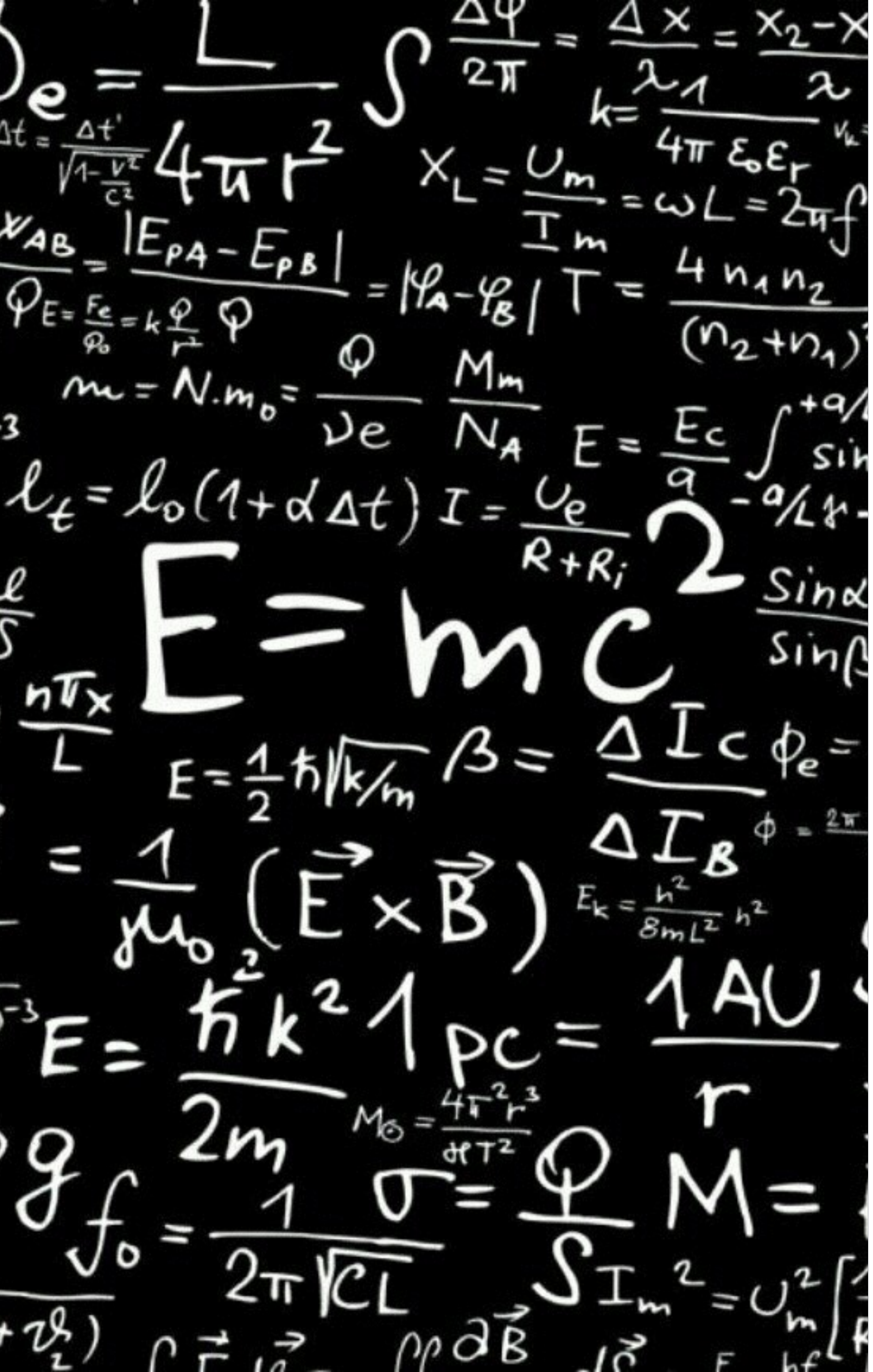
Day 3-4

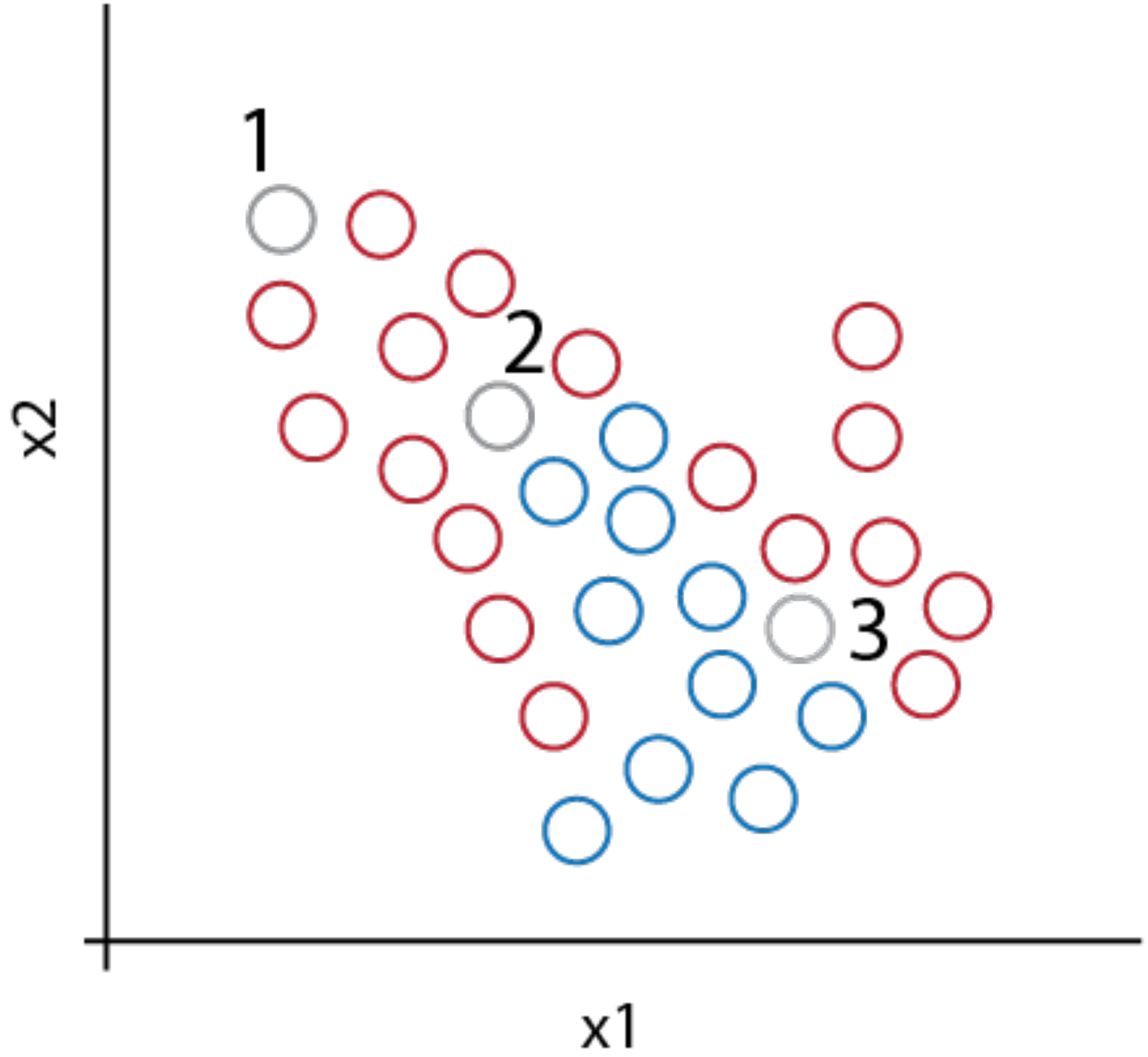
- Day 3 Morning : Nearest-Neighbor Methods, Feature Selection
- Day 3 Afternoon : Recommender System, Unsupervised Learning
- Day 4 Morning : Neural Network
- Day 4 Afternoon : Advanced Concepts in Machine Learning



INSTANCE-BASED LEARNING

Intuition Quiz

Would you classify point 1, 2, 3 as blue or red. Fill in the table.



Pt	BLUE or RED
1	
2	
3	

IBL: How Decision is Made

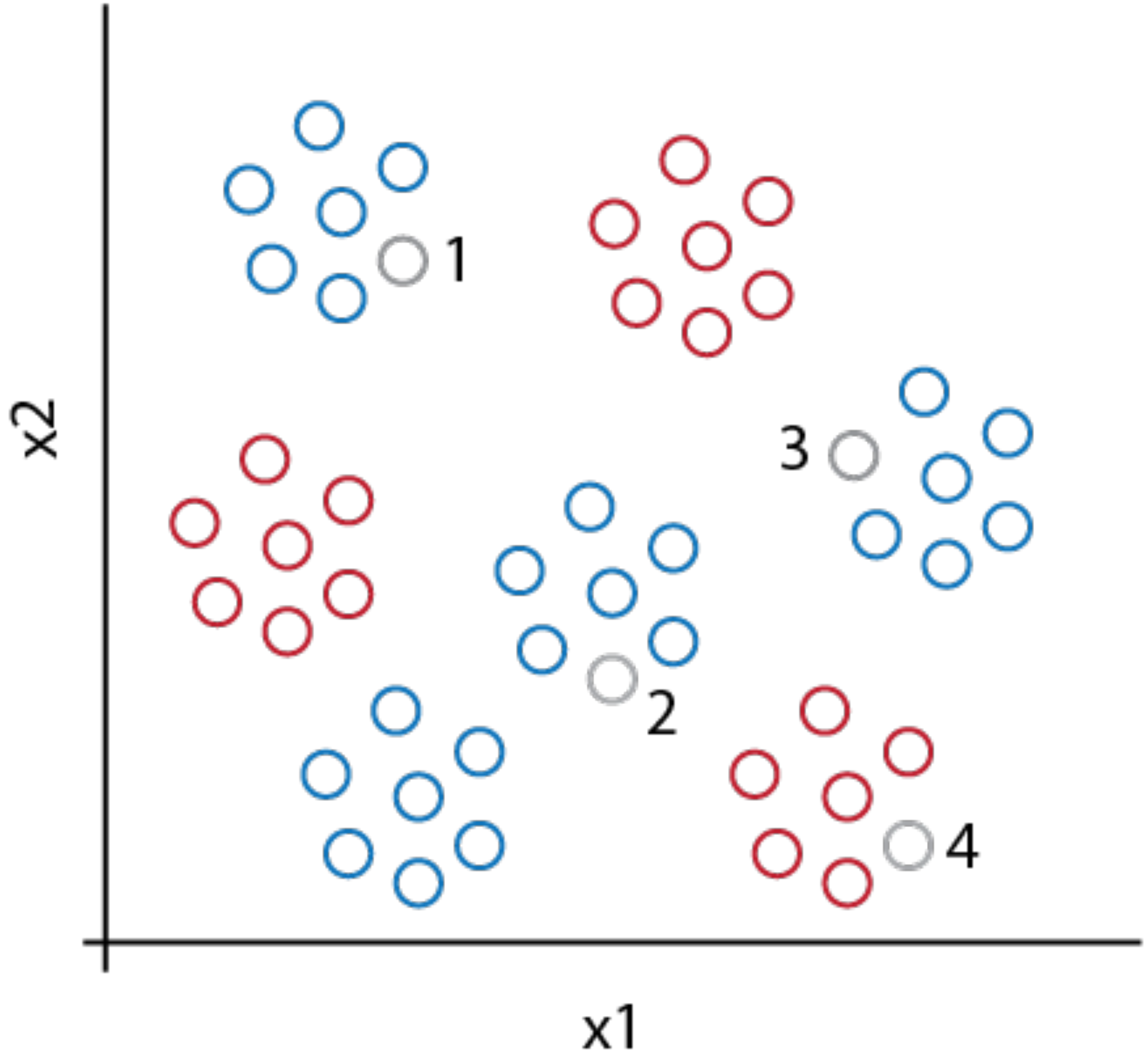
- Your source of knowledge is the similarity between two different data points. So you use similarity to make decisions such as classification and regression.
- You make decisions about one data point based on neighboring points.

Instance-based Learning

- Lazy algorithm: when you see your training set, you do nothing, just store them in the memory.
- When new sample comes you compare the new sample with the existing samples in the memory.
- Examples of algorithms in this family: nearest neighbor, kernel machines.

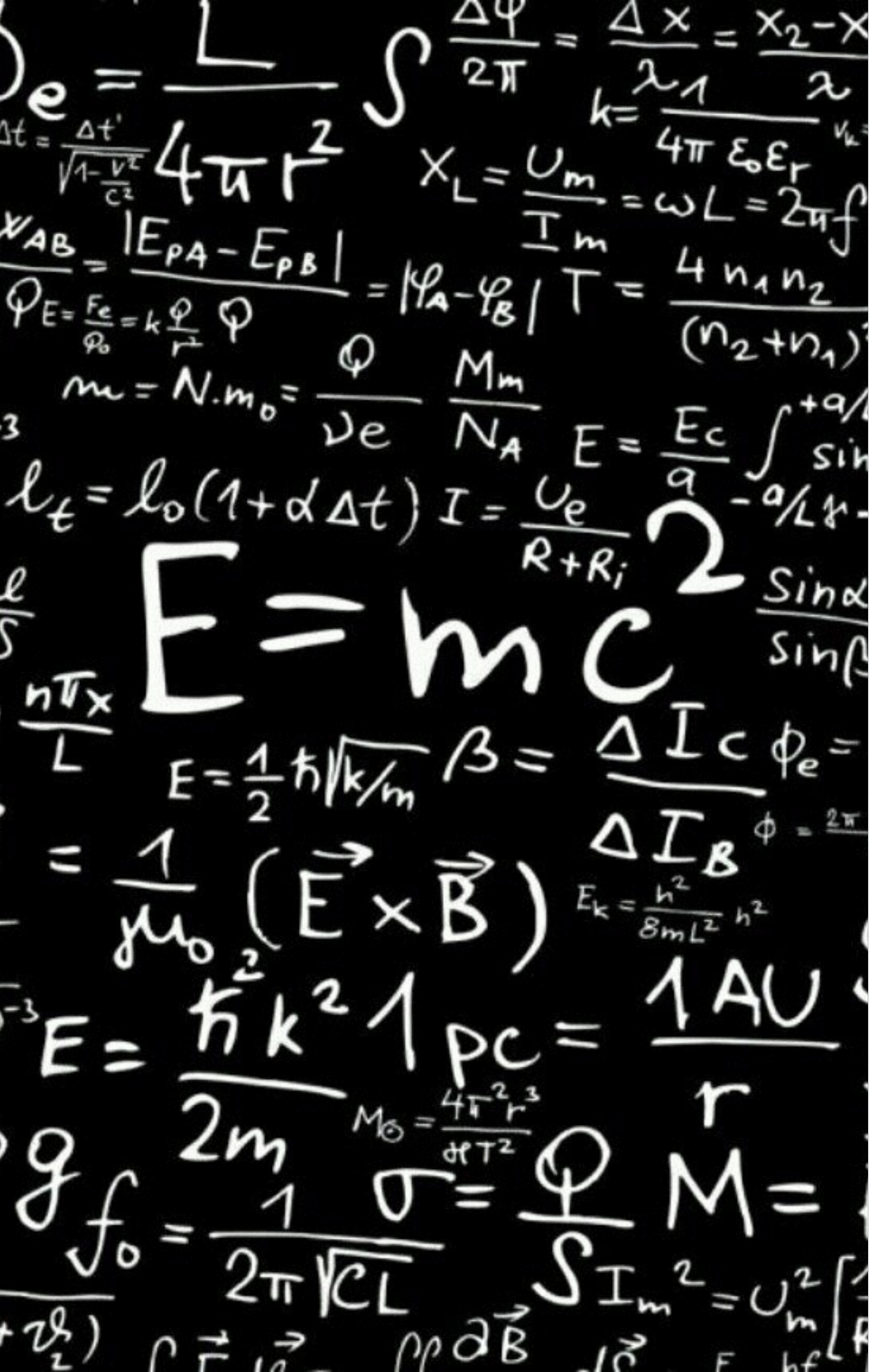
IBL - Nearest Neighbor Methods

- **Nearest neighbor:** when you see a new data point (x′), locate the nearest data point (x) and predict the label of x′ to be the same as label of x.



IBL - K-Nearest Neighbor Methods

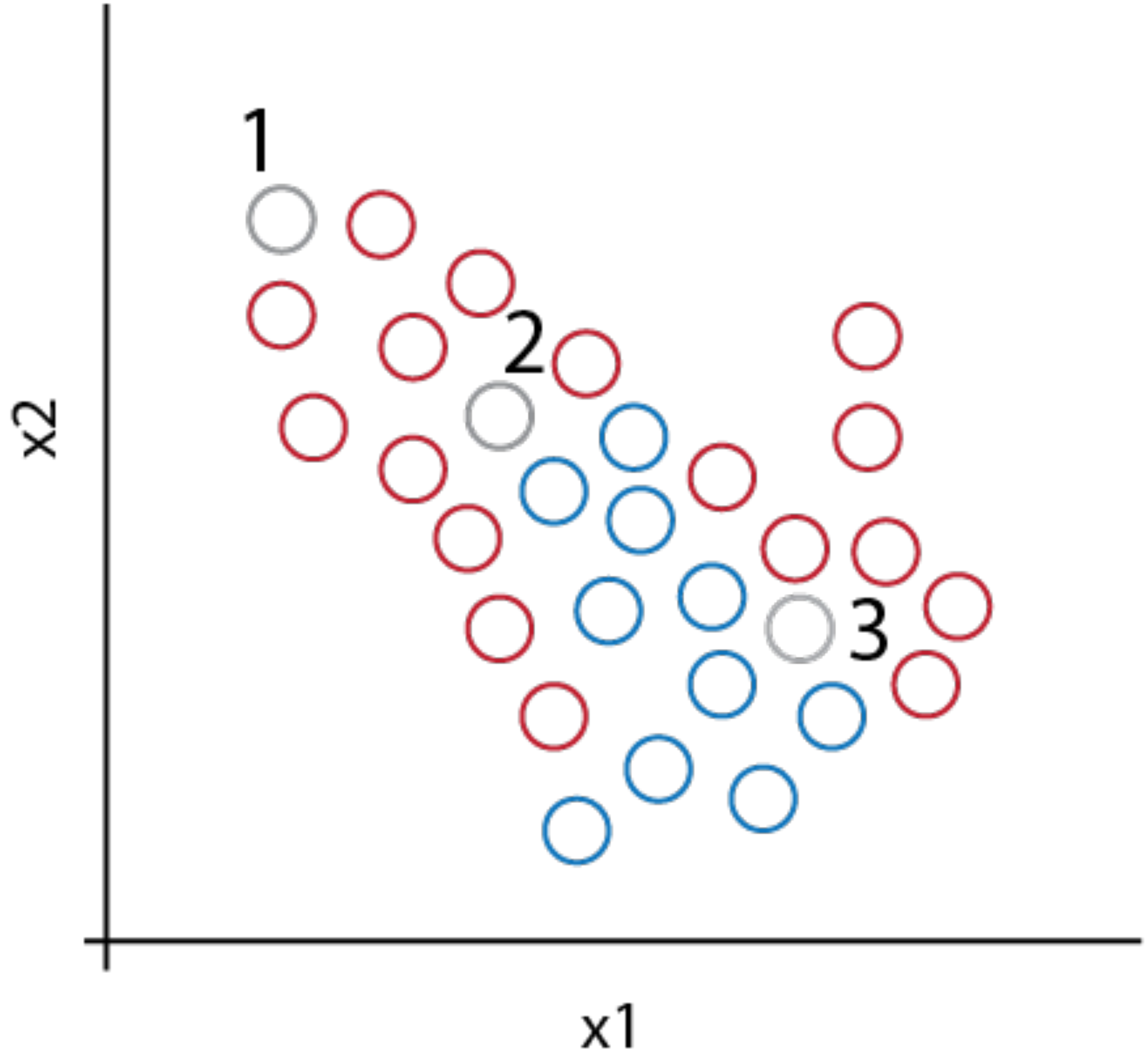
- **K-Nearest neighbor:** locate k nearest neighbors around x′.
  - For classification problem, let k neighbors vote for the right label of x′.
  - For regression problem, average the y values of all neighbors and predict that y as the label of x′.



K-NEAREST NEIGHBOR QUIZ

IBL - K-Nearest Neighbor Quiz

Use K-Nearest Neighbor Rule to classify point 1, 2, and 3 with different values of k.



Pt	k=1	k=2	k=3	k=4
1				
2				
3				

Pros and Cons

- Pros:
  - Training takes no time
  - Complex decision boundary is possible
  - Information is not lost
- Cons:
  - Query is slow (the more data the slower)
  - Storage space is huge
  - Easily fooled by irrelevant attributes

Distance and Similarity Metrics

- To determine whether two points are close, we use distance metrics.
- **Distance metrics** are the numerical value that tells you whether two points are close (low value) or far apart (high value).
- There are several ways to define distance metrics, such as euclidean distance, minkowski distance.
- **Similarity metrics** are the numerical value that tells you whether two points are close (very similar - high value) or far apart (very dissimilar - low value).
- Distance and similarity metrics are important in many ML models such as 'Support Vector Machine', 'K-Nearest Neighbor', 'K-Mean Clustering'

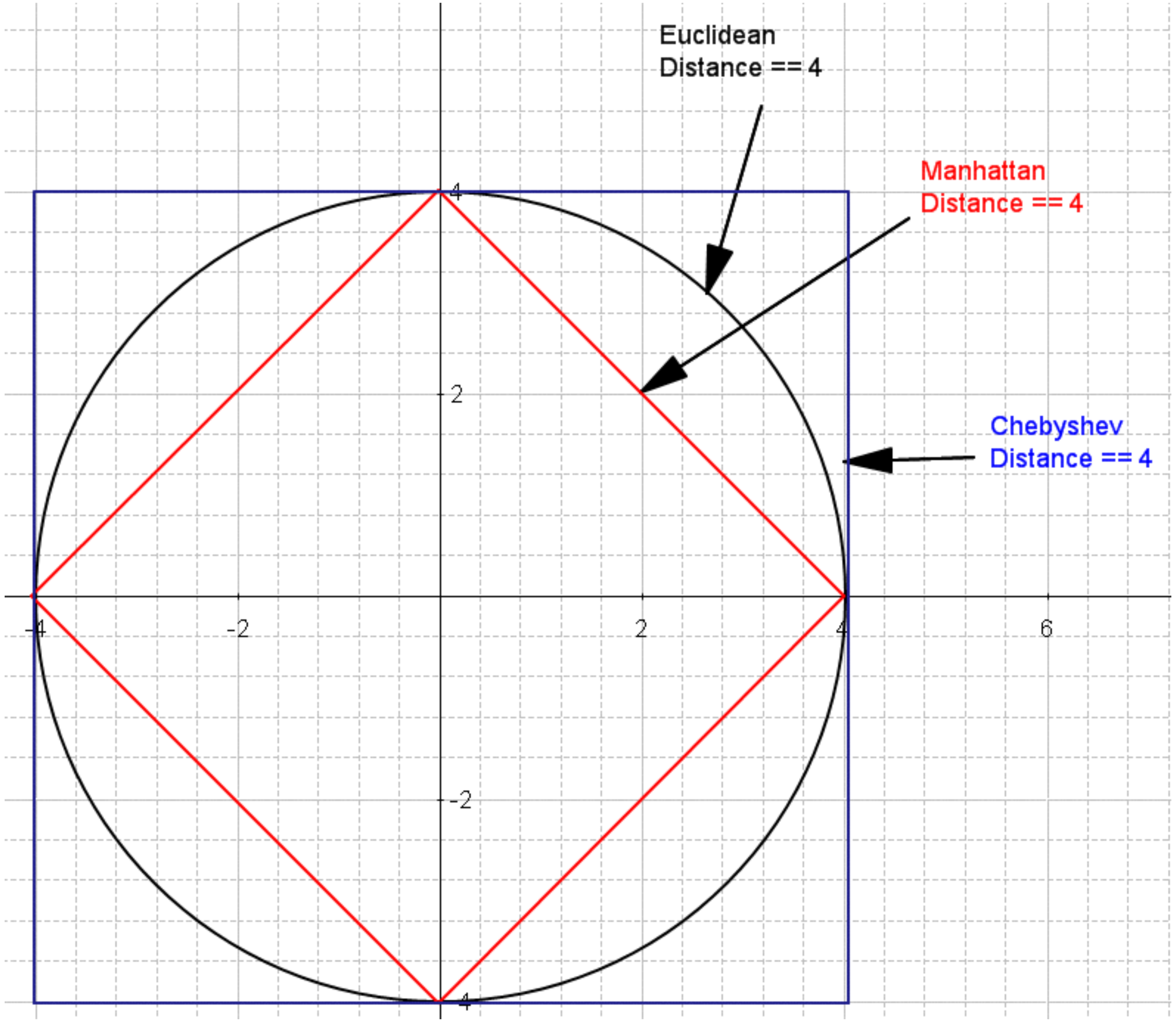
Distance Metrics for Real Value Features

- Euclidean Distance

  sqrt(sum((x - x′)^2))

- Manhattan Distance

  sum(|x - x′|)



Distance Metrics for Boolean Features

- Jaccard Distance

Feature	Me	My Dad
Man Barber	F	T
Toyota	T	T
MK	T	T
Water Park	T	F
Temple	F	T
Bar	F	F

N=6
NTT=2
NTF=1
NFT=2
NFF=1
NNEQ : number of non-equal dimensions

NNEQ = NTF + NFT = 3

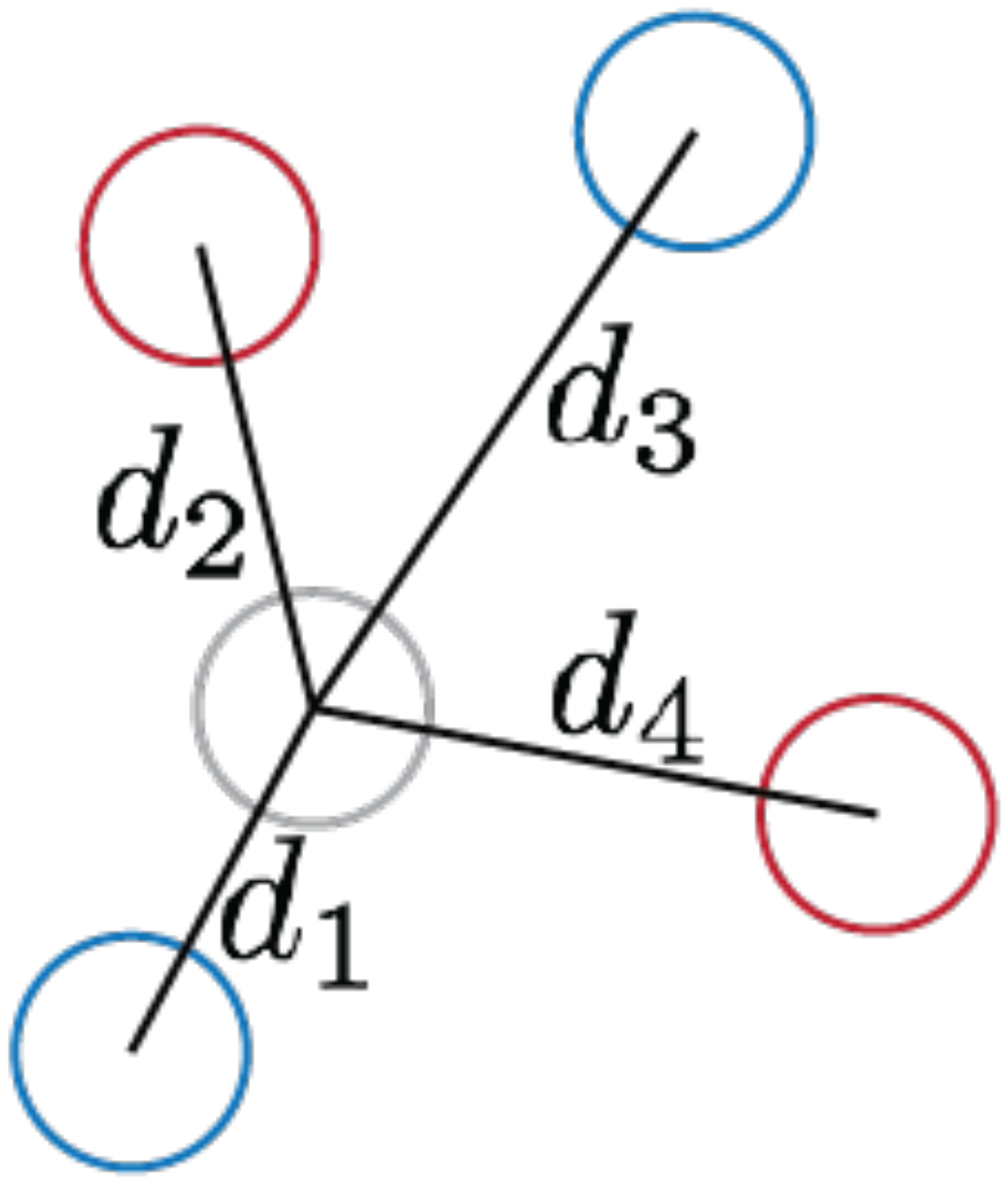
NNZ : number of nonzero dimensions

NNZ = NTF + NFT + NTT = 5

NNEQ / NNZ = 3/5 = 0.6

Using Distances as Weights

- **Neighbors who are closer to the target data point should get more say in the voting process.**



$$y' = \frac{w_1 y_1 + w_2 y_2 + w_3 y_3 + w_4 y_4}{w_1 + w_2 + w_3 + w_4}$$

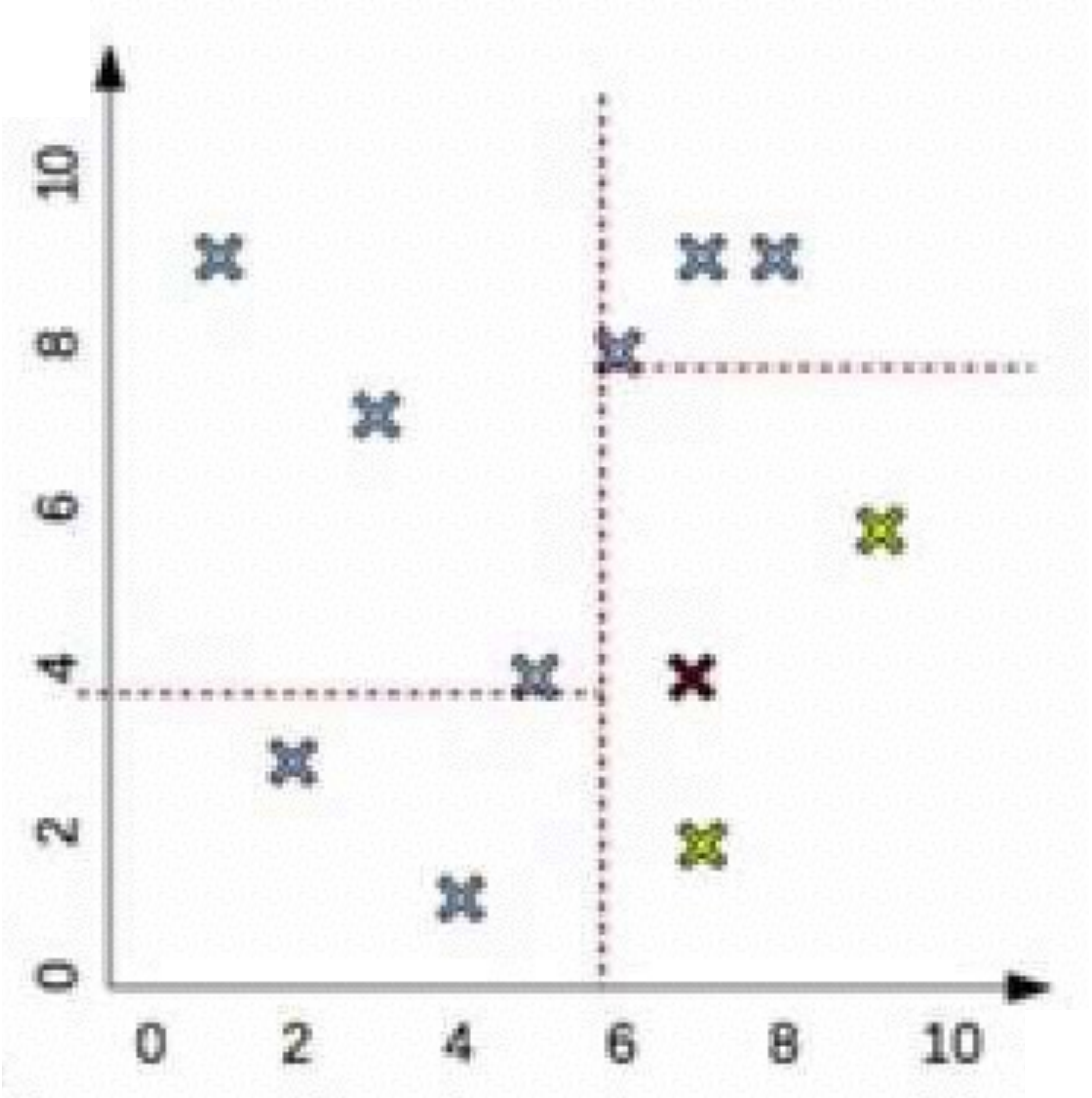
$$w_i = \frac{1}{d_i}$$

Searching for Nearest Neighbors

- **Brute force : when a new sample x′ appears, calculate the distance between x′ and all other points. Consider points with lowest distances for voting.**
- **Brute force is slowest, but the most accurate.**
- **If your data is sparse, then brute force is the right way.**
- **To speed up the search, you can use KD Tree or Ball Tree.**

K-D Tree

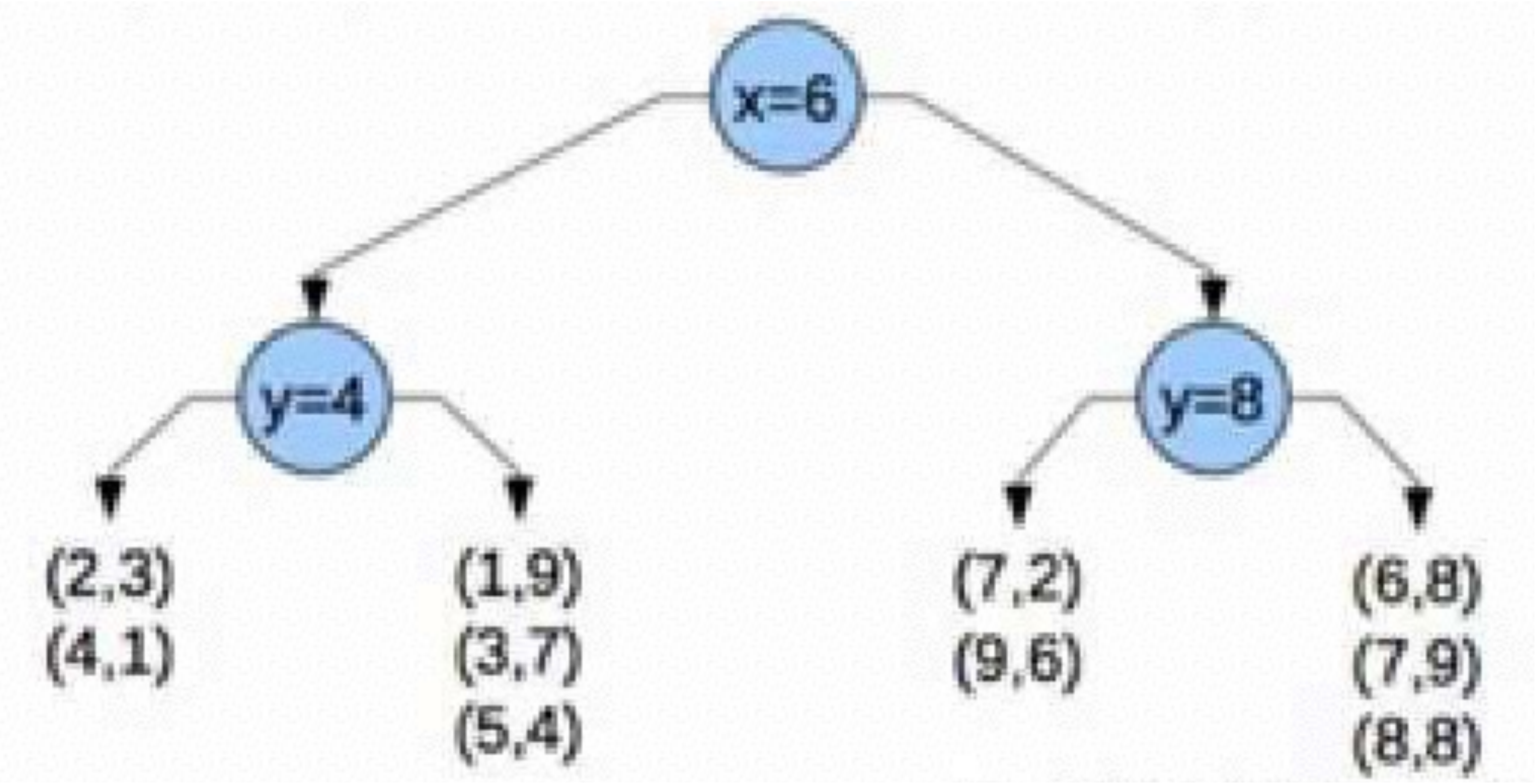
- Data = [(1,9), (2,3), (4,1), (3,7), (5,4), (6,8), (7,2), (8,8), (7,9), (9,6)]
- Say we want to search for nearest neighbors of point (7,4)

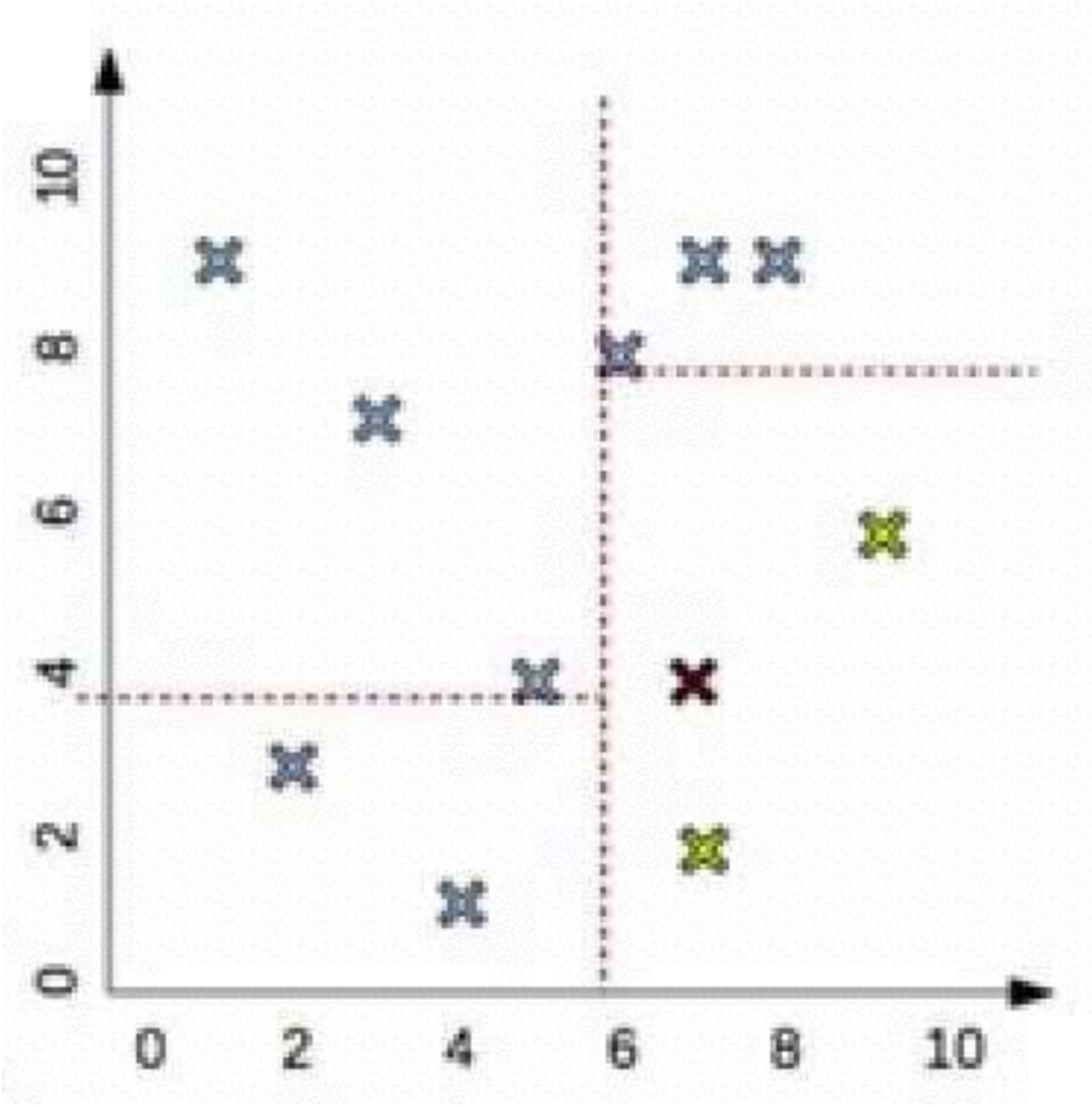


Credit: Victor Lavrenko

K-D Tree

- First, pick a random dimension (say x1) find median and split data. Repeat for other dimensions.

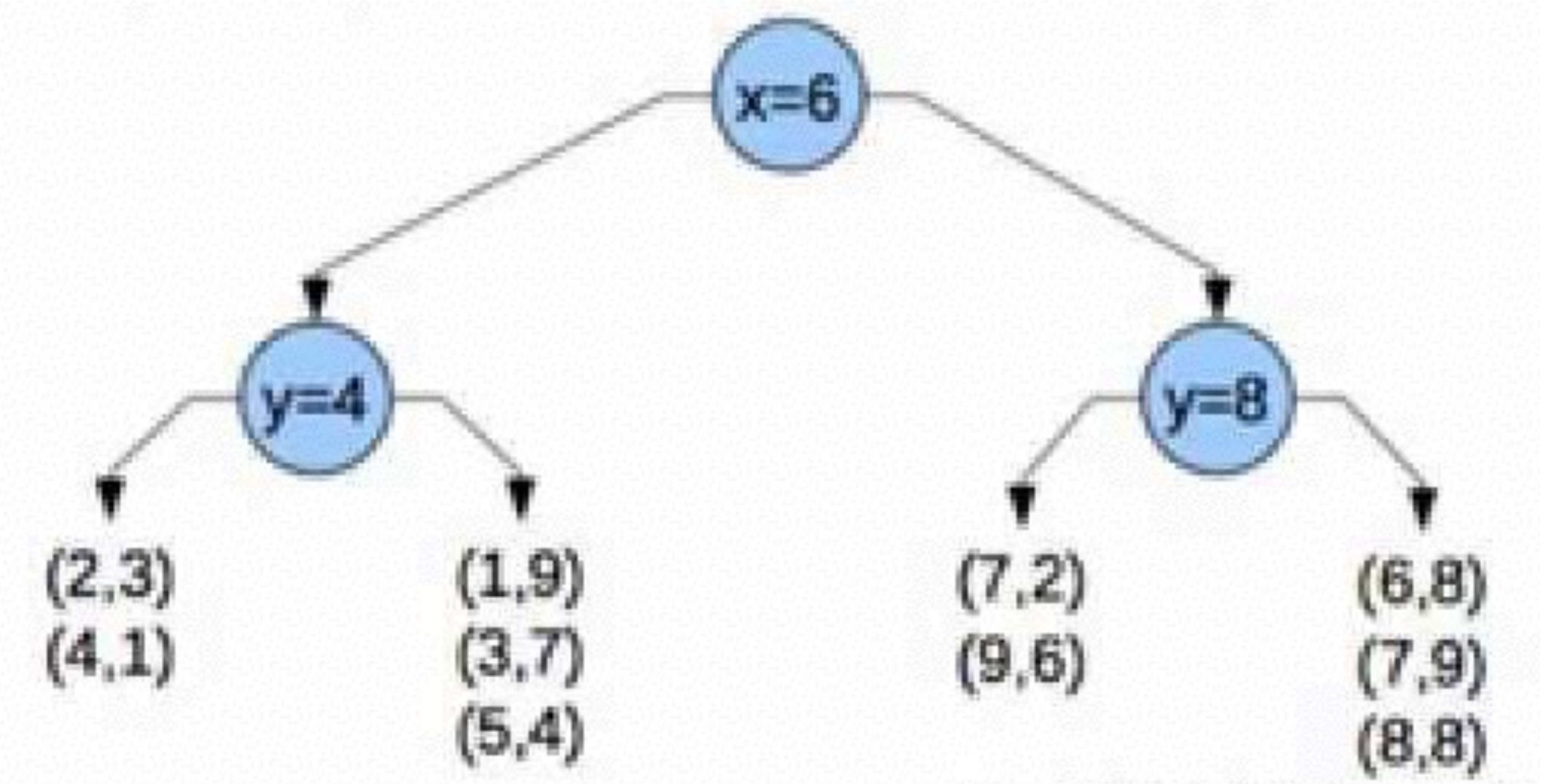


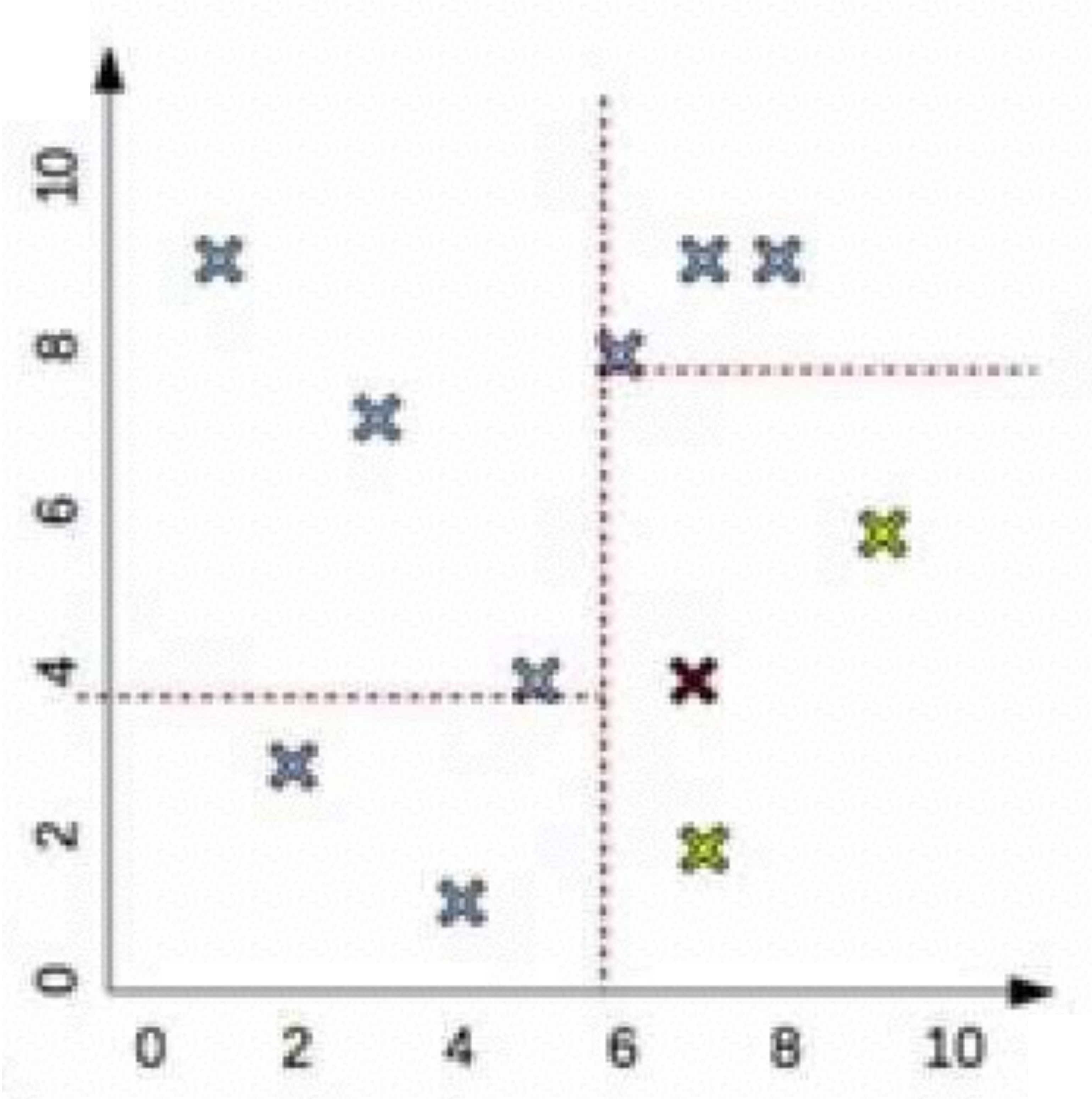


Credit: Victor Lavrenko

K-D Tree

- **Find region that contains (7,4) search for neighbors only in that region.**

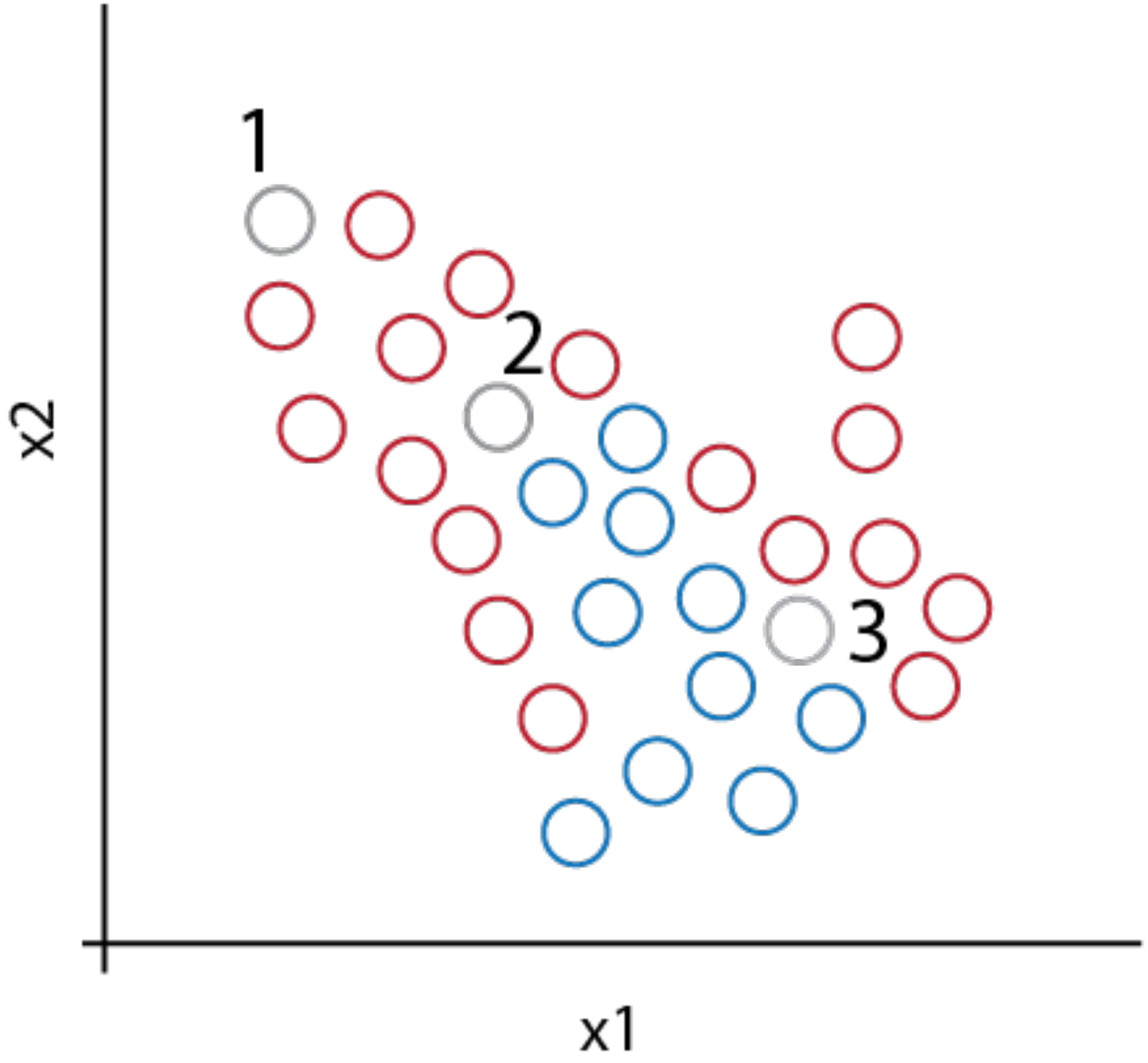




Credit: Victor Lavrenko

How to Avoid Overfitting

- Use k as an overfitting control.
  - If k is one, you are very susceptible to noise (overfitting).
  - If k is high, you are averaging over really large regions, you lose resolution (under fitting).



How to Avoid Overfitting

- Remove noisy instances prior to using nearest neighbor algorithm. Remove x if all nearest neighbors of x are in the opposite class.
- Form prototypes. If you observe lots and lots of very similar samples, lump them into a prototype by finding an average over all dimensions.

Text Classifier with KNN

K-NEAREST NEIGHBOR CODING LAB

- Why Feature Selection
- How to do feature selection

FEATURE SELECTION

Big Data

- Most problems you will face in the real world is gigantic.
    - Millions of rows
    - Hundreds or thousands of features
    - Your algorithm will take forever to run
- What can we do about it?
    - We might be able to look through all the features and manually select them.
    - But that would waste so much time and resources
    - So maybe do automated feature selection?

Feature Selection :
The process of selecting
the most relevant features
to be included in
the machine learning model

What Feature Matters Most?

- The algorithm predicts whether the email is spam or not. Which feature is most useful for the prediction?
  - Feature 1: whether the email contains the word ‘viagra’
  - Feature 2: whether the email is sent from a Nigerian Prince
  - Feature 3: whether the email is sent from one person to a massive amount of people
- For all 1000 features you have calculated, maybe only a few features are important.
- Feature selection algorithm gives you insight and interpretability of your model.

Curse of Dimensionality

- This is one of the most important problems in machine learning.
  - Imagine you have a large amount of features, each can have infinite number of values.
  - You will need an enormous amount of training data is required to ensure that there are several samples with each combination of values.
  - If you have limited samples, which do not cover the whole space, your model loses predictive power.

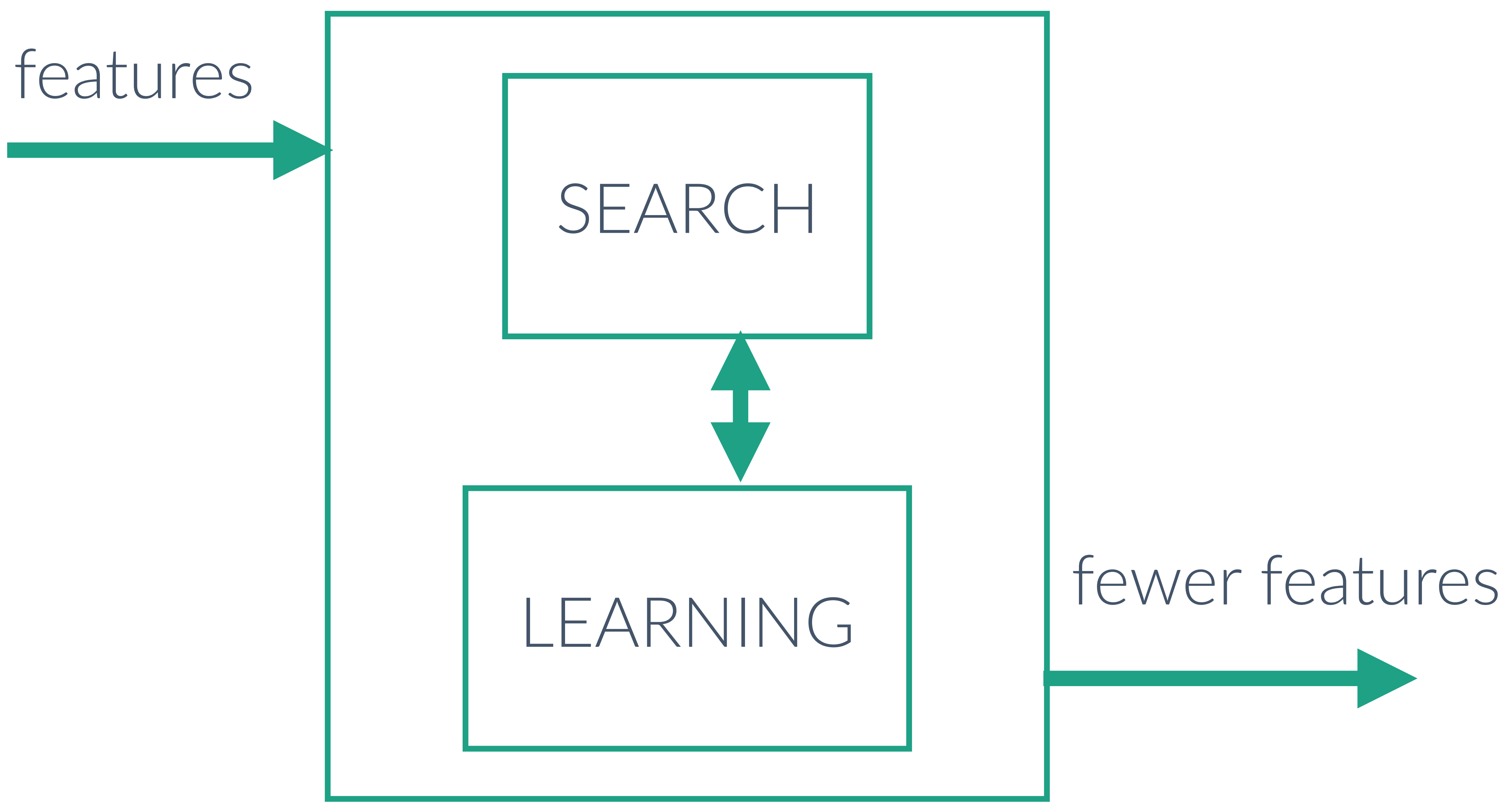
Curse of Dimensionality

- This is one of the most important problems in machine learning.
    - Take linear regression for example
    - The more features you have, the more parameters you need to fit the model
    - If you have 2 features, your solution space has 2 dimensions (small possible values)
    - If you have 1000 features, your solution space has 1000 dimensions (huge amount of possible values.
    - You algorithm can take a lot of time to find solution.

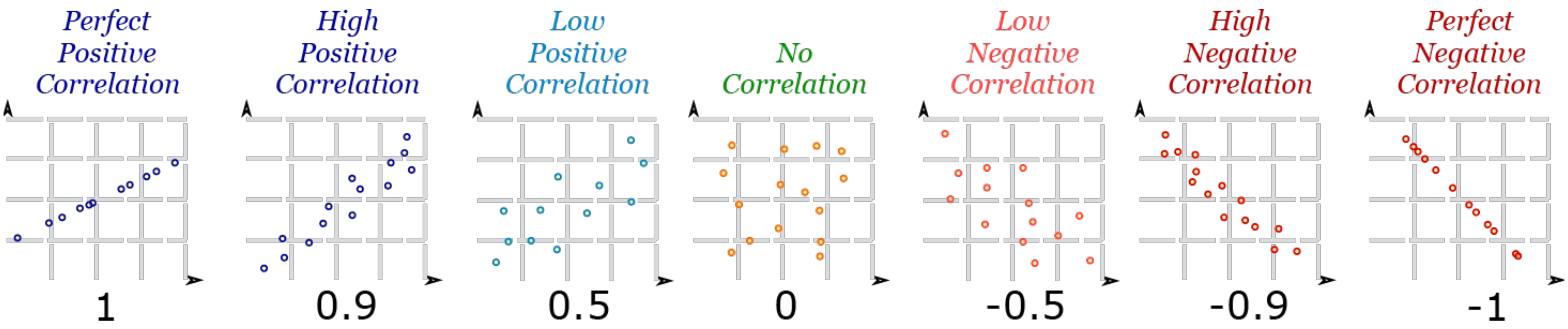
Feature Selection

Filtering Methods

Wrapping Methods



How to find a good feature



- You need to find features that have high correlation to your target.
  - Don't care whether it's a positive or negative correlation
  - The larger the number the better

How to find a good feature

- Correlation is not the only measure that tells you 'how much x is related to y' there are other measures we use.
- Such as:
  - ANOVA (Analysis of Variance)
  - Chi2
  - Mutual Information

Analysis of Variance

- Classification problem: Y can only be class 0 or class 1. Find variance of X within class and between classes.

$$\text{F-Value} = \frac{\text{Variance between classes}}{\text{Variance within class}}$$

V between class is **high**
V within class is **low**
F-Value is **high**
Feature X is **important**

V between class is **low**
V within class is **high**
F-Value is **low**
Feature X is **not important**

Analysis of Variance Quiz

Weight	Class
50	Adult
80	Adult
12	Children
30	Children
…	…

V between class is …
V within class is …
F-Value is …
Feature is …

Weight	Class
68	Thailand
75	China
80	Thailand
82	China
…	…

V between class is …
V within class is …
F-Value is …
Feature is …

Weight	Class
65	Human
1000	Animal
0.1	Animal
60	Animal
30	Human

V between class is …
V within class is …
F-Value is …
Feature is …

Sklearn Feature Selection

- f_classif : calculate analysis of variance between any x and y variable

  http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

- f_regress : calculate correlation between any x and y variable

  http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html