Data Pipelines for Data Analytics

Some supplemental slides – 2025-03-12

Analysis in the Pre-Digital Age vs. Analysis in the Digital Age

Then:

- only people could carry out the activity of analysis and the components of an analysis process

Then:

- a given analysis of a situation was typically a one-time, one-off activity
- a single person might carry out ‘an analysis’ and then move on

---

Now:

- we can distill the essence of an analysis process into an algorithm
- we can automate analytical activities and its supporting process
- we have analysis machines

Now:

- we can expect that we will probably want to repeat variations of the same analysis over and over on new data that is streaming in on a regular basis

Also now: Reasoning Machines*

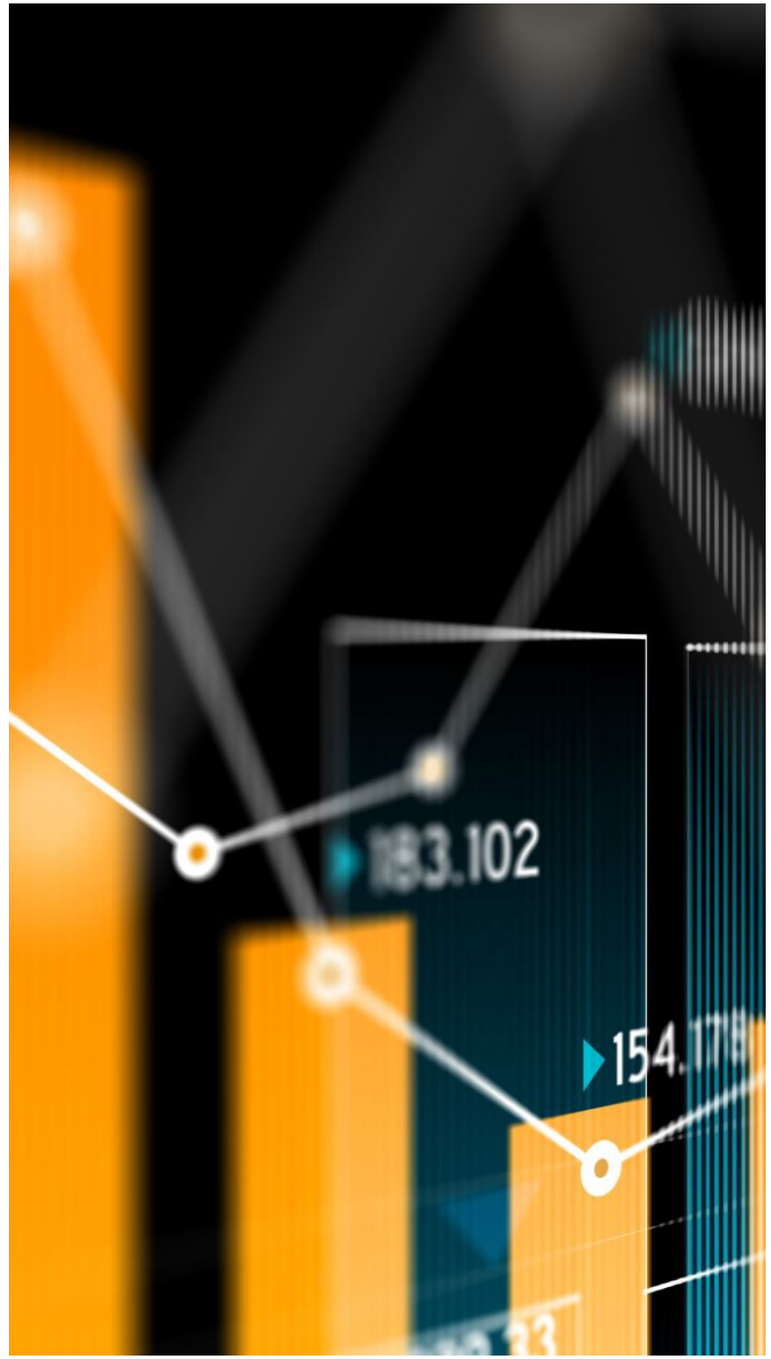
- Formalizing reasoning and analysis allows us to **automate it**, by programming it into computers.
- We can reframe reasoning as a process that **takes inputs** (premises/observations/ evidence) and **produces outputs** (conclusions).
- By automating this process, we can get machines to carry out reasoning for us.
- The result could(?) be more **reliable**, **dependable**, **consistent**.
- Data is fuel for these machines. BUT: **garbage in, garbage out**! Weak premises in, possibly false conclusion out!

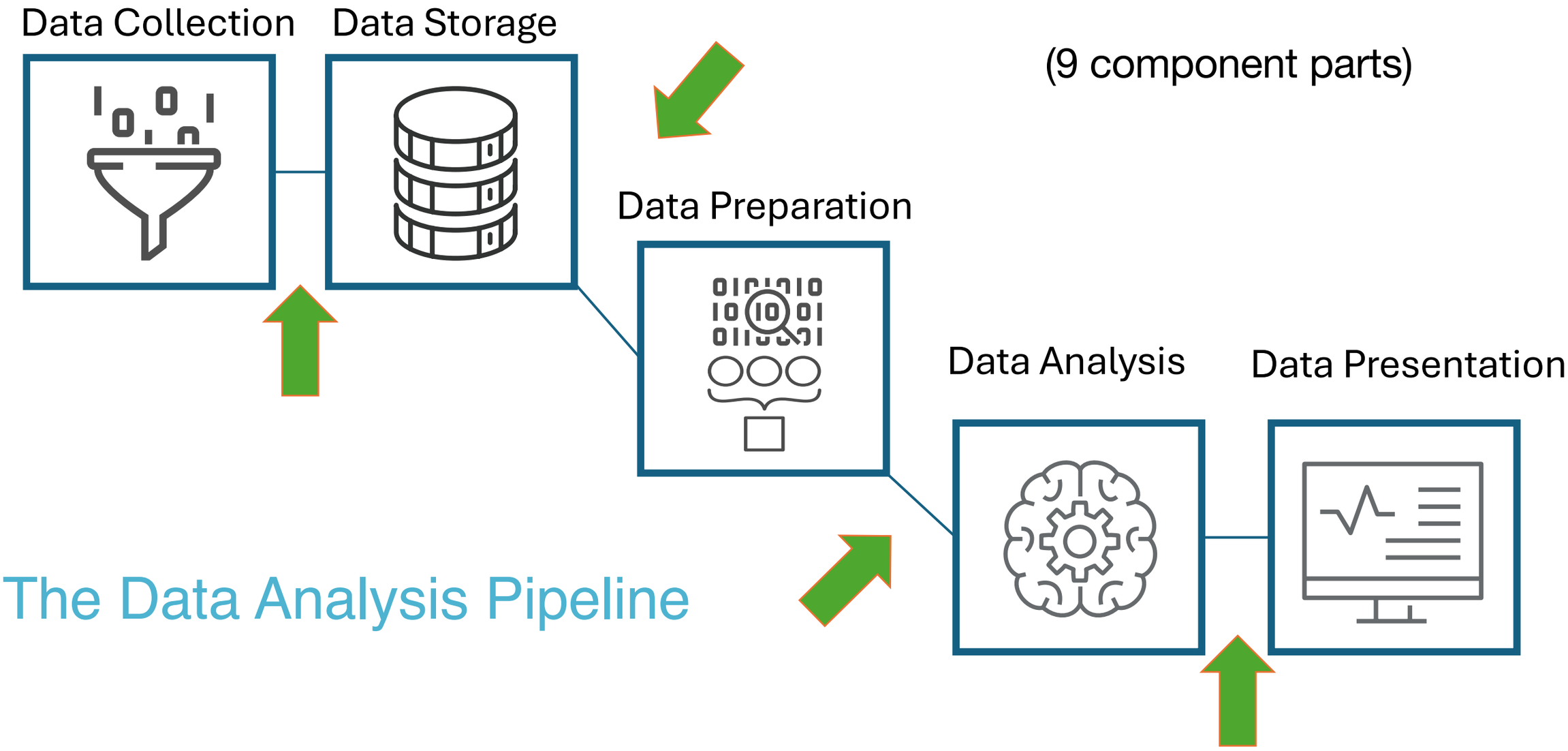
*Here I don't mean generative AI (e.g. ChatGPT). There are different types of AI that carry out reasoning tasks.

Increasing importance of data pipelines…

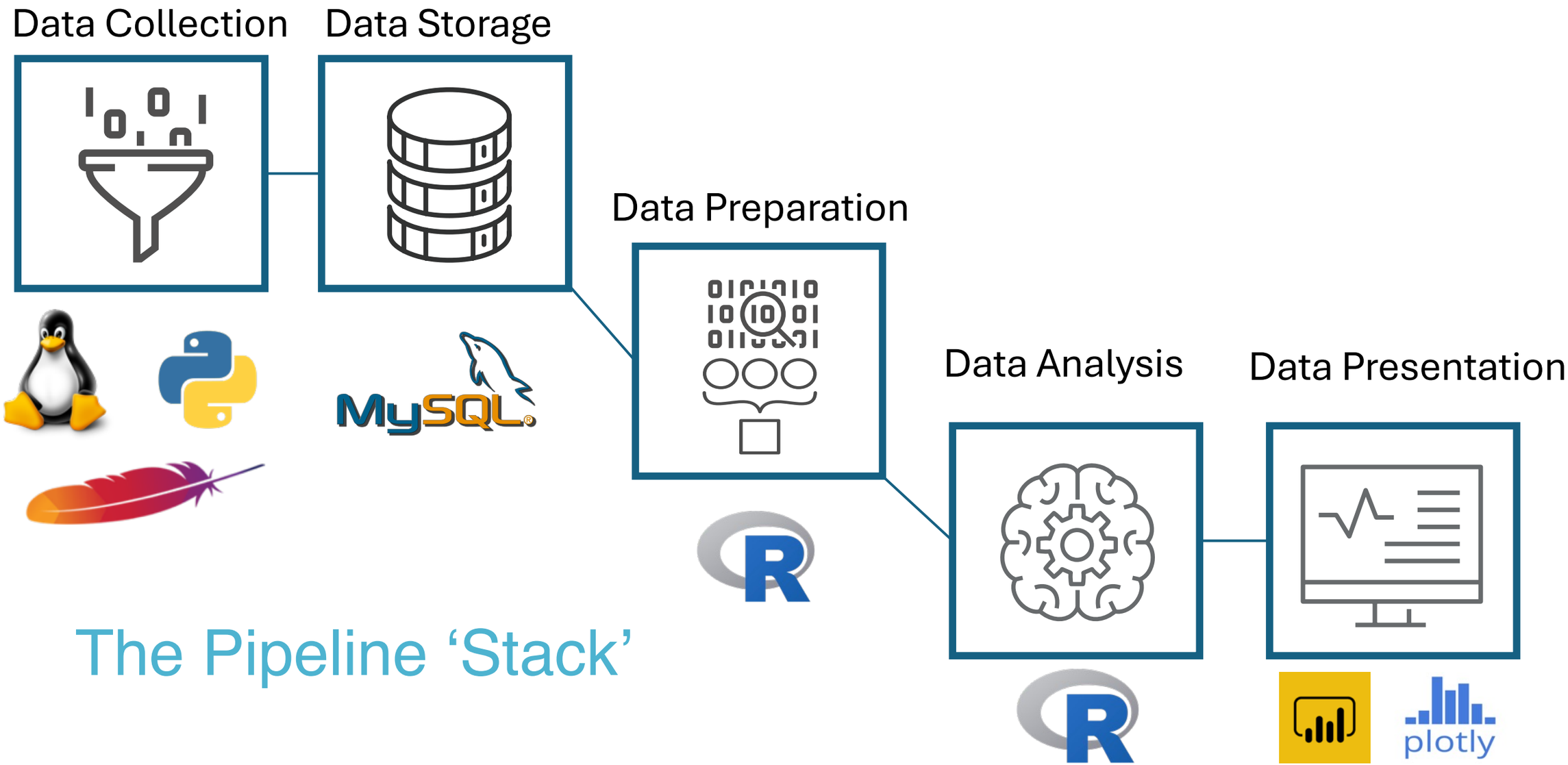
Desktop Data Analysis

- Business Intelligence needs are pushing the development of **desktop data analysis** tools and pipelines, such as:
  - PowerBI
  - Tableau
- Democratization of data + increase in data/digital literacy.
- This is likely going to push organizations forward as well.
- Not **necessarily** a substitute for ‘industrial’ or ‘professional’ data pipelines.

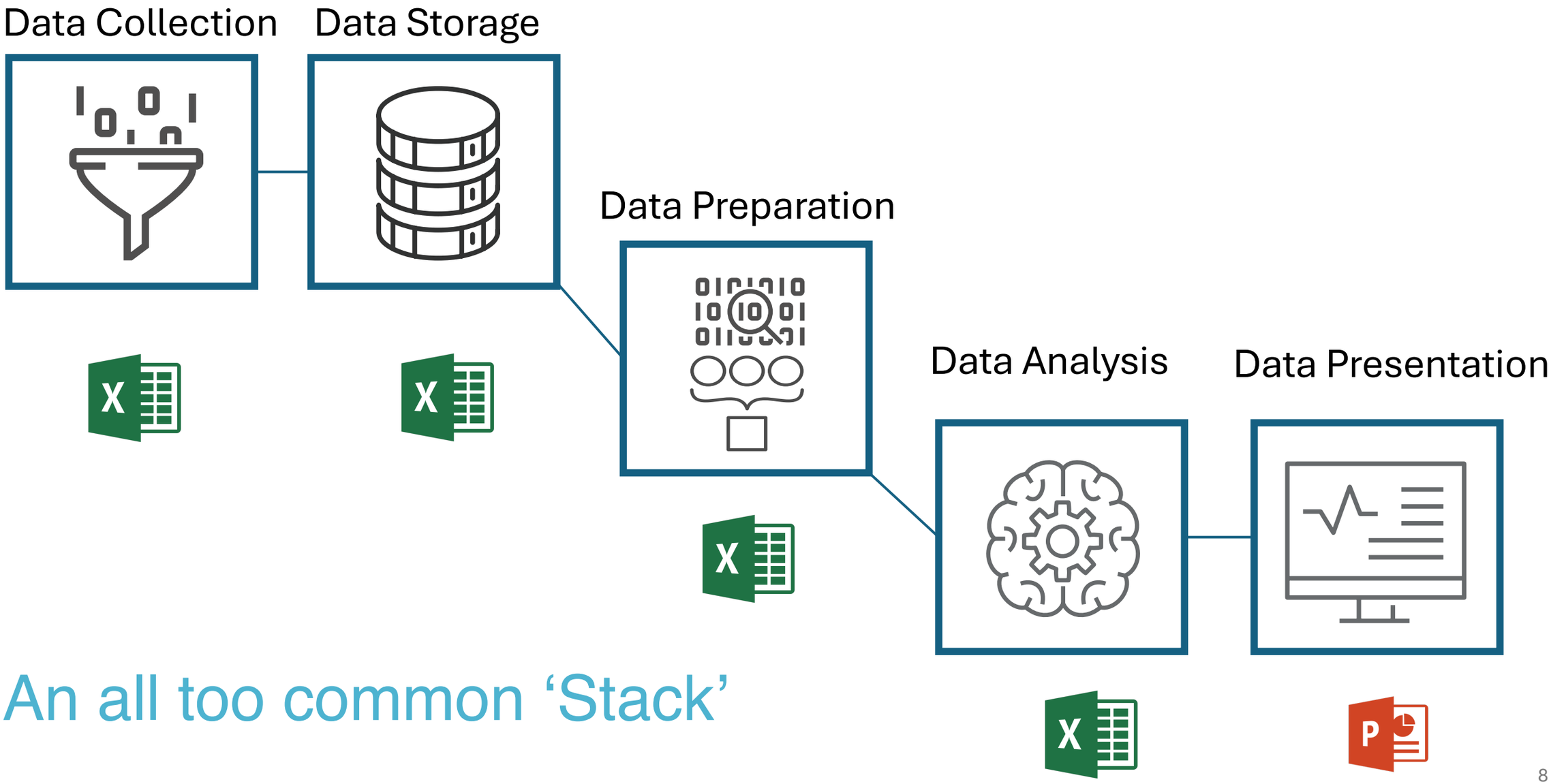




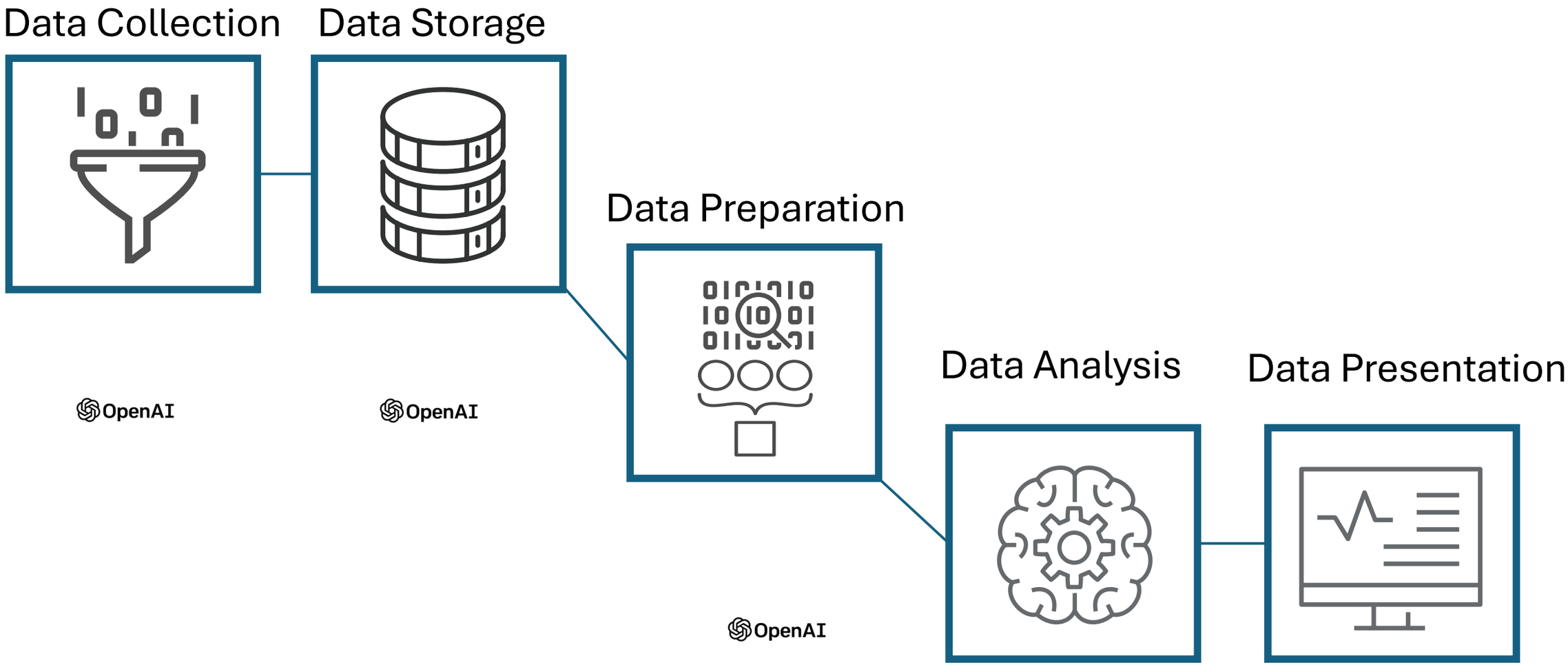
The Data Analysis Pipeline



The Pipeline ‘Stack’



An all too common ‘Stack’



A Stack of the Future?

Data Pipeline Technologies – Maturity Levels

	On-Premises (LAN)	Public Cloud	Private Cloud
Amateur	Shared Directory + Excel + Power Point + 'Desktop' Access	**Server Based End-to-End Automated Pipeline Tech:** On-Premises Azure, On-Premises IBM RedHat	Home Brewed Solutions using Servers Stood Up on Cloud – e.g., AWS, GCP
Semi-Pro	**Desktop DataScience:** Desktop PowerBI SQL-Lite (Desktop) MS Access Stand-Alone In-House DBMS – Read + Write		
Professional	**Server Based End-to-End Automated Pipeline Tech:** On-Premises Azure, On-Premises IBM RedHat	**End-to-end SaaS data pipelines:** e.g., COTS Pachyderm or more bespoke: e.g., SaaSCoder	**End-to-End Cloud Data Pipeline Infrastructure (Serverless/NoServer):** AWS, GCP, Azure

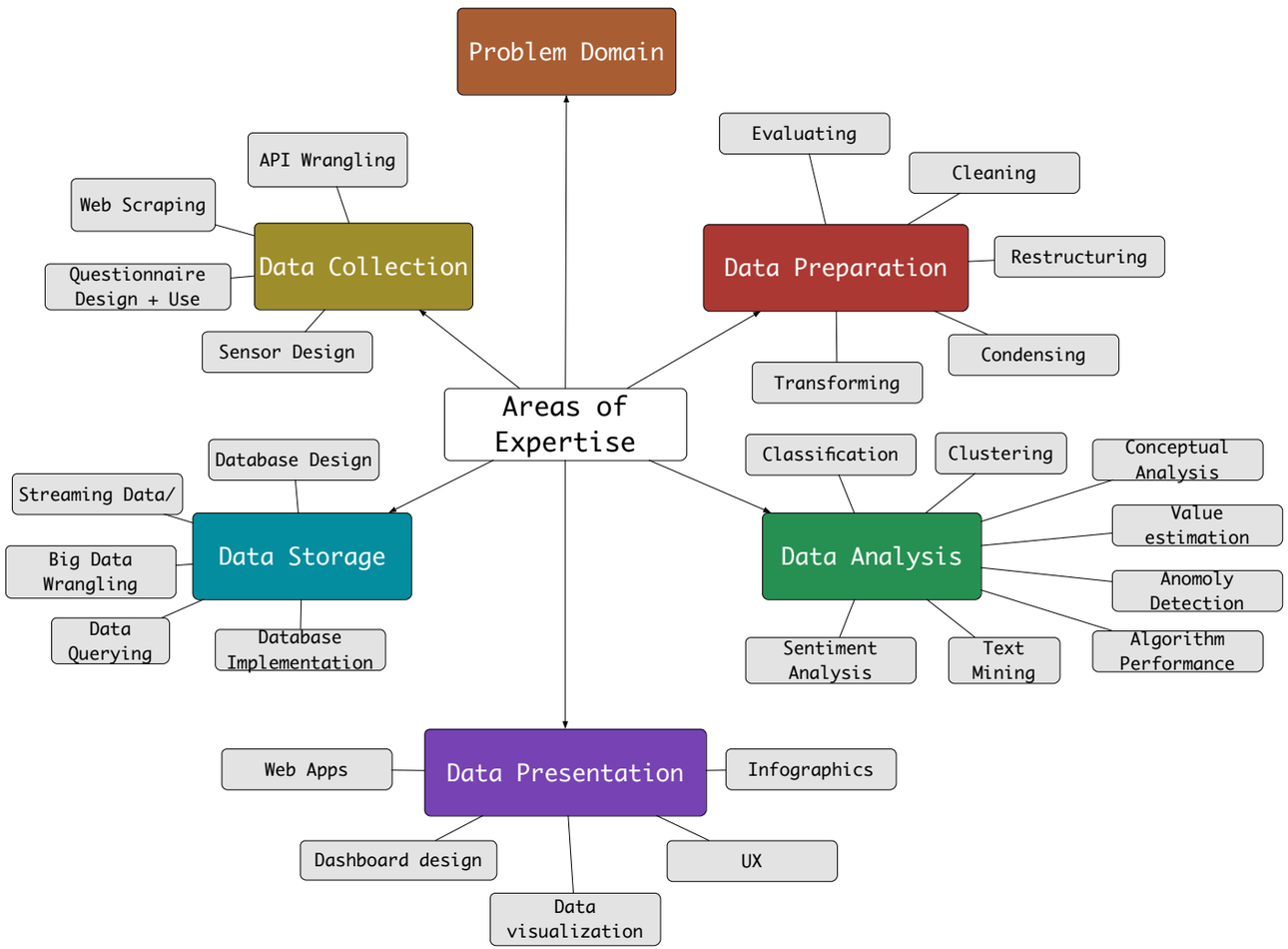
Pipeline Creation Phases

1. **Research + Design**
2. **Implementation**
3. **Testing**
4. **Production + Management**
5. **Back to Research + Design**

Data Scientist or Data Engineer?

- Common data analogy: “Data is the new oil.”
- **Data scientist:** expert on what flows through the pipeline – the data
- **Data engineer**: expert on how to build the pipeline itself – the IT infrastructure
- Do you need a data engineer or a data scientist?
- Trick question – you (probably) need both!

Data Science – A Team Sport*



*still the case with GenAI...

Coming back around to decision support

- What is produced by the pipeline (comes out the end of the pipeline) must reflect or represent reality (ground truth) adequately.
- AND ALSO the data presentation piece must be relevant to the decision makers.
- AND ALSO the data presentation piece must be understandable by and impactful for the decision makers.
- If not, all of the work put into the pipeline may not be worth it!