From Raw Data to AI Gold: Crafting the Perfect Data Pipeline

Whats the big deal about AI anyway?

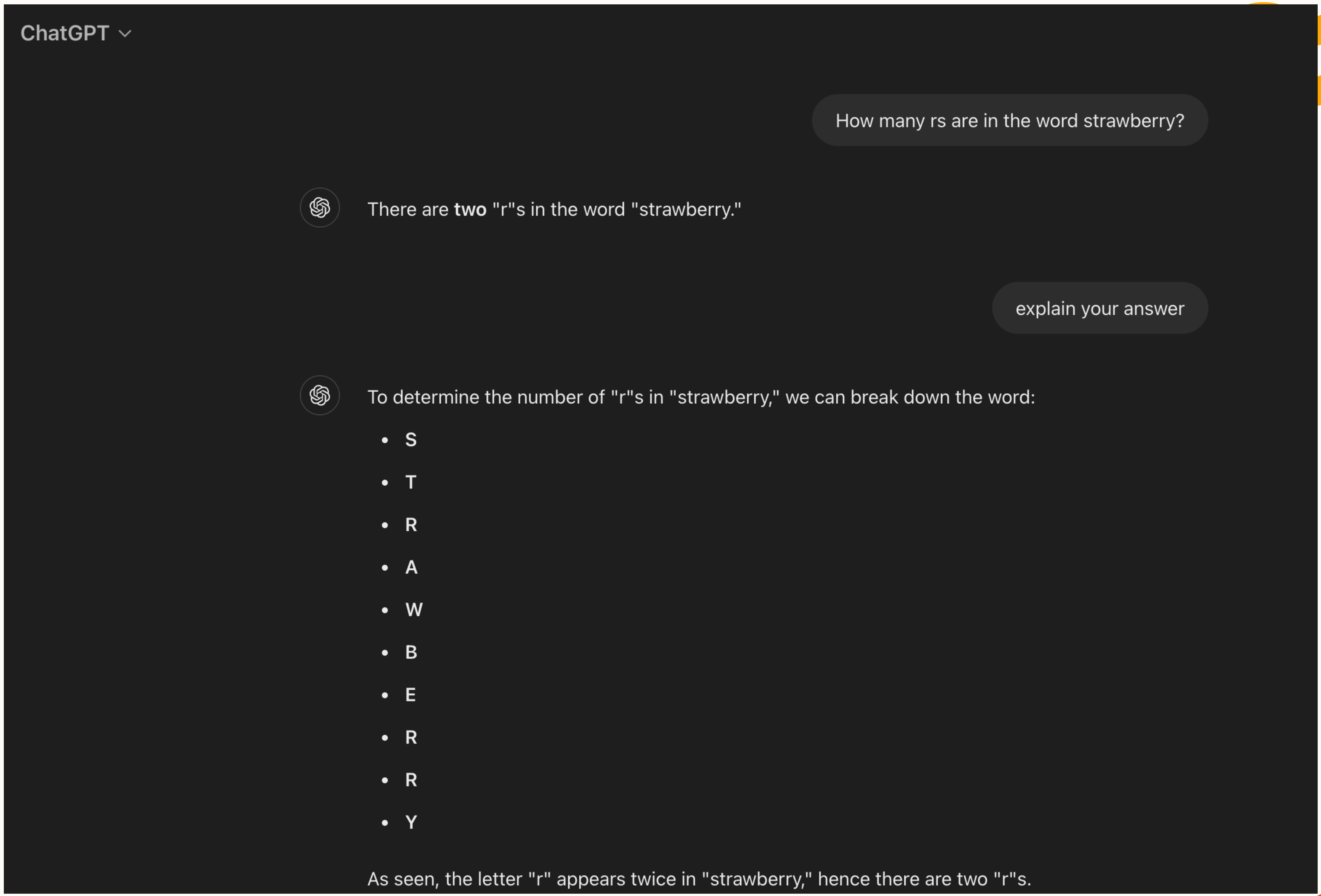
Music: அருகிலே

Volvo AI ad

Creativity and time savings on your behalf!

But what if AI takes over?

Not so fast...



How many rs are in the word strawberry?

There are **two** "r"s in the word "strawberry."

explain your answer

To determine the number of "r"s in "strawberry," we can break down the word:

- S
- T
- R
- A
- W
- B
- E
- R
- R
- Y

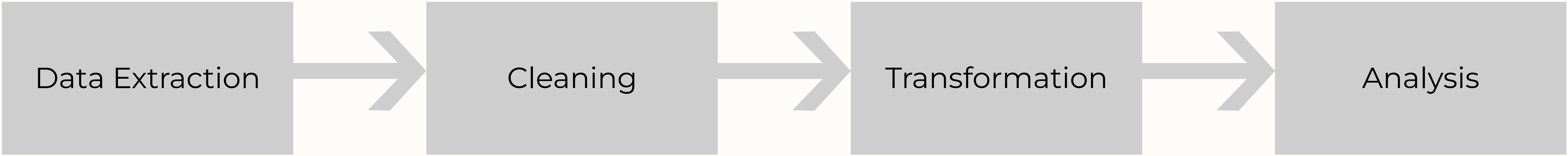
As seen, the letter "r" appears twice in "strawberry," hence there are two "r"s.

So what do we do about it? Welcome Data the Superhero!

But raw data is messy...
(like her hair!)
and unstructured :(

And so...behind every great AI application is a solid data pipeline!

What is a Data Pipeline?



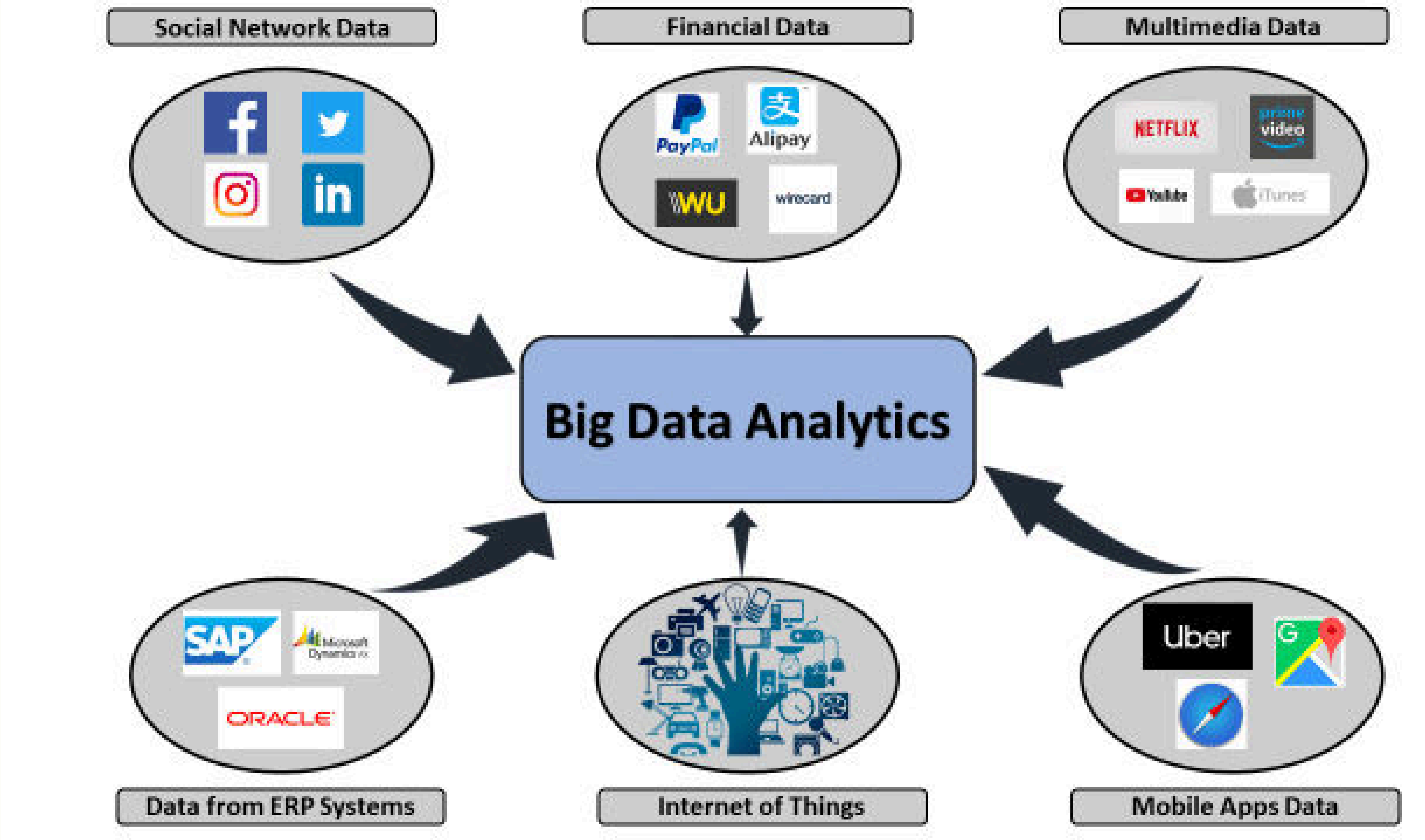
Data Extraction

Cleaning

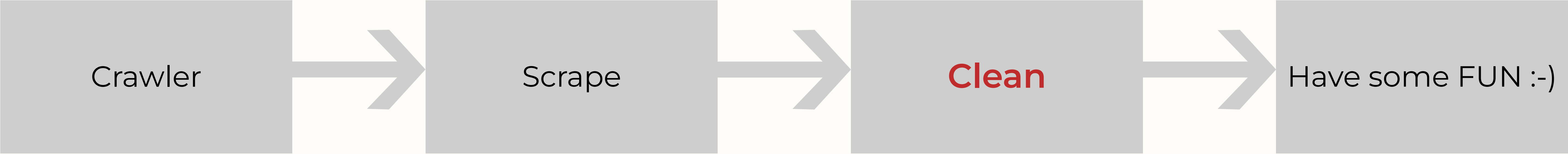
Transformation

Analysis

Data Extraction: Data Sources



Scrape, Clean and Transform



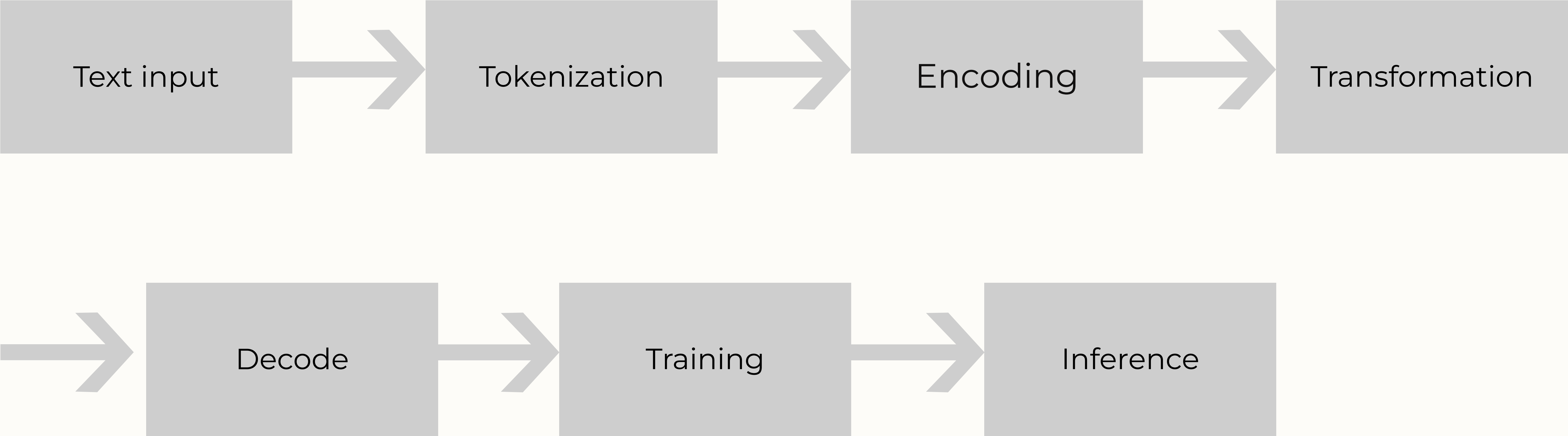
Crawler

Scrape

Clean

Have some FUN :-)

✳ What happens in a LLM?



Text input → Tokenization → Encoding → Transformation

→ Decode → Training → Inference

References

https://github.com/gayatrib/DataPipeline/tree/main

https://github.com/kanishk307/IndianFoodDatasetGeneration/blob/main/Code/

https://github.com/gayatrib/CS410Project/blob/master/CS420Project.pptx

https://cobusgreyling.medium.com/run-a-small-language-model-slm-local-offline-1f62a6cbdaef

Thanks!

Do you have any questions?

gayatri.sriram@gmail.com
+91 98868 90210

LinkedIn