Welcome to CME 250 Introduction to Machine Learning!

Spring 2020 – Online version

April 21th 2020

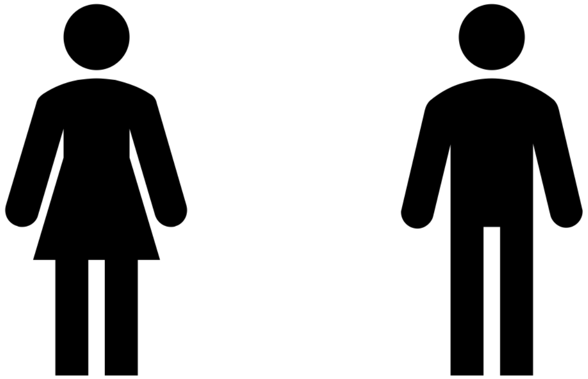
Stanford | Institute for Computational & Mathematical Engineering

Today's schedule

- Practical example of data exploration
- Intro to Supervised Learning
  - K-Nearest Neighbors
  - Linear Regression
- Example: Imputation - Dealing with Missing Data

Let’s get to know each other…

Breakout room



You

Another student

Name

Location

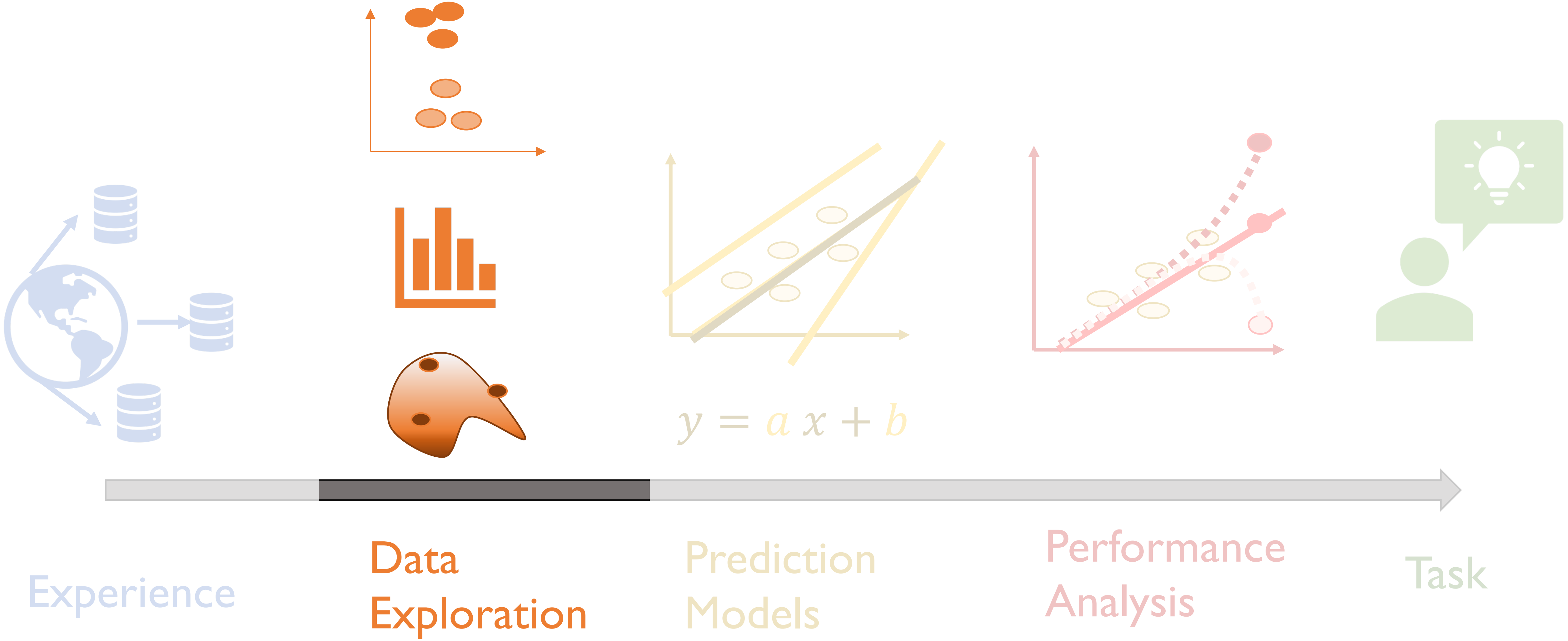
Department

Year

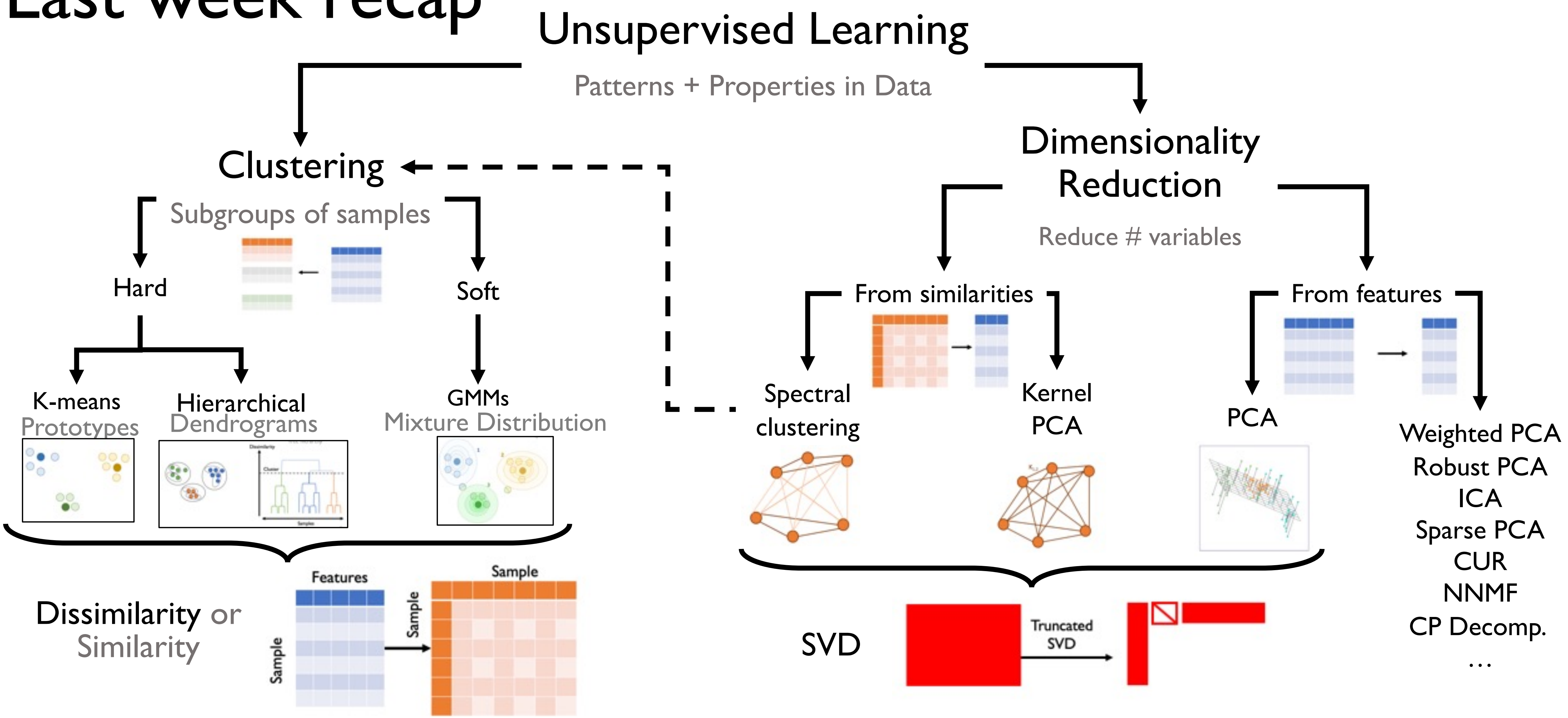
Have you discovered a new TV series, book ?

3 mins

Chat/Audio/Video

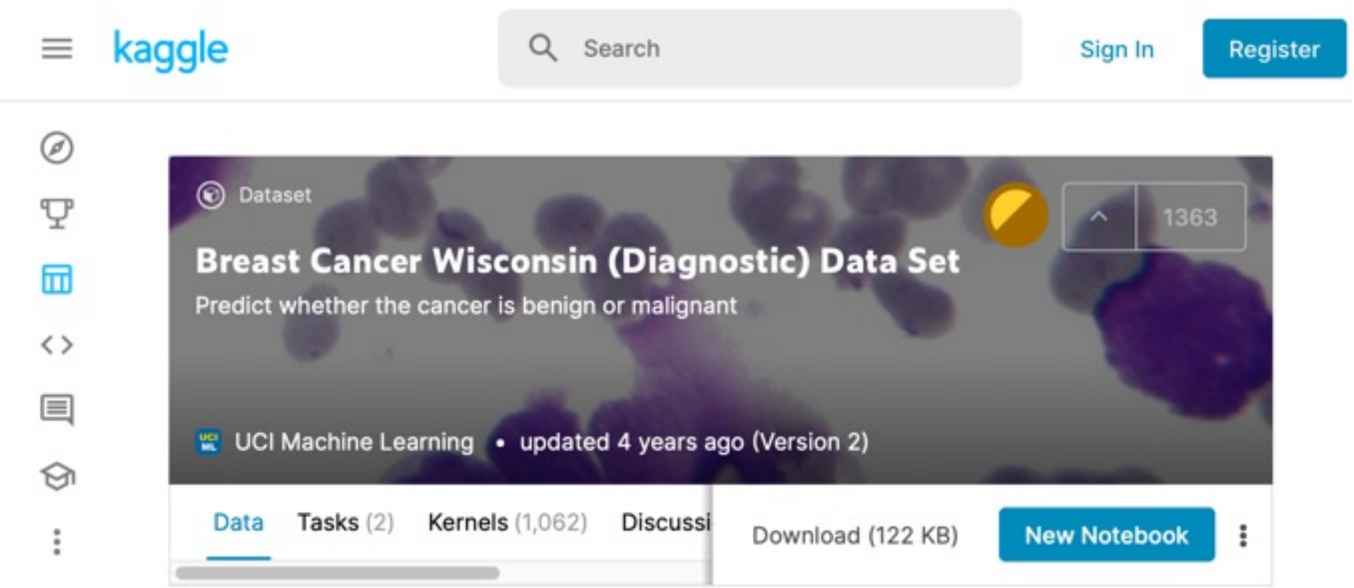

$y = a\,x + b$
Experience
Data Exploration
Prediction Models
Performance Analysis
Task

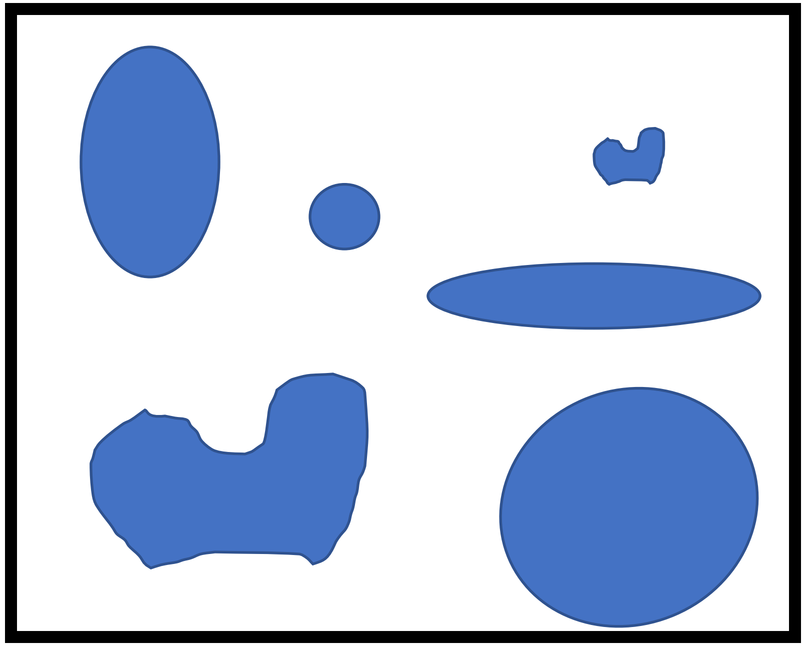
Last week recap



Example of Data Exploration :

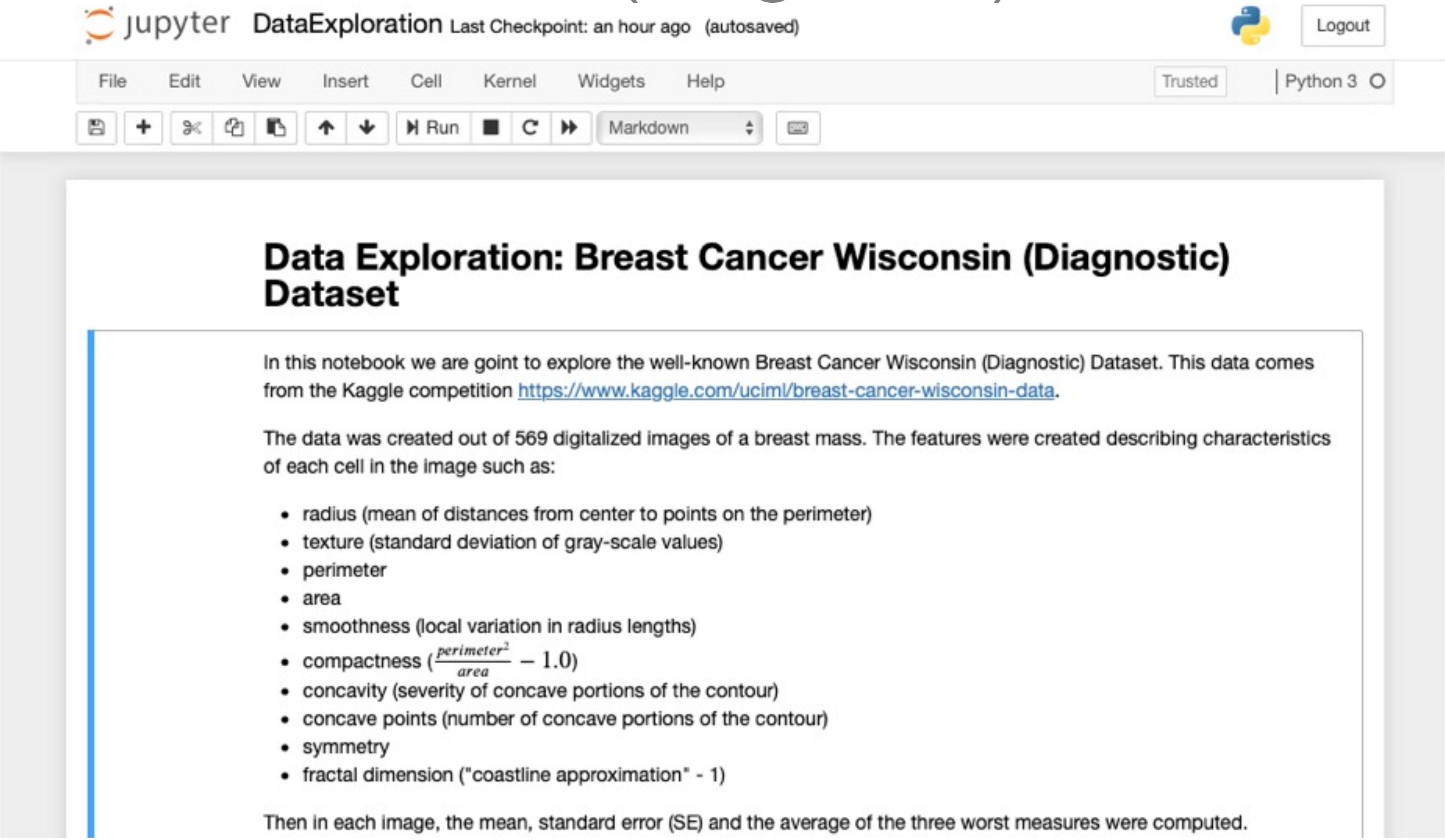
Breast Cancer Wisconsin (Diagnostic) Dataset





https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Example of Data Exploration :
Breast Cancer Wisconsin (Diagnostic) Dataset


jupyter DataExploration Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Data Exploration: Breast Cancer Wisconsin (Diagnostic) Dataset

In this notebook we are goint to explore the well-known Breast Cancer Wisconsin (Diagnostic) Dataset. This data comes from the Kaggle competition https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

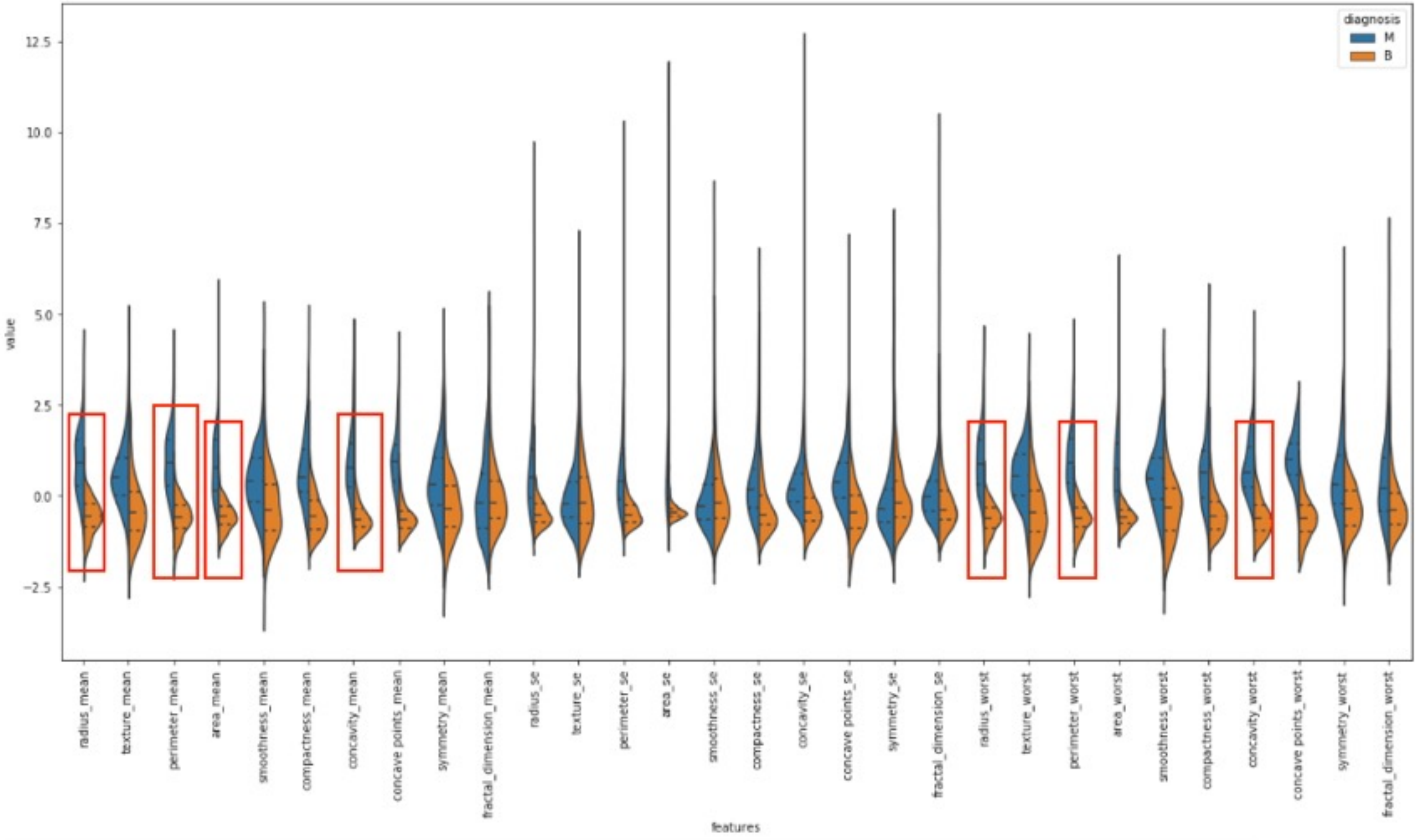
The data was created out of 569 digitalized images of a breast mass. The features were created describing characteristics of each cell in the image such as:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ($\frac{perimeter^2}{area} - 1.0$)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

Then in each image, the mean, standard error (SE) and the average of the three worst measures were computed.

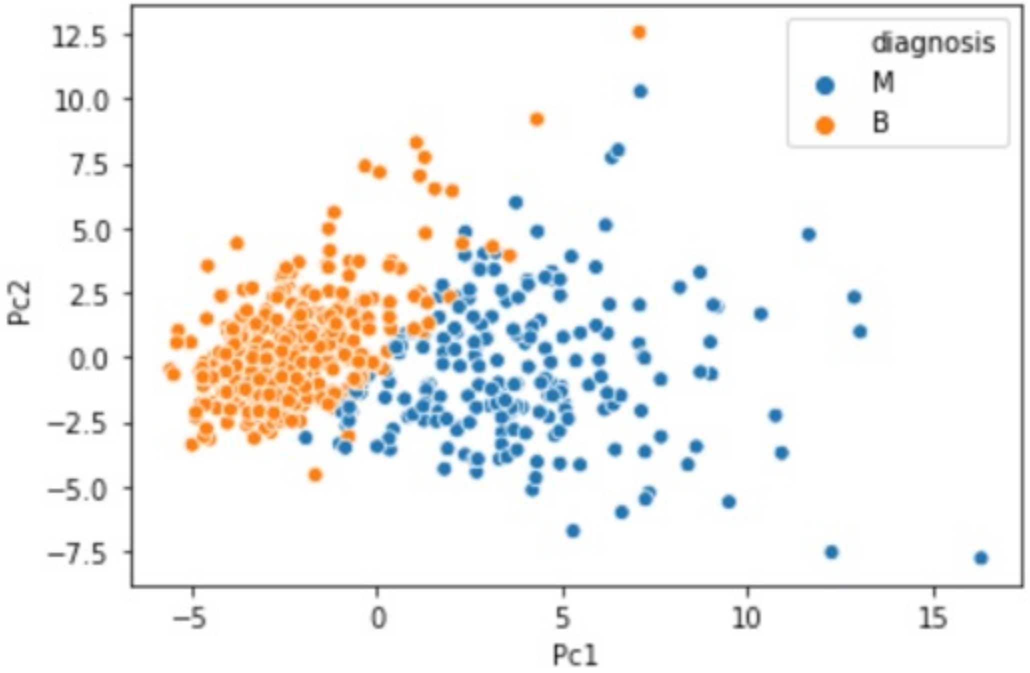
Notebook + Data in course resources

Example of Data Exploration :
Breast Cancer Wisconsin (Diagnostic) Dataset

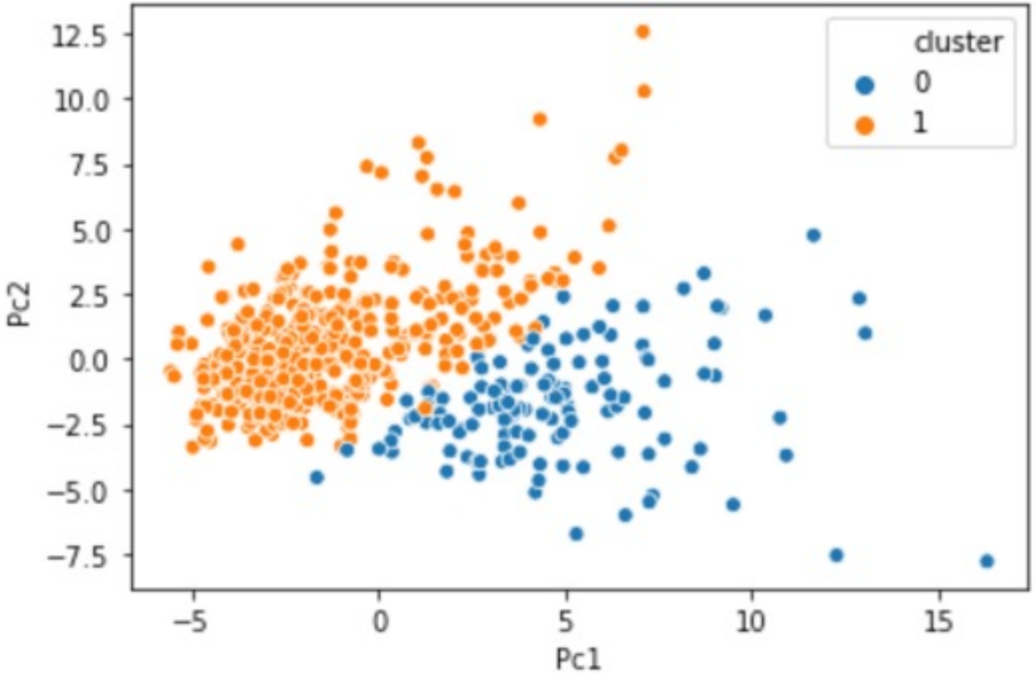


Notebook + Data in course resources

Example of Data Exploration :
Breast Cancer Wisconsin (Diagnostic) Dataset



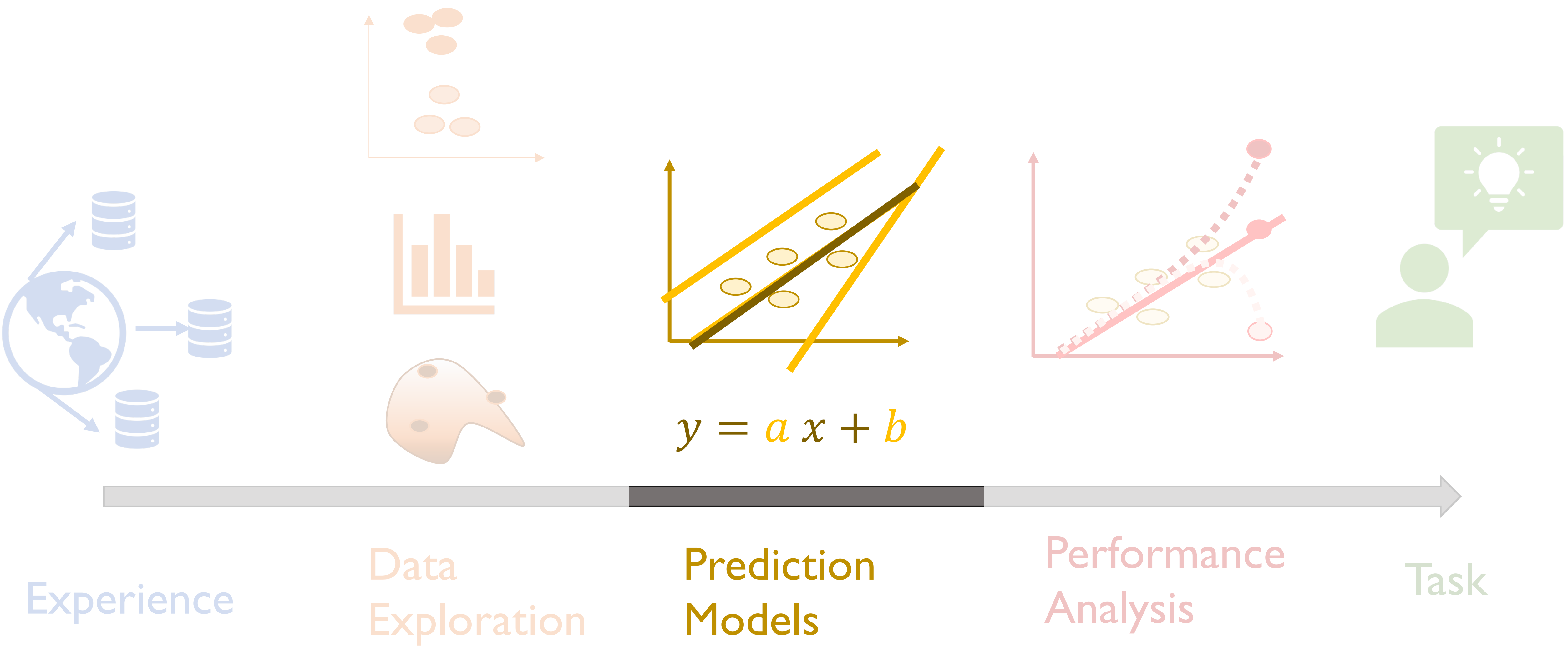
True Labels



K-means Clustering

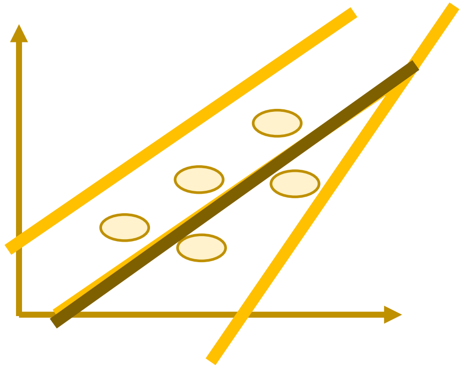
How do we predict labels?

Notebook + Data in course resources



$y = a\,x + b$
Experience
Data Exploration
Prediction Models
Performance Analysis
Task

Supervised Learning Part I:
K-Nearest Neighbors & Linear Regression



$$y = a\,x + b$$

Prediction
Models

Introduction to Statistical Learning
Chapter 3: Linear Regression

Elements Statistical Learning
Chapter 3.2: Linear Regression
Chapter 13.3: K-Nearest-Neighbor Classifier

Supervised Learning

"Learn by example"

Features

Inputs | Outputs

Given | Sample

$$\left(x_1^{(1)}, x_2^{(1)}, \dots, x_{p-1}^{(1)}, x_p^{(1)}\right) \xrightarrow{\hat{f}} y^{(1)}$$

$$\left(x_1^{(2)}, x_2^{(2)}, \dots, x_{p-1}^{(2)}, x_p^{(2)}\right) \longrightarrow y^{(2)}$$

$$\dots$$

$$\left(x_1^{(n)}, x_2^{(n)}, \dots, x_{p-1}^{(n)}, x_p^{(n)}\right) \longrightarrow y^{(n)}$$

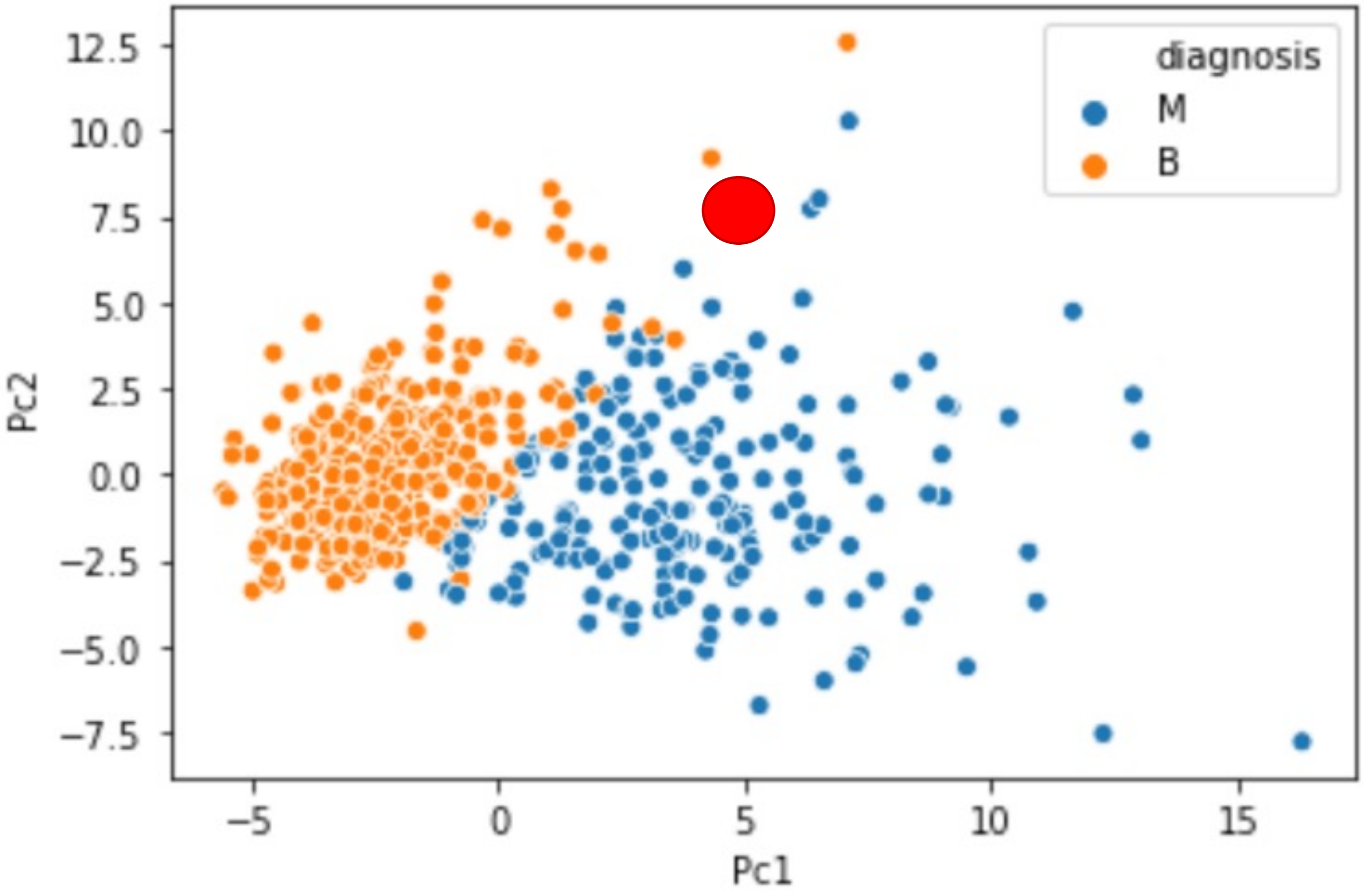
Categorical
Classification

Quantitative
Regression

We want to predict

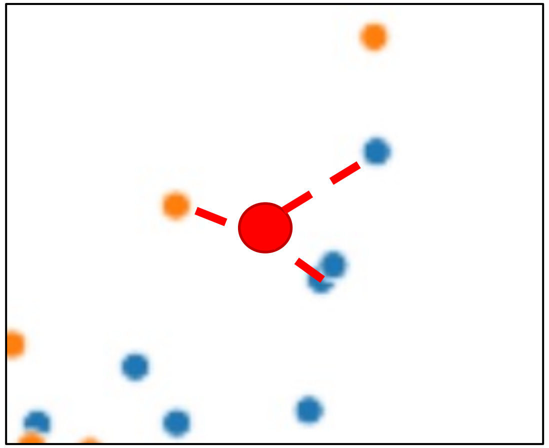
$$(x_1, x_2, \dots, x_{p-1}, x_p) \xrightarrow{\hat{f}} ?$$

Supervised Learning



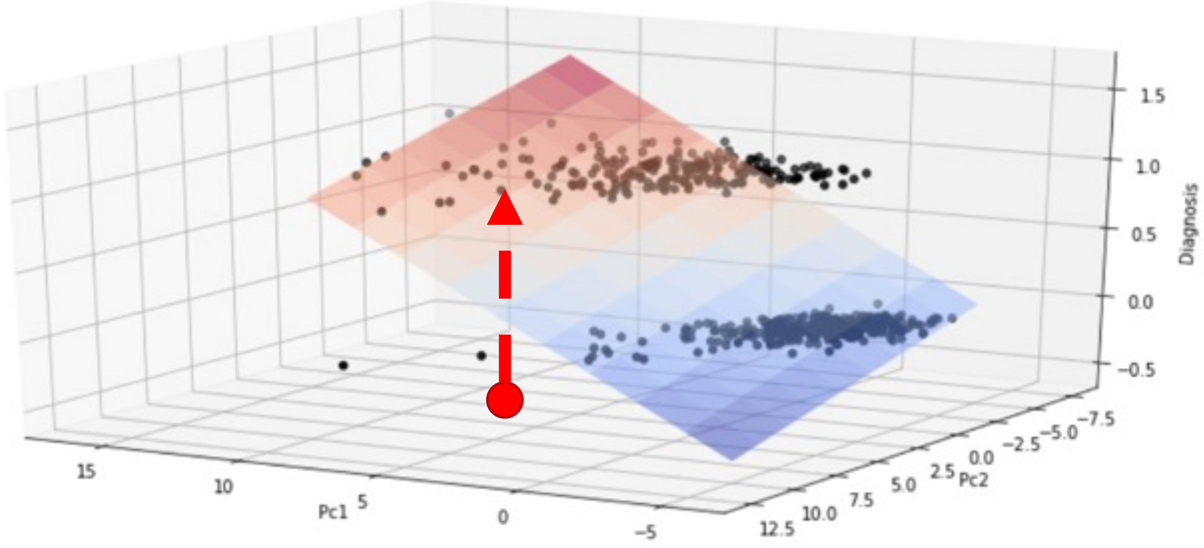
What is the diagnosis for this sample?

Types of supervised learning



Non-parametric

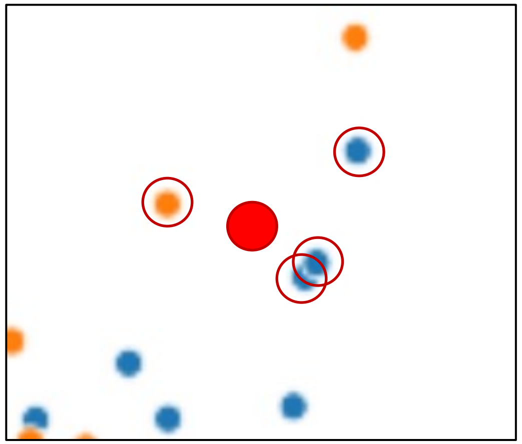
Fit local model for each data

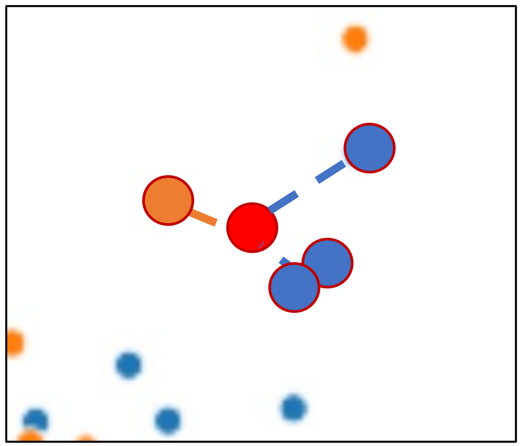


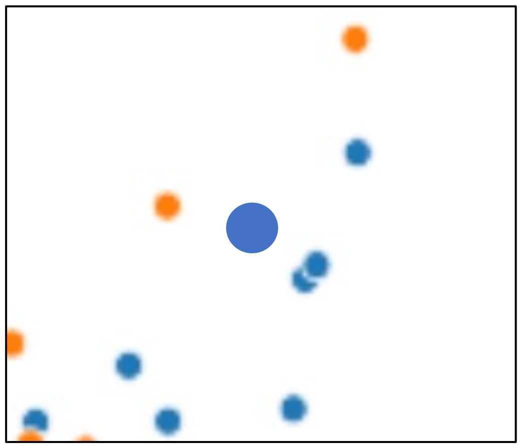
Parametric

Fit parameters of model for all the data

K-Nearest Neighbors (Non- parametric)







1) Find k nearest neighbors

2) Y categorical:
Most common

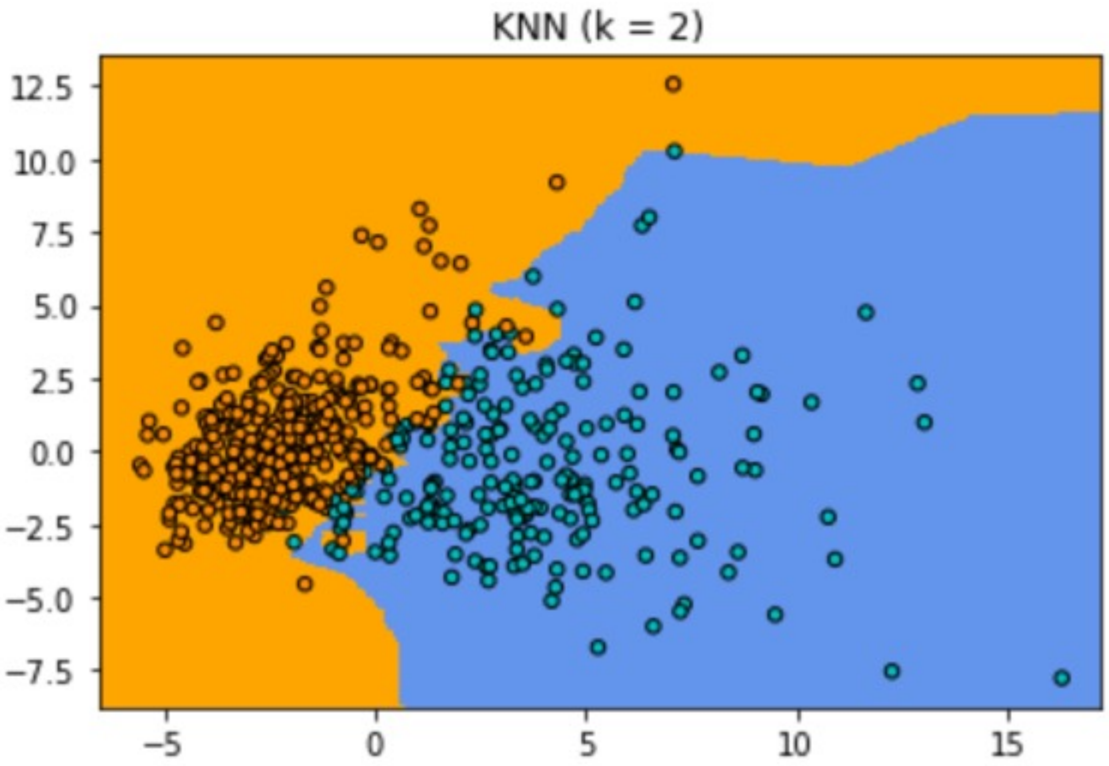
Y quantitative:

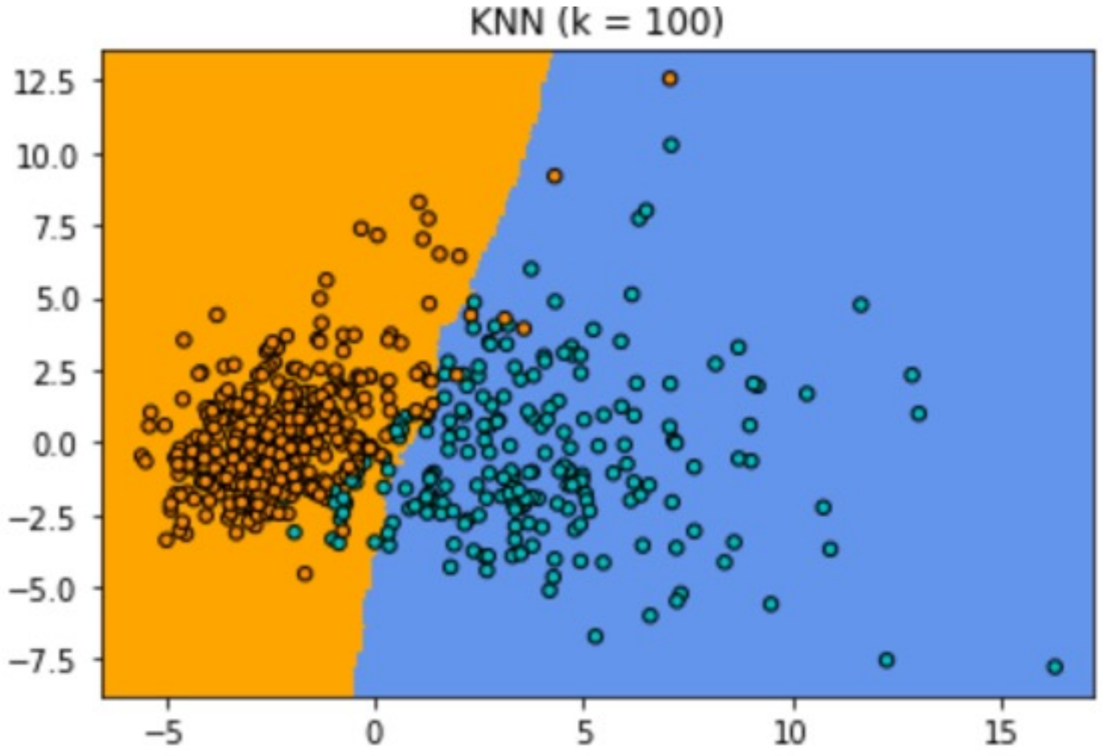
$$\frac{1}{K}\sum_{j\in N_k(\bullet)} y^{(j)}$$

3) Assign new value

* Nearest depends on similarity, usually Euclidean

K-Nearest Neighbors: Breast Cancer





Challenges K-Nearest Neighbors

Easy to implement

Flexible

Choose k (Next class)

Choose weights

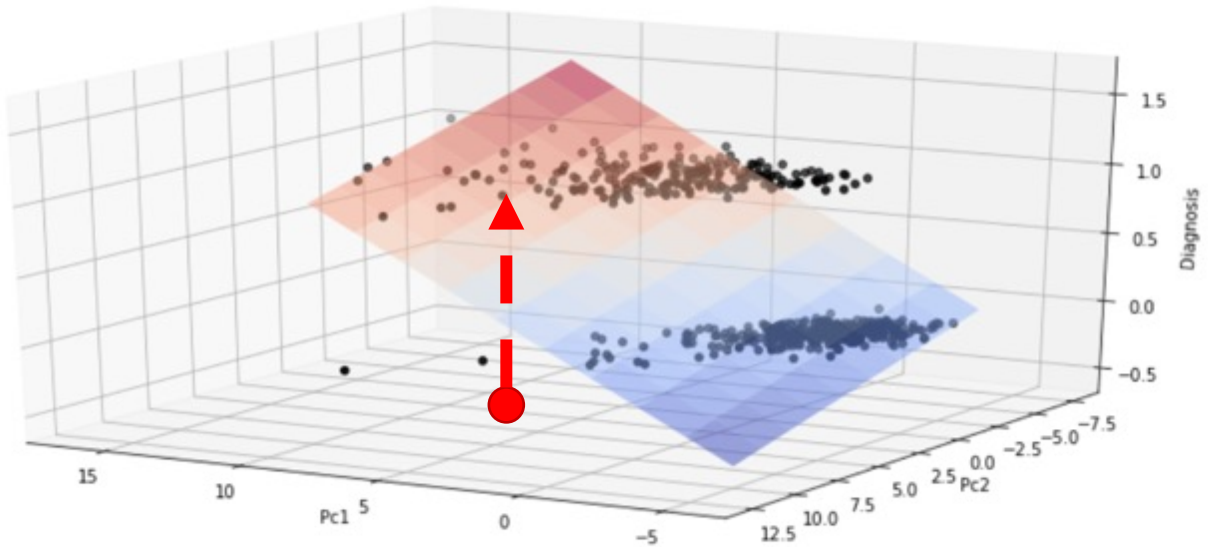
Computationally Expensive: compute k-NN for each sample

Dependent on distance

Sensitive to imbalanced data sets

Linear Regression

"Simplest model that we could assume"

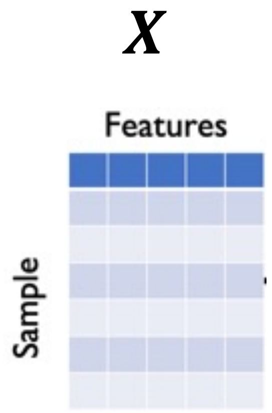


$$y \approx f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

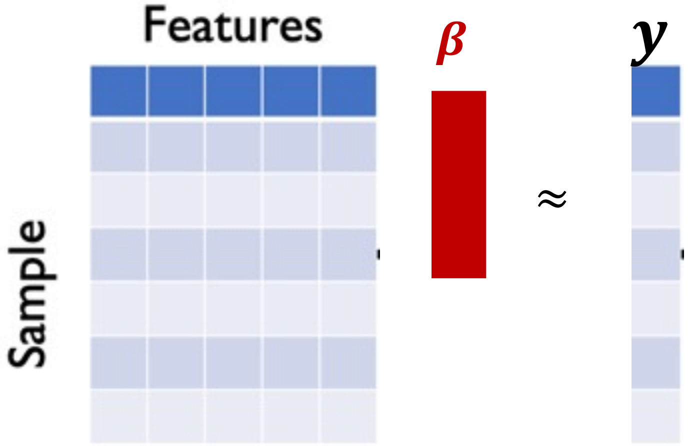
Train : Find $\beta_0, \beta_1, \ldots, \beta_p$

Linear Regression: Finding Coefficients

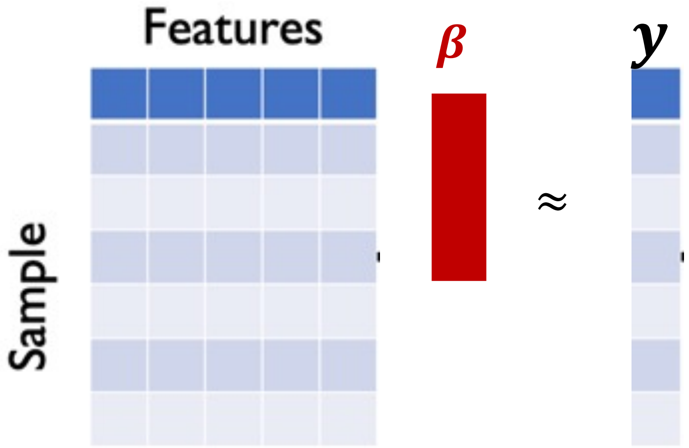
$$y \approx f(x) = \color{red}{\beta_0} + \color{red}{\beta_1} x_1 + \cdots + \color{red}{\beta_p} x_p$$



$$\left\{ \begin{array}{l} \left(1, x_1^{(1)}, x_2^{(1)}, \dots, x_{p-1}^{(1)}, x_p^{(1)}\right) \longrightarrow y^{(1)} \\ \left(1, x_1^{(2)}, x_2^{(2)}, \dots, x_{p-1}^{(2)}, x_p^{(2)}\right) \longrightarrow y^{(2)} \\ \qquad \dots \\ \left(1, x_1^{(n)}, x_2^{(n)}, \dots, x_{p-1}^{(n)}, x_p^{(n)}\right) \longrightarrow y^{(n)} \end{array} \right\}$$



Linear Regression: Finding Coefficients



Solve the least-squares problem

$$\min_{\alpha} \|\boldsymbol{X}\beta - y\|_2^2$$

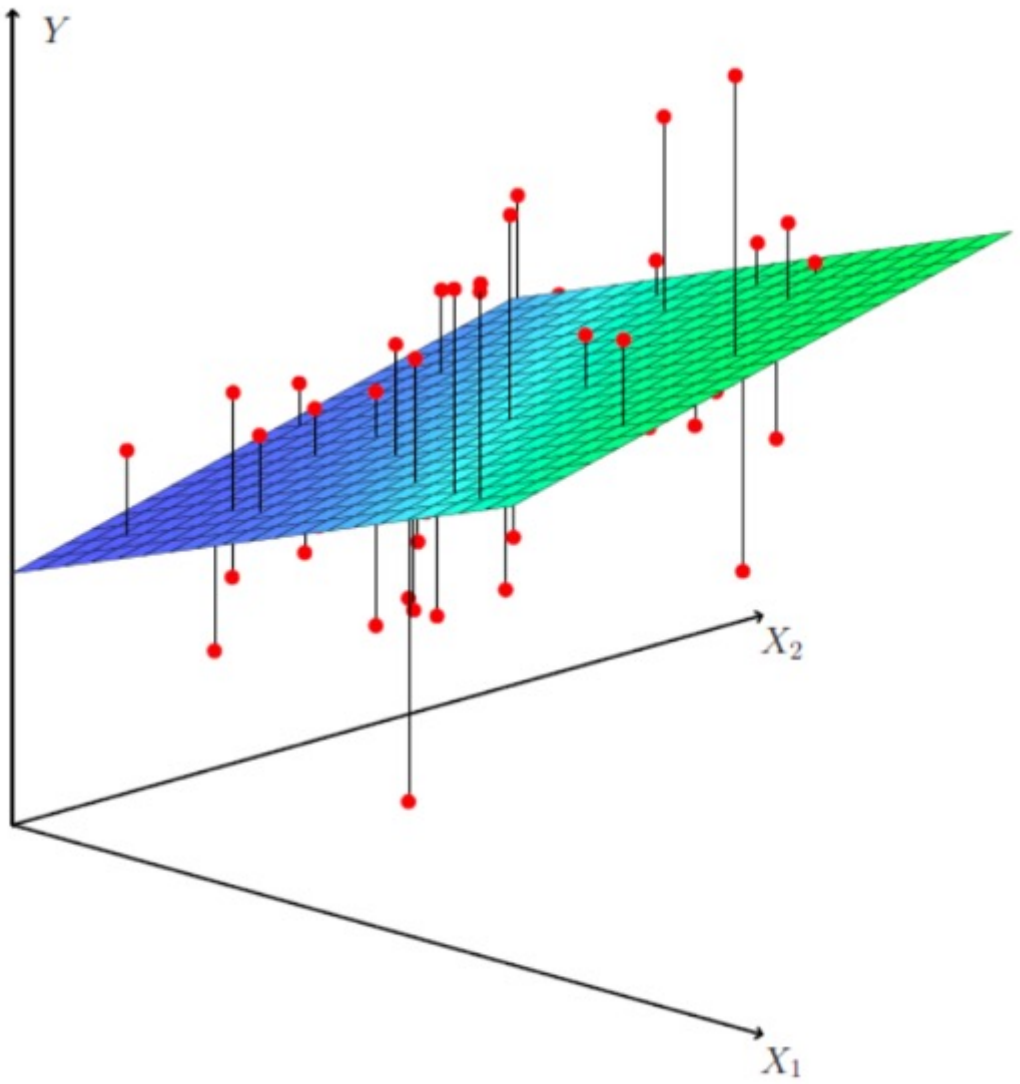
where

$$\|\boldsymbol{X}\beta - y\|_2^2 = \sum_{i=1}^{n} \left(\overbrace{\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} - y^{(i)}}^{\text{residual}} \right)^2$$

Solution: Normal equations

$$\beta = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T y$$

Linear Regression: What does $||\boldsymbol{X}\beta - y||_2^2$ mean?



$$||\boldsymbol{X}\beta - y||_2^2 = \sum_{i=1}^{n} \left(f\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Linear Regression: Relationship with N(0,1)

If we assume $Y \sim Normal$

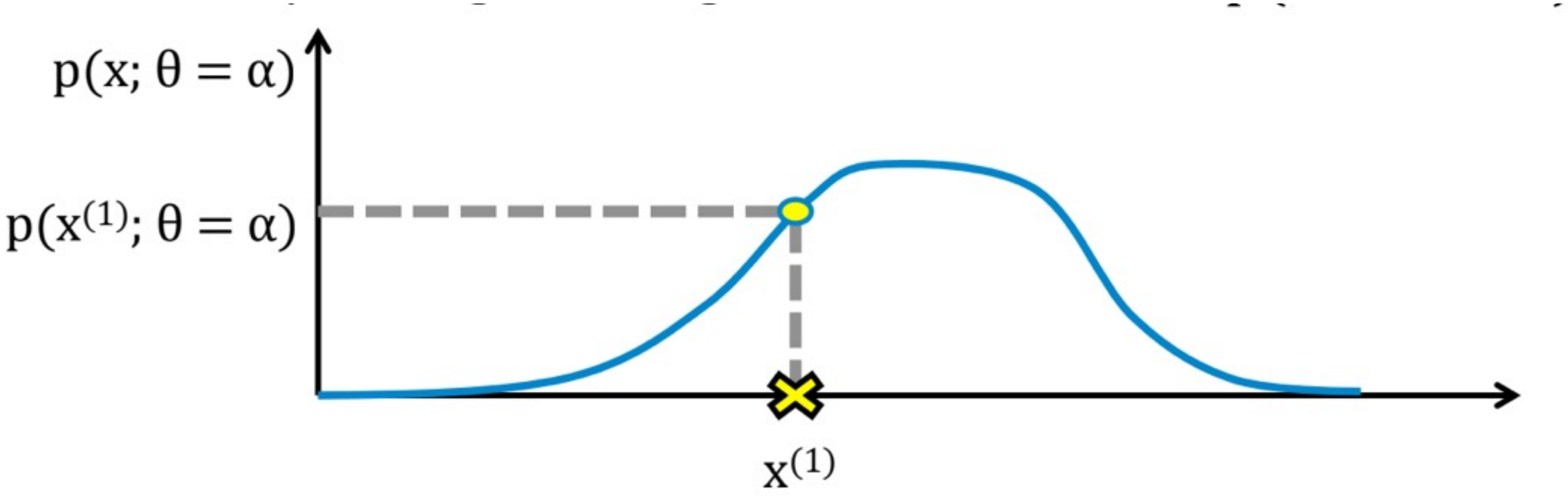
$$E[Y|X] = f(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Then the Maximum Likelihood Estimator of β is the solution of

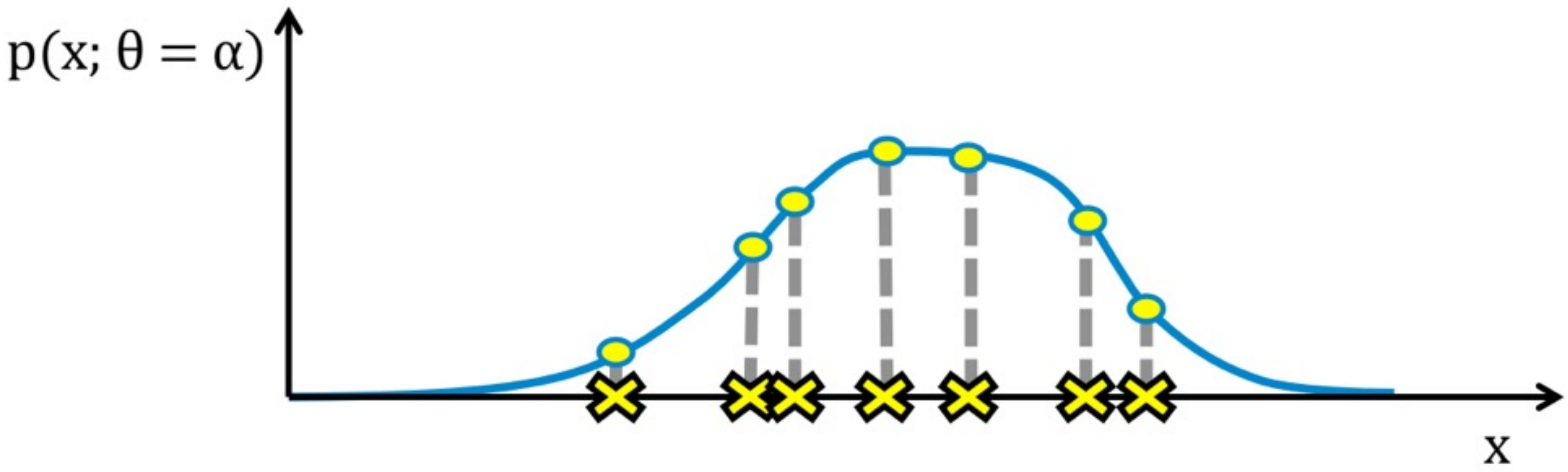
$$\min_{\alpha} ||\boldsymbol{X}\beta - y||_2^2$$

In Stats, performance = check normality assumptions

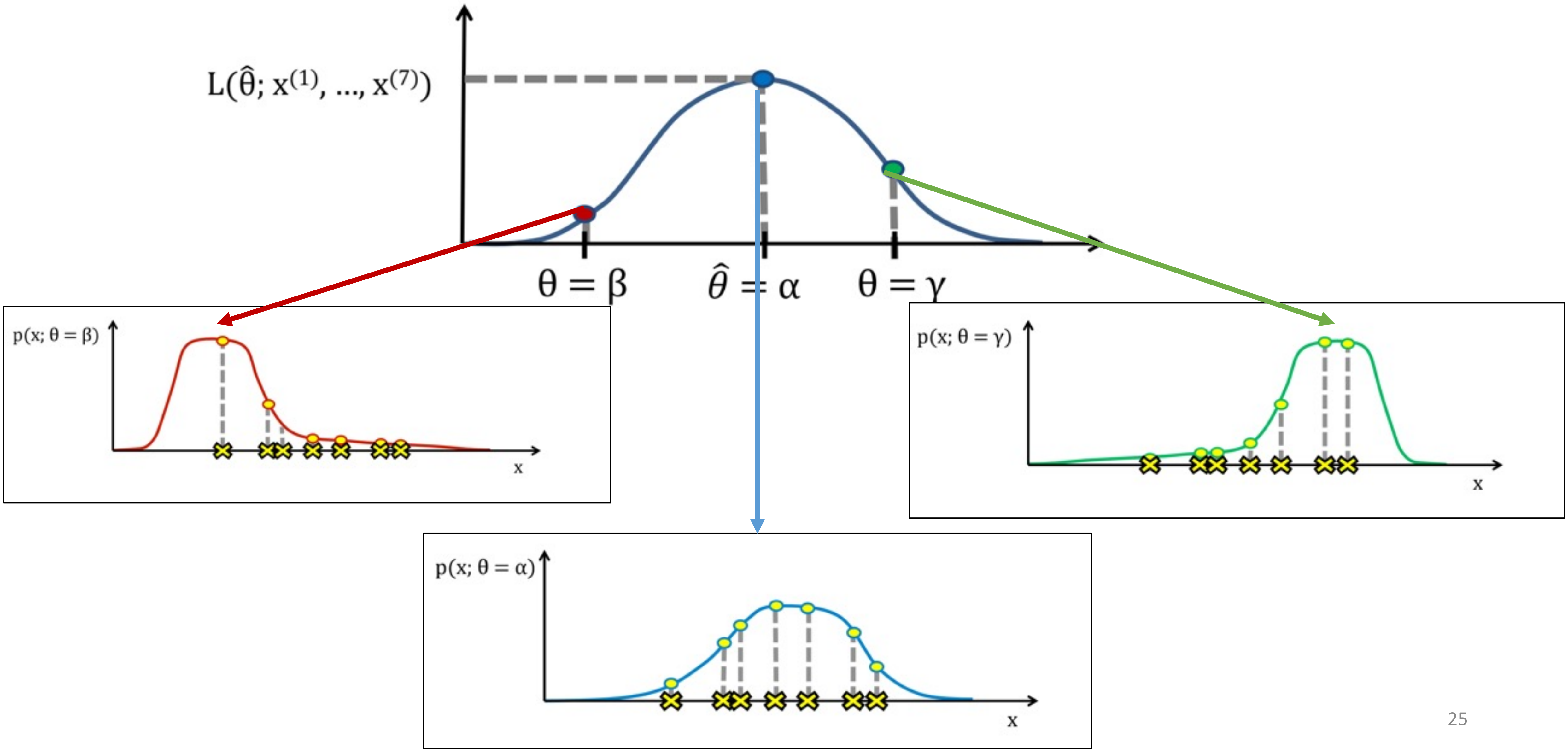
Recall Maximum Likelihood



Recall Maximum Likelihood



Recall Maximum Likelihood

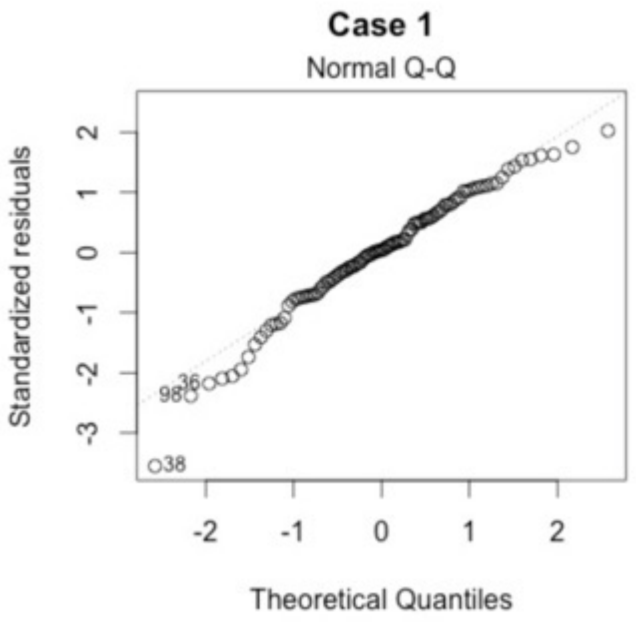


Linear Regression: Relationship with N(0,1)

$$R^2 = 1 - \frac{\sum_i \left(f\left(x^{(i)}\right) - y^{(i)}\right)^2}{\sum_i (\bar{y} - y^{(i)})^2}$$

Proportion of variability in Y explained by X

Residual plots



Significance, Confidence Interval

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

ISL Table 3.6

https://data.library.virginia.edu/diagnostic-plots/

Linear Regression: Relationship with N(0,1)

Inference

Prediction

Linear Regression: Relationship with N(0,1)

Inference

Relationships

between X and Y

STATS

Interpretability / Significance of coefficient

Prediction

Generalize f to

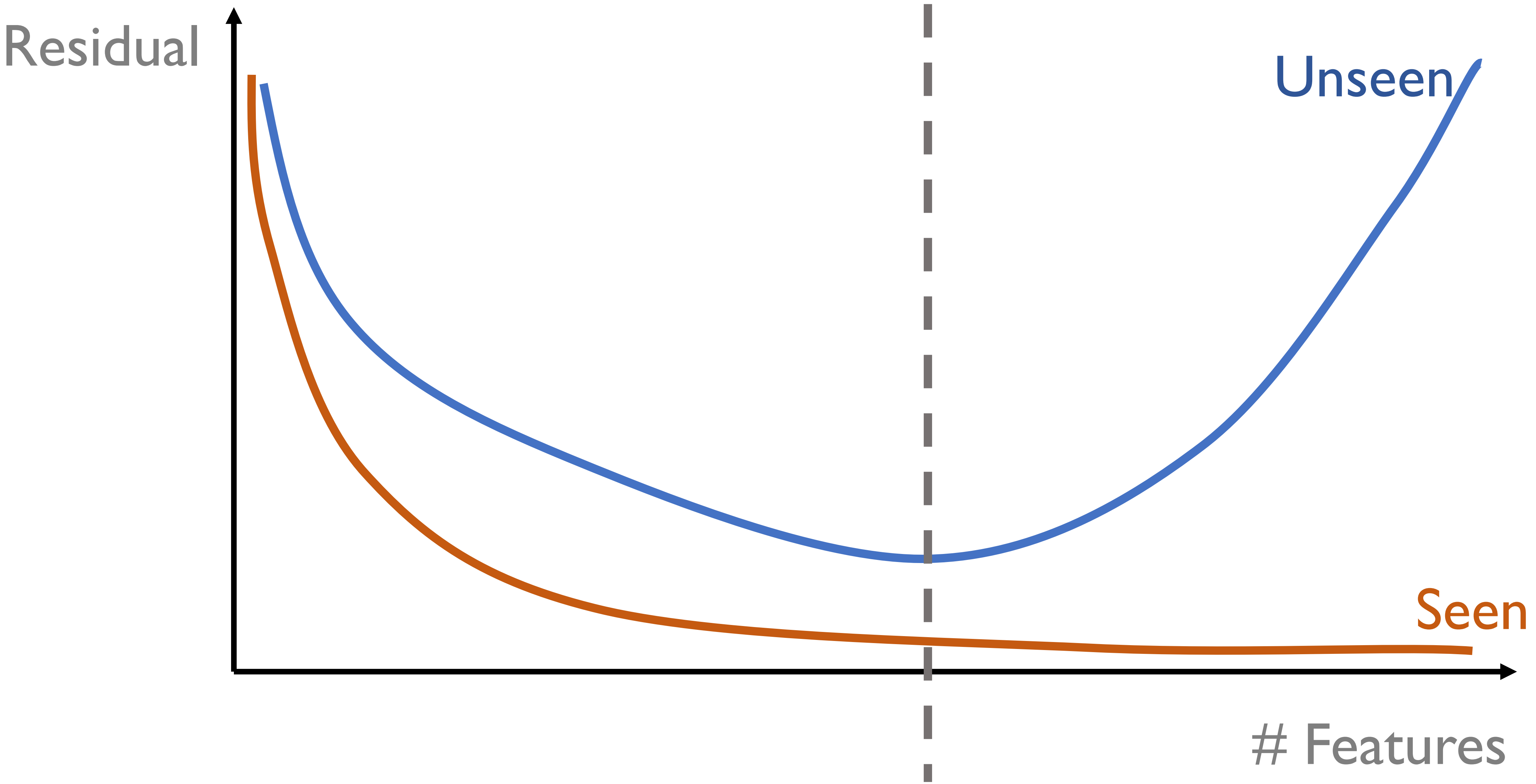
unseen X

ML

Minimize Residual

Linear Regression: Relationship with N(0,1)

Minimize residual (next class)



Linear Regression: Breast Cancer

1) Transform Categorical into Quantitative variables

M ⟶ 1

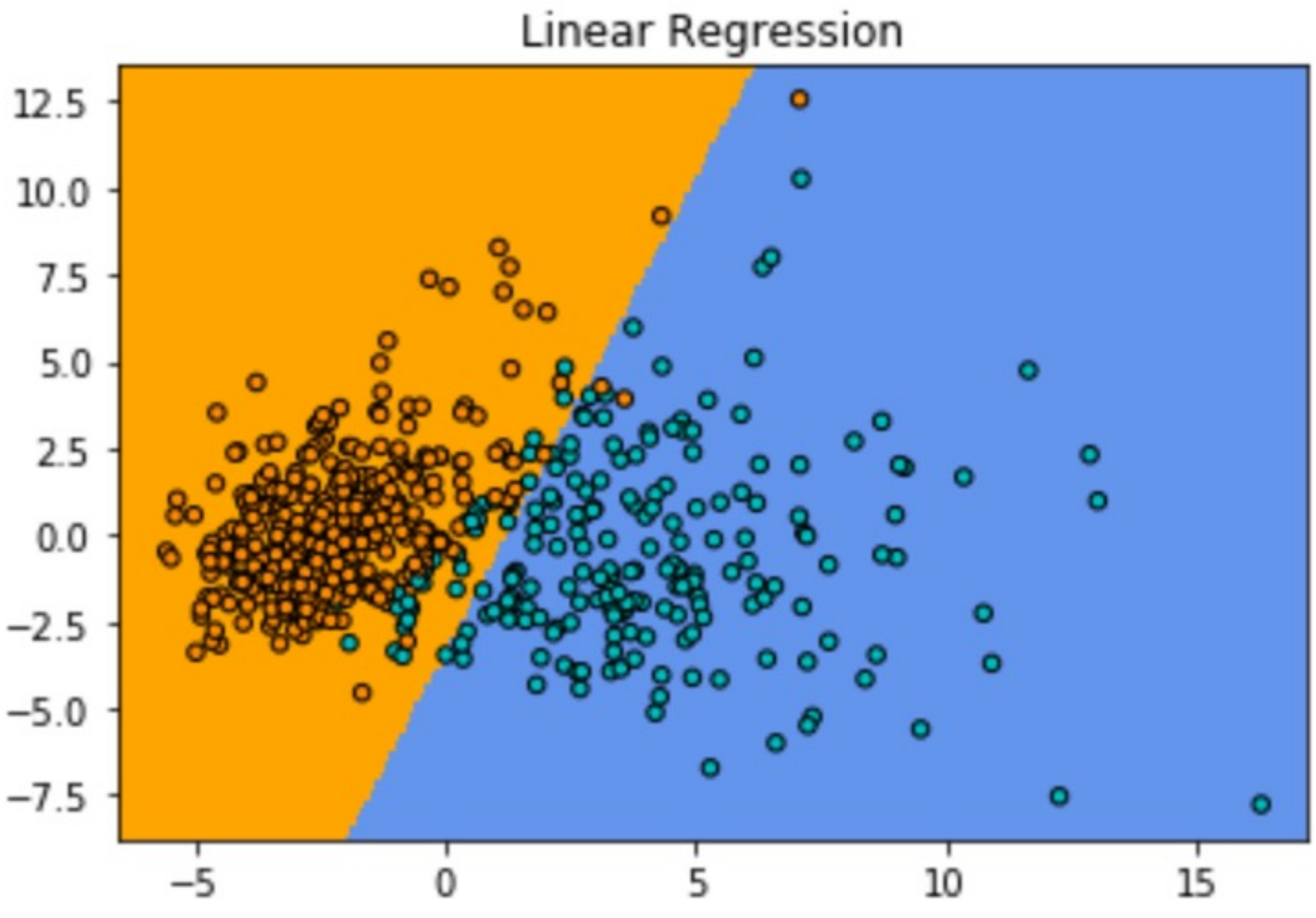
- - - - - - - - - - - - - - 0.5

B ⟶ 0

2) Solve least squares problem

We use 1st and 2nd Principal Components

Linear Regression: Breast Cancer



Challenges Linear Regression

Simple model

Interpretable coefficients

Good results with small data sets

Too simple model

- Add non linearities
- Assume other distributions

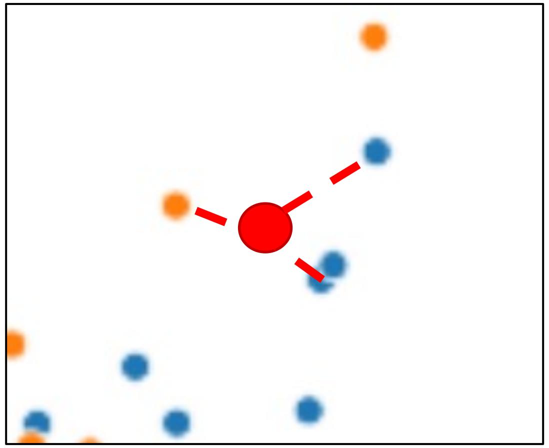
Is it useful for classification?

Feature Selection

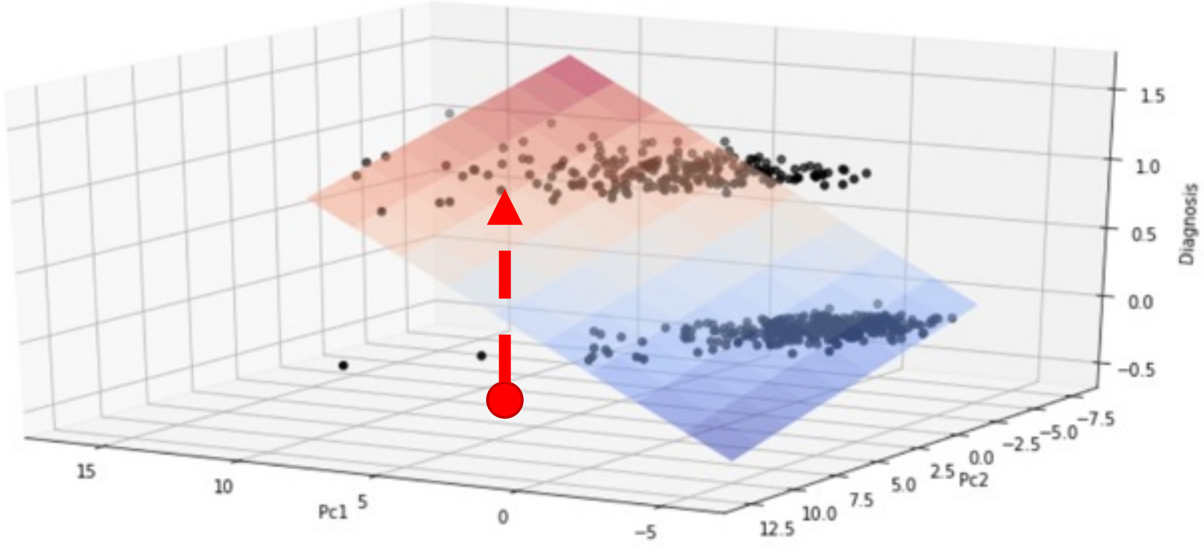
Sensitive to outliers

Poor extrapolation

Types of supervised learning



K-Nearest Neighbors



Linear Regression

Can we use them to deal with missing values?

Missing Values

Option 1

Remove samples from training

Option 2

Impute values based in other samples

Option 3

Use **models** that handle missing values

OK if we have enough samples

MCAR: Missing Completely at Random

CART: Classification and Regression Tree

Imputation

Variable type → Classification
Variable type → Regression

Option 1

Replace by **mean** (most common category) in feature

Option 2

Assign value using **KNN** computing with rest of variables

Option 3

Assign value using **Linear Regression** computing with rest of variables

https://scikit-learn.org/stable/modules/impute.html