Business Anaytics

M. Tech QROR – 2nd yr (2024)

Introduction to Analytics & Data Quality

Prasun Das

SQC & OR Unit

Indian Statistical Institute

e-mail: prasun@isical.ac.in

Business Problems

Business: An organization involved in trade of goods, services, or both to consumers.

Problem: A perceived gap between (existing state, desired state), or a deviation from a norm, standard, or status quo.

Business Problem (*format*): "The problem of W affects X, the impact of which is Y, so a successful solution would be to Z."

Four key elements of a ***problem statement*:**

1. Root cause (W)
2. Affected stakeholders/product users (X)
3. Impact of the issues (Y)
4. A successful solution must include (Z)

Example:

- W: customers smoking in a public place
- X: other non-smoking group of customers
- Y: low customer satisfaction, passive smoking, increased cleaning cost.
- Z: increase awareness of smoking & smoking effects, air pollution and impose warning/rules to eliminate smoking from the place.

Business Intelligence vs. Business Analytics

Business Intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning. Intelligence is querying, mining, reporting, OLAP, and alert tools (dashboards) that can answer questions (https://www.tableau.com/learn/articles/business-intelligence) using Data Warehouse and visualization thereof. It, of course, combines data analytics tools & techniques.

Business Analytics focuses on understanding business performance based on data and statistical/quantitative methods used for explanatory, predictive and optimization modeling. Analytics may be used as input for human decisions or may drive fully automated decisions.

Business Analytics: Domain

1. Understanding Business Problem
2. Databases and Data Warehouses (Information management)
3. Data Processing (Quality) / Visualization (Diagnostic / Descriptive Analytics)
4. Mining, Modelling and/or Optimization (Predictive and/or Prescriptive Analytics)
5. Making Business Decisions
6. Implementing Business Solutions (Deployment)

Business domain: HR, Sales and Marketing, Operations, Logistics and Supply Chain, Retail, Health care, Banking and Financial services, Manufacturing, Sports and so on…

Business Analytics – Types

Analytic Excellence Leads to Better Decisions

Business Analytics – Types

1. **Descriptive Analytics:** gain insights/information from historical/raw data.
2. **Diagnostics Analytics:** does cause-and-effect analysis (RCA/FMEA/HIRA).
3. **Predictive analytics:** predictive modeling using statistical, data-mining and machine learning techniques with the goal to identify a model or set of models that can be used to predict some response of interest.
4. **Prescriptive analytics:** recommend decisions using optimization, simulation etc. It requires a predictive model with two additional components: actionable data and a feedback system that tracks the outcome produced by the action taken.

Business Analytics – in greater details

Purpose	Type of Analytics	Tools / Techniques
How good in the Data?	Preparatory Analytics	Data Quality rules/scores, missing data estimation, outliers detection etc.
What happened?	Descriptive Analytics	Data profiling (ETL), Data mining, Descriptive statistics etc.
Why and when it happened?	Diagnostic Analytics	Control charts, Hypothesis testing, ANOVA etc.
How did it happen?	Causal Analytics	C&E analysis, FMEA
What will happen?	Predictive Analytics	Modeling – clustering / classification / prediction / forecasting etc.
How can we improve?	Prescriptive Analytics	DOE, Optimization, Decision making, Simulation etc.
How confident can we be?	Reliability based Analytics	Reliability models, Risk analysis etc.

In Short,

Business Analytics ≡ KDD + DM + DM

Non-trivial extraction of implicit, previously unknown and potentially useful information from data, or the search for relationships and global patterns that exist in databases followed by business decision making.

KDD: Knowledge Discovery in Databases

DM: Data Mining

DM: Decision Making

DATA AND ITS QUALITY

Data

- Data are facts and figures about any characteristic of individuals.
- Example:

  Color of Maruti Suzuki car is Red

  Prabal has got 95 marks in Mathematics

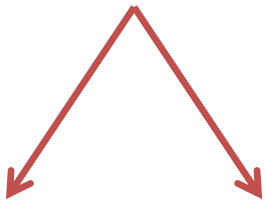
  Microstructure of DP Steel contains two phases

TYPES OF DATA

Attribute Data (Qualitative)

- *Students – Good or Naughty*
- *Road Accidents – Minor, Severe or Fatal*
- *Passing rate – Low, Medium or High*
- *Fraud detection – Guilty or Not guilty*

Variable Data (Quantitative)



Discrete Data

- *Countable & finite*

Continuous Data

- *All possible values over real line*

Variable Data (Quantitative)

Discrete

- *Number of footfalls per day in a retail store*
- *Number of accidents in a month*
- *Number of cars produced in a fortnight*
- *Number of people landed in a day at Jorhat airport*
- *Number of telephone calls in a week*

Continuous

- *Depth of bore well in a district*
- *Failure rate of loom machine*
- *%marks in B.tech CSE course*
- *Cooling rate of metallic plate*
- *Age of 1st year students*

DATA: Analytics Perspective

Structured Data: relational databases and spreadsheets.

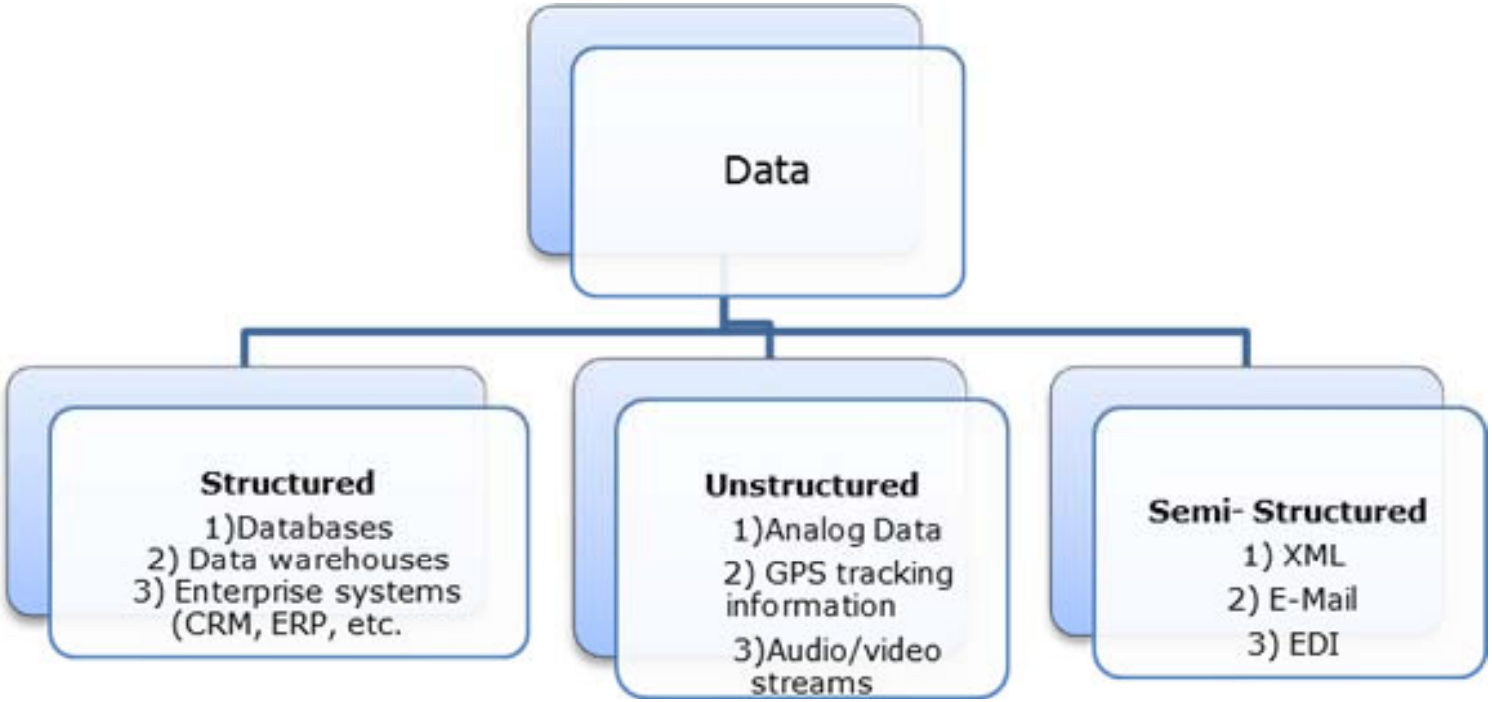
Examples: fields for phone numbers, pin codes, Aadhar card numbers etc., in general all quantitative and to some extent qualitative data.

Unstructured Data: NOT relational and doesn't fit into pre-defined database.

Examples: e-mail messages, word processing documents, videos, audios, webpages, photos, graphic images, streaming data, pdf files, ppt presentations, blog entries, wikis.

Most of the data (~80%) in any organization are UNSTRUCTURED.

DATA: Analytics Perspective



Semi-structured data don't reside in a relational database but do have some organizational properties. For ex., e-mails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.

Data Science Domain

The fundamental concepts of Data Science span the process of envisioning the problem, to applying data science techniques, to deploying the results to decision-making (DDD: Data-driven Decision-making).

Brynjolfsson, Erik, Hitt, Lorin M. and Kim, Heekyung Hellen (2011), Strength in Numbers: How Does Data-Driven Decision making Affect Firm Performance? http://dx.doi.org/10.2139/ssrn.1819486 (Findings: firms that adopt DDD have productivity 5-6% higher than what would be expected given their other investments and information technology usage.)

Note: Data engineering and processing, "Big Data" *per se*, are critical to support data science.

Data Science and Data Mining: Interchangeable words…

Data Science is a set of fundamental principles that guides the extraction of knowledge from the data.

Data Mining is the extraction of knowledge from data, via technologies, that incorporate these principles.

As a term, "data science" often is applied more broadly than the traditional use of "data mining", but data mining techniques provide some of the clearest illustrations of the principles of data science.

A Data Science Framework

Consider an Example: Customer Churn / Customer Retention in Telecommunication or Financial business

Other areas: direct marketing, online advertising, credit scoring, financial trading, fraud detection, product recommendation and so on.

Business Problem → Business Solution → Deployment

Phenomenon Understanding

Business Problem: Customer Churn

Q1. Whether interested in classifying fraction of customers towards churn?

Q2. Whether interested in computing propensity of churn for new group / individuals?

Q3. Whether interested in finding the reasons (root causes) for churn?

Q4. Whether interested in comparing the churn status with competitors?

Q5. Whether interested in forecasting churn rate over time?

Informative Attributes:

1. Demographic variables: age, gender, location, income, length with the company etc.
2. Usage variables: variables regarding network, talk time, sms, usage
3. QRC variables: variables regarding Query, Requests and Complaints
4. Variables regarding customers other connections
5. Connection variables: type of connection, and other features in the connection

Data Science : Fundamental Concepts

Concept-1: Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages **(problem of structured thinking about analytics)**.

Concept-2: From a large mass of data, information technology can be used to find informative descriptive attributes of entities of interest **(problem of correlation and feature selection/extraction)**.

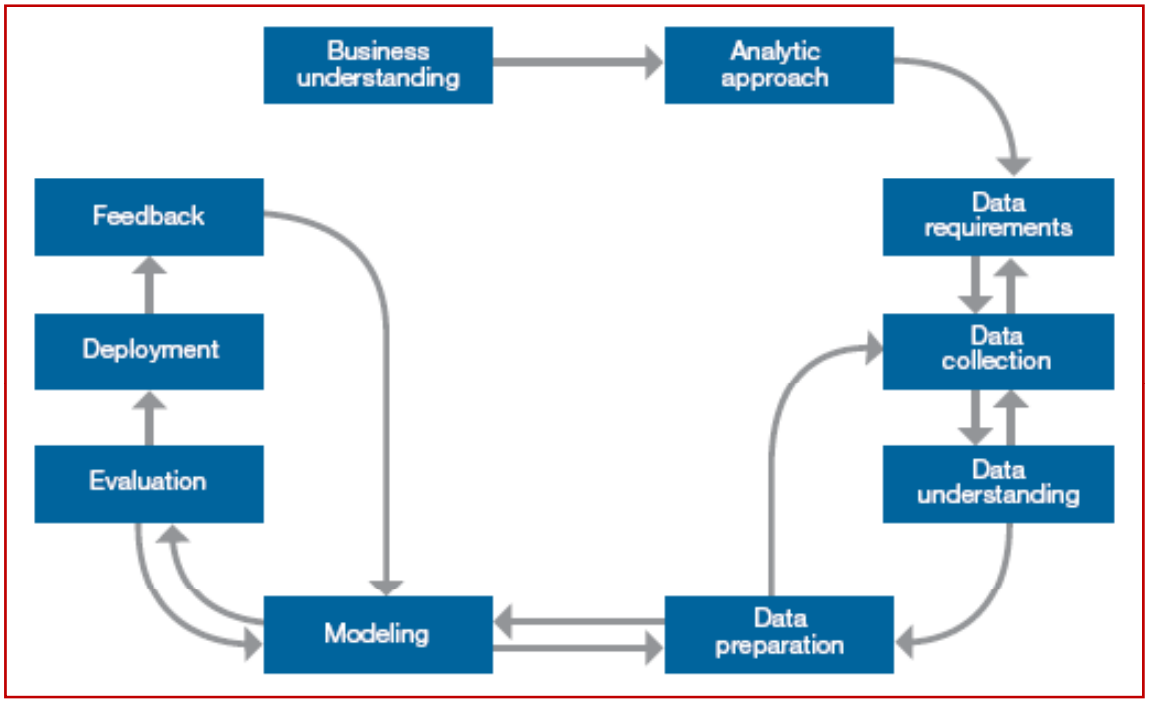
Concept-3: If you look too hard at a set of data, you will find something – but it might not generalize beyond the data you're looking at **(problem of overfitting)**.

Concept-4: Formulating data mining solutions and evaluating the results involves thinking carefully about the context in which they will be used **(problem of deployment policy)**.

Data Science : Other Issues

- Parametric (linear models) vs. Non-parametric approach (nearest neighbor models).
- Flexibility / Complexity vs. Interpretability (linear models to NN/ANN models)
- Explanatory Analytics (interpretable models) vs. Predictive Analytics (flexible models) and its combination (sequential/hybrid)
- Training, Validation and Test Errors
- Model Validation – Significance, Adequacy, Accuracy, Generalizability
- Overfitting (low bias for training data but high variance for test data)

Data Science : Foundational 10-stage Methodology



A **methodology** provides the data scientist with a framework for how to proceed with whatever methods, processes and heuristics will be used to obtain answers or results, independent of particular technologies or tools.

Data – BIG in nature: Big Data

Every day, we create 2.5 quintillion (10^{18}) bytes of data and even more…

According to the latest estimate, **328.77 million TB (~0.33 ZB)** of data are created each day.

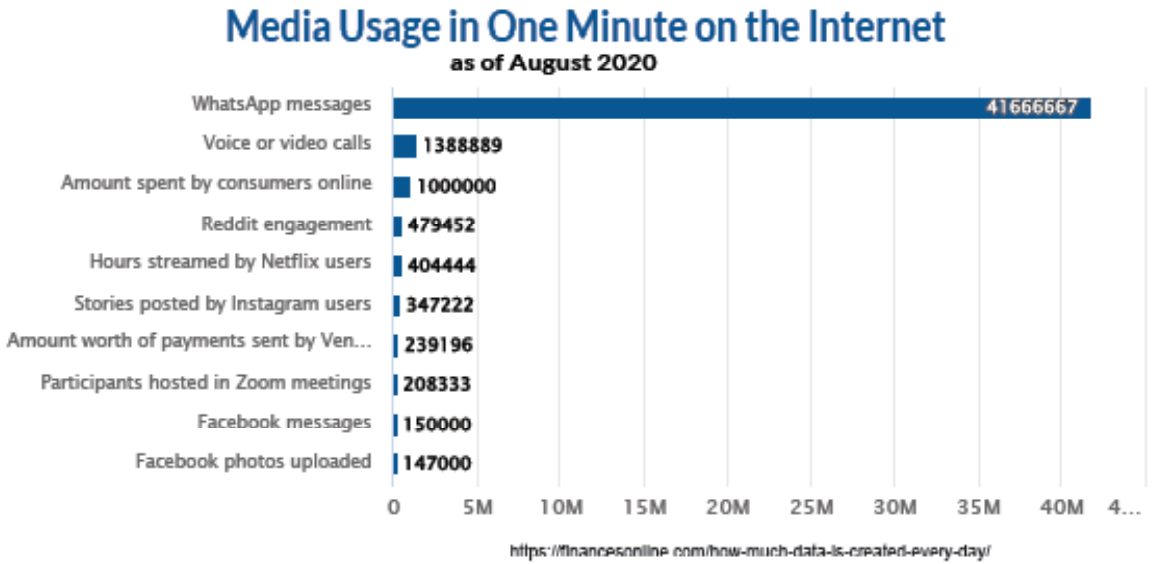
Source is everywhere: posts to social media sites, digital pictures and videos, sensors used to gather climate information, satellite images, and cell phone GPS signals to name a few.

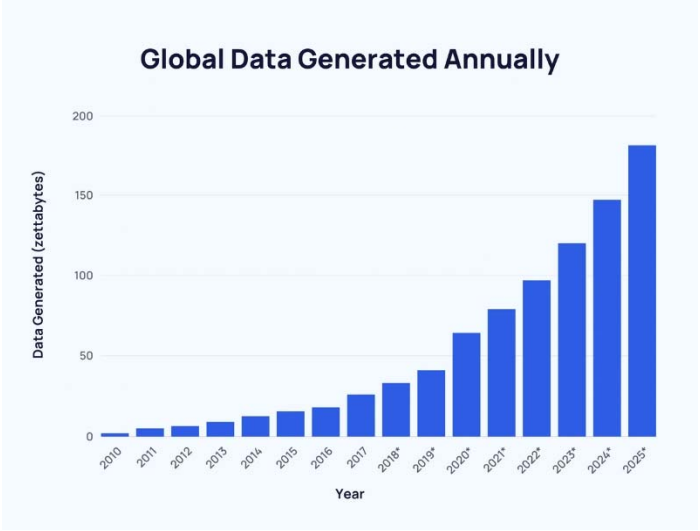
This data is "**big data**"

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Usage of Data

- By the end of 2020, **44 zettabytes** made up the entire digital universe
- In 2021, people created **2.5 quintillion bytes** of data every day.
- In 2022, **91%** of Instagram users **engaged with brand videos**.
- In 2022, users sent around **650 million Tweets** per day.
- In 2022, **333.2 billion emails** were sent every day.
- In 2023, 120 **zettabytes** data are generated.
- By 2025, **200+ zettabytes** of data will be in **cloud storage** around the globe.



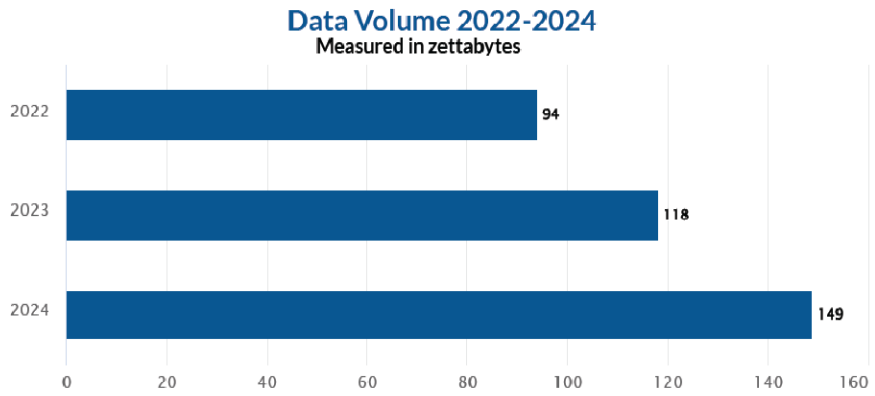


https://explodingtopics.com/blog/data-generated-per-day#how-much

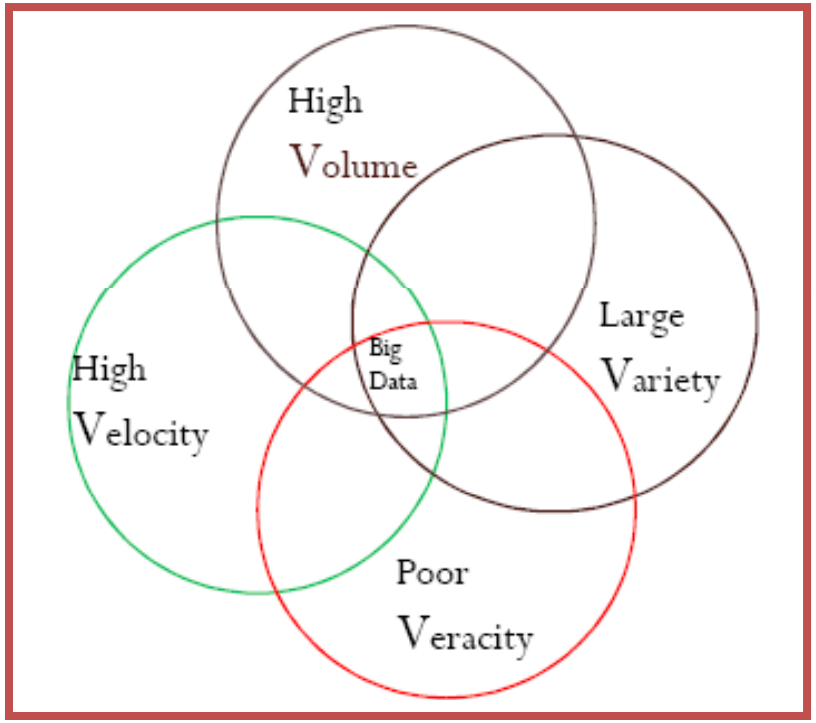
Data Statistics for the Future

In 2022, the world produced and consumed **97 ZB**. That is almost unimaginable data, which will only compound with the rise of the number of IoT connected devices.

- 463 ZB of data will be created every day by 2025. (Raconteur, 2020)
- Big Data is going to be worth $229.4 billion by 2025. (Strategic Tech Investor, 2021)
- The internet population in 2023 will be 66% of the world's total population. (Cisco, 2020)
- Each person will have an average of 1.6 networked mobile devices and connections come 2023. (Cisco, 2020)
- On average, every person will have 3.6 total networked devices and connections. (Cisco, 2020)
- The fixed broadband speed will be 110 Mbps up from 46 Mbps in 2018. (Cisco, 2020)
- Wi-Fi speed is predicted to go up to 92 Mbps. (Cisco, 2020)
- There will be more connected devices in 2023—three times the size of the global population. (RCR Wireless News, 2020)
- According to another prediction, there will be 43 billion IoT-connected devices. (McKinsey & Company, 2019)



Big Data – 4 ‘V’s



Big Data – Key Features

+ Increase in storage capabilities

+ Increase in processing power

+ Availability of data

+ Cheaper Hardware

+ Better Value-for-Money for Businesses

The **structured databases** are not well suited for storing and processing **Big data** for which *elements of computing* (such as storage, memory, processing and bandwidth) are becoming cheaper → making it more economical to conduct expensive data-intensive approaches to analyzing information.

Challenges in Big Data

- *High* **Volume,** *High* **Velocity,** *Large* **Variety,** *Poor* **Veracity**
- Non-random samples (*failure of inferential models*)
- Mixture data (*multiple populations*)
- Real time analysis on Streaming data (*speed vs. accuracy*)
- Scalability of algorithms [O(N^2), O(N^3)]
- Analysis with semi-structured data (*images+audio+text+video*)
- Protecting Privacy and Confidentiality (*ethical?; type-2 error?*)
- High Dimensional data [width too large as compared to depth]
- **Data Quality**

Big Data – Volume

- Enterprises are acquiring very large volume of data through variety of sources
- Some examples of use:
  - ✓Analysing Twitter data

    Terabytes of Tweets are created each day which can be used for improved product sentiment analysis
  - ✓Predict power consumption (a study in Norway)

    Convert billions of annual meter readings into better predict power consumption say every hour / minute.

Big Data – Velocity

- For time-sensitive processes such as catching fraud, preventing accidents, giving life saving medication etc. big data must be used as it streams into an enterprise in order to maximize its value.
- Some examples of use:
  - Scrutinize millions of credit card transactions each day to identify potential fraud
  - Analyze billions of daily call detail records in real time to predict customer churn faster
  - In ICU, analyze blood chemistry / ECG readings in real time to deliver life saving medication

Big Data – Variety

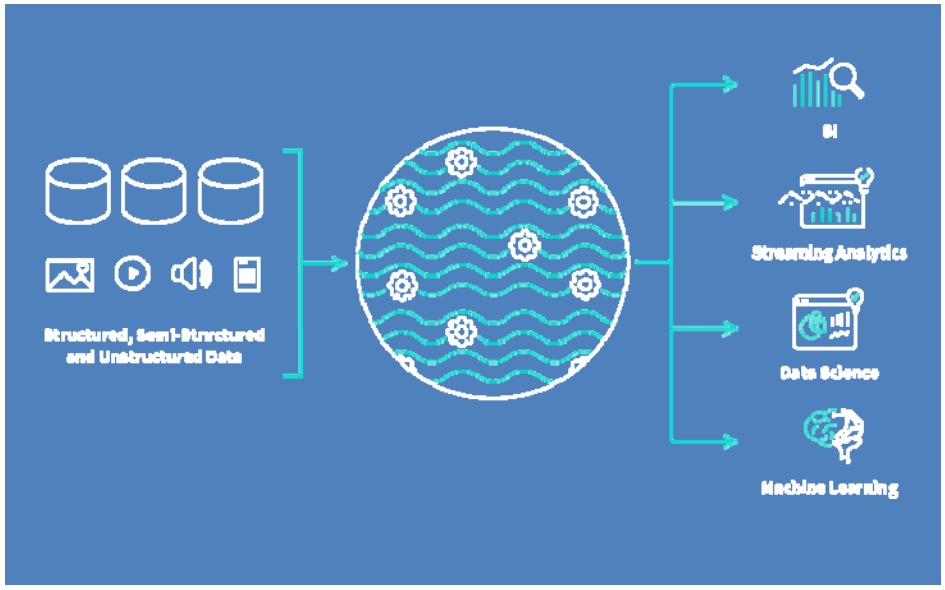
- Big data can be of any type - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.
- Some examples of use:
  - ✓ Monitor live video feeds from surveillance cameras to identify potential threats
  - ✓ Utilize image, audio, video and web information about a customer to give better product usage trainings, safety tips and recommendations.

Big Data – Veracity

- **Accuracy** is a big concern in Big Data. There is no easy way to segregate good data from bad.
- Some concerns (reliability/truthfulness):
  - ✓ Among thousands of reviews of hotels which ones are authentic and which ones are not?
  - ✓ How to find out the TRUTH from thousands of product reviews
  - ✓ How to identify a rumor from a informed communication?

Data Lake vs. Data Warehouse

A **data lake** is a repository that stores all of your organization's data — both structured and unstructured. Think of it as a massive storage pool for data in its natural, raw state (like a lake). A **data lake architecture** can handle the huge volumes of data that most organizations produce without the need to structure it first. Data stored in a data lake can be used to build data pipelines to make it available for data analytics tools to find insights that inform key business decisions.



Data Lake vs. Data Warehouse

A **data warehouse** is a repository for business data. However, unlike a data lake, only highly structured and unified data lives in a data warehouse to support specific business intelligence and analytics needs.

The large majority of organizations primarily use data warehouses and the clear trend is toward cloud data warehouses. Data lakes are typically used by data scientists for machine learning and exploration of flat files.

Still, many organizations use both a data lake and a data warehouse to cover the spectrum of their data storage needs.

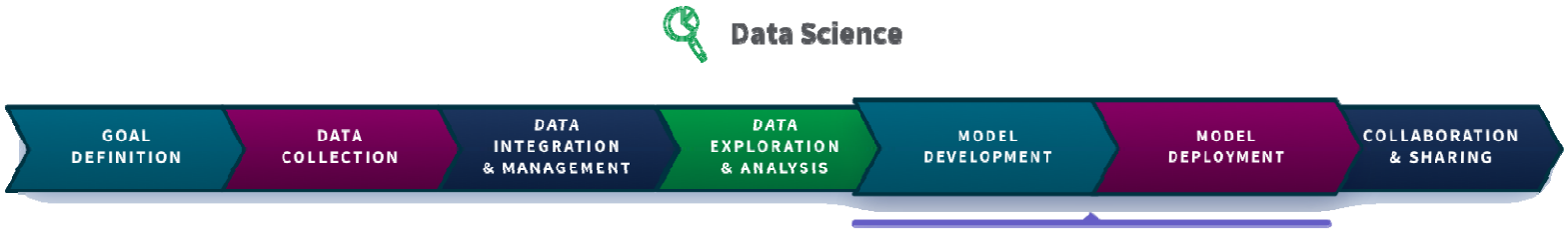
Data Lake vs. Data Warehouse

	Data Lake	Data Warehouse
1. Data Storage	A data lake contains all an organization's data in a raw, unstructured form, and can store the data indefinitely — for immediate or future use.	A data warehouse contains structured data that has been cleaned and processed, ready for strategic analysis based on predefined business needs.
2. Users	Data from a data lake — with its large volume of unstructured data — is typically used by data scientists and engineers who prefer to study data in its raw form to gain new, unique business insights.	Data from a data warehouse is typically accessed by managers and business-end users looking to gain insights from business KPIs, as the data has already been structured to provide answers to pre-determined questions for analysis.
3. Analysis	Predictive analytics, machine learning, data visualization, BI, big data analytics.	Data visualization, BI, data analytics.
4. Schema	Schema is defined after the data is stored in a data lake vs data warehouse, making the process of capturing and storing the data faster.	In a data warehouse, the schema is defined before the data is stored. This lengthens the time it takes to process the data, but once complete, the data is at the ready for consistent, confident use across the organization.
5. Processing	ELT (Extract, Load, Transform). In this process, the data is extracted from its source for storage in the data lake, and structured only when needed.	ETL (Extract, Transform, Load). In this process, data is extracted from its source(s), scrubbed, then structured so it's ready for business-end analysis.
6. Cost	Storage costs are fairly inexpensive in a data lake vs data warehouse. Data lakes are also less time-consuming to manage, which reduces operational costs.	Data warehouses cost more than data lakes, and also require more time to manage, resulting in additional operational costs.

https://www.qlik.com/us/data-lake/data-lake-vs-data-warehouse

Data Science vs. Data Analytics

Data Science is the application of tools, processes, and techniques towards combining, preparing and examining large datasets and then using programming, statistics, machine learning and algorithms to design and build new data models.



Data analytics is the use of tools and processes to combine, prepare and analyze datasets to identify patterns and develop actionable insights.

DATA QUALITY

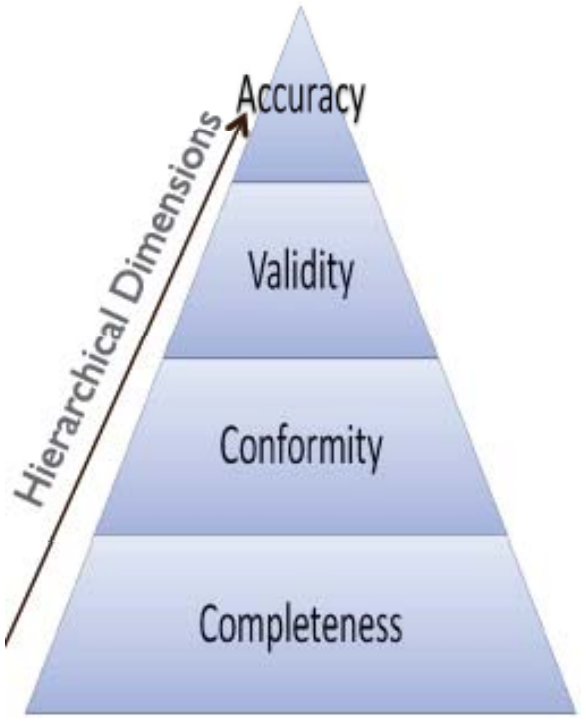
What is Data Quality ?

Data quality represents the **reliability** and **effectiveness** of data. It means the data can be used in various purposes like operations, decision making, planning etc.

The **objective** of Data Quality Assessment System is to

- ✓Identify the various components of data quality and the measurable and controllable quality characteristics.
- ✓Provide methods to measure and monitor these characteristics.
- ✓Develop systems to ensure that the quality of data achieves the expected performance levels.
- ✓Ensure smooth capturing of data under any dynamic set-up upto the granular most level.
- ✓Ensure the efficiency of system in such a manner so that the impact of assessment can be made easily even under any changed circumstances.

Dimensions of Data Quality

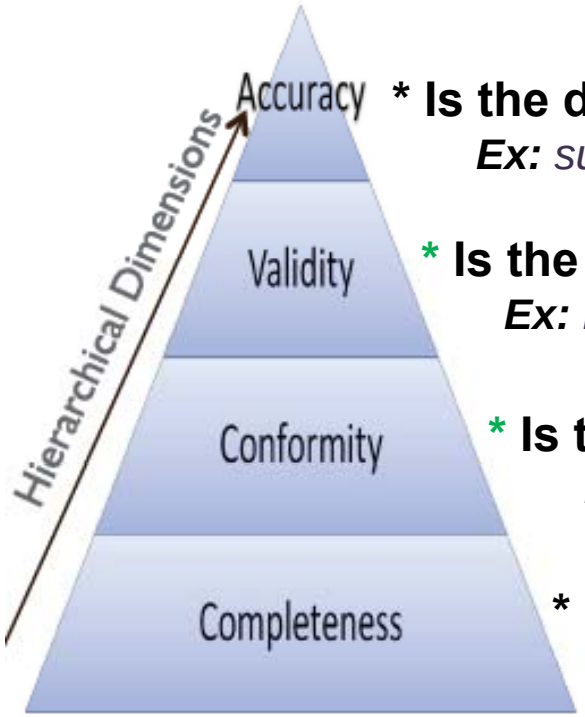


Conventional View
Data Quality = Accuracy of data
New View
Data Quality = Beyond accuracy

Data Quality : Challenges under Big Data

DATA must be used in various purposes like Operations, Decision Making, Planning etc. with lot more confidence (**reliability** and **effectiveness**).

MAJOR COMPONENTS:



* **Is the data verifiably correct?**
  Ex: *subject matter expert, customer validation, hard-copy record*

* **Is the data from a set of expected values?**
  Ex: *Pin Codes must be from list for all Indian codes*

* **Is the data in the correct format?**
  Ex: *A customer name must be in letters and not numbers*

* **Do we have all of the data elements?**
  Ex: *A Social Security is needed to complete a loan application.*

Dimensions of Data Quality

Completeness: Do we have all of the data elements?

- The measure of the presence of core source data elements that, exclusive of derived fields, must be present in order to complete a given business process.
- ***Ex:*** *A Social Security is needed to complete a loan application.*

Conformity: Is the data in the correct format?

- The measure of a data element's adherence to required formats, data types, field lengths.
- ***Ex:*** *A customer name must be in letters and not numbers and dates must be in year/month/day format.*

Validity: Is the data from a set of expected values?

- The data corresponds to reference tables, lists of values from golden sources documented in metadata, value ranges.
- ***Ex:*** *Zip Codes must be from list for all Indian codes for Indian customers.*

Accuracy: Is the data verifiably correct?

- The measure of whether the value of a given data element is correct and reflects the real world as viewed by a valid real world source.
- Business Rules are needed to confirm Accuracy
- ***Ex:*** *Subject matter expert, customer validation, hard-copy record*

More Components of Data Quality

Definition—each data element should have clear meaning and acceptable values.

Completeness – All required data must be present in a record.

Accuracy – Closeness of result of observations to the true values or values accepted as true.

Timeliness – most frequent data quality dimensions that must be managed. Data may be accurate and complete at the time of entry, but the real world construct make the data change overtime, thus the same data may be accurate and incomplete

Accessibility – should be available to authorized persons when and where needed.

Granularity – the extent to which a large entity is sub-divided

Confidentiality – depends on accessibility and security on data management at different levels.

Reliability – consistent and representative across volume.

Legibility – should be readable at least

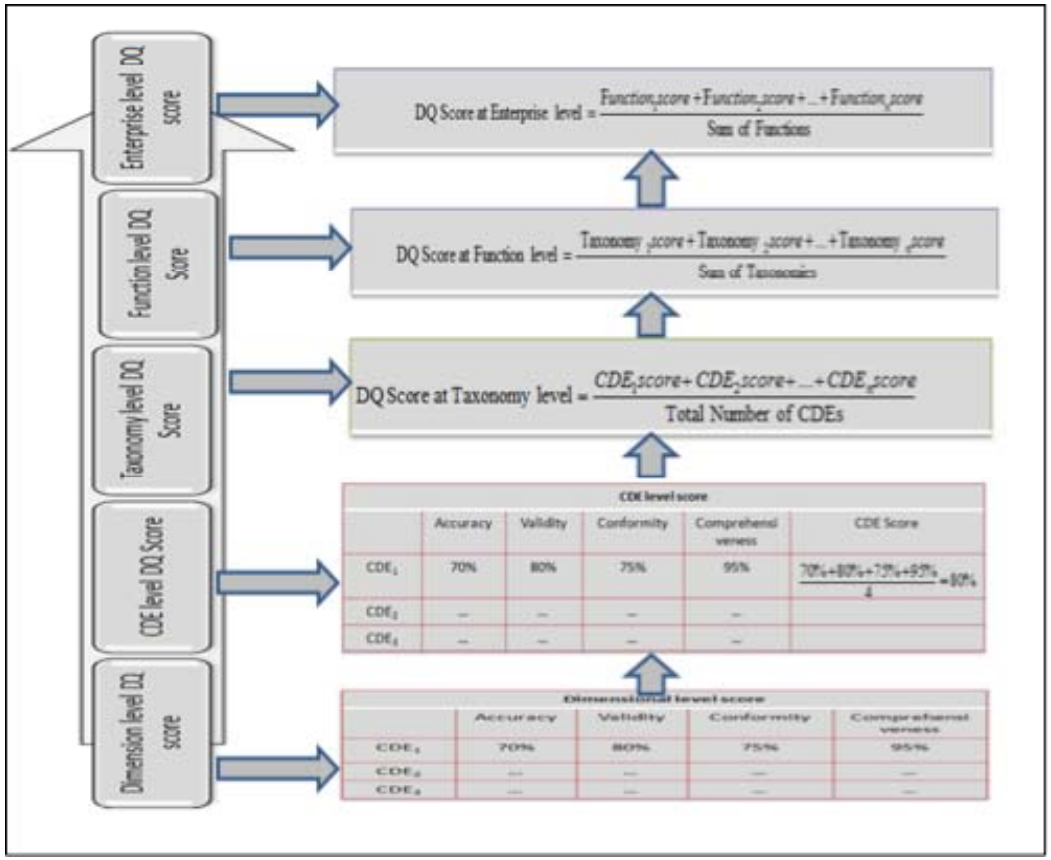
Usefulness – pertinent and useful.

Are these DATA?

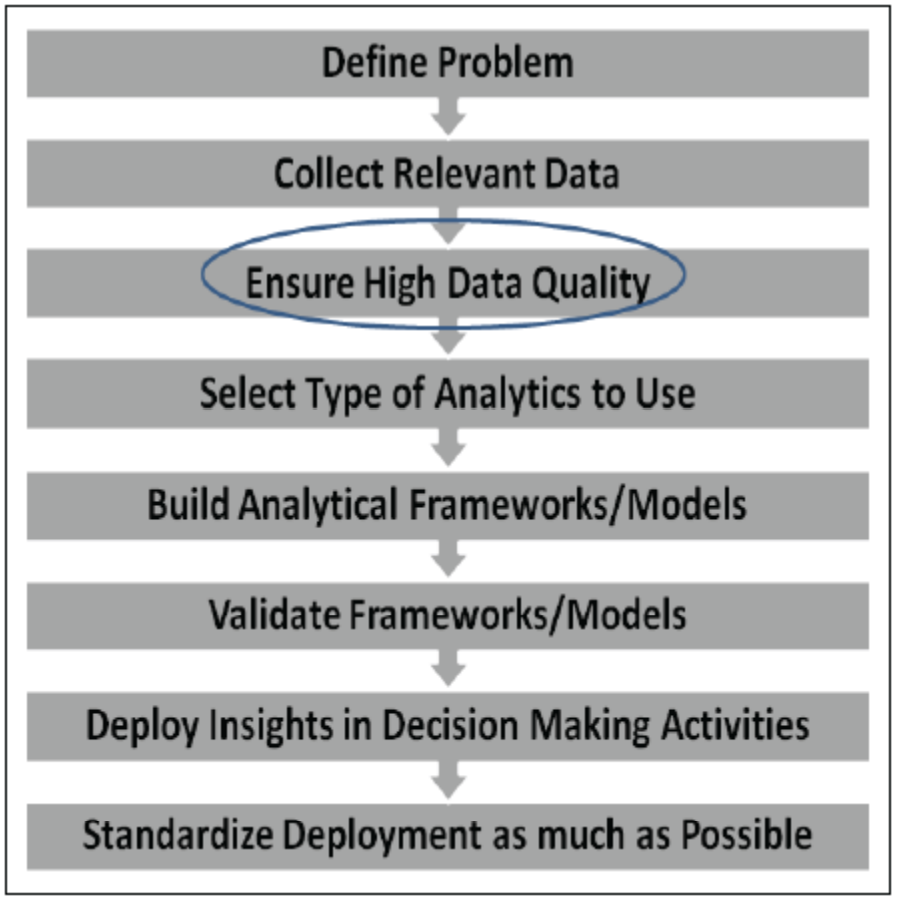
Is it a "Quality" Data ?

Measuring Data Quality: DQ Score

- DQ is measured at Critical Data Element (CDE) level based on dimensions such as completeness, conformance, validity and accuracy.
- A CDE is defined as the data that is “critical to success” in a specific business area (line of business, shared service or function), or the data “required to get the job done”.
- CDE level scores can be rolled up to enterprise level to assess overall DQ score



Execution of Analytics



High-Quality Data + Reliable Analytics ≡ Timely and Effective Decisions

List of Books

WILEY

Home / Engineering & Materials Science / Industrial Engineering / Quality Control

Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality

Rajesh Jugulum

ISBN: 978-1-118-34232-9

304 pages
February 2014

Carlo Batini and Monica Scannapieca: *Data Quality: Concepts, Methodologies and Techniques*

Redman, T. C. 2001. *Data Quality. The Field Guide. Boston: Digital Press*