DATASCI 510

Data Science: Process and Tools

Lesson 09

Supervised Learning

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Reflections

The beauty of mathematics only shows itself to more patient followers

Maryam Mirzakhani, 1977-2017

Lesson 09 Agenda

- Finish Lesson 8 (Jupyter Notebook) *** due to power outage last week ***
- ML Overview
- Supervised Learning Data Flow
- Break
- Supervised Learning Schema
- Metrics: Accuracy, RMSE, MAE, R^2
- Lesson_09_a_Supervised_Learning.ipynb
- Break
- Lesson_09_b_Accuracies.ipynb
- Lesson_09_c_assignment.ipynb
- Interview Question

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

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

Overview: Machine Learning

- Machine learning uses algorithms that learn from data to analyze patterns and infer outcomes
- Training data is data used to teach a supervised or unsupervised learning model

Overview: Unsupervised Learning

- Unsupervised learning investigates patterns in the data
- Segmentation / Clustering is unsupervised learning to organize data into groups
- Anomaly Detection is unsupervised learning to find data that differ from the rest
- Principal Component Analysis (PCA) reorganizes data so that earlier dimensions have more variance

Overview: Supervised Learning

- Supervised learning attempts to mimic an expert in predicting "expert" values (labels) that can be either numbers or categories
- Regression is supervised learning to predict a numeric value
- Classification is supervised learning to predict a category
- Test data is data used to evaluate a supervised learning model
- Predict means to apply a supervised learning model although sometimes it is also used (incorrectly) for unsupervised learning
- Has a special variable: expert label, label, target, outcome, target outcome, y, result, known result, etc.

Machine learning Mapped to Tasks

We manufacture consumables on hundreds of production lines with thousands of machines. A machine break down leads to unscheduled expensive maintenance with prolonged production stoppage. We introduce machine learning to reduce maintenance costs and production stoppage.

Technology	Business Goal	Question that is Answered
Anomaly Detection (Unsupervised)	I want to see if my machine is behaving normally. I want an intelligent alarm that tells me that my engine is not behaving correctly.	Is a specific machine behaving normally?
Clustering / Segmentation (Unsupervised)	I want my maintenance efforts to be more targeted. That is why, I want to group machines into clusters that behave similarly.	Which machines are like each other in terms of maintenance and repair requirements?
Time Series Regression (Supervised)	I have created a graph that plots the number of machine failures week over the last 4 years. I want to project that graph into the future to estimate how many repair technicians I will need.	When will we experience peak periods of technician need and machine failures?
Estimation / Regression (Supervised)	Before I engage a repair technician for proactive maintenance, I want to estimate the repair costs and downtime costs.	How much will the repair of a specific machine cost?
Classification (Supervised)	I have too many machines that would benefit from proactive maintenance. I need to prioritize those machines that have the highest probability of failure.	What is the probability that a specific machine fails within the next two weeks?

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

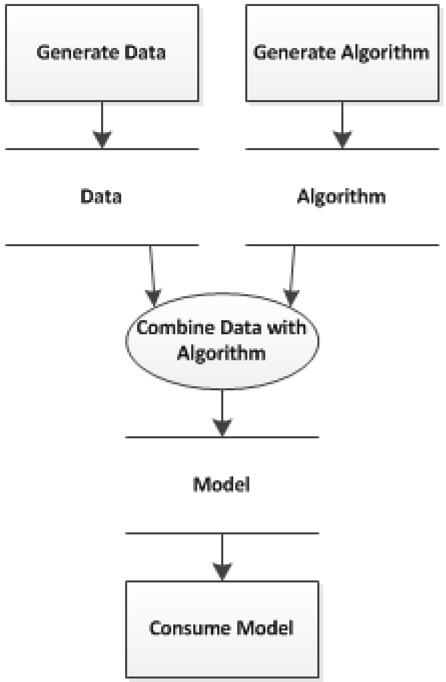
Data Flow in Supervised Learning

FROM DATA TO PREDICTIONS

- How do we get from data to predictions?

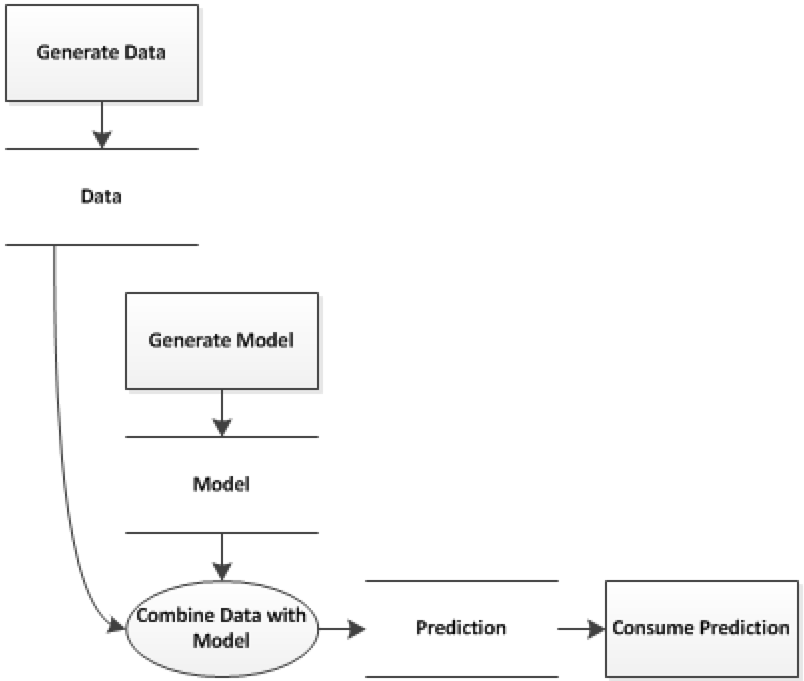
Data → ? → Predictions

FROM DATA TO PREDICTIONS



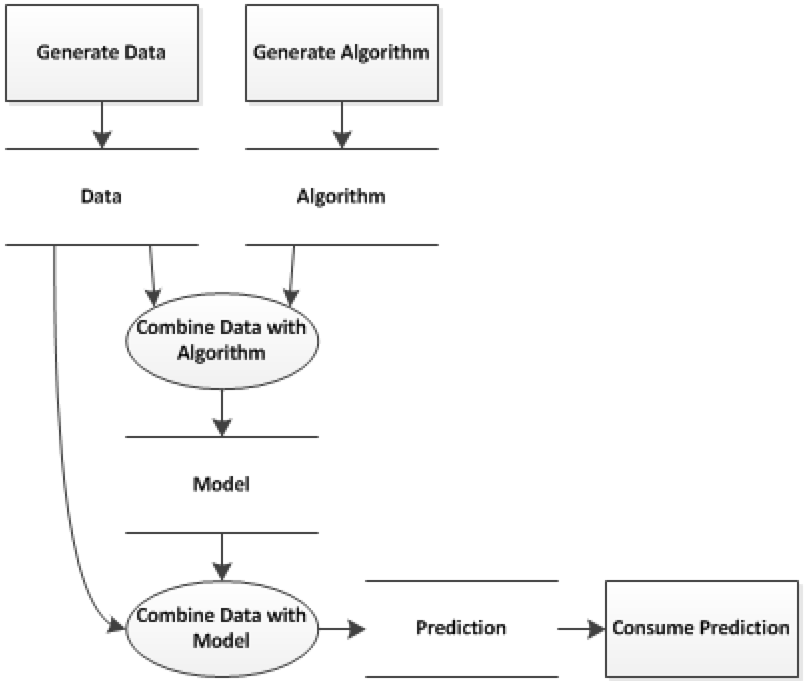
Training Data + Algorithm → Model

FROM DATA TO PREDICTIONS



Model + Operational Data → Prediction

FROM DATA TO PREDICTIONS



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

FROM DATA TO PREDICTIONS

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

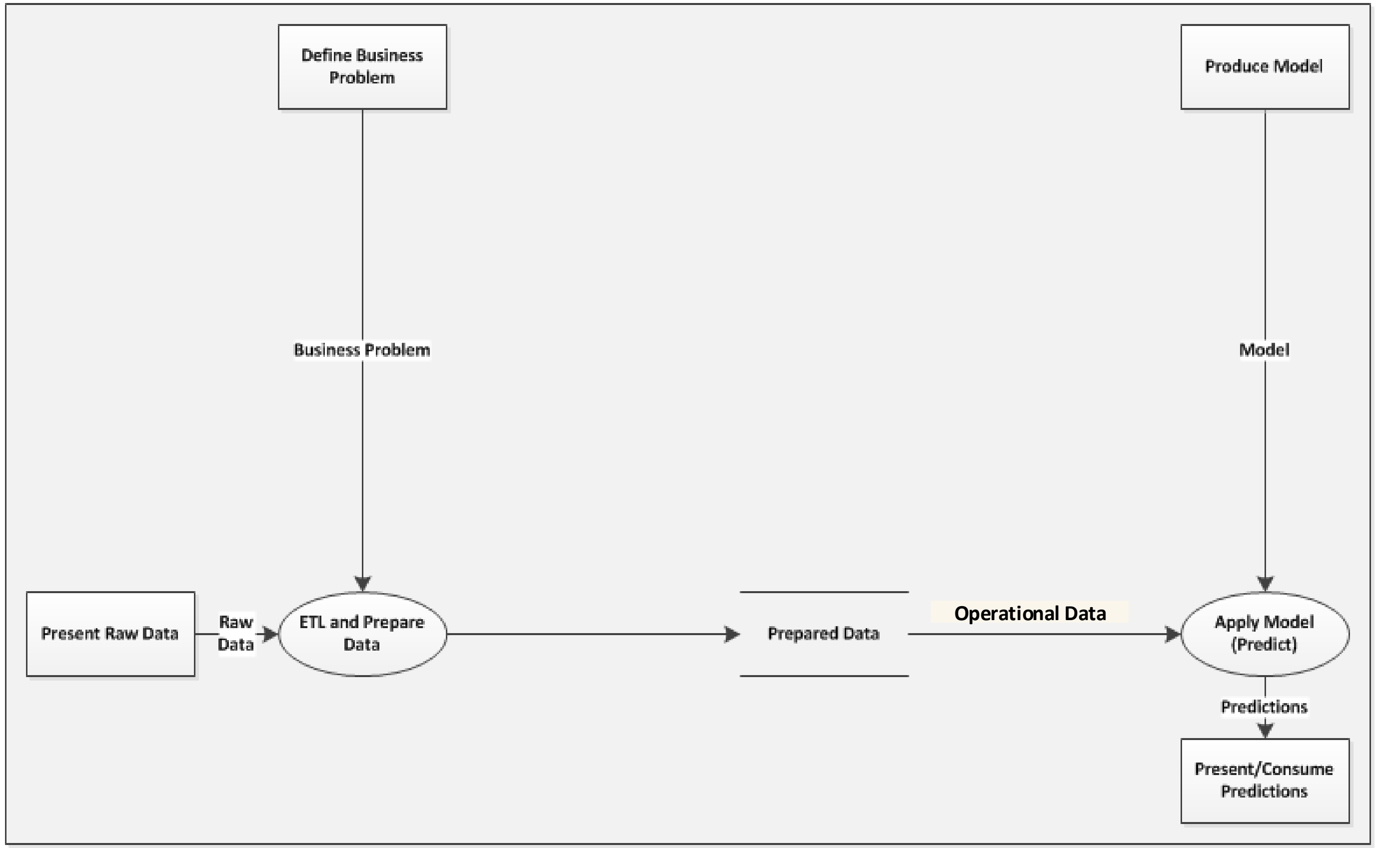
Training Data + Algorithm → Model
Model + Operational Data → Prediction

DFD OF SUPERVISED LEARNING

MODEL ACTS ON DATA

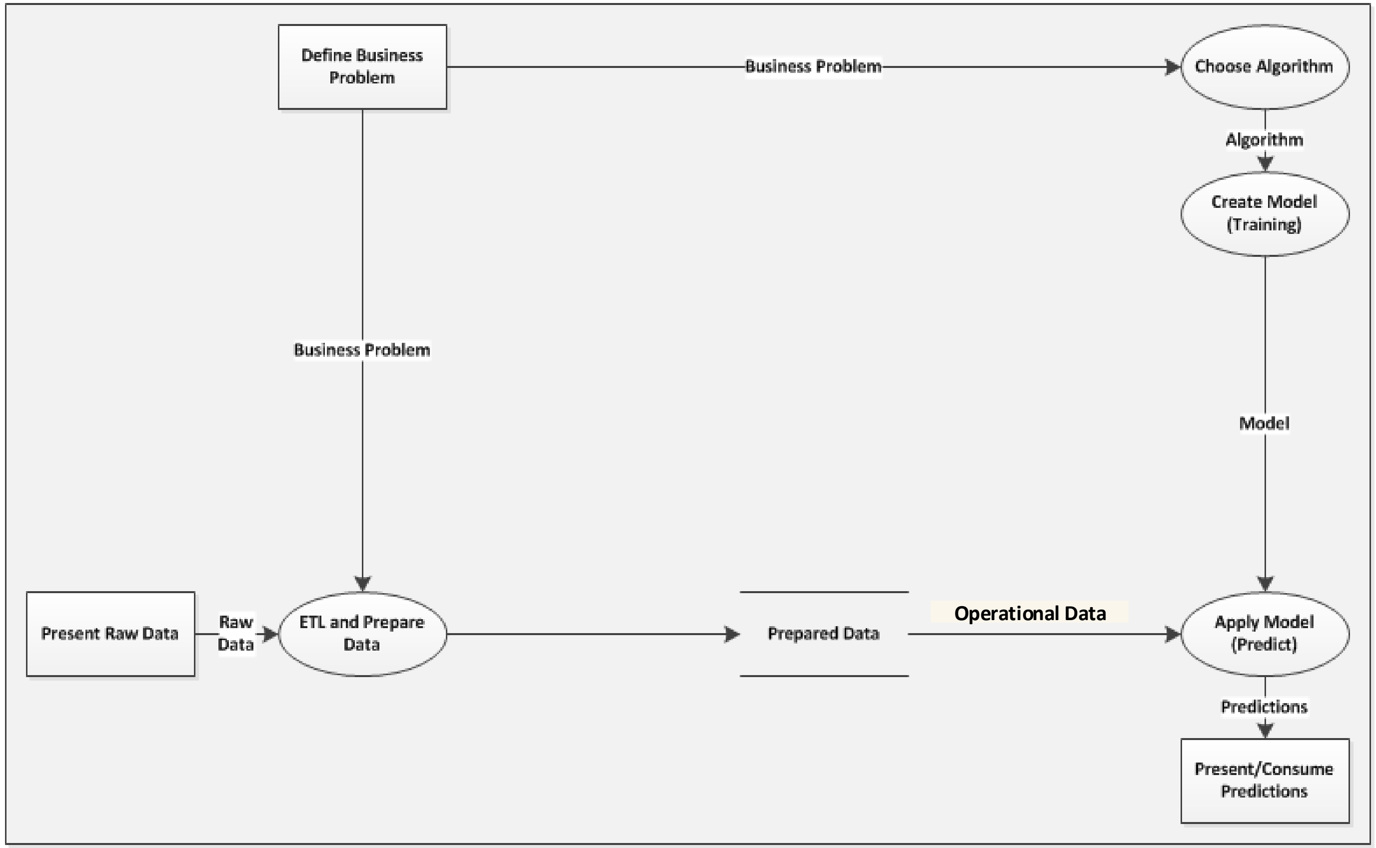
Model + Data → Prediction

DATA ETL AND PREPARATION DRIVEN BY BUSINESS PROBLEM



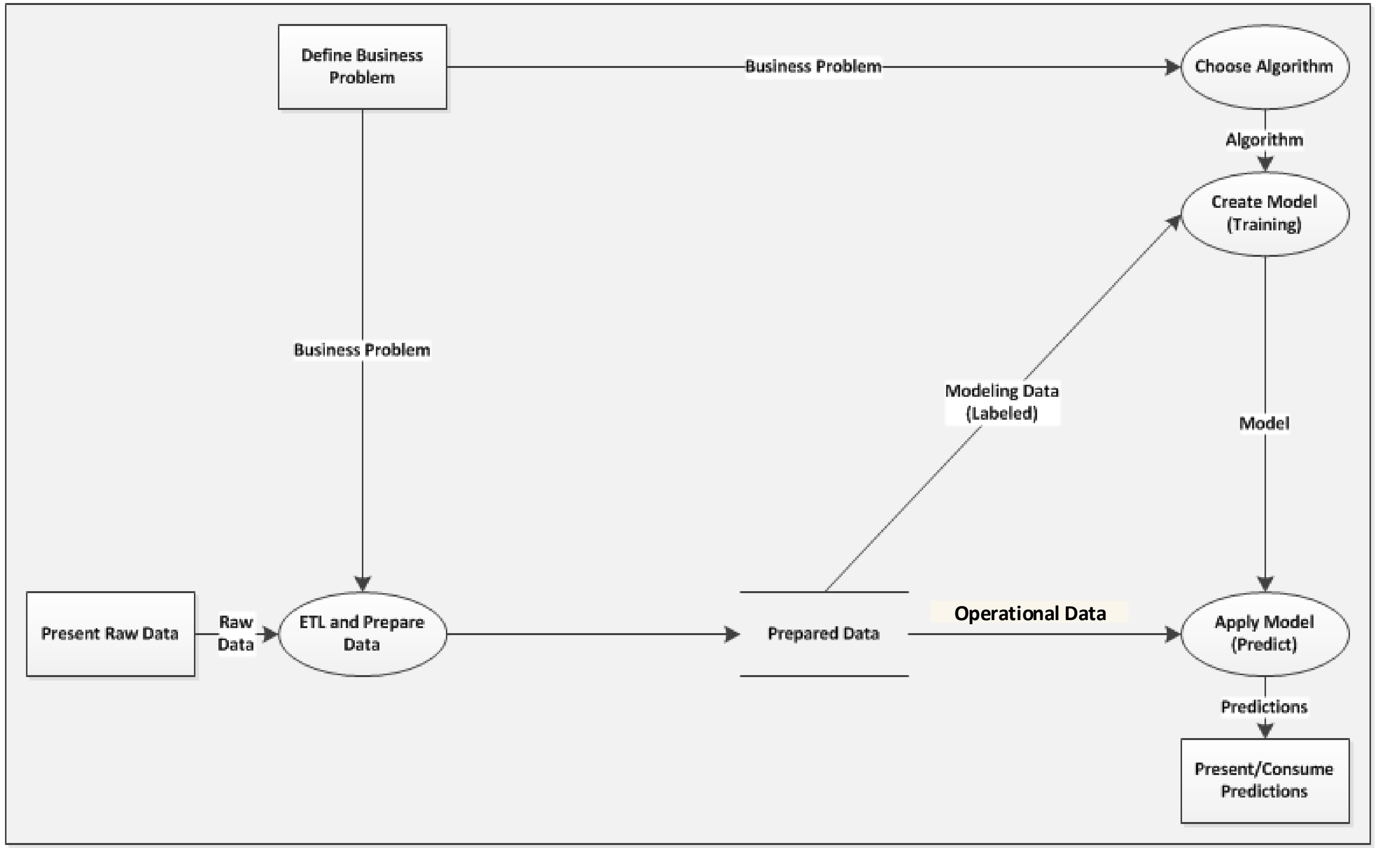
Business Problem determines ETL and Data Prep

ALGORITHM CHOICE DRIVEN BY BUSINESS PROBLEM



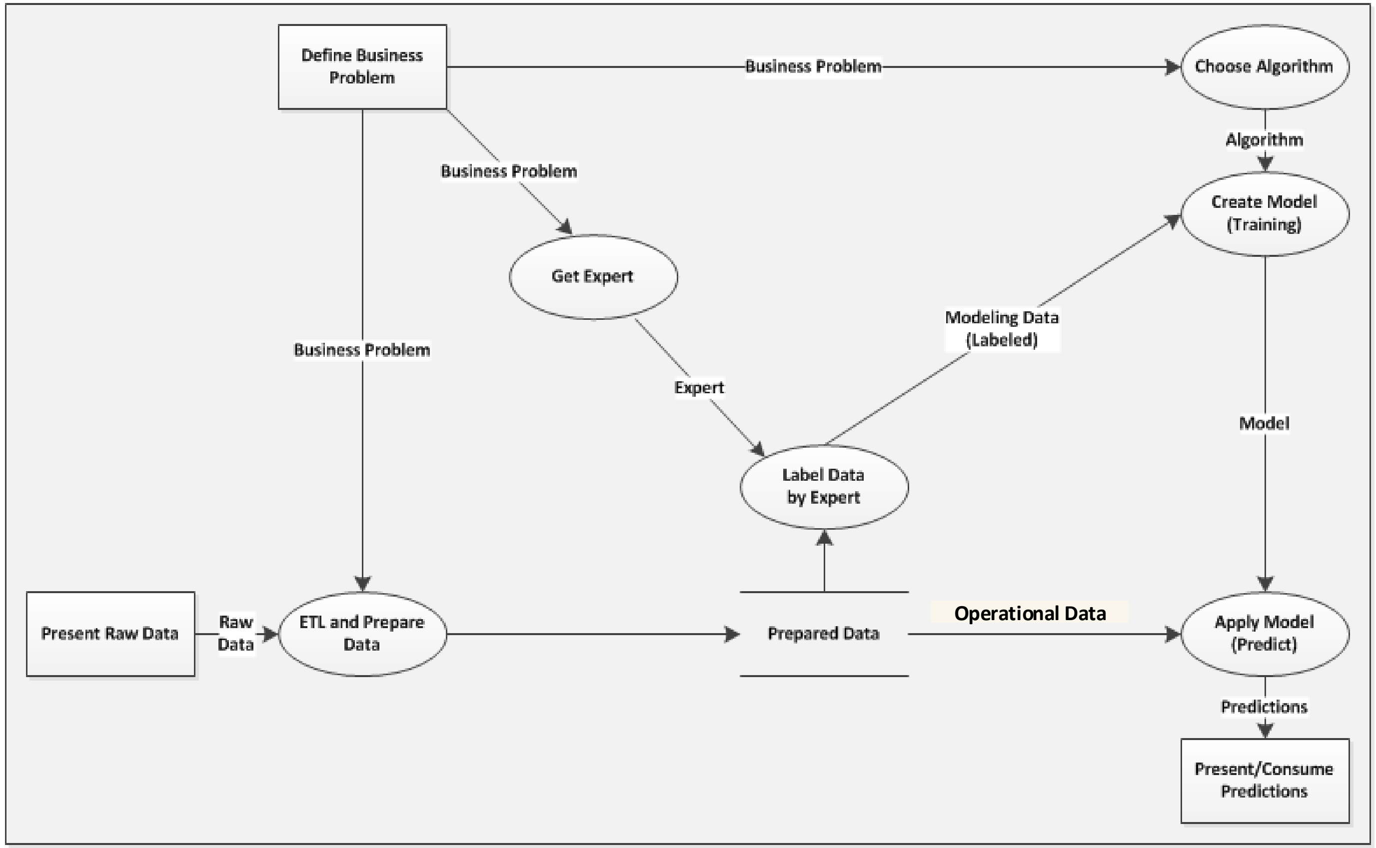
Business Problem determines the choice of Algorithm.

MODEL CREATION NEEDS DATA



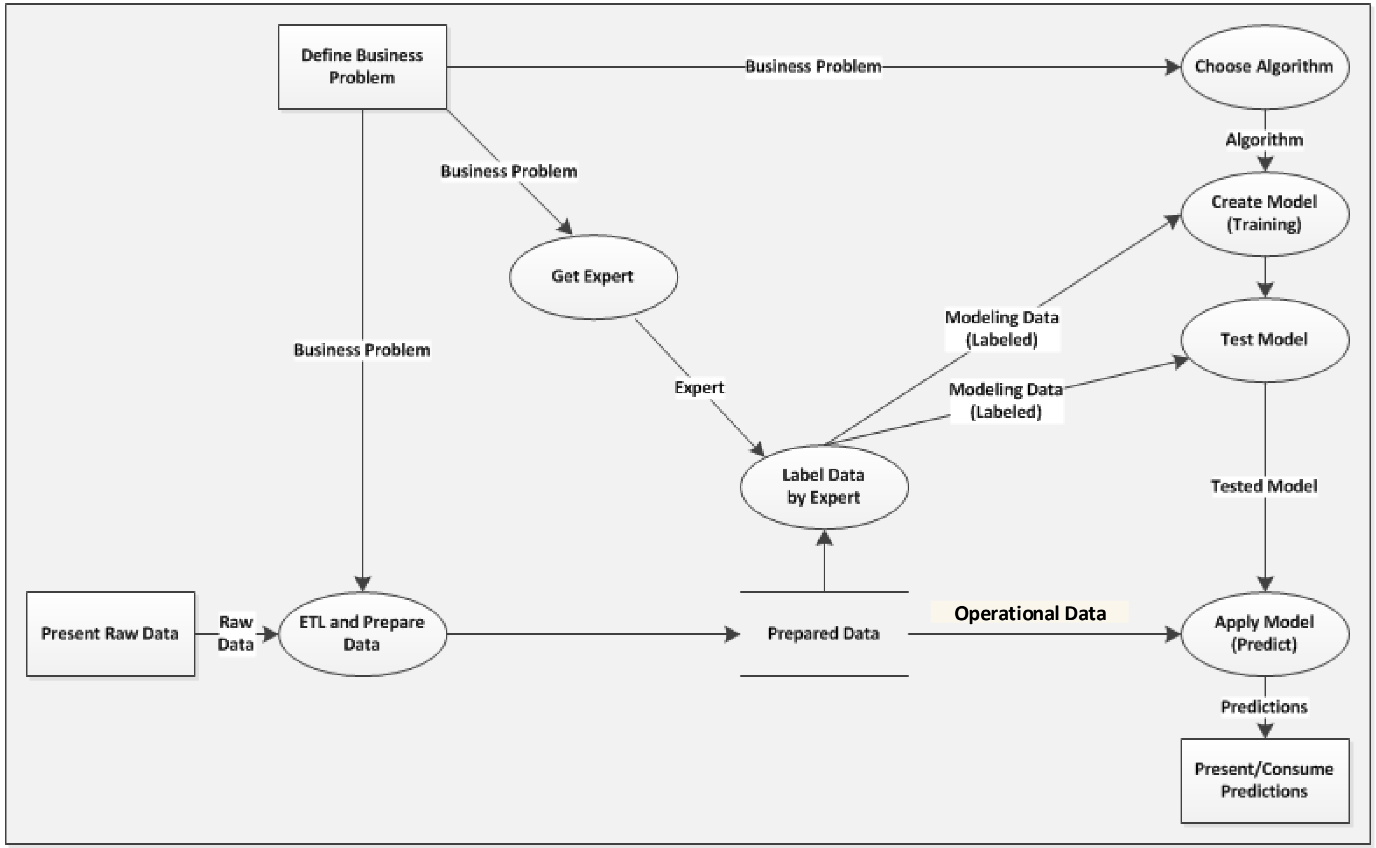
Data + Algorithm → Model

SUPERVISED TRAINING NEEDS DATA LABELED WITH OUTCOMES



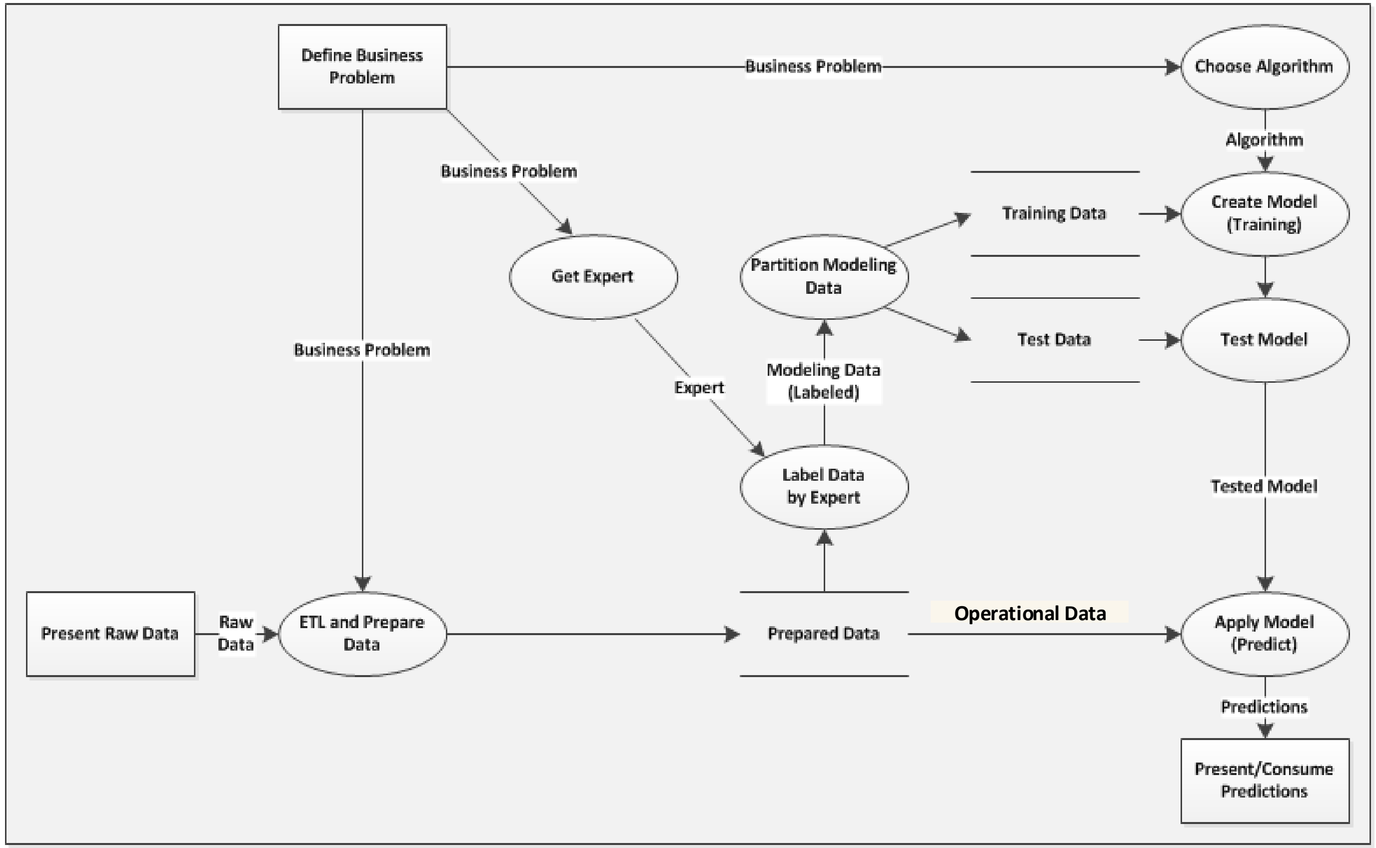
Supervised Learning requires expert labeling of data.

MODELS NEED TO BE TESTED



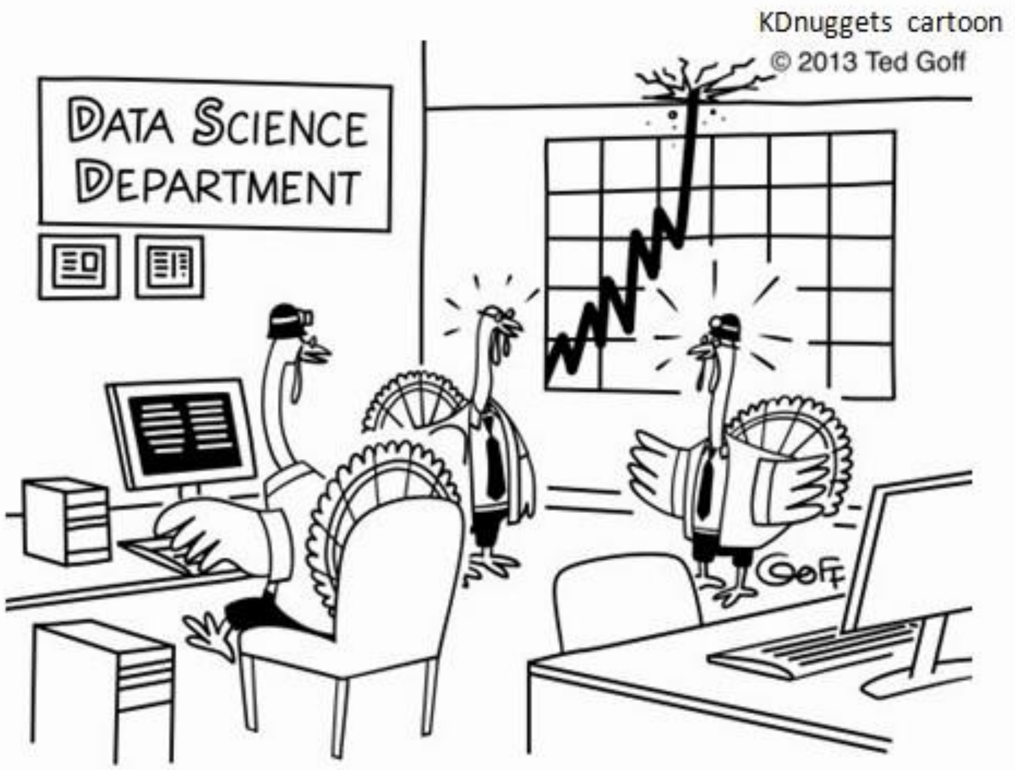
Do not trust predictions from an un-tested model!

TRAINING & TESTING OF MODEL USE DIFFERENT DATA



Do not test a model using training data!

Break



"I don't like the look of this. Searches for gravy and turkey stuffing are going through the roof !"

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

SUPERVISED LEARNING SCHEMA

SUPERVISED LEARNING SCHEMA

Rectangular Dataset (aka table)

- Modeling Dataset
  - Vertical Partition (Both Input and Output are needed for training and testing)
    - Input columns
    - Output column (target, outcome) is categorical for classification or numeric for regression
  - Horizontal partition of modeling data into training and test data
- Operational (Incremental) data
  - Schema is same as modeling data, except:
    - Input columns are used to predict target outcome
    - No output column (target, outcome)
    - Not partitioned into training and test data

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.

Machine Learning Terms and Concepts

Machine Learning

- Supervised Learning: Requires expert labels
  - Regression: Teach a machine to predict a numeric label
  - Classification: Teach a machine to predict a categorical label

- Unsupervised Learning: Does not make use of expert labels
  - Clustering
  - PCA
  - Anomaly Detection

Machine Learning Terms and Concepts

Training

- Supervised Learning: Requires expert labels
  - Features can be categorical and/or numeric
  - Labels:
    - **Numeric** labels are used as examples to teach a **regression** to predict a numeric label
    - **Categorical** labels are used as examples to teach a **classification** to predict a numeric label
- Unsupervised Learning: Does not make use of expert labels
  - No expert labels: The data, the algorithm, and the hyper parameters determine the outcome

Testing

- Supervised Learning: Data is split between training and testing
- Unsupervised Learning: No testing data. No concept of accuracy. No testing

Machine Learning Terms and Concepts

Predictions

- Supervised Learning
  - hat notation ($\hat{y}$ or y_hat). For Example:
    - **y_hat = supervisedLearner.predict(X)**
  - soft prediction determines the probability of a class
    - E.g. The refrigerator has a 25% chance of breaking down in the next 2 months
  - hard prediction in a classification predicts the class like: Positive vs Negative
    - E.g. The refrigerator is predicted to breakdown in the next two months
  - A soft prediction has more information than a hard prediction. You can get a hard prediction from a soft prediction but not vice versa
  - A hard prediction is equivalent to a soft prediction that is compared to a threshold, typically at 50%. For Example:
    - 0 to 49.9% is Negative
    - 50 to 100% is Positive
- Unsupervised Learning: No Predictions! (There is matching which is somewhat similar)

Machine Learning Terms and Concepts

- **Regularization**
  - Don’t train models to over-think their predictions
  - Penalize your model for using too many coefficients or making too many decisions
  - By simplifying the model, the model is more generally applicable. i.e, the model is “generalized”

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Some metrics in ML

Classification: hard vs. soft predictions

- Hard classifiers only return the prediction for the class
- Soft classifiers return the "probability" or confidence that prediction is in each class
  - example: 3 classes - cat, dog, squirrel
  - hard prediction: dog, soft prediction: [0.12, 0.84, 0.04]
- Most classifiers return soft predictions
- We can later set a threshold and obtain hard predictions
- Most performance metrics depend on the choice of the threshold

Performance measures for classification

> The easiest way to measure a classifier's performance is accuracy:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{number of predictions}}$$

> Works both for **binary** and **multi-class** classification

> We can use **weighted accuracy** to emphasize certain classes

> Accuracy is not a smooth function, so ML algorithms use alternative functions to optimize over, such as **cross-entropy**

Regression: residuals

> Measuring performance in **regression** is more straight-forward

> **Errors** or **residuals** are the difference between the model's predicted value and actual value (ground truth) for each row

> We use the **hat notation**: $\widehat{Y}_i$ is the prediction at i^{th} row and Y_i is the actual, so $Y_i - \widehat{Y}_i$ is the **error** or the **residual**

> A good model should **minimize error** on the test data, and the error should look like **random noise** (nothing left to predict)

Performance measures for regression

- Root Mean Squared Error: **RMSE** = $\sqrt{\frac{1}{N}\sum_i (Y_i - \widehat{Y}_i)^2}$
- Mean Absolute Error: **MAE** = $\frac{1}{N}\sum_i |Y_i - \widehat{Y}_i|$
- Coefficient of Determination: **R^2**
  - Is 1 minus the ratio of $\sum_i (Y_i - \widehat{Y}_i)^2$ over $\sum_i (Y_i - \bar{Y})^2$ where $\bar{Y}$ is the mean of Y
  - R^2 represents proportion of variation explained by the model
  - $R^2 = 0$ means a totally useless model and $R^2 = 1$ means a perfect model
- **Adjusted R^2**: lowers R^2 in proportion of the number of features (discouraging us to just keep adding useless features)
  - $R^2 = 1 - (1-R^2) * \frac{n-1}{n-p-1}$
  - where n and p are respectively the number of data points and the number of features

Break


HOW DID OUR AD PERFORM ON YOUR NETWORK?
GREAT!
YOUR AWESOMENESS METRIC IS UP 22%,
YOUR COOL FACTOR METRIC IS UP 36%,
YOUR SMILEY-FACE METRIC IS UP 58%,
AND A.M.W.T.M.U. IS UP A RECORD 85.4%.
WHAT'S A.M.W.T.M.U.?
ANOTHER METRIC WE TOTALLY MADE UP.
TOM FISH BURNE

Interview question

- Assume you are at one end of a tunnel, with a perfectly rectangular surface and need to move to the other end of the tunnel
- You are told that there are a given number of mines on the surface of this tunnel and their exact emplacements are known
- Furthermore, you know that it is safe to walk around the mines, as long as you keep a distance of r from the location of the mines
- Propose an algorithm to determine whether a safe path exists through the tunnel