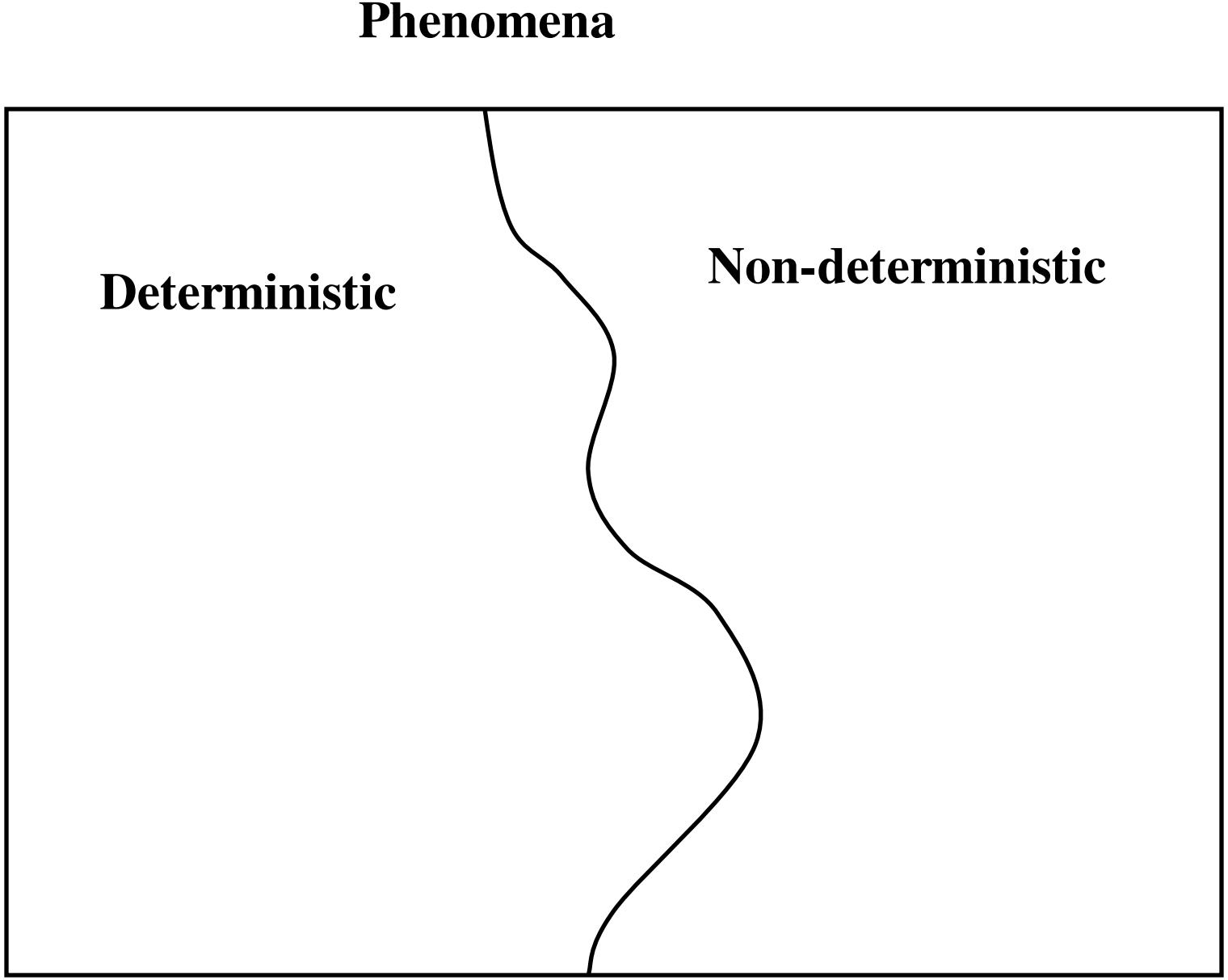
Probability Theory

Probability – Models for random phenomena


Phenomena
Deterministic
Non-deterministic

Deterministic Phenomena

- There exists a mathematical model that allows "***perfect***" prediction the phenomena's outcome.
- Many examples exist in Physics, Chemistry (the exact sciences).

Non-deterministic Phenomena

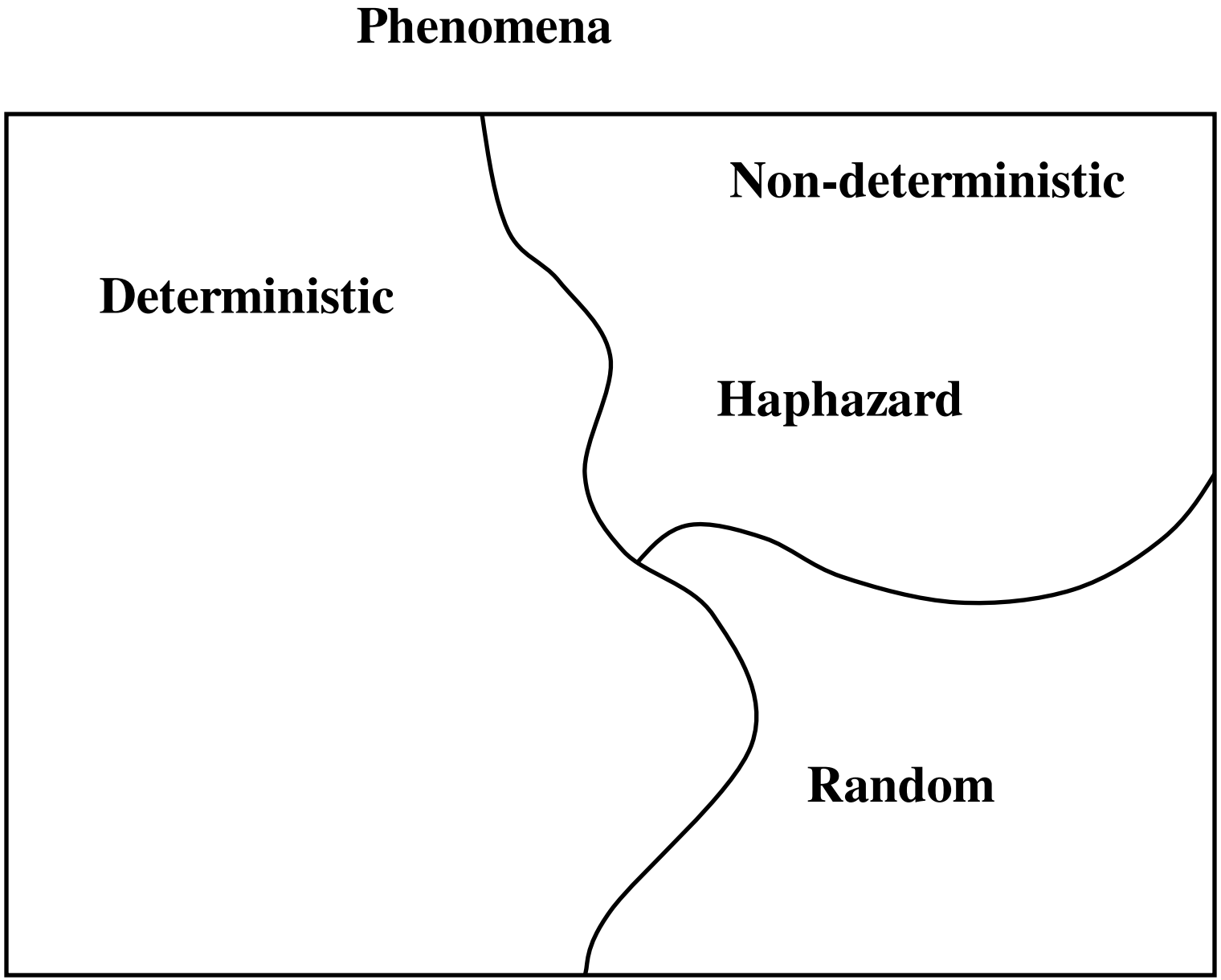
- **No** mathematical model exists that allows "***perfect***" prediction the phenomena's outcome.

Non-deterministic Phenomena

- may be divided into two groups.

1. **Random phenomena**
   - Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

2. **Haphazard phenomena**
   - unpredictable outcomes, but no long-run, exhibition of statistical regularity in the outcomes.


Phenomena
Non-deterministic
Deterministic
Haphazard
Random

Haphazard phenomena

- unpredictable outcomes, but no long-run, exhibition of statistical regularity in the outcomes.
- Do such phenomena exist?
- Will any non-deterministic phenomena exhibit long-run statistical regularity eventually?

Random phenomena

- Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

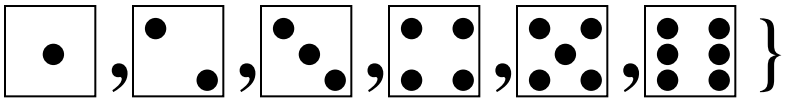
Examples

1. Tossing a coin – outcomes $S = \{\textbf{Head, Tail}\}$

   Unable to predict on each toss whether is Head or Tail.

   In the long run can predict that 50% of the time heads will occur and 50% of the time tails will occur

2. Rolling a die – outcomes

$S = \{$  $\}$

Unable to predict outcome but in the long run can one can determine that each outcome will occur 1/6 of the time.

Use symmetry. Each side is the same. One side should not occur more frequently than another side in the long run. If the die is not balanced this may not be true.

Definitions

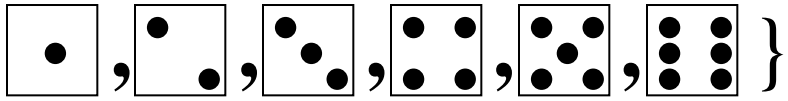
The sample Space, *S*

The **sample space**, *S*, for a random phenomena is the set of all possible outcomes.

Examples

1. Tossing a coin – outcomes $S = \{$**Head, Tail**$\}$

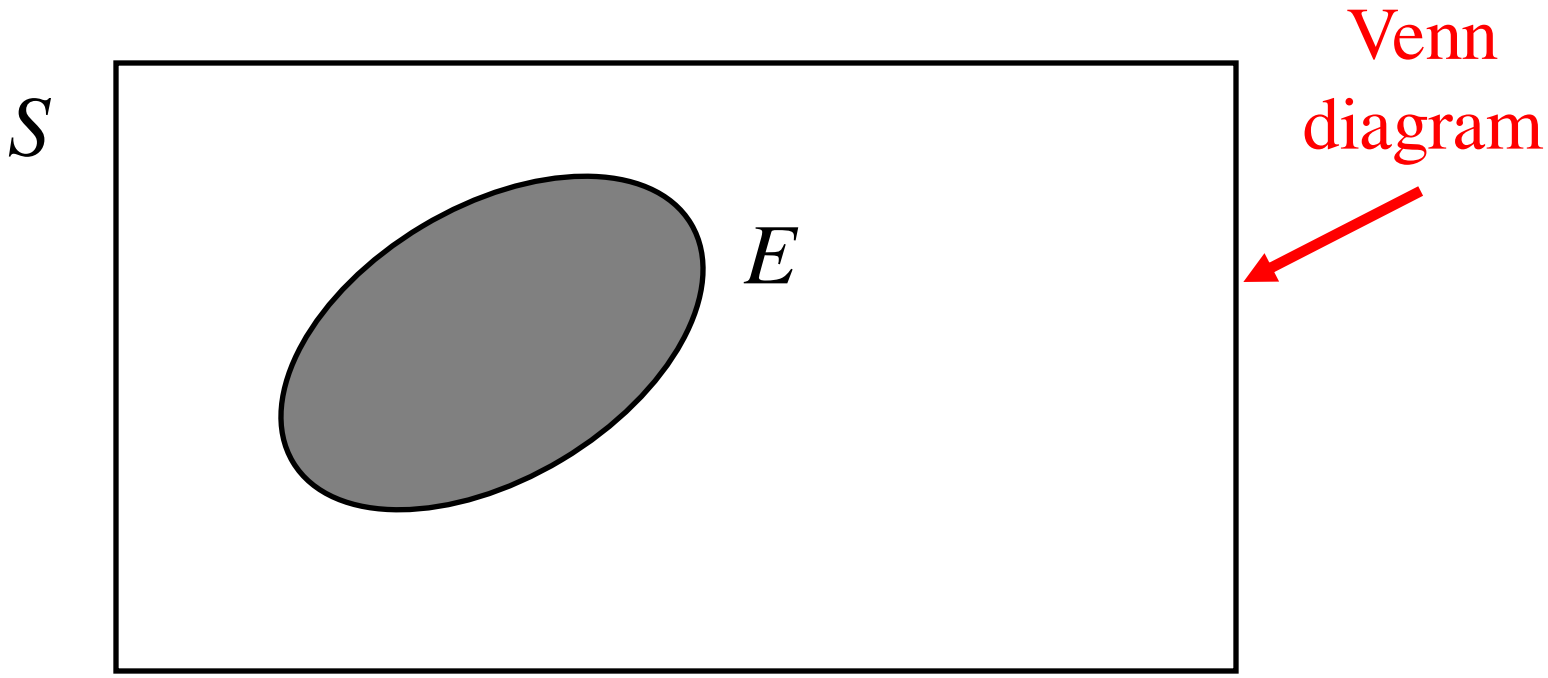
2. Rolling a die – outcomes

$S = \{$  $\}$

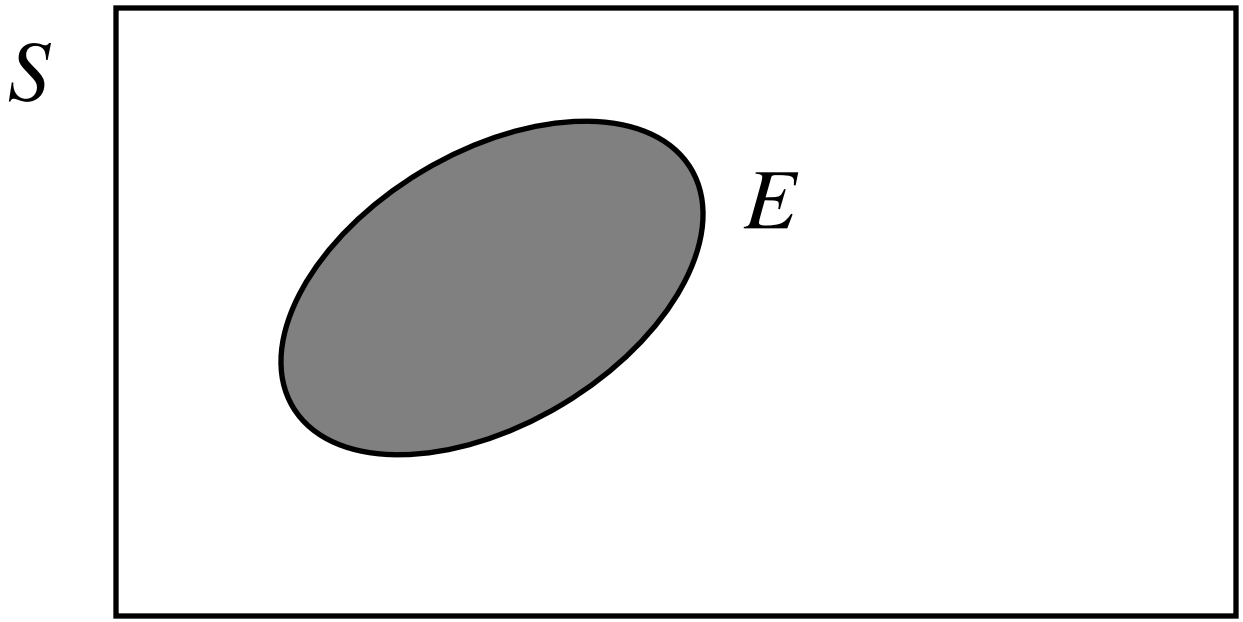
$= \{1, 2, 3, 4, 5, 6\}$

An Event , *E*

The **event**, *E*, is any subset of the **sample space**, *S*. i.e. any set of outcomes (not necessarily all outcomes) of the random phenomena

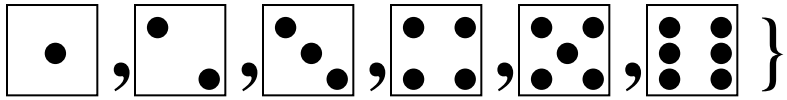


The **event**, *E*, is said to **have occurred** if after the outcome has been observed the outcome lies in *E*.



Examples

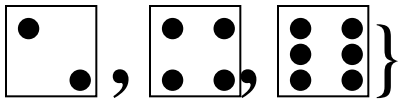
1. Rolling a die – outcomes

$S = \{$  $\}$

$= \{1, 2, 3, 4, 5, 6\}$

E = the event that an even number is rolled

$= \{2, 4, 6\}$

$= \{$  $\}$

Special Events

The Null Event, The empty event - ϕ

$\phi = \{\ \}$ = the event that contains no outcomes

The Entire Event, The Sample Space - S

S = the event that contains all outcomes

The empty event, ϕ , never occurs.

The entire event, S, always occurs.

Probability

Definition: probability of an Event *E*.

Suppose that the sample space $S = \{o_1, o_2, o_3, \ldots o_N\}$ has a finite number, N, of oucomes.

Also each of the outcomes is equally likely (because of symmetry).

Then for any event E

$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

Note: the symbol $n(A) =$ no. of elements of A

Thus this definition of $P[E]$, i.e.

$$P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$$

Applies only to the special case when

1. The sample space has a finite no.of outcomes, and
2. Each outcome is equi-probable

   If this is not true a more general definition of probability is required.

Maximum Likelihood

Much estimation theory is presented in a rather ad hoc fashion. Minimising squared errors seems a good idea but why not minimise the absolute error or the cube of the absolute error?

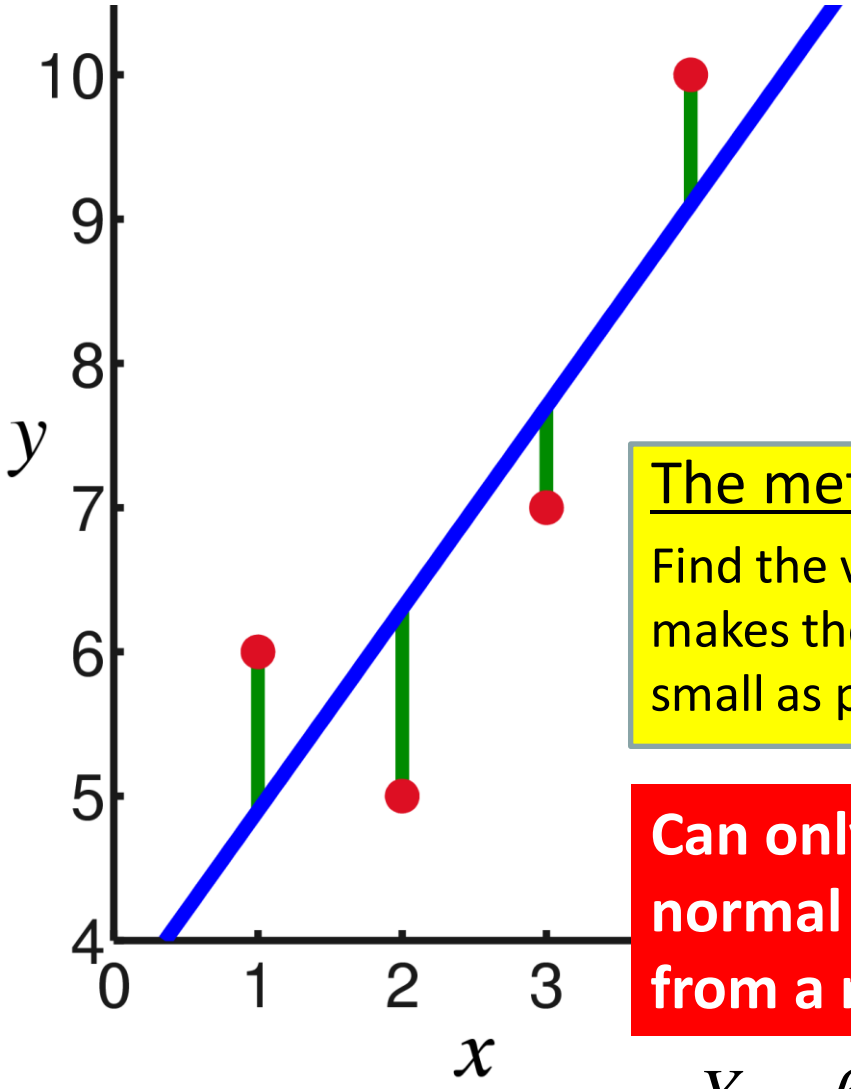
The answer is that there is an underlying approach which justifies a particular minimisation strategy conditional on certain assumptions.

This is the maximum likelihood principle.

The idea is to assume a particular model with unknown parameters, we can then define the probability of observing a given event conditional on a particular set of parameters. We have observed a set of outcomes in the real world. It is then possible to choose a set of parameters which are most likely to have produced the observed results.

This is maximum likelihood. In most cases it is both consistent and efficient. It provides a standard to compare other estimation techniques.

The method of least-squares



Model for the expectation (fixed part of the model):

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

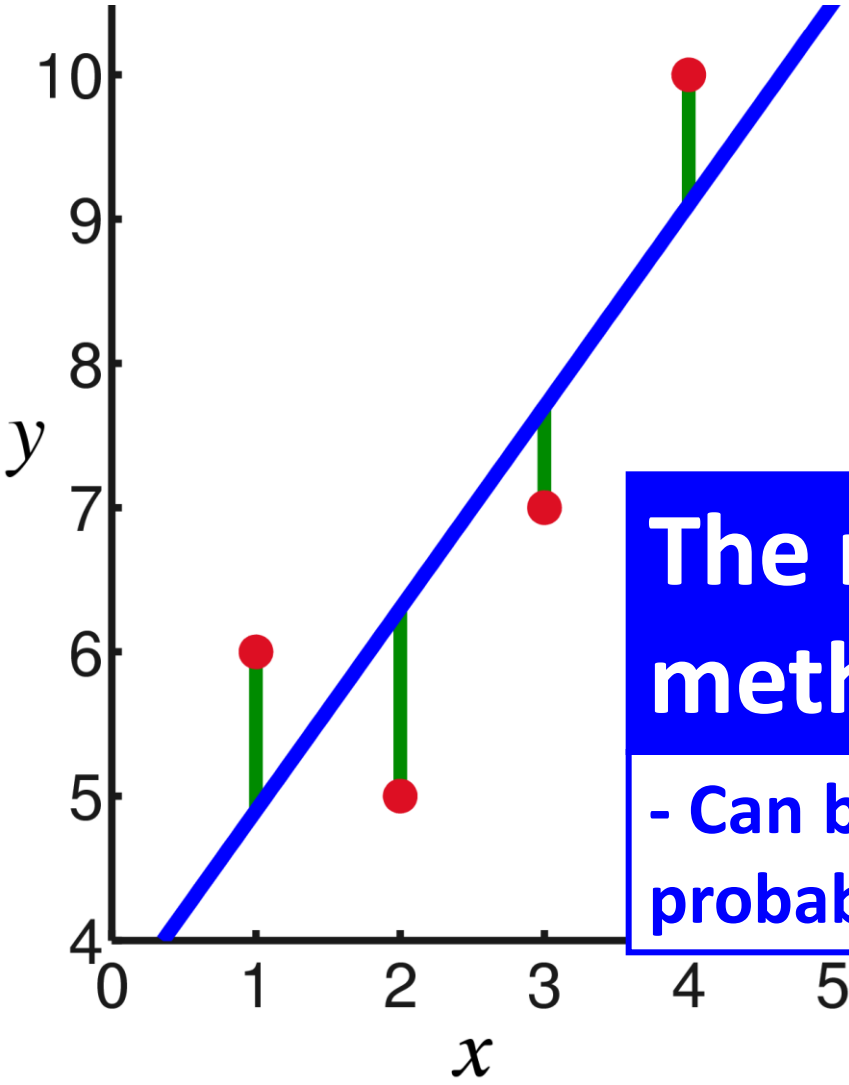
Residuals: $r_i = y_i - E[Y_i]$

The method of least-squares:

Find the values for the parameters (β_0 and β_1) that makes the sum of the squared residuals (Σr_j^2) as small as possible.

Can only be used when the error term is normal (residuals are assumed to be drawn from a normal distribution)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma)$$



Model for the expectation (fixed part of the model):

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

Residuals: $r_i = y_i - E[Y_i]$

The maximum likelihood method is more general!

- Can be applied to models with any probability distribution

The maximum likelihood

Example:

We want to estimate the probability, p, that individuals are infected with a certain kind of parasite.

Ind.:	Infected:	Probability of observation:
1	1	p
2	0	$1-p$
3	1	p
4	1	p
5	0	$1-p$
6	1	p
7	1	p
8	0	$1-p$
9	0	$1-p$
10	1	p

The maximum likelihood method (discrete distribution):

1. Write down the probability of each observation by using the model parameters
2. Write down the probability of all the data

$$\Pr(\text{Data} \mid p) = p^6(1-p)^4$$

3. Find the value parameter(s) that maximize this probability

The maximum likelihood

Example:

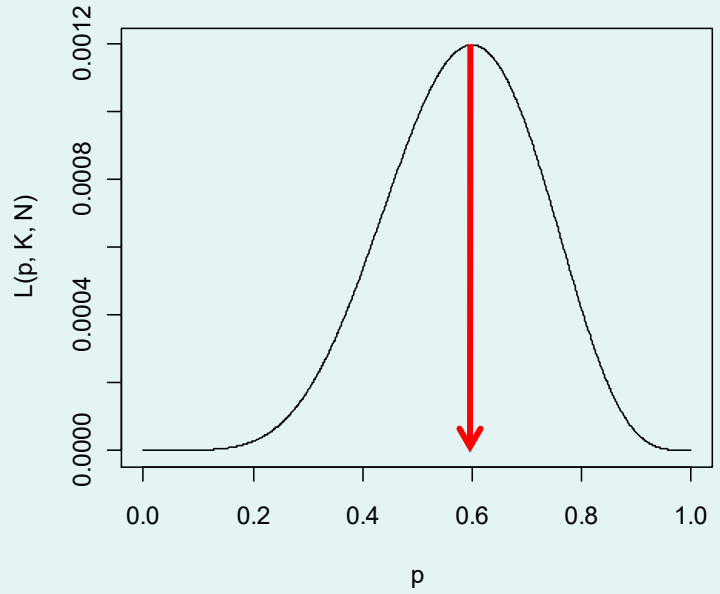
We want to estimate the probability, p, that individuals are infected with a certain kind of parasite.

Ind.:	Infected:	Probability of observation:
1	1	p
2	0	$1-p$
3	1	p
4	1	p
5	0	$1-p$
6	1	p
7	1	p
8	0	$1-p$
9	0	$1-p$
10	1	p

Likelihood function:

$$L(p) = \Pr(\text{Data} \mid p) = p^6(1-p)^4$$

- Find the value parameter(s) that maximize this probability



Statistical Approaches

- Mean
- Variance
- Skewness
- Kurtosis

Variance

- The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean.

- Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

Population variance

- When you have collected data from every member of the population that you're interested in, you can get an exact value for population variance.

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

Sample Variance

- When you collect data from a sample, the sample variance is used to make estimates or inferences about the population variance.

$$s^2 = \frac{\Sigma (X - \bar{x})^2}{n - 1}$$

Concept of Skewness

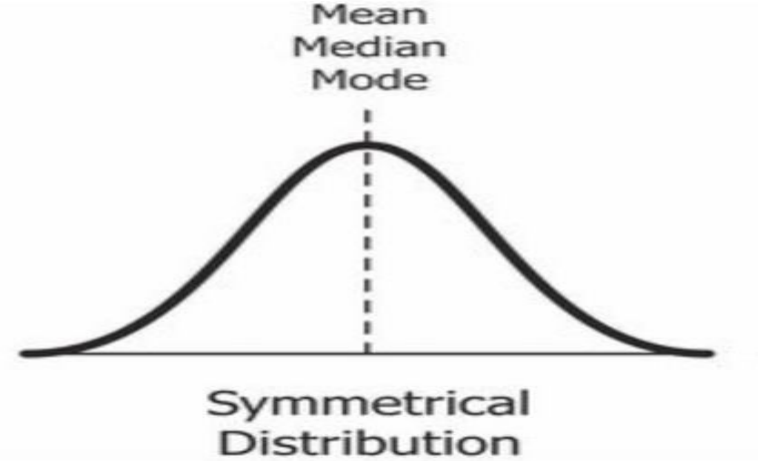
A distribution is said to be skewed-when the mean, median and mode fall at different position in the distribution and the balance (or center of gravity) is shifted to one side or the other i.e. to the left or to the right.

Therefore, the concept of skewness helps us to understand the relationship between three measures-

- **Mean.**
- **Median.**
- **Mode.**

Symmetrical Distribution

- A frequency distribution is said to be symmetrical if the frequencies are equally distributed on both the sides of central value.
- A symmetrical distribution may be either bell – shaped or U shaped.
- In symmetrical distribution, the values of mean, median and mode are equal i.e. **Mean=Median=Mode**



Skewed Distribution

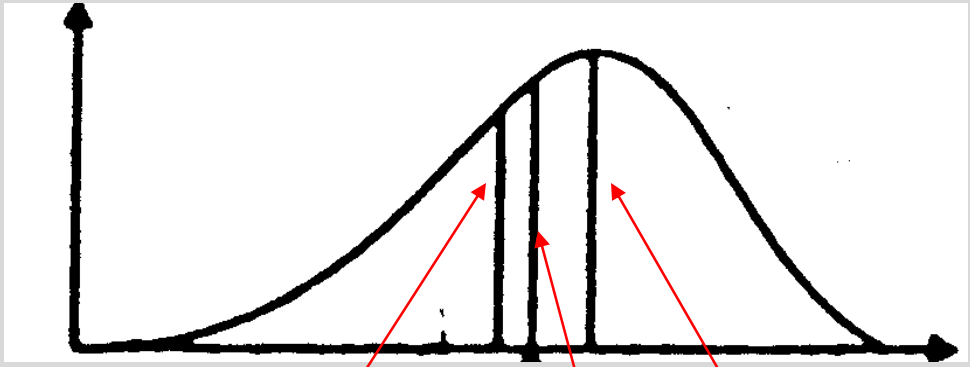
- A frequency distribution is said to be skewed if the frequencies are not equally distributed on both the sides of the central value.

- A skewed distribution may be-

  - **Positively Skewed**
  - **Negatively Skewed**

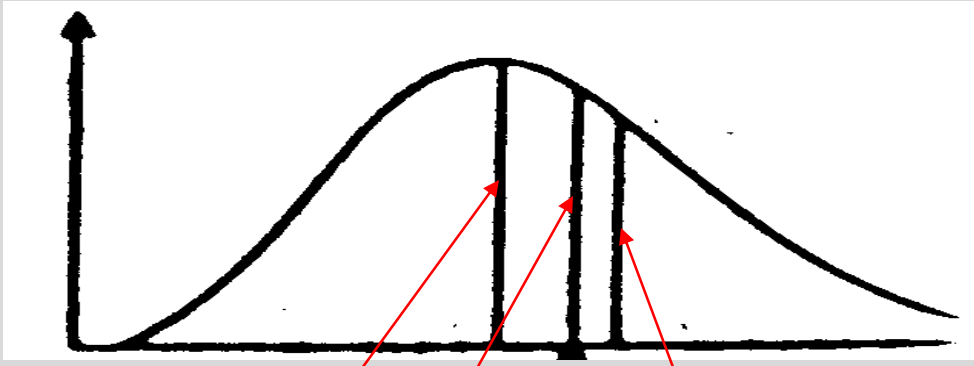
Skewed Distribution

- **<u>Negatively Skewed</u>**
- In this, the distribution is skewed to **the left (negative)**
- Here, **Mode** exceeds Mean and Median.



Mean<Median<Mode

- **<u>Positively Skewed</u>**
- In this, the distribution is skewed to **the right (positive)**
- Here, **Mean** exceeds Mode and Median.



Mode<Median<Mean

Tests of Skewness

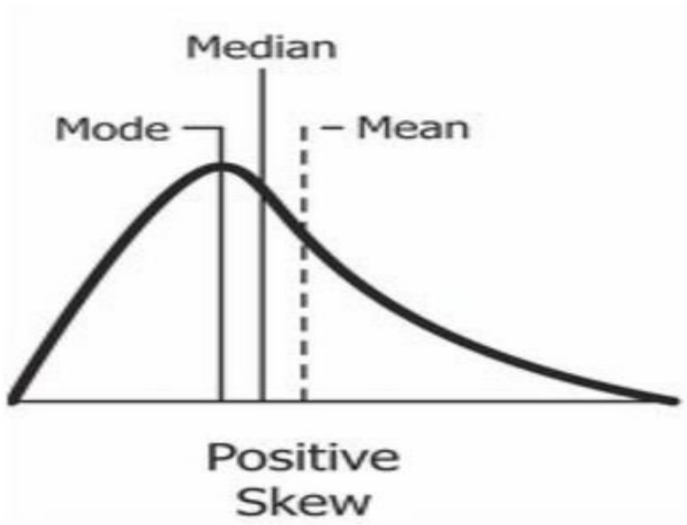
In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

- The values of mean, median and mode do not coincide.
- When the data are plotted on a graph they do not give the normal bell shaped form i.e. when cut along a vertical line through the center the two halves are not equal.
- The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
- Quartiles are not equidistant from the median.
- Frequencies are not equally distributed at points of equal deviation from the mode.

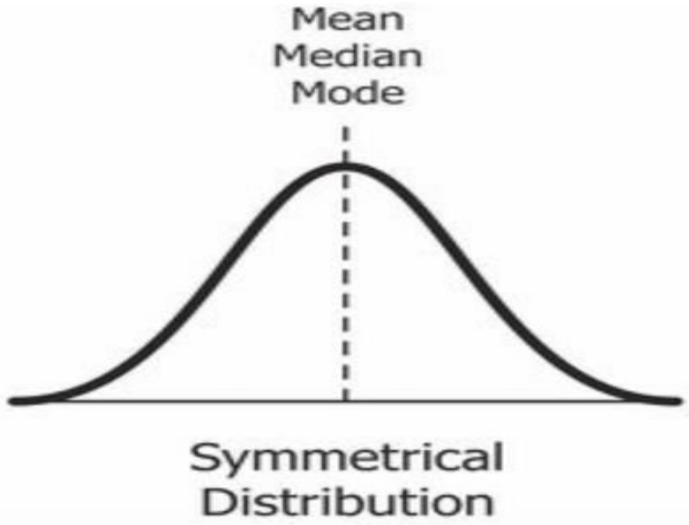
Graphical Measures of Skewness

- Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry.
- Positive or negative skewness can be detected graphically (as below) depending on whether the right tail or the left tail is longer but, we don't get idea of the magnitude
- Hence some statistical measures are required to find the magnitude of lack of symmetry

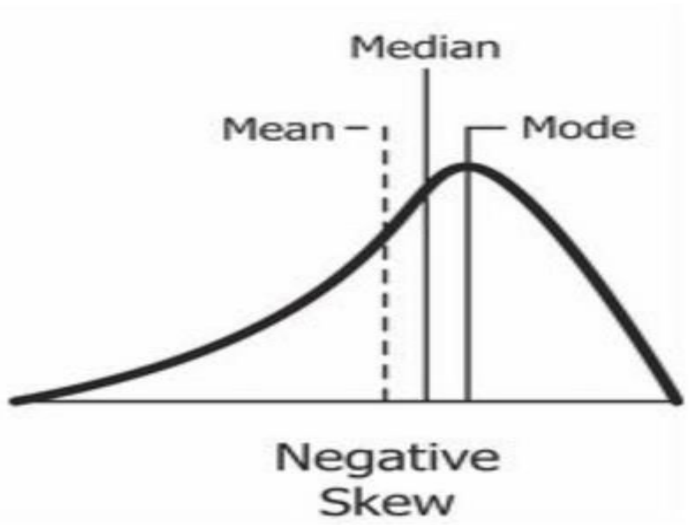
Mean> Median> Mode



Mean=Median=Mode



Mean<Median<Mode



Statistical Measures of Skewness

Absolute Measures of Skewness

Following are the absolute measures of skewness:

- Skewness (Sk) = Mean – Median
- Skewness (Sk) = Mean – Mode
- Skewness (Sk) = (Q3 - Q2) - (Q2 - Q1)

Relative Measures of Skewness

There are four measures of skewness:

- β and γ Coefficient of skewness
- Karl Pearson's Coefficient of skewness
- Bowley's Coefficient of skewness
- Kelly's Coefficient of skewness

β and γ Coefficient of Skewness

- Karl Pearson defined the following **β and γ coefficients** of skewness, based upon the second and third central moments:-

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

- β_1 as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative.
- This drawback is removed if we calculate Karl Pearson's Gamma coefficient γ_1 which is the square root of β_1 ie

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Karl Pearson's Coefficient of Skewness……01

- This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$SK_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Where,

SK_P = Karl Pearson's Coefficient of skewness,

σ = standard deviation.

Normally, this coefficient of skewness lies between -3 to +3.

Karl Pearson's Coefficient of Skewness…..02

In case the mode is indeterminate, the coefficient of skewness is:

$$SK_P = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\sigma}$$

Now this formula is equal to

$$SK_P = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The value of coefficient of skewness is **zero**, when the distribution is **symmetrical**.
The value of coefficient of skewness is **positive**, when the distribution is **positively skewed**.
The value of coefficient of skewness is **negative**, when the distribution is **negatively skewed**.

Bowley's Coefficient of Skewness……01

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

$$SK_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

Where,

SK_B = Bowley's Coefficient of skewness,

Q_1 = Quartile first Q_2 = Quartile second

Q_3 = Quartile Third

Bowley's Coefficient of Skewness…..02

The above formula can be converted to-

$$SK_B = \frac{Q_3 + Q_1 - 2\text{Median}}{(Q_3 - Q_1)}$$

The value of coefficient of skewness is **zero**, if it is a **symmetrical distribution**.
If the value is **greater than zero**, it is **positively skewed** distribution.
And if the value is **less than zero**, it is **negatively skewed** distribution.

Kelly's Coefficient of Skewness…..01

Kelly developed another measure of skewness, which is based on percentiles and deciles.

The formula for measuring skewness is based on percentile as follows:

$$SK_k = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Where,

SK_K = Kelly's Coefficient of skewness,

P_{90} = Percentile Ninety.

P_{50} = Percentile Fifty.

P_{10} = Percentile Ten.

Kelly's Coefficient of Skewness…..02

This formula for measuring skewness is based on percentile are as follows:

$$SK_k = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$$

Where,

SK_K = Kelly's Coefficient of skewness,

D_9 = Deciles Nine.

D_5 = Deciles Five. D_1 = Deciles one.

Moments:

- In Statistics, moments is used to indicate peculiarities of a frequency distribution.
- The utility of moments lies in the sense that they indicate different aspects of a given distribution.
- Thus, by using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the peakedness of the curve.
- The moments about the actual arithmetic mean are denoted by μ.
- The first four moments about mean or central moments are following:-

Moments:

Moments around Mean

For ungrouped data, $\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$

For grouped data, $\mu_r = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \overline{x})^r$

where $n = \sum_{i=1}^{k} f_i$ and $\overline{x} = \sum_{i=1}^{k} f_i x_i$

Moments around any Arbitrary No

For ungrouped data, $\mu'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$

For grouped data, $\mu'_r = \frac{1}{n}\sum_{i=1}^{n} f_i x_i^r$

where, $n = \sum_{i=1}^{k} f_i$

Conversion formula for Moments

1st moment: $\mu_1 = 0$ **(Mean)**

2nd moment: $\mu_2 = \mu_2' - \mu_1'^2$ **(Variance)**

3rd moment: $\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$ **(Skewness)**

4th moment: $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$ **(Kurtosis)**

Two important constants calculated from μ_2, μ_3 and μ_4 are:-

β_1 (read as beta one)	β_2 (read as beta two)
• β_1 is used to measures of skewness. • It is defined as: $$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$	• β_2 is used to measures Kurtosis. • It is defined as: $$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Kurtosis

•Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess.

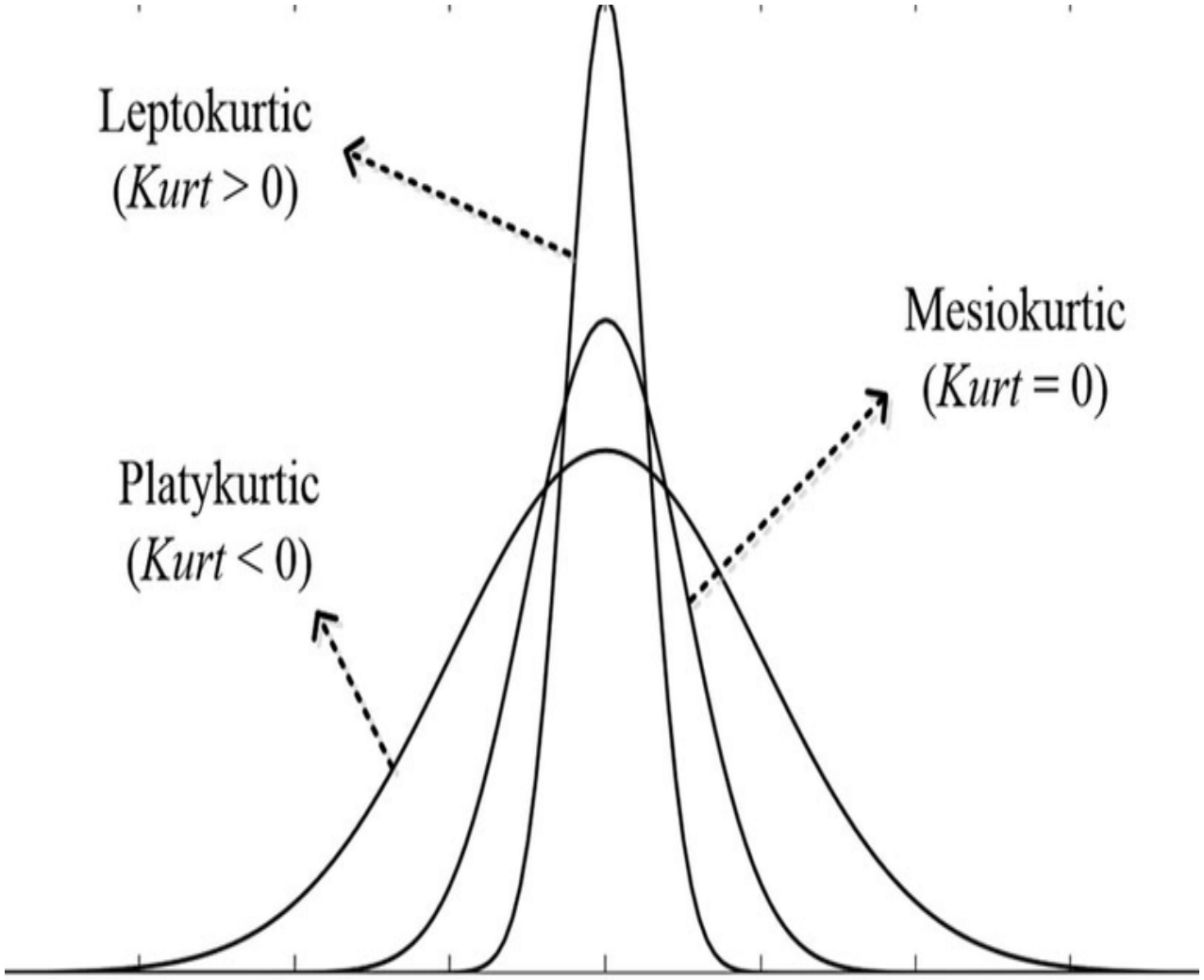
•While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution.

•Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are:-

- **Leptokurtic**
- **Mesokurtic**
- **Platykurtic**

Kurtosis

- When the peak of a curve becomes relatively high then that curve is called **Leptokurtic.**
- When the curve is flat-topped, then it is called **Platykurtic**.
- Since normal curve is neither very peaked nor very flat topped, so it is taken as a basis for comparison.
- This normal curve is called **Mesokurtic**.



Measure of Kurtosis

- There are two measure of Kurtosis:

- **Karl Pearson's Measures of Kurtosis**

- **Kelly's Measure of Kurtosis**

Karl Pearson's Measures of Kurtosis

Formula

- For calculating the kurtosis, the second and fourth central moments of variable are used
- For this, following formula given by Karl Pearson is used:

$$\boldsymbol{\beta_2 = \frac{\mu_4}{\mu_2^2}}$$

- Or

$$\textbf{Kurtosis } (\gamma 2) = \left(\frac{\mu_4}{\mu_2^2}\right) - 3$$

where,

$\boldsymbol{\mu_2}$ = Second order central moment of distribution

$\boldsymbol{\mu_4}$ = Fourth order central moment of distribution

Result:

- If $\boldsymbol{\beta_2} = 3$ or $\boldsymbol{\gamma 2} = 0$, then curve is said to be mesokurtic;
- If $\boldsymbol{\beta_2} < 3$ or $\boldsymbol{\gamma 2} < 0$, then curve is said to be platykurtic;
- If $\boldsymbol{\beta_2} > 3$ or $\boldsymbol{\gamma 2} > 0$, then curve is said to be leptokurtic;

Kelly's Measure of Kurtosis

Formula

- Kelly has given a measure of kurtosis based on percentiles.
- The formula is given by :-

$$\beta_2 = \frac{P_{75} - P_{25}}{P_{90} - P_{10}}$$

where,

P_{75}, P_{25}, P_{90}, and P_{10} are 75th, 25th , 90th and 10th percentiles of dispersion respectively.

Result:

- If $\beta_2 > 0.26315$, then the distribution is platykurtic.
- If $\beta_2 < 0.26315$, then the distribution is leptokurtic.

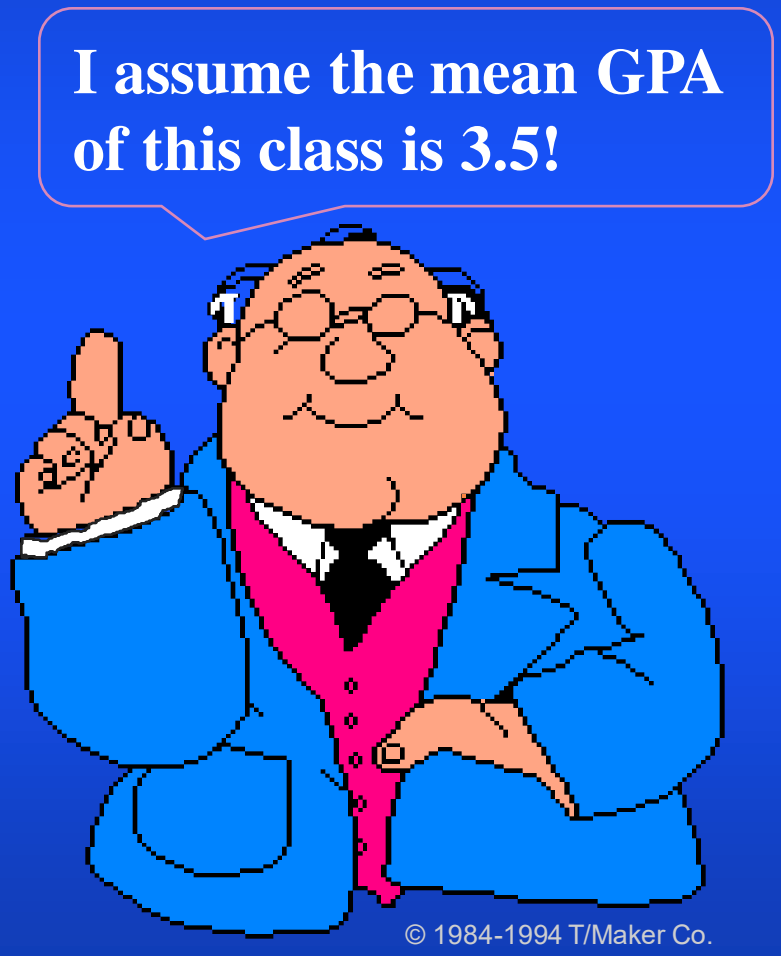
Chapter Topics

- Hypothesis Testing Methodology
- Z Test for the Mean (σ Known)
- p-Value Approach to Hypothesis Testing
- Connection to Confidence Interval Estimation
- One Tail Test
- t Test of Hypothesis for the Mean
- Z Test of Hypothesis for the Proportion

What is a Hypothesis?

A hypothesis is an assumption about the population parameter.

- **A parameter is a Population mean or proportion**
- **The parameter must be identified before analysis.**



The Null Hypothesis, H_0

- States the Assumption (numerical) to be tested

  e.g. The average # TV sets in US homes is at least 3 (H_0: $\mu \geq 3$)

- Begin with the assumption that the null hypothesis is TRUE.

  (Similar to the notion of innocent until proven guilty)

  - Refers to the Status Quo
  - Always contains the ' = ' sign

- The Null Hypothesis may or may not be rejected.

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
  e.g. The average # TV sets in US homes is less than 3 (H_1: $\mu < 3$)
- Challenges the Status Quo
- Never contains the '=' sign
- The Alternative Hypothesis may or may not be accepted

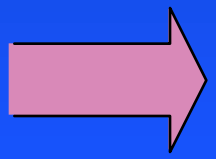
Identify the Problem

Steps:

- **State the Null Hypothesis (H_0: $\mu \geq 3$)**
- **State its opposite, the Alternative Hypothesis (H_1: $\mu < 3$)**
  - **Hypotheses are mutually exclusive & exhaustive**
  - **Sometimes it is easier to form the alternative hypothesis first.**

Hypothesis Testing Process

Assume the population mean age is 50. (Null Hypothesis)



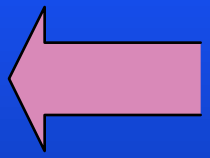
Population

Is $\overline{X} = 20 \cong \mu = 50$?

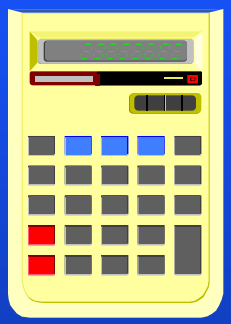
No, not likely!

REJECT

Null Hypothesis



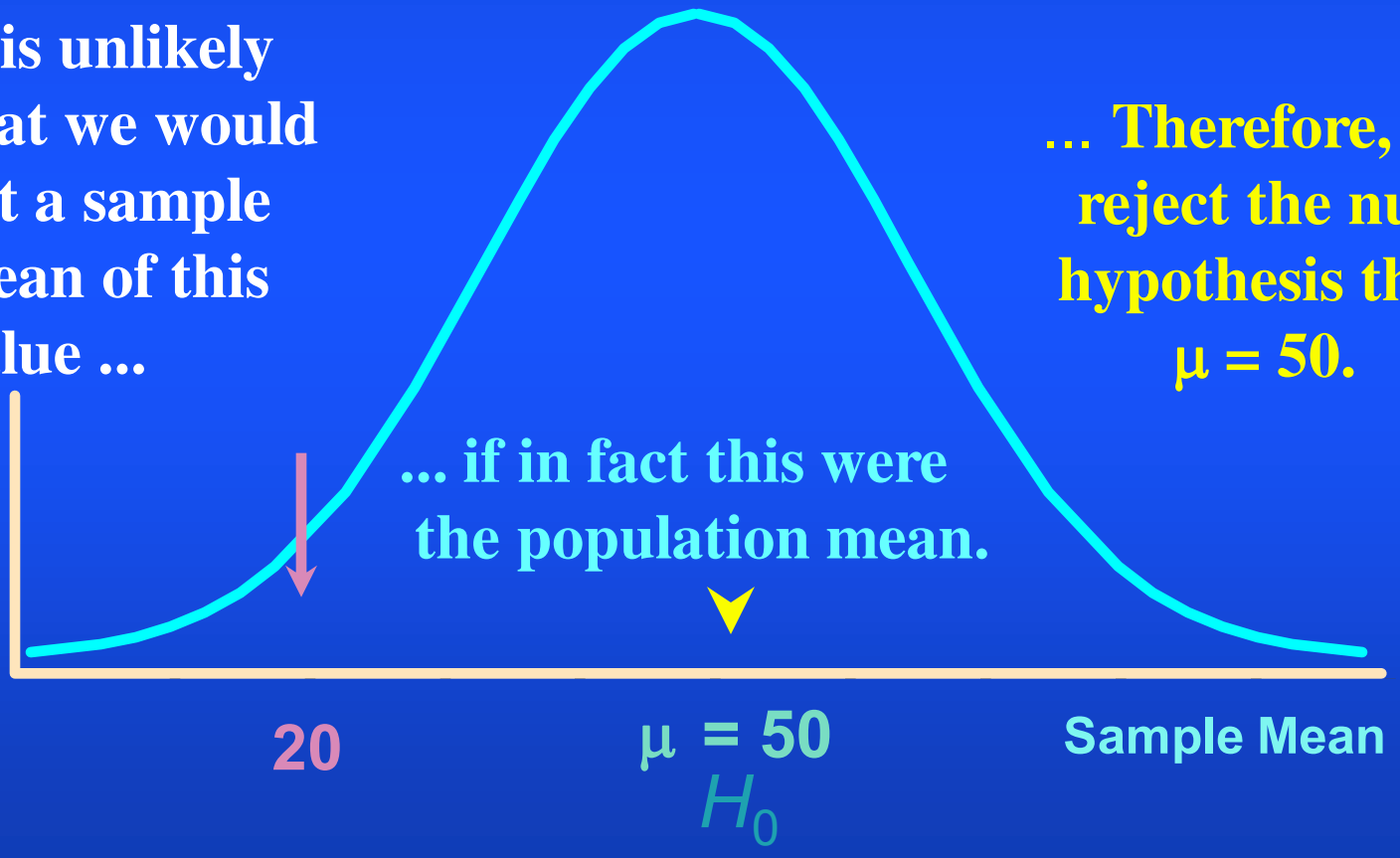
The Sample Mean Is 20



Sample

Reason for Rejecting H_0

Sampling Distribution



Level of Significance, α

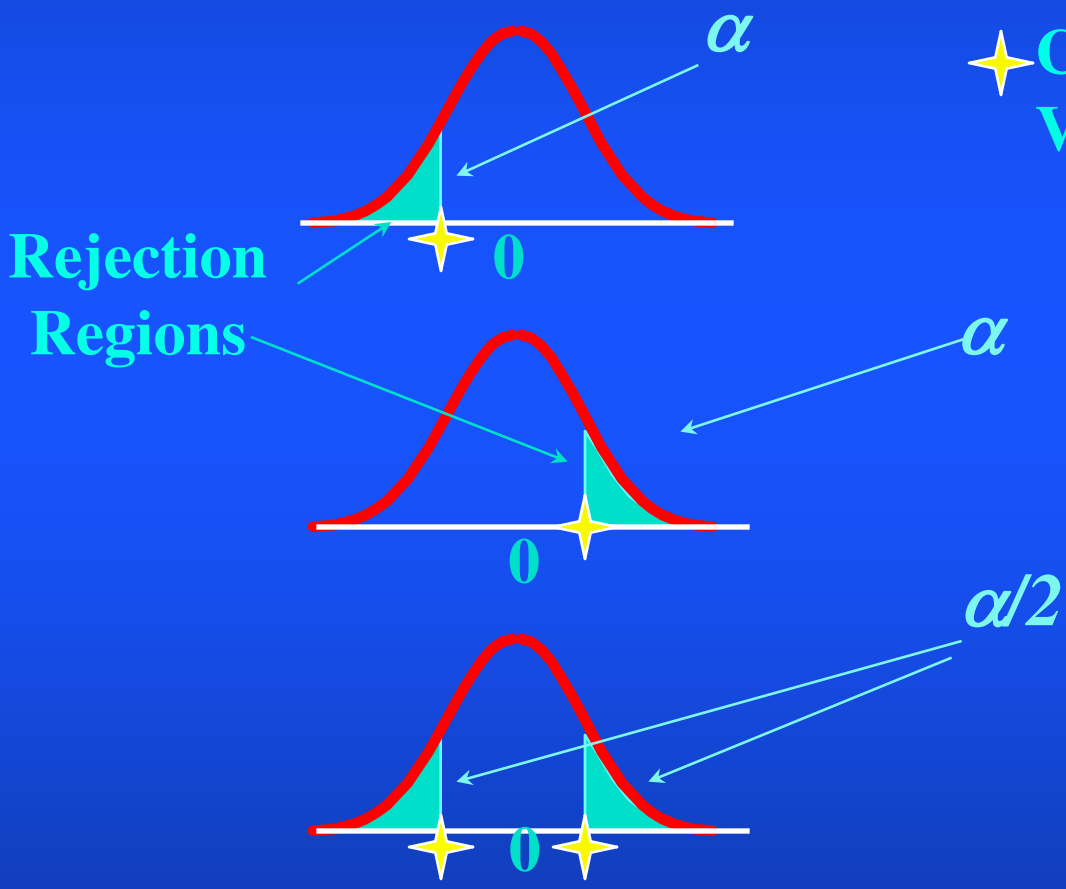
- Defines Unlikely Values of Sample Statistic if Null Hypothesis Is True
  - Called Rejection Region of Sampling Distribution
- Designated α (alpha)
  - Typical values are 0.01, 0.05, 0.10
- Selected by the Researcher at the Start
- Provides the Critical Value(s) of the Test

Level of Significance, α and the Rejection Region

$H_0: \mu \geq 3$
$H_1: \mu < 3$

$H_0: \mu \leq 3$
$H_1: \mu > 3$

$H_0: \mu = 3$
$H_1: \mu \neq 3$



Errors in Making Decisions

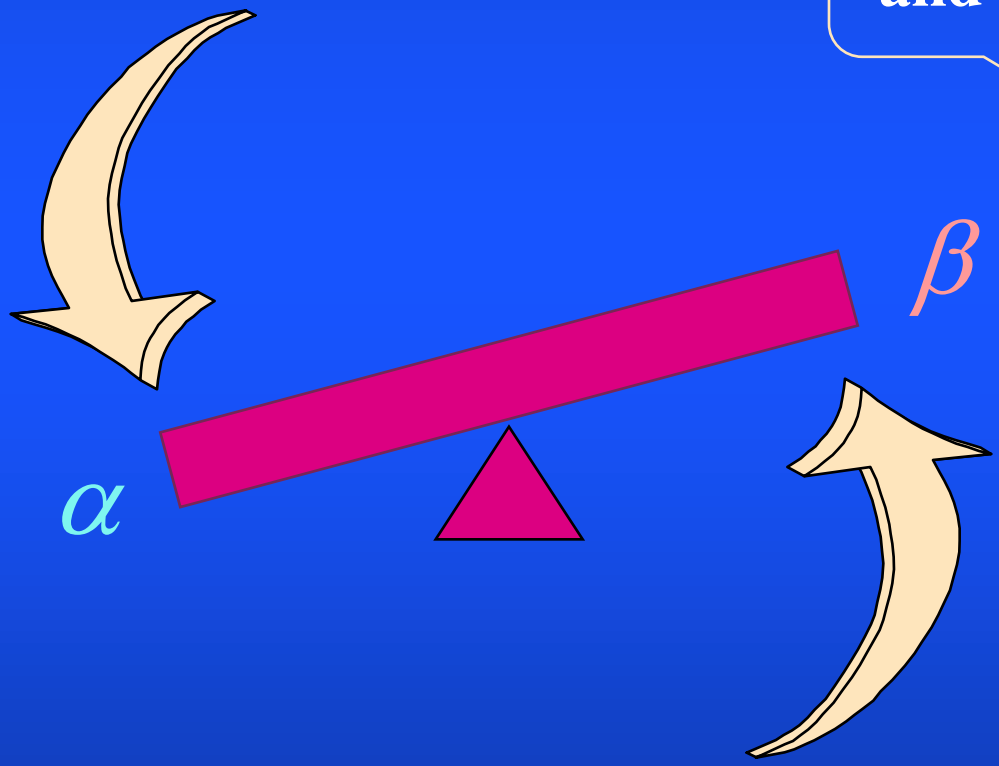
- **Type I Error**
  - Reject True Null Hypothesis
  - Has Serious Consequences
  - Probability of Type I Error Is α
    - Called Level of Significance
- **Type II Error**
  - Do Not Reject False Null Hypothesis
  - Probability of Type II Error Is β (Beta)

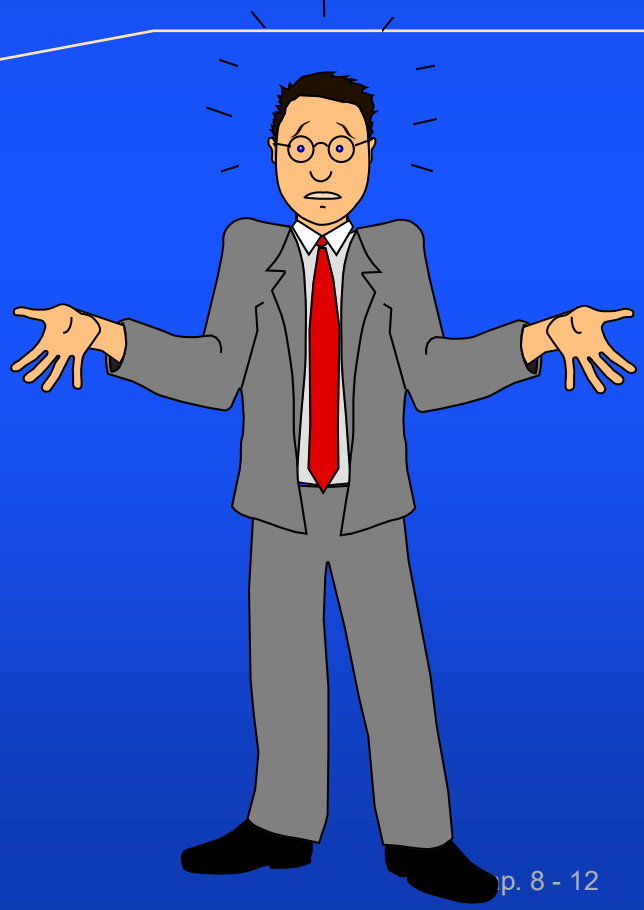
Result Possibilities

H_0: Innocent

Jury Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Verdict	Innocent	Guilty	**Decision**	H_0 True	H_0 False
Innocent	**Correct**	**Error**	Do Not Reject H_0	**1 - α**	**Type II Error (β)**
Guilty	**Error**	**Correct**	Reject H_0	**Type I Error (α)**	**Power (1 - β)**

α & β Have an Inverse Relationship





Z-Test Statistics (σ Known)

- **Convert Sample Statistic (e.g., $\bar{X}$) to Standardized Z Variable**

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Test Statistic

- **Compare to Critical Z Value(s)**
  - **If Z test Statistic falls in Critical Region, Reject H_0; Otherwise Do Not Reject H_0**

p Value Test

- **Probability of Obtaining a Test Statistic More Extreme ($\leq$ or $\geq$) than Actual Sample Value Given H_0 Is True**
- **Called Observed Level of Significance**
  - **Smallest Value of a H_0 Can Be Rejected**
- **Used to Make Rejection Decision**
  - **If *p* value $\geq \alpha$, Do Not Reject H_0**
  - **If *p* value $< \alpha$, Reject H_0**

Hypothesis Testing: Steps

Test the Assumption that the true mean # of TV sets in US homes is at least 3.

1.	State H_0	$H_0: \mu \geq 3$
2.	State H_1	$H_1: \mu < 3$
3.	Choose α	$\alpha = .05$
4.	Choose n	n = 100
5.	Choose Test:	*Z Test (or p Value)*

Hypothesis Testing: Steps
(continued)

Test the Assumption that the average # of TV sets in US homes is at least 3.

6. Set Up Critical Value(s)	***Z = -1.645***
7. Collect Data	***100 households surveyed***
8. Compute Test Statistic	***Computed Test Stat.= -2***
9. Make Statistical Decision	***Reject Null Hypothesis***
10. Express Decision	***The true mean # of TV set is less than 3 in the US households.***

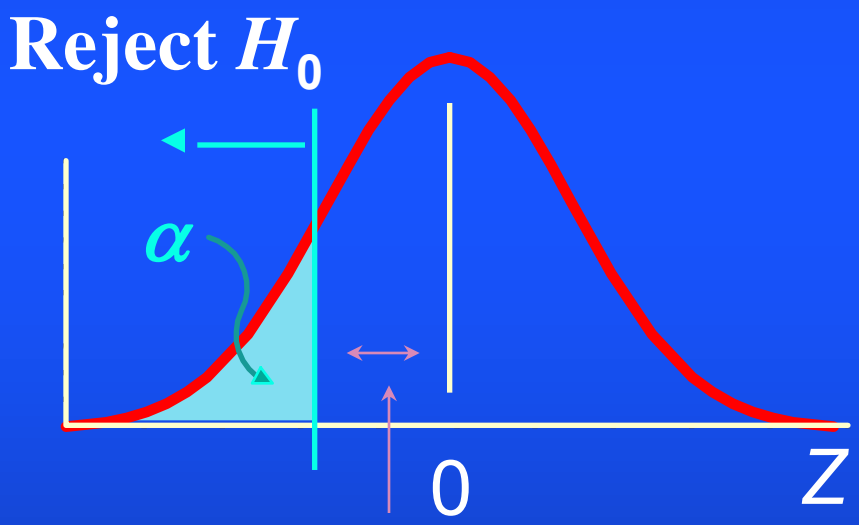
One-Tail Z Test for Mean (σ Known)

- **Assumptions**
  - Population Is Normally Distributed
  - If Not Normal, use large samples
  - Null Hypothesis Has ≤ or ≥ Sign Only
- *Z* Test Statistic:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

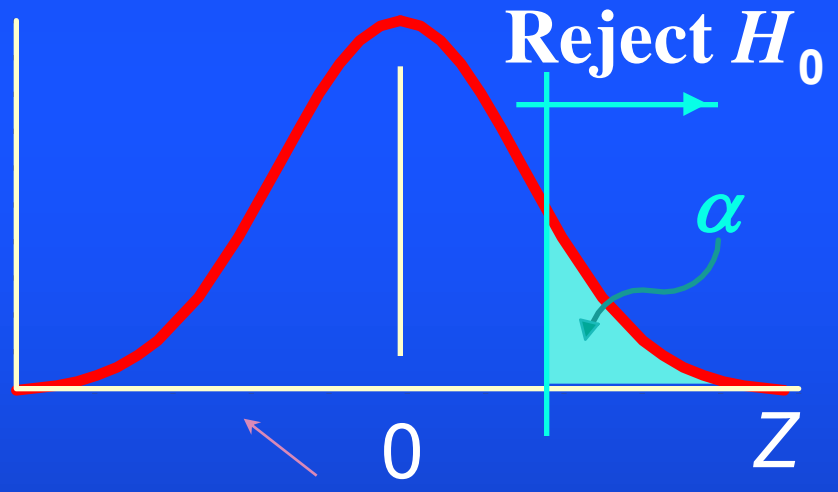
Rejection Region

H_0: $\mu \geq 0$
H_1: $\mu < 0$



Must Be *Significantly* Below $\mu = 0$

H_0: $\mu \leq 0$
H_1: $\mu > 0$



Small values don't contradict H_0
Don't Reject H_0!

Example: One Tail Test

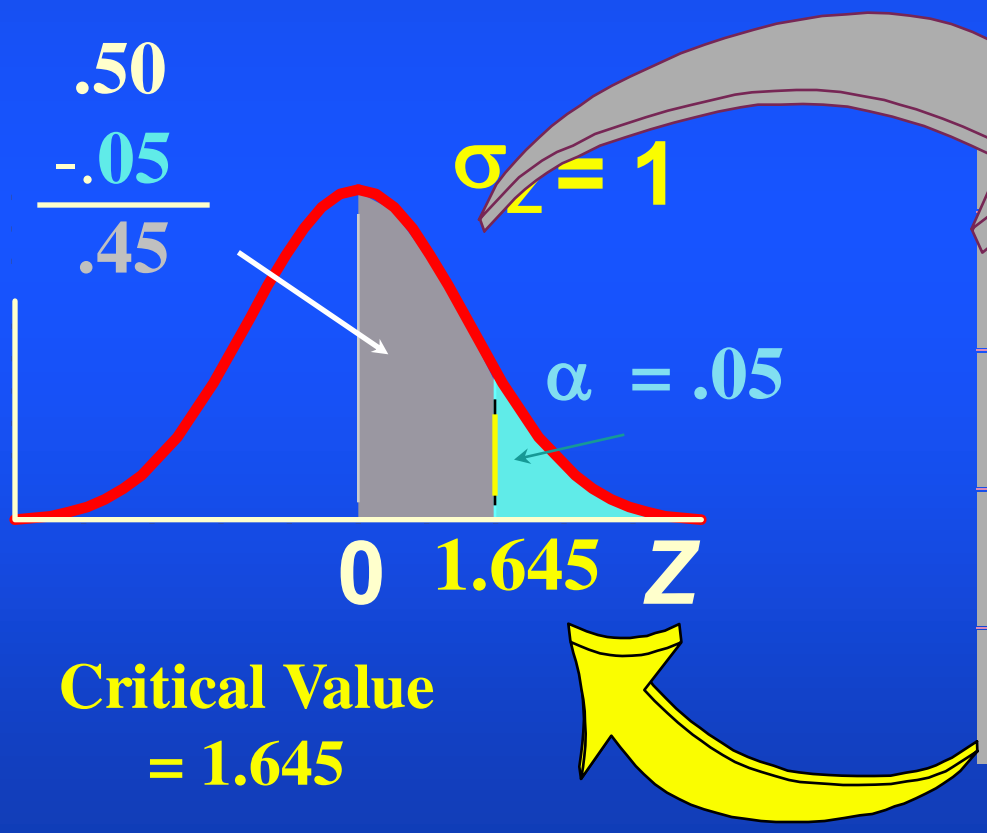
Does an average box of cereal contain **more than** **368** grams of cereal? A random sample of **25** boxes showed $\bar{X} = 372.5$. The company has specified σ to be **15** grams. Test at the α=**0.05** level.

H_0: $\mu \leq 368$

H_1: $\mu > 368$

Finding Critical Values: One Tail

What Is *Z* Given $\alpha = 0.05$?



Critical Value = 1.645

Standardized Normal Probability Table (Portion)

	.04	.05	.06
1.6	**.5495**	**.5505**	.5515
1.7	.5591	.5599	.5608
1.8	.5671	.5678	.5686
1.9	.5738	.5744	.5750

Example Solution: One Tail

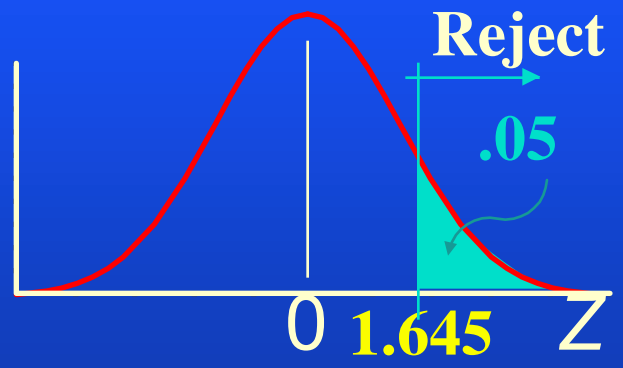
H_0: $\mu \leq 368$

H_1: $\mu > 368$

$\alpha = 0.025$

$n = 25$

Critical Value: 1.645



Test Statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = 1.50$$

Decision:

Do Not Reject at $\alpha = .05$

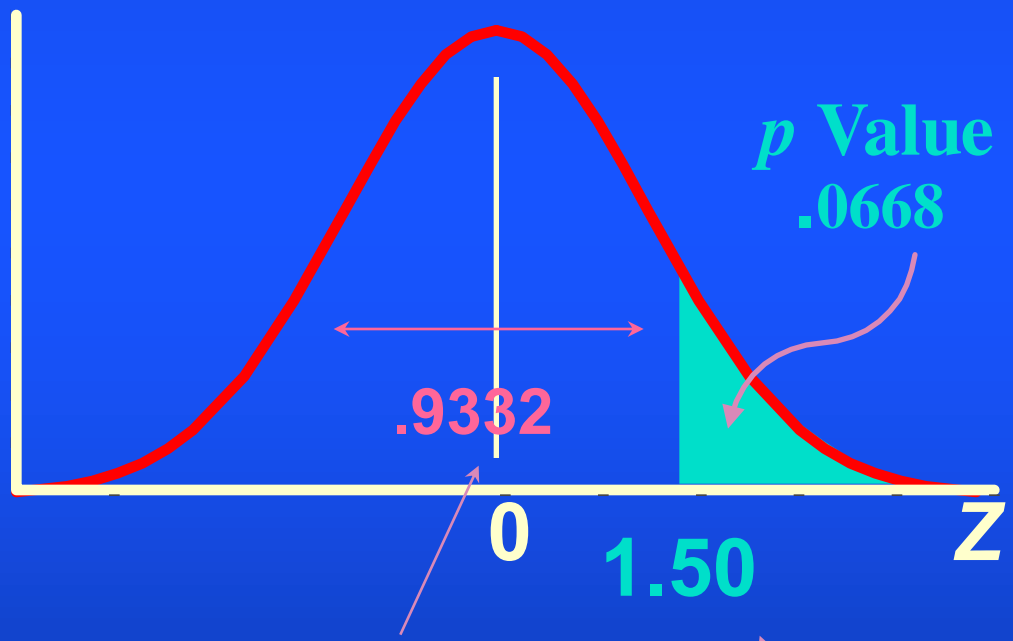
Conclusion:

No Evidence True Mean Is More than 368

p Value Solution

p Value is $P(Z \geq 1.50) = 0.0668$

Use the alternative hypothesis to find the direction of the test.



1.0000
- .9332
.0668

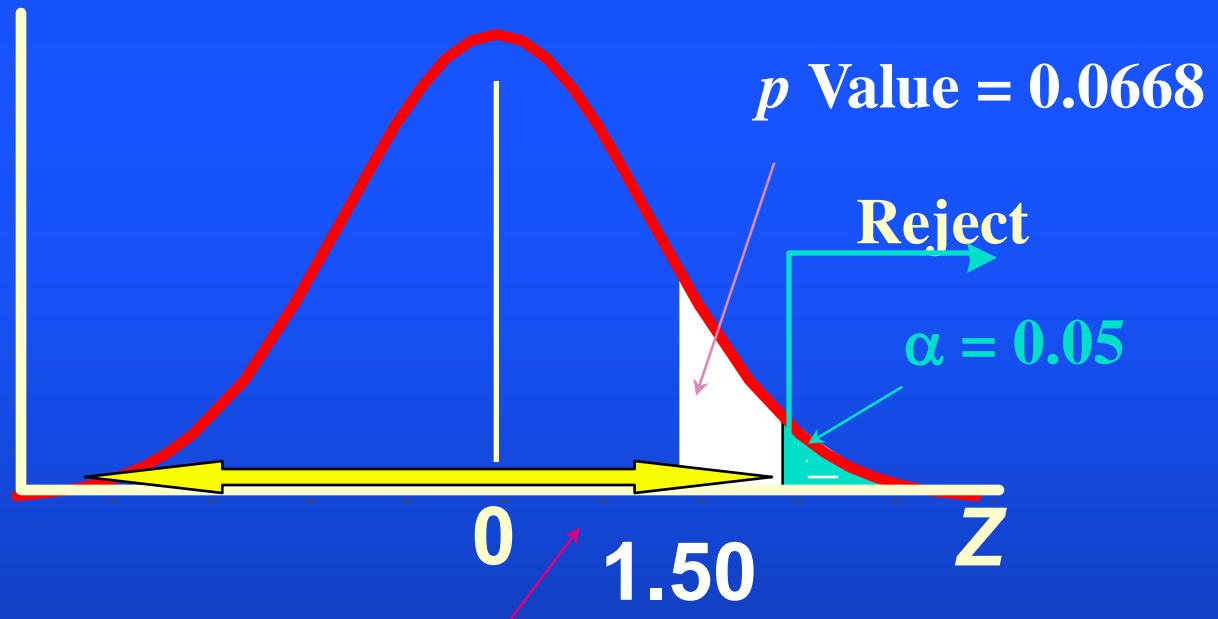
From *Z* Table: Lookup 1.50

Z Value of Sample Statistic

p Value Solution

(*p* Value = 0.0668) ≥ (α = 0.05).

Do Not Reject.



Test Statistic Is In the Do Not Reject Region

Example: Two Tail Test

Does an average box of cereal contains 368 grams of cereal? A random sample of 25 boxes showed $\bar{X} = 372.5$. The company has specified σ to be 15 grams. Test at the α=0.05 level.

H_0: $\mu = 368$

H_1: $\mu \neq 368$

Example Solution: Two Tail

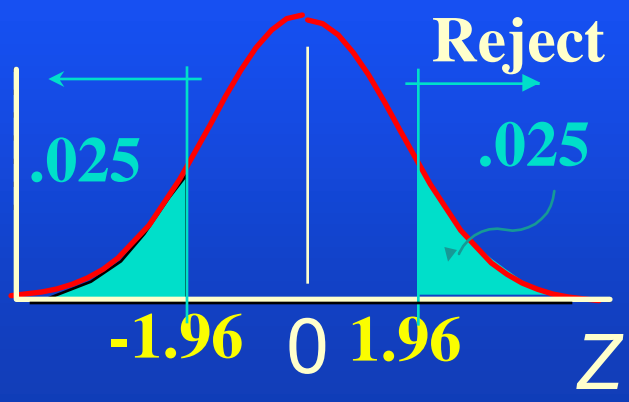
H_0: $\mu = 386$

H_1: $\mu \neq 386$

$\alpha = 0.05$

$n = 25$

Critical Value: ±1.96



Test Statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{3725 - 368}{15/\sqrt{25}} = 1.50$$

Decision:

Do Not Reject at $\alpha = .05$

Conclusion:

No Evidence that True Mean Is Not 368

Connection to Confidence Intervals

For $\bar{X} = 372.5$oz, $\sigma = 15$ and $n = 25$,

The 95% Confidence Interval is:

$$372.5 - (1.96)\,15/\sqrt{25} \text{ to } 372.5 + (1.96)\,15/\sqrt{25}$$

or

$$366.62 \le \mu \le 378.38$$

If this interval contains the Hypothesized mean (368), we do not reject the null hypothesis.

It does. Do not reject.

t-Test: σ Unknown

Assumptions

- Population is normally distributed
- If not normal, only slightly skewed & a large sample taken

Parametric test procedure

t test statistic

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Example: One Tail t-Test

Does an average box of cereal contain **more than** 368 grams of cereal? A random sample of **36** boxes showed $\bar{X} = 372.5$, and $\sigma = 15$. Test at the α=**0.01** level.

σ is not given,

H_0: $\mu \leq 368$
H_1: $\mu > 368$

Example:Z Test for Proportion

- **Problem:** A marketing company claims that it receives 4% responses from its Mailing.

- **Approach:** To test this claim, a random sample of 500 were surveyed with 25 responses.

- **Solution:** Test at the $\alpha = .05$ significance level.

Z Test for Proportion: Solution

$H_0: p = .04$

$H_1: p \neq .04$

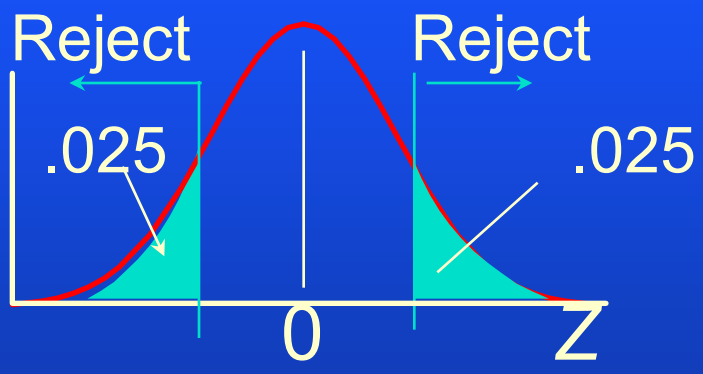
$\alpha = .05$

$n = 500$

Test Statistic:

$$Z \cong \frac{p - p_S}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.04 - .05}{\sqrt{\frac{.04(1-.04)}{500}}} = 1.14$$

Critical Values: ± 1.96



Decision:

Do not reject at $\alpha = .05$

Conclusion:

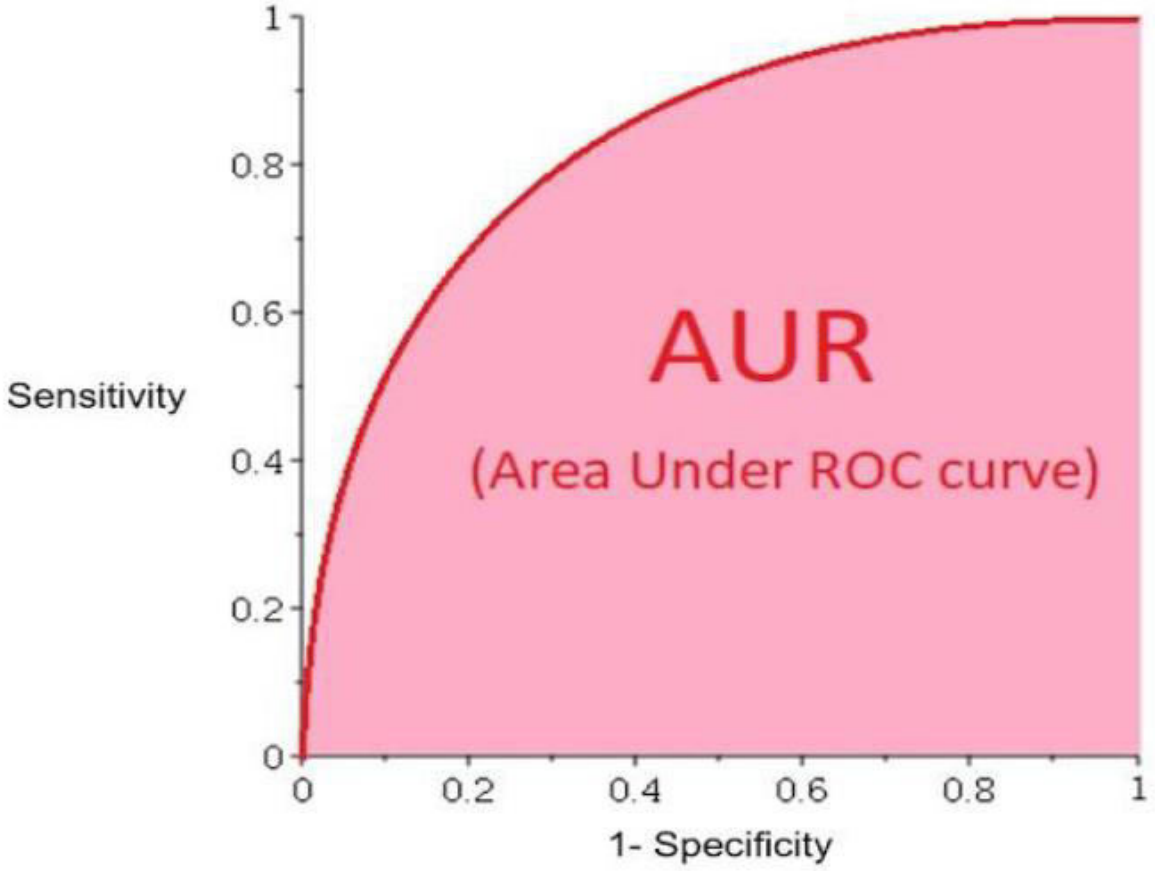
We do not have sufficient evidence to reject the company's claim of 4% response rate.

Model Selection

Splitting dataset

- Split the dataset in 70/15/15 fashion
- Have a different dataset for validation
- In case the model is not fitting well, there might be errors of problem selection
- If training error is high there might underfitting
- Small error in training and a significant validation and testing error , we have issue variance.

Area under the curve



Cross-Validation

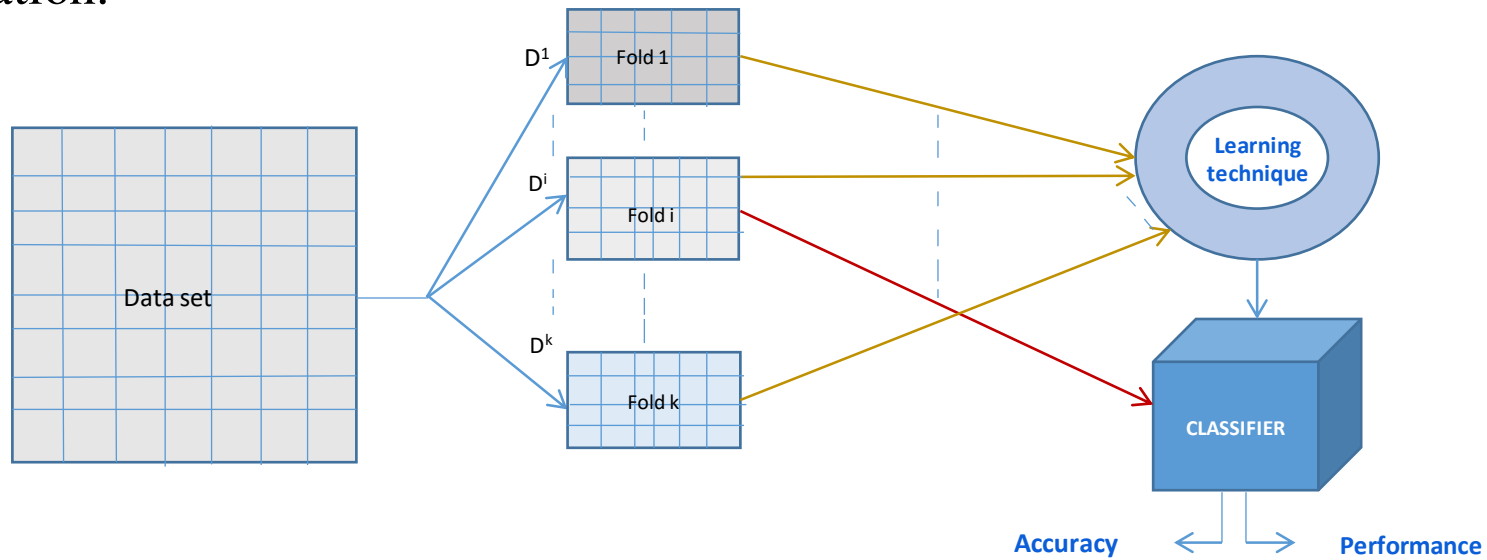
- The main drawback of Random subsampling is, it does not have control over the number of times each tuple is used for training and testing.
- Cross-validation is proposed to overcome this problem.
- There are two variations in the cross-validation method.
  - k-fold cross-validation
  - *N*-fold cross-validation

k-fold Cross-Validation

- Dataset consisting of N tuples is divided into k (usually, 5 or 10) equal, mutually exclusive parts or folds $(D_1, D_2, \ldots, D_k)$, and if N is not divisible by k, then the last part will have fewer tuples than other (k-1) parts.

- A series of k runs is carried out with this decomposition, and in i^{th} iteration D_i is used as test data and other folds as training data
  - Thus, each tuple is used same number of times for training and once for testing.

- Overall estimate is taken as the average of estimates obtained from each iteration.



N-fold Cross-Validation

- In *k*-fold cross-validation method, $\frac{k-1}{N}$ part of the given data is used in training with *k*-tests.
- *N*-fold cross-validation is an extreme case of *k*-fold cross validation, often known as "Leave-one-out" cross-validation.
- Here, dataset is divided into as many folds as there are instances; thus, all most each tuple forming a training set, building *N* classifiers.
- In this method, therefore, *N* classifiers are built from *N-1* instances, and each tuple is used to classify a single test instances.
- Test sets are mutually exclusive and effectively cover the entire set (in sequence). This is as if trained by entire data as well as tested by entire data set.
- Overall estimation is then averaged out of the results of *N* classifiers.

N-fold Cross-Validation : Issue

- So far the estimation of accuracy and performance of a classifier model is concerned, the *N*-fold cross-validation is comparable to the others we have just discussed.
- The drawback of *N*-fold cross validation strategy is that it is computationally expensive, as here we have to repeat the run *N* times; this is particularly true when data set is large.
- In practice, the method is extremely beneficial with very small data set only, where as much data as possible to need to be used to train a classifier.

Accuracy Estimation

- We have learned how a classifier system can be tested. Next, we are to learn the metrics with which a classifier should be estimated.

- There are mainly to things to be measured for a given classifier
  - Accuracy
  - Performance

- **Accuracy estimation**
  - If N is the number of instances with which a classifier is tested and p is the number of correctly classified instances, the accuracy can be denoted as

$$\epsilon = \frac{p}{N}$$

  - Also, we can say the error rate (i.e., is misclassification rate) denoted by $\bar{\epsilon}$is denoted by

$$\bar{\epsilon} = 1 - \epsilon$$

Accuracy : True and Predictive

- Now, this accuracy may be true (or absolute) accuracy or predicted (or optimistic) accuracy.

- True accuracy of a classifier is the accuracy when the classifier is tested with all possible unseen instances in the given classification space.
  - However, the number of possible unseen instances is potentially very large (if it is not infinite)
  - For example, classifying a hand-written character
  - Hence, measuring the true accuracy beyond the dispute is impractical.

- Predictive accuracy of a classifier is an accuracy estimation for a given test data (which are mutually exclusive with training data).
  - If the predictive accuracy for test set is $\in$ and if we test the classifier with a different test set it is very likely that a different accuracy would be obtained.

  - The predictive accuracy when estimated with a given test set it should be acceptable without any objection

Predictive Accuracy

Example 11.1 : Universality of predictive accuracy

- Consider a classifier model M^D developed with a training set D using an algorithm M.
- Two predictive accuracies when M^D is estimated with two different training sets T_1 and T_2 are

  $(M^D)_{T1} = 95\%$

  $(M^D)_{T2} = 70\%$

- Further, assume the size of T_1 and T_2 are

  $|T_1| = 100$ records

  $|T_2| = 5000$ records.

- Based on the above mentioned estimations, neither estimation is acceptable beyond doubt.

Predictive Accuracy

- With the above-mentioned issue in mind, researchers have proposed two heuristic measures
  - Error estimation using **Loss Functions**
  - Statistical Estimation using **Confidence Level**
- In the next few slides, we will discus about the two estimations

Error Estimation using Loss Functions

- Let T be a matrix comprising with N test tuples

$$\begin{bmatrix} X_1 & y_1 \\ X_2 & y_2 \\ & \\ X_N & y_N \end{bmatrix}_{N\times(n+1)}$$

where X_i (i = 1, 2, …, N) is the n-dimensional test tuples with associated outcome y_i.

- Suppose, corresponding to (X_i, y_i), classifier produces the result (X_i, y_i')
- Also, assume that $(y_i - y_i')$ denotes a difference between y_i and y_i' (following certain difference (or similarity), (e.g., $(y_i - y_i') = 0$, if there is a match else 1)
- The two loss functions measure the error between y_i(the actual value) and y_i'(the predicted value) are

Absolute error: $|y_i - y_i'|$

Squred error: $|y_i - y_i'|^2$

Error Estimation using Loss Functions

- Based on the two loss functions, the test error (rate) also called generalization error, is defined as the average loss over the test set T. The following two measures for test errors are

  Mean Absolute Error (MAE): $\frac{\sum_{i=1}^{N} |y_i - y_i'|}{N}$

  Mean Squared Error(MSE):): $\frac{\sum_{i=1}^{N} (y_i - y_i')^2}{N}$

- Note that, MSE aggregates the presence of outlier.
- In addition to the above, a relative error measurement is also known. In this measure, the error is measured relative to the mean value $\tilde{y}$ calculated as the mean of y_i ($i = 1, 2, \ldots, N$) of the training data say D. Two measures are

  Relative Absolute Error (RAE: $\frac{\sum_{i=1}^{N} |y_i - y_i'|}{\sum_{i=1}^{N} |y_i - \tilde{y}|}$

  Relative Squared Error (RSE): $\frac{\sum_{i=1}^{N} (y_i - y_i')^2}{\sum_{i=1}^{N} (y_i - \tilde{y})^2}$

Performance Estimation of a Classifier

- Thus, when the classifier classified a test data set with imbalanced class distributions, then predictive accuracy on its own is not a reliable indicator of a classifier's effectiveness.
- This necessitates an alternative metrics to judge the classifier.
- Before exploring them, we introduce the concept of Confusion matrix.

Confusion Matrix

- A confusion matrix for a two classes (+, -) is shown below.

	C1	C2
C1	True positive	False negative
C2	False positive	True negative

	+	-
+	++	+-
-	-+	--

- There are four quadrants in the confusion matrix, which are symbolized as below.
  - True Positive (TP: f_{++}) : The number of instances that were positive (+) and correctly classified as positive (+v).
  - False Negative (FN: f_{+-}): The number of instances that were positive (+) and incorrectly classified as negative (-). It is also known as **Type 2 Error**.
  - False Positive (FP: f_{-+}): The number of instances that were negative (-) and incorrectly classified as (+). This also known as **Type 1 Error**.
  - True Negative (TN: f_{--}): The number of instances that were negative (-) and correctly classified as (-).

Confusion Matrix

Note:

- N_p = TP (f_{++}) + FN (f_{+-})
  = is the total number of positive instances.

- N_n = FP(f_{-+}) + Tn(f_{--})
  = is the total number of negative instances.

- $N = N_p + N_n$
  = is the total number of instances.

- (TP + TN) denotes the number of correct classification

- (FP + FN) denotes the number of errors in classification.

- For a perfect classifier FP = FN = 0, that is, there would be no Type 1 or Type 2 errors.

Confusion Matrix

Example 11.4: Confusion matrix

A classifier is built on a dataset regarding Good and Worst classes of stock markets. The model is then tested with a test set of 10000 unseen instances. The result is shown in the form of a confusion matrix. The result is self explanatory.

Class	Good	Worst	Total	Rate(%)
Good	6954	46	7000	99.34
Worst	412	2588	3000	86.27
Total	**7366**	**2634**	**10000**	**95.52**

Predictive accuracy?

Confusion Matrix for Multiclass Classifier

- Having m classes, confusion matrix is a table of size $m \times m$, where, element at (i, j) indicates the number of instances of class i but classified as class j.
- To have good accuracy for a classifier, ideally most diagonal entries should have large values with the rest of entries being close to zero.
- Confusion matrix may have additional rows or columns to provide total or recognition rates per class.

Confusion Matrix for Multiclass Classifier

Example 11.5: Confusion matrix with multiple class

Following table shows the confusion matrix of a classification problem with six classes labeled as C_1, C_2, C_3, C_4, C_5 and C_6.

Class	C1	C2	C3	C4	C5	C_6
C1	52	10	7	0	0	1
C2	15	50	6	2	1	2
C3	5	6	6	0	0	0
C4	0	2	0	10	0	1
C5	0	1	0	0	7	1
C_6	1	3	0	1	0	24

Predictive accuracy?

Confusion Matrix for Multiclass Classifier

- In case of multiclass classification, sometimes one class is important enough to be regarded as positive with all other classes combined togather as negative.

- Thus a large confusion matrix of m*m can be concised into 2*2 matrix.

Example 11.6: *m*×*m* CM to 2×2 CM

- For example, the CM shown in Example 11.5 is transformed into a CM of size 2×2 considering the class C_1 as the positive class and classes C_2, C_3, C_4, C_5 and C_6 combined together as negative.

Class	+	-
+	52	18
-	21	123

How we can calculate the predictive accuracy of the classifier model in this case?
Are the predictive accuracy same in both Example 11.5 and Example 11.6?

Performance Evaluation Metrics

- We now define a number of metrics for the measurement of a classifier.
  - In our discussion, we shall make the assumptions that there are only two classes: + (positive) and – (negative)
  - Nevertheless, the metrics can easily be extended to multi-class classifiers (with some modifications)

- **True Positive Rate (TPR):** It is defined as the fraction of the positive examples predicted correctly by the classifier.

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{f_{++}}{f_{++}+f_{+-}}$$

  - This metrics is also known as ***Recall***, ***Sensitivity*** or ***Hit rate***.

- **False Positive Rate (FPR):** It is defined as the fraction of negative examples classified as positive class by the classifier.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{f_{-+}}{f_{-+}+f_{--}}$$

  - This metric is also known as ***False Alarm Rate***.

Performance Evaluation Metrics

- **False Negative Rate (FNR):** It is defined as the fraction of positive examples classified as a negative class by the classifier.

$$FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = \frac{f_{+-}}{f_{++} + f_{+-}}$$

- **True Negative Rate (TNR):** It is defined as the fraction of negative examples classified correctly by the classifier

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = \frac{f_{--}}{f_{--} + f_{-+}}$$

  - This metric is also known as ***Specificity***.

Performance Evaluation Metrics

- **Positive Predictive Value (PPV):** It is defined as the fraction of the positive examples classified as positive that are really positive

$$PPV = \frac{TP}{TP + FP} = \frac{f_{++}}{f_{++} + f_{-+}}$$

  - It is also known as *Precision*.

- **F_1 Score (F_1):** Recall (r) and Precision (p) are two widely used metrics employed in analysis, where detection of one of the classes is considered more significant than the others.

  - It is defined in terms of (r or TPR) and (p or PPV) as follows.

$$F_1 = \frac{2r.p}{r + p} = \frac{2TP}{2TP + FP + FN}$$

$$= \frac{2f_{++}}{2f_{++} + f_{+-} + f_{-+}} = \frac{2}{\frac{1}{r} + \frac{1}{p}}$$

Note

- F_1 represents the harmonic mean between recall and precision
- High value of F_1 score ensures that both Precision and Recall are reasonably high.

Performance Evaluation Metrics

- More generally, F_β score can be used to determine the trade-off between Recall and Precision as

$$F_\beta = \frac{(\beta+1)rp}{r+\beta p} = \frac{(\beta+1)TP}{(\beta+1)TP+\beta FN+FP}$$

- Both, Precision and Recall are special cases of F_β when $\beta = 0$ and $\beta = 1$, respectively.

$$F_\beta = \frac{TP}{TP+FP} = Precision$$

$$F_\alpha = \frac{TP}{TP+FN} = Recall$$

Performance Evaluation Metrics

- A more general metric that captures Recall, Precision as well as is defined in the following.

$$F_\omega = \frac{\omega_1 TP + \omega_4 TN}{\omega_1 TP + \omega_2 FP + \omega_3 FN + \omega_4 TN}$$

Metric	ω_1	ω_2	ω_3	ω_4
Recall	1	1	0	1
Precision	1	0	1	0
F_β	β+1	β	1	0

Note

- In fact, given *TPR*, *FPR*, *p* and *r*, we can derive all others measures.
- That is, these are the universal metrics.

Predictive Accuracy (ε)

- It is defined as the fraction of the number of examples that are correctly classified by the classifier to the total number of instances.

$$\varepsilon = \frac{TP + TN}{P + N}$$

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

$$= \frac{f_{++} + f_{--}}{f_{++} + f_{+-} + f_{-+} + f_{--}}$$

- This accuracy is equivalent to F_w with $w_1 = w_2 = w_3 = w_4 = 1$.

Error Rate ($\bar{\varepsilon}$)

- The error rate $\bar{\varepsilon}$ is defined as the fraction of the examples that are incorrectly classified.

$$\bar{\varepsilon} = \frac{FP + FN}{P + N}$$

$$= \frac{FP + FN}{TP + TN + FP + FN}$$

$$= \frac{f_{+-} + f_{-+}}{f_{++} + f_{+-} + f_{-+} + f_{--}}$$

Note

$\bar{\varepsilon} = 1 - \varepsilon$.

Analysis with Performance Measurement Metrics

- Based on the various performance metrics, we can characterize a classifier.
- We do it in terms of TPR, FPR, Precision and Recall and Accuracy
- **Case 1: Perfect Classifier**

When every instance is correctly classified, it is called the perfect classifier. In this case, $TP = P$, $TN = N$ and CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \frac{P}{P} = 1$$

$$F_1 Score = \frac{2 \times 1}{1+1} = 1$$

$$Accuracy = \frac{P+N}{P+N} = 1$$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	0	N

Analysis with Performance Measurement Metrics

- **Case 2: Worst Classifier**

  When every instance is wrongly classified, it is called the worst classifier. In this case, $TP = 0$, $TN = 0$ and the CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{0}{N} = 0$$

$F_1 Score$ = Not applicable as $Recall + Precision = 0$

$$\text{Accuracy} = \frac{0}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	0	P
	-	N	0

Analysis with Performance Measurement Metrics

- **Case 3: Ultra-Liberal Classifier**

  The classifier always predicts the + class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{P}{P+N}$$

$$F_1 Score = \frac{2P}{2P+N}$$

$$\text{Accuracy} = \frac{P}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	N	0

Analysis with Performance Measurement Metrics

- **Case 4: Ultra-Conservative Classifier**

  This classifier always predicts the - class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \text{Not applicable}$$

$$(\text{as } TP + FP = 0)$$

$$F_1 Score = \text{Not applicable}$$

$$\text{Accuracy} = \frac{N}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	0	p
	-	0	N

Predictive Accuracy versus TPR and FPR

- One strength of characterizing a classifier by its *TPR* and *FPR* is that they do not depend on the relative size of *P* and *N*.
  - The same is also applicable for *FNR* and *TNR* and others measures from CM.
- In contrast, the *Predictive Accuracy*, *Precision*, *Error Rate*, F_1 *Score*, etc. are affected by the relative size of *P* and *N*.
- *FPR*, *TPR*, *FNR* and *TNR* are calculated from the different rows of the CM.
  - On the other hand Predictive Accuracy, etc. are derived from the values in both rows.
- This suggests that *FPR*, *TPR*, *FNR* and TNR are more effective than *Predictive Accuracy*, etc.

Pattern and Anomaly Detection Introduction

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of pattern recognition and anomaly detection
- Gain knowledge on example of polynomial curve fitting
- Learn about probability theory architecture and working model
- Understand Information theory

What is pattern?

- Pattern is all about it in this digital age.

- A pattern can be either visually identified or mathematically detected via the implementation of algorithms.

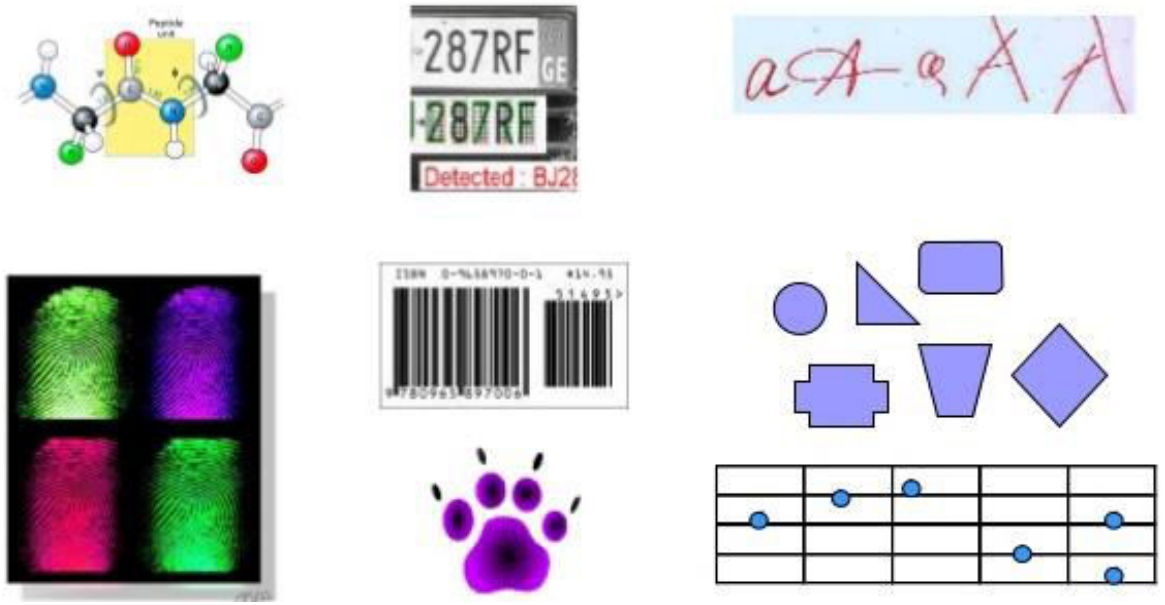


Figure: Pattern definition

Source: https://images.app.goo.gl/NmRdihnyFymA23uRA

What is pattern recognition?

- As per Wikipedia, pattern recognition is the automated recognition of patterns and regularities in data.

- It has applications in statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Pattern recognition techniques

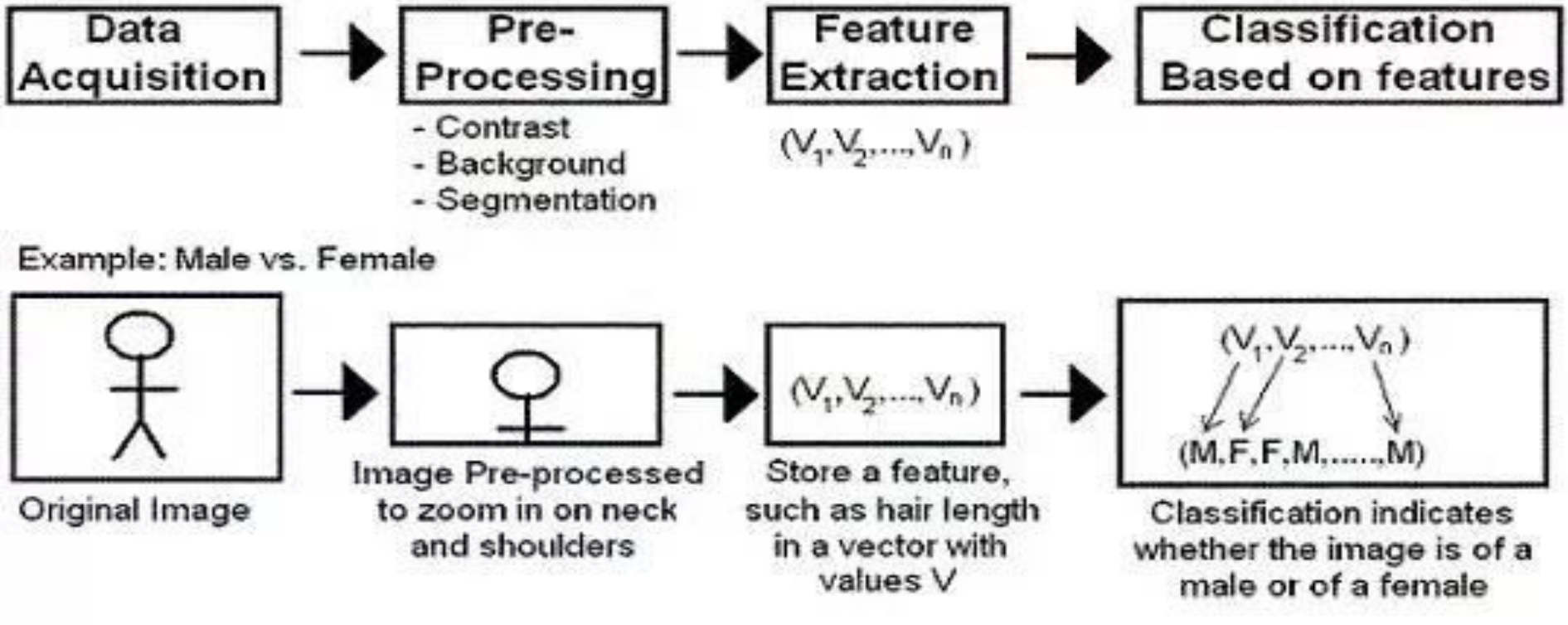


Figure: Pattern recognition process

Source: https://images.app.goo.gl/3x6gZVVao9u3vV8w7

Training and learning in pattern recognition

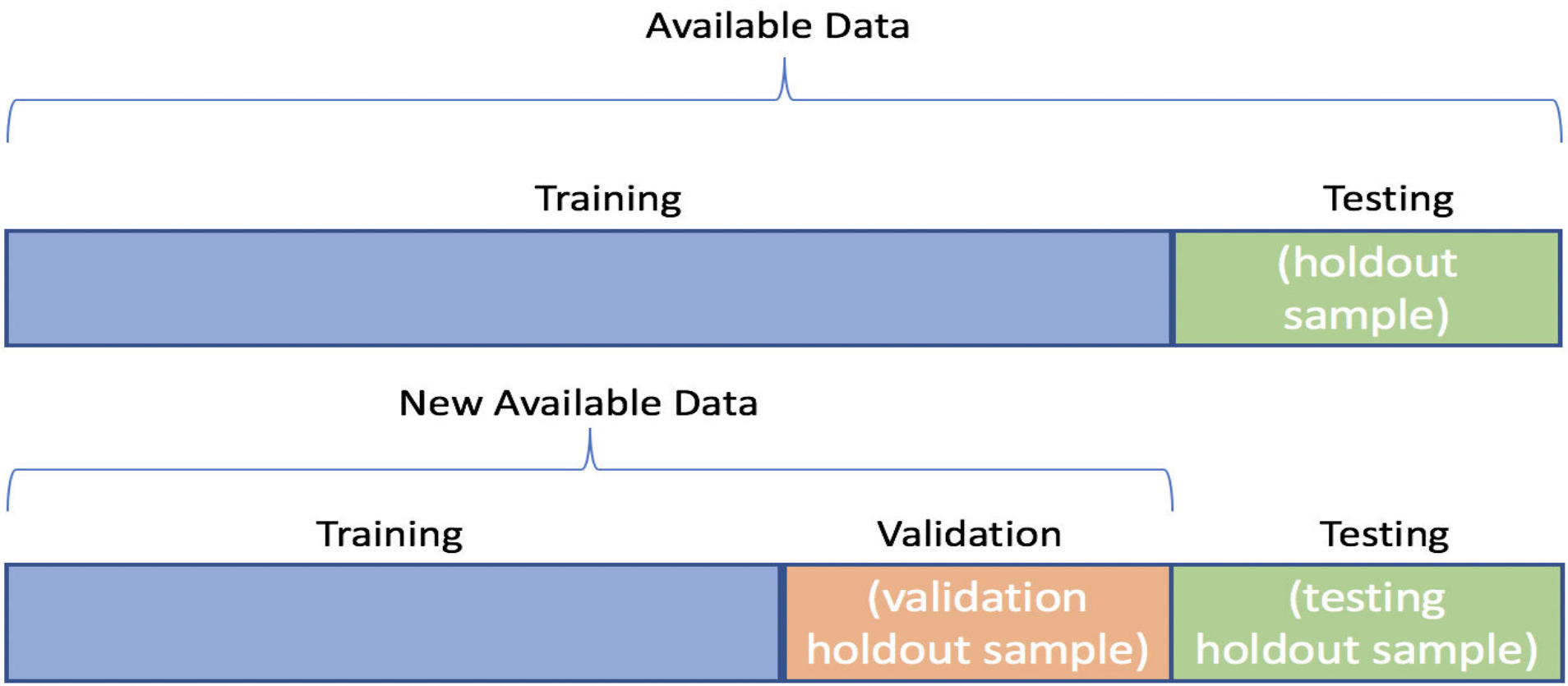


Figure: Training and testing dataset

Source: https://images.app.goo.gl/vetePfKYS2t7vmX99

Pattern recognition applications

- A model is an ideal concrete entity, or theoretical idea. The definition of the animal is an illustration when thinking regarding animal groups.
- The definition of the ball is a trend while speaking about various styles about balls.
- The groups can be baseball, cricket game, table tennis match, in sample case balls.
- Before approaching a new species, the species class has to be established.
- Choosing attributes and describing patterns is a very critical phase in classifying the layout.
- Effective presentation requires the use of non-discriminatory attributes, and the sample classification computational pressure.

Pattern recognition use cases

- Customer research and stock market analysis.
- Chat bots, NLP with text generation, text analysis, text translation.
- Optical Character Recognition (OCR), document classification and signature verification.
- Image recognition, visual search, face recognition.
- Voice recognition and ai assistants.
- Recommendation sentiment analysis, audience research.

What is anomaly detection?

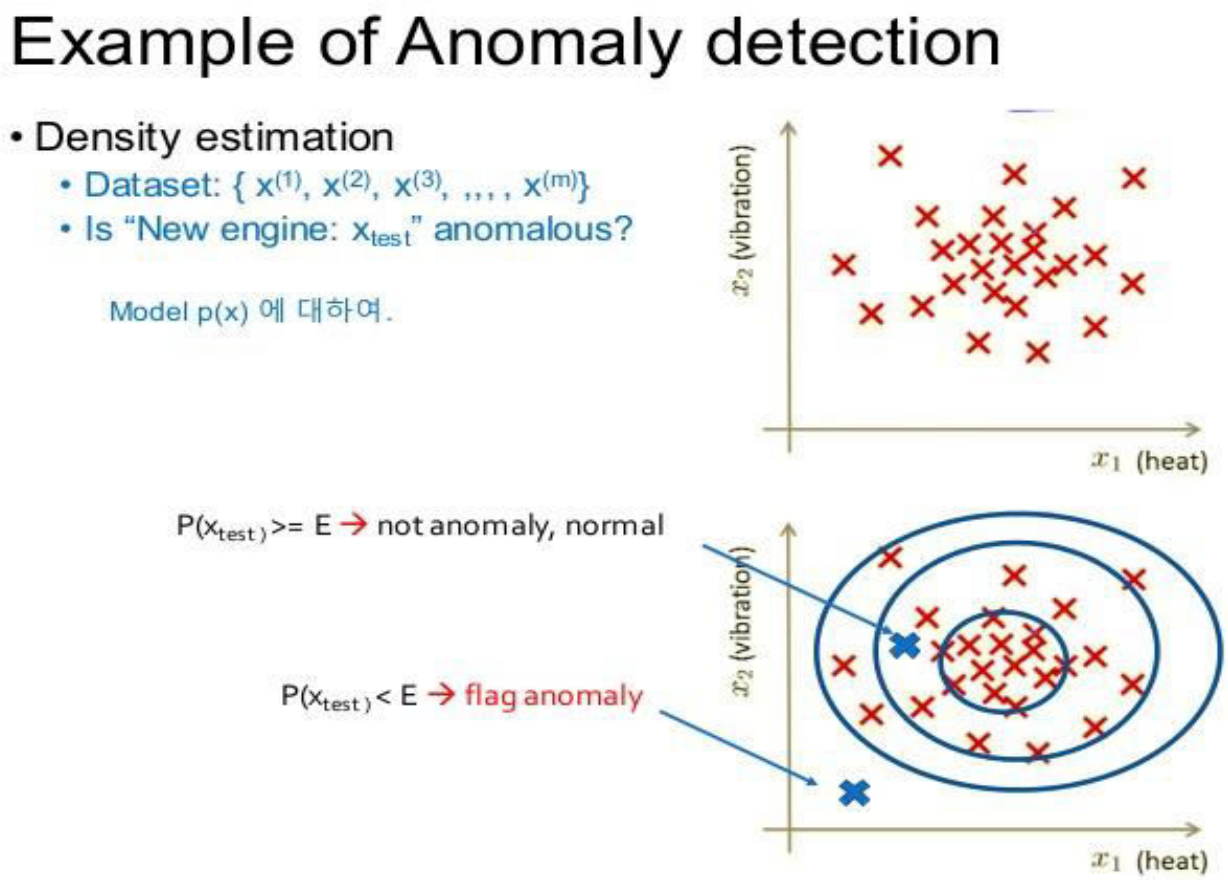


Figure: Anomaly detection example

Source: https://images.app.goo.gl/WpR4Rk1Xth1Fj4rt6

What are some other practical uses for anomaly detection?

- Traffic dropped or spiked.
- Transactions or revenue dropped.
- Traffic from social media increased or decreased.
- Traffic from organic search increased or decreased.

How is anomaly detection calculated over time?

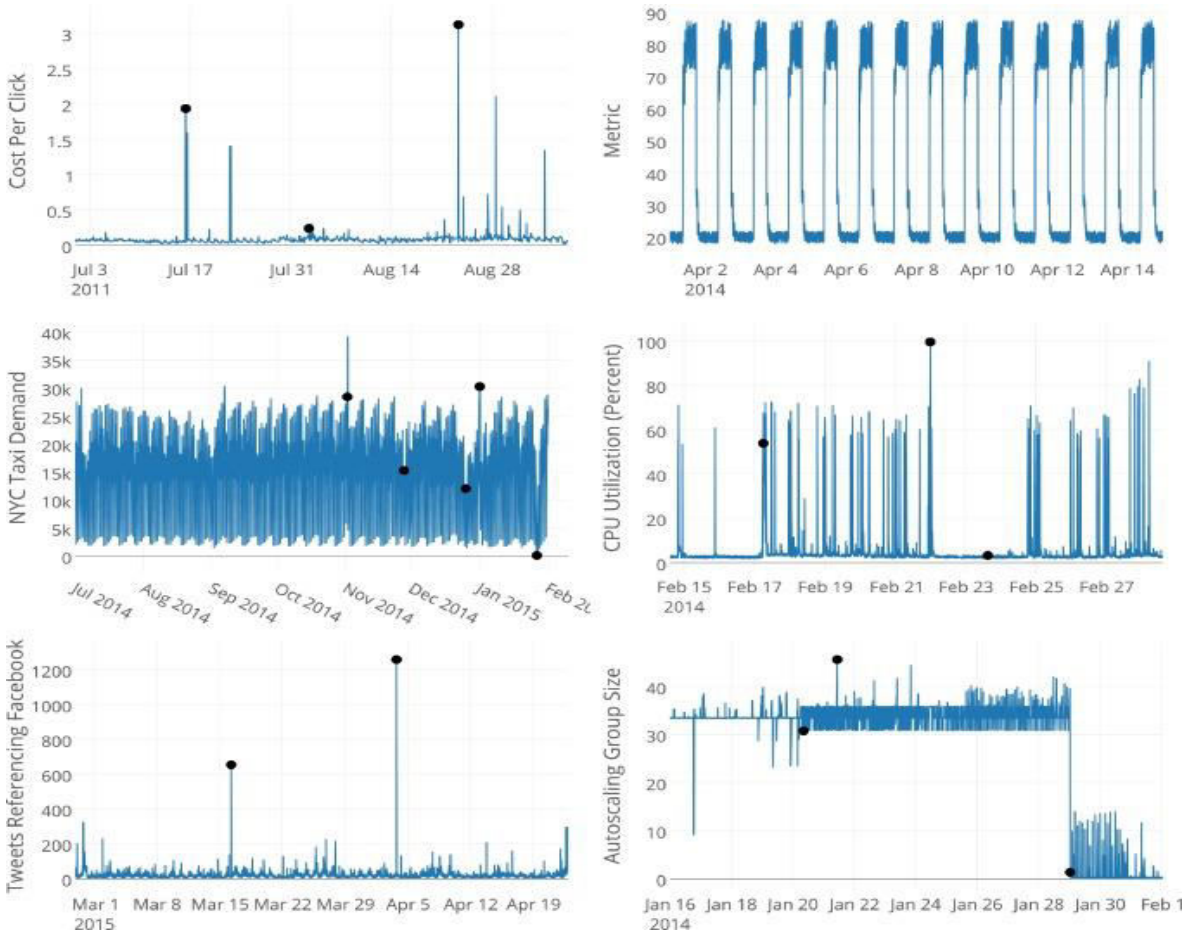


Figure: CPU time frame for anomaly detection

Source: https://images.app.goo.gl/a5wk76ZUmKZLj63v6

Self evaluation: Exercise 1

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 1: Polynomial curve fitting.

Key point for AI and ML anomaly detection

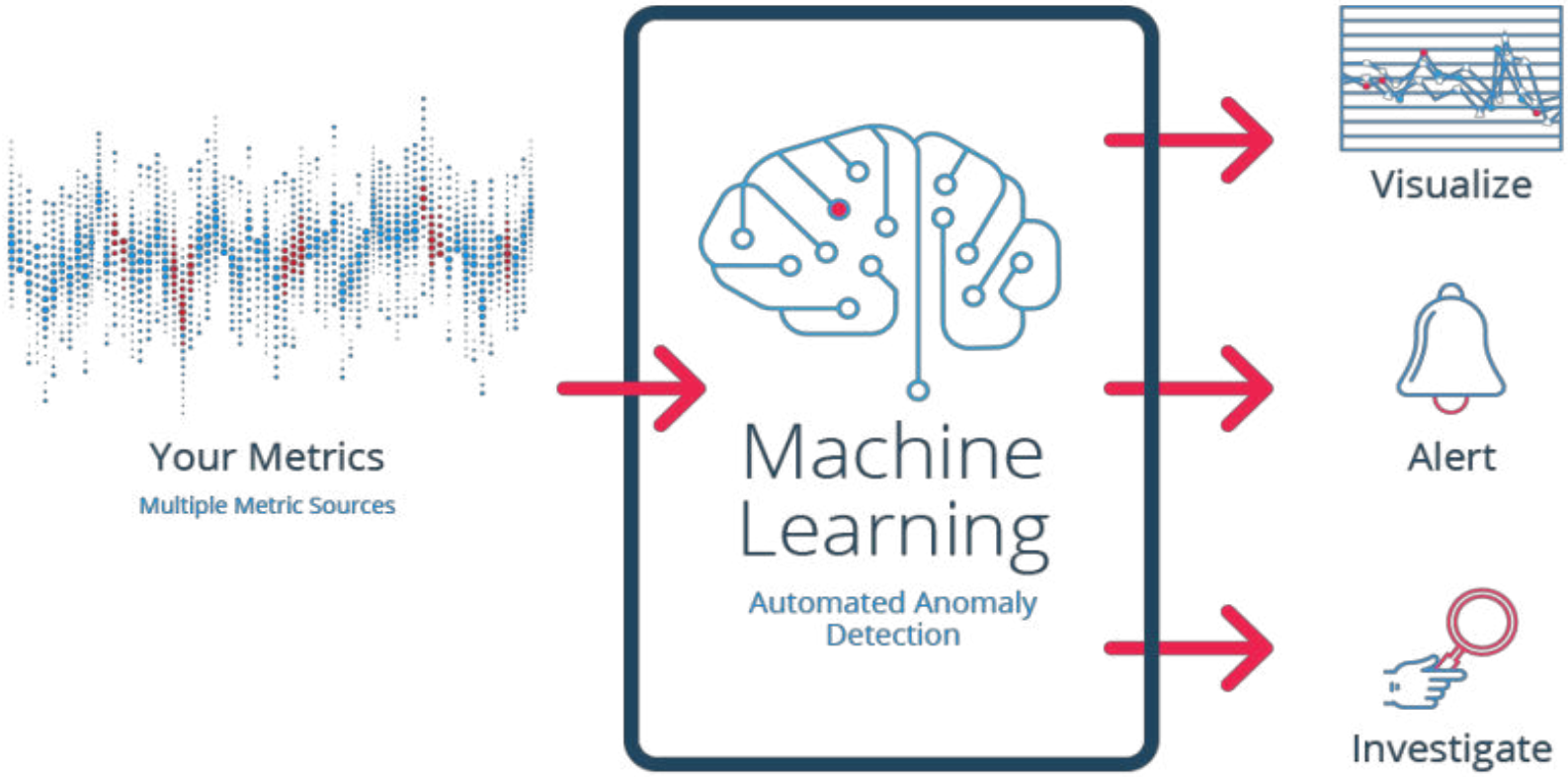


Figure: Anomaly detection flow with Machine learning

Source: https://images.app.goo.gl/ewiV18h8cpAkEyCL9

Tasks for artificial intelligence

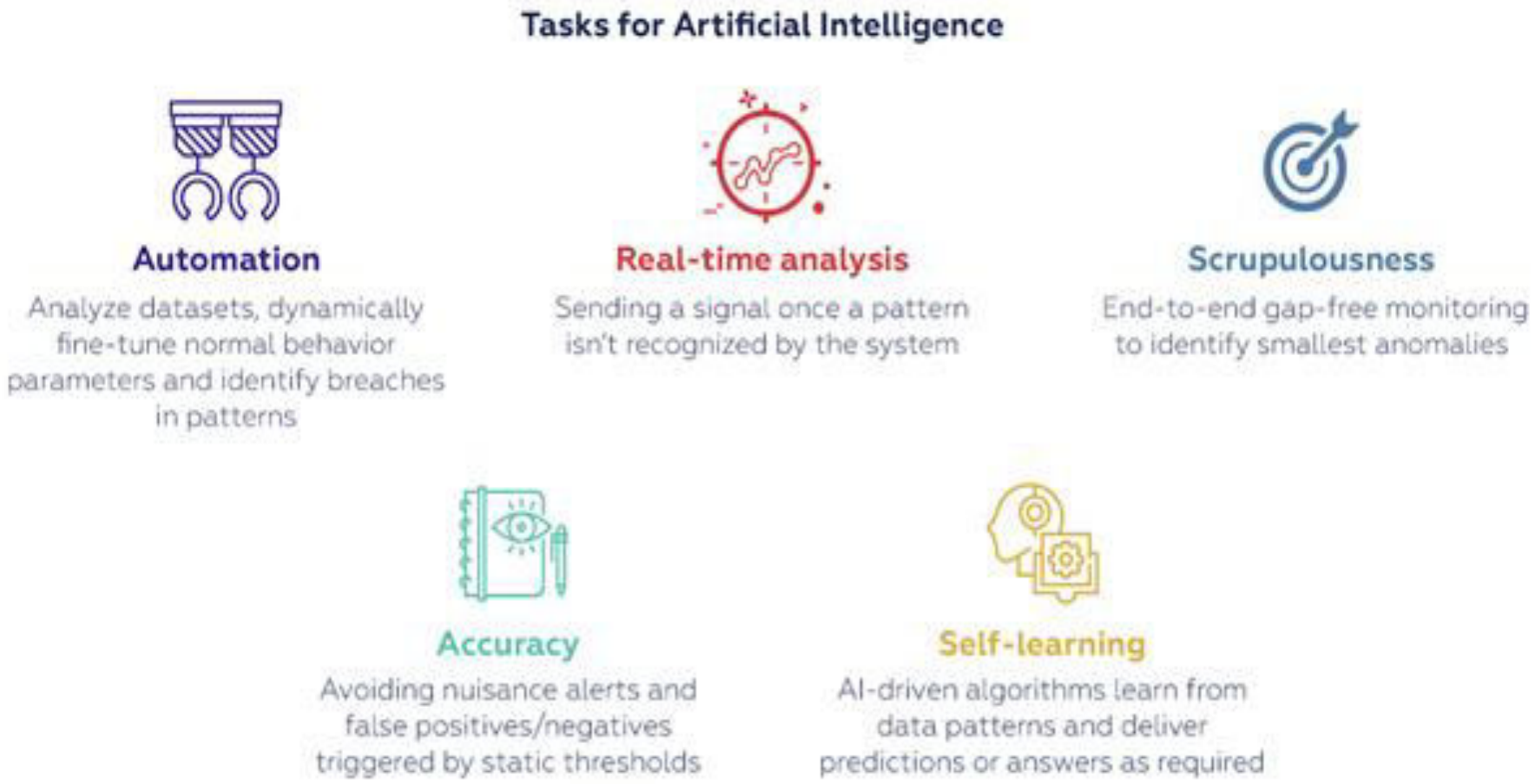


Figure: AI Task

Source: https://images.app.goo.gl/F9Z664sSWR53yGqLA

AI system learning process (1 of 2)

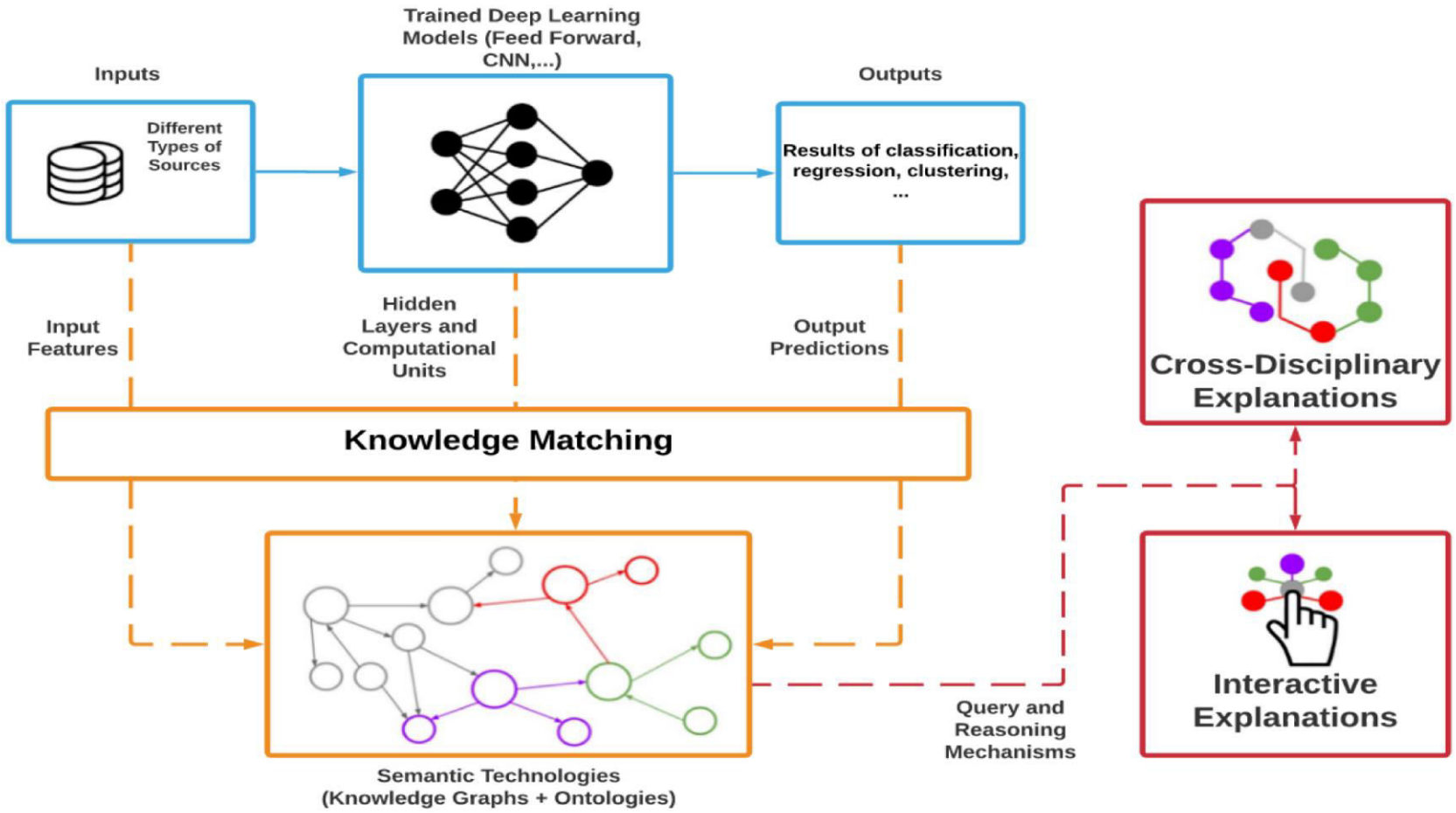


Figure: AI Learning Process

Source: https://images.app.goo.gl/dBtnk7CPRtqe7xM59

AI system learning process (2 of 2)

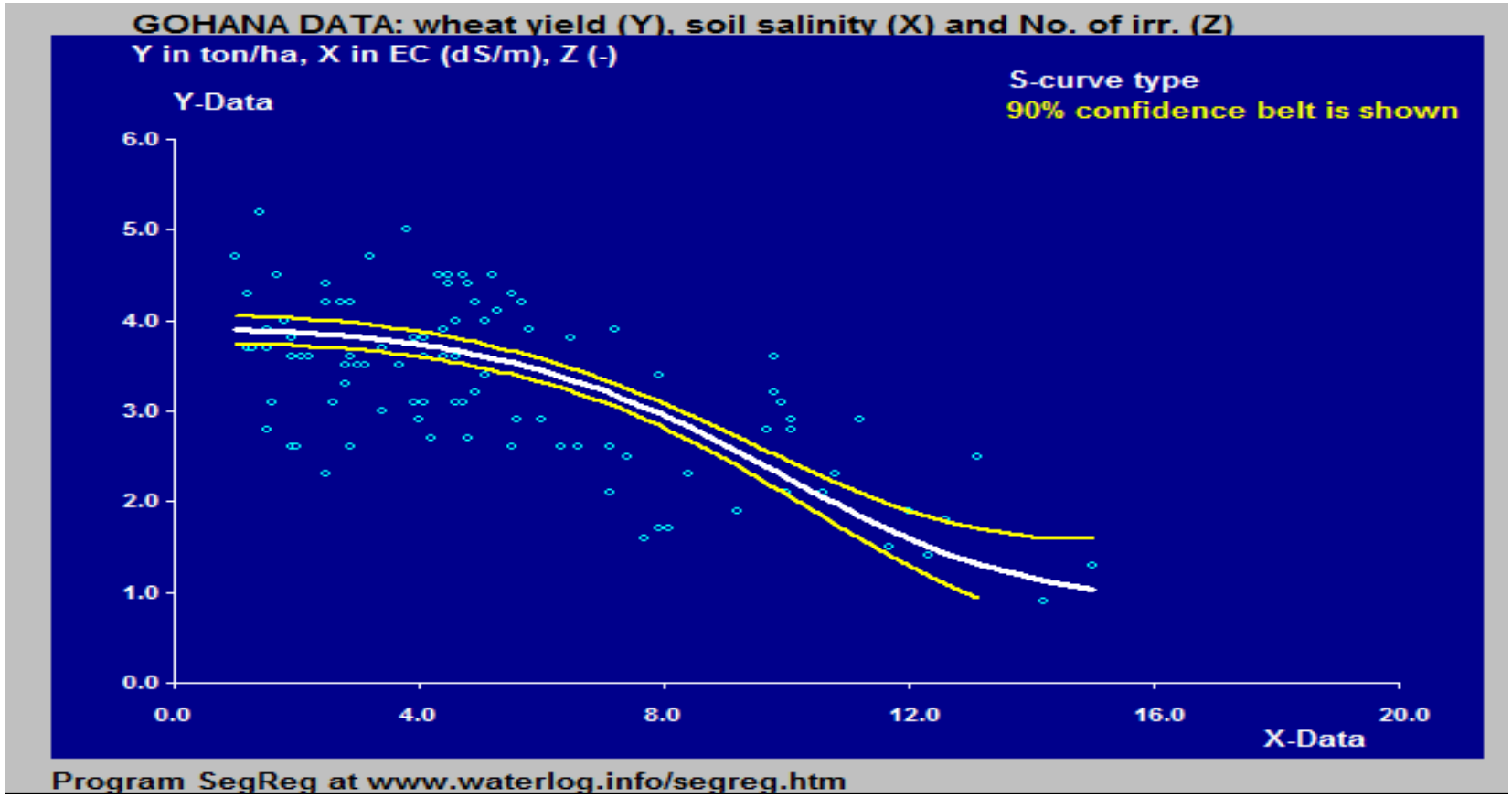


Figure: Relation between wheat yield and soil salinity

Source: https://upload.wikimedia.org/wikipedia/commons/thumb/4/46/Gohana_inverted_S-curve.png/560px-Gohana_inverted_S-curve.png

Self evaluation: Exercise 2

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 2: Probability and distribution.

Test to geometric requirements for curves algebraic

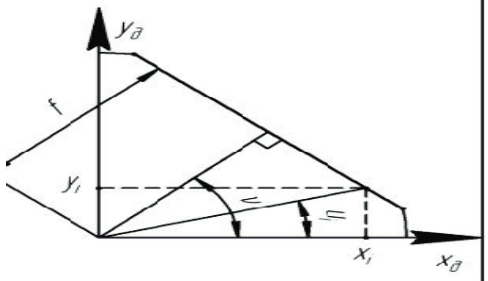
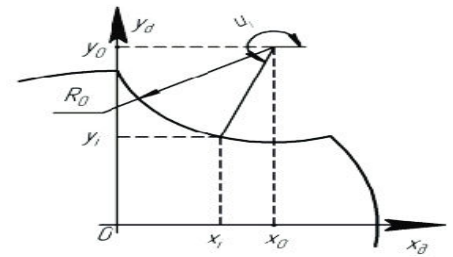
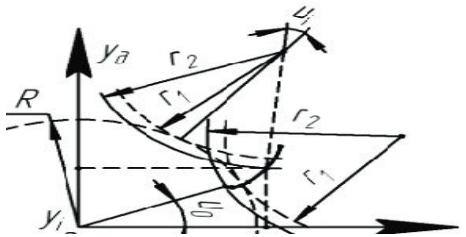
Scheme of curve plotting	Equation of a curve in a parametrical look
1	2
	Straight line: $x_i = \frac{f}{\cos(u_i - \nu)} \cdot \cos(u_i)$ $y_i = \frac{f}{\cos(u_i - \nu)} \cdot \sin(u_i)$
	Circle arch: $x_i = x_0 + R_0 \times \cos(u_i)$ $y_i = y_0 + R_0 \times \sin(u_i)$
	Epicycloid: $x_i = (R + r_1)\cos(u_i + u_0) -$ $-r_2 \cos\left[\left(\frac{R + r_1}{r_1}\right)(u_i + u_0)\right]$ $y_i = (R + r_1)\sin(u_i + u_0) -$ $-r_2 \sin\left[\left(\frac{R + r_1}{r_1}\right)(u_i + u_0)\right]$

Figure: Algebraic curves in a parametrical form used for creation of the forming line in the edge section of different rated surfaces

Source: https://images.app.goo.gl/Rn75dzFLCv9y6nu5A

Curves matched to data points (1 of 2)

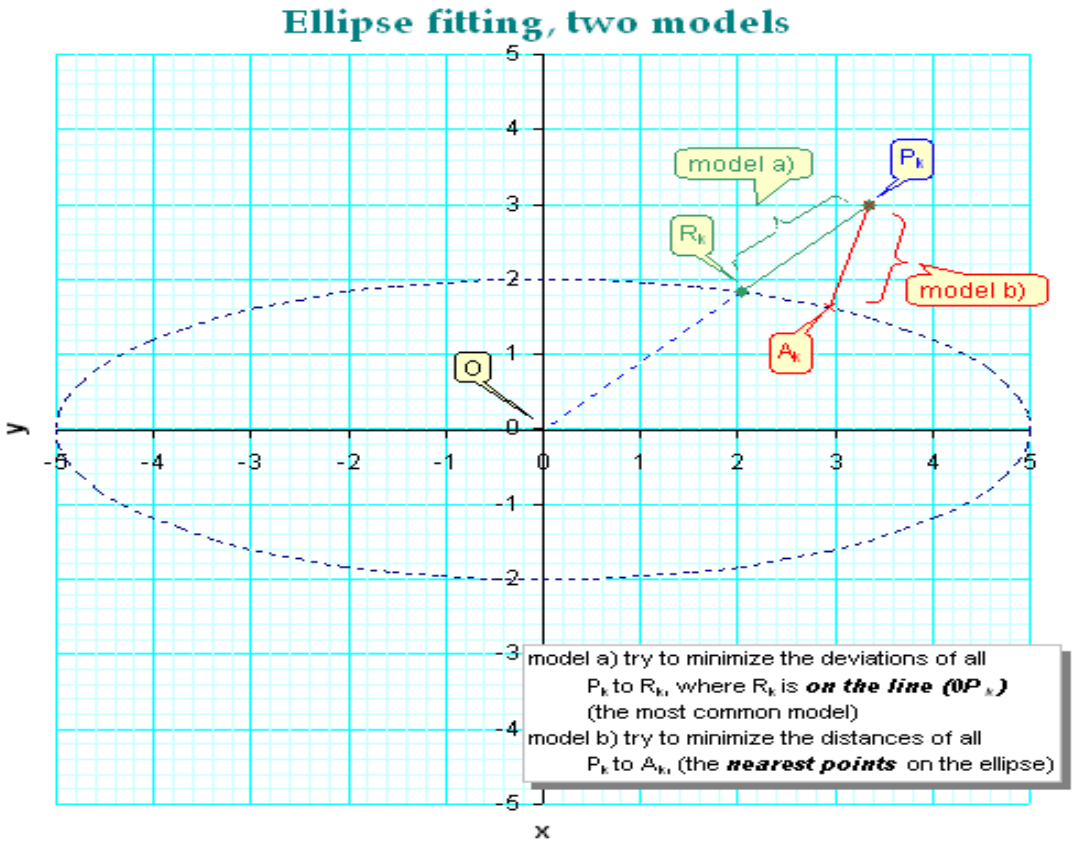


Figure: Different models of ellipse fitting

Source: https://lh3.googleusercontent.com/HkRug5Yd6SlGy0AkSgLZ9FYwrq3Os5jeSoEiHqg5ft1se9C8uSUcXjY9p3yfYfhg13eyUA=s86

Curves matched to data points (2 of 2)

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.977411977					
R Square	0.955334173					
Adjusted R Square	0.952025593					
Standard Error	13.21744878					
Observations	30					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	100887.8743	50443.93714	288.7444885	5.95206E-19	
Residual	27	4716.925714	174.7009524			
Total	29	105604.8				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	21.92	10.57395903	2.073017301	0.047839767	0.224028187	43.61597181
Month	-24.5485714	6.91777293	-3.54862348	0.001441196	-38.742669	-10.3544738
MonSq	8.057142857	0.967418193	8.328500456	6.14519E-09	6.072164688	10.04212103

Figure: Linear regression output

Figure: Quadratic regression output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.91683485					
R Square	0.840586142					
Adjusted R Square	0.83489279					
Standard Error	24.52030396					
Observations	30					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	88769.93143	88769.93143	147.6434502	1.1088E-12	
Residual	28	16834.86857	601.2453061			
Total	29	105604.8				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-53.28	10.20861661	-5.21912048	1.52374E-05	-74.1914032	-32.3685968
Month	31.85142857	2.621330755	12.15086212	1.1088E-12	26.48187593	37.22098121

Case study: Anomaly detection with IBM Watson

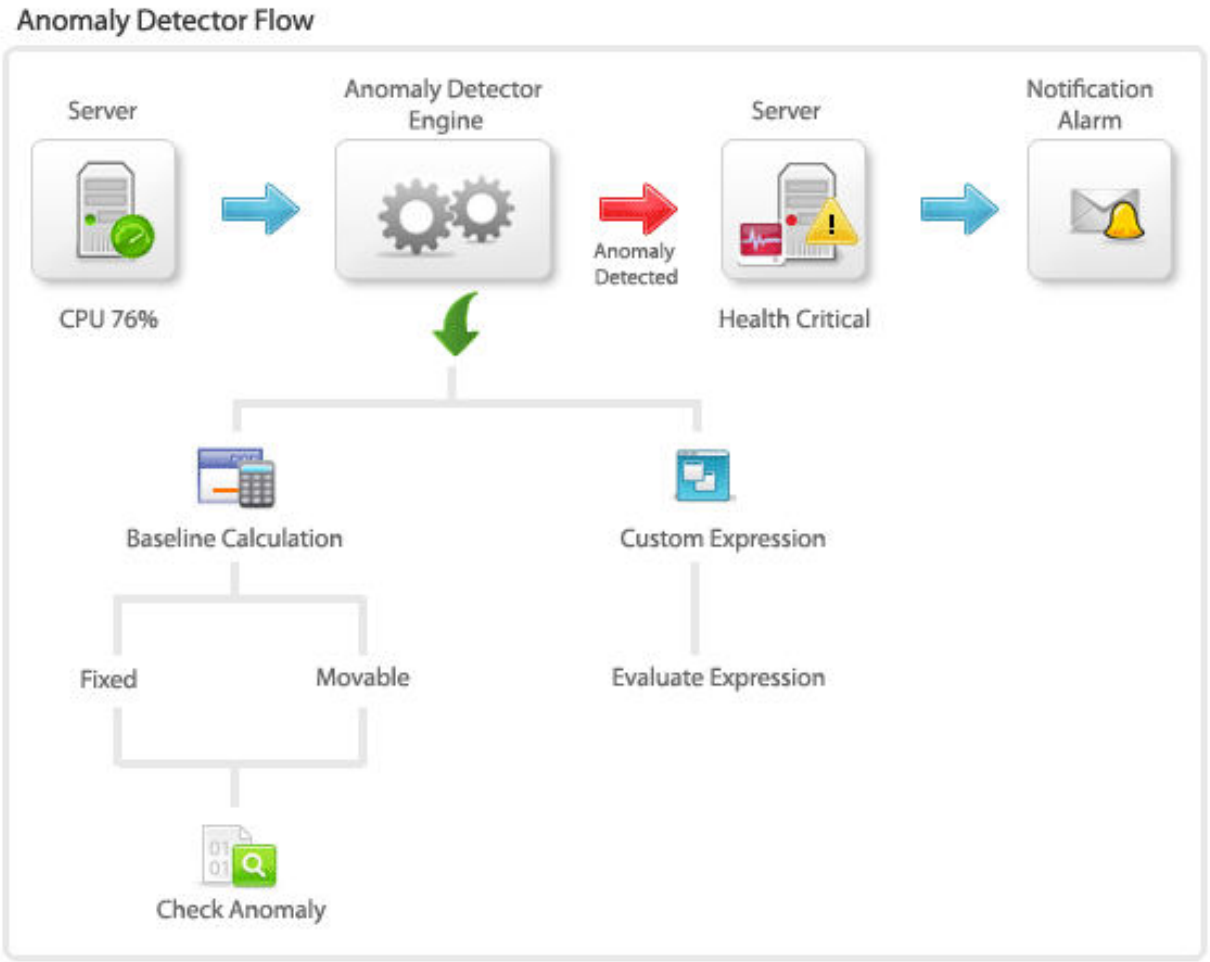


Figure: Anomaly detection workflow engine

Source: https://images.app.goo.gl/ukrNQnHbjnP5XXKf6

Self evaluation: Exercise 3

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 3: Simple linear regression.

Probability theory (1 of 2)

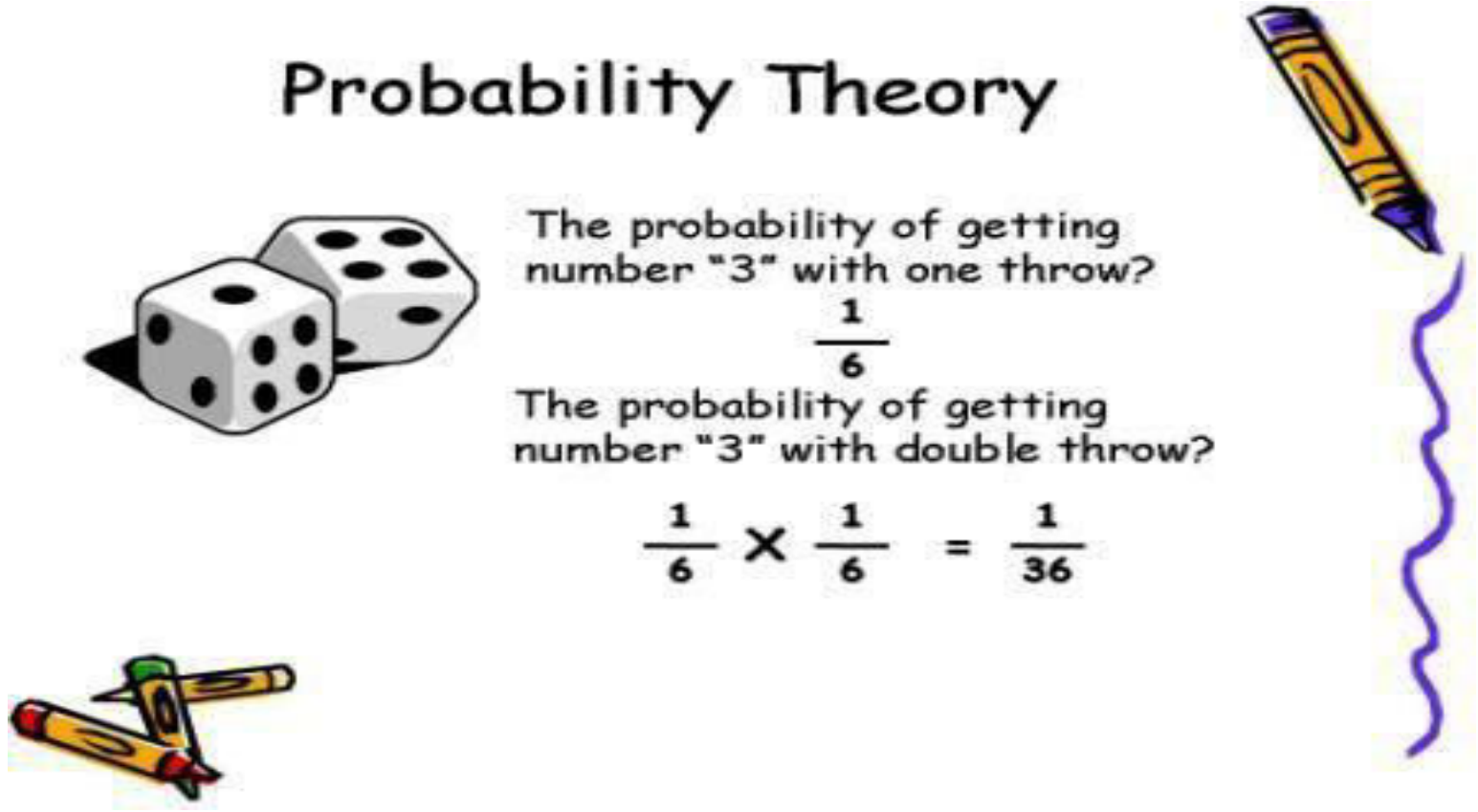


Figure: Probability Theory

Source: https://images.app.goo.gl/DKudJzQZCPZEQyvt6

Probability theory (2 of 2)

- Sample Space: 12 There are 12 marbles total (4+5+1+2 = 12)

Probability= Total Possible outcome

- P (black) = 2/12 = 1/6 There are 2 black marbles in the bag, 12 is your sample space.

.

- P (blue) = 4/12 = 1/3 There are 4 blue marbles in the bag , 12 is your sample space.

- P (blue or black) = 6/12= 1/2 4 blue + 2 black = 6 , 12 is your sample space.

- P (not green) = 11/12 There's 1 green. So 12-1 = 11 that aren't green,12 is your sample space.

- P (not purple) = 1

- I will select a marble that is not purple because there are no purple marbles in the bag. Whenever the chance of something occurring is definite, the probability is i.

Maximum likelihood theory and estimation (1 of 2)

- The estimate of density is the issue of evaluating the distribution of likelihood for a sub-set of a sample in a question domain.

- Two Important concepts:
  - Probability density estimation problem.
  - Maximum likelihood calculation.

Maximum likelihood theory and estimation (2 of 2)

- Suppose that we are given a sequence (x1….xn) of IID random variables and a priori distribution of it is given by Us wish to find the MAP estimate of it. Note that the normal distribution is its own conjugate prior, so we will be able to find a closed-form solution analytically.

- The function to be maximized is then given by:

$$f(\mu)f(x \mid \mu) = \pi(\mu)L(\mu) = \frac{1}{\sqrt{2\pi}\sigma_m}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_m}\right)^2\right)\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma_v}\right)^2\right),$$

- Which is equivalent to minimizing the following function of it:

$$\sum_{j=1}^{n}\left(\frac{x_j-\mu}{\sigma_v}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_m}\right)^2.$$

- Thus, we see that the MAP estimator for p is given by:

$$\hat{\mu}_{\mathrm{MAP}} = \frac{\sigma_m^2 n}{\sigma_m^2 n+\sigma_v^2}\left(\frac{1}{n}\sum_{j=1}^{n}x_j\right)+\frac{\sigma_v^2}{\sigma_m^2 n+\sigma_v^2}\mu_0 = \frac{\sigma_m^2\left(\sum_{j=1}^{n}x_j\right)+\sigma_v^2\mu_0}{\sigma_m^2 n+\sigma_v^2}.$$

- Which turns out to be a linear interpolation between the prior mean and the sample mean weighted by their respective covariance's. The case of is called a non-informative prior and leads to an ill-defined a priori probability distribution.

Self evaluation: Exercise 4

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.
- You are instructed to write the following activities using python code.
- Exercise 4: Multiple linear regression.

Model selection (1 of 2)

- The MDL (Minimum Description Length) statistic is calculated as follows:

$$MDL = L(h) + L(D \mid h)$$

- Where h is the model, D is the predictions made by the model, L(h) is the number of bits required to represent the model, and L(D | h) is the number of bits required to represent the predictions from the model on the training dataset.

- The score as defined above is minimized, e.g., the model with the lowest MDL is selected.

- The number of bits required to encode (D | h) and the number of bits required to encode (h) can be calculated as the negative log-likelihood. For example:

$$MDL = -\log(P(\text{theta})) - \log(P(y \mid X, \text{theta}))$$

- Or the negative log-likelihood of the model parameters (theta) and the negative log-likelihood of the target values (y) given the input values (X) and the model parameters (theta).

Model selection (2 of 2)

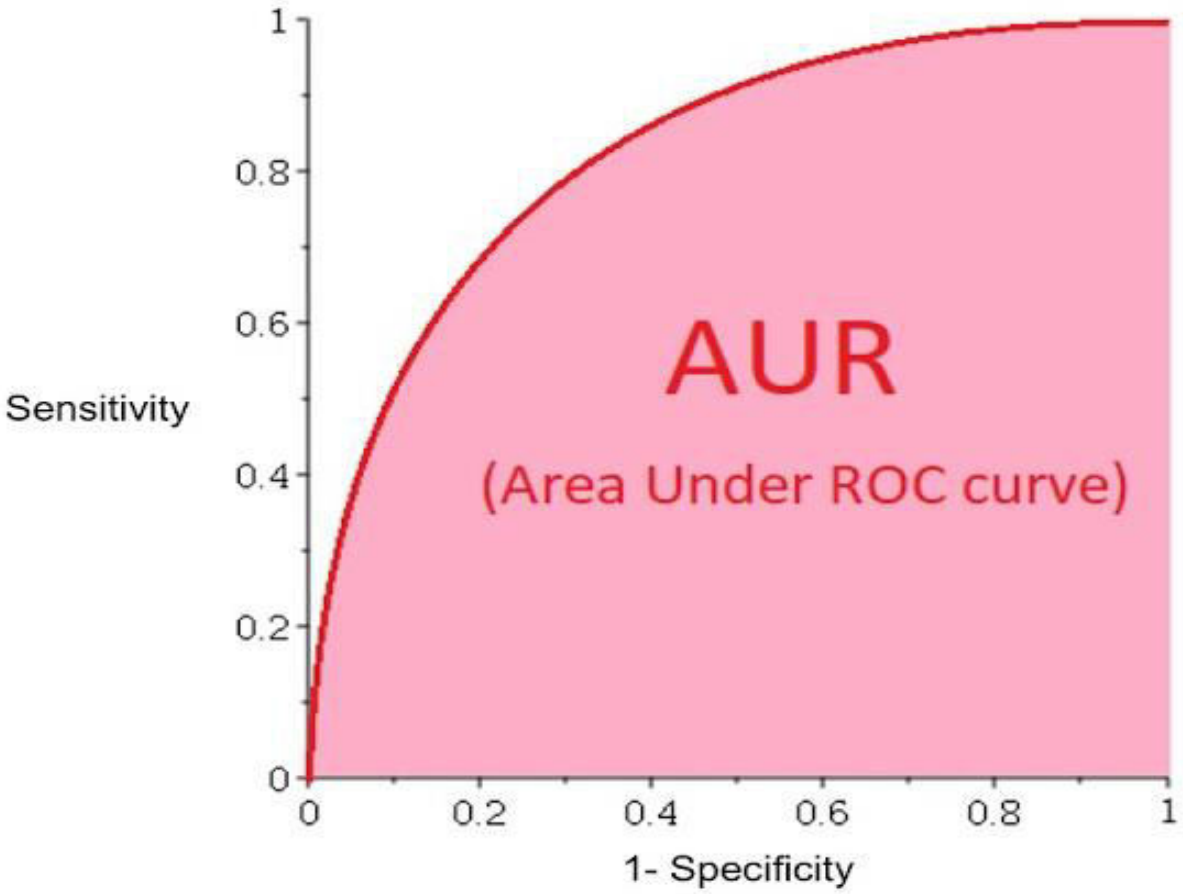


Figure: AU ROC

Source: https://images.app.goo.gl/iMXwj7jLYADko1uCA

Matrices of uncertainty (confusion matrices)

- A few other metrics are computed from these values:

- Accuracy: How often is the classifier correct? $\left(\frac{TP+TN}{total}\right)$
- Misclassification rate (or "error rate"): How often is the classifier wrong? $\left(\frac{FP+FN}{total} = 1 - accuracy\right)$
- Recall (or "sensitivity" or "true positive rate"): How often are positive-labeled samples predicted as positive? $\left(\frac{TP}{num\ positive\text{-}labeled\ examples}\right)$
- False positive rate: How often are negative-labeled samples predicted as positive? $\left(\frac{FP}{num\ negative\text{-}labeled\ examples}\right)$
- Specificity (or "true negative rate"): How often are negative-labeled samples predicted as negative? $\left(\frac{TN}{num\ negative\text{-}labeled\ examples}\right)$
- Precision: How many of the predicted positive samples are correctly predicted? $\left(\frac{TP}{TP+FP}\right)$
- Prevalence: How many labeled-positive samples are there in the data? $\left(\frac{num\ positive\text{-}labeled\ examples}{num\ examples}\right)$

Loss of logging (log-loss)

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Figure: Log Loss formula

Source: https://images.app.goo.gl/UpFaWENnNrSm935R9

Rate for F1 (F1 score)

- The F1 score is the weighted average accuracy and warning, even the balanced F-score or F-measure.

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Figure: F1 score

Metric selection

- Metric selection is more complex for biased groups (or strongly predetermined bias data).
- For example, you have a dataset with only 0.5% of the data in category 1.
- You run your experiment and remember that 99.5 percent of the tests are correctly graded.

Hyperparameter selection (1 of 2)

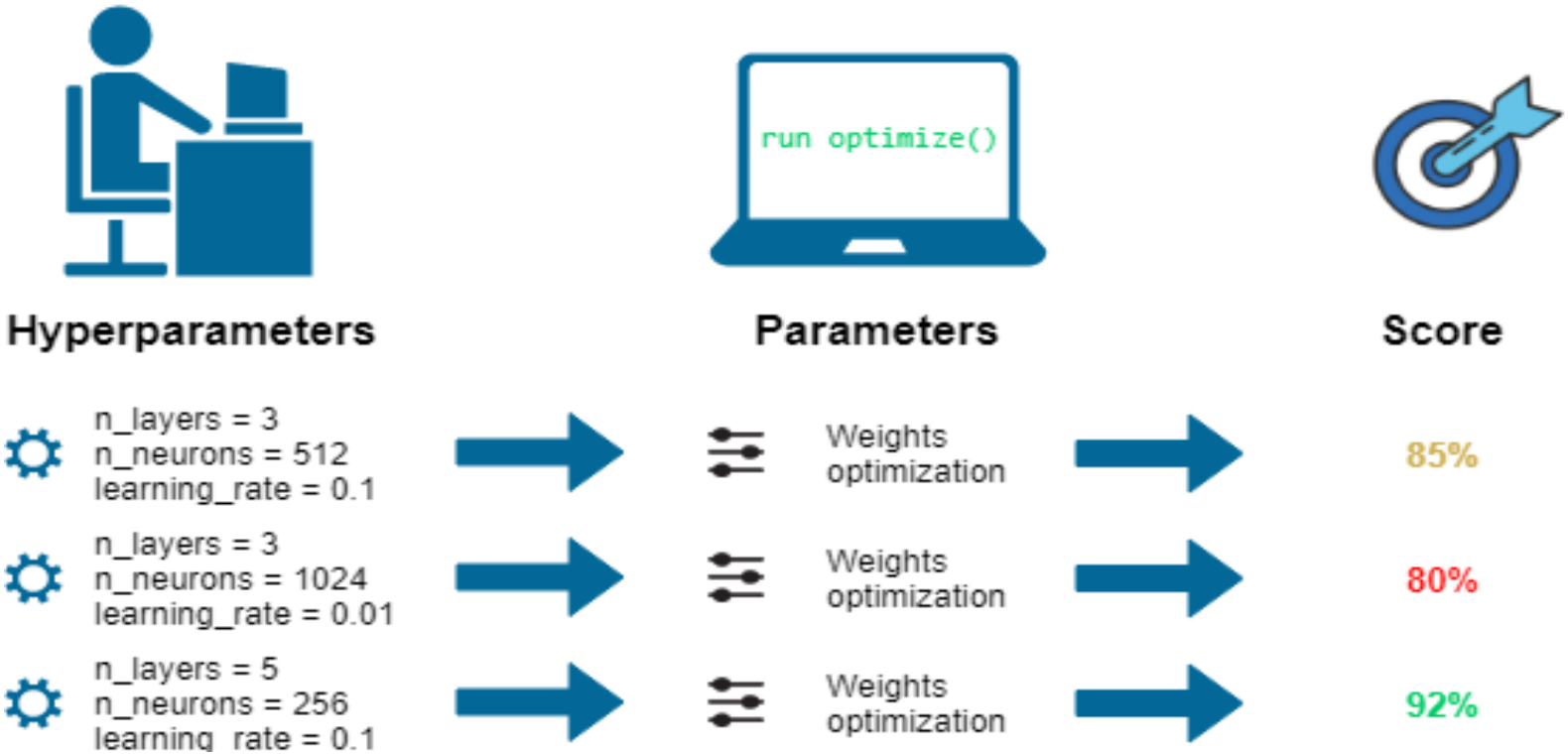


Figure: Hyperparameter selection

Source: https://images.app.goo.gl/o3DaxpbGLPVRxKnPA

Hyperparameter selection (2 of 2)

- Optimization of Bayesian hyperparameter.

- There are two parts:
  - Exploration: Test the feature with the most uncertain outcome in collection of hyperparameters.
  - Operating: Test this function on a set of high-value hyperparameters.

The problem with high dimensionality

- The dimension of a problem refers to the number of input variables (actually, degrees of freedom).

- The exponential increase in data required to densely populate space as the dimension increases.

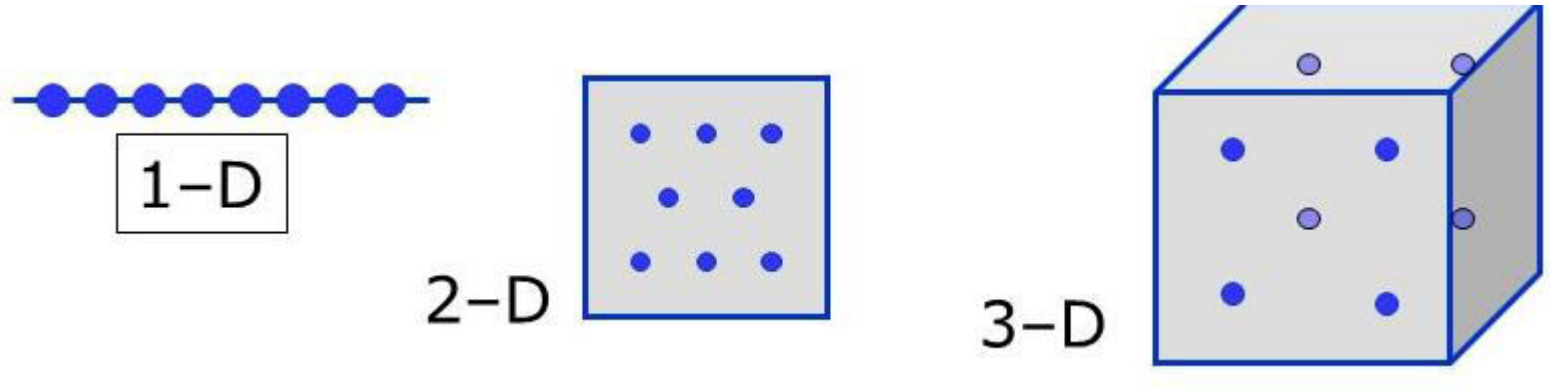


Figure: The Problem with High Dimensionality

Source: https://images.app.goo.gl/T6QLvXses34dSeXXA

Information theory

Information:

$$I(x) = -\log P(x). \tag{3.48}$$

Entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]. \tag{3.49}$$

KL divergence:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]. \tag{3.50}$$

Figure: Information theory Formulas

Source: https://images.app.goo.gl/DnbEL1fyXk23ory49

Self evaluation: Exercise 5

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 5: Logistic regression model.

Checkpoint (1 of 2)

Multiple choice questions:

1. The recalled output in pattern association problem depends on?
   a) Nature of input-output
   b) Design of network
   c) Both input & design
   d) None of the mentioned

2. What is the objective of feature maps?
   a) To capture the features in space of input patterns
   b) To capture just the input patterns
   c) Update weights
   d) To capture output patterns

3. Use of nonlinear units in the feedback layer of competitive network leads to concept of?
   a) Feature mapping
   b) Pattern storage
   c) Pattern classification
   d) None of the mentioned

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. The recalled output in pattern association problem depends on?
   a) Nature of input-output
   b) Design of network
   c) Both input & design
   d) None of the mentioned

2. What is the objective of feature maps?
   a) To capture the features in space of input patterns
   b) To capture just the input patterns
   c) Update weights
   d) To capture output patterns

3. Use of nonlinear units in the feedback layer of competitive network leads to concept of?
   a) Feature mapping
   b) Pattern storage
   c) Pattern classification
   d) None of the mentioned

Checkpoint (2 of 2)

Fill in the blanks:

1. _____learning is involved in pattern clustering task.
2. If the weight matrix stores the given patterns, then the network becomes _____.
3. Activation models are _____.
4. Information theory is using in _____ detection.

True or False:

1. From given input-output pairs pattern recognition model should capture characteristics of the system? True/False
2. Can system be both interpolative & accretive at same time? True/False
3. Does pattern classification belong to category of non-supervised learning? True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Unsupervised** learning is involved in pattern clustering task.
2. If the weight matrix stores the given patterns, then the network becomes **auto associative memory.**
3. Activation models are **deterministic.**
4. Information theory is using in **pattern** detection.

True or False:

1. From given input-output pairs pattern recognition model should capture characteristics of the system? **True**
2. Can system be both interpolative & accretive at same time? **False**
3. Does pattern classification belong to category of non-supervised learning? **False**

Question bank

Two mark questions:

1. What is pattern detection?
2. What is information theory?
3. What is linear regression model?
4. What is the math formula for curve designing?

Four mark questions:

1. What is the difference between pattern and anomaly detection?
2. What is polynomial curve fitting?
3. Describe high dimensionality problems.
4. Describe information theory components.

Eight mark questions:

1. Explain model selection techniques.
2. Explain probability theory in details.

Unit summary

After completing this unit, you should be able to:

- Understand the concept of pattern recognition and anomaly detection
- Gain knowledge on example of polynomial curve fitting
- Learn about probability theory architecture and working model
- Understand Information theory

Statistical Approaches for Pattern Recognition

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of probability distributions
- Gain knowledge on example of statistical approaches
- Understand linear models for regression
- Learn about linear models for classification

Understanding statistics

- Statistics is a method of statistical research utilizing computational models, descriptions, and excerpts for theoretical or real-life data studies.

- Statistics is the analysis of how to draw inference from the evidence, feedback, and conclusions. Such statistical indicators involve.

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

T-test

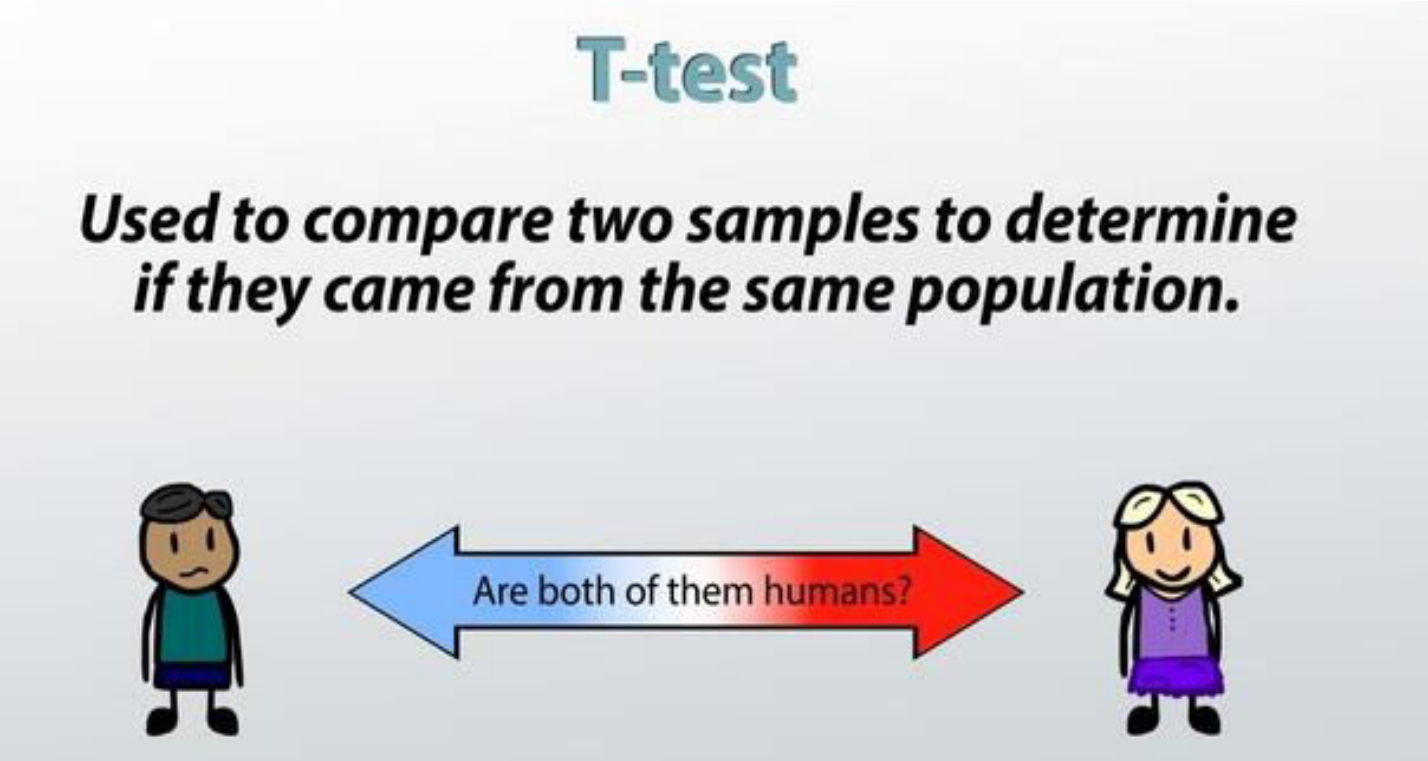


Figure: T-test example

Source: https://images.app.goo.gl/PAJFuF1hSWmx9y5o7

Z-test

- It is a method that tests how two population variables differ in the defined variances and the sample size.
- It is believed that the test statistics have a normal distribution.
- To do an accurate z-test, irritating parameters such as standard deviation need to be understood.
- Z-statistical or z-value is a sum that represents the amount, over or below the general population of standard variations generated by the Z-test.

Self evaluation: Exercise 6

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 6: Polynomial regression for classification.

Z-test and t-test difference

- A typical and simplified method of statistical research is a z-test that measures the statistical validity of a sample mean to the predicted mean population but needs awareness of the standard model variance, which is not always feasible.

- The t-test is a more practical form of study since it needs just the standard deviation of the sample in comparison to the norm of the population.

P-value

- In mathematics, the p-value is the chance of producing outcomes as extreme as the findings of a mathematical experiment test obtained, given the null hypothesis is accurate.

- A lower p-value indicates better support for the alternate hypothesis.

Descriptive statistics

- Descriptive statistics are short descriptive equations summarizing a certain collection of results, which may either reflect the whole population or a subset of a community.

- Descriptive statistics are divided into core pattern measurements and volatility measurements (spread).

- Measures of central tendency include the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables, and the kurtosis and skewness.

Self evaluation: Exercise 7

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection , it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 7: Neural networks.

Type I error

- Type I error always use If the null hypothesis is dismissed, even though it is true and should not be dismissed during the hypothesis testing phase.

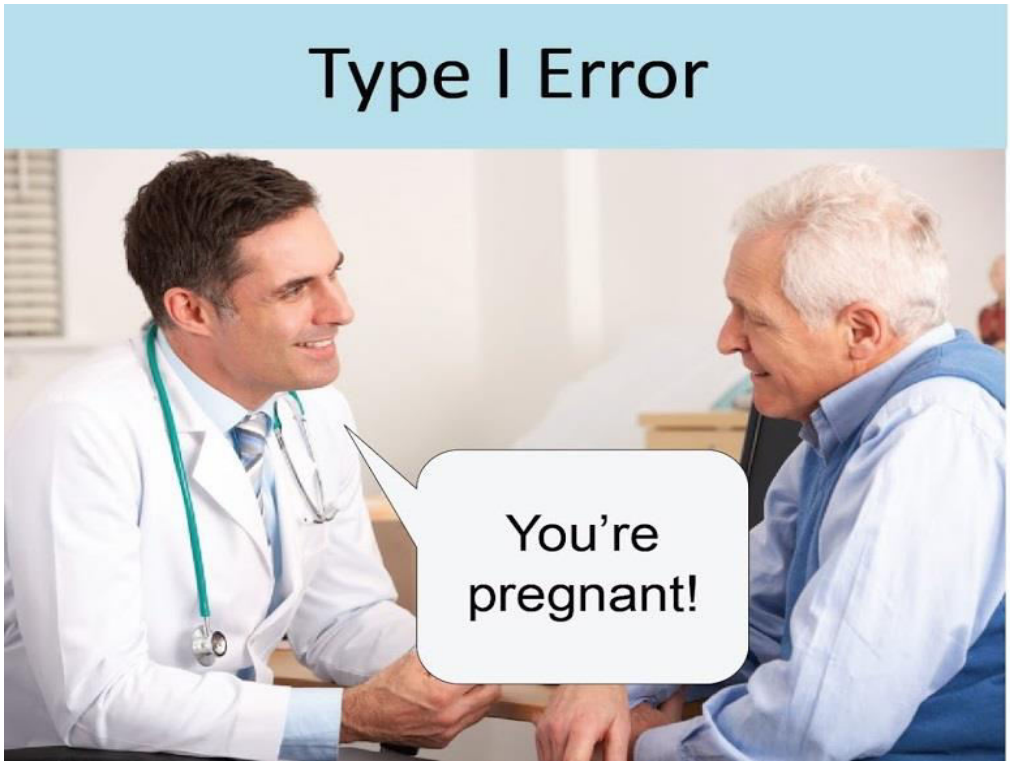


Figure: Type I error

Source: https://images.app.goo.gl/h3GAMnAgikgkrQ3GA

Type II error

- Throughout mathematical research, the dismissal of a real null hypothesis is a Type I, while the error of type II defines the mistake that happens when a null hypothesis is not discarded and is simply incorrect.

- Or put things another way, things generate a false statement. The fallacy denies the alternate explanation, although it does not arise out of chance.

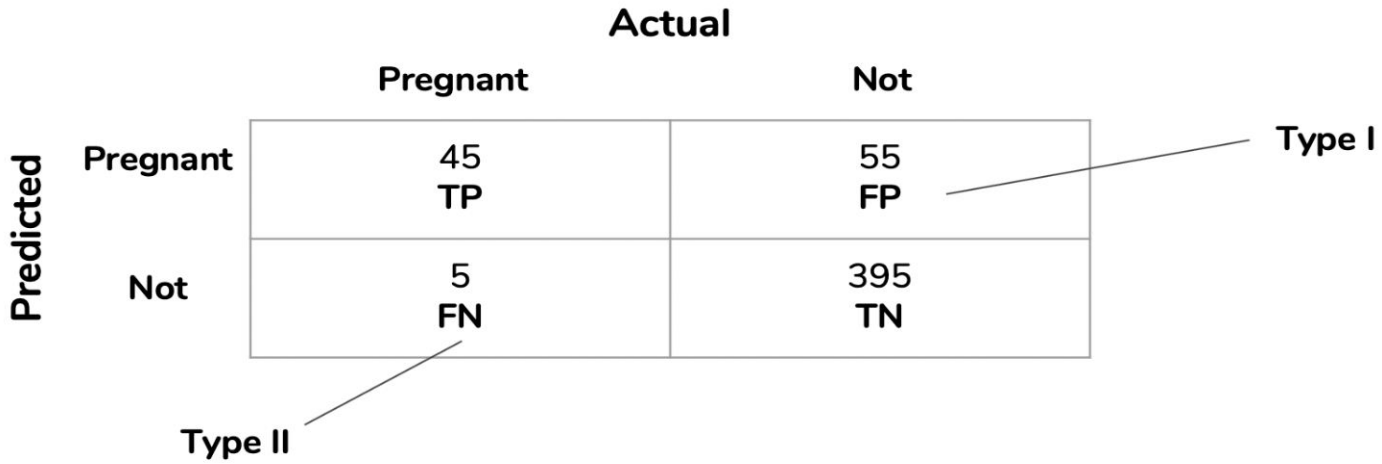


Figure: Confusion Matrix

Source: https://images.app.goo.gl/vjRF7VdSvThqaGbY9

Differences between type I and type II errors

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Figure: type I and II error

Source: https://images.App.Goo.Gl/sr5bvbv93kw9hm378

Null hypothesis

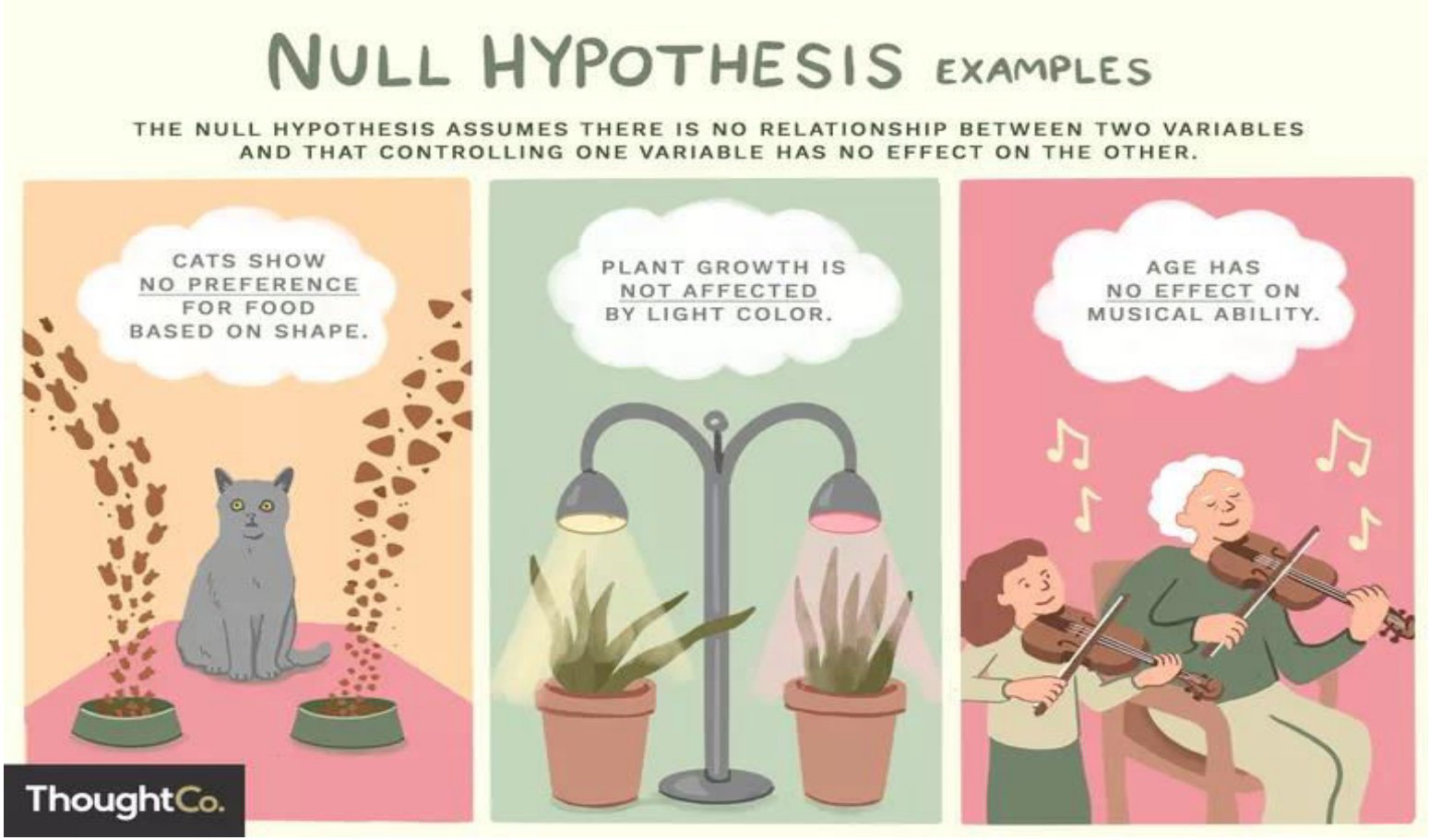


Figure: Null Hypothesis

Source: https://www.thoughtco.com/null-hypothesis-examples-609097

Statistical significance

- Statistical importance is an analyst's conviction that the findings in the data cannot be interpreted by chance alone.

- The tool by which the analyst makes the decision is mathematical hypothesis testing.

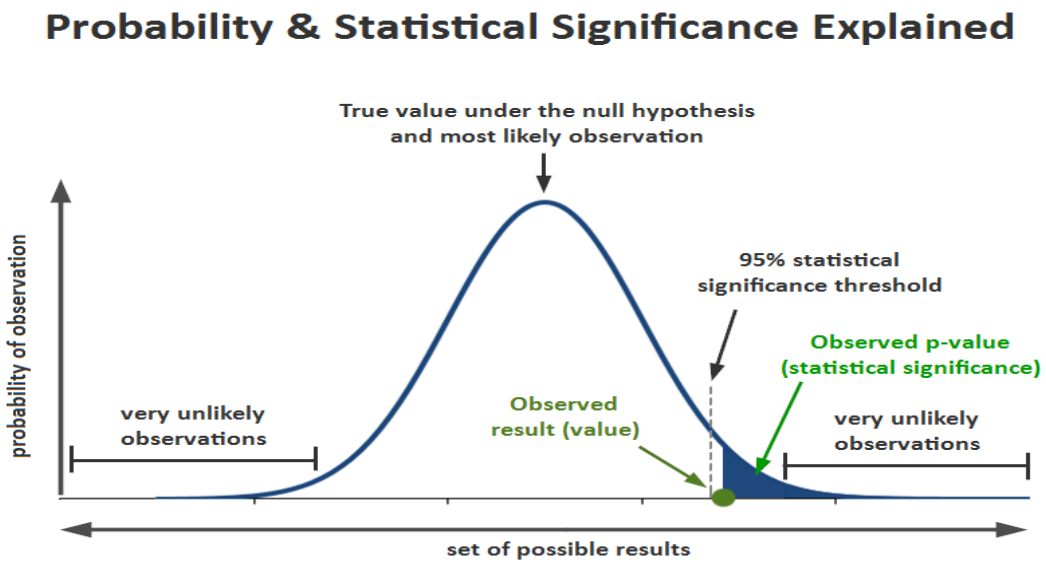


Figure: Probability and Significance overview

Source: https://images.app.goo.gl/jAxPTMP9VVxHhdHs6

Self evaluation: Exercise 8

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 8: Sparse kernel machines

Hypothesis testing

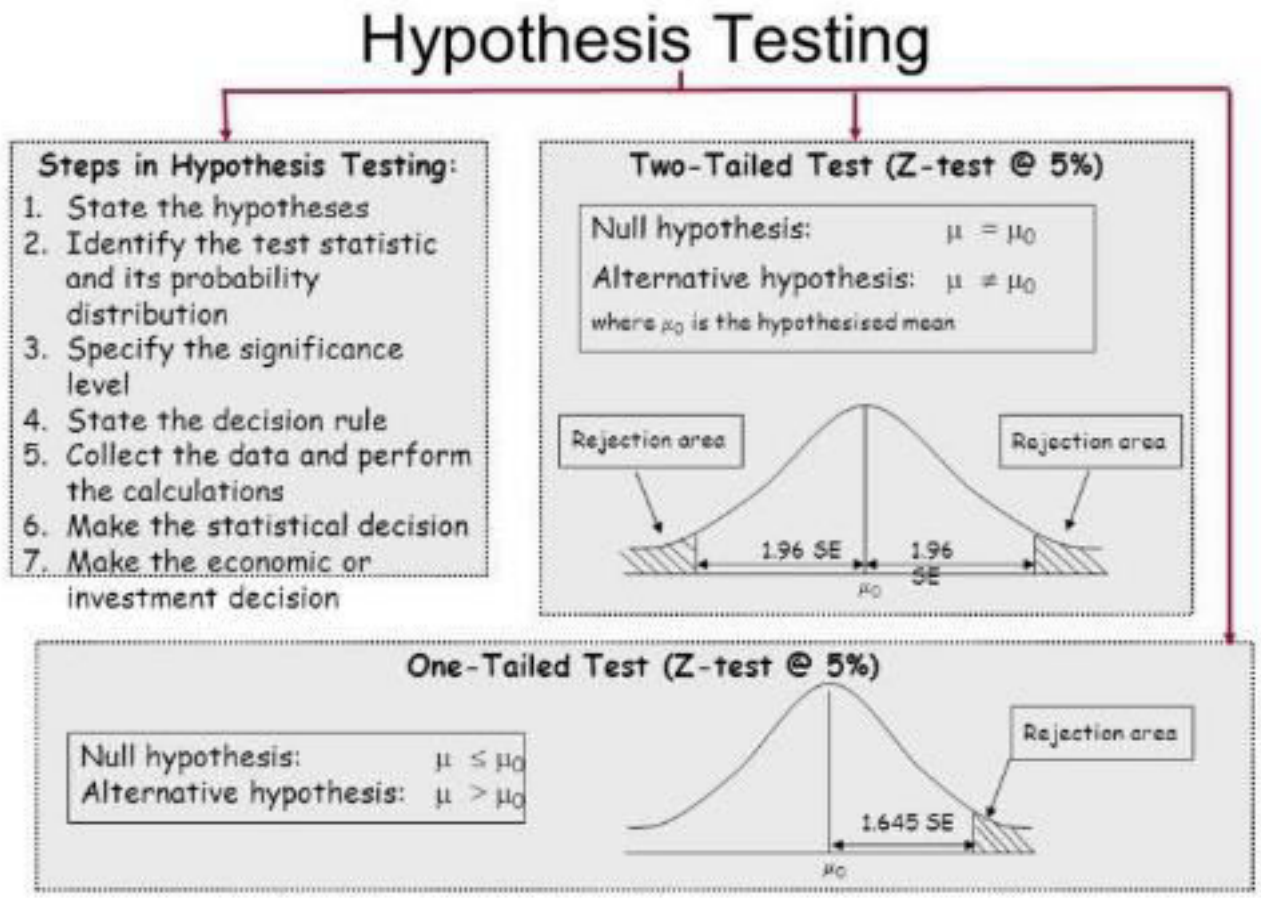


Figure: Hypothesis testing Process

Source: https://images.app.goo.gl/DcJE1FPHEm9xLeT89

Four steps of hypothesis testing

- The first step is to state the two assumptions so that only one hypothesis is right.
- The next step is to establish a method of analysis to analyze the data.
- The third stage is the implementation of the software and the realistic evaluation of the sample data.
- The fourth and final stage is to determine and dismiss the null hypothesis or to suggest that the null hypothesis is plausible if the information is given.

Real-world example of hypothesis testing

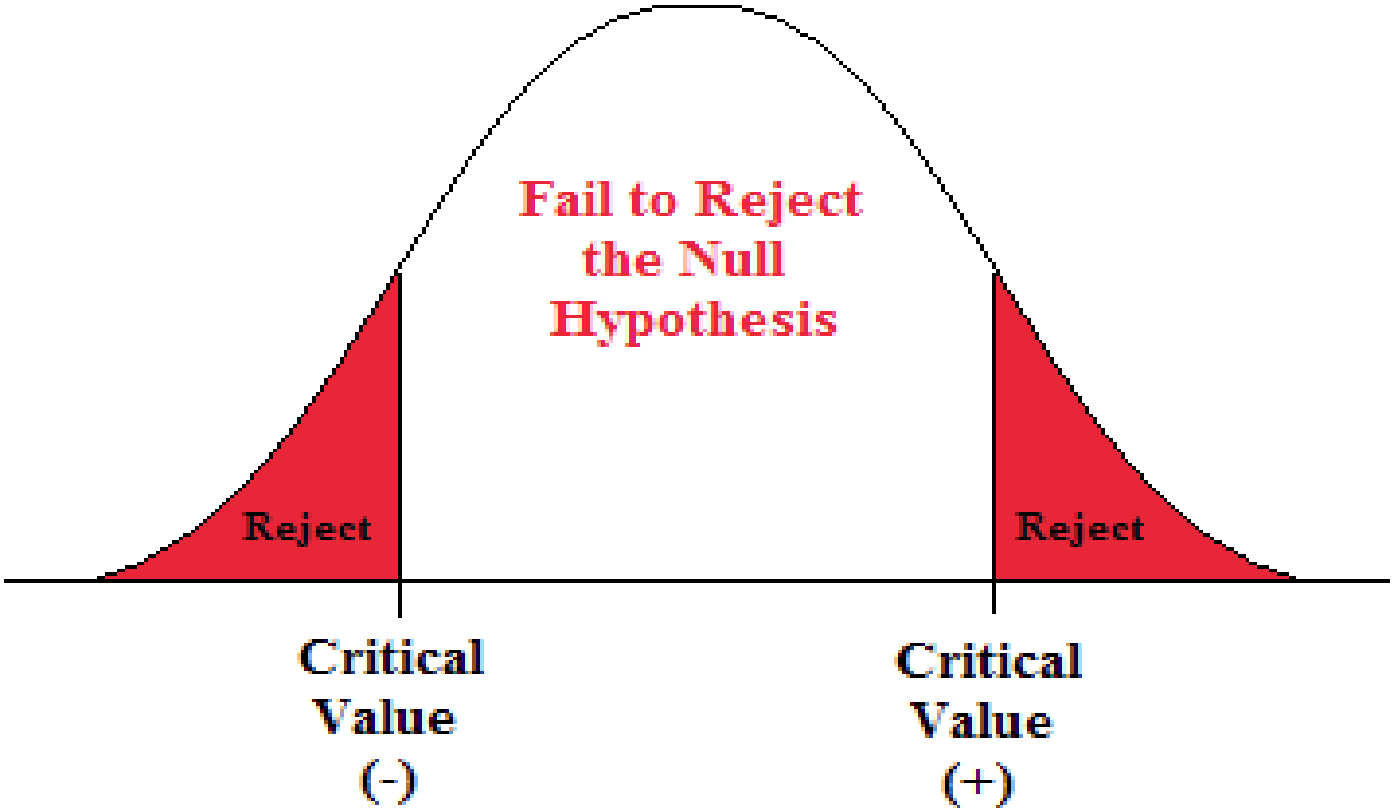


Figure: Beta Risk

Source: https://images.app.goo.gl/d8JcqwQ4Lws3oXN19

Bonferroni test

- A Bonferroni test is multiple forms of comparison used for statistical evaluation. Finally, a result can emerge from many hypothesis experiments with different variables, suggesting the dependent variable's statistical significance, although none exists.

Check of one-tailed

- A one-tail test is a statistical test in which a distribution's critical area is unilateral so that the value is either greater or lower than a certain value but not both.

- If the test sample falls into the critical unilateral zone, an alternative hypothesis rather than the null hypothesis is accepted.

Probability distributions

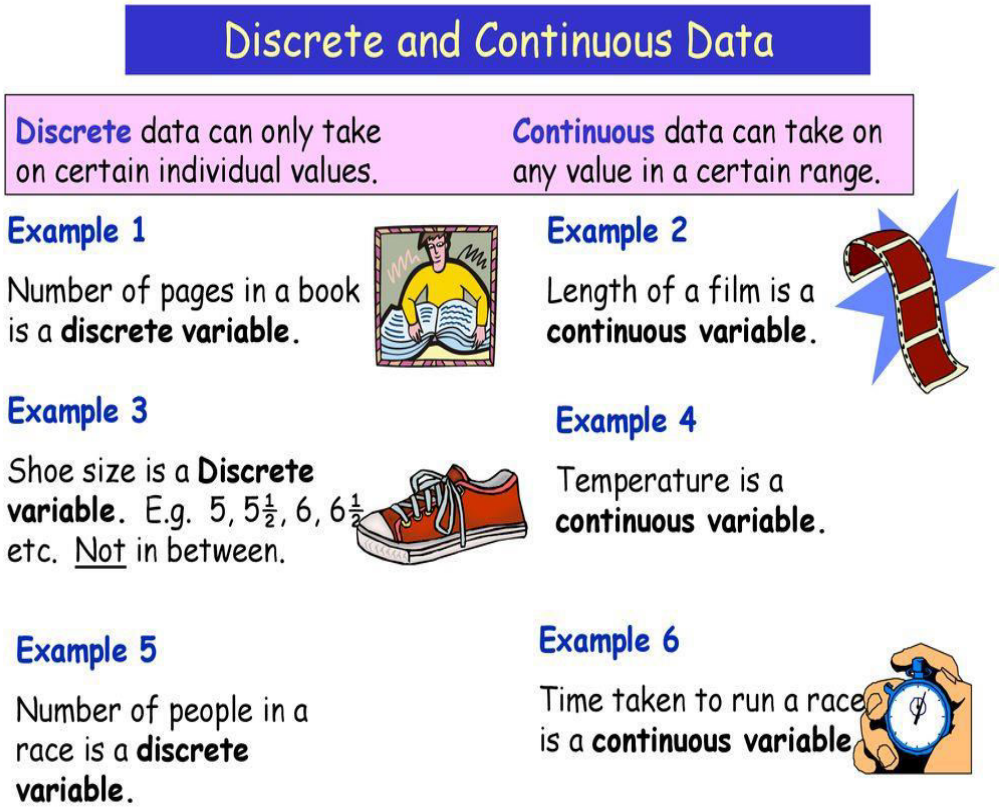


Figure: Discrete data and continuous data

Source: https://images.app.goo.gl/JdPuZvCNUxvNRV9S7

Self evaluation: Exercise 9

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 9: Sampling methods for pattern recognition.

Types of distributions

- **Bernoulli distribution:**
  - Bernoulli equation defines events with precisely two real-life outcomes. Several illustrations of such activities are as follows: a team wins a tournament or not, a student passes or fails an assessment and a roll-out dice shows either a 6 or another number.
  - Only two possible tests, namely 1 (success) and 0 (failure) have a Bernoulli distribution and only one analysis.

- Therefore, a random X variable with a Bernoulli distribution will have value 1 at the probability of success (p), and value 0 at the probability of failure(q or 1-p).

- An incident of the head here is a success, and a tail occurrence is a deception. Having a head = 0.5 = probability to get a neck because only two of these scenarios are likely.

Regression models

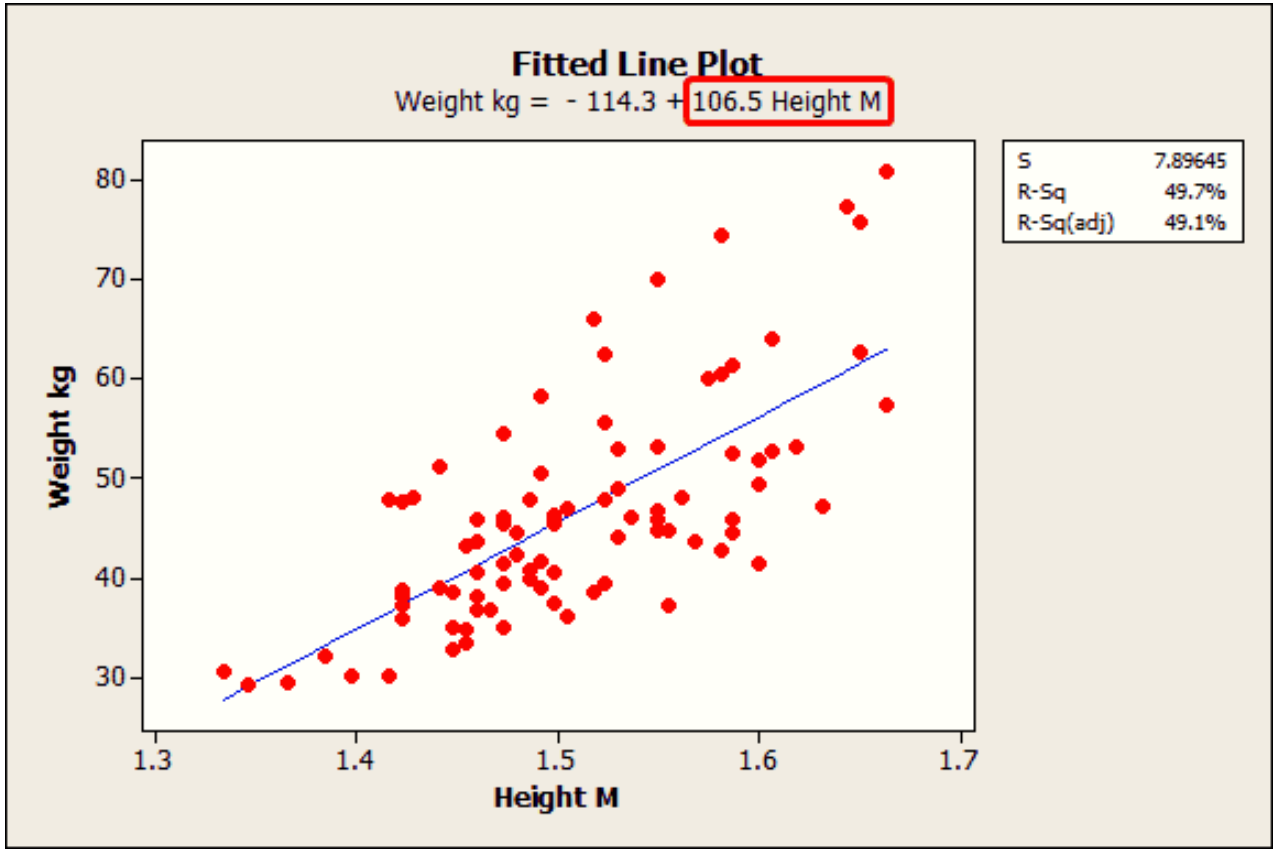


Figure: Regression Analysis

Source: https://images.app.goo.gl/yAcBF7tq1zwh4gWaA

Self evaluation: Exercise 10

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 10: Decision tree.

Types of regression

Type of regression	Dependent variable and its nature	Independent variable and its nature	Relationship between variables
Simple linear	One, continuous, normally distributed	One, continuous, normally distributed	Linear
Multiple linear	One, continuous	Two or more, may be continuous or categorical	Linear
Logistic	One, binary	Two or more, may be continuous or categorical	Need not be linear
Polynomial (logistic) [multinomial]	Non-binary	Two or more, may be continuous or categorical	Need not be linear
Cox or proportional hazards regression	Time to an event	Two or more, may be continuous or categorical	Is rarely linear

Figure: Regression types

Source: https://images.app.goo.gl/59CM8MmjMP1sGPQJA

How to select the best model for regression?

- A research institute asks its students to perform linear regression-whether the outcome is constant.

- When you have a conditional regression logistics requirement! However, the more options you have, the easier it is to pick one.

Common questions

- How many regression types do we have?
- How much mathematical knowledge is required to understand regression?
- Ridge vs. Lasso Regression - what is the difference?
- Which types of problems can be solved using regression?
- What are the major challenges faced by regression techniques?
- Is Regression Analysis relevant in the industry?
- Which programming language works best for regression?

Linear models for classification

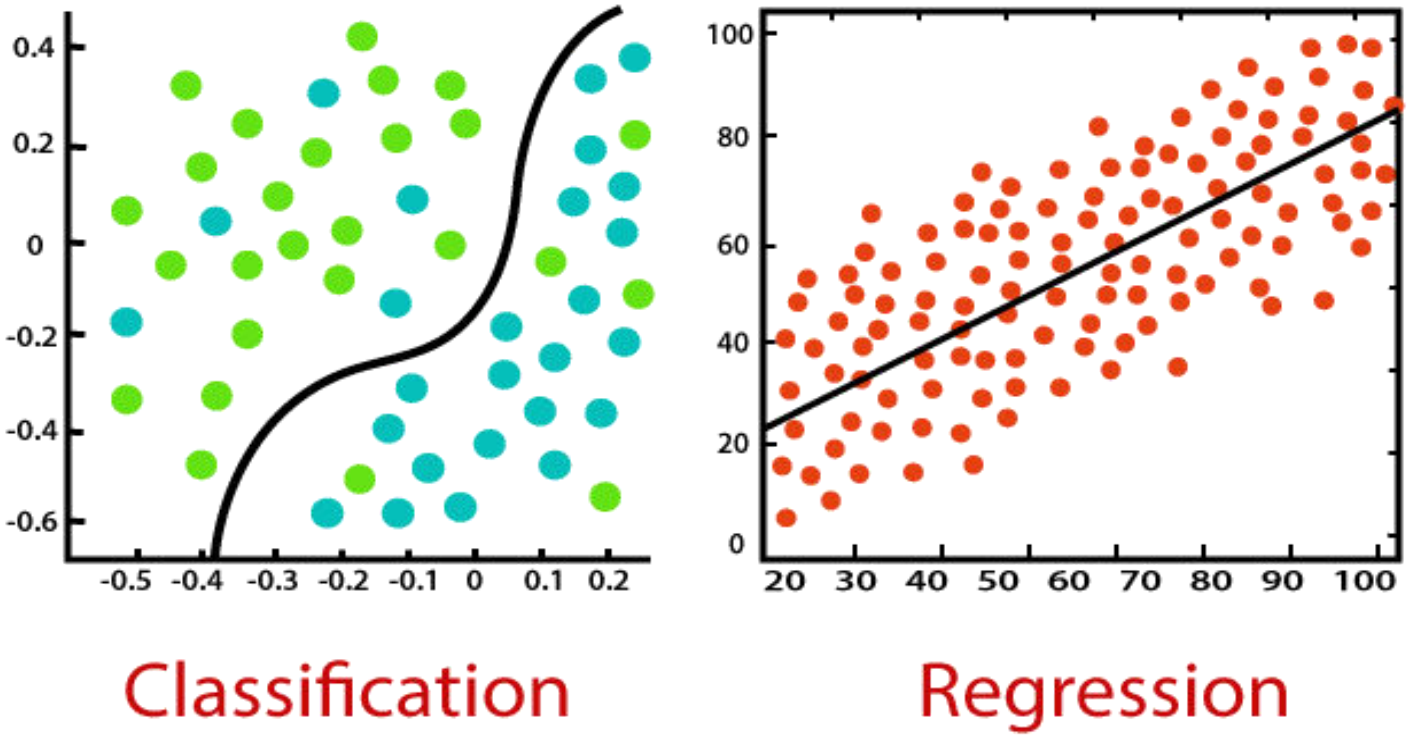


Figure: Classification and regression Model

Source: https://images.app.goo.gl/fzwieTpyn2DHXAwB9

Example of positive linear regression

- Price elasticity research.
- Risk evaluation in an insurance company.
- Sports analysis.

Checkpoint (1 of 2)

Multiple choice questions:

1. Memory decay affects what kind of memory?
   a) Short tem memory in general
   b) Older memory in general
   c) Can be short term or older
   d) None of the mentioned

2. How is pattern information distributed?
   a) It is distributed across the weights
   b) It is distributed in localized weights
   c) It is distributed in certain proactive weights only
   d) None of the above

3. What are the requirements of learning laws?
   a) Learning should be able to capture more & more patterns
   b) Convergence of weights
   c) All the mentioned
   d) None of the above

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. Memory decay affects what kind of memory?
   - **a) Short tem memory in general**
   - b) Older memory in general
   - c) Can be short term or older
   - d) None of the mentioned

2. How is pattern information distributed?
   - **a) It is distributed across the weights**
   - b) It is distributed in localized weights
   - c) It is distributed in certain proactive weights only
   - d) None of the above

3. What are the requirements of learning laws?
   - a) Learning should be able to capture more & more patterns
   - b) Convergence of weights
   - **c) All the mentioned**
   - d) None of the above

Checkpoint (2 of 2)

Fill in the blanks:

1. _____factors affect the performance of learner system does not include?
2. In language understanding, the levels of knowledge that does not include ___.
3. _____consists of the categories which does not include structural units.
4. A search algorithm takes _____as an input and returns ____as an output.

True or False:

1. In pattern mapping problem in neural nets, is there any kind of generalization involved between input & output? True/False
2. Linear neurons can be useful for application such as interpolation, is it true? True/False
3. Does pattern classification & grouping involve same kind of learning? True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Good data structures** factors affect the performance of learner system does not include?
2. In language understanding, the levels of knowledge that does not include **Empirical**.
3. **A model of language** consists of the categories which does not include structural units.
4. A search algorithm takes **problem** as an input and returns **solution** as an output.

True or False:

1. In pattern mapping problem in neural nets, is there any kind of generalization involved between input & output? **True**
2. Linear neurons can be useful for application such as interpolation, is it true? **True**
3. Does pattern classification & grouping involve same kind of learning? **False**

Question bank

Two mark questions:

1. What is probability distributions ?
2. What are the components of probability distributions ?
3. List any 3 types linear models for regression.
4. What is y=mx+c formula for linear regression?

Four mark questions:

1. What is multiple regression model?
2. Describe r-squared method.
3. Describe classification techniques.
4. Describe any 3 types of classification methods.

Eight mark questions:

1. Explain linear models for classification.
2. Explain probability distributions with chart.

Unit summary

Having completed this unit, you should be able to:

- Understand the concept of probability distributions
- Gain knowledge on example of statistical approaches
- Understand linear models for regression
- Learn about linear models for classification

Welcome to:

Machine Learning Approaches for Pattern Recognition

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of neural networks and kernel methods
- Learn example of sparse kernel machines and graphical models
- Gain knowledge on sampling methods for pattern recognition
- Understand pattern recognition in sequential data

Neural networks

- A neural network is an algorithm series that attempts to understand simple connections in a data set using a computer that simulates the role of the brain.
- In this case, neural networks refer to biological or artificial neuronal structures.
- In the world of finance, neural networks are helpful in developing processes, including time series forecasts, algorithmic trading, and classification of stocks, credit risk modeling and the creation of proprietary price indices.

How neural networks learn?

Supervised	Unsupervised	Semi-Supervised	Reinforcement
• Data has **known labels** or output	• Labels or output unknown • Focus on **finding patterns and gaining insight** from the data	• Labels or output known for **a subset of** data • A blend of supervised and unsupervised learning	• Focus on **making decisions** based on previous experience • Policy-making with feedback
• Insurance underwriting • Fraud detection	• Customer clustering • Association rule mining	• Medical predictions (where tests and expert diagnoses are expensive, and only part of the population receives them)	• Game AI • Complex decision problems • Reward systems

Figure: Type of learning

Source: https://images.app.goo.gl/yUyUcXku5R6fg9sK9

Neural networks examples

- Handwriting recognition.
- Financial-market forecasting.
- Social media images analysis.
- Diagnostic imaging for diagnosing cancer.

Neural networks use cases

- Space.
- Automotive.
- Electronics.
- Manufacturing.
- Mechanics.
- Robotics.
- Telecom.
- Banking.

Kernel methods

- Kernels or kernel approaches (also called kernel functions) are collections of different forms of algorithms used to evaluate trends. Use a linear classifier, they are utilized to solve a nonlinear problem.

- Kernel techniques are employed in SVM (Support Vector Machines) and is used in problems of classification and regression.

- The SVM utilizes what is called a "kernel trick" where the data is converted and an appropriate boundary for potential outputs is sought.

Self evaluation: Exercise 11

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 11: Random forest.

Sparse kernel machines use cases

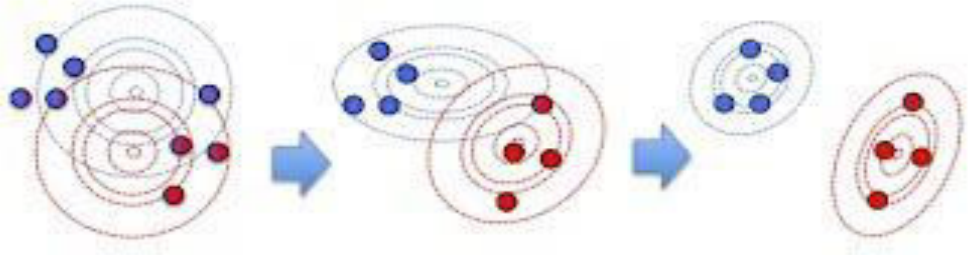
- Facial recognition.
- Categorization of text and hypertext.
- Identification of photographs.
- Bioinformatics.
- Protein folding and remote detection of homology.
- Recognition of handwriting.
- GPC.

Graphical models

- Some popular machine learning problems involve the classification of discrete, different data points.
- For example, if a picture contains a cat or a dog, predict which digit it will be from 0 to 9, if a handwritten character is included.
- There are some things that do not fit into the above structure. Each word is defined by its part of expression (a noun, a pronoun, a verb, an adjective, etc.) for example given a sentence "Machine learning I like.
- "As this basic illustration alone demonstrates, each term "learning" may be considered as a meaning-based noun or verb.
- This role is essential for many more complicated text activities such as the language, voice-to - text translation, etc.

Mixture models and EM

Gaussian Mixture Model



- Data with D attributes, from Gaussian sources $c_1 \ldots c_k$
  - how typical is $\mathbf{x}_i$ under source **c**: $P(\vec{x}_i \mid c) = \frac{1}{\sqrt{2\pi|\Sigma_c|}} \exp\left\{-\frac{1}{2}(\vec{x}_i - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}_i - \vec{\mu}_c)\right\}$ where $\sum_a \sum_b (x_{ia} - \mu_{ca})\left[\Sigma_c^{-1}\right]_{ab} (x_{ib} - \mu_{cb})$
  - how likely that $\mathbf{x}_i$ came from **c**: $P(c \mid \vec{x}_i) = \frac{P(\vec{x}_i \mid c) P(c)}{\sum_{c=1}^{k} P(\vec{x}_i \mid c) P(c)}$
  - how important is $\mathbf{x}_i$ for source **c**: $w_{ic} = P(c \mid \vec{x}_i) / \left(P(c \mid \vec{x}_1) + \ldots + P(c \mid \vec{x}_n)\right)$
  - mean of attribute **a** in items assigned to **c**: $\mu_{ca} = w_{c1} x_{1a} + \ldots + w_{cn} x_{na}$
  - covariance of **a** and **b** in items from **c**: $\Sigma_{cab} = \sum_{i=1}^{n} w_{ci} \left(x_{ia} - \mu_{ca}\right)\left(x_{ib} - \mu_{cb}\right)$
  - prior: how many items assigned to **c**: $P(c) = \frac{1}{n}\left(P(c \mid \vec{x}_1) + \ldots + P(c \mid \vec{x}_n)\right)$

Figure: Mixture Models and EM

Source: https://images.app.goo.gl/9TQmWETJYvpDJi7w5

Bayesian networks: Directed graphical models

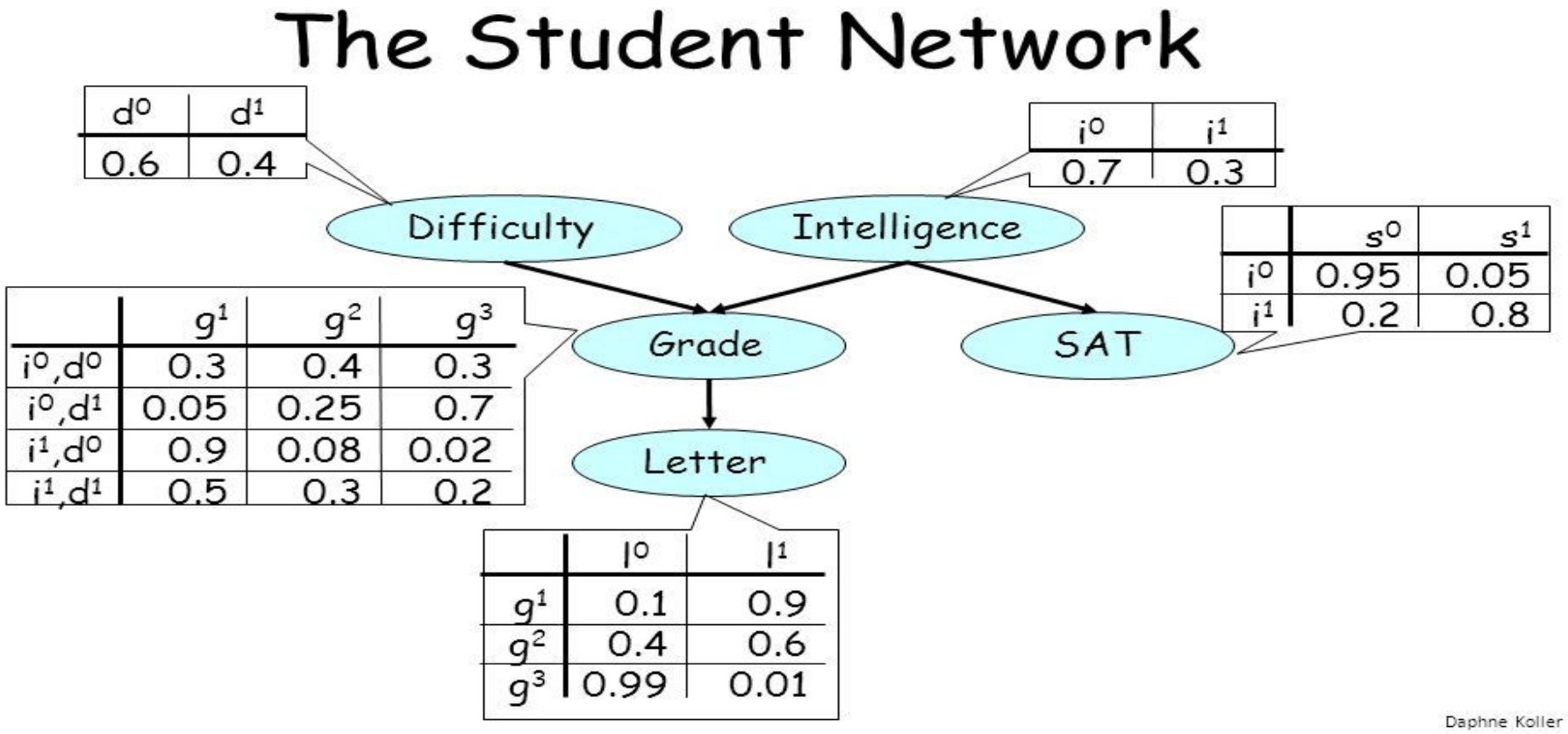


Figure: Bayesian networks: Directed graphical models

Source: https://images.app.goo.gl/qECFiT5QUHJAS7U38

Conditional probability distributions

- CPDs with "difficulties" and "knowledge," since they do not depend on the other variables, are relatively simple.

- In theory, the tables reflect the probability of other variables, as values 0 or 1.

- The numbers in each table must be 1, as you may have noted.

Potential functions

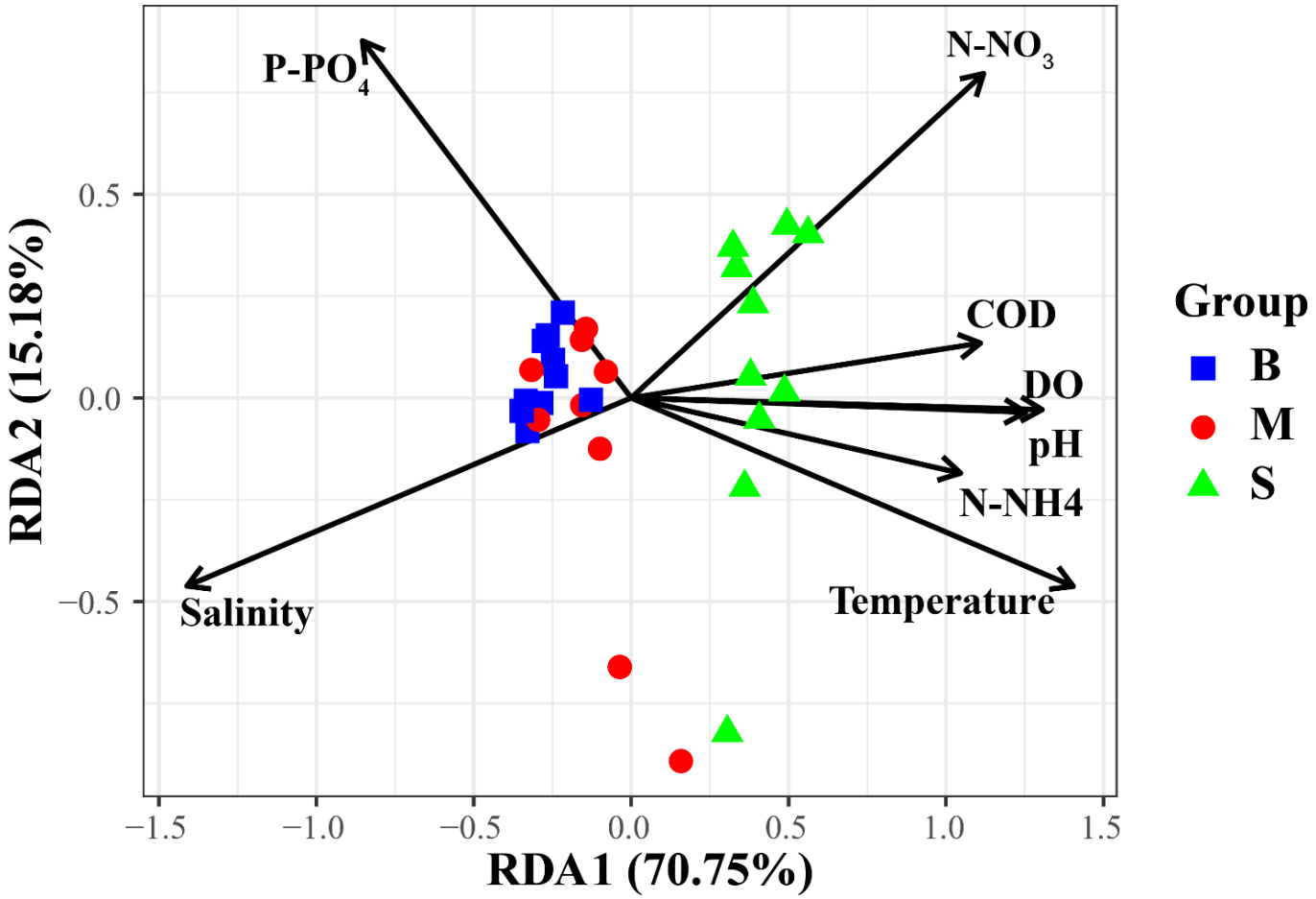


Figure: Potential Functions

Source: https://images.app.goo.gl/crfvoAiNu3bruX8E8

Self evaluation: Exercise 12

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 12: SVM.

Conditional independences

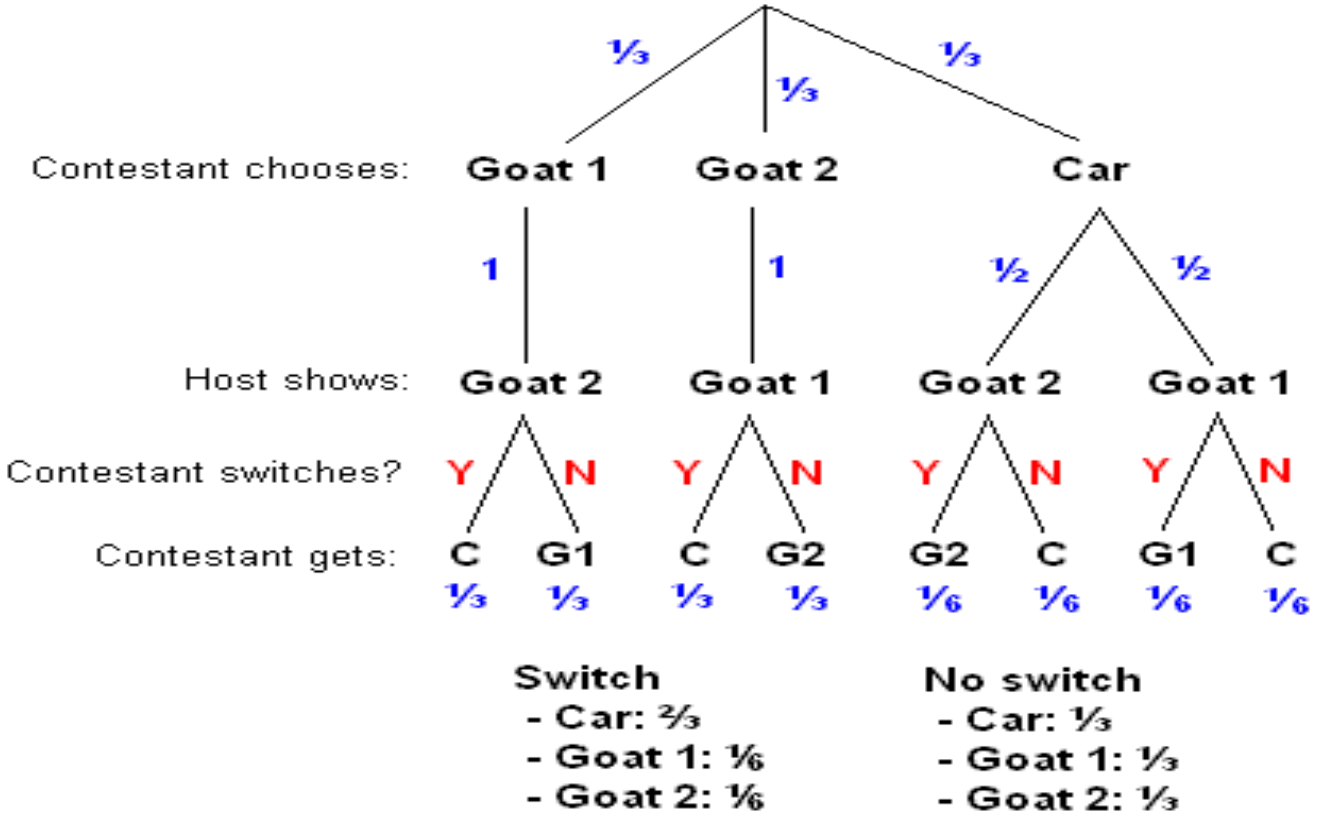


Figure: Conditional Independences

Source: https://images.app.goo.gl/DEUAMZEwBBKR1CN58

Sampling methods for pattern recognition

- Sampling is a method to gather population data based on statistics from a portion of the community (sample) without the need to look at individuals.

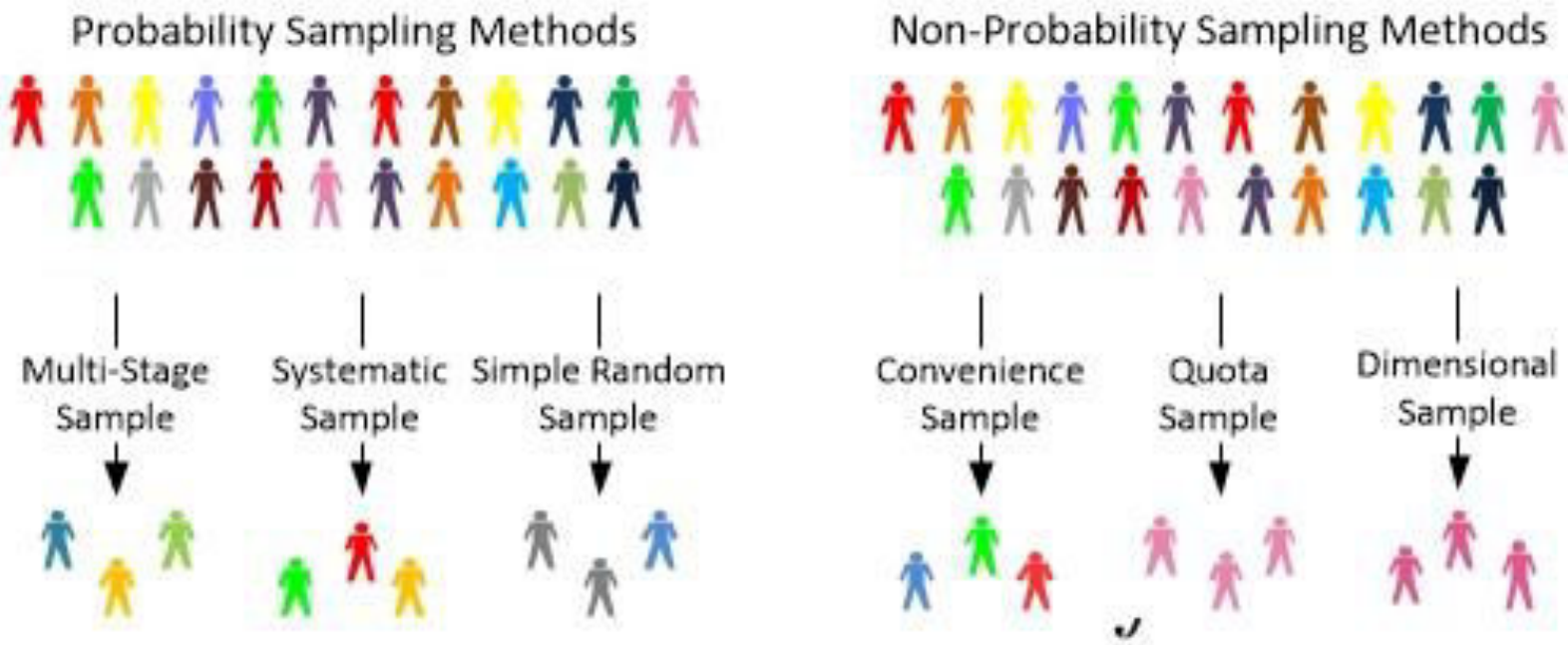


Figure: Sampling Methods

Source: https://images.app.goo.gl/xbLo6pyTzPRrszEs9

Continuous latent variables

	Manifest variables	
Latent variables	**Continuous**	**Categorical**
Continuous	Factor analysis	Item response theory
Categorical	Latent profile analysis	Latent class analysis

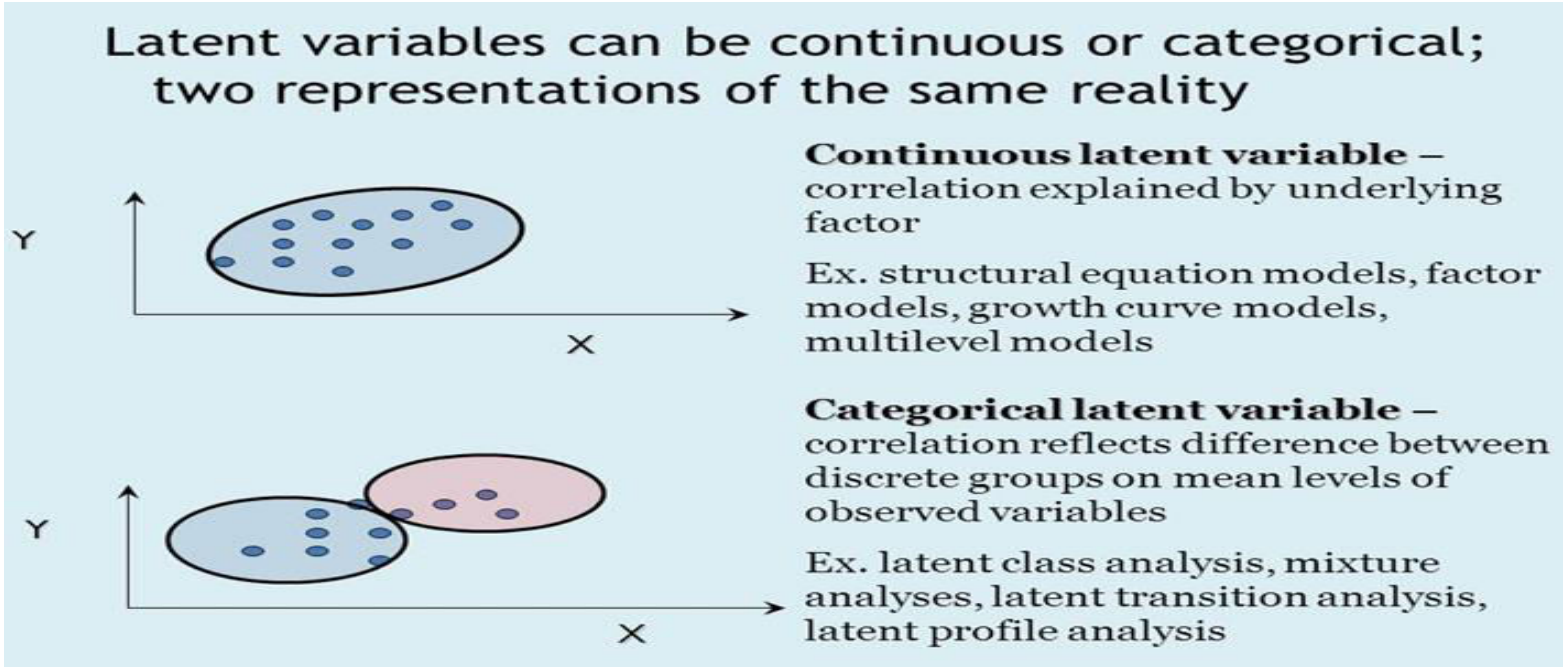


Figure: Continuous Latent Variables

Source: https://images.app.goo.gl/PjPtHiRmSCW6J7ej6

Self evaluation: Exercise 13

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 13: Local Outlier Factor (LOF).

Combining models for pattern recognition

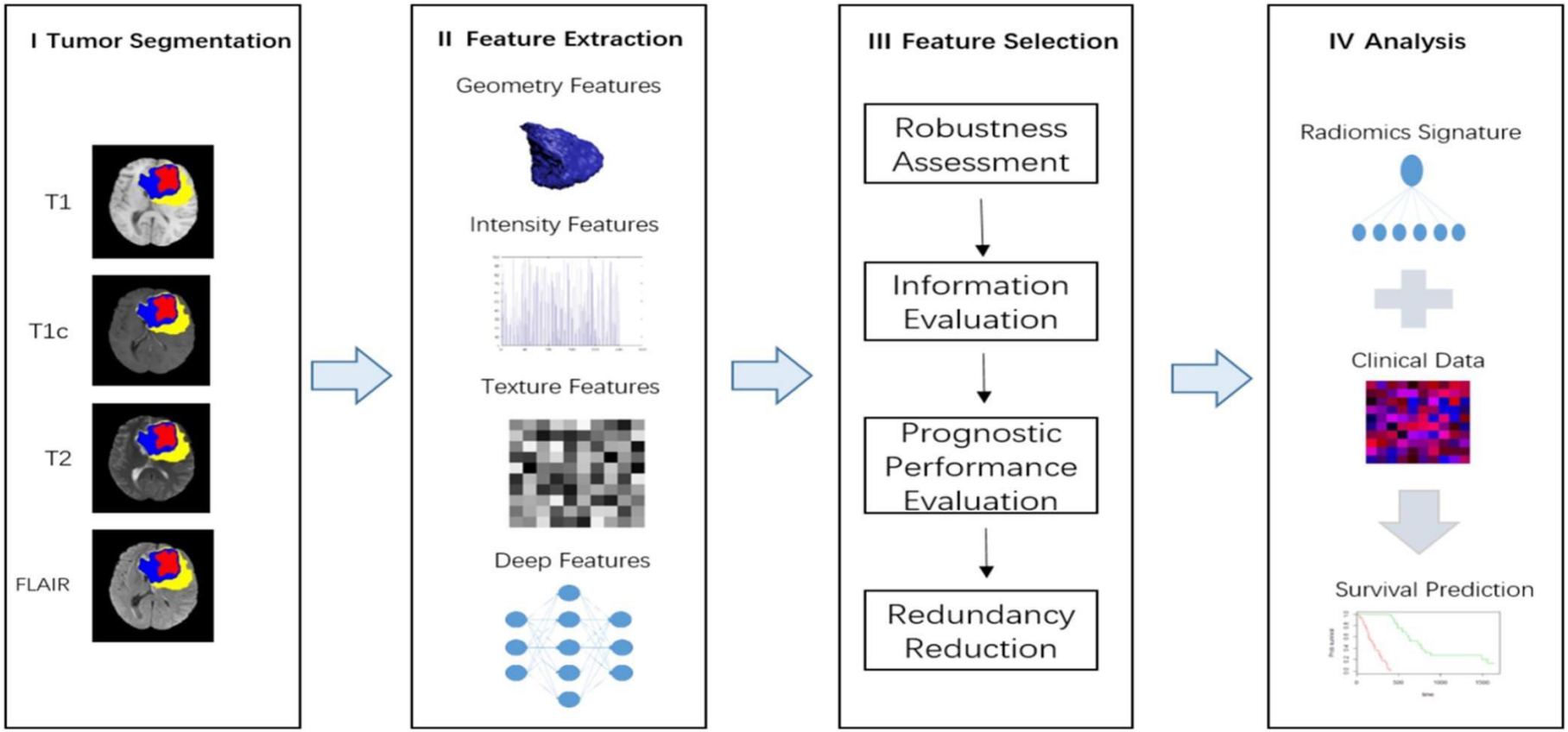


Figure: Combining Models for Pattern Recognition

Source: https://images.app.goo.gl/zct722Vc3vptf1KU9

Markov chain Monte Carlo

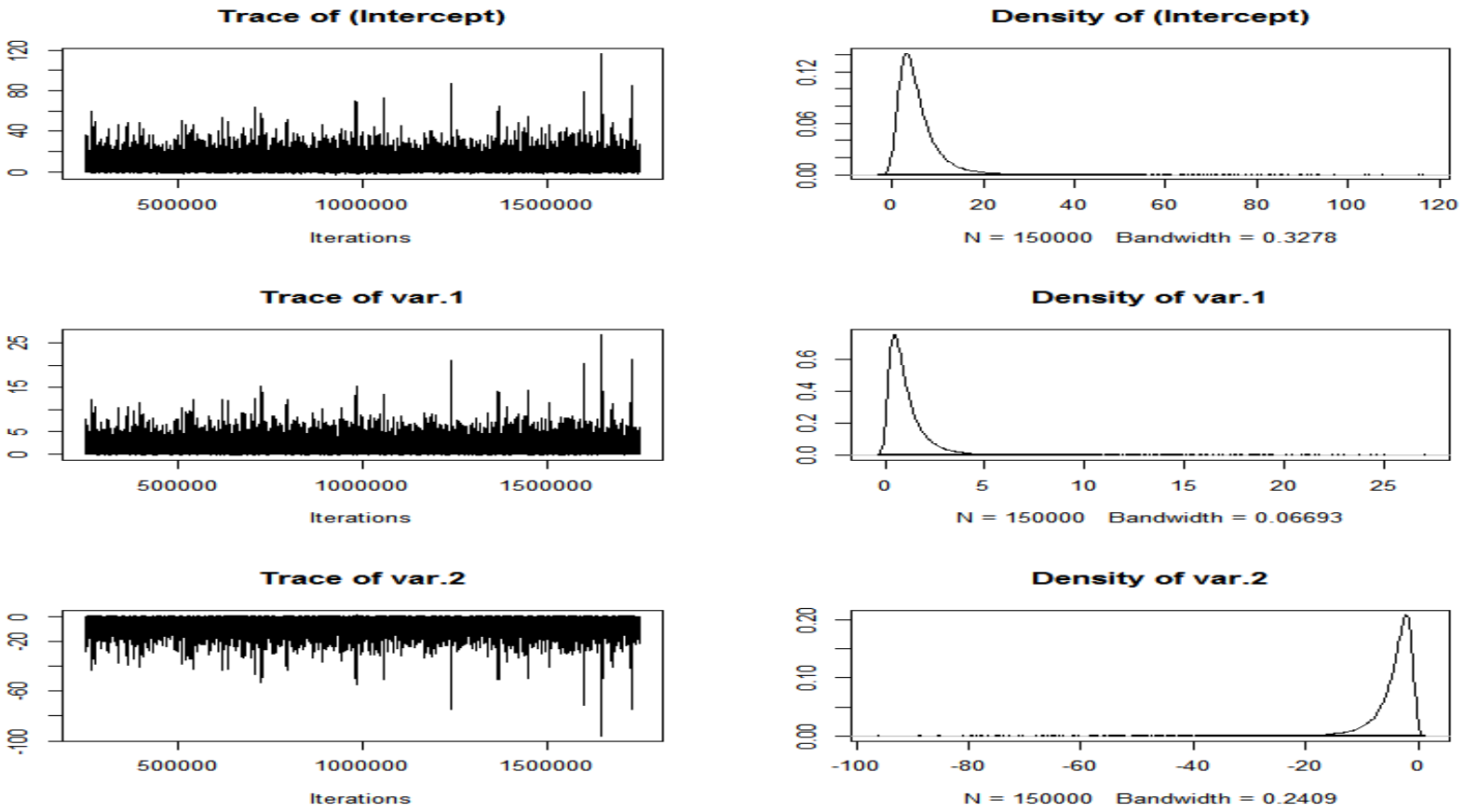


Figure: Markov chain Model

Source: https://images.app.goo.gl/9pTxRVtBURiuar6n8

The K-means algorithm

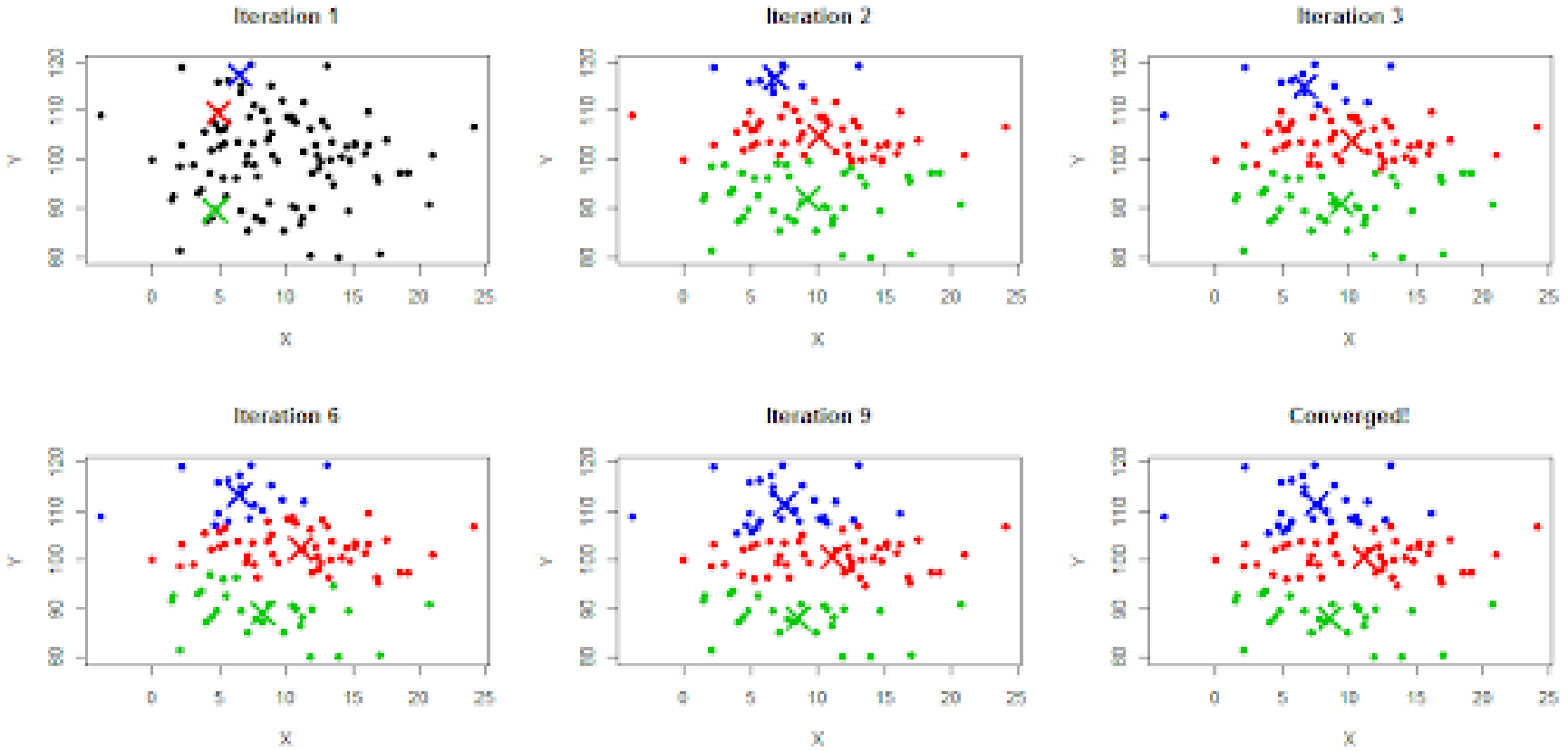


Figure: K-Means Algorithm

Source: https://images.app.goo.gl/27BUWSqpXkXMvG5u5

Applications of K-means

- Classification document.
- Delivery store optimization.
- Crime locations identifying.
- Segmentation of customers.
- Statistical review football team.
- Detection of fraud in insurance.
- Rideshare data analysis.

Checkpoint (1 of 2)

Multiple choice questions:

1. High entropy means that the partitions in classification are
   a) Pure
   b) Not pure
   c) Useful
   d) Useless

2. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?
   a) 12
   b) 24
   c) 48
   d) 72

3. Which of the following is NOT supervised learning?
   a) PCA
   b) Decision tree
   c) Linear regression
   d) Naive Bayesian

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. High entropy means that the partitions in classification are
   a) Pure
   b) Not pure
   c) Useful
   d) Useless

2. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?
   a) 12
   b) 24
   c) 48
   d) 72

3. Which of the following is NOT supervised learning?
   a) PCA
   b) Decision tree
   c) Linear regression
   d) Naive Bayesian

Checkpoint (2 of 2)

Fill in the blanks:

1. High entropy means that the partitions in classification are _____.
2. _____ is NOT supervised learning.
3. The _____ methods recognize clusters based on density function distribution.
4. Attributes are statistically _____ of one another given the class value.

True or False:

1. Stochastic gradient descent performs less computation per update than batch gradient descent. True/False
2. To classify job applications into two categories and to detect the applicants who lie in their applications using density estimation to detect outliers we can use generative classifiers. True/False
3. A good way to pick the number of clusters k, used for k-Means clustering is to try multiple values of k and choose the value that minimizes the distortion measure. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. High entropy means that the partitions in classification are **not pure**.
2. **PCA** is NOT supervised learning.
3. The **density-based clustering** methods recognize clusters based on density function distribution.
4. Attributes are statistically **dependent** of one another given the class value.

True or False:

1. Stochastic gradient descent performs less computation per update than batch gradient descent. **True**
2. To classify job applications into two categories and to detect the applicants who lie in their applications using density estimation to detect outliers we can use generative classifiers. **True**
3. A good way to pick the number of clusters k, used for k-Means clustering is to try multiple values of k and choose the value that minimizes the distortion measure. **False**

Question bank

Two mark questions:

1. Define precision and recall.
2. Explain how a ROC curve works.
3. How is KNN different from k-means clustering?
4. What is the difference between supervised and unsupervised machine learning?

Four mark questions:

1. What's the trade-off between bias and variance?
2. Why is "Naive" Bayes naive?
3. Describe the difference between L1 and L2 regularization.
4. What is the difference between type I and type II error?

Eight mark questions:

1. What is a Fourier transform?
2. What is the difference between a generative and discriminative model?

Unit summary

Having completed this unit, you should be able to:

- Understand the concept of neural networks and kernel methods
- Learn example of sparse kernel machines and graphical models
- Gain knowledge on sampling methods for pattern recognition
- Understand pattern recognition in sequential data

Anomaly Detection & Anomaly Detection Approaches

Unit objectives

After completing this unit, you should be able to:

- Understand the applications of anomaly detection
- Learn about example of classification-based methods
- Gain knowledge on nearest neighbor-based approach
- Gain an insight into clustering based methods
- Understand statistical approach, graph and model-based approach for ML implementation

What are anomalies?

- The detection of anomalies is a method used to detect unusual patterns not in accordance with appropriate behavior, called outliers.

- This includes many business applications:
  - Intrusion detection(recognizing unusual anomalies in network traffic activity that may indicate a hack).
  - System health monitoring (recognizing a cancerous tumor growing in an MRI scan).
  - Fraud detection in transactions with credit card to failure detection in operational environments.

Applications of anomaly detection

- Incursion/Intrusion discovery.
- Monitoring of frauds/ crimes.
- Healthcare informatics.
- Detection of industrial damage.
- Image processing.

Related use cases

- Rare class mining.
- Chance discovery.
- Novelty detection.
- Exception mining.
- Removal of noise.

Types of input data

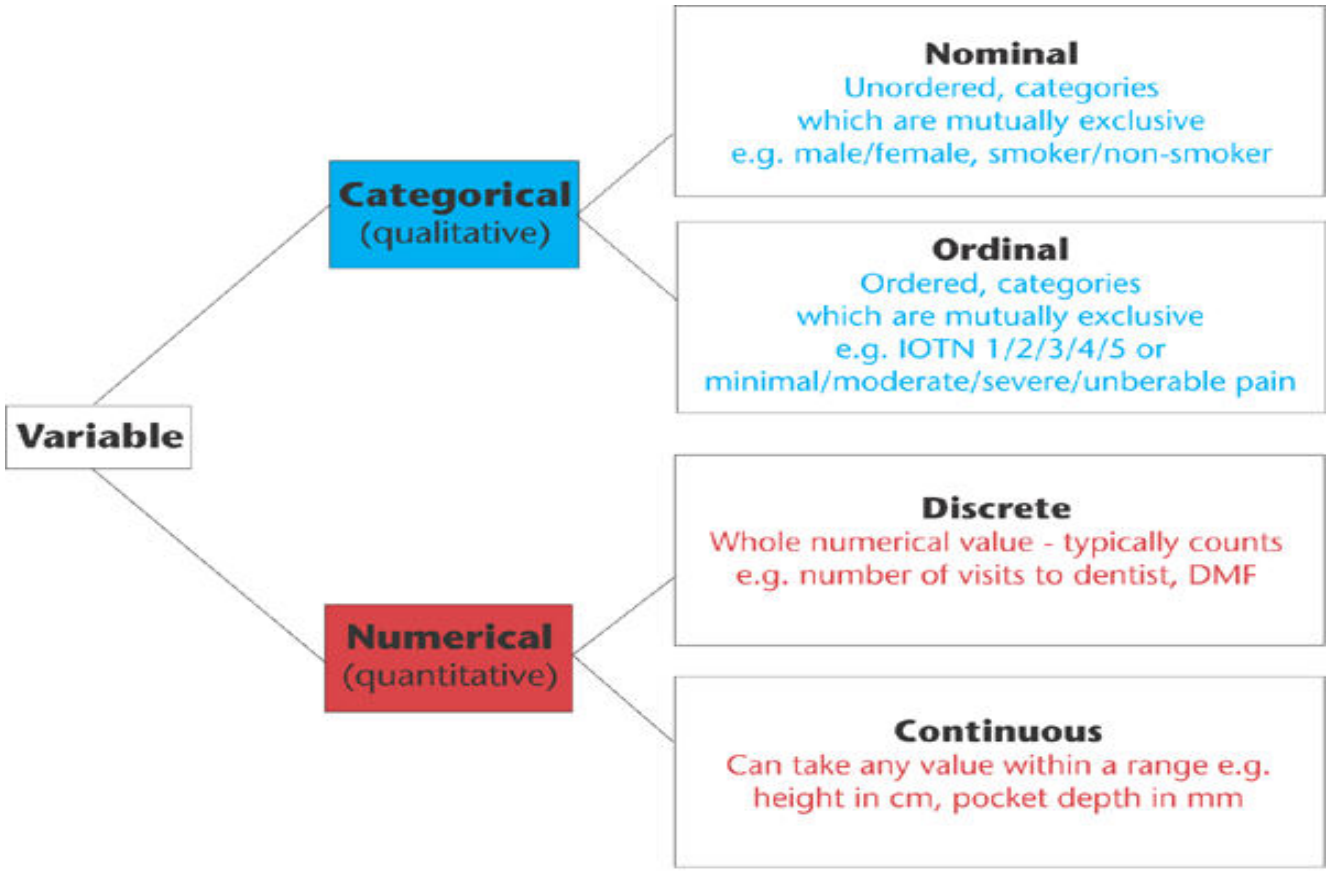


Figure: Types of Input Data

Source: https://images.app.goo.gl/r1NtrYL4DiHSu2y69

Types of anomalies

- Anomalies can be classified into the following three categories:
  - Point anomalies:
    - If one component is aberration, it is an anomaly purpose against all the other items.
    - It is the easiest type of phenomenon and it is studied by other studies.
    - The variations of points are considered in O1 and O2.
  - Contextual anomalies:
    - If in any given sense the entity is aberrant. just here is it a conceptual abnormality (also known as conditional anomaly).
  - Collective anomalies:
    - If irregularity is seen with a few other objects connected to other particles.
    - Such scenario, even the set of artifacts should not be aberrant.

Evaluation of an anomaly detector

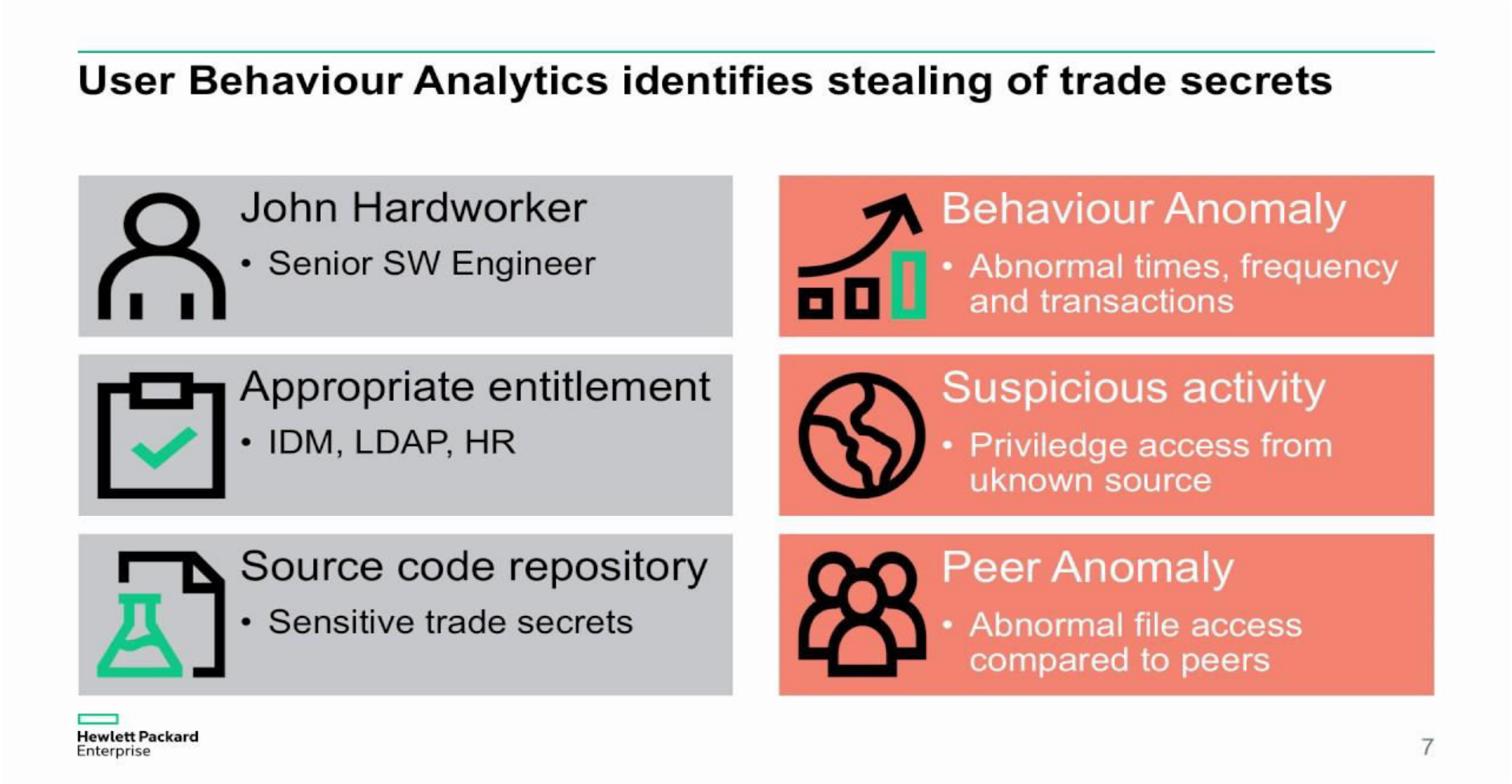


Figure: UBA Analysis

Source: https://images.app.goo.gl/S4XWcMZNQKmTec2Z7

Taxonomy of approaches

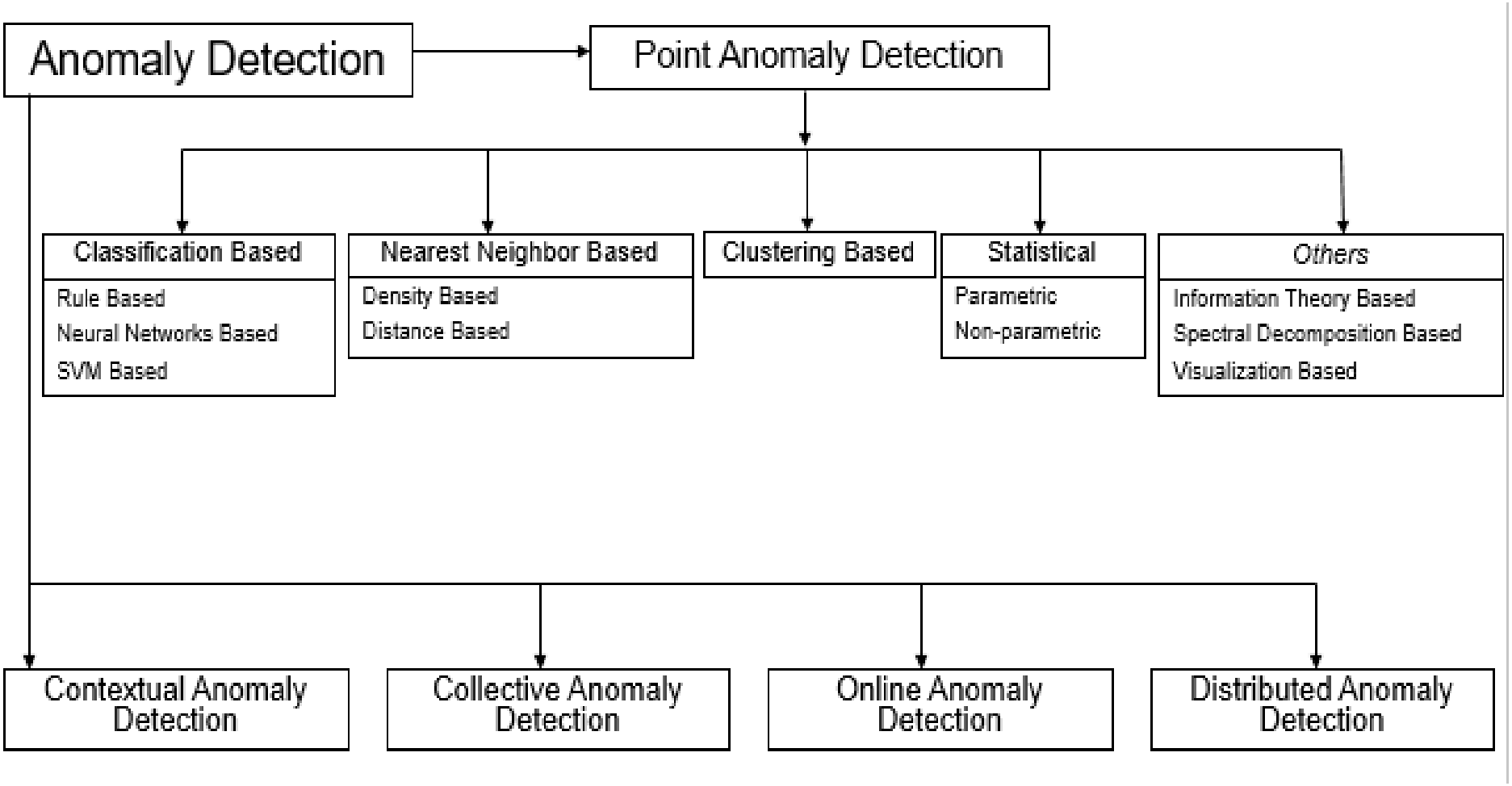


Figure: Approaches of taxonomy

Classification based

- The primary reason: Create a categorization models based on defined dataset for usual (and outlier (uncommon) occurrences to identify any mysterious new incident.

- Slanted (extremely unbalanced) class distributions will be treated by classification methods.

- Classification:
  - Methods for controlled sorting.
  - Need regular category & aberrant type understanding.
  - Define the structure to distinguish natural from established abnormalities.

- Strategies of semi-controlled grouping:
  - Need ordinary class awareness only.
  - Use the adapted detection data to predict behavior pattern and then recognize any abnormalities behavior as unusual.

Classification use cases

- Interception of web traffic-certain.
- Video classification.
- Classification of photographs.
- Classification of speech.

Supervised classification techniques

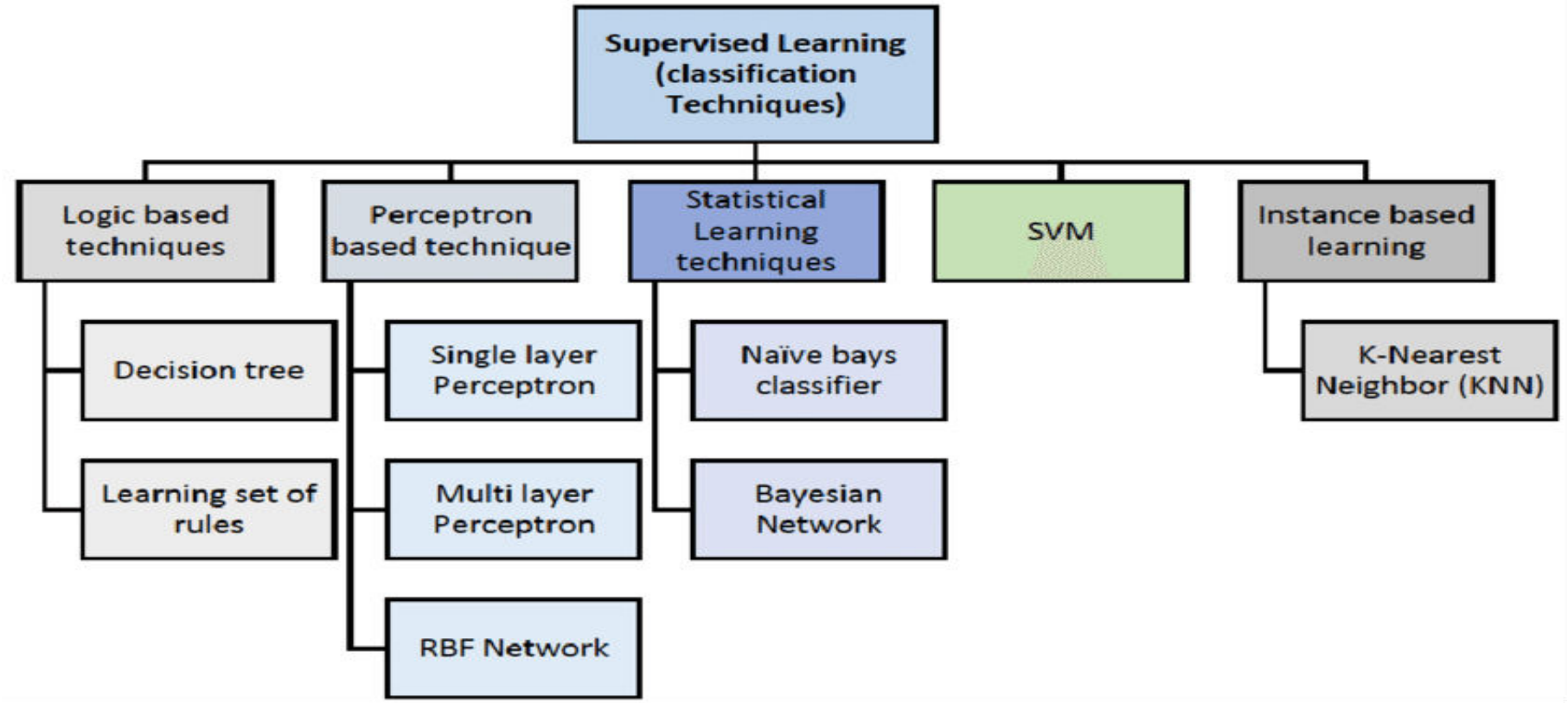


Figure: Supervised Classification Techniques

Source: https://images.app.goo.gl/PyZtjRJLce2FDQ6i7

Self evaluation: Exercise 14

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 14: Cluster based Local Outlier Factor (CBLOF).

Nearest neighbor based techniques

- Relevant presumption:
  - Standard points are in proximity, and anomalies are far away from other points.
  - 2 step common method:
    - For every data record, compute community.
    - To assess if or not the data reports are anomalous in the community.

- Categories:
  - Distance based methods: Anomalies are by far the most isolated reference points.
  - Density based methods: Data points in low-density regions are exceptions.

Self evaluation: Exercise 15

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 15: Local Density Cluster based Outlier Factor (LDCOF).

Self evaluation: Exercise 16

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 16: Local Correlation Integral (LOCI).

Others model techniques

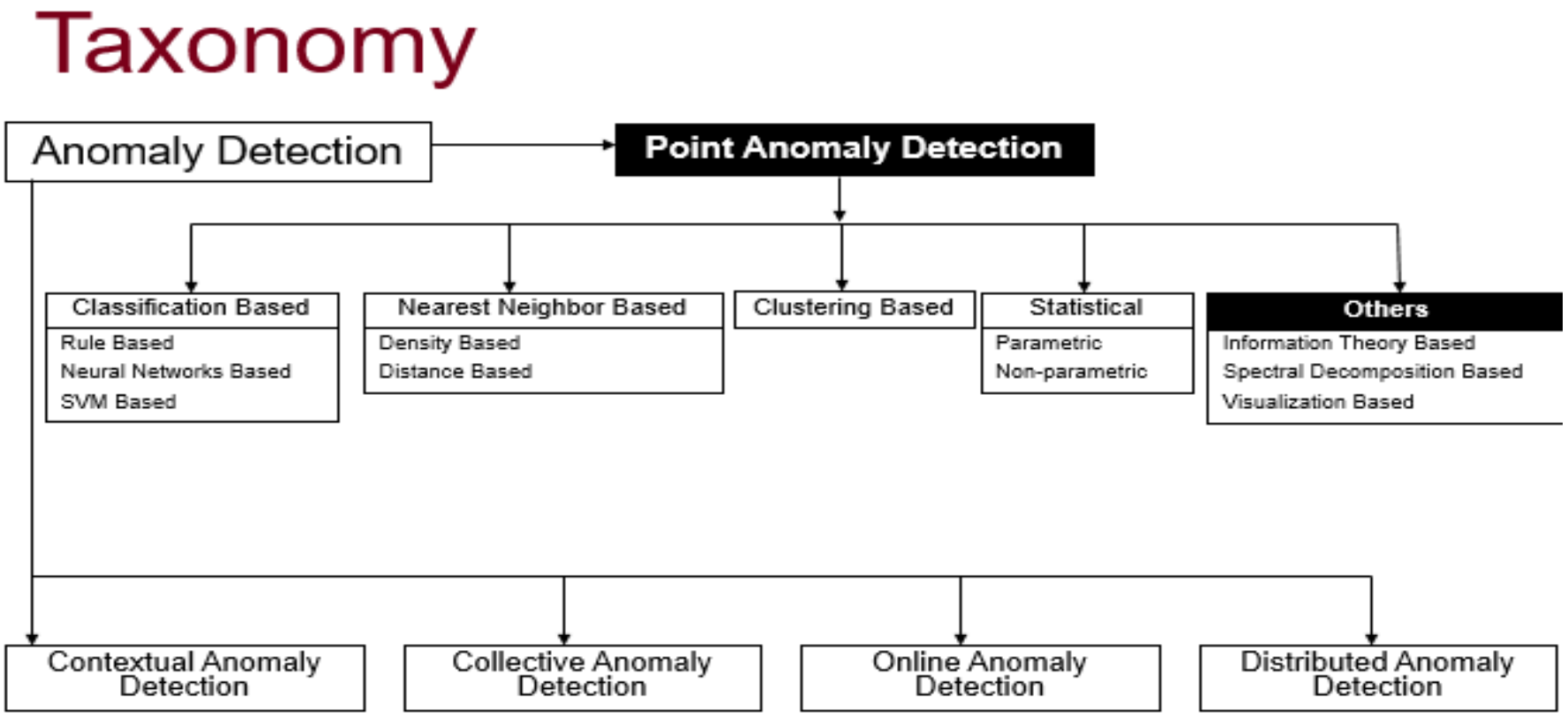


Figure: Others model techniques

Information theory

- In information theory, the main purpose is for one person (a transmitter) to send a message to another (the recipient) over a path.

- To do that, the transmitter sends a sequence of partial messages (possibly one) that provide hints to the original message.

- The details value in each of these incomplete communications is an indicator of how ambiguous it is for the receiver. For starters:
  - A partial message which reduces in half the number of possibilities transmits a bit of message details.
  - If, for example, the transmitter wanted to submit the output of a randomly chosen digit to the recipient, the partial response of "the number is odd" would give a bit of details.

Contextual anomaly based

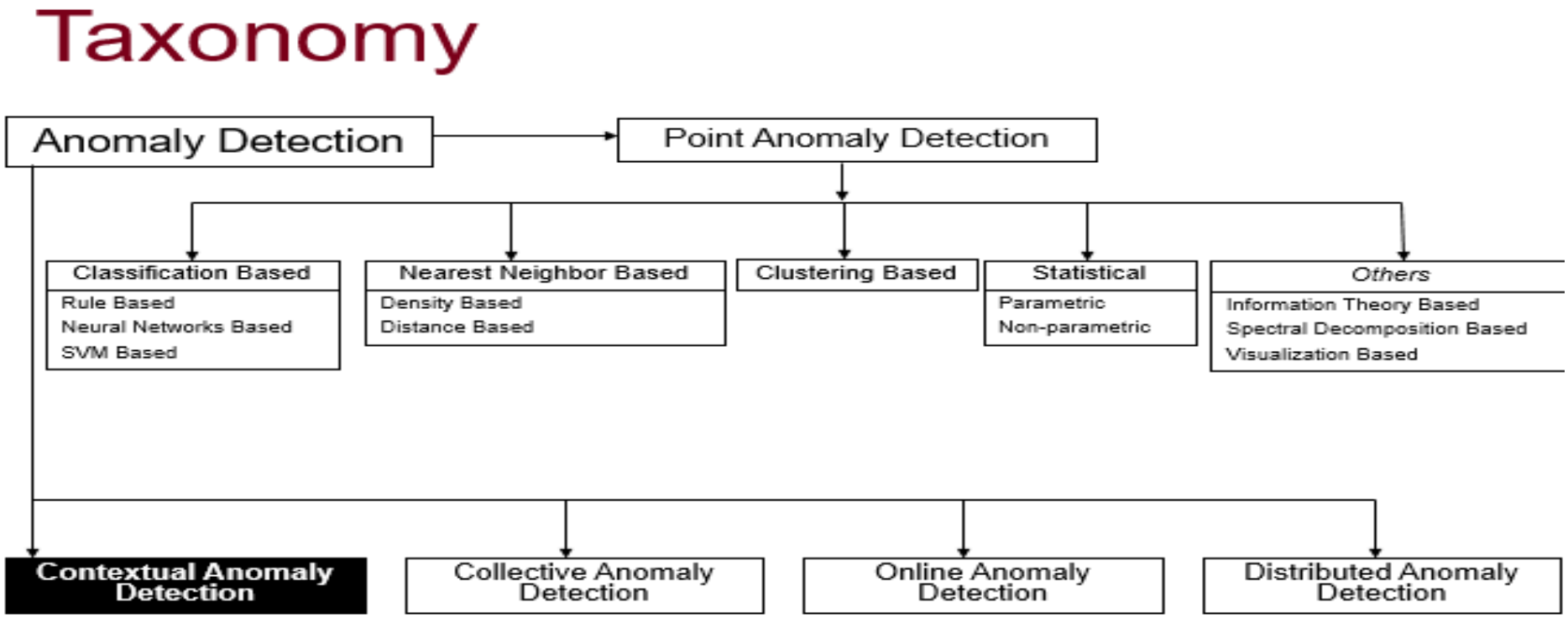


Figure: Contextual anomaly

Self evaluation: Exercise 17

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 17: Influenced Outlierness (INFLO).

Collective anomaly detection

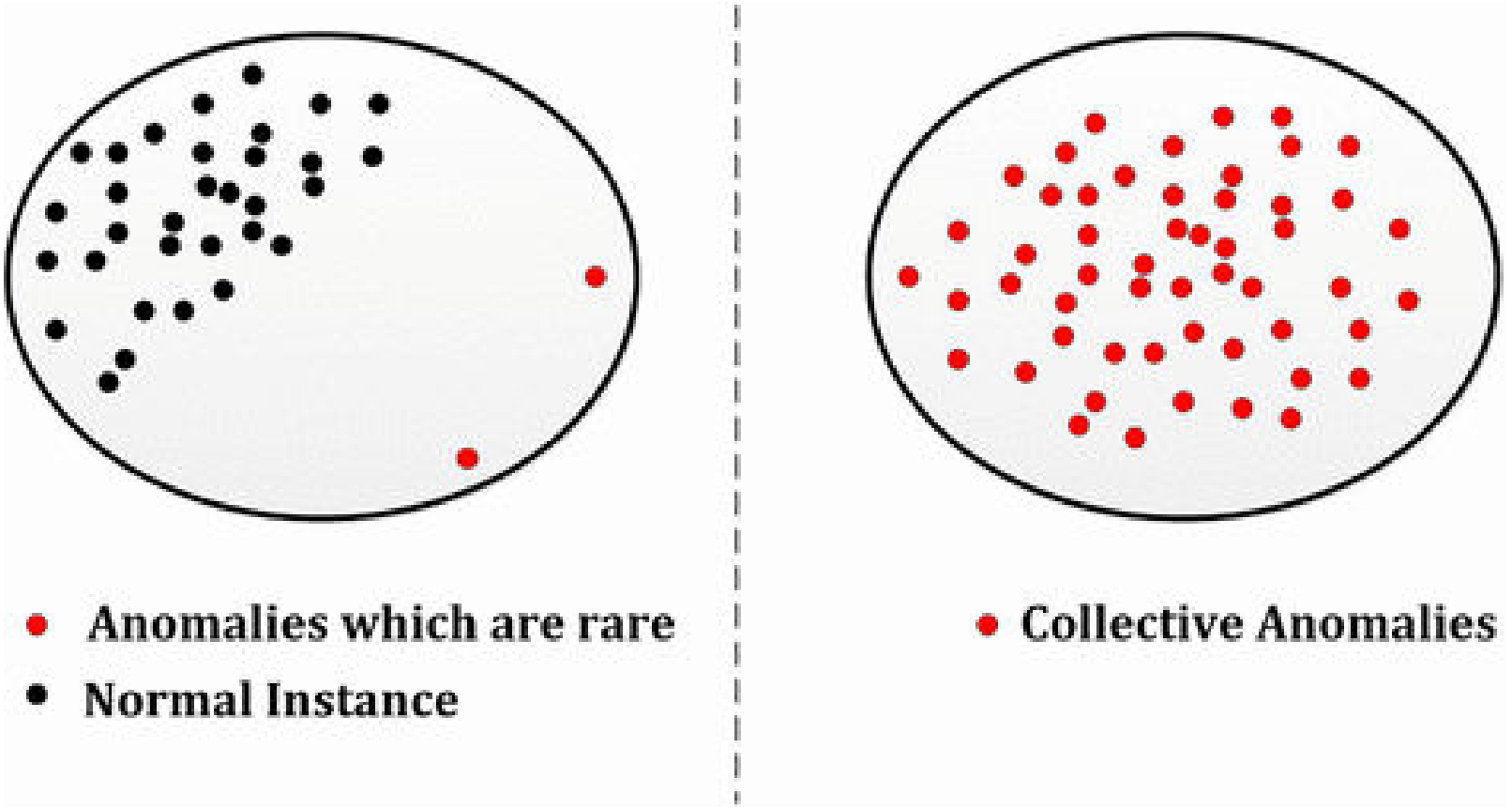


Figure: Collective Anomaly Detection

Source: https://images.app.goo.gl/Uw64MLJ8k1ewrun37

On-line based model

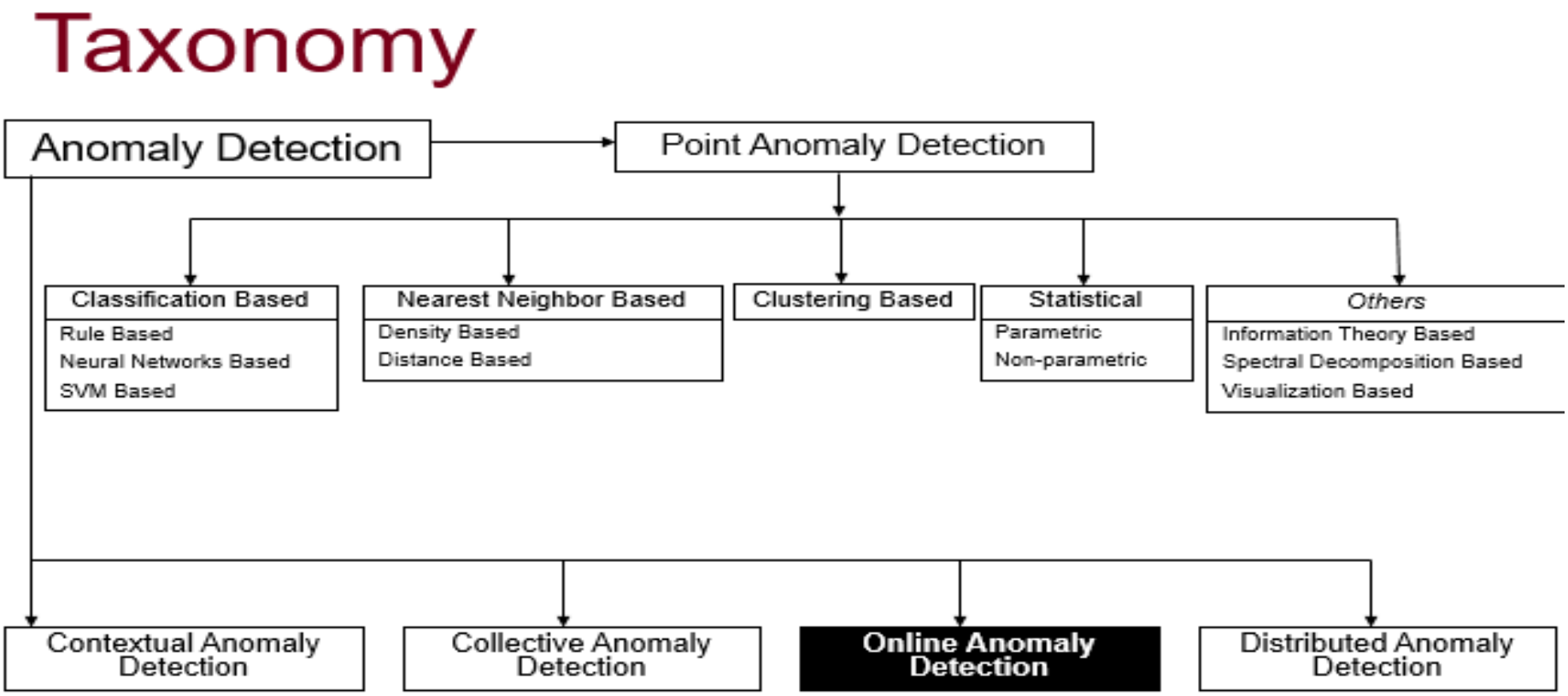


Figure: On-line based model

Distributed anomaly detection

- Data could originate from various sources in several phenomenon object tracking:
  - Intrusion prevention network.
  - Misuse by payment card.
  - Health of aircraft.

- Examination of information from a data position will un-detect errors that take place at many points concurrently:
  - In such hierarchical structures, abnormalities can be observed by combining information from predefined intervals on identified abnormalities in order to identify irregularities at national level in a complicated web.

- Good efficiency and decentralized architectures are required for connection and anomaly incorporation.

Self evaluation: Exercise 18

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.
- You are instructed to write the following activities using python code.
- Exercise 18: Local Outlier Probability (LoOP).

IDS analysis strategy

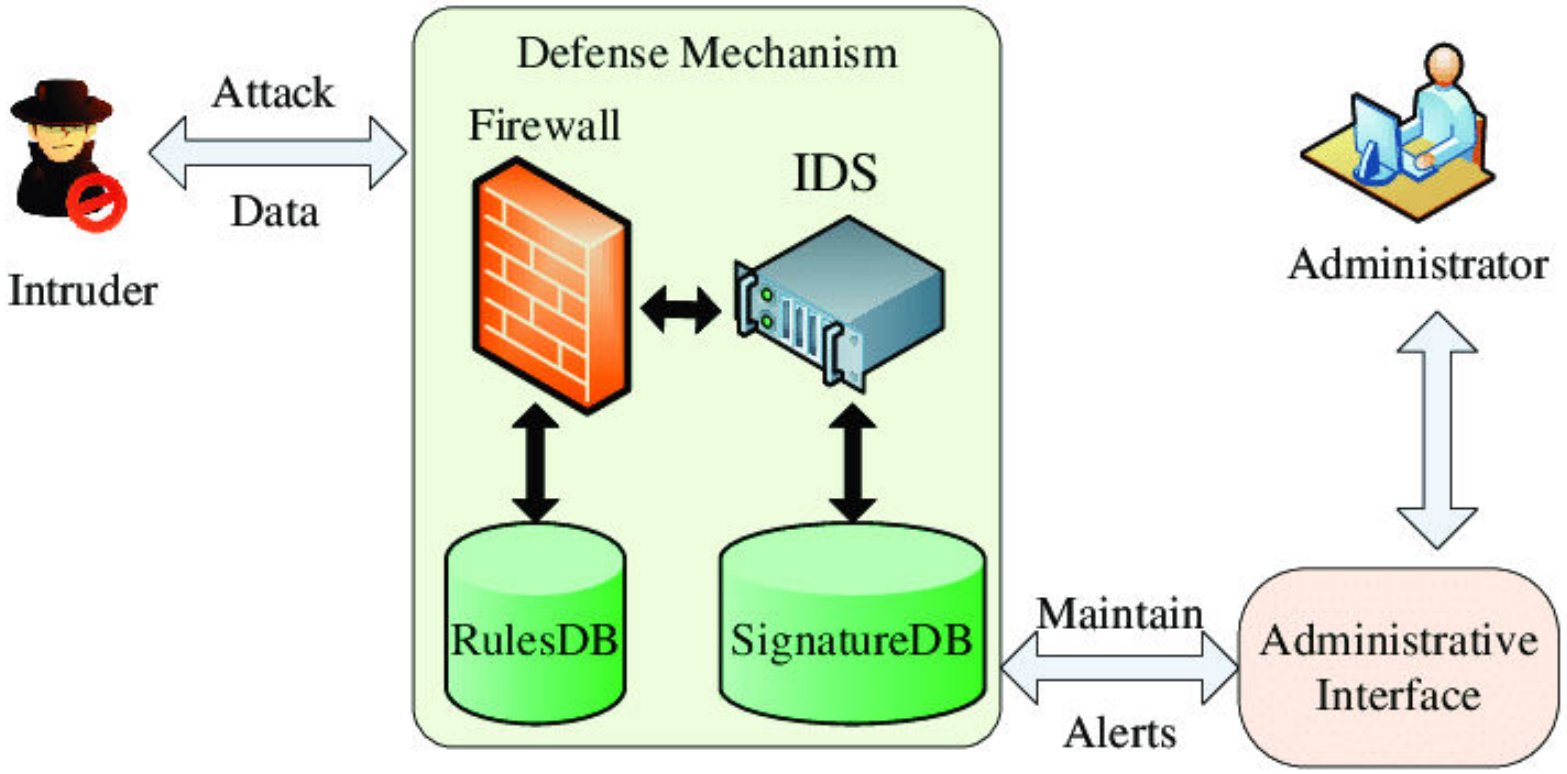


Figure: IDS

Source: https://images.app.goo.gl/gzuRHTQH2FAVeuVN8

Self evaluation: Exercise 19

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.
- You are instructed to write the following activities using python code.
- Exercise 19: Connectivity based Outlier Factor (COF).

Checkpoint (1 of 2)

Multiple choice questions:

1. What is unsupervised learning?
   a) Features of group explicitly stated
   b) Number of groups may be known
   c) Neither feature & nor number of groups is known
   d) None of the mentioned

2. What is plasticity in neural networks?
   a) Input pattern keeps on changing
   b) Input pattern has become static
   c) Output pattern keeps on changing
   d) Output is static

3. What are the tasks that cannot be realized or recognized by simple networks?
   a) Handwritten characters
   b) Speech sequences
   c) Image sequences
   d) All of the mentioned

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. What is unsupervised learning?
   a) Features of group explicitly stated
   b) Number of groups may be known
   c) Neither feature & nor number of groups is known
   d) None of the mentioned

2. What is plasticity in neural networks?
   a) Input pattern keeps on changing
   b) Input pattern has become static
   c) Output pattern keeps on changing
   d) Output is static

3. What are the tasks that cannot be realized or recognized by simple networks?
   a) Handwritten characters
   b) Speech sequences
   c) Image sequences
   d) All of the mentioned

Checkpoint (2 of 2)

Fill in the blanks:

1. Expectation maximization is an algorithm ____ in machine learning.
2. The only examples necessary to compute ____in an SVM is support vectors.
3. Averaging out the predictions of multiple _____will drastically reduce the variance.
4. The presence of _____(which leads to overfitting) is not generally a problem with weak classifiers.

True or False:

1. MAP estimates are equivalent to the ML estimates when the prior used in the MAP is a uniform prior over the parameter space. True/False
2. Because decision trees learn to classify discrete-valued outputs instead of real-valued functions it is impossible for them to over fit. True/False
3. If $P(A|B) = P(A)$ then $P(A \cap B) = P(A)P(B)$. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. Expectation maximization is a **clustering** algorithm in machine learning.
2. The only examples necessary to compute **f(x)** in an SVM is support vectors.
3. Averaging out the predictions of multiple **classifiers** will drastically reduce the variance.
4. The presence of **over-training** (which leads to overfitting) is not generally a problem with weak classifiers.

True or False:

1. MAP estimates are equivalent to the ML estimates when the prior used in the MAP is a uniform prior over the parameter space. **True**
2. Because decision trees learn to classify discrete-valued outputs instead of real-valued functions it is impossible for them to over fit. **False**
3. If $P(A|B) = P(A)$ then $P(A \cap B) = P(A)P(B)$. **True**

Question bank

Two marks question:

1. How would you handle an imbalanced dataset?
2. What's the F1 score? How would you use it?
3. Which is more important to you model accuracy, or model performance?
4. How is a decision tree pruned?

Four marks question:

1. How would you handle an imbalanced dataset?
2. When should you use classification over regression?
3. Name an example where ensemble techniques might be useful.
4. How do you ensure you're not overfitting with a model?

Eight marks question:

1. How would you evaluate a logistic regression model?
2. What's the "kernel trick" and how is it useful?

Unit summary

Having completed this unit, you should be able to:

- Understand the applications of anomaly detection
- Learn about example of classification-based methods
- Gain knowledge on nearest neighbor-based approach
- Gain an insight into clustering based methods
- Understand statistical approach, graph and model-based approach for ML implementation

Real-world problems

Unit objectives

After completing this unit, you should be able to:

- Understand the network intrusion detection
- Gain knowledge on anomaly detection in big data
- Understand anomaly detection for autonomous robots

Network intrusion detection

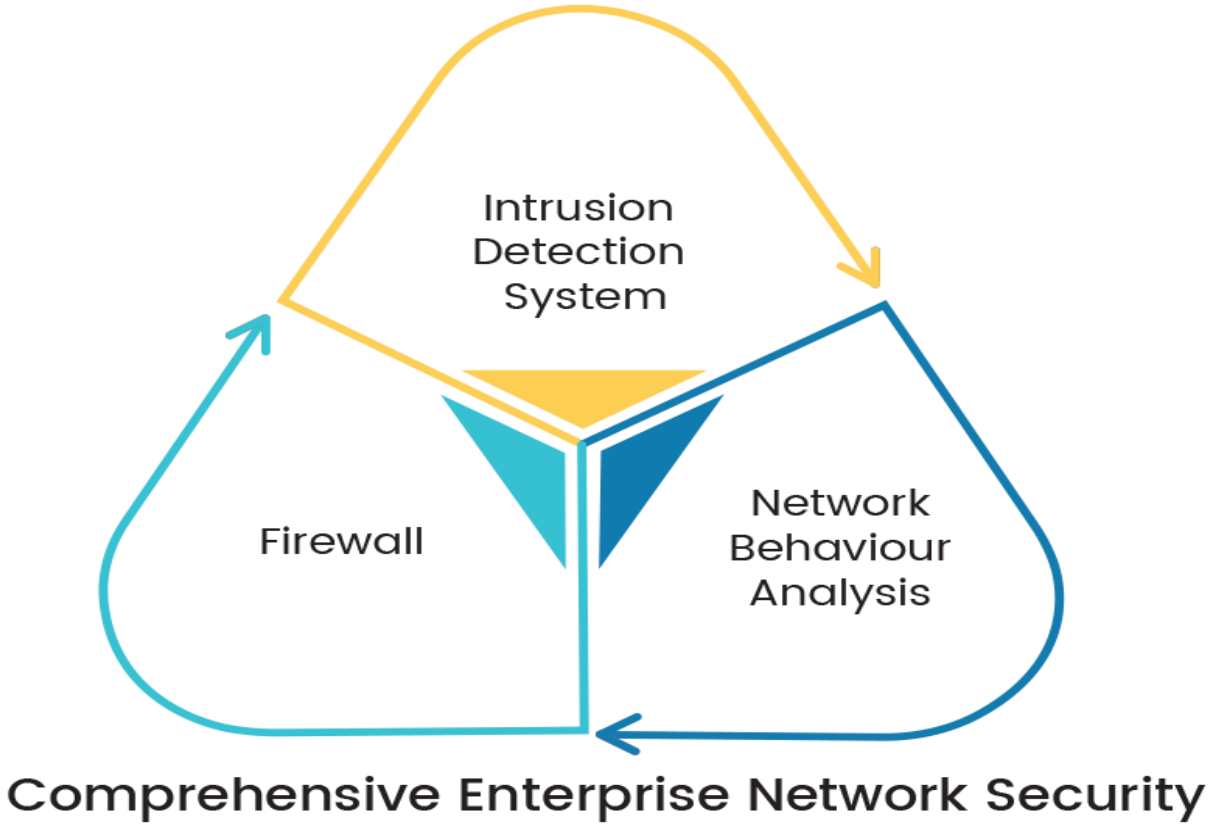


Figure: Network intrusion detection

Source: https://images.app.goo.gl/ZWkhczDWBdatVbmJ8

Understanding of IDS core operation

- A computer or software program carry out these useful roles is an intrusion detection system (IDS):
  - Evaluates a full cyber threat network infrastructure.
  - Senses a cyber threat immediately as it happens.
  - Implements an anti-attack preventive measure (intrusion protection systems) easily.
  - Submit files to a monitoring department or operator.

How an IDS works?

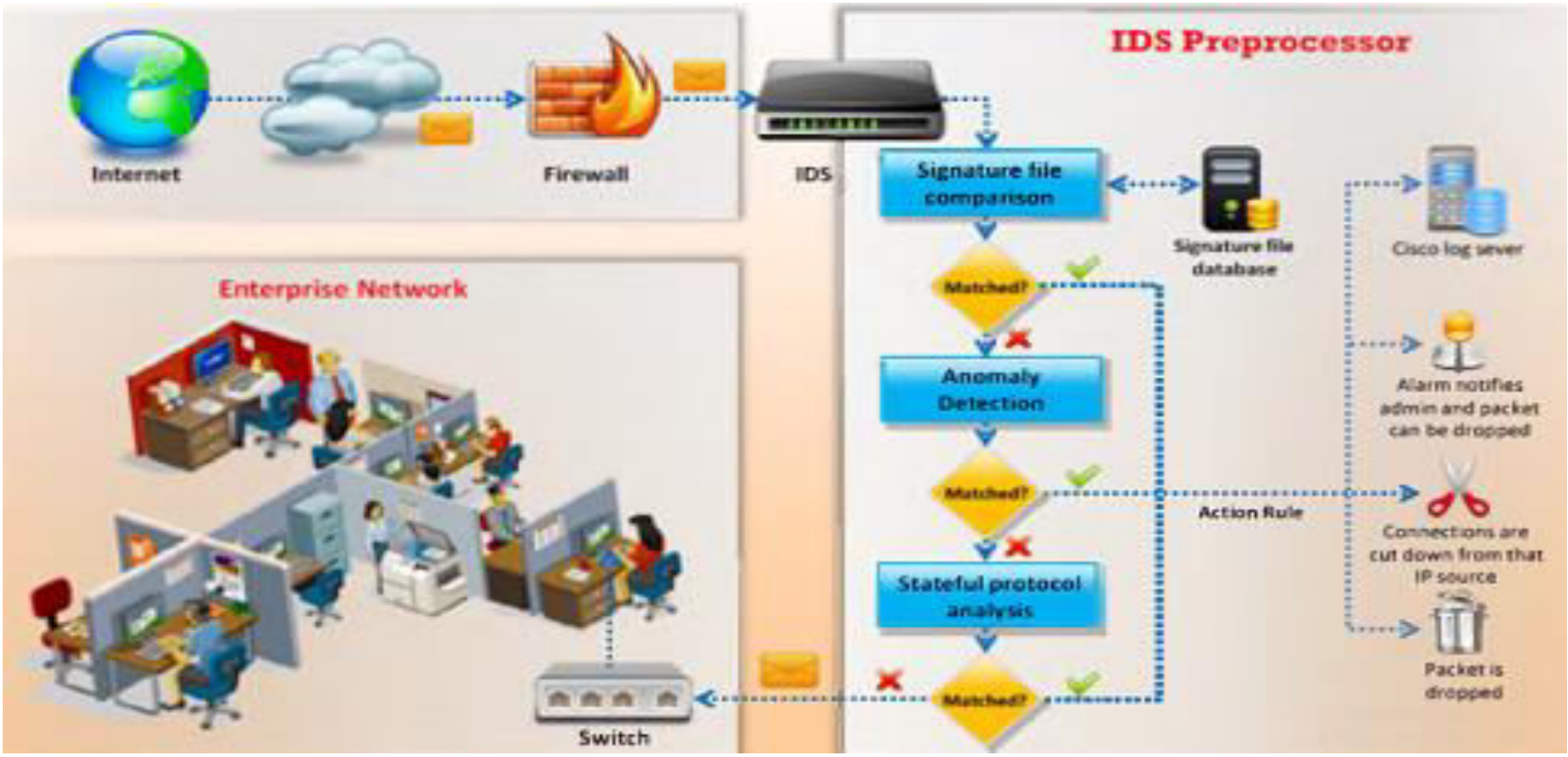


Figure: IDS working model

Source: https://images.app.goo.gl/FaBeWcdTYnL18rBD8

Types of intrusion detection systems

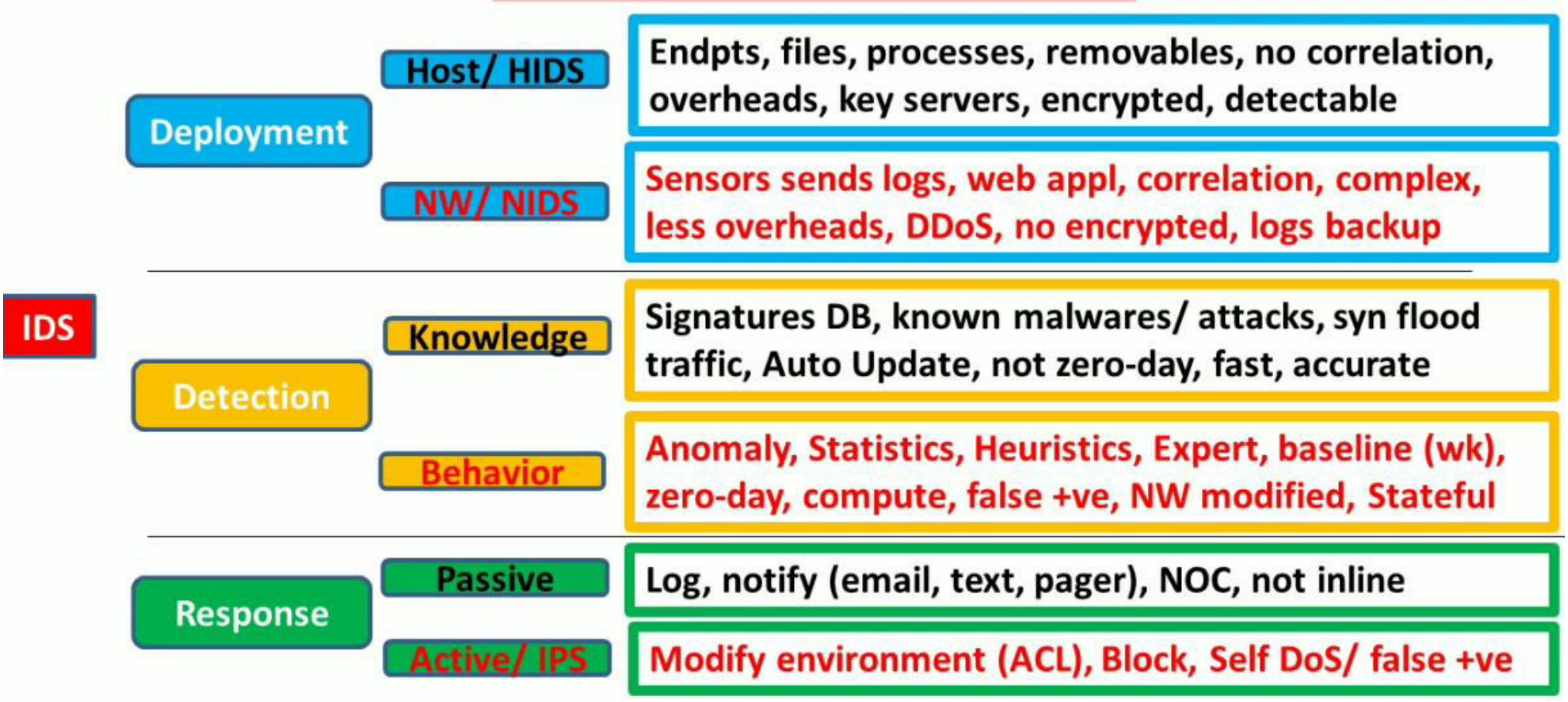


Figure :IDS categories

Source: https://images.app.goo.gl/kg6Hf9kXFMeCxxFfA

Self evaluation: Exercise 20

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 20: OpenCV (Object Detection with CAM).

Fundamental concerns of intrusion detection systems

- IDS is not scalable.
- False positives & negatives.
- Experienced administrators required.
- Encrypted packets.
- Protocol-based attacks.
- Ongoing updates.

Intrusion detection vs. intrusion prevention

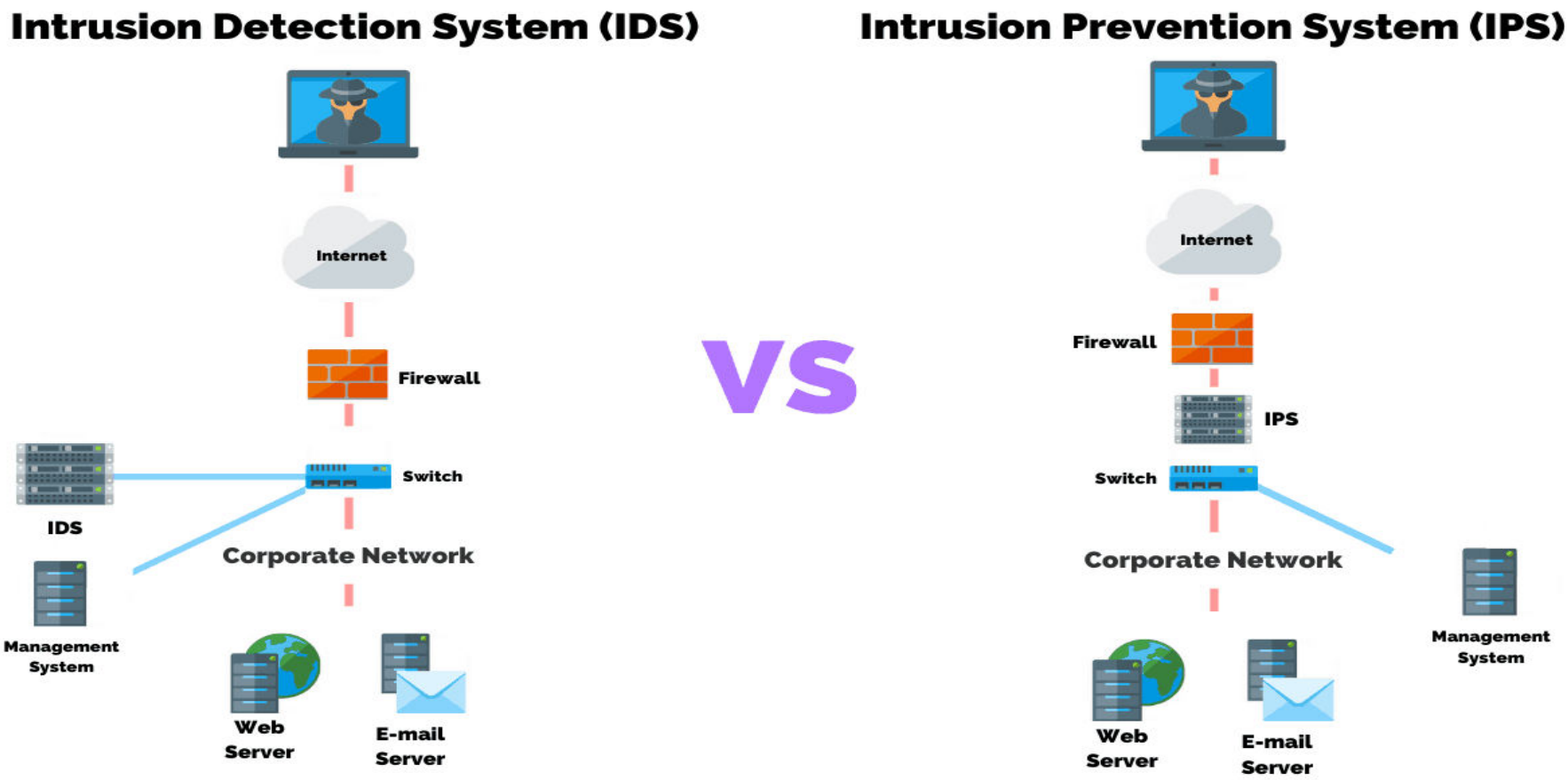


Figure: Intrusion Detection vs. Intrusion Prevention

Source: https://images.app.goo.gl/fXqCTJYByNixx6vT6

The future of IDS

Network Traffic Processing

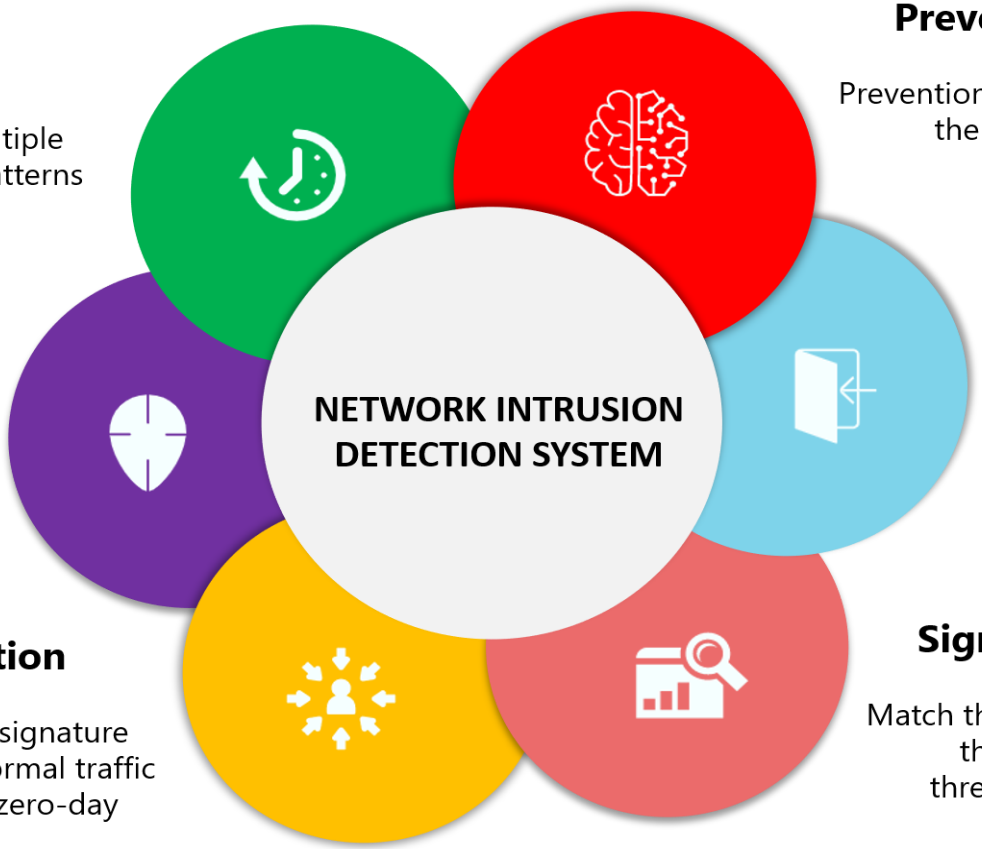
Convert the traffic into multiple network traffic parameter patterns (Signatures)

Prevention System

Prevention System acts based on the threat detected

Threat Reporting

Report/Log the threat



Threat Classification

Classify the threat based on Signature matching

Anomaly Detection

Match the input network signature against already available normal traffic signatures to detect the zero-day attacks

Signature Matching

Match the input signature against the already available threat signature patterns

Figure: Securing IoT with Intrusion Detection Systems

Source: https://images.app.goo.gl/yMGXBCuSCMGAHZuM9

Anomaly detection in big data

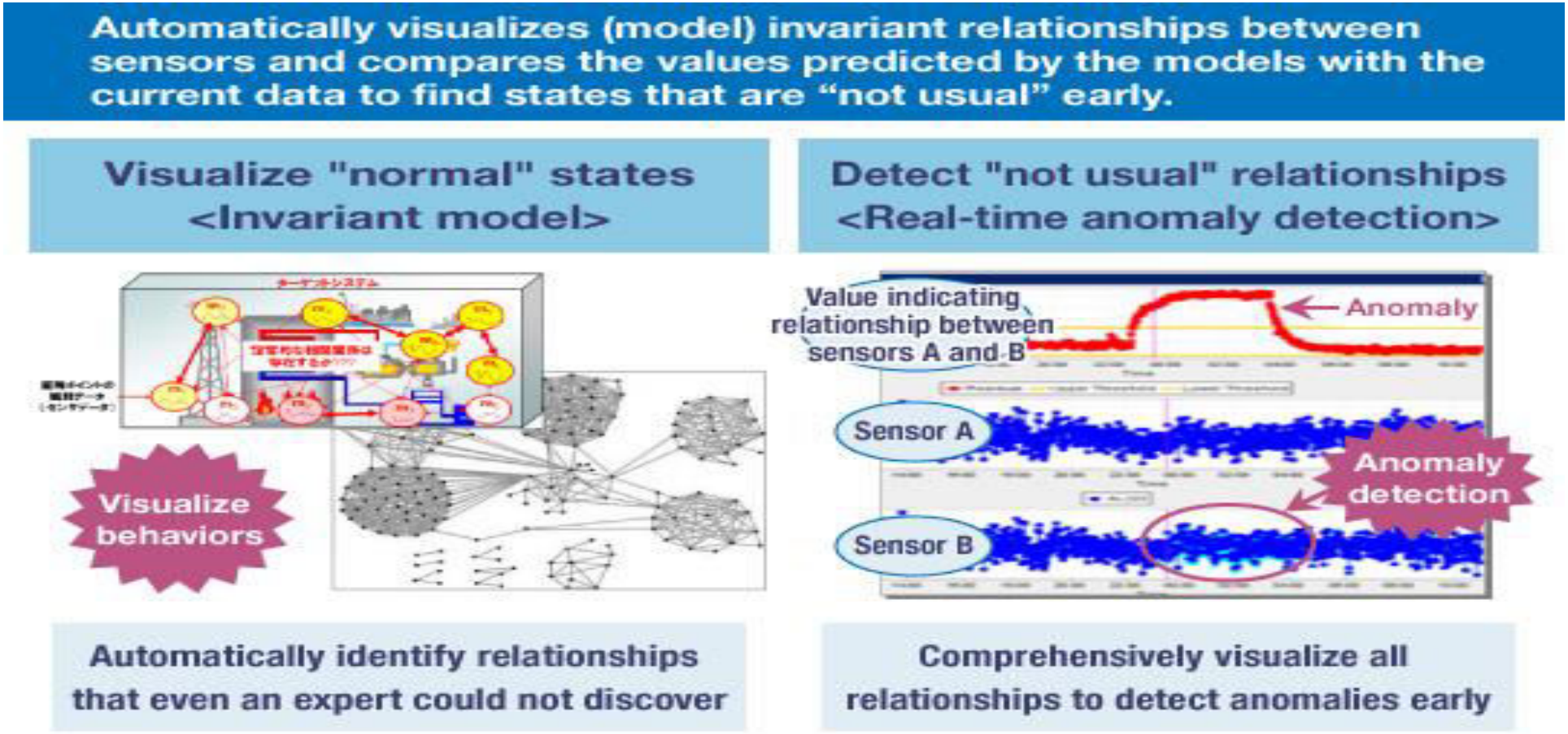


Figure: Bigdata visualization anomalies

Source: https://images.app.goo.gl/9SS6RTaHGHJXkLzk7

Key attributes of advanced anomaly detection

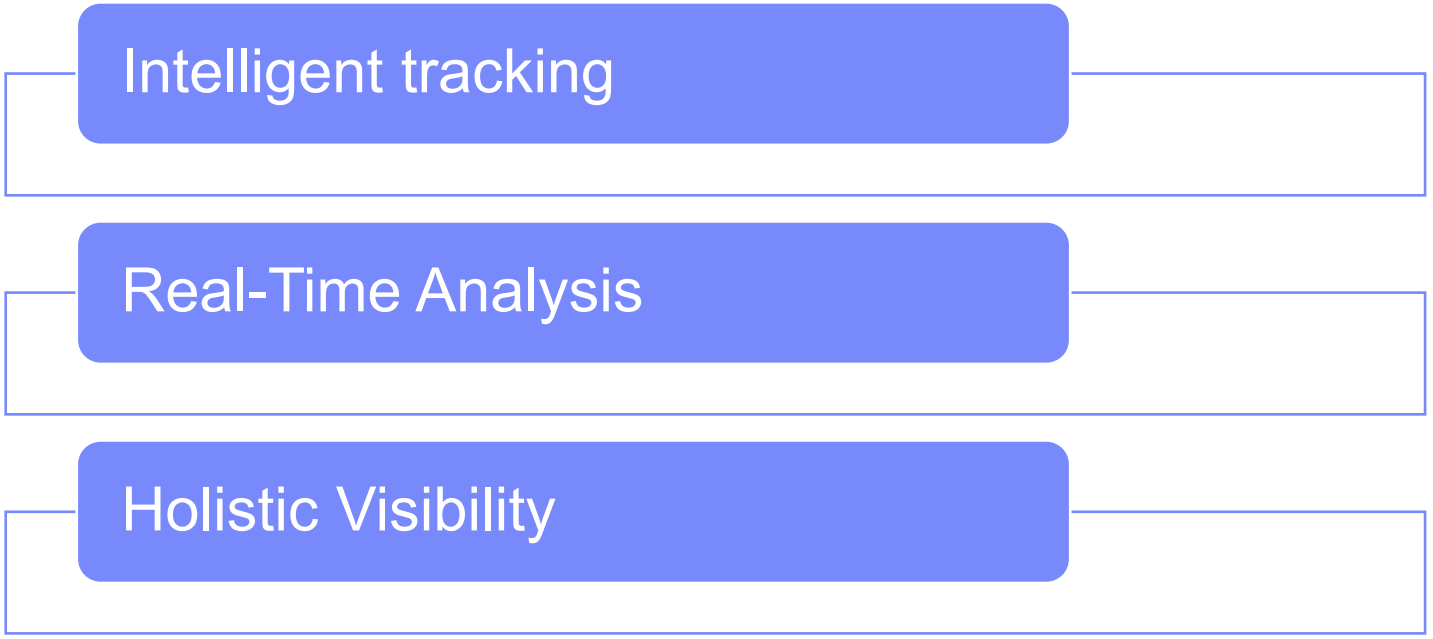


Figure: Key attributes of advanced anomaly detection

Self evaluation: Exercise 21

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 21: OpenCV (Object Detection with Video).

The real-world impact of anomaly detection

Figure: Robot dog reminds park goers about social distancing | Coronavirus

Source: https://images.app.goo.gl/RWug9Bhv44zNNJWz9

Anomaly detection on 5G: Possibilities and opportunities

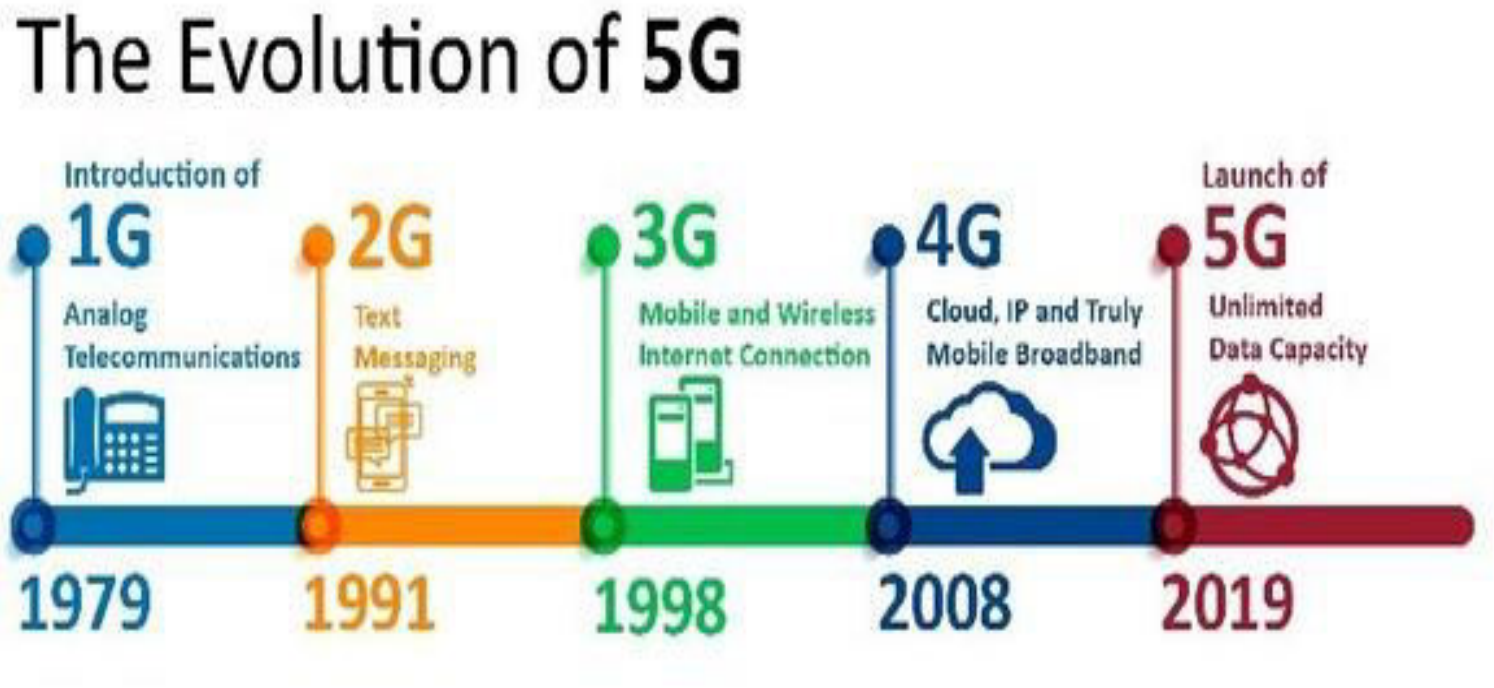


Figure: The Evolution of 5G

Source: https://images.app.goo.gl/43eqdfsmck7Cbhoa7

Real time anomaly detection in docker, Hadoop cluster

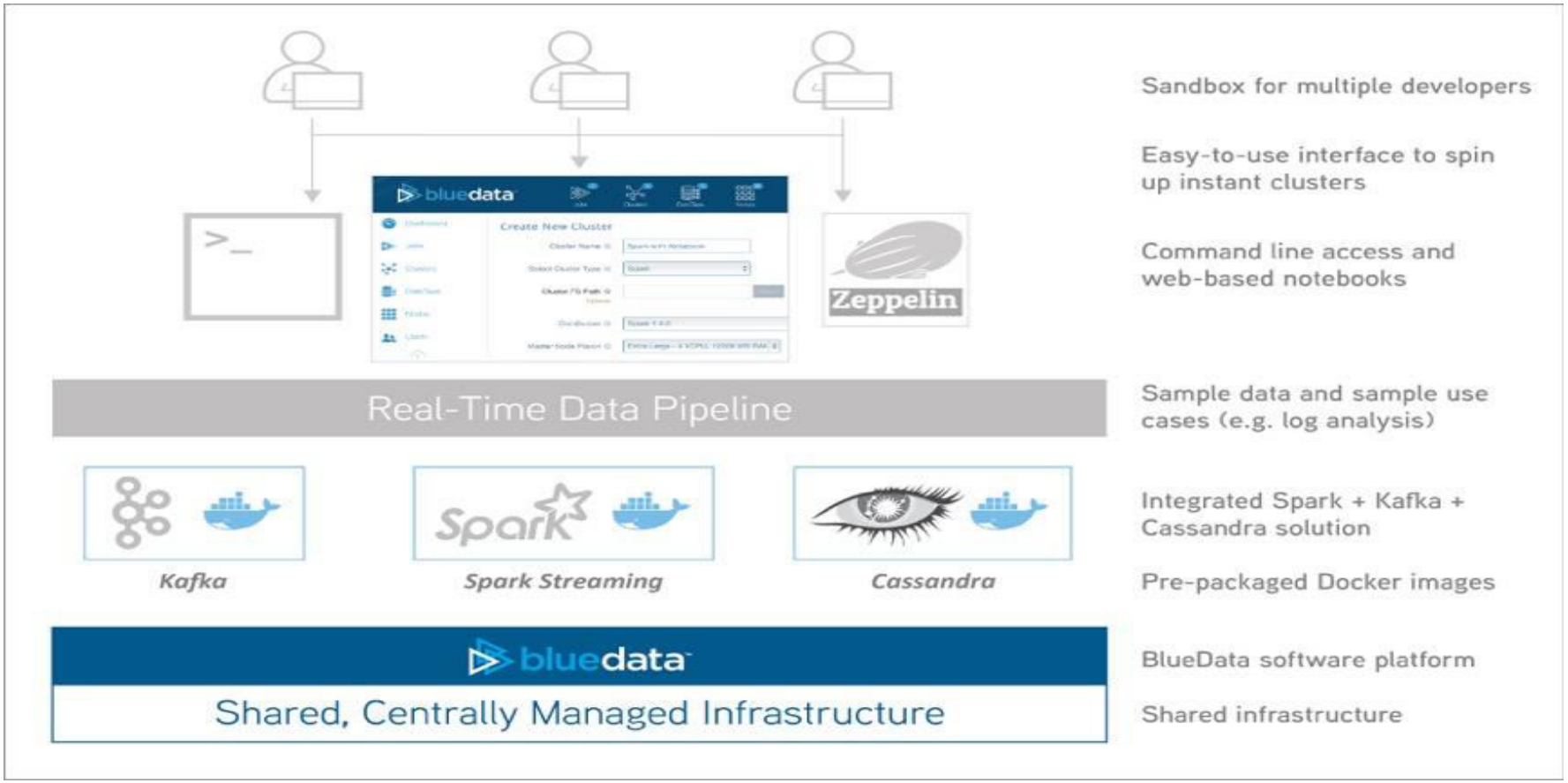


Figure: Real Time Anomaly Detection in Docker, Hadoop cluster

Source: https://images.app.goo.gl/u9hxmmh17NomAnJV8

Anomaly detection in IoT

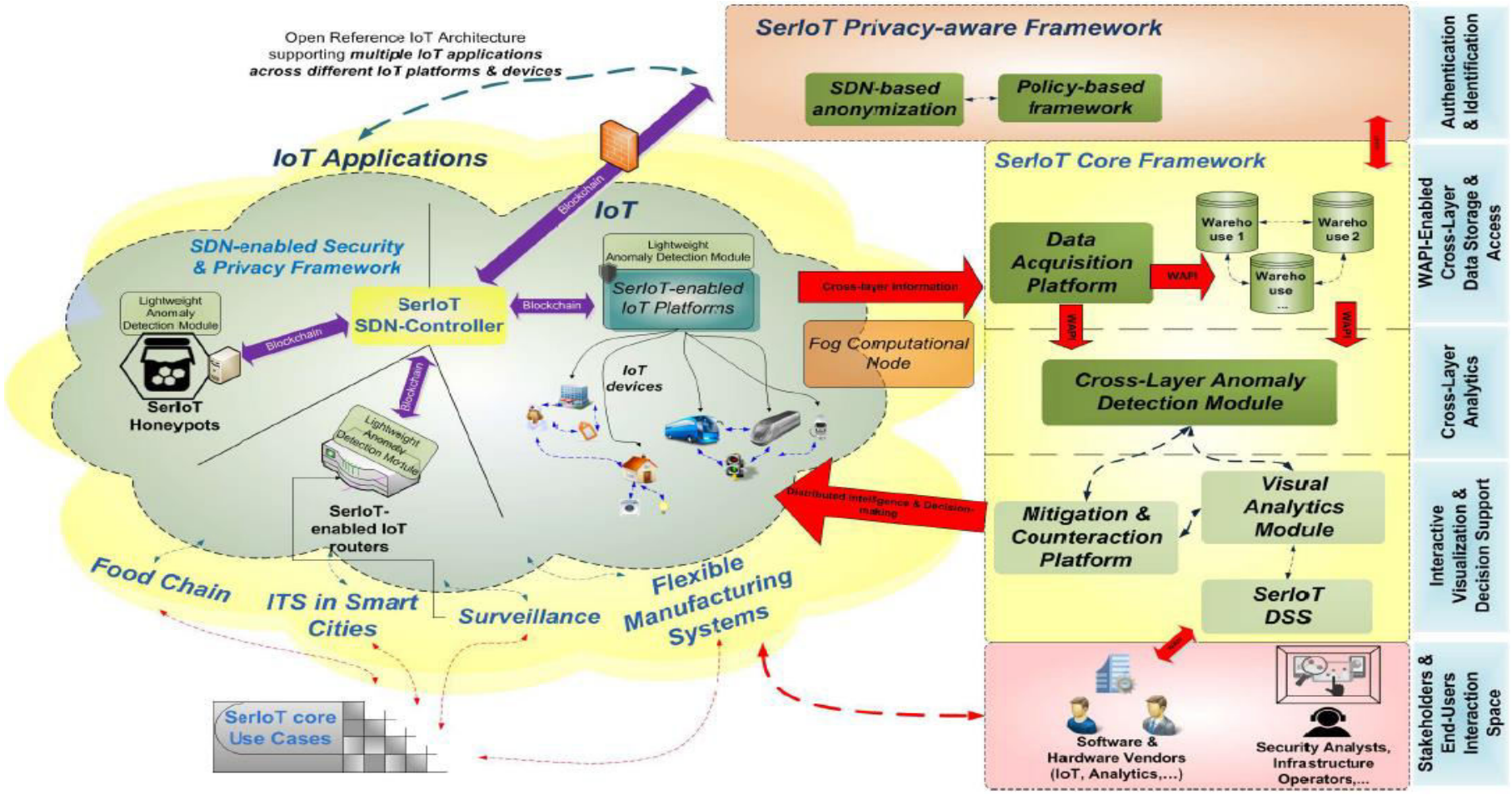


Figure: Anomaly Detection in IoT

Source: https://images.app.goo.gl/ayCxMUueCDq9XE2N8

Detection of deviations in deep learning time series results

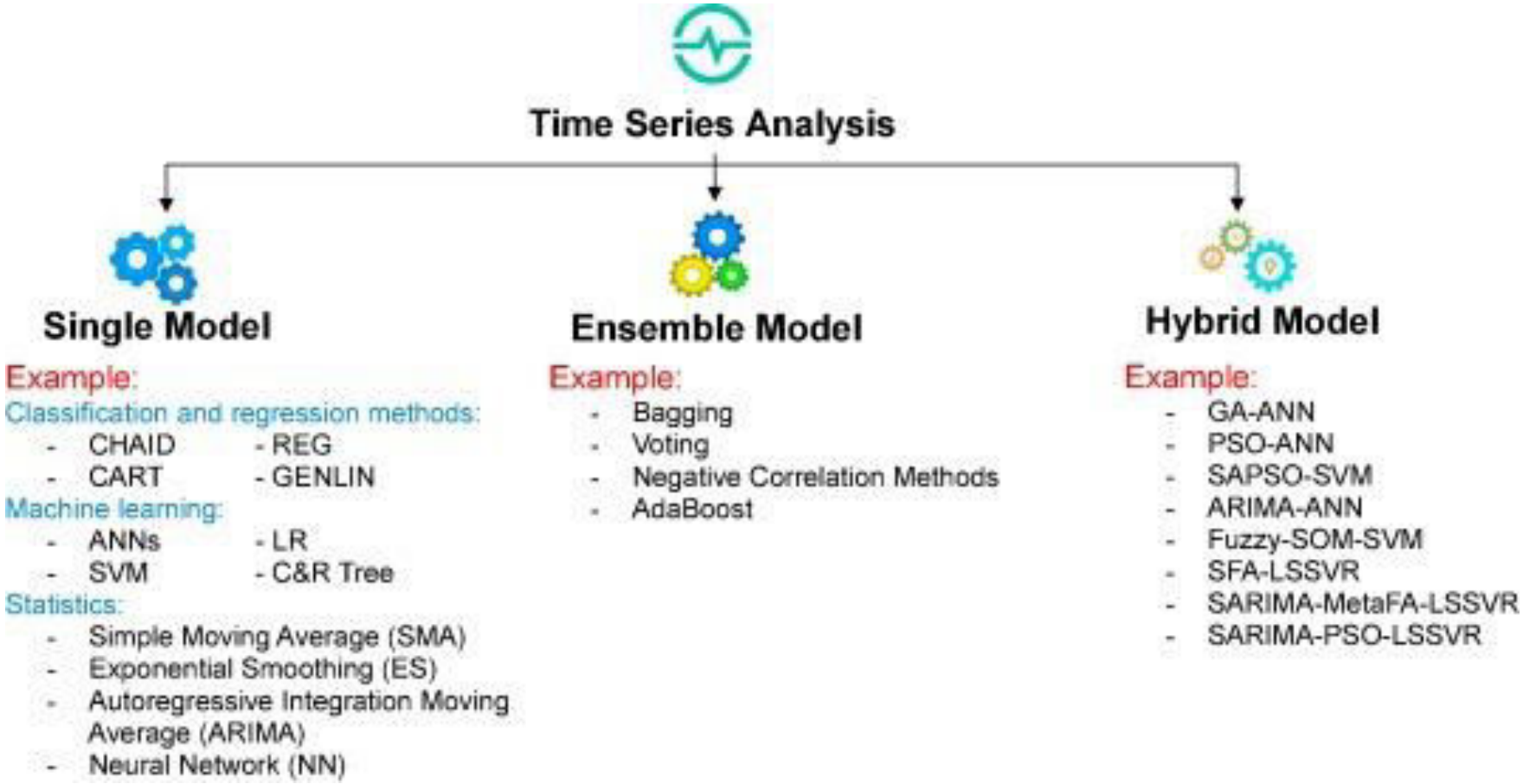


Figure: Detection of deviations in Deep Learning time series results

Source: https://images.app.goo.gl/2UUTc7Eq8w2pDDrV6

Self evaluation: Exercise 22

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 22: OpenCV (Color Filtration).

Anomaly detection use cases

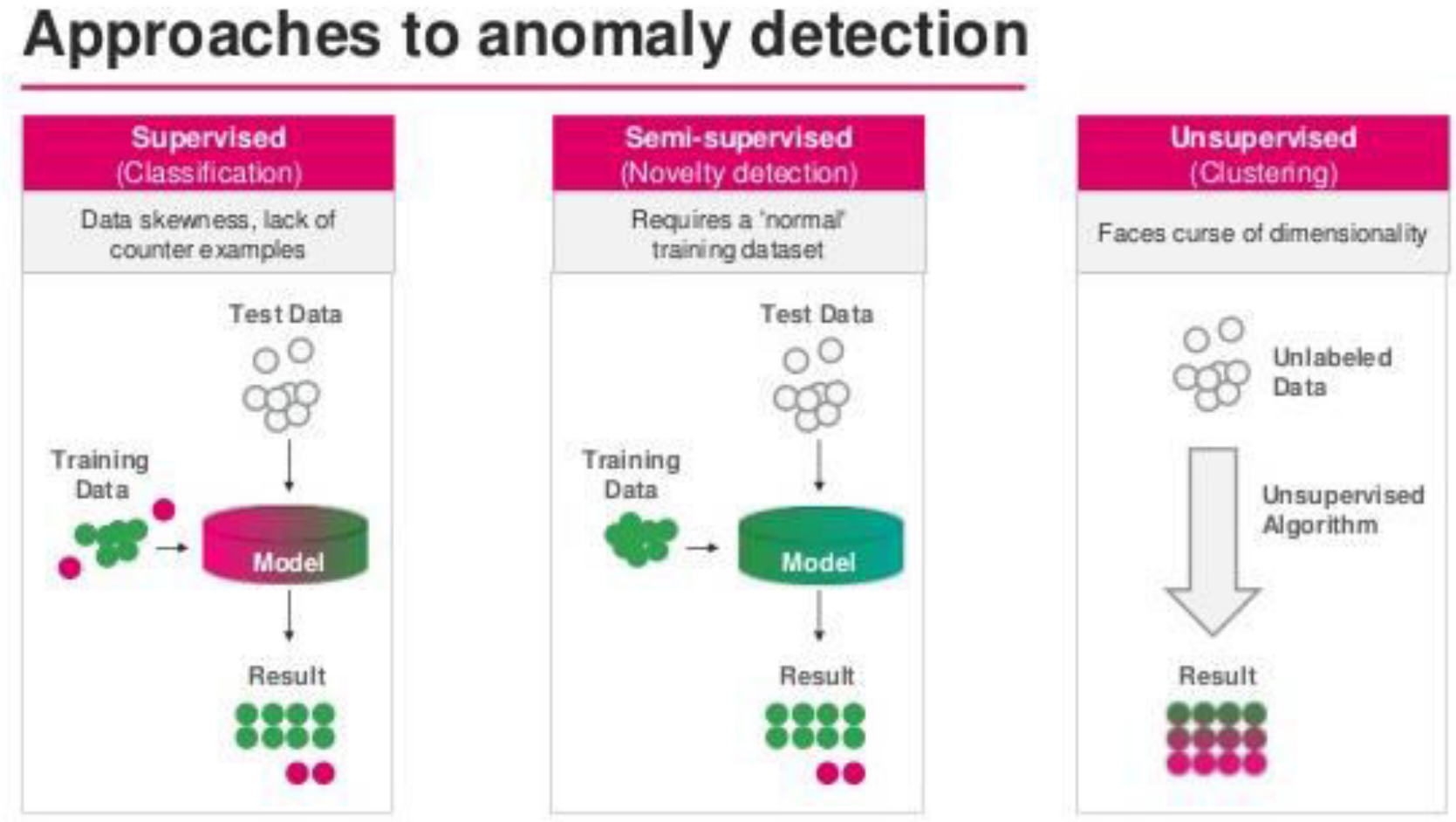


Figure: Approaches to anomaly detection

Source: https://images.app.goo.gl/KQnkiHCvv6Ryqa4V8

Anomaly detection with time series forecasting

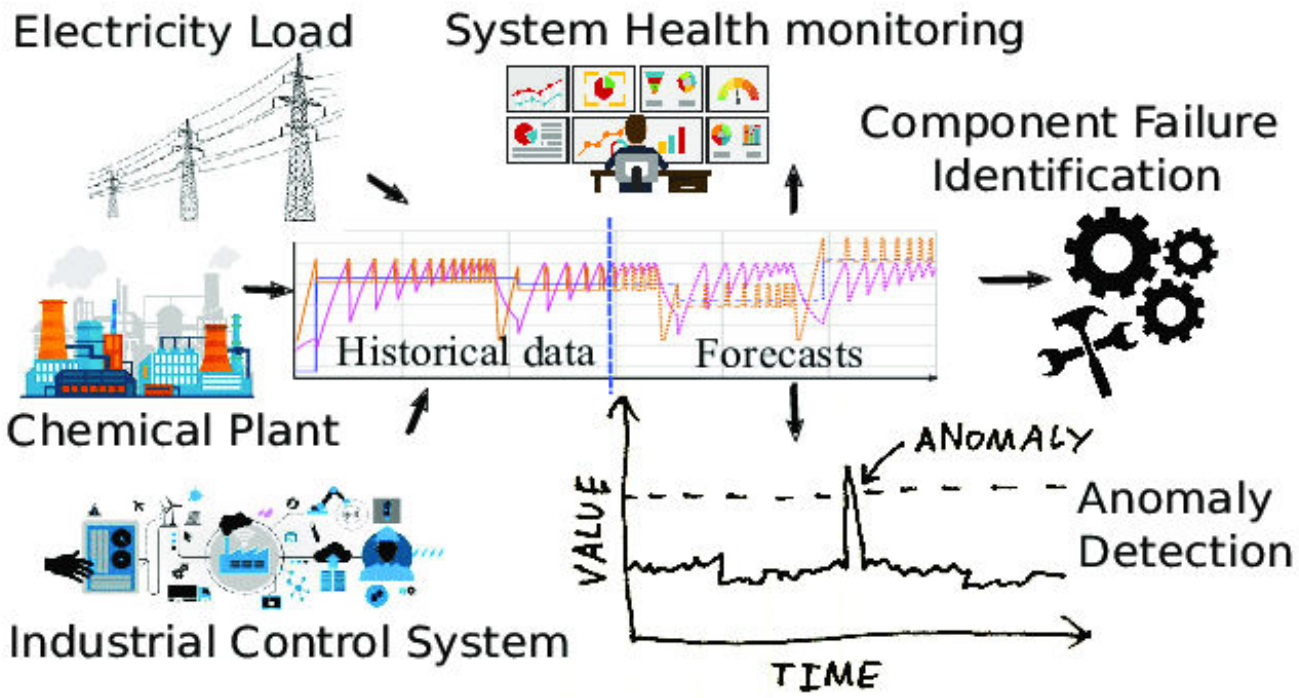


Figure: Anomaly Detection with Time Series Forecasting

Source: https://images.app.goo.gl/9NhyxHrSHdVr6nB78

Self evaluation: Exercise 23

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 23: OpenCV (Object Detection with haar cascade).

What is time series analysis?

- Time period analysis is also time period research. In attempt to elucidate the framework and role provided by the time series, the study is carried out.

- A mathematical model may be conveniently established by knowing the time series process, such that additional forecasts, surveillance and management could be made.

Time series data models

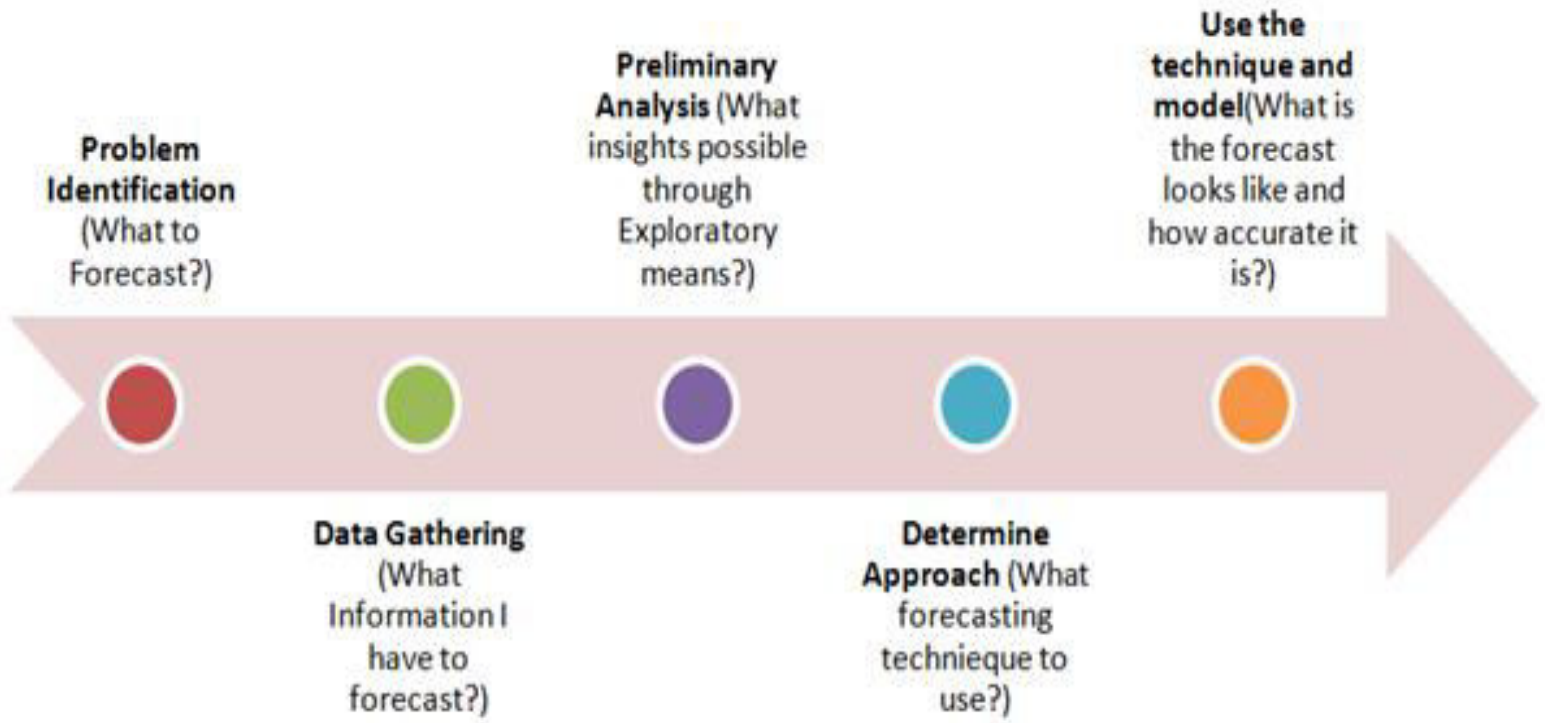


Figure: ARIMA forecasting process

Source: https://images.app.goo.gl/RZ4yJXcDvFZqv7R96

Self evaluation: Exercise 24

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

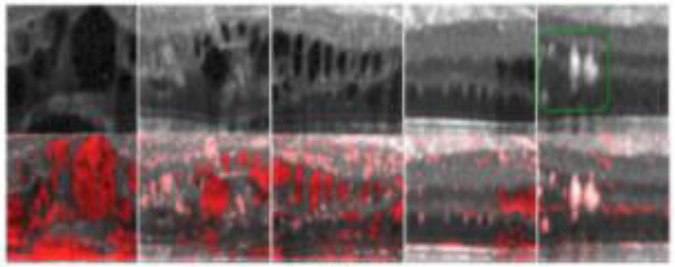
- Exercise 24: Graph theory.

How to find anomaly in time series data?

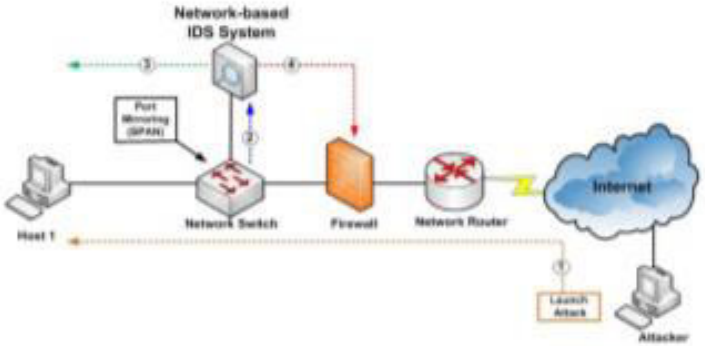
Figure: Anomaly in time series data

Anomaly detection using machine learning

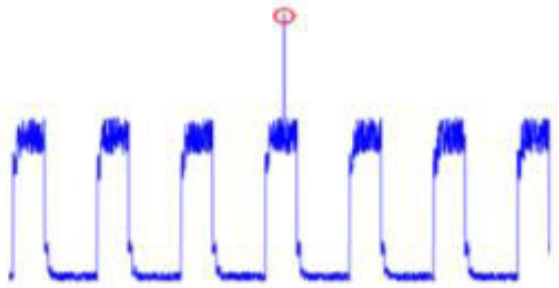
(a) Illegal Traffic Flow detection



(b) Detecting Retinal Damage



(c) Cyber-Network Intrusion detection



(d) Internet Of Things (IoT) Big-Data Anomaly detection

Figure: Anomaly Detection using Machine Learning

Source: https://images.app.goo.gl/3qwPTMzDRUUU7Ett6

Anomaly detection using deep learning

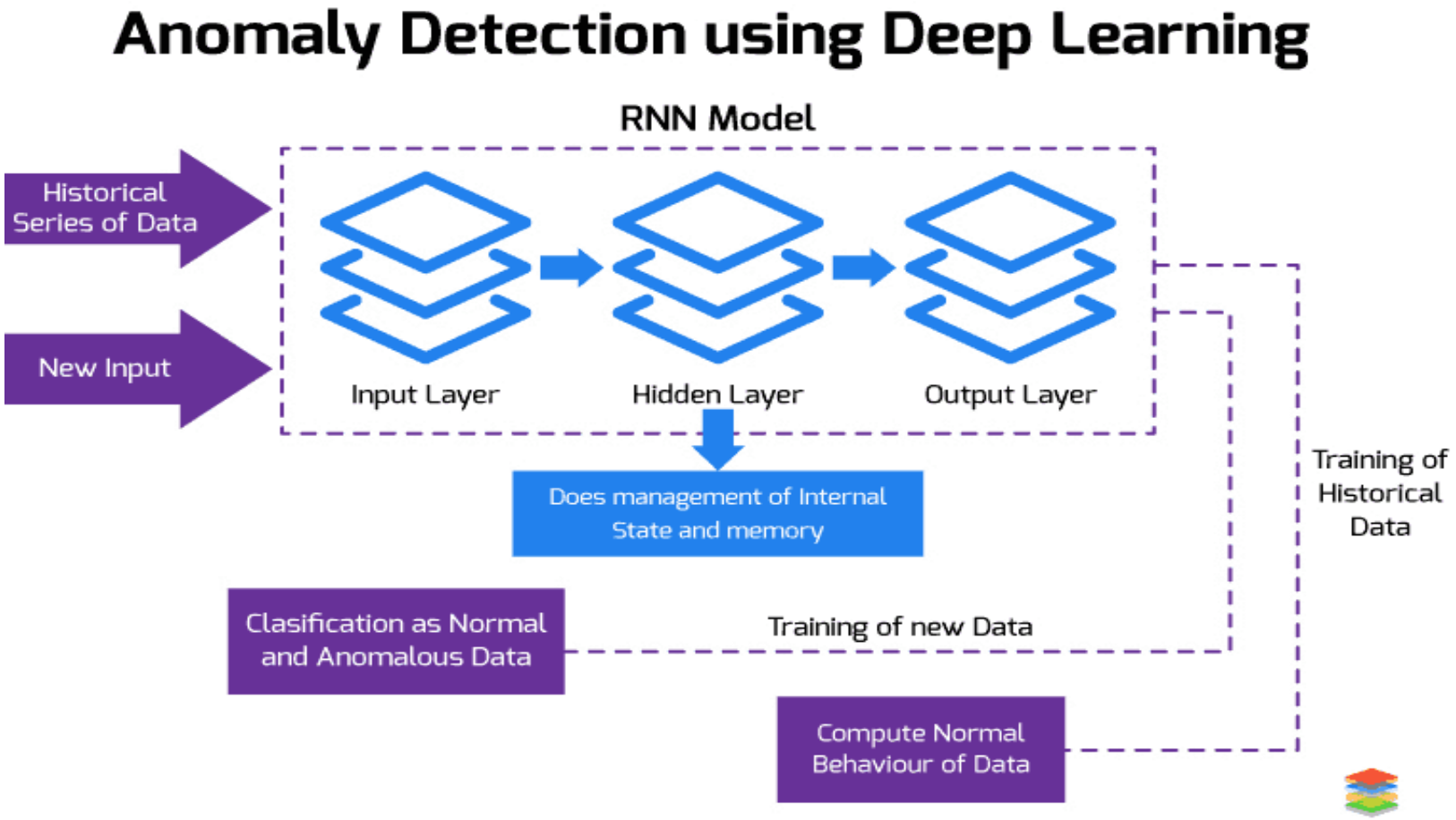


Figure: Anomaly Detection using Deep Learning

Source: https://images.app.goo.gl/NrZJ3oR11XrByVZA8

Anomaly detection for an e-commerce pricing system

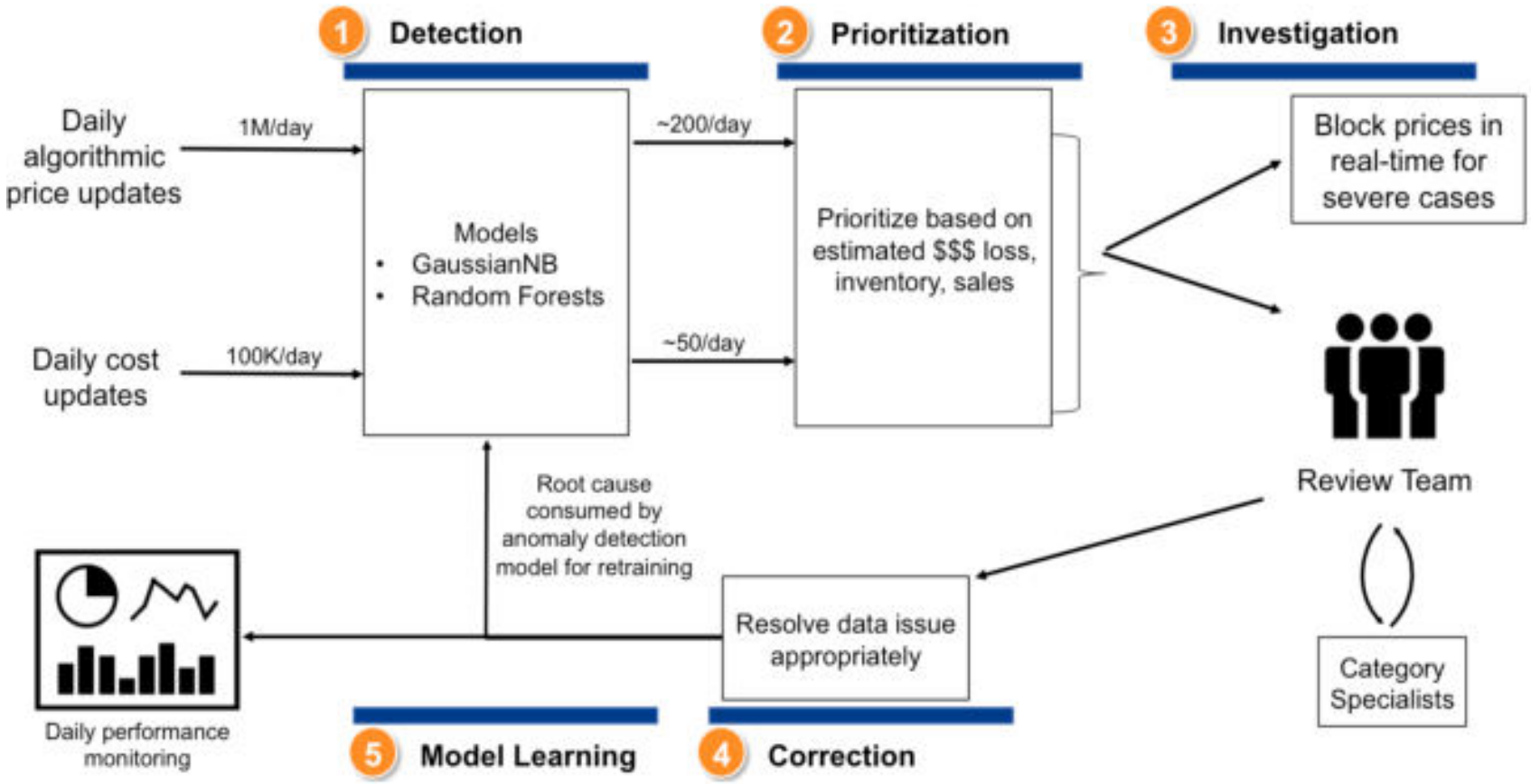


Figure: Anomaly Detection for an E-commerce Pricing System

Source: https://images.app.goo.gl/MtbnBSJFoQha7vkp9

IBM's Watson AIOps automates IT anomaly detection and remediation

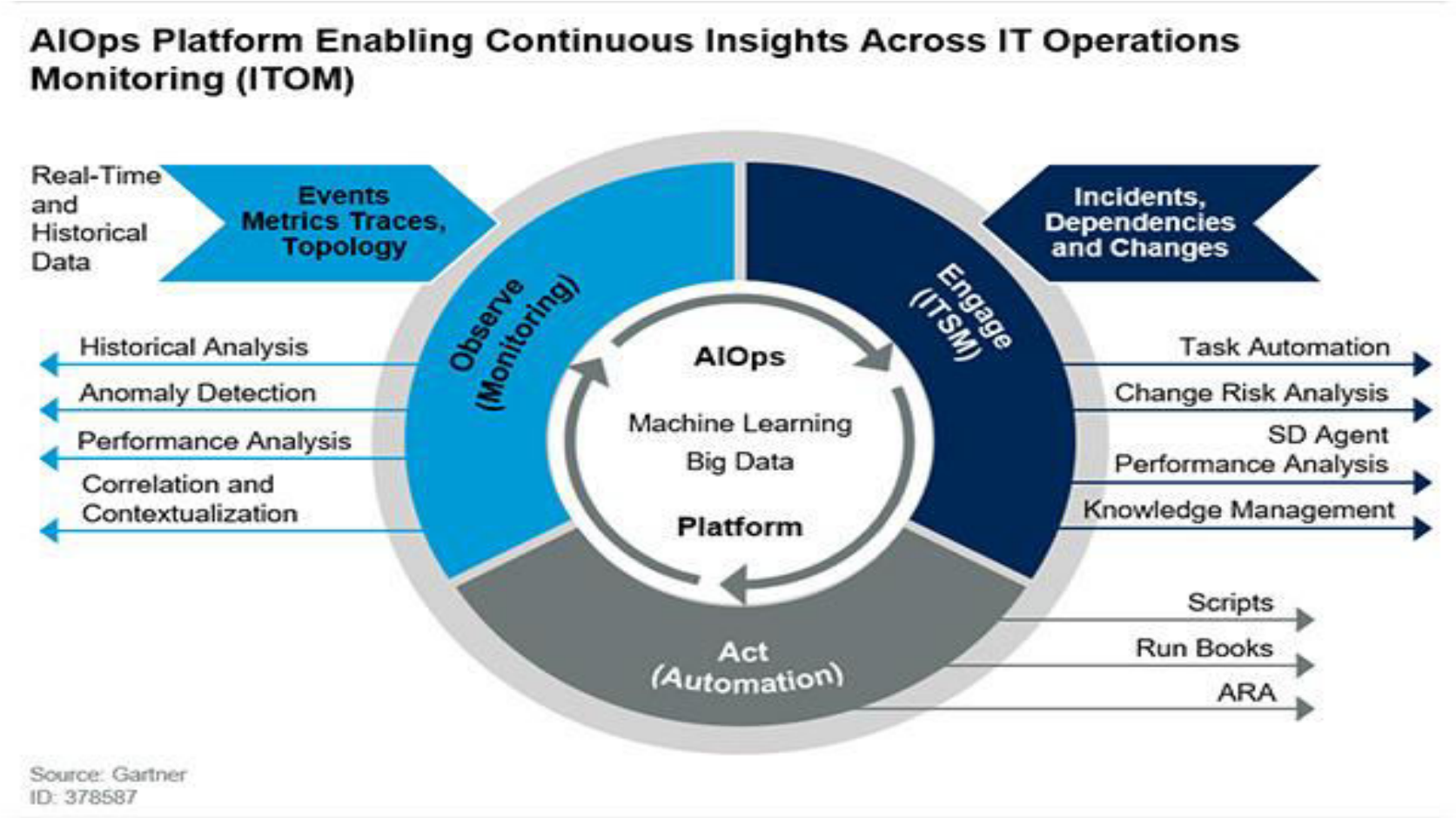


Figure: AIOps

Source: https://images.app.goo.gl/KC2fhvxbxnX6fFzS7

Self evaluation: Exercise 25

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.
- You are instructed to write the following activities using python code.
- Exercise 25: GUI for pattern detection.

Checkpoint (1 of 2)

Multiple choice questions:

1. The numerical output of a sigmoid node in a neural network:
   a) Is unbounded, encompassing all real numbers
   b) Is unbounded, encompassing all integers
   c) Is bounded between 0 and 1
   d) Is bounded between -1 and 1

2. What would you do in PCA to get the same projection as SVD?
   a) Transform data to zero mean
   b) Transform data to zero median
   c) Not possible
   d) None of these

3. Regarding bias and variance, which of the following statements are true?
   a) Models which overfit have a high bias
   b) Models which overfit have a low bias
   c) Models which underfit have a high variance
   d) Models which underfit have a low variance

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. The numerical output of a sigmoid node in a neural network:
   a) Is unbounded, encompassing all real numbers
   b) Is unbounded, encompassing all integers
   c) Is bounded between 0 and 1
   d) Is bounded between -1 and 1

2. What would you do in PCA to get the same projection as SVD?
   a) Transform data to zero mean
   b) Transform data to zero median
   c) Not possible
   d) None of these

3. Regarding bias and variance, which of the following statements are true?
   a) Models which overfit have a high bias
   b) Models which overfit have a low bias
   c) Models which underfit have a high variance
   d) Models which underfit have a low variance

Checkpoint (2 of 2)

Fill in the blanks:

1. In k nearest neighbors, k=1 increases the _____.
2. K-means _____ is guaranteed to converge to a local minimum.
3. _____(CDF) is a method to describe the distribution of random variables.
4. In kernelized SVMs, the _____ K has to be positive semi-definite.

True or False:

1. A cumulative distribution function (CDF) cannot be less than 0 or bigger than 1. True/False
2. K-Means Clustering is guaranteed to converge (i.e., terminate). True/False
3. Nearest neighbors is a parametric method. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. In k nearest neighbors, k=1 increases the **complexities**.
2. K-means **clustering** is guaranteed to converge to a local minimum.
3. **Cumulative Distribution Function (CDF)** is a method to describe the distribution of random variables.
4. In kernelized SVMs, the **kernel matrix** K has to be positive semi-definite.

True or False:

1. Acumulative distribution function (CDF) cannot be less than 0 or bigger than 1. **True**
2. K-Means Clustering is guaranteed to converge (i.e., terminate). **True**
3. Nearest neighbors is a parametric method. **False**

Question bank

Two mark questions:

1. What is overfitting, and how can you avoid it?
2. What is ‘training set’ and ‘test set’ in a machine learning model? How much data will you allocate for your training, validation, and test sets?
3. How do you handle missing or corrupted data in a dataset?
4. What is semi-supervised machine learning?

Four mark questions:

1. How can you choose a classifier based on a training set data size?
2. Describe the confusion matrix with respect to machine learning algorithms.
3. What is a false positive and false negative and how are they significant?
4. What are the three stages of building a model in machine learning?

Eight mark questions:

1. What is deep learning in modern businesses?
2. What are the applications of supervised machine learning in modern businesses?

Unit summary

Having completed this unit, you should be able to:

- Understand the network intrusion detection.
- Gain knowledge on anomaly detection in big data.
- Understand anomaly detection for autonomous robots.