

# Data Science Workshop

## An Experiential Journey with Data to Inspire *Your* Work

Maureen Norton

Neeraj Madan

Mwai Kalengamaliro

July 27<sup>th</sup>, 2022

# Data Science Workshop

## An Experiential Journey with Data to Inspire *Your* Work

### Data Science Workshop

02:00 – 04:30 US ET

#### Data Science Workshop: An Experiential Journey with Data to Inspire Your Work

The Data Science Virtual Workshop, "An Experiential Journey with Data to Inspire Your Work", will make you think differently about data and how it can solve problems! This workshop includes surprising use cases that will make you think differently about data, sometimes laugh and hopefully inspire your own work to discover actionable insights in the mounds of data available. The use cases and introductory material will be followed by a hands-on experiential journey addressing a common challenge across industries – how to improve the customer experience. The most valuable part of this workshop is that it is designed to help you gain experience and relate it to your work – so at the end you have a plan of action on how you can make data more useful in your organization to solve a key challenge.

"Improving Customer Experiences with Real-Time Insights", will be used as an example during the workshop. This experiential session will include a step by step journey based on how data science is helping companies to predict the customer experience journey and proactively address the issues, leading to the improvement of Net Promoter Score. The session will also highlight the importance of using AI Canvas, CRISP-DM (Cross Industry Standard Process for Data Mining) and Agile in Data Science projects.

The methodology involves consuming historical NPS data; using machine learning and artificial intelligence to identify the most important features and creating an algorithm to predict the customer experience.

#### Pre-Workshop Setup Instructions

It is important to have the *setup done prior to the workshop* so that the time during the session is focused on the content and experiential journey. If you encounter difficulty during the set up, reach out to Neeraj Madan or George Stark with your questions at [ds-workshop@opengroup.org](mailto:ds-workshop@opengroup.org).

Please note that while there are many tools that could be used for building predictive analytics and machine learning solutions, this workshop will be based on IBM Watson Studio for the experiential session.

**Facilitators:** Maureen Norton, Neeraj Madan, and Mwai Kalengamaliro

# Agenda

Section	Time (US PT)
<b>Getting Started</b> <i>a. Session Introduction and Expectation Setting</i> <i>b. Data Science Introduction</i> <i>c. Predictive Analytics and Machine Learning Solutions</i> <i>d. Setup Environment and Create a Project</i> <<Break (10 mins)>>	<b>02:00 pm – 02:50 pm</b>
<b>Hands on Experiential Journey (Net Promoter Score Example)</b> <i>a. Business understanding: Exercise 1: Identify an opportunity in your business context and document</i> <i>b. Data understanding: Exercise 2: What data set would you gather to work the problem statement</i> <i>c. Data preparation: Exercise 3: How would you prepare the dataset and what challenges do you foresee?</i> <<Break (10 mins)>> <i>d. Modeling: Exercise 4: What modeling techniques would you attempt and why?</i> <i>e. Evaluation: Exercise 5: What metrics would you use to evaluate your model performance?</i> <i>f. Deployment: Exercise 6: How do you plan to consume the outputs of the model?</i>	<b>03:00 pm – 04:30 pm</b>

Let's talk about data

Is there a source of data that has information about

- ANY topic
- ANY where
- ANY time

Is there a source that has their finger on the pulse of what people think at any moment in time?



Twitter

# Soggy Fries



Let's talk about data

What other types  
of data can be  
used to drive  
deeper insights?



# WEATHER

The image features a dramatic sky scene. The bottom half is filled with large, white, fluffy clouds. Above the clouds is a bright blue horizon line. The top half of the image is a dark blue sky with faint, thin, curved lines that resemble orbits or weather patterns. The word "WEATHER" is written in large, white, bold, sans-serif capital letters across the middle of the image, partially overlapping the clouds and the blue sky.



# Four Common Data Science Models

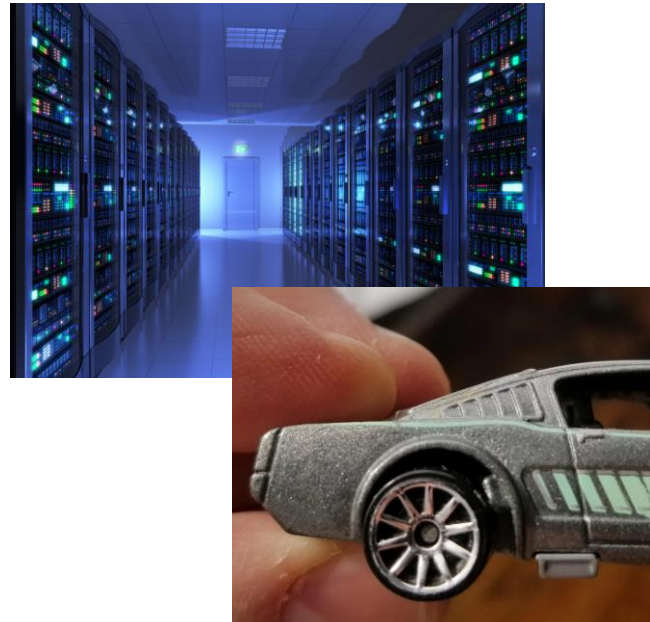
## Risk Assessment

Create a “Screening Model” to identify “threats”. Threats can be any sort of fraudulent activity (e.g., credit transaction, passenger screening, ability to purchase, altered video/photo, Fake/Real news)



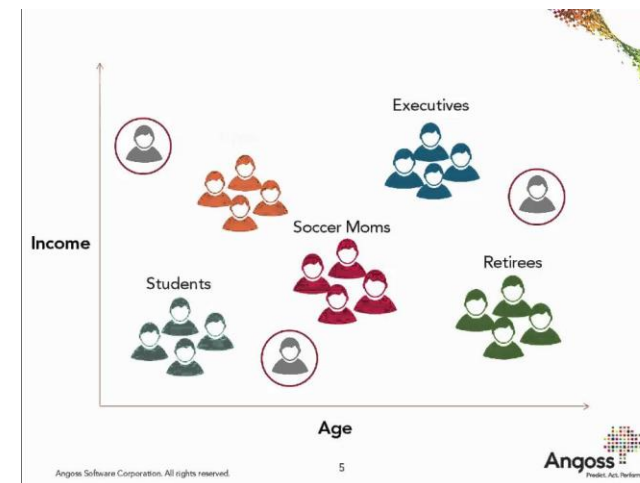
## Quality/Defect

Identify problematic components, defects in a product (e.g., server failure, castings, ATM Machines)



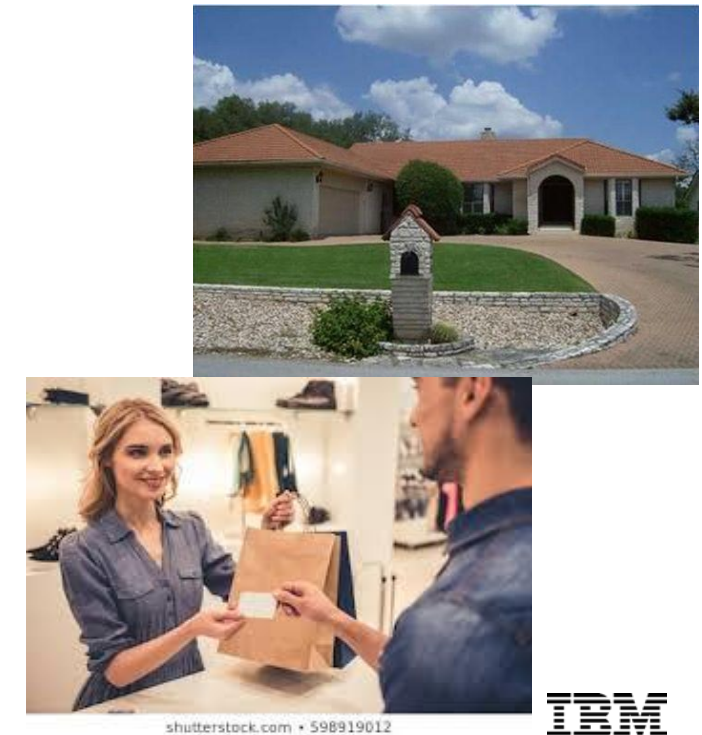
## Business Value/ Customer Satisfaction

Identifies which customers have the most potential business value based on their characteristics and activities. Which customers are likely to be happy? Which will be promoters?

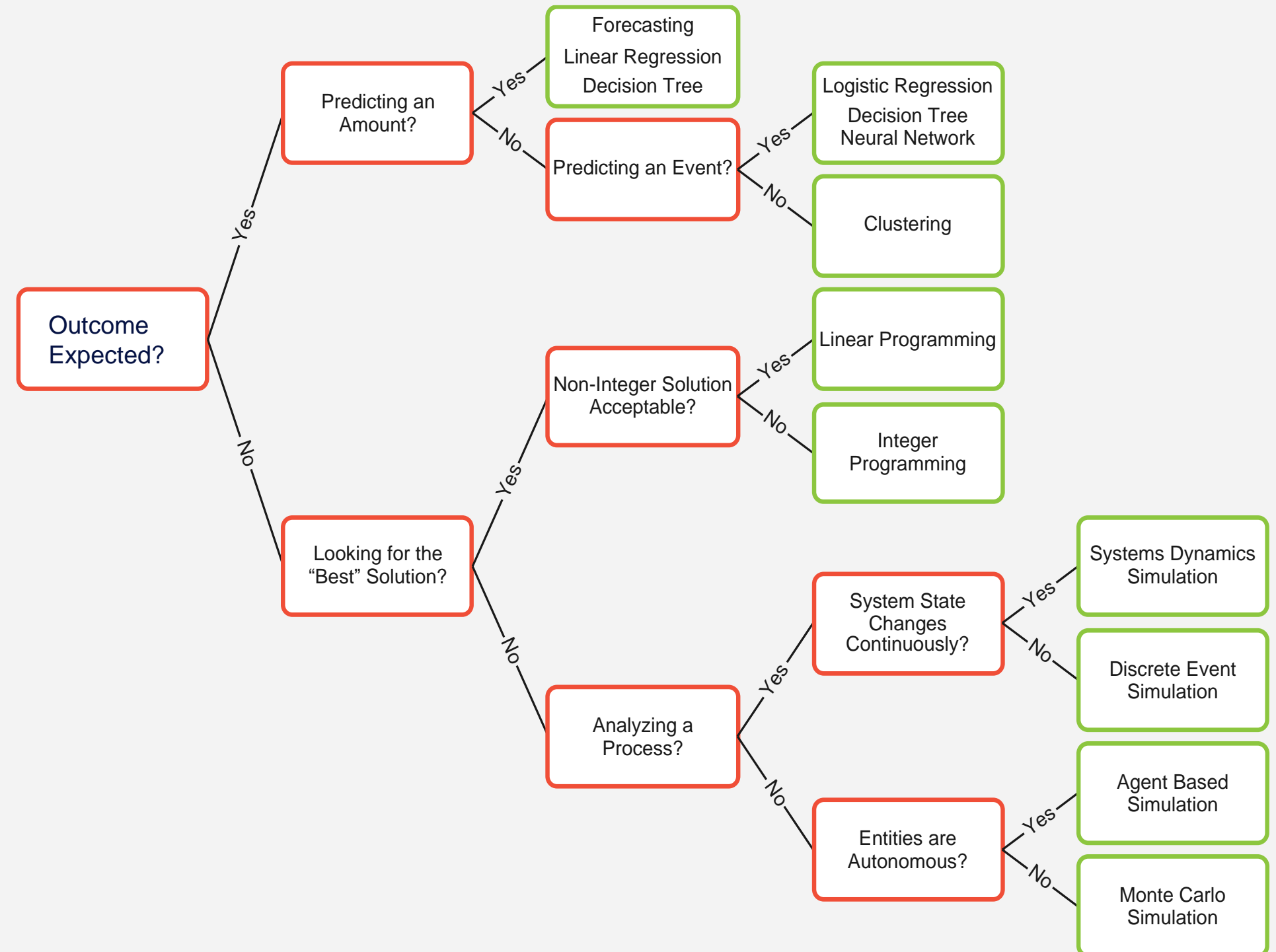


## Price/Cost/Value

Predict value (e.g., home/ rental prices, value of retail transaction, number of issues, etc.)

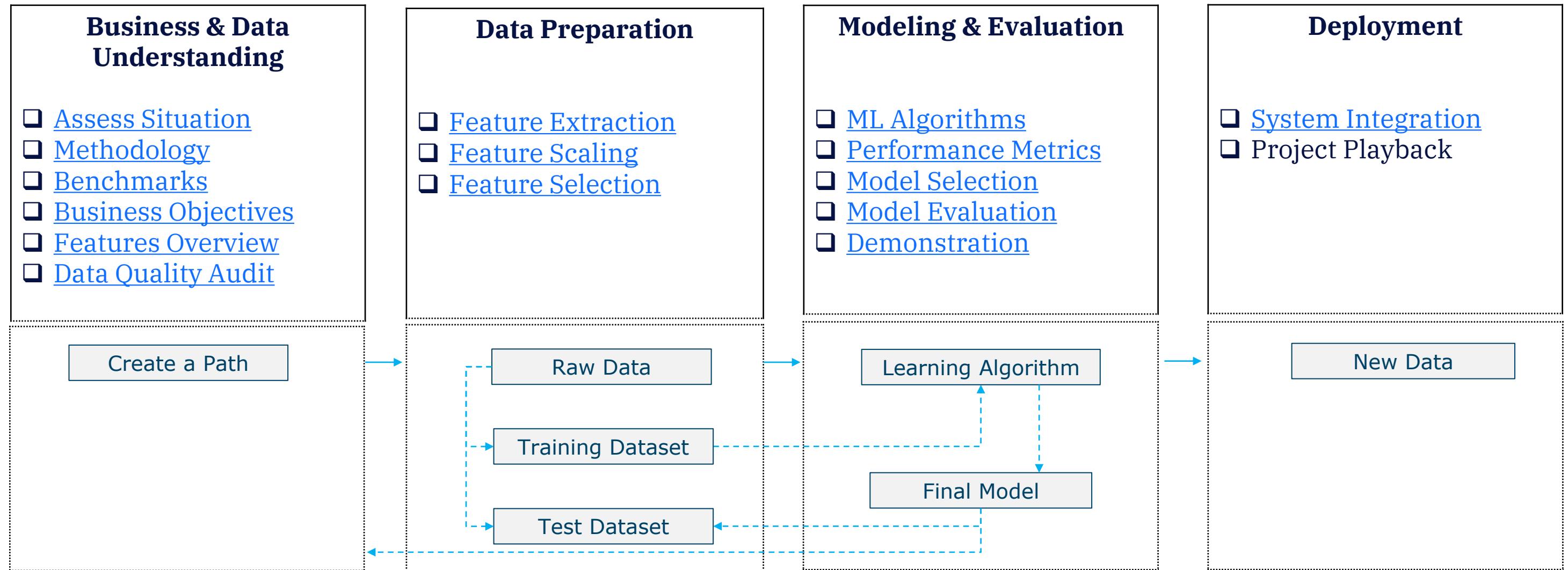


# Choosing the right analytic approach



# Framework: Roadmap to Building Machine Learning System

"Essentially, all models are wrong, but some are useful."--- Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.



# Predictive Analytics and Machine Learning Solutions

*While there are many tools that could be used for building predictive analytics and machine learning solutions (see below for examples), this workshop will be based on IBM Watson Studio for the experiential learning session.*

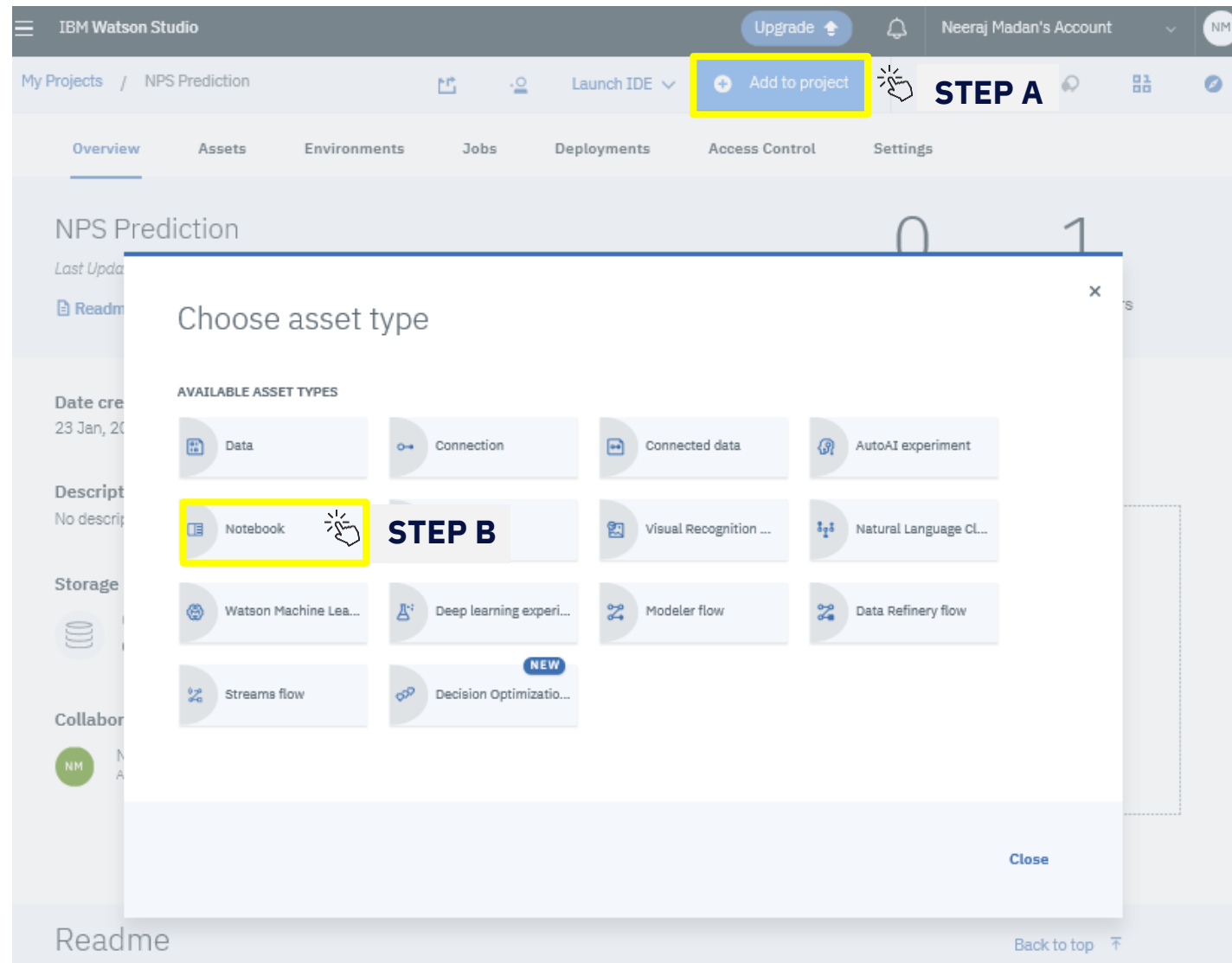
To name a few:

1. IBM Watson Studio
2. SAS Advanced Analytics
3. RapidMiner
4. Amazon SageMaker
5. Azure Machine Learning Studio (Microsoft)
6. Google Cloud AI Platform

# Audience Poll

# Create a Project

## Add a notebook





# Create a Project

## Import a notebook from GitHub

The screenshot shows the 'New notebook' form in IBM Watson Studio. The form is divided into two main sections: 'New notebook' and 'Select runtime'. The 'New notebook' section has three tabs: 'Blank', 'From file', and 'From URL'. The 'From URL' tab is selected and highlighted with a yellow box, with a lightbulb icon and 'STEP A' next to it. Below the tabs, there is a 'Name' field with a yellow box around it containing the text '(Please type "NPS Prediction")', a lightbulb icon, and 'STEP B'. Below the 'Name' field is a 'Description (optional)' field with a placeholder 'Type your Description here' and a '500 characters remaining' indicator. The 'Select runtime' section has a 'Select runtime' label and a yellow box around the text 'Default Python 3.8 XS (2 vCPU 8 GB RAM)', with a lightbulb icon and 'STEP C' next to it. Below this, there is a 'Notebook URL' label and a yellow box around the text 'Use the link highlighted below', with a lightbulb icon and 'STEP D (Enter Notebook URL Link)' next to it. The top of the page shows the IBM Watson Studio header with a menu icon, the text 'IBM Watson Studio', an 'Upgrade' button, a notification bell, the user's name 'Neeraj Madan's Account', and a profile icon 'NM'. The breadcrumb navigation shows 'My Projects / NPS Prediction / Add Notebook'.

IBM Watson Studio

Upgrade

Neeraj Madan's Account

My Projects / NPS Prediction / Add Notebook

New notebook

Blank From file **From URL** **STEP A**

Name

(Please type "NPS Prediction") **STEP B**

26 characters remaining

Description (optional)

Type your Description here

500 characters remaining

Select runtime

Default Python 3.8 XS (2 vCPU 8 GB RAM) **STEP C**

The selected runtime has 1 vCPU and 4 GB RAM and is free.  
[Learn more](#) about capacity unit hours and Watson Studio pricing plans.

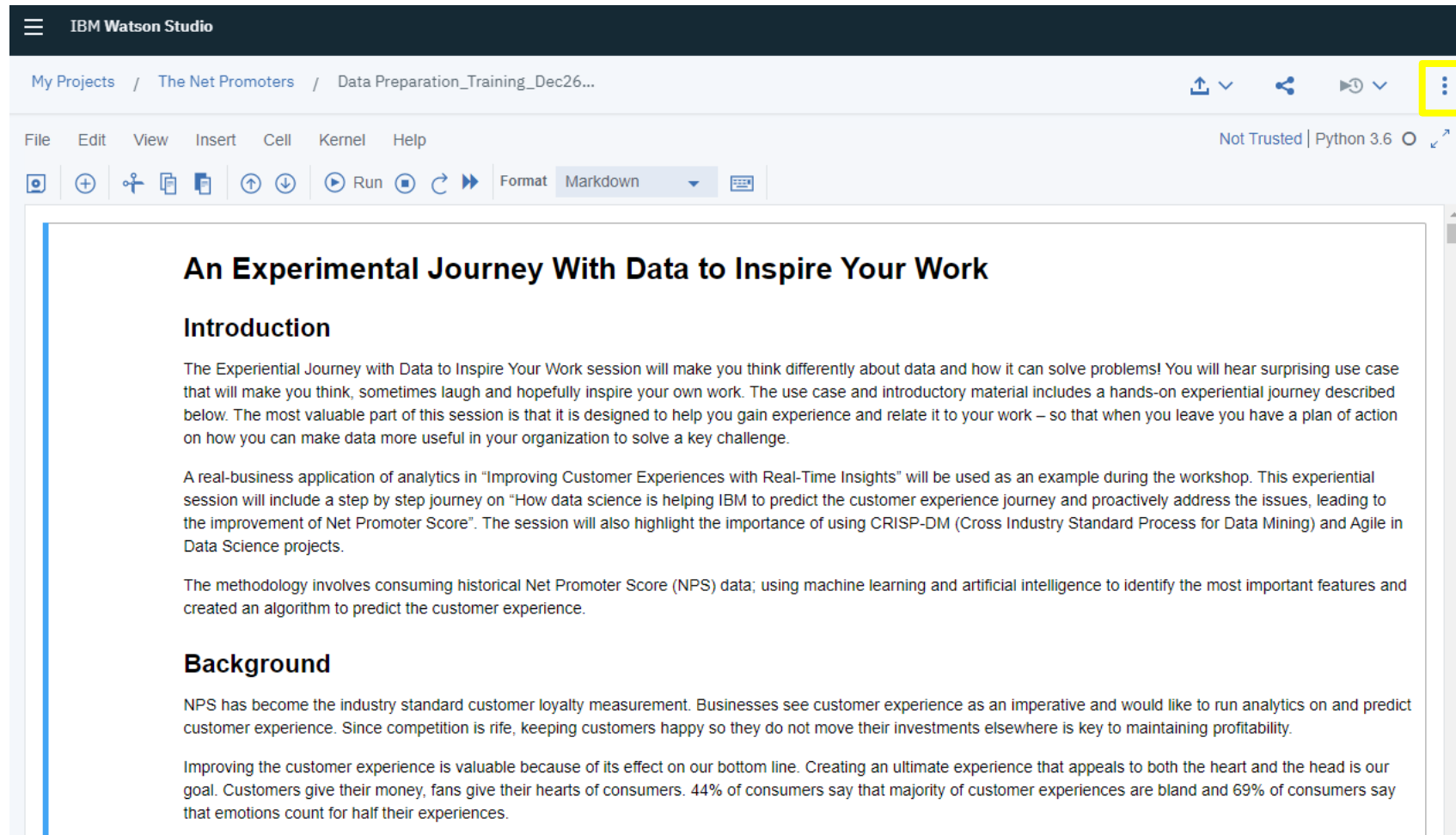
Notebook URL

Use the link highlighted below **STEP D** (Enter Notebook URL Link)

Notebook URL: <https://github.com/neemadan/An-Experiential-Journey-With-Data-to-Inspire-Your-Work/blob/master/An%20Experimental%20Journey%20With%20Data%20to%20Inspire%20Your%20Work.ipynb>

# Create a Project

## Let's get started!



The screenshot displays the IBM Watson Studio web interface. At the top, a dark blue header bar contains the IBM logo and the text 'IBM Watson Studio'. Below this, a breadcrumb trail shows the navigation path: 'My Projects' / 'The Net Promoters' / 'Data Preparation\_Training\_Dec26...'. To the right of the breadcrumb, there are icons for upload, share, and a dropdown menu, with the dropdown menu icon highlighted by a yellow square. Below the header, a menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', and 'Help'. To the right of the menu bar, the text 'Not Trusted | Python 3.6' is displayed. Below the menu bar, a toolbar contains various icons for document management, including a plus sign, a minus sign, a document icon, a run icon, and a format dropdown menu set to 'Markdown'. The main content area is a document editor with a blue vertical margin on the left. The document title is 'An Experimental Journey With Data to Inspire Your Work'. The document content includes an 'Introduction' section with two paragraphs and a 'Background' section with two paragraphs.

### An Experimental Journey With Data to Inspire Your Work

#### Introduction

The Experiential Journey with Data to Inspire Your Work session will make you think differently about data and how it can solve problems! You will hear surprising use case that will make you think, sometimes laugh and hopefully inspire your own work. The use case and introductory material includes a hands-on experiential journey described below. The most valuable part of this session is that it is designed to help you gain experience and relate it to your work – so that when you leave you have a plan of action on how you can make data more useful in your organization to solve a key challenge.

A real-business application of analytics in “Improving Customer Experiences with Real-Time Insights” will be used as an example during the workshop. This experiential session will include a step by step journey on “How data science is helping IBM to predict the customer experience journey and proactively address the issues, leading to the improvement of Net Promoter Score”. The session will also highlight the importance of using CRISP-DM (Cross Industry Standard Process for Data Mining) and Agile in Data Science projects.

The methodology involves consuming historical Net Promoter Score (NPS) data; using machine learning and artificial intelligence to identify the most important features and created an algorithm to predict the customer experience.

#### Background

NPS has become the industry standard customer loyalty measurement. Businesses see customer experience as an imperative and would like to run analytics on and predict customer experience. Since competition is rife, keeping customers happy so they do not move their investments elsewhere is key to maintaining profitability.

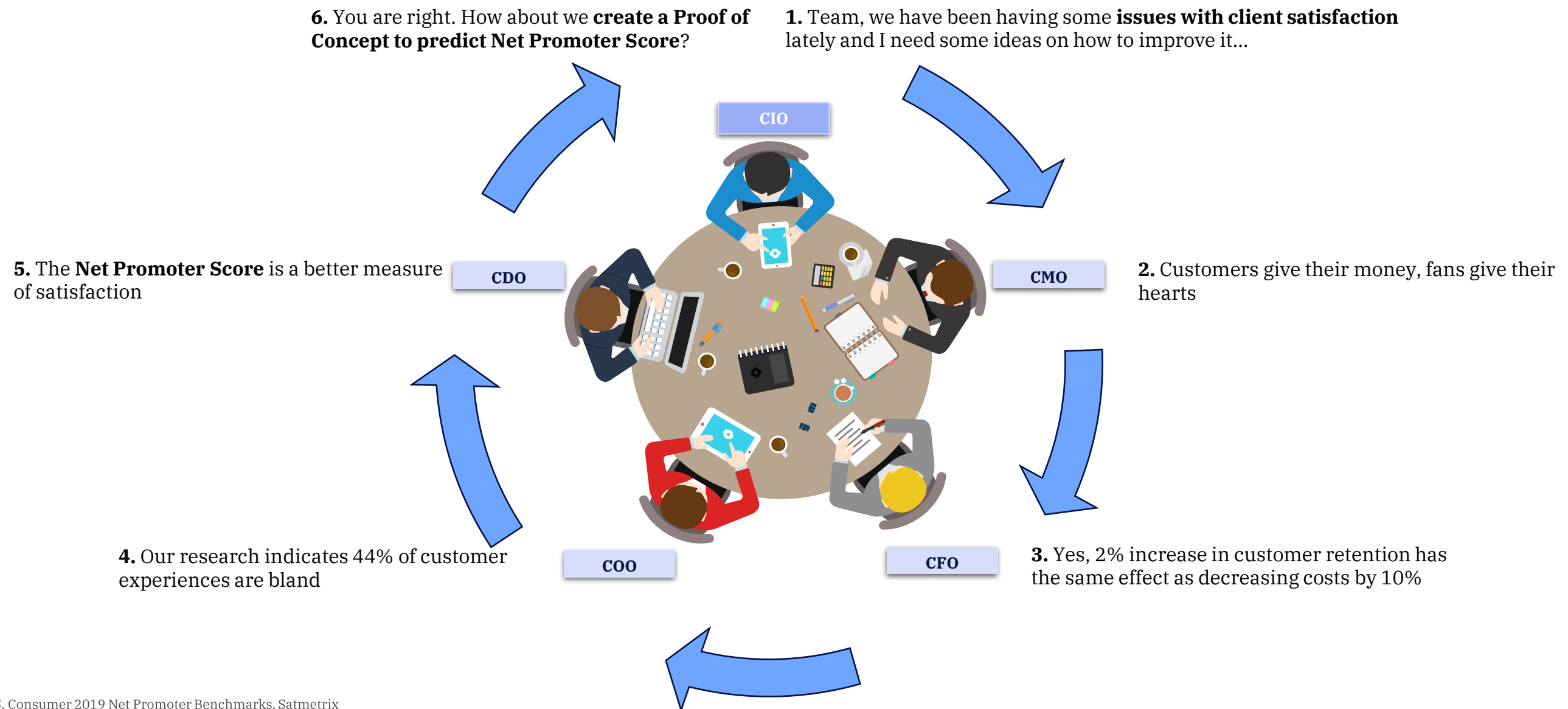
Improving the customer experience is valuable because of its effect on our bottom line. Creating an ultimate experience that appeals to both the heart and the head is our goal. Customers give their money, fans give their hearts of consumers. 44% of consumers say that majority of customer experiences are bland and 69% of consumers say that emotions count for half their experiences.

# Take away

Now, I am able to

- ✓ Create/ setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to Building a Machine Learning System

# Behind the Scenes: Let us talk about Improving Customer Experience with Real-Time Insights



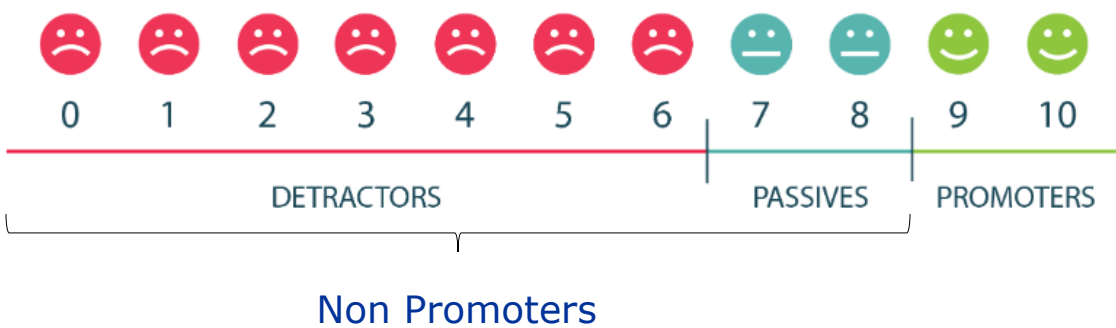
Source: U.S. Consumer 2019 Net Promoter Benchmarks, Satmetrix

# Assess Situation

In year 2020, **The Company** world-wide **supported 500,000 cases** which were created in multiple platforms.

The **Net Promoter Survey (NPS)** response rate was **15%**.

**60% cases were non-promoters and 40% were promoters.**



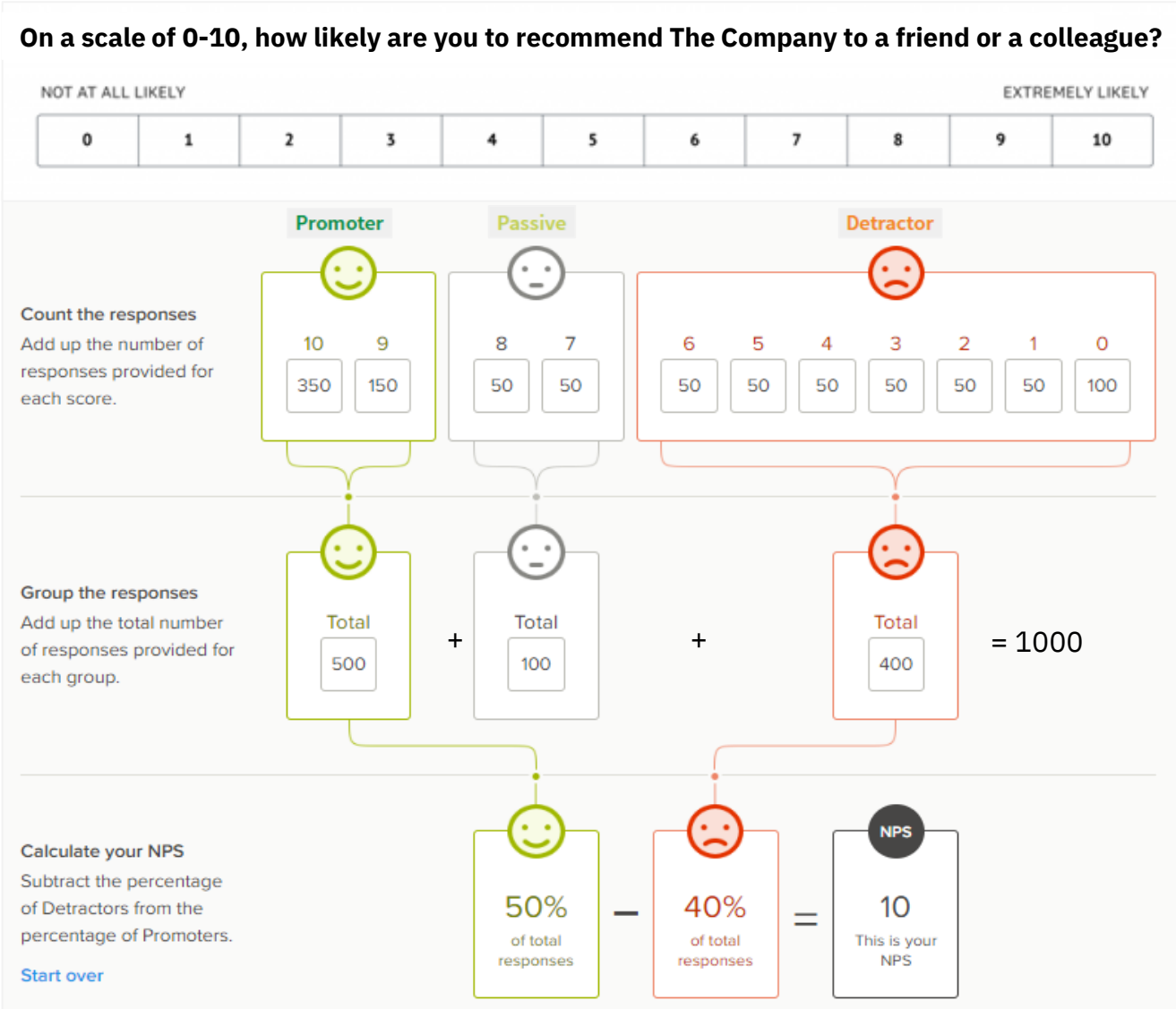
Note: The numbers highlighted above are crafted for this workshop.

# Methodology

**Net Promoter** has become the industry standard customer loyalty measurement. Businesses see customer experience as an imperative.

On a scale of 0-10, how likely would you recommend [brand/ support] to a friend or colleague?

Calculating NPS score is as simple as tallying up your responses and subtracting the percentage of detractors from the percentage of promoters. The score is a whole number that ranges from -100 to 100, and indicates customer happiness with our brand experience.

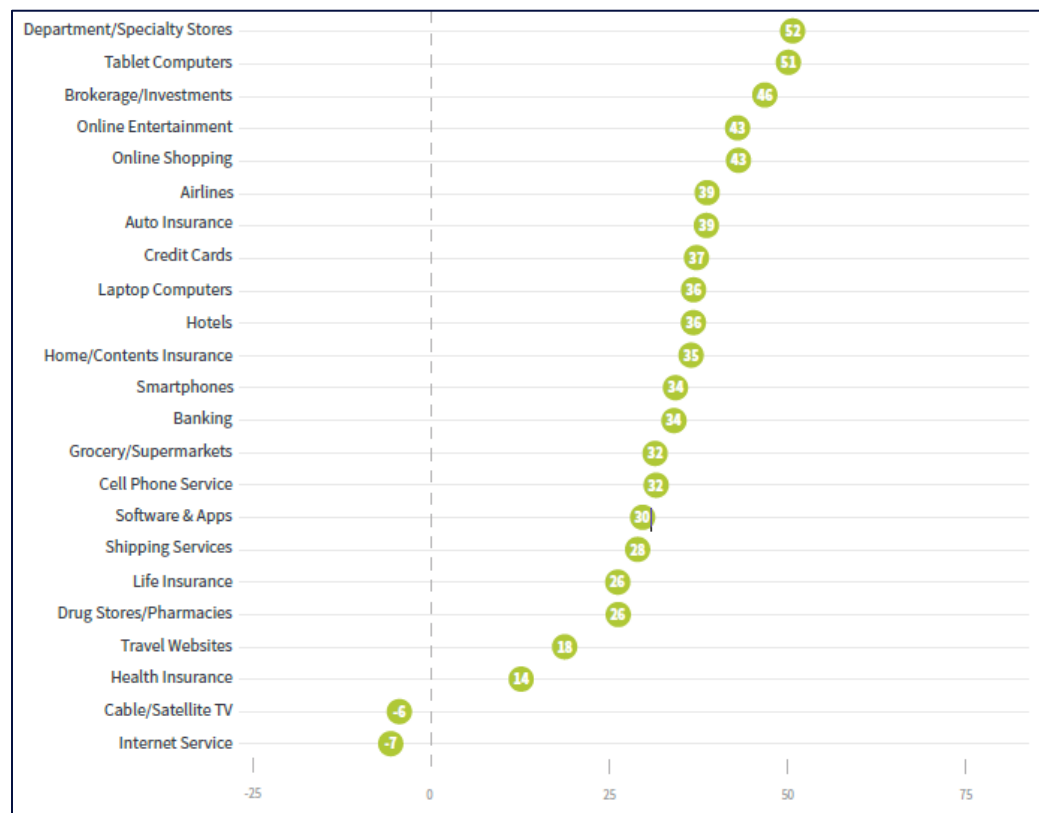


Source: How to Use Net Promoter to Drive Business Growth, Satmetrix and [NPS Calculator](#)

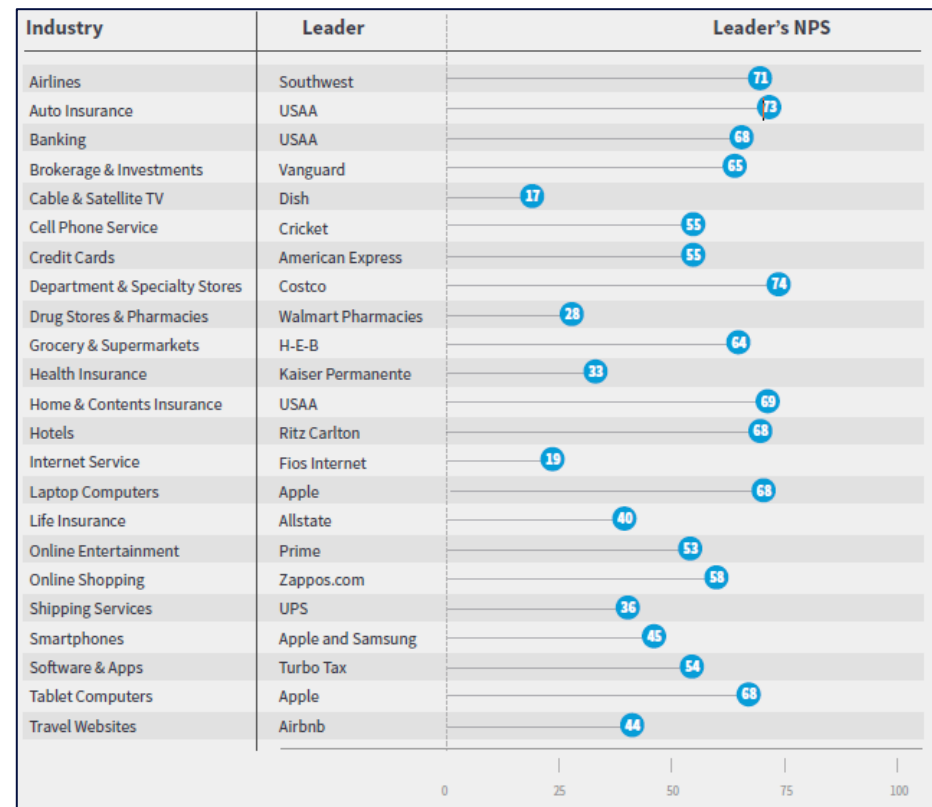


# Benchmarks (Average NPS by Industry and Leaders)

Net Promoter Scores vary widely by industry, as you can see from the average scores for 23 industries.



Knowing what similar companies have achieved helps us to set realistic goals for improvement, and realism is key to the long-term success of your program.



Source: U.S. Consumer 2019 Net Promoter Benchmarks, Satmetrix



# Business Objectives

**Goal:** Improve the Net Promoter Score by identifying potential non promoters ahead of time and proactively address customer issues

**Approach:** Consume historical NPS data; Use machine learning and artificial intelligence to identify the most important features and select an algorithm to predict the non promoters

**Desired Result:** Create a capability to share the top candidates for non-promoter surveys with The Company to proactively address customer issues.





# Exercise

1. Identify a data science opportunity in your business context and document.

Consider using this template...

As a <role>, I would like to <direction> the <target variable> for <scope> by <amount> in <timeframe>.

Role = End User

Direction = improve/reduce or increase/decrease

Target Variable = fraud, risk, customer satisfaction, volume, effort, price, cost, availability, productivity, revenue, etc.

Scope = section of the business of interest

Amount = value or percent

Timeframe = weeks, months, years



# Take-away

Now, I am able to

- ✓ Setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to build a Machine Learning System
- ✓ Assess the situation, understand the methodology, identify the benchmarks, and define business objectives


# Features Overview

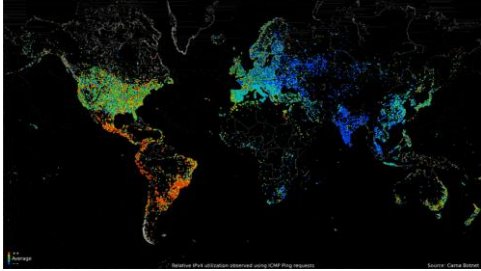
Time


Geography


Money

Sentiment & Emotions



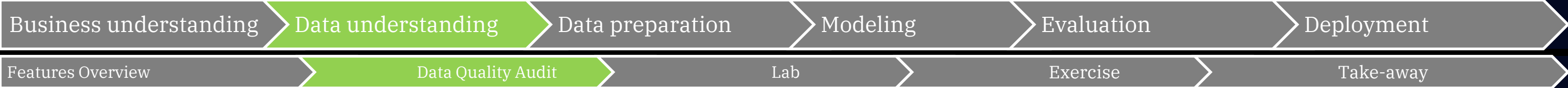






<u>TIME</u>	<u>LOCATION</u>	<u>MONEY</u>	<u>SENTIMENT &amp; EMOTIONS</u>
Day of Week	Country	Lifetime Spent	Sentiment
Time Window (Prime or Non-Prime)	Geography	Monthly Recurring Spend	Emotion (Anger, Disgust, Fear, Joy, and Sadness)
Age of Account	Region		
Meaningful Update			

Other Features: Assignment Count, Support Plan, Account Type, Severity, Technology, Case Origination Source, Case Origination User Type, Tribe, Catalog



# Data Quality Audit

## Pandas-Profiling Report (Things to check)

- ☐ Number of observations, features, and type
- ☐ Large Number of Distinct Values (High Cardinality)
- ☐ Correlation
- ☐ Missing values
- ☐ End goal is to review the data and improve Data Quality

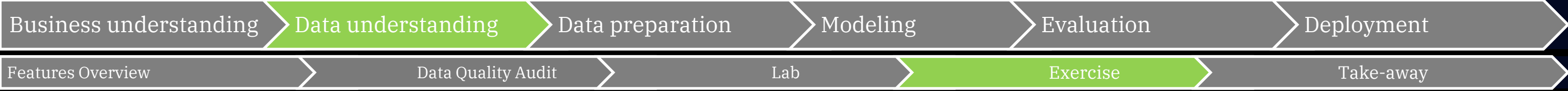




# Lab: Instructions

Run the following section in the notebook.

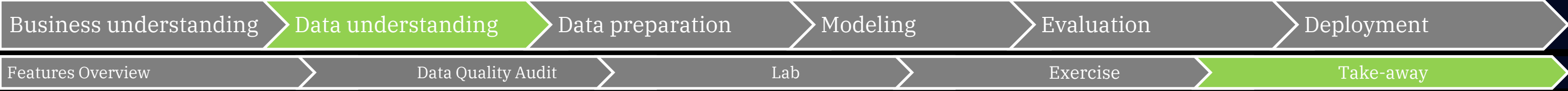
1. Introduction to Notebooks
2. Load packages and verify the version
3. Data Exploration
  - Load and read the files from GitHub
  - Explore the data and perform quality audit



# Exercise

2. What data set would you gather to work on the problem statement?

Be sure to include the feature that you believe would influence the outcome, the definition, type, and range (e.g., weather, forecasted rain probability at 10 am, percent, 0.0 to 1.0) of that feature, and the data source that would provide it (weather.com).



# Take away

Now, I am able to

- ✓ Setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to build a Machine Learning System
- ✓ Assess the situation, understand the methodology, identify the benchmarks, and define business objectives
- ✓ Introduction to Notebook, load packages, and verify the versions
- ✓ Explore the data set and perform quality audit



# Extract, Scale, and Select Features (1 of 2)

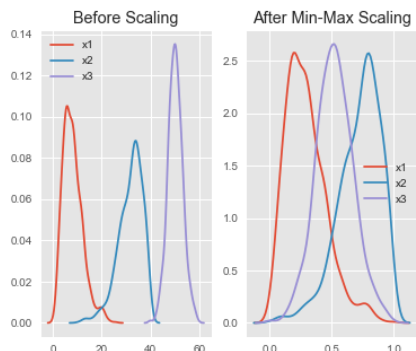
## Feature Extraction

### One Hot Encoding



- Other Techniques:
- OrdinalEncoding
  - LabelEncoder
  - BinaryEncoder
  - Hashing Encoder
  - Target/Mean Encoding
  - Autoencoders

## Feature Scaling



- Other Techniques:
- StandardScaler
  - RobustScaler
  - Normalizer

## Feature Selection

### Dimensionality Reduction



- Other Techniques:
- Pairwise Correlation
  - Multi- collinearity
  - Principal Component Analysis,
  - Cluster Analysis, Correlation (with the target)
  - Forward/ Backward/ Stepwise selection
  - LASSO
  - Tree-based selection



# Extract, Scale, and Select Features (2 of 3)

## Why (a, b, and c) and (x, y, and z) are commonly used as mathematical placeholders?

The letters at the end of the alphabet, viz., x, y, z, etc. are to denote unknown variables, while those at the start of the alphabet, a, b, c, etc. denote constants, was first highlighted in *La Géométrie* in year 1637.



# Extract, Scale, and Select Features (3 of 3)

**Why are the input variables denoted by an upper-case 'X' whereas the target variable is a lower-case 'y'?**

In terms of Linear Algebra, it is extremely common to use capital Latin letters for matrices (e.g., design matrix  $X$ ) and lowercase Latin letters for vectors (response vector  $y$ ).

A vector is a list of numbers (can be in a row or column), A matrix is an array of numbers (one or more rows, one or more columns).





# Lab: Instructions

4. Feature Extraction

5. Feature Scaling

6. Feature Selection



# Exercise

3. How would you prepare the dataset and what challenges do you foresee?

**Now, as you have shortlisted the input features to prepare a data model, document the transformation steps (extract, scale, and select) you would apply on features to prepare the data-set.**

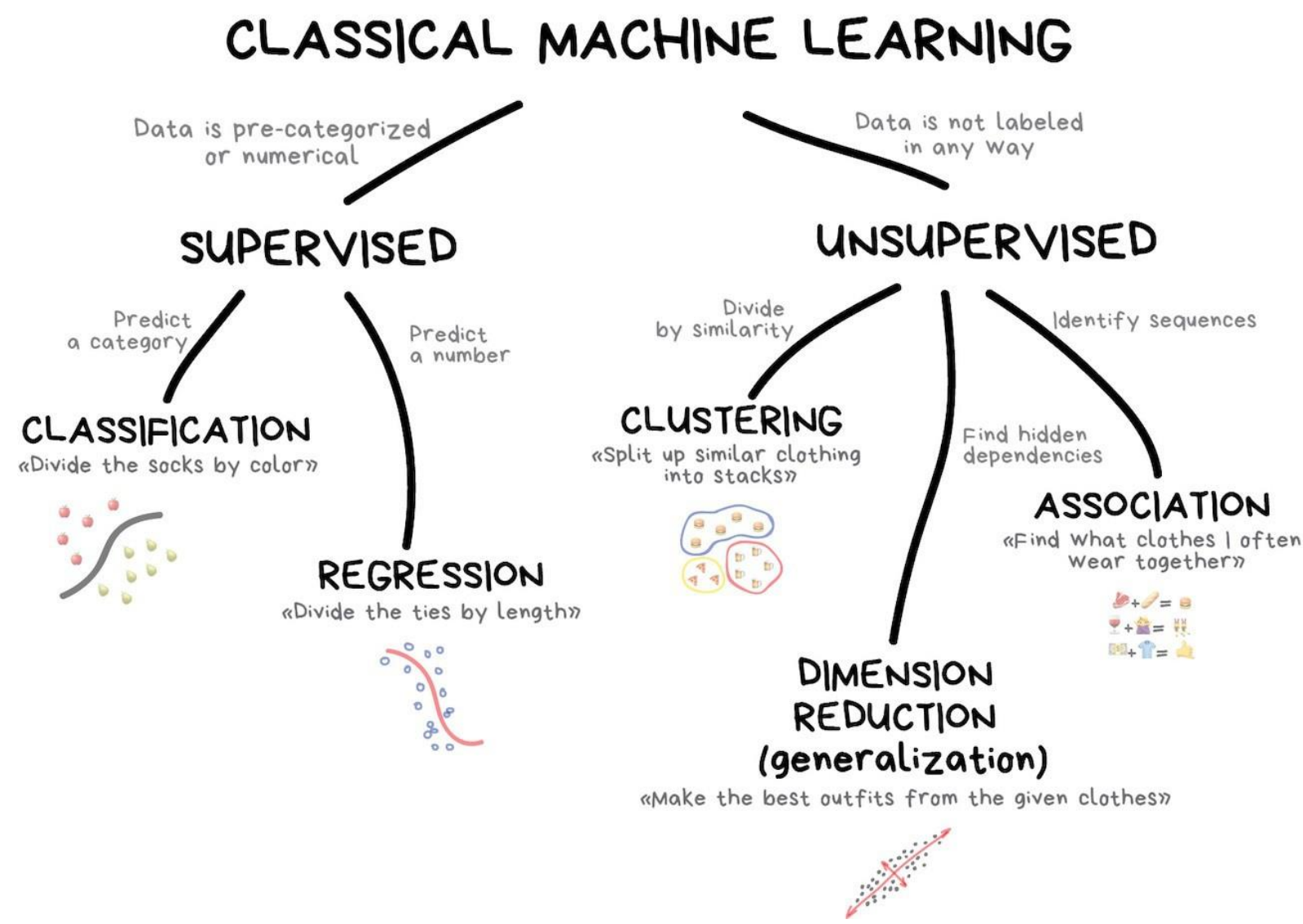


# Take away

Now, I am able to

- ✓ Setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to build a Machine Learning System
- ✓ Assess the situation, understand the methodology, identify the benchmarks, and define business objectives
- ✓ Introduction to Notebook, load packages and verify the versions
- ✓ Explore the data set and perform quality audit
- ✓ Extract, scale, and select features for the data model

# Machine Learning Algorithm Selection (1 of 2)



Source: [Machine Learning for Everyone](#)



# Machine Learning Algorithm Selection (2 of 2)

Supervised Learning		Unsupervised Learning		
Classification	Regression	Clustering	Dimensionality Reduction	Association

1. Spam filtering

2. Fraud detection

3. Stock price forecasts

4. House Price
5. Customer Segmentation

6. Topic modeling and similar document search

7. To place the products on the shelves





# Machine Learning Algorithm Selection (2 of 2)

Supervised Learning		Unsupervised Learning		
Classification	Regression	Clustering	Dimensionality Reduction	Association
Spam filtering Fraud detection	Stock price forecasts House Price	Customer Segmentation	Topic modeling and similar document search	To place the products on the shelves

Selection

Performance Metrics

Evaluation

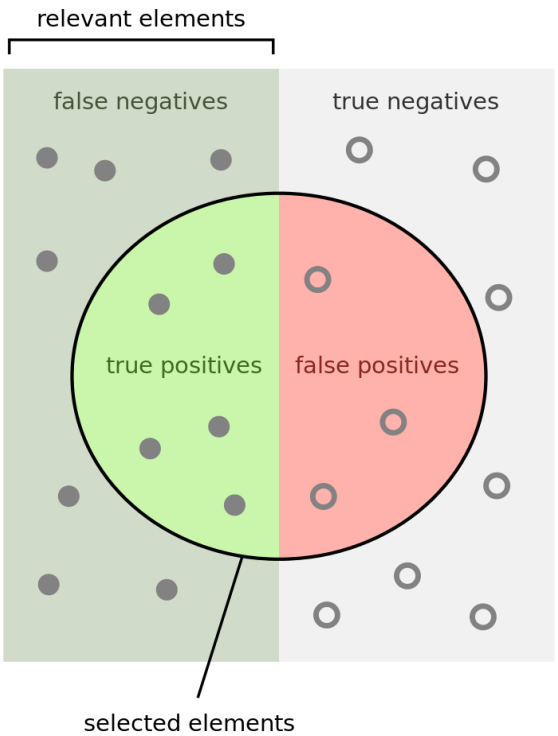
Demonstration

Lab

Exercise

Exercise

# Measure Model Performance



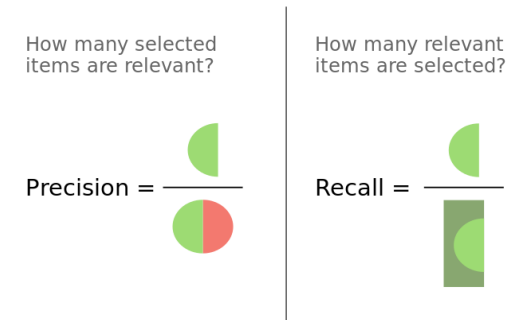
	Predicted:		
	NO	YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



## What is Confusion Matrix?

Describes the performance of a classification model on a set of test data for which true values are known.

- Accuracy:** Overall, how often is the classifier correct?  
 $(TP+TN)/total = (100+50)/165 = 0.91$
- Precision:** When it predicts yes, how often is it correct?  
 $(TP/predicted\ yes = 100/110 = 0.91)$
- Recall:** When it's yes, how often does it predict yes?  
 $(TP/actual\ yes = 100/105 = 0.95)$
- F1:** harmonic mean of precision and recall:  $(2 * precision * recall) / (precision + recall)$



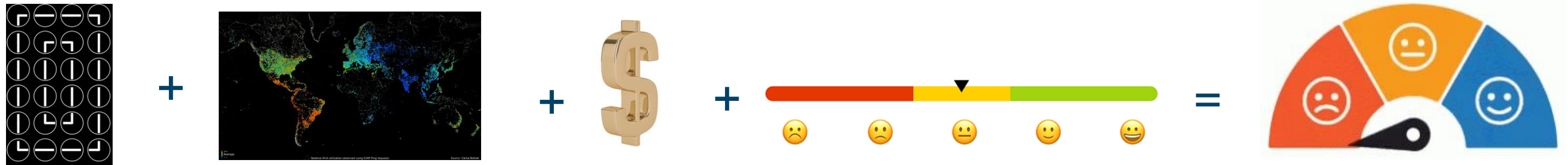


# Model Evaluation

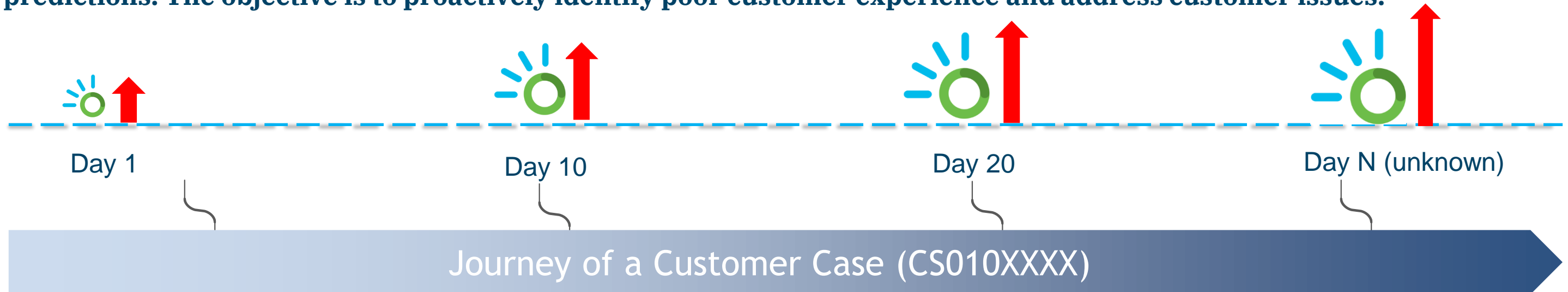
	Metrics					
Model Name	Accuracy		Precision	Recall	F1 Score	
LogisticRegression	⬆️	65%	63%	59%	↗️	58%
SGDClassifier	➡️	55%	59%	58%	➡️	55%
SVM	↗️	60%	59%	60%	↗️	59%
KNeighborsClassifier	↗️	56%	53%	53%	➡️	53%
GaussianProcessClassifier	⬆️	62%	57%	55%	➡️	53%
MultinomialNB	⬆️	62%	59%	57%	↗️	57%
DecisionTreeClassifier	↗️	56%	54%	54%	➡️	54%
RandomForestClassifier	↗️	60%	55%	54%	➡️	53%
GradientBoostingClassifier	⬆️	67%	64%	62%	⬆️	62%
VotingClassifier	⬆️	63%	59%	55%	➡️	52%
MLPClassifier	⬆️	62%	31%	50%	⬇️	38%



# Demonstration



**The algorithm consumes multiple signals (time, geography, spend, and sentiments) and gives the non promoter predictions. The objective is to proactively identify poor customer experience and address customer issues.**





# Lab: Instructions

Run the following section in the notebook.

7. Split data into train and test sets

8. Measure Model Performance

9. Evaluate and Select Model



# Exercise

4. What modeling techniques would you attempt, and metrics would you use to evaluate the model performance?

**Consider the type of output you are generating and choose an appropriate technique for the model. Remember, accuracy, precision, recall, F1 and think about “explainability” of the model results.**



# Take away

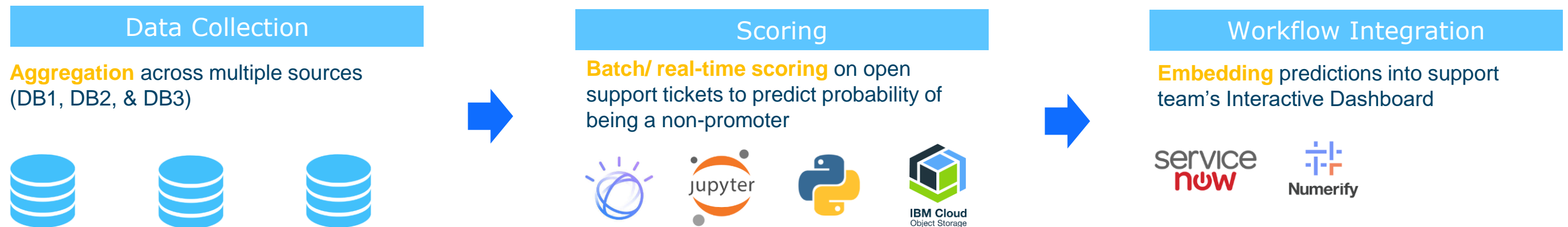
Now, I am able to

- ✓ Setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to build a Machine Learning System
- ✓ Assess the situation, understand the methodology, identify the benchmarks, and define business objectives
- ✓ Introduction to Notebook, load packages and verify the versions
- ✓ Explore the data set and perform quality audit
- ✓ Extract, scale, and select features for the data model
- ✓ Split data into train & test sets, select model, evaluate performance metrics, and demonstrate

# Proposed Solution

The model developed as a part of the hack uses artificial intelligence and machine learning to predict Non-Promoters on historical data pattern. Key aspects of approach include,

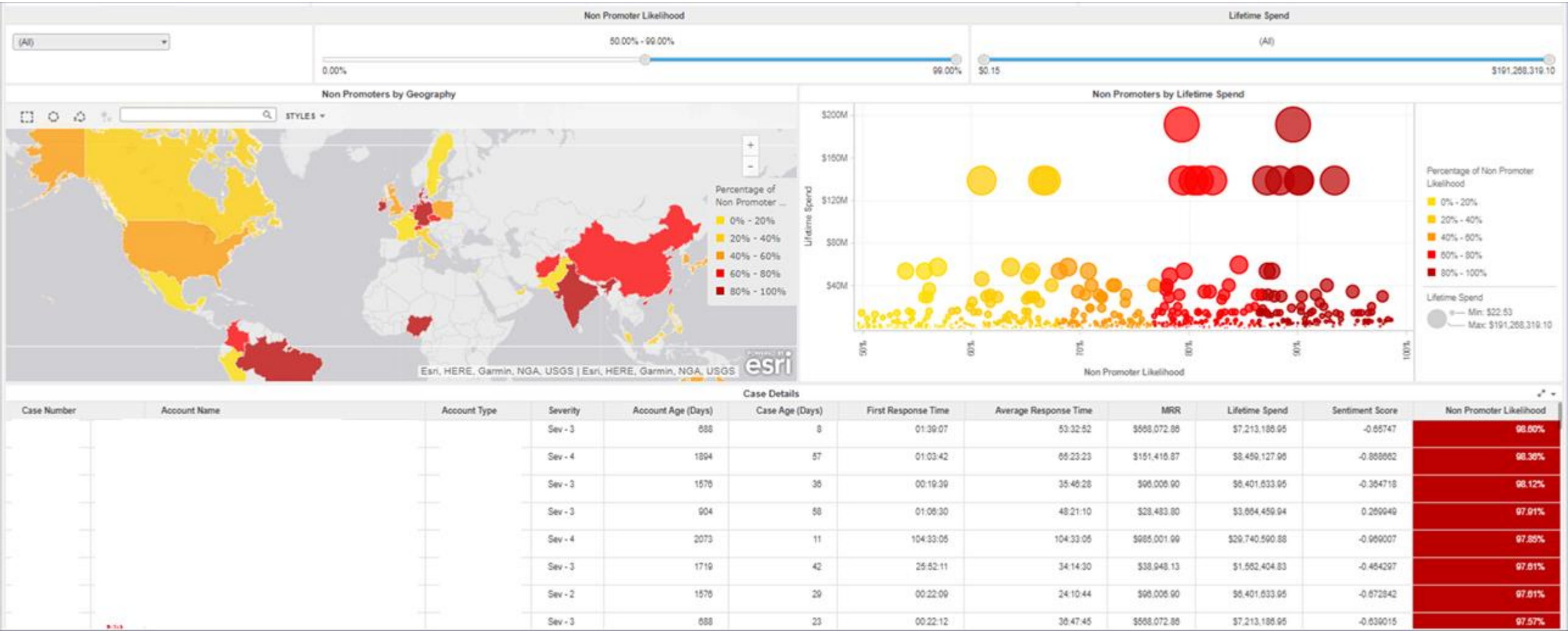
1. Watson NLP to create additional features from customer conversation logs
2. Machine Learning algorithms to produce the predictions



# Solution Integration

“Ideas are easy, Execution is everything.” John Doerr

## NPS Predictions Board



Note: This visual highlighted above is crafted for workshop purpose.



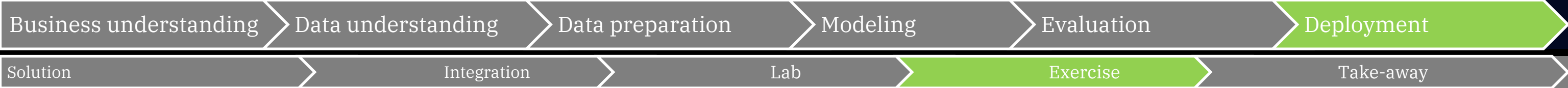
# Lab: Instructions

Run the following section in the notebook.

10. Save the Model

11. Deploy the Model

12. Predict Cases

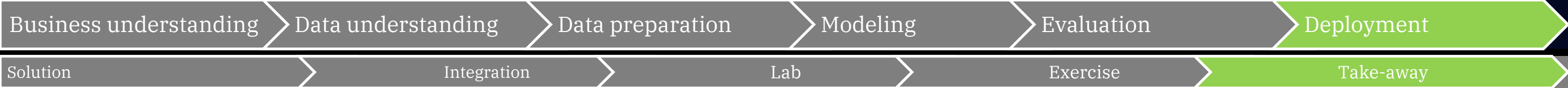


# Exercise

5. How do you plan to consume the outputs of model?

**Consider dashboards, reports, visualizations or automated decisions. Describe how the end user interacts with results.**



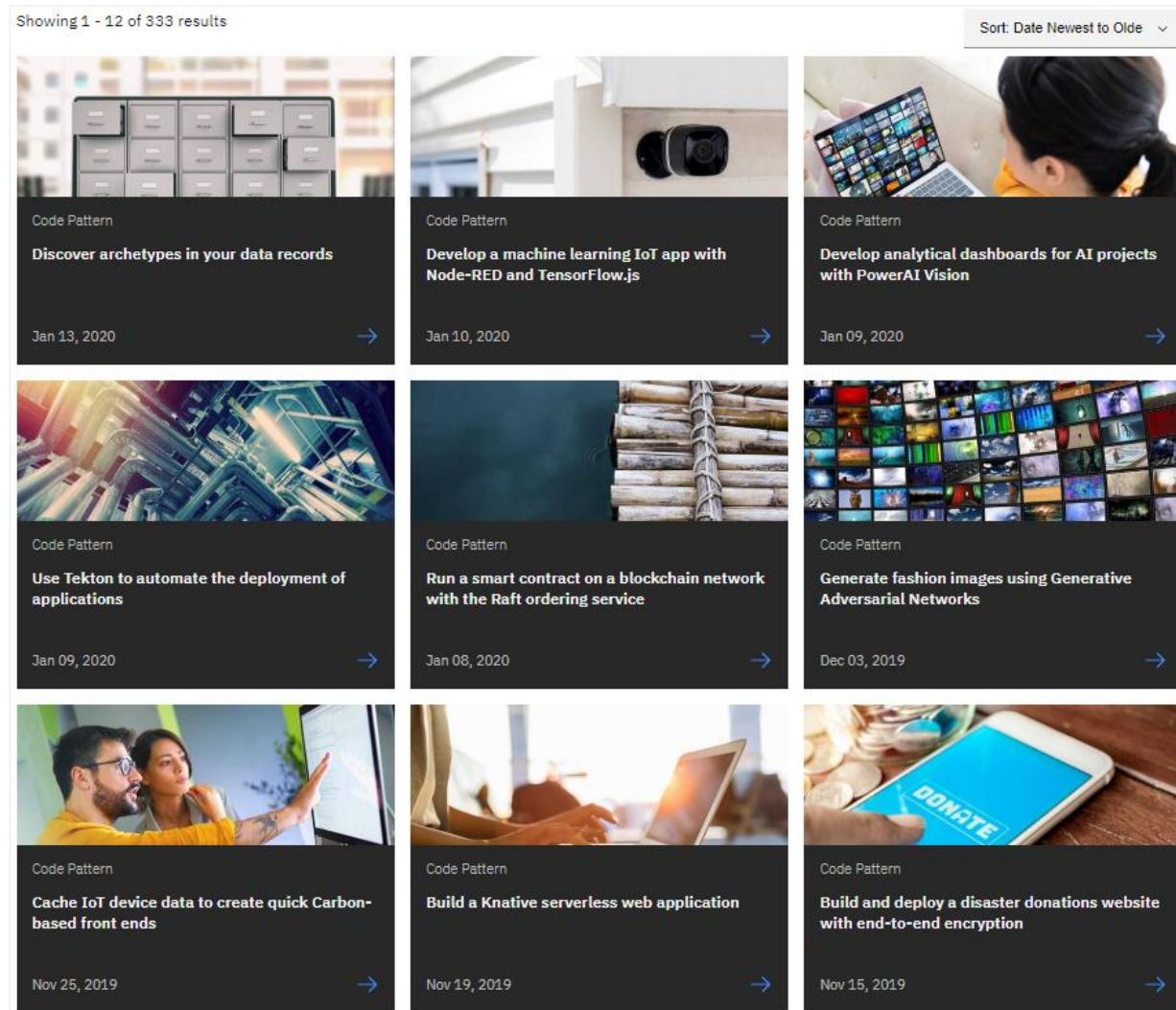


# Take away

Now, I am able to




- ✓ Setup the Data Science environment on IBM Cloud
- ✓ Learn the Roadmap to build a Machine Learning System
- ✓ Assess the situation, understand the methodology, identify the benchmarks, and define business objectives
- ✓ Introduction to Notebook, load packages and verify the versions
- ✓ Explore the data set and perform quality audit
- ✓ Extract, scale, and select features for the data model
- ✓ Split data into train & test sets, select model, evaluate performance metrics, and demonstrate
- ✓ Generate ideas on how to consume the predictions and integrate solution in business systems

# IBM Code Patterns



Link: <https://developer.ibm.com/patterns/>

# Contact Us

	<p>Name: Maureen Norton Email Id: <a href="mailto:nortonm@us.ibm.com">nortonm@us.ibm.com</a> LinkedIn: <a href="https://www.linkedin.com/in/maureennorton/">https://www.linkedin.com/in/maureennorton/</a></p>
	<p>Name: Mwai Kalengamaliro Email Id: <a href="mailto:mkaleng@us.ibm.com">mkaleng@us.ibm.com</a> LinkedIn: <a href="https://www.linkedin.com/in/mwai-kalengamaliro/">https://www.linkedin.com/in/mwai-kalengamaliro/</a></p>
	<p>Name: Neeraj Madan Email Id: <a href="mailto:nmadan@us.ibm.com">nmadan@us.ibm.com</a> LinkedIn: <a href="https://www.linkedin.com/in/neerajmadan/">https://www.linkedin.com/in/neerajmadan/</a></p>

# Thank you



© marketoonist.com

