Data Modeling

Data Modeling

- Data Modeling
  - Statistical Data Modeling,
  - Computational Data Modeling,
- Bonferroni's principle

Data Modeling

- Data Modeling is formulating data in a particular structure so that it can help in easy reporting in future.
- It helps in analyzing data that helps in meeting business requirements.
- It is a set of activity and techniques involved in understanding the structure of an organization and also propose solution that enables the organization to achieve its objectives.
- The data model will normally consist of entity types, attributes, relationships, integrity rules, and the definitions of those objects.
- It is used as the start point for interface or database design.

Importance of Data Model

- Data model portrays a better understanding of business requirements.
- It helps in creation of a robust design and easy to rework.
- A qualified data model helps in providing better consistency across all projects of an enterprise.
- It improves data quality.
- Less data movements .
- Less movement implies less maintenance.
- Re-use of data model saves the entire efforts put in to design an existing model.

Statistical Data Modeling

- A representative smaller version of the data collected.
- It should summarize the data as closely as possible (be 'a good fit').
- We cannot measure a population, so the best we can do is make generalizations from a sample to a population using a representative summary, i.e. a statistical model.
- Example, we use statistical models every day without realizing it, the simplest summary model for numerical data is a mean, while for categorical data is a proportion.

Computational Modeling

- Computational modeling is the use of computers to simulate and study the behavior of complex systems using mathematics, physics and computer science.
- A computational model takes the form of an **algorithm.**
- A computational model contains numerous variables that characterize the system being studied.
- Simulation is done by adjusting each of these variables alone or in combination and observing how the changes affect the outcomes
- The results of model simulations help researchers make predictions about what will happen in the real system that is being studied in response to changing conditions.
- Modeling can expedite research by allowing scientists to conduct thousands of simulated experiments by computer in order to identify the actual physical experiments that are most likely to help the researcher find the solution to the problem being studied.

Bonferroni’s Principle

Bonferroni's Principle is an informal presentation of a statistical theorem that states if your method of finding significant items returns significantly more items that you would expect in the actual population, you can assume most of the items you find with it are bogus. This essentially means that an algorithm or method we think is useful for finding a particular set of data actually returns more false positives as it returns larger portions of the data than should be within that category.

Different Types of Data

- Structured Data
- Unstructured Data
- Natural Language Data
- Machine-Generated Data
- Graph-based data
- Audio, Video and images data
- Steaming Data

Structured Data

- It resides in a fixed field within a record.
- Stored in tables in databases or excel files.
- SQL is used to manage and query data that resides in databases.

Unstructured Data

- It is not easy to fit unstructured data into a fixed field within a record.
- Example: emails, docs etc

Natural Language Data

- It is special type of unstructured data
- Its challenging to process as it requires knowledge of specific data science techniques and linguistics
- Had success in NER, topic recognition, summarization, text completion, SA but models trained on one-domain don’t generalize to other domain.
- Not able to decipher the meaning of every piece of text.

Machine Generated data

- This is the information created by a computer, process, application or other machine without human intervention.
- Machine generated data is becoming a major data resource .
- Eg. Web server logs, call detail records, network event logs, and telemetry.

Graph Based or Network data

- Graph data focuses on the relationship or adjacency of objects.
- The graph structure uses nodes, egdes and properties to represent and store graphical data.
- Graph based data is a natural way to represent social networks, and its structure allows to calculate specific metrics such as influence of a person and the shortest path between two people.

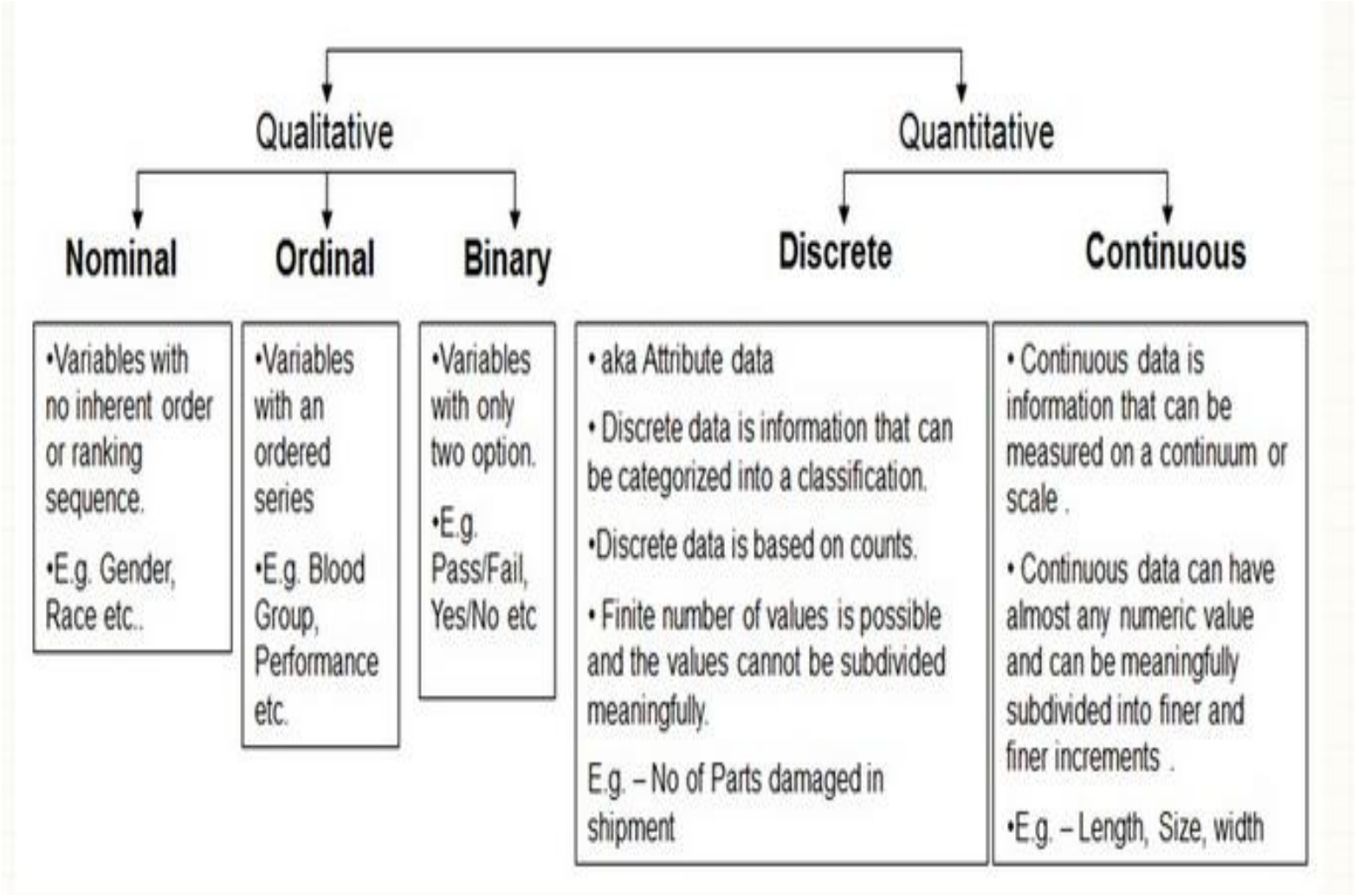
Audio, Image and Video data

- Audio, image and video are data types that pose specific challenges to a data scientist.
- Recognizing objects in pictures turn out to be challenging for computers.
- High speed cameras at stadiums capture ball and athlete movements to calculate in real time, the path taken by a defender relative to two baselines.

Streaming data

- The data flows into the system when an event happens instead of being loaded into a data store in a batch.
- Example, “What’s trending? on twitter, live sports or music events.

Data types


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Data Preprocessing