Machine Learning

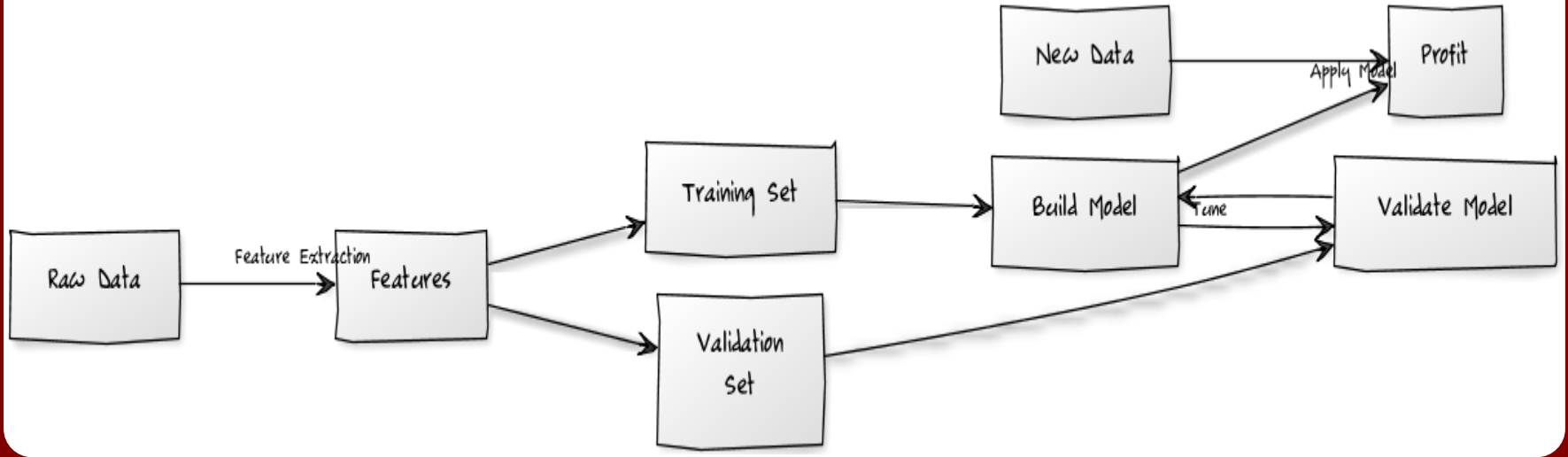
an introduction

Machine Learning

A branch of artificial intelligence in which a computer generates rules based on raw data that has been fed into it.

Supervised Learning

Takes a known set of inputs (samples) with an expected set of outputs (targets) and builds a predictive model that generates statistically similar outputs for new input.



Supervised Learning

Spam Filtering

OCR

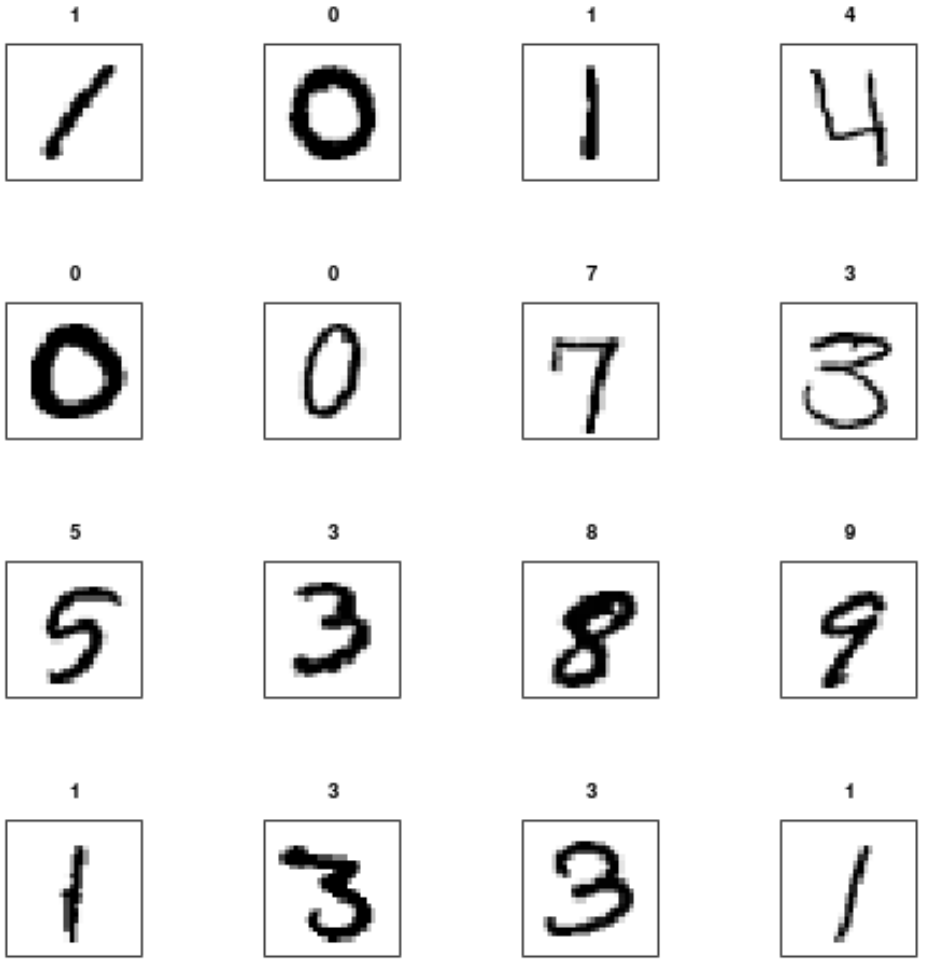
Fraud/Abuse Detection

Price Modeling

Types of Supervised Learning

Classification:

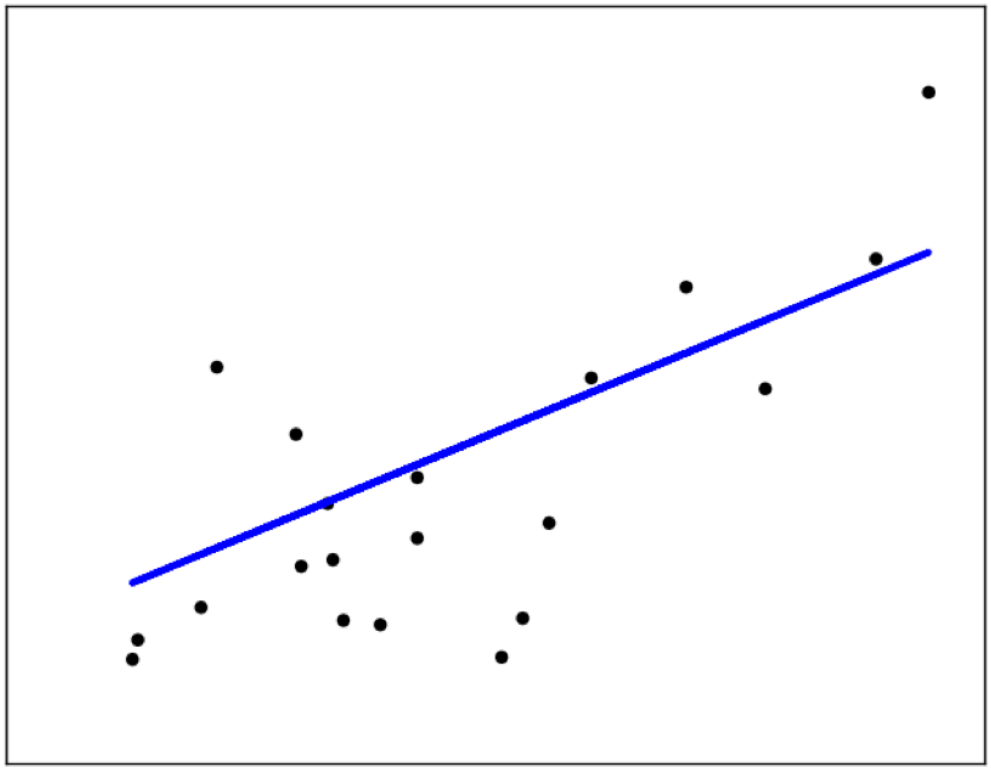
Choose from a finite set of classes (labels) based on a set of input features.



Types of Supervised Learning

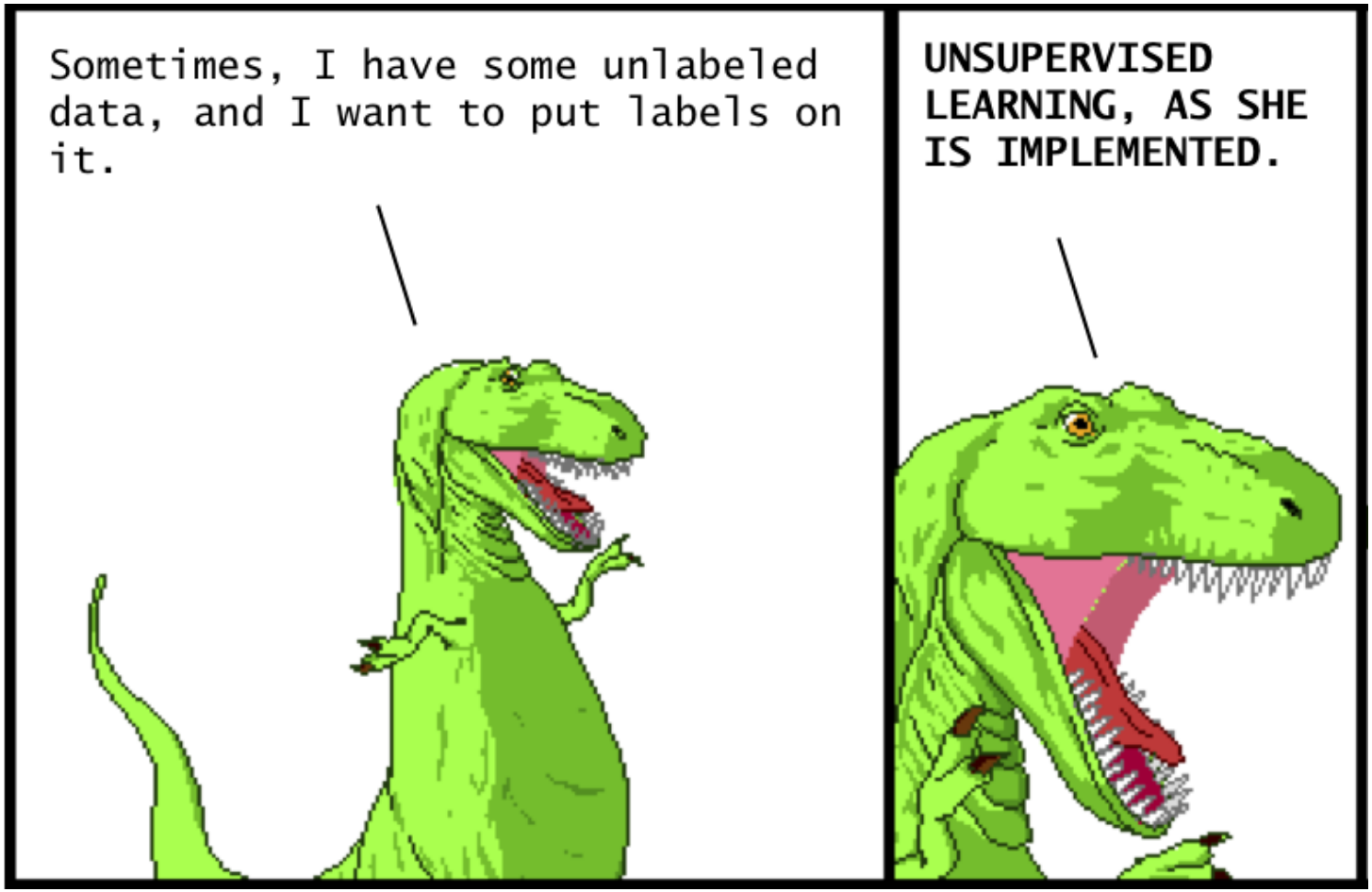
Regression:

Predict a continuously varying variable based on a set of input features.



Unsupervised Learning

Trying to find structure in unlabeled data.



Unsupervised Learning

Recommendation Systems

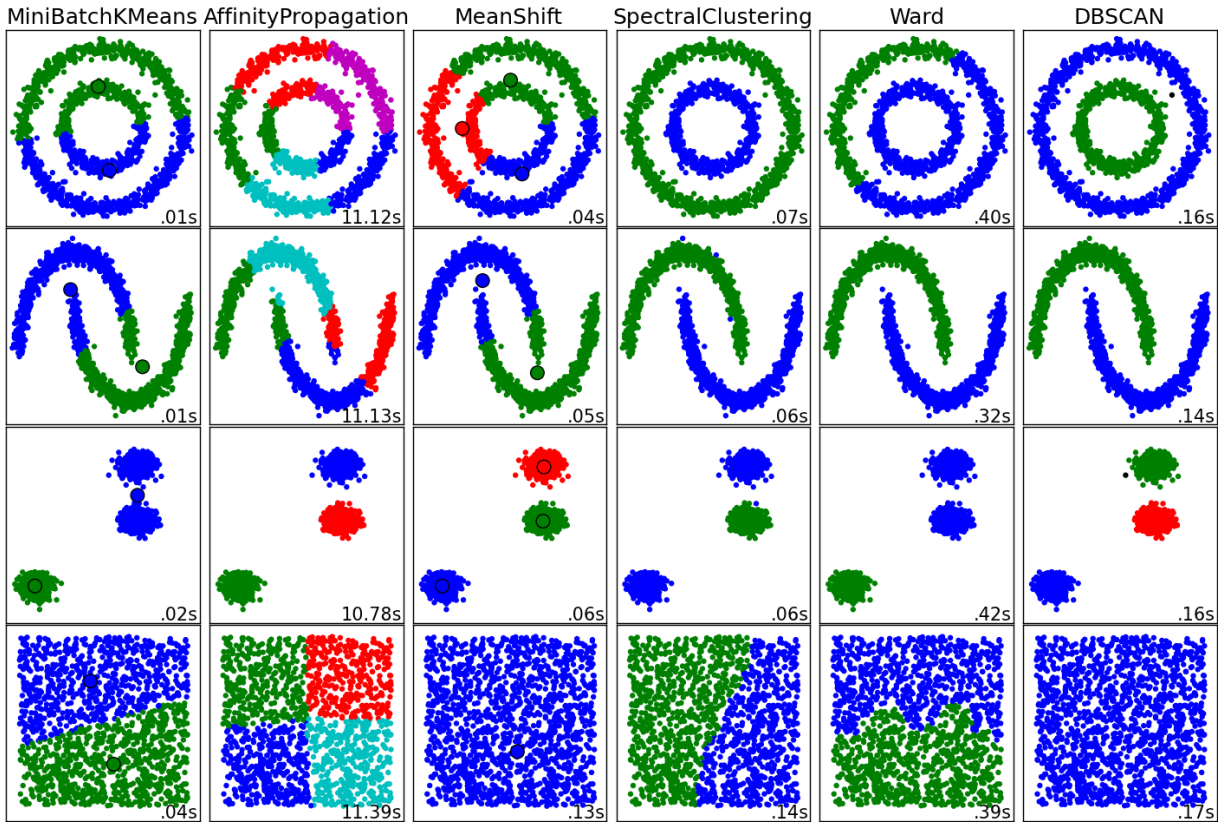
Crime Analysis

Market Research

Types of Unsupervised Learning

Clustering:

Assign unlabeled data to groups of similar data (clusters).



Other types of Machine Learning

Semi-Supervised Learning: Few samples are labeled and many are not labeled.

Reinforcement Learning: Useful for problems where the correct actions/reward are not known in advance.

Main steps

Obtain

Scrub

Explore

Model

i**N**terpret

Model Evaluation

Depending on the type of algorithm, there are many metrics/scores you can use to evaluate your model.

Classification metrics (e.x. accuracy, f-score, matthews correlation coefficient)

Regression metrics (e.x. explained variance, mean-squared error, mean-absolute error, R^2)

Overfitting & Cross-validation

Overfitting is when you model fits the input data well, but has poor performance when presented with new data.

One good way of avoiding overfitting is to use cross-validation.

Cross-validation usually involves splitting your samples into several groups of training and test data.

Further Resources

Coursera ML Course.

Machine Learning by Tom Mitchell

Popular ML frameworks:

- **sklearn** (*scikit-learn python*)
- **WEKA** (*java, has a nice GUI*)
- **R** (*Not a framework popular for data mining*)
- **Julia** *(Not a framwork, popular for real-time data mining)*
- **Matlab** *(Popular computing toolset that many people have experience with)*
- **Mahout** *(Scalable ML libraries. Built on Hadoop)*

Questions?

http://nlp.cs.