

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

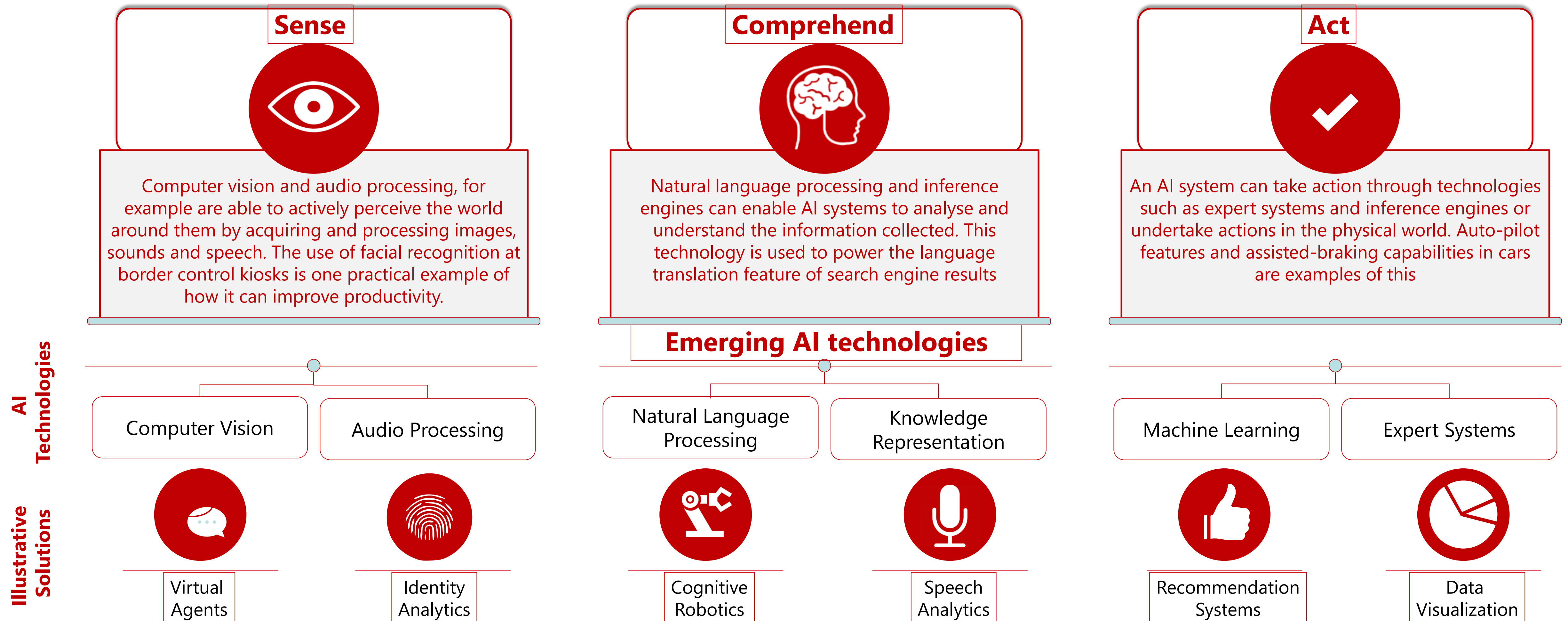


Agenda

- 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



Source: [Accenture: Why artificial intelligence is the future of growth](#), April 2016

#StrataData

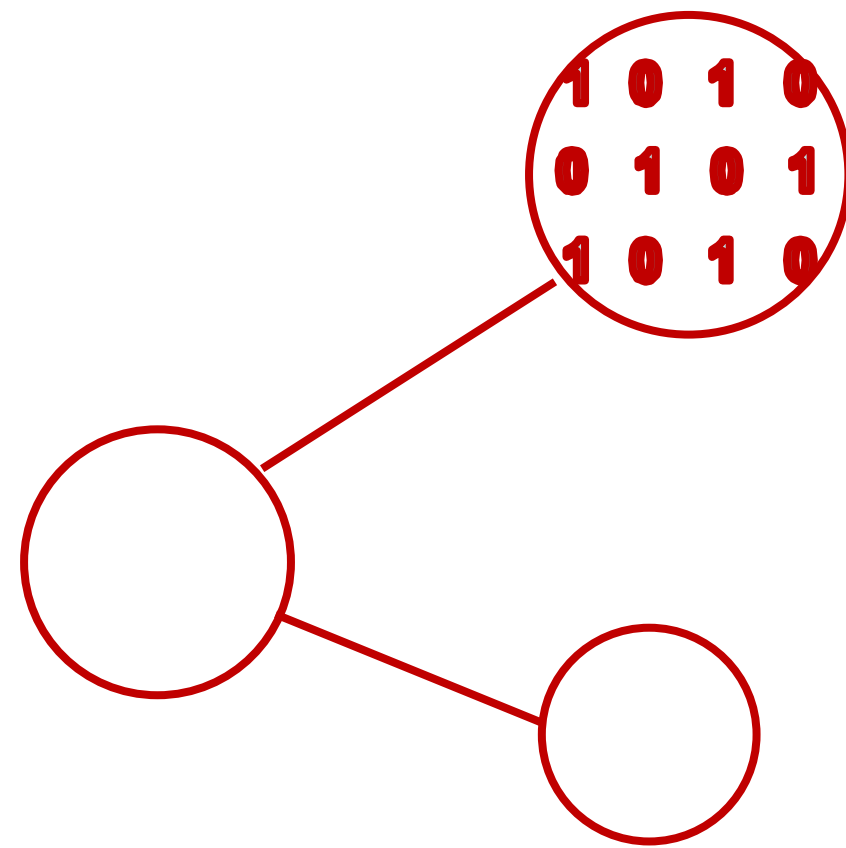
 @frlazzeri

 @mathew_jaya

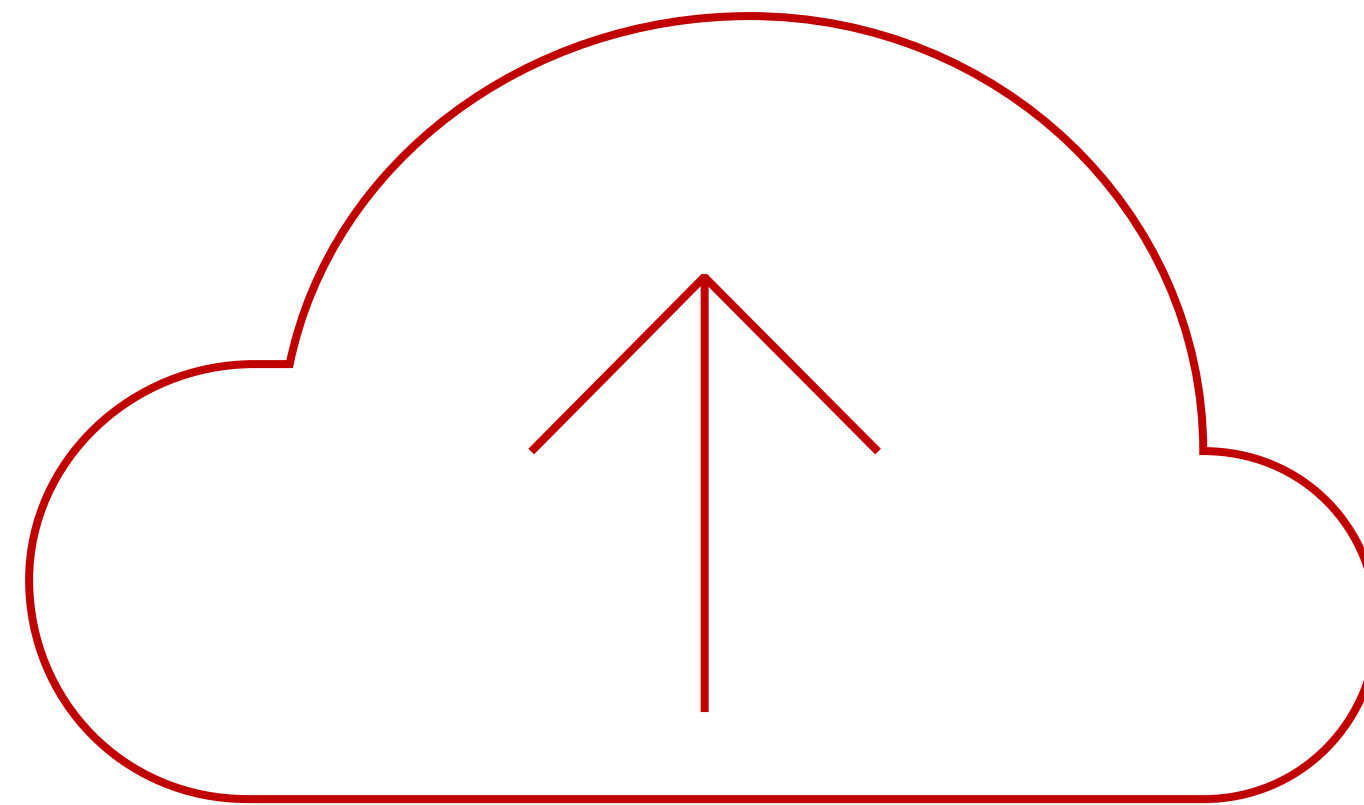
Strata
DATA CONFERENCE

Why AI?

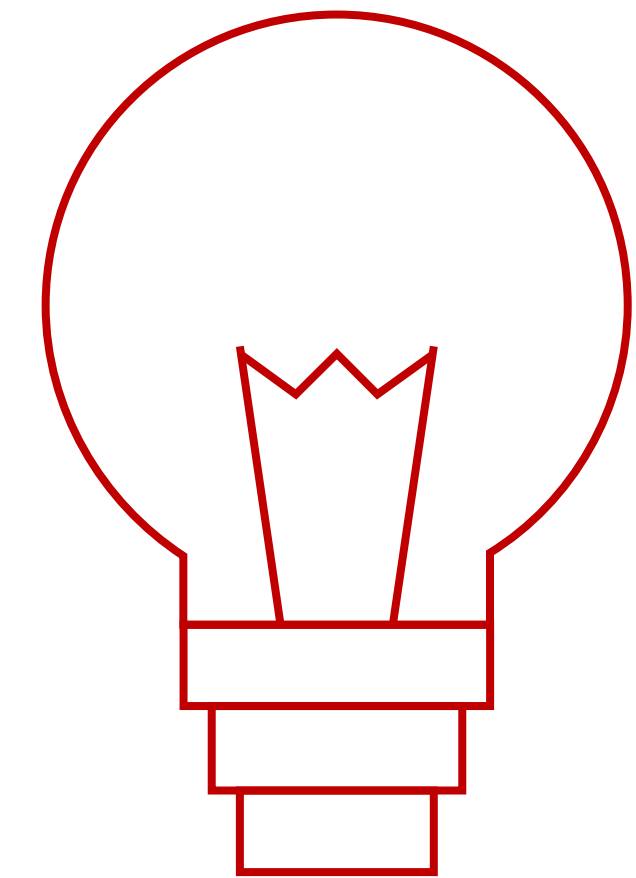
Three major trends are converging



Data

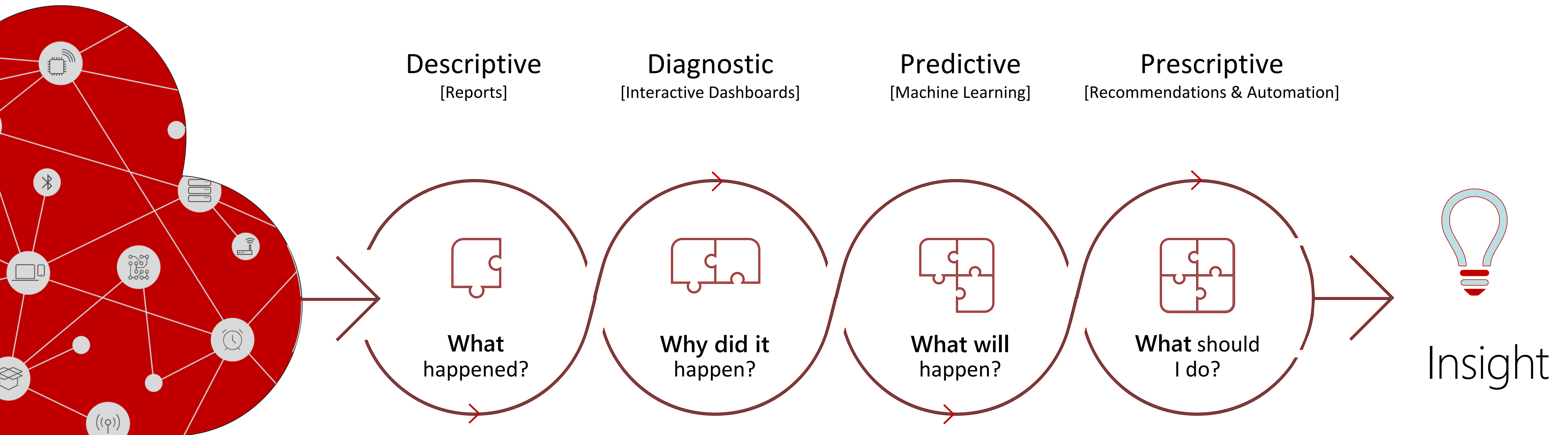


Cloud

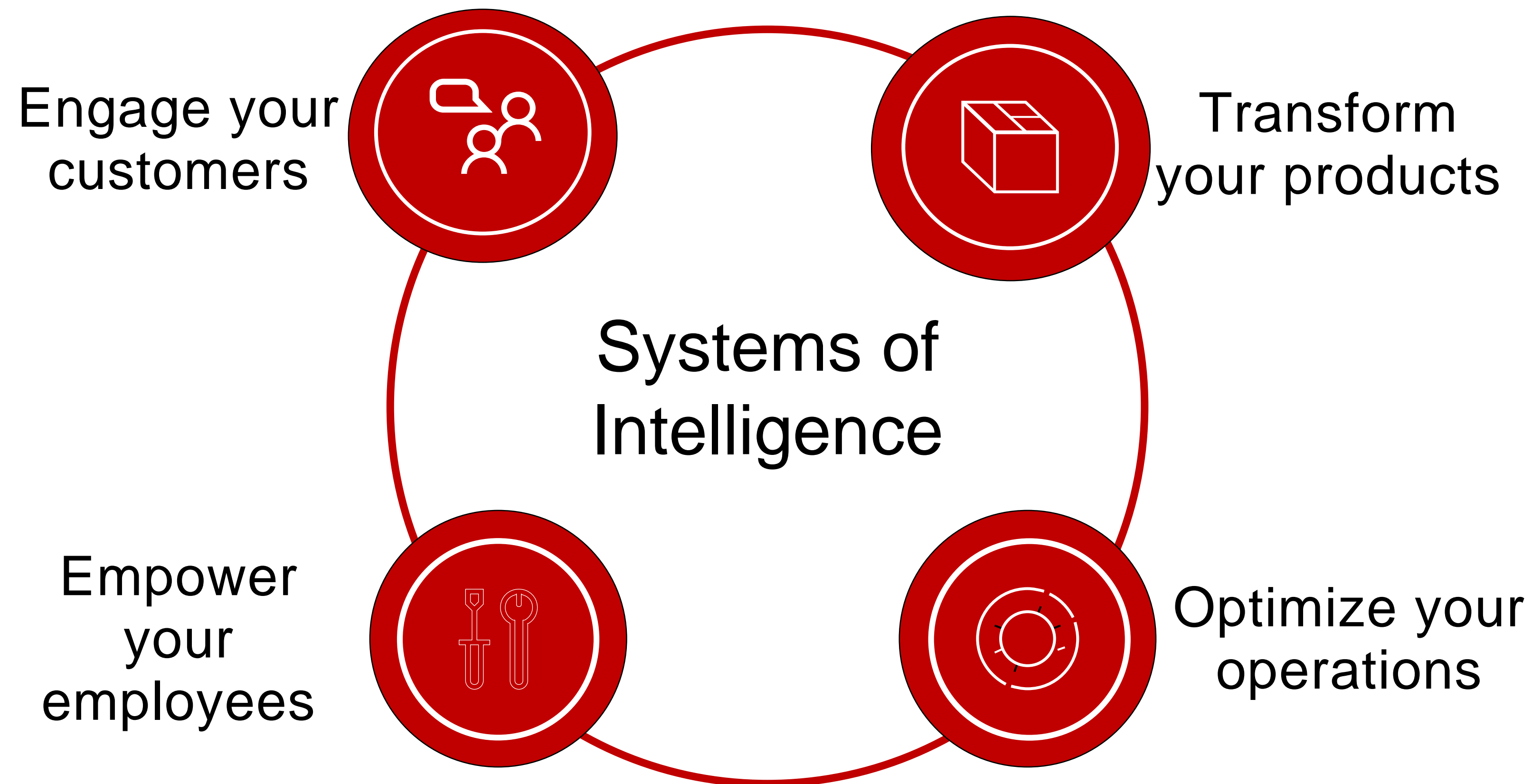


Intelligence

From data to decisions and actions



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

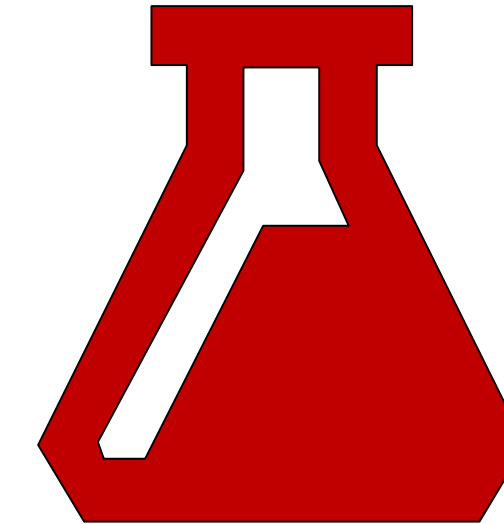
aka.ms/AICustomModels



Cognitive
Services



Custom Services
Bring your Own Data



Custom
AI Models



Flexibility

Being Obsessed with Data

Question
is sharp

E.g. Predict
whether
component X will
fail in the next Y
days

Data
measures
what you
care about

E.g. Identifiers at the
level you are
predicting, relevant
data collected &
feature engineering
using domain
knowledge

Data is
accurate

E.g. Failures are
really failures,
human labels on root
causes

Data is
connected

E.g. Machine
information linkable
to usage information

A lot of
data

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**

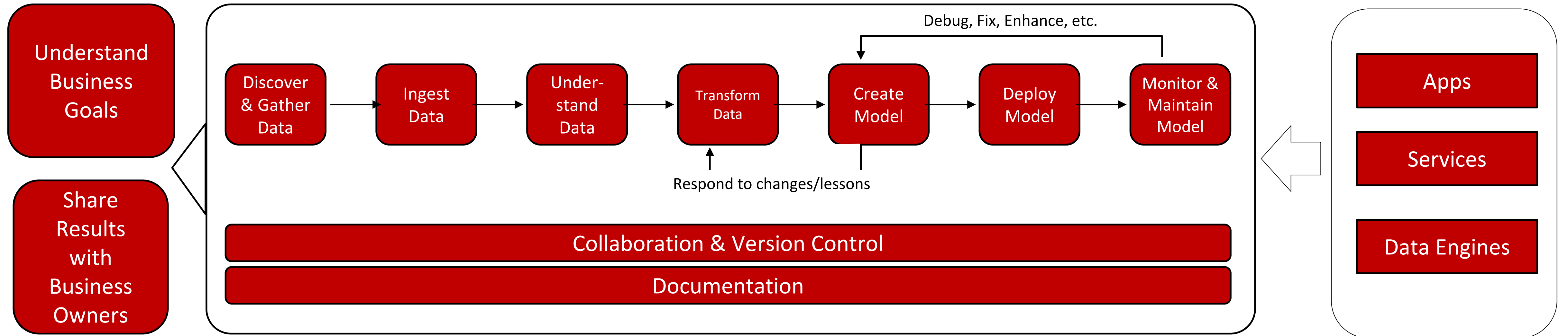
Quantify the
**Metric Value
Improvement**
useful for
customer scenario

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

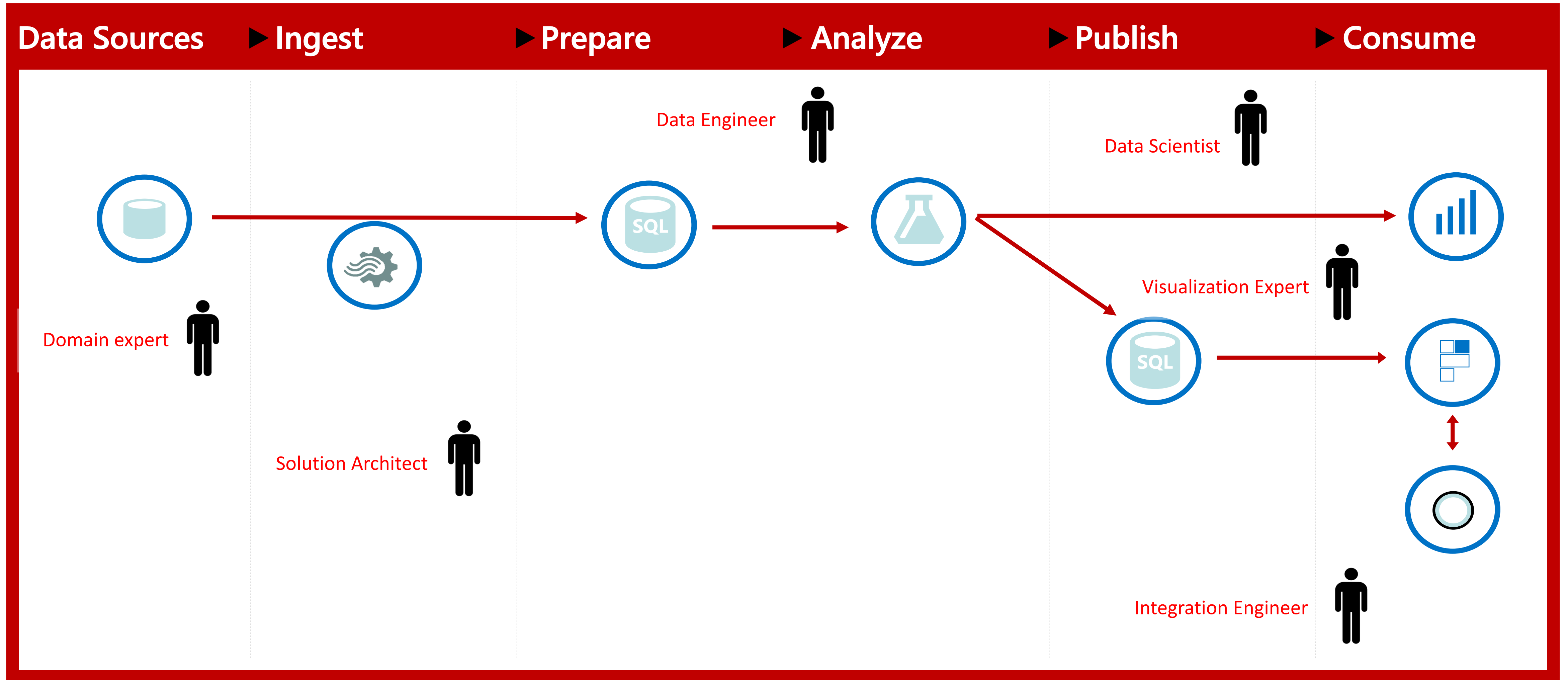
Consume a
model

Team Data Science Process

aka.ms/TeamDataScience

Project Manager 

Executive
Sponsorship –
IT & Business 




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




Loan Credit Risk with SQL Server

Using SQL Server 2016 with R Services, a lending institution can make use of predictive analytics to reduce number of loans they offer to those borrowers most likely to default, increasing the profitability of their loan ...

1.3K 343 24 days ago

Microsoft




Personalized Offers

In today's highly competitive and connected environment, modern businesses can no longer survive with generic, static online content. Furthermore, marketing strategies using traditional tools are often expensive, ...

4.6K 399 3 months ago

Microsoft




Campaign Optimization with SQL Server

This solution demonstrates how to build and deploy a machine learning model with SQL Server 2016 with R Services to recommend actions to maximize the purchase rate of leads targeted by a campaign.

8.1K 1.2K 2 hours ago

Microsoft




Campaign Optimization with Azure HDInsight Spark Clusters

This solution demonstrates how to build and deploy a machine learning model with Microsoft R Server on Azure HDInsight Spark clusters to recommend actions to maximize the purchase rate of leads targeted by a...

1.2K 157 18 days ago

Microsoft



Predicting Length of Stay in Hospitals

This solution enables a predictive model for Length of Stay for in-hospital admissions. Length of Stay (LOS) is defined in number of days from the initial admit date to the date that the patient is discharged from any ...

10K 1.2K 25 days ago

Microsoft




Demand Forecasting and Price Optimization

Pricing is recognized as a pivotal determinant of success in many industries and can be one of the most challenging tasks. Companies often struggle with several aspects of the pricing process, including accurately fo...

4.4K 670 3 months ago

Microsoft




Quality Assurance

Quality assurance systems allow businesses to prevent defects throughout their processes of delivering goods or services to customers. Building such a system that collects data and identifies potential problems alone...

1.9K 313 3 months ago

Microsoft




Telemetry Analytics

Super computers have moved out of the lab and are now parked in our garage! These cutting-edge automobiles contain a myriad of sensors, giving them the ability to track and monitor millions of events every second...

9.2K 1.4K 3 months ago

Microsoft




Demand Forecasting

Accurately forecasting spikes in demand for products and services can give a company a competitive advantage. This solution focuses on demand forecasting within the energy sector...

9.5K 1.6K 3 months ago

Microsoft



Predictive Maintenance

This Predictive Maintenance solution monitors aircraft and predicts the remaining useful life of aircraft engine components

11K 2.1K 3 months ago

Microsoft

Microsoft Learning

aka.ms/MicrosoftAlLearning

Microsoft Virtual Academy | Courses ▾ Search all courses 🔍

Free Microsoft training delivered by experts

Developers IT Pros Data Pros

Windows 10 Cloud Development Game Development Web Development Database Development

C# / XAML Visual Studio For Beginners Mobile App Development

Browse all developer courses ➔

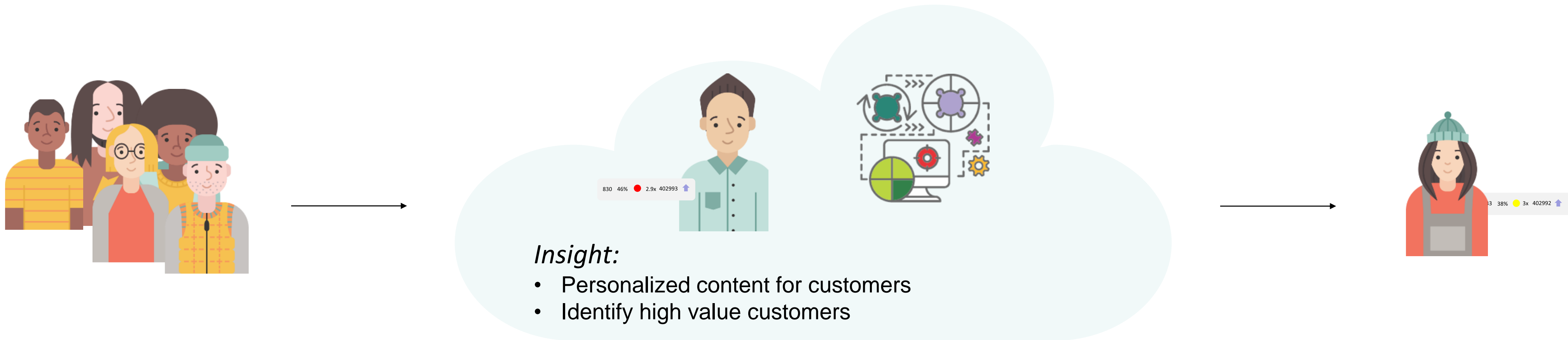
Learning Paths
Curated collections of courses to help you build skills. Complete a learning path to earn a badge you can share with others.

Learn more

MICROSOFT VIRTUAL ACADEMY
Microsoft
Complete a Learning Path to Earn this Badge
Learning Path Completed

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

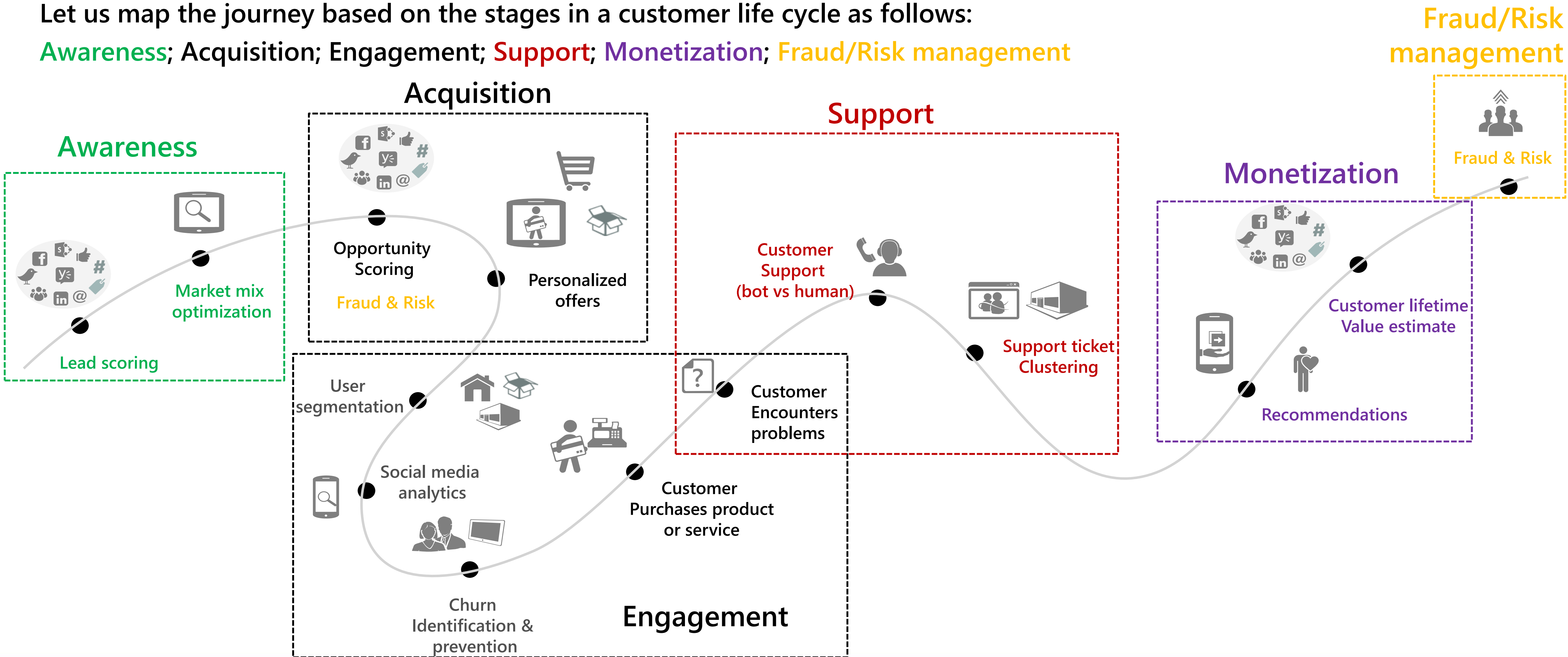
Intelligent Action

- 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

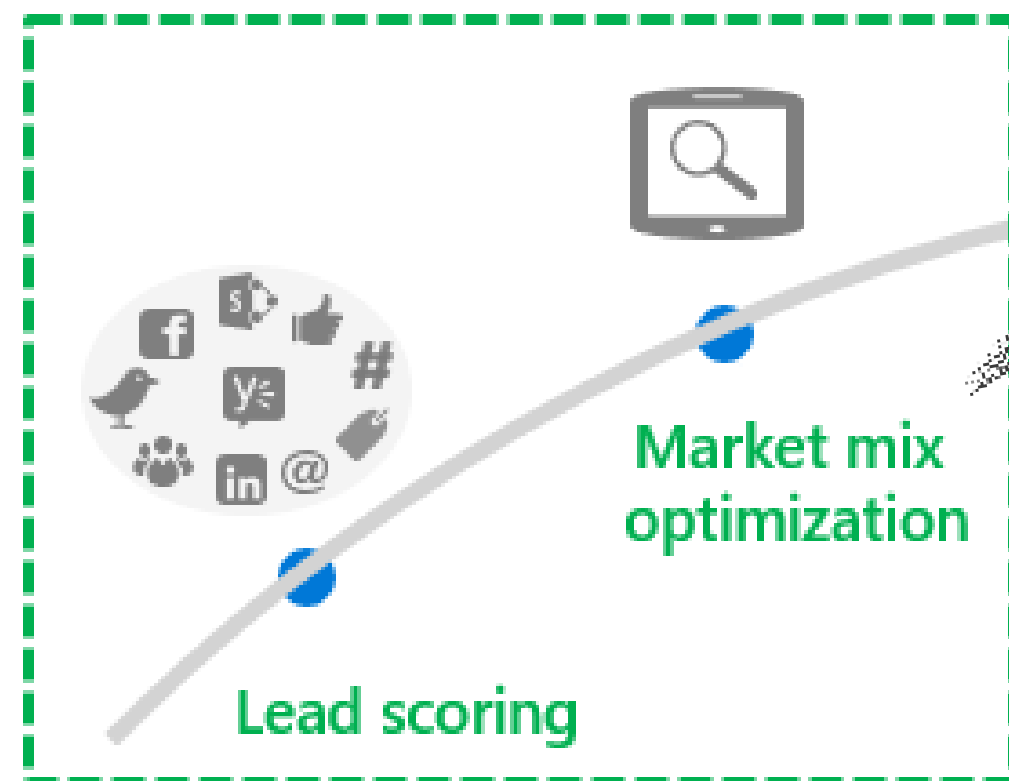
Let us map the journey based on the stages in a customer life cycle as follows:

Awareness; Acquisition; Engagement; **Support**; Monetization; **Fraud/Risk management**



Marketing cycle – Awareness & Acquisition

Awareness



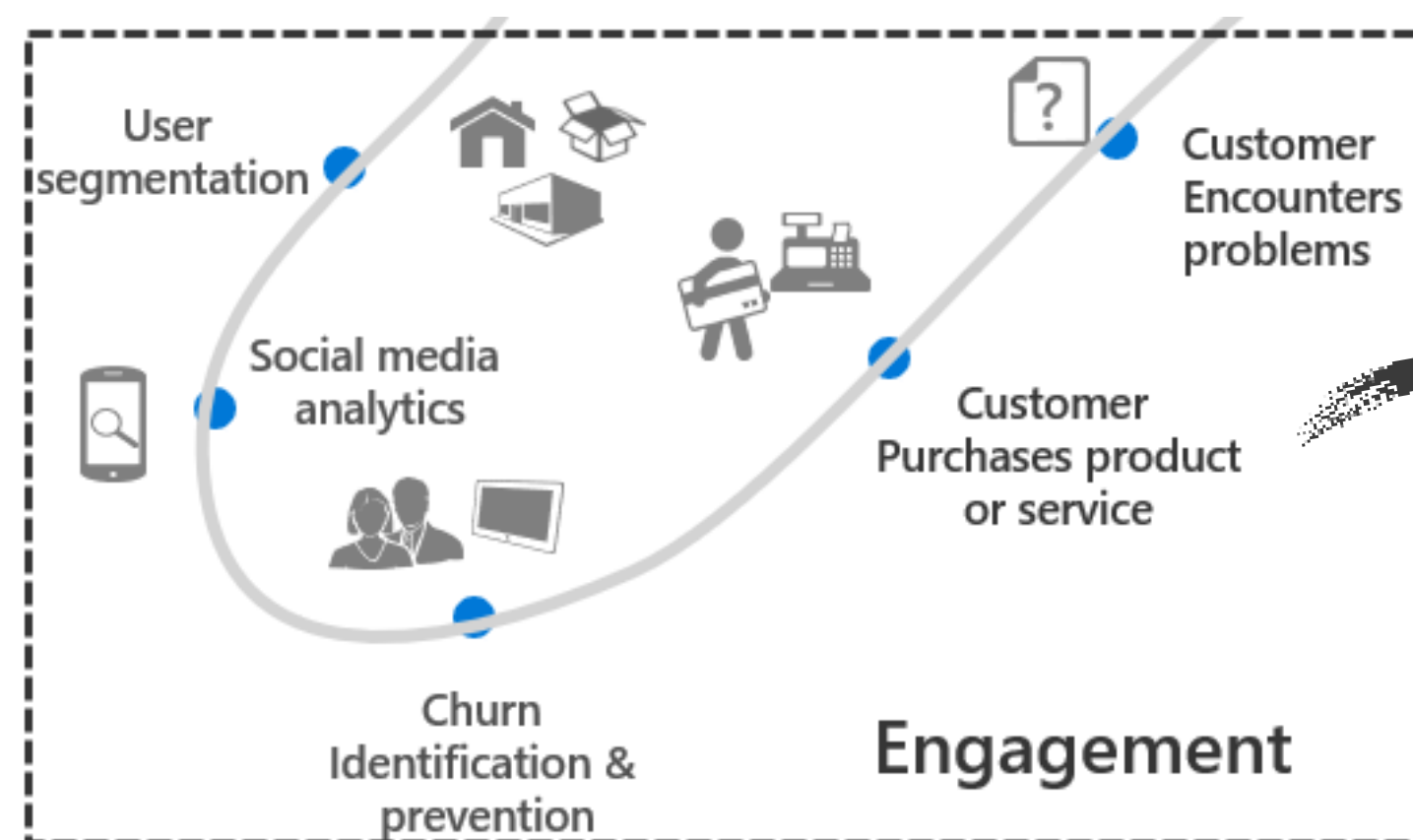
- ✓ Marketing mix optimization for omni-channel budget optimization
- ✓ ML based lead-scoring models

Acquisition

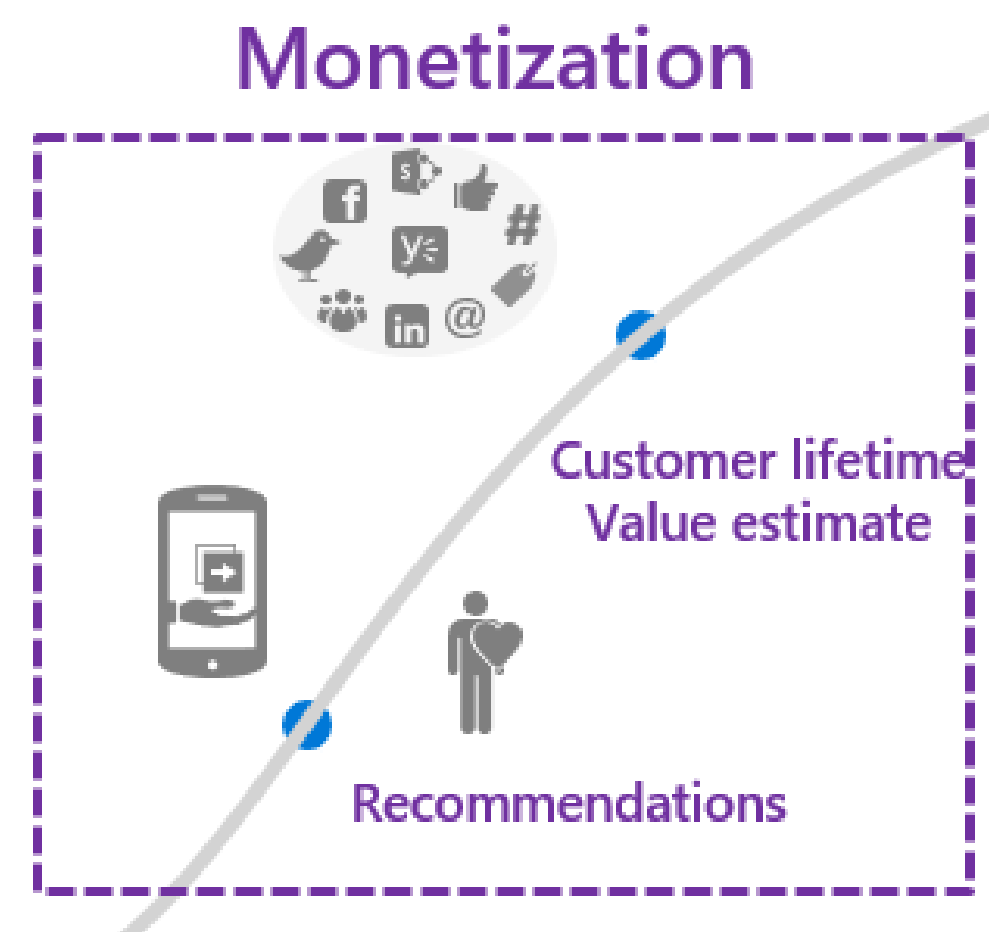


- ✓ Opportunity scoring can help target users who are most likely to make a purchase
- ✓ AI powered content creation

Marketing cycle – Engagement & Monetization



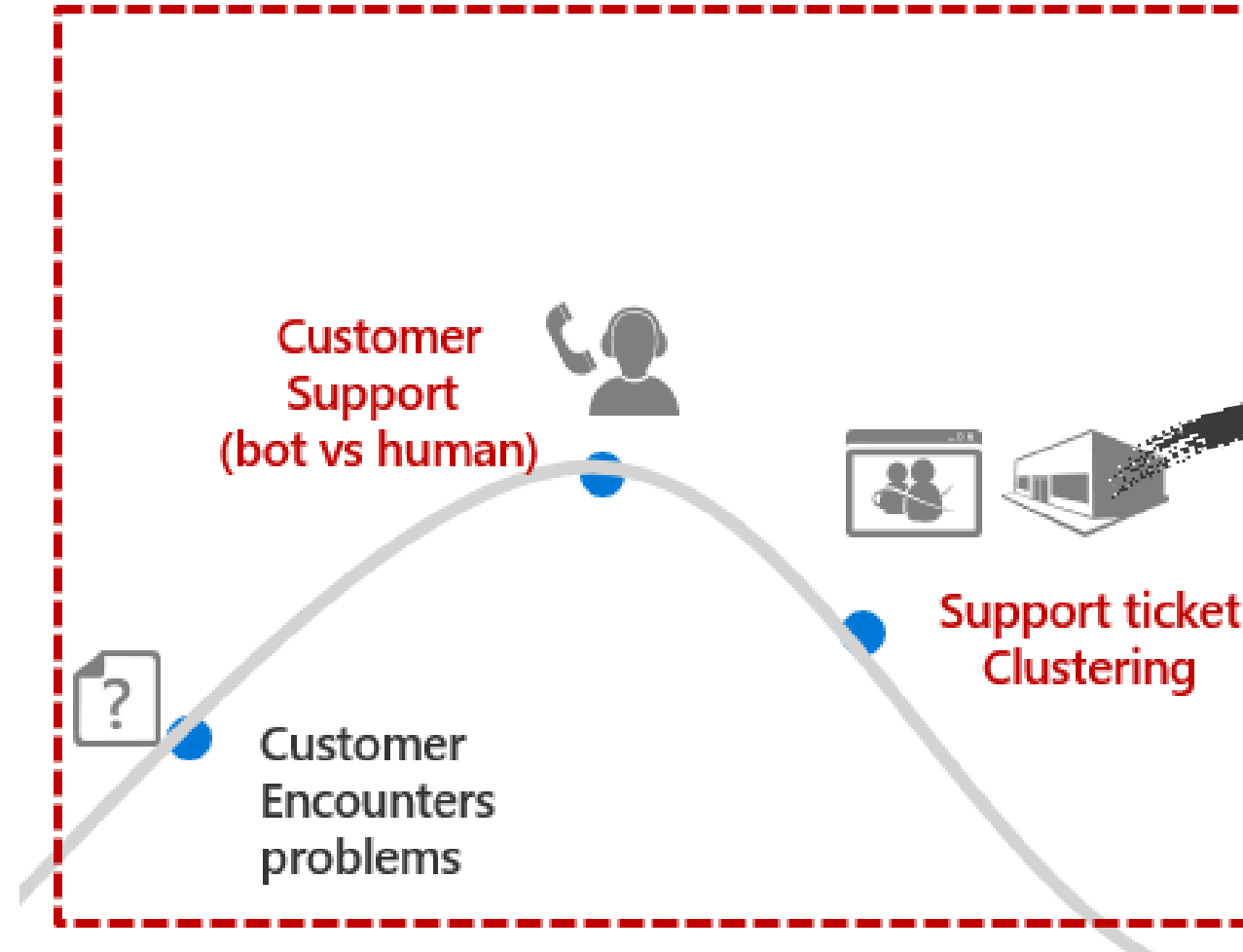
- ✓ Customer segmentation based on their browsing/purchase patterns
- ✓ ML based models can help determine customers who are most likely to churn in the near future based on their behavior



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

Marketing cycle – Support & Fraud

Support



- ✓ 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

Fraud/Risk management



- ✓ 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

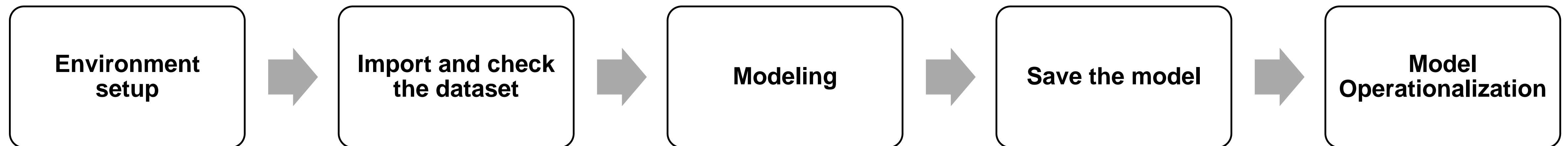
Credit Card Fraud Detection Data Set

- 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 $V1, V2, \dots, V28$: 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

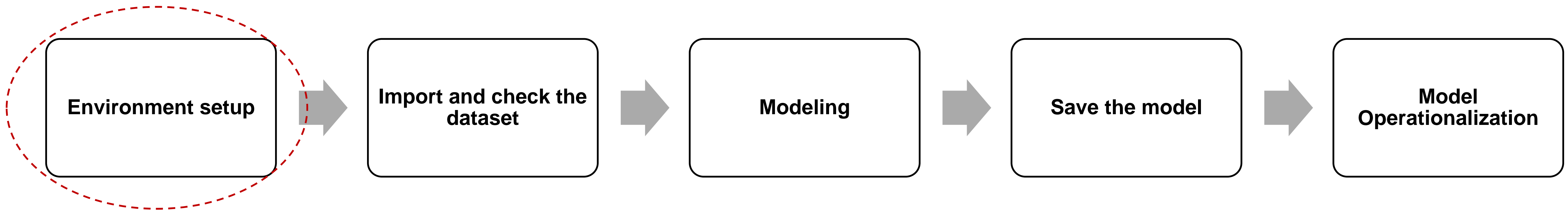
Time	V1	V2	V3	V4
0	-1.3598071336738	-0.0727811733098497	2.53634673796914	1.37815522427443
0	1.19185711131486	0.26615071205963	0.16648011335321	0.448154078460911
1	-1.35835406159823	-1.34016307473609	1.77320934263119	0.379779593034328
1	-0.966271711572087	-0.185226008082898	1.79299333957872	-0.863291275036453
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

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

The screenshot shows a Jupyter Notebook interface with the title 'Strata_DL_Sept2018'. The notebook contains three code cells under the heading 'Environment setup'. The first cell imports basic modules like os, keras, shutil, and json. The second cell imports data manipulation and machine learning libraries like pandas, numpy, datetime, sklearn, and keras. The third cell imports tensorflow, sklearn, keras, and other modules for model training and evaluation.

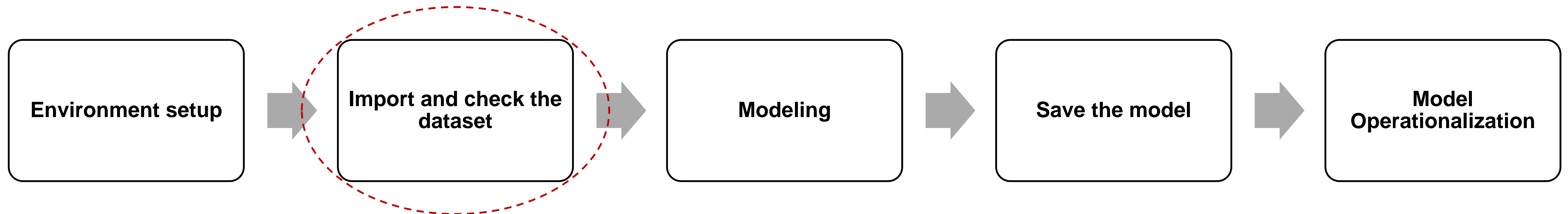
```
In [ ]: # Import necessary components
import os
import keras
import shutil
import json

In [ ]: import re
import pandas as pd
import numpy as np
import datetime

from sklearn import preprocessing
from sklearn.metrics import confusion_matrix, recall_score, precision_score
from keras.models import Sequential
from keras.layers import Dense, Dropout, LSTM, Activation
from math import ceil

In [ ]: import pickle
from scipy import stats
import tensorflow as tf
from sklearn.model_selection import train_test_split
from keras.models import Model, load_model
from keras.layers import Input, Dense
from keras.callbacks import ModelCheckpoint, TensorBoard
from keras import regularizers
```

Demo



- Import the dataset and check the distributions of the variables

jupyter Strata_DL_Sept2018 Last Checkpoint: 2 minutes ago (autosaved) Python [conda env:py35] Logout

File Edit View Insert Cell Kernel Widgets Help Trusted

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.

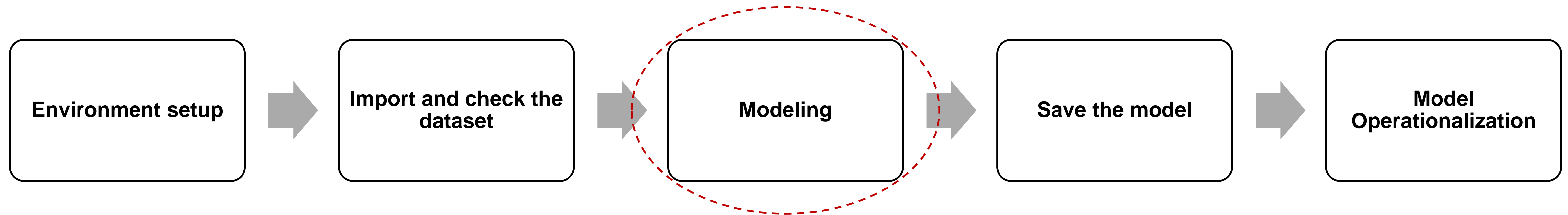
```
In [ ]: # Check data statistics
print(cc.describe())
```

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) Python [conda env:py35]

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)

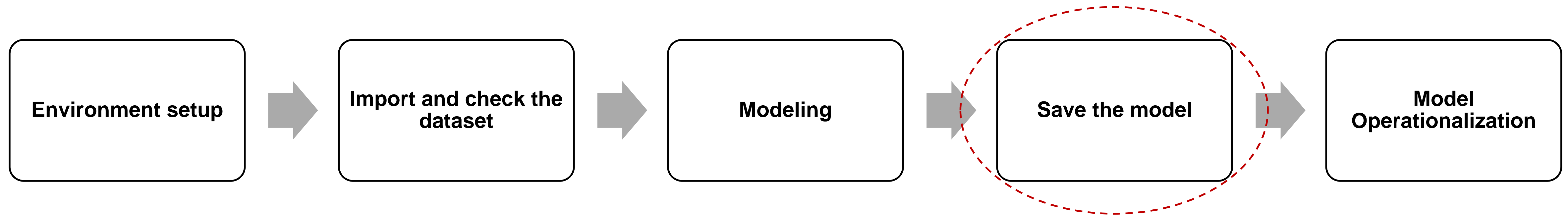
In [ ]: X_test[1]

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]

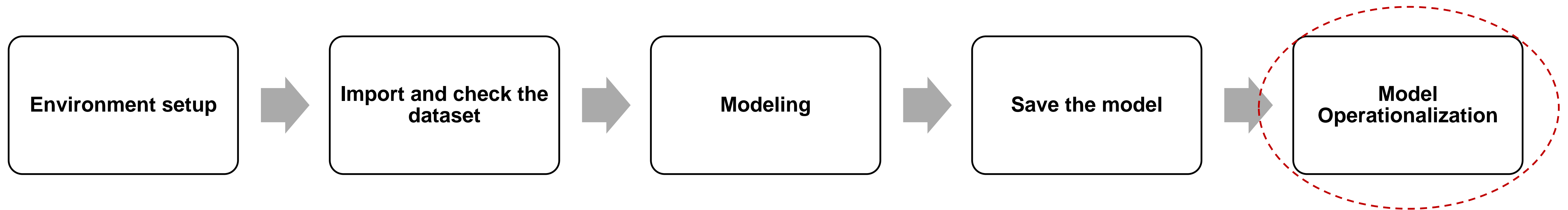
Saving the model

```
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

jupyter Strata_DL_Sept2018 Last Checkpoint: 4 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python [conda env:py35]

Deployment

There are multiple options to operationalize a model, this is entirely dependent on the tools used.

Once the assets (model, schema file, scoring script etc.) are stored, we can download them into a deployment compute context for operationalization on an Azure web service. For this scenario, we will deploy this on our local context. We demonstrate how to setup this web service this through a CLI window opened in the AML.

Create a model management endpoint

Create a modelmanagement under your account. We will call this automodelmanagement. The remaining defaults are acceptable. `az ml account modelmanagement create --location <ACCOUNT_REGION> --resource-group <RESOURCE_GROUP> --name automodelmanagement` You can find the subscription name or subscription id through the (<https://portal.azure.com>) under the resource group you'd like to use.

Check environment settings

Show what environment is currently active: `az ml env show`

If nothing is set, we setup the environment with the existing model management context first: `az ml env setup --location <ACCOUNT_REGION> --resource-group <RESOURCE_GROUP> --name automodelmanagement` using the same and in the previous section.

Then set the current environment: `az ml env set --resource-group <RESOURCE_GROUP> --cluster-name automodelmanagement`

Check that the environment is now set: `az ml env show`

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/@curiously/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