



Data Science

Won Kim
2022



Course Objectives

- Learn the Big Data/Machine Learning Process and key elements of each step of the process.
- Learn some Data Mining/Machine Learning Algorithms and Result Evaluation Methods.
- Learn to use Python and Python-based Platforms to carry out a Big Data Project.
- Prepare to take the Machine Learning course and the Deep Learning course.



Course Contents

- 10 Lectures
- 3+ Labs
- 2 Exams
- 2-4 Quizzes
- 3 Programming Homework
- 4-6 Written Homework
- Term Project and Presentation



Course Grading Policy

- Exams
 - mid-term: 150
 - final: 250
- Term Project: 150
- Labs: 100
- Homework: 200
- Attendance 150

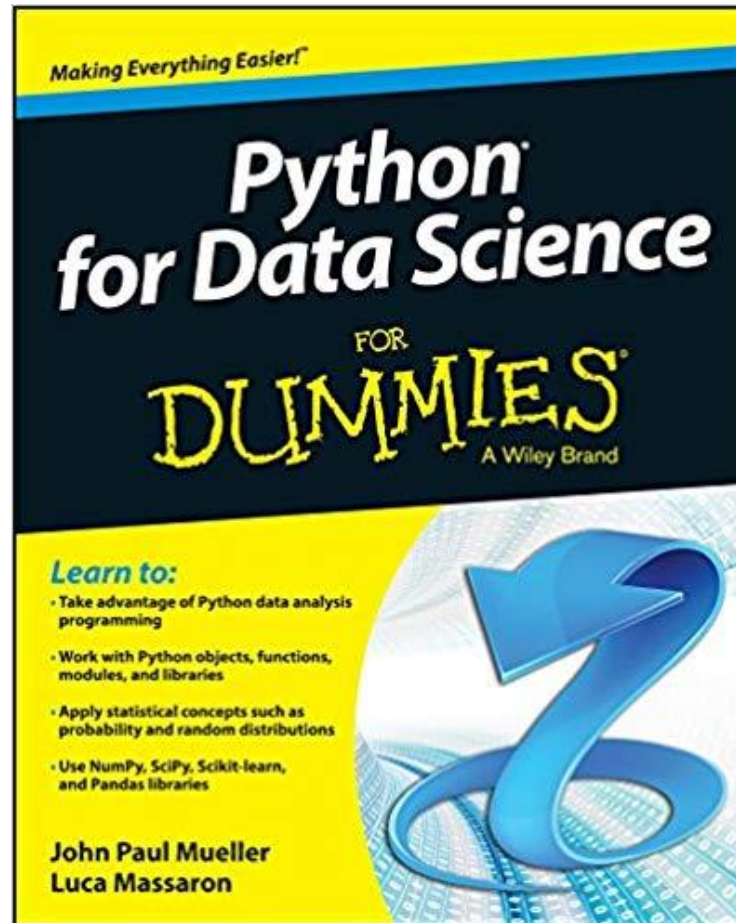


Textbook

- None

Suggested Reference

- **Python for Data Science (for Dummies)**
 - John Paul Mueller and Luca Massaron (John Wiley & Sons, 2015)



Suggested References

- Naver Blog posts
 - “김원의 SW중심세상”
 - Big Data, AI, UIUX, SW교육, 산업



[규제개혁] `개·망·신 3法`...
핀테크 스타트업 10곳중 8곳 과도한 금융규제
성장 막아 '금융결제 시스템 폐쇄적 운영' 관련
금 업체 시장진입 ...
2019. 1. 16.

[UIUX] 지하철 UX 미흡...
지하철 UIUX가 매우 우수하지만, 좀 미흡
만 면도 있습니다. 이 포스트에서는 두 가
지를 살펴 보겠습니다.B...
2019. 1. 10.





About the Course PPT

- Created by studying over 100 PPT/PDF files and blogs/writings on the Internet.
- They are referenced in Acknowledgments.
- Some of them were used with no change or minor changes.
- Many others were used to a small extent.



The Title of This Course Is “Data Science”

- But the scope of this course is broader than just “Data Science”.
- The course is about “Big Data End-to-End Process”.
- This course and the Machine Learning course are two common courses in 4 of the 5 “depth tracks” in the 3rd year curriculum of the School of Computing.
- (e.g.) curriculum of the “Big Data” track
 - 1. Data Science
 - 2. Machine Learning (= Data Mining)
 - 3. Deep Learning (= Neural Networks)
 - 4. Big Data Platforms



Note (1/2)

- This course will be taught by two professors.
- However, the same lecture materials, same homework problems, and same exams will be used.



Note (2/2)

- Our School of Computing (SW Department + AI Department) has adopted self-directed learning methods in all computer-science courses (except 1st year courses).
- Professors will become more of coaches than lecturers.
- We are doing this because life is a continuous learning process, and we believe students will find, after graduation, training in school on self-directed learning very valuable.
- There are two key elements to the self-directed learning.
 - MOOC and Active Learning



MOOC (massively online open courseware)

- At least 3 weeks' classes will be conducted using MOOCs.
- Students can take the classes from anywhere any time (within one week of the posting of the MOOC).
- There will be exercise problems and assignments (due in one week).
- Students can ask questions online and receive answers within 24 hours.
- This method has been used very successfully for 6 semesters for all SW ELITE courses and SW Basic courses for all non-Computer Science students in Gachon University.



Active Learning

- At least 2 weeks' classes will be conducted using an active learning method.
- The instructor will give a brief introduction to one or more topics, and the students will learn them in depth on their own, and then submit a report and present using PPT in the next class.
- ** This learning model has been used successfully for several years in the "Software Industry Seminar" course.
- ** One semester is not really sufficient to teach everything in one course (e.g., Data Science, Machine Learning), and some should be left to students to learn on their own.



SLAM (Software Learning and Assessment Mini-Courses) Platform

- Our School of Computing will build a self-directed and adaptive learning platform (named SLAM).
 - (The plan is to have a working platform by February 2022.)
- Students will be asked to submit good learning materials found in their active learning, and selected materials will be uploaded to SLAM for other students to use later.
- We envision SLAM to be a symbol of Gachon University's education excellence, and strongly encourage all SW/AI major students to contribute to its building.
- The School will reward strong contributions to the building of SLAM (both the software and contents).



Machine Learning MOOC Classes (1/2)

- 5 Topics will be on MOOC.
 - NumPY, Matplotlib
 - Data Preprocessing 2&3
 - Correlation and Regression
 - Clustering Algorithms
 - Evaluation Metrics
- The professors will record different topics, in order to share the load; that is, the professor who records the lecture may not be the same one who teaches the class. But the contents are identical.



Course Outline

- Overview of Data Science
- Big Data End-to-End Process
- Big Data Preprocessing
- Learning Models
- Model Evaluation



Roadmap: This Class

- **Confusion**
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Terminology

- Big Data, Data Mining, Machine Learning
 - Big data means either a Big data itself, or analysis of big data.
 - Big data analysis uses data mining/machine learning techniques
 - Data mining and machine learning are about the same; uses data to extract useful knowledge.
- Machine Learning vs. Deep Learning
 - (broad) Machine learning includes traditional machine learning and deep learning
 - (narrow) Deep learning uses neural network; machine learning does not.



Confusion by Non-Tech People

- “Every organization must do it.”
- “You must have big data (large amount of data) to do it.”
- “This is AI (Artificial Intelligence).”
 - “All you need is big data. The AI software will give you results automatically.”



Confusion by Tech People

- “Big Data is so big, it cannot be processed using conventional technologies.”
 - 3Vs (volume, variety, velocity) – Doug Laney of Gartner Group, 2001
- “For Big Data, you need Hadoop/Spark.”
- “For Big Data, you must use NoSQL.”
 - “You cannot use SQL and RDBMS”
- “For Big Data, you must use Data Mining or Machine Learning.”



Caused By

- Mass media reporting
 - 99%+ of the audience is non-technical
 - (“AI is magic.” “Machine learning is automatic and magic.”)
- Marketing by large tech corporations and marketing industry
 - exaggerates product capabilities, advertises non-existing products, attaches popular terms to existing products,...
- Tech people who don’t know, but bravely make statements based on general knowledge



A Joke About All the Confusion – (Simon Matthews)

- “Everyone talks about Big Data.”
- “But nobody knows how to do it.”
- “But everyone thinks everyone else is doing it.”
- “So everyone says he is doing it, too.”

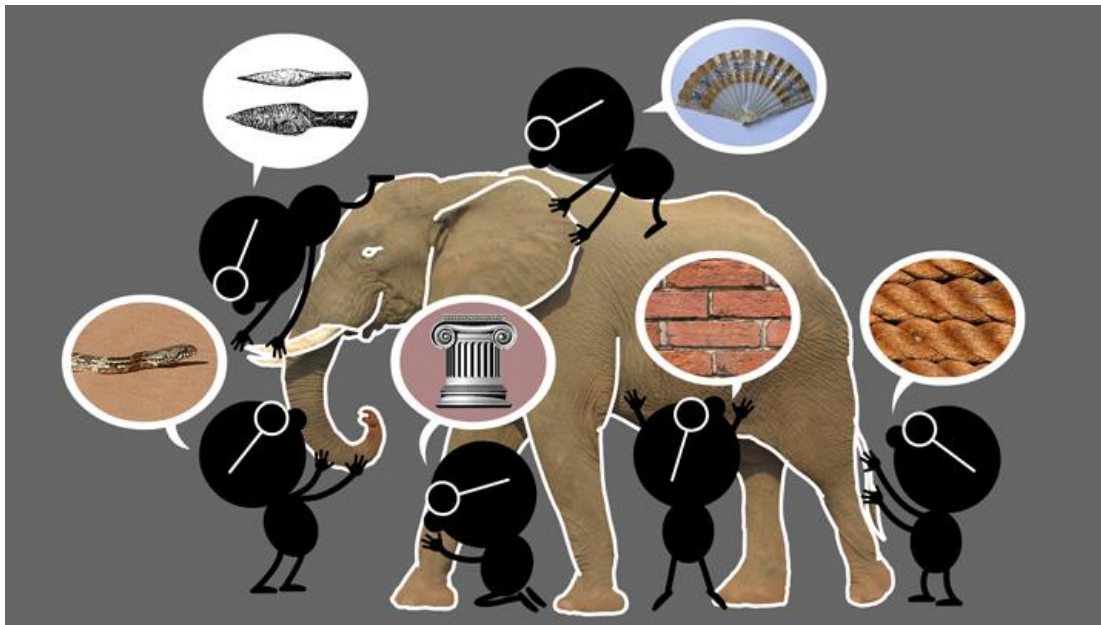


Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check

What Big Data Is NOT

- Is not a breakthrough new technology
- Is not for every organization
- Does not require Hadoop/Spark & NoSQL
- Is not AI, Data Mining, SNS opinion mining, automatic, etc.





Simple but Realistic Definition (1/2)

- Big Data
 - Process and techniques to discover from data answers or insights to solve a business problem
- Data Science
 - Goal is the same as big data
 - But currently focuses on machine learning and deep learning (The scope should expand)



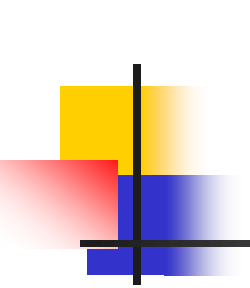
Simple but Realistic Definition (2/2)

- Big Data deals with data in its native form.
- But Machine Learning and Deep Learning require all data to be converted into numerical data and put into a single 2-dimensional array (matrix) form.
- This is because Machine Learning and Deep Learning aim to understand the meaning of data in order to classify or cluster related data. (Big Data does not.)



A Few Things Missing in the Definition?

- **Big Data**
 - Process and techniques to discover from data answers or insights to solve a business problem.
- Why is there no mention of “big” (data)?
- Why is there no “data mining/machine learning” in the Definition?
- Why is there no “Hadoop/Spark” and “NoSQL” in the Definition?



Why There Is No “Big Data” In the Definition (1/2)

- “Big” data is not really new. There were “big” data even before the Web and Internet companies.
- Besides, there is no rule/guideline about “how big” is “Big”, “Middle”, and “Small”.

Some Very Big Data (that have existed before the term “Big Data”) (1/2)

- Governments of China, US, Russia, Japan, ...
 - Census, tax service, military, police data...
- Echelon Spy System (run by the US, UK, Canada, Australia, New Zealand, and Japan)
 - intercepts all email and fax communications globally





Some Very Big Data (2/2)

- Corporations with a huge customer base
 - China Mobile, NTT, NTT Docomo, AT&T, Verizon...
 - WalMart, COSTCO, Target,...
 - (Google, Facebook, Baidu, Apple, Amazon, Zoom, Microsoft...)



Why There Is No “Big” (Data) In the Definition (2/2)

- The most important thing is “helping businesses discover from data answers or insights to solve business problems”.
- This can be done even with “small” data.
 - Can you think about some examples?
- Also, most often, only a small part of “big” data is necessary. (e.g., year 2005 data from the past 30 years data)
- Further, most often, a small sample of “big” data is used (and yields the same result).
 - Can you think about some examples?



Big Data Challenges

- Diverse Data Sources
 - corporate data warehouses, Web, smartphones, communication networks, sensor networks, satellites, ...
- Diverse Data Types
 - alphanumeric, string, semi-structured documents(e.g. email, form), free-form text, photos, images, videos, audios, time series data,...
- Diverse Data Semantics
 - strong typing, relationships, sentiment, new speak
- Performance, Scalability, Availability (24x7) Requirements
 - real-time response; huge number of users



Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- **What are the application areas of Big Data?**
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Famous Application Examples

- Amazon Alexa, iPhone Siri
- Google Translate
- IBM Watson (why did this fail?)
- Facebook targeted ad
- Amazon book recommendations
- Netflix movie recommendations



Application Areas (1/3)

- Business
 - marketing analysis
 - customer segmentation, mailing effectiveness, marketing campaign, pricing
 - customer behavior analysis
 - churn, purchase patterns and profitability, sentiment
 - fraud detection
 - credit card, insurance claims
 - business risk analysis
 - sales analysis
 - (including) related items purchase patterns
 - manufacturing and sales office location analysis
 - employee performance, assignment analysis



Application Areas (2/3)

- Governance
 - election, public opinion analysis
- Medicine
 - DNA sequence analysis
 - medical image analysis
- Security
 - anti-terrorism, crime analysis and prevention



Application Areas (3/3)

- Categorization and Summarization of Text
 - emails, media articles, problem reports, ...
 - Internet search result snippets
- Others
 - prevention of automobile parts simultaneous failures
 - music genre automatic analysis
 - student performance analysis



Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- **What is the Big Data Process?**
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check

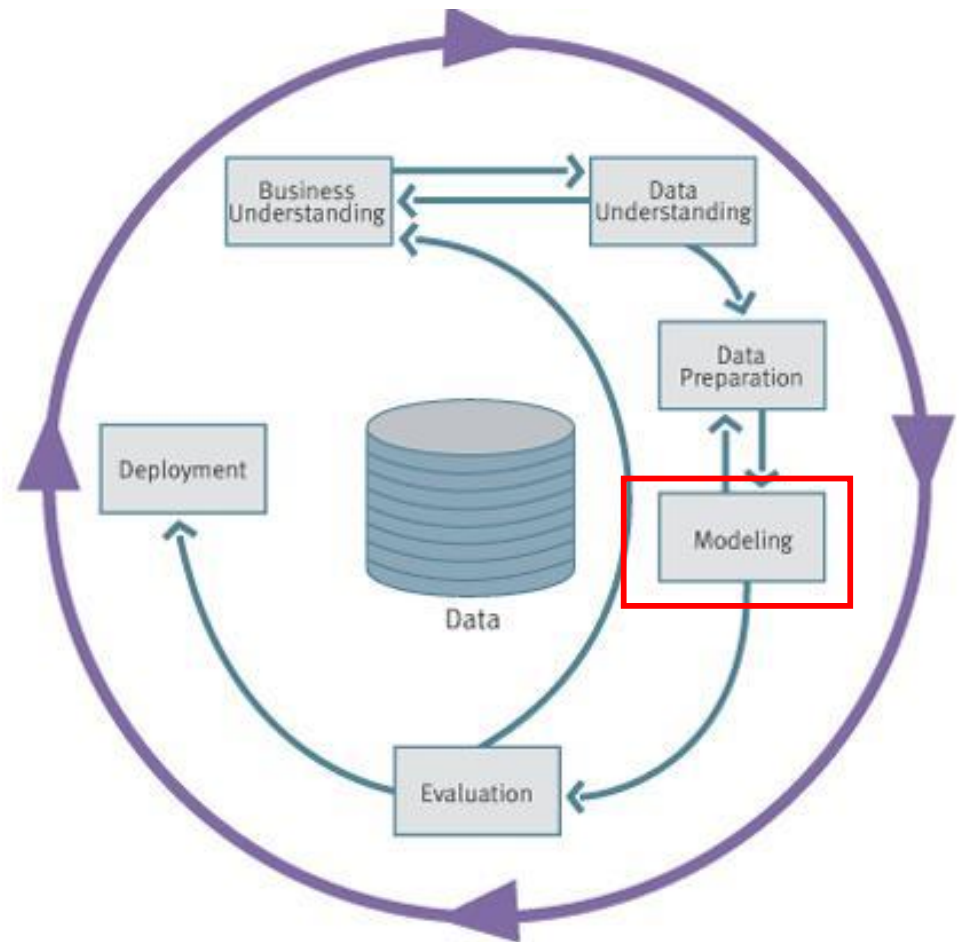


Big Data End-to-End Process

- Objective Setting
- Data Curation
- Data Inspection
- Data Preprocessing
- Data Analysis (Modeling)
- Evaluation of the Results
- Deployment

A Graphical View of the Process

- The term “Modeling” refers to “learning models” in machine learning, and is too narrow..
- The term “data analysis” includes use of machine learning, RDB queries, text documents, images, etc.





Iterative Big Data Process (in Pseudo Code ^^)

```
while (1) {  
    define/tune business needs;  
    if (failure) break;  
    estimate ROI; /* time, manpower, cost */  
    if (failure) break;  
    collect and explore data;  
    if (failure) break;  
    prepare data for analysis;  
    analyze data  
    evaluate results;  
    if (success) break; }  
if (success) {  
    estimate ROI;  
    if (positive)  
        apply results to business;
```



Software Used for Big Data

- Database management system
- Data migration and data warehousing system
- OLAP system
- Data preprocessing software
- Statistical analysis software, Excel
- Data analysis results visualization software
- data mining/machine learning software



Data Preprocessing (Preparation)

- 80% of the time and efforts needed for the entire end-to-end process
- Exploratory data analysis
- Data cleaning
- Data format changes
- Data restructuring (including feature engineering)
- Data value changes
- Data reduction

Significance of Data Preprocessing

- If Olympic 100 meter dash final is the Analysis, everything before the final is Data Preprocessing.
 - all regular competitions and Olympic-qualifying competition
 - Olympic preliminary competitions





Consequences of Poor Data Preprocessing

- Useless Results
 - “Divorced women like purple color.”
 - “Students with good grades live in Seoul.”
- Unnecessary costly computations
 - excessive compute time and resources



Must Evaluate the Results of Analysis

- Algorithms should not be trusted 100%.
- Results of Analysis (e.g., by Machine Learning or Deep Learning) must be combined with and/or cross-checked against other data.
- (example) cross checking against marketing analysis data
 - competitors' activities
 - changes in general economic circumstances
 - occurrence of special events, etc.



Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- **What does Big Data require?**
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Requirements for Big Data (1/2)

- **Business needs**
- Data
 - enough data appropriate for generating the results for the business needs
- Computers and software
 - for storing, preparing, and analyzing data; and validating and interpreting the analysis results
- Team of trained people



Example Business Needs

- Minimize the financial loss (by the card company) due to use of stolen credit card
- Minimize customer churn (by an OTT company)
- Increase the effectiveness of targeted online advertising (by a shirt company)
- Predict product failure (by an automobile company)
- Predict employee performance (by a midsize company)
- ...



Requirements for Big Data (2/2)

- Team of trained people
 - Who can develop the business needs, and apply the results back to the business needs
 - Who understand the meaning of data stored/collected
 - Who can code and use the software for data preprocessing, analysis, and results evaluation
 - Who can protect the security and privacy of data
 - Who can develop data curation and data quality policies and practices
 - Who can manage the big data project and the team



Big Data Is NOT for Every Organization

- Big Data is costly.
- To do big data, policies and procedures must be in place for data curation, data management, and data quality control.
- Organizations that do Big Data in-house need to satisfy the following conditions.
 - expect great benefits from big data
 - the benefits are much greater than the cost, and
 - have strong need to protect their data
- All others will either outsource big data to consulting companies, or not do big data at all.



A Short Discussion

- There is a big restaurant. The owner wants to change the menu. Would this require a big data analysis? If so, what kind of data would be required? How can such data be obtained?



The Limited Role of Big Data

- Big Data can provide answers and insights in solving a business problem.
- However, usually it has a limited role in solving the business problem.
- Often business management, investment, etc. are required to solve the business problem.
- Example: to stop customer churning
 - Change pricing policy, develop a new product, acquire a company, etc.



Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Existing Technologies

- file systems, database systems
 - OLAP, data warehouses
 - report generation, decision support
 - dashboard, visualization
 - statistical analysis, data mining/machine learning
 - information retrieval
 - text processing
 - image processing
-
- **** Data mining/machine learning software are only a part of these technologies**



Data Mining/Machine Learning Algorithms

- Classification Algorithms
 - decision trees, Bayesian networks, support vector machines, k-nearest neighbors, neural networks
- Regression Algorithms
 - linear regression, multiple linear regression
 - non-linear regression
- Clustering Algorithms
 - k-means, EM (expectation maximization), DBSCAN, OPTICS, hierarchical agglomerative clustering, ...
- Market Basket Analysis (Association Rules Discovery)
 - apriori algorithm, FP tree



Types of Data That Can Be Analyzed

- formatted alphanumeric data
- semi-structured data (email, form that can include alphanumeric data, short text)
- free-form text
- time series data
- images/photos, video
- sound, audio
- ...

Data Presented in Table Form

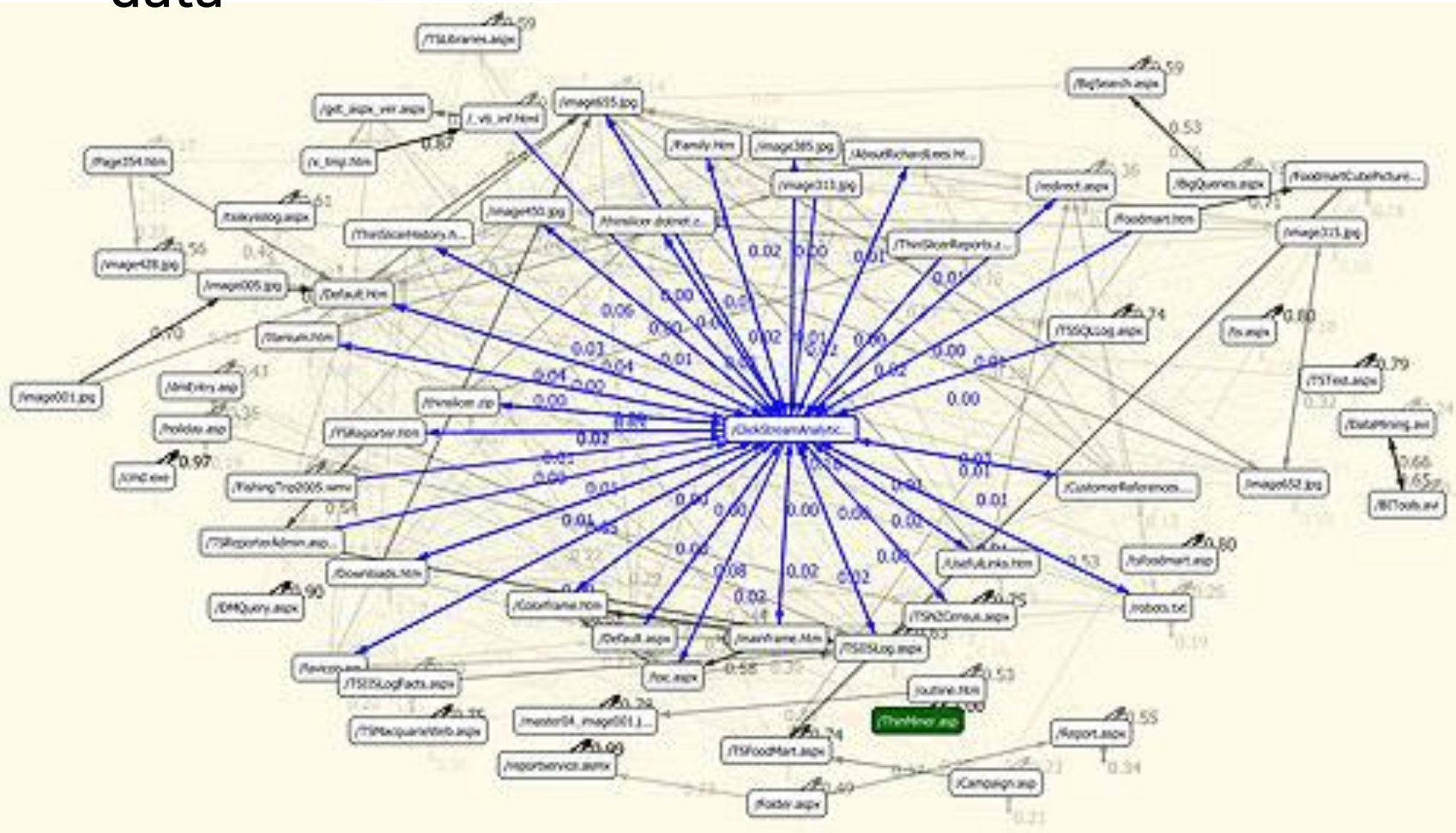
Main products	Finished products		Parts and components	
Electronics				
Office machines	7511	Typewriters, cheque-writing machines	7591	Parts of and accessories suitable for 7511, 751.8
	7512	Calculating machines, cash registers	7599	Parts of and accessories suitable for 751.2, 752
	7518	Office machines, n.e.s.		
Automatic data processing (ADP) machines	7521	Analogue & hybrid data processing machines	7599	Parts of and accessories suitable for 7512, 752
	7522	Complete digital data processing machines		
	7523	Complete digital central processing units		
	7524	Digital central storage units, separately consigned		
	7525	Peripheral units, incl. control & adapting units		
	7528	Off-line data processing equipment n. e. s.		
Television, radio-broadcast receivers, gramophones and telecom equipment	7611	Television receivers, colour	7649	Parts of apparatus of 76 (including TV, radio, gramophones and telecom equipment)
	7612	Television receivers, monochrome		
	7621	Radio-broadcast receivers for motor vehicles		
	7622	Radio-broadcast receivers portable, incl. sound rec.		
	7628	Other radio-broadcast receivers		
	7631	Gramophones & record players, electric		
	7638	Other sound recorders and reproducers		
	7648	Telecommunications equipment		
Thermionic, cold & photo-cathode valves (semiconductors)	7761	Television picture tubes, cathode ray	7768	Piezo-electric crystals, mounted, parts of 776
	7762	Other electronic valves and tubes		
	7763	Diodes, transistors, similar semi-conductor devices		
	7764	Electronic microcircuits		
Automotive				
Automobiles	7810	Passenger motor cars, for transport of passengers & goods	7841	Chassis fitted with engines for motor vehicles
	7821	Motor vehicles for transport of goods/materials	7842	Bodies for the motor vehicles of 722/
	7822	Special purpose motor lorries and vans	781/ 782/ 783	Other parts & accessories of motor vehicles
	7831	Public-service type passenger motor vehicles etc.		
	7832	Road tractors and semi-trailers		
Car engines	7132	Internal combustion piston engines for propelling vehicles	7139	Parts of internal combustion piston engines

Source: UN Comtrade.

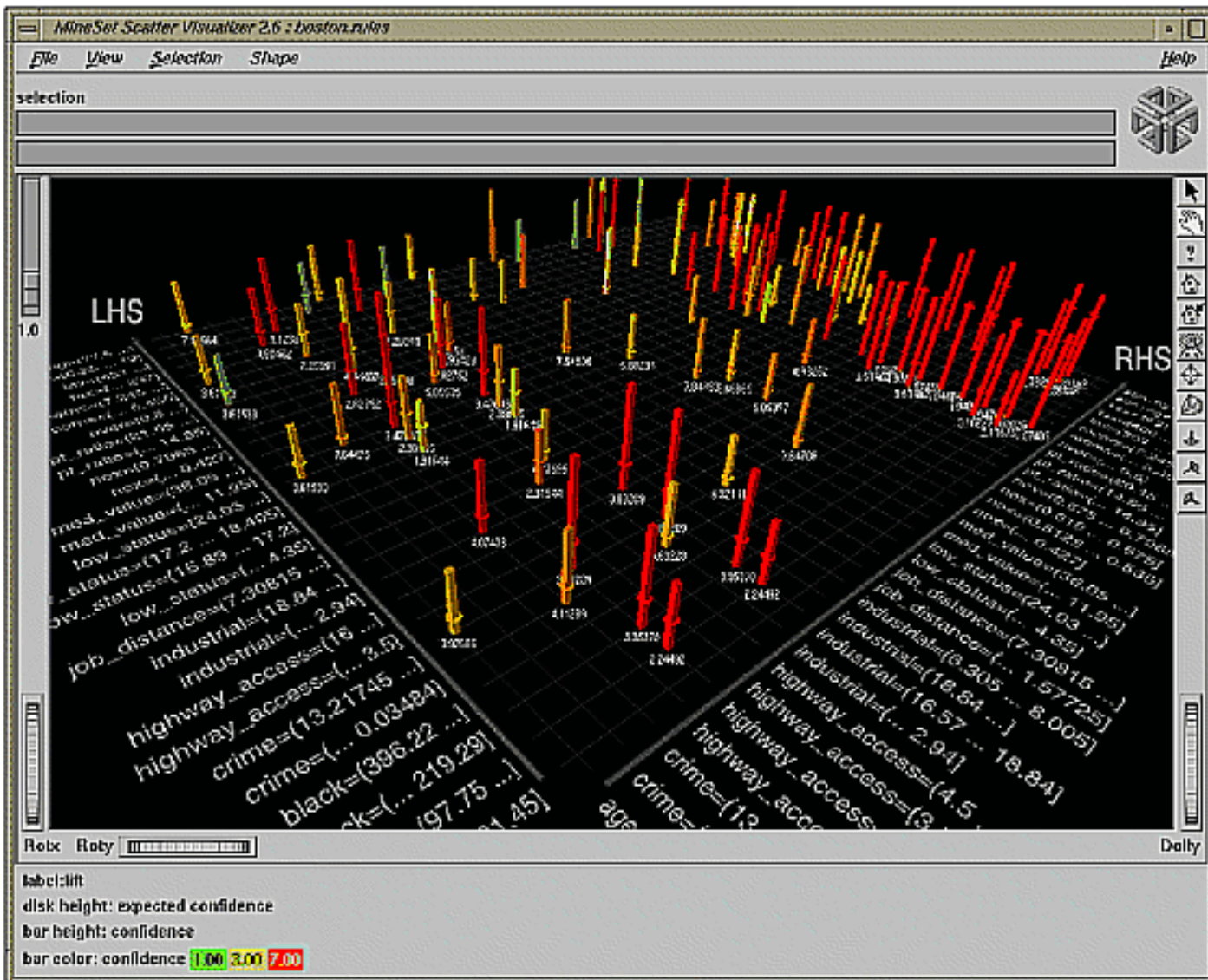
Note: The names of the items are taken directly, with some abbreviation, from SITC Rev 2.

Data Visualization

- Graphically represents complex relationships among data



Visualization – Association Rules



Visualization – Dash Board



D/BM Online Monitor

DM - D/BM - D/BM Blog - Support / Contact - About Online Monitor





Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Big Data Platforms

- What is a big data platform?
 - Software to store and manage data, develop big data applications (analytics), and run the applications
 - Python and R libraries
 - RDB, data warehouse
 - Open source software frameworks
 - NoSQL, Apache Hadoop/Spark



Python Libraries

- NumPy – linear algebra and array processing
- Pandas– data preprocessing and exploration
- Matplotlib (Seaborn) – data exploration by visualization
- Scikit-Learn – for machine learning
- TensorFlow (Keras), Pytorch – for deep learning



NoSQL

- Limited database management system
(accurate name should be) Web Data Store
- About 150, developed for Web applications,
and available as open source software
- Different from RDB
 - non-normalized relations (collections), and eventual consistency
- Focus on performance, scalability and
availability in a distributed computer network



Some NoSQL Systems – Moving Targets

- Structured Row-Based Data Stores
 - (Google) Big Table, MegaStore, Spanner
 - (Amazon) SimpleDB
 - MongoDB, CouchDB, Terrastore
(provides some ad hoc query facilities)
 - Hbase (Big Table), Cassandra
- Key-Value Stores (* not really a database *)
 - (Amazon) Dynamo, Voldemort, Riak, membase, membrain, memcached, Amazon S3



Co-Existence of RDB and NoSQL

- Can use both platforms
 - relational database platform
 - Hadoop/NoSQL platform
 - Make use of map-reduce engine
- Hadoop Sqoop
 - import of relational data to Hadoop
 - export of Hadoop data to relational database



Roadmap: This Class

- Confusion
- What is Big Data (Data Science)?
- What are the application areas of Big Data?
- What is the Big Data Process?
- What does Big Data require?
- What technologies can be used for Big Data?
- What are Big Data platforms?
- Reality check



Machine Learning/Deep Learning Is NOT Magic

- (People have to do all the work of data preprocessing, before using the machine learning/deep learning algorithms.)
- People have to determine the business needs.
- People have to acquire data appropriate for the business needs, and explore the data.
- People have to do data preprocessing to put the data in a form that the machine learning algorithms require.
- People have to select the machine processing algorithms appropriate for the data and the problem to solve.
- People have to evaluate the results, and experiment and tune parameters and data repeatedly to achieve accurate results.



Written Homework 1: Big Data Story

- Study the IBM Watson big-data application, and answer the following.
 - What is it? What is its business objective?
 - Describe the types and amount of data used.
 - Describe the technologies used (big data, machine learning, deep learning).
 - Is it a business success or failure? What is the reason for its success/failure?
 - If it is a failure, what is the lesson to be learned from it?



Written Homework 2

- Find and briefly describe 1 big data application of your choice.
- You may choose any large corporation or a government branch that has a huge amount of data. (The following is just a hint.)
 - Google, Facebook(Meta), Apple, Microsoft, Netflix, Zoom, NC Soft, WalMart, CostCo, China Mobile, Verizon, T-Mobile, Wikipedia,...
- Describe the types and amounts of data it has.
- Describe how the organization uses the data to increase its revenue/profits/customer satisfaction.



Written Homework: Spec

- Purpose: to have you think critically and dig deeper, rather than believing everything you read on the Internet.
- Submit: post to the CyberCampus in PPT format (not WORD or PDF)
- Due: 9:00 p.m. the day before the next class
- ** Late homework will not be accepted. Don't live dangerously. Try to submit the homework by around 8:00 p.m. to prevent a mishap.)



Individual Programming Homework: Python Basic

- Submit a single WORD file containing the source code and the output screen capture for both problems.
- Due: post to the CyberCampus by 9:00 p.m., the day before the next class.
- (* It would be best to put the Written Homework PPT file and the Programming Homework WORD file into a single zip file. *)
- (* the file name should be just your name. Nothing else.)



Ex-1. Check Prime Number

- Define a function *isPrime* with a parameter n
 - Check if n is a prime number
 - Divide n by all integer i in $[2, \sqrt{n}]$, and check if the remainder is zero
 - If zero for any i , n is not prime, then print i and return
 - If not zero for all i , n is prime
- Write a program using isPrime
 - Read an integer number n in $[2, 32767]$
 - Call isPrime with n to check prime



Ex-2. Dictionary Generation

- Define a function *makeDict* with two parameters K & V
 - K is a tuple of unique subjects (keys)
 - e.g., ('Korean', 'Mathematics', 'English')
 - V is a tuple of (possibly redundant) numeric points (values)
 - e.g., (90.3, 85.5, 92.7)
 - K and V have the same size s
 - Make a dictionary D of size s with K and V
 - Each entry in D is made of the key and the value in K and V
 - e.g., {'Korean': 90.3, 'Mathematics': 85.5, 'English': 92.7}
 - Return D
- Write a program
 - Make two tuples K and V and call *makeDict* to get D
 - For every key in K , check if the value obtained from D is correct



End of Class
