Introduction to Machine Learning

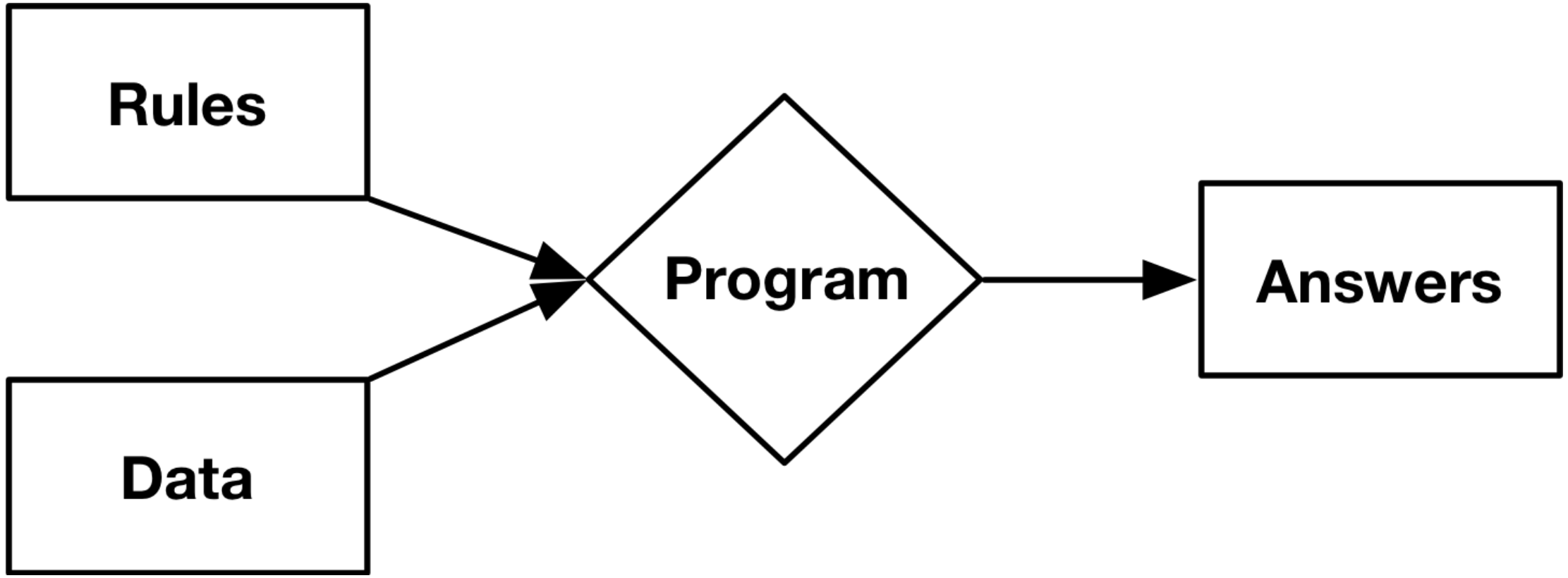
CS3300 Data Science

RJ Nowling

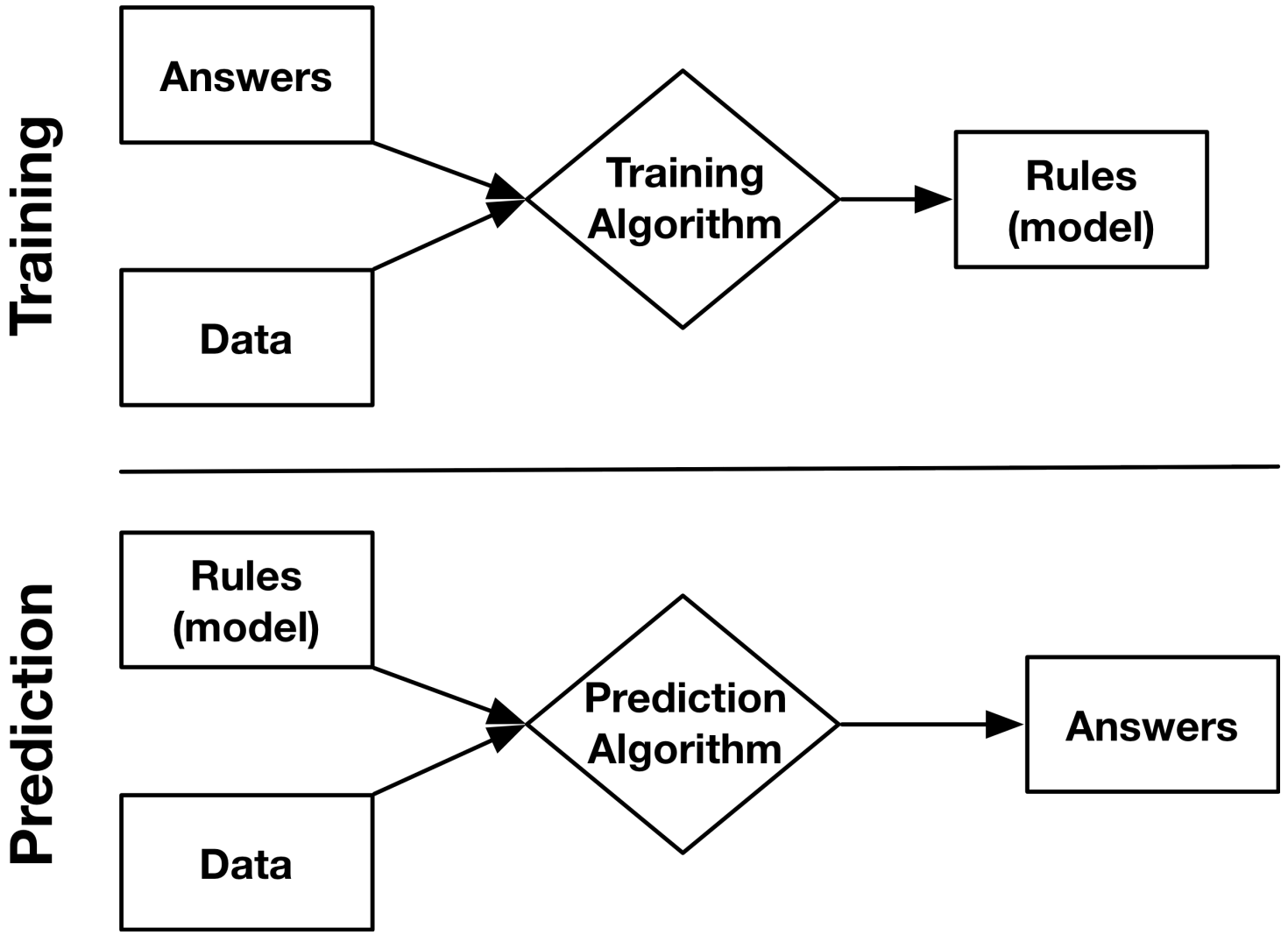
Readings

- Chapter 7
- Section 11.5

Traditional Programming



Machine Learning



Machine Learning

- Goal: automated decision making
  - Is this a cat or a dog?
- Reduce the time / cost to make decisions
- More consistent decision making
- More accurate decision making

Social Media

- "Timelines" on Facebook, "What you missed" on Twitter
- Filter, recommend "relevant" content – move away from purely chronological display
- User engagement important to retention
  - Revenue depends on ad impressions
  - Can't show ads if users aren't on the platform
  - More usage -> collect more data -> show more relevant ads
- Users leave when bored or frustrated
- Need to quickly and continually connect users to engaging content

Copyright and SOPA

- User-generated content: free, produced at massive scale
- Stop Online Piracy Act (SOPA) – introduced 2011/2012
- Would have made social media companies liable for copyright violations in user-generated content
- Problematic: social media companies could be sued for enormous amounts of money even for small violations
- Would have to hire people to review every new post, video, etc. – prohibitive due to cost and speed -> drive companies out of business
- SOPA was prevented from becoming law

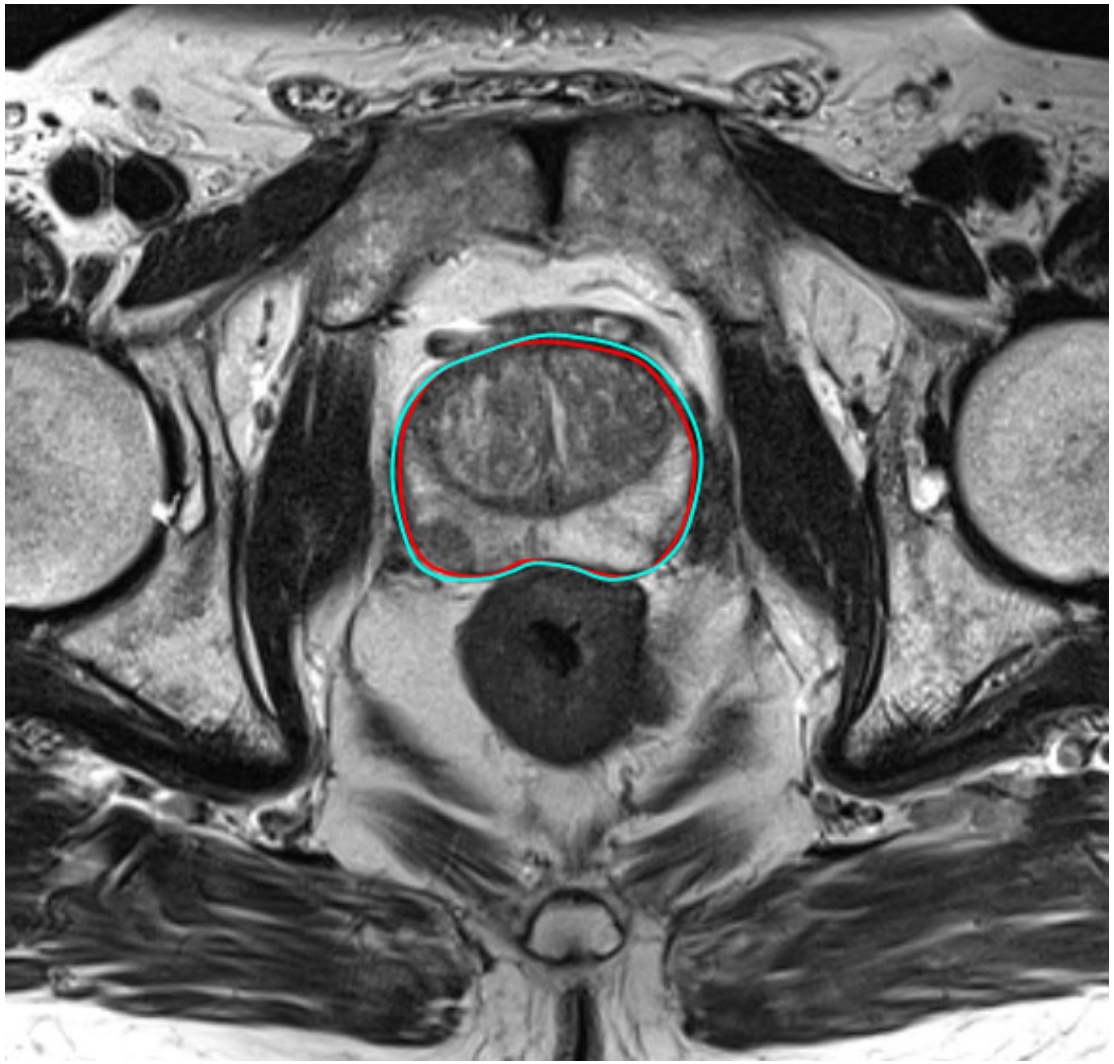
Copyright and SOPA

- Companies now use "layered" process
  - Machine learning tags to tag copyright violations in new content
  - Reject content predicted to violate copyright; let other content pass
  - Humans receive and can process appeals from users and copyright violation notices from companies
- Machine learning is less accurate but can make decisions quickly
- Users can still post content almost immediately, while reviewers only see a small fraction of overall content
- Reduce copyright infringement but in a sustainable way

Digital Advertising

- Traditional print, TV, and radio advertising can only broadcast to a wide segment of customers
- Digital adverting can be customized for each user
- More direct feedback on effectiveness of advertising by tracking clicks and purchases
- Machine learning is used in real time to:
  - Predict which companies you'll buy from
  - Which ads you'll click on

Biomedical Image Segmentation

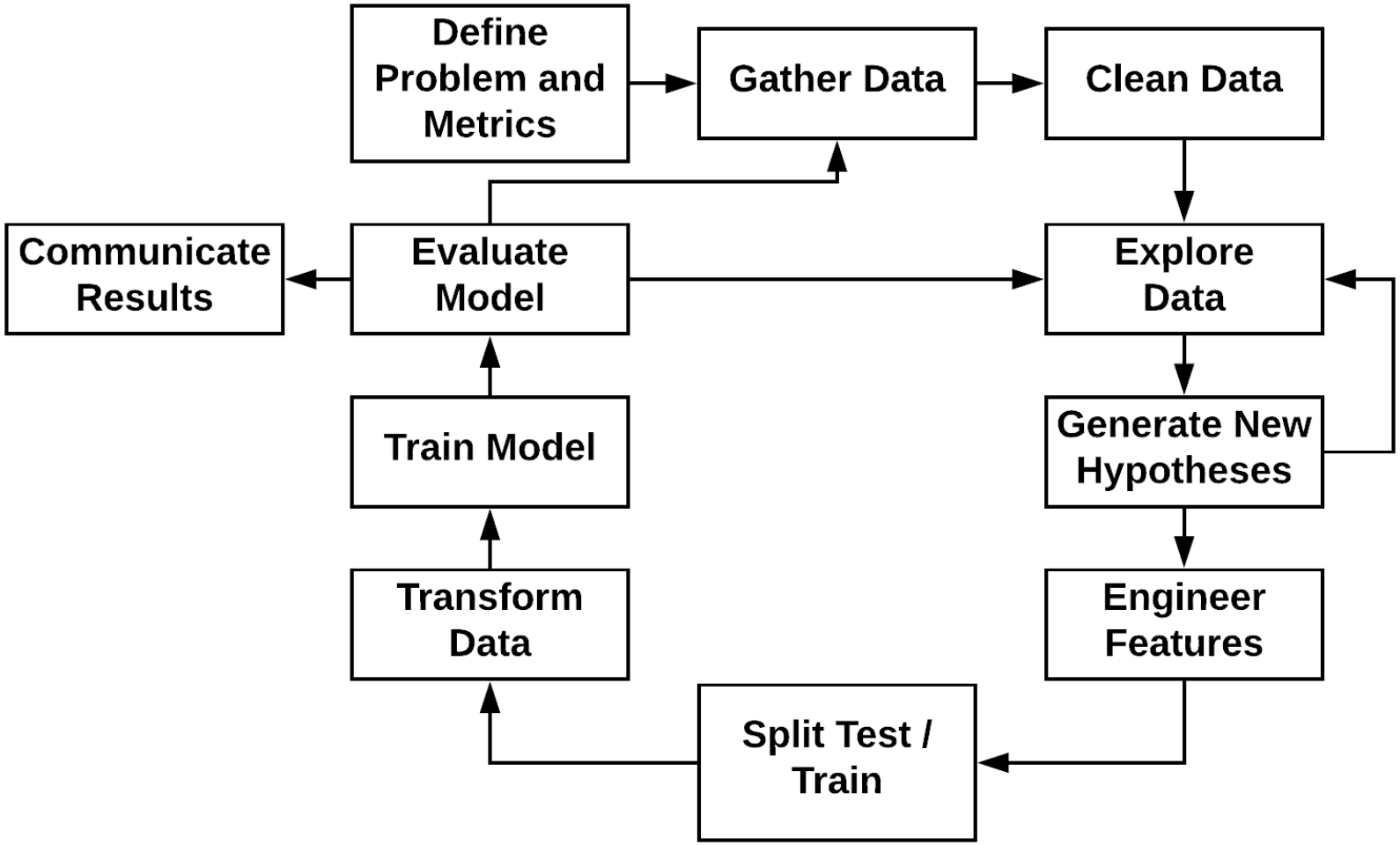


Courtesy of Dr. Peter LaViolette's group
at the Medical College of Wisconsin

Books to Read

- *The Zap Gun*
  by Philip K. Dick
  - Speculates about "perfect" weapons that can target a specific person and only that person.
- *Prediction Machines: The Simple Economics of Artificial Intelligence*
  by Agrawal, Gans, and Goldfarb
  - Reframe key innovation of ML / AI as a "drop in the cost of prediction"
  - Increased productivity – new business models
  - Reduces uncertainty – enabling new business strategies

Machine Learning Process



Define a Problem

- We want to predict the sale price for real estate transactions.

  Regression

- We want to predict whether the animal in a picture is a cat or dog.

  Classification

A Good Problem Definition

Does

- Specific
- Defines the variable to be predicted and its type
- Optionally, describes the metric used to evaluate the predictions

Does Not

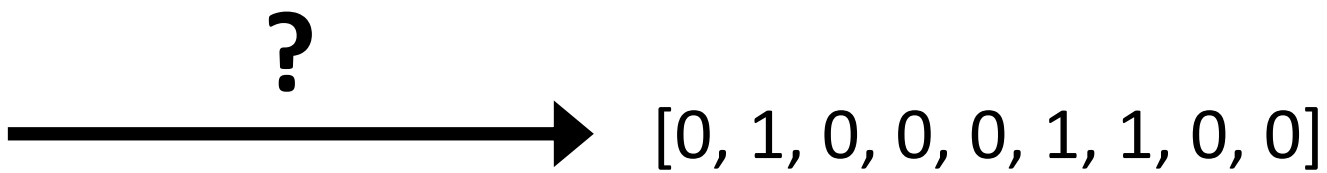
- List a specific machine learning algorithm

Feature Engineering

Dear Class,

I've uploaded the new homework to D2L. It's due on Monday. I'm looking forward to seeing your solutions!

RJ



Evaluation

- We evaluate models by comparing predictions to known values
- We use metrics to quantify the amount of error
- The choice of metric depends on:
  - The type of machine learning problem (e.g., classification or regression)
  - Whether some types of errors are more important than others

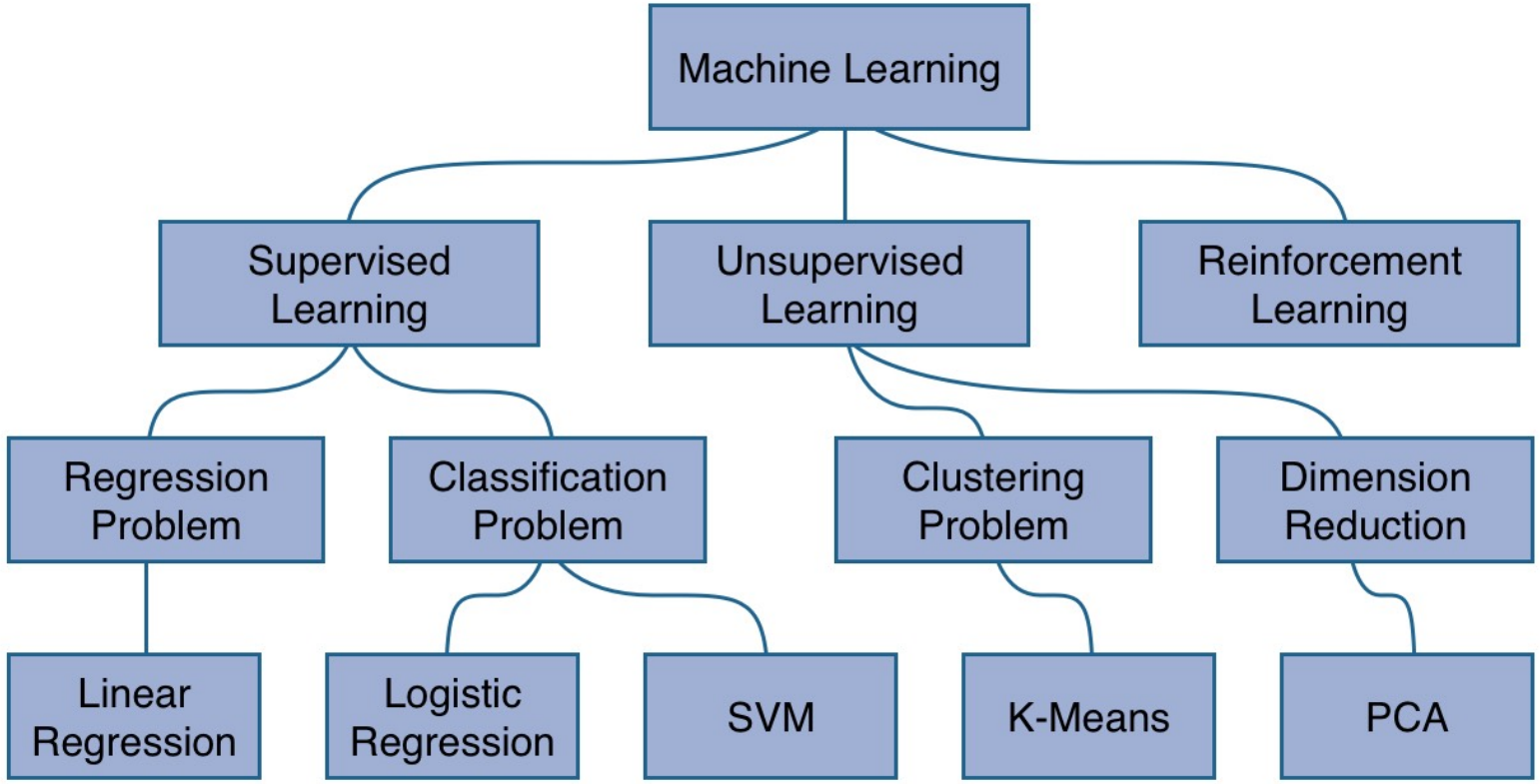
Experimental Setup

- We train machine learning models to make predictions about unseen data
- To realistically evaluate a model, we need to design an experiment that accurately reflects how we intend to use the model

Supervised Machine Learning

- Supervised
  - We are trying to predict the values of a variable
  - We train the model using samples with known values
  - We apply the model to samples for which the value is not known
- Classification -- predict a categorical value
  - Does the picture contain a cat or a dog?
- Regression -- predict a continuous value
  - How much money will the house sell for?

Machine Learning



https://www.robinwieruch.de/machine-learning-javascript-web-developers/

Common Regression Algorithms

- Linear Regression
- Polynomial Regression
- Multivariate Adaptive Regression Splines (MARS)
- k-Nearest Neighbors (kNN) Regression
- Random Forest Regression
- Neural Network Regression

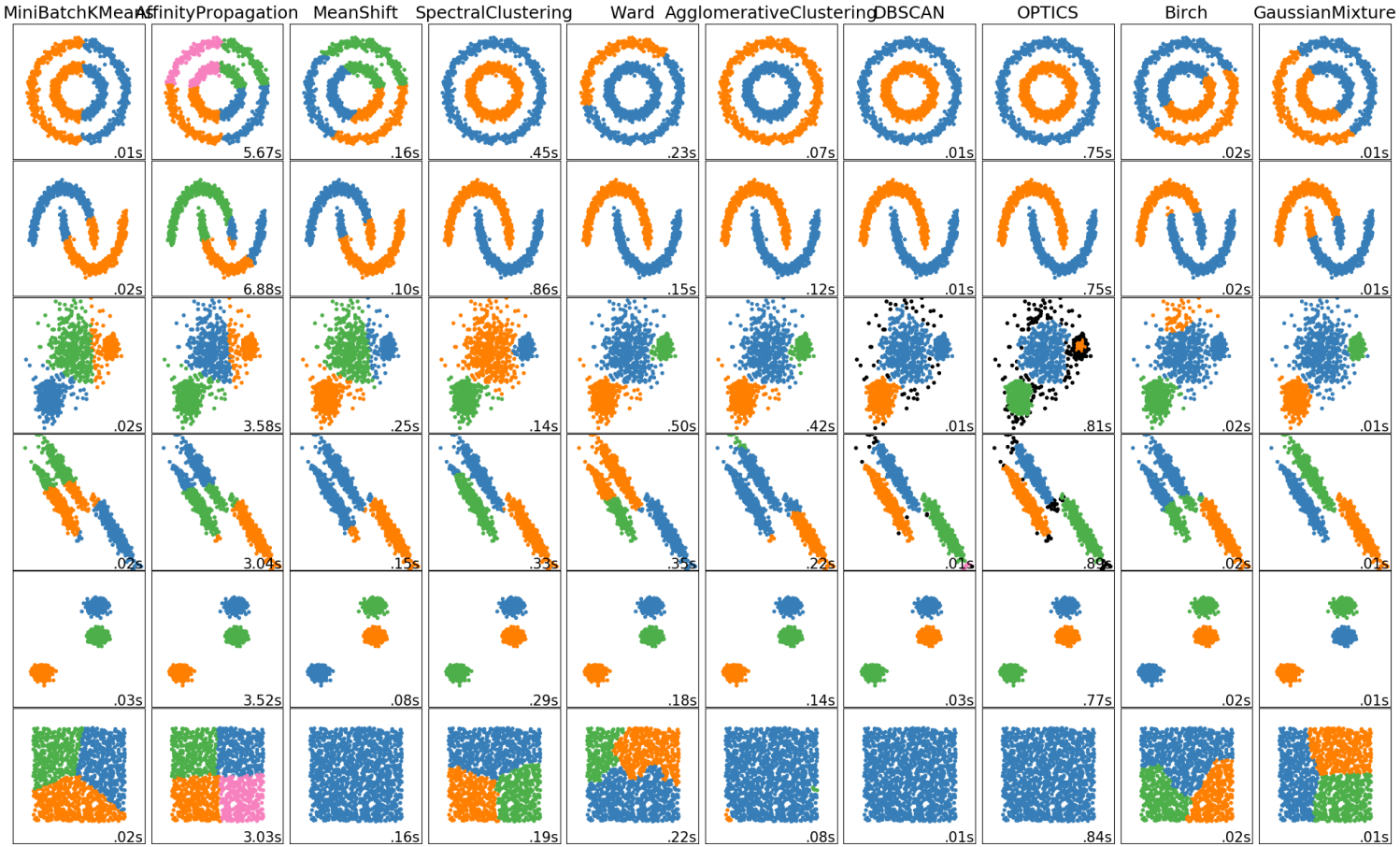
Common Classification Algorithms

- Logistic Regression
- k-Nearest Neighbors (kNN)
- Random Forests
- Support Vector Machines (SVMs)
- Naïve Bayes
- Neural Networks

Unsupervised Machine Learning

- Used when you have data that does not have an output variable
- Often used in an exploratory setting to identify relationships between samples or patterns in the data
- Common approaches include:
  - Clustering – grouping samples into a discrete set of clusters
  - Dimensionality Reduction – reducing the number of variables by grouping variables

Clustering



https://scikit-learn.org/stable/modules/clustering.html

Common Clustering Algorithms

- k-Means
- k-Medoids
- Ward's Hierarchical Clustering
- Gaussian Mixture Models
- DBSCAN
- Spectral Clustering

Dimensionality Reduction

- Principal Component Analysis (PCA)
- Non-negative Matrix Factorization (NMF)
- Uniform Manifold Approximation and Projection (UMAP)
- Variational Autoencoder (VAEs)

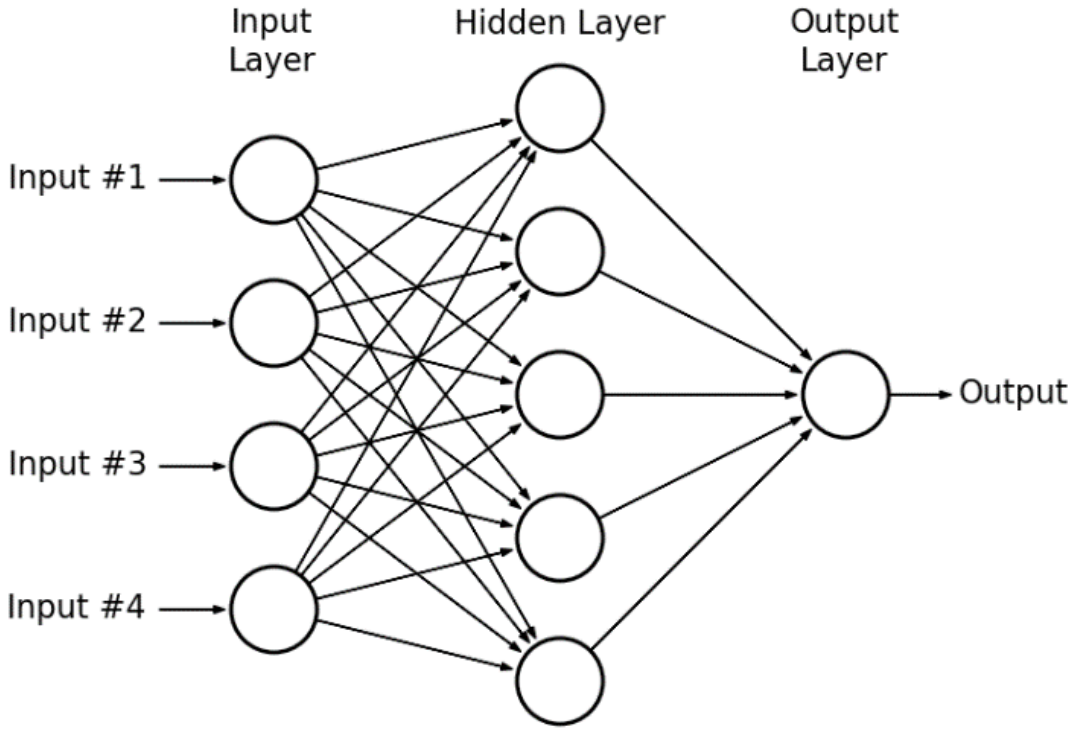
Scikit-Learn

- Popular Python library for machine learning
- One of the most widely used for machine learning, period
- Key innovations:
  - High quality software with regular releases
  - Includes a large array of algorithms – hard to find something missing!
  - Great user documentation
  - Fit nearly all models into a unified API

http://scikit-learn.org/stable/user_guide.html

Artificial Neural Networks (ANNs)

- Artificial neural networks consists of layers of neurons
- Each layer usually has the same type of neurons
- The outputs of the neurons in the previous layer are used as the inputs of the neurons in the next layer



Artificial Neural Networks (ANNs)

- Invented in 1943 by Warren McCulloch and Walter Pitts
- Inspired by neurons in animal brains

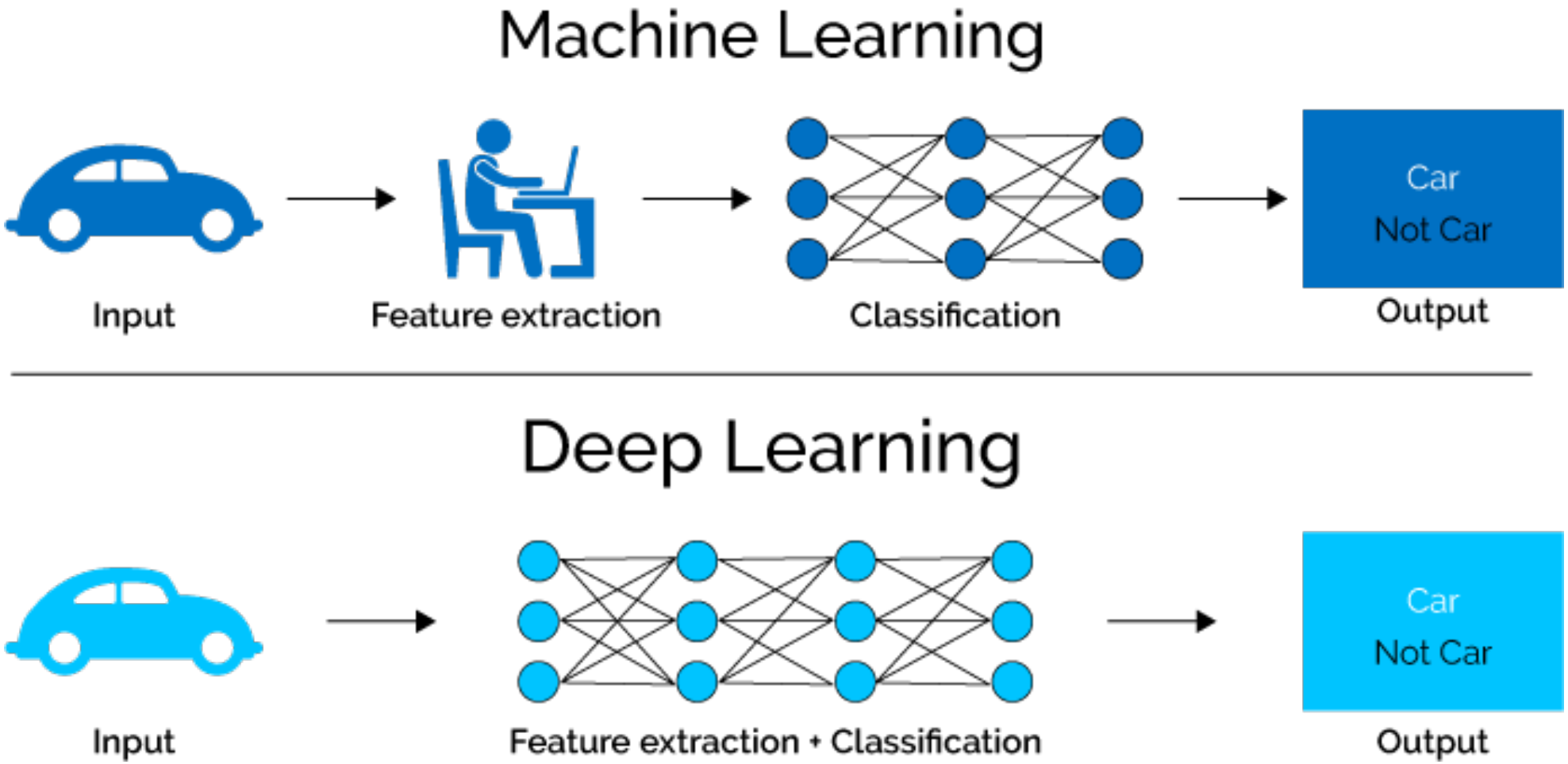
Artificial Neural Networks (ANNs)

- Despite their long history, ANNs have only recently out-performed other machine learning models
- In 2012, a convolutional neural network called AlexNet outperformed all competitions in the ImageNet competition
  - In 2011, a good classification error was 25%
  - AlexNet achieved an error rate of 16% -- a 10% improvement and significant step forward
  - By 2015, deep learning models were out performing humans
- Deep learning models are ANNs with many layers

Artificial Neural Networks (ANNs)

- Contributing factors that enabled deep learning:
  - Increase in available data – deep learning methods require a massive amount of data to avoid over-fitting
  - Increase in computational power – GPUs are well-suited to training deep learning models
  - Algorithm advances – small changes to the math solved the "vanishing gradient" problem which had previously prevented using more than one or two hidden layers
  - Easy-to-use, high-quality, and free software frameworks that run on commodity hardware

Deep Learning



https://www.upwork.com/hiring/for-clients/log-analytics-deep-learning-machine-learning/

What You'll Learn

- Unsupervised learning for exploratory data analysis
- Defining classification and regression problems
- Experimental setup for training and evaluating models
- Feature engineering and selection
- Evaluation metrics
- We will focus on applications rather then methods

What You Won't Learn

- Details of machine learning algorithms
- More than one example ML model per problem type
- Deep learning techniques
- Techniques for handling images, audio, and other complex data types
- Good news! You'll learn this in the Machine Learning and Deep Learning courses