

# Data Engineering

LTAT.02.007

Ass Prof. Riccardo Tommasini

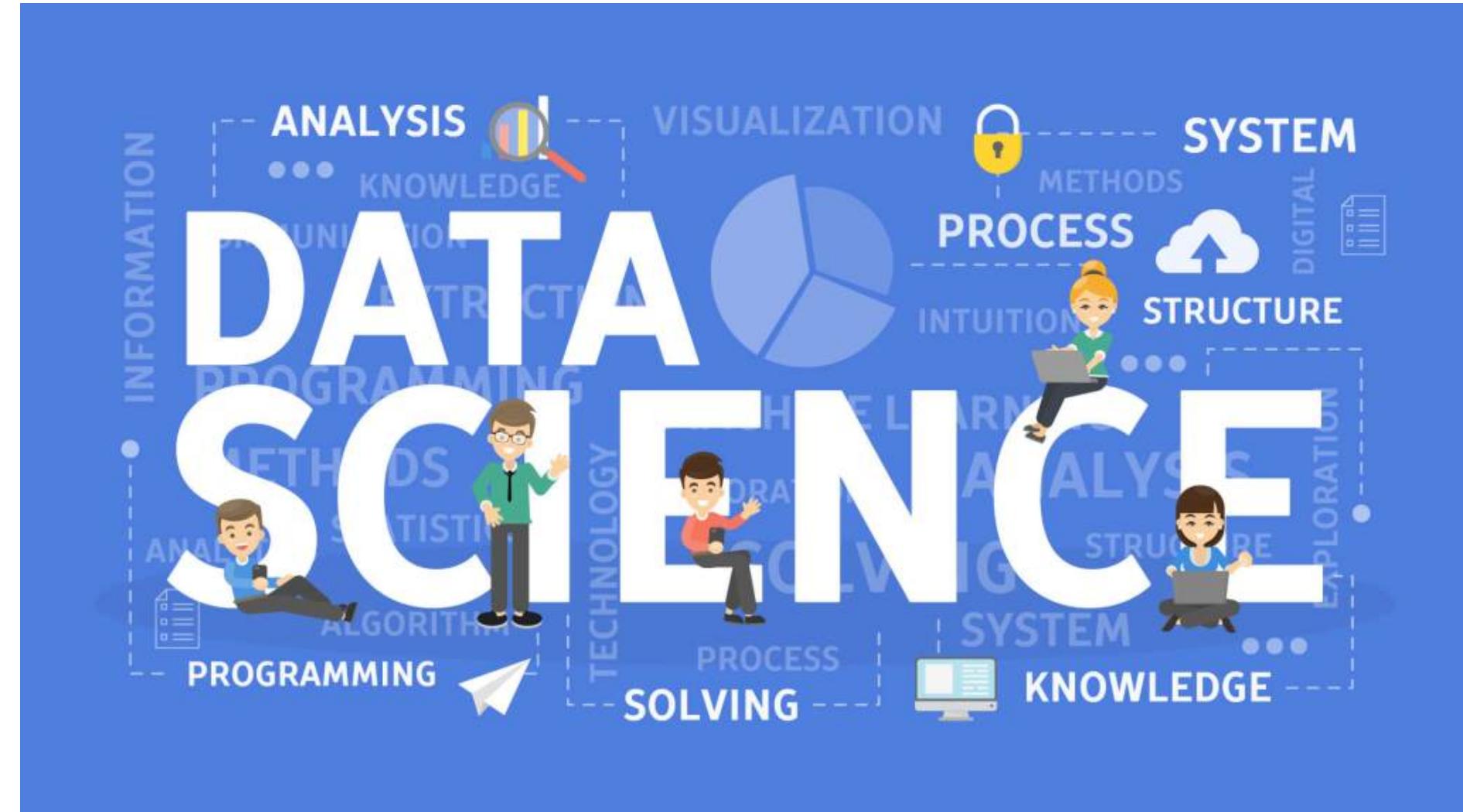
Assistants: **Fabiano Spiga, Mohamed Ragab, Hassan Eldeeb**

- <https://courses.cs.ut.ee/2020/dataeng>
- Forum
- Moodle



# Data Engineer

# Data Science<sup>01</sup>



<sup>01</sup> Source

# Roles in a Data Science Project<sup>02</sup>



<sup>02</sup> <http://emanueledellavalle.org/slides/dspm/ds4biz.html#25>

# Roles in a Data Science Project<sup>02</sup>



<sup>02</sup> <http://emanueledellavalle.org/slides/dspm/ds4biz.html#25>

## The Data Engineer

A dedicated specialist that maintain data available and usable by others (Data Scientists).<sup>03</sup>

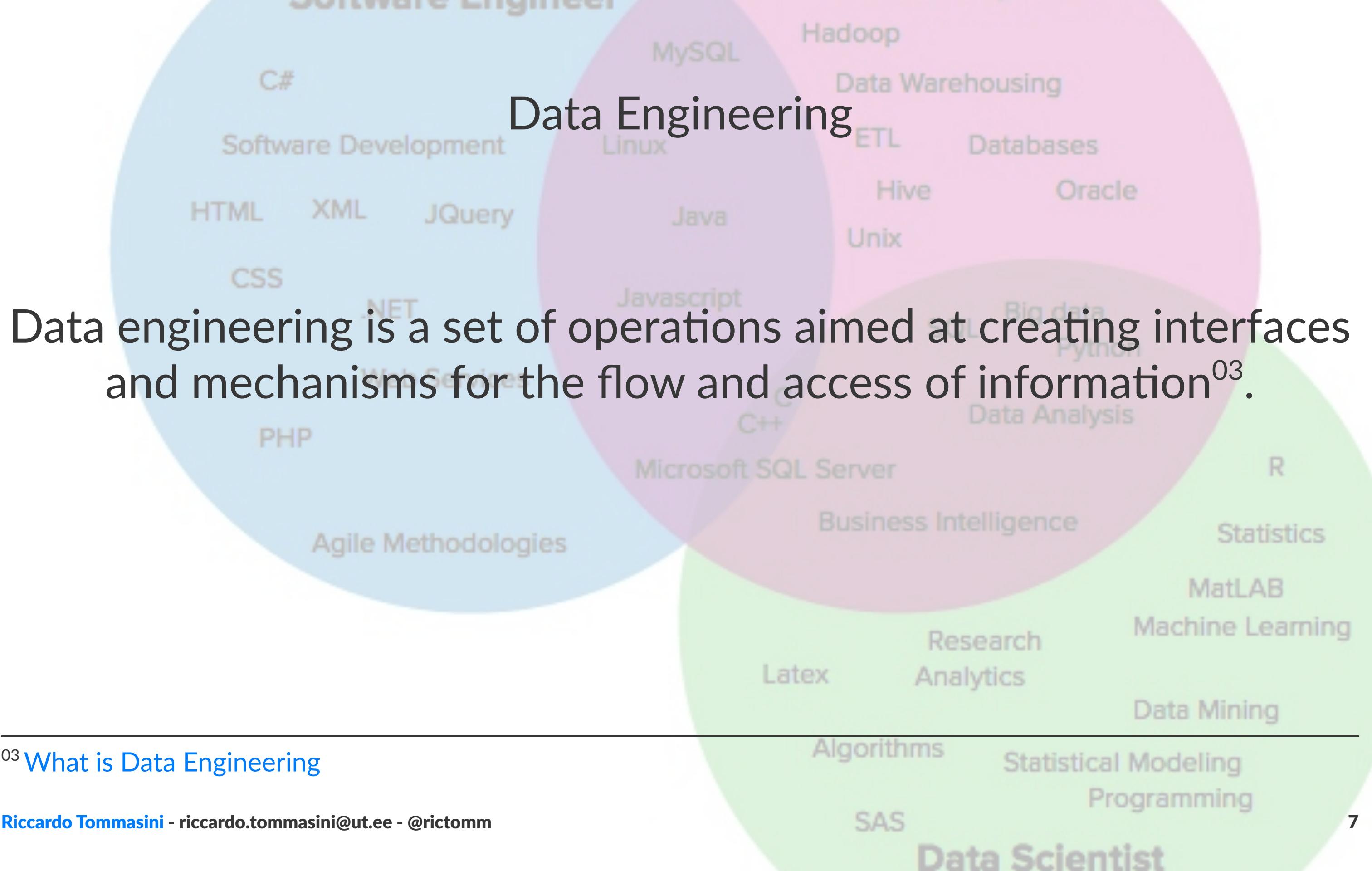
Data engineers set up and operate the organization's data infrastructure preparing it for further analysis by data analysts and scientists.<sup>03</sup>

Data engineering field could be thought of as a superset of business intelligence and data warehousing that brings more elements from software engineering.<sup>04</sup>

---

<sup>03</sup> [What is Data Engineering](#)

<sup>04</sup> [Source: The Rise of Data Engineer](#)



<sup>03</sup> [What is Data Engineering](#)

⤳ You Retweeted

 **Seth Rosen** @sethrosen · Apr 20

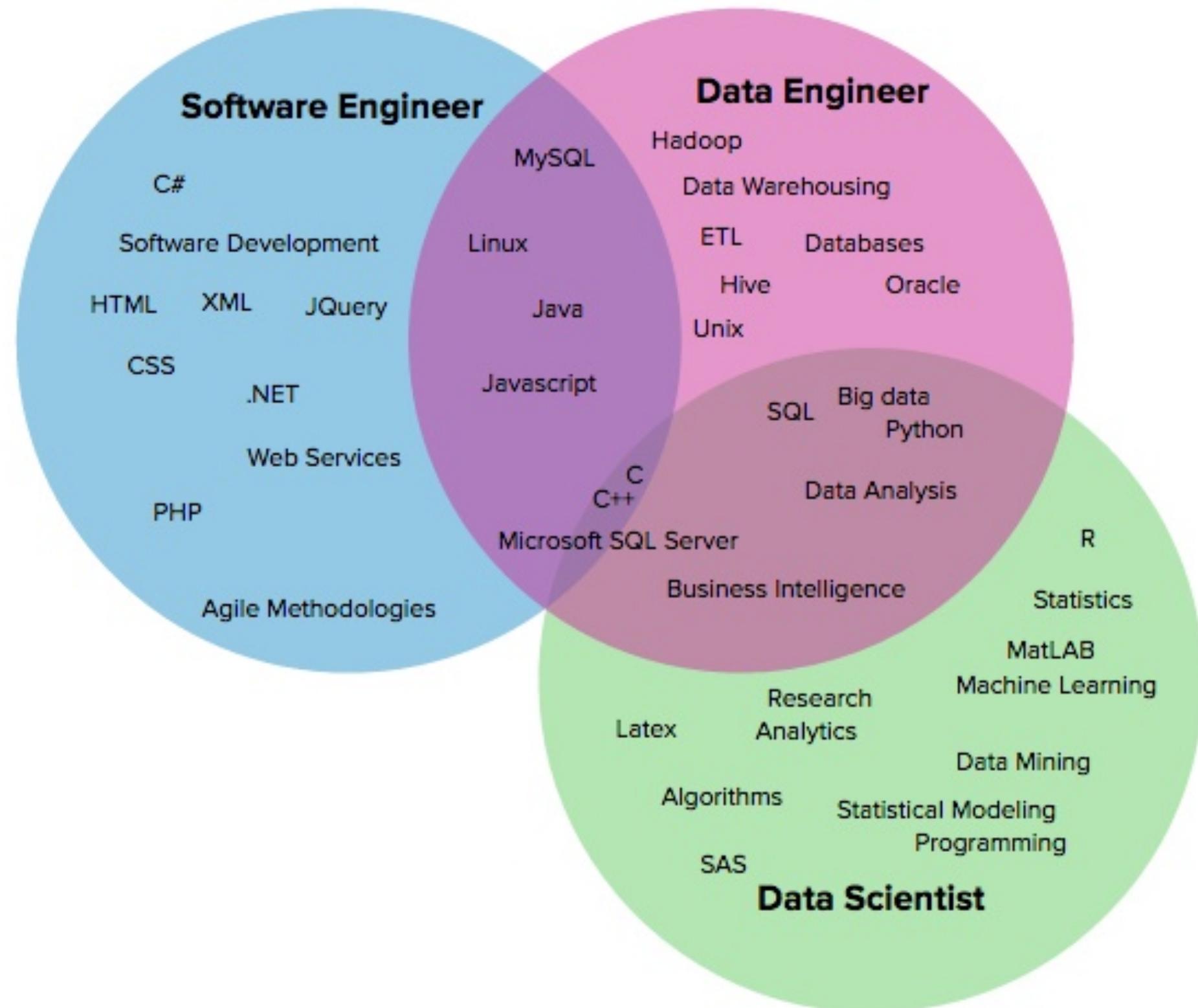
Them: Can you just quickly pull this data for me?

Me: Sure, let me just:

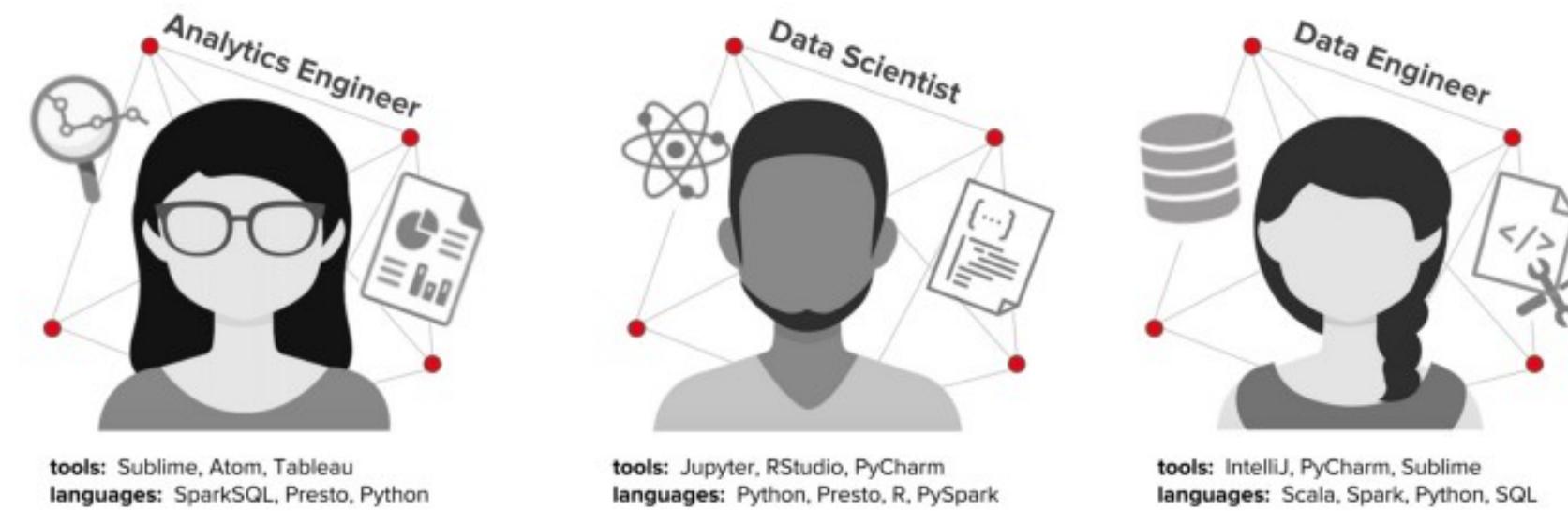
```
SELECT * FROM  
some_ideal_clean_and_pristine.table_that_you_think_exists
```

323 4.4K 28K ↑

[Show this thread](#)



# Netflix's Perspective<sup>05</sup>



<sup>05</sup> Netflix Innovation

# The Knowledge Scientist<sup>06</sup>



---

<sup>06</sup> [The Manifesto](#)

# Google's Two-Cents

## Professional Data Engineer

A Professional Data Engineer enables data-driven decision making by collecting, transforming, and publishing data. A Data Engineer should be able to design, build, operationalize, secure, and monitor data processing systems with a particular emphasis on security and compliance; scalability and efficiency; reliability and fidelity; and flexibility and portability. A Data Engineer should also be able to leverage, deploy, and continuously train pre-existing machine learning models.

The Professional Data Engineer exam assesses your ability to:

- ✓ Design data processing systems
- ✓ Build and operationalize data processing systems
- ✓ Operationalize machine learning models
- ✓ Ensure solution quality

[Register](#)

[FAQs](#)

This exam is available in English and Japanese.



# Philosophy of (Data) Science<sup>07</sup>



---

<sup>07</sup> Data as Fact

# What is Data?



# Oxford Dictionary

*Data [uncountable, plural] facts or information, especially when examined and used to find out things or to make decisions.*<sup>08</sup>

---

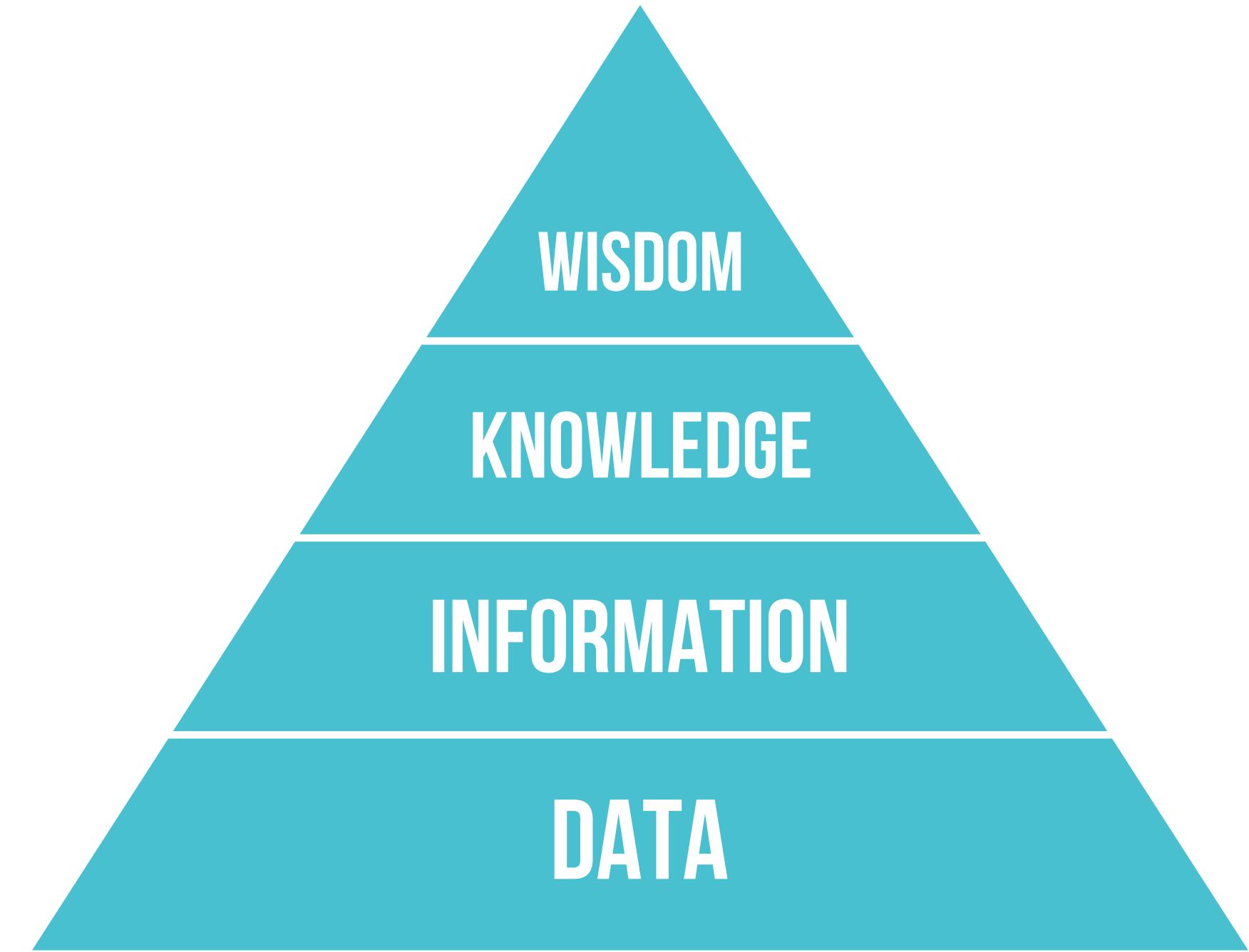
<sup>08</sup> Def

# Wikipedia

Data (treated as singular, plural, or as a mass noun) is any sequence of one or more symbols given meaning by specific act(s) of interpretation<sup>09</sup>

---

<sup>09</sup> Data in Computing)



## Data Warehouse: A Traditional Approach:

A data warehouse is a copy of transaction data specifically structured for query and analysis. – **Ralph Kimball**

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.-- **Bill Inmon**

---

<sup>03</sup> [What is Data Engineering](#)

# Data Warehouse vs Data Bases

Surprisingly, Data Warehouse isn't a regular database.

- A database normalizes data separating them into tables and avoiding redundancies
- It supports arbitrary workload and complex queries
- do not store multiple versions of data
- a Data Warehouse uses few tables to improve performance and analytics.
- a Data Warehouse allows simple queries
- supports versioning for complex analysis

# Data Pipeline

A Data pipeline is a sum of tools and processes for performing data integration<sup>03</sup>

Constructing data pipelines is the core responsibility of data engineering.

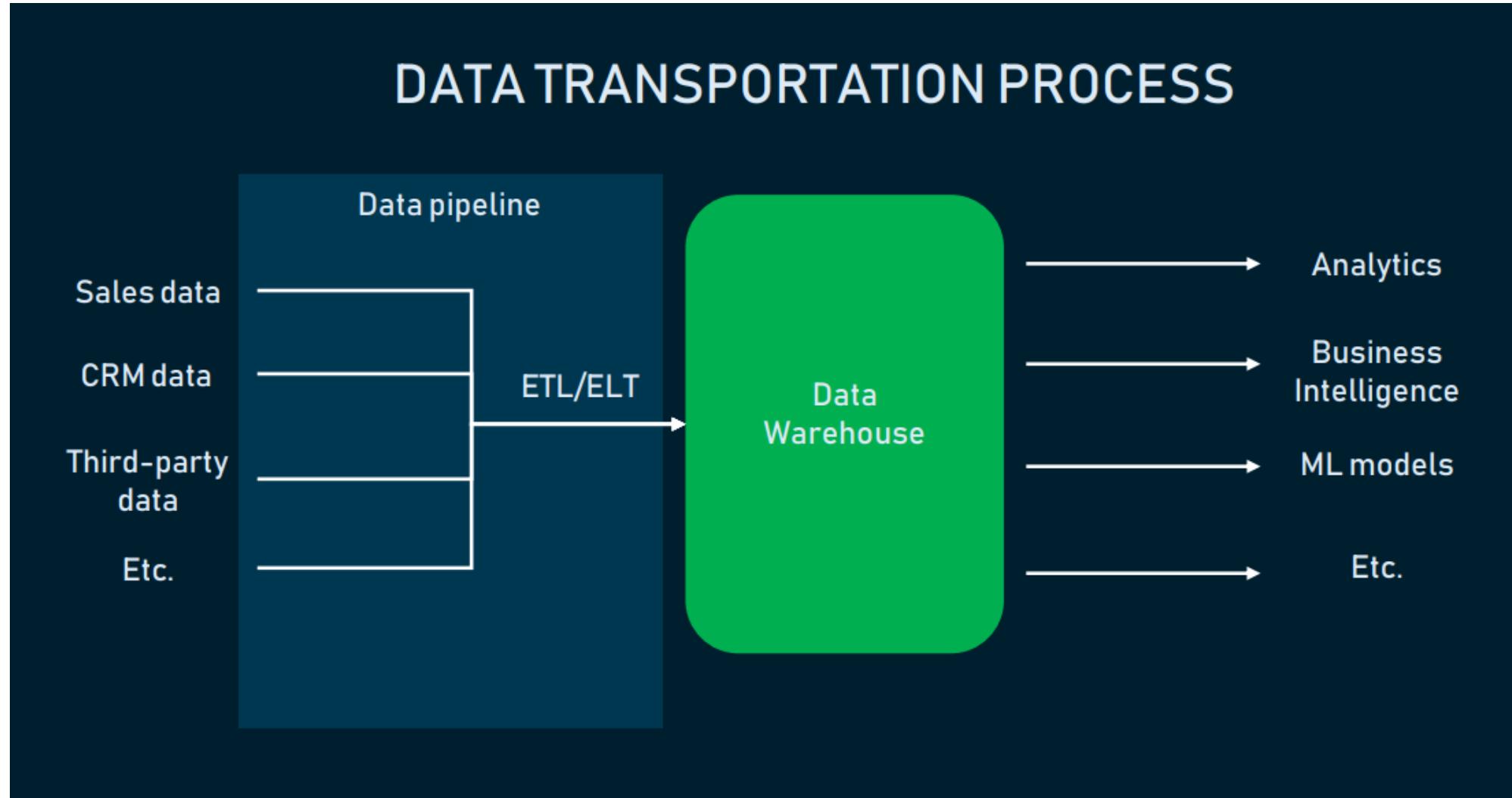
---

<sup>03</sup> [What is Data Engineering](#)

# Data pipelines are used for

- moving data to the cloud or to a data warehouse
- data wrangling
- data integration

# Transporting data from sources into a warehouse<sup>010</sup>



<sup>010</sup> Source

# Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines<sup>011</sup>

- **SQL-centric Pipelines** uses SQL dialects from Presto or Hive. Pipelines (ETLs) are defined in a declarative way, and almost everything centers around SQL and tables.
- **JVM-centric Pipelines** uses languages like Java or Scala and often involves thinking data transformation in an imperative manner, e.g. in terms of key-value pairs.
- Drag & Drop...

---

<sup>011</sup> we are focusing on ETL

# Skill Set: SQL mastery<sup>03</sup>

If english is the language of business, SQL is the language of data.

- SQL/DML/DDL primitives are simple enough that it should hold no secrets to a data engineer. Beyond the declarative nature of SQL, she/he should be able to read and
- understand database execution plans, and have an understanding of what all the steps are,
- understand how indices work,
- understand the different join algorithms

---

<sup>03</sup> [What is Data Engineering](#)

## Skill Set: Data modeling<sup>03</sup>

For a data engineer, entity-relationship modeling should be a cognitive reflex, along with a clear understanding of normalization, and have a sharp intuition around denormalization tradeoffs.

The data engineer should be familiar with dimensional modeling and the related concepts and lexical field.

---

<sup>03</sup> [What is Data Engineering](#)

# But...

# Engineers Shouldn't (only) Write (SQL-based) ETL<sup>012</sup>

- Unless you need to process over many petabytes of data, or you're ingesting hundreds of billions of events a day, most technologies have evolved to a point where they can trivially scale to your needs.
- Unless you need to push the boundaries of what these technologies are capable of, you probably don't need a highly specialized team of dedicated engineers to build solutions on top of them.

---

<sup>012</sup> [JeffMagnusson, 2016](#)

# If Not (only) ETL, Then...What?<sup>013</sup>

Data Engineers are still a critical part of any high-functioning data team.

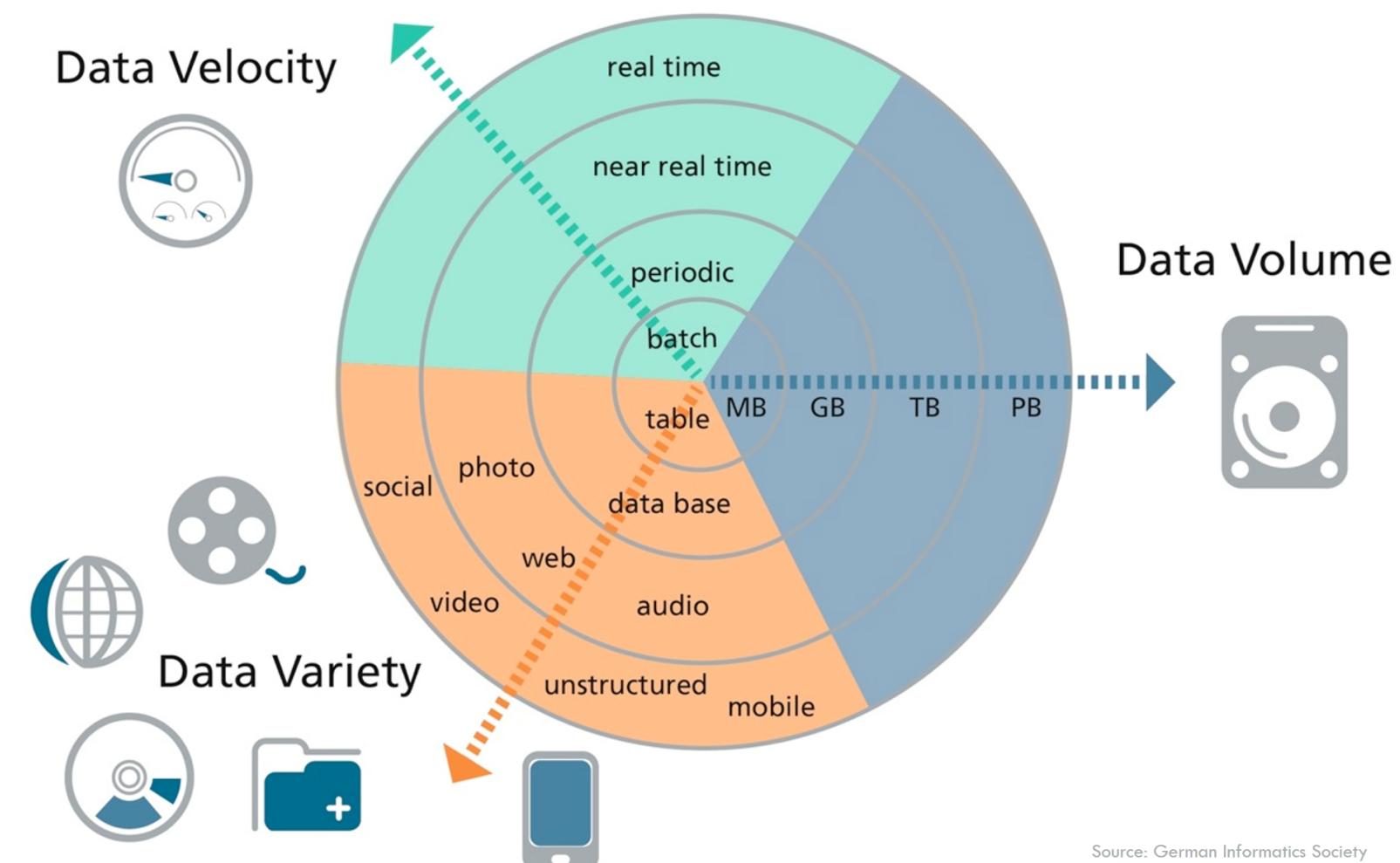
- managing and optimizing core data infrastructure,
- building and maintaining custom ingestion pipelines,
- supporting data team resources with design and performance optimization, and
- building non-SQL transformation pipelines.

---

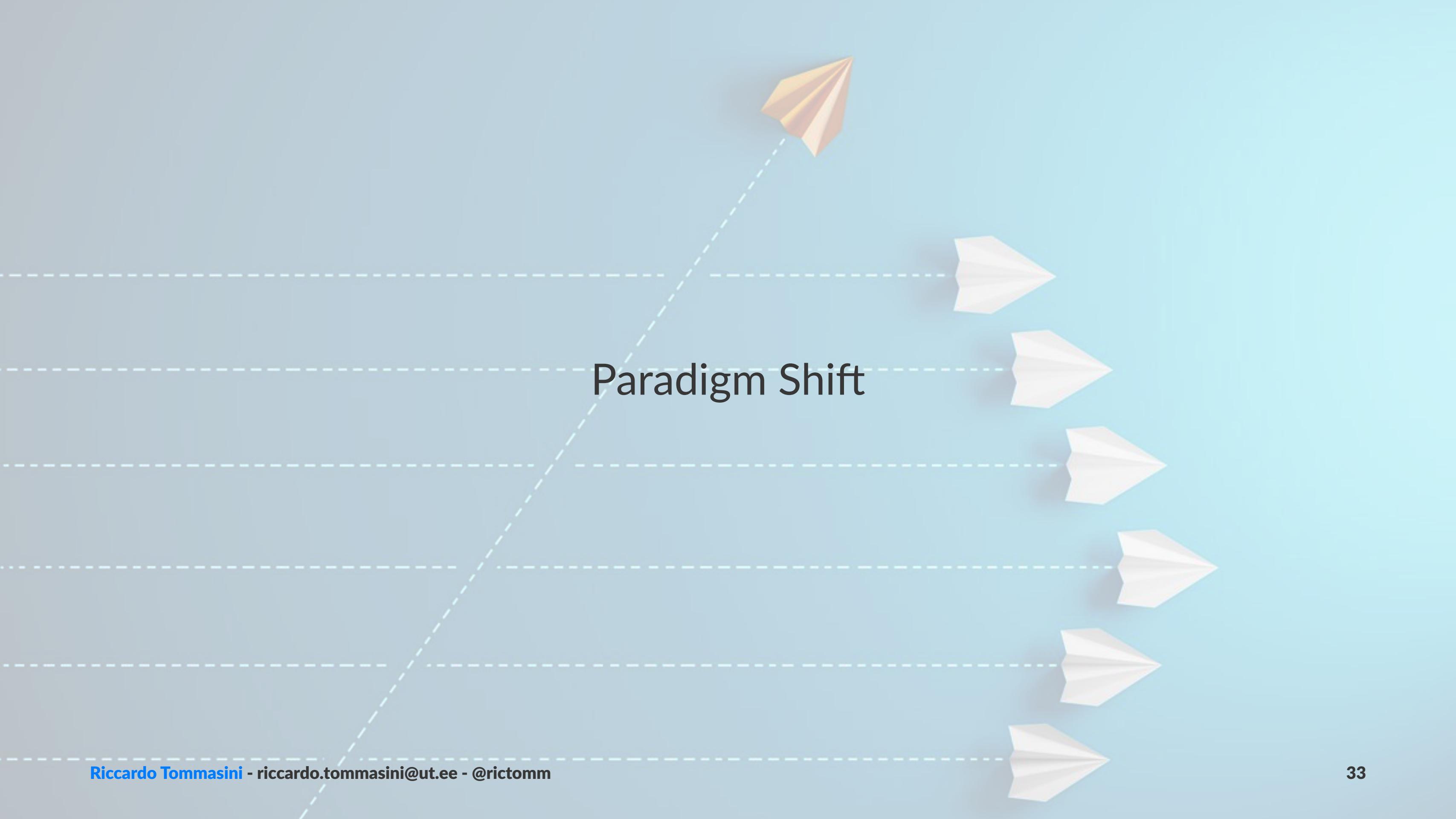
<sup>013</sup> [TristanHandy, 2019](#)

# Big Data

# Challenges 014



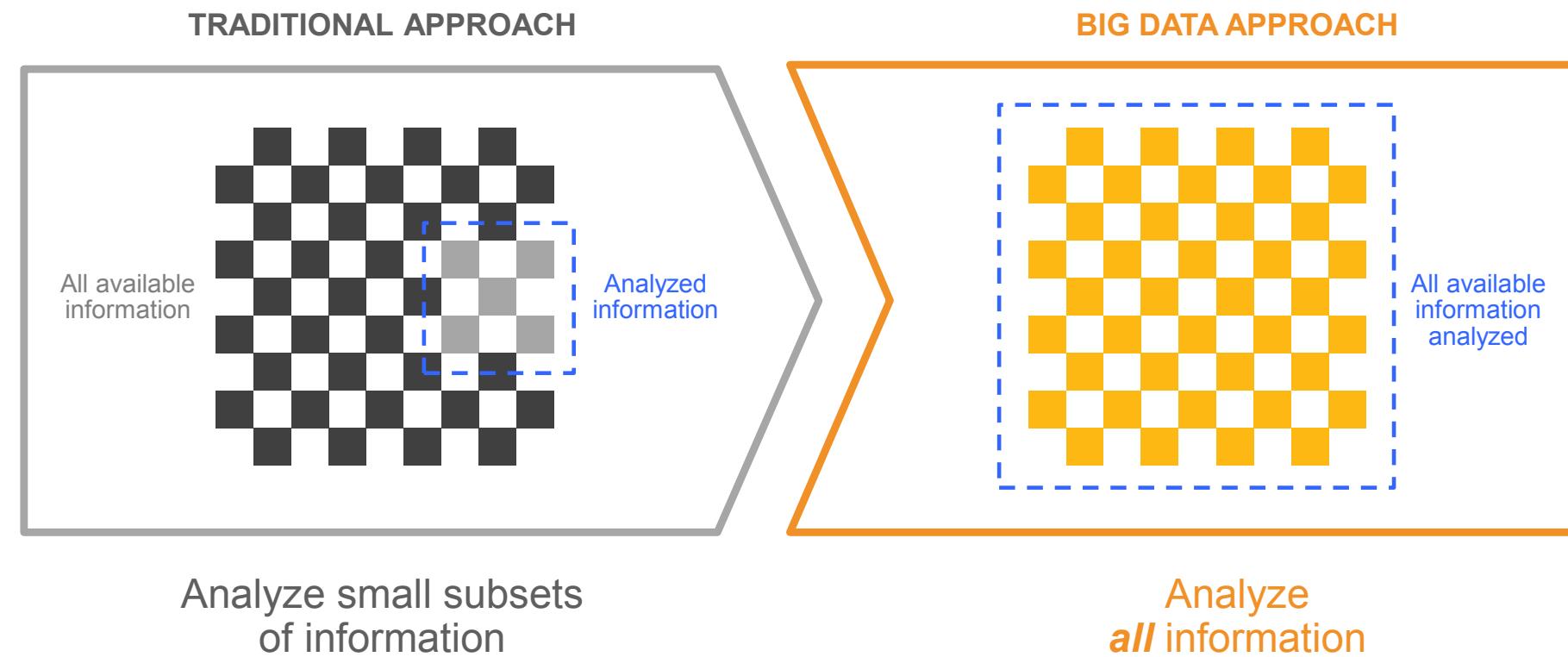
014 Lanely, 2001



Paradigm Shift

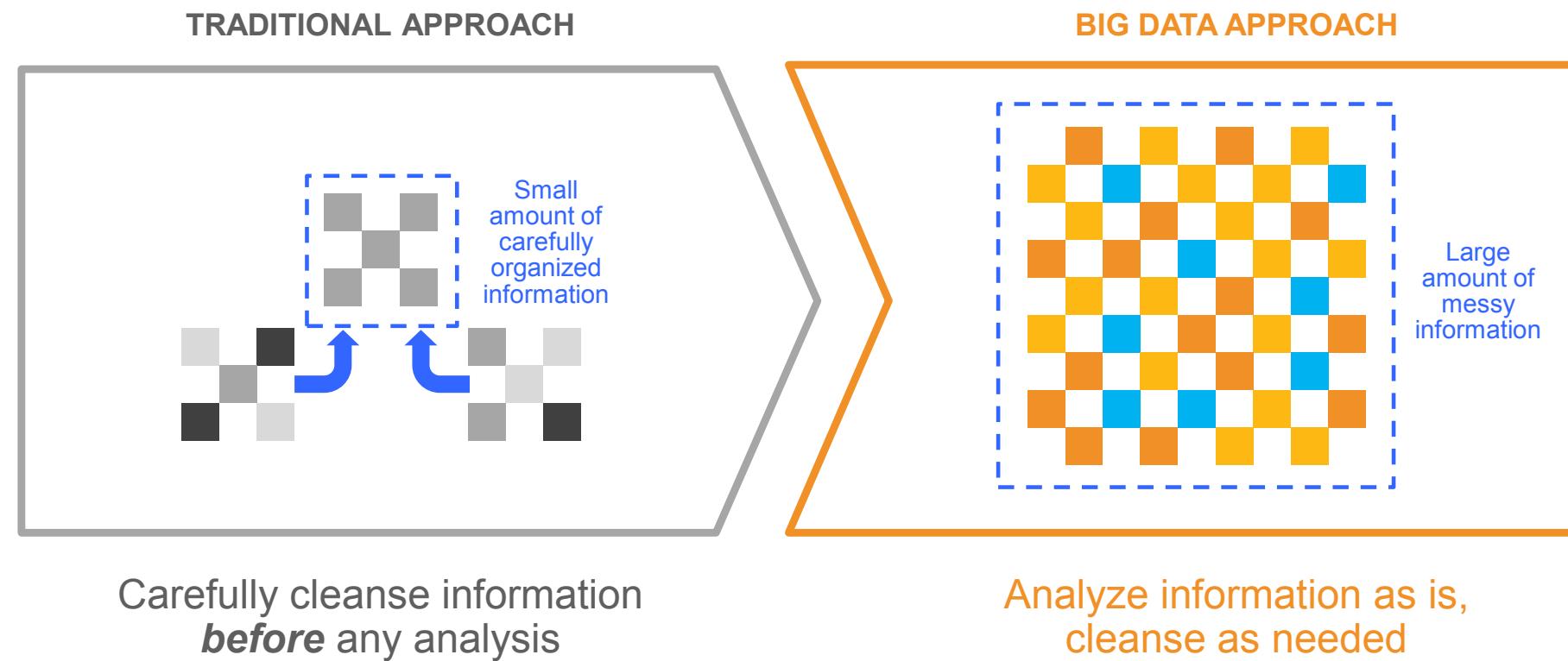
## Paradigm shifts enabled by big data

### Leverage more of the data being captured



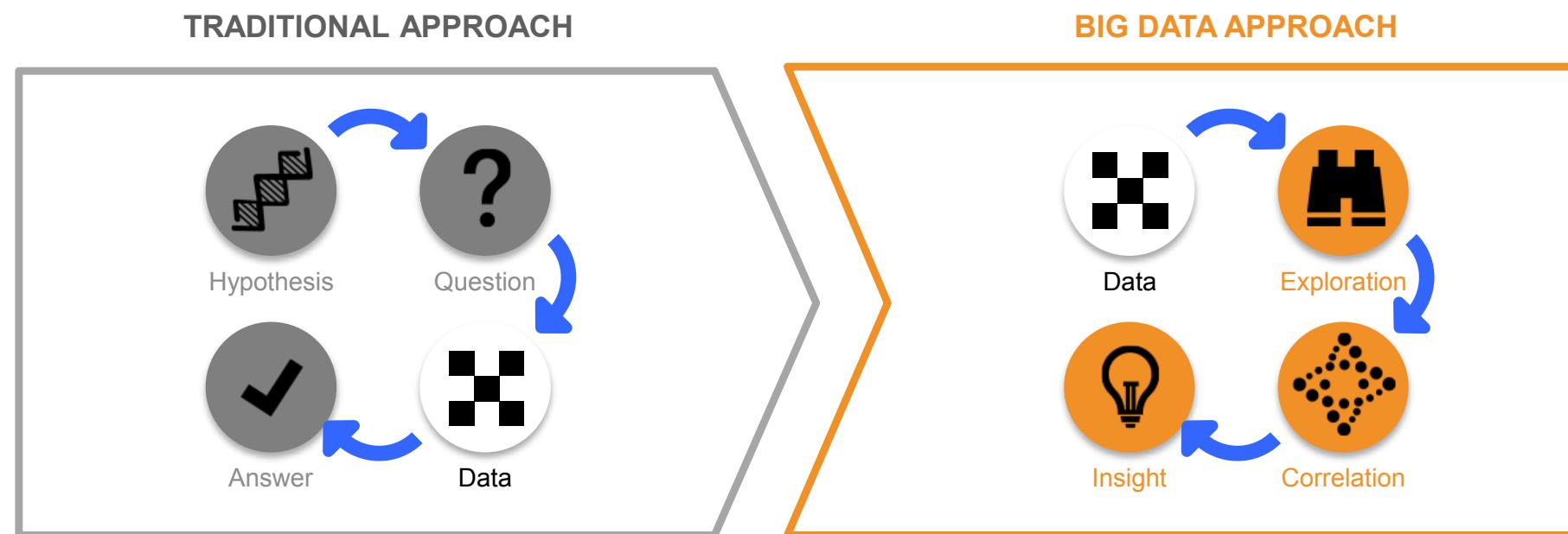
## Paradigm shifts enabled by big data

### Reduce effort required to leverage data



## Paradigm shifts enabled by big data

Data leads the way—and sometimes correlations are good enough



Start with hypothesis and test against selected data

Explore **all** data and identify correlations

## Paradigm shifts enabled by big data

### Leverage data as it is captured



Analyze data *after* it's been processed and landed in a warehouse or mart

Analyze data *in motion* as it's generated, in real-time

# Data Lake

A Data lake is a vast pool of raw data (i.e., data as they are natively, unprocessed). A data lake stands out for its high agility as it isn't limited to a warehouse's fixed configuration<sup>03</sup>.

---

<sup>03</sup> [What is Data Engineering](#)

## HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

### STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools



1

The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.



### UNSTRUCTURED DATA

1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools



2

The reservoir of water is a dataset, where you run analytics on all the data.

3

The outflow of water is the analyzed data.

4

Through this process, you are able to “sift” through all the data quickly to gain key business insights.

Full Infographic

# DATA WAREHOUSE

- Structured Data
- Schema On Write
- Data Pipelines: Extract-Transform-Load
- Processing Model: Batch

VS

# DATA LAKE

- Unstructured Data
- Schema on Read
- Data Pipelines: Extract-Load-Transform
- Processing Model: Streaming

# (Big) Data Engineer

In the context of Big Data, a data engineer must focus on **distributed systems**, and **programming languages** such as Java and Scala is recommended.

## New Tasks

Since data lake are taking data from a wide range of systems, data can be in **structured** or **unstructured** formats, and usually **not clean**, e.g., with missing fields, mismatched data types, and other data-related issues.

Therefore data engineers are challenged with the task of wrangling, cleansing, and integrating data.

# The Future of Data Engineering

- there has been a significant shift toward real-time data pipelines
- Increased connectivity between data sources and the data warehouse
- Self-service analytics via smart tools, made possible by data engineering
- Automation of Data Science functions
- Hybrid data architectures spanning on-premise and cloud environments