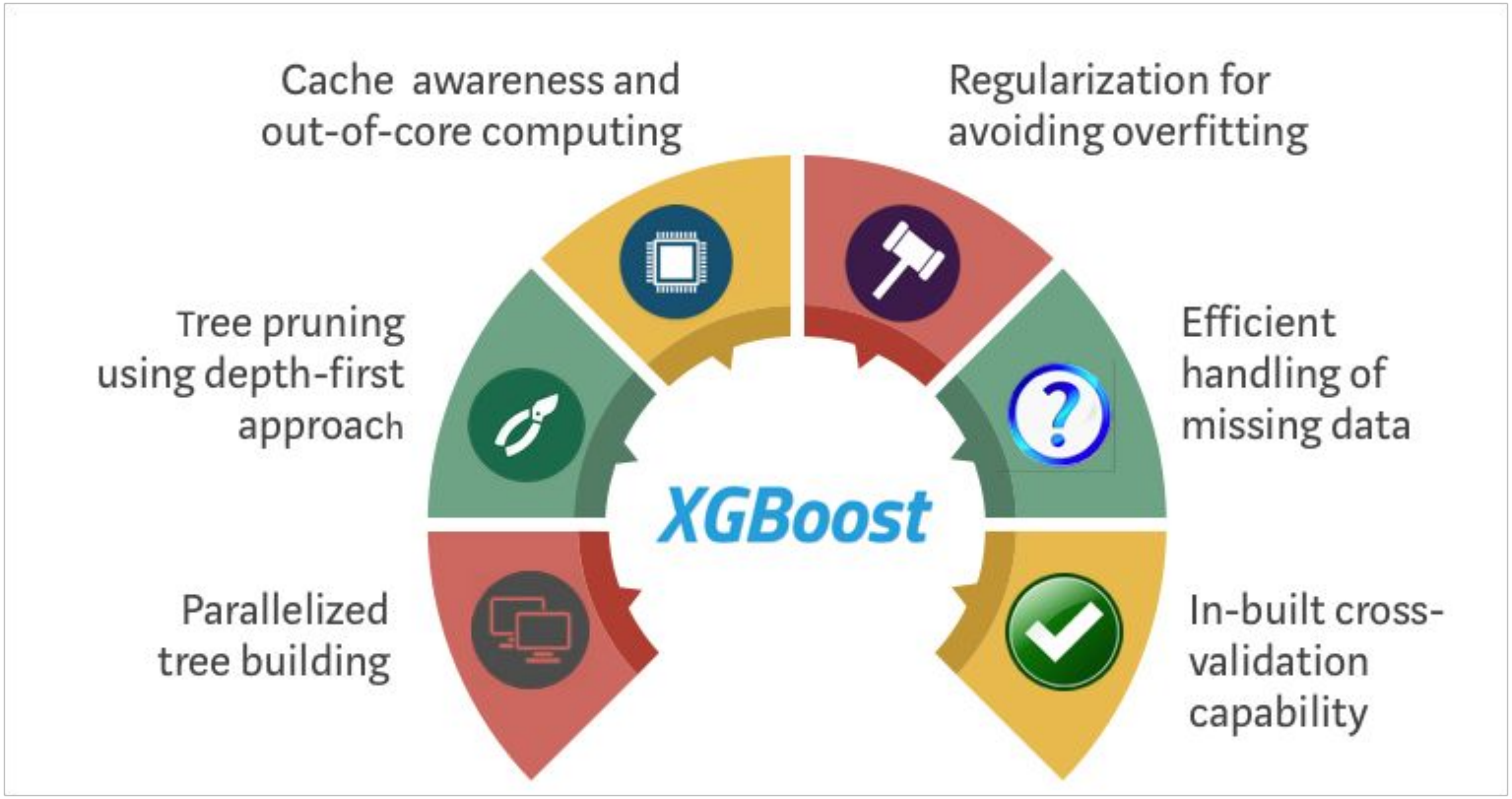
Random Forest: Bagging + Random Subset Per Split
Feature Importance: Average impurity decrease

Boosting: AdaBoost, Gradient Boosting, XGBoost



Gradient Boosting: XGBoost

- Normal boosting is easy to overfit => regularisation
- Use stochastic gradient descent (technically Newton-Raphson variant)
- Many efficiency improvement



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