



Lecture 5

Introduction to Practical Statistics

Machine Learning with Python



“AI is the new electricity”

-Andrew Ng

“Data is the new Oil”

-Clive Humby



Statistics Vs Machine Learning

Statistics

- ★ Hypothesis Testing
- ★ Experimental Design
- ★ ANOVA Method
- ★ Linear Regression
- ★ Logistic Regression
- ★ PCA
- ★ Factor Analysis
- ★ Discriminant Analysis

Machine Learning

- ★ Decision Trees
- ★ Neural Networks
- ★ SVM
- ★ Clustering Methods
- ★ Association Rules
- ★ Feature Selection
- ★ Visualization
- ★ Dimensionality Reduction
- ★ Graphical Models

Key Statistical Techniques used in Machine Learning

Causality

Relationship between two events -- cause and the other it's effect as its consequence

Correlation

The extent to which two or more variables are related to one another.

The degree ranging from $[-1,1]$

Clustering

Assigning observations or records into separate groups so that objects within each cluster are similar in some manner and objects across clusters are dissimilar.

Dependent Variable

The variable whose value is unknown and to be explained or predicted.

Independent Variable

Predictors or variables that are used to predict a dependent variable.

Factor Analysis

Reduces huge number of variables of interest to study to a handful of composite metrics that capture the essence of hundreds of variables

Hypothesis

A hypothesis for any given problem is a quantifiable, testable and statistical view of the question in hand used to test the validity of a conclusion.

Hypothesis is of two types-

- ★ Null Hypothesis
 - ★ Alternate Hypothesis
-

Hypothesis

A hypothesis for any given problem is a quantifiable, testable and statistical view of the question in hand used to test the validity of a conclusion.

Hypothesis is of two types-

- ★ Null Hypothesis
- ★ Alternate Hypothesis

Null Hypothesis

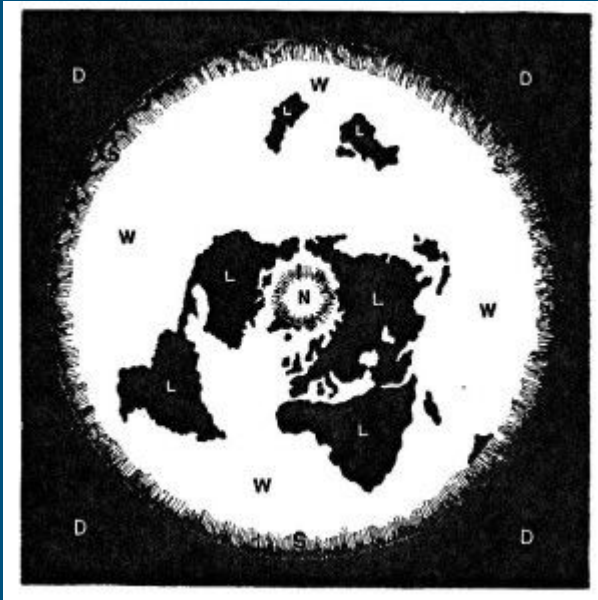
- Researchers work to reject, nullify or disprove the null hypothesis
- It is called null because it is supposed to be nullified

Alternate Hypothesis

- Opposite of Null Hypothesis
- Researchers come up with an alternate hypothesis, one that they think explains the phenomenon and then work to reject the null hypothesis.
- This is done to ensure there are no flaws in your hypothesis.

Example

Earlier, People believed the world was flat.



Null Hypothesis : The world is flat

Alternate Hypothesis : The world is round

Copernicus first disproved the null hypothesis.

And then proved the alternate!

In order to change people's thinking, one has to prove that their thinking is **wrong**.

Other Examples of Null Hypothesis

?

⇒ Newly proposed vaccine offers no protection

Examples of Alternate Hypothesis

?

⇒ Alternate vaccines can cut no. of infected individuals in half

Sample Problem

A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.

Step 1

⇒ Find the alternate hypothesis from the problem.
I expect the average recovery period to be greater than 8.2 weeks.

Step 2

⇒ Convert the hypothesis into math

$$H_1: \mu > 8.2$$

Step 3

⇒ State what will happen if alt hypothesis doesn't come true

$$H_0: \mu \leq 8.2$$

Sample Problem

A researcher is studying the effects of a radical exercise program on knee surgery patients. There is a good chance the therapy will improve recovery time, but there's also the possibility it will make it worse. Average recovery times for knee surgery patients is 8.2 weeks.

Step 1: State what will happen if the experiment doesn't make any difference. That's the null hypothesis—that nothing will happen.

In this experiment, if nothing happens, then the recovery time will stay at 8.2 weeks.

$$H_0: \mu = 8.2$$

Step 2: Figure out the alternate hypothesis. What happens if our experiment makes a difference?

$$H_1: \mu \neq 8.2$$

Exercise

A healthcare provider saw that 48% percent of their members received their flu shot in a recent year. The healthcare provider tried a new advertising strategy in the following year, and they took a sample of members to test if the proportion who received their flu shot had changed.

$$H_0 : p = 48\%$$

$$H_a : p > 48\%$$

(where p is the proportion of members who received the flu shot)

$$H_0 : p = 48\%$$

$$H_a : p \neq 48\%$$

(where p is the proportion of members who received the flu shot)

$$H_0 : \mu = 48\%$$

$$H_a : \mu > 48\%$$

(where μ is the average number of members who received the flu shot)

Hypothesis Testing

Hypothesis testing is about testing to see whether the stated hypothesis is acceptable or not.

Hypothesis Testing Procedure

Set up a Hypothesis



Set up a Suitable Significance Level



Determine a Suitable Test Statistic



Determine the Critical region



Perform Computations



Decision-Making

Hypothesis Testing Procedure



The first step is to establish the hypothesis to be tested. Statistical hypothesis is an Assumption about the value of an unknown parameter. This is done while constructing Null and Alternative Hypothesis about the population

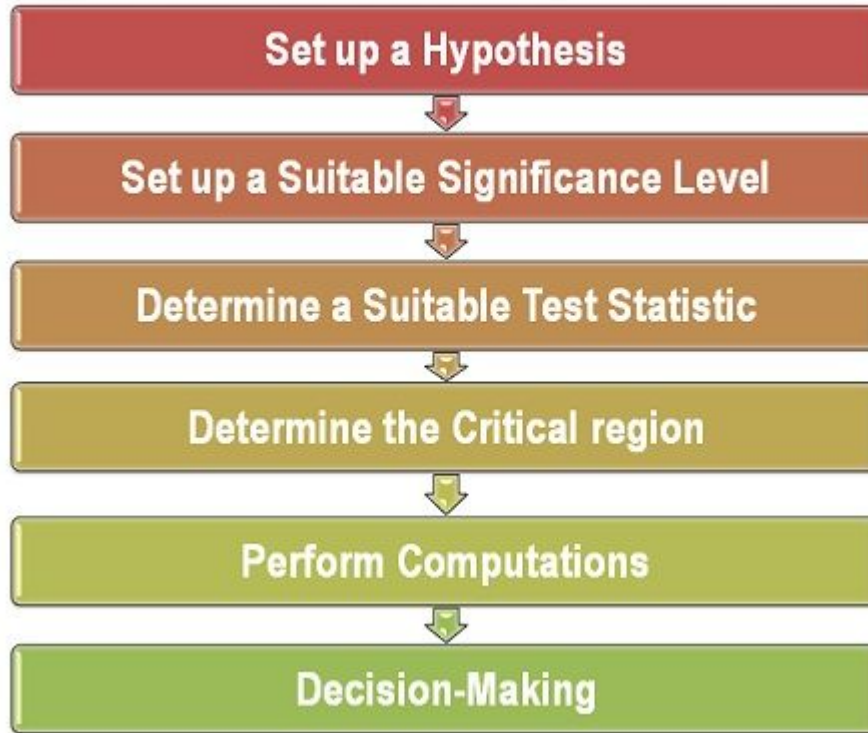
Hypothesis Testing Procedure



Next, Decide the level of significance, i.e. a confidence level with which the null hypothesis is accepted or rejected. The significance level is denoted by ' α '

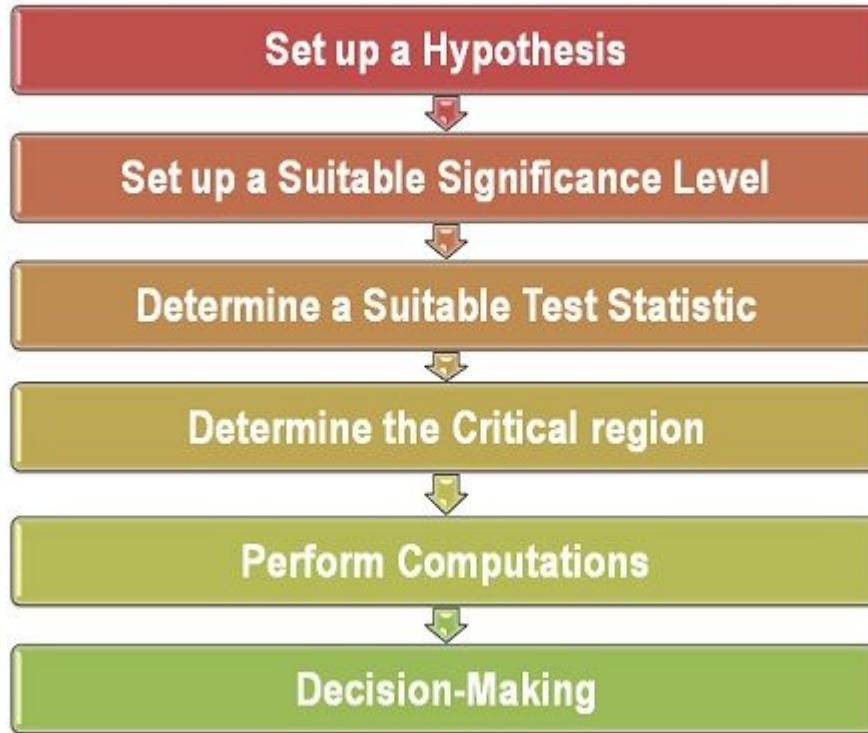
In practice, we either take 5% or 1% level of significance.

Hypothesis Testing Procedure



The next step is to determine a suitable test statistic and its distribution.

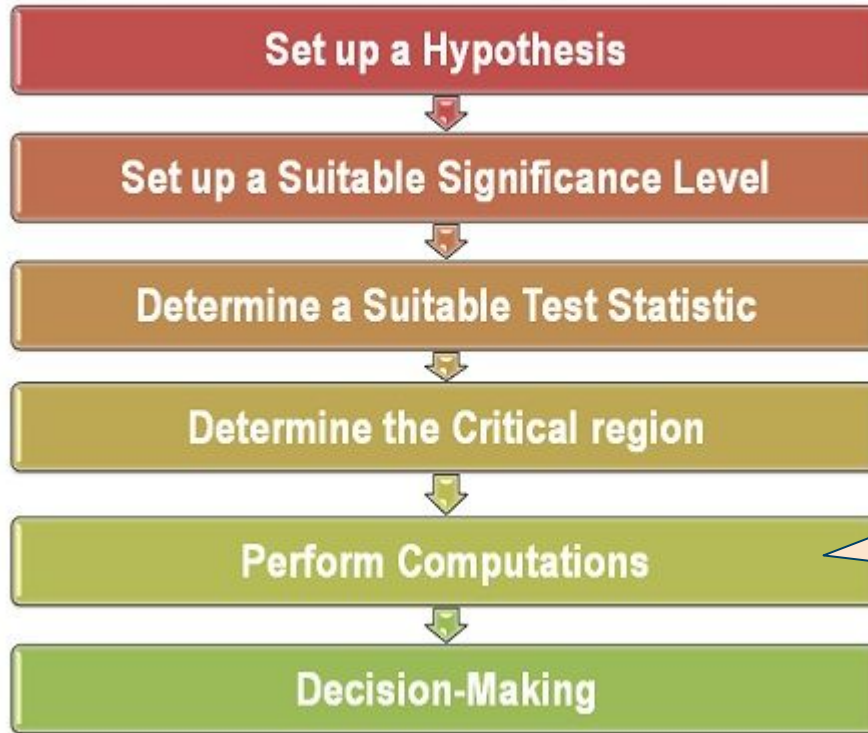
Hypothesis Testing Procedure



Decide which values to the test statistic will lead to the acceptance of H_0 and which will lead to its rejection.

The values that lead to rejection of H_0 is called the critical region.

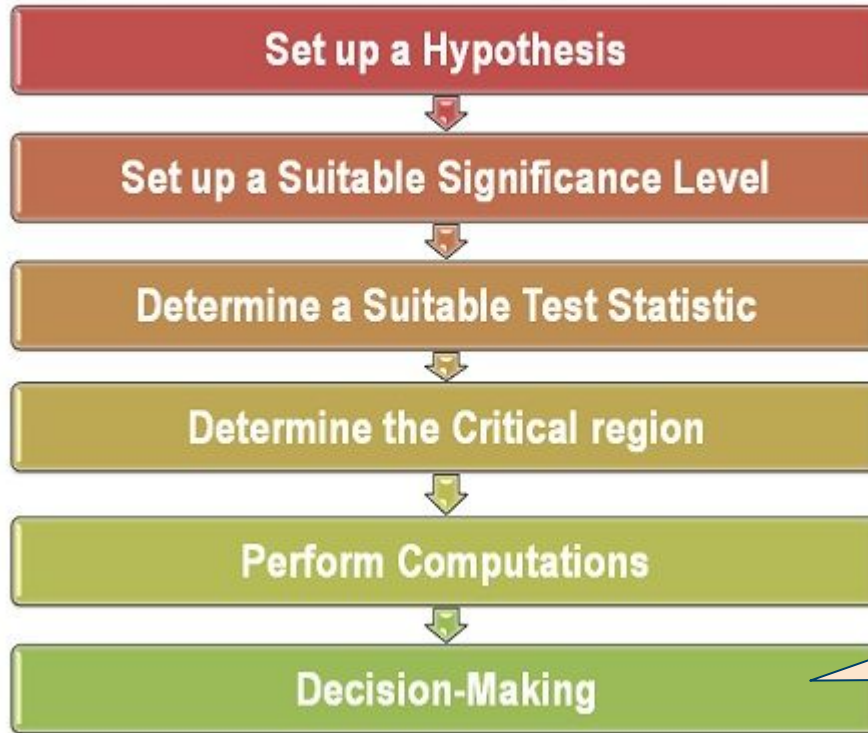
Hypothesis Testing Procedure



Compute several values for the random sample of size 'n.'

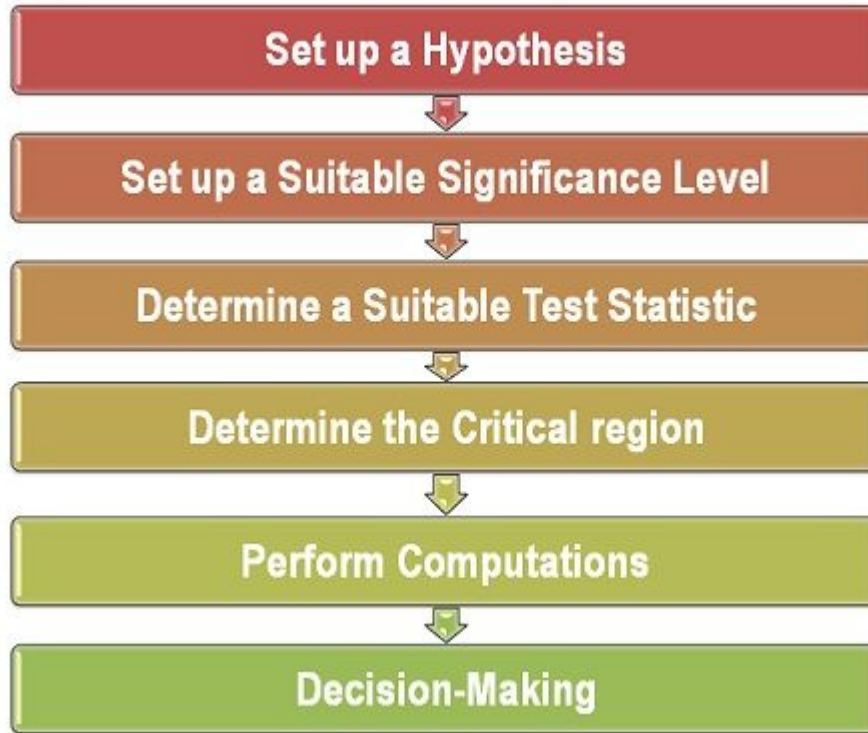
Then apply the formula of the test statistic to check whether the sample results falls in the acceptance region or the rejection region.

Hypothesis Testing Procedure



The statistical conclusions can be drawn. The decision that the null hypothesis is accepted or rejected depends on whether the computed value falls in the acceptance region or the rejection region.

Hypothesis Testing Procedure



It is necessary to follow these steps systematically so that the results obtained are accurate and do not suffer from any statistical error

Type of Error

(Type 1 : false positive)

Type 1 \Rightarrow Probability of falsely rejecting the hypothesis when it is \rightarrow true

Conclusion: Manufacture wastes resources on developing an ineffective drug

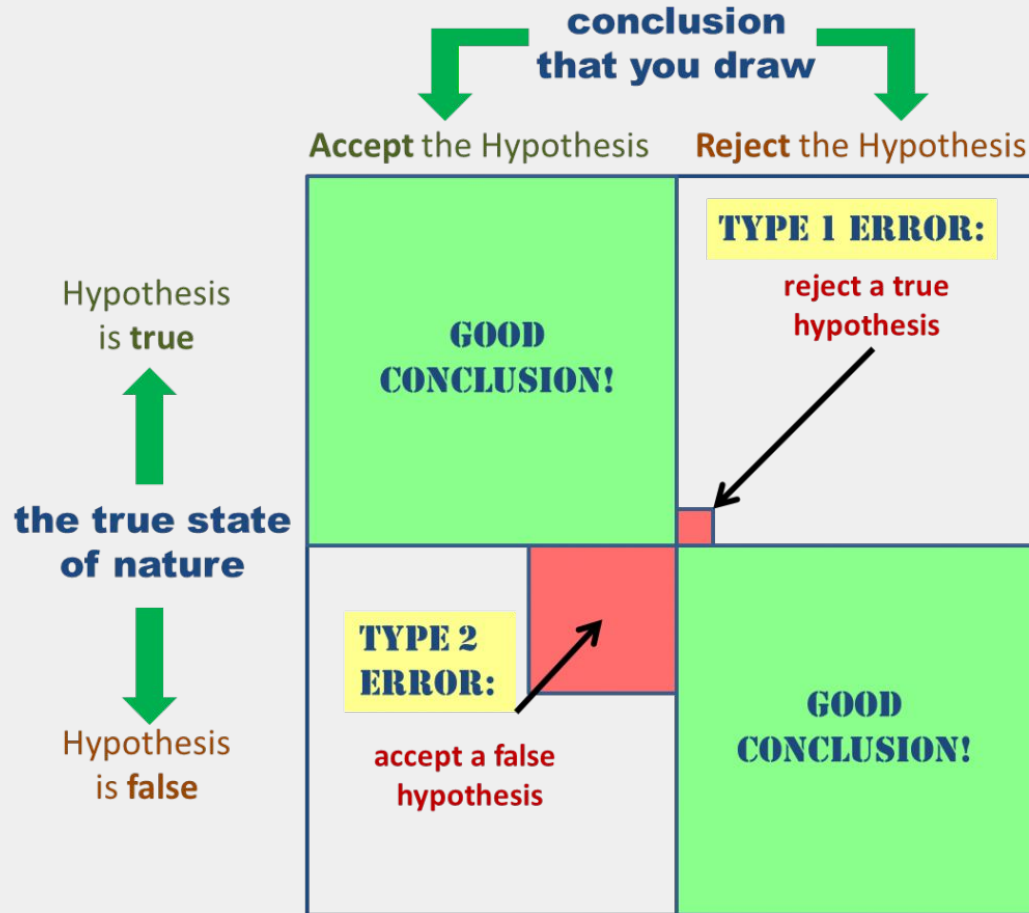
or, manufacturer misses opportunity to build and make profit on an effective drug.

Type of Error

(Type 2 : false negative)

Type 1 \Rightarrow Probability of falsely accepting the null hypothesis when alternate hypothesis is \rightarrow true

e.g., exposing large number of people to a potentially lethal drug as non-harmful. Patients and family suffer and manufacturers are sued.



Significance Level

(alpha)

The significance level indicates the amount of evidence required to accept that an event is unlikely to have arisen by chance and is therefore statistically significant.

Typically,

$$\alpha = 0.05 \text{ (5\%)} \text{ or } 0.01 \text{ (1\%)}$$

p -value

p-value is calculated during Hypothesis Testing

It gives the probability of the occurrence of data under the assumption that H_0 is true.

Small p-value is indication of genuine data thus rejecting the null hypothesis.

if $p < \alpha \Rightarrow H_0 \rightarrow \text{rejected}$

$p \geq \alpha \Rightarrow H_0 \rightarrow \text{Do not Reject Null}$

p -value

A small p-value (≤ 0.05) indicates strong evidence against the null hypothesis so you reject Null.

A large p-value (> 0.05) indicates weak evidence against the null hypothesis so you do not reject Null.

Linear Regression

Using a single independent variable to depict a dependent variable.

R Squared

Coefficient of Determination

A statistical measure of how close the data is to the fitted regression line.

The better this value, the better the model.

Its value lies between $[0,1]$

Data Analysis

Finding consistent patterns or relationships among the variables.

Choosing the right statistical model for the problem.

Testing the hypothesis for significant findings.

Variation

Accounting for the possibility of errors arising from the instrument, observer or applied formulae.

Sample

A sample is any proper subset of a population. A small sample can give a distracted view of population.

Samples are chosen carefully in a random manner to avoid sample bias.

Population

Comprises all the members recorded correctly in your observations.

Resampling

K-fold resampling is one in which we subdivide the data into k roughly equal sized parts, then repeat the modelling process k times leaving one section out each time for validation purposes.

JackKnife

Sub divide data into as many parts as there are observations. Leave one observation to upto half the sample and use the remaining $(n-1)$ observations as training set.

Data Types

- ★ Continuous
- ★ Discrete
- ★ Categorical
- ★ Binary
- ★ Ordinal

Data Structure

- ★ Data Frame
- ★ Time Series Data Record
- ★ Spatial for Mapping
Locations
- ★ Graph for Connections
Network

Central Tendency of Data

- ★ Mean
- ★ Weighted Mean
- ★ Median
- ★ Trimmed Mean
- ★ Robust
- ★ Outlier

Variability of Data

- ★ Variation
- ★ Standard Deviation
- ★ Mean Absolute Deviation
- ★ Median Absolute Deviation
- ★ Range
- ★ Order Statistics
- ★ Percentile
- ★ Interquartile Range (IQR)

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

The standard deviation formula is the square root of the variance.

Mean Absolute Deviation Formula

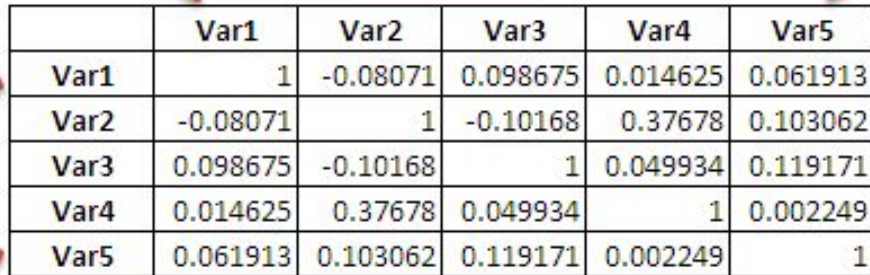
$$\frac{\sum |x - \bar{x}|}{n}$$

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - M_j(x_j)|$$

Correlation Matrix

A Table where variables are shown on both x-axis and y-axis (row & columns) and the cell values are correlations between the variables.

5 X 5 Matrix



	Var1	Var2	Var3	Var4	Var5
Var1	1	-0.08071	0.098675	0.014625	0.061913
Var2	-0.08071	1	-0.10168	0.37678	0.103062
Var3	0.098675	-0.10168	1	0.049934	0.119171
Var4	0.014625	0.37678	0.049934	1	0.002249
Var5	0.061913	0.103062	0.119171	0.002249	1

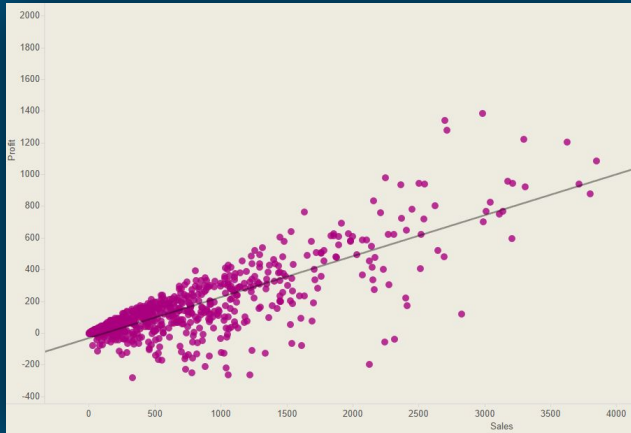
Pearson Correlation Coefficient (r)

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

A correlation coefficient quantifies the relationship of two or more variables between [-1,1].

In linear correlation the coefficient quantifies the strength and direction of the correlation between the variables.

Positive Correlation

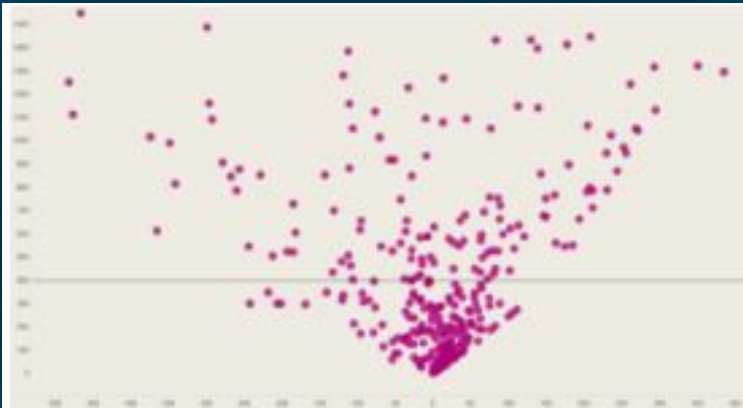


$r=1$ is total positive correlation

Variable having strong positive correlation are nearer to 1.

I.e., as x increases, y increases

No Correlation



0 is no correlation

Variable having no correlation or weak correlation are less than 0.5.

$r \sim 0$ suggest independent variables.

Negative Correlation



$r = -1$ is negatively correlated with each other

Variable negatively correlated are nearer to -1 .

$r \sim 0$ suggest independent variables.

Scatterplot



It is used to visualize the relationship between two measured data variables representing x-axis with one variable and y-axis with another.

Ideal for moderately sized data points (~5k)

Contingency Table

A tally of counts between 2 or more categorical variable e.g., Grade A-G of Loans vs Outcome of Loan Payment.

Example of Contingency Table

A simple 2 x 2 Contingency Table

Groups	Dogs	Cats	Total
Males	42	10	52
Females	9	39	48
Total	51	49	100

Marginal Totals

Variables involved in the experiment and tabulated in contingency table are called marginal totals.

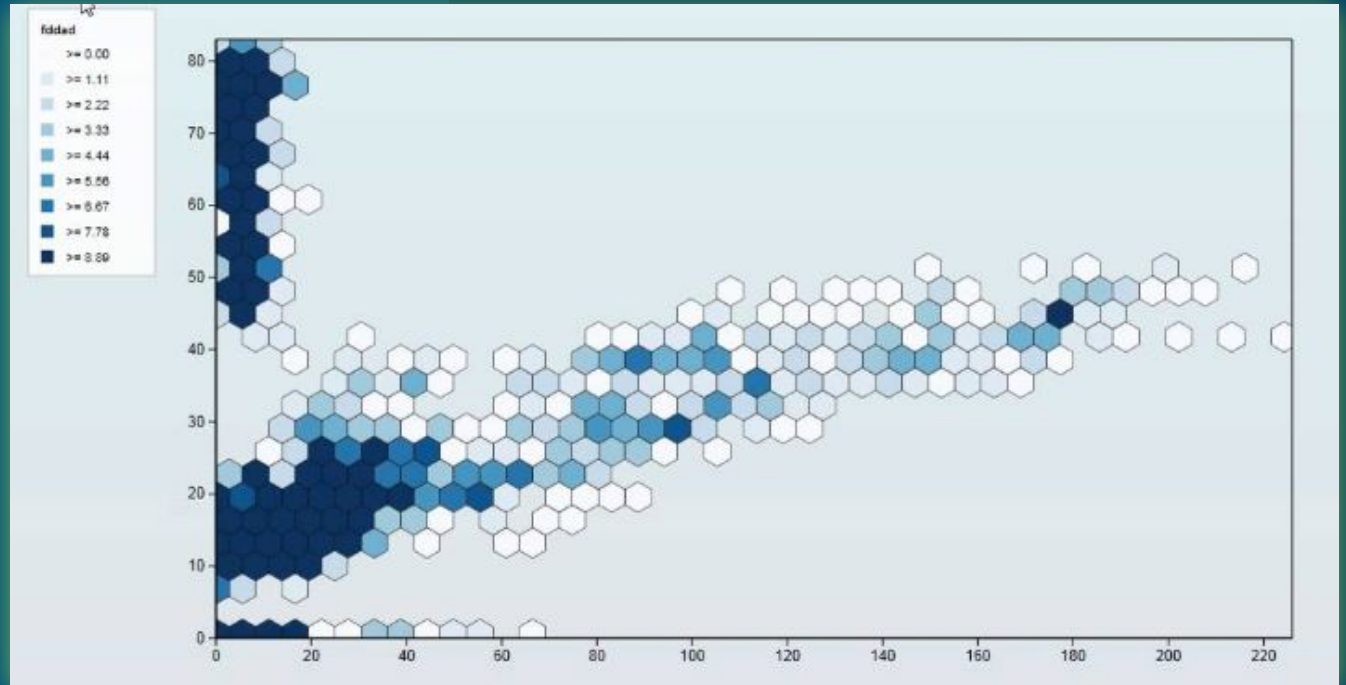
Data Distributions

- ★ Boxplot and Whiskers
(based on IQR)
- ★ Frequency table (bins)
- ★ Histograms
- ★ Density Estimates
- ★ Barchart
- ★ Pie Chart (frequency
plotted as wedges)

Hexagonal Binning

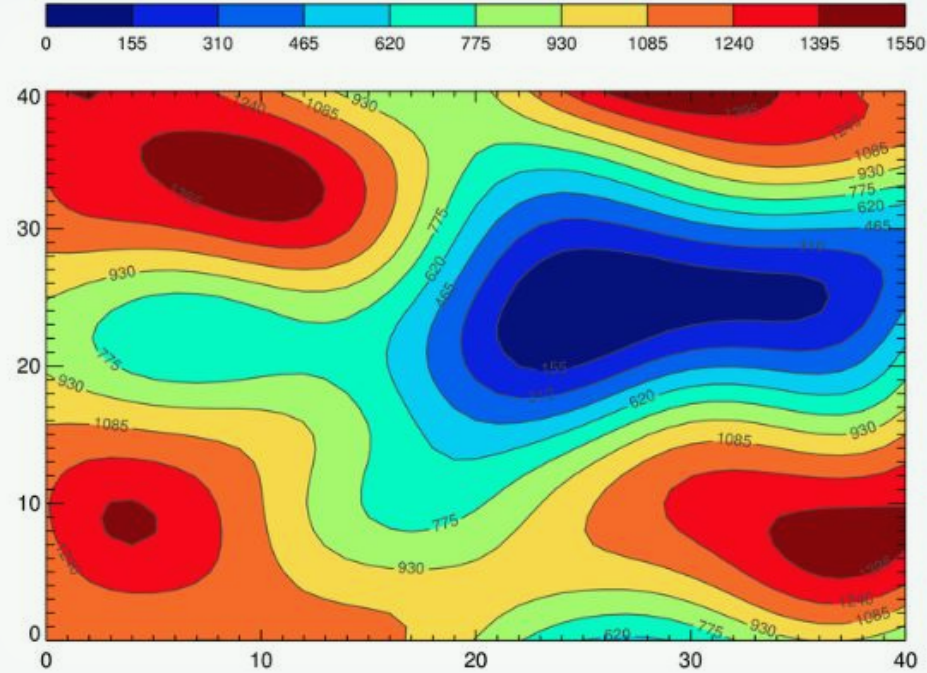
Plot of two numeric variables with records binned into hexagons making denser area darker.

Hexagonal Binning



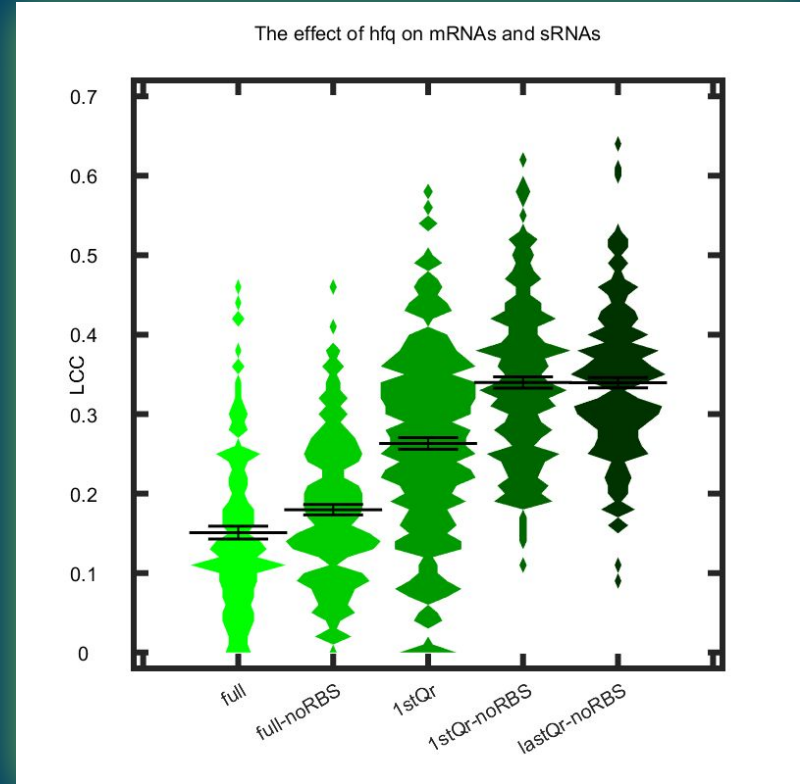
Contour Plot

A plot showing density of two numeric variables on a heat map.

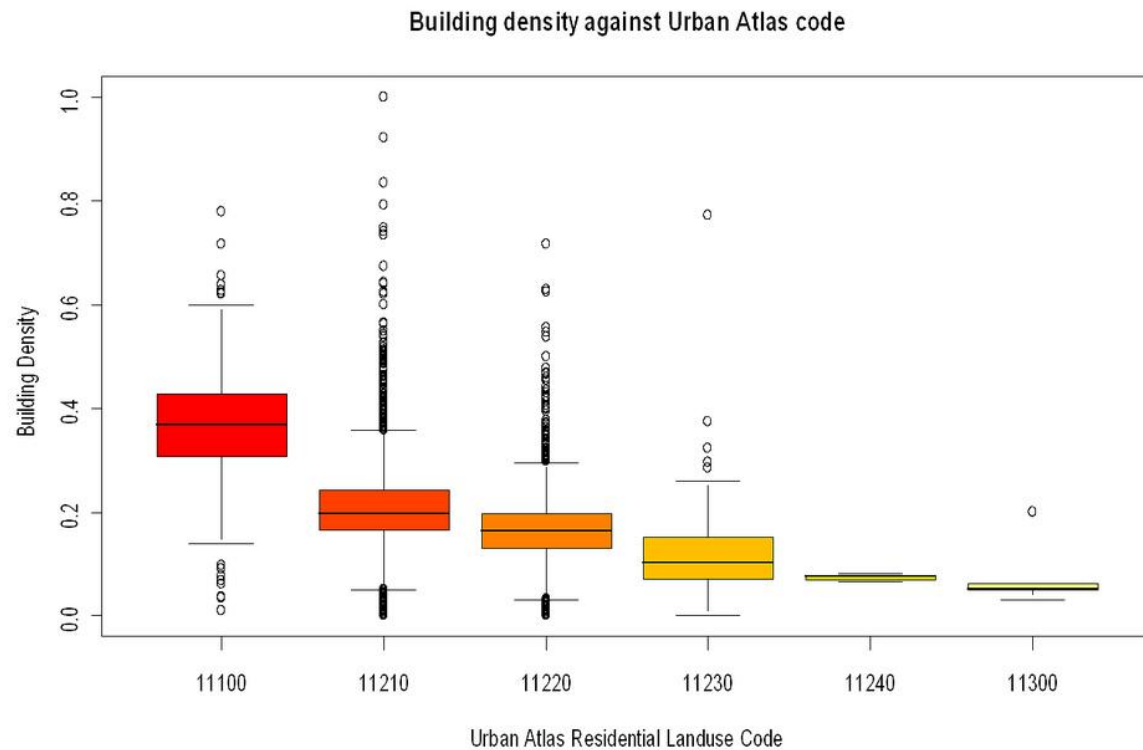


Violin Plot

Showing density estimates



Box Plot



Common Terminologies

Sample Size :	n
Size of Population:	N
Standard Dev:	σ
Variance :	σ^2
Standard Error:	<u>σ/\sqrt{n}</u>

Bootstrap Sample

Sample taken with replacement
from an observed data set.

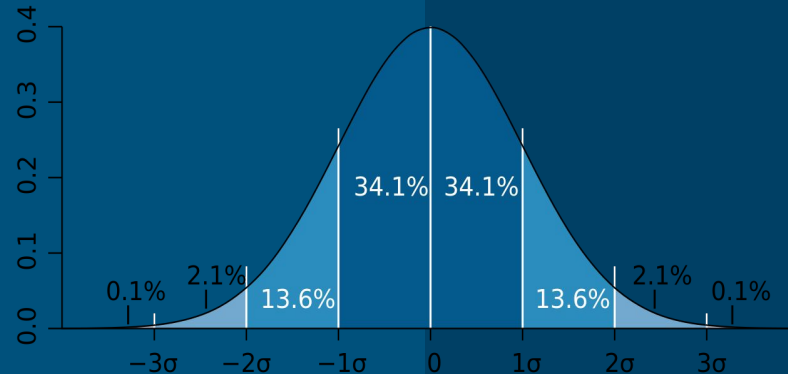
Confidence Interval

Given a sampling procedure and population, what is the probability that sample distribution will occur within a fixed range everytime.

Say, 90% confidence interval means it will enclose the true mean 90% of the time when constructed repeatedly in the same fashion with the same population.

QQ Plot

A plot to visualize how close a sample distribution is to the normal distribution?



Degrees of Freedom

The number of independent values or quantities which can be assigned to a statistical distribution.

Student's t Distribution

A method of hypothesis testing
when the test statistic follows a
normal distribution

Lambda

Rate at which events occur per
unit of time and space.

Poisson Distribution

Frequency Distribution of number
of events in sampled units of time
or space.

Exponential Distribution

Frequency Distribution of time or
distance from one event to the
next event.

Parametric Test Vs Non Parametric Test

- ★ Parametric
 - One which has complete information about the population parameter.
 - Mean is used to calculate central tendency
- ★ Non-Parametric
 - When Researcher has no clue of the population parameter
 - Median is used.

ANOVA Test

Analysis of Variance

STEPS:

1. Combine all the data together in a single box.
2. Shuffle and draw out x resamples (no. of groups) of five values each (no. Of record in each group in original data).
3. Record the mean of each of the following groups.
4. Record the sampled variance among the four group means
5. Repeat steps 2-4 R times (iterations).
6. P-value is the proportion of observed variance to sample variance.

One way test

A hypothesis test that counts in only one direction.

E.g., A is better than B

Two way test

A hypothesis test that counts
p-value in two directions.

E.g. for, $A > B$

&

for, $A < B$

Chi Square

(goodness of fit)

A Statistical test that checks the goodness of fit or discrepancy between the expected theoretical values and observed values.

Wilcoxon test

Used for testing unequal variances among multiple groups.

e.g., if patients have been assigned at random to various treatment groups significant difference in the treatment effect can be gauged if there is a variation in any parameter of the treatment.

Overfitting

Over fitting the model with too many noisy parameters, so that the more models you run, the greater the probability that something will emerge just by chance.

Outliers

(Anomaly Detection)

An outlier is an object that deviates significantly from the general average within a dataset. I

It is numerically distant from the rest of the data and therefore indicates that something unusual is going on that requires additional analysis.

ANOVA

Analysis of Variance

A statistical test of whether
the means of more than two
groups are all equal

ANOVA Test

Analysis of Variance

A single hypothesis test of the overall variance among multiple groups.

Fisher's Exact Test

(No sampling)

An exhaustive test that has a procedure to enumerate all possible arrangements or permutations that can occur.

Tabulate their frequencies and determine exactly how extreme the observed results are from expected output.

Points to Remember



Know Your
Objectives



Know the origin
of your Data



Assign probabilities
to check validity of
predictions



Know Your
assumptions
before starting
analysis

Machine Learning

The field of computer science related to the development and use of algorithms to enable machines to learn from what they are doing and become better over time.

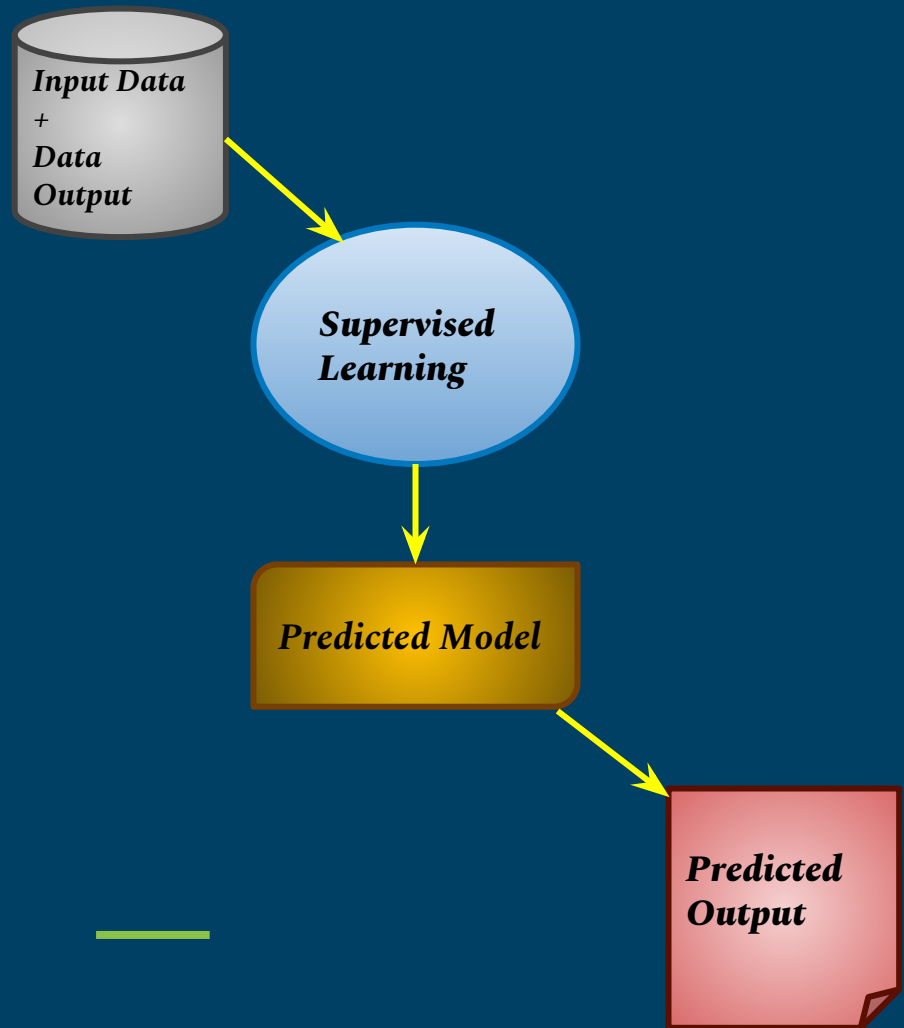
Machine Learning Techniques

→ Supervised Learning

→ Unsupervised Learning

—

Supervised Learning



Supervised Learning

Finding a Machine Learning Function $f(x)$ from labeled data to decipher the relationship between independent variables and dependent variables.

Supervised Learning

Labelled data is one that has both independent variable(s) and dependent variable(s).

Applications of Supervised Learning

- Classification.
- Regression

Pattern Recognition

(Supervised learning)

Identifying patterns in data via algorithms to make predictions of new data coming from the same source.

Supervised Learning

Training vs Test

Cat1	Cat2	Cat3	Cat4	Target
------	------	------	------	--------

x1	x2	x3	x4	y1
----	----	----	----	----

x1	x2	x3	x4	y2
----	----	----	----	----

x1	x2	x3	x4	??
----	----	----	----	----

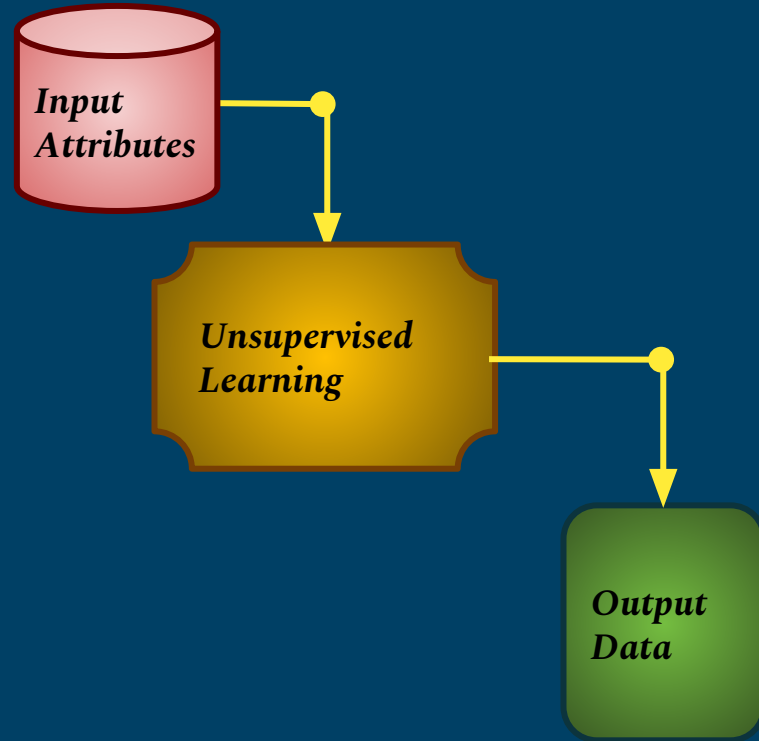
x1	x2	x3	x4	??
----	----	----	----	----

Examples of Supervised Learning

- Images with image labels.
- Predicting time of the day based on given temperature using prior knowledge from weather forecast datalog.

Unsupervised Learning

(No Training)



Unsupervised Learning

Deriving ML inferences or insights by exploring the data itself.

Unsupervised Learning

The target variable (labelled class) is not present in unsupervised learning dataset.

No use of prior knowledge.

No relationships to define.

Applications of Unsupervised Learning

- Dimensionality Reduction Techniques.
- Clustering
- Association Analysis.

Examples of Supervised Learning

- Watching a cricket match for first time without any prior knowledge of the game.
- Reducing the number of questions in an IQ Test to predict the IQ of a person.

Evaluating Model Performance

$$\text{Accuracy} = \frac{\text{(No. of correctly predicted cases)}}{\text{(Total number of predictions)}}$$

- ❑ Ideal Accuracy = 1
- ❑ Accuracy = 0.5 ? ?

Recommender Systems

An algorithm that analyzes a user's purchases and actions on an e-commerce site and uses that data to recommend complementary products.

Regression Analysis

A statistical technique for defining the dependency between continuous variables. It assumes a one way causal effect from a variable to the response of another.

Predictive Analysis

The most valuable analysis as it helps predict what someone is likely to do or how someone will behave in the near future. It is used to identify both risks and opportunities.

Predictive Modeling

The process of developing a model to predict a trend or outcome.

Query Analysis

The process of analyzing a search query for the purpose of optimizing it for the best possible results.

Routing Analysis

Using many different variables to find the optimal route for a certain means of transport in order to decrease fuel costs and increase efficiency.

Real Time Data

Data that is created, processed, stored, analyzed and visualized within milliseconds of its creation.

Scalable Machine Learning Systems

The ability of a system to maintain acceptable level of performance as work-load or scope increases.

Text Analysis

The application of statistical linguistic and machine learning techniques on text based sources to derive meaning or insight.

Time Series Analysis

The process of analyzing well defined data obtained through repeated measurements of time. The data has to be well defined and measured at successive points in time spaced at identical time intervals.

Topological Data Analysis

Focusing on the shape of complex data and identifying clusters and any statistical significant that is present within that data.

Visualization

Visualizations are complex graphs that can include many variables of data while still remaining understandable and readable.

