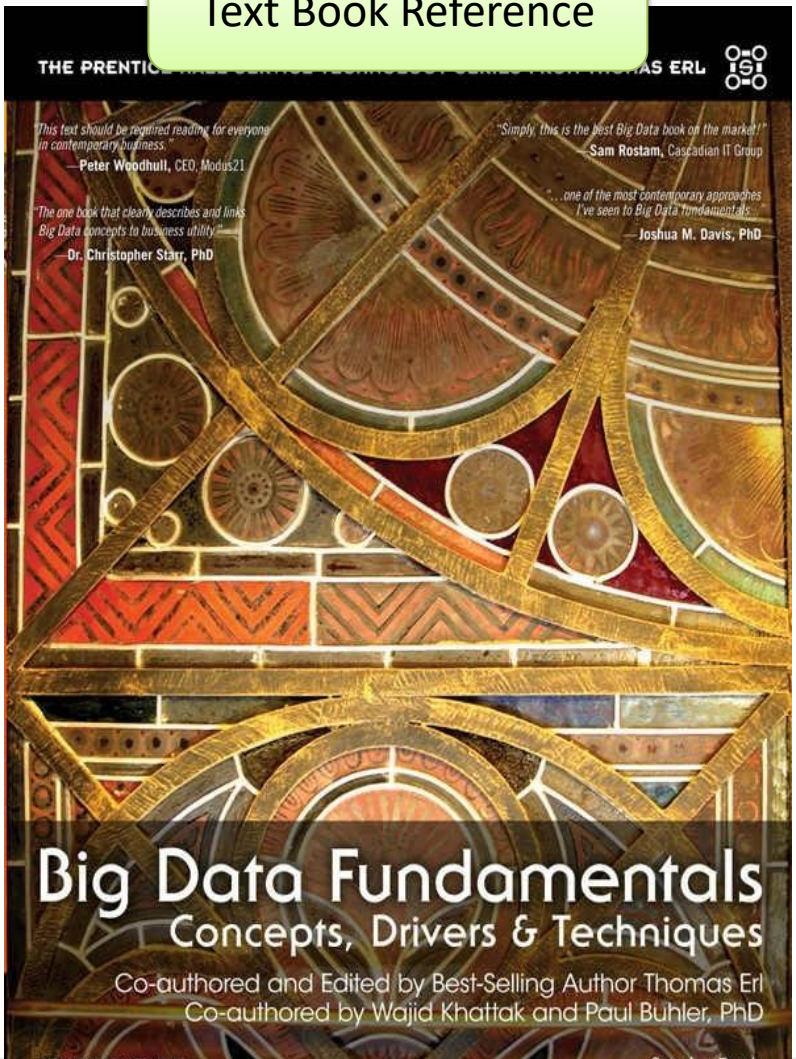
Big Data Fundamentals

Adityo Hidayat, S.Kom, M.B.A., CISA

References

Text Book Reference



Actual Case Studies

transjakarta

- DB Size **282 GB**
- DB Size **+14G /month**
- Avg Trx **287rb /day**

Indonesia Stock Exchange
Bursa Efek Indonesia

- DB Size **4,8GB compressed**
- DB Size **16GB uncompressed**
- Avg Trx **255rb /day**

Actual Big Data Experience

Course Outline

Concepts

Analysis & Analytics

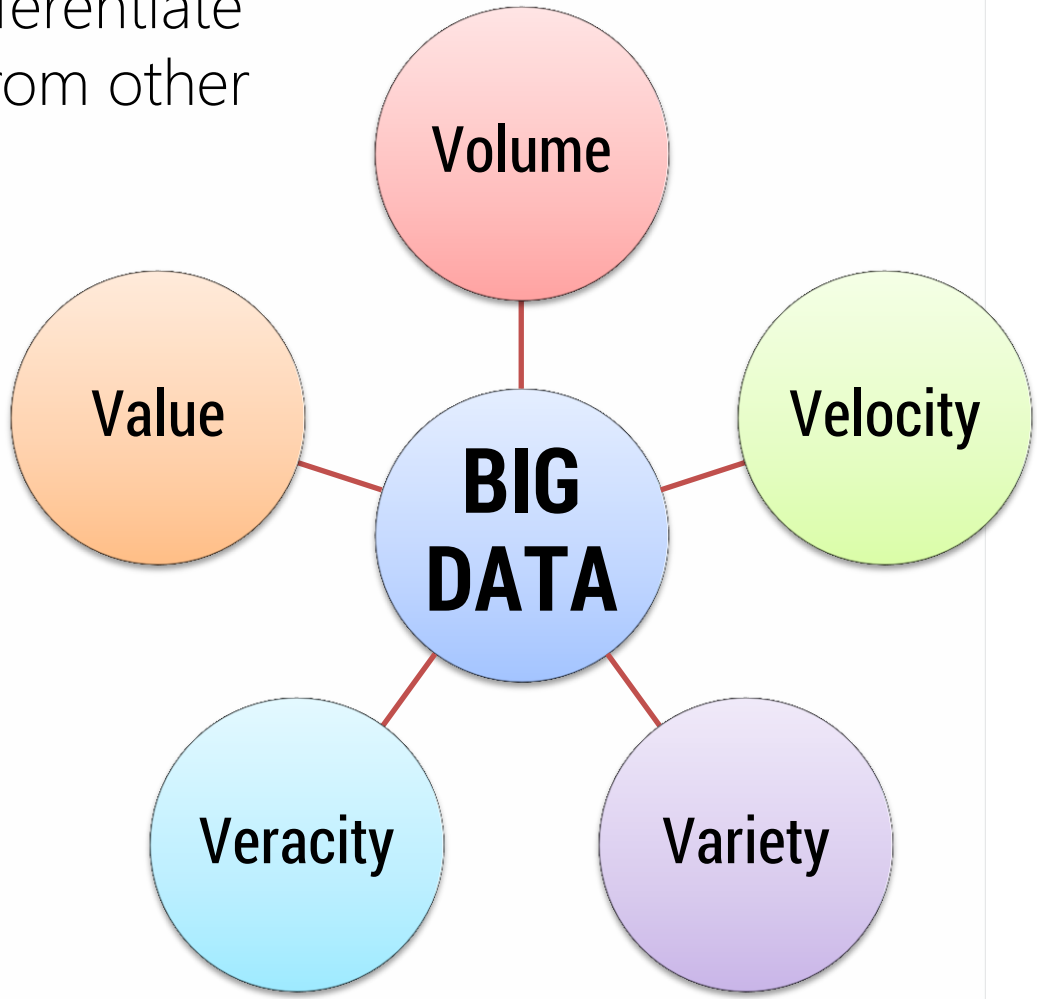
Big Data Technology

Big Data Processing

Analysis Methods

Big Data Characteristics

This characterstics helps differentiate data categorized as “BIG” from other forms of data.



Volume

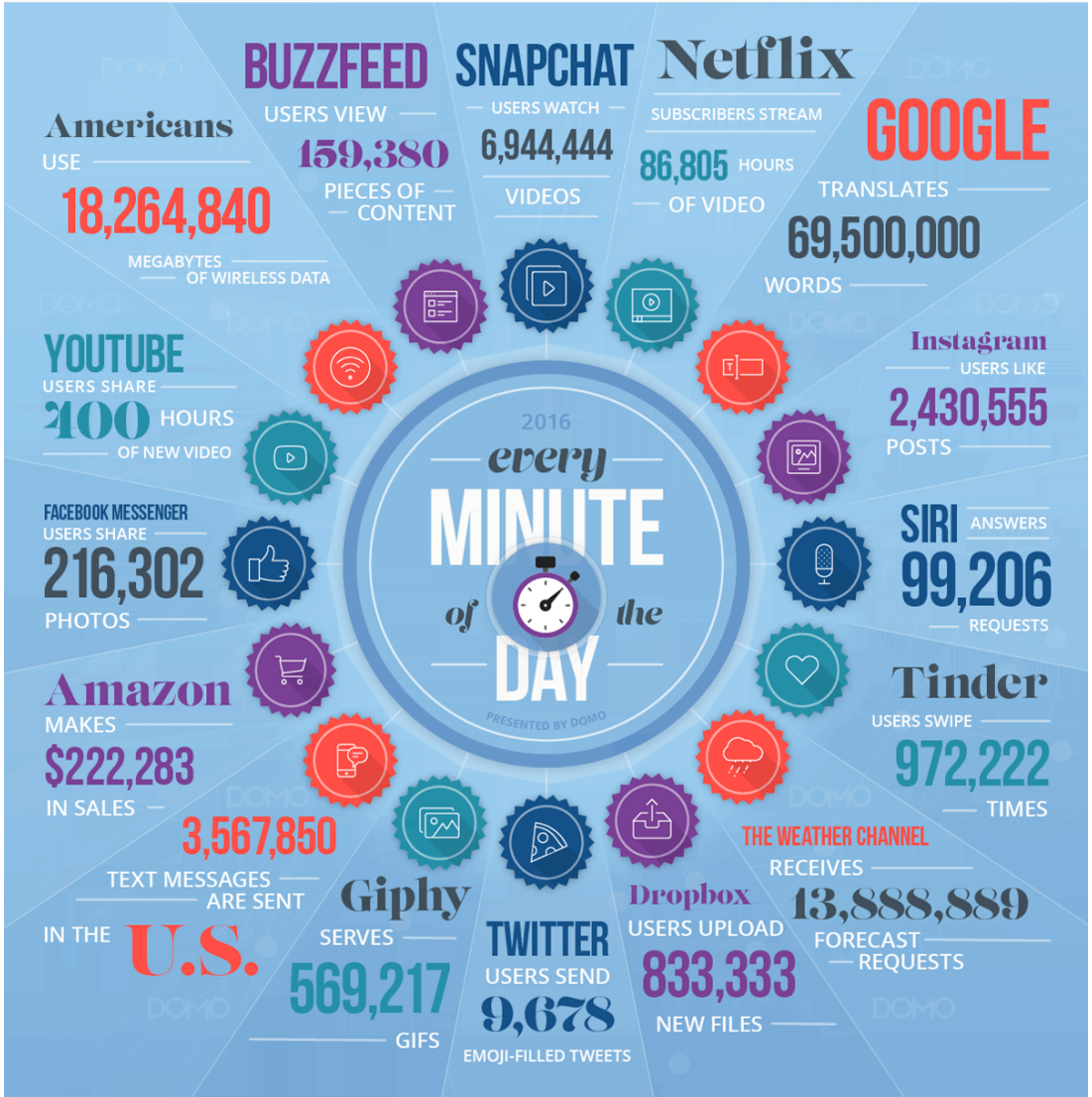
- The anticipated volume of data that is processed by Big Data solutions is substantial and ever-growing.
- Typical data sources that are responsible for generating high data volumes can include:
  - online transactions, such as point-of-sale and banking
  - scientific and research experiments
  - Large Millimeter/Submillimeter Array telescope
  - sensors, such as GPS sensors, RFIDs, smart meters and telematics
  - social media, such as Facebook and Twitter

Velocity

- In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time.
- Coping with the fast inflow of data requires the enterprise to design highly elastic and available data processing solutions and corresponding data storage capabilities.
- Depending on the data source, velocity may not always be high. For example, MRI scan images are not generated as frequently as log entries from a high-traffic webserver.

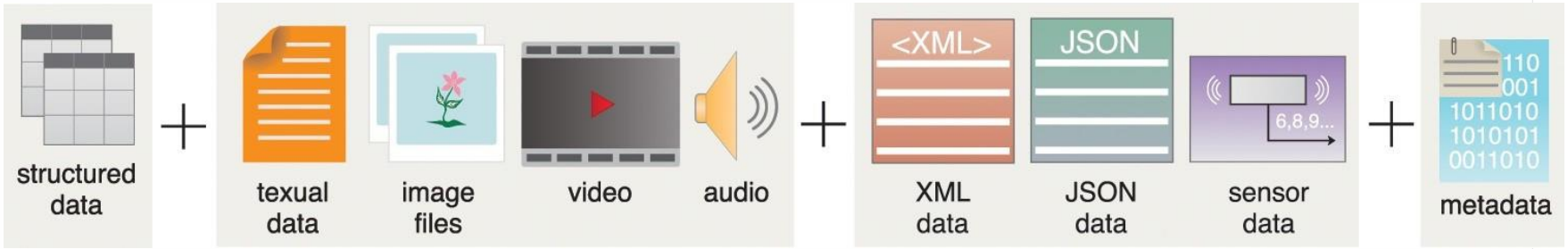
DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.



Variety

- Data variety refers to the multiple formats and types of data that need to be supported. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.

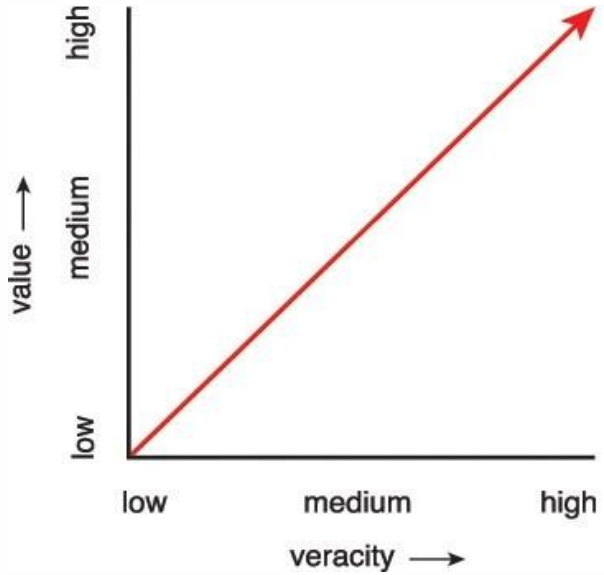


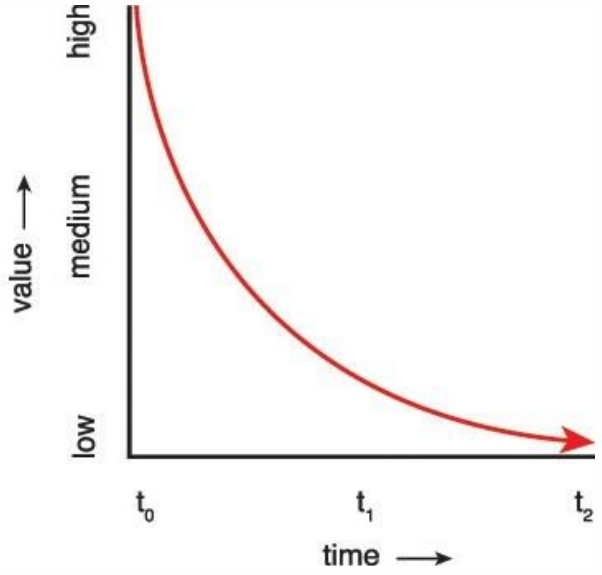
Veracity

- Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.
- Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information.
- Data that is acquired in a controlled manner, for example via online customer registrations, usually contains less noise than data acquired via uncontrolled sources, such as blog postings.

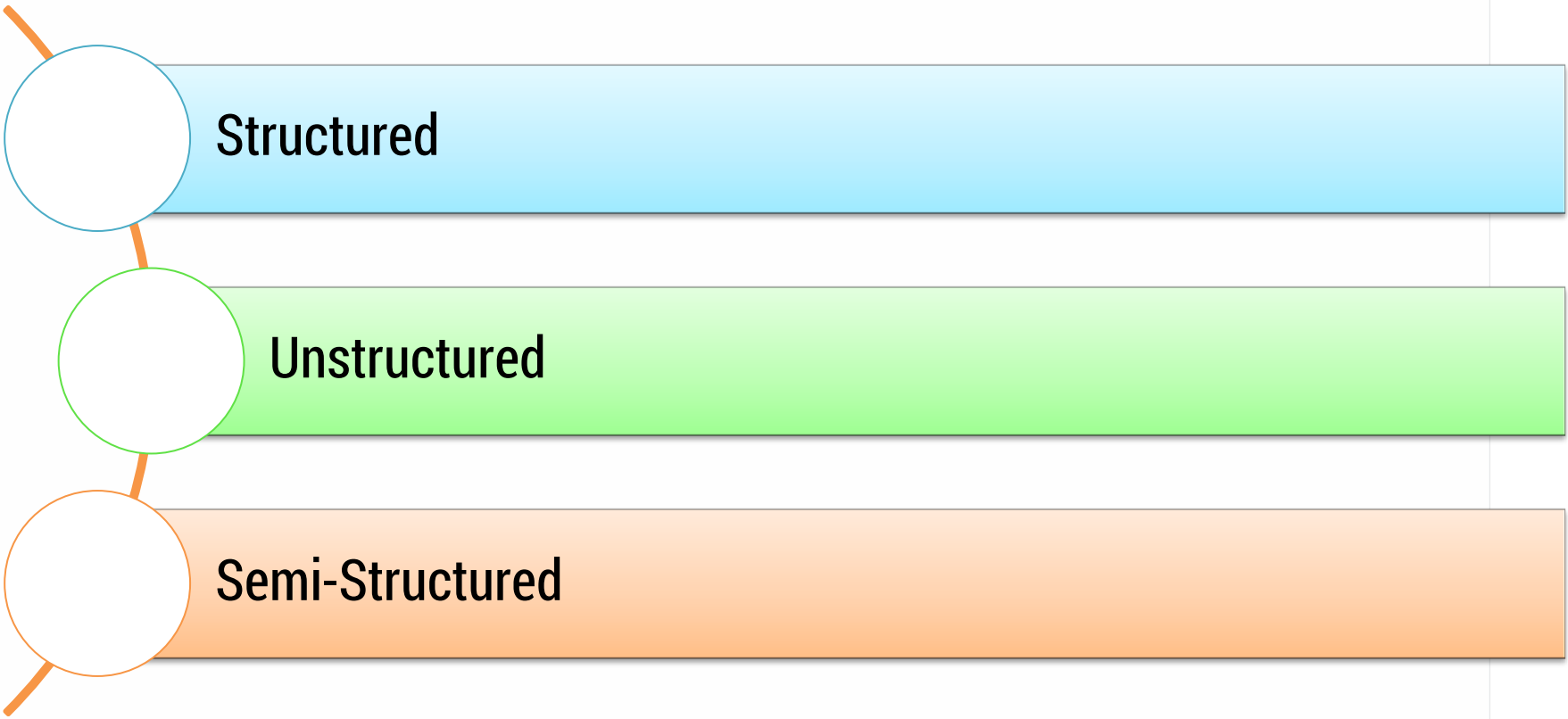
Value

- Value is defined as the usefulness of data for an enterprise.
- The value characteristic is intuitively related to the veracity. The higher the data fidelity, the more value it holds.
- Value is also dependent on how long data processing takes because analytics results have a shelf-life; for example, a 20 minute delayed stock quote has no value for making a trade compared to a quote that is 20 milliseconds old.





Types of Data Sources

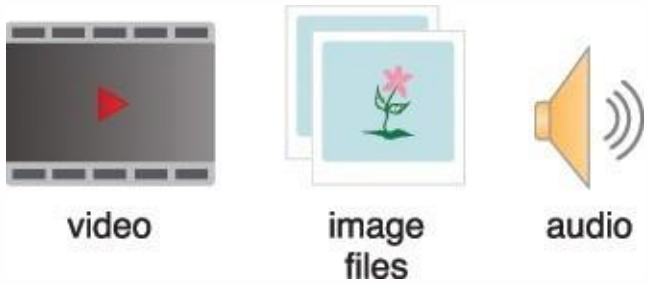


Structured Data

- Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database.
- Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems.
- Examples of this type of data include banking transactions, invoices, and customer records.

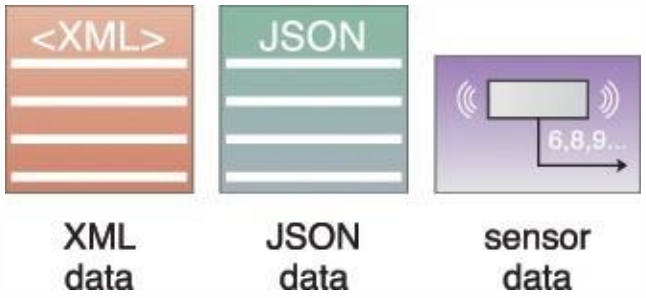
Unstructured Data

- Data that does not conform to a data model or data schema is known as unstructured data.
- Unstructured data cannot be directly processed or queried using SQL. If it is required to be stored within a relational database, it is stored in a table as a Binary Large Object (BLOB).
- Alternatively, a Not-only SQL (NoSQL) database is a non-relational database that can be used to store unstructured data alongside structured data.



Semi-Structured Data

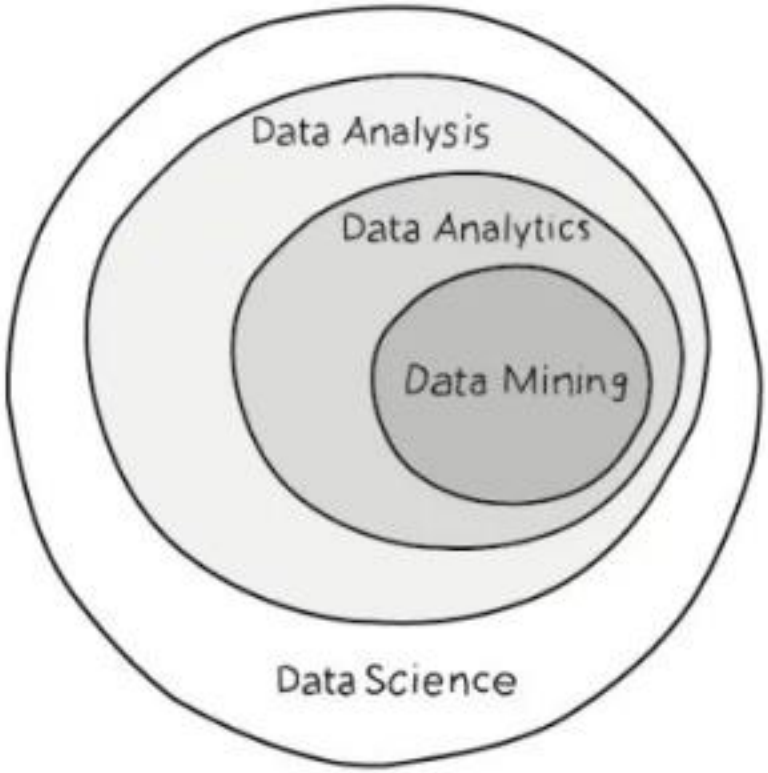
- Semi-structured data has a defined level of structure and consistency, but is not relational in nature. This kind of data is commonly stored in files that contain text.
- XML and JSON files are common forms of semi-structured data. Due to the textual nature of this data and its conformance to some level of structure, it is more easily processed than unstructured data.



Metadata

- Metadata provides information about a dataset's characteristics and structure. This type of data is mostly machine-generated and can be appended to data.
- The tracking of metadata is crucial to Big Data processing, storage and analysis because it provides information about the pedigree of the data and its provenance during processing.
- Examples of metadata include:
  - XML tags providing the author and creation date of a document
  - attributes providing the file size and resolution of a digital photograph
- Big Data solutions **rely on metadata**, particularly when processing **semi-structured** and **unstructured** data.

Analysis & Analytics



https://www.quora.com/What-is-the-difference-between-Analytics-and-analysis-data-analytics

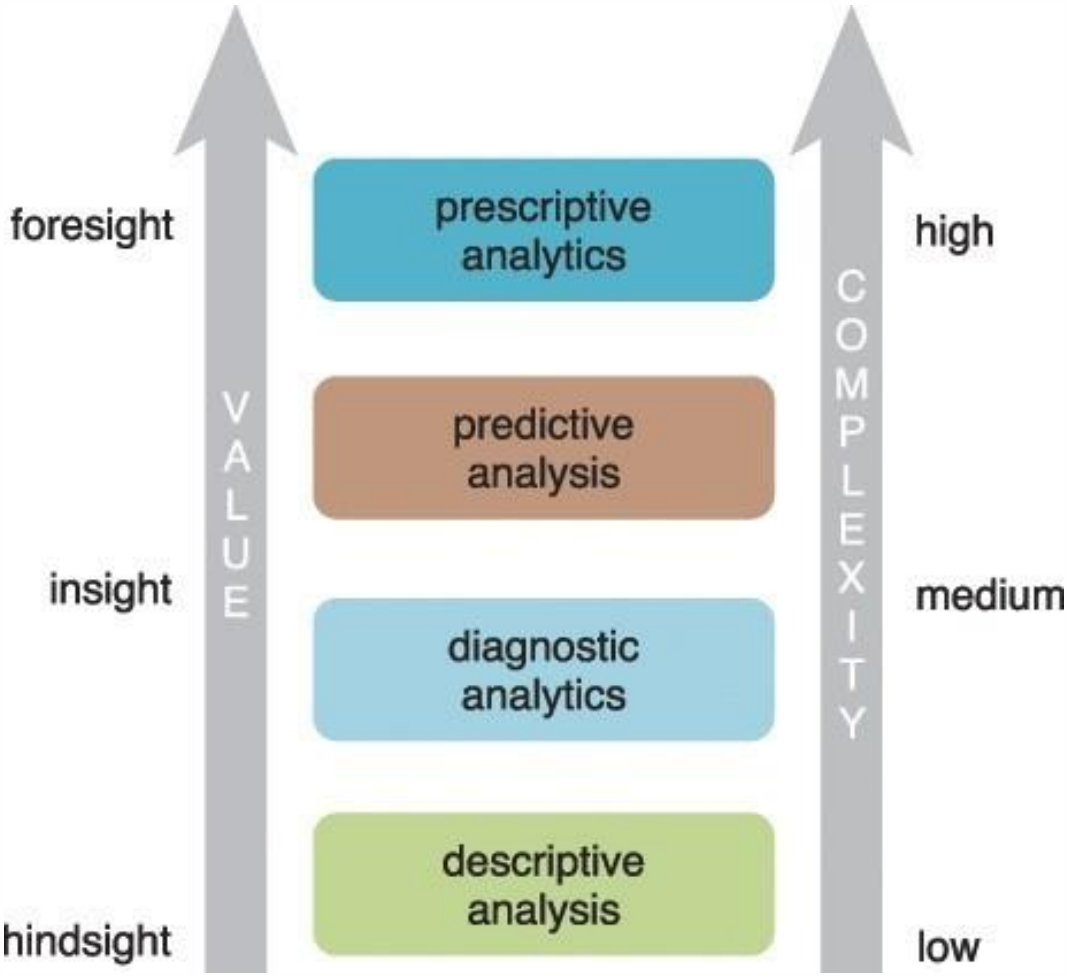
Data Analysis

- Data analysis is the **process of examining data to find facts, relationships, patterns, insights and/or trends**.
- Example: Analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the daily temperature.
- The results of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information.

Data Analytics

- Data analytics is a discipline that includes the **management of the complete data lifecycle**, which encompasses **collecting, cleansing, organizing, storing, analyzing and governing** data.
- Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone.
- In Big Data environments, data analytics has developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different sources.

Categories of Data Analytics



Descriptive Analytics

- Descriptive analytics are carried out to answer questions about events that have already occurred.

What was the sales volume over the past 12 months?

What is the number of support calls received as categorized by severity and geographic location?

What is the monthly commission earned by each sales agent?

Diagnostic Analytics

- Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
- Diagnostic analytics usually require collecting data from multiple sources and storing it in a structure that lends itself to performing **drill-down** and roll-up analysis

Why were Q2 sales less than Q1 sales?

Why have there been more support calls originating from the Eastern region than from the Western region?

Why was there an increase in patient re-admission rates over the past three months?

Predictive Analytics

- Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future.
- Predictions are made based on patterns, trends and exceptions found in historical and current data.

What are the chances that a customer will default on a loan
if they have missed a monthly payment?

If a customer has purchased Products A and B,
what are the chances that they will also purchase Product C?

What will be the patient survival rate
if Drug B is administered instead of Drug A?

Prescriptive Analytics

- Prescriptive analytics build upon the results of predictive analytics by **prescribing actions that should be taken**.
- Various outcomes are calculated, and the best course of action for each outcome is suggested. The approach shifts from explanatory to advisory and can include the simulation of various scenarios.

When is the best time to trade a particular stock?

Among three drugs, which one provides the best results?

Analysis Methods

Quantitative Analysis

Data Mining

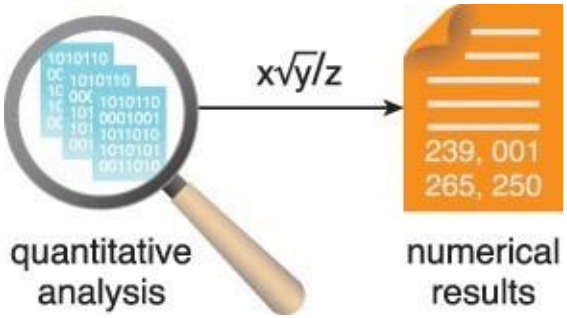
Statistical Analysis

Machine Learning

Semantic Analysis

Quantitative Analysis

- Data analysis technique that focuses on **quantifying** the patterns and correlations found in the data.
- Produces numerical results.
- Analysis results are absolute in nature and can therefore be used for numerical comparisons.



Data Mining

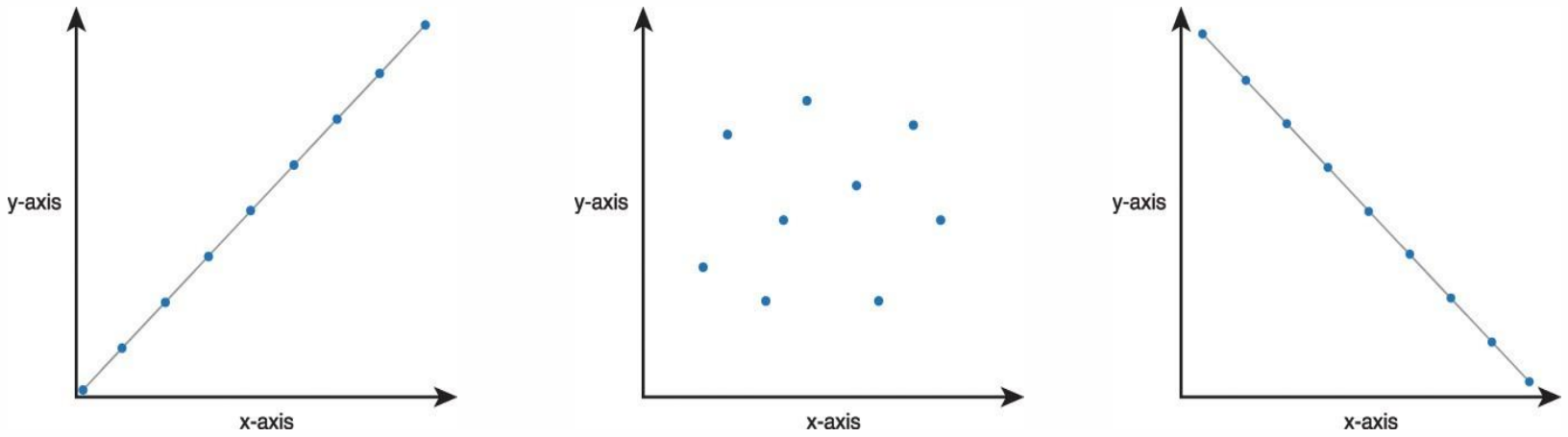
- Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets.
- In relation to Big Data analysis, data mining generally refers to **automated, software-based techniques that sift through massive datasets to identify patterns and trends**.
- Involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns. Data mining forms the basis for **predictive analytics** and **business intelligence** (BI).

Statistical Analysis

- Statistical analysis uses statistical methods based on mathematical formulas as a means for analyzing data.
- This type of analysis is commonly used to describe datasets via summarization, such as providing the mean, median, or mode of statistics associated with the dataset. It can also be used to infer patterns and relationships within the dataset, such as **regression** and **correlation**.

Statistical Analysis - Correlation

- Used to determine whether two variables are related to each other. Expressed as a coefficient between –1 to +1



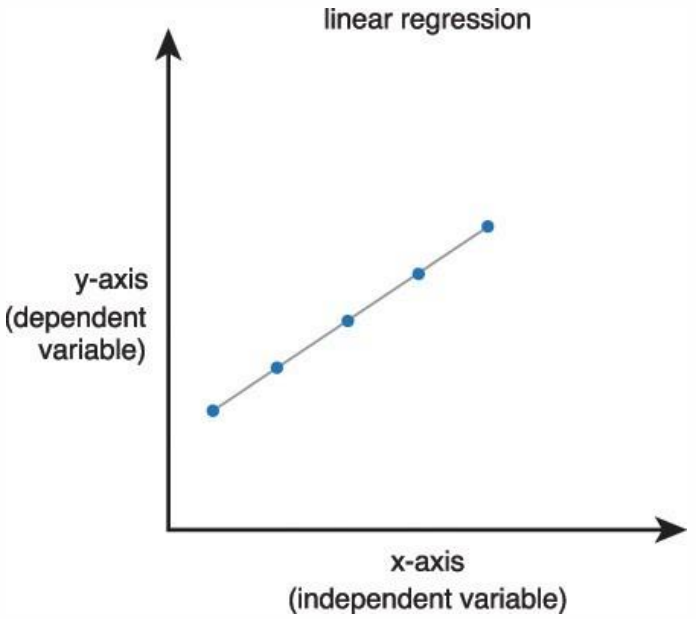
Does distance from the sea affect the temperature of a city?

Do students who perform well at elementary school perform equally well at high school?

To what extent is obesity linked with overeating?

Statistical Analysis - Regression

- Correlation can first be applied to discover if a relationship exists.
- Regression can then be applied to **predict the values** of the dependent variable



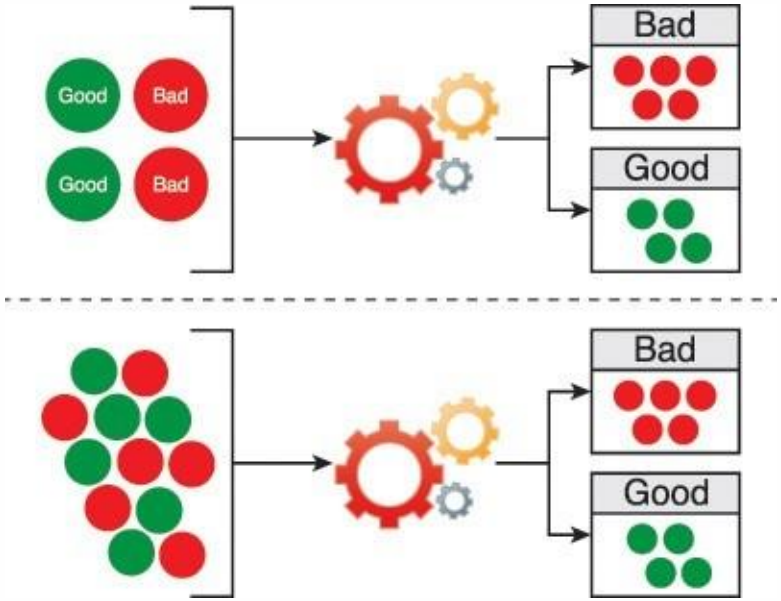
What will be the temperature of a city that is 250 miles away from the sea?

What will be the grades of a student studying at a high school based on their primary school grades?

What are the chances that a person will be obese based on the amount of their food intake?

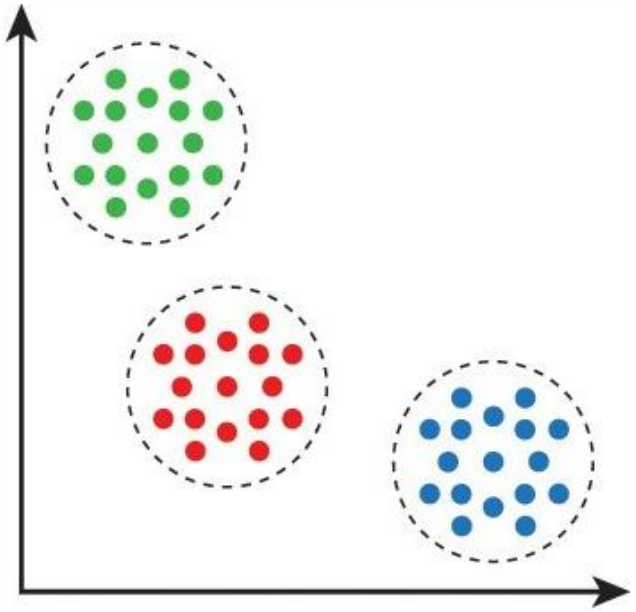
Machine Learning

Classification



Do the medical test results for the patient indicate a risk for a heart attack?

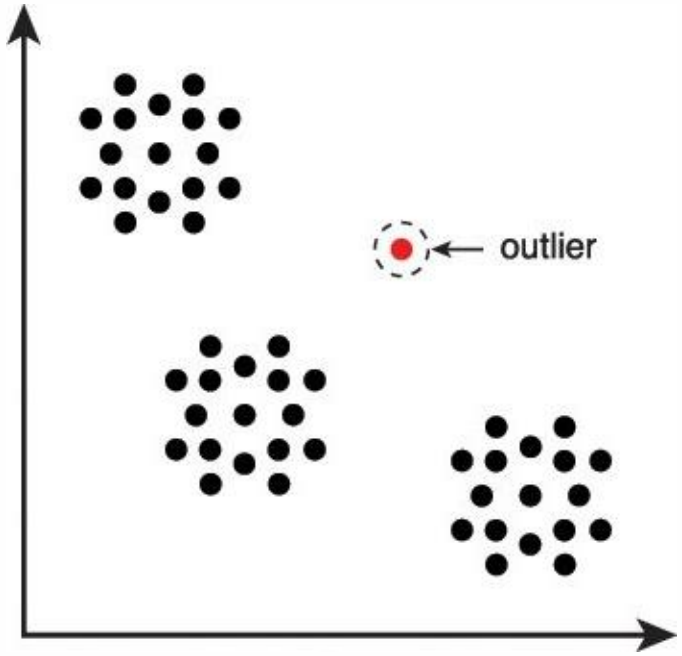
Clustering



How many groups of customers exist based upon similar purchase history?

Machine Learning

Outlier Detection



Is there a particular strain of virus that does not respond to medication?

- Used to identify anomalies, abnormalities and deviations that can be advantageous, such as opportunities, or unfavorable, such as risks.
- Fraud Detection Example: A set of known fraudulent transactions is first fed into the outlier detection algorithm. After training the system, unknown transactions are then fed into the outlier detection algorithm to predict if they are fraudulent or not.

Machine Learning

Filtering

- Filtering is the automated process of finding relevant items from a pool of items.
- Items can be filtered either based on a user's own behavior or by matching the behavior of multiple users.
- **Recommender system: Based on the similarity of the users' behavior** (likes, rating, historical purchases), items are filtered for the target user.

Which holiday destinations can be recommended based on the travel history of a tourist?

Which news articles should be displayed based on a user's interest?

Semantic Analysis

Natural Language Processing (NLP)

- Text and speech recognition
- Example: the food company employs NLP to **transcribe customer calls** into text data that are then mined for commonly recurring reasons of customer dissatisfaction.

How can grammatical mistakes be automatically identified?

How can a system that can correctly understand different accents of English language be designed?

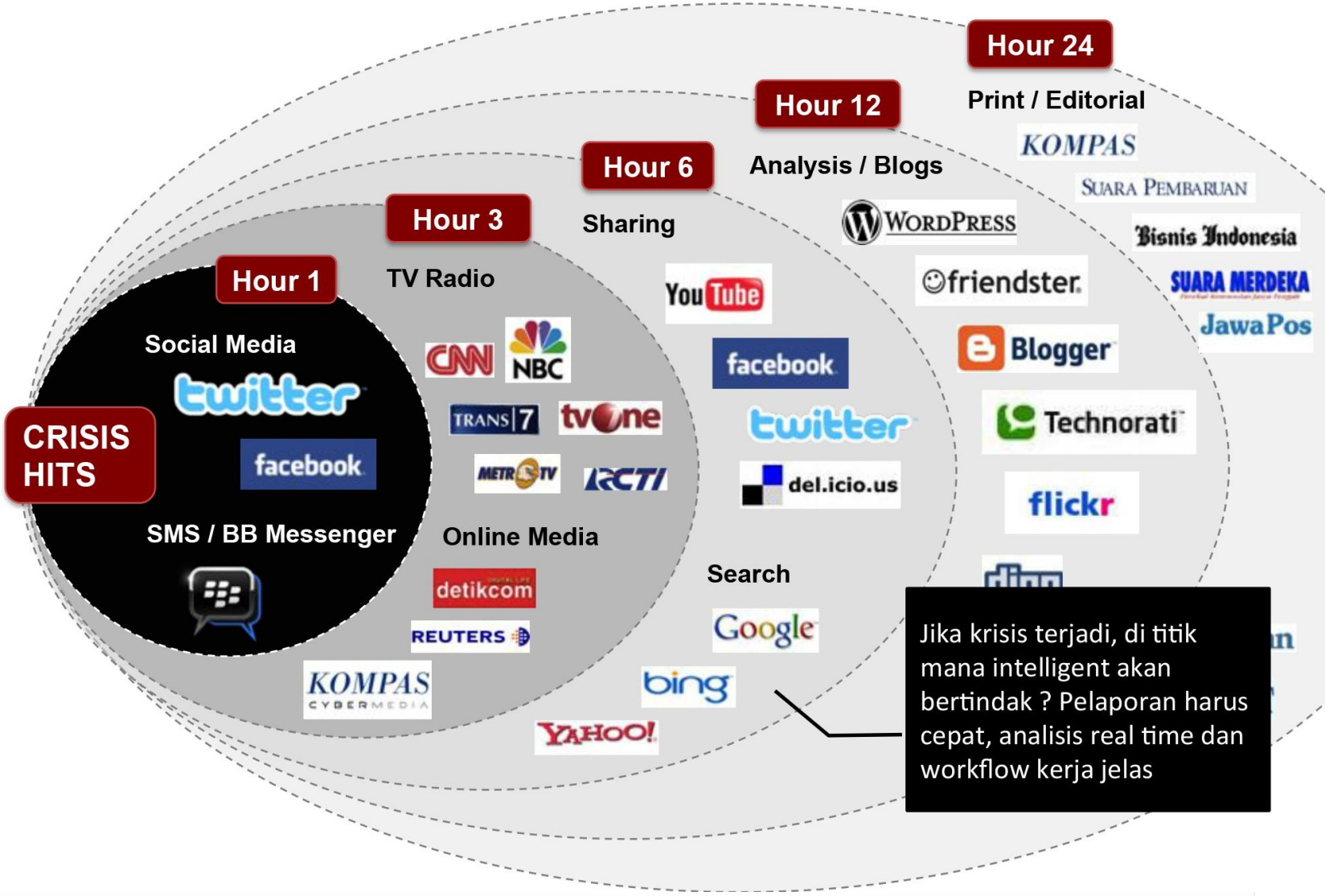
Sentiment Analysis

- Determining the bias or emotions of individuals.
- Example: identifying customer satisfaction or dissatisfaction early, gauging product success or failure, and spotting new trends.

Which contestant is a likely the winner of a singing contest?

Can customer churn be measured by social media comments?

Semantic Analysis



Aha!

Mari kita diskusikan...