Introduction to Data Science

Lecture 04; April 20th, 2015

Ernst Henle
ErnstHe@UW.edu
Skype: ernst.predixion

Agenda

- Social Interactions and Announcements
  - LinkedIn
  - Philosophy of Instruction
- Review
- Quiz 04a (MATLAB and normalization)
- In-Class Exercise: KMeans in R
- Break
- Data Structures
- Quiz 04b (Machine Learning terminology)
- Data and Models in Supervised Learning
- Break
- Data and Models in Supervised Learning (cntd)

Social Interactions and Announcements

- LinkedIn
  - Discussions
- Personalized Feedback

- Announcements:
  - Guest Lecture: May 11th 1-hour by Ben Olsen on "Design Concepts for Visualization"
  - Guest Lecture: Tentative May 18th 1-hour by Marius Marcu "Business Aspects of Data Science"

Review

- Normalization
  - Homework: simpleKMeansFinished.m
    - Questions in items 1-3
    - Code
  - Questions:
    - Why were the centroids normalized by the standard deviation and mean of the points and not the centroids?
    - Why where the points not de-normalized?
- Clustering
  - Real world example

Review: Homework Questions

1. Answer these questions:
   a. Why is normalization important in K-means clustering?
      Answer: So that the dimensions (data attributes) have similar scales.
   b. How do you encode categorical data in a K-means clustering?
      Answer: Category attributes are binarized
   c. Why is clustering un-supervised learning as opposed to supervised learning?
      Answer: The algorithm is not told what is observed or what the "goal" is. There is no expert label.

Review: Homework Questions

2. Given the following: simpleAssignToCentroids assigns the 17th point to a centroid by measuring the distance of the 17th point to each centroid. The centroid with the smallest distance to the 17th point is the point's centroid. How does simpleKMeans know which centroid was chosen for the 17th point? (Answer in one sentence or less by describing the data structure)

Open Octave and simpleKMeans.m

Answer:

simpleAssignToCentroids returns a vector where the value at index i is the cluster number (like 1, 2, or 3) for point(i, :); In this case i is 17

Review: Homework Questions

3. Given the following: simpleDetermineCentroids determines centroid for cluster 2 by finding the mean of all points that belong to cluster 2. How does simpleKMeans know which returned centroid is the one for cluster 2? (Answer in one sentence or less by describing the data structure) .

Answer:

simpleDetermineCentroids returns a matrix called centroids where row i is the centroid for cluster i. In this case i is 2.

Review: MATLAB/Octave

- Start Octave
- Open simpleKMeansFinished.m
- Run
  - simpleKMeansTests.m
  - testSimpleKMeans_zScore.m

Review: Homework Normalization

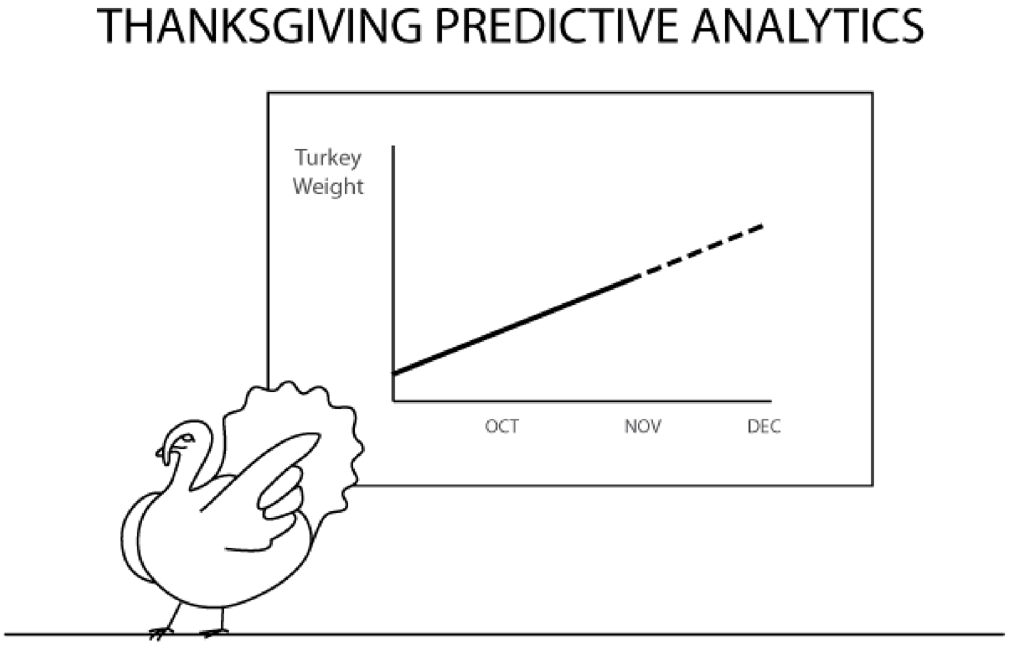
```
% Parameters for normalization and de-normalization
% Determine the mean and standard deviation of the points in both dimensions
meanPoints = mean(points);
sigma = std(points) ;

% Normalize points
% for each dimension for each point subtract away its mean and then divide by the standard deviation
points = (points .- meanPoints ) ./ sigma;

% Normalize Centroids
% For each dimension and centroid subtract away the mean of the dimension
% and then divide by the standard deviation of the dimension
centroids = (centroids .- meanPoints ) ./ sigma;
```

Review: Homework De-normalization

```
% Denormalization
% for each dimension for each centroid multiply by standard deviation and then add mean
centroids = (centroids .* sigma) .+ meanPoints ;
```

Quiz 04a

- UW Data Science 2015 Quiz 04
- https://catalyst.uw.edu/webq/survey/ernsthe/268522
- You may want to use Octave/MATLAB during the Quiz.



timoelliott.com

"I see no reason why excellent growth shouldn't continue..."

In-Class Exercise and Homework Assignment

Write K-Means in R: Kmeans_Skeleton.R

- Write a version of K-Means in R and name the file KMeans.R. The function signatures should be the same as those in Kmeans_Skeleton.R, Specifically, implement
  - **KMeans <- function(observations = sampleObservations, clusterCenters = centersGuess)**
  - **findLabelOfClosestCluster <- function(observations = sampleObservations, clusterCenters=centersGuess)**
  - **calculateClusterCenters <- function(observations=sampleObservations, clusterLabels=labelsRandom)**
- You can use Kmeans_Skeleton.R as a template and replace all lines that say: "**Put code in place of this line**". Execute the built in tests and verify that your code works:
  - **ClusterPlot()**
  - **findLabelOfClosestCluster()**
  - **calculateClusterCenters()**
  - **KMeans()**

Break

- **Big Data Humor: Top 10 Ways You Know You're a Data Scientist**
  - http://inside-bigdata.com/2013/10/28/big-data-humor-top-10-ways-know-youre-data-scientist/

Data Structures

Terminology and Concepts (1)

- Data
  - Dataset is a set of Data. A set implies a commonality. The commonality is expressed as a type or a relation.
  - A data type provides structure and meaning to the data. Just like there is no such thing as un-structured data, there is no such thing as un-typed data. Data can be insufficiently typed and structured.
- Rectangular Data
  - Datasets are often 2D matrices, which are organized into rows and columns. The column and row order is not important .
  - Columns are named with a header; A columns may be also referred to as an attribute or field. The number of columns is often called the dimensionality of the data.
  - Rows are not named. A row is often referred to as a case or observation. Number of rows in a category is called support.
- Data dimensionality
  - A data frame or a table can be considered a sparse multi-dimensional matrix
  - The dimensionality for un-supervised learning is #columns
  - The dimensionality for supervised learning is #columns - 1 because one column represents the value and not the dimension. This structure is very similar to a star schema

Terminology and Concepts (2)

- Predictive Analytics (Machine Learning , Artificial Intelligence)
  - Algorithms (often called Methods)
    - Supervised Learning
      - Classification
      - Estimation
    - Unsupervised Learning
      - Clustering
      - Association (Market-basket analysis)
      - Anomaly detection
    - Forecasting (Time Series)

Terminology and Concepts (3)

- Supervised Learning Algorithms
  - Classification Algorithms predict classes or categories
    - Logistic Regression (Deterministic)
    - Decision Trees (Deterministic)
    - Naïve Bayes (Deterministic)
    - Neural Net (Non-Deterministic)
  - Estimation Algorithms predict continuous (numeric) values
    - Generalized Linear Modeling abbreviated: GLM (Deterministic)
      - Linear Regression
      - Logistic Regression
    - Regression Trees (Deterministic)
    - Neural Net (Non-Deterministic)

Terminology and Concepts (4)

- Un-Supervised Learning Algorithms
  - Segmentation Algorithms, also called Clustering, create clusters or segments. These clusters can be thought of as categories.
    - Mixture of Gaussians aka Probabilistic (Deterministic)
    - Hierarchical (Deterministic)
    - K-Means (Non-Deterministic)
  - Association Algorithms associate or link items by a common attribute called the transaction ID.
    - Market Basket Analysis (Deterministic)
    - Affinity Analysis (Deterministic)
  - Anomaly Detection is used to find unusual or anomalous data like outliers

Terminology and Concepts (5)

- Forecasting (Time Series) is used to estimate future values based on past behaviors.
  - ARIMA / Auto ARIMA
  - Survival Analysis

Major types of Data Sets

- **Univariate**
- **Rectangular**
- **Time Series**
- **Nested**
- **Graphs (later in the course)**

Univariate (1)

- A collection of data. The data do not have a particular order. Example: Students' age. This type of data is often (mistakenly) called unstructured data, especially when the values are strings of indeterminate length. (Ragged Array)
- Example usage: anomaly detection.

Univariate (2)

Parent Income
40,000
53,000
60,000

Rectangular Data (1)

- The data set has columns and rows. Each cell has a value or is null.
- A Rectangular dataset is often called a matrix, data frame, or table.
- Example usage: classifications and estimations

Rectangular Data (2)

- Columns have descriptive headers like: Name, Age, Height, Weight of each student.
- Columns are also called attributes and fields.
- All values within a column have the same data type

Rectangular Data (3)

- Rows generally do not have names. If a row has a name, then the names could be considered another column.
- Rows are also called observations or cases
- The number of rows in a category is called support.

Rectangular Data (4)

ID	IQ	Parent Income	Moral Support	Gender	College Plans
835	107	40,000	Yes	Female	Applied
016	99	53,000	Yes	Male	Applied
490	105	60,000	No	Male	Did not apply

Time Series (1)

- A rectangular data set where the independent variable is time. The observations are sorted by time.
- Example usage: forecasting.

Time Series (2)

Date	Red Wine Sales	White Wine Sales	Rose Sales
1/22/13	$103.00	$300.50	$19.00
1/23/13	$35.50	$204.00	$44.00
1/24/13	$217.50	$74.50	$80.00

Nested (1)

- A rectangular data set where the rows have a table. Such a table can have a flat representation.
- Example usage: associations (shopping basket analyses).

Nested (2)

Transaction ID	**Item**
1	Milk
	Sugar
2	Lumber
3	Milk
	Sugar
	Flour

Nested (3)

Transaction ID	Item
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

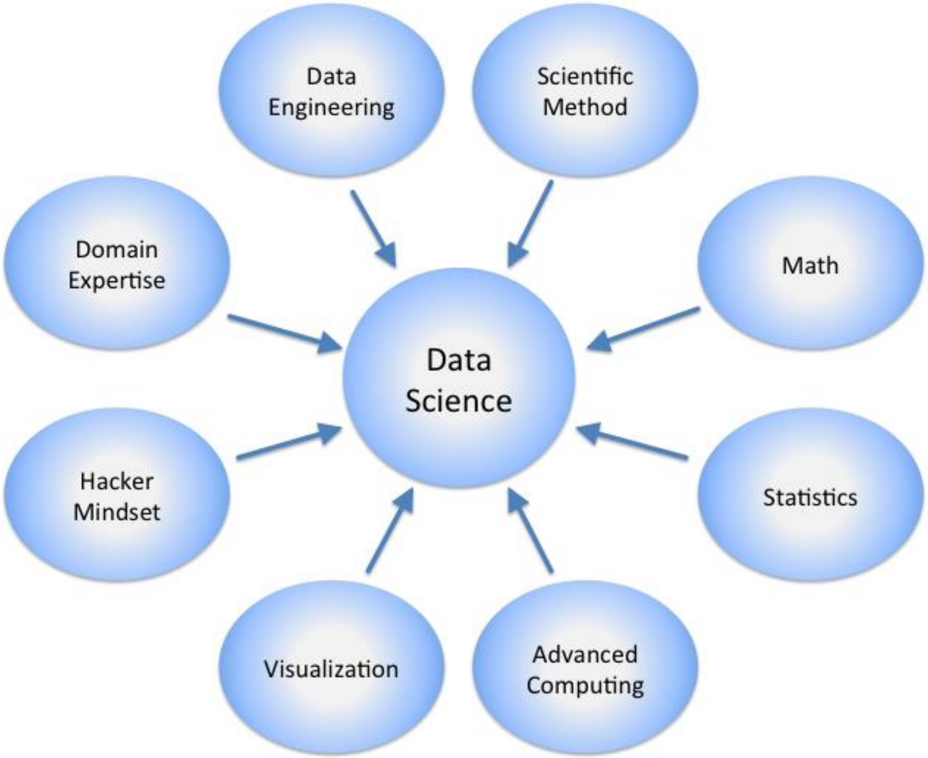
Nested (4)

Transaction ID	Item
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

Data Structures

Quiz 04b

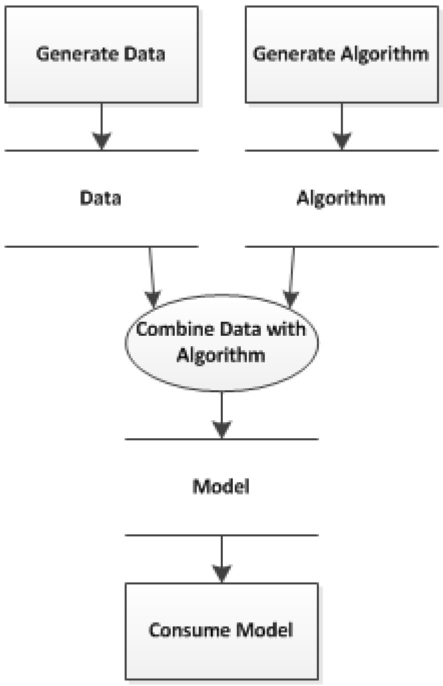
- Data Science UW 2015 Quiz 04b
- https://catalyst.uw.edu/webq/survey/ernsthe/268525
- Check your answers with others. Use a search engine to clarify new terms. We did not cover everything in class!



Data and Models in Supervised Learning

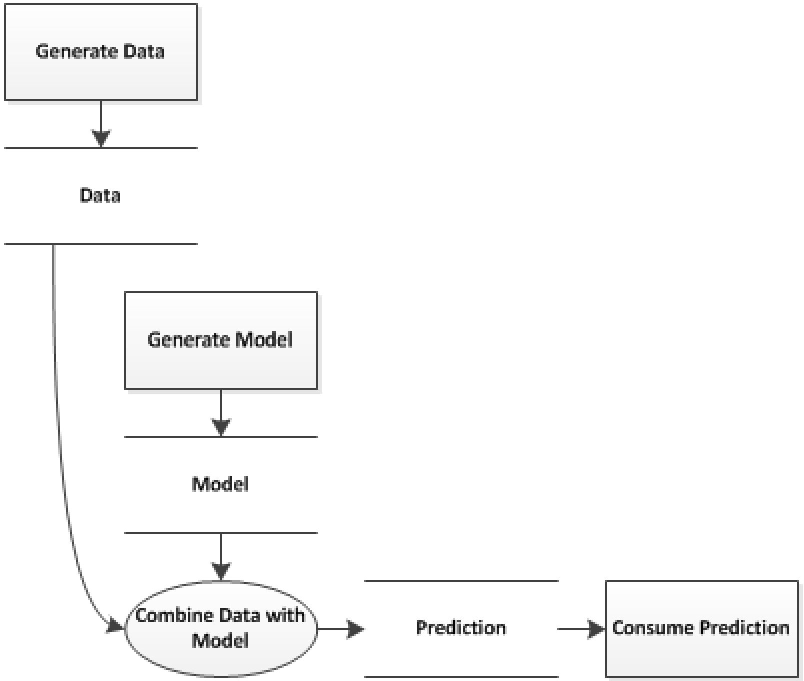
From Data to Predictions (0)

From Data to Predictions (1)



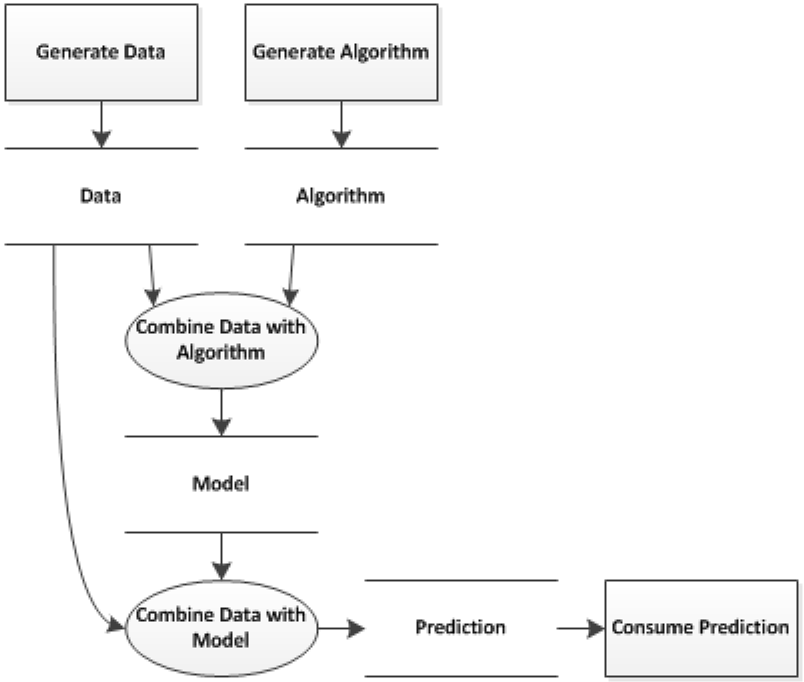
Data + Algorithm → Model

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

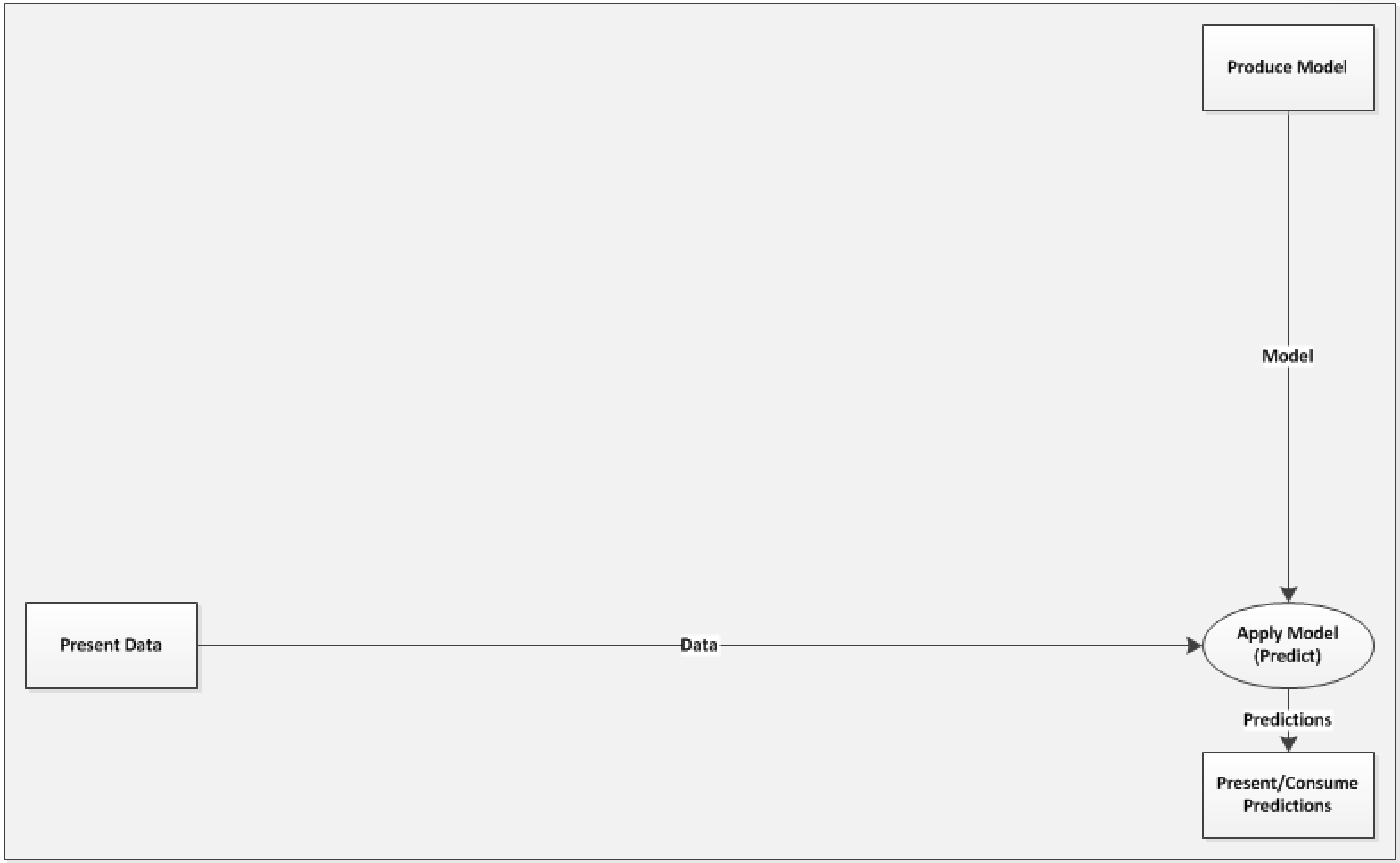
- A model or hypothesis is (best response)
  - a combination of test data and training data
  - a predictor based on data and algorithm
  - a falsification of a theory
  - a verified theory as long as the model was not falsified
- A model applied to new data leads to a (best response)
  - Prediction
  - 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

Break

- Colbert on Predictive Analytics
  - http://www.colbertnation.com/the-colbert-report-videos/408981/february-22-2012/the-word---surrender-to-a-buyer-power

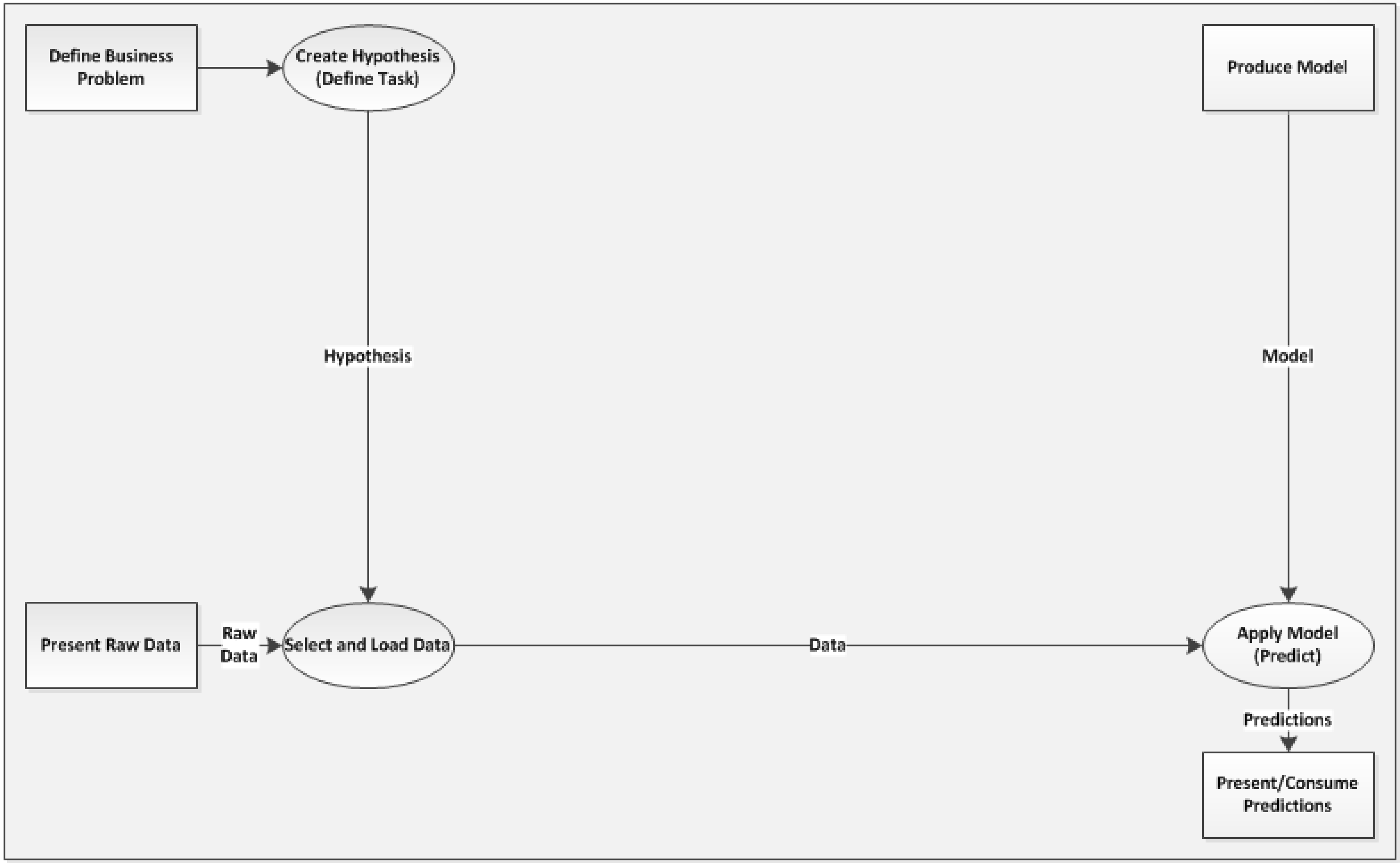
(0) DFD of Supervised Learning

(1) Model Acts on Data



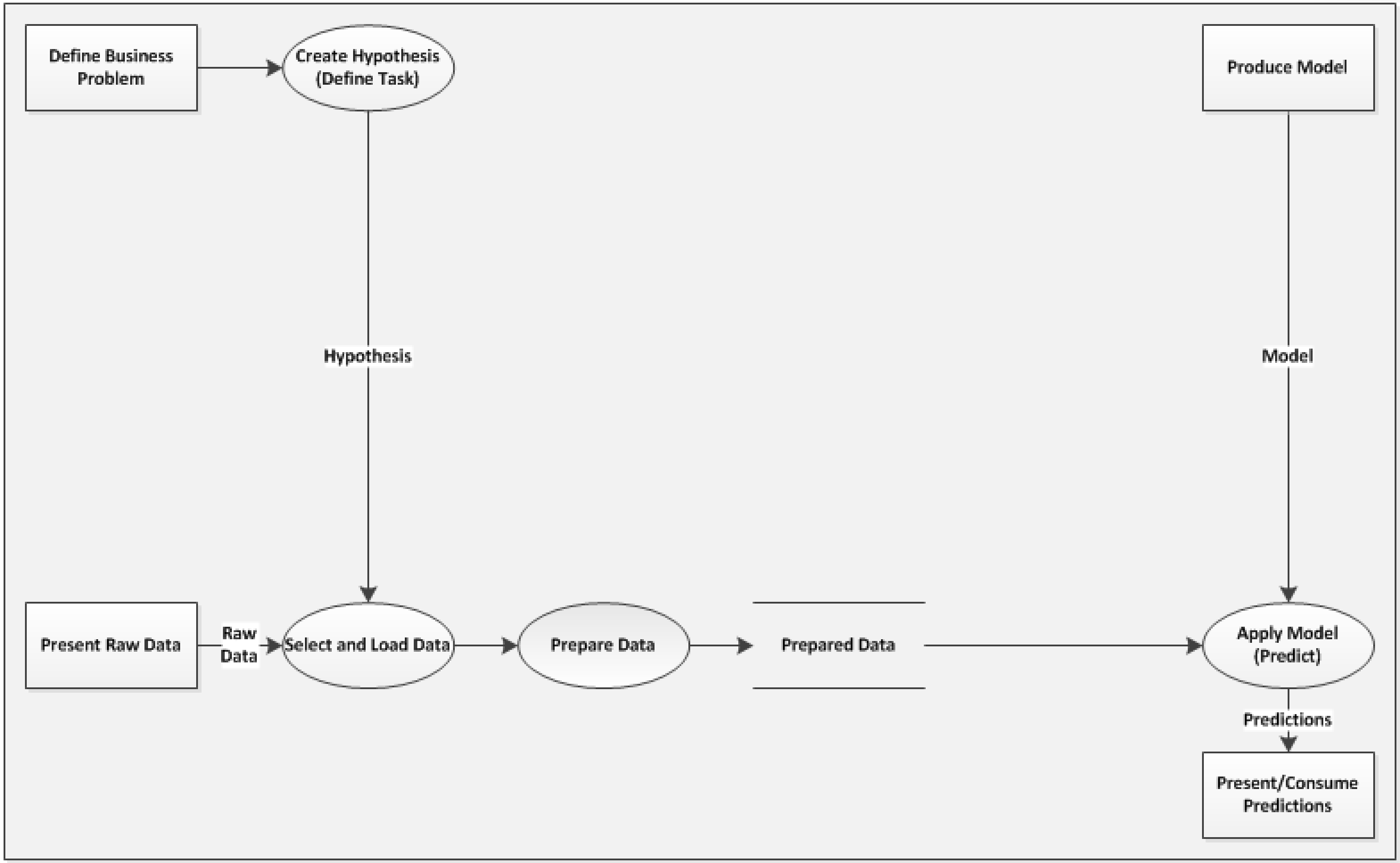
Model + Data → Prediction

(2) Data Selection Reflects Hypothesis / Business Problem



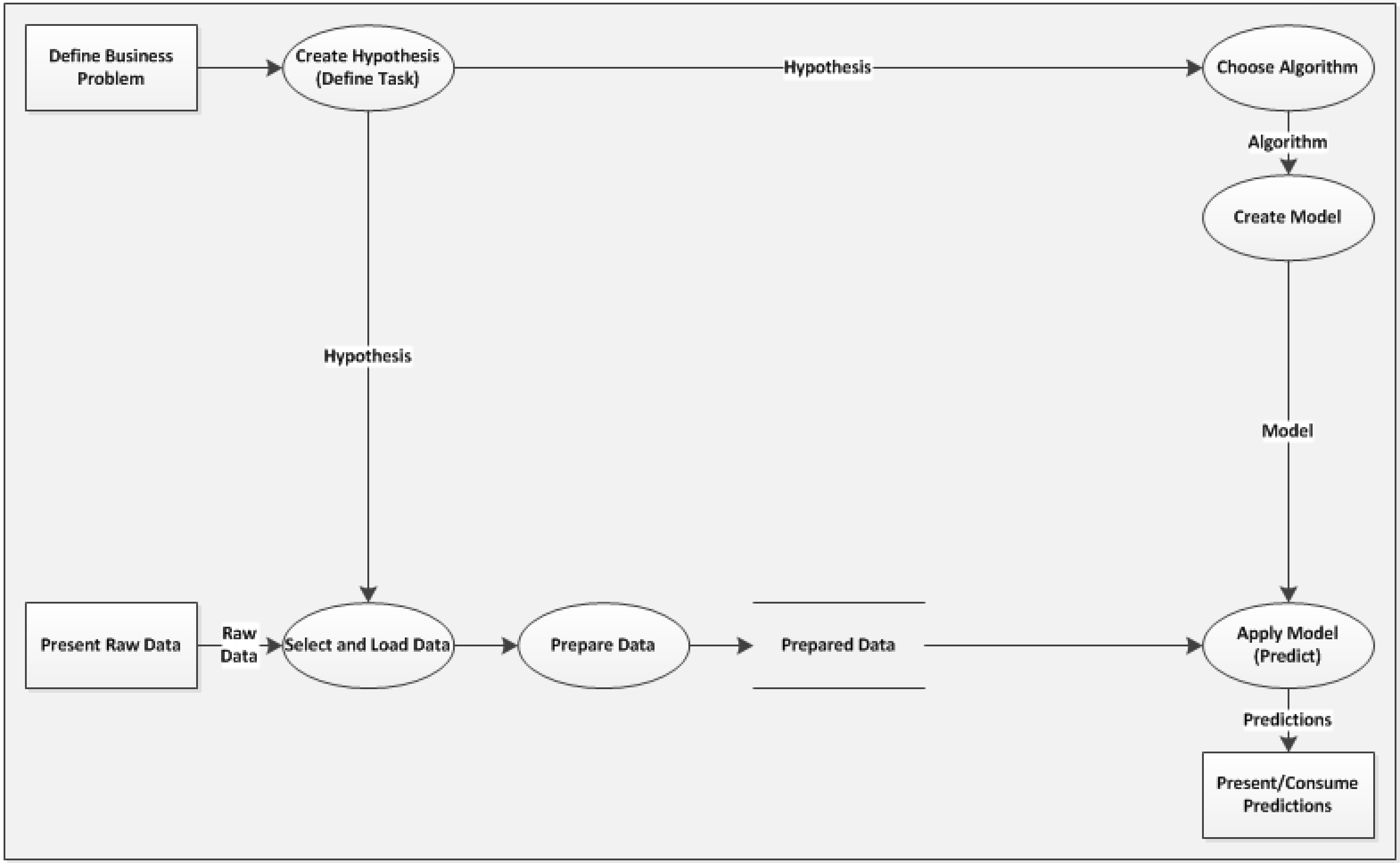
Hypothesis determines what data are loaded

(3) Data Needs Preparation



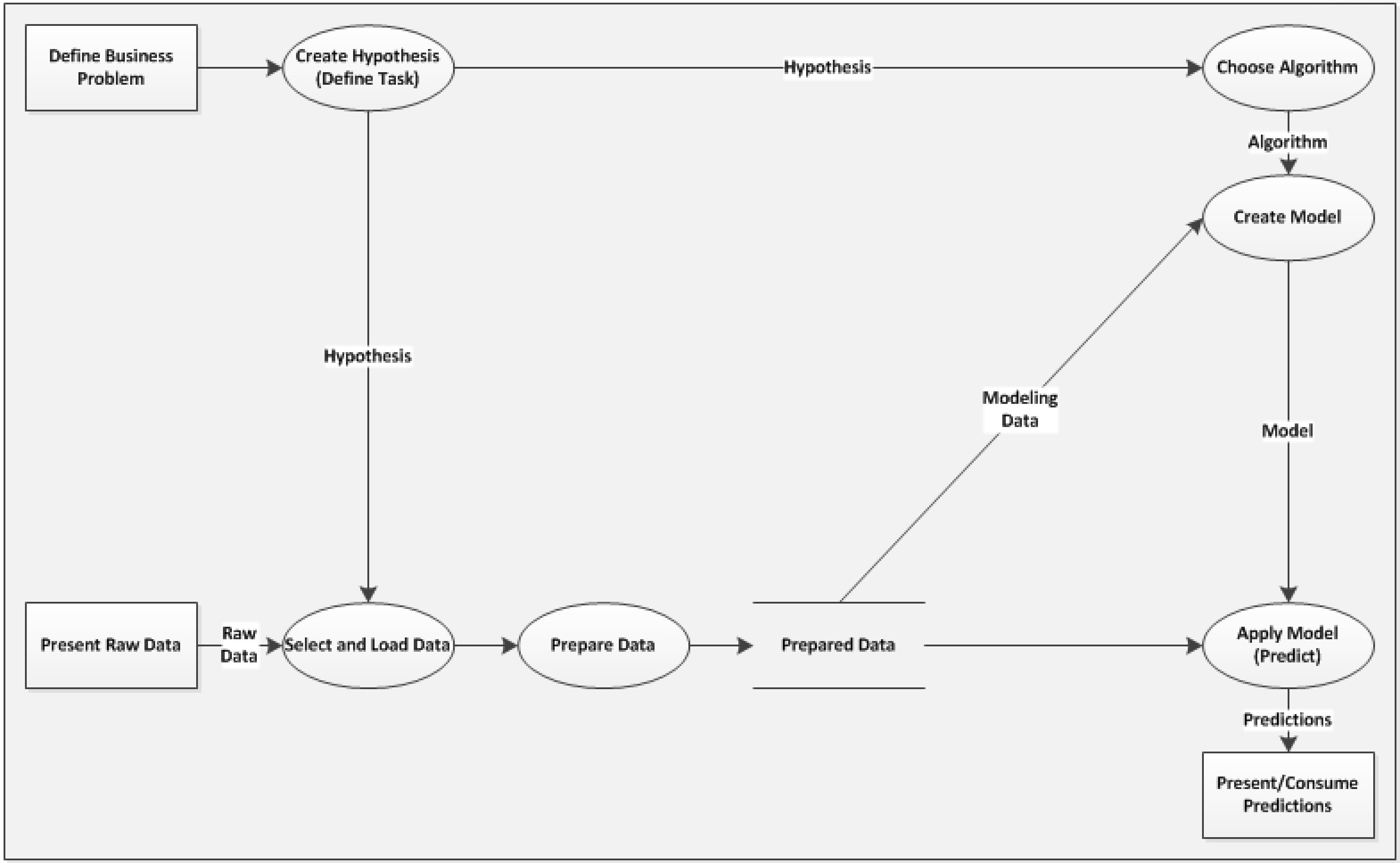
Data need to be prepared for use by a model.

(4) Model Creation Reflects Hypothesis / Business Problem



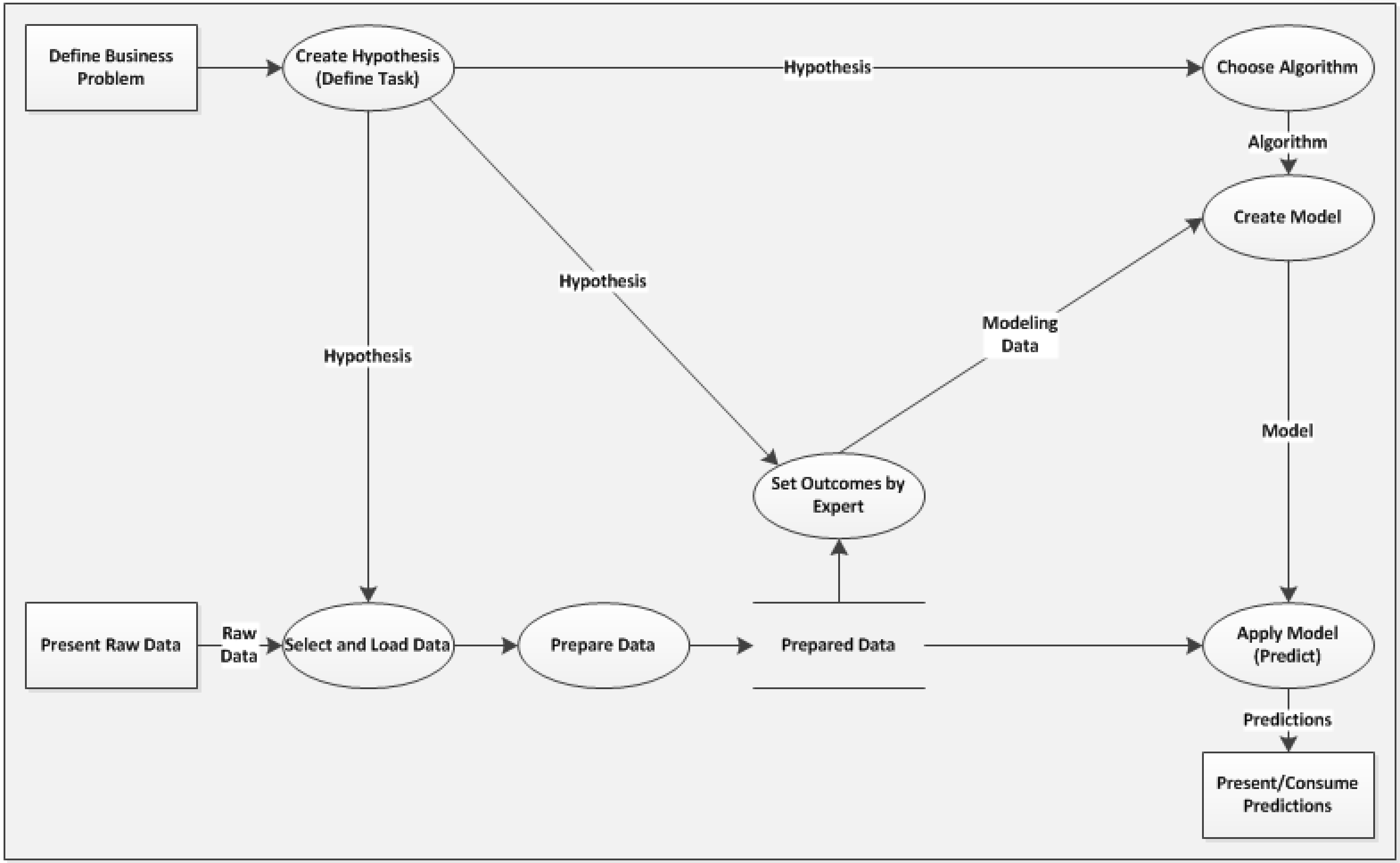
Hypothesis determines the choice of Algorithm.

(5) Model Creation needs Data



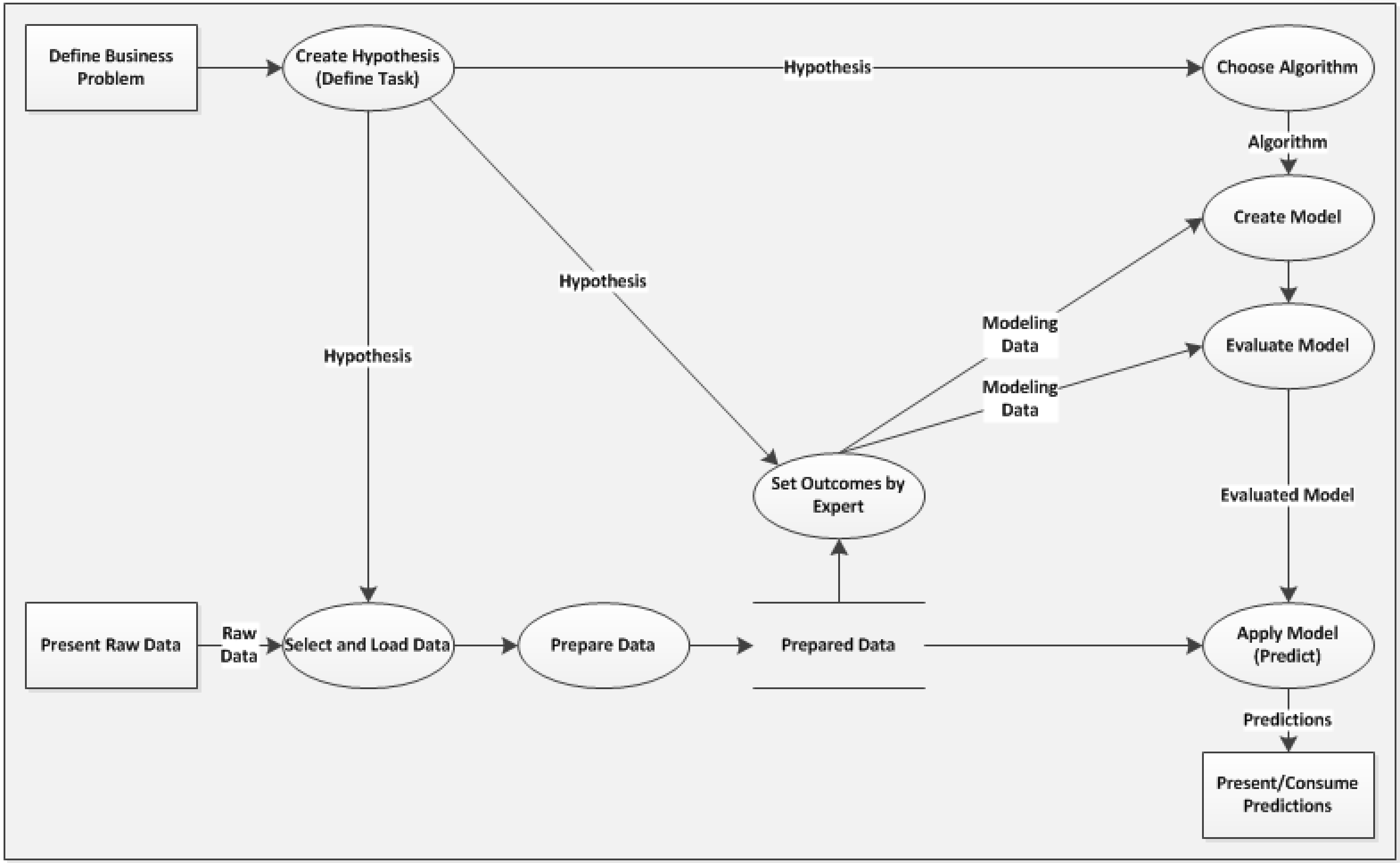
Data + Algorithm → Model

(6) Supervised Training needs Data Labeled with Outcomes



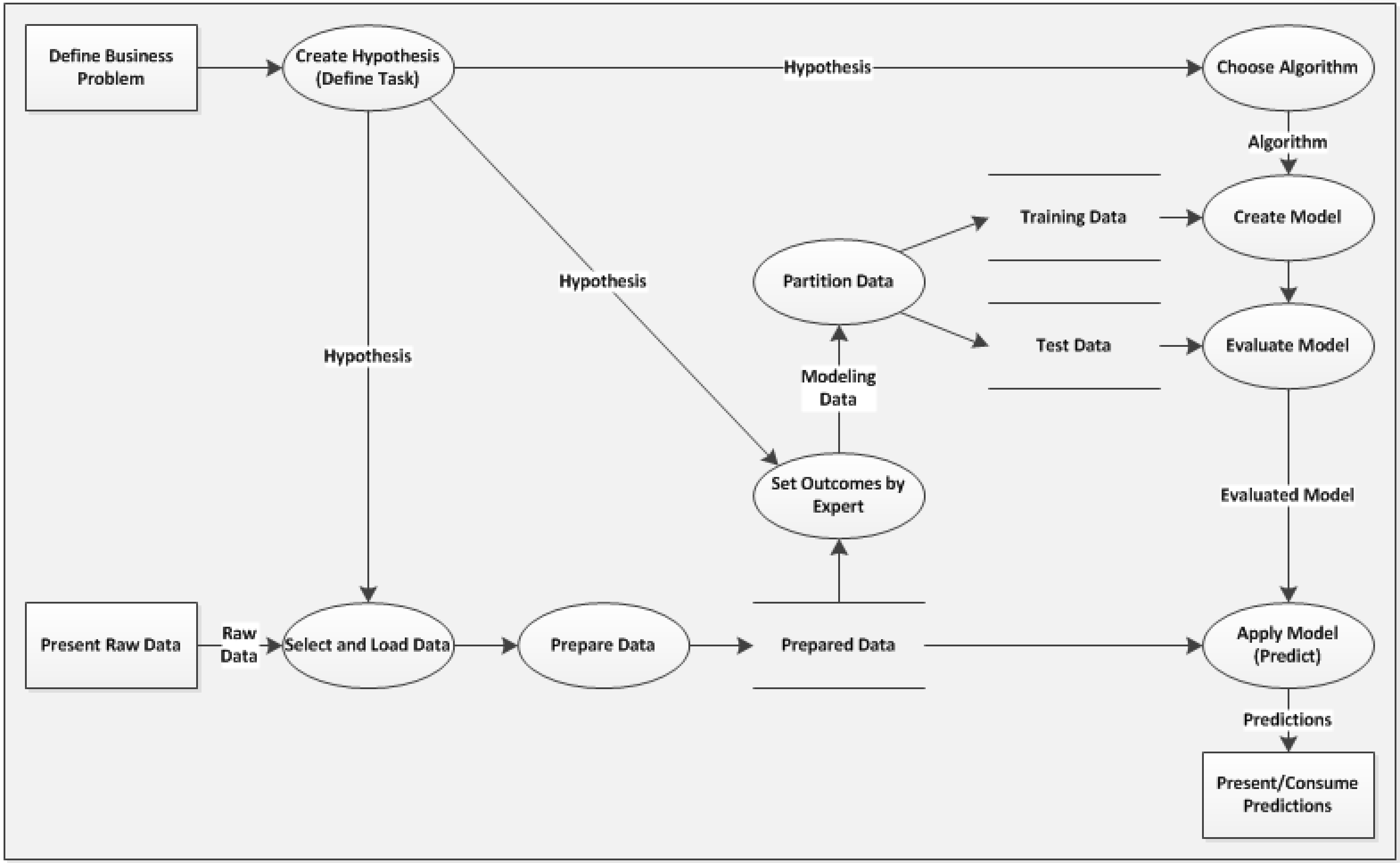
Supervised Learning requires expert labeling of data.

(7) Models need to be Evaluated



Do not trust predictions from an un-tested model!

(8) Creation & Evaluation of Model may not use same Data



Do not test a model using training data!

Data and Models in Supervised Learning

Assignments

1. K-Means in R
   1. Review the slide: “In-Class Exercise and Homework Assignment”
   2. Copy Kmeans_Skeleton.R to Kmeans.R. **Complete Kmeans.R**. Make sure you get the test results.
   3. Submit the completed KMeans.R to Catalyst by Saturday 11:00 PM.
2. Preparation for the next weeks:
   1. Take a look at the part of DataScience02b.R that is titled: A glimpse into what we will do in future lessons (Do not submit anything for this item -- just play with it)
   2. Reading Assignment
      1. Read: AFewUsefulThingsToKnowAboutMachineLearning.pdf http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
      2. http://en.wikipedia.org/wiki/Supervised_learning
      3. http://en.wikipedia.org/wiki/Unsupervised_learning
      4. Links found in the feedback to the questions in Quiz 04b

Introduction to Data Science