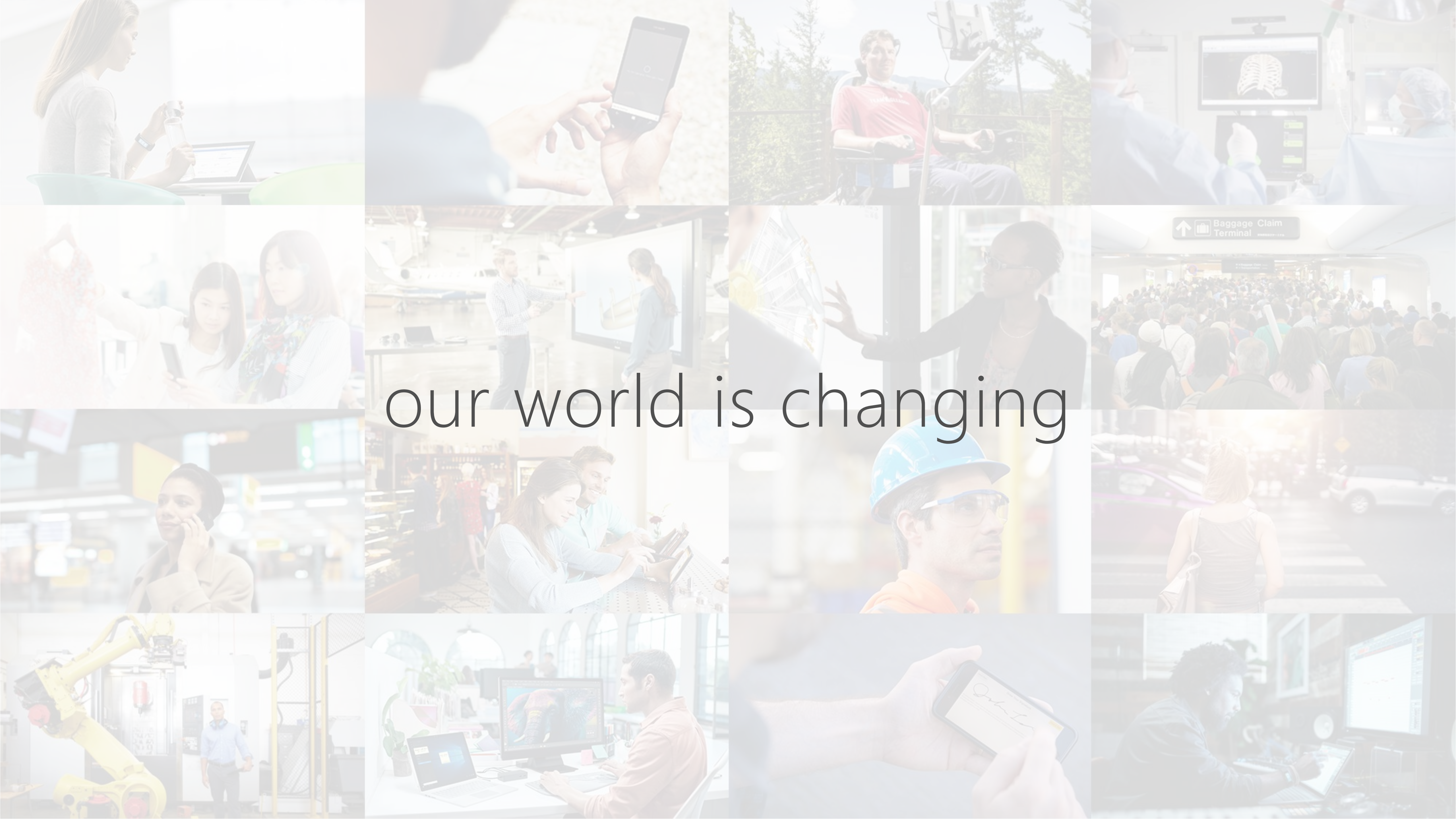
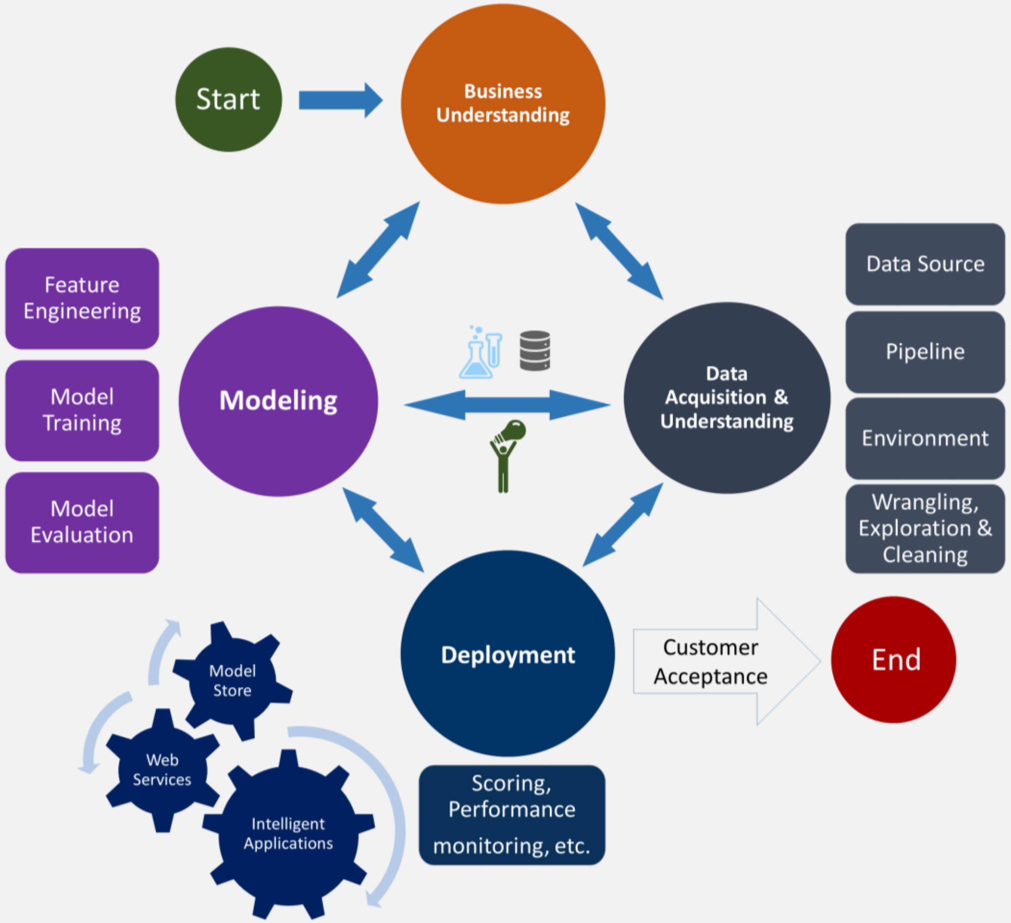
Model Interpretability and Data Drift

Nicholas Moore



our world is changing

Data Science Lifecycle



Team Data Science Process: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview there is a lot of hype around model creation, but not model maintenance

The moment you put a model in production, it starts degrading.

Covariate Shift

Changes in the distribution of independent variables

Concept Drift

Changes in the relationship between independent and target variables

Model Interpretability

Understanding the process a model uses to arrive at an outcome

MACHINE LEARNING BASICS

SUPERVISED MACHINE LEARNING

Classification

Categorical outputs

Regression

Continuous outputs

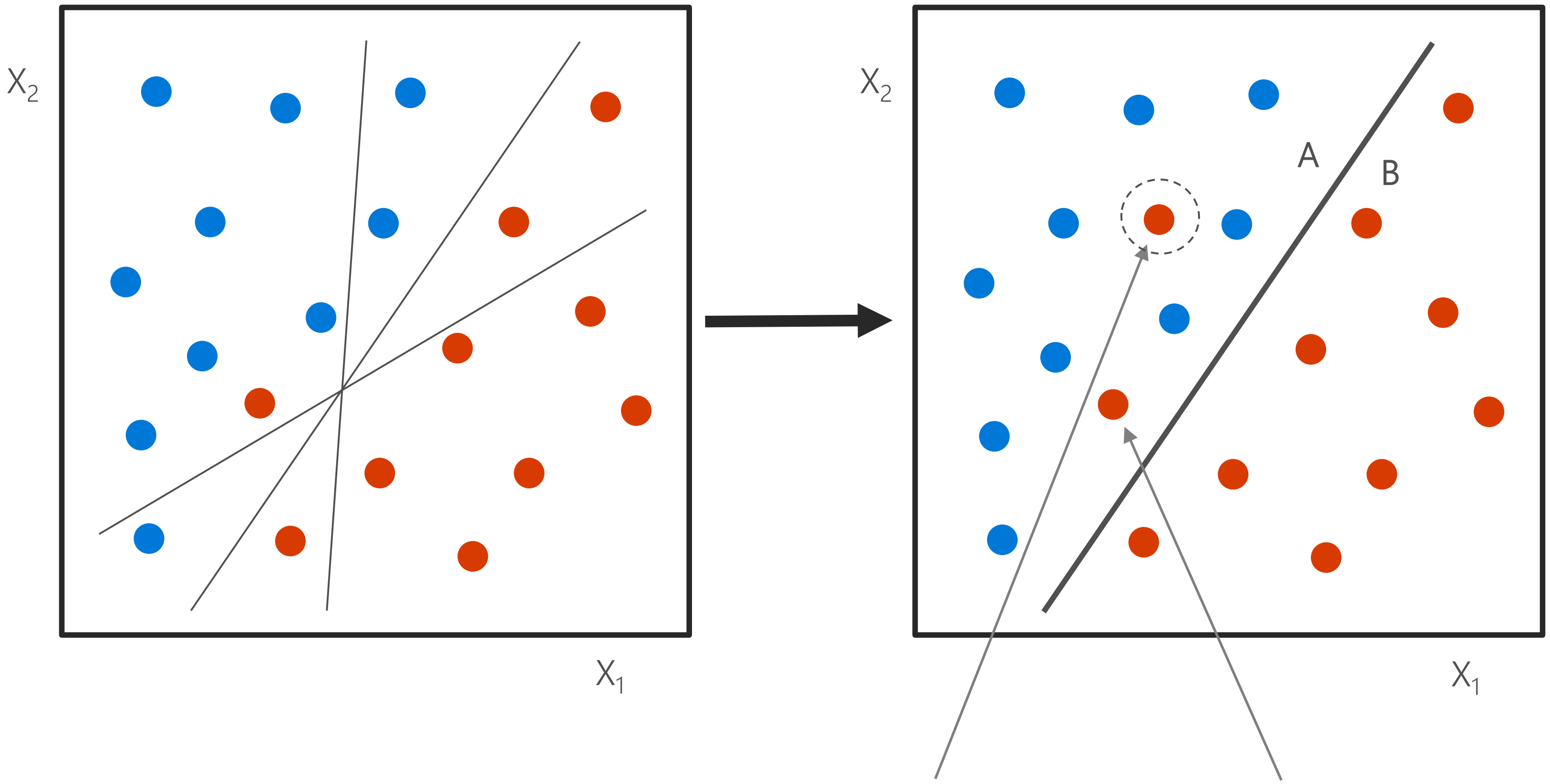
Reproduce outputs from a training data set by
creating a rule that maps inputs to outputs

MODEL CREATION PROCESS

CLASSIFICATION

Iterativly find the best boundry to separate classes

Use boundry to classify new data into a class

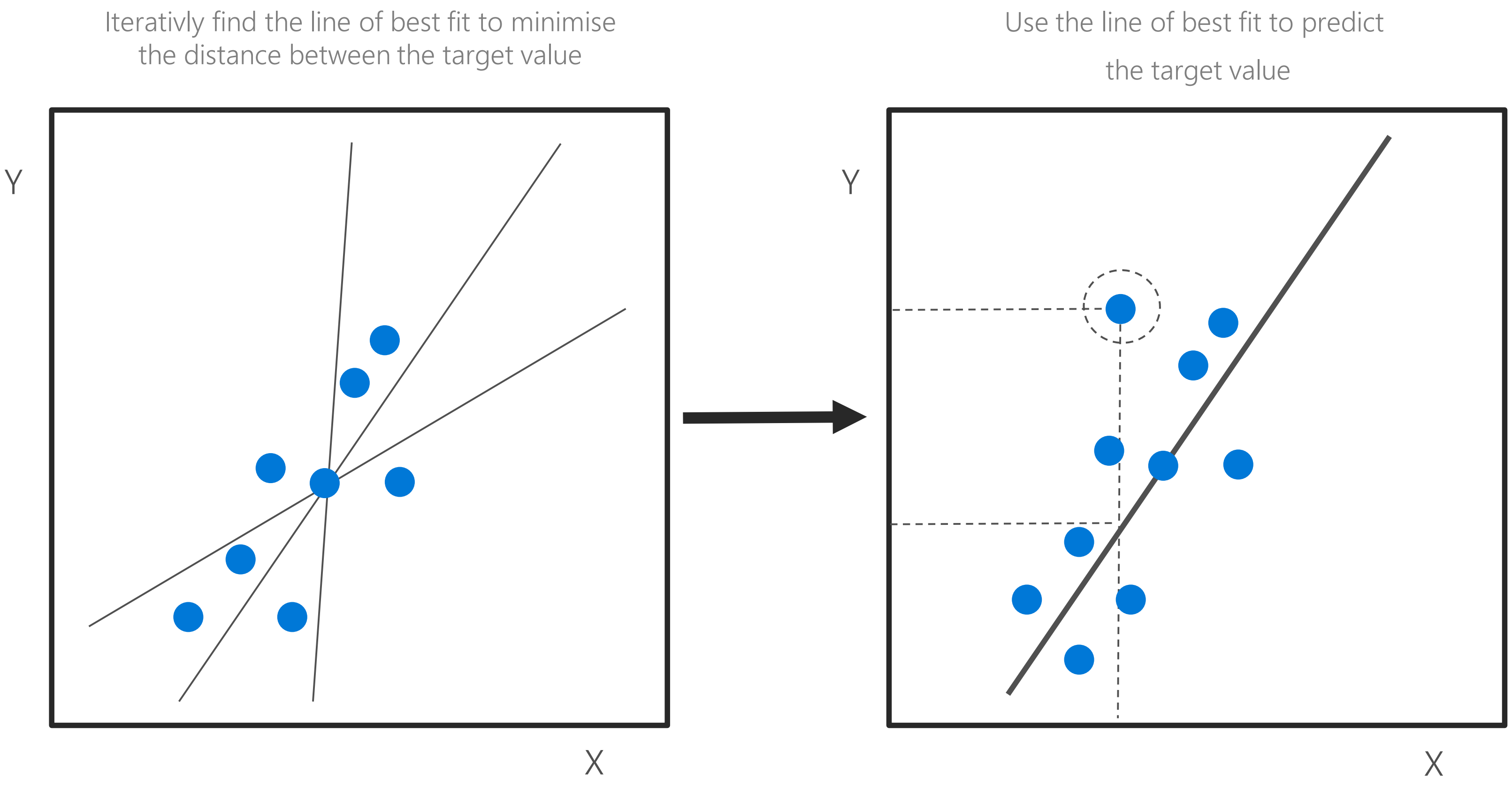


New data point classified as class A

Class B incorrectly classified as class A

MODEL CREATION PROCESS

REGRESSION


Iterativly find the line of best fit to minimise the distance between the target value

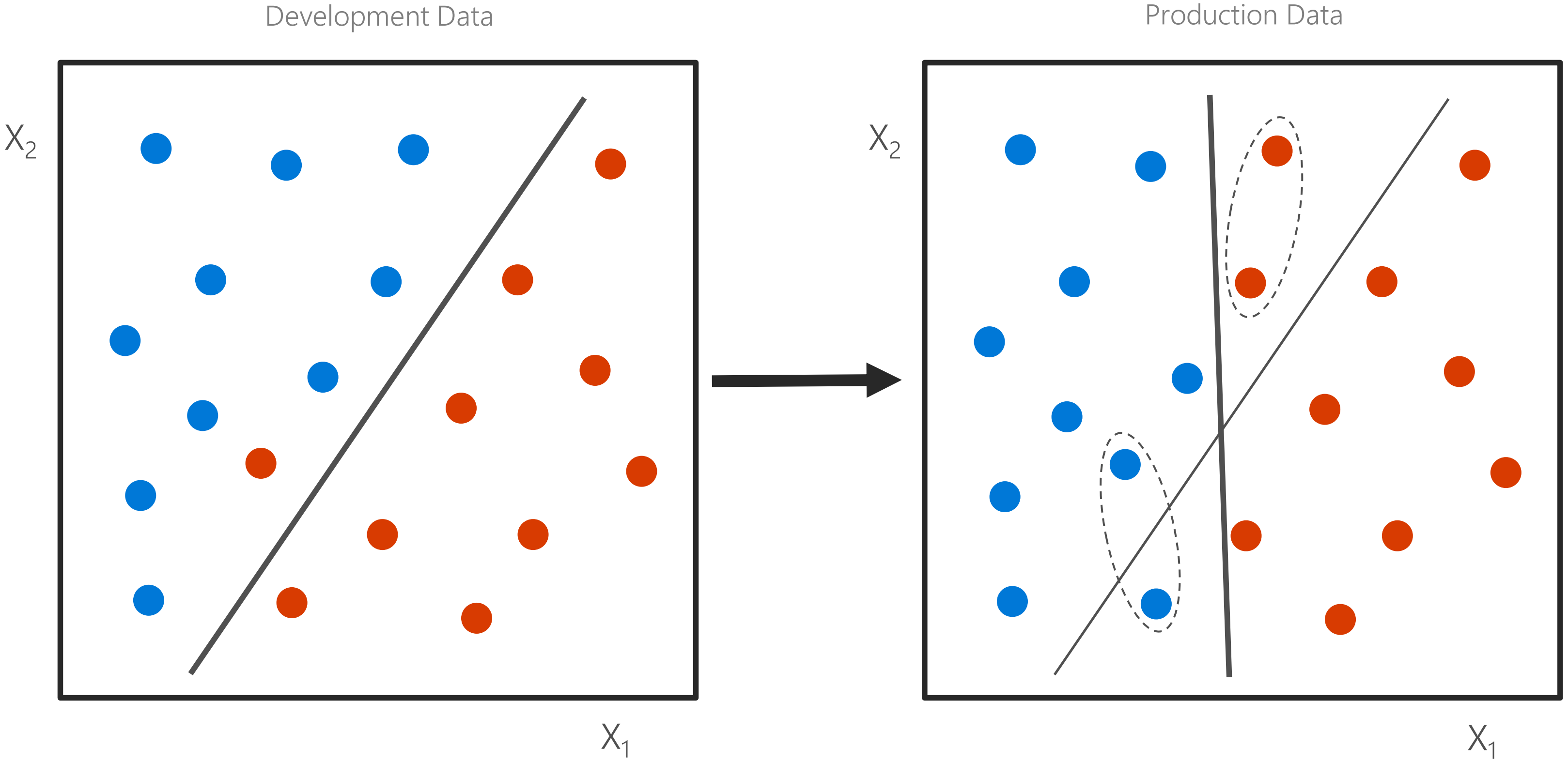
Use the line of best fit to predict

the target value

DATA DRIFT

CONCEPT DRIFT

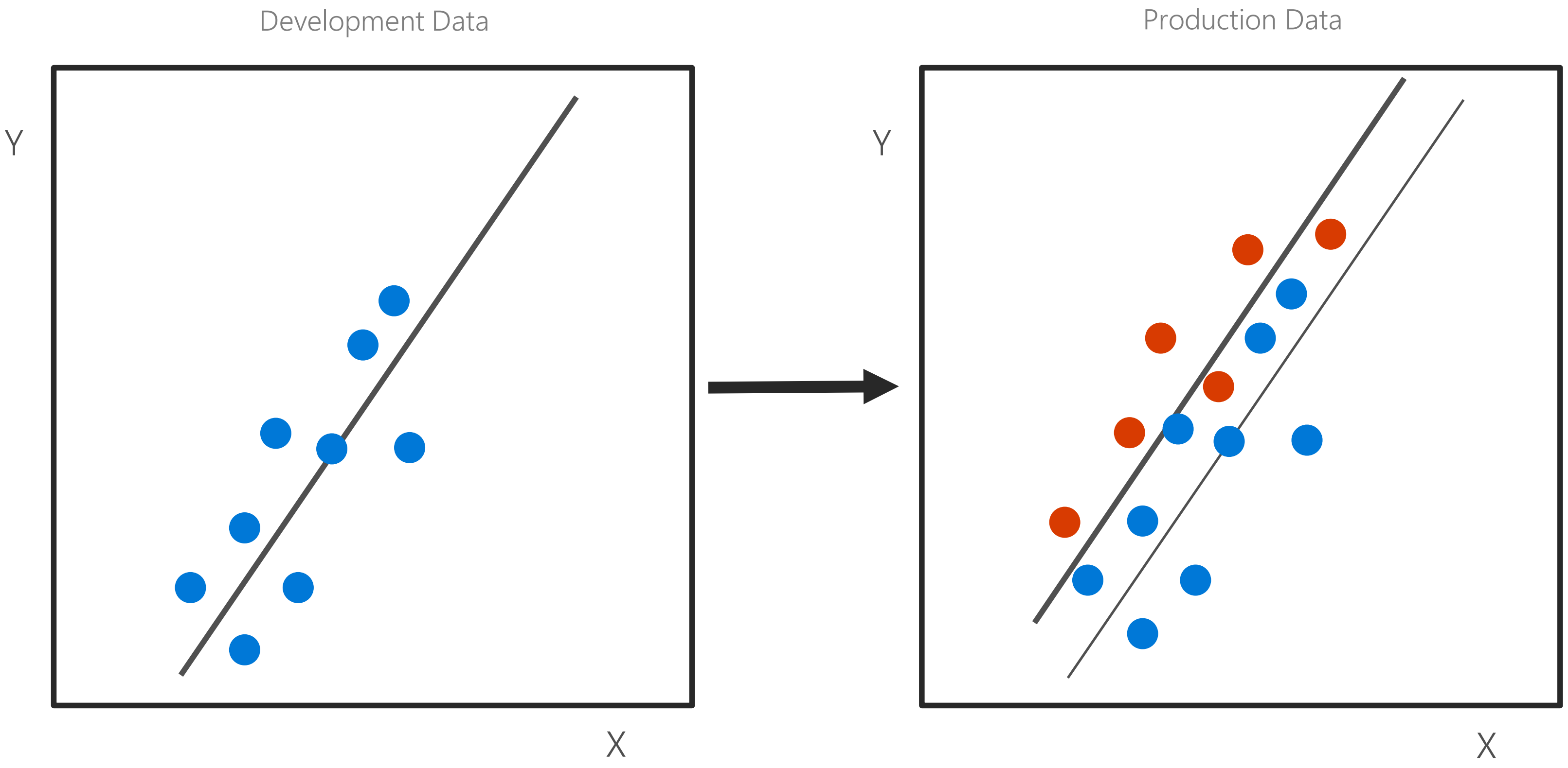
CLASSIFICATION



Change in the relationship between independent
and target variables in the underlying problem overtime

CONCEPT DRIFT

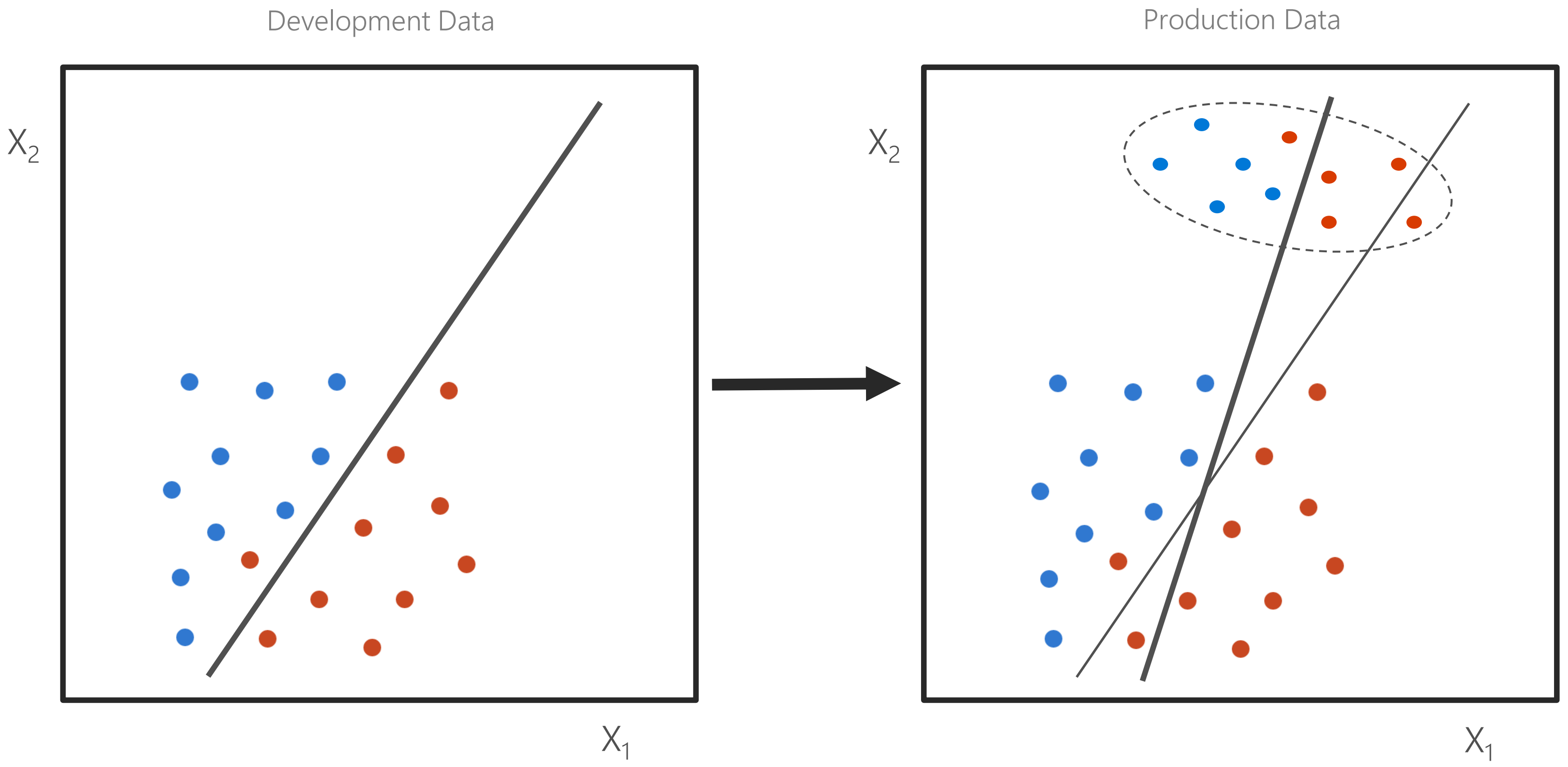
REGRESSION



Change in the relationship between independent and target variables in the underlying problem overtime

COVARIATE SHIFT

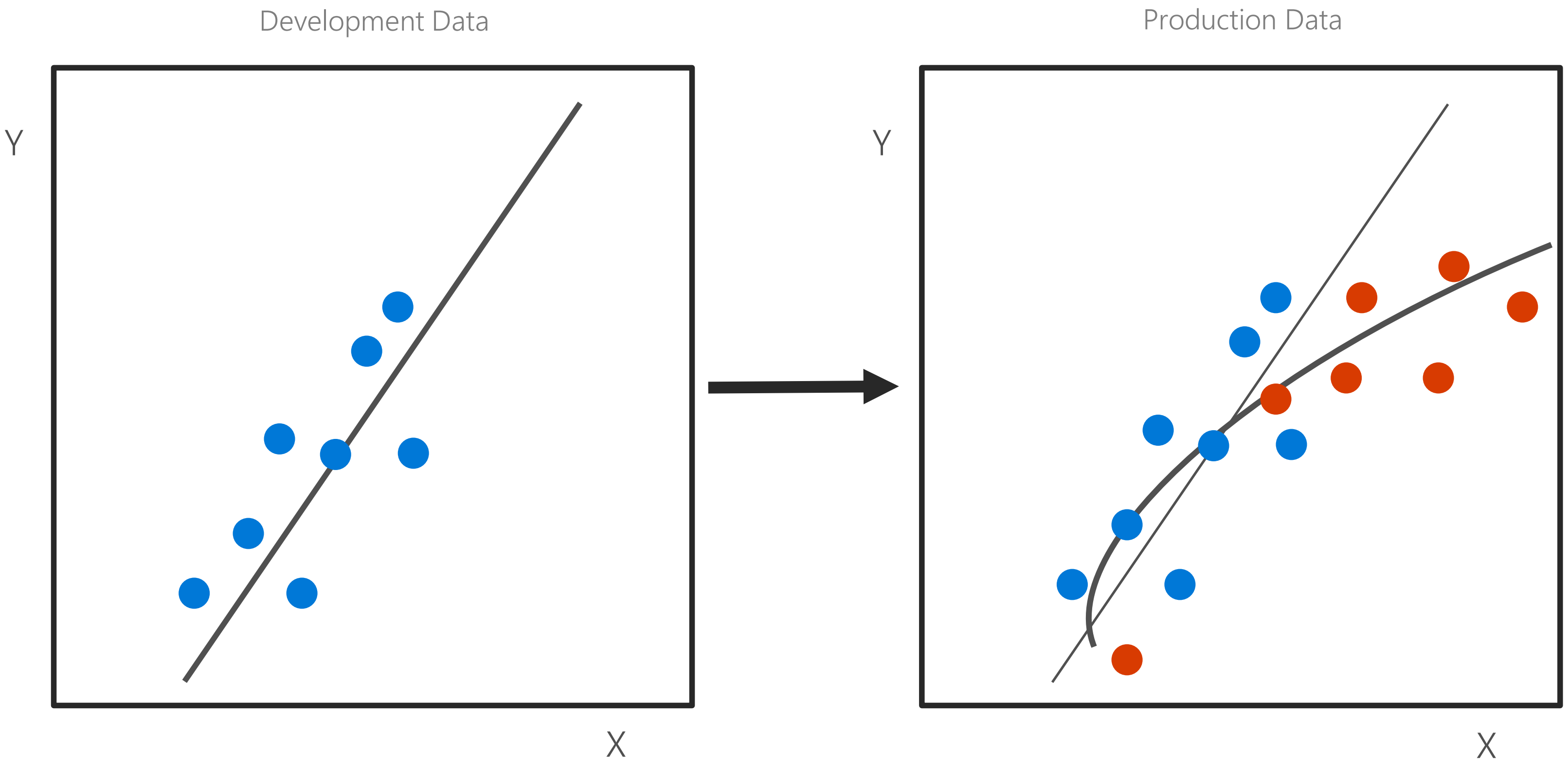
CLASSIFICATION



Change in the distribution of the independent variables in the underlying problem overtime

COVARIATE SHIFT

REGRESSION



Change in the distribution of the independent variables in the underlying problem overtime

Why is this a problem?

- ⚠ Production model no longer fit-for-purpose
- ⚠ Can be difficult to detect
- ⚠ Requires models to be monitored in production
- ⚠ Changes can be gradual, cyclical, or abrupt
- ⚠ Increased model maintenance

How to address the problem?

- ✓ Periodically re-fit or update the model
- ✓ Monitor distribution of independent variables in production
- ✓ Weight data
- ✓ Periodically assess the performance of your predictions
- ✓ Learn the change

EXAMPLES OF MODEL DRIFT

CONCEPT DRIFT

Identifying patterns of people who
commit fraud or hacking computer networks

New regulations introduced to save electricity
which influences predicting demand

New production procedures which
impact models designed to assess
quality of products / produce

COVARIATE SHIFT

Changes in common words or
meaning of works overtime

Ware and tear of sensors over time in
predictive maintenance solutions

Different accents in
speech-to-text solutions

MODEL INTERPRETABILITY

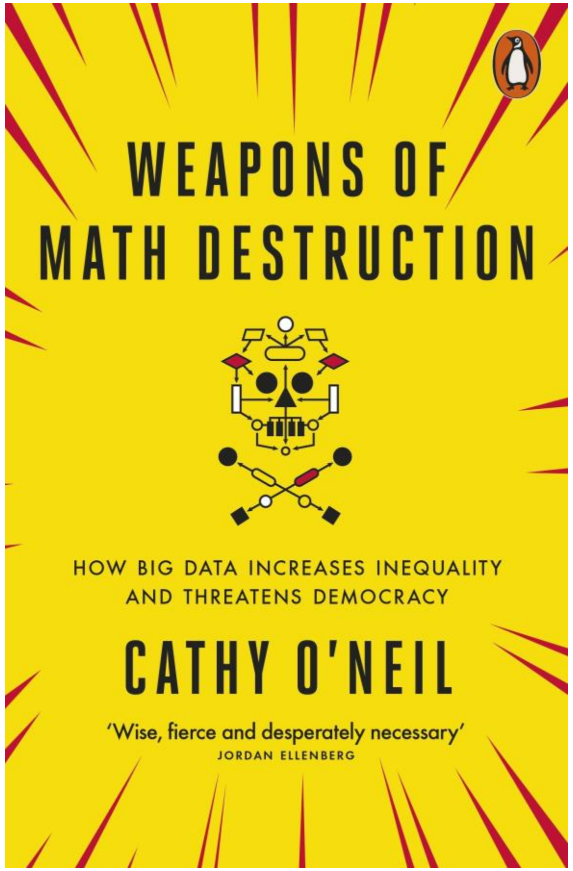
Fair and Accountable ML

What outcomes do machine learning models influence:

Will a person get a loan or credit card from a lending institution?

Will a teacher be fired based on their teaching evaluation score?

Will a prisoner be released on parole based on likelihood to reoffend?


WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
'Wise, fierce and desperately necessary'
JORDAN ELLENBERG

Why is it important?

Model interpratability aims to understand the process a machine learning model uses to make predictions

- ✓ Feature causality
- ✓ Inform data collection and feature engineering
- ✓ Model debugging
- ✓ Building trust with business stakeholders
- ✓ Regulation and auditability

INTERPRETABILITY OF ALGORITHMS

WHITE-BOX MODELS

Simpler computation
Less predictive capability
Easier to understand

Linear Regression
Logistic Regression
Decision Trees
Naïve Bayes
K-Nearest Neighbors

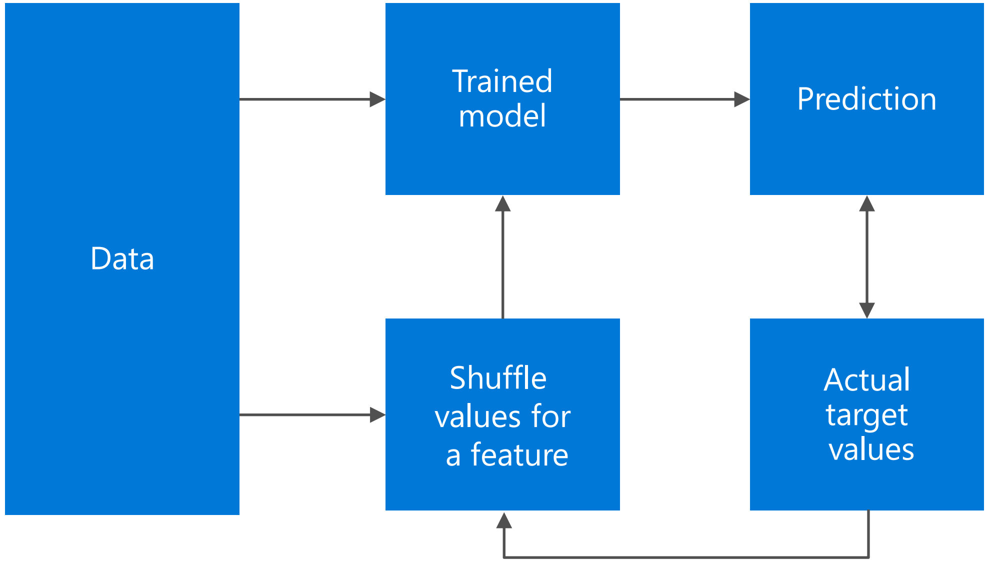
BLACK-BOX MODELS

High computational complexity
Emphasis on predictive capability
Difficult to understand

Ensembles
Kernel-based SVM
Neural Networks
Deep Learning

Model Interpretability Techniques

PERMUTATION IMPORTANCE



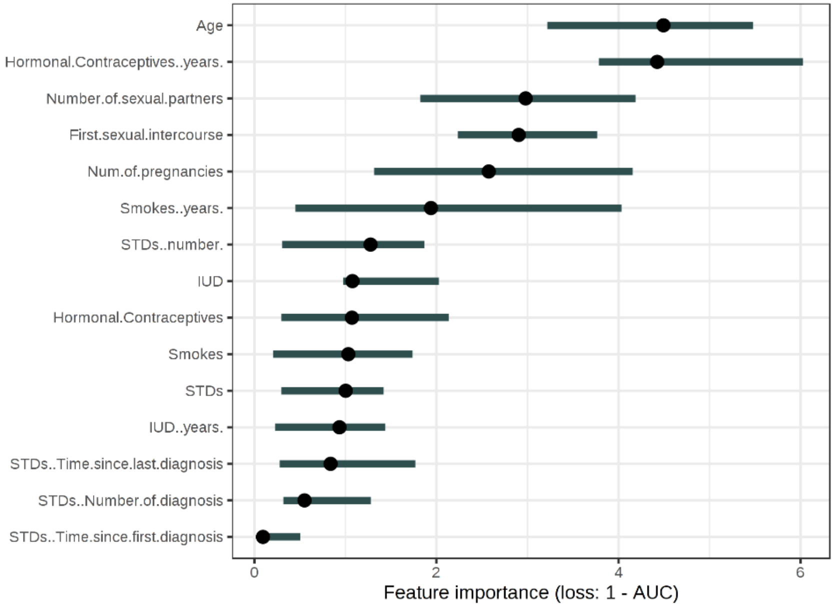
Data

Trained model

Prediction

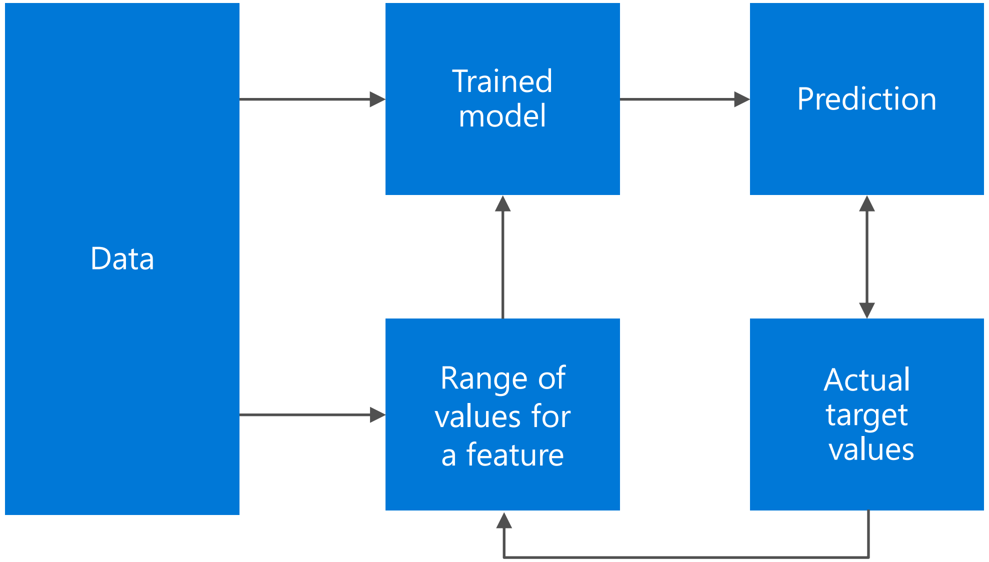
Shuffle values for a feature

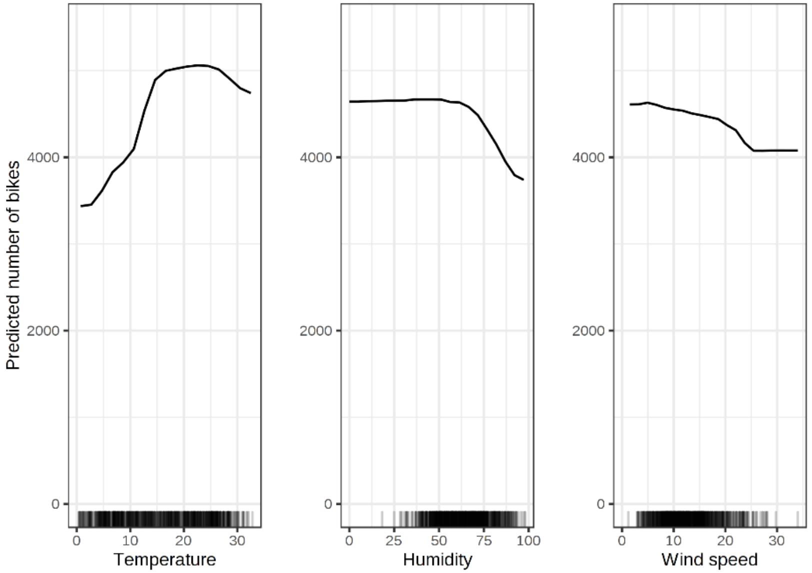
Actual target values



Model Interpretability Techniques

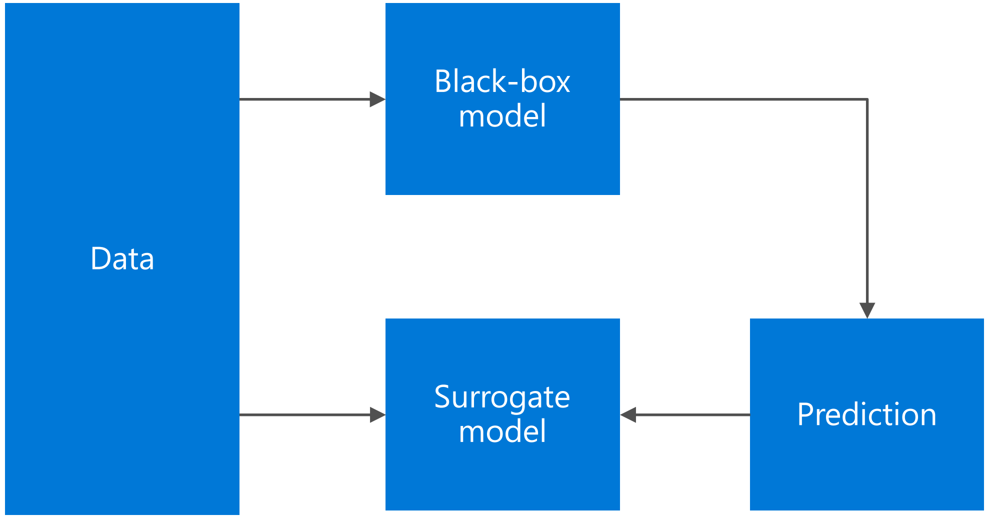
PARTIAL DEPENDENCE PLOT

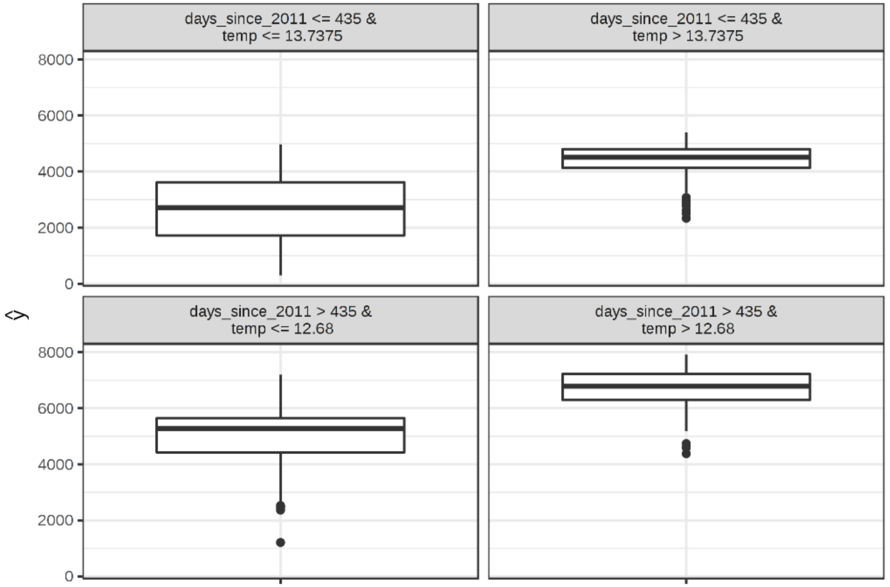




Model Interpretability Techniques

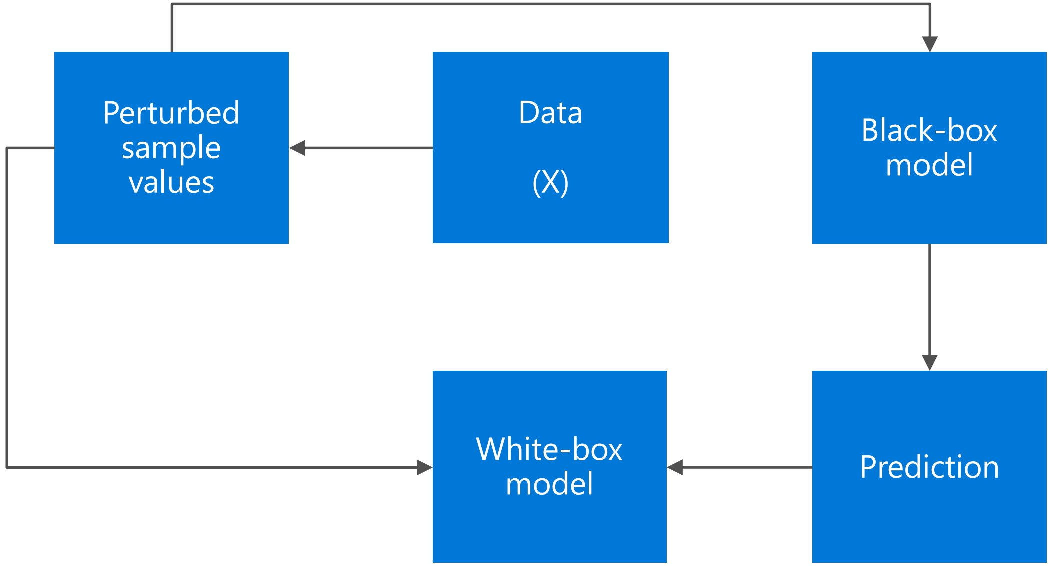
GLOBAL SURROGATE MODELS





Model Interpretability Techniques

LOCAL INTERPRETABLE MODEL AGNOSTIC EXPECTATIONS (LIME)


Perturbed sample values

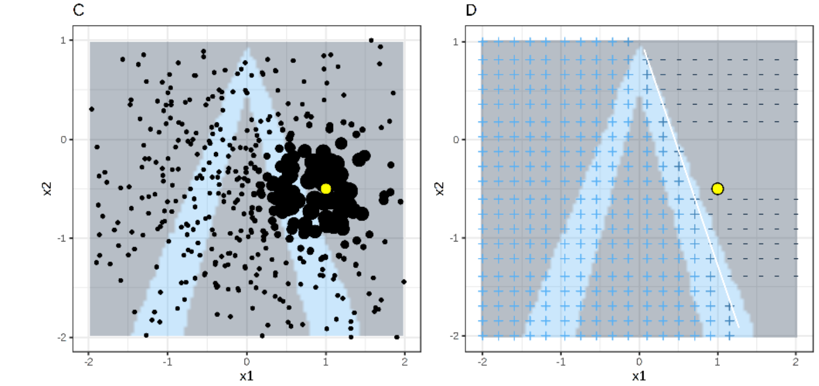
Data

(X)

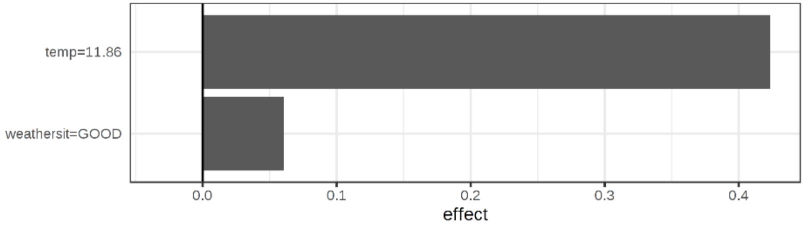
Black-box model

Prediction

White-box model

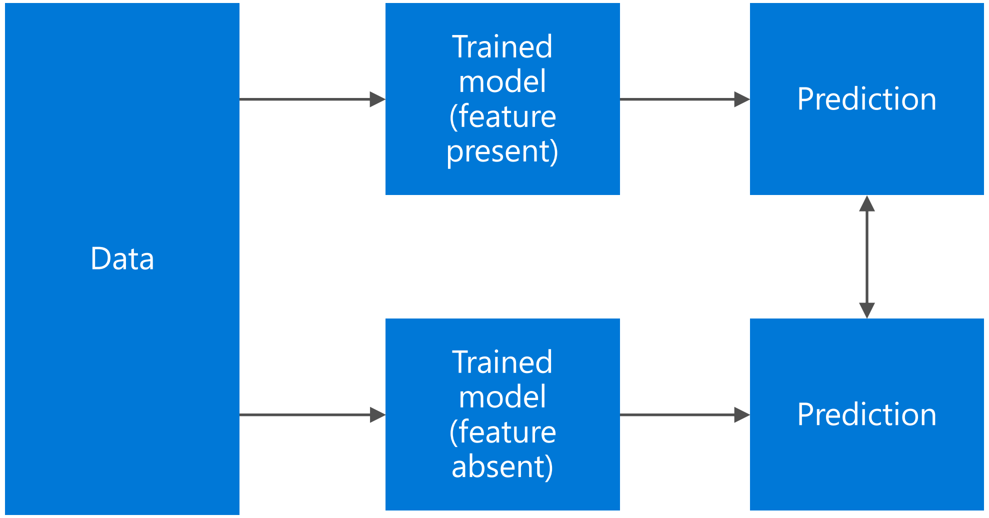


Actual prediction: 0.89
LocalModel prediction: 0.44



Model Interpretability Techniques

SHAPLEY ADDITIVE EXPLANATIONS (SHAP)



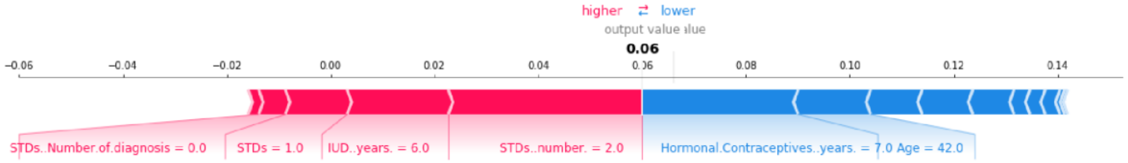
Data

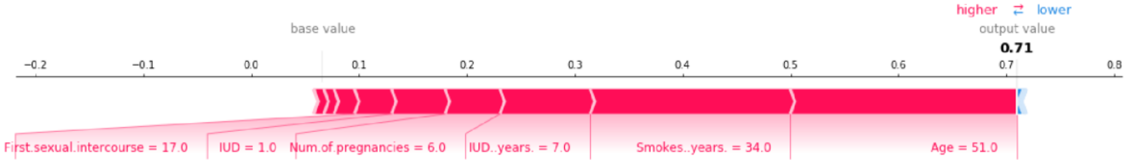
Trained model (feature present)

Prediction

Trained model (feature absent)

Prediction





SOLUTIONS FROM MICROSOFT AZURE

Azure Machine Learning service

Set of Azure Cloud Services + Python SDK

That enables you to:

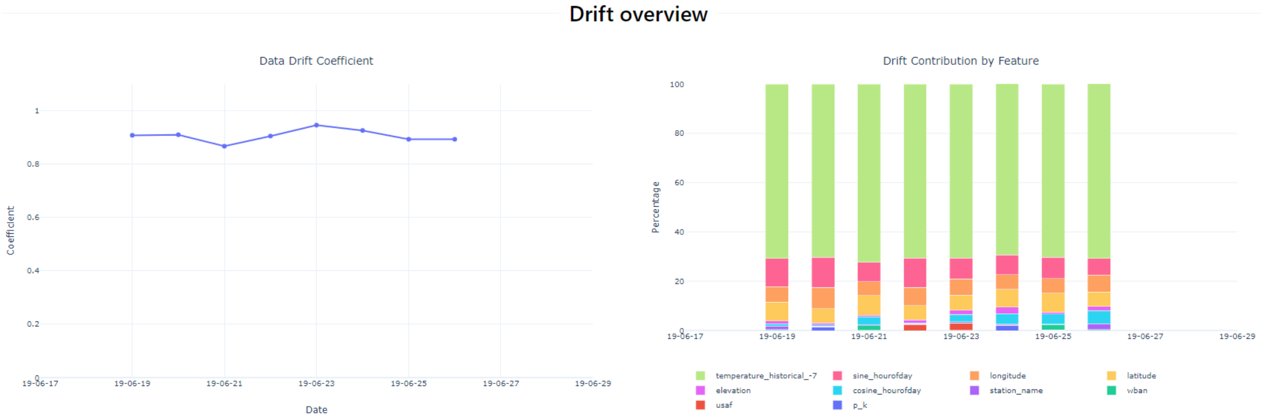
- ✓ Prepare Data
- ✓ Build Models
- ✓ Train Models

- ✓ Manage Models
- ✓ Track Experiments
- ✓ Deploy Models

Azure Machine Learning Service

DATA DRIFT

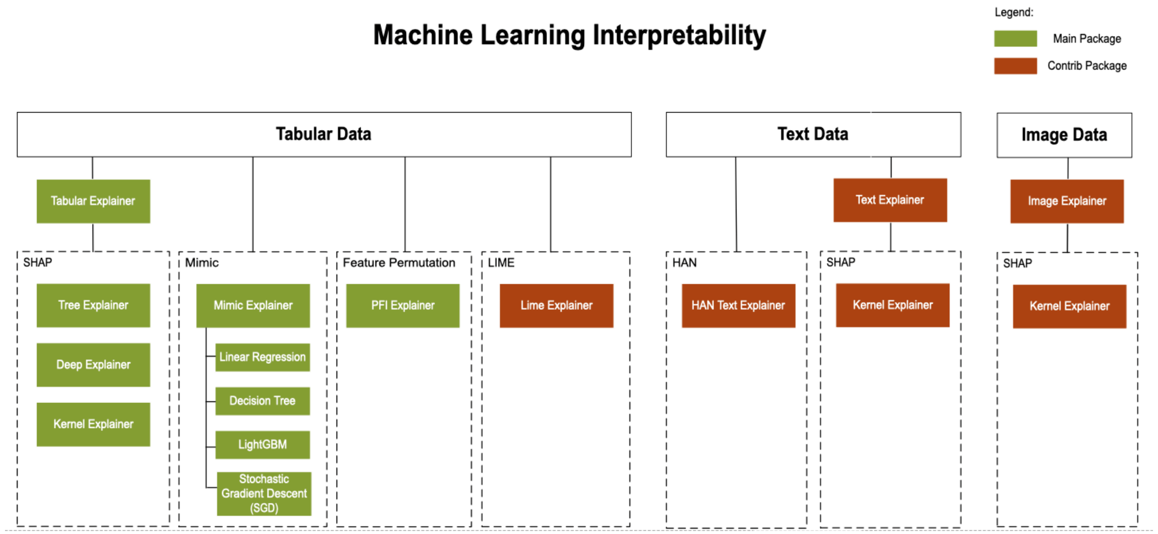
- ○ Measures the magnitude of data drift, called the drift coefficient.
- ○ Measures the data drift contribution by feature, informing which features caused data drift
- ○ Measures distance metrics - currently Wasserstein and Energy Distance are computed
- ○ Measures distributions of features - currently kernel density estimation and histograms. Send alerts to data drift by email



Azure Machine Learning Service

MODEL INTEPRETABILITY

- Feature importance values for both raw and engineered features
- Interpretability on real-world datasets at scale, during training and inference
- Interactive visualizations to aid you in the discovery of patterns in data and explanations at training time
- Explain machine learning models globally on all data, or locally on a specific data point using the state-of-art technologies in an easy-to-use and scalable fashion



Machine Learning Interpretability
Legend:
🟩 Main Package
🟥 Contrib Package
Tabular Data
Tabular Explainer
SHAP
Tree Explainer
Deep Explainer
Kernel Explainer
Mimic
Mimic Explainer
Linear Regression
Decision Tree
LightGBM
Stochastic Gradient Descent (SGD)
Feature Permutation
PFI Explainer
LIME
Lime Explainer
Text Data
Text Explainer
HAN
HAN Text Explainer
SHAP
Kernel Explainer
Image Data
Image Explainer
SHAP
Kernel Explainer