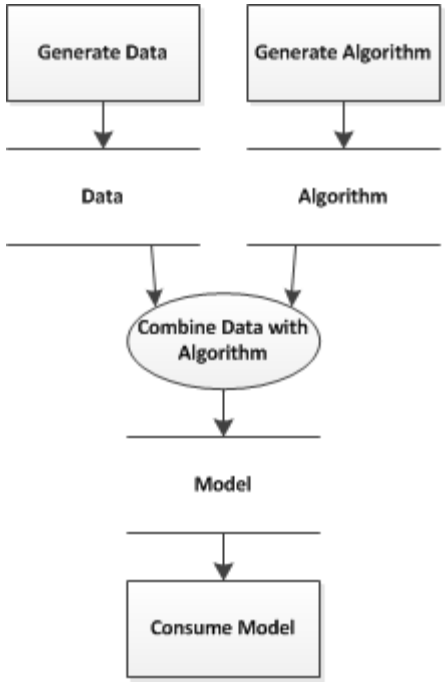
Data and Models in Supervised Learning

© 2017 Ernst Henle. All rights reserved

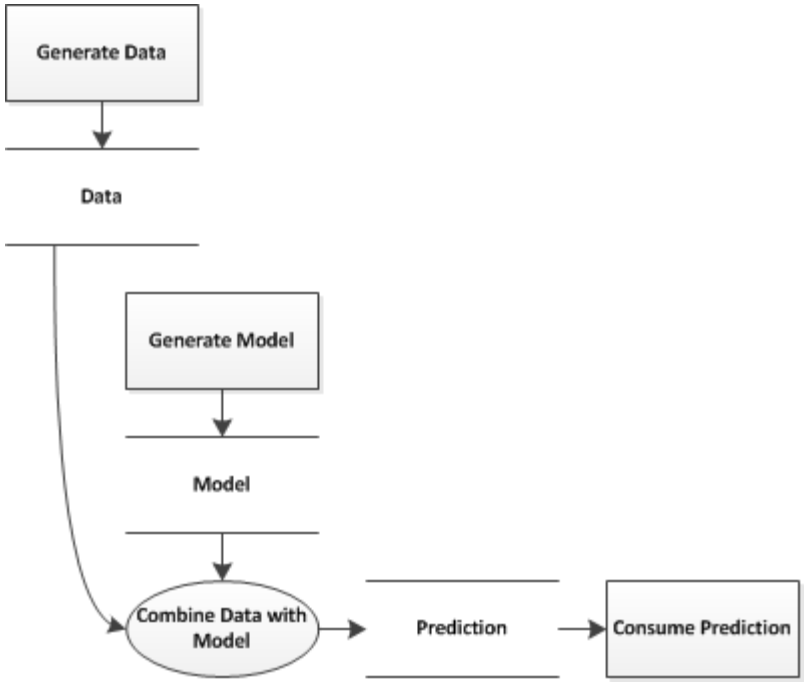
From Data to Predictions (0)

From Data to Predictions (1)



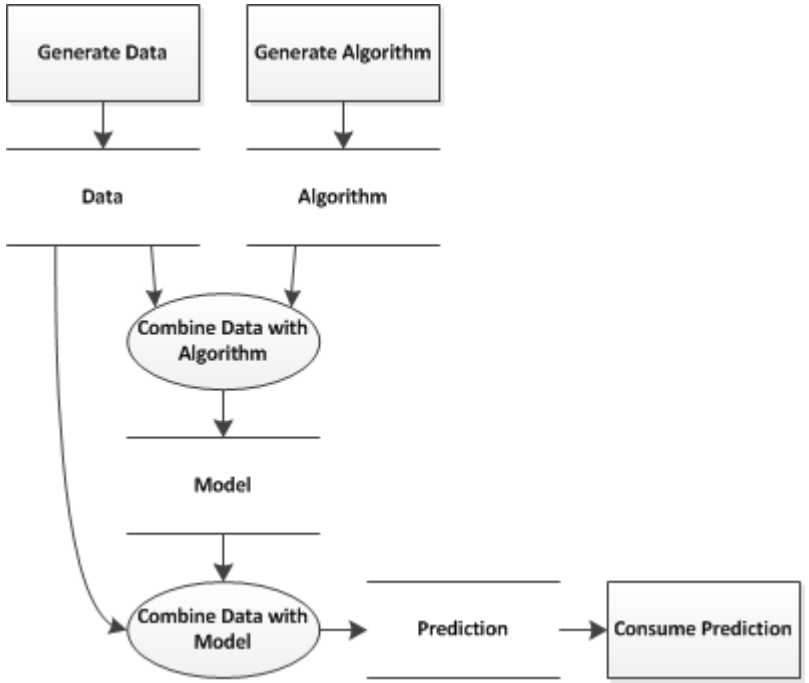
Data + Algorithm → Model

From Data to Predictions (2)



Model + Data → Prediction

From Data to Predictions (3)



Data + Algorithm → Model
Model + Data → Prediction

From Data to Predictions (4)

- Pseudo Assignments (Derivations):
  - Data + Algorithm → Model
  - Model + Data → Prediction

- Create Model from Algorithm and Data
  - Example Algorithm: Logistic Regression
    - Create Model: model <- glm(formula, data=trainSet, family="binomial")
- Predict from Model and Data
    - Predict: prediction <- predict(model, newdata=testSet, type="response")

Data + Algorithm → Model
Model + Data → Prediction

From Data to Predictions (5) Review

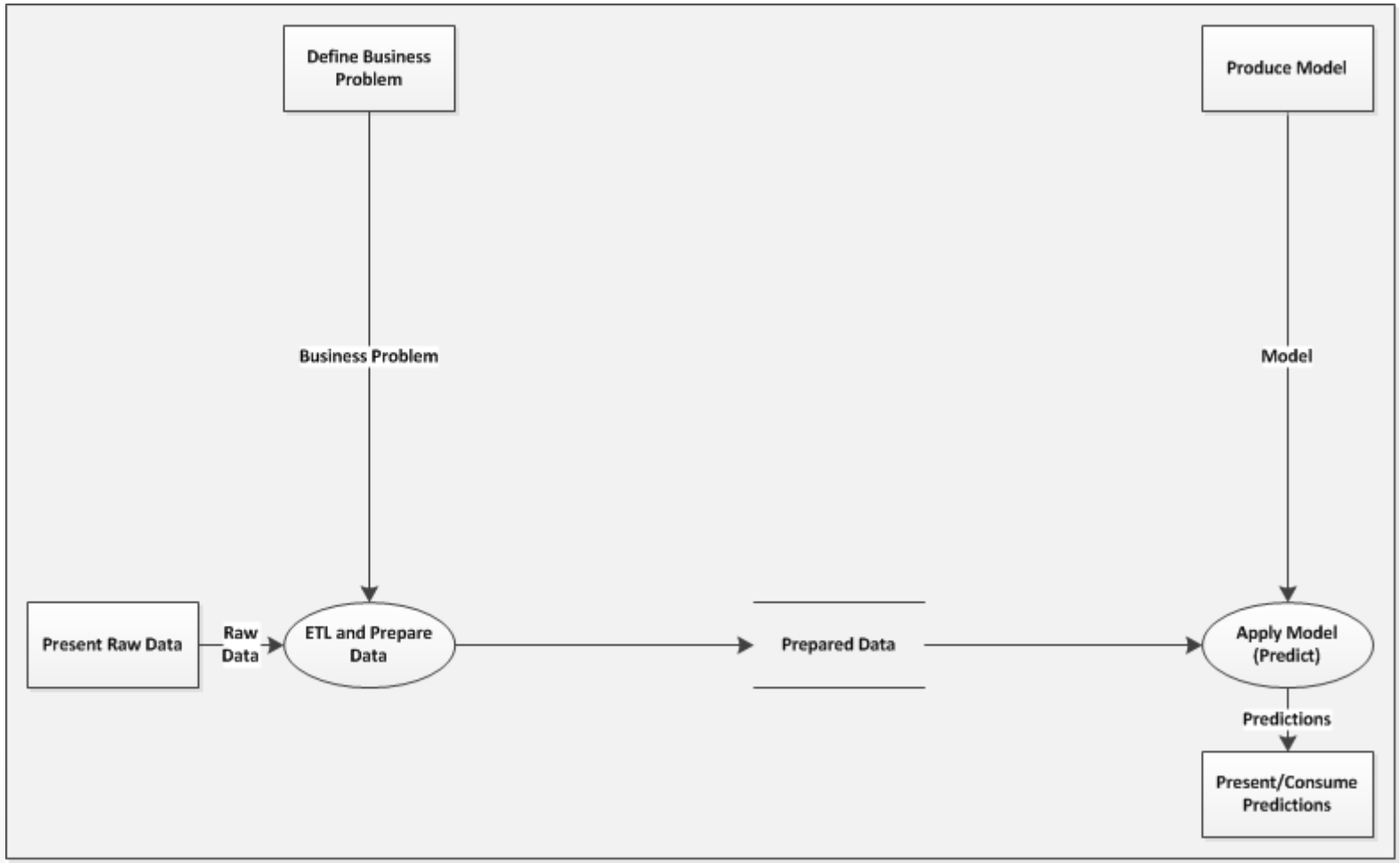
- A model or hypothesis is (best response)
  - a combination of test data and training data
  - a predictor based on data and algorithm
  - a falsification of a theory
  - a verified theory as long as the model was not falsified
- A model applied to new data leads to a (best response)
  - Prediction
  - Falsification / Verification
  - Hypothesis
  - errors
- A model applied to test data leads to a (best response)
  - Prediction
  - Falsification / Verification
  - Hypothesis
  - errors
- A hypothesis that cannot be tested
  - is a law if the data are consistent
  - is an untested hypothesis
  - is not a hypothesis
  - is a theory

(0) DFD of Supervised Learning

(1) Model Acts on Data

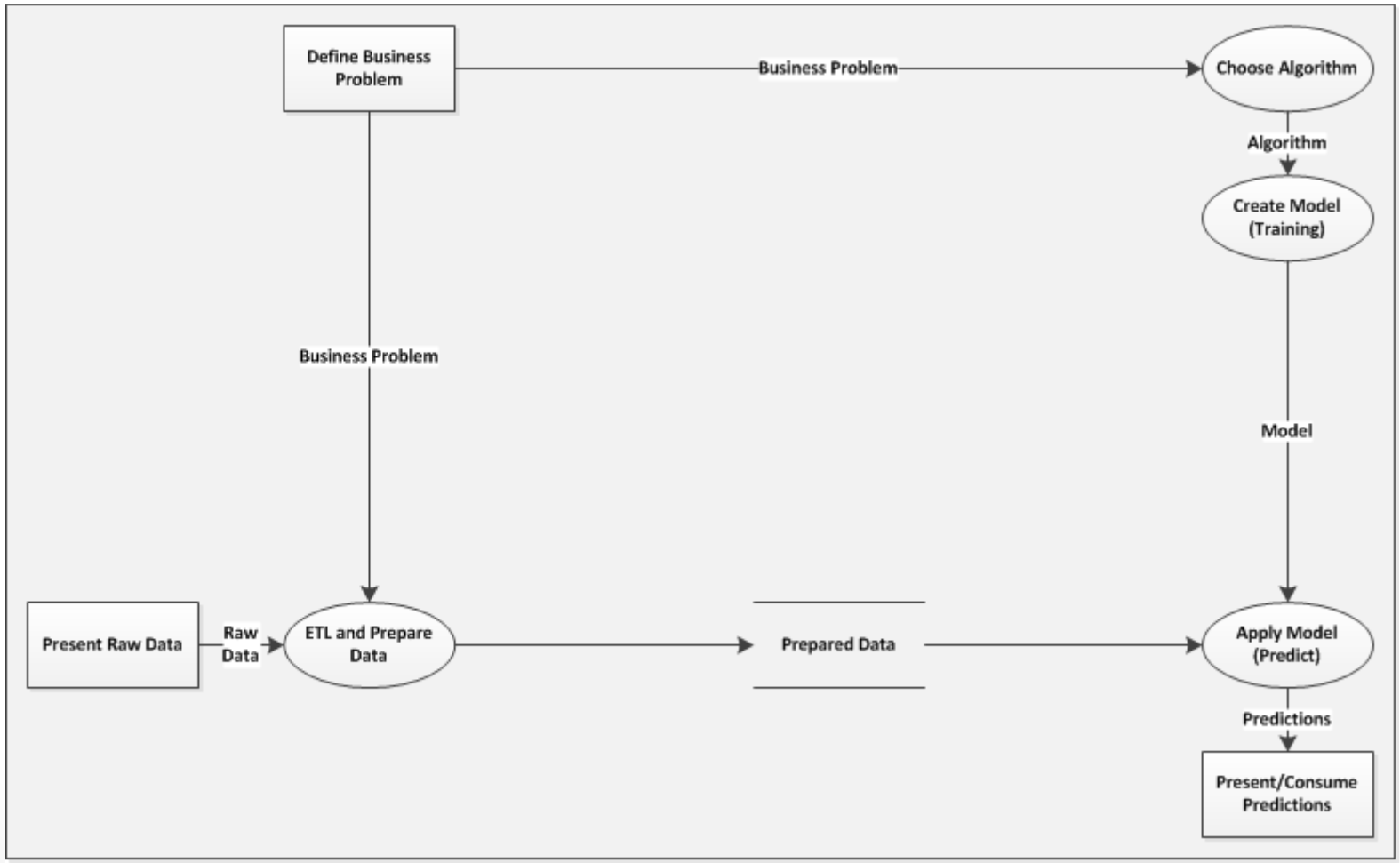
Model + Data → Prediction

(2) Data ETL and Preparation driven by Business Problem



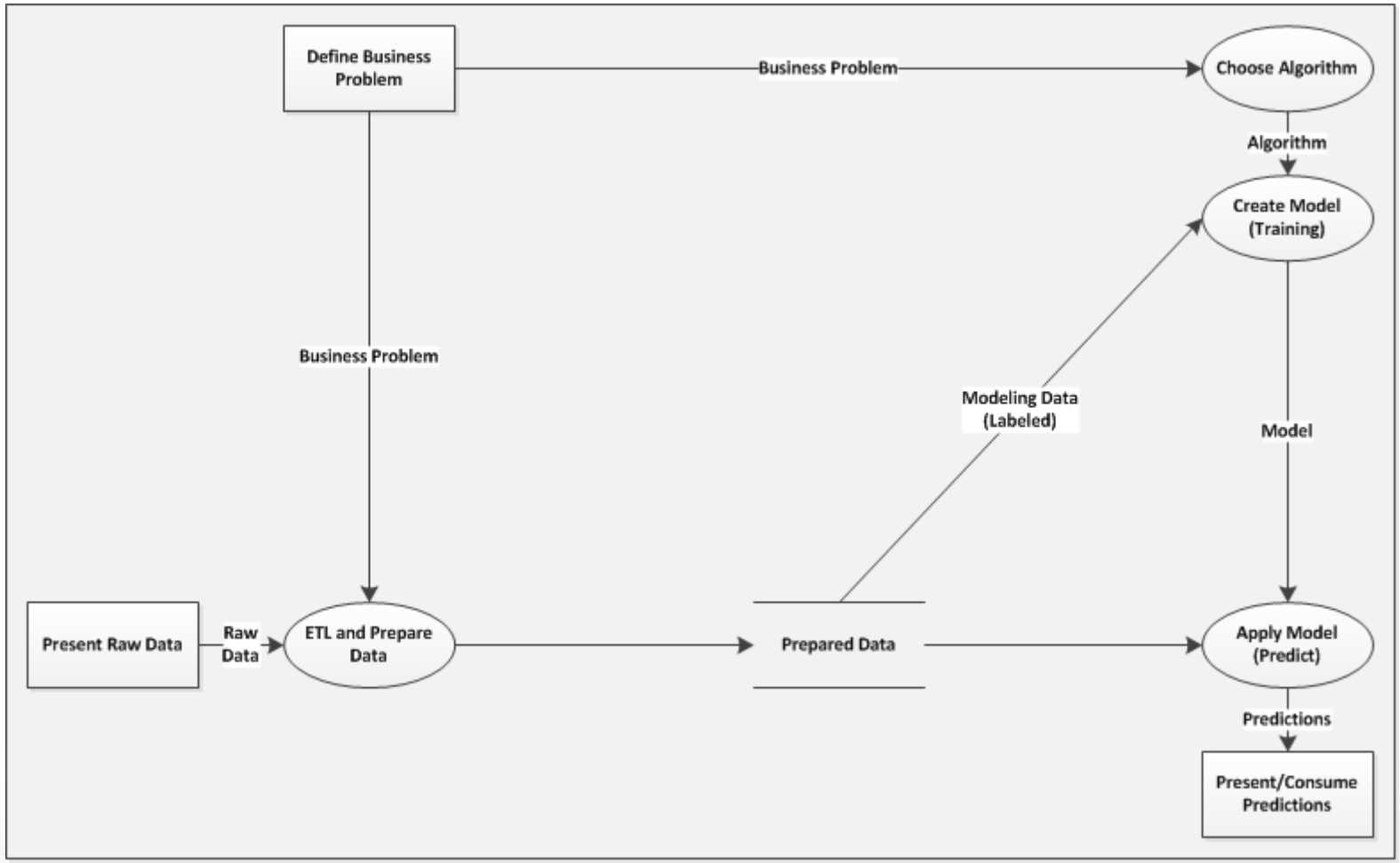
Business Problem determines ETL and Data Prep

(3) Algorithm choice driven by Business Problem



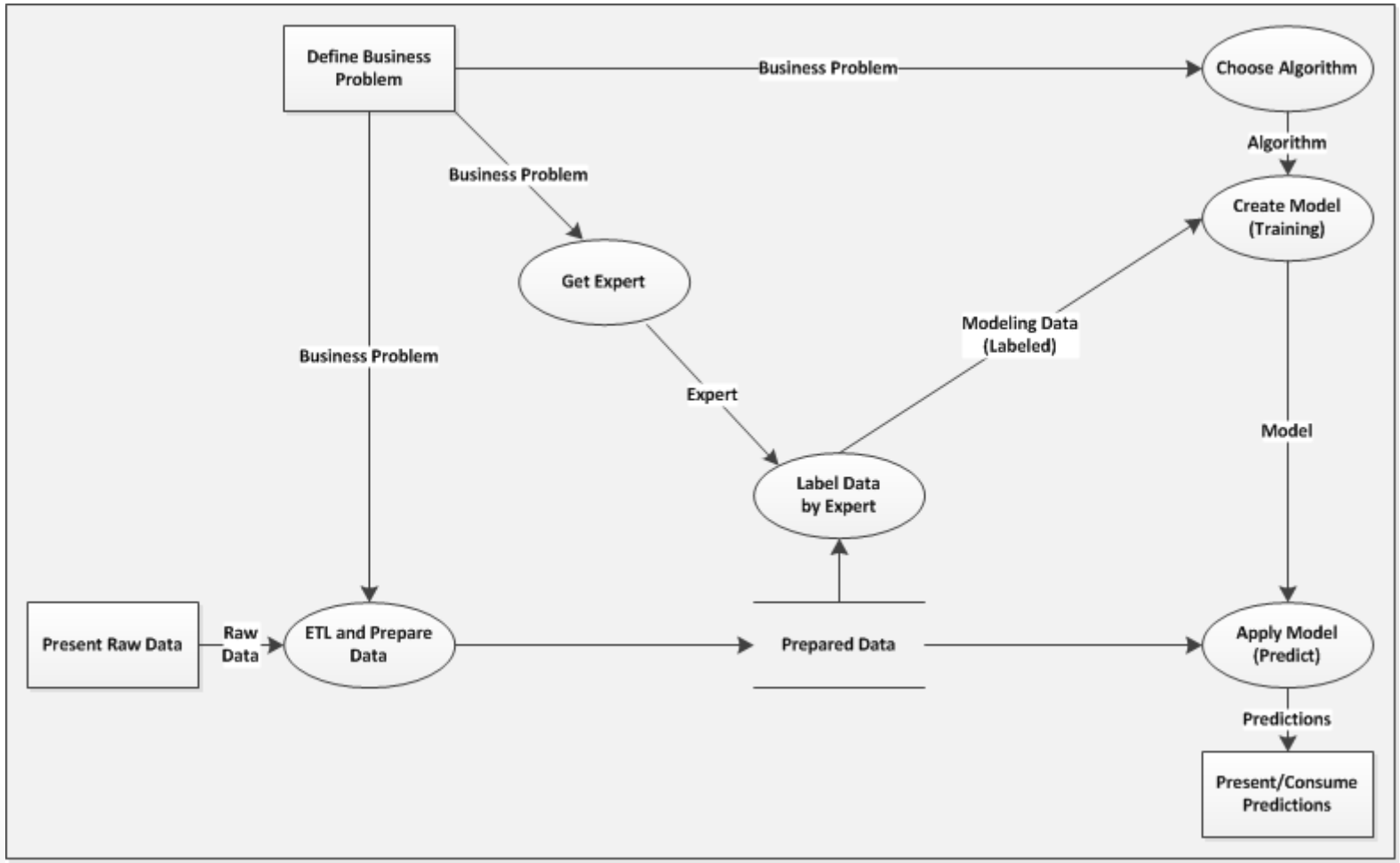
Business Problem determines the choice of Algorithm.

(4) Model Creation needs Data



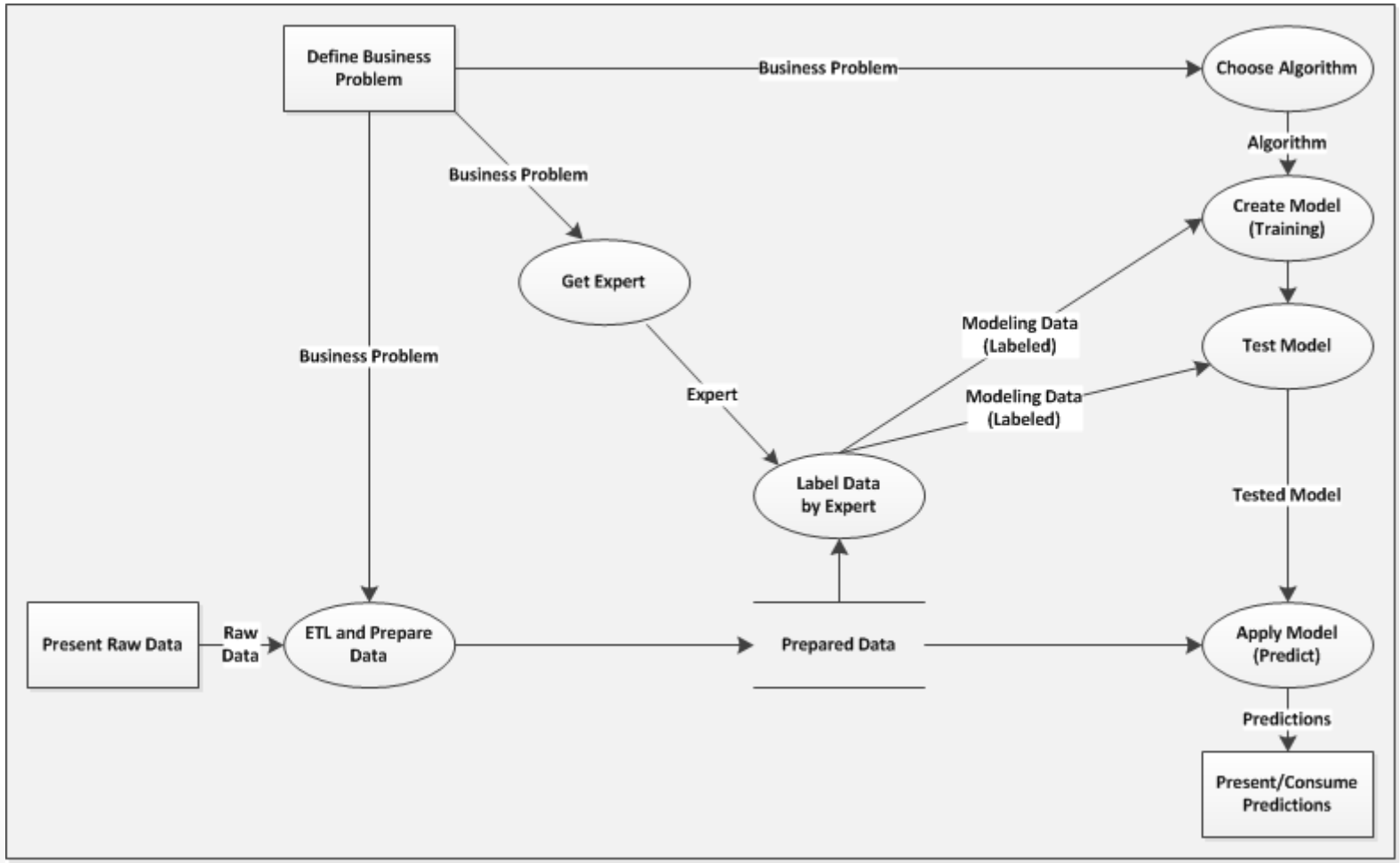
Data + Algorithm → Model

(5) Supervised Training needs Data Labeled with Outcomes



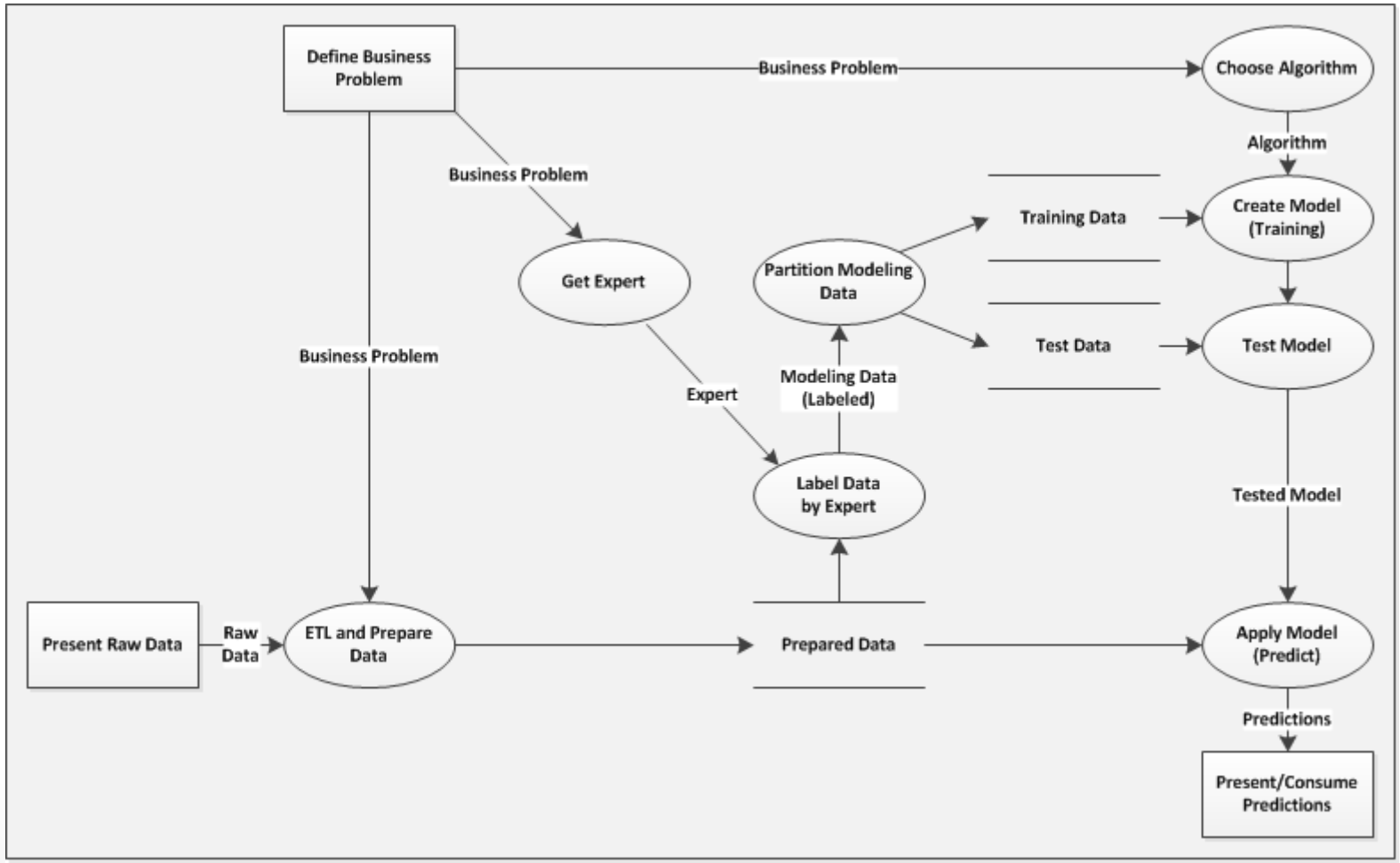
Supervised Learning requires expert labeling of data.

(6) Models need to be Tested



Do not trust predictions from an un-tested model!

(7) Training & Testing of Model use different Data



Do not test a model using training data!

Data and Models in Supervised Learning

© 2017 Ernst Henle. All rights reserved

Classification Schema

© 2017 Ernst Henle. All rights reserved

Classification Schema (0)

- Rectangular Modeling Dataset
  - Schema
    - Input columns
    - Output column (target column, outcome)
      - Classification: Category Column
  - Partition For Training and Test Data
  - Incremental Data
- Algorithm
  - Classification
    - Logistic Regression
    - Neural Network
    - Decision Tree
    - Naïve Bayes

Classification Schema (1)

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

Classification Schema (2)

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.

Classification Schema (3)

Column	Column	Column	Column 4	Column 5	Column 6	Column 7
			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
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

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".

Target Outcome

Classification Schema (4)

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

Classification Schema (5)

Column 1	Column 2	Column 3	Colur 4			
330-3141	Seaborg	Good	0.12			
330-3150	Seaborg	No	0.98			
330-3202	Seaborg	Yes	0.24			
415-2008	Seaborg	Yes	0.25			
415-2081	Seaborg	Bad	0.24			
415-2796	Seaborg		0.41			
415-2799	Seaborg	Yes	0.92			
415-2913	Seaborg	Yes	0.37			
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

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.

Random or Keys

Target Outcome

Classification Schema (6)

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

Classification Schema (7)

Column 1	Column 2	Column 3	Colur 4			
330-3141	Seaborg	Good	0.12			
330-3150	Seaborg	No	0.98			
330-3202	Seaborg	Yes	0.24			
415-2008	Seaborg	Yes	0.25			
415-2081	Seaborg	Bad	0.24			
415-2796	Seaborg		0.41			
415-2799	Seaborg	Yes	0.92			
415-2913	Seaborg	Yes	0.37			
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

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.

Constant

Random or Keys

Target Outcome

Classification Schema (8)

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

Classification Schema (9)

Column 1	Column 2	Column 3	Colur 4			
330-3141	Seaborg	Good	0.12			
330-3150	Seaborg	No	0.98			
330-3202	Seaborg	Yes	0.24			
415-2008	Seaborg	Yes	0.25			
415-2081	Seaborg	Bad	0.24			
415-2796	Seaborg		0.41			
415-2799	Seaborg	Yes	0.92			
415-2913	Seaborg	Yes	0.37			
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

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 .

Random or Keys

Constant

Proxy

Target Outcome

Classification Schema (10)

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

Classification Schema (11)

Column	Column	Column	Column 4	Column 5	Column 6	Column 7
			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
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

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.

Random or Keys

Constant

Proxy

Continuous Input

Categorical Input

Binary Input

Target Outcome

Classification Schema (12)

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

Classification Schema (13)

	Column 4	Column 5	Column 6	Column 7
	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

Classification Schema (14)

Column 4	Column 5	Column 6	Column 7
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

Classification Schema (15)

Outcome ~ Input 1 + Input 2 + Input 3

Input 1	Input 2	Input 3	Outcome
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

Classification Schema (16)

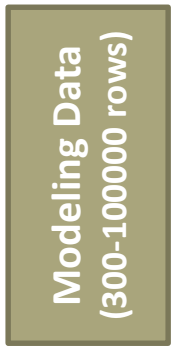
Outcome ~ Input 1 + Input 2 + Input 3

Modeling Data
(300-100000 rows)

Input 1	Input 2	Input 3	Outcome
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

Classification Schema (17)

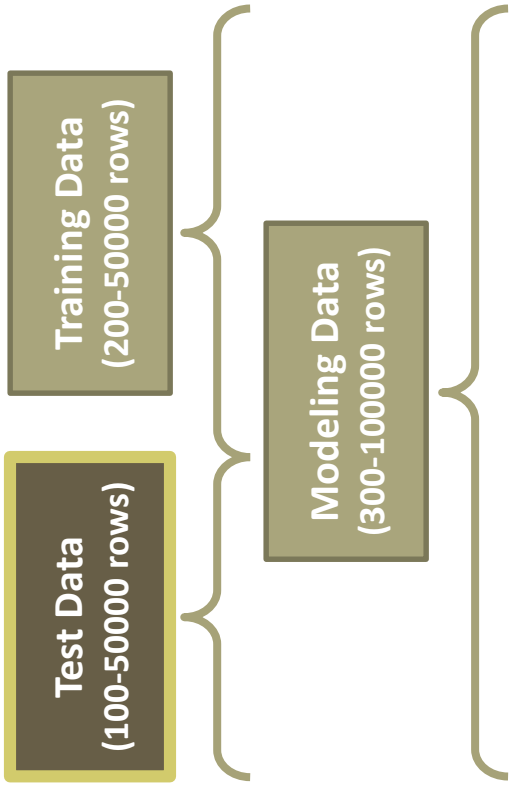
Outcome ~ Input 1 + Input 2 + Input 3



Input 1	Input 2	Input 3	Outcome
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

Classification Schema (18)

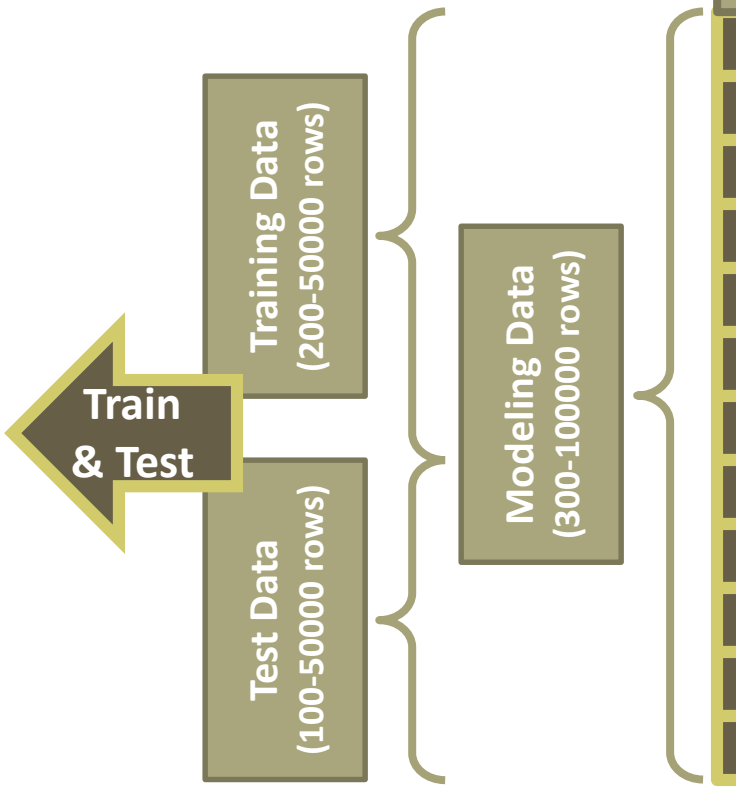
Outcome ~ Input 1 + Input 2 + Input 3



Input 1	Input 2	Input 3	Outcome
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**

Classification Schema (19)

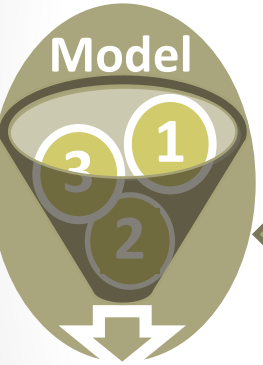
Outcome ~ Input 1 + Input 2 + Input 3



Input 1	Input 2	Input 3	Outcome
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

Classification Schema (20)

Outcome ~ Input 1 + Input 2 + Input 3



Train & Test

Training Data (200-50000 rows)

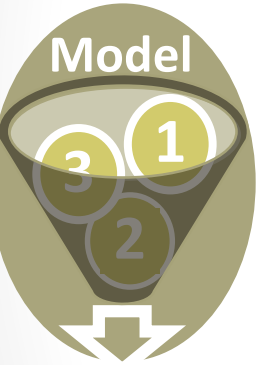
Test Data (100-50000 rows)

Modeling Data (300-100000 rows)

Input 1	Input 2	Input 3	Outcome
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

Classification Schema (21)

Outcome ~ 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".

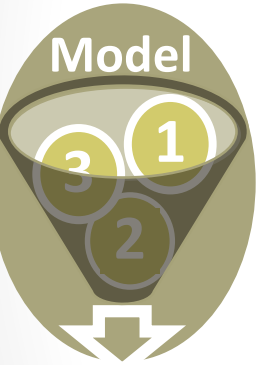
Input 1	Input 2	Input 3	Outcome
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	
0.321	blue	F	
0.098	green	F	No
0.765	red	T	No
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

Target Outcome

Operational Data (1- ∞ rows)

Classification Schema (22)

Outcome ~ Input 1 + Input 2 + Input 3



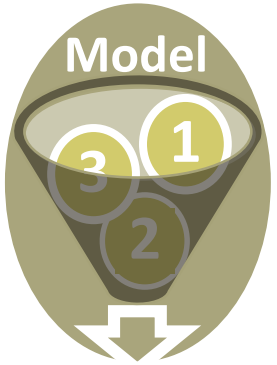
Target Outcome

Operational Data (1- ∞ rows)

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

Classification Schema (23)

Outcome ~ Input 1 + Input 2 + Input 3

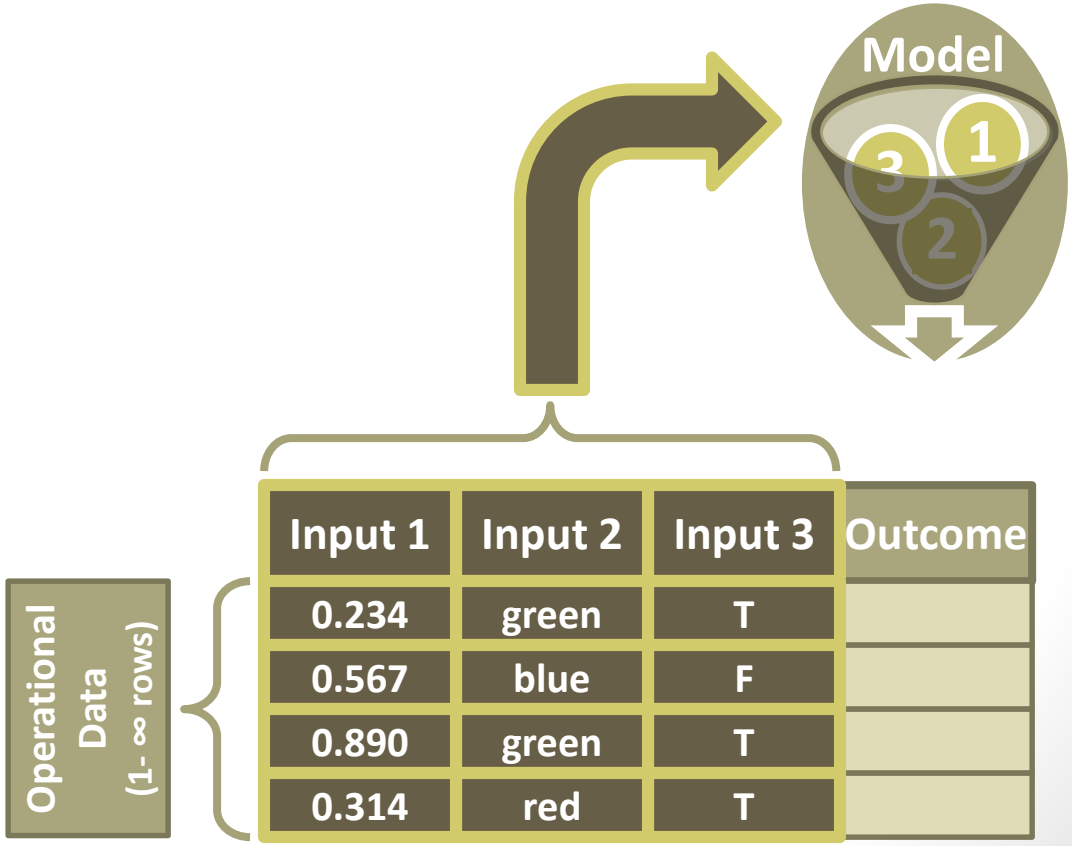


Operational Data (1- ∞ rows)

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

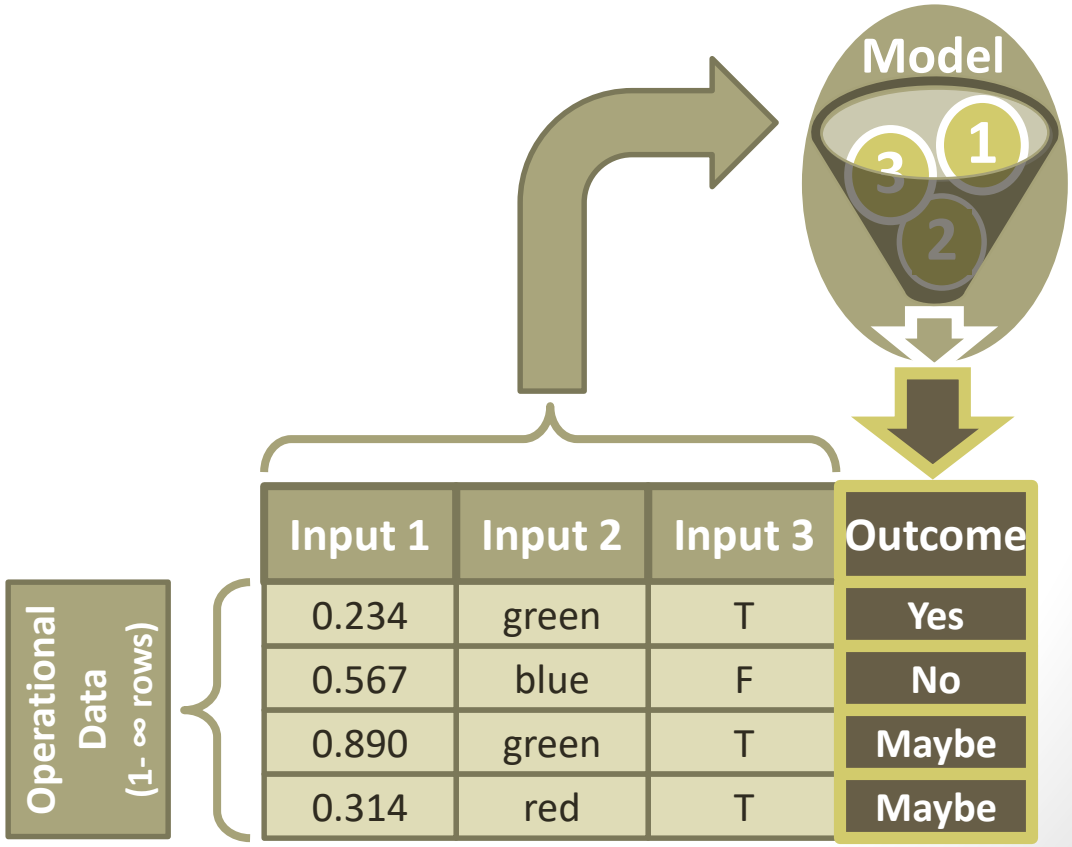
Classification Schema (24)

Outcome ~ Input 1 + Input 2 + Input 3



Classification Schema (25)

Outcome ~ Input 1 + Input 2 + Input 3



Classification Schema (26)

- 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, 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.
- Proxy Column
  - A proxy column is a column that predicts too well. It is too good to be true. Something from the target leaked. This is also called target leakage. The leaked information is "not fair" to use in modeling. Values for that attribute will not be available when you want to predict the target outcome from operational data.
- Key Column
  - In principle, a key column should not affect the model's prediction. The relationship between a key and any other attribute should be random. In practice, the algorithm will find a pattern in the key column and train on this pattern. This pattern is likely to be fortuitous, that means: random. The pattern will not hold for test data or when the model is applied. As a consequence, the key column will affect the model in a bad way.
- Constant Column
  - A constant column should have no affect on the model's predictions. The constant column may increase computation time and cause other problems. It is standard practice to remove all constant columns prior to modeling.

Classification Schema

© 2017 Ernst Henle. All rights reserved