Know Your data
& Build Predictive Modeling

IBM CODE

Meshael AlMuhanna, Unified Governance & Integration Technical Specialist.

Hissah AlMuneef, Cloud Developer Advocate

Meaad AlRshoud, Cloud Developer Advocate

IBM

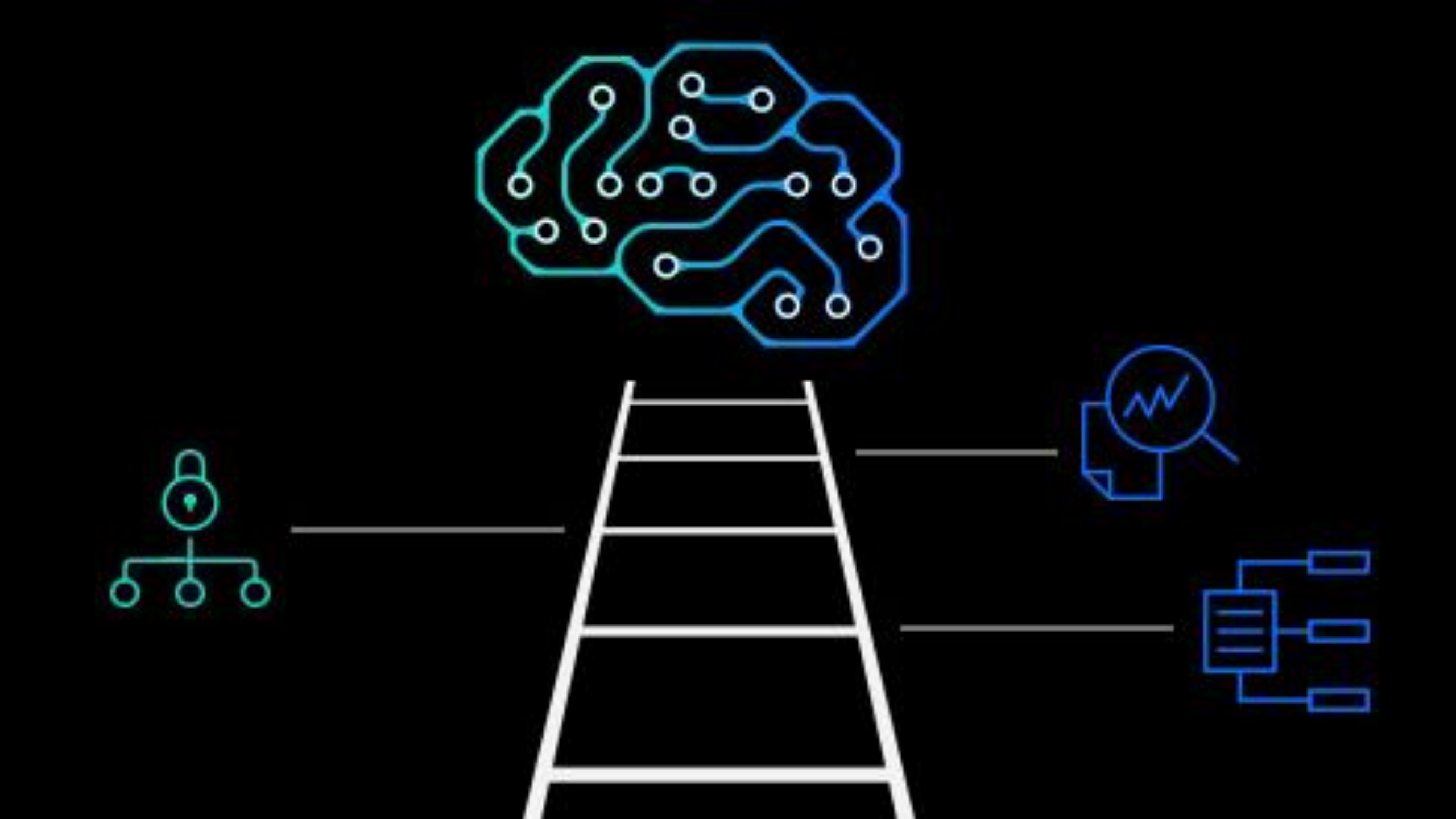
❖ Agenda:

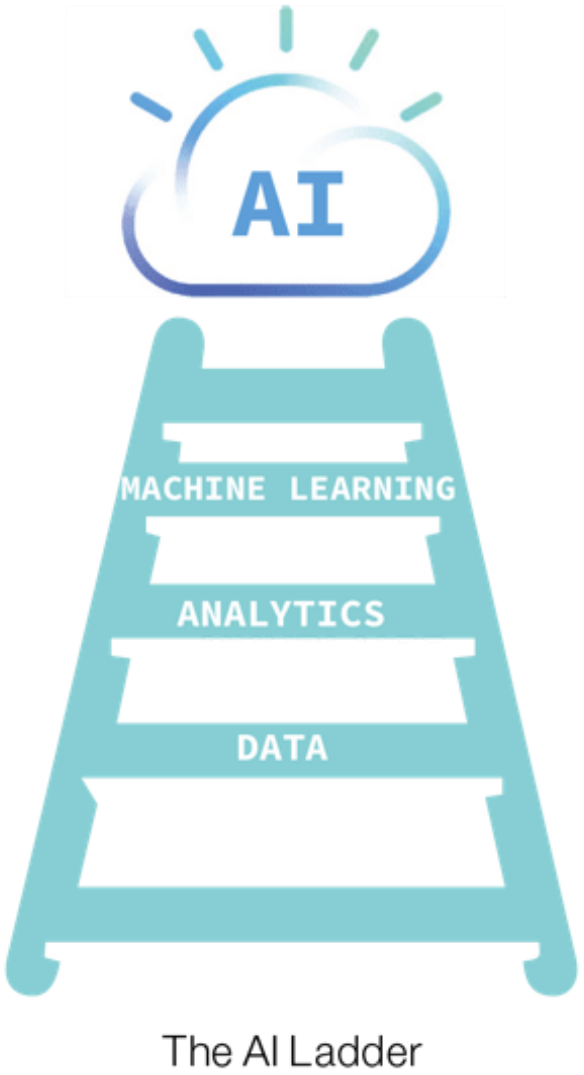
- IBM's AI ladder.

- Demonstration of data quality and ETL tools.

- Watson Studio overview.

- Predictive model use case.





The AI Ladder

IBM's Steps to Successful AI Journey

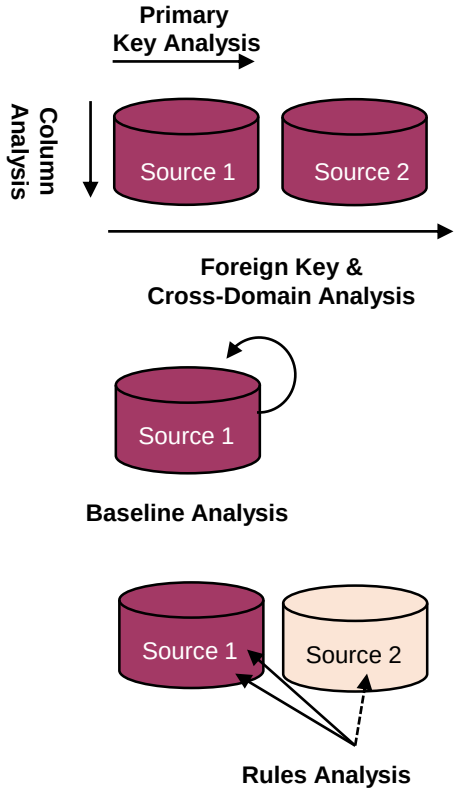
IBM platforms deliver the capabilities our clients need

Collect	Organize	Analyze
Hybrid Data Management	Unified Governance & Integration	Data Science & Business Analytics
		
▪ Collect all types of data, structured and unstructured ▪ Includes all open sources of data ▪ Leverages a single platform with a common application layer ▪ Write once and deploy anywhere	▪ Satisfy all matters of finding, cataloging and masking data ▪ Integrates fluid data sets ▪ Delivers built-in compliance ▪ Leverages advanced machine learning capabilities	▪ Delivers descriptive, prescriptive and predictive insights across all types of data ▪ Empowers all your teams and their unique use cases ▪ Enables advanced analytics and data science methods
Offerings: • Db2 & Db2 Warehouse • Db2 Event Store • Integrated Analytics System • Big SQL & Hortonworks Hadoop	*Offerings:* • Information Server • Data Replication • Master Data Management • Optim & StoredIQ	*Offerings:* • SPSS & DSX • Cognos & Watson Analytics • Watson Explorer • Planning Analytics

User & application independence across on premise, private cloud, and public cloud

Understand the Quality of Data Sources

- Data Quality Score: estimate the proportion of reliable data values in the given dataset.
- Run Quality Scanner to calculate quality score
- Declare the type of problem to scan and how many passes over the data
- Findings will be all aggregated
- Score will be calculated



Out of The Box Problem Detected

- Missing Values
  - Check missing values where Null values are not expected

- Uniqueness Violation
  - Check duplicate values

- Invalid Format
  - Checks for values
- Inconsistency Detection
  - Checks for values have different use of case

- Suspect Outlier
  - Checks for values that seem not to be of the same domain as other

- Violation of Correlation
  - Finds correlation between columns

- Data Rule Violation
  - Runs analysis against defined data rules

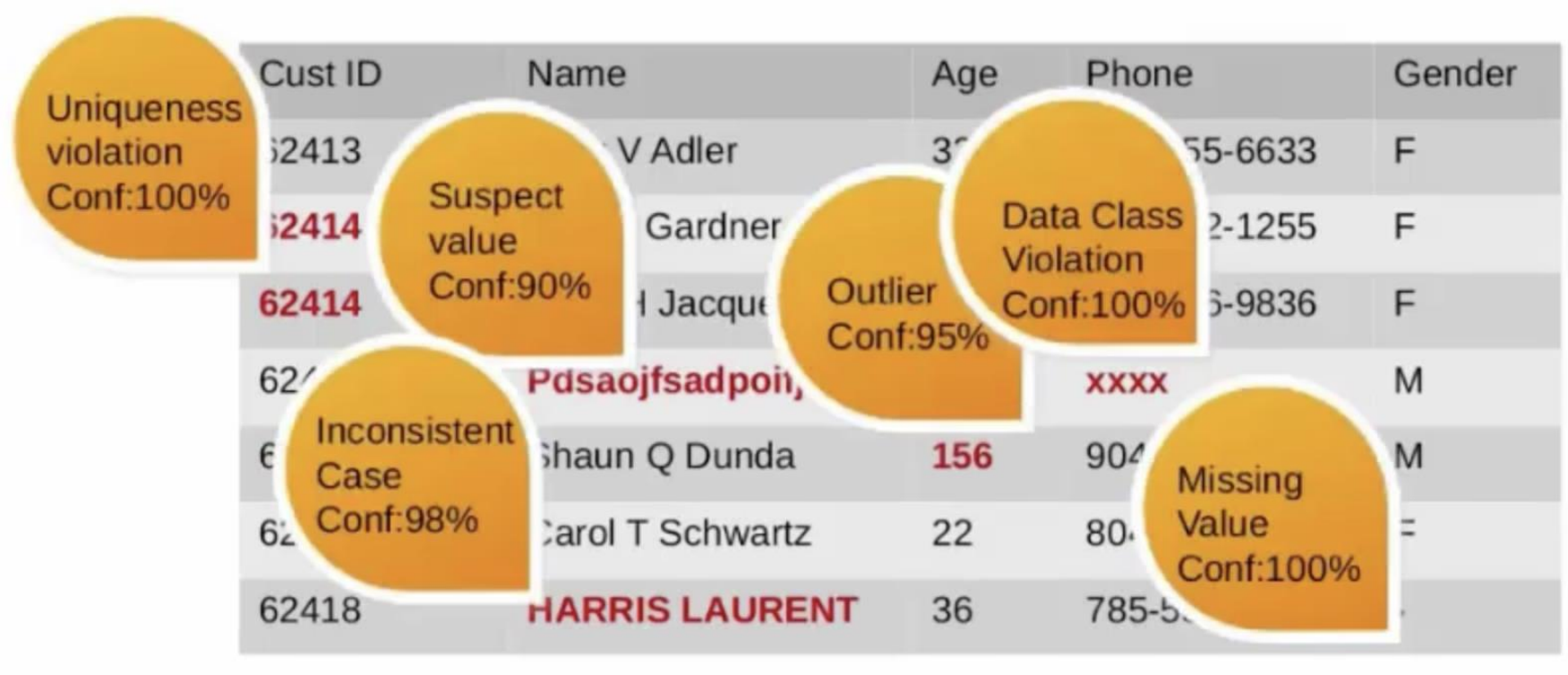
EXAMPLE

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	Pdsaojfsadpoifj	46	xxxx	M
62416	Shaun Q Dunda	156	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	HARRIS LAURENT	36	785-555-5835	-

EXAMPLE

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	**Pdsaojfsadpoifj**	46	**xxxx**	M
62416	Shaun Q Dunda	**156**	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	**HARRIS LAURENT**	36	785-555-5835	-

EXAMPLE


Uniqueness violation Conf:100%
Suspect value Conf:90%
Outlier Conf:95%
Data Class Violation Conf:100%
Inconsistent Case Conf:98%
Missing Value Conf:100%
Cust ID
Name
Age
Phone
Gender
62414
62414
156
xxxx
HARRIS LAURENT
62418
36

EXAMPLE

Data Set Score: 80%

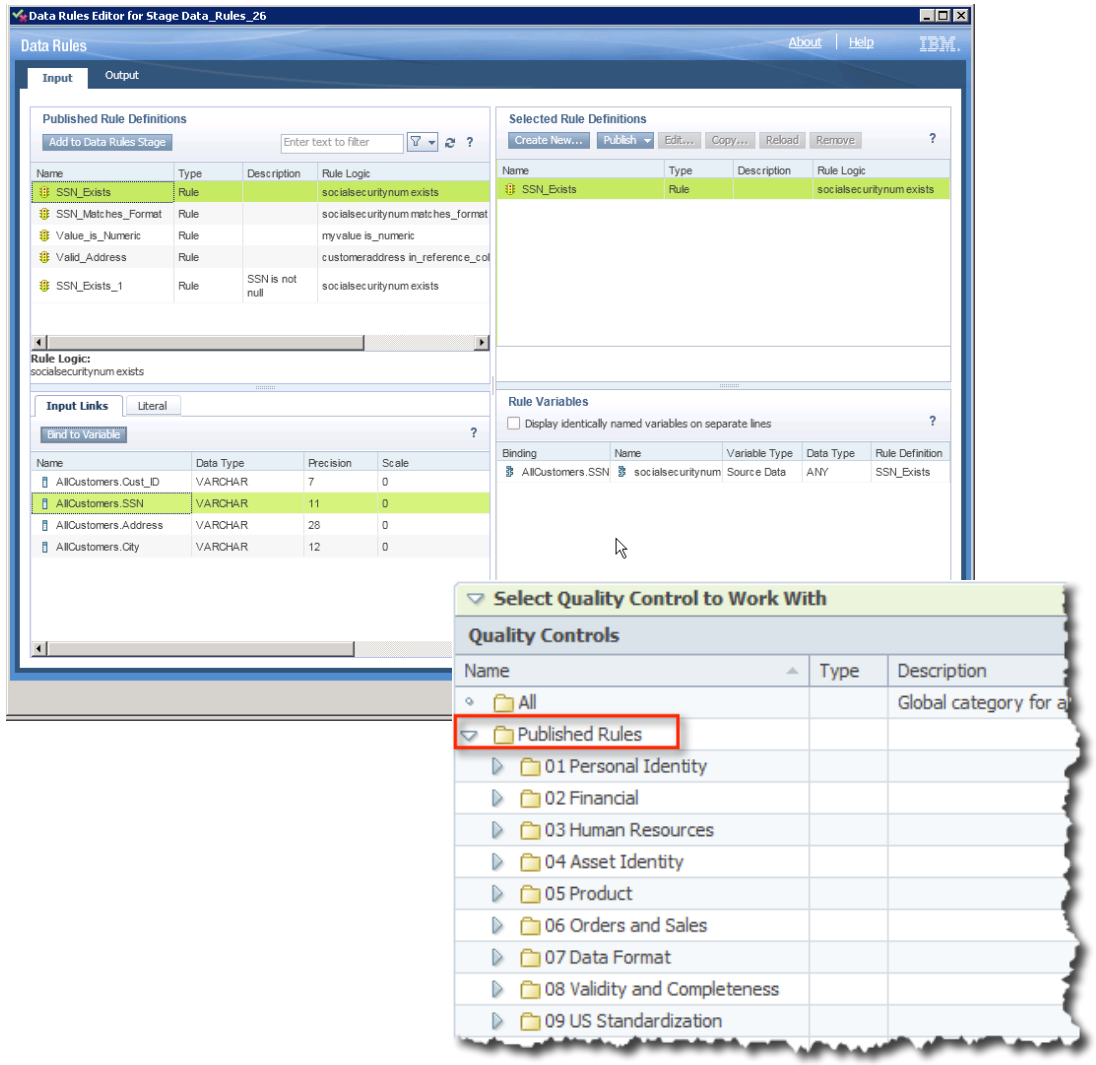
Score: 71%	Score: 73%	Score: 86%	Score: 85%	Score: 85%
Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	**Pdsaojfsadpoifj**	46	**xxxx**	M
62416	Shaun Q Dunda	**156**	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	**HARRIS LAURENT**	36	785-555-5835	-

Problems have been Identified, What’s Next?

Fix Identified Quality Issues

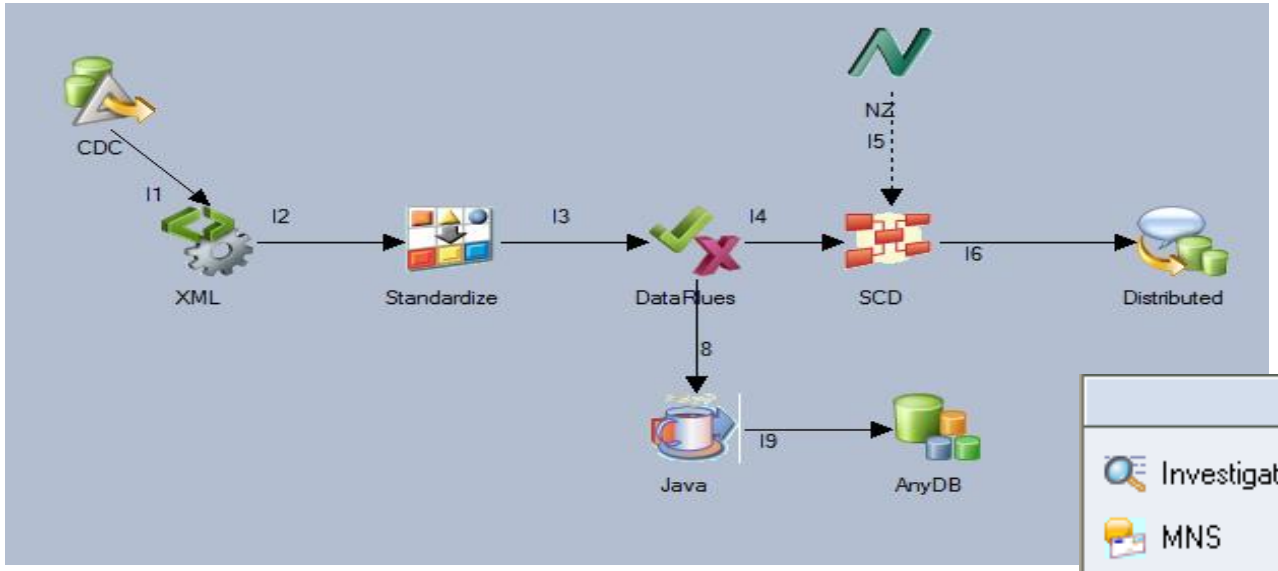
Examples of Rules:

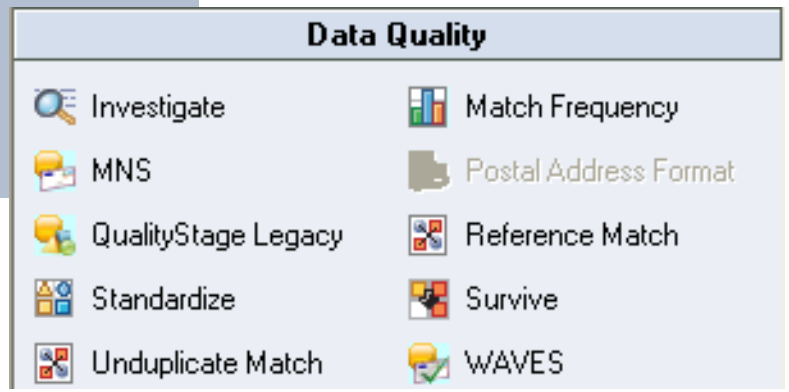
- The Gender field must be populated and must be in the list of accepted values
- The Social Security Number must be numeric and in the format 999-99-9999
- If Date of Birth Exists AND Date of Birth > 1900-01-01 and < TODAY
  Then Customer Type Equals ‘P’
- The Bank Account Branch ID is valid in the Branch Reference master list



Enforce Quality on data

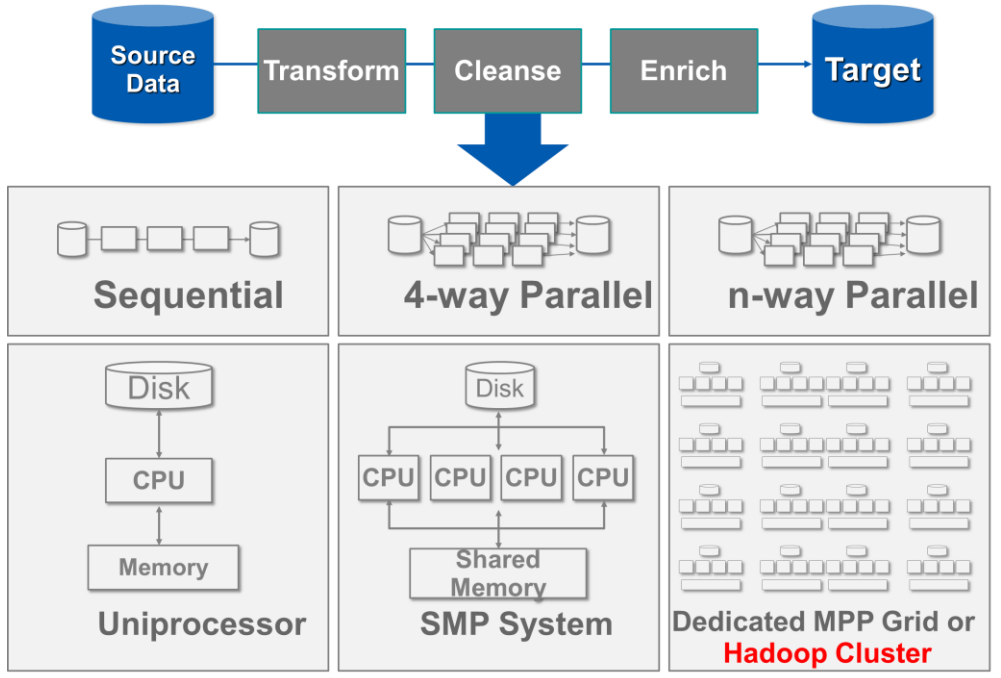
Fully integrated ETL & Data qualities capabilities

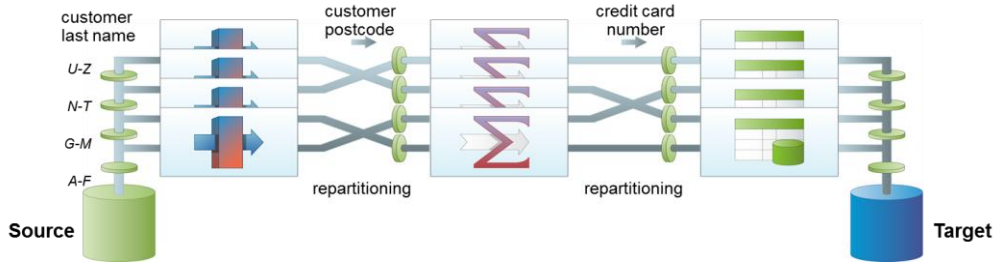




Why Information Integration is Important?

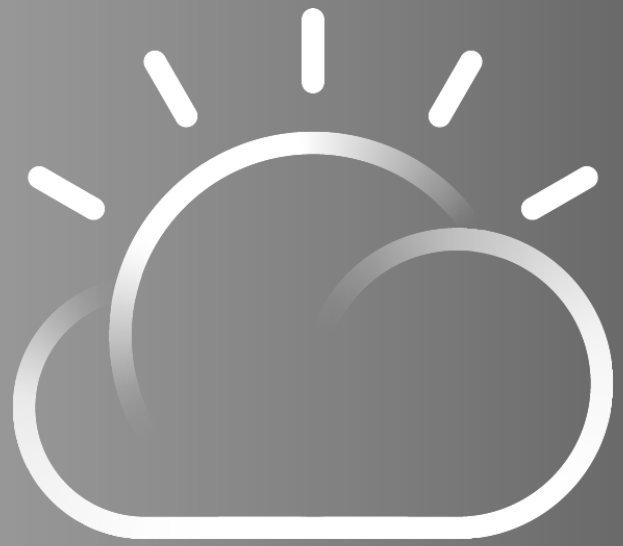
Performance





Connectivity

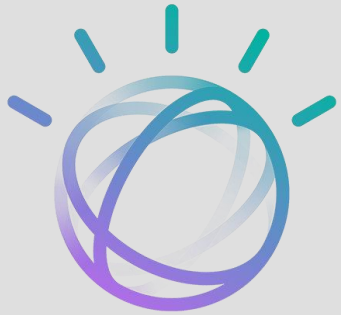
Predict Loan Eligibility Using SPSS in Watson Studio



IBM Cloud

Machine Learning

IBM Watson



Watson Studio

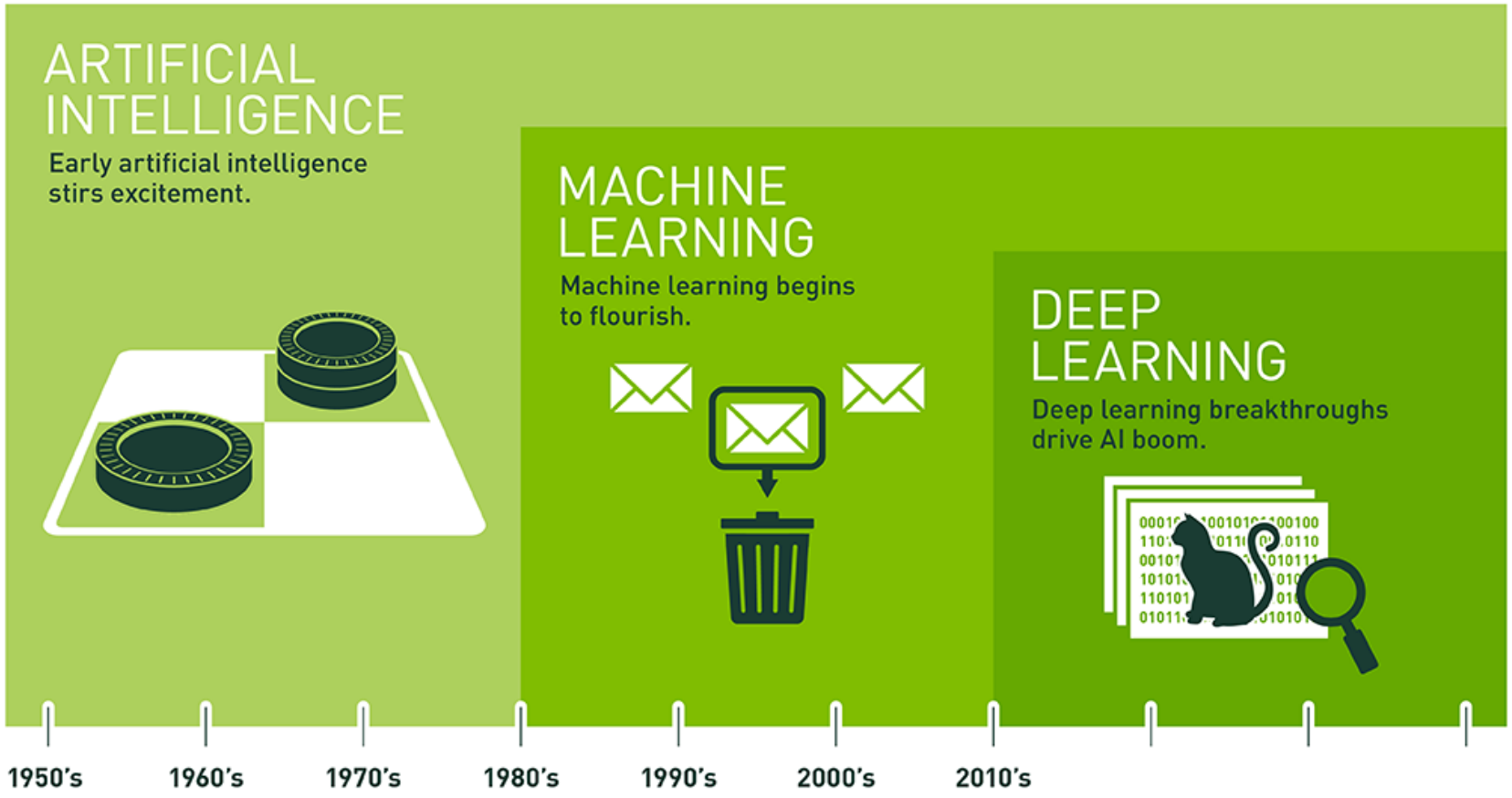
Loan Eligibility Predictive Model

IBM

Machine Learning

IBM **Code**

Concept



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Artificial Intelligence

Netflix

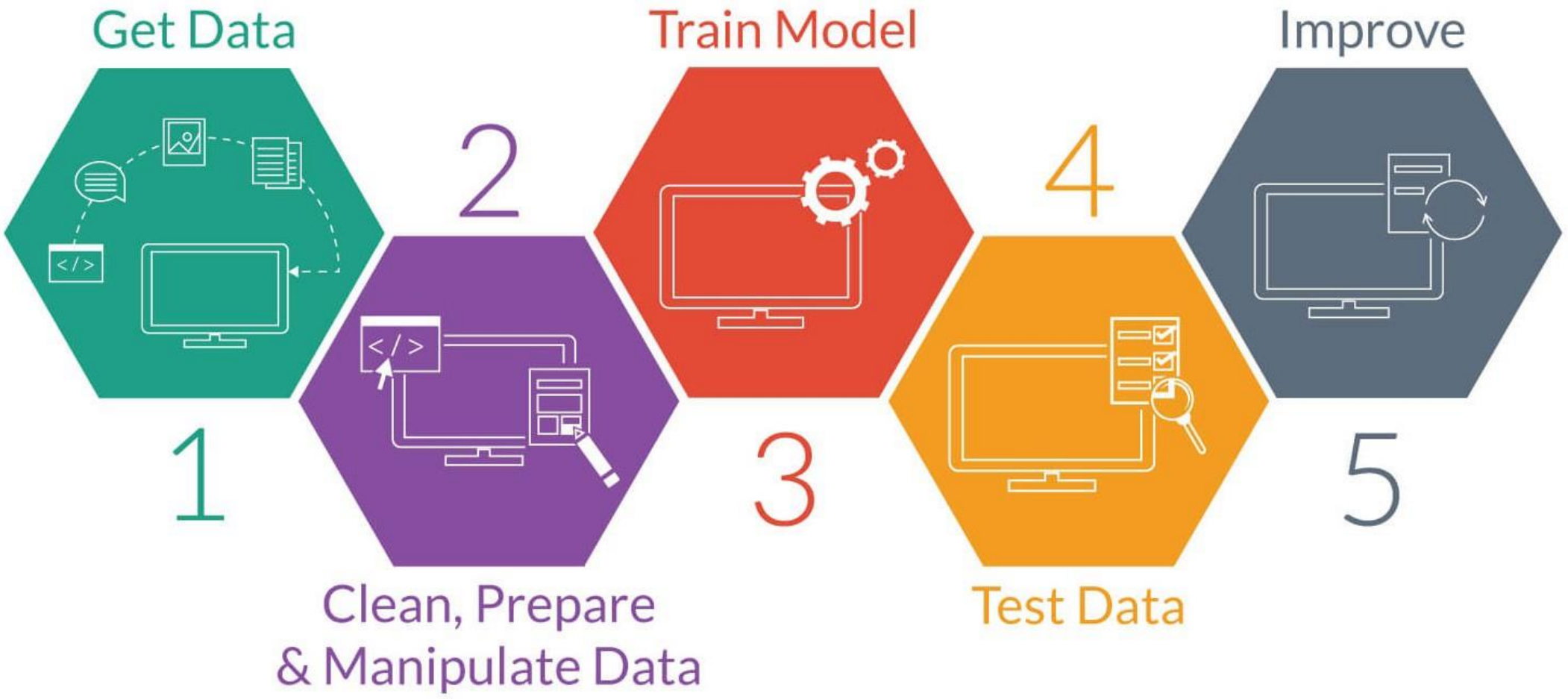
Machine learning is integral to Netflix's video recommendation engine. The company has valued the ROI of these algorithms at £1 billion a year due to their impact on customer retention.

PayPal

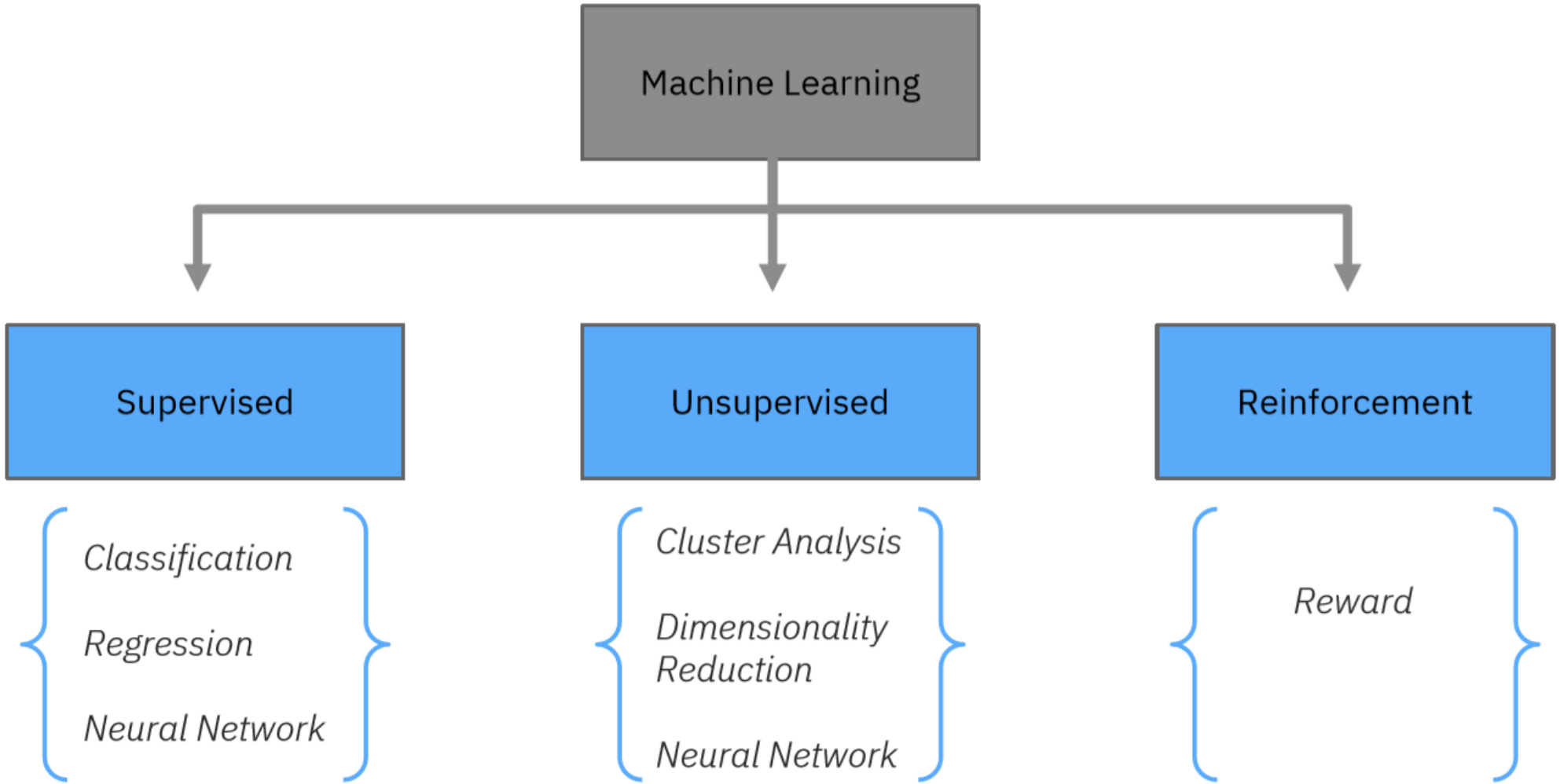
The online payment platform uses machine learning algorithms to combat fraud. By implementing deep learning techniques, PayPal analyses vast quantities of customer data and evaluates risk accordingly.

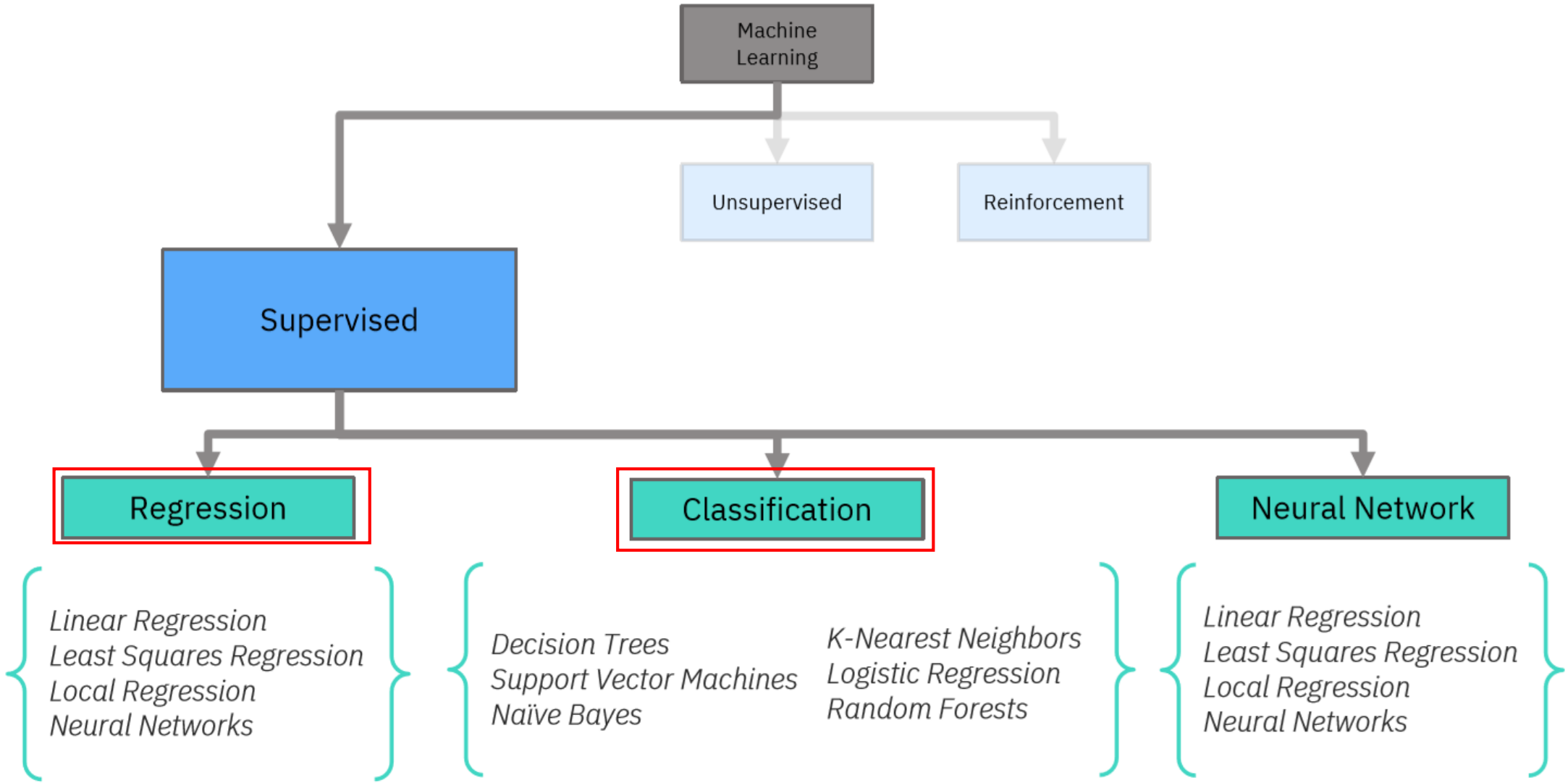
Machine Learning

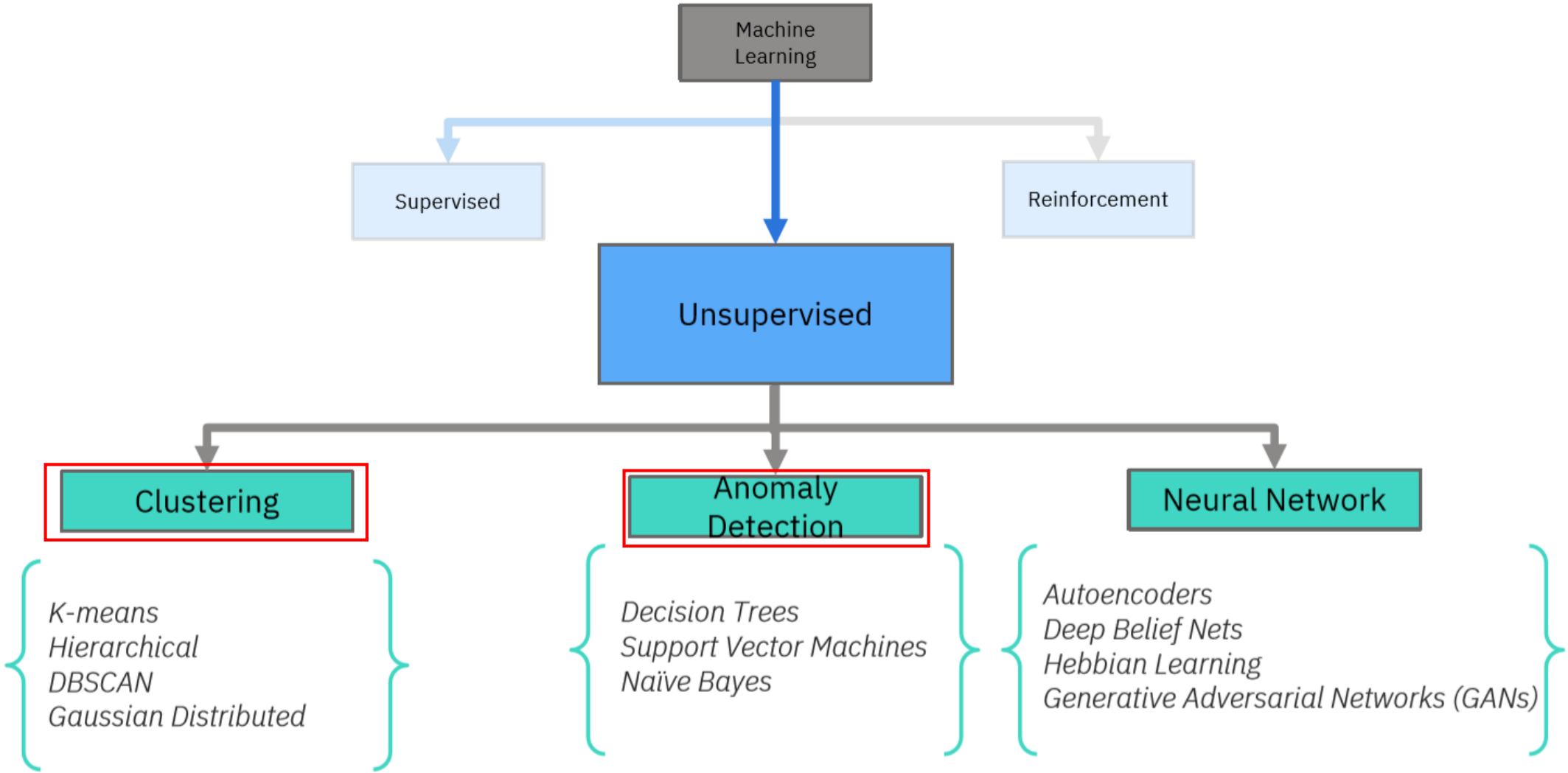
Methodology



IBM Code

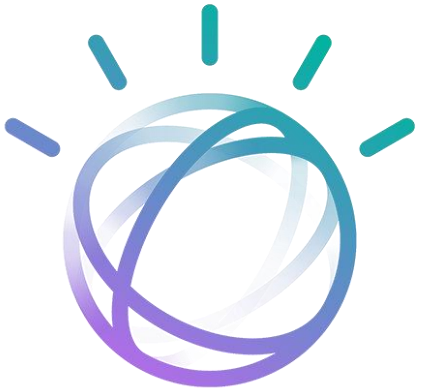

Machine Learning
Supervised
Unsupervised
Reinforcement
Classification
Regression
Neural Network
Cluster Analysis
Dimensionality Reduction
Neural Network
Reward


Machine Learning
Unsupervised
Reinforcement
Supervised
Regression
Classification
Neural Network
Linear Regression
Least Squares Regression
Local Regression
Neural Networks
Decision Trees
Support Vector Machines
Naïve Bayes
K-Nearest Neighbors
Logistic Regression
Random Forests
Linear Regression
Least Squares Regression
Local Regression
Neural Networks


Machine Learning
Supervised
Reinforcement
Unsupervised
Clustering
Anomaly Detection
Neural Network
K-means
Hierarchical
DBSCAN
Gaussian Distributed
Decision Trees
Support Vector Machines
Naïve Bayes
Autoencoders
Deep Belief Nets
Hebbian Learning
Generative Adversarial Networks (GANs)

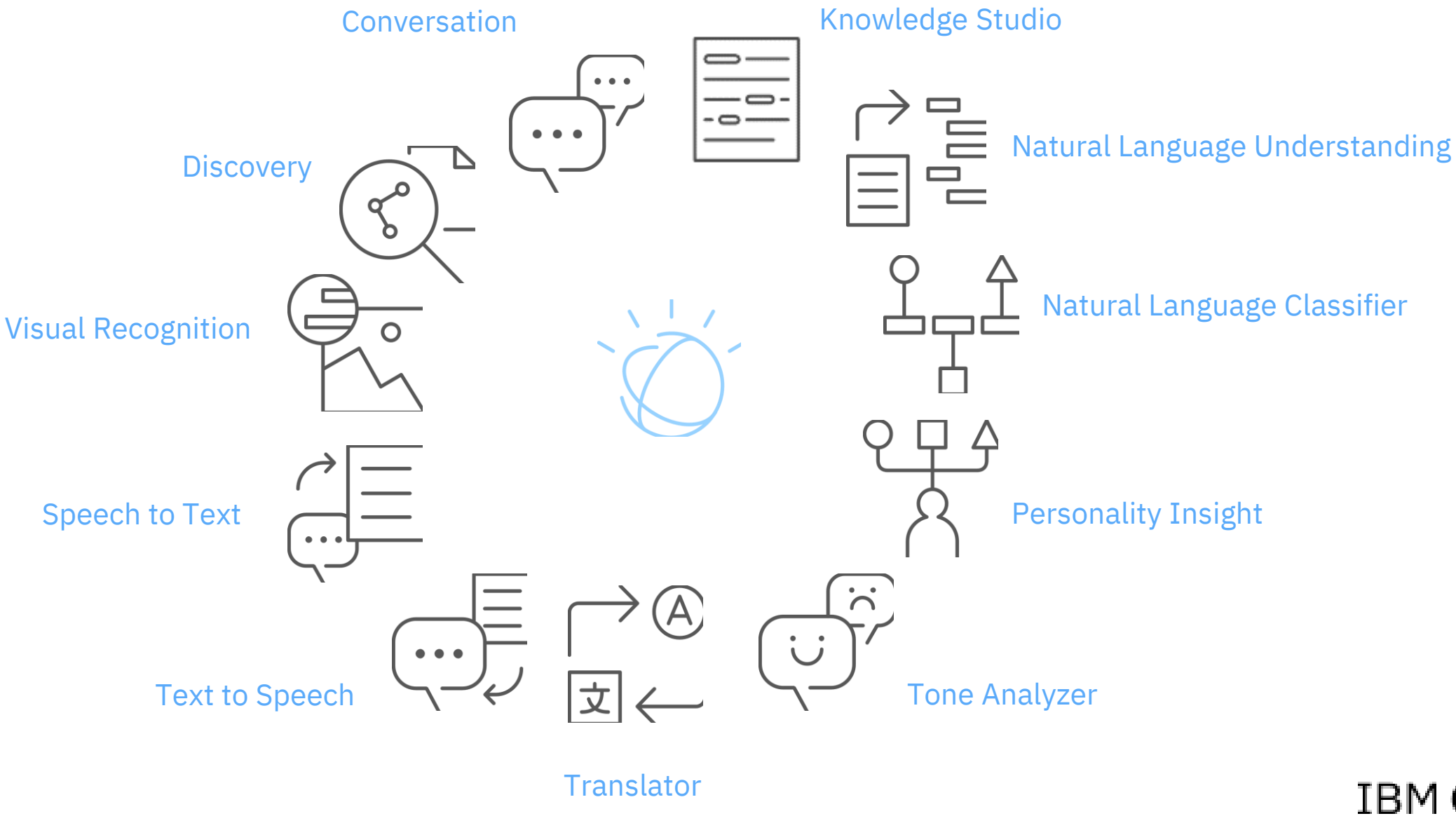
IBM

Watson is AI for Business



IBM **Code**

With Watson:

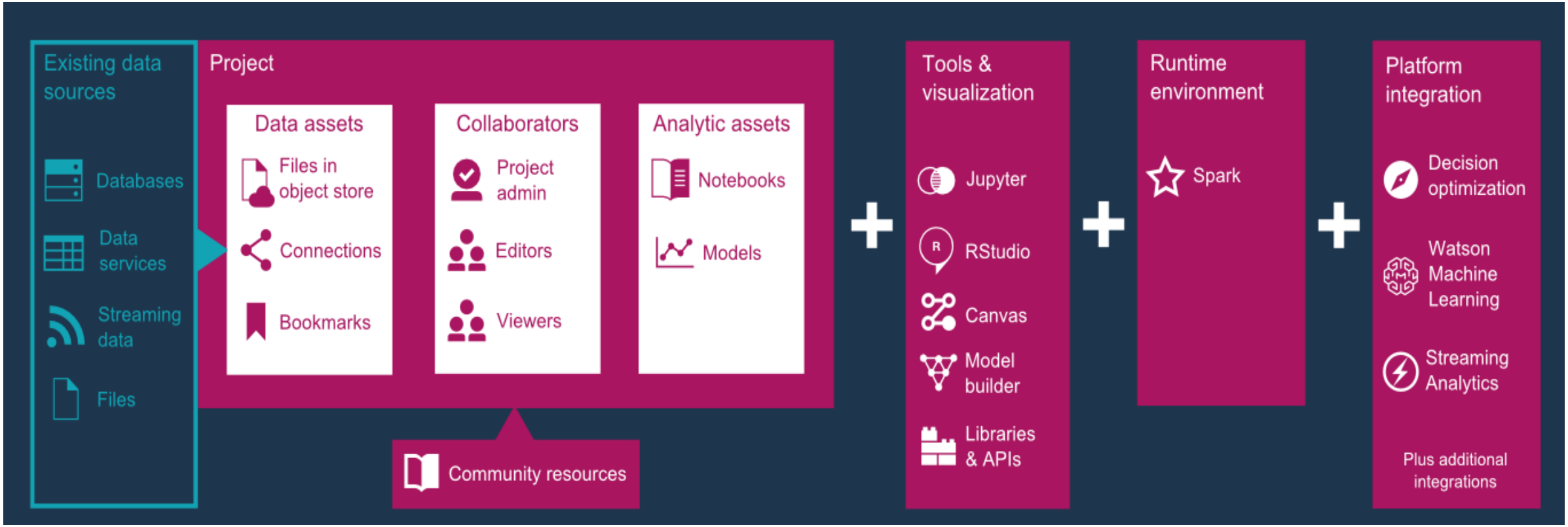


IBM Code

Watson Studio

IBM Code

Watson Studio



Predict Loan Eligibility Using SPSS in Watson Studio

IBM Code

Problem Statement

Loans Company wants to automate the loan eligibility process based on customer detail provided while filling online application form.

Data

Not Feature

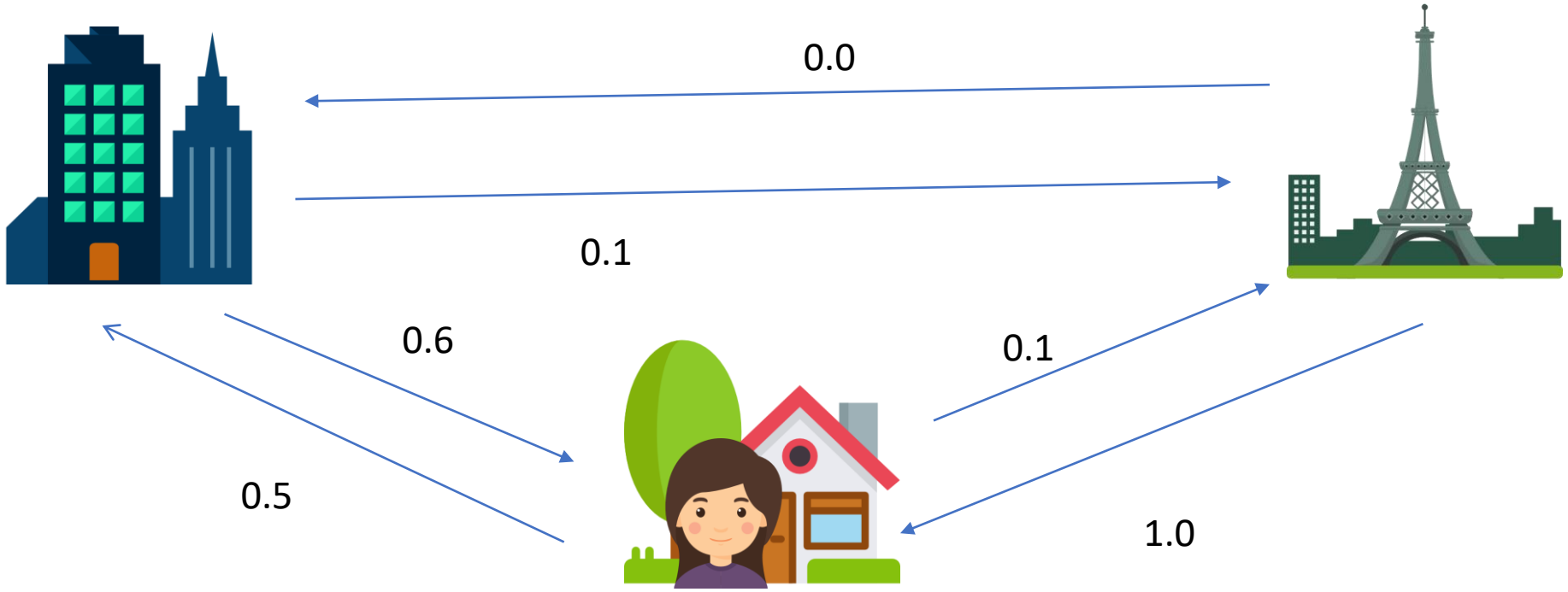
Features

Loan_ID String	Gender String	Married String	Dependents String	Education String	Self_Employed String	ApplicantIncome String
LP001002	Male	No	0	Graduate	No	5849
LP001003	Male	Yes	1	Graduate	No	4583
LP001005	Male	Yes	0	Graduate	Yes	3000

Class

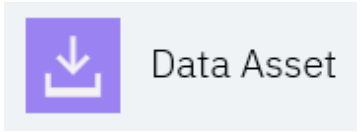
CoapplicantIncome Decimal	LoanAmount Decimal	Loan_Amount_Term Decimal	Credit_History Decimal	Property_Area String	Loan_Status String
0	146.412162	360	1	Urban	Y
1508	128	360	1	Rural	N
0	66	360	1	Urban	Y

Bayes Net



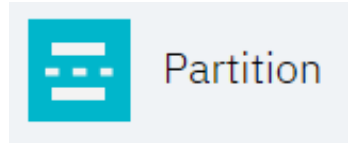
Steps to Solution…

1. Import our Data using **Data Asset** node.

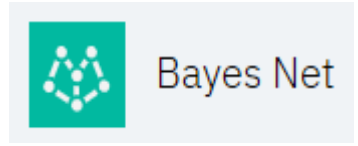


2. Configures variables type using **Types** node.

3. Split our data for training and testing sets using **Partition** node.



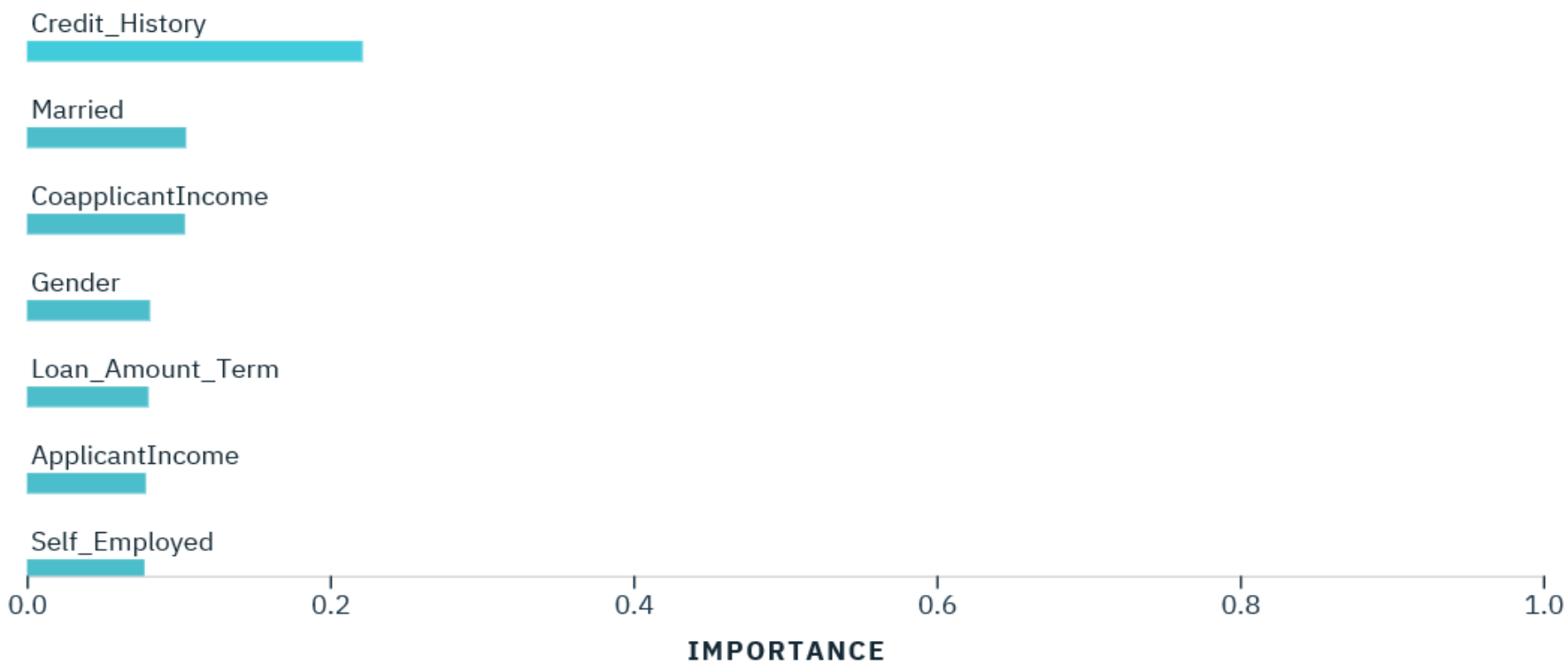
4. Build a probability model using Bayesian Network algorithm by the **Bayes Net** node.



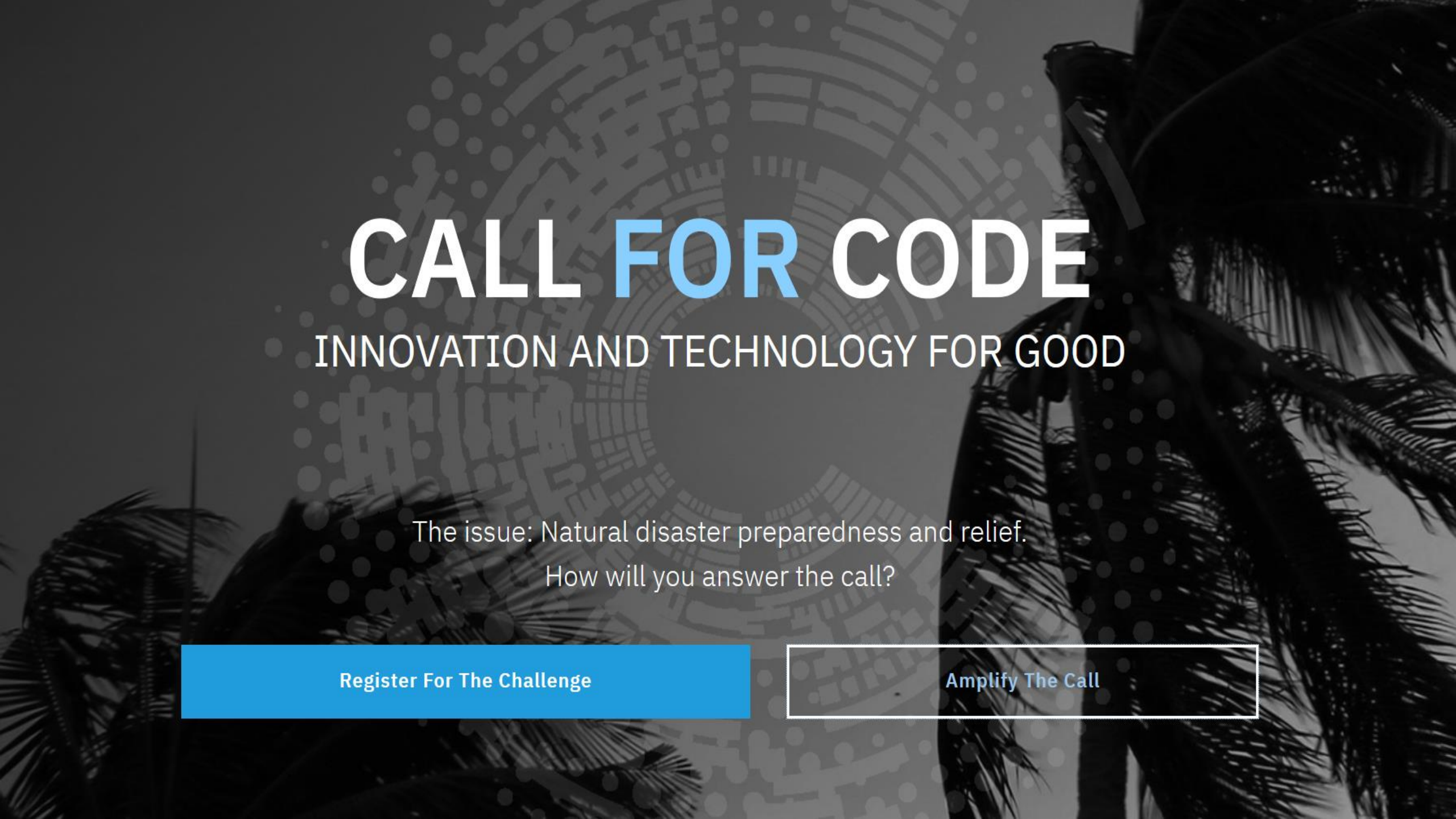
5. Try other models ! Why not !

Predictor Importance

TARGET : LOAN_STATUS



IBM Code

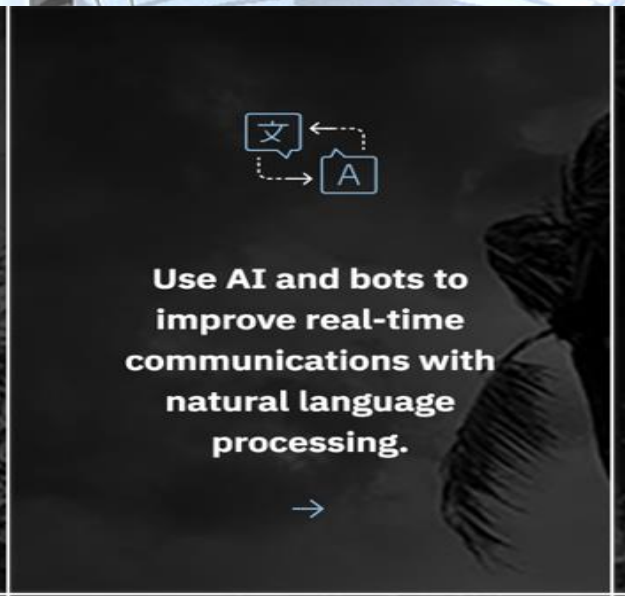

CALL FOR CODE
INNOVATION AND TECHNOLOGY FOR GOOD
The issue: Natural disaster preparedness and relief.
How will you answer the call?
Register For The Challenge
Amplify The Call

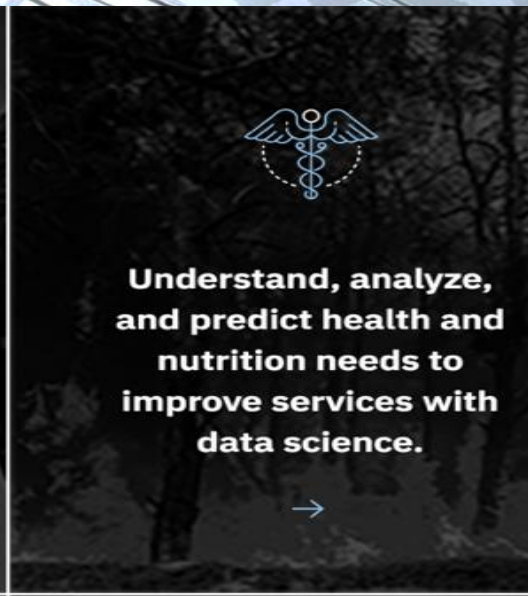
Get Started

Call for Code

Commit for a **CAUSE**. Push for **CHANGE**.

Call for Code Website:
https://developer.ibm.com/callforcode/
Challenge Details:
https://callforcode.org/challenge/





Resources

Learn – develop – connect

IBM Code (developer.ibm.com/code)
IBM Developer Works (ibm.com/developerworks)
GitHub (github.com/DevExCodeHub)

Learning Lab - Coursera - Udacity - more