

# Lecture 1: Medical Databases

CSCI6410/EPAH6410/CSCI4148

Finlay Maguire ([finlay.maguire@dal.ca](mailto: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

**CONFIDENTIAL**

**PART A – PRESENT HEALTH HISTORY (continued)**

IV. GENERAL HEALTH, ATTITUDE AND HABITS (continued)

Please mark if you have any changes in your health or habits. If yes, please explain:

Marital status: No \_\_\_\_\_ Yes \_\_\_\_\_  
Job: No \_\_\_\_\_ Yes \_\_\_\_\_  
Residence: No \_\_\_\_\_ Yes \_\_\_\_\_  
Employment: No \_\_\_\_\_ Yes \_\_\_\_\_  
Are you having any legal problems or conflicts with the law? No \_\_\_\_\_ Yes \_\_\_\_\_

Have any blood relatives had any of the following diseases? If no, indicate relationship (mother, brother, etc.)

Relationship Age, If Living Age At Death State of Health Or Cause of Death Family Members

Father \_\_\_\_\_ Mother \_\_\_\_\_ Brother \_\_\_\_\_ Sister \_\_\_\_\_ Spouse \_\_\_\_\_ Children \_\_\_\_\_

Asthma Diabetes Cataracts Blood Disease Epilepsy Hypertension Tuberculosis Gout High Blood Pressure Heart Disease Mental Problems Suicide

**PART B – PAST HISTORY**

Please give the following information about your immediate family members.

Relationship Age, If Living Age At Death State of Health Or Cause of Death Family Members

Father \_\_\_\_\_ Mother \_\_\_\_\_ Brother \_\_\_\_\_ Sister \_\_\_\_\_ Spouse \_\_\_\_\_ Children \_\_\_\_\_

Have any blood relatives had any of the following diseases? If no, indicate relationship (mother, brother, etc.)

Relationship Age, If Living Age At Death State of Health Or Cause of Death Family Members

Father \_\_\_\_\_ Mother \_\_\_\_\_ Brother \_\_\_\_\_ Sister \_\_\_\_\_ Spouse \_\_\_\_\_ Children \_\_\_\_\_

Asthma Diabetes Cataracts Blood Disease Epilepsy Hypertension Tuberculosis Gout High Blood Pressure Heart Disease Mental Problems Suicide

**PART C – BODY SYSTEMS REVIEW**

ME: Please answer questions 1 through 12, then answer question 13. If you do not know the answer, please start on question 13.

1. Do you have any physical problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
2. Do you have any social problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
3. Have you ever had a cold? No \_\_\_\_\_ Yes \_\_\_\_\_  
4. Do you have any eye problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
5. Do you have any ear problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
6. Do you have any dental problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
7. Do you have any heart problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
8. Do you have any lung problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
9. Do you have any kidney problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
10. Do you have any liver problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
11. Do you have any gallbladder problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
12. Do you have any intestinal problems? No \_\_\_\_\_ Yes \_\_\_\_\_  
13. Do you have any other problems? No \_\_\_\_\_ Yes \_\_\_\_\_

**ANDRUS/CLINI-REC® HEALTH HISTORY QUESTIONNAIRE**

Chart No. \_\_\_\_\_ Date of Birth \_\_\_\_\_  
Name \_\_\_\_\_ Marital Status \_\_\_\_\_  
Occupation \_\_\_\_\_ Physician \_\_\_\_\_ City \_\_\_\_\_

**PART A – PRESENT HEALTH HISTORY**

**CURRENT MEDICAL PROBLEMS**

Please list the medical problem for which you came to see the doctor. About when did they begin?

Problems \_\_\_\_\_ Date Began \_\_\_\_\_

What concerns you most about these problems?

If you are being treated for any other illnesses or medical problems by another physician, please describe the problems and write the name of the physician and address of the medical facility treating you.

Physician or Medical Facility \_\_\_\_\_ City \_\_\_\_\_

**PART B – MEDICATIONS**

Please list all medications you are now taking, including those you buy without a doctor's prescription (such as aspirin, cold tablets or vitamin supplements).

Medication \_\_\_\_\_

**PART C – ALLERGIES AND SENSITIVITIES**

List anything you are allergic to such as certain foods, medications, dust, chemicals, or soaps, household items, pollen, bee stings, insect bites, or any other substances that may affect you.

Allergic To \_\_\_\_\_ Effect \_\_\_\_\_ Allergic To \_\_\_\_\_ Effect \_\_\_\_\_

**PART D – GENERAL HEALTH, ATTITUDE AND HABITS**

How is it been with your health recently? Poor \_\_\_\_\_ Fair \_\_\_\_\_ Good \_\_\_\_\_ Excellent \_\_\_\_\_  
Health has been: Poor \_\_\_\_\_ Fair \_\_\_\_\_ Good \_\_\_\_\_ Excellent \_\_\_\_\_

Has your appetite changed? Decreased \_\_\_\_\_ Increased \_\_\_\_\_ Stayed same \_\_\_\_\_  
Drowsy \_\_\_\_\_ Very tired \_\_\_\_\_  
Are you constipated much of the time? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you feel you have trouble sleeping? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you usually have trouble sleeping? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you smoke? Less than I need \_\_\_\_\_ All I need \_\_\_\_\_  
Smoked \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_ How many years? \_\_\_\_\_  
Cigarettes \_\_\_\_\_ Cigars \_\_\_\_\_ Pipe \_\_\_\_\_  
Have you ever smoked? No \_\_\_\_\_ Yes \_\_\_\_\_ How many years? \_\_\_\_\_  
Do you drink beer? No \_\_\_\_\_ Yes \_\_\_\_\_ How many days per week? \_\_\_\_\_  
Do you drink alcoholic beverages? Alcohol \_\_\_\_\_ No \_\_\_\_\_ I drink \_\_\_\_\_ Beers \_\_\_\_\_ Glasses of Wine \_\_\_\_\_  
Do you regularly wear dentures? Prior problem: No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you regularly wear dentures? Dentures \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_  
**DO YOU**  
Feel nervous? \_\_\_\_\_ Nervous \_\_\_\_\_ Insomniac \_\_\_\_\_  
Find it hard to concentrate? \_\_\_\_\_  
Lose your temper? \_\_\_\_\_  
Tire easily? \_\_\_\_\_  
Have any sexual problems? \_\_\_\_\_  
**DO YOU**  
Ever feel like you can't control your actions? \_\_\_\_\_  
Feel bored with your work? \_\_\_\_\_  
Use marijuanna? \_\_\_\_\_  
Use "hard drugs"? \_\_\_\_\_  
Do you want to talk to the doctor about a personal matter? \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_

Page 2 © 1976, 1983 Biomed Systems Inc.  
Printed in U.S.A.

**CONFIDENTIAL**

ANDRUS/CLINI-REC® HEALTH HISTORY QUESTIONNAIRE  
Today's Date \_\_\_\_\_  
Name \_\_\_\_\_ Marital Status \_\_\_\_\_  
Occupation \_\_\_\_\_ Physician \_\_\_\_\_ City \_\_\_\_\_

**PART A – PRESENT HEALTH HISTORY**

**CURRENT MEDICAL PROBLEMS**

Please list the medical problem for which you came to see the doctor. About when did they begin?

Problems \_\_\_\_\_ Date Began \_\_\_\_\_

What concerns you most about these problems?

If you are being treated for any other illnesses or medical problems by another physician, please describe the problems and write the name of the physician and address of the medical facility treating you.

Physician or Medical Facility \_\_\_\_\_ City \_\_\_\_\_

**PART B – MEDICATIONS**

Please list all medications you are now taking, including those you buy without a doctor's prescription (such as aspirin, cold tablets or vitamin supplements).

Medication \_\_\_\_\_

**PART C – ALLERGIES AND SENSITIVITIES**

List anything you are allergic to such as certain foods, medications, dust, chemicals, or soaps, household items, pollen, bee stings, insect bites, or any other substances that may affect you.

Allergic To \_\_\_\_\_ Effect \_\_\_\_\_ Allergic To \_\_\_\_\_ Effect \_\_\_\_\_

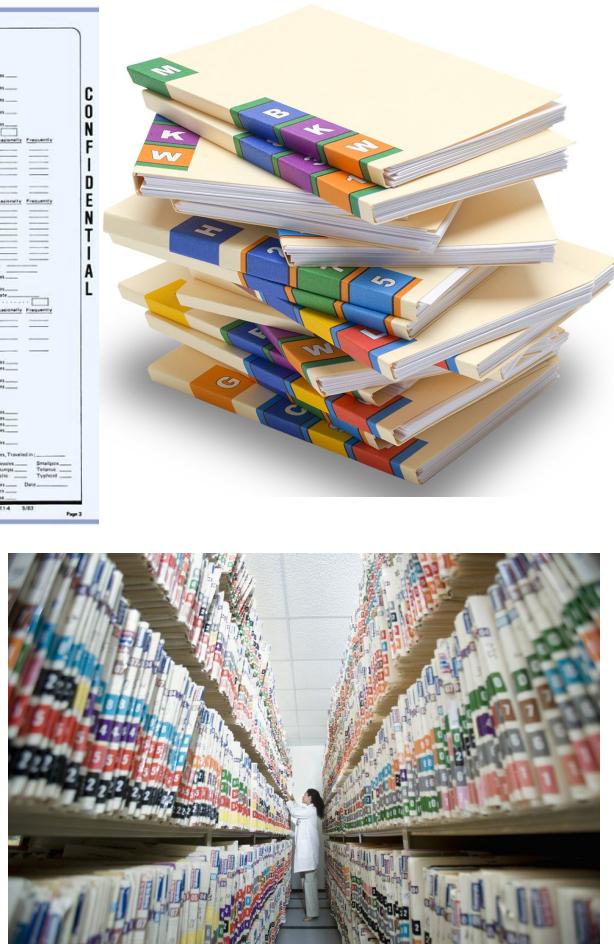
**PART D – GENERAL HEALTH, ATTITUDE AND HABITS**

How is it been with your health recently? Poor \_\_\_\_\_ Fair \_\_\_\_\_ Good \_\_\_\_\_ Excellent \_\_\_\_\_  
Health has been: Poor \_\_\_\_\_ Fair \_\_\_\_\_ Good \_\_\_\_\_ Excellent \_\_\_\_\_

Has your appetite changed? Decreased \_\_\_\_\_ Increased \_\_\_\_\_ Stayed same \_\_\_\_\_  
Drowsy \_\_\_\_\_ Very tired \_\_\_\_\_  
Are you constipated much of the time? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you feel you have trouble sleeping? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you usually have trouble sleeping? No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you smoke? Less than I need \_\_\_\_\_ All I need \_\_\_\_\_  
Smoked \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_ How many years? \_\_\_\_\_  
Cigarettes \_\_\_\_\_ Cigars \_\_\_\_\_ Pipe \_\_\_\_\_  
Have you ever smoked? No \_\_\_\_\_ Yes \_\_\_\_\_ How many years? \_\_\_\_\_  
Do you drink beer? No \_\_\_\_\_ Yes \_\_\_\_\_ How many days per week? \_\_\_\_\_  
Do you drink alcoholic beverages? Alcohol \_\_\_\_\_ No \_\_\_\_\_ I drink \_\_\_\_\_ Beers \_\_\_\_\_ Glasses of Wine \_\_\_\_\_  
Do you regularly wear dentures? Prior problem: No \_\_\_\_\_ Yes \_\_\_\_\_  
Do you regularly wear dentures? Dentures \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_  
**DO YOU**  
Feel nervous? \_\_\_\_\_ Nervous \_\_\_\_\_ Insomniac \_\_\_\_\_  
Find it hard to concentrate? \_\_\_\_\_  
Lose your temper? \_\_\_\_\_  
Tire easily? \_\_\_\_\_  
Have any sexual problems? \_\_\_\_\_  
**DO YOU**  
Ever feel like you can't control your actions? \_\_\_\_\_  
Feel bored with your work? \_\_\_\_\_  
Use marijuanna? \_\_\_\_\_  
Use "hard drugs"? \_\_\_\_\_  
Do you want to talk to the doctor about a personal matter? \_\_\_\_\_ No \_\_\_\_\_ Yes \_\_\_\_\_

Page 2 © 1976, 1983 Biomed Systems Inc.  
Printed in U.S.A.

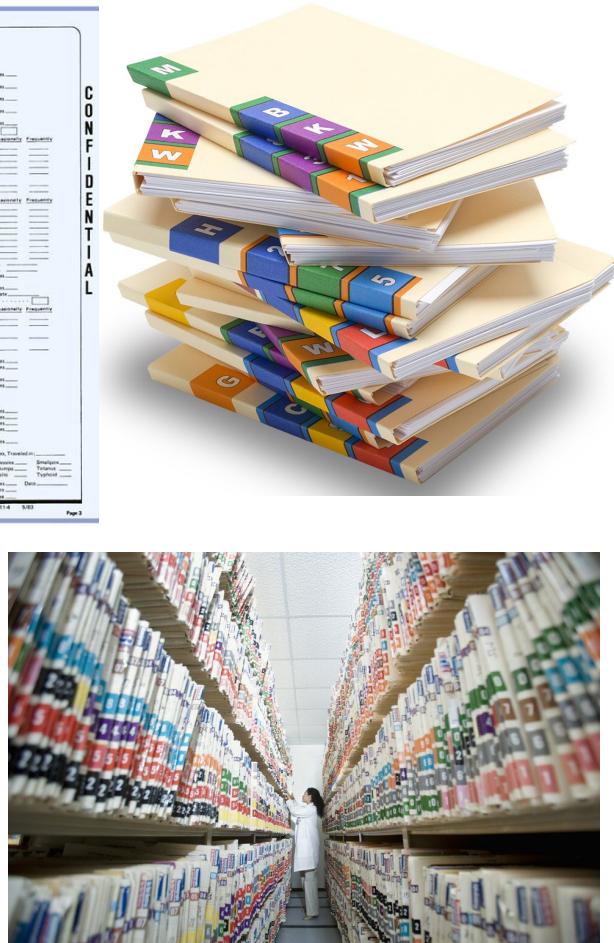
**CONFIDENTIAL**



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

**PART A - PRESENT HEALTH HISTORY (continued)**

IV. GENERAL HEALTH, ATTITUDE AND HABITS (continued)

Please mark with an (X) any of the following if you are not certain when an illness started.

**PART B - PAST HISTORY**

Please give the following information about your immediate family members. If you do not know the relationship, please indicate relationship (mother, brother, etc.)

**PART C - BODY SYSTEMS REVIEW**

ME: Please answer questions 1 through 12, then answer question 13. If you do not know the answer to a question, please start on question 6.

1. Do you have any blood relatives had any of the following diseases? If so, indicate relationship (mother, brother, etc.)

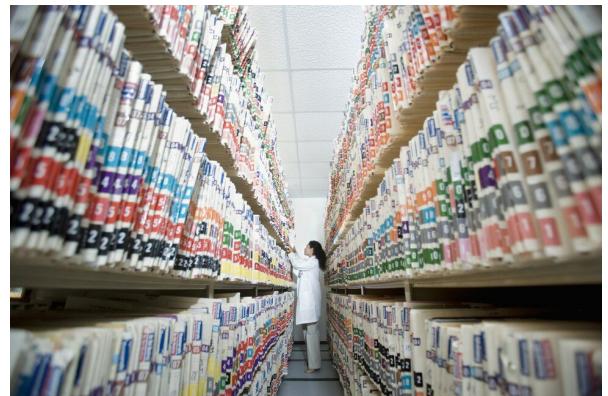
2. Do you have any personal problems? If so, list them below.

3. Do you have any social problems? If so, list them below.

4. Do you eat fast foods or junk food? If so, list them below.

5. Do you drink beer, wine or liquor? If so, list them below.

6. Do you smoke cigarettes or smoking in your residence? If so, how many? Do you wish to discuss any special problems with the doctor?



# Databases (broadly) are ordered collections of data

Examples include:

- Medical Charts
- Phone Book
- Dictionaries
- Spreadsheet

Ordering:

- Index
- Defined fields
- Standardisation

**PART A – PRESENT HEALTH HISTORY (continued)**

IV. GENERAL HEALTH, ATTITUDE AND HABITS (continued)

Have you generally had any changes in your health recently? If yes, please explain:

Marital status: No \_\_\_\_\_ Yes \_\_\_\_\_ If yes, please explain:

Job: \_\_\_\_\_

Residence: \_\_\_\_\_

Employment: \_\_\_\_\_

Are you having any legal problems or conflicts with the law? No \_\_\_\_\_ Yes \_\_\_\_\_

**PART B – PAST HISTORY**

Please give the following information about your immediate family members. Age, if living, \_\_\_\_\_ Age at Death \_\_\_\_\_ State of Health Or Cause of Death \_\_\_\_\_ Family Members \_\_\_\_\_

Father \_\_\_\_\_ Mother \_\_\_\_\_ Brother \_\_\_\_\_ Sister \_\_\_\_\_ Spouse \_\_\_\_\_ Children \_\_\_\_\_

Have any blood relatives had any of the following diseases? If so, indicate relationship (mother, brother, etc.)

Asthma \_\_\_\_\_ Diabetes \_\_\_\_\_ Cataracts \_\_\_\_\_

Blood Disease \_\_\_\_\_ Disease \_\_\_\_\_ Epilepsy \_\_\_\_\_ Tuberculosis \_\_\_\_\_

Phenylketonuria \_\_\_\_\_ Gout \_\_\_\_\_ High Blood Pressure \_\_\_\_\_ Mental Problem \_\_\_\_\_

Suicide \_\_\_\_\_

**PART C – BODY SYSTEMS REVIEW**

ME: Please answer questions 1 through 12, then ask your doctor to answer questions 13 through 20. Please start on question 6.

1. Do you have any pain? \_\_\_\_\_

2. Do you have any skin problems? \_\_\_\_\_

3. Have you ever had a fever? \_\_\_\_\_

4. Do you have any eye problems? \_\_\_\_\_

5. Do you have any ear problems? \_\_\_\_\_

6. Do you have any nose or throat problems? \_\_\_\_\_

7. Do you have any heart problems? \_\_\_\_\_

8. Do you have any lung problems? \_\_\_\_\_

9. Do you have any kidney problems? \_\_\_\_\_

10. Do you have any liver problems? \_\_\_\_\_

11. Do you have any gallbladder problems? \_\_\_\_\_

12. Do you have any bowel problems? \_\_\_\_\_

13. Do you have any bladder problems? \_\_\_\_\_

14. Do you have any blood problems? \_\_\_\_\_

15. Do you have any bone problems? \_\_\_\_\_

16. Do you have any muscle problems? \_\_\_\_\_

17. Do you have any joint problems? \_\_\_\_\_

18. Do you have any nerve problems? \_\_\_\_\_

19. Do you have any brain problems? \_\_\_\_\_

20. Do you have any spinal problems? \_\_\_\_\_

**ANDRUS/CLINI-REC® HEALTH HISTORY QUESTIONNAIRE**

Chart No. \_\_\_\_\_ Date of Birth \_\_\_\_\_ Marital Status \_\_\_\_\_

Occupation: \_\_\_\_\_ Physician or Medical Facility: \_\_\_\_\_ City: \_\_\_\_\_

**PART A – PRESENT HEALTH HISTORY**

**CURRENT MEDICAL PROBLEMS**

Please list the medical problem for which you came to see the doctor. About when did they begin?

Problems: \_\_\_\_\_ Date Began: \_\_\_\_\_

What concerns you most about these problems?

If you are being treated for any other illnesses or medical problems by another physician, please describe the problems and write the name of the physician and the address.

Physician or Medical Facility: \_\_\_\_\_ City: \_\_\_\_\_

**PART B – MEDICATIONS**

Please list all medications you are now taking, including those you buy without a doctor's prescription (such as aspirin, cold tablets or certain supplements).

Medication: \_\_\_\_\_

**PART C – ALLERGIES AND SENSITIVITIES**

List anything that you are allergic to such as certain foods, medications, dust, chemicals, or soaps, household items, pollen, bee stings, etc. Please list how each affects you.

Allergic To: \_\_\_\_\_ Effect: \_\_\_\_\_ Allergic To: \_\_\_\_\_ Effect: \_\_\_\_\_

**PART D – GENERAL HEALTH, ATTITUDE AND HABITS**

How is your general health? \_\_\_\_\_

How has it been most of your life? \_\_\_\_\_

Health has been: Poor \_\_\_\_\_ Fair \_\_\_\_\_ Good \_\_\_\_\_ Excellent \_\_\_\_\_

Has your appetite changed? \_\_\_\_\_

Appetite: Decreased \_\_\_\_\_ Increased \_\_\_\_\_ Stayed same \_\_\_\_\_

Are you drinking much of the time? \_\_\_\_\_

Drinking: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you usually have **more** sleep than you need? \_\_\_\_\_

Sleep: Decreased \_\_\_\_\_ Increased \_\_\_\_\_

Do you smoke? \_\_\_\_\_

Smokes: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you drink? \_\_\_\_\_

Drinks: Less than I need \_\_\_\_\_ As I need \_\_\_\_\_ More than I need \_\_\_\_\_

Do you ever smoke? \_\_\_\_\_

Smoked: No \_\_\_\_\_ Yes \_\_\_\_\_

How many years? \_\_\_\_\_

Do you drink alcohol? \_\_\_\_\_

Alcohol: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you drink alcoholic beverages? \_\_\_\_\_

Alcohol: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you drink beer? \_\_\_\_\_

Beer: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you drink wine? \_\_\_\_\_

Wine: No \_\_\_\_\_ Yes \_\_\_\_\_

Do you drink coffee or tea a day? \_\_\_\_\_

Coffee or Tea: No \_\_\_\_\_ Yes \_\_\_\_\_

DO YOU \_\_\_\_\_

Feel nervous? \_\_\_\_\_

Find it hard to concentrate? \_\_\_\_\_

Forget things? \_\_\_\_\_

Love your temper? \_\_\_\_\_

Are you irritable? \_\_\_\_\_

Tire easily? \_\_\_\_\_

Do you feel relaxing? \_\_\_\_\_

Have any sexual problems? \_\_\_\_\_

DO YOU \_\_\_\_\_

Ever feel like you can't get enough sleep? \_\_\_\_\_

Feel bored with your work? \_\_\_\_\_

Use marijuanna? \_\_\_\_\_

Use "hard" drugs? \_\_\_\_\_

Do you want to talk to the doctor about a personal matter? \_\_\_\_\_

Yes \_\_\_\_\_ No \_\_\_\_\_

Printed on Recycled Paper

© 1976, 1983 Biomed Systems Inc. 1976-1983

Page 2

**PART E – BODY SYSTEMS REVIEW**

ME: Please answer questions 1 through 12, then ask your doctor to answer questions 13 through 20. Please start on question 6.

1. Do you have any pain? \_\_\_\_\_

2. Do you have any skin problems? \_\_\_\_\_

3. Have you ever had a fever? \_\_\_\_\_

4. Do you have any eye problems? \_\_\_\_\_

5. Do you have any ear problems? \_\_\_\_\_

6. Do you have any nose or throat problems? \_\_\_\_\_

7. Do you have any heart problems? \_\_\_\_\_

8. Do you have any lung problems? \_\_\_\_\_

9. Do you have any kidney problems? \_\_\_\_\_

10. Do you have any liver problems? \_\_\_\_\_

11. Do you have any gallbladder problems? \_\_\_\_\_

12. Do you have any bowel problems? \_\_\_\_\_

13. Do you have any bladder problems? \_\_\_\_\_

14. Do you have any blood problems? \_\_\_\_\_

15. Do you have any bone problems? \_\_\_\_\_

16. Do you have any muscle problems? \_\_\_\_\_

17. Do you have any joint problems? \_\_\_\_\_

18. Do you have any nerve problems? \_\_\_\_\_

19. Do you have any brain problems? \_\_\_\_\_

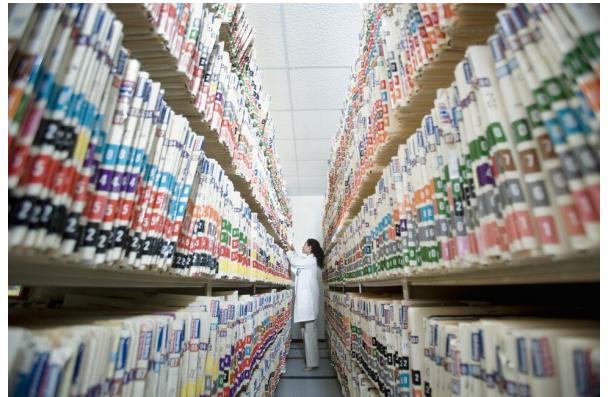
20. Do you have any spinal problems? \_\_\_\_\_

**CONFIDENTIAL**

Printed on Recycled Paper

© 1976, 1983 Biomed Systems Inc. 1976-1983

Page 3

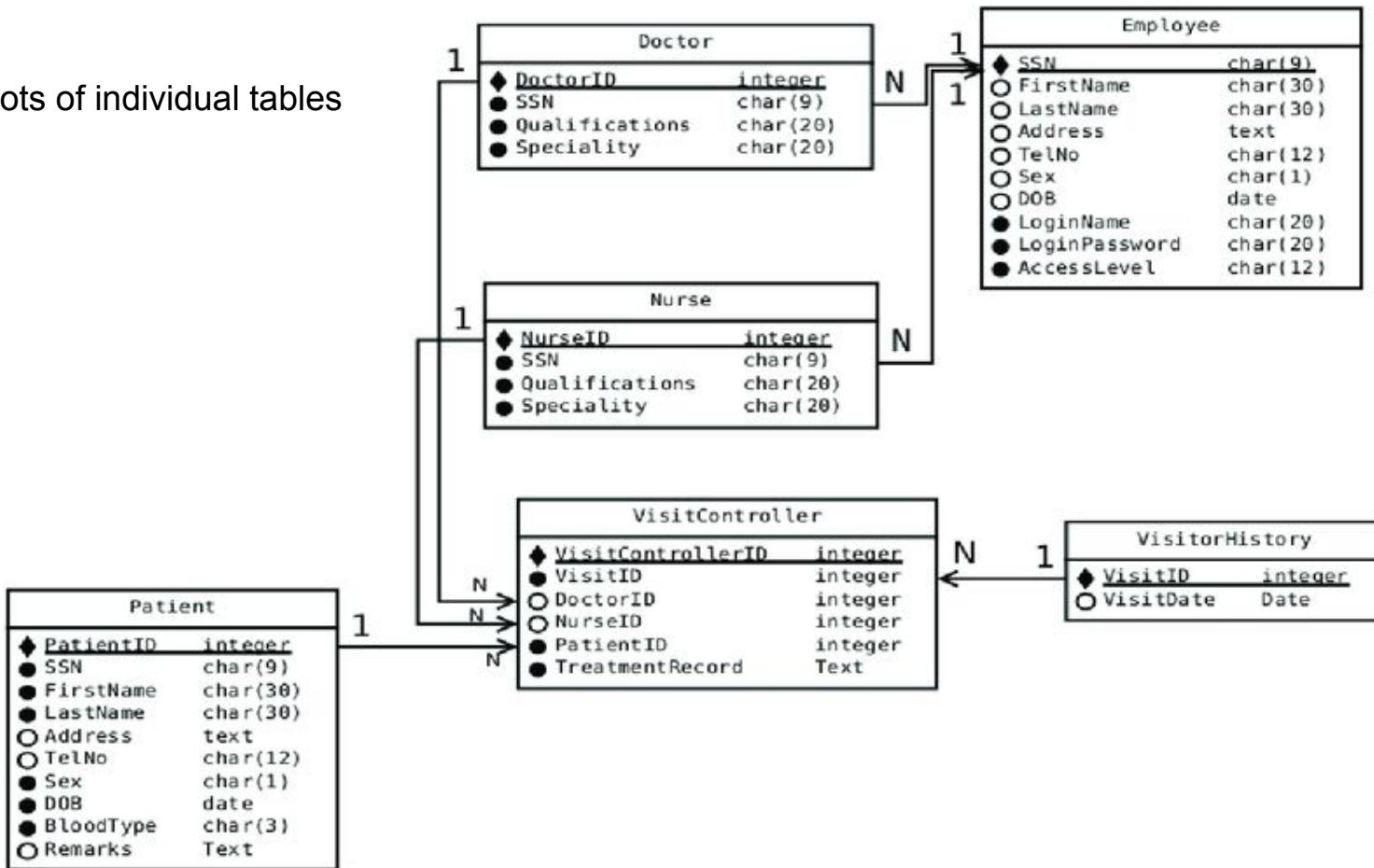


Organisation make some tasks easier/harder:

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

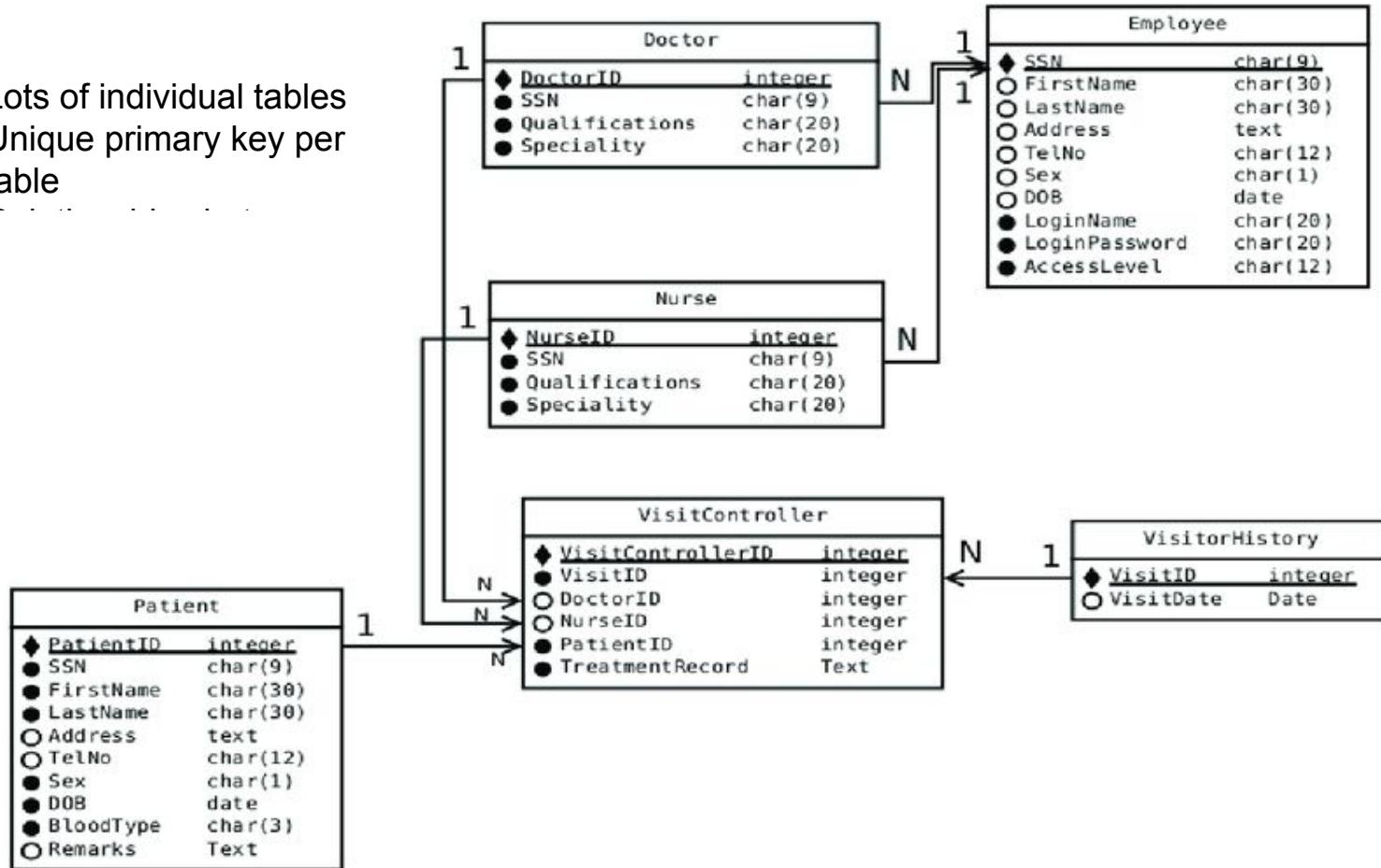
# Most Common Type: Relational Databases

- Lots of individual tables



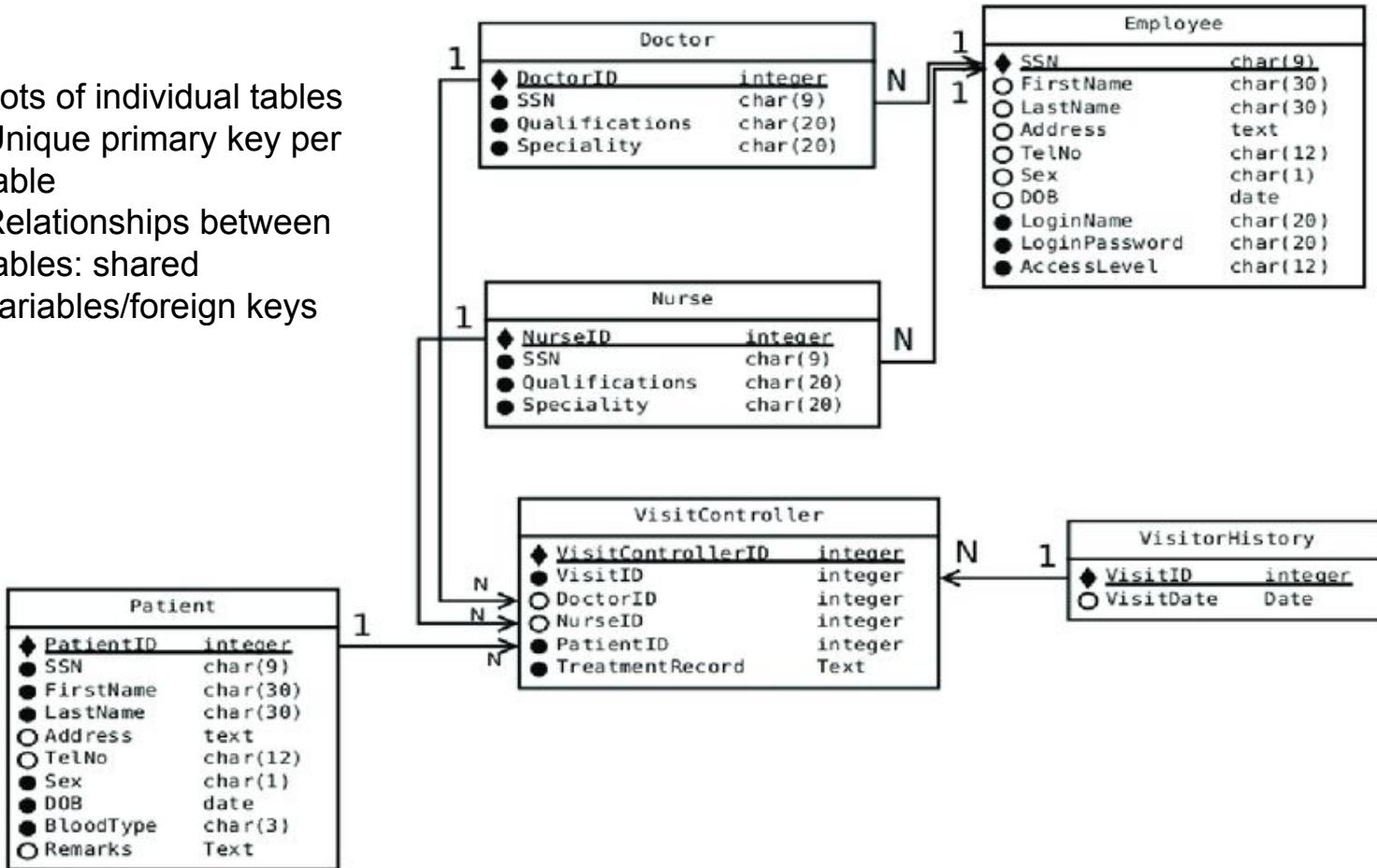
# Most Common Type: Relational Databases

- Lots of individual tables
- Unique primary key per table



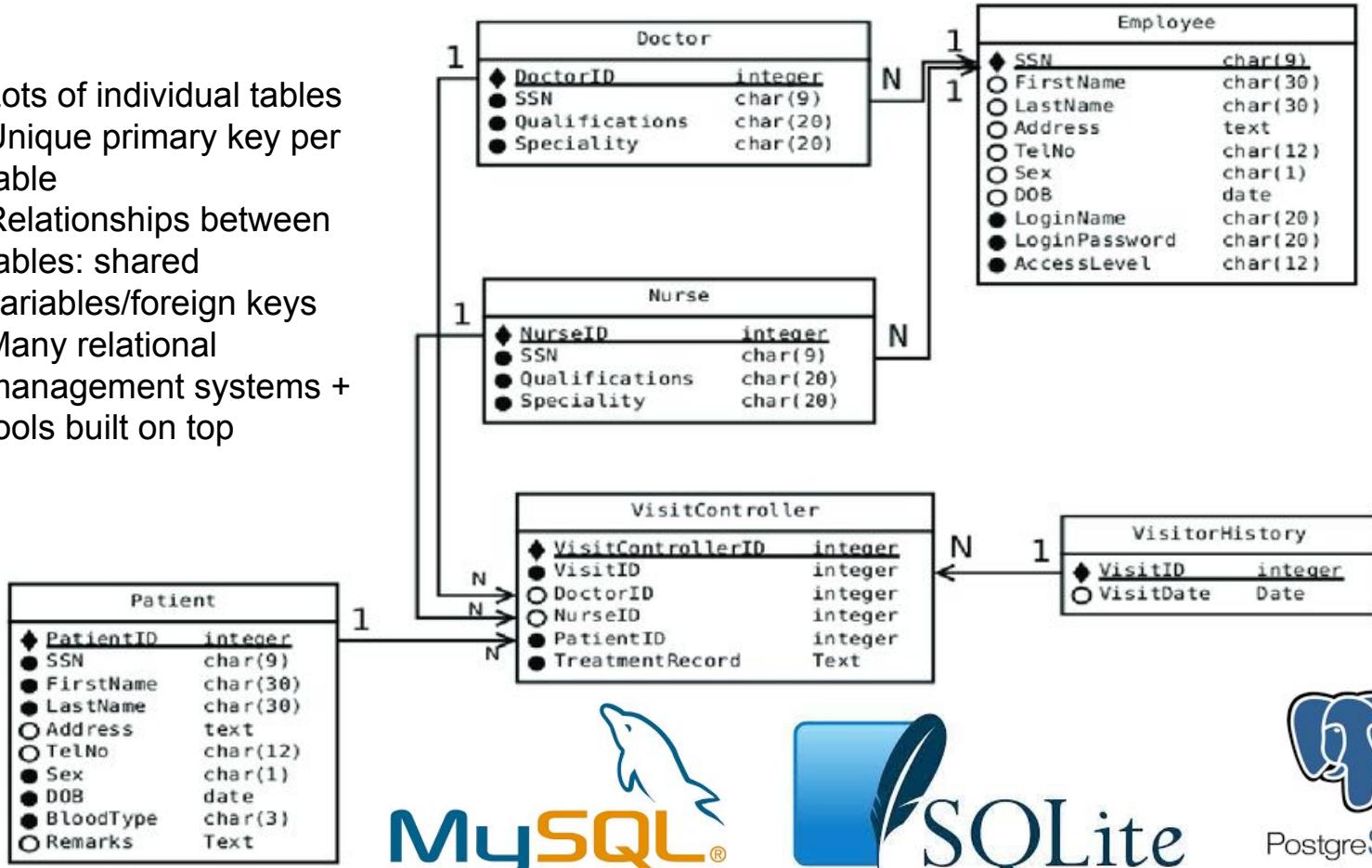
# Most Common Type: Relational Databases

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



# 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



# Queried using Structured Query Language (SQL)

- Non-procedural Language
- Standardised/powerful/flexible

# Queried using Structured Query Language (SQL)

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

# Queried using Structured Query Language (SQL)

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

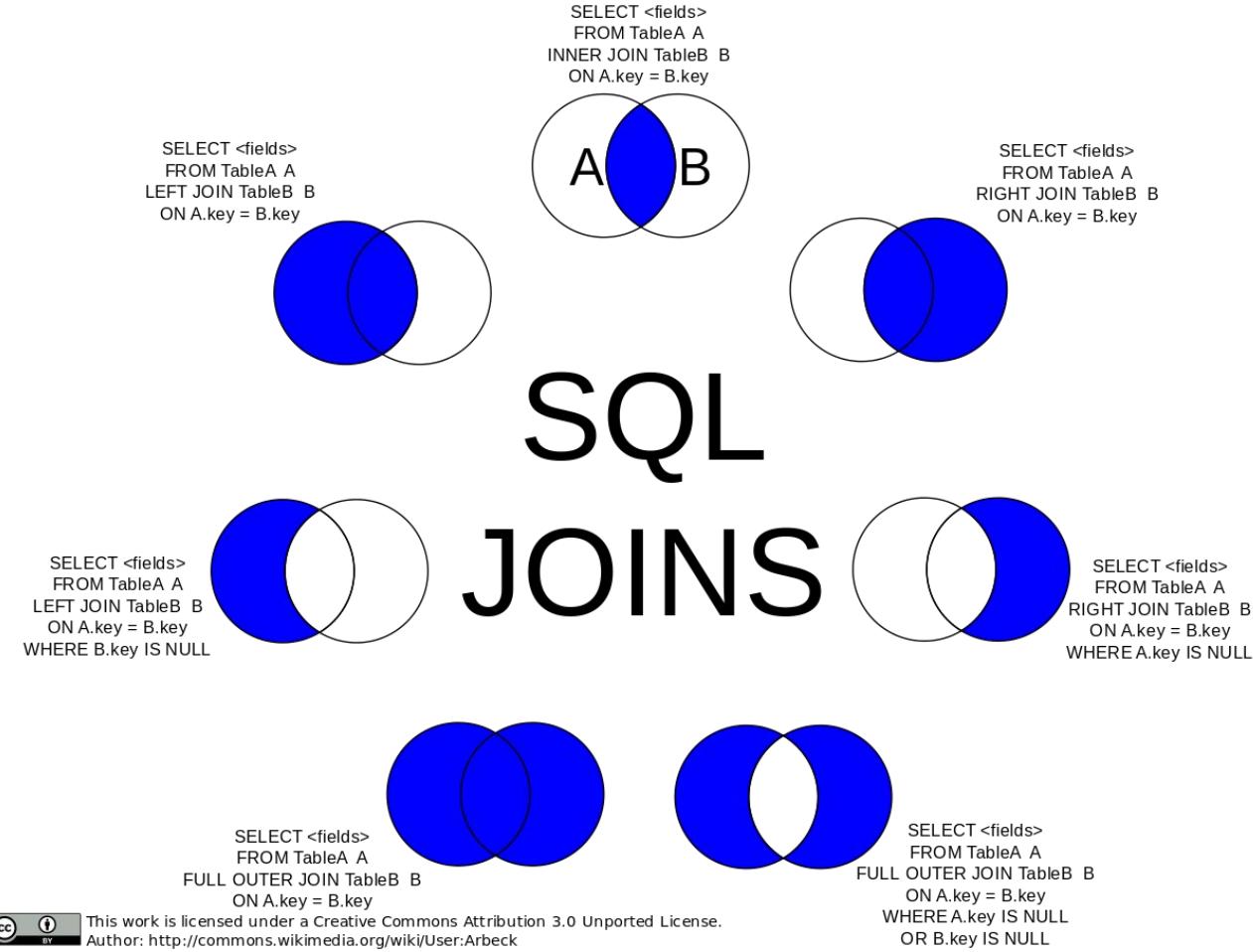
```
flights %>%
  select(contains("delay")) %>%
  show_query()
#> <SQL>
#> SELECT `dep_delay`, `arr_delay`
#> FROM `nycflights13::flights`
```

```
flights %>%
  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`)
```

```
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`
```

# SQL enables complex joins/queries

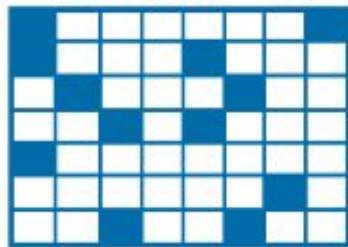


Are all databases relational?

# Non-Relational Databases

- Less common than relational in medicine

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



**Column based**

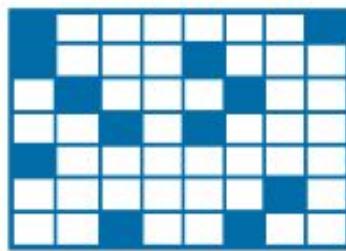


Google  
Big Query

# Non-Relational Databases

- Less common than relational in medicine

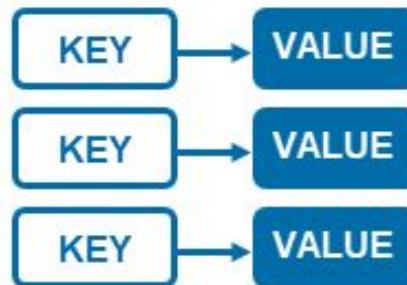
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



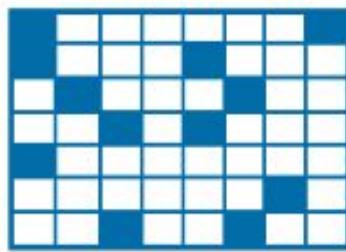
Key-value



# Non-Relational Databases

- Less common than relational in medicine

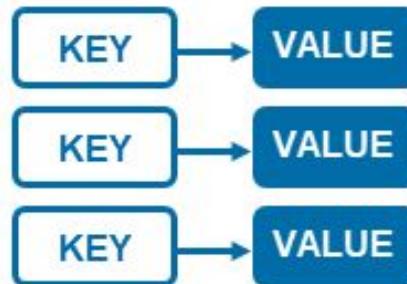
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



Key-value



Graph



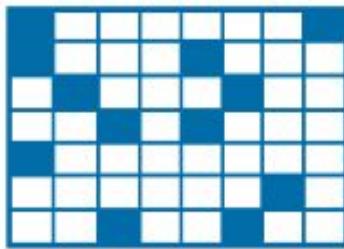
# Non-Relational Databases

- Less common than relational in medicine
- Querying can be... very easy or very complicated

Find me the homepage of anyone known by Tim Berners-Lee.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

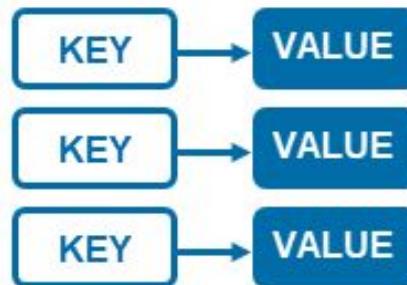
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



Key-value



redis



Graph



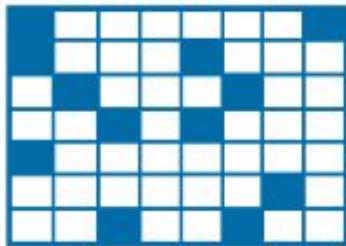
# Non-Relational Databases

- Less common than relational in medicine
- Querying can be... very easy or very complicated

Find me the homepage of anyone known by Tim Berners-Lee.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

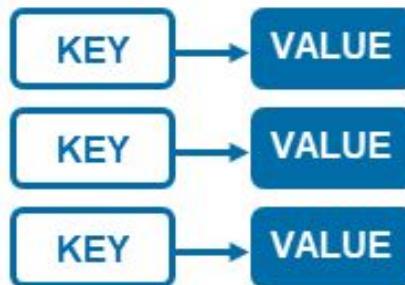
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



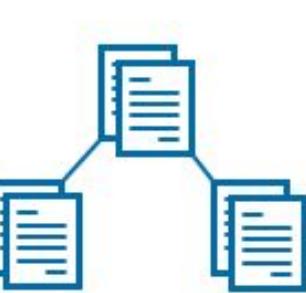
Key-value



redis



Graph



Document



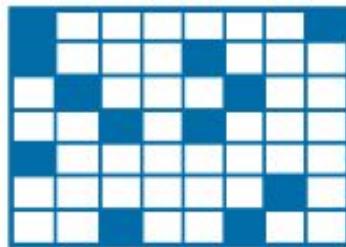
# Non-Relational Databases

- Less common than relational in medicine
- Querying can be... very easy or very complicated
- Mostly for very large datasets:
  - User data / security audit data
  - Medical image data

Find me the homepage of anyone known by Tim Berners-Lee.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

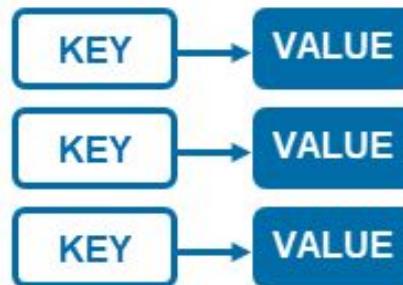
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



Key-value



redis



Graph



neo4j



Document



mongoDB

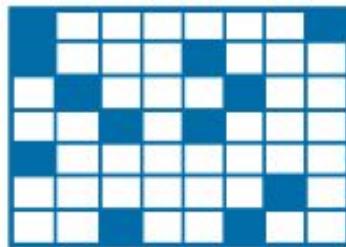
# Non-Relational Databases

- Less common than relational in medicine
- Querying can be... very easy or very complicated
- Mostly for very large datasets:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies

Find me the homepage of anyone known by Tim Berners-Lee.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

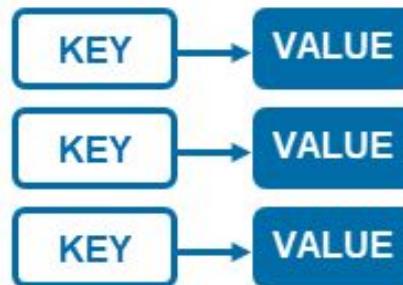
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



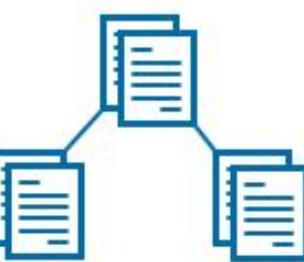
Key-value



redis



Graph



Document



mongoDB

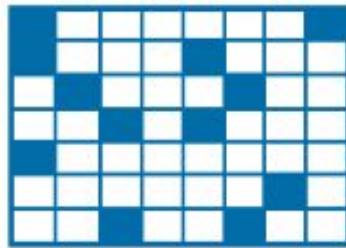
# Non-Relational Databases

- Less common than relational in medicine
- Querying can be... very easy or very complicated
- Mostly for very large datasets:
  - User data / security audit data
  - Medical image data
- Or unusual data structures:
  - Contact tracing
  - Ontologies
- Or both:
  - Social media data

Find me the homepage of anyone known by Tim Berners-Lee.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

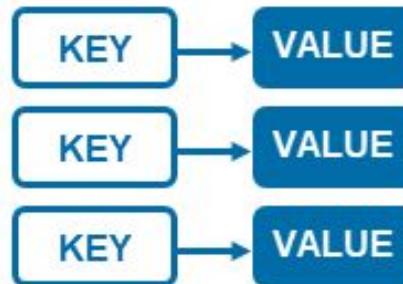
<https://phoenixnap.com/kb/database-types>



Column based



Google  
Big Query



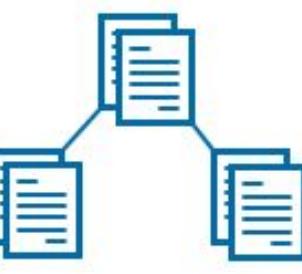
Key-value



redis



Graph



Document



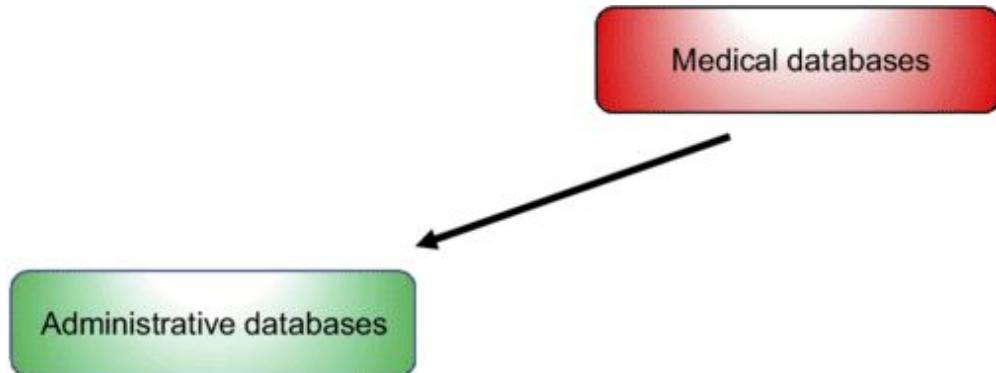
What are medical databases?

# Many types of database

Medical databases

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

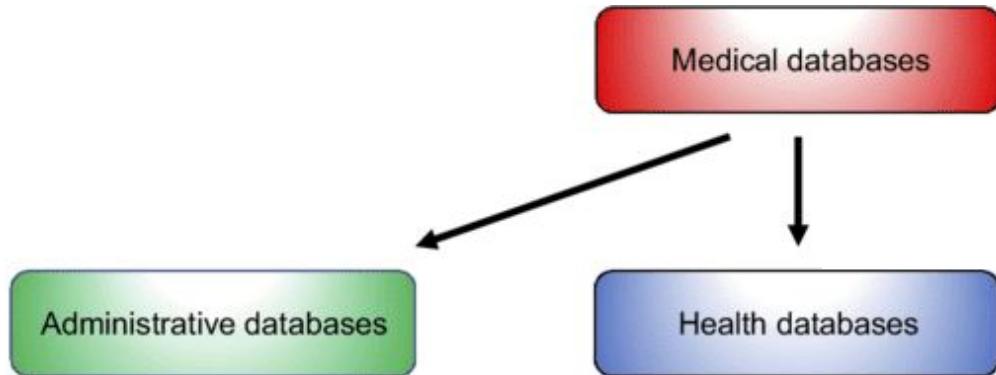
# Many types of database



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

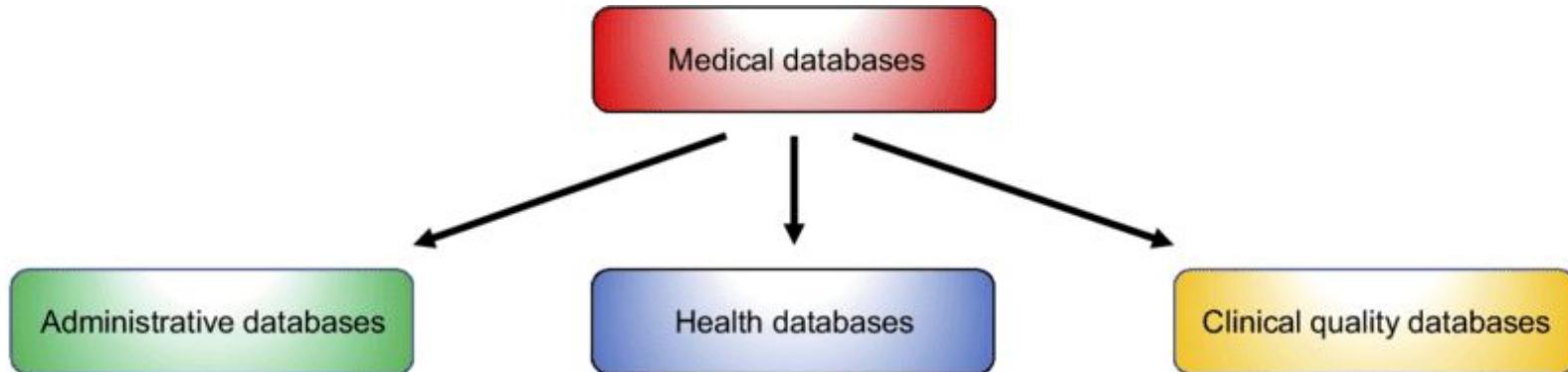
■ Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic

# Many types of database



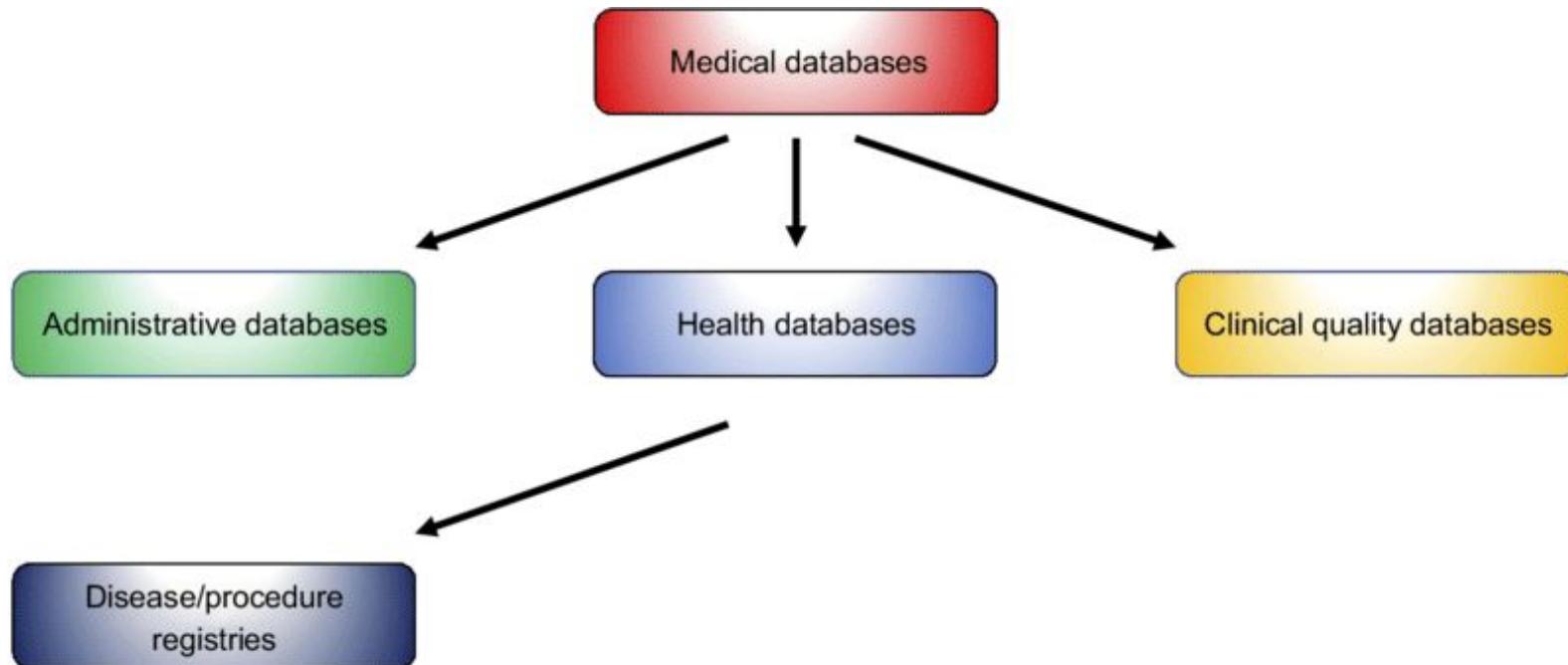
- All types of registries and databases that contain health-related data
- Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic
- Register health data for the purpose of surveillance and research

# Many types of database



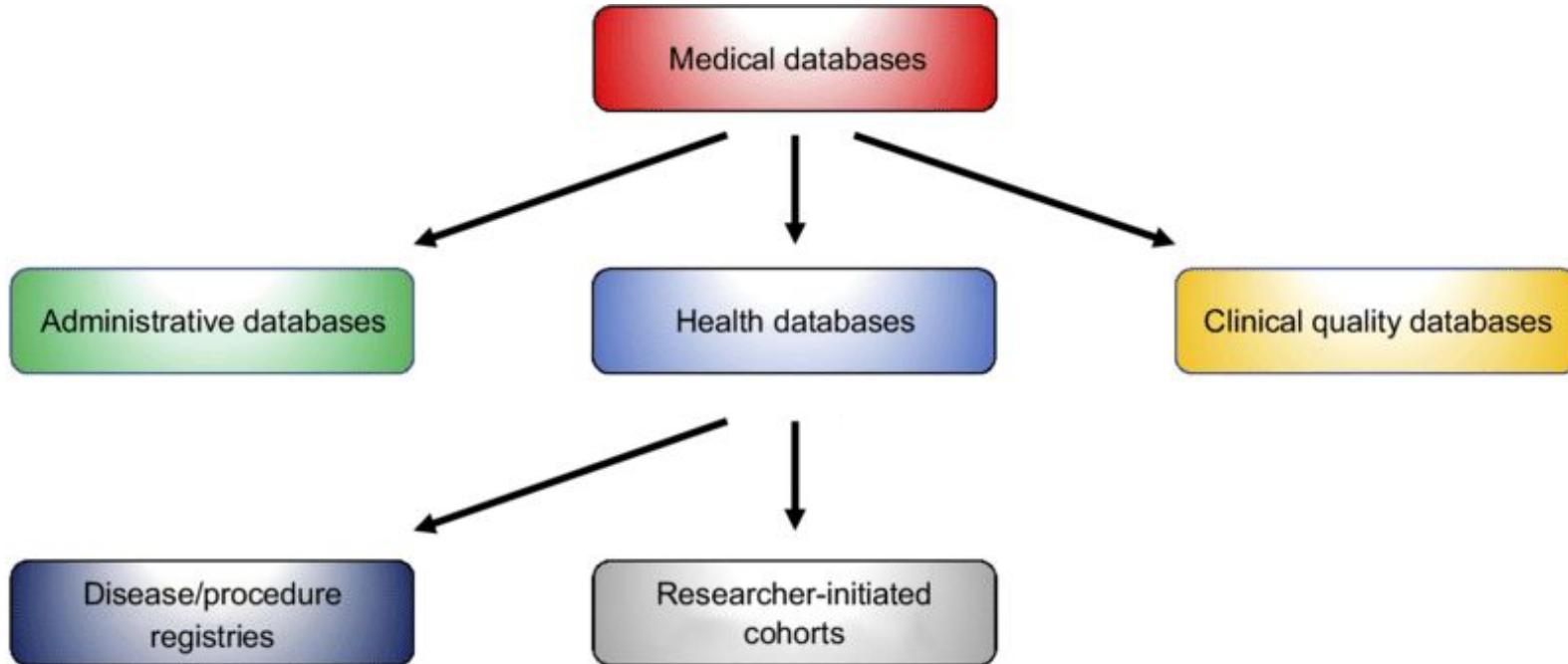
- All types of registries and databases that contain health-related data
- Register individuals according to geographic area, health insurance program, or attendance at a particular hospital or clinic
- Register health data for the purpose of surveillance and research
- Register detailed clinical data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

# 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

# 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

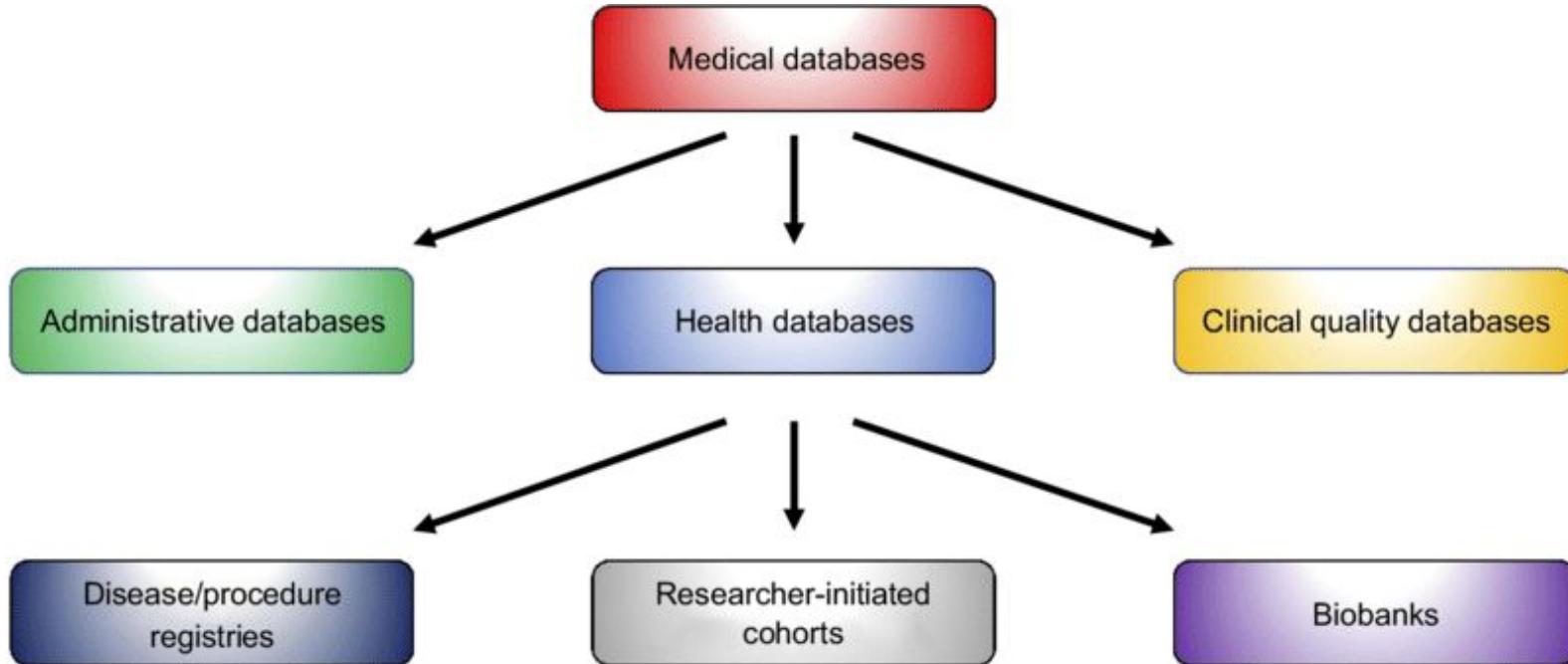
■ Register health data for the purpose of surveillance and research

■ 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)

# 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)

# 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

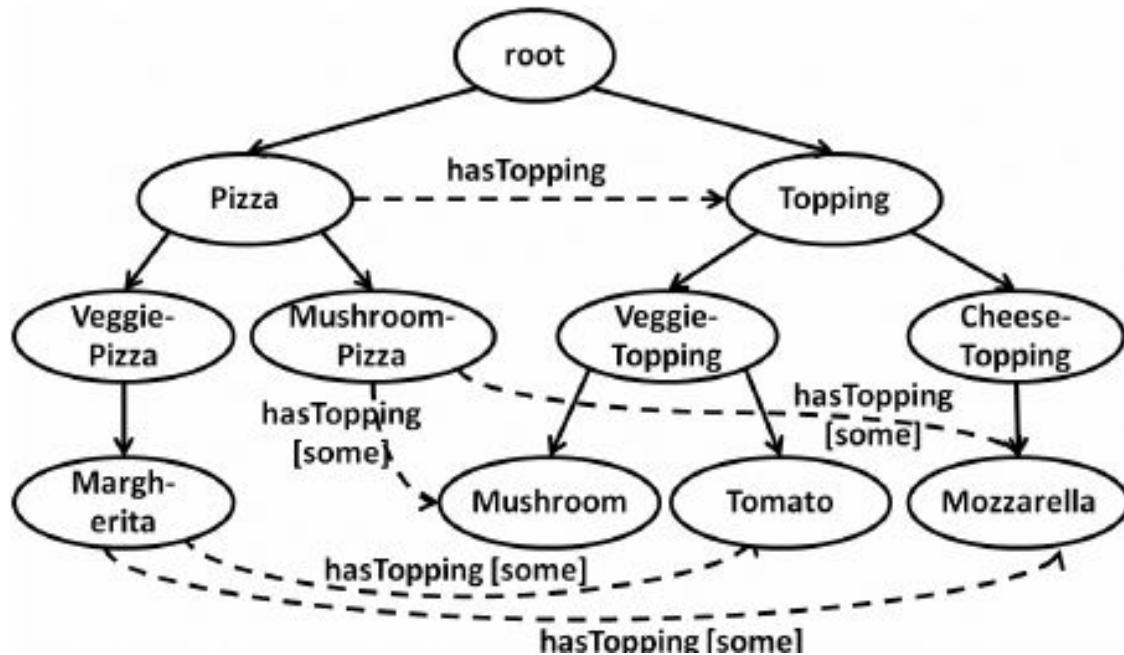
Generalisability



How do medical databases try to handle standardisation?

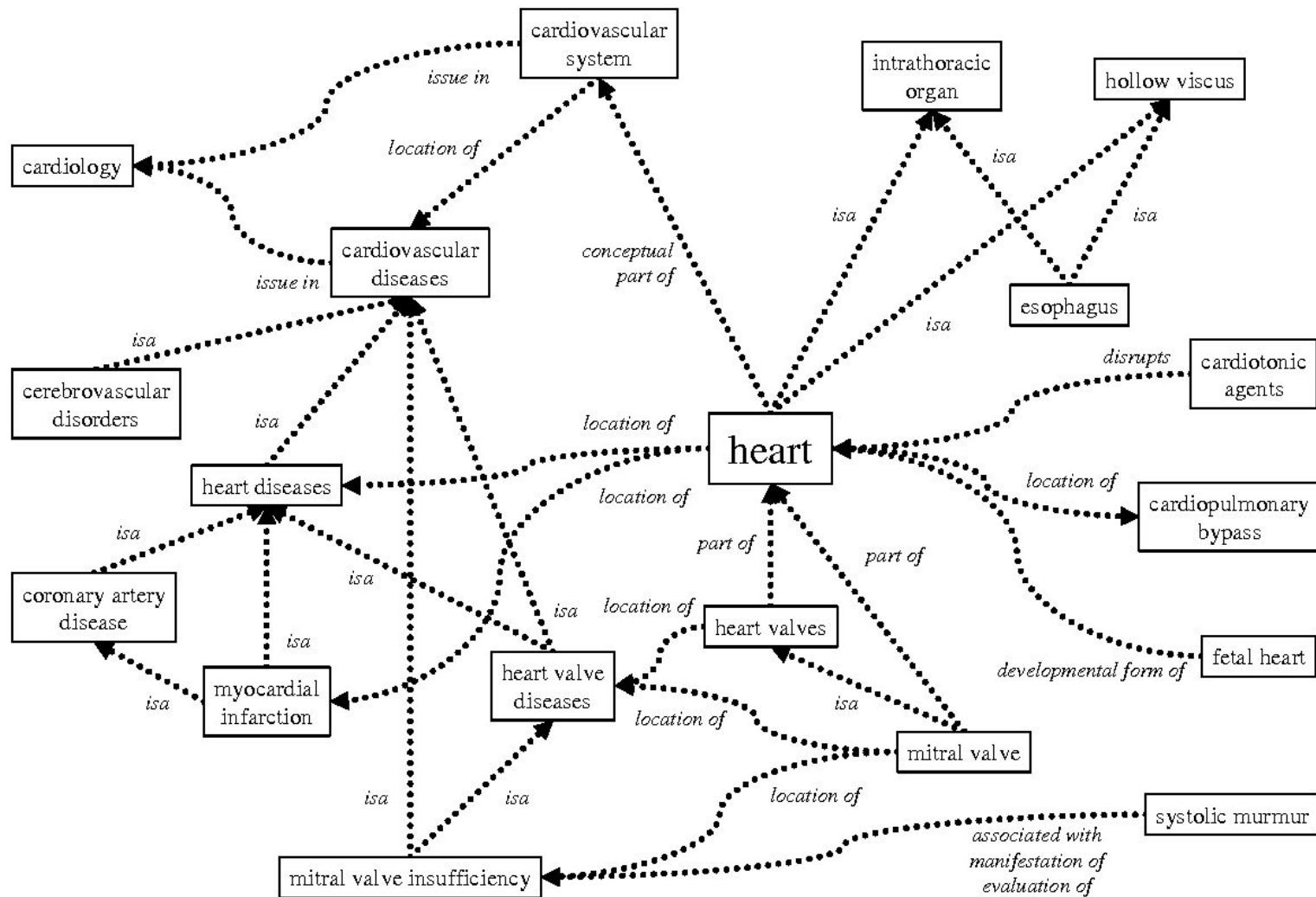
# Ontologies for standardisation

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

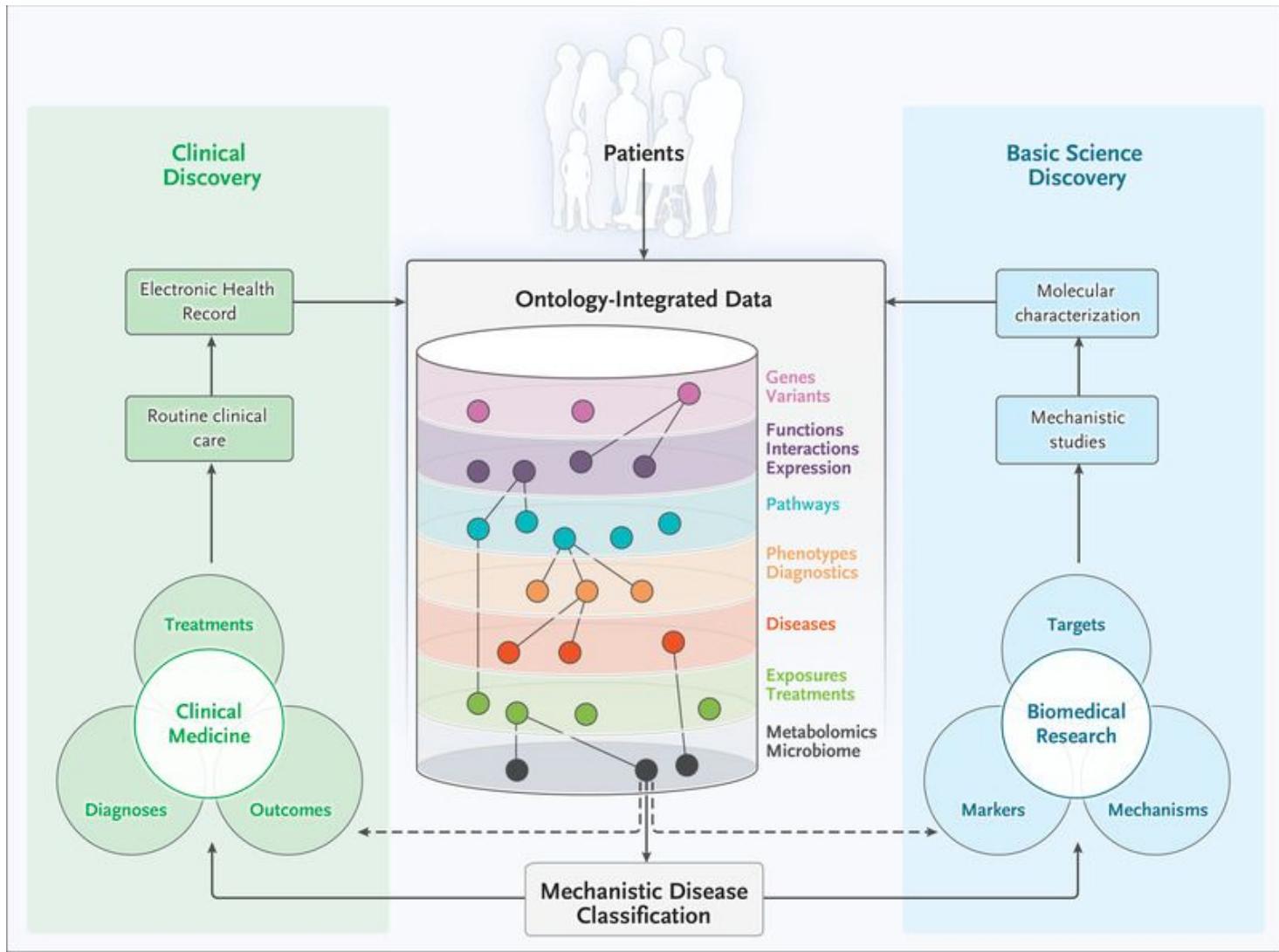


<https://www.researchgate.net/figure/Example-pizza-ontology-represented-as-a-graph-G-a-and-a-changed-versi>  
on-of-the-pizza\_fig1\_236842047

# Medical Ontologies



# Ontologies for linking diverse types of data



# 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 <ul style="list-style-type: none"><li>• 3 -5 characters</li><li>• First character is numeric or alpha</li><li>• Characters 2-5 are numeric</li></ul>	ICD-10-CM <ul style="list-style-type: none"><li>• 3 -7 characters</li><li>• Character 1 is alpha</li><li>• Character 2 is numeric</li><li>• Characters 3 – 7 can be alpha or numeric</li></ul>
Procedure Structure	ICD-9-CM <ul style="list-style-type: none"><li>• 3-4 characters</li><li>• All characters are numeric</li><li>• All codes have at least 3 characters</li></ul>	ICD-10-PCS <ul style="list-style-type: none"><li>• ICD-10-PCS has 7 characters</li><li>• Each can be either alpha or numeric</li><li>• Numbers 0-9; letters A-H, J-N, P-Z</li></ul>

[https://www.cdc.gov/nchs/icd/icd10cm\\_pcs\\_background.htm](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 <ul style="list-style-type: none"><li>• 3 -5 characters</li><li>• First character is numeric or alpha</li><li>• Characters 2-5 are numeric</li></ul>	ICD-10-CM <ul style="list-style-type: none"><li>• 3 -7 characters</li><li>• Character 1 is alpha</li><li>• Character 2 is numeric</li><li>• Characters 3 – 7 can be alpha or numeric</li></ul>
Procedure Structure	ICD-9-CM <ul style="list-style-type: none"><li>• 3-4 characters</li><li>• All characters are numeric</li><li>• All codes have at least 3 characters</li></ul>	ICD-10-PCS <ul style="list-style-type: none"><li>• ICD-10-PCS has 7 characters</li><li>• Each can be either alpha or numeric</li><li>• Numbers 0-9; letters A-H, J-N, P-Z</li></ul>

[https://www.cdc.gov/nchs/icd/icd10cm\\_pcs\\_backround.htm](https://www.cdc.gov/nchs/icd/icd10cm_pcs_backround.htm)

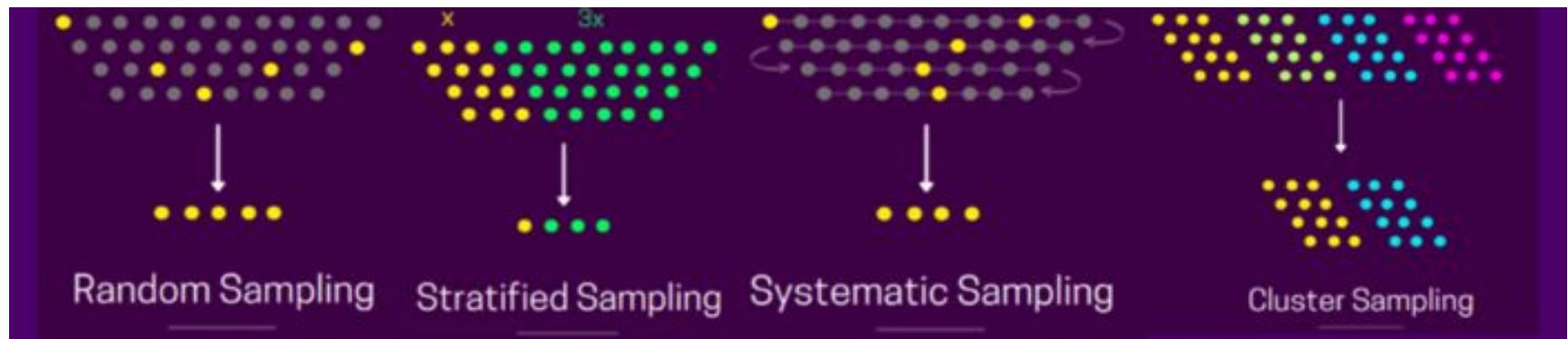
How do we sample for medical databases?

# Sampling strategy

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

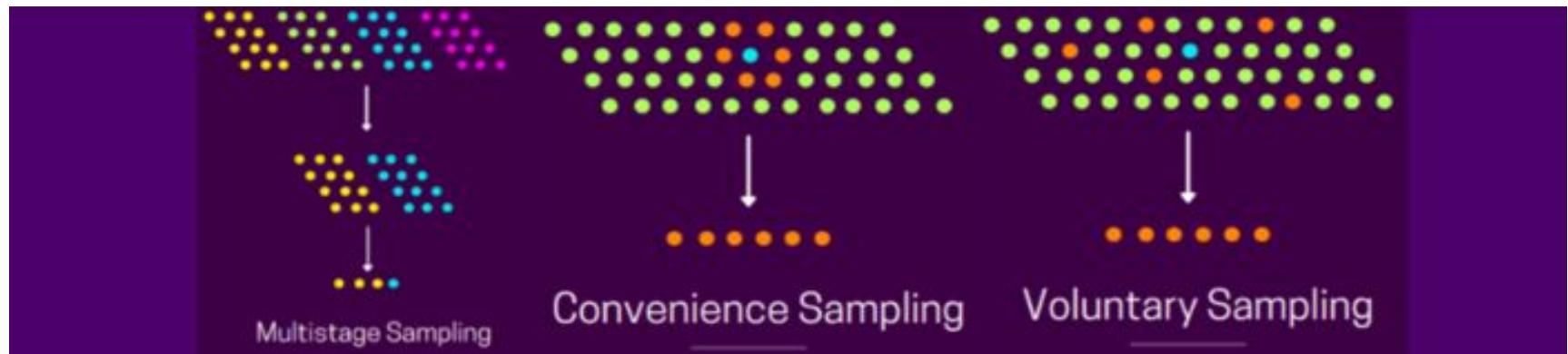
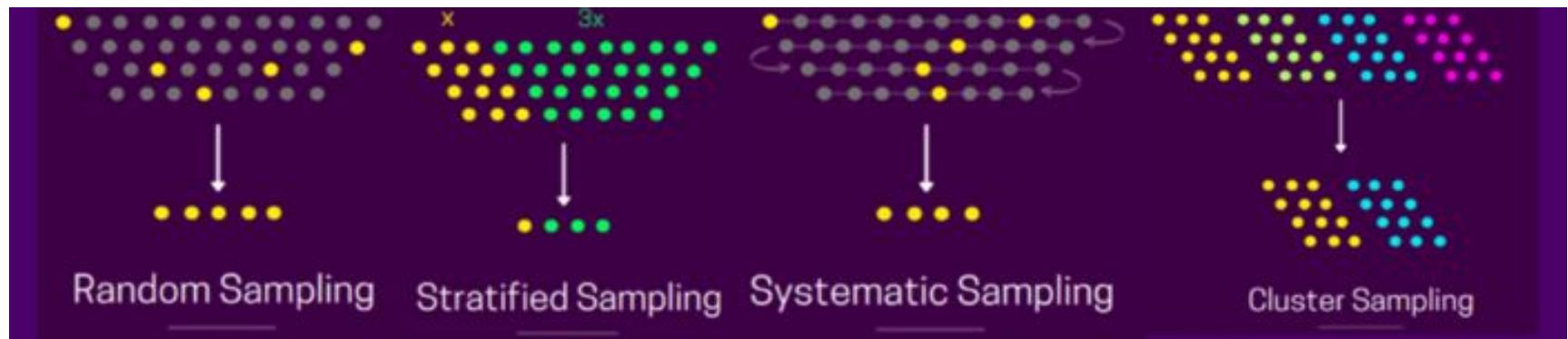
# Sampling strategy

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



# Sampling strategy

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

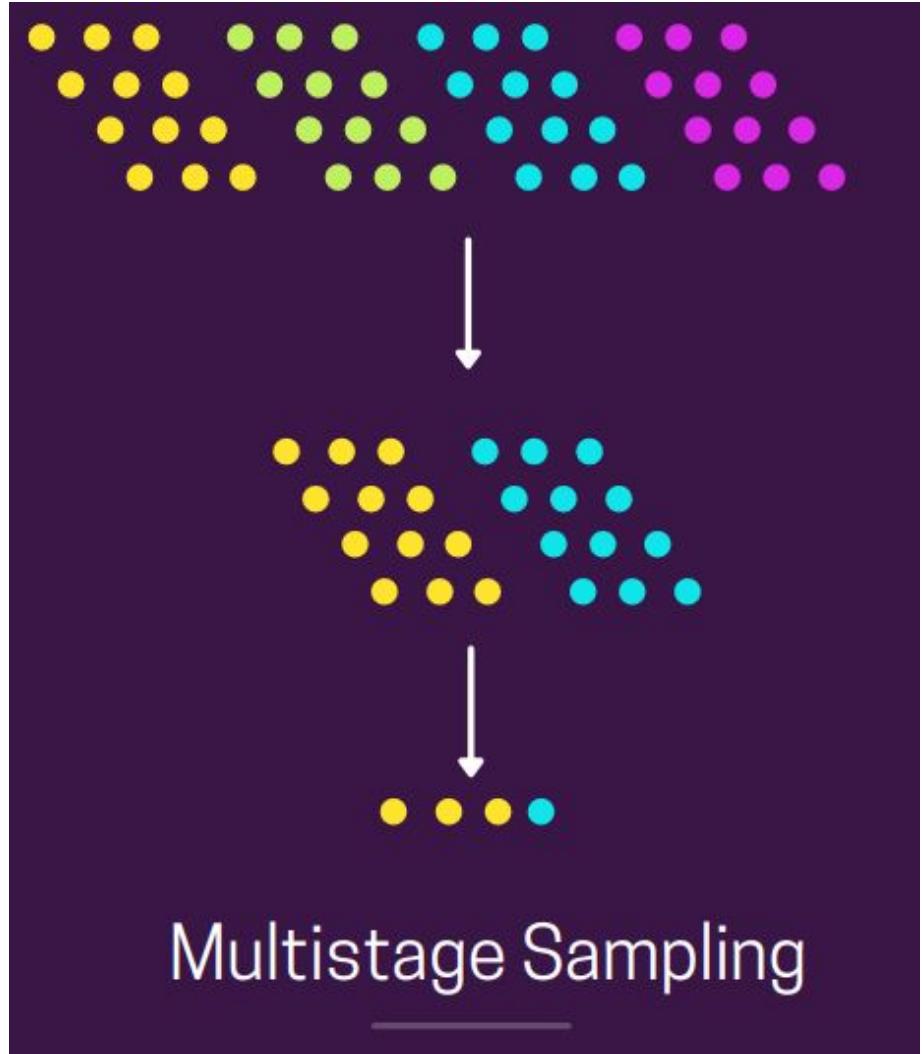


# Survey weights

- 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.
- Often not directly supported by machine learning libraries (`sample_weights` implemented for some models)

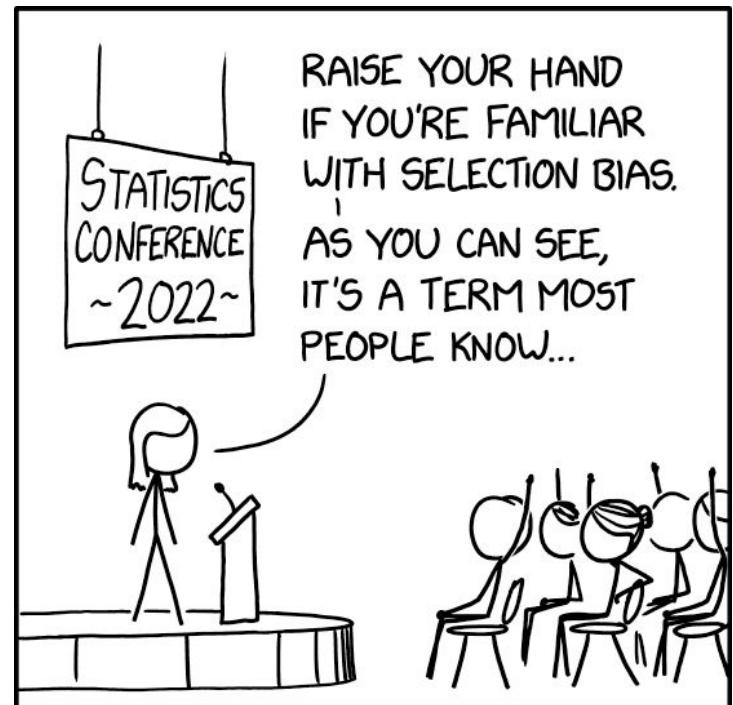


# Types of weights

- 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

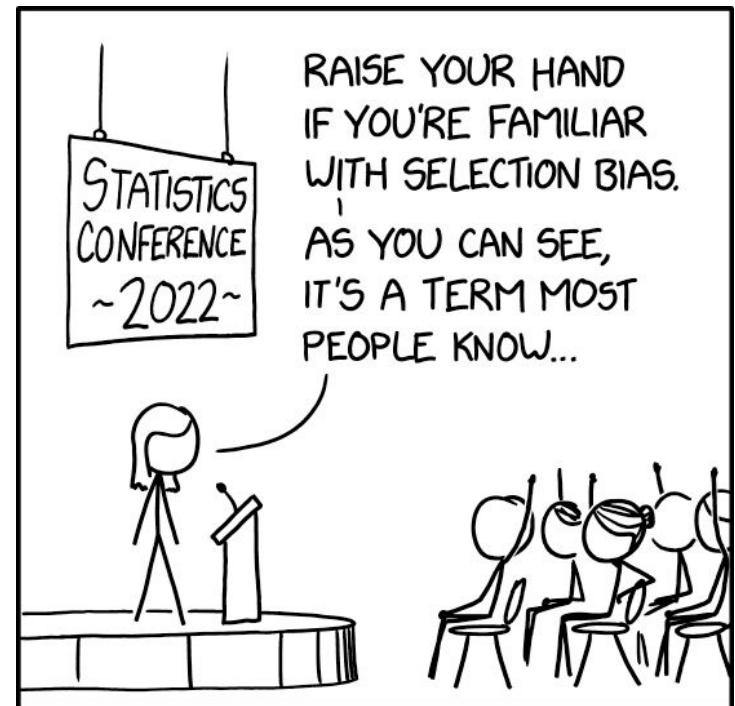
# Types of weights

- 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



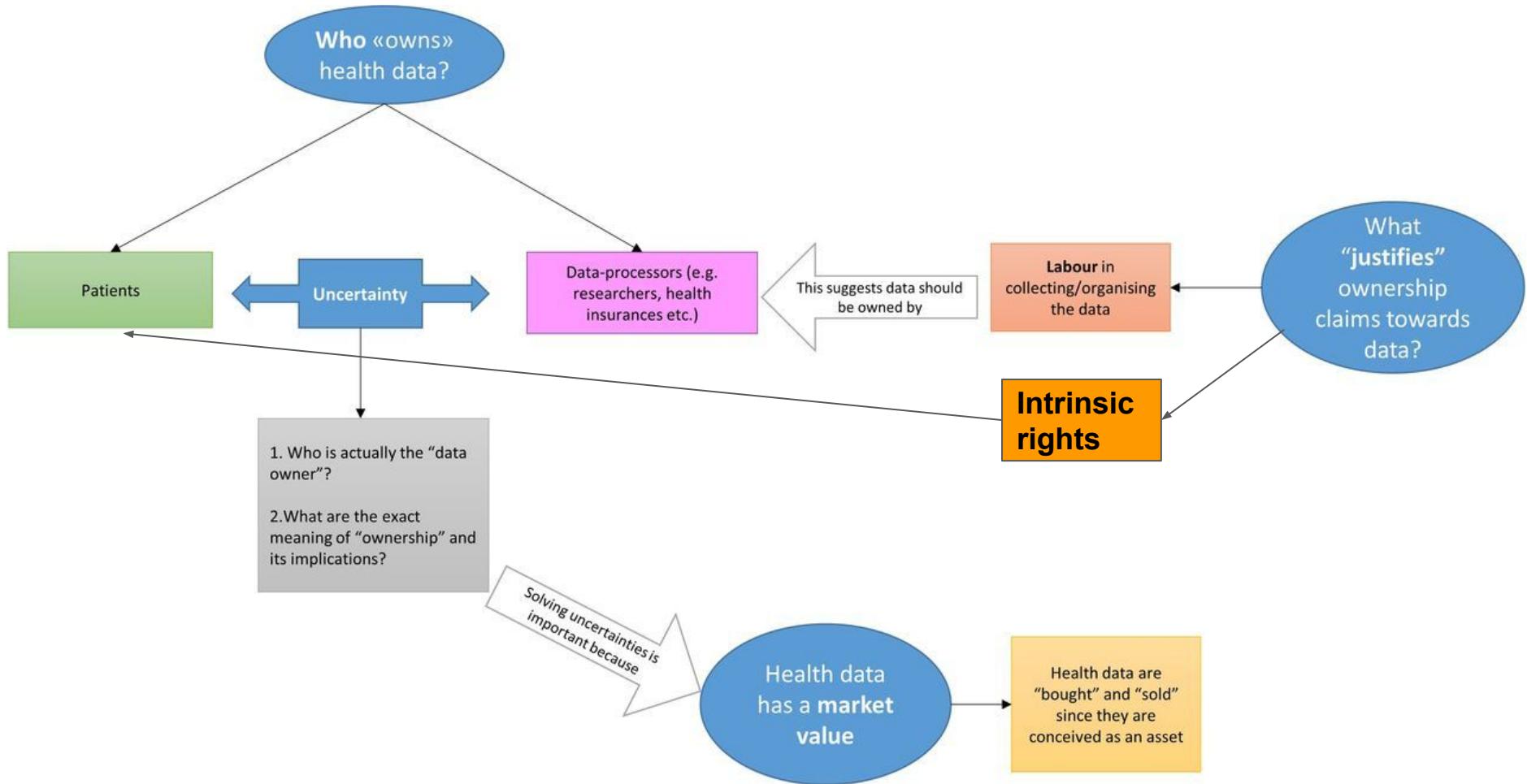
# Types of weights

- 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:



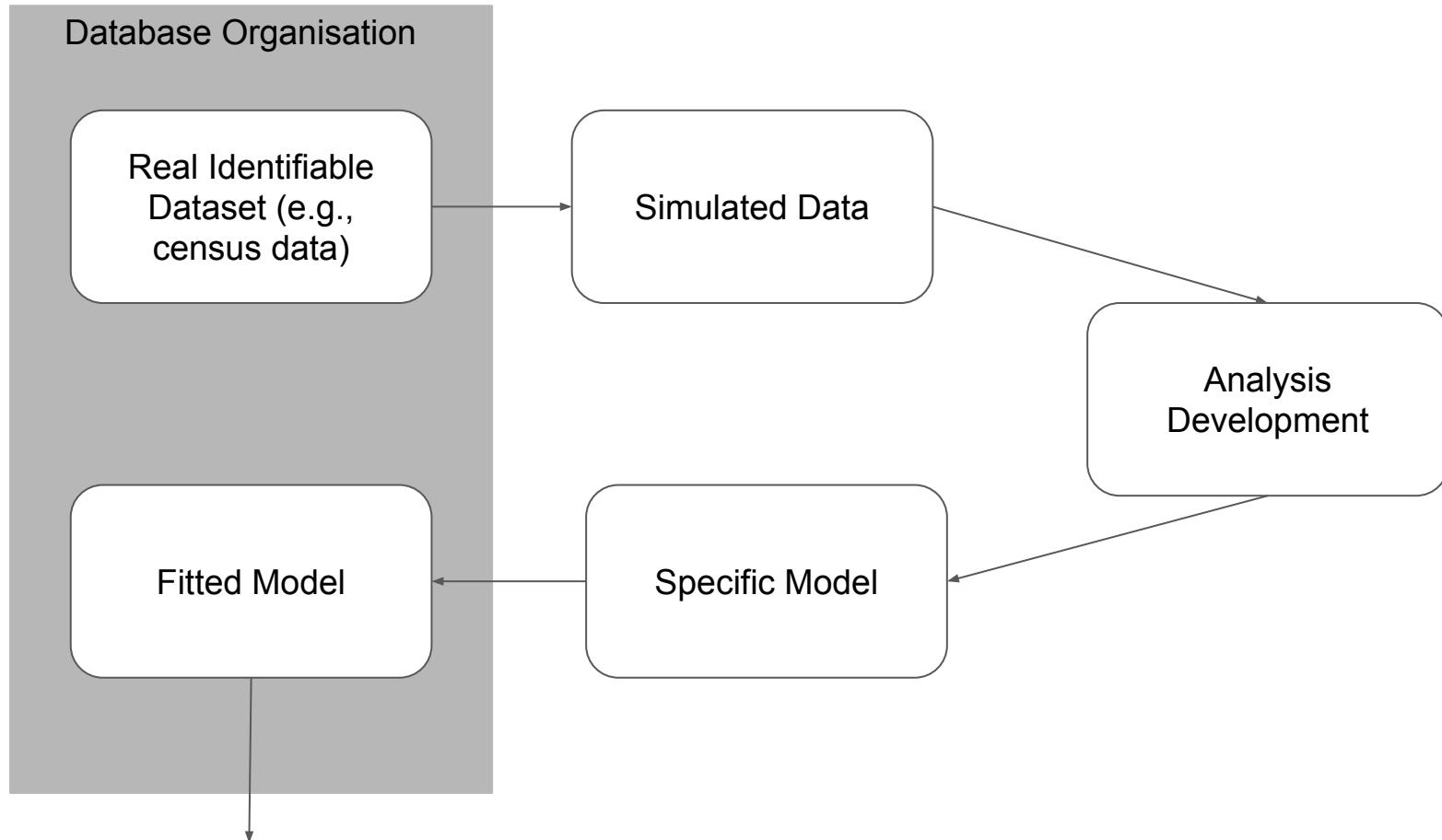
Who actually owns this data?

# Data Ownership is Difficult



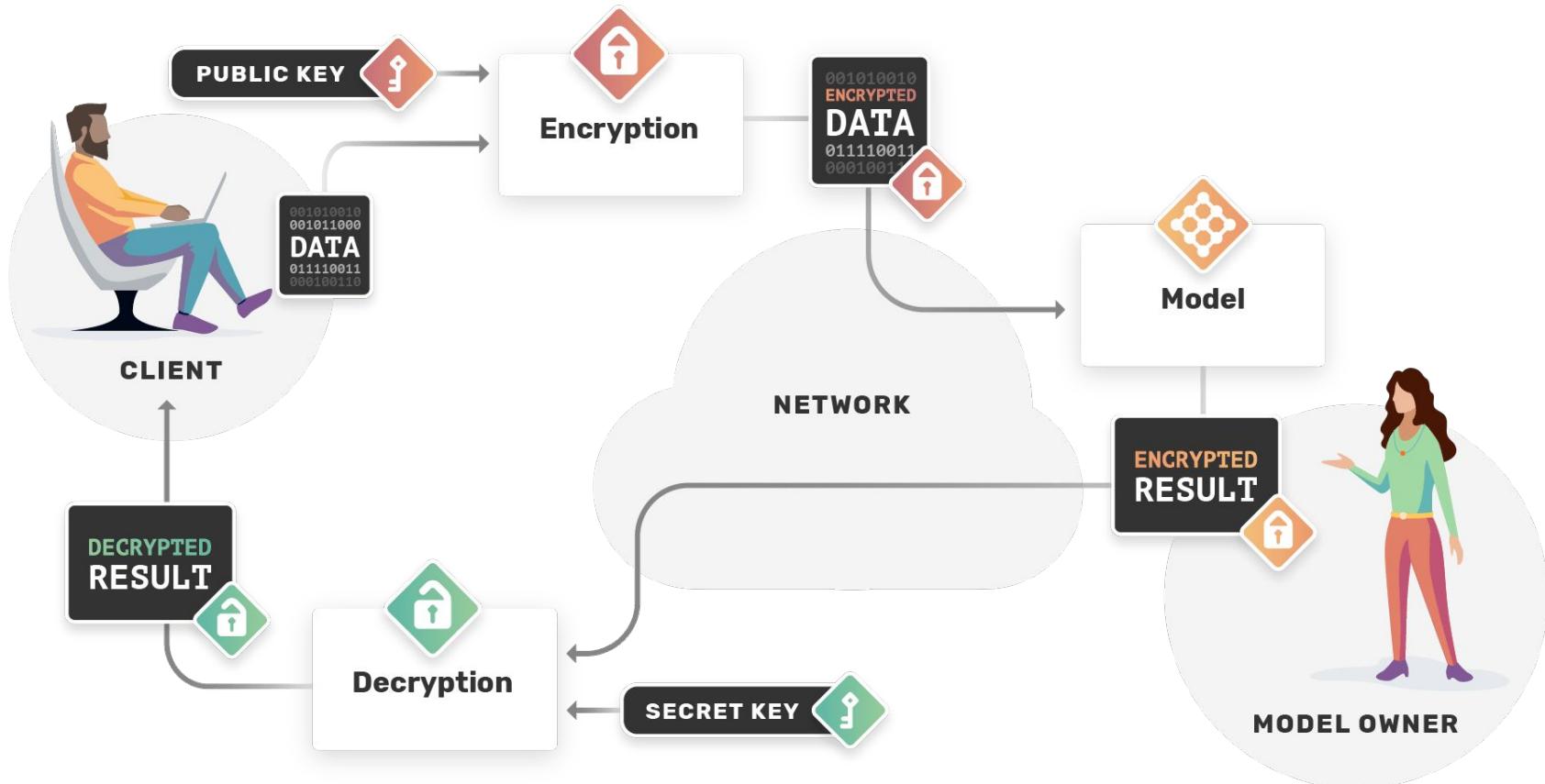
How do you protect privacy in these databases?

# No direct data access



# Shared data but encrypted: homomorphic encryption

Partial to fully homomorphic encryption



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

# Data privacy is a continuum

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE
DIRECT IDENTIFIERS			
INDIRECT IDENTIFIERS			
SAFEGUARDS and CONTROLS			
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)

# Indirectly identifiable: Pseudonymous Data

	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS
<b>DIRECT IDENTIFIERS</b> Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)			
<b>INDIRECT IDENTIFIERS</b> Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)			
<b>SAFEGUARDS and CONTROLS</b> Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals			
	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls

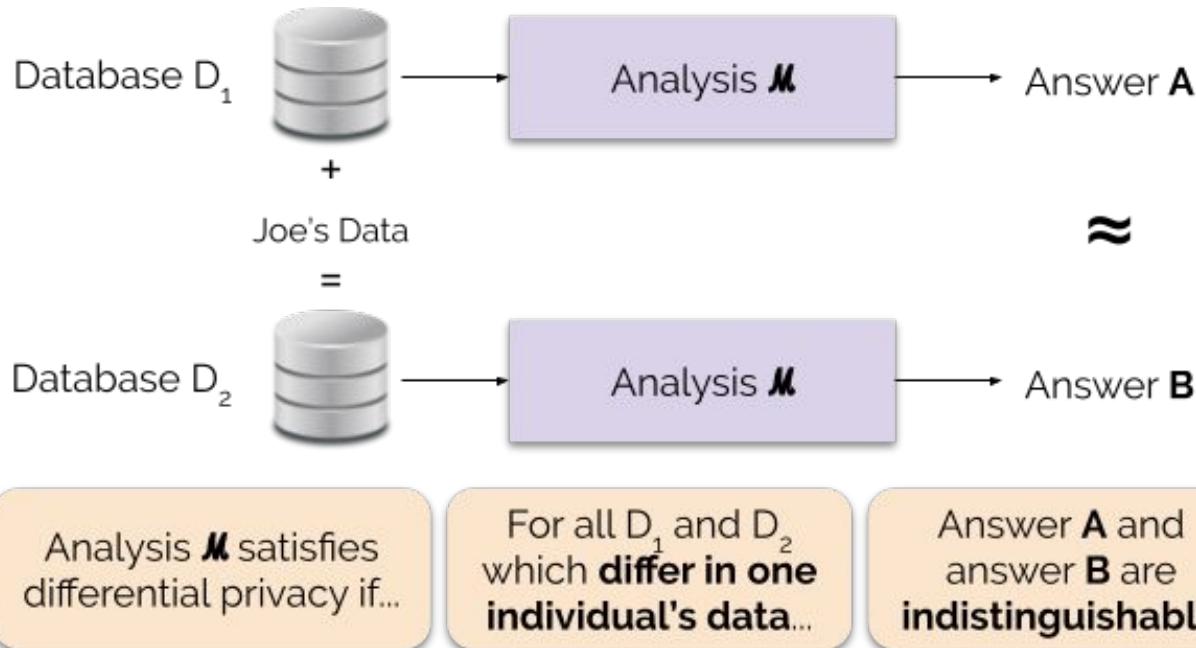
# Identifiers removed/broken: De-Identified Data

	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED
DIRECT IDENTIFIERS		
INDIRECT IDENTIFIERS		
SAFEGUARDS and CONTROLS		
	LIMITED or NONE IN PLACE	CONTROLS IN PLACE
	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls

# Non-identifiability Guarantee: Anonymous Data

	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS		
INDIRECT IDENTIFIERS		
SAFEGUARDS and CONTROLS		
Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals		
 ELIMINATED or TRANSFORMED		 ELIMINATED or TRANSFORMED
 ELIMINATED or TRANSFORMED		 ELIMINATED or TRANSFORMED
 NOT RELEVANT due to nature of data		 NOT RELEVANT due to high degree of data aggregation
For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)		Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

# Differential privacy: no singling out individuals



# Differential privacy: no singling out individuals



Analysis  $M$  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  $D_1$

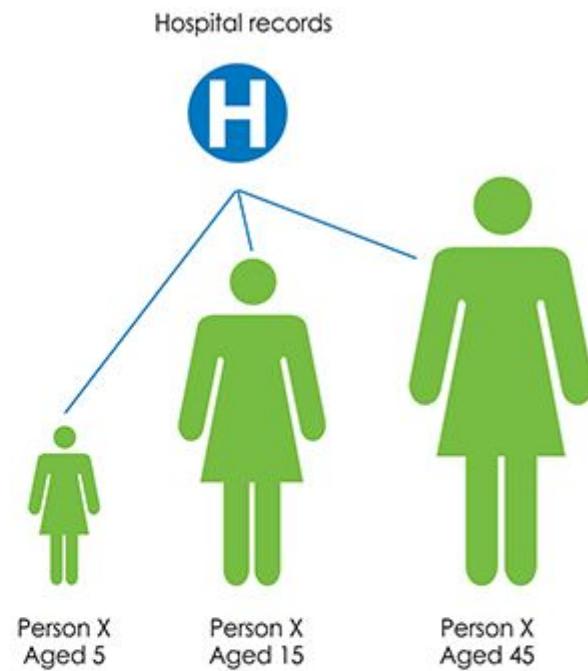
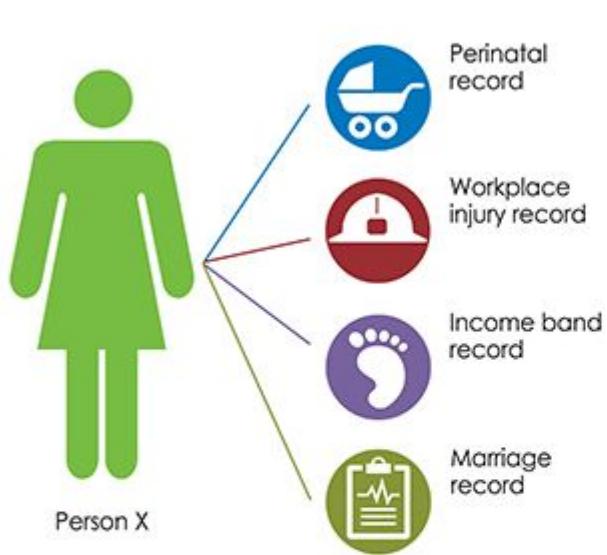
$$\frac{\Pr[M(D_1) \in O]}{\Pr[M(D_2) \in O]} \leq e^\epsilon$$

Probability of seeing output  $O$  on input  $D_2$

Indistinguishability:  
bounded ratio of probabilities

# 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

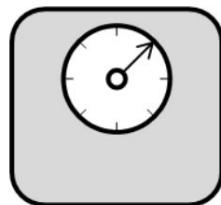
**Let's take a short break!**

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

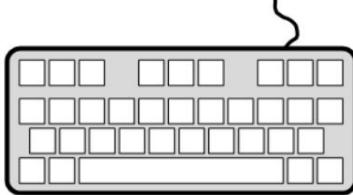
# Data Cleaning: even “simple” fields can be a nightmare

## Data Quality

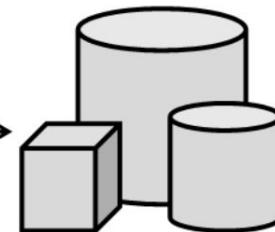
Actual value:  
200.6 lbs.



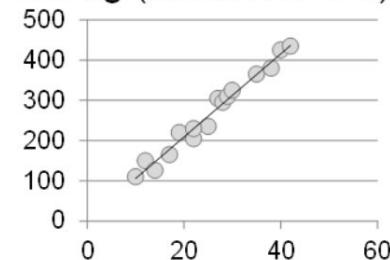
Recorded value:  
200 lbs.



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
- 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*

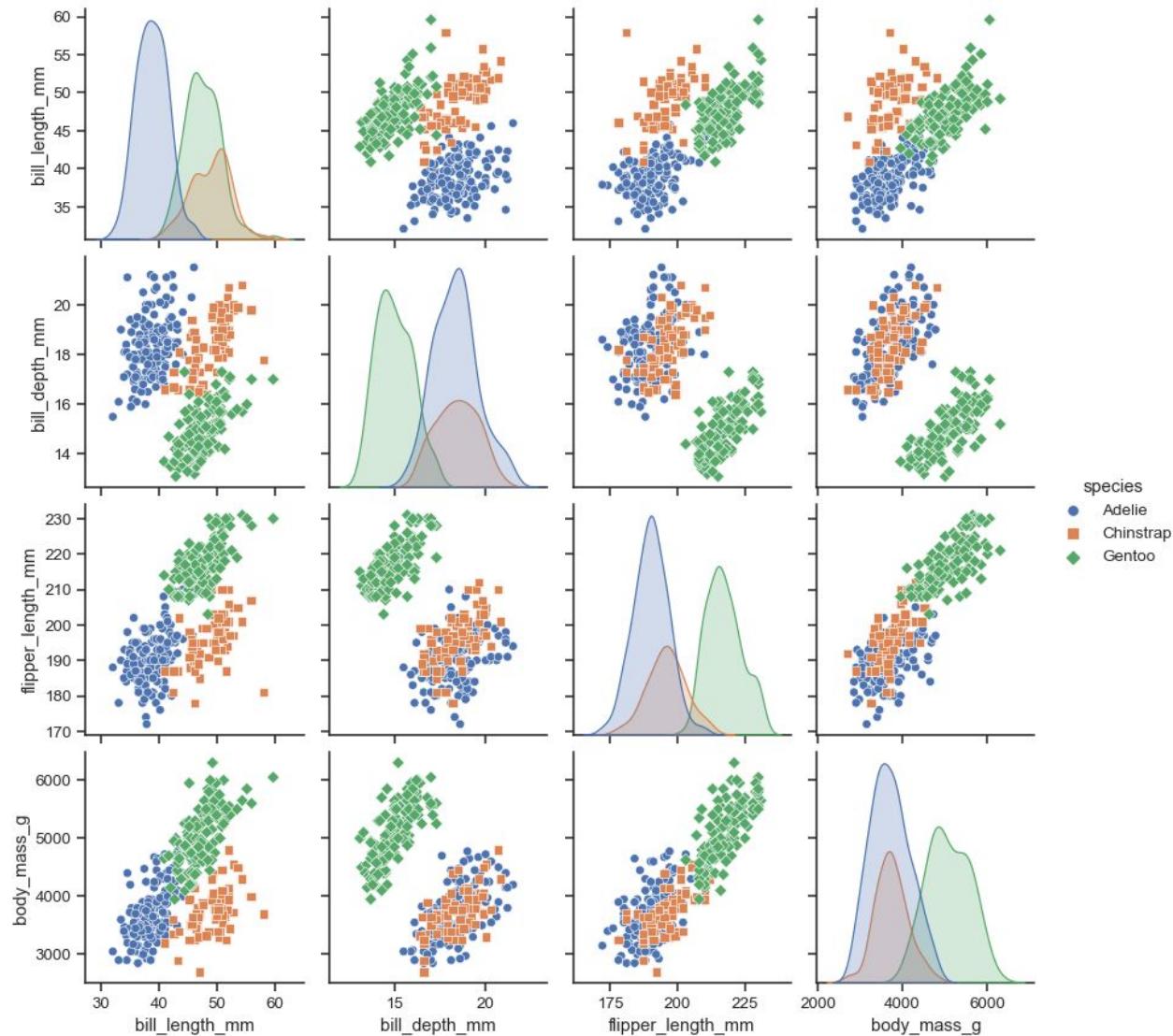
Slide from Dr. Hadi Kharrazi

9 months & >25 rules to clean weight



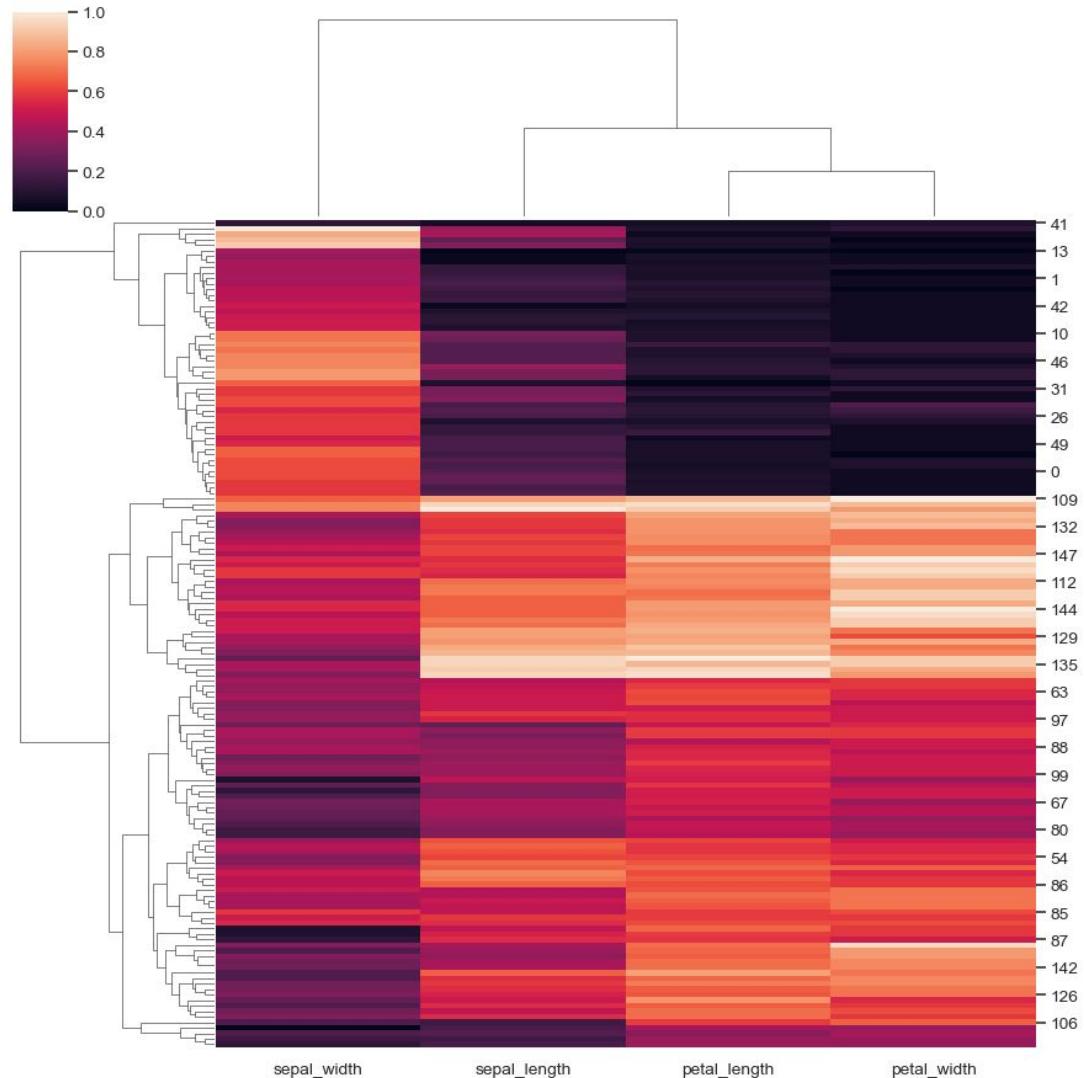
# Exploratory Data Analysis

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



# Exploratory Data Analysis

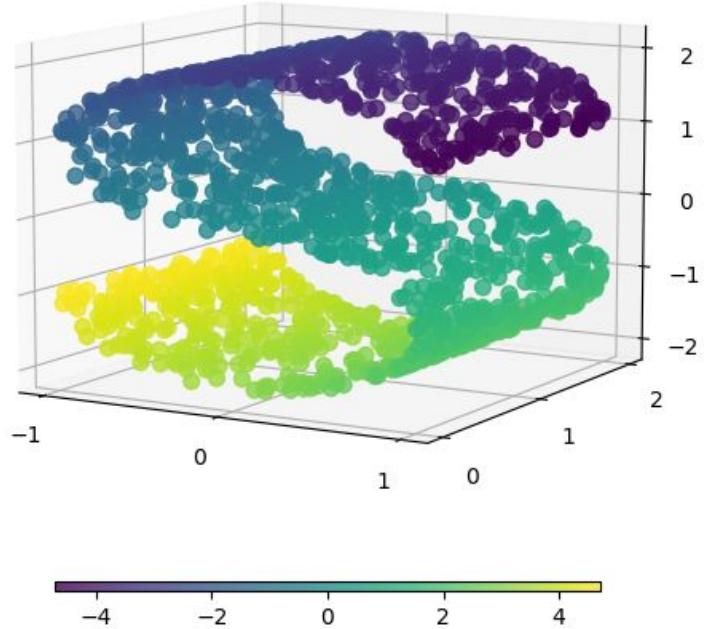
- Individual variable distributions
- Pairwise variable distributions
- Distributions relative to variable(s) of interest
- Hierarchical clustering of variables
- Point analysis of extreme values



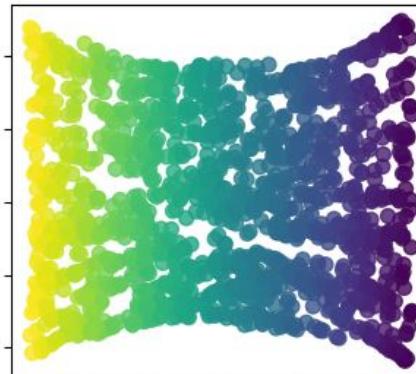
How do I look at all the data together?

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

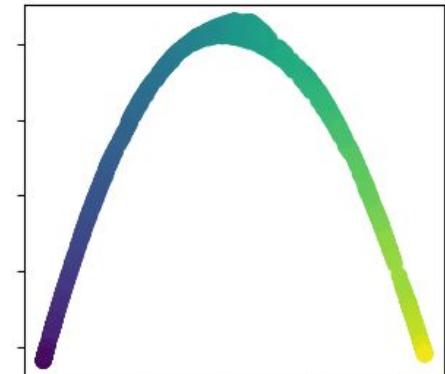
Original S-curve samples



Isomap Embedding



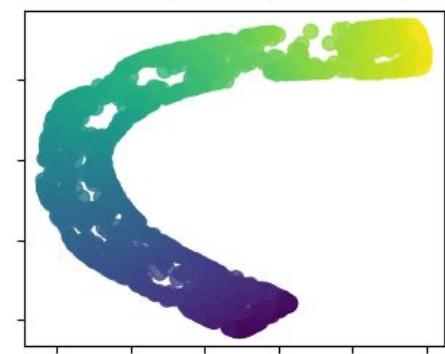
Spectral Embedding



Multidimensional scaling

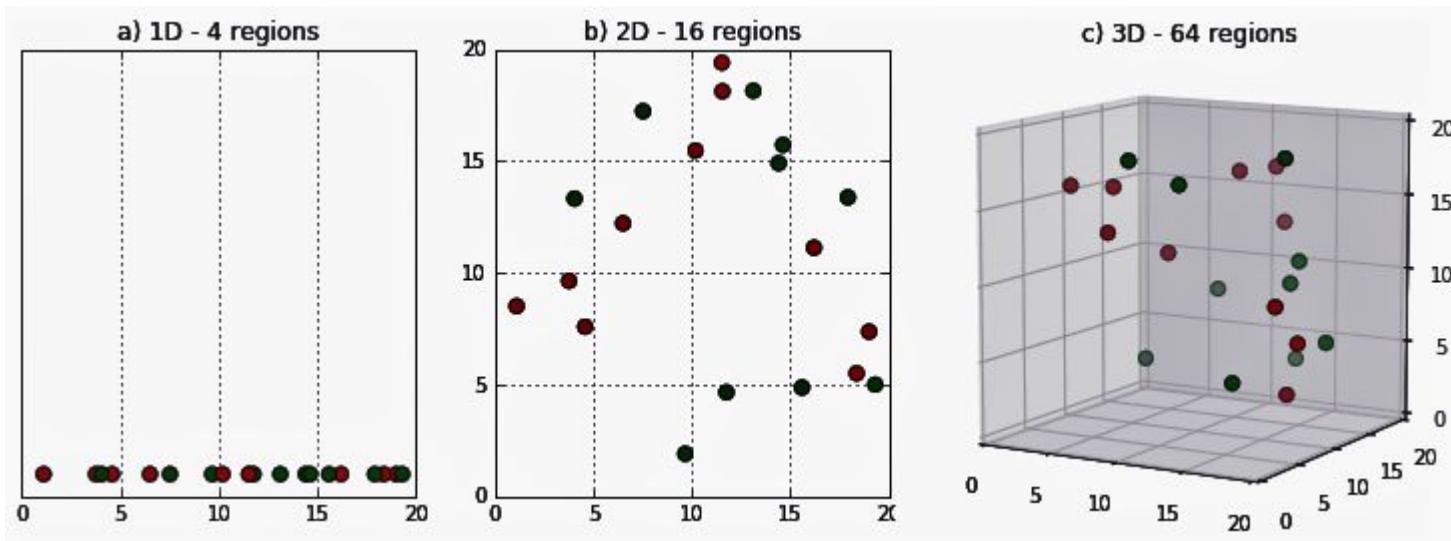


T-distributed Stochastic Neighbor Embedding



Why is this hard?

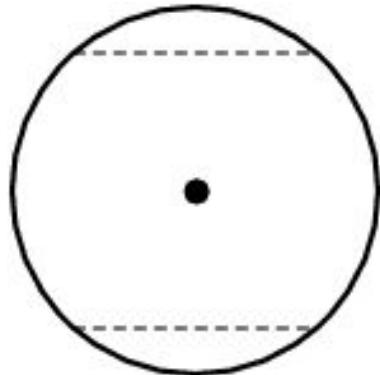
# High dimensional data is sparse



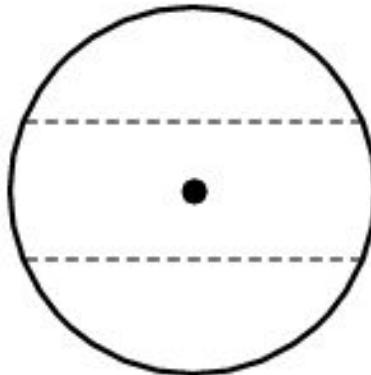
<https://medium.com/analytics-vidhya/the-curse-of-dimensionality-and-its-cure-f9891ab72e5c>

# High dimensional space is counterintuitive

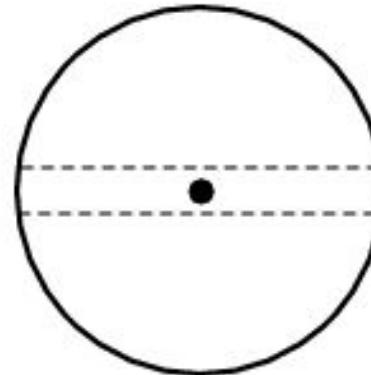
Orthogonality -> Band-size to capture 99% of the volume of a sphere:



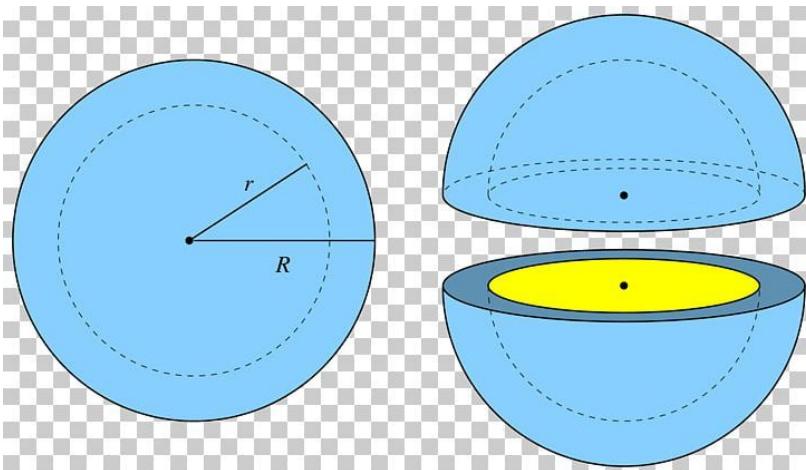
$$d = 2$$



$$d = 10$$



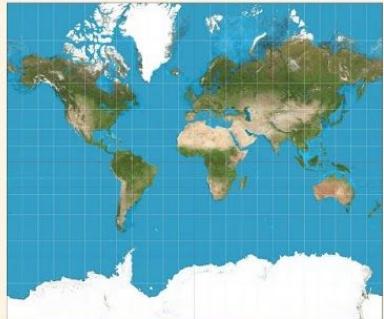
$$d = 100$$



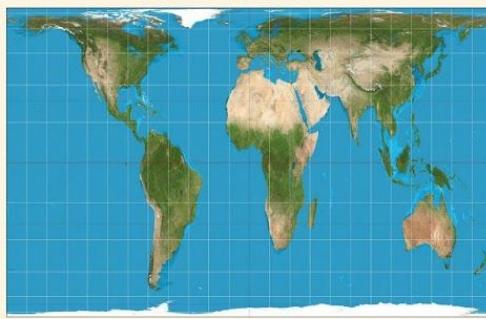
Mass becomes increasingly “shell-like”

# No representation is perfect

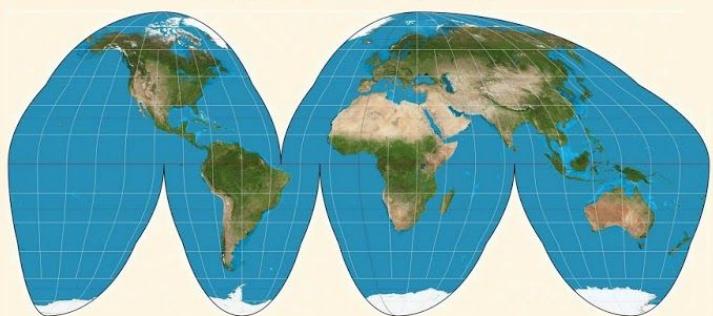
MERCATOR



GALL-PETERS



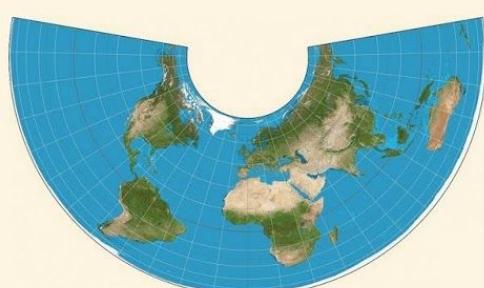
GOODE-HOMOLOSCINE



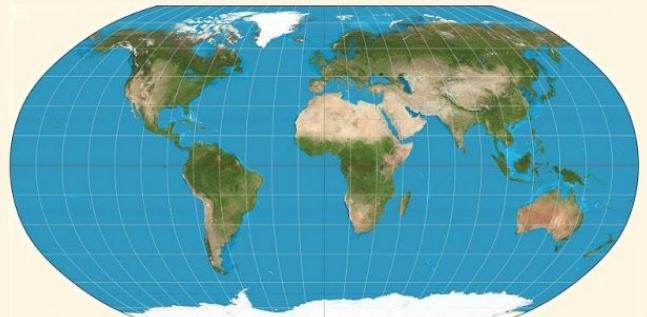
WATERMELON



ALBERS



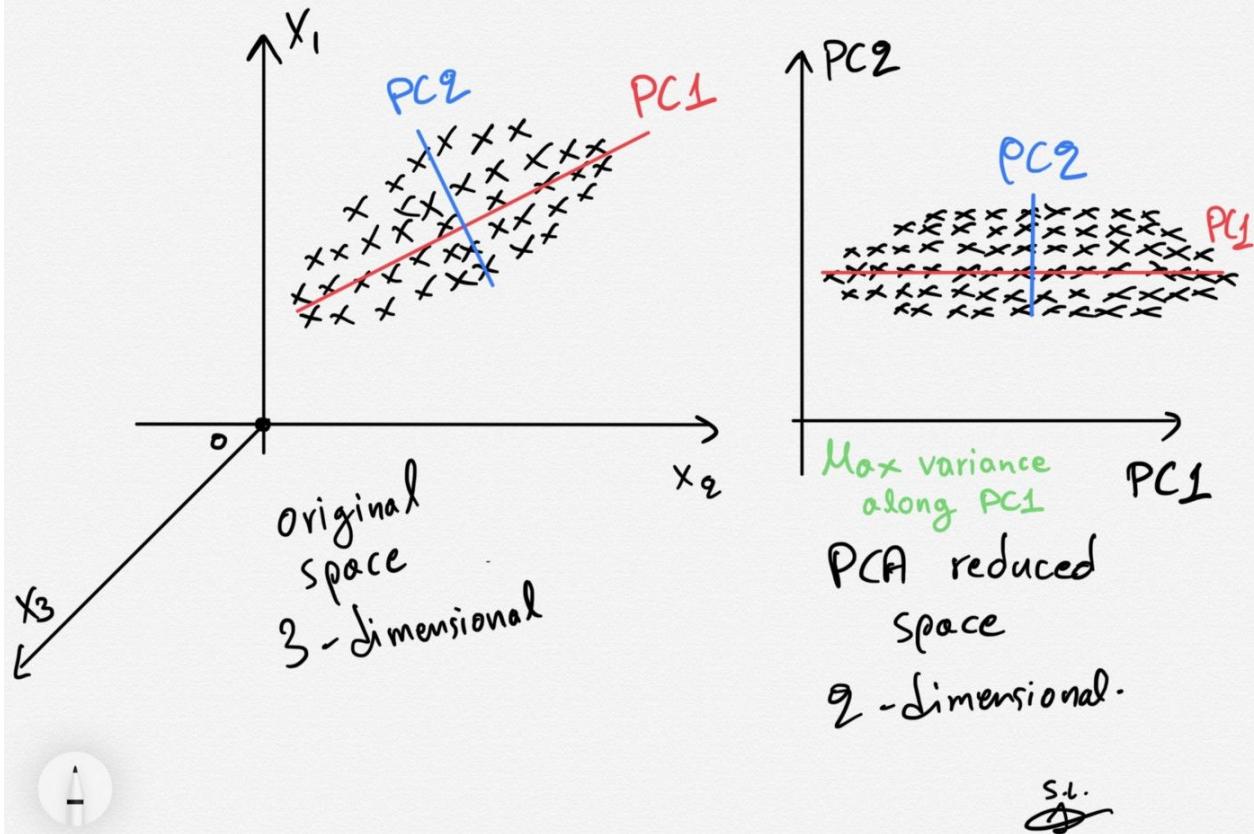
ROBINSON



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
- 
- | Dimensions | Percentage of explained variances |
|------------|-----------------------------------|
| 1          | 41.2%                             |
| 2          | 18.4%                             |
| 3          | 12.4%                             |
| 4          | 8.2%                              |
| 5          | 7%                                |
| 6          | 4.2%                              |
| 7          | 3%                                |
| 8          | 2.7%                              |
| 9          | 1.6%                              |
| 10         | 1.2%                              |
- What variables contribute most to PCs? BiPlot

# MultiDimensional Scaling (MDS): Distances

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

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

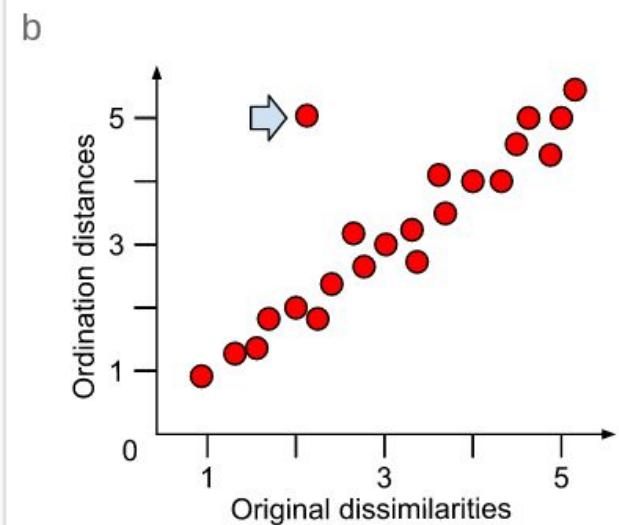
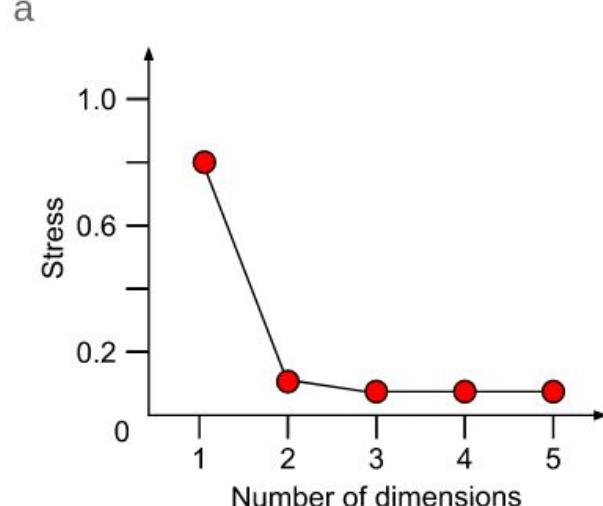
Where  $x_1, \dots, x_N$  are data points with their new set of coordinates in lower dimensional space.

$d_{ij}$  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.

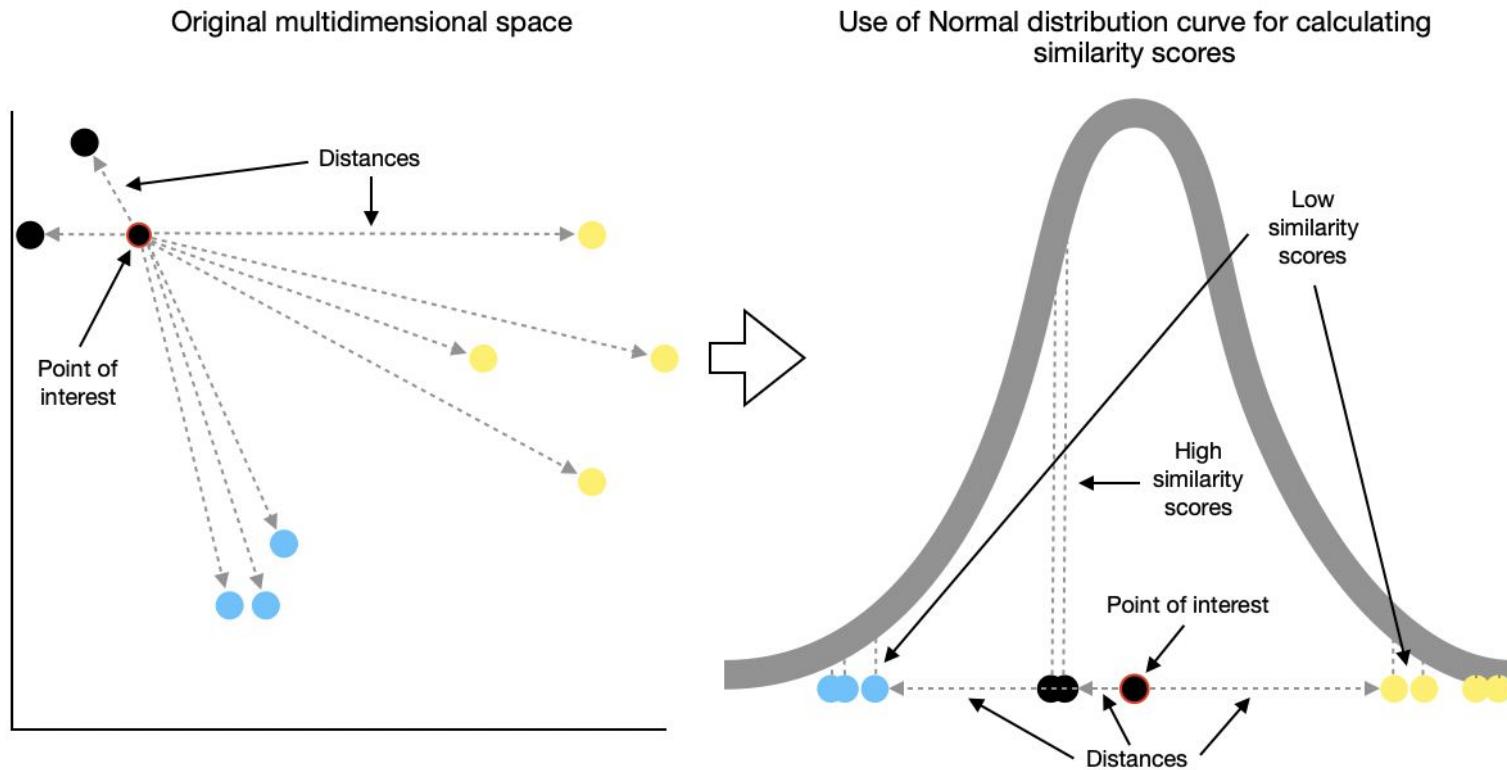
The closer the value of  $\|x_i - x_j\|$  is to  $d_{ij}$  the

Non-Metric: Ranks

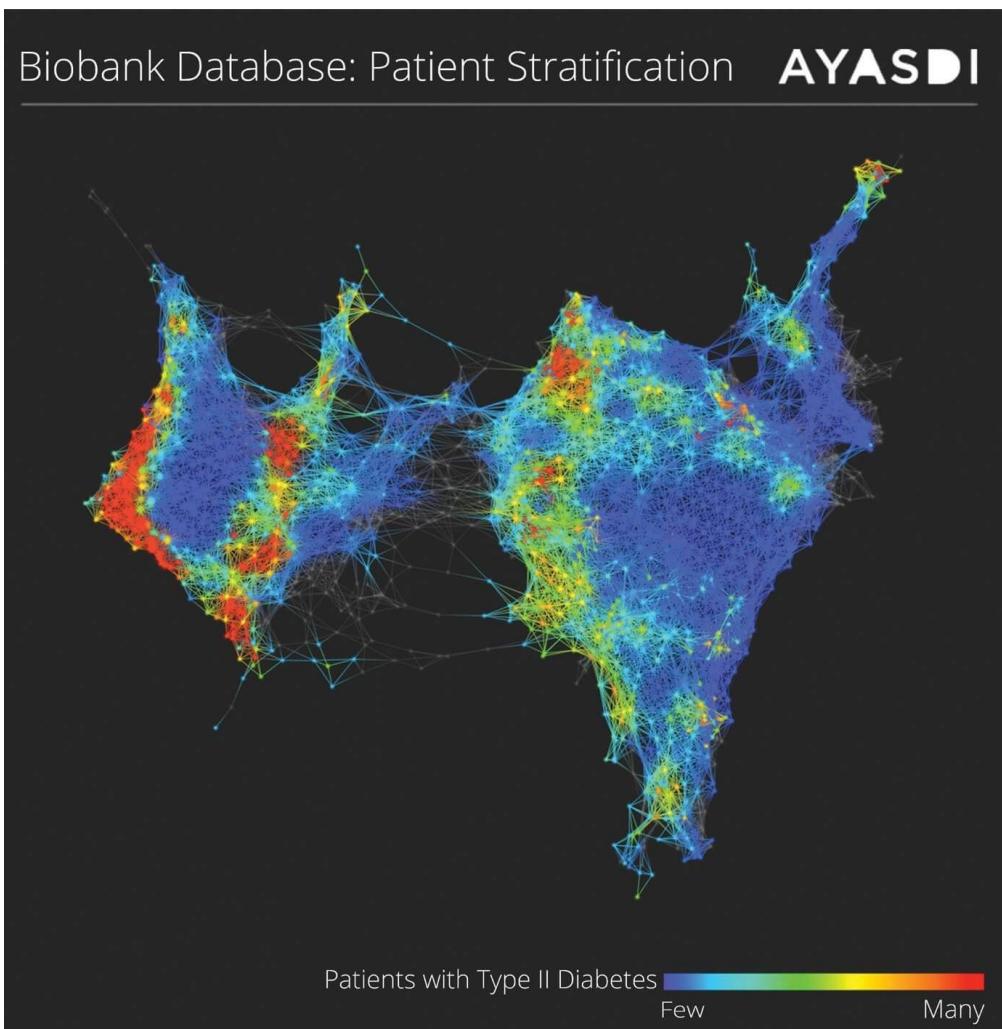


# t-SNE/UMAP: Probabilities

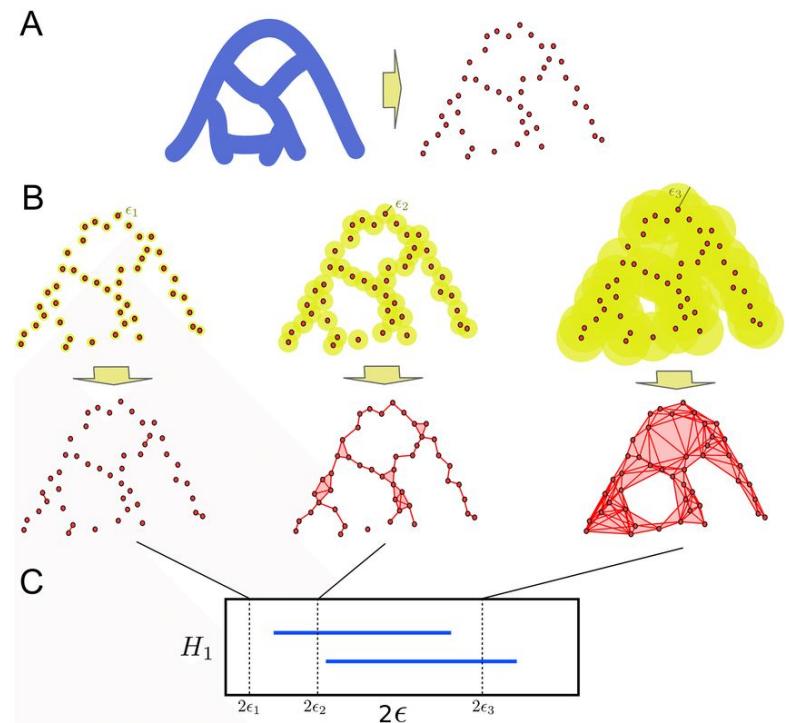
- Pairwise probability distribution in all dimensions
- Pairwise probability distribution in few dimensions
- Stochastic minimisation of KL divergence between distributions



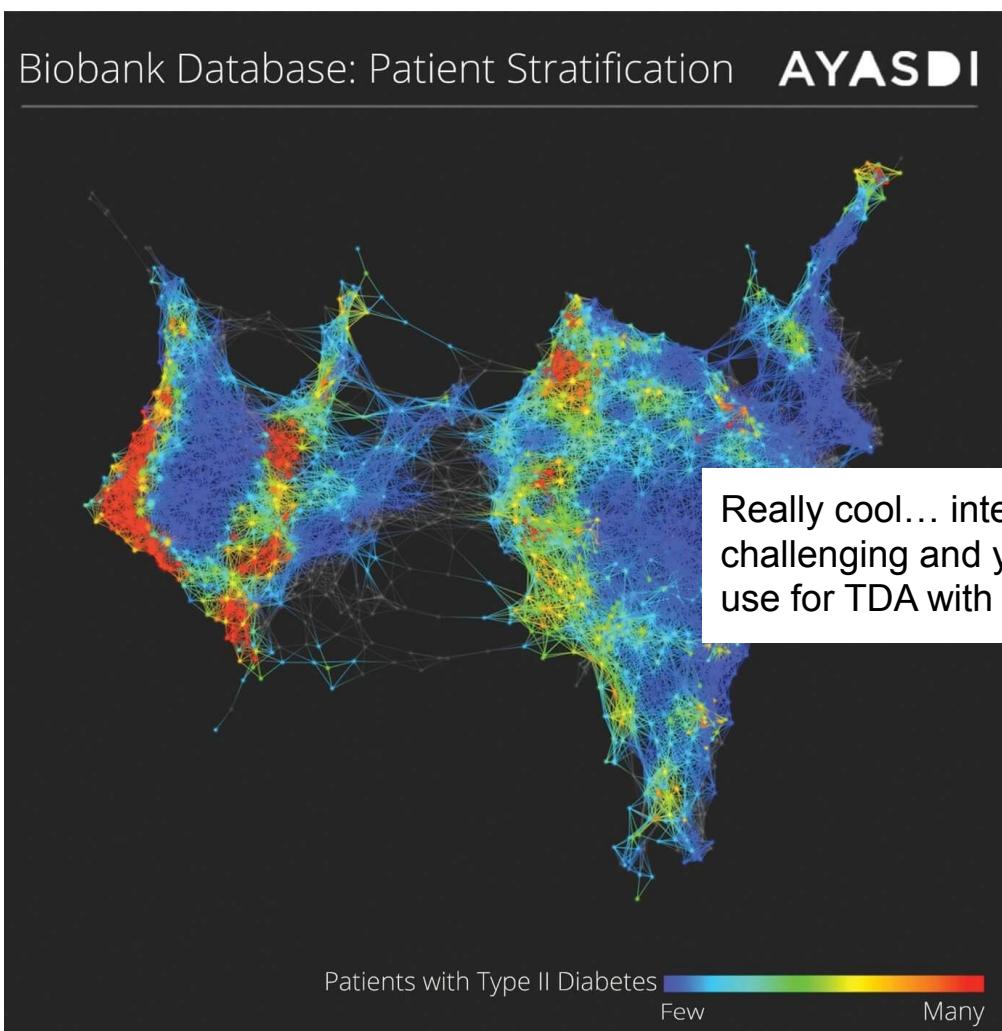
# Topological Data Analysis



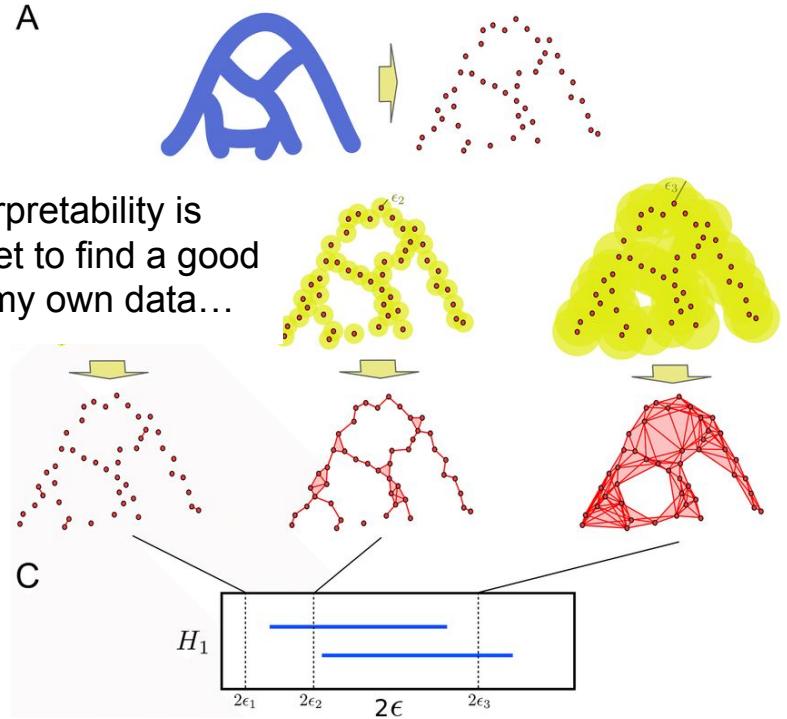
- Point clouds  $\rightarrow$  increase radius  $\rightarrow$  simplicial complexes  $\rightarrow$  topological characteristics



# Topological Data Analysis

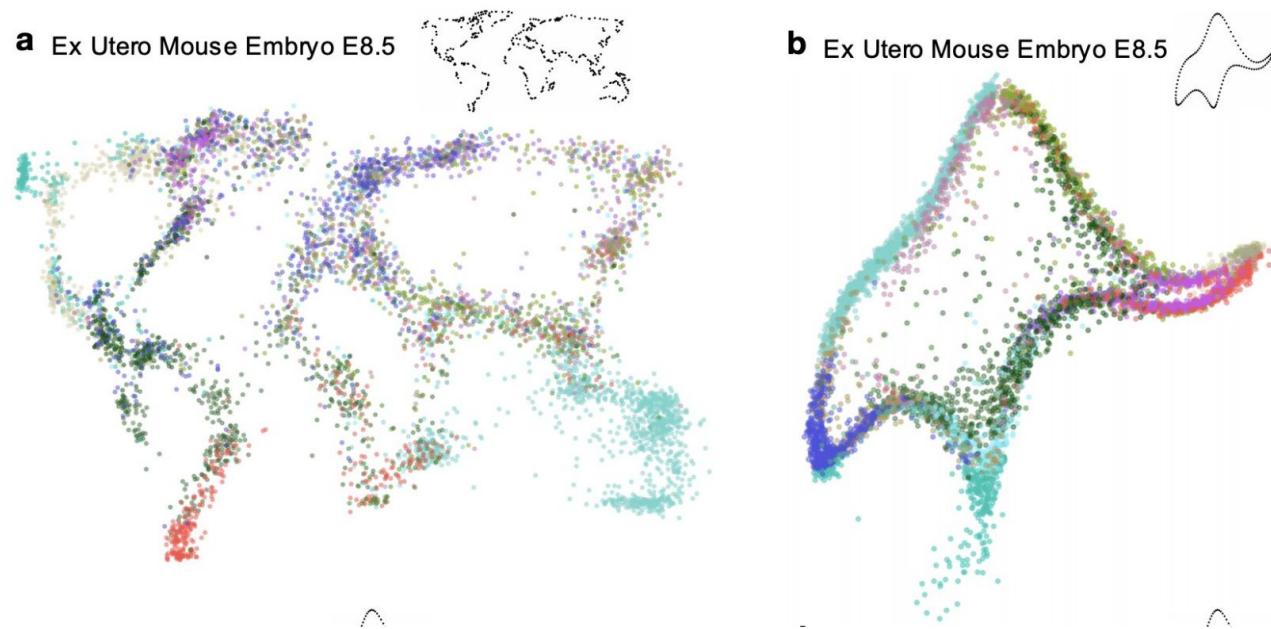


- Point clouds  $\rightarrow$  increase radius  $\rightarrow$  simplicial complexes  $\rightarrow$  topological characteristics



# Avoid over-interpreting single plots

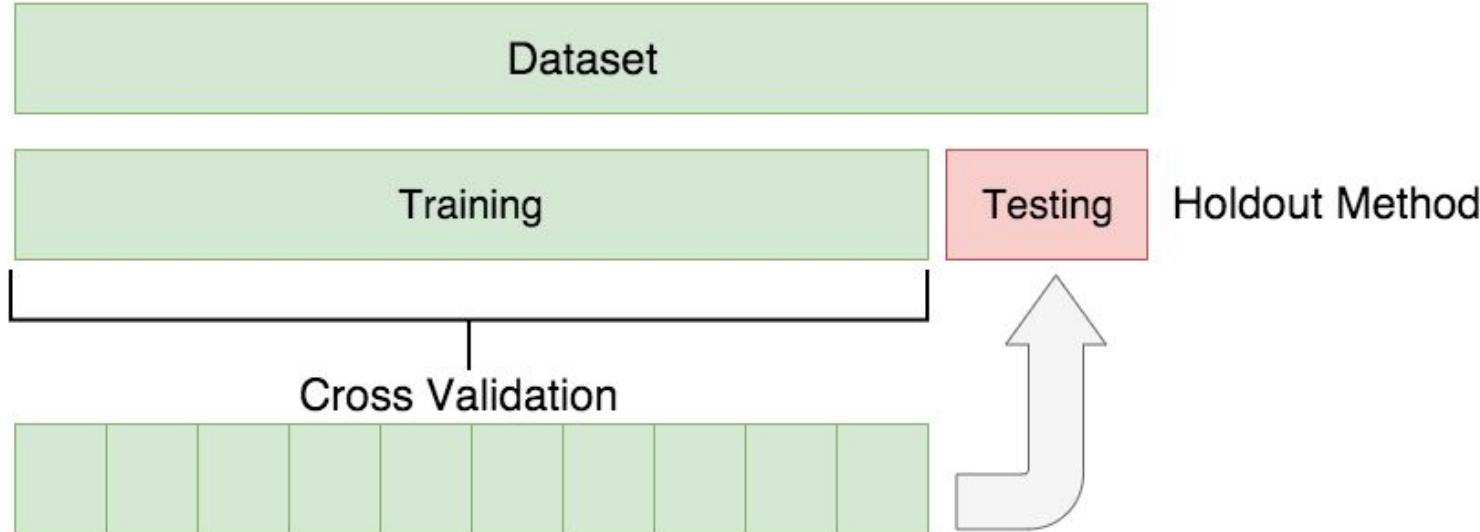
- Sensitive to hyperparameters
- Beware analysing these non-linear projections
- Can contribute to confirmation bias



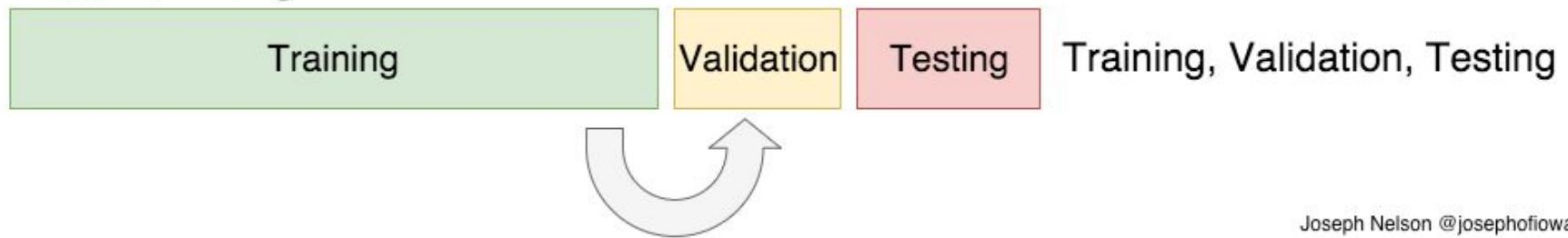
<https://www.biorxiv.org/content/10.1101/2021.08.25.457696v3>

# Predicting using tabular data

# Overfitting 101: Test-Train Split

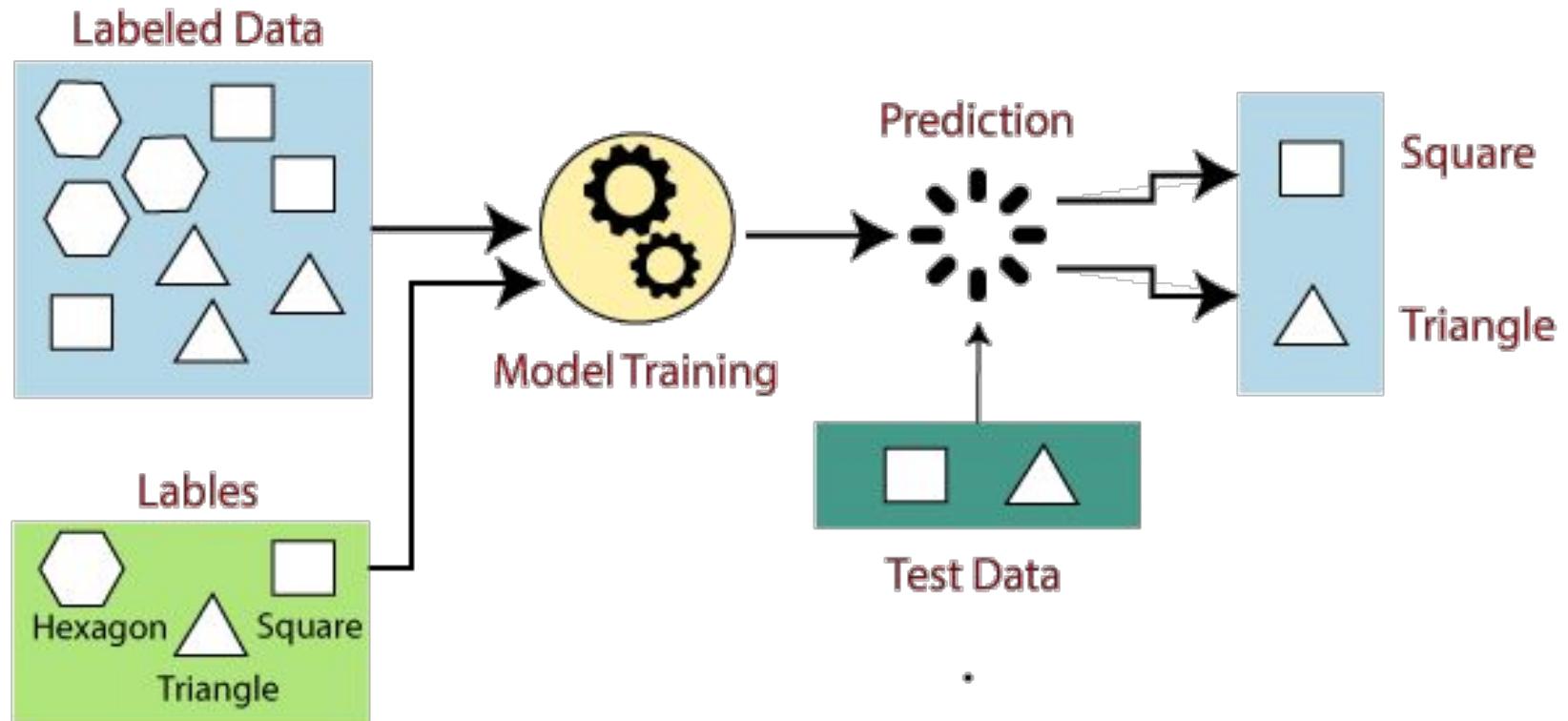


Data Permitting:



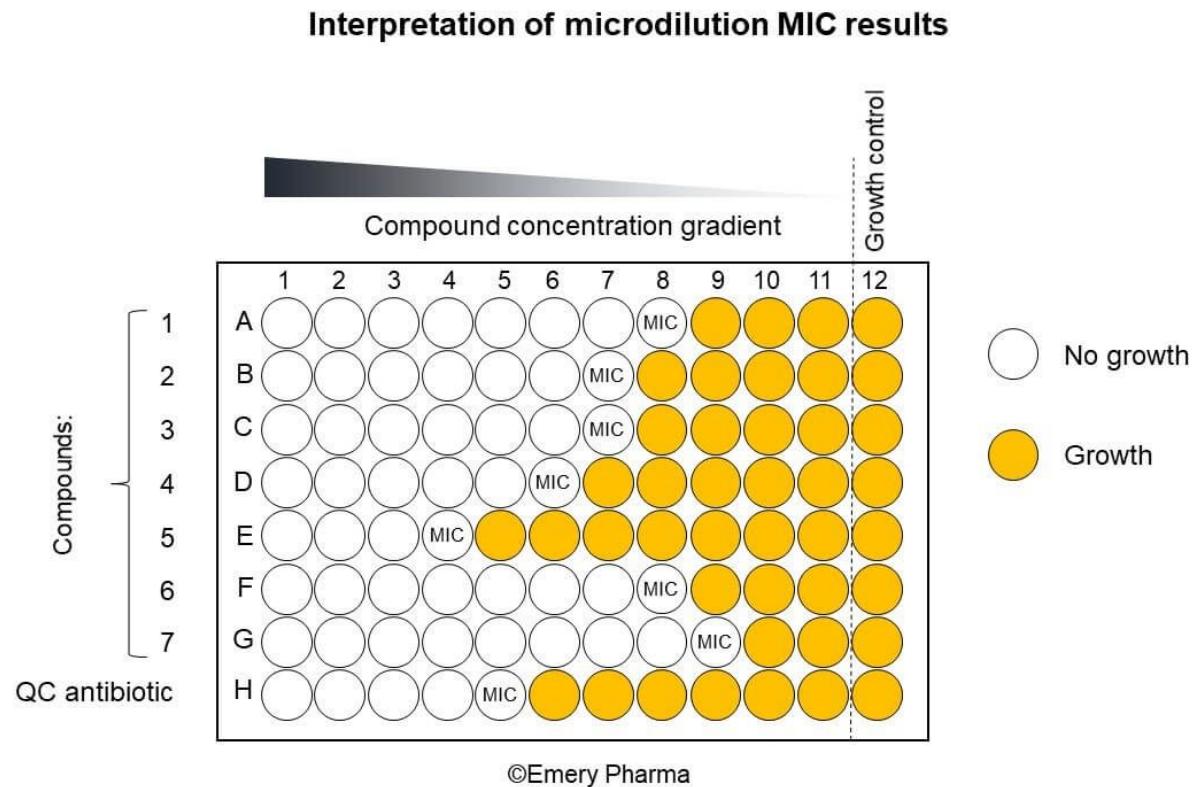
Joseph Nelson @josephofiowa

# 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

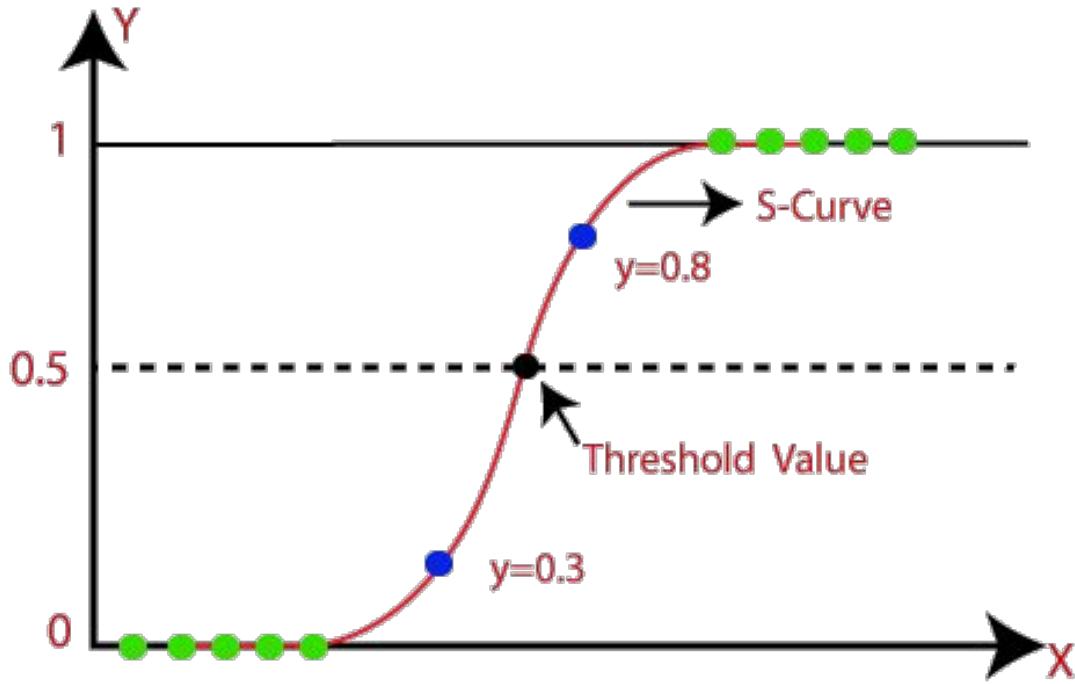


# Start simple: linear regression

Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon	
Simple regression: $\text{Im}(y \sim 1 + x)$	<b>y is independent of x</b> P: One-sample t-test N: Wilcoxon signed-rank	<code>t.test(y)</code> <code>wilcox.test(y)</code>	$\text{Im}(y \sim 1)$ $\text{Im}(\text{signed\_rank}(y) \sim 1)$	✓ for $N > 14$	One number (intercept, i.e., the mean) predicts <b>y</b> . - (Same, but it predicts the <i>signed rank</i> of <b>y</b> .)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	<code>t.test(y1, y2, paired=TRUE)</code> <code>wilcox.test(y1, y2, paired=TRUE)</code>	$\text{Im}(y_2 - y_1 \sim 1)$ $\text{Im}(\text{signed\_rank}(y_2 - y_1) \sim 1)$	✓ for $N > 14$	One intercept predicts the pairwise $y_2 - y_1$ differences. - (Same, but it predicts the <i>signed rank</i> of $y_2 - y_1$ .)	
	<b>y ~ continuous x</b> P: Pearson correlation N: Spearman correlation	<code>cor.test(x, y, method='Pearson')</code> <code>cor.test(x, y, method='Spearman')</code>	$\text{Im}(y \sim 1 + x)$ $\text{Im}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for $N > 10$	One intercept plus <b>x</b> multiplied by a number (slope) predicts <b>y</b> . - (Same, but with <i>ranked x</i> and <b>y</b> )	
	<b>y ~ discrete x</b> P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	<code>t.test(y1, y2, var.equal=TRUE)</code> <code>t.test(y1, y2, var.equal=FALSE)</code> <code>wilcox.test(y1, y2)</code>	$\text{Im}(y \sim 1 + G_2)^A$ $\text{gls}(y \sim 1 + G_2, \text{weights}=\dots)^B$ $\text{Im}(\text{signed\_rank}(y) \sim 1 + G_2)^A$	✓ ✓ for $N > 11$	An intercept for <b>group 1</b> (plus a difference if <b>group 2</b> ) predicts <b>y</b> . - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of <b>y</b> .)	
Multiple regression: $\text{Im}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	<code>aov(y ~ group)</code> <code>kruskal.test(y ~ group)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_N)^A$ $\text{Im}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_N)^A$	✓ for $N > 11$	An intercept for <b>group 1</b> (plus a difference if $\text{group} \neq 1$ ) predicts <b>y</b> . - (Same, but it predicts the <i>rank</i> of <b>y</b> .)	
	P: One-way ANCOVA	<code>aov(y ~ group + x)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_N + x)^A$	✓	- (Same, but plus a slope on <b>x</b> ). Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous <b>x</b> .	
	P: Two-way ANOVA	<code>aov(y ~ group * sex)</code>	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2*S_2 + G_3*S_3 + \dots + G_N*S_K)$	✓	Interaction term: changing <b>sex</b> changes the <b>y ~ group</b> parameters. Note: $G_{2..N}$ is an <i>indicator (0 or 1)</i> for each non-intercept levels of the <b>group</b> variable. Similarly for $S_{2..K}$ for <b>sex</b> . The first line (with $G_j$ ) is main effect of group, the second (with $S_j$ ) for sex and the third is the <b>group</b> $\times$ <b>sex</b> interaction. For two levels (e.g. male/female), line 2 would just be "S <sub>2</sub> " and line 3 would be $S_2$ multiplied with each $G_j$ .	[Coming]
	Counts ~ discrete x N: Chi-square test	<code>chisq.test(groupXsex_table)</code>	<b>Equivalent log-linear model</b> <code>glm(y ~ 1 + G<sub>2</sub> + G<sub>3</sub> + ... + G<sub>N</sub> + S<sub>2</sub> + S<sub>3</sub> + ... + S<sub>K</sub> + G<sub>2</sub>*S<sub>2</sub> + G<sub>3</sub>*S<sub>3</sub> + ... + G<sub>N</sub>*S<sub>K</sub>, family=...)^A</code>	✓	Interaction term: (Same as Two-way ANOVA.) Note: Run <code>glm</code> using the following arguments: <code>glm(model, family=poisson())</code> . As linear-model, the Chi-square test is $\log(y) = \log(N) + \log(a) + \log(b) + \log(a\beta)$ where $a, b$ are proportions. See more info in <a href="#">the accompanying notebook</a> .	Same as Two-way ANOVA
N: Goodness of fit	<code>chisq.test(y)</code>	<code>glm(y ~ 1 + G<sub>2</sub> + G<sub>3</sub> + ... + G<sub>N</sub>, family=...)^A</code>	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA	

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

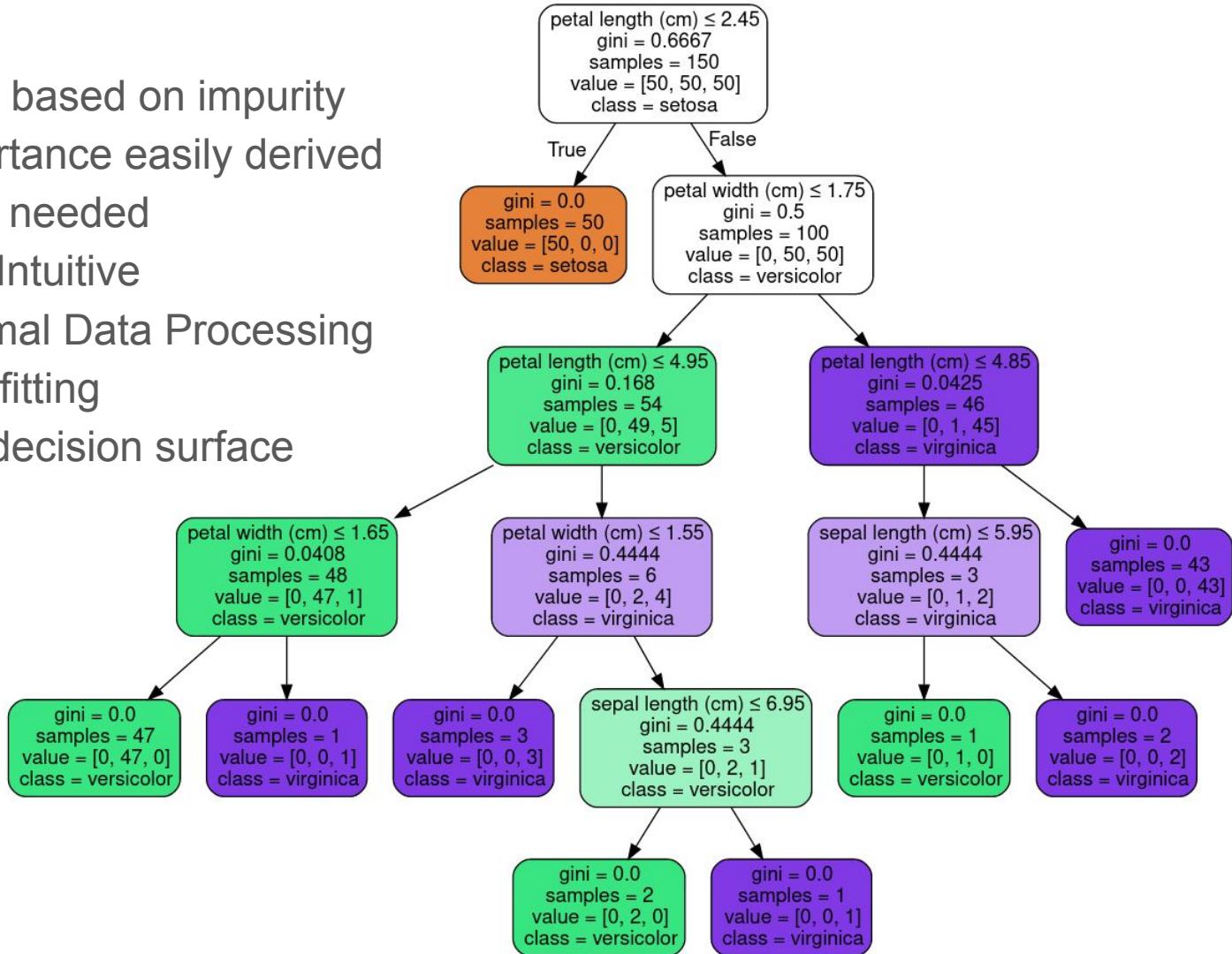
# Add a sigmoid for classification



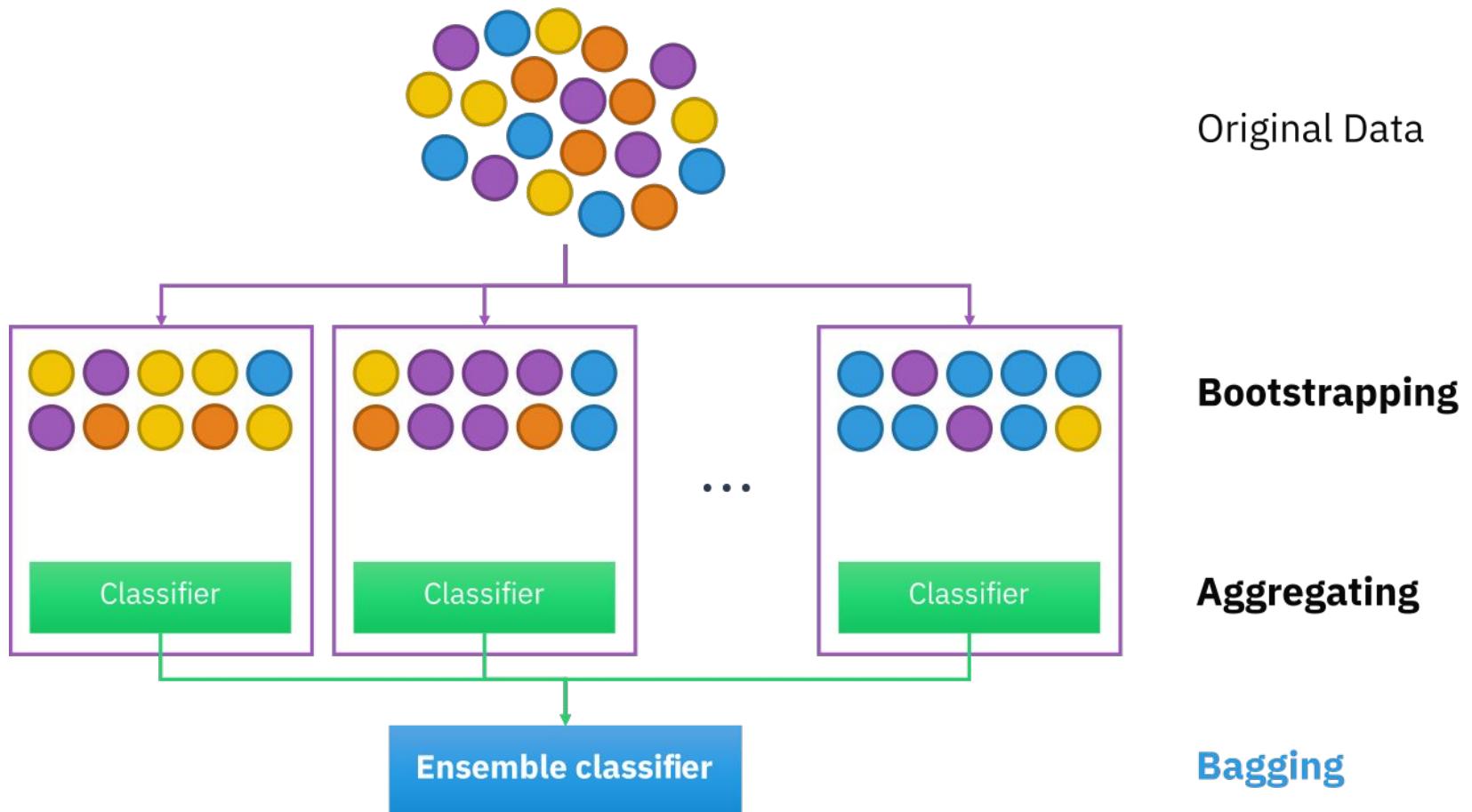
- Crude measure of feature importance (model coefficients)
- Specific feature selection can be a good idea
- Support for regularisation (Lasso/L1  $\rightarrow$  sparsity vs Ridge/L2  $\rightarrow$  minimal vs ElasticNet  $\rightarrow$  balance)
- Statistics has developed much better practices for treatment/interpretation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

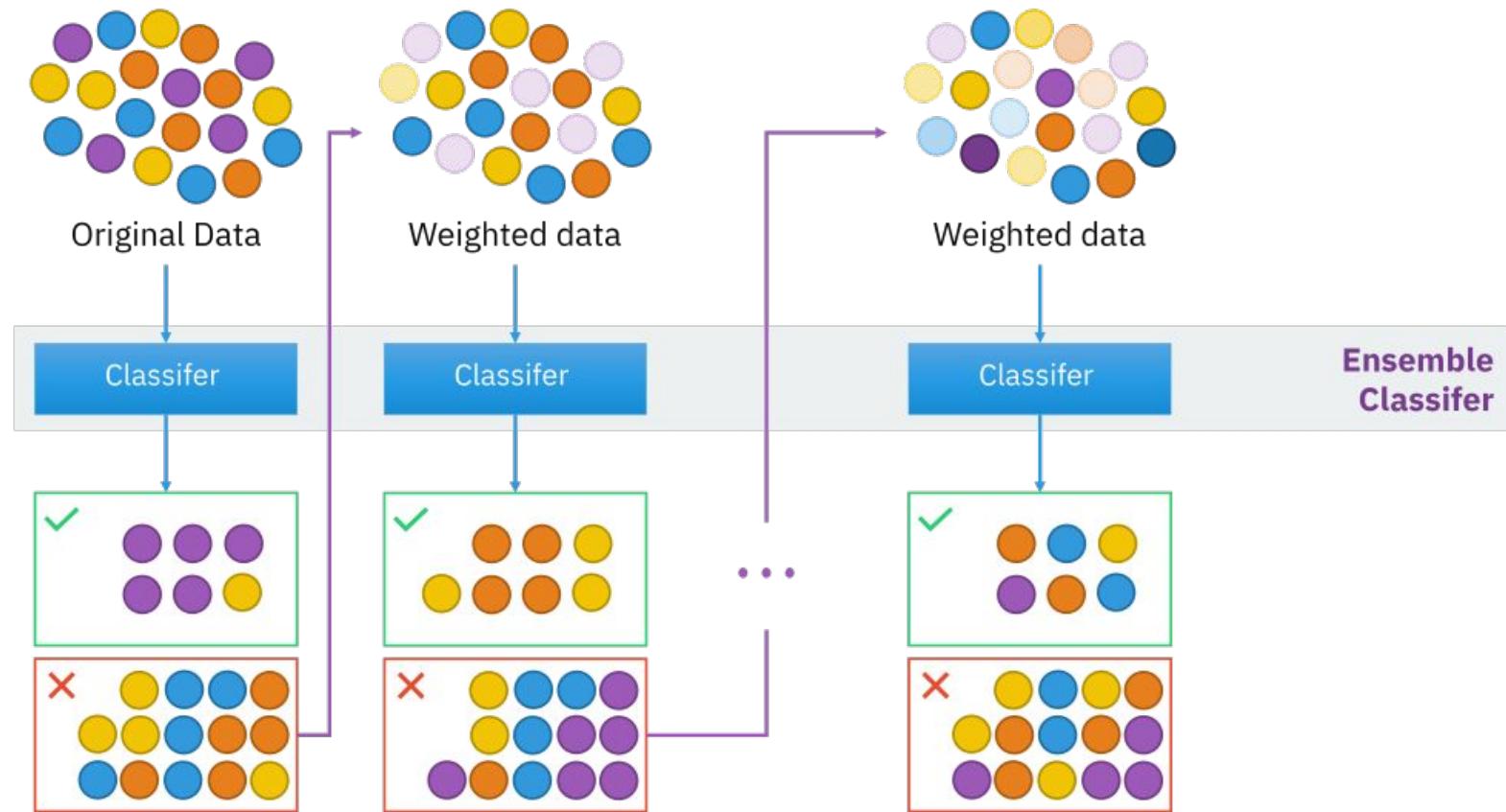


# Many Decision Trees: Bagging



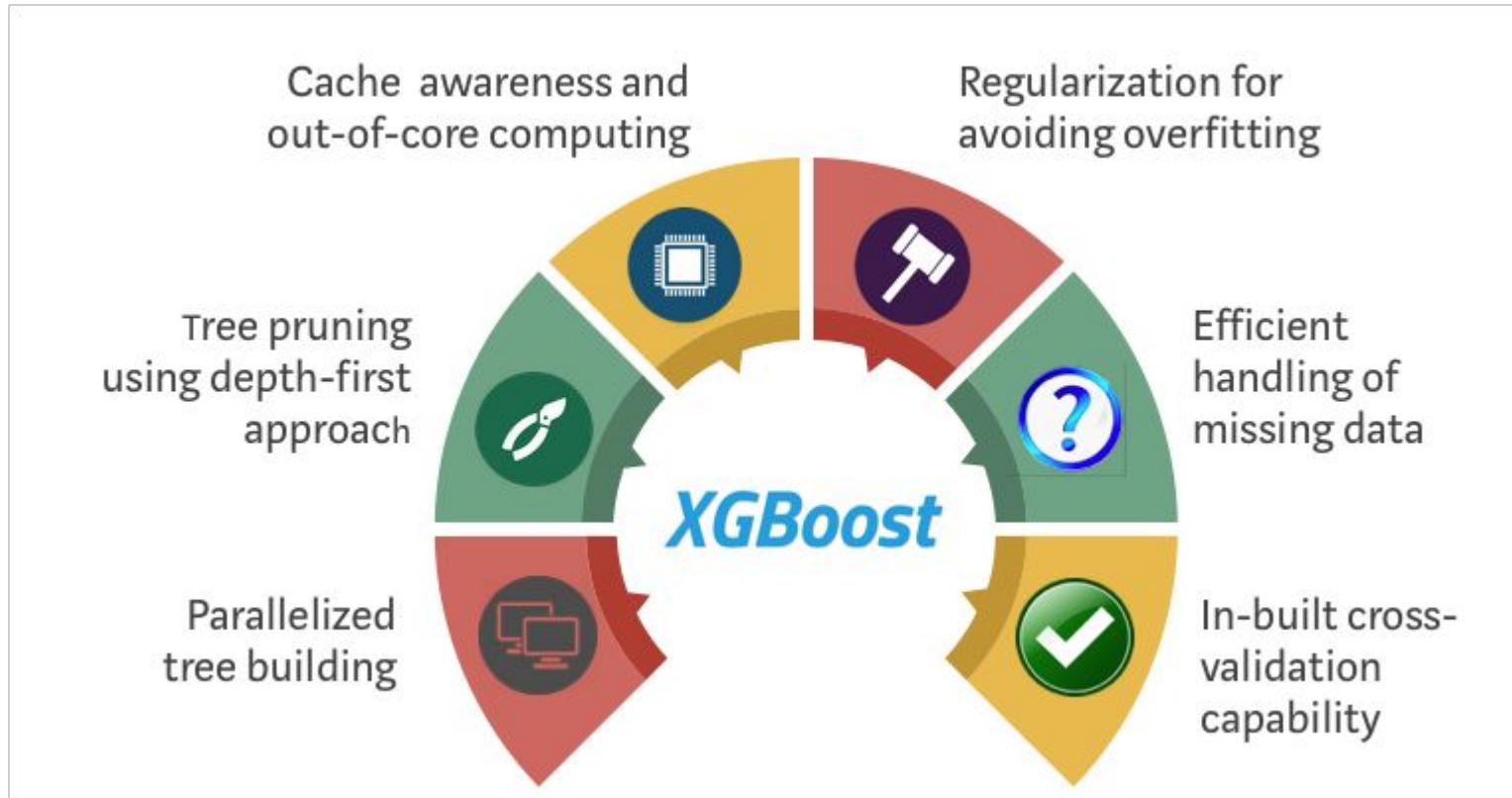
Random Forest: Bagging + Random Subset Per Split  
Feature Importance: Average impurity decrease

# Boosting: AdaBoost



# Gradient Boosting: XGBoost

- Fix on pseudo-residuals instead of weights
- Use stochastic gradient descent



# 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