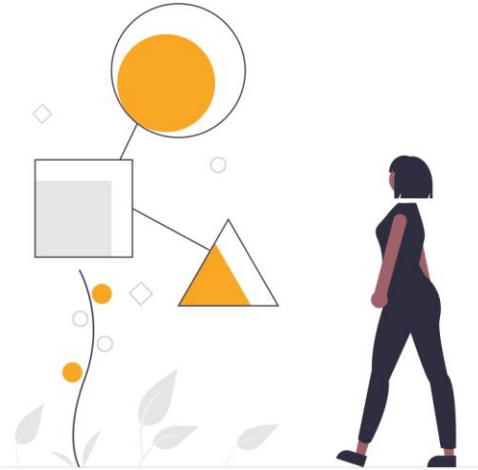
1. Introduction to Data Analytics

Part 1

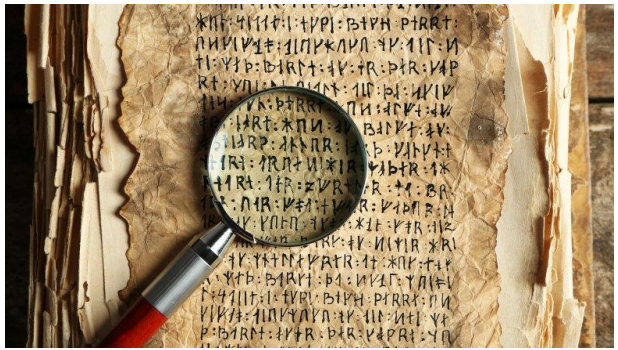
Khalil Israfilzada, PhD
Faculty of Economics and Management
Vytautas Magnus University
Kaunas, 2025



Why Data is Important?

Have you ever been fascinated with ancient languages, perhaps those now known as “**dead**” languages? The complexity of these languages can be mesmerizing, and the best part about them is the extent to which ancient peoples went to preserve them.

Using **ink** made from burned **wood**, **water**, and oil they **copied** the **text** to **papyrus paper**. Some used tools to chisel the text into pottery or stone.



What is the commonality between dead languages and business analytics?

Why Data is Important?

Over 5,000 years ago, the ancient Mesopotamians started to record **quantities on clay tablets**. They partitioned the tablet into **rows** and **columns**. Within each cell, they drew a picture of the type of item and made holes indicating the quantity of it. Each type of item had its own standard pictographic representation, making this ledger language one of the earliest form of human writing we've discovered. It's called **"Proto-cuneiform"** because it later evolved into a complete written language called "Cuneiform".

In other words, the ancients invented Excel before Word!

Why Data is Important?

When it comes to business, product and market data can provide an edge over the competition. That makes this data worth its **weight in gold** (maybe oil?). Important data can include weather, trends, customer tendencies, historical events, outliers, products, and anything else relevant to an aspect of business. What is different about today is how data can be stored. It no longer has to be hand-copied to papyrus or chiselled into stone. It is an automatic process that requires very little human involvement and can be done on a massive scale.

Why Data is Important?

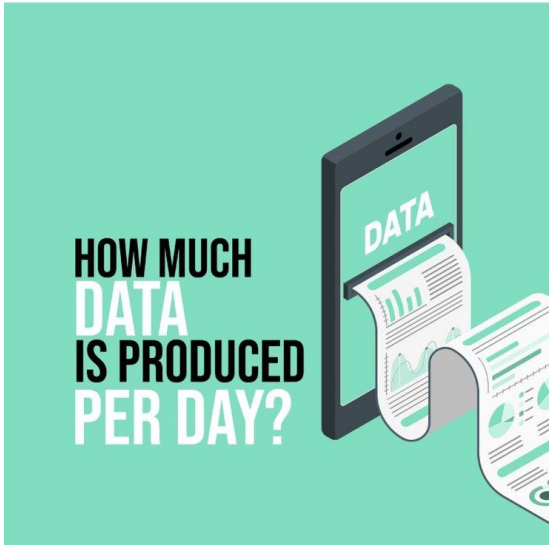
Today, gathering data to help you better understand your customers and business is relatively easy. In fact, it's become so easy there's **the danger of having too much data to deal with.**

In a recent article, data and analytics guru Bernard Marr said:

"While the average small business has less self-generated data than big players. . .this doesn't mean big data is off limits. In fact, in many ways, big data is more suited to small businesses, because **they're generally more agile and able to act more quickly on data-driven insights**."

Big Data Overview

Data is created constantly, and at an ever-increasing rate. **Mobile phones**, **social media**, **imaging technologies** to determine a **medical diagnosis**—all these and more **create new data**, and that **must be stored** somewhere for **some purpose**. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.



Big Data Overview

- Gartner is an independent analysis firm that reports on the technology sector. They phrased one need for the data in several comments on their website including the need for members of an organization ***to speak the same language.***

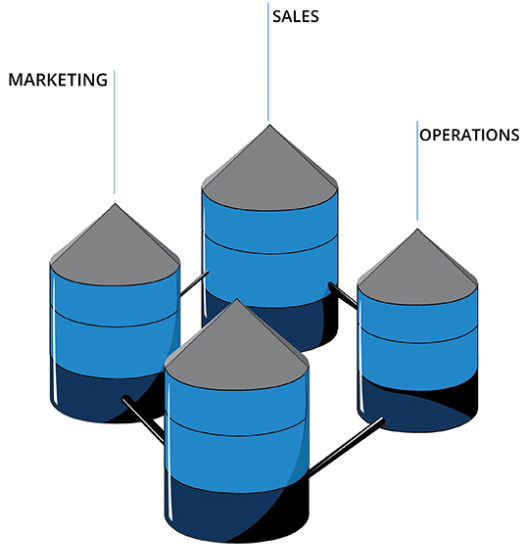
Gartner®

Imagine an organization where the marketing department speaks French, the product designers speak German, the analytics team speaks Spanish and **no one speaks a second language**... That's essentially how **a data-driven business functions** when **there is no data literacy.**

Big Data Overview

- We have spent hundreds of billions of dollars collecting data, but most of it sits in silos. Silos of data never analyzed, never touched again. Not only a **sunk cost** to acquire but also the cost to maintain, backup or archive – the cost is tremendous. Everyone has talked about monetizing this data, but few are very successful.

Data Silos = More Team Members Taking Longer to Achieve Less



To be clear, it is not technology that is hampering progress.

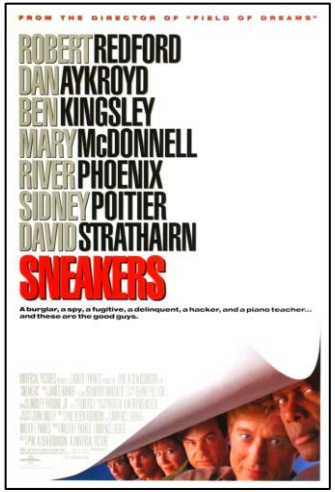
It is lack of vision, human capital, and execution.

Big Data Overview

Cosmo: **The world isn't run by weapons anymore, or energy, or money. It's run by little ones and zeroes, little bits of data. It's all just electrons.**

Sneakers (1992), movie

Data costs include the **cost to acquire it**, the cost to **store** it, the **legal liability** of **keeping** it, the potential **risk** of a **data breach**. We normally think that data is cheap, but when you consider the **total cost of ownership**, it is **really quite expensive.**



Big Data Overview

- **Technology Keeps Raging, but We Need More Than Technology to Be Successful**

Technology is raging and has been for several years and **data is the new oil.** Much of the growth in the last few years could be loosely described as **"creating value from data."** Value could mean increasing sales revenue, reducing avoidable costs, improving patient satisfaction, targeting high-value customers and prospects, creating policy for the broadest social good and much, much more.

Big Data Overview

- There is a HUGE overlap in the areas where value is created from data. This is not by accident as we will explain. But, for now take a look at the dizzying list of overlapping subject areas:

- Business Intelligence (BI)
- Visual BI, Analytics
- Visual Analytics
- Business Analytics
- Data Analytics
- Predictive Analytics
- Prescriptive Analytics
- Advanced Analytics
- Big Data
- Data Science

- Text Analytics
- Graph Analytics
- Social Analytics
- Network Analytics
- Modern Analytics
- Directed Acyclic Graph (DAG) Analytics
- Statistics
- Optimization
- Data Mining
- Data Modeling
- ML

- Decision Science
- Enterprise or Business Decision Management
- Business Process Management
- Data Engineering
- AI
- Computational Intelligence
- Management Science
- Linear and Mathematical Programming
- Deep Learning, Informatics
- Decision Science
- Many others

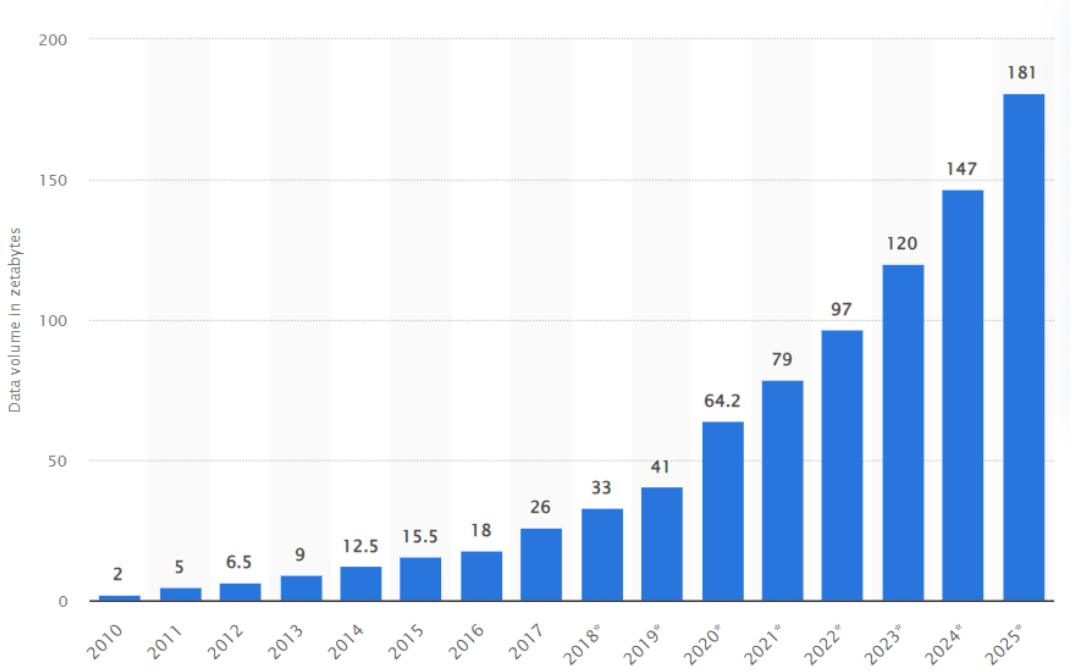
Big Data Overview

Therefore, there is a need for simplification, for **"One Global Term"** that applies to each of these. It is believed ***analytics*** is the appropriate umbrella term that captures the spirit of all these methods; the term ***analytics*** will be used frequently when it is not focusing on a specific form.

! **Analytics** is the process of *extracting* and *creating* information from raw data by *filtering*, *processing*, *categorizing*, *condensing* and *contextualizing* the data.

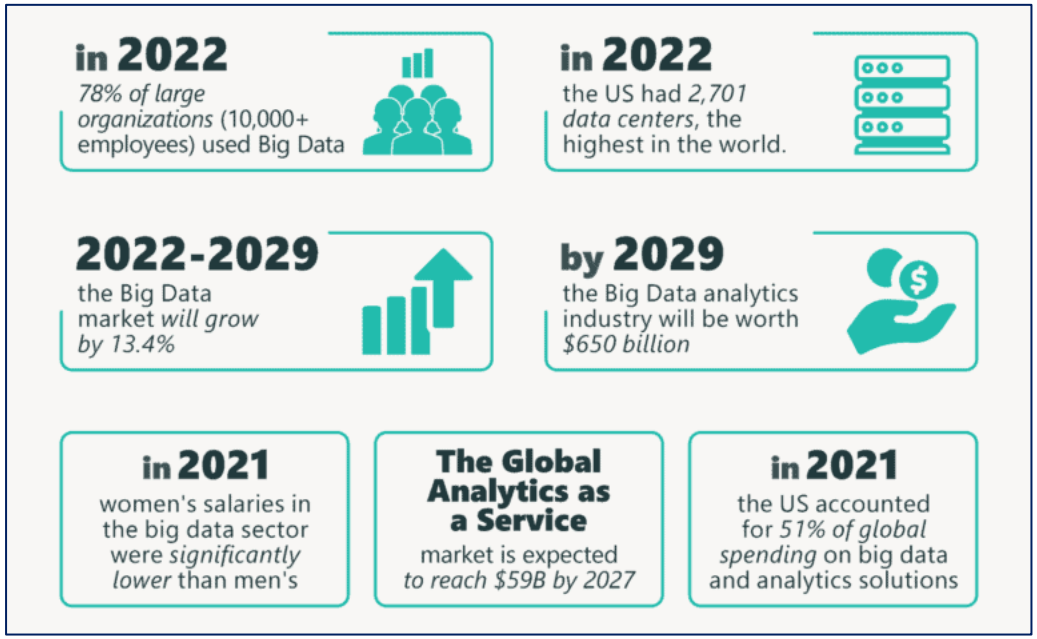
Data and Analytics Explosion

The volume of data collected is growing exponentially with no end in sight. In November 2018 there were 5 billion consumers that interacted with data, but by 2025 it will be 6 billion or 75% of the world's population. In 2025 there will be 150 billion devices creating data in real time.

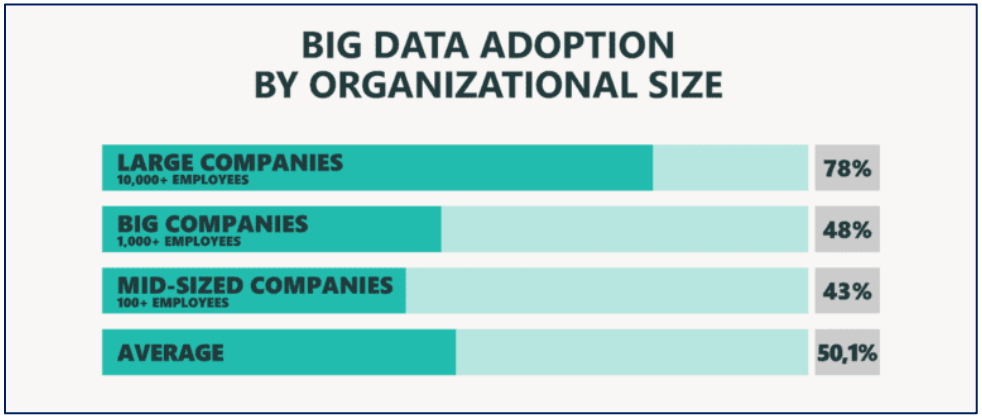


Data and Analytics Explosion

The market was worth $240 billion in 2021 and is projected to grow considerably over the next few years to around **$650 billion in 2029**. From 2020 to 2024, the ratio of unique data to duplicated data is predicted to decrease gradually from 1:9 to 1:10.



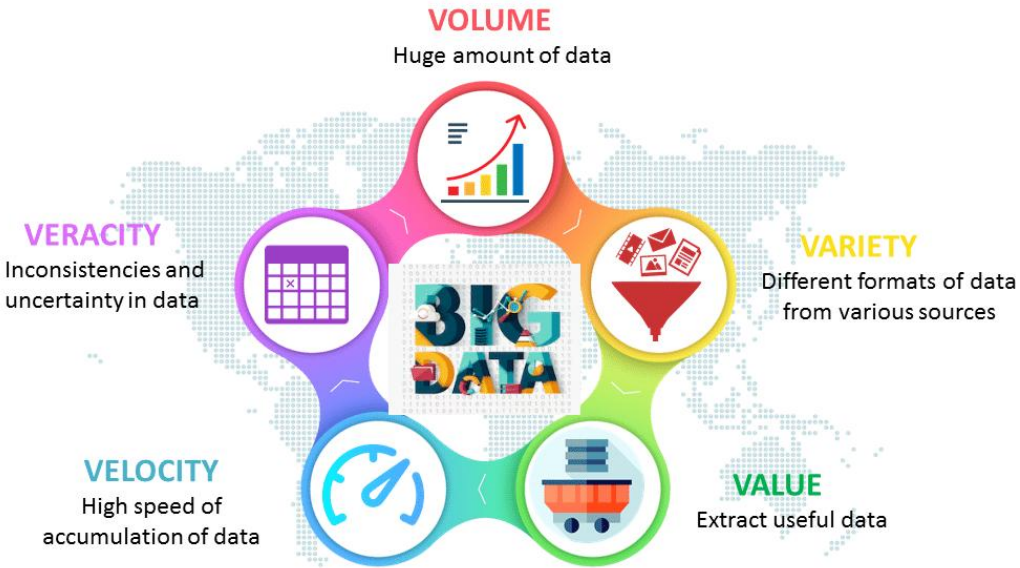
Data and Analytics Explosion



FUN FACT! According to these numbers, there is a probability that **YOU** will end up working for a company that has implemented big data analytics. The probability is approximately **FIFTY** per cent.

Characteristics of Big Data

As with anything huge, we need to make proper categorizations in order to improve our understanding. As a result, features of big data can be characterized by **five Vs.: volume, variety, velocity, value,** and **veracity.**



Characteristics of Big Data

Volume. Big data is a form of data whose volume is so large that it would not fit on a single machine therefore specialized tools and frameworks are required to store process and analyze such data.

Velocity. Velocity of data refers to how fast the data is generated. Data generated by certain sources can arrive at very high velocities, for example, social media data or sensor data.

Variety. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.

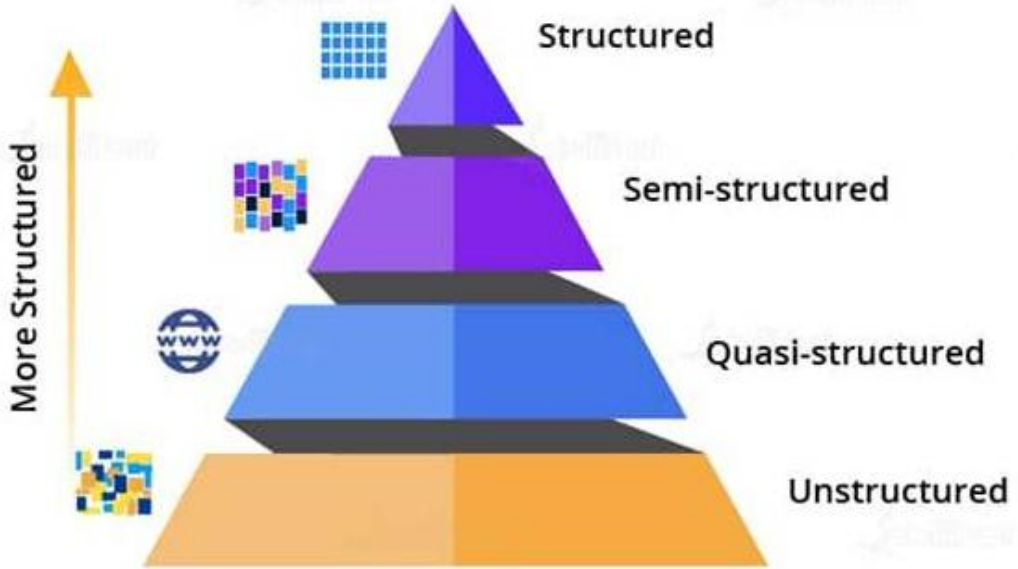
Characteristics of Big Data

Veracity. Veracity refers to the trustworthiness and quality of the data. If the data is not trustworthy and/or reliable, then the value of Big Data remains unquestionable. Veracity refers to how accurate is the data. To extract value from the data, the data needs to be cleaned to remove noise.

Value. Value of data refers to the usefulness of data for the intended purpose. The end goal of any big data analytics system is to extract value from the data.

Data Structures

Big Data comes in many forms; it can be Structured, Semi-Structured, Quasi-Structured or Unstructured as in figure shows the Data structure types and the data growth accordingly.



Data Structures

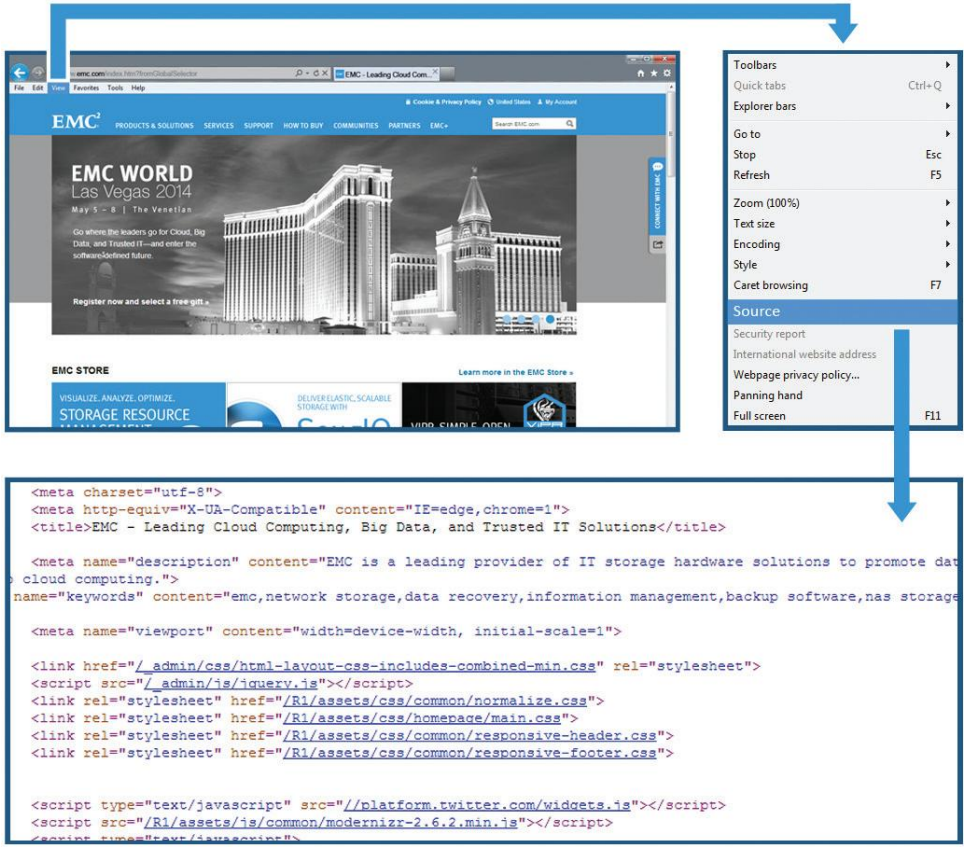
1. Structured Data. Structured data is organized and formatted in a way that it's easily searchable in databases, generally in rows and columns, facilitating straightforward analysis.

Marketing teams can harness structured data to segment their audience and orchestrate targeted campaigns based on various customer attributes like **age**, **location**, and **purchasing patterns**, thus enabling **data-driven decision-making and strategy planning**.

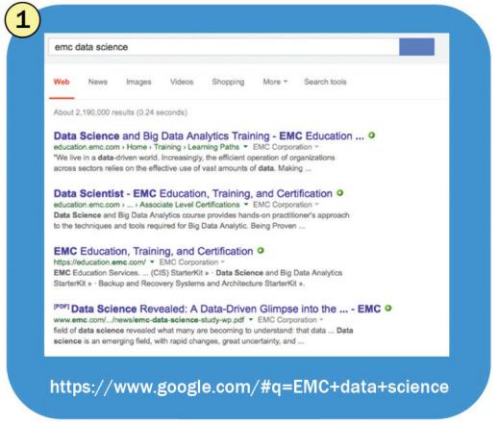
SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

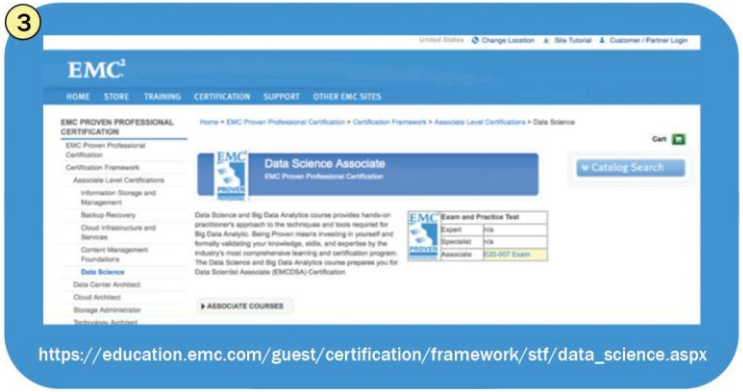
Data Structures

2. Semi-Structured Data. Semi-structured data, although not organized in rows and columns, has some elements of structure, such as tags and hierarchies, which facilitate its analysis to a certain extent.



Data Structures





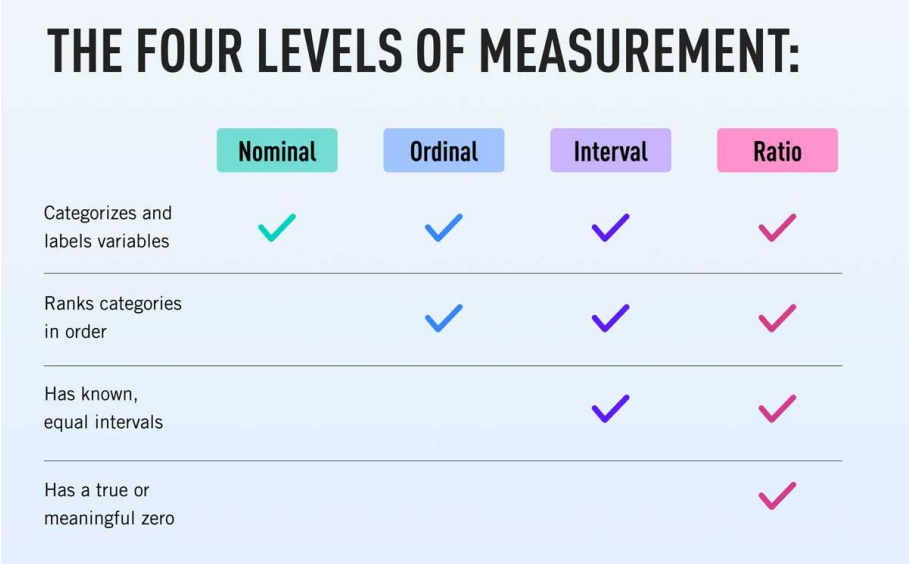
3. Quasi-Structured Data. Quasi-structured data, a type of data that doesn't follow a fixed format or structure but exhibits some levels of organization or patterns that can be extracted and analyzed with specific tools.

Data Structures

4. Unstructured Data. Unstructured data is information that doesn't adhere to a specific form or structure, encompassing a variety of data types such as text, images, and videos, which require advanced tools for effective analysis.

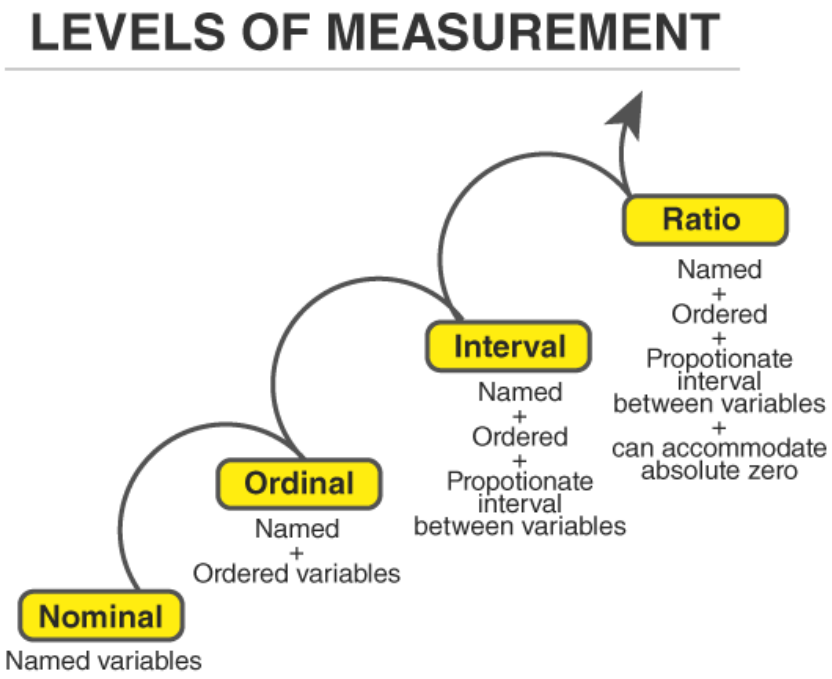
Properties and scales of measurement

Scales of measurement is how variables are defined and categorised. Psychologist **Stanley Stevens** developed the four common scales of measurement: **nominal**, **ordinal**, **interval** and **ratio**. Each scale of measurement has properties that determine how to properly analyse the data. The properties evaluated are **identity**, **magnitude**, **equal intervals** and a **minimum value of zero**.

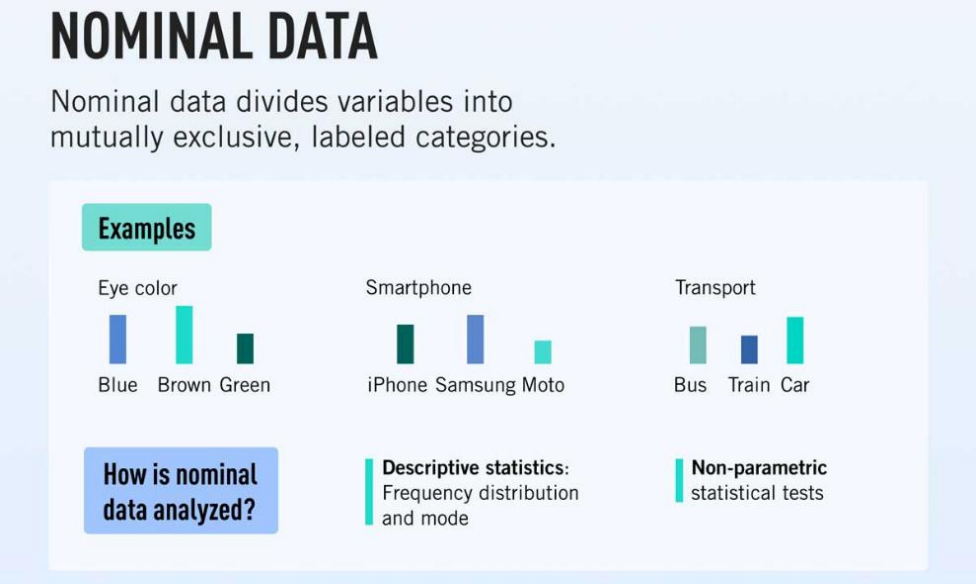


Properties and scales of measurement

Levels of Measurements. There are four different scales of measurement. The data can be defined as being one of the four scales. The four types of scales are:

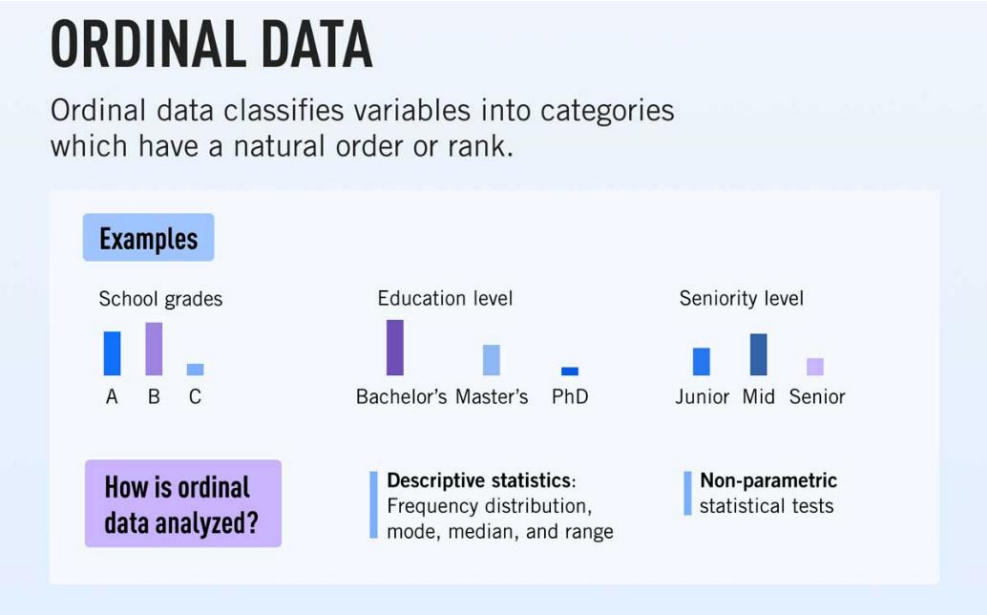


Properties and scales of measurement



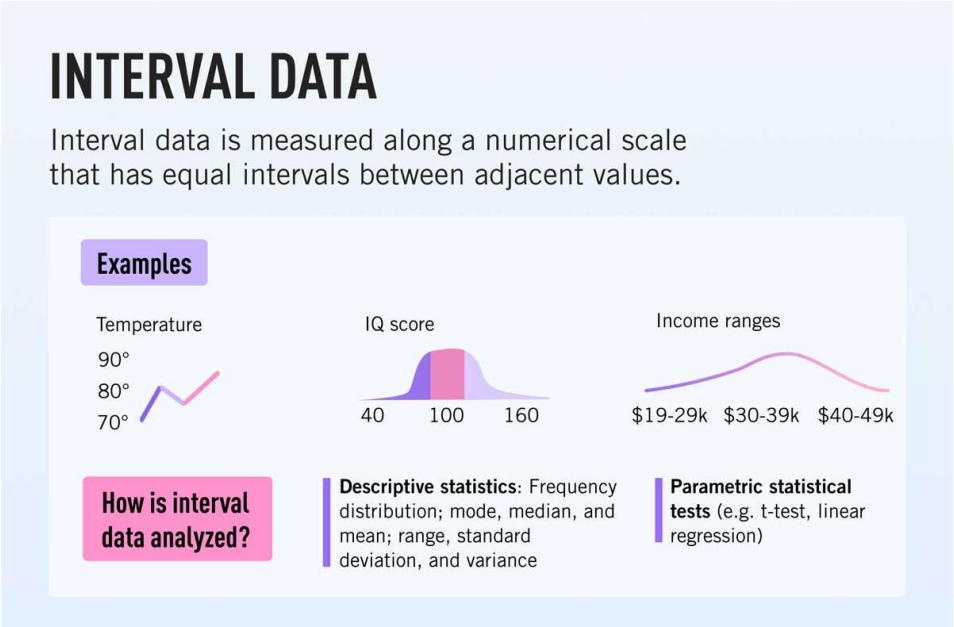
The nominal scale, also known as the categorical scale, is one of the simplest scales of measurement used in statistics and data analytics. A nominal scale is the 1st level of measurement scale in which the numbers serve as "**tags**" or "**labels**" to classify or identify the objects.

Properties and scales of measurement



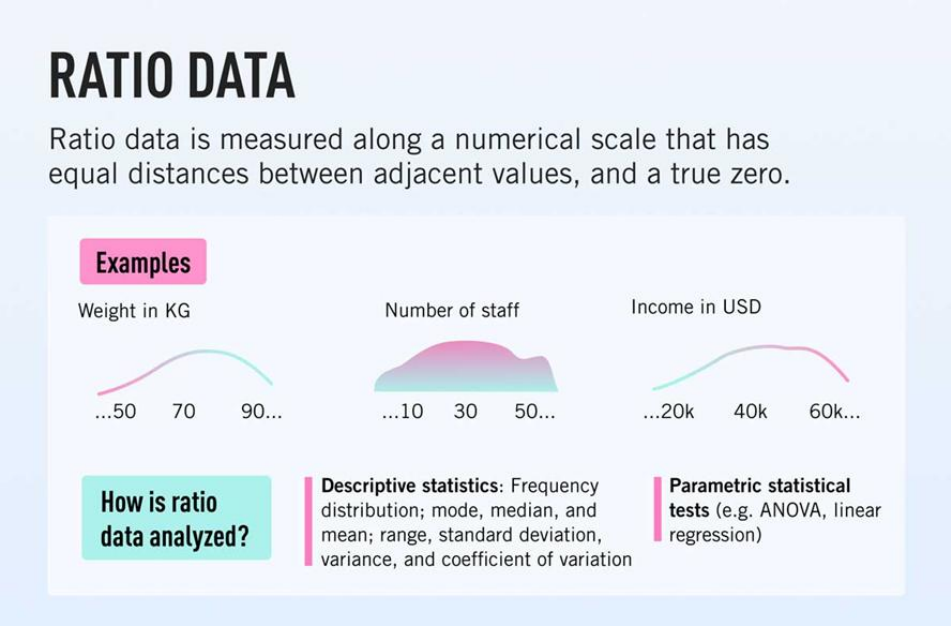
The ordinal scale of measurement is used to categorize data and indicate the relative order of the items being measured. Unlike nominal scales, ordinal scales provide clear directions regarding the hierarchy or order of categories.

Properties and scales of measurement



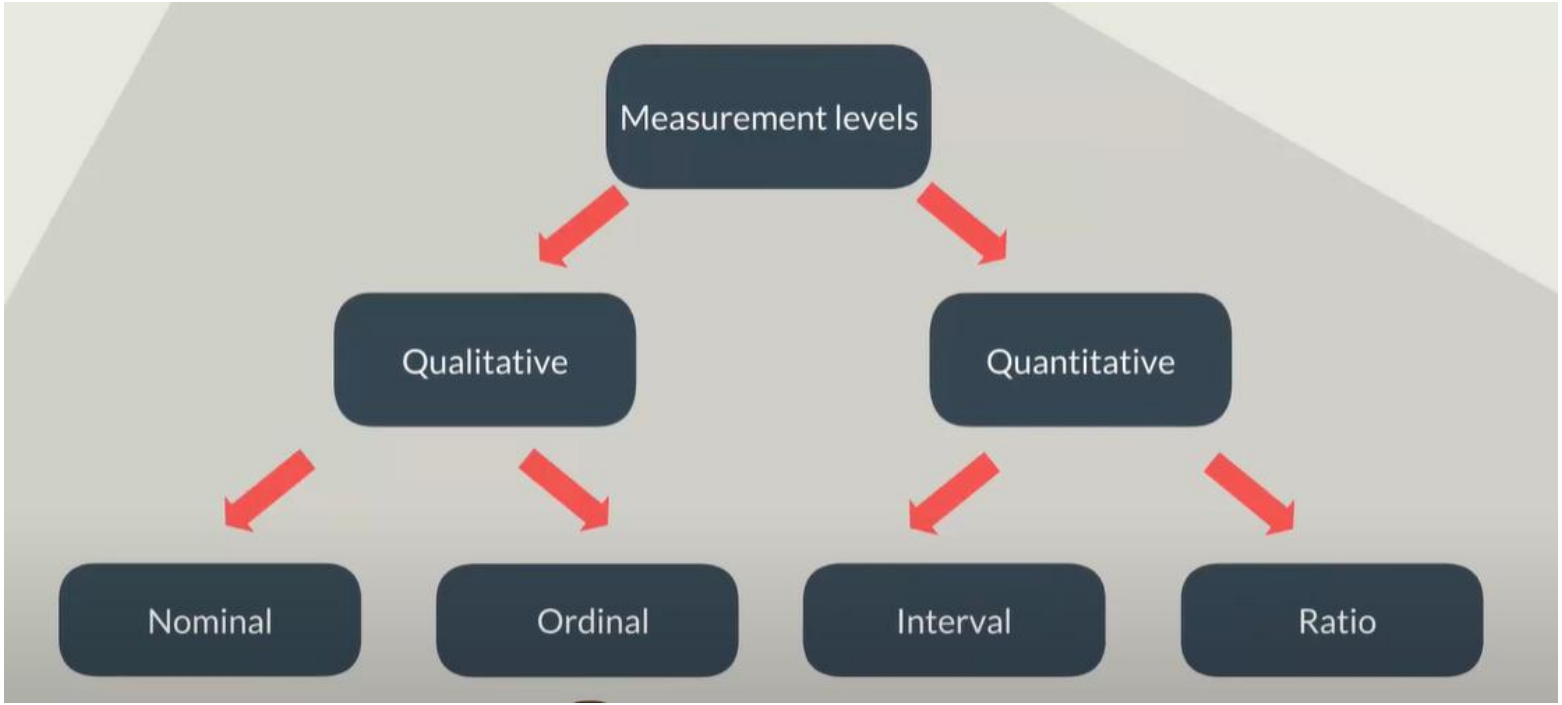
The interval scale of measurement represents a step further in complexity from the ordinal scale, incorporating a standardized scale of measurement that allows for the determination of the exact distances between scale points.

Properties and scales of measurement



The ratio scale of measurement is the highest level of measurement, which shares all the characteristics of the interval scale and also includes a true zero point. This allows for a wide range of statistical analyses to be conducted, including the calculation of ratios, which is not possible with interval data.

Data at the highest level: qualitative and quantitative



ANY QUESTIONS?