A Day in the Life of a Data Scientist: How do we train our teams to get started with AI?

Francesca Lazzeri & Jaya Mathew
@frlazzeri @mathew_jaya

Strata
DATA CONFERENCE

Agenda

- What is Artificial Intelligence (AI)?
- Why AI?
- How to get started with AI
- Understanding the ML workflow
- Suggested tools for AI development
- AI usage in marketing
- Fraud management use case

What is AI?

What is AI? – To sense, comprehend and act

Sense

Computer vision and audio processing, for example are able to actively perceive the world around them by acquiring and processing images, sounds and speech. The use of facial recognition at border control kiosks is one practical example of how it can improve productivity.

Comprehend

Natural language processing and inference engines can enable AI systems to analyse and understand the information collected. This technology is used to power the language translation feature of search engine results

Act

An AI system can take action through technologies such as expert systems and inference engines or undertake actions in the physical world. Auto-pilot features and assisted-braking capabilities in cars are examples of this

AI Technologies

Emerging AI technologies

- Computer Vision
- Audio Processing
- Natural Language Processing
- Knowledge Representation
- Machine Learning
- Expert Systems

Illustrative Solutions

Virtual Agents

Identity Analytics

Cognitive Robotics

Speech Analytics

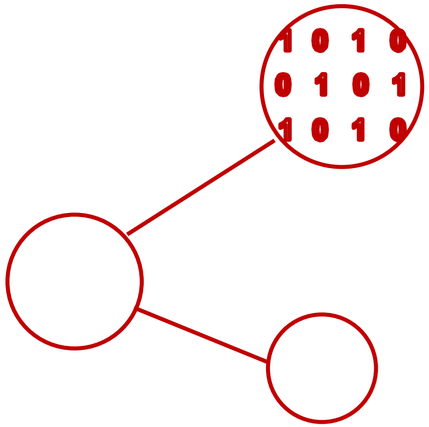
Recommendation Systems

Data Visualization

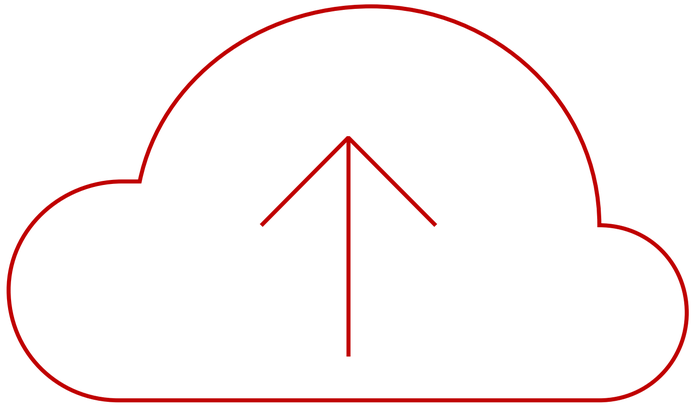
Source: Accenture: Why artificial intelligence is the future of growth, April 2016

Why AI?

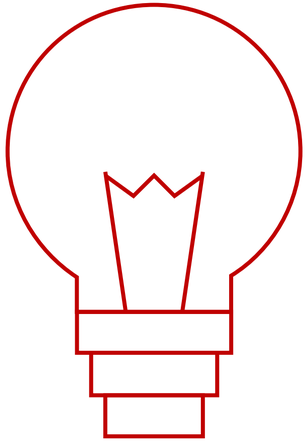
Three major trends are converging



Data

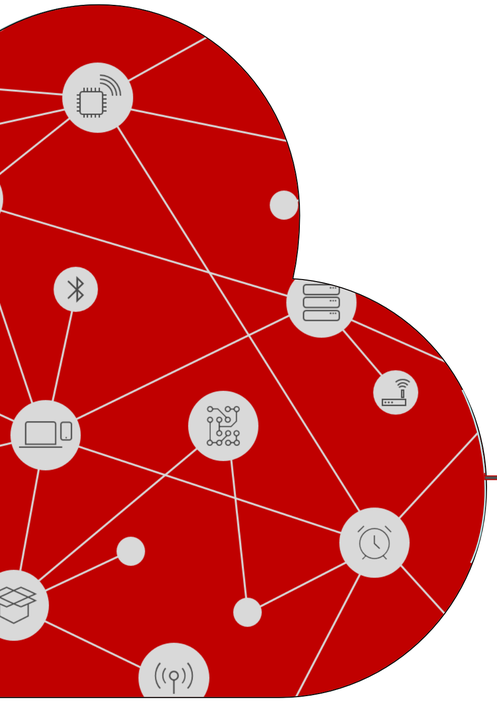


Cloud



Intelligence

From data to decisions and actions



Descriptive
[Reports]

What happened?

Diagnostic
[Interactive Dashboards]

Why did it happen?

Predictive
[Machine Learning]

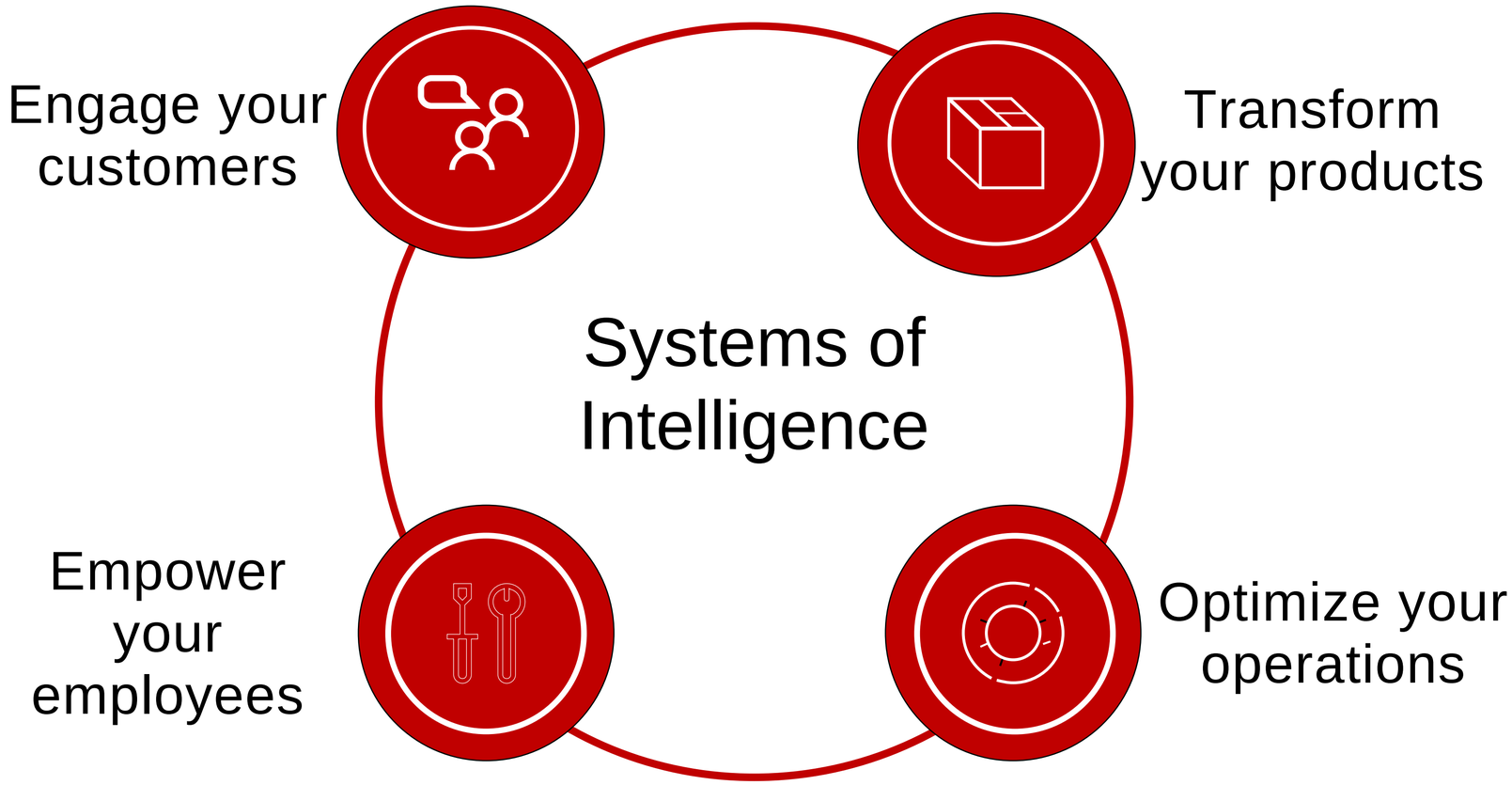
What will happen?

Prescriptive
[Recommendations & Automation]

What should I do?

Insight

Digital transformation is driving new business value



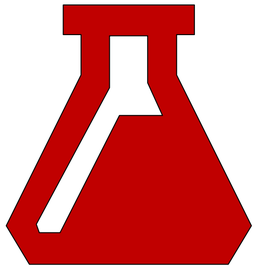
How to get started with AI

I want to use AI – How can I get started?

aka.ms/AICognitiveServices
aka.ms/AICustomModels

Cognitive
Services

Custom Services
Bring your Own Data



Custom
AI Models

Flexibility

@frlazzeri @mathew_jaya

Being Obsessed with Data

Question is sharp	Data measures what you care about	Data is accurate	Data is connected	A lot of data
E.g. Predict whether component X will fail in the next Y days	E.g. Identifiers at the level you are predicting, relevant data collected & feature engineering using domain knowledge	E.g. Failures are really failures, human labels on root causes	E.g. Machine information linkable to usage information	E.g. Will be difficult to predict failure accurately with few examples

Asking the right questions

Business scenario	Key decision	Data Science question
Energy forecasting	Should I buy or sell energy contracts?	What will be the long/short-term demand for energy in a region?
Customer churn	Which customers should I prioritize to reduce churn?	What is probability of churn within X days for each customer?
Personalized marketing	What product should I offer first?	What is the probability that customer will purchase each product?
Product feedback	Which service/product needs attention?	What is social media sentiment for each service/product?

Defining Performance Metrics

Establish a **Qualitative Objective**

Translate into **Quantifiable Metric**

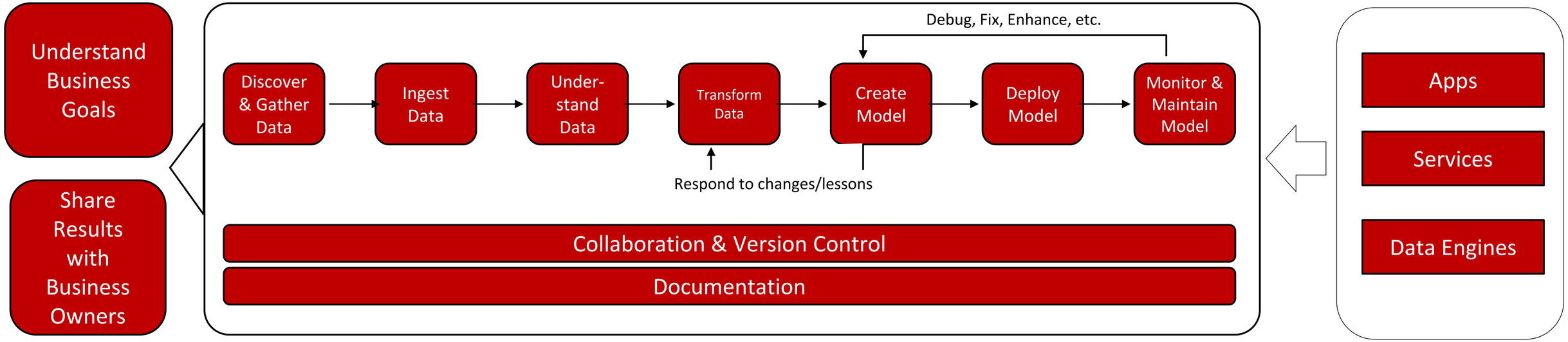
Quantify the **Metric Value Improvement** useful for customer scenario

Establish a **Baseline**

Establish how to measure the improvement in the **Data Science Metric**

Understanding the ML workflow

Sample ML workflow



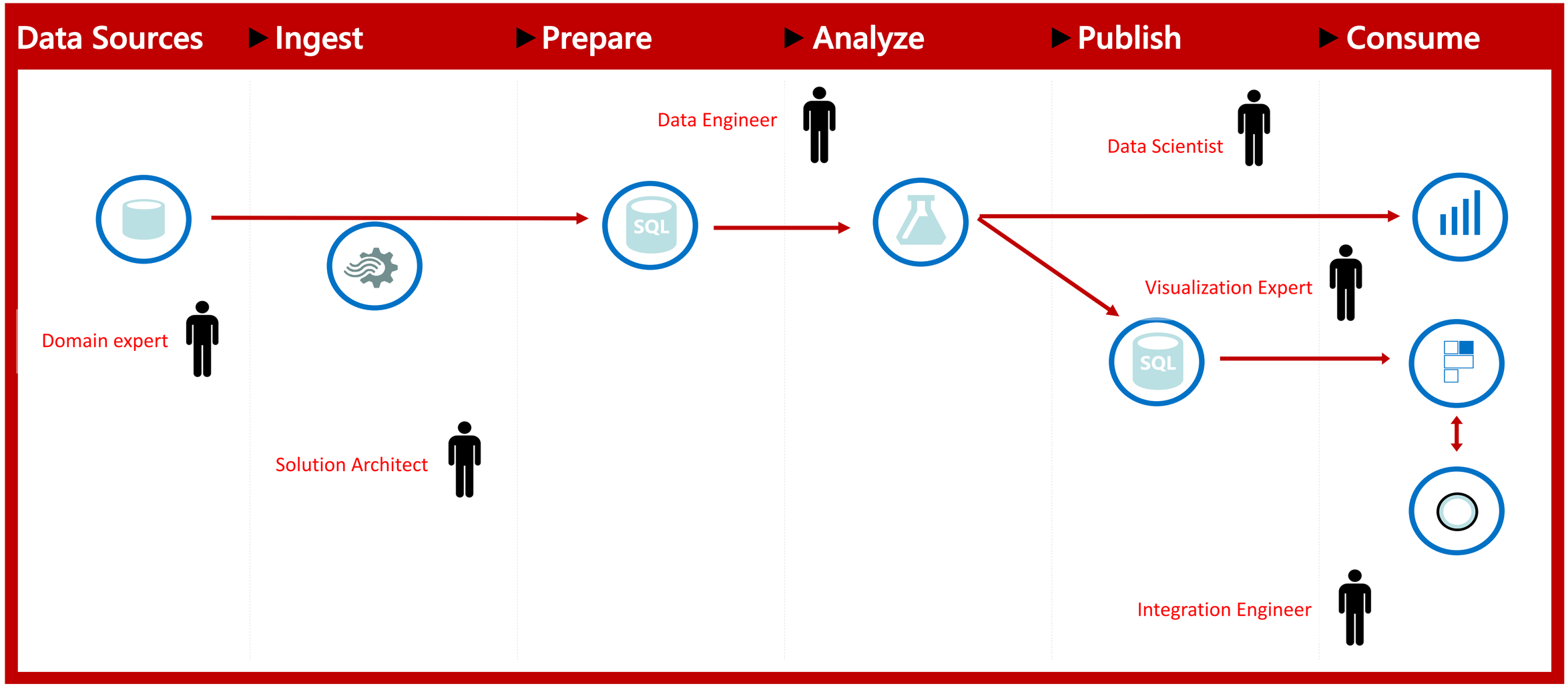
Build a model
Science

Publish a model
Operations

Consume a model

Team Data Science Process

aka.ms/TeamDataScience



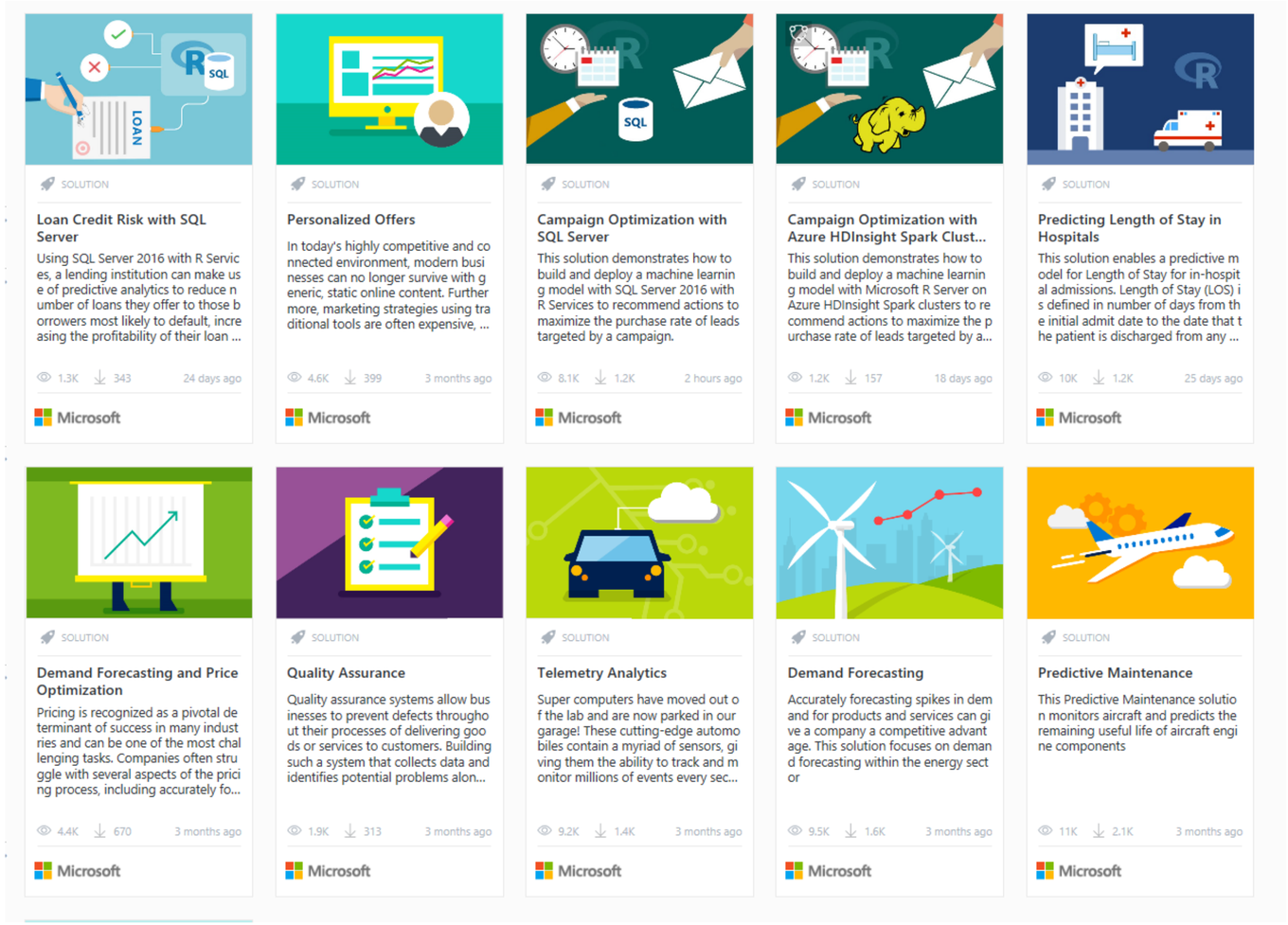
Suggested tools for AI development

Example Tools / Dev Environments

- **Azure Machine Learning Studio:** Serverless collaborative drag-and-drop tool for graphical machine learning development
- **Azure Machine Learning Services:** Visual AI powered data wrangling, experimentation, and lifecycle management
- **Visual Studio Code Tools for AI:** Build, debug, test, and deploy AI with Visual Studio Code on Windows and Mac
- **Azure Notebooks:** Organize your datasets and Jupyter Notebooks in a centralized library for Data Science and Analysis
- **Deep Learning Virtual Machine:** A pre-configured environment for deep learning using GPU instances

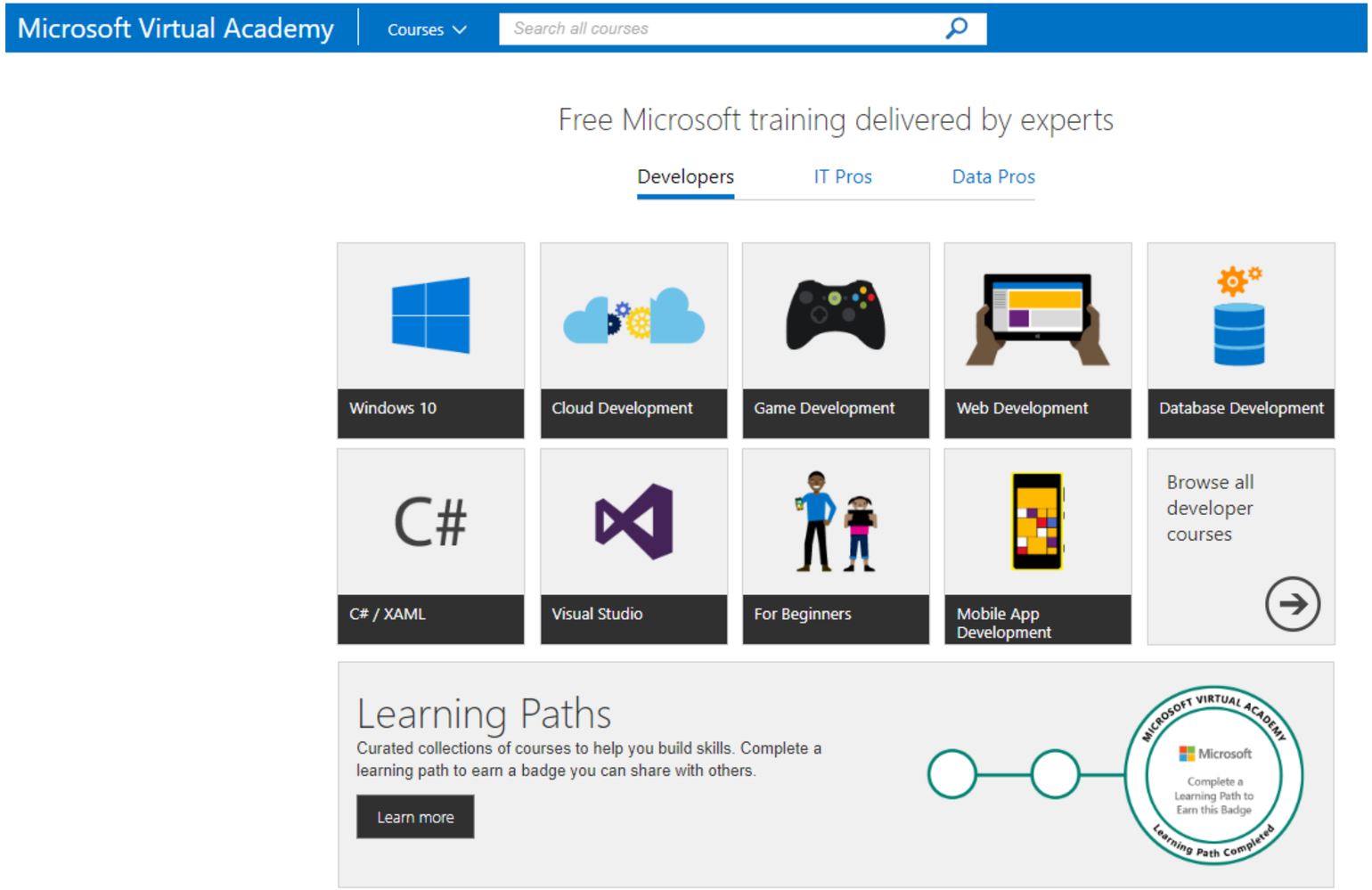
AI Solution Templates

aka.ms/AzureAIGallery



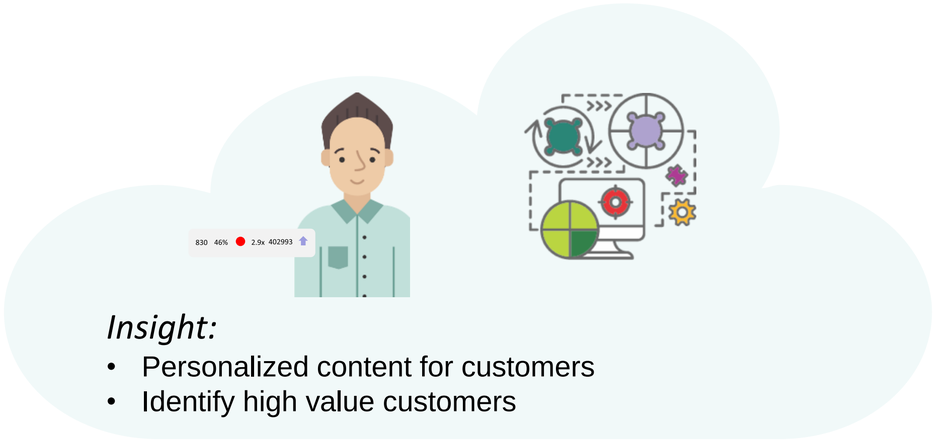
Microsoft Learning

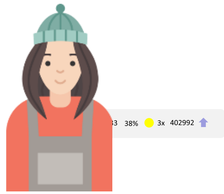
aka.ms/MicrosoftAILearning



AI usage in marketing

How can AI help in Marketing & Personalization?





Input Data

- In store
- Online activity
- Social media
- Past campaign performance

ML & AI

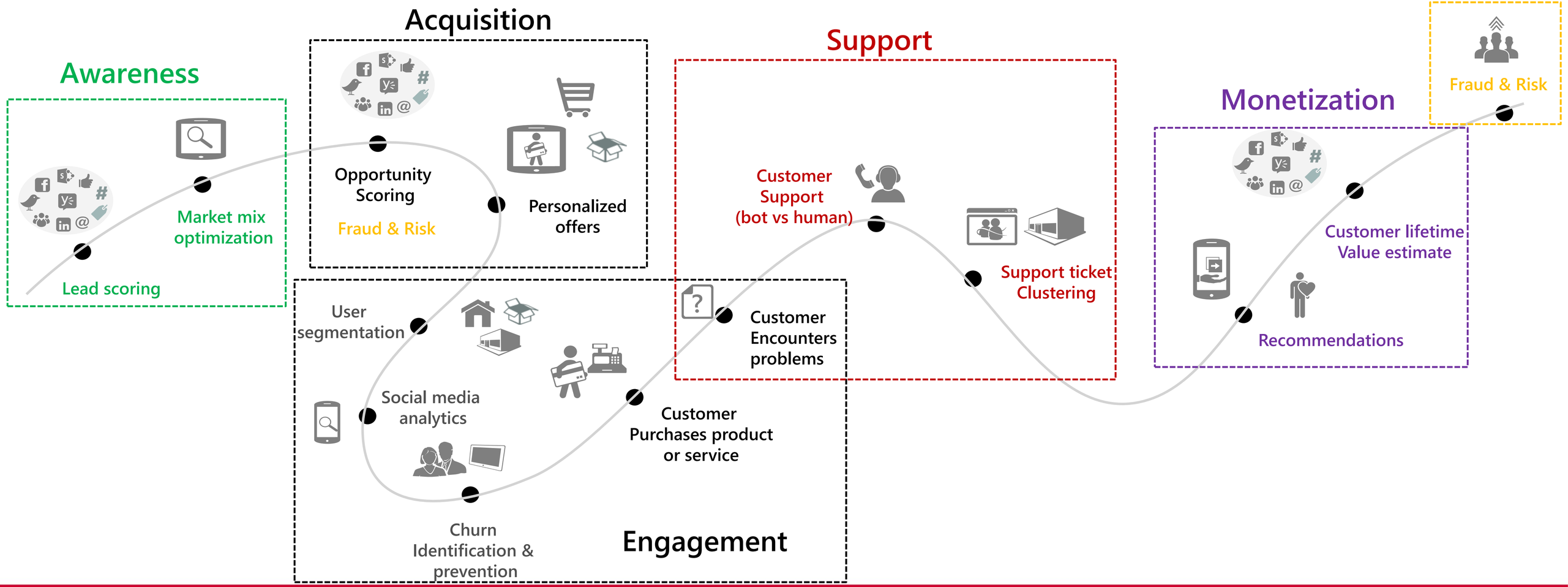
- Learn customer omni-channel preference
- Personalize web/email experience
- Preemptive chatbots to answer customer queries
- Monitor for cart abandonment, churn, retention

Intelligent Action

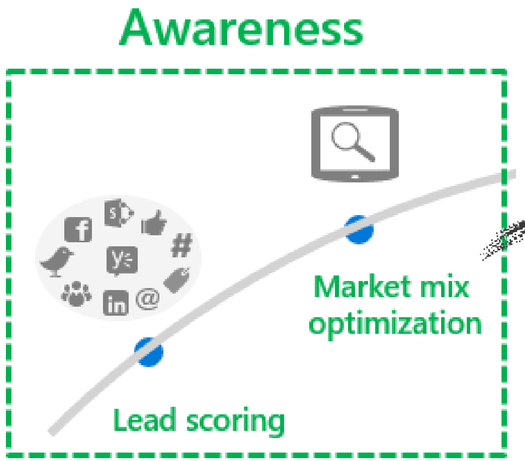
- Personalized website experience
- Adaptive product pricing, offers for cross/upsell
- Premium loyalty programs and service experience
- Predictive customer service via social media, chatbots

Marketer's AI Journey Map

Let us map the journey based on the stages in a customer life cycle as follows:
Awareness; Acquisition; Engagement; Support; Monetization; Fraud/Risk management



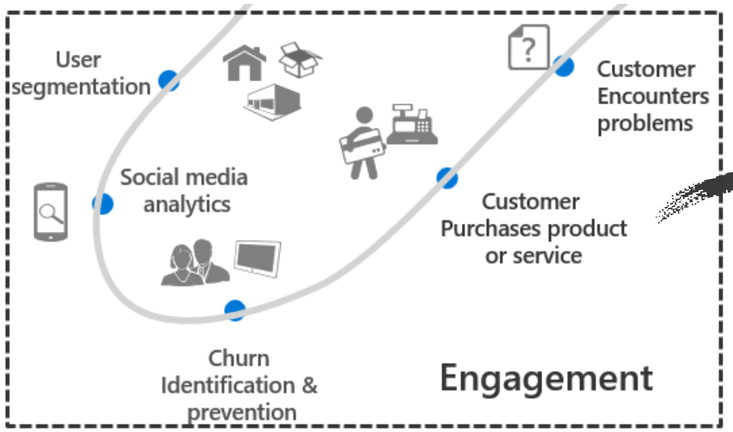
Marketing cycle – Awareness & Acquisition



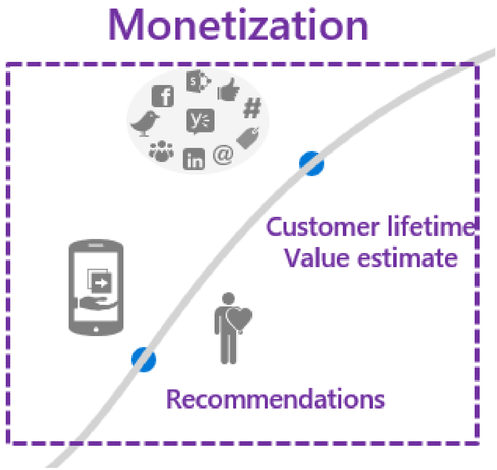
- ✓ Marketing mix optimization for omni-channel budget optimization
- ✓ ML based lead-scoring models

- ✓ Opportunity scoring can help target users who are most likely to make a purchase
- ✓ AI powered content creation

Marketing cycle – Engagement & Monetization

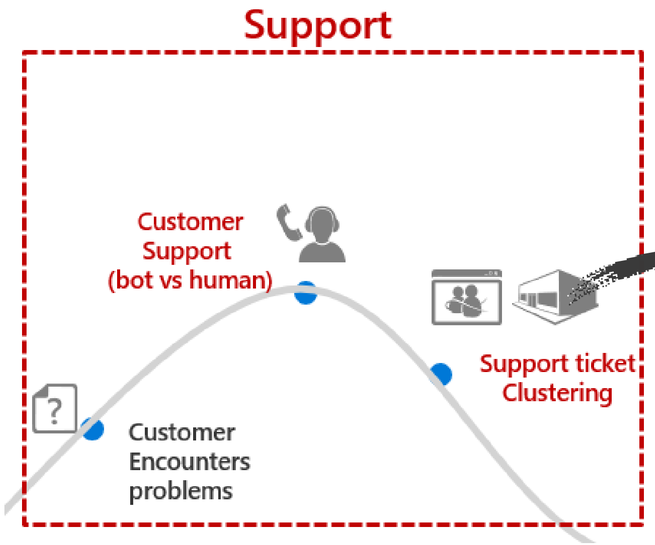


- ✓ Customer segmentation based on their browsing/purchase patterns
- ✓ ML based models can help determine customers who are most likely to churn in the near future based on their behavior



- ✓ AI powered personalized recommendations
- ✓ Using all the information available on customer's buying, browsing patterns determine the Lifetime value (LTV)

Marketing cycle – Support & Fraud



- ✓ Automated answers based on a knowledge base can help decrease call center costs and predict staffing needs
- ✓ Support ticket clustering can help the team find solutions for customer complains quickly and even report similar problems back to the product teams

- ✓ While acquiring customers, external customer data can be used to determine customers who are likely to be risky for the business in the long run
- ✓ ML & AI models can be built to predict if a customer is likely to end up in bankruptcy, being delinquency which results in non-payment for the products/services

Fraud management use case

Credit Card Fraud Detection Data Set

- Dataset used:
  - contains transactions made by credit cards in September 2013 by European cardholders
  - presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions
- Features *V1, V2, ... V28*: are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'
- Feature *Time*: contains the seconds elapsed between each transaction and the first transaction in the dataset
- Feature *Amount*: is the transaction Amount
- Feature *Class*: is the response variable and it takes value 1 in case of fraud and 0 otherwise

Time	V1	V2	V3	V4
0	-1.3598071336738	-0.072781173309849 7	2.53634673796914	1.37815522427443
0	1.19185711131486	0.26615071205963	0.16648011335321	0.448154078460911
1	-1.35835406159823	-1.34016307473609	1.77320934263119	0.379779593034328
1	-0.966271711572087	-0.185226008082898	1.79299333957872	-0.863291275036453
2	-1.15823309349523	0.877736754848451	1.548717846511	0.403033933955121
2	-0.425965884412454	0.960523044882985	1.14110934232219	-0.168252079760302
4	1.22965763450793	0.141003507049326	0.0453707735899449	1.20261273673594
7	-0.644269442348146	1.41796354547385	1.0743803763556	-0.492199018495015

@frlazzeri @mathew_jaya

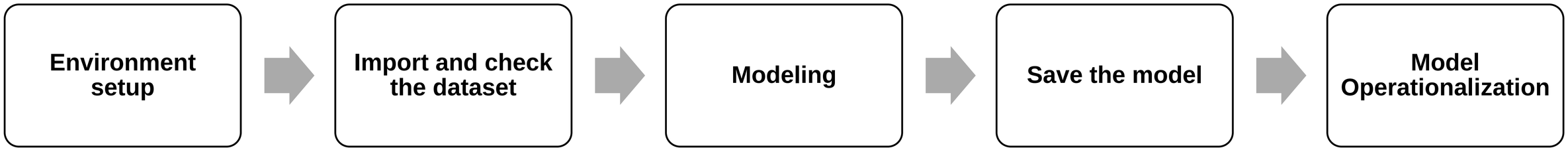
Data Source: https://www.kaggle.com/mlg-ulb/creditcardfraud

Demo

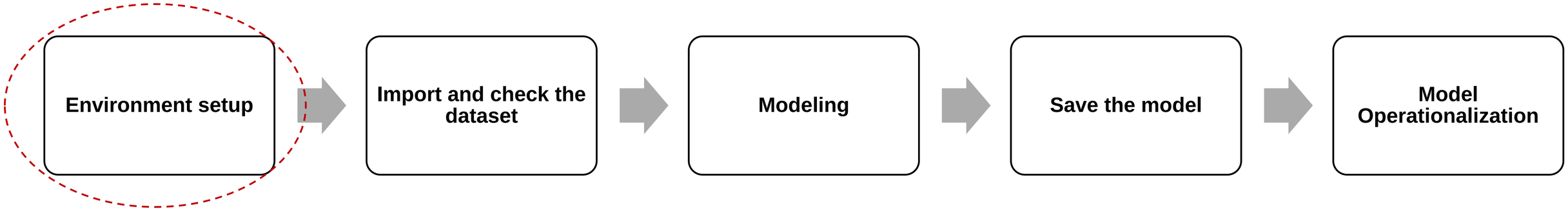
- The sample code was tested and run using the Jupyter notebook environment on a remote Azure VM (Standard F8s (8 vcpus, 16 GB memory))

- The sample code is available at the following GitHub location: https://github.com/jayamathew/Codebase/tree/master/conferences

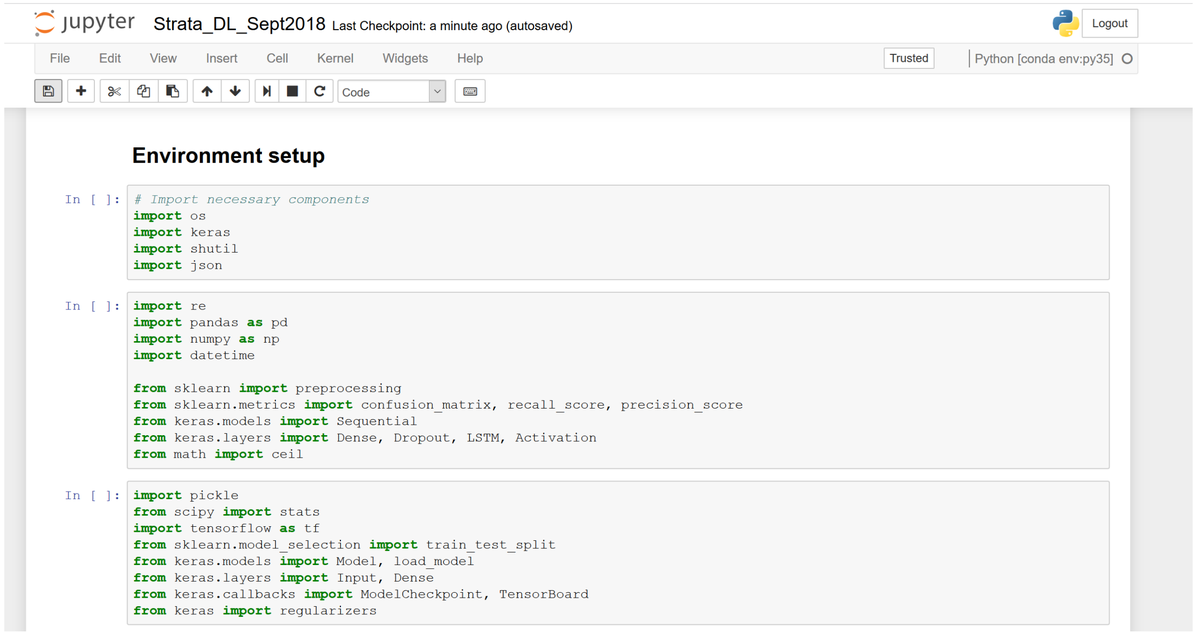
- The outline of the code is as follows:



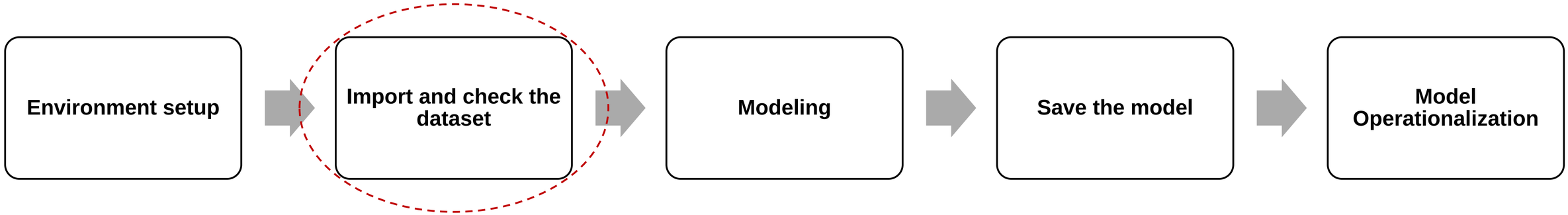
Demo



- Import the necessary libraries and provide credentials to access the data



Demo



- Import the dataset and check the distributions of the variables

jupyter Strata_DL_Sept2018 Last Checkpoint: 2 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python [conda env:py35]

Code

Import the Credit card data set

In [ ]:
```
# Check the path
aml_dir
```

In [ ]:
```
# Ingest the dataset
cc = pd.read_csv('C://dsvm//notebooks/creditcard.csv')
```

After data ingestion from Blob, check to see the various columns and number of rows/columns of the dataset.

In [ ]:
```
# Check sample data
cc.head(1)
```

In [ ]:
```
# Check the number of rows/columns
cc.shape
```

Now that the data is properly imported, check the descriptive statistics of the columns in the dataset.

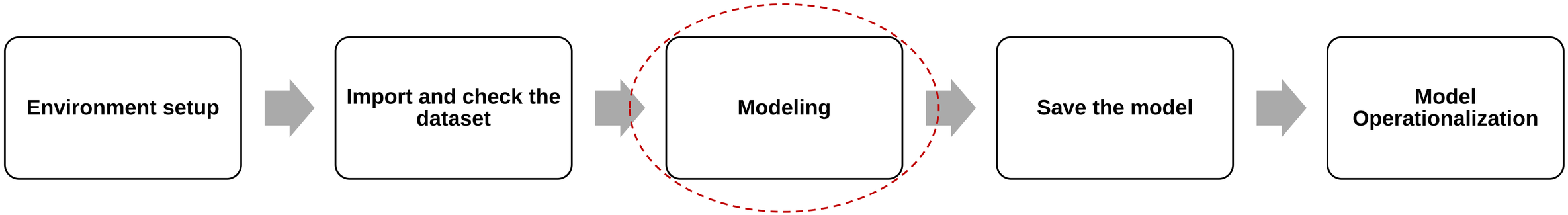
In [ ]:
```
# Check data statistics
print(cc.describe())
```

Here we visualize and access the distribution of the variable 'Class'. This is the variable which indicates whether a transaction was fraud/normal.

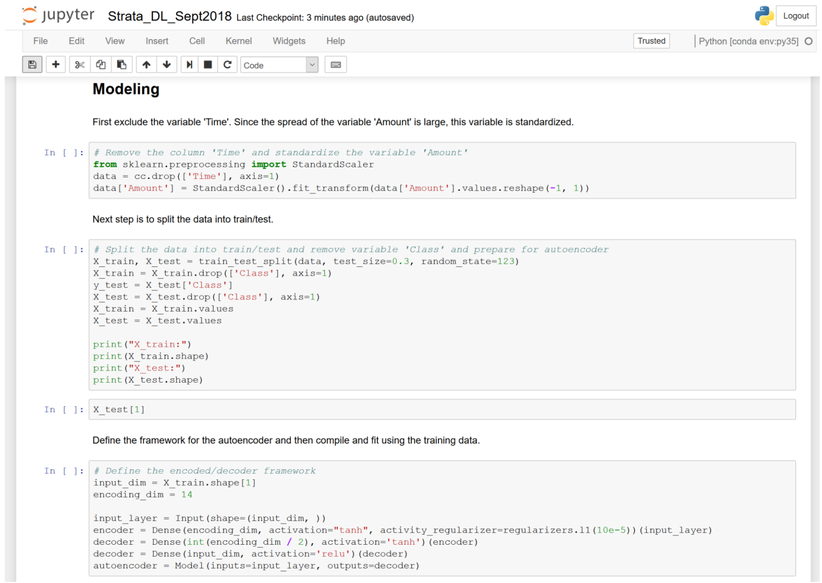
In [ ]:
```
from matplotlib import pyplot as plt
```

In [ ]:
```
# Variable class is used for the classification of entries as fraud/non-fraud, check the distribution of the variable
class_freq = pd.value_counts(cc['Class'], sort = True)
class_freq.plot(kind = 'bar', rot=0)
plt.title("Class Frequency")
plt.xlabel("Class")
plt.ylabel("Frequency");
```

Demo



- Build any model and tune hyper parameters (if needed)


jupyter Strata_DL_Sept2018 Last Checkpoint: 3 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python [conda env:py35]

Code

Modeling

First exclude the variable 'Time'. Since the spread of the variable 'Amount' is large, this variable is standardized.

In [ ]:
```
# Remove the column 'Time' and standardize the variable 'Amount'
from sklearn.preprocessing import StandardScaler
data = cc.drop(['Time'], axis=1)
data['Amount'] = StandardScaler().fit_transform(data['Amount'].values.reshape(-1, 1))
```

Next step is to split the data into train/test.

In [ ]:
```
# Split the data into train/test and remove variable 'Class' and prepare for autoencoder
X_train, X_test = train_test_split(data, test_size=0.3, random_state=123)
X_train = X_train.drop(['Class'], axis=1)
y_test = X_test['Class']
X_test = X_test.drop(['Class'], axis=1)
X_train = X_train.values
X_test = X_test.values

print("X_train:")
print(X_train.shape)
print("X_test:")
print(X_test.shape)
```

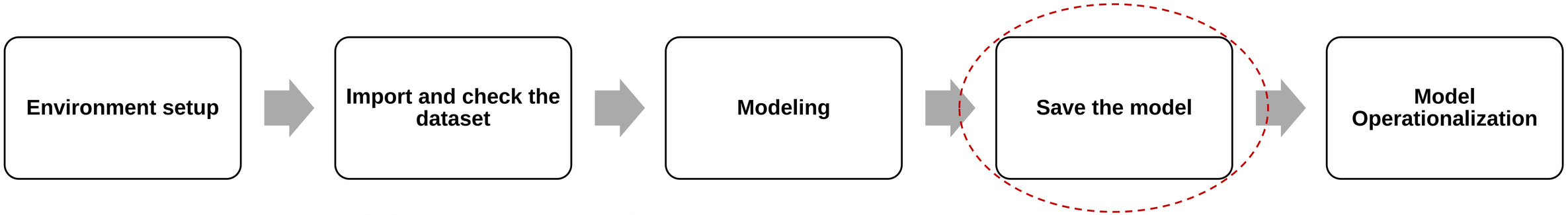
In [ ]:
```
X_test[1]
```

Define the framework for the autoencoder and then compile and fit using the training data.

In [ ]:
```
# Define the encoded/decoder framework
input_dim = X_train.shape[1]
encoding_dim = 14

input_layer = Input(shape=(input_dim, ))
encoder = Dense(encoding_dim, activation="tanh", activity_regularizer=regularizers.l1(10e-5))(input_layer)
decoder = Dense(int(encoding_dim / 2), activation='tanh')(encoder)
decoder = Dense(input_dim, activation='relu')(decoder)
autoencoder = Model(inputs=input_layer, outputs=decoder)
```

Demo



- Save the best model for operationalization

jupyter Strata_DL_Sept2018 Last Checkpoint: 4 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python [conda env:py35]

Code

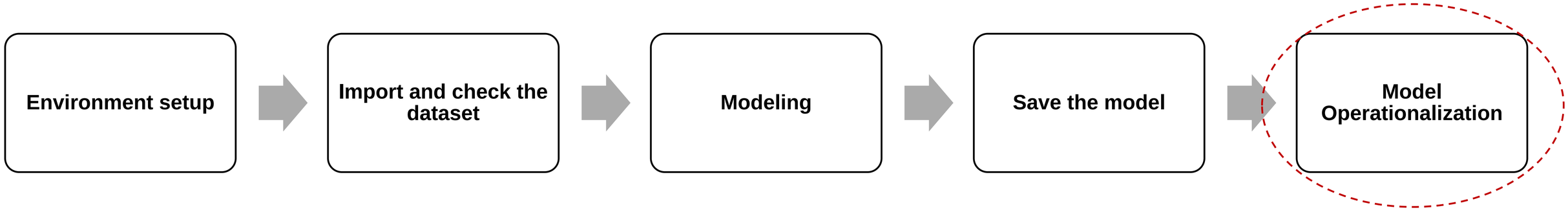
Saving the model

In [ ]:
```
autoencoder
```

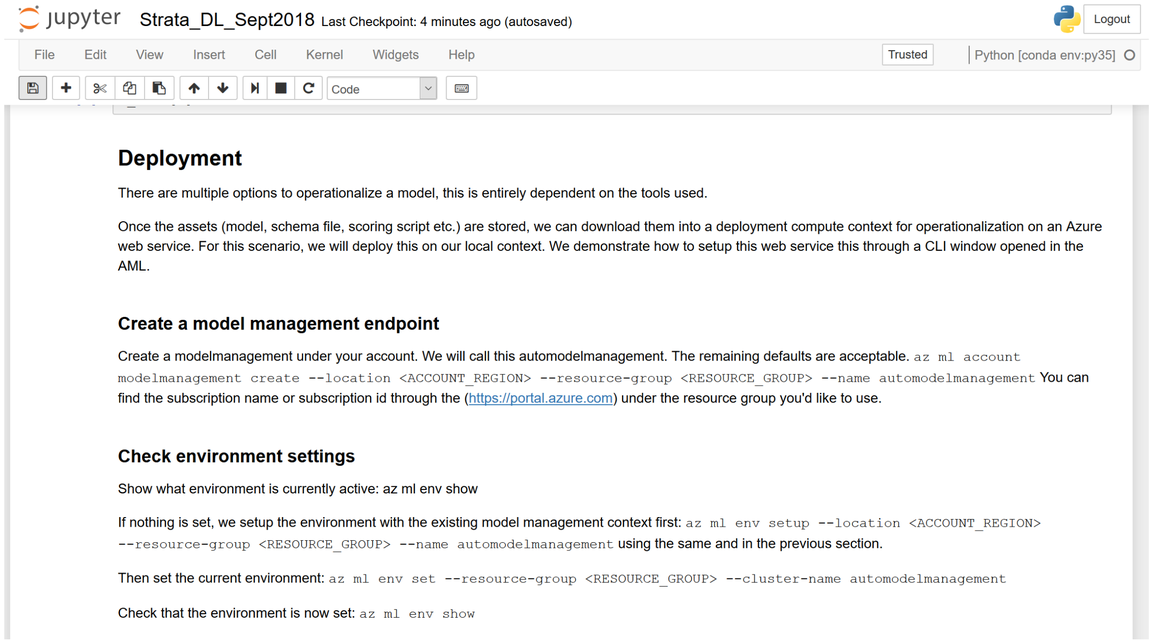
In [ ]:
```
# Save the model for operationalization: https://machinelearningmastery.com/save-load-keras-deep-learning-models/
from keras.models import model_from_json
import os
import h5py
from sklearn import datasets

# save model
# serialize model to JSON
model_json = autoencoder.to_json()
with open("C://dsvm//notebooks/autoencoder.json", "w") as json_file:
    json_file.write(model_json)
# serialize weights to HDF5
autoencoder.save_weights("C://dsvm//notebooks/autoencoder.h5")
print("Model saved")
```

Demo



- Create the necessary functions for model operationalization using any tool of choice



Links to get started with AI

- Cognitive Services: https://aka.ms/AICognitiveServices
- Azure ML: https://aka.ms/AICustomModels
- Azure Machine Learning Studio: https://aka.ms/AzureStudio
- Azure Machine Learning Services: https://aka.ms/AMLServices
- Visual Studio Code Tools for AI: https://aka.ms/VSCodeToolsAI
- Azure Notebooks: https://aka.ms/AzureJNotebooks
- Preconfigured Virtual Machines: https://aka.ms/AzureVirtualMachines
- Deep Learning Virtual Machine: https://aka.ms/AzureDSVM
- Team Data Science Process: https://aka.ms/TeamDataScience
- Data Source for demo: https://www.kaggle.com/mlg-ulb/creditcardfraud
- Blog Post by Venelin Valkov: https://medium.com/@curiousily/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd
- GitHub location for demo: https://github.com/jayamathew/Codebase/tree/master/conferences
- Deep Learning Book by Ian Goodfellow, Yoshua Bengio, Aaron Courville: http://www.deeplearningbook.org/

Thank You!

Francesca Lazzeri & Jaya Mathew
@frlazzeri **@mathew_jaya**

Strata
DATA CONFERENCE