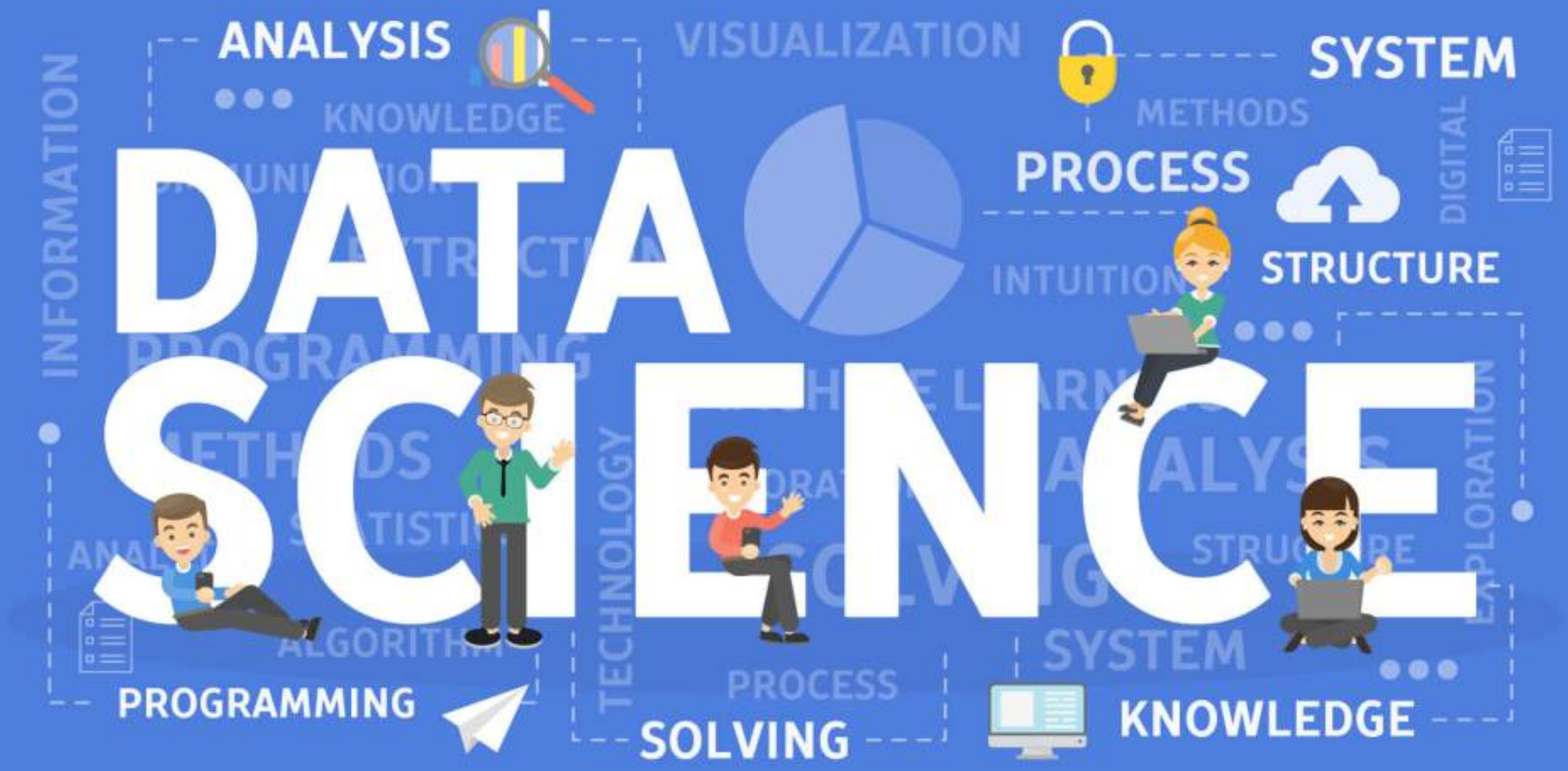
[IF-5-OT7:TD] Foundation of data engineering

MCF Riccardo Tommasini

http://rictomm.me

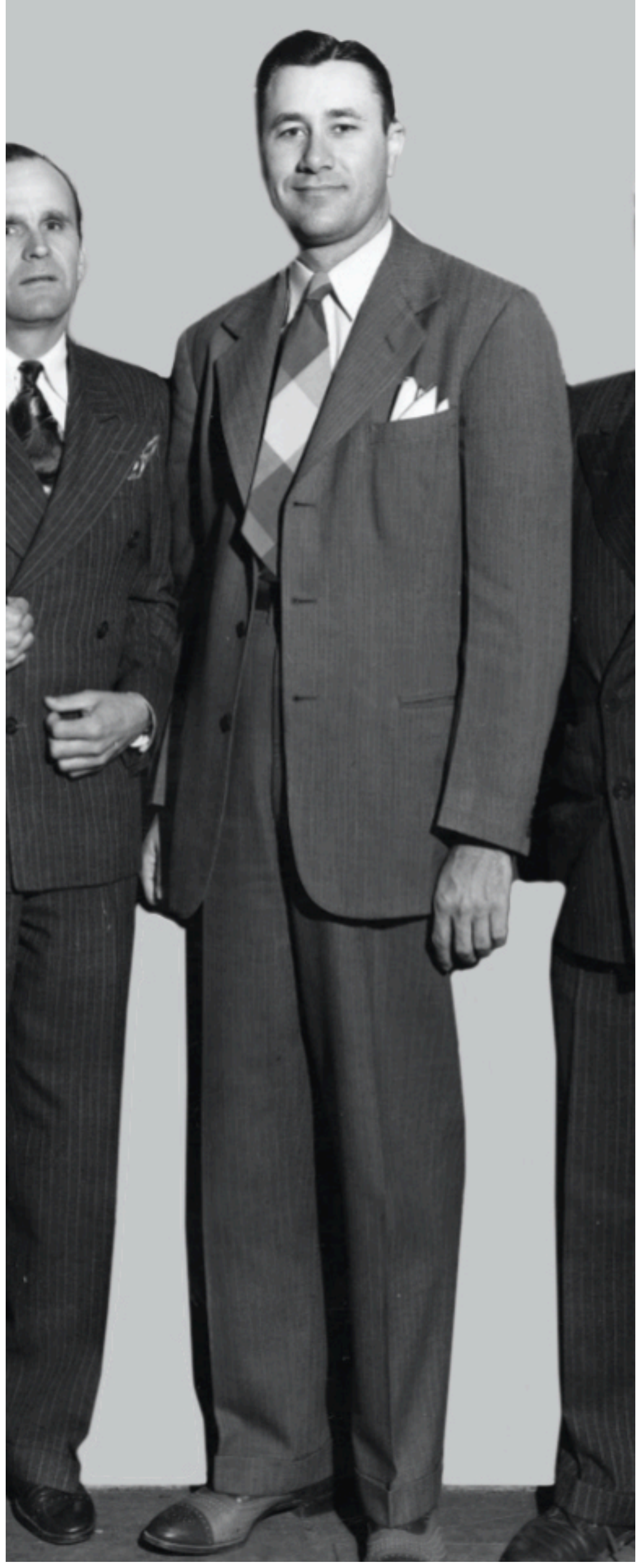
riccardo.tommasini@insa-lyon.fr


ANALYSIS
VISUALIZATION
SYSTEM
KNOWLEDGE
METHODS
INFORMATION
DIGITAL
DATA
PROCESS
STRUCTURE
INTUITION
PROGRAMMING
SCIENCE
METHODS
TECHNOLOGY
ANALYSIS
STATISTICS
STRUCTURE
EXPLORATION
ALGORITHM
SYSTEM
PROGRAMMING
PROCESS
SOLVING
KNOWLEDGE

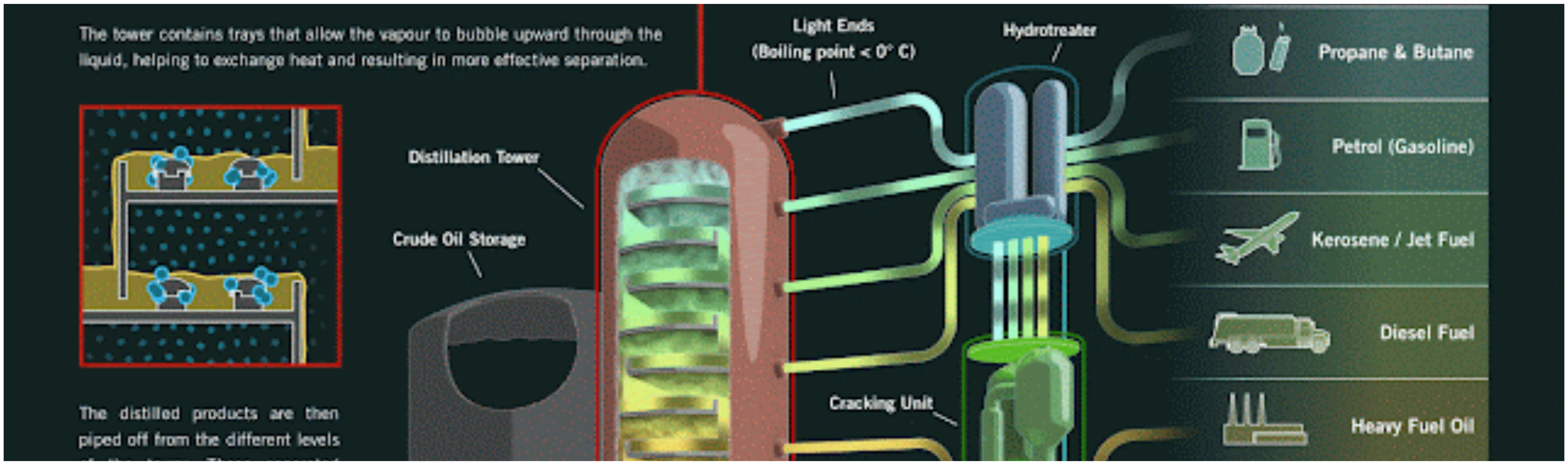
Quote

“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]


The tower contains trays that allow the vapour to bubble upward through the liquid, helping to exchange heat and resulting in more effective separation.
Distillation Tower
Crude Oil Storage
The distilled products are then piped off from the different levels
Light Ends (Boiling point < 0° C)
Hydrotreater
Cracking Unit
Propane & Butane
Petrol (Gasoline)
Kerosene / Jet Fuel
Diesel Fuel
Heavy Fuel Oil

...refining crude oil

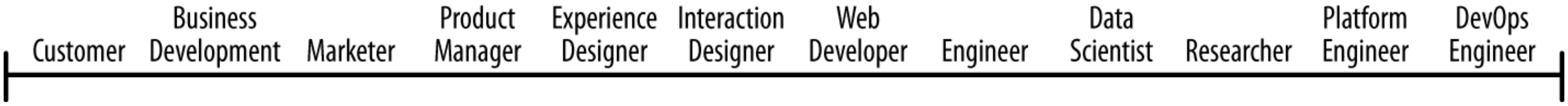
[01] Source

Discovering what we don’t know from data
- Obtaining predictive, actionable insight from data
- Creating Data Products that have business impact now
- Communicating relevant business stories from data
- Building confidence in decisions that drive business value

Data Engineering is...

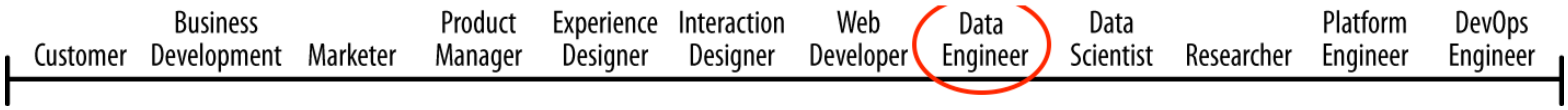
...build the refinery.

Roles in a Data Science Project[02]



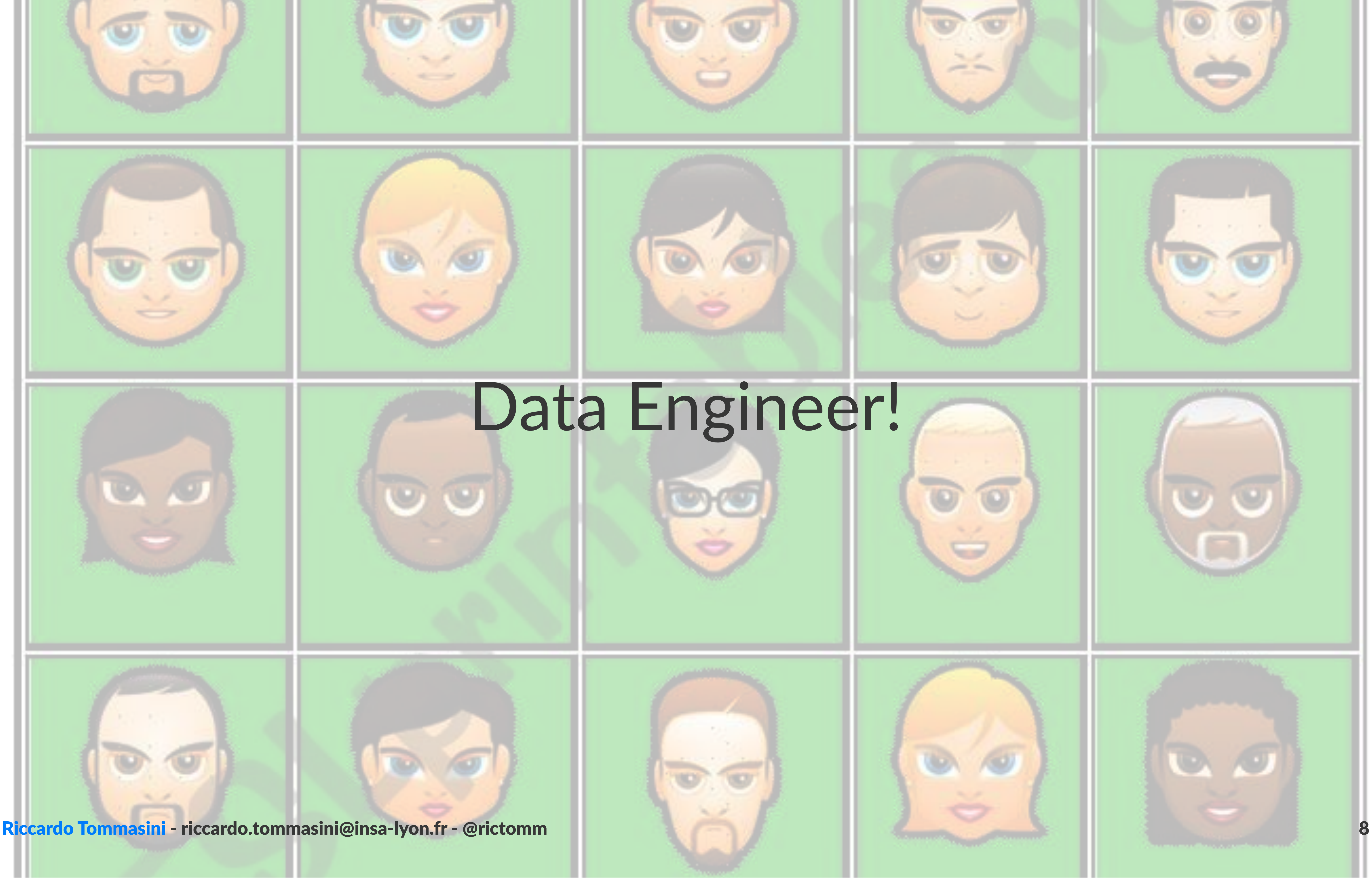
[02] http://emanueledellavalle.org/slides/dspm/ds4biz.html#25

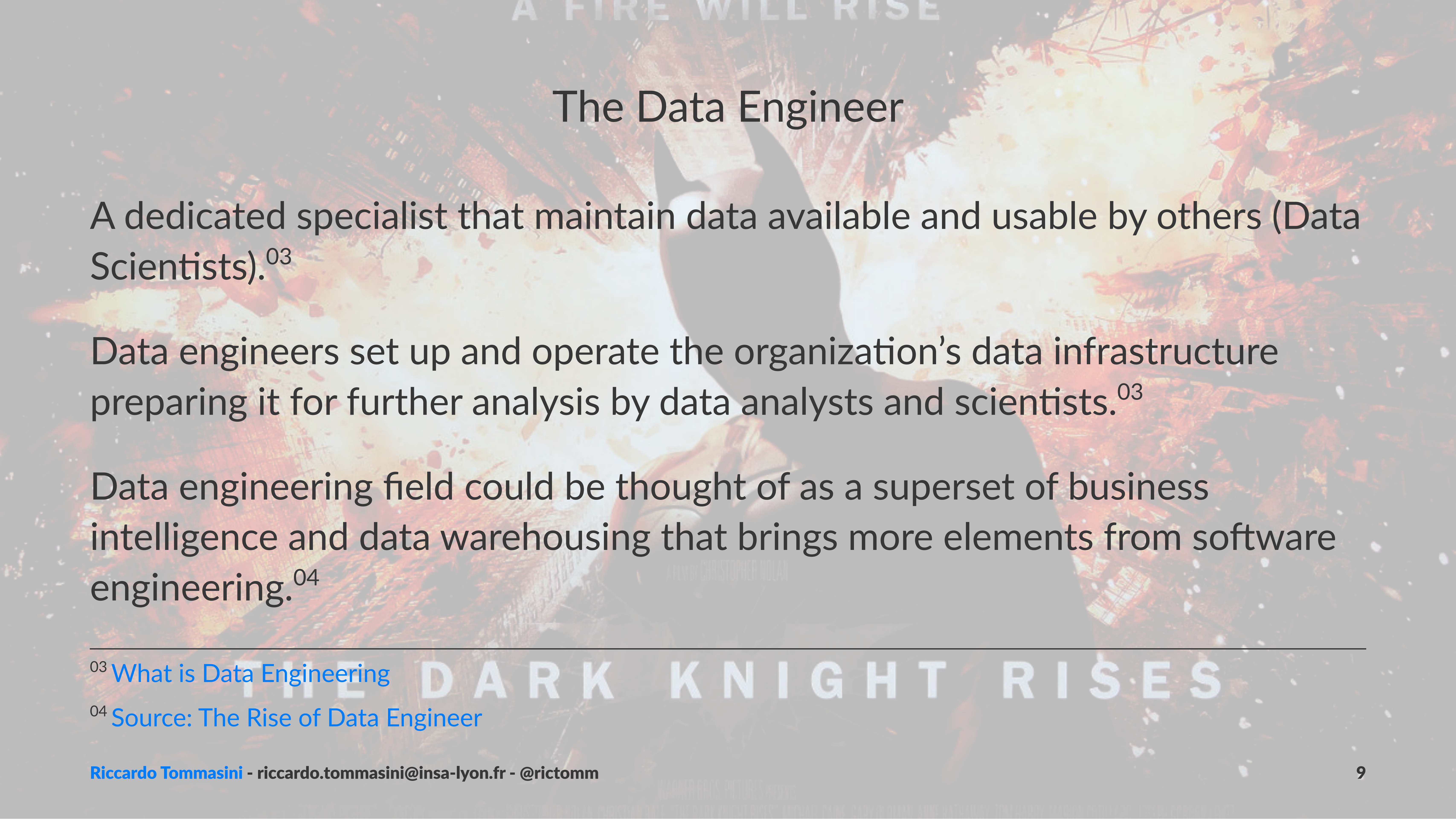
Roles in a Data Science Project[02]



[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]

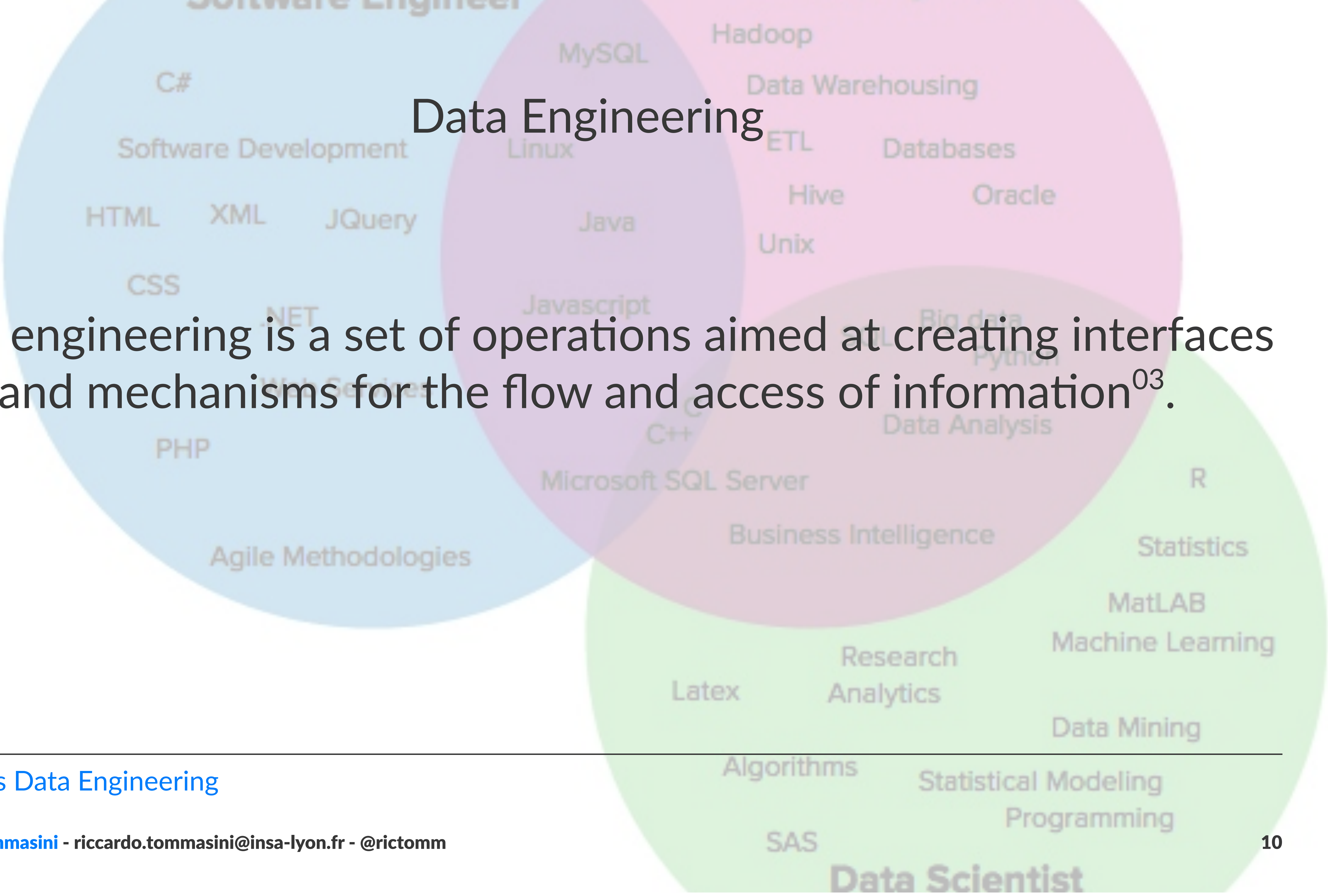
Data engineering field could be thought of as a superset of business intelligence and data warehousing that brings more elements from software engineering.[04]

[03] What is Data Engineering

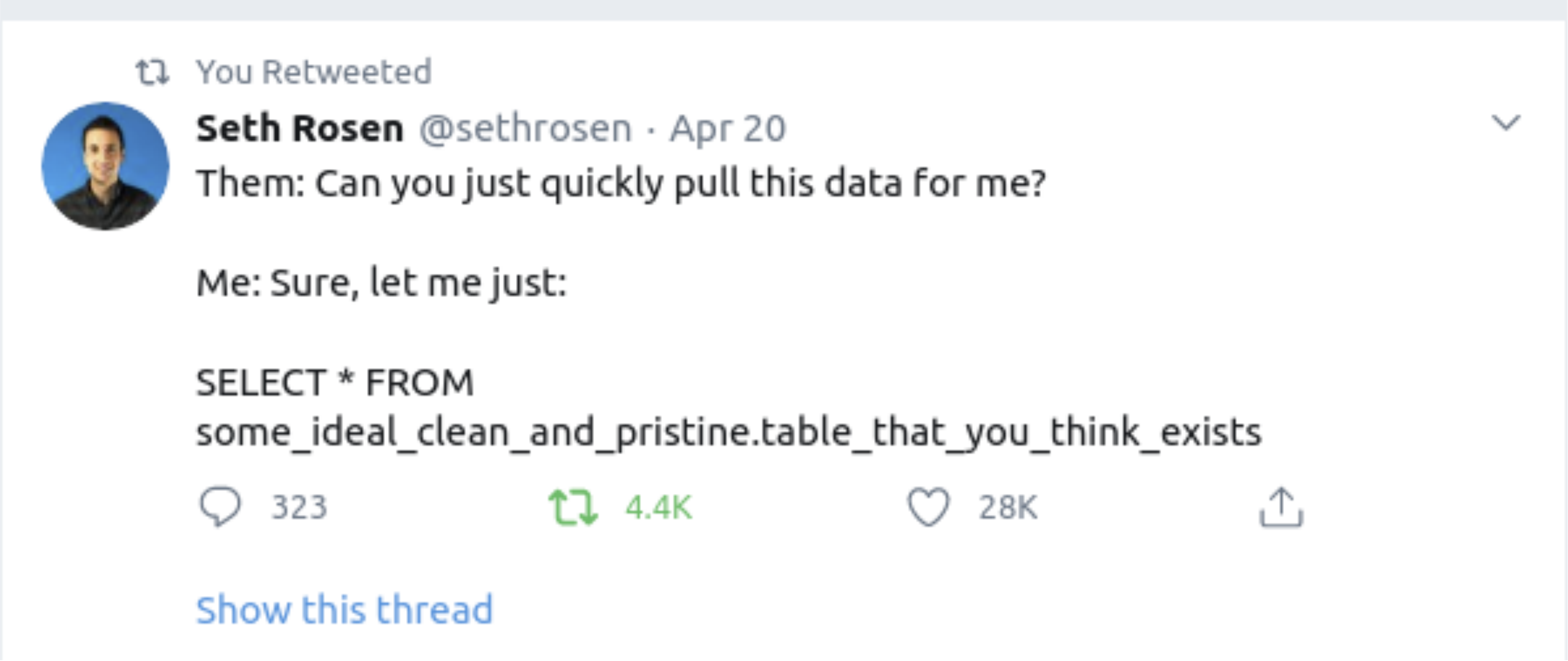
[04] Source: The Rise of Data Engineer

Data Engineering

Data engineering is a set of operations aimed at creating interfaces and mechanisms for the flow and access of information[03].



[03] What is Data Engineering



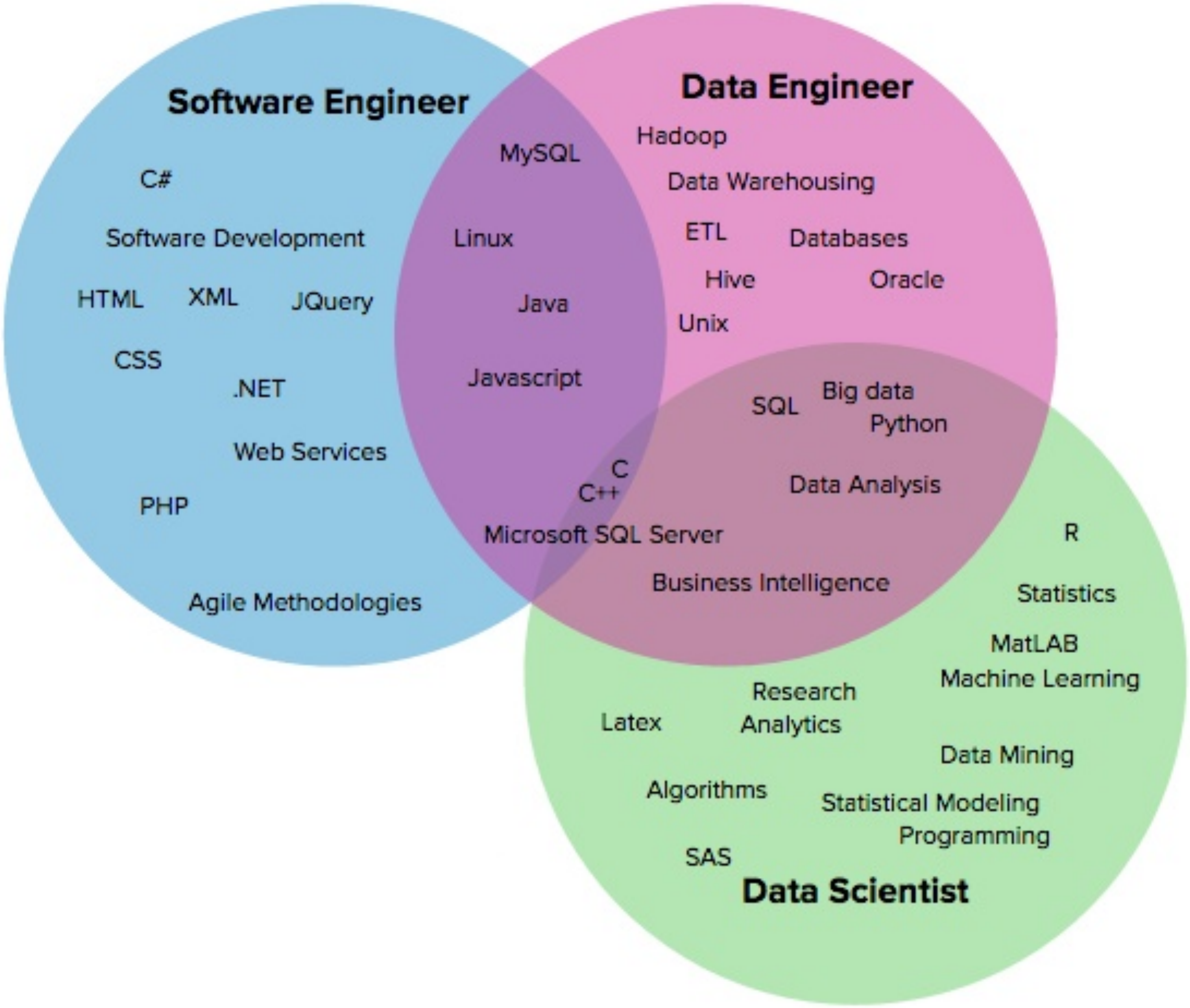
You Retweeted

Seth Rosen @sethrosen · Apr 20

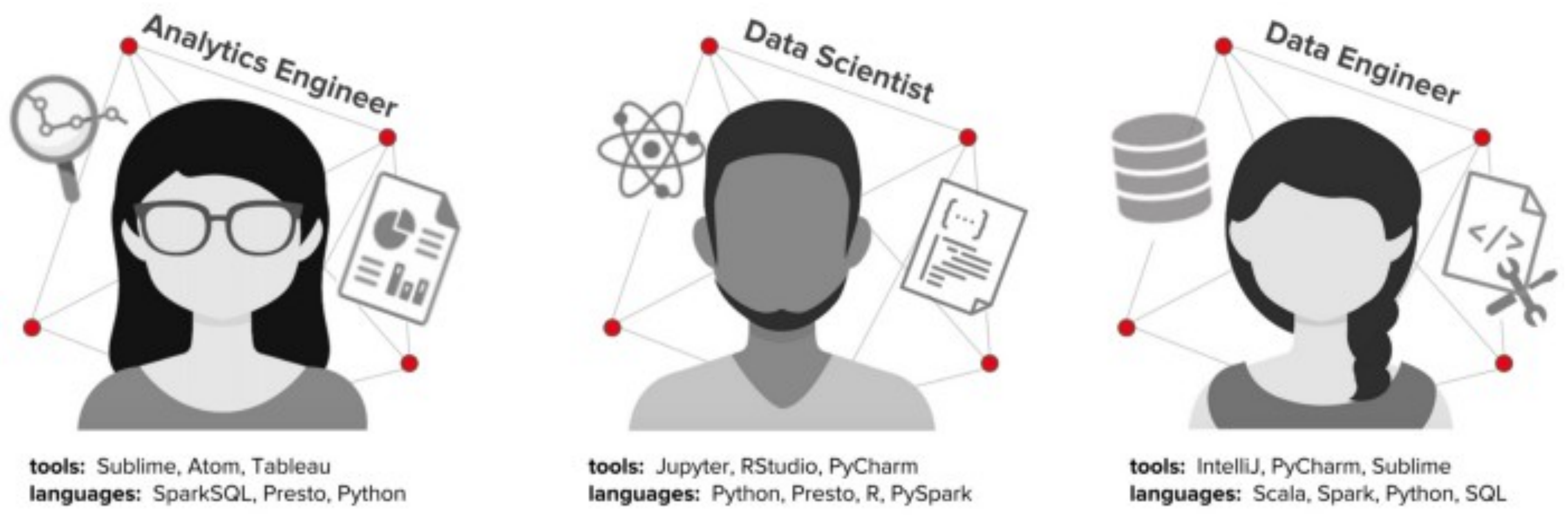
Them: Can you just quickly pull this data for me?

Me: Sure, let me just:

SELECT * FROM
some_ideal_clean_and_pristine.table_that_you_think_exists


Software Engineer
Data Engineer
Data Scientist
C#
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
MySQL
Linux
Java
Javascript
Hadoop
Data Warehousing
ETL
Databases
Hive
Oracle
Unix
SQL
Big data
Python
C
C++
Data Analysis
Microsoft SQL Server
Business Intelligence
R
Statistics
MatLAB
Machine Learning
Research
Analytics
Latex
Data Mining
Algorithms
Statistical Modeling
Programming
SAS

Netflix's Perspective[05]

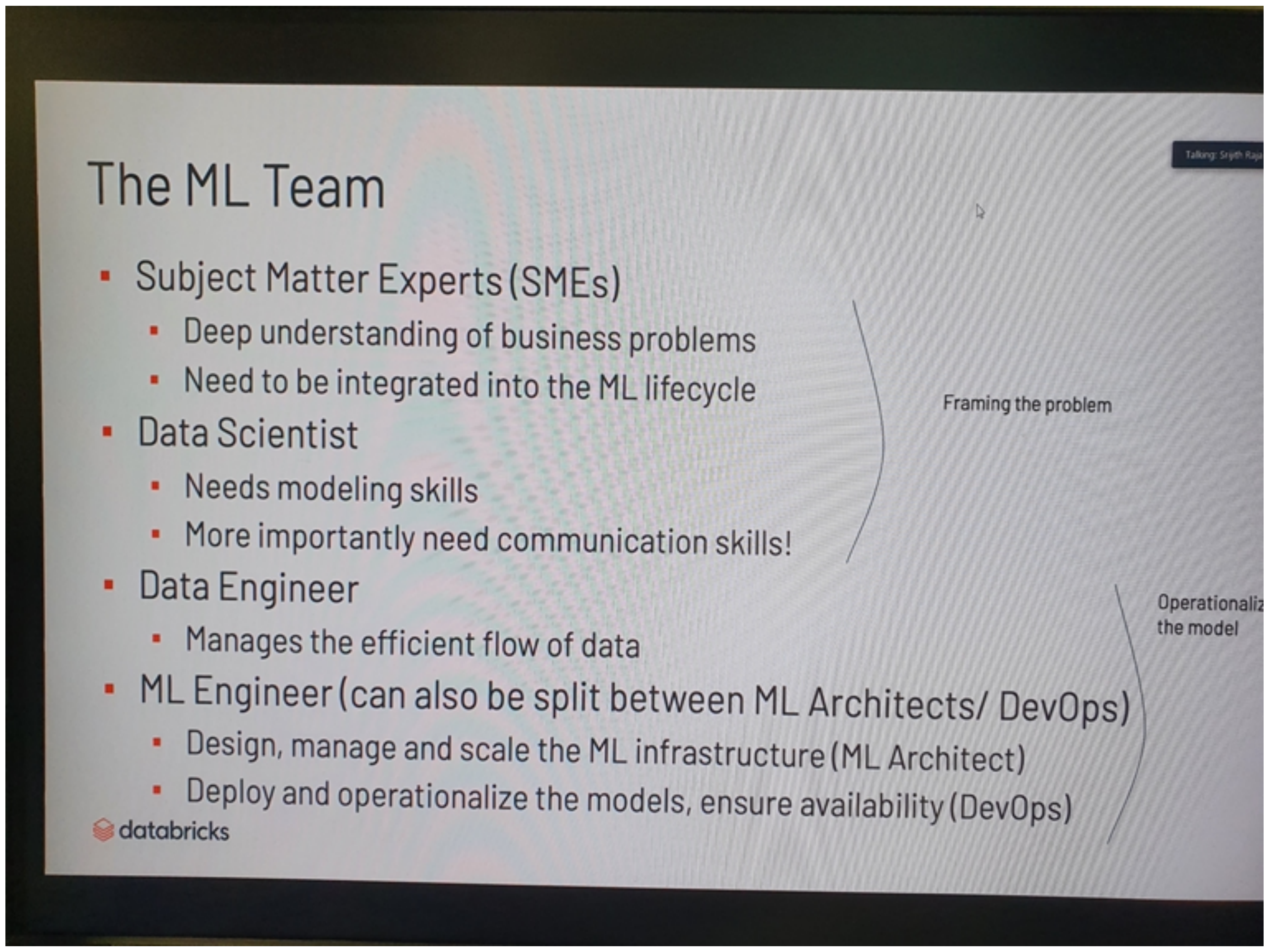

Analytics Engineer
tools: Sublime, Atom, Tableau
languages: SparkSQL, Presto, Python
Data Scientist
tools: Jupyter, RStudio, PyCharm
languages: Python, Presto, R, PySpark
Data Engineer
tools: IntelliJ, PyCharm, Sublime
languages: Scala, Spark, Python, SQL

- a data engineer might create a new aggregate of a dataset containing trillions of streaming events
- analytics engineer might use that aggregate in a new report on global streaming quality
- a data scientist might build a new streaming compression model reading the report

[05] Netflix Innovation

each of these workflows has multiple overlapping tasks:

Data Bricks


Talking: Srijith Raj

The ML Team

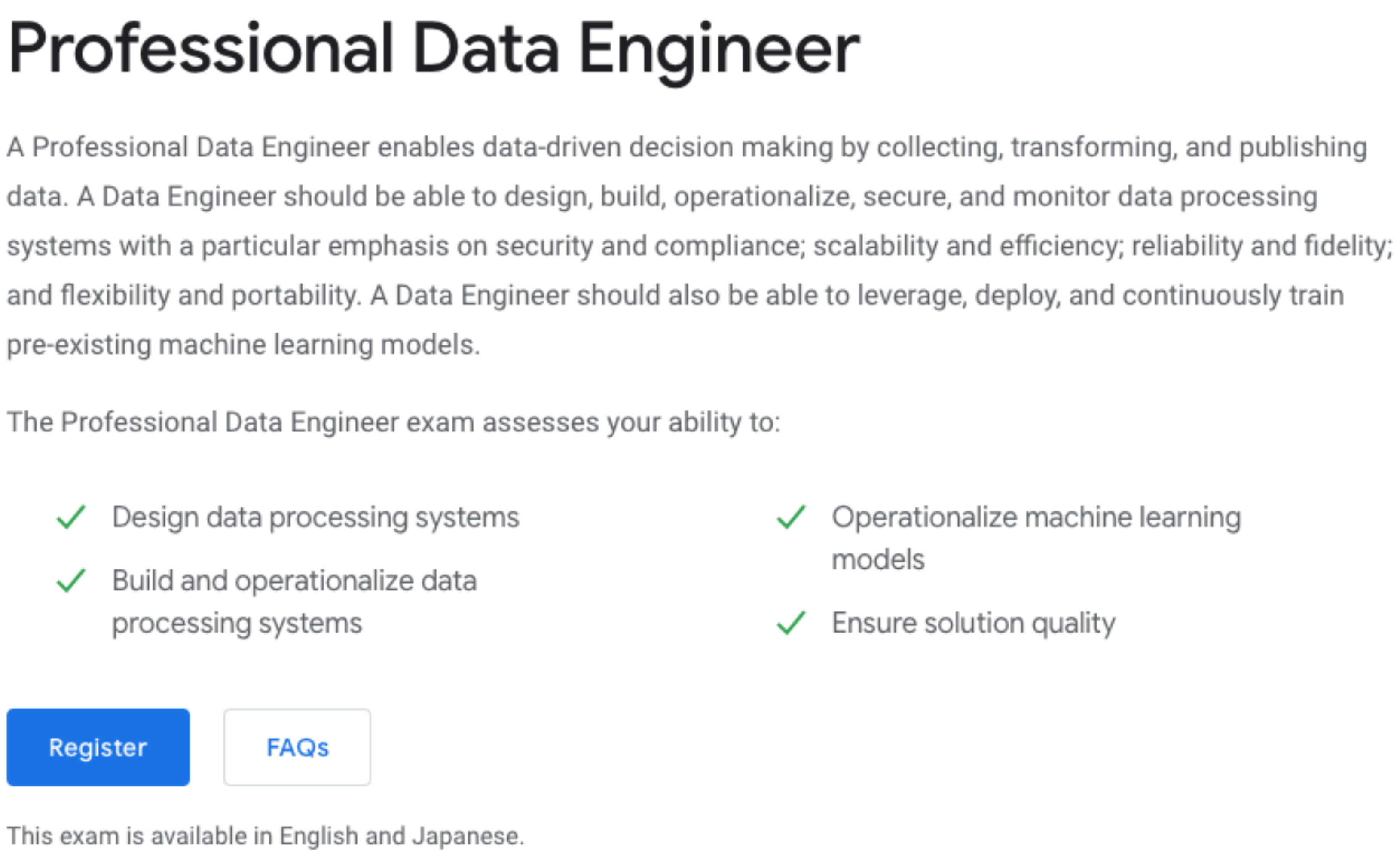
- Subject Matter Experts (SMEs)
  - Deep understanding of business problems
  - Need to be integrated into the ML lifecycle
- Data Scientist
  - Needs modeling skills
  - More importantly need communication skills!
- Data Engineer
  - Manages the efficient flow of data
- ML Engineer (can also be split between ML Architects/ DevOps)
  - Design, manage and scale the ML infrastructure (ML Architect)
  - Deploy and operationalize the models, ensure availability (DevOps)

Framing the problem

Operationaliz
the model

databricks

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

This exam is available in English and Japanese.

The Knowledge Scientist[06]

[06] The Manifesto

Philosophy of (Data) Science[07]

[07] Data as Fact

Nowdays we deal with a number of data from different domains.

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]

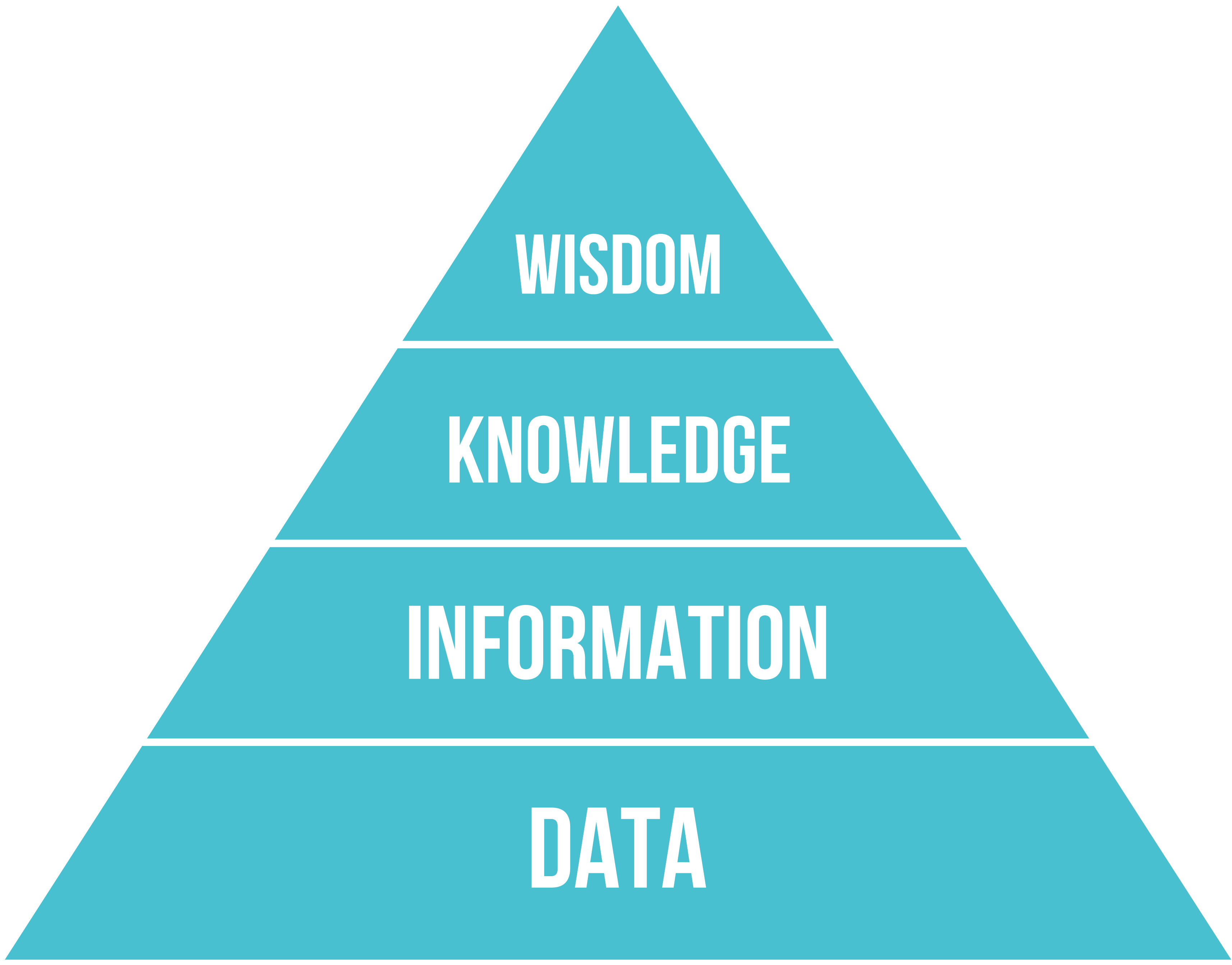
[08] Def

Wikipedia

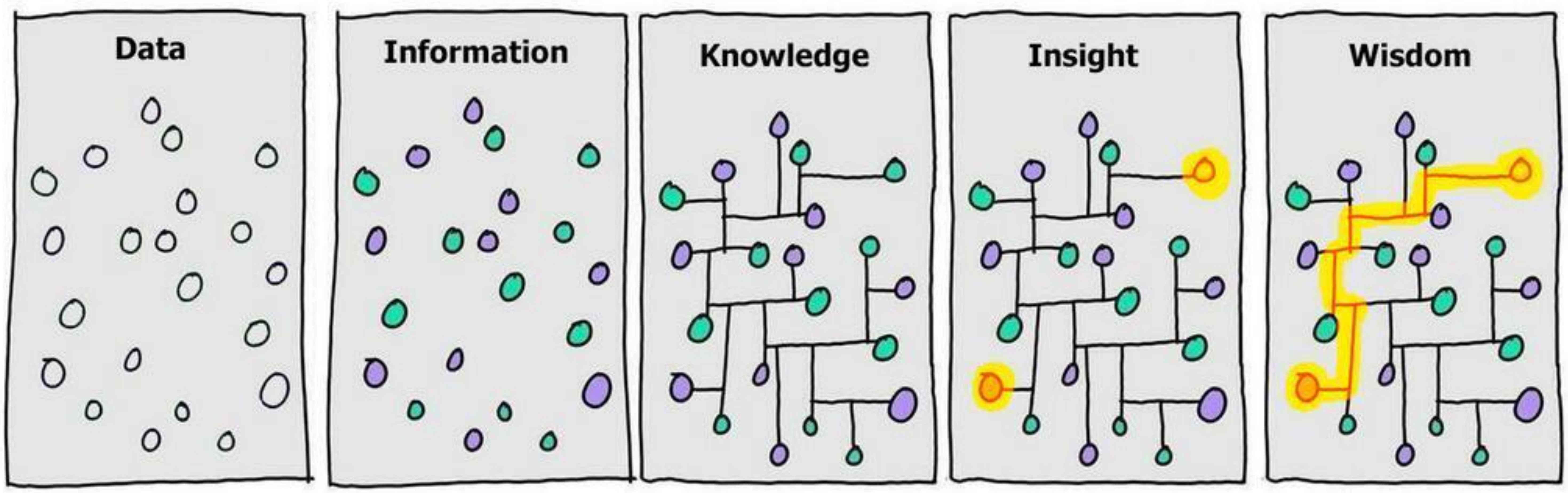
Data (treated as singular, plural, or as a mass noun) is any sequence of one or more symbols given meaning by specific act(s) of interpretation [09]

[09] Data in Computing)

DIKW Pyramid

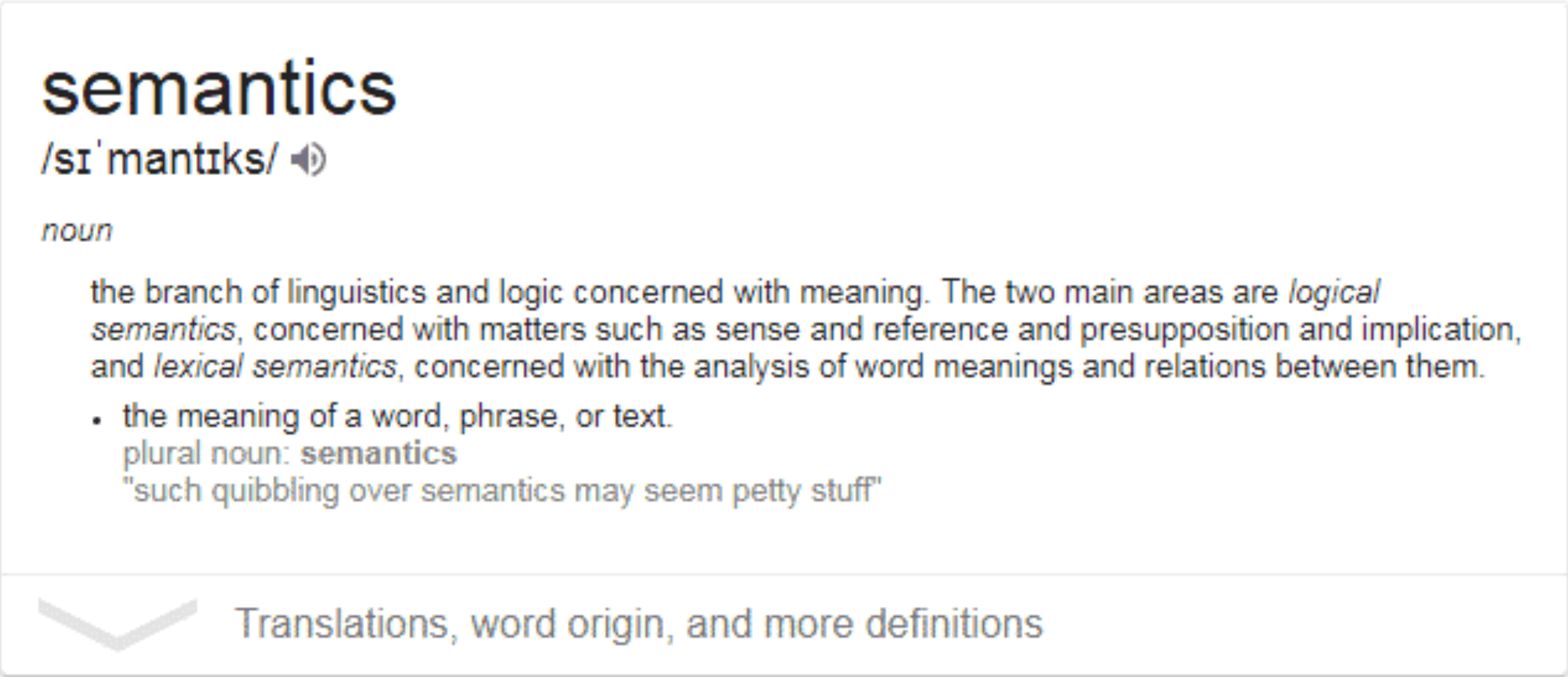


Graph View



Data about data

Data Semantics



semantics

/sɪˈmantɪks/

noun

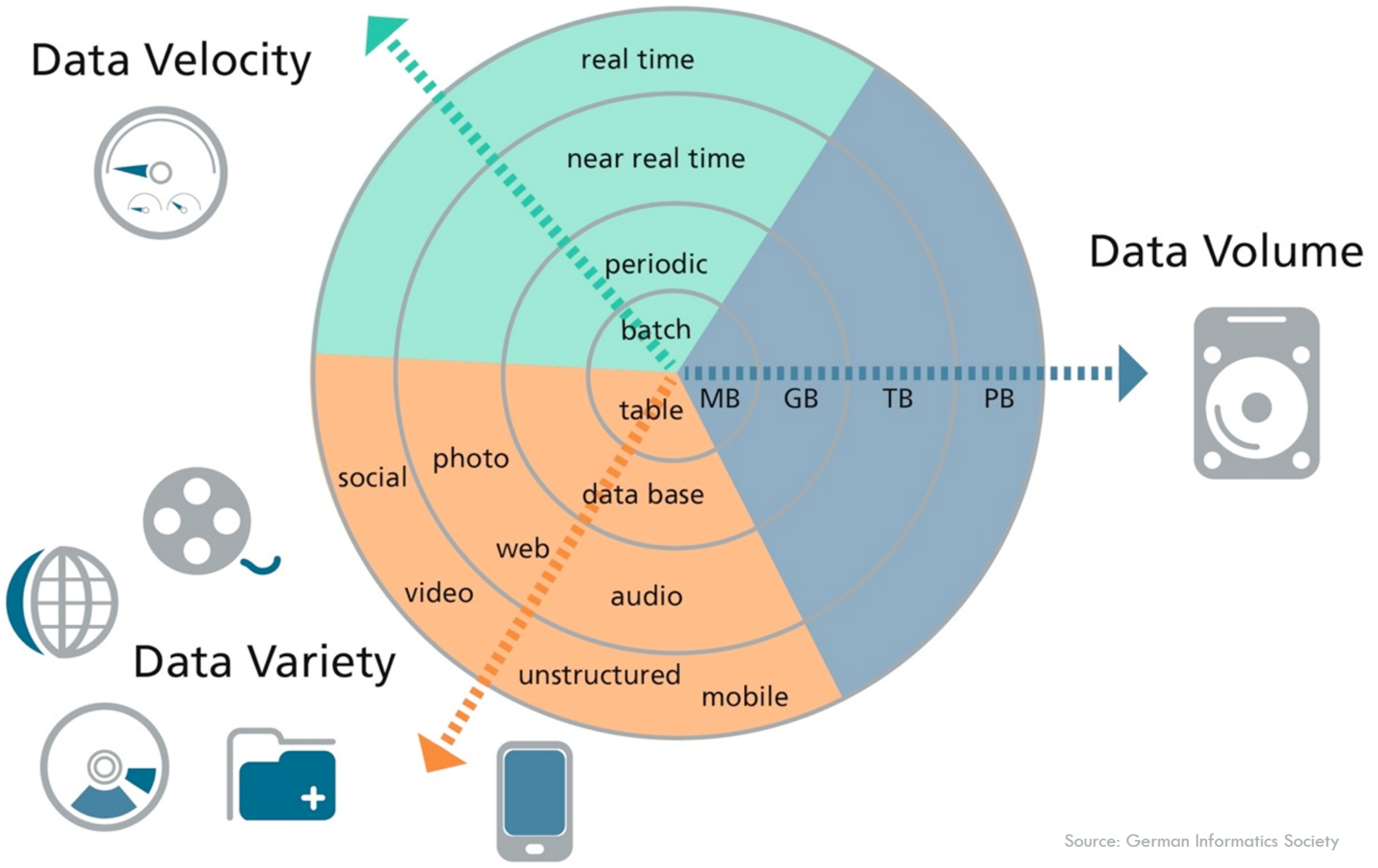
the branch of linguistics and logic concerned with meaning. The two main areas are *logical semantics*, concerned with matters such as sense and reference and presupposition and implication, and *lexical semantics*, concerned with the analysis of word meanings and relations between them.

- 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

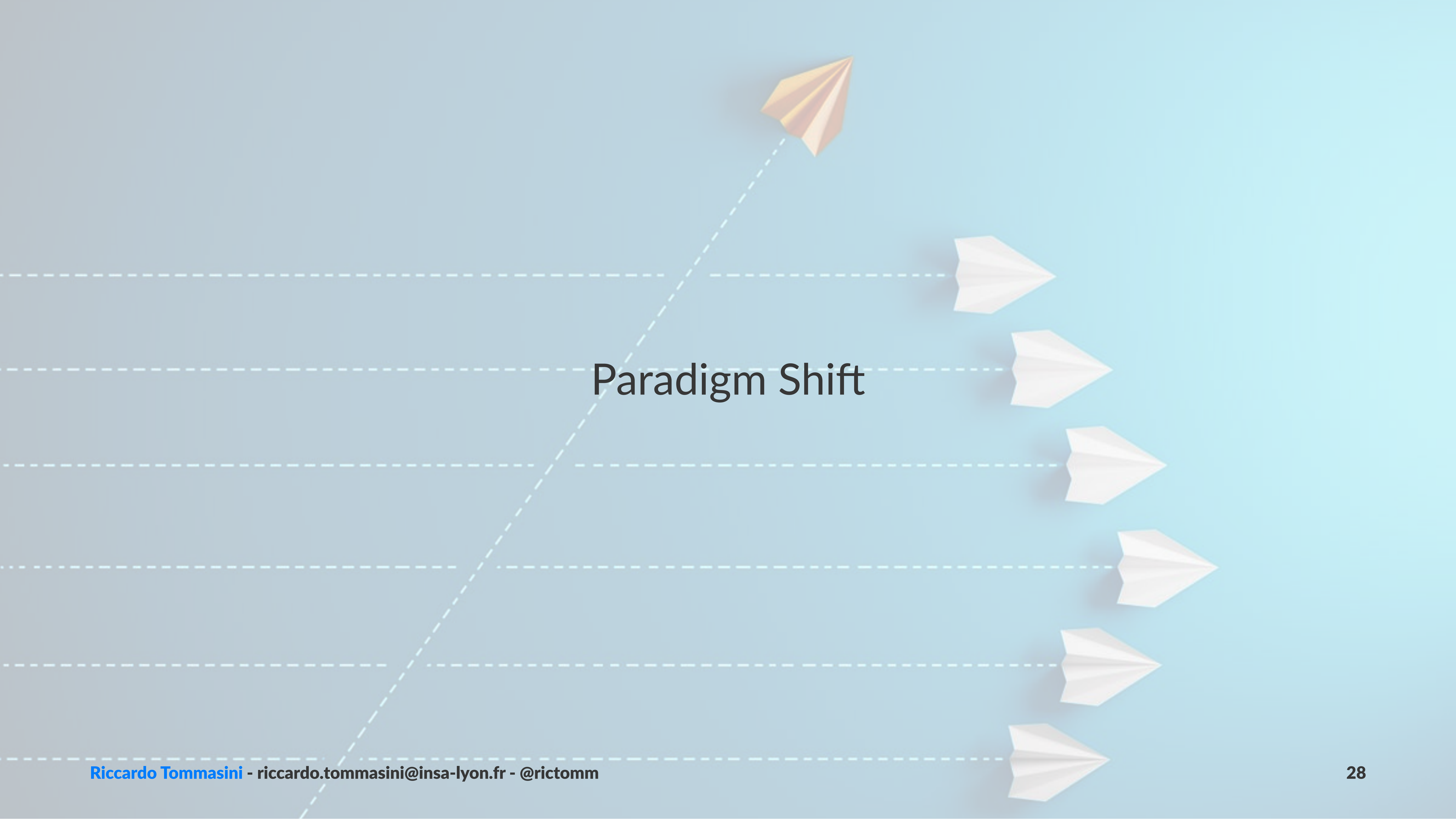
Big Data

Challenges [014]

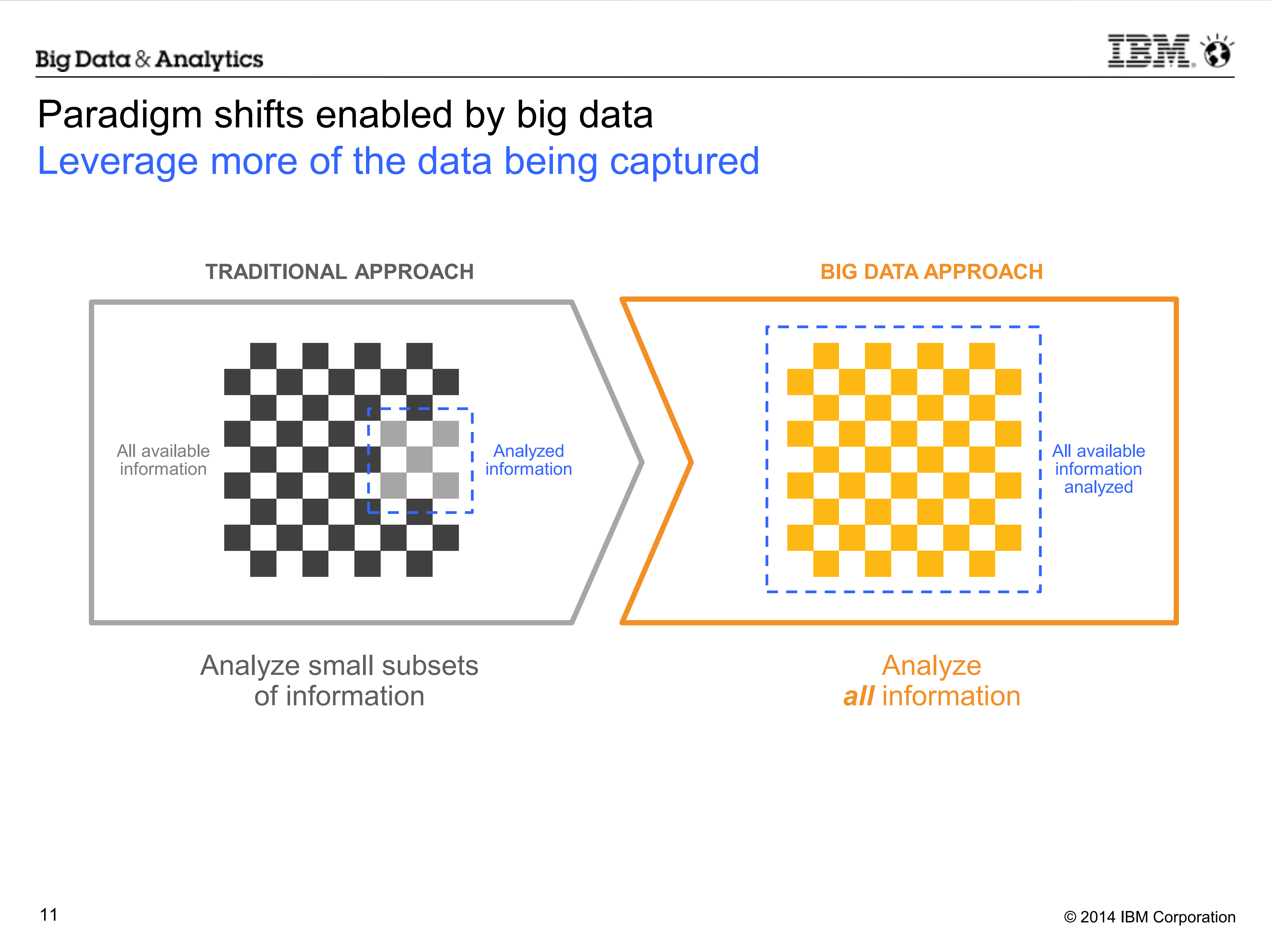


Source: German Informatics Society

[014] Lanely, 2001



Paradigm Shift



Big Data & Analytics

Paradigm shifts enabled by big data

Leverage more of the data being captured

TRADITIONAL APPROACH

All available information

Analyzed information

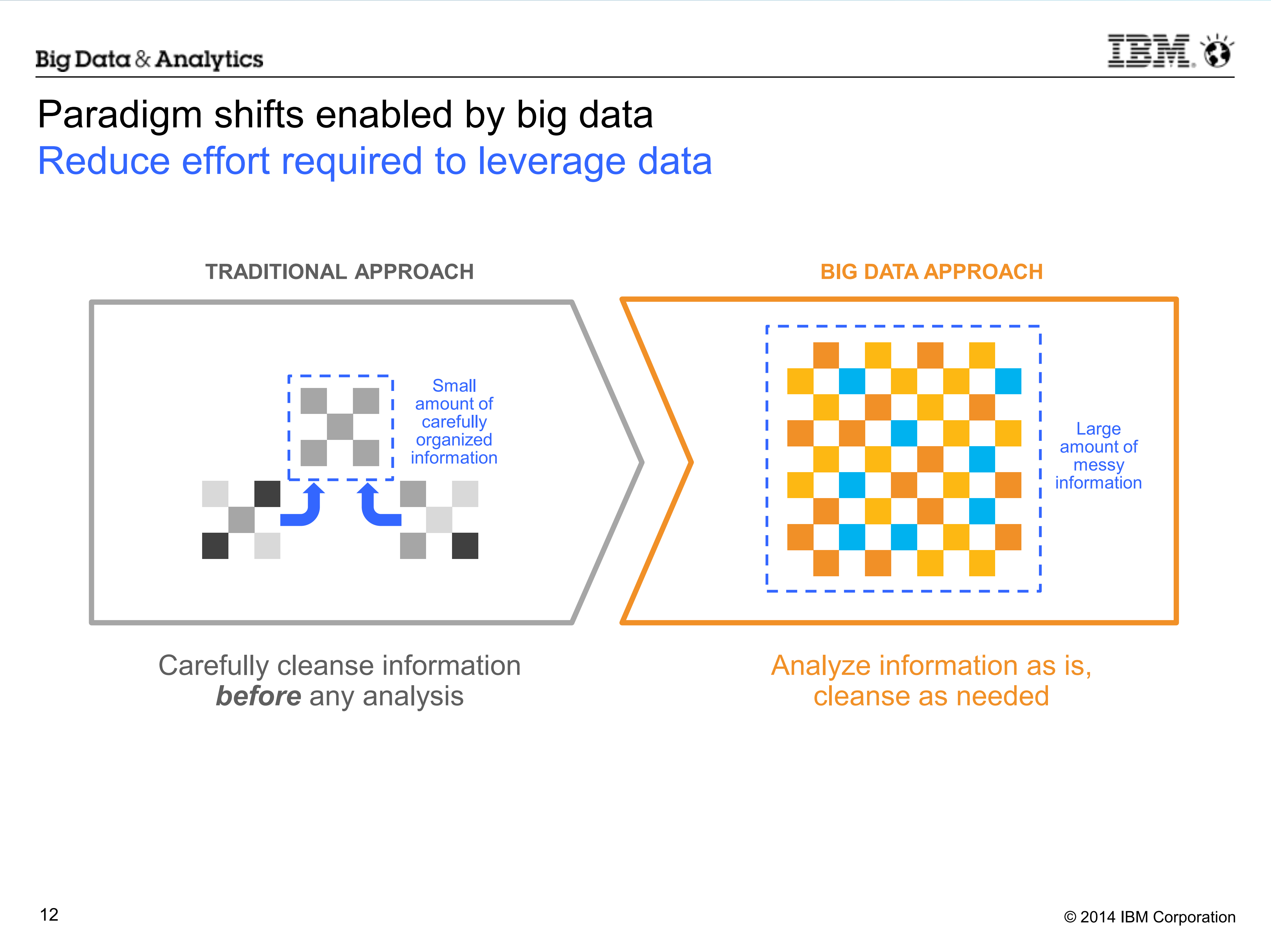
Analyze small subsets of information

BIG DATA APPROACH

All available information analyzed

Analyze *all* information

11



Big Data & Analytics

Paradigm shifts enabled by big data

Reduce effort required to leverage data

TRADITIONAL APPROACH

Small amount of carefully organized information

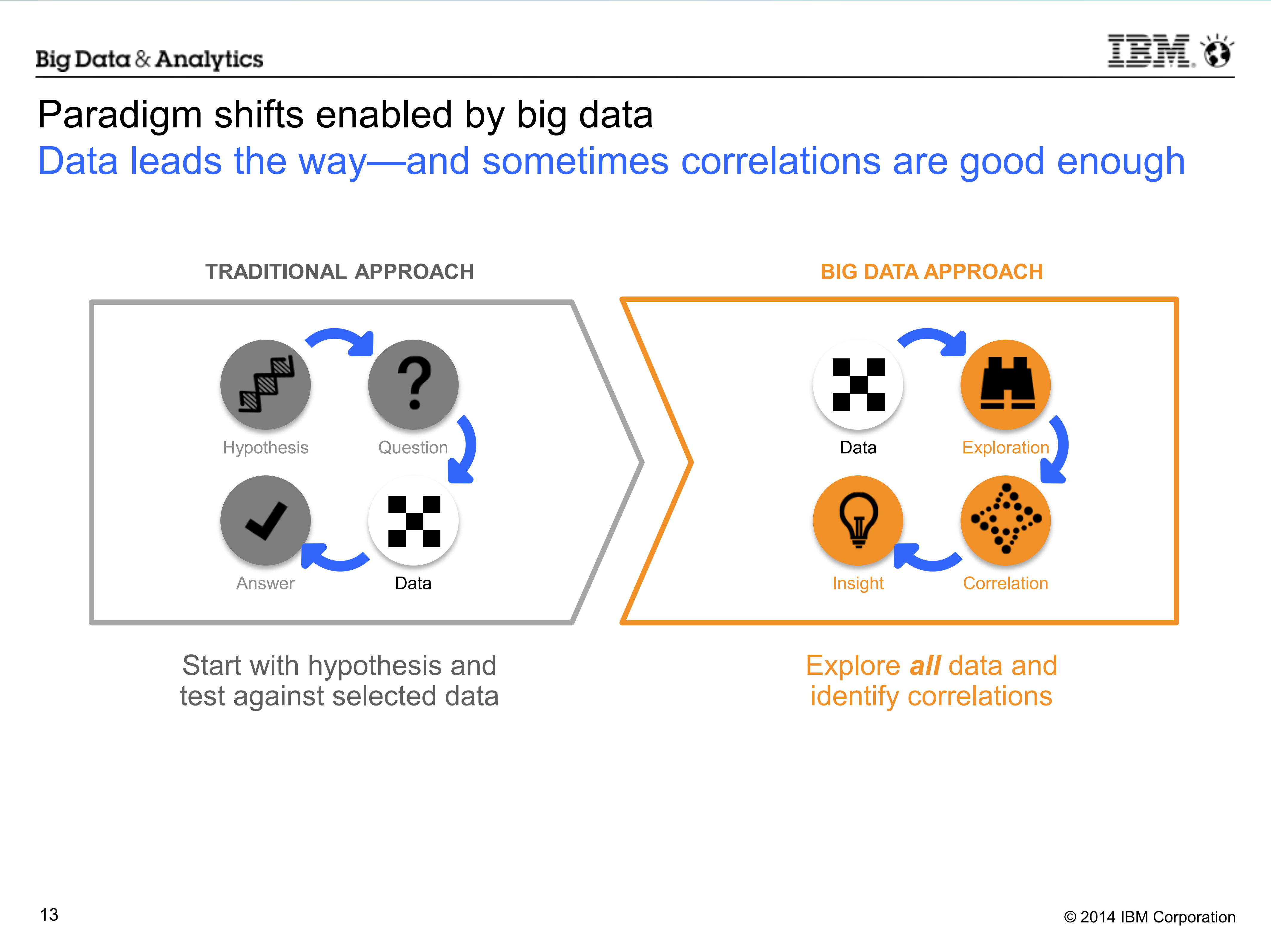
Carefully cleanse information *before* any analysis

BIG DATA APPROACH

Large amount of messy information

Analyze information as is, cleanse as needed

12



Big Data & Analytics

IBM

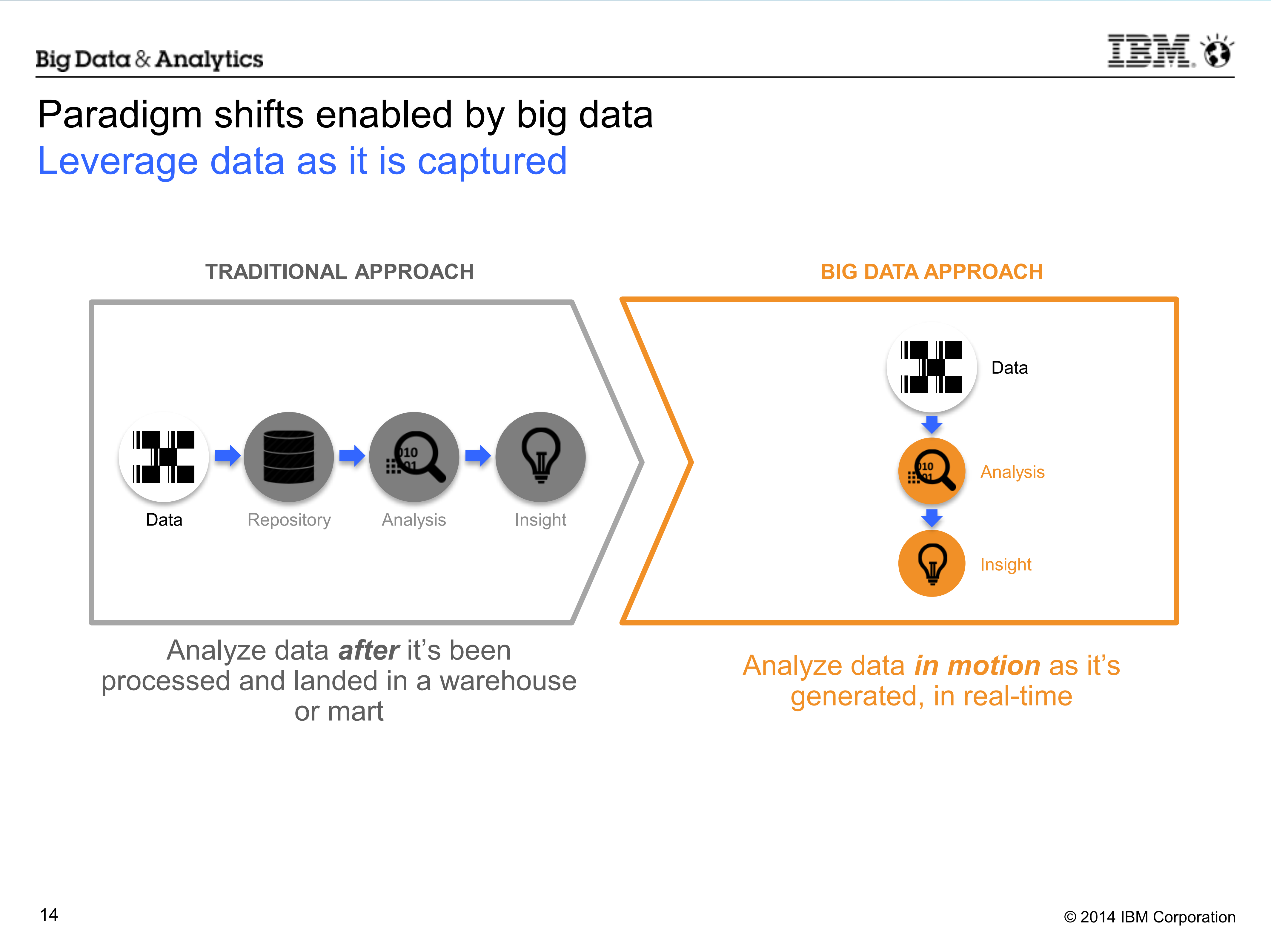
Paradigm shifts enabled by big data

Data leads the way—and sometimes correlations are good enough

Start with hypothesis and test against selected data

Explore ***all*** data and identify correlations

13



Big Data & Analytics

Paradigm shifts enabled by big data

Leverage data as it is captured

TRADITIONAL APPROACH

Data → Repository → Analysis → Insight

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

BIG DATA APPROACH

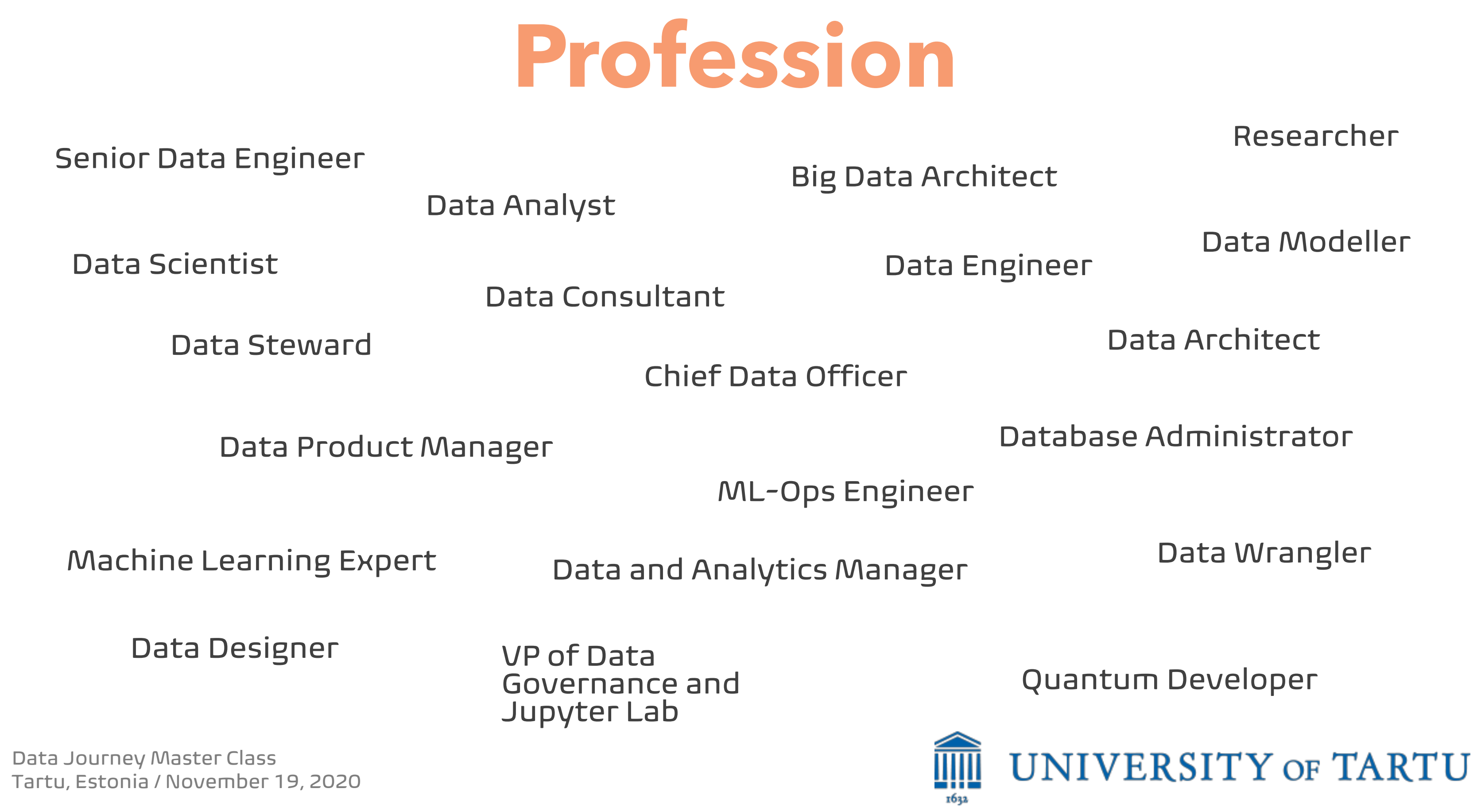
Data → Analysis → Insight

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

14

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

Senior Data Engineer

Researcher

Big Data Architect

Data Analyst

Data Scientist

Data Modeller

Data Engineer

Data Consultant

Data Steward

Data Architect

Chief Data Officer

Data Product Manager

Database Administrator

ML-Ops Engineer

Machine Learning Expert

Data Wrangler

Data and Analytics Manager

Data Designer

VP of Data
Governance and
Jupyter Lab

Quantum Developer

Data Journey Master Class
Tartu, Estonia / November 19, 2020

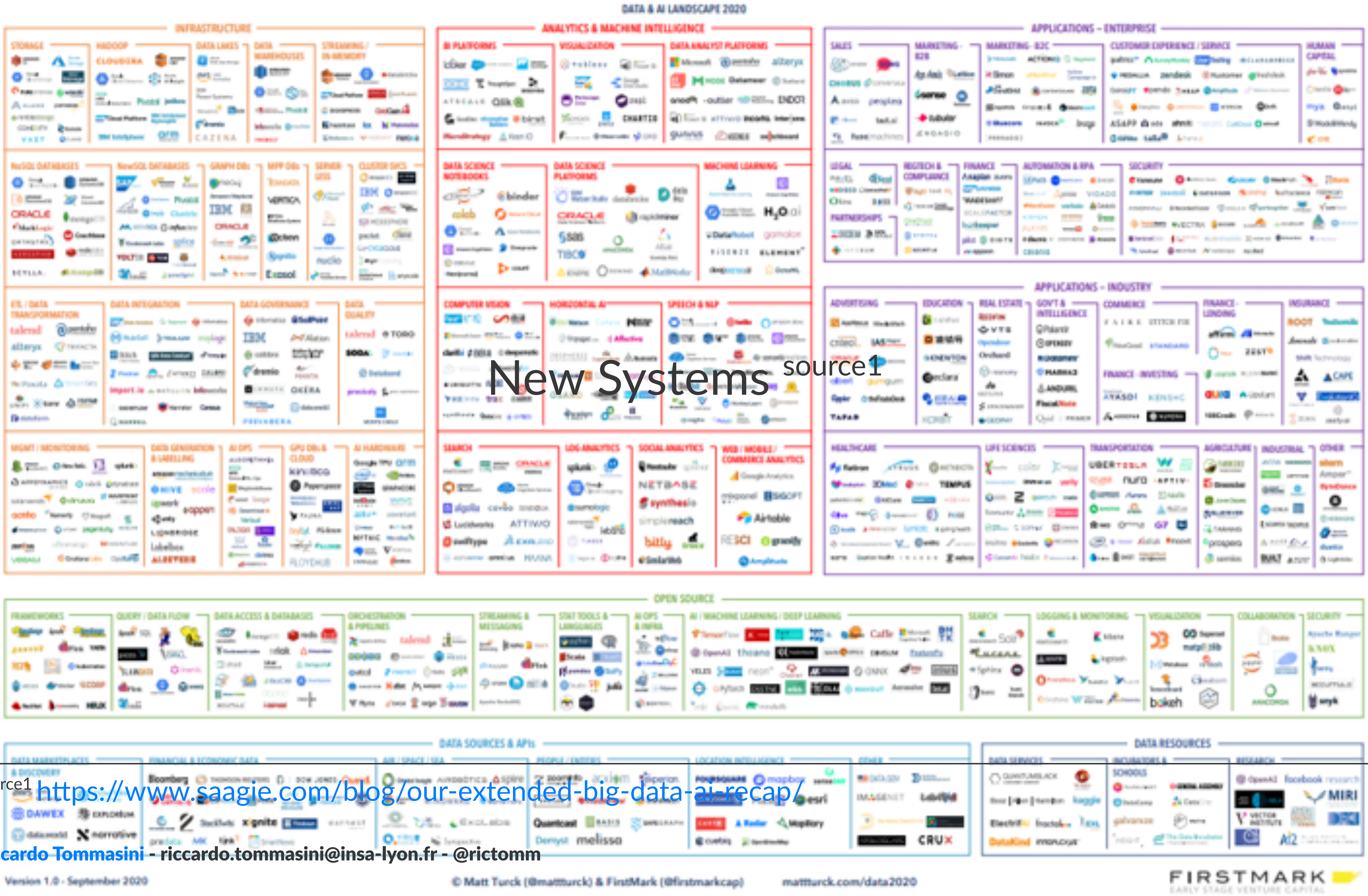
UNIVERSITY OF TARTU

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.

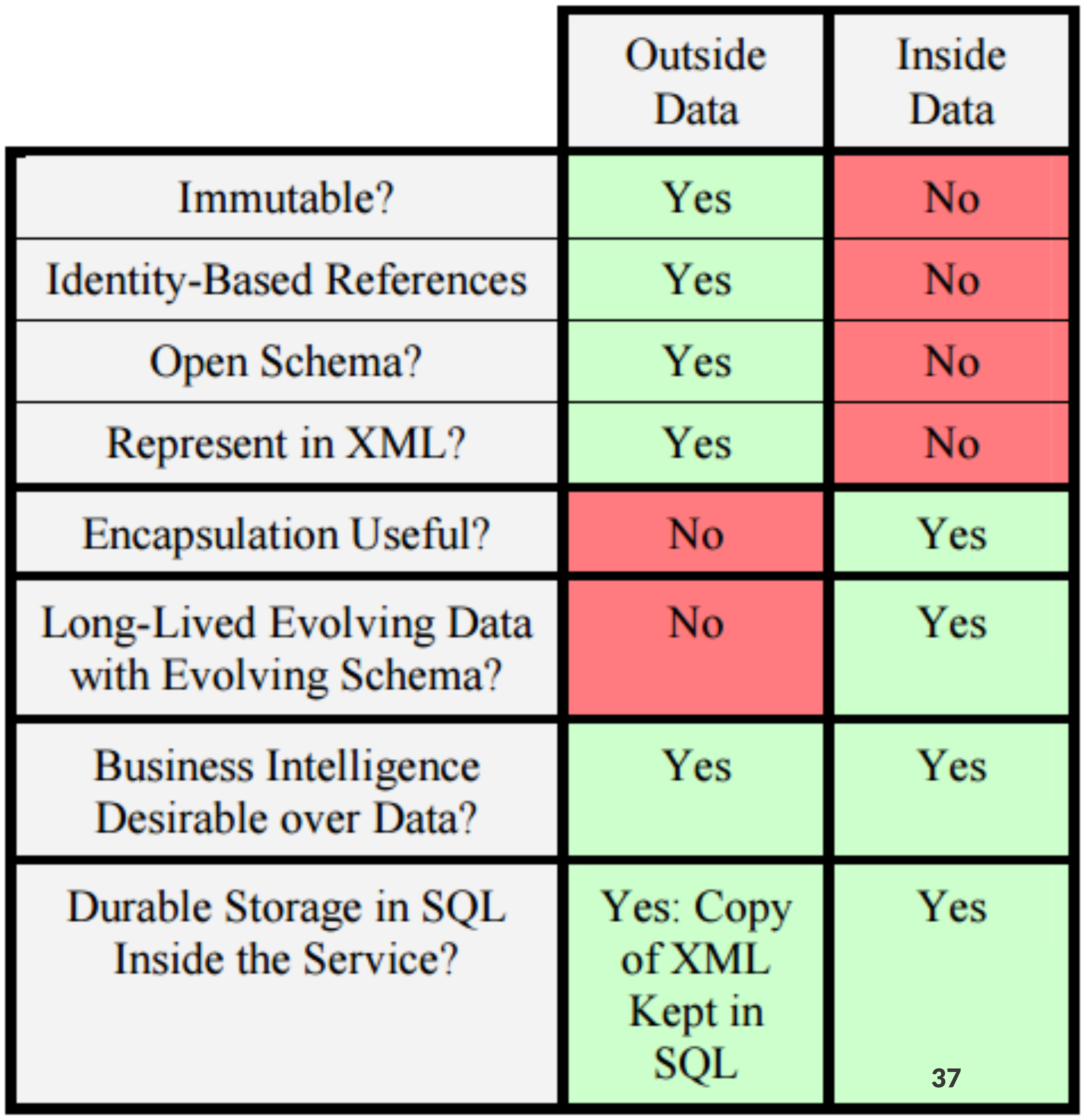
New Systems source1


DATA & AI LANDSCAPE 2020
INFRASTRUCTURE
ANALYTICS & MACHINE INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
Version 1.0 - September 2020
© Matt Turck (@mattturck) & FirstMark (@firstmarkcap)
mattturck.com/data2020
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

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

Where is Data?

Data on the Inside vs Data on the Outside [pt]



	Outside Data	Inside Data
Immutable?	Yes	No
Identity-Based References	Yes	No
Open Schema?	Yes	No
Represent in XML?	Yes	No
Encapsulation Useful?	No	Yes
Long-Lived Evolving Data with Evolving Schema?	No	Yes
Business Intelligence Desirable over Data?	Yes	Yes
Durable Storage in SQL Inside the Service?	Yes: Copy of XML Kept in SQL	Yes

[pt] Data on the Outside vs Data on the Inside *Pat Helland, CIDR 2005*

The Outside (WEB)

International ecosystem of applications and services that allows us to search, aggregate, combine, transform, replicate, cache, and archive the information that underpins society.

The Web is the result of millions of simple, small-scale interactions between agents and resources that use the founding technologies of HTTP and URI. [121]

The Web is a set of widely commoditised servers, proxies, caches, and content delivery networks [an engineers]

[121] Architecture of the World Wide Web, Volume One

Resources

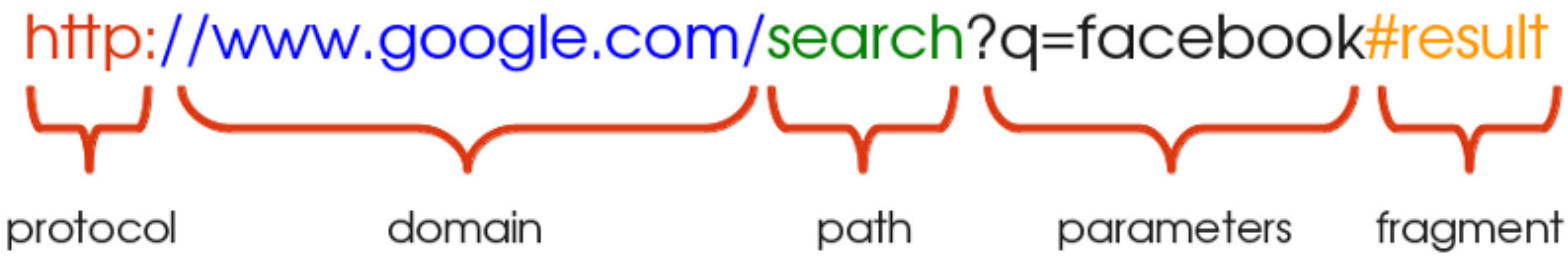
Resources are the fundamental building blocks

Anything we can expose, i.e., documents, images, videos, audio, devices, people, things...

We can represent them by abstracting the useful information and identifying using a Uniform Resource Identifier (URI)

URIs

URL format (RFC 2396):
scheme:[//[user:password@]host[:port]][/]path[?query][#fragment]



E.G.:
git@github.com:nodejs/node.git
mongodb://root:pass@localhost:27017/TestDB?options?replicaSet=test
http://example.com

Representation

Access to a resource is mediated by a representation

This separation is convenient to promote loose-coupling between server (producers) and client (consumers)

Multiple Views and Content Negotiation are the basis for interoperability

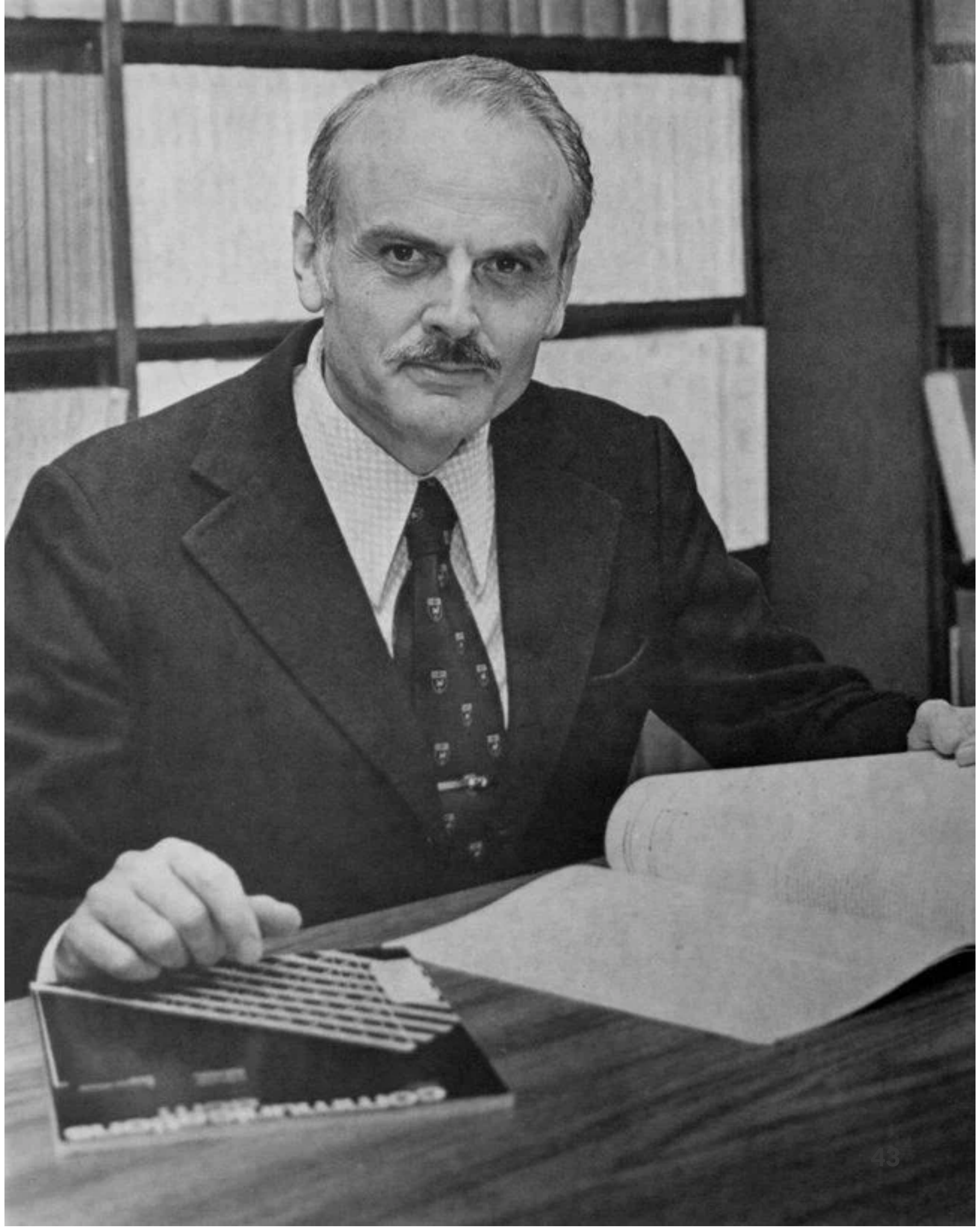
Protocols: HTTP

- GET
  - Uniform Interface
  - read-only operation
  - idempotent
- POST
  - like a resource upload
  - idempotent
- DELETE
  - remove resources
  - idempotent
- HEAD
  - HEAD is like GET except it returns only a response code
- PUT
  the only non-idempotent and unsafe operation
  is allowed to modify the service in a unique way
- OPTIONS is used to request information about the communication options of the resource

Databases Management Systems

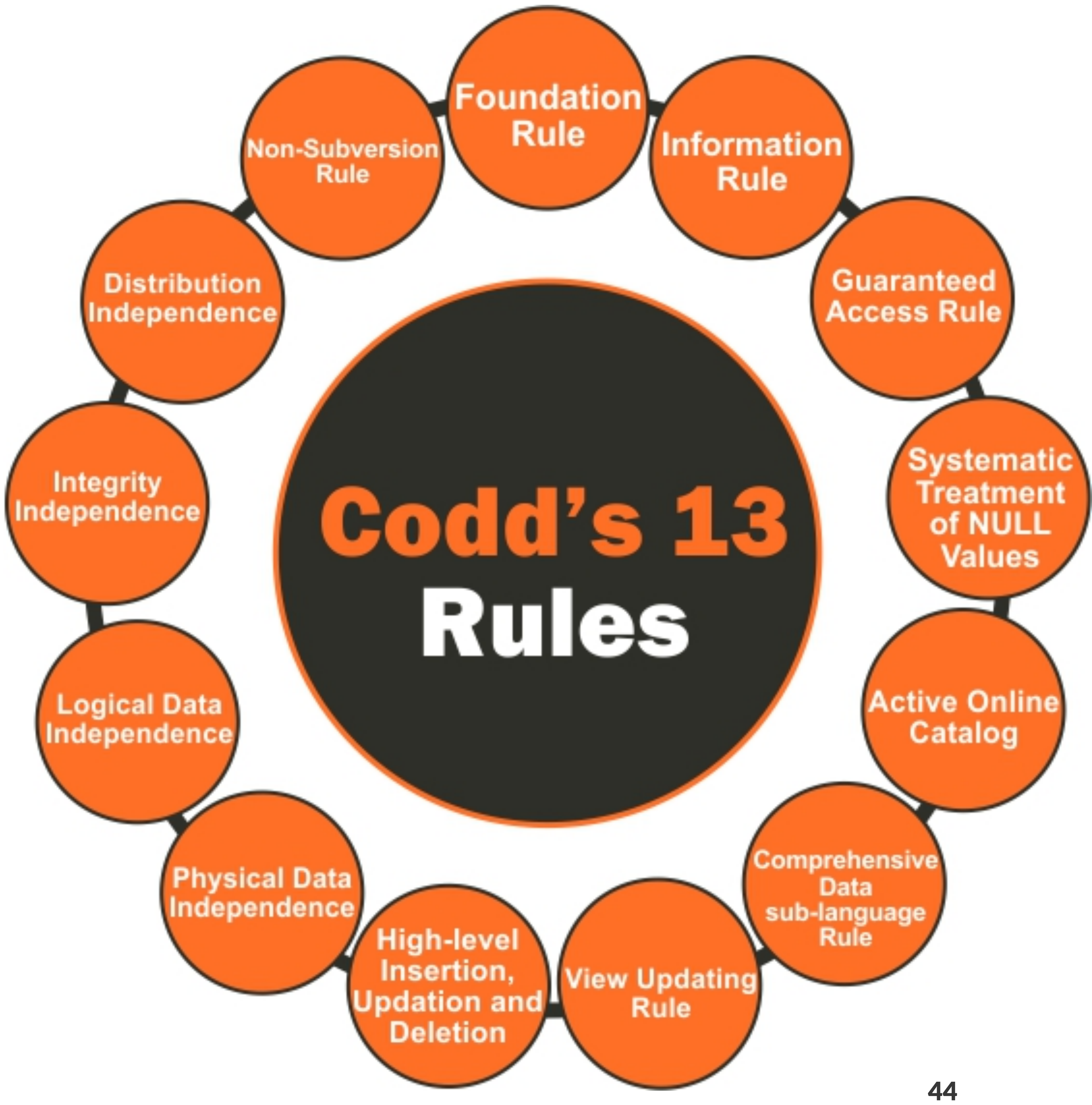
A database is **an organised collection of structured information, or data**, typically stored electronically in a computer system

Several kind of DBSMs exist. We will survey some of them. It is interesting to know that Edgar F. Codd defined 12+1 rules that make a DBMS relational link

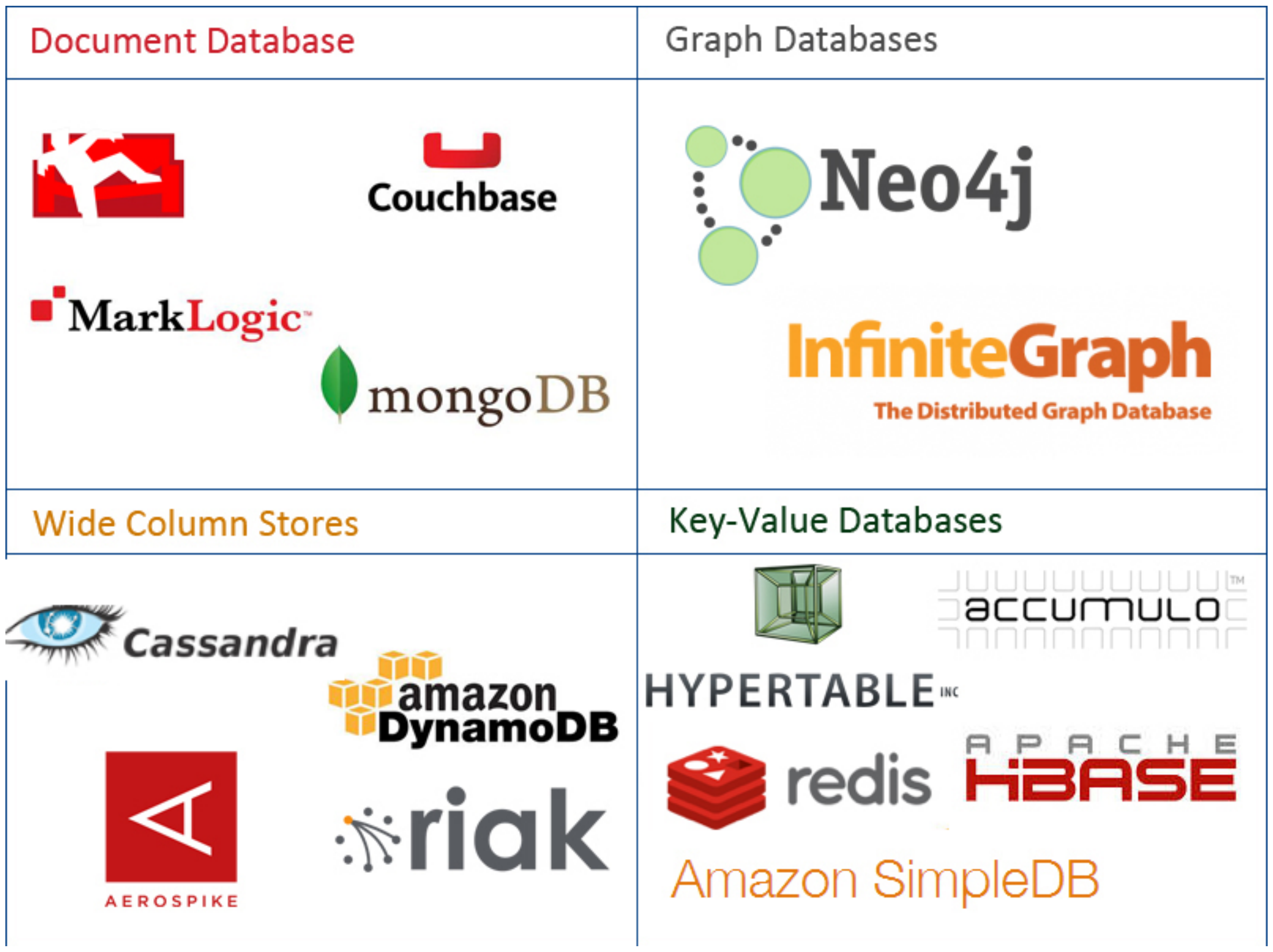


Relational DBMS

- It must be relational as a database and as a management system
- All data should be in table form
- All data should be accessible without ambiguity


Codd's 13 Rules
Foundation Rule
Information Rule
Guaranteed Access Rule
Systematic Treatment of NULL Values
Active Online Catalog
Comprehensive Data sub-language Rule
View Updating Rule
High-level Insertion, Updation and Deletion
Physical Data Independence
Logical Data Independence
Integrity Independence
Distribution Independence
Non-Subversion Rule

NoSQL Familty



Document Database	Graph Databases
Couchbase MarkLogic mongoDB	Neo4j InfiniteGraph The Distributed Graph Database
Wide Column Stores	**Key-Value Databases**
Cassandra amazon DynamoDB AEROSPIKE riak	HYPERTABLE INC accumulo redis APACHE HBASE Amazon SimpleDB

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

- A data warehouse is a central repository where raw data is transformed and stored in query-able forms.[03]
- Data Warehouse are still relevant today and their maintenance is part of Data Engineers' resposibilities.
- The warehouse is created with structure and model first before the data is loaded and it is called schema-on-write.

[03] What is Data Engineering

Data Warehouse vs Data Bases

Surprisingly, Data Warehouse isn't a regular database.

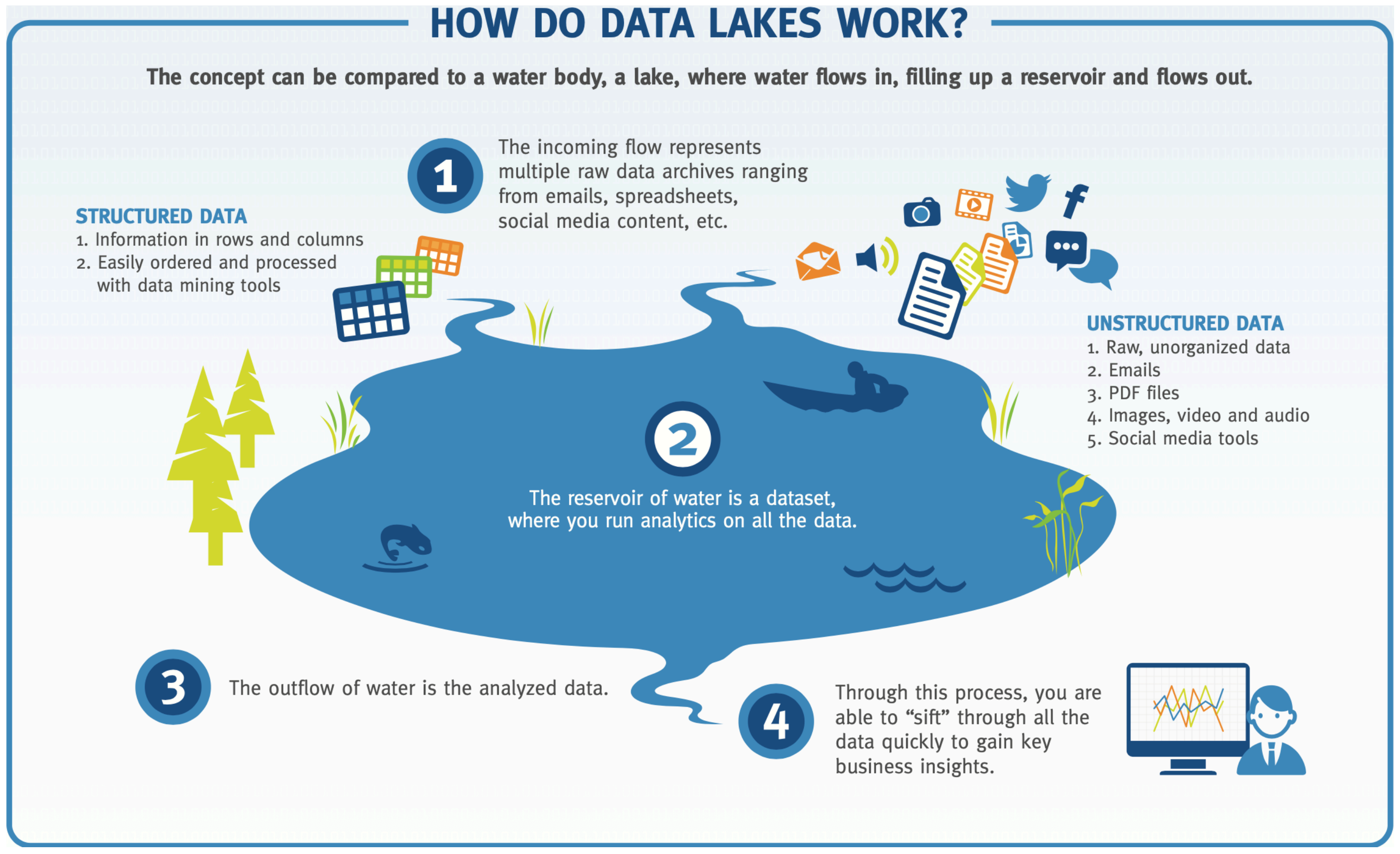
- A database normalizes data separating them into tables and avoiding redundancies
- It supports arbitrary workload and complex queries
- do not store multiple versions of data

- a Data Warehouse uses few tables to improve performance and 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 configuration[03].

[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.

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.

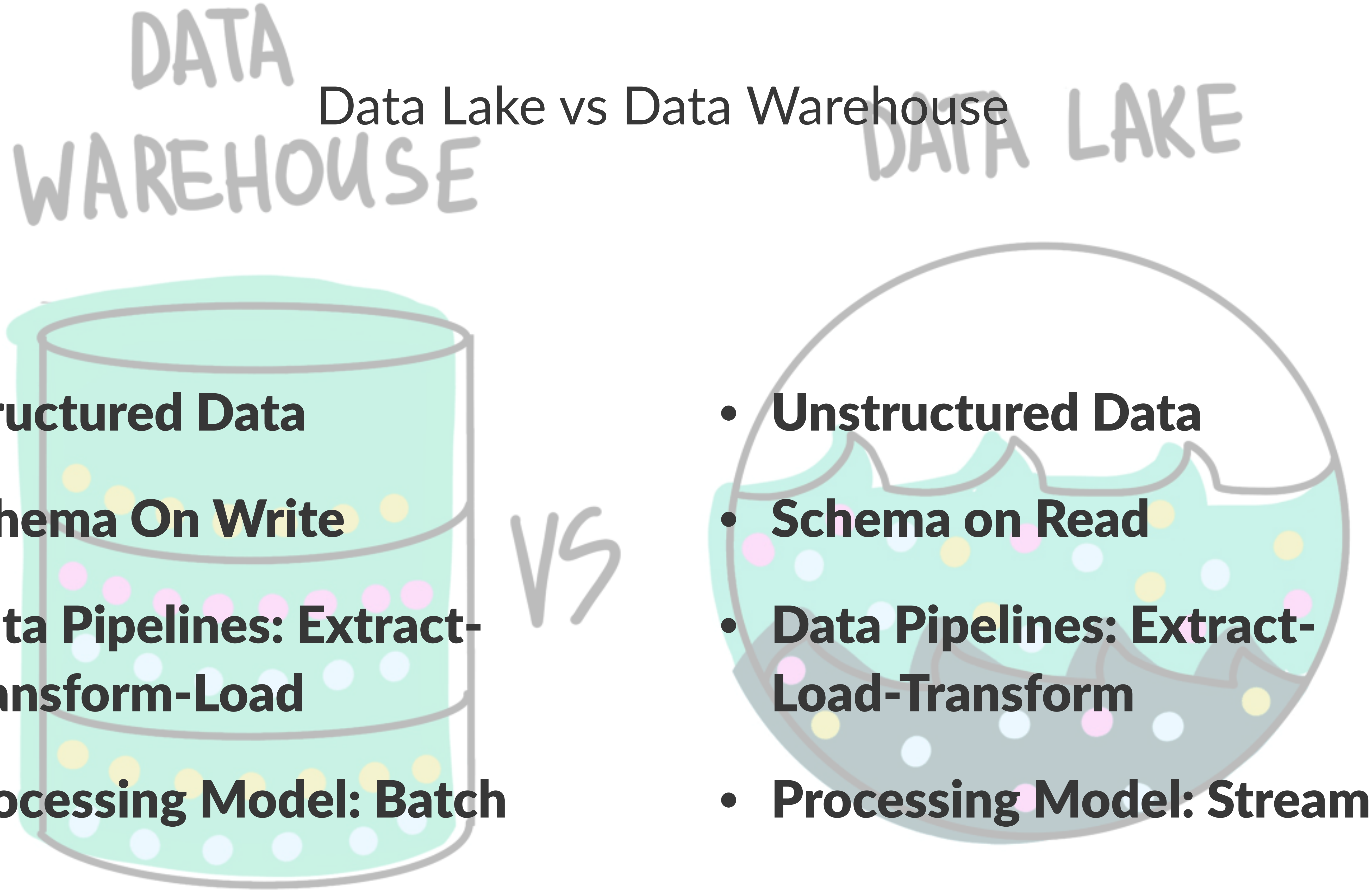
3 The outflow of water is the analyzed data.

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

Full Inforgraphic

- In Data Lake, the raw data is loaded as-is, when the data is used it is given structure, and it is called schema-on-read.
- Data Lake gives engineers the ability to easily change.
- In practice, Data Lake is a commercial term so don't sweat it.

Data Lake vs Data Warehouse



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

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