Lecture 8: Data considerations for clinical machine learning

BIODS388/BIOMED388

Anuj Pareek MD PhD, Mars Huang PhD Student

11/05/2020

Outline

1. **Types of EHR Data**
2. Challenges when working with clinical data
3. How Much Data is Needed?
4. Quality of data

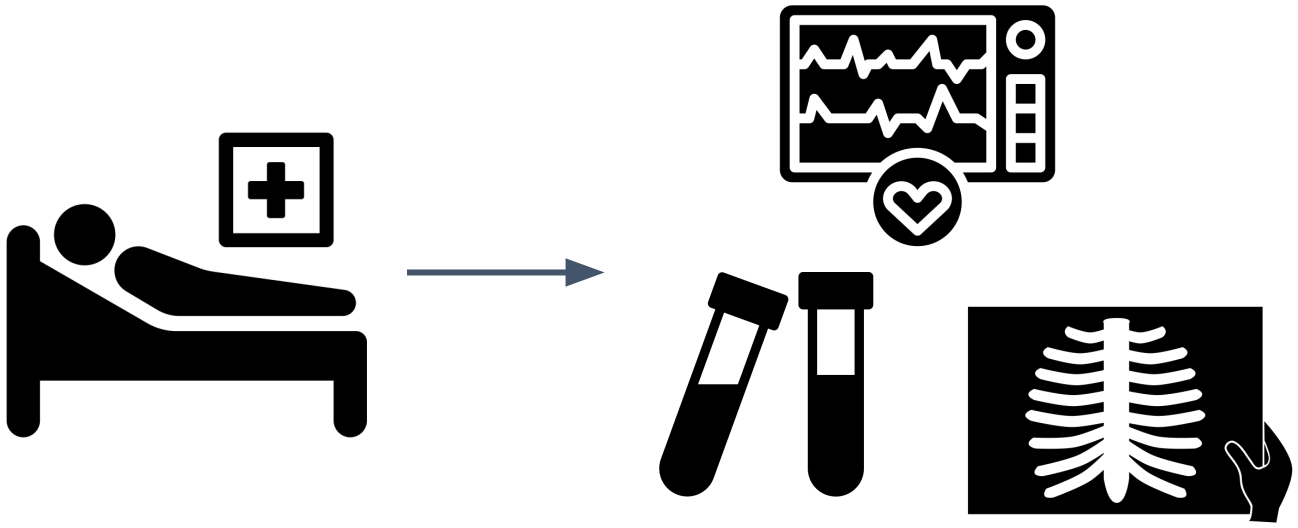
Adopt a timeline view of data

Patient admission

- Date
- Chief complaint
- Symptoms/Current

Timeline

Adopt a timeline view of data



Patient admission

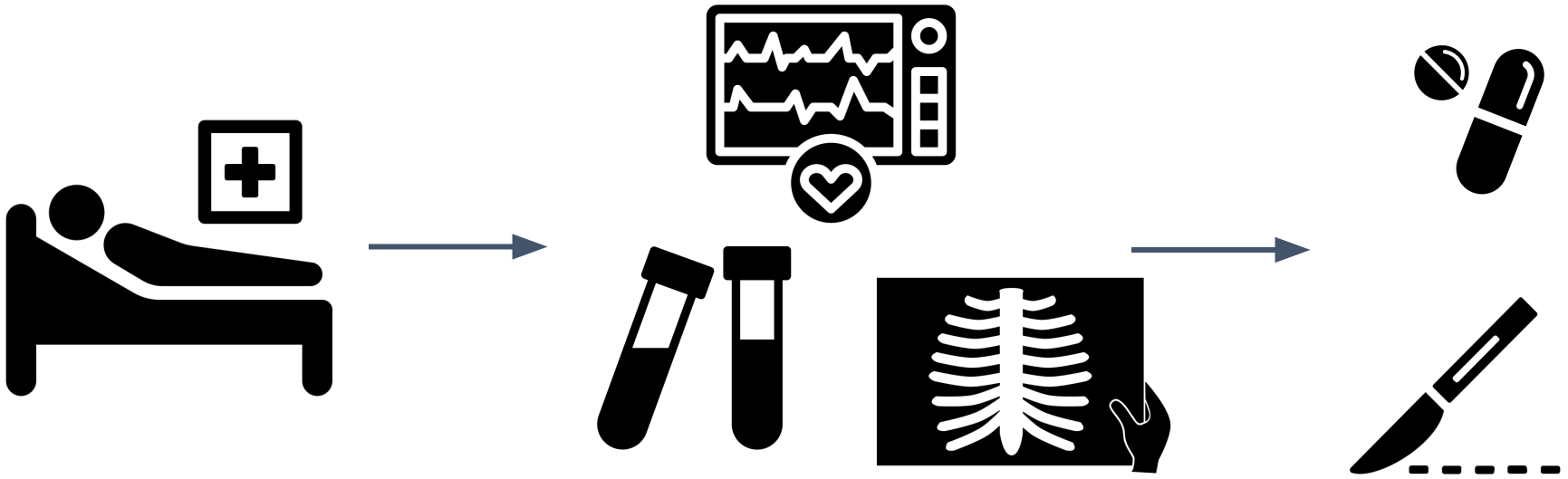
- Date
- Chief complaint
- Symptoms/Current

Monitoring, Labs, Imaging

- Vitals
- EKG
- Blood count
- X-ray

Timeline

Adopt a timeline view of data



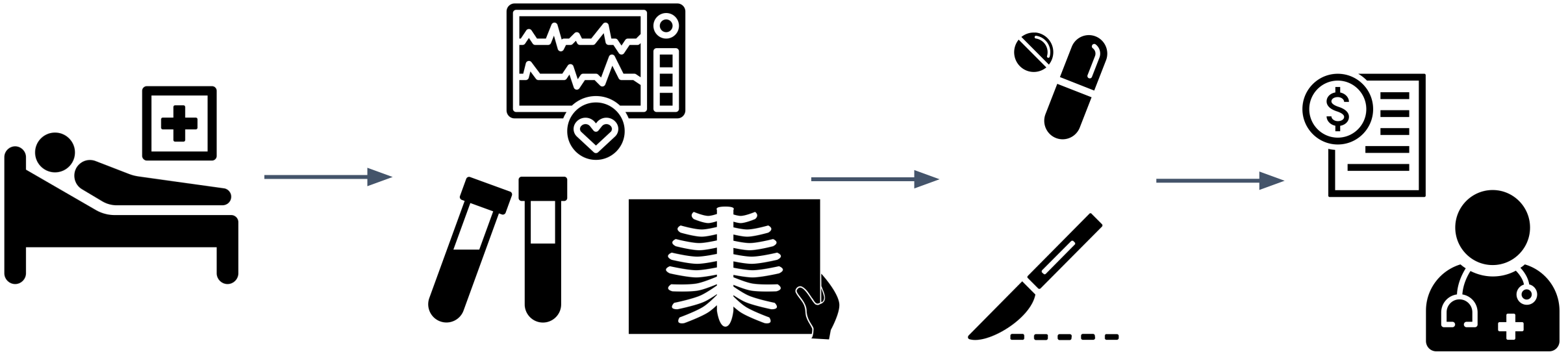
Patient admission
- Date
- Chief complaint
- Symptoms/Current

Monitoring, Labs, Imaging
- Vitals
- EKG
- Blood count
- X-ray

Medication, Procedures
- Drug dosage
- Surgery

Timeline

Adopt a timeline view of data



Patient admission
- Date
- Chief complaint
- Symptoms/Current

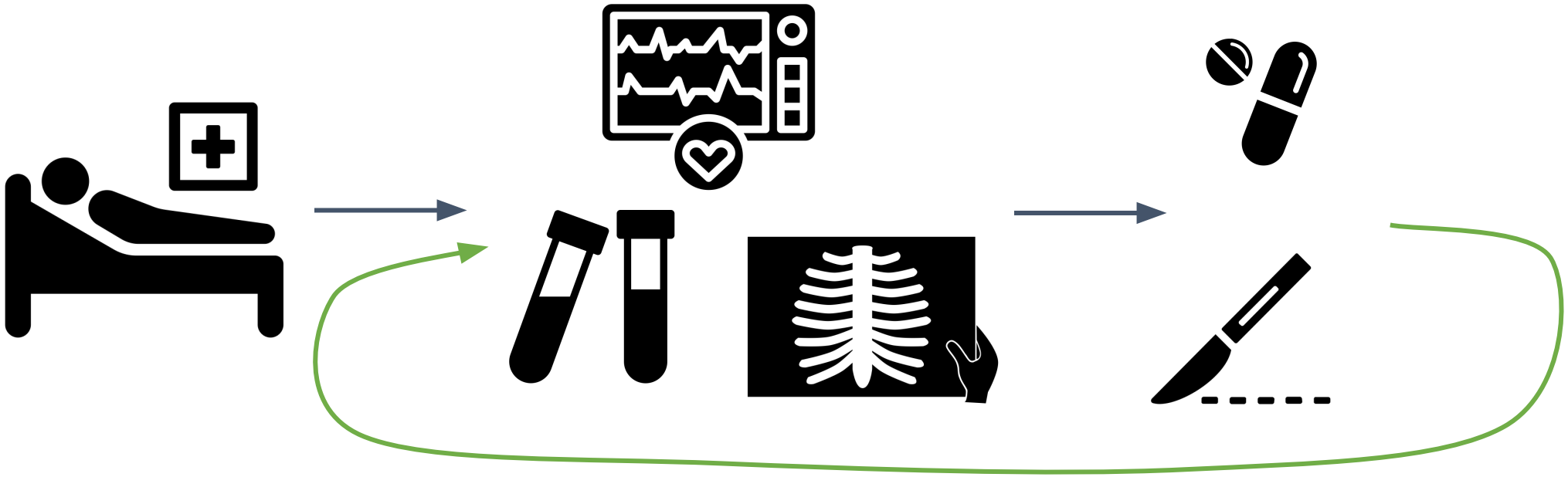
Monitoring, Labs, Imaging
- Vitals
- EKG
- Blood count
- X-ray

Medication, Procedures
- Drug dosage
- Surgery
- Biopsy

Diagnosis, Discharge
- Billing codes
- ICD codes
- Discharge date

Timeline

Adopt a timeline view of data



Patient admission
- Date
- Chief complaint
- Symptoms/Current

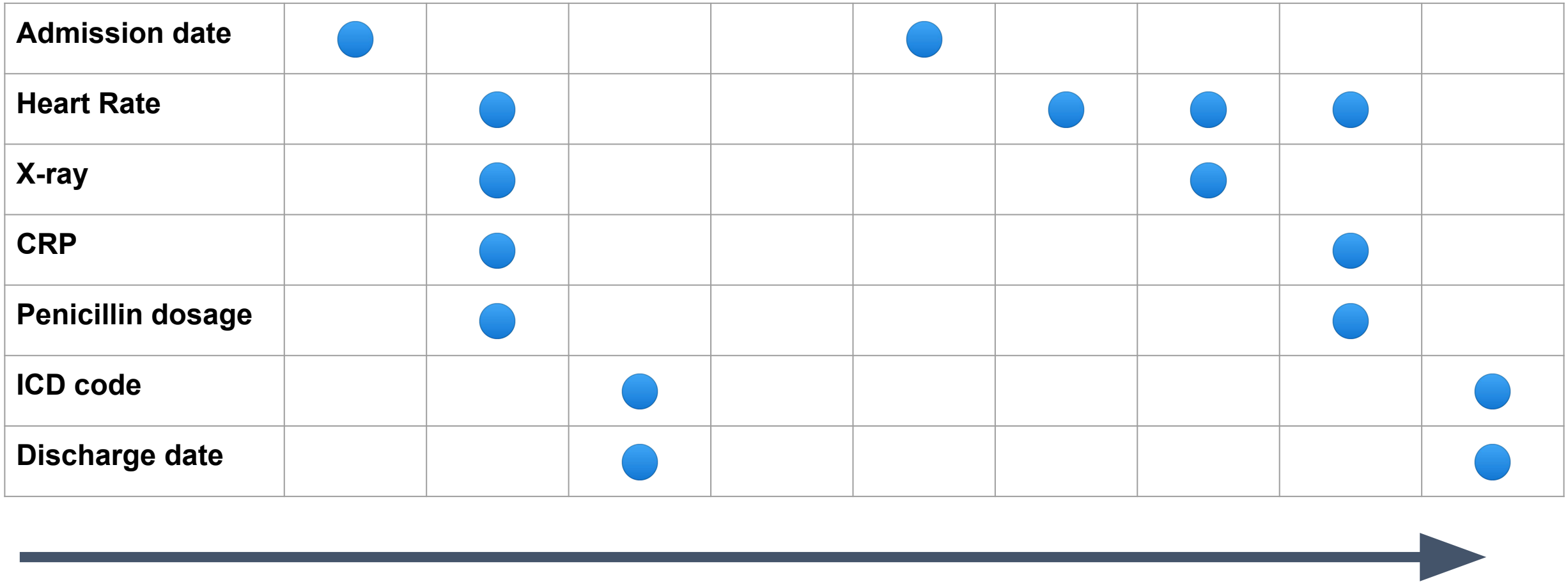
Monitoring, Labs, Imaging
- Vitals
- EKG
- Blood count
- X-ray

Medication, Procedures
- Drug dosage
- Surgery
- Biopsy

Diagnosis, Discharge
- Billing codes
- ICD codes
- Discharge date

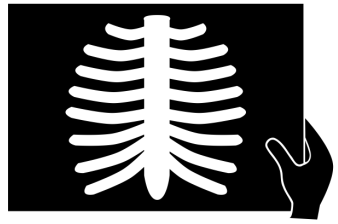
Timeline

EHR data visualized in a matrix for a single patient



Some medical data types: Medical imaging

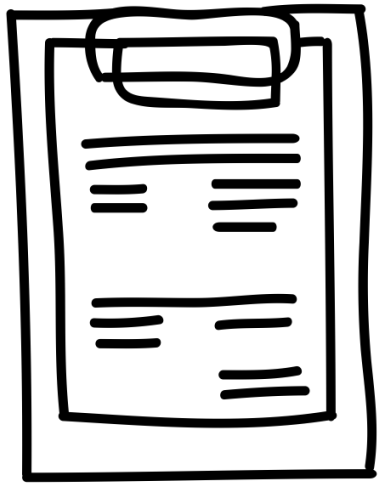
- **Radiology and Nuclear Medicine** use DICOM format
- DICOM consists of metadata and image data



- *Metadata*: Patient name, scanner model, date of image acquisition, contrast dosage etc.
- *Image data*: 2D or 3D array of points. Usually single channel (B&W)
- **Pathology** use high resolution slide scans
- Scans are images; 2D array in three channels (RGB color)

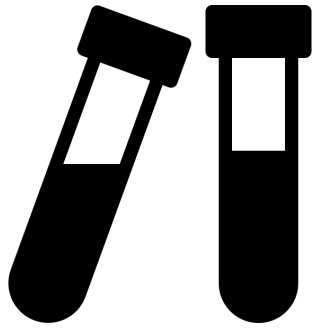
Some medical data types: Medical Notes

- Free-text
- Written by healthcare staff; physicians, nurses, assistants etc.
- *Unstructured approach:* Notes are input data for NLP-models/ transformers
- *Structured approach*: Use NLP tools to extract structured features from notes



Some medical data types: Laboratory/Blood values

- Numerical values from assays; i.e. 43
- Many different kinds of tests;
  Complete blood count, CRP, hemoglobine, urine-albumin etc.
- Specific thresholds for abnormal values
- Used in scoring and metrics
- Laboratory tests are often repeated and a single test-value should not be considered static for a patient
- Classical Machine Learning models can work really well



Some medical data types: ICD-codes

- International Classification of Diseases
- There is both ICD-9, ICD-10, ICD-11 (gets updated)
- Look up ICD-10 here: https://icd.who.int/browse10/2019/en

  Example: B34.2 Coronavirus infection, unspecified site

- Used for billing purposes, and not always accurate

Outline

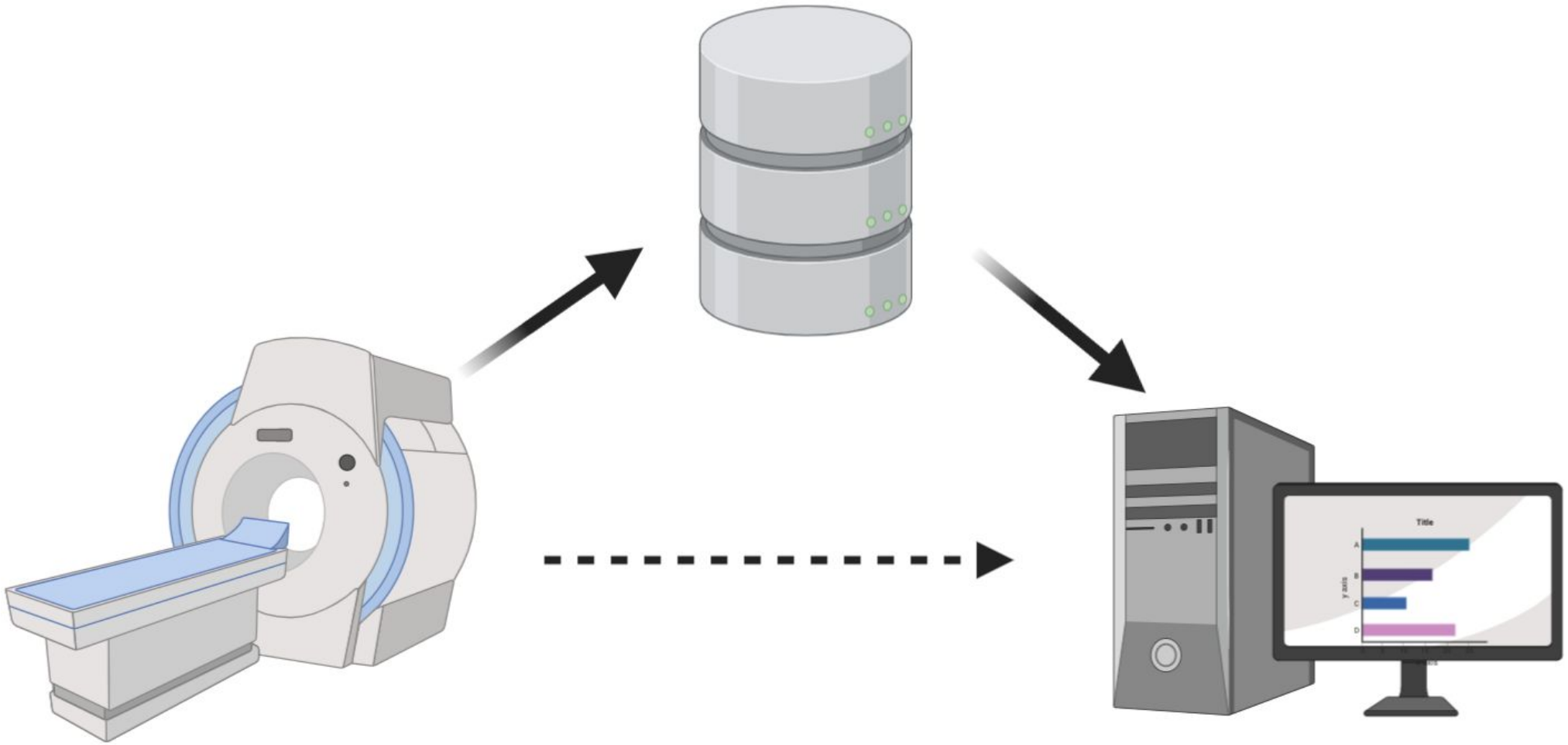
1. Types of EHR Data
2. **Challenges when working with clinical data**
3. How Much Data is Needed?
4. Assessment of quality of data

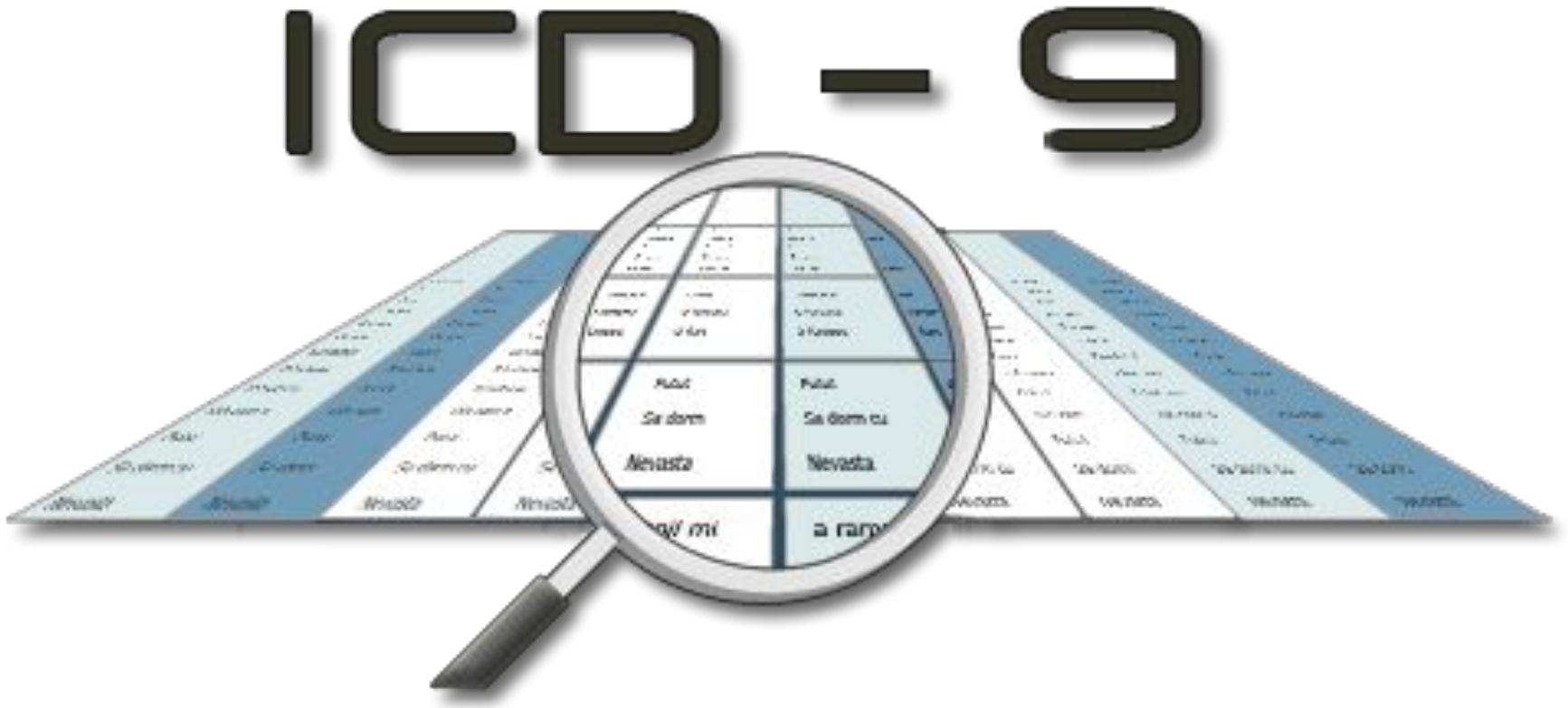
Considerations and challenges when working with clinical data

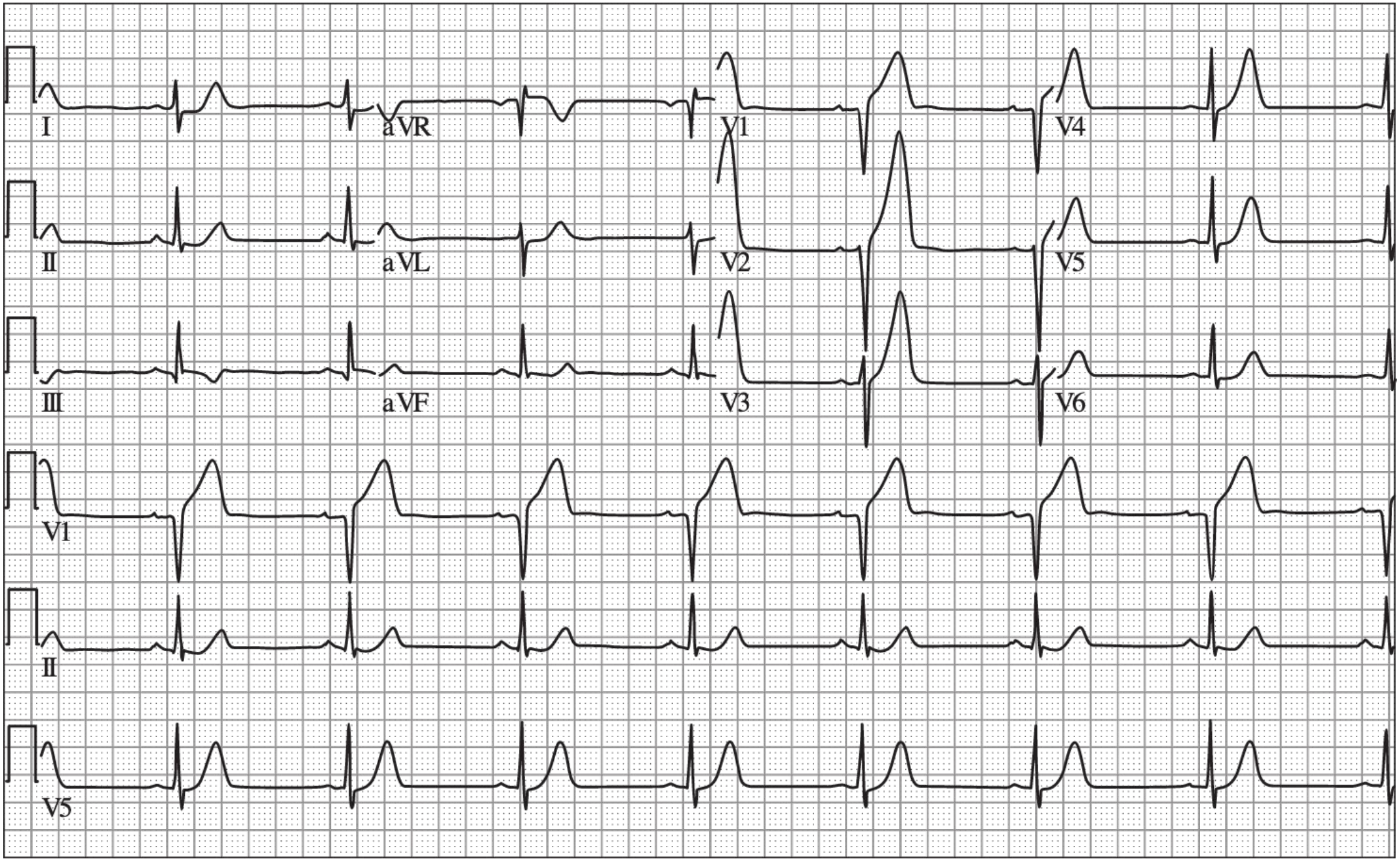
1. Data relevance
2. Incomplete data
3. Imbalanced data

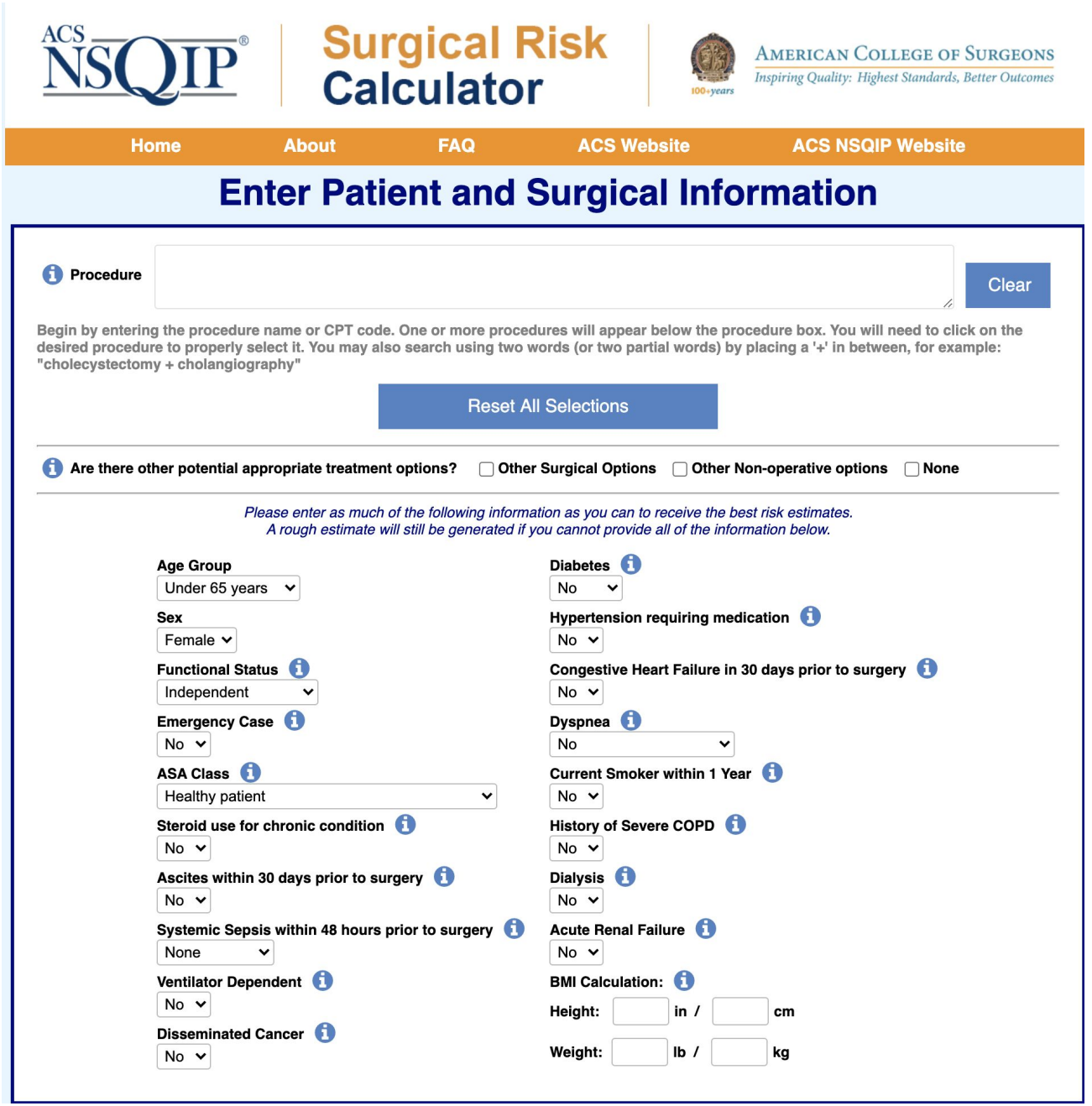
Considerations and challenges when working with clinical data

1. **Data relevance**
2. Incomplete data
3. Imbalanced data


Title
A
B
C
D

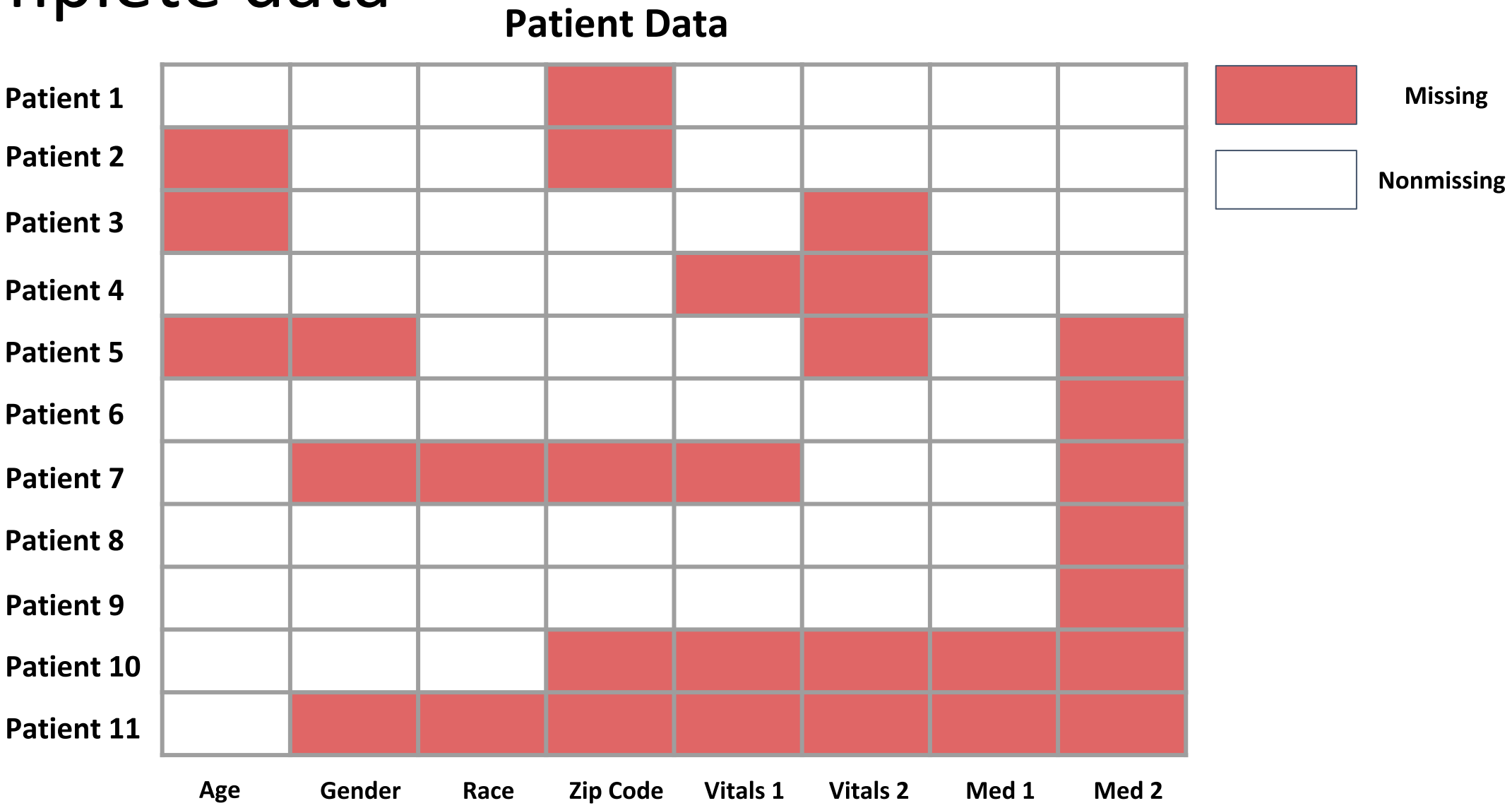

ICD - 9


I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
V1
II
V5

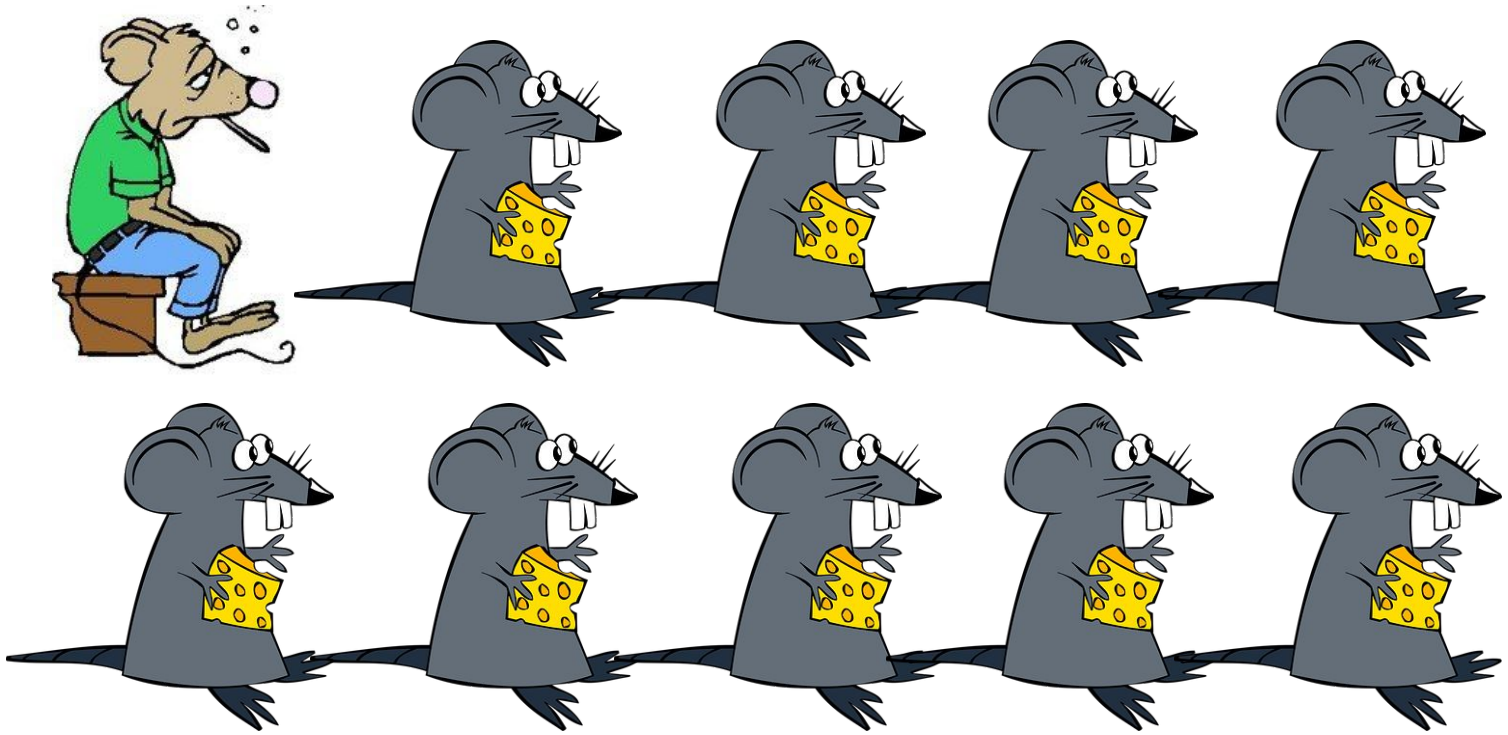


http://riskcalculator.facs.org/RiskCalculator/PatientInfo.js

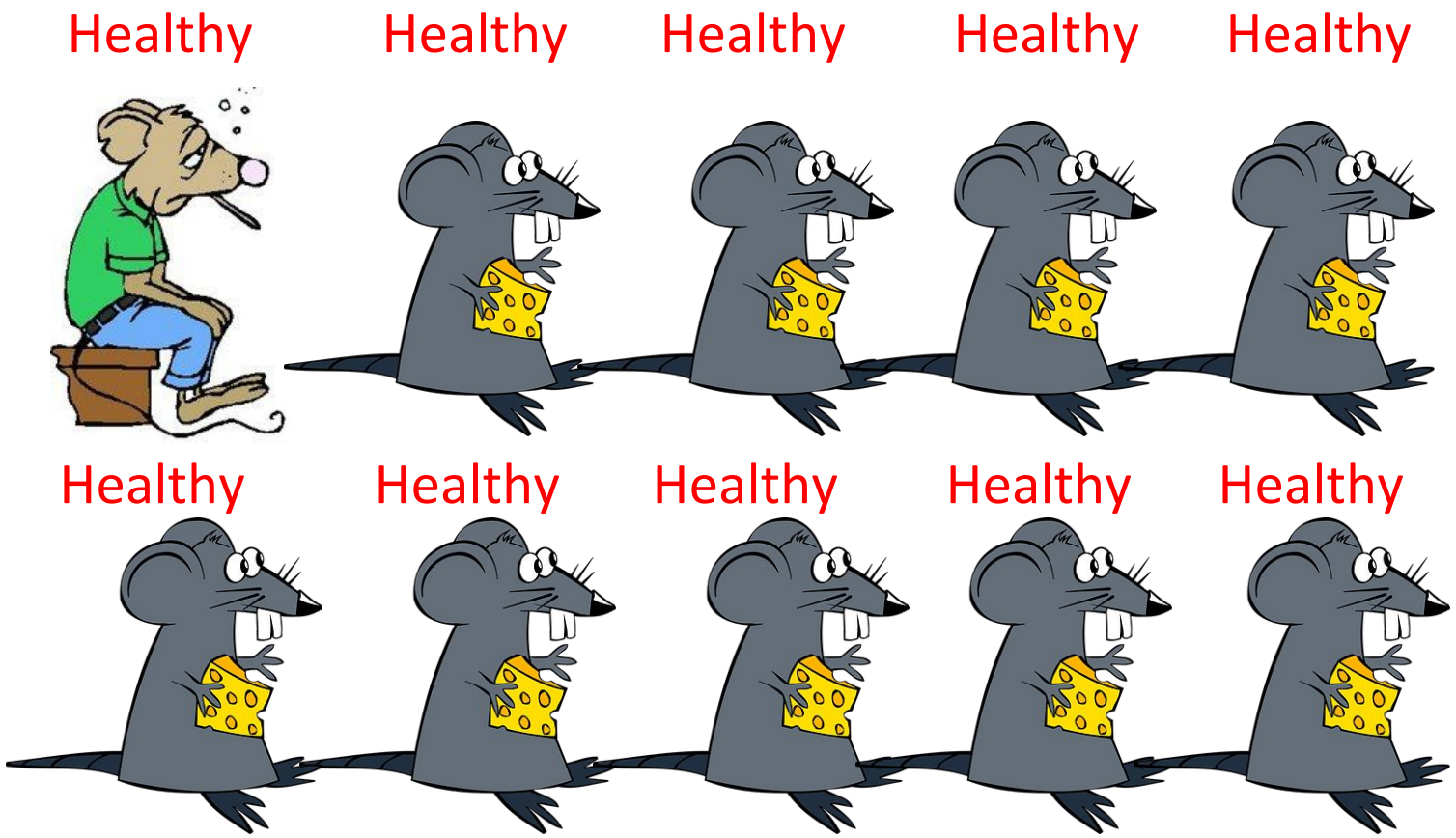
Incomplete data



Class imbalance



Why imbalance data is a challenge?



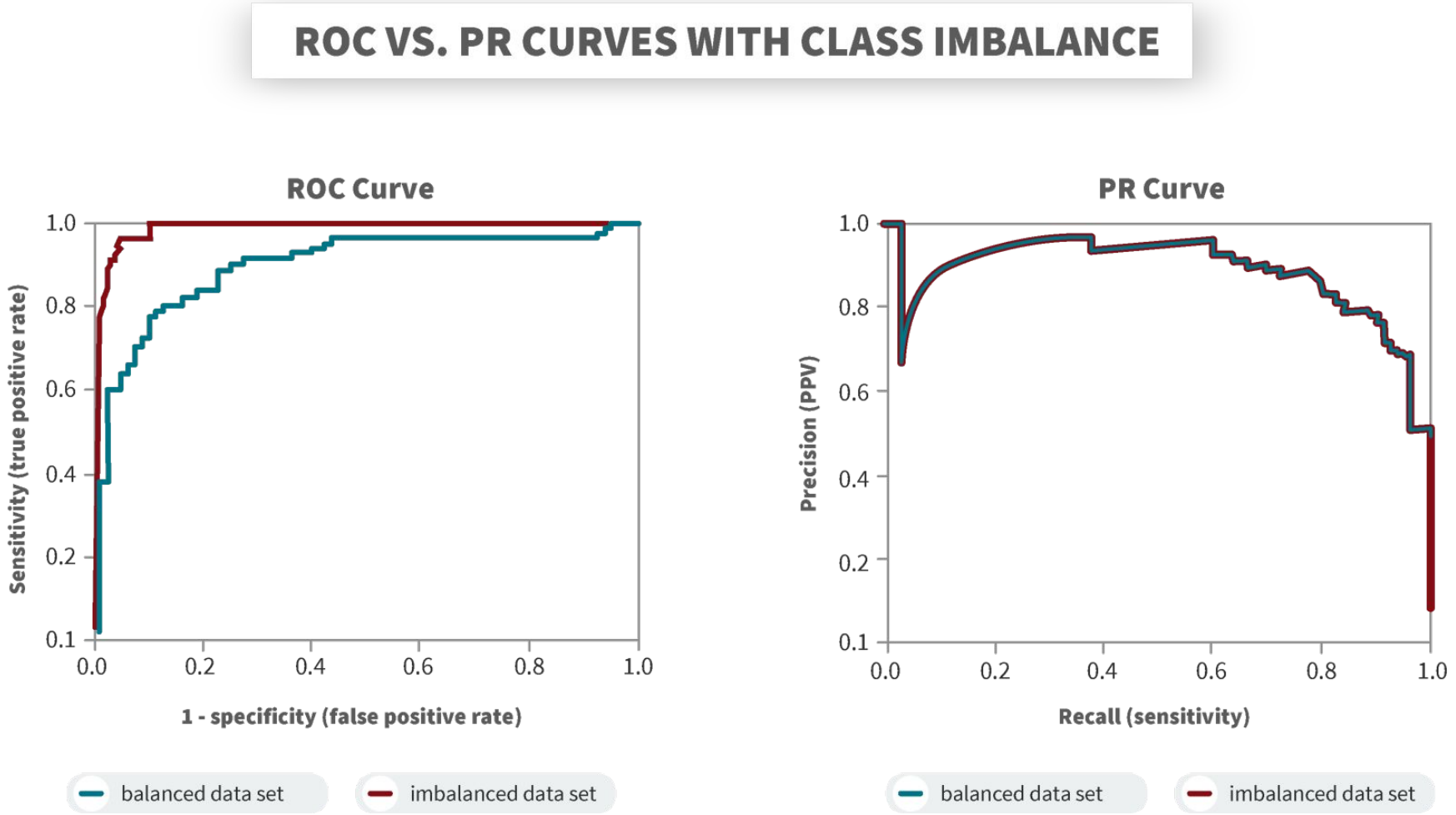
Model Accuracy = 90%!

Ways to address imbalance data

1. Choose better metrics
2. Sampling training set
3. Choose better models
4. Choose better loss functions

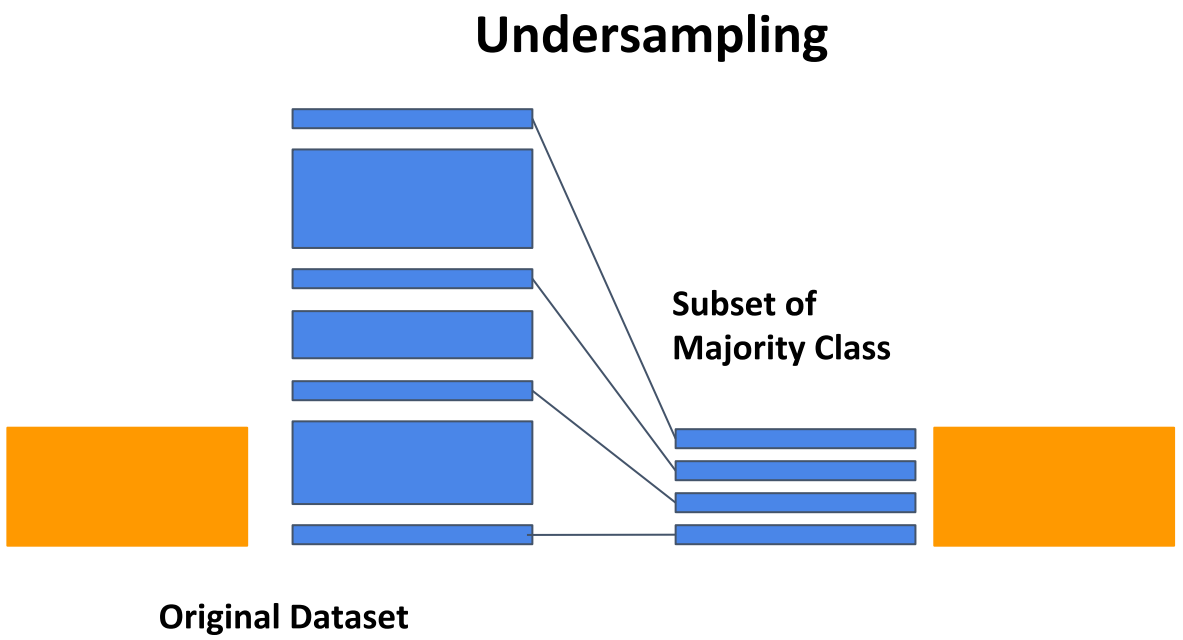
How to address the class imbalance problem:

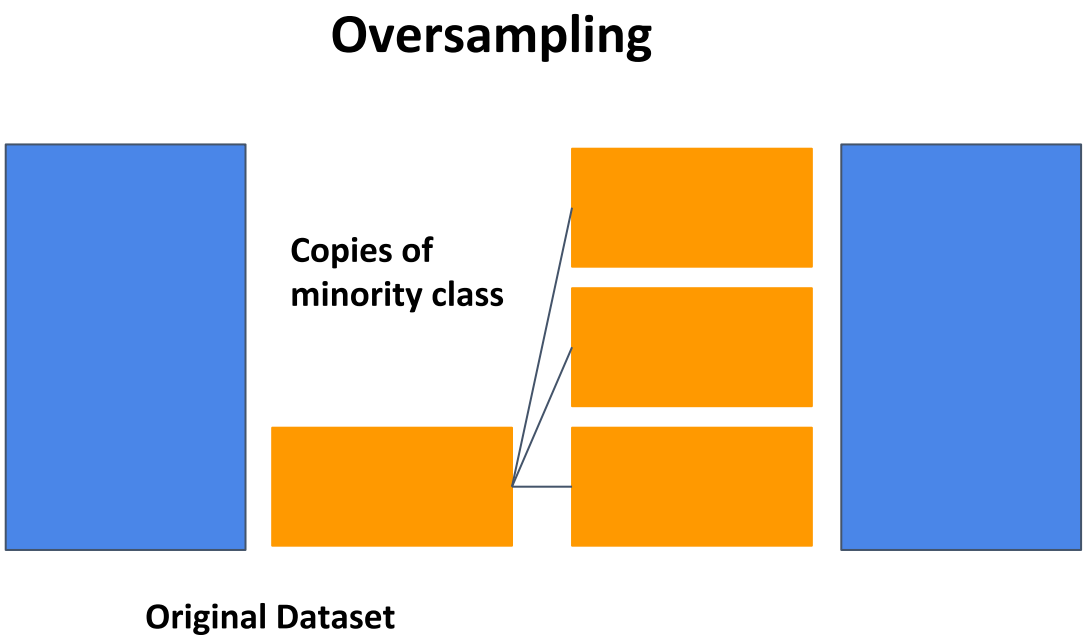
- choosing better metrics



How to address the class imbalance problem:

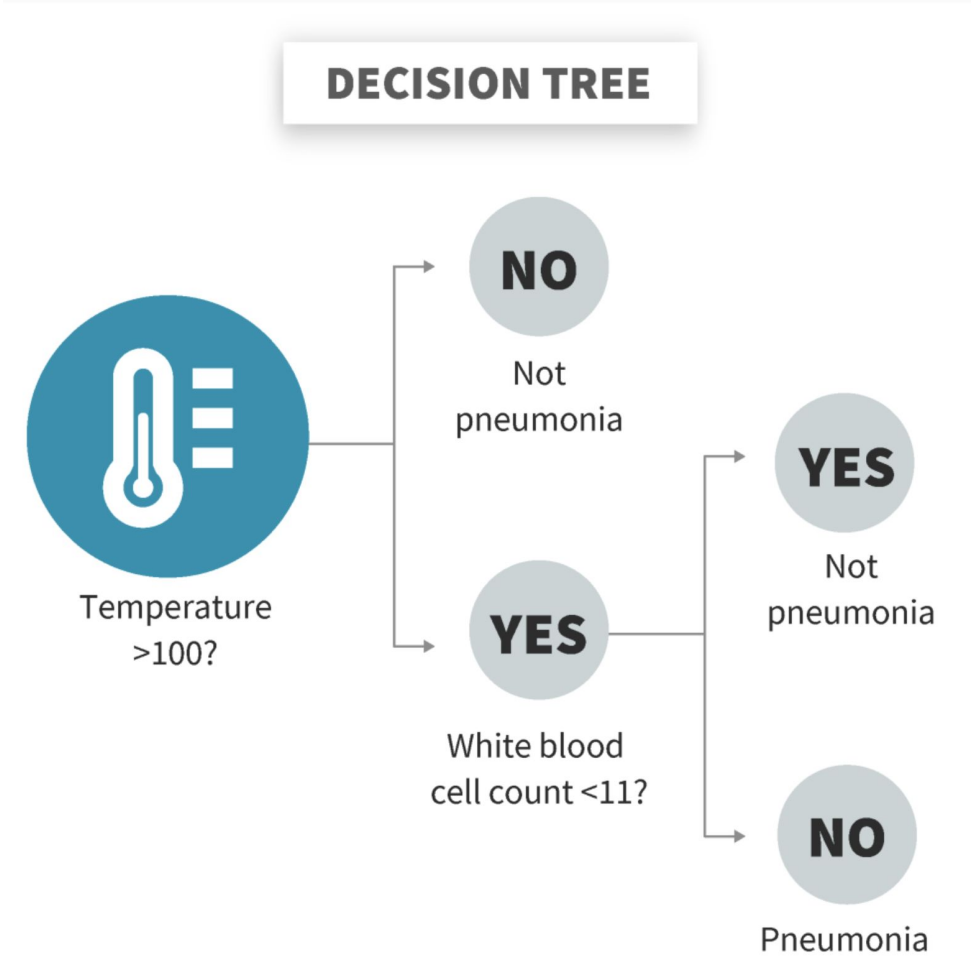
- Sampling training set





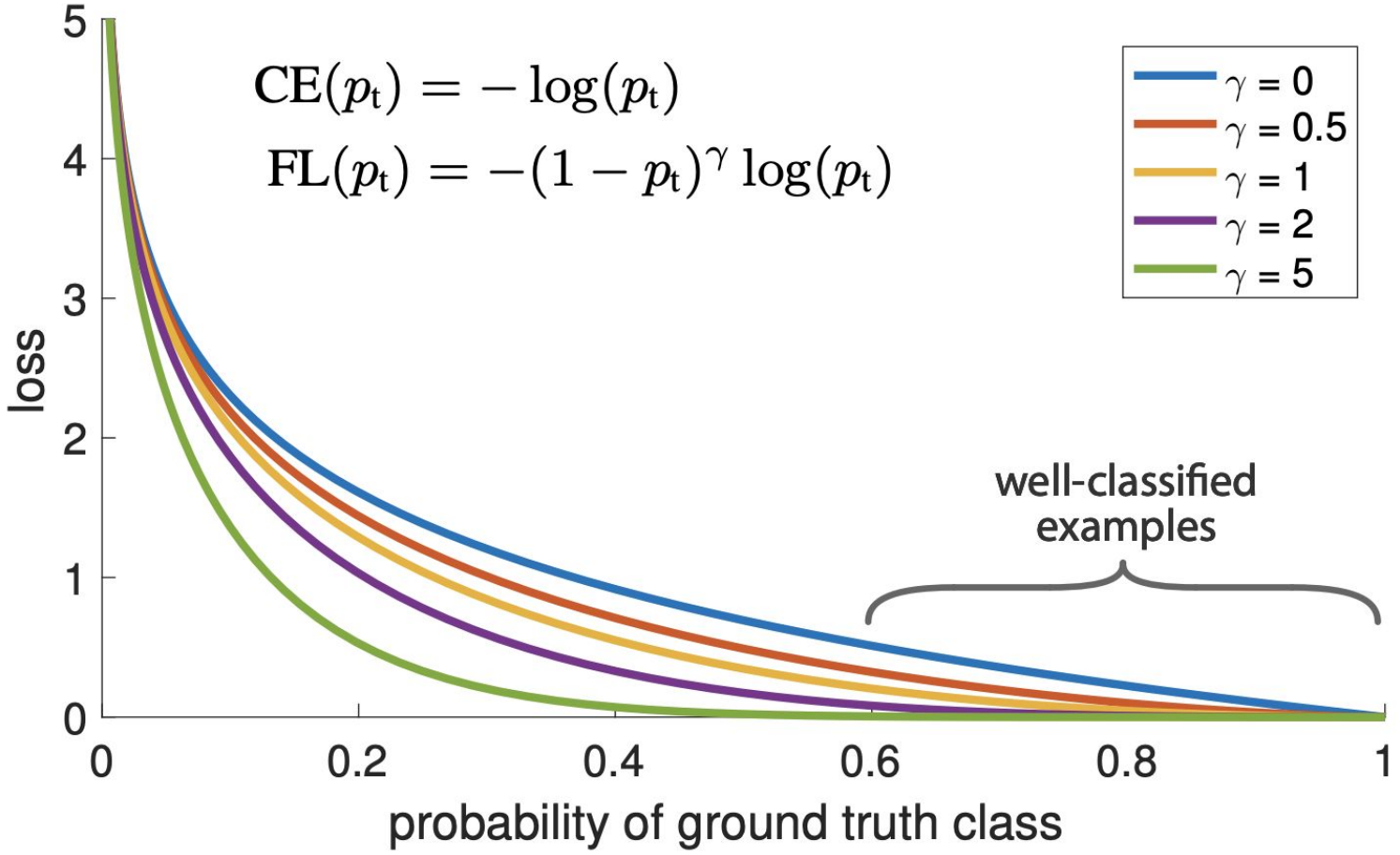
How to address the class imbalance problem:

- Choosing better model



How to address the class imbalance problem:

- Choosing better loss function



Outline

1. Types of EHR Data
2. Challenges when working with clinical data
3. **How Much Data is Needed?**
4. Assessment of quality of data

Why consider getting more data?

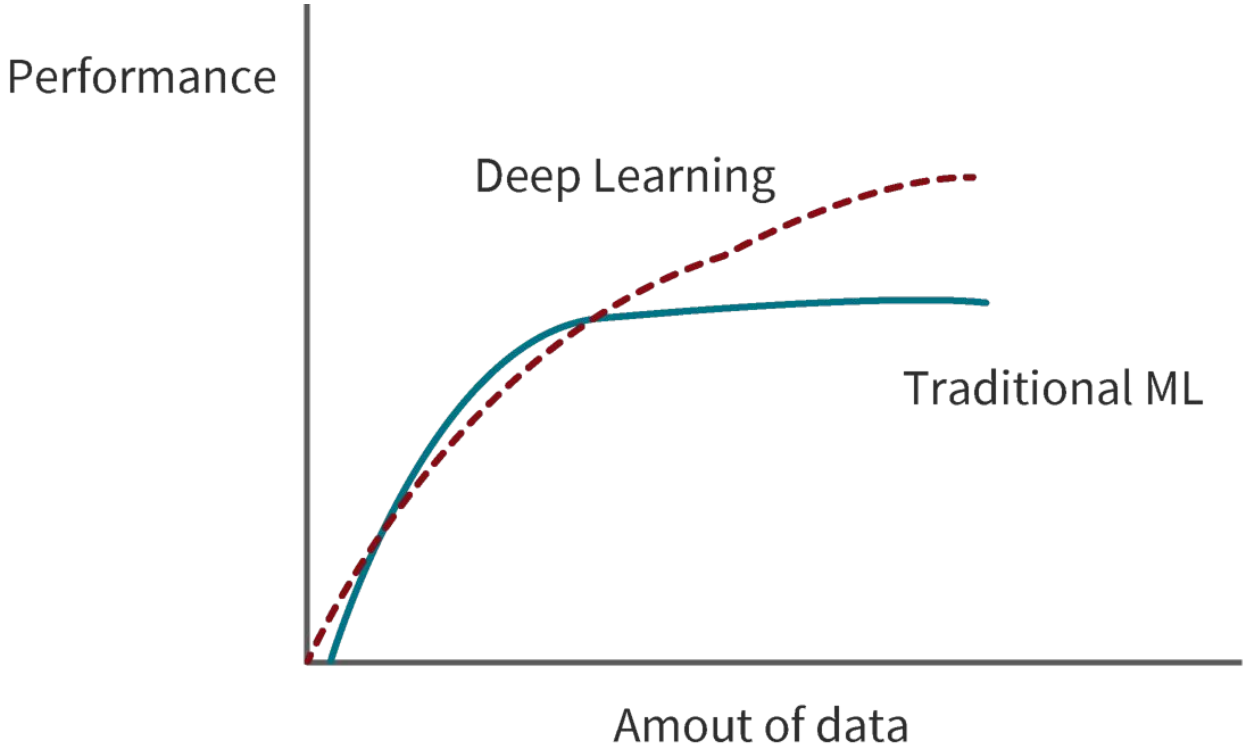
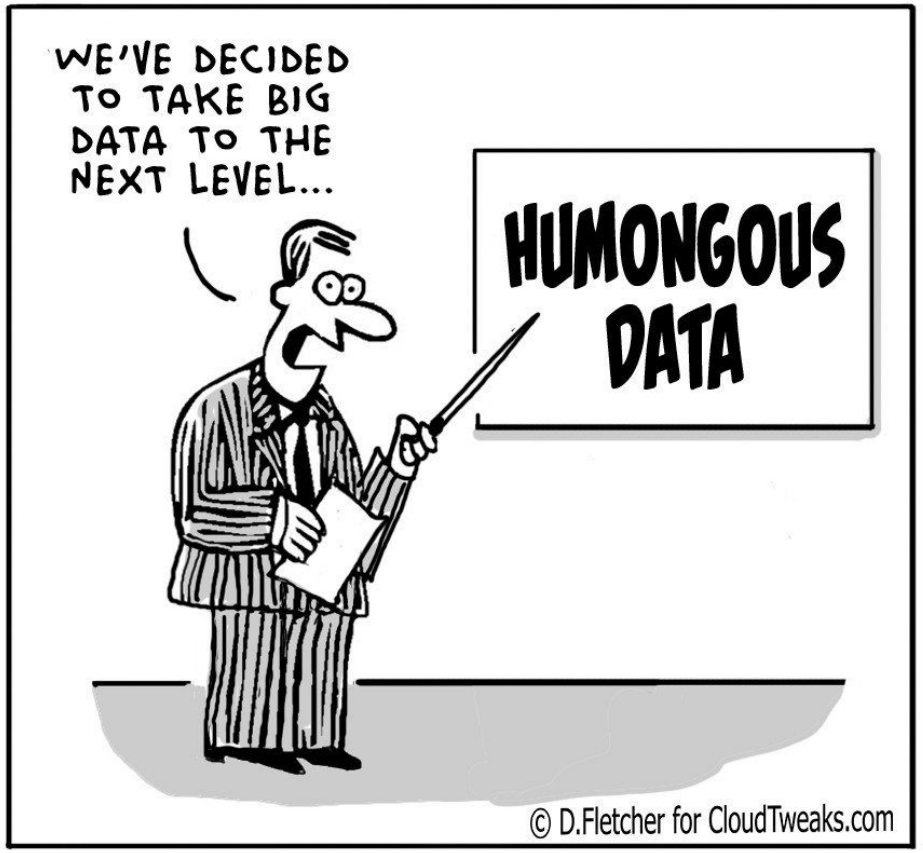


Figure shows how the performance of machine learning algorithms changes with increasing data size in the case of traditional machine learning algorithms (regression, etc.) and in the case of deep learning.
Specifically, for traditional machine learning algorithms, performance grows according to a power law and then reaches a plateau. This plateau is often higher with deep learning approaches for certain tasks and is one of the reasons for widespread use for applications with large data availability

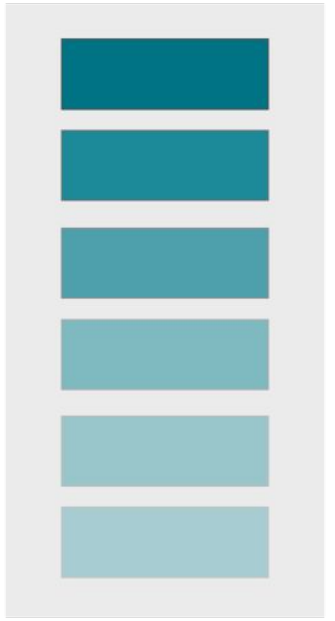
Too much data?



Bloodletting was used to "treat" a wide range of diseases, becoming a standard treatment for almost every ailment

Risk of using historical clinical data

2 months retrospective EHR data



12 months of retrospective EHR data

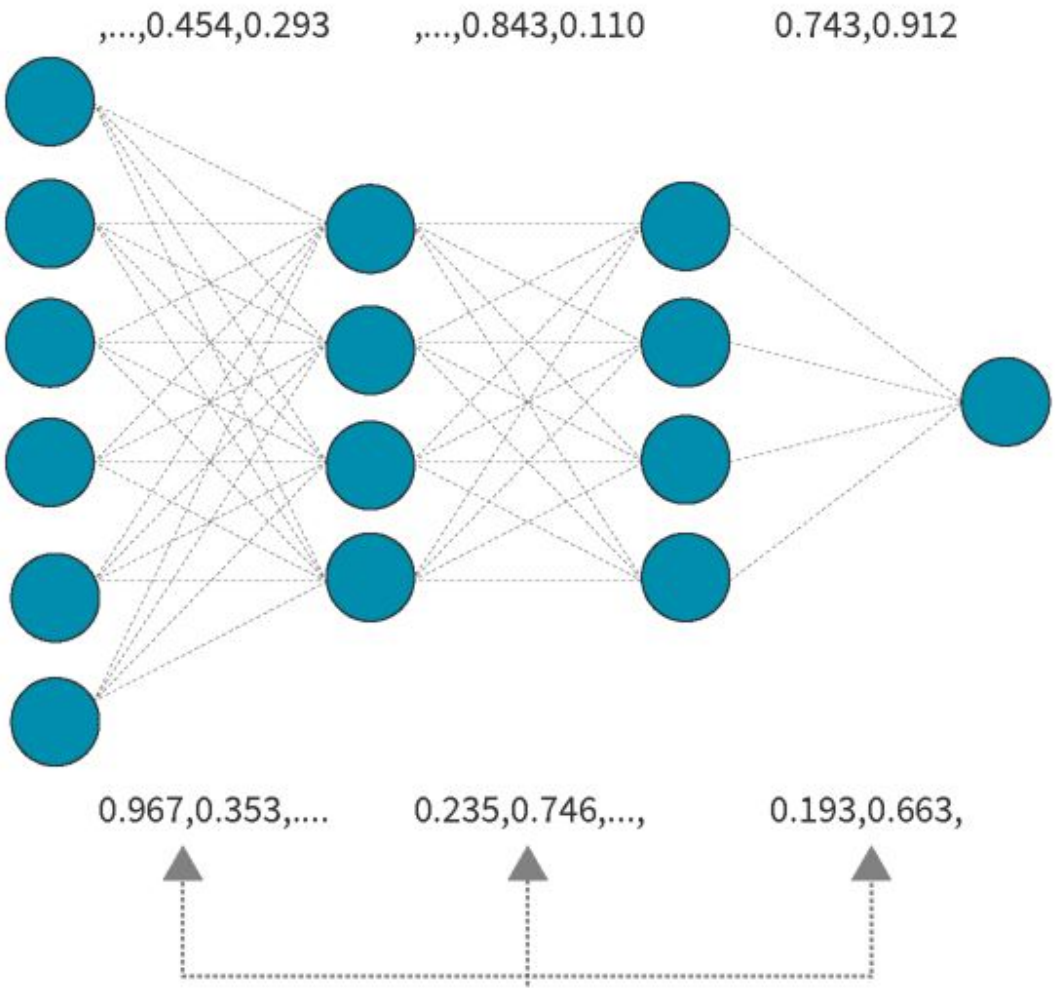
Model performance with increasing amounts of retrospective data

Outline

1. Types of EHR Data
2. Challenges when working with clinical data
3. How Much Data is Needed?
4. **Quality of data**

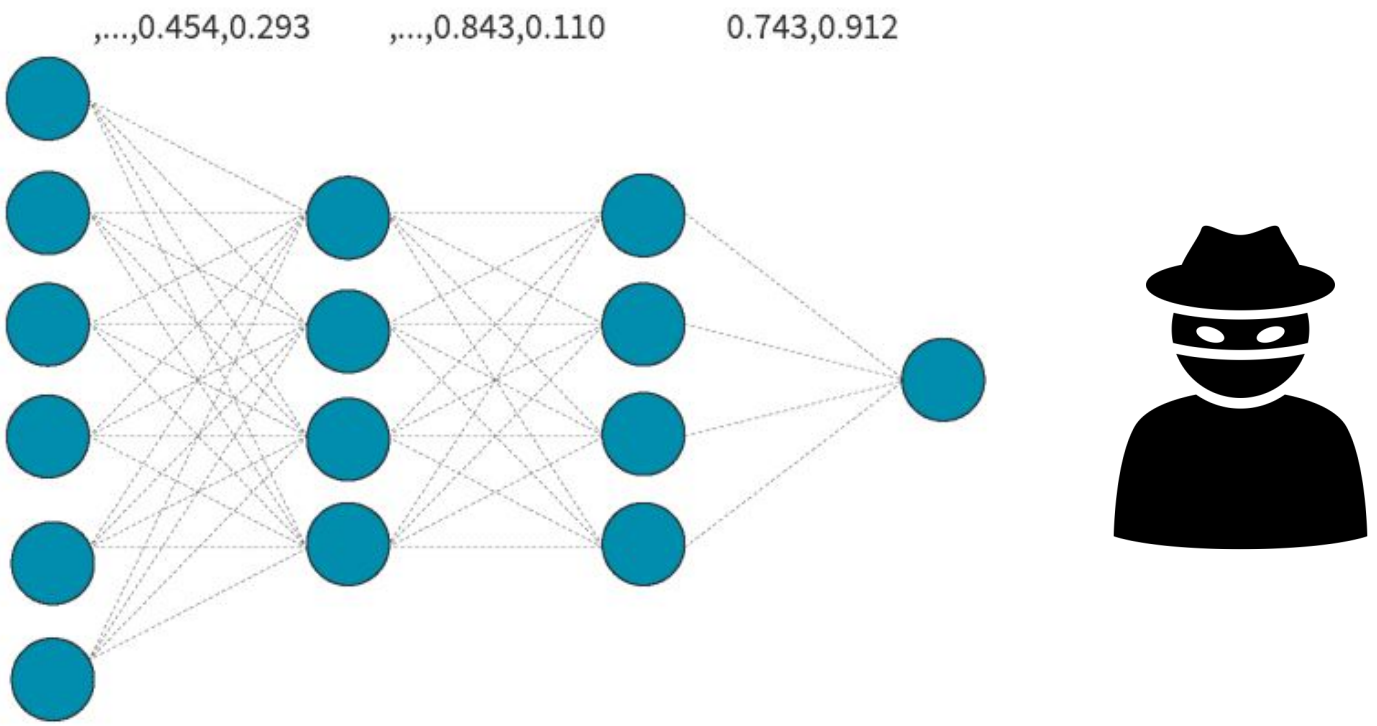
Risk of using poor data

- Garbage in garbage out



Risk of using poor data

- Garbage in garbage out

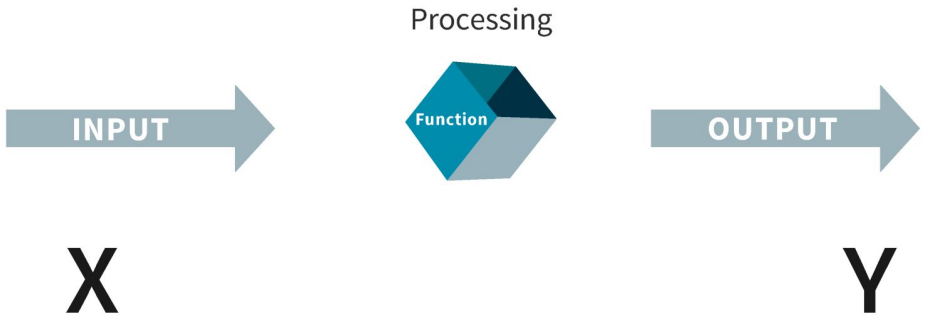


- Check: "*Automated Inference on Criminality Using Face Images*". 2017

  https://arxiv.org/abs/1611.04135

Define Labels and Ground truth

- Be mindful of how each data-point, label and ground truth are created
- Document how each data-type is extracted or curated and make your data reproducible.
- Relevant medical definitions and ontologies are important
- Keep a log of potential biases that may have occurred during data extraction
- Consideration of the ground truth and how well it matches reality is a critical task in labeling data.



Inaccurate Ground truths

- *Example*: Mortality can be straightforward on a timeline of days, but how exact are we on a timeline of minutes?

- *Example*: Hypertension or diabetes; rely on numerical cutoffs that have changed over time in medical practice. Consideration of data “shelf life” is very important, also for treatment and change in reporting standards.

- *Discussion Example*: How would you define pneumonia diagnosis for your patient selection?

Inaccurate Ground truths

2007 Infectious Diseases Society of America/American Thoracic Society Criteria for Defining Severe Community-acquired Pneumonia

Validated definition includes either one major criterion or three or more minor criteria

Minor criteria

- Respiratory rate ≥ 30 breaths/min
- Pa_{O2}/F_{IO2} ratio ≤ 250
- Multilobar infiltrates
- Confusion/disorientation
- Uremia (blood urea nitrogen level ≥ 20 mg/dl)
- Leukopenia* (white blood cell count < 4,000 cells/μl)
- Thrombocytopenia (platelet count < 100,000/μl)
- Hypothermia (core temperature < 36°C)
- Hypotension requiring aggressive fluid resuscitation

Major criteria

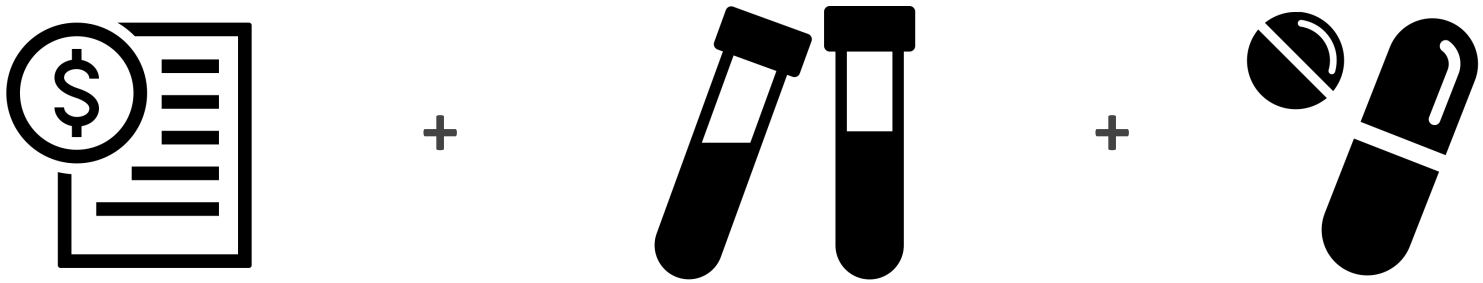
- Septic shock with need for vasopressors
- Respiratory failure requiring mechanical ventilation

Understand Noise in Data

- Get help from domain experts for understanding your dataset and its labels, including associated uncertainties and biases
- Get multiple expert reviewers to label data and check agreement between multiple reviewers. This gives an understanding of the uncertainty or noise of labels
- Report potential biases and interobserver (and intraobserver) agreement

Handling Noise in Data

- Get multiple expert reviewers to label data/ground truths and use average or majority voting. This reduces noise.
- Get labels on a subset of data to confirm the quality of a surrogate label (i.e. ICD codes for diagnosis)
- Combine multiple noisy labels for a more accurate label; i.e. instead of only ICD-code, also require a positive test and administration of medication. This comes at the expense of narrowing the cohort and possible selection bias.
- After adding multiple noisy labels you can compare with a subset of data hand-labeled by experts



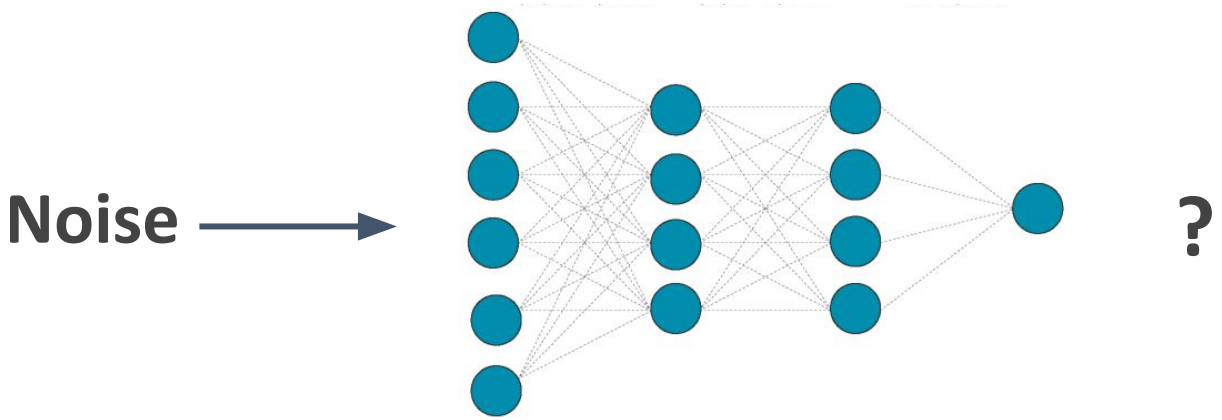
Data quality vs quantity

- Stanford team built a predictive model based on EHR data for acute and chronic disease prediction

- Researchers experimented with adding noise and more data to learn about the effects of the models performance on a ground truth test set.

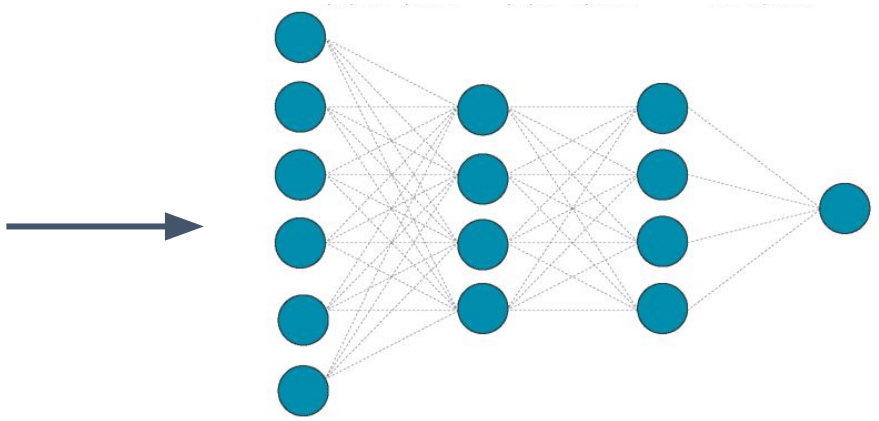
- n=2000, 100% accurate ground truths in training →benchmark performance
- n=3000, 90% accurate ground truths in training →benchmark performance
- n=4000, 85% accurate ground truths in training →benchmark performance

- Noise in data can sometimes have a regularizing effect and prevent model overfitting.



You can train a model with noisy data

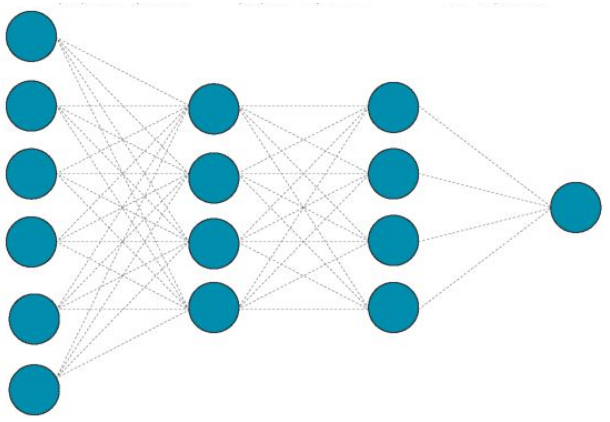
- Important concept; you can overcome data label noise by increasing data volume.
- If you have a scalable way to label large datasets without much effort, even if its noisy, you can overcome the issues of noise with volume.
- In the Stanford study, 10% noise required 50% more data, and 15% noise required 100% more data.
- These numbers for extra training data may not hold for other datasets and data types, but the concept of addressing noise with more training data is very important

More data 

Weak Supervision

- Weak supervision is the ML technique where noisy or imprecise labels are used to provide supervision signal for in a supervised machine learning setting.
- This approach removes the burden of obtaining ‘expensive’ hand-labeled data sets
- Important to remember not to have noisy / weak labels in test set. Why?

“Weak” training labels →



Summary

Today we covered:

- Types of EHR Data
- Challenges when working with clinical data
- How Much Data is Needed?
- Assessment of quality of data

Coming up next time: **Team-based design and evaluation of clinical machine learning applications**