Everything Starts with Data

Week 4 64/1

Announcement

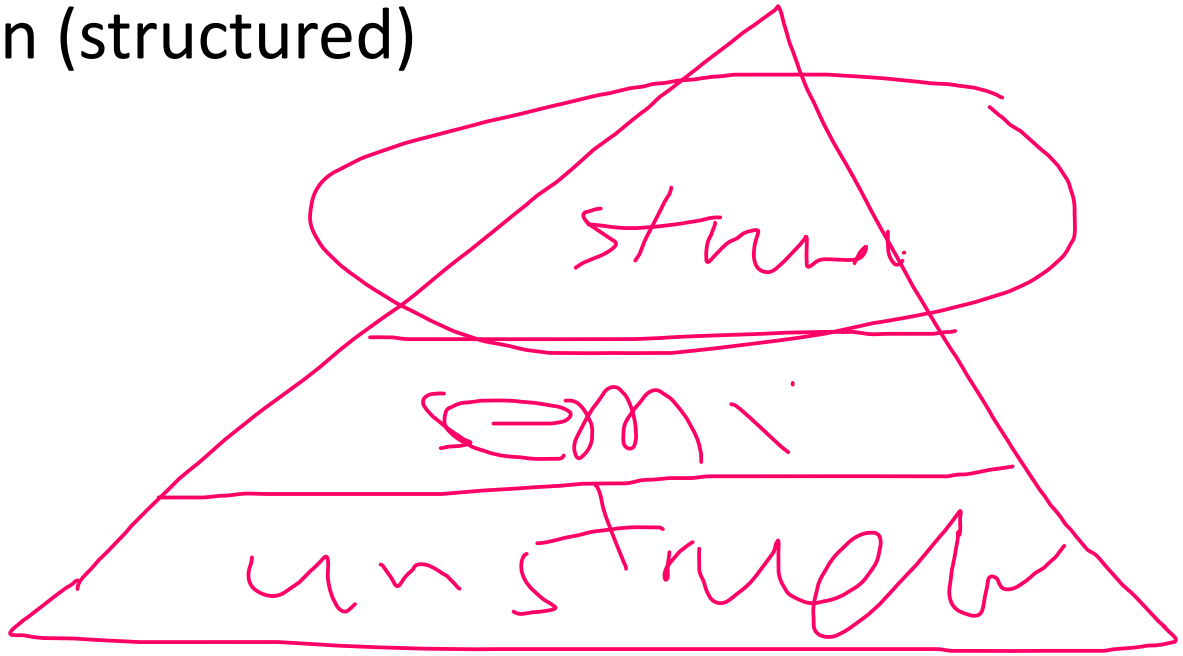
- Double check google classroom
  - Make sure you use actual first name and last name
  - When you submit any work, please write your SID, firstname and lastname.

Where are we now?

Week	Topics
1	Data Scientist Foundation
2	Basic data analytics: KDD
3	Basic data analytics: Data to Data Product
4	What is Data (Str - eg. nomi, unstr - img, text)
5	Dataset (Basic manipulation)
6	Data quality (e.g., outlier, inconsistency, duplication, etc.)
7	Processes - history, e.g., turn kdd to crisp-dm -> modern
8	Processes - in action
Midterm	

Agenda

- Unstructured data(i.e., basic image, text)
- Basic data manipulation (structured)



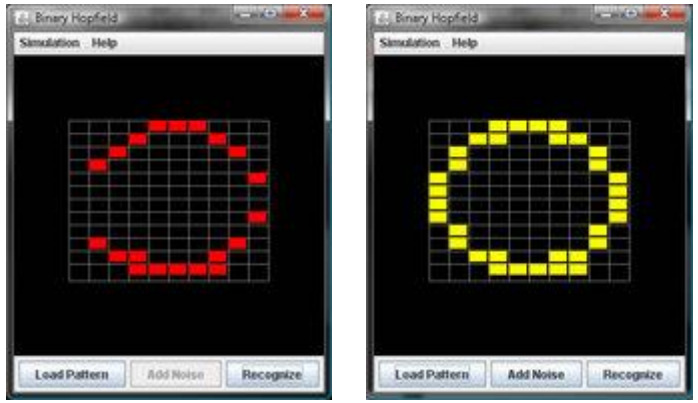
Unstructured data- Image data

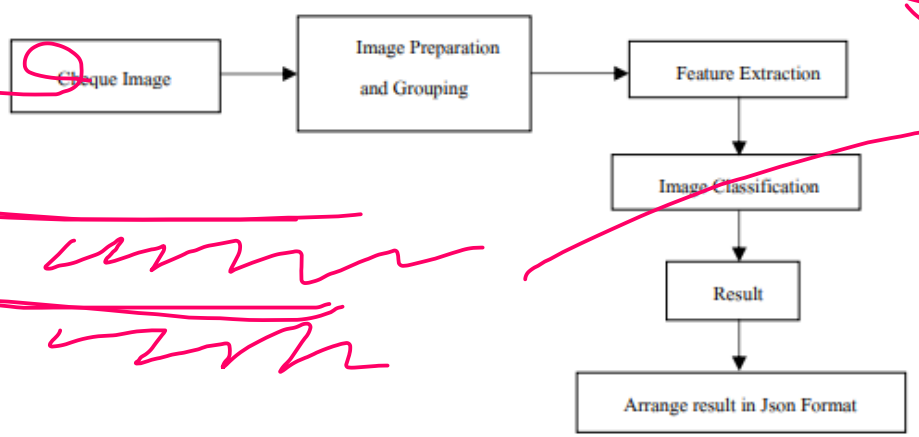
- http://cmtourism.org/ds/CSC5542/p5
- Binary Hopfield Network to recognize image

Fig. 2. An example results of morphology method

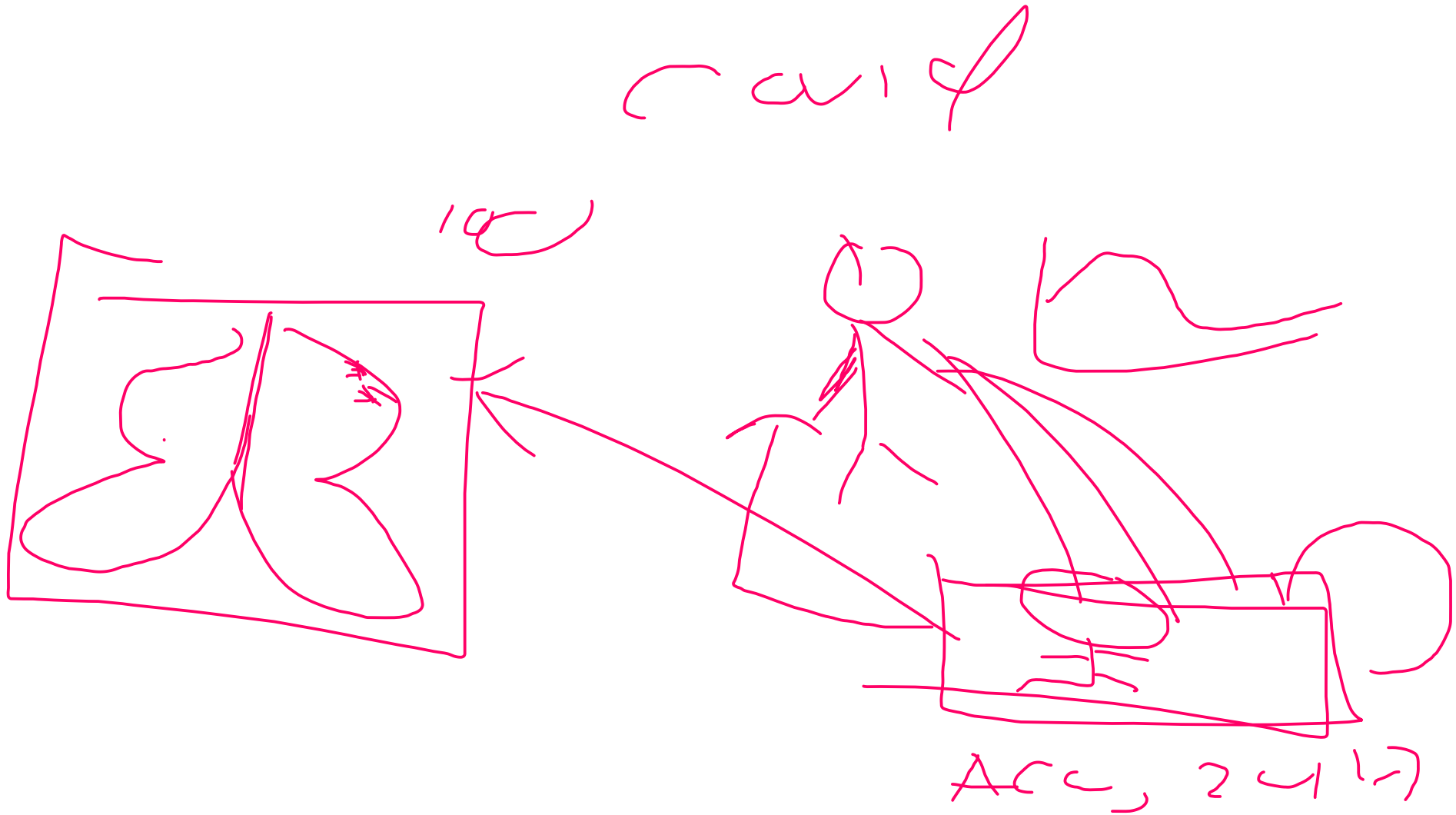
Table 1. Parameter sets for Hyperopt [11].

Classifier	Hyperparameter Set
KNN	Number of neighbors = {1,3,5}, Metric = { euclidean, manhattan }
SVM	C = {1,10,100,1000}, Kernel = { linear, rbf }
GBM	Number of estimators = { 2,4,8,…,256 }, Learning rate = {0.001,0.01,0.1}





C. Ngamsom, P. Phannachitta(2020)
https://datascience.cmu.ac.th/storage/articles/16.pdf



Unstructured data - Textual data

- Fact- 80% of data from internet in textual data
- Speech, text databases, etc.
- **Linguistic approach** – syntax, morphology, semantic, analysis stylistics, etc.
- **NLP approach** – Taggers, parses, spell checkering, spell correction, stop word lists, etc.

WebCrawler / scraper

- https://www.tripadvisor.com/
- Beautiful soup
- Scrapy
- Xpath

WebCrawler / scraper

- Requirements
  - Name of the products
  - URL of the product
  - Item code
  - Nutrition detail per 100g
    - Energy in kilocalories
    - Energy in kilojoules
    - Fat
    - Saturates
    - Fibre
    - Salt

WebCrawler / scraper

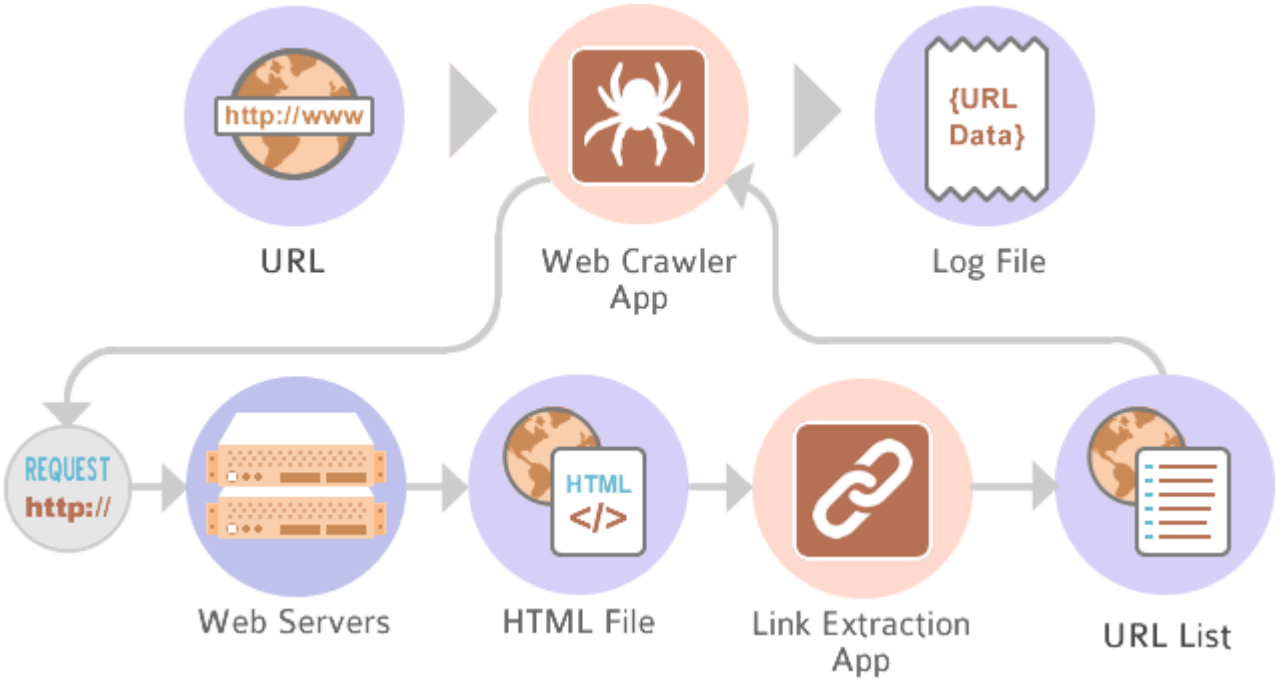


Image source: https://www.geekboots.com/story/what-is-web-spider-and-how-does-it-works

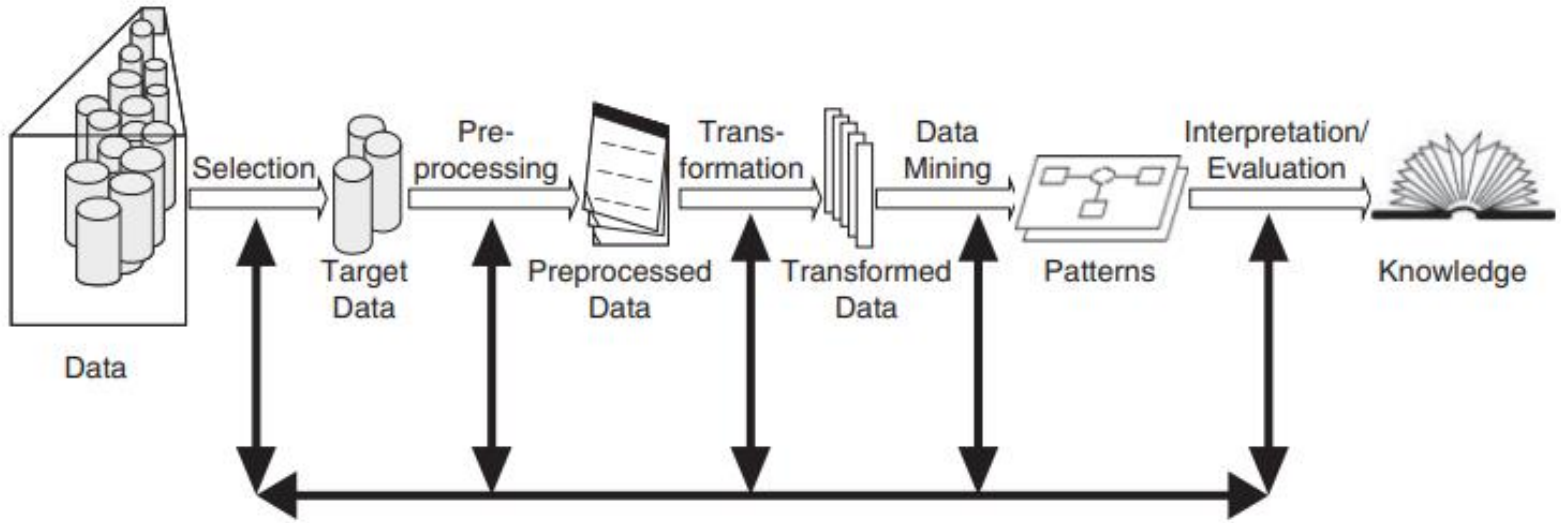


Figure 5 Overview of the steps constituting the knowledge discovery in databases (KDD) process (Fayyad *et al.*, 1996b)

Recall Processes of KDD

1. Learning application domain (initial selection)
2. Data Cleaning
3. Data Integration – where multiple data sources may be combined (heterogenous info. sources)
4. Data transformation
5. Data reduction / feature selection
6. Selecting function of data mining/ml
    1. Prediction/ classification/ associate / clustering

Recall Processes of KDD (Cont.)

7. Selecting the mining / machine learning algorithms
   Depends on the 6 step
8. Evaluation of the data mining/ml algorithm
9. Result interpretation – visualization of the model, main finding, etc.
10. Action (use of discover knowledge -> public policies, intelligent systems)

Data cleaning

- 60-70% of the time spending on cleaning data in the Data Mining processes
- “57% of data scientists regard cleaning and organizing data as the least enjoyable part of their work and 19% say this about collecting data sets” (Forbes, 2016)
- Ignored the sample/case and variables/features
  - case that contain more than 15 % of miss values should be ignored.
  - Variables missing at least 10 % of data were candidates for deletion
  - Ignore the sample, usually perform when target class is missing

Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#1852a116f637, (Accessed, August 2018)

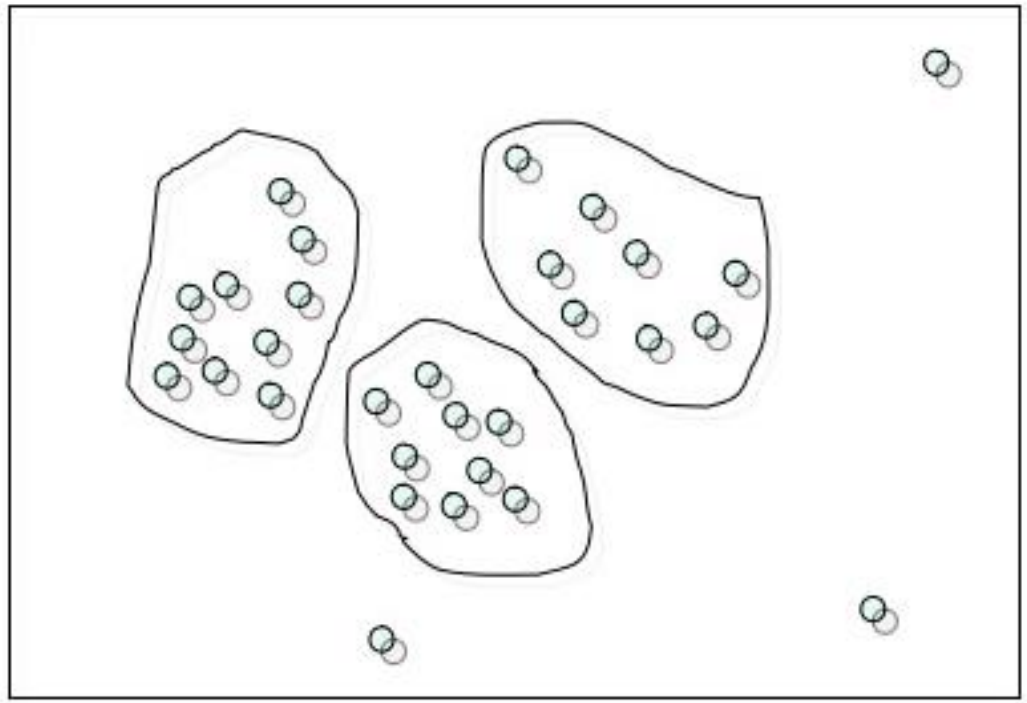
Data Cleaning – (cont.)

- Variables that are Missing At Random (MAR)
  - Use imputation methods
    - Mean or mode substitution (easy to implement)
- Identify outlier and extreme values
    - Binning approach – the most basic technique (sort data to equal bin, then smooth by mean or median
    - Semi-Automated approach - Automate script and domain expert to correct inconsistent data.
    - Clustering approach – using clustering algorithm to group common values,
- Use domain knowledge expert to correct the missing value
  - E.g. Let the weather climate expert comes to check those values.

Binning approach – Smooth data

- Sort data : age = [3,7, 8, 13, 22, 22, 26, 22, 26, 28, 30, 37]
- Partition into equal-depth / equal frequency bins: bin_number = 3
  - Bin1 = [3, 7, 8, 13]
  - Bin 2 = [22, 22, 22, 26]
  - Bin 3 = [26,28, 30 27]
- Smooth by mean: bin1= [10, 10, 10, 10] , bin 2 = [23, 23, 23, 23], bin3 = [30, 30, 30, 30]
- Smooth by bin boundaries: bin1 = [3, 3, 3 13], bin2 = [22, 22, 26,26], bin3= [26,26,26, 37]

Data Cleaning – using Cluster analysis (Identify outliers/extreme values)



Data transformation

- Normalization
  - Scaling attribute values to fall with in a specified range (binning again)
  - Contract or replace with new attribute
    - e.g. measure 3 times, we construct average of the three column
    - e.g. we replace Celsius to Fahrenheit
    - Replace target variable majority vote
  - Dealing with derived attribute (e.g. date of employee)

Data transformation -Continuous value

Data transformation (cont.)

- Normalization using
  - Decimal Scaling (for NN, SVM) – move the decimal point of value of attribute
  - Min-max function (for NN, SVM) – move the attribute value in the specific range.
  - Select normalization techniques depends on machine learning algorithm and nature of data set (try and try till you get good results)

$$s' = \frac{s - Min}{Max - Min} \qquad z = \frac{x - \mu}{\sigma}$$

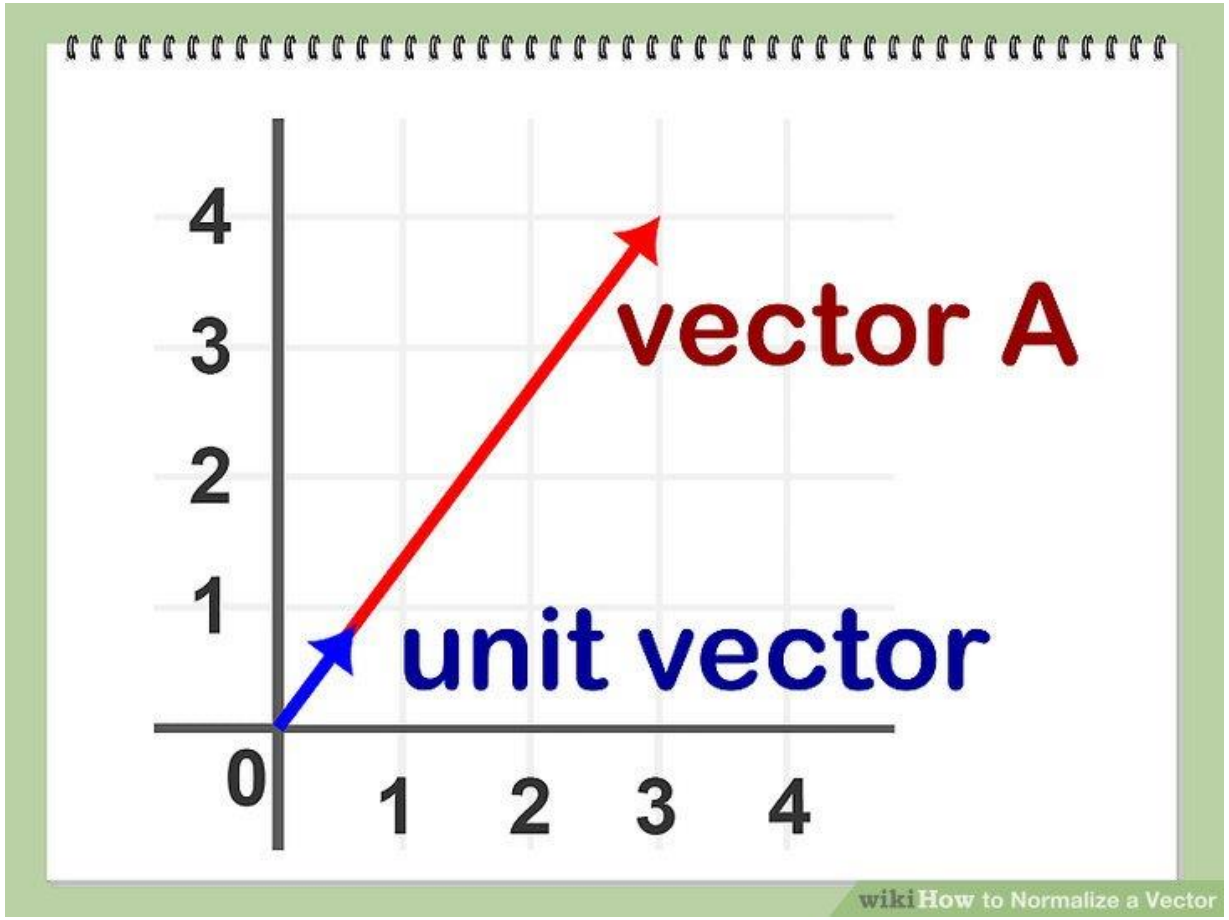
Decimal scaling normalization

- Suppose that the recorded values of x range from − 986 to 917.
- The maximum absolute value of x is 986.
- To normalize by decimal scaling, we therefore divide each value by 1,000
- so that −986 normalizes to −0.986 and 917 normalizes to 0.917.

Min-max normalization

- Suppose that the minimum and maximum values for the feature income are 12,000 and 98,000, respectively. We would like to map income to the range 0.0,1.0 . By min-max normalization function. a value of $73,600 for income is transformed to:
-
$$s' = \frac{s - Min}{Max - Min}$$
- $\frac{73600-12000}{98000-12000}(1.0 - 0) + 0 = 0.716$

Unit Vector normalize



When your feature have large range,
e.g. dot product can return and overflow,
So, scaling the vector can be benefit.

$$\vec{u} = \frac{\vec{v}}{|\vec{v}|} = \frac{(3,4)}{\sqrt{3^2+4^2}} = \frac{(3,4)}{5} = \left(\frac{3}{5}, \frac{4}{5}\right).$$

Data transformation-Discrete value

Data transformation- encoding

- Nominal features (one-of-n encoding)
- Ordinal features (Thermometer encoding)

	1 d4_23	2 a3_5	3 e1_2_5	4 d2
1	0	0	3	8
2	0	0	3	1
3	0	0	2	8
4	0	0	2	1
5	0	0	3	8
6	0	0	2	8
7	0	0	2	8
8	0	0	1	8
9	1	0	3	8
10	0	0	1	8
11	1	0	3	1
12	1	0	1	8
13	0	1	3	1
14	1	1	3	8
15	0	NaN	3	1

	1 d4_23	2 a3_5	3 e1_2_5_v1	4 e1_2_5_v2	5 e1_2_5_v3
1	0	0	0	0	1
2	0	0	0	0	1
3	0	0	0	1	0
4	0	0	0	1	0
5	0	0	0	0	1
6	0	0	0	1	0
7	0	0	0	1	0
8	0	0	1	0	0
9	1	0	0	0	1
10	0	0	1	0	0

One-of n/one-hot encoding

- Categorical variables need to be converted into forma that could provided machine learning algorithms to perform better

Compay name	Type	Price
BMW	1	220,000
FORD	2	780,000
Toyoya	3	670,000
Toyoya	3	640,000

CN_1	CN_2	CN_3	Price
1	0	0	220,000
0	1	0	780,000
0	0	1	670,000
0	0	1	640,000

Thermometer Encoding

Compay name	Type	Price
BMW	1	220,000
FORD	2	780,000
Toyoya	3	670,000
Toyoya	3	640,000

CN_1	CN_2	CN_3	Price
0	0	1	220,000
0	1	1	780,000
1	1	1	670,000
1	1	1	640,000

Data transformation – LibSVM format

- Every data mining software require specific data format in order to use in data mining processes.
- Transform data to the LibSVM format.
- <target> <index 1>:<value 1> <index 2>:<value 2>...<index n>.

```
-1 3:1 11:1 14:1 19:1 39:1 42:1 55:1 64:1 67:1 73:1 75:1 76:1 80:1 83:1
-1 3:1 6:1 17:1 27:1 35:1 40:1 57:1 63:1 69:1 73:1 74:1 76:1 81:1 103:1
-1 4:1 6:1 15:1 21:1 35:1 40:1 57:1 63:1 67:1 73:1 74:1 77:1 80:1 83:1
-1 5:1 6:1 15:1 22:1 36:1 41:1 47:1 66:1 67:1 72:1 74:1 76:1 80:1 83:1
-1 2:1 6:1 16:1 22:1 36:1 40:1 54:1 63:1 67:1 73:1 75:1 76:1 80:1 83:1
-1 2:1 6:1 14:1 20:1 37:1 41:1 47:1 64:1 67:1 73:1 74:1 76:1 82:1 83:1
-1 1:1 6:1 14:1 22:1 36:1 42:1 49:1 64:1 67:1 72:1 74:1 77:1 80:1 83:1
-1 1:1 6:1 17:1 19:1 39:1 42:1 53:1 64:1 67:1 73:1 74:1 76:1 80:1 83:1
-1 2:1 6:1 18:1 20:1 37:1 42:1 48:1 64:1 71:1 73:1 74:1 76:1 81:1 83:1
+1 5:1 11:1 15:1 32:1 39:1 40:1 52:1 63:1 67:1 73:1 74:1 76:1 78:1 83:1
```

Data transformation – Weka format

```
% This is a toy example, the UCI weather dataset.
% Any relation to real weather is purely coincidental

@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

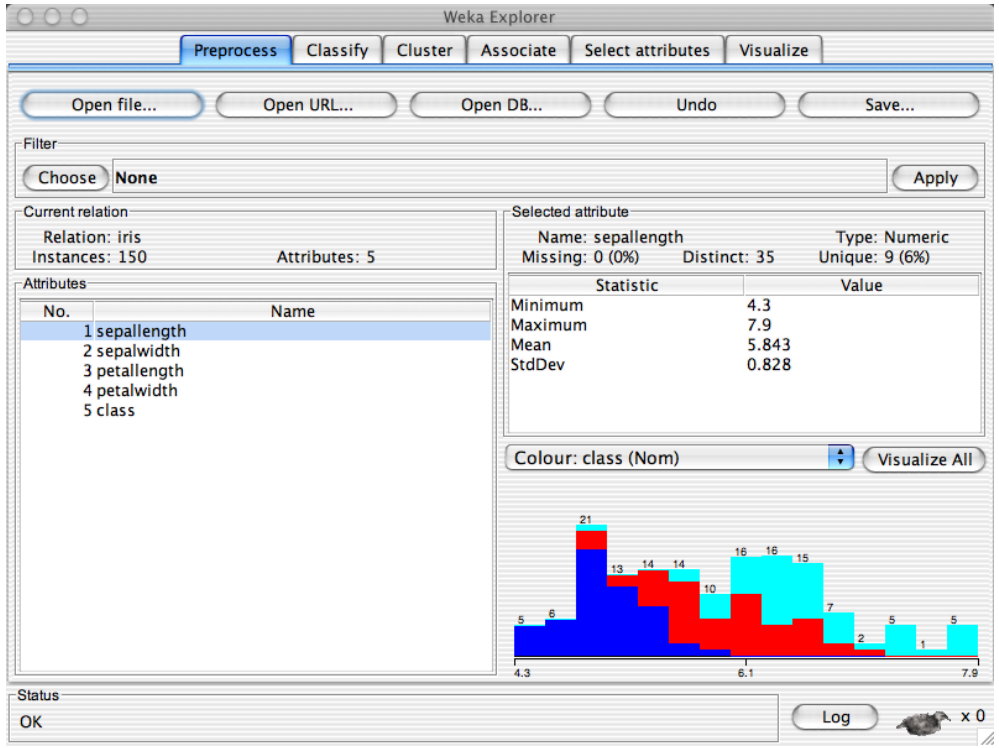
Comment

Dataset name

Attributes

Target / Class variable

Data Values



Data reduction

- Reducing the number of attributes/features
  - Domain knowledge
  - Data Discretization
  - Feature selection method
    - Removing irrelevant attributes
    - Removing redundant attributes
  - Dimensionality reduction: create new features that are a combination of the older features
    - Principle **C**omponent **A**nalysis (PCA) – search for the lower dimensional space that can best represent the data
  - Sampling – select a subset of sample in data to be analyzed.

Data Discretization

- Sub set of data reduction, it diminishes data from large of domain of numerical to categorical values.
- To convert a large number of values into smaller one, so constructing model can **faster** (learning speed)
- Many machine learning algorithms expected discrete values as input - , ID3 decision tree, or Mutual Information features selection algorithms.
- Usually lead to a loss of information

Data Mining/ML Techniques

- Rule-based – the extraction of useful if-else from database on statistical significance
- Nearest Neighbor – A simple classification techniques that classifies a new record based on the most similar to it in a historical database
- Linear Regression – y=mx+b
- Decision Tree – graph or tree like model, provide all possible choices.
- Support Vector Machine – use of hyper plane to separate data points in 2 to n dimensions.
- Artificial neural networks – non linear predictive models mimic our biological neural networks.
- Ensemble learnings
- Etc.

Data Analytical tools

The Most popular data science tools for techies (Kaggle 2020, bigdata.black)

Courtesy of bigdata.black

Data Mining – Introduction (cont.)

- In Business, SAS + Taradata are dominated
- SAS Enterprise Miner
- Teradata Analytic
- Data scient spent (1-2 week on data)
- Less time in a day

Why SAS?

- SAS = Statistical Package + DBMS + Programming Language
- Easy to use – drag and drop component, no coding parts. (e.g. Window vs Linux)
- Graphical Capabilities – so much visualization
- Advancement in tools – keep update, beat almost other tools.
- Job scenario – 72% of analytical market use SAS

Why not SAS for this course?

- No money
- Most data mining tools are commercial tools
- The most popular and free (open source) is Weka

Data Mining Software (open source)

- Weka - https://www.cs.waikato.ac.nz/ml/weka/
- Rapid miner - https://rapidminer.com/

What is Weka?

- Waikato Environment for Knowledge Analysis
- It’s a data mining/ machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand
- It is an open source software issued under the GNU General Public License.
- Pure Java

Weka Installation

- Install JRE
- Download from
- https://www.cs.waikato.ac.nz/ml/weka/downloading.html
- Current version is 3.8
- Support multiple platform: Window, MacOS, Linux

Weka features

- 49 data preprocessing tools
- 76 classification/ regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 10 search algorithms for feature selection

Weka GUI

- Three graphical type of user interface
  - The Explorer (exploratory data analysis)
  - The Experimenter (experiment environment)
  - The KnowledgeFlow (new process model inspired interface)
  - The Workbench
  - Simple CLI (Command prompt)

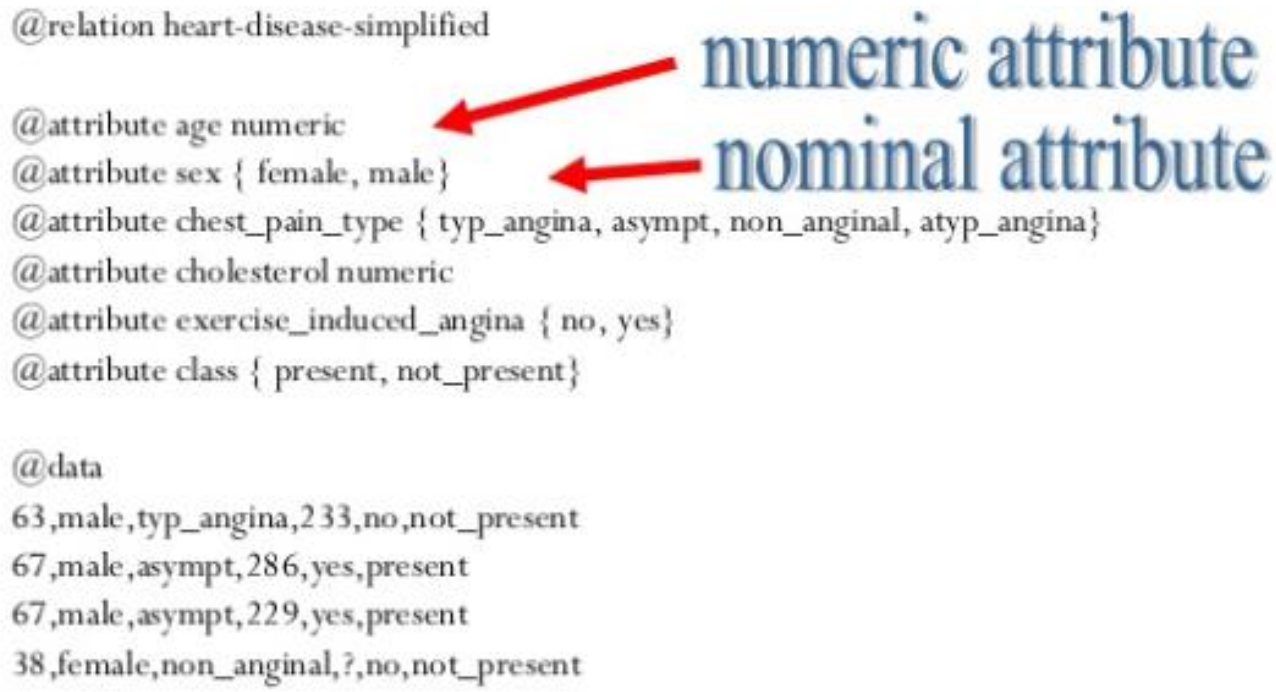
ARFF Files

- Weka has its own file format
- A dataset has to start with a declaration of its name
  - @relation name
- @attribute attribute_name specification
  - If an attribute is nomimal, specification contains a list of the possible attribute values in curly brackets:
  - @attribute nominal_attribute {first, second, third}
  - If an attribute is numeric, specification is replaced by the keyword numeric:
  - @attribute numerica_attribute numeric
- After the attribute declarations, the actual data is introduced by a tag:
  - @data

Sample of ARFF file

```
@relation weather
@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { TRUE, FALSE }
@attribute play { yes, no }
@data
sunny, 85, 85, FALSE, no
sunny, 80, 90, TRUE, no
overcast, 83, 86, FALSE, yes
rainy, 70, 96, FALSE, yes
rainy, 68, 80, FALSE, yes
rainy, 65, 70, TRUE, no
overcast, 64, 65, TRUE, yes
sunny, 72, 95, FALSE, no
sunny, 69, 70, FALSE, yes
rainy, 75, 80, FALSE, yes
sunny, 75, 70, TRUE, yes
overcast, 72, 90, TRUE, yes
overcast, 81, 75, FALSE, yes
rainy, 71, 91, TRUE, no
```

Sample of ARFF file



Weka: Explorer

- Preprocess: Choose and modify the data being acted on.
- Classify: Train and test learning schemes that classify or perform regression
- Cluster: Learn clusters for the data
- Associate: Learn association rules for the data
- Select attributes: Select the most relevant attributes in the data
- Visualize: View and interactive 2D plot of the data

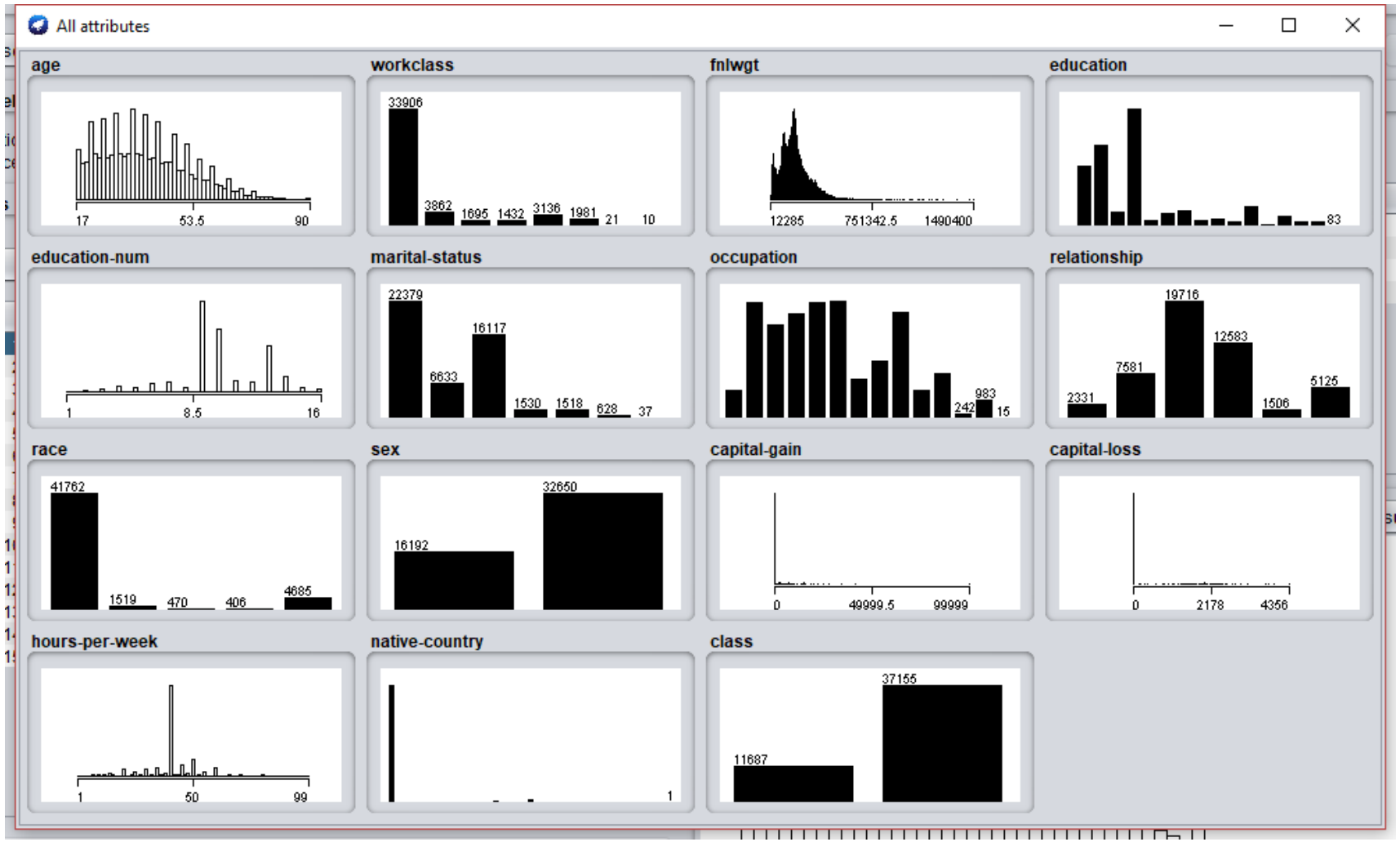
Data set – Census Income Dataset

- Download from UCI machine learning repository
- https://archive.ics.uci.edu/ml/datasets/census+income
- Prediction task is to determine whether a person makes over 50K a year.
- Convert using Python library, *csv2arff or you can google online conversion tools (there are fews)*
- Convert using *weka.core.converters.CSVLoader*

Data set – Census Income Dataset (cont.)

- **Data Understanding**
- 48,842 instances
- 15 attributes and their types
  - **Nominal**: workclass, education, marital-status, occupation, relationship, race, sex, native-country
  - **Continuous**: age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week:
- Salary – class target=2 (<= 50K and > 50K)

Data set – Census Income Dataset (cont.)



Workshop 4

- Workshop 4 will be post by tonight.
- The due date will be next Sunday at midnight.

References

- Anon. n.d. “Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems): Amazon.Co.Uk: Ian H. Witten, Eibe Frank, Mark A. Hall: 9780123748560: Books.” Retrieved August 20, 2018 (https://www.amazon.co.uk/Data-Mining-Practical-Techniques-Management/dp/0123748569/ref=sr_1_6?ie=UTF8&qid=1534755332&sr=8-6&keywords=data+mining).
- Jr, Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2013. *Multivariate Data Analysis*. 7 edition. Harlow: Pearson.