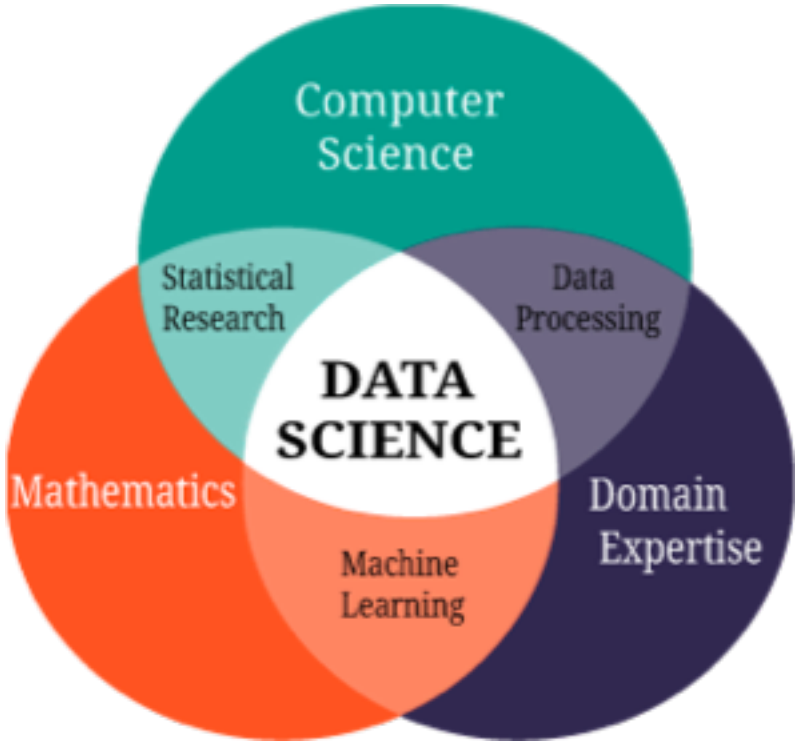
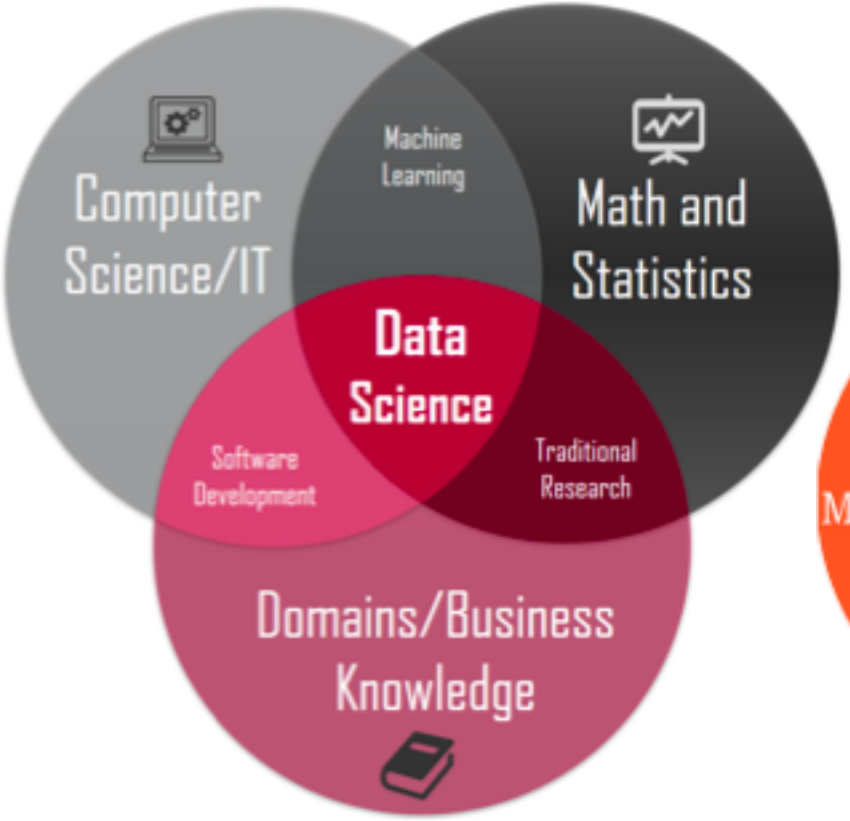
Basic Data Analysis

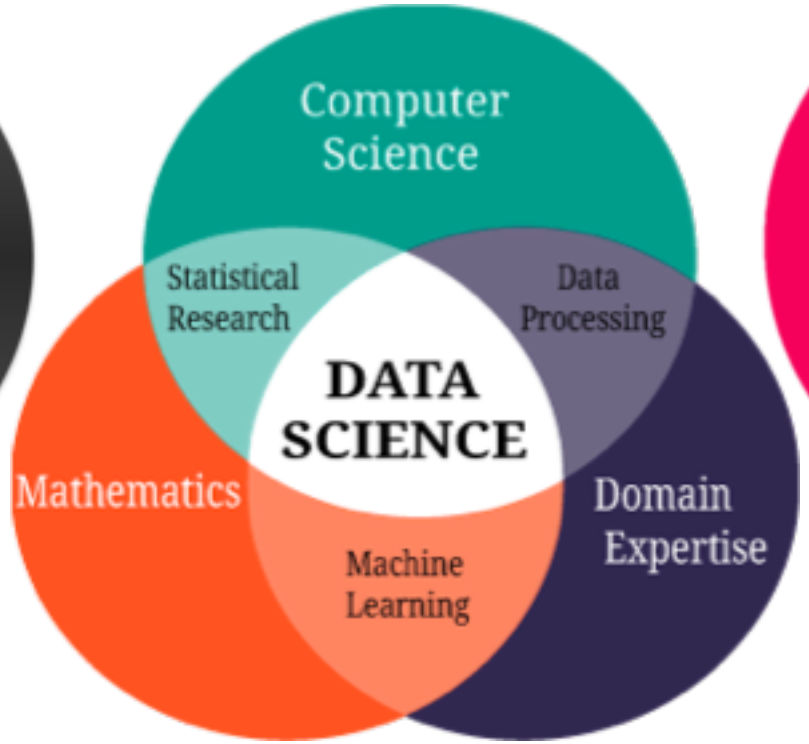
(or as some people like to call it ‘Data Science’…)

What is ‘Data Science’?



What is ‘Data Science’?

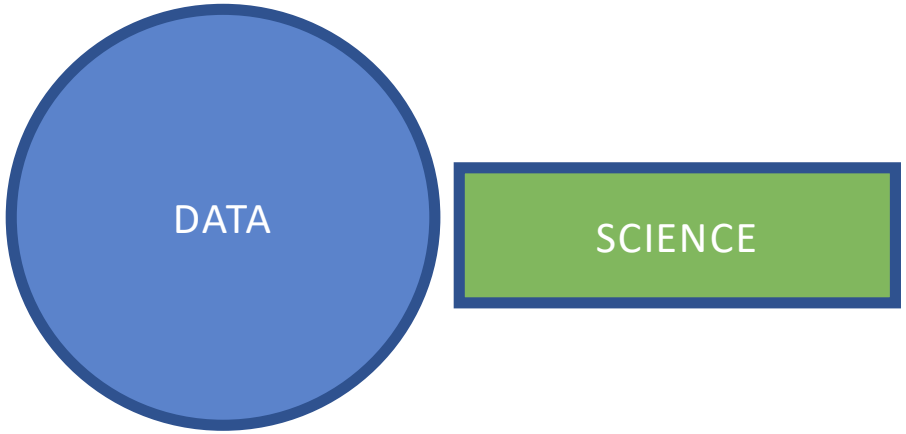




What is ‘Data Science’?

SCIENCE

What is ‘Data Science’?



What is ‘Data Science’?

What are the typical questions when we get a lot of data?

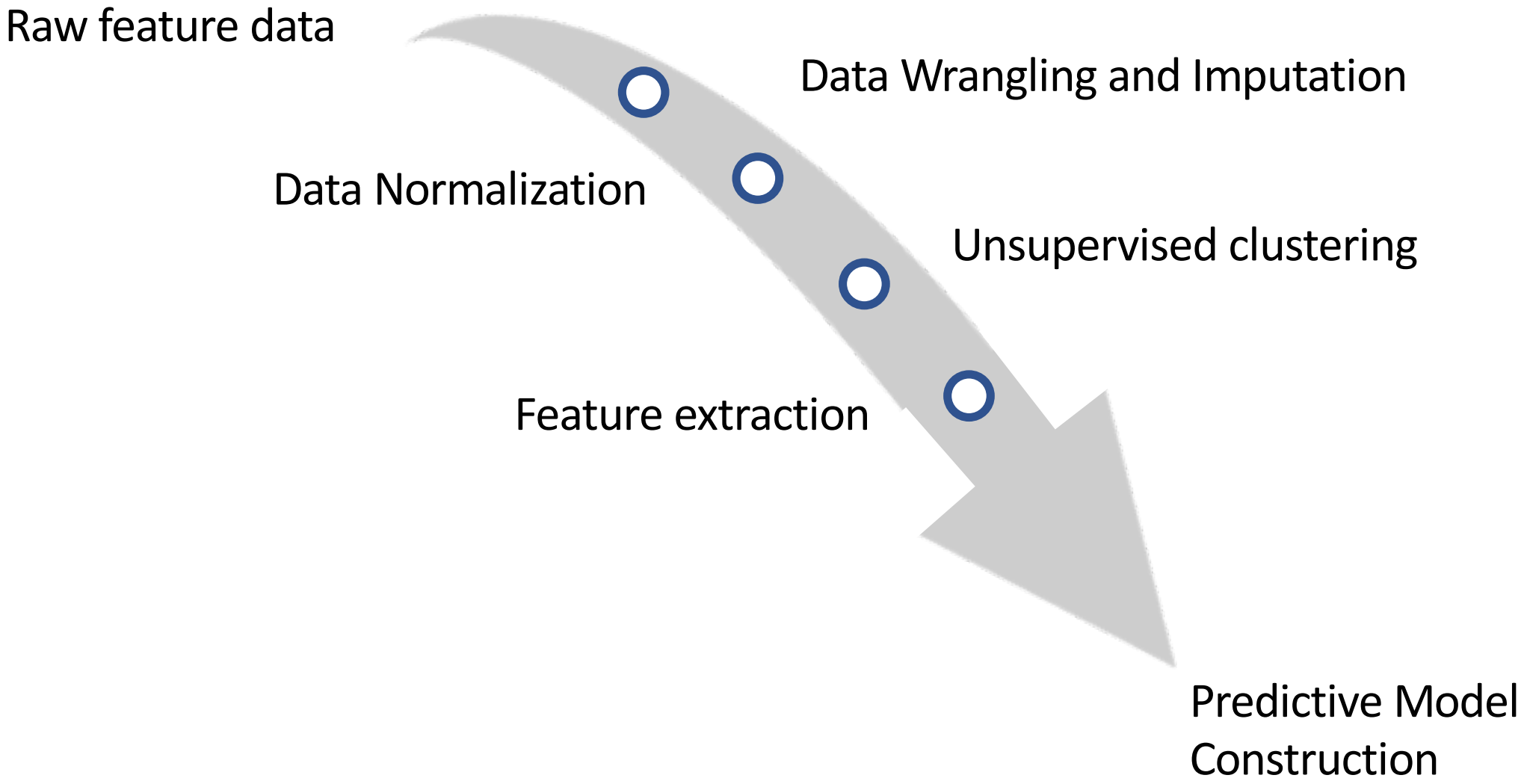
- Are there unknown sample groups defined by the data?
  - unsupervised analysis (i.e. hierarchical clustering)

- What are the important features that associate with the data pattern ?
  - feature extraction (i.e. generalized linear regression)

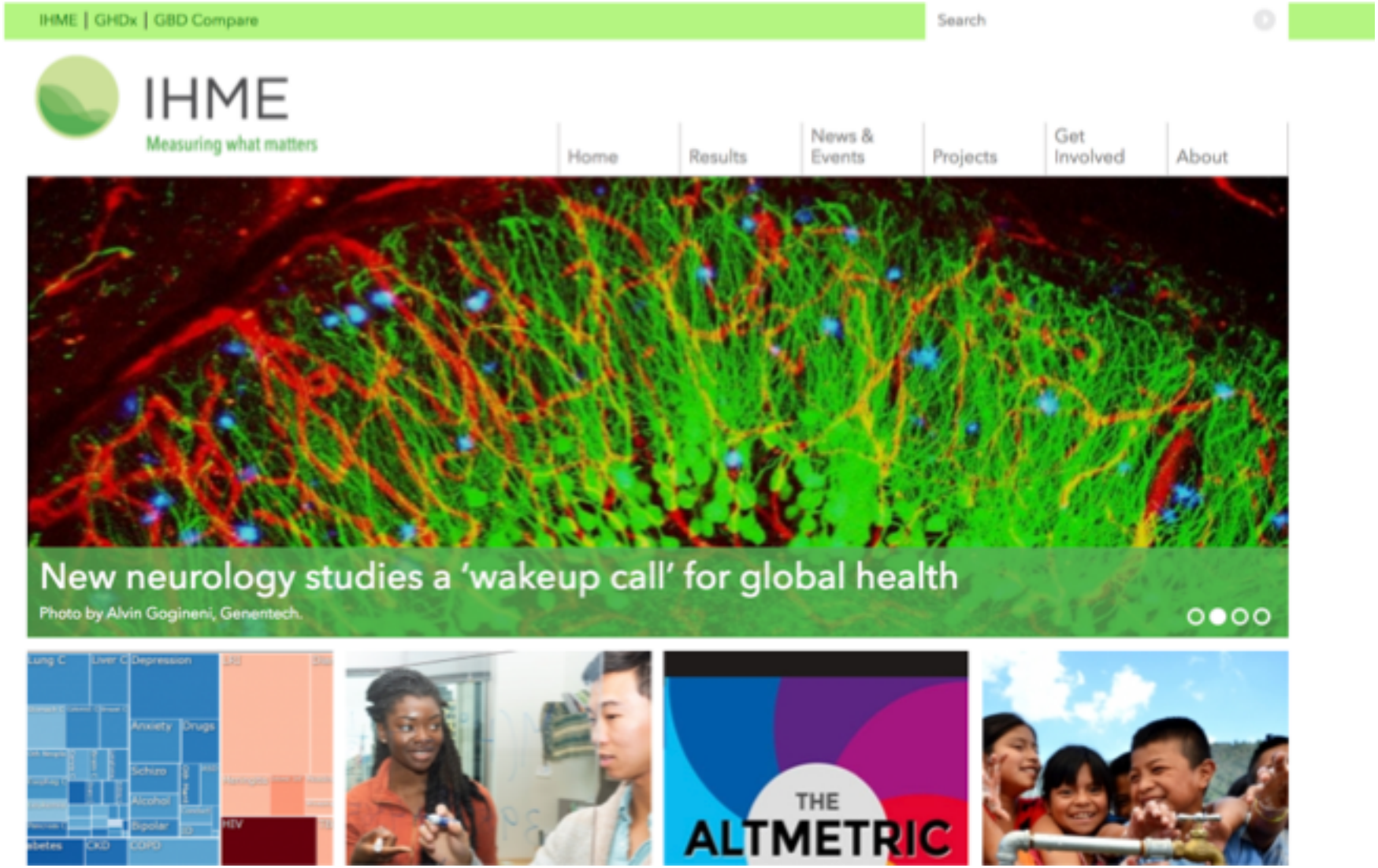
- Can the selected features be used in a predictive model ?

Generalized Data Analysis Workflow

When you know very little about the data at hand



The Institute of Health Matrix and Evaluation



http://www.healthdata.org/

The Global Health Data Exchange (GHDx)

IHME | GHDx | GBD Compare

Search Login

GHDx

Global Health Data Exchange

Discover the World's Health Data

Home Countries Series and Systems Organizations Keywords IHME Data About the GHDx Help

Home > IHME Data

GBD Results Tool

Default results are deaths and DALYs for 2017 with trends since 1990. Refer to the **GBD Results Tool User Guide** for help with common questions and troubleshooting. **Download additional GBD 2017 results** from the GHDx.

Terms defined | **Codebook** | **Tools Overview**

Base	Single / Change / PoD	**Context**	Cause	**Measure**	Add/Remove... (2)
Location	Add/Remove... (1)	**Age**	Add/Remove... (1)	**Sex**	Add/Remove... (1)
Year	Add/Remove... (1)	**Metric**	Add/Remove... (3)	**Cause**	Add/Remove... (1)

Search | Permalink | Download CSV

MEASURE	LOCATION	SEX	AGE	CAUSE	METRIC	YEAR	VAL	UPPER	LOWER
Deaths	Global	Both sexes	All Ages	All causes	Number	2017	55,945,729.74	56,516,734.27	55,356,403.54
Deaths	Global	Both sexes	All Ages	All causes	Percent	2017	100.00	100.00	100.00
Deaths	Global	Both sexes	All Ages	All causes	Rate	2017	732.23	739.70	724.52

http://ghdx.healthdata.org/gbd-results-tool

Our Data Set

- Downloaded from GHDx
- 2017 all cause of death data from all countries
- 195 countries and 133 causes of death
- Represented as Percentage of Cause of Death by Country (‘standardization’)

Input Variables

raw_dt Raw data table as read from data files downloaded from Global Health Data Exchange

matrix_dt Table of Percentage of Death with x being cause and y being country

perc_dt Subsetted data table to include only percentage of death by cause

location_code Table of Numeric Code for countries and their region

Analysis Strategy

Load prepared data

- raw data downloaded from GHDx was store in *data/IHME-GBD_2017_DATA* and pre-processed and annotated by *data_preprocess.R*

Visualize Data distribution

- Feature reduction by Principle Component Analysis (PCA)
- Project on a Scatter Plot

unsupervised clustering

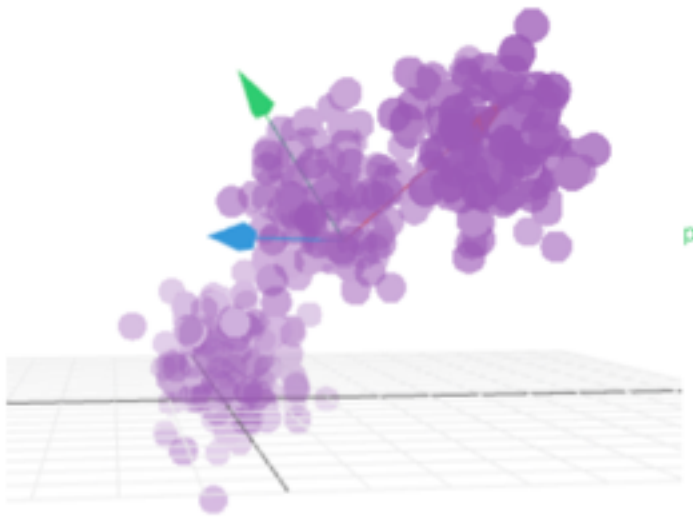
- Perform Hierarchical clustering using Euclidean Distance

Feature Extraction

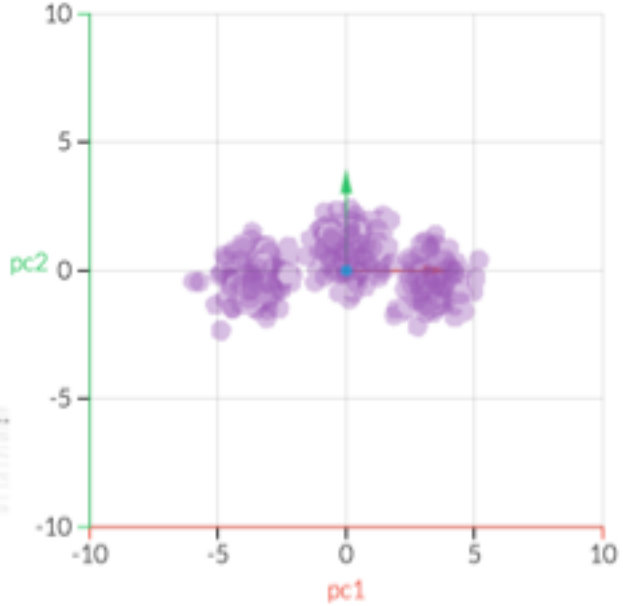
- linear regression to find the most common cause of death globally

Principle Component Analysis (PCA)

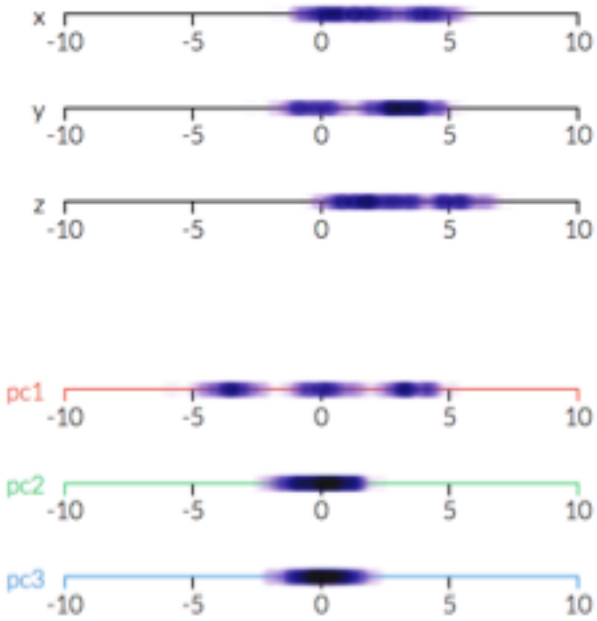
Feature reduction technique that convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (Principle Component or PC's)



3-D



2-D

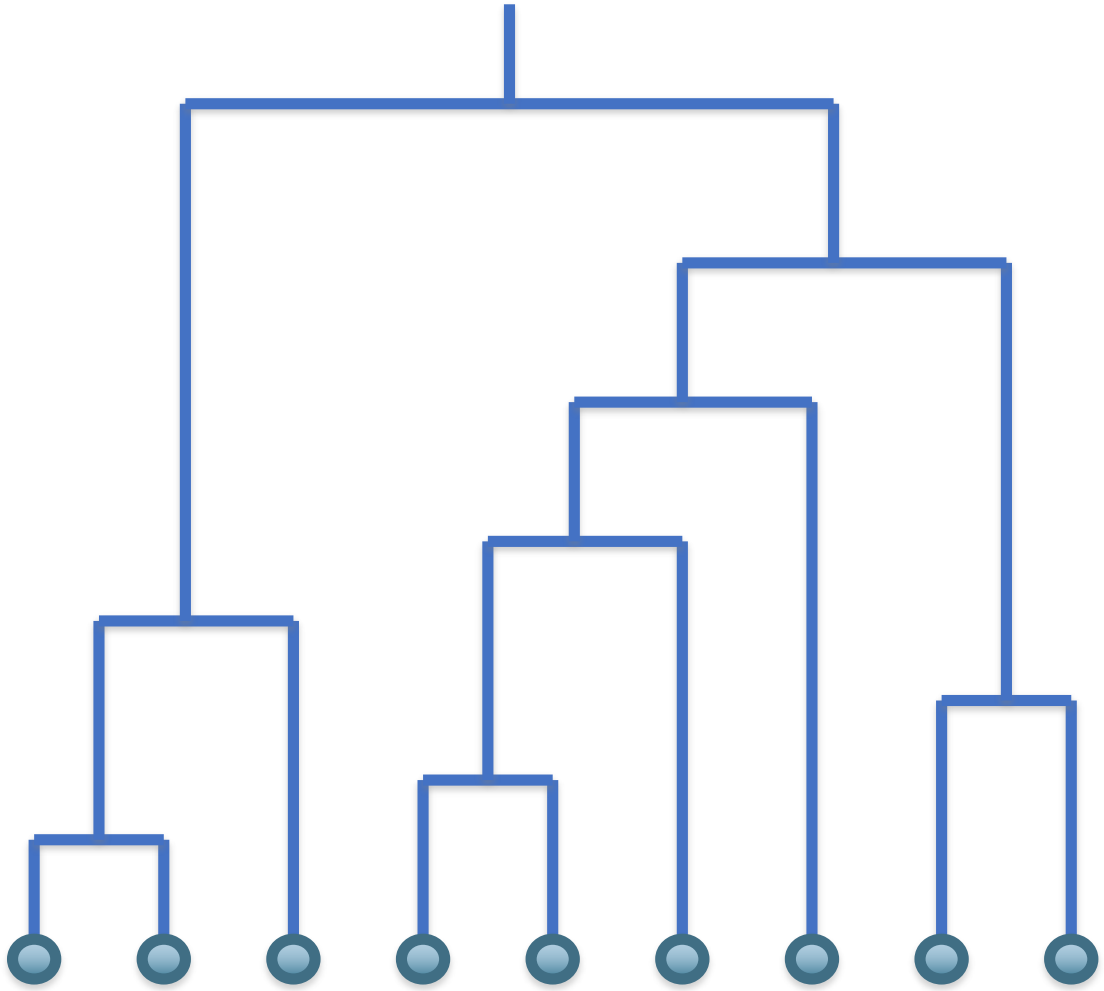


1-D

http://setosa.io/ev/principal-component-analysis/

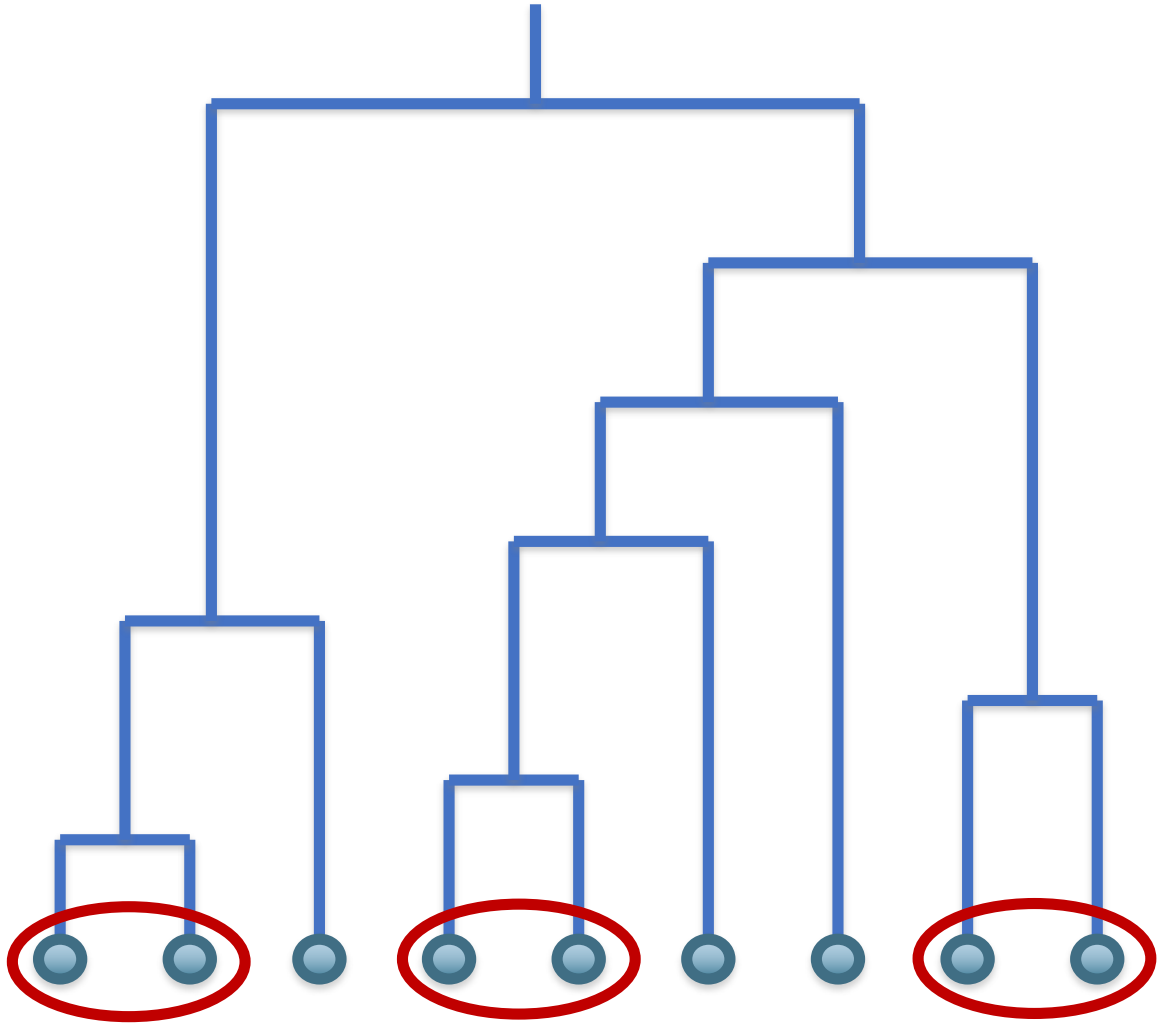
Hierarchical Clustering: Dendrogram

- Determine pairwise distance between all samples with each sample being its own cluster
- Connect closest pair of cluster until there is only one
- Cutting the dendrogram at a desired level to obtain desired number of clusters



Hierarchical Clustering: Dendrogram

- Determine pairwise distance between all samples with each sample being its own cluster
- Connect closest pair of cluster until there is only one
- Cutting the dendrogram at a desired level to obtain desired number of clusters



Hierarchical Clustering: Dendrogram

- Determine pairwise distance between all samples with each sample being its own cluster
- Connect closest pair of cluster until there is only one
- Cutting the dendrogram at a desired level to obtain desired number of clusters