Application of Big Data in Social Science

week 1

Heidi Hyeseung Choi
Fall 2022
HYSIS
heidichoi@hanyang.ac.kr

Today we will…

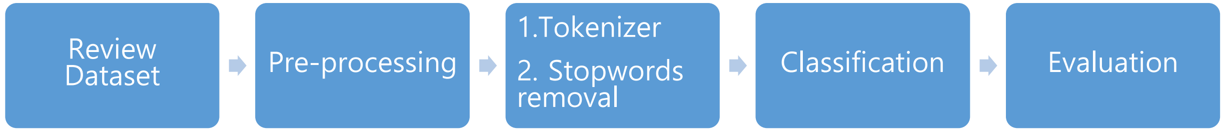
- Go through the lecture format, scope and level of the class, evaluation criteria for this course.
- Introduction of me
- Introduction on why we should care about data
- Go through the syllabus
- Install anaconda

Lecture format

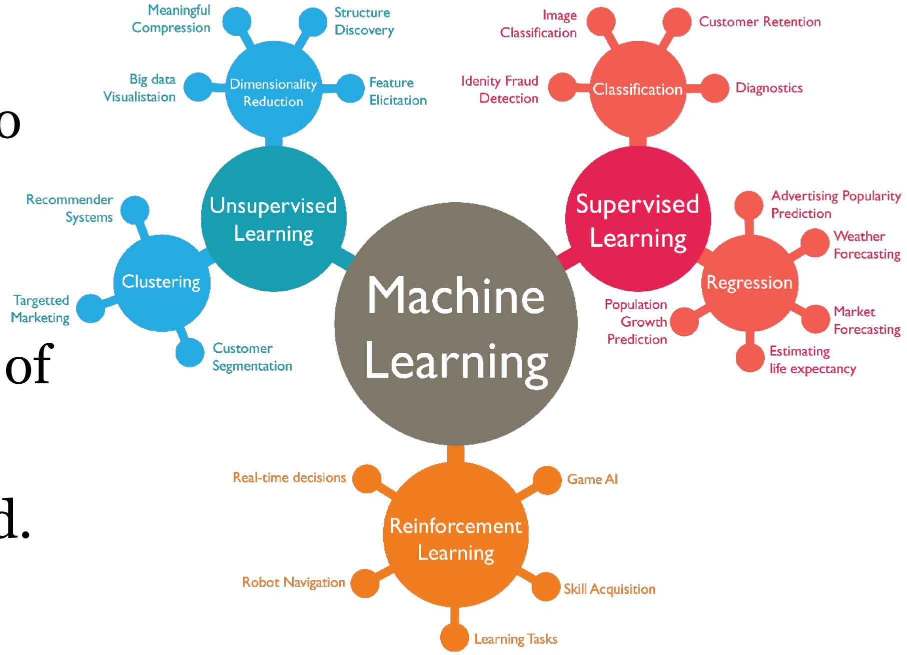
- OFFLINE lectures: Thursday 1:00pm – 4:00pm
- Consist of two parts: 1) lecture 2) coding
- We will use jupyter notebook throughout the course.
- We will take a short break in the middle of the course.
- Questions during the class?

Scope and level of the class

- DIS course: learn the very basics.
  - What kind of steps we need to take when you want to analyze using text data.
  - What different text analysis there are.
  - Be more familiar with terminologies.
  - Not go into real details (maths) of each algorithm etc.



- Those who are not so comfortable with Python, don't worry too much.
- Mainly on practical coding for social science majors.
- Depending on the level of the class, I may change on the depth of what we learn.
- Based on last year's experience, students still feel very confused.



Scope and level of the class

How familiar you should be with Python?

- basic datatypes: integers, floats, strings, lists, dictionaries..
- how to assign a variable.. a = 1
- if statement, for and while loops and functions
- assignment uses = and comparison uses ==.
- logical operators are words (and, or, not)
- importing libraries
- importing files
- Handling dataframes
- data manipulation using pandas, numpy
- a little bit of visualization using seaborn, matplotlib.
- if you come across something you do not know, we can go through together.

Python basics	Python data types
	variables and expressions
	conditional codes
	loops and iteration 1: for loops
	loops and iteration 2: while loops
	loops and iteration 3: nested loops
	functions
	Data manipulation
	Data visualization

Evaluation

Criteria	Ratio
Attendance	10%
Study Participation	10%
Mid-term exam	30%
Final exam	30%
Homework Assignment 1	10%
Homework Assignment 2	10%
Total	100%

Attendance & Study Participation

1. Attendance: 10 points

- the actual online or offline attendance
- must attend more than 10 classes to be graded.

2. Study participation: 10 points

- upload jupyter notebook on LMS.
- will give you enough time... :S

- For late uploads (up to 1 hour): 0.5 point deduction
- For absence or not uploading your file.

= > one point off from the entire attendance or study participation grade.

Case 1:
attended the class but did not upload the file:
-1 point from study participation grade.
attendance =>10 points,
study participation => 9 points

Case 2:
did not attend nor hand in the work :
-1 points off from attendance and study participation (total -2)

Evaluation

Criteria	Ratio
Attendance	10%
Study Participation	10%
Mid-term exam	30%
Final exam	30%
Homework Assignment 1	10%
Homework Assignment 2	10%
Total	100%

Mid term exam & Final exam

1. Mid-term exam (30%)

- To be held on week 9
- Offline unless stated otherwise
- Theoretical and coding questions

2. Final Exam (30%)

- To be held on week 16
- Offline unless stated otherwise
- Theoretical and coding questions

Evaluation

Criteria	Ratio
Attendance	10%
Study Participation	10%
Mid-term exam	30%
Final exam	30%
Homework Assignment 1	10%
Homework Assignment 2	10%
Total	100%

Homework Assignment 1 & 2

1. Homework Assignment 1 : 10 points

- will be given out in October
- will be based on what we learned during the class.

2. Homework Assignment 2: 10 points

- will be given out in December

Changed schedule ☺

Updated version

9.1	1	Introduction
9.8	2	Web Scraping
9.15	3	
9.22	4	Natural Language Processing
9.29	5	
10.6	6	Text Analysis
10.13	7	
10.20	8	Revision?
10.27	9	Mid term exam
11.3	10	Social Network Analysis
11.10	11	Machine Learning: Supervised Learning
11.17	12	
11.24	13	Machine Learning: Unsupervised Learning
12.1	14	
12.8	15	Data Visualization
12.15	16	Final Exam

I need to go on two business trips in October. So we will take a week off (yay ☺) or have a revision class.

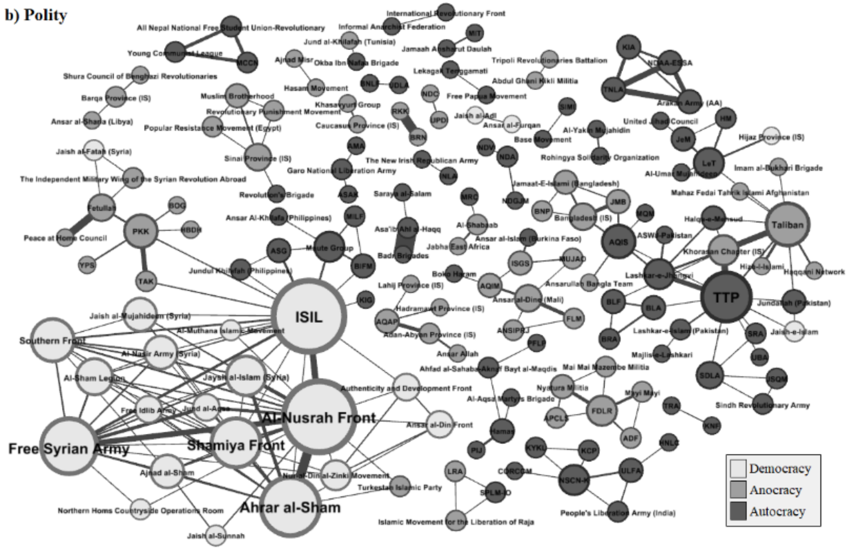
- I might have to upload video lecture in October.
- Mid term exam will be held on 9th week or if allowed, have it on 8th.

Instructor: Heidi Hyeseung Choi

Research Assistant Professor at Center for Global Circular Economy at Hanyang University

- HYDIS 06’ – International Politics track (B.A.)
- KAIST Graduate school of Future Strategy (M.A.)
- KAIST Graduate school of Future Strategy (Ph.D.)

- at Hanyang Division of Int’l Studies
  - Student council
  - Debate society
- before graduate school
  - NGO
  - MOFAT IFANS (Institute of Foreign Affairs and National Security)
- during graduate school
  - Center for Future Strategy, KAIST
  - Science and Technology Policy Institute (STEPI)
- at graduate school
  - Computational Social Science and Future Strategy Lab
  - Complex network theory & data science
  - International Security
    - network dynamics of terrorist group alliances



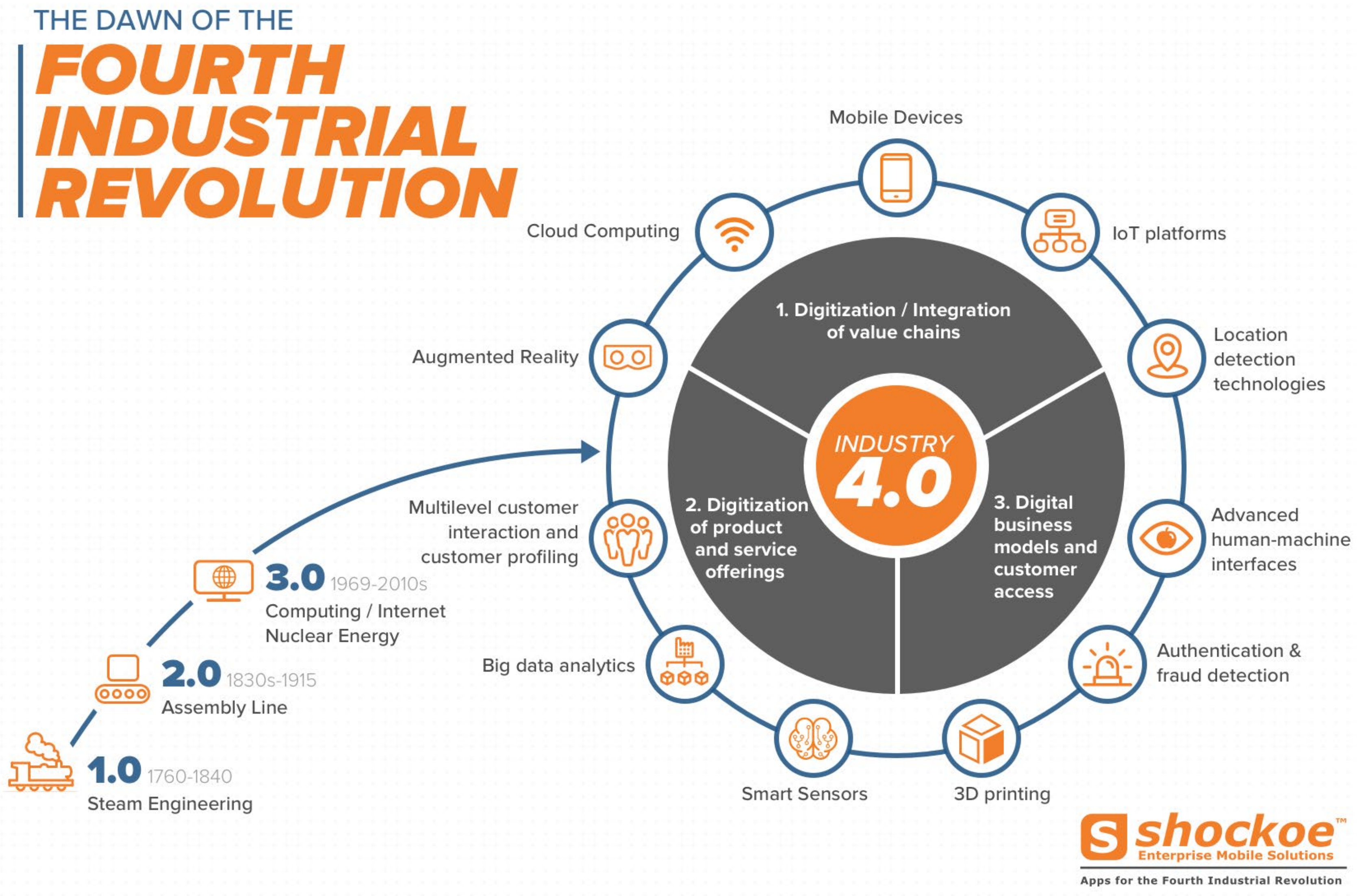
Apply data science on international relations

So…

Why should I suddenly have to care about all this data and coding?

I’m not a computer science major!!!

Background

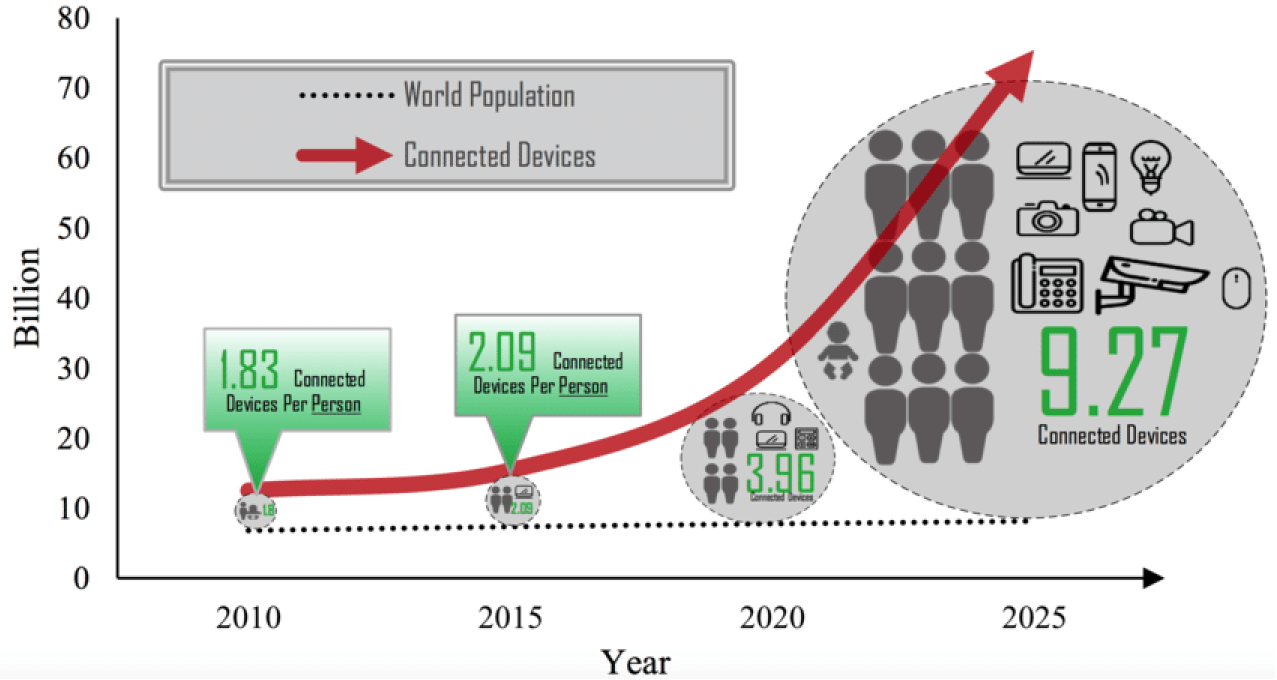

THE DAWN OF THE
FOURTH INDUSTRIAL REVOLUTION
Mobile Devices
IoT platforms
Cloud Computing
Location detection technologies
Augmented Reality
1. Digitization / Integration of value chains
2. Digitization of product and service offerings
3. Digital business models and customer access
INDUSTRY 4.0
Multilevel customer interaction and customer profiling
Advanced human-machine interfaces
3.0 1969-2010s
Computing / Internet
Nuclear Energy
Big data analytics
Authentication & fraud detection
2.0 1830s-1915
Assembly Line
1.0 1760-1840
Steam Engineering
Smart Sensors
3D printing
shockoe™
Enterprise Mobile Solutions
Apps for the Fourth Industrial Revolution

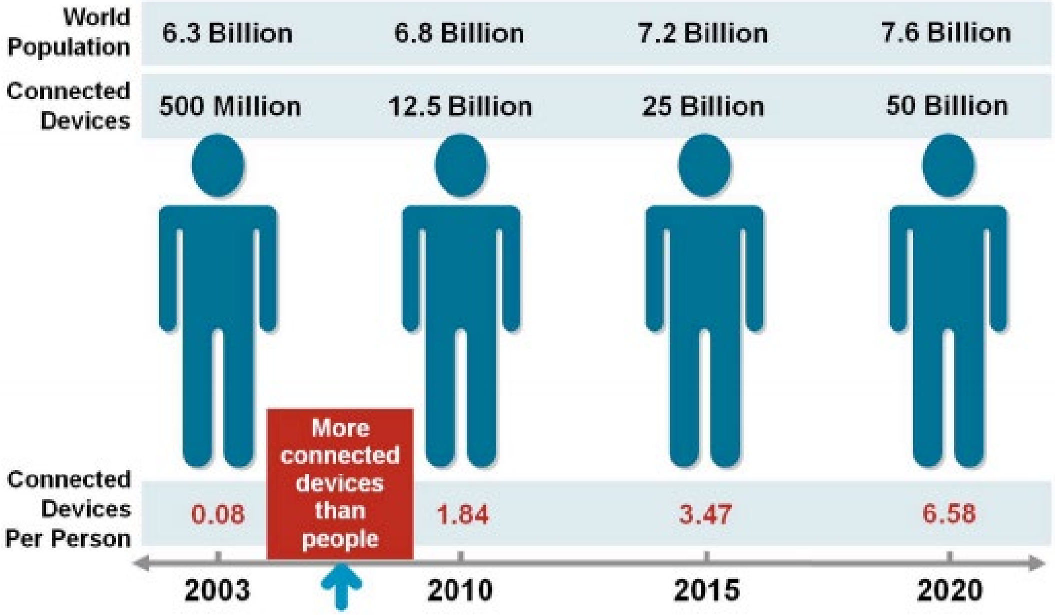
The 4th industrial revolution..

- Definition of the Fourth Industrial Revolution by WEF:
  - “A fundamental change in the way we live, work and relate to one another.
  - A new chapter in human development enabled by extraordinary technology advances commensurate with those of the first, second and third industrial revolutions.
  - These advances are merging the physical, digital and biological worlds in ways that create both huge promise and potential peril.
  - The Fourth Industrial Revolution is about more than just technology-driven change; it is an opportunity to help everyone, including leaders, policy-makers and people from all income groups and nations, to harness converging technologies in order to create an inclusive, human-centred future.”

Background: why the hype?

- Getting more and more connected now.





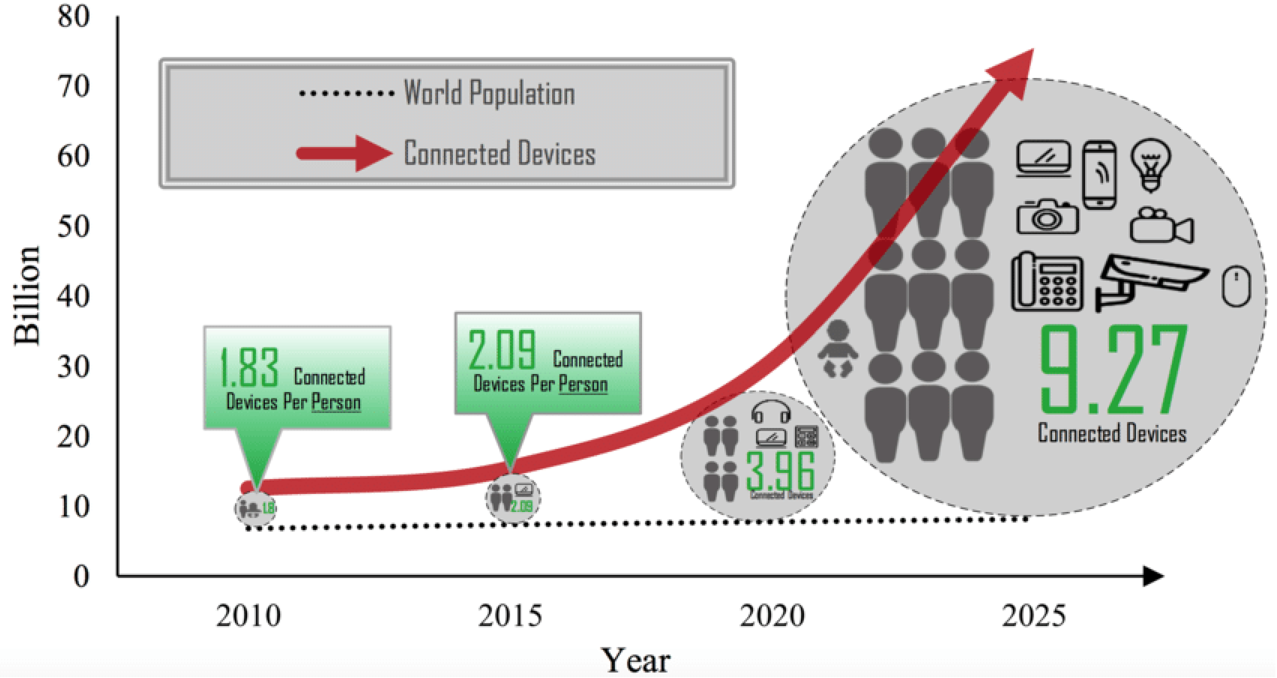
World Population	6.3 Billion	6.8 Billion	7.2 Billion	7.6 Billion
Connected Devices	500 Million	12.5 Billion	25 Billion	50 Billion
Connected Devices Per Person	0.08	1.84	3.47	6.58
	2003	2010	2015	2020

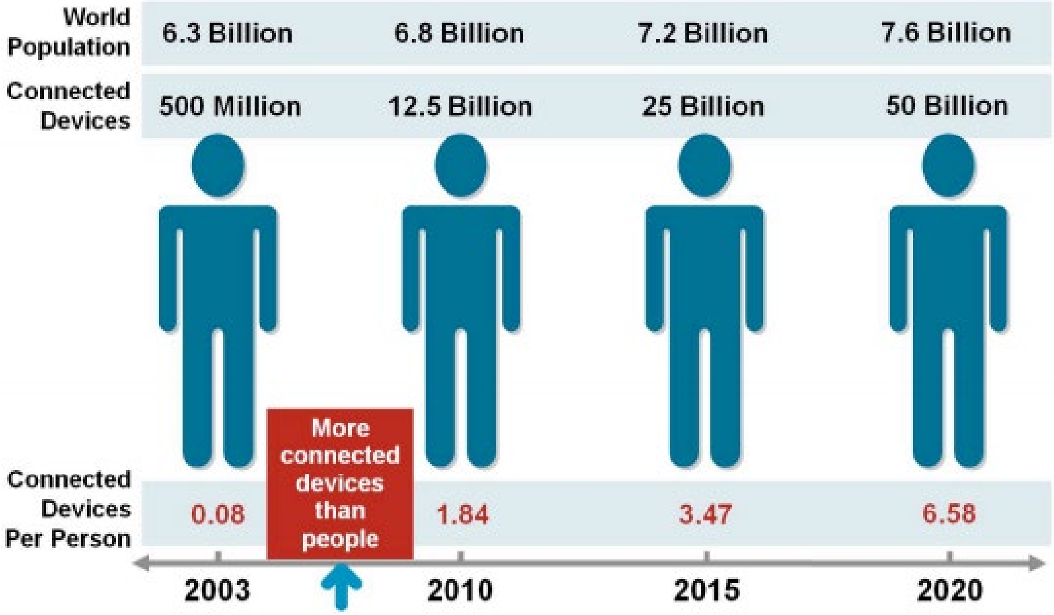
More connected devices than people

Source: Cisco IBSG, April 2011

Background: why the hype?

- Getting more and more connected now.





Source: Cisco IBSG, April 2011

more data available!

But still, why should I care?

Simply because it will directly affect your everyday life!

- AI
  - autonomous vehicles, Netflix, chatbots, translators etc.

- IoT
  - wearables.. Fitness trackers, GPS tracking, virtual glasses, health
  - Smart homes.. Lights, air purifiers, robotic vacuum cleaner, gas
  - Jobs!
  - And this online courses too

A brief video on smart homes and how your life will be affected.

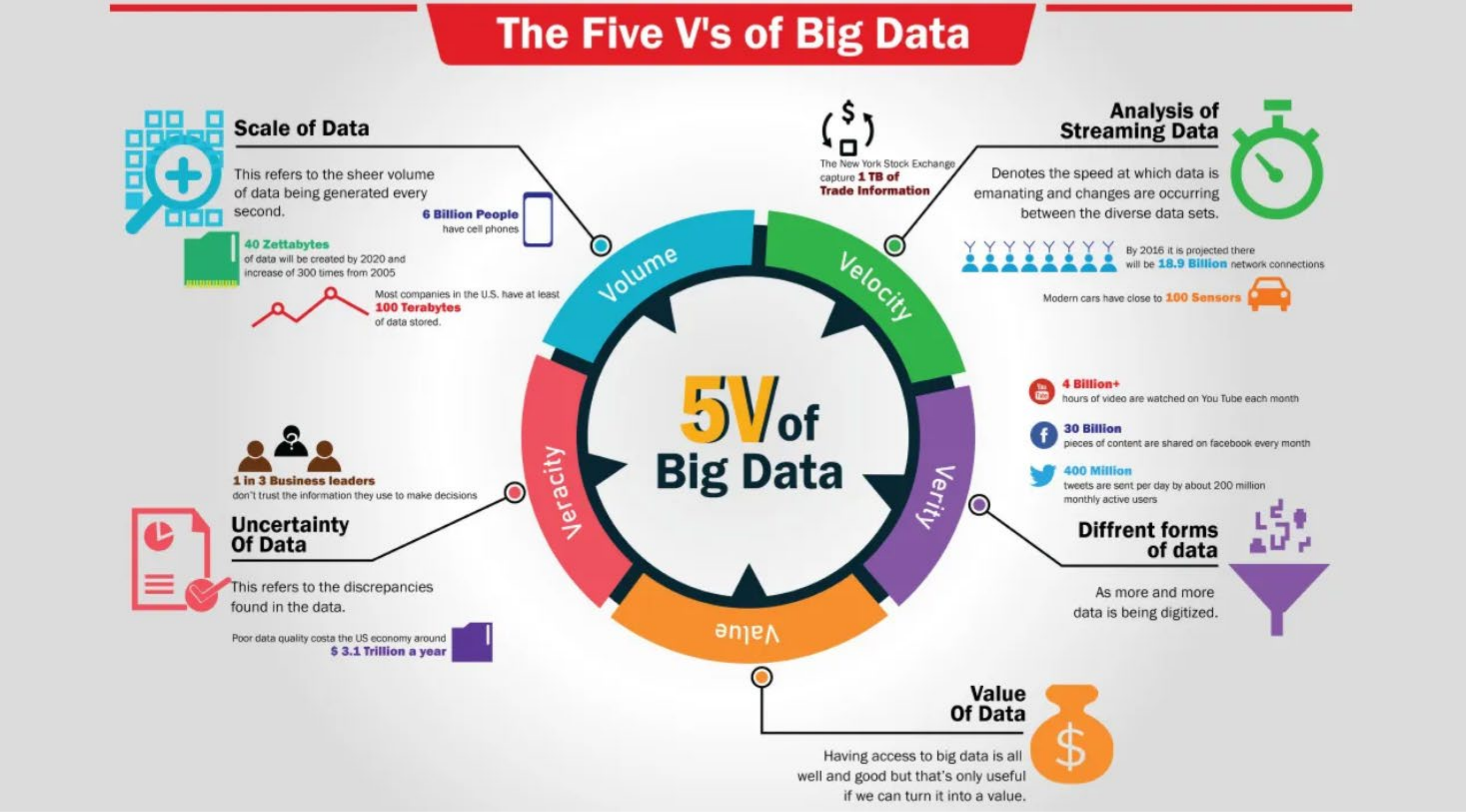
What is Big Data?

- What made it possible?
  - Availability of data from various sources.
    - eg) social media, internet, IoT devices.
  - Advancing data storage technology
- No agreed definition of the exact size of data to be ‘big data’.
- The 5Vs of Big data

Volume	The amount/size of data
Velocity	How fast the data is being updated
Variety	Diversity of data types
Veracity	Accuracy or trustful of dataset
Value	Usefulness (significance) of gathered data

- Big Data Analytics is the process of storing, transforming, analyzing, and processing large amount of data (a.k.a. big data) to generate valuable insights.

The 5Vs of Big Data



The Five V's of Big Data

5V of Big Data

Volume

Velocity

Verity

Value

Veracity

Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.

The New York Stock Exchange capture **1 TB of Trade Information**

Analysis of Streaming Data

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.

By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**

4 Billion+ hours of video are watched on You Tube each month

30 Billion pieces of content are shared on facebook every month

400 Million tweets are sent per day by about 200 million monthly active users

Diffrent forms of data

As more and more data is being digitized.

Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.

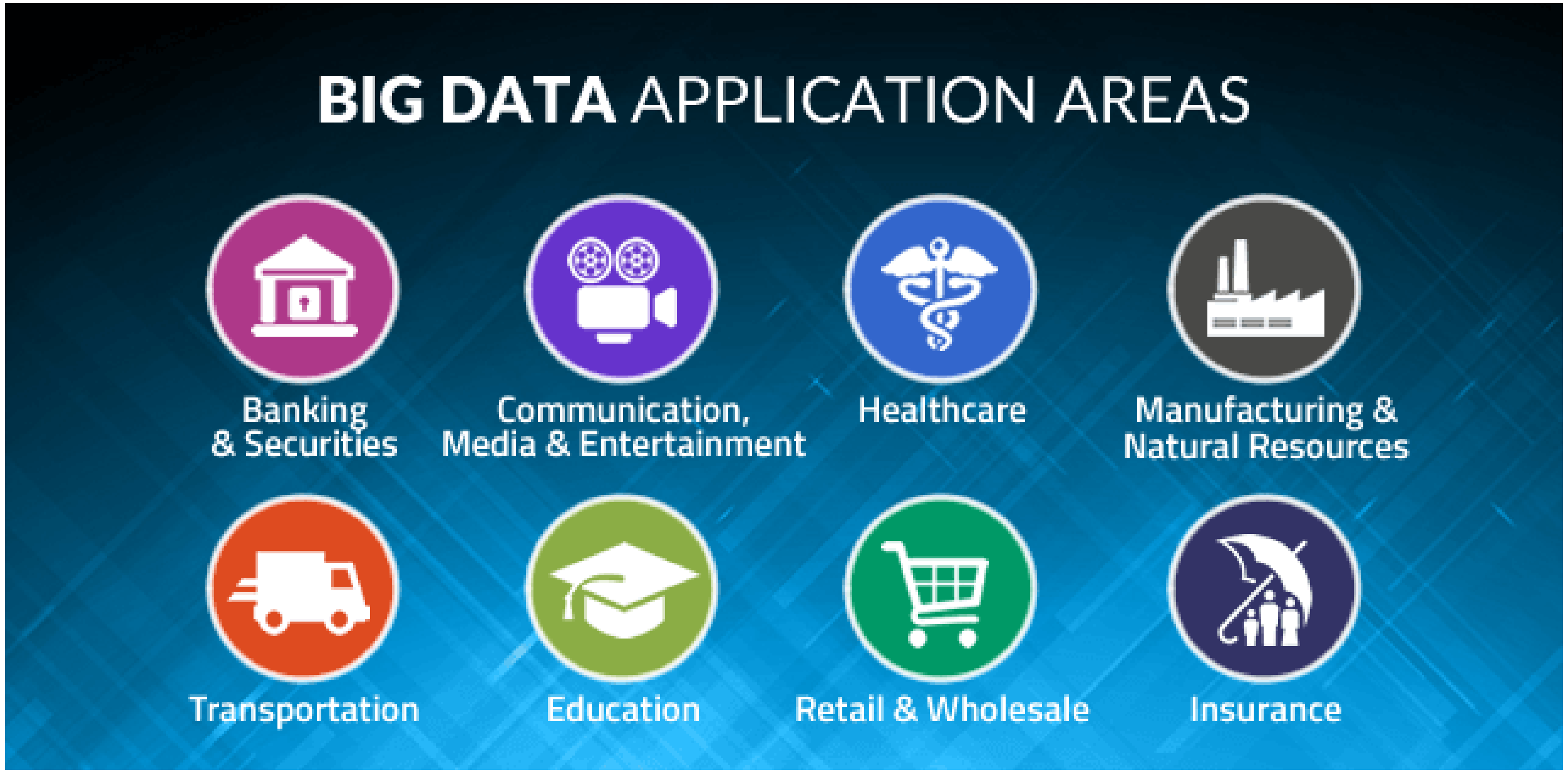
1 in 3 Business leaders don't trust the information they use to make decisions

Uncertainty Of Data

This refers to the discrepancies found in the data.

Poor data quality costs the US economy around **$ 3.1 Trillion a year**

Big data applications in various fields.


BIG DATA APPLICATION AREAS
Banking & Securities
Communication, Media & Entertainment
Healthcare
Manufacturing & Natural Resources
Transportation
Education
Retail & Wholesale
Insurance

1. Big Data in Banking Sector

- Use different clustering techniques to take important decisions (opening up new branch etc)
- Association algorithm can be applied to predict the amount of cash needed to be present in a certain branch at a specific time of the year.
- Develop banking platforms so that bank operations can be done online.
- Machine learning and AI are used to detect fraudulent activities and reporting directly to related companies.
- Handle, store and analyze massive amount of bank data and ensure its security.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

2. Big Data in Casino Business

- Quickly identify the most popular games to increase the number of similar machines to attract more customers.
- Analyze routes of the customers or order of the machines that they approach to for relocation of machines.
- Encourage customers to visit their casino again by analyzing their information.
- Make sure that casino earns profit by changing probability of customers' earnings.
- Automatically detect games that are not so popular so that they can increase their performances.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

3. Big Data in Restaurants

- Collect essential data from customers (gender, age, preferred menu, time of the day etc) to find new possibilities and develop new menu.
- Evaluate data to predict customer behaviors, such as which menu customers tend to buy together (chicken and coke? Salad?)
- Find hidden patterns and similarities to help restaurants to determine their potential customers.
- Use smart inventory or stock management system to help managers to keep track of the resources.
- Use data to predict the specific time of the day so that they can prepare food according to the demand.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

4. Big Data in Media & Entertainment

- Gather information and demand of the individuals.
- Identify the device and the most effective time to view data for analysis.
- Use big data to identify and set target groups for media companies
- Identify reasons behind subscribing and unsubscribing a content.
- Choose the place where artists promote their performances.
- Analyze the devices of individuals to adjust the screen size, OS etc to place their songs or videos.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

5. Applications of Big Data in Tourism

- Analyze data that travelers provide on social media
- Gather information on credit or debit card information for quick purchases and quick identification of the travelers.
- Use information on geo-location, traffic and weather, travel agencies can offer benefits tailored to particular customers.
- Can help to enhance customer service and customer's buying habits by analyzing past information.
- Increase efficiency of customers by relating to airplane and travel packages.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

6. Big Data Applications in Healthcare

- Using wearable digital devices, big data can monitor patients and send reports to the associated doctors.
- Big data can evaluate symptoms and identify any many diseases at the early stages.
- Can keep the sensitive records secured and store huge amount of data efficiently. Availability of medical database has also played a major role.
- Big data can save lives by analyzing the behavior and health condition of the patients.
- Big data applications can also foretell the location where there is a chance of dengue or malaria spreading.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

7. Big Data in e-commerce

- Can collect data and customer requirements even before the official operation has started.
- Creates a high performing marketing model and set a startup apart from the existing and become successful.
- Ecommerce owners can identify the most viewed products and the pages that appeared the maximum number of time.
- Evaluates customers behavior and suggests similar products. It increases the number of sales and generates revenue.
- If any product is added to cart but was not ultimately bought by a customer, big data can automatically send a promotional offer to that particular customer.
- Big data applications can generate a sorted report depending on the visitor's age, gender, location, and so on.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

8. Big Data to Ensure National Security

- The governments collect the information of all citizens, and this data is stored into a database for many purposes.
- Data science is implemented on these databases to extract meaningful information alongside a hidden relationship between datasets.
- Can evaluate the density of the population in a specific location and identify the possible threatening situations even before anything has occurred.
- Security officers can use this dataset to find any criminal and detect fraudulent activities in any area of the country.
- Besides, related personnel can predict the potential outspread of any virus or diseases and take necessary actions to prevent.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

9. Big Data in Education

- Can store, manage, analyze the large datasets that include students records. Maintaining security using big data is also quotable.
- Big data can make sure if the question papers do not get leaked before the examinations.
- It provides influential data on classroom activities and helps in taking decisions for the organizations.
- Using high-resolution cameras, video footage and image processing, big data can evaluate student’s facial expression and can track their movements.
- Motivates students by identifying problems and rendering the best possible education to the children.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

10. Big Data in Digital Marketing

- Analyzes market, competitors and evaluate the business goal. It can identify the opportunities as well.
- Can find the existing social media users and target them based on demographics, gender, income, age, and interests.
- Generates reports after every ad campaign that includes the performance, audience engagements, and what could be done for generating better results also.
- Data science used for possible retargeting customers and transform into loyal clients.
- Focuses on highly searched topics and advice the business owners to execute them on content strategy to rank business's website higher on google.
- Can create lookalike audiences using the existing audience database to target similar clients and earn the profits.

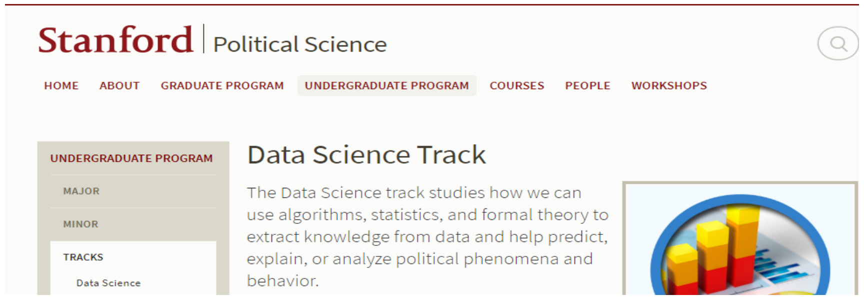
Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

11. Big Data Application in Government Sector

- The government can access daily functional information considering particular indecent or topic.
- Can help to identify the areas that need attention and analyze to improve the current situation.
- Using big data, governments can easily reach to public demand and act accordingly.
- Big data helps to monitor the decisions taken by the government and evaluate the results
- Besides, can predict any terrorist attack and take necessary action to prevent unwanted conditions.

Source: https://www.omni-academy.com/best-big-data-applications-examples-qatar/

Even in political science? International relations?



Data Science Track

The Data Science track studies how we can use algorithms, statistics, and formal theory to extract knowledge from data and help predict, explain, or analyze political phenomena and behavior.

- Academics: qualitative analysis vs quantitative analysis
- Quantitative analysis: statistical analysis using data
- Big data + political science
- Which data?
  - social data (eg. twitter feeds, facebook, Instagram, newspaper comments)
  - geospatial locations and mobility (eg. mobile phone GPS, satellite images)
  - other traditional data (eg. economic data, trade data, insurgents data, war data)
- Analyze and forecast political phenomena and behavior
- This became possible due to increased number of available data and computational power.
- Increasing number of universities around the world is incorporating data into political science.

So…

what we will be covering throughout the semester..

weekly course schedule: before the mid-term exam

9.1	1	**Introduction**
9.8	2	**Web Scraping**
9.15	3	
9.22	4	**Natural Language Processing**
9.29	5	
10.6	6	**Text Analysis**
10.13	7	
10.20	8	**Revision?**
10.27	9	**Mid term exam**

Web scraping and NLP

Web scraping

- Web scraping is literally scraping a large amount of data from websites
- Would be better than copy-and-paste work ☺

Natural Language Processing (NLP)

- Helping computers understand human’s natural language.
- Goal of NLP is to read, decipher, understand and make sense of the human language.
- NLP is a difficult problem in computer science.
  - Most of the data that we scrap off the internet are unstructured.
  - need to pass rules to computers which can be difficult for computers to understand.
  - abstract words or sarcastic remark?
- Word segmentation, parsing, sentence breaking.

weekly course schedule: before the mid-term exam

9.1	1	**Introduction**
9.8	2	**Web Scraping**
9.15	3	
9.22	4	**Natural Language Processing**
9.29	5	
10.6	6	**Text Analysis**
10.13	7	
10.20	8	**Revision?**
10.27	9	**Mid term exam**

Text analysis

- Classify information from texts, such as keywords, names, tweets, emails etc.
- Learn different text analysis techniques to analyze text data.
- Word frequency method: counting the most frequently occurring words in a given text.
- Identifying words that commonly co-occur.
    - eg) ‘coronavirus disease’, ‘social security’
- Identify relations among words.
- Sentiment analysis: detecting emotions and classifying text as positive, negative or neutral.

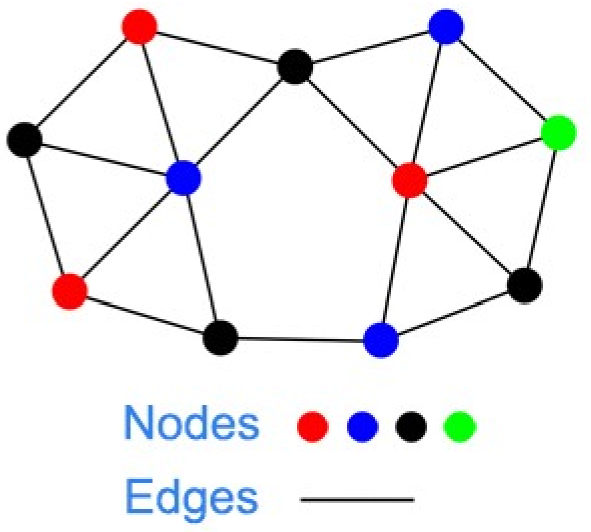
Text analysis and visualization

weekly course schedule: after the mid-term exam

11.3	10	**Social Network Analysis**
11.10	11	**Machine Learning: Supervised Learning**
11.17	12	
11.24	13	**Machine Learning: Unsupervised Learning**
12.1	14	
12.8	15	**Data Visualization**
12.15	16	**Final Exam**

Network Analysis

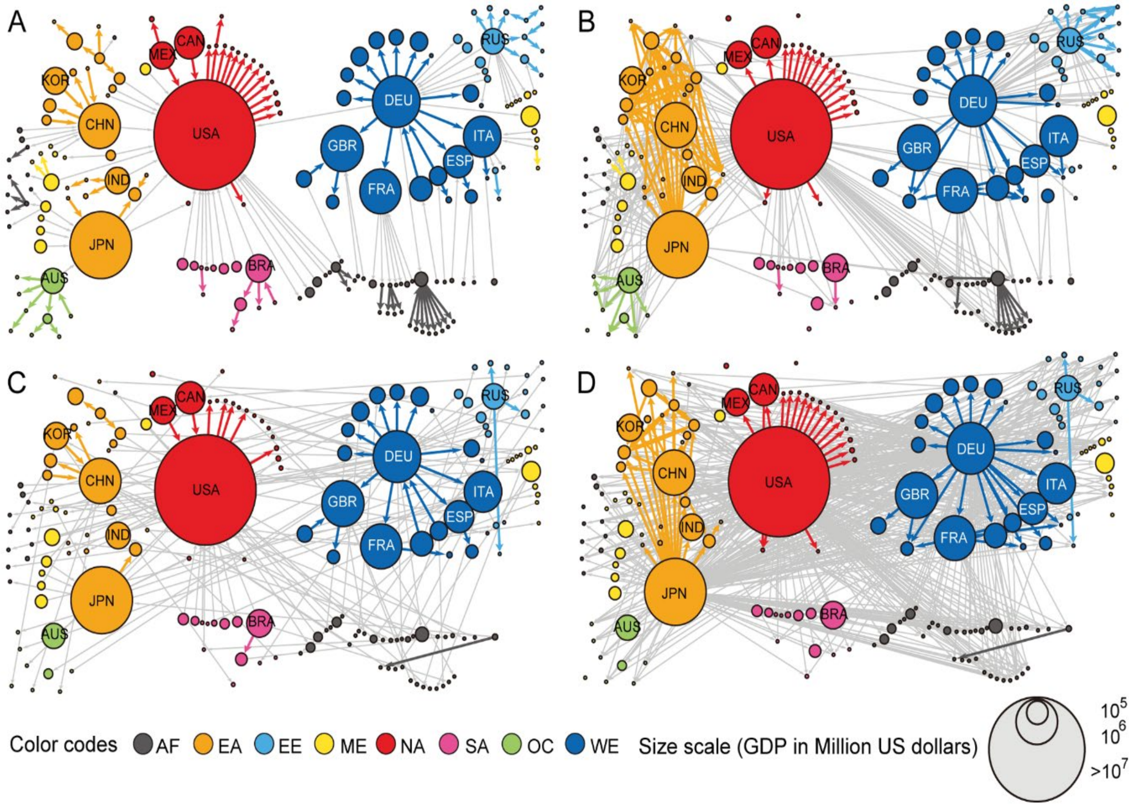
- analyzing social structures through networks and graph theory.
- network consists of nodes and edges that connect them.
- eg) business network, friends network, country alliance network, disease transmission etc.



ComputerHope.com

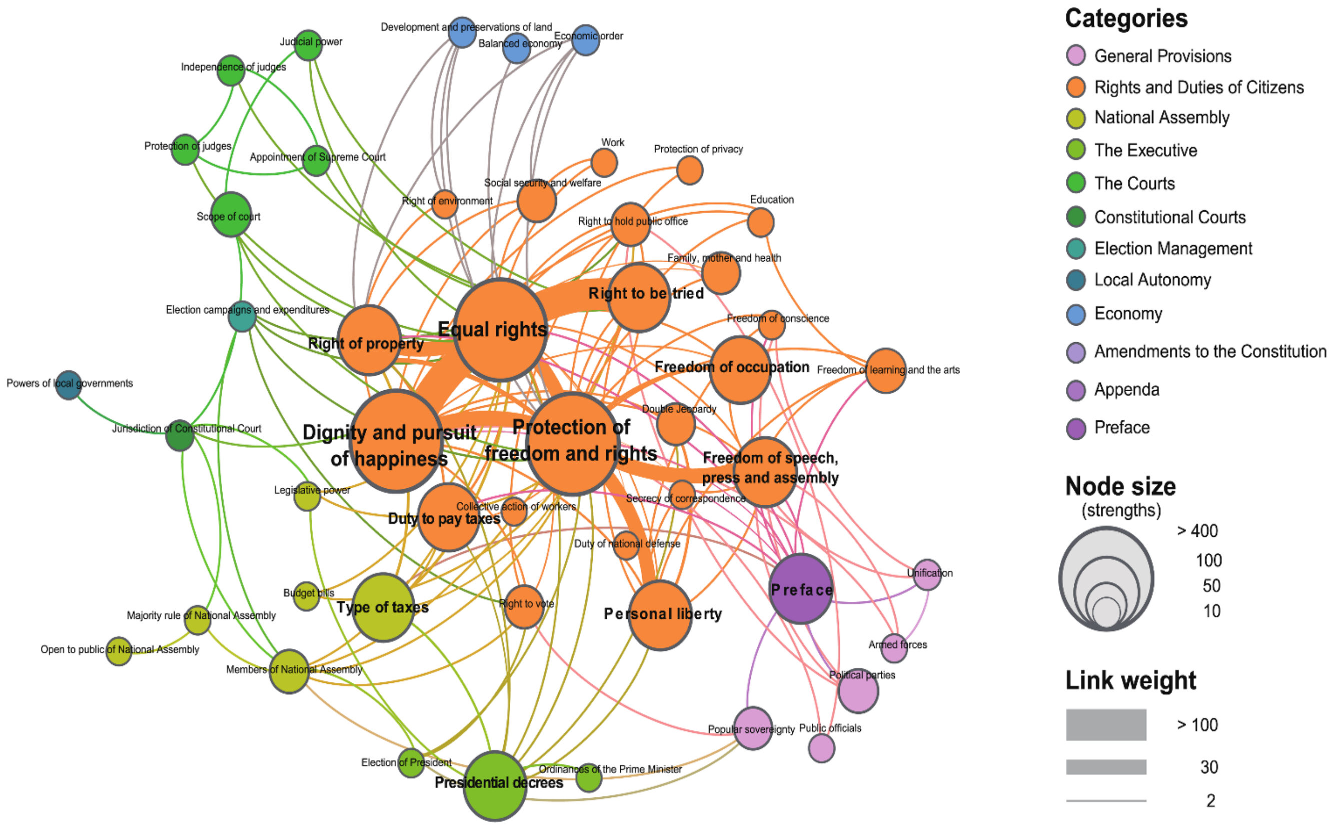
Network Analysis

Global patterns of crisis spreading during economic crises



Impact of the topology of global macroeconomic network on the spreading of economic crises, PLoS ONE 6 e18443 (2011) [K.-M. Lee, J.-S. Yang, G. Kim, J. Lee, K.-I. Goh, I. Kim]

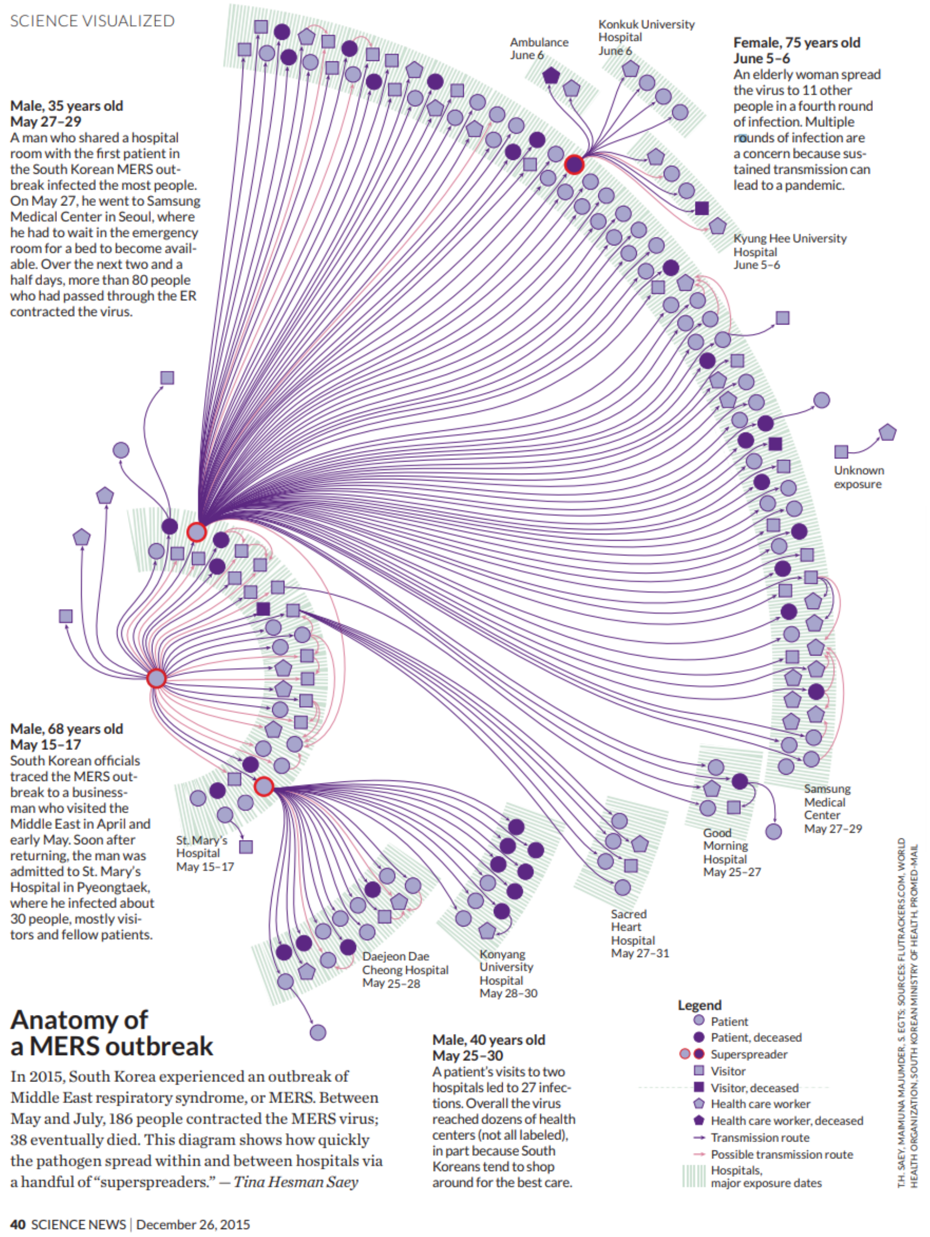
The backbone of the Constitutional legal network (CLN).



Network structure reveals patterns of legal complexity in human society: The case of the Constitutional legal network, PLoS ONE 14, e0209844 (2019) [B. Lee, K.-M. Lee. J.-S. Yang]

Network Analysis

- Analyze the spread of virus and find who affected the most, using different centrality measures.



SCIENCE VISUALIZED

Male, 35 years old
May 27–29
A man who shared a hospital room with the first patient in the South Korean MERS outbreak infected the most people. On May 27, he went to Samsung Medical Center in Seoul, where he had to wait in the emergency room for a bed to become available. Over the next two and a half days, more than 80 people who had passed through the ER contracted the virus.

Female, 75 years old
June 5–6
An elderly woman spread the virus to 11 other people in a fourth round of infection. Multiple rounds of infection are a concern because sustained transmission can lead to a pandemic.

Male, 68 years old
May 15–17
South Korean officials traced the MERS outbreak to a businessman who visited the Middle East in April and early May. Soon after returning, the man was admitted to St. Mary's Hospital in Pyeongtaek, where he infected about 30 people, mostly visitors and fellow patients.

Male, 40 years old
May 25–30
A patient's visits to two hospitals led to 27 infections. Overall the virus reached dozens of health centers (not all labeled), in part because South Koreans tend to shop around for the best care.

Anatomy of a MERS outbreak

In 2015, South Korea experienced an outbreak of Middle East respiratory syndrome, or MERS. Between May and July, 186 people contracted the MERS virus; 38 eventually died. This diagram shows how quickly the pathogen spread within and between hospitals via a handful of "superspreaders." — *Tina Hesman Saey*

T.H. SAEY, MAIMUNA MAJUMDER, S. EGTS; SOURCES: FLUTRACKERS.COM, WORLD HEALTH ORGANIZATION, SOUTH KOREAN MINISTRY OF HEALTH, PROMED-MAIL

40 SCIENCE NEWS | December 26, 2015

weekly course schedule: after the mid-term exam

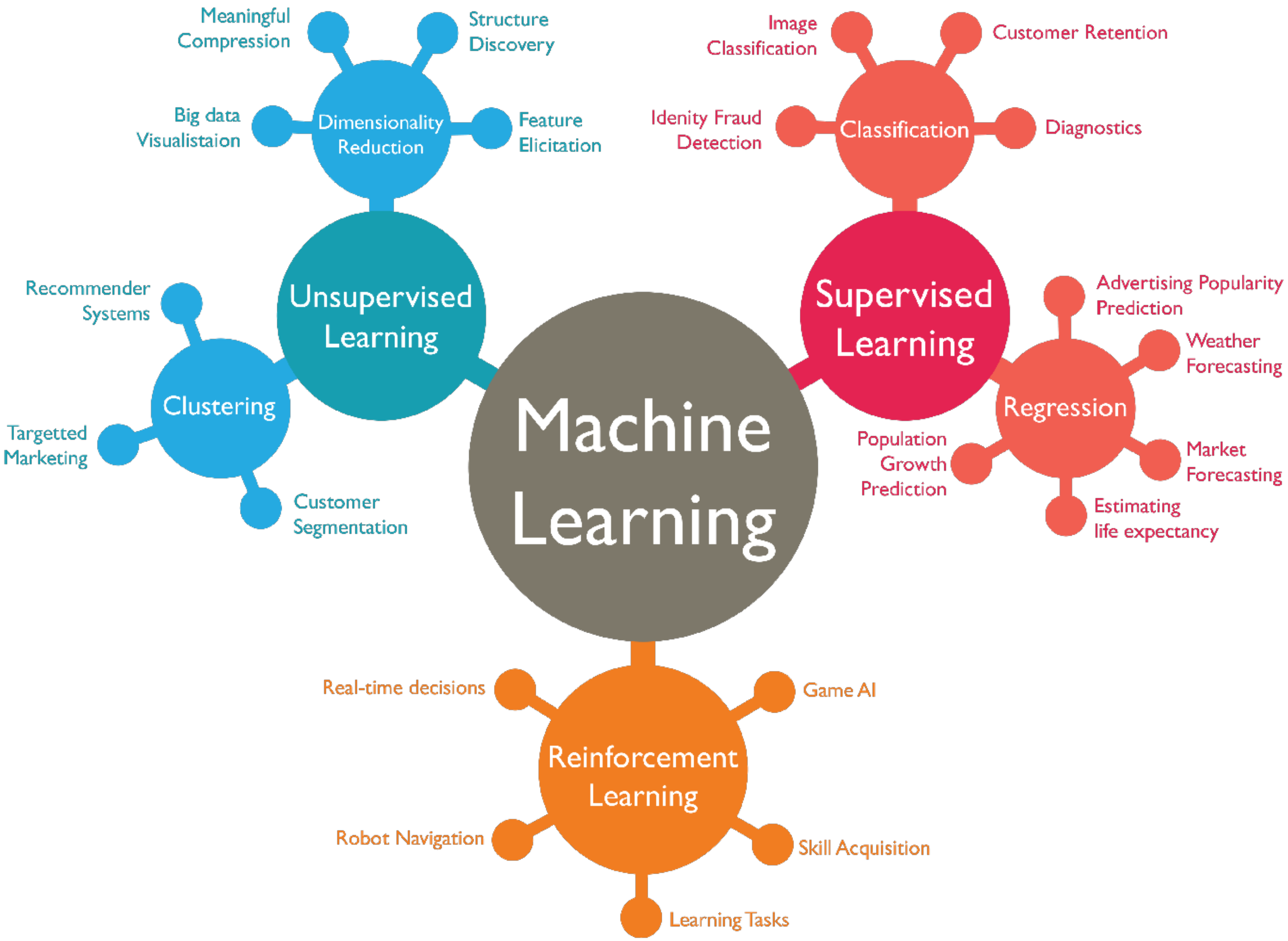
11.3	10	**Social Network Analysis**
11.10	11	**Machine Learning: Supervised Learning**
11.17	12	
11.24	13	**Machine Learning: Unsupervised Learning**
12.1	14	
12.8	15	**Data Visualization**
12.15	16	**Final Exam**

Machine Learning (ML)

> “*Machine Learning algorithms enable the computers*
> *to learn from data, and even improve themselves,*
> *without being explicitly programmed.*”
>
> \- Arthur Samuel (1901-1990)

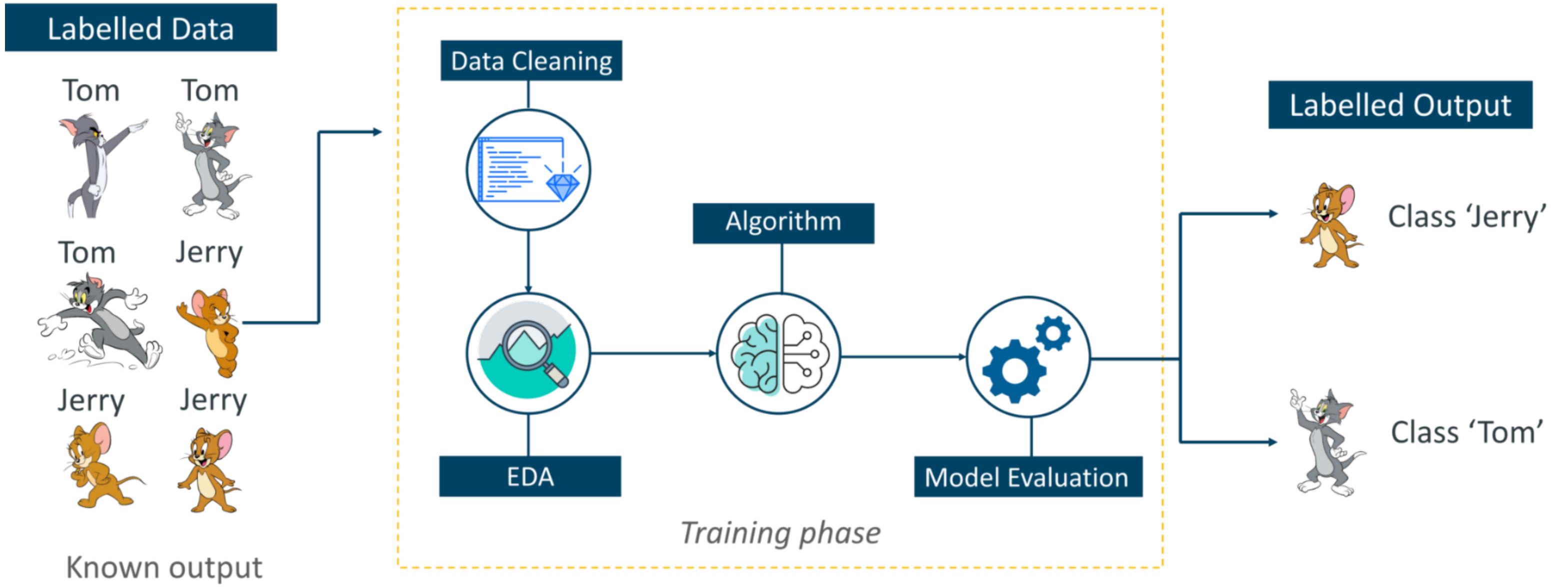
- develop algorithms that make machines learn to do something without being explicitly programmed.
  - eg) self-driving cars, speech recognition, effective web search, fitness trackers
- provide computers the ability to automatically learn and improve from experience
- to do.. prediction, image recognition, speech recognition, medical diagnoses.. etc

Machine Learning (ML): the big picture



Machine Learning (ML)

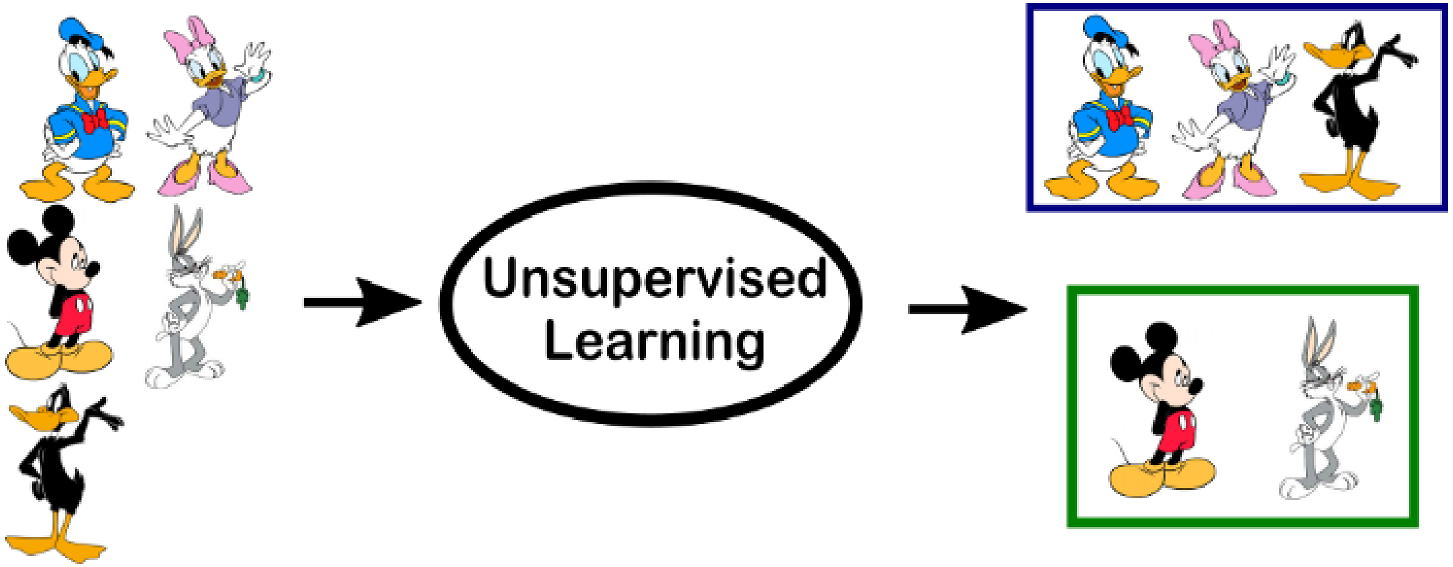
- supervised machine learning:
  - algorithm is provided with set of data where it has the answer for each of the input values.
  - apply what has been learned in the past to new data to predict future events.
  - Types of supervised learning: Regression, classification
  - Applications: risk evaluation, forecast sales



source: https://becominghuman.ai/an-introduction-to-machine-learning-33a1b5d3a560

Machine Learning (ML)

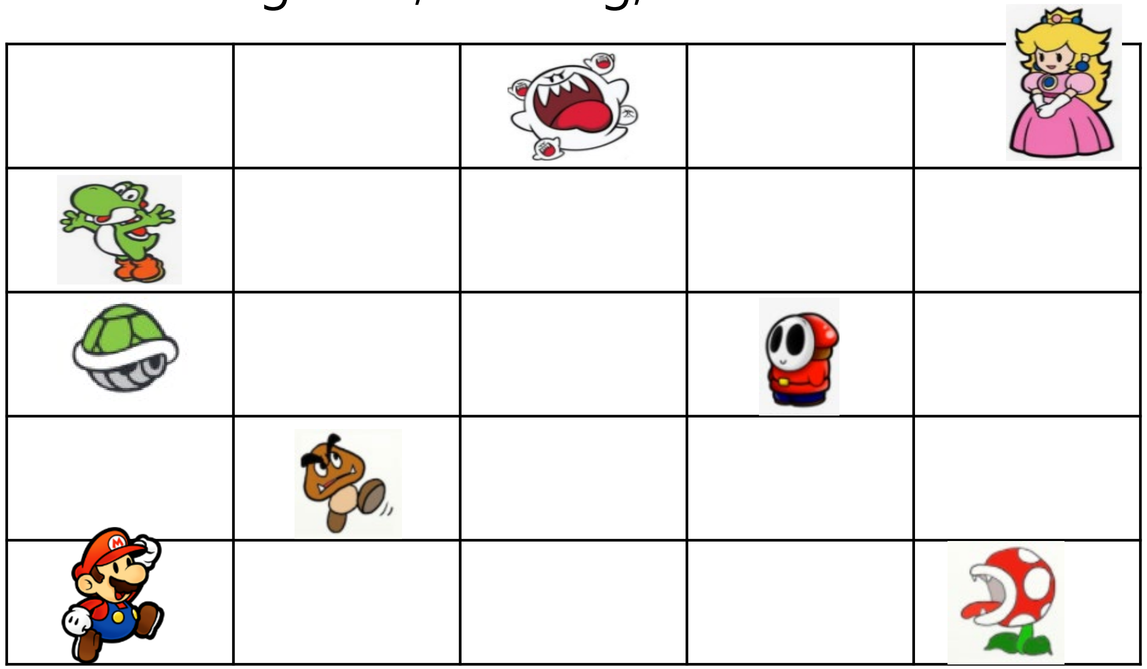
- unsupervised machine learning:
  - When data provided is neither classified nor labeled.
  - Self-organized learning used to find hidden patterns from unlabeled data.
  - Provide the machine with data and ask to look for hidden feature and cluster.
  - Types of unsupervised machine learning: clustering, dimension reduction
  - Applications: Recommendation system, anomaly dectection.



source: https://becominghuman.ai/an-introduction-to-machine-learning-33a1b5d3a560

Machine Learning (ML)

- Reinforcement machine learning:
  - Neither based on supervised or unsupervised learning.
  - Enable an agent to learn in an interactive environment by trial and error.
  - Algorithm learns to react to an environment on their own.
  - The agent gets the **reward**(appreciation) on success but will **not receive any reward** or appreciation on failure. In this way, the agent learns from the environment.
  - Best solution is decided based on the maximum reward.
  - Applications: Self Driving Cars, Gaming, Healthcare

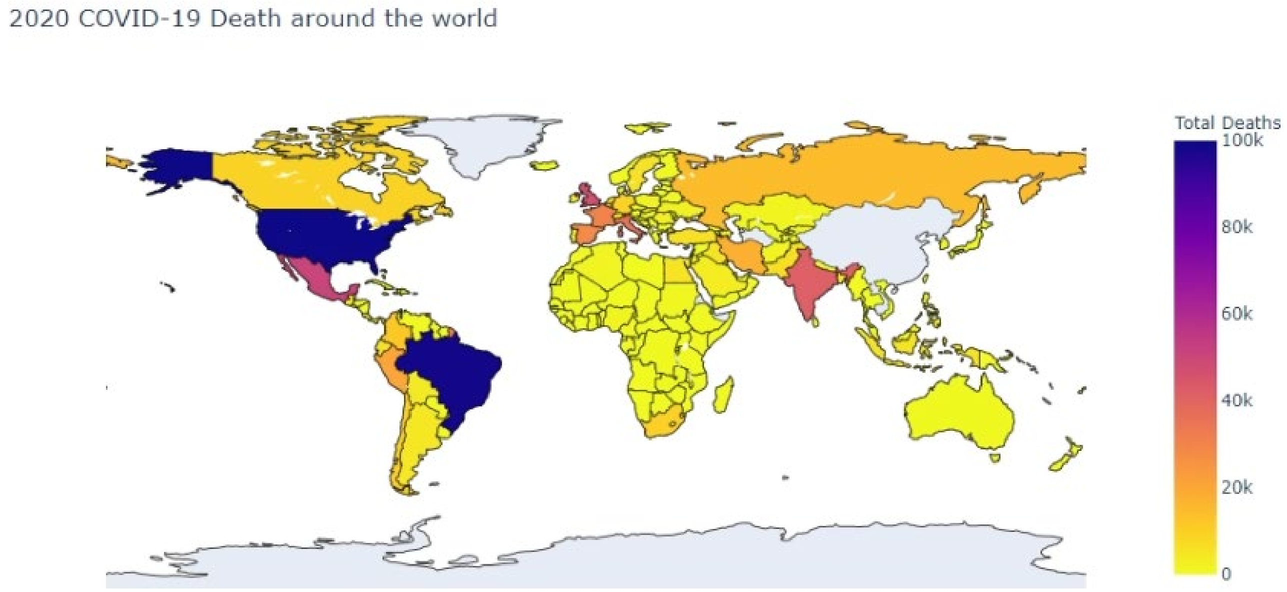


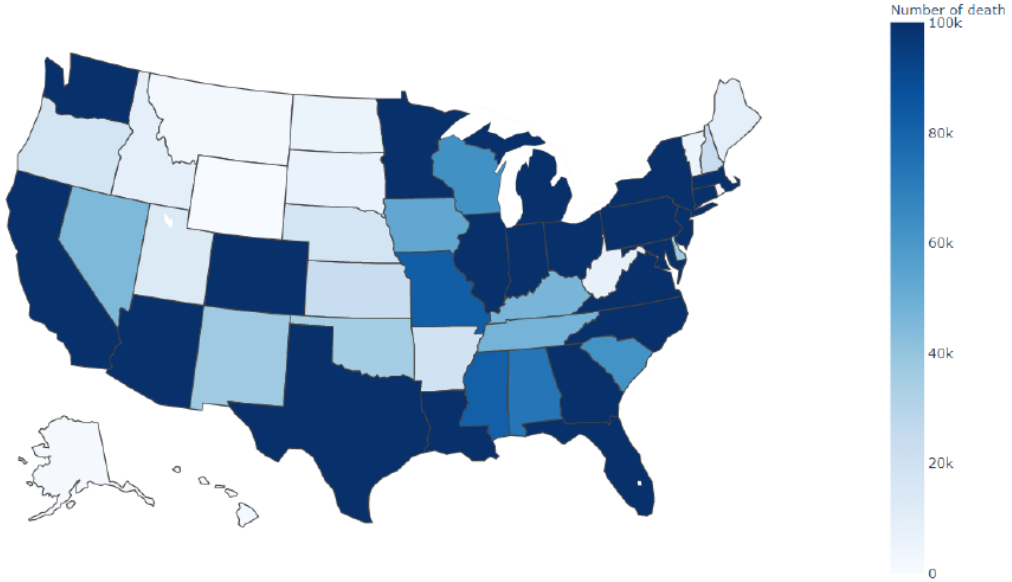
weekly course schedule: after the mid-term exam

11.3	10	**Social Network Analysis**
11.10	11	**Machine Learning: Supervised Learning**
11.17	12	
11.24	13	**Machine Learning: Unsupervised Learning**
12.1	14	
12.8	15	**Data Visualization**
12.15	16	**Final Exam**

Data Visualization

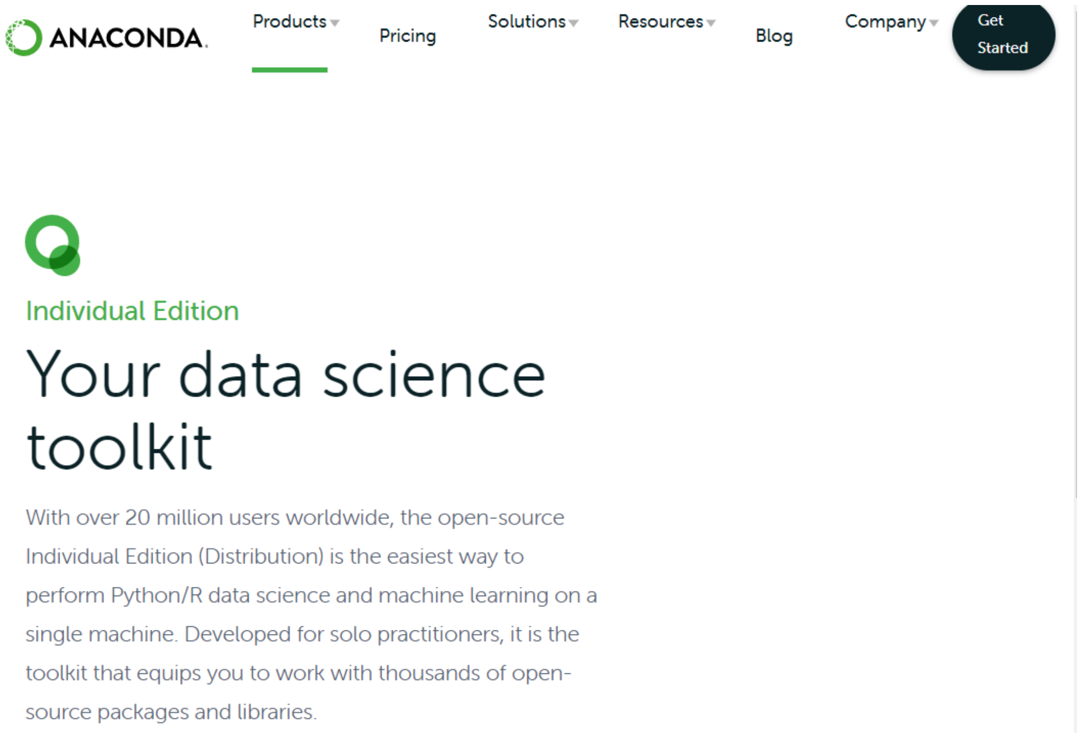
- How well we visualize our data is important in data science.
- We will have some fun visualizing our data, so that you can actually use these graphs for your own purpose!





What you need to do before next class

- Install anaconda!



Individual Edition

Your data science toolkit

With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries.

- Play around with Jupyter Notebook!

Always remember!

I may not be able to answer all questions that students may have during the class.

If you have any questions,
you can ask me via email.

Hope what you learn in this course
is actually useful to you. ☺