PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Machine Learning

MACHINE LEARNING

Overview

- Machine Learning includes supervised, unsupervised, and semi-supervised (reinforcement) learning from historical (training) data
- Unsupervised learning finds patterns in the data without direction by an expert.
- Supervised learning attempts to mimic an expert by learning from expertly labeled data.

MACHINE LEARNING

Overview of Unsupervised Learning

Clustering, Anomaly Detection, and PCA are examples of Unsupervised Learning

- Clustering or Segmentation groups data points together
- Anomaly detection finds data points that are different
- PCA reorganizes numeric data. Each point is mapped to a new location (in a lower dimension).

MACHINE LEARNING

Overview of Supervised Learning

Classification and Regression are examples of Supervised Learning.

- Classifications predict categories. Each case in the training data, like a row in a table, was labeled with a category (not a number)
- Regressions predict numeric values. Each case in the training data, like a row in a table, was labeled with a numeric value.

Phases of a predictive analytics model

Training: feeding a machine learning algorithm some data so that it can learn from it and come up with a reliable generalization (representation) of the data

Testing (Supervised Learning): using data with unknown targets (to the particular model) and measuring how much the model's predictions align with the actual targets

Deployment: putting a model into production, to be used with unknown targets

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

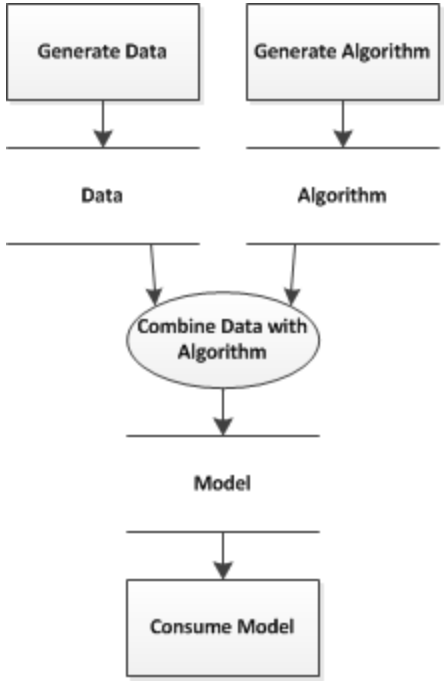
Data Flow in Supervised Learning

FROM DATA TO PREDICTIONS

> How do we get from data to predictions?

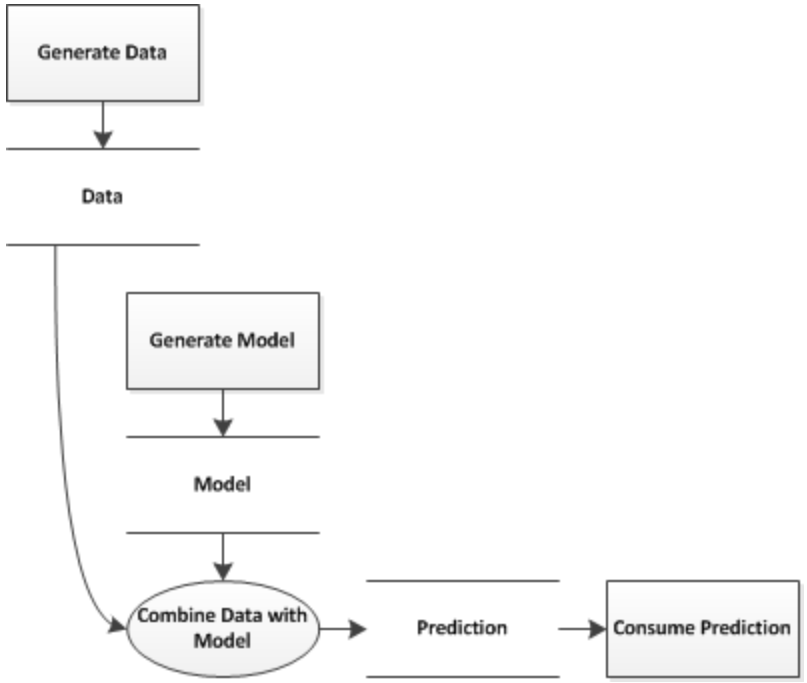
Data → ? → Predictions

FROM DATA TO PREDICTIONS



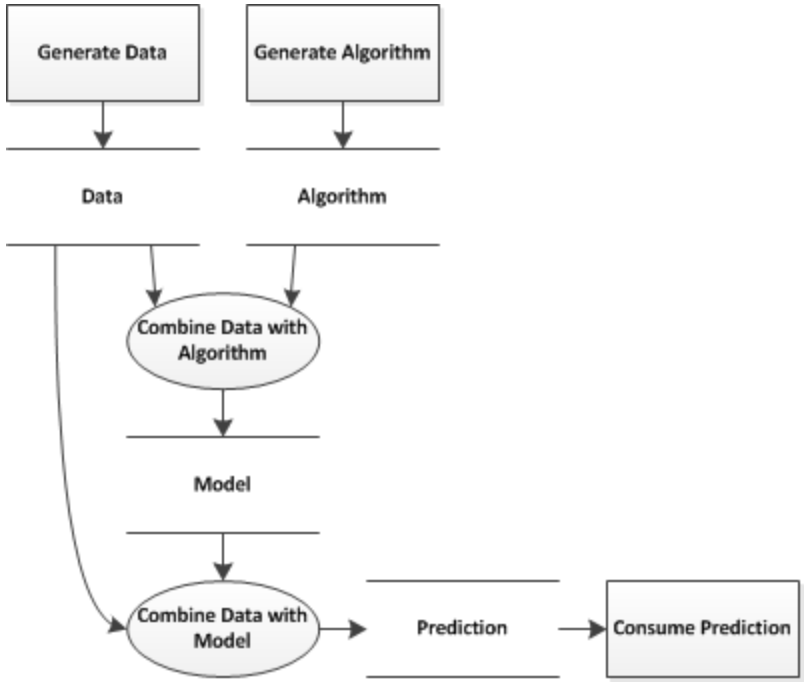
Training Data + Algorithm → Model

FROM DATA TO PREDICTIONS



Model + Operational Data → Prediction

FROM DATA TO PREDICTIONS



Training Data + Algorithm → Model
Model + Operational Data → Prediction

FROM DATA TO PREDICTIONS

- Pseudo Assignments (Derivations):
  - Training Data + Algorithm → Model
  - Model + Operational Data → Prediction
- Create Model from Algorithm and Data
  - Example Create Logistic Regression
    - model = LogisticRegression()
    - model.fit(OldInputs, OldTarget)
- Predict from Model and Data
    - prediction = model.predict(NewInputs)
    - The prediction are for “new” target values

Training Data + Algorithm → Model
Model + Operational Data → Prediction

FROM DATA TO PREDICTIONS

Some Algorithms for Supervised Learning

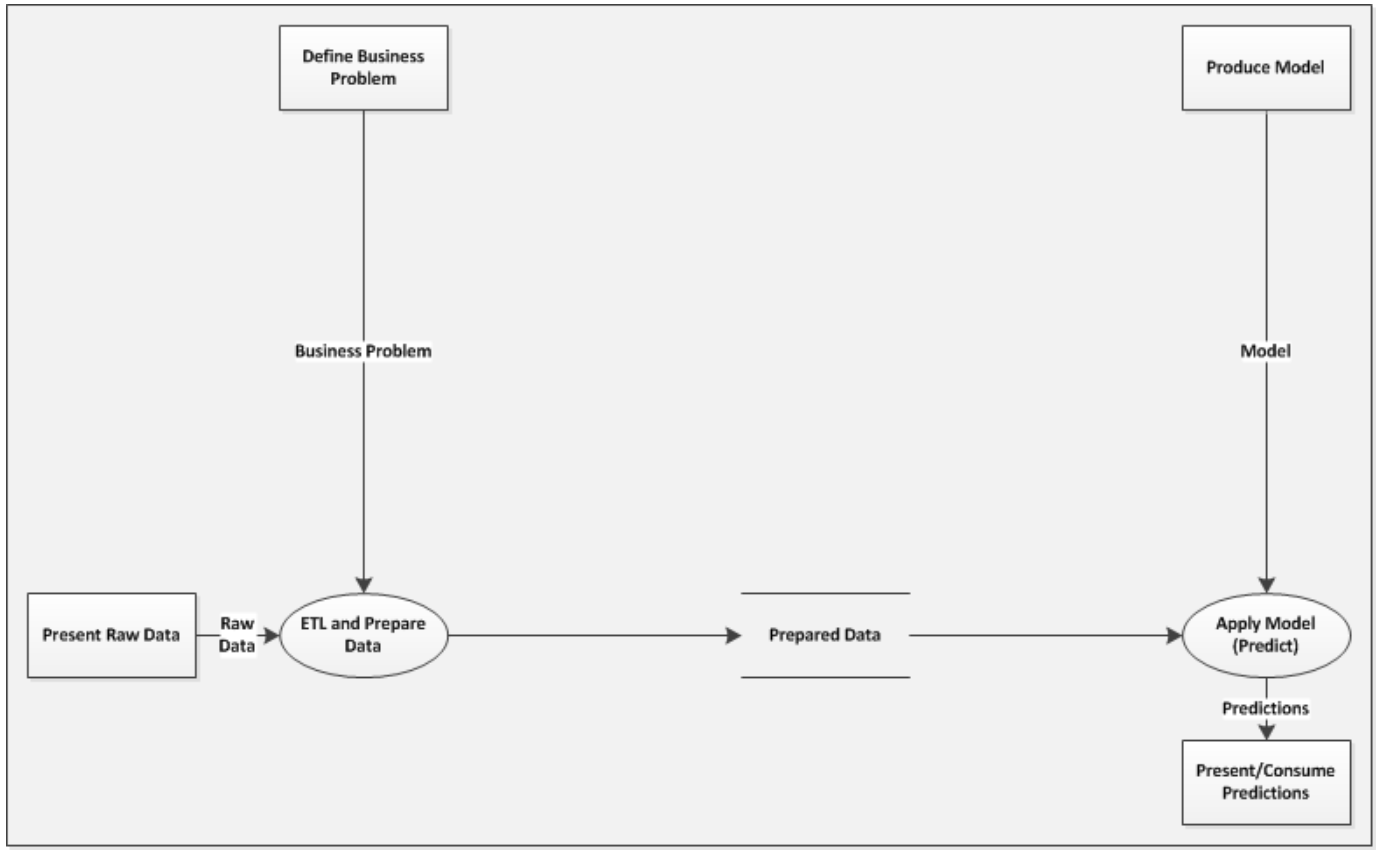
- Classification
  - Logistic Regression
  - Neural Network
  - Decision Tree
  - Naïve Bayes
- Regression
  - Linear Regression
  - Regression Trees
  - Neural Network

DFD OF SUPERVISED LEARNING

MODEL ACTS ON DATA

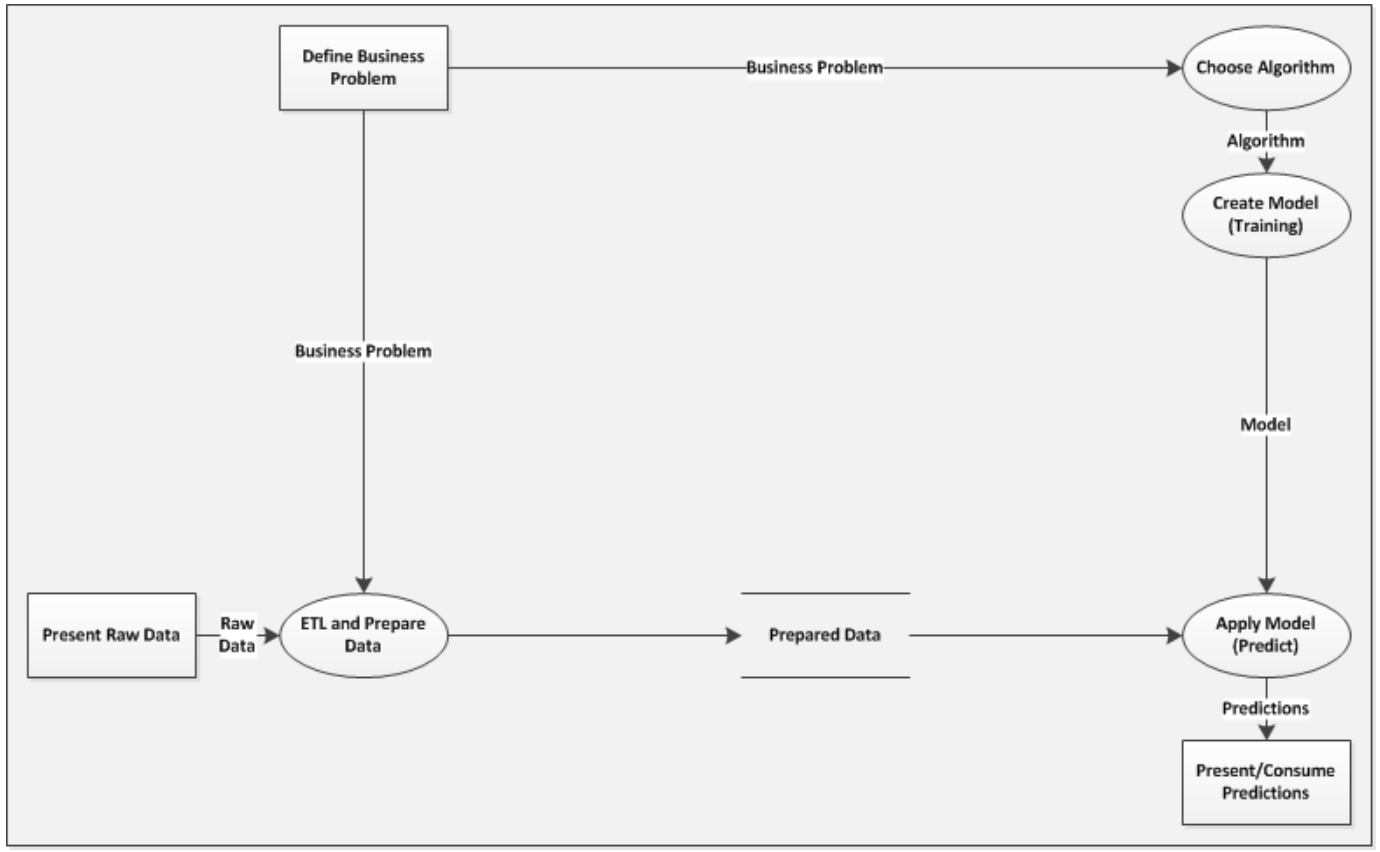
Model + Data → Prediction

DATA ETL AND PREPARATION DRIVEN BY BUSINESS PROBLEM



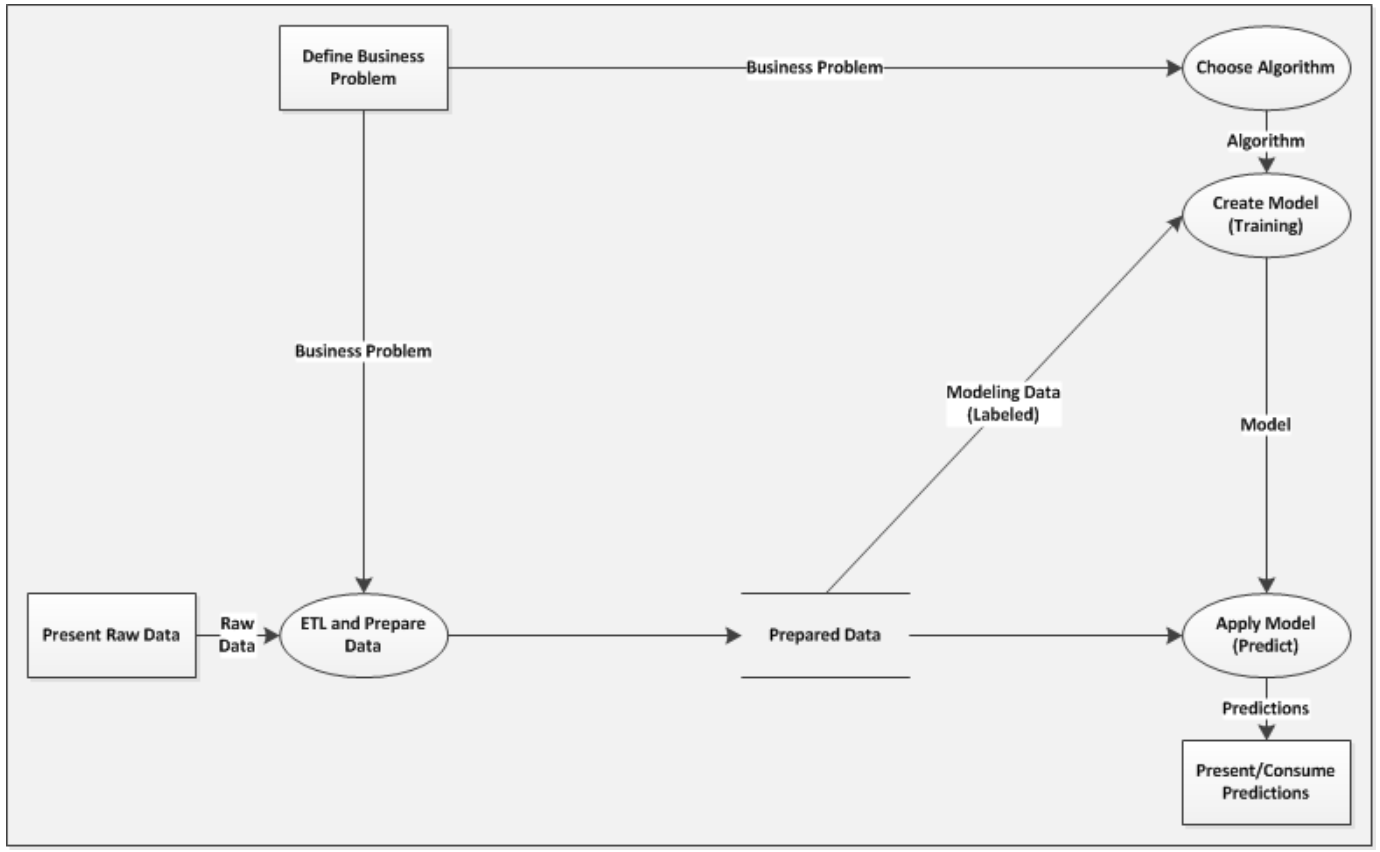
Business Problem determines ETL and Data Prep

ALGORITHM CHOICE DRIVEN BY BUSINESS PROBLEM



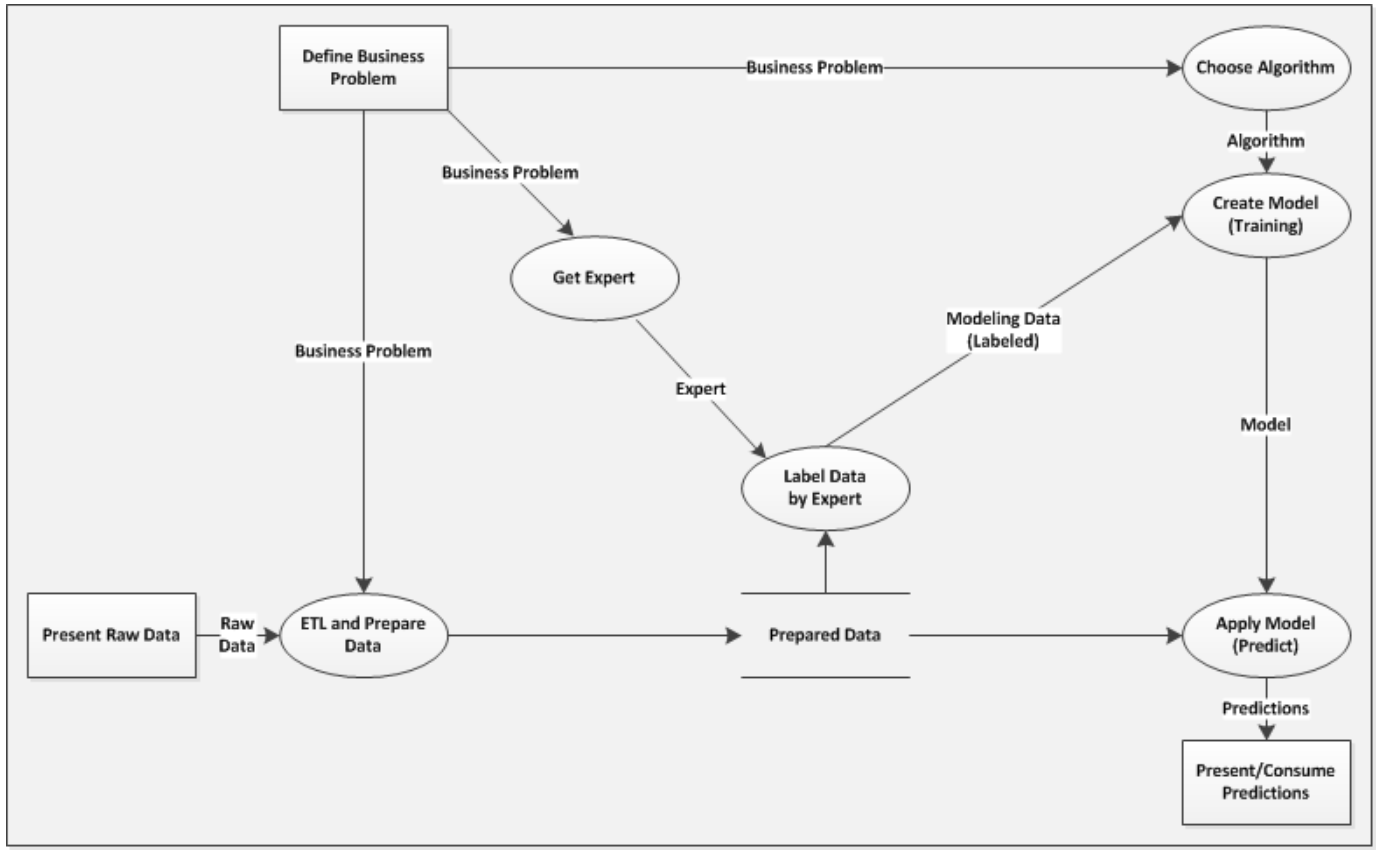
Business Problem determines the choice of Algorithm.

MODEL CREATION NEEDS DATA



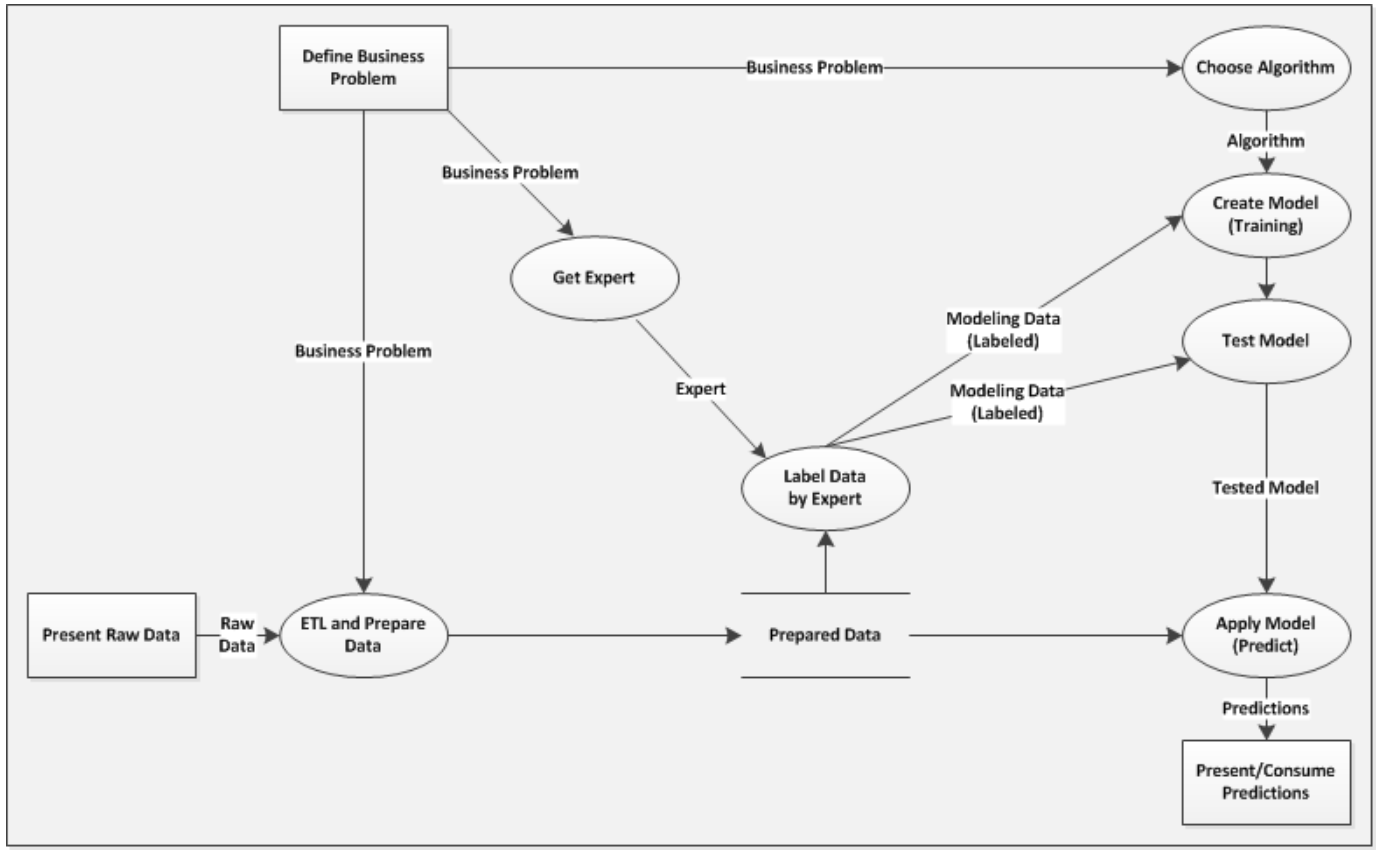
Data + Algorithm → Model

SUPERVISED TRAINING NEEDS DATA LABELED WITH OUTCOMES



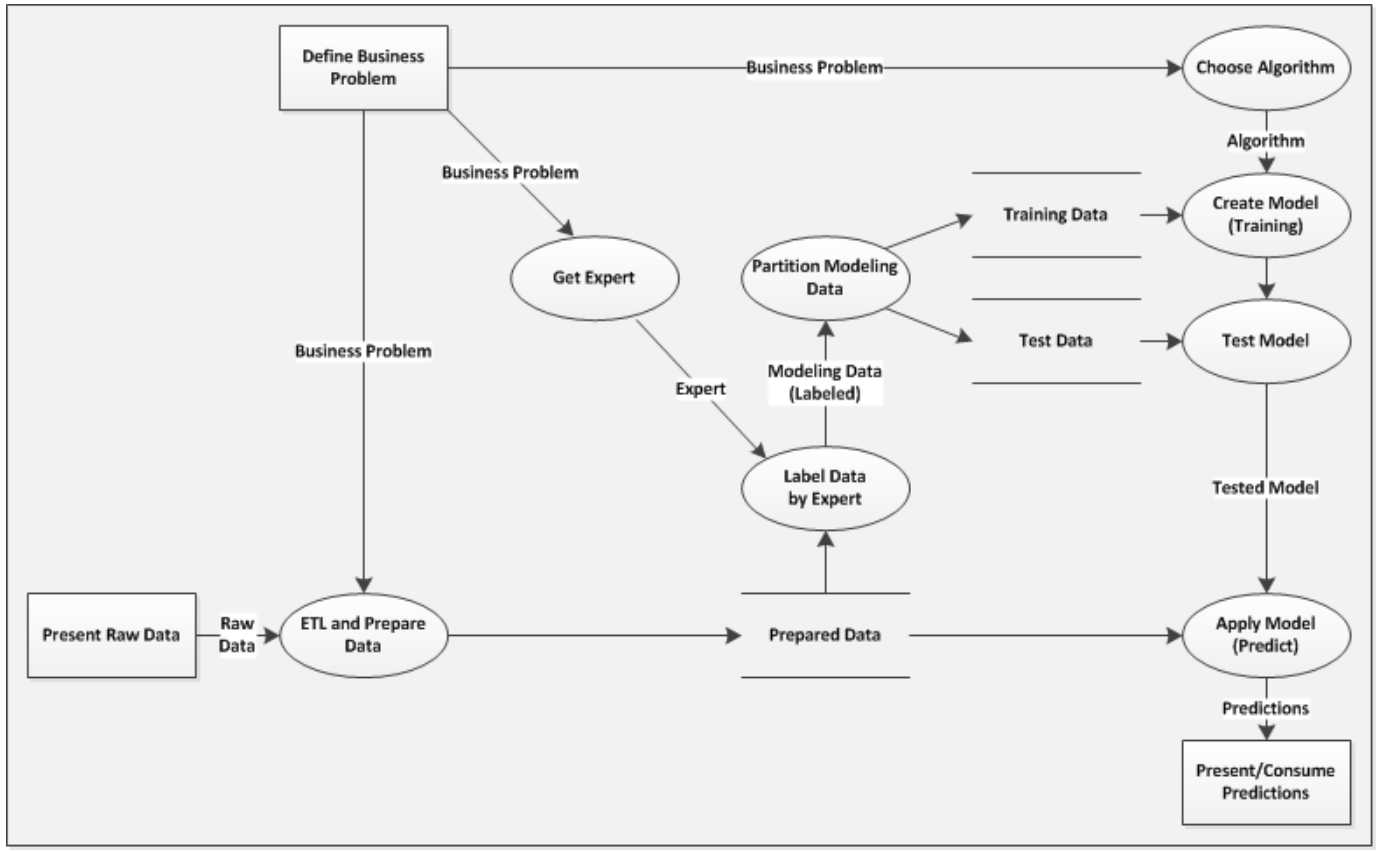
Supervised Learning requires expert labeling of data.

MODELS NEED TO BE TESTED



Do not trust predictions from an un-tested model!

TRAINING & TESTING OF MODEL USE DIFFERENT DATA



Do not test a model using training data!

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

SUPERVISED LEARNING SCHEMA

SUPERVISED LEARNING SCHEMA

- Modeling Dataset
  - Rectangular Dataset (aka table)
  - Schema
    - Input columns
    - Output column (target, outcome)
      - Classification: Category Column
      - Regression: Numeric Column
  - Horizontal partition of modeling data into training and test data
- Incremental data has same schema as modeling data, except:
  - Incremental data does not have the output column (target, outcome)
  - Incremental data is not partitioned into training and test data

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Here is a rectangular dataset. The table has columns with headers and the data in each column have the same datatype. The data have been prepared and are ready for modeling.

SUPERVISED LEARNING SCHEMA

Elsewhere, I have new data that do not contain the target outcome. I want to predict categorical values, like these, from this new data. For each row in the new data, I want to use the values from the other columns in the same row to predict the value in the missing column. This predicted value is called the "Target Outcome".

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Target Outcome

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Target Outcome

SUPERVISED LEARNING SCHEMA

Keys and random data should not be used as inputs for predictive analytics. Random data may appear to have patterns, but those patterns are fortuitous and will not be available when needed for predictions. Keys may contain patterns, but these patterns are deceptive and may also not be available when needed.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Target Outcome

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Target Outcome

SUPERVISED LEARNING SCHEMA

Columns with constant data are unnecessary. In general, they will not affect the algorithm and therefore the model will be the same. But, they distract from the task. Also, they may increase memory and processing requirements.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Constant

Target Outcome

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Constant

Target Outcome

SUPERVISED LEARNING SCHEMA

A proxy column is a column that was created after the "target" was observed. The proxy contains information that would not be available for predictions. The proxy column correlates well with the target .

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Constant

Proxy

Target Outcome

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Constant

Proxy

Target Outcome

SUPERVISED LEARNING SCHEMA

Some inputs to supervised learning are continuous attributes, like integers, floats and time.

Some inputs to supervised learning are categories, like strings, binned numbers, and factors.

Some inputs to supervised learning are binary attributes, like categories with only two states and binarized multi-state categories.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No

Random or Keys

Constant

Proxy

Continuous Input

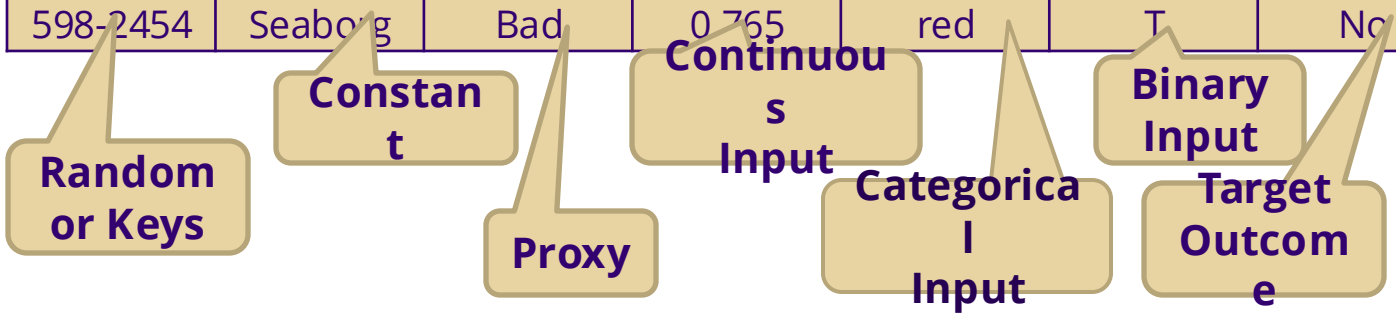
Categorical Input

Binary Input

Target Outcome

SUPERVISED LEARNING SCHEMA

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
330-3141	Seaborg	Good	0.123	red	T	Yes
330-3150	Seaborg	No	0.987	green	T	No
330-3202	Seaborg	Yes	0.245	blue	F	Yes
415-2008	Seaborg	Yes	0.254	blue	T	Yes
415-2081	Seaborg	Bad	0.244	blue	F	No
415-2796	Seaborg		0.415	green	F	Maybe
415-2799	Seaborg	Yes	0.925	red	T	Yes
415-2913	Seaborg	Yes	0.376	green	F	Yes
415-3659	Seaborg	Bad	0.615	green	T	No
595-8413	Seaborg		0.321	blue	F	Maybe
598-1243	Seaborg	No	0.098	green	F	No
598-2454	Seaborg	Bad	0.765	red	T	No



SUPERVISED LEARNING SCHEMA

	Input 1	Input 2	Input 3	Target
	0.123	red	T	Yes
	0.987	green	T	No
	0.245	blue	F	Yes
	0.254	blue	T	Yes
	0.244	blue	F	No
	0.415	green	F	Maybe
	0.925	red	T	Yes
	0.376	green	F	Yes
	0.615	green	T	No
	0.321	blue	F	Maybe
	0.098	green	F	No
	0.765	red	T	No

Continuous Input

Categorical Input

Binary Input

Target Outcome

SUPERVISED LEARNING SCHEMA

Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

Continuous Input

Categorical Input

Binary Input

Target Outcome

SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3

Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

SUPERVISED LEARNING SCHEMA

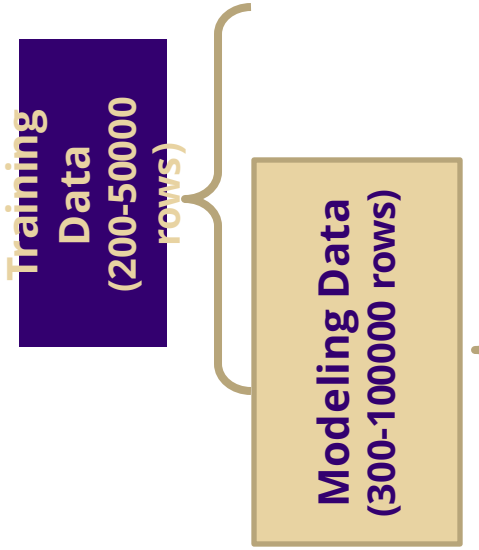
Target from Input 1, Input 2, Input 3

Modeling Data (300-100000 rows)

Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

SUPERVISED LEARNING SCHEMA

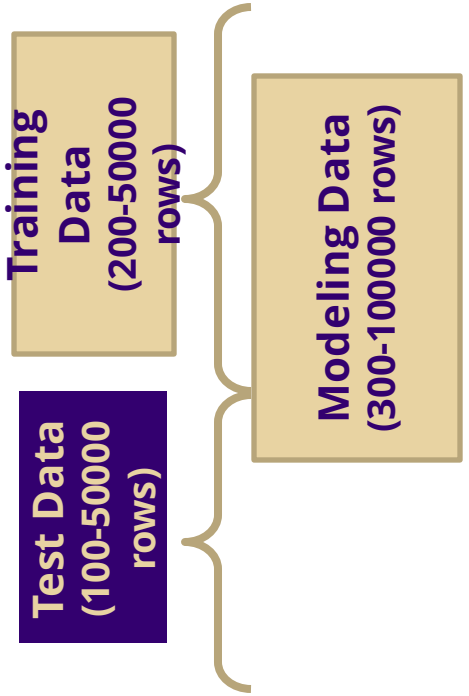
Target from Input 1, Input 2, Input 3



Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

SUPERVISED LEARNING SCHEMA

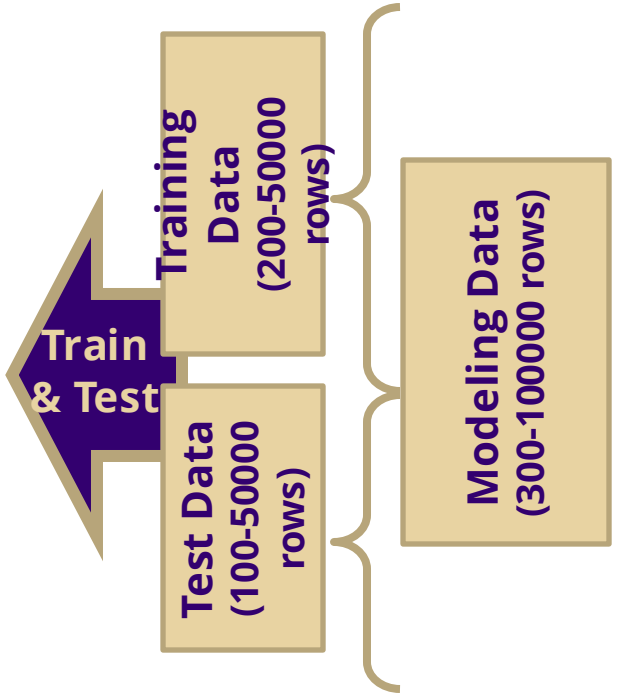
Target from Input 1, Input 2, Input 3



Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

SUPERVISED LEARNING SCHEMA

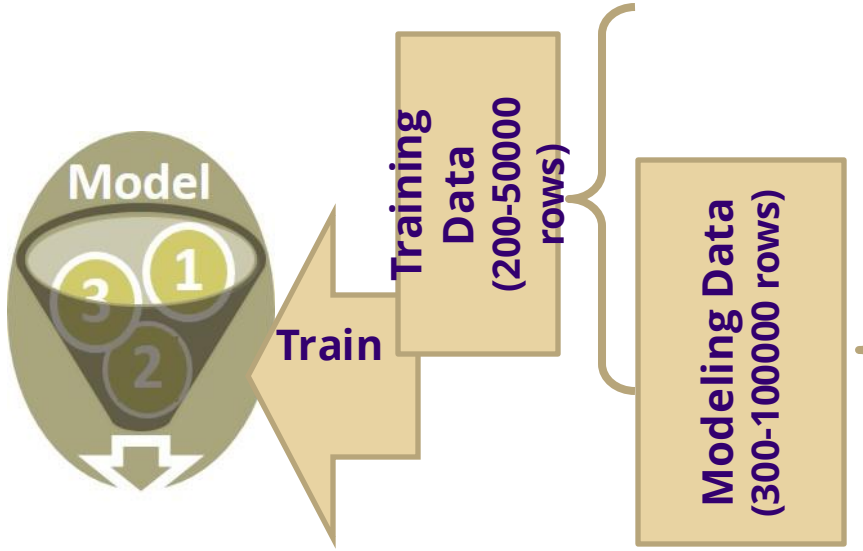
Target from Input 1, Input 2, Input 3



Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No

SUPERVISED LEARNING SCHEMA

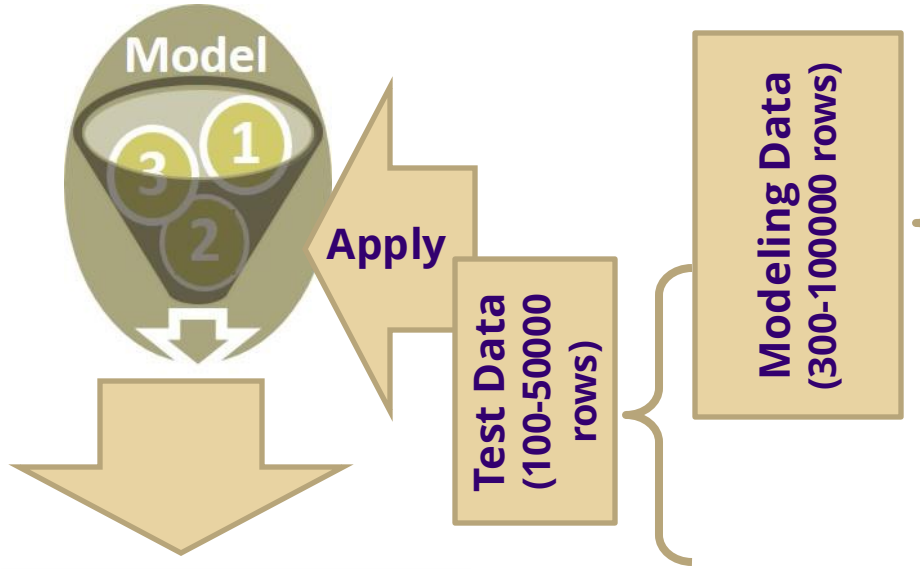
Target from Input 1, Input 2, Input 3



Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes

SUPERVISED LEARNING SCHEMA

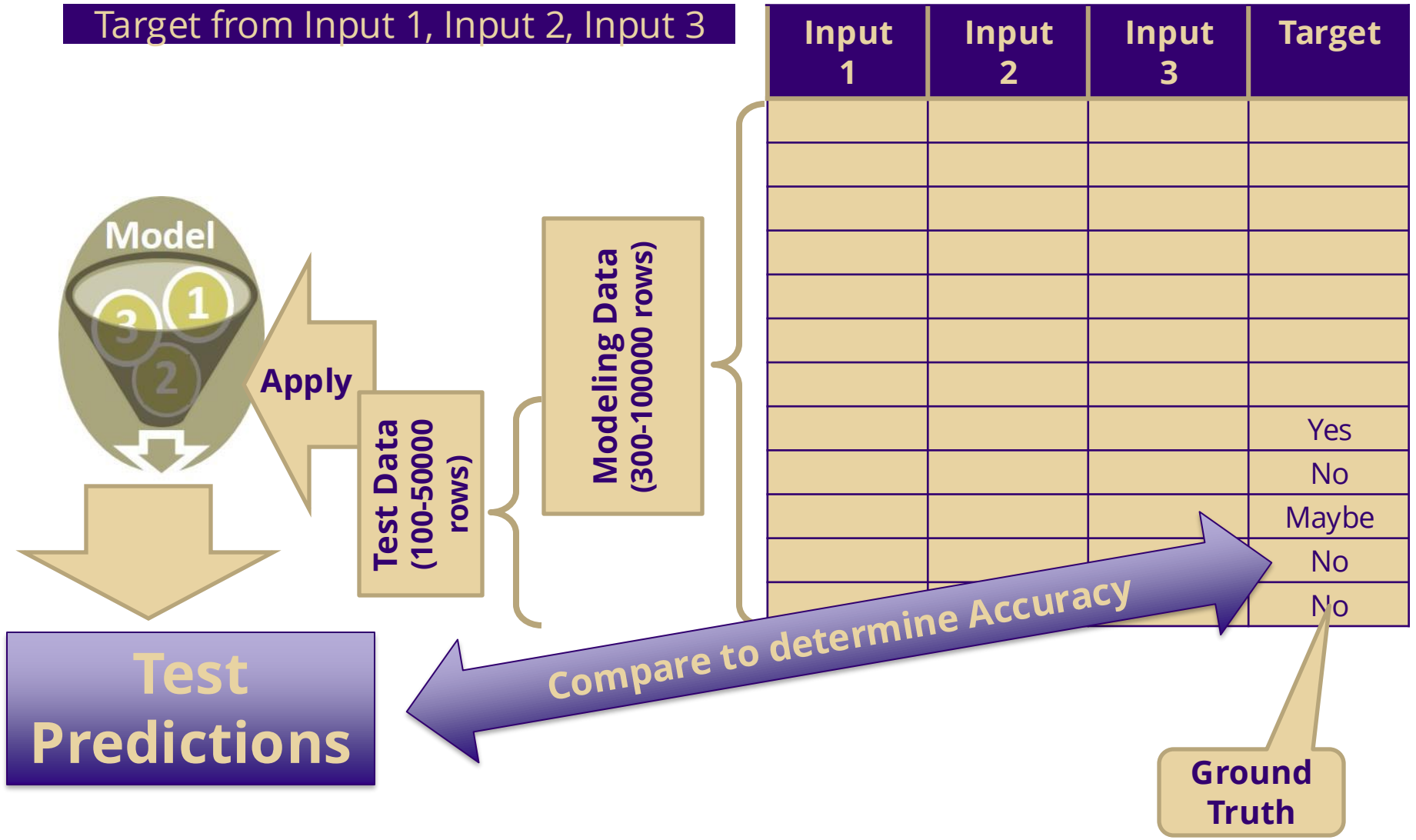
Target from Input 1, Input 2, Input 3



Input 1	Input 2	Input 3	Target
0.376	green	F	
0.615	green	T	
0.321	blue	F	
0.098	green	F	
0.765	red	T	

Test Predictions

SUPERVISED LEARNING SCHEMA



SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3

Elsewhere, I have new data that do not contain the target outcome. I want to predict categorical values, like these, from this new data. For each row in the new data, I want to use the values from the other columns in the same row to predict the value in the missing column. This predicted value is called the "Target Outcome".

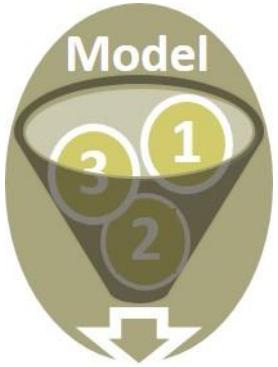
Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

Operational Data (1- ∞ rows)

Unknown Target Outcome

SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



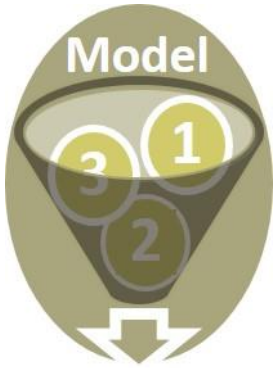
Operational Data (1- ∞ rows)

Input 1	Input 2	Input 3	Target
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

Unknown Target Outcome

SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



Deploy Model to Predict Target Outcome

Unknown Target Outcome

Operational Data (1- ∞ rows)

Input 1	Input 2	Input 3	Target
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3

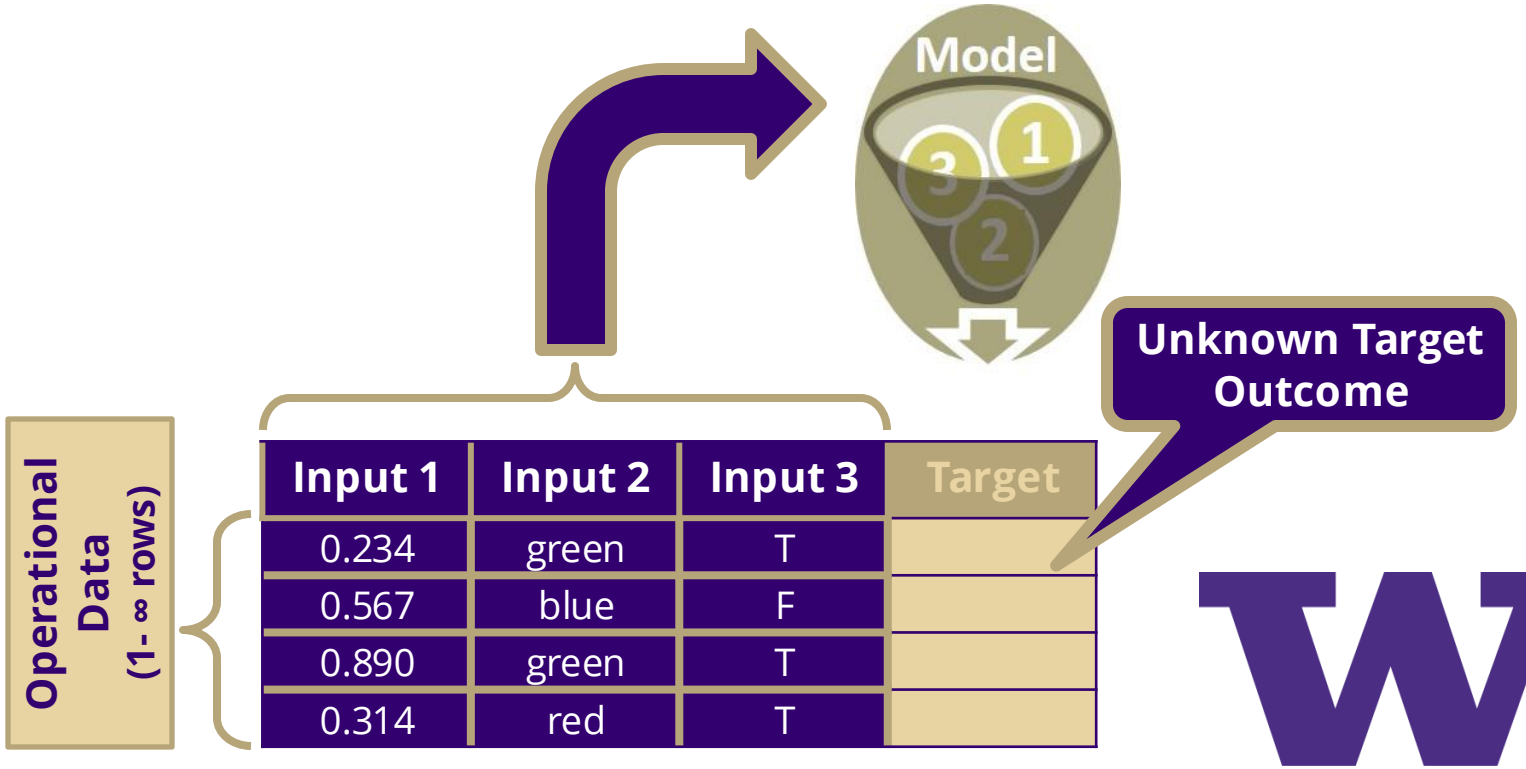
Unknown Target Outcome

Operational Data (1- ∞ rows)

Input 1	Input 2	Input 3	Target
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

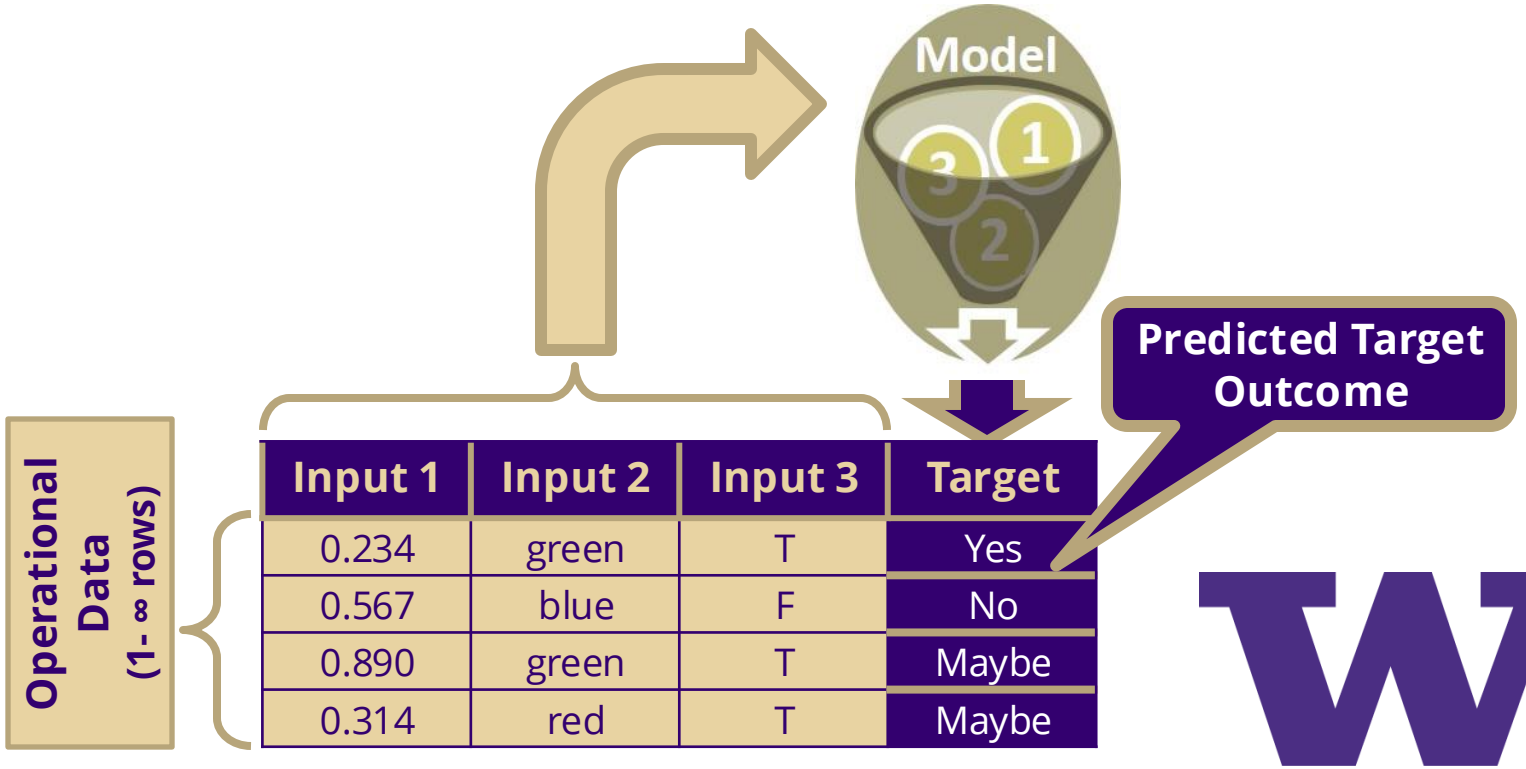
SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



SUPERVISED LEARNING SCHEMA

- Attributes
  - All the columns are attributes
- Input Column
  - Input columns are columns that can help predict the outcome. Input columns can be of type binary, ordinal, numeric, or category.
- Target Outcome
  - The term "Target Outcome" is redundant. The outcome is the target and vice versa. The target or outcome is the output of a predict function. Providing target or outcome values during modeling makes the process supervised. Creating a model using a outcome is called supervised learning.