Data Science Workshop
An Experiential Journey with Data to Inspire ***Your*** Work

George Stark
Maureen Norton
Neeraj Madan
January 29, 2020

The Open Group Conference
Making Data Useful
San Antonio, Texas
January 29, 2020
9:00 – 12:30

Data Science Workshop

An Experiential Journey with Data to Inspire *Your* Work

January 27, 2020 | January 28, 2020 | **January 29, 2020** | January 30, 2020

Data Science Workshop

9:00 AM - 12:30 PM	**An Experiential Journey With Data to Inspire Your Work** 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 cases that will make you think, sometimes laugh and hopefully inspire your own work. The use cases and introductory material will be followed by 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 NPS data; using machine learning and artificial intelligence to identify the most important features and created an algorithm to predict the customer experience. **Facilitators:** Neeraj Madan, Maureen Norton, George Stark

IBM

The Open Group Conference
Making Data Useful
San Antonio, Texas
January 29, 2020
9:00 – 12:30

Agenda

Section	Time
Getting Started ***a. Session Introduction and Expectation Setting*** ***b. Data Science Introduction*** *a. Let's Talk About Data* *b. Common Business Models* *c. Data Science Techniques* *d. Participant Workbook*	**9:00 am – 10:00 am**
Predictive Analytics and Machine Learning Solutions **IBM Watson Studio and Machine Learning (Introduction and Setup)** **<< Break >>**	**10:00 am – 10:40 am**
Hands on Experiential Journey (Net Promoter Score Example) ***a. Business understanding:*** *Exercise 1: Identify an opportunity in your business context and document* ***b. Data understanding:*** *Exercise 2: What data set would you gather to work the problem statement* ***c. Data preparation:*** *Exercise 3: How would you prepare the dataset and what challenges do you foresee?* **<<Break (15 mins)>>** ***d. Modeling:*** *Exercise 4: What modeling techniques would you attempt and why?* ***e. Evaluation:*** *Exercise 5: What metrics would you use to evaluate your model performance?* ***f. Deployment:*** *Exercise 6: How do you plan to consume the outputs of the model?*	**10: 45 am – 12:30 pm**

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

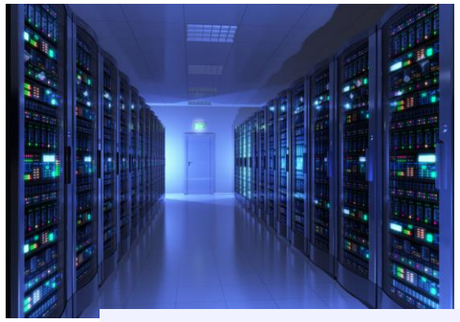
Four Common Data Science Models

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)

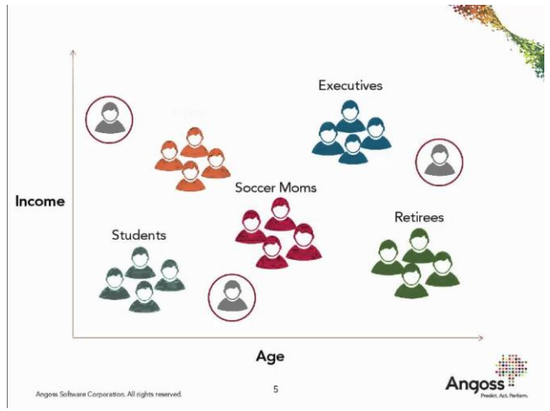
Quality/Defect Prediction

Identify problematic components, predict number of defects in a product (e.g., code, castings, compounds, raw materials, ATM Machines)



Business Value/ Customer Satisfaction

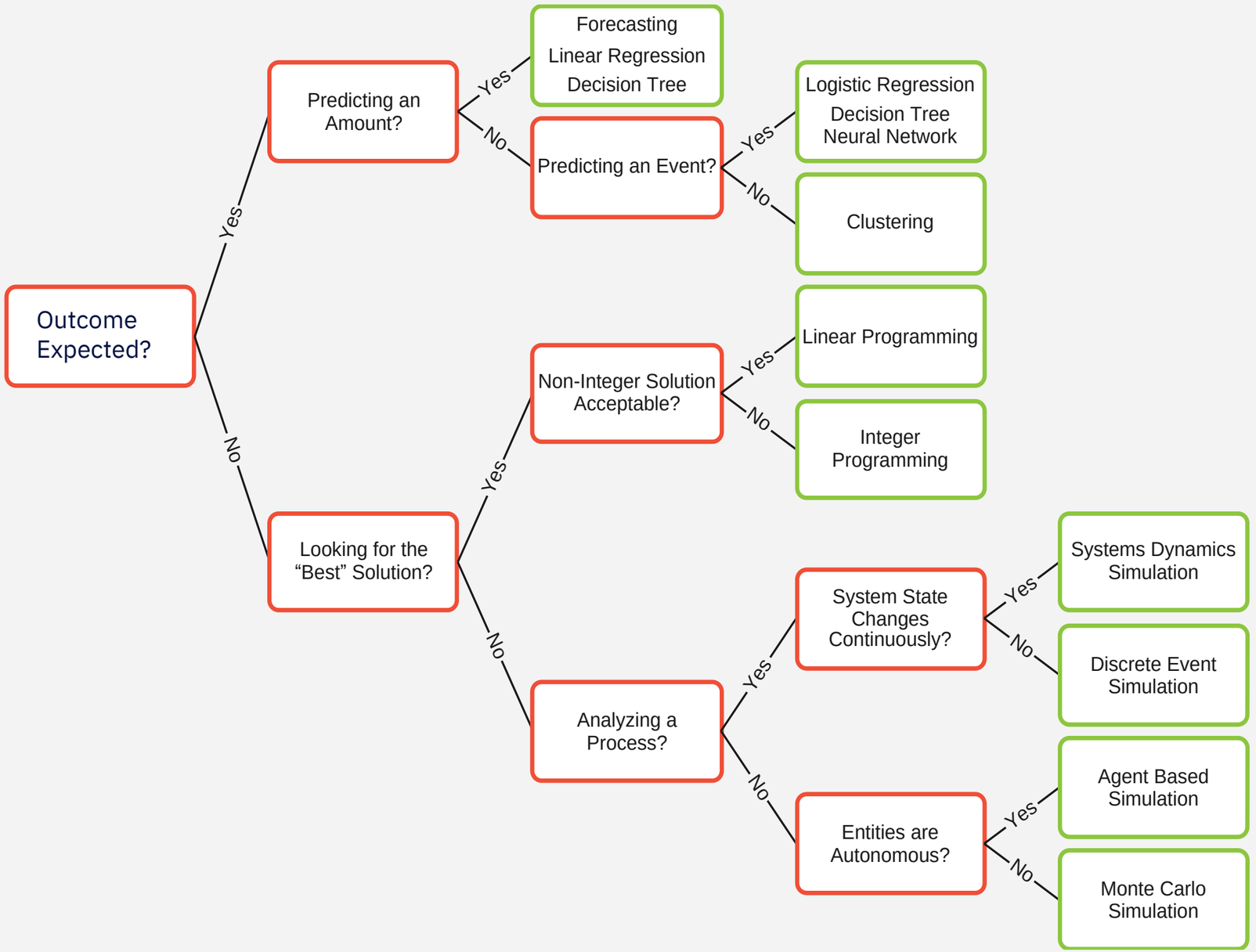
Create a classification algorithm that accurately identifies which customers have the most potential business value based on their characteristics and activities. Which customers are likely to be happy? Which will be promoters?



Price/Cost/Value

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

Choosing the right analytic approach



Data Science Project One-Pager

Project Description	Business Lead: Last, First Data Science Lead: Last, First Architect: Last, First

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment

✓ Current Business Situation
✓ Decision Maker Identification
✓ Data-Mining Success Criteria from a Business Perspective
  ✓ Cupcake Deliverable: “Classify a transaction with class of profit”
  ✓ Wedding Cake Deliverable: “Predict with 80% Confidence Interval (CI) the GP of a transaction”
	❑ Existing Data Sources
	  ✓ Customer Loyalty
	  ❑ Sales
	  ❑ Users (via API)
	  ❑ Financial
	  ❑ SKU
	❑ Additional Data
	  ❑ MagicHat
		❑ Merging
		❑ Feature Selection
		❑ Aggregation
		❑ Feature Engineering
		  ❑ OneHotEncoding
		  ❑ TF-IDF
		  ❑ Principal Component Analysis (PCA)
		❑ Data Quality Strategy
		  ❑ Drop
		  ❑ Imputation
		  ❑ Mean
		  ❑ Median
		  ❑ ML-based (KNN)
		❑ Test/Train Split
		  ❑ K-fold CV
			❑ Model Types
			  ❑ Random Forest Regressor
			  ❑ KNN-based ensemble
			  ❑ Generalized Additive Models (GAM)
			❑ Model Requirements
			  ❑ L1/L2 Regularization
			❑ Ensemble Pipeline
			❑ LIME-interpolation
				❑ Hyper parameter Validation
				  ❑ Validation Curve
				  ❑ GridSearch CV
				❑ Visualizations
				❑ Evaluation Metrics
				  ❑ Classification
				  ❑ Accuracy
				  ❑ Precision
				  ❑ Recall
				  ❑ Confusion Matrix
				  ❑ Regression
				  ❑ R2
				  ❑ RMSE
				❑ Business Feedback
					❑ Summary Findings
					❑ Integration
					❑ Model Monitoring
					❑ Model Maintenance
					❑ Enhancements
					  ❑ Recommendations
					  ❑ UI/UX Design Thinking
					  ❑ Cognitive

Predictive Analytics and Machine Learning Solutions

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

"The Forrester Wave™: Multimodal Predictive Analytics And Machine Learning Solutions, Q3 2018", Forrester Research, September 2018

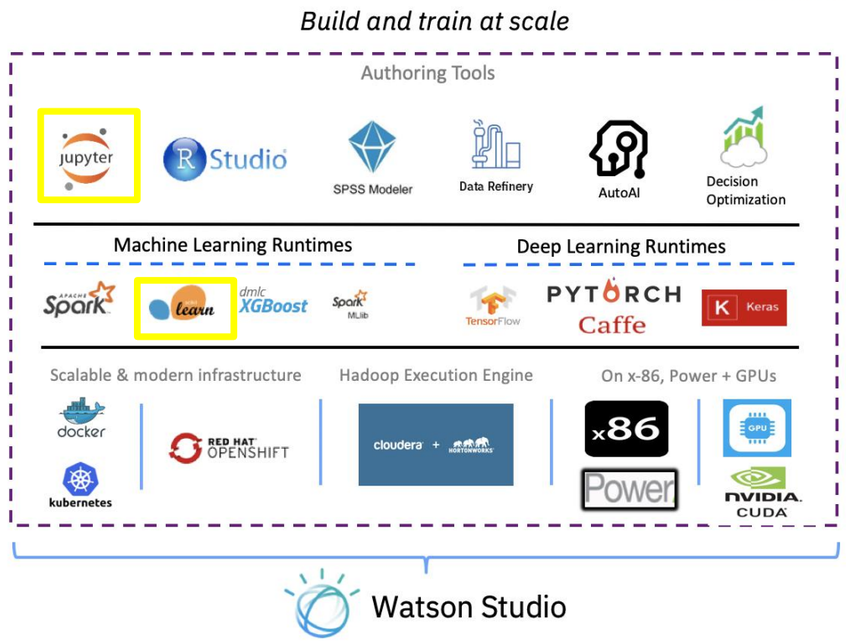
IBM Watson Studio

(Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business)

Analyze any data, no matter where it lives

Empower your entire organization with notebooks, visual productivity, and automation tools

One platform to rule them all from discovery to production

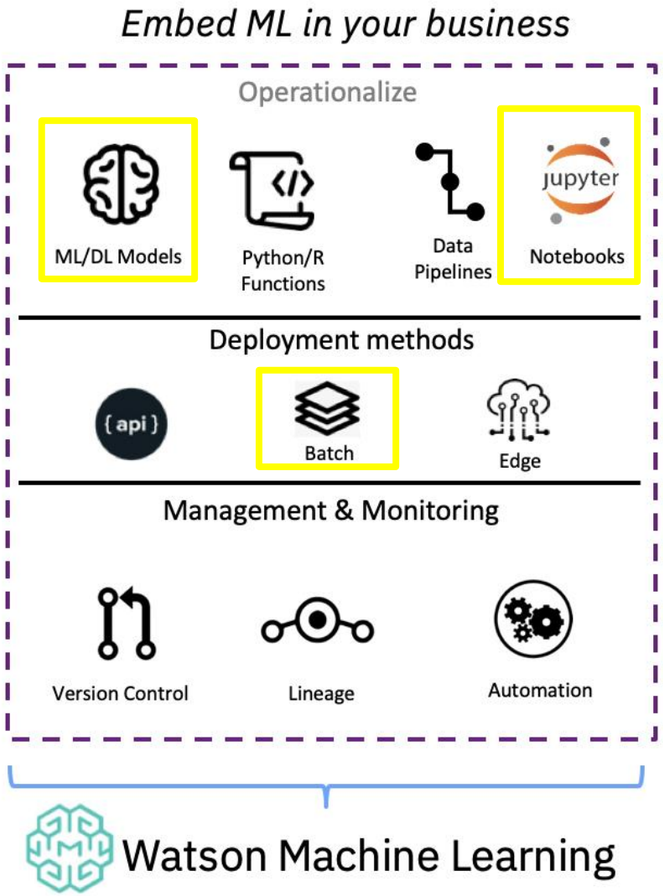


IBM Watson Machine Learning
(Embed Machine Learning and Deep Learning in your Business)

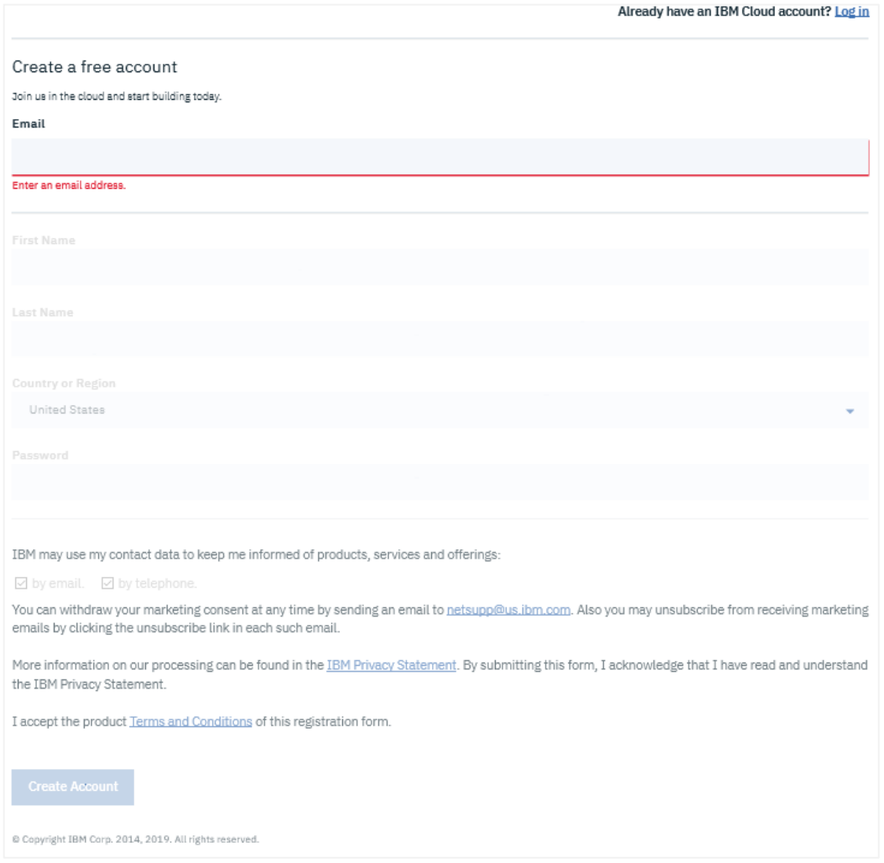
Deploy and Manage Models

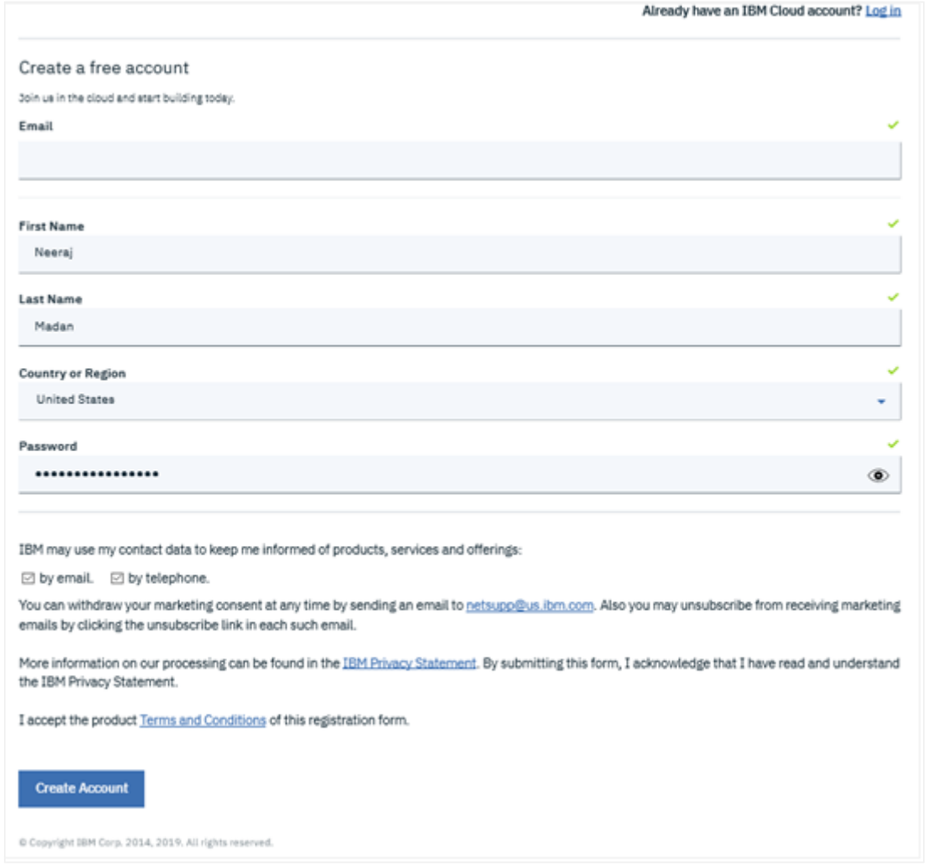
Intelligent Model Operations

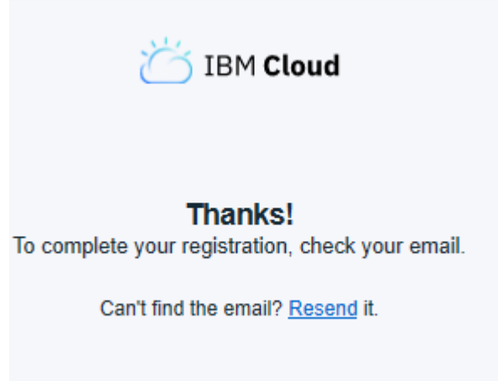
Accelerate Compute Intensive Workloads



Getting Started: IBM Watson Studio Setup
(Step 1: Sign up / Log into IBM Cloud - https://ibm.biz/BdqDmG)

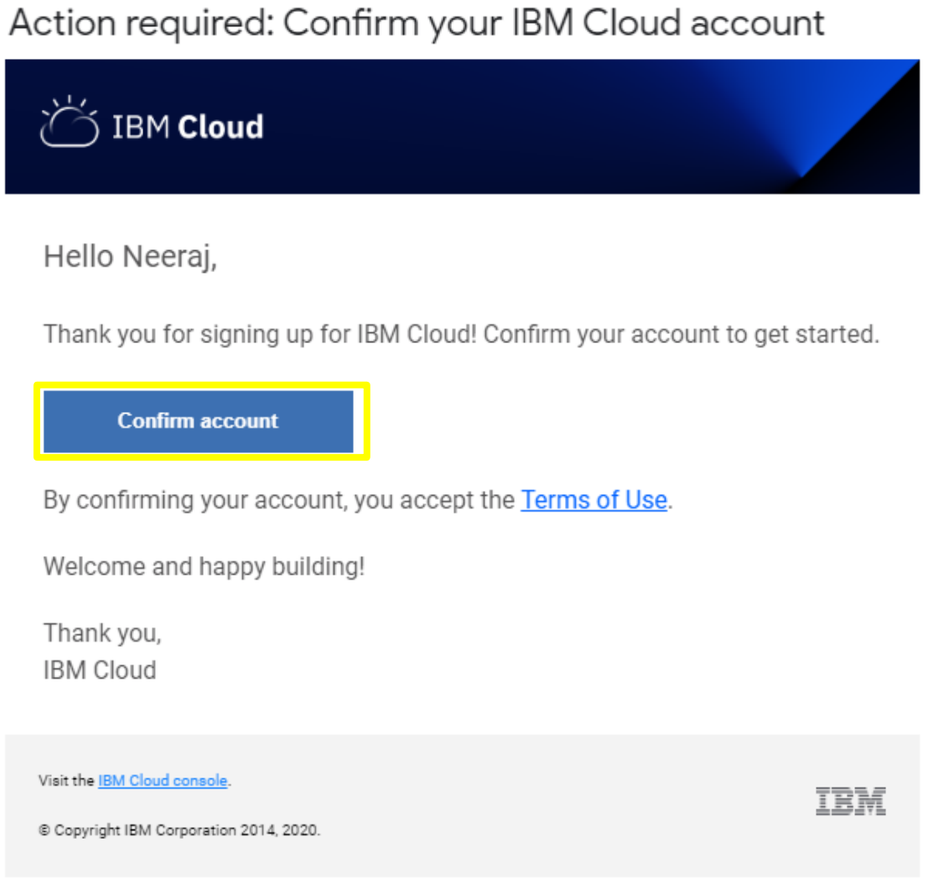


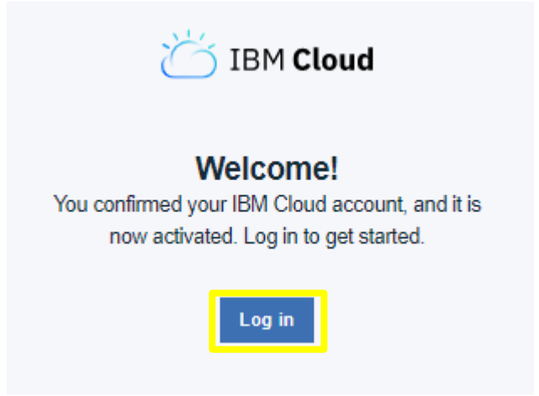


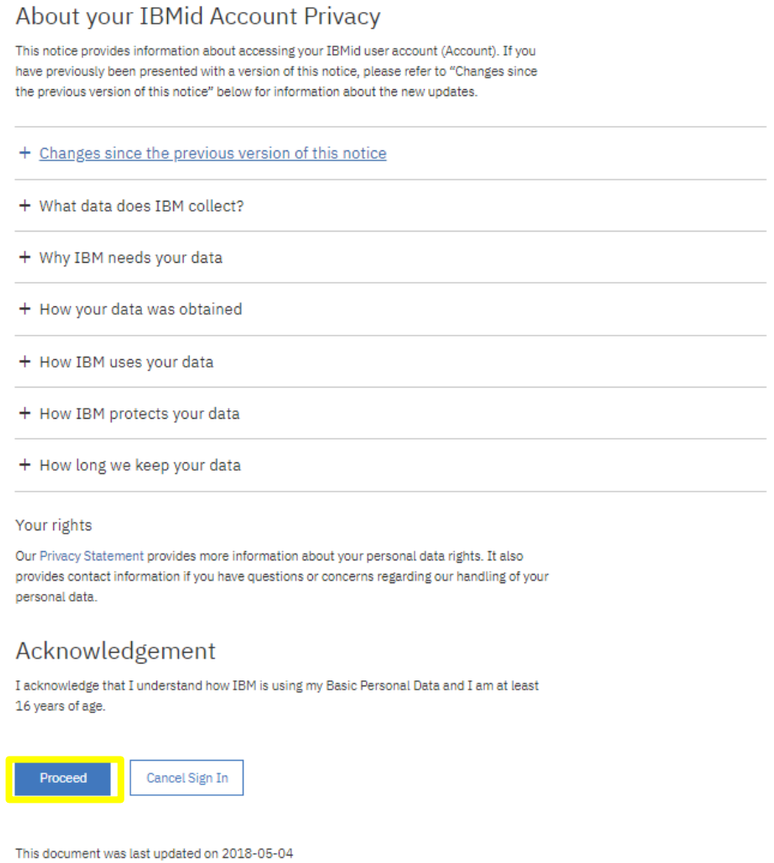


IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

Getting Started: IBM Watson Studio Setup
(Step 2: Confirm your account (Email verification)

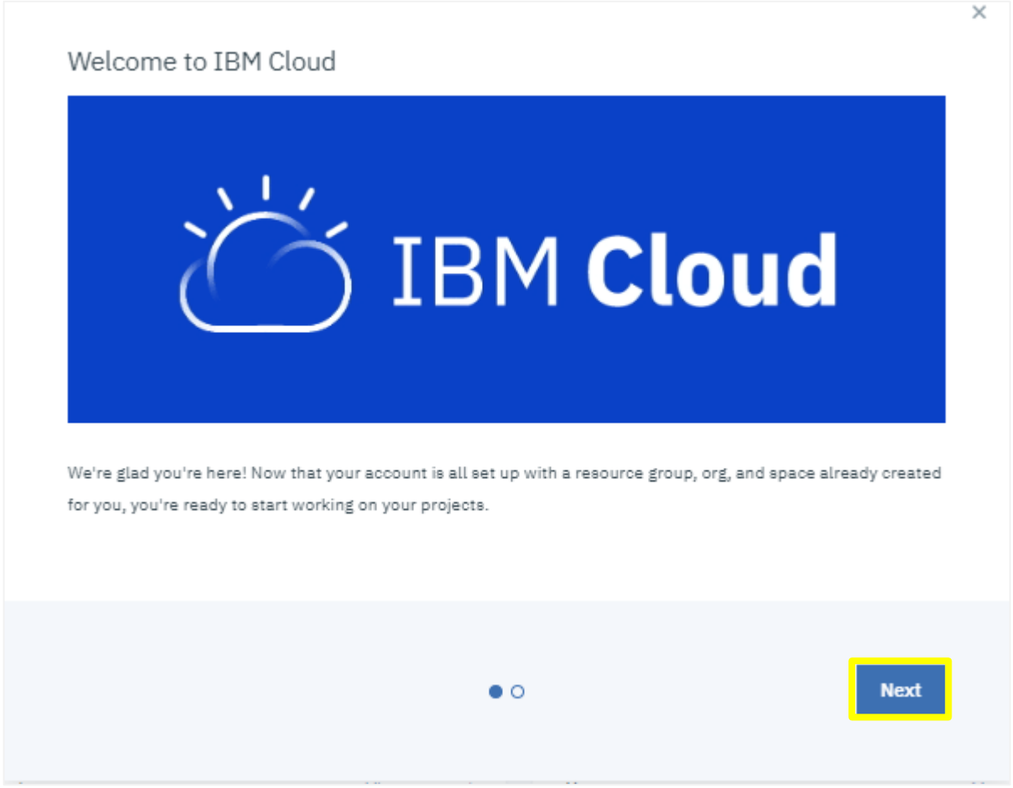


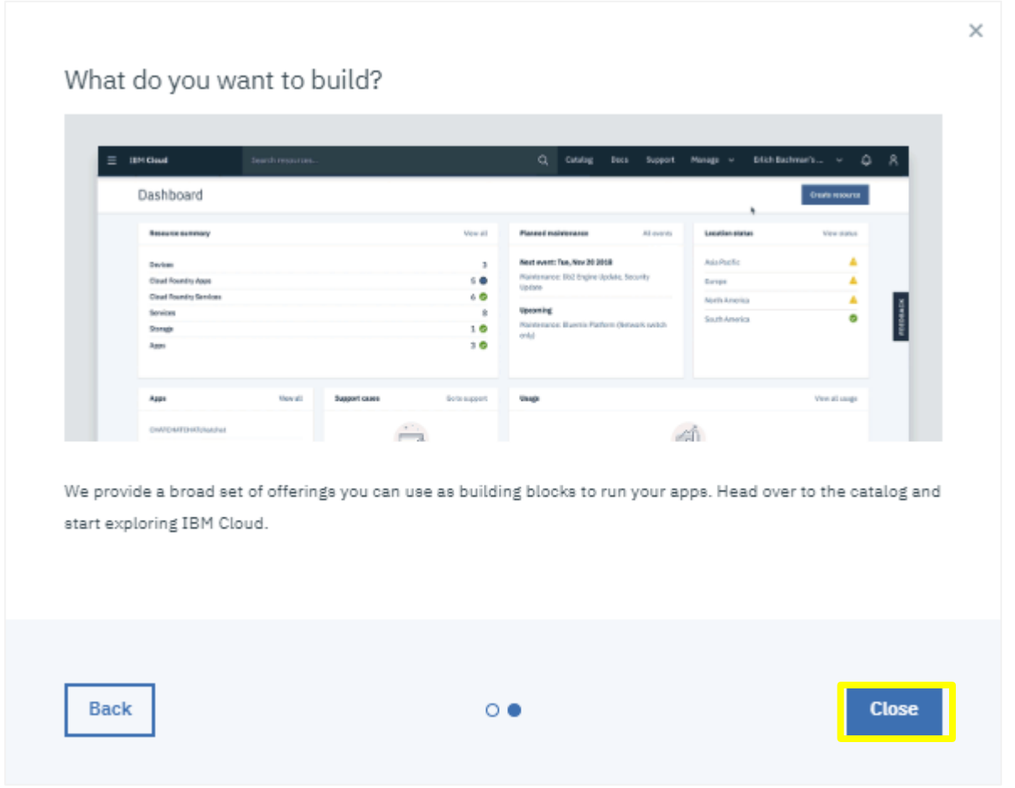




IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

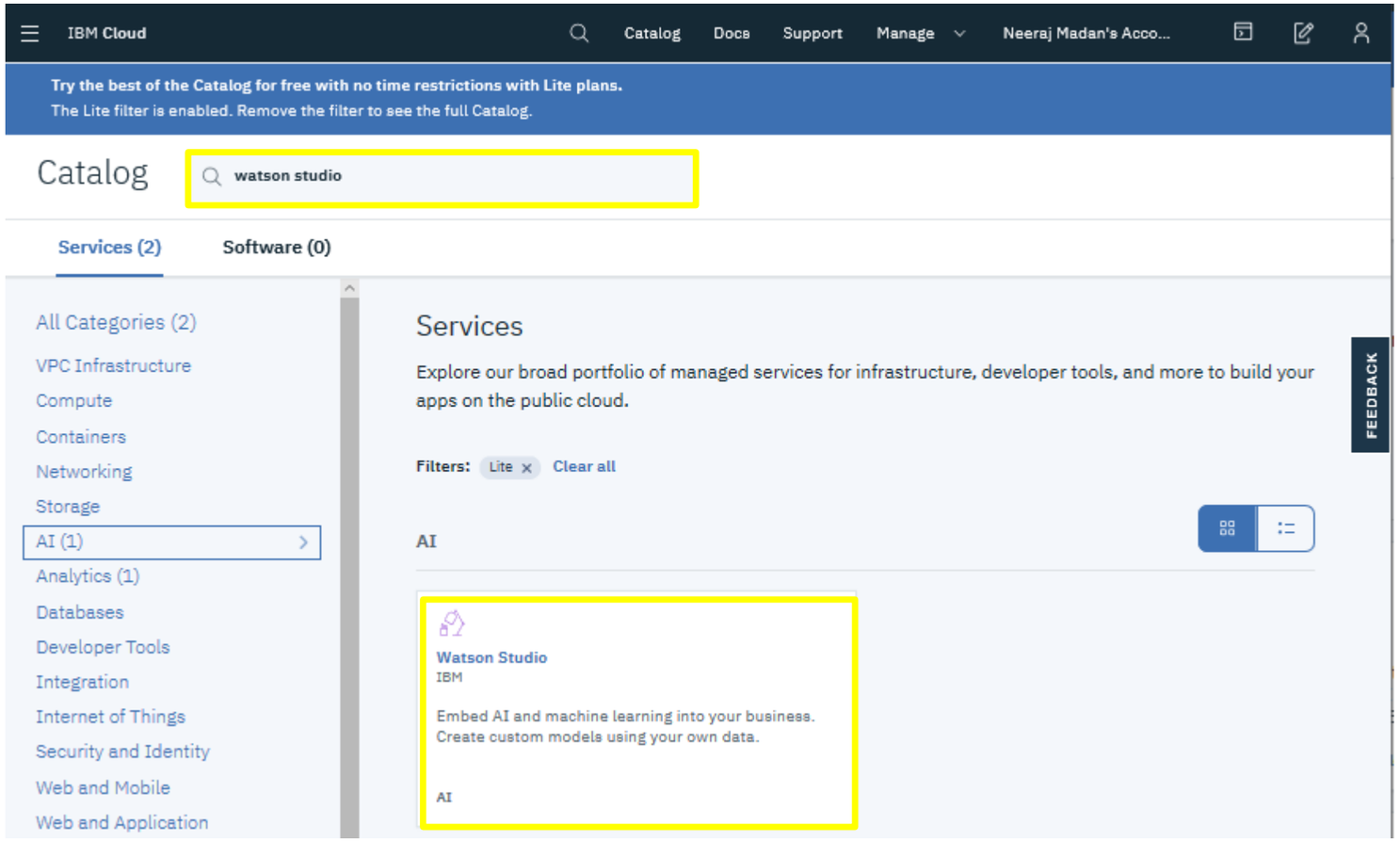
Getting Started: IBM Watson Studio Setup
(Step 3: Welcome to IBM Cloud)





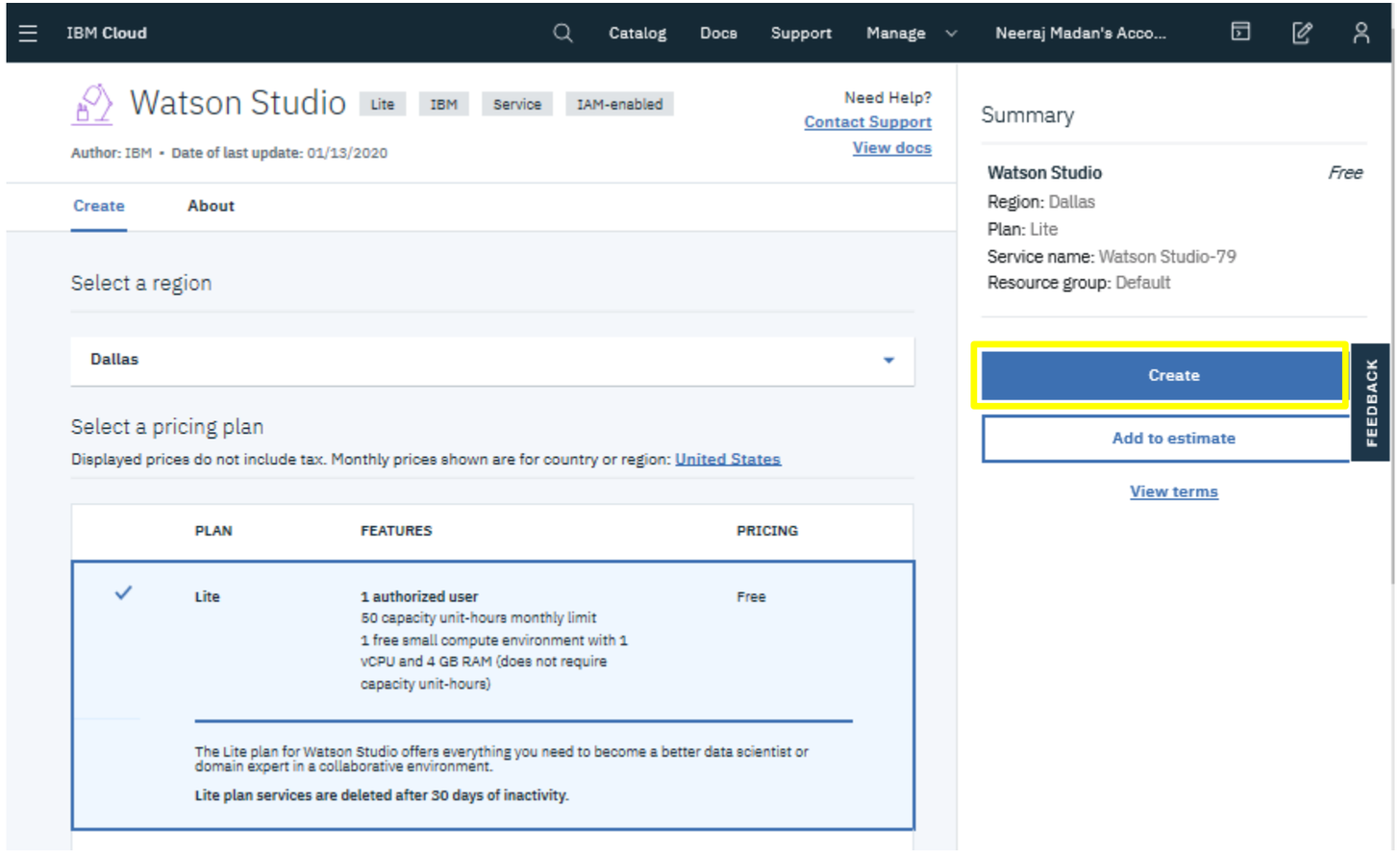
[IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog]

Getting Started: IBM Watson Studio Setup
(Step 4: Locate Watson Studio in the Catalog)



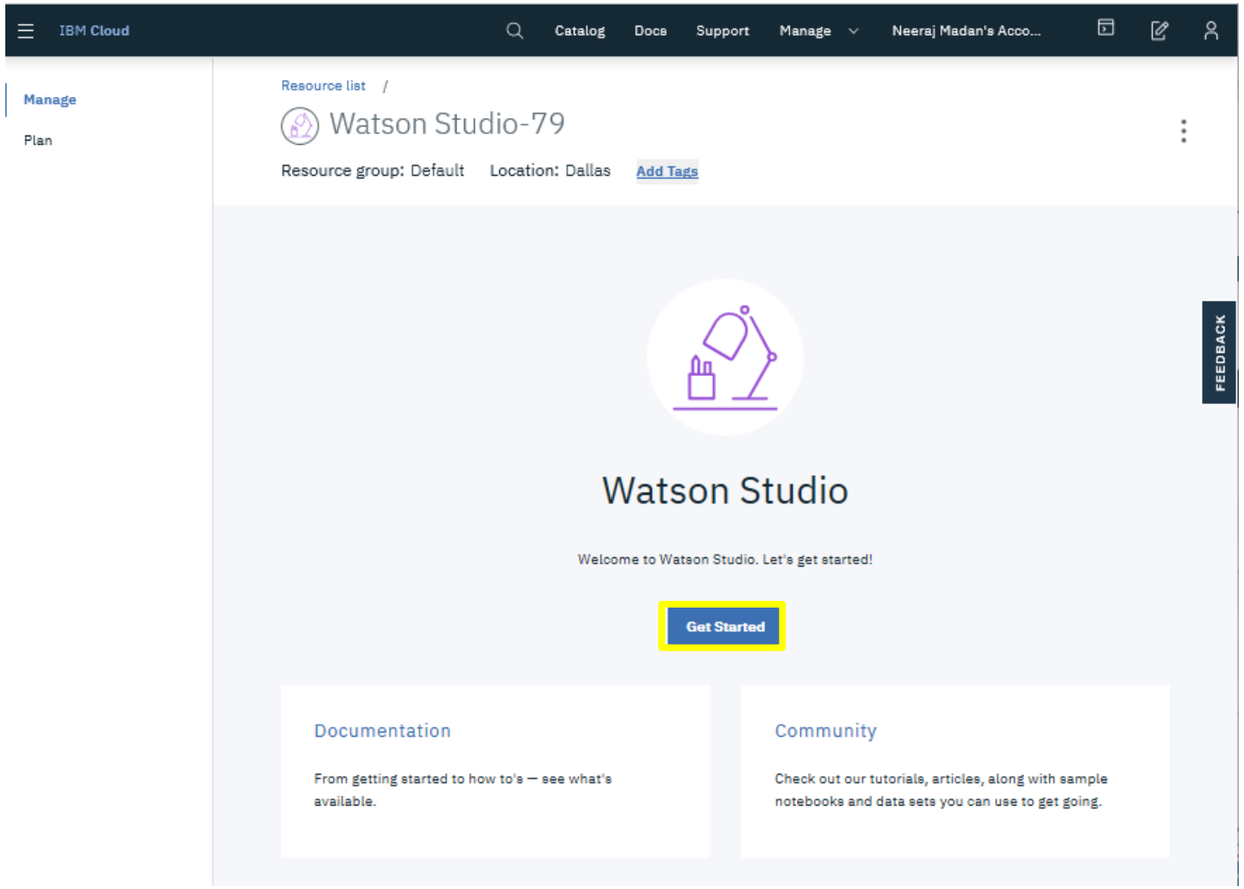
[IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog]

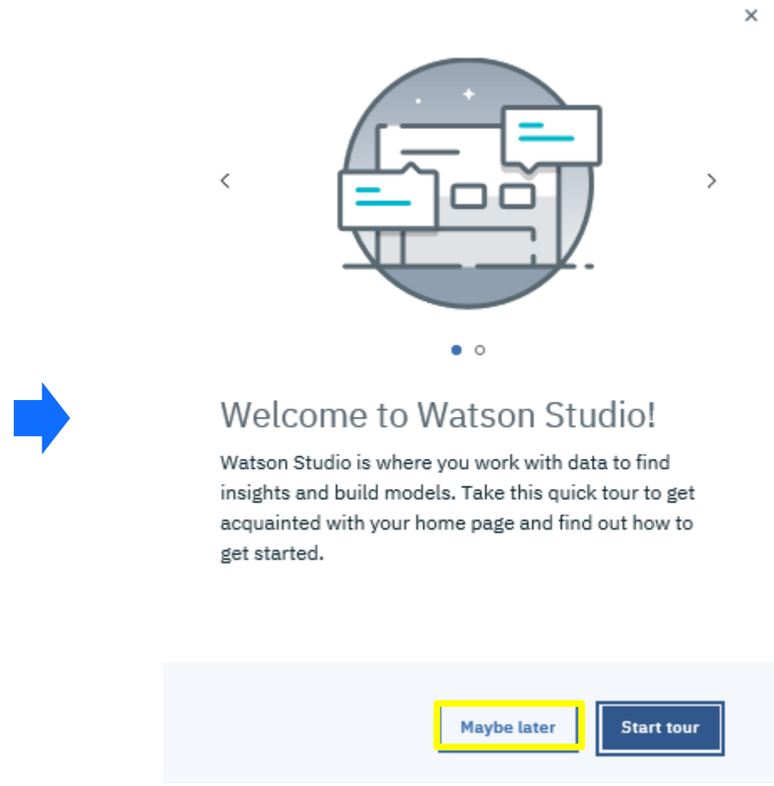
Getting Started: IBM Watson Studio Setup
(Step 5: Create a Watson Studio Service)



IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

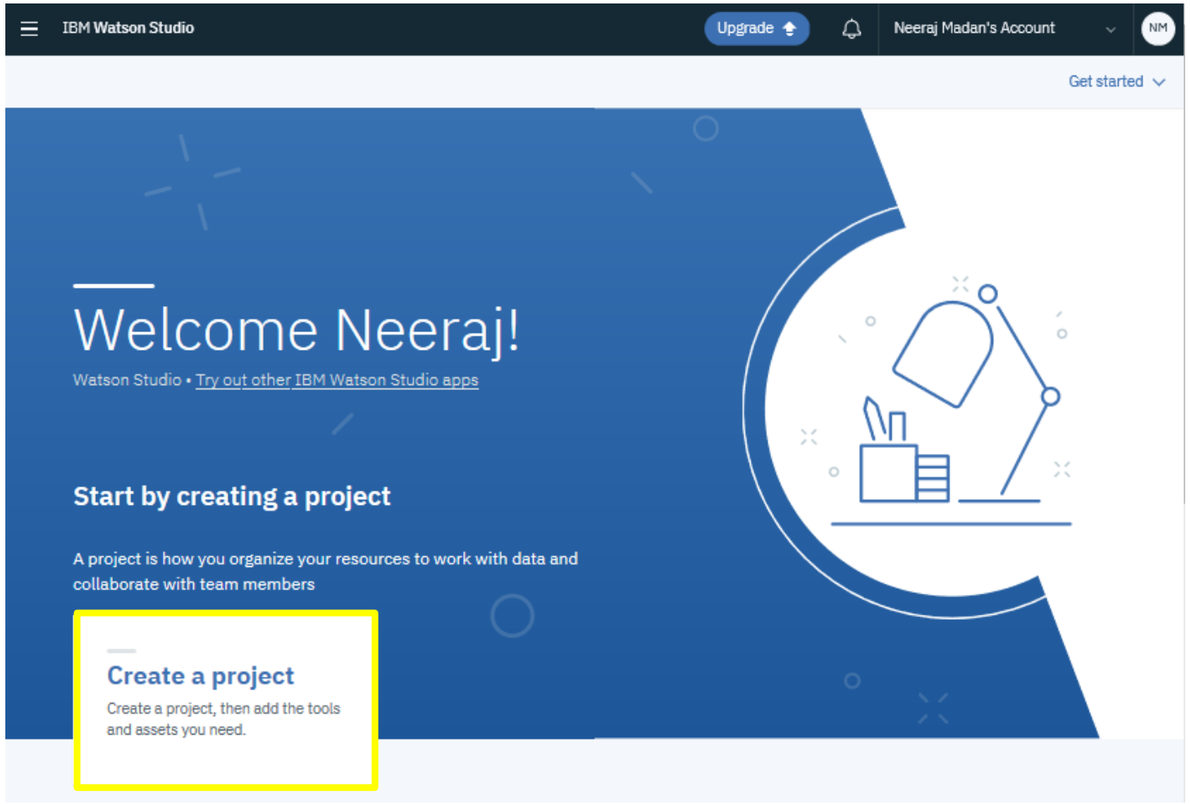
Getting Started: IBM Watson Studio Setup
(Step 6: Launch Watson Studio)

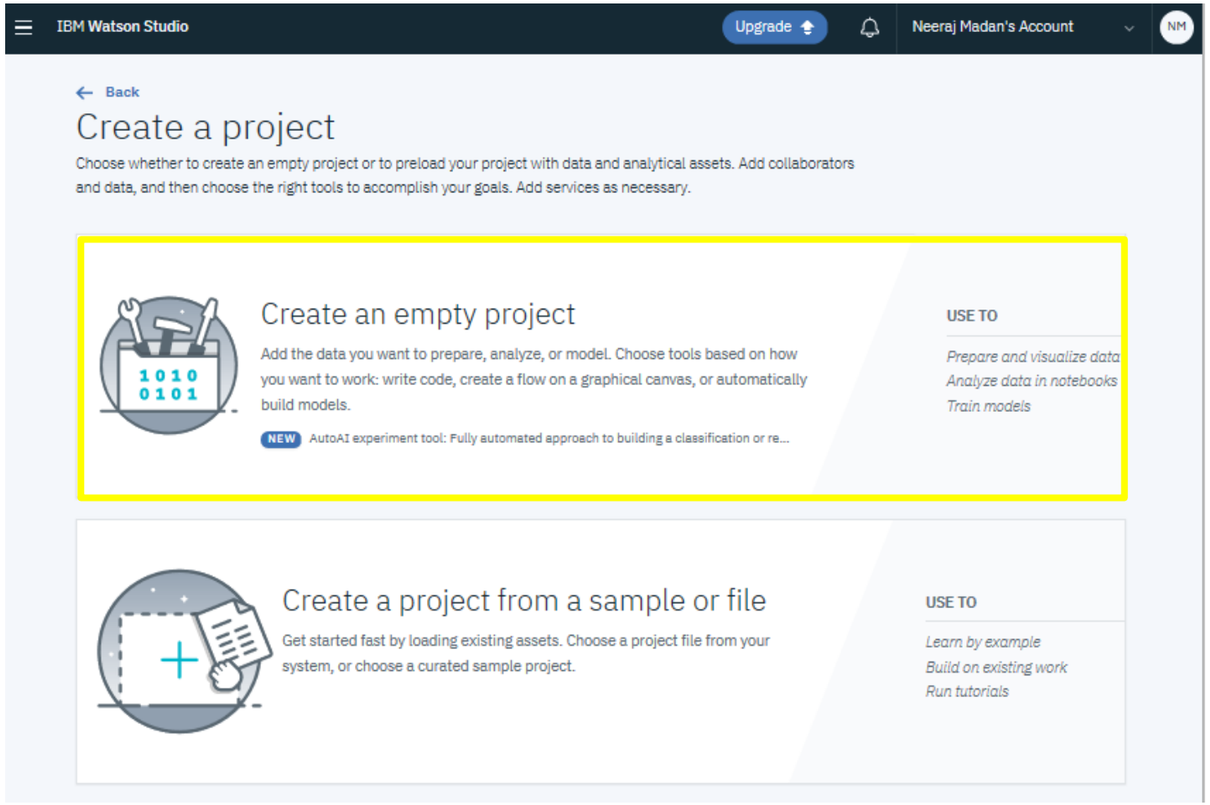




[IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog]

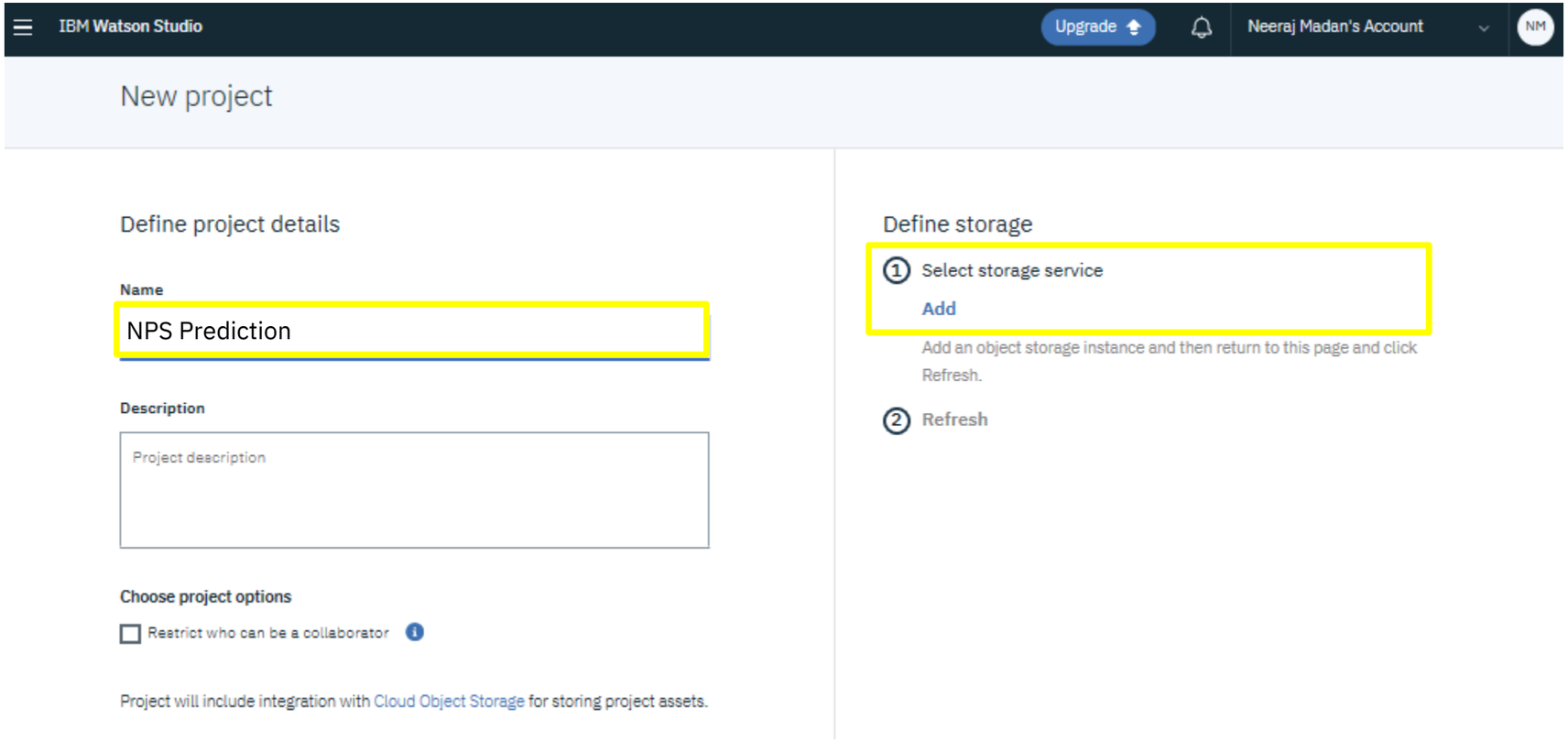
Getting Started: IBM Watson Studio Setup
(Step 7: Create a project and pick an empty template)





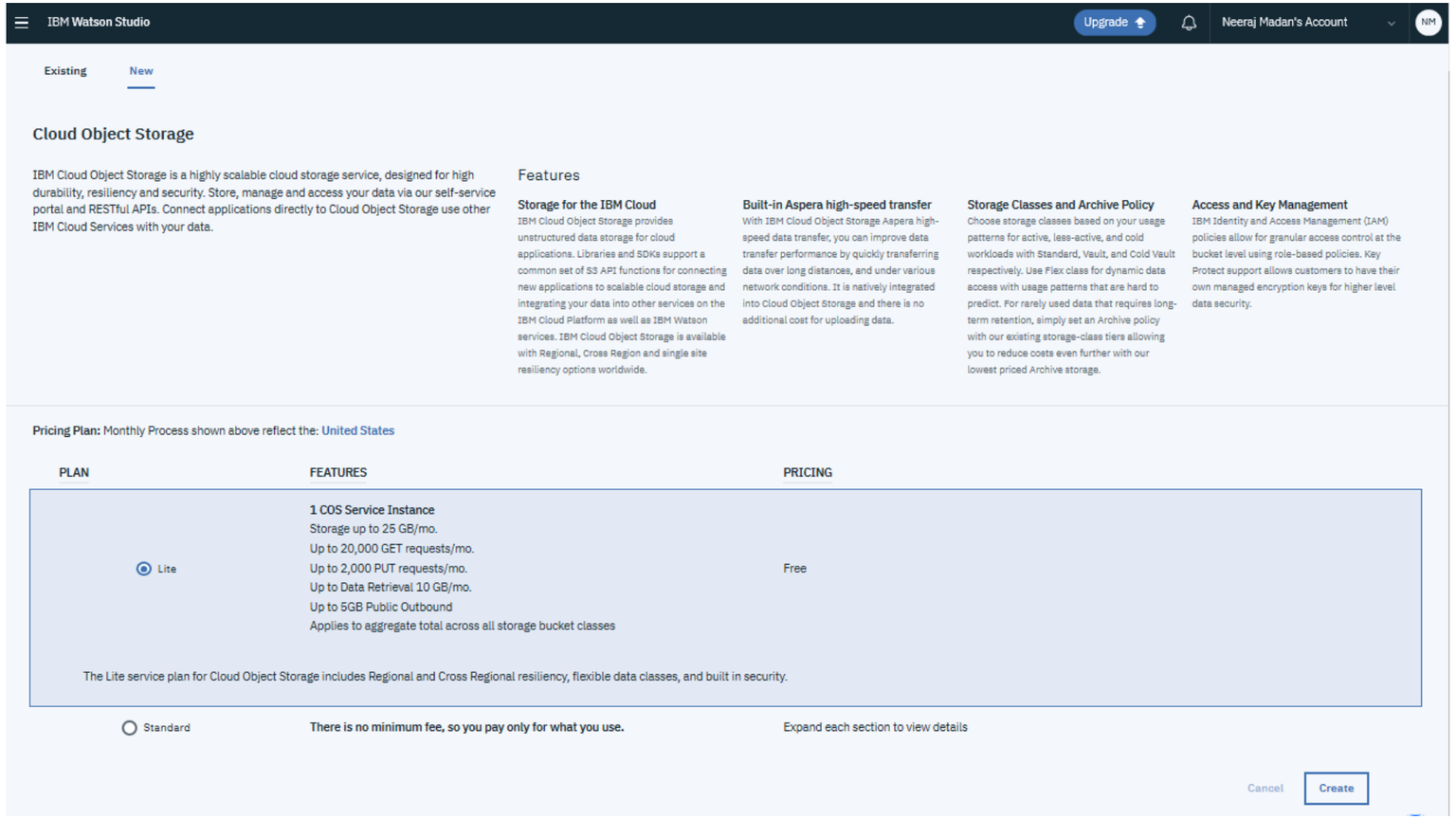
IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

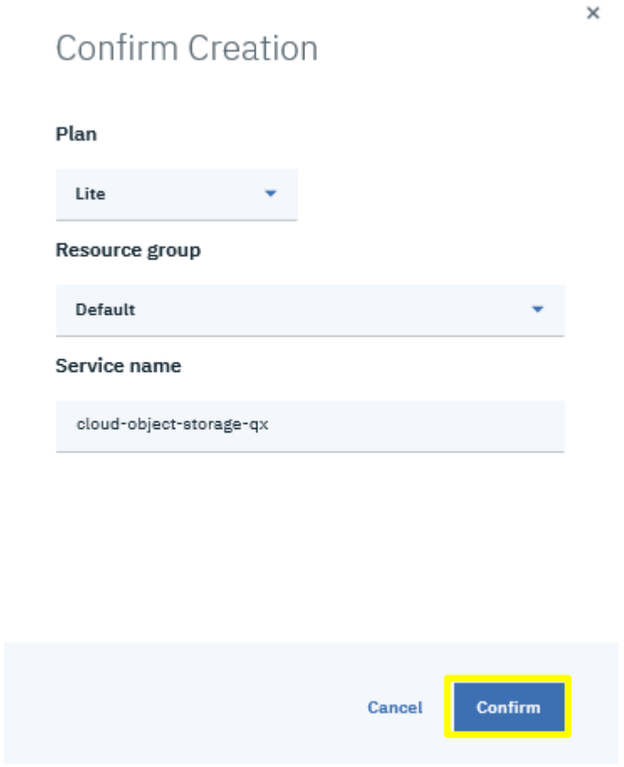
Getting Started: IBM Watson Studio Setup
(Step 8a: Project and Storage Setup)



IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

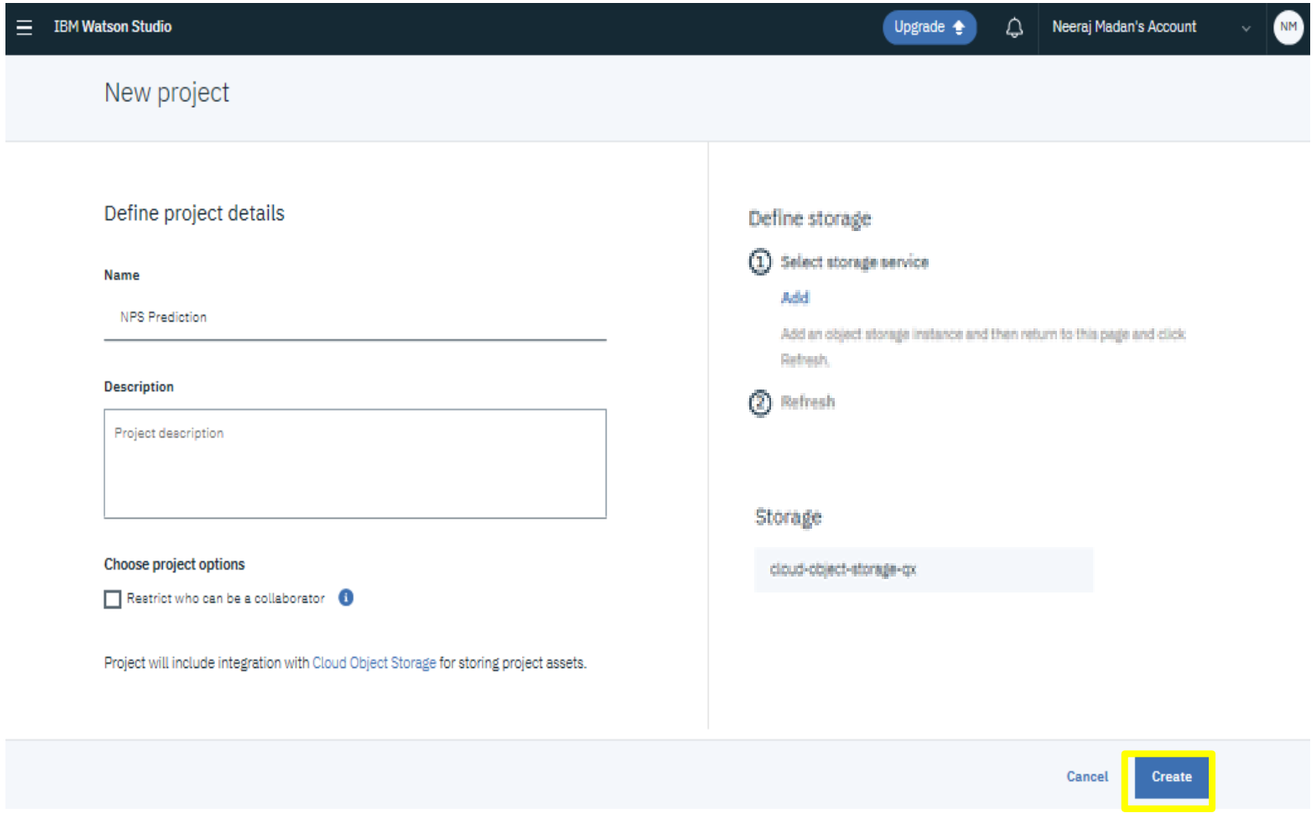
Getting Started: IBM Watson Studio Setup
(Step 8b: Project and Storage Setup)

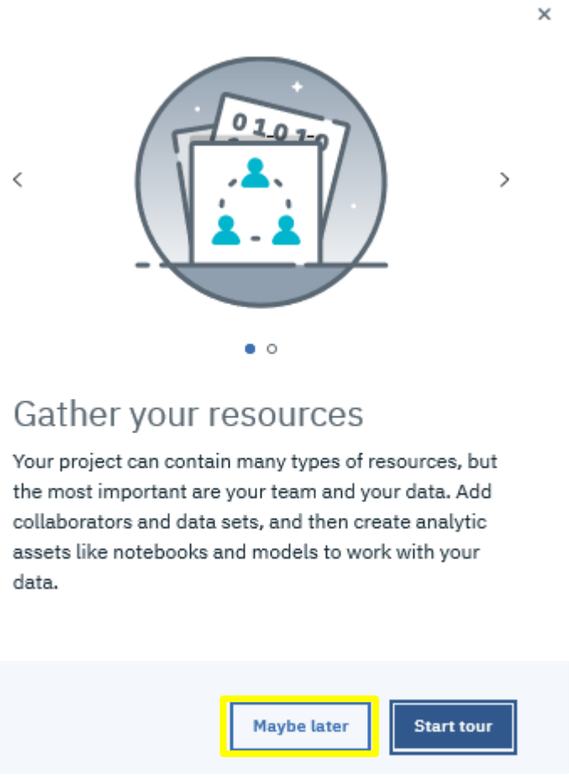




[IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog]

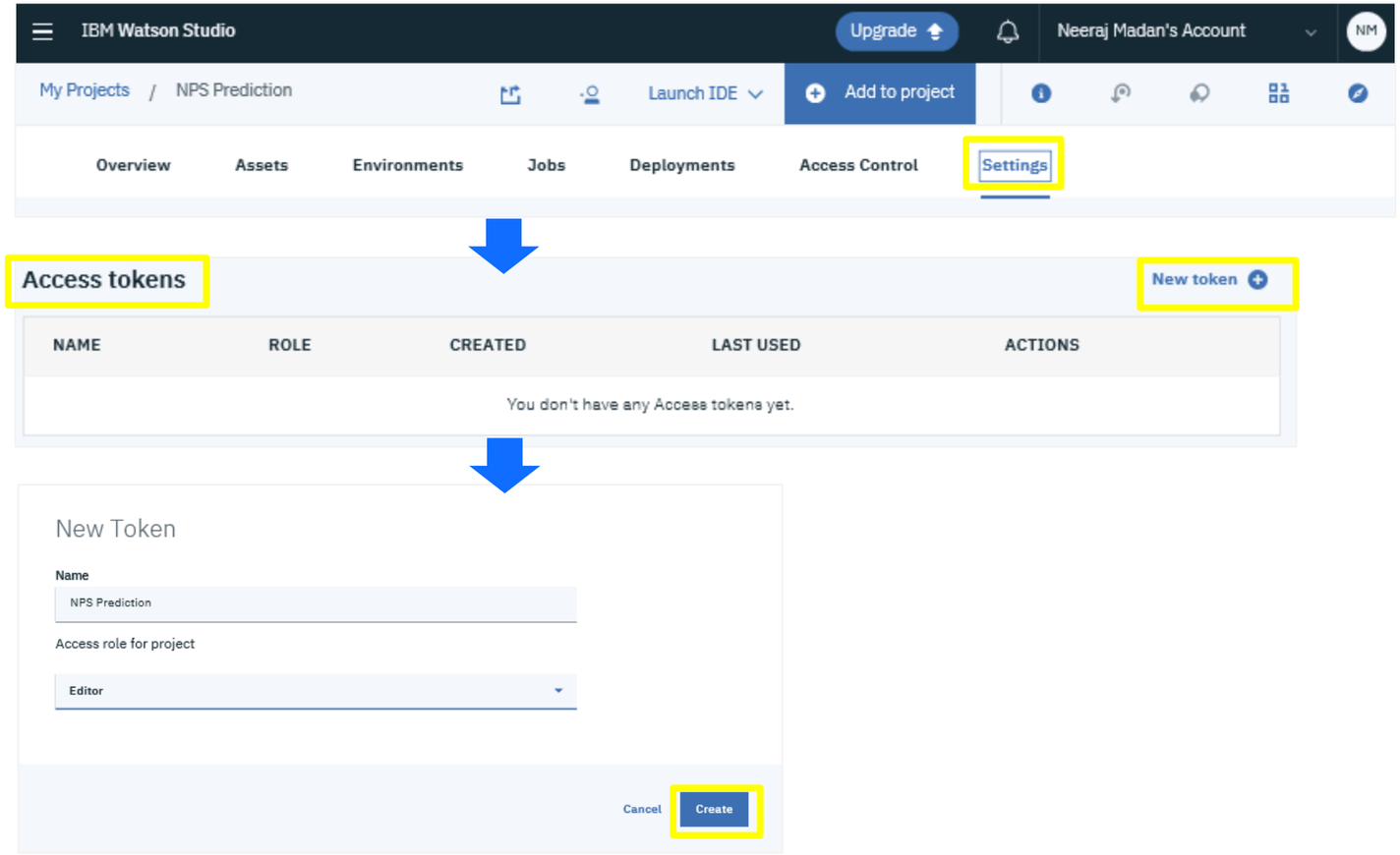
Getting Started: IBM Watson Studio Setup
(Step 8c: Project and Storage Setup)





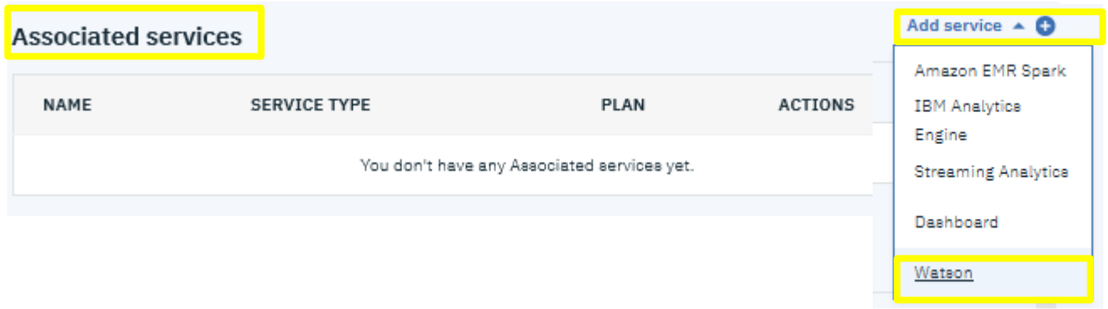
[IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog]

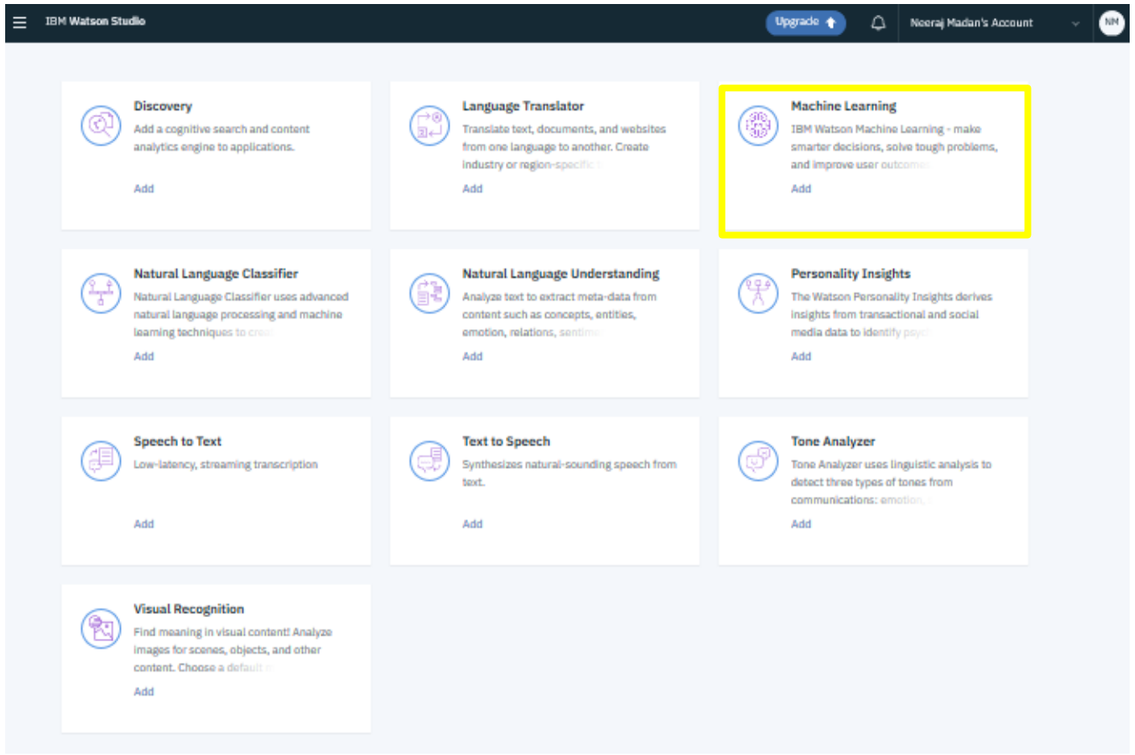
Getting Started: IBM Watson Studio Setup
(Step 9a: Create a project token)



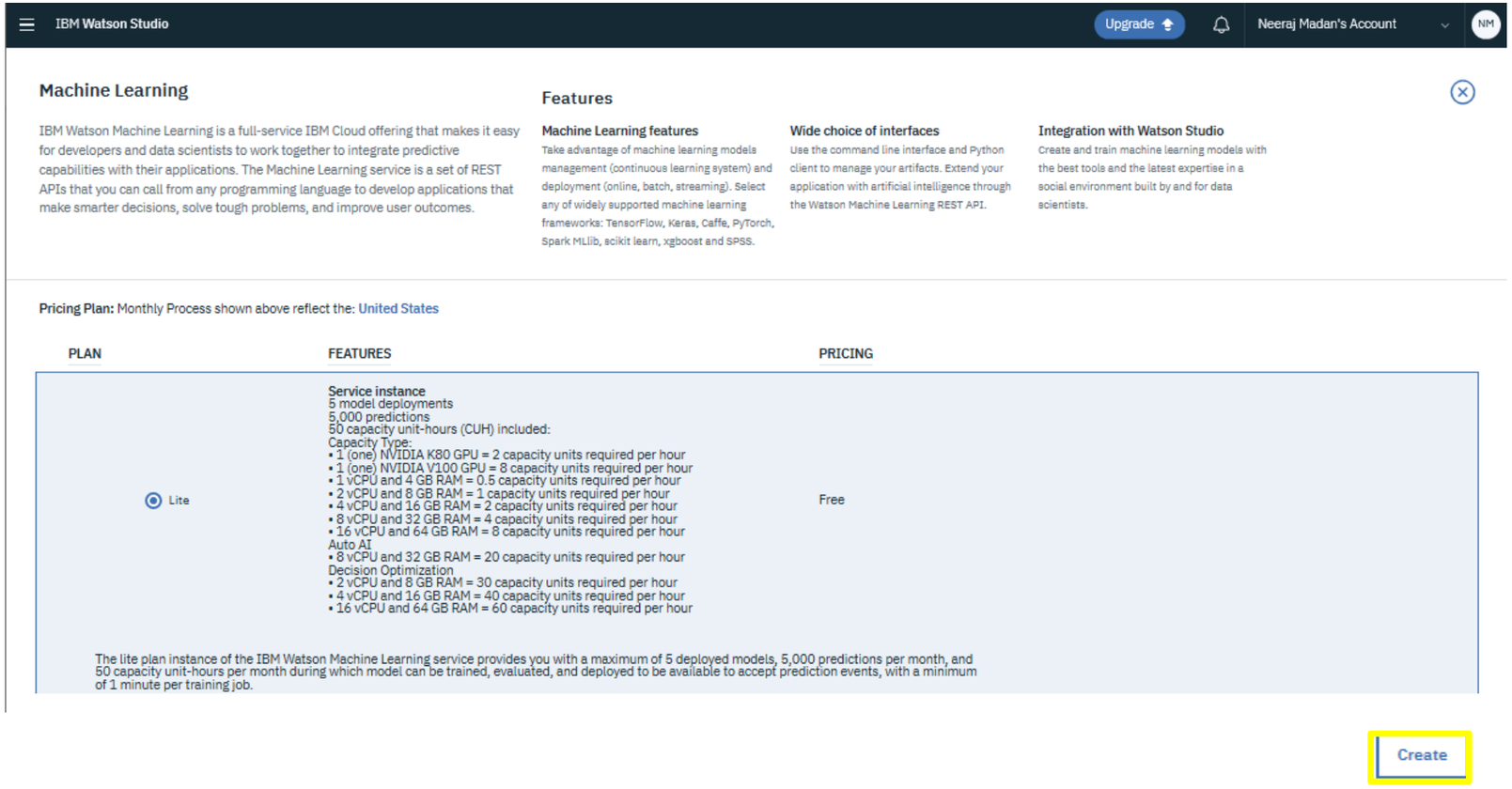
IBM Watson: Sign up for Watson Studio and Watson Knowledge Catalog

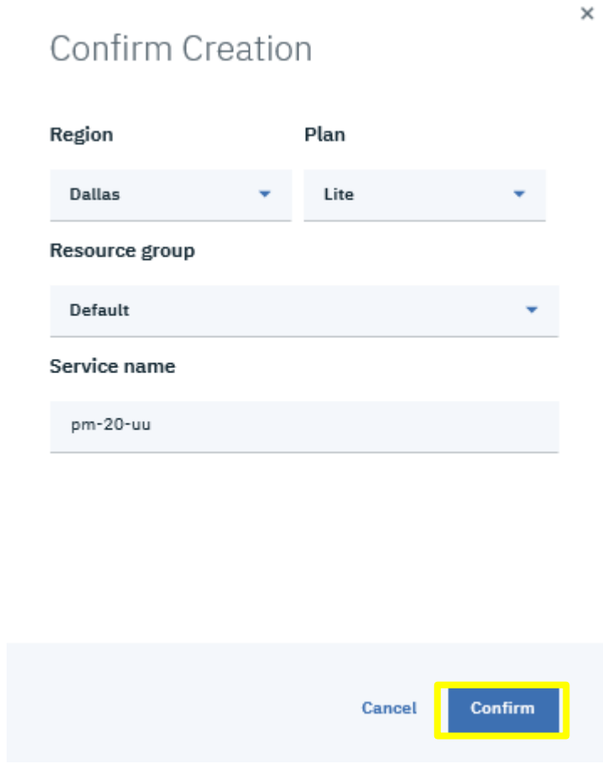
Getting Started: IBM Watson Machine Learning Setup
(Step 9b: Add Watson Machine Learning Service)



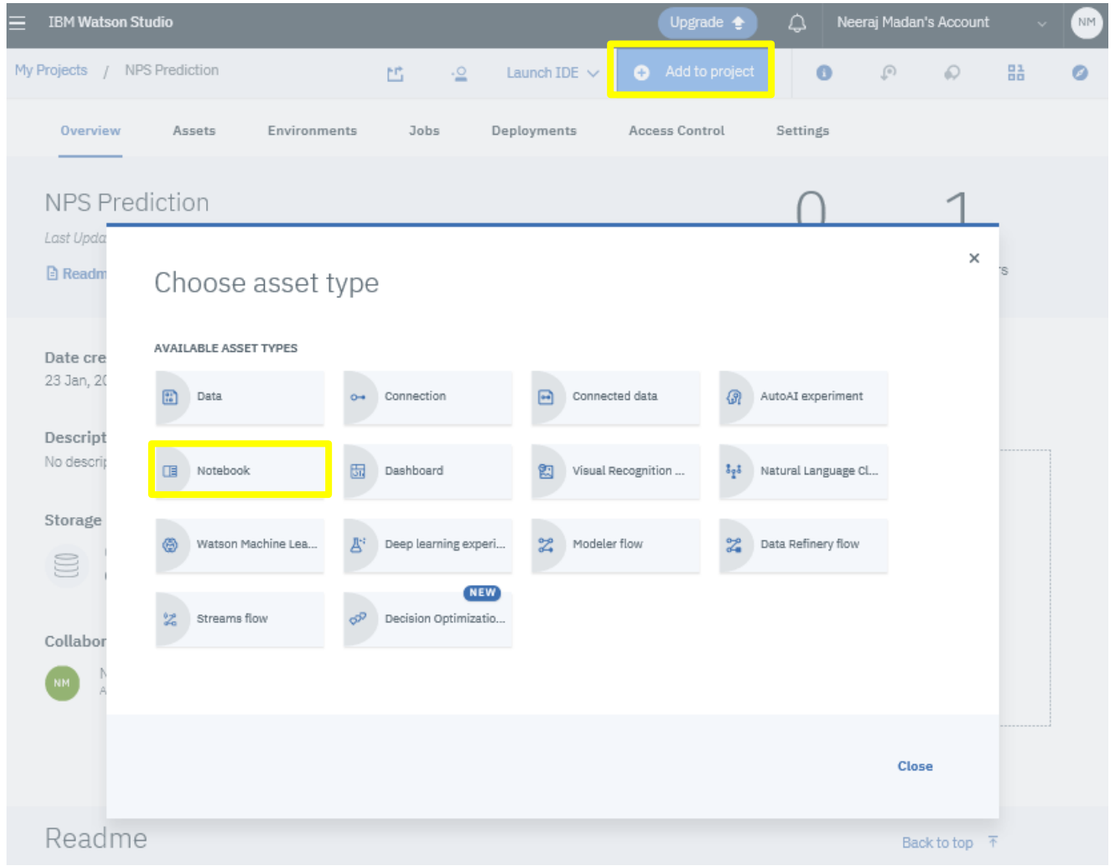


Getting Started: IBM Watson Machine Learning Setup
(Step 9c: Add Watson Machine Learning Service)

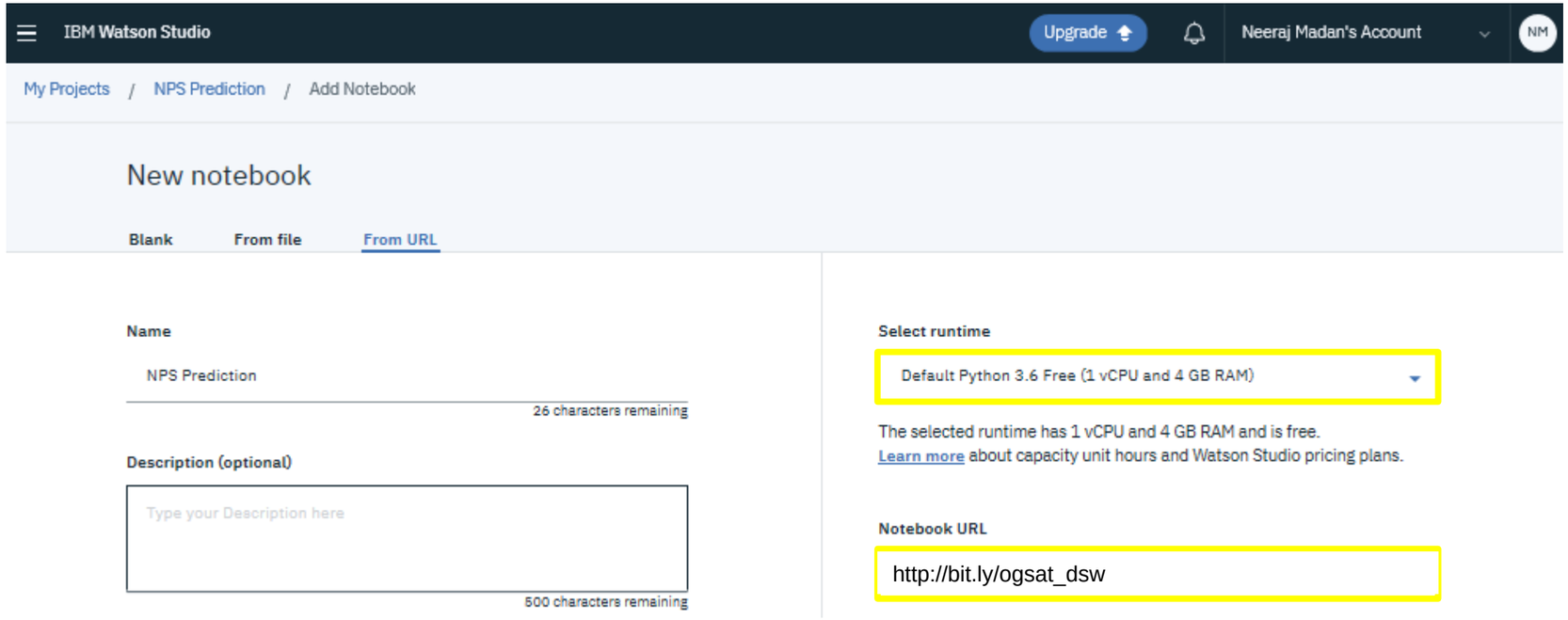




Getting Started: Load Notebook
(Step 10a: Add a notebook to your project)

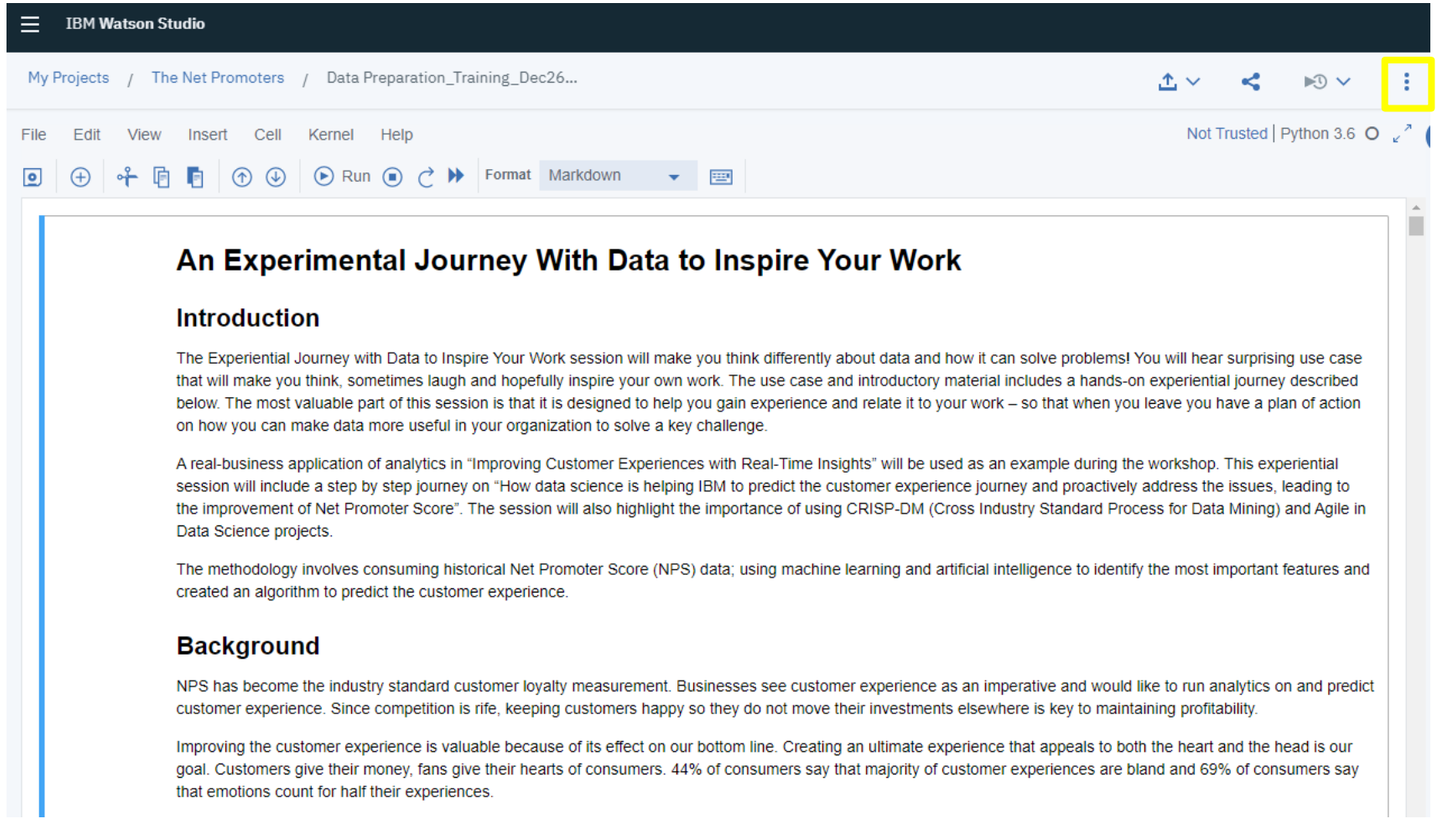


Getting Started: Load Notebook
(Step 10b: Import a notebook, get the link from github)

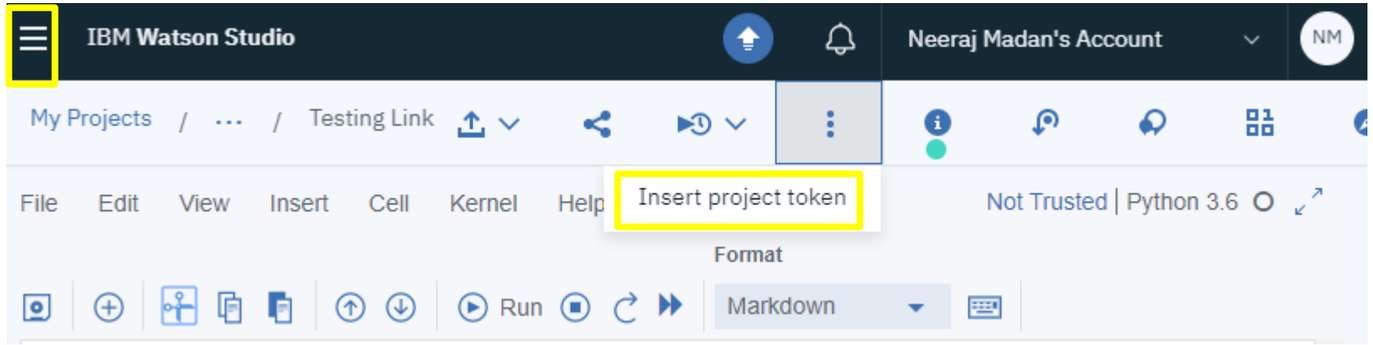


Link: http://bit.ly/ogsat_dsw
GitHub Link: http://bit.ly/OGSAT2020_GitHub

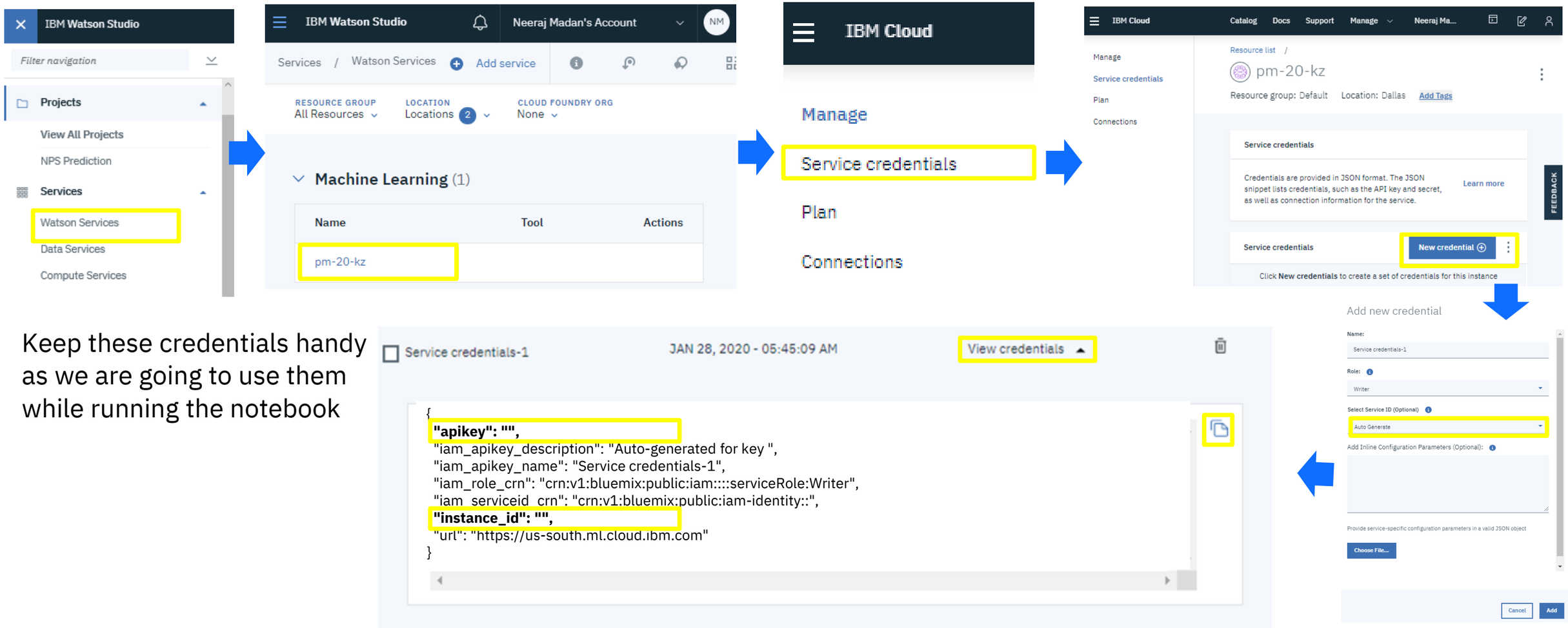
Getting Started: Load Notebook
(Step 10c: Let's get started!)



Getting Started: Load Notebook
(11. Update your credentials in the Notebook)



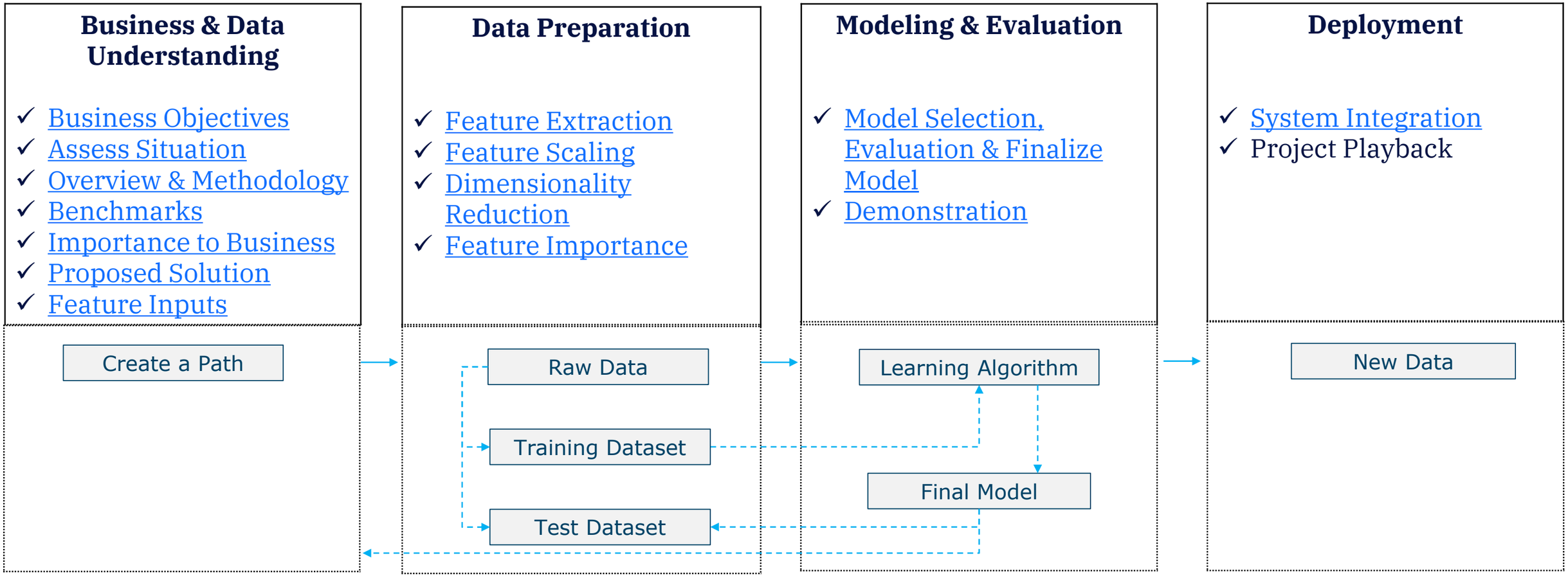
Getting Started: Load Notebook
(12. Machine Learning Credentials)



Keep these credentials handy as we are going to use them while running the notebook

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.



Source: A roadmap for building machine learning systems from Python Machine Learning: Perform Python Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow by Sebastian Raschka; Vahid Mirjalili and CRISP-DM

Take away

Now, I am able to

- ✓ create/ setup the Data Science environment on IBM Cloud
- ✓ learn/ re-use the roadmap to build a machine learning system

Business Objectives

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

Methodology: Consumed historical NPS data; Used machine learning and artificial intelligence to identify the most important features and created an algorithm to predict the non promoters.

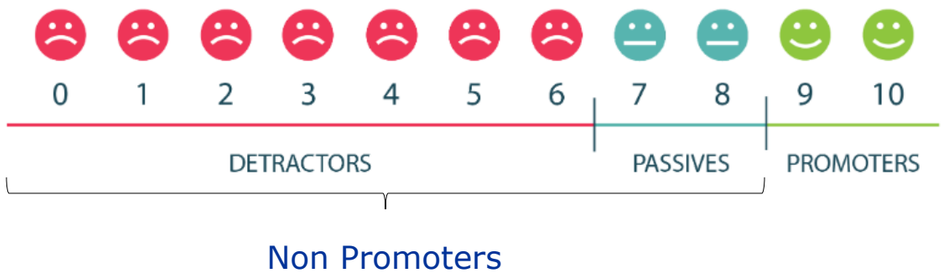
Desired Result: To create an insert into production environment (ticketing system) to indicate top candidates for non-promoter surveys.

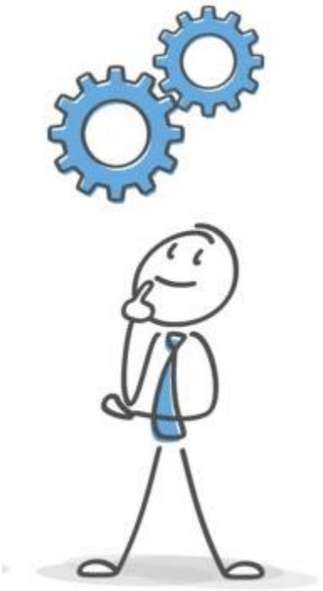
Assess Situation

In year 2018, IBM 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.





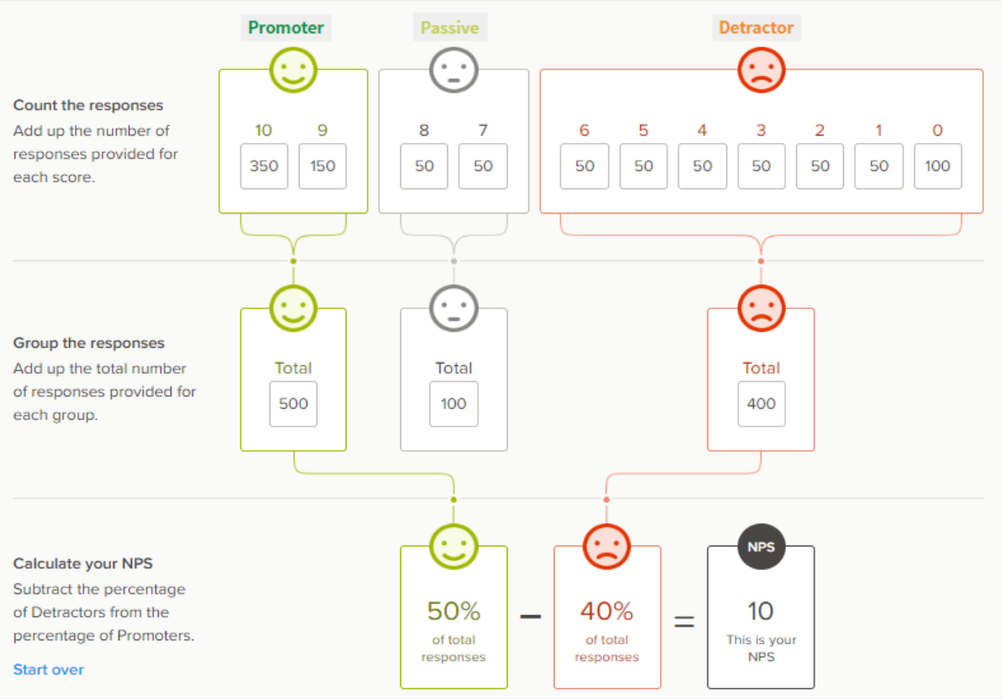
Note: The numbers highlighted above are crafted for this workshop.

Overview and 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?

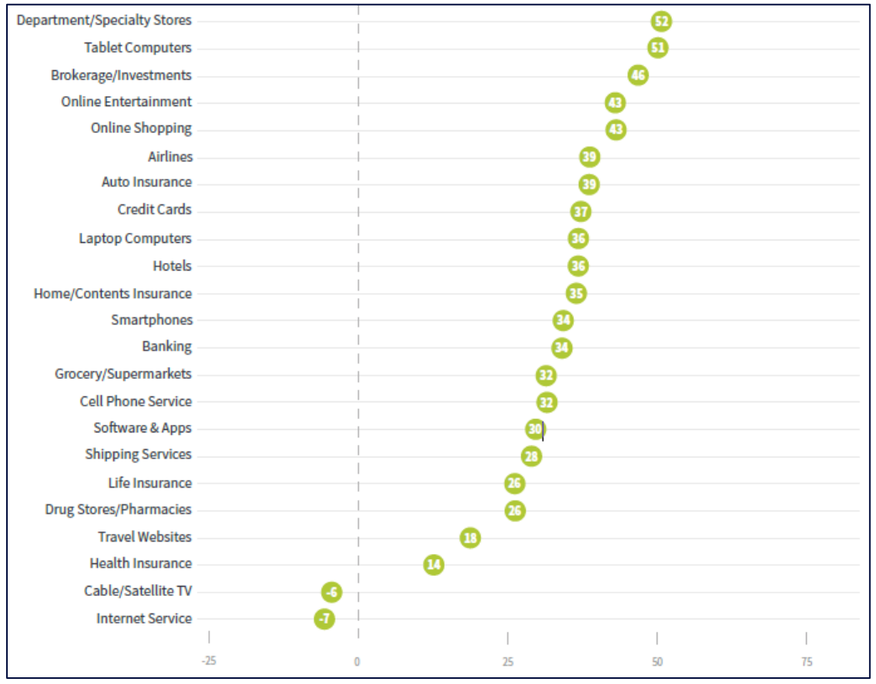
Calculating NPS score is as simple as tallying up your responses and subtracting the percentage of detractors from the percentage of promoters. The score is a whole number that ranges from -100 to 100, and indicates customer happiness with our brand experience.



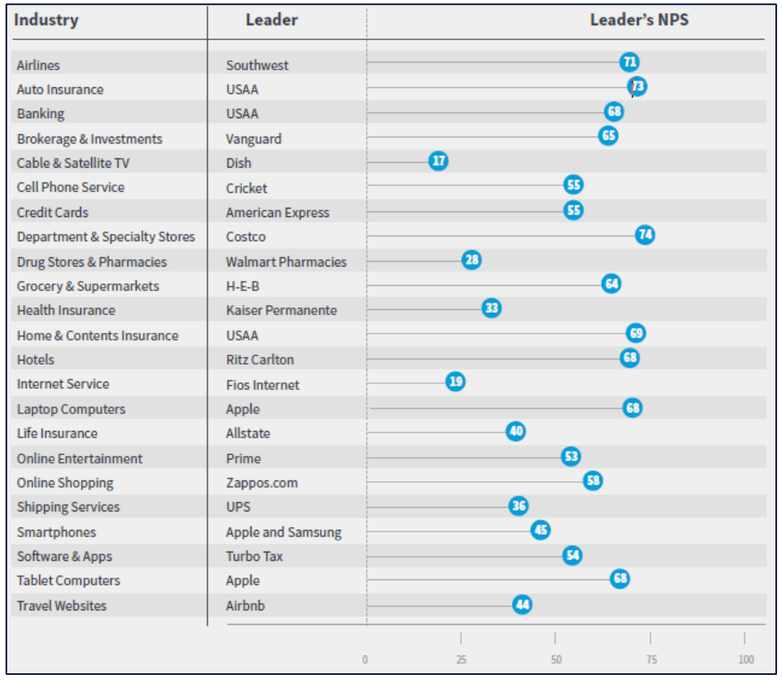
Source: How to Use Net Promoter to Drive Business Growth, Satmetrix and NPS Calculator

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

Importance to Business

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.

91% of marketing leaders believe that in two years companies will be competing primarily on the basis of the customer experience (Gartner)

2% increase in customer retention has the same effect as decreasing costs by 10%

5X Acquiring new customers can cost as much as 5X more than satisfying and retaining current customers

44% of consumers say that majority of customer experiences are bland

69% of consumers say that emotions count for over half their experiences

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

Exercise 1

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

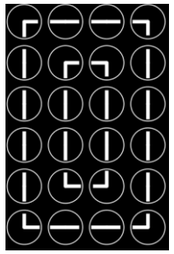
Take away

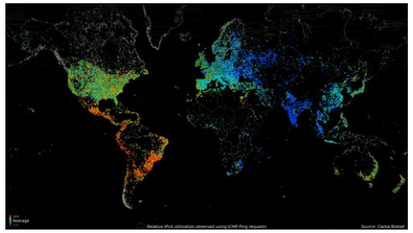
Now, I am able to

- ✓ setup the Data Science environment on IBM Cloud
- ✓ re-use the roadmap to build a machine learning system
- ✓ use a framework to assess the situation, define business objectives, and create a baseline

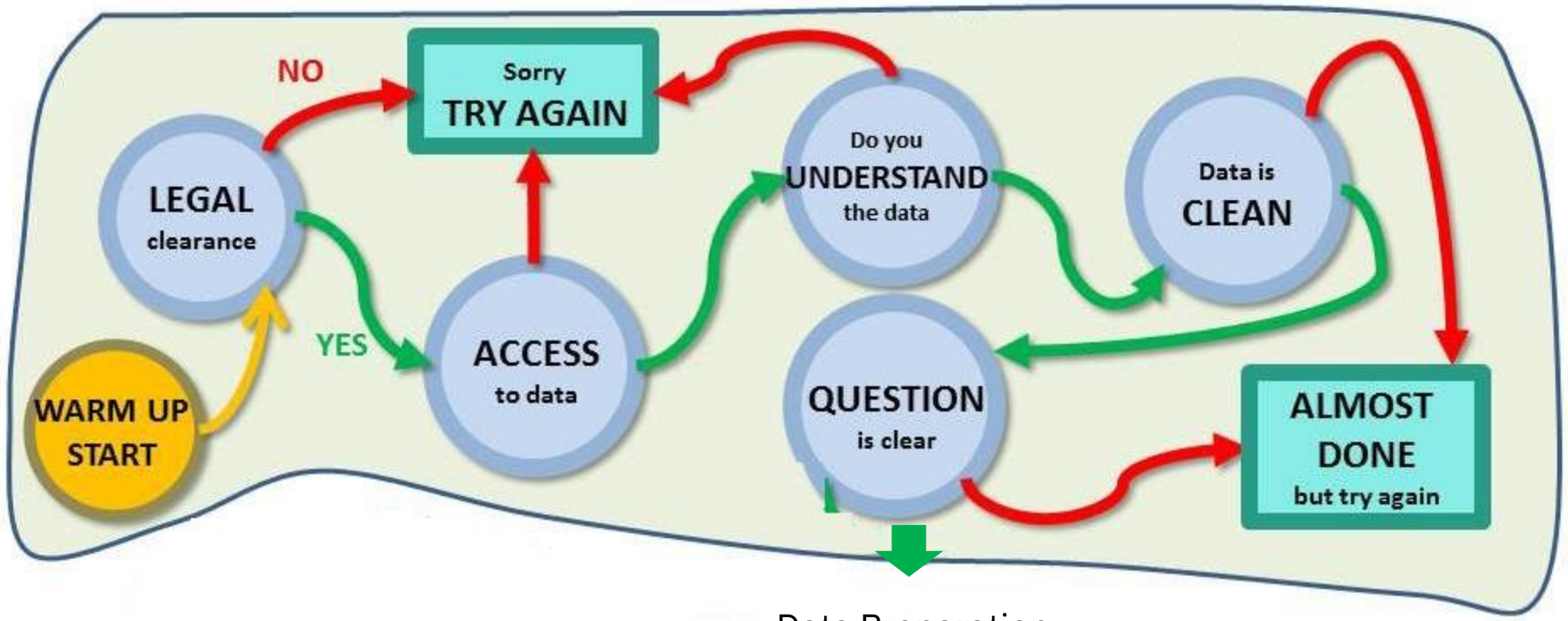
Feature Inputs

Time Based	Geography	Money	Sentiment/ Emotions





Easy-Peasy



Data Preparation

An Extended Version Of The Scikit-Learn Cheat Sheet

Exercise 2

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

Lab 1

Run the following section in the notebook.

1. Load packages, libraries and verify the version
2. Explore the data and perform quality audit

Take away

Now, I am able to

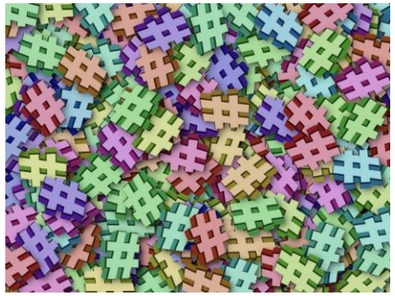
- ✓ setup the Data Science environment on IBM Cloud
- ✓ re-use the roadmap to build a machine learning system
- ✓ use a framework to assess the situation, define business objectives, and create a baseline
- ✓ load the packages, libraries, and verify the version.
- ✓ explore the data set and perform quality audit

Extract, Scale, and Reduce Dimensions

Feature Extraction

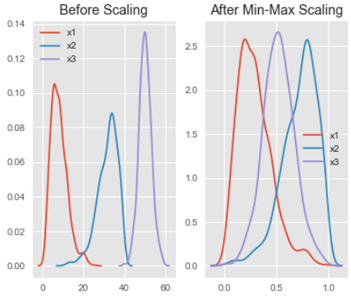
One Hot Encoding

Hashing Encoder



Feature Scaling

MinMax Scaler



Dimensionality Reduction

Percent Missing Value

Amount of Variation

The Hitchhiker's Guide to Feature Extraction, Feature Scaling with scikit-learn by Ken Keen, and A Practical Guide to Dimensionality Reduction Techniques

Exercise 3

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

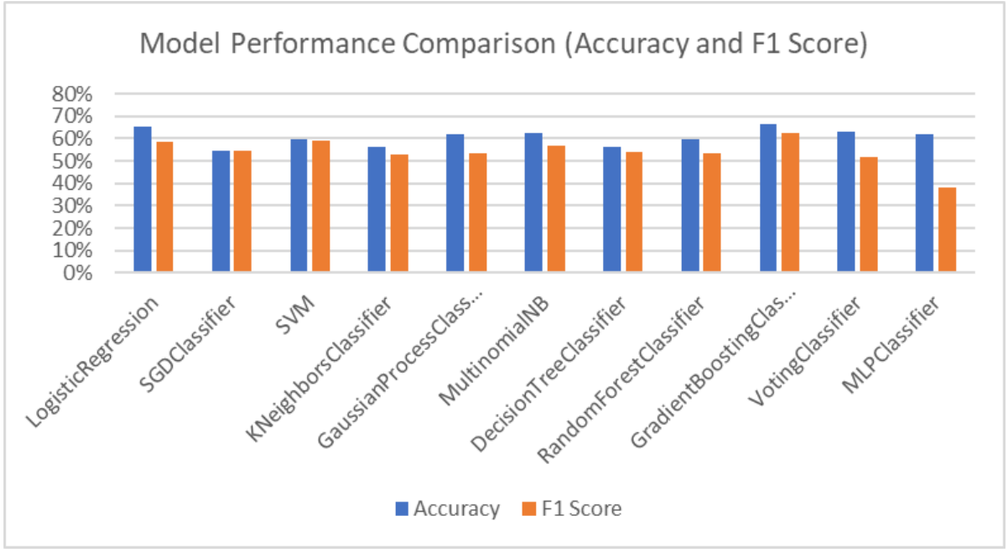
Take away

Now, I am able to

- ✓ setup the Data Science environment on IBM Cloud
- ✓ re-use the roadmap to build a machine learning system
- ✓ use a framework to assess the situation, define business objectives, and create a baseline
- ✓ load the packages, libraries, and verify the version.
- ✓ explore the data set and perform quality audit
- ✓ prepare the data for modelling, extract the features, scale the data, and reduce the dimensions

Model Selection and 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 2

Run the following section in the notebook.

7. Split data into train and test sets

8. Model Selection

9. Performance Metric

10. Evaluation

Exercise 4 & 5

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

Take away

Now, I am able to

- ✓ setup the Data Science environment on IBM Cloud
- ✓ re-use the roadmap to build a machine learning system
- ✓ use a framework to assess the situation, define business objectives, and create a baseline
- ✓ load the packages, libraries, and verify the version.
- ✓ explore the data set and perform quality audit
- ✓ prepare the data for modelling, extract the features, scale the data, and reduce the dimensions
- ✓ Split data into train & test sets, select model, and evaluate performance metrics

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 includes,

1. Watson NLP to create additional features from customer conversation logs

2. Machine Learning algorithms to come up with the predictions

Data Collection

Aggregation across multiple sources (ServiceNow, IMS, & Medallia)

Scoring

Batch/ real-time scoring on open support tickets to predict probability of being a non-promoter

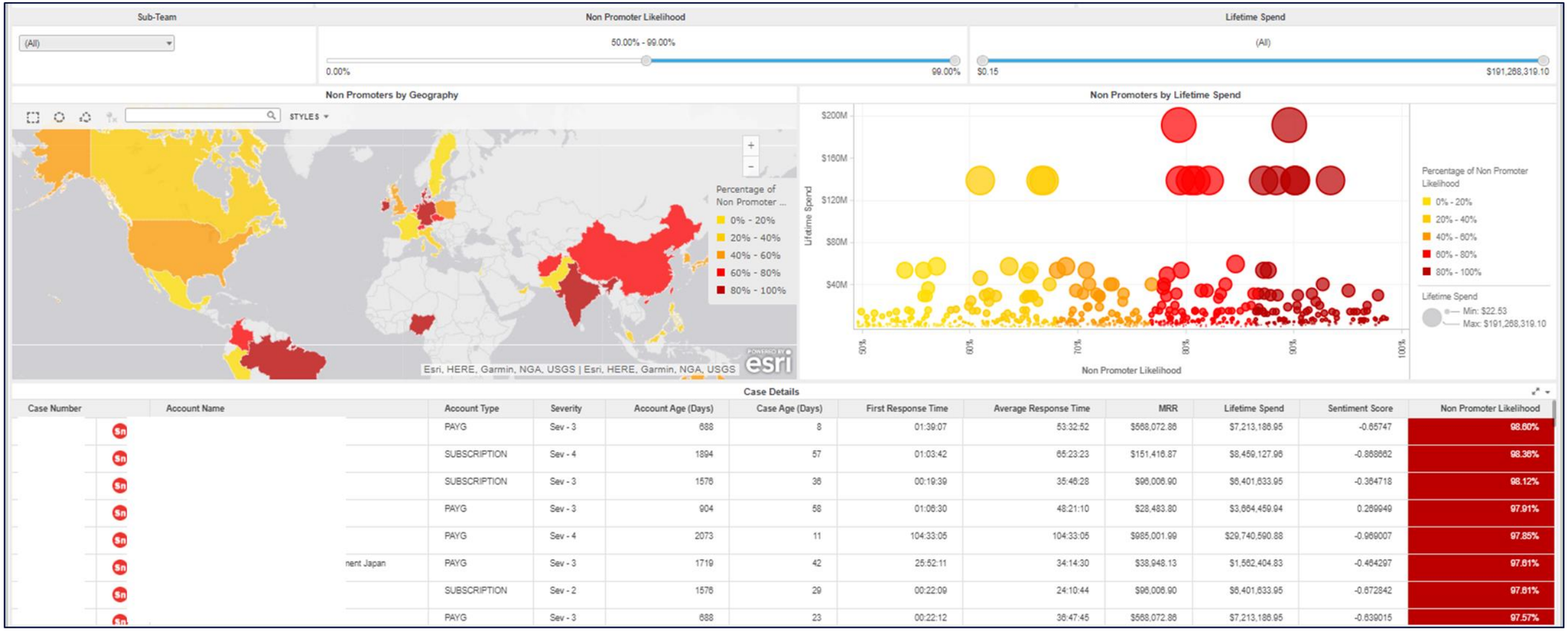
Workflow Integration

Embedding predictions into support team's Interactive Dashboard

Solution Integration

"Ideas are easy, Execution is everything." John Doerr

NPS Predictions Board



Case Number	Account Name	Account Type	Severity	Account Age (Days)	Case Age (Days)	First Response Time	Average Response Time	MRR	Lifetime Spend	Sentiment Score	Non Promoter Likelihood
		PAYG	Sev - 3	688	8	01:39:07	53:32:52	$568,072.86	$7,213,186.95	-0.65747	98.60%
		SUBSCRIPTION	Sev - 4	1894	57	01:03:42	65:23:23	$151,416.87	$8,459,127.96	-0.868662	98.36%
		SUBSCRIPTION	Sev - 3	1576	36	00:19:39	35:46:28	$96,006.90	$6,401,633.95	-0.384718	98.12%
		PAYG	Sev - 3	904	58	01:06:30	48:21:10	$28,483.80	$3,864,459.94	0.269949	97.91%
		PAYG	Sev - 4	2073	11	104:33:05	104:33:05	$985,001.99	$29,740,590.88	-0.969007	97.85%
	nent Japan	PAYG	Sev - 3	1719	42	25:52:11	34:14:30	$38,948.13	$1,562,404.83	-0.464297	97.61%
		SUBSCRIPTION	Sev - 2	1576	29	00:22:09	24:10:44	$96,006.90	$6,401,633.95	-0.672842	97.61%
		PAYG	Sev - 3	688	23	00:22:12	36:47:45	$568,072.86	$7,213,186.95	-0.639015	97.57%

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

Lab 3

Run the following section in the notebook.

11. Deployment

Exercise 6

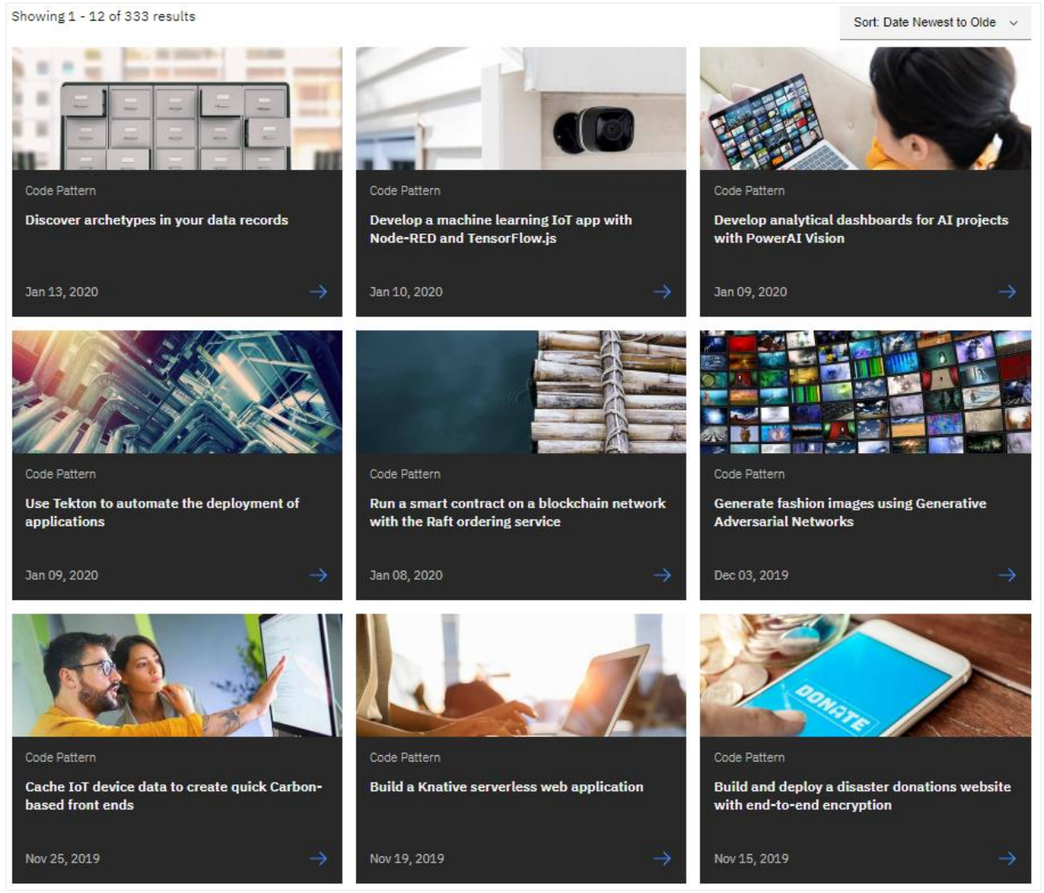
How do you plan to consume the outputs of model?

Take away

Now, I am able to

- ✓ setup the Data Science environment on IBM Cloud
- ✓ re-use the roadmap to build a machine learning system
- ✓ use a framework to assess the situation, define business objectives, and create a baseline
- ✓ load the packages, libraries, and verify the version.
- ✓ explore the data set and perform quality audit
- ✓ prepare the data for modelling, extract the features, scale the data, and reduce the dimensions
- ✓ Split data into train & test sets, select model, and evaluate performance metrics
- ✓ 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

	Name: Maureen Norton Email Id: nortonm@us.ibm.com LinkedIn: https://www.linkedin.com/in/maureennorton/
	Name: George Stark Email Id: gstark@us.ibm.com LinkedIn: https://www.linkedin.com/in/gestark/
	Name: Neeraj Madan Email Id: nmadan@us.ibm.com LinkedIn: https://www.linkedin.com/in/neerajmadan/