Week 4

Artificial Intelligence Program

Infrastructure and Architecture

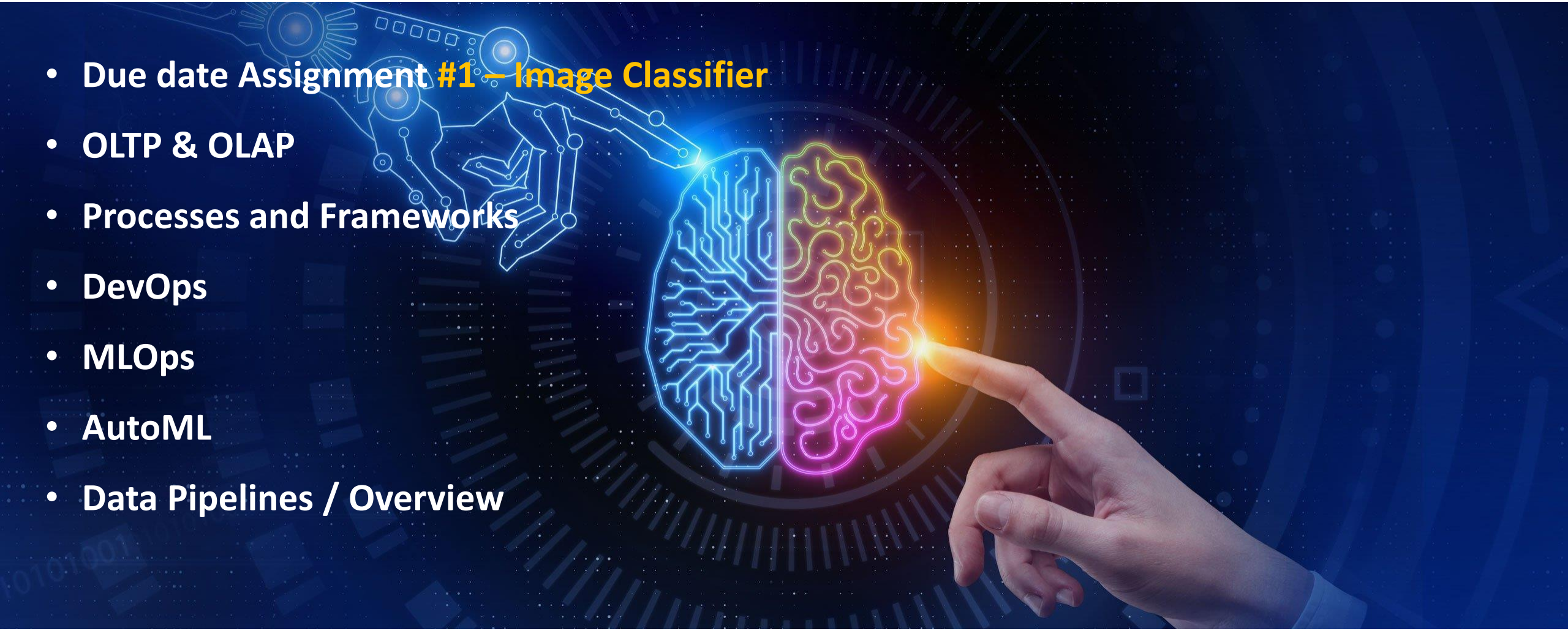
> Agenda // Program

Assignments [50%]
EXAMS [40%]
C# Challenges [10%]

WEEK	SUBJECT	ASSIGNMENT / TO BE DELIVERED	CHALLENGES
2	Intro / AI Function / Enablers		
3	Infra and Architecture / On-prem vs. Cloud / CSPs		C 1
4	Data Pipeline / Processes / Framework / AutoML	#1 Image Classifier **[5%]**	
5	Data Pipeline / Processes / Framework / AutoML		C 2
6	More Data / SSIS / ADF / Data Quality	#2 Machine Learning Studio **[10%]**	
7	Azure services – Intro **EXAM 1 [20%]**		C 3
8	**READING WEEK**	**NO CLASSES**	
9	Azure Cognitive Services 1		4.1
10	Azure Cognitive Services 2	#3 Draw your own Architecture **[5%]**	4.2
11	Azure Cognitive Services 3		4.3
12	Azure Cognitive Services 4	#4 Azure Pipeline / Sentiment Analysis **[10%]**	4.4
13	AWS Academy // Cloud Foundations		
14	AWS Academy // Machine Learning	#5 AWS Academy // Cloud Foundations **[10%]**	
15	Enterprise Architecture **EXAM 2 [20%]**	#6 AWS Academy // Machine Learning **[10%]**	

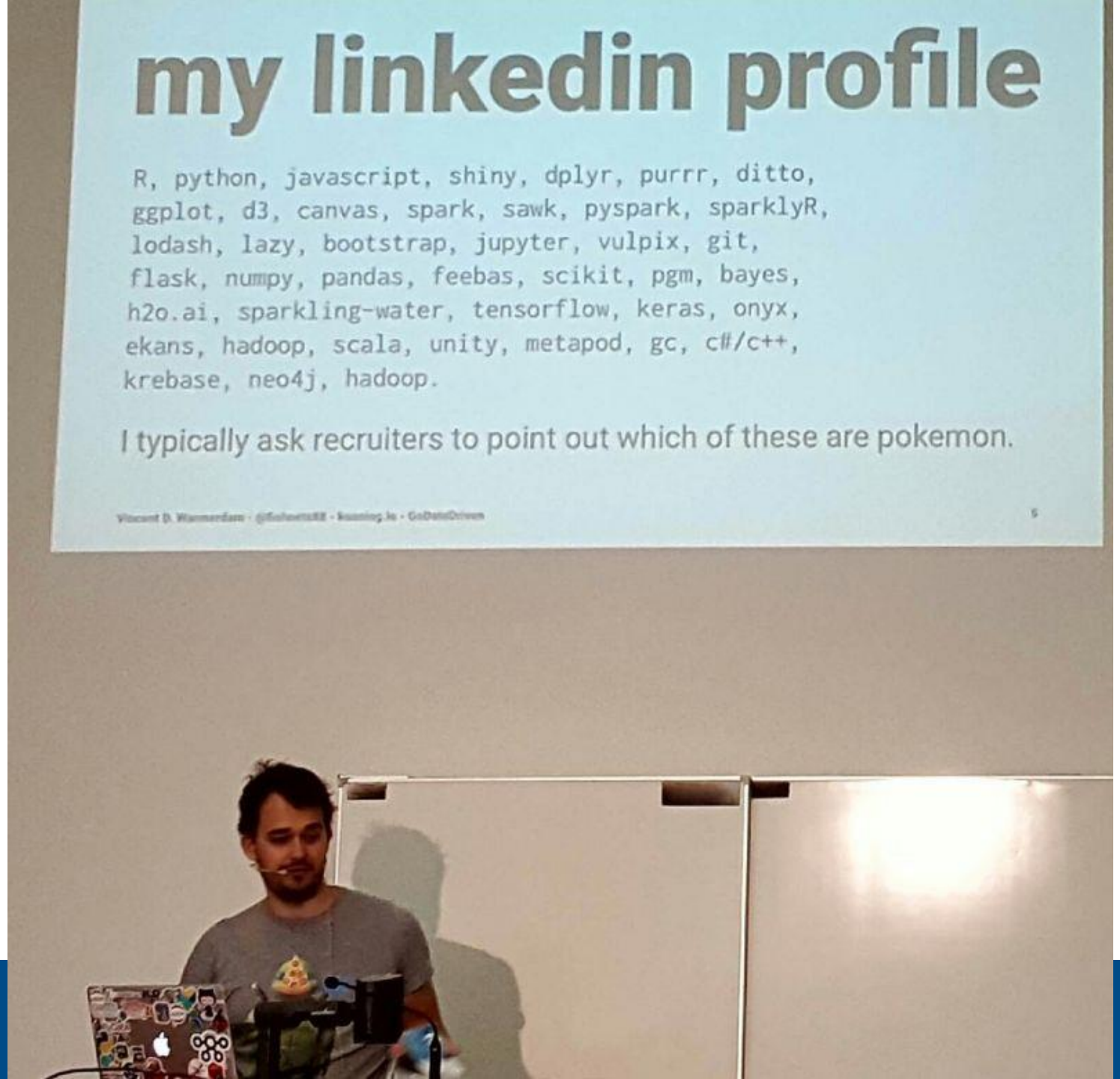
aws academy

> Agenda (4)



- Due date Assignment #1 – Image Classifier
- OLTP & OLAP
- Processes and Frameworks
- DevOps
- MLOps
- AutoML
- Data Pipelines / Overview

Which Of These Are Pokemon?



Georgian

Data Pipeline

OLTP & OLAP

EXAM

> OLTP & OLAP

OLTP stands for

On-Line Transactional processing

OLAP stands for

On-Line Analytical Processing

OLTP characteristics:

- Many transactions
- Latency sensitive
- Small payloads
- Balanced read/write or Heavy write workloads

OLAP characteristics:

- Few transactions
- Throughput sensitive
- Large (return) payloads
- Heavy read workloads (including full table scans)

https://hackr.io/blog/olap-vs-oltp

EXAM

> OLTP & OLAP

What is OLTP? **TRANSACTIONAL**

An OLTP system captures and maintains transaction data in a database. Each transaction involves individual database records made up of multiple fields or columns.

Examples include banking and credit card activity or retail checkout scanning.

In OLTP, the emphasis is on fast processing, because OLTP databases are read, written, and updated frequently. If a transaction fails, built-in system logic ensures data integrity.

What is OLAP? **ANALYTICAL**

OLAP applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and business intelligence projects. In OLAP, the emphasis is on response time to these complex queries. Each query involves one or more columns of data aggregated from many rows. Examples include year-over-year financial performance or marketing lead generation trends. OLAP databases and data warehouses give analysts and decision-makers the ability to use custom reporting tools to turn data into information. Query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact the accuracy of business intelligence insights.

> OLTP & OLAP

EXAM

OLTP // TRANSACTIONAL

- Frequent Read/Write/Update
- Customer Impact
  - ERP Solutions
  - Payment Processing
  - Business support solutions

OLAP // ANALYTICAL

- Historical and Aggregated Data
- NO Customer Impact
  - Datawarehouse solutions
  - Business Intelligence
  - Back Office solutions

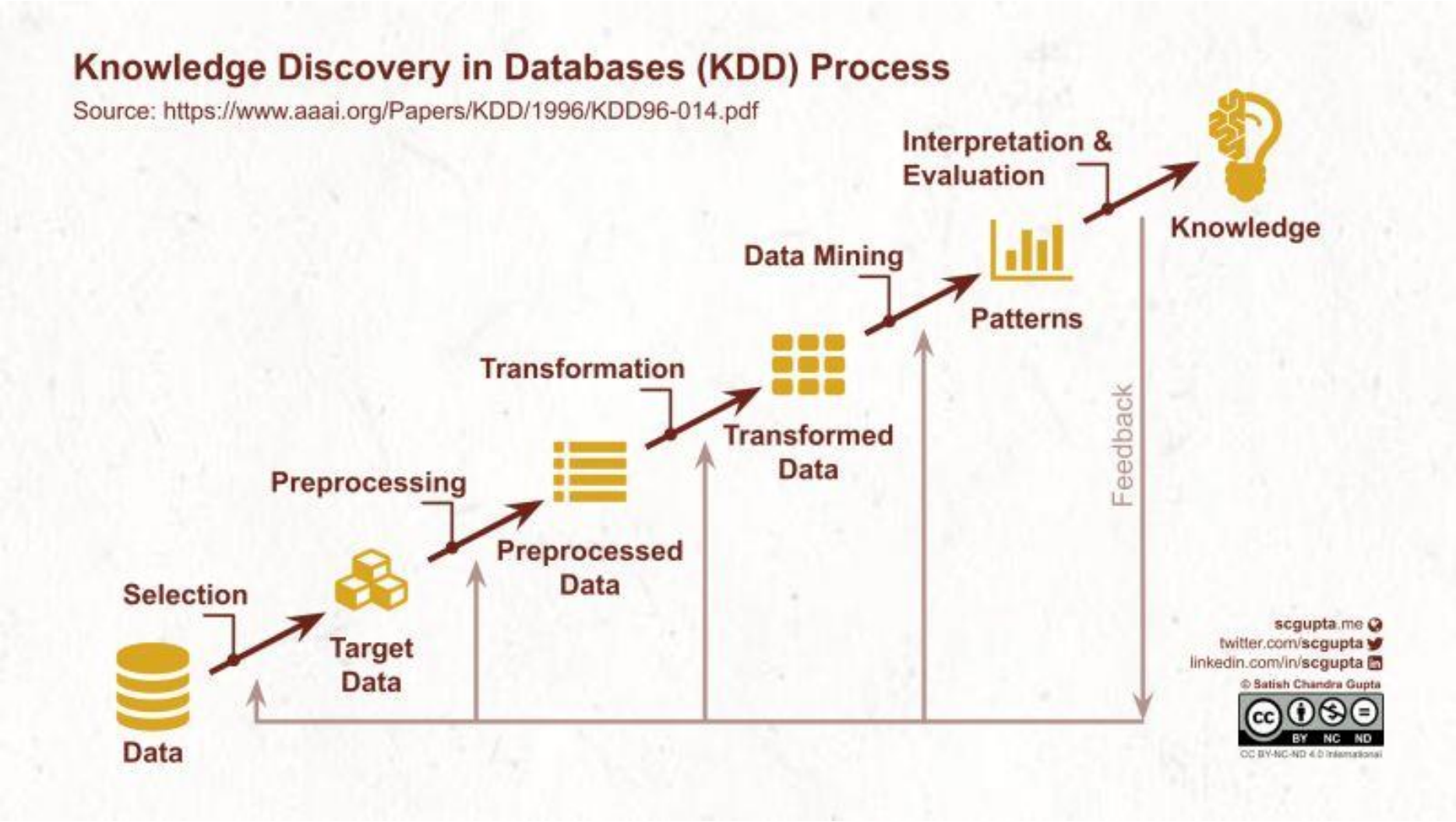
BASIS FOR COMPARISON	OLTP	OLAP
Basic	It is an online transactional system and manages database modification.	It is an online data retrieving and data analysis system.
Focus	Insert, Update, Delete information from the database.	Extract data for analyzing that helps in decision making.
Data	OLTP and its transactions are the original source of data.	Different OLTPs database becomes the source of data for OLAP.
Transaction	OLTP has short transactions.	OLAP has long transactions.
Time	The processing time of a transaction is comparatively less in OLTP.	The processing time of a transaction is comparatively more in OLAP.
Queries	Simpler queries.	Complex queries.
Normalization	Tables in OLTP database are normalized (3NF).	Tables in OLAP database are not normalized.
Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequently modified. Hence, data integrity is not affected.

https://techdifferences.com/difference-between-oltp-and-olap.html

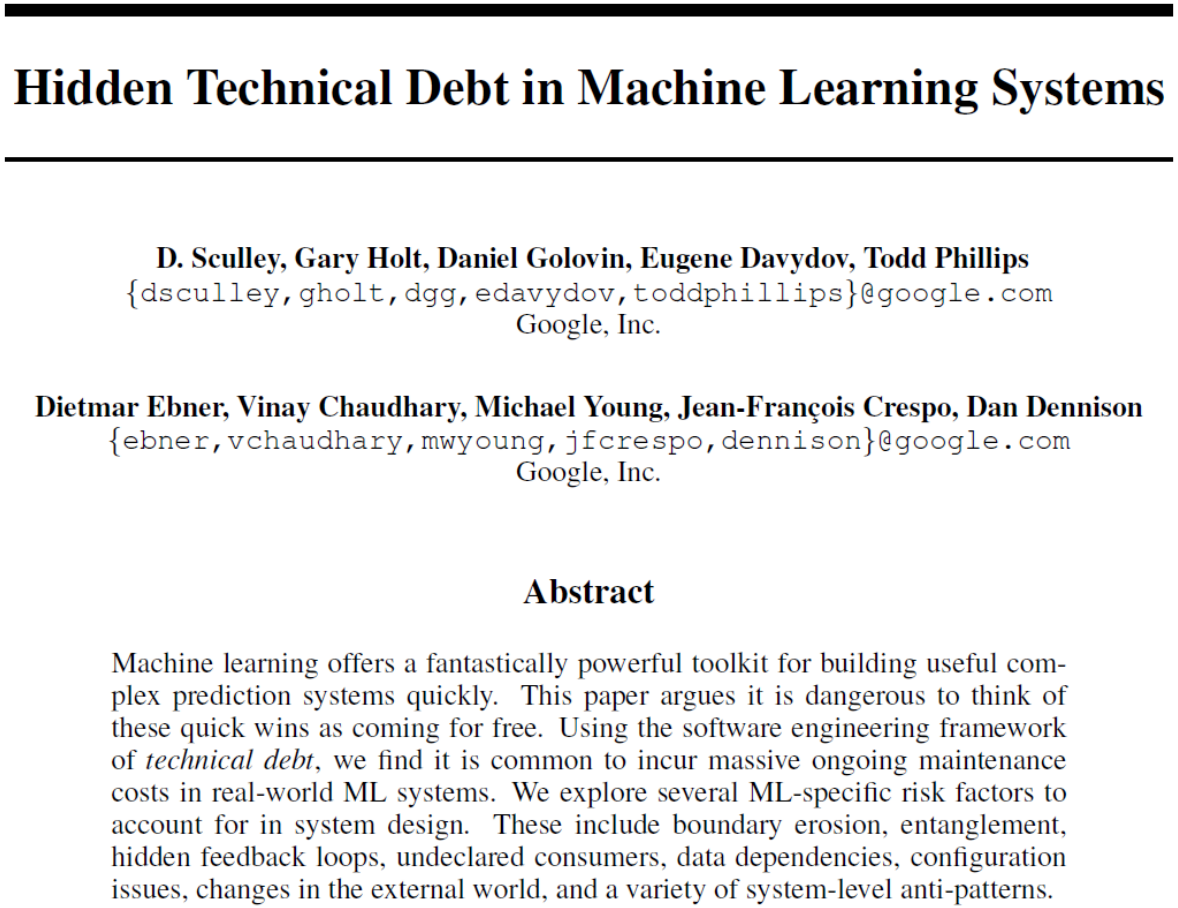
MLOps

Machine Learning Ops

> MLOps


Knowledge Discovery in Databases (KDD) Process
Source: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf
Interpretation & Evaluation
Knowledge
Data Mining
Patterns
Transformation
Transformed Data
Preprocessing
Preprocessed Data
Selection
Target Data
Data
Feedback
scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta
© Satish Chandra Gupta
BY NC ND
CC BY-NC-ND 4.0 International

> DevOps and MLOps



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.

This discipline is relatively new and started in 2015 with the publication of the article "***Hidden Technical Debt in Machine Learning Systems***".

This article basically highlights the costs and efforts to maintain a Machine Learning solution due to the technical debt generated by this kind of solution.

> DevOps and MLOps | What is technical debt?

Technical debt (also known as design debt or code debt, but can be also related to other technical endeavors) is a concept in software development that reflects the implied cost of additional rework caused by choosing an easy (limited) solution now instead of using a better approach that would take longer.

As with monetary debt, if technical debt is not repaid, it can accumulate 'interest', making it harder to implement changes. Unaddressed technical debt increases software entropy. Similarly, to monetary debt, technical debt is not necessarily a bad thing, and sometimes (e.g., as a proof-of-concept) is required to move projects forward.

On the other hand, some experts claim that the "technical debt" metaphor tends to minimize the ramifications, which results in insufficient prioritization of the necessary work to correct it.

As a change is started on a codebase, there is often the need to make other coordinated changes in other parts of the codebase or documentation. Changes required that are not completed are considered debt, and until paid, will incur interest on top of interest, making it cumbersome to build a project. Although the term is used in software development primarily, it can also be applied to other professions.

https://en.wikipedia.org/wiki/Technical_debt

> DevOps and MLOps

MLOps vs. ModelOps vs. AIOps

MLOps (or ModelOps) is a relatively new discipline, emerging under these names particularly in late 2018 and 2019. The two — MLOps and ModelOps — are, at the time this book is being written and published, largely being used interchangeably. However, some argue that ModelOps is more general than MLOps, as it's not only about machine learning models but any kind of model (e.g., rule-based models). For the purpose of this book, we'll be specifically discussing the machine learning model lifecycle and will thus use MLOps.

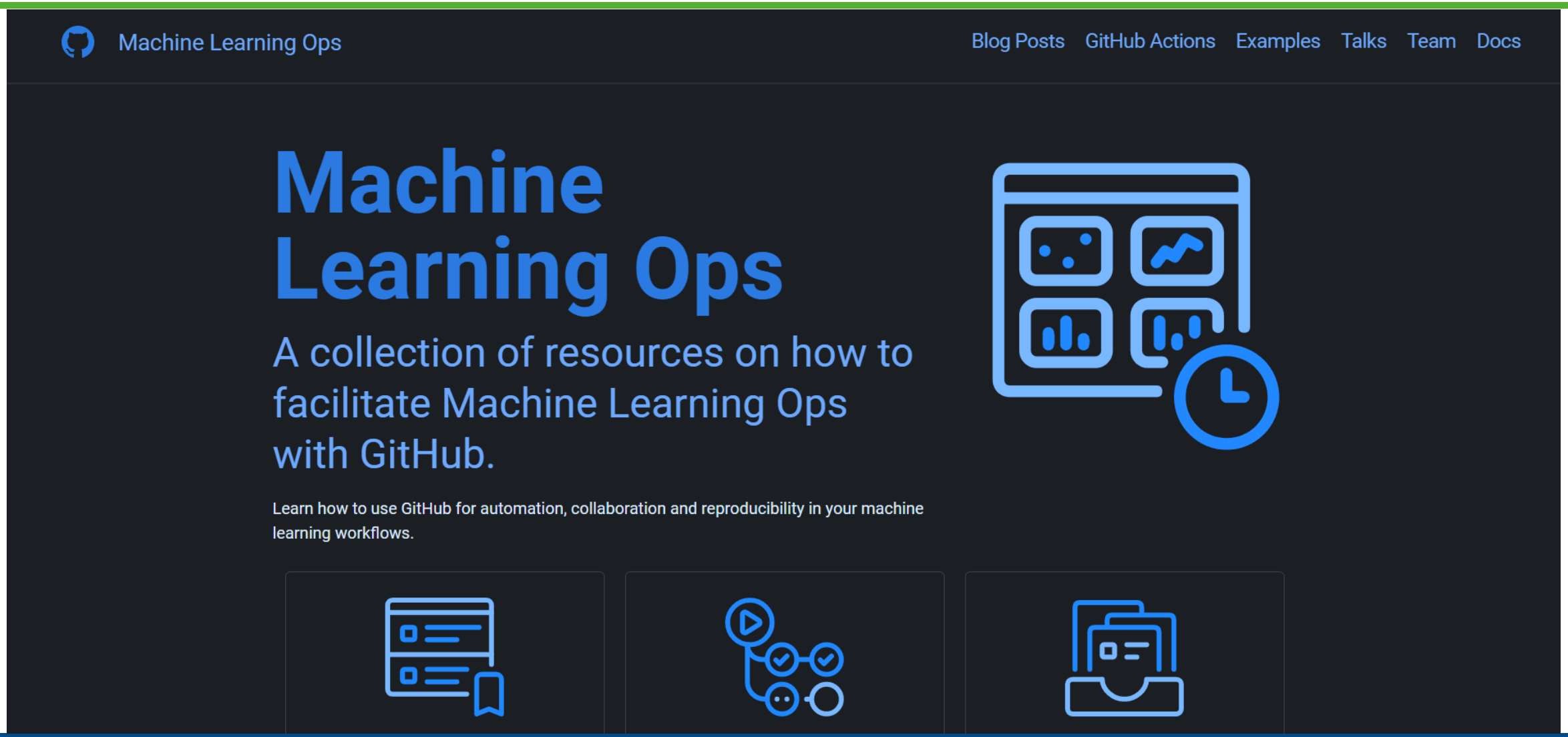
AIOps, though sometimes confused with MLOps, is another topic entirely and refers to the process of solving operational challenges through the use of artificial intelligence (i.e., AI for DevOps). An example would be a form of predictive maintenance but for network failures, alerting DevOps teams to possible problems before they arise. While important and interesting in its own right, AIOps is outside the scope of this book.

MLOps – Machine Learning Operations is a critical component of data science projects deployment in enterprise environments.

DevOps – The concept of DevOps that tries to pull closer the development area and the operational area (increasing collaboration and communication) tries to manage all the software changes and updates in the computer environment, focusing on continuous delivery and high quality.

Automation and trust between different teams and processes

> MLOps Process

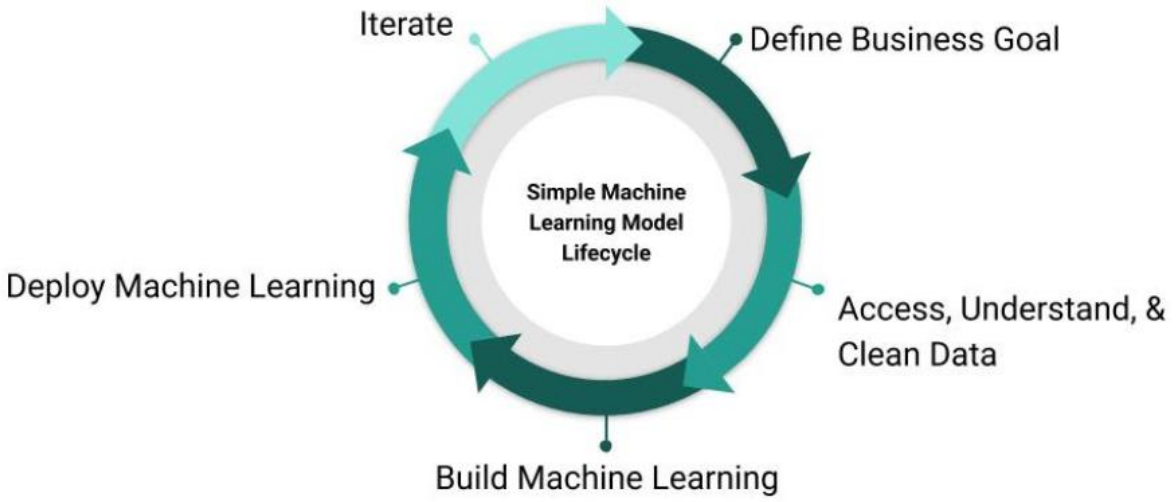


https://mlops.githubapp.com/

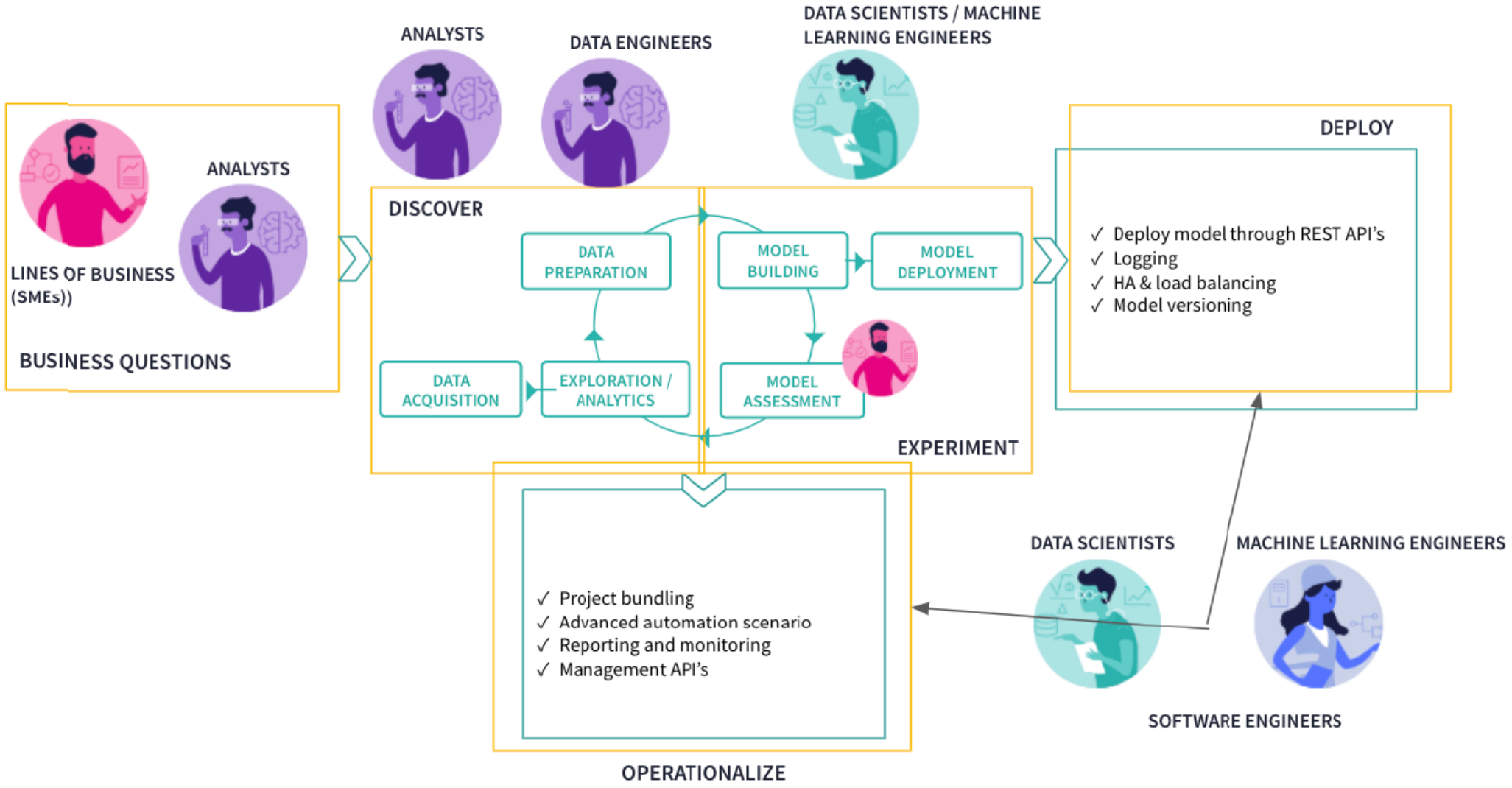
> MLOps Process

The development of multiple machine learning models and their deployment in production environment are relatively new. The number of different models has been managed at a small scale or simply a low interest about how to interpret their dependencies, versions, dependencies and of course the history told by a massive number of different models.

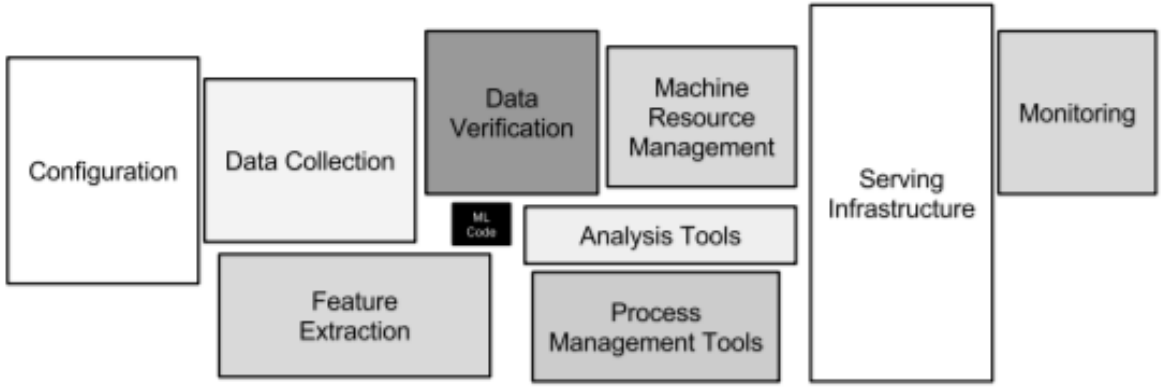
With automated processes these models, versions, became more critical and the need to control each of their variances becomes more important for the companies, specially managing risks and also understand the final results presented by this black boxes models.



> MLOps Process | The Reality



> MLOps Challenges

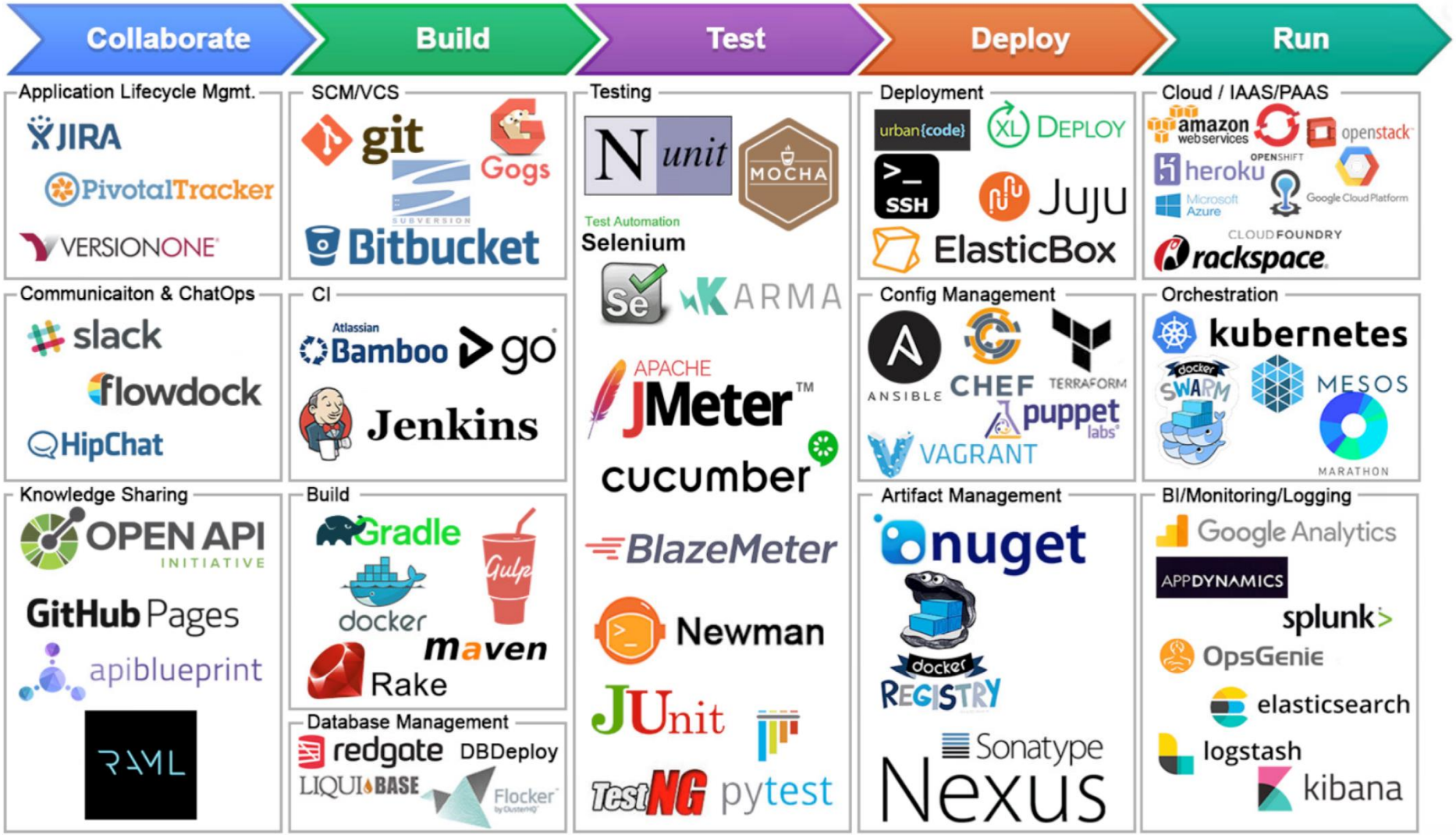


Main Challenges:

- Small part is ML code
- Data Management: Massive datasets
- Experiment: Integrated environments / data access
- Time to train: Adequate infra
- Testing: Bias and fairness

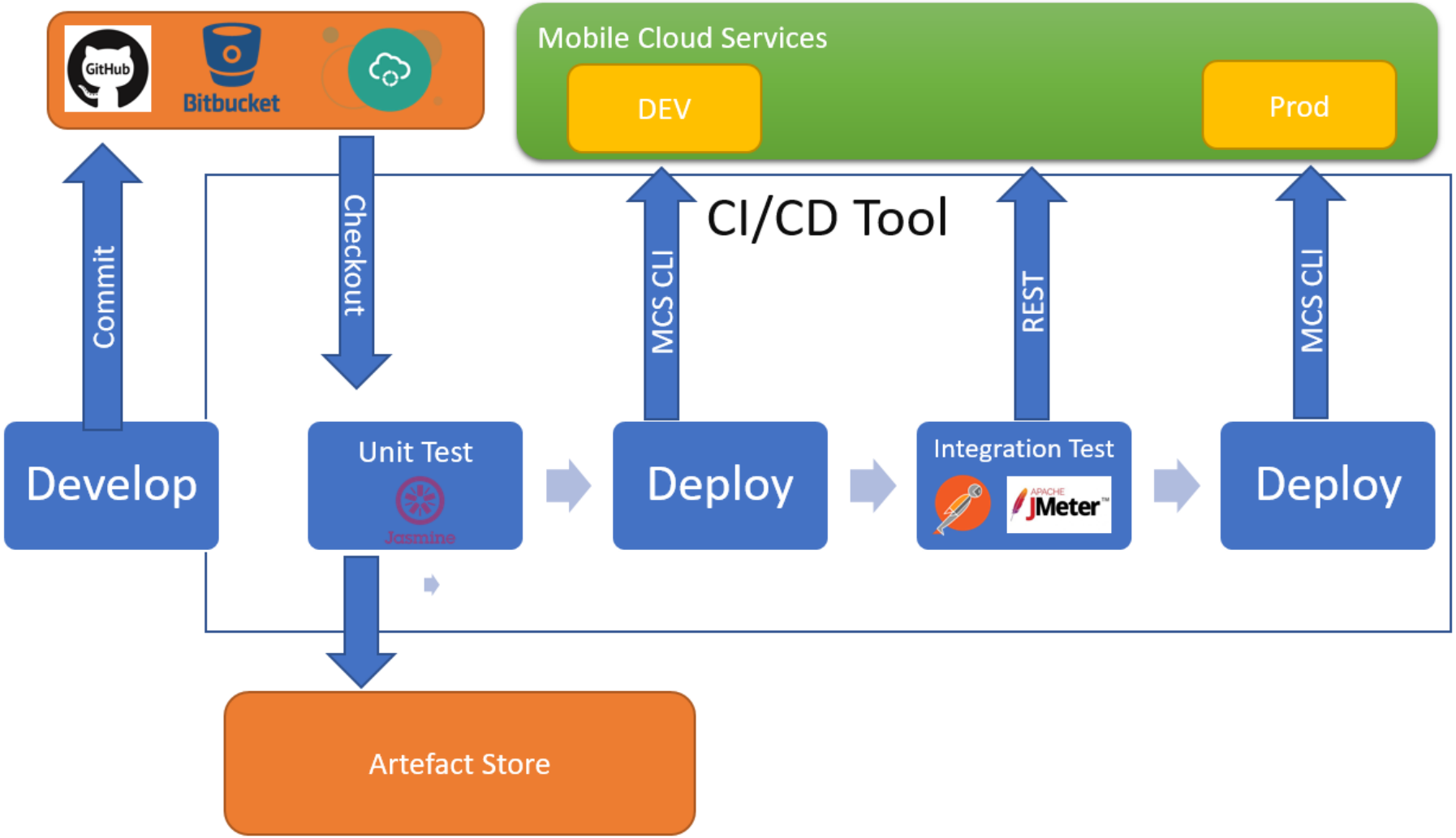
It is not Rocket science but is also not so simple as “Hello World”. Look for CI/CD tools.

> MLOps | CI/CD Tools



https://www.suntechnologies.com/blogs/ci-cd-pipelines-for-deploying-multiple-projects-best-ways-to-automate-build-and-deployment/

> MLOps | CI/CD Tools



https://www.ateam-oracle.com/cicd-with-mobile-cloud-services

> MLOps Process

The great difference between MLOps and DevOps:

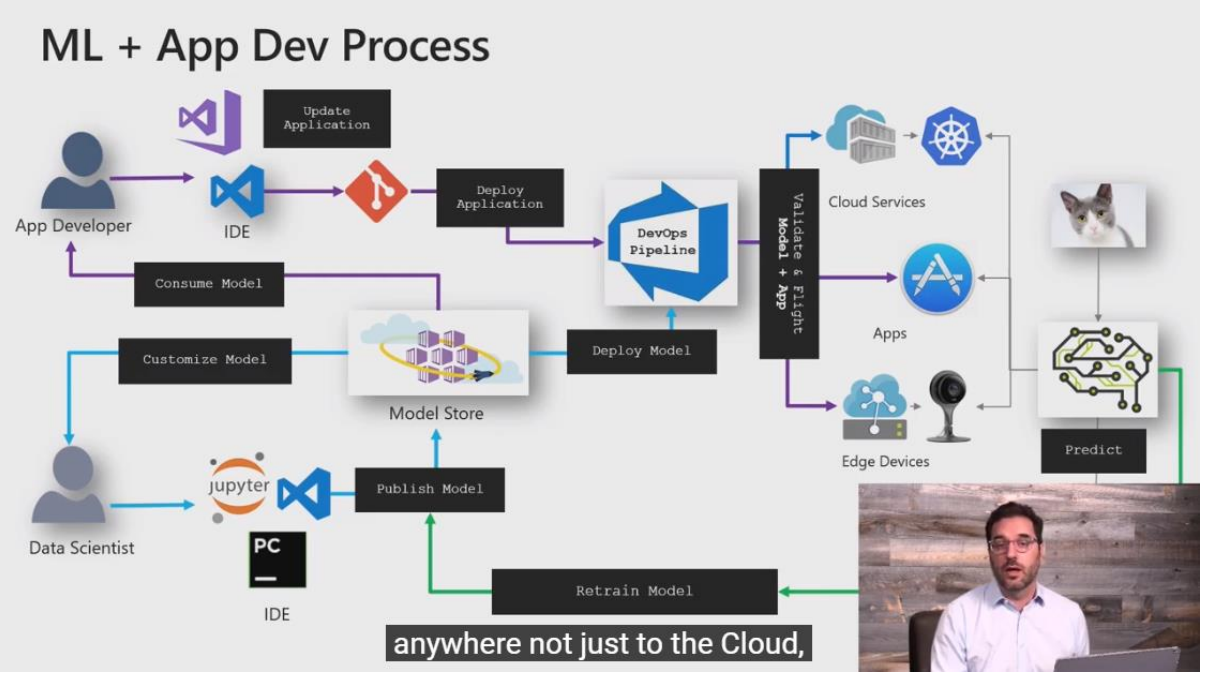
Deploying code is completely different than deploying machine learning models in a production environment.

Why?

Why MLOps is so important?

Because prediction models are only as good as the data they are trained on, which means the training data must be a good reflection of the data encountered in the production environment. If the production environment changes, then the model performance is likely to decrease rapidly.

> MLOps Process



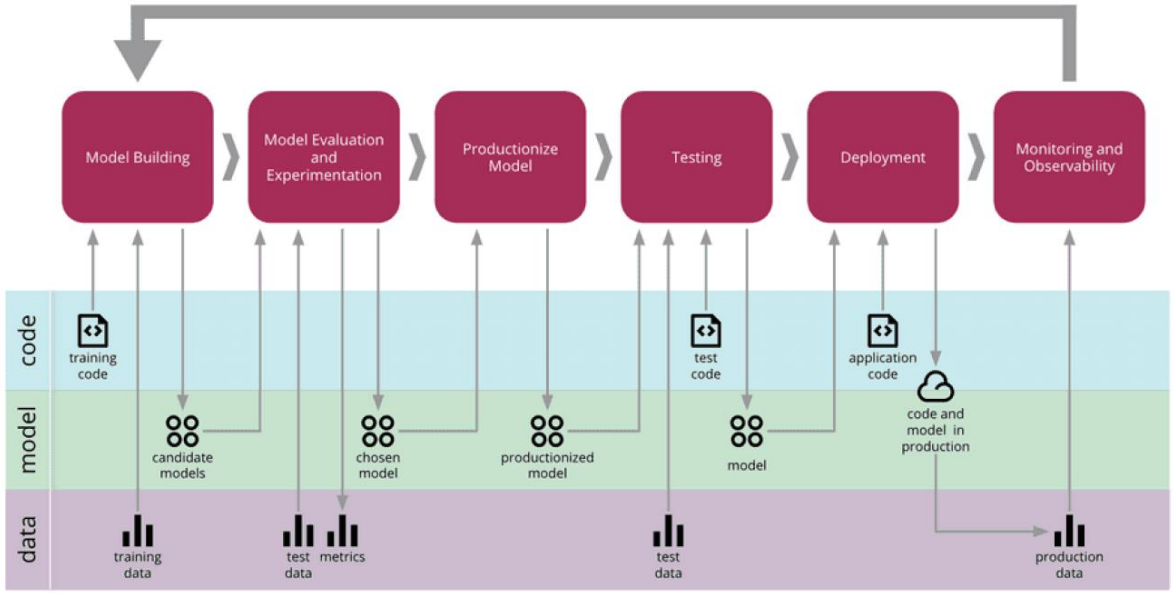
https://youtu.be/0MaHb070H_8

https://youtu.be/p4DYmaClkNY

> Other Important Aspects

Machine learning models need to be monitored at two levels:

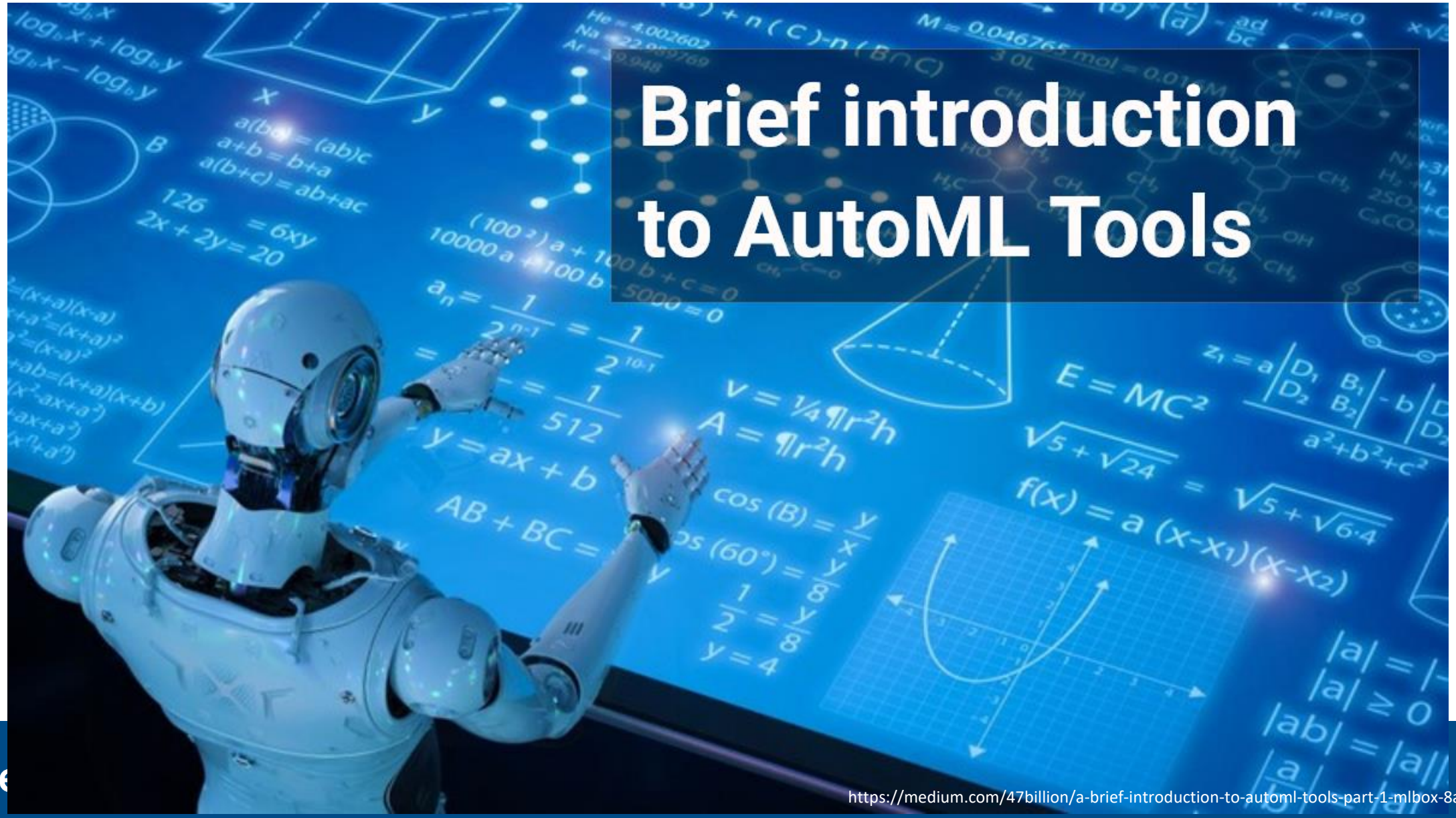
- At the resource level, including ensuring the model is running correctly in the production environment. Key questions include: Is the system alive? Is the CPU, RAM, network usage, and disk space as expected? Are requests being processed at the expected rate?
- At the performance level, meaning monitoring the pertinence of the model over time. Key questions include: Is the model still an accurate representation of the pattern of new incoming data, and is it still performing as well as during its design phase?



Auto ML

Auto Machine Learning

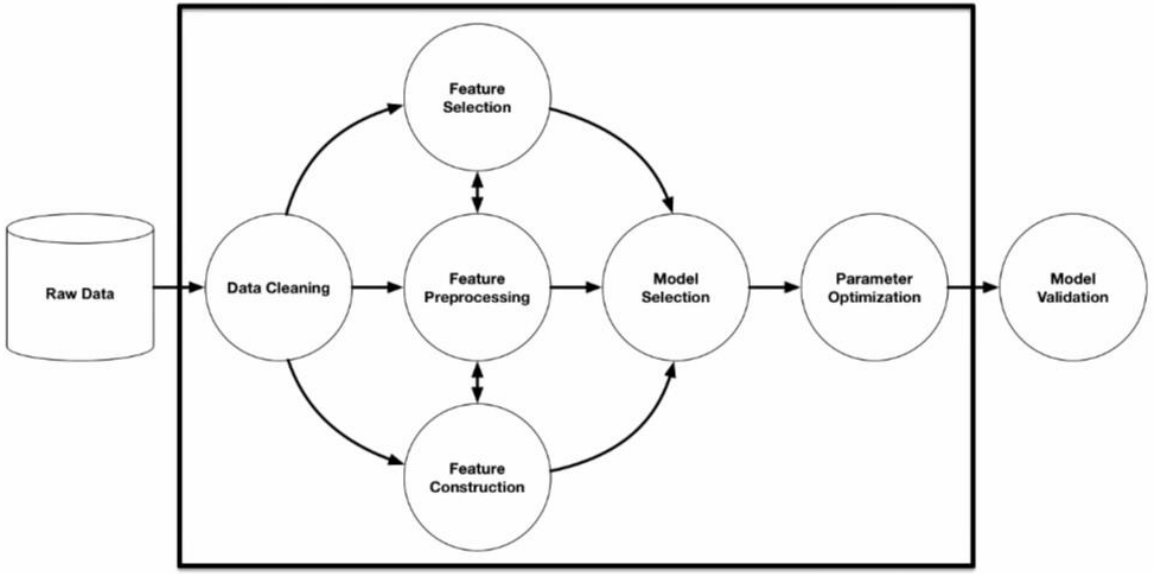
> ML | Auto ML



https://medium.com/47billion/a-brief-introduction-to-automl-tools-part-1-mlbox-8a447b8ff65

> ML | Auto ML | Intro

https://youtu.be/Rsg_XzgGqZw



Source: R. Olson et. al. (2016) "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science."

Introduction of AutoML

One cannot introduce AutoML without mentioning the machine learning project's life cycle, which includes data cleaning, feature selection/engineering, model selection, parameter optimization, and finally, model validation. As advanced as technology has become, the traditional data science project still incorporates a lot of manual processes and remains time-consuming and repetitive.

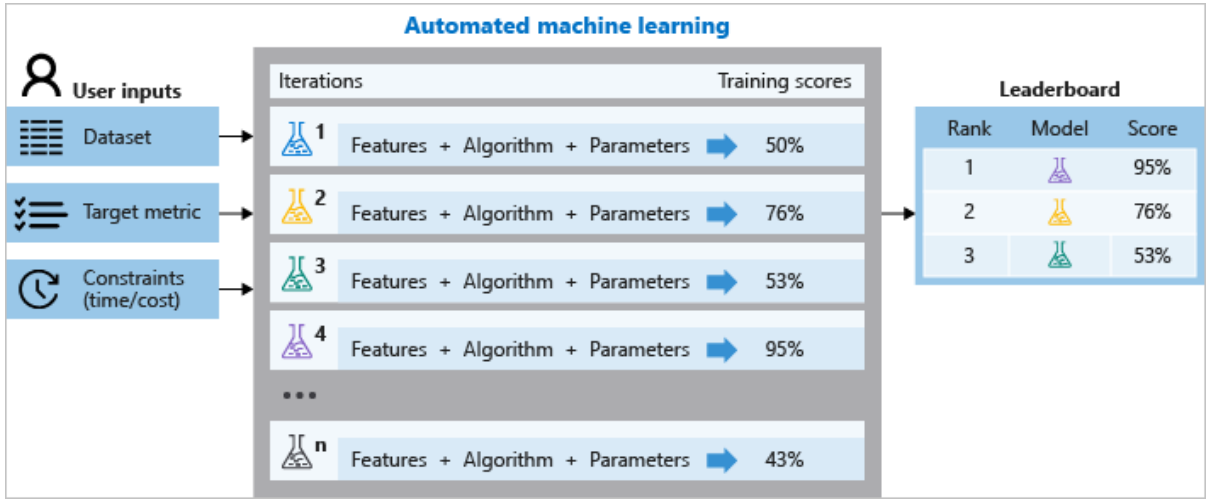
AutoML came into the picture to automate the entire process from data cleaning to parameter optimization. It provides tremendous value for machine learning projects in terms of both time savings and performance.

> ML | Auto ML

What is automated machine learning (AutoML)?

Automated machine learning, also referred to as automated ML or AutoML, is the process of automating the time consuming, iterative tasks of machine learning model development. It allows data scientists, analysts, and developers to build ML models with high scale, efficiency, and productivity all while sustaining model quality.

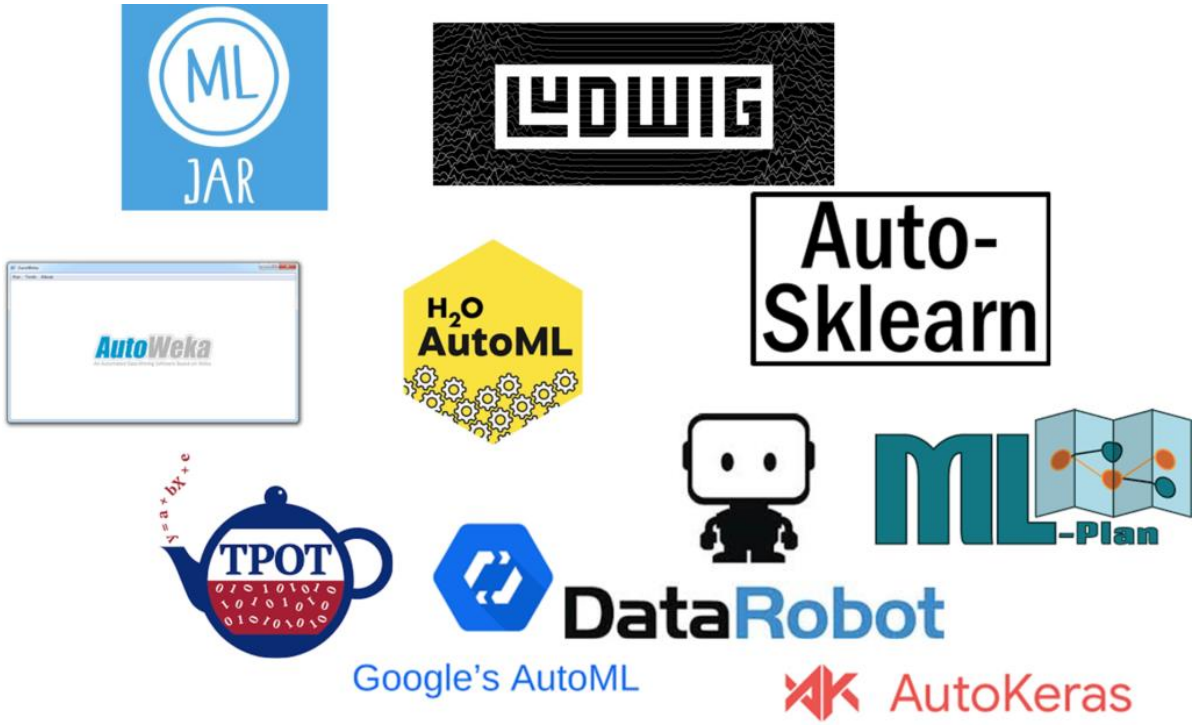
Traditional machine learning model development is resource-intensive, requiring significant domain knowledge and time to produce and compare dozens of models. With automated machine learning, you'll accelerate the time it takes to get production-ready ML models with great ease and efficiency.



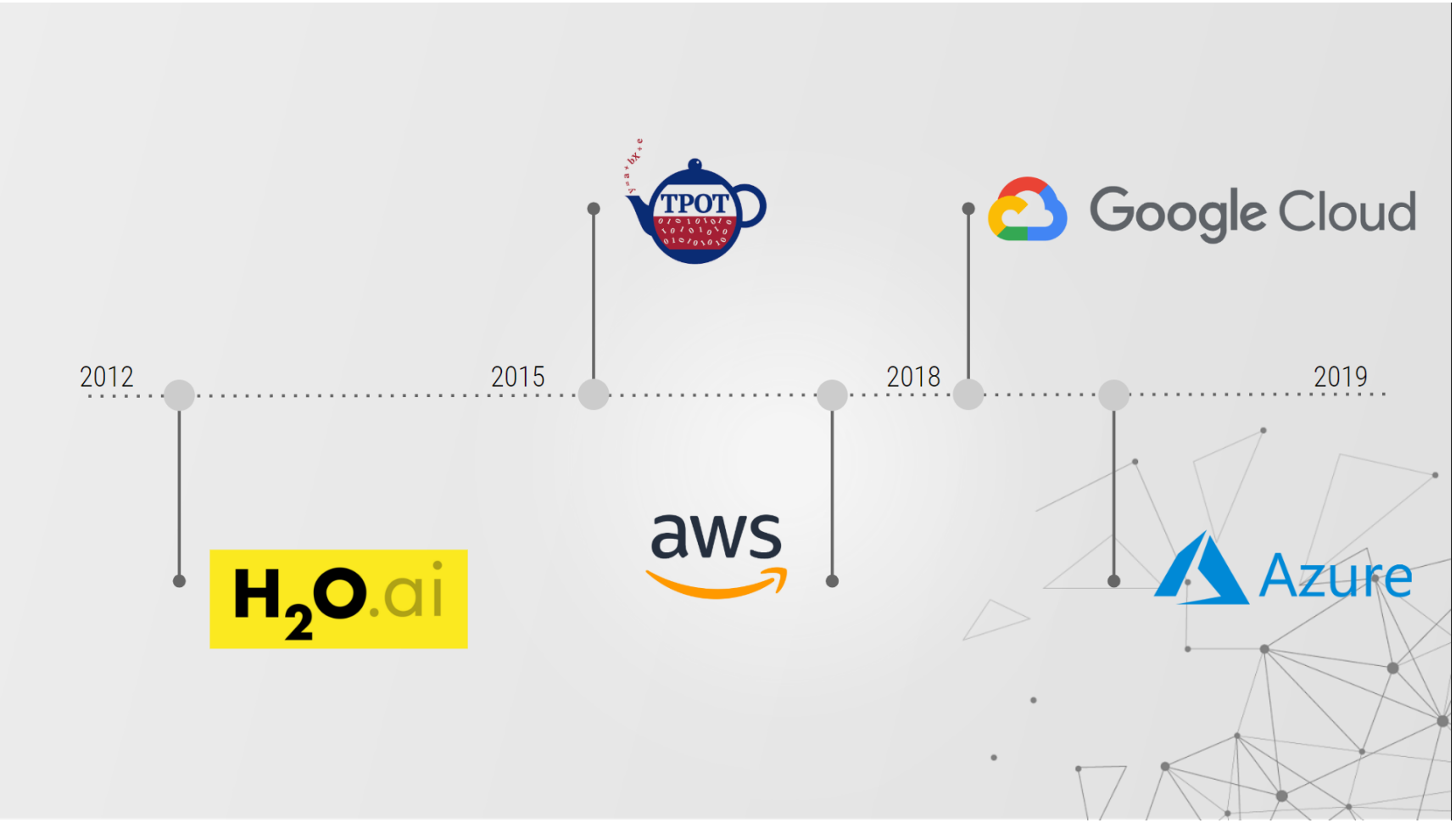
https://www.microsoft.com/en-us/videoplayer/embed/RE2Xc9t

https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml

> ML | Auto ML | Available Platforms

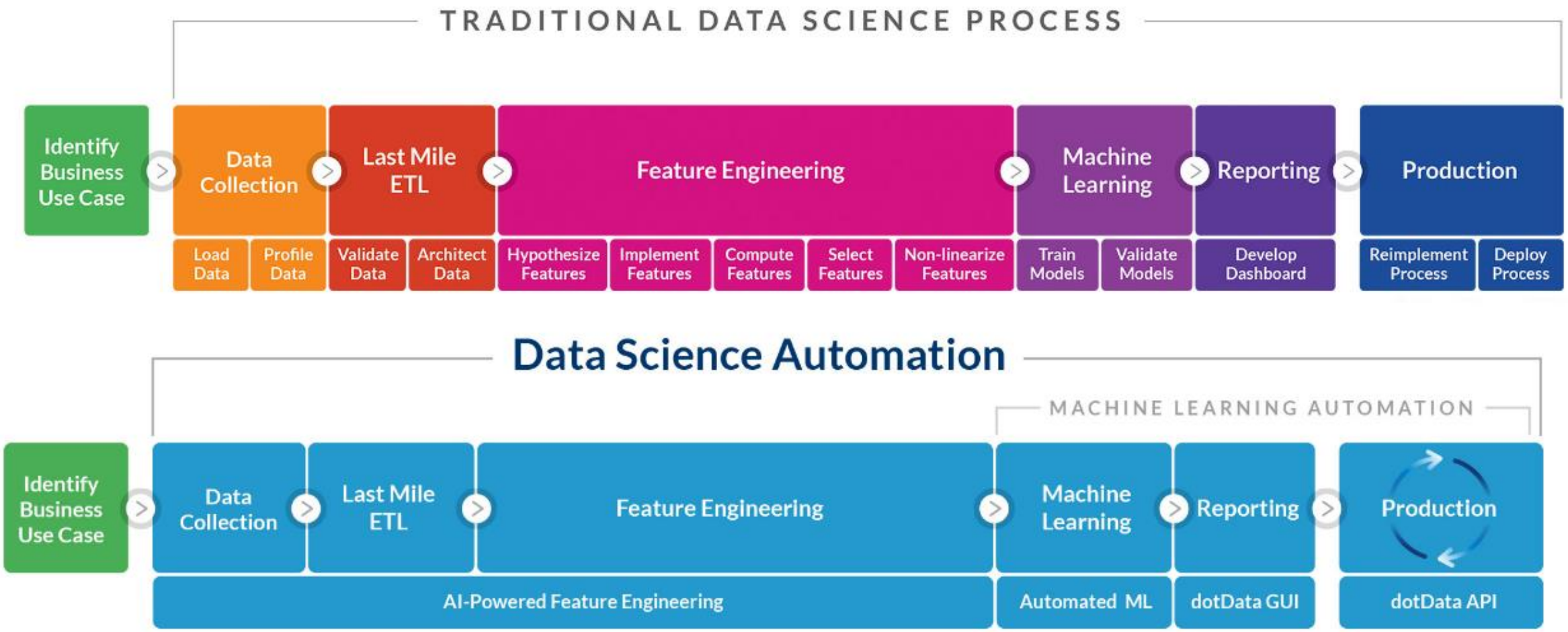


> ML | Auto ML | Available Platforms



https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> ML | Auto ML


TRADITIONAL DATA SCIENCE PROCESS
Identify Business Use Case
Data Collection
Last Mile ETL
Feature Engineering
Machine Learning
Reporting
Production
Load Data
Profile Data
Validate Data
Architect Data
Hypothesize Features
Implement Features
Compute Features
Select Features
Non-linearize Features
Train Models
Validate Models
Develop Dashboard
Reimplement Process
Deploy Process
Data Science Automation
MACHINE LEARNING AUTOMATION
Identify Business Use Case
Data Collection
Last Mile ETL
Feature Engineering
Machine Learning
Reporting
Production
AI-Powered Feature Engineering
Automated ML
dotData GUI
dotData API

> Data Scientists vs. AutoML Platforms

KDnuggets

When Will AutoML replace Data Scientists?

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms

KDnuggets

Dataset 1: SPEED DATING

Joseph Chin
Statistically insignificant

Azure

Azure AutoML
An ever-expanding cloud service to help organizations meet their data science needs

Aifaz Gowani
An outlier

Google Cloud

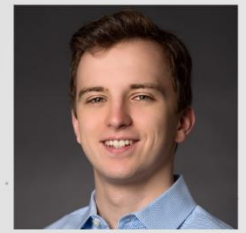
Google Cloud AutoML
Simple, secure and flexible ML service that lets you build pipelines and train custom models.

Dataset 2: ASHRAE

Matthew Peng
Placed in top 5 in Humana Data Science competition

Azure

Azure AutoML
While you are paying more than $40K for this degree, I am running at $2 per hour



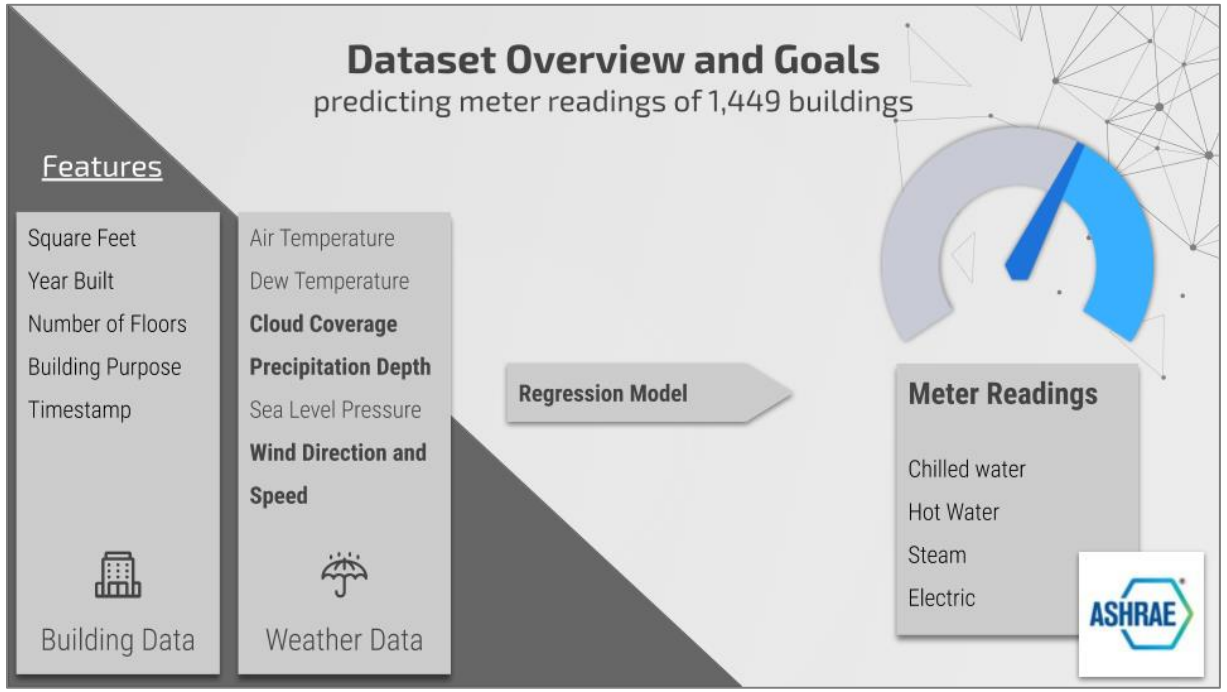
Gabriel James
Did not place top 5 in Humana Data Science competition

Google Cloud

Google Cloud AutoML
Slowly taking over your job with better everything

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms



https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms

KDnuggets

What we tried and what worked

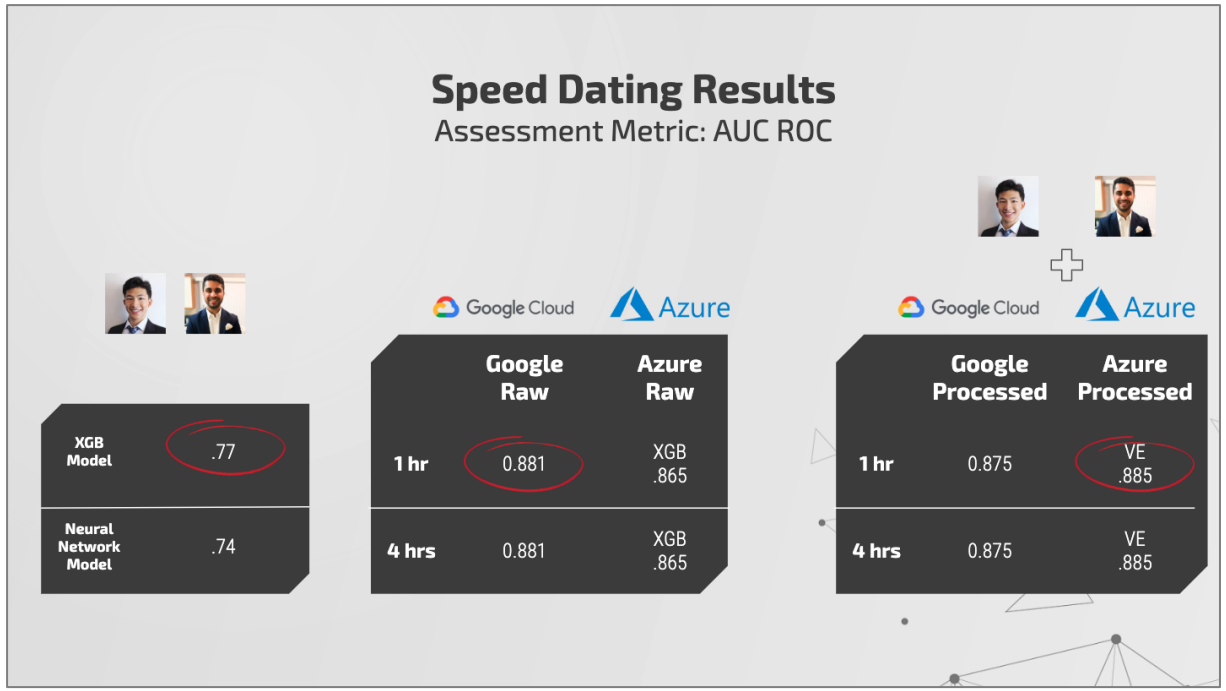
What we did		Effectiveness/ Local model	Effectiveness/ AutoML
Deal with missing value	Fill na for both numerical and categorical data	😀	😀
Clean up data	123 features → 67 features	😀	😀
SMOTE	Rebalancing highly imbalanced data	😀	😀
One Hot Encoding	For categorical features	😀	☹
Feature Engineering	With domain knowledge	😀	😀

What we tried and what worked

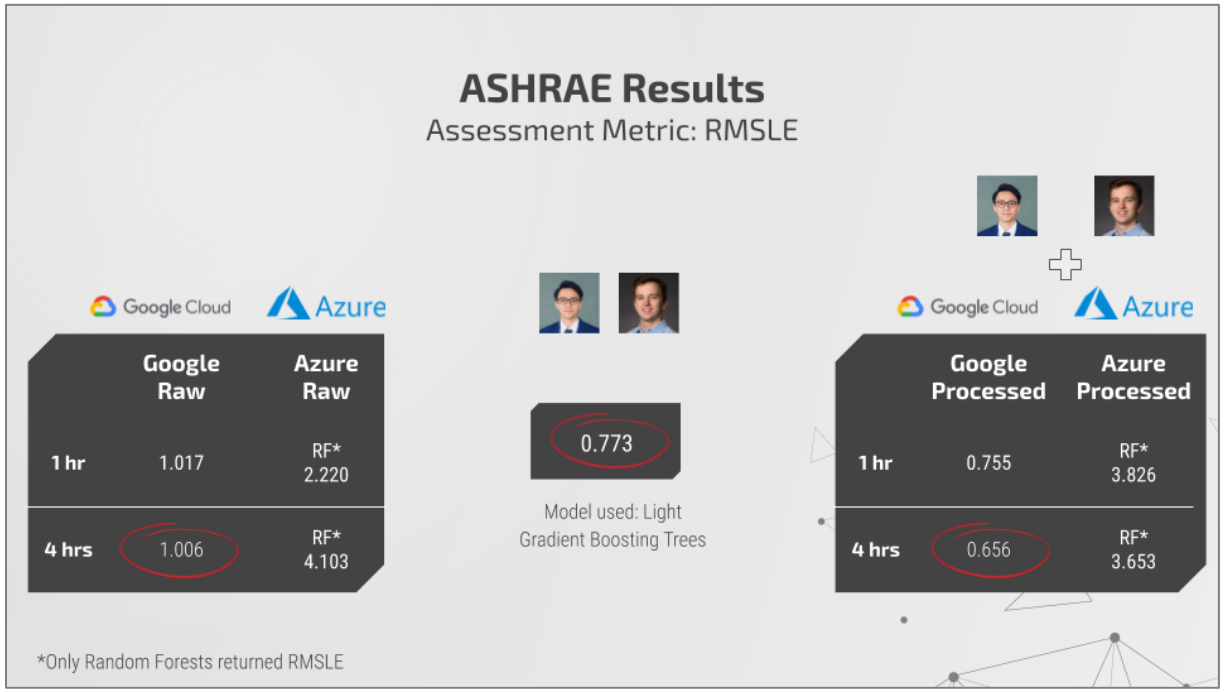
What we did		Effort required	Effectiveness/ Local model	Effectiveness/ AutoML
Handle missing values	Interpolation and Interpretation	A lot	😀	😀
Delete anomalous data	Delete site 0 data before June	A lot of EDA	😀	😀
Divide and conquer	Train a model for each building, i.e. train 1,499 models	Write a for loop	😀	
New features	Apply domain knowledge to add features such as holidays, weekend, lag terms # of features: **13 -> 32**	Think hard	😀	😀
Feature selection	Apply forward, backward and stepwise selection # of features: **32 -> 13**	Hours	😀	☹

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms



AUC - ROC curve (Area Under The Curve - Receiver Operating Characteristics)
It is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells **how much the model is capable of distinguishing between classes**. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.



RMSLE (Root Mean Squared Logarithmic Error)
It is a common metric for regression problems
This metric measures the **ratio between actual values and predicted values and takes the log of the predictions and actual values**. Use this instead of RMSE if an under-prediction is worse than an over-prediction. You can also use this when you don't want to penalize large differences when both of the values are large numbers.

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms

Takeaways from Speed Dating Dataset:

• Data Scientists can add value by providing well feature engineered datasets to AutoML platforms.

• Azure is more transparent in informing which model was used in the prediction; Google's model creation and selection information are proprietary.

• Google does not handle one-hot encoded variables well.

Takeaways from ASHRAE Dataset:

• Though AutoML is a powerful tool for prediction, it cannot preprocess data well enough to consistently outperform a human.

• A few extra hours of training can considerably increase the performance of an AutoML platform.

• Allow AutoML platforms to select features for you; otherwise, you run the risk of heavily restricting the platform's performance.

• Combining a data scientist's expertise of the business problem with AutoMLs' feature selection, feature preprocessing, model selection, and hyper-parameter tuning capabilities is a potent solution to deriving valuable insights and strong predictive results.

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms

Will AutoML replace Data Scientists / AI Engineers?

The answer is **NO**.

While AutoMLs are good at building models, they are still not capable of doing most of a data scientist's job. We still need data scientists to **define business problems**. We still need data scientists to apply their domain knowledge to generate more useful features. AutoML nowadays can only deal with limited types of problems such as classification and regression problems. Currently, they do not have the capability to build recommendation and ranking models. Most importantly, we still need data scientists to draw actionable insights from the data, and it can not be done by AutoMLs alone.

However, AutoMLs are still strong tools for data scientists to create values for their stakeholders.

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Data Scientists vs. AutoML Platforms

What is the main difference between a Data Scientist and an AutoML Solution?

> Data Scientists vs. AutoML Platforms

Some examples that might be worth considering.

• **Performance over interpretability -** Sometimes, the stakeholders may only care about the precision of models, and the interpretability is not the most crucial consideration.

• **Speed to production -** Both Google and Azure provide convenient ways to deploy your models into production.

• **Better use of your time -** Data scientists have a plethora of responsibilities that can be overwhelming. As a data scientist, time can be your scarcest resource. As a result, you can focus on tasks that produce the most value (sometimes spending time on preparing a fabulous presentation is worth more than improving 1% of the model's accuracy).

https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html

> Open Sources Alternatives

AutoWeka
An Automated Data Mining Software Based on Weka

https://youtu.be/m7kpIBGEdkI

AutoKeras

https://youtu.be/Hsw9z1GcAms

TPOT

https://youtu.be/QrJlj0VCHys

H2O.ai

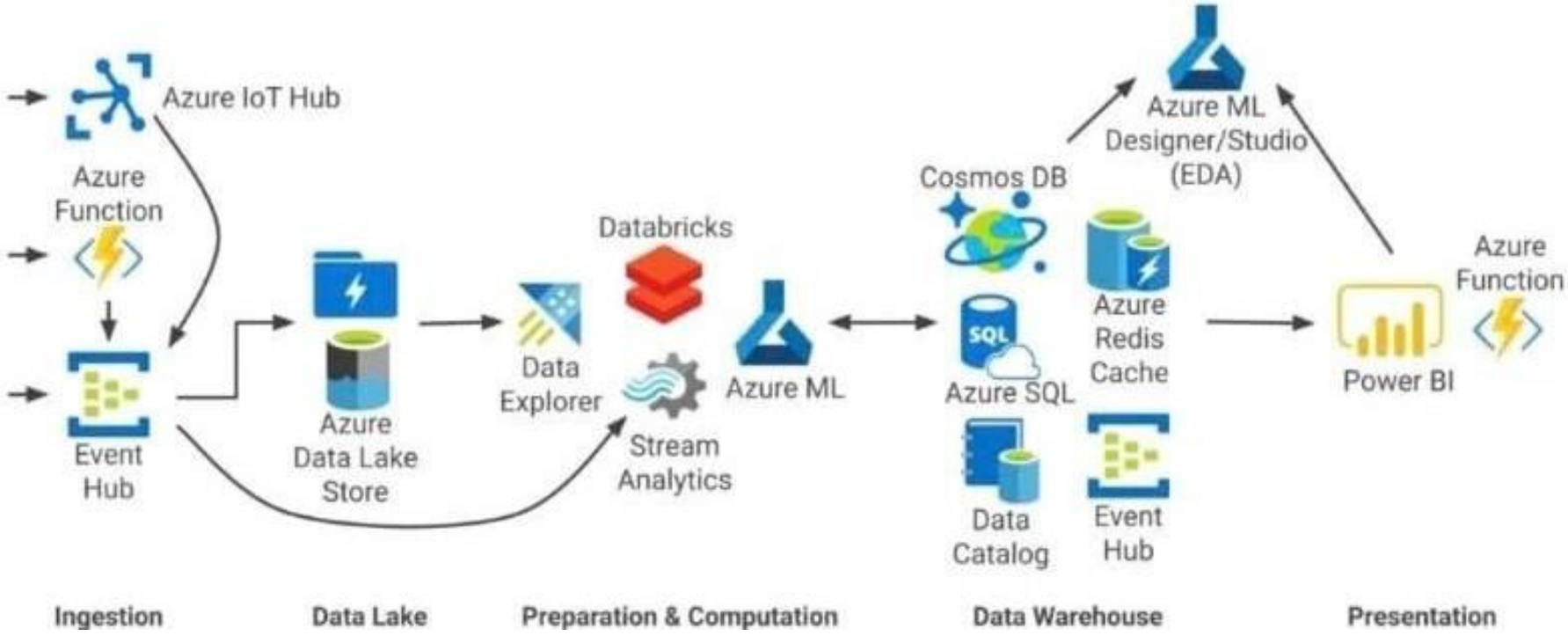
https://youtu.be/42Oo8TOl85I

salesforce
TransmogrifAI

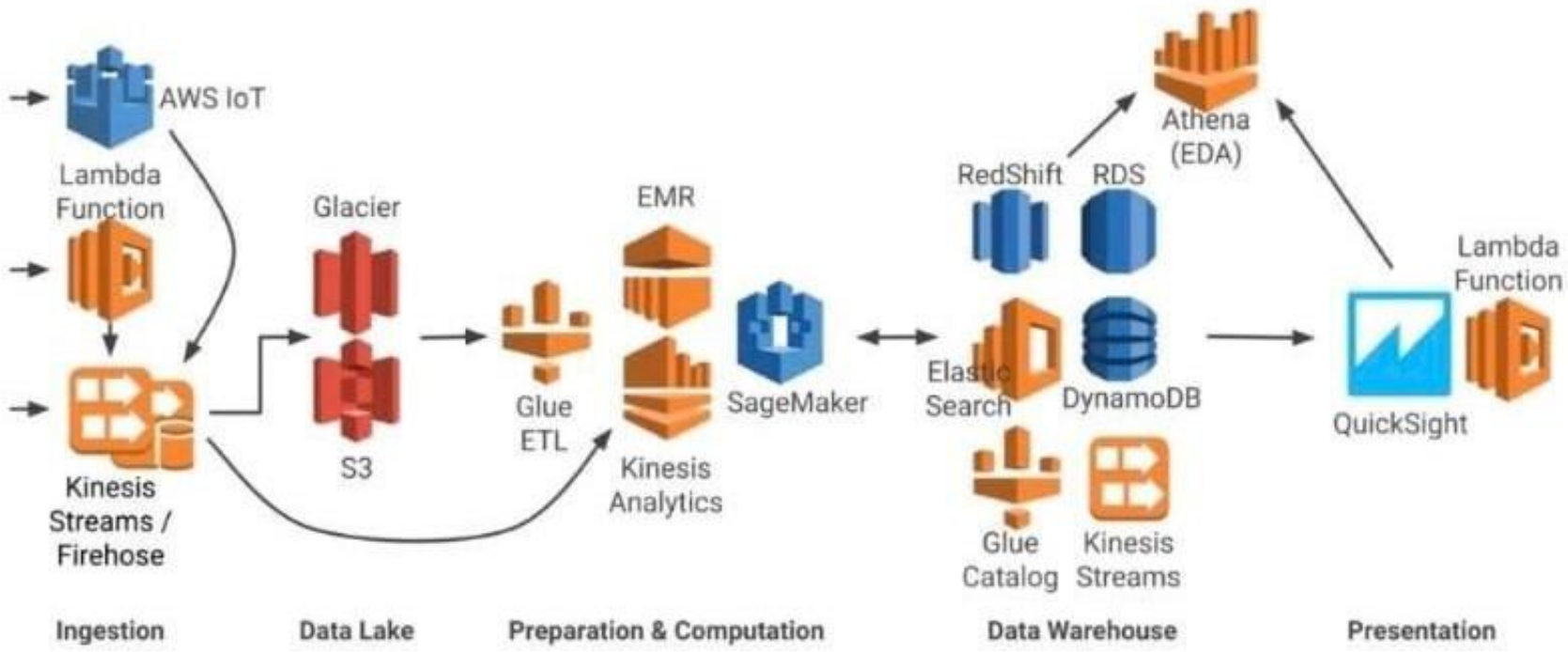
https://youtu.be/YDw1GieW4cw

Data Pipelines

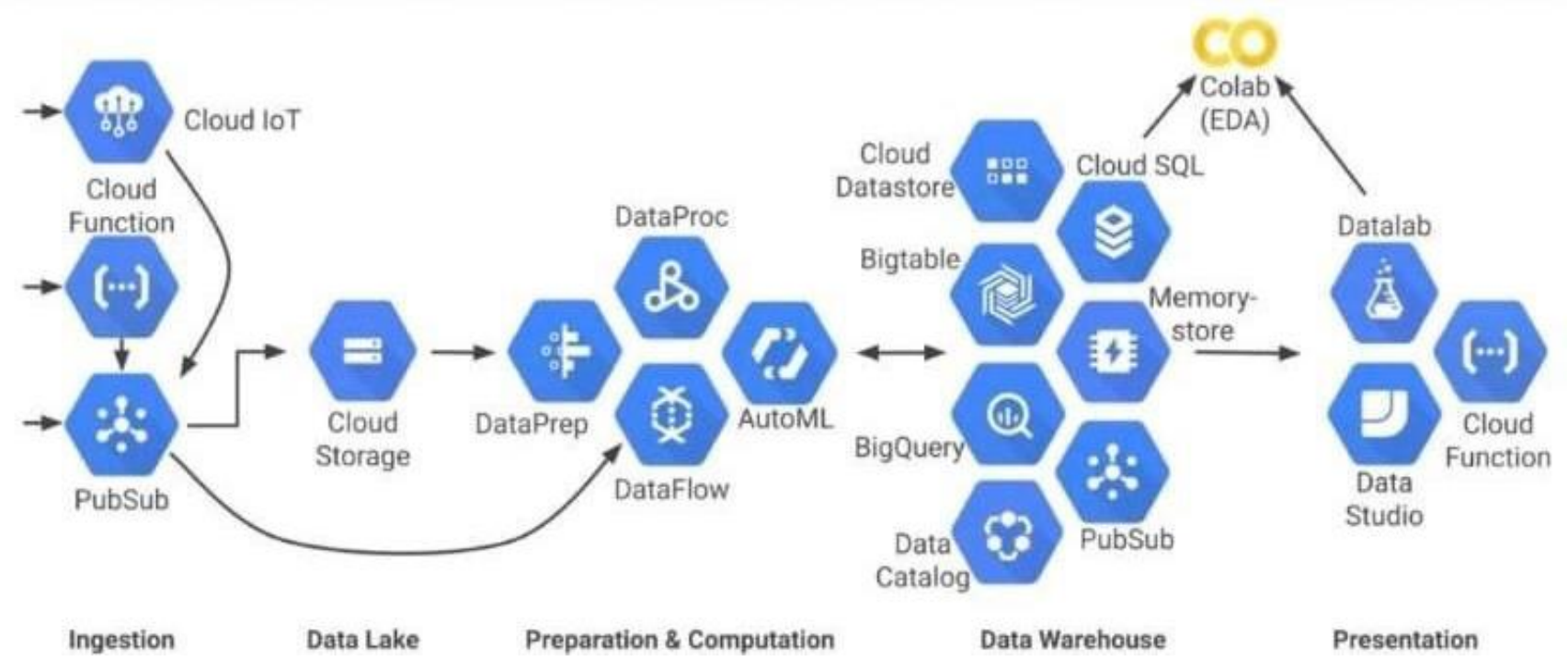
> Data Pipelines // Azure



> Data Pipelines // AWS



> Data Pipelines // GCP



References

> References (1)

- Big Data Analytics Program, 2019/2020 – Georgian College, Barrie, Ontario
- Git Hub, ss - https://mlops.githubapp.com/
- Informatica – Data Catalog – https://www.informatica.com/ca/products/data-catalog.html
- Google Research – Publication Data Management Challenges in Production Machine Learning, Alkis Polyzotis, Martin A. Zinkevich, Steven Whang, Sudip Roy - https://research.google/pubs/pub46178/
- https://research.google/pubs/pub46178/
- https://databricks.com/session_na20/an-approach-to-data-quality-for-netflix-personalization-systems
- Microsoft Azure Machine Learning Studio – Official Documentation, December 2020 - https://docs.microsoft.com/en-us/azure/machine-learning
- Microsoft Azure Machine Learning Cheat Sheet. Machine Learning Algorithm Cheat Sheet for Azure Machine Learning designer, December 2020. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet
- DataCookbook, Understanding Data Lineage, https://youtu.be/lUxgWb6WpF0
- Microsoft, What is Azure Data Catalog?, https://docs.microsoft.com/en-us/azure/data-catalog/overview
- Microsoft, Azure Data Platform End-to-End, Implement a Modern Data Platform Architecture, Official Material
- H2O.ai, RMSLE, website, https://www.h2o.ai/community/glossary/rmsle-root-mean-squared-logarithmic-error
- Kaggle, Understanding the metric: RMSLE, website, https://www.kaggle.com/carlolepelaars/understanding-the-metric-rmsle

> References (2)

- Big Data Analytics Program, 2019/2020 – Georgian College, Barrie, Ontario
- Kaggle, dataset: Forest Fires in Brazil, Luis Gustavo Modeli, https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil
- Data Iku, ebook, 2021 Trends: Where Enterprise AI is headed next?
- Microsoft, Success by Design Implementation Guide, First Edition, 2021
- Microsoft, Quickstart: Create your first data science experiment in Machine Learning Studio (classic), https://docs.microsoft.com/en-us/azure/machine-learning/classic/create-experiment
- Microsoft, Automated ML, https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml
- KD Nuggets, website, May 2021, https://www.kdnuggets.com/2020/02/data-scientists-automl-replace.html
- Microsoft, Tutorial 1: Predict credit risk, https://docs.microsoft.com/en-us/azure/machine-learning/classic/tutorial-part1-credit-risk
- Microsoft, Tutorial 2: Train credit risk models, https://docs.microsoft.com/en-us/azure/machine-learning/classic/tutorial-part2-credit-risk-train
- Microsoft, Tutorial 3: Deploy credit risk model, https://docs.microsoft.com/en-us/azure/machine-learning/classic/tutorial-part3-credit-risk-deploy
- Towards Data, Understanding AUC - ROC Curve, website, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

END OF DAY 4