



# **CIS 635 Knowledge Discovery & Data Mining**

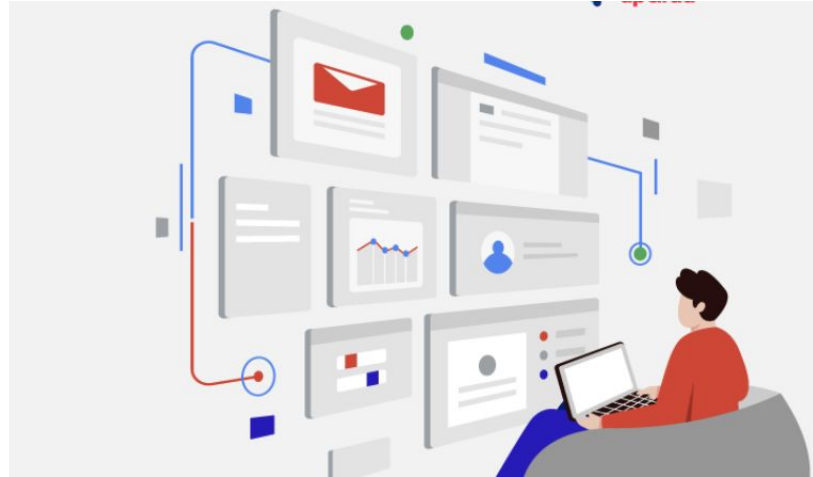
Data modalities - Introduction and Overview



# 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)



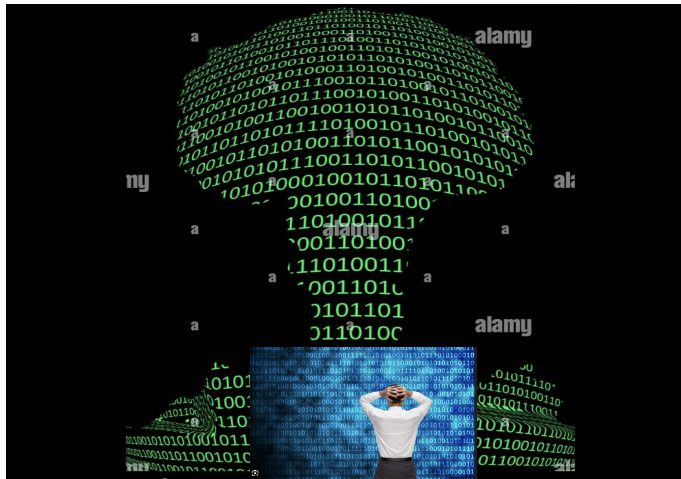
# Everyday observed (real life) data



- 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)

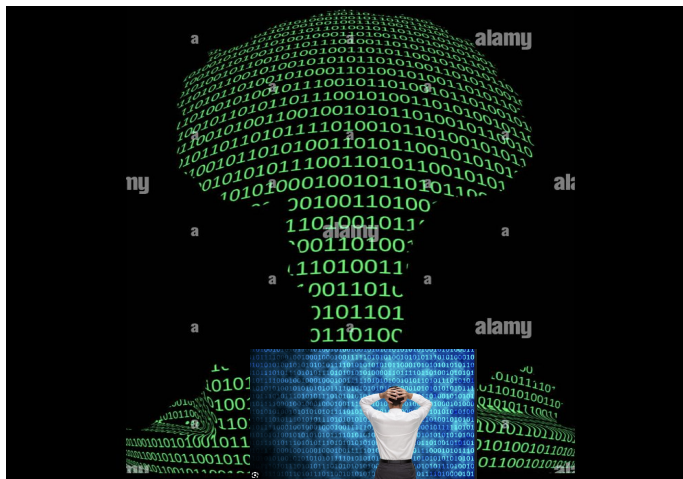


# Everyday observed (real life) data



- 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)

Detail Compact Column 10 of 13 columns

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
614 unique values	Male 80%	<div><div>398 65%</div><div>true</div></div>	0 56%	Graduate 78%	<div><div>82 13%</div><div>true</div></div>	<div><div>150</div><div>8</div></div>
	Female 18%	<div><div>213 35%</div><div>false</div></div>	1 17%	Not Graduate 22%	<div><div>500 81%</div><div>false</div></div>	
	Other (13) 2%	<div><div>3 0%</div><div>[null]</div></div>	Other (167) 27%		<div><div>32 5%</div><div>[null]</div></div>	
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**

Data Card Code (2) Discussion (0)

Loan\_Train.csv (38.01 kB)

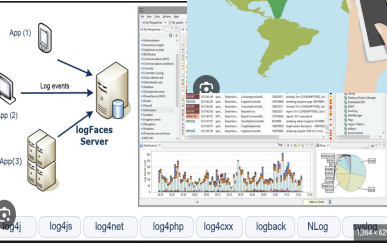
Detail Compact Column 10 of 13 columns

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
614 unique values	Male 80%	<div><div></div> true 398 65%</div>	0 56%	Graduate 78%	<div><div></div> true 82 13%</div>	<div><div></div> 150 8</div>
	Female 18%	<div><div></div> false 213 35%</div>	1 17%	Not Graduate 22%	<div><div></div> false 500 81%</div>	
	Other (13) 2%	<div><div></div> [null] 3 0%</div>	Other (167) 27%		<div><div></div> [null] 32 5%</div>	
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

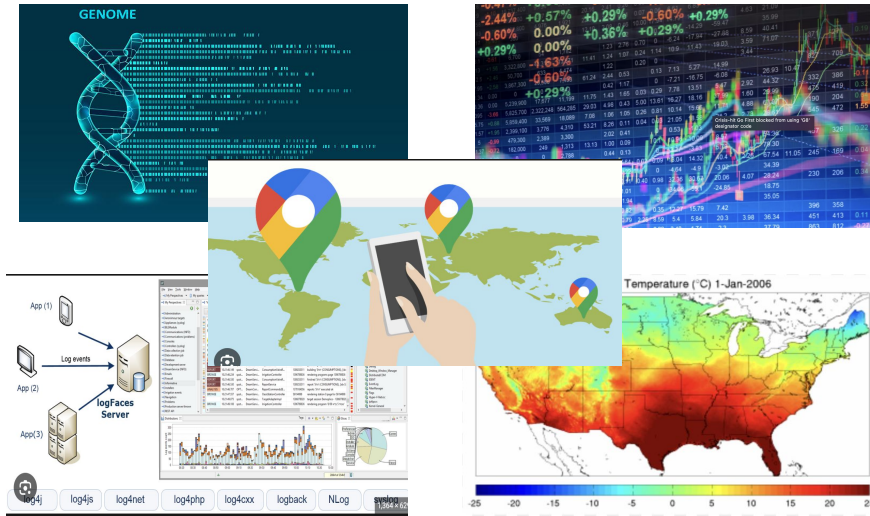
- Generally collected through **forms** (manual or online)
- Stored in **databases/spread-sheets** mainly
- Also popular the **.csv** file format

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

[Kaggle loan approval dataset](#)



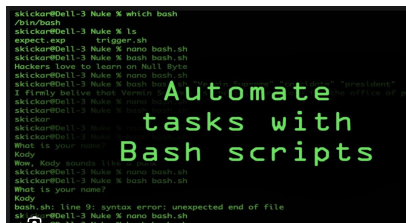
# Structured data



- There also other formats, or you can convert them to
- 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





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

# 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)

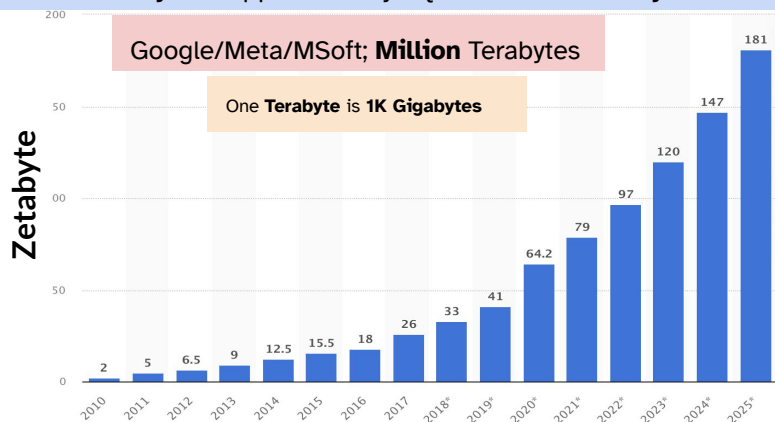


## Data Growth and the Trend



# Data Growth and the Trend

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

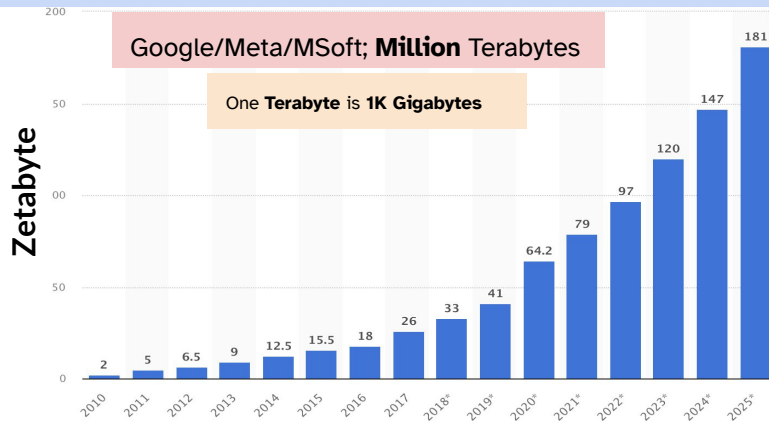


Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- **Exponential growth**
  - ***Before 2010 collectively relatively close to Zero Zetabyte***
- **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)

# Data Growth and the Trend

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



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

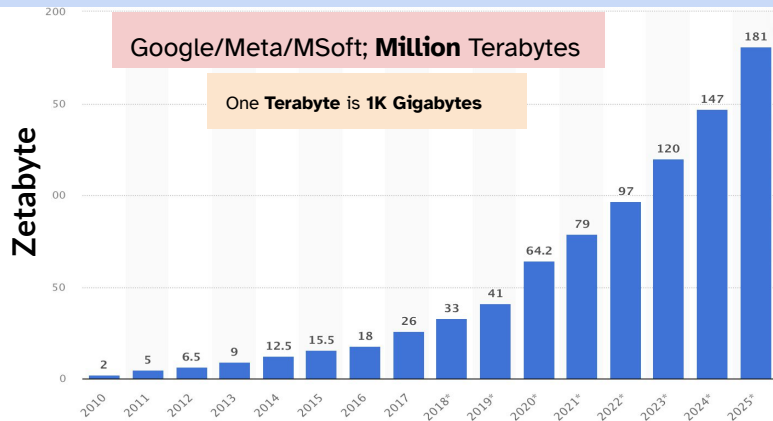
- Exponential growth
  - Before 2010 collected data close to Zero
- Mainly due to
  - Social media
  - IoT sensors



- Forecast might be underestimated as we expect more due to the recent Generative AI/LLM releases (ChatGPT for an example)

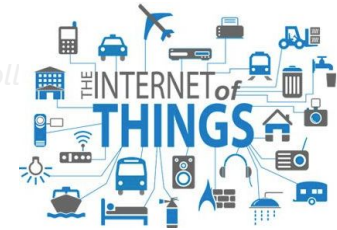
# Data Growth and the Trend

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



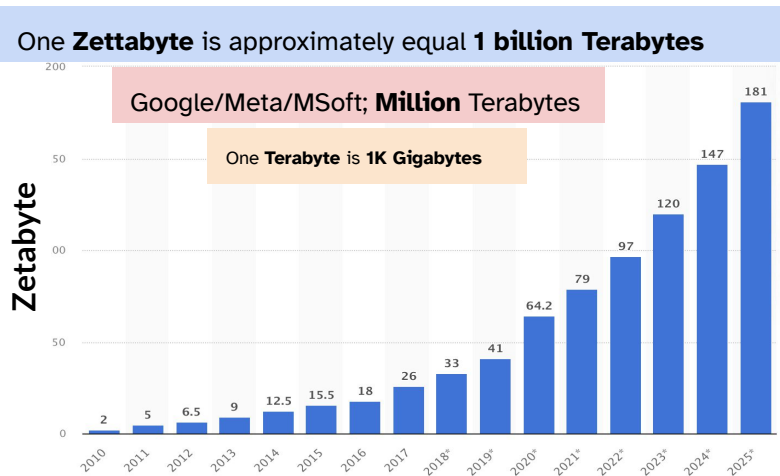
Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- 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)

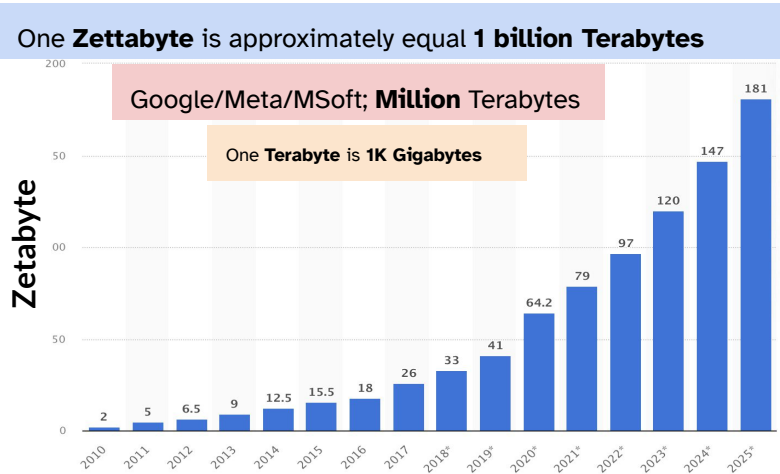
# Data Growth and the Trend



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)

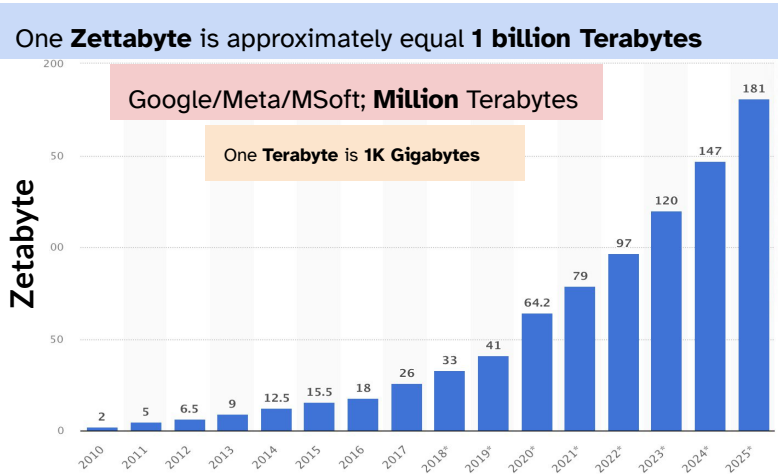
# Data Growth and the Trend



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

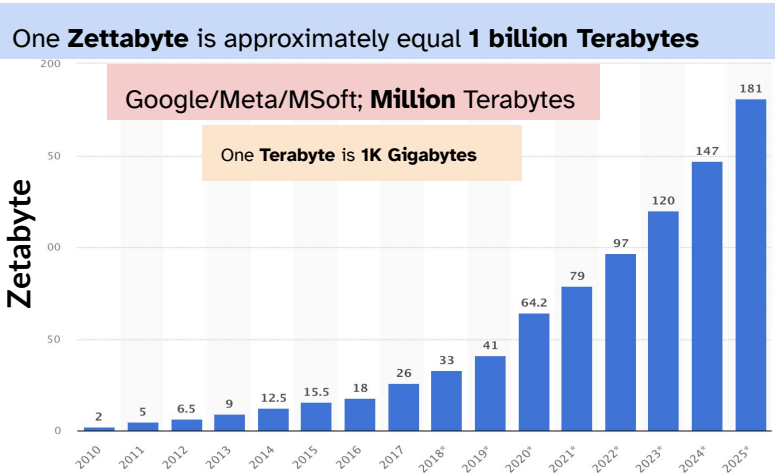
# Data Growth and the Trend



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



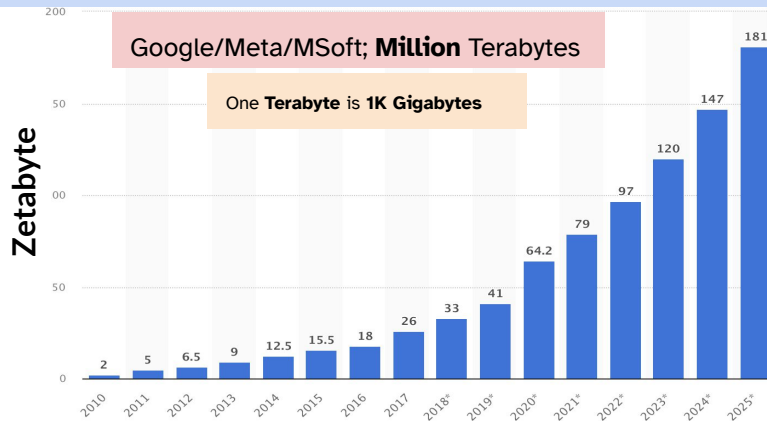
Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

- 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



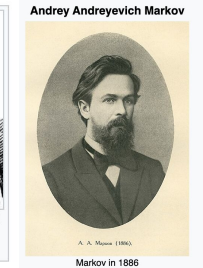
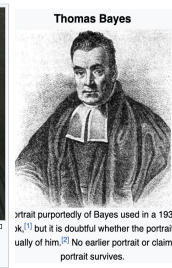
# Data Growth and the Trend

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



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





**QA**