Know Your data
& Build Predictive Modeling

IBM CODE

Meshael AlMuhanna, Unified Governance & Integration Technical Specialist.

Hissah AlMuneef, Cloud Developer Advocate

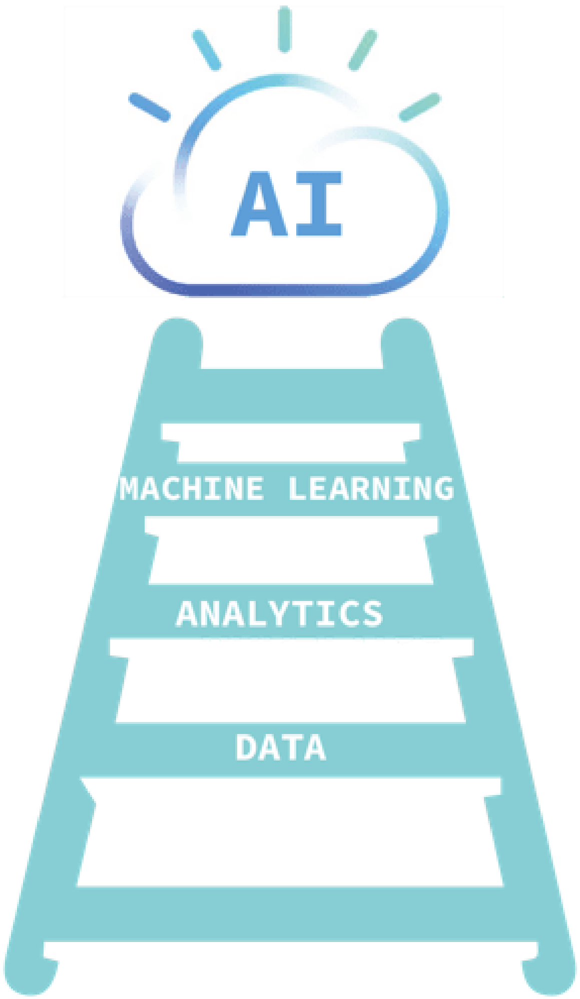
❖ Agenda:

- 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

Hybrid Data Management

- 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

Offerings:

- Db2 & Db2 Warehouse
- Db2 Event Store
- Integrated Analytics System
- Big SQL & Hortonworks Hadoop

Organize

Unified Governance & Integration

- Satisfy all matters of finding, cataloging and masking data
- Integrates fluid data sets
- Delivers built-in compliance
- Leverages advanced machine learning capabilities

Offerings:

- Information Server
- Data Replication
- Master Data Management
- Optim & StoredIQ

Analyze

Data Science & Business Analytics

- 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

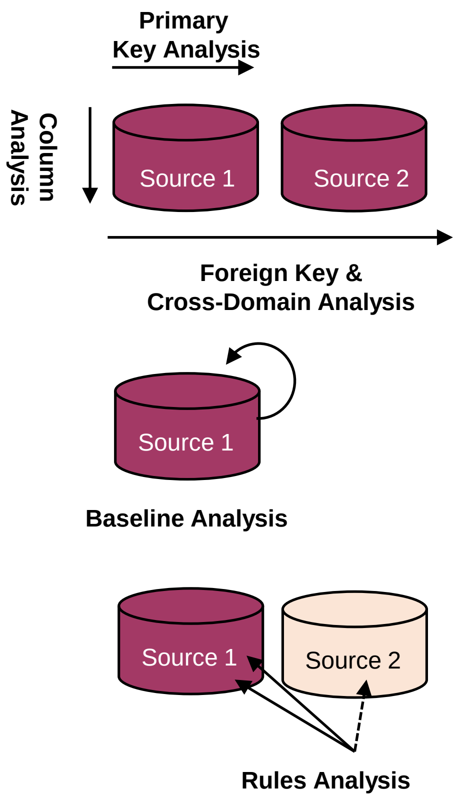
Offerings:

- 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

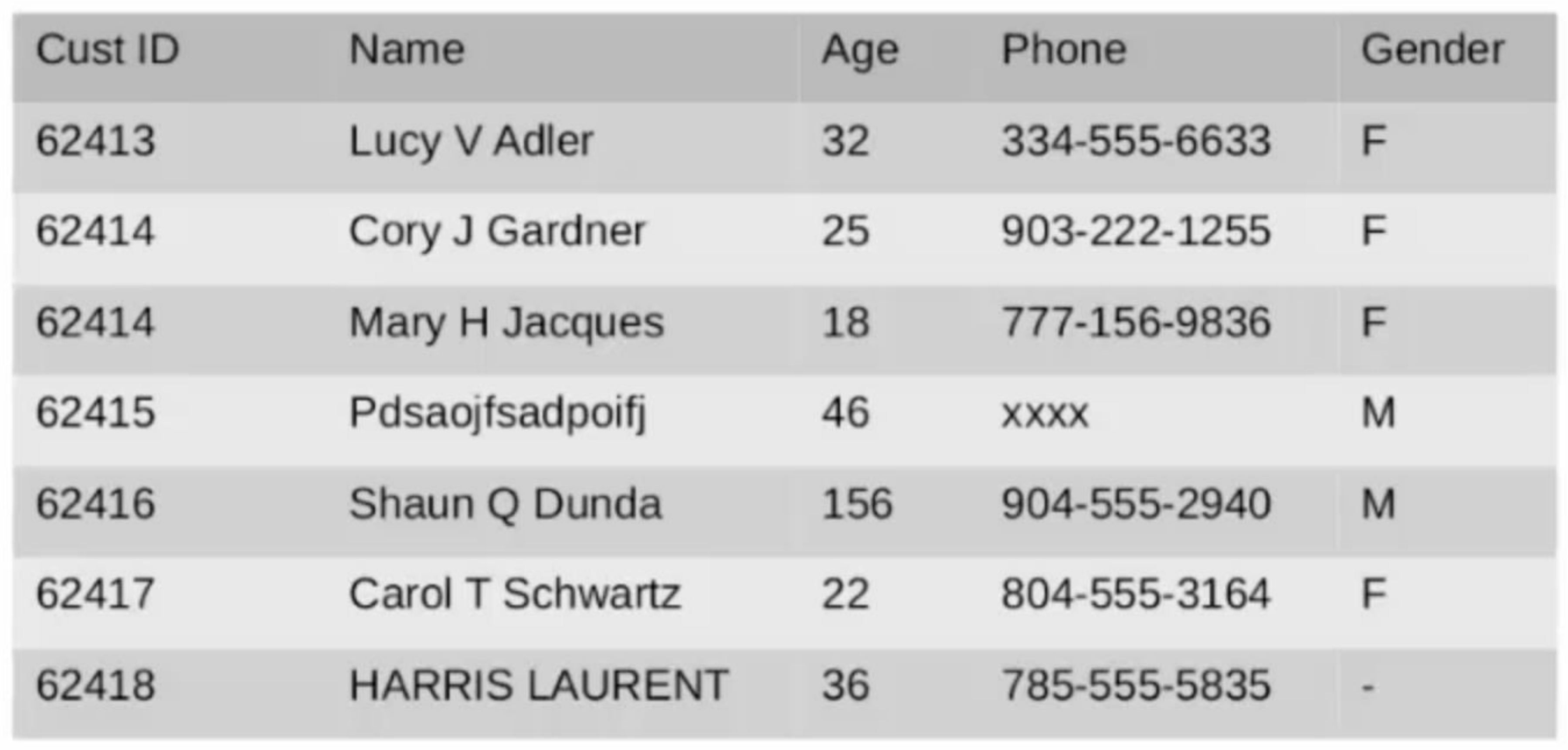
- Data Quality Score: estimate the proportion of reliable data values in the given dataset.
- Run Quality Scanner to calculate quality score
- Declare the type of problem to scan and how many passes over the data
- Findings will be all aggregated
- Score will be calculated



Out of The Box Problem Detected

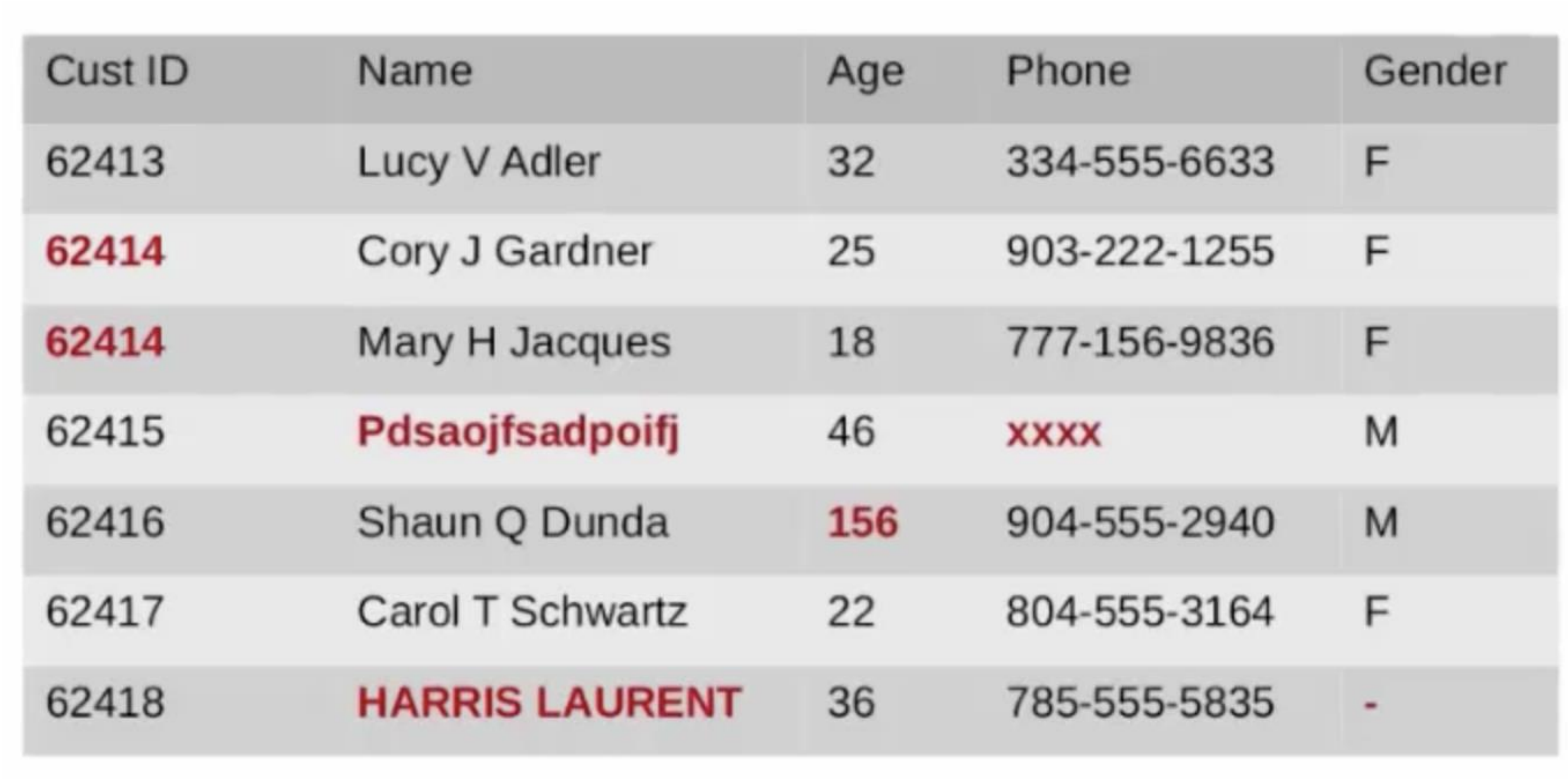
- Missing Values
  - Check missing values where Null values are not expected
- Uniqueness Violation
  - Check duplicate values
- Invalid Format
  - Checks for values
- Inconsistency Detection
  - Checks for values have different use of case
- Suspect Outlier
  - Checks for values that seem not to be of the same domain as other
- Violation of Correlation
  - Finds correlation between columns
- Data Rule Violation
  - Runs analysis against defined data rules

EXAMPLE



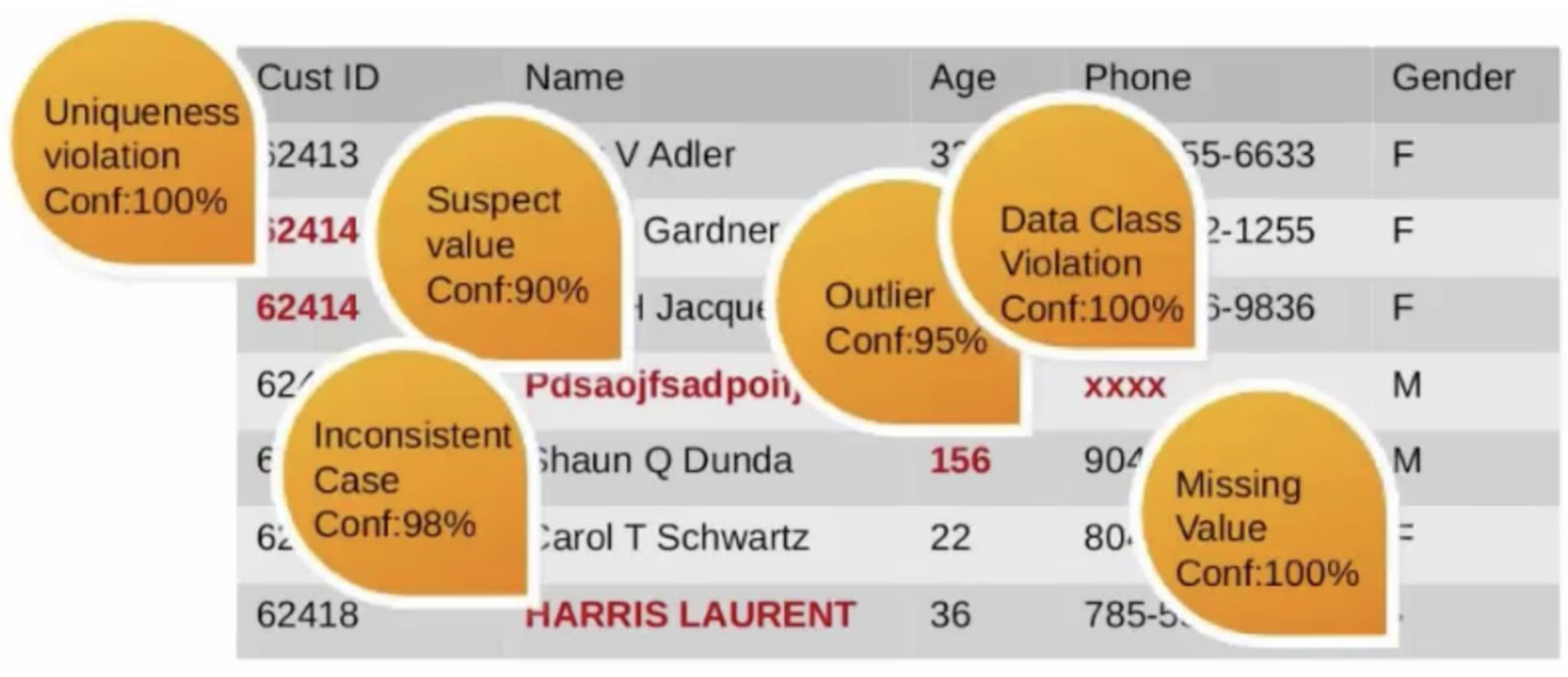
Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	Pdsaojfsadpoifj	46	xxxx	M
62416	Shaun Q Dunda	156	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	HARRIS LAURENT	36	785-555-5835	-

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	-

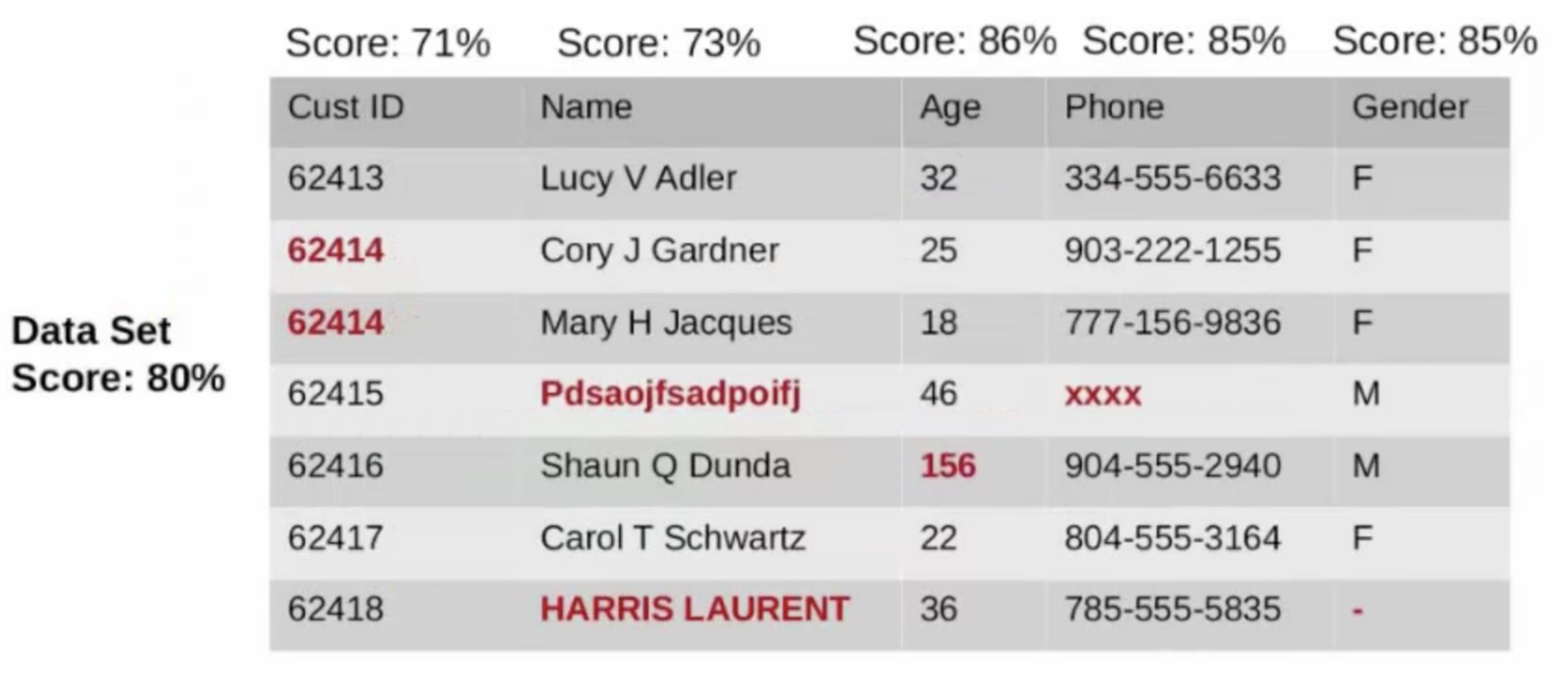
EXAMPLE



Cust ID	Name	Age	Phone	Gender
2413	V Adler	3	55-6633	F
2414	Gardner		2-1255	F
62414	Jacque		6-9836	F
62	Pdsaojfsadpoij		xxxx	M
6	haun Q Dunda	156	904	M
62	arol T Schwartz	22	80	F
62418	HARRIS LAURENT	36	785-5	

Uniqueness violation Conf:100%
Suspect value Conf:90%
Outlier Conf:95%
Data Class Violation Conf:100%
Inconsistent Case Conf:98%
Missing Value Conf:100%

EXAMPLE



Data Set Score: 80%

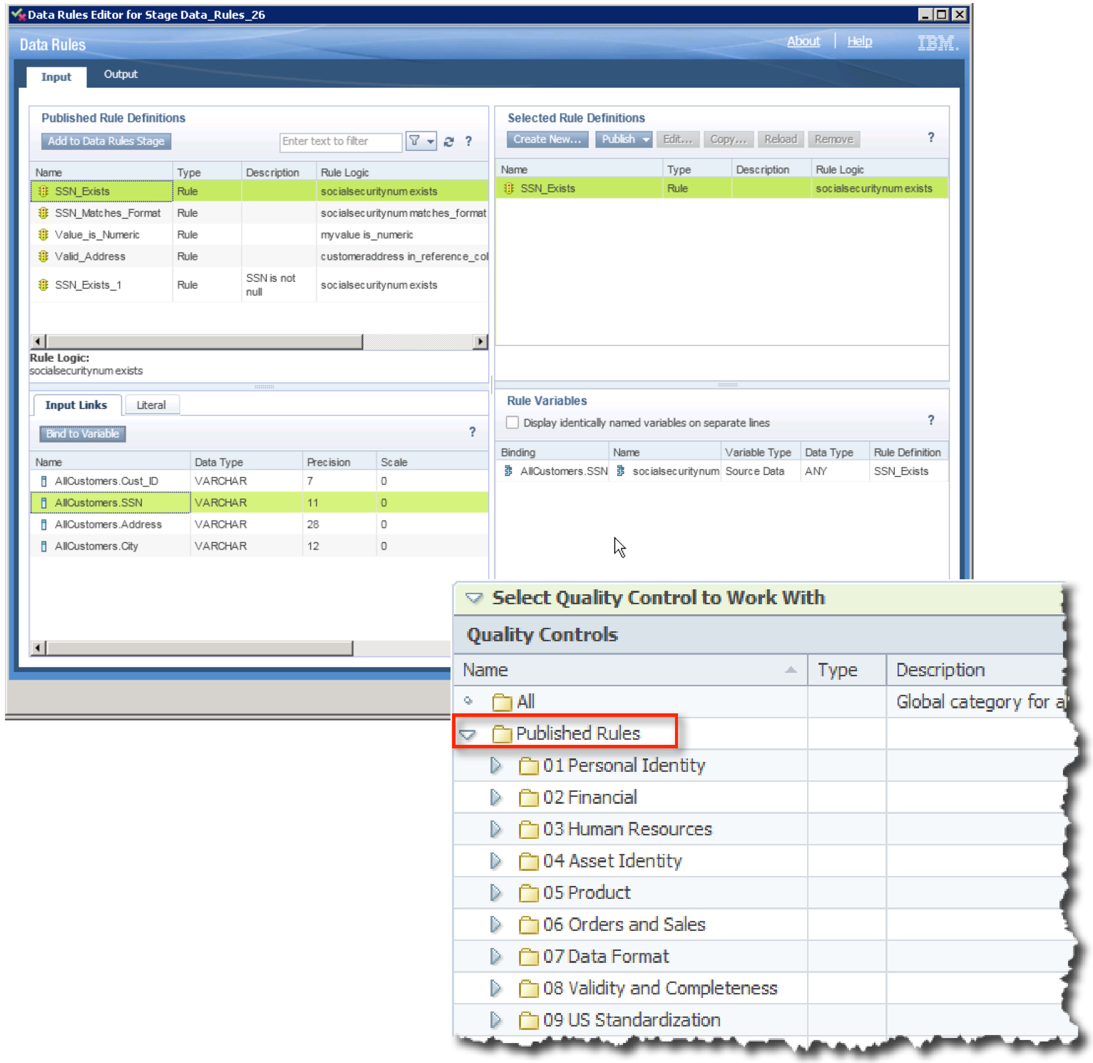
Score: 71%	Score: 73%	Score: 86%	Score: 85%	Score: 85%
Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	**Pdsaojfsadpoifj**	46	**xxxx**	M
62416	Shaun Q Dunda	**156**	904-555-2940	M
62417	Carol T Schwartz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



Data Rules Editor for Stage Data_Rules_26

Published Rule Definitions

Name	Type	Description	Rule Logic
SSN_Exists	Rule		socialsecuritynum exists
SSN_Matches_Format	Rule		socialsecuritynum matches_format
Value_is_Numeric	Rule		myvalue is_numeric
Valid_Address	Rule		customeraddress in_reference_col
SSN_Exists_1	Rule	SSN is not null	socialsecuritynum exists

Rule Logic:
socialsecuritynum exists

Name	Data Type	Precision	Scale
AllCustomers.Cust_ID	VARCHAR	7	0
AllCustomers.SSN	VARCHAR	11	0
AllCustomers.Address	VARCHAR	28	0
AllCustomers.City	VARCHAR	12	0

Selected Rule Definitions

Name	Type	Description	Rule Logic
SSN_Exists	Rule		socialsecuritynum exists

Rule Variables

□ Display identically named variables on separate lines

Binding	Name	Variable Type	Data Type	Rule Definition
AllCustomers.SSN	socialsecuritynum	Source Data	ANY	SSN_Exists

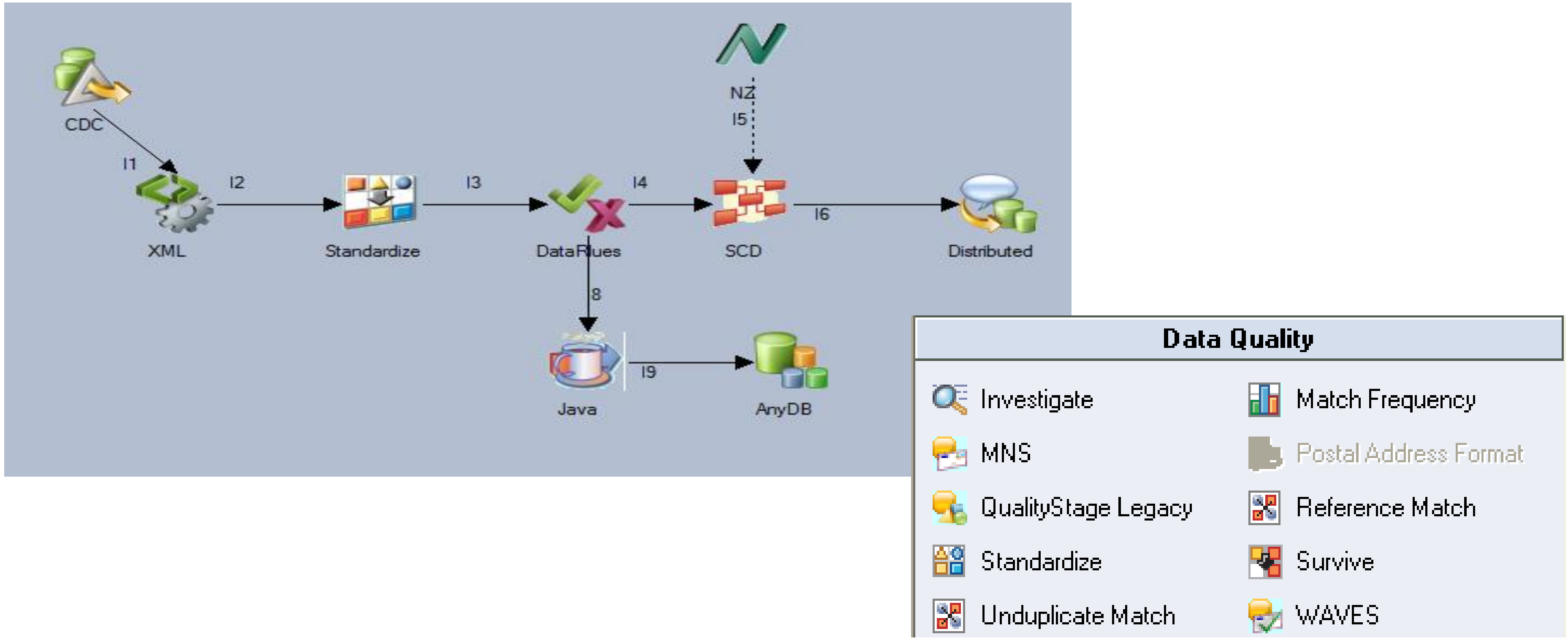
Select Quality Control to Work With

Quality Controls

Name	Type	Description
All		Global category for a
Published Rules		
01 Personal Identity		
02 Financial		
03 Human Resources		
04 Asset Identity		
05 Product		
06 Orders and Sales		
07 Data Format		
08 Validity and Completeness		
09 US Standardization		

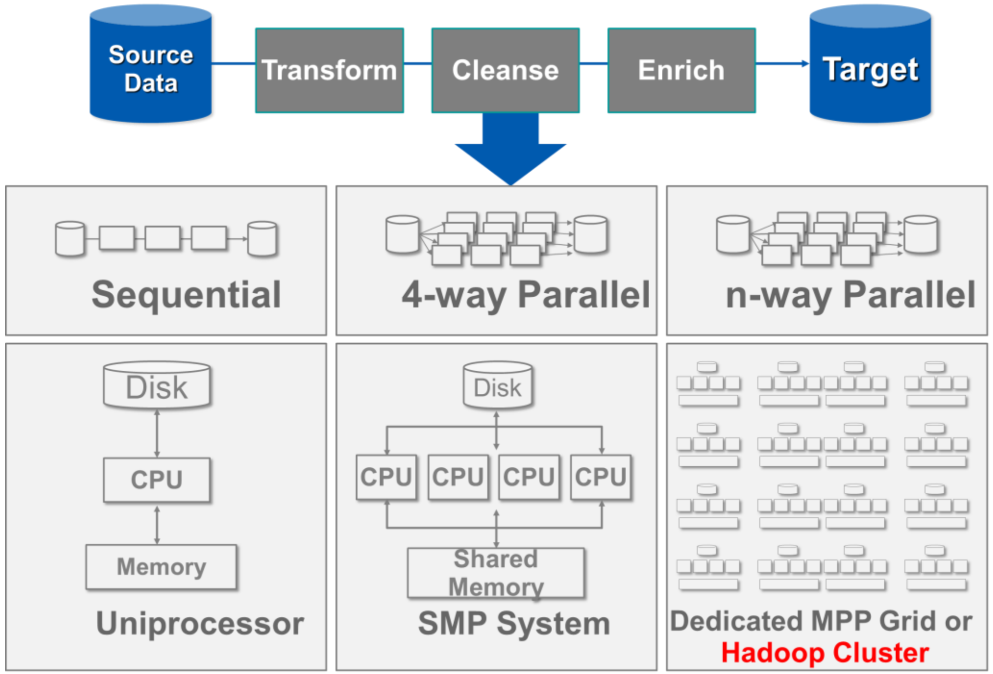
Enforce Quality on data

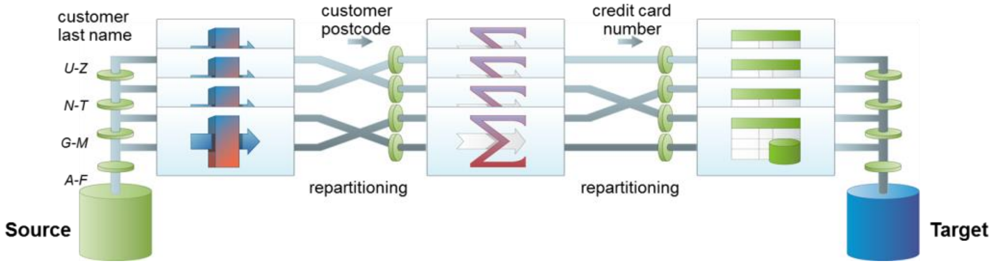
Fully integrated ETL & Data qualities capabilities



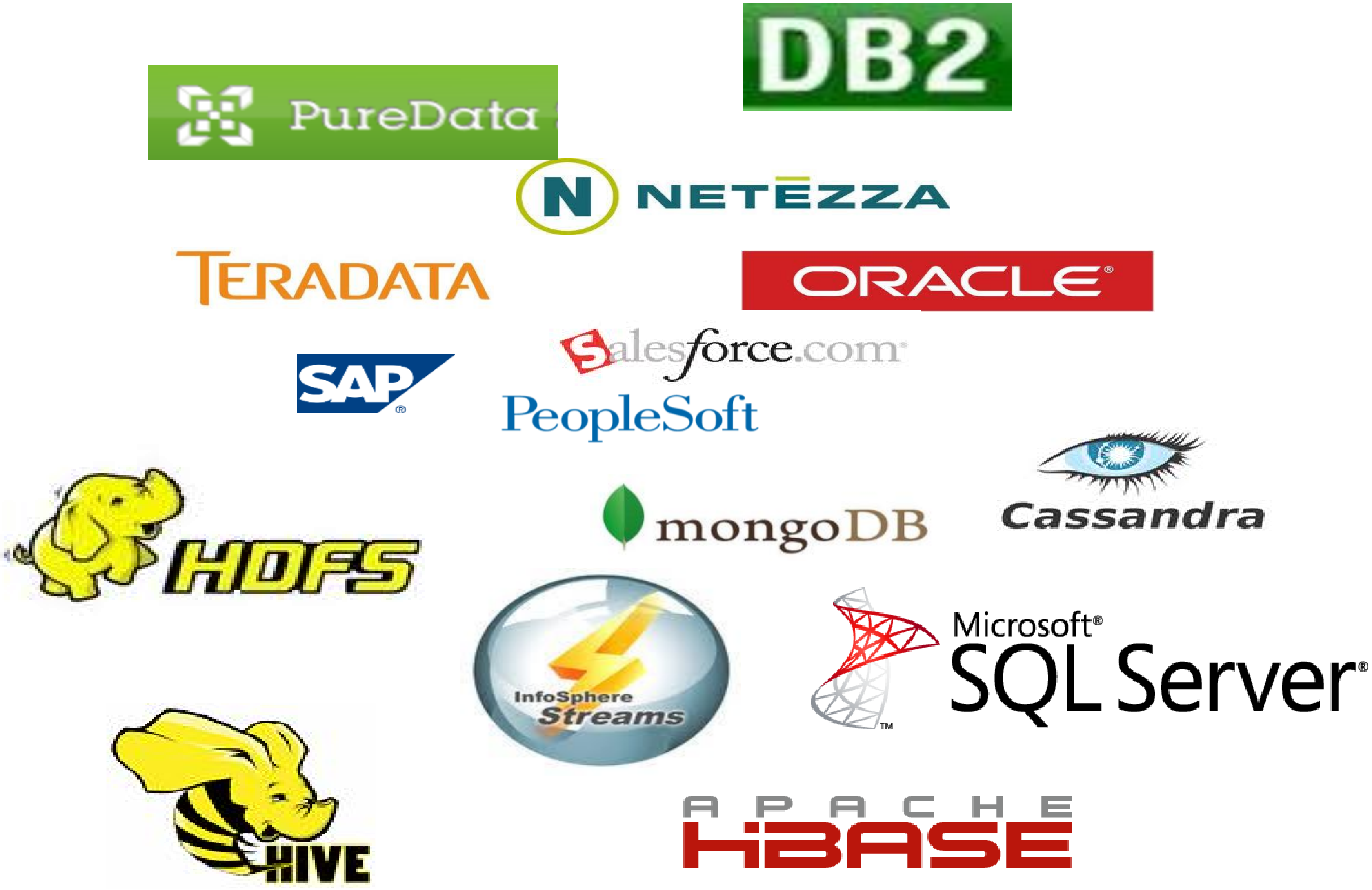
Why Information Integration is Important?

Performance





Connectivity



Predict Loan Eligibility Using SPSS in Watson Studio

Machine Learning

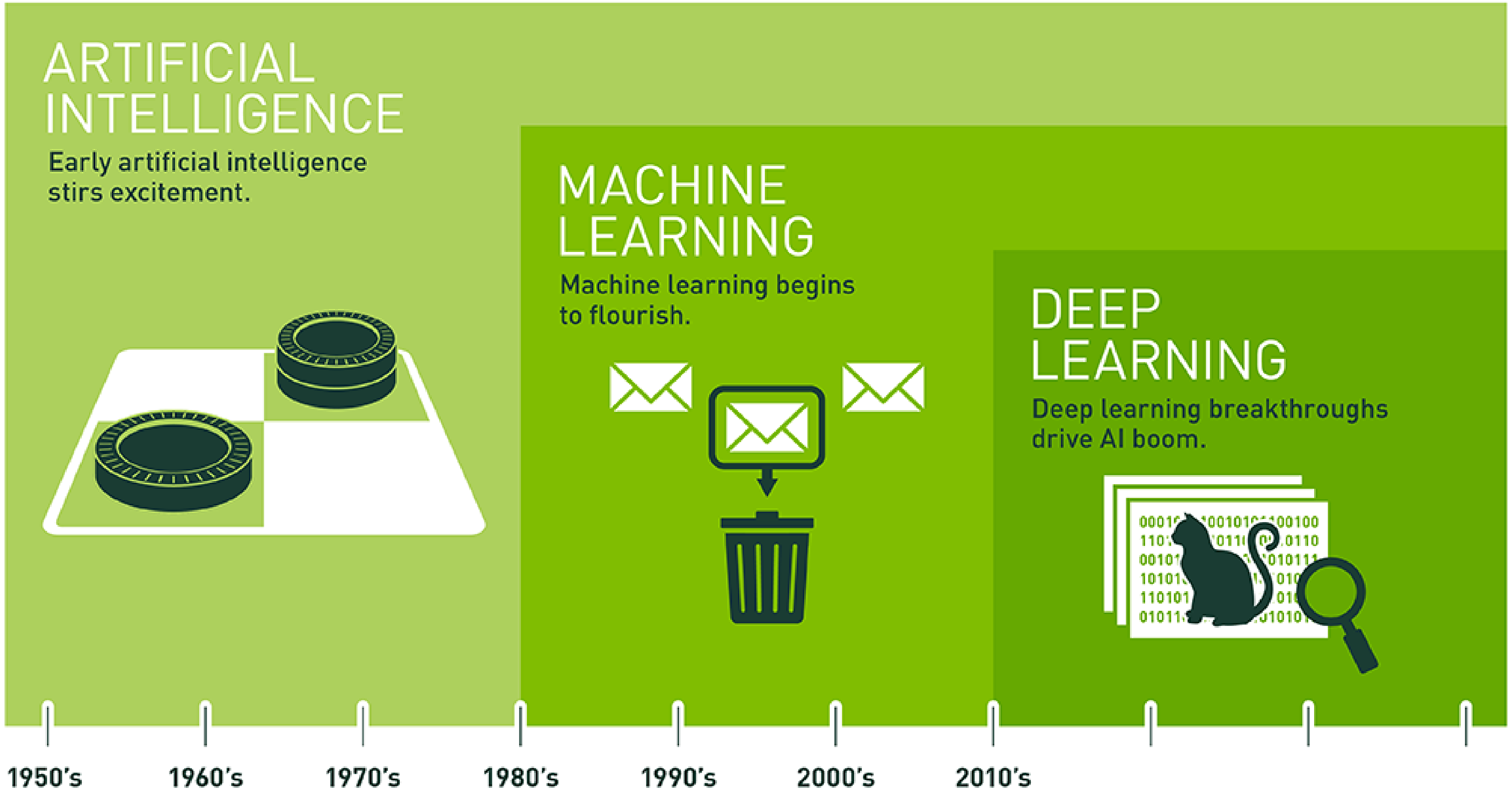
IBM Watson

Watson Studio

Loan Eligibility Predictive Model

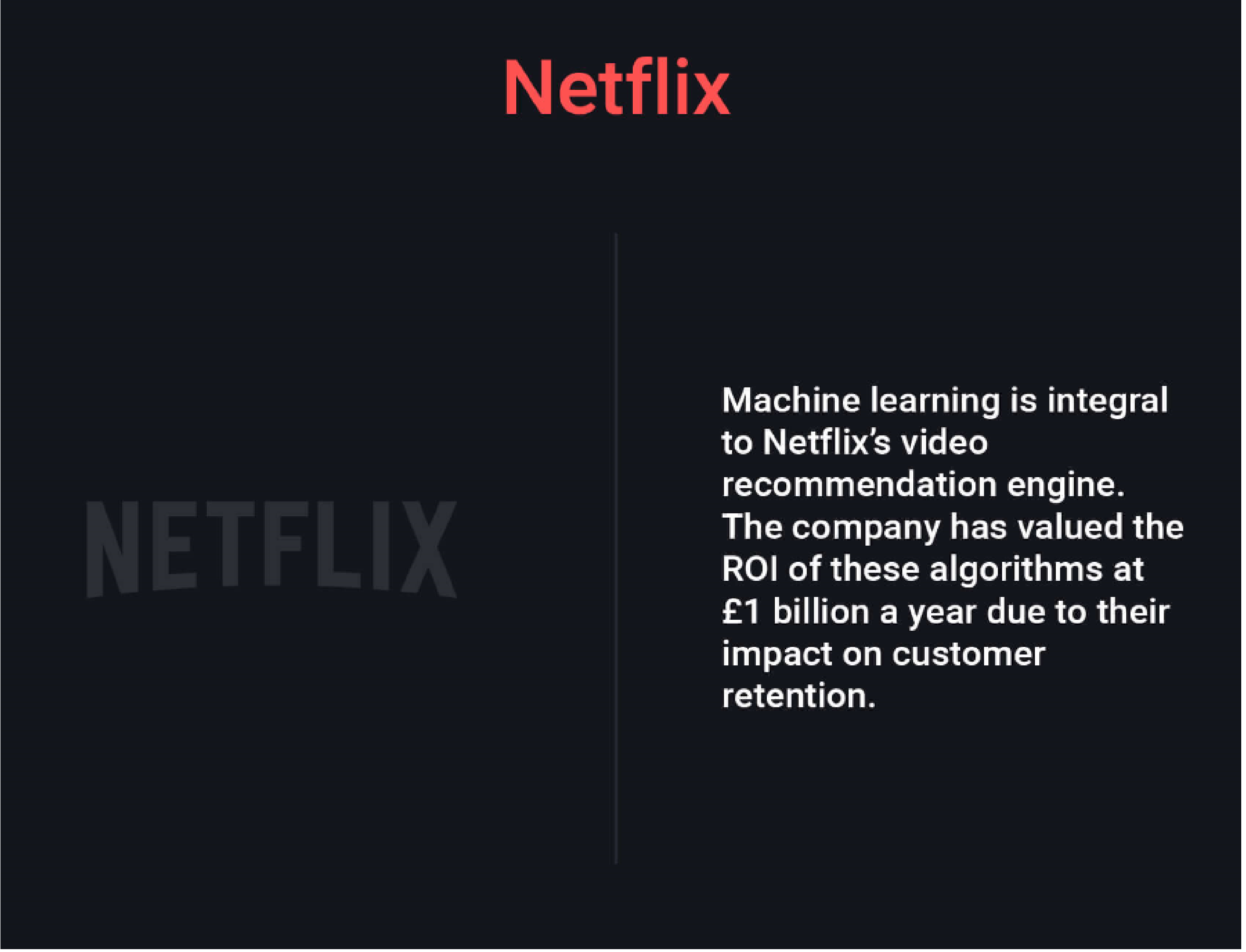
Machine Learning

Concept


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

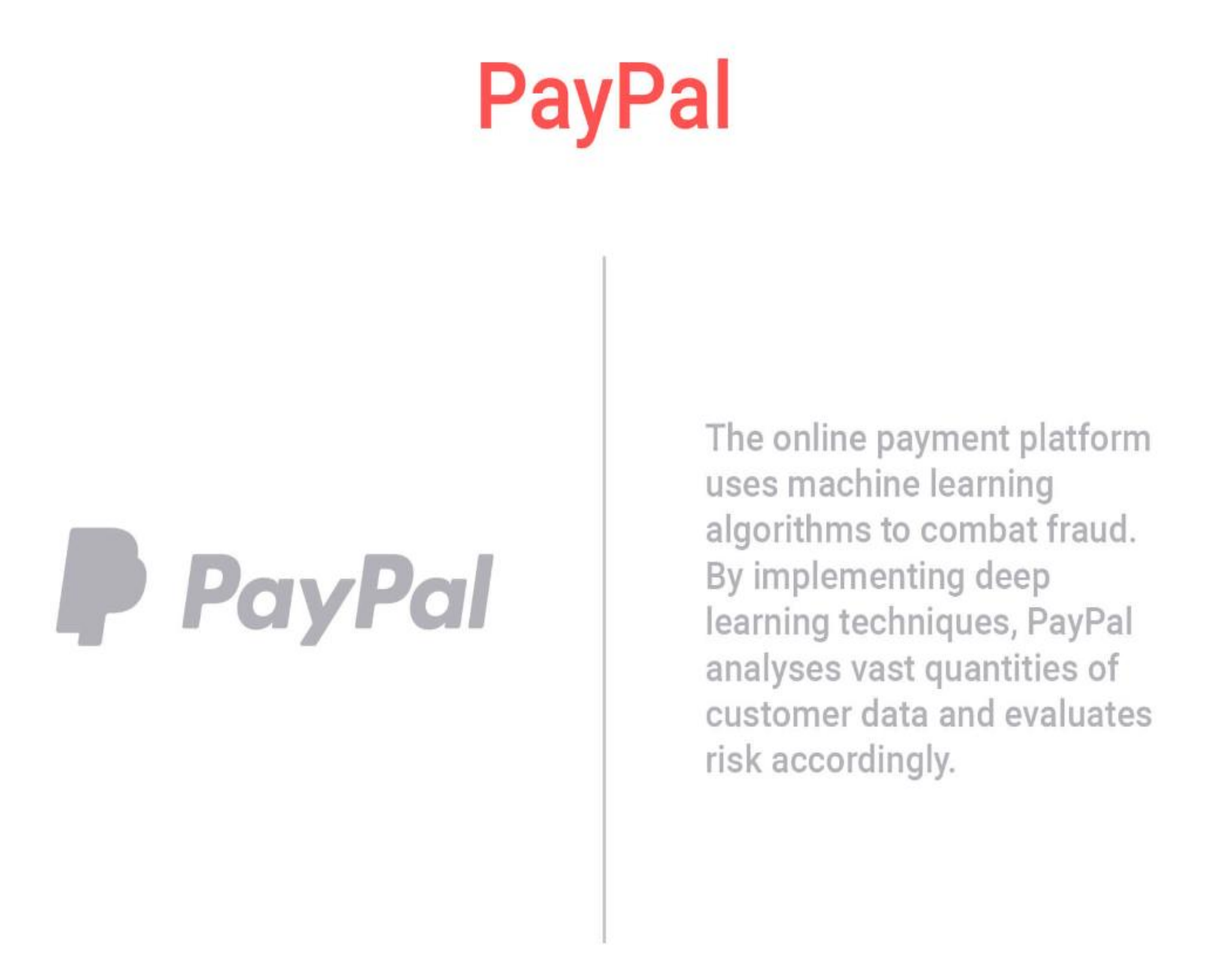
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

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.

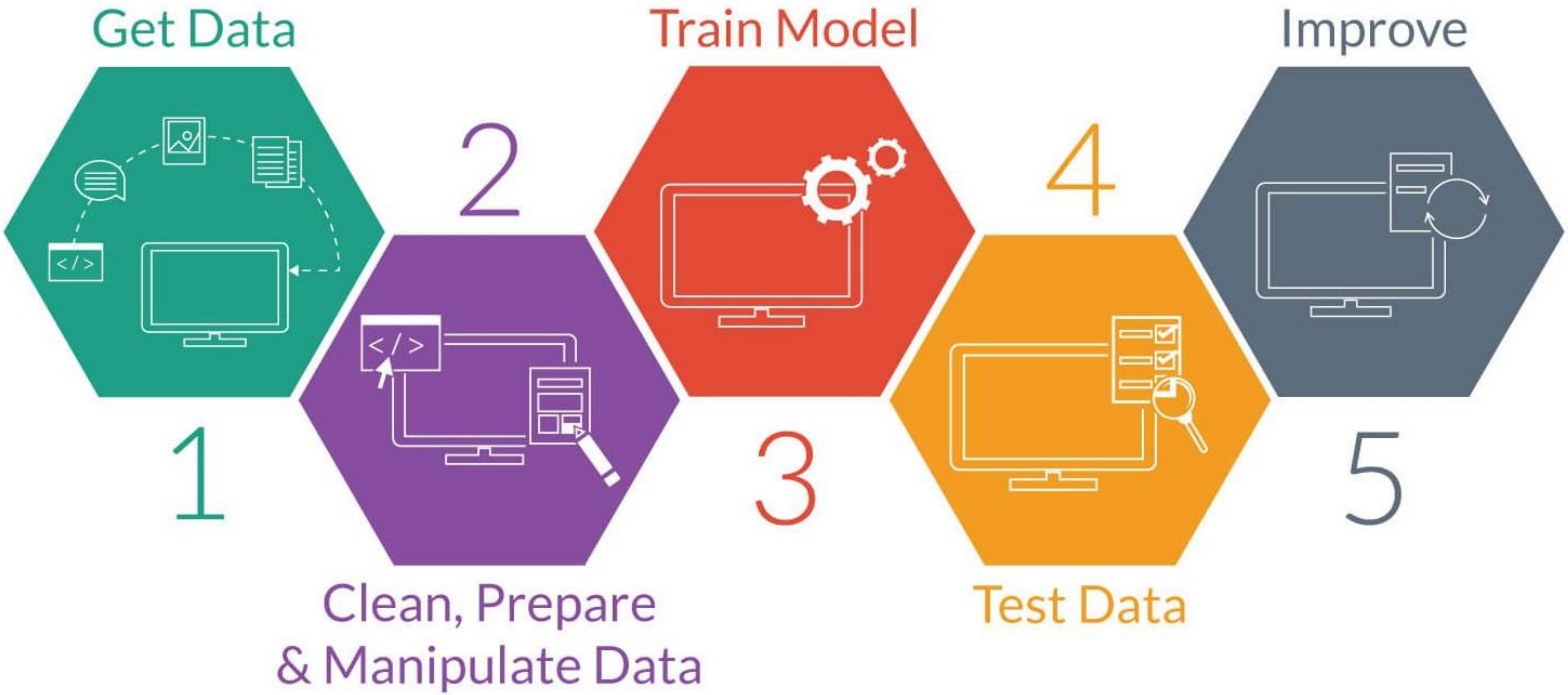


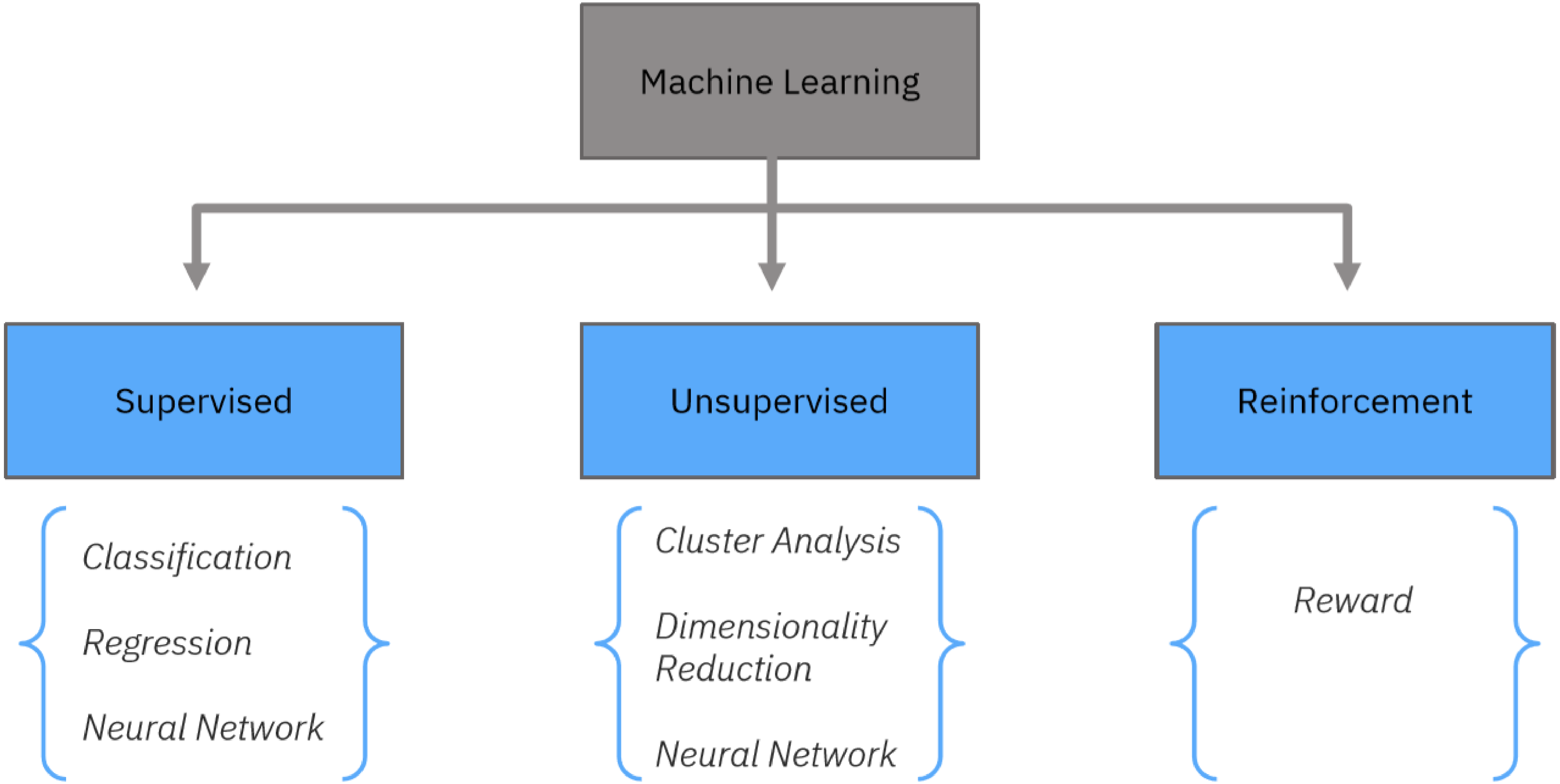
PayPal

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





Machine Learning

Supervised

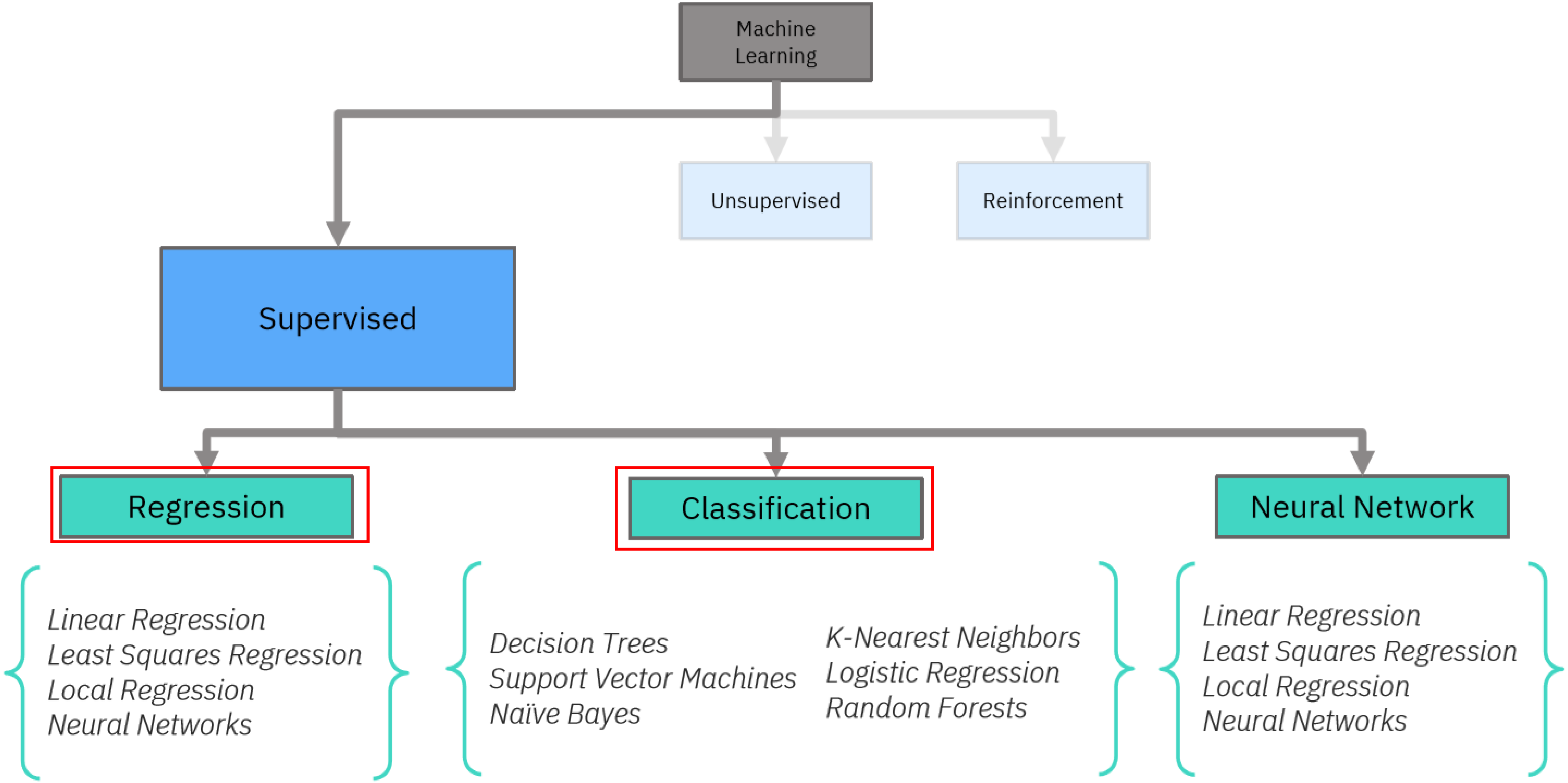
- Classification
- Regression
- Neural Network

Unsupervised

- Cluster Analysis
- Dimensionality Reduction
- Neural Network

Reinforcement

- Reward



Machine Learning

Supervised

Unsupervised

Reinforcement

Regression

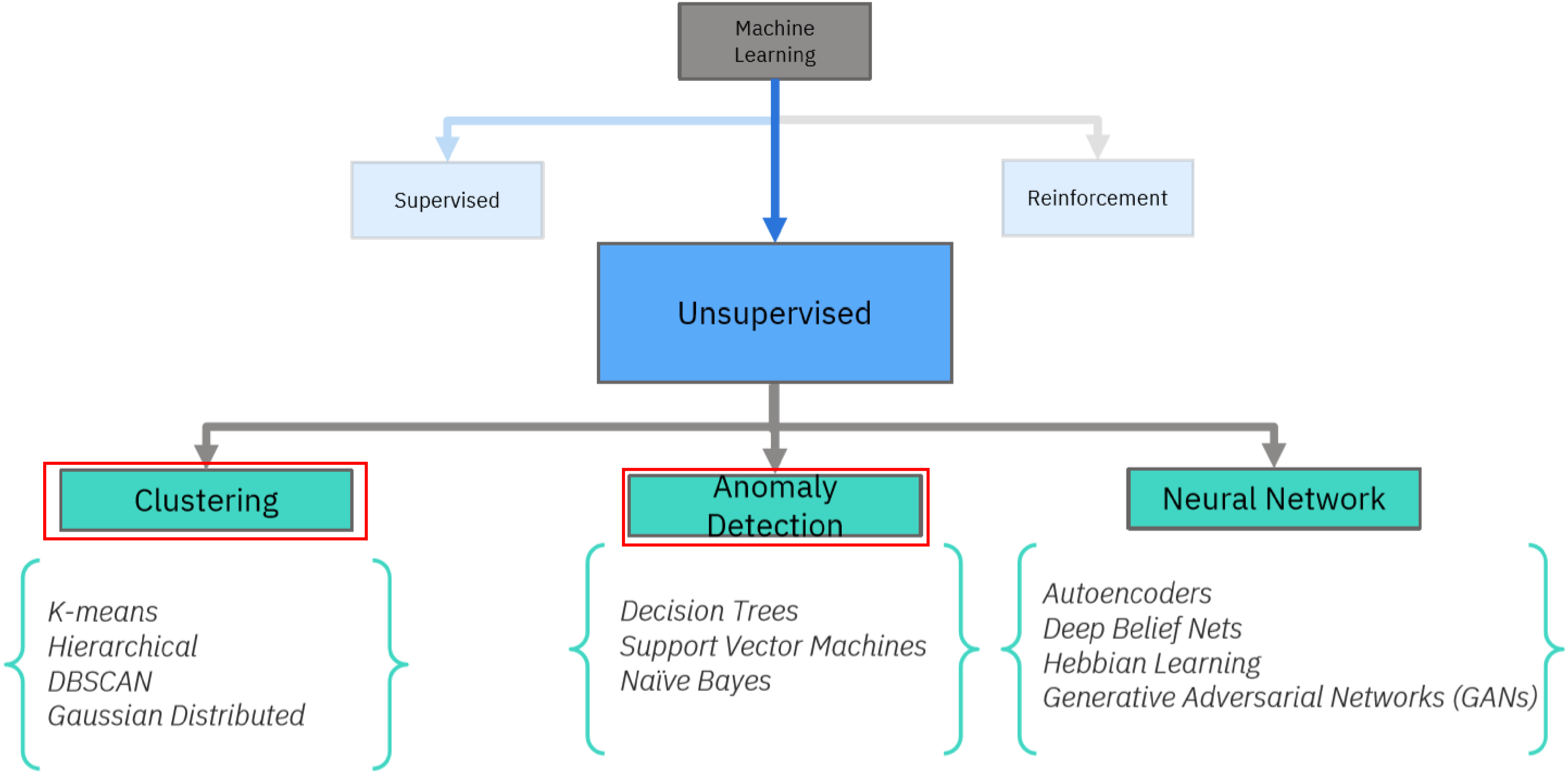
- *Linear Regression*
- *Least Squares Regression*
- *Local Regression*
- *Neural Networks*

Classification

- *Decision Trees*
- *Support Vector Machines*
- *Naïve Bayes*
- *K-Nearest Neighbors*
- *Logistic Regression*
- *Random Forests*

Neural Network

- *Linear Regression*
- *Least Squares Regression*
- *Local Regression*
- *Neural Networks*



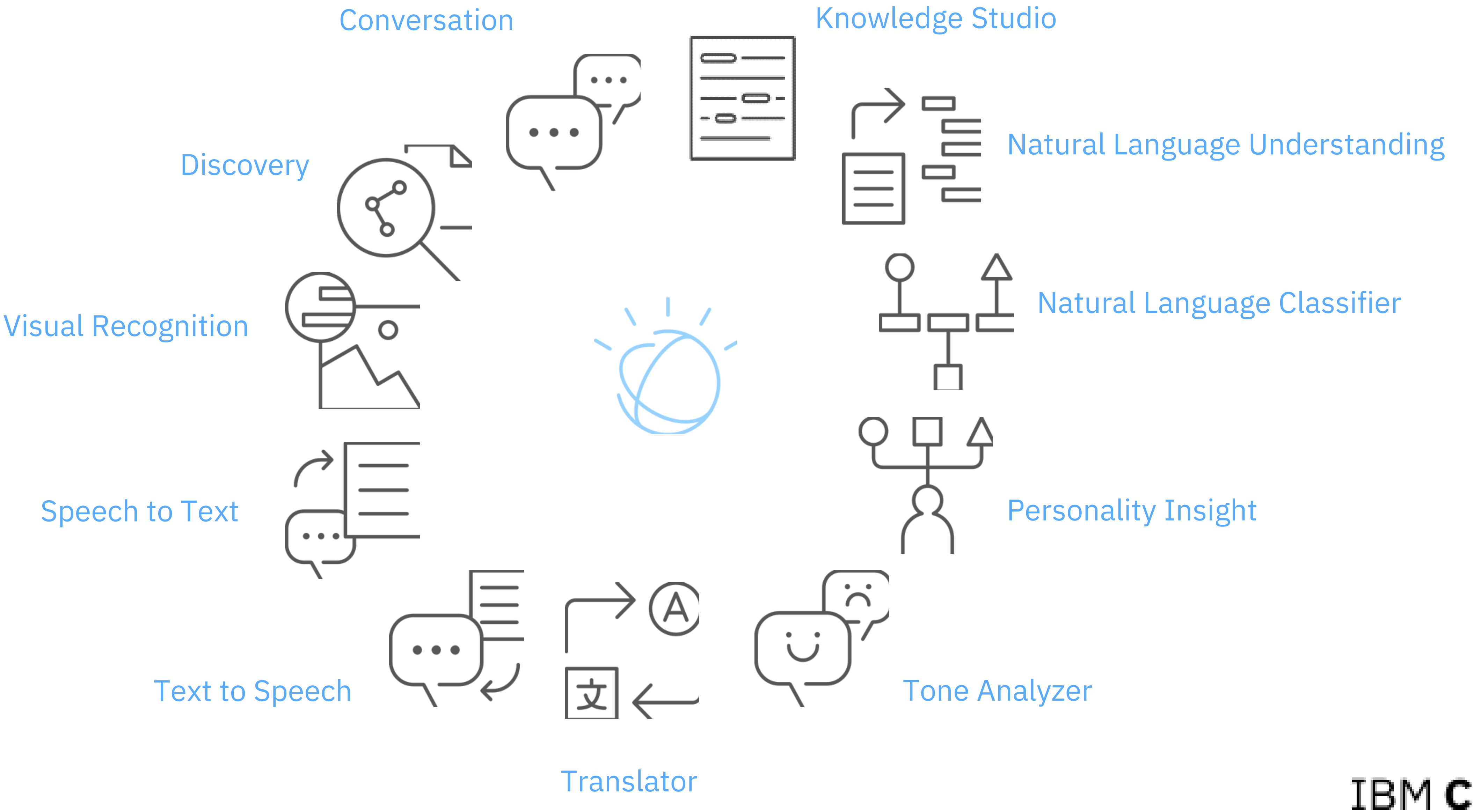
K-means
Hierarchical
DBSCAN
Gaussian Distributed

Decision Trees
Support Vector Machines
Naïve Bayes

Autoencoders
Deep Belief Nets
Hebbian Learning
Generative Adversarial Networks (GANs)

Watson is AI for Business

With Watson:



Conversation

Knowledge Studio

Natural Language Understanding

Natural Language Classifier

Personality Insight

Tone Analyzer

Translator

Text to Speech

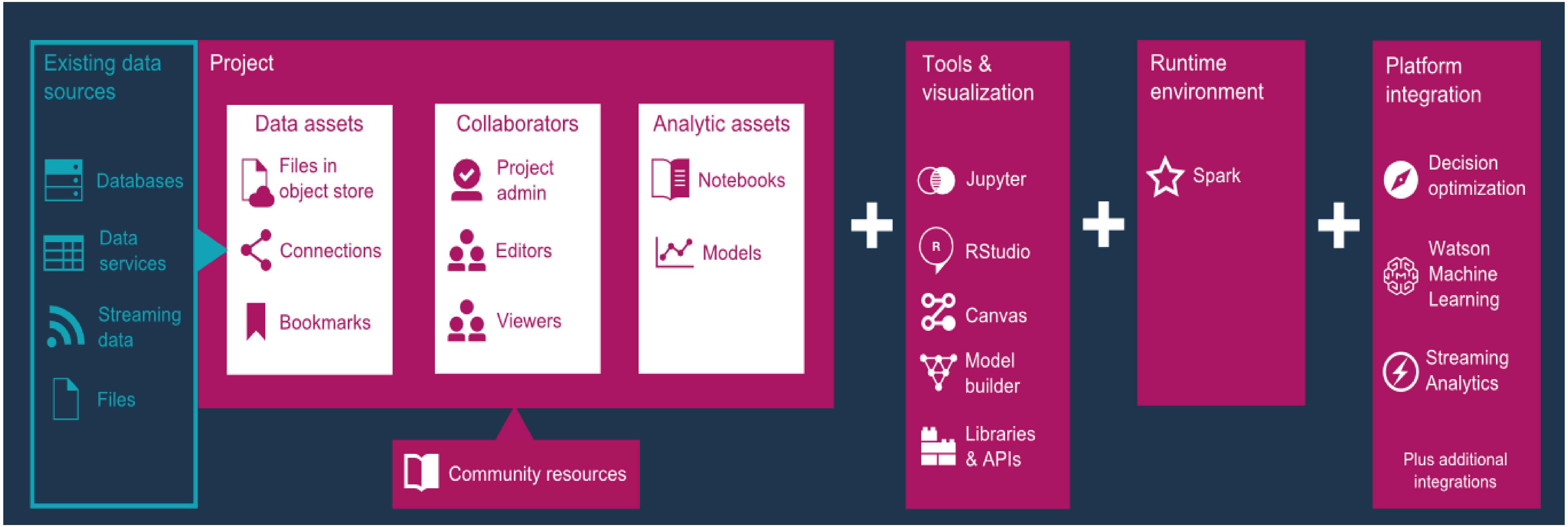
Speech to Text

Visual Recognition

Discovery

Watson Studio

Watson Studio


Existing data sources

Databases

Data services

Streaming data

Files

Project

Data assets

Files in object store

Connections

Bookmarks

Collaborators

Project admin

Editors

Viewers

Analytic assets

Notebooks

Models

Community resources

+

Tools & visualization

Jupyter

RStudio

Canvas

Model builder

Libraries & APIs

+

Runtime environment

Spark

+

Platform integration

Decision optimization

Watson Machine Learning

Streaming Analytics

Plus additional integrations

IBM Code

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

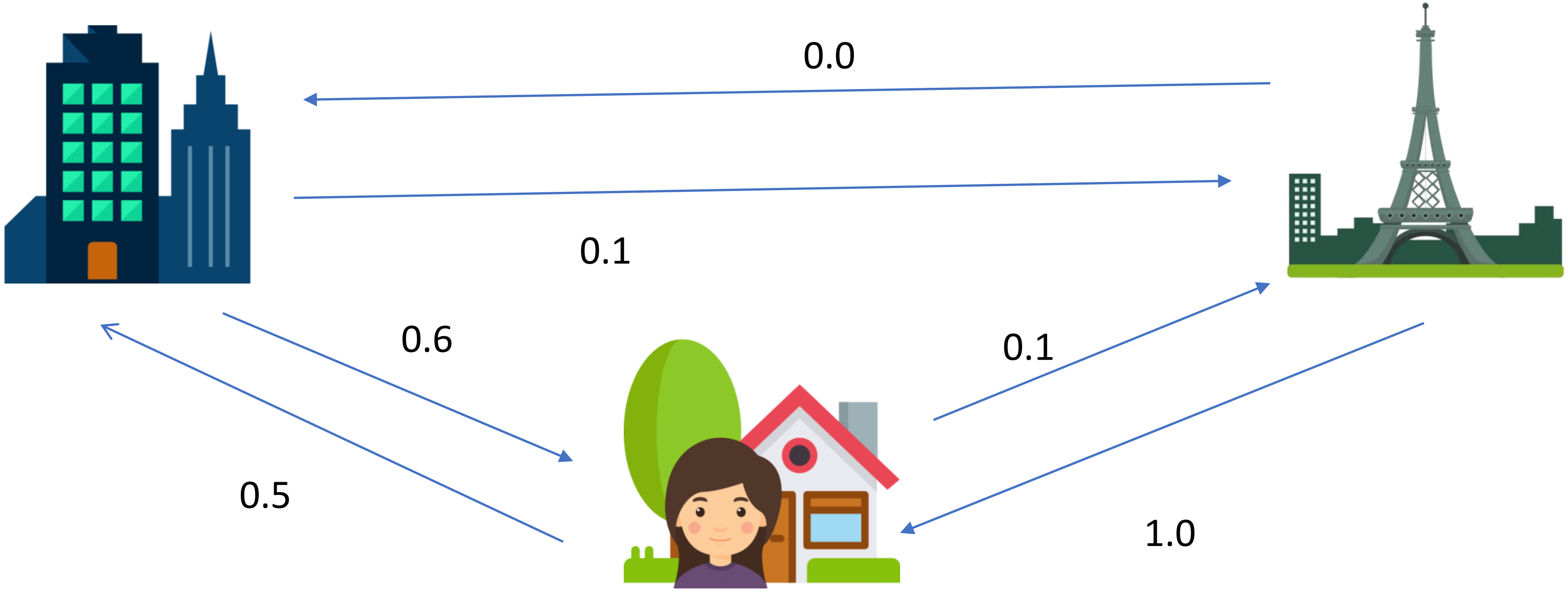
Features

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

Class

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

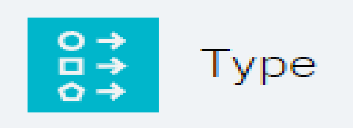
Bayes Net



Steps to Solution…

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

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



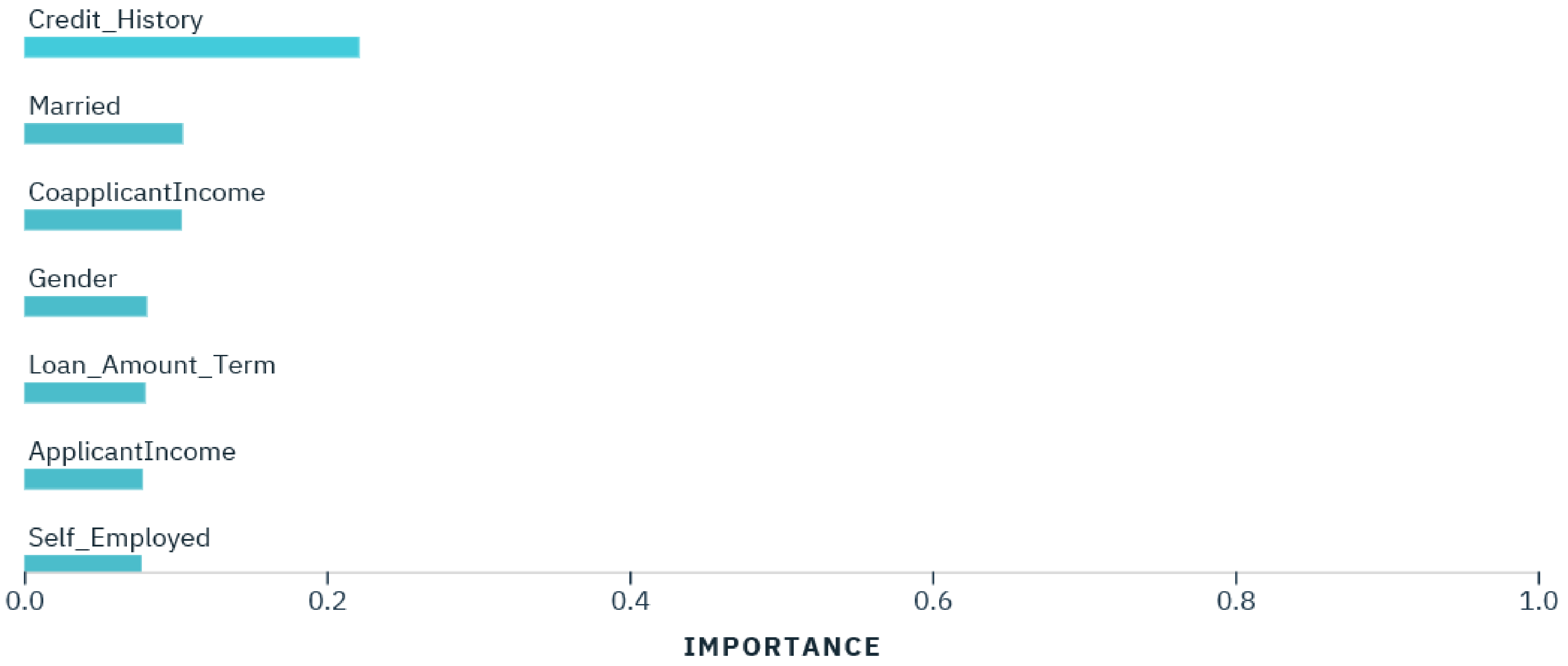
3. Split our data for training and testing sets using **Partition** node.

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



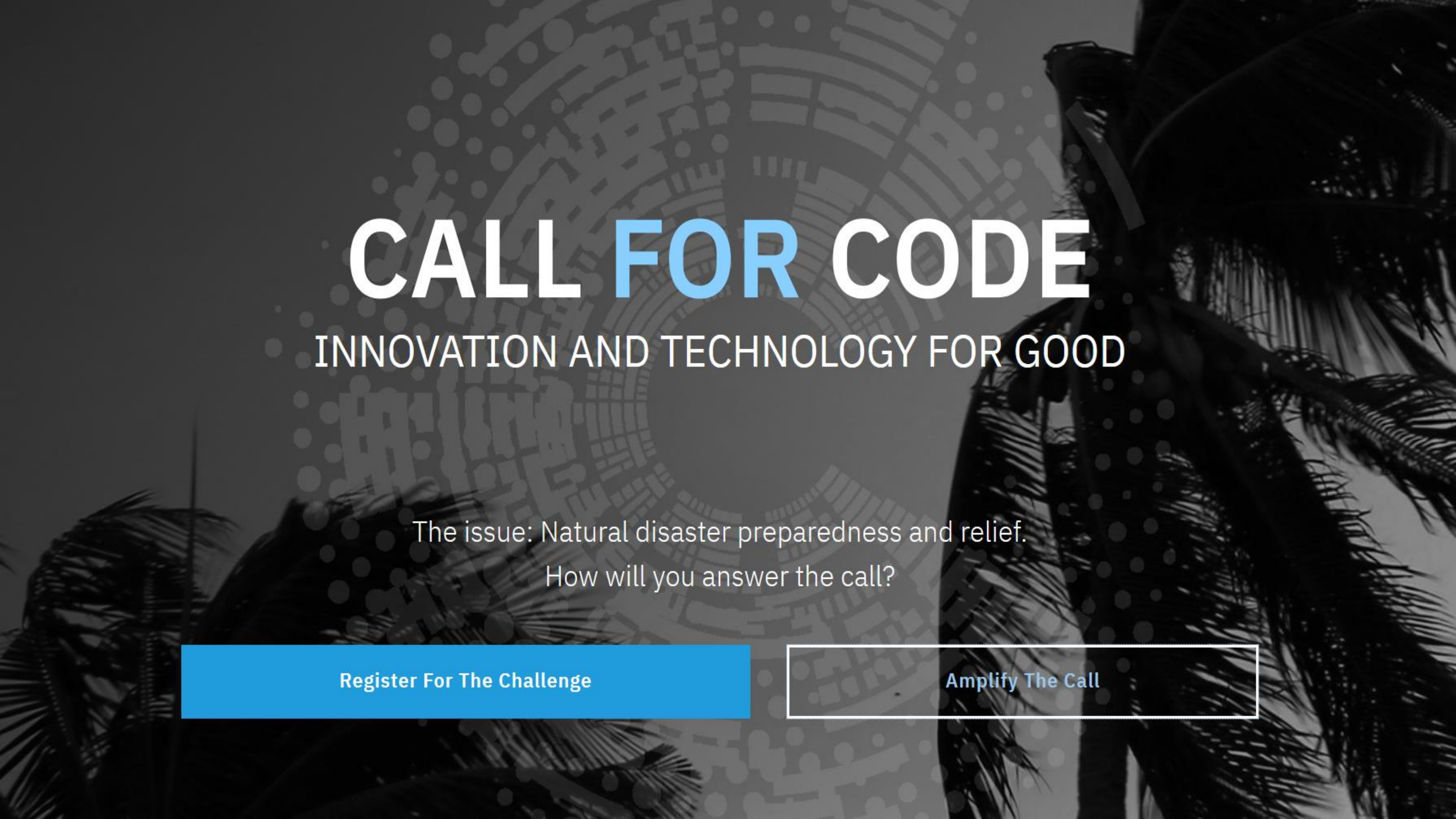
IBM Code

Please, Sign Up for IBM Cloud (Region)

https://ibm.biz/BdYmuL

GitHub

https://github.com/DevExCodeHub/Loan_eligibility_lab



CALL FOR CODE

INNOVATION AND TECHNOLOGY FOR GOOD

The issue: Natural disaster preparedness and relief.

How will you answer the call?

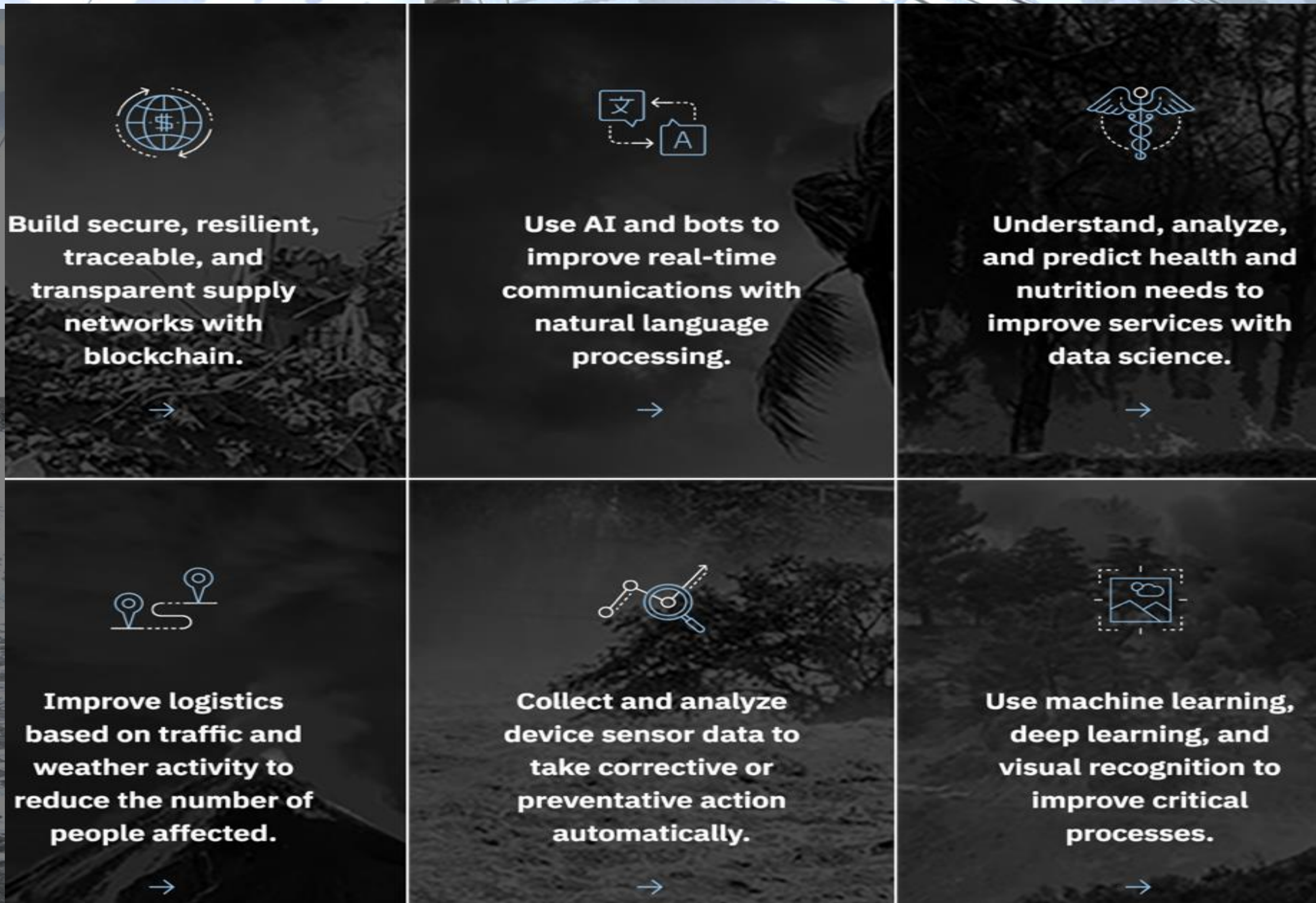
Get Started

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/



Build secure, resilient, traceable, and transparent supply networks with blockchain.
→

Use AI and bots to improve real-time communications with natural language processing.
→

Understand, analyze, and predict health and nutrition needs to improve services with data science.
→

Improve logistics based on traffic and weather activity to reduce the number of people affected.
→

Collect and analyze device sensor data to take corrective or preventative action automatically.
→

Use machine learning, deep learning, and visual recognition to improve critical processes.
→

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