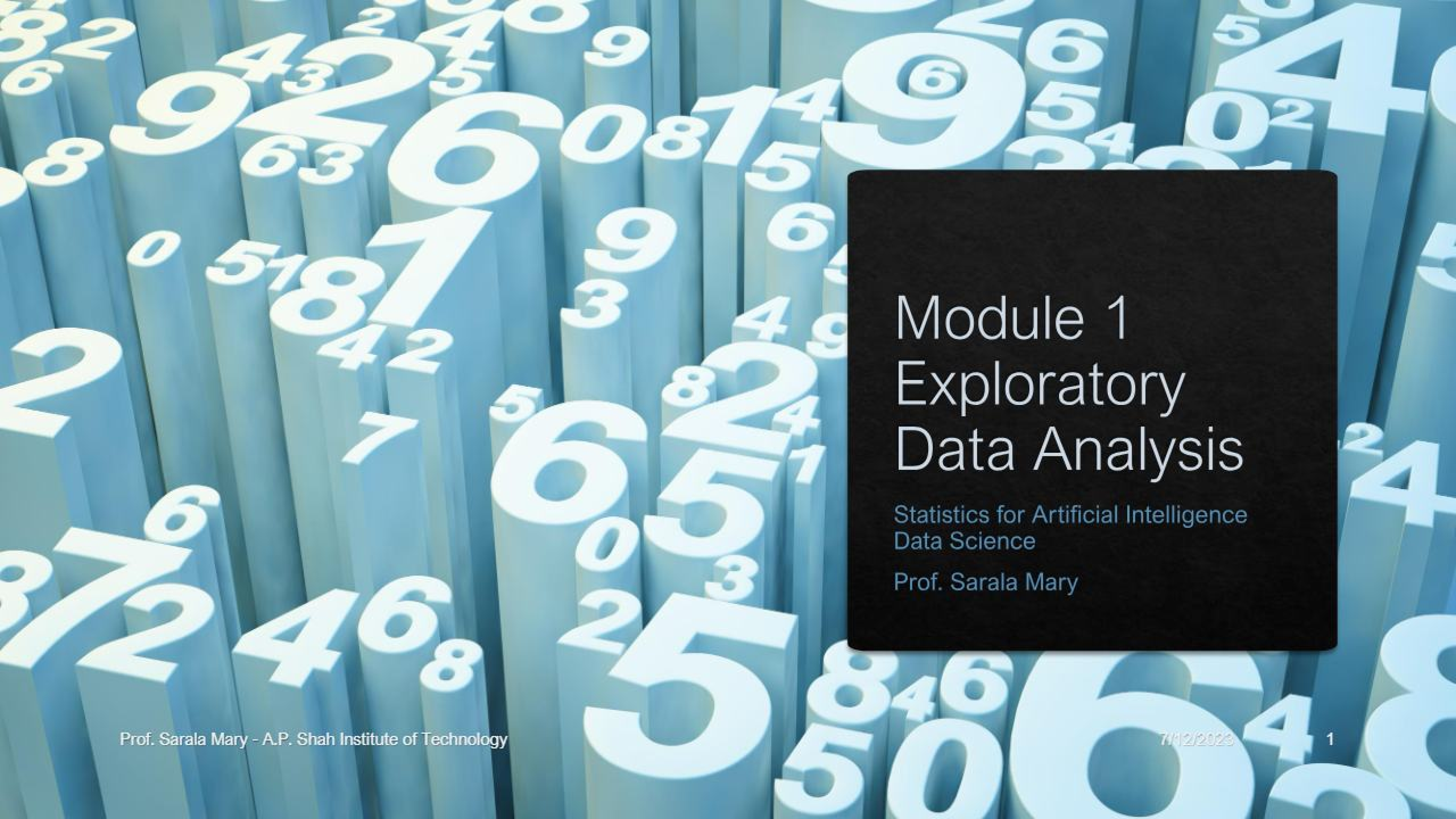
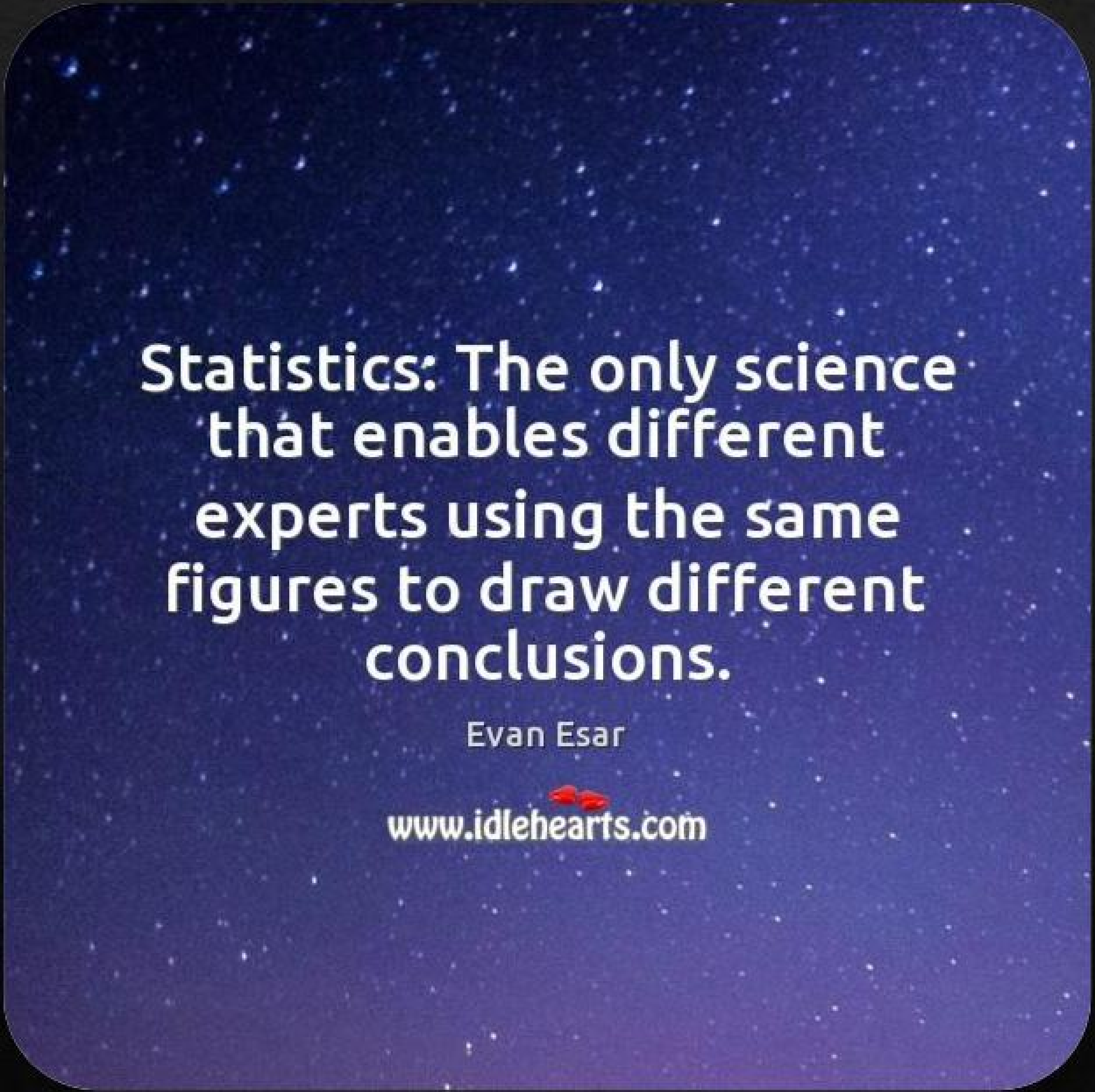


Module 1 Exploratory Data Analysis

Statistics for Artificial Intelligence Data Science

Prof. Sarala Mary



Statistics: The only science that enables different experts using the same figures to draw different conclusions.

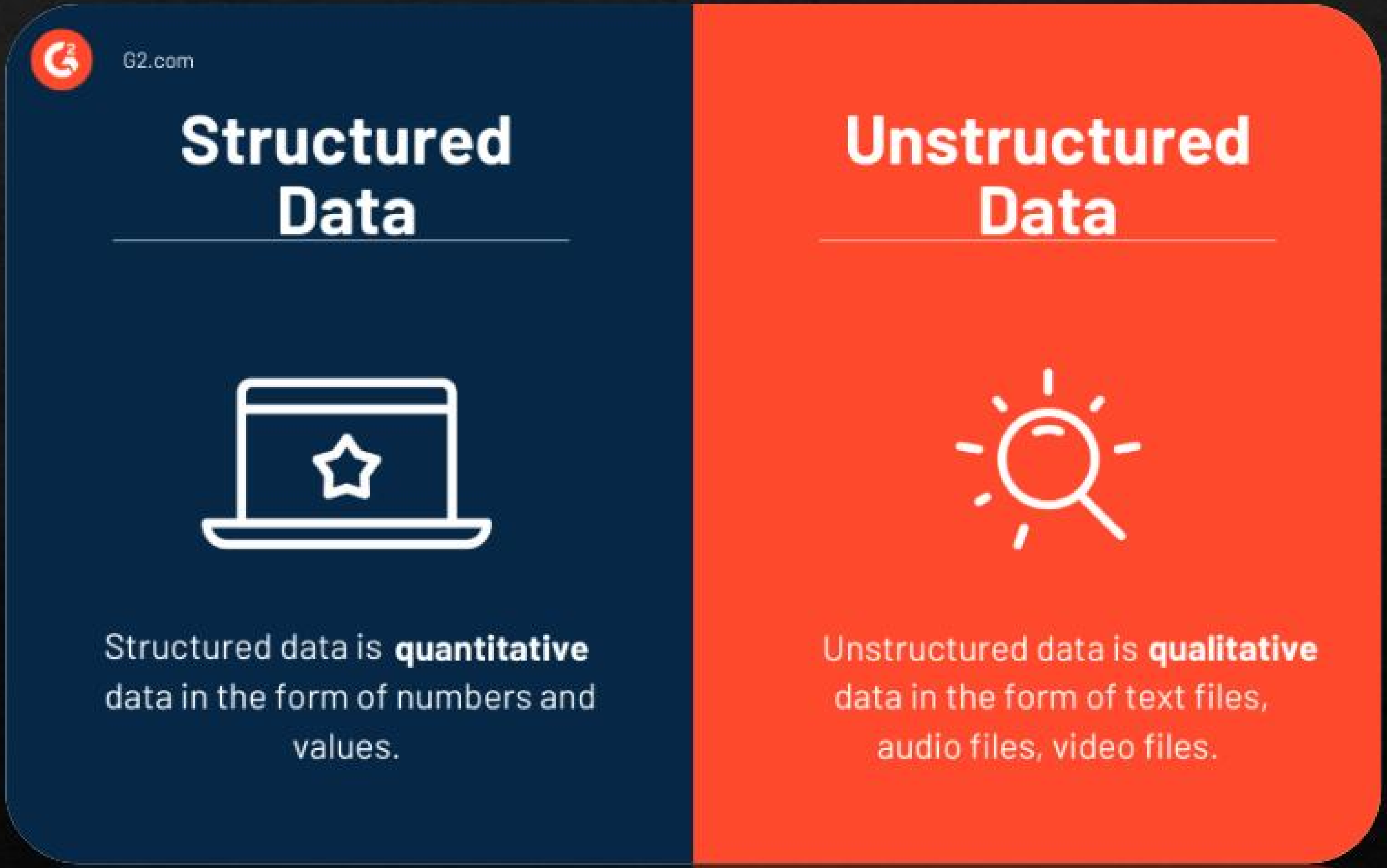
Evan Esar

www.idlehearts.com

What is Data?

- Data is a collection of facts, such as numbers, words, measurements, observations, or just descriptions of things.

Structured Data and Unstructured Data



G2.com

Structured Data

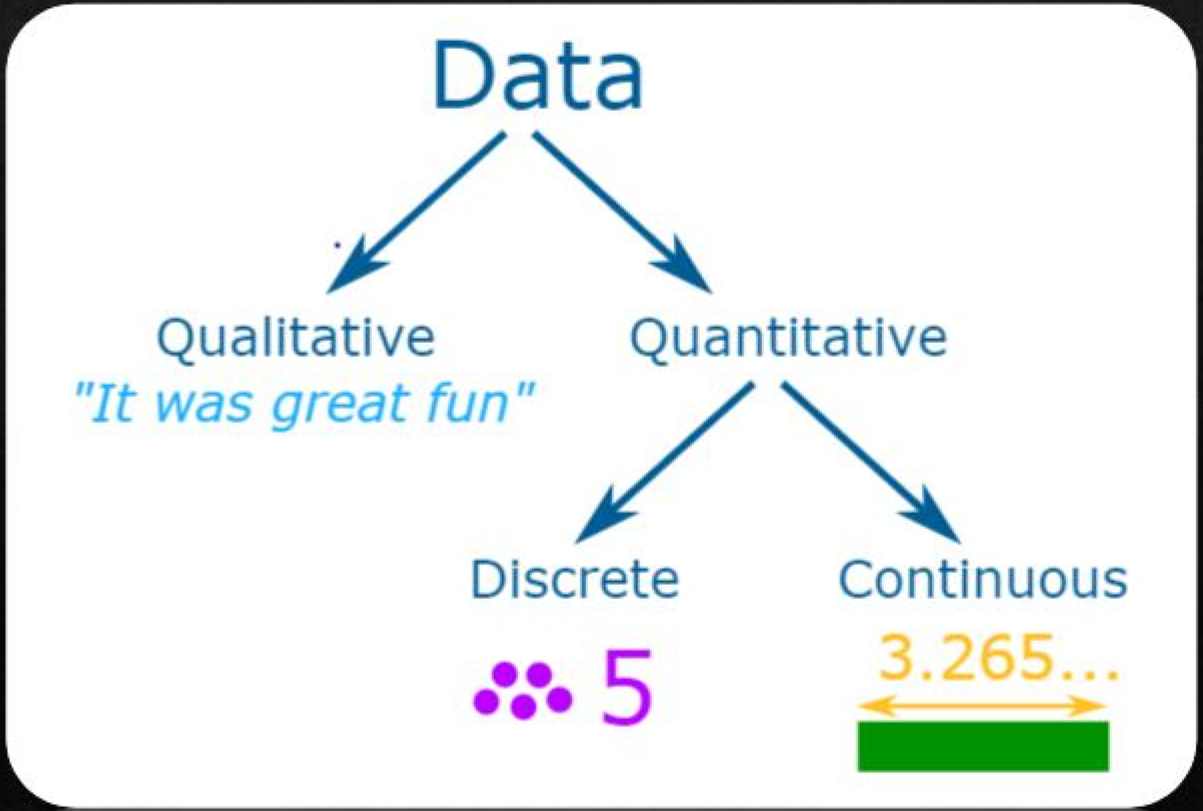
Structured data is **quantitative** data in the form of numbers and values.

Unstructured Data

Unstructured data is **qualitative** data in the form of text files, audio files, video files.

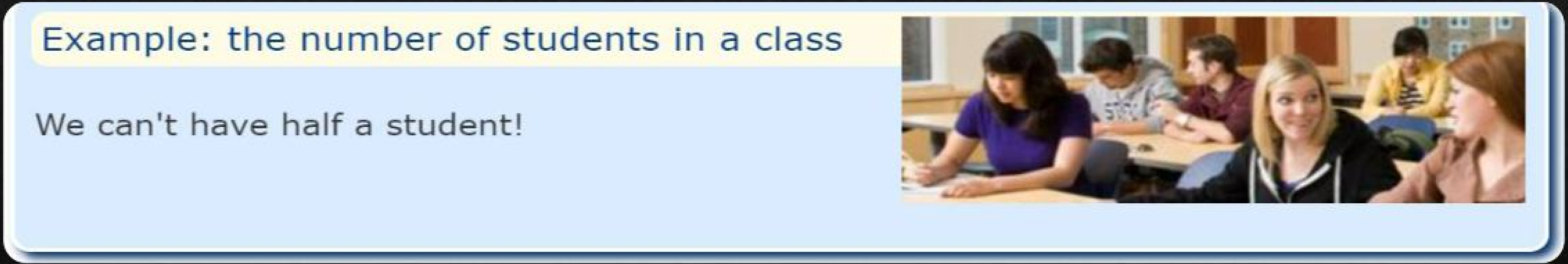
Qualitative vs Quantitative

- Qualitative data is descriptive information (it describes something)
- Quantitative data is numerical information (numbers)



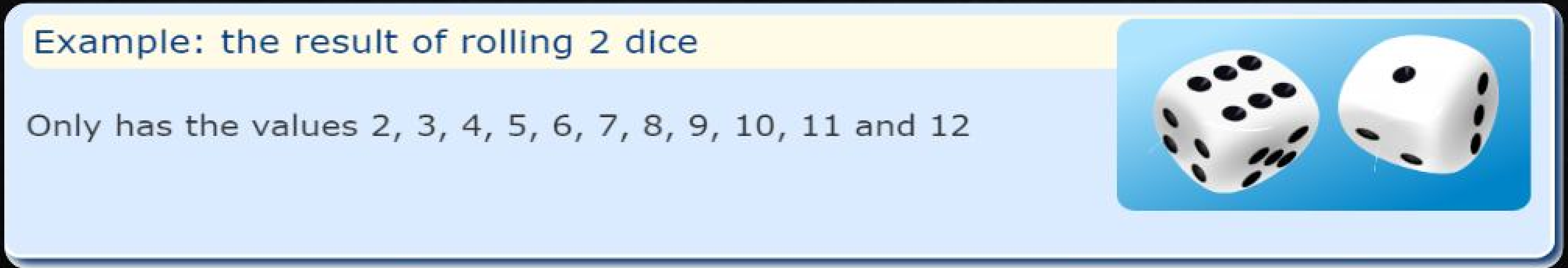
Discrete Data

- Discrete Data can only take certain values.



Example: the number of students in a class

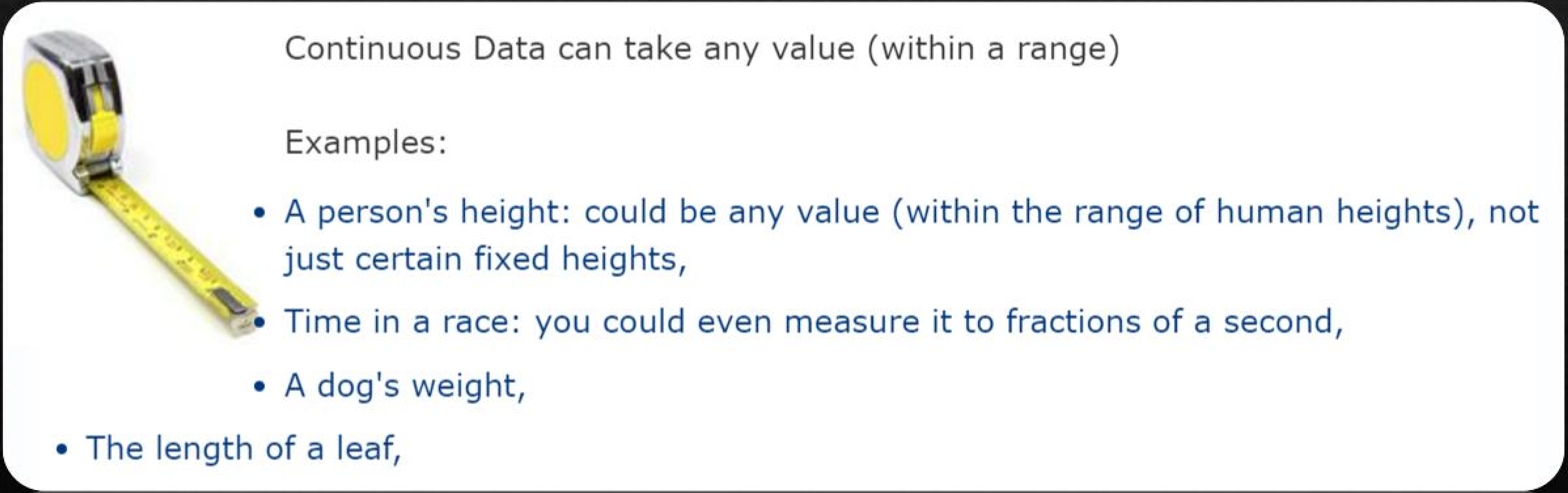
We can't have half a student!



Example: the result of rolling 2 dice

Only has the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12

Continuous Data

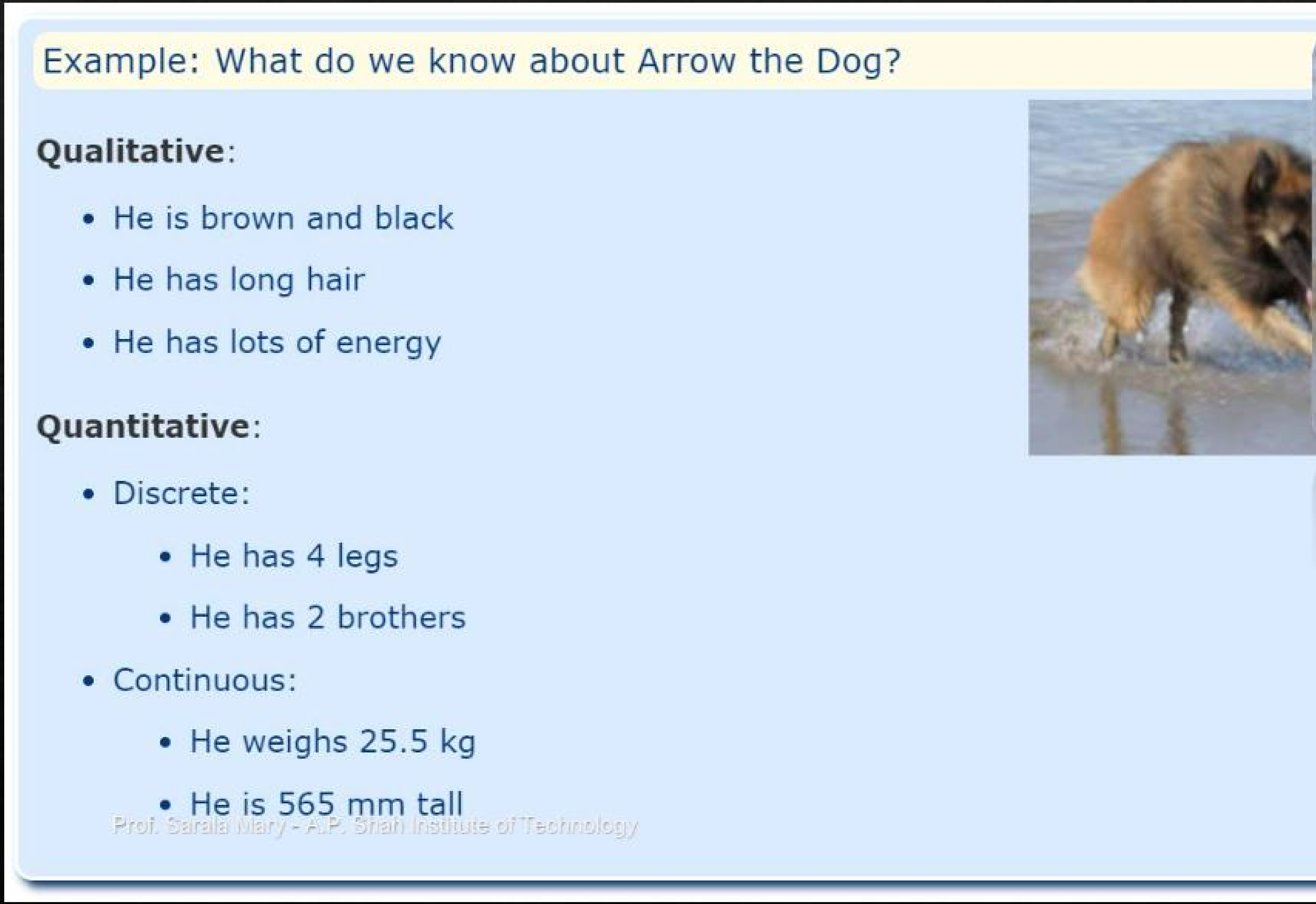


Continuous Data can take any value (within a range)

Examples:

- A person's height: could be any value (within the range of human heights), not just certain fixed heights,
- Time in a race: you could even measure it to fractions of a second,
- A dog's weight,
- The length of a leaf,

Example



Example: What do we know about Arrow the Dog?

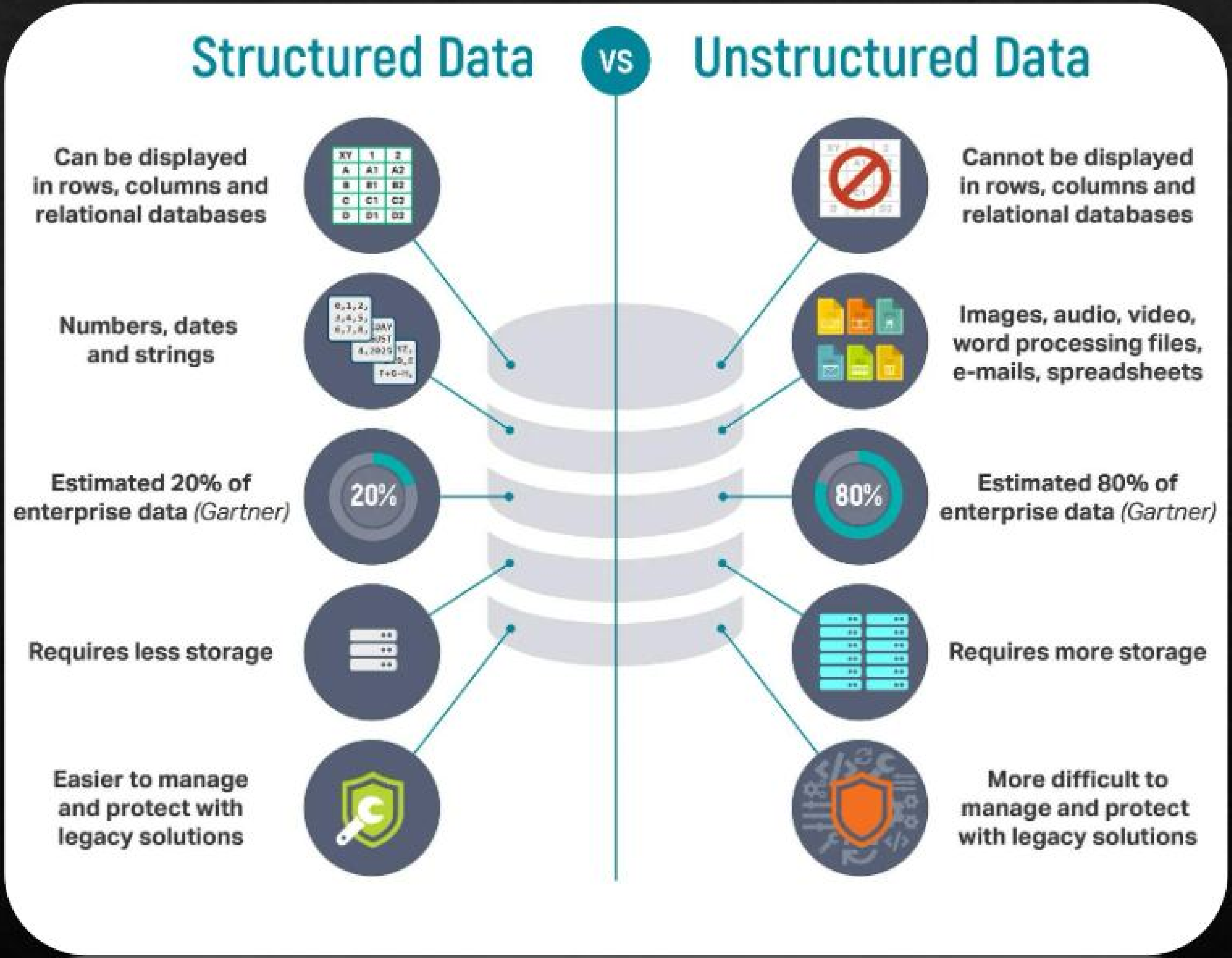
Qualitative:

- He is brown and black
- He has long hair
- He has lots of energy

Quantitative:

- Discrete:
  - He has 4 legs
  - He has 2 brothers
- Continuous:
  - He weighs 25.5 kg
  - He is 565 mm tall

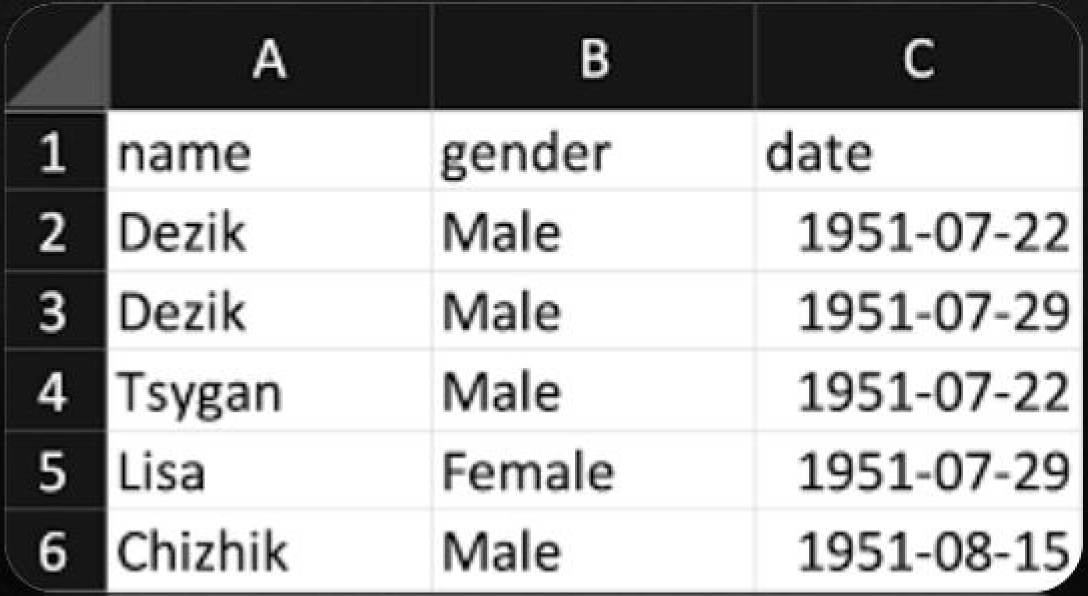
Structured Data and Unstructured Data



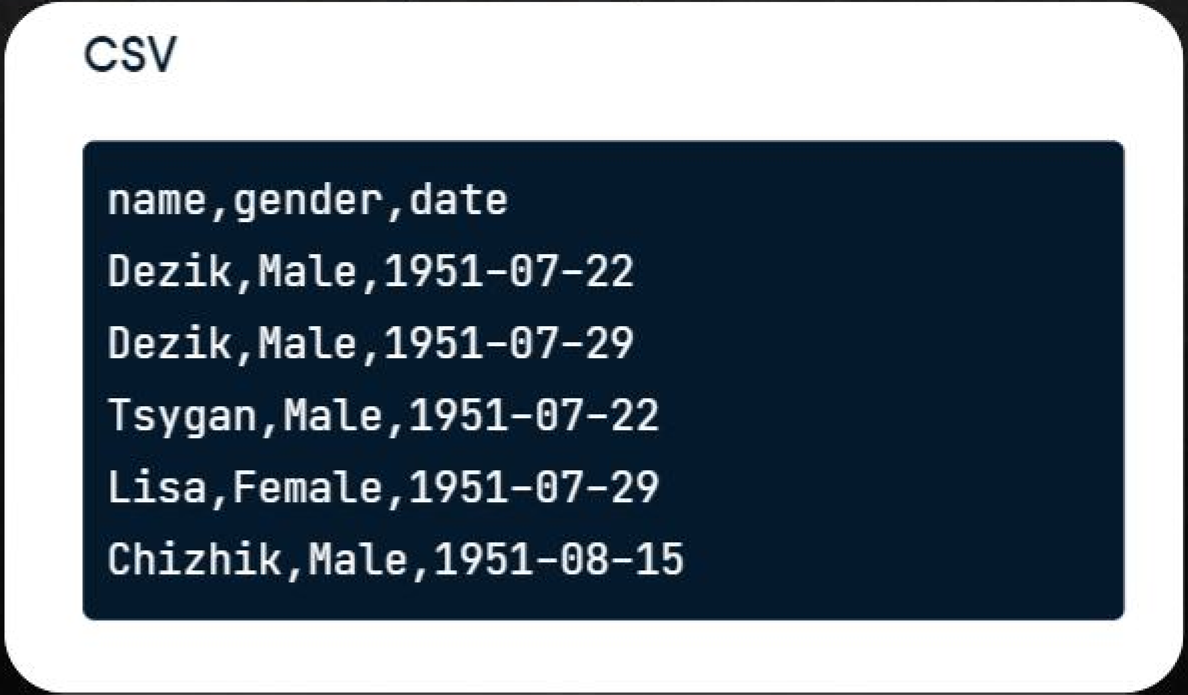
Structured Data	VS	Unstructured Data
Can be displayed in rows, columns and relational databases		Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings		Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*		Estimated 80% of enterprise data *(Gartner)*
Requires less storage		Requires more storage
Easier to manage and protect with legacy solutions		More difficult to manage and protect with legacy solutions

Rectangular Data

- Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables).



	A	B	C
1	name	gender	date
2	Dezik	Male	1951-07-22
3	Dezik	Male	1951-07-29
4	Tsygan	Male	1951-07-22
5	Lisa	Female	1951-07-29
6	Chizhik	Male	1951-08-15



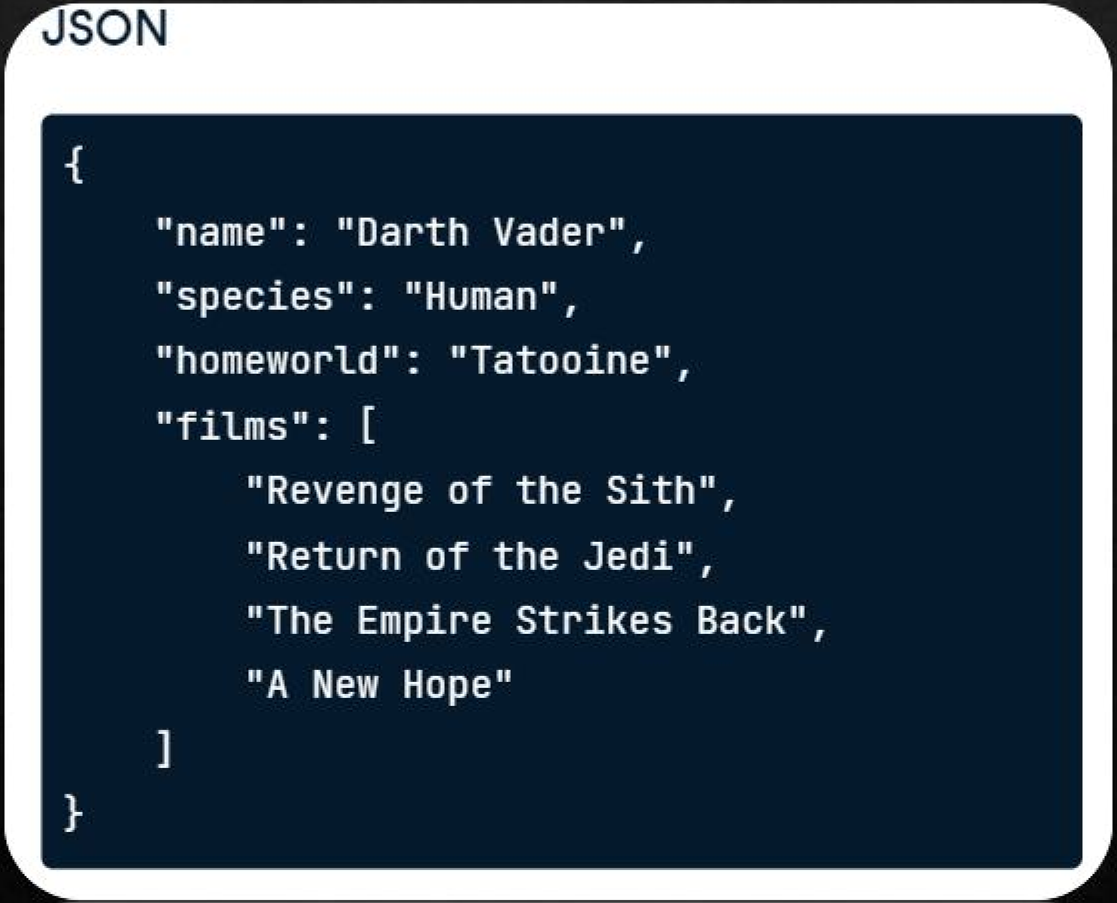
CSV

```
name,gender,date
Dezik,Male,1951-07-22
Dezik,Male,1951-07-29
Tsygan,Male,1951-07-22
Lisa,Female,1951-07-29
Chizhik,Male,1951-08-15
```

Key terms of Rectangular Data

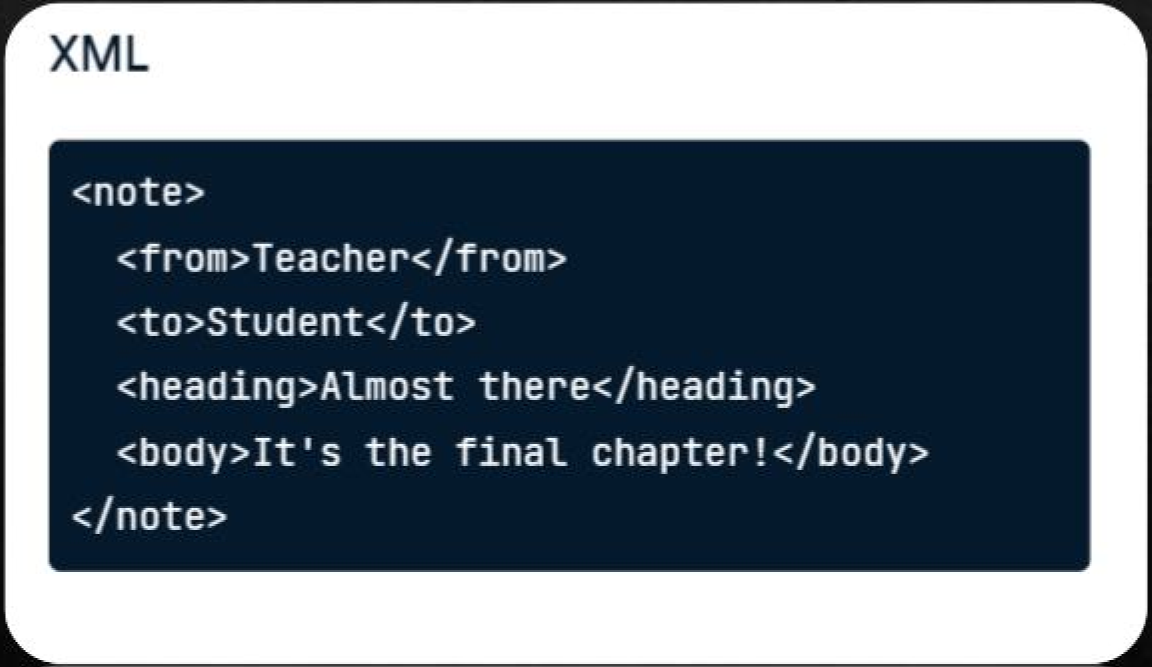
- Data frame - Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
- Feature - A column within a table is commonly referred to as a feature.
- Records - A row within a table is commonly referred to as a record.

Non – Rectangular Data



JSON

```
{
    "name": "Darth Vader",
    "species": "Human",
    "homeworld": "Tatooine",
    "films": [
        "Revenge of the Sith",
        "Return of the Jedi",
        "The Empire Strikes Back",
        "A New Hope"
    ]
}
```



XML

```
<note>
  <from>Teacher</from>
  <to>Student</to>
  <heading>Almost there</heading>
  <body>It's the final chapter!</body>
</note>
```