

Data Engineering

LTAT.02.007

Prof. Ahmed Awad

Ass. Prof Riccardo Tommasini

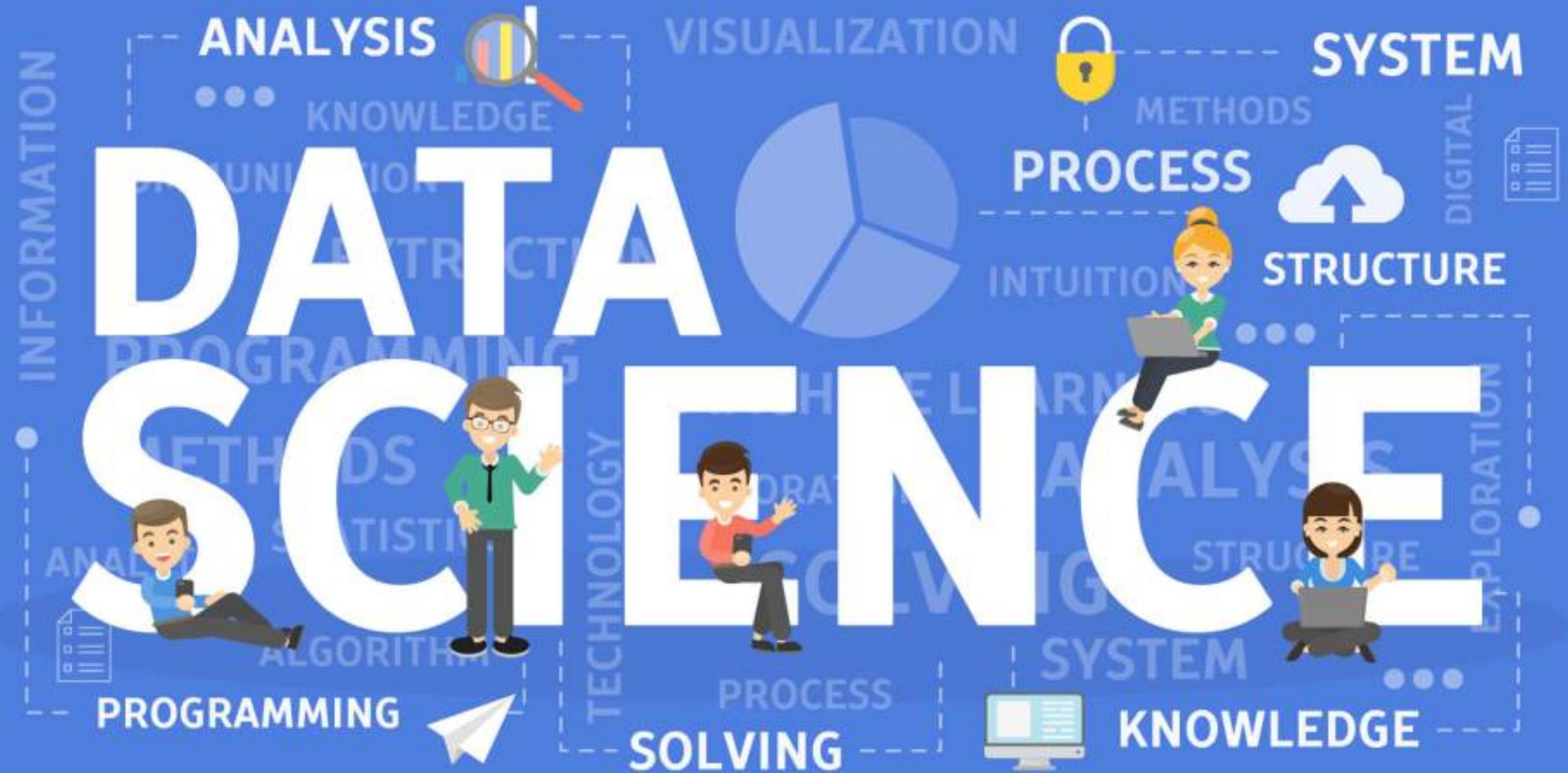
TAs: Kristo Raun, Fabiano Spiga, Mohamed Ragab



dataeng

Moodle





Course Intro

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

Course Intro

- Lectures

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading
 - 60% on 4 MCQs

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading
 - 60% on 4 MCQs
 - 40% on course project

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading
 - 60% on 4 MCQs
 - 40% on course project
 - Assignments are meant to help you work on your project and to get feedback

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading
 - 60% on 4 MCQs
 - 40% on course project
 - Assignments are meant to help you work on your project and to get feedback
 - SQL Assessment

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

Course Intro

- Lectures
 - Mostly F2F in 1037 with recording
 - Some will be online (especially delivered by Riccardo). We will use the Zoom link announced on Moodle
- Practice
 - All online, using the same zoom link announced on Moodle. Will be recorded
- Grading
 - 60% on 4 MCQs
 - 40% on course project
 - Assignments are meant to help you work on your project and to get feedback
 - SQL Assessment
 - More details on course page

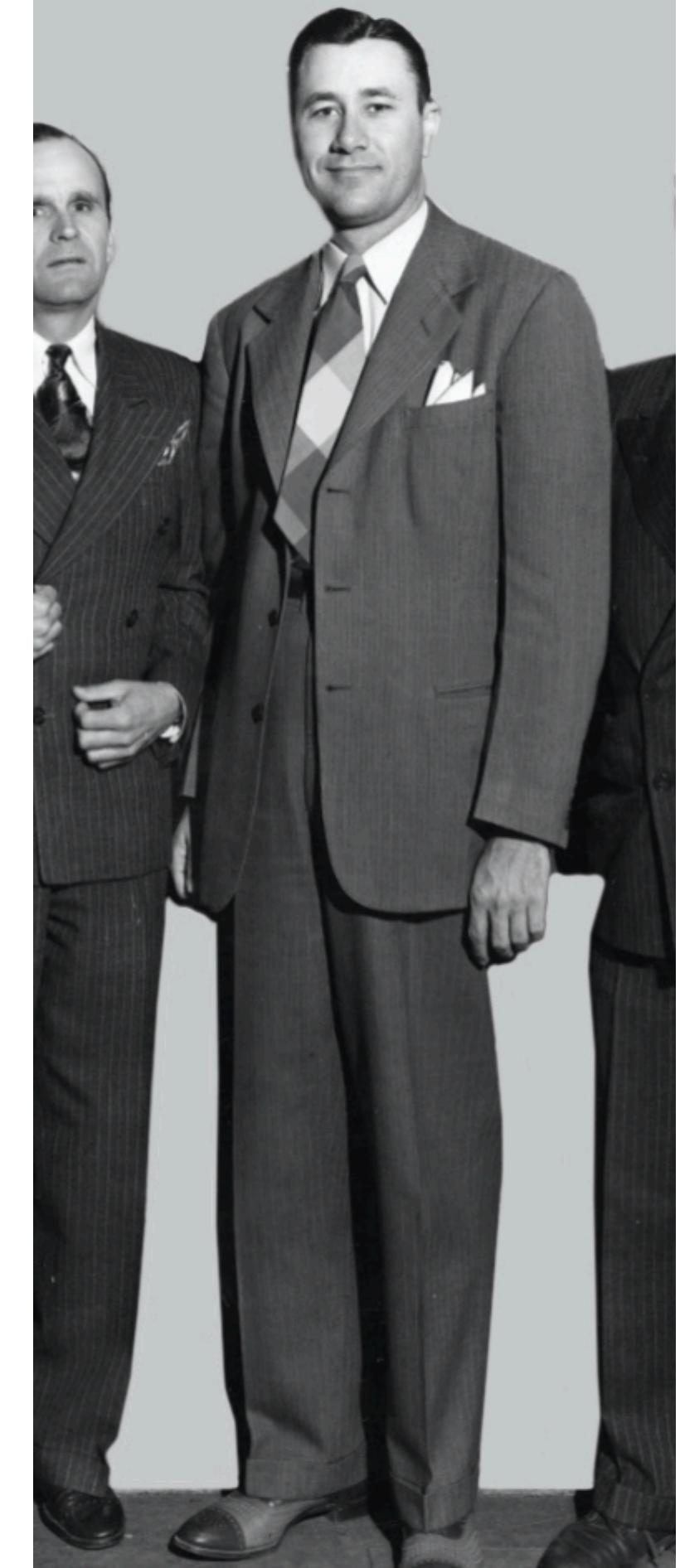
<https://courses.cs.ut.ee/2020/dataeng>

Quote

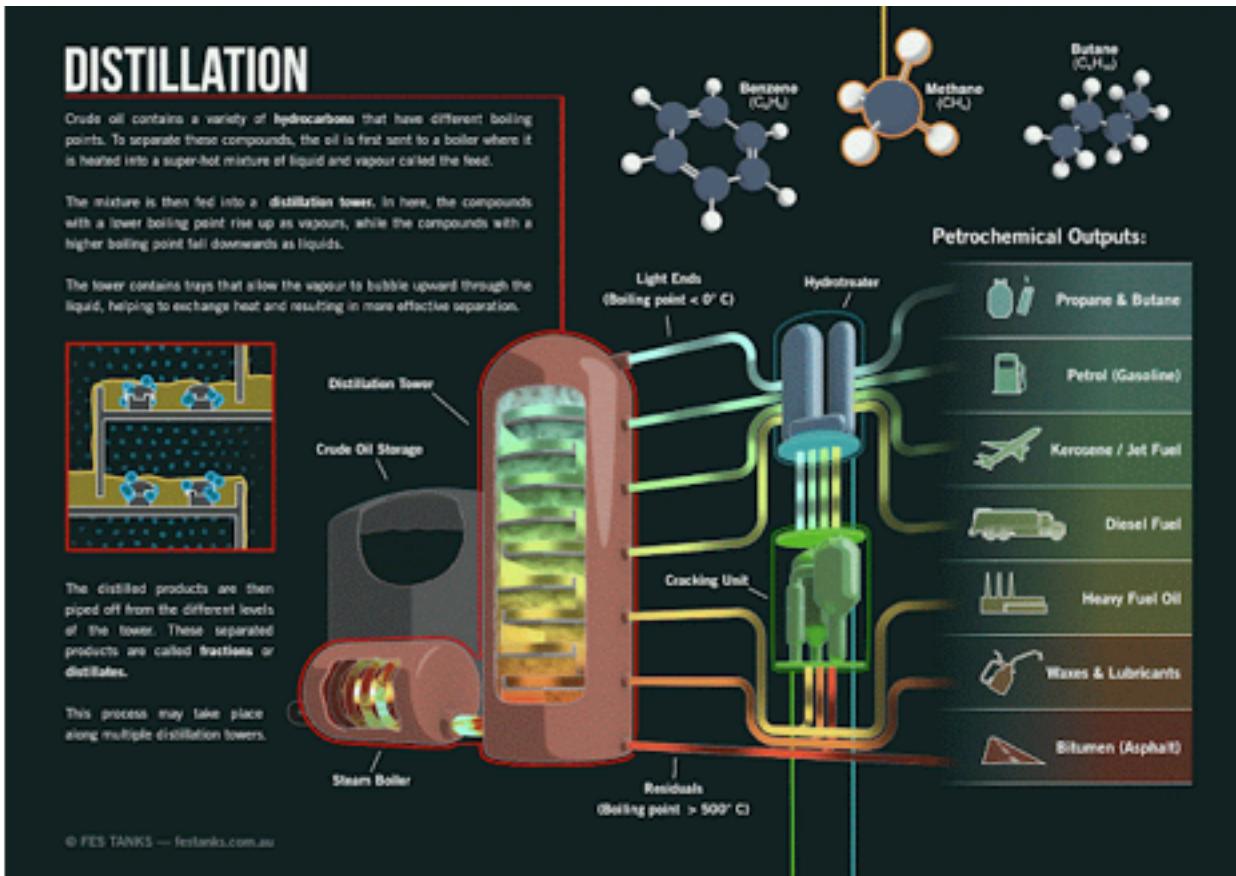
“A scientist can discover a new star, but he cannot make one.

He would have to ask an engineer to do it for him.”

– *Gordon Lindsay Glegg*



Data Science is...⁰¹



...refining crude oil

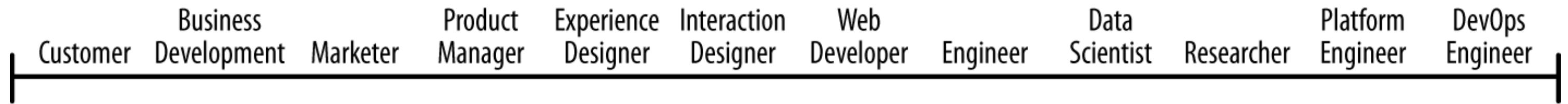
⁰¹ Source

Data Engineering is...



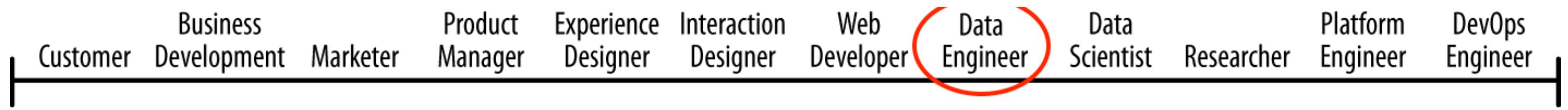
...build the refinery.

Roles in a Data Science Project⁰²



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

Roles in a Data Science Project⁰²



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



Data Engineer

The Data Engineer

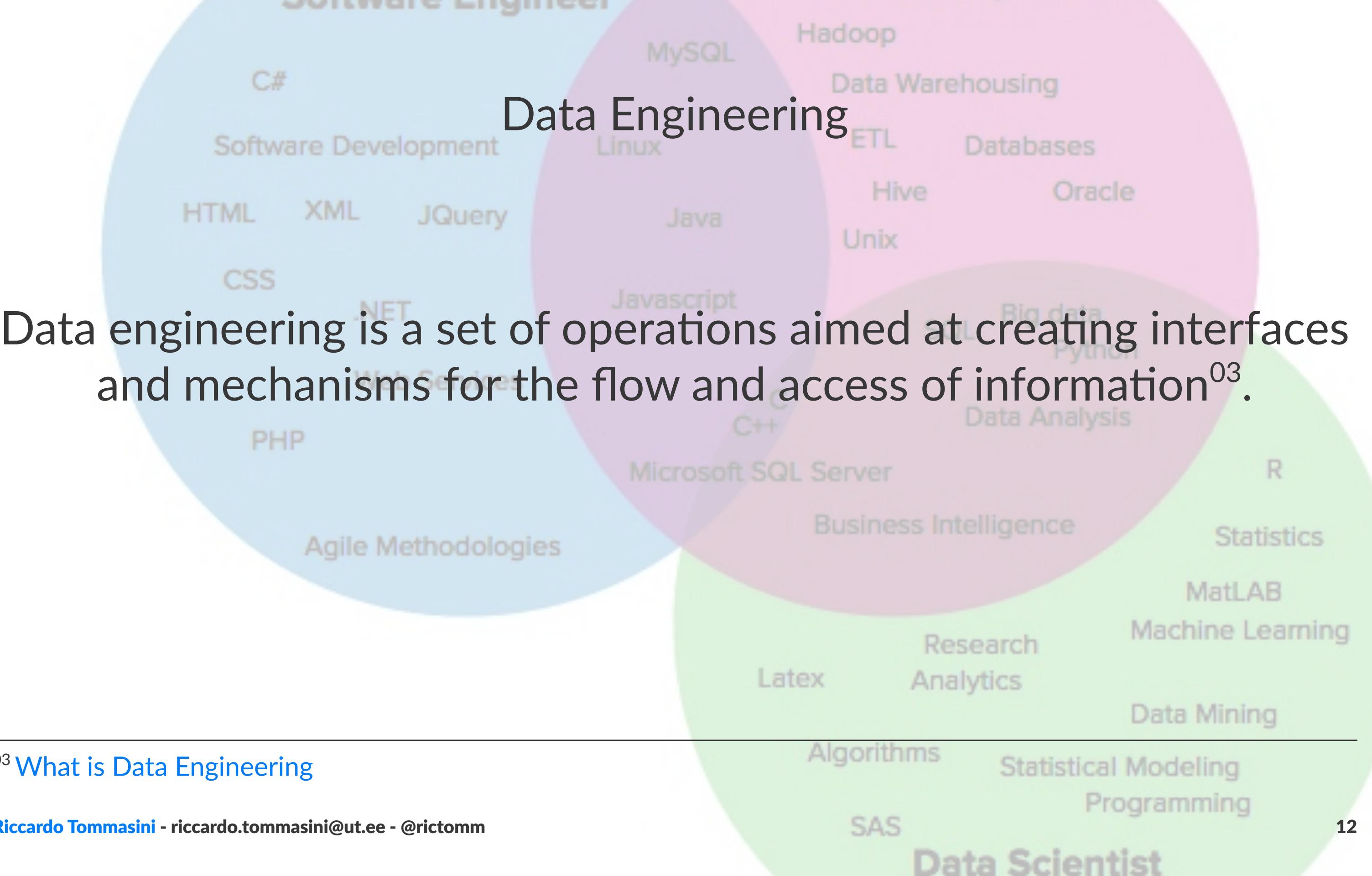
A dedicated specialist that maintain data available and usable by others (Data Scientists).⁰³

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

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

⁰³ [What is Data Engineering](#)

⁰⁴ [Source: The Rise of Data Engineer](#)



⁰³ [What is Data Engineering](#)

rt 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
```

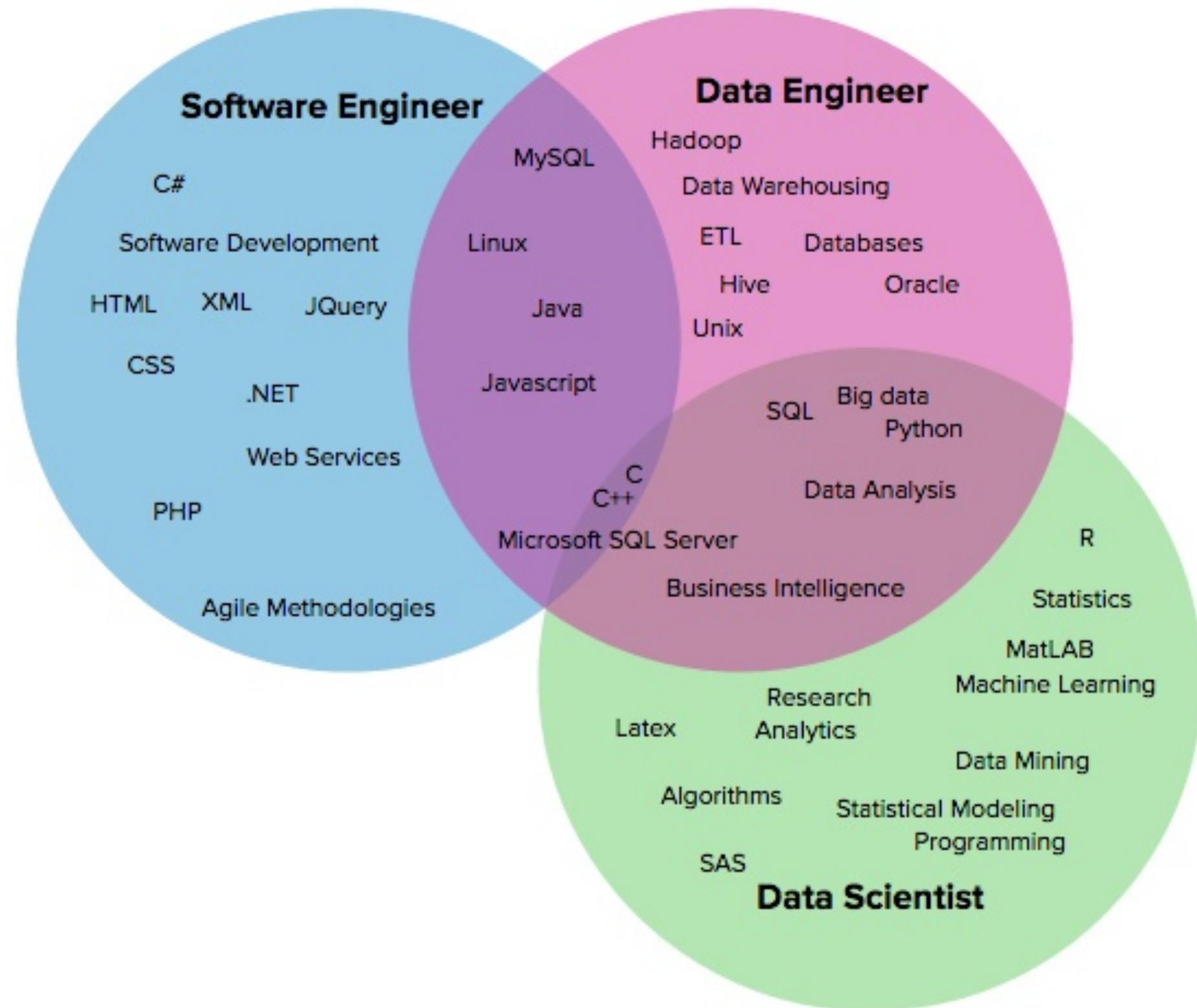
323

4.4K

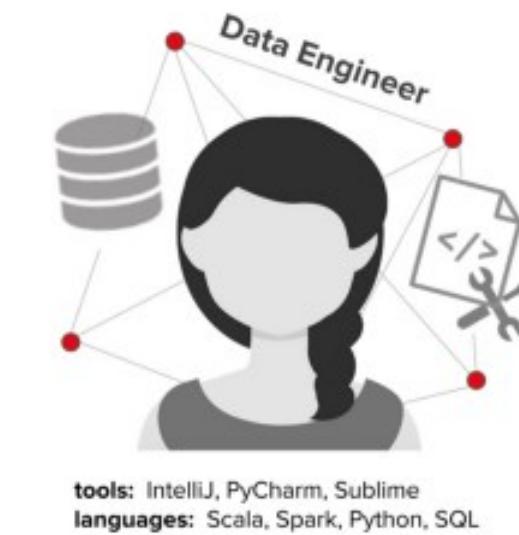
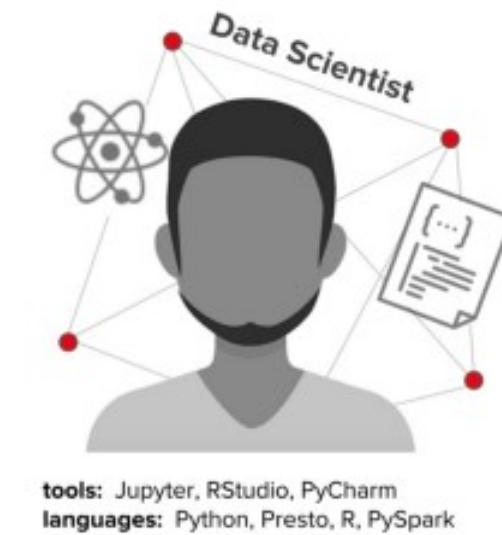
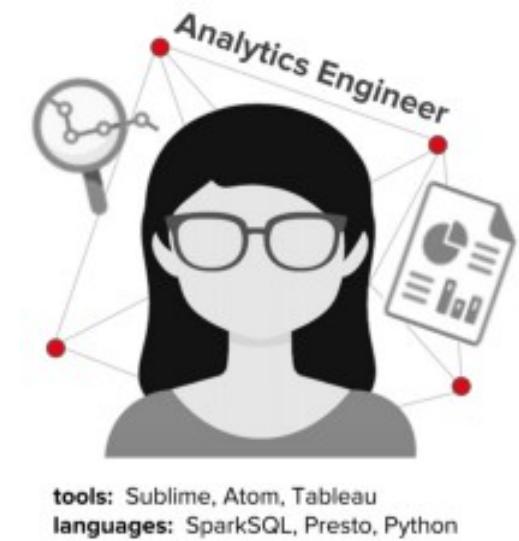
28K



[Show this thread](#)

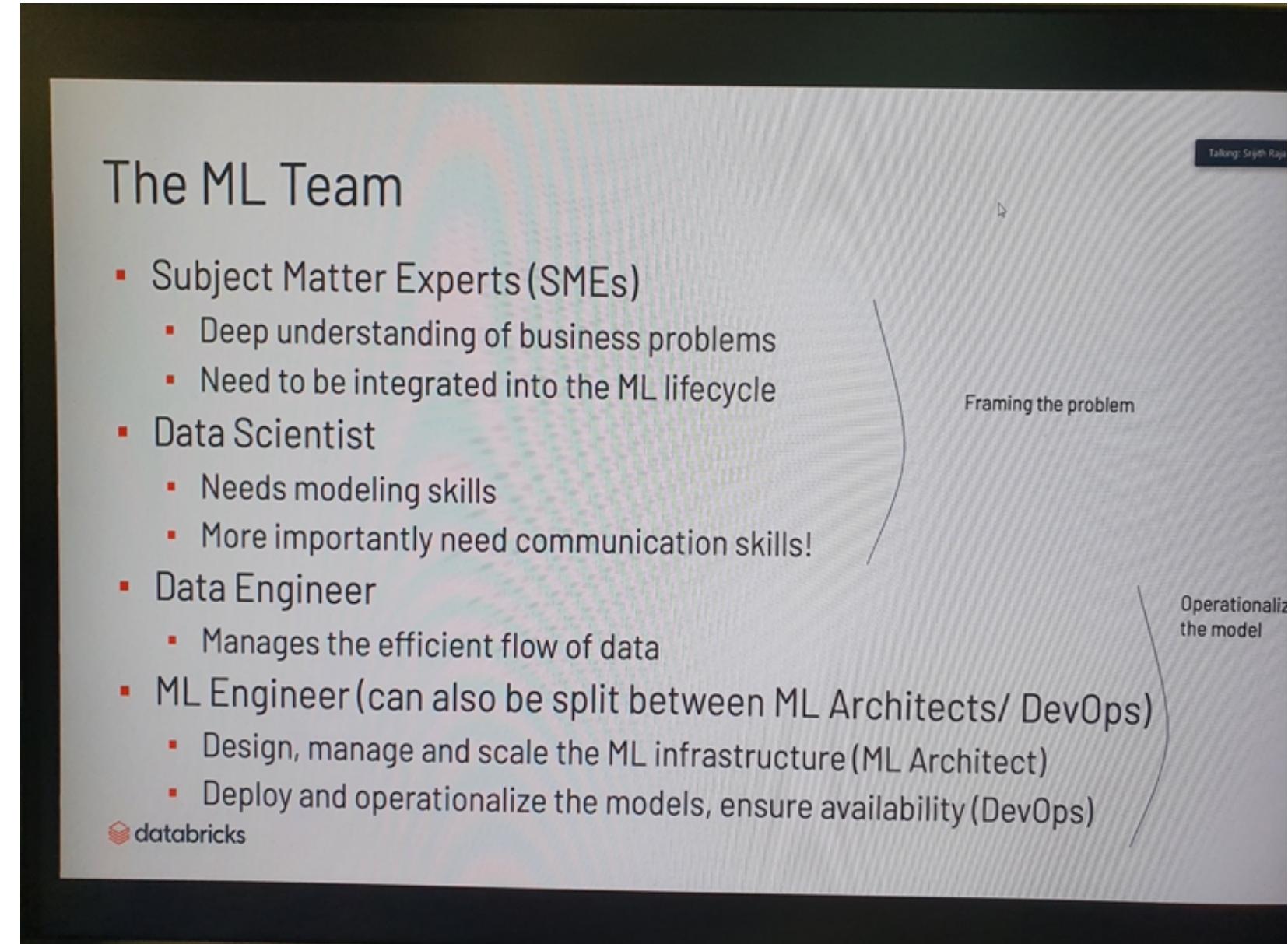


Netflix's Perspective⁰⁵



⁰⁵ Netflix Innovation

Data Bricks



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.

The Knowledge Scientist⁰⁶



⁰⁶ [The Manifesto](#)

Philosophy of (Data) Science⁰⁷



⁰⁷ 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.*⁰⁸

⁰⁸ Def

Wikipedia

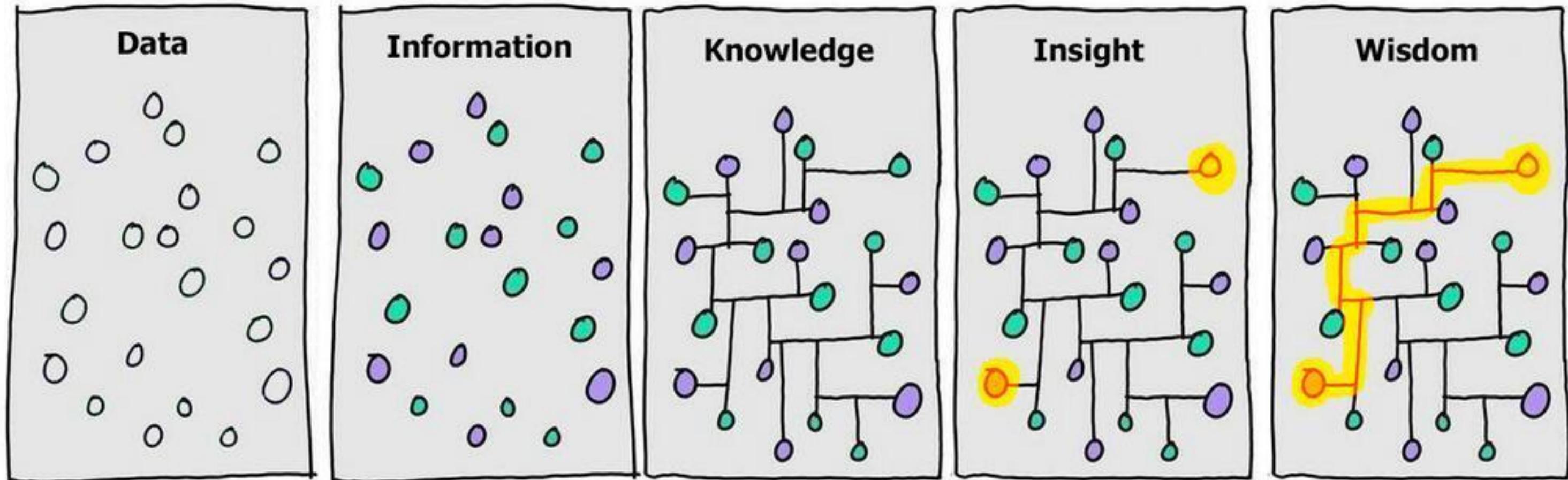
Data (treated as singular, plural, or as a mass noun) is any sequence of one or more symbols given meaning by specific act(s) of interpretation⁰⁹

⁰⁹ Data in Computing)

DIKW Pyramid



Graph View



Data about data



Data Semantics

semantics

/sɪ'mantɪks/ 

noun

the branch of linguistics and logic concerned with meaning. The two main areas are *logical semantics*, concerned with matters such as sense and reference and presupposition and implication, and *lexical semantics*, concerned with the analysis of word meanings and relations between them.

- the meaning of a word, phrase, or text.

plural noun: **semantics**

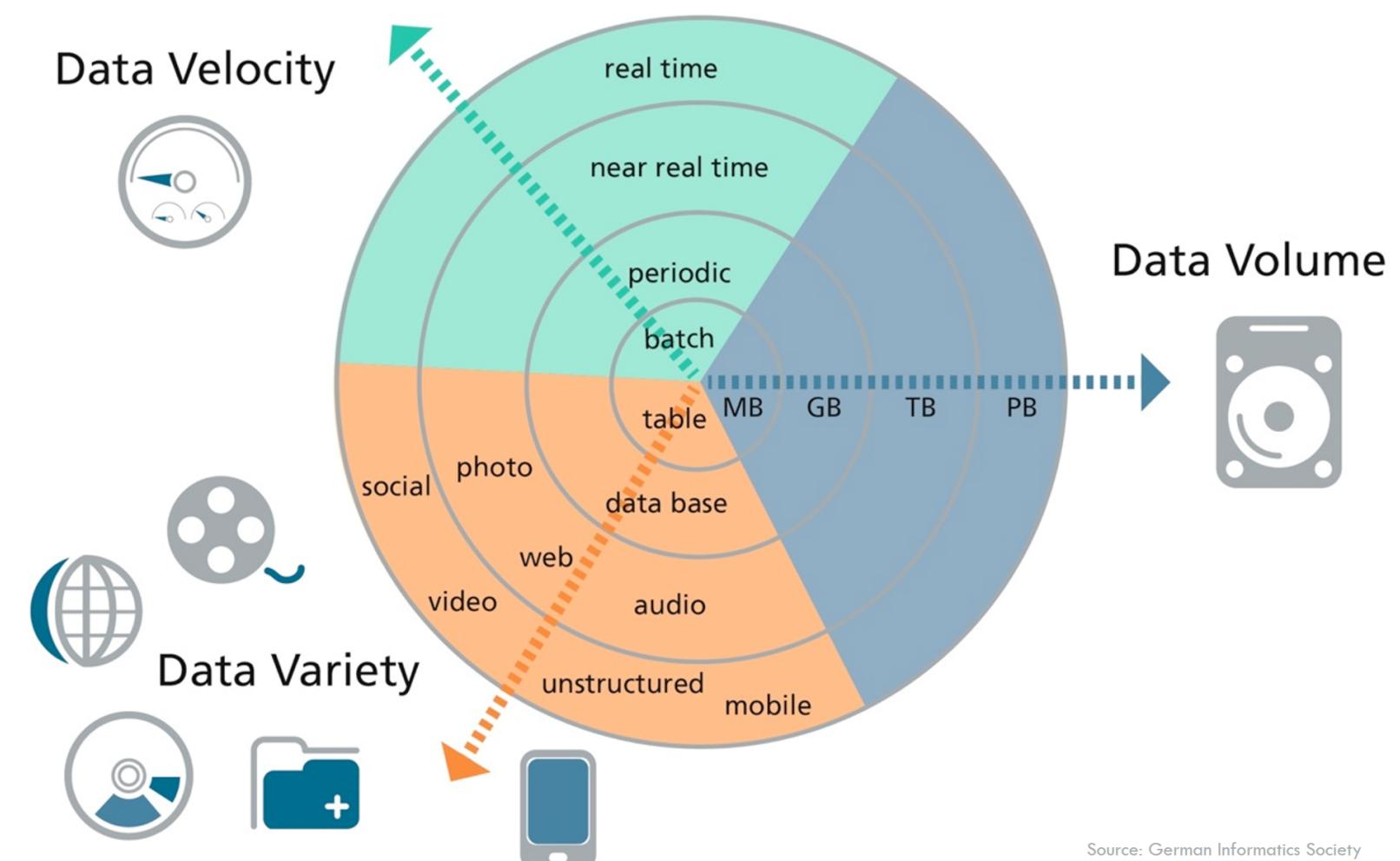
"such quibbling over semantics may seem petty stuff"



Translations, word origin, and more definitions

Big Data

Challenges 014

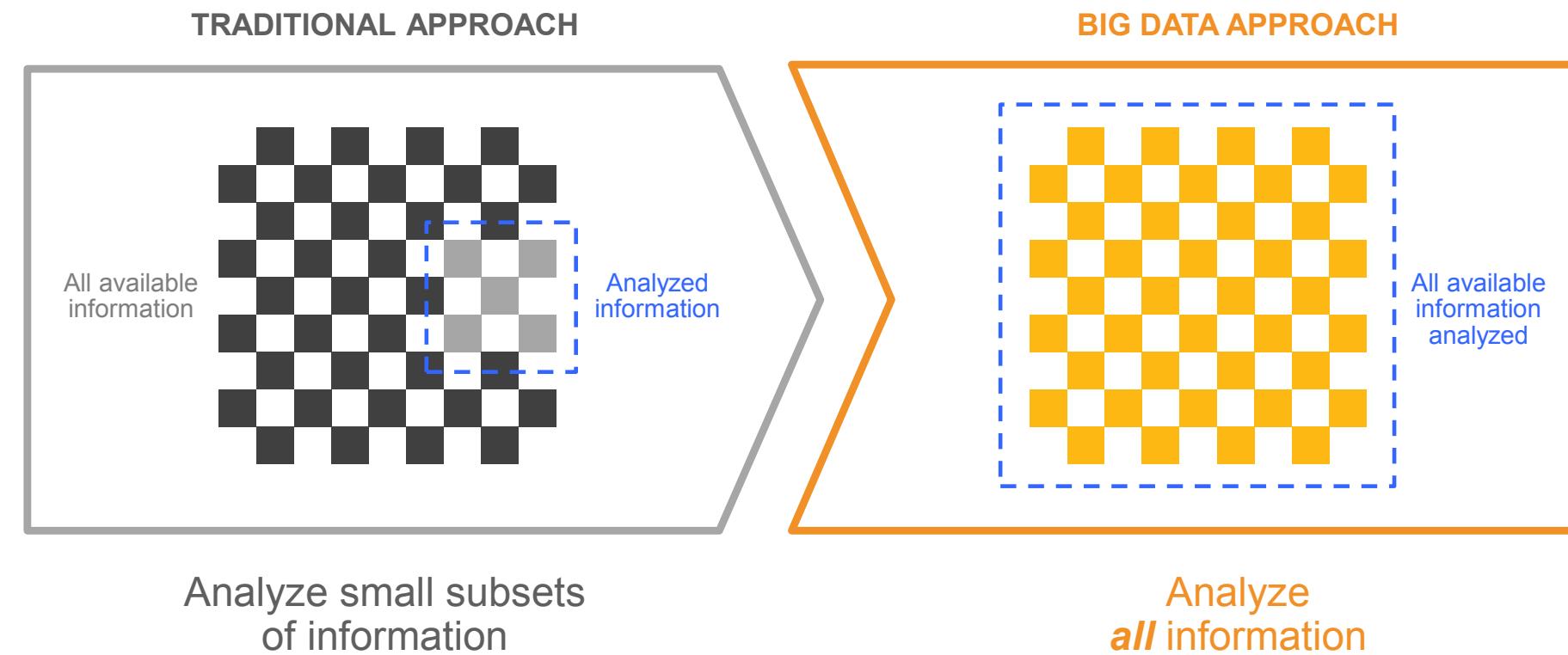


014 Lanely, 2001

Paradigm Shift

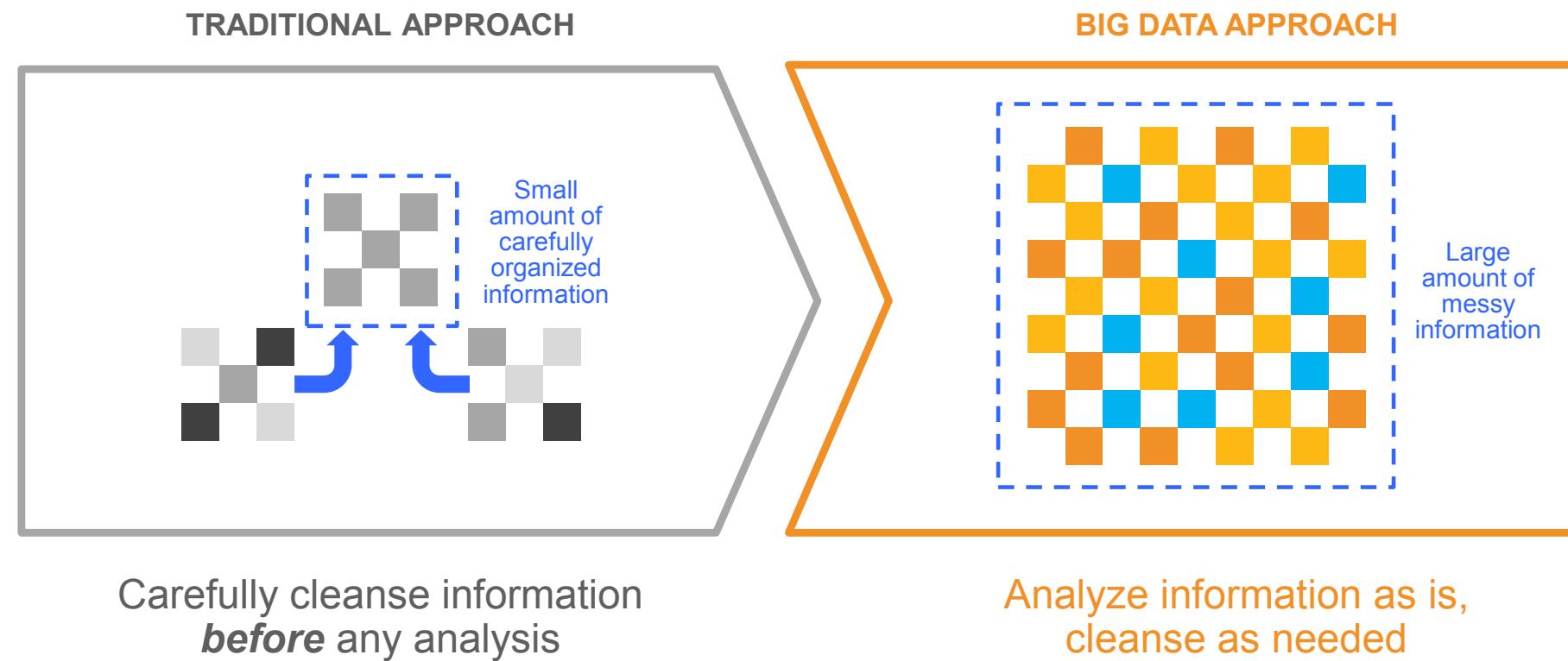
Paradigm shifts enabled by big data

Leverage more of the data being captured



Paradigm shifts enabled by big data

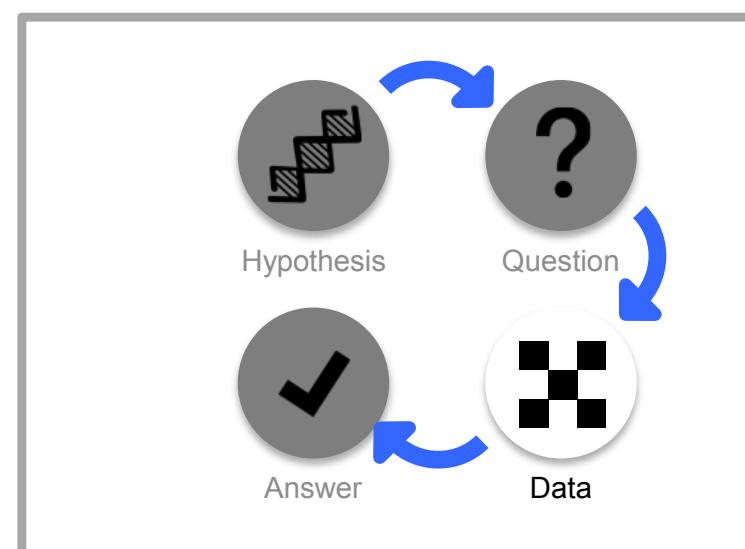
Reduce effort required to leverage data



Paradigm shifts enabled by big data

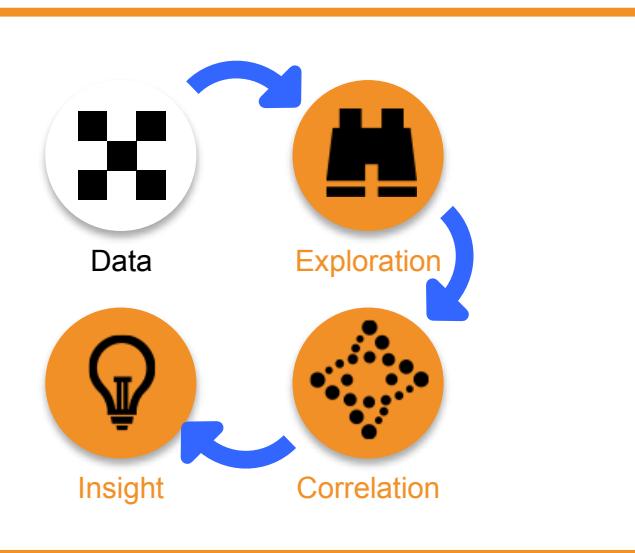
Data leads the way—and sometimes correlations are good enough

TRADITIONAL APPROACH



Start with hypothesis and test against selected data

BIG DATA APPROACH



Explore **all** data and identify correlations

Paradigm shifts enabled by big data

Leverage data as it is captured



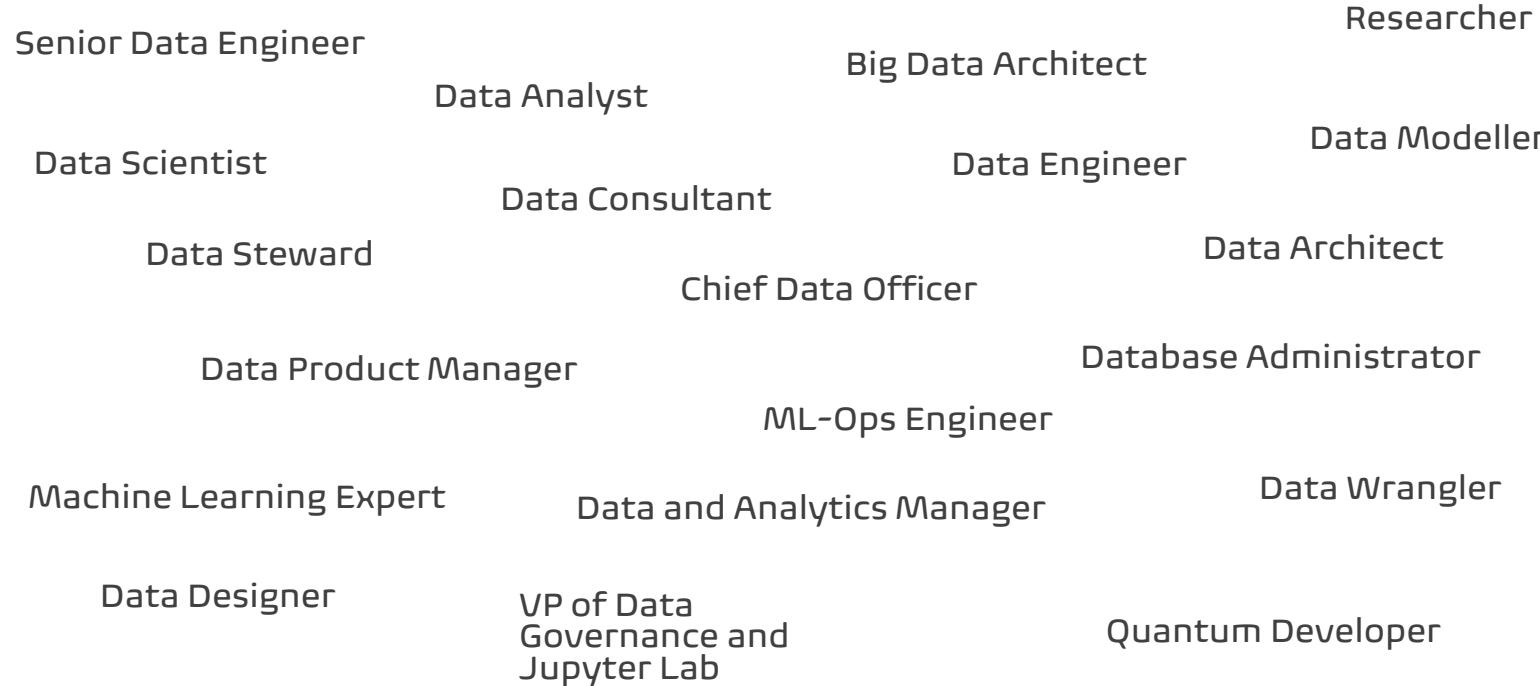
Analyze data *after* it's been processed and landed in a warehouse or mart

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

New Roles

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

Profession

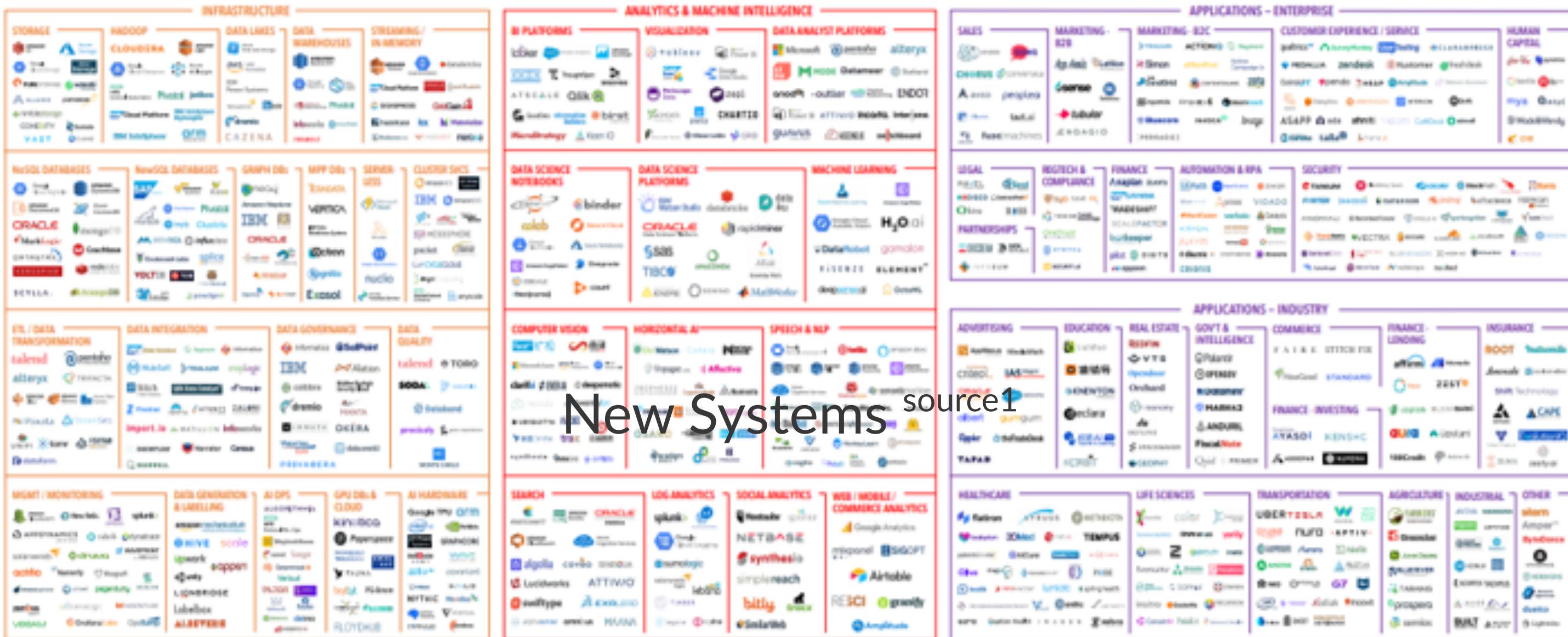


New Tasks

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.

DATA & AI LANDSCAPE 2020



New Systems



source¹ <https://www.saagie.com/blog/our-extended-big-data-ai-recap/>

Riccardo Tommasini - riccardo.tommasini@ut.ee - @rictomm

Infrastructure



Data Systems and Databases

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**

⁰³ [What is Data Engineering](#)

Data Warehouse vs Data Bases

Surprisingly, Data Warehouse isn't a regular database.

Data Warehouse vs Data Bases

Surprisingly, Data Warehouse isn't a regular database.

- A database normalizes data separating them into tables and avoiding redundancies

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

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

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

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 analytic.

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 analytic.
- a Data Warehouse allows simple queries

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 analytic.
- a Data Warehouse allows simple queries
- supports versioning for complex analysis

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 configuration⁰³.

⁰³ [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.



2

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



UNSTRUCTURED DATA

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



3

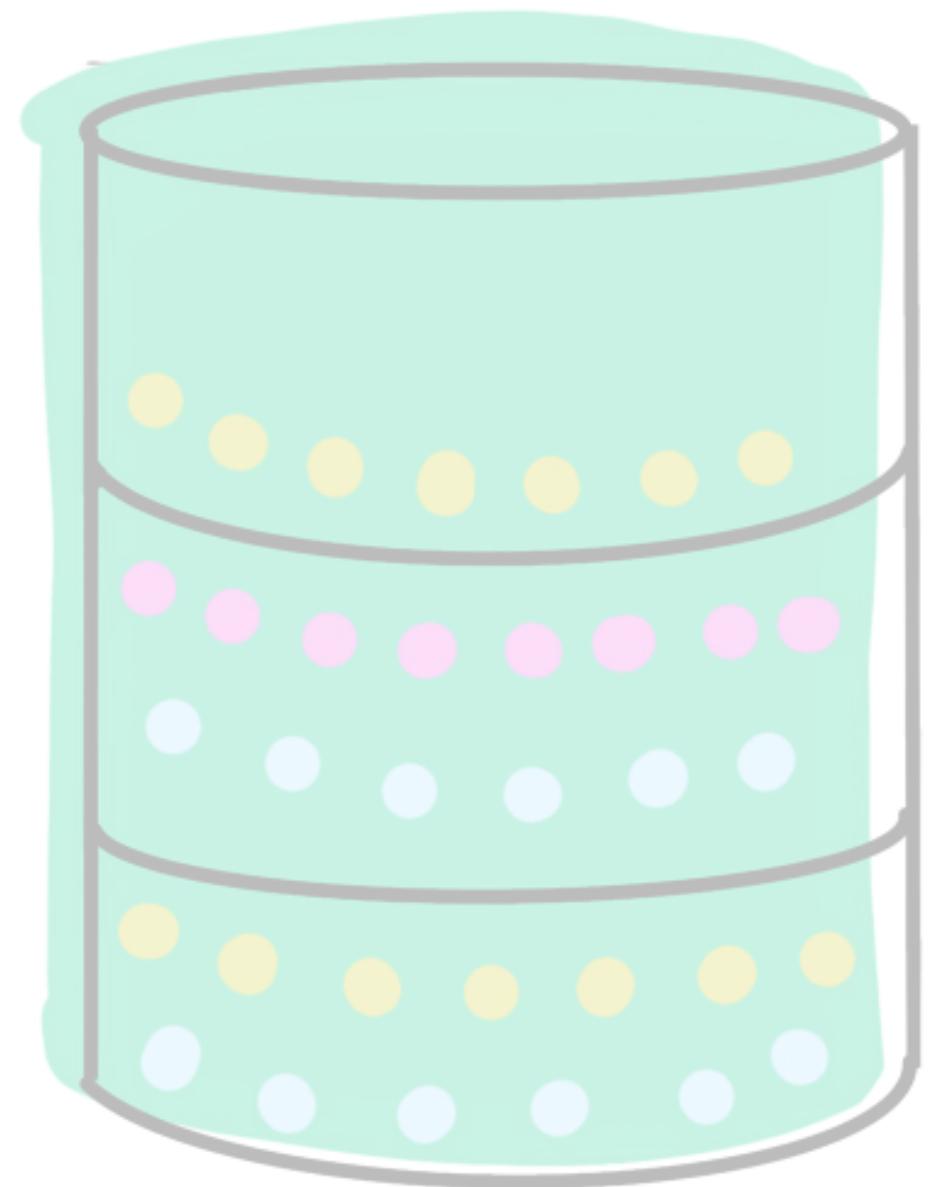
The outflow of water is the analyzed data.

4

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

Full Infographic

DATA WAREHOUSE



Data Lake vs Data Warehouse

DATA LAKE

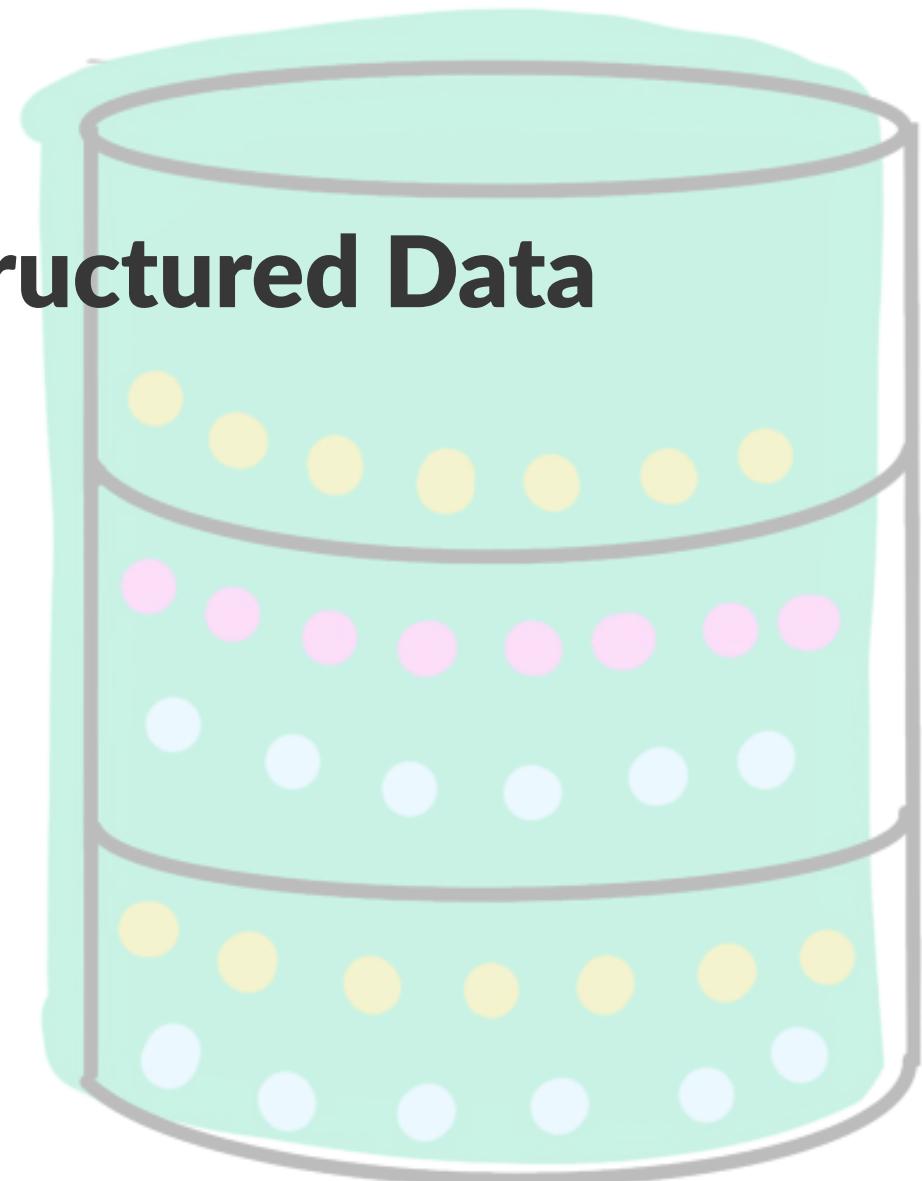
VS



DATA WAREHOUSE

Data Lake vs Data Warehouse

- Structured Data



VS



DATA WAREHOUSE

Data Lake vs Data Warehouse

- Structured Data
- Schema On Write



VS



DATA WAREHOUSE

Data Lake vs Data Warehouse

- Structured Data
- Schema On Write
- Data Pipelines: Extract-Transform-Load

VS

DATA LAKE



DATA WAREHOUSE

Data Lake vs Data Warehouse

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

VS

DATA LAKE



DATA WAREHOUSE

Data Lake vs Data Warehouse

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

VS

DATA LAKE



DATA WAREHOUSE

Data Lake vs Data Warehouse

DATA LAKE

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

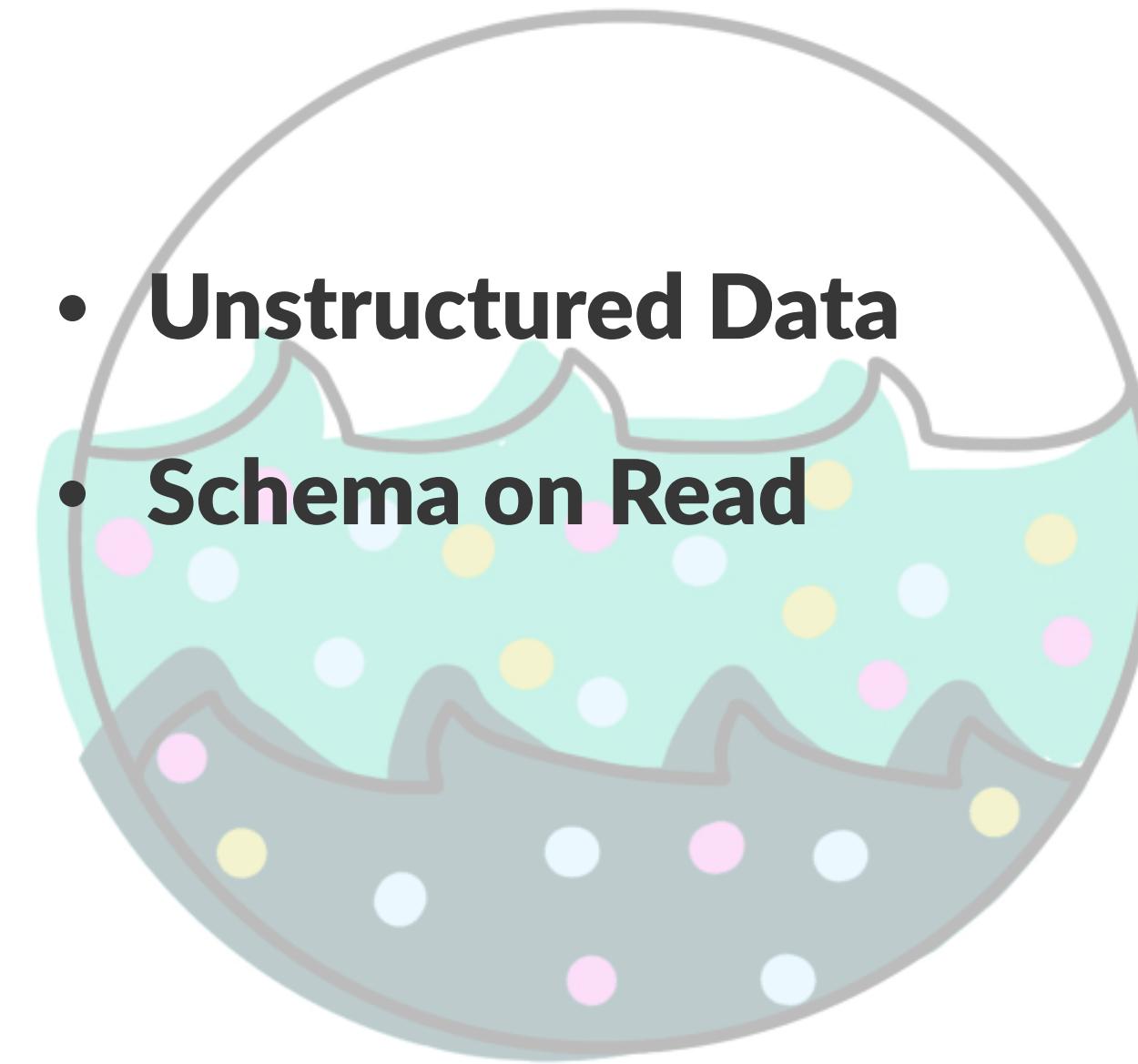
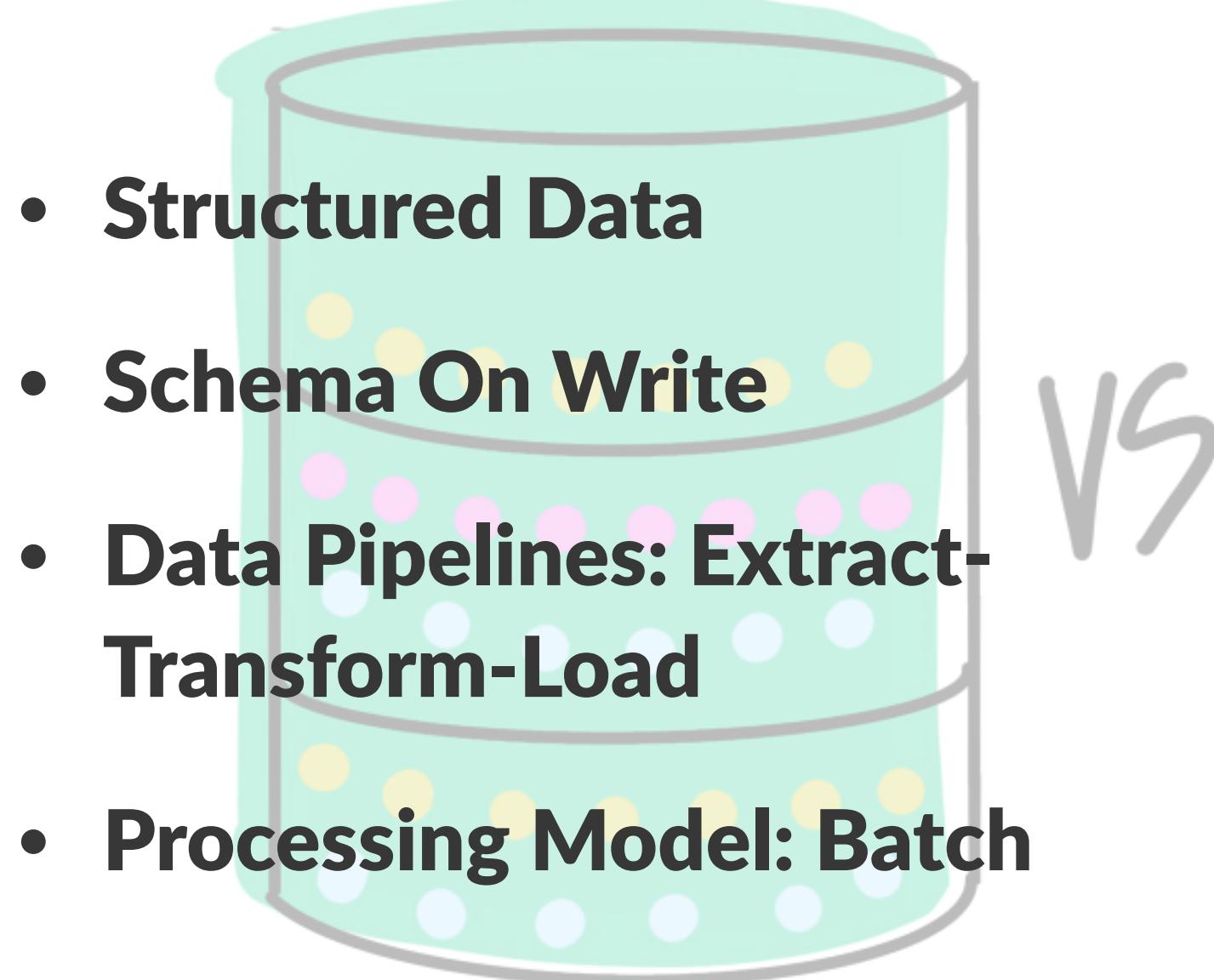
VS

- Unstructured Data

DATA WAREHOUSE

Data Lake vs Data Warehouse

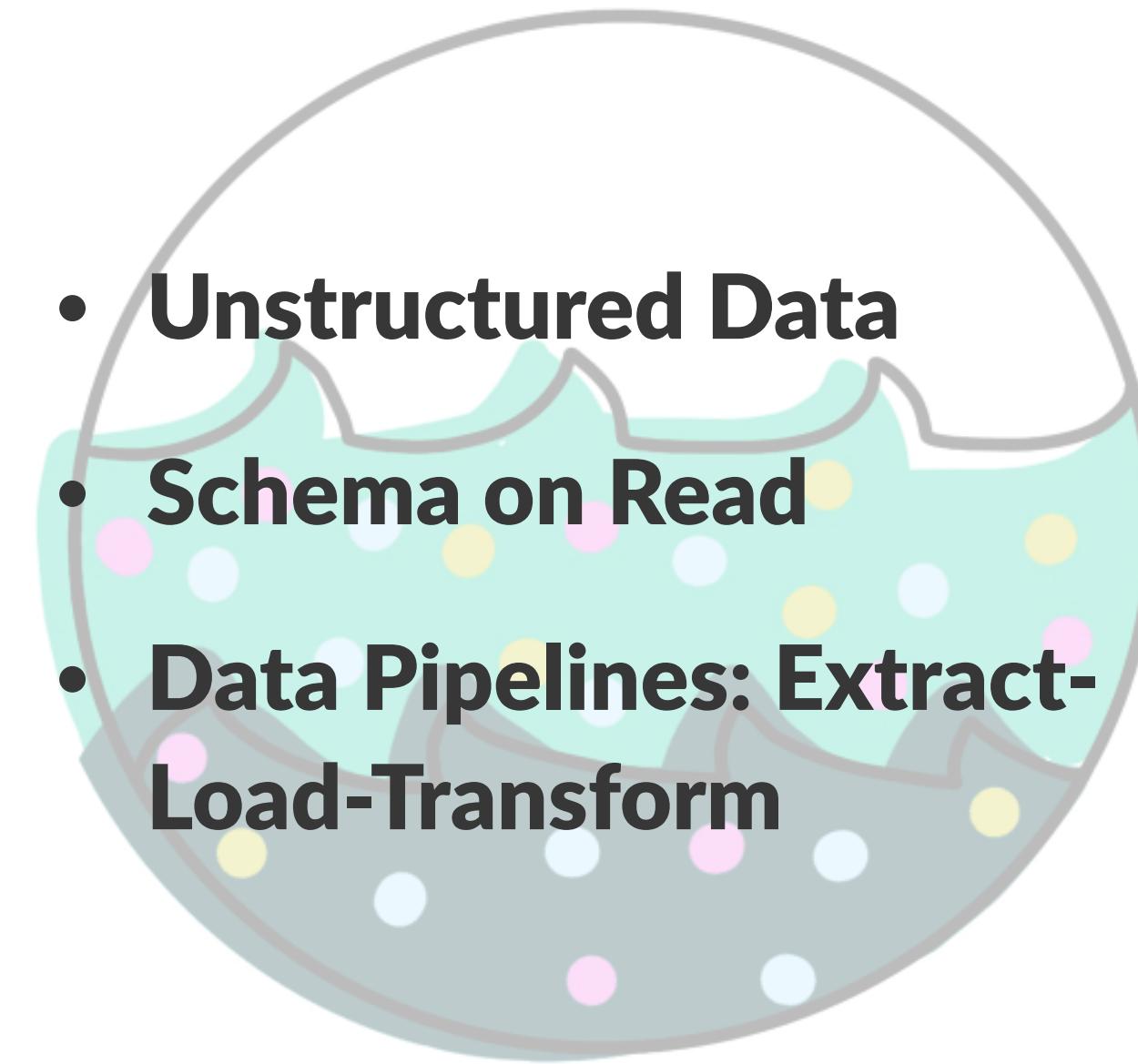
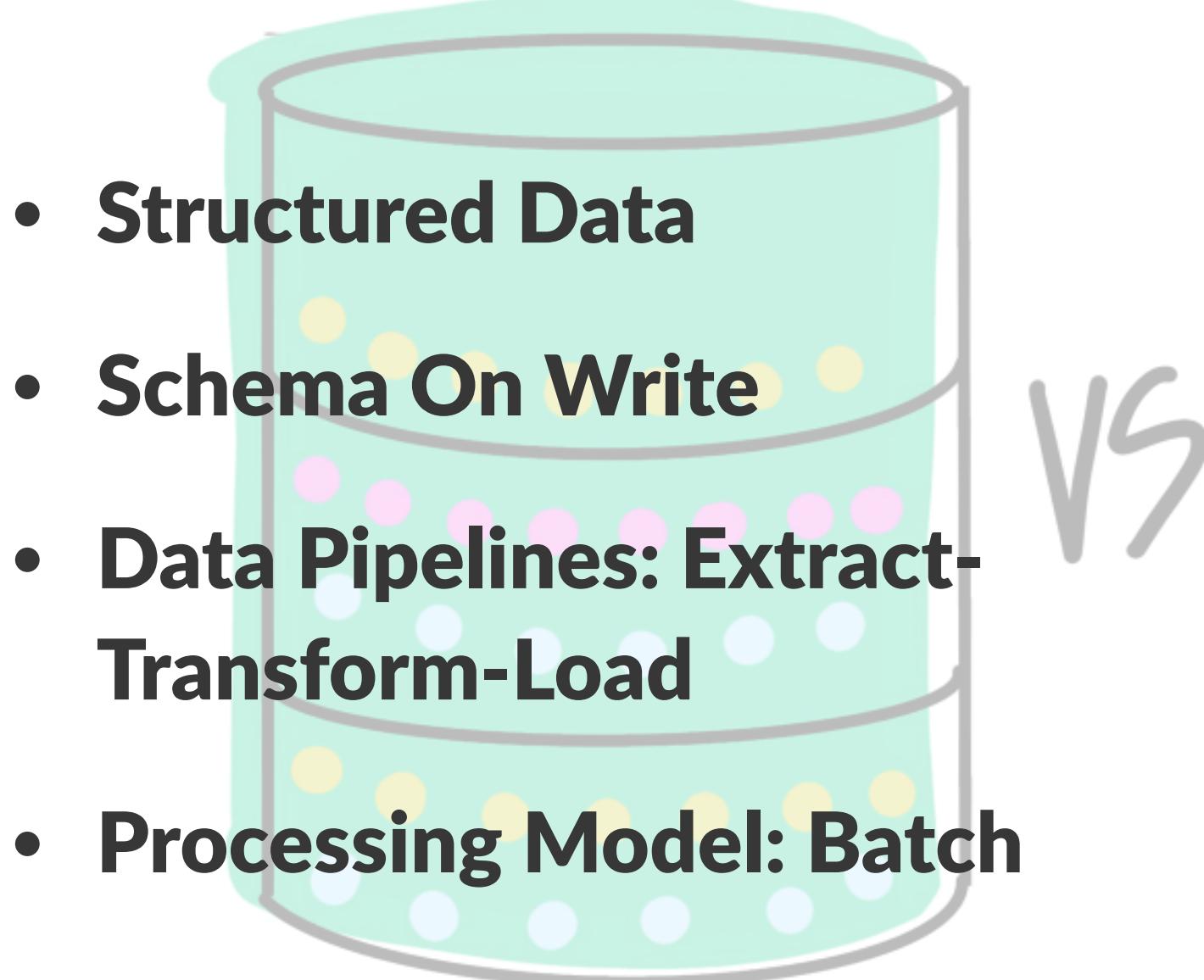
DATA LAKE



DATA WAREHOUSE

Data Lake vs Data Warehouse

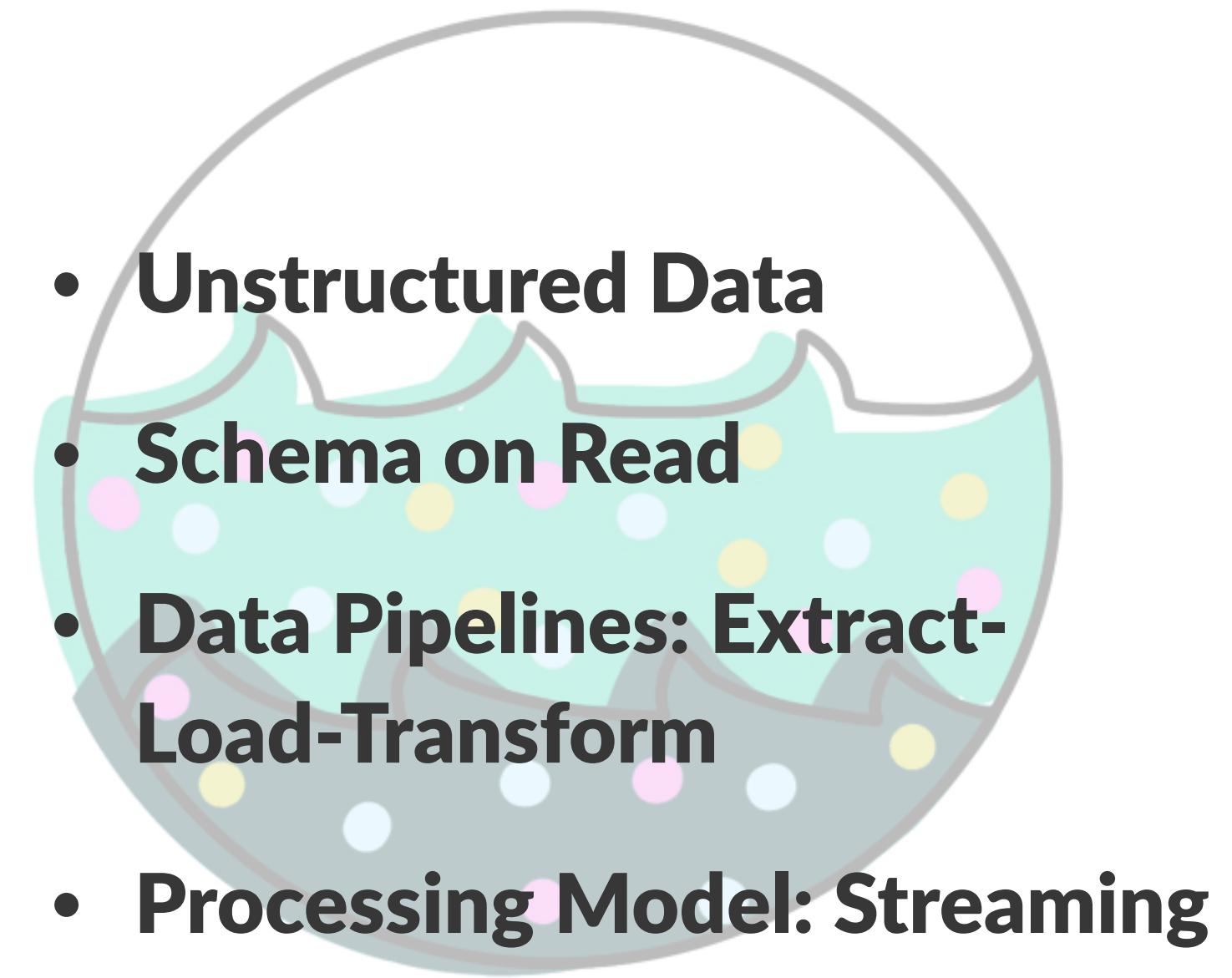
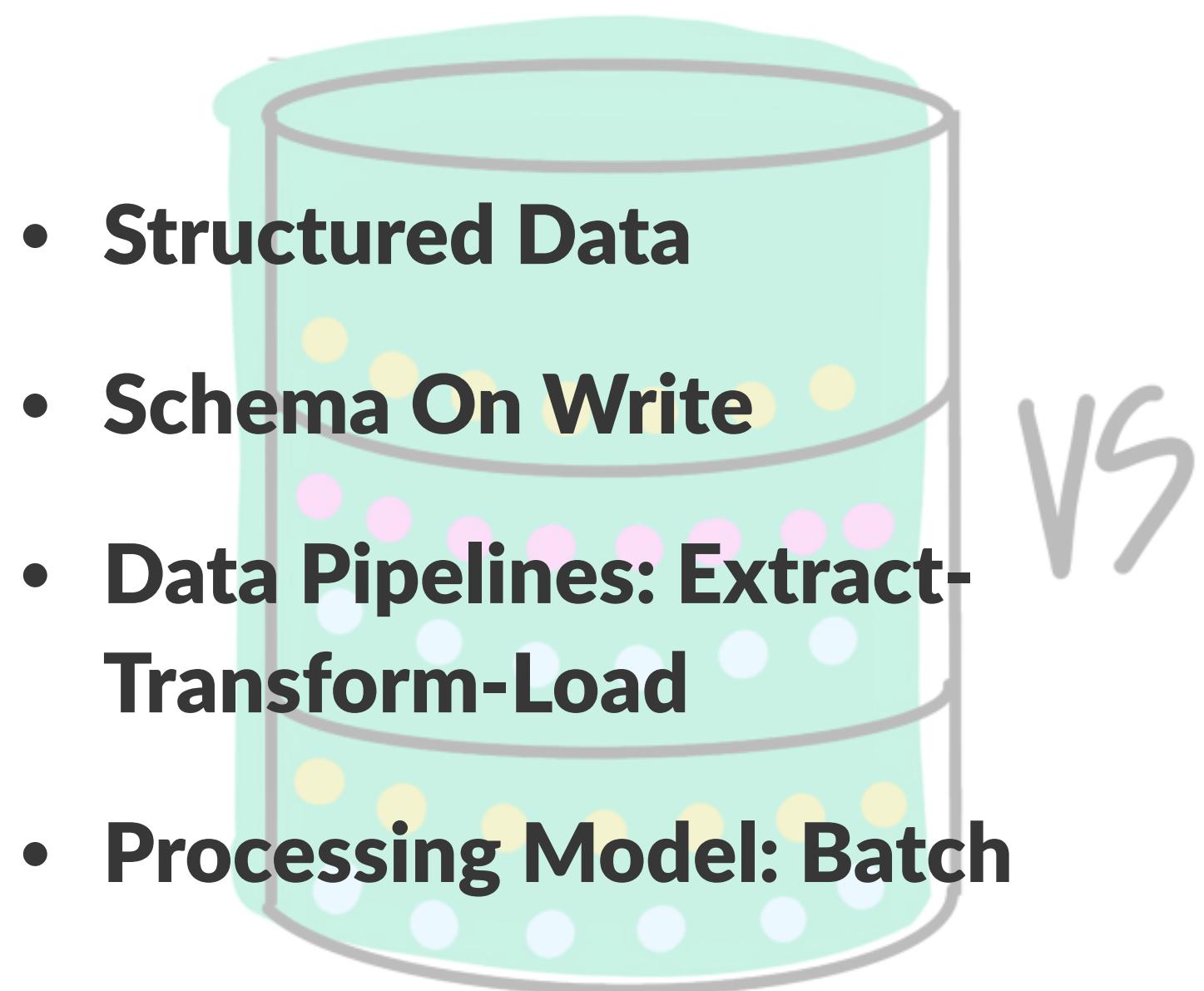
DATA LAKE



DATA WAREHOUSE

Data Lake vs Data Warehouse

DATA LAKE



Data Pipeline

A Data pipeline is a sum of tools and processes for performing data integration⁰³

Constructing data pipelines is the core responsibility of data engineering.

⁰³ [What is Data Engineering](#)

Data pipelines for moving data

Data pipelines for moving data

- data wrangling

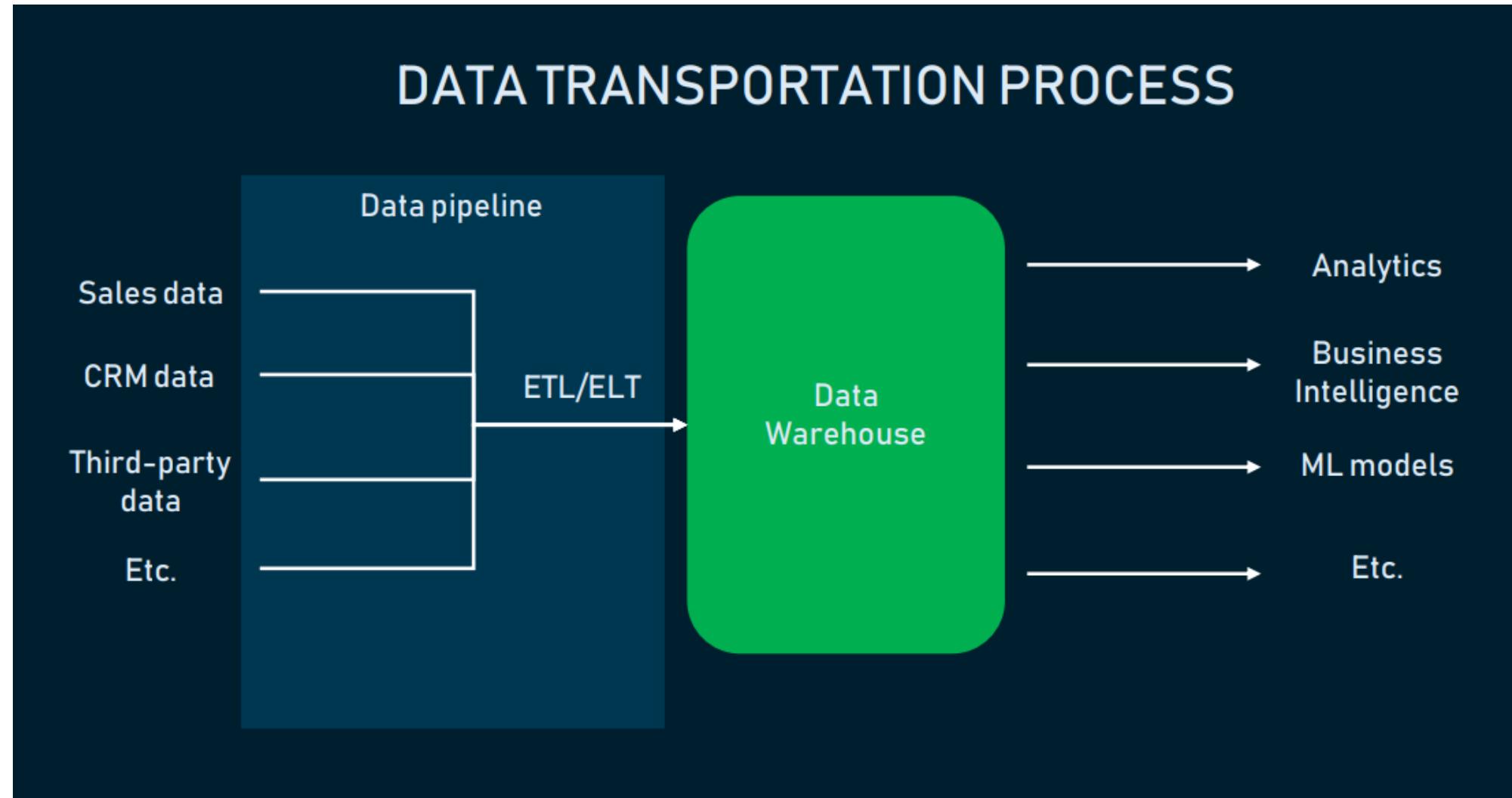
Data pipelines for moving data

- data wrangling
- data integration

Data pipelines for moving data

- data wrangling
- data integration
- data transformation

Transporting data from sources into a warehouse⁰¹⁰



⁰¹⁰ Source

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

⁰¹¹ we are focusing on ETL

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

- **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.

⁰¹¹ we are focusing on ETL

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

- **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.

⁰¹¹ we are focusing on ETL

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

- **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.

⁰¹¹ we are focusing on ETL

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

- **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.

⁰¹¹ we are focusing on ETL

Two Paradigms (and a half): SQL- v.s. JVM-Centric Pipelines⁰¹¹

- **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...

⁰¹¹ we are focusing on ETL

Skill Set: SQL mastery⁰³

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

⁰³ [What is Data Engineering](#)

Skill Set: SQL mastery⁰³

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

⁰³ [What is Data Engineering](#)

Skill Set: SQL mastery⁰³

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,

⁰³ [What is Data Engineering](#)

Skill Set: SQL mastery⁰³

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,

⁰³ [What is Data Engineering](#)

Skill Set: SQL mastery⁰³

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

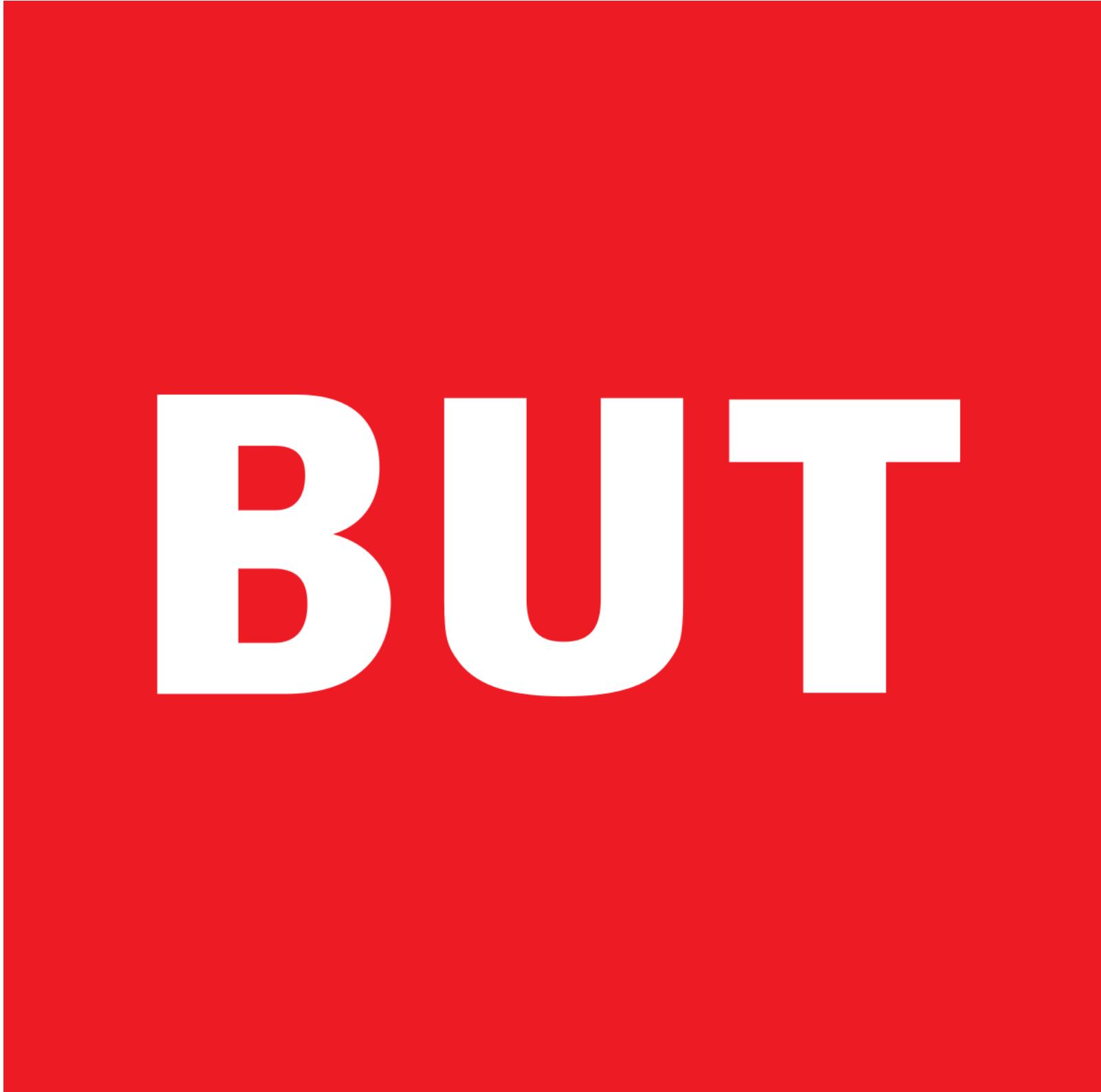
⁰³ [What is Data Engineering](#)

Skill Set: Data modeling⁰³

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.

⁰³ [What is Data Engineering](#)



BUT

Engineers Shouldn't (only) Write (SQL-based) ETL⁰¹²

⁰¹² [JeffMagnusson, 2016](#)

Engineers Shouldn't (only) Write (SQL-based) ETL⁰¹²

- 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.

⁰¹² [JeffMagnusson, 2016](#)

Engineers Shouldn't (only) Write (SQL-based) ETL⁰¹²

- 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.

⁰¹² [JeffMagnusson, 2016](#)

If Not (only) ETL, Then...What?⁰¹³

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.

⁰¹³ [TristanHandy, 2019](#)