BADM 4830 / BAIM 4200 Advanced Business Analytics

Module 5 - MLOps – Development, Production & Monitoring

Ivan Portilla
Jesus.Portilla@Colorado.edu
Portilla@gmail.com
github.com/jiportilla/giveback

IBM

Objectives of This Module

Upon completion of this module, you will understand:

Module 5: MLOPs – How - Weeks 11-12 (11/25/2021 Thanksgiving Day)

1. Developing Models
    a) Agile development
    b) Supervised vs Unsupervised Learning
    c) Tooling Review
    d) ML Python Libraries review, Numpy, scikit-learn, TensorFlow, Pandas
    e) Development Resources (stack overflow, forums, IBM galleries, slack, call4code)
    f) Best Practices
2. Preparing for Production
3. Deploying to Production
4. Monitoring & Feedback loop
5. Model Governance
6. Lab: Telco customer churn

Spring 2021
BADM 4830 / BAIM 4200 Advanced Business Analytics

- This course will give students the language, knowledge, and actionable methods to work alongside technical and non-technical members of your team to create AI solutions.

- Students will explore what it means to design artificial intelligence systems as a team, guided by a clear intent and a focus on people. This course will give you the framework and tools you need to recognize responsible AI design, align your team, and work with data sources to start building AI solutions.

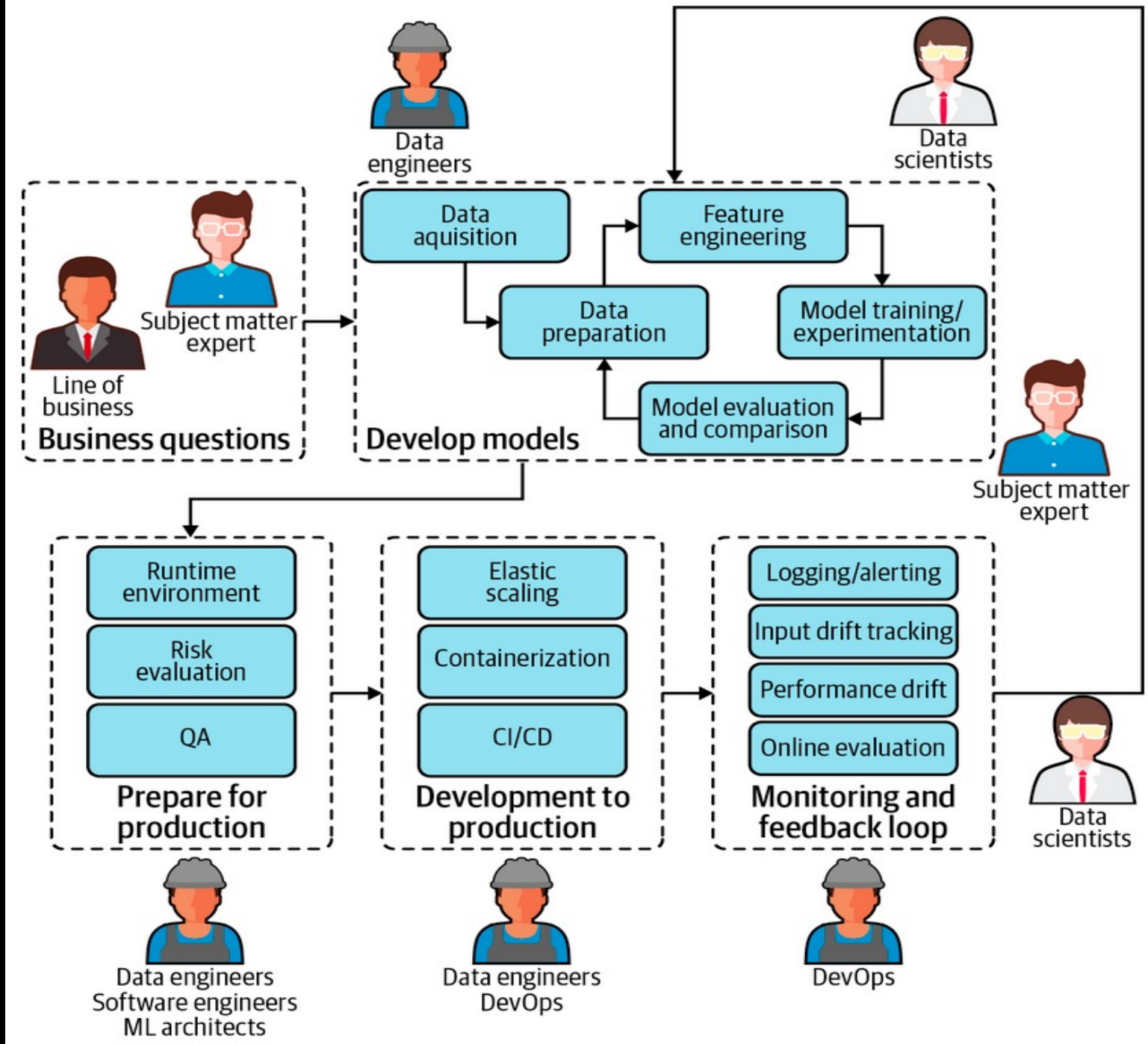
- Students will learn the tools, technology, and practices that enable cross-functional AI teams to efficiently deploy, monitor, retrain, and govern models in production systems.

Re-cap

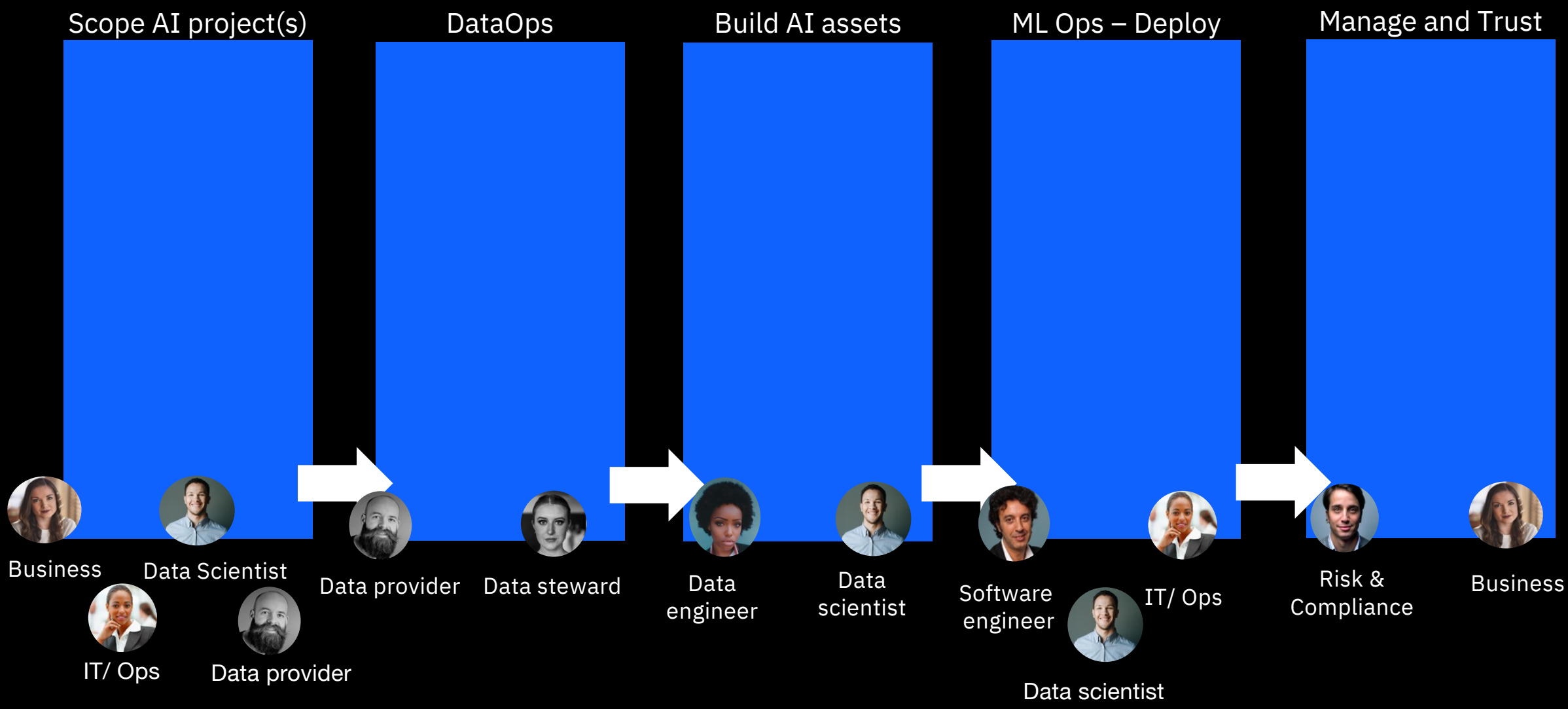
MLOps What & Why

1. MLOps in Action
2. Chatbot AI Lifecycle

Enterprise ML Lifecycle



Stages in Operationalizing AI



Stages in Operationalizing AI

Continuous integration / continuous delivery pipeline

Scope AI project(s)

Explore, prioritize, select use cases – feasibility vs impact

Detail selected use case(s) – KPIs, data, workflow, success criteria

Technical design: tools, infrastructure, data details, approach


Business

Data Scientist


IT/ Ops

Data provider

DataOps

All: Define sources & needs

Provider: Set up catalog for discovery, lineage, curation, access rules

Steward: Create data policies & access

Consumer: Find, understand, add, explore, review, share


Data provider

Data steward

Build AI assets

Connect data

Analyze data

Prepare data

Build models

Train models

Visualize & evaluate models

Test for Bias, Fairness


Data engineer

Data scientist

ML Ops – Deploy

Review, 3rd party oversight, unit tests, validation, approve deployable version

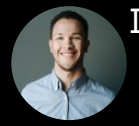
Deploy model to ML runtime engine

Monitor & evaluate model execution

Manage against thresholds

Validation reports


Software engineer



Data scientist

IT/ Ops

Manage and Trust

Configure monitoring
Configure integration with other systems

Monitor and manage:

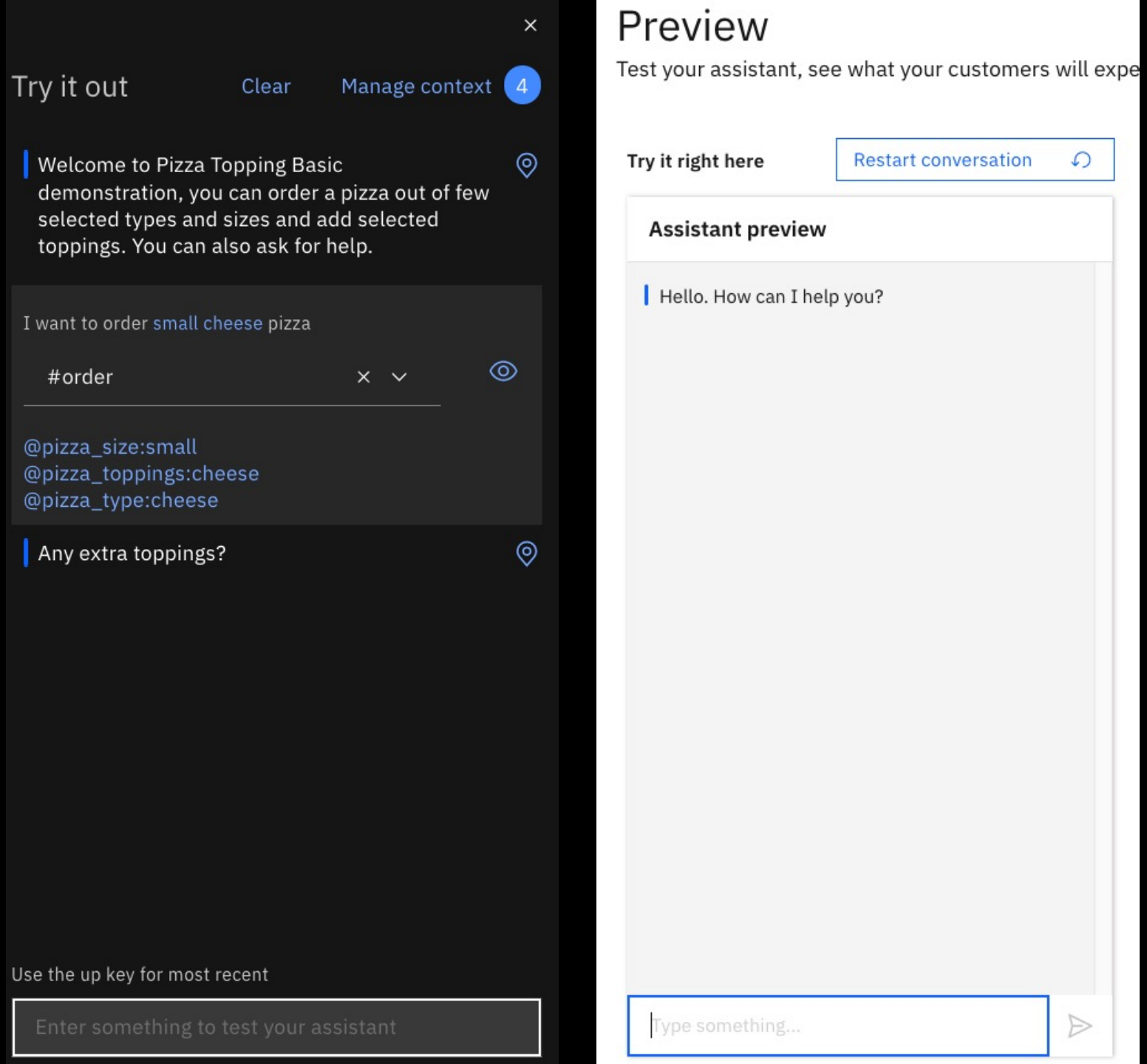
1. Quality
2. Performance
3. Custom metrics
4. Bias/fairness
5. Accuracy & Data Drift

Explain on demand


Risk & Compliance

Business

Chatbot Testing



Chatbot Testing

https://github.com/jiportilla/INSA

curl -X POST -u "apikey:{apikey}" --header "Content-Type:application/json" --data "{\"input\": {\"text\": \"Hello\"}}" "{url}/v2/assistants/{assistant_id}/message?version=2021-06-14"

In [17]:
```
import json
#import sys
#import os
#sys.path.append(os.path.join(os.getcwd(),'..','..'))
import ibm_watson
```

In [18]:
```
from ibm_watson import AssistantV2
from ibm_cloud_sdk_core.authenticators import IAMAuthenticator
```

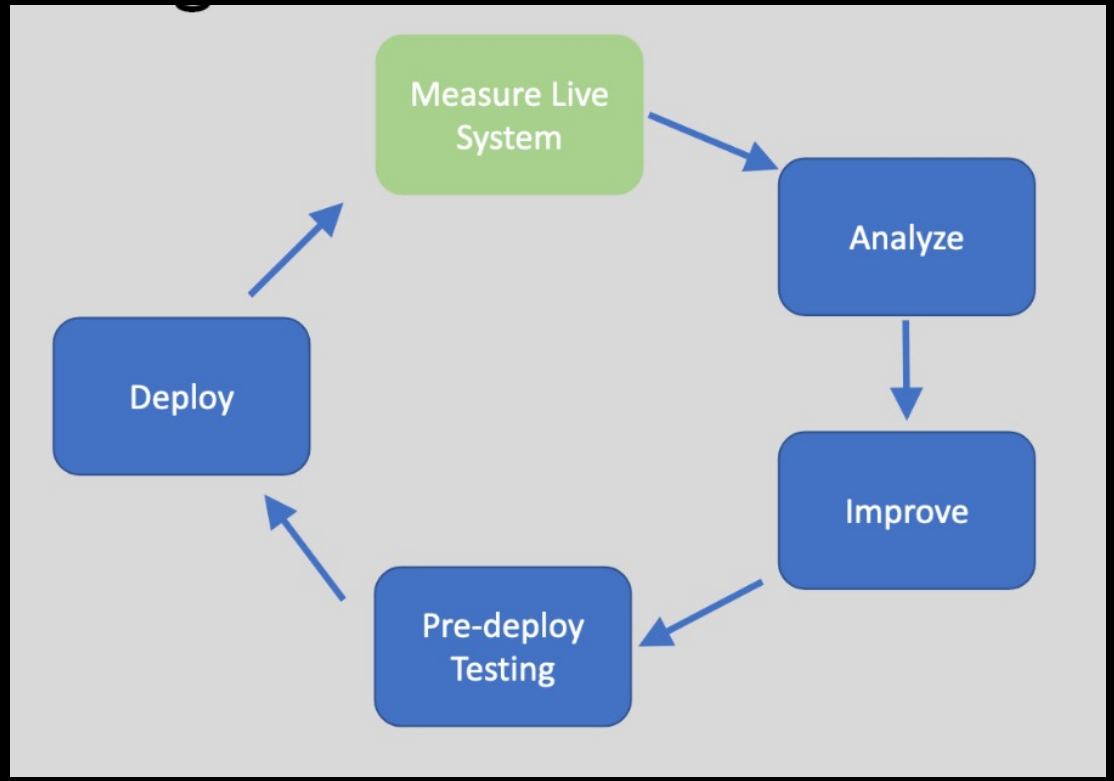
In [19]:
```
authenticator = IAMAuthenticator('{apikey}')
assistant = AssistantV2(
    version='2021-06-14',
    authenticator=authenticator
)

assistant.set_service_url('https://api.us-south.assistant.watson
7')

#
#assistant.set_disable_ssl_verification(True)
```

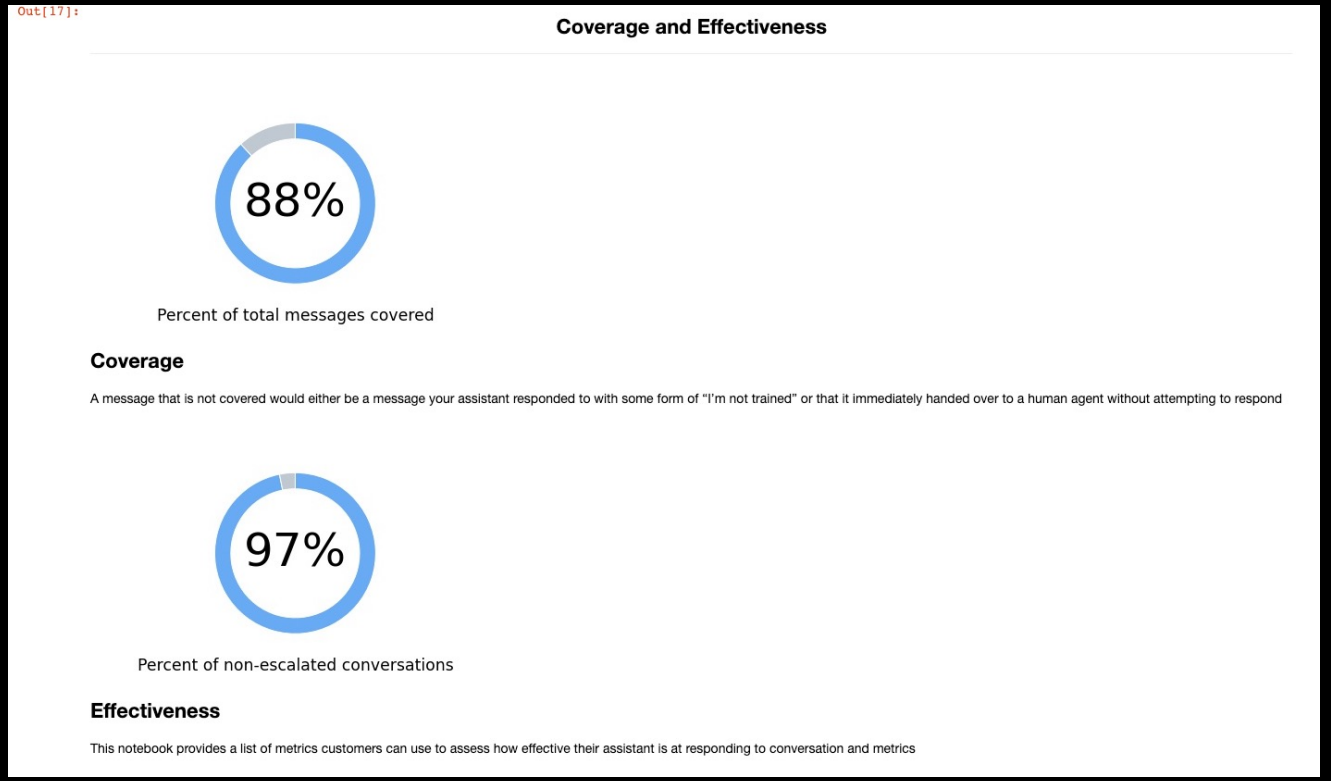
In [22]:
```
response = assistant.message_stateless(
    assistant_id='16f29638-5c96-48ff-b0aa-c42f679b9153',
    input={
        'message_type': 'text',
        'text': 'what is a field notice alert'
    }
).get_result()

print(json.dumps(response, indent=2))
```


```
{
  "output": {
    "intents": [
      {
        "intent": "Field_Notice",
        "confidence": 1
      }
    ],
    "entities": [
      {
        "entity": "Topics_FieldService",
        "location": [
```

Chatbot Testing

https://github.com/jiportilla/ic/blob/master/2021-Work/IC-Continuous%20Improvement%20Framework-Innovation-Center-v0.4.pdf





https://dataplatform.cloud.ibm.com/exchange/public/entry/view/133dfc4cd1480bbe4eaa78d3f635e568

Agenda

MLOps What & Why

1. ML Metrics
2. Trust & Transparency
3. Watson OpenScale

https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Precision and Recall

Precision is the fraction of retrieved instances that are relevant - What proportion of positive identifications was actually correct?

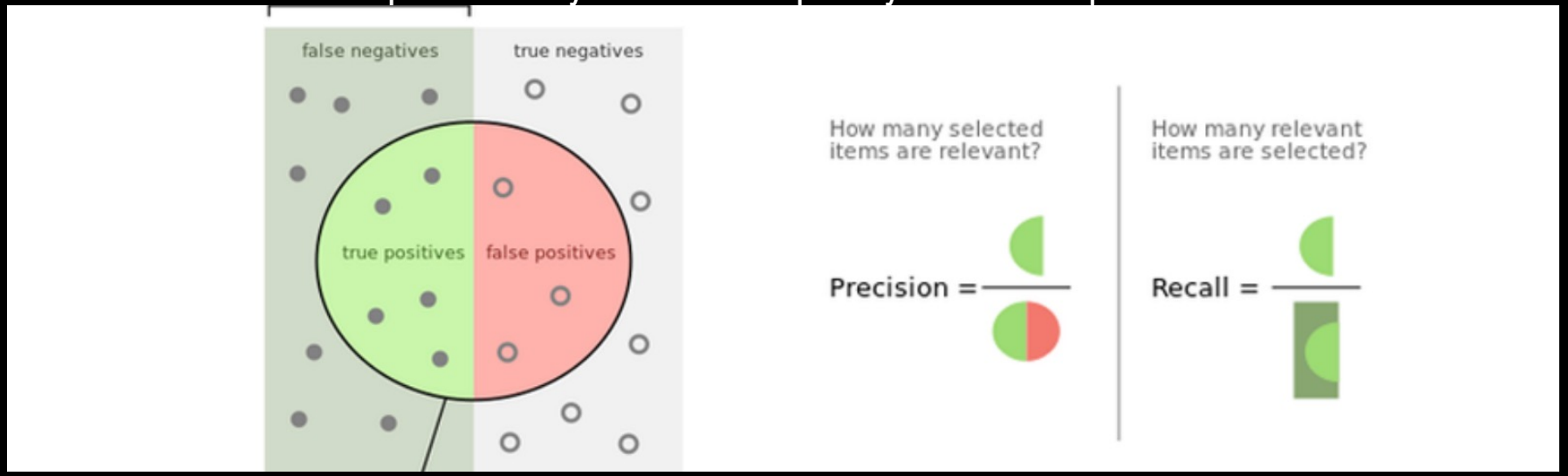
Recall is the fraction of relevant instances that are retrieved - What proportion of actual positives was identified correctly?

Precision (p) = # of aligned mentions / # of referenced mentions

Recall (r) = # of aligned mentions / # of total mentions

f1 is the harmonic mean of precision and recall – punishes you for a disparity between precision and recall

f1 = 2 * P * r / (p + r)



ROC and PR Curves

A receiver operating characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings.
The true-positive rate is also known as sensitivity

A PR Curve uses precision – how many true positives out of all that have been predicted as positives

PR is thought to be more informative when there is a high-class imbalance

Example of Precision and Recall

Suppose a computer program recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats

Of the 8 dogs identified,

- 5 actually are dogs – true positives
- The rest are cats – false positives
- The 7 Dogs that were not recognized - false negatives

The programs precision is …
The programs recall is …

f1 = 0.5

The programs precision is 5/8
The programs recall is 5/12

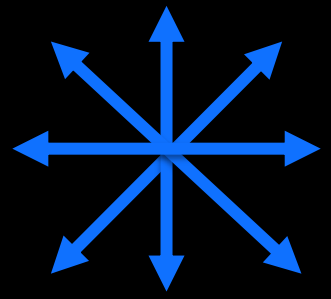
How can we trust an AI-infused recommendation?

Trustworthiness presents new challenges to operationalizing AI

Bias

Training data and AI models may be biased.

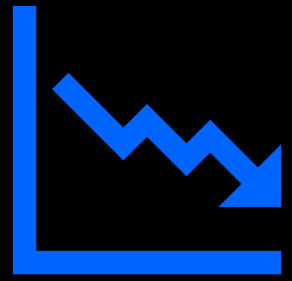
Are privileged groups at a systematic advantage compared to unprivileged groups?



Quality

Models need to perform well across the AI/ML lifecycle.

Are relevant performance metrics being monitored over time?



Drift

Changes in input data cause model to make inaccurate decisions.

Training data may not include data ranges or combinations seen in real life

Explainability

Traditional statistical models are simpler to interpret and explain.

At what point would the outcome have been different?

Trustworthy AI

https://learn.ibm.com/course/view.php?id=8717

The need for trust in AI has been of importance and one way of achieving it is through comprehending how machine learning models predict labels by various means throughout the AI application lifecycle.

https://www.ibm.com/training/collection/trustworthyai

OpenScale

Watson OpenScale Setup
A key component of a Governed MLOps solution is the ability to monitor AI models for accuracy, fairness, explainability and drift.

These capabilities deliver trustworthy AI which business leaders can safely adopt in their business processed and customer engagements.
With multiple client engagements, we have found that having the confidence to trust AI models is just as important, and sometime even more important, than the performance of the AI models.

Watson OpenScale, a component of Watson Studio and Cloud Pak for Data, is IBM's solution to deliver trustworthy AI and enable monitoring of AI models for fairness, explainability and drift.

Next, we show how to leverage Watson OpenScale to monitor the loan risk prediction model we deployed to with AutoAI.

https://youtu.be/l8V_Q_URrWc

https://youtu.be/aEmzG02Mk0U https://youtu.be/9R4Xd2UiNko

OpenScale

What is Watson OpenScale?
As more and more applications make use of (ML) and (DL) models, users find it difficult to comprehend some of the decisions made by these applications. Within an enterprise, users do not completely trust the application since it may be perceived to be inconsistent in its working. This impacts the adoption of artificial intelligence (AI) in order to innovate and grow the business.

Watson OpenScale is an open platform that addresses this and is designed to accelerate the adoption of trusted AI into applications.

Watson OpenScale

- Helps organizations confidently automate and operate AI at scale, with trust and transparency around AI processes, as the demand for AI grows.
- Provides enterprises with visibility into how AI is being built, used, and performs at the scale of the business.
- Can be embedded in AI models into existing or new business applications.
- Is available on IBM Cloud and IBM Cloud Private.

The solution makes it easier for data scientists, application developers, IT and AI operations teams, and business-process owners to collaborate in deploying, managing, monitoring, and updating AI assets in production.

OpenScale

https://developer.ibm.com/articles/the-ai-360-toolkit-ai-models-explained/

AI Fairness 360
This extensible open-source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle.
http://aif360.mybluemix.net/

AI Explainability 360
This extensible open-source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle.
http://aix360.mybluemix.net/

The Adversarial Robustness 360 Toolbox (ART), an open-source software library, supports both researchers and developers in defending deep neural networks against adversarial attacks, making AI systems more secure.

https://developer.ibm.com/open/projects/adversarial-robustness-toolbox/

Agenda

Check class to openscale instance

Load & explain the German risk model data

Show the autoAI experiment, deployment, test

Show openscale

- 3 subscription (prod, preProd, challenger)

Test impact of features impact

explanation

- Explain

Risk with 61.94 confidence level because of

Sex, Loan Amount, age

Run experiment with Inspect

- Inspect

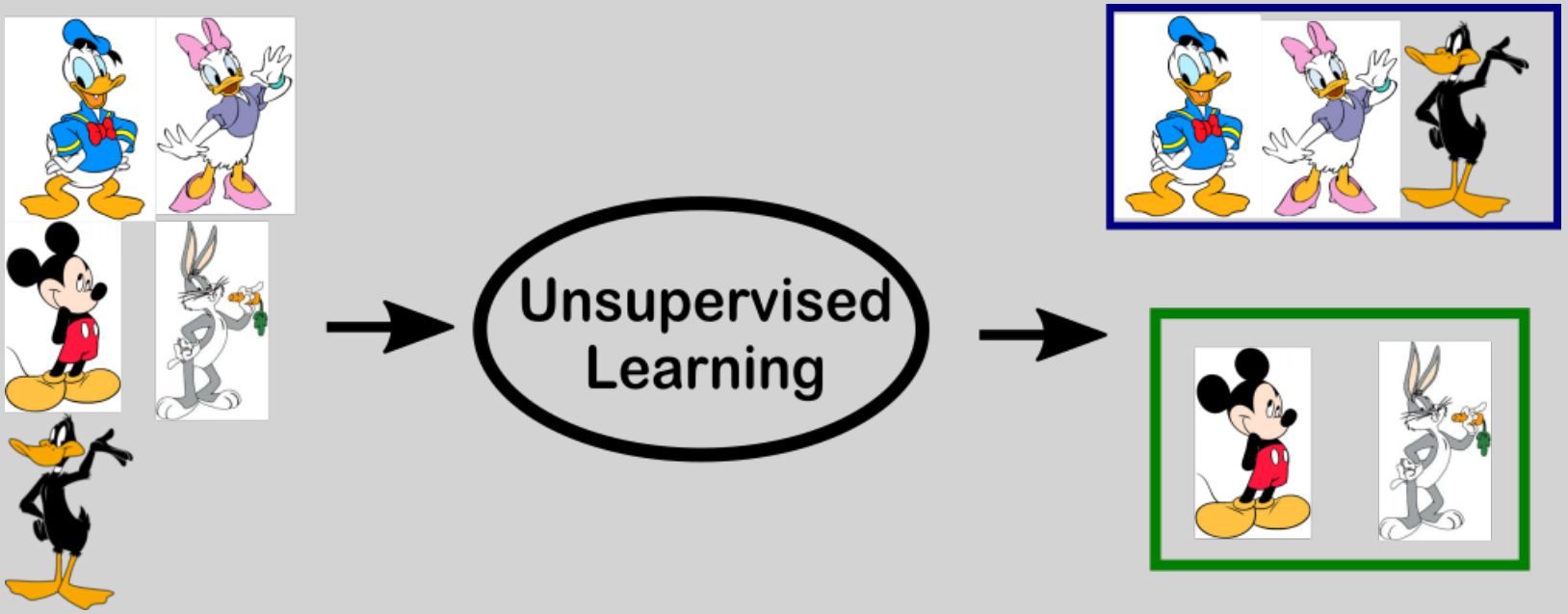
Sex - M

Age - 66

Loan Amount - 2,000

Unsupervised Learning

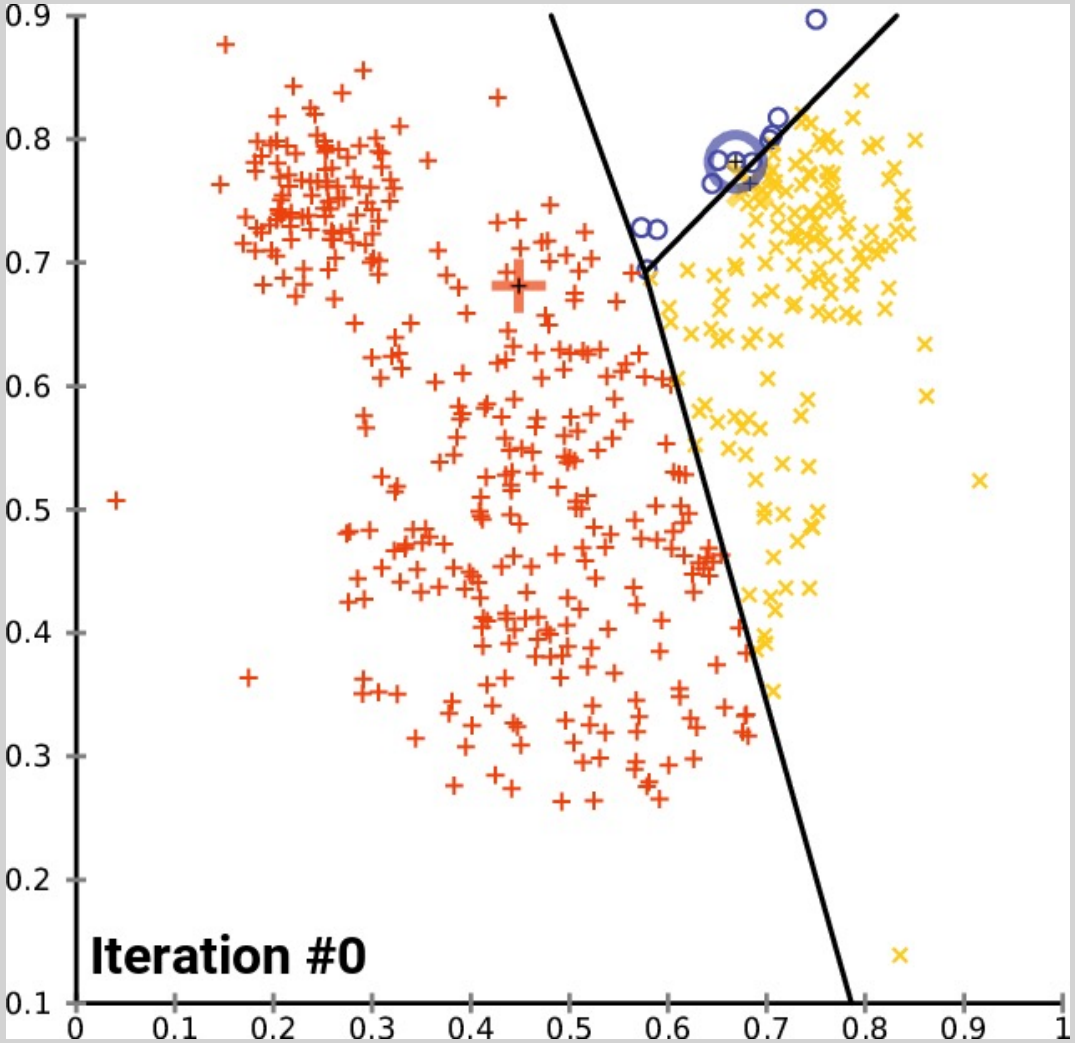
Machine learning task of inferring a function to describe hidden structure from “unlabeled” data



Common scenarios include:
- Data Exploration
- Outlier Detection
- Pattern Recognition

https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51

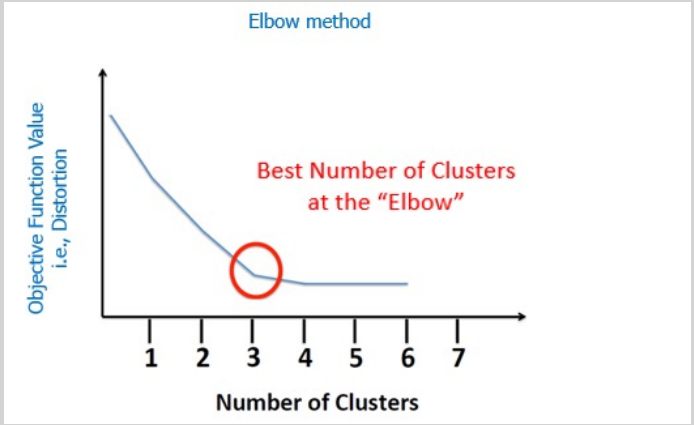
Unsupervised Learning



K-Means clustering:

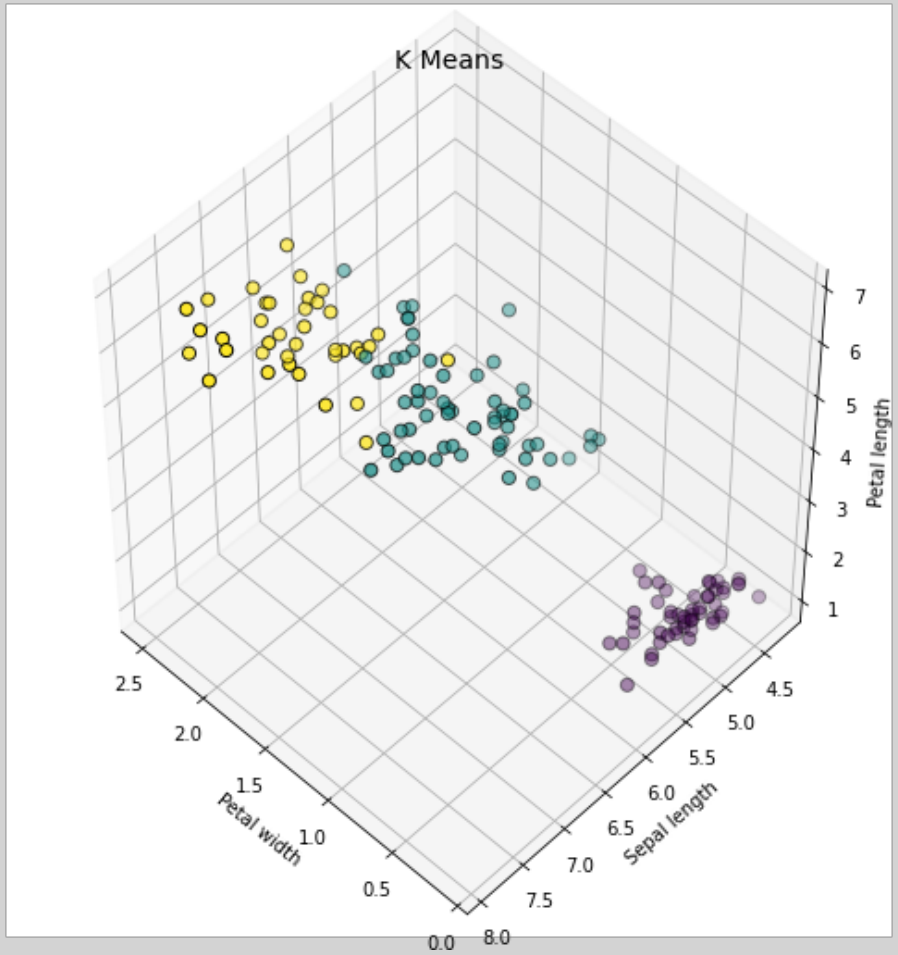
This algorithms involve you telling the algorithms how many possible cluster (or K) there are in the dataset.

The algorithm then iteratively moves the k-centers and selects the datapoints that are closest to that centroid in the cluster.



https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51

Unsupervised Learning




```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
%matplotlib inline
from sklearn import datasets#Iris Dataset
iris = datasets.load_iris()
X = iris.data#KMeans
km = KMeans(n_clusters=3)
km.fit(X)
km.predict(X)
labels = km.labels_#Plotting
fig = plt.figure(1, figsize=(7,7))
ax = Axes3D(fig, rect=[0, 0, 0.95, 1], elev=48, azim=134)
ax.scatter(X[:, 3], X[:, 0], X[:, 2],
c=labels.astype(np.float), edgecolor="k", s=50)
ax.set_xlabel("Petal width")
ax.set_ylabel("Sepal length")
ax.set_zlabel("Petal length")
plt.title("K Means", fontsize=14)
```

https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51

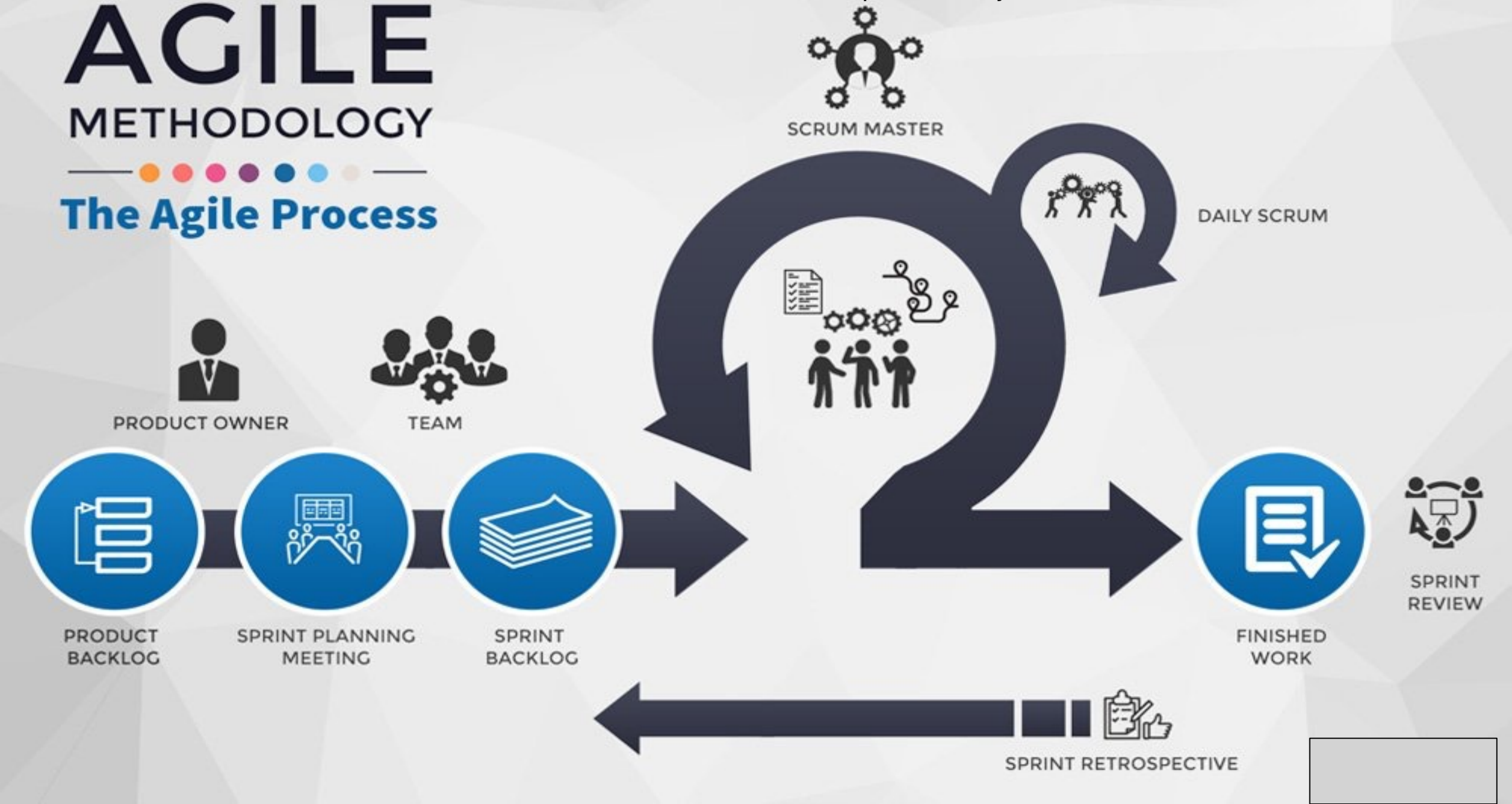
Demos - Lab

Agile Development

https://www.youtube.com/watch?v=OJflDE6OaSc

AGILE METHODOLOGY

The Agile Process



Project: Campus brochure

As a < type of user >, I want < some goal > so that < some reason >.

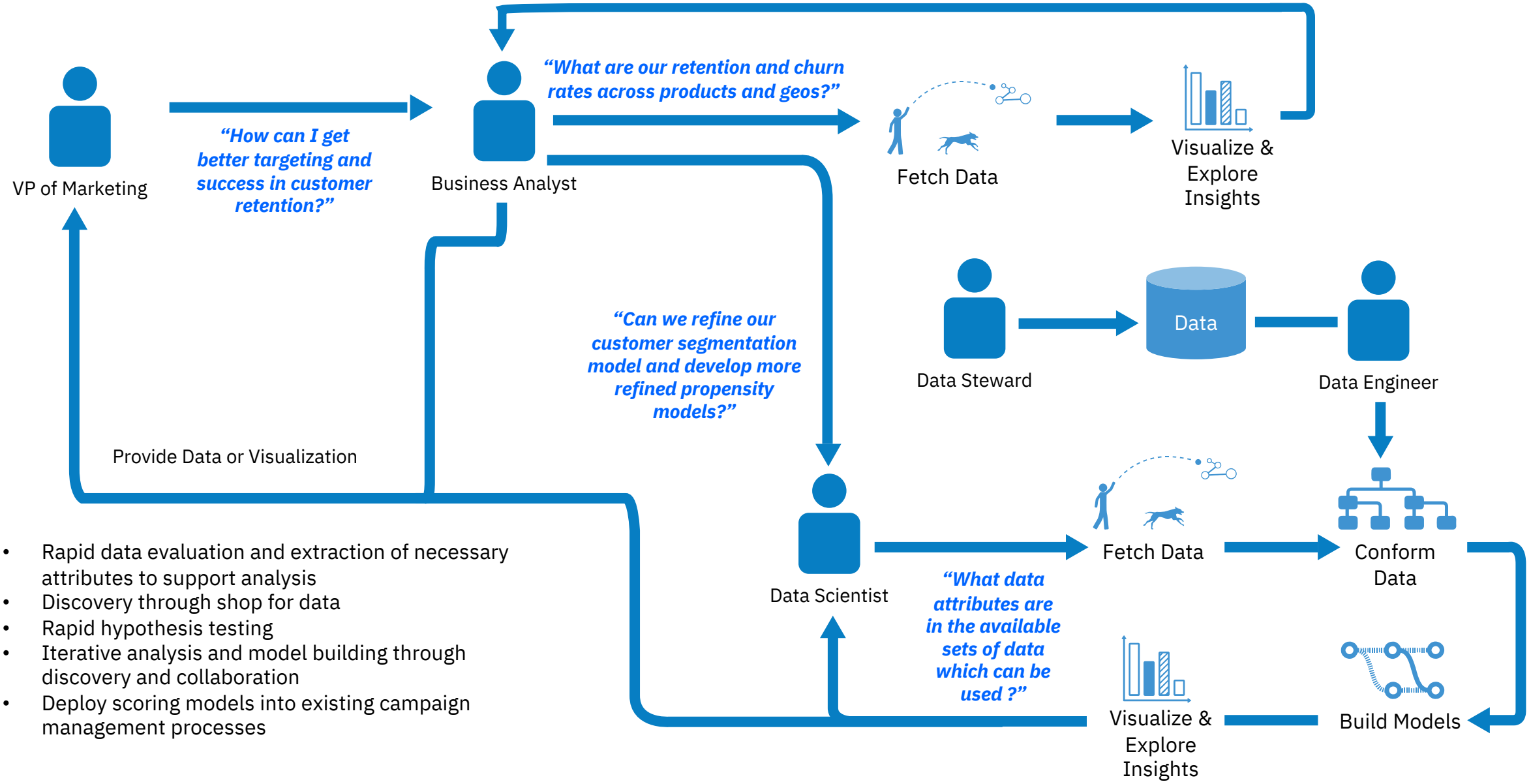
As a parent, I want to visit the main campus so my son can feel comfortable studying at CU

As a future student, I want to check out college life so that I can continue my education on safe environment

As a faculty, I want to get familiar with the buildings layout so that I can find my way around for my classes

https://www.mountaingoatsoftware.com/agile/user-stories

Sample business problem: Churn Analysis



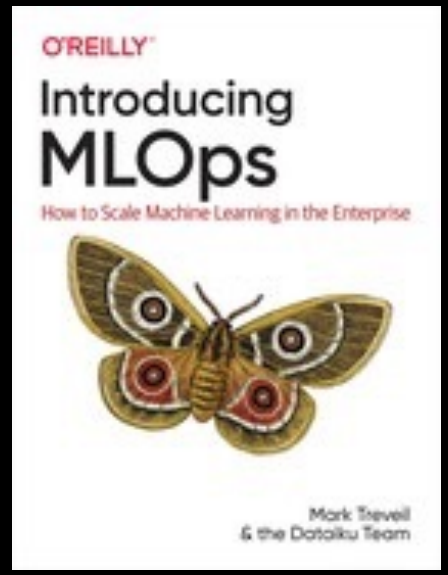
- Rapid data evaluation and extraction of necessary attributes to support analysis
- Discovery through shop for data
- Rapid hypothesis testing
- Iterative analysis and model building through discovery and collaboration
- Deploy scoring models into existing campaign management processes

Q&A

Ivan Portilla

ivanp@us.ibm.com

@iportilla



https://medium.com/inside-machine-learning/ai-ops-managing-the-end-to-end-lifecycle-of-ai-3606a59591b0