Microsoft Azure

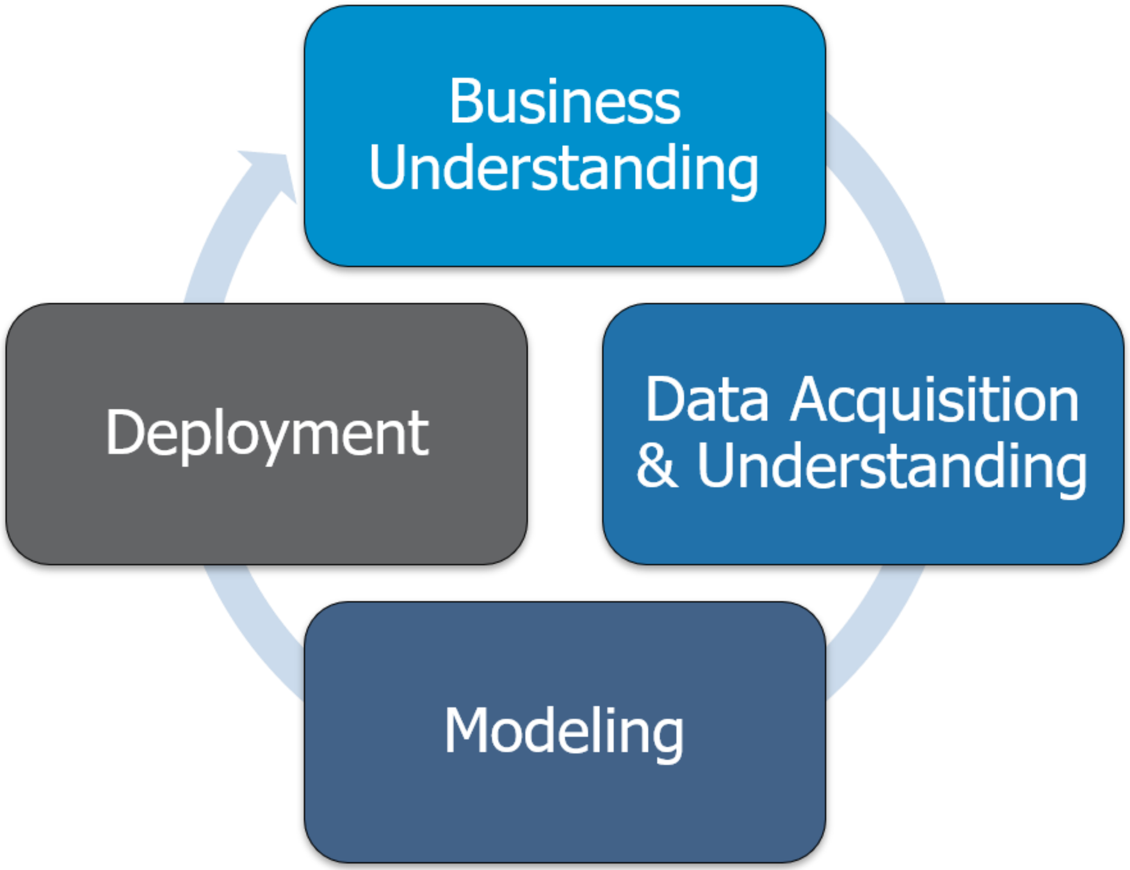
Data Science & Machine Learning

Developing Predictive & Prescriptive Analytic Services

Jon Tupitza
Sr. Cloud Solution Architect – Data & AI – Azure Data Scientist
Jon.Tupitza@Microsoft.com

The Microsoft: Team Data Science Process

Iterative & Exploratory: Largely Based on Conducting Experiments



Business Understanding
- Identify the Problem Domain
- Identify the Solution Scenario

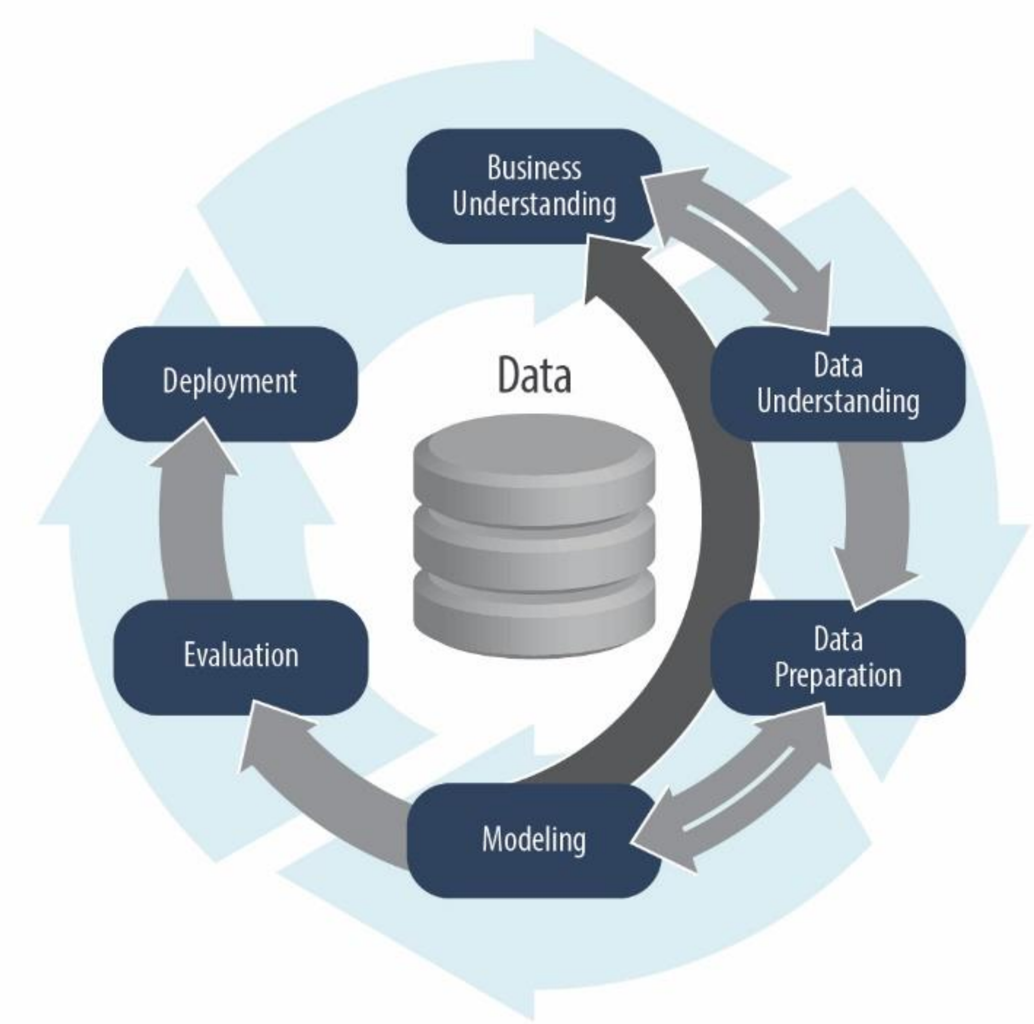
Acquire & Understand Data
- Load, Prepare & Explore Data
- Identify Influential Features

Develop Machine Learning Models
- Select & Engineer Features
- Train, Evaluate & Tune Models

Deployment
- Publish Models as Webservices
- Consume Models Visually and Programmatically

CRISP-DM: Cross-Industry Standard Process-Data Mining

First Introduced in 1996!



Business Understanding	• Identify the Problem Domain • Identify the Solution Scenario
Data Understanding	• Load and Explore Data • Identify Influential Features
Data Preparation	• Remove Duplicates & Nulls • Impute Missing Values • Select & Engineer Features
Modeling	• Train Models Using a Variety of Algorithms • Tune Hyper-parameters
Evaluation	• Test Models' Performance & Predictive Power • Cross-Validate to Appraise Goodness-of-Fit • Select Most Effective Model for Deployment
Deployment	• Publish Models On-premises or in the Cloud • Consume Models Visually & Programmatically

Machine Learning: Key Topics & Activities

Azure Machine Learning Features that Enable Machine Learning Productivity

Exploratory Data Analysis

- Univariate Analysis
- Feature Engineering
- Feature Importance

AutoML:

- Feature Permutation
- Algorithm Selection
- Hyperparameter Tuning

Training and Evaluation

- Model Selection
- Cross-Validation
- Hyperparameter Tuning

Explainability (Responsible AI)

- Partial Dependence
- Out-of-Sample Accuracy (Lift Chart)
- Sensitivity Analysis (Feature Impact)

Machine Learning : Paradigms

Supervised Learning
Classification and Regression

- Learn by historic example to predict the future
- Requires historic data (Observations where the outcome [a Label] is recorded)

Unsupervised Learning
Clustering

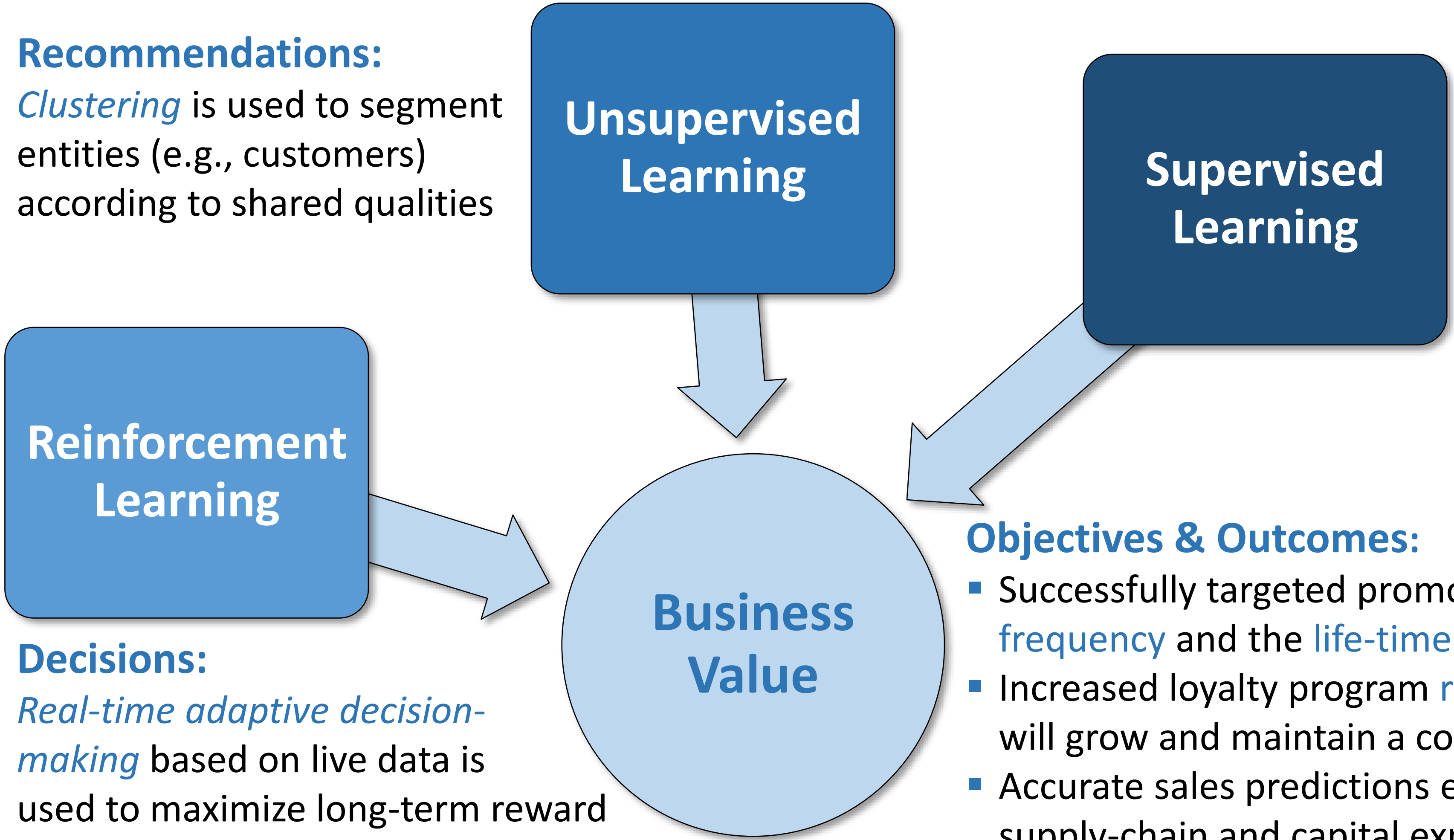
- Understanding the past by gaining insight from historic data
- Involves Feature-sets having no Labels (Observations where answers aren't known)
- Examine data to reveal its intrinsic [but hidden] structure
- Make recommendations by grouping people, things, or events together

Reinforcement Learning

- Studies the consequences of making decisions over time by using agents to interact with environments through trial-and-error
- Uses feedback to reinforce preferred actions over time by way of positive rewards
- Negotiates Exploitation versus Exploration to identify optimal rewards
- Continually adapts to their environments to maximize long-term reward

Machine Learning Paradigms: Strategic Interaction

Recommendations:
Clustering is used to segment entities (e.g., customers) according to shared qualities



Predictions:
- *Classification & Regression* use historic outcomes to predict future results
- *Time-series forecasts* are used to project outcomes that are used to measure accuracy of predictions

Decisions:
Real-time adaptive decision-making based on live data is used to maximize long-term reward

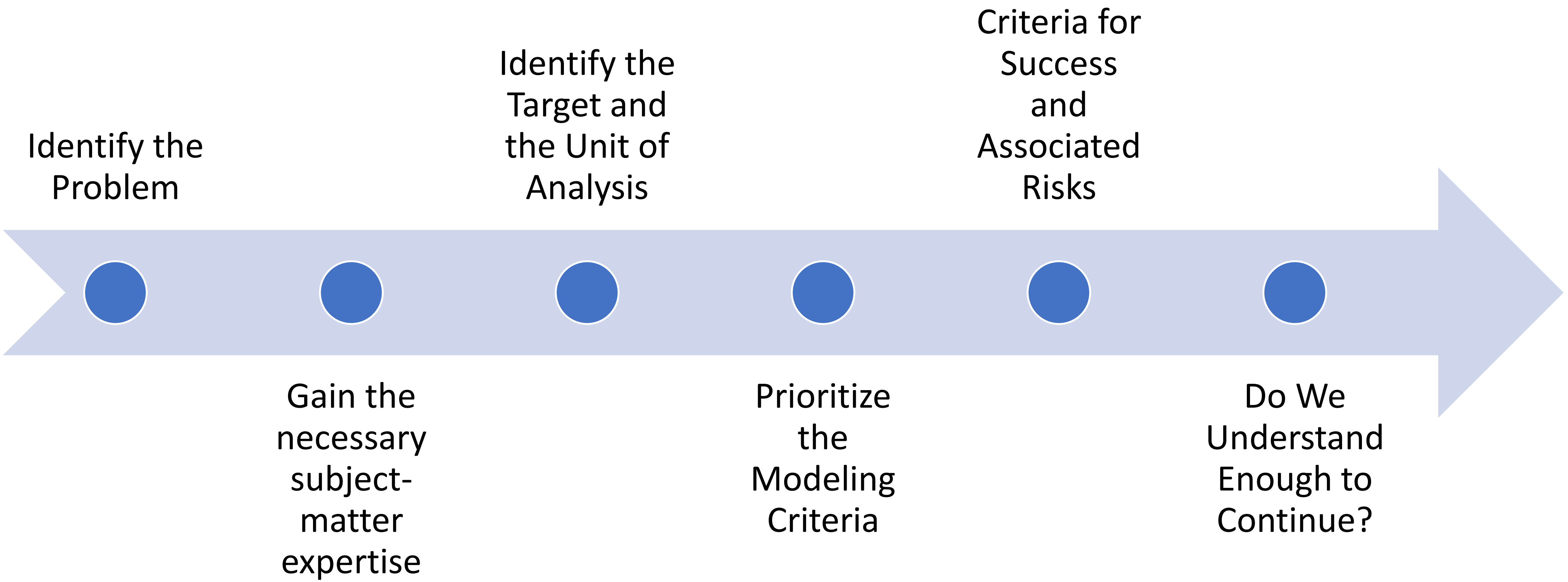
Objectives & Outcomes:
- Successfully targeted promotions increase purchase frequency and the life-time value of customers
- Increased loyalty program recruiting & retention will grow and maintain a company’s customer base
- Accurate sales predictions enable more effective supply-chain and capital expense management

Understanding the Business Case

Defining the Business Objectives for the Project

Business Understanding: Criteria for Success

If we don’t understand the question we need to answer, and its business value.. We’ll fail!



Business Understanding: Identify the Problem

Acquiring a Complete Understanding of the Situation in Business Terms

- State the Problem in the Language of the Business; Not in Technical Jargon
- Identify all actions and outcomes that might result
- Requires specificities including the number of parties that would be affected, and costs
- Requires impact to the bottom line to be specified

Insurance:

- Fulfilling accident claims cost us $125,000,000 last year, but we don't have a reliable way to determine which applicants represent a high probability of being involved in an accident.

Telecommunications:

- Customers who leave and take their business to our competition cost us $65,000,000 last year, but we can't reliably identify those customers who are at most risk for leaving.

HealthCare:

- Hospital readmissions cost us $75,000,000 last year, but we don't have a way to identify which patients are at most risk.

Acquiring & Understanding Data

Exploratory Data Analysis (EDA)

Data Understanding: Define the Unit of Analysis

What Does Each Row (Observation) Represent?

Retail Sales Data

Date	Customer ID	Store ID	Purchase Amount	Number of Items
10/02/2019	2037	27	$107.23	3
10/02/2019	2038	27	$99.50	2
10/02/2019	2037	27	$212.49	5
10/02/2019	2091	29	$37.04	2
10/02/2019	2302	29	$18.02	1

Data Understanding: Define the Target

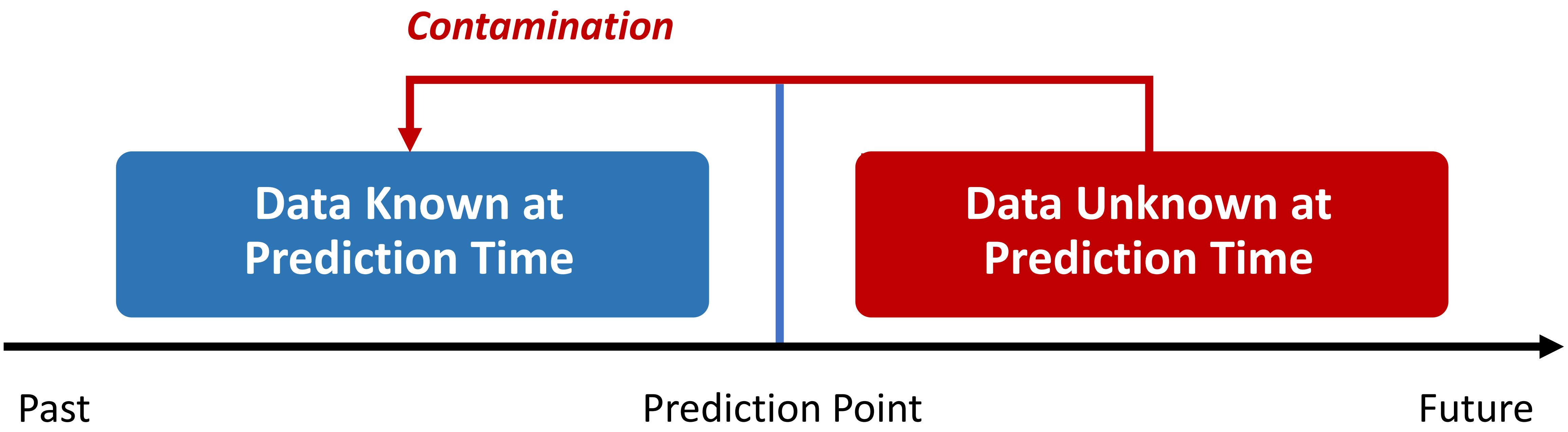
What Do You Want to Predict?

Retail Sales Data

Date	Customer ID	Store ID	Purchase Amount	Number of Items
10/02/2019	2037	27	$107.23	3
10/02/2019	2038	27	$99.50	2
10/02/2019	2037	27	$212.49	5
10/02/2019	2091	29	$37.04	2
10/02/2019	2302	29	$18.02	1

Data Understanding: Target Leakage

Data Not Known at the Time of Prediction



Data Understanding: Target Leakage

Data Not Known at the Time of Prediction

Loan Risk Data

Education	Married	Purpose	Late Payments	Annual Income	Loan Is Bad
1	Yes	Car Purchase	0	$107,000	0
3	No	Small Business	3	$99,000	1
1	Yes	House Purchase	5	$85,000	1
2	No	Marriage	1	$72,000	0
2	No	Debt Consolidation	0	$120,000	0

Demo: EDA with Azure Machine Learning Studio

Loading and Exploring Datasets

Prepare Data, Engineer & Select Features

Data Science & Machine Learning Development

Data Preparation: Feature Engineering

- Date/Time Parsing
- Impute Missing Values
- Scaling
- Encoding
- Generalization
- Discretization (Binning)
- Credibility Estimates

DateOfSale
03/07/2020
06/07/2020
09/02/2020
10/08/2020
12/07/2020

DayOfWeek	MonthOfYear	IsWeekend
Saturday	March	1
Sunday	June	1
Wednesday	September	0
Thursday	October	0
Monday	December	0

Age

Age Group

DateOfBirth
02/13/1969
03/05/1972
04/11/1984
05/21/1995
06/24/2002

Age	Age Group
51	Fifties
48	Forties
36	Thirties
25	Twenties
18	Teens

Feature Engineering: Numerical Features

- **Impute** missing values and create a flag to indicate imputed values:
  - Mean
  - Median

Price
7
null
5
NaN
9

Mean	Median	Imputed Flag
7	7	0
7	5	1
5	5	0
7	5	1
9	9	0

- **Scaling:**
  - **Standardize**: Rescale so mean (μ) = 0 and Standard Deviation (σ) = 1

$$z = (x_i - \mu)/\sigma$$

  - **Normalize**: Rescale so the range falls between 0 and 1

$$z = (x - \min(x) / \max(x) - \min(x)$$

Feature Engineering: Encoding Categorical Features

Because machine learning algorithms cannot interpret text data, categorical features must first be transformed (encoded) into numerical values.

One-Hot Encoding

Group
A
B
C
A
A

→

Group A	Group B	Group C
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0

aka, Dummy Features

Count Encoding

Group
A
B
C
A
A

→

Count
3
1
1
3
3

Ordinal Encoding

Group
A
B
C
A
A

→

Ordinal
0
1
2
0
0

Feature Engineering: Credibility Estimates

Categorical features often have unequal member distributions.
Credibility estimates help compensate for this imbalanced representation.

Target	Group	Credibility Estimate
0	A	3 * (0.33 – 0.4)
0	B	1 * (0 – 0.4)
1	C	1 * (0 – 0.4)
1	A	3 * (0.33 – 0.4)
0	A	3 * (0.33 – 0.4)

The more of a value we observe in a group, the more we trust that group's deviation from the overall mean.

$$\boldsymbol{count_k \times (\bar{y}_k - \bar{y})}$$

Feature Selection: Partial Dependence

A Univariate Method: The influence of each feature's influence is measured before modeling

Iteratively sets all observations in the column to each unique value contained in that column.

Then

Then observe the correlation each value of the column has to the response variable (target).

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

All Values Set to $20k

Loan	Income	IsBad
$11,200	**$20,000**	False
$10,000	**$20,000**	False
$8,000	**$20,000**	True
$16,000	**$20,000**	True
$4,000	**$20,000**	False

All Values Set to $65k

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$65,000**	False
$8,000	**$65,000**	True
$16,000	**$65,000**	True
$4,000	**$65,000**	False

Feature Selection: Permutation Importance

A Model-based Method: The performance of the model is measured before and after...

Randomly shuffling the values in each column, one-at-a-time, to break the correlation that each column has to the target variable

Then

Measuring the impact the change to each column's influence has upon the model's overall performance according to one of many applicable metrics:

Generates a stack-ranked list of features by their scores

Classification: Accuracy, Precision, or Recall

Regression: MAE, RMSE, RAE, RSE, R^2, Precision, Recall

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

Shuffle Column 1

Loan	Income	IsBad
$8,000	$108,000	False
$4,000	$65,000	False
$16,000	$20,000	True
$10,000	$110,000	True
$11,200	$155,000	False

Shuffle Column 2

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$155,000**	False
$8,000	**$108,000**	True
$16,000	**$20,000**	True
$4,000	**$110,000**	False

Demo: Data Preparation & Feature Engineering

Data Cleansing and Feature Engineering Features with Python

Microsoft