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

MCF Riccardo Tommasini

http://rictomm.me

riccardo.tommasini@insa-lyon.fr

Who I Am?

- PhD In 2015-2020
- Assistant Prof in Estonia (Data Management) 2019-2021
- Maître de conférencés at INSA 2021-now
- Member of LIRIS Lab

What do I Do?

Stream Processing

- real-time data analytics
- time-series analysis
- continuous data integration (hiring interns and PhD Students)

Graph Processing

- Graph Databases
- Incremental Graph Query Processing
- Graph repair and consistency

PS: Also working on Internet Memes at https://meme4.science

Course Intro

- Lectures
  - F2F Covering a range of "theoretical topics"
  - Two Guest Lectures (TBA)
  - Part of the MCQ
- TP
  - Planning to record 45/60 min videos to watch before TD
  - in-class Project sections (TP) in November
- Grading
  - 40% with MCQs based on TPs
  - 40% project (Team of 2/3)
  - 20% project presentation
  - 5% Bonus
- Infos:
  - Waiting for Moodle
  - Full Schedule out this week (sorry)

Project Theme

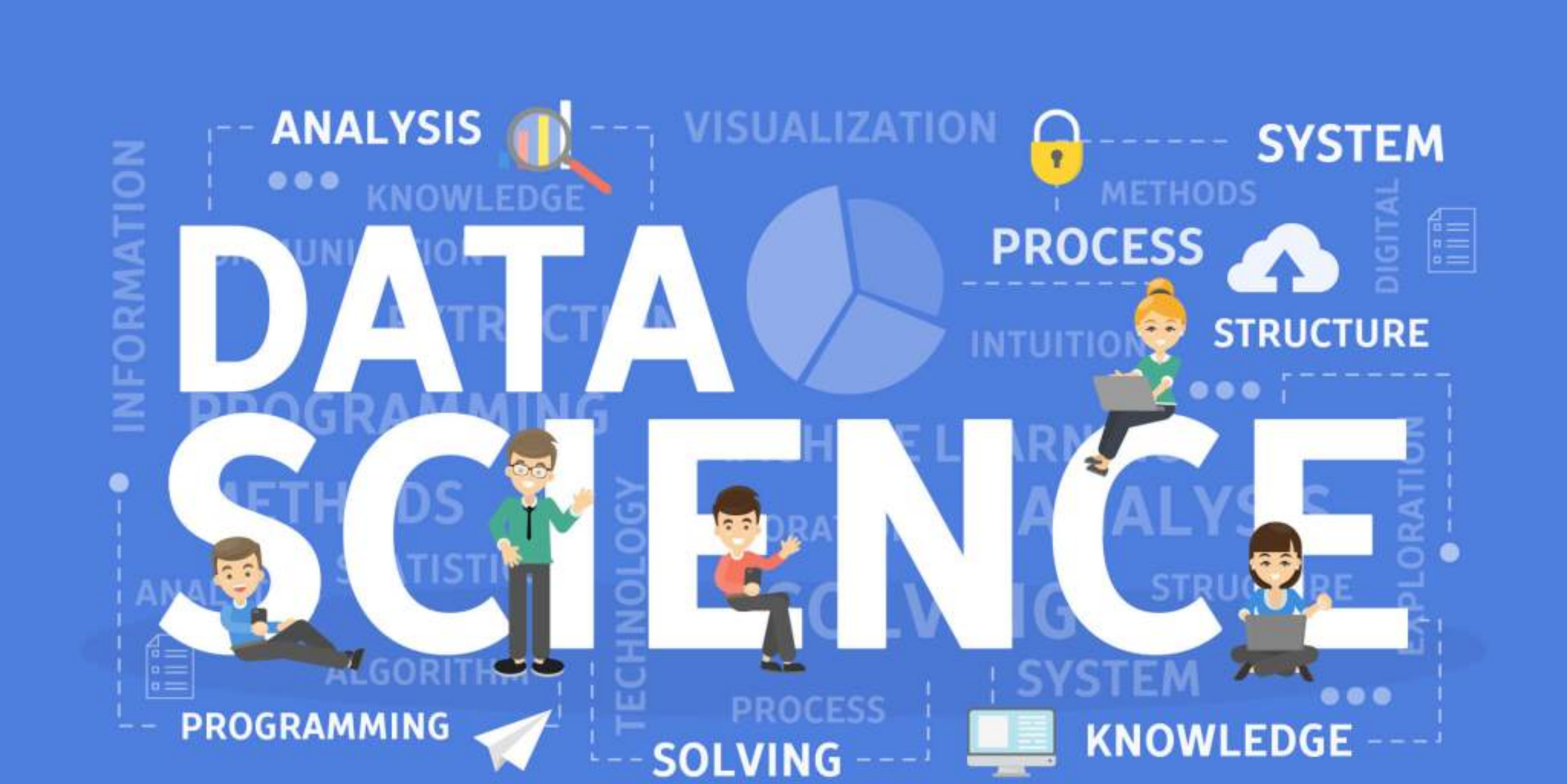
You will have to implement a data pipeline using Apache Airflow combining different databases (using) all orchestrated using docker (compose).

Pipeline covers all the steps of the data lifecycle (will see today)

Checklist for the pipeline will be published soon.

Topics:

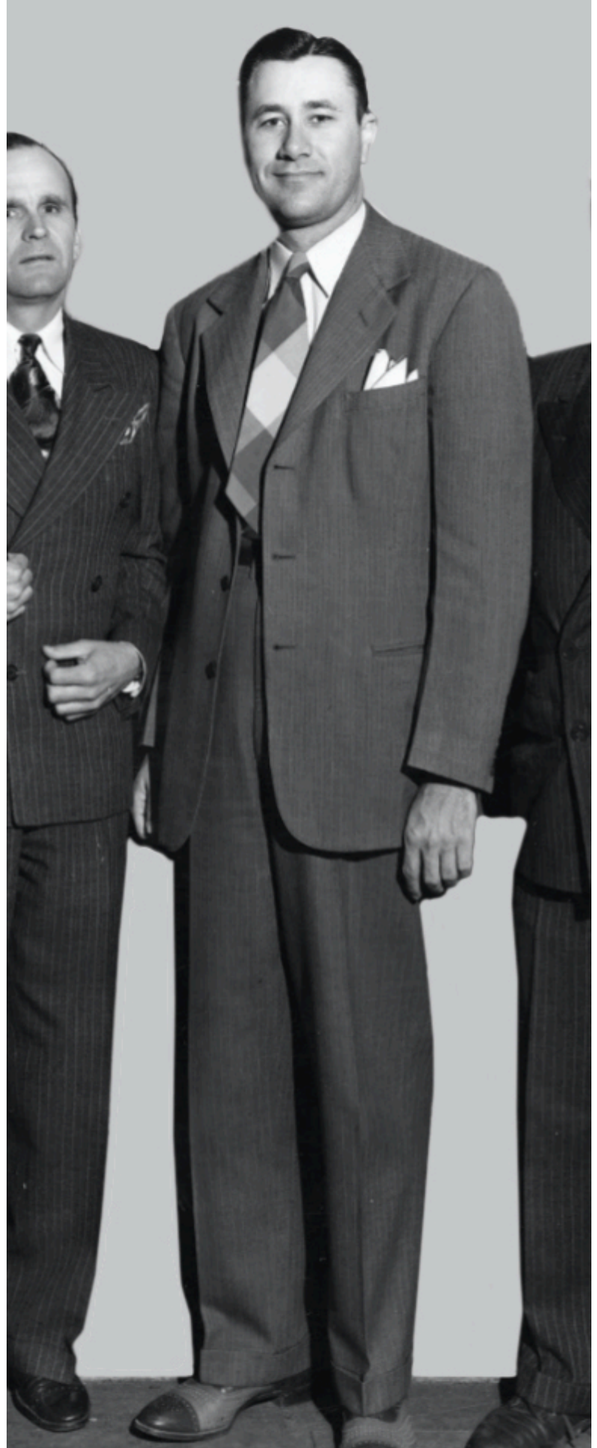
- Internet Memes
- Environmental Impact of Energy Sources
- Come with your proposal (at your own risk)


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

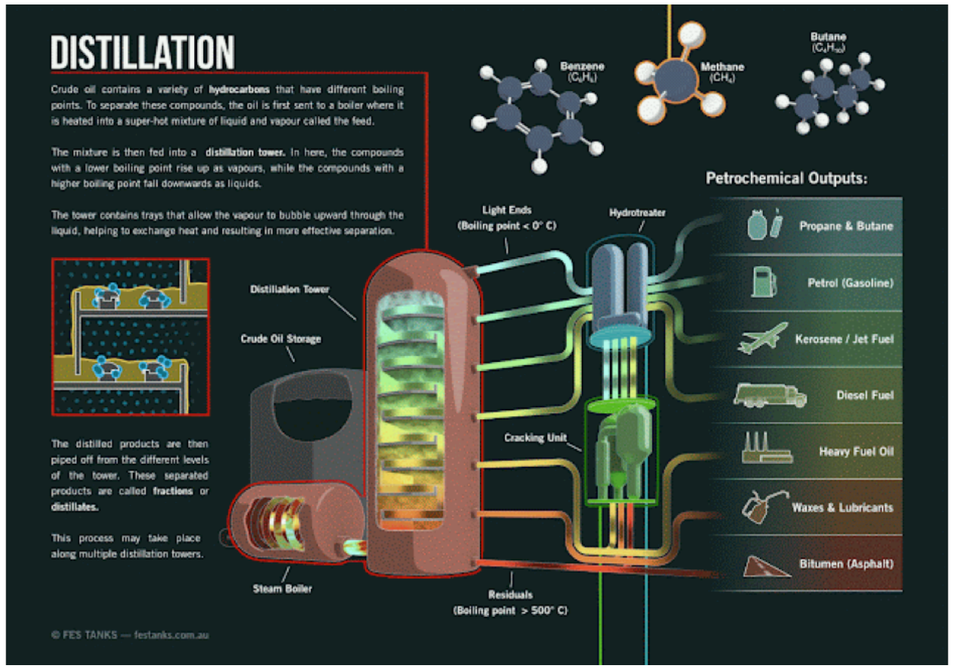
Quote

“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]



...refining crude oil

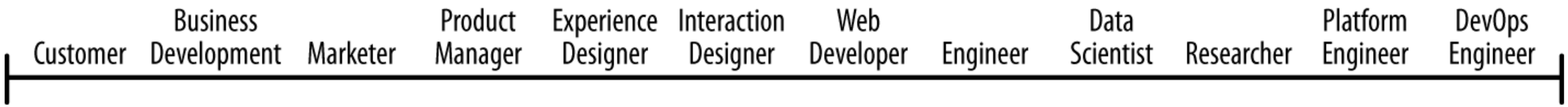
---

[01] Source

Data Engineering is...

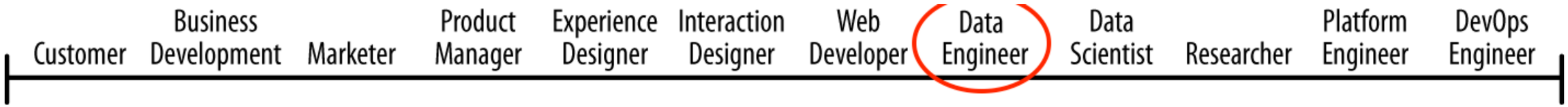
...build the refinery.

Roles in a Data Science Project[02]

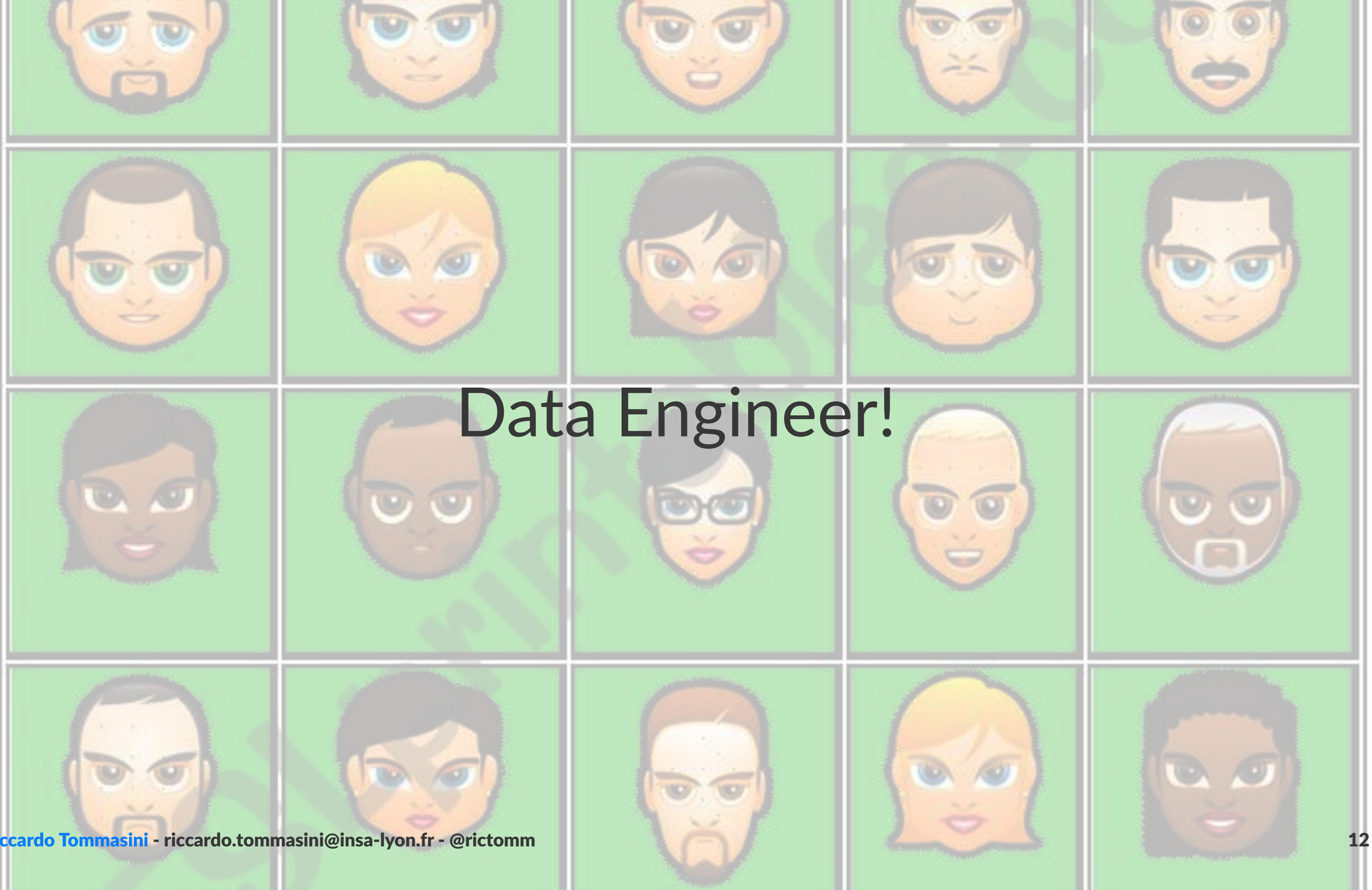


[02] http://emanueledellavalle.org/slides/dspm/ds4biz.html#25

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

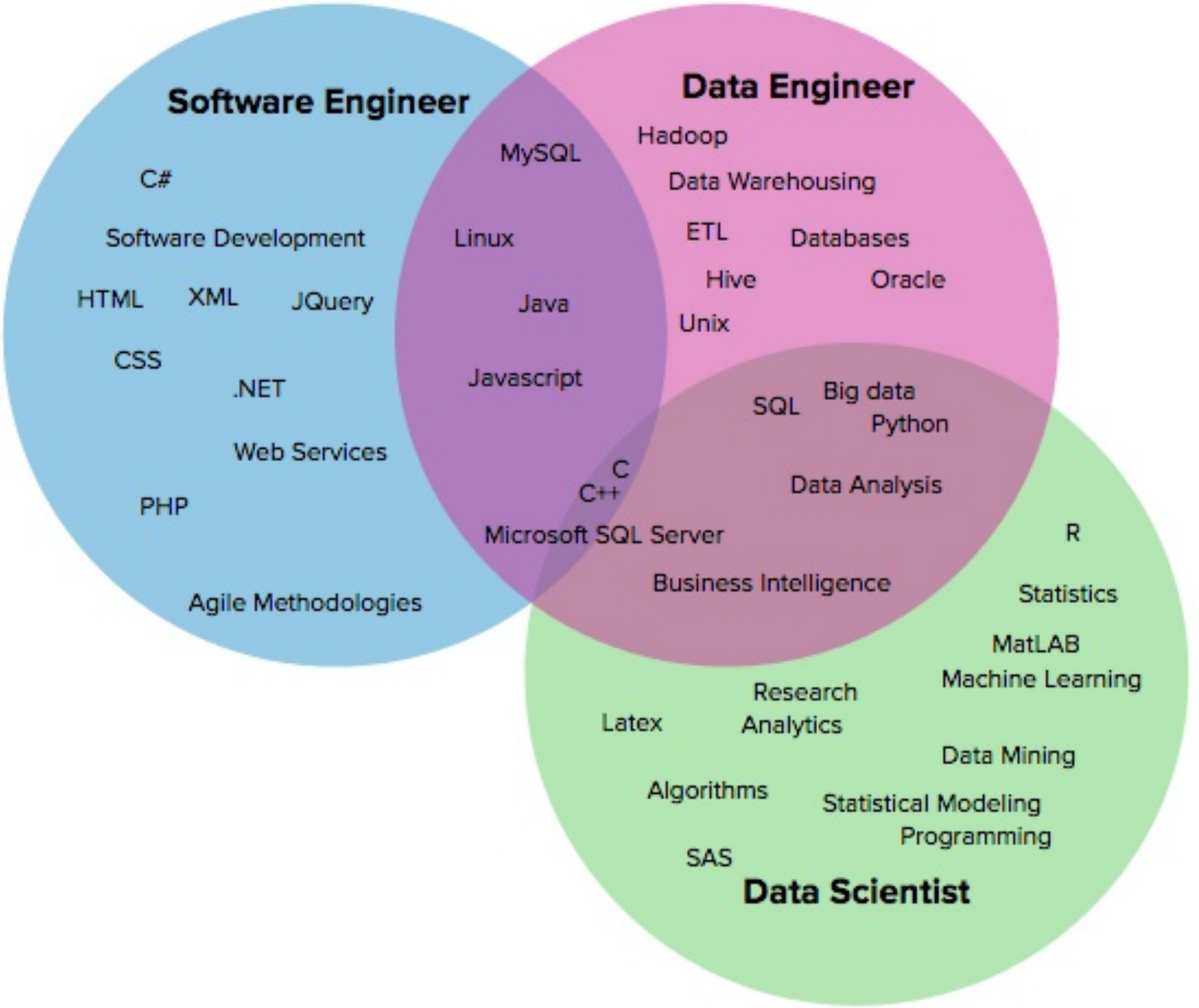
Them: Can you just quickly pull this data for me?

Me: Sure, let me just:

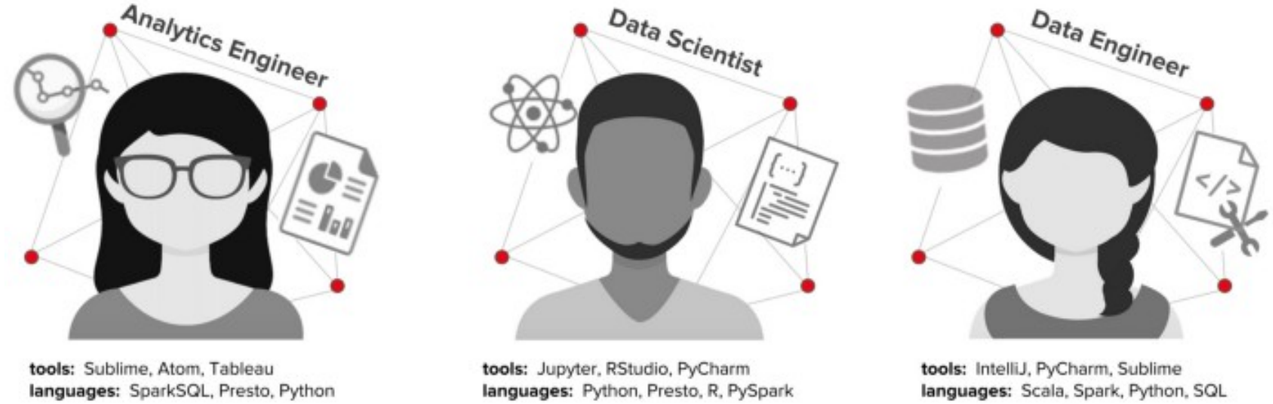
```
SELECT * FROM
some_ideal_clean_and_pristine.table_that_you_think_exists
```

323 4.4K 28K

Show this thread

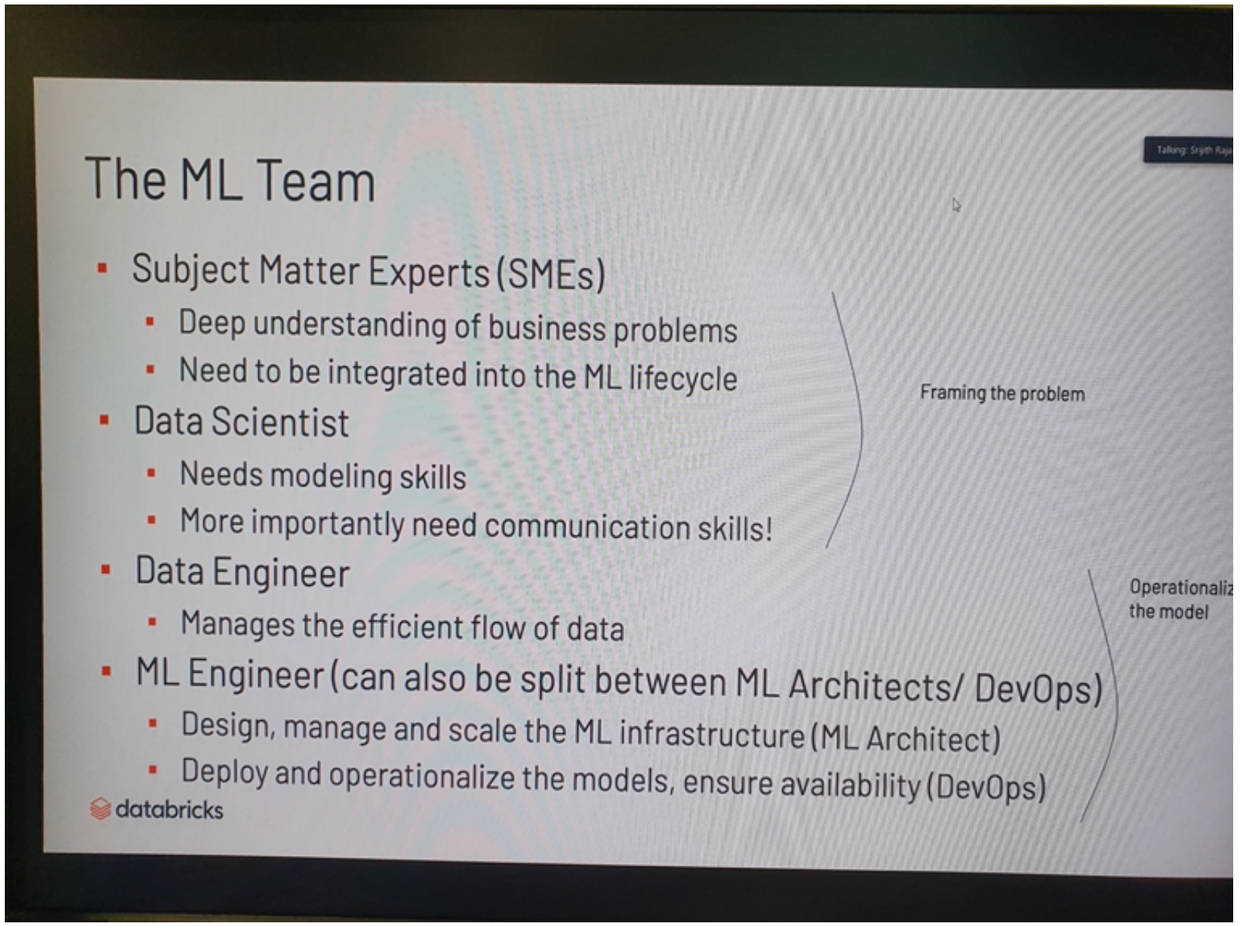

Software Engineer
C#
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
Data Engineer
MySQL
Hadoop
Data Warehousing
Linux
ETL
Databases
Hive
Oracle
Java
Unix
Javascript
SQL
Big data
Python
C
C++
Data Analysis
Microsoft SQL Server
R
Business Intelligence
Statistics
MatLAB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

Netflix's Perspective[05]



[05] Netflix Innovation

Data Bricks


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)
databricks
Framing the problem
Operationaliz
the model

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 Scientist[06]

[06] The Manifesto

Philosophy of (Data) Science[07]

[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]

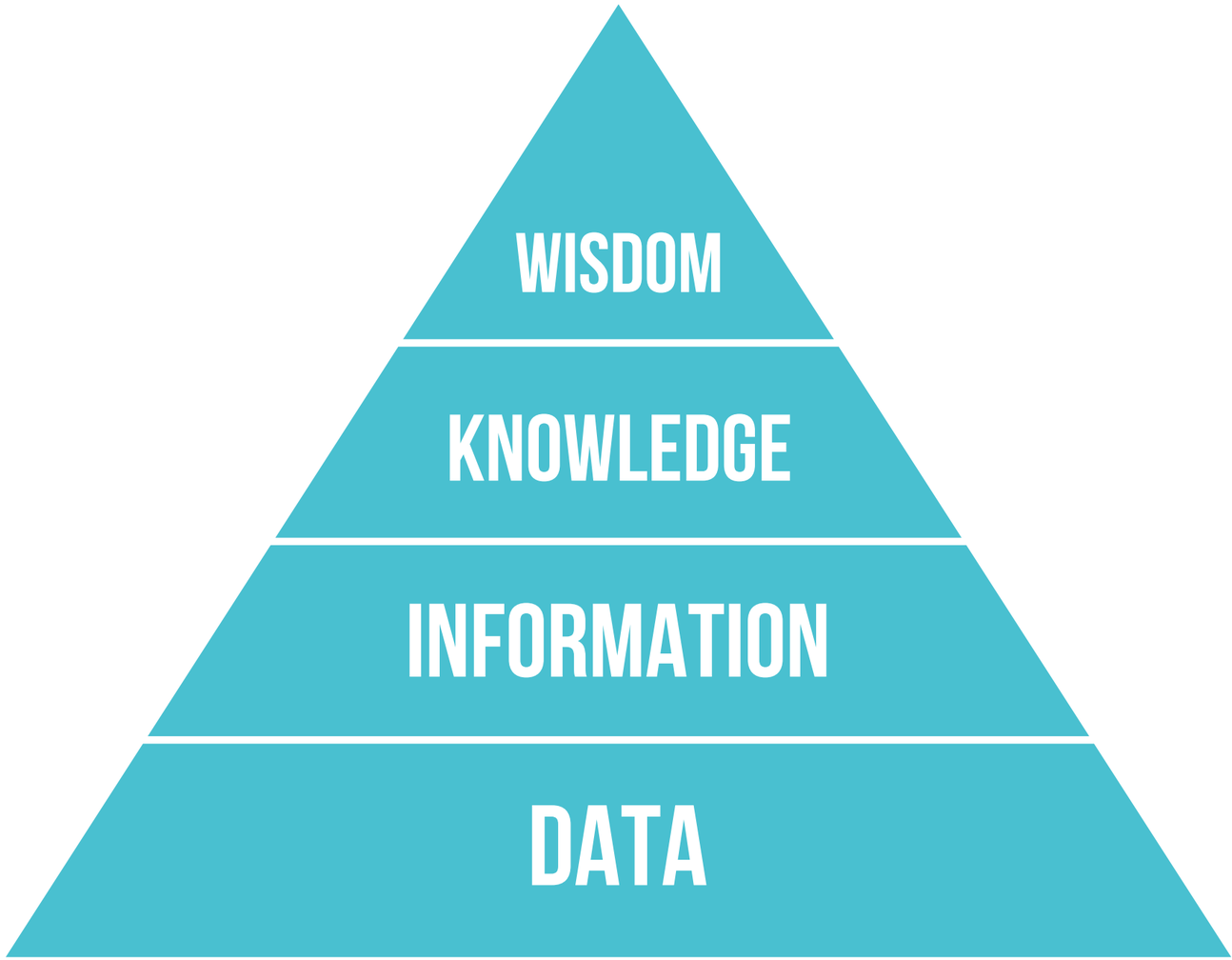
[08] Def

Wikipedia

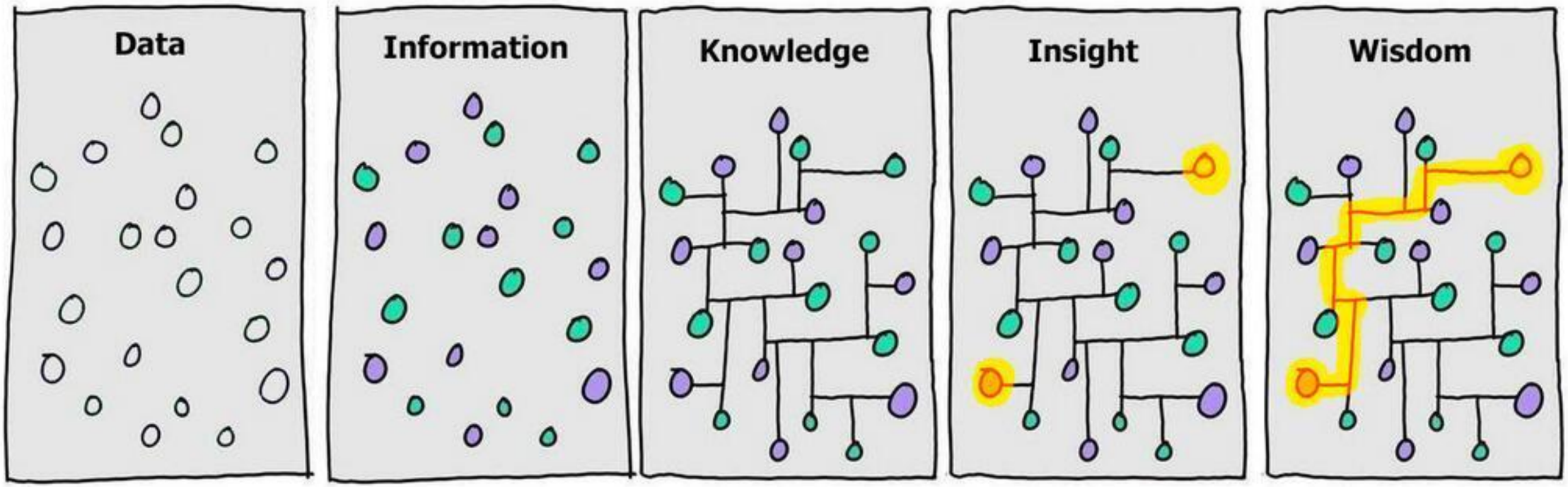
Data (treated as singular, plural, or as a mass noun) is any sequence of one or more symbols given meaning by specific act(s) of interpretation [09]

[09] Data in Computing)

DIKW Pyramid

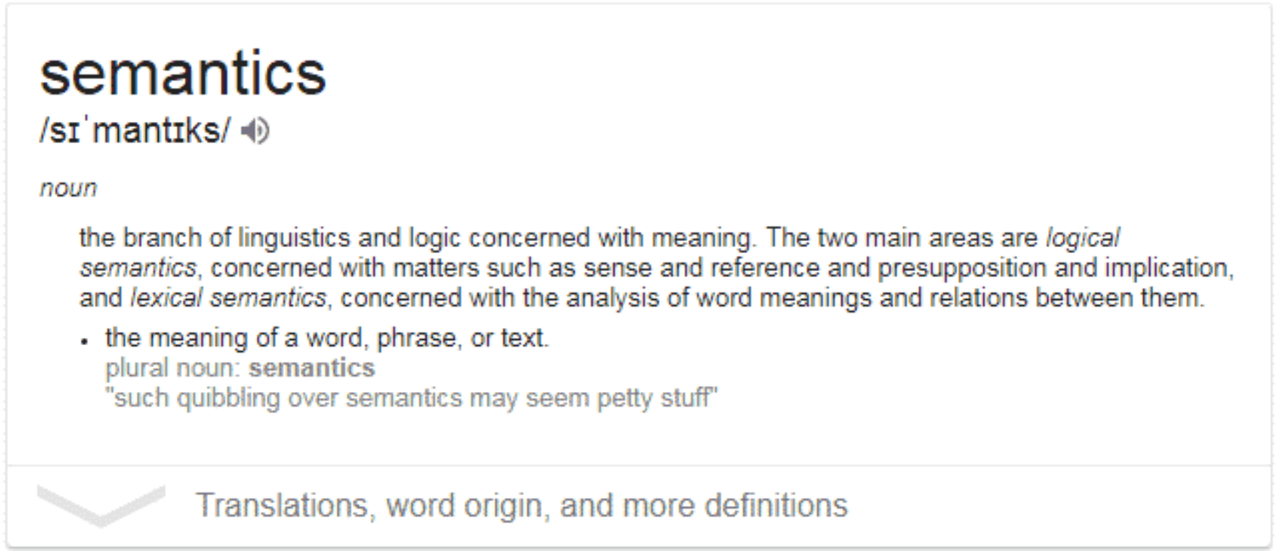


Graph View



Data about data

Data Semantics



semantics

/sɪˈmantɪks/

noun

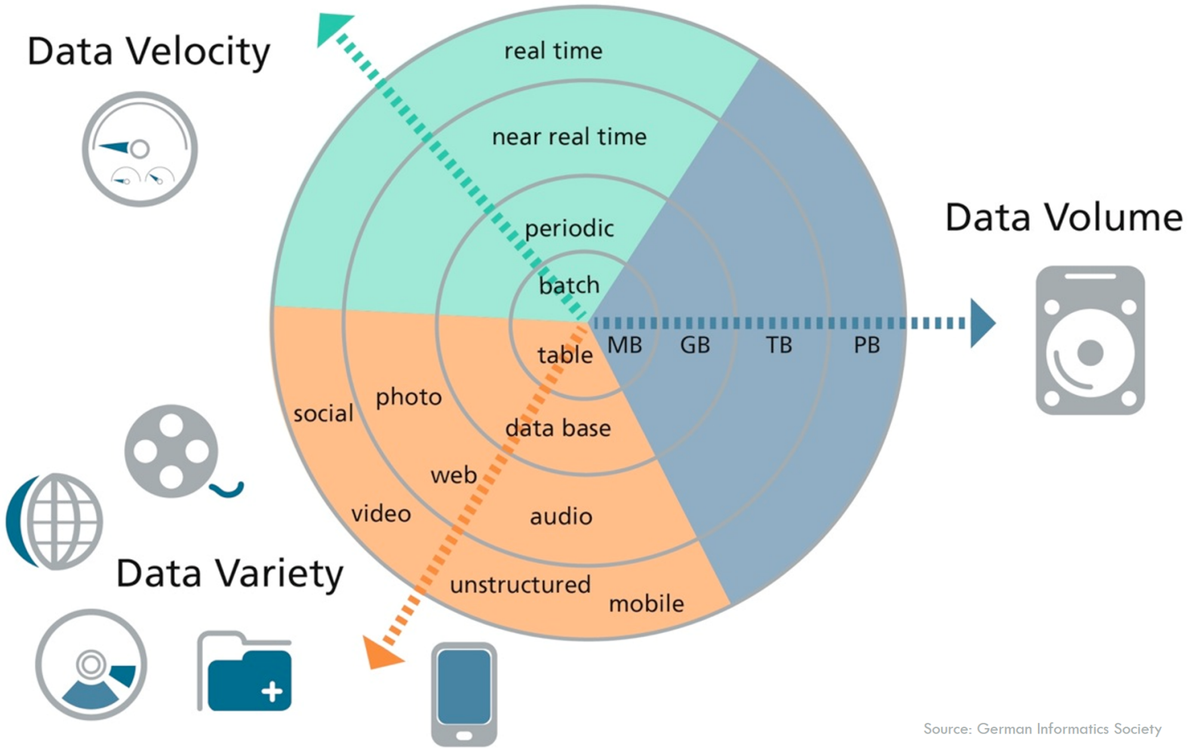
the branch of linguistics and logic concerned with meaning. The two main areas are *logical semantics*, concerned with matters such as sense and reference and presupposition and implication, and *lexical semantics*, concerned with the analysis of word meanings and relations between them.

- 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

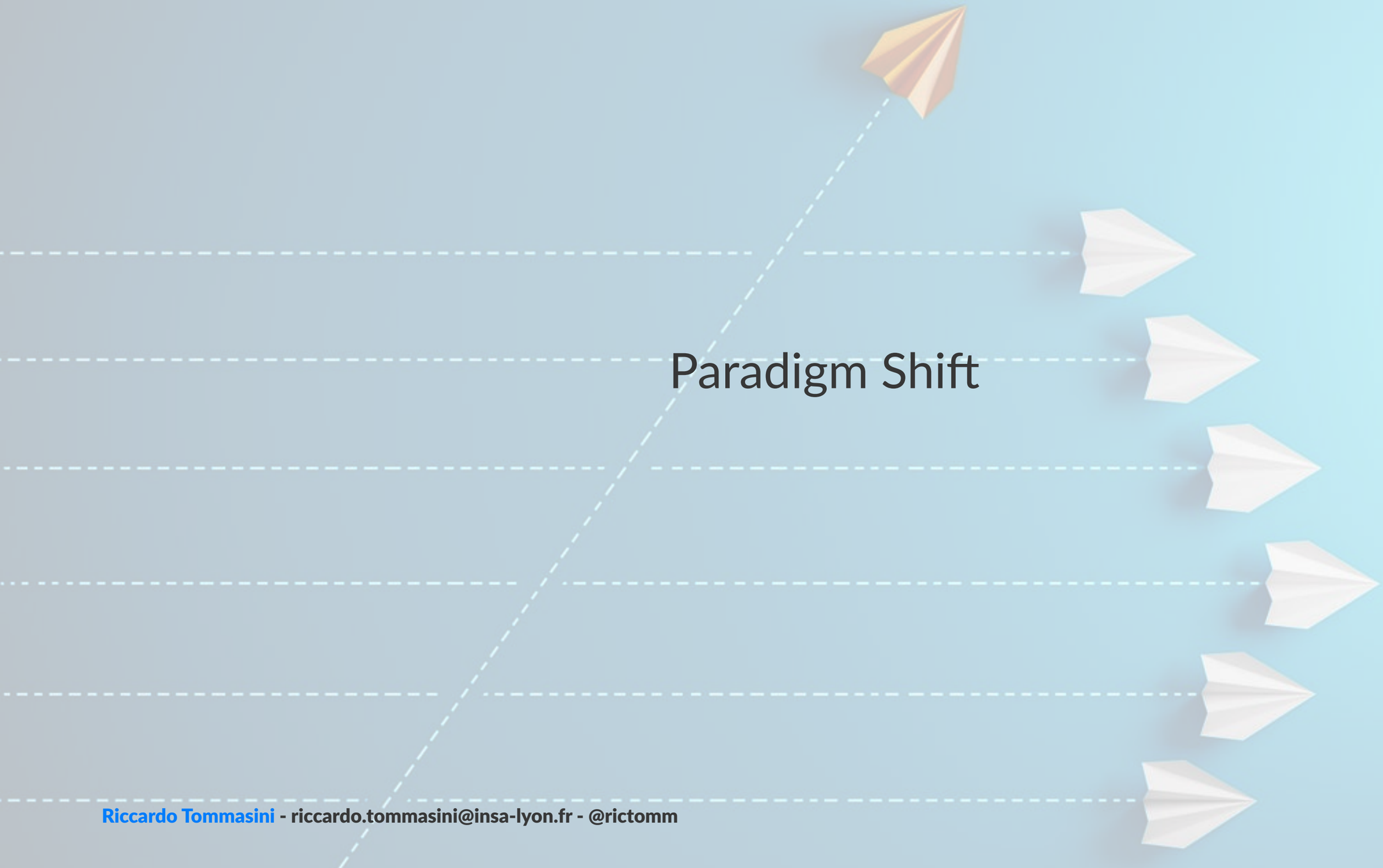
Big Data

Challenges [014]



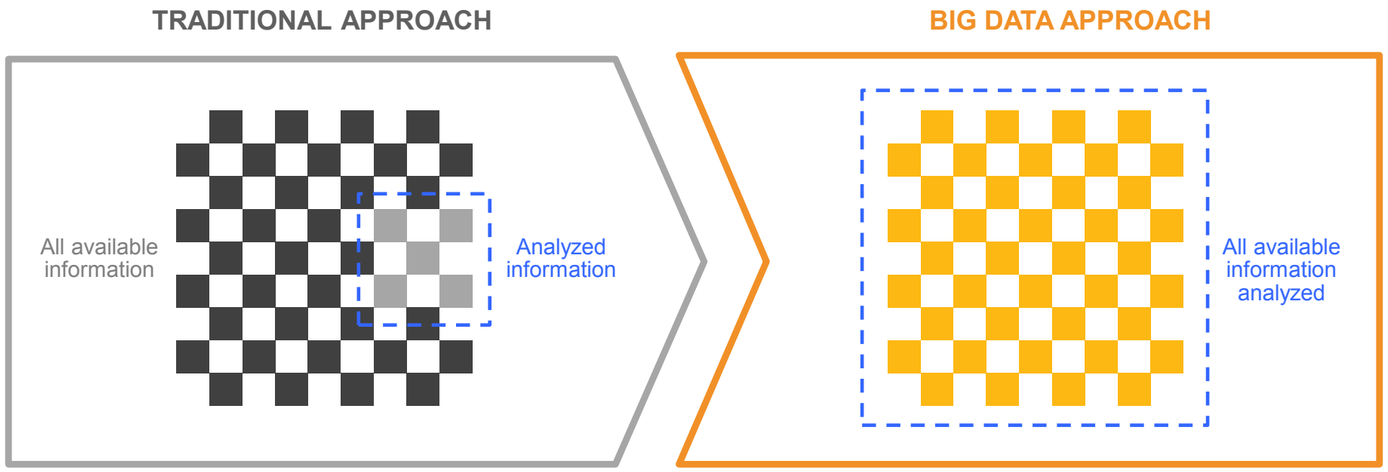
[014] Lanely, 2001

Paradigm Shift



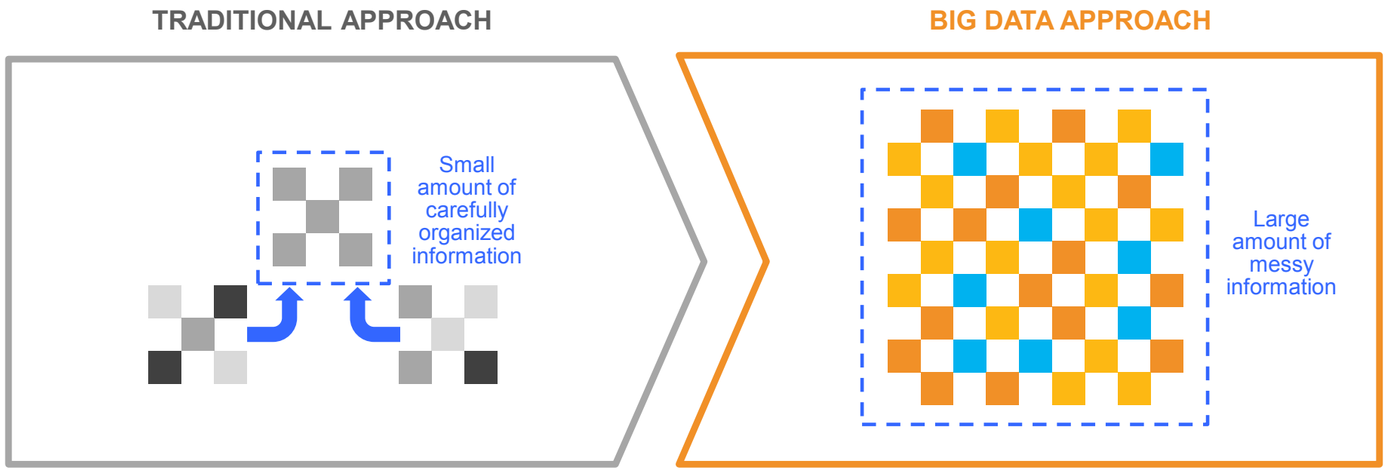
Paradigm shifts enabled by big data

Leverage more of the data being captured



Paradigm shifts enabled by big data

Reduce effort required to leverage data



Carefully cleanse information ***before*** any analysis

Analyze information as is, cleanse as needed

Paradigm shifts enabled by big data

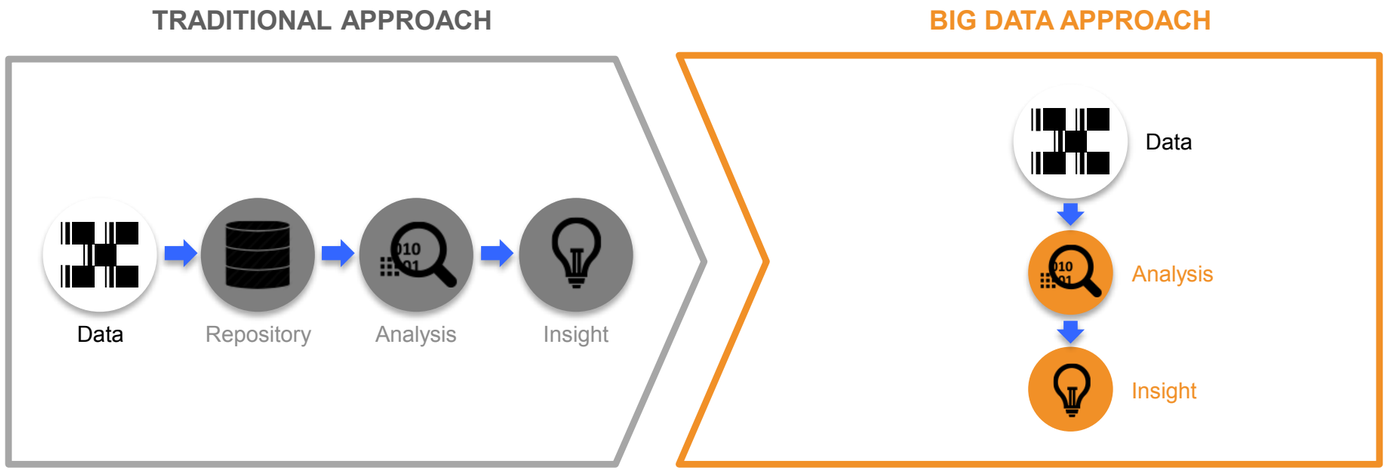
Data leads the way—and sometimes correlations are good enough

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

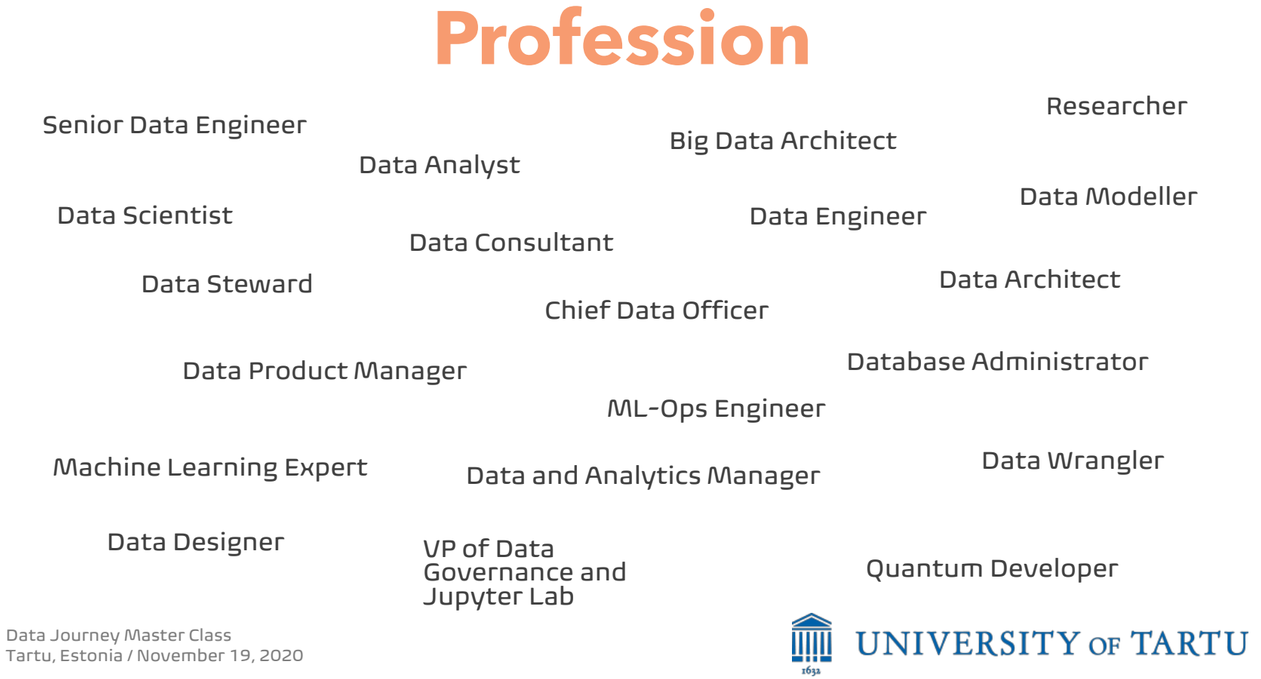


Analyze data ***after*** it's been processed and landed in a warehouse or mart

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

New Roles

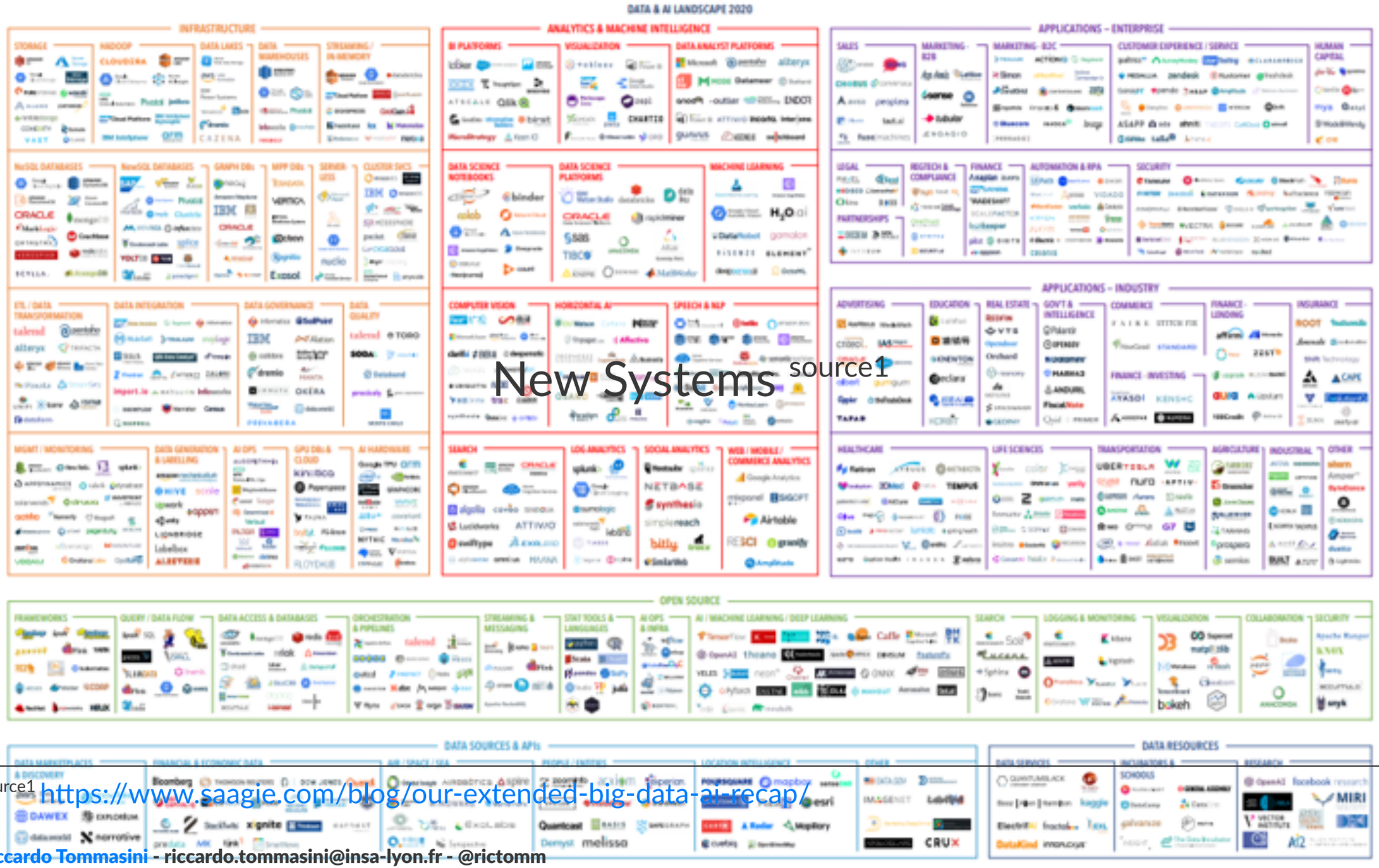
In the context of Big Data, a data engineer must focus on **distributed systems**, and **programming languages** such as Java and Scala.



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.



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

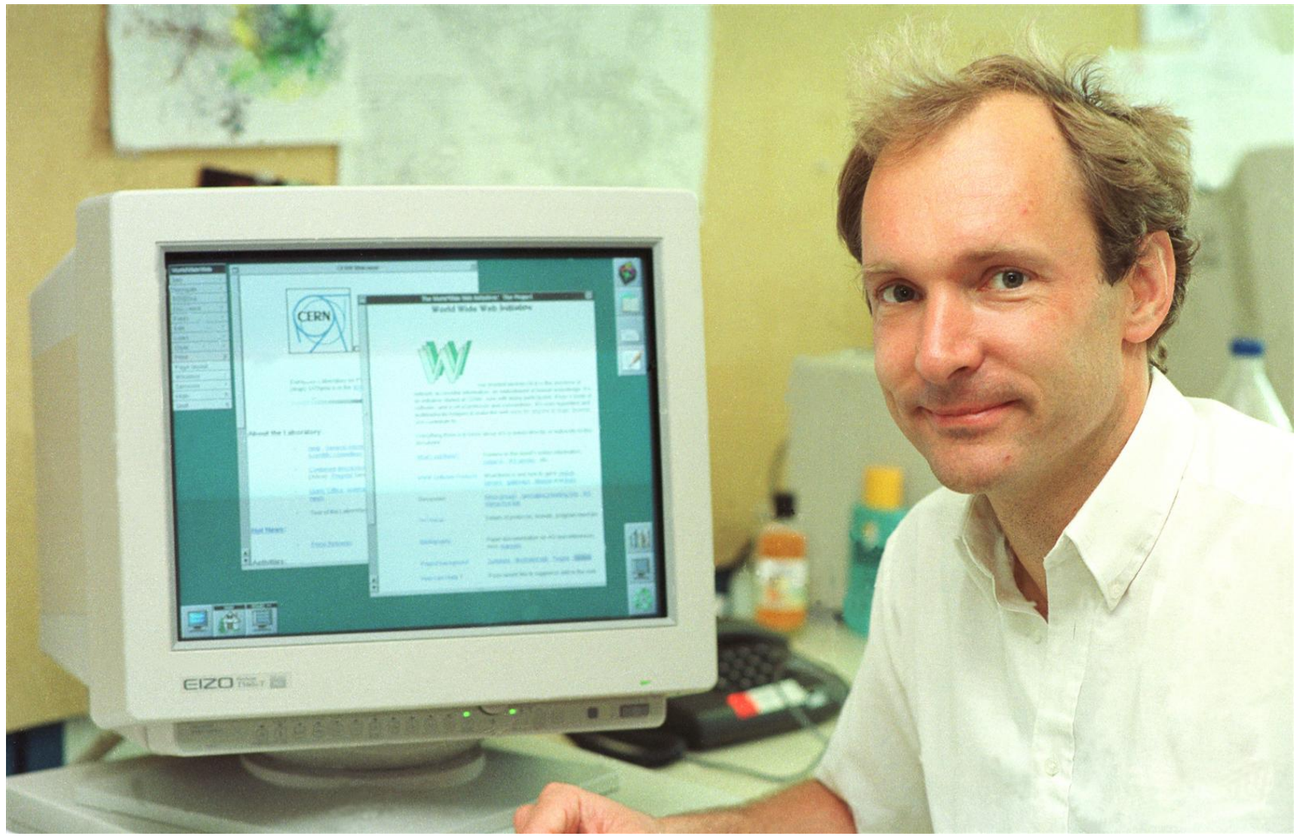
[pt] Data on the Outside vs Data on the Inside *Pat Helland, CIDR 2005*

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

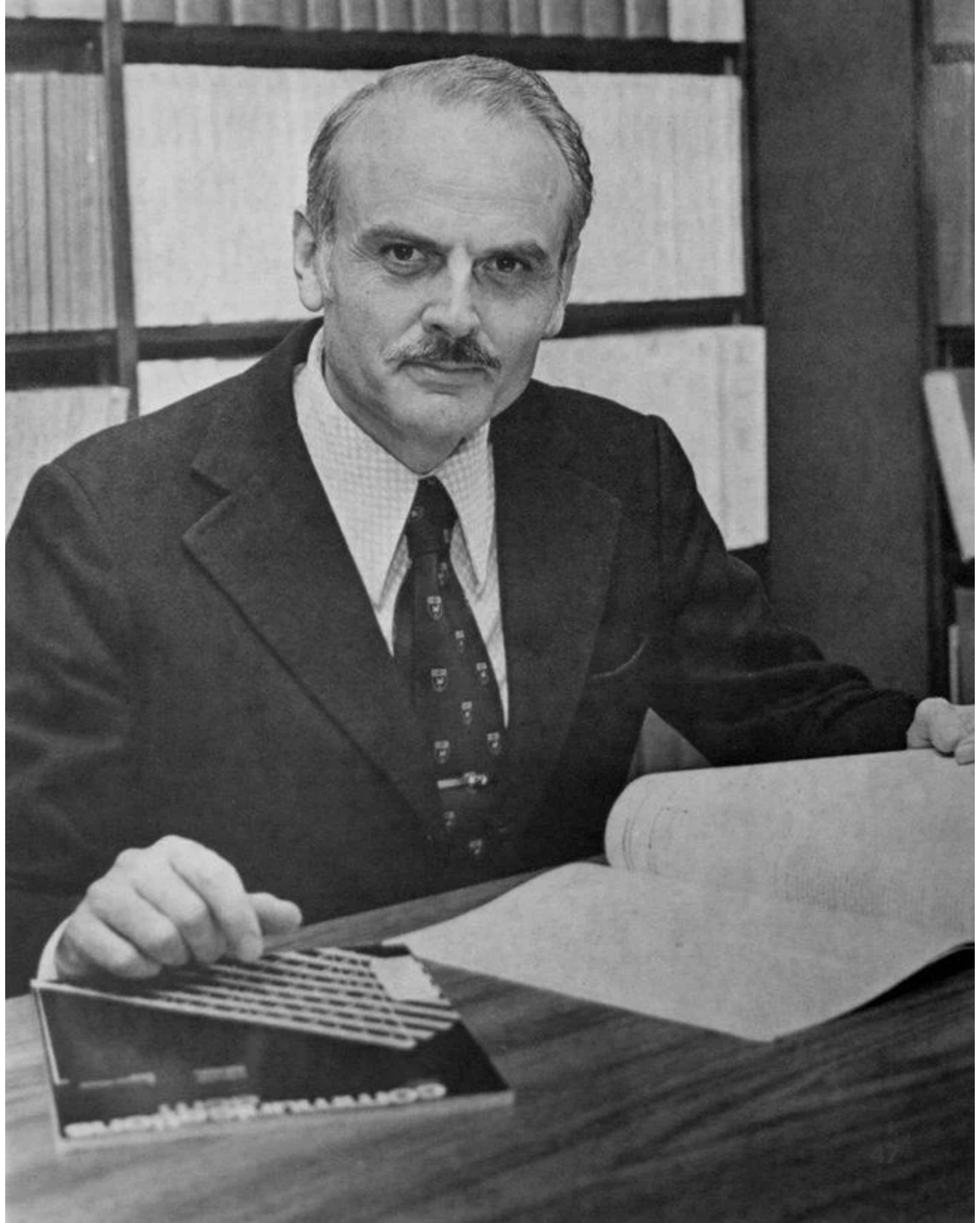
Protocols: HTTP

- 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 System

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



Riccardo Tommasini - riccardo.tommasini@insa-lyon.fr - @rictomm

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

Beyond Relational (No Only SQL)

Google, Amazon, Facebook, and DARPA all recognised that when you scale systems large enough, you can never put enough iron in one place to get the job done (and you wouldn't want to, to prevent a single point of failure).

Once you accept that you have a distributed system, you need to give up consistency or availability, which the fundamental transactionality of traditional RDBMSs cannot abide.
--Cedric Beust

The Reasons Behind

- **Queryability**: need for specialised query operations that are not well supported by the relational model
- **Schemaless**: desire for a more dynamic and expressive data model than relational
- **Flexibility**: need to accomodate the "schema on read" phylosophy

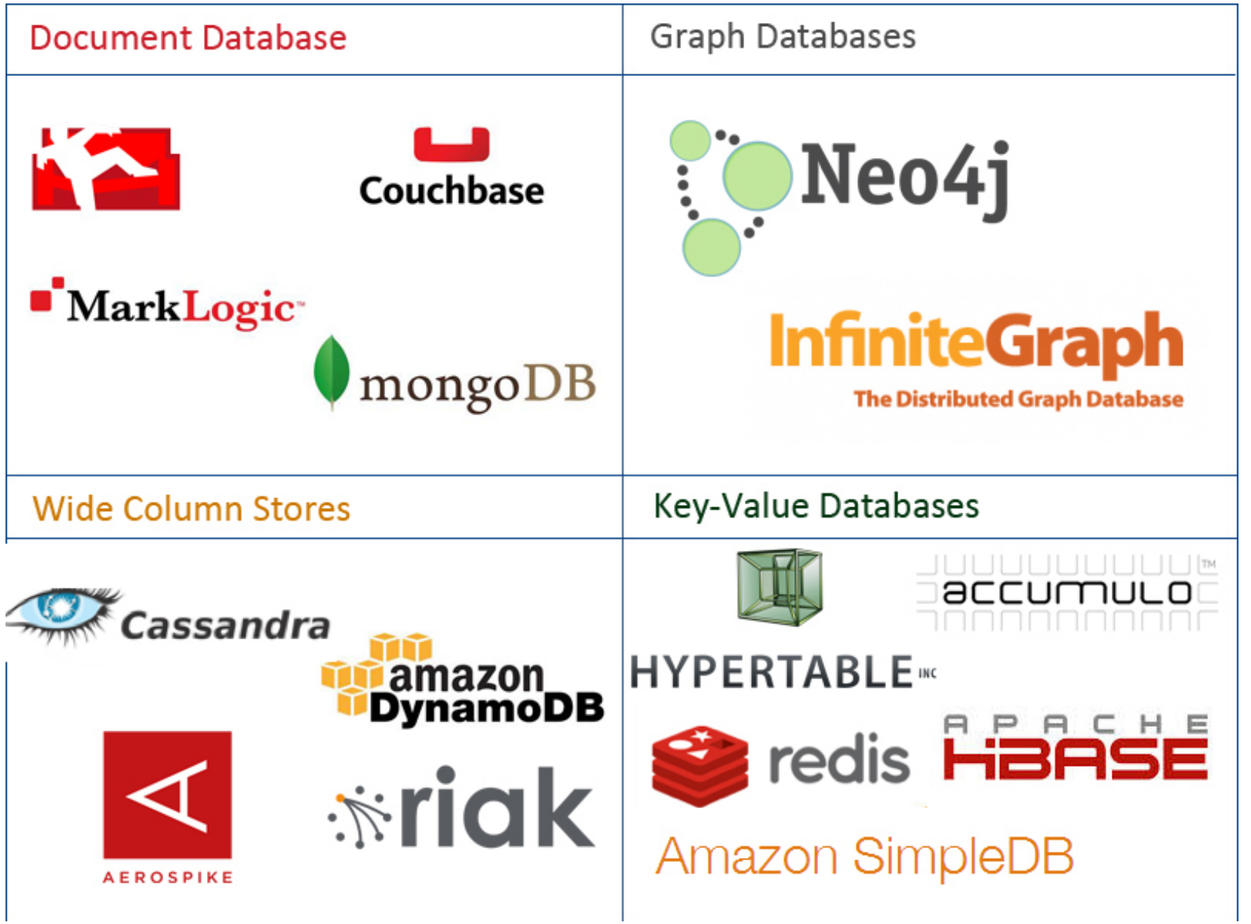
Object-Relational Mismatch

Most application development today is done in **object-oriented** programming languages

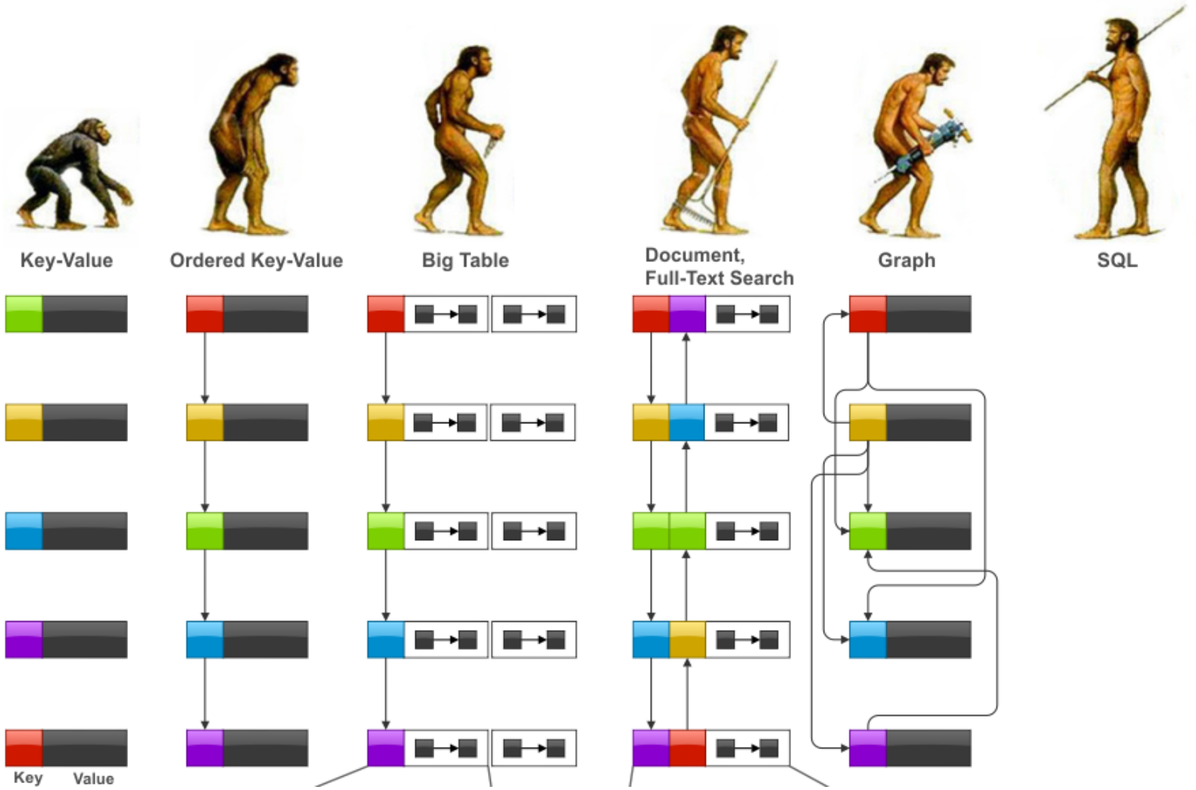
An **awkward translation** layer is required between the **objects** in the application code and the database model of **tables**, **rows**, and **columns**

Object-relational mapping (**ORM**) frameworks like **Hibernate** try to mild the mismatch, but they **can’t completely hide** the differences

NoSQL Familty



History of Data Models[5]



[5] by Ilya Katsov

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

Data Warehouse vs Data Bases

Surprisingly, Data Warehouse isn’t a regular database.

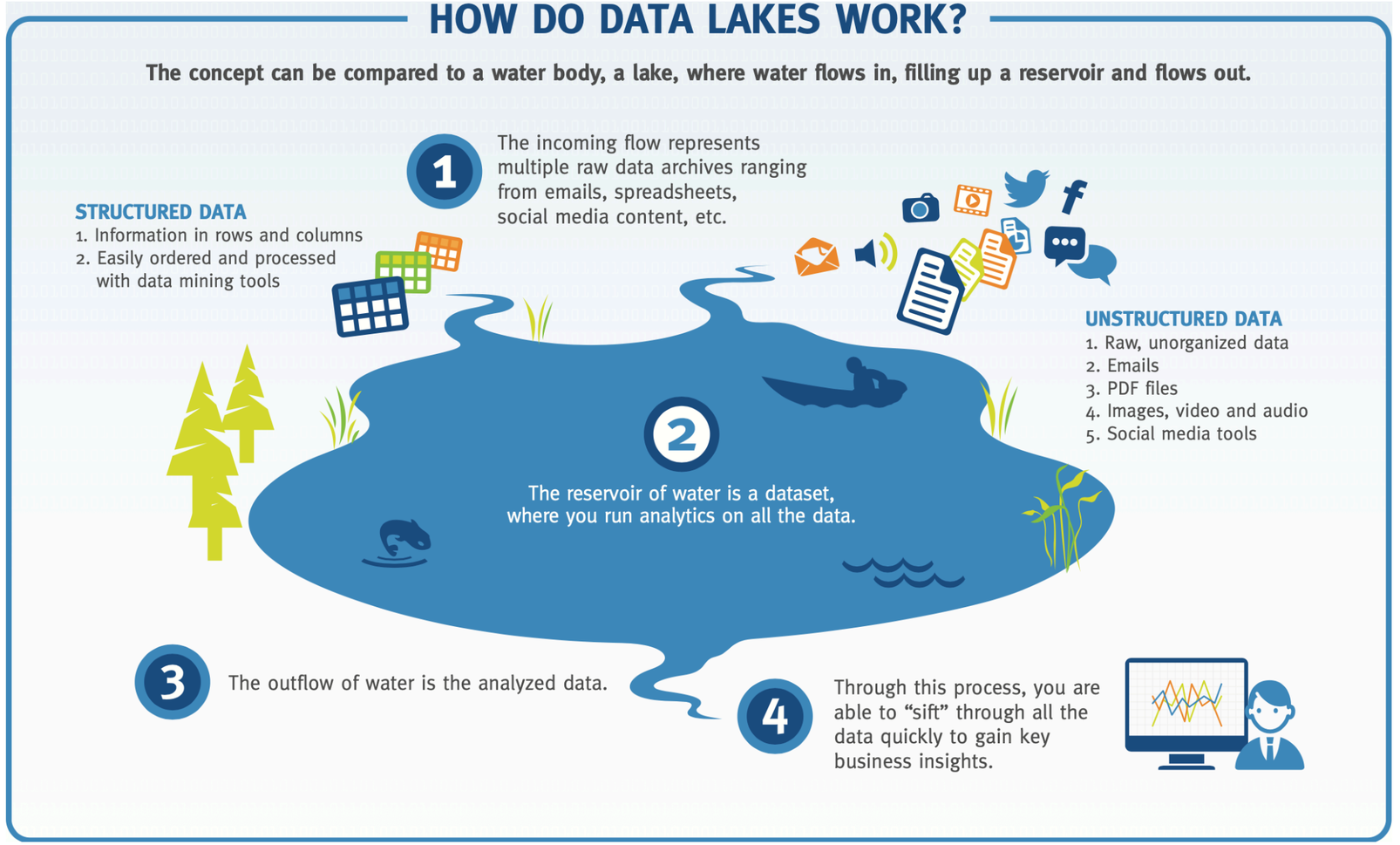
- A database normalizes data separating them into tables and avoiding redundancies
- It supports arbitrary workload and complex queries
- do not store multiple versions of data

- a Data Warehouse uses few tables to improve performance and analytic.
- a Data Warehouse allows simple queries
- supports versioning for complex analysis

Data Lake

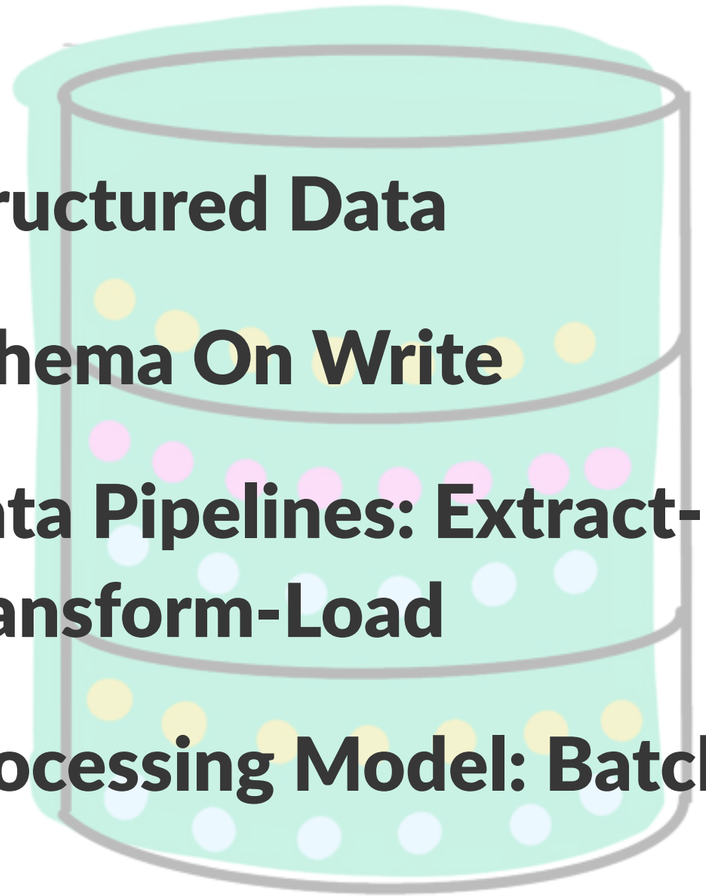
A Data lake is a vast pool of raw data (i.e., data as they are natively, unprocessed). A data lake stands out for its high agility as it isn't limited to a warehouse's fixed configuration[03].

[03] What is Data Engineering



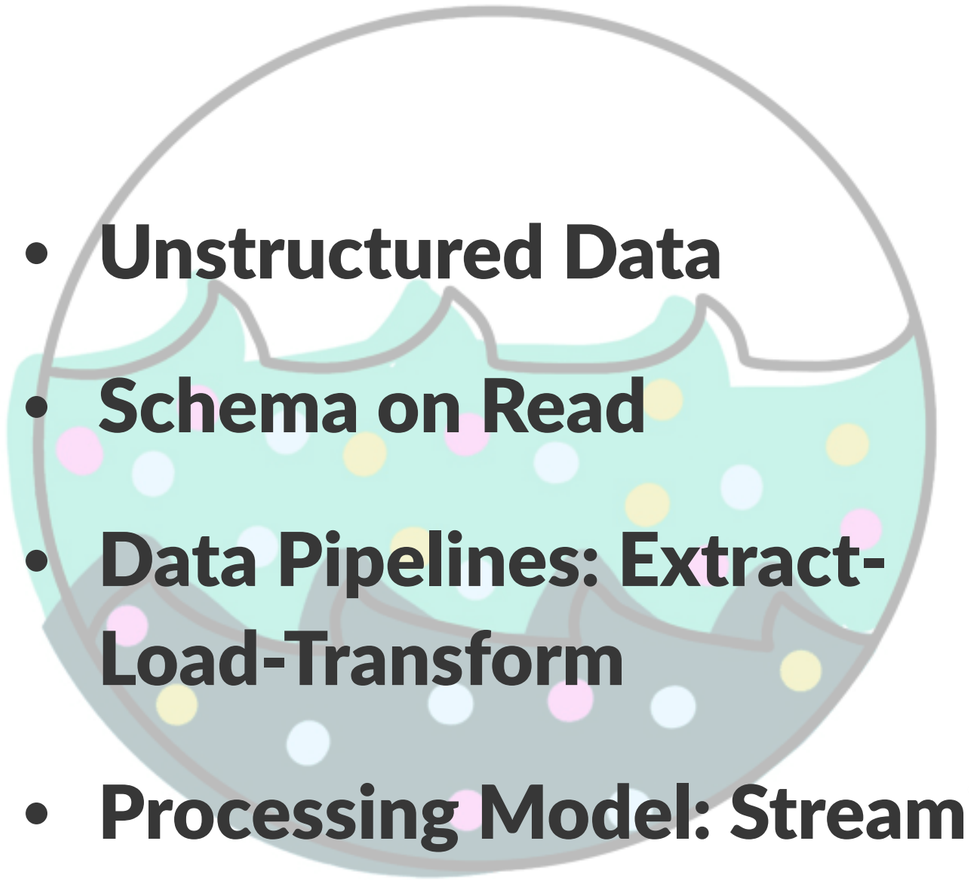
Full Inforgraphic

Data Lake vs Data Warehouse



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

VS

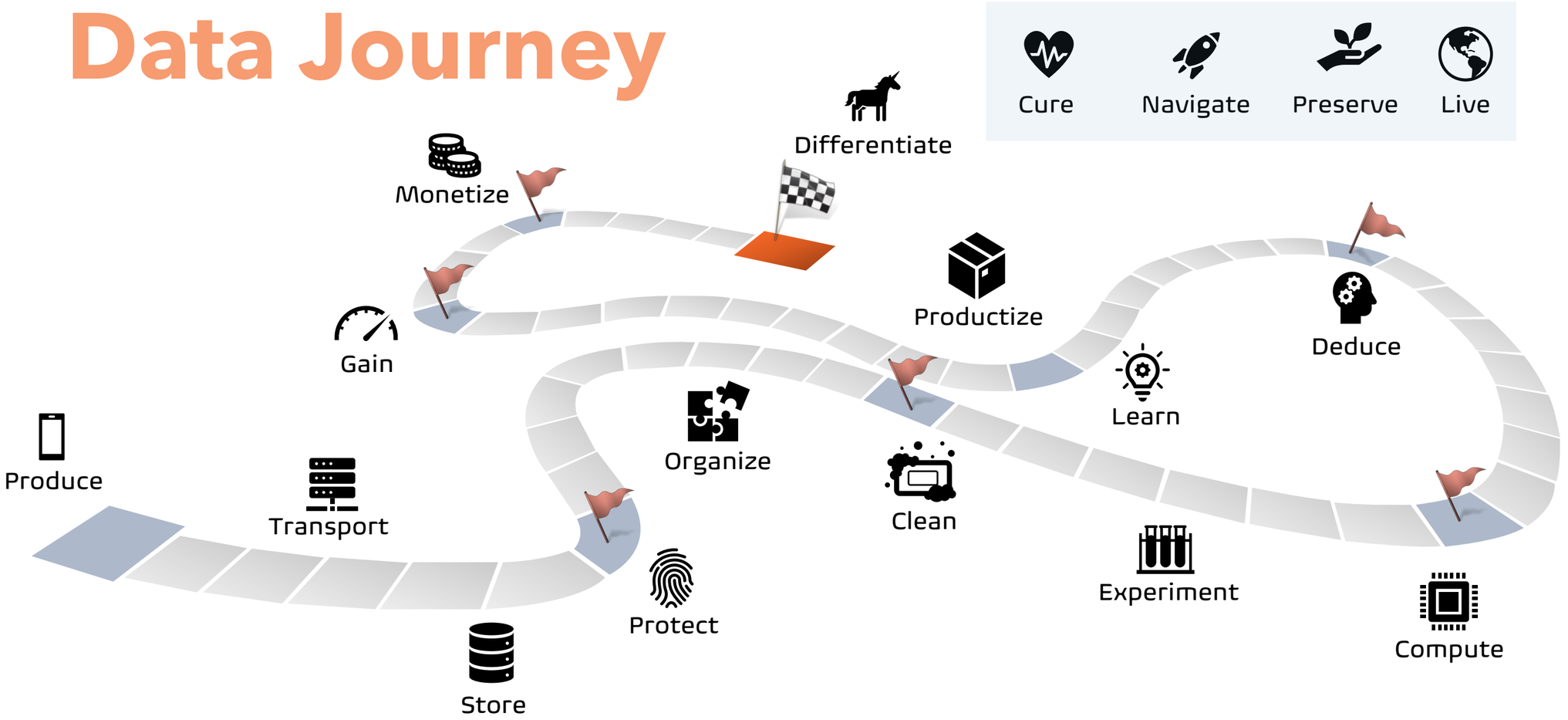


- **Unstructured Data**
- **Schema on Read**
- **Data Pipelines: Extract-Load-Transform**
- **Processing Model: Streaming**

Data Lifecycle[dl1]

[dl1] Curtesy of Herminio Velazquez

Data Journey



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

UNIVERSITY OF TARTU
1632

A 30000ft view

- Data Collection
- Data Wrangling
  - Data Shaping
  - Data Cleansing
  - Data Augmenting
- Data Analysis

Data Collection

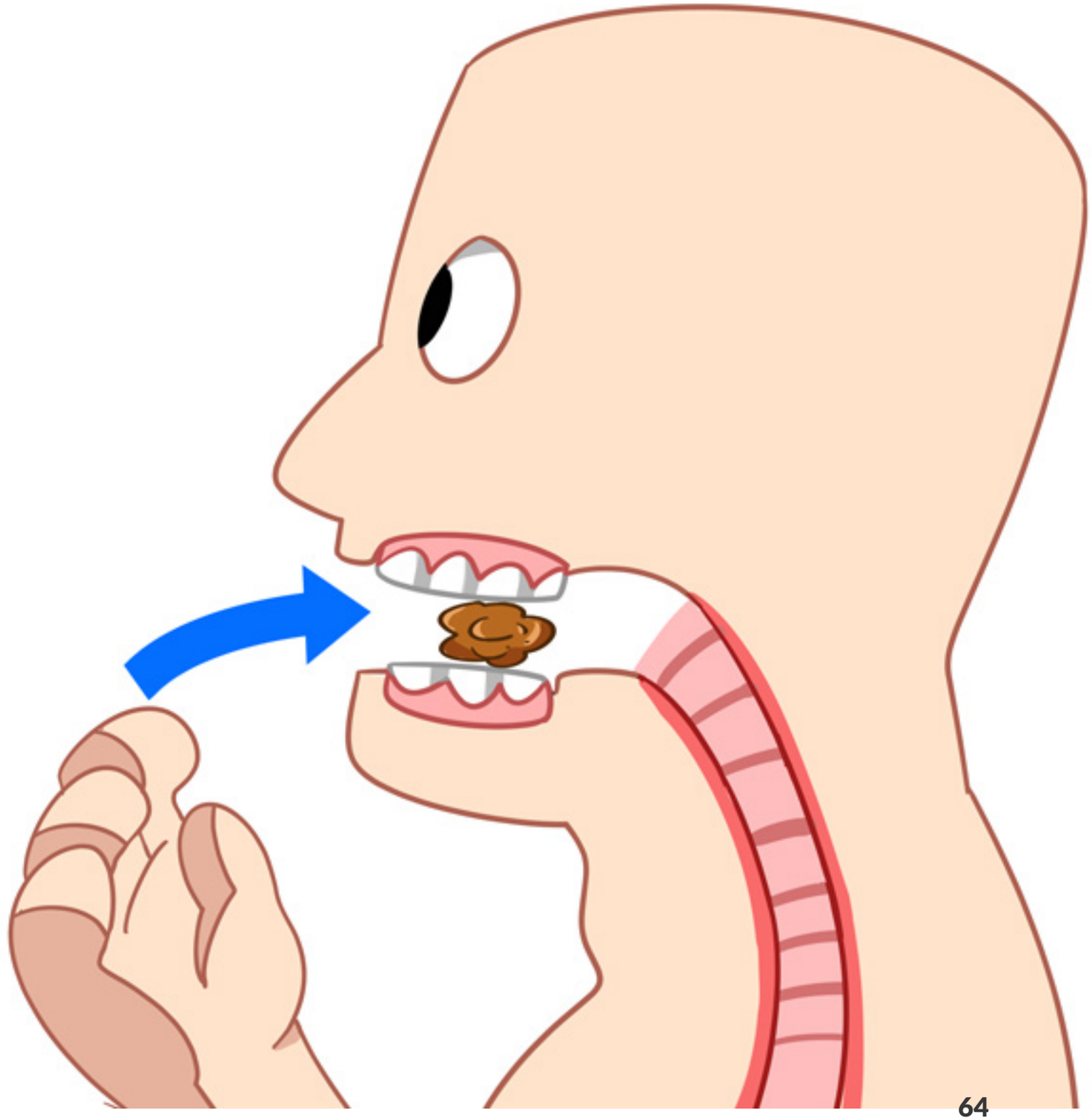
- We call Data Collection (aka Acquisition) the process of finding and accessing new data sources
- Happens **outside** the data warehouse/ lake and may involve different organizations
- Not to be confused with synonimns/ Data Ingestion, which is the process of filling our Data Warehouse/Lake with new data

Data Collection Examples

- Reading Files
- Data Crawling
- Accessing Databases
- Calling Web API
- Consuming WebSockets

Not to be confused with Data Ingestion

- Maintaining a Distributed File System
- Using a Distributed Message Queue
- Using a Publishing Subscribe System



Data Source Selection Criteria

- Credibility
- Completeness
- Accurateness
- Verifiability
- Currency
- Accessibility
- Compliance
- Cost
- Legal issues
- Security
- Storage
- Provenance

Data Wrangling

The process of transforming “raw” data into data that can be analyzed to generate valid actionable insights

Data scientists spend more time preparing data than analyzing them.

Isn't data science sexy?

Data Wrangling: an Iterative process

- Understand
- Explore
- Transform
- Augment
- Visualize

Data Processing

Pre

- Data Cleansing
- Data Integration
- Data Reduction

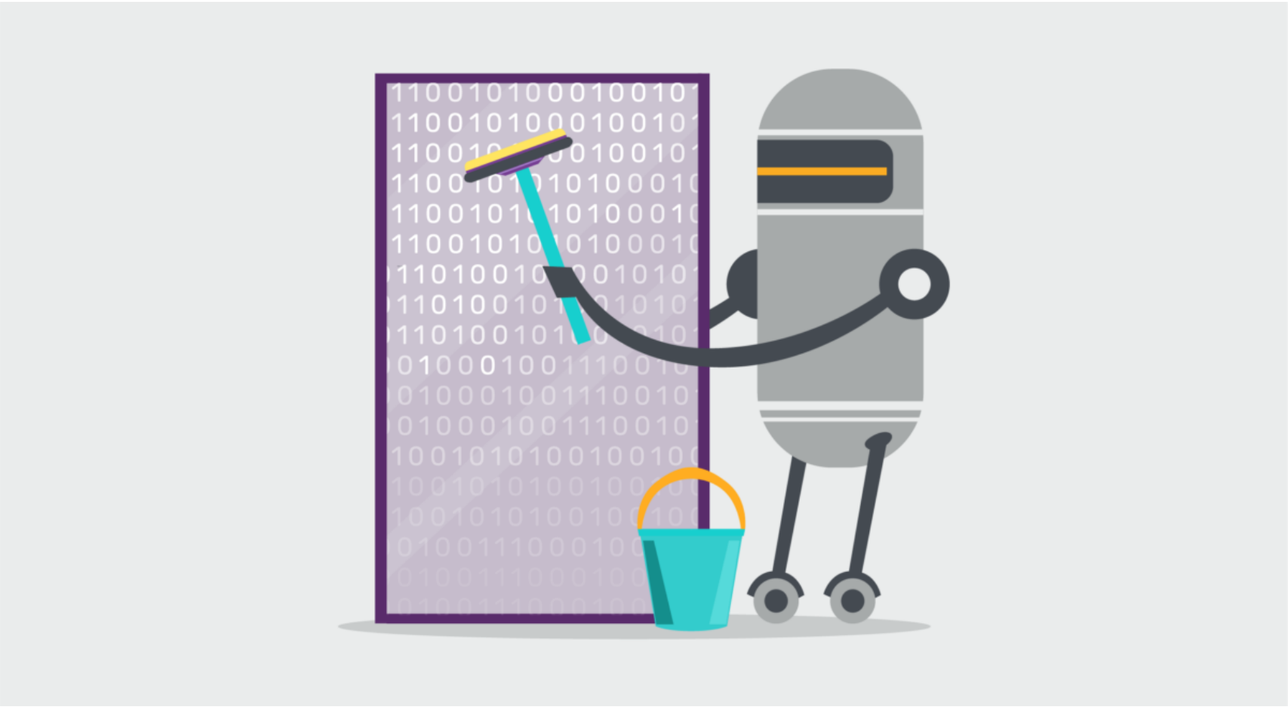
Post

- Data Tranformation
- Data Increase
- Data Visualization

Data Cleansing

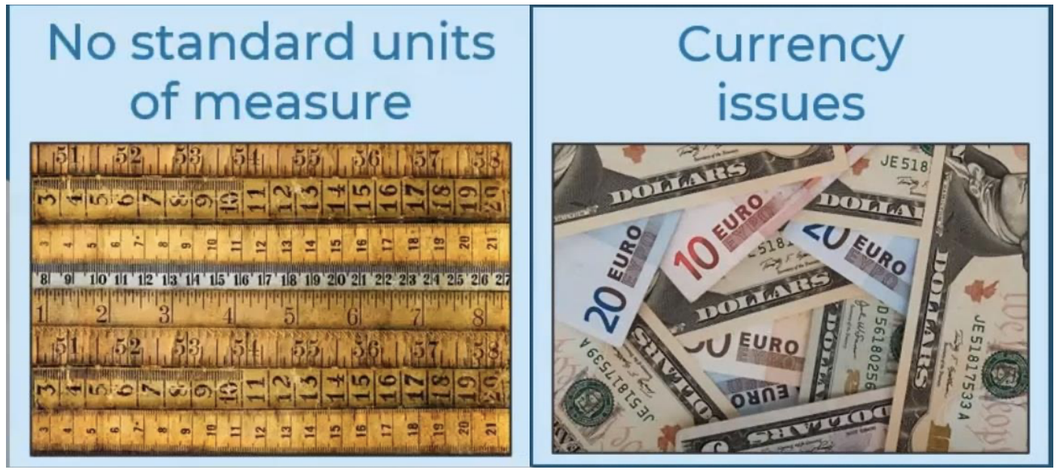
Data cleansing or ** data scrubbing** is the act of **detecting and correcting (or removing corrupt or inaccurate records)** from a dataset.

The term refers to identifying incomplete, incorrect, inaccurate, partial or irrelevant parts of the data and then replacing, modifying, filling in or deleting this dirty data.



Why is Data “Dirty” ?

- Dummy Values



Why is Data “Dirty” ?

- Dummy Values
- Absence of Data

SSN	Name (First, Initial, Last)	Address
334600443	Lisa Boardman	144 Wars St.
334600443	Lisa Brown	144 Ward St.
525520001	Ramon Bonilla	38 Ward St.
525250001	Raymond Bonilla	38 Ward St.
0	Diana D. Ambrosion	40 Brik Church Av.
0	Diana A. Dambrosion	40 Brick Church Av.
0	Colette Johnen	600 113th St. apt.5a5
0	John Colette	600 113th St. ap. 585
850982319	Ivette A Keegan	23 Florida Av.
950982319	Yvette A Kegan	23 Florida St.

Why is Data “Dirty” ?

- Dummy Values
- Absence of Data
- Cryptic Data

id	first_name	last_name	email	gender	value
1	Judy	Boness	jboness0@issuu.com	Non-binary	52
2	Nealy	Skepper	nskepper1@apache.org	Non-binary	35
3	Lanita	Hellewell	lhellewell2@usa.gov	Agender	20
4	Oswald	Gartside	ogartside3@fema.gov	Agender	2
5	Clemens	Polo	cpolo4@pagesperso-orange.fr	Polygender	31
6	Bliss	Smellie	bsmellie5@loc.gov	Non-binary	84
7	Isador	Phelps	iphelps6@msn.com	Genderqueer	26
8	Roseline	Stitwell	rstitwell7@tinyurl.com	Polygender	58
9	Elie	O'Hartigan	eohartigan8@spiegel.de	Genderqueer	60

Why is Data “Dirty” ?

- Dummy Values
- Absence of Data
- Cryptic Data
- Contradicting Data

id	first_name	last_name	email	age	birth date
1	Augustus	Blasi	ablasi0@ed.gov	15	10/23/2021
2	Ardath	Johnson	ajohnson1@clickbank.net	85	8/14/2021
3	Rowen	Whorlton	rwhorlton2@microsoft.com	33	1/1/2021
4	Bernete	Ashpital	bashpital3@businesswire.com	10	11/23/2021
5	Carny	Brixey	cbrixey4@aboutads.info	80	8/13/2021
6	Jacques	Colliber	jcolliber5@unblog.fr	3	7/16/2021
7	Enrica	Boldison	eboldison6@smh.com.au	25	11/12/2021
8	Dru	Risbridge	drisbridge7@weather.com	39	4/28/2021
9	Channa	Abramsky	cabramsky8@fema.gov	20	8/10/2021
10	Adelbert	Clorley	aclorley9@spiegel.de	58	3/5/2021
11	Eifine	Wagstaffe	fwagstaffea@homestead.com	5	6/28/2021

Why is Data “Dirty” ?

- Dummy Values
- Absence of Data
- Cryptic Data
- Contradicting Data
- Shared Field Usage
- Inappropriate Use of Fields
- Violation of Business Rules
- Non-Unique Identifiers

Data Integration

Data integration systems aims at providing a uniform access to a set of heterogeneous data sources.

Data integration systems aims at bridging the heterogeneity between the sources and produce a uniform query interface.

Why data needs to be integrated?

Data sources can differ on the data model (relational, hierarchical, semi-structured), on the schema level, or on the query-processing capabilities.

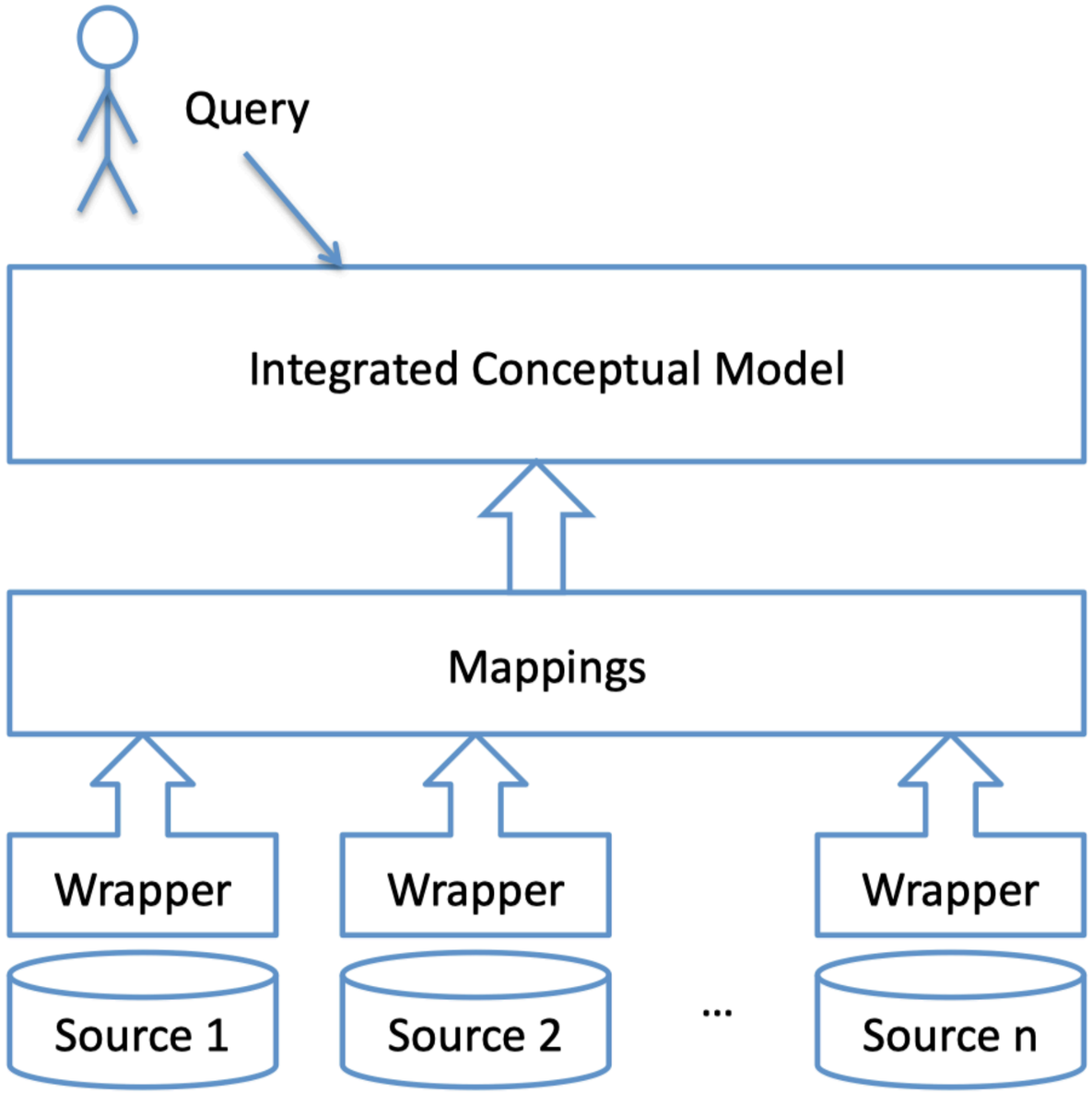
In a data integration architecture, these sources are queried by using a global schema, also called mediated schema, which provides a virtual view of the underlying sources.

Data integration strategies

Mappings are functions that link each source schema to the ICM.

They are also expressed using formal languages, and implemented into software (wrappers), which contains the data access logic.

Two main strategies:
- Global As View (GAV): The ICM is expressed in terms for the data source schemas
- Local As View (LAV): The ICM is expressed independently for the data source schemas



Data Transformation

Data Transformation refers to the process changing the shape of the data, typically for migrating data from one data system to another

Data transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.

Why we need Data Transformation

Different analytical tasks (e.g., queries) result either simpler to write or more efficient to run once executed within highly specialized systems.

Data transformation is often used to reshape data into forms that are more convenient.

Strategies for Data Transformation

- Batch
- Streaming

Data Reduction

Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

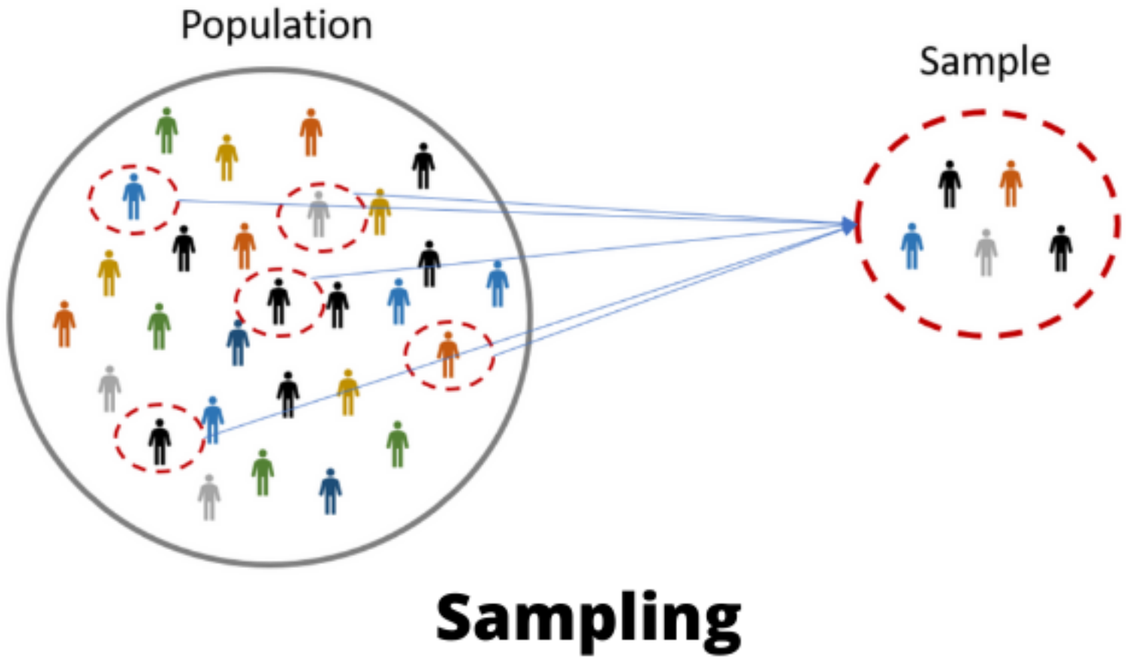
Why data reduction?

A database/data warehouse may store terabytes of data.

- Datasets do not fit in memory
- Complex data analysis may take a very long time to run
- Data can be overwhelming to interpret

Data reduction strategies (1/2)

- Numerosity reduction (some simply call it: Data Reduction)
  - Regression
  - Histograms, clustering, sampling
  - Data cube aggregation



HotCubator

Data reduction strategies (2/2)

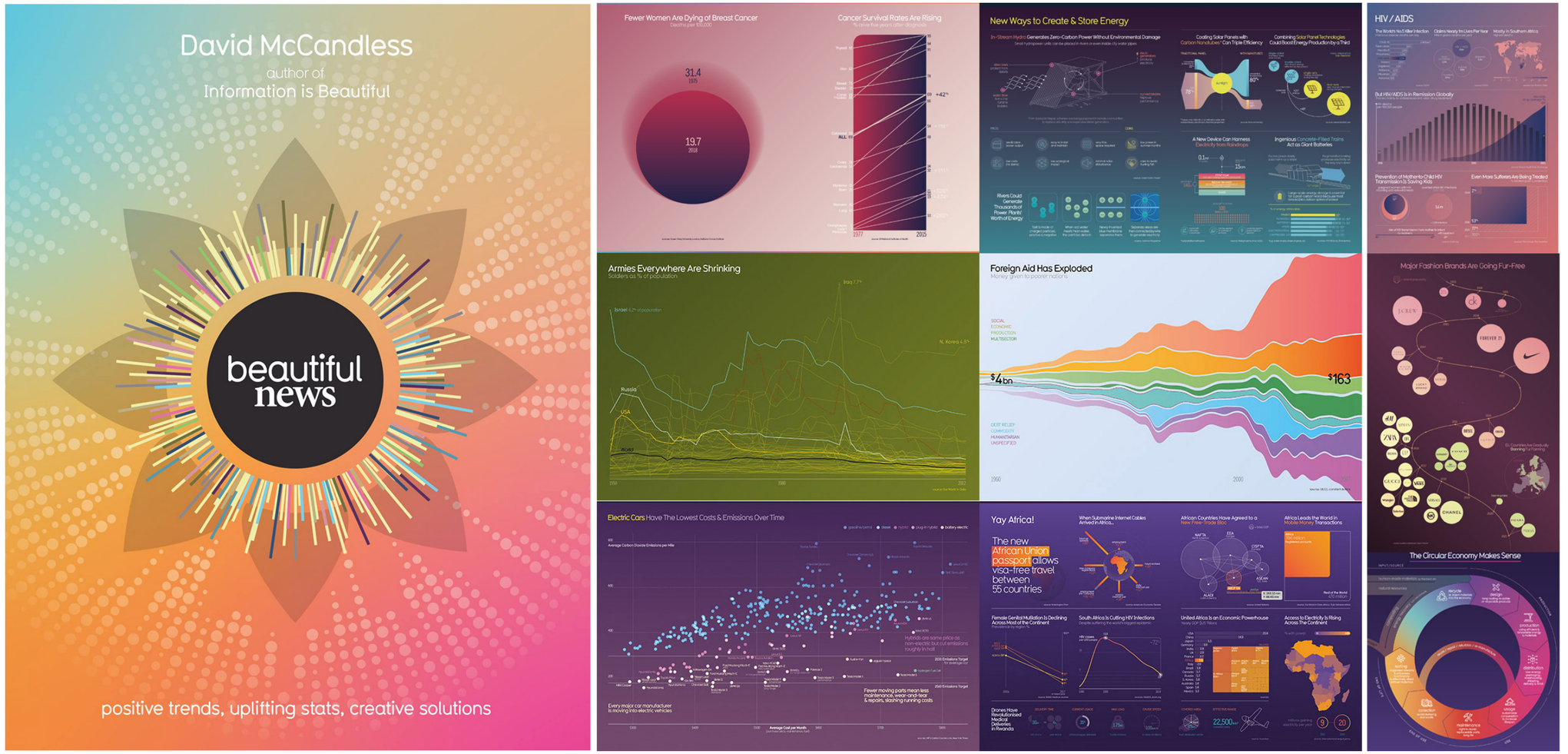
- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation

Data Summarization and Visualization

Data Summarization refers to the process of aggregating data to reach an higher level of abstraction.

In practice, it is a form of data reduction, which requires pre-analysis.

*Data Summarization** is tightly coupled with **Data Visualization**, i.e., the process of representing data in a form that is more easy to understand for the data scientist.


David McCandless
author of
Information is Beautiful
beautiful news
positive trends, uplifting stats, creative solutions
Fewer Women Are Dying of Breast Cancer
Deaths per 100,000
31.4
1975
19.7
2018
Cancer Survival Rates Are Rising
% who live five years after diagnosis
+42%
1977
2015
New Ways to Create & Store Energy
In-Stream Hydro Generates Zero-Carbon Power Without Environmental Damage
Coating Solar Panels with Carbon NanoTubes Can Triple Efficiency
Combining Solar Panel Technologies Could Boost Energy Production by a Third
A New Device Can Harness Electricity from Raindrops
Ingenious Concrete-Filled Trains Act as Giant Batteries
Rivers Could Generate Thousands of Power Plants' Worth of Energy
HIV / AIDS
But HIV/AIDS Is in Remission Globally
Prevention of Mother-to-Child HIV Transmission Is Saving Kids
Even More Sufferers Are Being Treated
Armies Everywhere Are Shrinking
Soldiers as % of population
Israel 6.2% of population
Iraq 7.7%
N. Korea 4.4%
Russia
USA
World
Foreign Aid Has Exploded
Money given to poorer nations
$4bn
$163
SOCIAL
ECONOMIC
PRODUCTION
MULTISECTOR
DEBT RELIEF
COMMODITY
HUMANITARIAN
UNSPECIFIED
1960
2000
Major Fashion Brands Are Going Fur-Free
Electric Cars Have The Lowest Costs & Emissions Over Time
Average Carbon Dioxide Emissions per Mile
Average Cost per Month
Hybrids are same price as non-electric but cut emissions roughly in half
2020 Emissions Target
for average car
2040 Emissions Target
Fewer moving parts mean less maintenance, wear-and-tear & repairs, slashing running costs
Every major car manufacturer is moving into electric vehicles
Yay Africa!
The new African Union passport allows visa-free travel between 55 countries
When Submarine Internet Cables Arrived in Africa...
African Countries Have Agreed to a New Free Trade Bloc
Africa Leads the World in Mobile Money Transactions
Female Genital Mutilation Is Declining Across Most of the Continent
South Africa Is Cutting HIV Infections
United Africa Is an Economic Powerhouse
Yearly GDP ($US Trillions)
Access to Electricity Is Rising Across The Continent
Drones Have Revolutionised Medical Deliveries in Rwanda
22,500
The Circular Economy Makes Sense

source

Data Increase (as opposite to Reduction)

Why we need increase the amount of data?

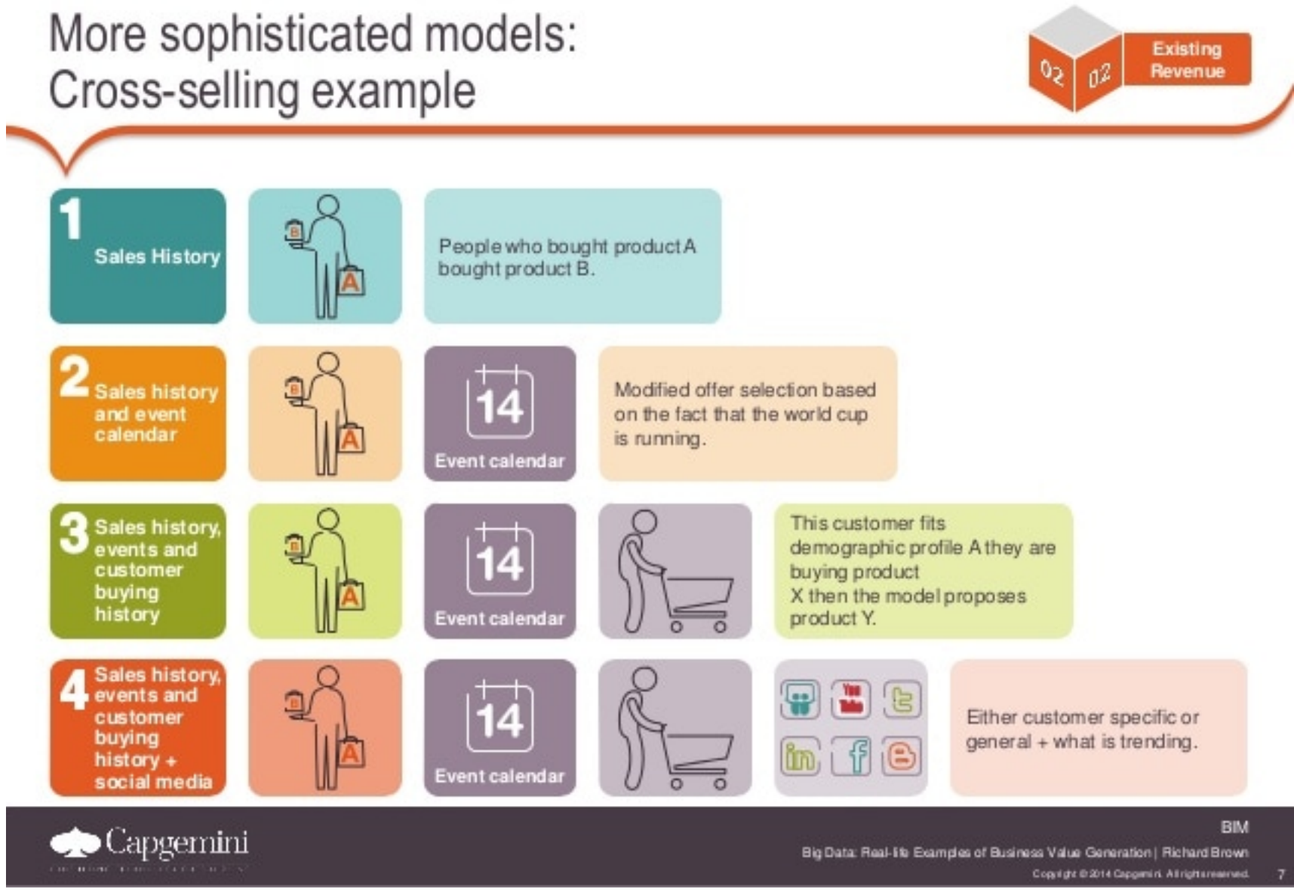
Data analytics tasks are often data or knowledge intensive.

Data intensive analytical tasks require to increase the amount of available data, e.g., for boosting the model training

Knowledge intensive tasks require contextual domain knowledge to increase the accuracy of the result.

Data Enrichment

Data enrichment refers to the process of appending or otherwise enhancing collected data with relevant context obtained from additional sources.



Data Agumentation

Data agumentation in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data.

Strategies for Data Transformation

- Batch
- Streaming

Moving Data (AKA Implement the Lifecycle)

Data Pipeline

A **data pipeline** is a set of data processing elements connected in series, where the output of one element is the input of the next one.

Data pipelines are used to shape, organises, and moves data to a destination for storage, insights, and analysis.

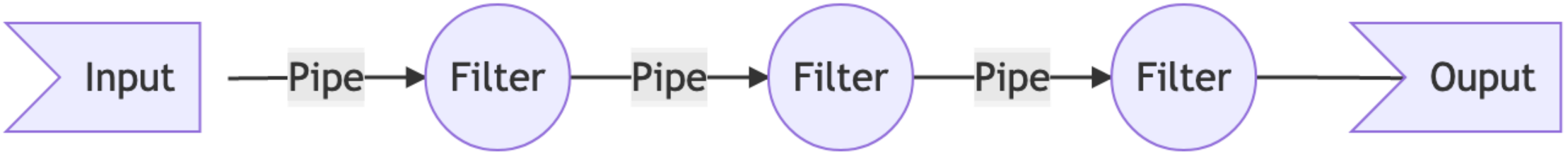
Modern data pipeline generalise the notion of ETL (extract, transform, load) to include data ingestion, integration, and movement across any cloud architecture and add additional layers of resiliency against failure.

Data pipelines Then

The concept of a pipeline began from the good old Unix "Pipe" symbol (|).

The output of one "process" (on left side of the pipe) to be given as "input" to another process (which was on the right side of the pipe).

Elements of a Data Pipeline



Pipes are connectors which send data from one component (filter) to another.

Filters do actual data "processing" (transformation/cleansing/scrubbing/munging... whatever)

Input or **Source** is the actual data source (database output/text file/SQL resultset/raw text)

Output or **Sink** is the final output at the end of this chain.

The elements of a pipeline are often executed in parallel or in time-sliced fashion; in that case, some amount of buffer **storage** is often inserted between elements.

Basic Operations of Data Pipelines

- access information of different data sources
- extract discrete data elements of the source data
- copy extracted data from a data source to another
- transform data
  - correct errors in data elements extracted from source data
  - standardize data in data elements based on field type
- join or merge (in a rule-driven way) with other data sources

Big Data Pipeline[1]

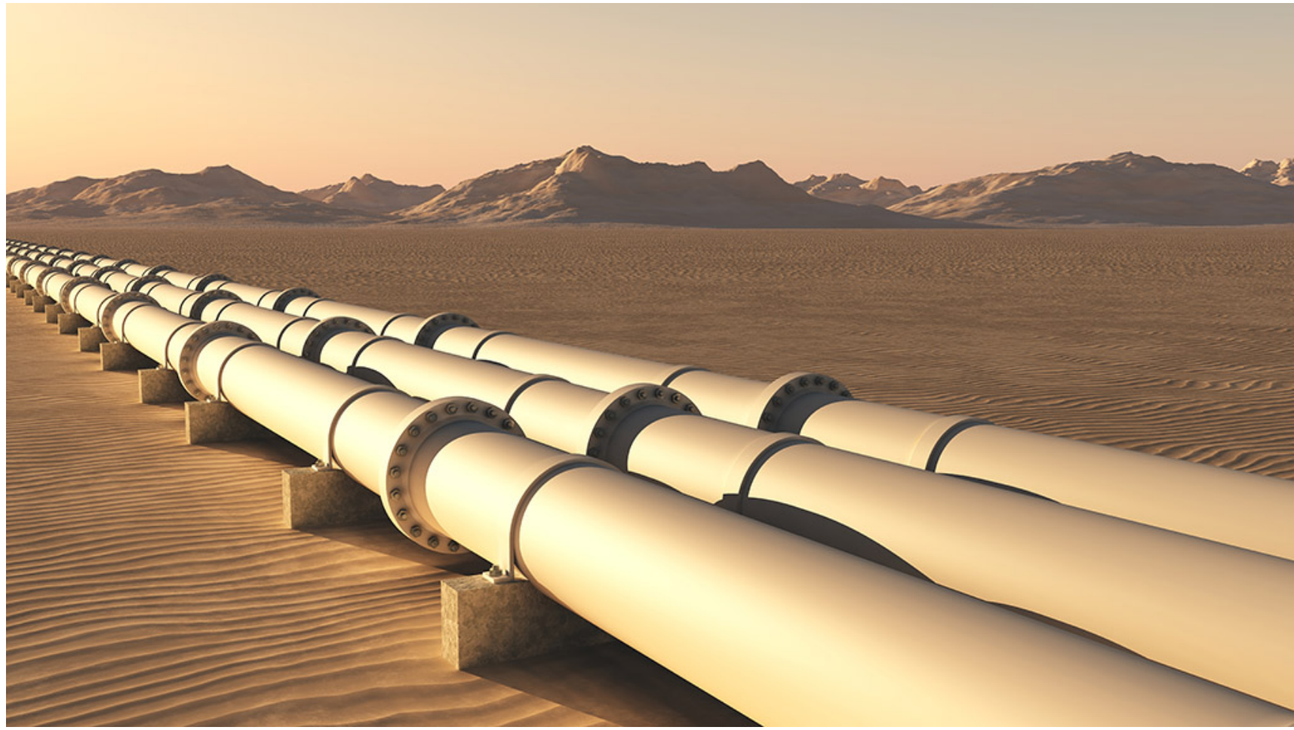
The big data world brings additional challenges, i.e., volume, variety and velocity, which forced a paradigm shift in data architectures.

[1] henceforth referred as data pipelines only

Concepts

Typically represented as Direct (Acyclic) Graphs (D(A)G)

Data pipelines open up the possibility of creating "workflows" which can help reuse, modularize and componentize data flows.



Data pipeline components

- **Storage and Ingestion Layers**, e.g., HDFS
  - Storage layers nowadays typically support polyglot persistence.
  - Message Buses that help move chunks of data (sometimes at blazing speeds) from one system to another, e.g., Kafka, RabbitMQ, Kinesis
  - Serialization Frameworks, e.g., Protocol Buffers, Avro,
- **(Event Stream) Processing frameworks**, e.g., Kafka Streams, Flink
  - JVM Based
  - SQL Based
- **Workflow Management Tools** that supervise the processes which run inside your data pipelines, e.g., Airflow, Luigi, Dagster
  - "orchestrating" systems
  - "choreographing" systems
- **Query layer**, e.g., NoSQL Datamarts
- **Analytics layer** (out of scope)

Implementing Pipelines[011]

- **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.
- Workflow Managers (API)
- Drag & Drop...

[011] we are focusing on ETL

Skill Set: SQL mastery[03]

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 modeling[03]

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) ETL[012]

- Unless you need to process over many petabytes of data, or you are 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 specialised 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 optimising core data infrastructure,
- building and maintaining custom ingestion pipelines,
- supporting data team resources with design and performance optimisation, and
- building non-SQL transformation pipelines.

[013] TristanHandy, 2019