

# THE DATA'S JOURNEY

Hypothesis



Validation



Cleaning & Normalization



Data Exploration



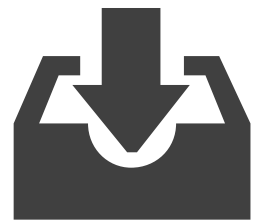
Data Collection



Analysis



# CONTINUING OUR JOURNEY



**DATA COLLECTION**



**EXPLORATORY  
DATA ANALYSIS**



**DATA ANALYSIS**

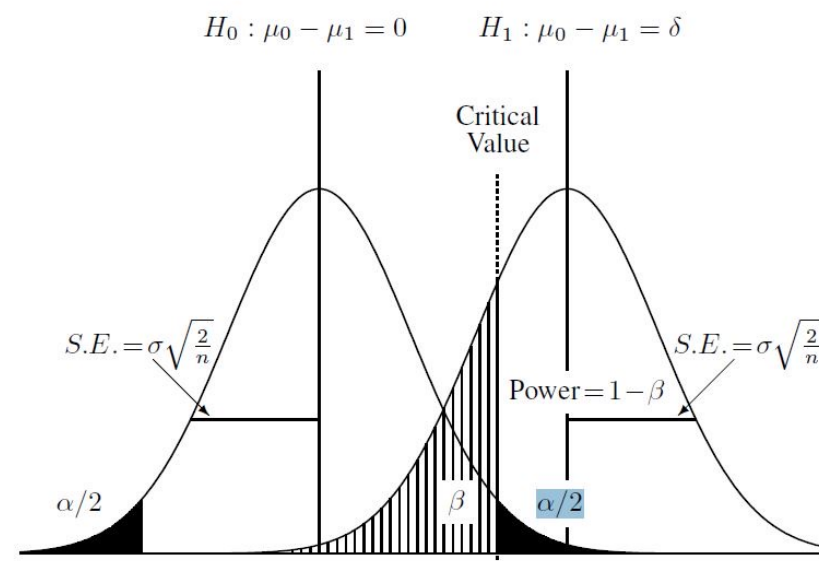


**MODEL EVALUATION**



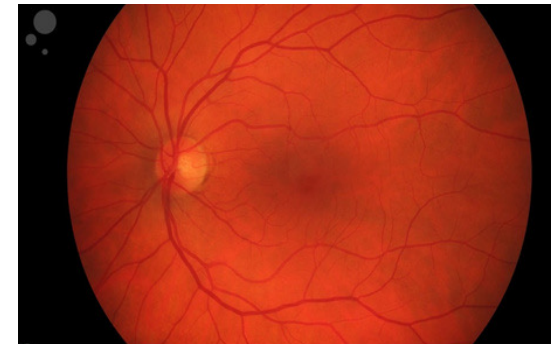
# DATA SCIENCE ANALYSIS

## HYPOTHESIS TESTING



**RQ:** Is there a significant difference in the variable of interest between two or more groups?

## PREDICTION/ CLASSIFICATION



**RQ:** Does this patient have a cataract?

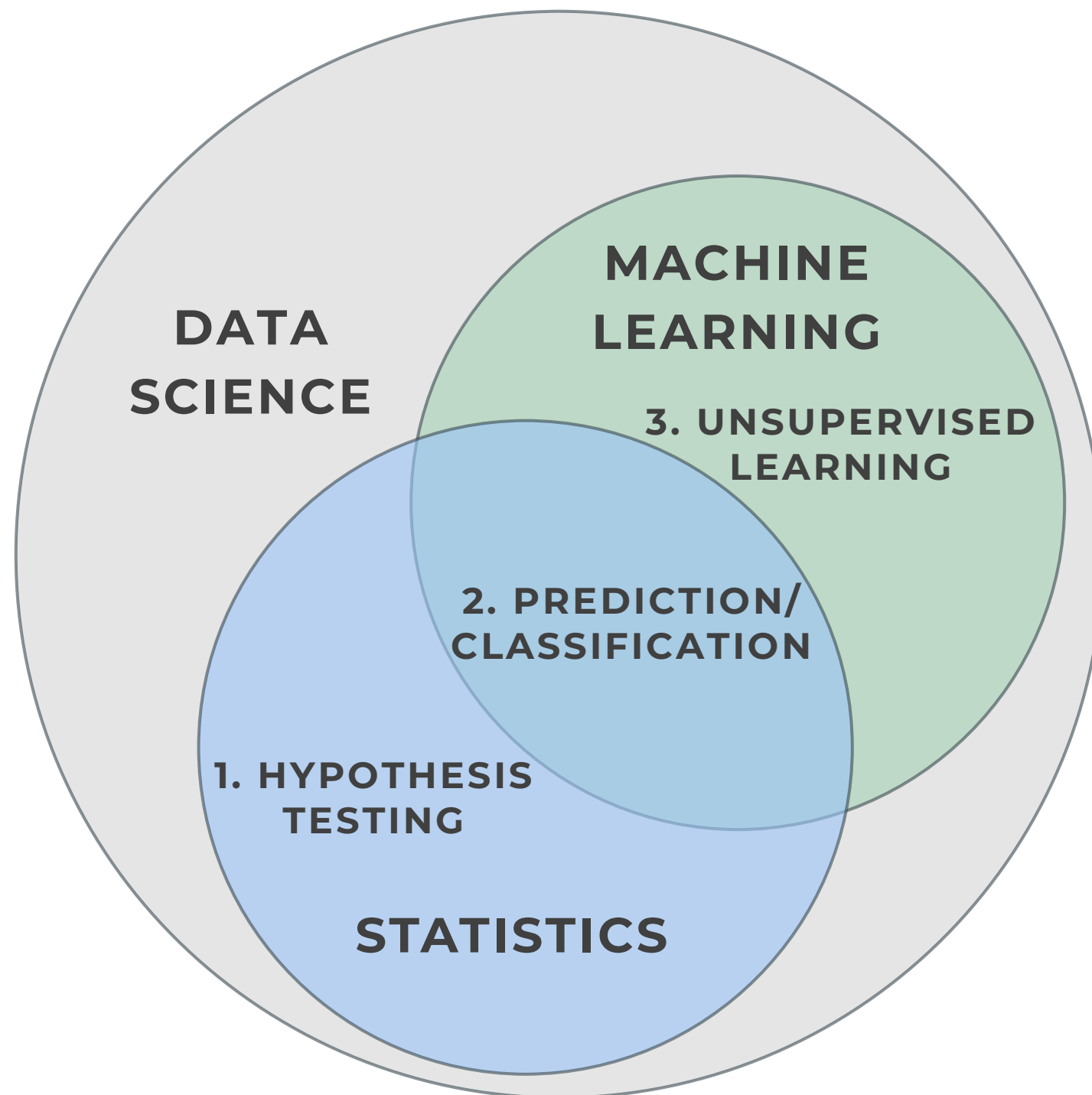
**RQ:** Can we estimate cancer risk based on genetic risk, smoking and age?

## UNSUPERVISED LEARNING



**RQ:** What sets the cancer cells apart from the healthy cells?  
Are there subtypes within one cell type?

# DATA SCIENCE ANALYSIS

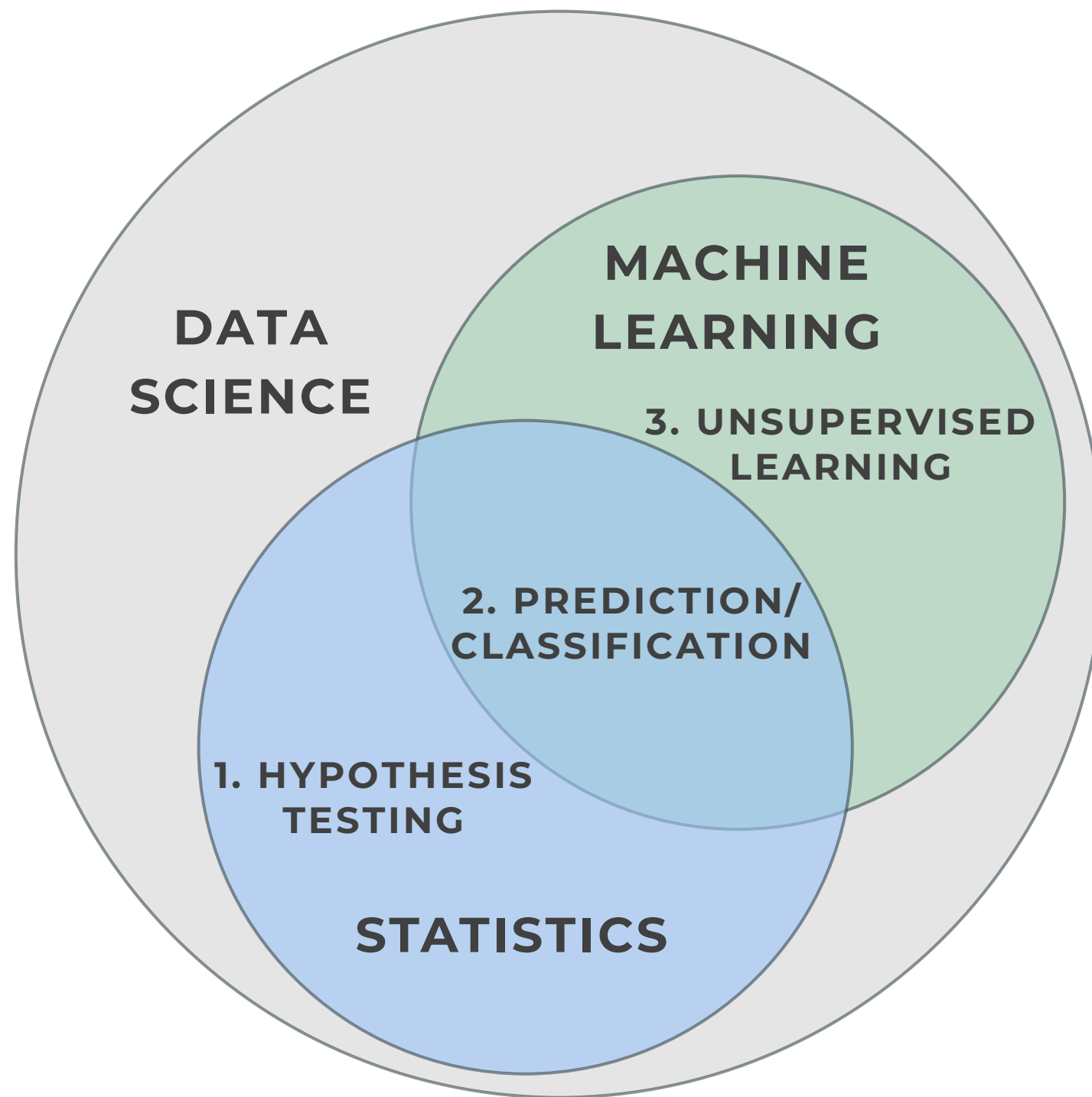


## STATISTICS:

- Focus is on inference\*, achieved through fitting of a probability model
- Quantitative measure of confidence of a 'true' effect (confidence interval)
- We have data/model assumptions which we must verify
- Statistical models can do predictions, but predictive accuracy is not their strength<sup>[1]</sup>. BUT interpretable.

[1] Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat Methods* **15**, 233–234 (2018).

# DATA SCIENCE ANALYSIS



## MACHINE LEARNING:

- Purpose is prediction/classification, finding patterns in large data
- Makes minimal assumptions about the data
- When the number of input variables exceed observation
- Appropriate for complicated nonlinear interactions
- Provide a degree of interpretability, but tends to sacrifice interpretability for predictive power[1].

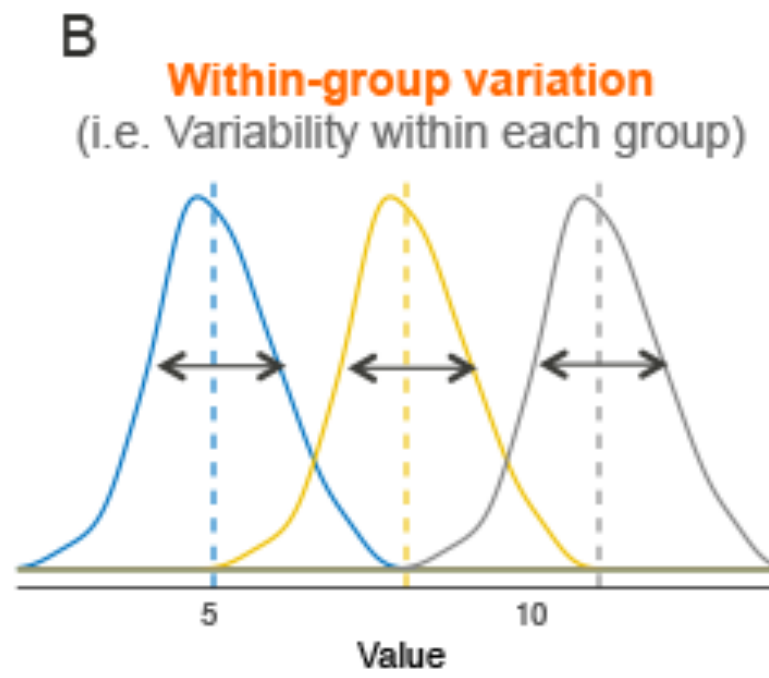
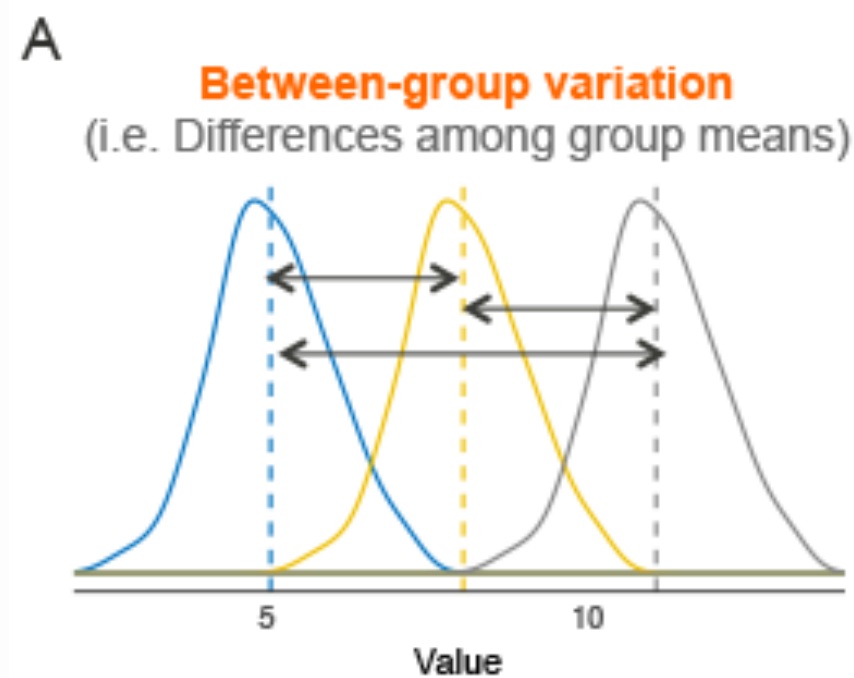
[1] Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat Methods* **15**, 233–234 (2018).

# HYPOTHESIS TESTING

## Questions that boil down to:

*Is there a difference in feature A between these two or more groups?*

- Is the bacterial load higher in colon swaps of cancer patients?
- Is the expression of gene A higher in tumor samples?
- Is there a difference in median height between men and women?



## A difference test:

- t-test
- ANOVA
- Fisher's exact test



# PREDICTIVE MODEL / CLASSIFICATION

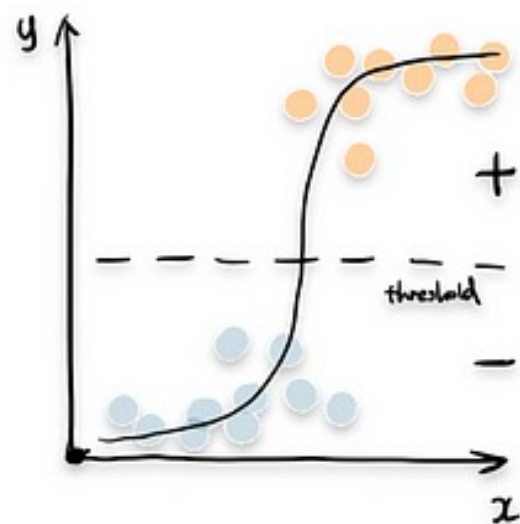
Using statistical/ML models to **predict outcomes / classify new data**.

Discovering and quantifying the relationships between predictor variables and outcome.

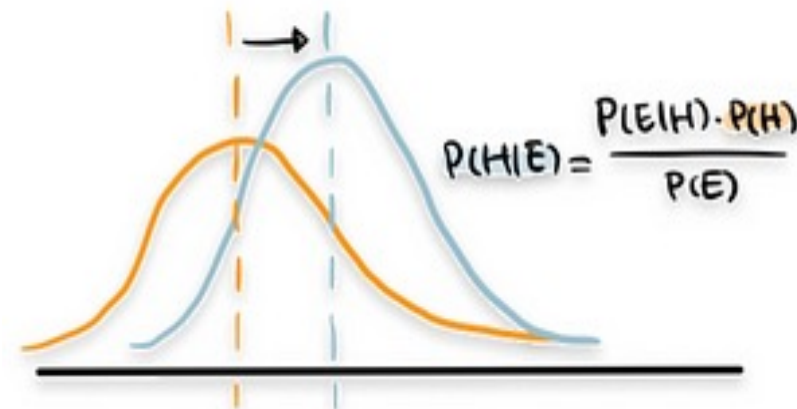
**Prediction (Regression)** == outcome is continuous (weight of newborn)

**Classification** == outcome is a class/group (cystic fibrosis or healthy)

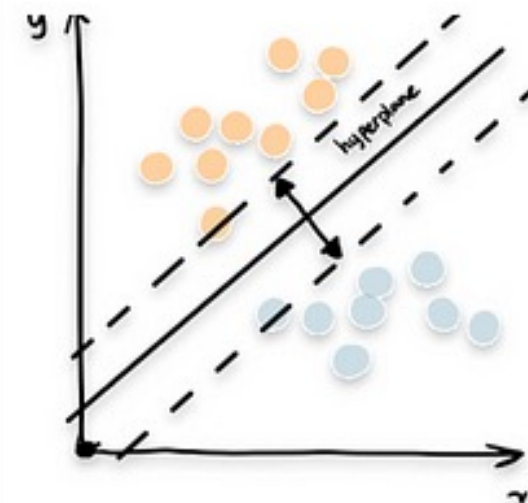
Logistic Regression



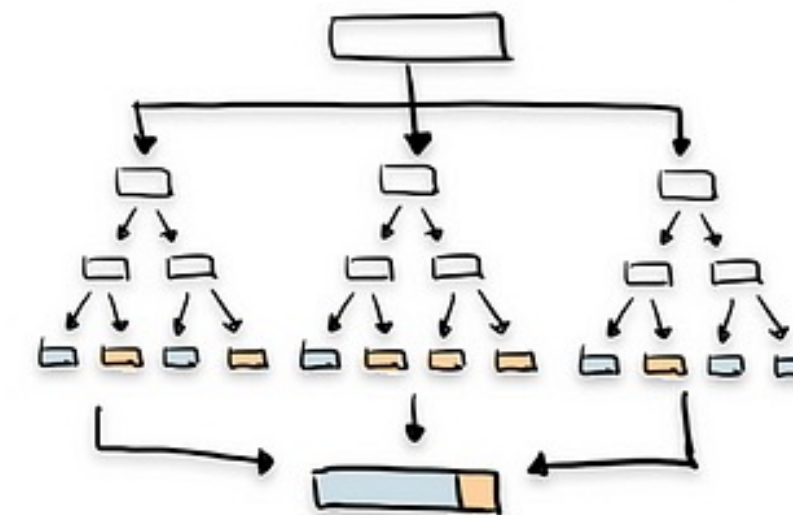
Naïve Bayes



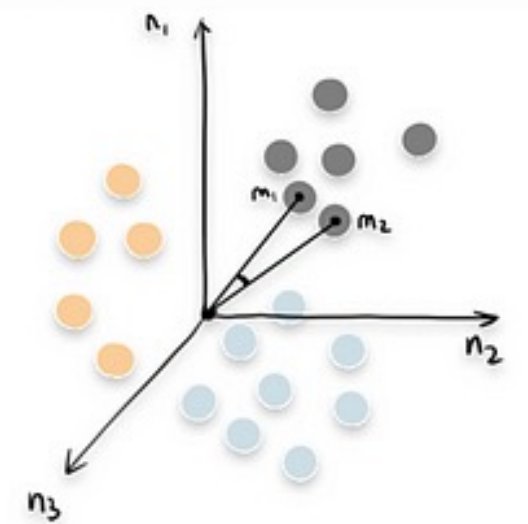
Support Vector Machine



Random Forest

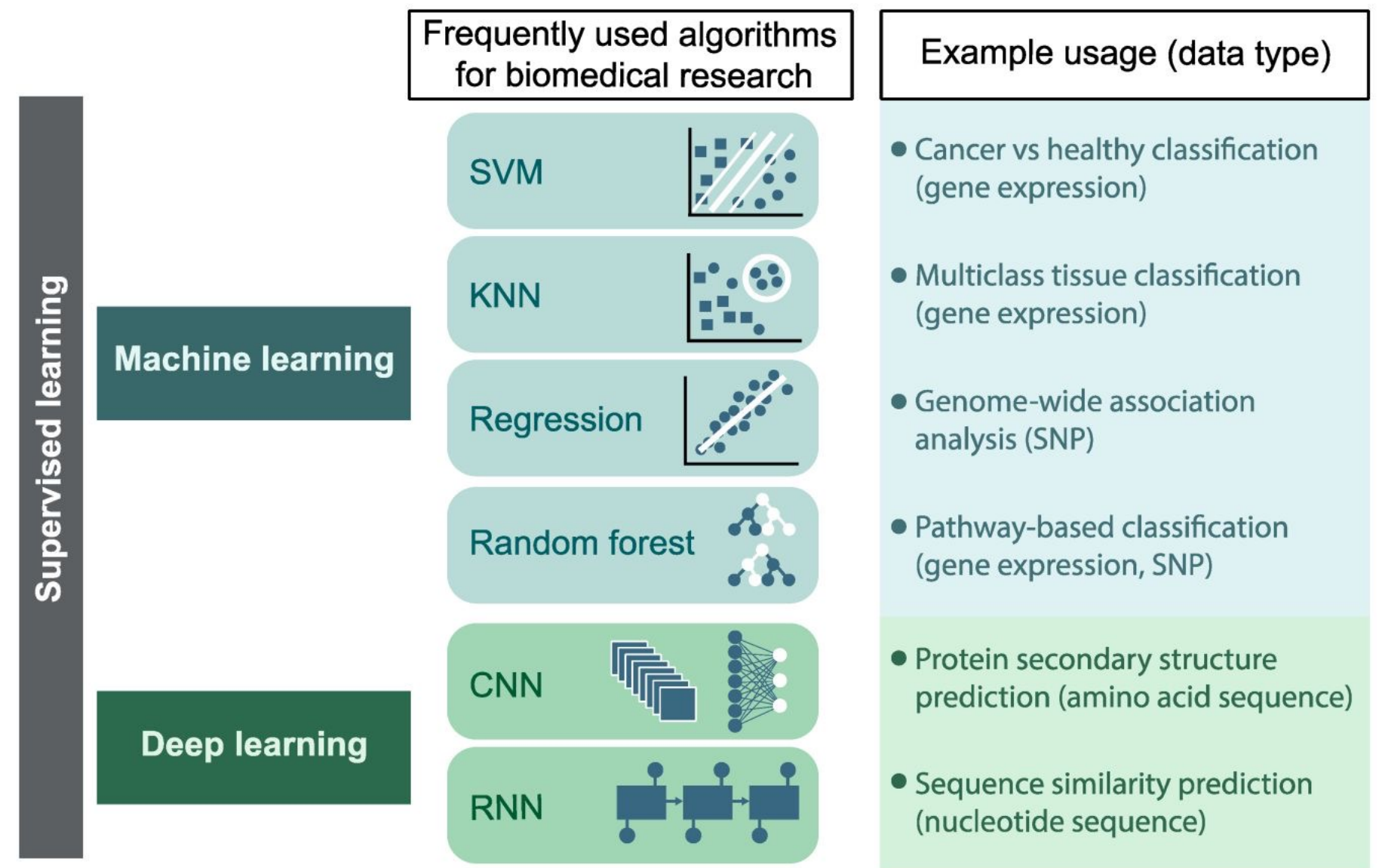


K-Nearest Neighbors



# SUPERVISED LEARNING

- **Supervised learning** is one category of machine learning models.
- **Supervised** means the ground truth is known.
- We have **labels** for the **outcome** when training our model, i.e. cancer or health sample





# CLASSIFICATION OF FRUIT - EXAMPLE

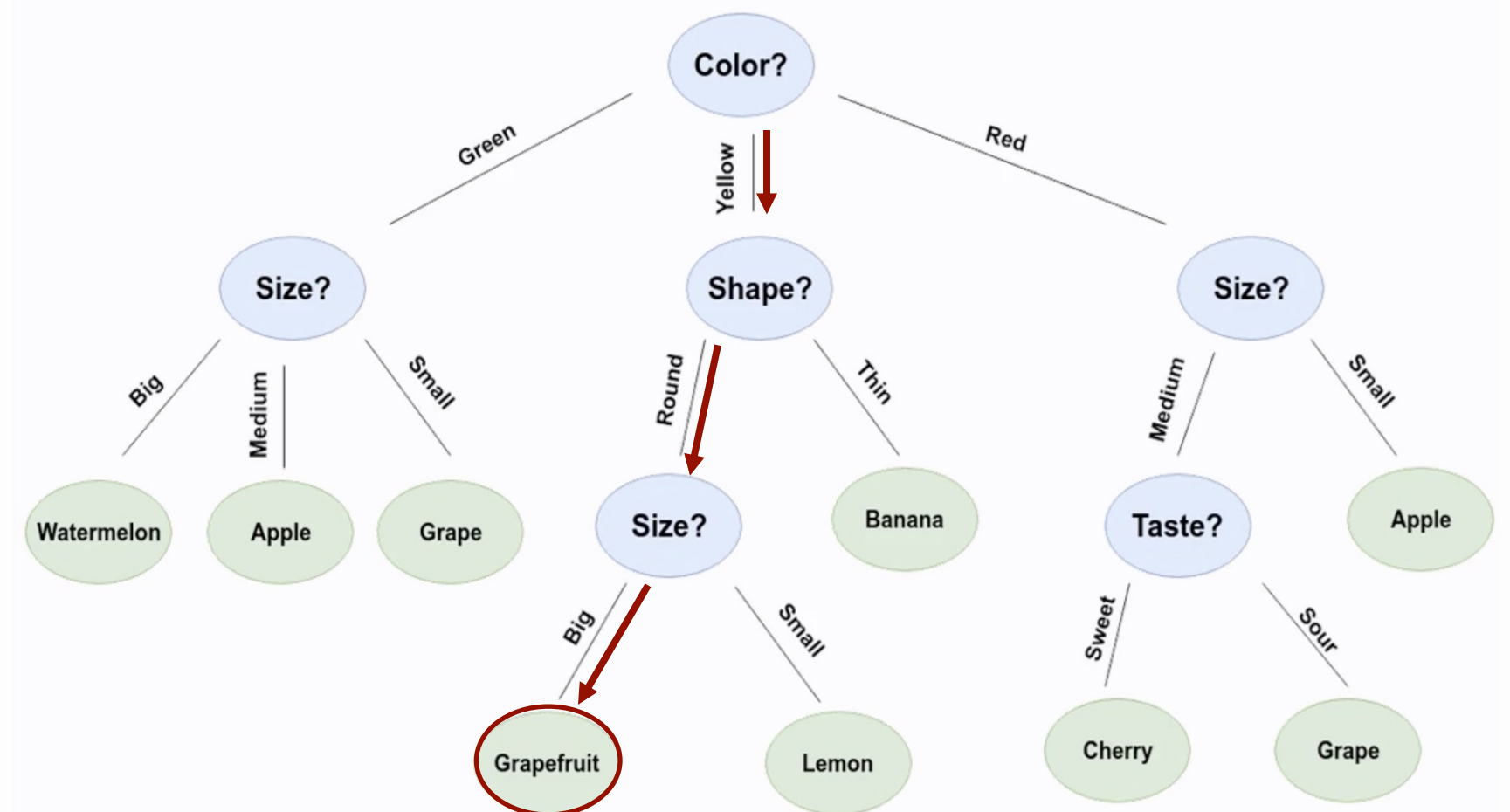
A model trained and evaluated using class labels is used to **classify** a new data point (label unknown).

- **Probability** (0.68 Grapefruit & 0.32 Lemon)
- **Response** (Grapefruit)

Some models have **feature importance ranking**.

What feature/variable is most important for classification, i.e. size, shape or color?

## Decision Tree – Fruit Classification



# GROUP DISCUSSION 3.1

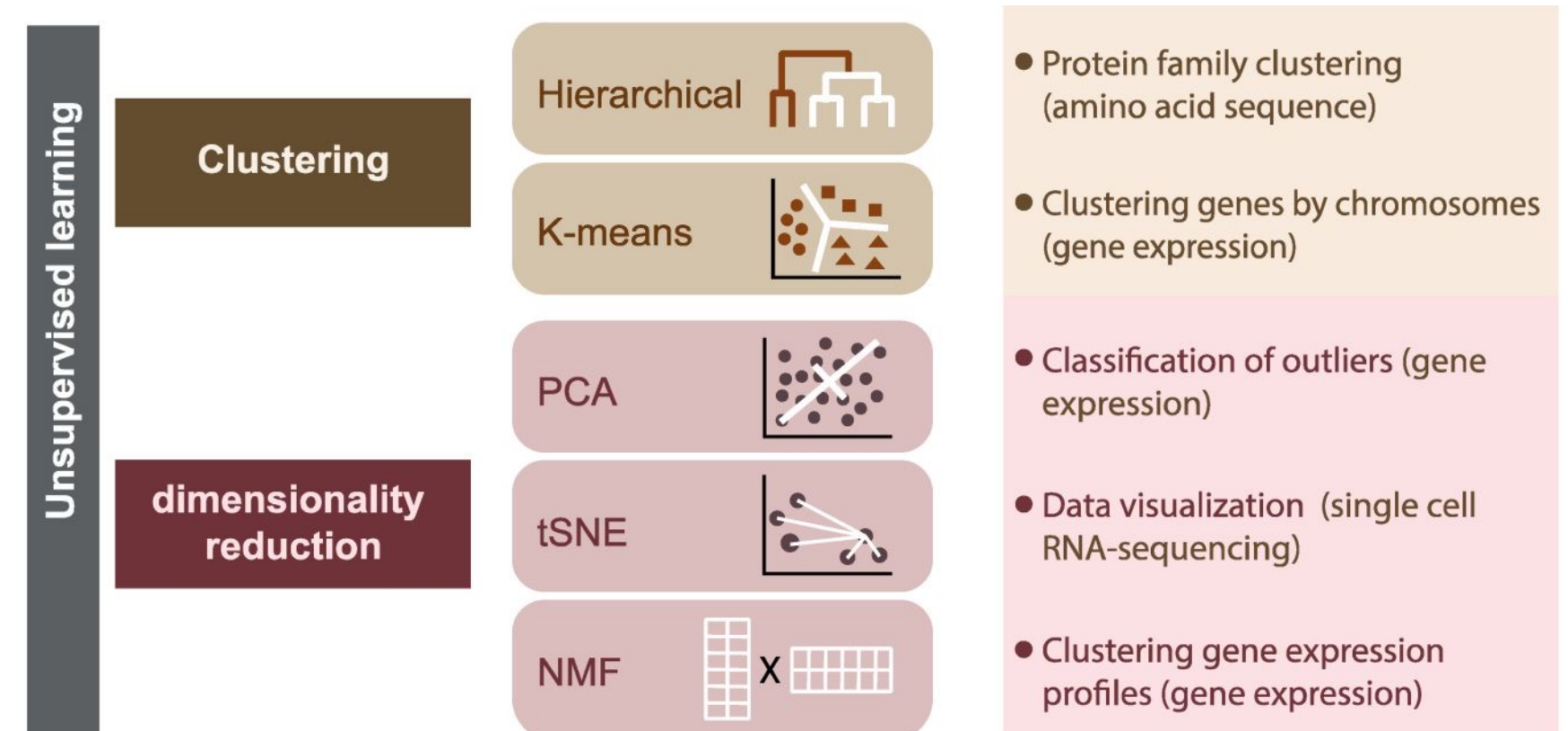
In your groups discuss:

- Within your field what type(s) of model(s) is most often used? (slide 8 for inspiration). Why do you think this model is favored, i.e. what scientific question are you trying to answer (look at slide 3).
- If you had no 'ground truth' (no labels) to use for model training, do you think you could investigate a classification problem regardless? If so, what would the scientific question be?



# UNSUPERVISED LEARNING

- **Unsupervised Classification:** Group observations into clusters.
- What if we do not know the groups our data partition into, **no labels**...
- A scientific question could be: *Do our observations stratify into groups and what data characteristics drive this partitioning?*
- For this we use **unsupervised learning** methods (PCA is one example).



# DIMENSIONALITY REDUCTION

**Dimensionality reduction (DR) methods** project data from a high dimensional space into a low dimensional space.

DR methods are **unsupervised**.

They can be either linear or non-linear.

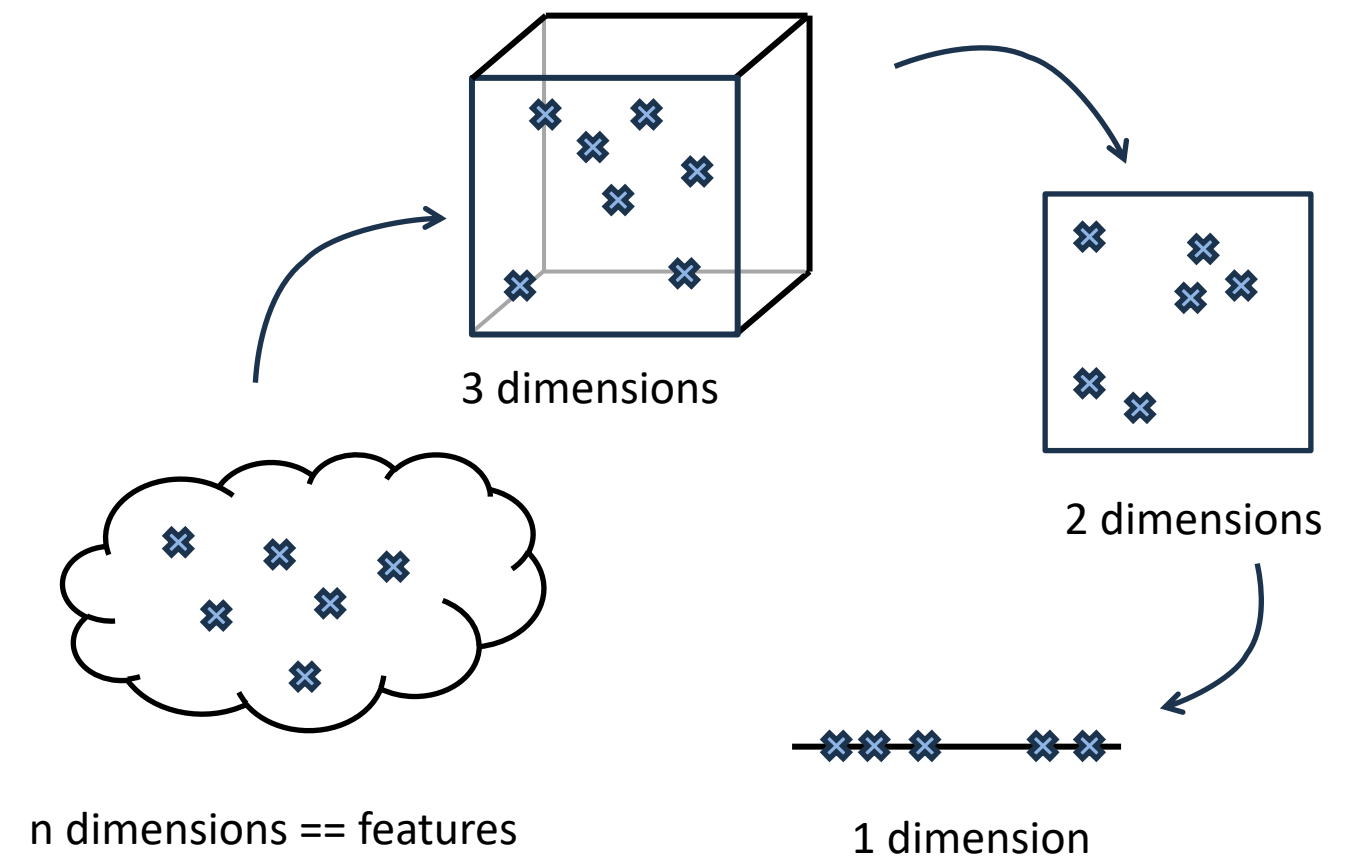
## Linear projection:

**PCA** = principal component analysis

## Non-linear projection:

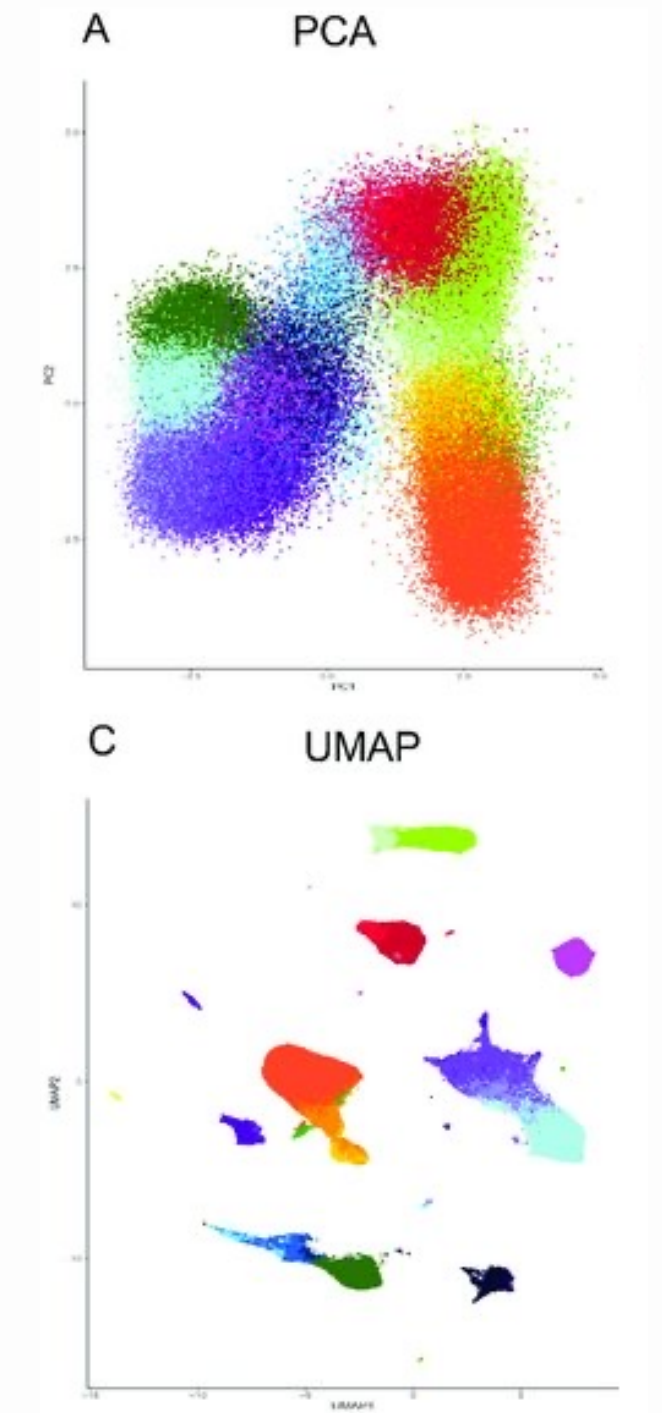
**UMAP** = Uniform Manifold Approximation and Projection

**T-SNE** = t-distributed stochastic neighbor embedding



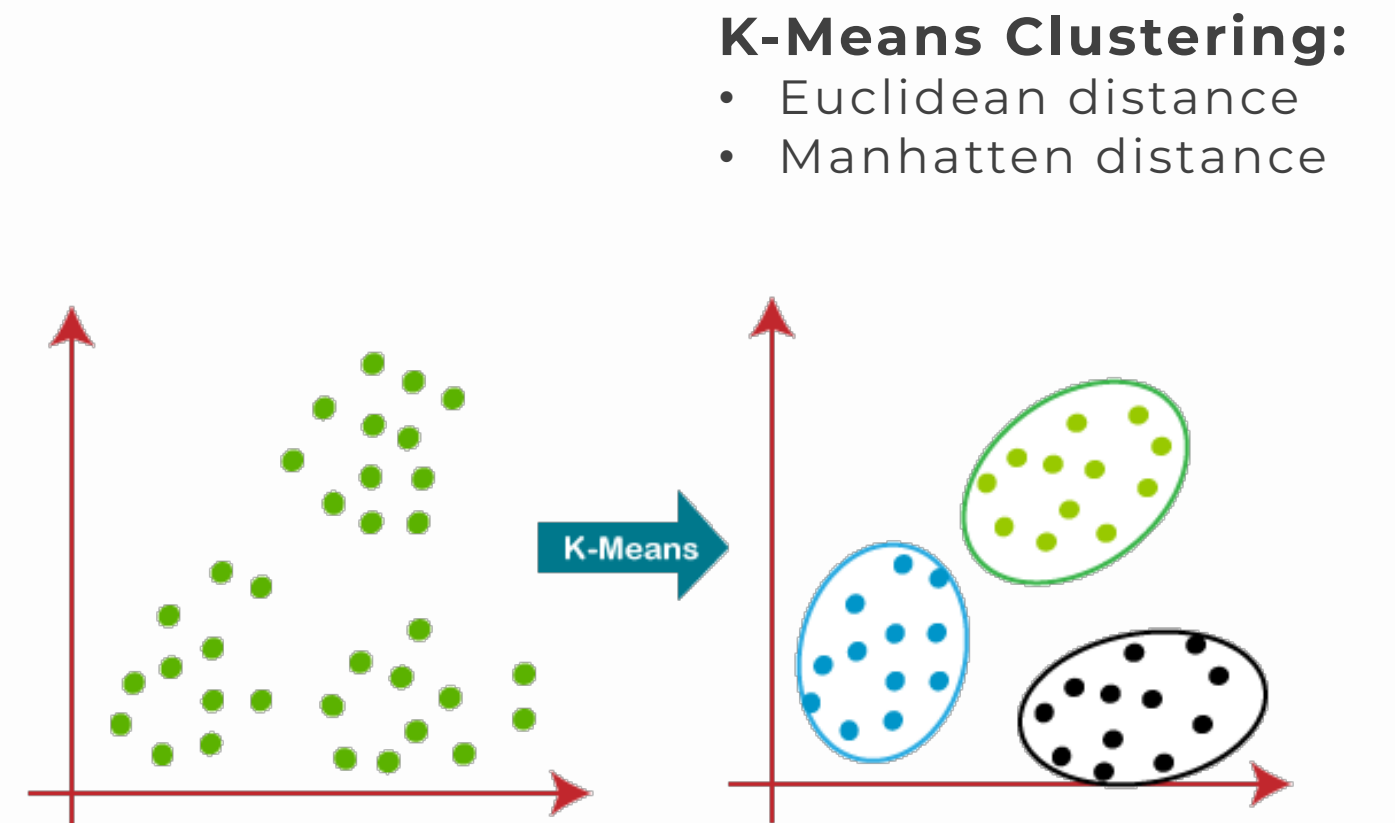
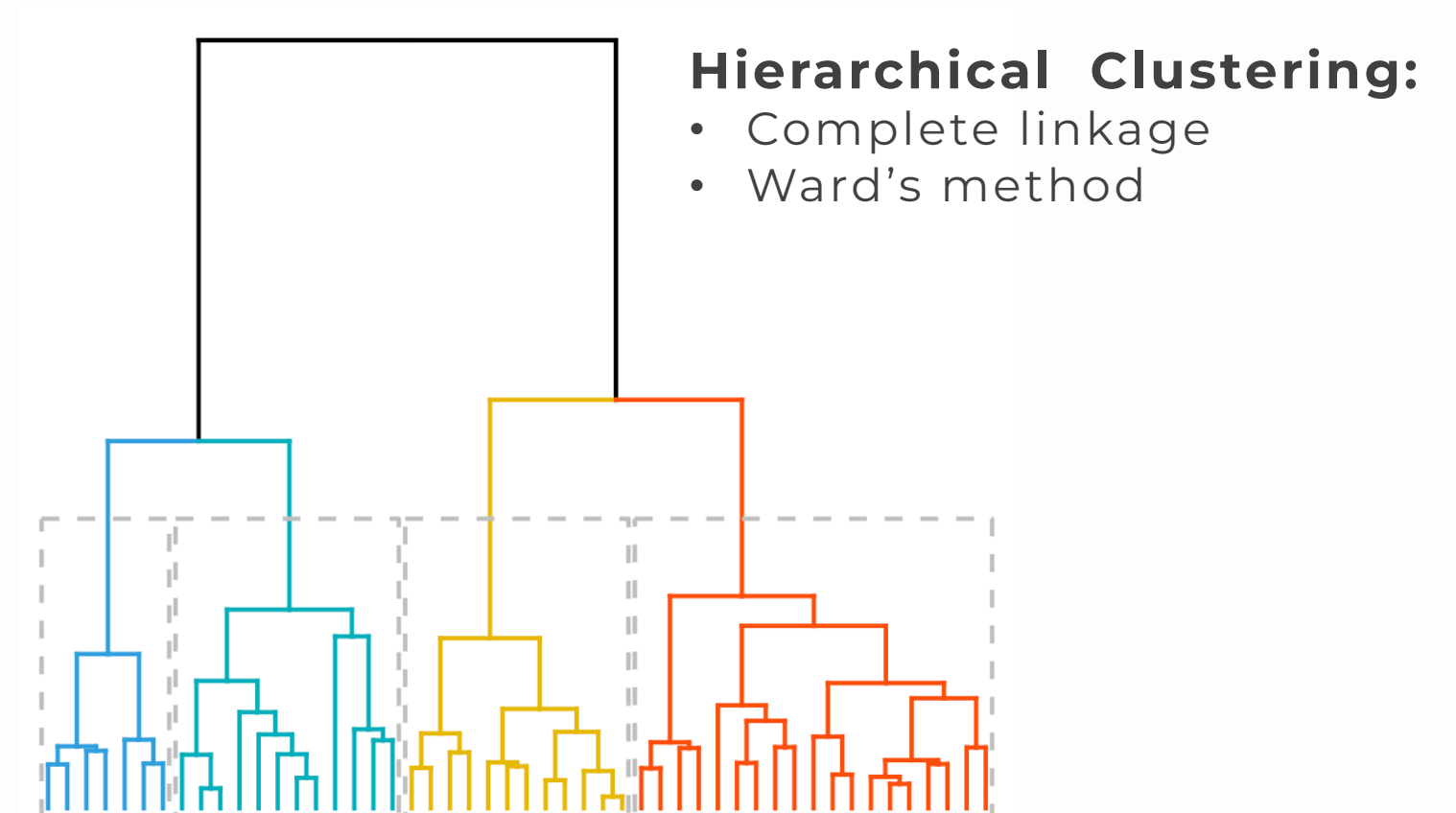
# VISUAL CLUSTERING BASED ON DR

- **DR** is not a method for clustering, **BUT** it is used in this way.
- Clustering based on DR == visual inspection of plot to identify clusters.
- Coloring observations by meta-information (i.e. cancer grade, drug treatment or cell type,...) **may** support ground truth.
- DR methods can have parameters which should be specified correctly as they will affect results.



# CLUSTERING

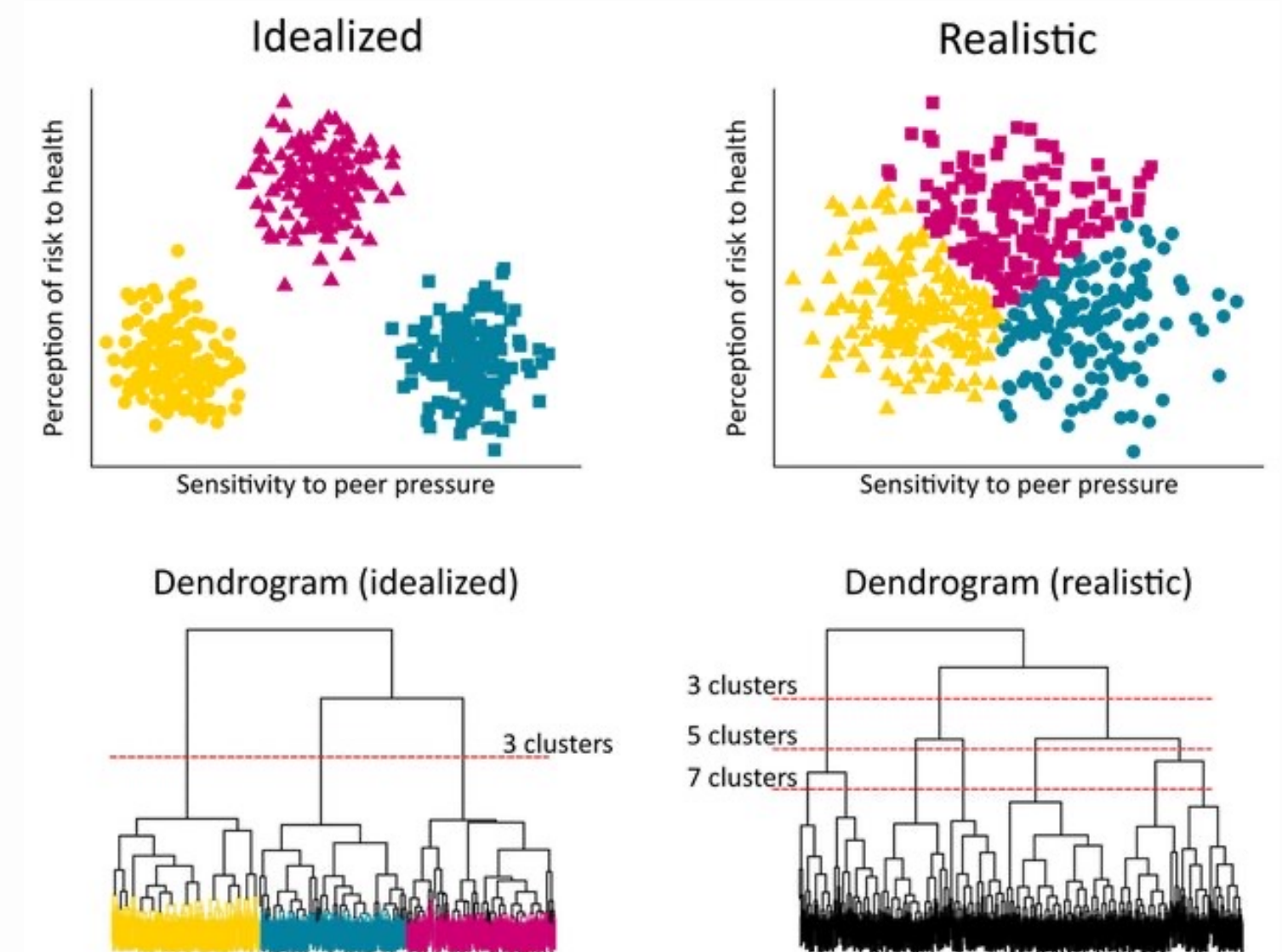
- **Clustering** == type of **unsupervised learning**.
- **Metric** to define **similarity** of observations.
- Items of a group share features **that we care about**.
- **RQ:**
  - Optimal number of clusters for this dataset?
  - What characterizes a certain cluster?
  - Which cluster does a new data point belong to?





# CHALLENGES

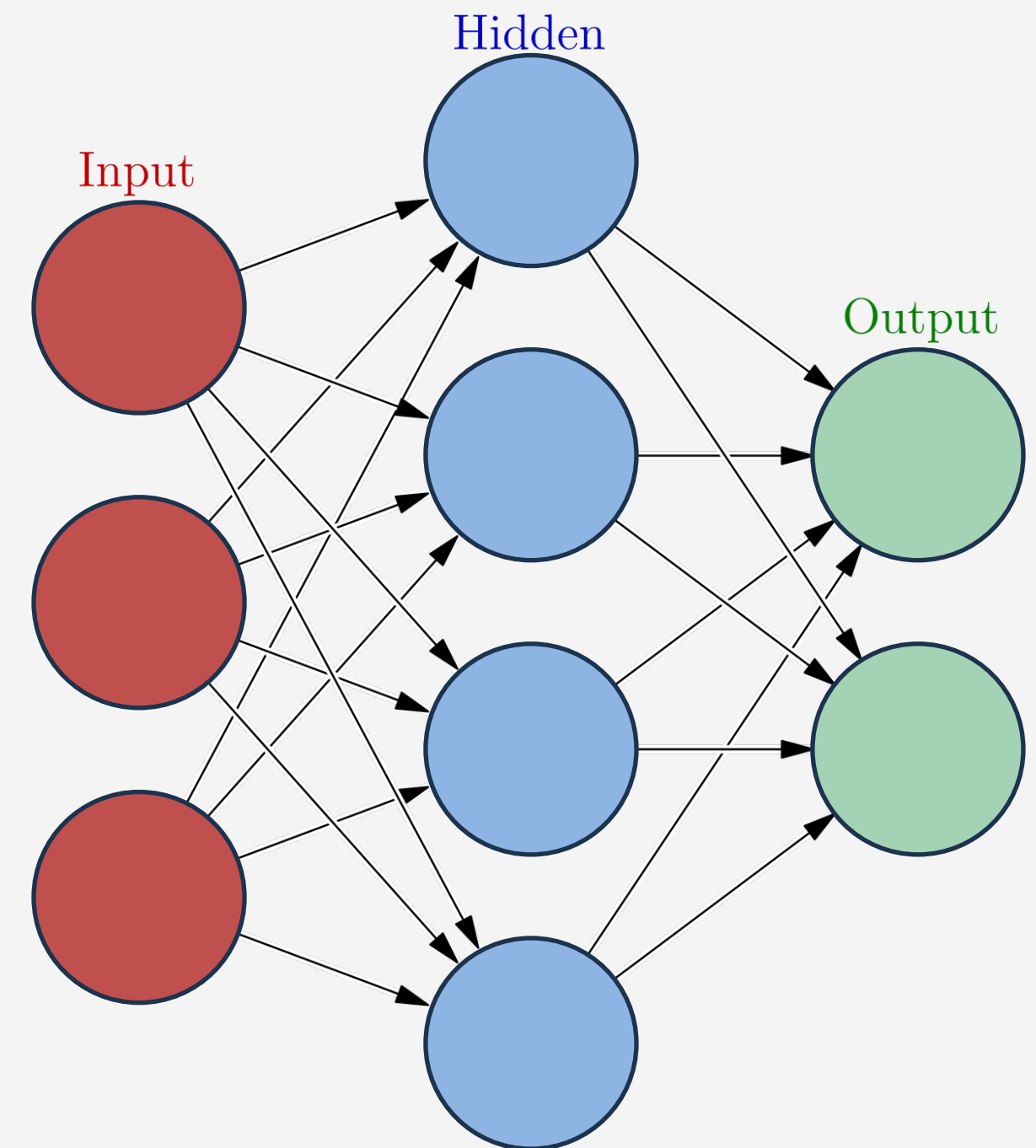
- How to define similarity?
- Human tendency to see clusters in randomness (clustering illusion)
- Algorithm may require the user to specify the number of clusters (chicken-and-egg problem)
- Non-linear clustering algorithms are powerful but can have unexpected behaviors!



Engl, Elisabeth, Peter Smittenaar, and Sema K. Sgaier. *Gates Open Research* 3 (2019).

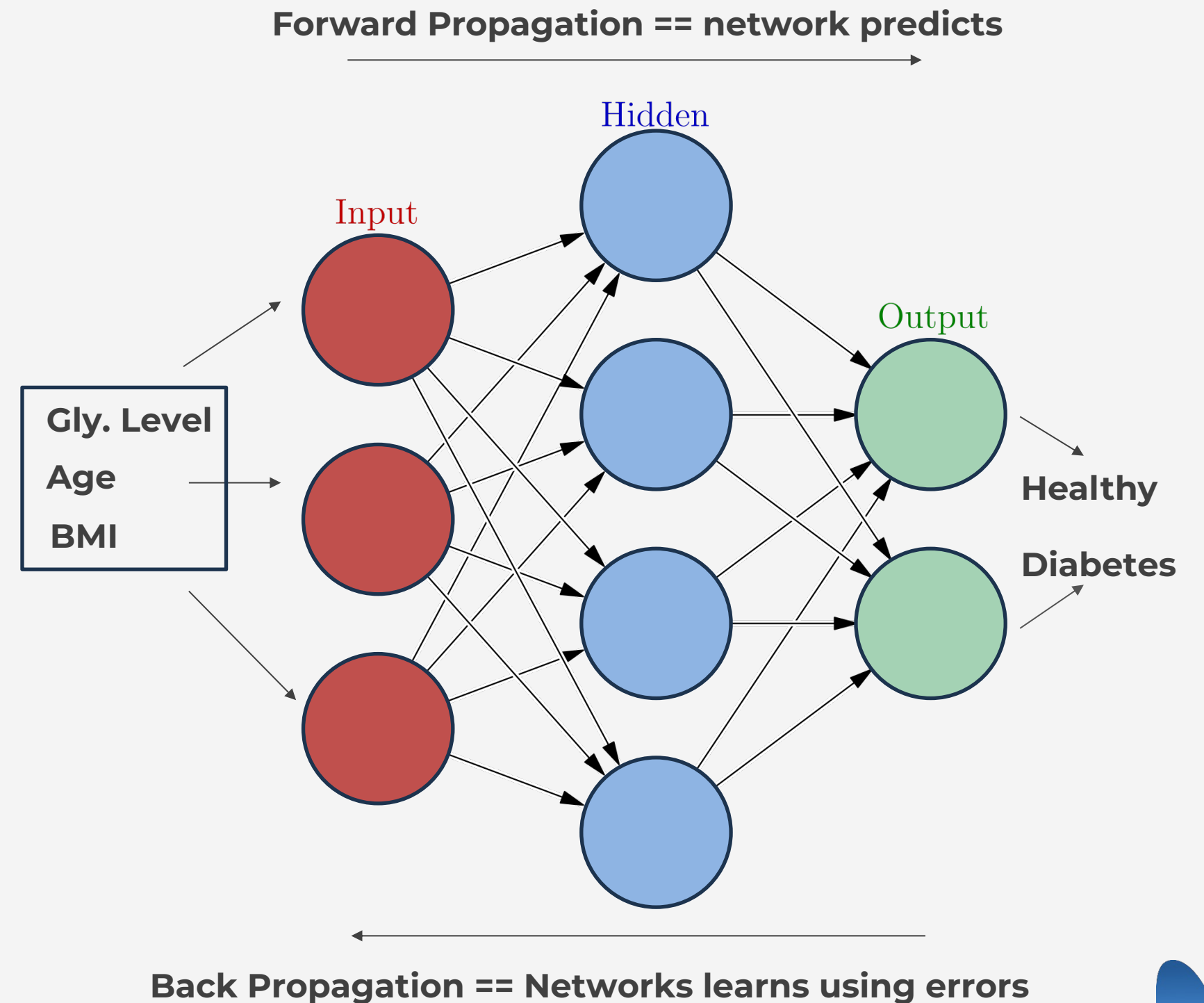
# NEURAL NETWORKS

- **Neural Networks (NN)** mimic the function and structure of the human brain
- NN can be either supervised or unsupervised
- We use NN for medical image analysis, genomics analysis, protein structure prediction, etc.
- NN can be difficult to understand in detail but not in broad strokes, let's try...



# NEURAL NETWORK FOR CLASSIFICATION

- **Input layer:** Data goes in (i.e. patient biometrics, expression data, medical images)
- **Hidden layer:** The model learns data patterns.
  - This is where the computation is performed.
  - Many hidden layers == deep neural network
- **Output layer:** Samples are classified as either cancer or healthy.



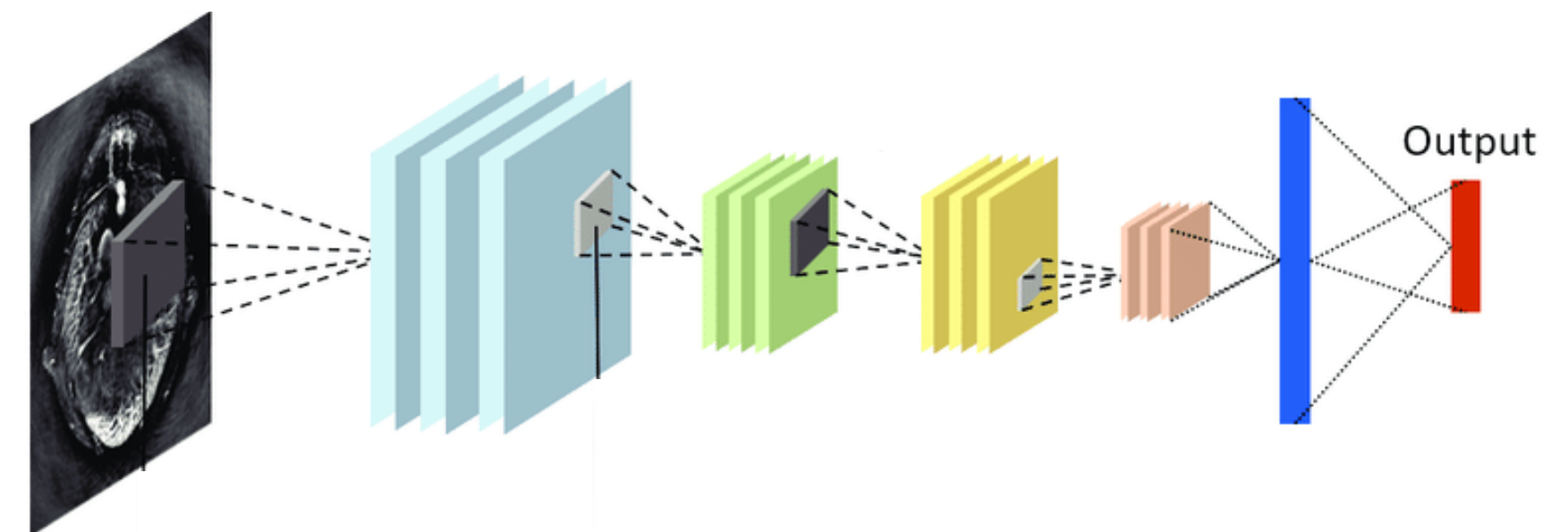
# MEDICAL IMAGE ANALYSIS



- There are different architectures of neural networks (NNs).
- Different neural networks are good for different data and tasks.
- NNs for **medical image** analysis are often *convolutional neural networks (CNNs)*.
- Transformer models are extremely powerful.

- **Many** Data Scientists do **NOT** implement their own NNs, but they use a trained one.
- Can use different programming languages to implement NNs in.

## CONVOLUTIONAL NEURAL NETWORK



Yang, Changchun, et al. "Review of deep learning for photoacoustic imaging." *Photoacoustics* 21 (2021): 100215.



# ChatGPT

- You may already have used or heard about **ChatGPT**
- The algorithm behind is a *Transformer Neural Network*
- This AI bot can do many things, also programming!

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

**You**  
How can I do k-means in python?

**ChatGPT**  
To perform k-means clustering in Python, you can use popular libraries like Scikit-Learn or SciPy. I'll provide you with an example using Scikit-Learn, which is one of the most widely used libraries for machine learning in Python.

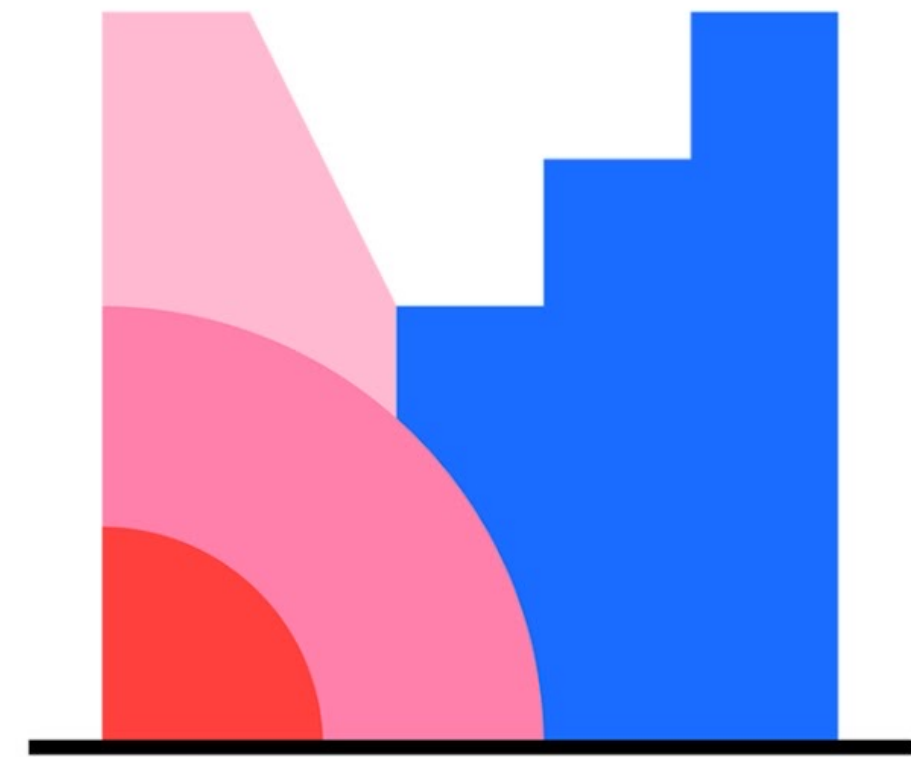
Here's a step-by-step guide to doing k-means clustering in Python using Scikit-Learn:

1. Import the necessary libraries:

```
python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

1. Create or load your dataset. Make sure it's in a suitable format, such as a NumPy array or a Pandas DataFrame.

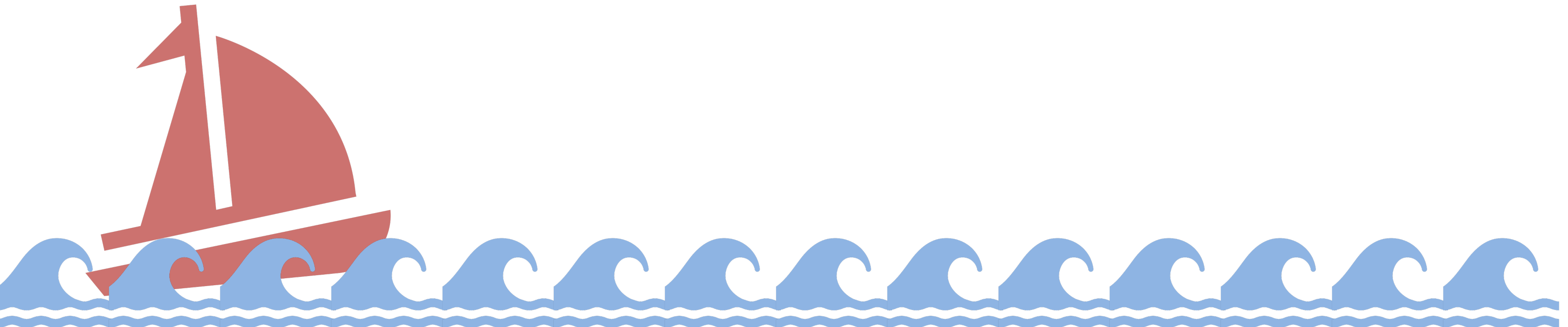




**Mentimeter**

**Q9-Q11:**

A couple of questions about data analysis and models.

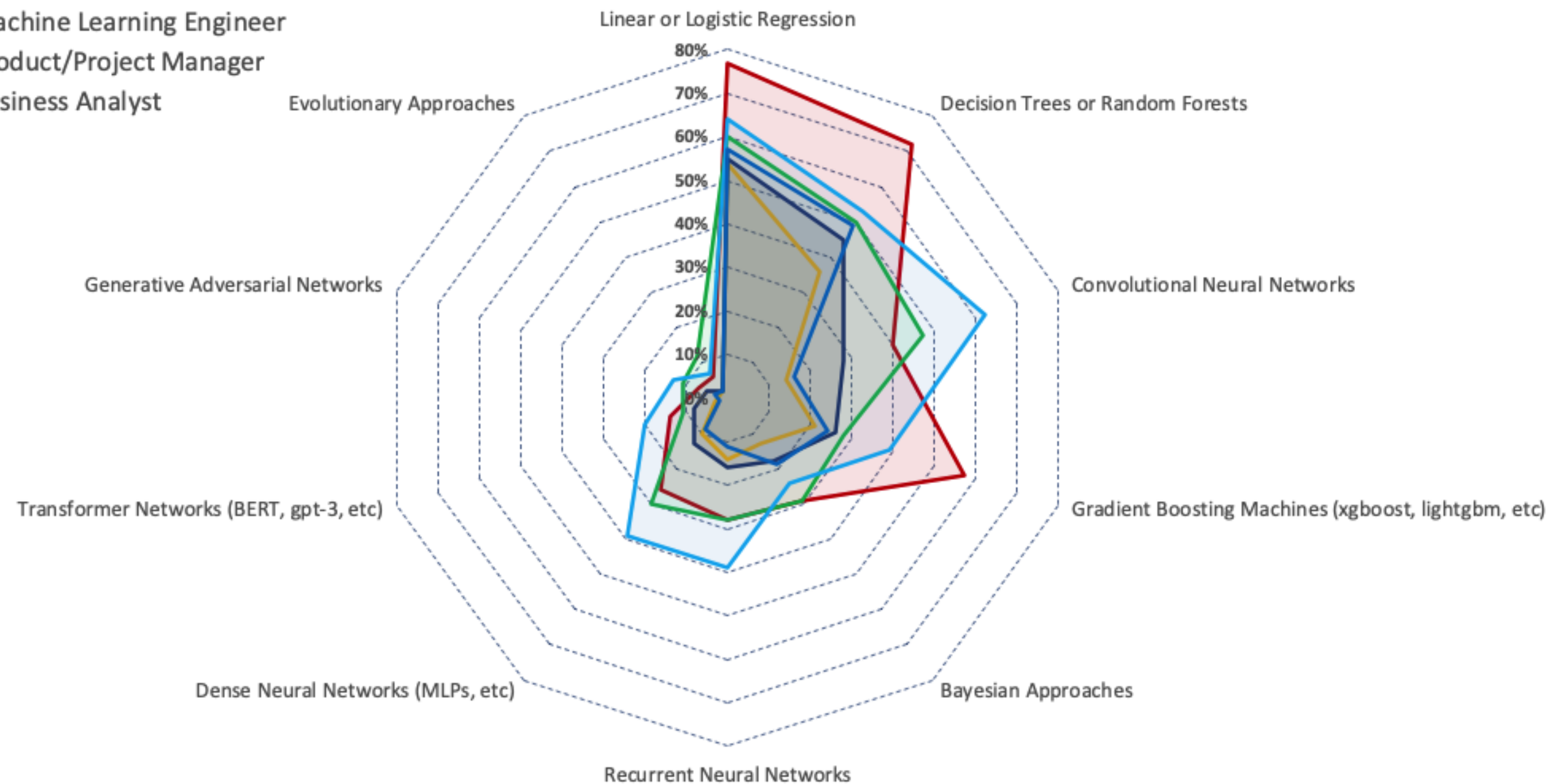




# Usage of ML Algorithms across Different Data Professional Roles

## Job Title

- Data Scientist
- Research Scientist
- DBA/Database Engineer
- Machine Learning Engineer
- Product/Project Manager
- Business Analyst



# WHICH APPROACH SHOULD WE USE?

## STATISTICAL ANALYSIS

- 'Simple' question
- Linear relationships
- Small dataset (few obs.)

The scientific questions asked are central to the methods chosen.

The **dataset size** guides choice of algorithm.

## MACHINE LEARNING

- High dimensional data
- Non-linear relationships

### SUPERVISED

Outcome is known

### UNSUPERVISED

Outcome is unknown

Fully data driven

## GROUP DISCUSSION 3.2

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

**Is there a difference in weight between mice of strain A and strain B?**

Further discuss:

**What types of variables (data types) do we have? Do you know of any test you could use to answer this question?**



## GROUP DISCUSSION 3.3

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

*You have protein abundance data from skin samples (~ 10.000 different protein species). These samples were collected from patients with psoriasis (normal adjacent -and affected skin) and from healthy controls.*

**Is protein abundance predictive of the skin phenotype? And if so, are the levels of all proteins equally predictive/informative?**



## GROUP DISCUSSION 3.4

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

**Is the amount of bacterial load in a swap of the oral epithelium (gums) based on skin type, diet and whether the person recently had antibiotic treatment?**

Further discuss:

**What are the outcome variable(s) and the explanatory variable(s) in this scenario?**



## GROUP DISCUSSION 3.5

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

*Gene expression data and biometrics (height, weight, age, etc.) from patients with colorectal cancer. You are interested in exploring if there are any potential subgroups of cancer patients within your dataset, in order to pair each subgroup with the appropriate healthy controls.*

**What type of analysis could you use for this?** *N.B while avoiding a fishing-expedition*





# BREAK

