Big Data & AI in Business

How to become a data driven business

Sessions 10&11: Data Strategy & AI Operations

David **G Pisano**

The **AI strategy** aims to obtain the maximum potential of this technology for our businesses by **solving all the barriers** to its implementation.

1 Business Value

Usage: What are the main drivers of AI value in each business?
AI ambition: How much do we believe AI can transform our business?
AI maturity: How innovative in the use of AI does the company need to be?

2 Data Access

Types of sources: What data will be needed at the company to develop AI?
Data Platform: What technological tools will we use to facilitate access to data at the company?
Data Governance: How are we going to ensure that they are of the right quality?

3 AI Operations

Organization: Where should we locate the AI specialists at the company?
Sourcing Model: Who develops and maintains the AI models?
Process industrialization: How efficient and scalable is AI development, deployment and maintenance?

4 AI Culture

Skills training & hiring: How can we train the company staff? How do we attract & retain talent?
Communication: how do we make it easier for everyone to understand what AI brings to the company?
Government: How do we guarantee the proper use of AI (security, privacy, ethics) at the company?

5

Are the legal/regulatory issues sufficiently resolved?

6

Is society ready to accept the transformation of the value proposition?

7

It may be legal, profitable and working - but is it ethical?

The **AI strategy** aims to obtain the maximum potential of this technology for our businesses by **solving all the barriers** to its implementation.

1 Business Value

Usage: What are the main drivers of AI value in each business?
AI ambition: How much do we believe AI can transform our business?
AI maturity: How innovative in the use of AI does the company need to be?

2 Data Access

Types of sources: What data will be needed at the company to develop AI?
Data Platform: What technological tools will we use to facilitate access to data at the company?
Data Governance: How are we going to ensure that they are of the right quality?

3 AI Operations

Organization: Where should we locate the AI specialists at the company?
Sourcing Model: Who develops and maintains the AI models?
Process industrialization: How efficient and scalable is AI development, deployment and maintenance?

4 AI Culture

Skills training & hiring: How can we train the company staff? How do we attract & retain talent?
Communication: how do we make it easier for everyone to understand what AI brings to the company?
Government: How do we guarantee the proper use of AI (security, privacy, ethics) at the company?

5

Are the legal/regulatory issues sufficiently resolved?

6

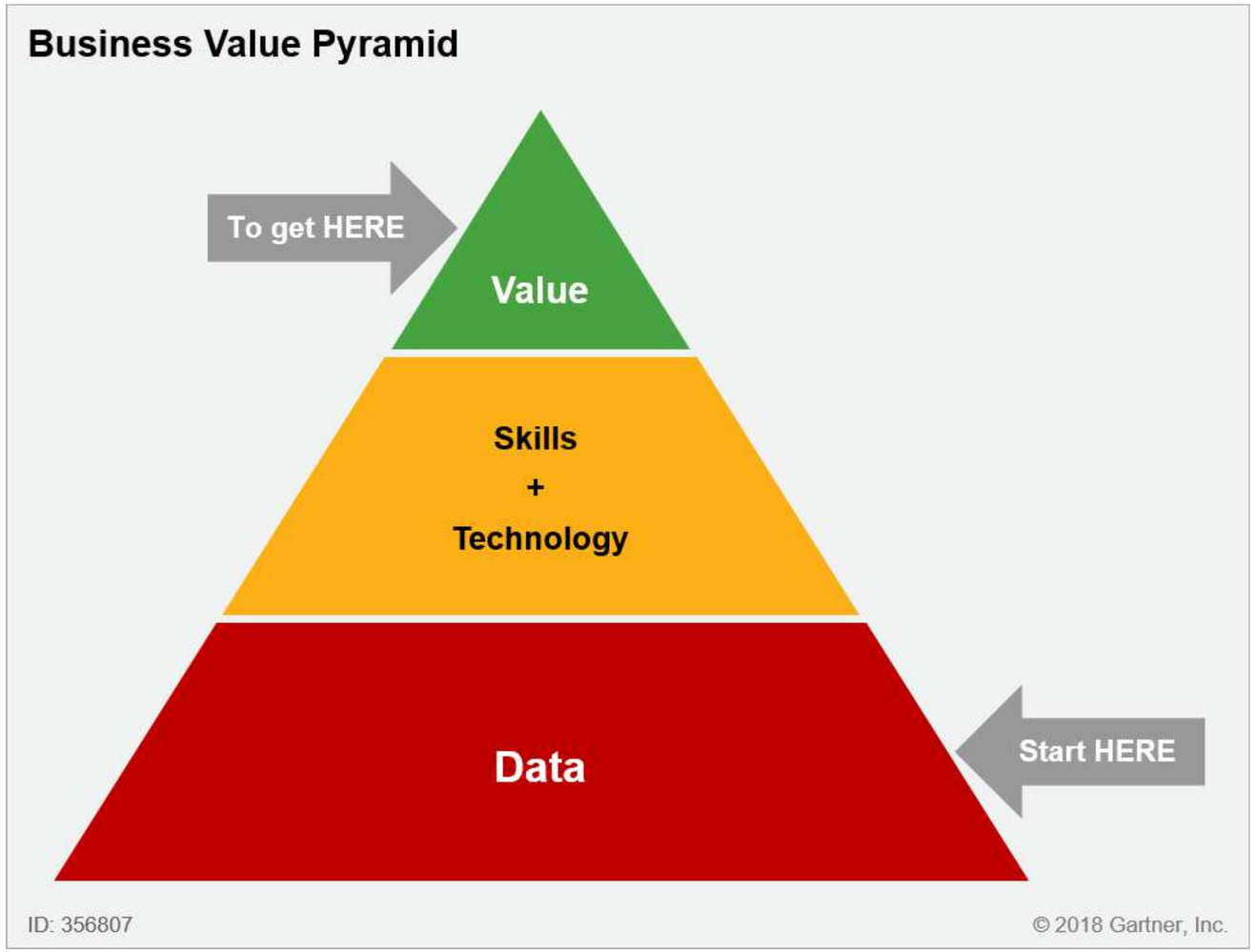
Is society ready to accept the transformation of the value proposition?

7

It may be legal, profitable and working - but is it ethical?


2nd Barrier:
Data

AI is the **evolution** of data through the various stages of analysis



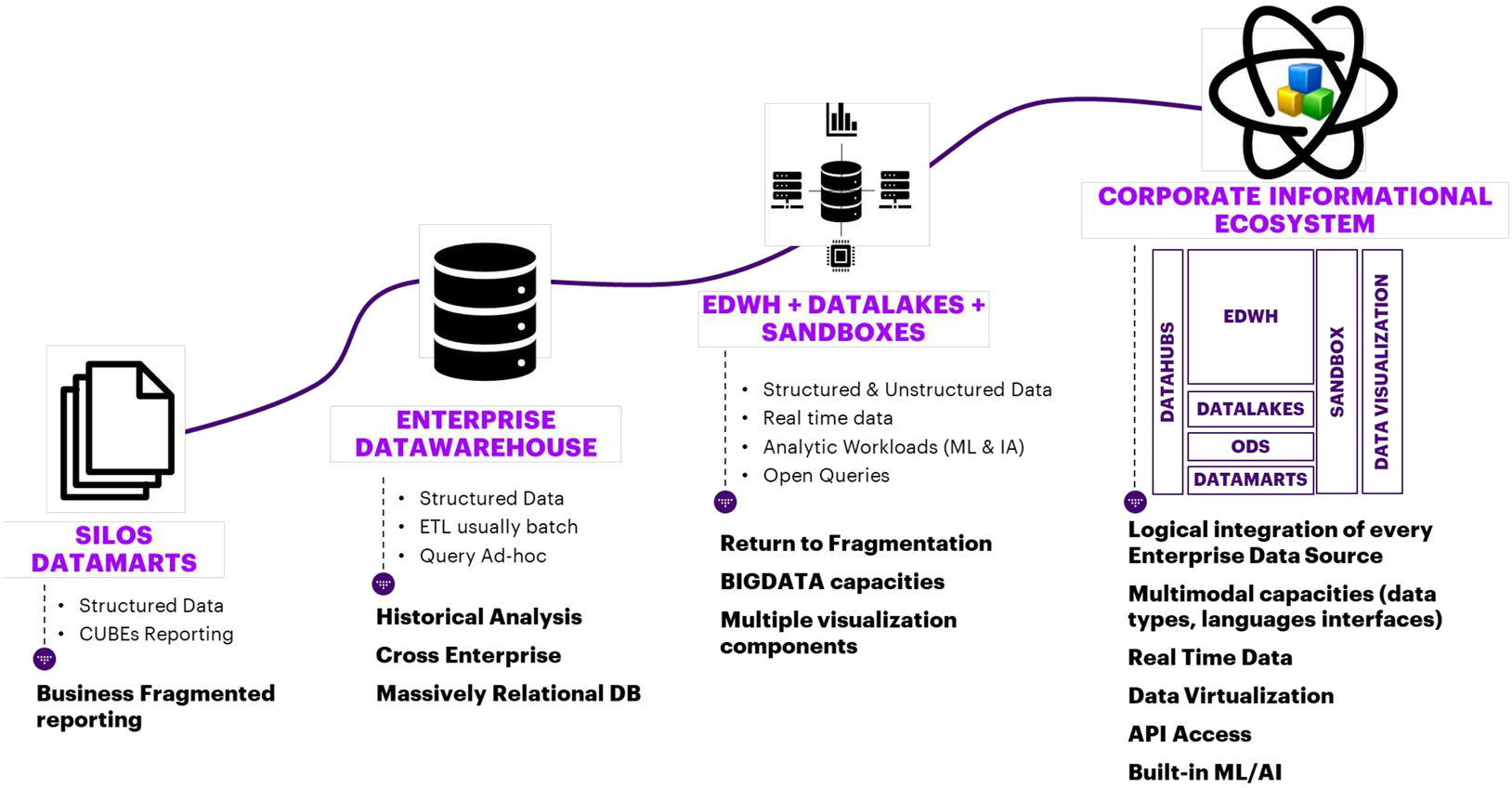
- Everyone talks about data, but no one emphasises the way **data needs to be managed systematically**, which allows **data to be used** by machine learning algorithms to **build intelligent systems** that **automate business processes**

- Think of **AI** as an **extension** of your **existing analytics** platform

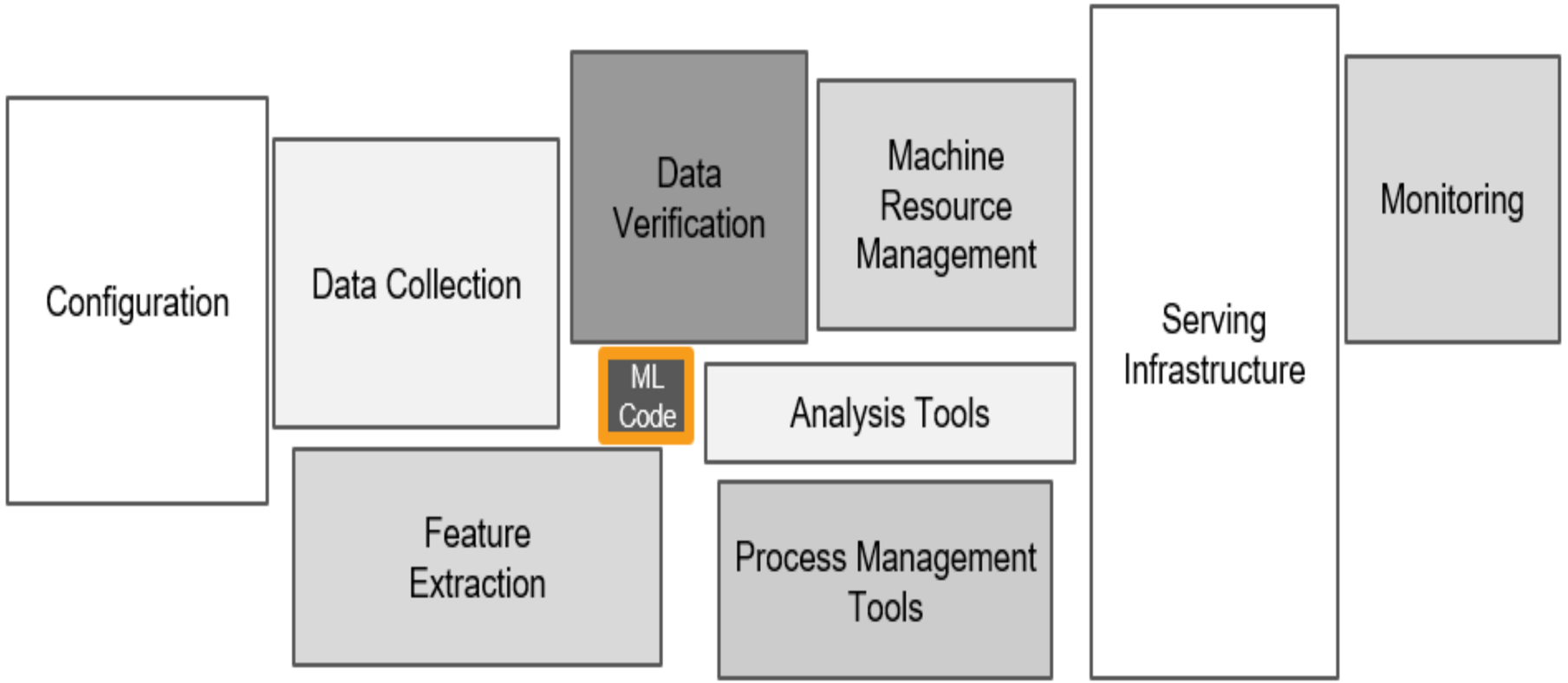
Data Access

For organizations that are more mature in the use of AI, data availability is the main problem to solve

O'Reilly 2020



Deployment of **AI/ML** is not (only) about the algorithms



A lot of the development work associated with AI/ML is not related to the AI/ML algorithms themselves. Rather, it is more concerned with the **preparation of data**, its **governance**, **performance** considerations and **delivery** of results according to service-level agreements (SLAs).

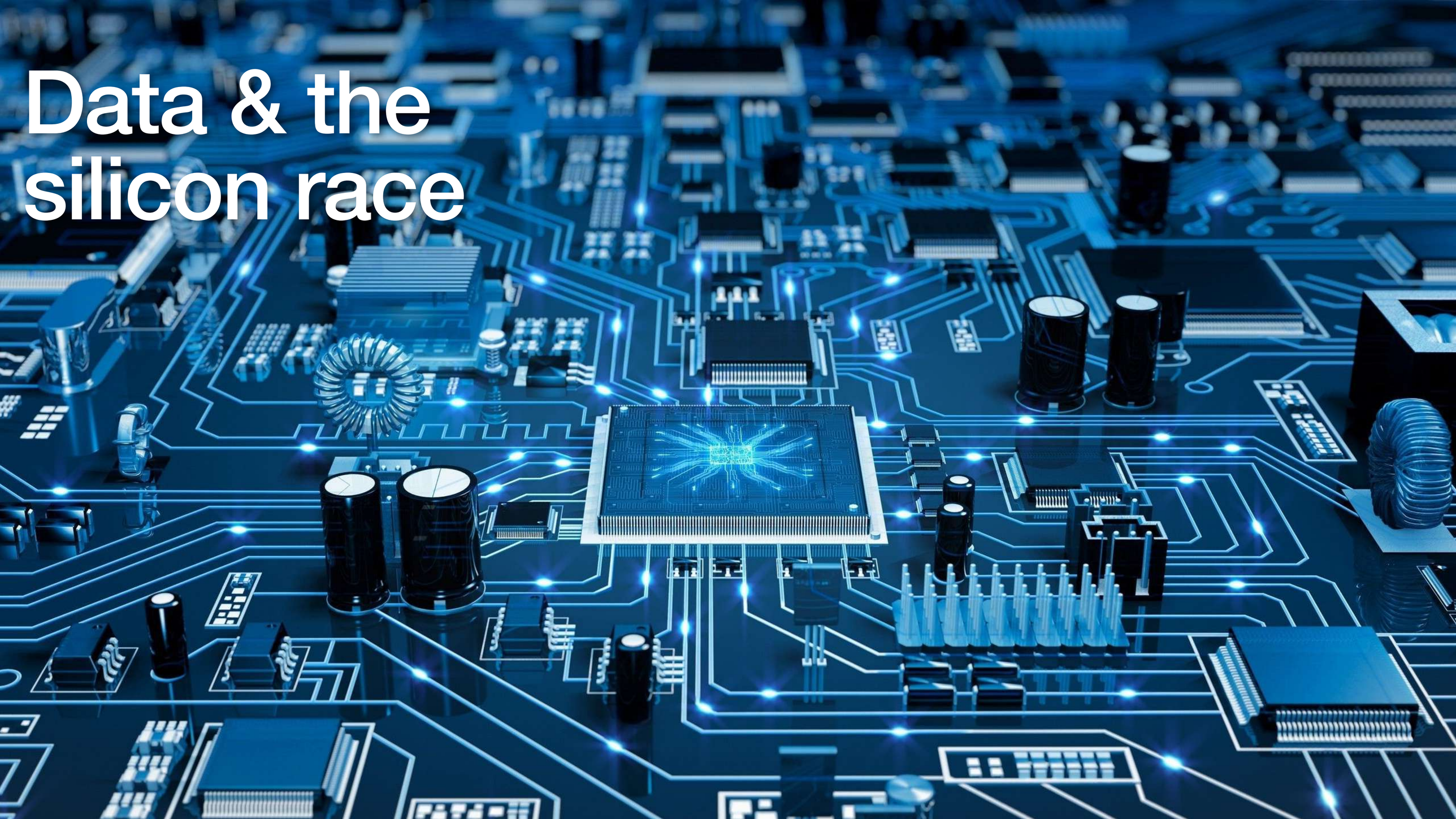
Data **Platform**

¿What is the data management platform (**DMP**) used for collecting and managing data? ¿It allows to use big data and artificial intelligence algorithms to process and analyze large data sets? ¿Can it grow and acquire new functionalities as the data grows and the business needs evolve?

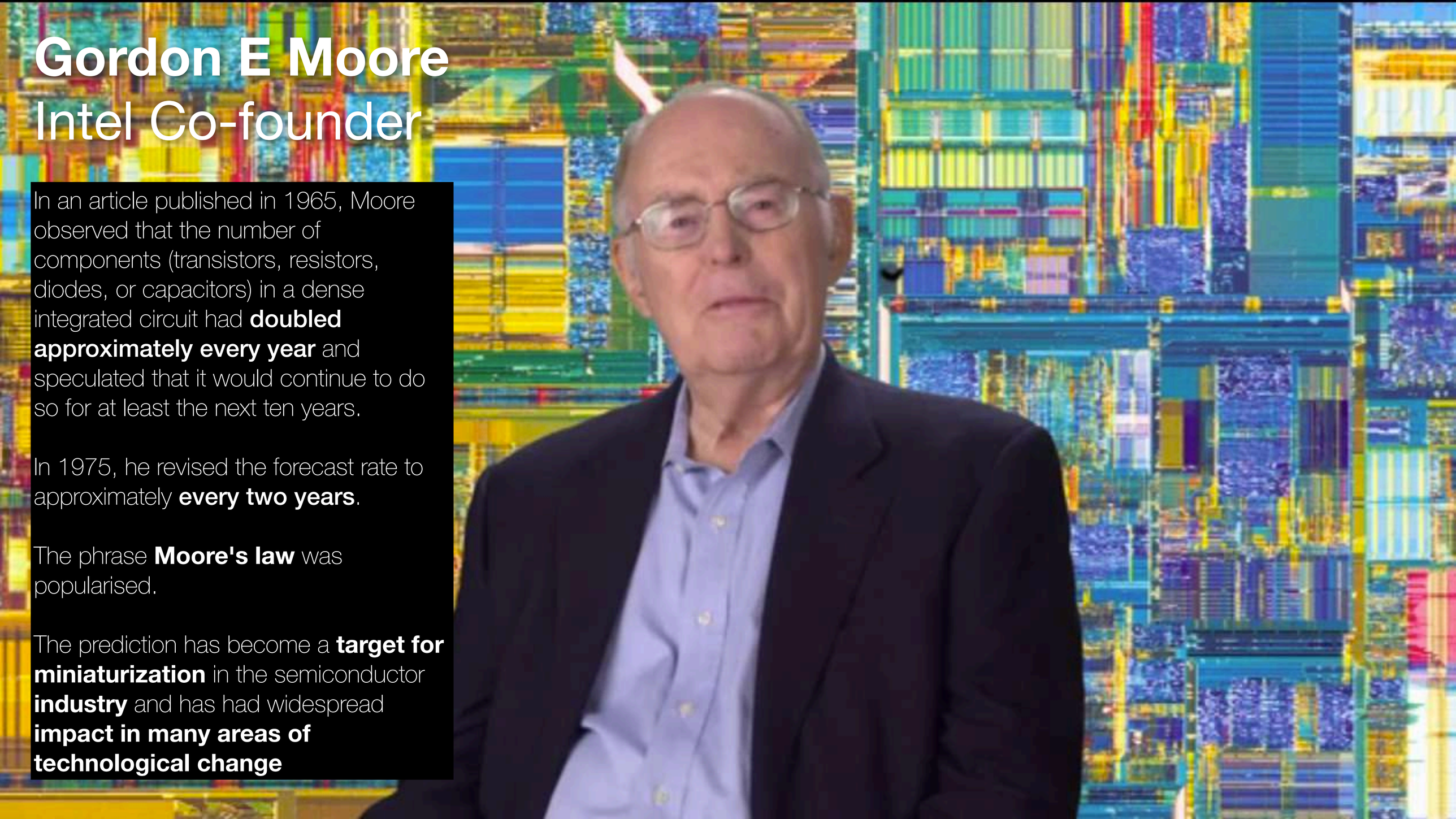
Data **Governance**

¿What are the processes and policies in place that ensure that high **data quality** exists through its complete lifecycle? ¿What data **controls** there are that support business objectives? ¿It is the data **available**, **usable**, **consistent** and **secure**?

Data & the silicon race



Gordon E Moore
Intel Co-founder



In an article published in 1965, Moore observed that the number of components (transistors, resistors, diodes, or capacitors) in a dense integrated circuit had **doubled approximately every year** and speculated that it would continue to do so for at least the next ten years.

In 1975, he revised the forecast rate to approximately **every two years**.

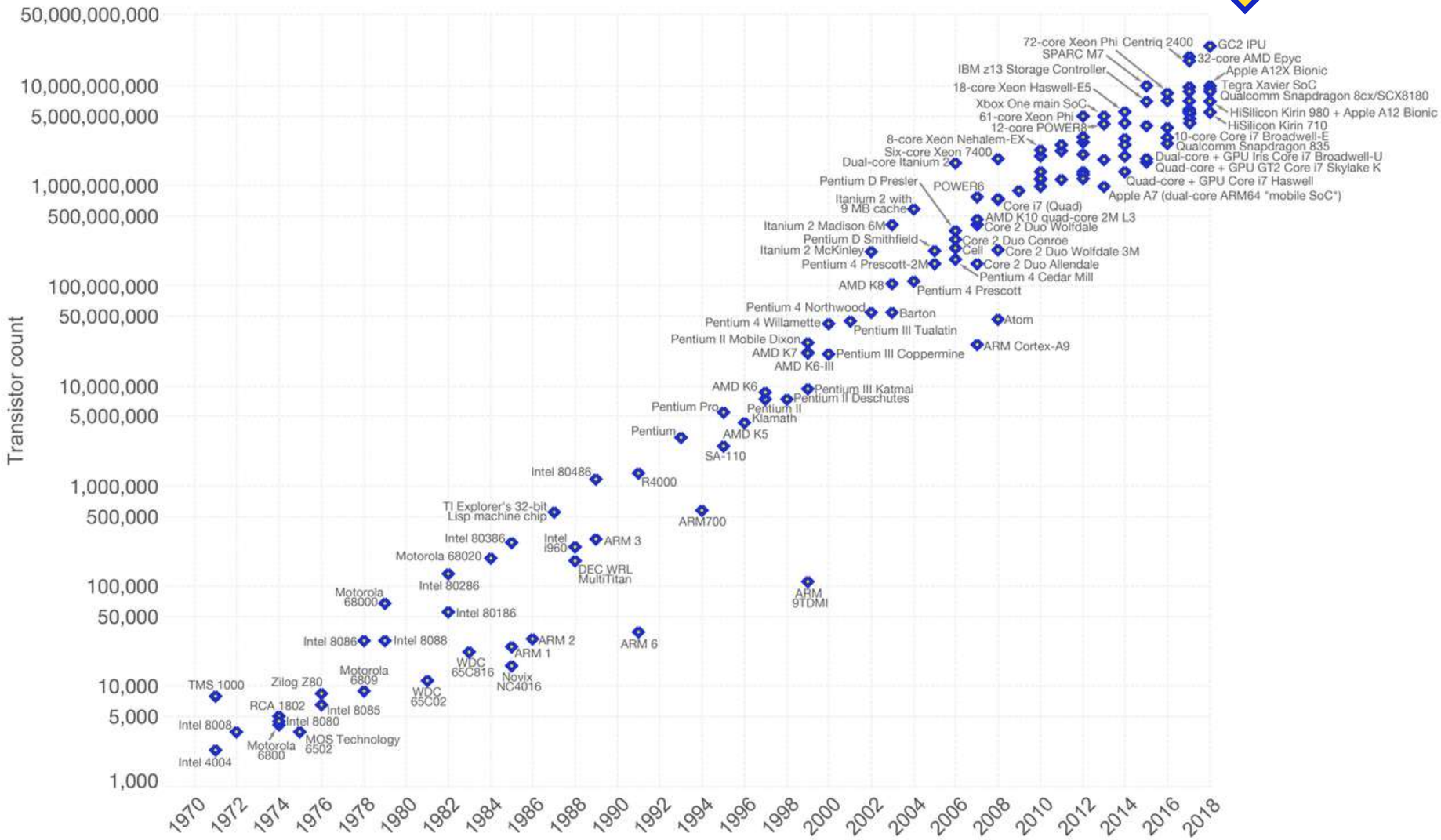
The phrase **Moore's law** was popularised.

The prediction has become a **target for miniaturization** in the semiconductor **industry** and has had widespread **impact in many areas of technological change**

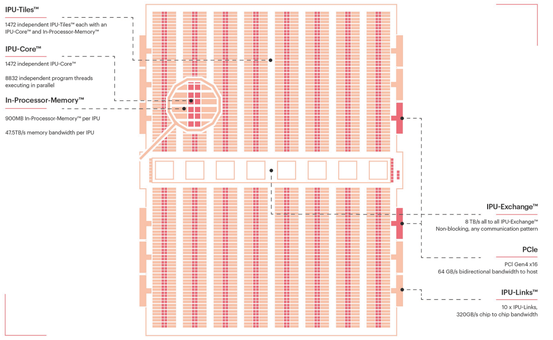
Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.

Our World in Data



Colossus™ MK2 GC200 IPU (2020)
59.4Bn transistors
1,472 cores
250 TFlops per chip



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

The **fifth** paradigm change: what will be the next one?

GENERATION OF COMPUTERS

1 GEN — FIRST-GENERATION COMPUTERS (1940-1956)
based on vacuum tubes.

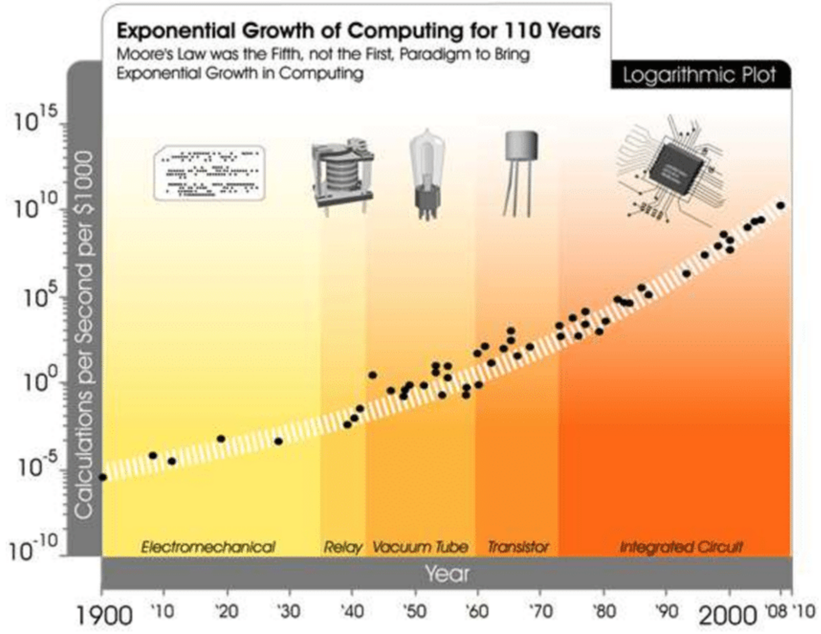
2 GEN — SECOND-GENERATION COMPUTERS (1956-1963)
transistors was used

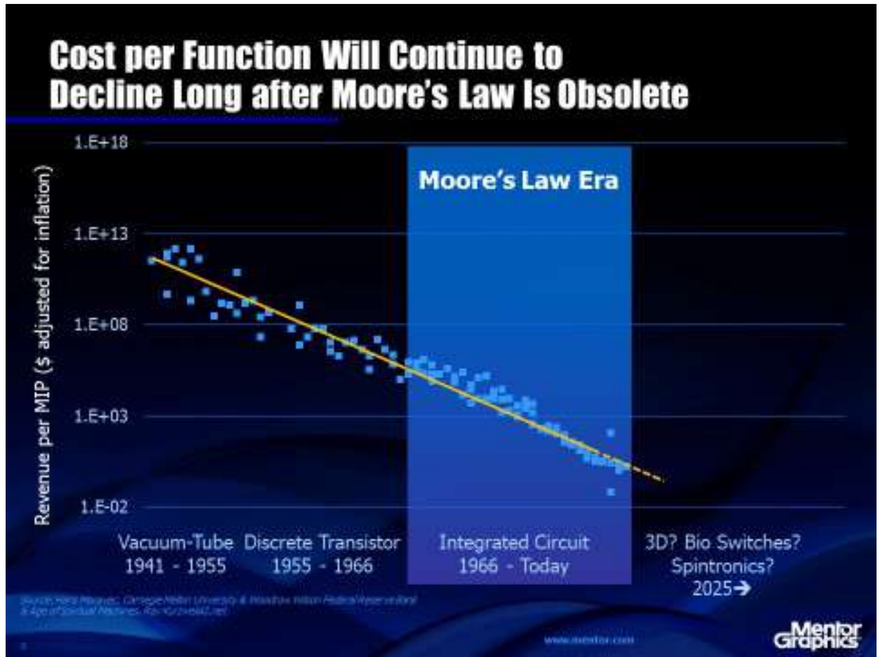
3 GEN — THIRD-GENERATION COMPUTERS (1964-1971)
Integrated Circuits (IC) was used

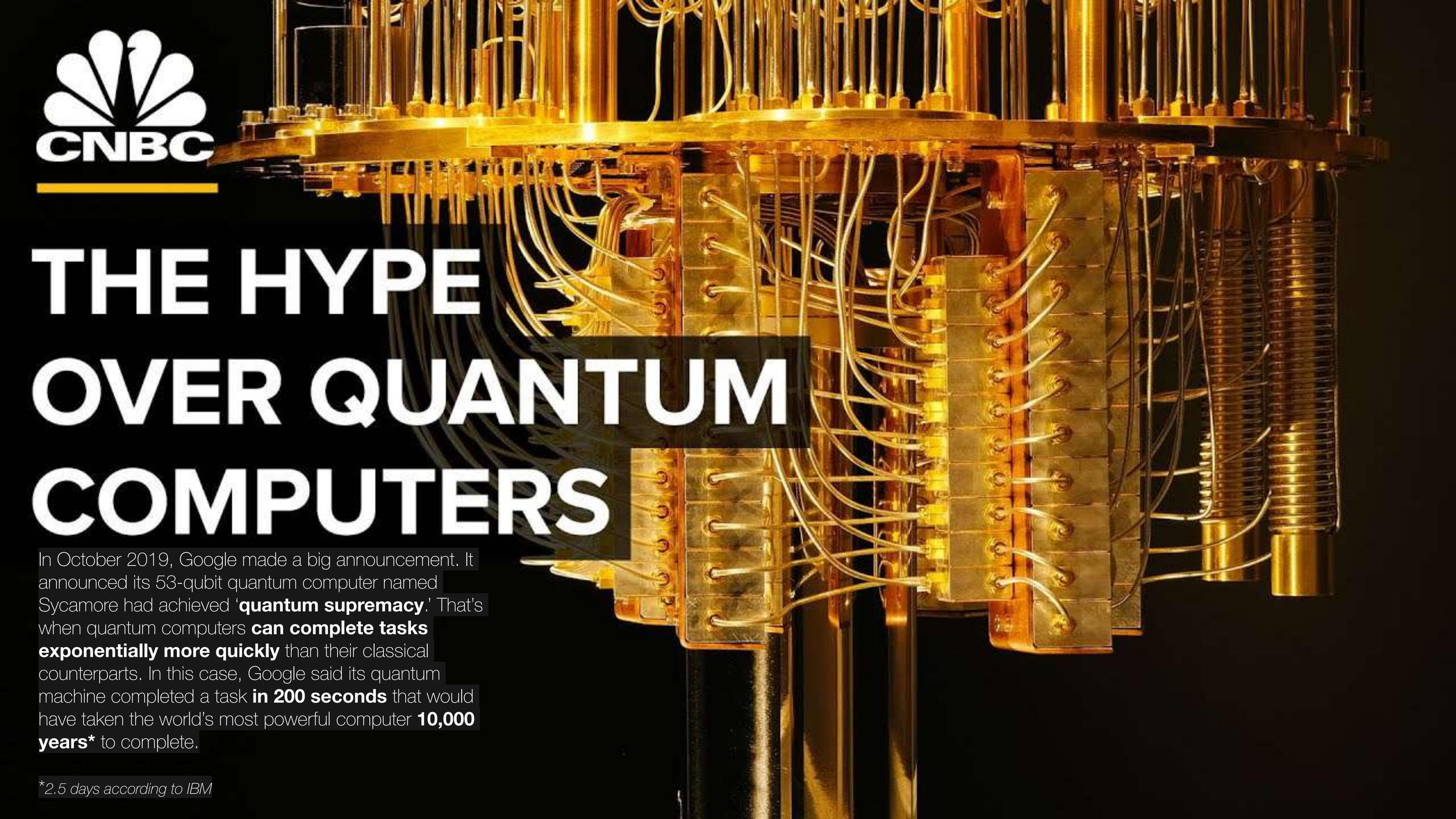
4 GEN — FOURTH GENERATION COMPUTERS (1971-1980)
Very Large Scale Integrated (VLSI) circuits was used

5 GEN — FIFTH-GENERATION COMPUTERS (1980-PRESENT)
ULSI (Ultra Large Scale Integration) technology was used

windowsground.com





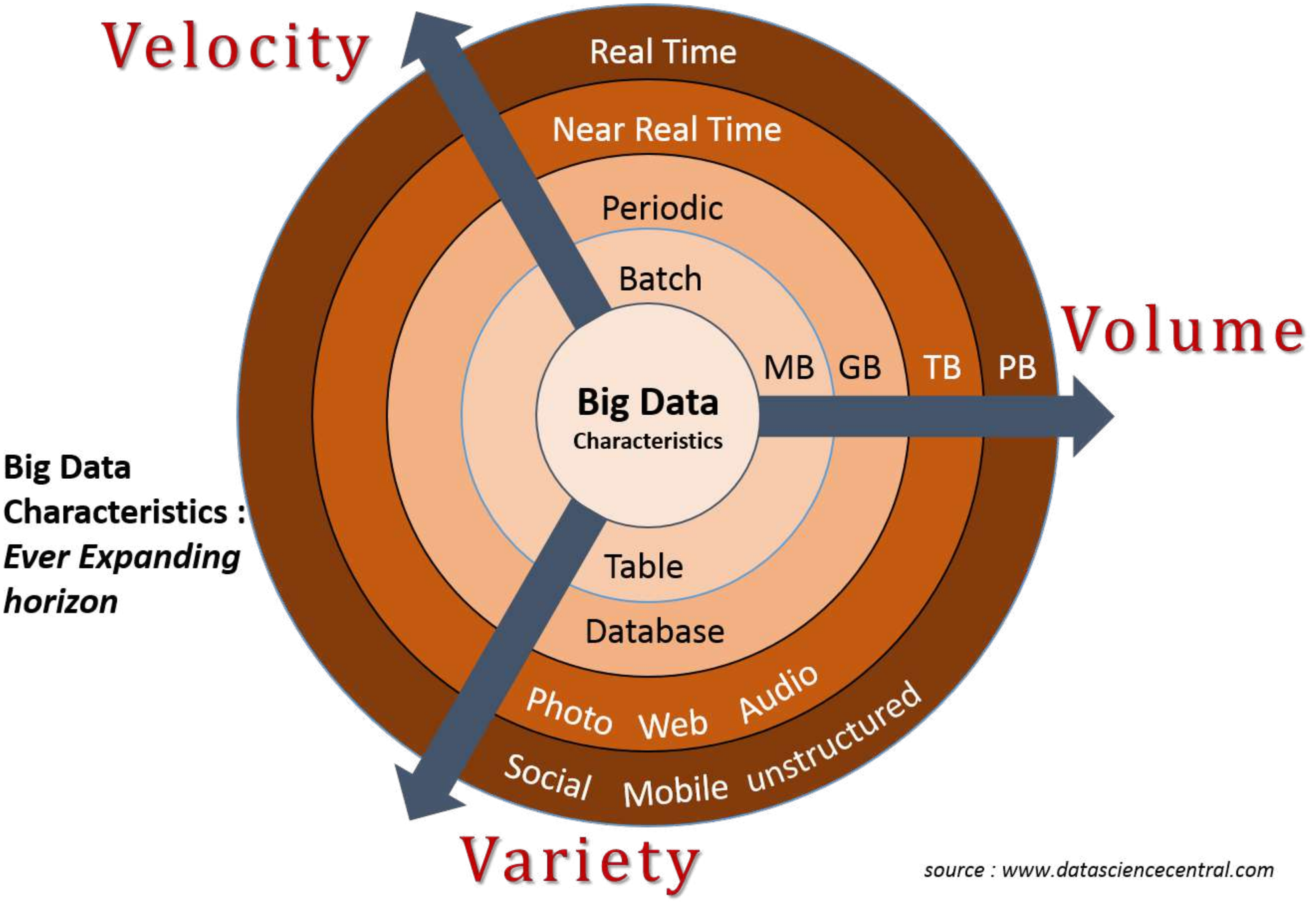


CNBC

THE HYPE OVER QUANTUM COMPUTERS

In October 2019, Google made a big announcement. It announced its 53-qubit quantum computer named Sycamore had achieved '**quantum supremacy**.' That's when quantum computers **can complete tasks exponentially more quickly** than their classical counterparts. In this case, Google said its quantum machine completed a task **in 200 seconds** that would have taken the world's most powerful computer **10,000 years*** to complete.

**2.5 days according to IBM*



Big Data Characteristics : ***Ever Expanding horizon***

source : www.datasciencecentral.com


Google
big data
Google Search
I'm Feeling Lucky

MapReduce: Big Data distribution

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean, Sanjay Ghemawat

OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA (2004), pp. 137-150

Motivation: Large Scale Data Processing

Many tasks: Process lots of data to produce other data

Want to use hundreds or thousands of CPUs

- ... but this needs to be easy

MapReduce provides:

- Automatic parallelization and distribution
- Fault-tolerance
- I/O scheduling
- Status and monitoring

Programming model

Input & Output: each a set of key/value pairs

Programmer specifies two functions:

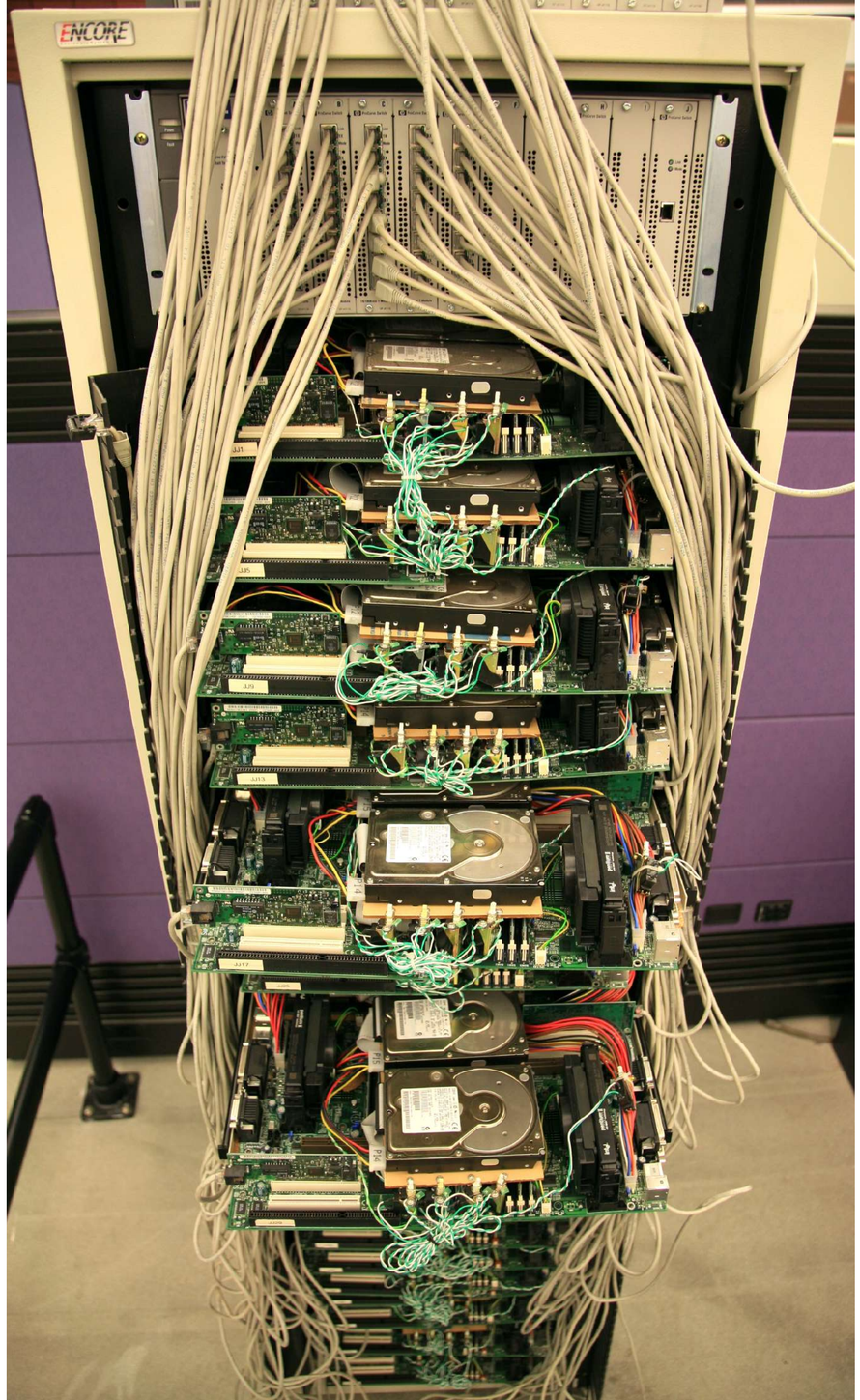
```
map (in_key, in_value) -> list(out_key, intermediate_value)
```

- Processes input key/value pair
- Produces set of intermediate pairs

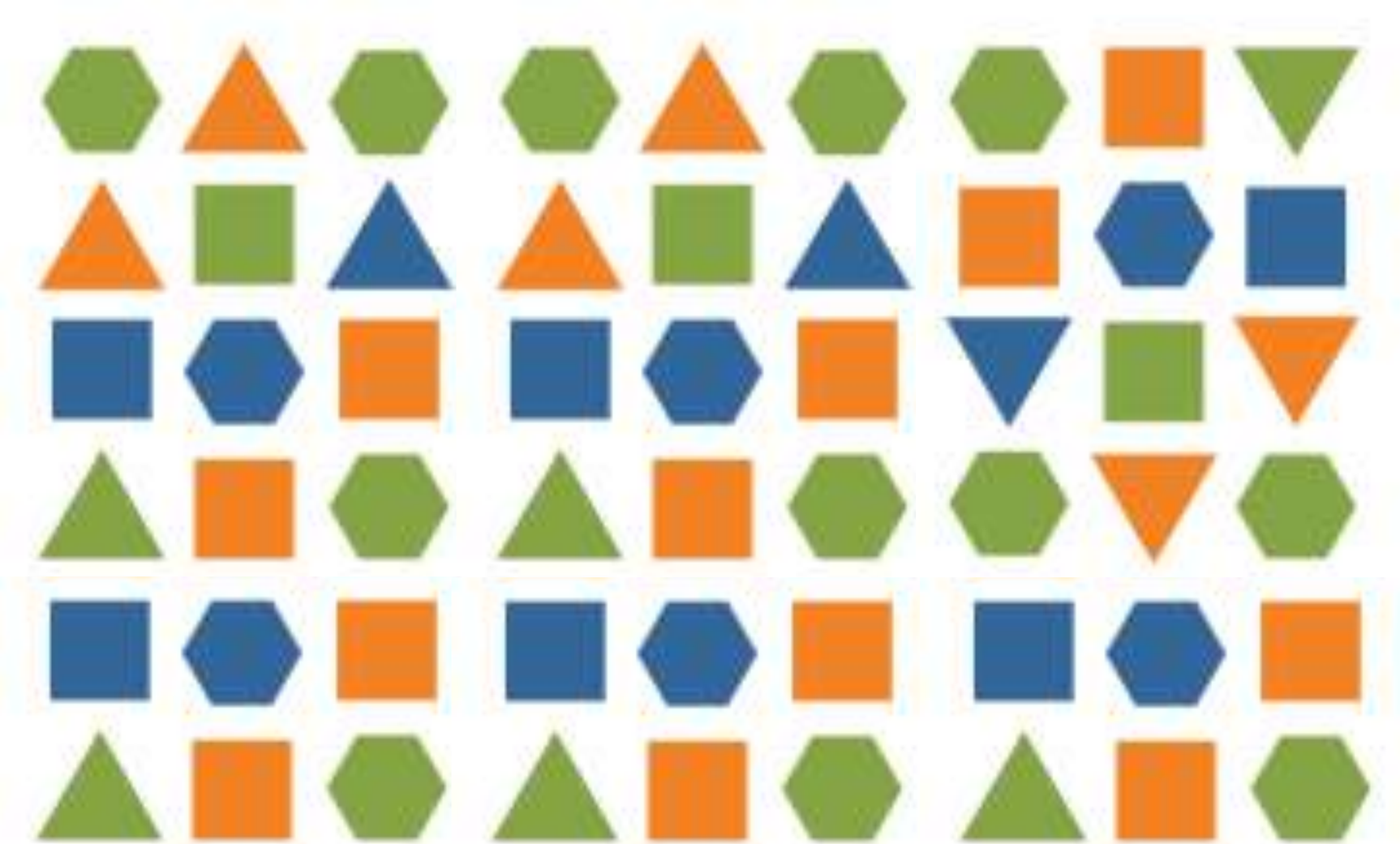
```
reduce (out_key, list(intermediate_value)) -> list(out_value)
```

- Combines all intermediate values for a particular key
- Produces a set of merged output values (usually just one)

Inspired by similar primitives in LISP and other languages



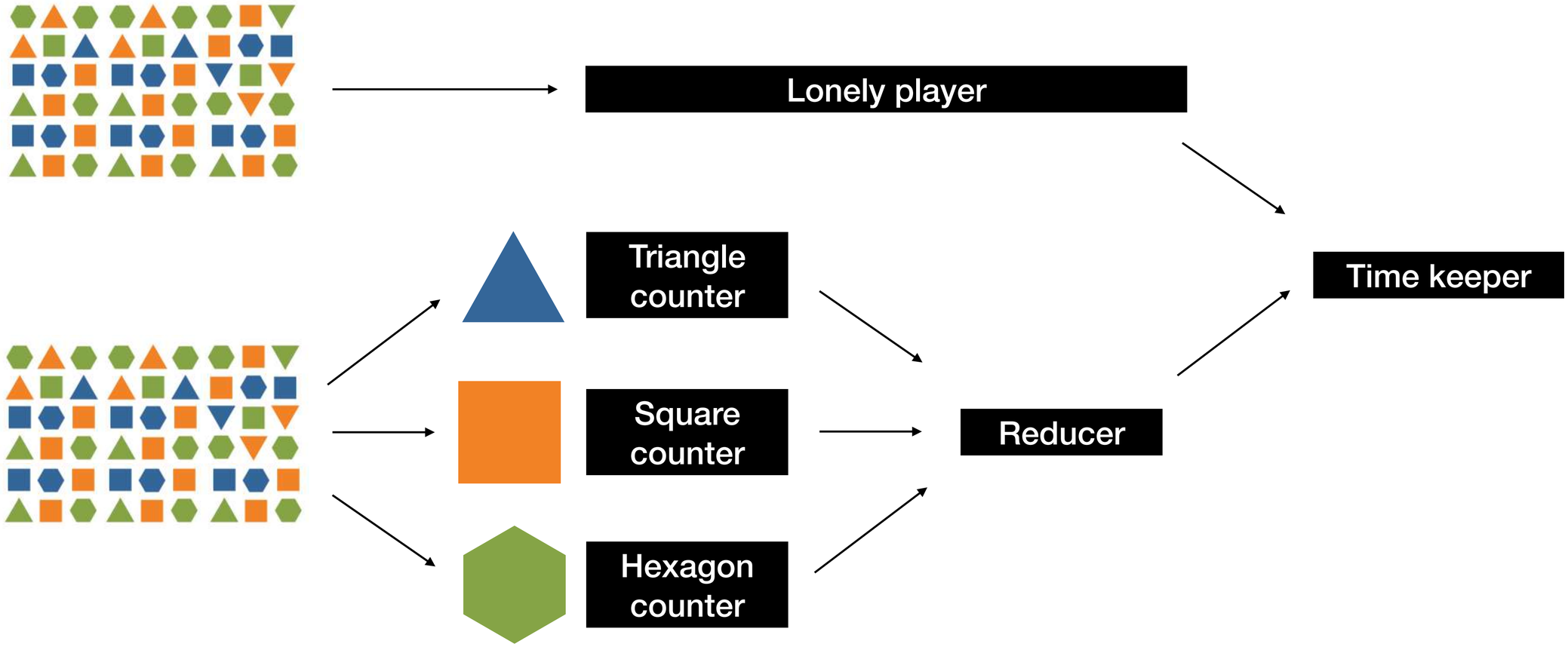
Map Reduce example

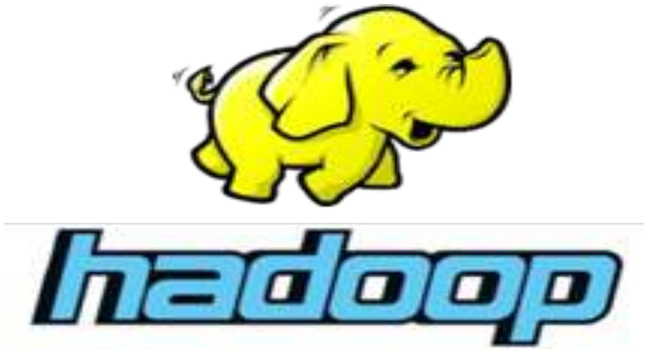


15 21 18

How to do it

Organize groups of 6 people
In each group assign the following roles:





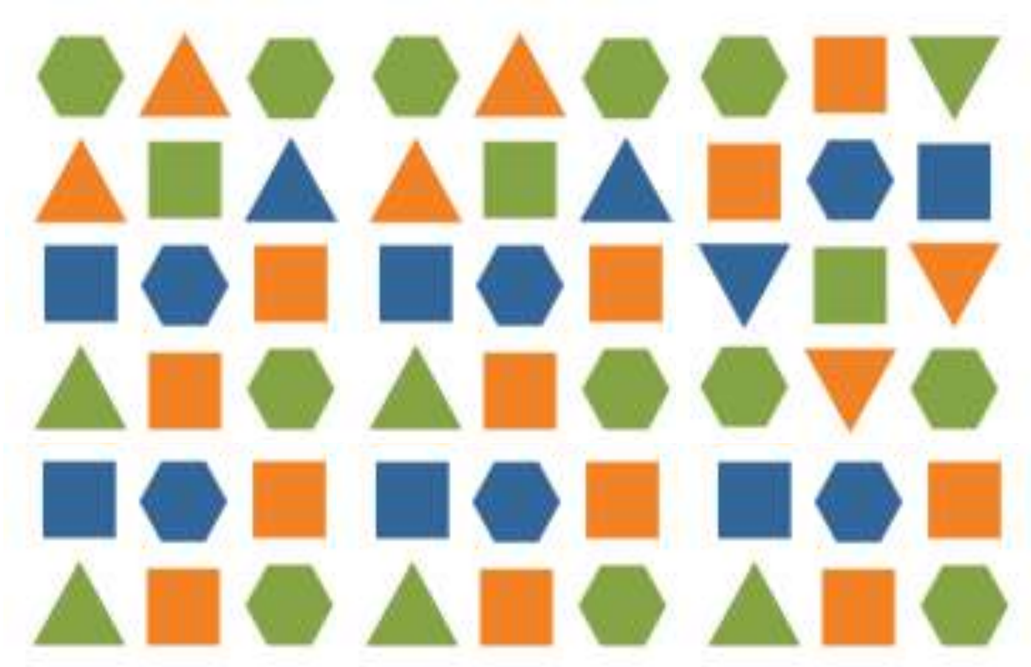
Doug Cutting, 2006

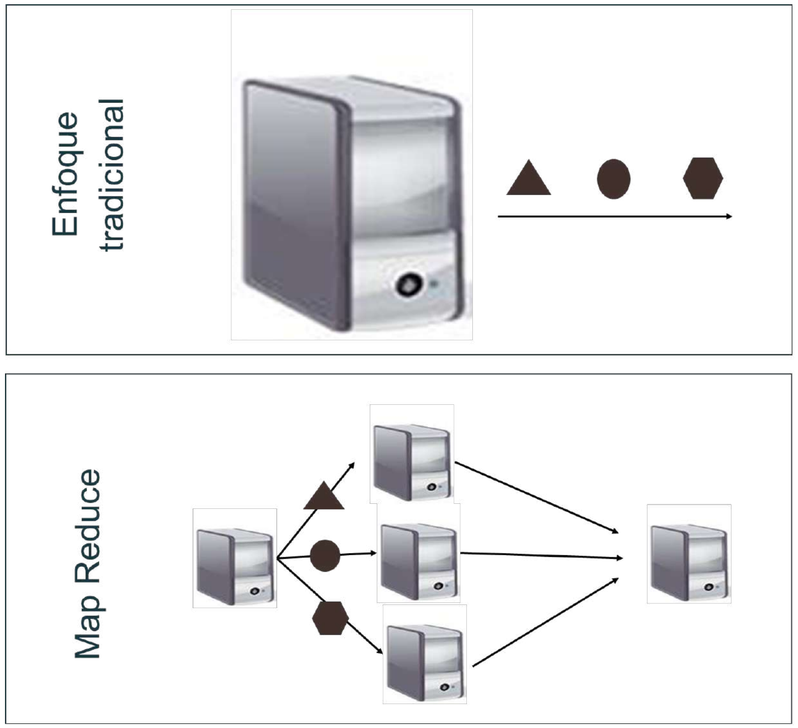
+

Splits the data between the nodes

Code is transferred to the nodes
Data is processed in parallel

DATA LOCALITY: The nodes manipulate the data they have access to

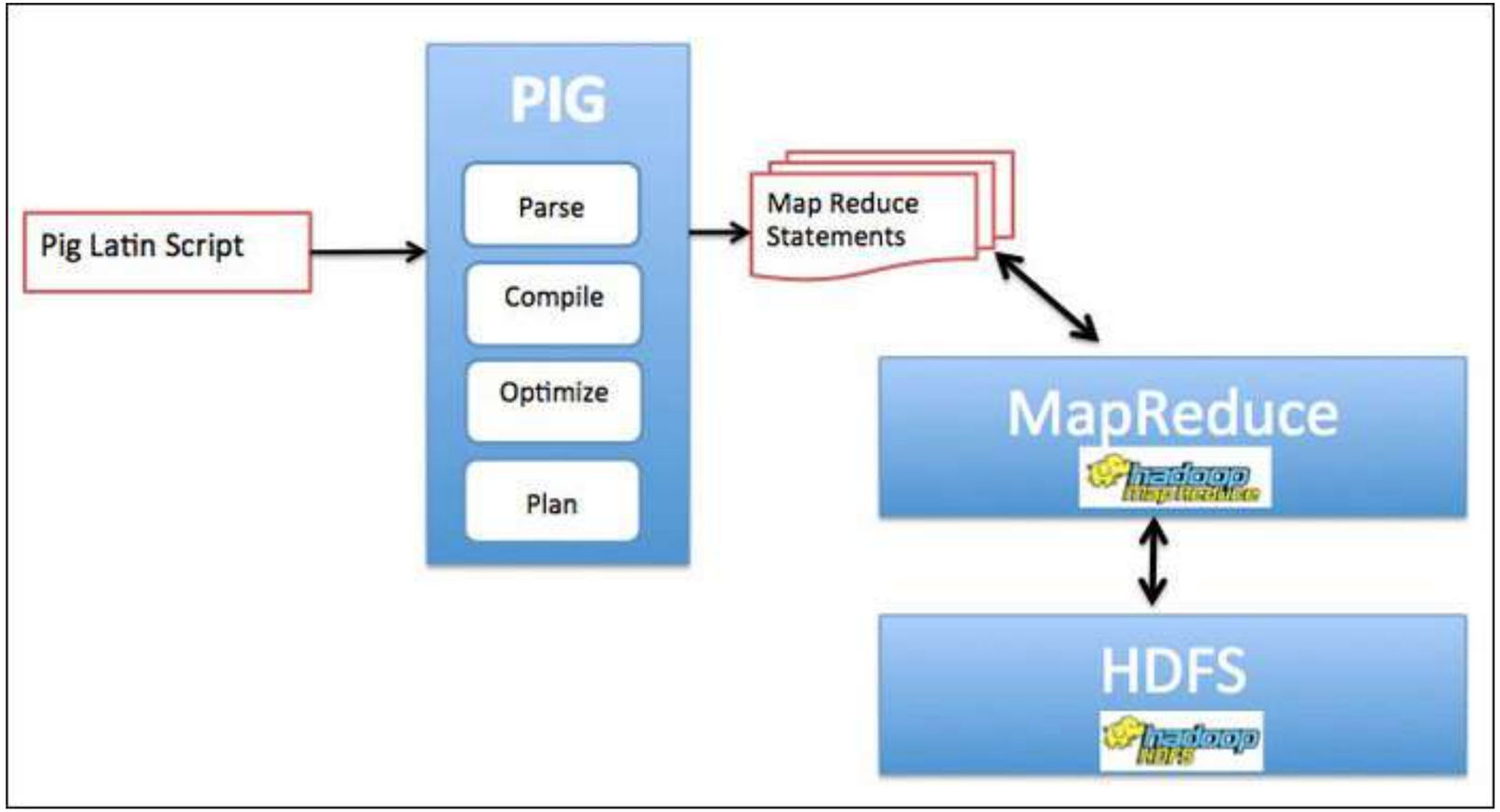




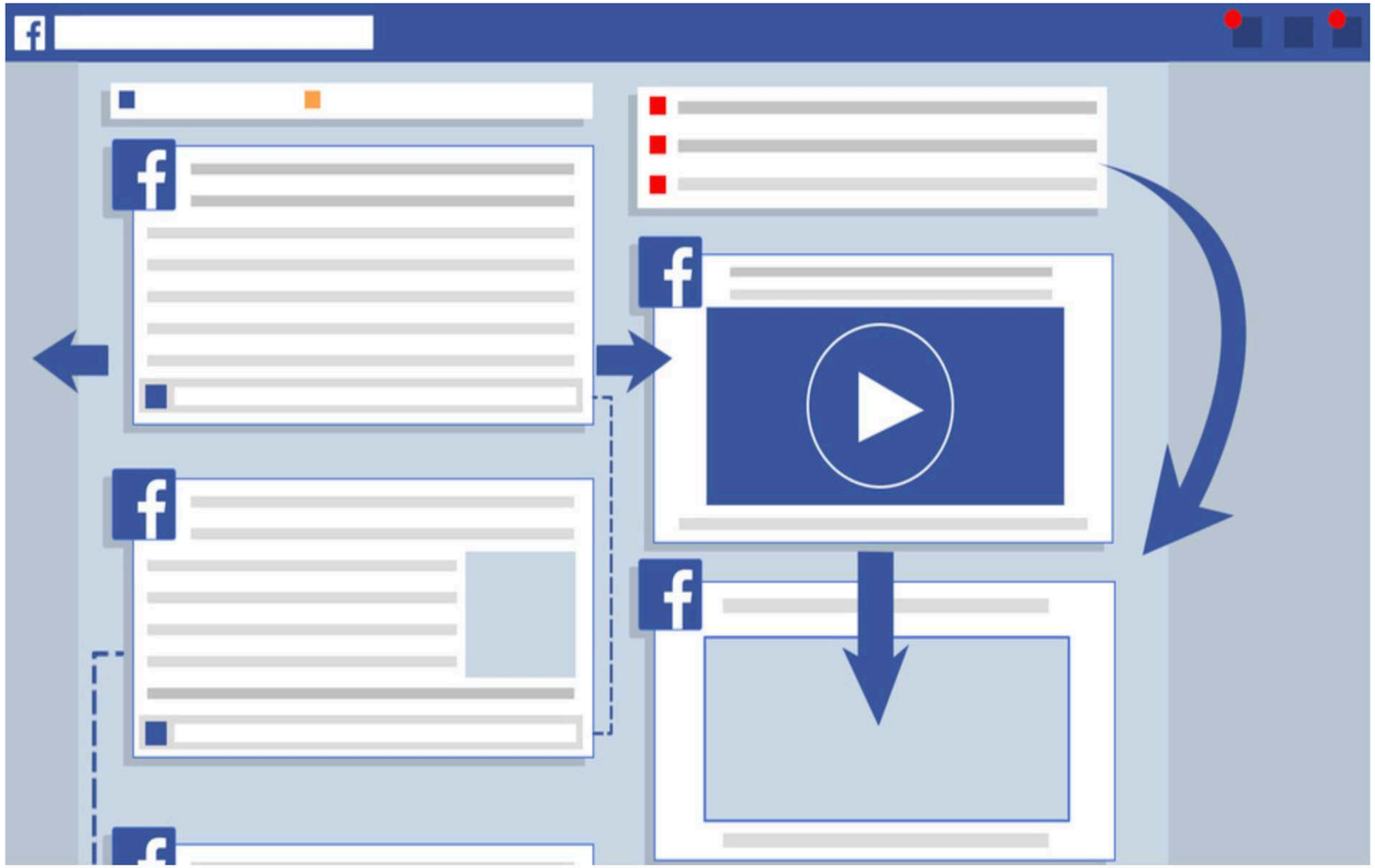
Apache Pig

Hadoop gets a programming language

- Makes it simpler to write MapReduce programs
- Pig Latin abstracts you from specific details and allows you to focus on the data processing



Distributed Database over Hadoop



APACHE STORM™

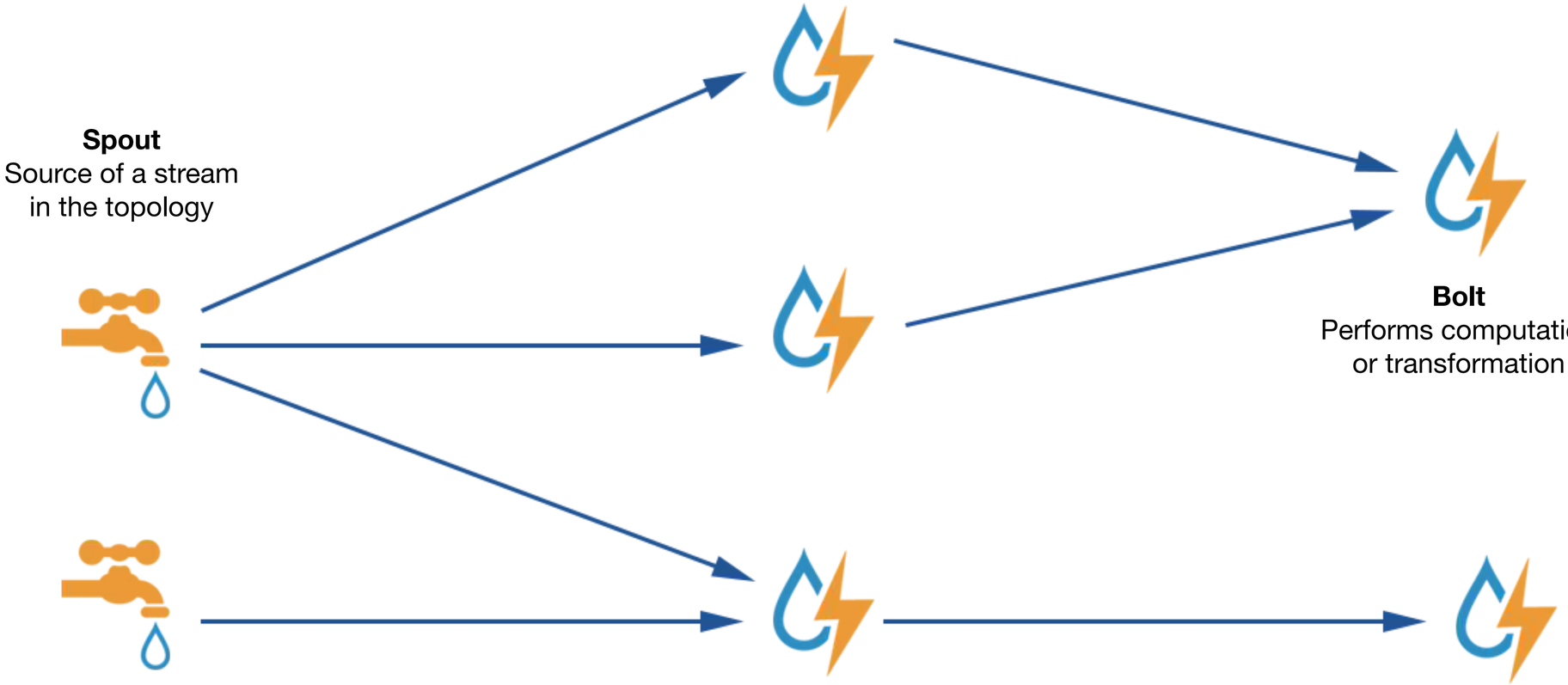
Distributed · Resilient · Real-time

Real-time engine for distributed stream processing

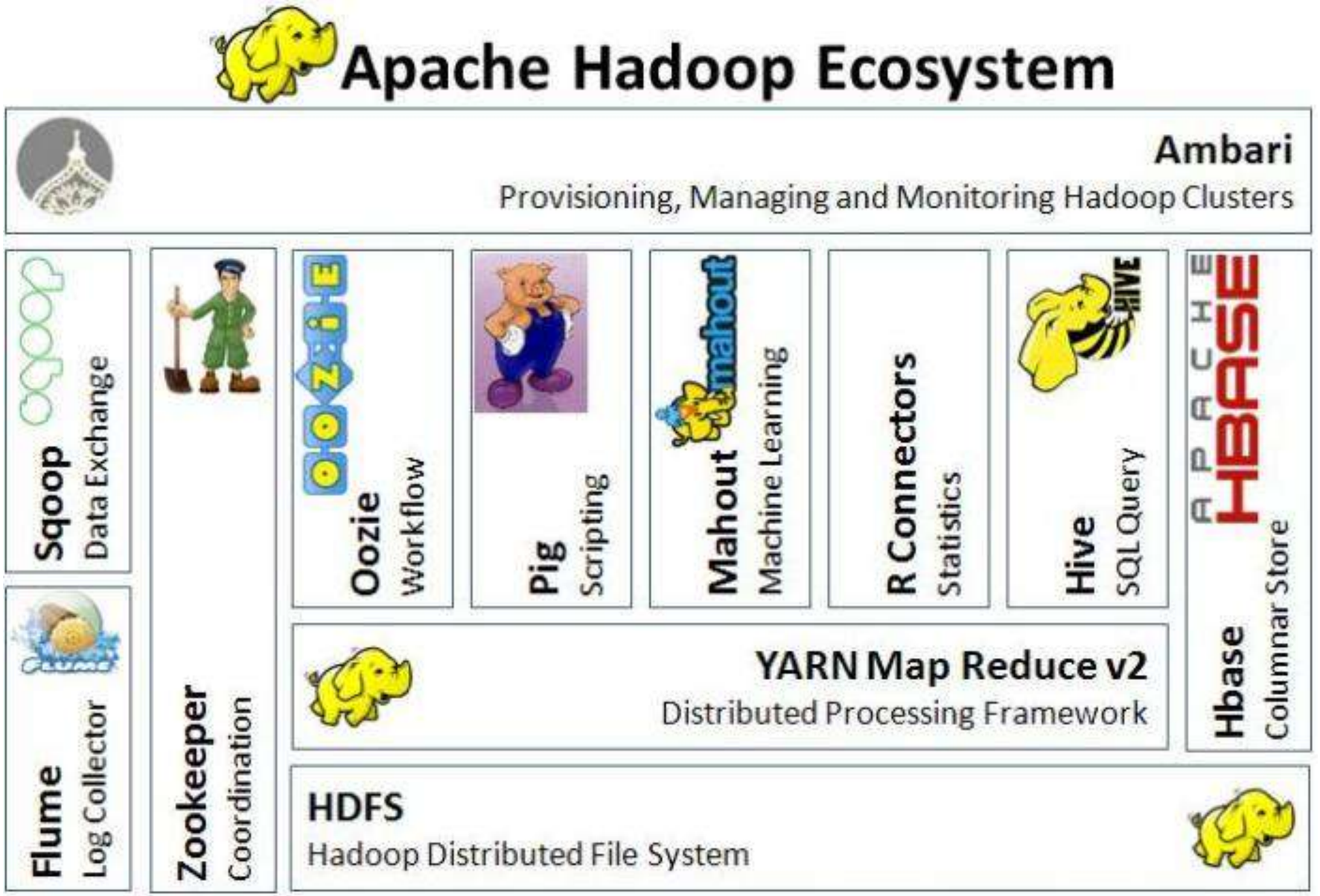
It did for realtime processing what Hadoop did for batch processing

Open sourced after

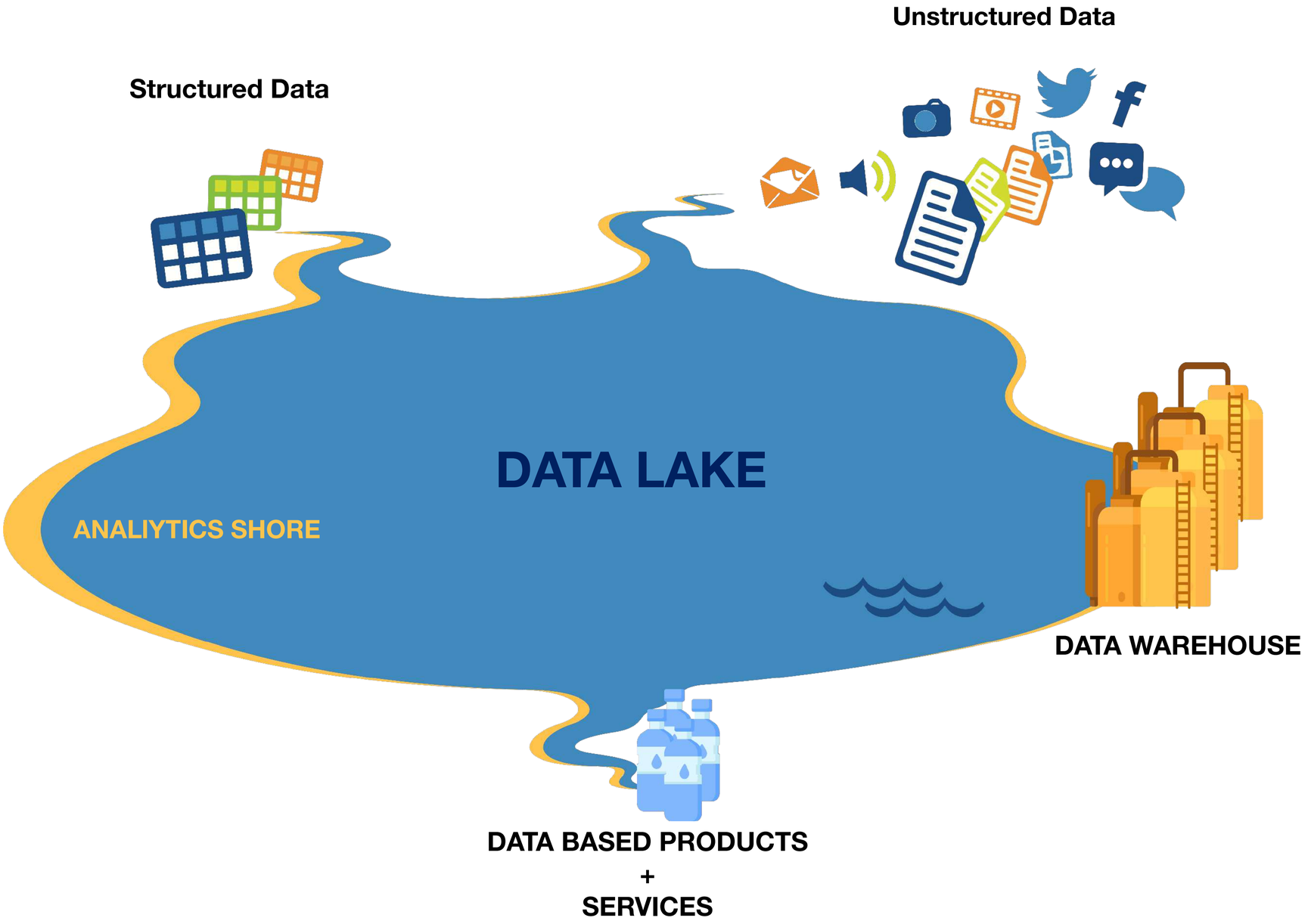
acquired by Twitter

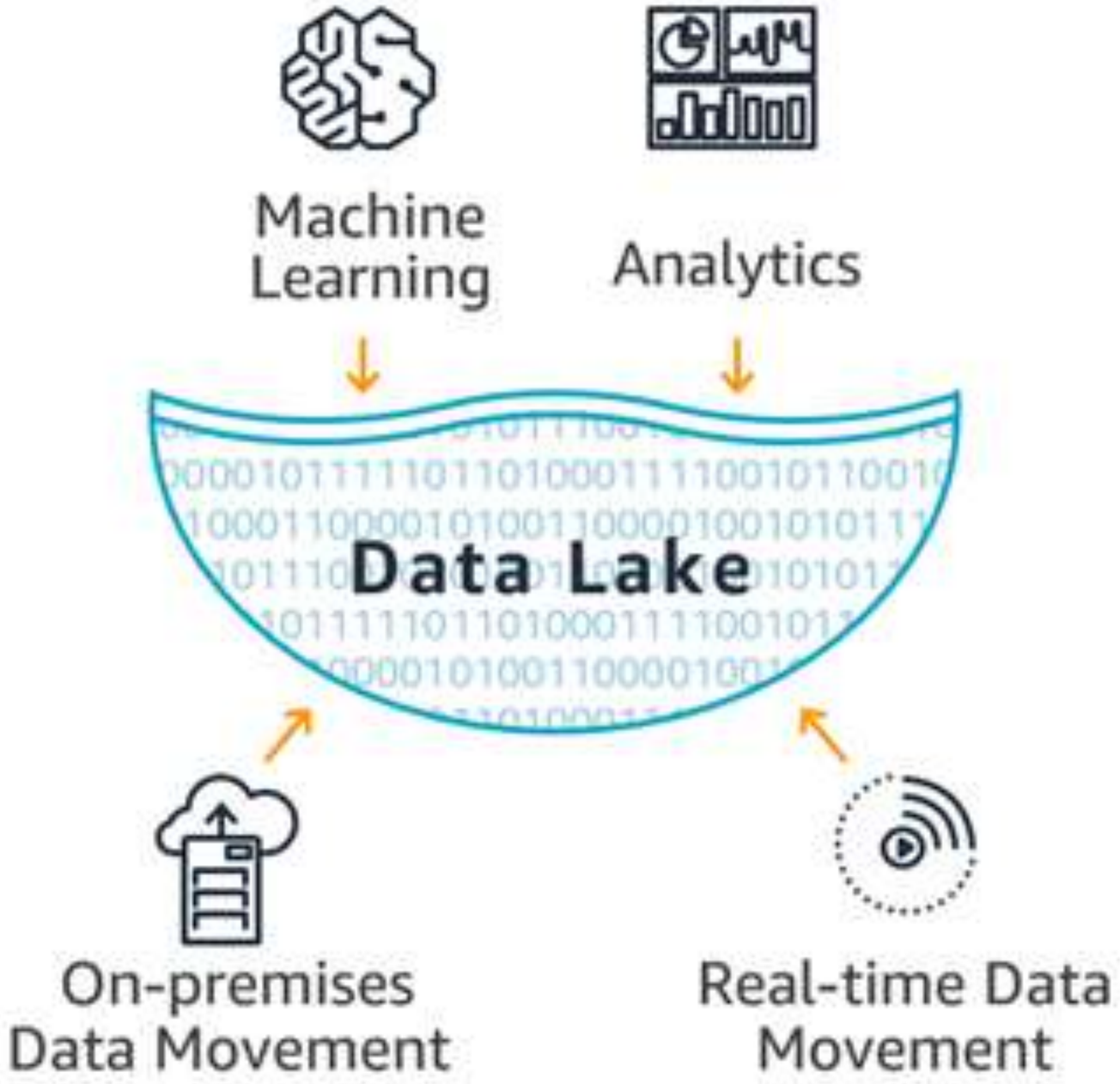


Hadoop as a Big Data Ecosystem




hadoop
VS
Spark


Unstructured Data
Structured Data
DATA LAKE
ANALIYTICS SHORE
DATA WAREHOUSE
DATA BASED PRODUCTS
+
SERVICES


Machine Learning
Analytics
Data Lake
On-premises Data Movement
Real-time Data Movement

Enterprise Ready

Cloud / On premises enterprise big data management platforms

CLOUDERA

(merged 2019)

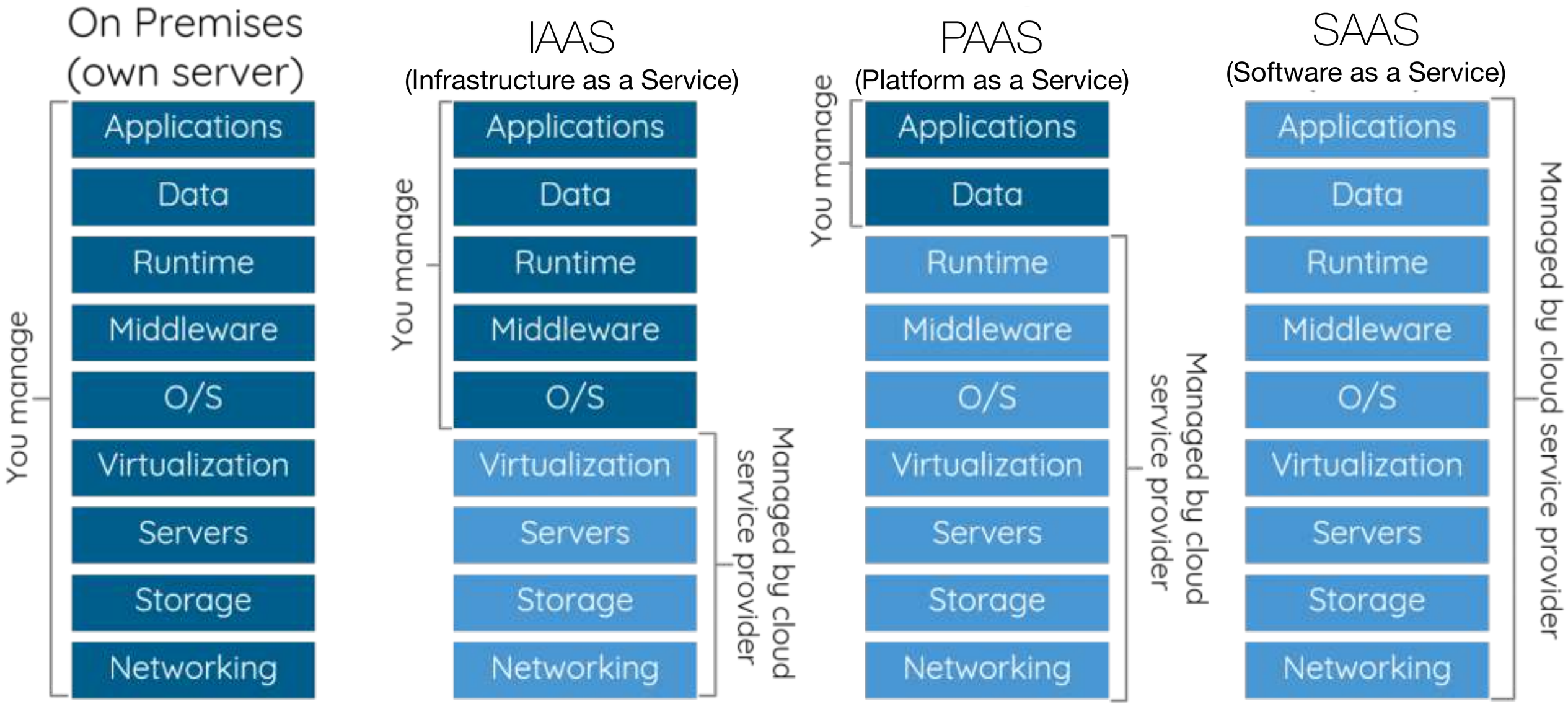
HORTONWORKS®

MAPR
DATA TECHNOLOGIES

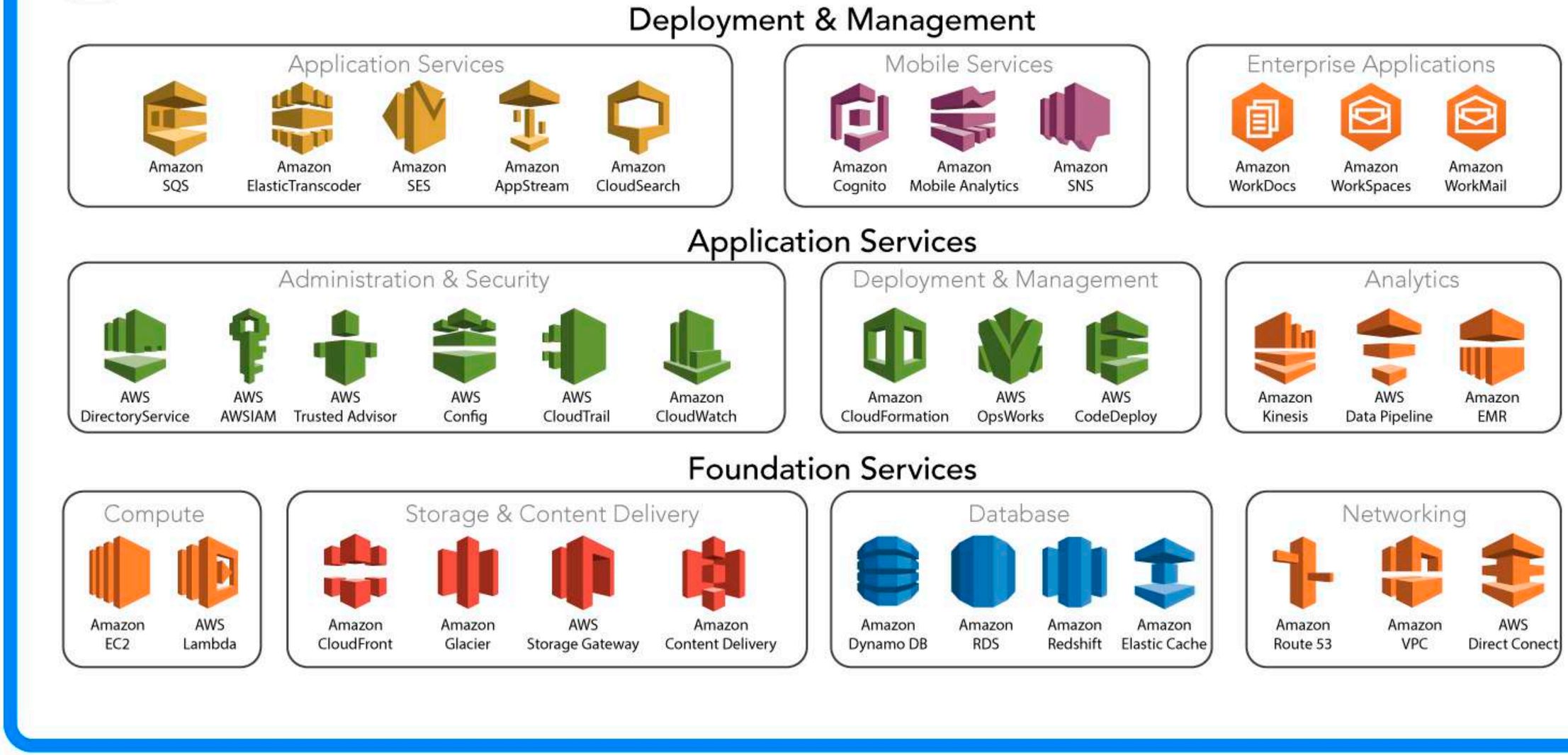
(shut down 2019)

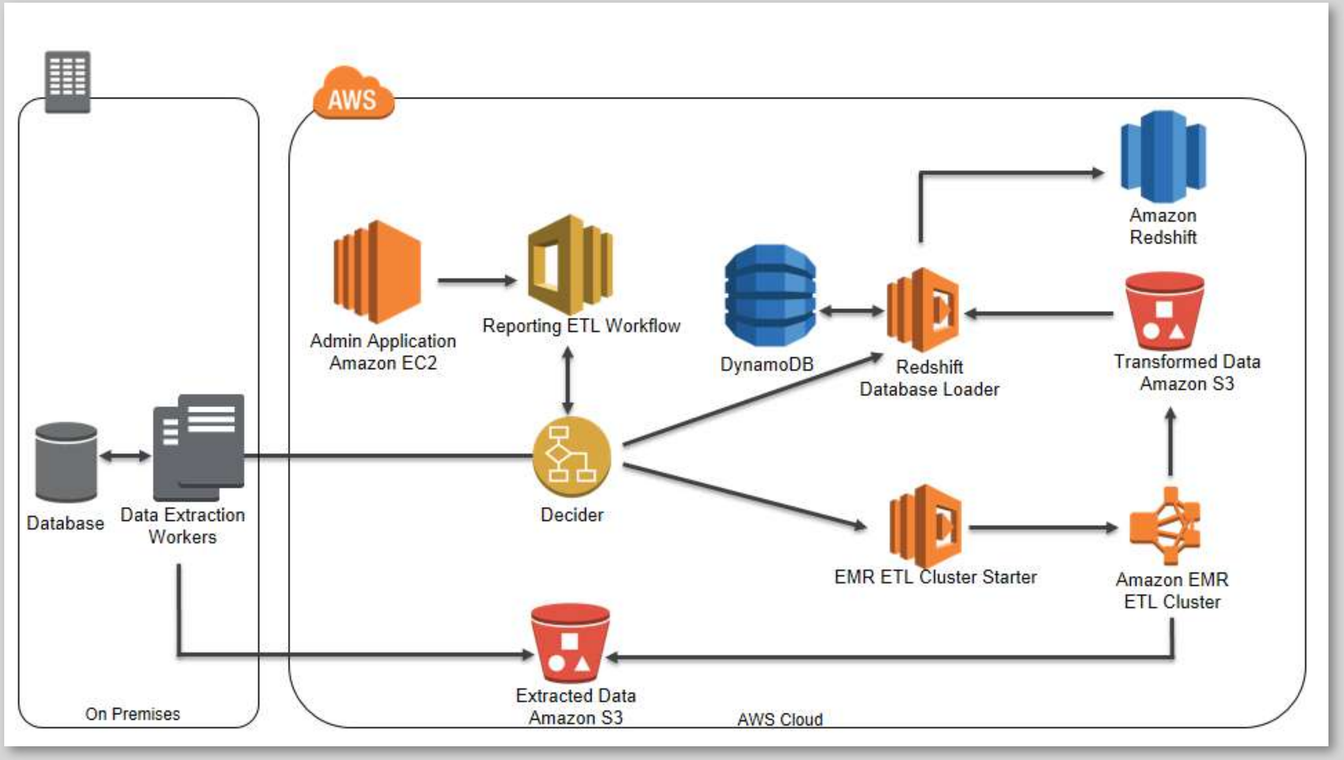
IBM

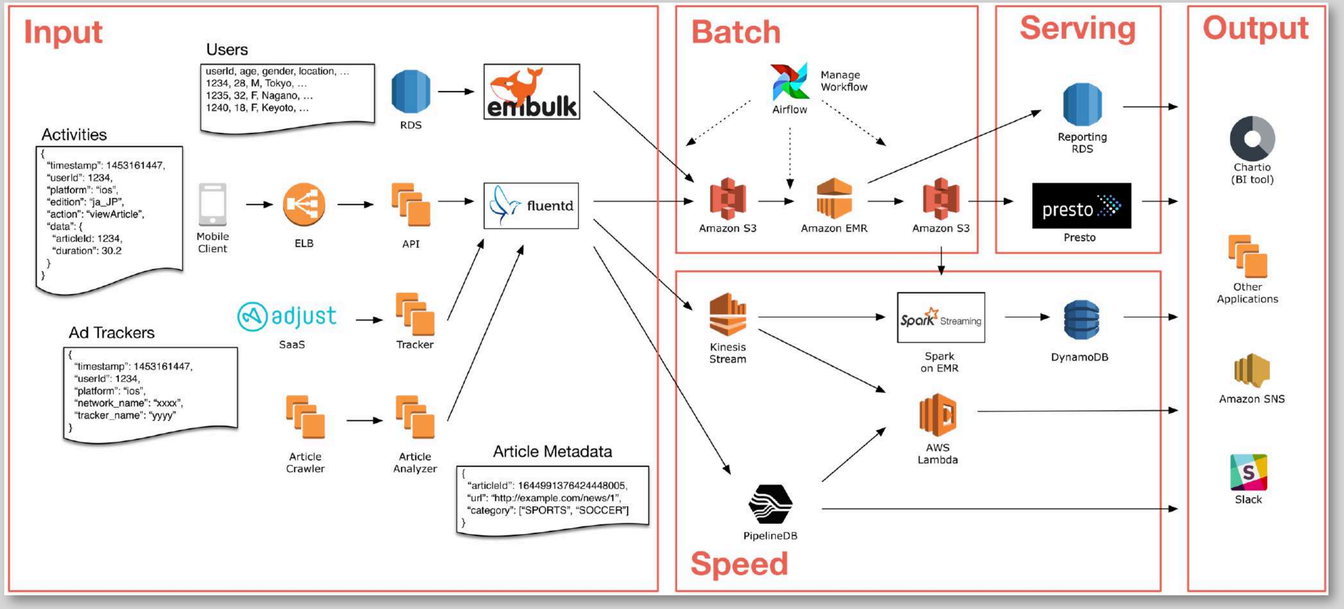

aws
amazon

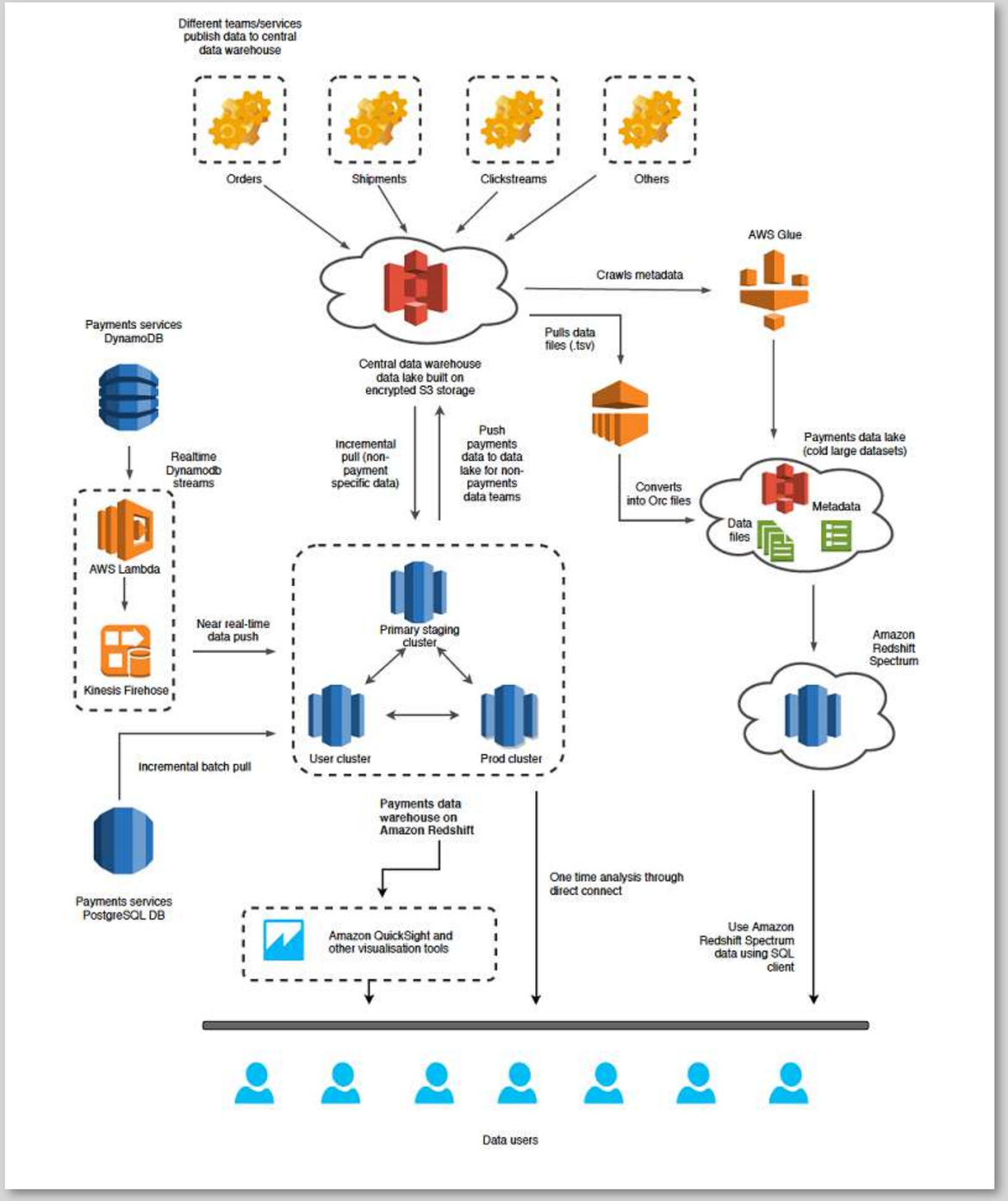

On Premises
(own server)
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You manage
IAAS
(Infrastructure as a Service)
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You manage
Managed by cloud service provider
PAAS
(Platform as a Service)
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You manage
Managed by cloud service provider
SAAS
(Software as a Service)
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Managed by cloud service provider

AWS Services




AWS
Admin Application
Amazon EC2
Reporting ETL Workflow
DynamoDB
Redshift
Database Loader
Amazon
Redshift
Transformed Data
Amazon S3
Database
Data Extraction
Workers
Decider
EMR ETL Cluster Starter
Amazon EMR
ETL Cluster
Extracted Data
Amazon S3
On Premises
AWS Cloud


Input
Users
userId, age, gender, location, ...
1234, 28, M, Tokyo, ...
1235, 32, F, Nagano, ...
1240, 18, F, Koyoto, ...
RDS
embulk
Activities
{
"timestamp": 1453161447,
"userId": 1234,
"platform": "ios",
"edition": "ja_JP",
"action": "viewArticle",
"data": {
"articleId: 1234,
"duration": 30.2
}
}
Mobile
Client
ELB
API
fluentd
adjust
SaaS
Tracker
Ad Trackers
{
"timestamp": 1453161447,
"userId": 1234,
"platform": "ios",
"network_name": "xxxx",
"tracker_name": "yyyy"
}
Article
Crawler
Article
Analyzer
Article Metadata
{
"articleId": 1644991376424448005,
"url": "http://example.com/news/1",
"category": ["SPORTS", "SOCCER"]
}
Batch
Manage
Workflow
Airflow
Amazon S3
Amazon EMR
Amazon S3
Kinesis
Stream
Spark Streaming
Spark
on EMR
DynamoDB
AWS
Lambda
PipelineDB
Speed
Serving
Reporting
RDS
presto
Presto
Output
Chartio
(BI tool)
Other
Applications
Amazon SNS
Slack


Different teams/services
publish data to central
data warehouse
Orders
Shipments
Clickstreams
Others
AWS Glue
Crawls metadata
Pulls data
files (.tsv)
Payments services
DynamoDB
Central data warehouse
data lake built on
encrypted S3 storage
Payments data lake
(cold large datasets)
Realtime
Dynamodb
streams
Incremental
pull (non-
payment
specific data)
Push
payments
data to data
lake for non-
payments
data teams
Converts
into Orc files
Metadata
Data
files
AWS Lambda
Kinesis Firehose
Near real-time
data push
Primary staging
cluster
User cluster
Prod cluster
Amazon
Redshift
Spectrum
Incremental batch pull
Payments data
warehouse on
Amazon Redshift
Payments services
PostgreSQL DB
Amazon QuickSight and
other visualisation tools
One time analysis through
direct connect
Use Amazon
Redshift Spectrum
data using SQL
client
Data users

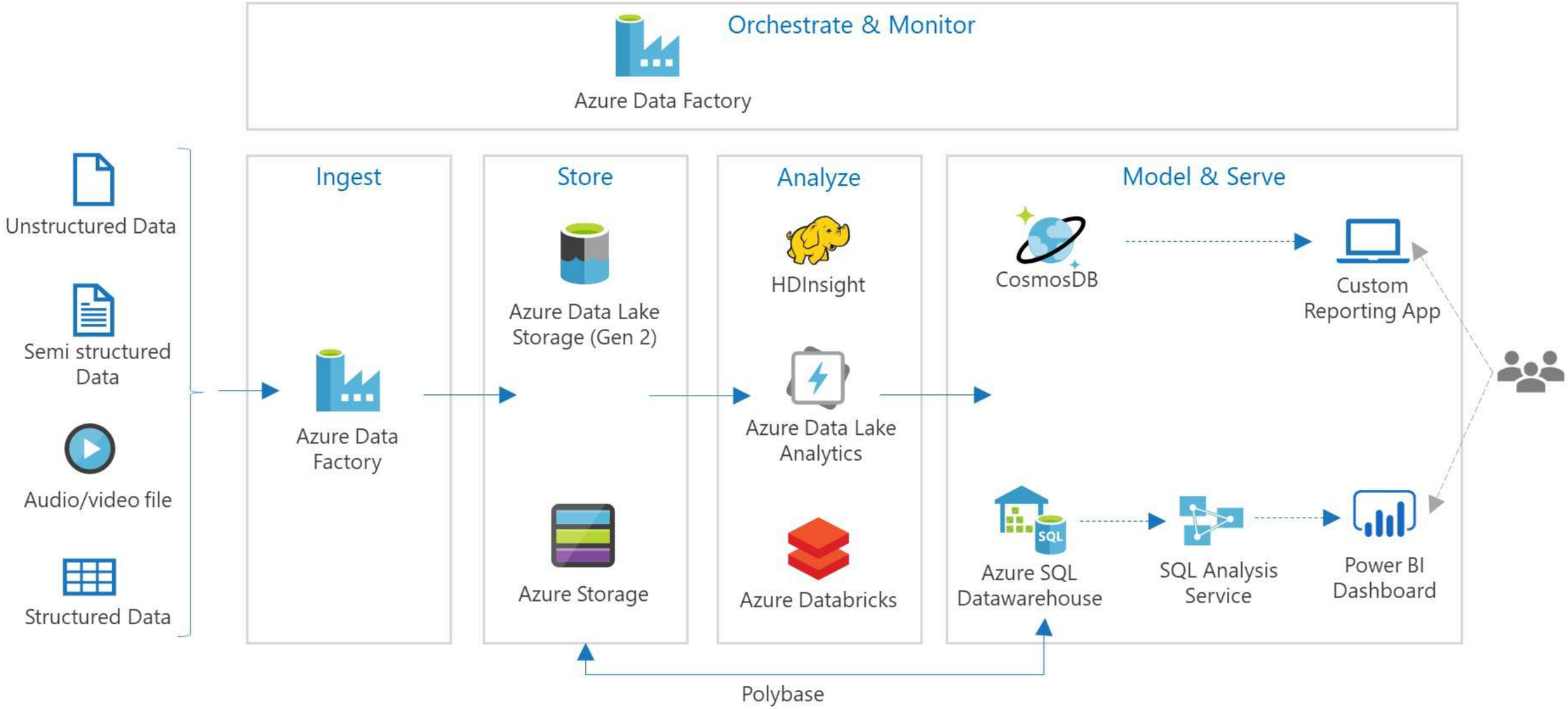
The Blockbuster Growth of Amazon's Cloud Business

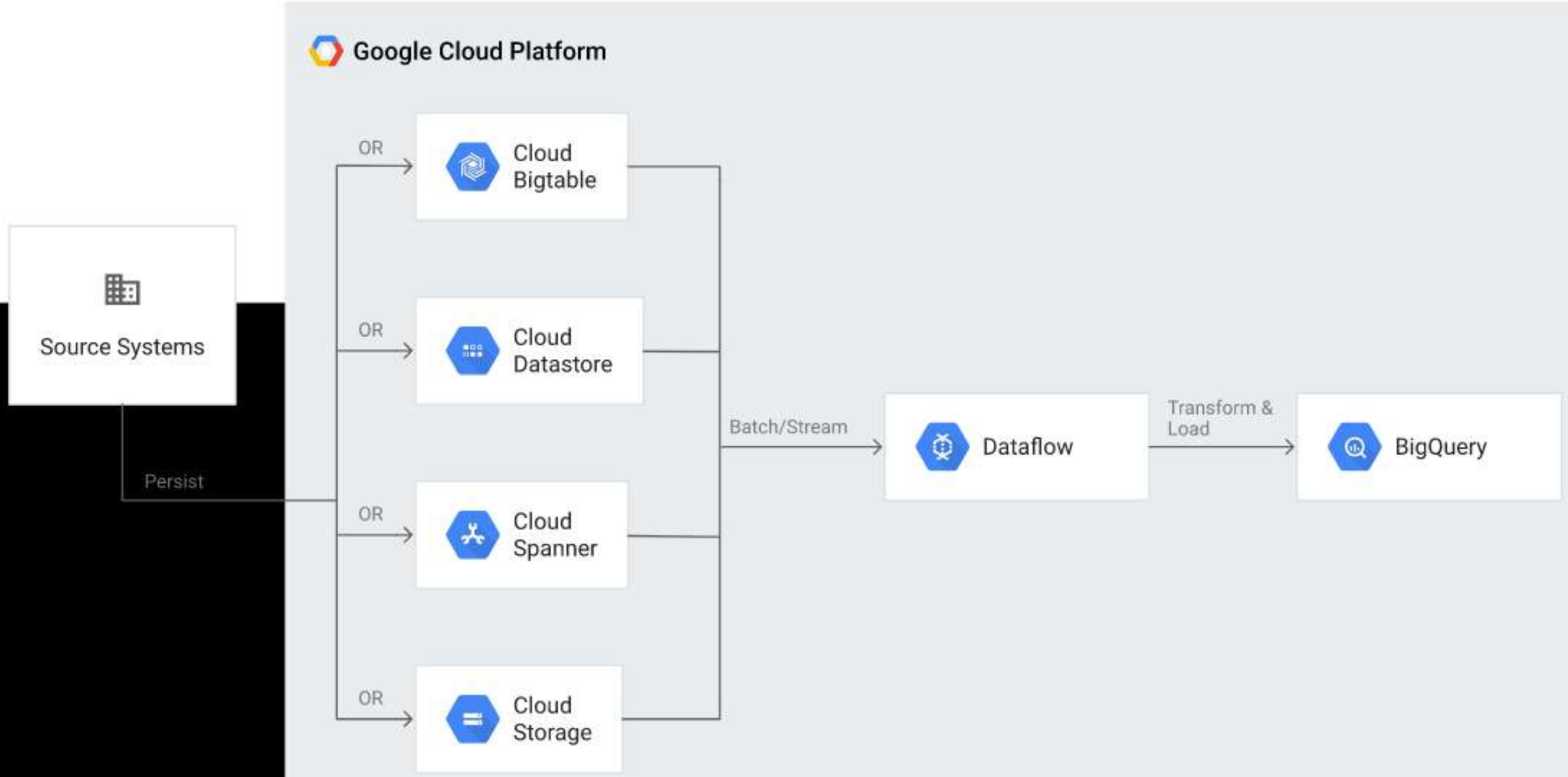
Quarterly revenue of Amazon Web Services*

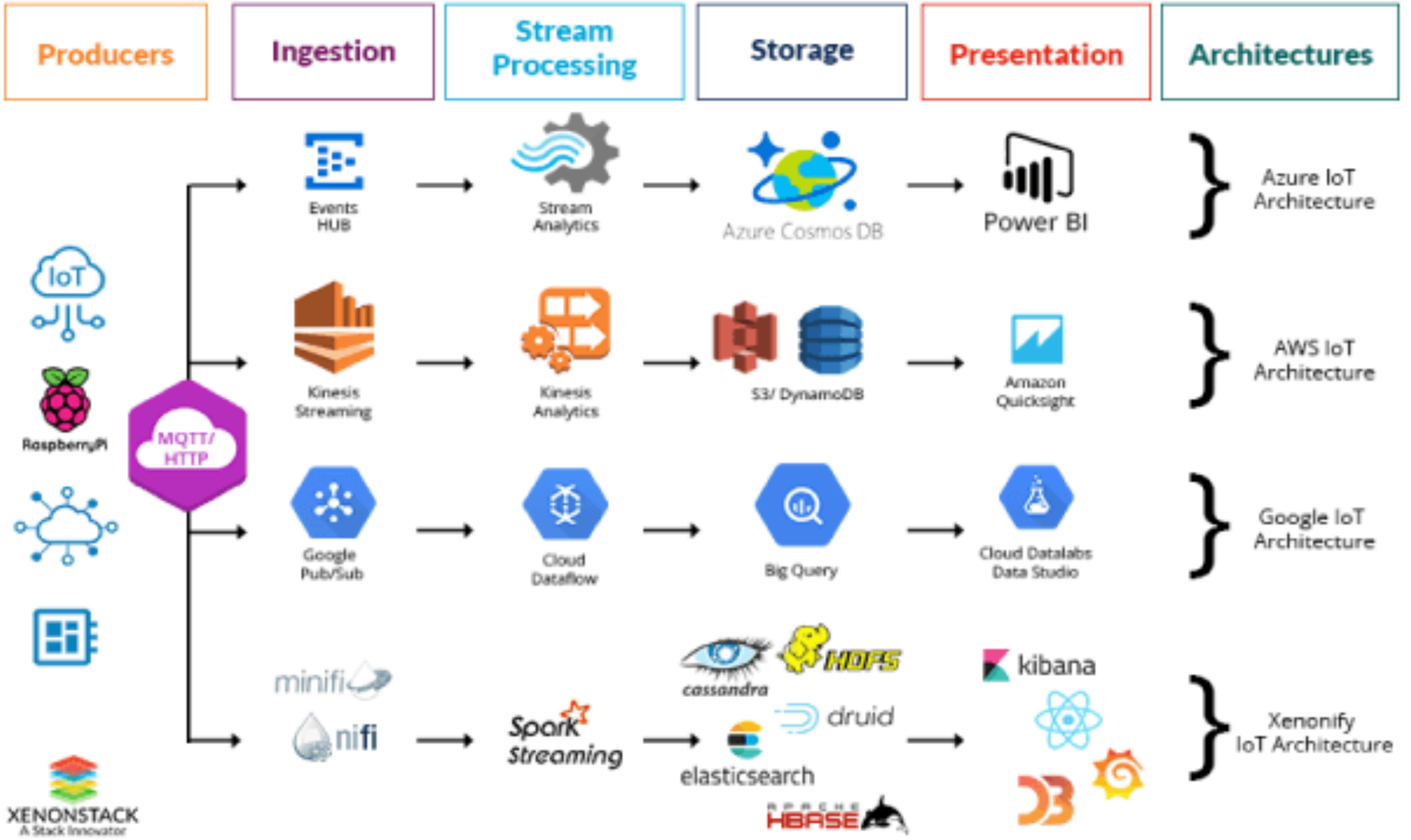
* Amazon Web Services (AWS) is Amazon's cloud platform, providing cloud infrastructure services to start-ups, enterprises and the public sector.

Source: Amazon

Microsoft Azure

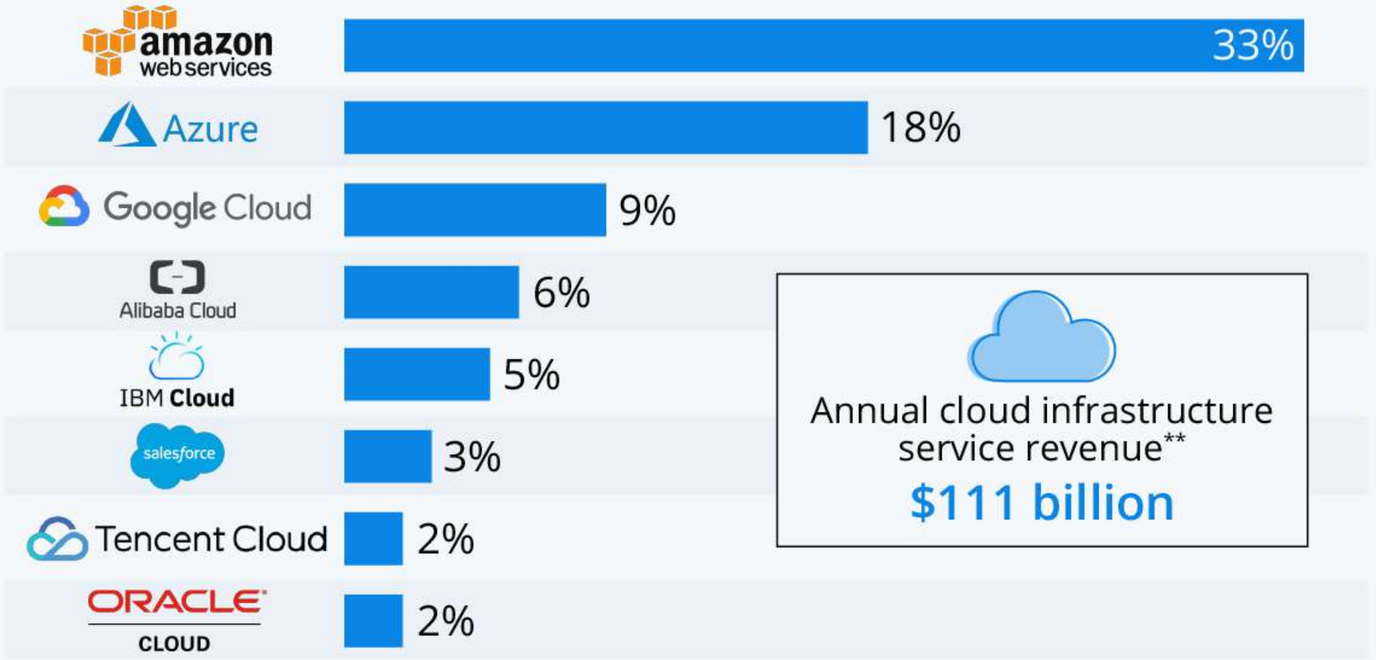



Google Cloud Platform
Source Systems
Persist
OR
Cloud Bigtable
OR
Cloud Datastore
OR
Cloud Spanner
OR
Cloud Storage
Batch/Stream
Dataflow
Transform & Load
BigQuery


Producers
Ingestion
Stream Processing
Storage
Presentation
Architectures
IoT
RaspberryPi
MQTT/ HTTP
Events HUB
Stream Analytics
Azure Cosmos DB
Power BI
Azure IoT Architecture
Kinesis Streaming
Kinesis Analytics
S3/ DynamoDB
Amazon Quicksight
AWS IoT Architecture
Google Pub/Sub
Cloud Dataflow
Big Query
Cloud Datalabs Data Studio
Google IoT Architecture
minifi
nifi
Spark Streaming
cassandra
HDFS
druid
elasticsearch
APACHE HBASE
kibana
Xenonify IoT Architecture
XENONSTACK
A Stack Innovator

Amazon Leads $100 Billion Cloud Market

Worldwide market share of leading cloud infrastructure service providers in Q2 2020*



* includes platform as a service (PaaS) and infrastructure as a service (IaaS) as well as hosted private cloud services

** 12 months ended June 30, 2020

Source: Synergy Research Group

CNIO
(On premise)

2010

2 years

SANITAS
(IaaS)

2016

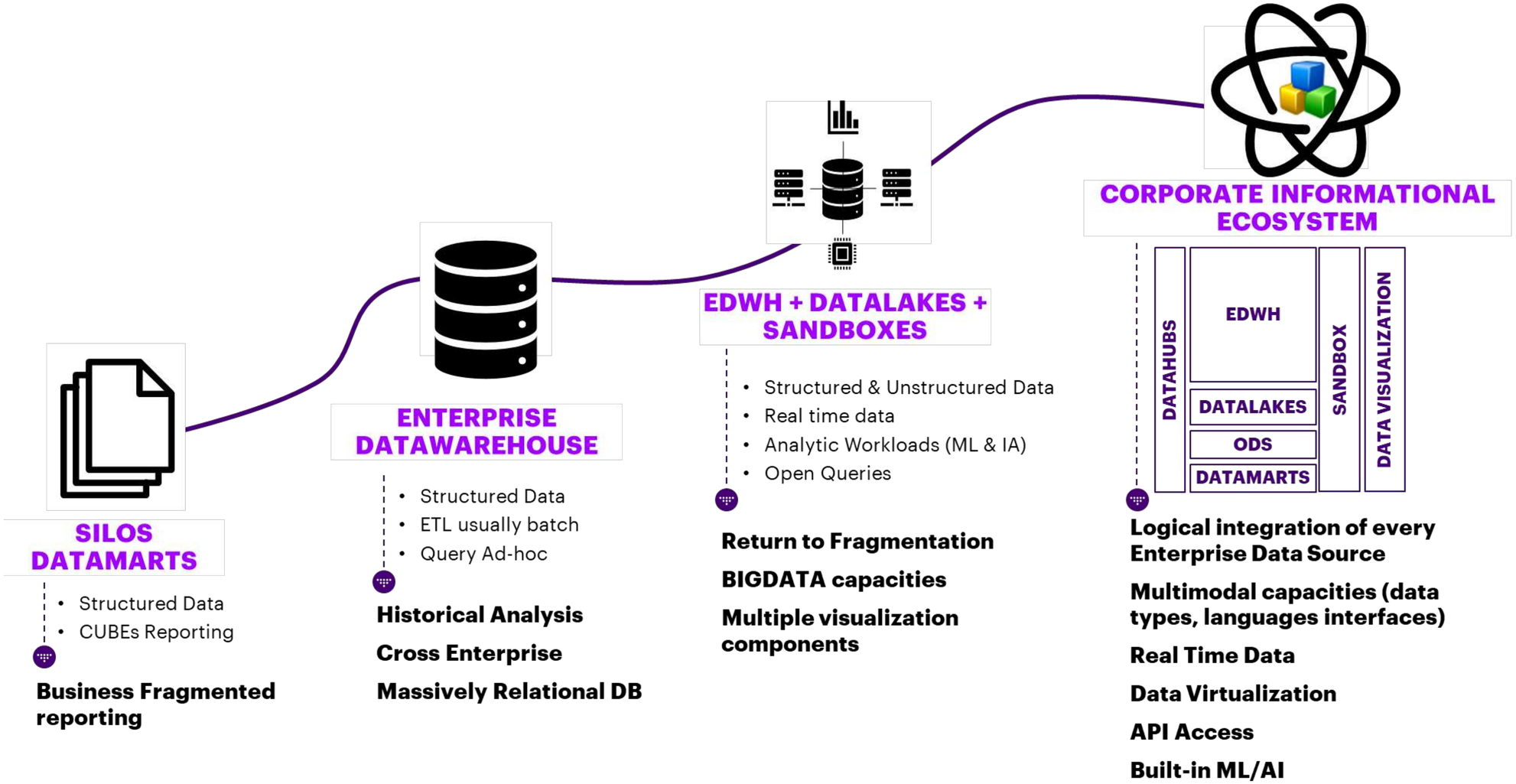
3 months

Genomcore
(SaaS)

2020

<1 hour

Data Platform evolution



ARIA

Our Technological Approach

Cloud infrastructure that yields flexibility, scalability, and rapid deployment

Azure

Reducing maintenance and operation costs

Mostly PaaS that makes possible to increase speed and decrease maintenance.

Is it built case by case selecting those component that will be reusable for other use cases

DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

DATA INGESTION	IoT HUB	DATA EXTRACTORS	External API Extractor
IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

DATA STORAGE	GOLDEN ZONE DATA LAKES	SILVER ZONE	
DATA INGESTION	IoT HUB	DATA EXTRACTORS	External API Extractor
IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

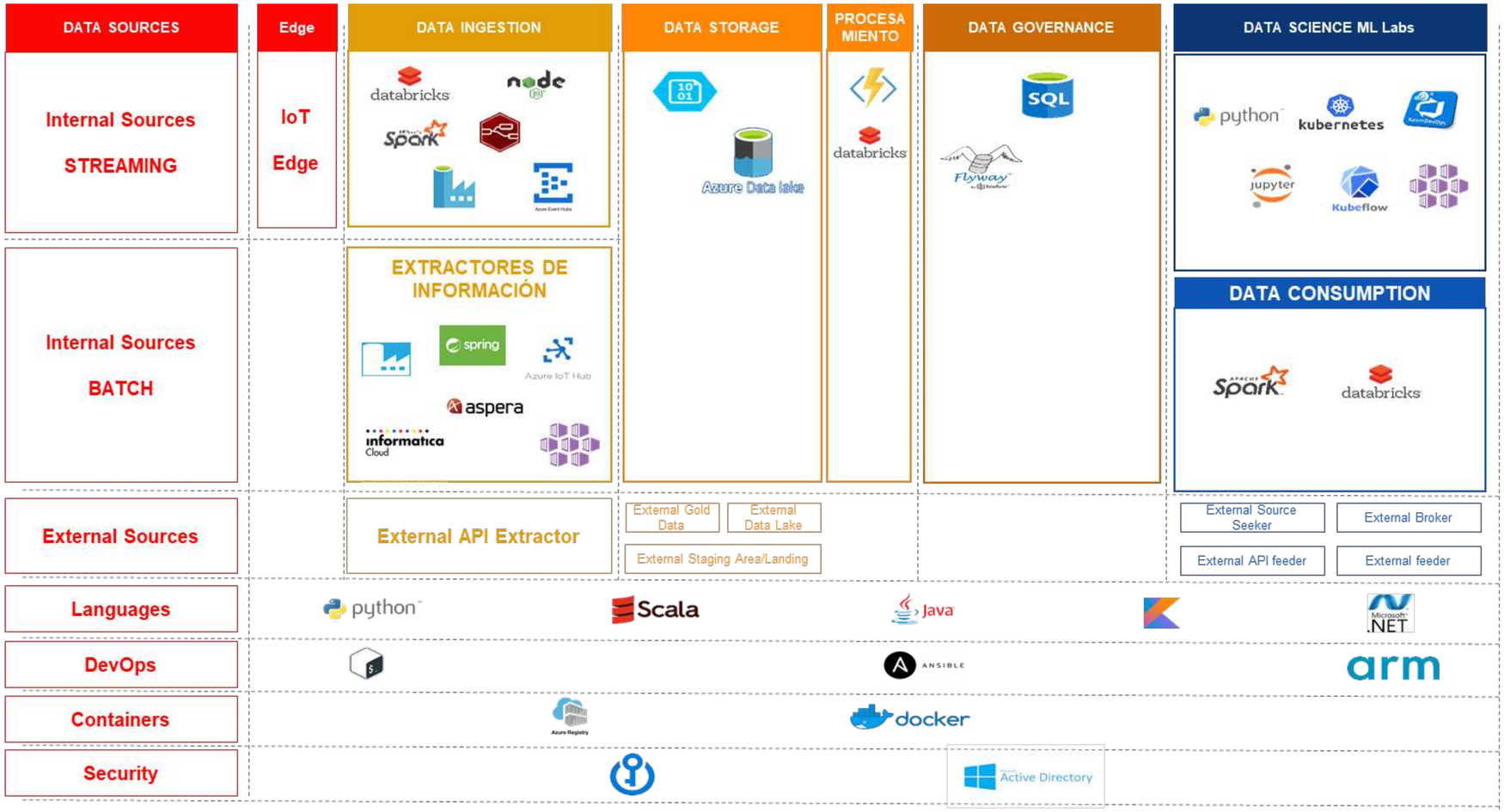
DATA GOVERNANCE	DATA DICTIONARY DATA QUALITY METADATA MANAGEMENT	DATA LINAGE ACCESS MANAGEMENT	
DATA STORAGE	GOLDEN ZONE DATA LAKES	SILVER ZONE	
DATA INGESTION	IoT HUB	DATA EXTRACTORS	External API Extractor
IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

ML LABS	DEVELOPMENT	TRAINING DEPLOYMENT	
DATA GOVERNANCE	DATA DICTIONARY DATA QUALITY	DATA LINAGE	
	METADATA MANAGEMENT	ACCESS MANAGEMENT	
DATA STORAGE	GOLDEN ZONE DATA LAKES	SILVER ZONE	
DATA INGESTION	IoT HUB	DATA EXTRACTORS	External API Extractor
IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

ARIA

DATA CONSUMPTION	INTERNAL API		EXTERNAL API
ML LABS	DEVELOPMENT	TRAINING	DEPLOYMENT
DATA GOVERNANCE	DATA DICTIONARY METADATA MANAGEMENT	DATA QUALITY	DATA LINAGE ACCESS MANAGEMENT
DATA STORAGE	GOLDEN ZONE DATA LAKES		SILVER ZONE
DATA INGESTION	IoT HUB	DATA EXTRACTORS	External API Extractor
IoT	IoT EDGE		
DATA SOURCES	Internal Sources STREAMING	Internal Sources BATCH	External data

ARIA


DATA SOURCES
Edge
DATA INGESTION
DATA STORAGE
PROCESAMIENTO
DATA GOVERNANCE
DATA SCIENCE ML Labs
Internal Sources STREAMING
IoT Edge
databricks
node
Spark
Azure Data lake
SQL
Flyway
python
kubernetes
jupyter
Kubeflow
Internal Sources BATCH
EXTRACTORES DE INFORMACIÓN
spring
Azure IoT Hub
aspera
informatica Cloud
DATA CONSUMPTION
Spark
databricks
External Sources
External API Extractor
External Gold Data
External Data Lake
External Staging Area/Landing
External Source Seeker
External Broker
External API feeder
External feeder
Languages
python
Scala
Java
.NET
DevOps
ANSIBLE
arm
Containers
Azure Registry
docker
Security
Active Directory

+40

Data **Platform**

¿What is the data management platform (**DMP**) used for collecting and managing data? ¿It allows to use big data and artificial intelligence algorithms to process and analyze large data sets? ¿Can it grow and acquire new functionalities as the data grows and the business needs evolve?

Data **Governance**

¿What are the processes and policies in place that ensure that high **data quality** exists through its complete lifecycle? ¿What data **controls** there are that support business objectives? ¿It is the data **available**, **usable**, **consistent** and **secure**?

Data **Governance**

Regulatory
Peace of Mind

DATA GOVERNANCE FOR ANALYTICS

DATA GOVERNANCE is a set of guidelines and mechanisms based on a series of principles and best practices that provide the company with the tools to guarantee the knowledge, security, quality and availability of information in accordance with organization policies, and with third parties.

GOALS:

INVENTORY
Have a complete information inventory avoiding information silos and duplicate data sets

LANGUAGE
Establish a common language in the company that allows to univocally identify a data avoiding duplication of information

QUALITY
Control and improve the quality of the information, taking care of its completeness, inconsistency and integrity.

COMPLIANCE
Comply with regulations, avoiding sanctions based on the traceability of information and the establishment of timely access controls.

SECURITY
Improve information security by establishing new roles and responsibilities in the company for the custody of information.

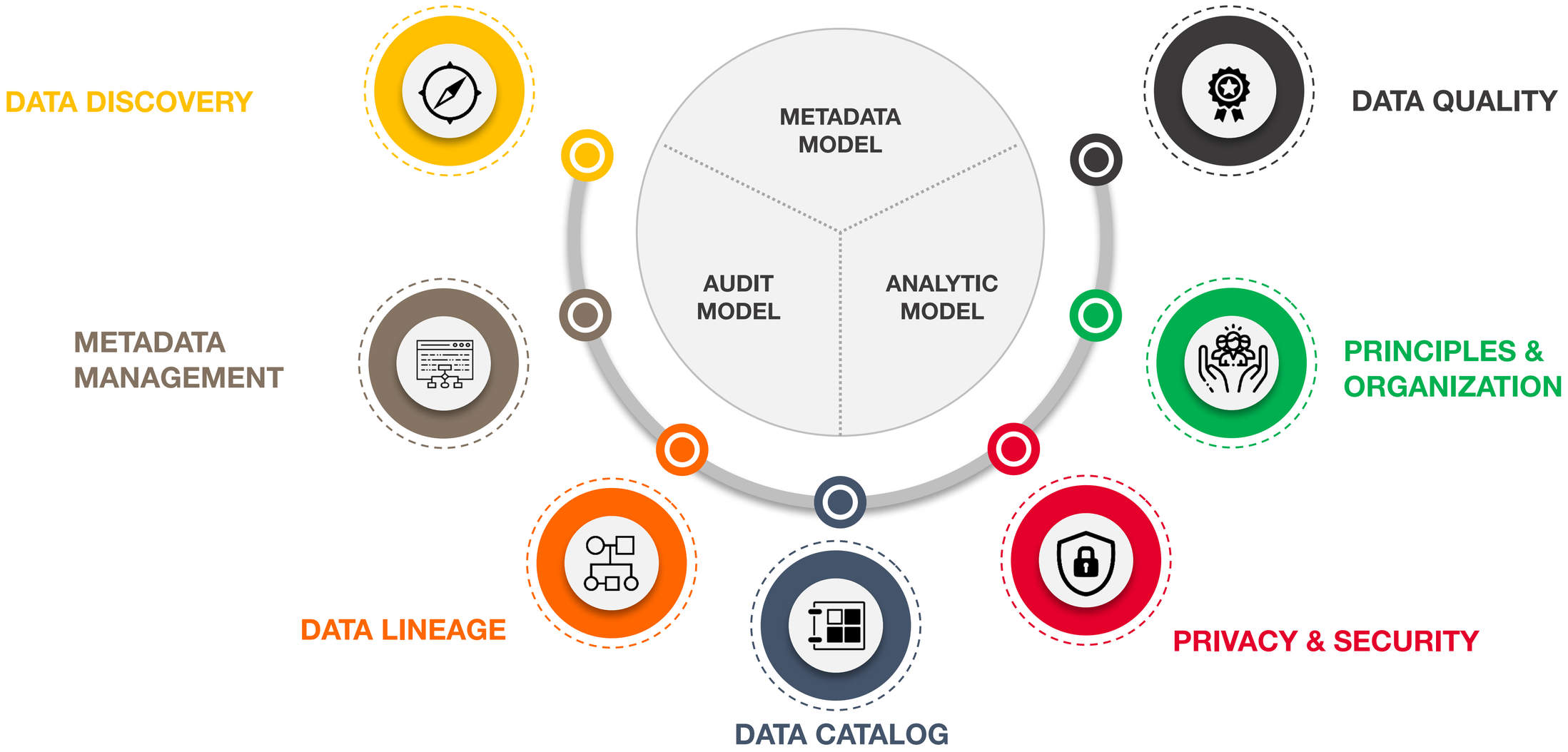
TRANSFERS
Establish information transfer policies including roles and responsibilities of the parties involved.

TIME TO MARKET
Optimize the execution of analytical projects avoiding the reprocessing of searches, transfers and cleaning of information.

DEMOCRATIZATION
Democratize access to information so that business decisions are made based on global information.

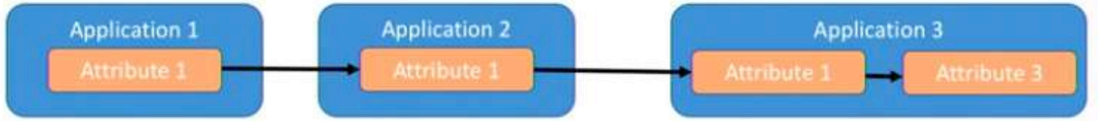
Data Governance for Analytics

Components Map

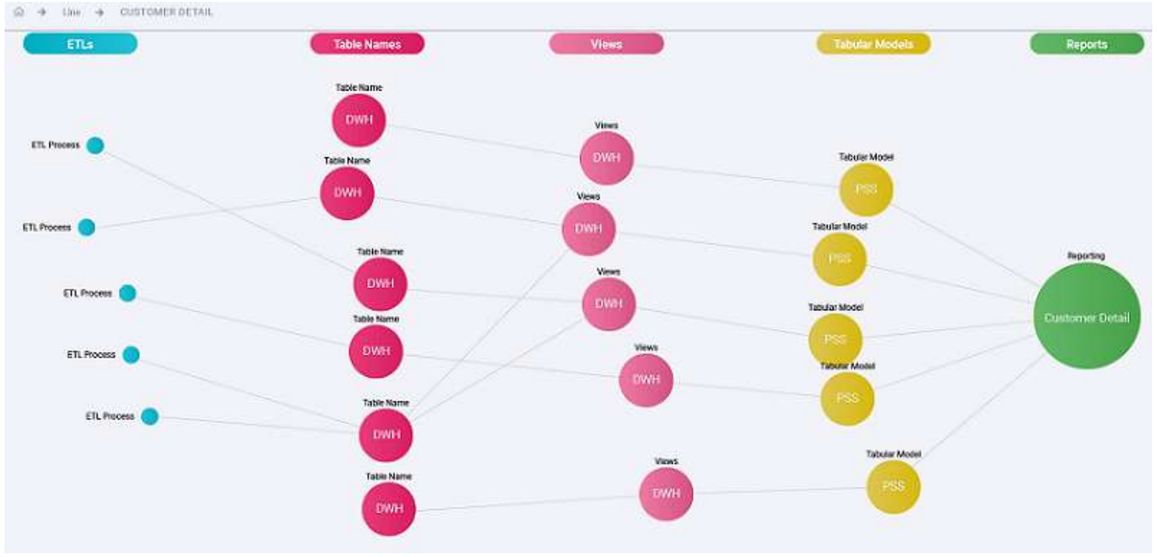


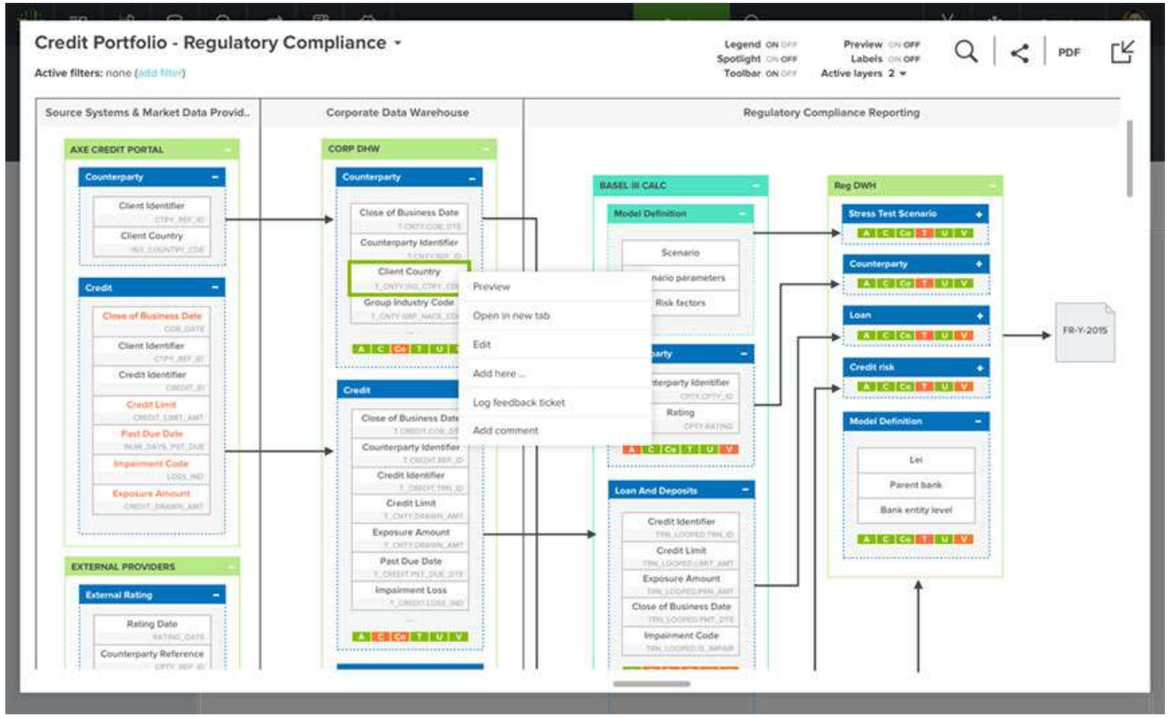
Data Lineage

What is Data Lineage?



"Data lineage is generally defined as a kind of data life cycle that includes the data's origins and where it moves over time. This term can also describe what happens to data as it goes through diverse processes. Data lineage can help with efforts to analyze how information is used and to track key bits of information that serve a particular purpose." - *Techopedia*



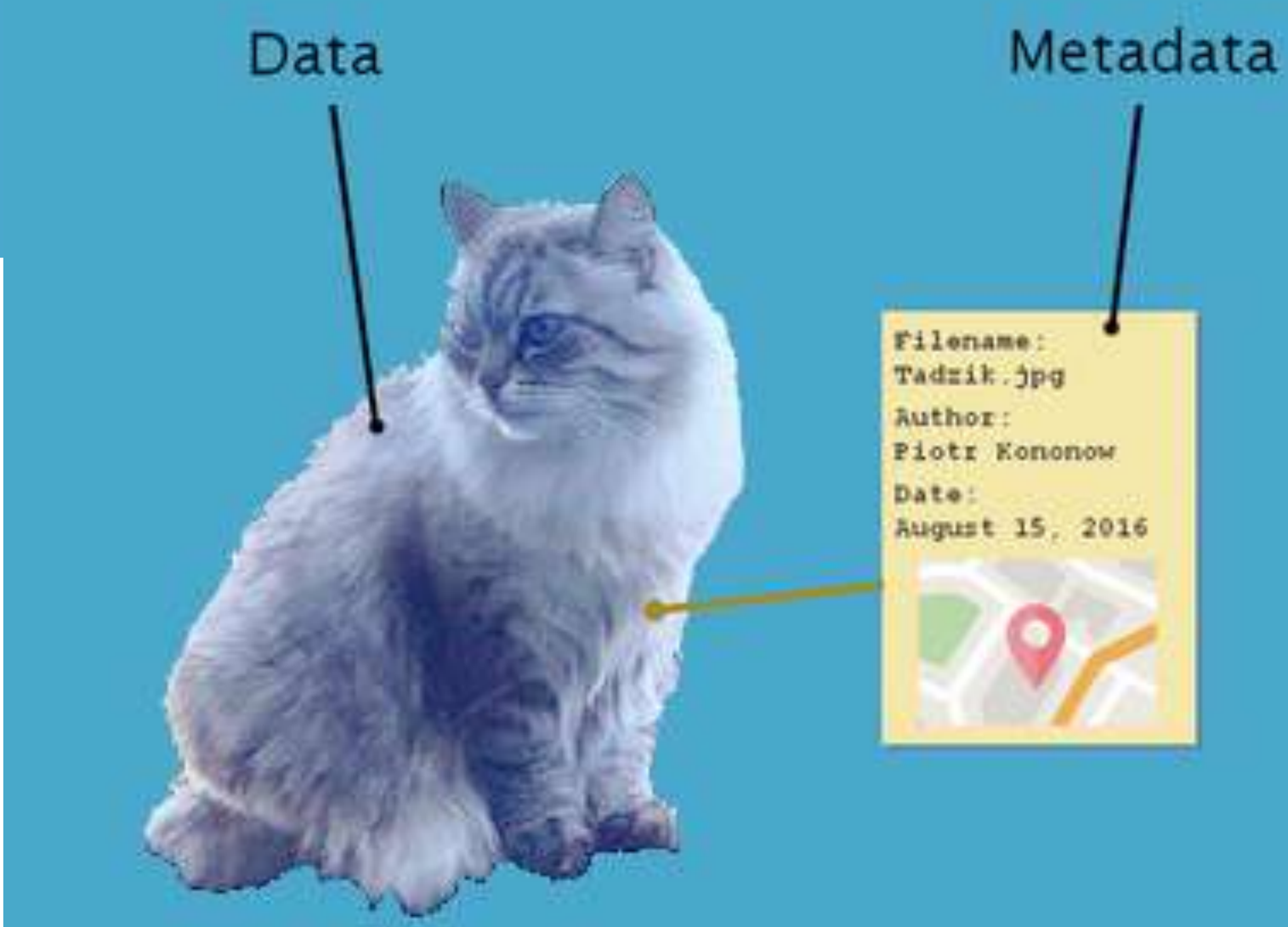


What is **Metadata**

Metadata is data about data

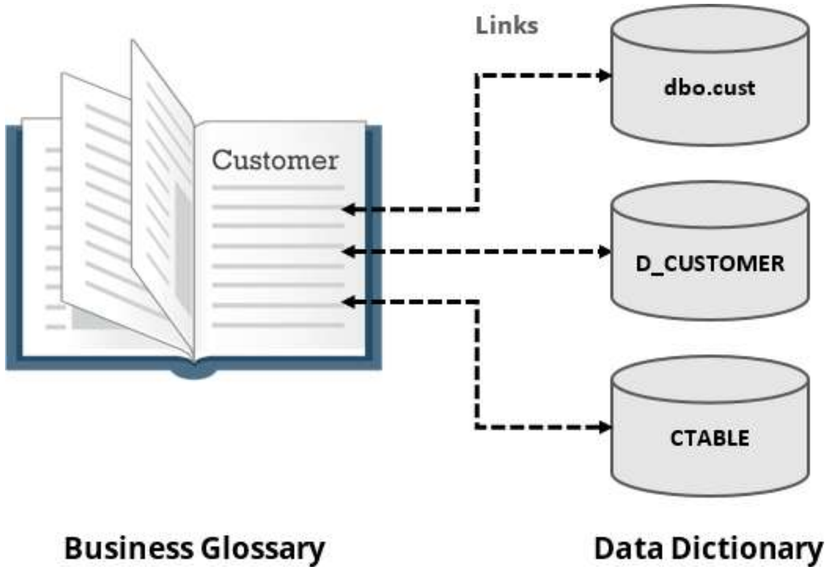
- Information about what sort of things a piece of data contains
- Describes other data
- Used for indexing, discovery & search
  - For example, it may say an entry is a diagnosis, but does not say what the diagnosis is
  - Useful for governance

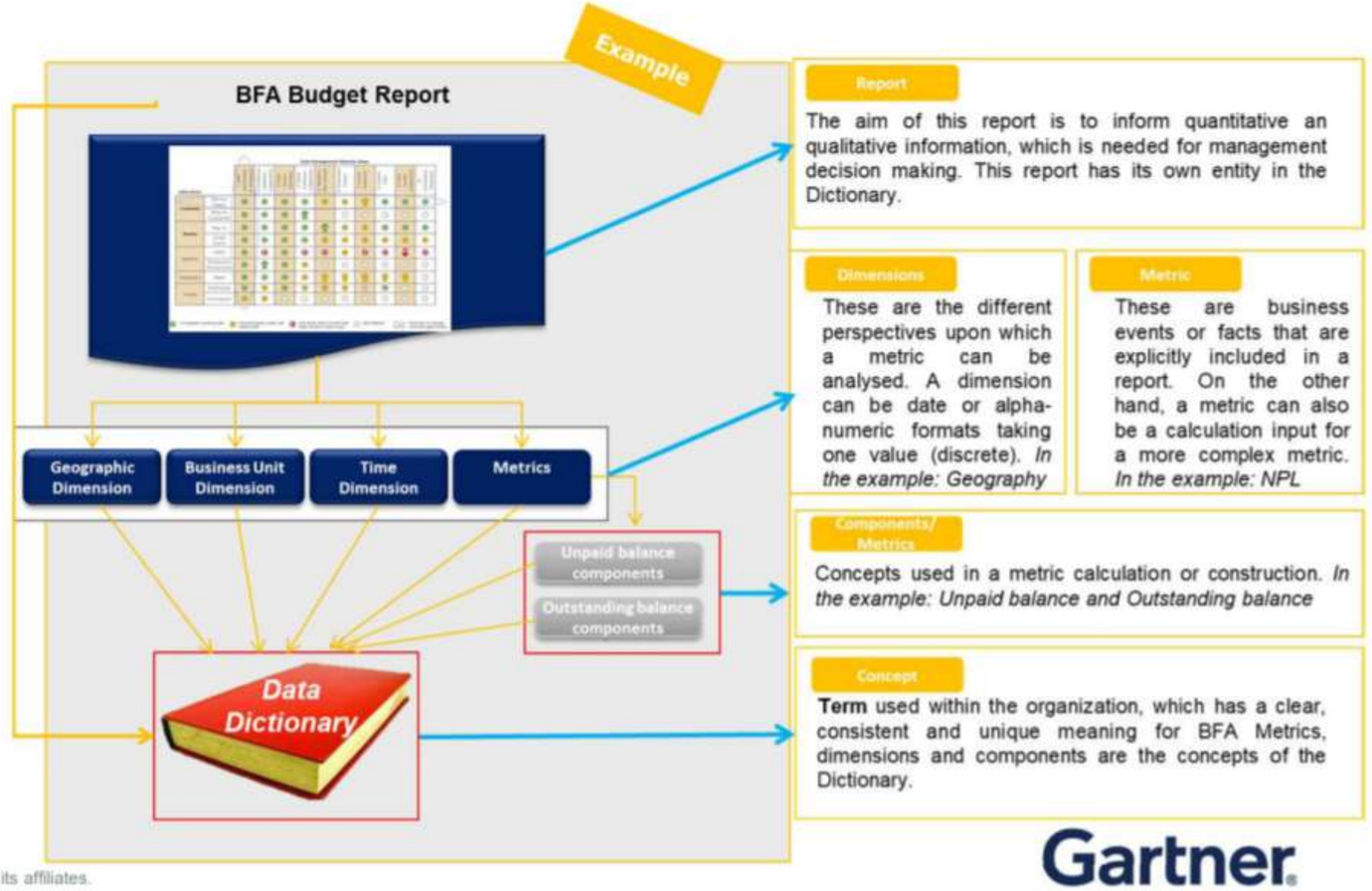
- It is **not** a summary or abstract
- Does **not** contain transaction data



Data **Catalog**, Data **Dictionary** and Business **Glossary**

	Summary	Scope	Owner
Business Glossary	Business Term Definitions	Enterprise or Division-Wide	Business
Data Dictionary	Technical Descriptions	Database-Specific	Technology
Data Catalog	Location Directory	Enterprise or Division-Wide	Technology



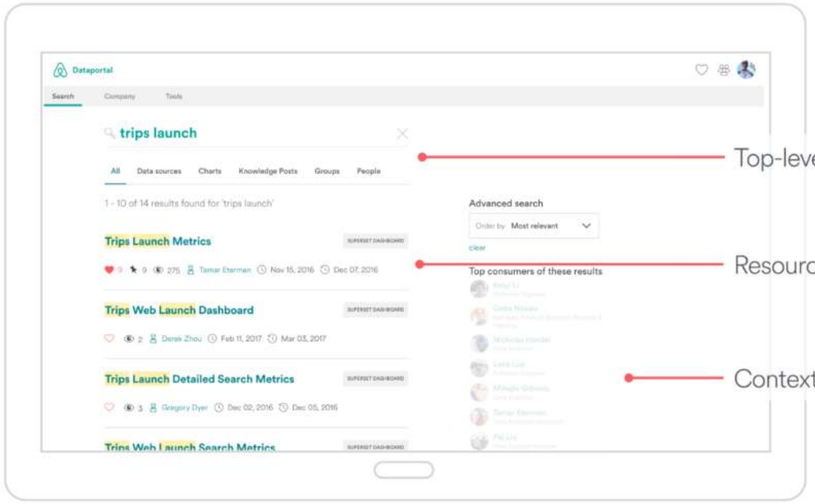


Gartner

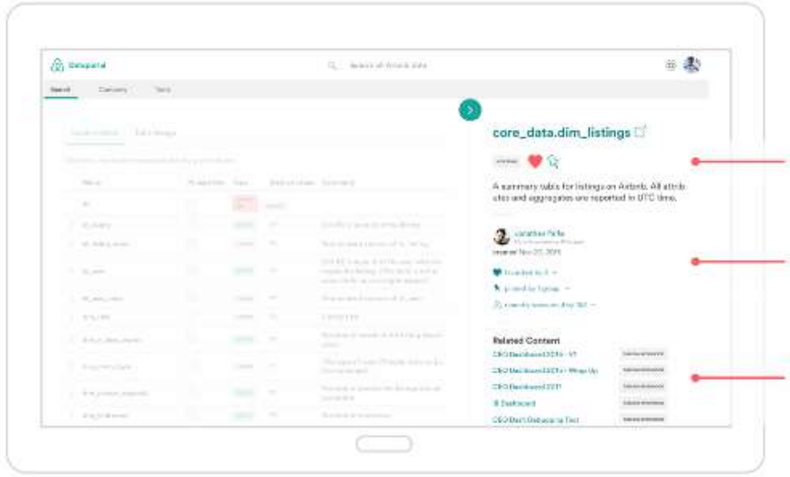
Airbnb Dataportal

Search all Airbnb **data**

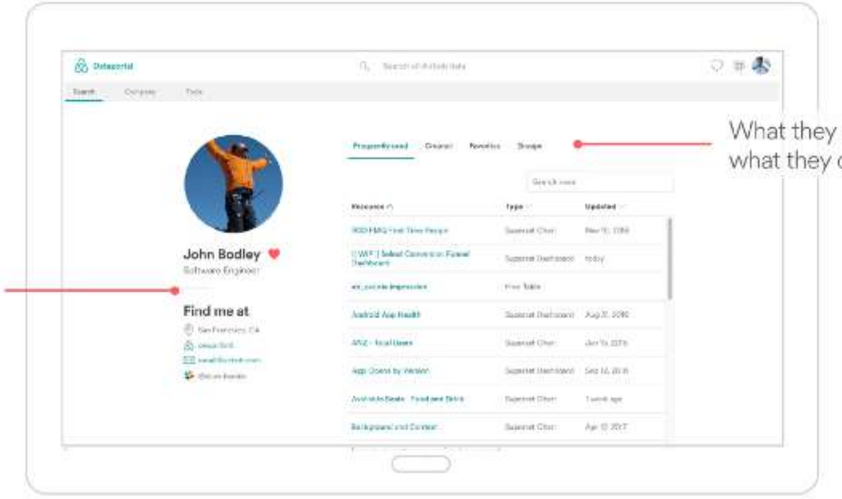
Search



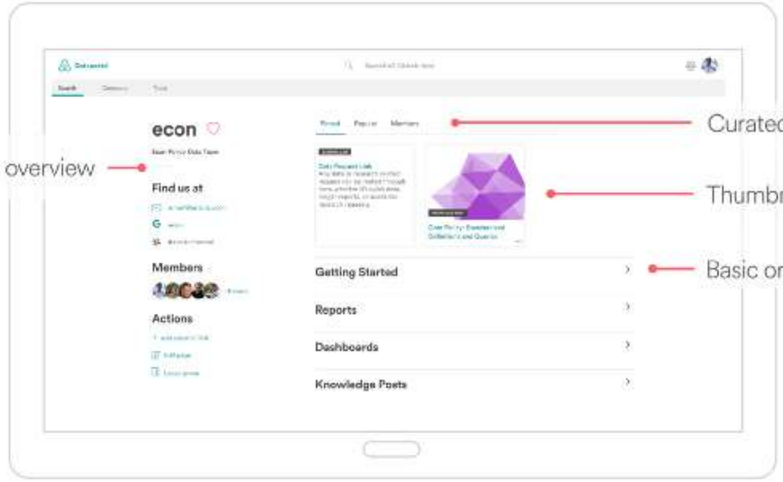
Context and metadata



Employee-centric Data

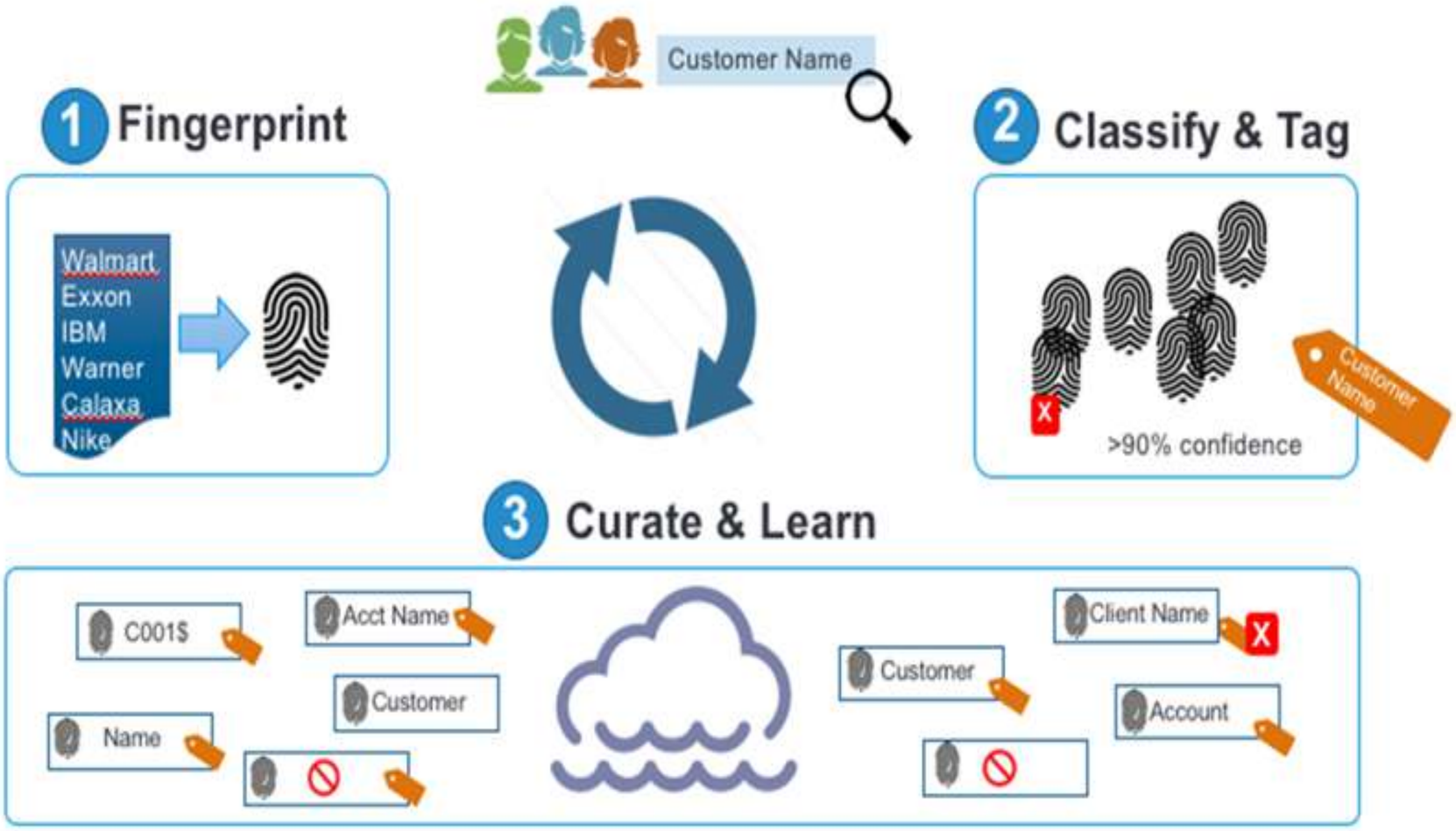


Team-centric Data



Automated **Data Detection**

One of the most exciting parts of the market today is the availability of **automated data and metadata profiling tools**. They are AI/ML tools that can be used to help discover the structure and content of incoming data.



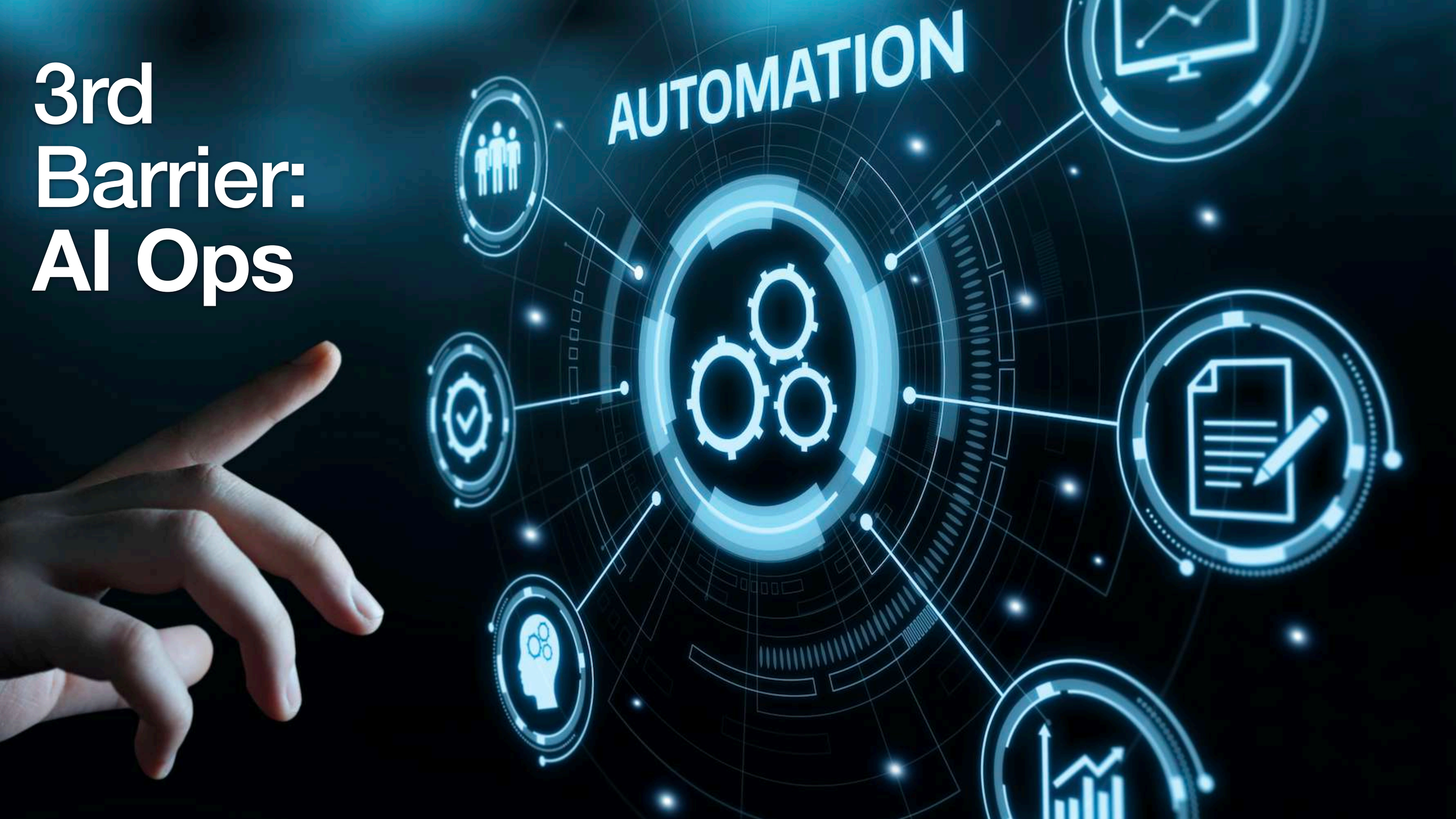
Data **Platform**

¿What is the data management platform (**DMP**) used for collecting and managing data? ¿It allows to use big data and artificial intelligence algorithms to process and analyze large data sets? ¿Can it grow and acquire new functionalities as the data grows and the business needs evolve?

Data **Governance**

¿What are the processes and policies in place that ensure that high **data quality** exists through its complete lifecycle? ¿What data **controls** there are that support business objectives? ¿It is the data **available**, **usable**, **consistent** and **secure**?

3rd Barrier: AI Ops



The **AI strategy** aims to obtain the maximum potential of this technology for our businesses by **solving all the barriers** to its implementation.

1 Business Value

Usage: What are the main drivers of AI value in each business?
AI ambition: How much do we believe AI can transform our business?
AI maturity: How innovative in the use of AI does the company need to be?

2 Data Access

Types of sources: What data will be needed at the company to develop AI?
Data Platform: What technological tools will we use to facilitate access to data at the company?
Data Governance: How are we going to ensure that they are of the right quality?

3 AI Operations

Organization: Where should we locate the AI specialists at the company?
Sourcing Model: Who develops and maintains the AI models?
Process industrialization: How efficient and scalable is AI development, deployment and maintenance?

4 AI Culture

Skills training & hiring: How can we train the company staff? How do we attract & retain talent?
Communication: how do we make it easier for everyone to understand what AI brings to the company?
Government: How do we guarantee the proper use of AI (security, privacy, ethics) at the company?

5

Are the legal/regulatory issues sufficiently resolved?

6

Is society ready to accept the transformation of the value proposition?

7

It may be legal, profitable and working - but is it ethical?

Why do 87% of data science projects never make it into production?

Venturebeat.com

But if this is a universal understanding, that AI empirically provides a competitive edge, **why do only 13% of data science projects**, or just one out of every 10, actually **make it into production**?

“One of the biggest [reasons] is sometimes **people think**, all I need to do is **throw money at a problem or put a technology in**, and **success comes out the other end**, and **that just doesn’t happen**,” Chapo said. “And we’re not doing it because we **don’t have the right leadership support,** to make sure we **create the conditions for success**.”

The other key player in the whodunit is **data**, Leff adds, which is a double edged sword — it’s what makes all of these analytics and capabilities possible, but **most organizations are highly siloed**, with owners who are simply not collaborating and leaders who are not facilitating communication.

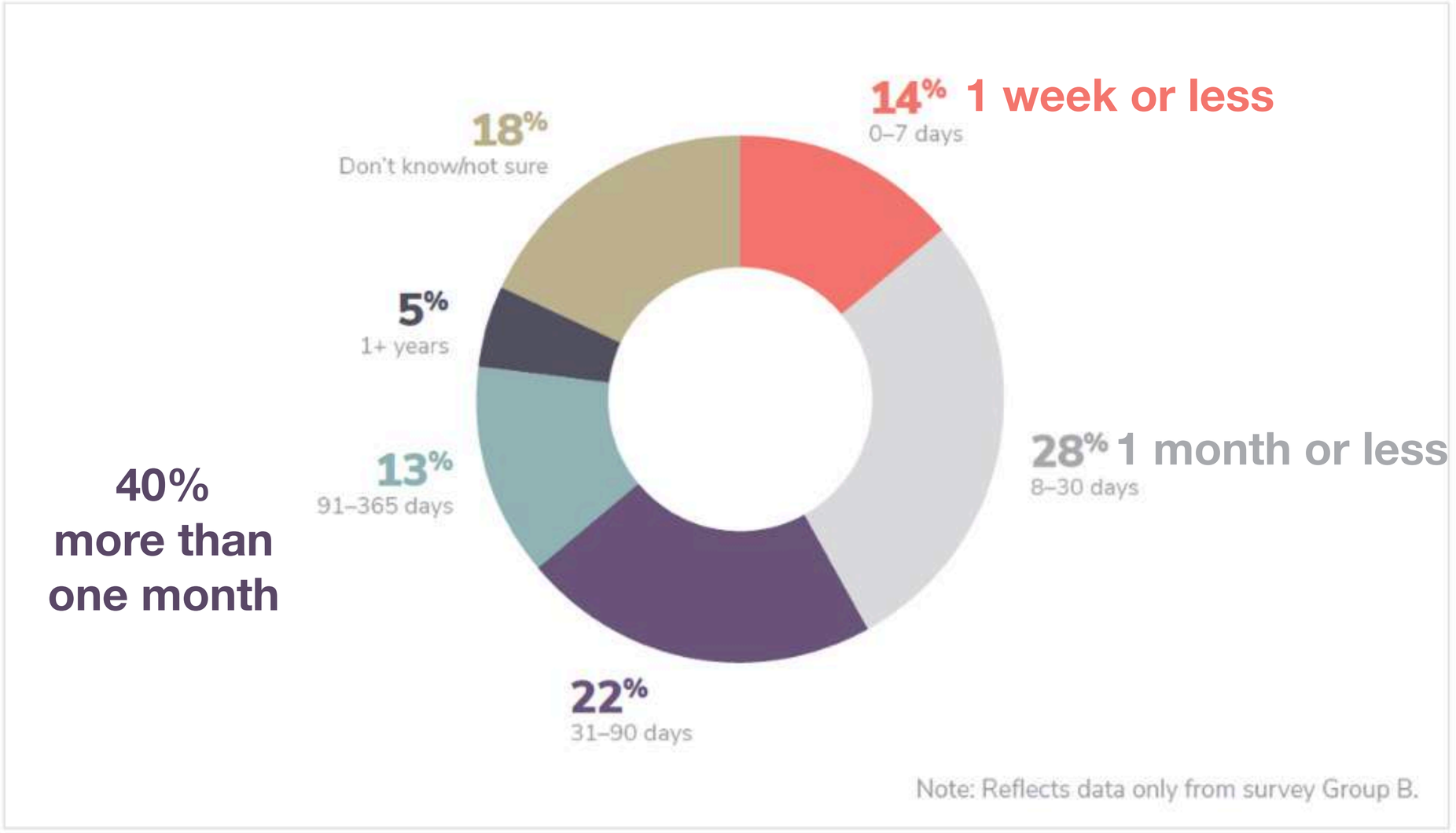
“I’ve had **data scientists** look me in the face and say **we could do that project**, but we **can’t get access to the data**,” Leff says. “And I say, **your management allows that to go on**?”

And the third issue, intimately connected to those silos, is the **lack of collaboration.** Data scientists have been around since the 1950s — and they were individuals sitting in a basement working behind a terminal. But **now that it’s a team sport**, and the importance of that work is now being embedded into the fabric of the company, it’s essential that every person on the team is able to collaborate with everyone else

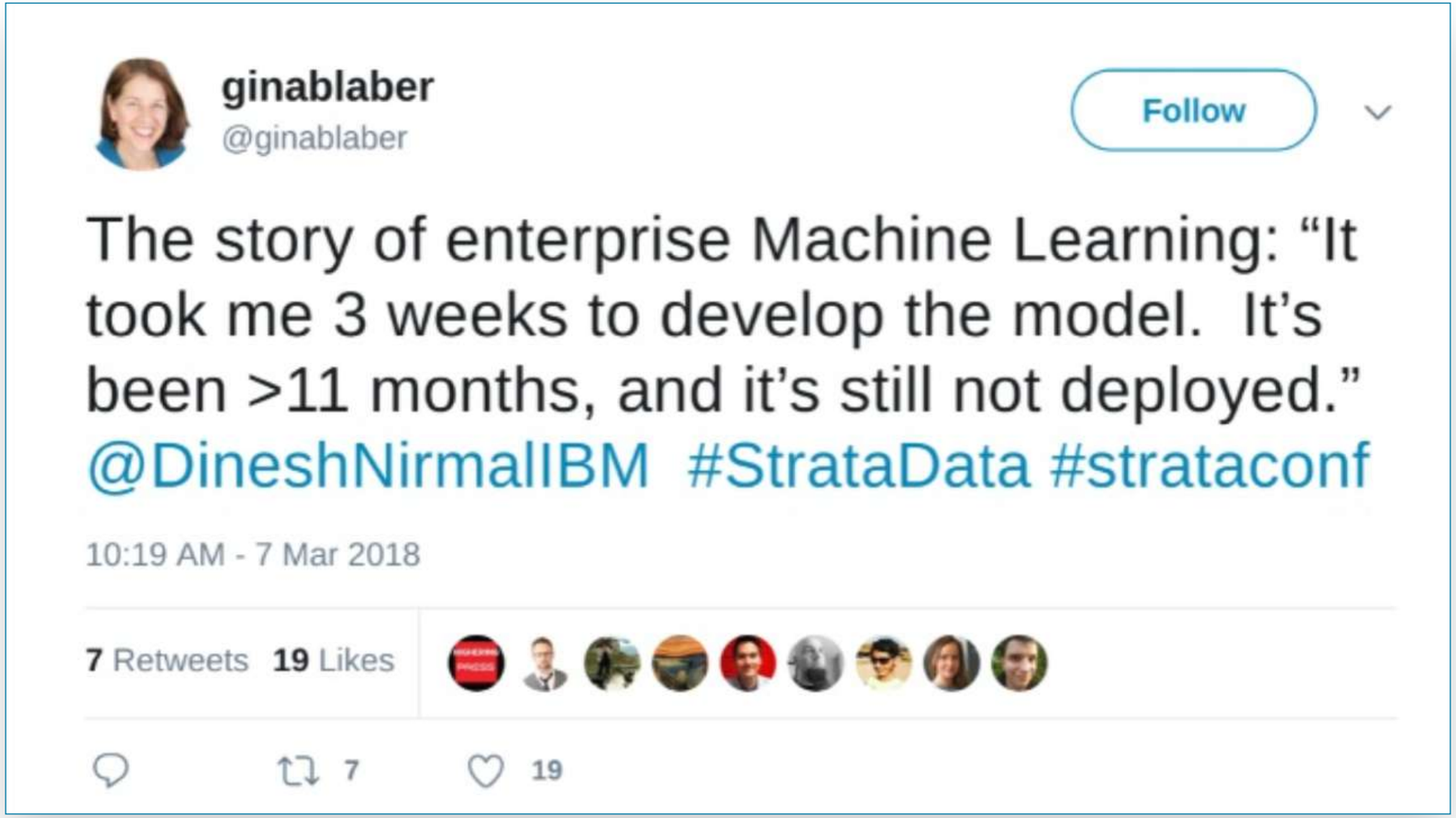
Add It Up: How Long Does a Machine Learning Deployment Take?

Lawrence E Hecht

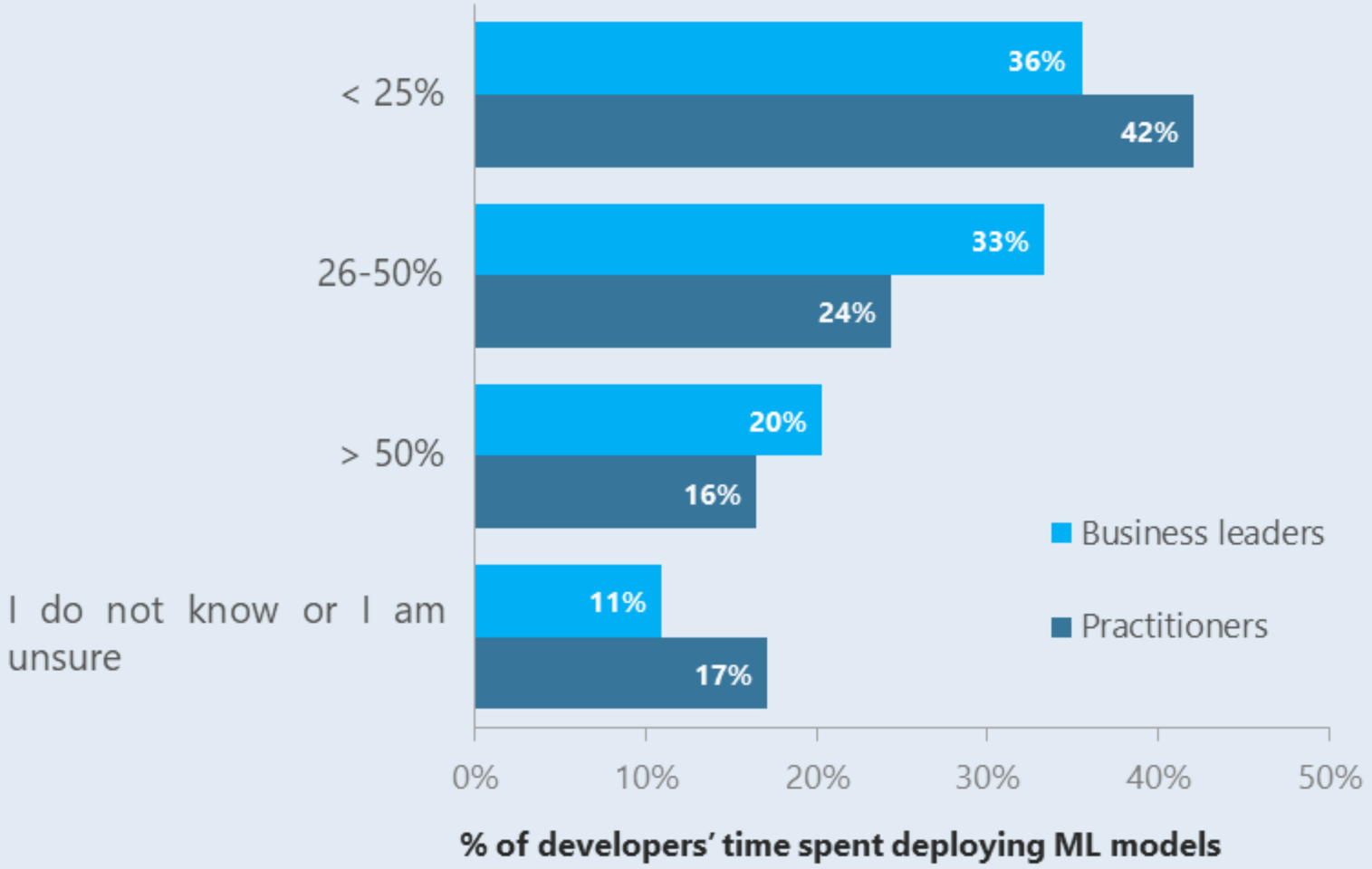
Machine learning model deployment timeline



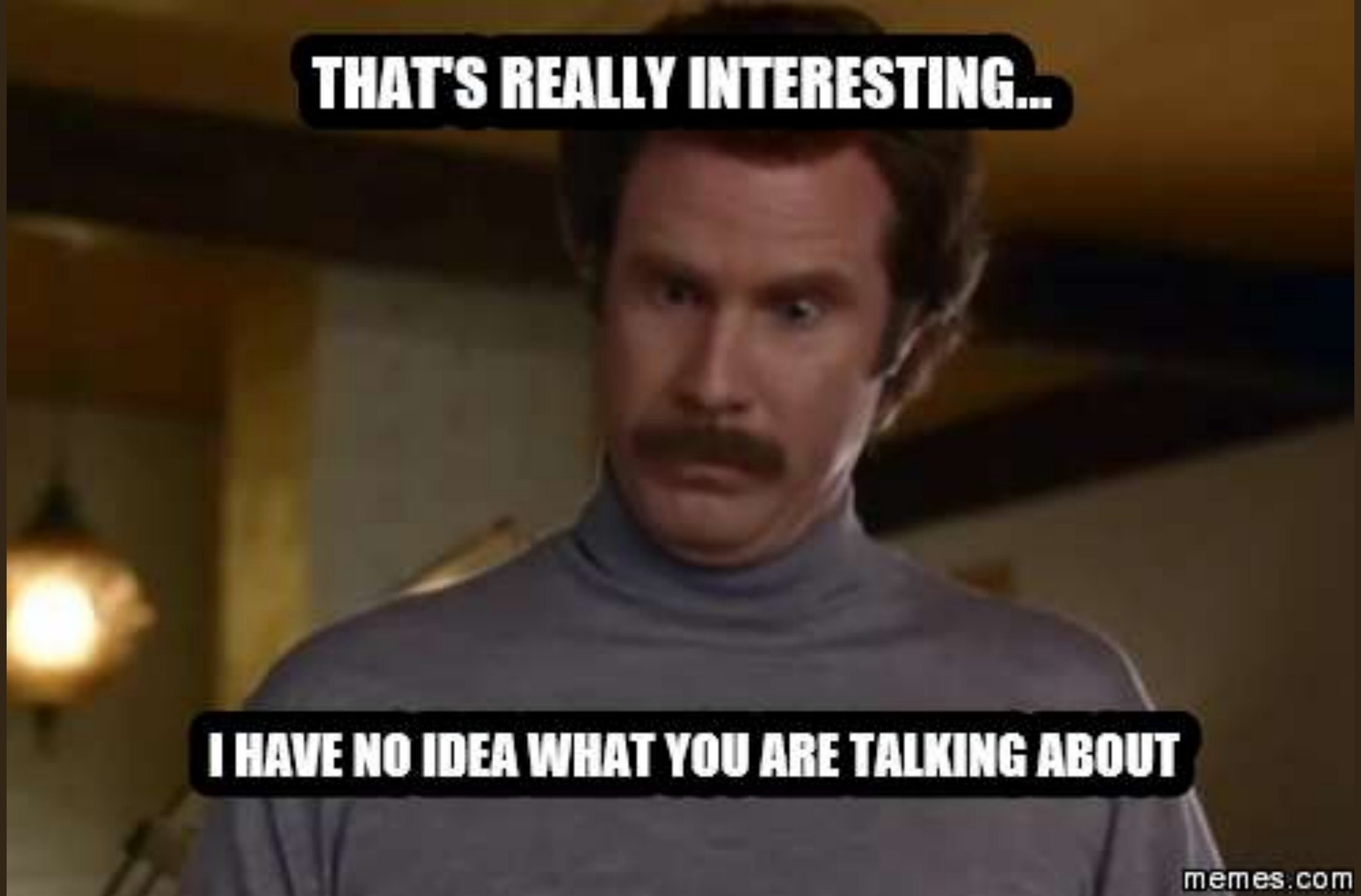
Source: Algorithmia's "2020 State of Enterprise ML". This question about how long it takes to deploy an ML model into production was only asked to a subset of respondents at a company that has an ML model production.

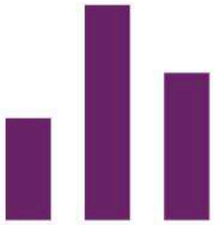

ginablaber
@ginablaber
Follow
The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf
10:19 AM - 7 Mar 2018
7 Retweets 19 Likes
7
19

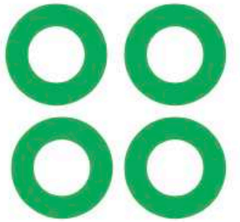
54% of Business Leaders Think Developers Spend >25% of Their Time Deploying ML Models

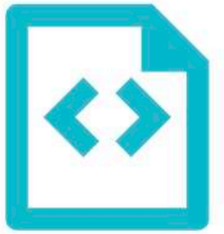


Source: Algorithmia's "2020 State of Enterprise Machine Learning". 745 respond people took the survey in October 2019. Q. What percentage of your developers' time is spent deploying ML models? Approximately 40% of were practitioners with job titles like data scientist and research engineer.


THAT'S REALLY INTERESTING...
I HAVE NO IDEA WHAT YOU ARE TALKING ABOUT
memes.com







Data

Schema

Sampling over Time

Volume

Model

Algorithms

More Training

Experiments

Hyperparameters

Code

Business Needs

Bug Fixes

Configuration

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

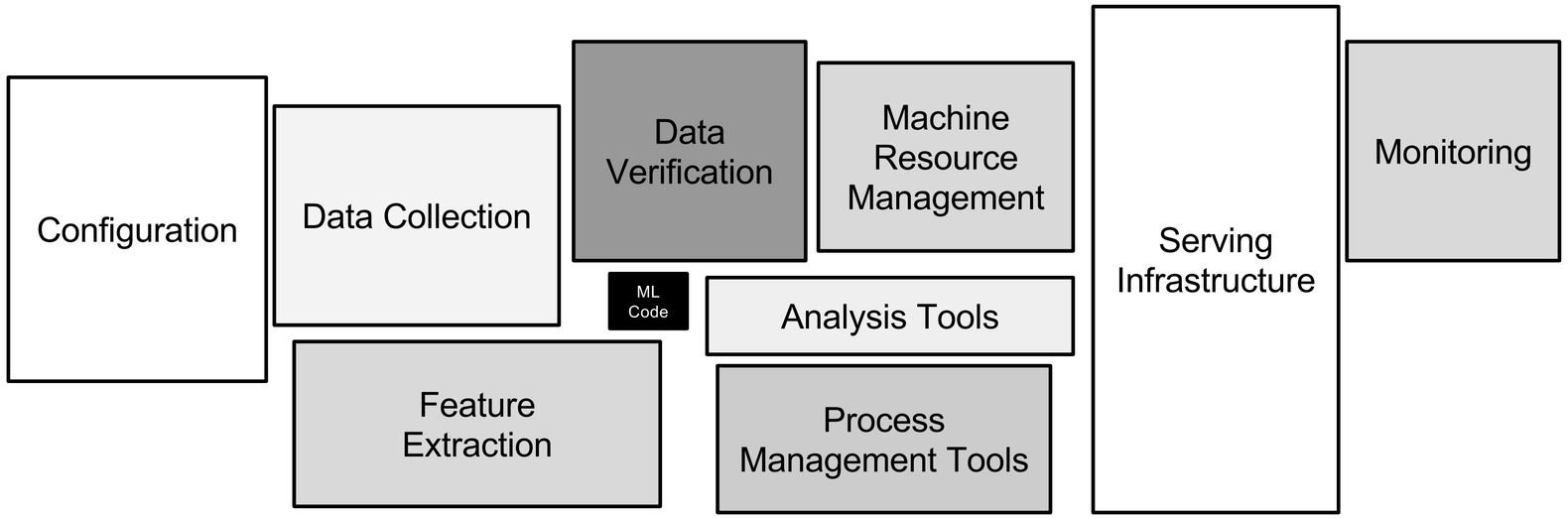
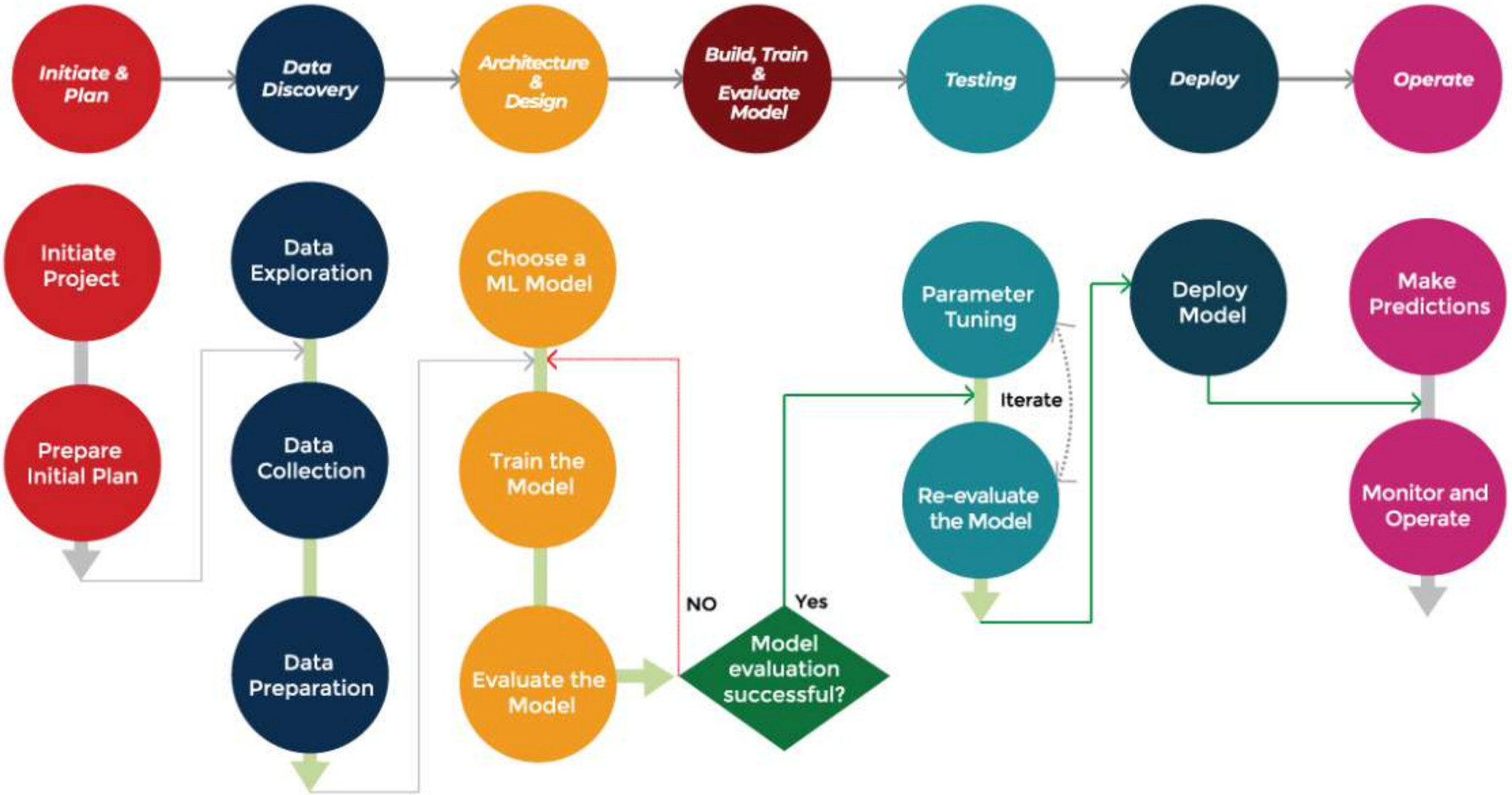
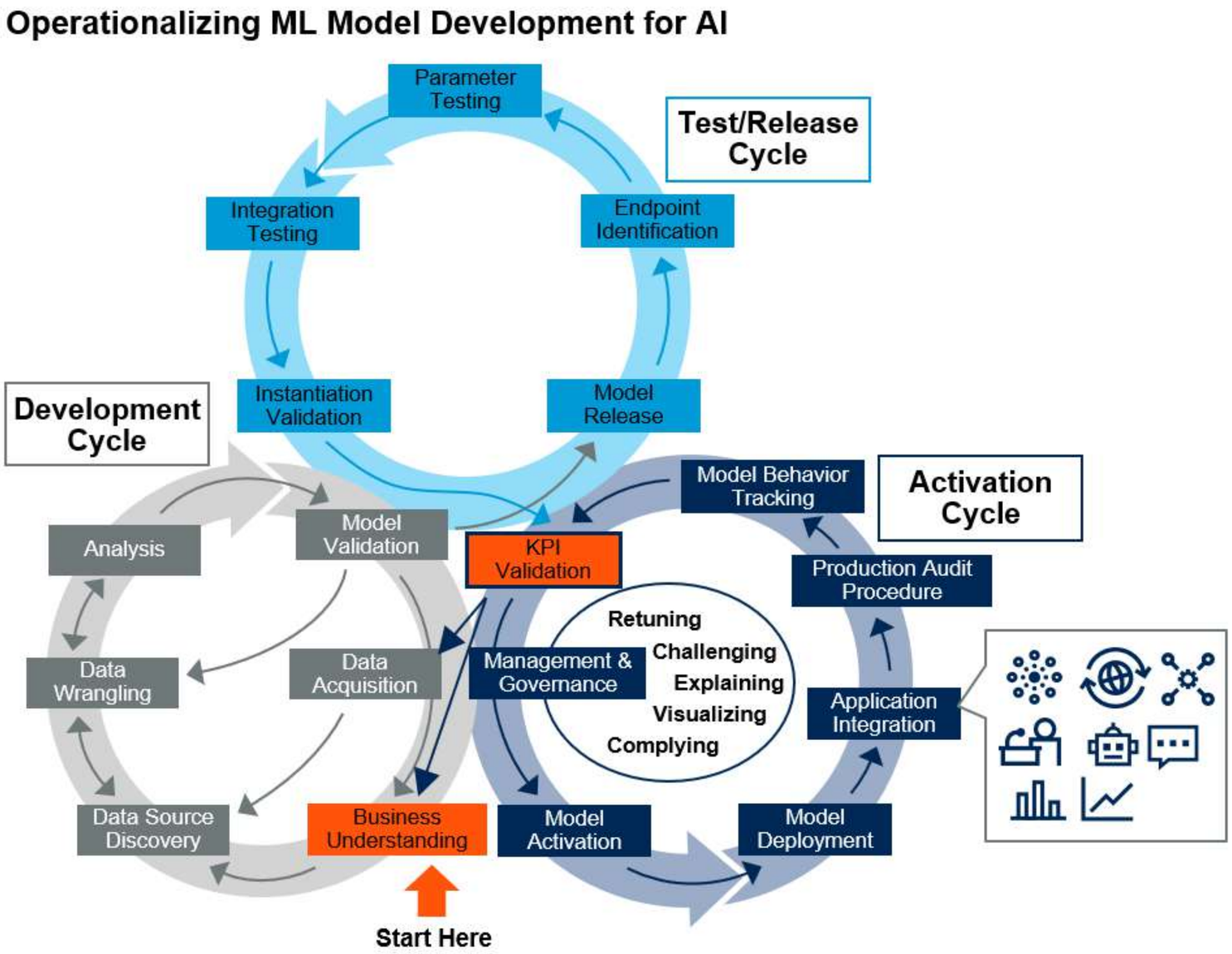


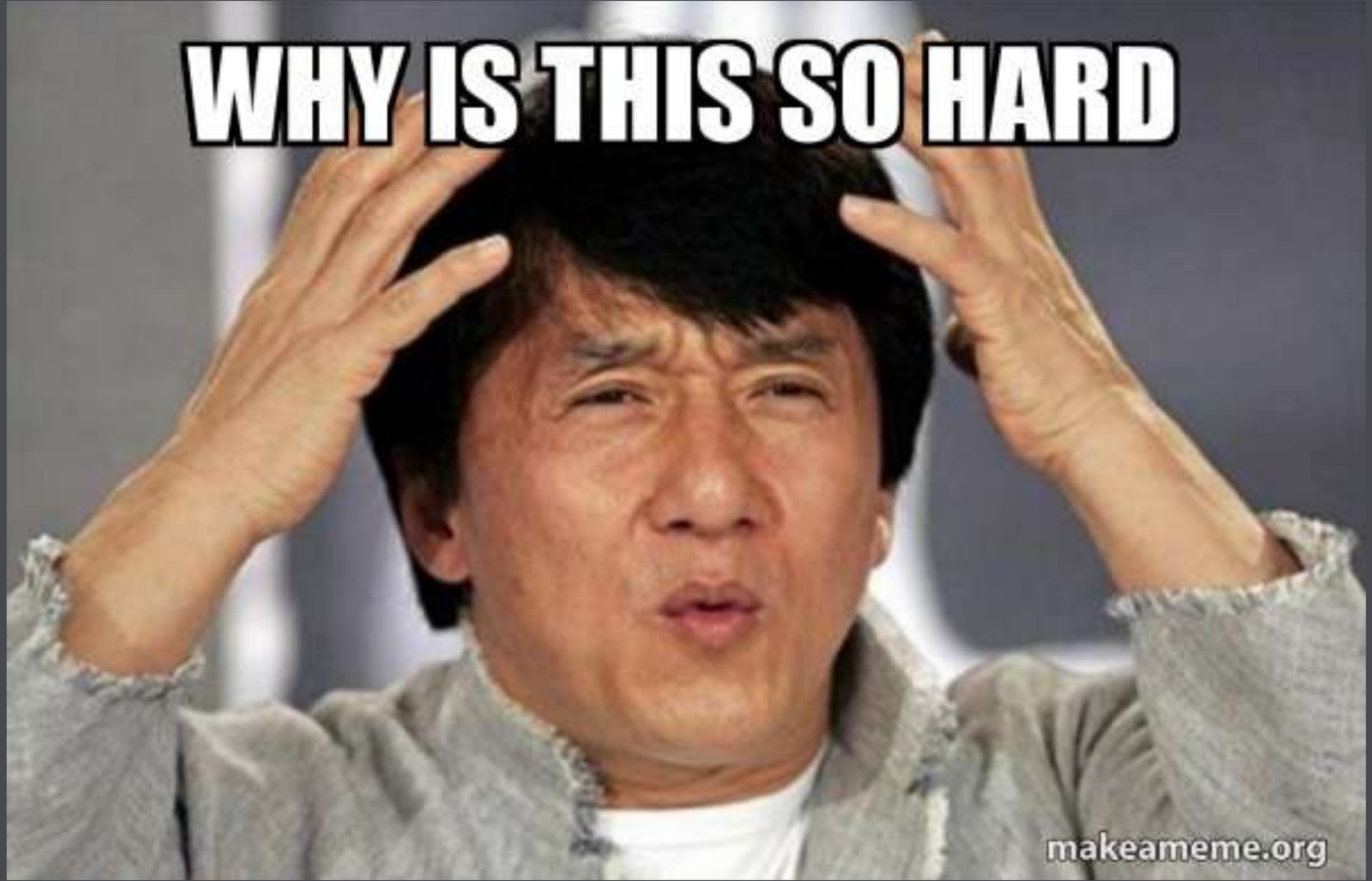
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

MLOps Pipeline

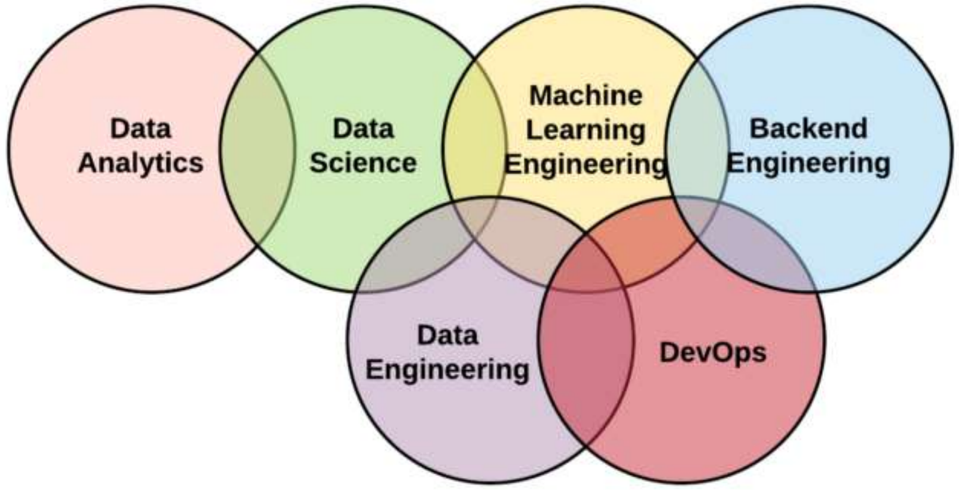


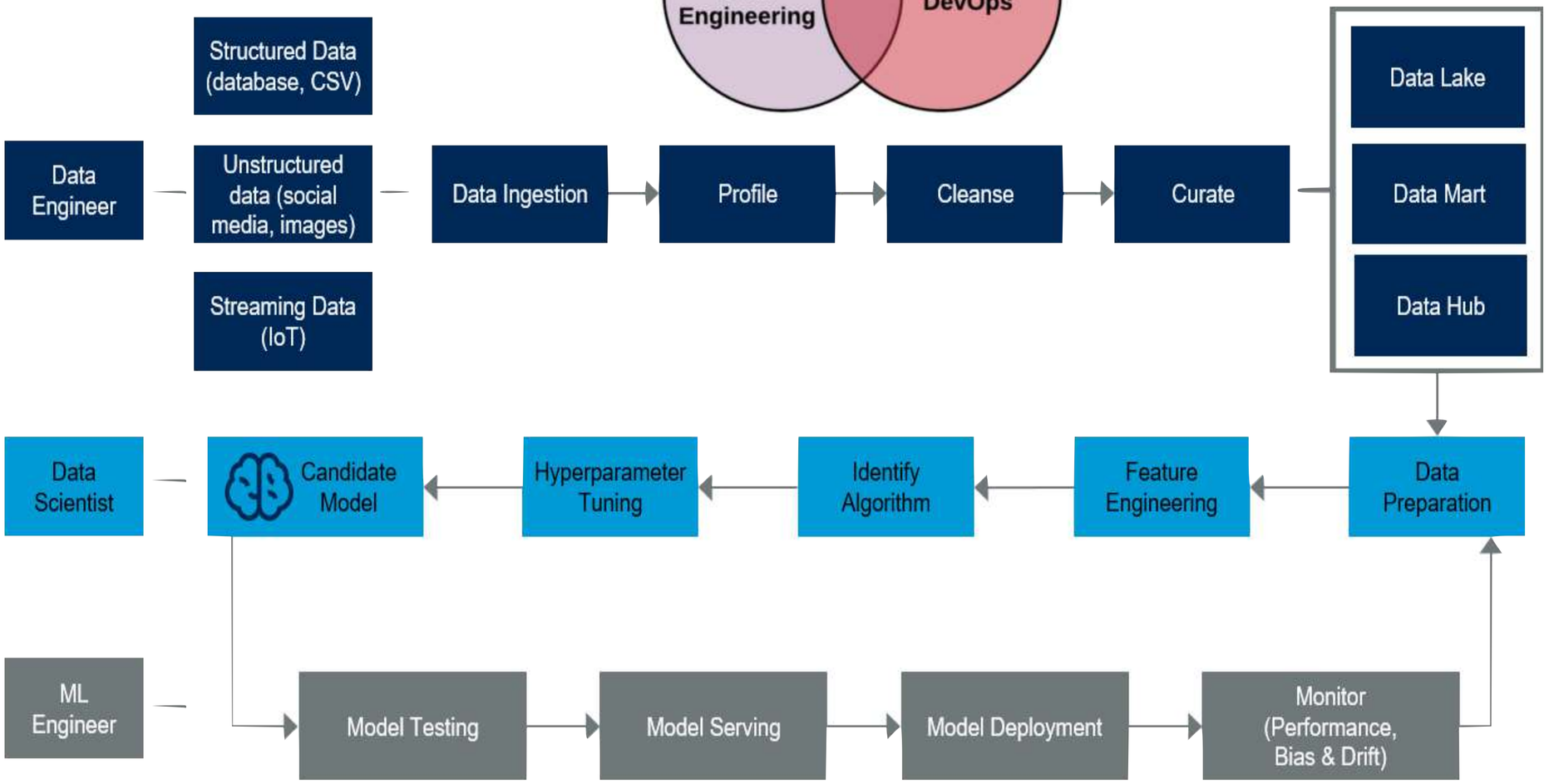
MLOps Pipeline




WHY IS THIS SO HARD
makeameme.org

Who
does what?





Communicatio

Busines

gement

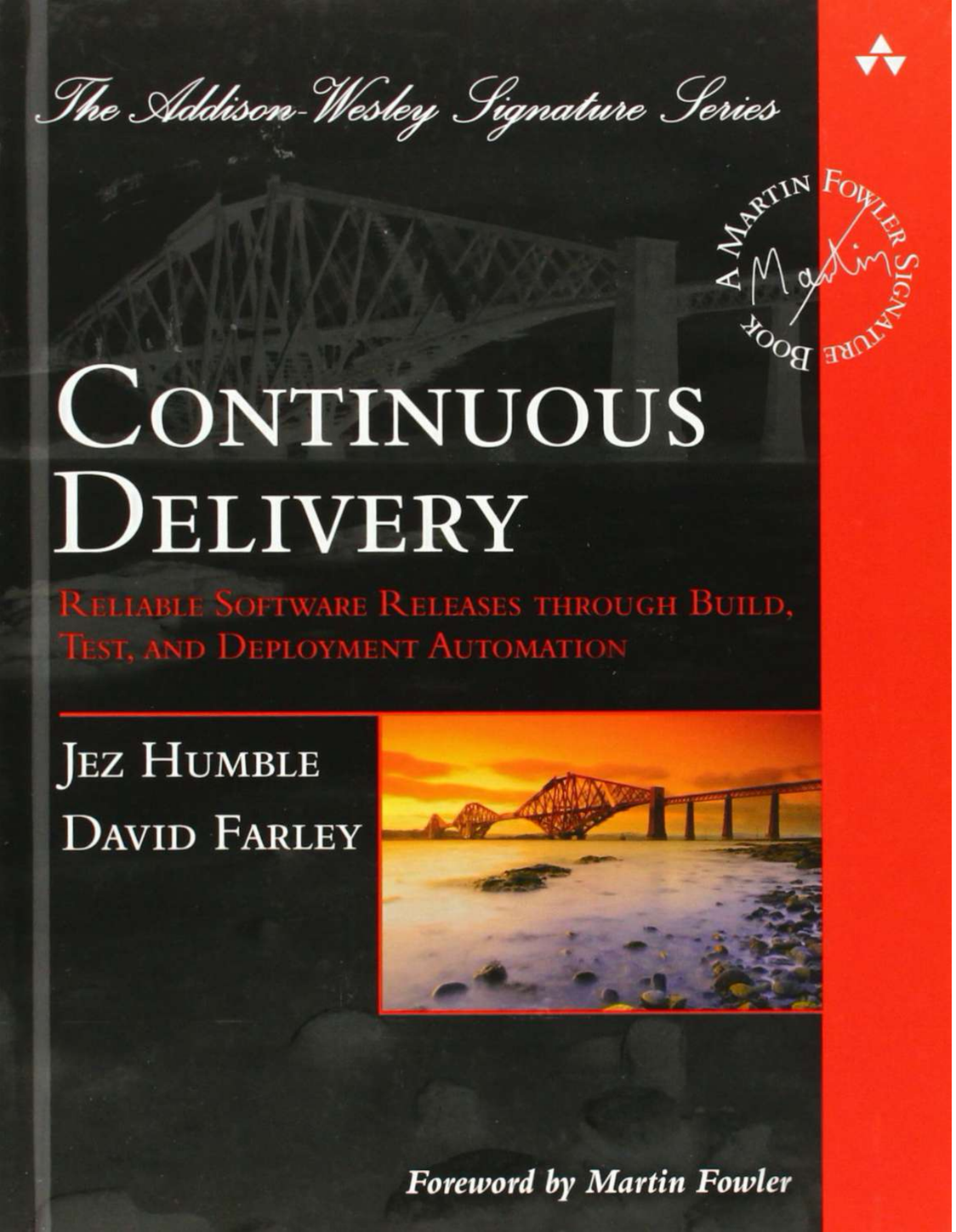
David G Pisano Retweeted

Florian Leitner @flowing 3 Jan 2019

Couldn't second this more; The hard part is finding engineers who can also build ML models. I believe pure data scientists who cannot engineer will only succeed in large companies, as they can more likely make use of employees who exclusively train models 24/7.

> **Emmanuel Ameisen** @mlpowered
>
> 2018 has been a continuing flurry of exciting work in Machine Learning. If you are interested in being part of the field in 2019, I've written about how some of the most impactful trends of 2018 will impact this year! medium.com/@emmanuelameis... pic.twitter.com/6aLQ63XNEs

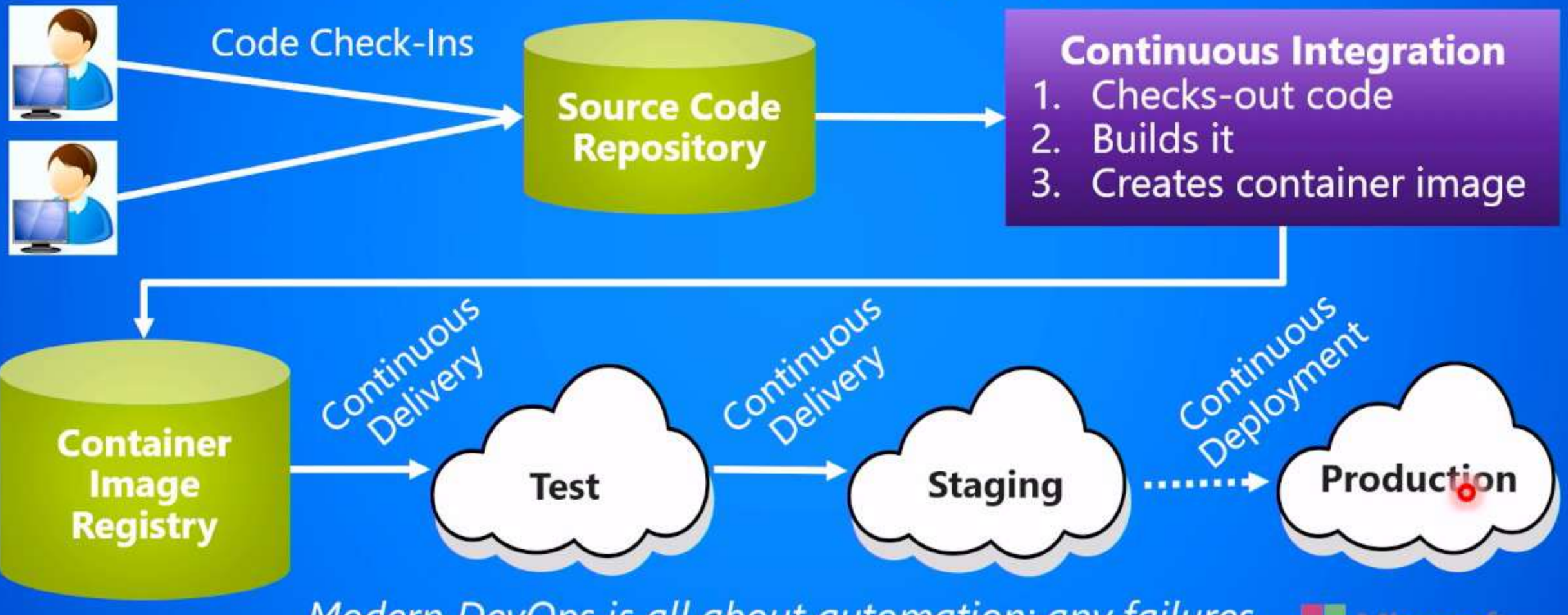
1 2



"Continuous Delivery is the ***ability to get changes of all types*** *— including new features, configuration changes, bug fixes, and experiments —* ***into production****, or into the hands of users,* ***safely and quickly*** *in a* ***sustainable way****"*

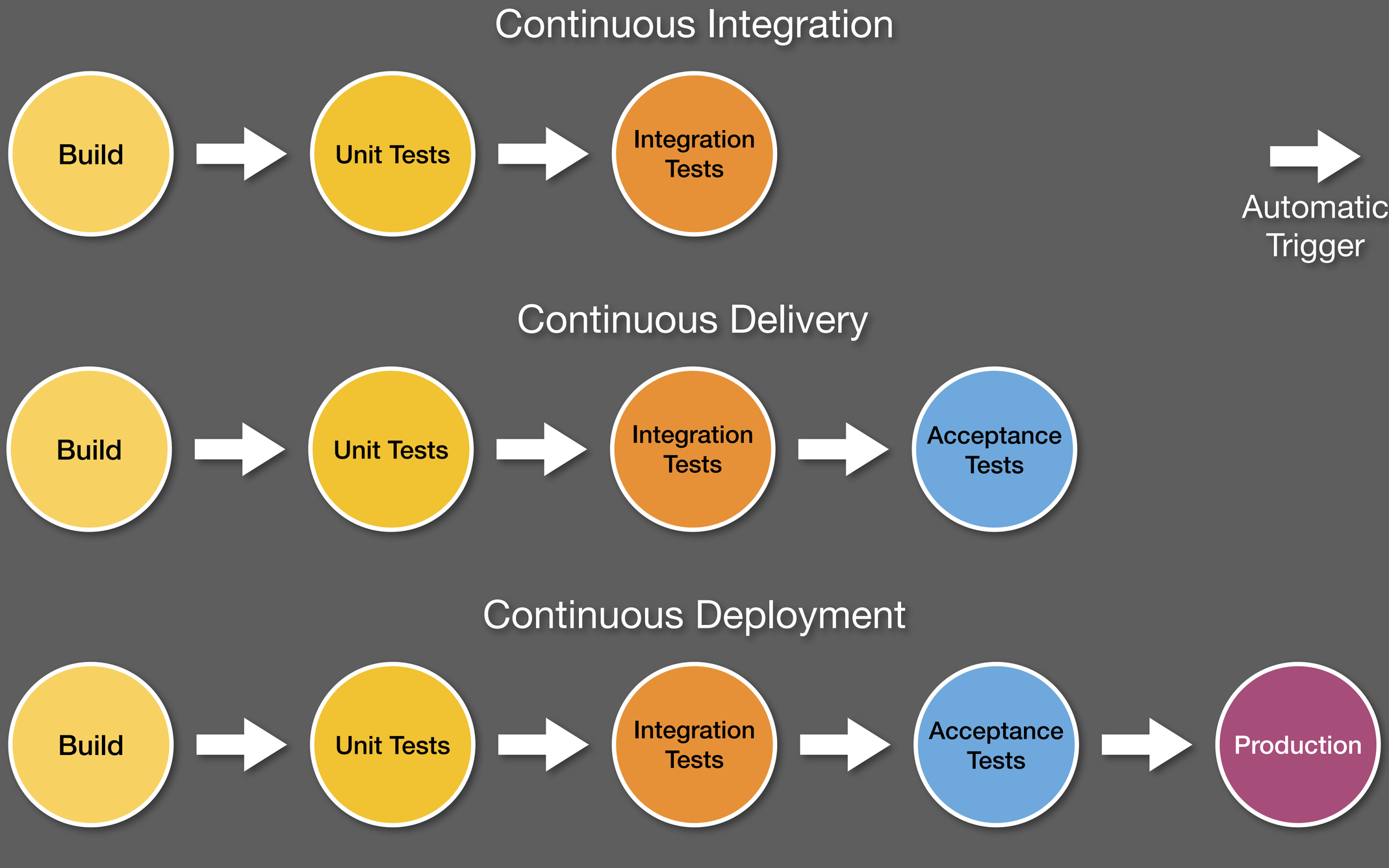
Jeff Humble and Dave Farley

CI: Continuous Integration
CD: Continuous Delivery, & Deployment



Modern DevOps is all about automation; any failures are sent back to developers & stops further progress

Microsoft

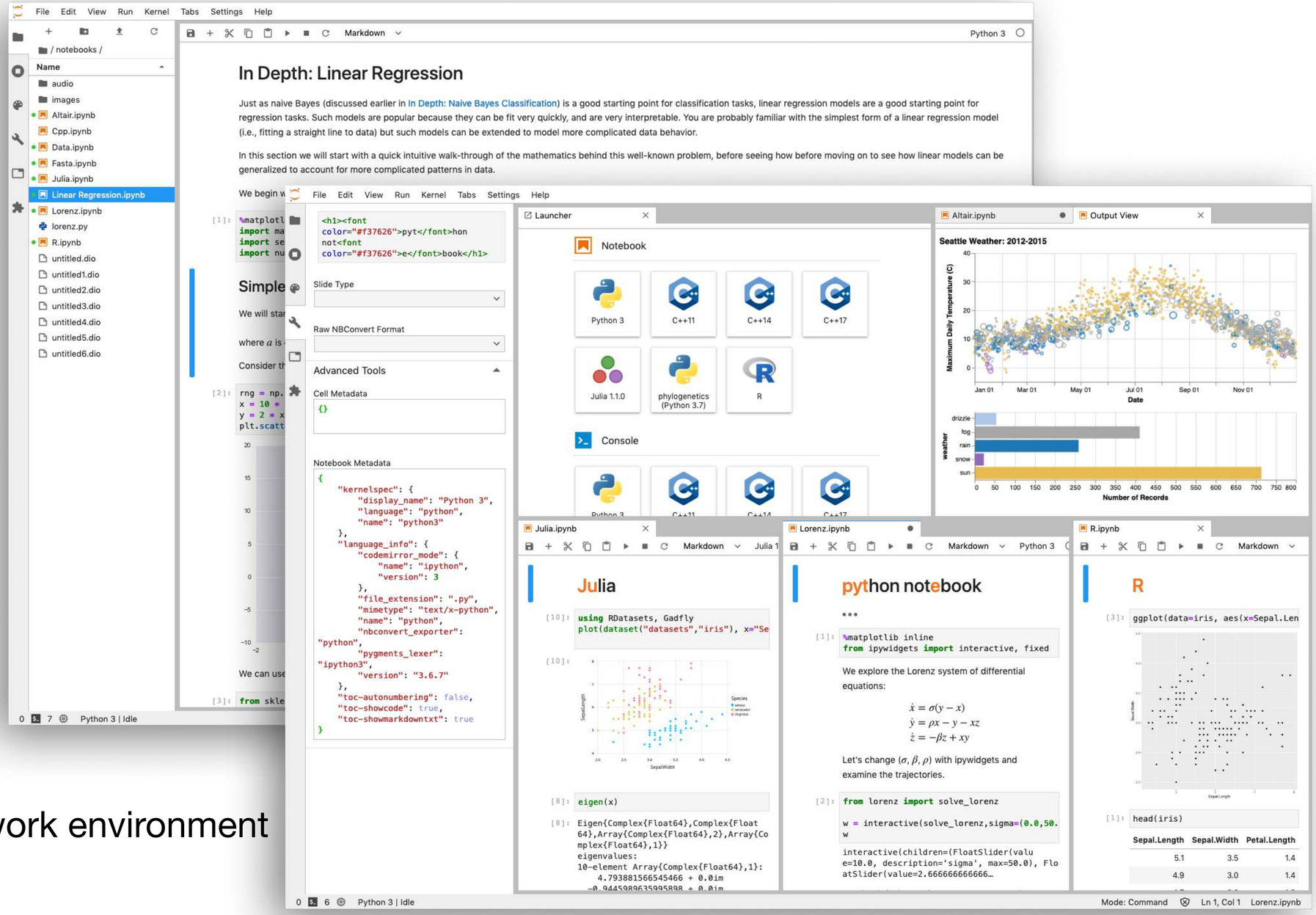

Continuous Integration
Build
Unit Tests
Integration Tests
Automatic Trigger
Continuous Delivery
Build
Unit Tests
Integration Tests
Acceptance Tests
Continuous Deployment
Build
Unit Tests
Integration Tests
Acceptance Tests
Production

DATA Engineer

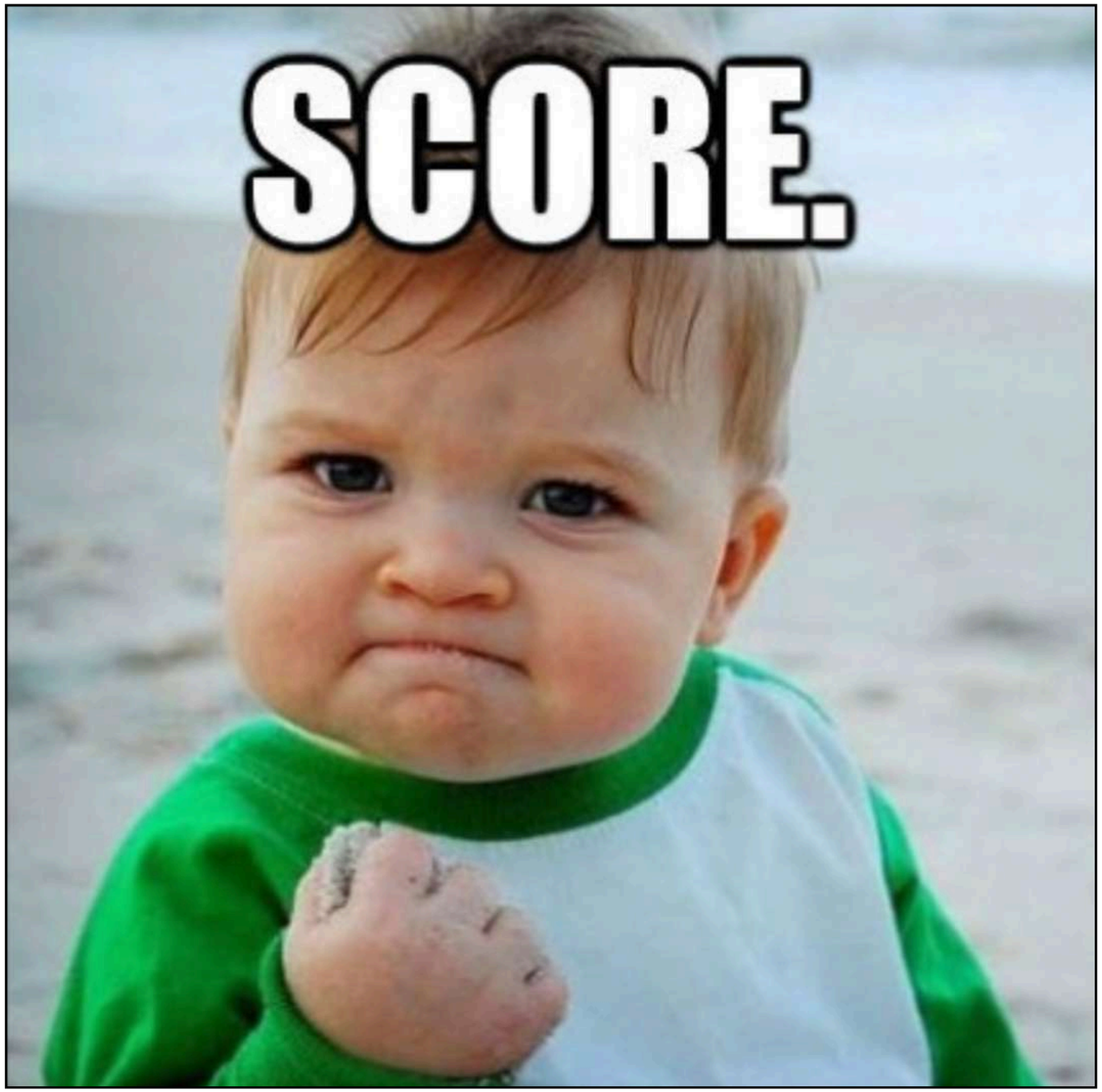
The GAP

DATA Scientist

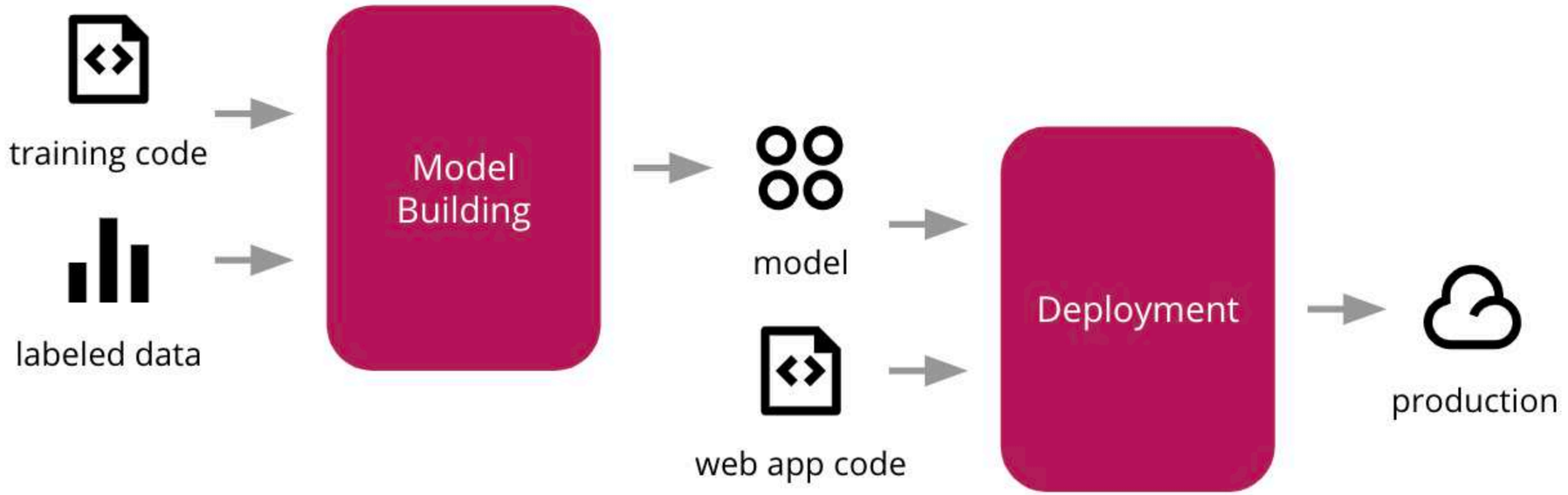
Data Scientists just want to Data Science

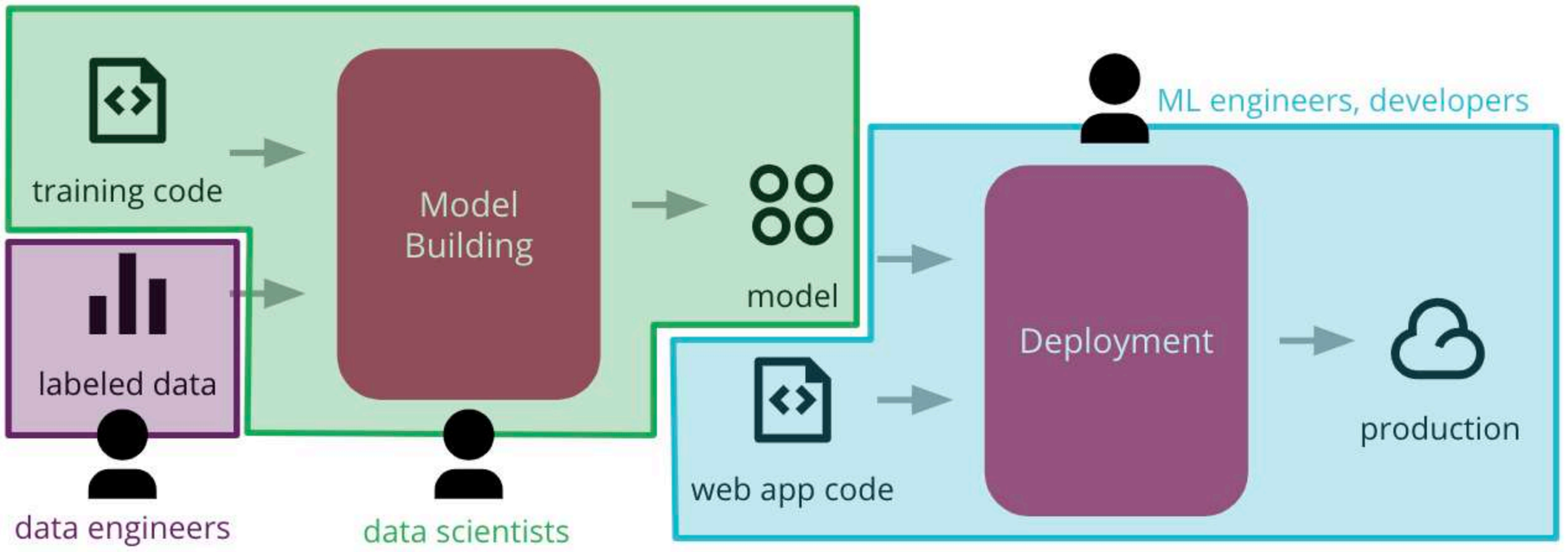


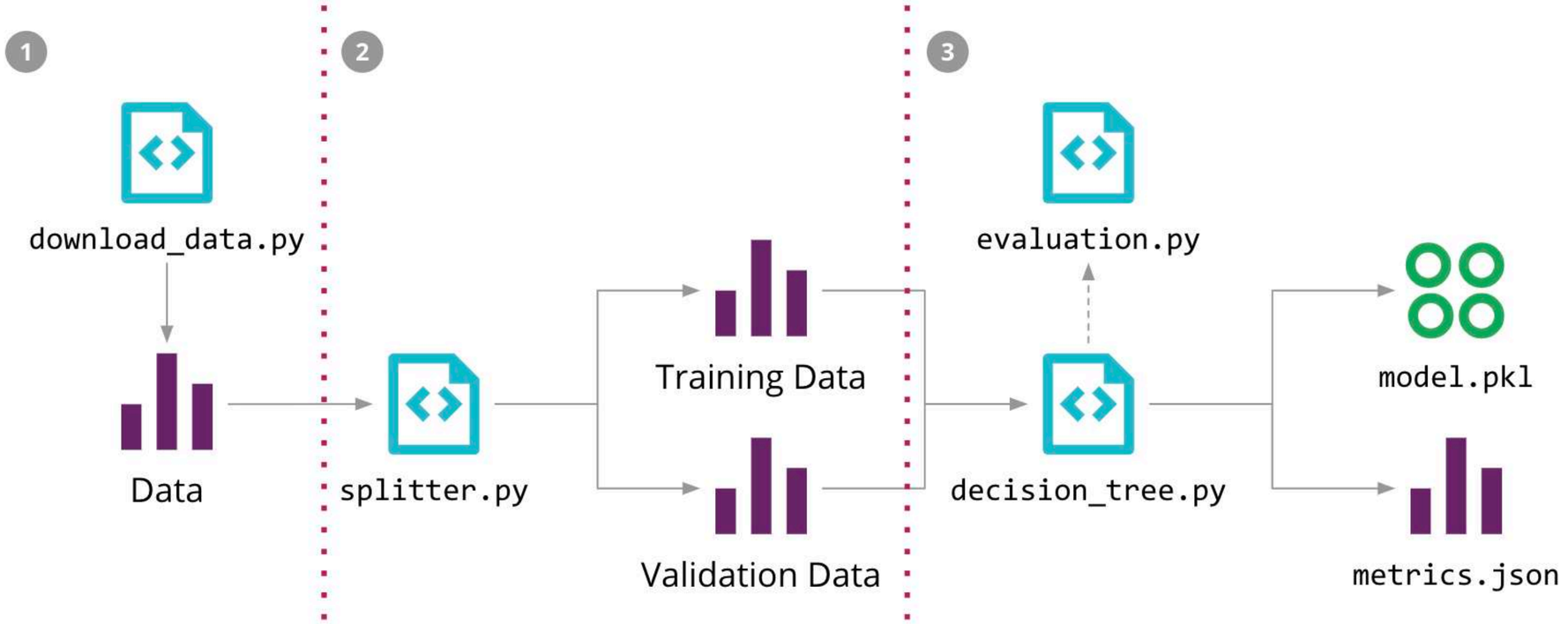
Typical data scientist work environment



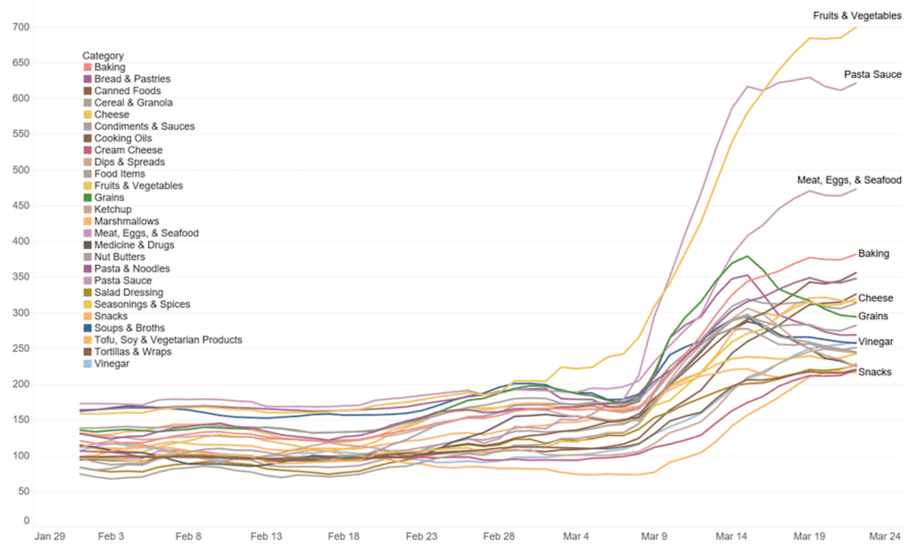
We’ve got the notebook into source control!


training code
labeled data
Model Building
model
web app code
Deployment
production


training code
labeled data
Model Building
model
data engineers
data scientists
ML engineers, developers
web app code
Deployment
production


1
2
3
download_data.py
Data
splitter.py
Training Data
Validation Data
evaluation.py
decision_tree.py
model.pkl
metrics.json

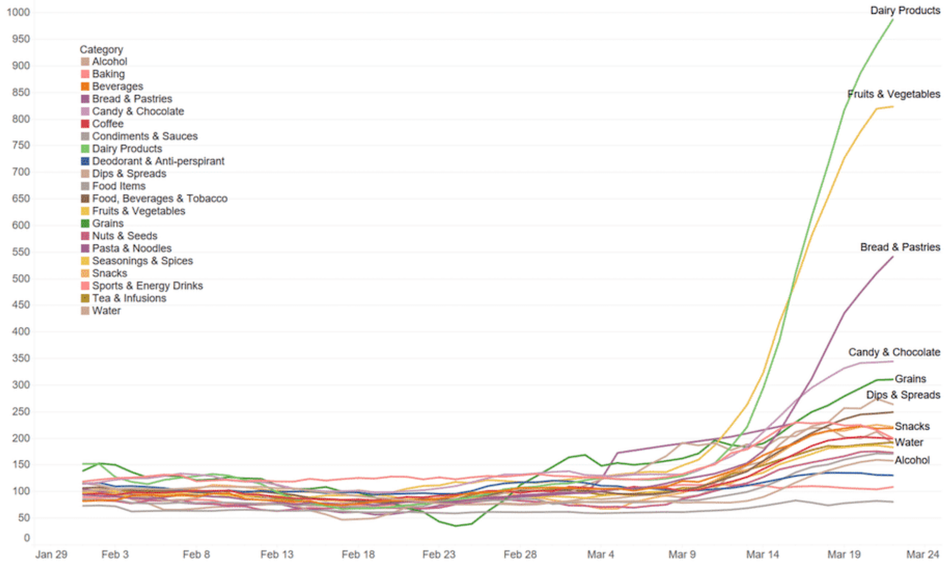
Consumers Now Using Online Providers For All Grocery Needs



Indexed sales by category compared to Average in Jan 1-28.
At least 5 retailers per category

criteo

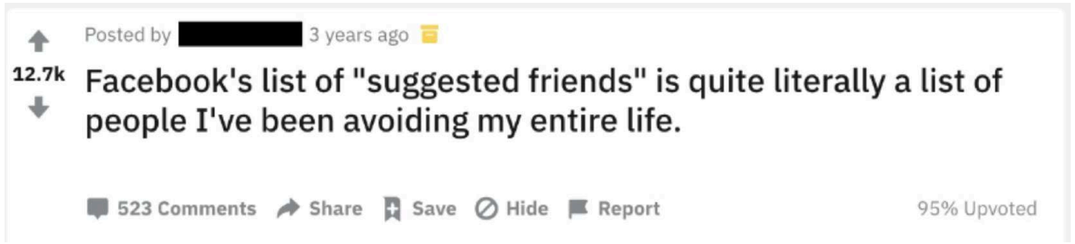
Sales of Dairy Products and Fruits & Vegetables Rise in March

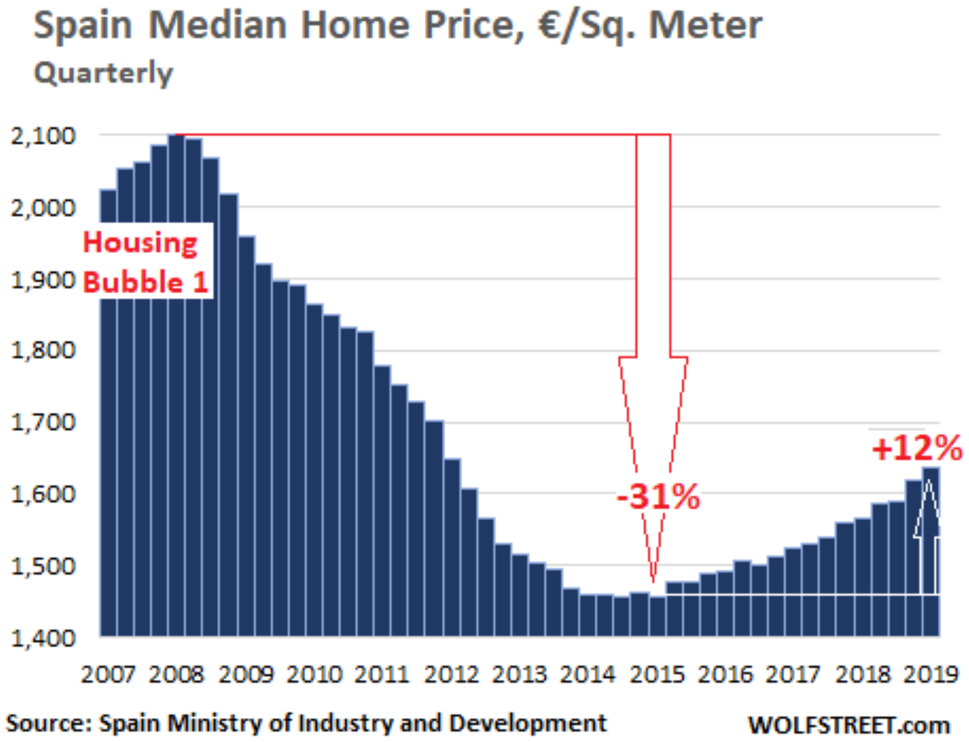


Indexed sales by category compared to Average in Jan 1-28.
At least 5 retailers per category

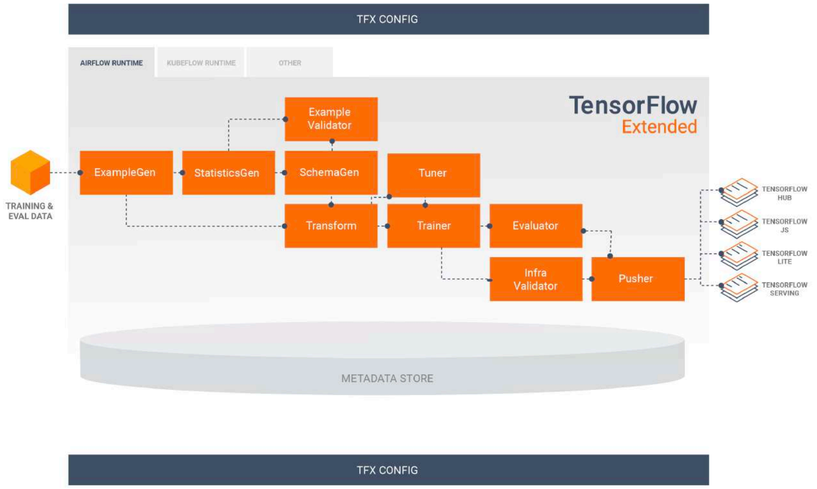
criteo

Why
is necessary to monitor and retrain?

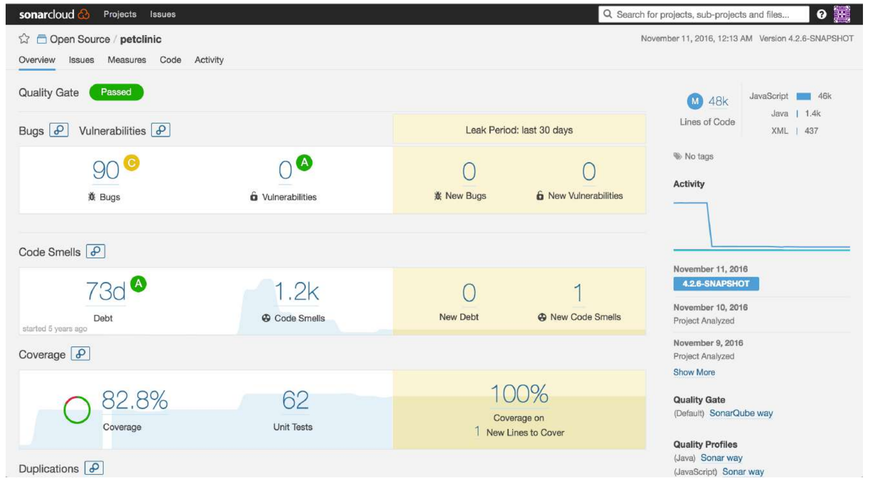




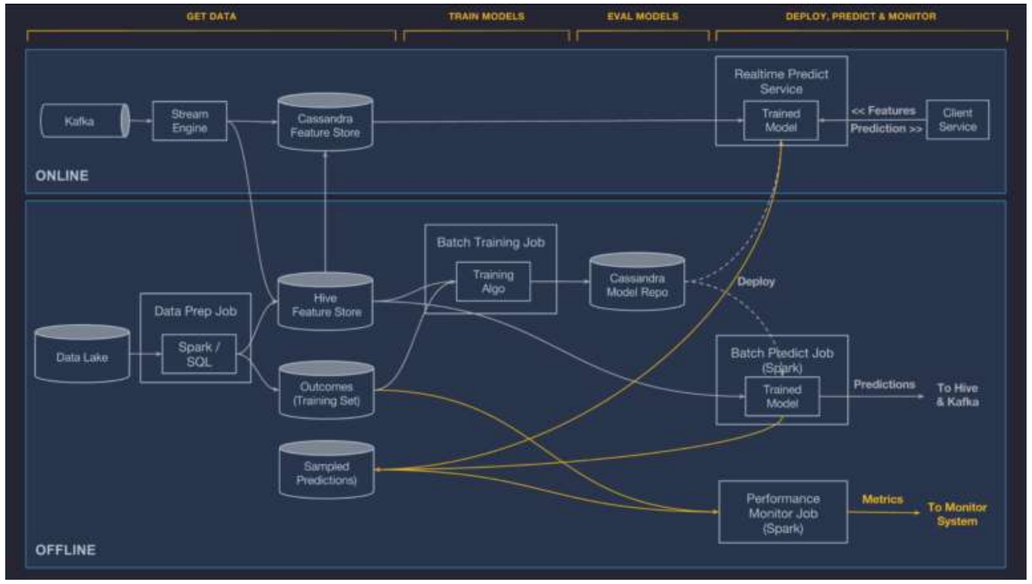
But I don’t work at a big company with thousand of ML engineers!

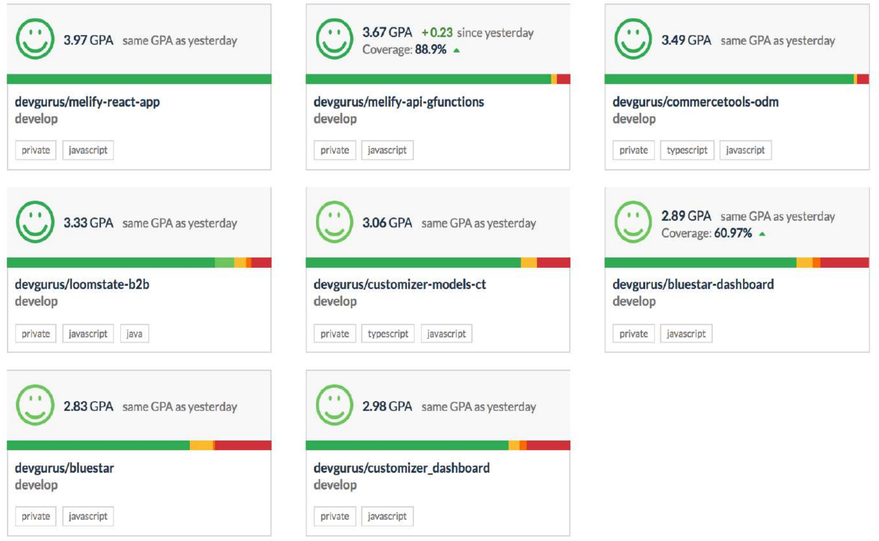


sonarqube



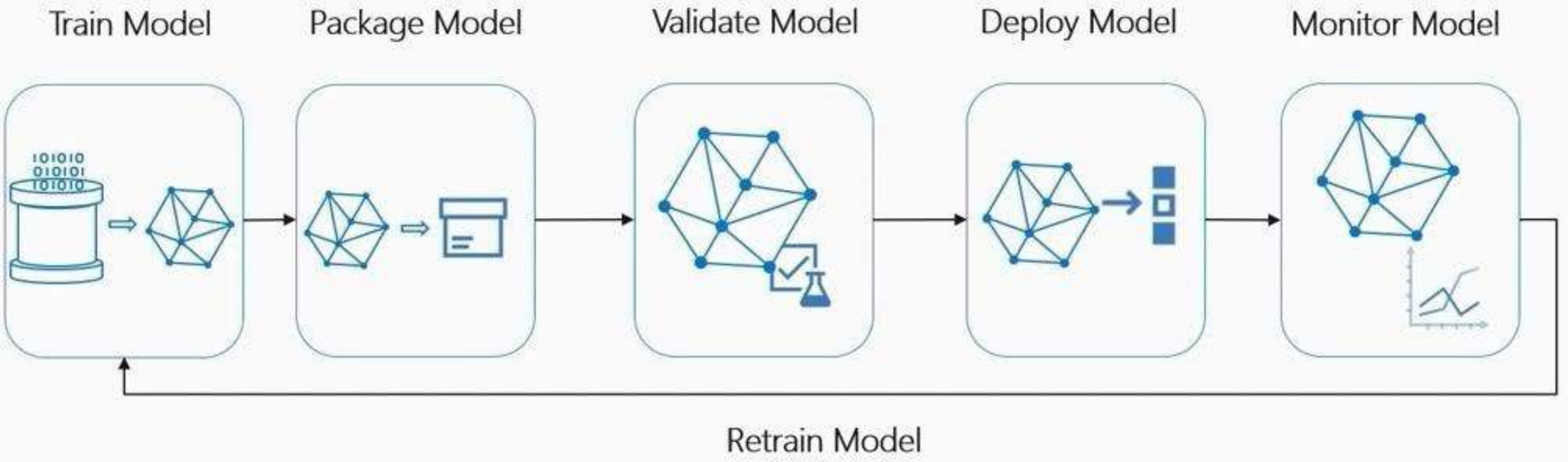
Michelangelo: Uber's Machine Learning Platform





What is the E2E ML lifecycle?

- **Develop & train model** with reusable ML pipelines
- **Package model** using containers to capture runtime dependencies for inference
- **Validate model behavior –** functionally, in terms of responsiveness, in terms of regulatory compliance
- **Deploy model -** to cloud & edge, for use in real-time / streaming / batch processing
- **Monitor model** behavior & business value, know **when to replace / deprecate a stale model**



MLOps Benefits

1. Reproducibility + Auditability

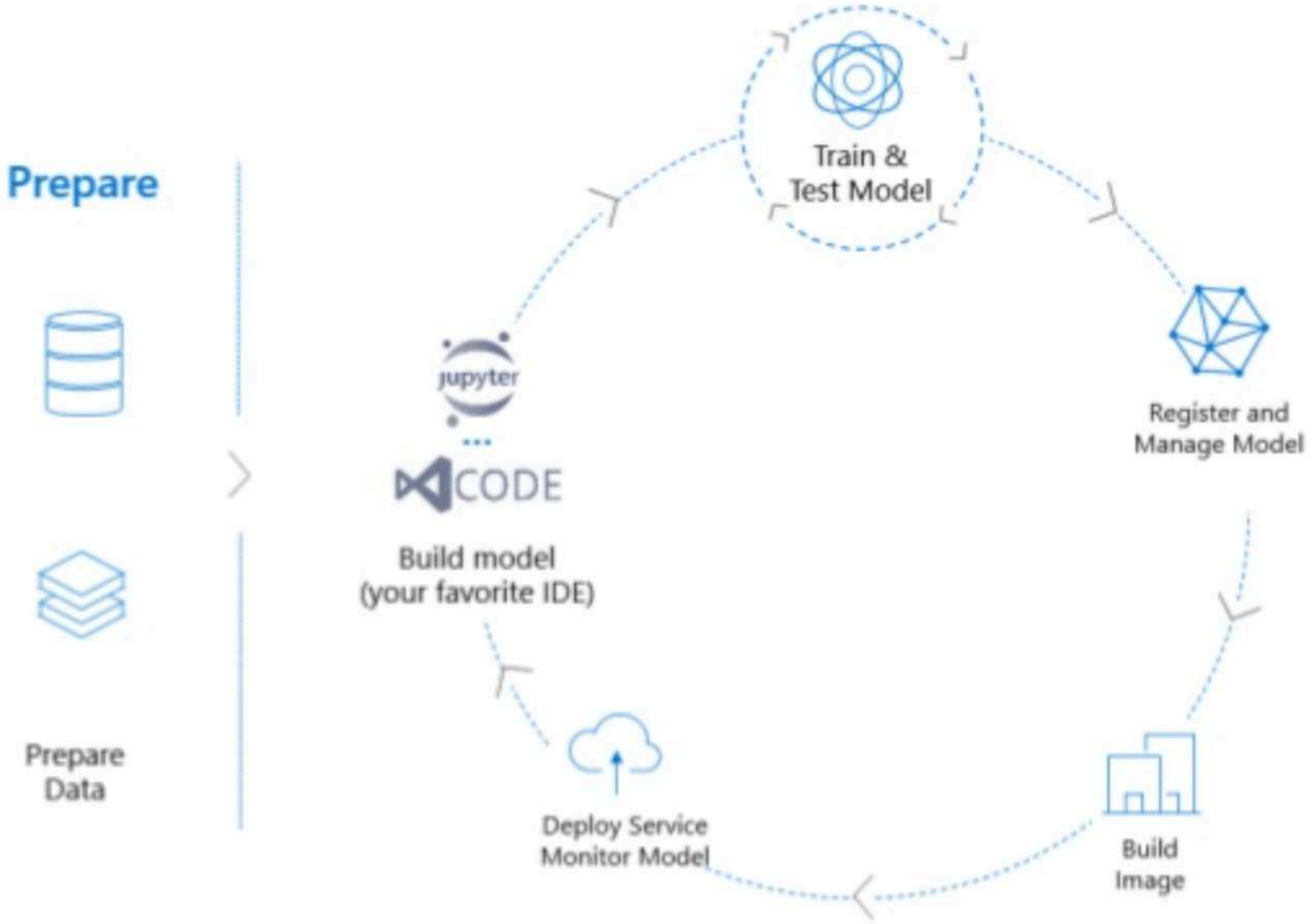
- Code drives generation and deployments
- Pipelines are reproducible and verifiable
- All artifacts can be tagged and audited

2. Validation

- SWE best practices for quality control
- Offline comparisons of model quality
- Minimize bias and enable explainability

3. Automation + Observability

- Controlled rollout capabilities.
- Live comparison of predicted vs. expected performance.
- Results fed back to watch for drift and improve model.



What's next?

Asynchronous Session 12

TRADITIONAL CORPORATION BUSINESS CASE

by Juanjo Casado


dgonzalezp@faculty.ie.edu