UNIT I INTRODUCTION

Introduction to Big Data – Issues and Challenges in the traditional systems - Evolution of Big Data – Four V's of Big Data – Big Data Use Cases and characteristics – Intelligent Data Analysis – Data Analytic Tools – Big Data Storage Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error – Random Sampling.

What is Data?
In computing, data is **information that has been translated into a form that is efficient for movement or processing**.
When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information.

What is Information?
Information is defined as classified or organized data that has some meaningful value for the user. Information is also the processed data used to make decisions and take action. Processed data must meet the following criteria for it to be of any significant use in decision-making:

- Accuracy: The information must be accurate.
- Completeness: The information must be complete.
- Timeliness: The information must be available when it's needed.

What is Big Data?
Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

What are the characteristics of big data?
Big data can be described by the following characteristics

- Volume - The quantity of data generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.
- Variety -The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.
- Velocity -In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- Variability- Inconsistency of the data set can hamper processes to handle and manage it.
- Veracity-The data quality of captured data can vary greatly, affecting the accurate analysis

Big Data has been described by some Data Management pundits (with a bit of a snicker) as "huge, overwhelming, and uncontrollable amounts of information." In 1663, John Graunt dealt with "overwhelming amounts of information" as well, while he studied the bubonic plague, which was currently ravaging Europe. Graunt used statistics and is credited with being the first person to use statistical data analysis. In the early 1800s, the field of statistics expanded to include collecting and analyzing data.

The evolution of Big Data includes a number of preliminary steps for its foundation, and while looking back to 1663 isn't necessary for the growth of data volumes today, the point remains that "Big Data" is a relative

term depending on who is discussing it. Big Data to Amazon or Google is very different than Big Data to a medium-sized insurance organization, but no less "Big" in the minds of those contending with it.

Such foundational steps to the modern conception of Big Data involve the development of computers, smart phones, the internet, and sensory (Internet of Things) equipment to provide data. Credit cards also played a role, by providing increasingly large amounts of data, and certainly social media changed the nature of data volumes in novel and still developing ways. The evolution of modern technology is interwoven with the evolution of Big Data.

The Foundations of Big Data

Data became a problem for the U.S. Census Bureau in 1880. They estimated it would take eight years to handle and process the data collected during the 1880 census, and predicted the data from the 1890 census would take more than 10 years to process. Fortunately, in 1881, a young man working for the bureau, named Herman Hollerith, created the Hollerith Tabulating Machine. His invention was based on the punch cards designed for controlling the patterns woven by mechanical looms. His tabulating machine reduced ten years of labor into three months of labor.

In 1927, Fritz Pfleumer, an Austrian-German engineer, developed a means of storing information magnetically on tape. Pfleumer had devised a method for adhering metal stripes to cigarette papers (to keep a smokers' lips from being stained by the rolling papers available at the time), and decided he could use this technique to create a magnetic strip, which could then be used to replace wire recording technology. After experiments with a variety of materials, he settled on a very thin paper, striped with iron oxide powder and coated with lacquer, for his patent in 1928.

During World War II (more specifically 1943), the British, desperate to crack Nazi codes, invented a machine that scanned for patterns in messages intercepted from the Germans. The machine was called Colossus, and scanned 5.000 characters a second, reducing the workload from weeks to merely hours. Colossus was the first data processor. Two years later, in 1945, John Von Neumann published a paper on the Electronic Discrete Variable Automatic Computer (EDVAC), the first "documented" discussion on program storage, and laid the foundation of computer architecture today.

It is said these combined events prompted the "formal" creation of the United States' NSA (National Security Agency), by President Truman, in 1952. Staff at the NSA were assigned the task of decrypting messages intercepted during the Cold War. Computers of this time had evolved to the point where they could collect and process data, operating independently and automatically.

The Internet Effect and Personal Computers

ARPANET began on Oct 29, 1969, when a message was sent from UCLA's host computer to Stanford's host computer. It received funding from the Advanced Research Projects Agency (ARPA), a subdivision of the Department of Defense. Generally speaking, the public was not aware of ARPANET. In 1973, it connected with a transatlantic satellite, linking it to the Norwegian Seismic Array. However, by 1989, the infrastructure of ARPANET had started to age. The system wasn't as efficient or as fast as newer networks. Organizations using ARPANET started moving to other networks, such as NSFNET, to improve basic efficiency and speed. In 1990, the ARPANET project was shut down, due to a combination of age and obsolescence. The creation ARPANET led directly to the Internet.

In 1965, the U.S. government built the first data center, with the intention of storing millions of fingerprint sets and tax returns. Each record was transferred to magnetic tapes, and were to be taken and stored in a central location. Conspiracy theorists expressed their fears, and the project was closed. However, in spite of its closure, this initiative is generally considered the first effort at large scale data storage.

Personal computers came on the market in 1977, when microcomputers were introduced, and became a major stepping stone in the evolution of the internet, and subsequently, Big Data. A personal computer could be used by a single individual, as opposed to mainframe computers, which required an operating staff, or some kind of time-sharing system, with one large processor being shared by multiple individuals. After the introduction of the microprocessor, prices for personal computers lowered significantly, and became described as "an affordable consumer good." Many of the early personal computers were sold as electronic kits, designed to be built by hobbyists and technicians. Eventually, personal computers would provide people worldwide with access to the internet.

In 1989, a British Computer Scientist named Tim Berners-Lee came up with the concept of the World Wide Web. The Web is a place/information-space where web resources are recognized using URLs, interlinked by hypertext links, and is accessible via the Internet. His system also allowed for the transfer of audio, video, and pictures. His goal was to share information on the Internet using a hypertext system. By the fall of 1990, Tim Berners-Lee, working for CERN, had written three basic IT commands that are the foundation of today's web:

- **HTML:** HyperText Markup Language. The formatting language of the web.
- **URL:** Uniform Resource Locator. A unique "address" used to identify each resource on the web. It is also called a URI (Uniform Resource Identifier).
- **HTTP:** Hypertext Transfer Protocol. Used for retrieving linked resources from all across the web.

In 1993, CERN announced the World Wide Web would be free for everyone to develop and use. The free part was a key factor in the effect the Web would have on the people of the world. (It's the companies providing the "internet connection" that charge us a fee).

The Internet of Things (IoT)

The concept of Internet of Things was assigned its official name in 1999. By 2013, the IoT had evolved to include multiple technologies, using the Internet, wireless communications, micro-electromechanical systems (MEMS), and embedded systems. All of these transmit data about the person using them. Automation (including buildings and homes), GPS, and others, support the IoT.

The Internet of Things, unfortunately, can make computer systems vulnerable to hacking. In October of 2016, hackers crippled major portions of the Internet using the IoT. The early response has been to develop Machine Learning and Artificial Intelligence focused on security issues.

Computing Power and Internet Growth

There was an incredible amount of internet growth in the 1990s, and personal computers became steadily more powerful and more flexible. Internet growth was based both on Tim Berners-Lee's efforts, Cern's free access, and access to individual personal computers.

In 2005, Big Data, which had been used without a name, was labeled by Roger Mougalas. He was referring to a large set of data that, at the time, was almost impossible to manage and process using the traditional business intelligence tools available. Additionally, Hadoop, which could handle Big Data, was created in 2005. Hadoop was based on an open-sourced software framework called Nutch, and was merged with Google's MapReduce. Hadoop is an Open Source software framework, and can process structured and unstructured data, from almost all digital sources. Because of this flexibility, Hadoop (and its sibling frameworks) can process Big Data.

Big Data Storage

Magnetic storage is currently one of the least expensive methods for storing data. Fritz Pfleumer's 1927 concept of striped magnetic lines has been adapted to a variety of formats, ranging from magnetic tape, magnetic drums, floppies, and hard disk drives. Magnetic storage describes any data storage based on a magnetized medium. It uses the two magnetic polarities, North and South, to represent a zero or one, or on/off.

Cloud Data Storage has become quite popular in recent years. The first true Cloud appeared in 1983, when CompuServe offered its customers 128K of data space for personal and private storage. In 1999, Salesforce offered Software-as-a-service (SaaS) from their website. Technical improvements within the internet, combined with falling data storage costs, have made it more economical for businesses and individuals to use the Cloud for data storage purposes. This saves organizations the cost of buying, maintaining, and eventually replacing their computer system. The Cloud provides a near-infinite amount of scalability, and is accessible anywhere, anytime, and offers a variety of services.

The Uses of Big Data

Big Data is revolutionizing entire industries and changing human culture and behavior. It is a result of the information age and is changing how people exercise, create music, and work. The following provides some examples of Big Data use.

- Big Data is being used in healthcare to map disease outbreaks and test alternative treatments.
- NASA uses Big Data to explore the universe.
- The music industry replaces intuition with Big Data studies.
- Utilities use Big Data to study customer behavior and avoid blackouts.
- Nikeuses health monitoring wearables to track customers and provide feedback on their health.
- Big Data is being used by cybersecurity to stop cybercrime.

Types of big data

Data comes in different forms. The fact be said, here are the three main categories it falls into.

- **Structured data** Data that can be stored, accessed and processed in the form of fixed-format is termed as 'structured data.' Since this data comes in a similar format, businesses get the maximum out of it by performing analysis. Various advanced technologies are also invented to extract data-driven decisions from structured data. However, the world is going towards an extent where the creation of structured data is ballooning too much as it has already reached the zettabytes mark.
- **Unstructured data** Any data that comes in an unknown form or structure falls under unstructured data. Processing unstructured data and analyzing them to get data-driven answers

is a challenging task as they are from different categories and outing them together will only make things worse. A heterogeneous data source containing a combination of simple text files, images, videos, etc. is an example of unstructured data.

- **Semi-structured data** Semi-structured data has both structured and unstructured data in it. We can see semi-structured data as structured in form, but it is actually not defined with table definition in relational DBMS. Web application data is an example of semi-structured data. It has unstructured data like log files, transaction history files, etc. OLTP systems are built to work with structured data wherein data is stored in relations.

Applications of big data

Business organisations are leveraging data to reach their maximum potential. Ever since technology took over big data analysis, business decisions are mostly based on predictive outcomes. Besides, big data is also contributing to personalized customer experiences at high-ends. Some of the important business applications of big data are listed below.

• Product development- Companies avail big data to anticipate customer demands. They build predictive models to see customer preference and provide relevant materials.

• Log analytics- Commercial and open-source log analytics provides the ability to collect, process and analyze massive log data without having to dump the data into relational databases and retrieving it through SQL queries.

• Security compliance- Big data helps you identify patterns in data that indicate fraud and aggregate large volumes of information to make regulatory reporting much faster.

• Recommendation engines- Big data, with its scalability and power to processes massive amounts of both unstructured and structured data enables companies to recommend the best option for customers based on their history.

Why is Big Data Important?

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

1. Cost Savings: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business. 2
2. . Time Reductions: The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.
3. 3. Understand the market conditions: By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.
4. 4. Control online reputation: Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.
5. 5. Using Big Data Analytics to Boost Customer Acquisition and Retention The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end,

loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.

6. 6. Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights BIG DATA ANALYTICS 4 Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.
7. 7. Big Data Analytics As a Driver of Innovations and Product Development Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Big Data Analytics Examples is used to generate various reports among those some examples are given below:

1. Fraud Management Report which is generally used in Banking Sectors to find the fraud transactions, hacking, unauthorized access to the account and so on.
2. Live Tracking Report which is generally used by Transport Sectors such as Meru, Ola, Uber, and Mega to track the vehicles, customer's requests, payment management, emergency alert and to find the daily needs and revenues and so on.
3. Sales Report and Future target and goal analysis which is mostly used by all sectors to analyze their sales, revenues, and needs of customers and also used to determine the future target and so on.
4. Many reports based on live data mostly used to manage live data in many entertainment sites, share market, real-time Sensex data etc.
5. Generate different types of alarms based on different activities like alarm generated by data Centre, various notifications Big Data Analytics Examples has been used here.
6. Google Analytics report where we can get how many user's visit counts, from which location the user is from, from which device the site is accessing and so on.
7. Many Health care Organization nowadays rapidly introduced Big Data predictive analytics to improve our daily life. It has been used to update many protocols of Healthcare Sectors and also used to improve the outcomes against entire populations.
8. Big Data Analytics Examples also played a vital role in many disaster situations. In the year April 2015 earthquake killed and also injured many peoples in Nepal. In this situation, North Carolina-based SAS has been developed by Analytics which has been played a massive role in rescue and relief operation.
9. Big Data Analytics examples has been used in Child Welfare also. In a neighborhood in London, an English Physician has been collected and used the huge data to originate the solutions against massive Cholera attack in 19^{th}
10. Big Data Analytics has been used in Online and Physical Security to identify the unauthorized activities, take various steps to prevent those attacks, introduced real-time monitoring to reduce fraud activities and also activating alarms against suspicious
    Actions.

What is Big Data Storage?

Big Data Storage is a new technology poised to revolutionize how we store data. The technology was first developed in the early 2000s when companies were faced with storing massive amounts of data that they could not keep on their servers.

The problem was that traditional storage methods couldn't handle storing all this data, so companies had to look for new ways to keep it. That's when Big Data Storage came into being. It's a way for companies to store large amounts of data without worrying about running out of space.

Big Data Storage Challenges

Big data is a hot topic in IT. Every month, more companies are adopting it to help them improve their businesses. But with any new technology comes challenges and questions, and big data is no exception.

The first challenge is how much storage you'll need for your extensive data system. If you're going to store large amounts of information about your customers and their behavior, you'll need a lot of space for that data to live.

It's not uncommon for large companies like Google or Facebook to have petabytes (1 million gigabytes) of storage explicitly dedicated to their big data needs, and that's only one company!

Another challenge with big data is how quickly it grows. Companies are constantly gathering new types of information about their customer's habits and preferences, and they're looking at ways they can use this information to improve their products or services.

As a result, big data systems will continue growing exponentially until something stops them. It means it's essential for companies who want to use this technology effectively to plan how they'll deal with it later on down the road when it becomes too much for them alone.

The challenges include capture, duration, storage, search, sharing, transfer, – analysis, and visualization.

• Big Data is trend to larger data sets

• due to the additional information derivable from analysis of a single large set of related data, – as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to

• "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Challenges of Big Data

The following are the five most important challenges of the Big Data

a) Meeting the need for speed In today's hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly.

b) Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

c) The challenge only grows as the degree of granularity increases. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly

d) Understanding the data

- It takes a lot of understanding to get data in the RIGHT SHAPE so that you can use
- visualization as part of data analysis.

e) Addressing data quality

- Even if you can find and analyze data quickly and put it in the proper context for the
- audience that will be consuming the information, the value of data for DECISIONMAKING PURPOSES will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis.

f) Displaying meaningful results

- Plotting points on a graph for analysis becomes difficult when dealing with extremely

- large amounts of information or a variety of categories of information.
- For example, imagine you have 10 billion rows of retail SKU data that you're trying to
- compare. The user trying to view 10 billion plots on the screen will have a hard time • seeing so many data points.
- . By grouping the data together, or "binning," you can more effectively visualize the data. f) Dealing with outliers
- The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text.
- Users can easily spot issues that need attention simply by glancing at a chart. Outliers typically represent about 1 to 5 percent of data, but when you're working with massive amounts of data, viewing 1 to 5 percent of the data is rather difficult
- We can also bin the results to both view the distribution of data and see the outliers.
- While outliers may not be representative of the data, they may also reveal previously unseen and potentially valuable insights.
- Visual analytics enables organizations to take raw data and present it in a meaningful way that generates the most value. However, when used with big data, visualization is bound to lead to some challenges.

Big Data Storage Key Considerations

Big data storage is a complicated problem. There are many things to consider when building the infrastructure for your big data project, but there are three key considerations you must consider before you move forward.

- Data velocity: Your data must be able to move quickly between processing centers and databases for it to be helpful in real-time applications.
- Scalability: The system should be able to expand as your business does and accommodate new projects as needed without disrupting existing workflows or causing any downtime.
- Cost efficiency: Because big data projects can be so expensive, choosing a system that reduces costs without sacrificing the quality of service or functionality is essential.

Finally, consider how long you want your stored data to remain accessible. If you're planning on keeping it for years (or even decades), you may need more than one storage solution.

Data Storage Technologies

Apache Hadoop, Apache HBase, and Snowflake are three big data storage technologies often used in the data lake analytics paradigm.

Hadoop

Hadoop has gained considerable attention as it is one of the most common frameworks to support big data analytics. A distributed processing framework based on open-source software, Hadoop enables large data sets to be processed across clusters of computers. Large data sets were initially intended to be processed and stored across clusters of commodity hardware.

What is HDFS

Hadoop comes with a distributed file system called HDFS. In HDFS data is distributed over several machines and replicated to ensure their durability to failure and high availability to parallel application.

It is cost effective as it uses commodity hardware. It involves the concept of blocks, data nodes and node name.

Where to use HDFS

- **Very Large Files:** Files should be of hundreds of megabytes, gigabytes or more.
- **Streaming Data Access:** The time to read whole data set is more important than latency in reading the first. HDFS is built on write-once and read-many-times pattern.
- **Commodity Hardware:**It works on low cost hardware.

Where not to use HDFS

- **Low Latency data access:** Applications that require very less time to access the first data should not use HDFS as it is giving importance to whole data rather than time to fetch the first record.
- **Lots Of Small Files:**The name node contains the metadata of files in memory and if the files are small in size it takes a lot of memory for name node's memory which is not feasible.
- **Multiple Writes:**It should not be used when we have to write multiple times.

HBase

With HBase, you can use a NoSQL database or complement Hadoop with a column-oriented store. This database is designed to efficiently manage large tables with billions of rows and millions of columns. The performance can be tuned by adjusting memory usage, the number of servers, block size, and other settings.

What is HBase

Hbase is an open source and sorted map data built on Hadoop. It is column oriented and horizontally scalable.

It is based on Google's Big Table.It has set of tables which keep data in key value format. Hbase is well suited for sparse data sets which are very common in big data use cases. Hbase provides APIs enabling development in practically any programming language. It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System.

Why HBase

- RDBMS get exponentially slow as the data becomes large
- Expects data to be highly structured, i.e. ability to fit in a well-defined schema
- Any change in schema might require a downtime
- For sparse datasets, too much of overhead of maintaining NULL values

Features of Hbase

- Horizontally scalable: You can add any number of columns anytime.
- Automatic Failover: Automatic failover is a resource that allows a system administrator to automatically switch data handling to a standby system in the event of system compromise
- Integrations with Map/Reduce framework: Al the commands and java codes internally implement Map/ Reduce to do the task and it is built over Hadoop Distributed File System.
- sparse, distributed, persistent, multidimensional sorted map, which is indexed by rowkey, column key,and timestamp.

- Often referred as a key value store or column family-oriented database, or storing versioned maps of maps.
- fundamentally, it's a platform for storing and retrieving data with random access.
- It doesn't care about datatypes(storing an integer in one row and a string in another for the same column).
- It doesn't enforce relationships within your data.
- It is designed to run on a cluster of computers, built using commodity hardware

What is HIVE

Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook.

Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL like queries called HQL (Hive query language) which gets internally converted to MapReduce jobs.

Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

Features of Hive

These are the following features of Hive:

- Hive is fast and scalable.
- It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
- It is capable of analyzing large datasets stored in HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It can operate on compressed data stored in the Hadoop ecosystem.
- It supports user-defined functions (UDFs) where user can provide its functionality.

Limitations of Hive

- Hive is not capable of handling real-time data.
- It is not designed for online transaction processing.
- Hive queries contain high latency.

Snowflake

Snowflake for Data Lake Analytics is an enterprise-grade cloud platform for advanced analytics applications built on top of Apache Hadoop. It offers real-time access to historical and streaming data from any source and format at any scale without requiring changes to existing applications or workflows. It also enables users to

quickly scale up their processing power as needed without having to worry about infrastructure management tasks such as provisioning

Implementing Big Data Techniques:

1. Association rule learning
2. Classification tree analysis
3. Genetic algorithms
4. Machine learning
5. Regression analysis
6. Sentiment analysis
7. Social network analysis

1. Association rule learning

Are people who purchase tea more or less likely to purchase carbonated drinks?

Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems.

Association rule learning is being used to help:

- place products in better proximity to each other in order to increase sales
- extract information about visitors to websites from web server logs
- analyze biological data to uncover new relationships
- monitor system logs to detect intruders and malicious activity
- identify if people who buy milk and butter are more likely to buy diapers

2. Classification tree analysis

Which categories does this document belong to?

Statistical classification is a method of identifying categories that a new observation belongs to. It requires a training set of correctly identified observations – historical data in other words.

Statistical classification is being used to:

- automatically assign documents to categories
- categorize organisms into groupings
- develop profiles of students who take online courses

3. Genetic algorithms

Which TV programs should we broadcast, and in what time slot, to maximize our ratings?

Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to "evolve" useful solutions to problems that require optimization.

Genetic algorithms are being used to:

- schedule doctors for hospital emergency rooms
- return combinations of the optimal materials and engineering practices required to develop fuel-efficient cars
- generate "artificially creative" content such as puns and jokes
-

4. Machine Learning

Which movies from our catalogue would this customer most likely want to watch next, based on their viewing history?

Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of "training data."

Machine learning is being used to help:

- distinguish between spam and non-spam email messages
- learn user preferences and make recommendations based on this information
- determine the best content for engaging prospective customers
- determine the probability of winning a case, and setting legal billing rates
-

5. Regression Analysis

How does your age affect the kind of car you buy?

At a basic level, regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

Regression analysis is being used to determine how:

- levels of customer satisfaction affect customer loyalty
- the number of supports calls received may be influenced by the weather forecast given the previous day
- neighbourhood and size affect the listing price of houses
- to find the love of your life via online dating sites
-

6. Sentiment Analysis

How well is our new return policy being received?

Sentiment analysis helps researchers determine the sentiments of speakers or writers with respect to a topic.

Sentiment analysis is being used to help:

- improve service at a hotel chain by analyzing guest comments
- customize incentives and services to address what customers are really asking for
- determine what consumers really think based on opinions from social media
-

7. Social Network Analysis

How many degrees of separation are you from Kevin Bacon?

Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.

Social network analysis is being used to:

- see how people from different populations form ties with outsiders
- find the importance or influence of a particular individual within a group
- find the minimum number of direct ties required to connect two individuals
- understand the social structure of a customer base
-

Whether your business wants to discover interesting correlations, categorize people into groups, optimally schedule resources, or set billing rates, a basic understanding of the seven techniques mentioned above can help Big Data work for you.

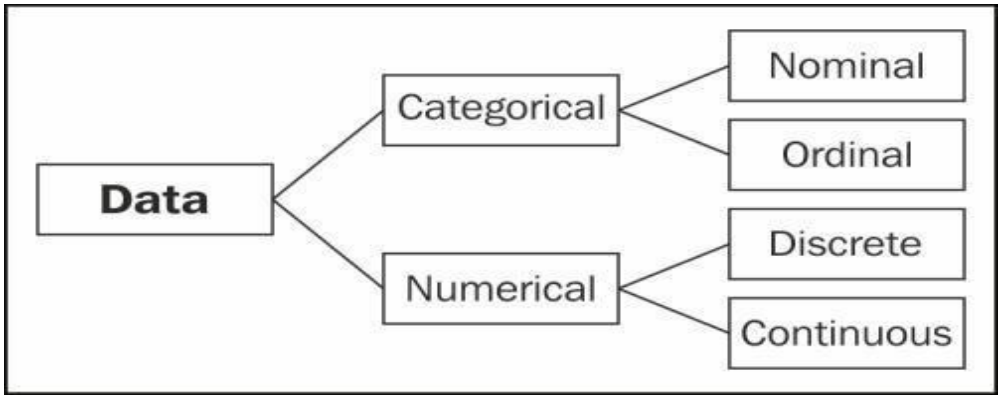
TYPES OF DATA

- In order to understand the nature of data it is necessary to categorize them into varioustypes.
- Different categorizations of data are possible.
- The first such categorization may be on the basis of disciplines, e.g., Sciences, Social Sciences, etc. in which they are generated.
- Within each of these fields, there may be several ways in which data can be categorized intotypes.

There are four types of data:

- Nominal
- Ordinal
- Interval
- Ratio

Each offers a unique set of characteristics, which impacts the type of analysis that can be performed.



The distinction between the four types of scales center on three different characteristics:

1. The **order** of responses – whether it matters or not
2. The **distance between observations** – whether it matters or is interpretable
3. The presence or inclusion of a **true zero**

Nominal Scales

Nominal scales measure categories and have the following characteristics:

- **Order:** The order of the responses or observations does not matter.
- **Distance:** Nominal scales do not hold distance. The distance between a 1 and a 2 is not the same as a 2 and 3.
- **True Zero:** There is no true or real zero. In a nominal scale, zero is uninterruptable.

Appropriate statistics for nominal scales: mode, count, frequencies

Displays: histograms or bar charts

Ordinal Scales

At the risk of providing a tautological definition, ordinal scales measure, well, order. So, our characteristics for ordinal scales are:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ordinal scales do not hold distance. The distance between first and second is unknown as is the distance between first and third along with all observations.
- **True Zero:** There is no true or real zero. An item, observation, or category cannot finish zero.

Appropriate statistics for ordinal scales: count, frequencies, mode

Displays: histograms or bar charts

Interval Scales

Interval scales provide insight into the variability of the observations or data.

Classic interval scales are Likert scales (e.g., 1 - strongly agree and 9 - strongly disagree) and Semantic Differential scales (e.g., 1 - dark and 9 - light).

In an interval scale, users could respond to "I enjoy opening links to th website from a company email" with a response ranging on a scale of values.

The characteristics of interval scales are:

- **Order:** The order of the responses or observations does matter.
- **Distance:** Interval scales do offer distance. That is, the distance from 1 to 2 appears the same as 4 to 5. Also, six is twice as much as three and two is half of four. Hence, we can perform arithmetic operations on the data.
- **True Zero:** There is no zero with interval scales. However, data can be rescaled in a

manner that contains zero. An interval scales measure from 1 to 9 remains the same as 11 to 19 because we added 10 to all values. Similarly, a 1 to 9 interval scale is the same a -4 to 4 scale because we subtracted 5 from all values. Although the new scale contains zero, zero remains uninterruptable because it only appears in the scale from the transformation.

Appropriate statistics for interval scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

Ratio Scales

Ratio scales appear as nominal scales with a true zero.
They have the following characteristics:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ratio scales do do have an interpretable distance.
- **True Zero:** There is a true zero.

Income is a classic example of a ratio scale:

- Order is established. We would all prefer $100 to $1!
- Zero dollars means we have no income (or, in accounting terms, our revenue exactly equals our expenses!)
- Distance is interpretable, in that $20 appears as twice $10 and $50 is half of a $100. For the web analyst, the statistics for ratio scales are the same as for interval scales.

Appropriate statistics for ratio scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

The table below summarizes the characteristics of all four types of scales.

	Nominal	Ordinal	Interval	Ratio
Order Matters	No	Yes	Yes	Yes
Distance Is Interpretable	No	No	Yes	Yes

Zero Exists	No	No	No	Yes

Most Popular Big Data Analytics Tools

Open-source big data analytics tools are intended to be publicly accessible and are typically managed and maintained by organizations with a specific mission. Let’s check some important big data processing tools.

Let's check some big data analytics tools examples and software used in big data analytics. Listed below are the top and the most popular tools for big data analytics:

1. APACHE Hadoop

Big data is being processed and stored using this Java-based open-source platform, and data can be processed efficiently and in parallel thanks to the cluster system. Data from one server can be processed by multiple structured and unstructured computers, and users of Hadoop can also access it across multiple platforms. Amazon, Microsoft, IBM, and other tech giants use it today as one of the best tools for big data analysis.

Features:

- Businesses can use this storage solution for free, and it is an efficient one.
- It can be installed on a bunch of disks or a JBOD of commodity hardware.
- The Hadoop Distributed File System (HDFS) provides quick access.
- Easily scales up to a large amount of data when it is distributed in small chunks.
- Easy to implement with MySQL, JSON, and highly flexible.

2. Cassandra

Data sets can be retrieved in large quantities using APACHE Cassandra, a distributed database with no SQL engine. A number of tech companies have praised it for its high availability and scalability without sacrificing speed or performance without compromising speed. It can handle petabytes of resources with almost no downtime and delivers thousands of operations every second. A public version of this best big data tool was created in 2008 by Facebook.

Features:

- With Cassandra, you can store data quickly and process it efficiently on efficient commodity hardware.
- Data can be structured, semi-structured, or unstructured, and users can change the data according to their requirements.
- Distributing data across multiple data centers is easy with replication.
- As soon as a node fails, it will be replaced immediately.

3. Qubole

Using ad-hoc analysis in machine learning, it fetches data from a value chain using open-source technology for big data analytics. Qubole provides end-to-end services in moving data pipelines with reduced time and effort. Configure Azure, AWS, and Google Cloud services simultaneously. As a result, cloud computing costs are also reduced by 50%.

Features:

- In order to target more acquisitions, Qubole offers predictive analysis.
- Multi-source data can be migrated to one location through this tool.
- Users can view real-time insights into their systems when it monitors their systems.

4. Xplenty

Using minimal code allows you to build a data pipeline. Its sales, marketing, and support solutions cover a broad spectrum of needs. In addition to providing ETL and ELT solutions, it features an interactive graphical interface. With Xplenty, you'll spend less on hardware and software and receive support via chat, email, telephonic and virtual meetings. Data can be processed for the application of big data analysis over the cloud and segregated using Xplenty.

Features:

- Integrated apps can be deployed on-premises and in the cloud.
- Regular verification of algorithms and certificates can be performed on the platform, along with SSL/TSL encryption.
- Databases, warehouses, and salesforce can receive and process data.

5. Spark

Data processing and numerous tasks can also be done on a large scale using Apache Spark. With the help of tools for big data, data can also be processed via multiple computers. Due to its easy-to-use APIs and its ability to handle multi-petabytes of data, it is widely used among data analysts. Spark is highly suitable for ML and AI today, which is why big tech giants are moving towards it now.

Features:

- Users can choose the language they wish to run in.
- Streaming can be handled by Spark using Spark Streaming.

6. MongoDB

This free, open-source platform, which came into the limelight in 2010, is a document-oriented (NoSQL) database that is used to store a large amount of information in a structured manner. As a result of MongoDB's support for multiple programming languages, such as Jscript, Python, and Ruby, it is extremely popular among developers.

Features:

- The backup function can be called back after writing or reading data from the master.
- Documents can be stored in a schema-less database.
- The Mongo database makes storing files easy without disrupting the stack.

7. Apache Storm

Small companies, especially those that do not have the resources for big data analytics, use storms, which are robust and user-friendly tools. Storm does not have any language barriers (programming) and can support any of them. With fault tolerance and horizontal scalability, it was designed to handle a pool of large data. Since Storm has a distributed real-time big data processing system, it leads the pack when it comes to real-time data processing. APACHE Storm is used in many tech giants' systems today. NaviSite, Twitter, and Zendesk are among the most well-known.

Features:

- With APACHE Storm, a single node is capable of processing up to 1 million messages per second.
- Even if a node is disconnected, Storm still processes the data.

8. SAS

It is one of the best tools used by data analysts today to create statistical models, and data scientists can manage data from several sources and mine, extract, or update it using it. Data can be accessed in SAS tables or Excel worksheets using Statistics Analytical System (SAS). In addition to that, SAS has also introduced new tools used in big data and products to get a stronger grip on artificial intelligence & machine learning.

Features:

- Data can be read from any format and is compatible with many programming languages, including SQL.
- Non-programmers will appreciate its easy-to-learn syntax as well as its vast libraries.

9. Data Pine

Since 2012, Datapine has been providing analytics for business intelligence (Berlin, Germany). Since its introduction, it has gained considerable popularity in a number of countries, especially in small- and medium-sized companies that need to extract data for monitoring purposes. Thanks to the enhanced user interface, anyone can access the data according to their requirements and choose between four price brackets, starting at $249 per month. Dashboards are available by function, industry, and platform.

Features:

Through the use of historical and current data, datapine provides forecasting and predictive analytics.

- Our BI tools and AI assistants are designed to cut down the manual chase.

10. Rapid Miner

Its purpose is to automate the design of data analytics workflows using visual tools. With this platform, users do not need to code in order to segregate data. Education technology, training, research, etc., are among the industries heavily using it today. Although it's open-source, it only supports 10000 data rows and one logical processor. ML models can be deployed to the web or mobile (only when the user interface is ready for real-time data collection) with the assistance of Rapid Miner.

What are statistical approaches?

Statistical approaches are model-based approaches such as a model is produced for the data, and objects are computed concerning how well they fit the model. Most statistical approaches to outlier detection are depends on developing a probability distribution model and considering how Iikely objects are below that model.

An outlier is an object that has a low probability concerning a probability distribution model of the data. A probability distribution model is produced from the data by computing the parameters of a user-defined distribution.

If the data is considered to have a Gaussian distribution, therefore the mean and standard deviation of the basic distribution can be measured by computing the mean and standard deviation of the data. The probability of every object below the distribution can be calculated.

A broad method of statistical tests based on been devised to identify outliers, or discordant observations, as they are known as in the statistical literature. Some of these discordancy tests are hugely specialized and consider a level of statistical knowledge further the capacity of this text.

Identifying the specific distribution of a data set − While several types of data can be defined by a small number of common distributions, including Gaussian, Poisson, or binomial, data sets with non-standard distributions are associatively common. Of course, if the wrong model is selected, therefore an object can be erroneously recognized as an outlier.

For instance, the data can be modeled as appearing from a Gaussian distribution, but can come from a distribution that has a larger probability (than the Gaussian distribution) of receiving values far from the mean. Statistical distributions with this kind of behavior are general in practice and called a heavy-tailed distributions.

The number of attributes used − Some statistical outlier detection techniques use to an individual attribute, but some techniques have been represented for multivariate data.

Mixtures of distributions − The data can be modeled as a combination of distributions, and outlier detection schemes can be produced based on such models. Although potentially more dynamic, such models are complex, both to learn and to use. For example, the distributions required to be identified earlier objects can be defined as outliers.

Statistical approaches to outlier detection have a firm foundation and constructed on standard statistical techniques, including computing the parameters of a distribution. When there is adequate knowledge of the data and the type of test that must be used these tests can be efficient. There are a broad method of statistical outlier’s tests for individual attributes. Fewer options are accessible for multivariate data, and these tests can implement poorly for high-dimensional record.

Introduction to Sampling and Resampling

Population is the set of all data points while Sample is the subset of Population.

Table of Content:

Sampling:

Sampling is a process of selecting group of observations from the population, to study the characteristics of the data to make conclusion about the population.

Example: Covaxin (a covid-19 vaccine) is tested over thousand of males and females before giving to all the people of country.

Types of Sampling:

Whether the data set for sampling is randomized or not, sampling is classified into two major groups:

- Probability Sampling
- Non-Probability Sampling

Probability Sampling (Random Sampling):

In this type, data is randomly selected so that every observations of population gets the equal chance to be selected for sampling.

Probability sampling is of 4 types:

- Simple Random Sampling
- Cluster Sampling
- Stratified Sampling
- Systematic Sampling

Non-Probability Sampling:

In this type, data is not randomly selected. It mainly depends upon how the statistician wants to select the data.

The results may or maynot be biased with the population.

Unlike probability sampling, each observations of population doesn't get the equal chance to be selected for sampling.

Non-probability sampling is of 4 types:

- Convenience Sampling
- Judgmental/Purposive Sampling
- Snowball/Referral Sampling
- Quota Sampling

Sampling Error:

Errors which occur during sampling process are known as Sampling Errors.

Or

Difference between observed value of a sample statistics and the actual value of a population parameters.

Mathematical Formula for Sampling Error:

$$Sampling\ Error\ = z \times \frac{\sigma}{\sqrt{n}}$$

Where,

z: $z - score\ value\ based\ on\ confidence\ interval\ (approx \approx 1.96)$

σ: $population\ standard\ deviation$

n: $sample\ size$

Sampling error can be reduced by:

- Increasing the sample size
- Classifying population into different groups

Advantage of Sampling:

- Reduce cost and Time
- Accuracy of Data
- Inferences can be applied to a larger population
- Less resource needed

Resampling:

Resampling is the method that consist of drawing repeatedly drawing samples from the population.

It involves the selection of randomized cases with replacement from sample.

Note: In machine learning resampling is used to improve the performance of the model.

Types of Resampling:

Two common method of Resampling are:

- K-fold Cross-validation
- Bootstrapping

K-fold cross-validation:

In this method population data is divided into k equal sets in which one set is considered as the test set for the experiment while all other set will be used to train the model.

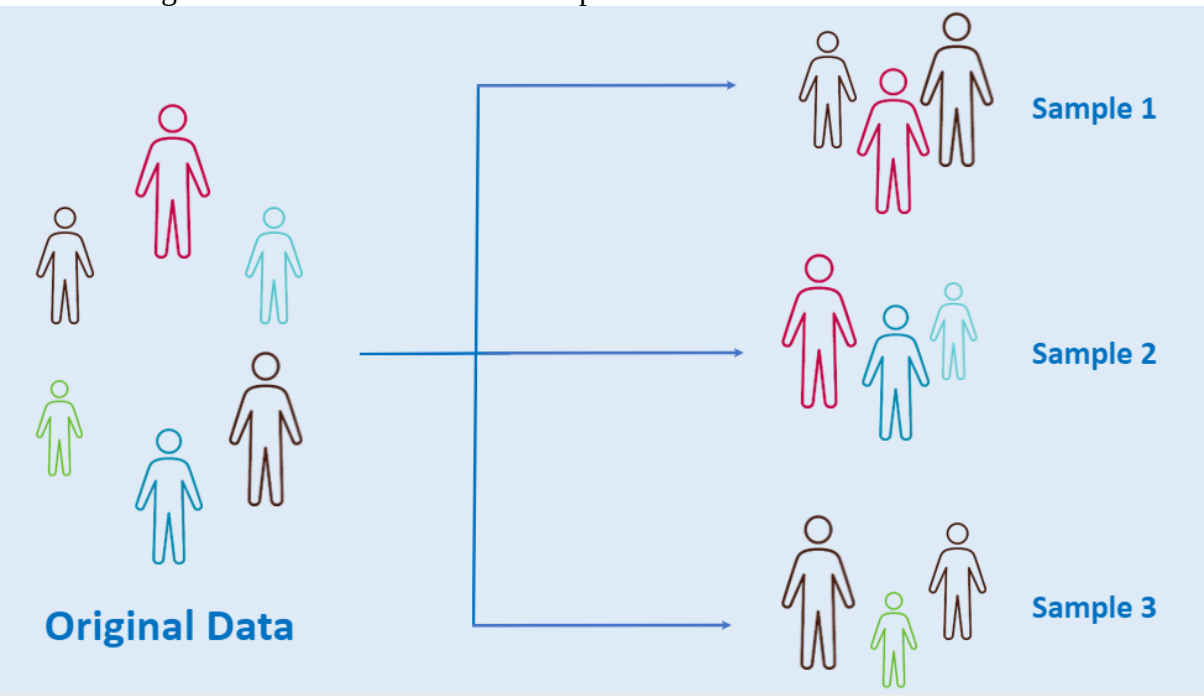
In first experiment, first set is considered as the test set and all other as trained set.

Process will be repeated k-time by choosing different sets as a test set.

Bootstrapping:

In bootstrapping, samples are drawn with replacement (i.e. one observation can be repeated in more than one group) and

the remaining data which are not used in samples are used to test the model.



Conclusion:

In this article we discussed about the introduction of sampling and resampling.

In predictive modelling problems sampling and resampling play an important role.

Ques 1. What is Sampling?

Ans 1: Sampling is a process of selecting group of observations from the population, to study the characteristics of the data to make conclusion about the population.

Ques 2. What is Resampling?

Ans 2. Resampling is the method that consist of drawing repeatedly drawing samples from the population.

It involves the selection of randomized cases with replacement from sample.

Introduction to Resampling methods

While reading about Machine Learning and Data Science we often come across a term called **Imbalanced Class Distribution** , generally happens when observations in one of the classes are much higher or lower than any other classes.

As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution. This problem is prevalent in examples such as Fraud Detection, Anomaly Detection, Facial recognition etc.

Two common methods of Resampling are –

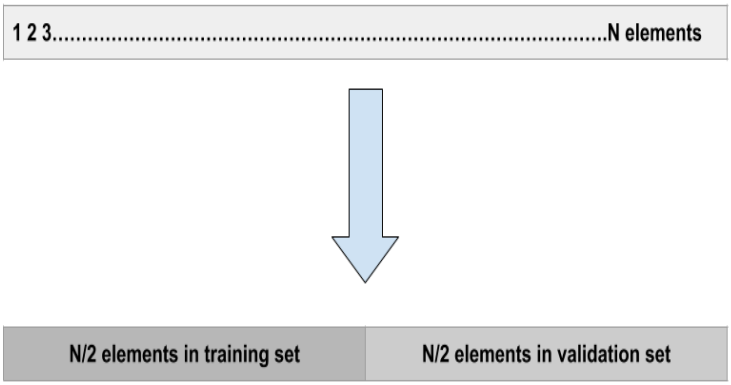
1. Cross Validation
2. Bootstrapping

Cross Validation –

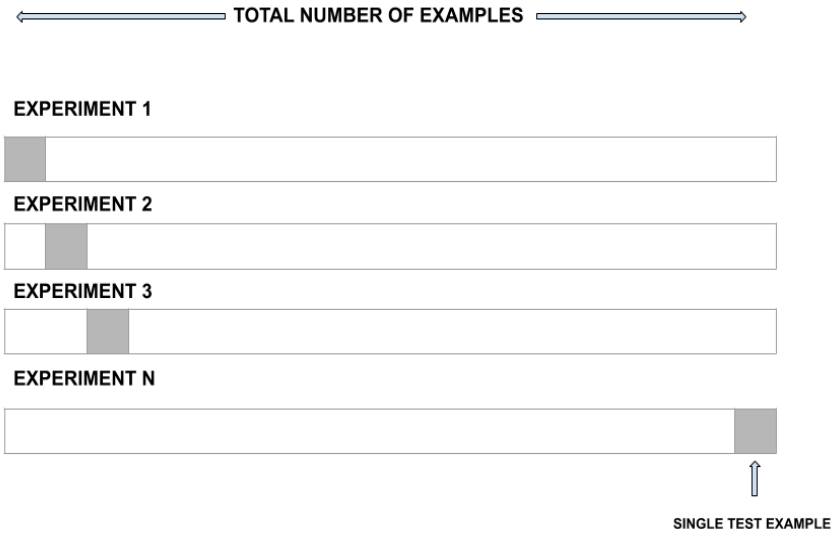
Cross-Validation is used to estimate the test error associated with a model to evaluate its performance.

Validation set approach:

This is the most basic approach. It simply involves randomly dividing the dataset into two parts: first a training set and second a validation set or hold-out set. The model is fit on the training set and the fitted model is used to make predictions on the validation set.

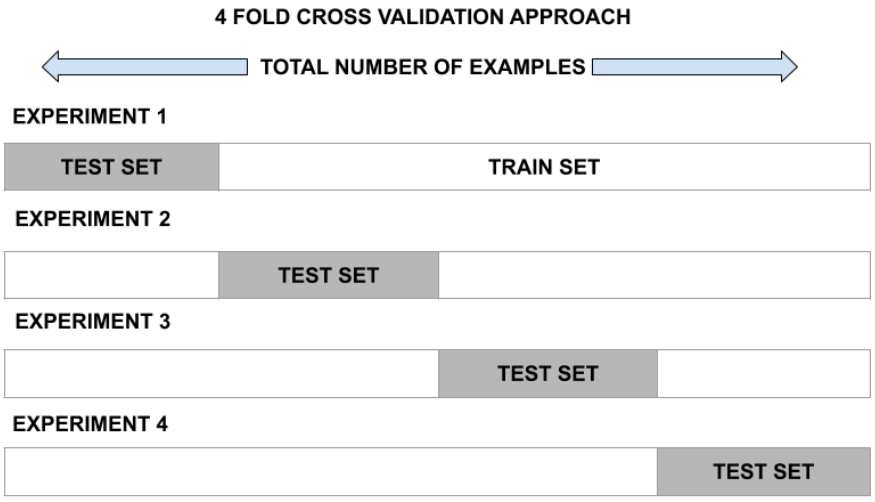


Leave-one-out-cross-validation:
LOOCV is a better option than the validation set approach. Instead of splitting the entire dataset into two halves only one observation is used for validation and the rest is used to fit the model.



k-fold cross-validation –

This approach involves randomly dividing the set of observations into k folds of nearly equal size. The first fold is treated as a validation set and the model is fit on the remaining folds. The procedure is then repeated k times, where a different group each time is treated as the validation set.



Bootstrapping –
Bootstrap is a powerful statistical tool used to quantify the uncertainty of a given model. However, the real power of bootstrap is that it could get applied to a wide range of models

where the variability is hard to obtain or not output automatically.