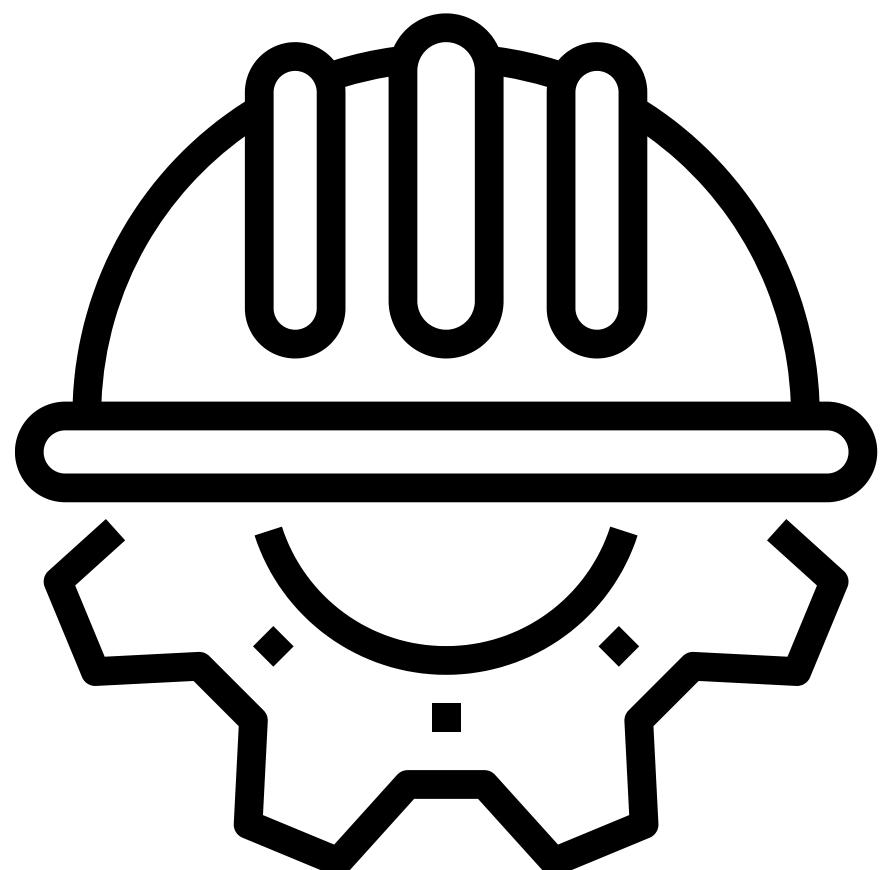


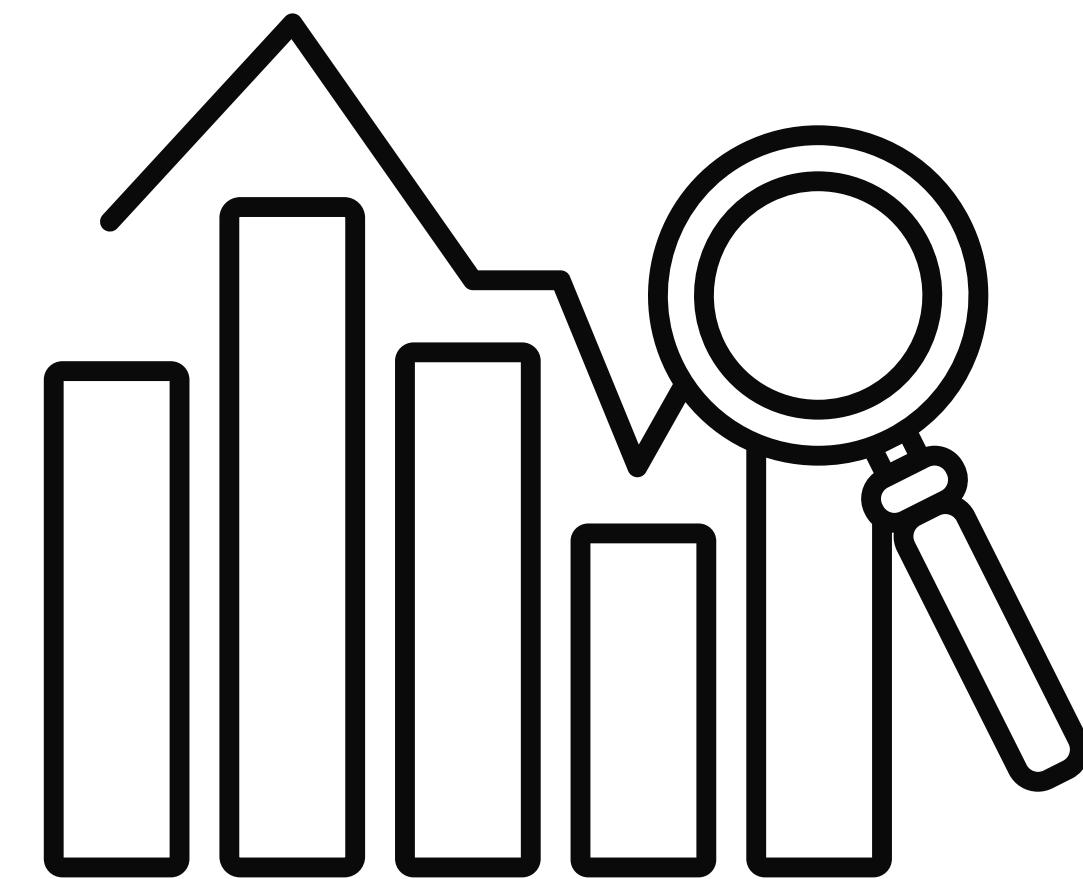
END TO END DATA PROCESS

ENG MARYAM SHERIF

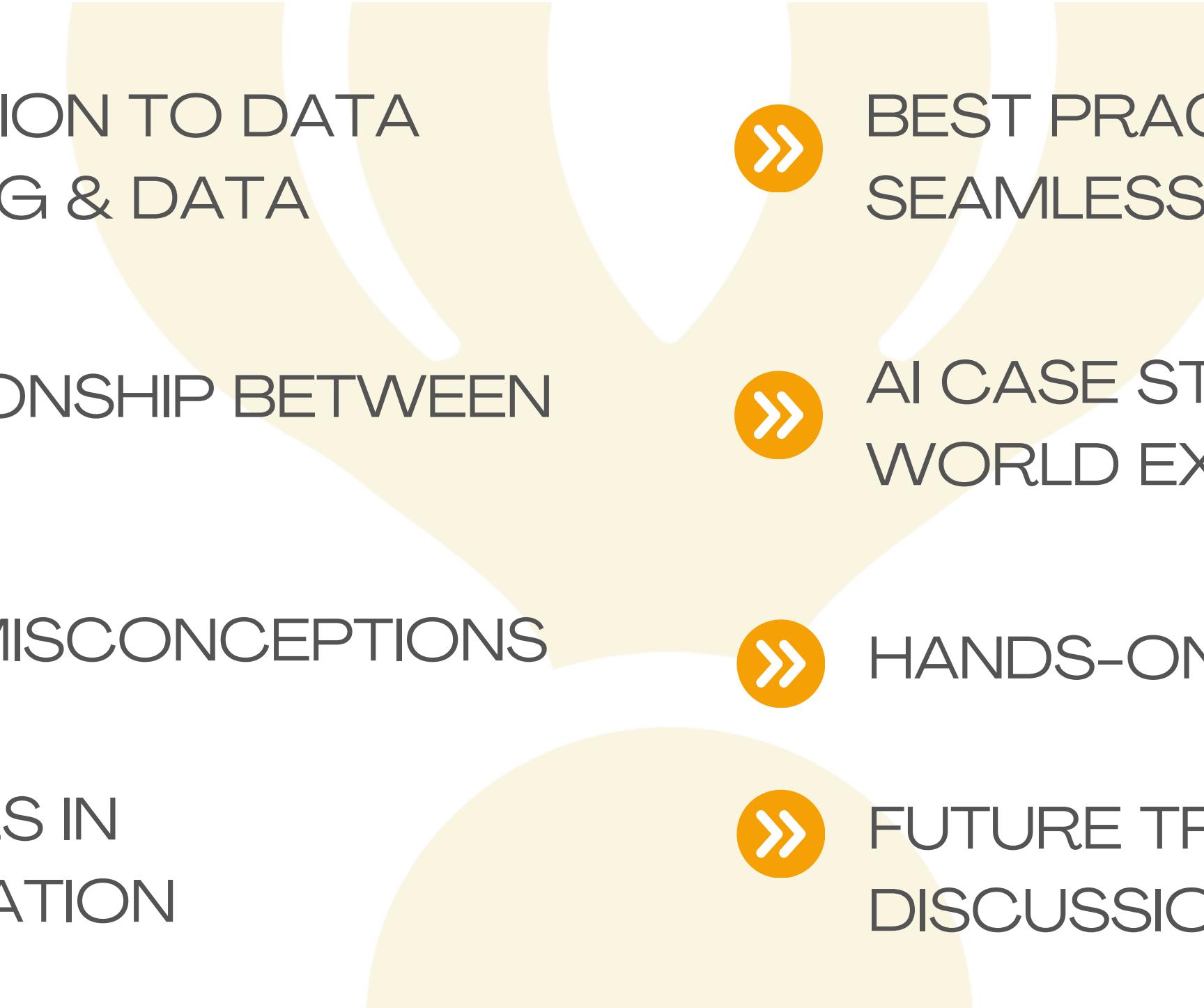
BRIDGING THE GAP BETWEEN DATA ENGINEERING AND DATA SCIENCE



VS



AGENDA

- 
- » INTRODUCTION TO DATA ENGINEERING & DATA SCIENCE
 - » THE RELATIONSHIP BETWEEN DE & DS
 - » COMMON MISCONCEPTIONS
 - » CHALLENGES IN COLLABORATION
 - » BEST PRACTICES FOR A SEAMLESS AI WORKFLOW
 - » AI CASE STUDIES & REAL-WORLD EXAMPLES
 - » HANDS-ON SESSION
 - » FUTURE TRENDS & OPEN DISCUSSION

INTRODUCTION TO DATA ENGINEERING & DATA SCIENCE

WHAT IS DATA ENGINEERING?



INTRODUCTION TO DATA ENGINEERING & DATA SCIENCE

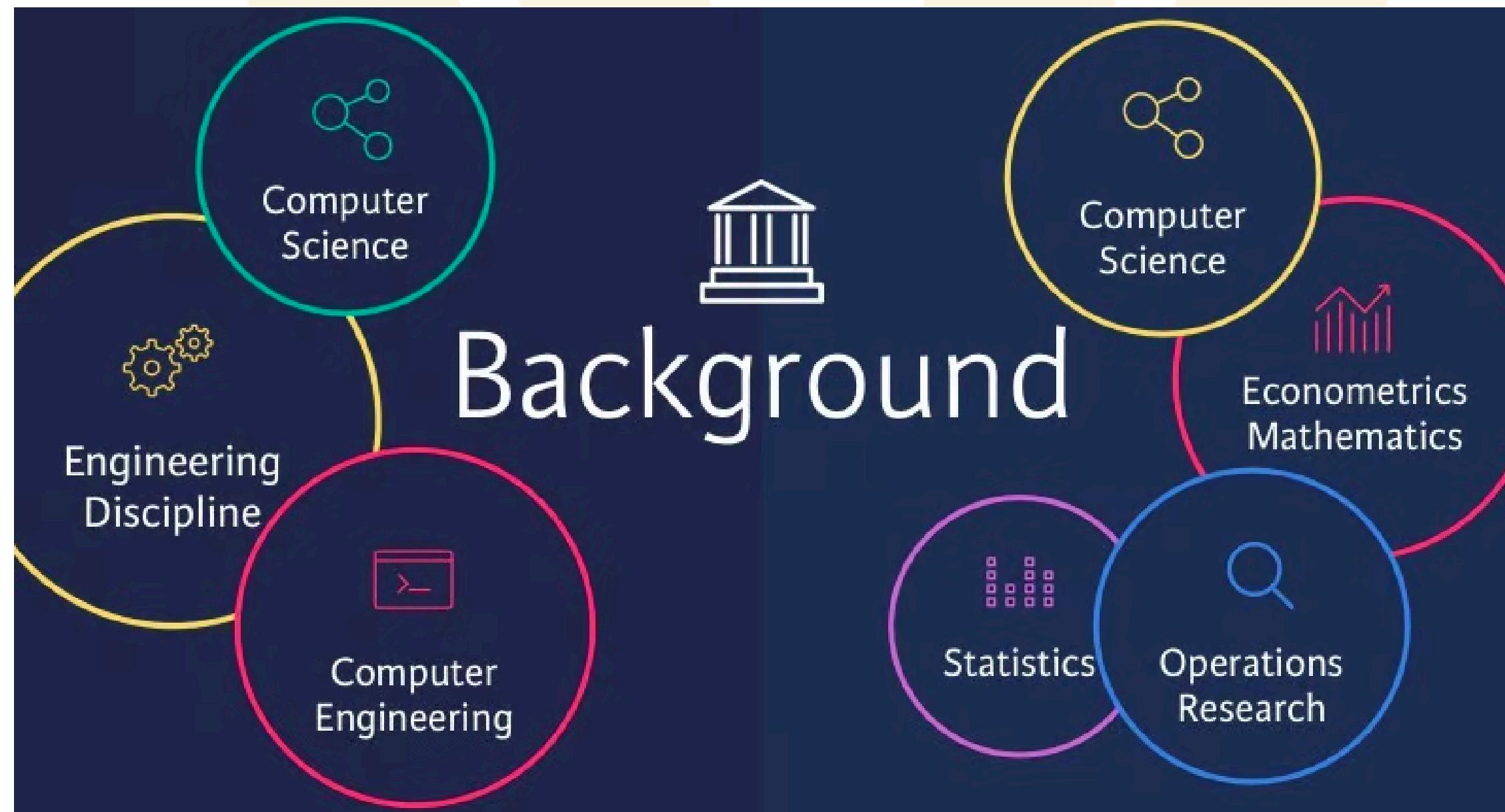
WHAT IS DATA SCIENCE?



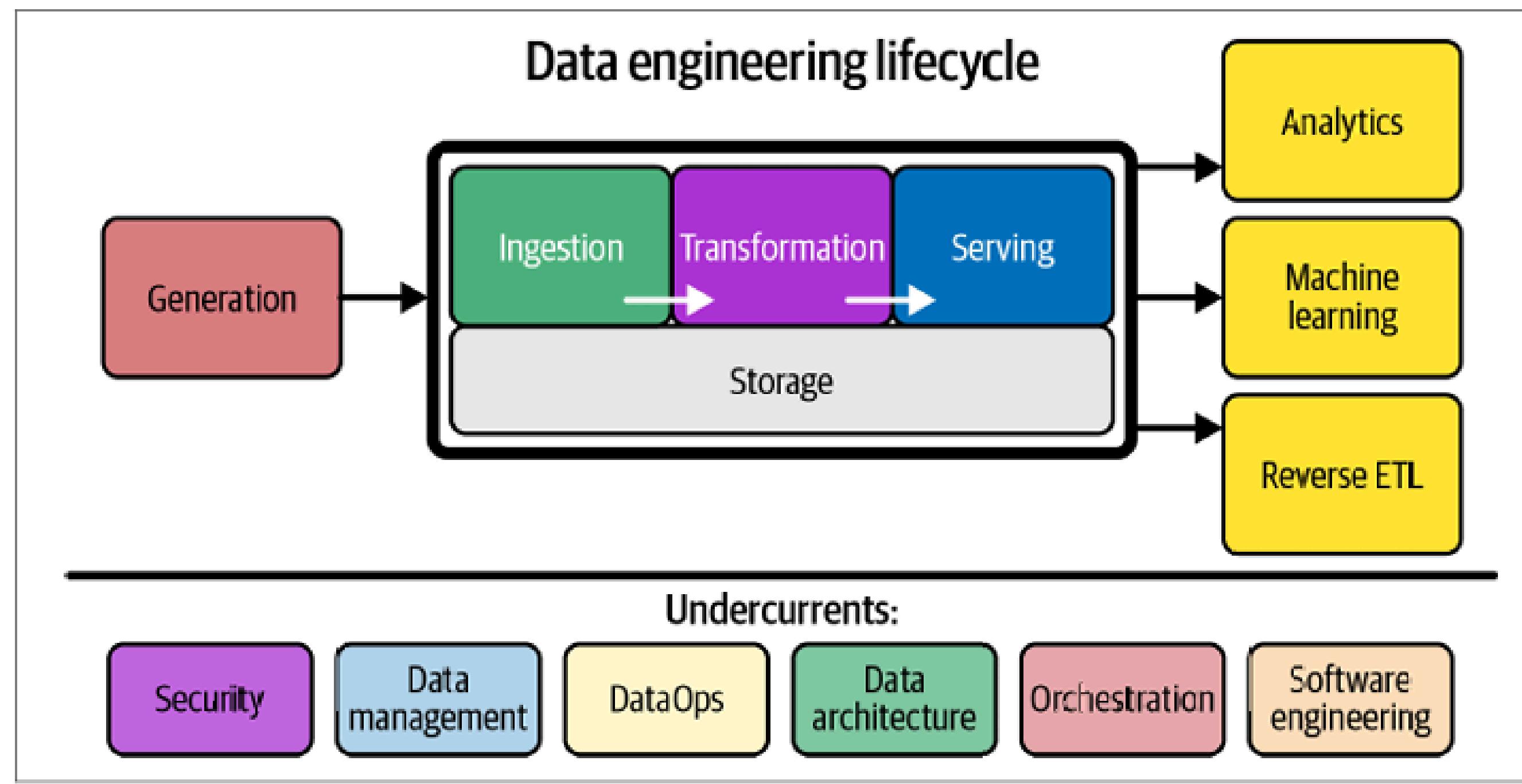
DATA SCIENCE VS DATA ENGINEERING: RESPONSIBILITIES



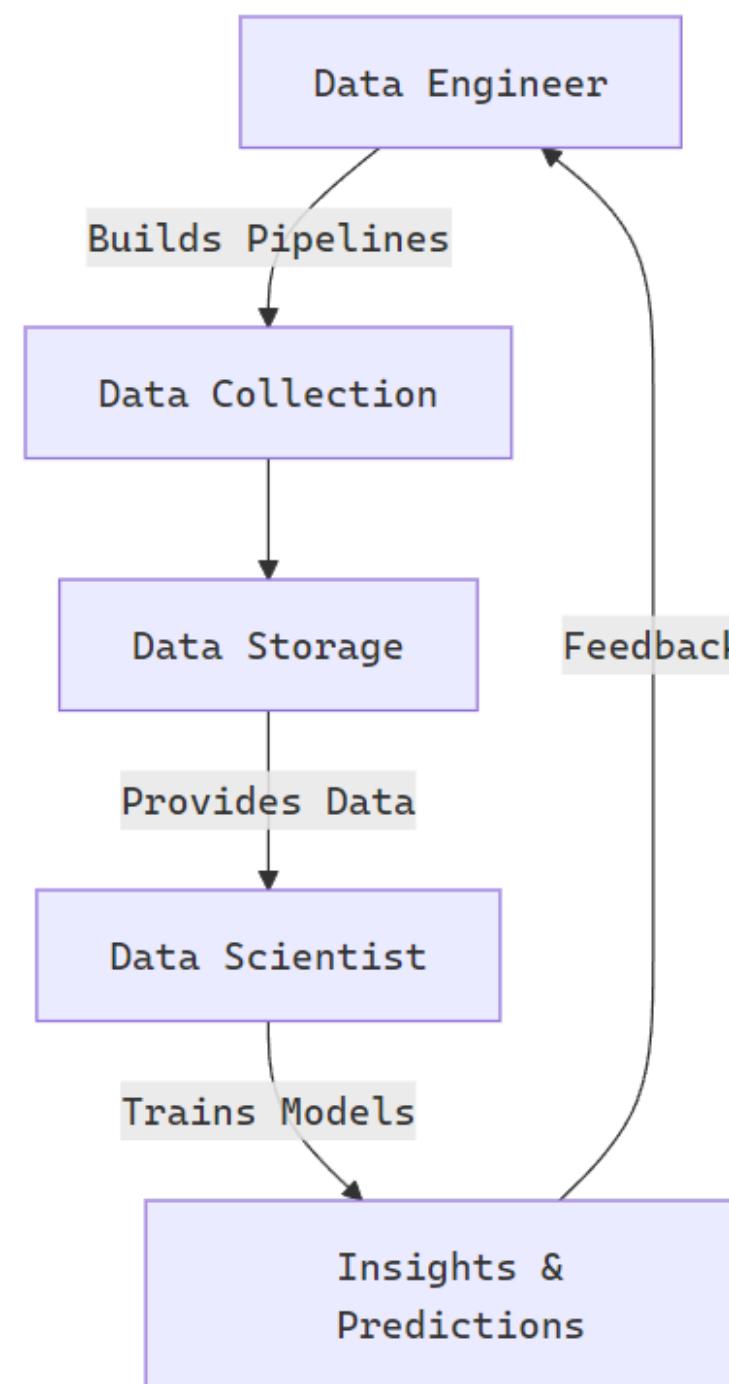
DATA SCIENCE VS DATA ENGINEERING: EDUCATIONAL BACKGROUND



THE DATA ENGINEERING LIFECYCLE



THE DATA ENGINEERING LIFECYCLE



"DES ENSURE DATA IS USABLE (CLEAN, STRUCTURED, ACCESSIBLE). DSS ENSURE DATA IS USEFUL (INSIGHTS, PREDICTIONS)."

DATA ENGINEERING VS DATA SCIENCE



COMMON MISCONCEPTIONS

DATA ENGINEERING MYTHS

» "DES JUST WRITE SQL AND ETL SCRIPTS."

» "DES DON'T NEED SOFTWARE ENGINEERING SKILLS."

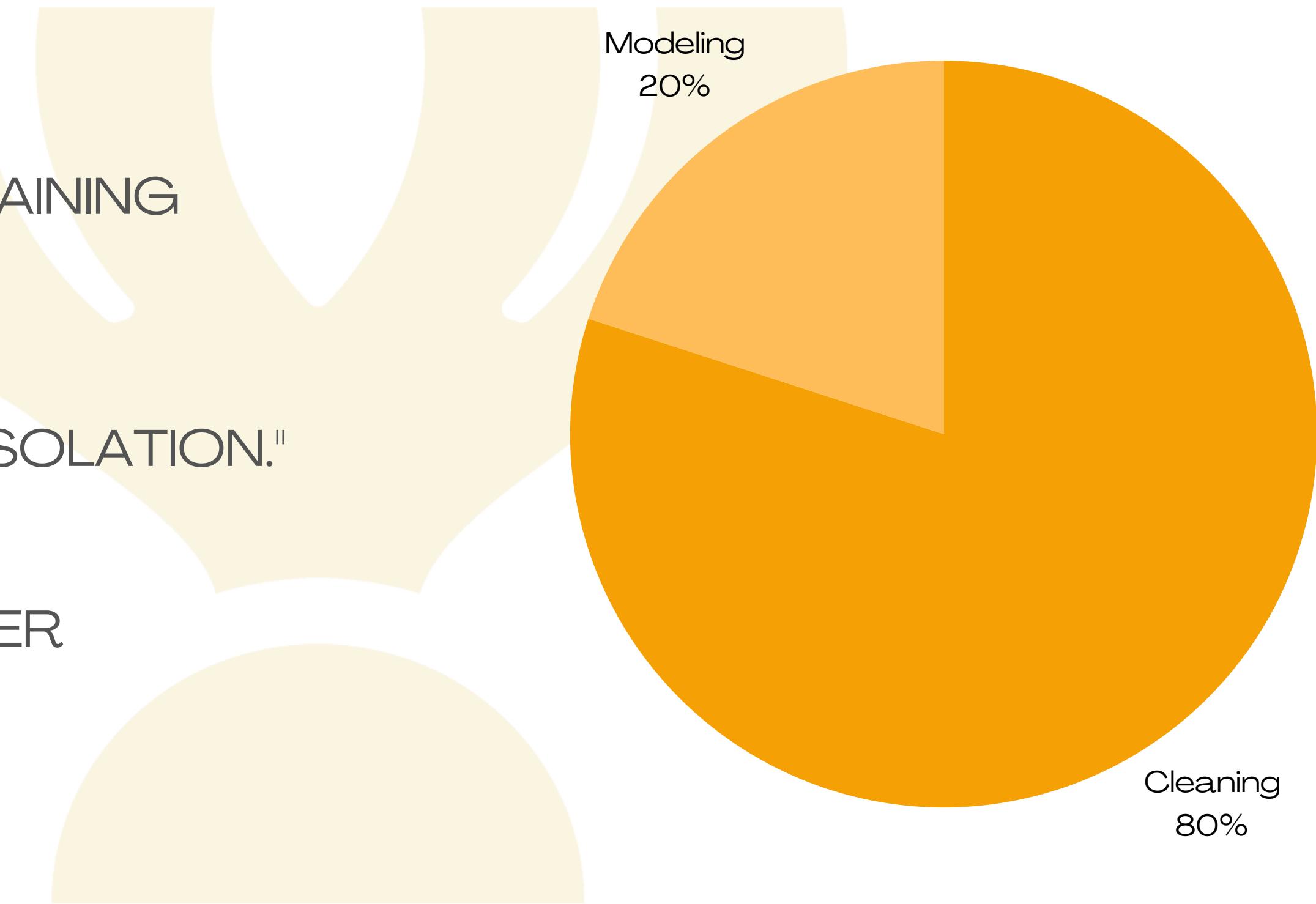
» "DE WORK IS DONE ONCE THE PIPELINE IS BUILT."



COMMON MISCONCEPTIONS

DATA SCIENCE MYTHS

- » "DS IS ALL ABOUT TRAINING FANCY MODELS."
- » "DSS CAN WORK IN ISOLATION."
- » "MORE DATA = BETTER MODELS."



CHALLENGES

COMMUNICATION GAPS



PREFERRED TOOLS



ITERATIVE PROCESSES

LET'S TAKE A BREAK



USE CASE

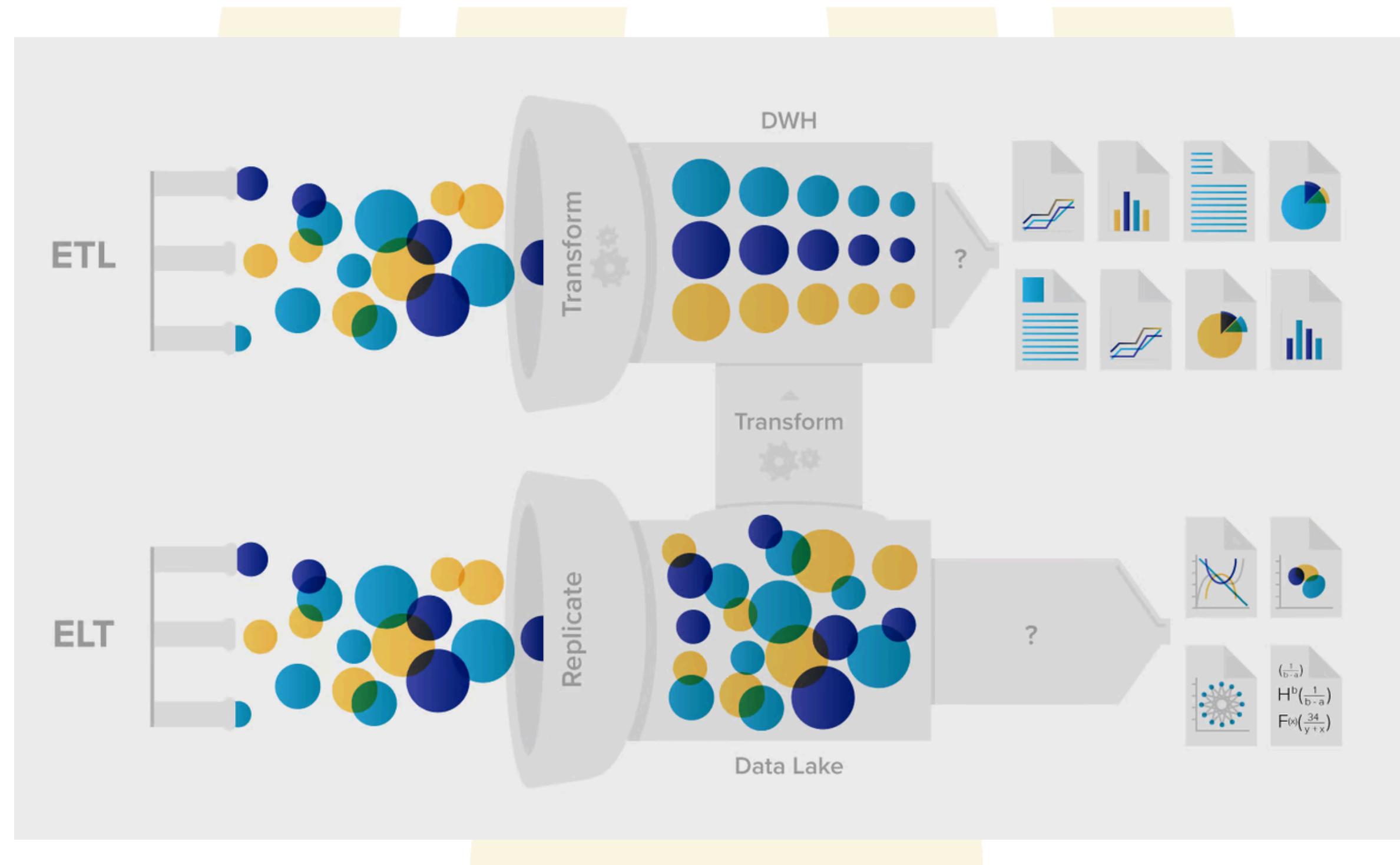


DATA WAREHOUSE VS DATA LAKE

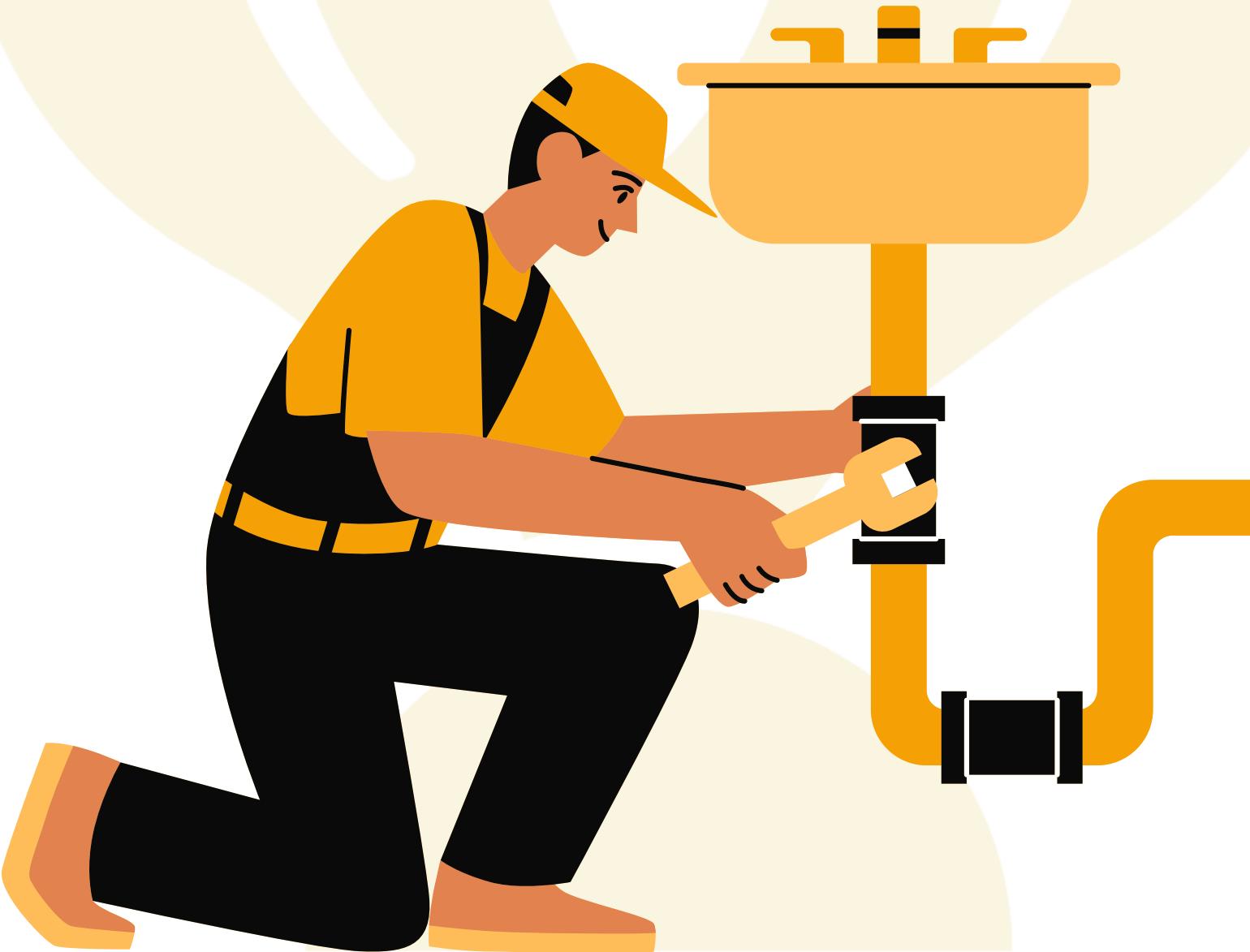
HOW TO CHOOSE BETWEEN THEM



ETL VS ELT



LET'S BUILD OUR PIPELINE



FUTURE TRENDS

LLMS WILL REVOLUTIONIZE DATA STACKS

The screenshot shows the Snowflake Copilot interface. On the left, a sidebar has icons for search, refresh, and data. The main area is titled "My Worksheet" with a "PUBLIC" share button and a "SP_WAREHOUSE" dropdown. A code editor window titled "CYBERSYN_US_PATENT_GRANTS.CYBERSYN" contains the following SQL query:

```
/*
Generated by Snowflake Copilot based on:
>Show me the names of the top 5 contributors with the most patents
*/
SELECT
    c.contributor_name,
    COUNT(p.patent_id) AS patent_count
FROM
    uspto_patent_contributor_relationships AS r
    JOIN uspto_contributor_index AS c ON r.contributor_id = c.contributor_id
    JOIN uspto_patent_index AS p ON r.patent_id = p.patent_id
GROUP BY
    c.contributor_name
ORDER BY
    patent_count DESC
LIMIT
    5;
```

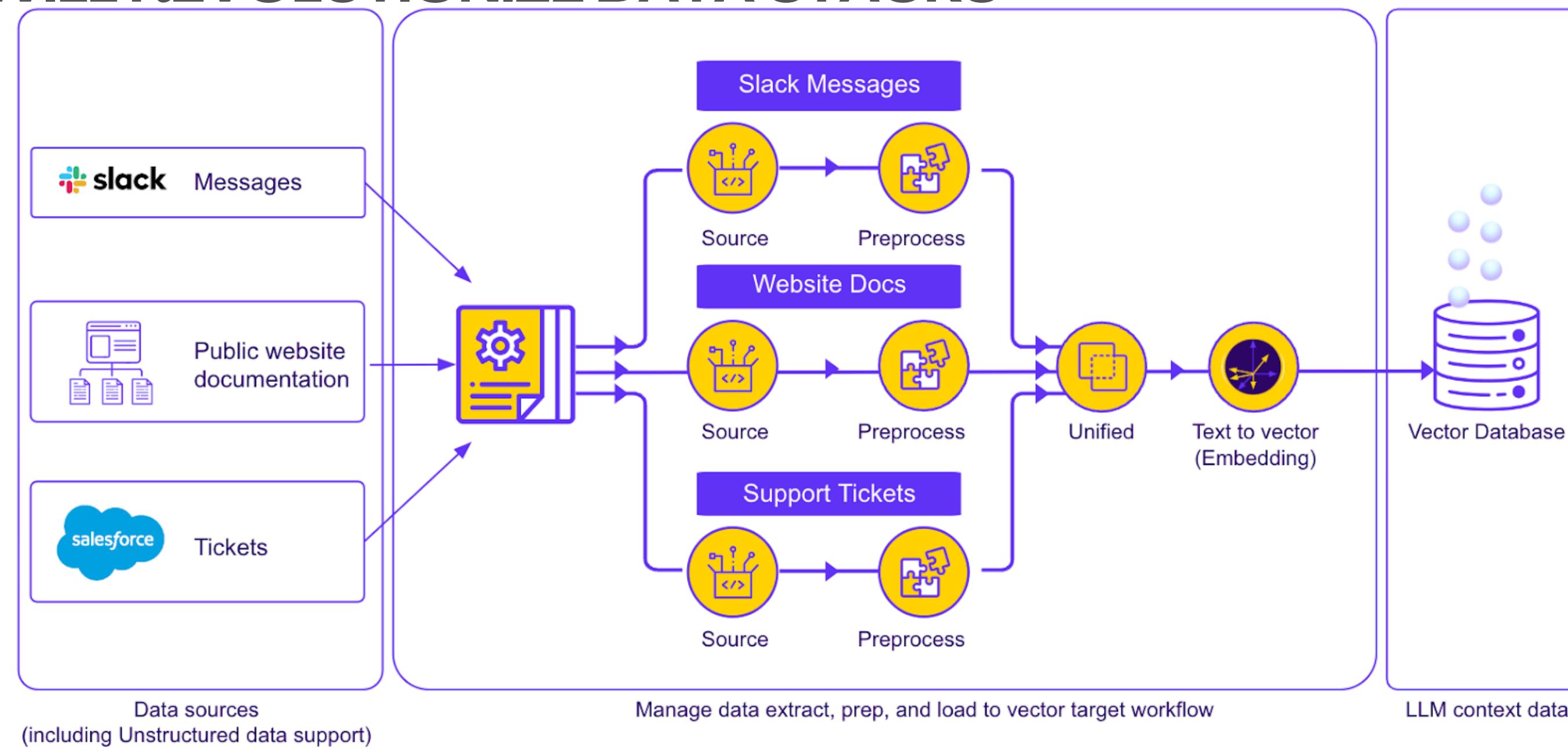
The "Results" tab is selected, displaying a table with the following data:

	CONTRIBUTOR_NAME	PATENT_COUNT
1	INTERNATIONAL BUSINESS MACHINES CORPORATION	218932
2	SAMSUNG ELECTRONICS CO., LTD.	159962
3	CANON KABUSHIKI KAISHA	112670
4	SONY CORPORATION	73697
5	INTEL CORPORATION	69264

On the right, a "Copilot" panel shows the generated SQL query with a "PREVIEW" button. It also includes a note: "Ordering the results by the patent count in descending order and limit the output to the top 5 contributors." Below the preview is a "SQL" section with the same query, a "Valid query" status, and "Add" and "Run" buttons. A message states: "This query will return the names of the top 5 contributors with the most patents." At the bottom, there's a text input field: "Ask a question about your data. Use @ to find tables and columns." and a "CYBERSYN_US_PATENT_GRANTS.CYBERSYN" dropdown.

FUTURE TRENDS

LLMS WILL REVOLUTIONIZE DATA STACKS



DISCUSSION TIME



REFERENCES

YOU WILL FIND ALL OF THE RESOURCES HERE

- ▶ [OBSIDIAN NOTE FOR THE REFERENCES](#)
- ▶ [GITHUB REPO AND DOCUMENTATION](#)





THANK YOU

FOR LISTENING

ANY QUESTIONS ?