**Unit I: Introduction To Big Data**

**Explain sampling with example**

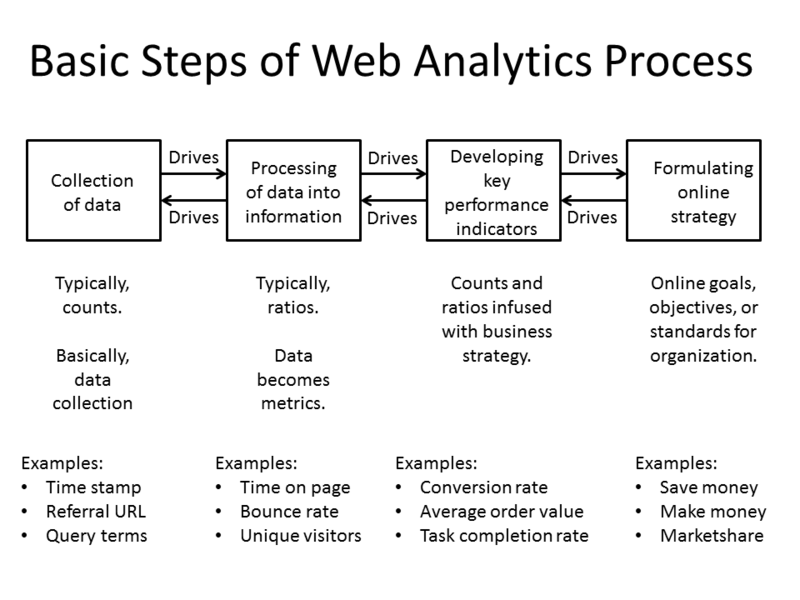
* Data sampling is a [statistical analysis](http://whatis.techtarget.com/definition/statistical-analysis) technique used to select, manipulate and analyze a representative subset of data points in order to identify patterns and trends in the larger [data set](http://whatis.techtarget.com/definition/data-set) being examined.
* Sampling allows [data scientists](http://searchbusinessanalytics.techtarget.com/definition/Data-scientist), predictive modelers and other data analysts to work with a small, manageable amount of data in order to build and run analytical models more quickly, while still producing accurate findings.
* Sampling can be particularly useful with data sets that are too large to efficiently analyze in full -- for example, in [big data analytics](http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics) applications.
* There are many different methods for drawing samples from data, depending on the data set and situation.
* Sampling can be based on [probability](http://whatis.techtarget.com/definition/probability), an approach that uses [random numbers](http://whatis.techtarget.com/definition/random-numbers) that correspond to points in the data set .
* This approach ensures that there is no correlation between points that are chosen for the sample.

**Example:**

* A search engine receives a stream of queries, and it would like to study the behavior of typical user.
* Suppose a user has issued search queries one time in the past month, d search queries twice, and no search queries more than twice. If we have a 1/10th sample, of queries, we shall see in the sample for that user an expected s/10 of the search queries issued once. Of the d search queries issued twice, only d/100 will appear twice in the sample; that fraction is d times the probability that both occurrences of the query will be in the 1/10th sample. Of the queries that appear twice in the full stream, 18d/100 will appear exactly once.
* To see why, note that 18/100 is the probability that one of the two occurrences will be in the 1/10th of the stream that is selected, while the other is in the 9/10th that is not selected.The correct answer to the query about the fraction of repeated searches is d/(s+d). However, the answer we shall obtain from the sample is d/(10s+19d).

**Web data and its analytics**

* **Web analytics** is the [measurement](https://en.wikipedia.org/wiki/Measurement), [collection](https://en.wikipedia.org/wiki/Data_collection), analysis and reporting of web [data](https://en.wikipedia.org/wiki/Data_(computing)) for purposes of understanding and optimizing [web usage](https://en.wikipedia.org/wiki/Web_usage).
* However, Web analytics is not just a process for measuring [web traffic](https://en.wikipedia.org/wiki/Web_traffic) but can be used as a tool for business and [market research](https://en.wikipedia.org/wiki/Market_research), and to assess and improve the effectiveness of a website. Web analytics applications can also help companies measure the results of traditional print or broadcast [advertising campaigns](https://en.wikipedia.org/wiki/Advertising_campaign).
* It helps one to estimate how traffic to a website changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a website and the number of page views. It helps gauge traffic and popularity trends which is useful for market research.



Basic Steps of Web Analytics Process

Most web analytics processes down to four essential stages or steps,which are:

* Collection of data: This stage is the collection of the basic, elementary data. Usually, this data is counts of things. The objective of this stage is to gather the data.
* Processing of data into information: This stage usually take counts and make them ratios, although there still may be some counts. The objective of this stage is to take the data and conform it into information, specifically metrics.
* Developing KPI: This stage focuses on using the ratios (and counts) and infusing them with business strategies, referred to as Key Performance Indicators (KPI). Many times, KPIs deal with conversion aspects, but not always. It depends on the organization.
* Formulating online strategy: This stage is concerned with the online goals, objectives, and standards for the organization or business. These strategies are usually related to making money, saving money, or increasing marketshare.

### **Web analytics technologies**

There are at least two categories of web analytics; *off-site* and *on-site* web analytics.

* Off-site web analytics refers to web measurement and analysis regardless of whether you own or maintain a website. It includes the measurement of a website's *potential* audience (opportunity), share of voice (visibility), and buzz (comments) that is happening on the Internet as a whole.
* On-site web analytics, the most common, measure a visitor's behavior once *on your website*. This includes its drivers and conversions; for example, the degree to which different [landing pages](https://en.wikipedia.org/wiki/Landing_page) are associated with online purchases. On-site web analytics measures the performance of your website in a commercial context. This data is typically compared against [key performance indicators](https://en.wikipedia.org/wiki/Key_performance_indicators) for performance, and used to improve a website or marketing campaign's audience response. [Google Analytics](https://en.wikipedia.org/wiki/Google_Analytics) and [Adobe Analytics](https://en.wikipedia.org/wiki/Adobe_Analytics) are the most widely used on-site web analytics service; although new tools are emerging that provide additional layers of information, including [heat maps](https://en.wikipedia.org/wiki/Heat_map) and [session replay](https://en.wikipedia.org/wiki/Session_replay).

**Processing tools for analytics of big data**

**1. Hadoop**

* Hadoop is one of the top frameworks in use today. So prevalent is it, that it has almost become synonymous with Big Data. But you already know about Hadoop, and MapReduce, and its ecosystem of tools and technologies including Pig, and Hive, and Flume, and HDFS. And all the others. Hadoop was first out of the gate, and enjoyed (and still does enjoy) widespread adoption in industry.
* If your data can be processed in batch, and split into smaller processing jobs, spread across a cluster, and their efforts recombined, all in a logical manner, Hadoop will probably work just fine for you.
* A number of tools in the Hadoop ecosystem are useful far beyond supporting the original MapReduce algorithm that Hadoop started as. Of particular note, and of a foreshadowing nature, is YARN, the resource management layer for the Apache Hadoop ecosystem. It can be used by systems beyond Hadoop, including Apache Spark.

**3. Flink**

* [Apache Flink](https://flink.apache.org/) is a streaming dataflow engine, aiming to provide facilities for distributed computation over streams of data. Treating batch processes as a special case of streaming data, Flink is effectively both a batch and real-time processing framework, but one which clearly puts streaming first.
* Flink provides a number of APIs, including a streaming API for Java and Scala, a static data API for Java, Scala, and Python, and an SQL-like query API for embedding in Java and Scala code. It also has its own machine learning and graph processing libraries.
* Flink has an impressive set of [additional features](https://flink.apache.org/features.html), including:
* High Performance & Low Latency
* Support for Event Time and Out-of-Order Events
* Exactly-once Semantics for Stateful Computations
* Continuous Streaming Model with Backpressure
* Fault-tolerance via Lightweight Distributed Snapshots

**4. Storm**

* [Apache Storm](http://storm.apache.org/) is a distributed real-time computation system, whose applications are designed as directed acyclic graphs.
* Storm is designed for easily processing unbounded streams, and can be used with any programming language. It has been benchmarked at processing over one million tuples per second per node, is highly scalable, and provides processing job guarantees.
* Unique for items on this list, Storm is written in Clojure, the Lisp-like functional-first programming language.
* Apache Storm can be used for real-time analytics, distributed machine learning, and numerous other cases, especially those of high data velocity.
* Storm can run on YARN and integrate into Hadoop ecosystems, providing existing implementations a solution for real-time stream processing. Five characteristics which make Storm ideal for real-time processing workloads are :
* Fast - benchmarked as processing one million 100 byte messages per second per node
* Scalable - with parallel calculations that run across a cluster of machines
* Fault-tolerant - when workers die, Storm will automatically restart them. If a node dies, the worker will be restarted on another node.
* Reliable - Storm guarantees that each unit of data (tuple) will be processed at least once or exactly once. Messages are only replayed when there are failures.
* Easy to operate - standard configurations are suitable for production on day one. Once deployed, Storm is easy to operate.

.

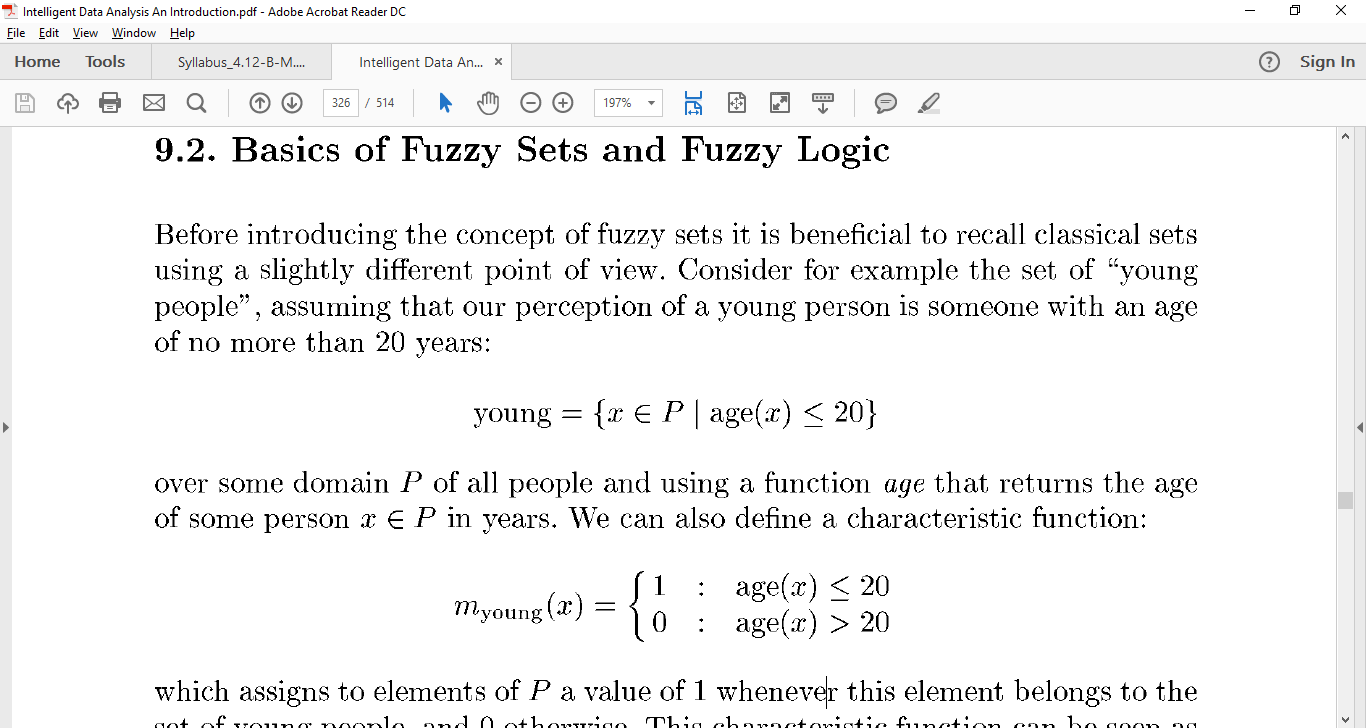
**5. Samza**

Finally, [Apache Samza](http://samza.apache.org/) is another distributed stream processing framework. Samza is built on [Apache Kafka](http://kafka.apache.org/) for messaging and YARN for cluster resource management. Its website provides the following overview of Samza:

* Simple API: Unlike most low-level messaging system APIs, Samza provides a very simple callback-based “process message” API comparable to MapReduce.
* Managed state: Samza manages snapshotting and restoration of a stream processor’s state. When the processor is restarted, Samza restores its state to a consistent snapshot. Samza is built to handle large amounts of state (many gigabytes per partition).
* Fault tolerance: Whenever a machine in the cluster fails, Samza works with YARN to transparently migrate your tasks to another machine.
* Durability: Samza uses Kafka to guarantee that messages are processed in the order they were written to a partition, and that no messages are ever lost.
* Scalability: Samza is partitioned and distributed at every level. Kafka provides ordered, partitioned, replayable, fault-tolerant streams. YARN provides a distributed environment for Samza containers to run in.
* Pluggable: Though Samza works out of the box with Kafka and YARN, Samza provides a pluggable API that lets you run Samza with other messaging systems and execution environments.
* Processor isolation: Samza works with Apache YARN, which supports Hadoop’s security model, and resource isolation through Linux CGroup

**Fuzzy logic**

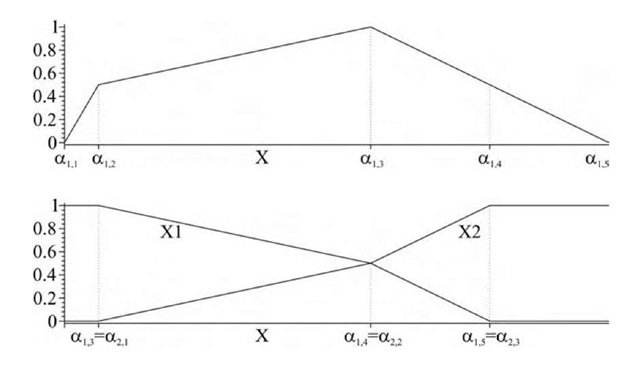
* **Fuzzy logic** is a form of [many-valued logic](https://en.wikipedia.org/wiki/Many-valued_logic) in which the [truth values](https://en.wikipedia.org/wiki/Truth_value) of variables may be any real number between 0 and 1.
* It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false.
* By contrast, in [Boolean logic](https://en.wikipedia.org/wiki/Boolean_algebra), the truth values of variables may only be the integer values 0 or 1.
* Fuzzy Logic is a logic system for reasoning that is approximate rather than exact.The fundamental unit of a fuzzy logic is the fuzzy set.
* Fuzzy logic has been applied to many fields, from [control theory](https://en.wikipedia.org/wiki/Control_theory) to [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence).
* [Classical logic](https://en.wikipedia.org/wiki/Classical_logic) only permits conclusions which are either true or false. However, there are also propositions with variable answers, such as one might find when asking a group of people to identify a color. In such instances, the truth appears as the result of reasoning from inexact or partial knowledge in which the sampled answers are mapped on a spectrum.
* Humans and animals often operate using fuzzy evaluations in many everyday situations. In the case where someone is tossing an object into a container from a distance, the person does not compute exact values for the object weight, density, distance, direction, container height and width, and air resistance to determine the force and angle to toss the object. Instead the person instinctively applies quick "fuzzy" estimates, based upon previous experience, to determine what output values of force, direction and vertical angle to use to make the toss.
* Both degrees of truth and [probabilities](https://en.wikipedia.org/wiki/Probability) range between 0 and 1 and hence may seem similar at first, but fuzzy logic uses degrees of truth as a [mathematical model](https://en.wikipedia.org/wiki/Mathematical_model) of *vagueness*, while [probability](https://en.wikipedia.org/wiki/Probability) is a mathematical model of *ignorance*.
* Take, for example, the concepts of "empty" and "full". The meaning of each of them can be represented by a certain [fuzzy set](https://en.wikipedia.org/wiki/Fuzzy_set). The concept of emptiness would be [subjective](https://en.wikipedia.org/wiki/Subjectivity) and thus would depend on the observer or designer. A 100 ml glass containing 30 ml of water may be defined as being 0.7 empty and 0.3 full, but another designer might, equally well, design a set [membership function](https://en.wikipedia.org/wiki/Membership_function_(mathematics)) where the glass would be considered full for all values down to 50 ml.



**Fuzzy model for data set**

* In fuzzy set theory **,** the grade of membership of a value x to a set S is defined through a membership function ji(x) that can take a value in the range [0, 1].
* The accompanying numerical attribute domain can be described by a finite series of MFs that each offers a grade of membership to describe x, which collectively form its concomitant fuzzy number.
* In this article, MFs are used to formulate linguistic variables for the considered attributes. These linguistic variables are made up of sets of linguistic terms which are defined by the MFs (see later).

**Figure 1. Example membership function and their use in a linguistic variable**

[](http://lh6.ggpht.com/_1wtadqGaaPs/TICj_YF0NiI/AAAAAAAAXyM/fA6_ZcrgPrM/s1600-h/tmp185182_thumb3.jpg)

**Pros and cons of FDT**

* Fuzzy Decision Trees Fuzzy decision trees (FDT) are particularly interesting for data mining and information retrieval because they enable the user to take into account imprecise descriptions of the cases, or heterogeneous values (symbolic, numerical, or fuzzy).
* Moreover, they are appreciated for their interpretability, because they provide a linguistic description of the relations between descriptions of the cases and decision to make or class to assign.
* The rules obtained through FDT make it easier for the user to interact with the system or the expert to understand, confirm or amend his own knowledge.
* Another quality of FDT is their robustness, since a small variation of descriptions does not drastically change the decision or the class associated with a case, which guarantees a resistance to measurement errors and avoids sharp differences for close values of the descriptions.

**Pros**

* **Fuzzy decision trees (FDTs)** benefit from the inductive learning approach that underpins their construction, to aid in the classification of objects based on their values over different attribute.
* Their construction in a fuzzy environment allows for the potential critical effects of imprecision to be mitigated, as well as brings a beneficial level of interpretability to the results found, through the decision rules defined.

**Cons**

* **As with the more traditional ‘crisp’ decision tree approaches,** there are issues such as the complexity of the results, in this case the tree defined.
* Future trends will surely include how FDTs can work on re-grading the complexity of the tree constructed, commonly known as pruning. Further, the applicability of the rules constructed, should see the use of FDTs extending in the range of applications it can work with.

**Regression Modelling**

* Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. For example, a regression model could be used to predict children's height, given their age, weight, and other factors.
* A regression task begins with a data set in which the target values are known. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data might track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.
* In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.
* The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide
* Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values

**Application**

* Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

**Time series**

* A time series is a series of [data points](https://en.wikipedia.org/wiki/Data_point) indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.
* Thus it is a sequence of [discrete-time](https://en.wikipedia.org/wiki/Discrete-time) data.
* Examples of time series are heights of ocean [tides](https://en.wikipedia.org/wiki/Tides), counts of [sunspots](https://en.wikipedia.org/wiki/Sunspots)
* Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.
* Time series *forecasting* is the use of a [model](https://en.wikipedia.org/wiki/Model_(abstract)) to predict future values based on previously observed values.
* While [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time.
* Time series analysis can be applied to [real-valued](https://en.wikipedia.org/wiki/Real_number), continuous data, [discrete](https://en.wiktionary.org/wiki/discrete) [numeric](https://en.wikipedia.org/wiki/Data_type#Numeric_types) data, or discrete symbolic data.

**Neural network**

* A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure during a learning phase.
* Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.
* In an artificial neural network, simple artificial nodes, called “neurons“, “neurodes”, “processing elements” or “units”, are connected together to form a network which mimics a biological neural network.
* It involves a network of simple processing elements that exhibit complex global behavior determined by the connections between the processing elements and element parameters.
* Artificial neural networks are used with algorithms designed to alter the strength of the connections in the network to produce a desired signal flow.
* Neural networks are also similar to biological neural networks in that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned.
* The term “neural network” usually refers to models employed in statistics, cognitive psychology and artificial intelligence.
* In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on statistics and signal processing. In some of these systems, neural networks or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive elements.
* While the more general approach of such adaptive systems is more suitable for real-world problem solving, it has far less to do with the traditional artificial intelligence connectionist models

# **Resampling**

* **Resampling** is the method that consists of drawing repeated [samples](http://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification/sampling/) from the original data samples. The method of Resampling is a nonparametric method of statistical inference.
* In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.
* Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample.
* Due to replacement, the drawn number of samples that are used by the method of resampling consists of repetitive cases.
* Resampling generates a unique sampling distribution on the basis of the actual data. The method of resampling uses experimental methods, rather than analytical methods, to generate the unique sampling distribution. The method of resampling yields unbiased estimates as it is based on the unbiased samples of all the possible results of the data studied by the researcher.

**Principal component analysis (PCA)**

* **Principal component analysis** (**PCA**) is a statistical procedure that uses an [orthogonal transformation](https://en.wikipedia.org/wiki/Orthogonal_transformation) to convert a set of observations of possibly correlated variables into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables called **principal components** (or sometimes, principal modes of variation).
* The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.
* In information processing such as pattern recognition, data compression and coding, image processing, high-resolution spectrum analysis, and adaptive beamforming, feature extraction or feature selection is necessary to deal with the large storage of raw data.
* Feature extraction is a dimensionality-reduction technique, mapping high-dimensional patterns onto a lower-dimensional space by extracting the most prominent features using orthogonal transforms. The extracted features do not have any physical meaning.
* In contrast, feature selection decreases the size of the feature set or reduces the dimension of the features by discarding the raw information according to a criterion.
* Orthogonal decomposition is a well-known technique to eliminate ill-conditioning. The Gram-Schmidt orthonormalization (GSO) is suitable for feature selection. This is due to the fact that the physically meaningless features in Gram-Schmidt space can be linked back to the same number of variables of the measurement space, thus resulting in no dimensionality reduction.
* The GSO procedure starts with QR decomposition of the transpose of the full feature matrix, 𝐗𝑇, where 𝐗=[𝐱1,𝐱2,…,𝐱𝑁]. QR decomposition can be performed by using the Householder transform or the Givens rotation [[1](https://www.hindawi.com/journals/isrn/2012/847305/#B50)], which is suitable for hardware implementation.
* The GSO transform can be used for feature subset selection; it inherits the compactness of the orthogonal representation and at the same time provides features that retain their original meaning.
* An orthogonal transform can decompose the correlations among the candidate features so that the significance of the individual features can be evaluated independently. Principal component analysis (PCA) is a well-known orthogonal transform that is used for dimensionality reduction.

**Stochastic search model**

* **Stochastic optimization** (**SO**) methods are [optimization](https://en.wikipedia.org/wiki/Optimization_(mathematics)) [methods](https://en.wikipedia.org/wiki/Iterative_method) that generate and use [random variables](https://en.wikipedia.org/wiki/Random_variable).
* For stochastic problems, the random variables appear in the formulation of the optimization problem itself, which involve random [objective functions](https://en.wikipedia.org/wiki/Objective_function) or random constraints.
* Stochastic optimization methods also include methods with random iterates. Some stochastic optimization methods use random iterates to solve stochastic problems, combining both meanings of stochastic optimization.
* Stochastic optimization methods generalize [deterministic](https://en.wikipedia.org/wiki/Deterministic_system_(mathematics)) methods for deterministic problems.
* Stochastic search and optimization techniques are used in a vast number of areas, including aerospace, medicine, transportation, and finance, to name but a few.
* Whether the goal is refining the design of a missile or aircraft, determining the effectiveness of a new drug, developing the most efficient timing strategies for traffic signals, or making investment decisions in order to increase profits, stochastic algorithms can help researchers and practitioners devise optimal solutions to countless real-world problems.

**2.Big data platform**

* Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
* It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure environment.

### **Advantages for Enterprises:**

Hadoop provides a cost effective storage solution for business.

* It facilitates businesses to easily access new data sources and tap into different types of data to produce value from that data.
* It is a highly scalable storage platform.
* Unique storage method of Hadoop is based on a distributed file system that basically ‘maps’ data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing.
* Hadoop is now widely used across industries, including finance, media and entertainment, government, healthcare, information services, retail, and other industries
* Hadoop is fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.
* Hadoop is more than just a faster, cheaper database and analytics tool. It is designed as a scale-out architecture that can affordably store all of a company’s data for later use.

**3.Report generation**

* A **report generator** is a computer program whose purpose is to take data from a source such as a [database](https://en.wikipedia.org/wiki/Database), [XML](https://en.wikipedia.org/wiki/XML) stream or a [spreadsheet](https://en.wikipedia.org/wiki/Spreadsheet), and use it to produce a document in a format which satisfies a particular human readership.
* Report generation functionality is almost always present in [database systems](https://en.wikipedia.org/wiki/Database_systems), where the source of the data is the database itself.

**Difference between reporting and analytics**

**Reporting** is “the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.”

Measuring core metrics and presenting them — whether in an email, a slidedeck, or online dashboard — falls under this category.

**Analytics** is “the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.”

Reporting provides you with information, analytics give you insights. Reporting raises questions, analytics attempts to answer them.

Both are valuable, but toward different purposes.

Analytics explains the “why?” and the “so what?”

How? Because it’s dynamic — on more than just a scale of time or interval. You should be able to contort the data to whatever you need. If what you see is a set of standard metrics, then it isn’t truly an analytic assessment (or an analytics product).

**Working of Report Generator**

* The Report Generator utility is used to create reports for applications and processes running in the system.
* The generated reports include dump files and event logs for the selected application or process. Usually, this information is required to analyze the application's behavior and you may need to send it to the application developers so that they can investigate the problem and find a solution faster.

This section describes how to generate reports using the Report Generator and includes the following items:

* https://support.smartbear.com/testcomplete/docs/_assets/commonImages/go.gif[Introducing Report Generator](https://support.smartbear.com/testcomplete/docs/testing-with/advanced/tracing-critical-errors/report-generator/about.html)

Describes the basic concepts of working with the Report Generator.

* https://support.smartbear.com/testcomplete/docs/_assets/commonImages/go.gif[Working With the Report Generator](https://support.smartbear.com/testcomplete/docs/testing-with/advanced/tracing-critical-errors/report-generator/working-with.html)

Contains a step-by-step instruction on how to create a report for the desired application or process using the Report Generator.

* https://support.smartbear.com/testcomplete/docs/_assets/commonImages/go.gif[Report Generator Pages](https://support.smartbear.com/testcomplete/docs/testing-with/advanced/tracing-critical-errors/report-generator/pages/index.html)

This section describes the pages of the Report Generator wizard in detail.

* https://support.smartbear.com/testcomplete/docs/_assets/commonImages/go.gif[Report Generator - System Requirements](https://support.smartbear.com/testcomplete/docs/testing-with/advanced/tracing-critical-errors/report-generator/system-requirements.html)

Describes the Report Generator’s system requirements.

* https://support.smartbear.com/testcomplete/docs/_assets/commonImages/go.gif[Report Generator Command Line](https://support.smartbear.com/testcomplete/docs/testing-with/advanced/tracing-critical-errors/report-generator/command-line.html)

Describes how you can work with the Report Generator from the command line.

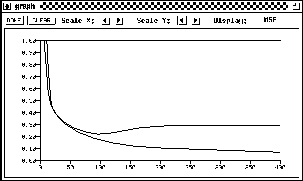
**Prediction and statistical interference**

* In [statistics](https://en.wikipedia.org/wiki/Statistics), prediction is a part of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference).
* One particular approach to such inference is known as [predictive inference](https://en.wikipedia.org/wiki/Predictive_inference), but the prediction can be undertaken within any of the several approaches to statistical inference.
* Indeed, one possible description of statistics is that it provides a means of transferring knowledge about a sample of a population to the whole population, and to other related populations, which is not necessarily the same as prediction over time.
* Statistical techniques used for prediction include [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) and its various sub-categories such as [linear regression](https://en.wikipedia.org/wiki/Linear_regression), [generalized linear models](https://en.wikipedia.org/wiki/Generalized_linear_model) ([logistic regression](https://en.wikipedia.org/wiki/Logistic_regression), [Poisson regression](https://en.wikipedia.org/wiki/Poisson_regression), [Probit regression](https://en.wikipedia.org/wiki/Probit_regression" \o "Probit regression)), etc. When these and/or related, generalized set of regression or [machine learning](https://en.wikipedia.org/wiki/Machine_learning) methods are deployed in commercial usage, the field is known as [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics).
* In many applications, such as time series analysis, it is possible to estimate the models that generate the observations.
* To use regression analysis for prediction, data are collected on the variable that is to be predicted, called the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) or response variable, and on one or more variables whose values are [hypothesized](https://en.wikipedia.org/wiki/Hypothesis) to influence it, called [independent variables](https://en.wikipedia.org/wiki/Independent_variable) or explanatory variables.

The principal reason why neural networks have attracted such interest, is the existence of learning algorithms for neural networks: algorithms that use data to estimate the optimal weights in a network to perform some task. There are three basic approaches to learning in neural networks

* Supervised learning uses a training set that consists of a set of pattern pairs: an input pattern and the corresponding desired (or target) output pattern. The desired output may be regarded as the ‘network’s ‘teacher” for that input. The basic approach in supervised learning is for the network to compute the output its current weights produce for a given input, and to compare this network output with the desired output. The aim of the learning algorithm is to adjust the weights so as minimize the difference between the network output and the desired output.
* Reinforcement learning uses much less supervision. If a network aims to perform that some task, then the reinforcement signal is a simple “yes” or “no” at the end of the task to indicate whether the task has been performed satisfactorily.
* Unsupervised learning only uses input data there is no training signal, unlike the previous two approaches. The aim of unsupervised learning is to make sense of some data set, for example clustering similar patterns together.

# **Generalization of Neural Networks**

     
**Figure:** Error development of a training and a validation set

One of the major advantages of neural nets is their ability to generalize. This means that a trained net could classify data from the same class as the learning data that it has never seen before. In real world applications developers normally have only a small part of all possible patterns for the generation of a neural net. To reach the best generalization, the dataset should be split into three parts:

* The **training set** is used to train a neural net. The error of this dataset is minimized during training.
* The **validation set** is used to determine the performance of a neural network on patterns that are not trained during learning.
* A **test set** for finally checking the over all performance of a neural net.