**CS311 Yoshii - Week 14B (Notes-14B) Big Data Analysis and Machine Learning and Beyond…**

**Big Data Analysis**

**Big Data** analysis is one of the most popular topics in computer science today. Once a vast amount of data is collected, how do you analyze the data quickly and accurately?

Big Data analysis is important in being able to **prediction and prevention.** Once we have the true Internet of Things (perhaps by2045), we will have a vast amount of data on everybody and everything. **Then, we will be able to predict many things before they happen!**

**Data Mining** == A fast computer will be used to discover interesting relations among things (i.e. find patterns). **We will not even tell the computer what to discover.** The computer will keep on finding patterns among things you never thought about before. *Research will change from verifying your theories to discovering brand new theories.*

How? **Heuristic** (does not work perfectly) algorithms are being developed in Artificial Intelligence to discover patterns in a vast amount of data. For example, by plotting data on a graph and finding clusters.

There are other method involving **statistics and probabilities (see CS571 notes).**

**Again, math is important in Computer Science**.

***How to Use Big Data for Unexpected Discovery among Things***

In one method, collected data is expressed as points on a multi-dimensional graph. And when **points** **form a cluster**, they are related.

**A simplest example**, embed sensors on everyone’s pens to collect all kinds of

data on everyone (e.g. time spent writing, time spent walking, holding pressure, etc.)

Plot them on a graph as follows:

Y axis - time writing

X axis – time walking/moving

Z axis - holding pressure

Etc.

Each person is a point on the graph. Do you see clusters?

For example, discover that people with high X and Y have low Z.

Not very useful in this case.

**We don’t have to wait for the IoT.**

The data used does not have to come from sensors. **One could go through all kinds**

**of existing databases** to plot all kinds of points and **find unexpected clusters**. For example, use

databases on cities and plot each city’s crime rate, amount of greens, number of coffee shops, etc.

**Maybe you will discover that the number of coffee shops is related to the crime rate.**

**End Results of Unexpected Discovery of Relations:**

* **Prediction: Predict something knowing the related things.**
* **Prevention: Help us change related things to achieve what we want to achieve.**

**How to Use Big Data to do Specific Prediction: Testing your theory**

Let’s say you already know what you want to predict based on what factors…

**With IoT, we can collerate the sensor data with an existing database:**

* An existing database on things you want to predict with identification info.
* Sensors collecting data only from these people/places.

**Example in Health with a Medical Database + Sensors:**

**Goal: predict what behaviors lead to a specific illness.**

1. You have a **medical database identifying patients with a specific illness (highblood pressure).**
2. You **collect data from these patients** using **sensors**:

* A sensor in a tooth brush will send data about how often and how long you brush your teeth.
* A sensor in your coffee machine will send data about how often and how much you drink coffee.
* A sensor on your plate will send data about what kinds of food you eat and how much.
* A sensor in your shoes will send data about how often you walk and for how long.
  + [ collected sensor data from billions of patients ]

🡺 **Find patterns o**n what kinds of behaviors are related to the illness these people have..

Useful End Results:

1. Prevention plans for the illness
2. Prediction on whether you will end up with the illness

What People Have Accomplished Already:

* a **medical database identifying patients with slow recovery after surgery**
* **collect data from these patients** using **sensors** in medical devices.
  + [ collected data from patients]

🡺 **Find patterns** on what kinds of symptoms are related to slow recovery.

**Example in Law Enforcement with Crime Database + Sensors:**

1. You have a crime database identifying **locations** with **several types of** **crimes**.
2. You **collect data from these locations** using **sensors:**

Sensors embedded in a variety of objects on streets (traffic signals, street signs, roads, etc.) will send data about

* The amount of traffic (foot, bicycles, cars)
* The day and time
* How dark (visibility)
* How many stores are open versus closed

[ collected sensor data from billions of locations ]

🡺 **Find patterns** on what kinds of things are related to the crimes.

Useful End Results:

1. Prevention plans for crimes
2. Prediction on where and what time certain types of crimes will occur today

What People Have Accomplished Already:

The above has already been done in Santa Cruz to predict locations of crimes. This has led to an increase in arrests and a decrease in crimes.

**Wheather is it is an unexpected discovery or to verify your theory, big data analysis will help you find patterns/relations among things to do prediction and prevention.**

**Machine Learning also Uses a Lot of Data**

Although we do not yet have sensors everywhere and we do not have a computer that will discover patterns among everything, people in Artificial Intelligence are working on the following **machine learning tasks** using a lot of data.

Machine Learning == making the computer learn.

There are many types of learning. Sometimes, the machine will learn to adjust its tactics on its own in learning by doing. E.g. I lost the game so I will never do this again.

In supervised machine-learning, we need to use a vast amount of **training data (examples**)

* + **Learning from Examples**

Instead of us telling the machine how to do it**, it discovers how by looking at the training examples. It is often used for “identification” and “categorization” problems.**

**Let’s try it. Can you learn from examples?**

**Yasuko – girl**

**Mariko – girl Junko - ??? Jiro - ???**

**Naomi – giri**

**Kaori – girl**

**Ichiro – boy**

**Taro – boy**

**Kento – boy**

**Shin – boy**

**In previous machine learning methods (plotting and SVM, decision trees, bayes probabilities), humans had to tell the machine which features to pay attention to (e.g. pay attention to the end of names). Tools have been written in Python.**

**With Deep Learning using neural nets, we do not have to. Check out tools such as TenSor.**

**Analysis of insects:** to be able to categorize bugs as good or bad.

Training Data: a vast number of insect images labeled as good or bad -> Have the computer learn how to categorize insects.

End result: feed an insect image -> the computer can categorize it as good or bad.

Usefulness: No-pesticide bug killer.

**Text mining of medical records:** to extract out relevant information (i.e. symptoms, test results, medications, diagnosis, etc.) from textual medical notes. It is much more difficult than you think because a variety of short-hand notations are used, and negations and questions are also included in the notes. To accomplish this task, **Bayesian statistics and probabilities are used. Math is important.**

Training Data: a vast number of medical records with words labeled by important categories (this is a symptom, this is the diagnosis, etc.) -> Have the computer learn how to categorize words in context from these examples so that they can be extracted out into a form.

End result: feed a medical record -> the computer can extract out the symptom and diagnosis words.

Usefulness: Extracted information can then be used to find patterns on what kinds of things are related to what illnesses.

**Analysis of facial expressions:** to be able to notice anger/frustration/stress, etc. even if the expression is for less than a second.

Training Data: a vast number of facial expressions each labeled by the expressed emotion -> Have the computer learn to recognize facial expressions.

End result: feed a facial expression -> The computer can determine the expressed emotion.

Usefulness: In law enforcement, the officers can “notice” a negative reaction immediately.

**Analysis of DNAs:** to beable to determine what part of the DNA is related to what human/animal feature.

Training Data: a vast amount of DNAs each labeled with an existence of a feature or not -> Have the computer learn what DNAs are related to the feature. (**Bio-informatics**)

End result: feed a DNA string -> The computer can determine the existence of a feature or not.

Usefulness: Is this person going to have a cancer?

**Analysis of MRI images:** to be able to “read” your mind (i.e. what image you are forming in your mind).

Training Data: a vast amount of MRI images each labeled by the item being visualized -> Have the computer learn what MRIs are related to which items.

End result: feed an MRI image -> The computer can determine what item you are thinking of.

Usefulness: ??

**Analysis of Student Answers:** to be able to predict student answers (**my project**)

Training Data: a vast amount of question-response pairs collected from students

End result: feed a question -> The computer can determine how the student will answer

Usefulness: able to help students better in Intelligent Tutorins Systems

**Other Advances related to CS Algorithms**

Although the following are not related to CS311, other advances in computing will change your world by 2045.

* **Nano machines** – we will create tiny molecules with sensors that can be injected into a person.

e.g. a molecule will detect certain types of cells and inform you and even “kill” the cells

* **3D printers to create human organs** – we will use a special 3D printer that uses human cells instead of plastic to create human organs.
* **Tutors for everything** – we will have a personal tutor for every student

Creating these things require not just the engineering knowledge, **but creation of algorithms**.

e.g. algorithms for detecting types of cells

e.g. algorithms for reading designs and generating a 3D object

e.g. algorithms for determining the best way to teach a particular person

***That is why you have to become good at developing algorithms.***

***You cannot just turn other people’s algorithms into C++.***

**In-Class Exercise (Week14B) - Big Data and Machine Learning - (MAX 2 people per team) 5 pts**

**Print your name first.**

**Do Week14B Big Data Exercise**

**The End of the Note Section:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***®Summarize here what you have learned in your own words and also write down your own thoughts/reactions/questions.***

***Email me now if you have any questions about what you read in this file.***