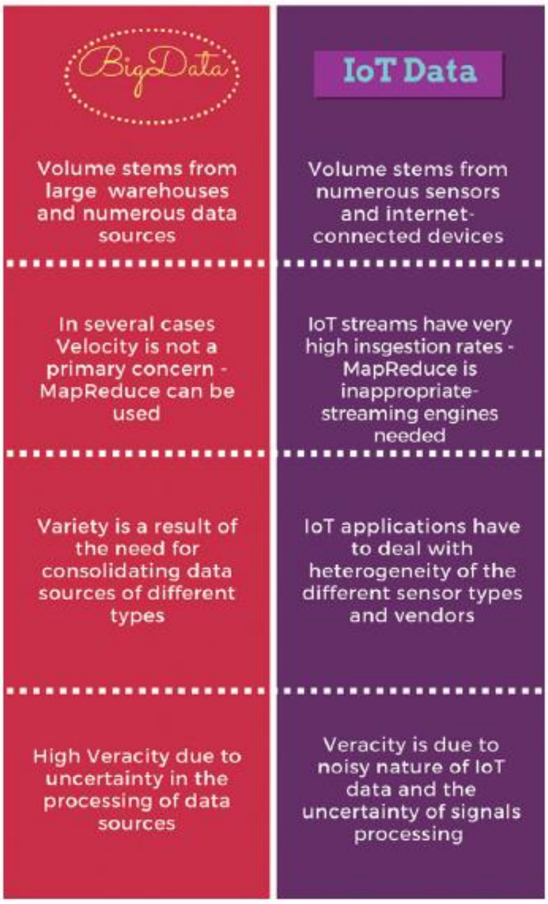
UNIT 5
DATA HANDLING & ANALYTICS

IoT Data and BigData

The rise of future internet technologies, including cloud computing and BigData analytics, enables the wider deployment and use of sophisticated IoT analytics applications, beyond simple sensor processing applications. It is therefore no accident that IoT technologies are converging with cloud computing and BigData analytics technologies towards creating and deploying advanced applications that process IoT streams. The integration of IoT data streams within cloud computing infrastructures enables IoT analytics applications to benefit from the capacity, performance and scalability of cloud computing infrastructures. In several cases, IoT analytics applications are also integrated with edge computing infrastructures, which decentralize processing of IoT data streams at the very edge of the network, while transferring only selected IoT data from the edge devices to the cloud. Therefore, it is



BigData	IoT Data
Volume stems from large warehouses and numerous data sources	Volume stems from numerous sensors and internet-connected devices
In several cases Velocity is not a primary concern - MapReduce can be used	IoT streams have very high insgestion rates - MapReduce is inappropriate- streaming engines needed
Variety is a result of the need for consolidating data sources of different types	IoT applications have to deal with heterogeneity of the different sensor types and vendors
High Veracity due to uncertainty in the processing of data sources	Veracity is due to noisy nature of IoT data and the uncertainty of signals processing

very common to deploy IoT analytics applications within edge and/or cloud computing infrastructures.

- **Volume**: IoT data sources (such as sensors) produce in most cases very large volumes of data, which typically exceed the storage and processing capabilities of conventional database systems.

- **Velocity**: IoT data streams have commonly very high ingestion rates, as they are produced continually, in very high frequencies and in several times in very short timescales.

- **Variety**: Due to the large diversity of IoT devices, IoT data sources can be very heterogeneous both in terms of semantics and data formats.

- **Veracity**: IoT data are a classical example of noise data, which are characterized by uncertainty. Therefore, systems, tools and techniques

IoT Analytics Lifecycle and Techniques

The IoT analytics lifecycle comprises the phases of data collection, analysis and reuse. In particular:

- **1st Phase – IoT Data Collection**: As part of this phase IoT data are collected and enriched with the proper contextual metadata, such as location information and timestamps. Moreover, the data are validated in terms of their format and source of origin. Also, they are validated in terms of their integrity, accuracy and consistency. Hence, this phase addresses several IoT analytics challenges, such as the need to ensure consistency and quality. Note that IoT data collection presents several peculiarities, when compared to traditional data consolidation of distributed data sources, such as the need to deal with heterogeneous IoT streams.

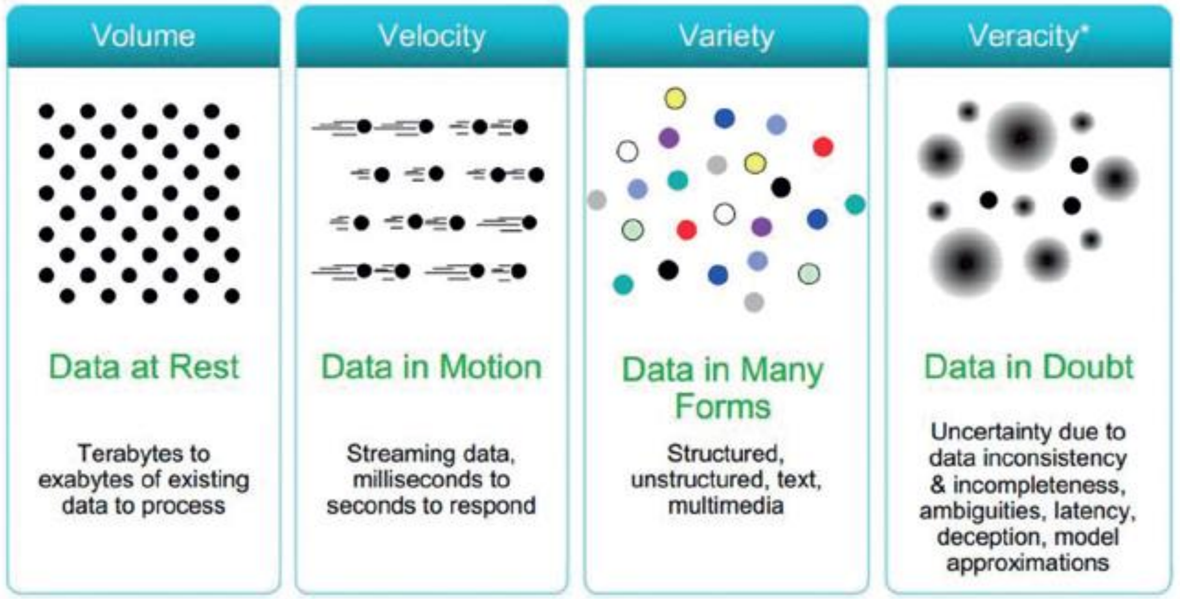
- **2nd Phase – IoT Data Analysis**:
This phase deals with the structuring, storage and ultimate analysis of IoT data streams. The latter analysis involves the employment of data mining and machine learning techniques such as classification, clustering and rules mining. These techniques are typically used to transform IoT data to actionable knowledge.

- **3rd Phase – IoT Data Deployment, Operationalization and Reuse**:
As part of this phase, the IoT analytics techniques identified in the previous steps are actually deployed, thus becoming operational. This phase ensures also the visualization of the IoT data/knowledge according to the needs of the application. Moreover, it enables the reuse of IoT knowledge and datasets across different applications.

Characteristics of IoT Generated Data

The volume and quality of the data generated by IoT devices is very different from the traditional transaction-oriented business data. Coming from millions of sensors and sensor-enabled devices, IoT data is more dynamic, heterogeneous, imperfect, unprocessed, unstructured and real-time than typical business data. It demands more sophisticated, IoT-specific analytics to make it meaningful.

As illustrated in Figure, the BigData is defined by 4 "Vs", which are Volume, Velocity, Variety and Veracity. The first V is for a large volume of data, not gigabytes but rather thousands of terabytes. The second V is referencing data streams and real-time processing. The third V is referencing the heterogeneity of the data: structure and unstructured, diverse data models, query language, and data sources.


Volume
Velocity
Variety
Veracity*
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

The fourth V is defining the data uncertainty, which can be due to data inconsistency, incompleteness, ambiguities, latency and lack of precise model. The IoT faces all 4 Vs of the BigData challenges. However the velocity is the main challenge: we need to process in real-time the data coming from IoT devices. For example, medical wearable such as Electro Cardio Graphic sensors produce up to 1000 events per second, which is a challenge for real-time processing. The volume of data is another important challenge. For example General Electric gathers each day 50 million pieces of data from 10 million sensors. A wearable sensor produces about 55 million data points per day. In addition, IoT also faces verity and veracity BigData challenges.

Data Analytic Techniques and Technologies

A cloud-based IoT analytics platform provides IoT-specific analytics that reduce the time, cost and required expertise to develop analytics-rich, vertical IoT applications. Platform's IoT-specific analytics uncover insights, create new information, monitor complex environments, make accurate predictions, and optimize business processes and operations. The applications of the IoT BigData Platform can be classified into four main categories i) deep understanding and insight knowledge ii) Real time actionable insight iii) Performance optimization and iv) proactive and predictive applications.

In the following we provide various technologies allowing building such an IoT analytics platform.

Batch Processing

Batch processing supposes that the data to be treated is present in a database. The most widely used tool for the case is **Hadoop MapReduce**. MapReduce is a programming model and Hadoop an implementation, allowing processing large data sets with a parallel, distributed algorithm on a cluster. It can run on inexpensive hardware, lowering the cost of a computing cluster. The latest version of MapReduce is YARN, called also MapReduce 2.0. **Pig** provides a higher level of programming, on top of MapReduce. It has its own language, PigLatin, similar to SQL. Pig Engine parses, optimizes and automatically executes PigLatin scripts as a series of MapReduce jobs on a Hadoop cluster. Apache **Spark** is a fast and general-purpose cluster computing system.

It provides high-levelAPIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It can be up to a hundred times faster than MapReduce with its capacity to work in-memory, allowing keeping large working datasets in memory between jobs, reducing considerably the latency. It supports batch and stream processing.

Stream Processing

Stream processing is a computer programming paradigm, equivalent to dataflow programming and reactive programming, which allows some applications to more easily exploit a limited form of parallel processing. **Flink** is a streaming dataflow engine that provides data distribution, communication and fault tolerance. It has almost no latency as the data are streamed in real-time (row by row). It runs on YARN and works with its own extended version of MapReduce.

Machine Learning

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. It is especially useful in the context of IoT when some properties of the data collected need to be discovered automatically. Apache Spark comes with its own machine learning library, called **MLib**.

It consists ofcommonlearning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction. Algorithms can be grouped in 3 domains of actions: Classification, association and clustering. To choose an algorithm, different parameters must be considered: scalability, robustness, transparency and proportionality.

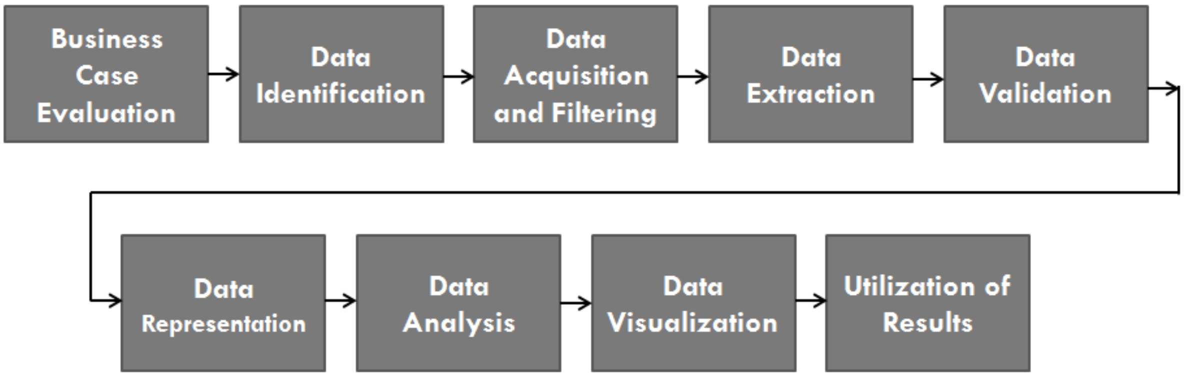
KNIME is an analytic platform that allows the user to process the data in a user-friendly graphical interface. It allows training of models and evaluation of different machine learning algorithms rapidly. If the workflow is already deployed on Hadoop, **Mahout,** a machine learning library can be used. Spark also has his own machine learning library called MLib.

H20 is a software dedicated to machine-learning, which can be deployed on Hadoop and Spark. It has an easy to use Web interface, which makes possible to combin BigData analytics easily with machine learning algorithm to train models.

Data Visualisation

Freeboard offers simple dashboards, which are readily useable sets of widgets able to display data. There is a direct Orion Fiware connector. Freeboard offers a REST API allowing controlling of the displays. **Tableau Public** is a free service that lets anyone publish interactive data to the web. Once on the web, anyone can interact with the data, download it, or create their own visualizations of it. No programming skills are required. Tableau allows the upload of analysed data from .csv format, for instance. The visualisation tool is very powerful and allows a deep exploration the data.

Kibana is an open source analytics and visualization platform designed to work with Elasticsearch. Kibana allows searching, viewing, and interacting with data stored in Elasticsearch indices. It can perform advanced data analysis and visualize data in a variety of charts, tables, and maps.



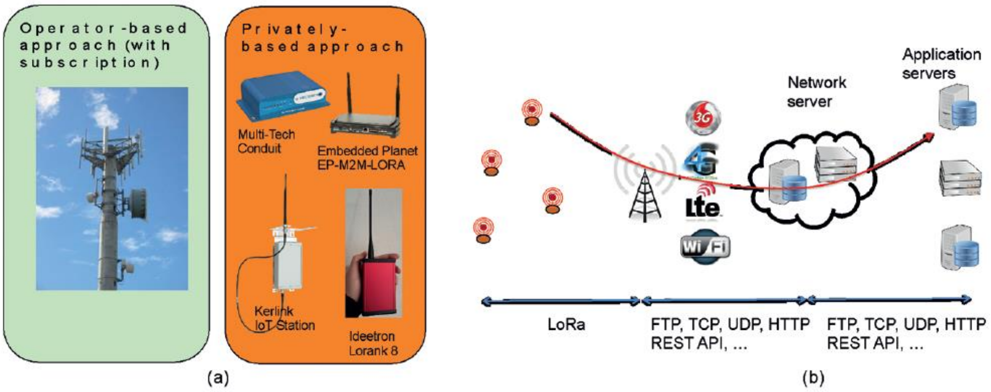
Data Acquisition:

1. Data Collection Using Low-power, Long-range Radios

Regarding the deployment of IoT devices in a large scale, it is still held back by technical challenges such as short communication distances. Using the traditional mobile telecommunications infrastructure is still very expensive (e.g., GSM/GPRS, 3G/4G) and not energy efficient for autonomous devicesthat must run on battery for months. During the last decade, low-power but short-range radio such as IEEE 802.15.4 radio have been considered by the WSNcommunity with multi-hop routing to overcome the limited transmission range. While such short-range communications can eventually be realized on smart cities infrastructures where high node density with powering facility can be achieved, it can hardly be generalized for the large majority of surveillance applications that need to be deployed in isolated or rural environments. Future 5G/LTE standards do have the IoT orientation but these technologies and standards are not ready yet while the demand is already high.

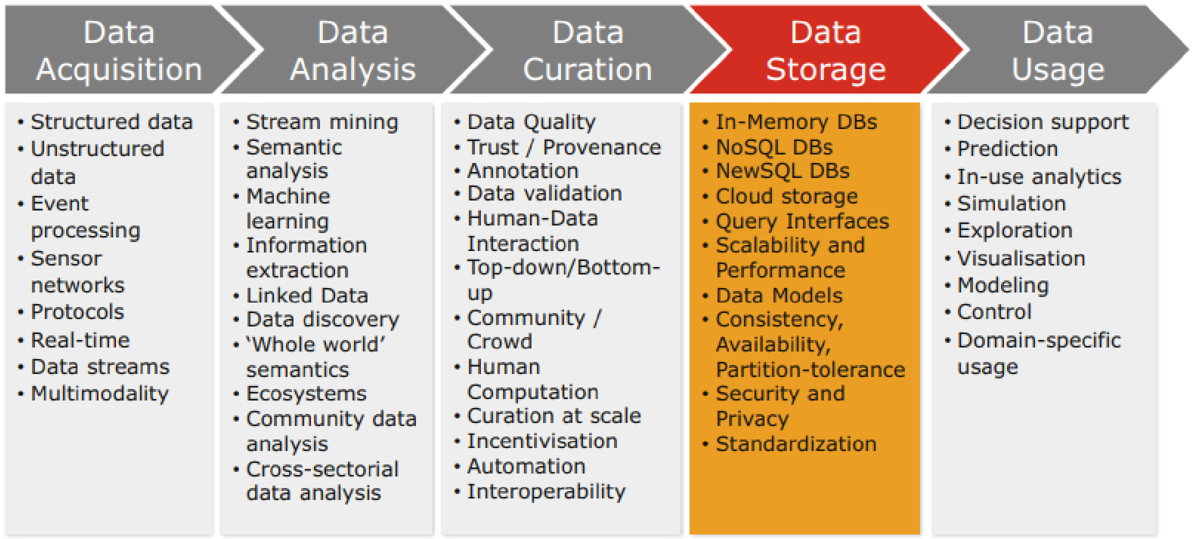
Architecture and Deployment

The deployment of LPWAN (both operator-based and privately-owned scenarios) is centred on gateways that usually have Internet connectivity as shown in Figure Although direct communications between devices are possible, most of IoT applications


Operator-based approach (with subscription)
Privately-based approach
Multi-Tech Conduit
Embedded Planet EP-M2M-LORA
Kerlink IoT Station
Ideetron Lorank 8
(a)
Network server
Application servers
3G
4G
lte
Wi Fi
LoRa
FTP, TCP, UDP, HTTP REST API, ...
FTP, TCP, UDP, HTTP REST API, ...
(b)

follow the gateway-centric approach with mainly uplink traffic patterns. In this typical architecture data captured by end-devices are sent to a gateway which will push data to well identified network servers. Then application servers managed by end-users could retrieve data from the network server. If encryption is used for confidentiality, the application server can be the place where data could be decrypted and presented to end-users.

Big Data Storage:



Data Acquisition

- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis

- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- 'Whole world' semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation

- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage

- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage

- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

Big data storage systems typically address the volume challenge by making use of distributed, shared nothing architectures. This allows addressing increased storage requirements by scaling out to new nodes providing computational power and storage. New machines can seamlessly be added to a storage cluster and the storage system takes care of distributing the data between individual nodes transparently.

Key Insights for Big Data Storage

As a result of the analysis of current and future data storage technologies, a number of insights were gained relating to data storage technologies. It became apparent that big data storage has become a commodity business and that scalable storage technologies have reached an enterprise-grade level that can manage virtually unbounded volumes of data

- **Potential to Transform Society and Businesses across Sectors:** Big data storage technologies are a key enabler for advanced analytics that have the potential to transform society and the way key business decisions are made. This is of particular importance in traditionally non-IT-based sectors such as energy. While these sectors face non-technical issues such as the lack of skilled big data experts and regulatory barriers, novel data storage technologies have the potential to enable new value-generating analytics in and across various industrial sectors.

- **Lack of Standards Is a Major Barrier:** The history of NoSQL is based on solving specific technological challenges which lead to a range of different storage technologies. The large range of choices coupled with the lack of standards for querying the data makes it harder to exchange data stores as it may tie application specific code to a certain storage solution.

- **Open Scalability Challenges in Graph-Based Data Stores:** Processing data based on graph data structures is beneficial in an increasing amount of applications. It allows better capture of semantics and complex relationships with other pieces of information coming from a large variety of different data sources, and has the potential to improve the overall value that can be generated by analysing the data.

Data Storage Technologies

During the last decade, the need to deal with the data explosion (Turner et al. 2014) and the hardware shift from scale-up to scale-out approaches led to an explosion of new big data storage systems that shifted away from traditional relational database models. These approaches typically sacrifice properties such as data consistency in order to maintain fast query responses with increasing amounts of data. Big data stores are used in similar ways as traditional relational database management systems, e.g. for online transactional processing (OLTP) solutions and data warehouses over structured or semi-structured data. Particular strengths are in handling unstructured and semi-structured data at large scale. unstructured and semi-structured data at large scale.

This section assesses the current state-of-the-art in data store technologies that are capable of handling large amounts of data, and identifies data store related trends. Following are differing types of storage systems:

- **Distributed File Systems:** File systems such as the Hadoop File System (HDFS) (Shvachko et al. 2010) offer the capability to store large amounts of unstructured data in a reliable way on commodity hardware. Although there are file systems with better performance, HDFS is an integral part of the Hadoop framework (White 2012) and has already reached the level of a de-facto standard. It has been designed for large data files and is well suited for quickly ingesting data and bulk processing.

- **NoSQL Databases**: Probably the most important family of big data storage technologies are NoSQL database management systems. NoSQL databases use data models from outside the relational world that do not necessarily adhere to the transactional properties of atomicity, consistency, isolation, and durability (ACID).

- **NewSQL Databases:** A modern form of relational databases that aim for comparable scalability as NoSQL databases while maintaining the transactional guarantees made by traditional database systems.

- **Big Data Querying Platforms:** Technologies that provide query facades in front of big data stores such as distributed file systems or NoSQL databases. The main concern is providing a high-level interface, e.g. via SQL3 like query languages and achieving low query latencies.

Hadoop:

. Hadoop provides storage for Big Data at reasonable cost

Storing Big Data using traditional storage can be expensive. Hadoop is built around commodity hardware. Hence it can provide fairly large storage for a reasonable cost. Hadoop has been used in the field at Peta byte scale. One study by Cloudera suggested that Enterprises usually spend around $25,000 to $50,000 dollars per tera byte per year. With Hadoop this cost drops to few thousands of dollars per tera byte per year. And hardware gets cheaper and cheaper this cost continues to drop.

Hadoop allows to capture new or more data

Some times organizations don't capture a type of data, because it was too cost prohibitive to store it. Since Hadoop provides storage at reasonable cost, this type of data can be captured and stored. One example would be web site click logs. Because the volume of these logs can be very high, not many organizations captured these. Now with Hadoop it is possible to capture and store the logs

With Hadoop, you can store data longer

To manage the volume of data stored, companies periodically purge older data. For example only logs for the last 3 months could be stored and older logs were deleted. With Hadoop it is possible to store the historical data longer. This allows new analytics to be done on older historical data. For example, take click logs from a web site. Few years ago, these logs were stored for a brief period of time to calculate statics like popular pages ..etc. Now with Hadoop it is viable to store these click logs for longer period of time.

Hadoop provides scalable analytics

There is no point in storing all the data, if we can't analyze them. Hadoop not only provides distributed storage, but also distributed processing as well. Meaning we can crunch a large volume of data in parallel. The compute framework of Hadoop is called Map Reduce. Map Reduce has been proven to the scale of peta bytes.

EXAMPLES OF BIG DATA ANALYTICS

Let us consider several examples of companies that are using big data analytics. The examples illustrate the use of different sources of big data and the different kinds of analytics that can be performed. Introducing a New Coffee Product at Starbucks Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong. The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and niche coffee forum discussion groups to assess customers’ reactions. By mid-morning, Starbucks discovered that although people liked the taste of the coffee, they thought that it was too expensive. Starbucks lowered the price, and by the end of the day all of the negative comments had disappeared. Compare this fast response with a more traditional approach of waiting for the sales reports to come in and noticing that sales are disappointing. A next step might be to run a focus group to discover why.

Perhaps in several weeks Starbucks would have discovered the reason and responded by lowering the price. Drilling for Oil at Chevron Each drilling miss in the Gulf of Mexico costs Chevron upwards of $100 million. To improve its chances of finding oil, Chevron analyzes 50 terabytes of seismic data. Even with this, the odds of finding oil have been around 1 in 5. In the summer of 2010, because of BP’s Gulf oil spill, the federal government suspended all deep water drilling permits. The geologists at Chevron took this time to seize the opportunity offered by advances in computing power and storage capacity to refine their already advanced computer models. With these enhancements, Chevron has improved the odds of drilling a successful well to nearly 1 in 3, resulting in tremendous cost savings. Monitoring Trucks at U.S. Xpress U.S. Xpress is a transportation company. Its cabs continuously stream more than 900 pieces of data related to the condition of the trucks and their locations [Watson and Leonard, 2011].

This data is stored in the cloud and analyzed in various ways, with information delivered to various users, from drivers to senior executives, on iPads and other tablet computers. For example, when a sensor shows that a truck is low on fuel, the driver is directed to a filling station where the price is low. If a truck appears to need maintenance, drivers are sent to a specific service depot. Routes and destinations are changed to ensure that orders are delivered on time.

Types of Big Data Analytics

Prescriptive Analytics

The most valuable and most underused big data analytics technique, prescriptive analytics gives you a laser-like focus to answer a specific question. It helps to determine the best solution among a variety of choices, given the known parameters and suggests options for how to take advantage of a future opportunity or mitigate a future risk. It can also illustrate the implications of each decision to improve decision-making. Examples of prescriptive analytics for customer retention include next best action and next best offer analysis.

- Forward looking
- Focused on optimal decisions for future situations
- Simple rules to complex models that are applied on an automated or programmatic basis
- Discrete prediction of individual data set members based on similarities and differences
- Optimization and decision rules for future events

Diagnostic Analytics

Data scientists turn to this technique when trying to determine why something happened. It is useful when researching leading churn indicators and usage trends amongst your most loyal customers. Examples of diagnostic analytics include churn reason analysis and customer health score analysis. Key points:

- Backward looking
- Focused on causal relationships and sequences
- Relative ranking of dimensions/variable based on inferred explanatory power)
- Target/dependent variable with independent variables/dimensions
- Includes both frequentist and Bayesian causal inferential analyses

Descriptive Analytics

This technique is the most time-intensive and often produces the least value; however, it is useful for uncovering patterns within a certain segment of customers. Descriptive analytics provide insight into what has happened historically and will provide you with trends to dig into in more detail. Examples of descriptive analytics include summary statistics, clustering and association rules used in market basket analysis. Key points:

- Backward looking
- Focused on descriptions and comparisons
- Pattern detection and descriptions

- MECE (mutually exclusive and collectively exhaustive) categorization
- Category development based on similarities and differences (segmentation)

Predictive Analytics

The most commonly used technique; predictive analytics use models to forecast what might happen in specific scenarios. Examples of predictive analytics include next best offers, churn risk and renewal risk analysis.

- Forward looking
- Focused on non-discrete predictions of future states, relationship, and patterns
- Description of prediction result set probability distributions and likelihoods
- Model application
- Non-discrete forecasting (forecasts communicated in probability distributions)

Outcome Analytics

Also referred to as consumption analytics, this technique provides insight into customer behavior that drives specific outcomes. This analysis is meant to help you know your customers better and learn how they are interacting with your products and services.

- Backward looking, Real-time and Forward looking
- Focused on consumption patterns and associated business outcomes
- Description of usage thresholds
- Model application

EXAMPLES OF BIG DATA ANALYTICS:

Let us consider several examples of companies that are using big data analytics. The examples illustrate the use of different sources of big data and the different kinds of analytics that can be performed

Introducing a New Coffee Product at Starbucks

Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong. The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and niche coffee forum discussion roups to assess customers' reactions. By mid-morning, Starbucks discovered that although people liked the taste
of the coffee, they thought that it was too expensive. Starbucks lowered the price, and by the end of the day all of the negative comments had disappeared.

Compare this fast response with a more traditional approach of waiting for the sales reports to come in and noticing that sales are disappointing. A next step might be to run a focus group to discover why. Perhaps in several weeks Starbucks would have discovered the reason and responded by lowering the price.

Drilling for Oil at Chevron

Each drilling miss in the Gulf of Mexico costs Chevron upwards of $100 million. To improve its chances of finding oil, Chevron analyzes 50 terabytes of seismic data. Even with this, the odds of finding oil have been around 1 in 5. In the summer of 2010, because of BP's Gulf oil spill, the federal government suspended all deep water drilling permits. The geologists at Chevron took this time to seize the opportunity offered by advances in computing power and storage capacity to refine their already advanced computer models. With these enhancements, Chevron has improved the odds of drilling a successful well to nearly 1 in 3, resulting in tremendous cost savings.

Monitoring Trucks at U.S. Xpress

U.S. Xpress is a transportation company. Its cabs continuously stream more than 900 pieces of data related to the condition of the trucks and their locations [Watson and Leonard, 2011]. This data is stored in the cloud and analysed in various ways, with information delivered to various users, from drivers to senior executives, on iPads and other tablet computers. For example, when a sensor shows that a truck is low on fuel, the driver is directed to a filling station where the price is low. If a truck appears to need maintenance, drivers are sent to a specific service depot Routes and destinations are changed to ensure that orders are delivered on time.

Statistical Models/ Methods:

The recent methodologies for big data can be loosely grouped into three categories: resampling-based, divide and conquer, and online updating. To put the different methods in a context, consider a dataset with *n* independent and identically distributed observations, where *n* is too big for standard statistical routines such as logistic regression

1. Subsampling-based methods/ Models:

1.1 Bags of little bootstrap:

It is a combination of subsampling, the *m*-out-of-*n* bootstrap, and the bootstrap to achieve computational efficiency. BLB consists of the following steps. First, draw *s* subsamples of size *m* from the original data of size *n*. For each of the subsets, draw *r* bootstrap samples of size *n* instead of *m*, and obtain the point estimates and their quality measures (e.g., confidence interval) from the *r* bootstrap samples. Then, the *s* bootstrap point estimates and quality measures are combined (e.g., by average) to yield the overall point estimates and quality measures. In summary, BLB has two nested procedures: the inner procedure applies the bootstrap to a subsample, and the outer procedure combines these multiple bootstrap estimates.

1.2 Leveraging

In a leveraging method, one samples a small proportion of the data with certain weights (subsample) from the full sample, and then performs intended computations for the full sample using the small subsample as a surrogate. The key to success of the leveraging methods is to construct the weights, the nonuniform sampling probabilities, so that influential data points are sampled with high probabilities

1.3 Mean log-likelihood

The method uses Monte Carlo averages calculated from subsamples to approximate the quantities needed for the full data. Motivated from minimizing the Kullback–Leibler (KL) divergence, they approximate the KL divergence by averages calculated from subsamples. This leads to a maximum mean log-likelihood estimation method.

1.4 Subsampling-based MCMC

As a popular general purpose tool for Bayesian inference, Markov chain Monte Carlo (MCMC) for big data is challenging because of the prohibitive cost of likelihood evaluation of every datum at every iteration. Liang and Kim (2013) extended the mean log-likelihood method to a bootstrap Metropolis–Hastings (MH) algorithm in MCMC. The likelihood ratio of the proposal and current estimate in the MH ratio is replaced with an approximation from the mean log-likelihood based on $k$bootstrap samples of size m.

2. Divide and conquer

A divide and conquer algorithm (which may appear under other names such as divide and recombine, split and conquer, or split and merge) generally has three steps: 1) partitions a big dataset into K blocks; 2) processes each block separately (possibly in parallel); and 3) aggregates the solutions from each block to form a final solution to the full data.

2.1 Aggregated estimating equations

For a linear regression model, the least squares estimator for the regression coefficient β for the full data can be expressed as a weighted average of the least squares estimator for each block with weight being the inverse of the estimated variance matrix. The success of this method for linear regression depends on the linearity of the estimating equations in β and that the estimating equation for the full data is a simple summation of that for all the blocks

2.2Majority voting

consider a divide and conquer approach for generalized linear models (GLM) where both the sample size n and the number of covariates p are large, by incorporating variable selection via penalized regression into a subset processing step. More specifically, for p bounded or increasing to infinity slowly, (p_n not faster than $o(e^{n_k})$, while model size may increase at a rate of $o(n_k)$), they propose to first randomly split the data of size *n* into *K* blocks (size $n_k = O(n/K)$). In step 2, penalized regression is applied to each block separately with a sparsity-inducing penalty function satisfying certain regularity conditions.

Analysis of Variance:

Analysis of Variance (ANOVA) is used to compare mean between two or more items. It's a statistical method that yields values that can be tested to determine whether a significant relation exists between variables.

Example:

- A car company wishes to compare the average petrol consumption of three similar models of cars and has six vehicles available for each model. It follows a 6×3 matrix, columns have cars and rows have models. Here, we compare the average petrol consumption.
- A teacher is interested in comparing the average percentage marks attained in the examinations of five different subjects and the marks are available for eight students, who have completed each examination. If the teacher wants to compare the mean average % of marks between all students of five different subjects, for comparing the mean between two entities we use Analysis of Variance.

Two-way Analysis of Variance

Let's take an example of a case which has elements such as Observation, Gender, Dosage with 16 observations of each. They all must be numerical since mean and variance is being used.

Here in Gender, we have to convert into dummy variable which involves assigning numbers like 1 and O for male and female. But LSS of variance can only be applied on quantitative data.

ANOVA is a particular form of statistical hypothesis test heavily used in the analysis of experiment data. A statistical hypothesis test is a method of making decision using data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, assuming the truth of the null hypothesis. A statistically significant result, when a probability (p-value) is less than a threshold (significance level), justifies the rejection of the null hypothesis but only if the prior probability of the null hypothesis is not high.

One-way Analysis of Variance

The above table has elements such as Df & Sum Sq which are an integral part of the One-way Analysis of Variance.

Df(Degree of Freedom) – In a statistical point of view, let's say data is end point with no statistical constraints. Here, the Degree of Freedom is N. When mean of N data is 1,000, the degree of freedom would be N-1. If there are more statistical constraints then degree of freedom will be N-2 and so on.

Sum Sq (Sum of Square)– It's a way of calculating variation. When we talk about variation, it's always calculated between value and mean.

ANOVA is a synthesis of several ideas and is used for multiple purposes. As a consequence, it is difficult to define concisely or precisely. It is used in logistic regression as well. It's not only used for calculating mean but also checking the different model performance. F-Test is used to compare the variation between the explained variance and unexplained variance. In ANOVA, we take the F-Test based on the within group variation to between group variation.

Dispersion:

Dispersion is used to measure the variability in the data or to see how spread out the data is. It measures how much the scores in a distribution vary from the typical score. So once we know the typical score of a set of values, we might want to know "how typical" it is of the entire group of data, or how much the scores vary from that typical score. This is measured by dispersion. When dispersion is low, the central tendency is more accurate or more representative of the data as majority of the data points are near the typical value, thus resulting in low dispersion and vice versa.

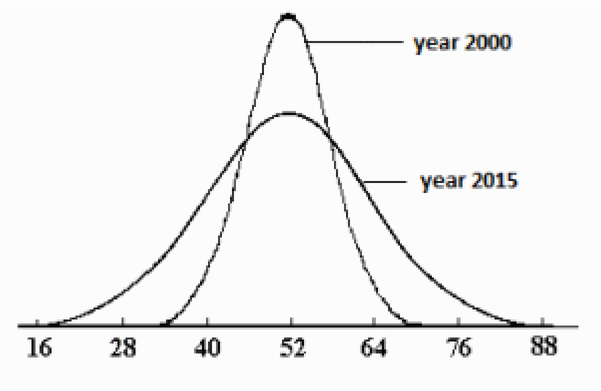


fig: Dispersion

Measuring dispersion is important to validate the claims like – "The rich are getting richer, and poor are getting poorer". Assume that we measured the incomes of a random sample of population in the year 2000 and again in the year 2015. In this case, mean is not important to answer the question, because it can be the case that the two distributions have the same mean but different dispersion (i.e. spread) of data which is more important to answer this question.

Let the distribution of income in our random samples from year 2000 and 2015 be represented by the two curves as shown in the diagram. We see that the two curves for the year 2015 and year 2000 have the same mean i.e. 52, but the curve for year 2015 is more spread out as compared to that for year 2000. Thus, as compared to 2000, there were more people in 2015 with higher and lower incomes which validates our claim.

Dispersion can be measured in a number of ways:

1. **Range** – The range of a dataset gives the difference between the largest and smallest value. Therefore, the range only takes the two most extreme values into account and tells nothing about what falls in the middle.
   Range = (Max. Value – Min. Value)

2. **Variation Ratio** – When data is measured at the nominal level, we can only compare the size of each category represented in the dataset. Variation Ratio (VR) tells us how much variation there is in the dataset from the modal category. A low VR indicates that the data is concentrated more in the modal category and dispersion is less, whereas a high VR indicates that the data is more distributed across categories, thereby resulting into high dispersion.
   Variation Ratio = 1 – (#data points in modal category / #total data points in all categories)

3. **Mean Deviation** – The mean deviation gives us a measure of the typical difference (or deviation) from the mean. If most data values are very similar to the mean, then the mean deviation score will be low, indicating high similarity within the data. If there is great variation among scores, then the mean deviation score will be high, indicating low similarity within the data. Let Xi be the observed value of data point, X(bar) be the mean and N be the total data points.

$$MD = \frac{\sum |x_i - \bar{x}|}{N}$$

4. **Standard Deviation** – The standard deviation of a dataset gives a measure of how each value in a dataset varies from the mean. This is very similar to the mean deviation, and indeed, gives us quite similar information, substantively. Let Xi be the observed value of data point, X(bar) be the mean and N be the total data points.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Thus, range tells us how far the extreme points of distribution are, and variation ratio tells us about the distribution of data in different categories relative to the modal category. Also, standard deviation and mean deviation tells us how data points in the distribution varies w.r.t the mean.

Regression analysis :

Regression analysis is used to estimate the strength and direction of the relationship between variables that are *linearly* related to each other. Two variables X and Y are said to be *linearly* related if the relationship between them can be written in the form

$$Y = mX + b$$

where
m is the *slope,* or the change in Y due to a given change in X
b is the *intercept,* or the value of Y when $X = 0$

As an example of regression analysis, suppose a corporation wants to determine whether its advertising expenditures are actually increasing profits, and if so, by how much. The corporation gathers data on advertising and profits for the past 20 years and uses this data to estimate the following equation:

$$Y = 50 + 0.25X$$

where

Y represents the annual profits of the corporation (in millions of dollars).

X represents the annual advertising expenditures of the corporation (in millions of dollars).

In this equation, the slope equals 0.25, and the intercept equals 50. Because the slope of the regression line is 0.25, this indicates that on average, for every $1 million increase in advertising expenditures, profits rise by $.25 million, or $250,000. Because the intercept is 50, this indicates that with no advertising, profits would still be $50 million.

This equation, therefore, can be used to forecast future profits based on planned advertising expenditures. For example, if the corporation plans on spending $10 million on advertising next year, its expected profits will be as follows:

$$Y = 50 + 0.25X$$

$$Y = 50 + 0.25(10) = 50 + 2.5 + 52.5$$

Hence, with an advertising budget of $10 million next year, profits are expected to be $52.5 million.

Precision:

Precision refers to the level of measurement and exactness of description in a GIS database. Precise locational data may measure position to a fraction of a unit. Precise attribute information may specify the characteristics of features in great detail. It is important to realize, however, that precise data – no matter how carefully measured – may be inaccurate. Surveyors may make mistakes or data may be entered into the database incorrectly.

- The level of precision required for particular applications varies greatly. Engineering projects such as road and utility construction require very precise information measured to the millimeter or tenth of an inch. Demographic analyses of marketing or electoral trends can often make do with less, say to the closest zip code or precinct boundary.
- Highly precise data can be very difficult and costly to collect. Carefully surveyed locations needed by utility companies to record the locations of pumps, wires, pipes and transformers cost $5-20 per point to collect.

High precision does not indicate high accuracy nor does high accuracy imply high precision. But high accuracy and high precision are both expensive.

Be aware also that GIS practitioners are not always consistent in their use of these terms. Sometimes the terms are used almost interchangeably and this should be guarded against.

Two additional terms are used as well:

- **Data quality** refers to the relative accuracy and precision of a particular GIS database. These facts are often documented in data quality reports.
- **Error** encompasses both the imprecision of data and its inaccuracies.