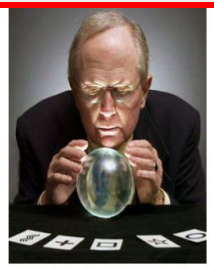
Multivariate Methods

CLASSIFICATION & ORDINATION

roadmap

First Tutorial

Second Tutorial



1. Input matrix (CSV-File)

2. Pre-processing
- Abundance-threshold
- Normalization to themeasurement
- Z-Score Normalization

2b. Test for normal distribution

	A	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
2	Protein 1	20	49	75	22	6	44	35	34
3	Protein 2	39	90	74	55	92	89	67	55
4	Protein 3	87	41	92	57	96	42	89	5
5	Protein 4	21	27	2	25	48	89	48	79
6	Protein 5	58	39	82	12	57	88	82	15
7	Protein 6	38	72	54	7	86	56	87	91
8	Protein 7	13	46	44	76	54	67	87	17
9	Protein 8	38	50	19	63	90	88	13	50
10	Protein 9	96	29	45	34	23	98	59	53
11	Protein 10	12	60	27	26	1	80	15	51
12	Protein 11	89	93	8	9	8	45	35	62
13	Protein 12	96	21	37	25	14	45	81	16

3. Check for Correlation?

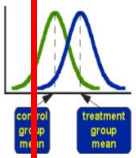
Lineare correlation???

yes

no

son correlation

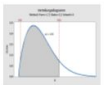
3b. Spearmans Rank correlation



Check for significant differences?

2 Groups

Normal distribution???



yes

no

Welch T-test (two-side, unequal variances)

Mann-Whitney Test

Differences of Groups

ANOSIM

Permanova

Check for patterns

Non-supervised learning

Grouping

Reduktion of Complexity

Hierachical Clustering

PCA

PCoA

NMDs

TMDS

Supervised learning

Decision Tree

GOALS :

Overview : Normalization & Group Comparison

Moving Ahead - Multivariate Methods : Supervised Learning & Unsupervised Learning, Ordination & Classification

Ordination: PCA, PCoA & NMDS

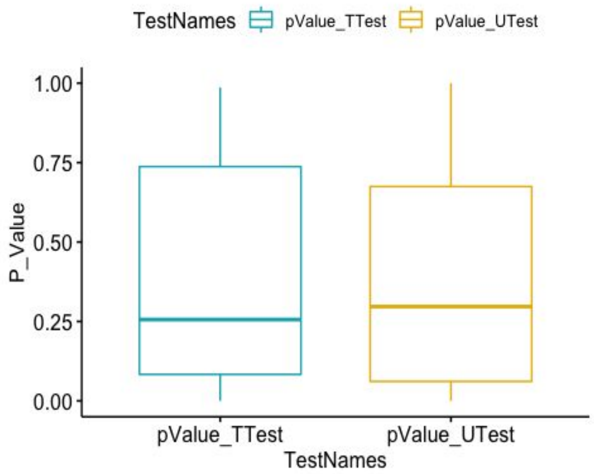
OVERVIEW

- what are we upto?
- keywords

  NORMALIZATION? GROUP COMPARISON? WHY ALL THE FUSS?
- previously unclarified

  p-value, w-value, Bonferoni Correction T-Test, Benjamin Hochberger Correction T-Test

what are we upto?



why?

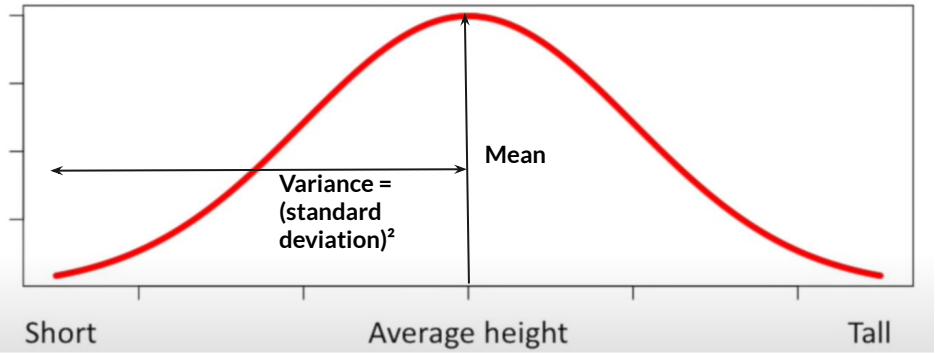
to **visualize** large amounts of complex **data** is easier than poring over spreadsheets or reports. ... **Data visualization** can also: **Identify areas that need attention or improvement.**

how?

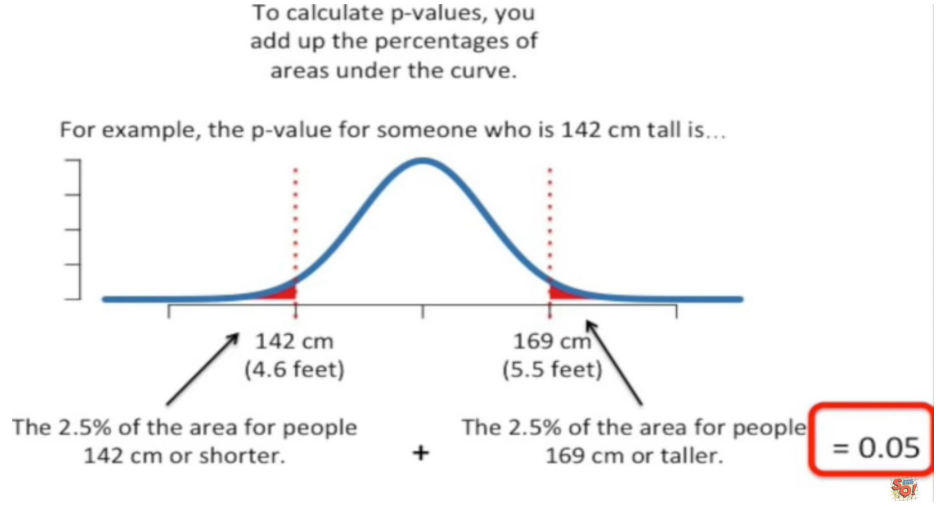
Statistical Tools through **R :**

- Normalization
- Group Comparison **(T-Test, PERMANOVA etc.)**
- Multivariate Methods **(Clustering, Ordination)**

keyword : normalization



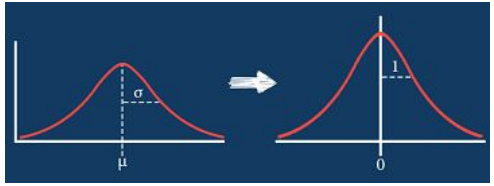
p-value

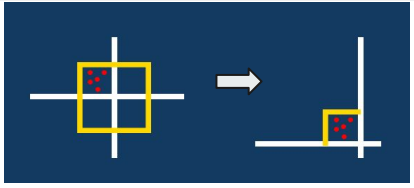


why bother?

- Robust **visualization** of a data or data variable - possible to create null hypothesis and test them
- **data normalization** when seeking for **relations**
- as part of data preparation for **machine learning**. The goal of **normalization** is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values
- Easy to **compare** data or data variables

how?

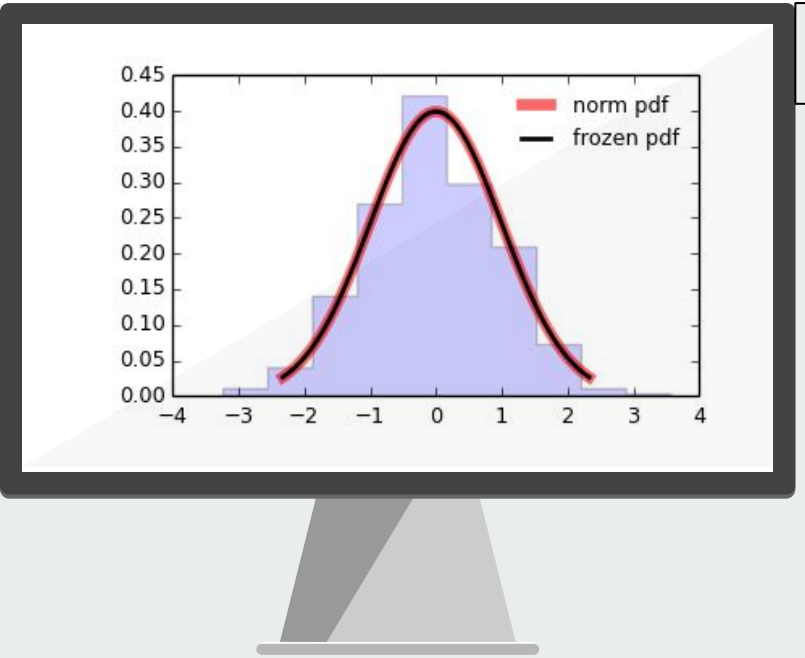




Name(ID)	**Age**	**Height**	**Gender** (1=f, 2=m, 3=other)	**Education Level** (0=Bachelor, 1= Master, 2= Post Doc)	**Class Label : Teacher(1) or Student(0)**
Robert	30	6.1	m(2)	Post Doc(2)	Teacher(1)
Julian	26	6.3	m(2)	Master(1)	Student(0)
Danial	25	5.8	m(2)	Master(1)	Student(0)
Max	26	5.9	m(2)	Master(1)	Student(0)
Faizan	23	6.0	m(2)	Master(1)	Student(0)
Abdullah	27	5.8	m(2)	Master(1)	Student(0)
Ammar	26	5.9	m(2)	Master(1)	Student(0)
Rahul	25	5.8	m(2)	Master(1)	Student(0)
Mean	**26**	**5.95**	**2**	**1.125**	

Name(ID)	**Age**		**Height**		**Gender** (1=f, 2=m, 3=other)	**Education Level** (0=Bachelor, 1= Master, 2= Post Doc)	**Class Label : Teacher(1) or Student(0)**
Robert	30	1	6.1	3/5	m(2)	Post Doc(2)	Teacher(1)
Julian	26	3/7	6.3	1	m(2)	Master(1)	Student(0)
Danial	25	2/7	5.8	0	m(2)	Master(1)	Student(0)
Max	26	3/7	5.9	1/5	m(2)	Master(1)	Student(0)
Faizan	23	0	6.0	2/5	m(2)	Master(1)	Student(0)
Abdullah	27	4/7	5.8	0	m(2)	Master(1)	Student(0)
Ammar	26	3/7	5.9	1/5	m(2)	Master(1)	Student(0)
Rahul	25	2/7	5.8	0	m(2)	Master(1)	Student(0)
Mean	**26**	3/7	**5.95**	0.3	**2**	**1.125**	

Test for Normality: Shapiro-Wilk Test




```
> shapiro.test(matrix$BE_03)

        Shapiro-Wilk normality test

data:  matrix$BE_03
W = 0.38432, p-value = 1.103e-14
```

Assumption Checks ▼

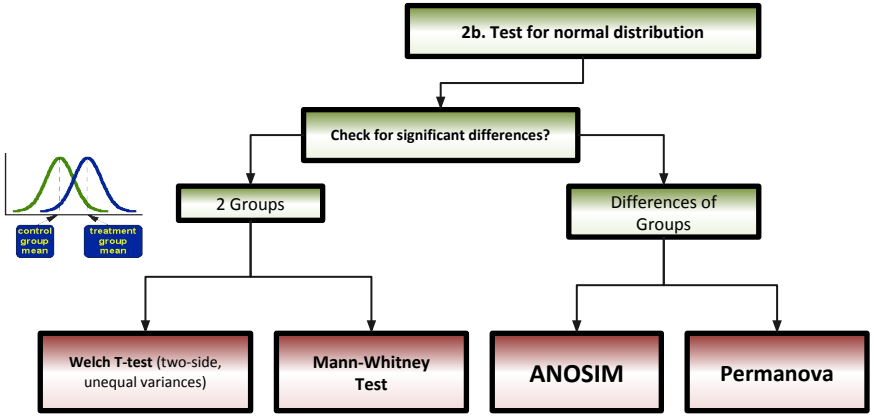
Test of Normality (Shapiro–Wilk) ▼

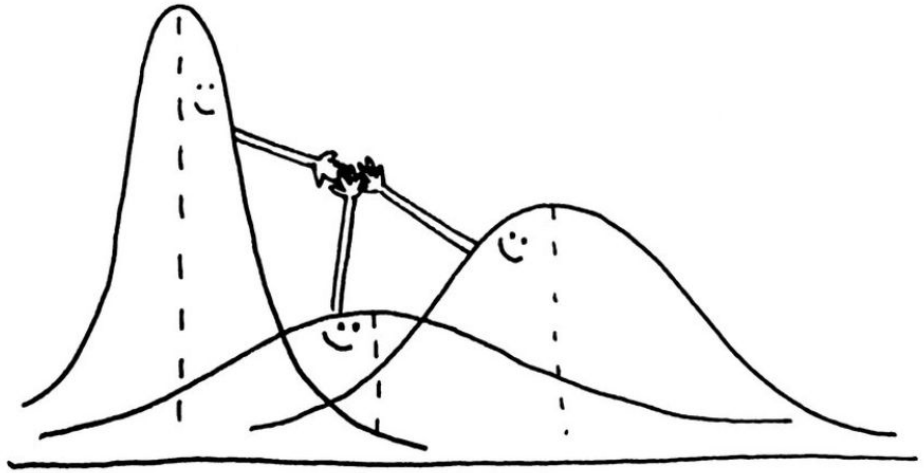
	W	p
Difference	0.938	0.325

Note. Significant results suggest a deviation from normality.

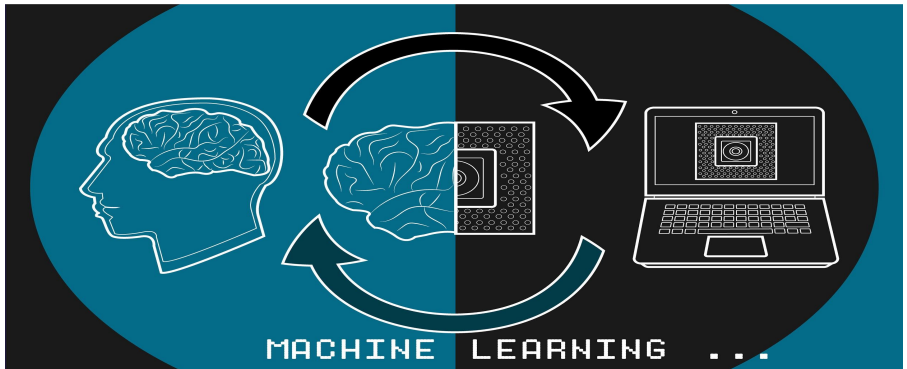
- Using w-value, we create a NULL hypothesis
  - *if W is very small then the distribution is probably not normally distributed*
- If $P < 0.05$, we reject the NULL Hypothesis

So now that we have data(normalized), what next?





- check for **Significant Differences (Group Comparison)**
  - between 2 or more groups
    - T-Test & U-Test
    - ANOSIM & PERMANOVA
    - ANOVA & Kruskal-Walis Test
- infer **Knowledge** out of dataset and/or **prove hypothesis**



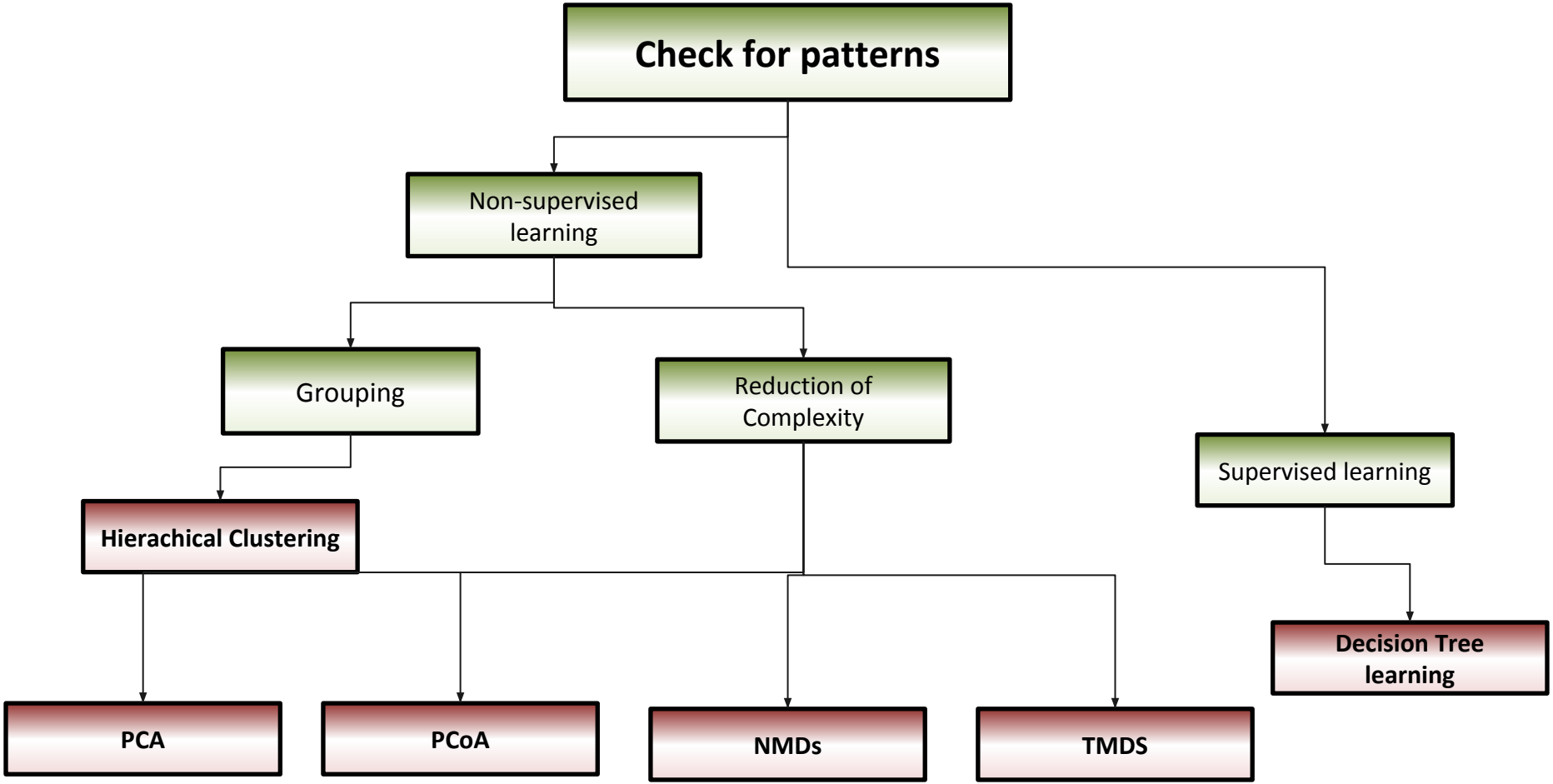
and why is this important?

can DATA be NOT Normalized & still make sense??

Multivariate Methods : Ordination & Classification

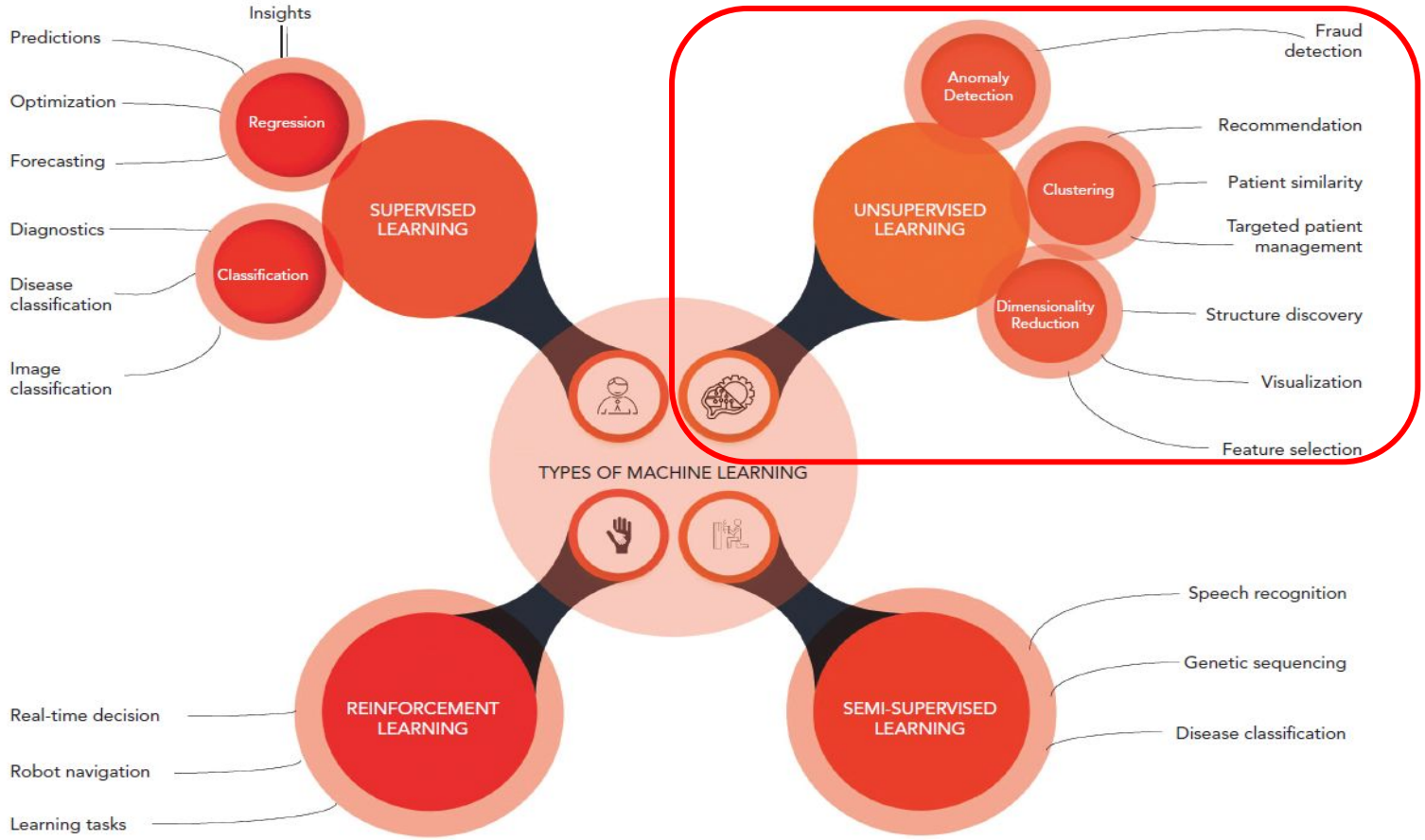
- unsupervised learning vs supervised learning
- Ordination
  - Grouping
    - Clustering
  - Dimension/Complexity Reduction
    - PCA
    - PCoA
    - NMDS
    - CCA

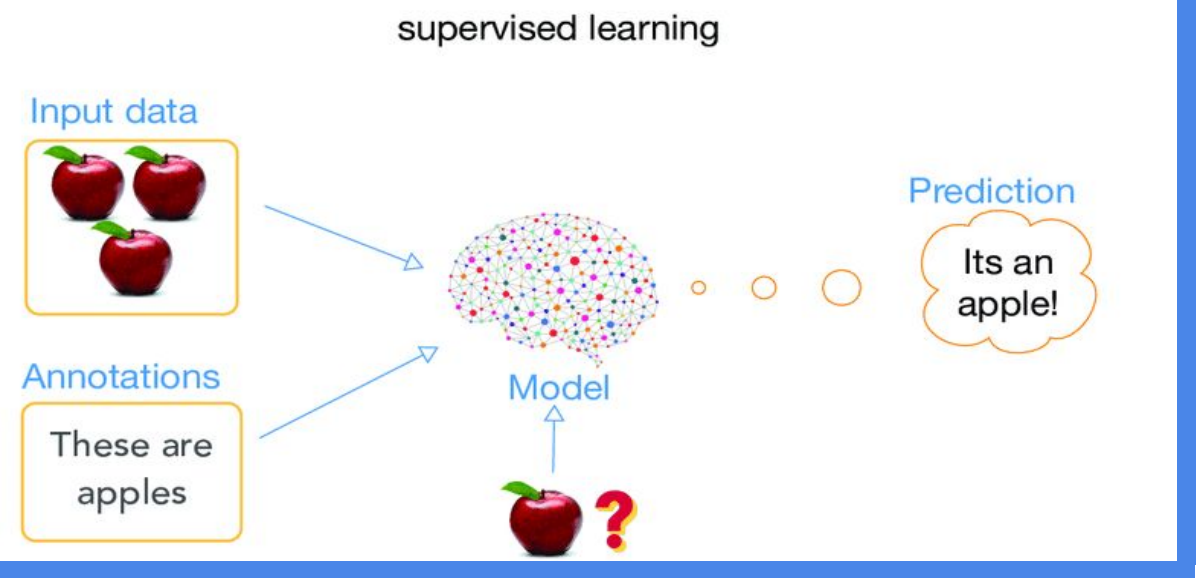
roadmap

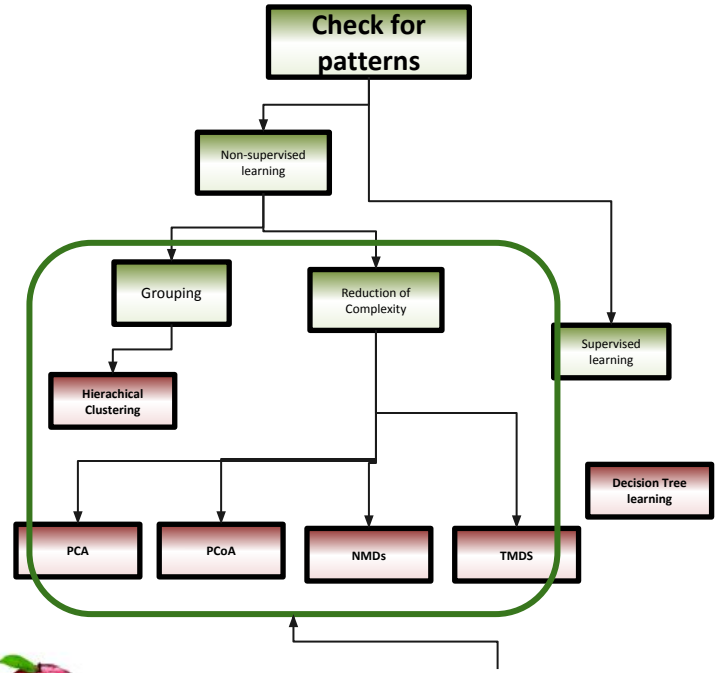

Check for patterns
Non-supervised learning
Grouping
Reduction of Complexity
Supervised learning
Hierachical Clustering
Decision Tree learning
PCA
PCoA
NMDs
TMDS

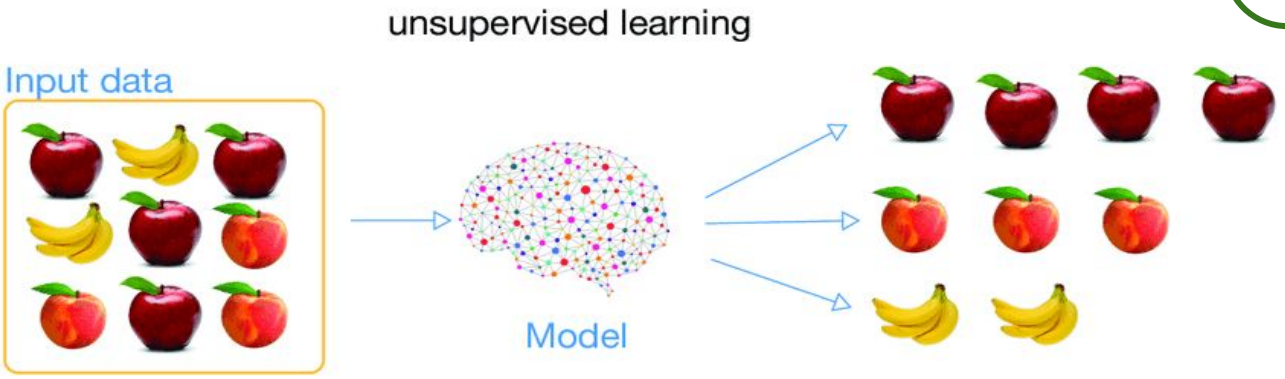
types of Machine Learning

Figure 1: Types of Machine Learning with Examples of Respective Use









ORDINATION

unsupervised learning vs/& **supervised learning**

what is DATA to a Machine??

unsupervised learning

- grouping
  - **Clustering**
- reduction of Dimension and/or Complexity
  - Principal Component Analysis **(PCA)**
  - Principal Coordinate Analysis **(PCoA)**
  - Non Metric MultiDimensional Scaling **(NMDS)**
  - Canonical Correspondence Analysis **(CCA)xx**

why?

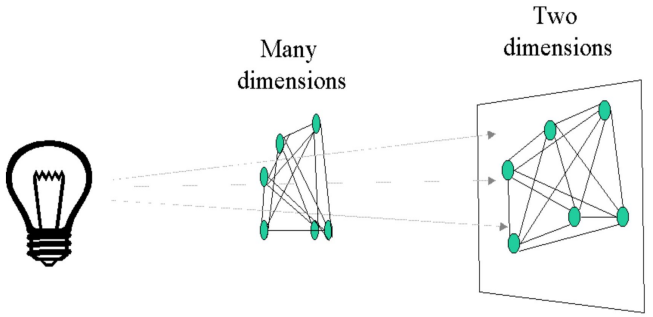
to find **Similarities** & **Recommendations**

Structure Discovery, Feature Selection & **Visualization**

how?

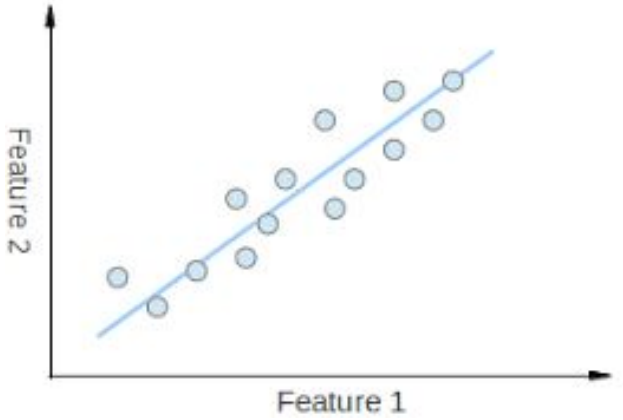
ordination (an unsupervised approach)

Ordination is a collective term for multivariate techniques which summarize a multidimensional dataset in such a way that when it is projected onto a low dimensional space, any intrinsic pattern the data may possess becomes apparent upon visual inspection.

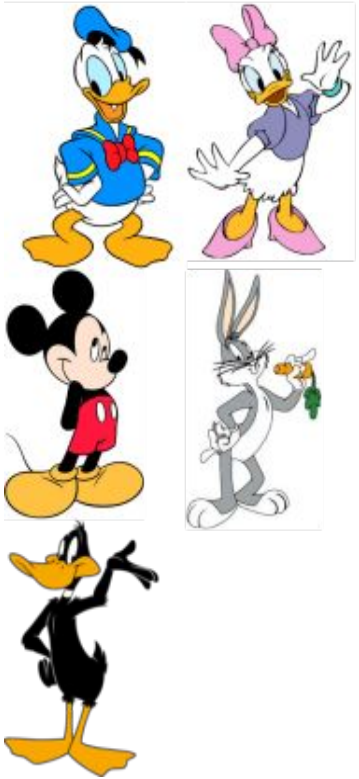


why?

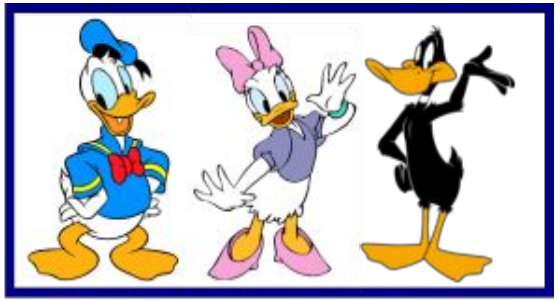
Ordination can be used on the analysis of any set of multivariate objects.

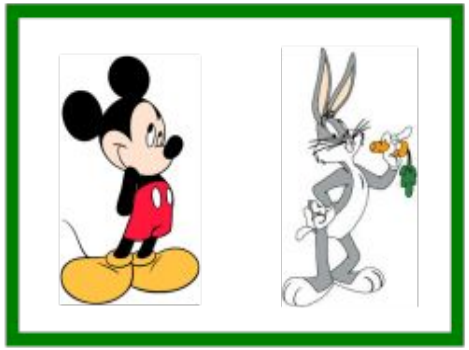


how?



Unsupervised Learning



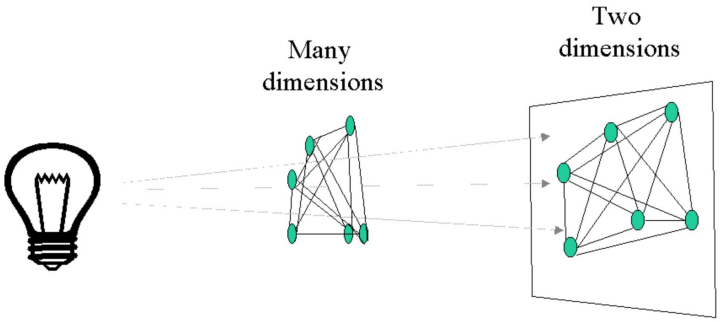


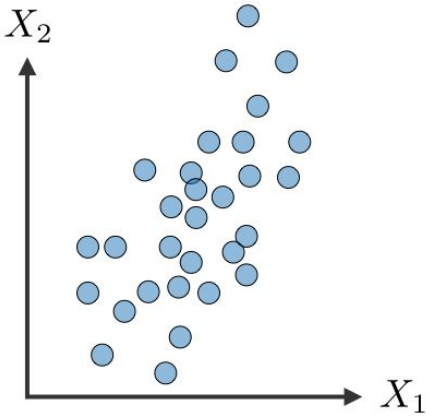
unsupervised learning: CLUSTERING

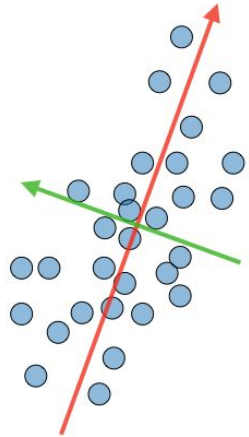
Ordination

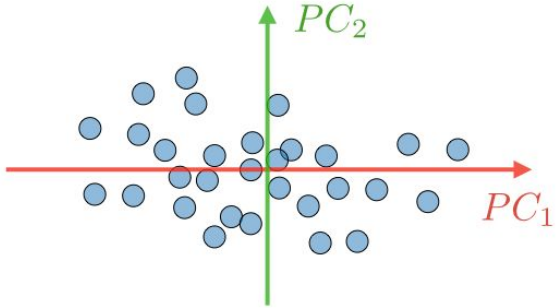
- Dimension Reduction
  - **PCA** (Principal Component Analysis)
  - **PCoA** (Principal Coordinates Analysis)
  - **NMDS** (Non metric Multidimensional Scaling)

PCA (Principal Component Analysis)





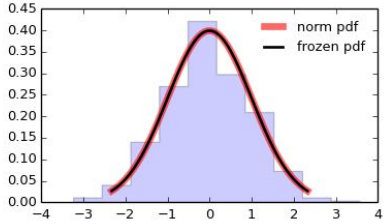


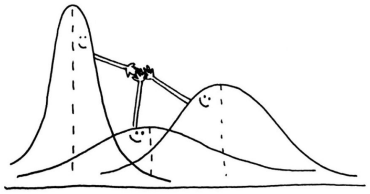


Data in feature space → Find principal components → Data in **p**rincipal **c**omponents space

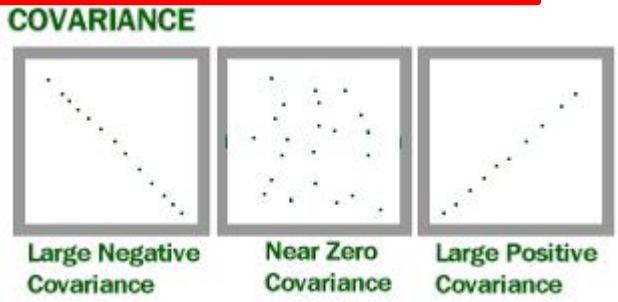
Steps (PCA)

1. Normalize the **Dataset**

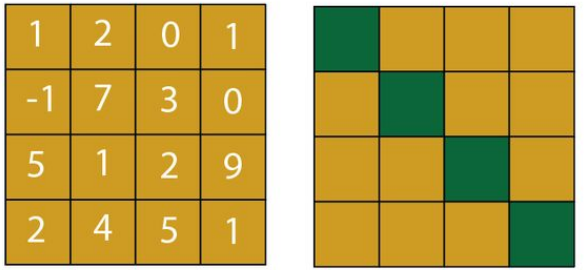




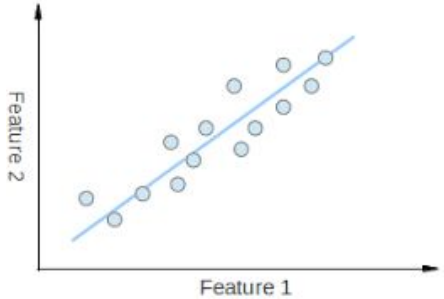
2. Compute **Covariance Matrix**



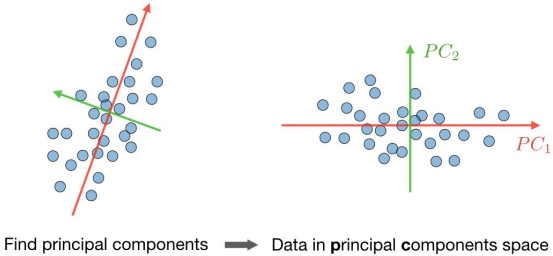
3. Perform **Eigen Decomposition**



4. Determine **Principal Component**



4. Compute **Transformation**

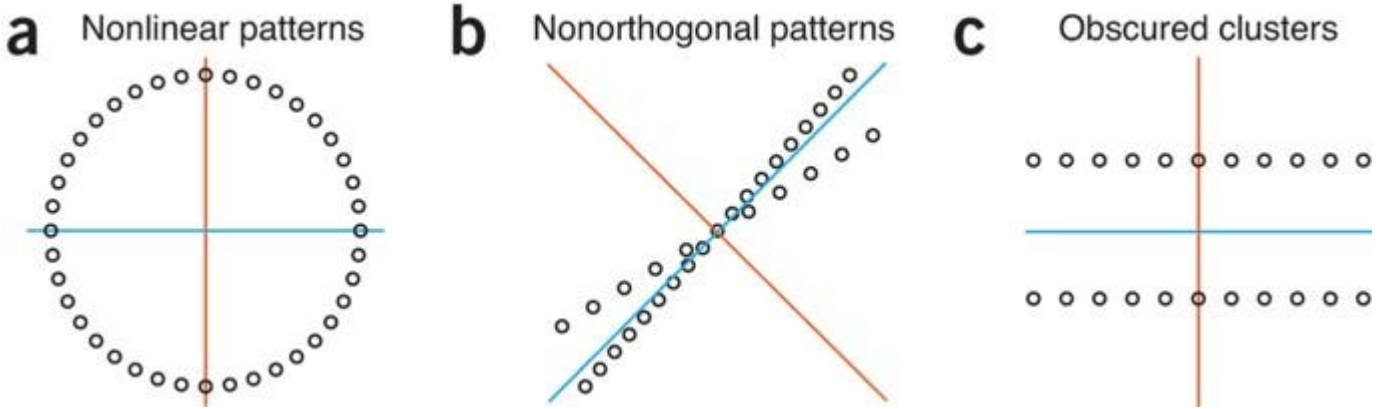


6. VISUALIZATION

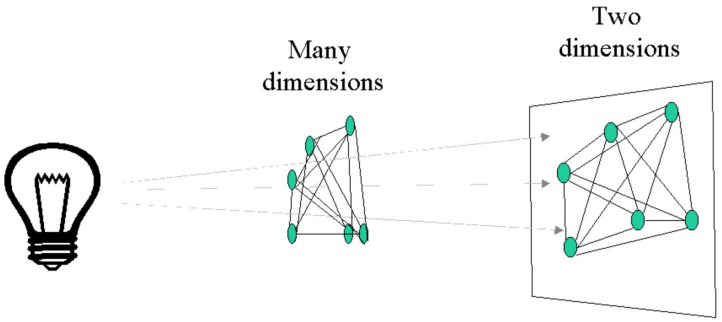
importance(PCA)

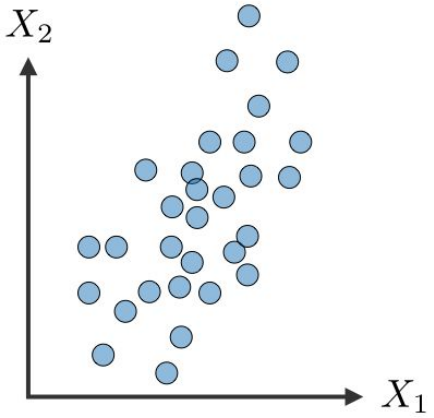
PCA helps you discover correlations & interpret your data, but it will not always find the important patterns.

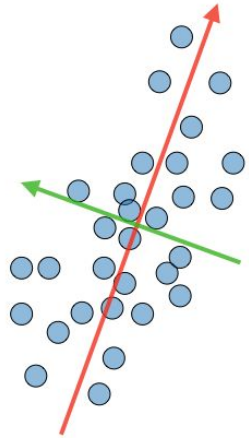
Principal component analysis (PCA) **simplifies the complexity in high-dimensional data while retaining trends and patterns.** It does this by transforming the data into fewer dimensions, which act as summaries of features

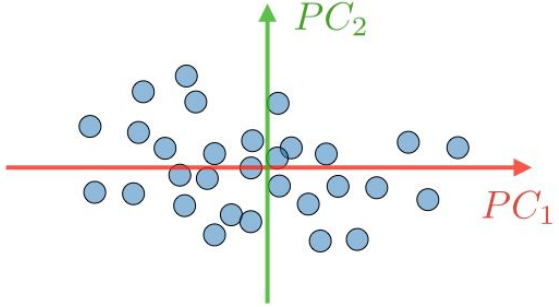


PCoA (Principal Component Analysis)/ metric multidimensional scaling



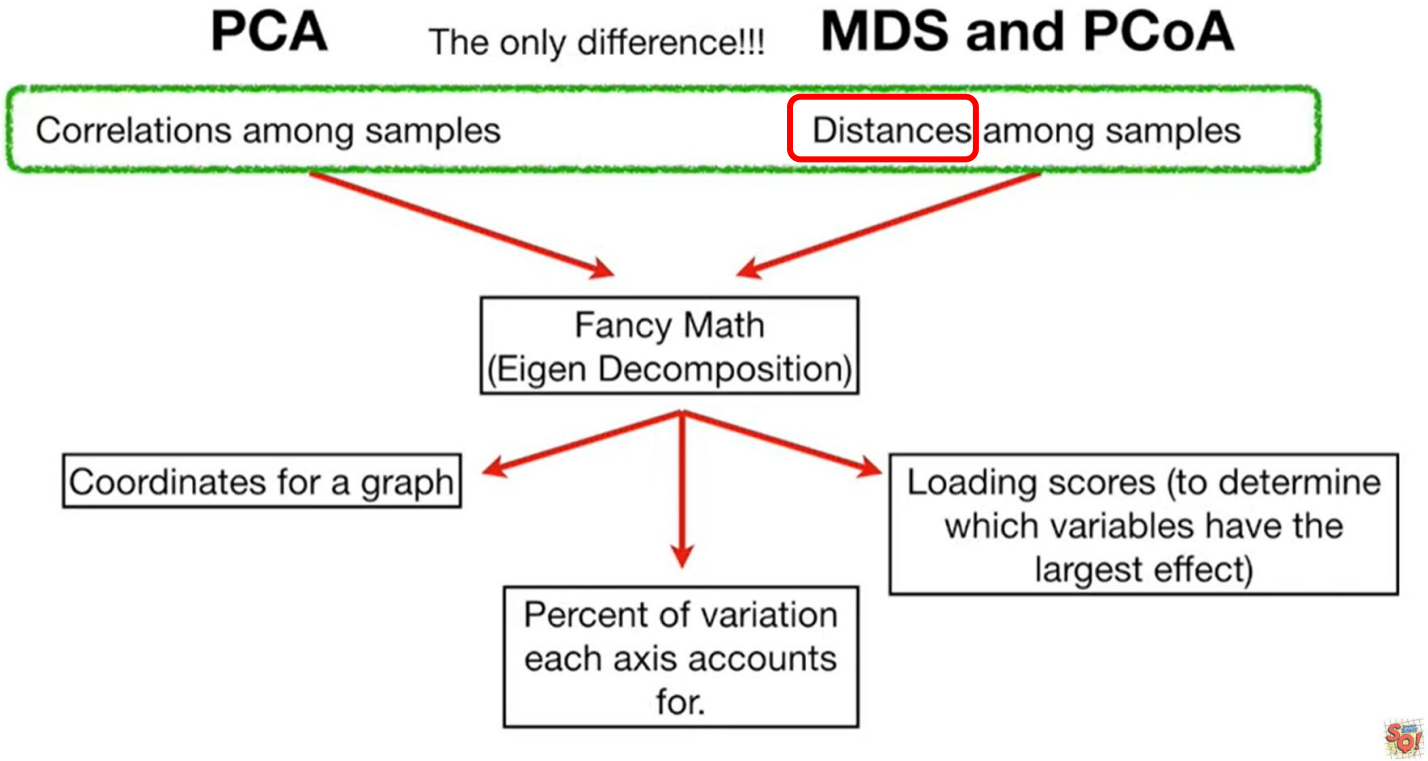




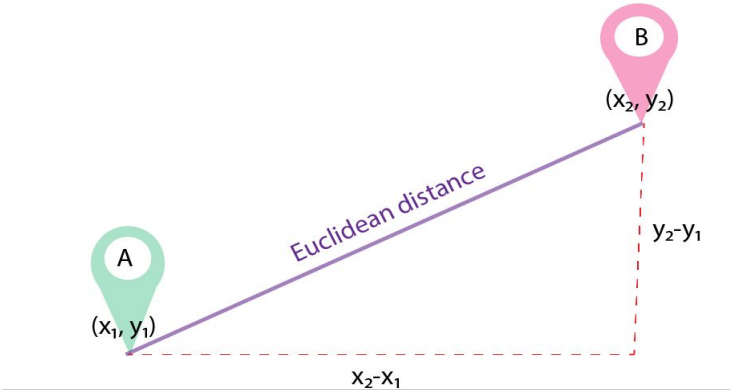


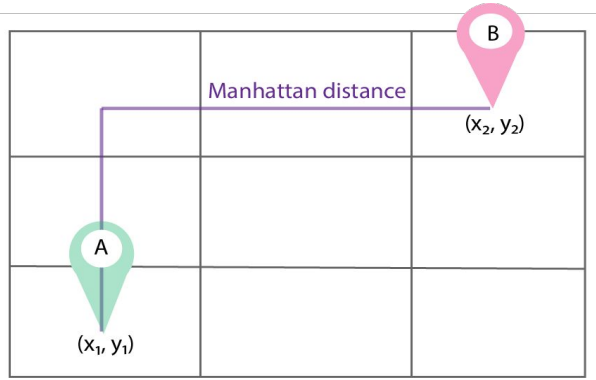
Data in feature space → Find principal components → Data in **p**rincipal **c**omponents space

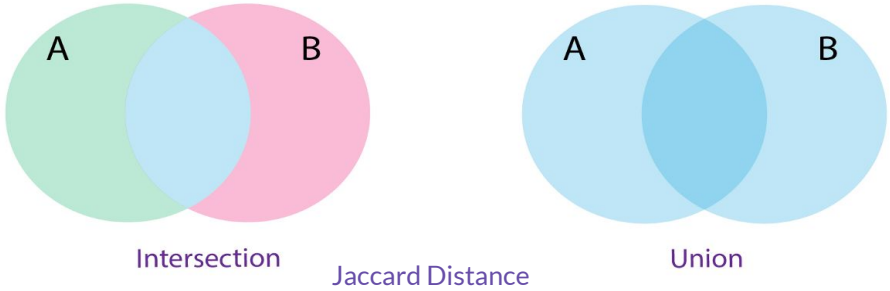
PCoA vs PCA

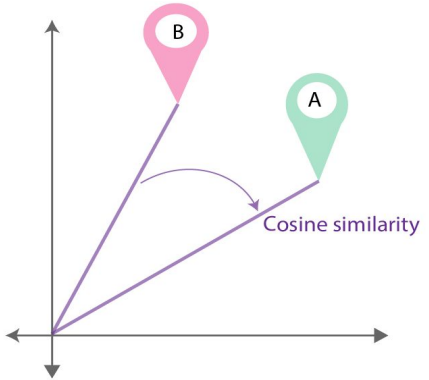

PCA
The only difference!!!
MDS and PCoA
Correlations among samples
Distances among samples
Fancy Math
(Eigen Decomposition)
Coordinates for a graph
Loading scores (to determine which variables have the largest effect)
Percent of variation each axis accounts for.

Distance/ Proximity Measures





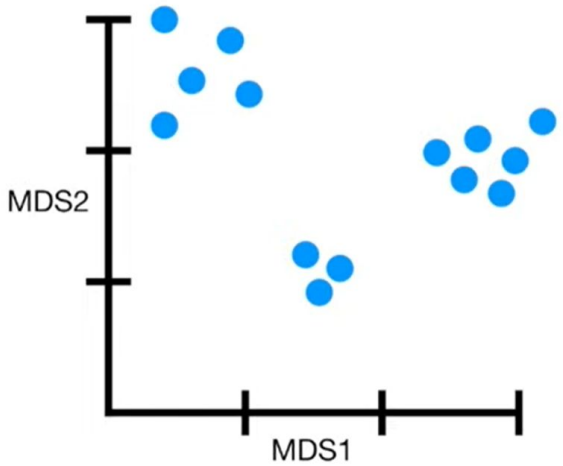


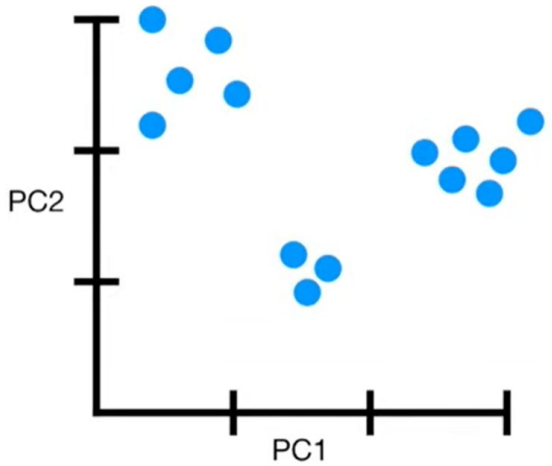


PCoA vs PCA

IF we use Euclidean Distance in PCoA, the graph would be similar to a PCA graph

In other words, clustering based on **minimizing the linear distances is the same maximizing the linear correlations.**





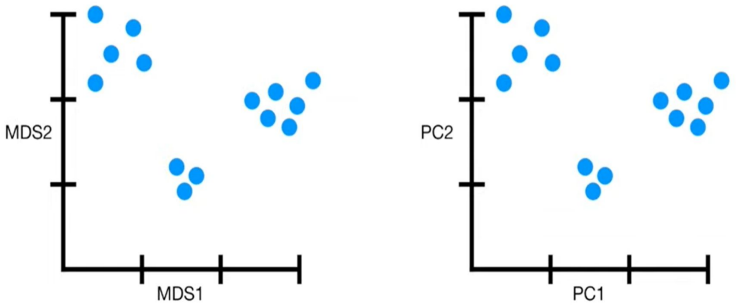
importance(PCoA)

As with other ordination techniques such as PCA and CA, PCoA produces a set of uncorrelated (orthogonal) axes to summarise the variability in the data set.

While PCoA is suited to handling a wide range of data, information concerning the original variables cannot be recovered.

How do I interpret a PCA/PCoA plot?

Interpreting the plots



1. There is Principal Component/Coordinate for each dimensions
    a. If we have "n" variables, we would have "n" Principal Components/Coordinates
2. PC1/PCoA1 would span the direction of most variation

    PC2/PCoA2 would span in the direction of 2nd most variation

    .

    .

    .

    PC"n"/PCoA"n" would span in the direction of "n"th most variation
3. Each axis has an eigenvalue whose magnitude indicates the amount of variation captured in that axis

Name(ID)	**Age**		**Height**		**Gender** (1=f, 2=m, 3=other)	**Education Level** (0=Bachelor, 1= Master, 2= Post Doc)	**Class Label : Teacher(1) or Student(0)**
Robert	30	1	6.1	3/5	m(2)	Post Doc(2)	Teacher(1)
Julian	26	3/7	6.3	1	m(2)	Master(1)	Student(0)
Danial	25	2/7	5.8	0	m(2)	Master(1)	Student(0)
Max	26	3/7	5.9	1/5	m(2)	Master(1)	Student(0)
Faizan	23	0	6.0	2/5	m(2)	Master(1)	Student(0)
Abdullah	27	4/7	5.8	0	m(2)	Master(1)	Student(0)
Ammar	26	3/7	5.9	1/5	m(2)	Master(1)	Student(0)
Rahul	25	2/7	5.8	0	m(2)	Master(1)	Student(0)
Mean	**26**	3/7	**5.95**	0.3	**2**	**1.125**	

Questions?

NMDS (Non-metric Multidimensional Scaling)

- Fundamentally different than PCA, CA (and DCA)
- Ordination based on **ranks** rather than **distance**
    - Axes determined by rearranging points in an iterative way to find the “best” ordered solution
- Avoids assumption of linear relationships among variables

if not distance then..

How to place objects initially?

- Random Placement
- Placement according to a PCA result
- Placement according to geographic distances
- Placement by moving from high to low dimensionality

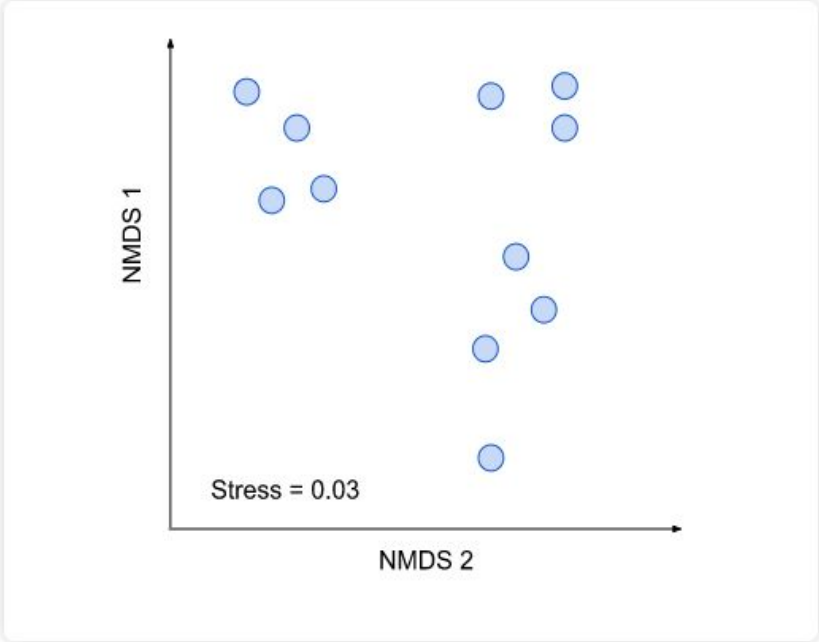


Figure 1: Schematic of a non-metric multidimensional scaling plot. Points represent objects. Objects that are more similar to one another are ordinated closer together. The axes are arbitrary as is the orientation of the plot. Stress values should always accompany an NMDS ordination.