AI and Machine Learning
Introduction

Daniel Hardt

CBS

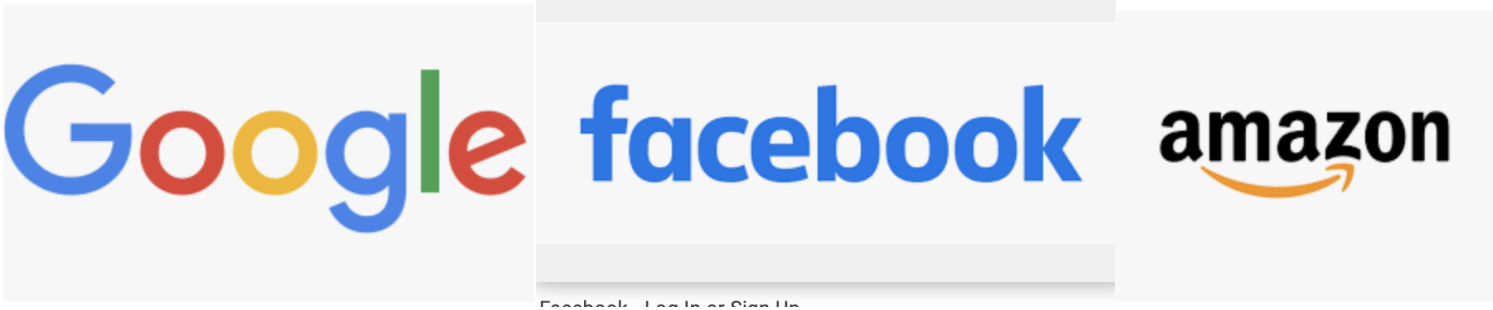
31 January 2022

Copenhagen
Business School
HANDELSHØJSKOLEN

Machine learning is the power behind AI

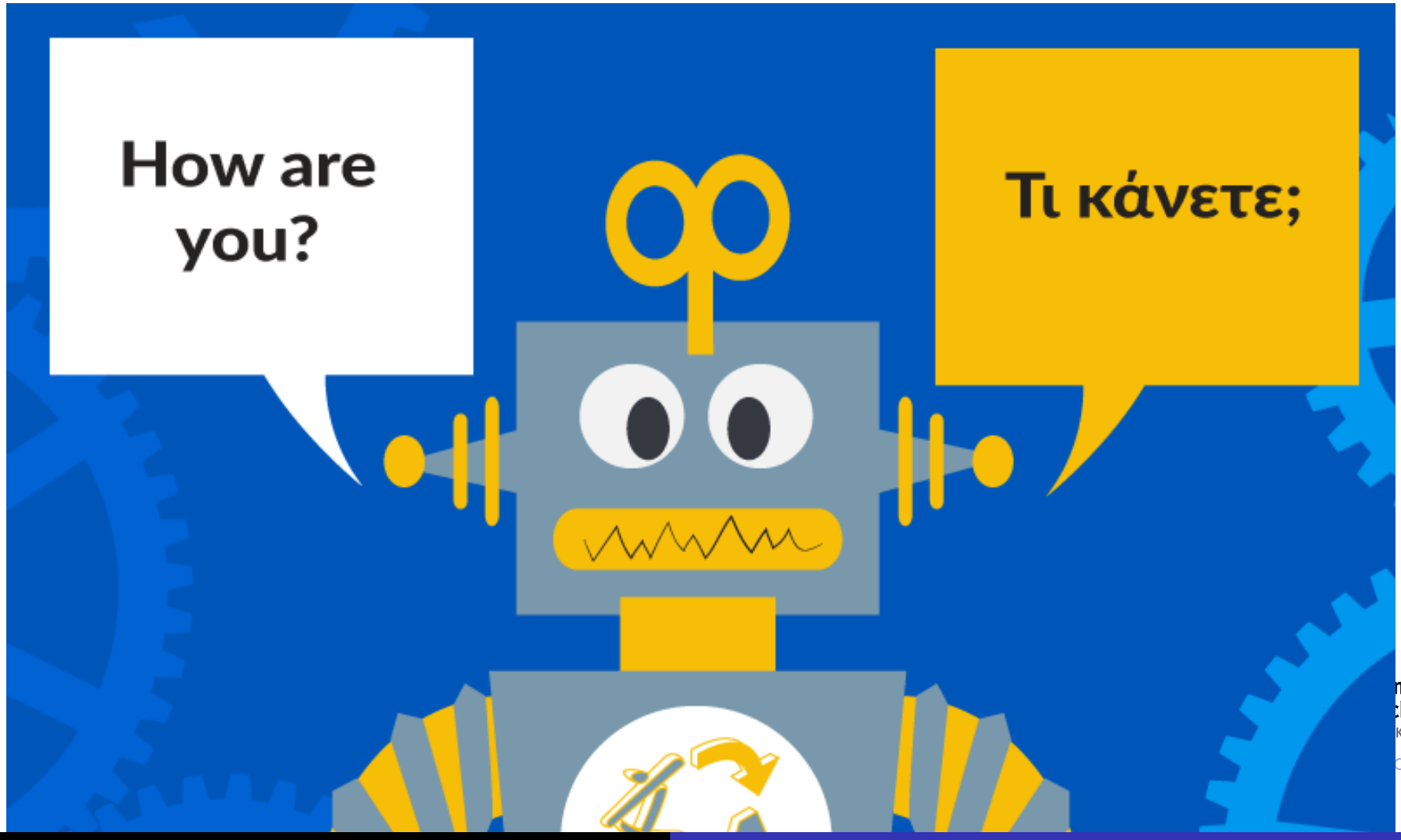
AI is the power behind the most important companies in the world

Big Tech: AI and ML



Intelligent Search

Machine Translation



Question Answering

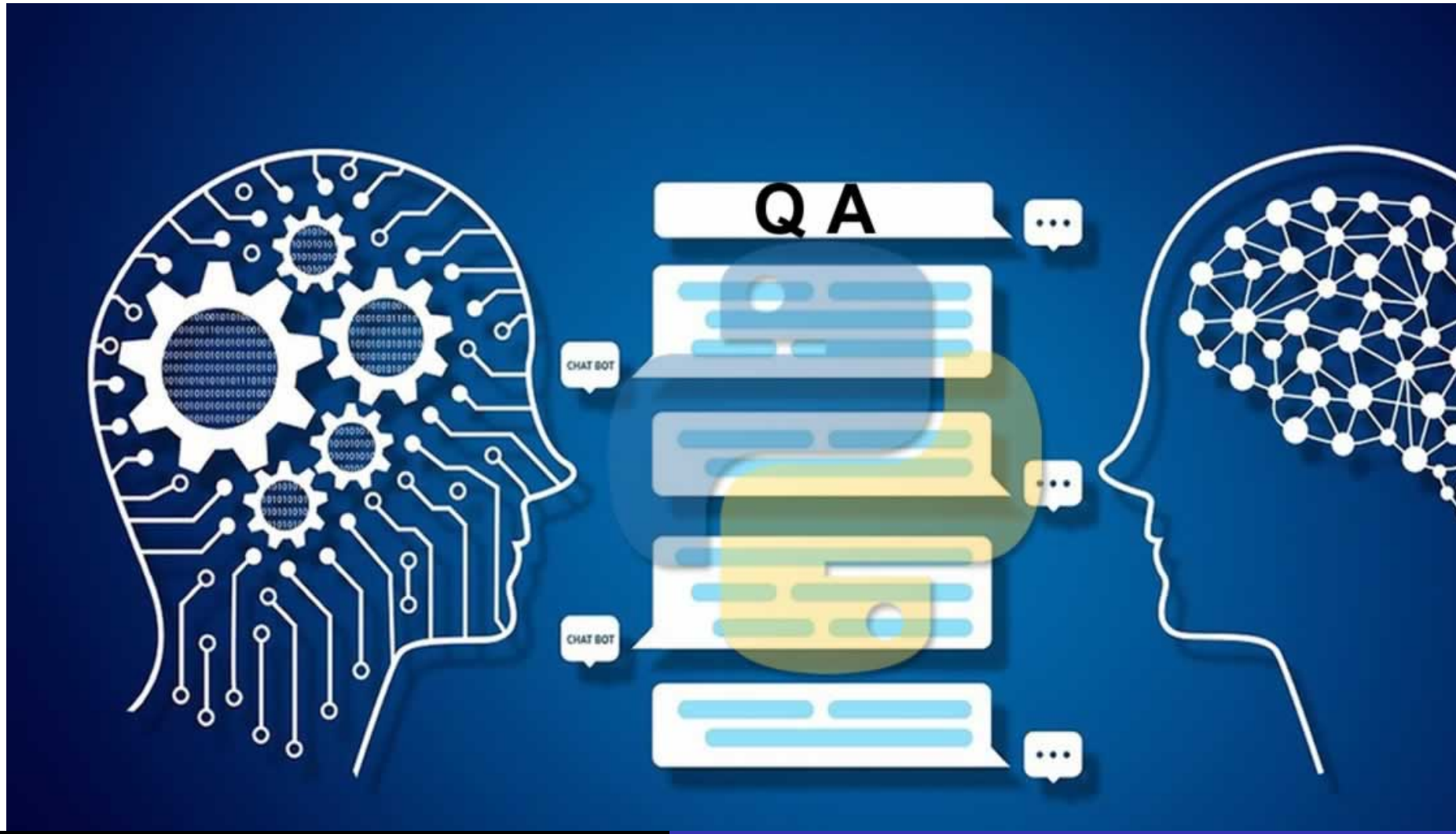
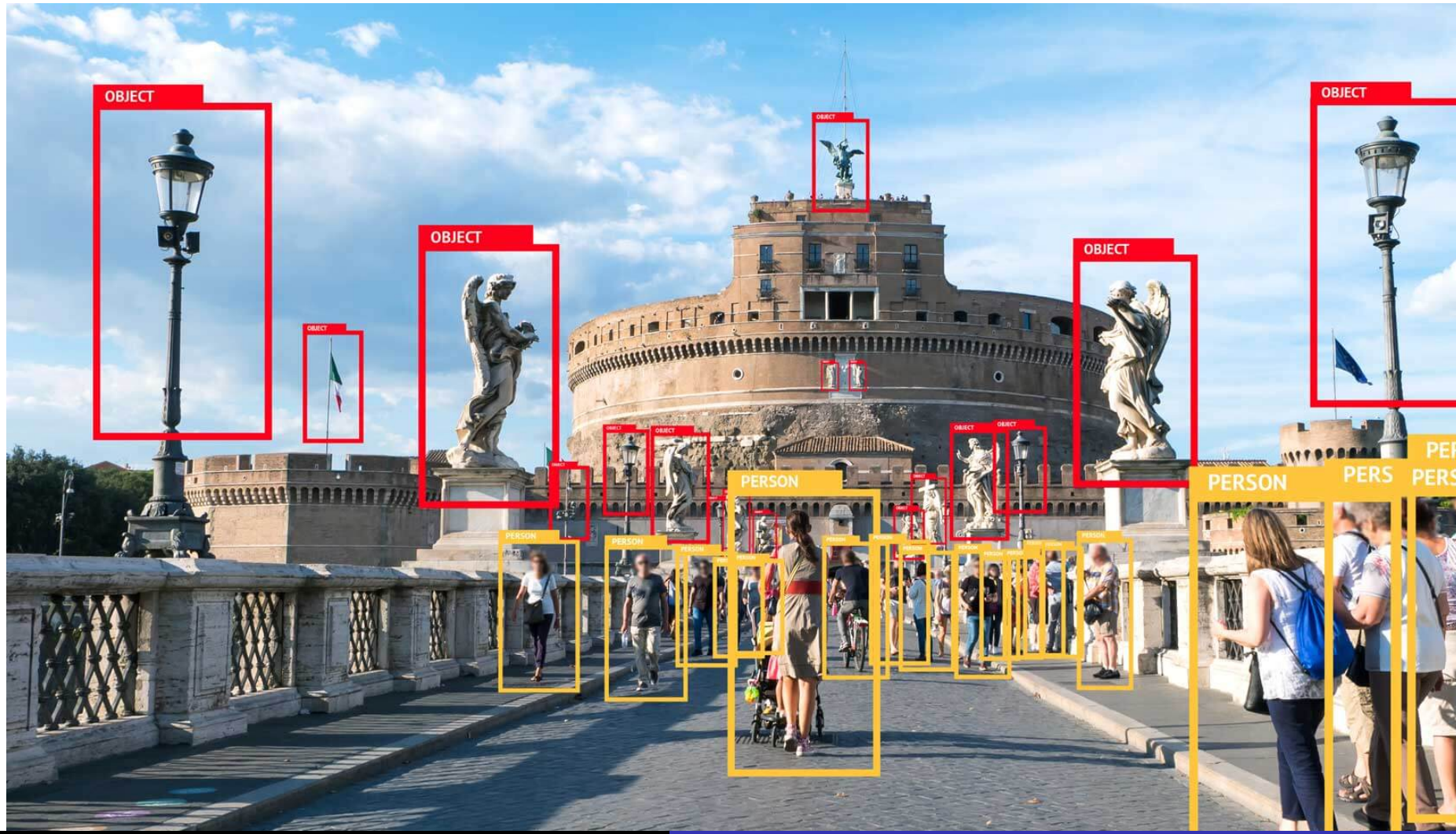
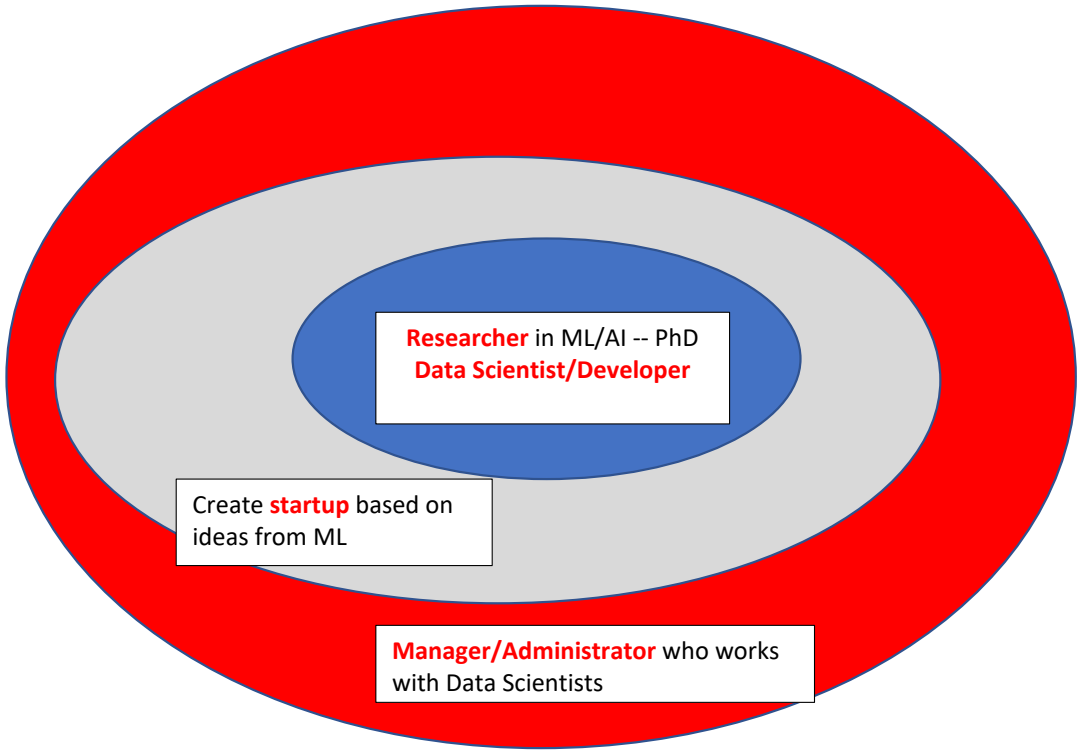


Image Recognition



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.

This class builds on *Big Data Management*

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

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!

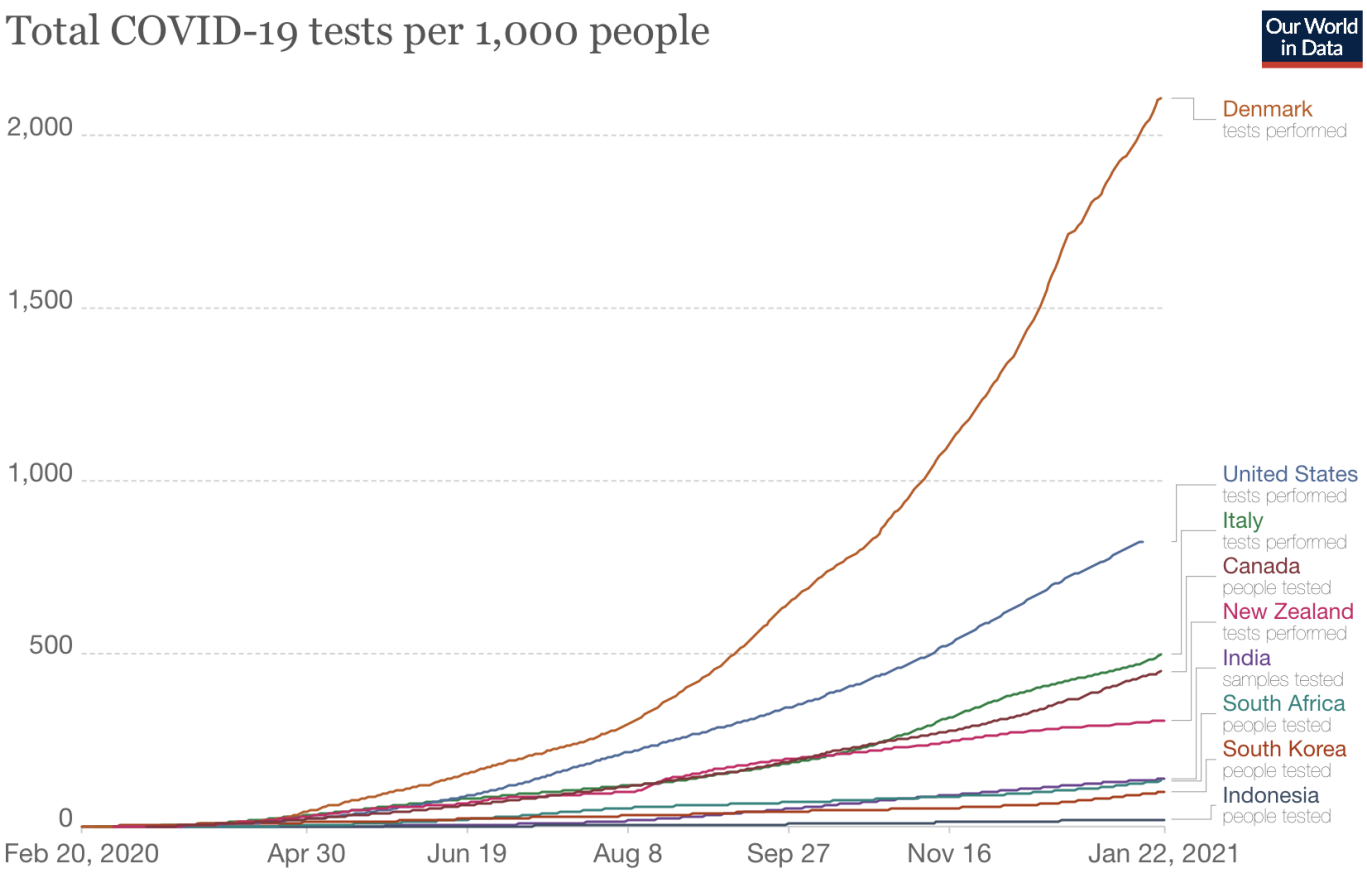
How to get Interesting Datasets

Social Media:

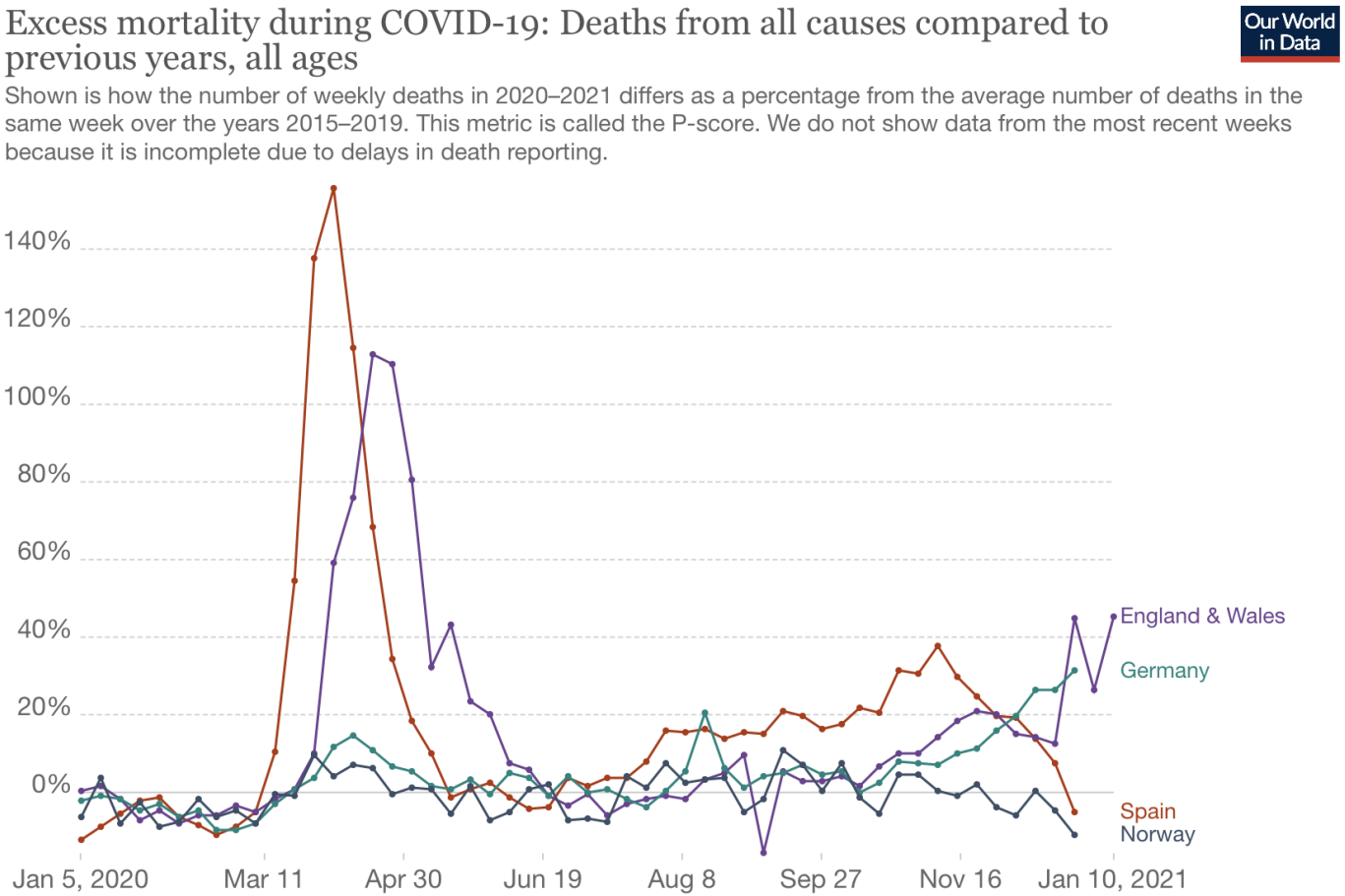
How to get Interesting Datasets

Our World in Data

Our World in Data



Our World in Data



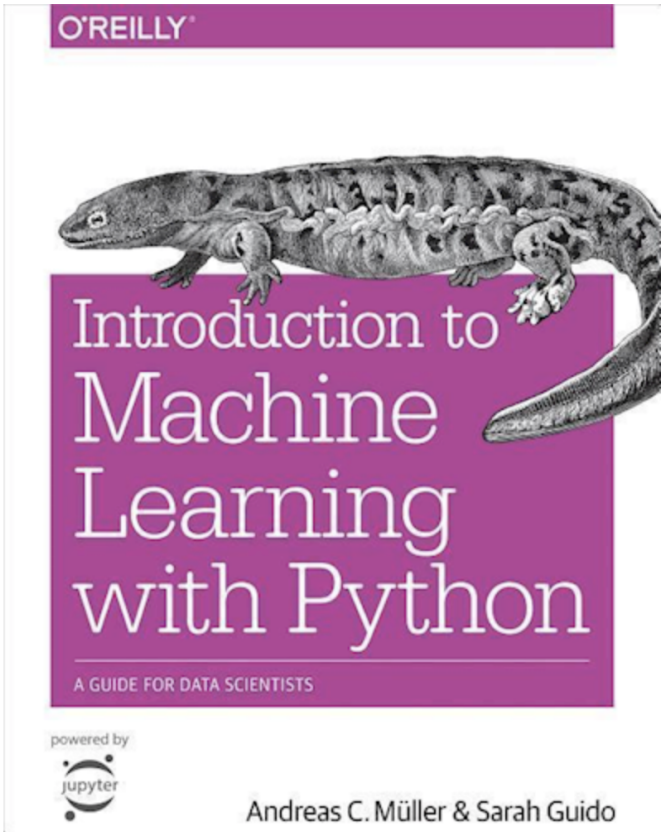
Interesting Datasets

- 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?
  - ...

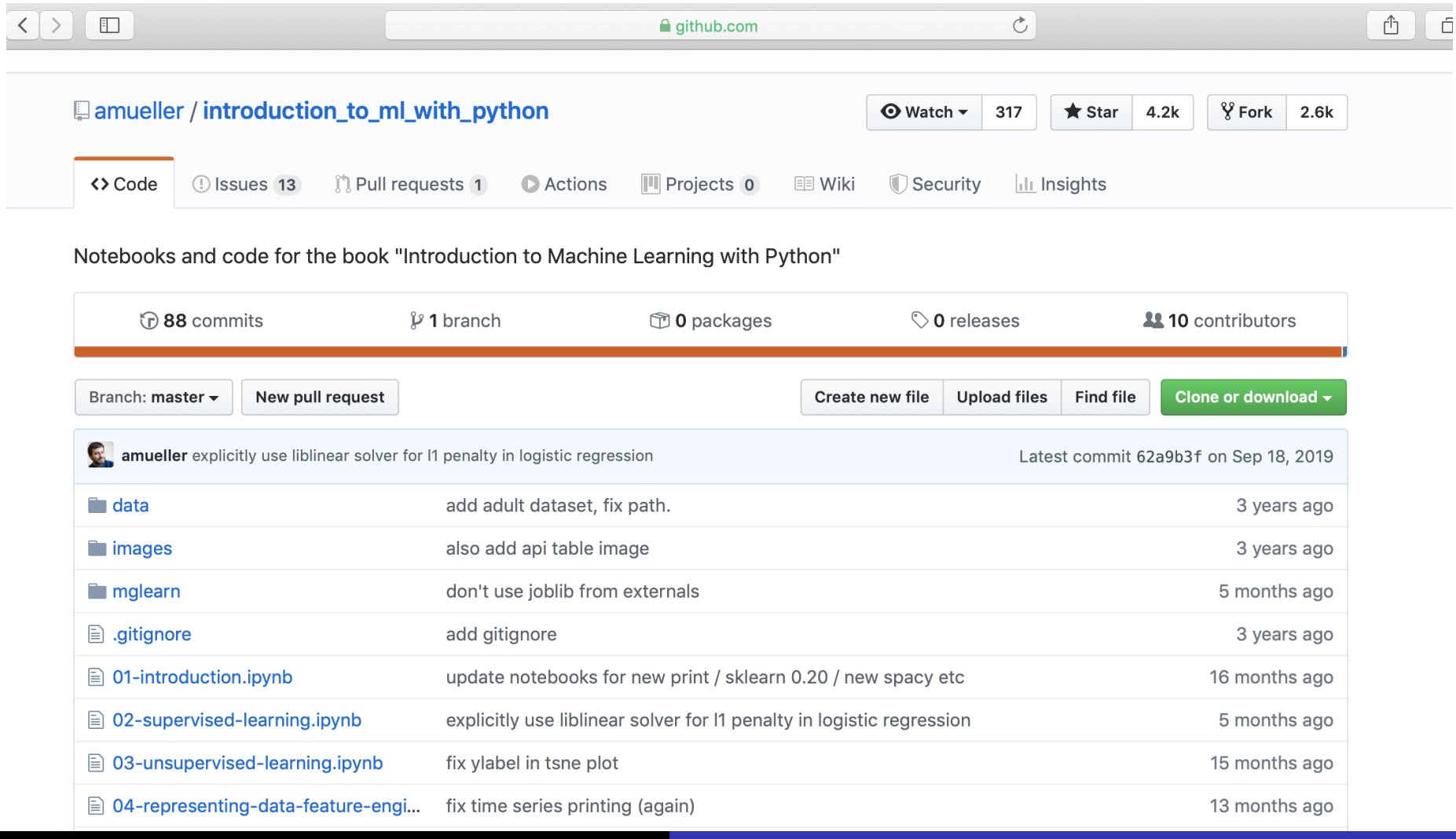
Weekly Sessions

- 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



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

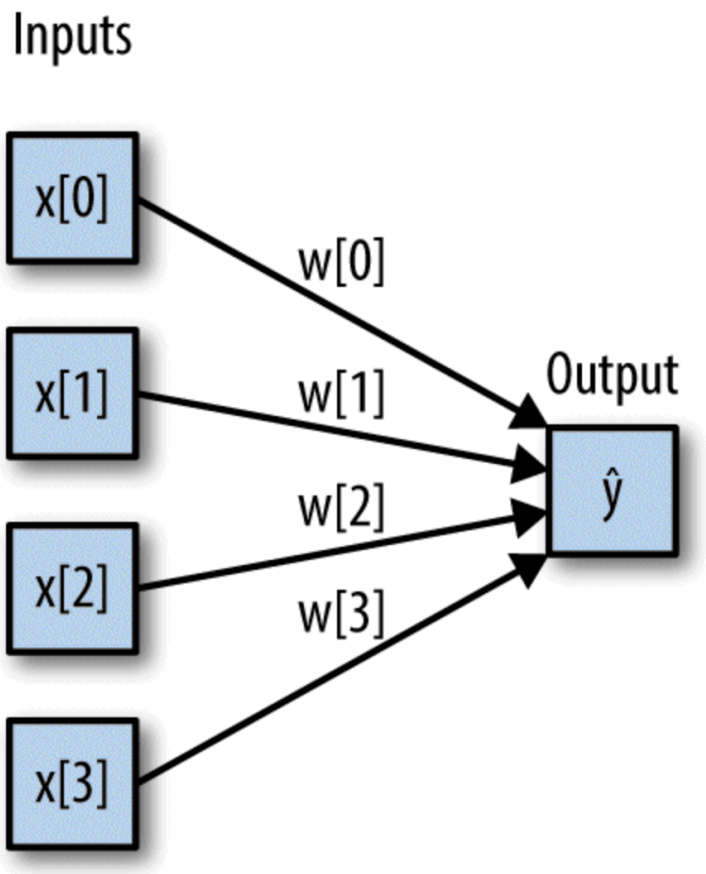
- number of trees
- max_features

Linear Models Again

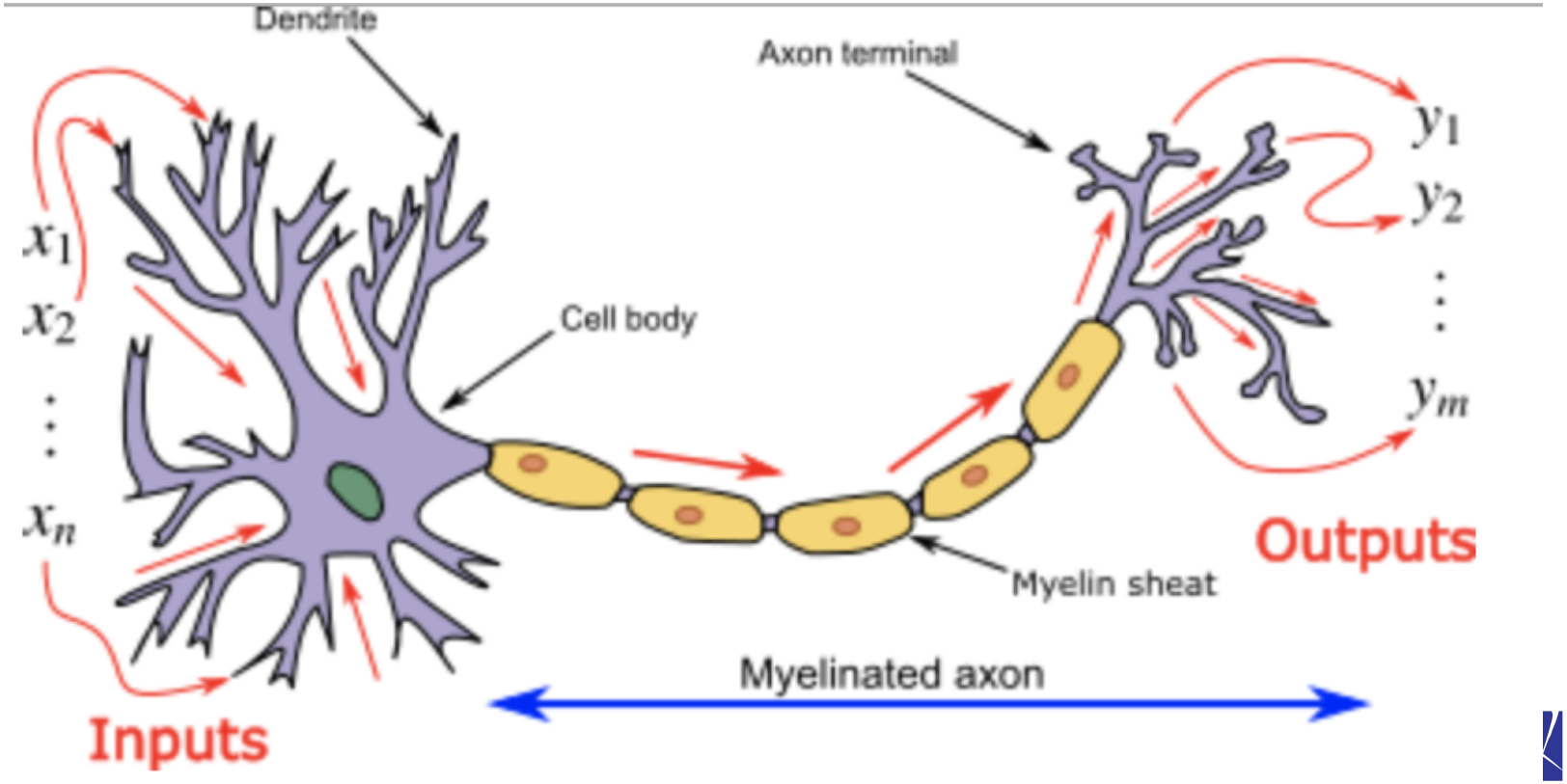
- $\hat{y} = w[0] * x[0] + w[1] * x[1] + \ldots$

A weighted sum of inputs

The Perceptron



Neurons in the Brain



Multilayer Perceptron/Neural Network

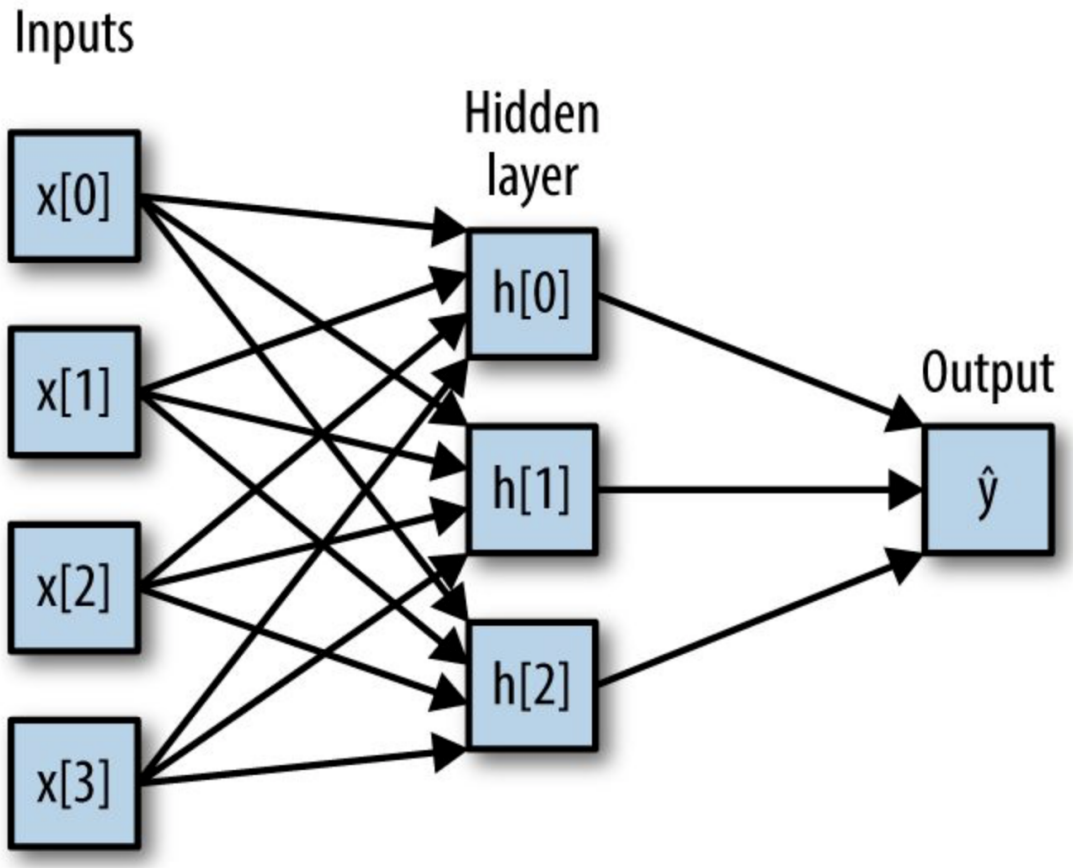


Figure 2-45. Illustration of a multilayer perceptron with a single hidden layer

Activation Functions

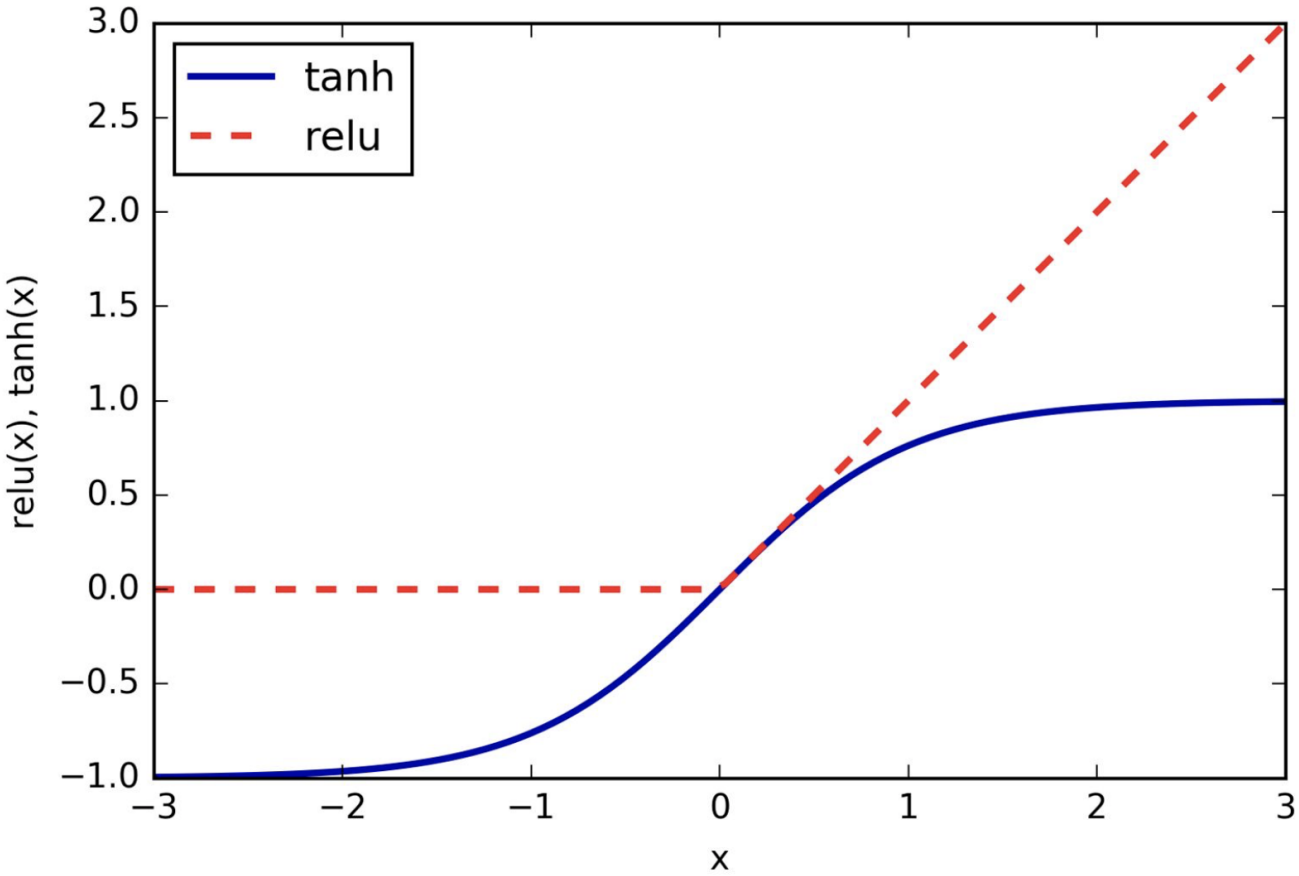


Figure 2-46. The hyperbolic tangent activation function and the rectified linear activation function

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