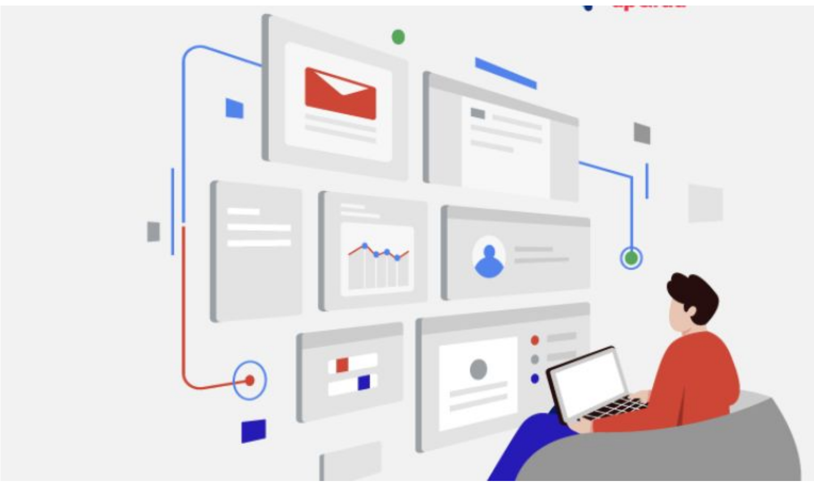
CIS 635 - 04 Knowledge Discovery & Data Mining

Data modalities - Introduction and Overview

Outline

- Everyday observed data and their modalities and types
- Data Trends, and where DS fits

Everyday observed (real life) data



Everyday observed (real life) data

- We have seen an **explosion of data** in recent days, especially in last few years!
- The transition from the Analog to the Digital World has made this Big Shift (was unimaginable a few decades ago)

Everyday observed (real life) data

- We have seen an **explosion of data** in recent days, especially in last few years!
- The transition from the Analog to the Digital World has made this Big Shift (was unimaginable a few decades ago)
- It’s very difficult to **segregate them into disjoint types** as data come in different **format** and in different **modalities**.

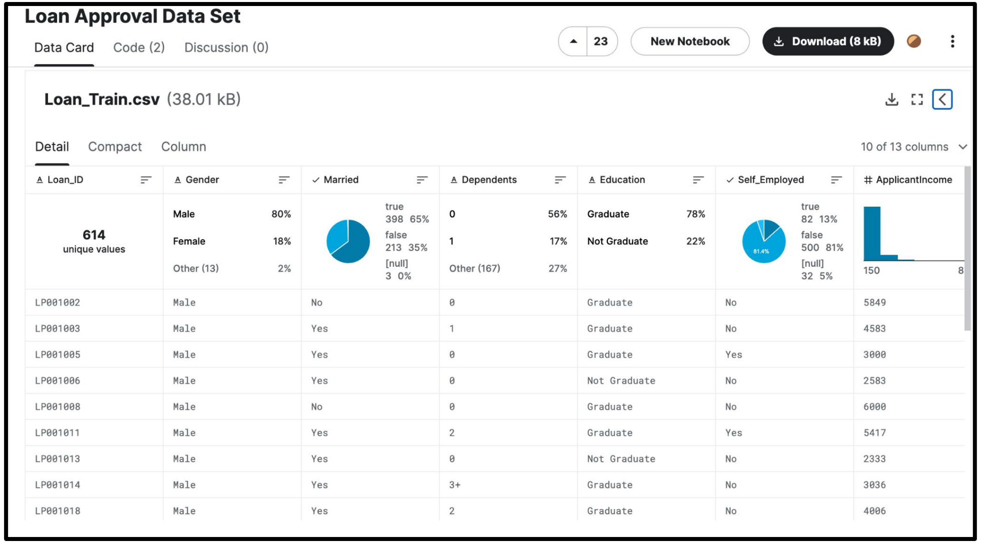
Everyday observed (real life) data

- We have seen an **explosion of data** in recent days, especially in last few years!
- The transition from the Analog to the Digital World has made this Big Shift (was unimaginable a few decades ago)
- It’s very difficult to **segregate them into different types** as data come in different **format** and in different **modalities**.
- We will try a **simplified approach** to **group** them and try to cover most (NOT ALL) data we perceive in our everyday life.

Everyday observed (real life) data

- Structured data
- Un(and Semi) structured data

Structured data



Loan Approval Data Set

Data Card | Code (2) | Discussion (0)

Loan_Train.csv (38.01 kB)

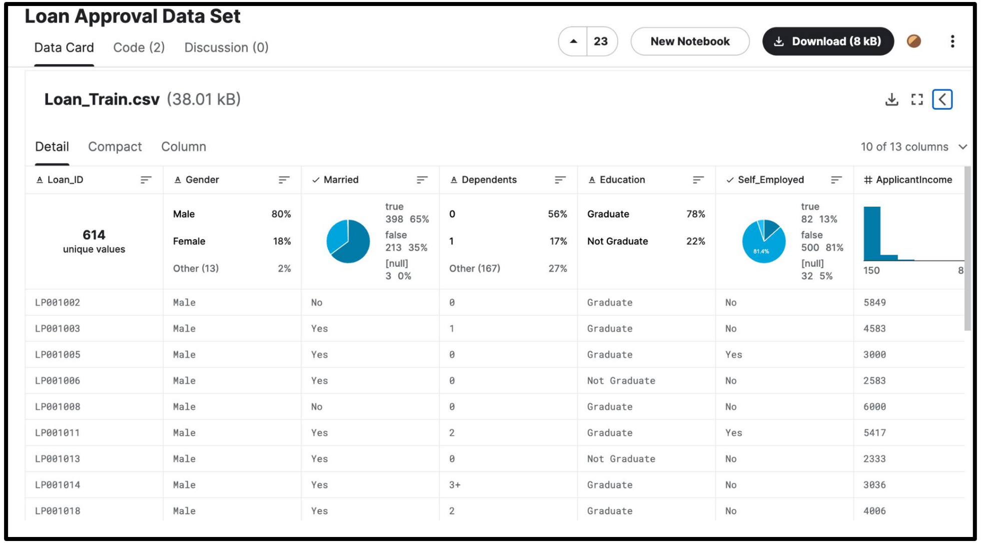
Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
614 unique values	Male 80% Female 18% Other (13) 2%	true 398 65% false 213 35% [null] 3 0%	0 56% 1 17% Other (167) 27%	Graduate 78% Not Graduate 22%	true 82 13% false 500 81% [null] 32 5%	150 8
LP001002	Male	No	0	Graduate	No	5849
LP001003	Male	Yes	1	Graduate	No	4583
LP001005	Male	Yes	0	Graduate	Yes	3000
LP001006	Male	Yes	0	Not Graduate	No	2583
LP001008	Male	No	0	Graduate	No	6000
LP001011	Male	Yes	2	Graduate	Yes	5417
LP001013	Male	Yes	0	Not Graduate	No	2333
LP001014	Male	Yes	3+	Graduate	No	3036
LP001018	Male	Yes	2	Graduate	No	4006

- Generally organized in **tables** and collected through filling **forms** (manual or online)
- Stored in **databases/spread -sheets** mainly
- Also popular the **.csv** file format

- Opening a bank account
- University registration
- Gmail
- Amazon account
- Your health profile

Kaggle loan approval dataset

Structured data



Loan Approval Data Set

Loan_Train.csv (38.01 kB)

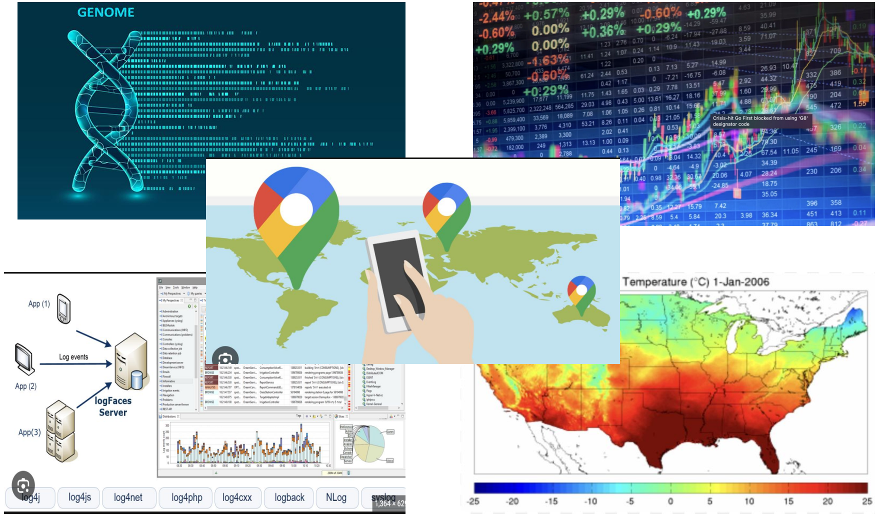
Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
614 unique values	Male 80% Female 18% Other (13) 2%	true 398 65% false 213 35% [null] 3 0%	0 56% 1 17% Other (167) 27%	Graduate 78% Not Graduate 22%	true 82 13% false 500 81% [null] 32 5%	150 8
LP001002	Male	No	0	Graduate	No	5849
LP001003	Male	Yes	1	Graduate	No	4583
LP001005	Male	Yes	0	Graduate	Yes	3000
LP001006	Male	Yes	0	Not Graduate	No	2583
LP001008	Male	No	0	Graduate	No	6000
LP001011	Male	Yes	2	Graduate	Yes	5417
LP001013	Male	Yes	0	Not Graduate	No	2333
LP001014	Male	Yes	3+	Graduate	No	3036
LP001018	Male	Yes	2	Graduate	No	4006

- Generally collected through **forms** (manual or online)
- Stored in **databases/spread-sheets** mainly
- Also popular the **.csv** file format

- Opening a **bank account**
- **University registration**
- **Gmail, Azure, and/or Amazon** account
- Your immigration, health, social media **profile**

Kaggle loan approval dataset

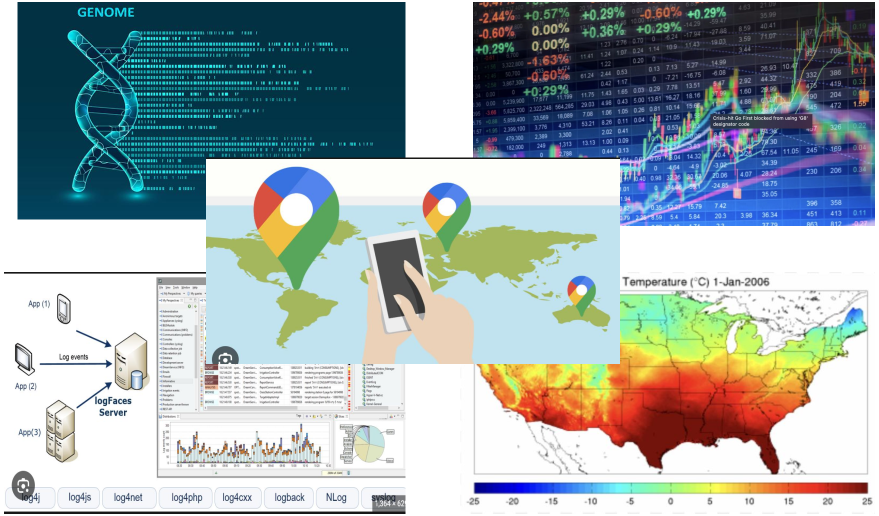
Structured data



- There also other formats, or you can convert them to
- Some are stand alone, while others are sequences and/or series
- Software generated logs

- Genomics data
- Stock prices
- Your CC history
- Weather data
- Google maps

Structured data

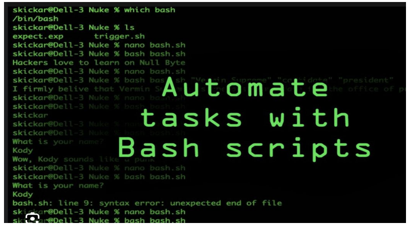


- There also other formats, or you can convert them to
- Software generated logs
- Some are stand alone, white others are sequences and/or series

- Genomics data
- Stock prices
- Your CC history
- Weather data
- Google maps

Un/Semi Structured data

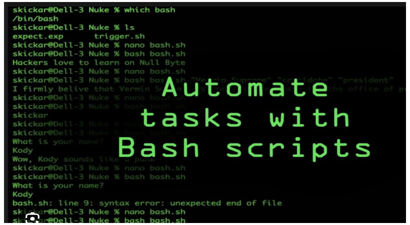
- Free forms
- Stored in data lake/warehouses mainly
- **Languages**: sequence of strings (semantics)
- **Audio**: Language + Acoustics; Music
- **Image** : Visual representation of the world
- **Video**: Sequence of images



```
skickar@Dell-3 Nuke % which bash
/bin/bash
skickar@Dell-3 Nuke % ls
expect.exp      trigger.sh
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
Hackers love to learn on Null Byte
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh "Vermin Supreme" "candidate" "president"
I firmly belive that Vermin Supreme is the best candidate for the office of p
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
skickar
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
Wow, Kody sounds like a punk
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
bash.sh: line 9: syntax error: unexpected end of file
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
```

- Some are sequence, while others are stand alone
- Social media data (emotions, vives)

Un/Semi Structured data



```
skickar@Dell-3 Nuke % which bash
/bin/bash
skickar@Dell-3 Nuke % ls
expect.exp      trigger.sh
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
Hackers love to learn on Null Byte
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh "Vermin Supreme" "candidate" "president"
I firmly belive that Vermin S[illegible]the office of p
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
skickar
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
Wow, Kody sounds like a punk
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
bash.sh: line 9: syntax error: unexpected end of file
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
```

- Free forms
- Stored in data lake mainly?
- Languages: sequence of strings
- Audio: Language + Acoustics; Music
- Image : Visual representation of the world
- Video: Sequence of images

- Some are sequences, while others are stand alone
- Social media data (emotions, vives)

Un/Semi Structured data

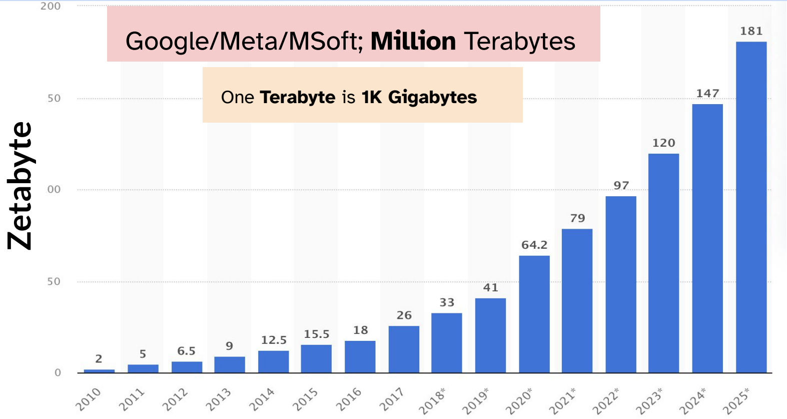
- Free forms
- Stored in data lake mainly?
- Languages: sequence of strings
- Audio: Language + Acoustics; Music
- Image : Visual representation of the world
- Video: Sequence of images

- Some are sequence, while others are stand alone
- **Social media** data (discussions, messages, emotions, vives)

Data Growth and the Trend

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**



Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

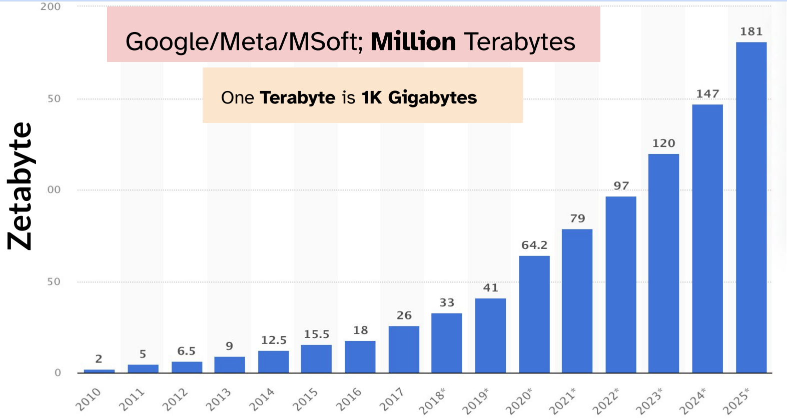
Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Exponential growth**
  - ○ ***Before 2010 collectively relatively close to Zero***
- Mainly due to
  - ○ IoT sensors
  - ○ Social media
  - ○
- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

statista

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**



Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

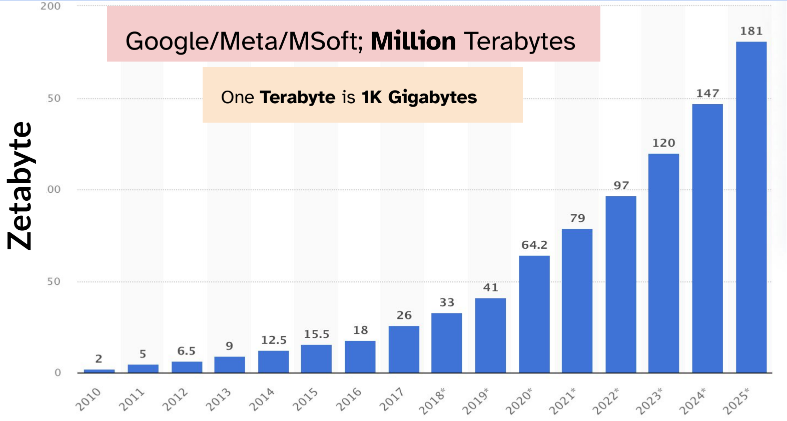
Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Exponential growth**
  - ***Before 2010 collectively relatively close to Zero***
- Mainly due to
  - IoT sensors
  - Social media
  -
- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

statista

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**



Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

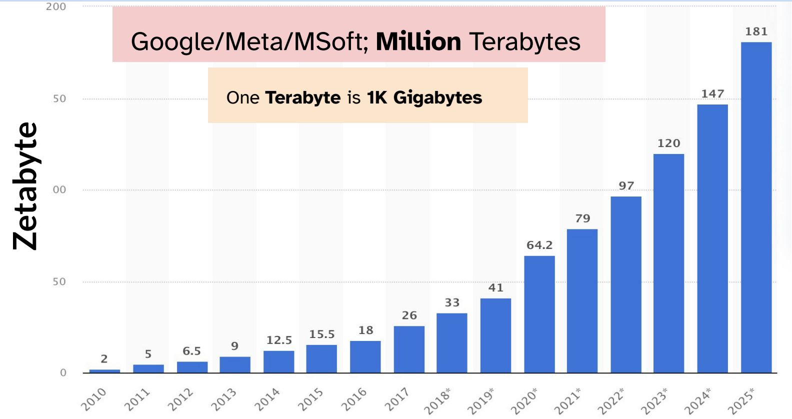
[Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025]

- Exponential growth
  - ○ *Before 2010 colle*
    close to Zero
- **Mainly due to**
  - ○ **Social media**
  - ○ IoT sensors
  - ○
- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

statista

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**



Google/Meta/MSoft; **Million** Terabytes

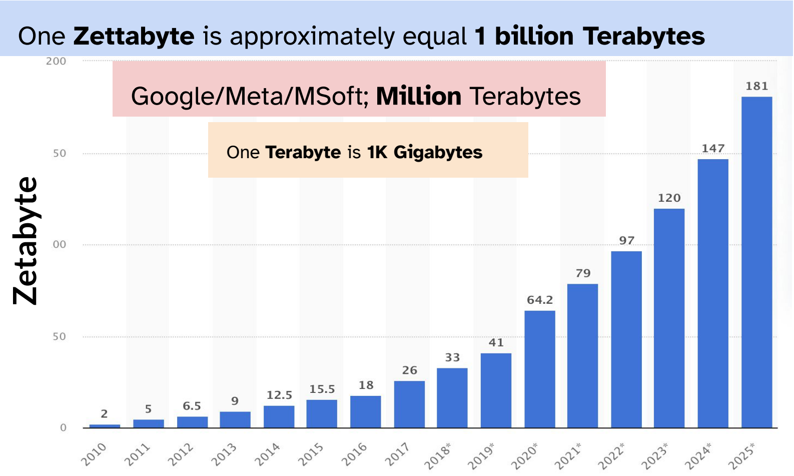
One **Terabyte** is **1K Gigabytes**

[Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025]

- Exponential growth
  - *Before 2010 coll close to Zero*
- **Mainly due to**
  - Social media
  - **IoT sensors**
- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

statista

Data Growth and the Trend



One **Zettabyte** is approximately equal **1 billion Terabytes**

Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

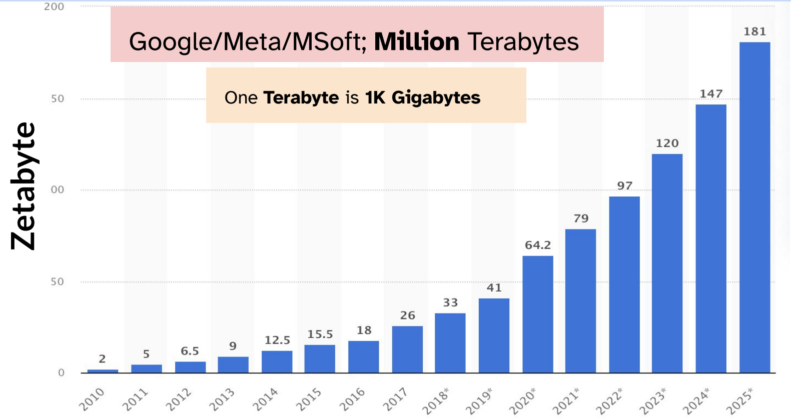
[Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025]

- Exponential growth
    - *Before 2010 collectively relatively close to Zero*
- **Mainly due to**
    - Social media
    - **IoT sensors**
- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

statista

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**

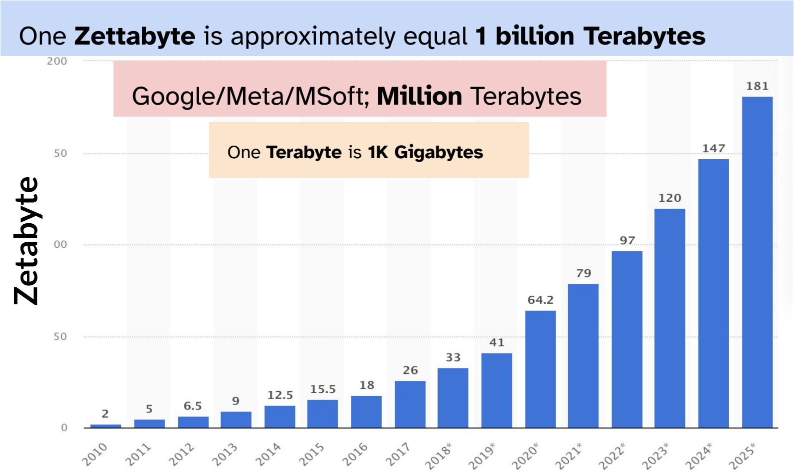

Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Data Science has emerged as an important discipline**
- We need to tackle new challenges, and many more are yet to come
- We will be learn about modern day **pioneers** in our class moving forward
- However, most the basic principles relates to basics of Math, Statistics and Probability theories. **Tribute** to

Data Growth and the Trend



One **Zettabyte** is approximately equal **1 billion Terabytes**

Google/Meta/MSoft; **Million** Terabytes

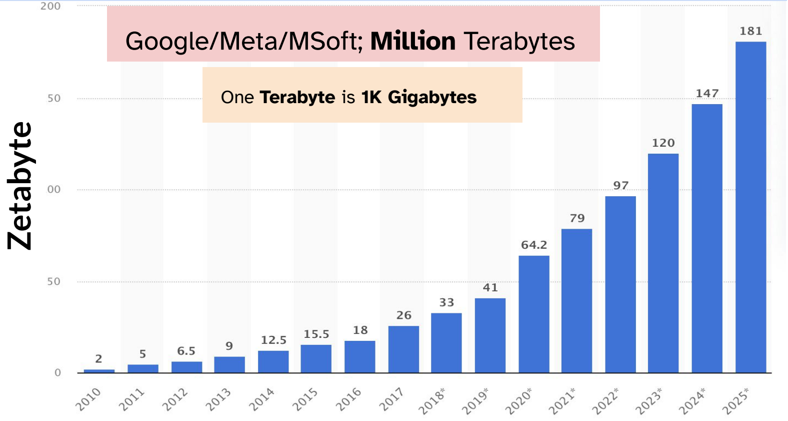
One **Terabyte** is **1K Gigabytes**

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Data Science has emerged as an important discipline**
- We need to tackle new challenges, and many more are yet to come
- We will be learn about modern day **pioneers** in our class moving forward
- However, most the basic principles relates to basics of Math, Statistics and Probability theories. **Tribute** to

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**



Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

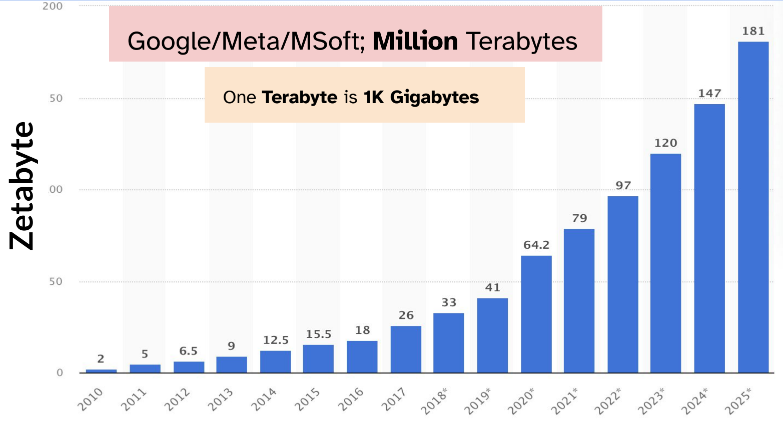
- **Data Science has emerged as an important discipline**
- We need to tackle new challenges, and many more are yet to come
- We will be learn about modern day **pioneers** in our class moving forward
- However, most the basic principles relates to basics of Math, Statistics and Probability theories. **Tribute** to

Data Growth and the Trend

One **Zettabyte** is approximately equal **1 billion Terabytes**

Google/Meta/MSoft; **Million** Terabytes

One **Terabyte** is **1K Gigabytes**



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Data Science** has emerged as an important discipline
- We need to tackle new challenges, and many more are yet to come
- We will be learn about modern day **pioneers** in our class moving forward
- However, most the basic principles relates to basics of Math, Statistics and Probability theories. **Tribute** to

Bernoulli's *Ars Conjectandi* was the first work that dealt with probability theory as currently understood.

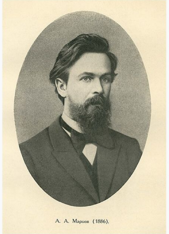
Carl Friedrich Gauss made major contributions to probabilistic methods leading to statistics;

Thomas Bayes

ortrait purportedly of Bayes used in a 1936 ok,[1] but it is doubtful whether the portrait is ually of him.[2] No earlier portrait or claimed portrait survives.

Gerolamo Cardano (16th century)

Andrey Andreyevich Markov



Markov in 1886

QA