Slide 3: Introduction

Why is big data emerging now? Because now we have the technology to digitize all existing knowledge, whereas this thing did not exist earlier. Information is coming from all different kinds of sources and it's a byproduct of our increasing use of technology. In 2020, data volume is expected to be 40 zettabytes. In comparison, 40 zettabytes is the amount of grains of sand on earth * 75. There's so much data processed that in the last 2 years, that we've processed more data than in the past 3000 years.

Slide 4: Analogies

Here are some analogies associated with big data. The data collected by Copernicus (astronomical data), was used to map the skies. The microscope opened up the invisible world. The electron microscope opened up the atomic world. Finally, big data acts as a microscope for the super visible world. Within the sea of data that we're collecting, we use algorithms to parse through this data.

Slide 5 Emerging Big Data

Previous, whenever we learn something, we'd write it down, then it becomes knowledge. Now, The knowledge is inside a pile of information,

and we need to extract it. As we become more digitized, everything becomes a data point. There’s so much data collected that it can’t be normally used. It’s up to us to parse out data, extrapolate it, in order to contextualizes it and give meaning to it

Slide 6 Why is Big Data Happening Now?

There are 4 major reasons why the big data phenomenon is occurring. First, the development of the internet have allows data to be transferred instantaneously and to anywhere in the world. The internet has allowed collaboration, and acts as a medium for data to be transported. The second reason is the ubiquity of small scale devices. Sensors are becoming cheaper, more efficient, and everywhere. The availability of these sensors allows for a vast sum of data to be collected. Three, the accelerating data storage capacity and computing power cost. Moore's law, cloud computing, and distributed platforms have allowed data to be stored and processed. And four, big data has increased the popularity of machine learning. In turn, these machine learning models requires more data to compute.

Slide 7 Data Case: Social Media

This is a case study for how/what social media data could be collected and who uses it. In social media, you could collect data about interests, likes, posts, messages, friends, photographs, relationships, status. birthday, work and educational history...etc. There could be a lot of people that could use this data. For example, Facebook use it's data to build a dating feature within their application. Advertisements could use it to sell products related to your interests. 3rd parties like Cambridge Analytica could use social media data to spread targeted political campaigns.

Slide 8 Data Case: Vending Machine

This is a case study for how/what vending machine data could be collected and who uses it. Vending machine data could be collected in various ways. For example, what products sell the most, the payment information about people (if they use electronic payment), time/days when purchases peak or fade, locations of vending machines where people buy these products, locations of products inside the machines which were picked up. People who might be interested in this data includes producers deciding where to place physical stores/distribution centers, etc. so as to maximize their profits. Designers could also use this data to sell a certain product.

Slide 9 Case Study: Flu Prediction

The case study with Google's flu prediction model occured because of big data. Initially, doctors would have to report whether their patients have the flu, then the result would be compiled throughout the country. That data is used to determine if there's a flu outbreak. This process was slow and inefficient, typically takes around 2 weeks to complete. Google realized that people would search for flu related keywords if they're sick. So they tracked and modeled who/where flu related searches are coming from. From this large source of data, Google is then able to predict where a flu outbreak had occurred with high accuracy. However, one issue with this approach is the influence of outside factors. For example, news reports are talking about the flu, then searches for the flu might increase.

Slide 10 Unique Case Studies

Big data has allowed us to uncover fairness issues. Big data has been used for disaster response in Haiti. Big data has been used in uprisings in Tunisia. Big Target patterns can tell more about ourselves than us. In a case study from Target, they were analyzing buying patterns from a customers predicted a pregnancy before she told her family.

Slide 11 Concerns with Big Data

Anything that's going to change the world, by definition should have the ability to change it for the worse. There are multiple issues with the data revolution. For example, if somebody is listening in. like the big brother effect, or if a device collects more information than intended. There are huge implications for how we interact with machines.

Deep Learning

Slide 13: Deep learning Classification

What is deep learning? Deep learning is a class of machine learning algorithms. First, we need to classify the problem in deep learning. There are a few ways to classify deep learning. Based on the logic, we can create a decision tree to help us to represent the algorithm in deep learning. It will provide an easier way for us to understand the basic logic. Then, there is the random forest classifier, just like the name " forest," which combined with lots of individual decision trees that operate together. Since there are many different trees work together, there is usually have a lower chance of having individual errors. A neural network is a set of algorithms, that inspired by the human neural system, and designed to recognize patterns.

Neural network help us classify the deep learning better. It gives us an easier way to manage and store the data.

Slide 14: Initial Problems

A multi-layer perceptron is a neural network with only fully connected layers. And is one kind of neural network architecture. The XOR is the problem if using a neural network to predict the outputs based on the XOR logic. The situation with the XOR problem is that regardless of what lines you draw, there'll never be a separation between the positives and negatives. If deep learning classifiers aren't able to identify such a basic grouping, how are they able to classify complex shapes?

Slide 15: Activation Functions

In neural networks, the activation functions are used to get the output of the node. We mainly talk about the non-linear activation functions here, since it is the most used activation functions. It makes it easy for the model to generalize a variety of data and to differentiate between different outputs.

Slide 16: Perceptron

Perceptron is a single layer neural network. It helps to classify the given input data. Perceptron consists of 4 parts: 1, input values; 2, weights and bias; 3, net sum; 4, activation function. We can consider perceptrons as a small piece of neural network, since they work the same way.

Slide 17: Bias variance trade-off

Bias is the difference between the average prediction of our model and the value that we are trying to predict. High bias usually mean high error on training and test data. Variance is the variability of model prediction for a given data point. High variance usually means a pretty good performance on training data, but poor results in data testing. For the model, the process to find the balance between bias and variance is called" bias-variance trade-off'.

Slide 18: Gradient Descent

Gradient descent is an optimization algorithm used to find the values of parameters of functions in neural networks. In the 3-dimensional graph we covered in class, we moved from the highest part of the graph to the bottom part. The line represents the steepest descent from the graph, and it decreases the cost function the quickest way. Back propagation is used to adjust each wright in the network in percentage to overall error, and we can use it the reduced each wright’s error.