Wharton
University of Pennsylvania

Principles of Machine Learning

Nick Lind

Wharton Analytics Fellows

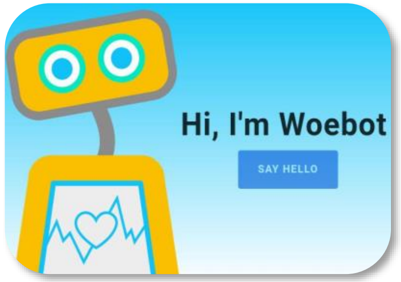
Modern Applications of AI Research

AI can be broken down into roughly five distinct research areas originating from the Total Turing test

Artificial Intelligence

Machine Vision	Natural Language Processing	Machine Learning	Robotics	Expert Systems

FarmTech
Prospera



Therapist Chatbots
Woebot

Pro Gaming
OpenAI

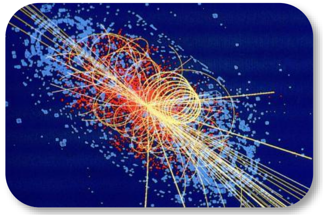
Robot Bankers
Commonwealth Bank of Australia

Warehouse Ops
Amazon

Introduction to Machine Learning

Whether you realize it or not, you are impacted by machine learning every single day

1. Machine learning algorithms are responsible for many **scientific and business model innovations**

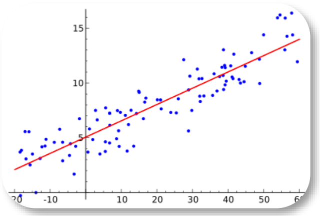


Discovery of Higgs boson

2. **Deep learning** architectures (e.g., recurrent, convolutional neural networks) are **to thank for a variety of modern technologies**

Siri, Alexa, and Cortana

3. Contrary to popular belief, **not all ML algorithms are obscure and overly-complex**

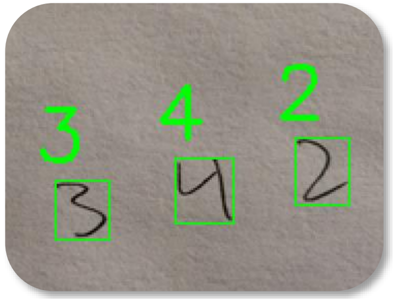


Linear Regression

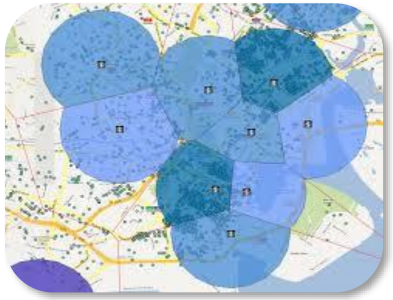
Four Types of Machine Learning

There are four types of problems that we aim to solve with ML, and each requires a different approach to learning and deployment

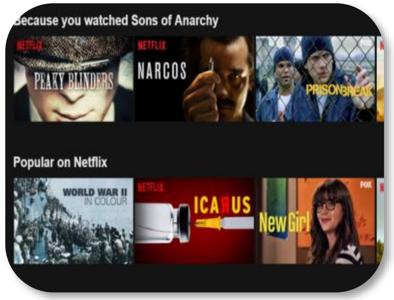
Machine Learning			
Supervised	Unsupervised	Semi-Supervised	Reinforcement
Generate predictions by training on **labeled datasets**	**Expose** and **visualize** hidden relationships and anomalies in **unlabeled datasets**	**Generate predictions** using a **small amount of labeled data** within a larger pool of unlabeled data	**Create an agent** capable of taking environmental actions to **maximize utility** over time



Handwriting Recognition



Geospatial Market Segmentation



Interactive Recommendations

Self-Driving Vehicles

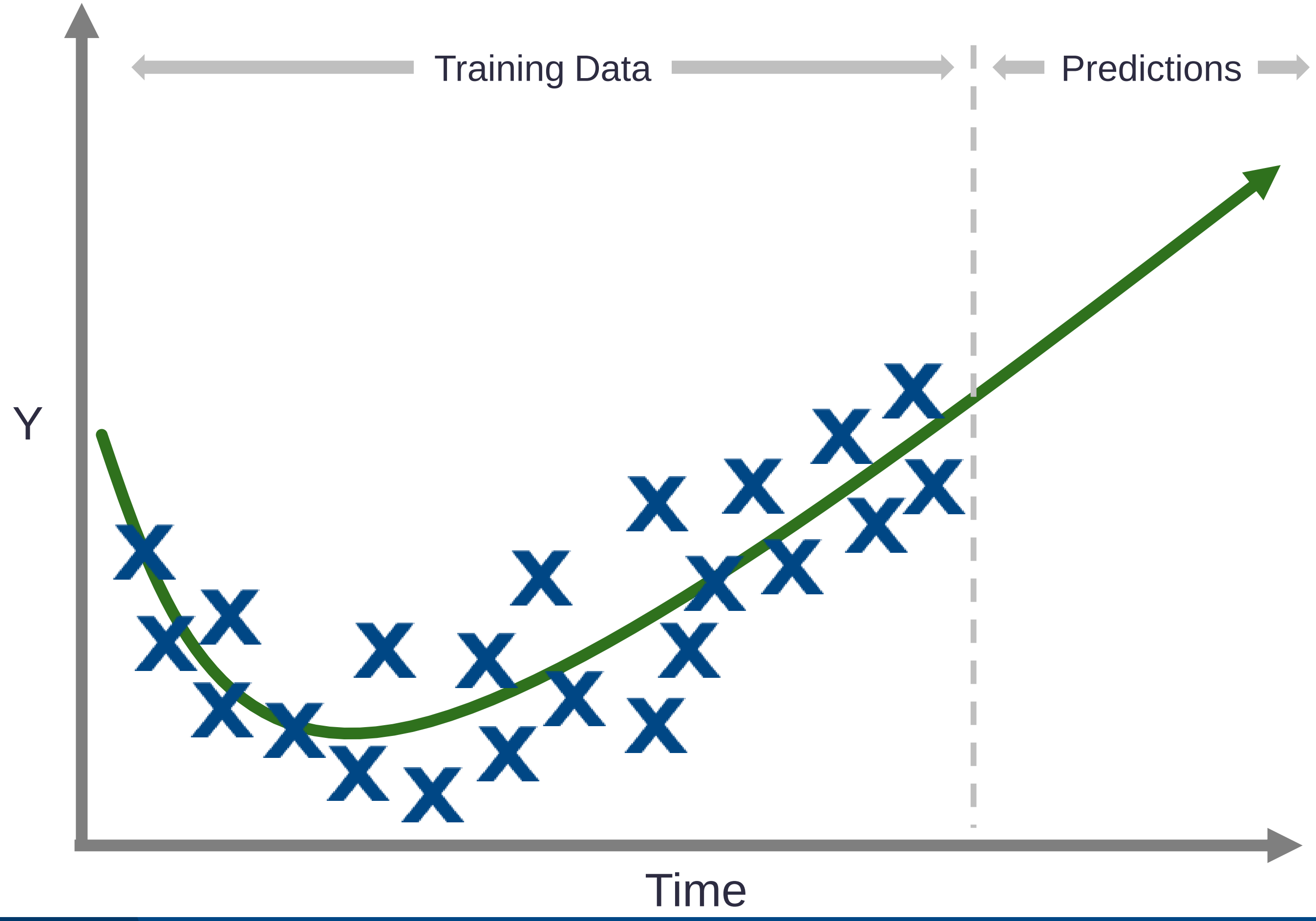
Principles of

Machine Learning

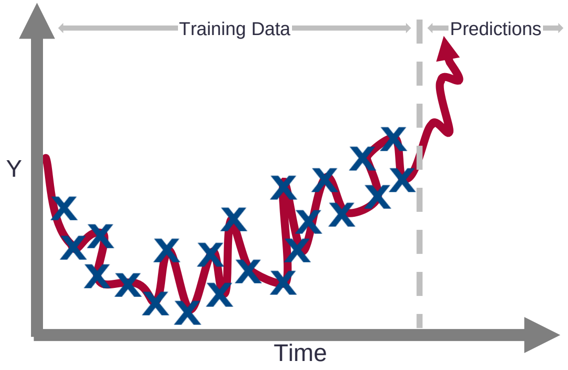
Principle #1: Generalization

When we train a machine to think, we are mostly concerned with how well it can predict the future. This often means that we need to restrain the complexity of our model to improve its ability to generalize

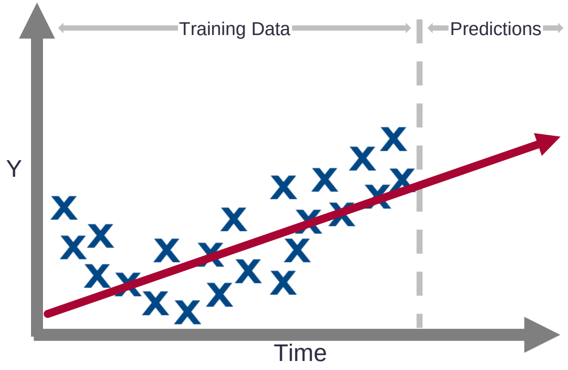
Strong Fit



Overfitting (High Variance)



Underfitting (High Bias)



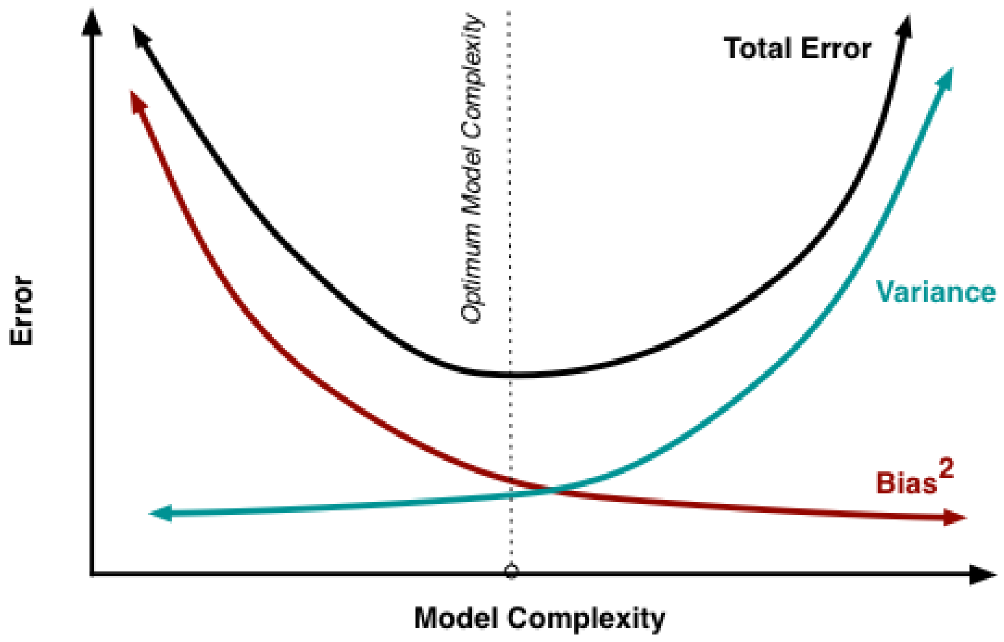
Principle #2: No Free Lunch

Unfortunately, there is no ‘magical’ algorithm that will solve all of our problems. Generating accurate predictions requires a thorough understanding of the underlying behaviors at play within our data

For a given problem, pick the right algorithms…

Supervised		Semi-Supervised
Regression	Classification	Clustering
Linear Regression	Logistic Regression	K-Nearest Neighbors
Multivariate Linear Reg.	Multinomial Logistic Reg.	HCA
Random Forests		PCA
Gradient Boosted Machines		LLE
Support Vector Machines		t-SNE
Multi-Layer Neural Networks		LDA
Recurrent Neural Networks		DBSCAN
Convolutional Neural Networks		Autoencoders
…		….

… to optimize the bias-variance trade-off

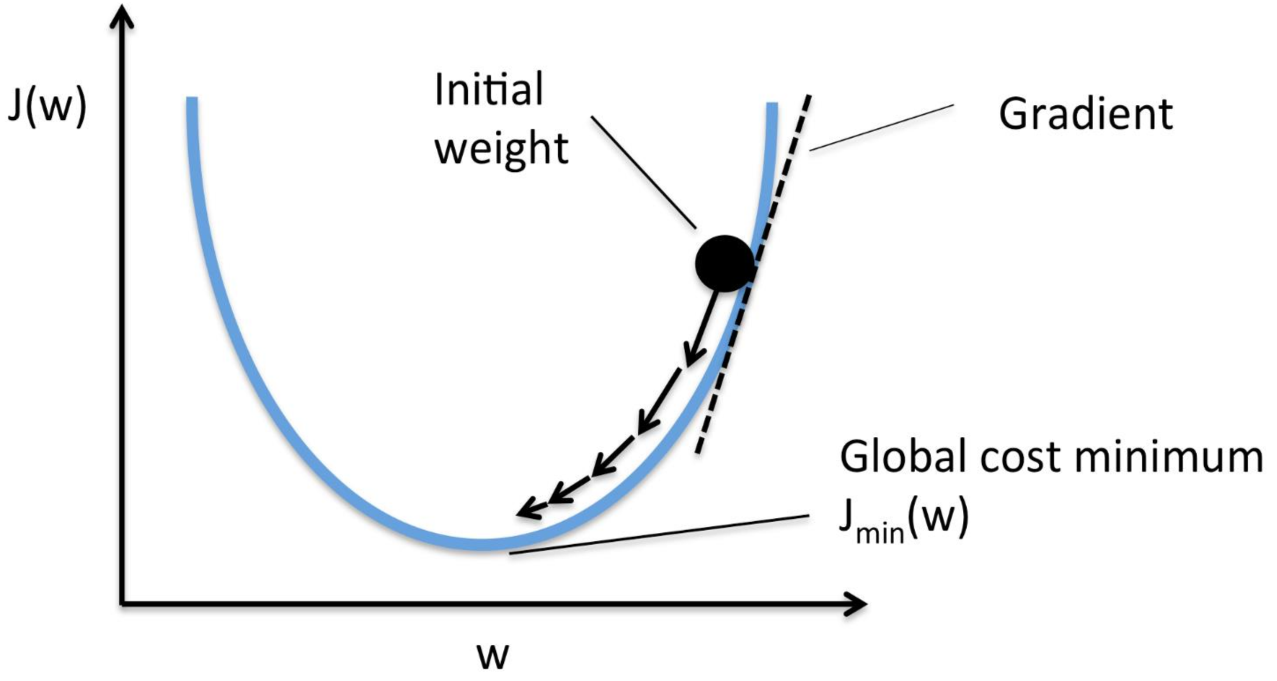


Graphic Source: Scott Fortmann-Roe

Principle #3: Occam's Razor

When two algorithms present similar results, there are many reasons why we should prefer the simpler of the two

Gradient Descent 101



Advantages of Simplicity

1. **Significantly less costly to compute** due to their relatively simple cost functions, accelerating insight to action
2. **Less likely to encounter optimization issues** when working in lower-dimensional spaces
3. Final solutions are generally **easier to interpret**, **visualize**, and **understand**

Graphic Source: *[Sebastian Raschka]*

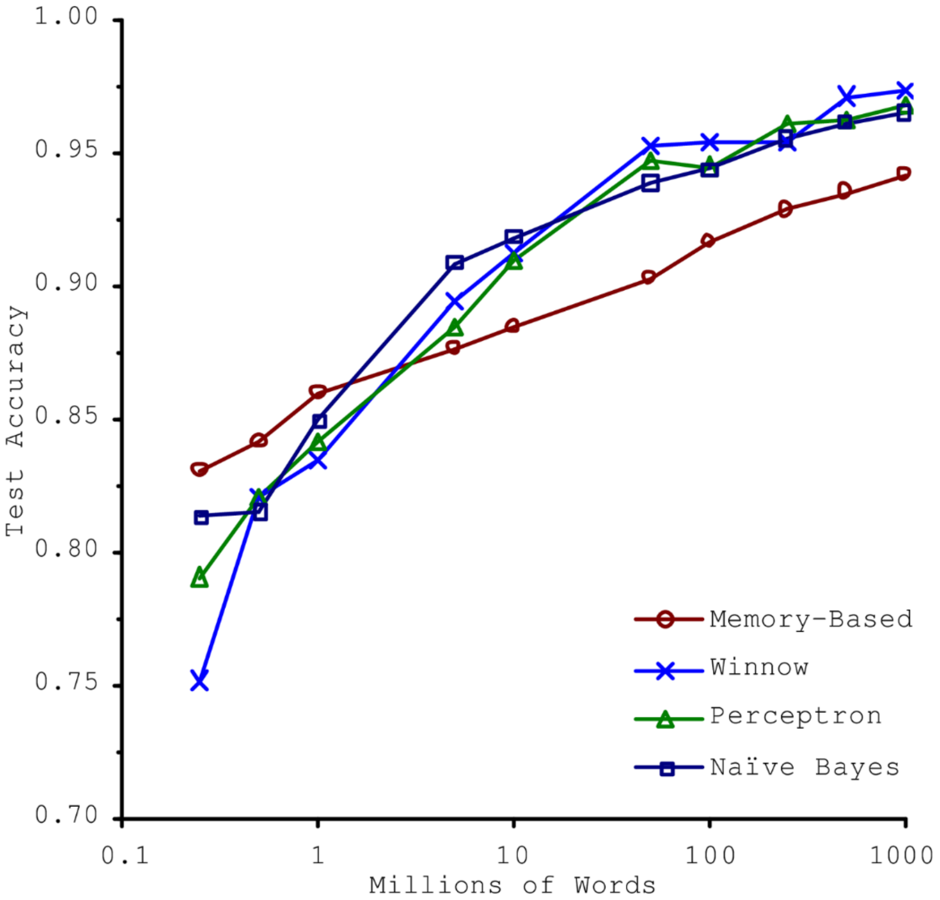
Principle #4: More Data > More Complex Algorithms

Using fancy algorithms is just one piece of the data science puzzle. Including new features or increasing the volume of data available for training will substantially improve your results.

*"**We don't have better algorithms** than anyone else; we just **have more data**"*

- Peter Norvig, Google

"The Unreasonable Effectiveness of Data"



Principle #5: Cross-Validation

In the same way that we cannot determine a drug's effectiveness by only testing it on a single patient, we need to examine our model using multiple data samples (called 'folds') to evaluate performance

Cross-Validation 101

k-fold Cross-Validation Error

Used to evaluate how well each model generalizes an independent data set

$$E = \frac{1}{10}\sum_{i=1}^{10} E_i$$

Graphic Source: [Karl Rosaen]

Principle #6: Algorithmic Diversity

Algorithmic diversity is key to predictive success; the combination of many simple models (“weak learners”) can outperform much more complex algorithms (“strong learners”)

Ensembling 101

1. Initial models are developed independently
2. Model predictions are compiled together using majority voting
3. Biases cancel out, giving us a much more accurate final model