



DATASCI 510

Data Science: Process and Tools

Lesson 02



© 2023 Ernst Henle




Lesson 2 Agenda

- Announcements
- Data Flow Diagrams: Rules of a Data Flow Diagram
- Data Flow Diagrams: Digital Pathology Example
- Lesson 02 In-class Quiz#1
- Lesson 02 Assignment
- Break
- Machine Learning
- Data Flow in Supervised Learning
- Break
- Supervised Learning Schema
- Lesson 02 In-class Quiz#2
- Time permitting: Data as Sparse Multi-Dimensional Matrices


Announcements

- No class on Thursday May 11th 2023
- Office hours from 9:00 to 11:00 AM on Sundays
- No office hours on Sunday May 14th 2023
- Guest lecture on using IoT data in legal cases



Data Flow Diagrams

A How-to for Assignment 2

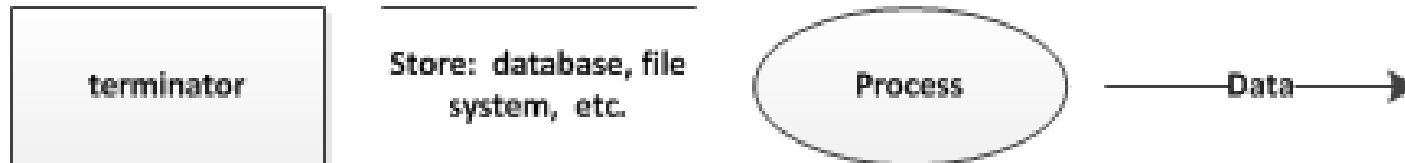


Data Flow

- Required for Data Processing
- SSADM specifies [Data Flow Diagrams \(DFD\)](#)

Four components of a DFD:

- Terminator
- Store
- Process
- Data Flow

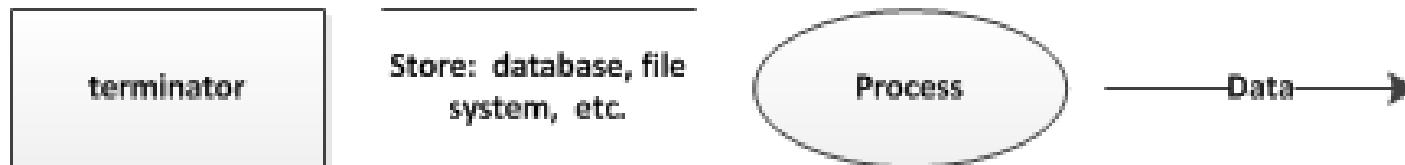


What is a Data Flow Diagram?

A defined language in the structured systems analysis and design method (SSADM).

–for describing processes that involve movement and transformation of data.

Dataflow diagrams (DFD,) define processes and generally do not reference components. DFDs processes are easily related to data science tasks.



DFD Symbols

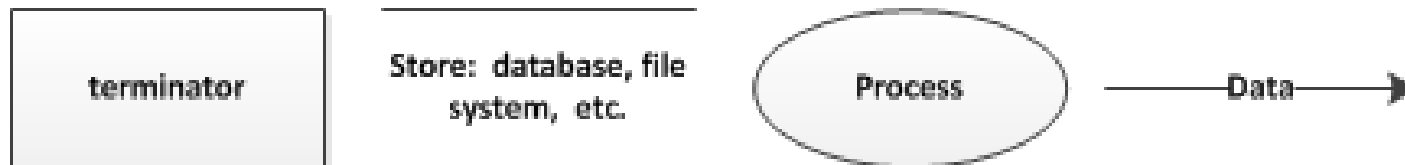
Complete Rectangles = start or terminate process

—either generate or consume data.

Rectangles without sides = stores, like databases.

Ellipse = a process that transforms data.

Arrow = data (including dataflow)



Data Flow Diagram Example

**Global human activity as measured by cell phone
camera use**

DFD Example: Image Aggregation Story

Describe, in a few sentences, a data science task that interests you.

1. Data are extracted and processed from images on cell phones
2. The processed data are combined
3. The combined data are used to derive meaning, like: Map human activity to times and locations.

Construct a data flow diagram that depicts the data processing that is required to complete the task

DFD: Image Aggregation Steps

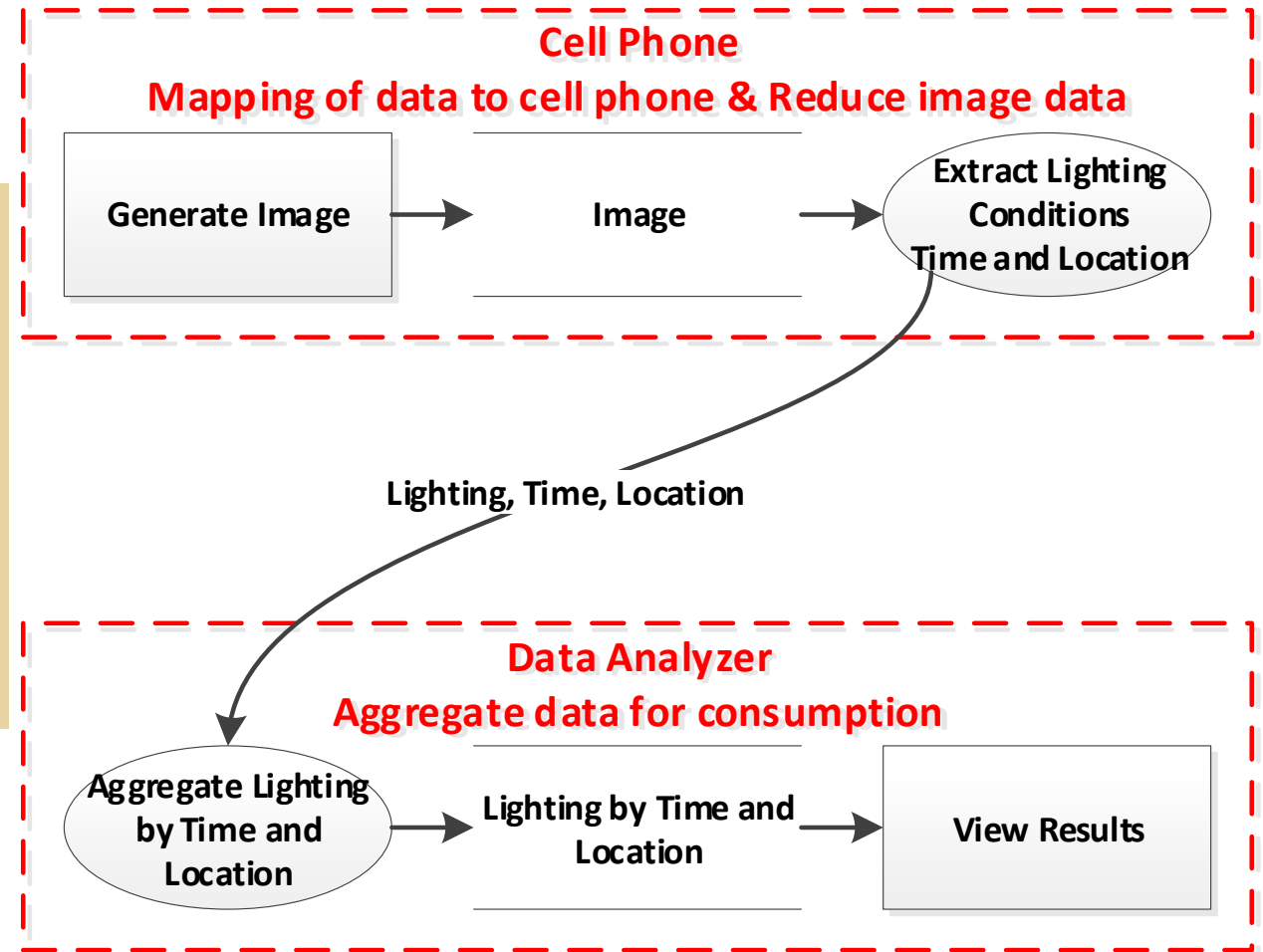
Collect and aggregate cell phone camera images

1. The image is taken (Image is mapped to cell phone)
2. Image is associated with cell location and time
3. The image characteristics are extracted (Data Reduction)
4. Image characteristics are sent with time and location
5. Image characteristics are collected and aggregated by location and time
6. The data are viewed and interpreted as human activity

DFD: Image Aggregation Steps

Collect and aggregate cell phone camera images

1. The image is taken (Image is mapped to cell phone)
2. Image is associated with cell location and time
3. The image characteristics are extracted (Data Reduction)
4. Image characteristics are sent with time and location
5. Image characteristics are collected and aggregated by location and time
6. The data are viewed and interpreted as human activity





A closer look at DFD

Understanding the components

Data Flow: DFD Arrow

An arrow represents data or data flow. The arrow is labeled by the name of the data. Example:



An arrow is necessary to connect the other data flow components. Every data flow component must have at least one arrow.

Data Flow Practice: DFD Arrow

Which example is correct?

—————Eat—————▶

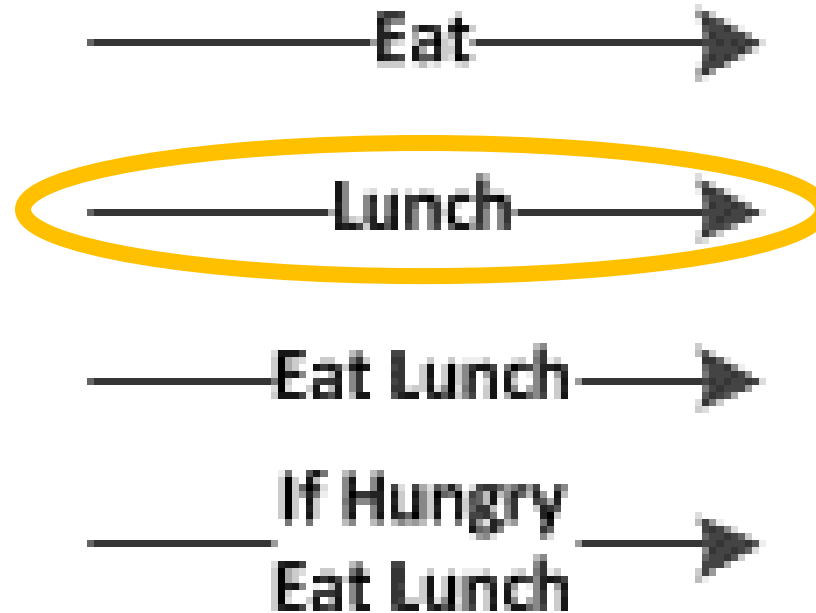
—————Lunch—————▶

—————Eat Lunch—————▶

—————If Hungry
Eat Lunch—————▶

Data Flow Practice: DFD Arrow

Which example is correct?



Data Flow: DFD Process

Represented by an ellipse

Takes in data from one or more data sources, transforms the data, and then outputs the data.

- A process must have at least one input arrow
- A process must have at least one output arrow.

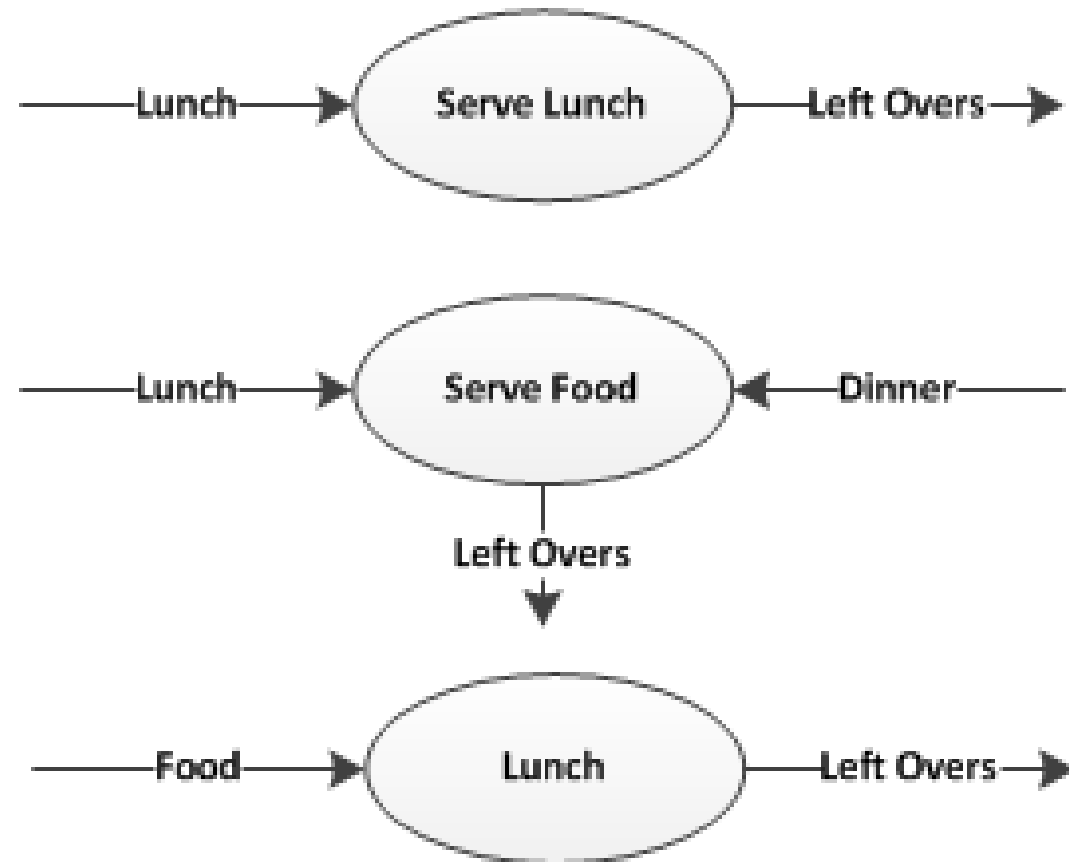
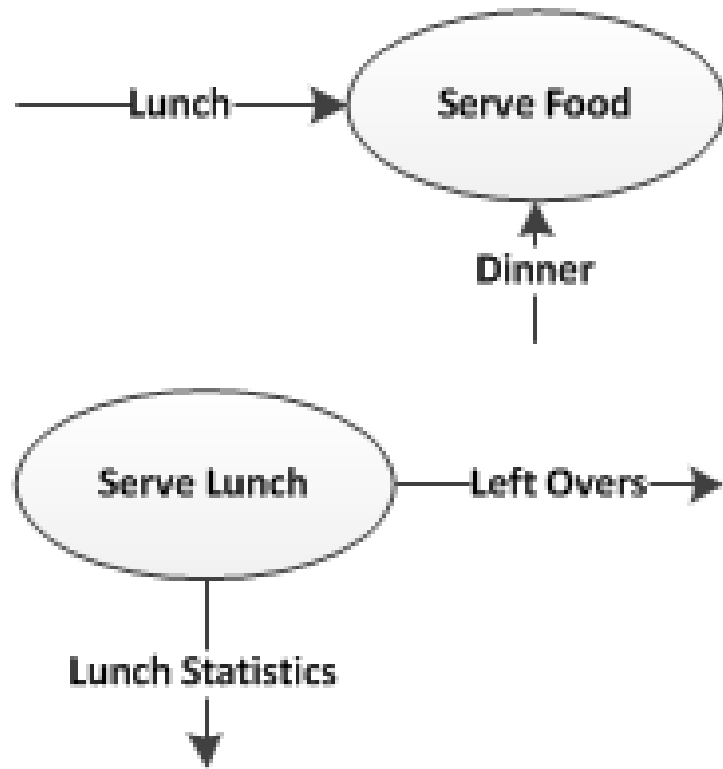
A process is labeled with a verb, like “Brighten”

Example:



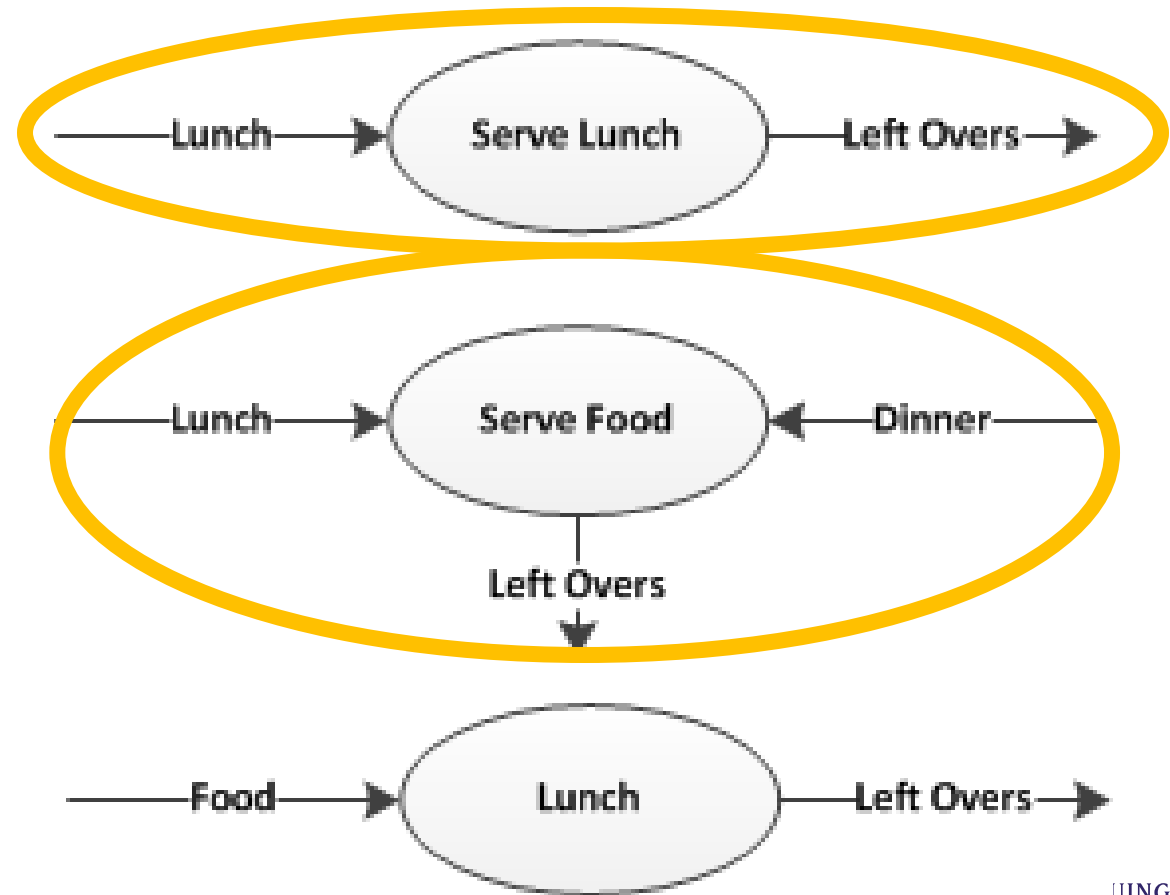
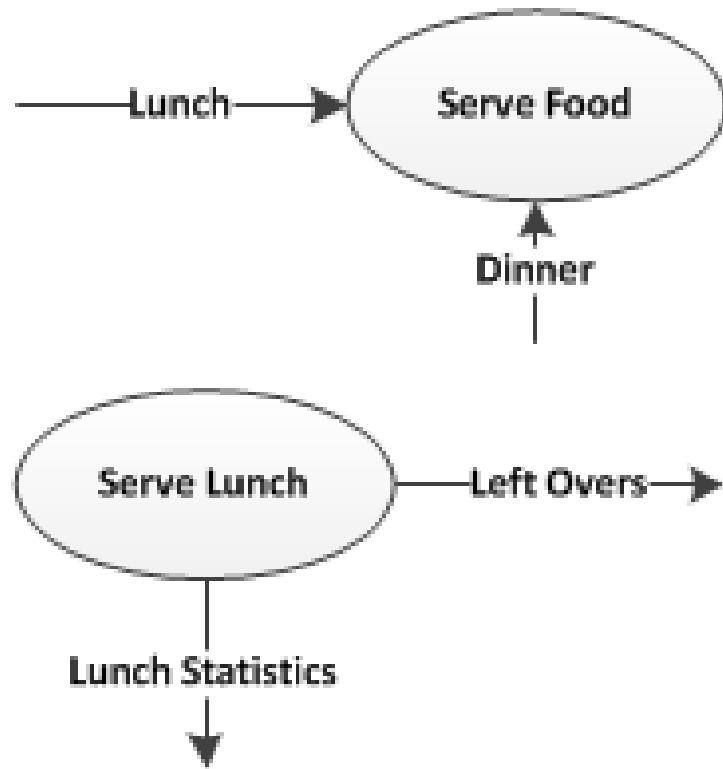
Data Flow Practice: DFD Process

Which of these are correct?



Data Flow Practice: DFD Process

Which of these are correct?



Data Flow: DFD Terminator

Represented by a rectangle with all four sides drawn.

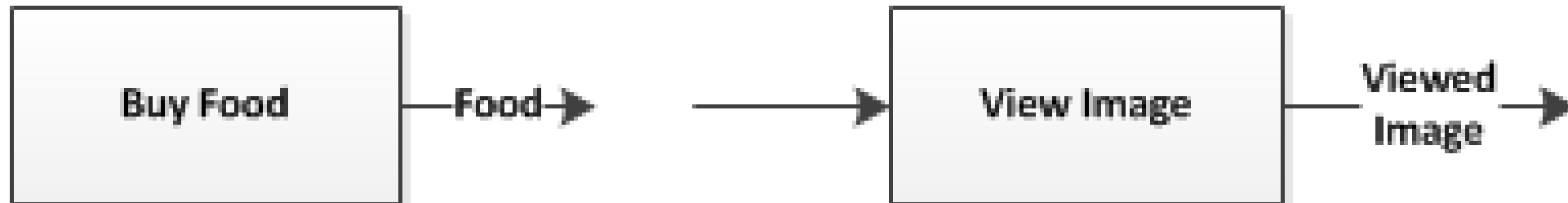
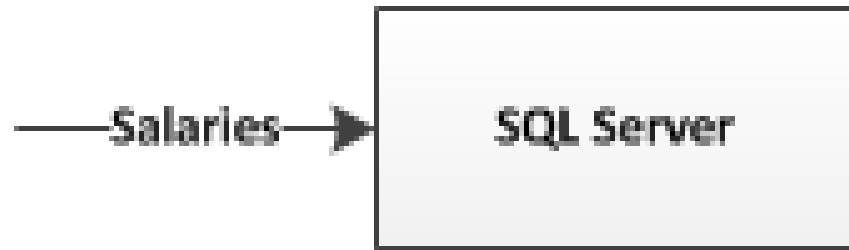
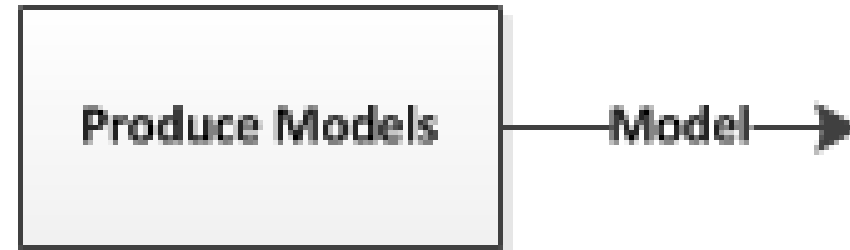
A process that either generates or consumes data. This process may reference a component like: Get data from Internet or View data in Monitor

Example:



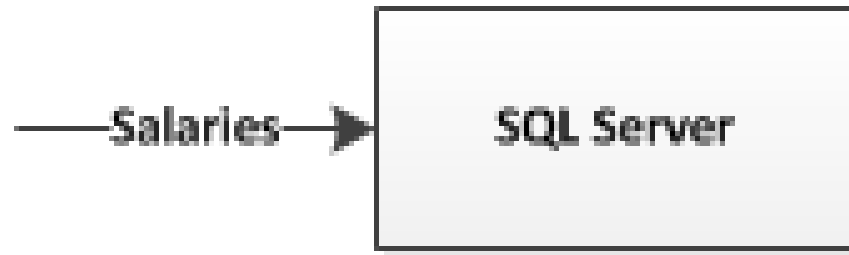
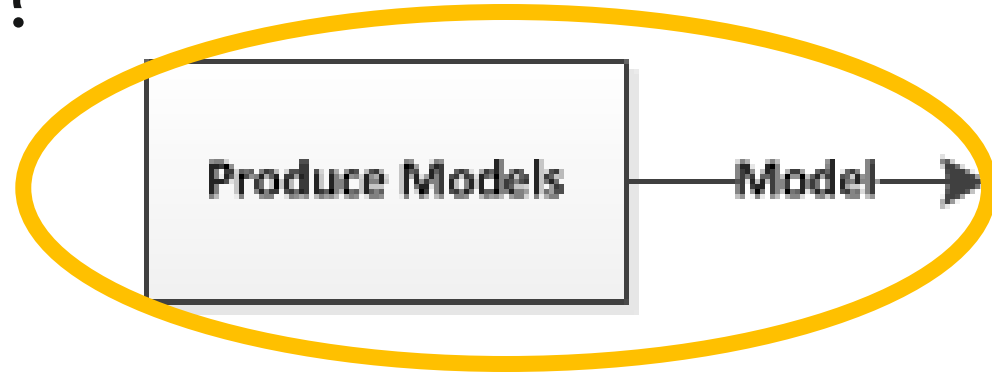
Data Flow Practice: DFD Terminator

Which of these are correct?



Data Flow Practice: DFD Terminator

Which of these are correct?



Data Flow: DFD Store

Represented by a rectangle that is missing the right-hand side or both the right- and left-hand sides.

A place where the data is persisted. Typical stores are text files, websites, and relational data bases.

- A store has at least one input arrow
- A store has at least one output arrow
- Typically, the input and output arrows are not labeled.

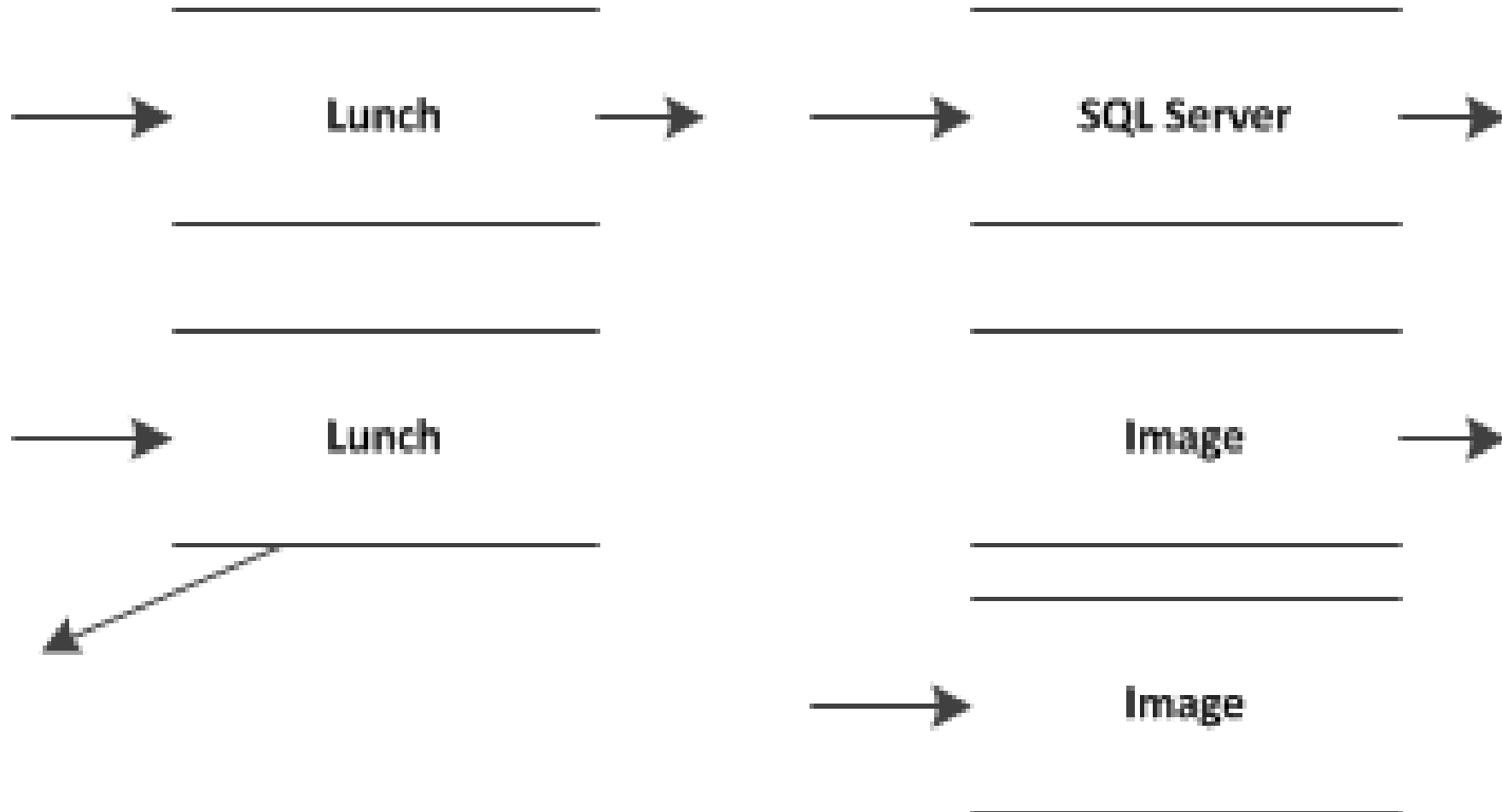
The name of the store describes the nature of the data (not the nature of the data base)

Example:



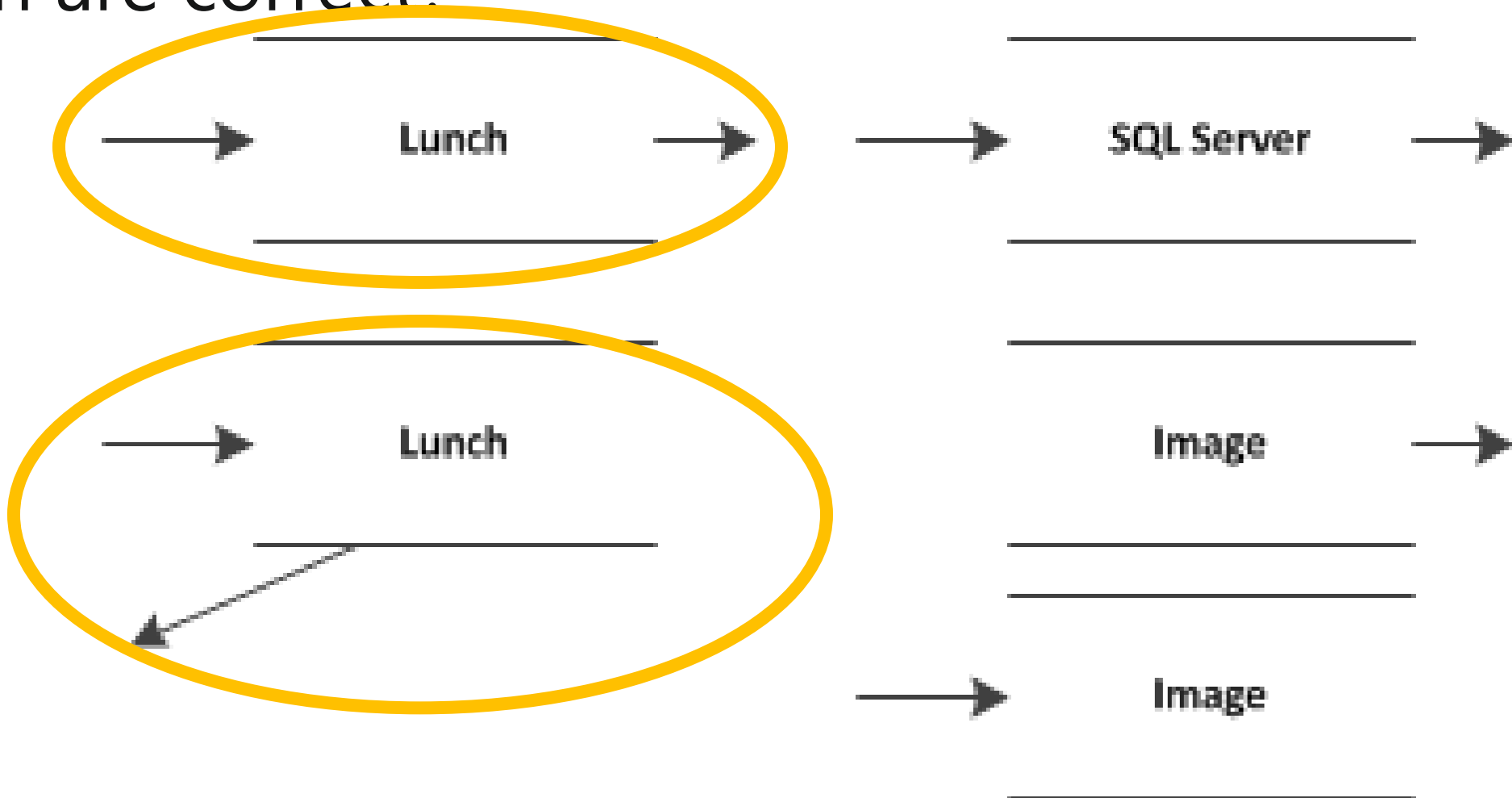
Data Flow Practice: DFD Store

Which are correct?



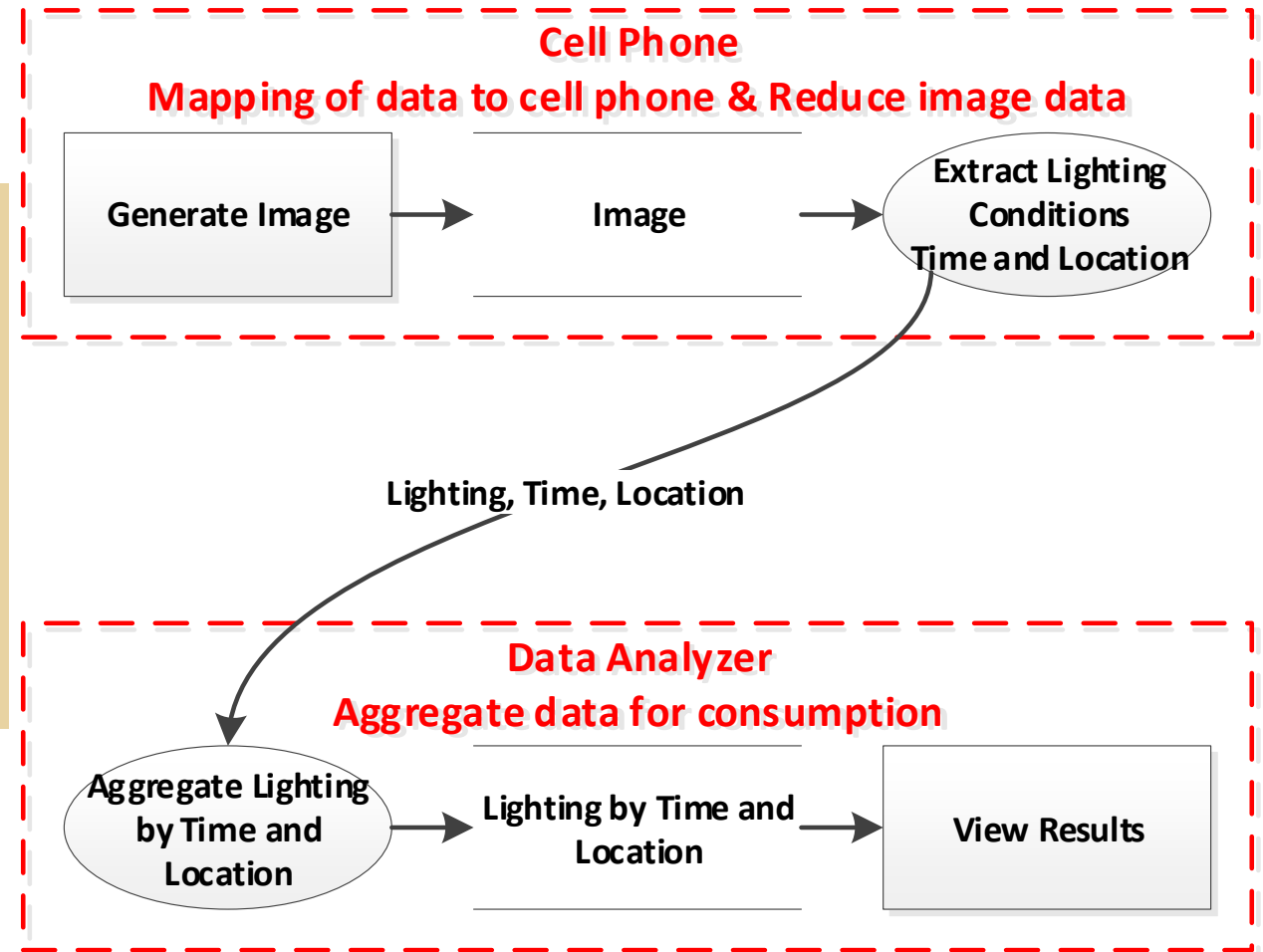
Data Flow Practice: DFD Store

Which are correct?



DFD: Image Aggregation Steps

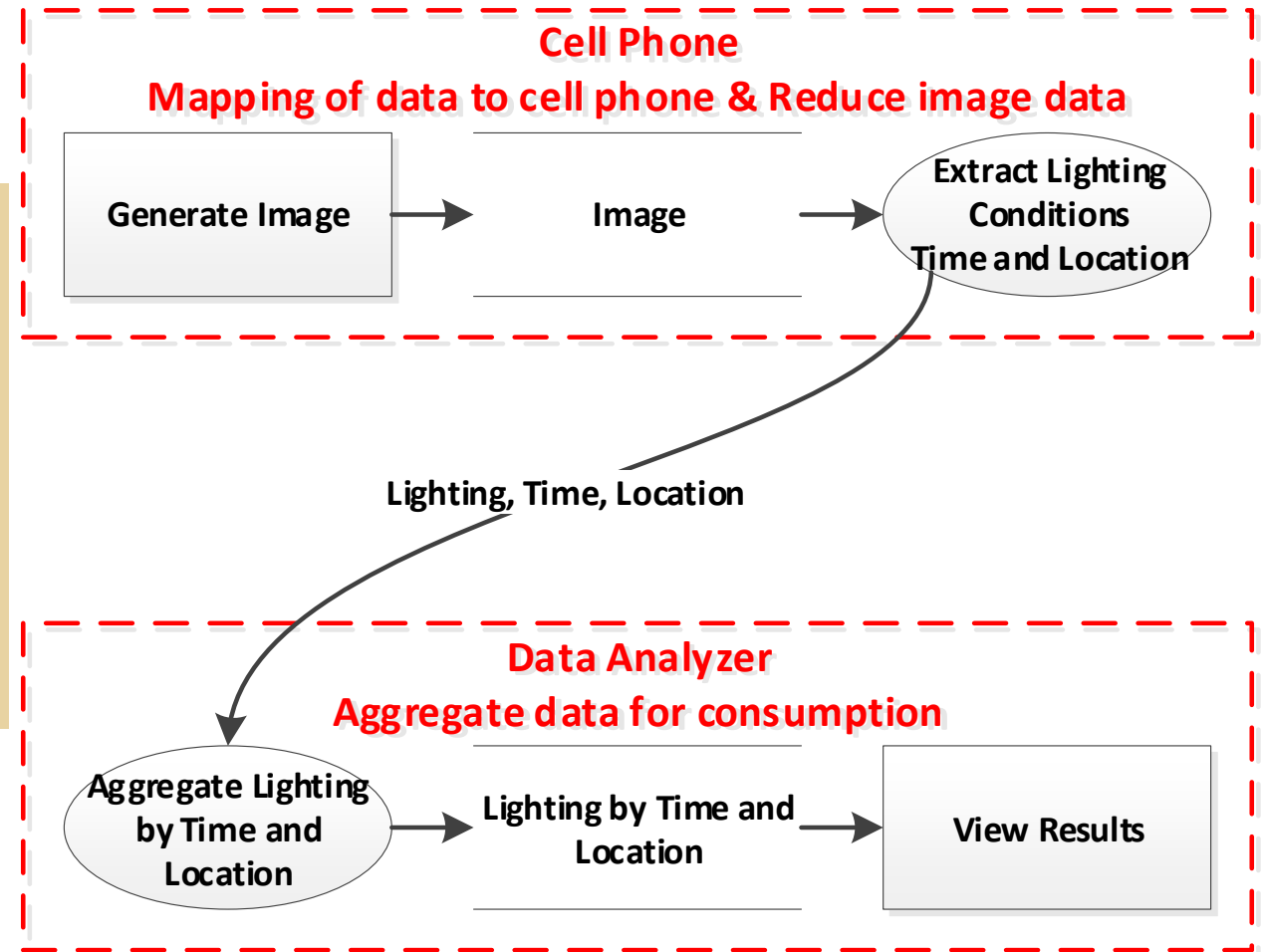
Revisit the image aggregation example



DFD: Image Aggregation Steps

Some Rules

- Terminator labels need to contain a verb
- Stores are labeled with nature of data
- Stores do not process data
- Arrows should be labeled with nature of data
- Arrows going in or out of stores need not be labeled
- Processes need to be labeled with a verb
- No component names (system, database, camera, analyzer, etc.) for DFD component labels

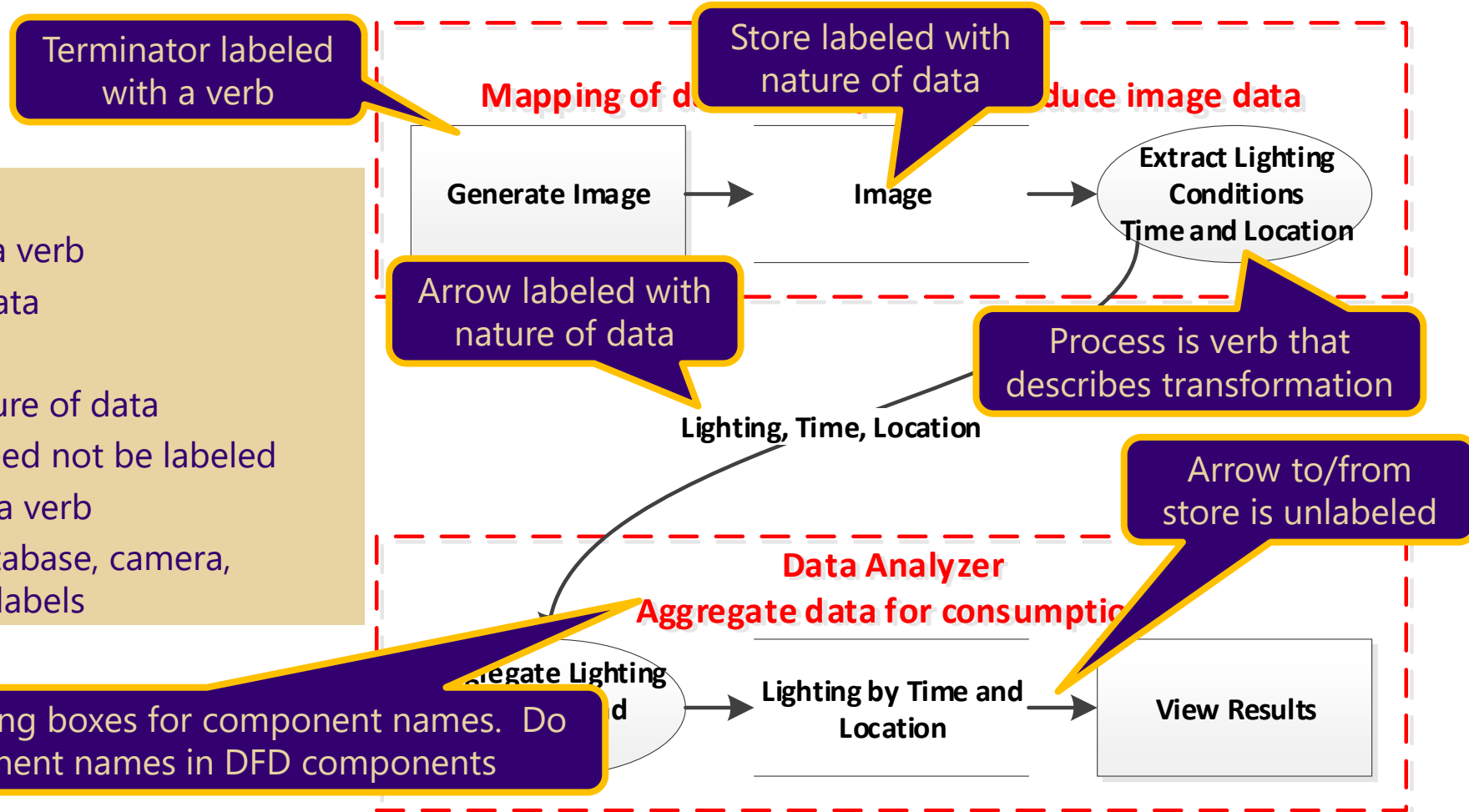


DFD: Image Aggregation Steps

Some Rules

- Terminator labels need to contain a verb
- Stores are labeled with nature of data
- Stores do not process data
- Arrows should be labeled with nature of data
- Arrows going in or out of stores need not be labeled
- Processes need to be labeled with a verb
- No component names (system, database, camera, analyzer, etc.) for DFD component labels

Use security bounding boxes for component names. Do not use component names in DFD components



DFD: Digital Pathology

An Example

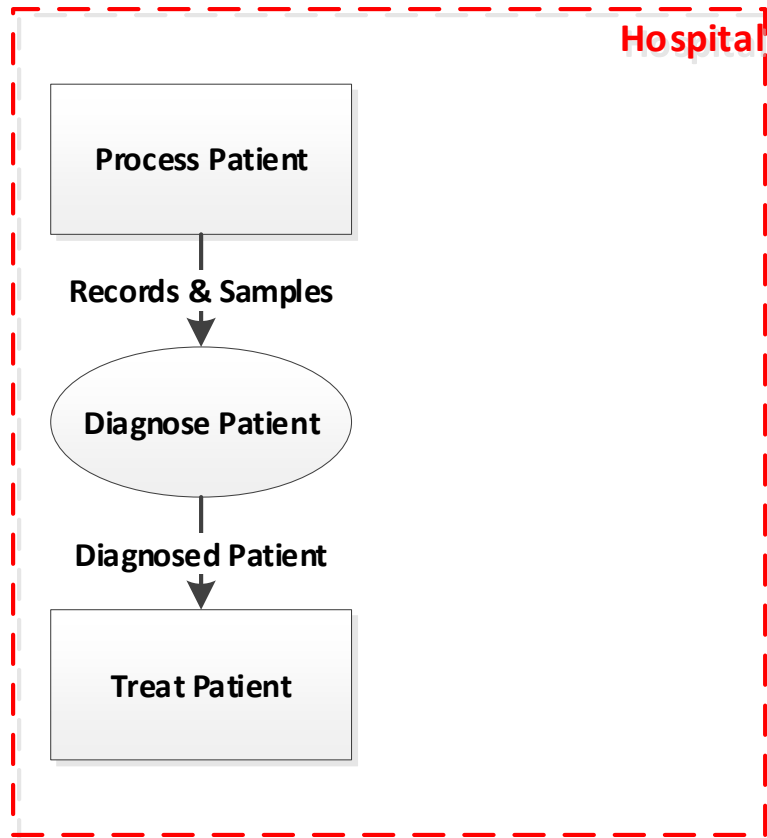
DFD Example: Digital Pathology

Many blood disorders manifest themselves through easily recognizable morphological changes, but the affected cells may be as few as one in a hundred thousand.

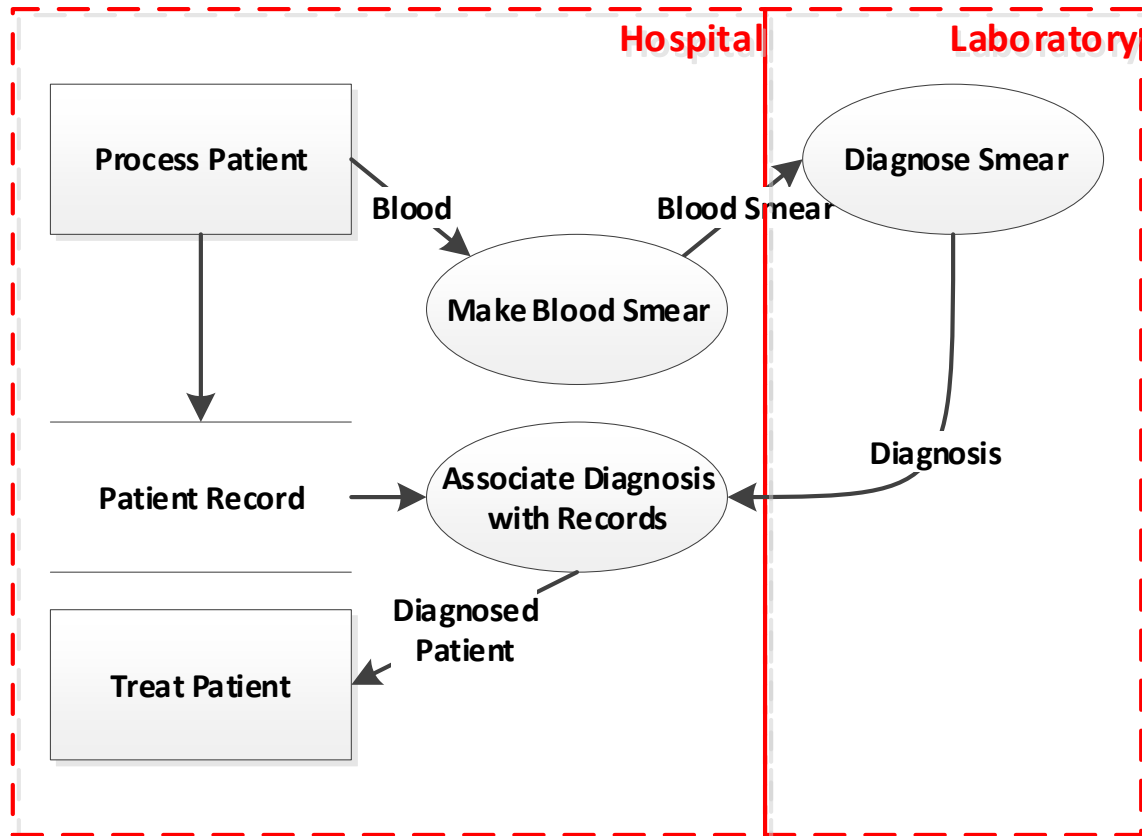
Given the scarcity and cost of pathologists, it is not possible to routinely screen for these blood disorders. We would like to find an automated way of diagnosing such disorders.

We use a pathologist to score aberrant cells and correlate these findings with shape characteristics determined by image segmentation.

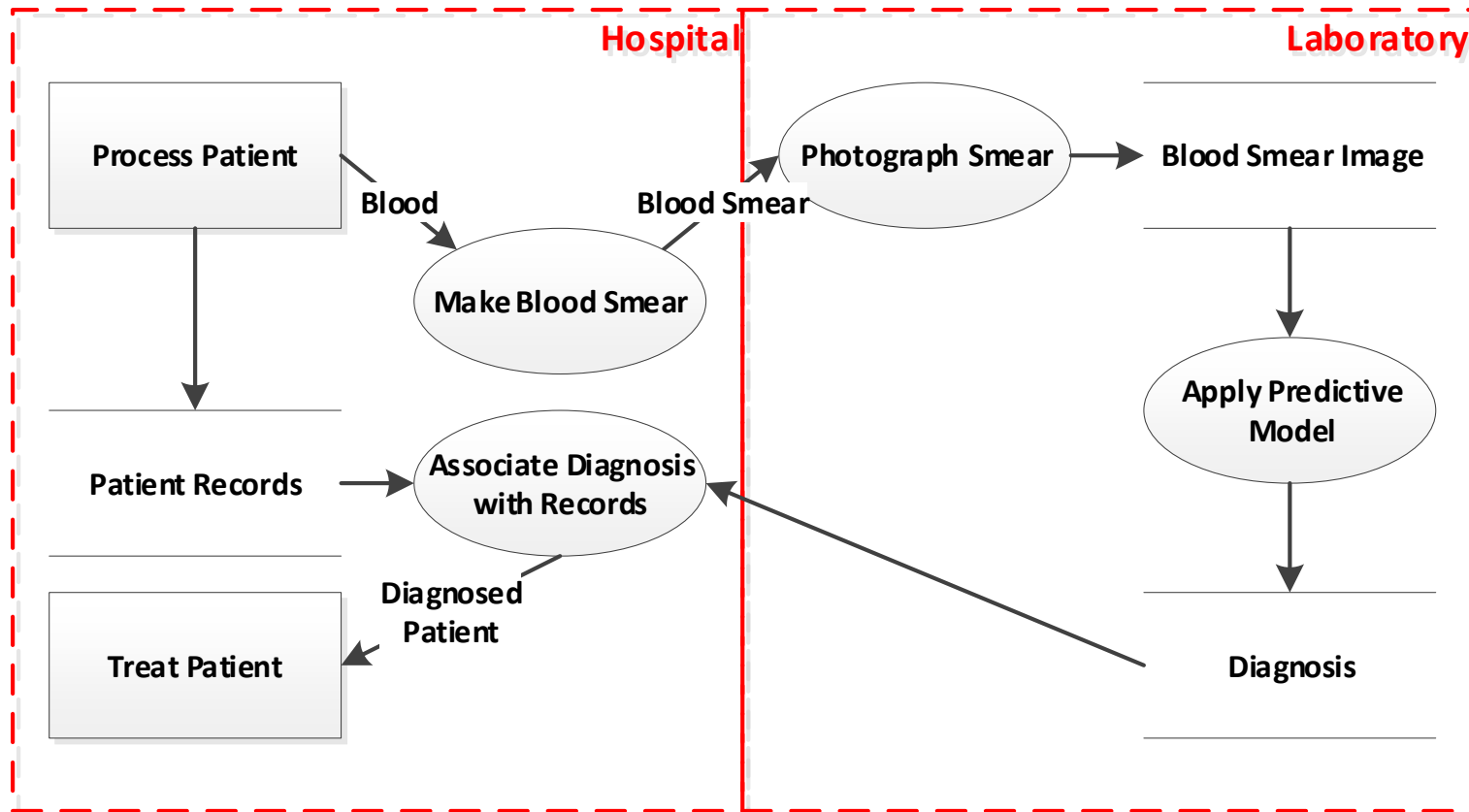
DFD Example: Digital Pathology



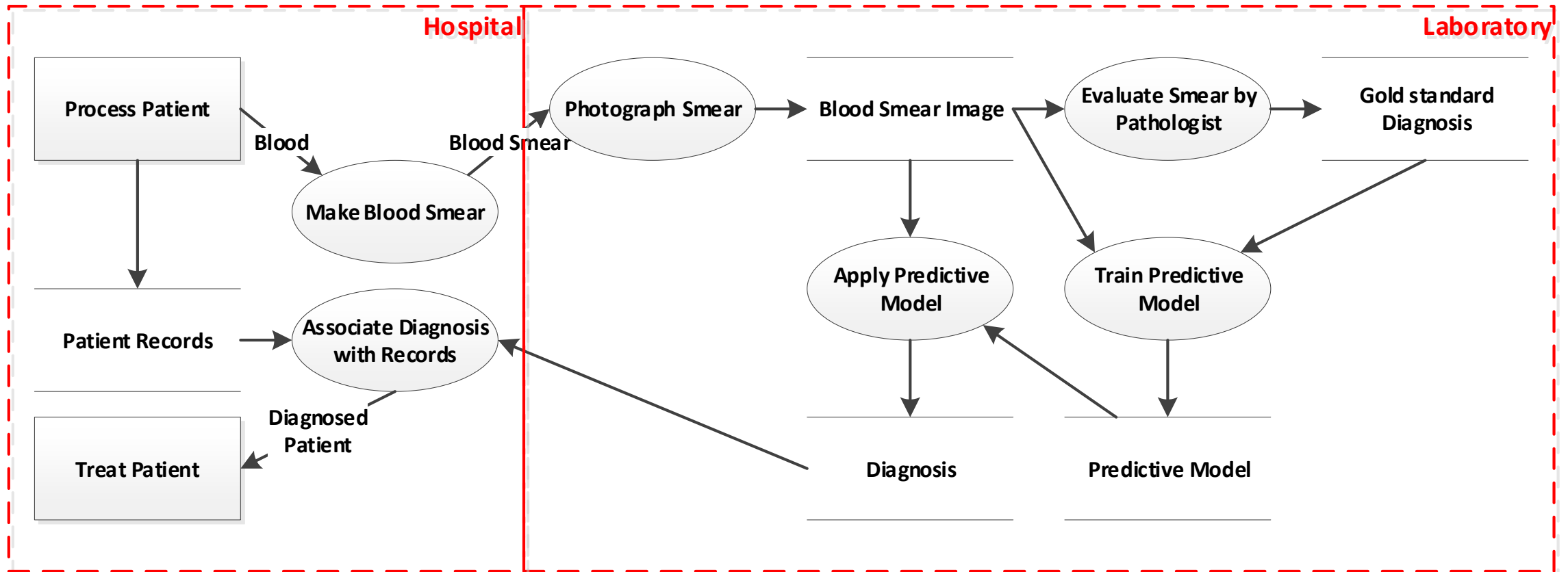
DFD Example: Digital Pathology



DFD Example: Digital Pathology



DFD Example: Digital Pathology



Lesson 02 In-class Quiz#1

Lesson 02 Assignment

Write a short description of a data science use case. Create a data flow diagram (DFD) for the use case and save it as a pdf or word doc. The data flow diagram should make use of all 4 DFD components and a security boundary. The DFD will be primarily graded on how well you followed the rules of writing a DFD.

You can select a use case from here:

- Kaiser Permanente: [Tracking pandemics](#)
- Northwestern: [Sports analytics](#)
- Mount Sinai: [Risk models around population health](#)
- Thomson Reuters: [Quantitative finance](#)
- Wells Fargo: [Chatbots](#)
- Google: [User Experience](#)
- Spotify: [User Experience](#)
- Uber: [Forecasting](#)
- Zocdoc: [Healthcare marketplace](#)
- Pythian: [Predictive maintenance](#)
- Astro Digital: [Food economy](#)
- Airbnb: [Personalization](#)

Lesson 02 Assignment

Report on use case

Write a short description (at least 4 sentences) about a use case containing the following information:

1. Describe the business problems that the use case tries to solve.
2. Describe the data science methods employed, like (un)supervised learning.
3. Describe the steps in the data science cycle from lecture 1 that were involved.
4. Describe challenges or questions faced during the steps of the data science process.

Lesson 02 Assignment

- DFD Figure Caption: Add explanations (1 to 3 sentences) for every DFD component in your DFD. These explanations must be in the DFD figure caption/legend and not the DFD itself.=
- DFD arrows: Each arrow connects one single DFD component to another single DFD component. The arrows are labeled with a noun that describes what the data are. The labels avoid the word "data".
- DFD stores: The stores are drawn as open rectangles. The stores have both in-bound and out-bound arrows. The stores are labeled with a noun that describes what the data. The labels avoid the word "data".
- DFD terminators: The terminators are drawn as complete rectangles. The starting terminator(s) have only out-bound arrow(s). The label(s) of the starting terminator(s) mention which data are found/created/selected/generated/etc. The ending terminator(s) have only in-bound arrow(s). The label(s) of the ending terminator(s) mention how the resulting data are presented/used/consumed/etc. The labels avoid the word "data".
- DFD process: Processes are drawn as ellipses. The processes have both in-bound and out-bound arrows. The processes are labeled with a verb. The labels avoid the word "data".
- DFD logic and readability: The DFD is easily understood by a fellow data scientist who is new to the use case.

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.

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 model 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

Data Flow in Supervised Learning



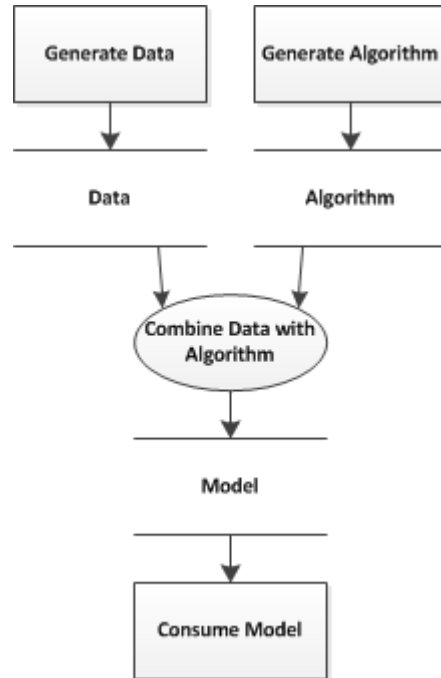
FROM DATA TO PREDICTIONS

> How do we get from data to predictions?

Data → ? → Predictions



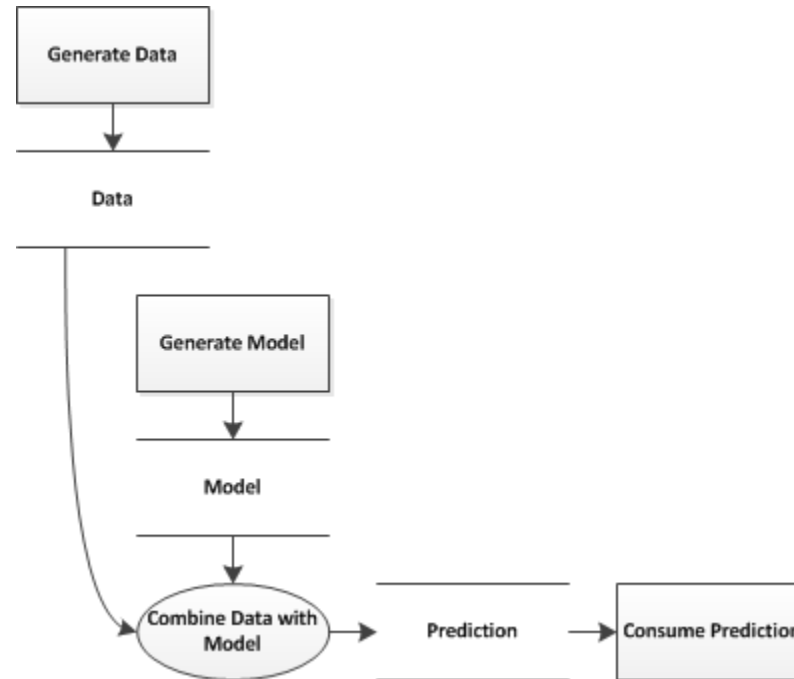
FROM DATA TO PREDICTIONS



Training Data + Algorithm → Model



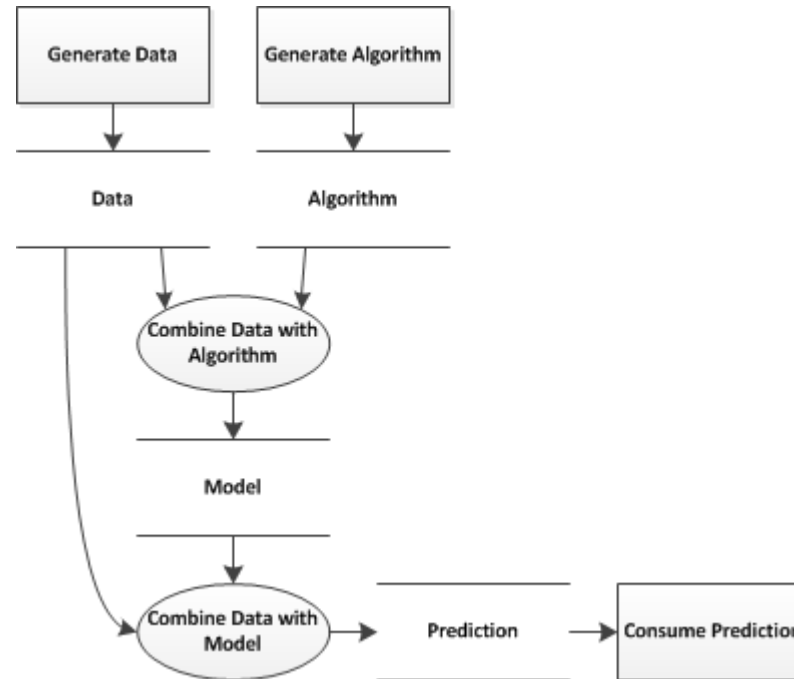
FROM DATA TO PREDICTIONS



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



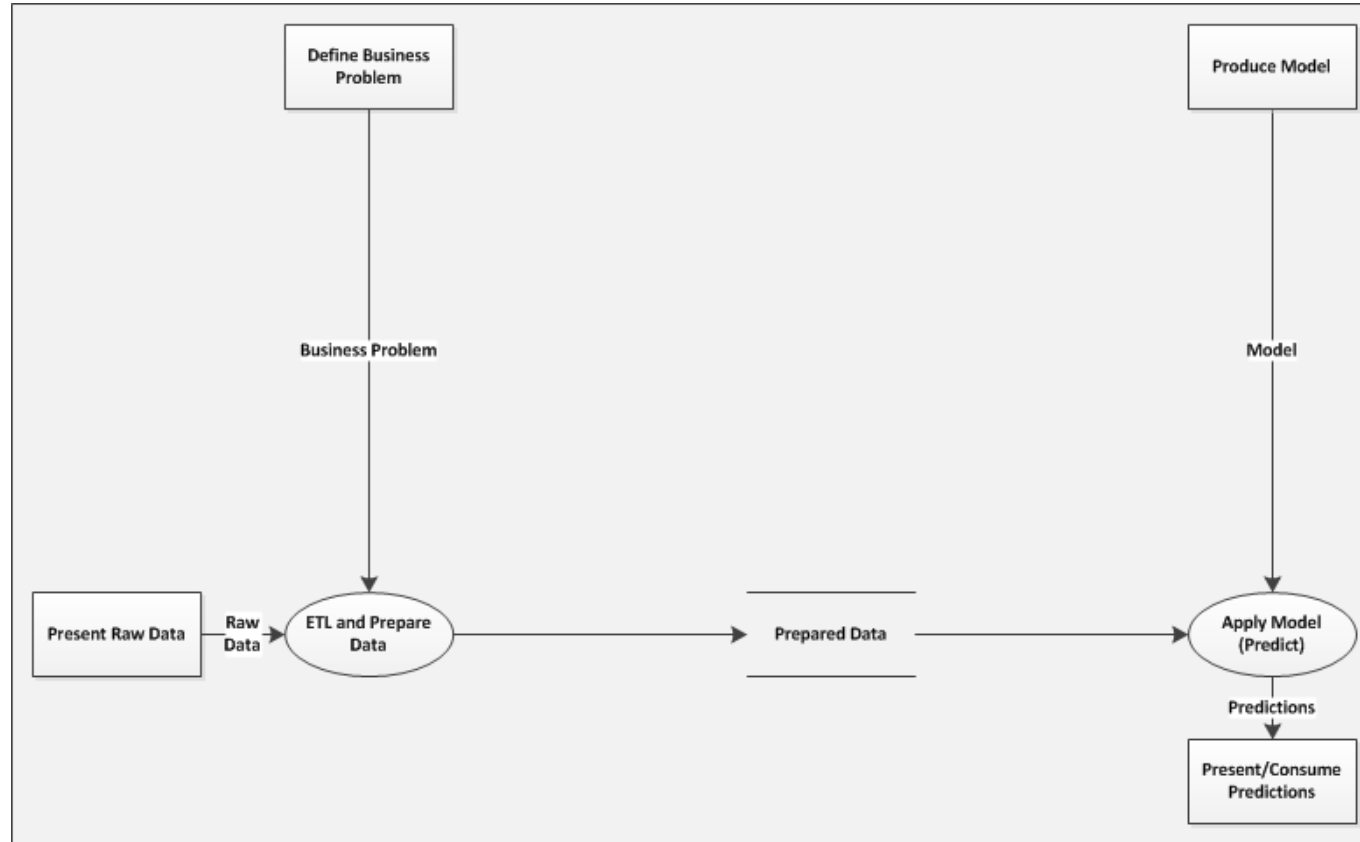
MODEL ACTS ON DATA



Model + Data → Prediction



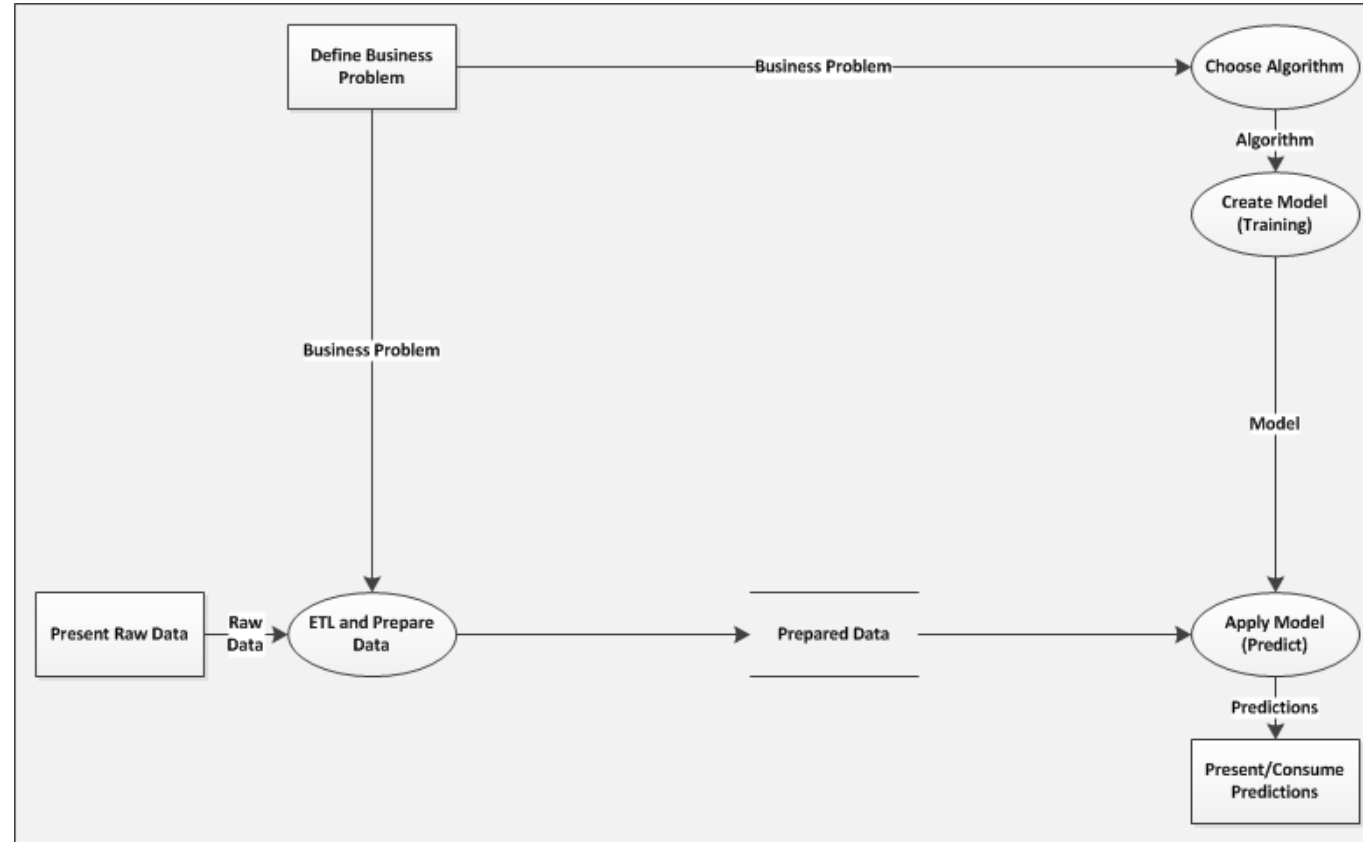
DATA ETL AND PREPARATION DRIVEN BY BUSINESS PROBLEM



Business Problem determines ETL and Data Prep



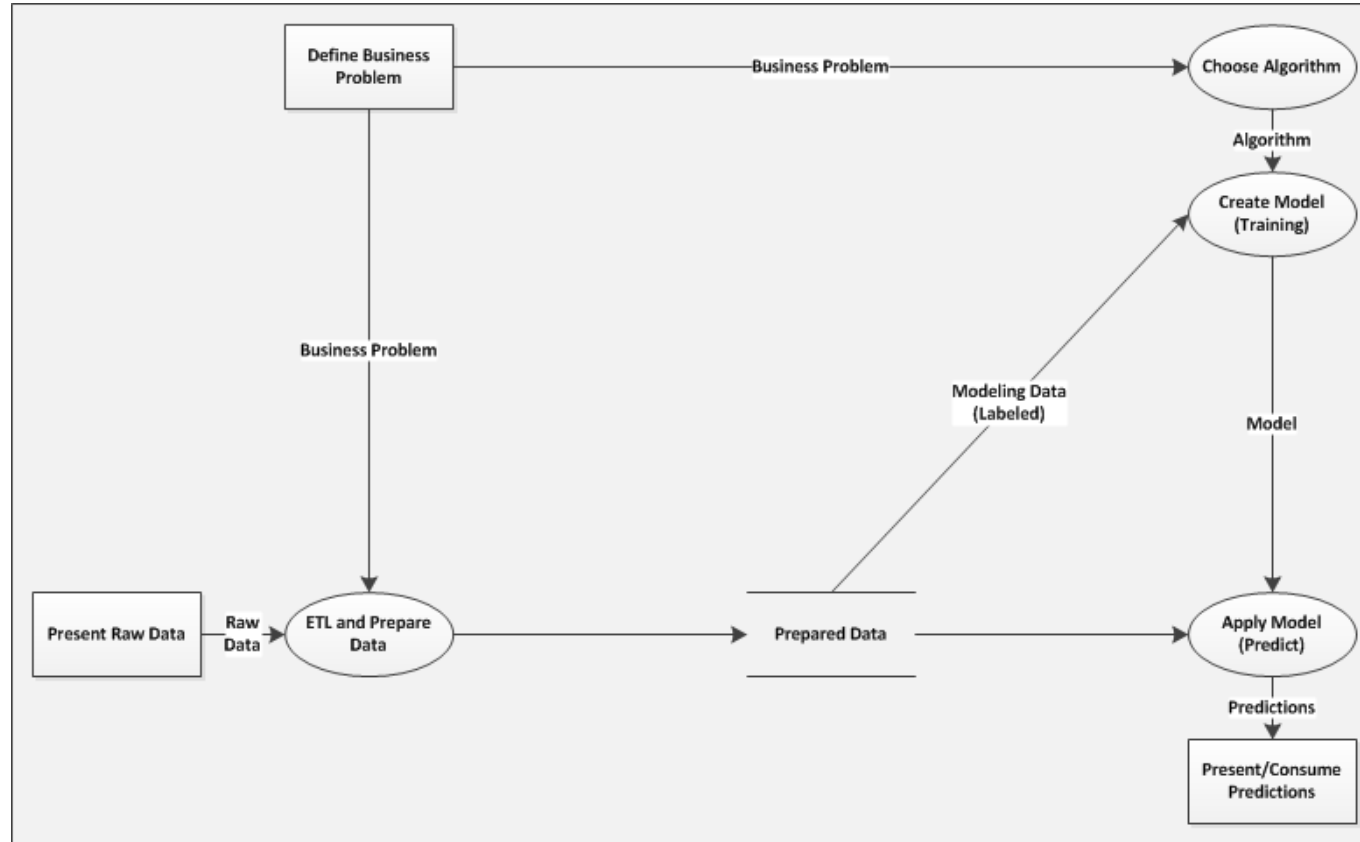
ALGORITHM CHOICE DRIVEN BY BUSINESS PROBLEM



Business Problem determines the choice of Algorithm.



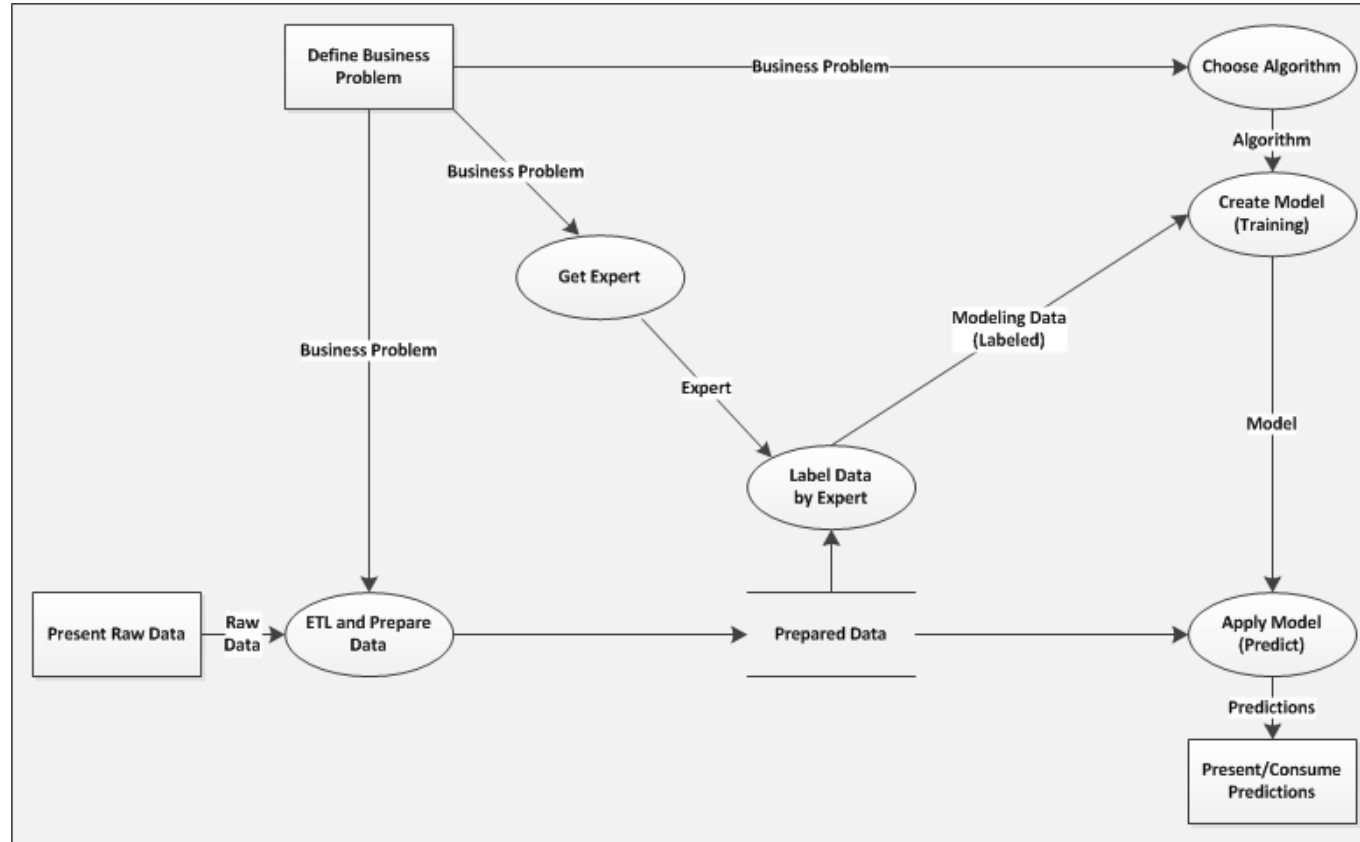
MODEL CREATION NEEDS DATA



Data + Algorithm → Model



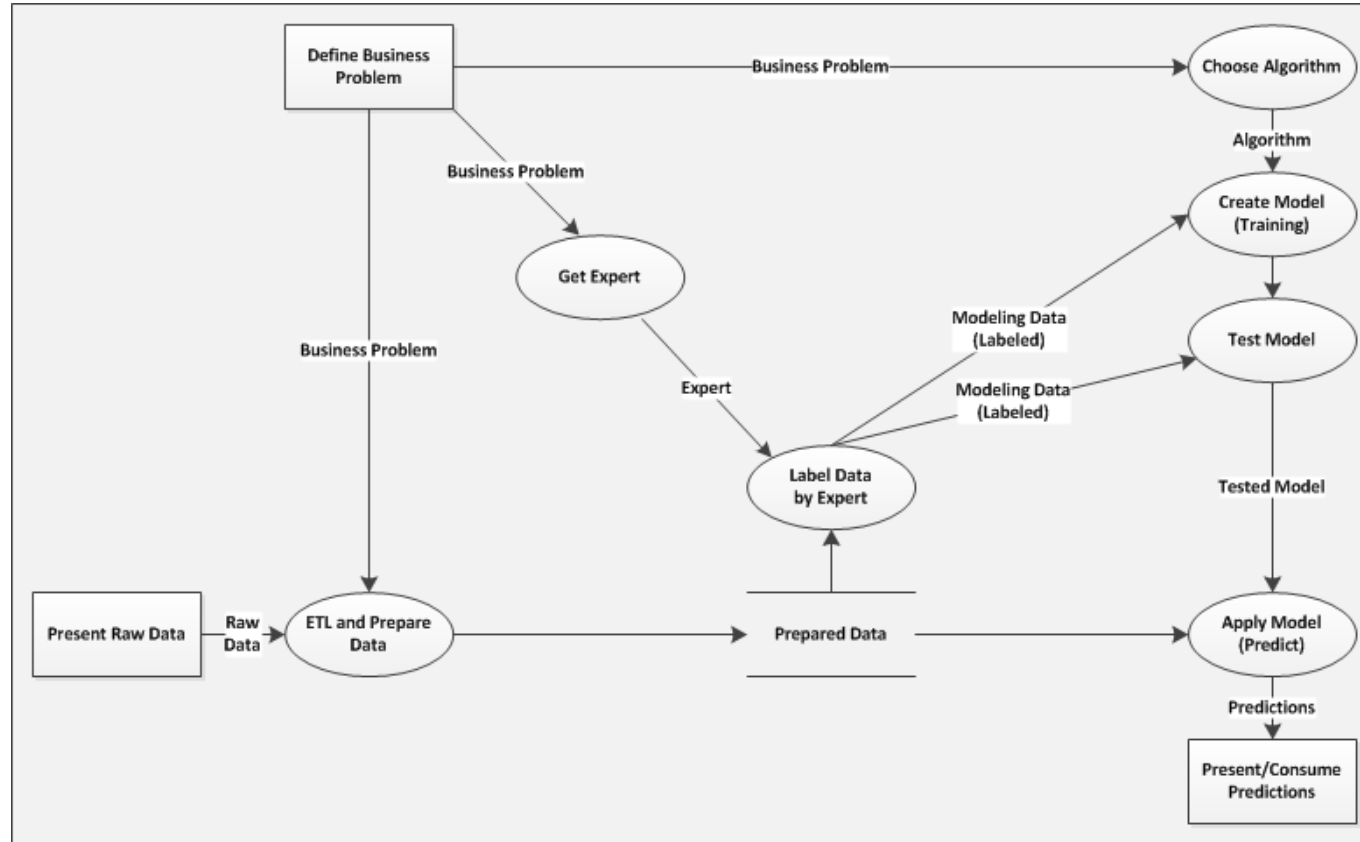
SUPERVISED TRAINING NEEDS DATA LABELED WITH OUTCOMES



Supervised Learning requires expert labeling of data.



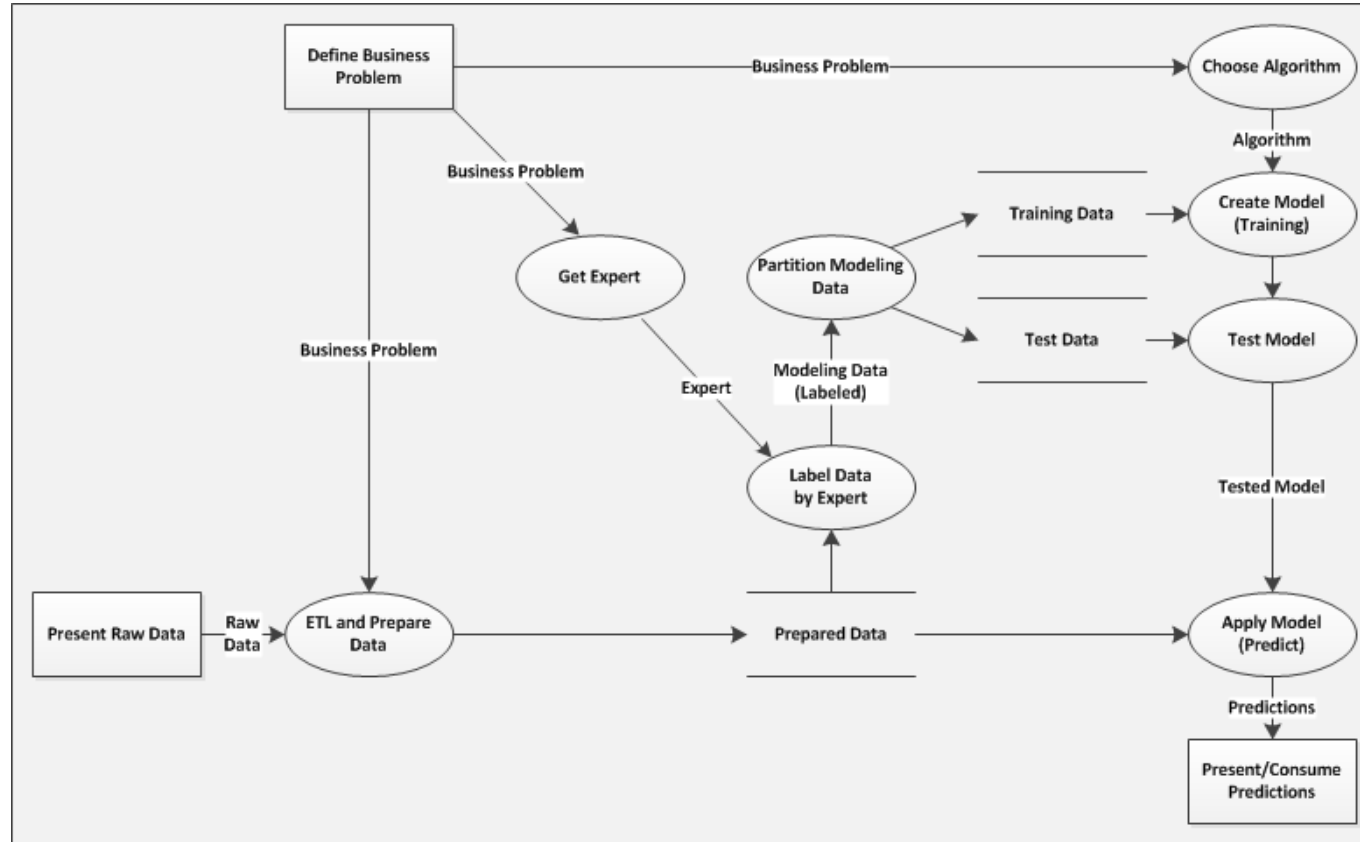
MODELS NEED TO BE TESTED



Do not trust predictions from an un-tested model!



TRAINING & TESTING OF MODEL USE DIFFERENT DATA



Do not test a model using training data!



Break



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

Random
or Keys

Constant

Proxy

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**

**Binary
Input**

**Categorical
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**

**Binary
Input**

**Categorical
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

Training Data
(200-50000
rows)

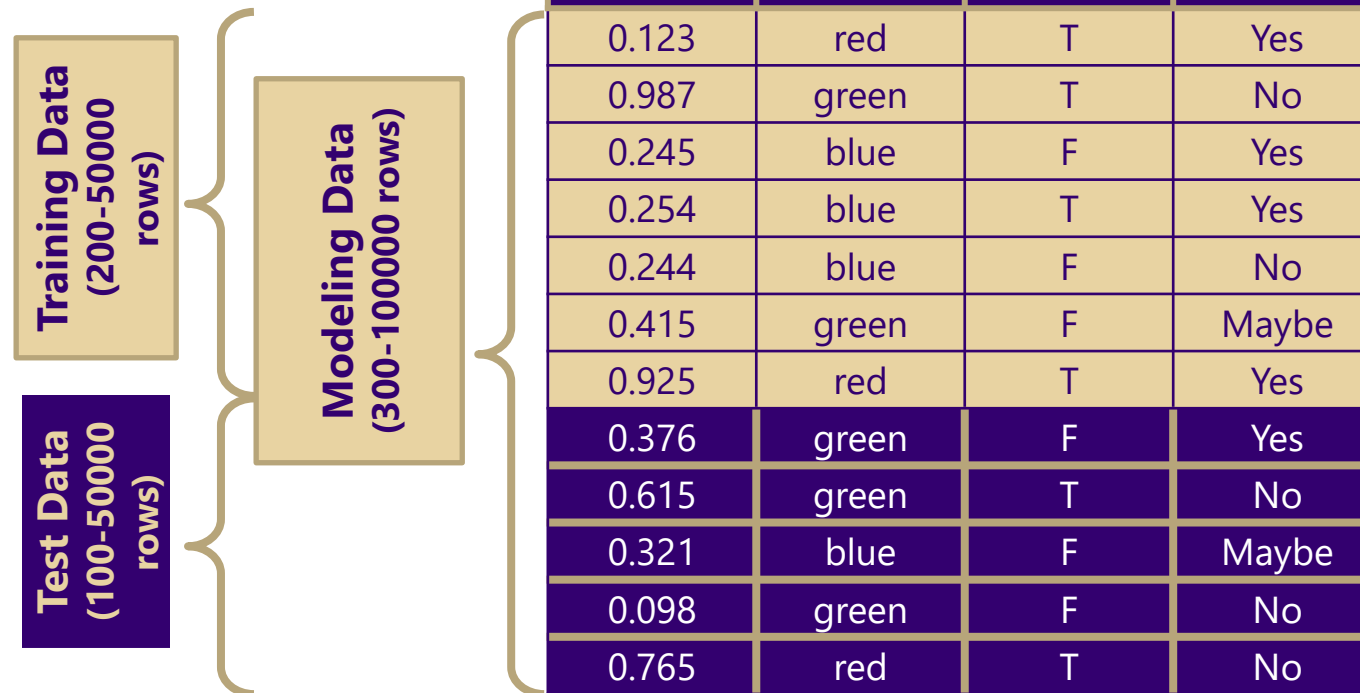
Modeling Data
(300-100000 rows)

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



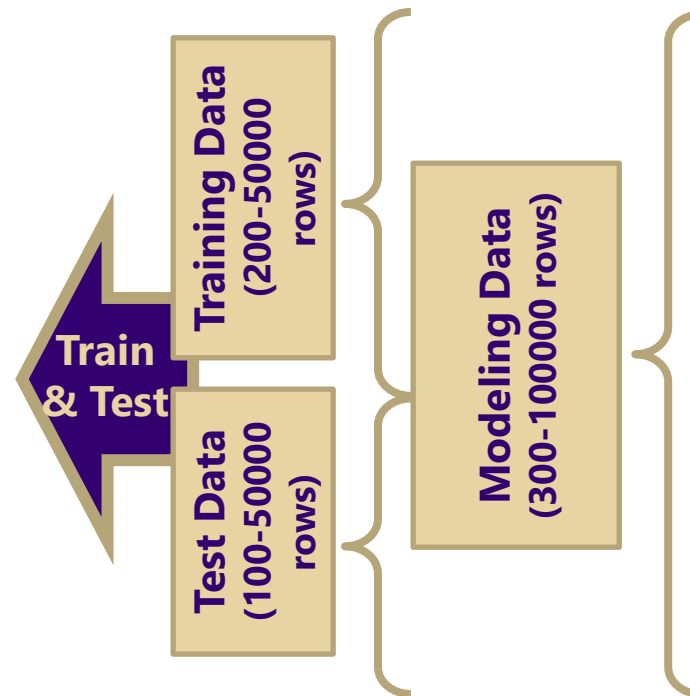
SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3

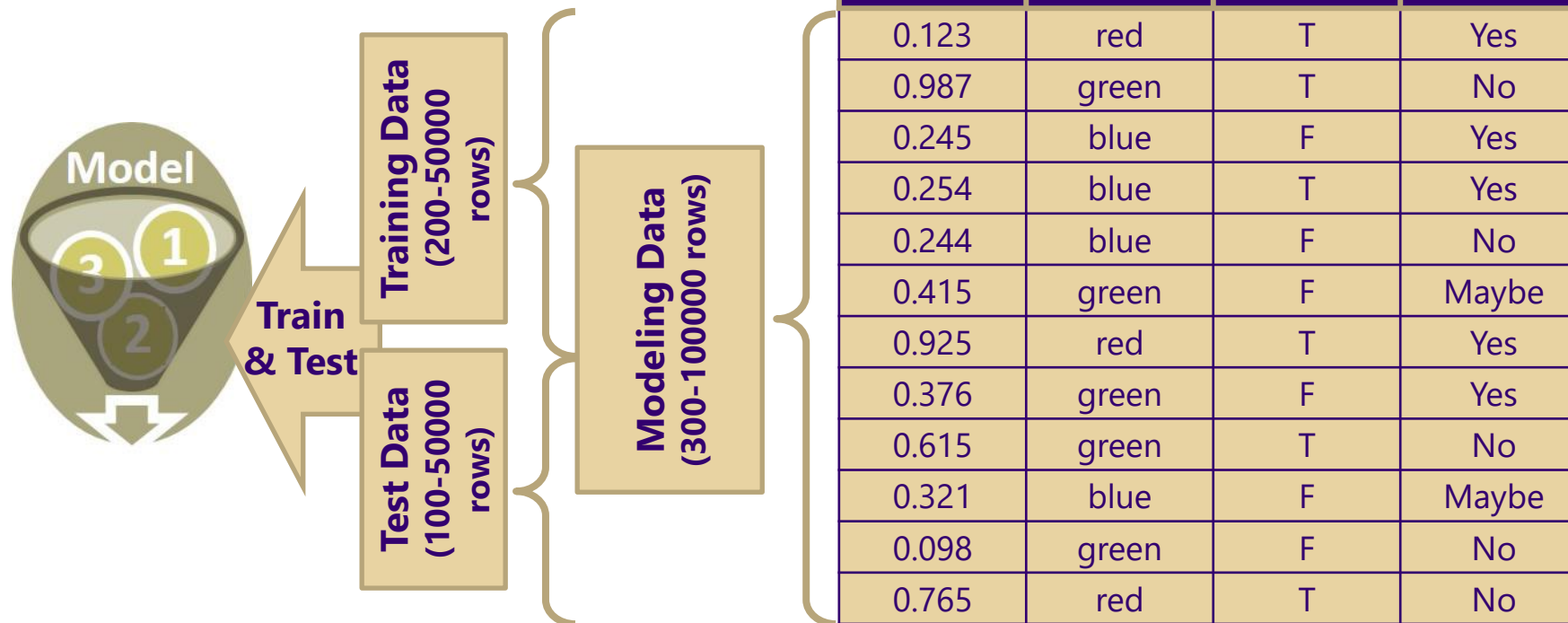


Input 1	Input 2	Input 3	Target
0.123	red	T	Yes
0.987	green	T	No
0.245	blue	F	Yes
0.254	blue	T	Yes
0.244	blue	F	No
0.415	green	F	Maybe
0.925	red	T	Yes
0.376	green	F	Yes
0.615	green	T	No
0.321	blue	F	Maybe
0.098	green	F	No
0.765	red	T	No



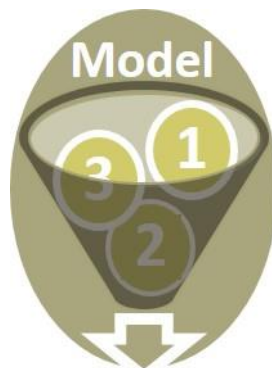
SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



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".

Operational Data
(1 - ∞ 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
0.234	green	T	Unknown Target Outcome
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

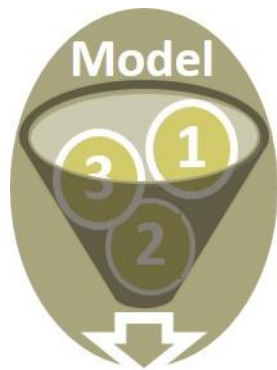


Operational Data (1 - ∞ rows)	Input 1	Input 2	Input 3	Target
	0.234	green	T	Unknown Target Outcome
	0.567	blue	F	
	0.890	green	T	
	0.314	red	T	



SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



Deploy Model to Predict Target Outcome

Operational
Data
(1 - ∞ rows)

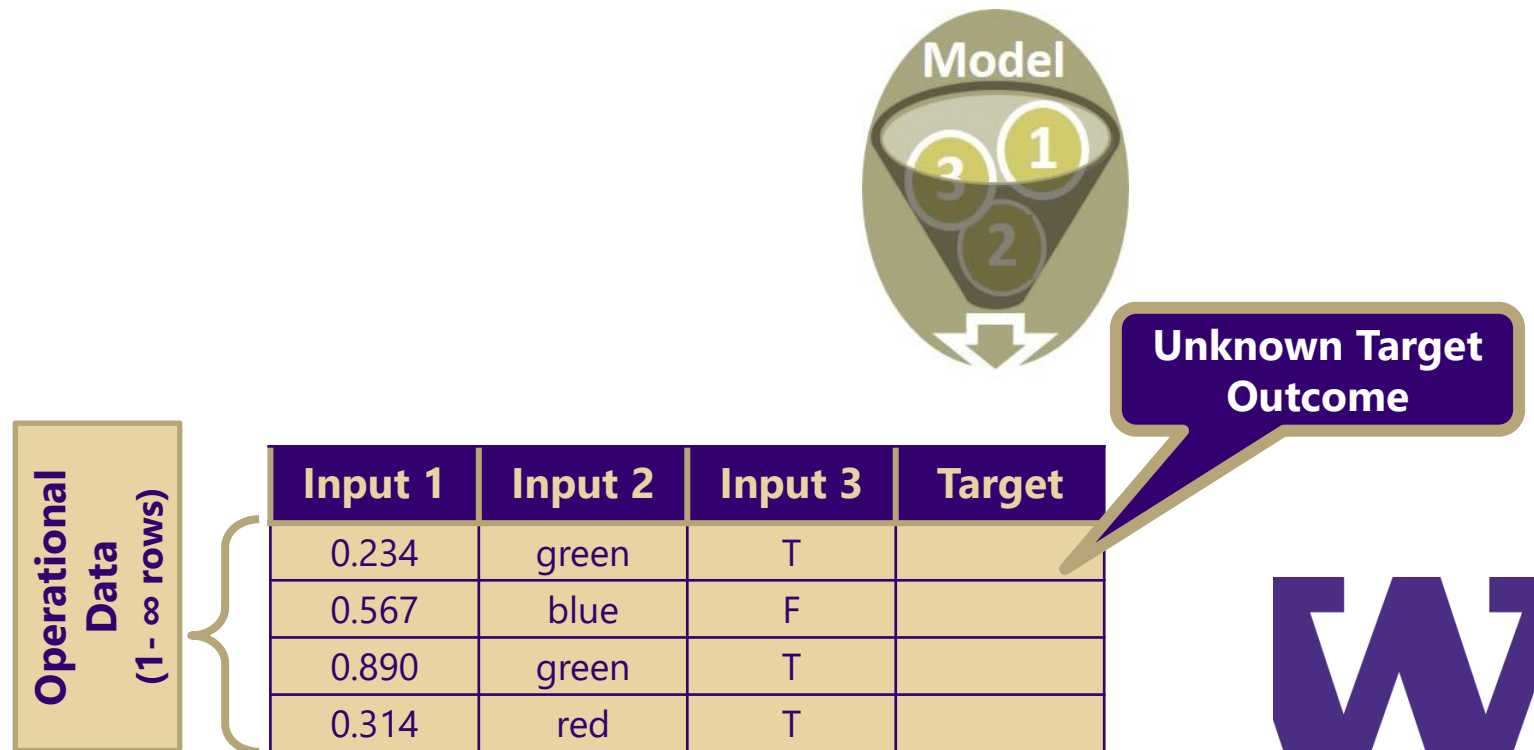
Input 1	Input 2	Input 3	Target
0.234	green	T	
0.567	blue	F	
0.890	green	T	
0.314	red	T	

Unknown Target Outcome



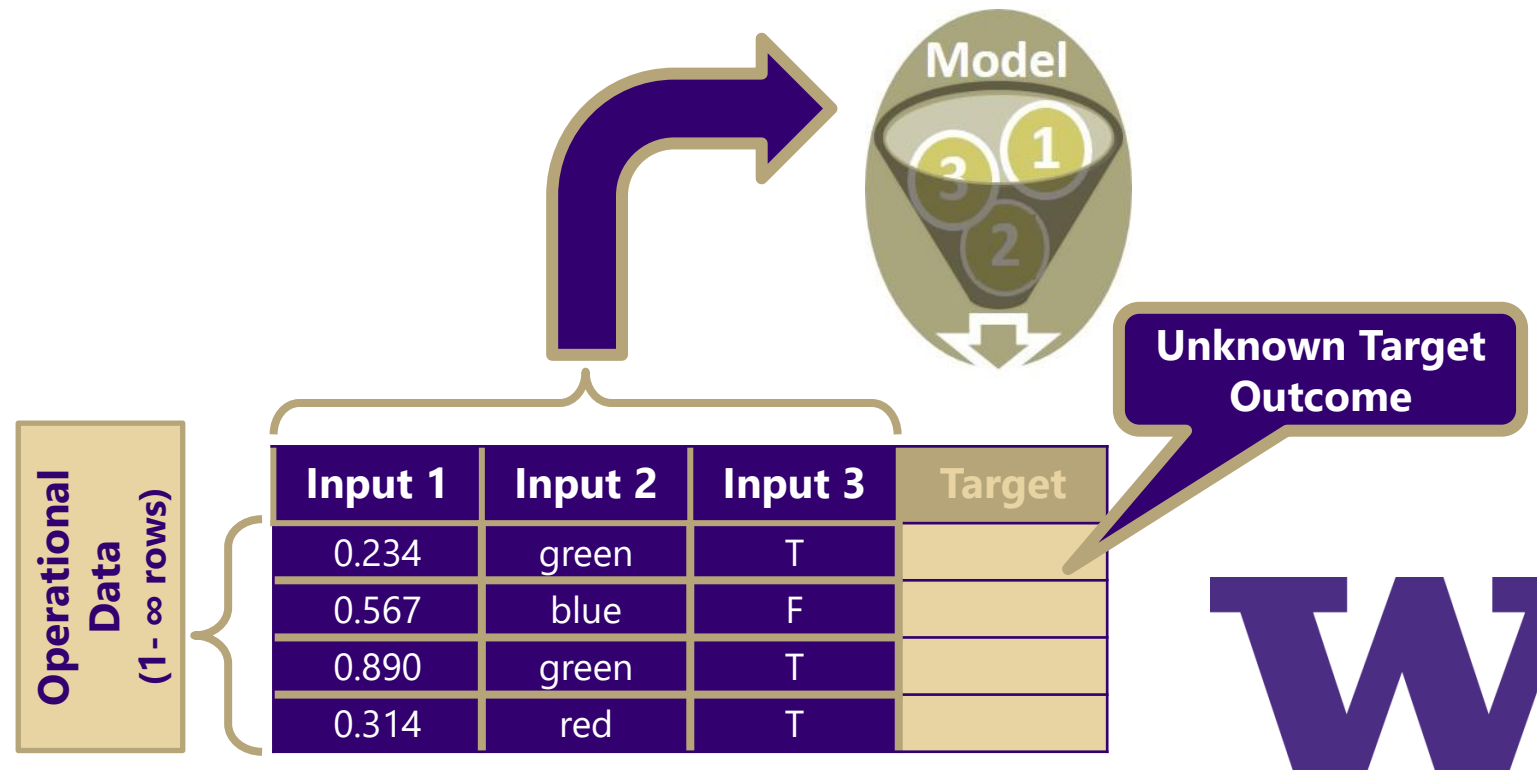
SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



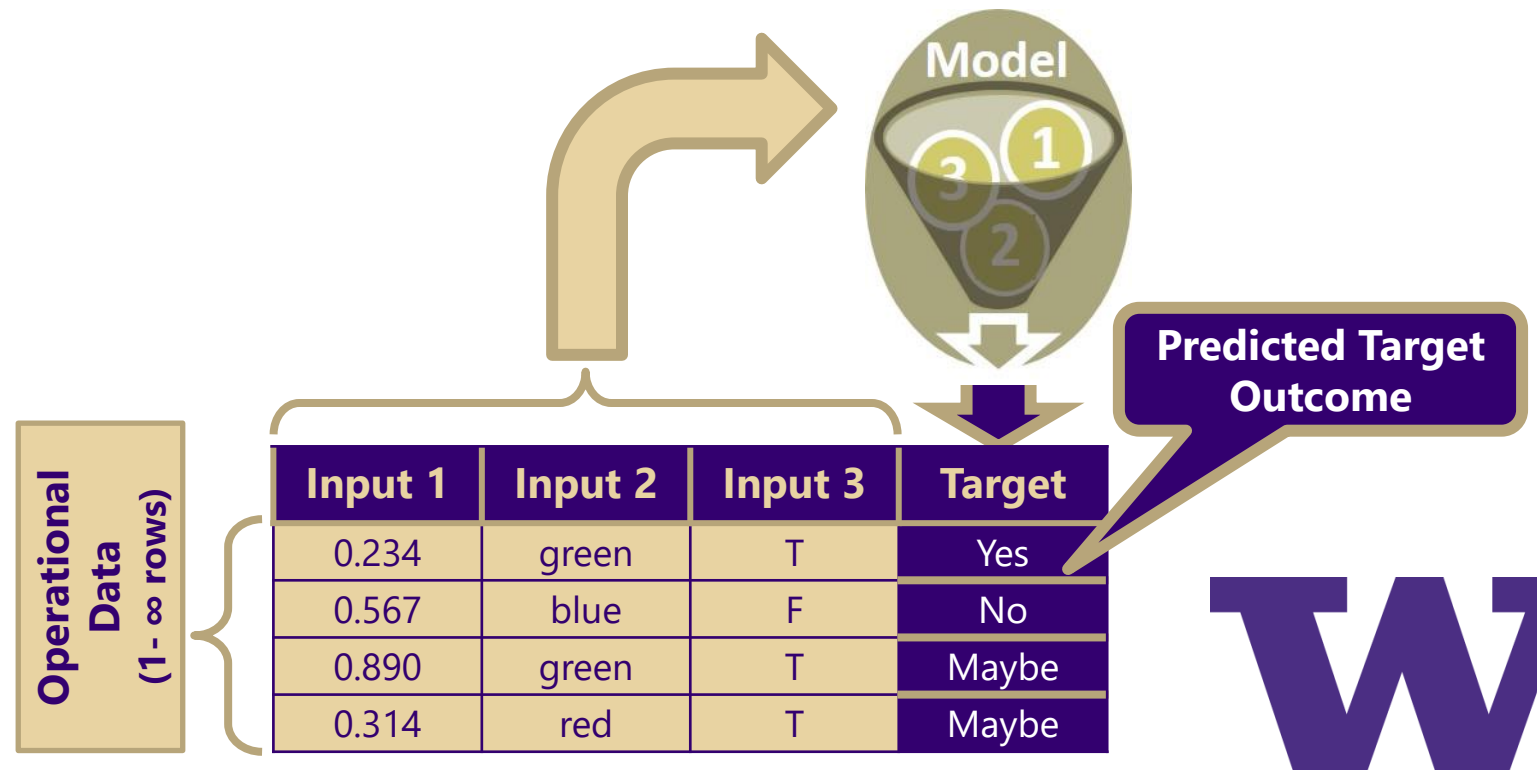
SUPERVISED LEARNING SCHEMA

Target from Input 1, Input 2, Input 3



SUPERVISED LEARNING SCHEMA

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.



Lesson 02 In-class Quiz#2