

# AI and Machine Learning Introduction

Daniel Hardt

CBS

31 January 2022

**AI, Machine Learning, and Business**  
AI and Machine Learning  
Taking this Class  
More Powerful Models  
Takeaways

**AI and Machine Learning**  
Machine Learning and CBS

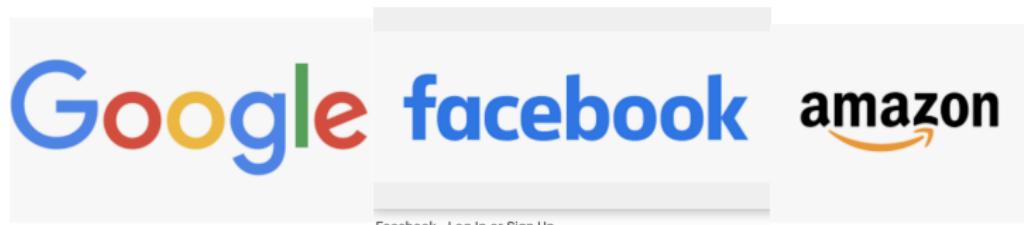
# Machine learning is the power behind AI

**AI is the power behind the  
most important companies  
in the world**

AI, Machine Learning, and Business  
AI and Machine Learning  
Taking this Class  
More Powerful Models  
Takeaways

AI and Machine Learning  
Machine Learning and CBS

# Big Tech: AI and ML

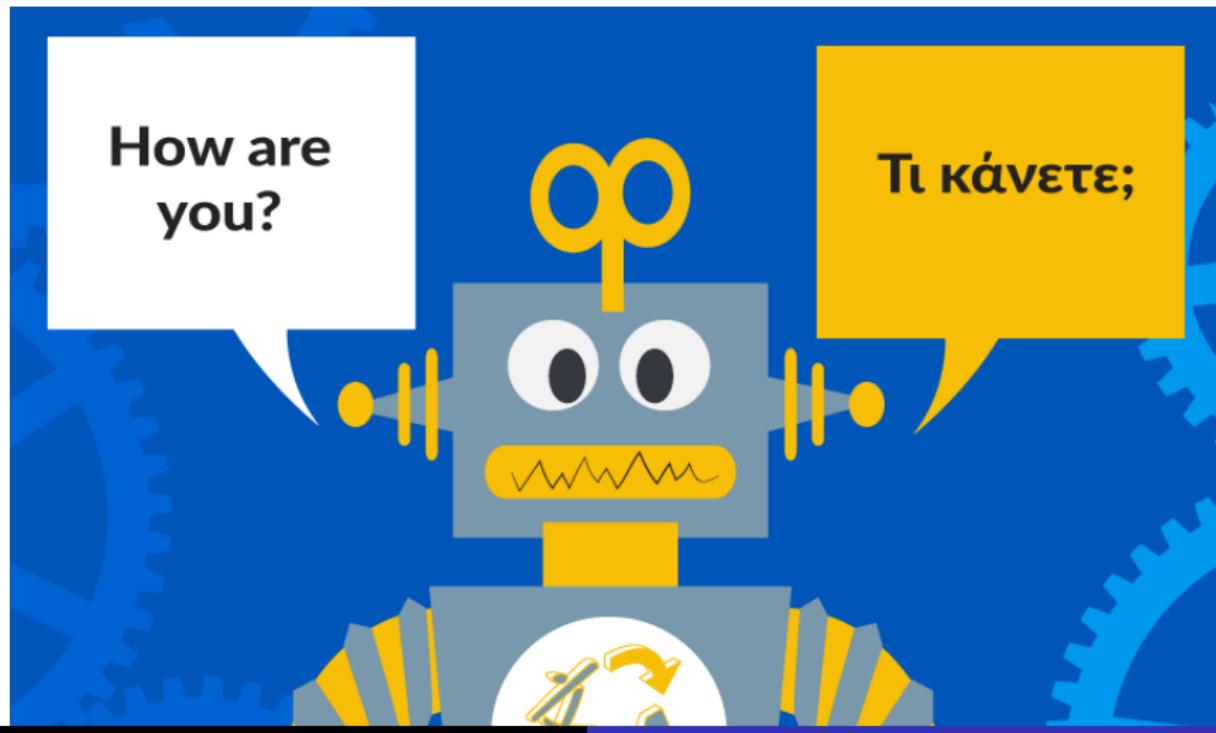


# Intelligent Search

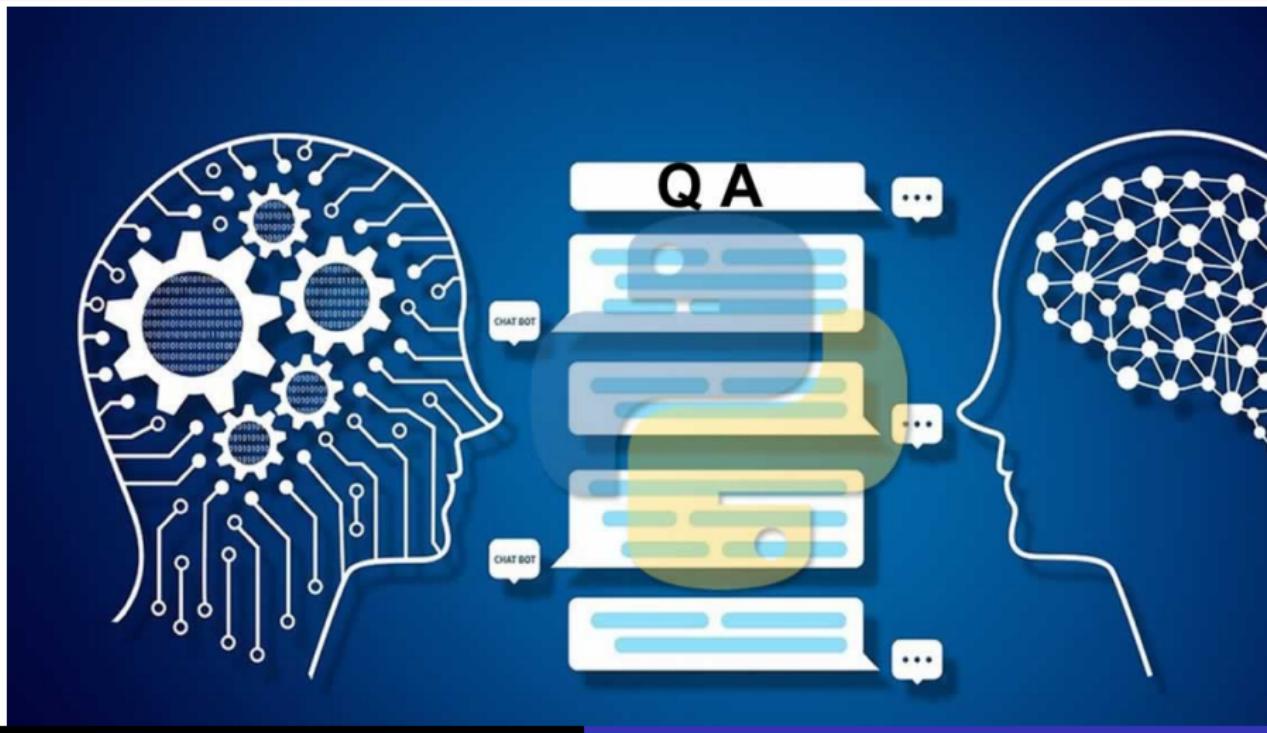
Google



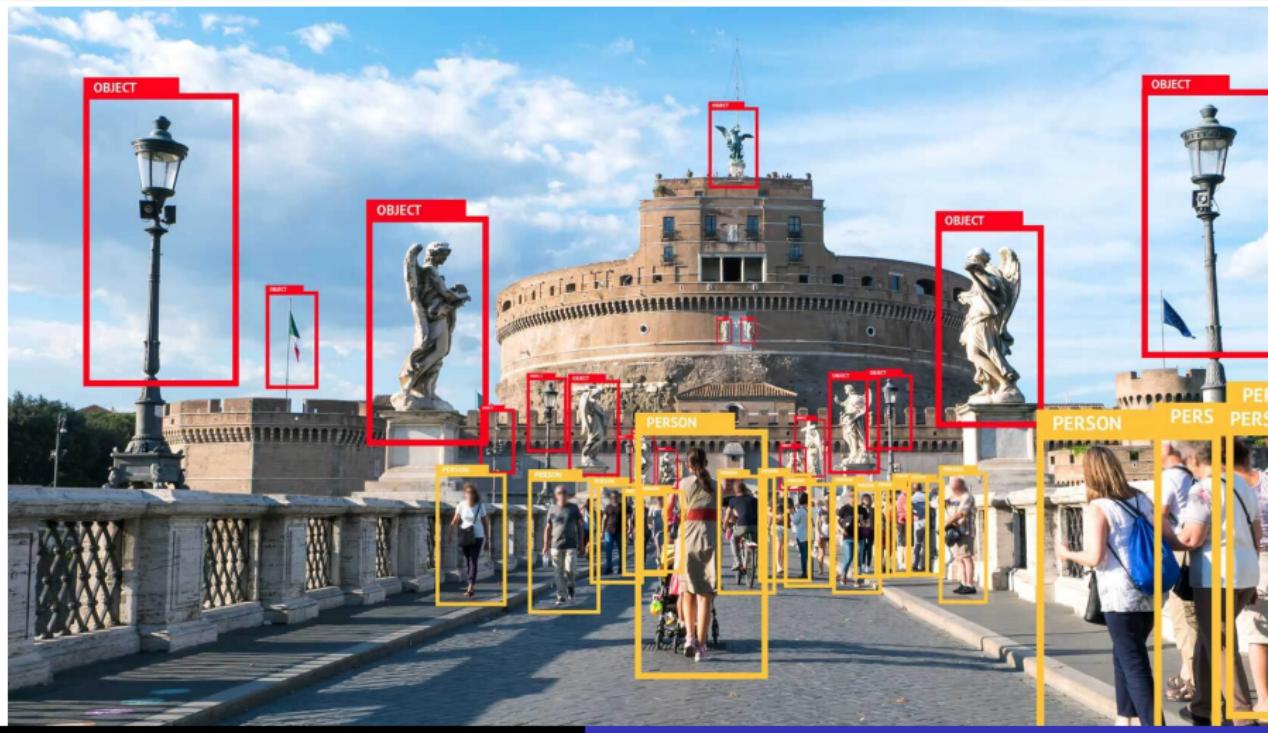
# Machine Translation



# Question Answering



# Image Recognition

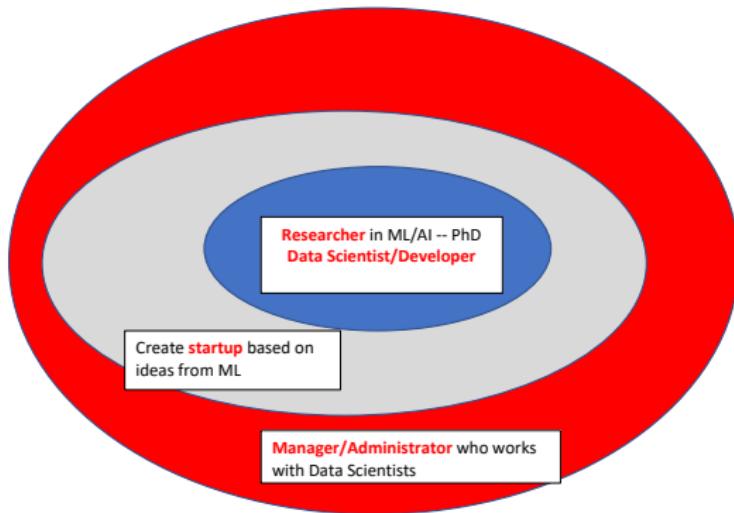


**AI, Machine Learning, and Business**  
AI and Machine Learning  
Taking this Class  
More Powerful Models  
Takeaways

**AI and Machine Learning**  
**Machine Learning and CBS**

# Why AI/Machine Learning at CBS?

# Why ML at CBS?



**Everyone:** understand how AI and ML will affect business, the economy, society and our common future.

AI, Machine Learning, and Business  
**AI and Machine Learning**  
Taking this Class  
More Powerful Models  
Takeaways

**Review: Basics of Machine Learning**  
Going further with Machine Learning  
Business and ML  
Class Project  
Collecting Data

This class builds on *Big Data Management*

AI, Machine Learning, and Business  
**AI and Machine Learning**  
Taking this Class  
More Powerful Models  
Takeaways

**Review: Basics of Machine Learning**  
Going further with Machine Learning  
Business and ML  
Class Project  
Collecting Data

# Review of Machine Learning Basics

# Classification and Regression Models

- Linear models for regression and classification
- Tree models for regression and classification
- Other models: kNN, Naive Bayes

# Basic Concepts

- Split training and test data
- Look at class distribution, compare results with baseline
- *Tune* models to reduce complexity, avoid overfitting

# More Key Concepts

- Evaluating Models:
  - Look at class distribution, compare results with baseline
  - Metrics: Accuracy, Precision, Recall
  - Cost of different types of errors
  - Expected Value
- Insight into Models and Domain:
  - Feature importance
  - Coefficients

AI, Machine Learning, and Business  
**AI and Machine Learning**  
Taking this Class  
More Powerful Models  
Takeaways

Review: Basics of Machine Learning  
**Going further with Machine Learning**  
Business and ML  
Class Project  
Collecting Data

# Going further with Machine Learning

# More Powerful Models

- Random forest and other *ensemble* models
- Multilayer Perceptron (MLP)/Neural Networks
- Reinforcement Learning

# Going More in Depth

- More detail about models
- Tuning: Systematic search for the best hyperparameter settings for a model
- Systematic ways to understanding the importance of model features
- Evaluating: exploring model metrics systematically

# ML Models and Business Value

- Expected Value Framework: what is the right metric for a given business problem?
- Classifier Thresholds: how “careful” should a classifier be, based on the business context?

# Pre-trained models: a new opportunity

- Pre-trained Models in Language and Vision create a new opportunity to build very powerful models with a small amount of training data
- **Language:**
  - BERT (Google)
  - GPT-3 (OpenAI)
- **Image Classification:**
  - VGG-16
  - ResNet

# Project

- Select/construct dataset
- Pose *interesting* questions, with societal/business relevance
- Produce a **research-style paper**
  - Detailed comparisons to relevant recent work
  - Propose ways to build on / improve recent work
  - Present new results and discuss in light of previous work

# The Secret of Machine Learning

- The most important thing is not
  - the model you use
  - or how much training data there is
  - or how you tune it
  - ...
- It's how **interesting** the dataset is!

## How to get Interesting Datasets

### Social Media:



AI, Machine Learning, and Business  
**AI and Machine Learning**  
Taking this Class  
More Powerful Models  
Takeaways

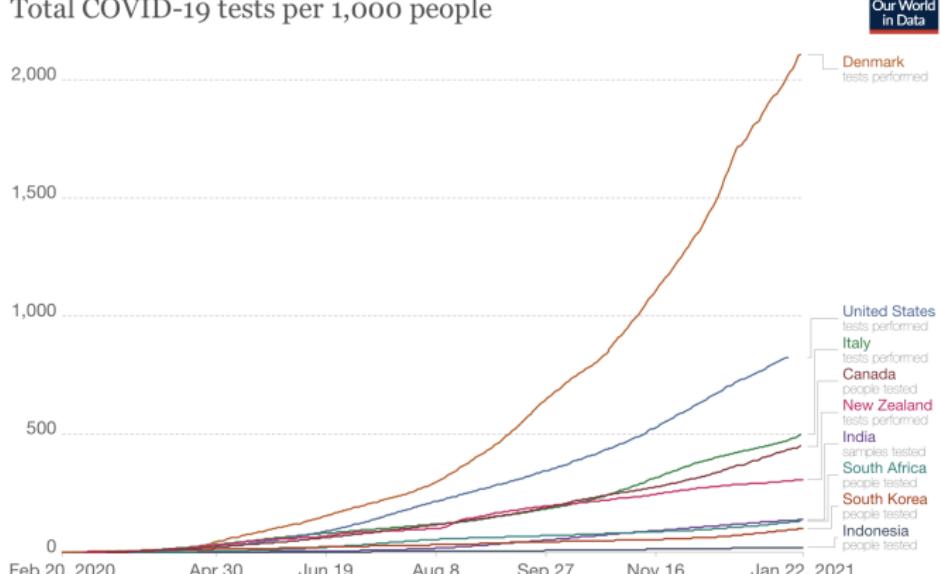
Review: Basics of Machine Learning  
Going further with Machine Learning  
Business and ML  
Class Project  
**Collecting Data**

## How to get Interesting Datasets

# Our World in Data

# Our World in Data

Total COVID-19 tests per 1,000 people



Source: Official sources collated by Our World in Data

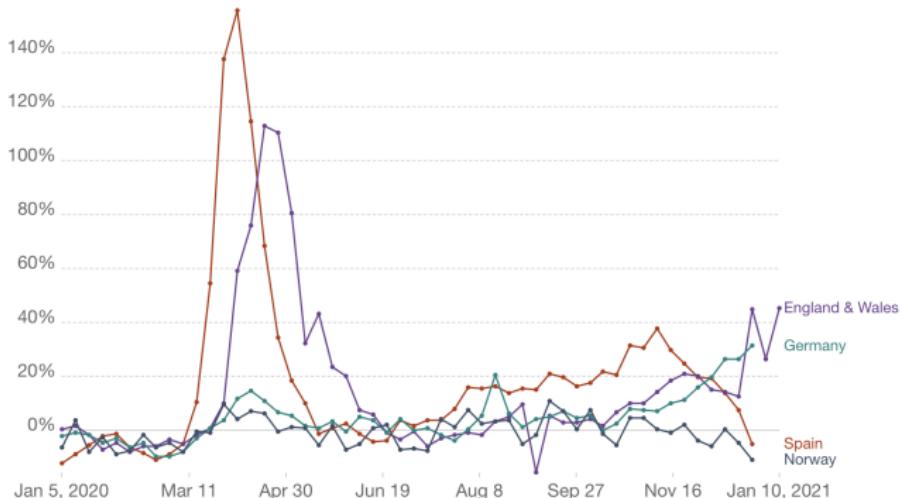
Note: Comparisons of testing data across countries are affected by differences in the way the data are reported. Details can be found at our Testing Dataset page.

# Our World in Data

Excess mortality during COVID-19: Deaths from all causes compared to previous years, all ages



Shown is how the number of weekly deaths in 2020–2021 differs as a percentage from the average number of deaths in the same week over the years 2015–2019. This metric is called the P-score. We do not show data from the most recent weeks because it is incomplete due to delays in death reporting.



Source: Human Mortality Database (2021), UK Office for National Statistics (2020)

[OurWorldInData.org/coronavirus](http://OurWorldInData.org/coronavirus) • CC BY

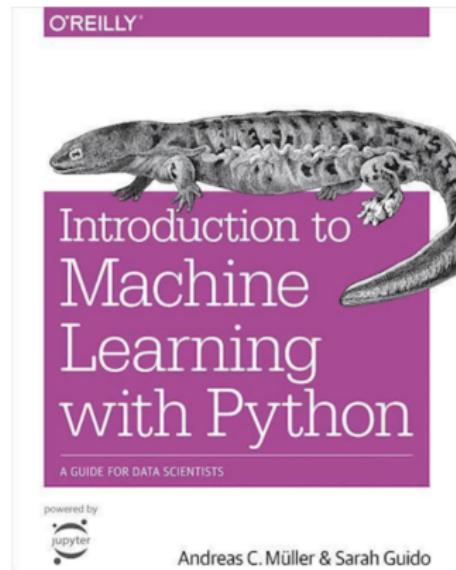
# Interesting Datasets

- Combining different datasets, to pose questions like:
  - How does sentiment on Twitter correspond to Covid-19 case levels?
  - How are different travel destinations described on Reddit?
  - Does weather affect the level of crime in different areas?
  - Does AirBnb data in different areas tell you about social conditions in different cities?
  - ...

# Weekly Sessions

- Readings
- Lecture
- Activities
  - Programming/building models
  - Finding and constructing interesting datasets
  - Developing project ideas
  - Submit results each week
- Feedback
  - Written feedback on submitted results
  - Feedback on project ideas and workplan
- Syllabus

# Main Reading



# Online Resources for Book

github.com

[amueller / introduction\\_to\\_ml\\_with\\_python](#)

Watch 317 ⭐ Star 4.2k Fork 2.6k

Code Issues 13 Pull requests 1 Actions Projects 0 Wiki Security Insights

Notebooks and code for the book "Introduction to Machine Learning with Python"

88 commits 1 branch 0 packages 0 releases 10 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

amueller	explicitly use liblinear solver for l1 penalty in logistic regression	Latest commit 62a9b3f on Sep 18, 2019
	data	add adult dataset, fix path.
	images	also add api table image
	mlearn	don't use joblib from externals
	.gitignore	add gitignore
	01-introduction.ipynb	update notebooks for new print / sklearn 0.20 / new spacy etc
	02-supervised-learning.ipynb	explicitly use liblinear solver for l1 penalty in logistic regression
	03-unsupervised-learning.ipynb	fix ylabel in tsne plot
	04-representing-data-feature-engi...	fix time series printing (again)

AI, Machine Learning, and Business  
AI and Machine Learning  
**Taking this Class**  
More Powerful Models  
Takeaways

# Final Project

# Interesting Dataset

# Interesting Questions!

# Use ML techniques studied in class

# Best practices for training, tuning and evaluating models

# Take advantage of large pre-trained models

Systematically explore  
different model evaluation  
metrics to connect to  
business value

Write research-style paper,  
where you connect with  
current research

# Get Feedback!

- Weekly lecture and discussion
- Weekly office hours
- Instructors
- Group members
- Classmates

# Random Forest

- Build many decision trees
- Each tree differs in random ways
  - Select different **data points** used to build tree
  - Select different **features** in split tests

## Data for Random Forest

- For each tree, create **bootstrap** sample
- Randomly select  $n$  items from orginal dataset, allowing repetitions
- Each tree will have same size dataset, but randomly different, because of repetitions

# Random Feature Subsets for Random Forest

- Parameter **max\_features**
- Select random subset of features of size max\_features
- If max\_features is high, more chance of overfitting

# Parameters for Random Forest

- number of trees
- max\_features

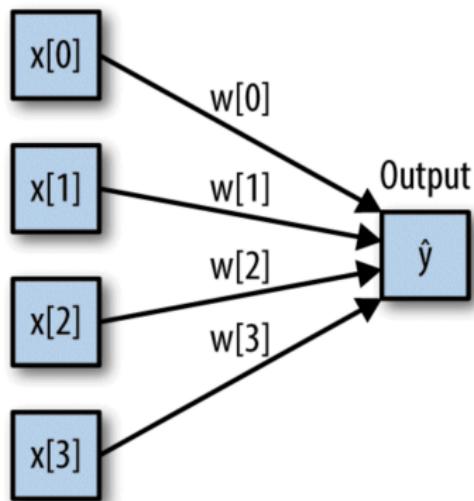
# Linear Models Again

- $\hat{y} = w[0] * x[0] + w[1] * x[1] + \dots$

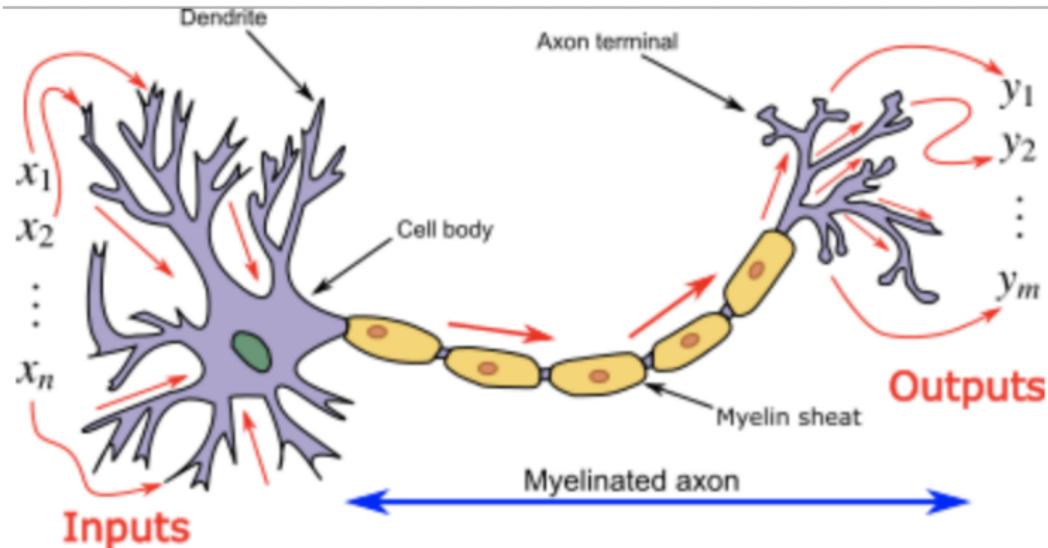
A weighted sum of inputs

# The Perceptron

Inputs



# Neurons in the Brain



# Multilayer Perceptron/Neural Network

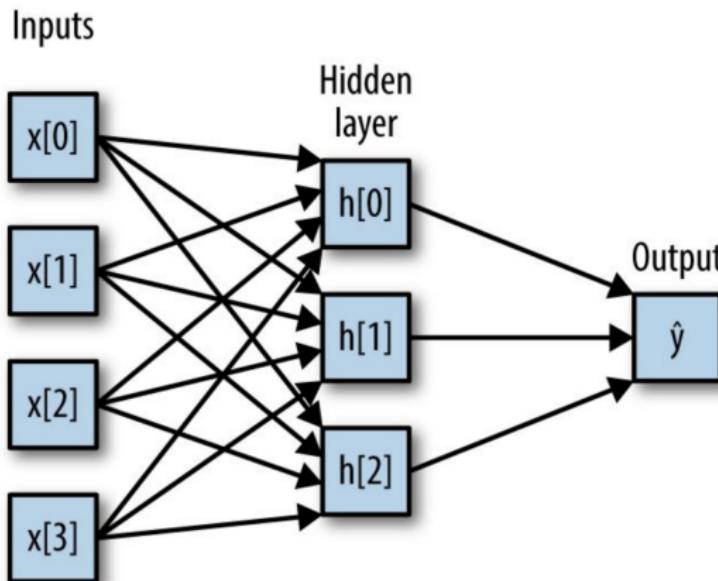
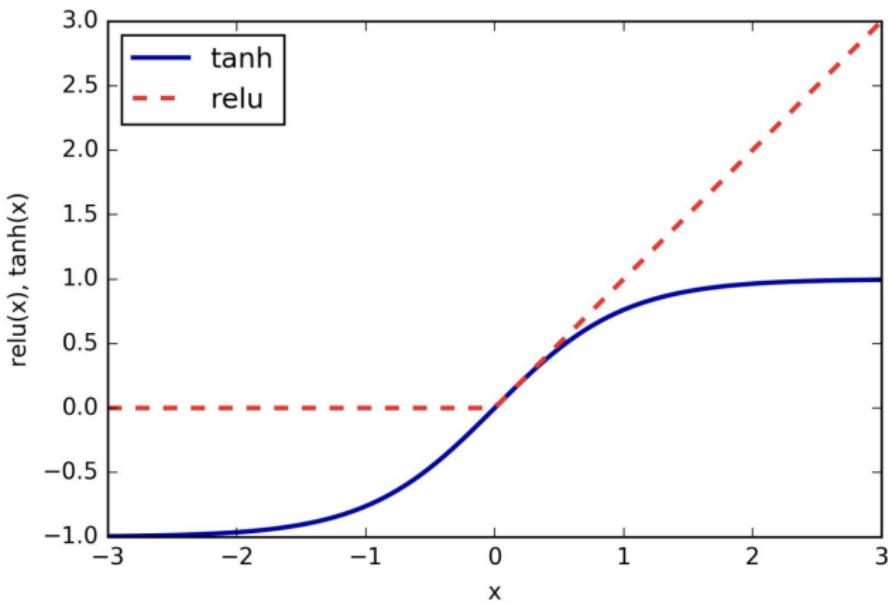


Figure 2-45. Illustration of a multilayer perceptron with a single hidden layer

# Activation Functions



**Figure 2-46.** The hyperbolic tangent activation function and the rectified linear activation function

# Neural Network – Equation

- $h[0] = \tanh(w[0, 0] * x[0] + w[1, 0] * x[1] + w[2, 0] * x[2] + w[3, 0] * x[3] + b[0])$
- $h[1] = \tanh(w[0, 1] * x[0] + w[1, 1] * x[1] + w[2, 1] * x[2] + w[3, 1] * x[3] + b[1])$
- $h[2] = \tanh(w[0, 2] * x[0] + w[1, 2] * x[1] + w[2, 2] * x[2] + w[3, 2] * x[3] + b[2])$
- $\hat{y} = v[0] * h[0] + v[1] * h[1] + v[2] * h[2] + b$

# Tuning Neural Networks

- Many parameters to adjust
  - Number of hidden layers
  - Number of units in each hidden layer
  - Regularization
- Scaling of inputs is important

# Takeaways

- Building on basics of ML from Big Data Management
- More powerful models
- Explore different ways to evaluate models – Expected Value, thresholds
- Large pre-trained models for language and vision
- Final project – research style paper