Course	Python for Data Science
Term	
Class	
Date	
Chapter. Topic	

What is data science?

Siva R Jasthi

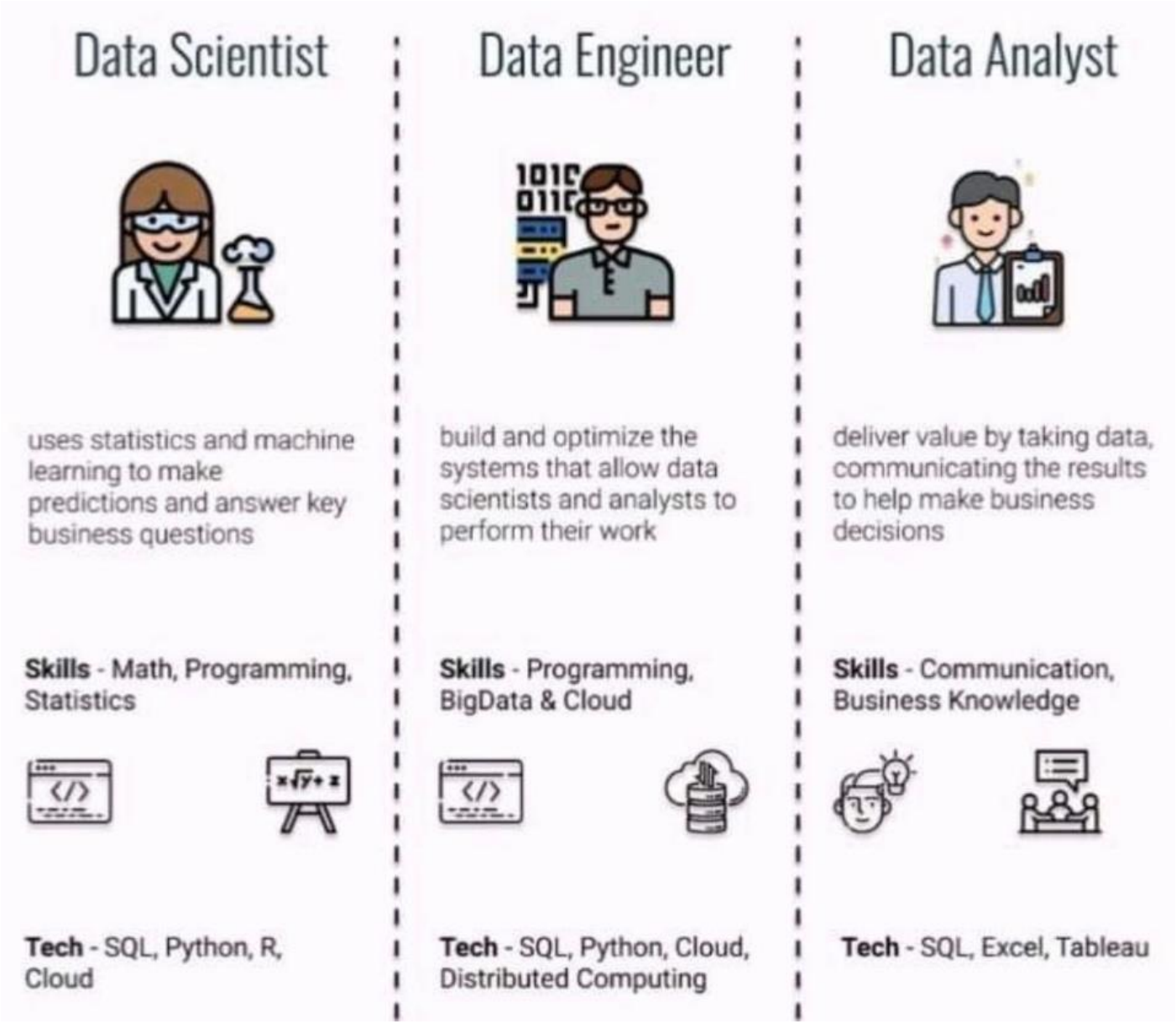
Computer Science and Cybersecurity

Metropolitan State University

Outline

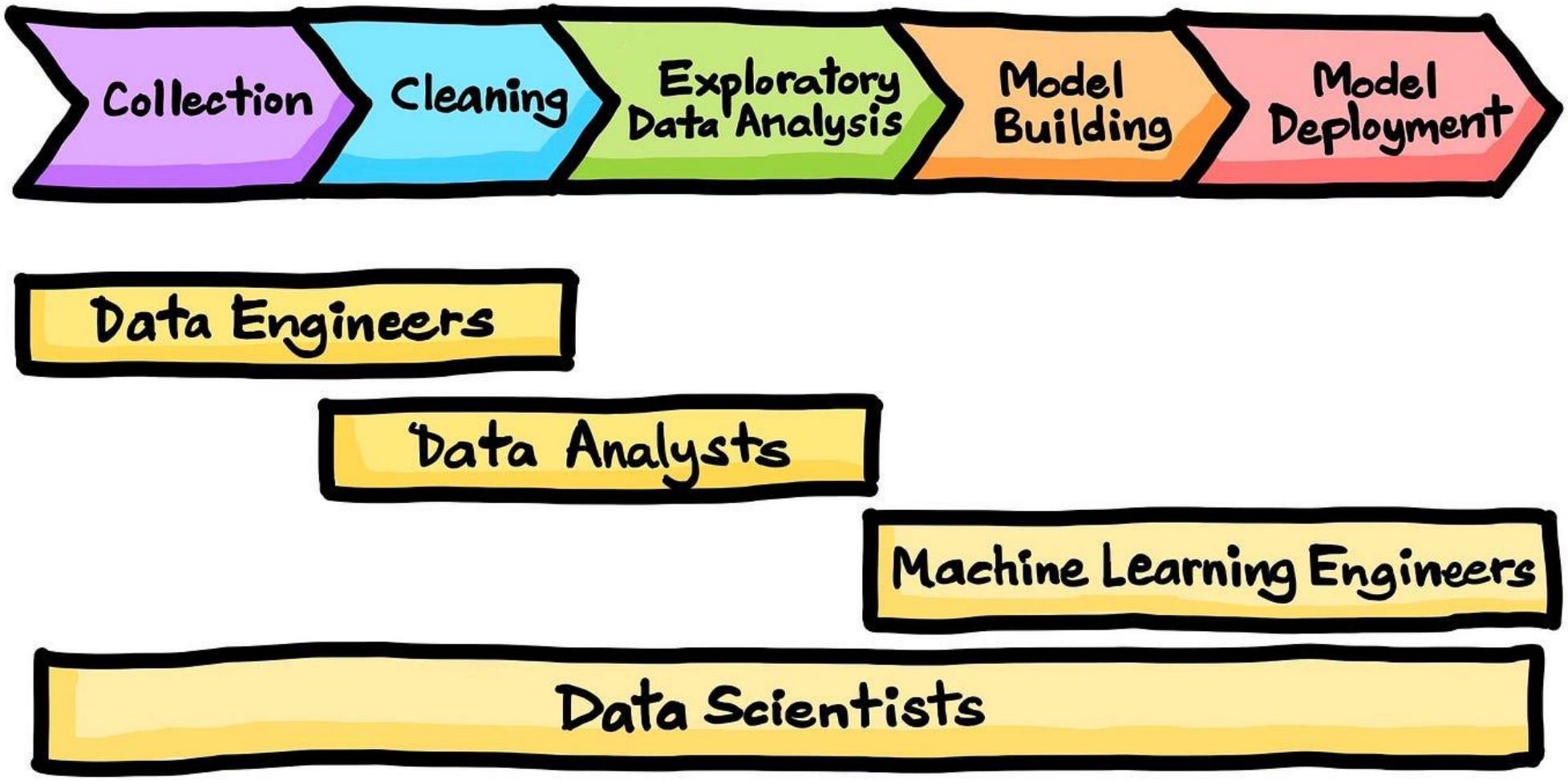
- What is Data Science?
- Why Data Science?
- Data Science Process
- Data Sources
- Data Types

Data _____?



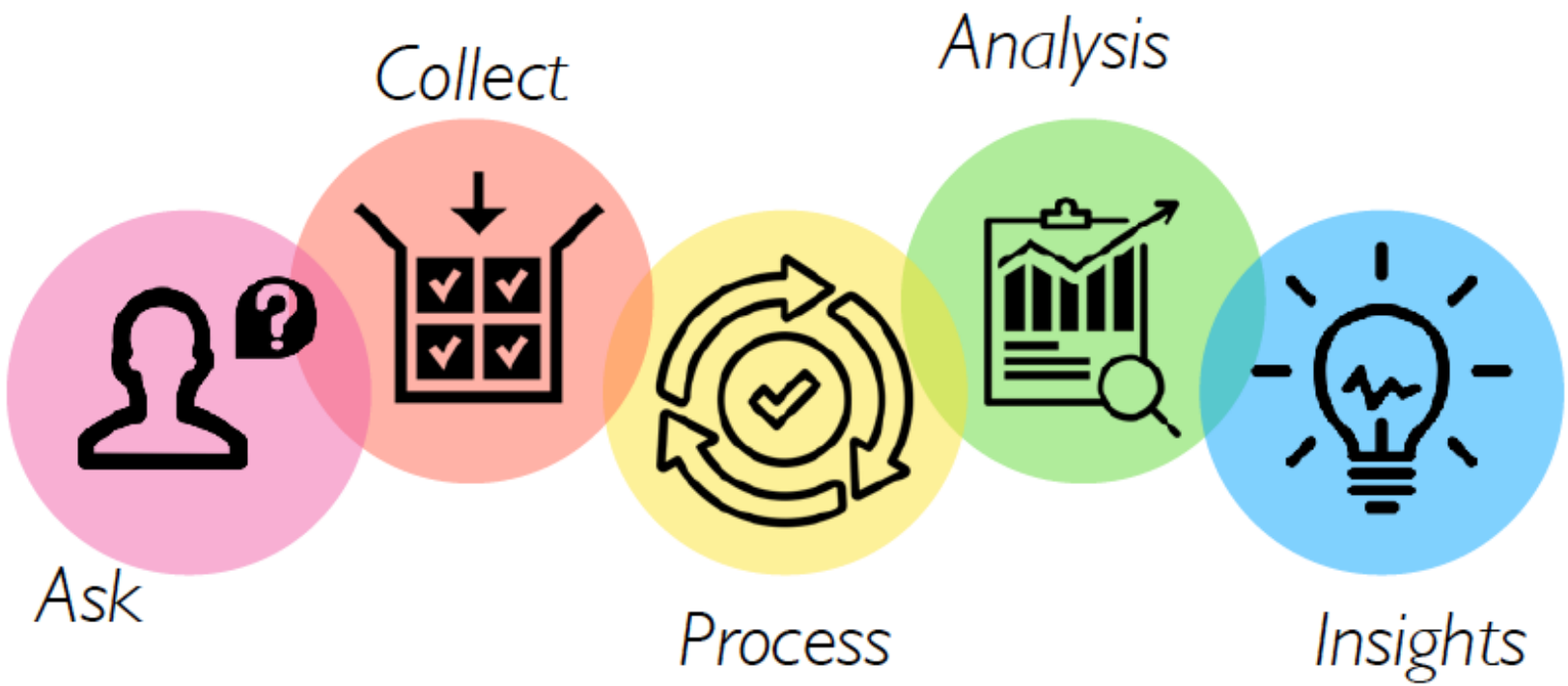
- Data Scientist
- Data Analyst
- Data Engineer
- ML Engineer

Data Science Process



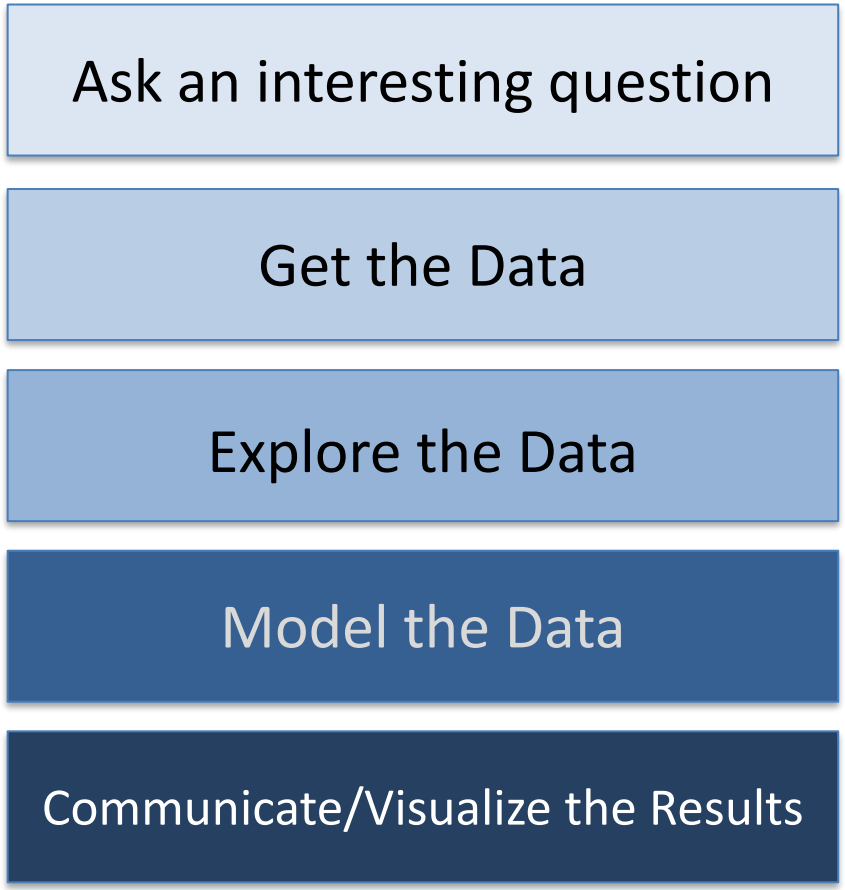
https://dataprofessor.beehiiv.com/p/the-data-science-process

DATA SCIENCE **PROCESS**



https://www.yellowslice.in/bed/data-science-design-thinking-process-of-amalgamation-to-perfection/

The Data Science Process



Some interesting questions

Ask an interesting question

Sports: "Which factors contribute most to a basketball team's success?“

Environment: "Is air pollution worse in urban areas compared to rural areas?“

Entertainment: "What types of movies are most popular during different times of the year?“

Health: "Does eating breakfast affect students' performance in school?“

Weather: "How does weather impact people's mood or productivity?“

Social Media: "What time of day gets the most engagement on social media?“

Shopping: "Do people spend more money on shopping during weekends compared to weekdays?

Some interesting questions (Contd.)

Ask an interesting question

1. Education: "Do people with higher levels of education (college vs. high school) have higher average salaries?“
2. College Majors: "Which college majors lead to the highest-paying jobs?“
3. Job Market: "How does the unemployment rate vary by education level?“
4. Degree Importance: "Is there a strong correlation between having a college degree and job satisfaction?“
5. Job Opportunities: "What industries hire the most college graduates?“
6. College vs. Trade School: "Do trade school graduates find jobs faster than college graduates?“
7. Job Stability: "Are people with college degrees less likely to change jobs compared to those without degrees?"

How to get the Data?

Get the Data

- You create it

- creating it from scratch

- Someone else provides it, all pre-packaged for you (e.g., files)

- sheroes_data.xlsx

- Someone else gives you access through an API or Key to get the data

https://www.projectabcd.com/api/getinfo.php?id=50 (gets the JSON data)

- Someone else has content on web and you scrape it.

https://www.projectabcd.com/display_the_dress.php?id=50 (html)

How to get the Data? Google the question

Get the Data

College Majors: "Which college majors lead to the highest-paying jobs?”

- https://www.kaggle.com/code/suugaku/dataquest-visualizing-earnings-by-college-major/input

Some popular data sources

See the document “Where to get data for Data Science and Machine Learning?”

Data Sources

Where to get the Data for Data Science and Machine Learning projects?

1. Kaggle

- Description: One of the largest platforms for data science and machine learning competitions, Kaggle offers a vast collection of datasets across various domains.
- Website: https://www.kaggle.com/datasets
- Popular Uses: Competitions, projects, and learning resources.

2. UCI Machine Learning Repository

- Description: A long-standing resource for machine learning datasets, hosted by the University of California, Irvine. It provides hundreds of datasets used in academic research.
- Website: https://archive.ics.uci.edu/ml/index.php
- Popular Uses: Academic research, training models, teaching.

3. Google Dataset Search

- Description: A specialized search engine from Google that helps you find datasets across the web. It indexes datasets from multiple sources and repositories.
- Website: https://datasetsearch.research.google.com/
- Popular Uses: Finding niche and domain-specific datasets.

4. AWS Open Data Registry

- Description: Amazon provides a collection of publicly available datasets, which can be accessed and analyzed via their cloud services.
- Website: https://registry.opendata.aws/
- Popular Uses: Large-scale datasets for cloud-based machine learning.

Types of Data Sources

1. Structured Data Sources

•**Relational Databases (RDBMS):** Data stored in structured tables, e.g., MySQL, PostgreSQL, Microsoft SQL Server, and Oracle.

•**Data Warehouses:** Centralized repositories of integrated data from multiple sources, often used for analytics, e.g., Amazon Redshift, Google BigQuery, Snowflake.

•**Spreadsheets:** Simple structured data stored in formats like Excel (.xls), CSV, or Google Sheets.

2. Unstructured Data Sources

•**Text Data:** Documents, emails, and chat logs, often stored in file formats like .txt, .pdf, or sourced from APIs (social media, web scraping).

•**Multimedia Data:** Images, audio, and video files sourced from cameras, sensors, or streaming platforms.

•**Social Media Data:** Platforms like Twitter, Facebook, and Instagram, providing unstructured data through APIs for text, images, and videos.

•**Log Files:** System or application logs that record events, errors, or user activity (e.g., web server logs).

Types of Data Sources

3. Semi-Structured Data Sources

- **JSON and XML Files:** Used for API responses, web data, or configuration files, where data is not fully structured but follows a pattern.
- **NoSQL Databases:** Databases that handle semi-structured or unstructured data, such as MongoDB, Cassandra, and HBase.

4. Real-Time Data Sources

- **Streaming Data:** Data that arrives in continuous flows from sensors, financial tickers, social media streams, and other real-time systems. Examples include Apache Kafka, AWS Kinesis, and Azure Stream Analytics.
- **IoT (Internet of Things) Devices:** Data collected from interconnected devices like smart appliances, wearables, or industrial sensors.

Types of Data Sources

5. Open Data Sources

•**Public Datasets:** Free-to-use datasets provided by governments, international organizations, and research institutions, such as datasets from Kaggle, UCI Machine Learning Repository, or government portals (data.gov).

•**Open APIs:** APIs provided by organizations or platforms for free or paid access to data, such as weather data, financial data, or healthcare data (e.g., NASA APIs, World Bank APIs).

6. Web Data

•**Web Scraping:** Extracting data from websites, typically unstructured or semi-structured, using tools like BeautifulSoup, Scrapy, or Selenium.

•**RSS Feeds:** Syndicated web content in XML format, often used to gather news, blogs, or updates

Types of Data Sources

7. Cloud Data Sources

- **Cloud Storage Services:** Services like AWS S3, Google Cloud Storage, and Azure Blob Storage for storing large datasets in the cloud.
- **Cloud Databases:** Managed databases and data platforms provided by cloud providers, such as Google BigQuery, Amazon RDS, and Microsoft Azure SQL.

8. Geospatial Data Sources

- **GIS Data:** Data with geographic or spatial components, often from GPS systems, satellites, or geographic information systems (GIS). Examples include OpenStreetMap, Google Maps API, and ESRI datasets.
- **Remote Sensing Data:** Data collected from sensors, drones, or satellites (e.g., Landsat satellite data).

Types of Data Sources

9. Sensor Data

- **IoT and Wearable Devices:** Data collected from physical devices, including smartwatches, fitness trackers, or smart home systems.
- **Industrial Sensors:** Used in manufacturing, agriculture, or healthcare to monitor conditions and collect data.

Types of Data Sets: (1) Record Data

- Relational records
  - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

- Transaction data

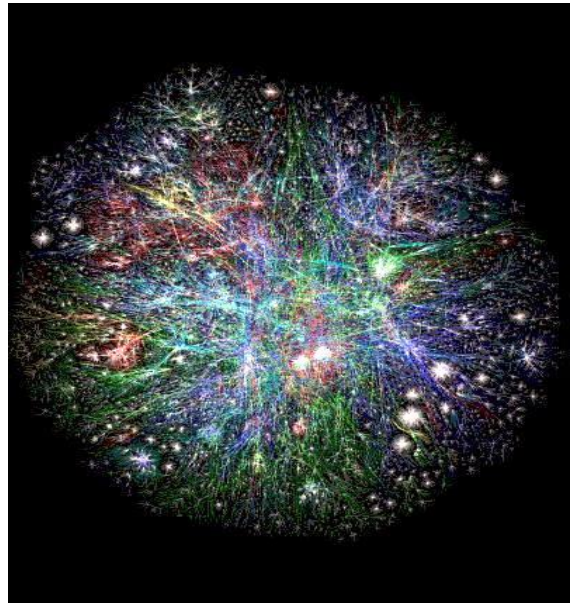
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

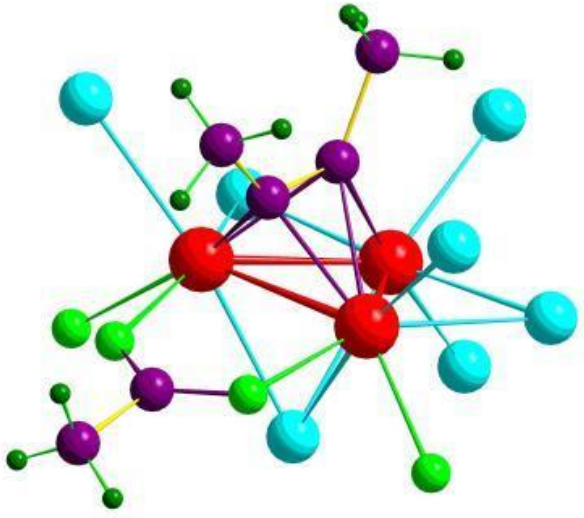
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Document data: Term-frequency vector (matrix) of text documents

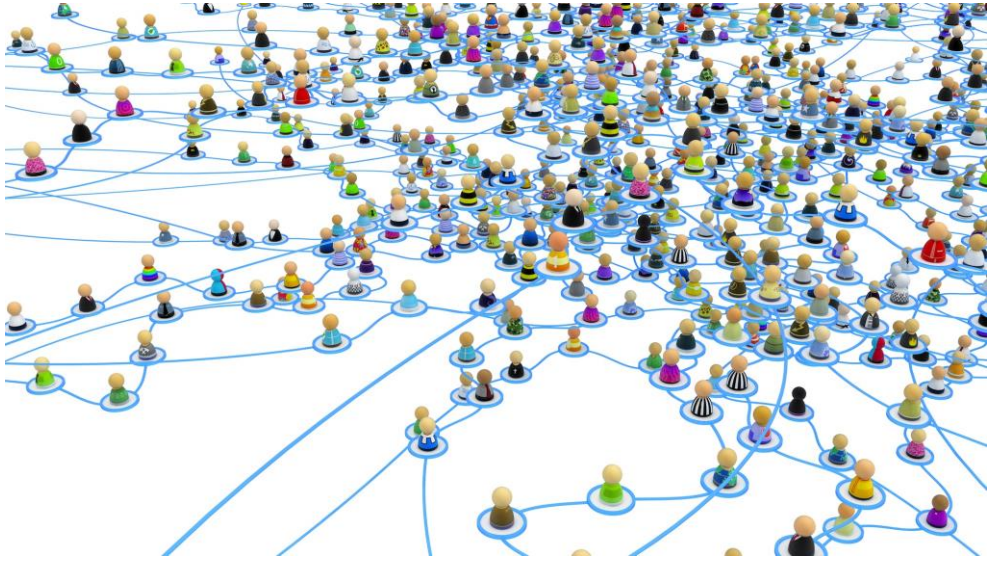
Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web



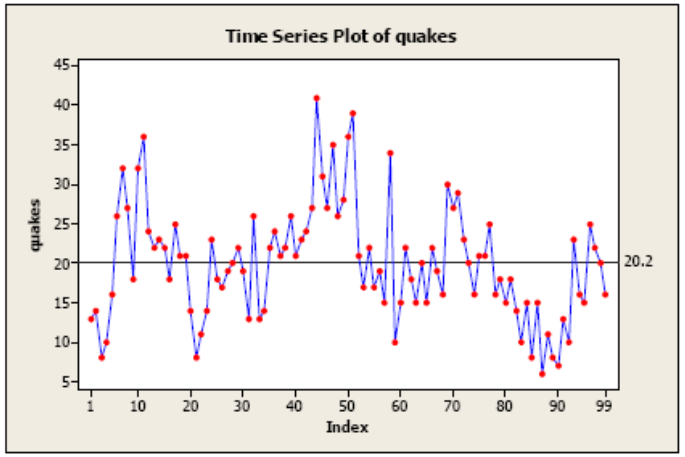


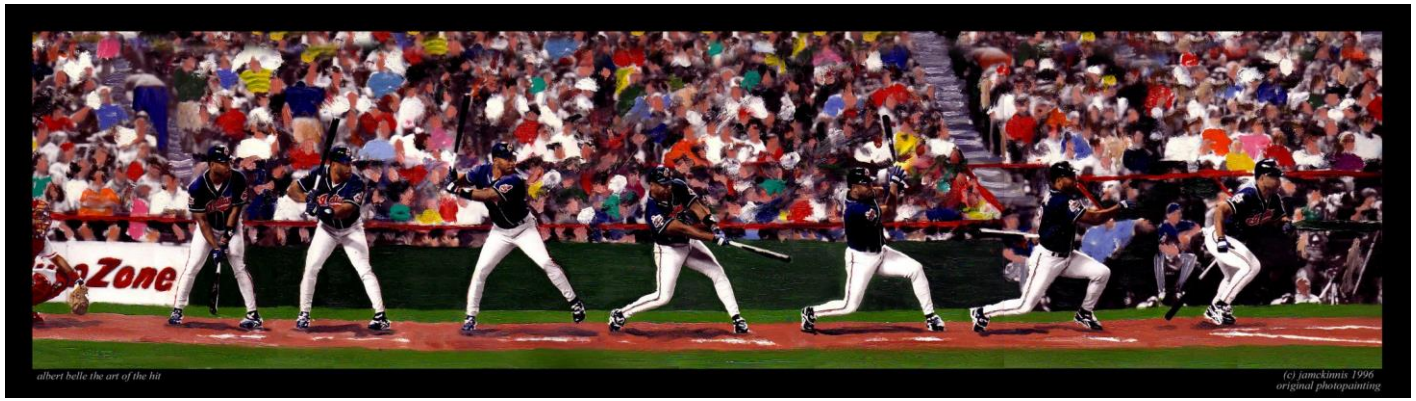
- ❑ Molecular Structures
- ❑ Social or information networks



Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series



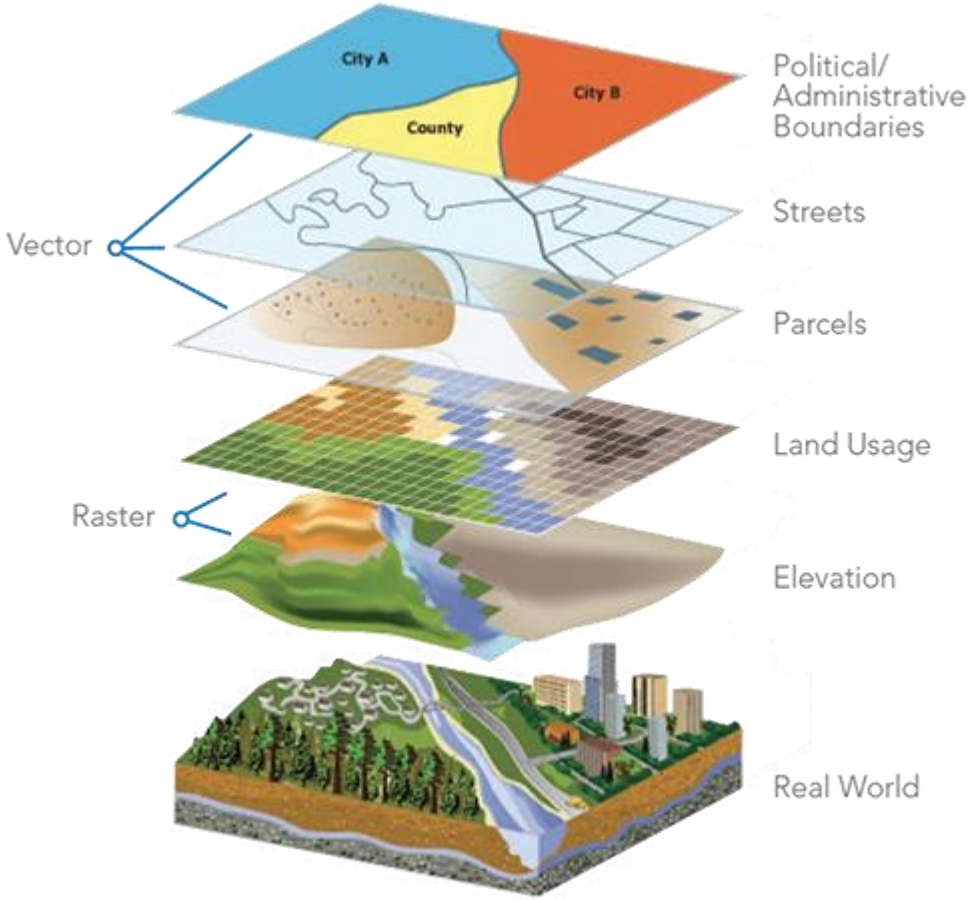


- Sequential Data: transaction sequences

- Genetic sequence data

Types of Data Sets: (4) Spatial, image and multimedia Data

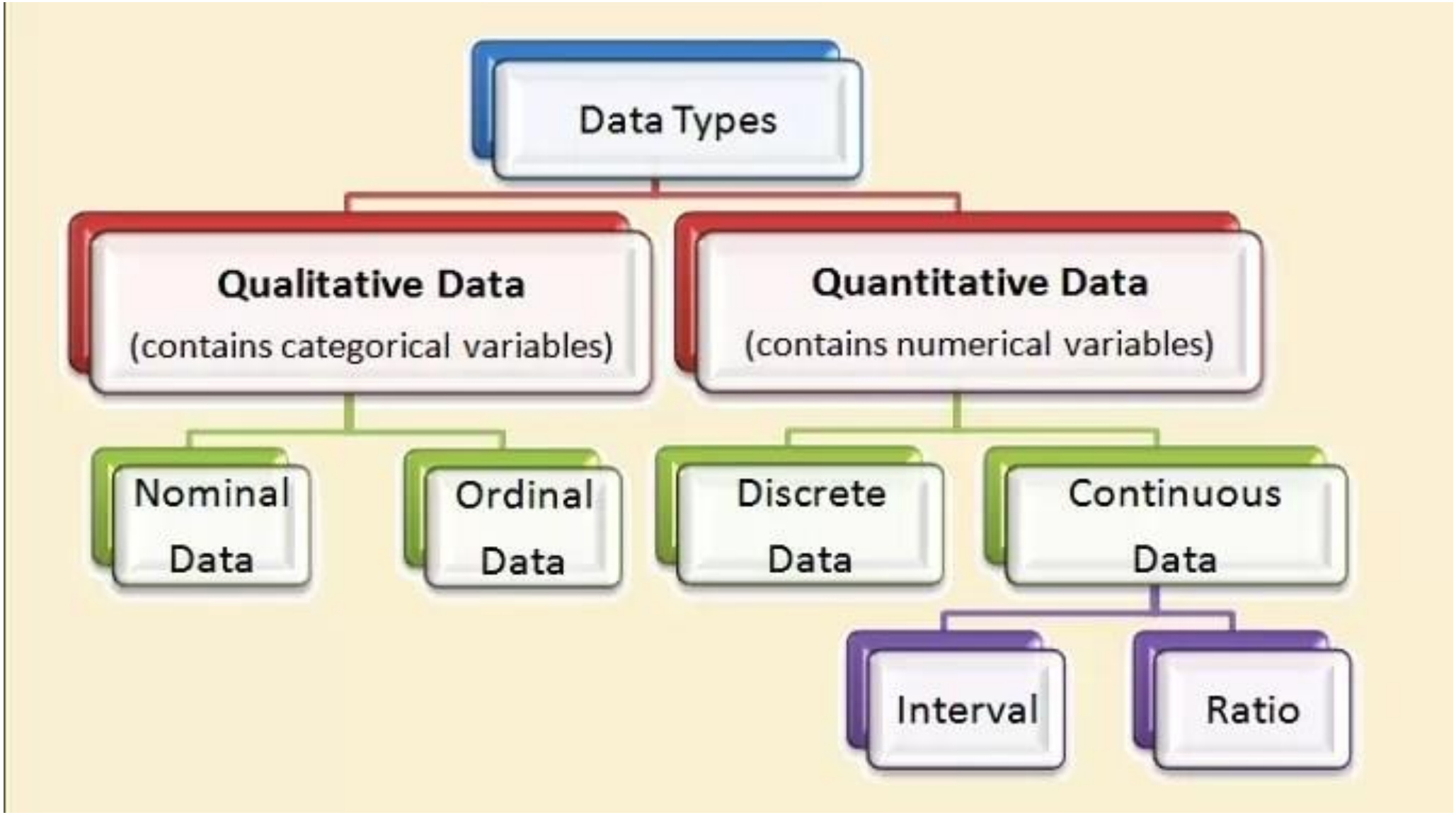
- Spatial data: maps



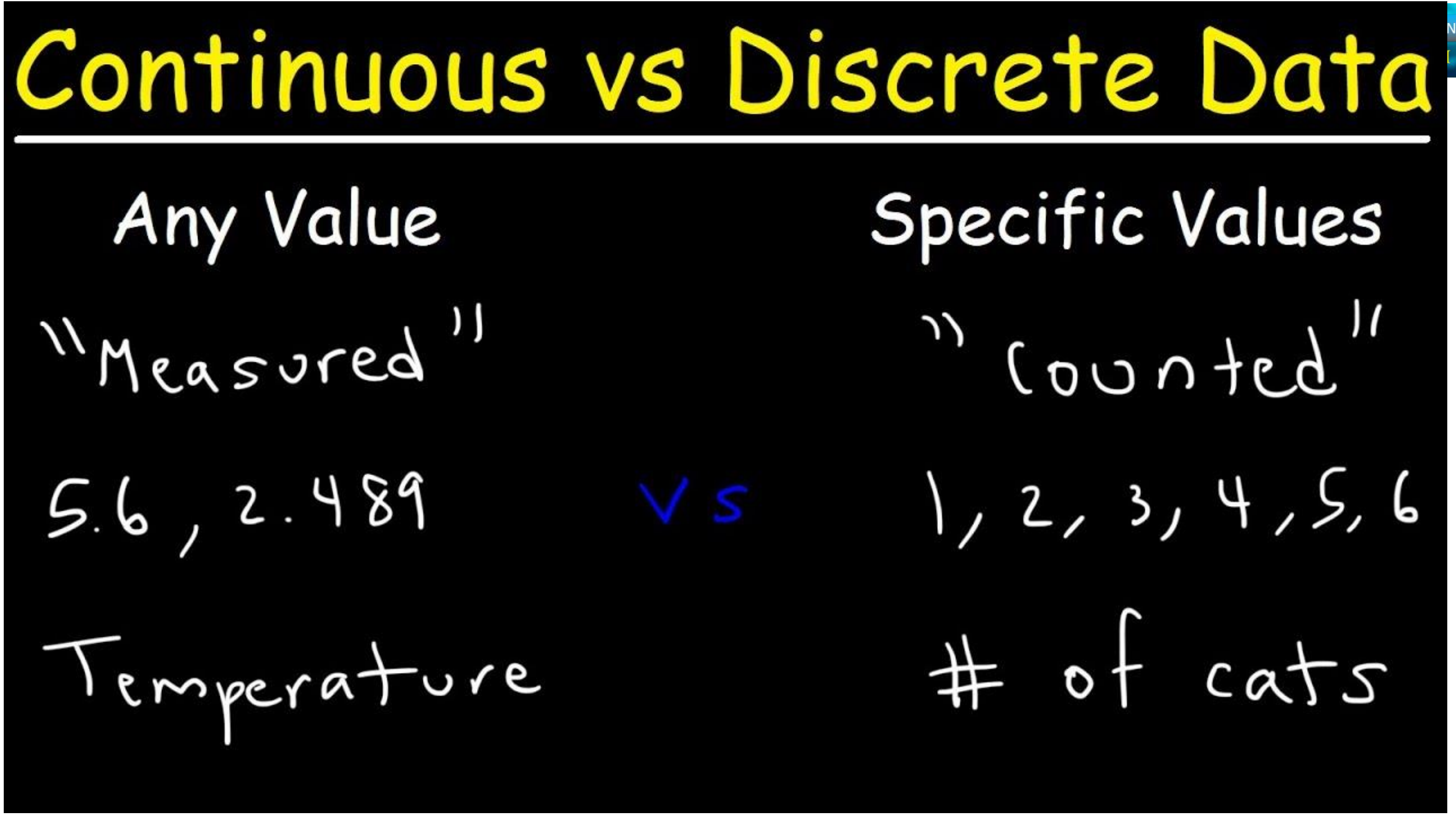
- Image data:

- Video data:

Data Types


Data Types
Qualitative Data
(contains categorical variables)
Quantitative Data
(contains numerical variables)
Nominal Data
Ordinal Data
Discrete Data
Continuous Data
Interval
Ratio

https://microbiologynote.com/difference-between-quantitative-and-qualitative-data/



https://www.youtube.com/watch?v=cz4nPSA9rlc

NOMINAL DATA

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color

Smartphone

Transport

How is nominal data analyzed?

Descriptive statistics: Frequency distribution and mode

Non-parametric statistical tests

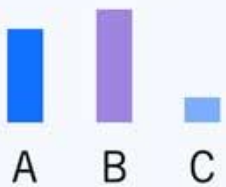
https://careerfoundry.com/en/blog/data-analytics/what-is-nominal-data/

ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades



Education level

Seniority level

How is ordinal data analyzed?

Descriptive statistics: Frequency distribution, mode, median, and range

Non-parametric statistical tests

https://careerfoundry.com/en/blog/data-analytics/what-is-ordinal-data/

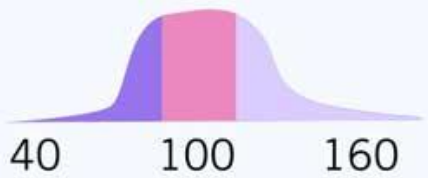
INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature

IQ score



Income ranges

How is interval data analyzed?

- **Descriptive statistics**: Frequency distribution; mode, median, and mean; range, standard deviation, and variance
- **Parametric statistical tests** (e.g. t-test, linear regression)

https://careerfoundry.com/en/blog/data-analytics/what-is-interval-data/

Why is temperature data called “interval” data?

Temperature measurements in Celsius are considered **interval data** because they meet the following criteria:

1. **Equal Intervals**: The difference between any two consecutive values (e.g., 15°C to 20°C) is consistent. The interval between each unit of measurement is the same, making it possible to perform meaningful mathematical operations like addition and subtraction.

2. **No True Zero**: Interval data lacks a true zero point that indicates the absence of the variable being measured. In the Celsius scale, 0°C does not represent the complete absence of temperature, but rather the freezing point of water. This distinguishes interval data from **ratio data**, which does have a meaningful zero point (e.g., weight, where 0 kg means no weight).

Therefore, while you can calculate differences between temperatures in Celsius, you cannot multiply or divide them meaningfully (e.g., you can’t say 20°C is twice as warm as 10°C), which is why it's classified as **interval** rather than **ratio** data.

RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG

...50 70 90...

Number of staff

...10 30 50...

Income in USD

...20k 40k 60k...

How is ratio data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

https://careerfoundry.com/en/blog/data-analytics/what-is-ratio-data/

THE FOUR LEVELS OF MEASUREMENT:

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

https://careerfoundry.com/en/blog/data-analytics/what-is-interval-data/

Attribute Types

- **Nominal:** categories, states, or "names of things" – Not ordered(smaller or greater)
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important/opportunity
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal: There is order

- Values have a meaningful order (ranking) but magnitude between successive values is not known
- *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
    - Measured on a scale of **equal-sized units**
    - Values have order
        - E.g., *temperature in C˚or F˚, calendar dates(gradual process)*
    - No true zero-point
- **Ratio**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
        - e.g., *temperature in Kelvin, length, counts, monetary quantities(Double money then what you have)*

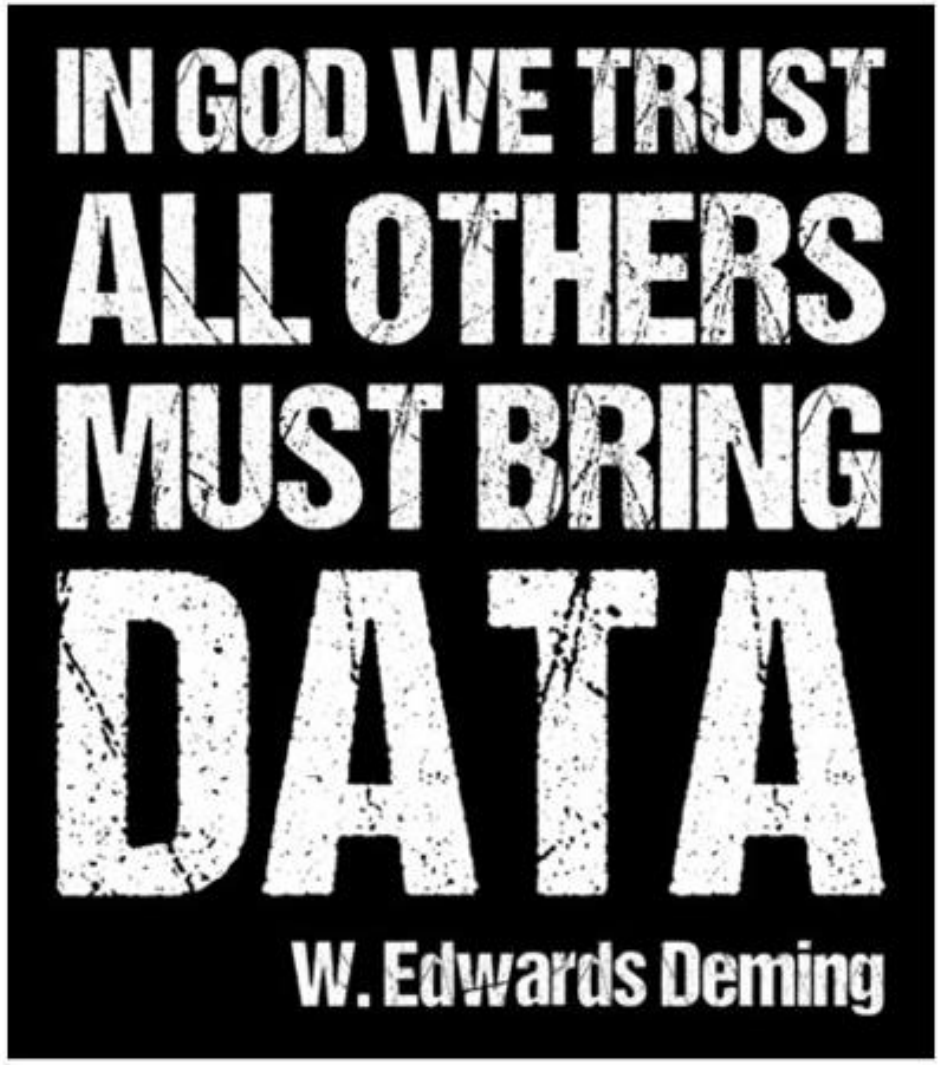
Discrete vs. Continuous Attributes

- **Discrete Attribute: Certain precision**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes(no decimal values), profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Summary


IN GOD
WE TRUST
ALL OTHERS
MUST BRING
DATA


IN GOD WE TRUST
ALL OTHERS
MUST BRING
DATA


IN GOD WE TRUST
ALL OTHERS
MUST BRING
DATA
W. Edwards Deming

Thank You.