

# 데이터 마이닝 특강

## Practice session

[Introduction of datamining]

Junseok Park

2018-08-22



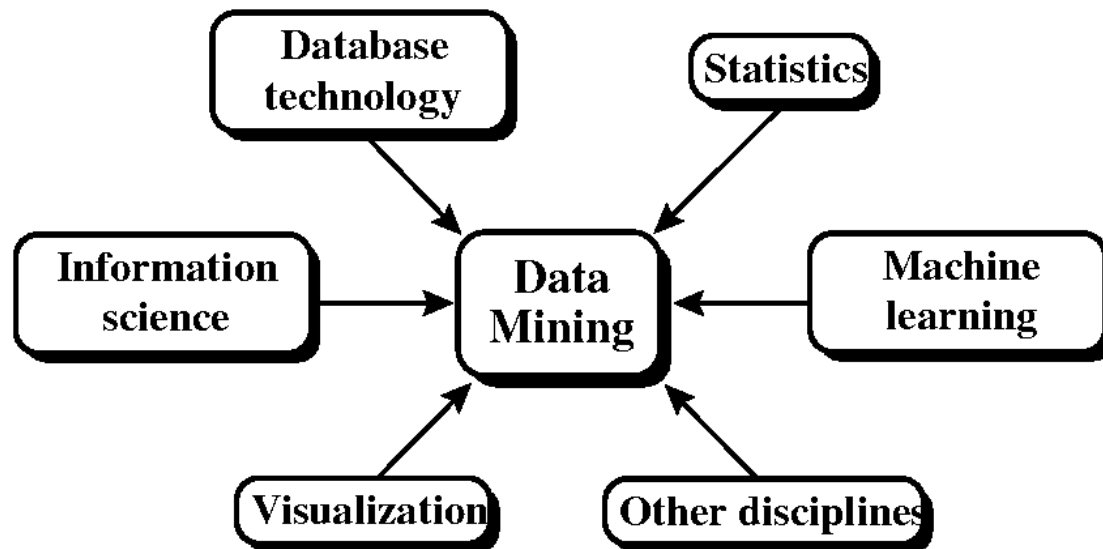
# **Introduction of datamining**

# Introduction

---

- **Data mining**

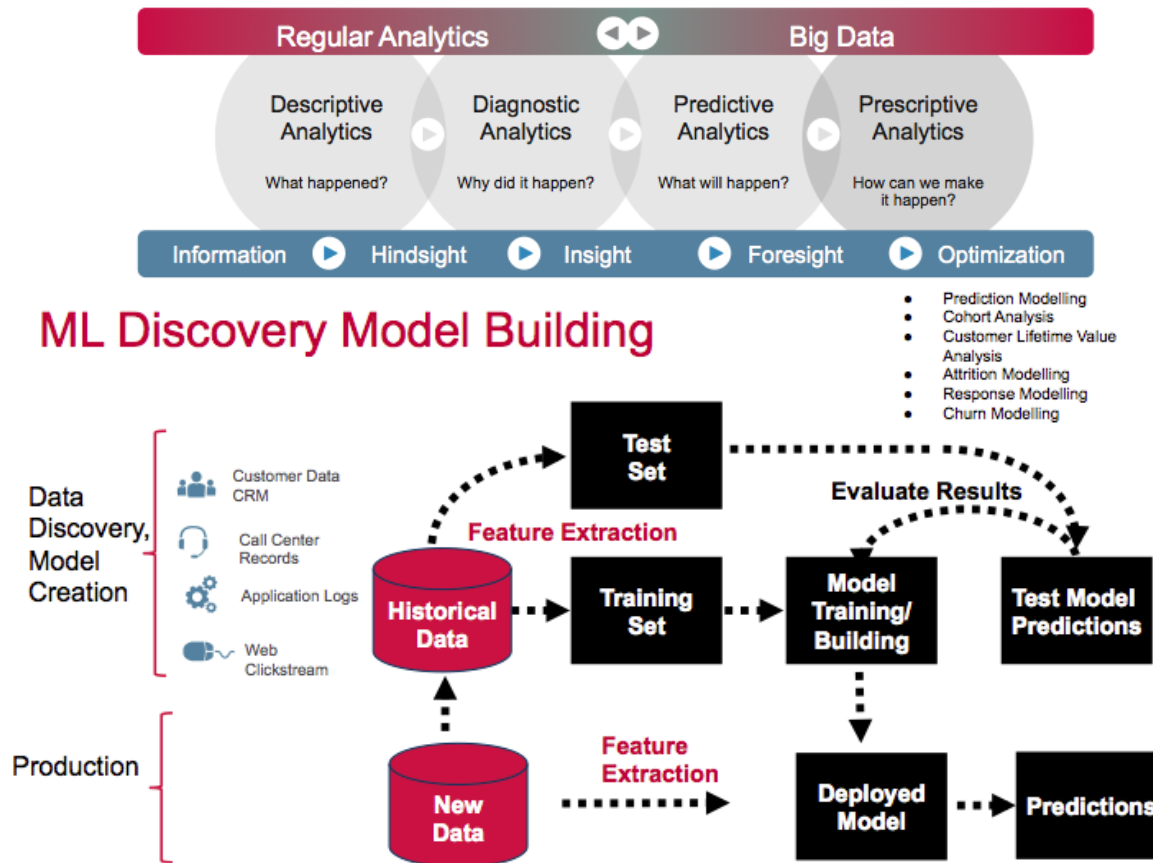
- knowledge discovery from data
- To use huge amount of data to discover interesting patterns or knowledge
- Alternative names
  - KDD : Knowledge discovery from data
  - Machine learning (Mainly in artificial intelligence)



Larose, Daniel T., and Chantal D. Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

# Application Example (Cont'd)

- Big data use cases in telecom



source from <http://datasciencegyan.com/big-data-use-cases-in-telecom/>

# Application Example (Cont'd)

---

- **Object detection API**

- models learned via Neural Architecture Search, instance segmentation support and models trained on new datasets such as Open Images

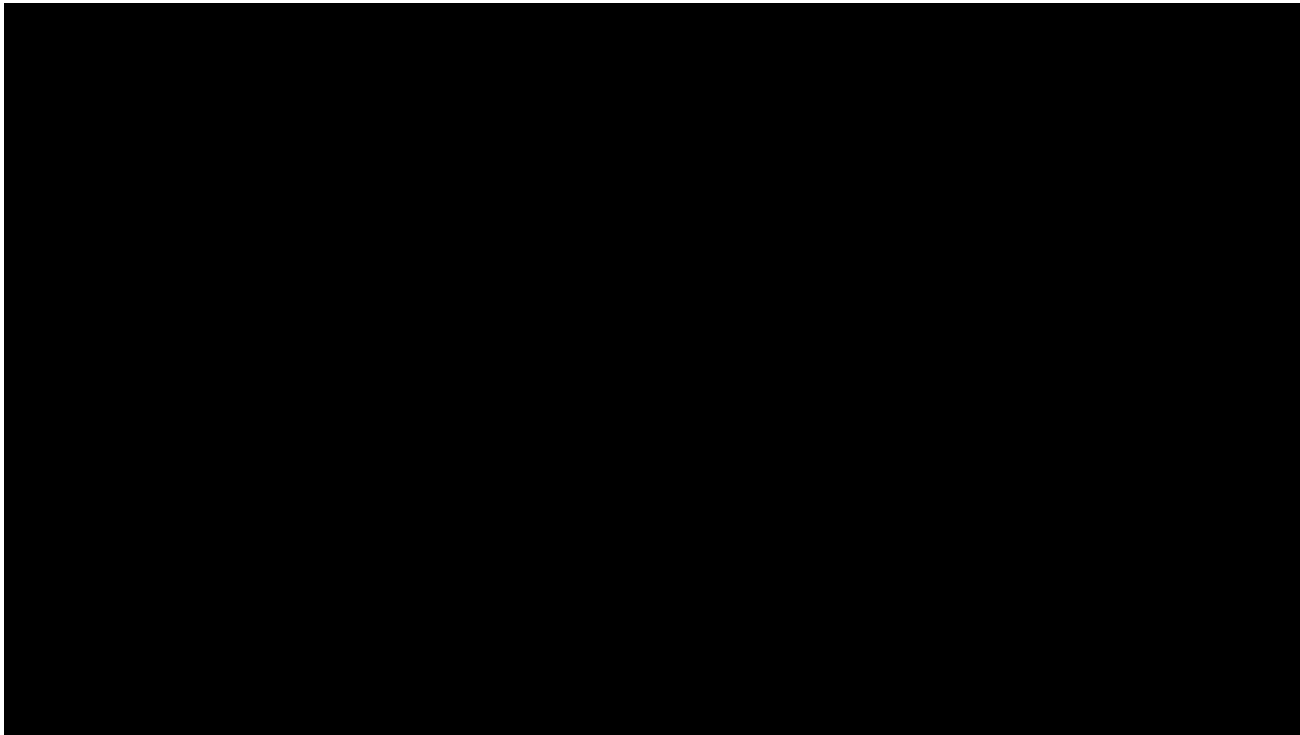


Source from Jonathan Huang ([ai.googleblog.com](https://ai.googleblog.com))

# Application Example

---

- **Virtual assistant**
  - mimic a human voice to book an appointment by phone  
(<https://youtu.be/wqhxVwXI6q8>)

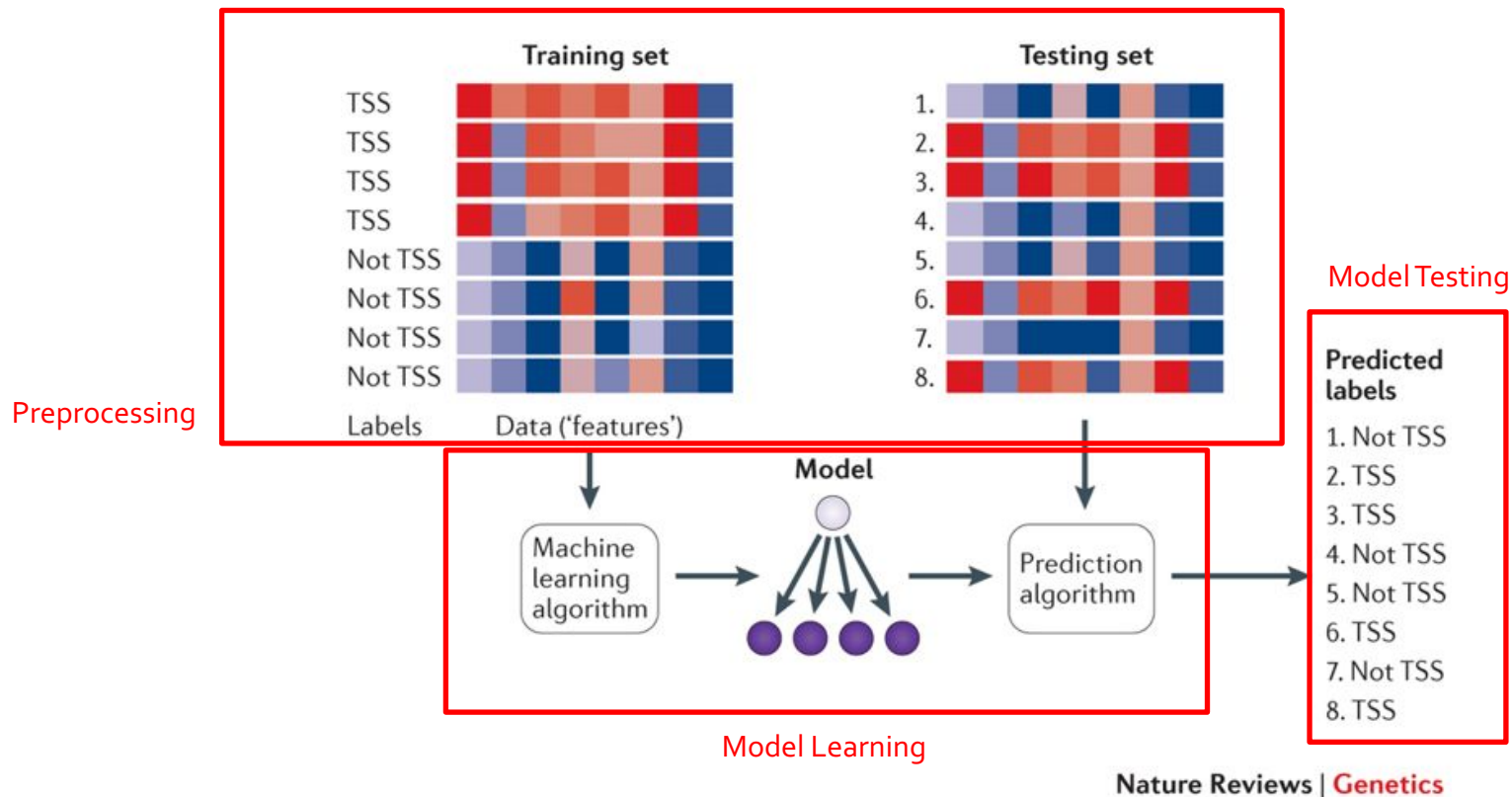


Source from Google's robot assistant now makes eerily lifelike phone calls for you, Olivia Solon, theguardian.com, 2018

# Objective

- **Objective of the practice session**

- Learn about how to build model and understand hyperparameters

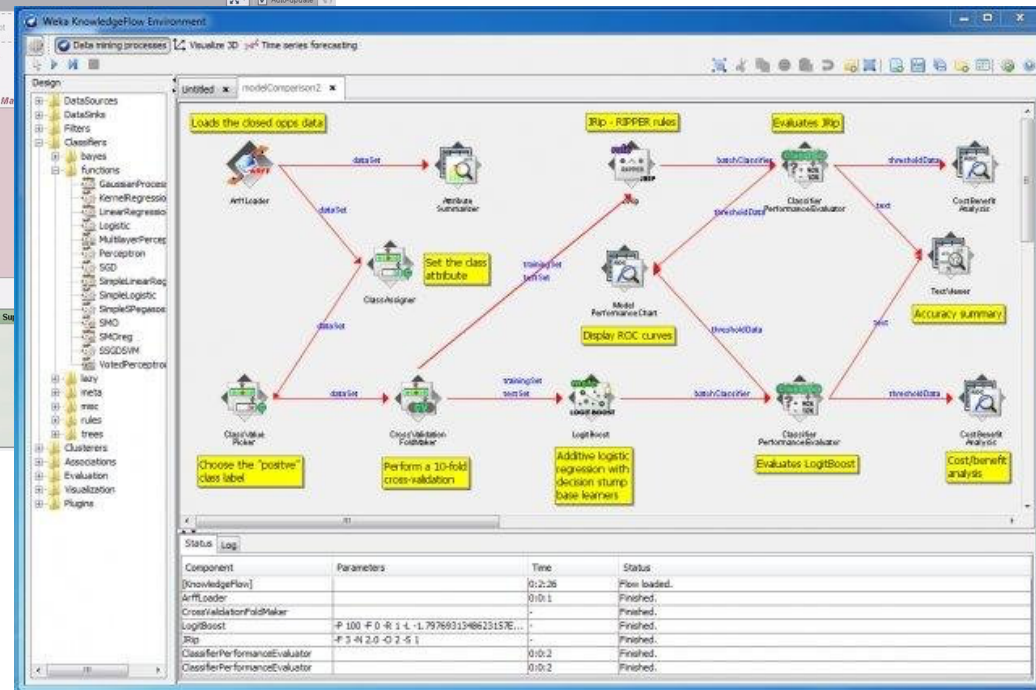
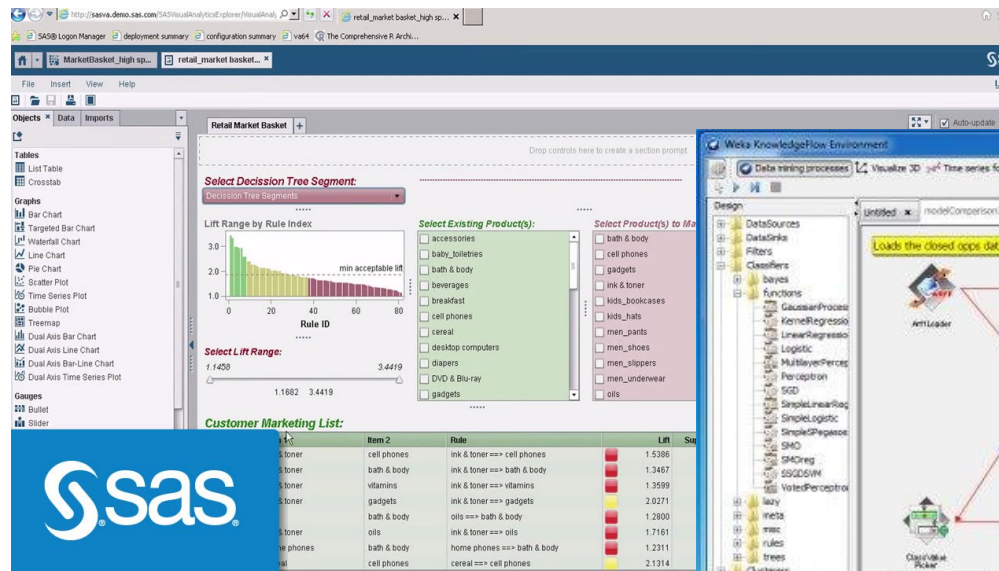


A canonical example of a machine learning application<sup>[1]</sup>

<sup>[1]</sup> Libbrecht MW, Noble WS, 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16(05/07/online), 321. DOI= <http://dx.doi.org/10.1038/nrg3920>.

# Data mining tools

- Various data mining tools in convenient approaches





# Practice Preparation (Cont'd)

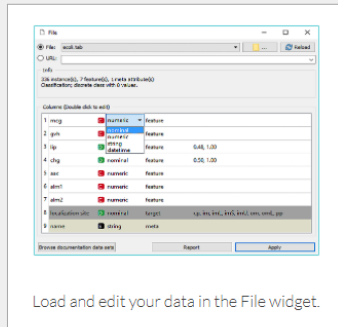
- **Orange**

- Open source machine learning and data visualization
- <https://orange.biolab.si>

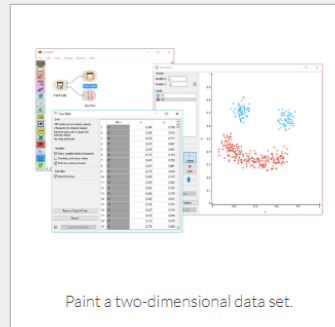


Home Screenshots Download Docs Blog Training [Donate](#)

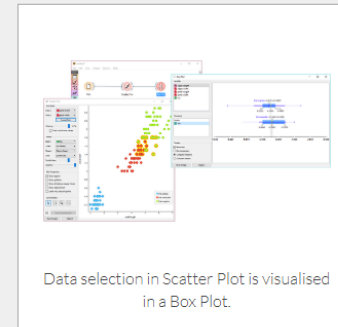
## Screenshots



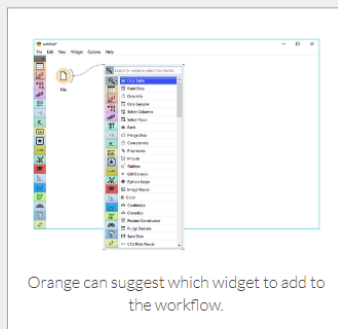
Load and edit your data in the File widget.



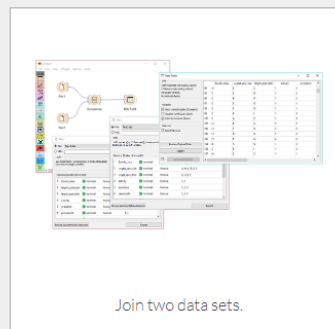
Paint a two-dimensional data set.



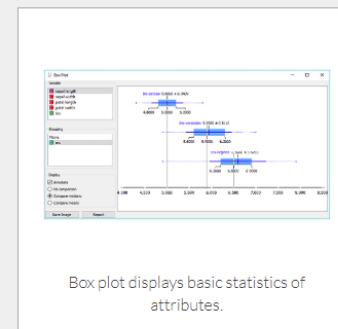
Data selection in Scatter Plot is visualised in a Box Plot.



Orange can suggest which widget to add to the workflow.



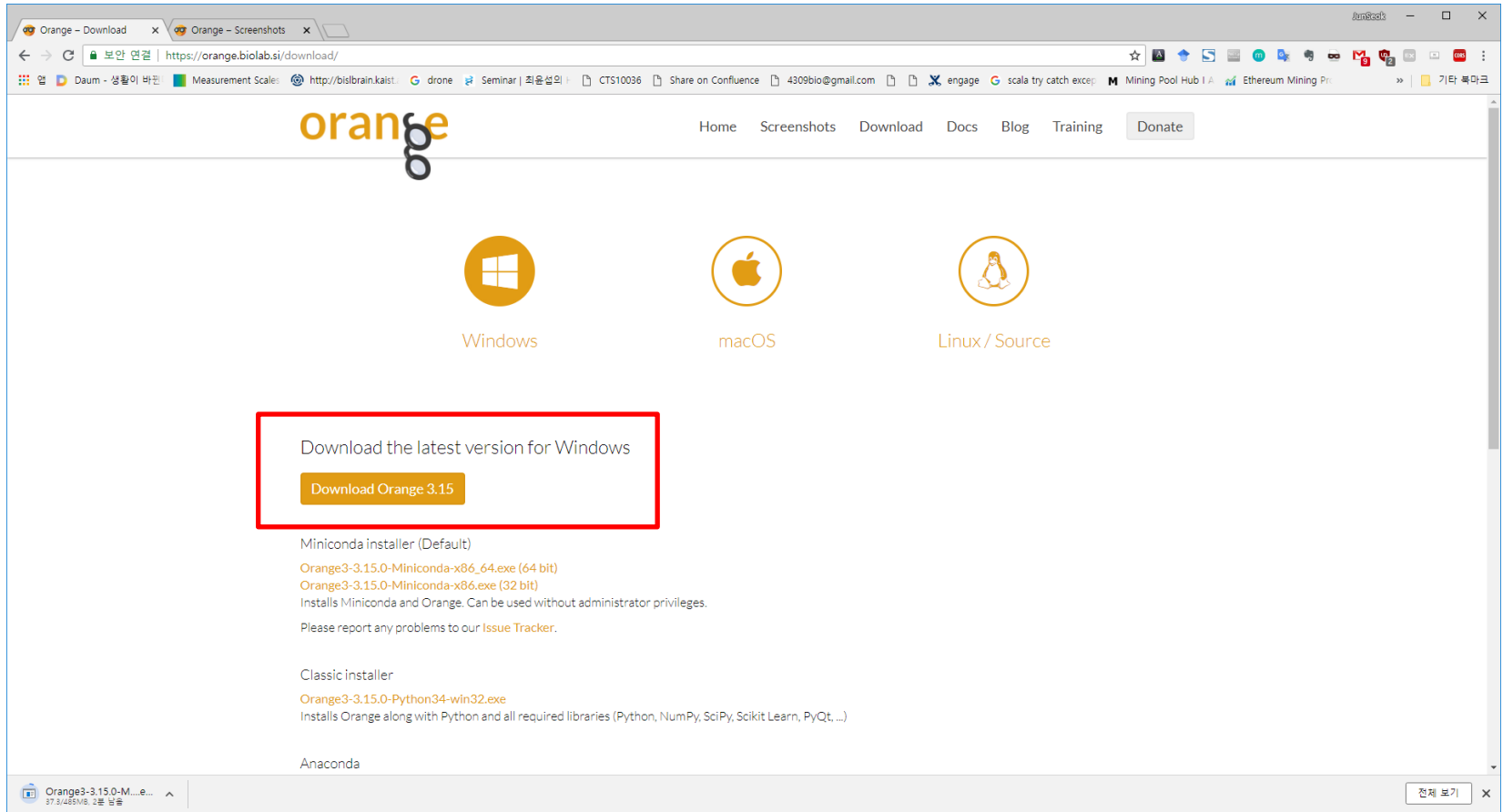
Join two data sets.



Box plot displays basic statistics of attributes.

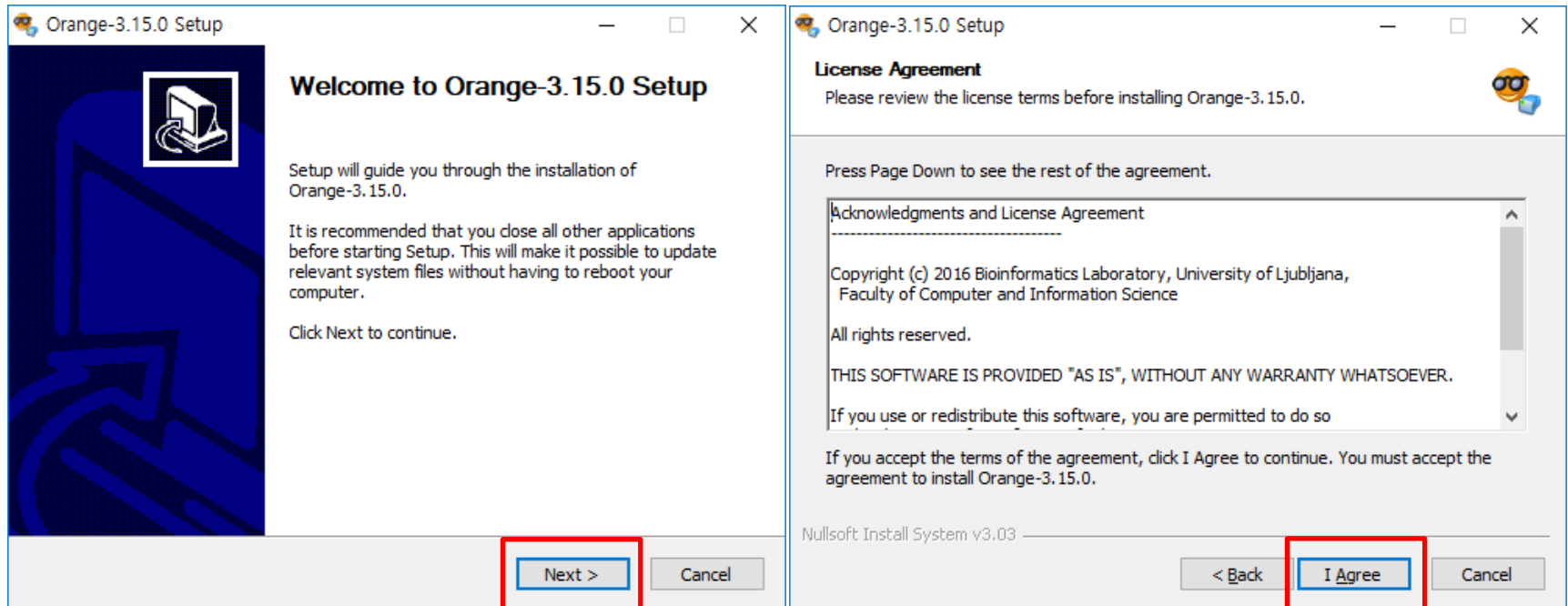
# Practice Preparation (Cont'd)

- Download orange



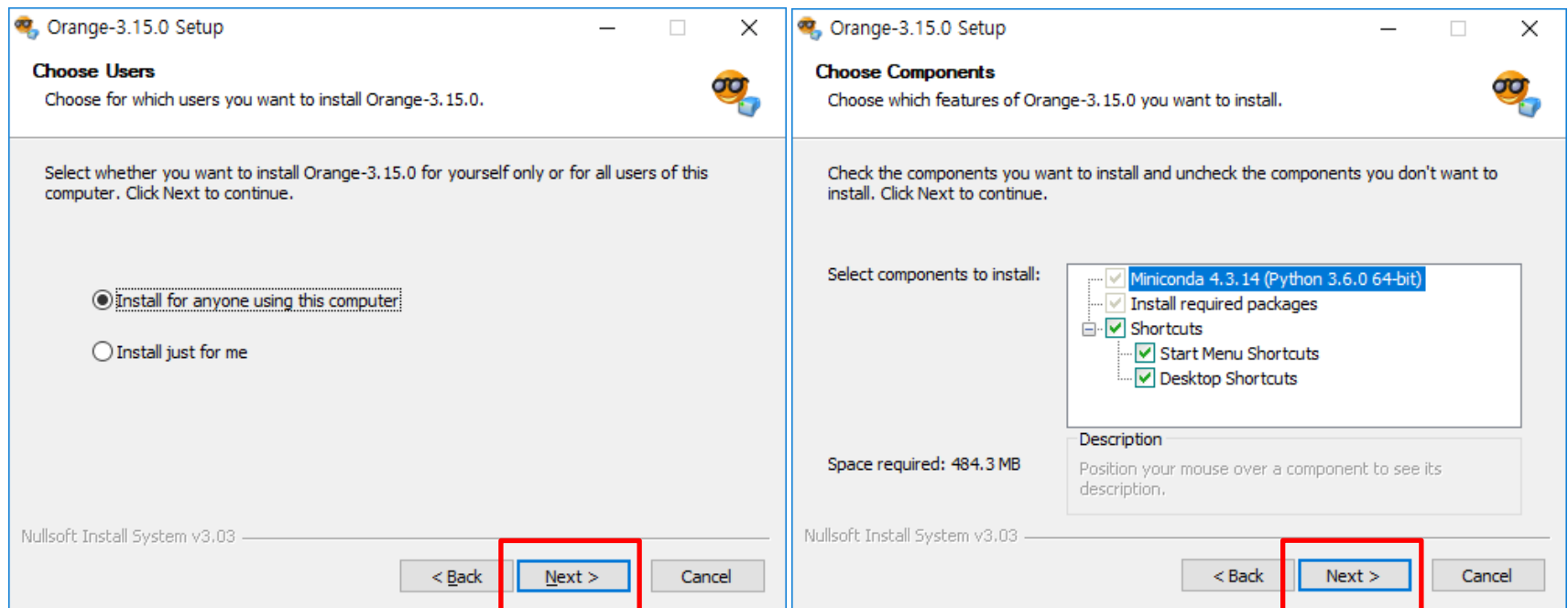
# Practice Preparation (Cont'd)

- Install orange (1/8)



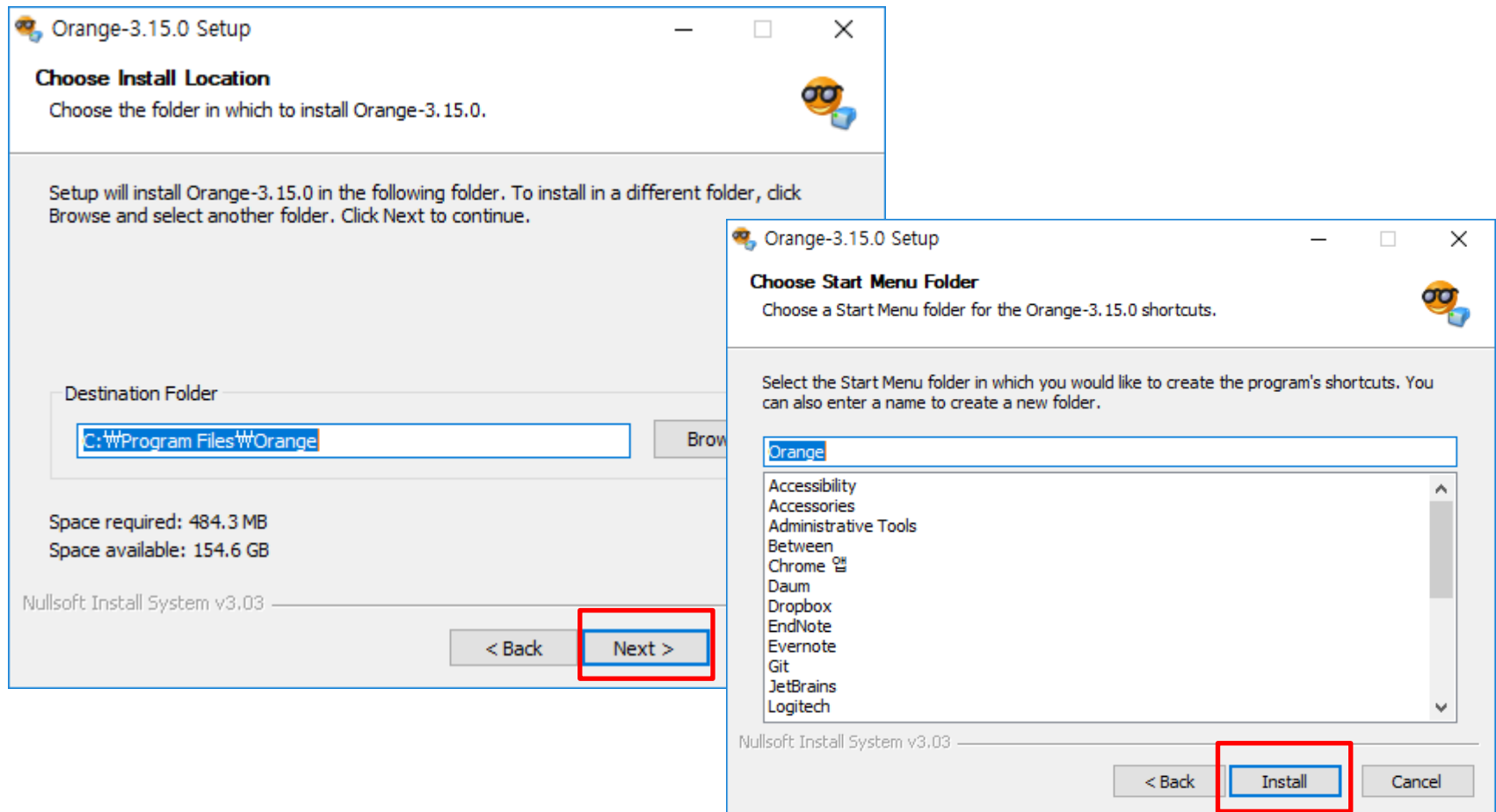
# Practice Preparation (Cont'd)

- Install orange (2/8)



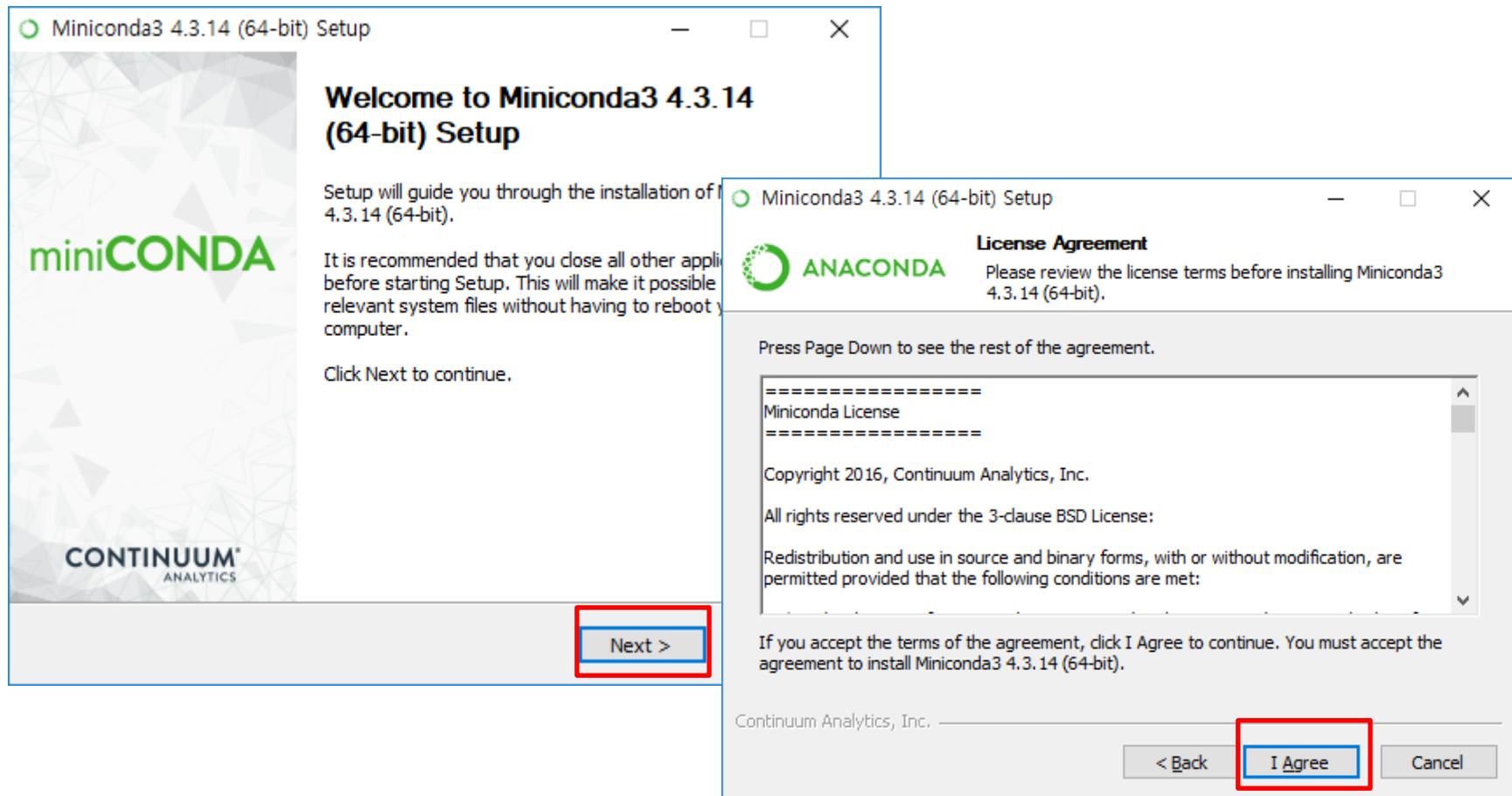
# Practice Preparation (Cont'd)

- Install orange (3/8)



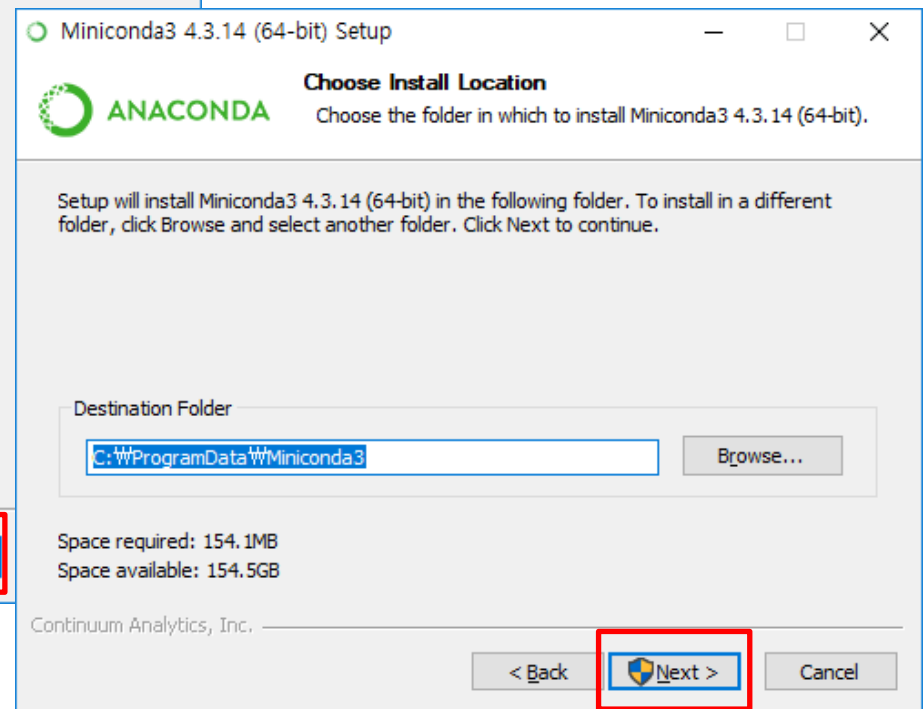
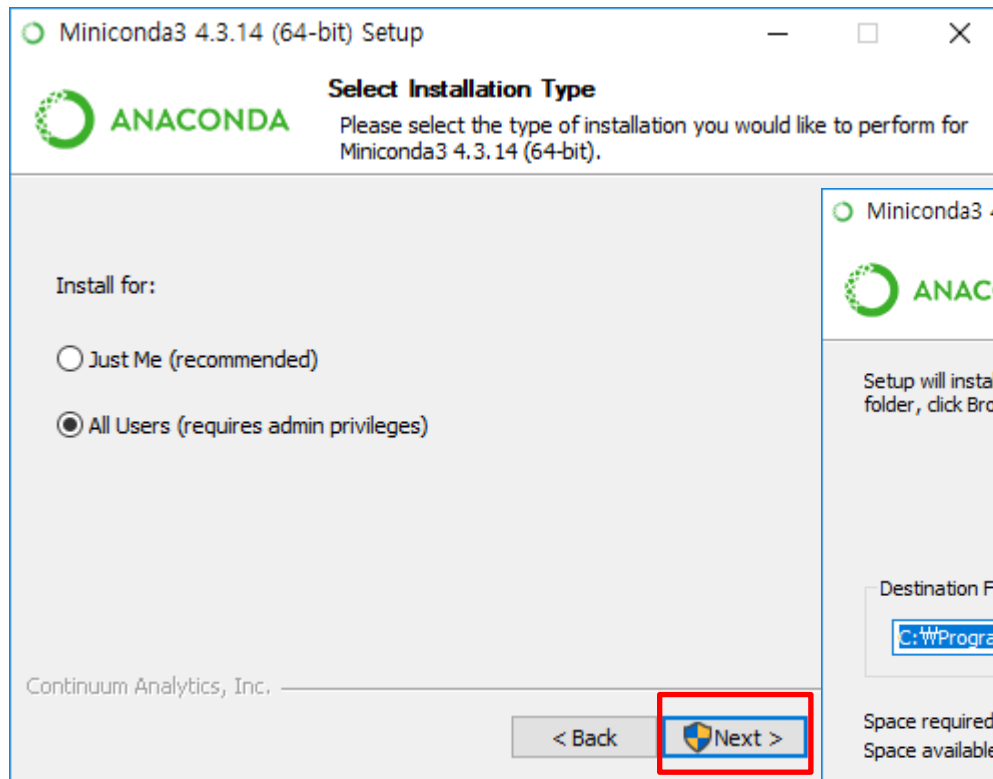
# Practice Preparation (Cont'd)

- Install orange (4/8)



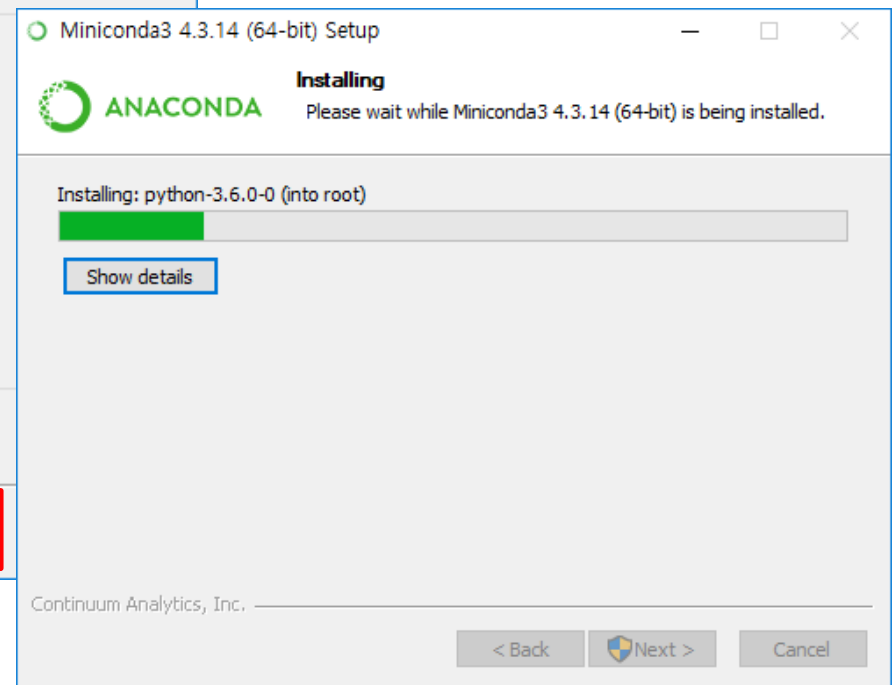
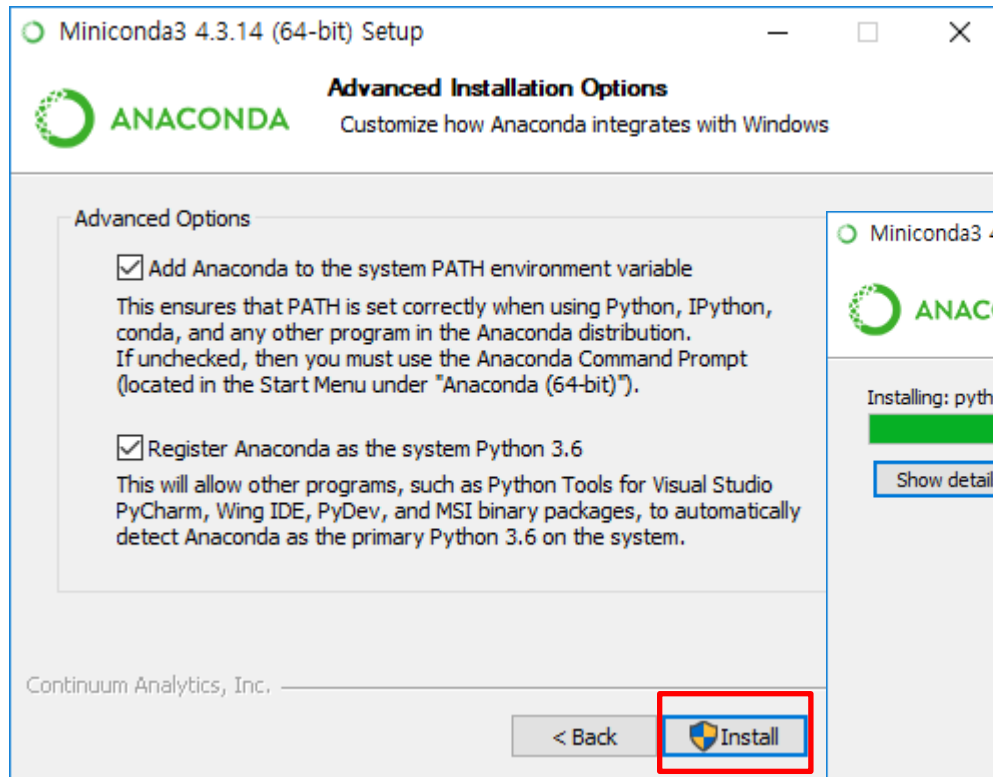
# Practice Preparation (Cont'd)

- Install orange (5/8)



# Practice Preparation (Cont'd)

- Install orange (6/8)



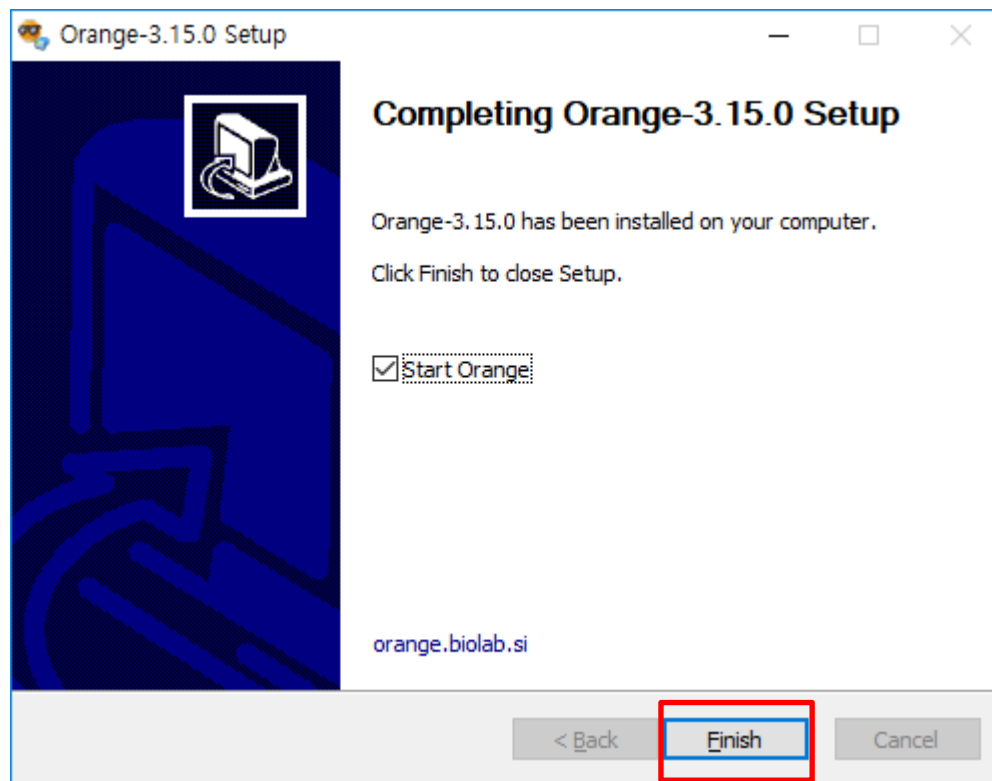




# Practice Preparation

---

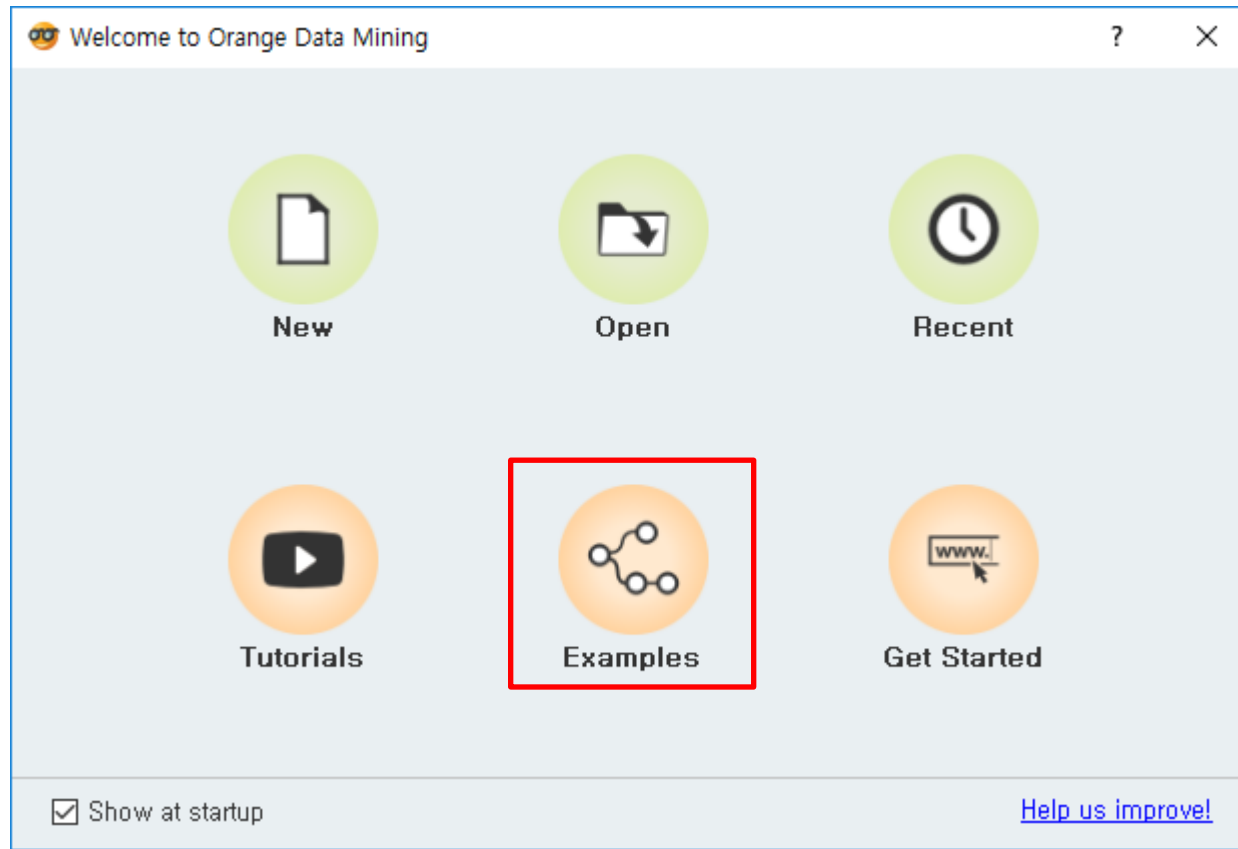
- Install orange (8/8)



# Demo practice (Cont'd)

---

- Tutorials for the first time
  - <https://orange.biolab.si/getting-started>



# Demo practice (Cont'd)

- The preloaded data mining workflows

- Hierarchical clustering : a method of clustering analysis which seeks to build a hierarchy of clusters

**Example Workflows**

### Hierarchical Clustering

The workflow clusters the data items in Iris dataset by first estimating the distances between data instances. Distance matrix is passed to hierarchical clustering, which renders the dendrogram. Select different parts of the dendrogram to further analyze the corresponding data.

Notice that hierarchical clustering can only handle small datasets, that is, those that contain only a small number of data instances. For larger datasets the distance matrix may get too big and may not fit in the memory. An alternative method that can consider such datasets is k-means clustering.

**Path:** C:\Program Files\Orange3\lib\site-packages\Orange3\addons\application\workflows\310-clustering.ows

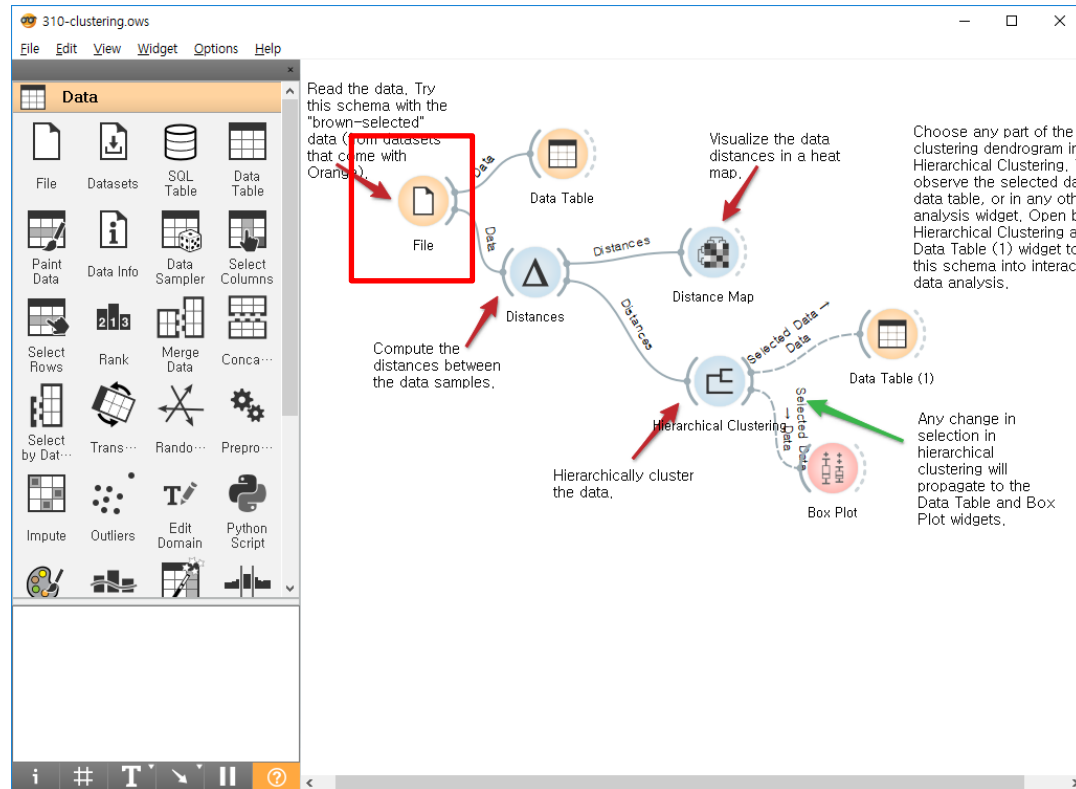
Interactive Visualizations   Visualization of Data Subsets   Classification Tree   Principal Component Analysis   **Hierarchical Clustering**   Feature Ranking

Open Cancel

# Demo practice (Cont'd)

- **Orange canvas**

- In orange, data mining workflows consist of computational components called widget
- Widgets do all the work and exchange information
  - File widget : send data to the data table widget and distance widget
  - Distance widget : compute the distances between the data samples



# Demo practice (Cont'd)

- Input data on the file widget

- chose *iris.tab* from the list of pre-installed data

The screenshot shows the Orange3 File widget interface on the left and a Windows file explorer window on the right. The File widget is set to 'File' mode with 'brown-selected.tab' selected. A red box highlights the folder icon button next to the filename. The 'Info' section shows 186 instances, 79 features, and 1 meta attribute, with a classification of categorical class with 3 values. The 'Columns' table lists 5 features: alpha 0, alpha 7, alpha 14, alpha 21, and alpha 28, all numeric. The file explorer shows the 'Orange > datasets' directory with 'iris.tab' selected. A red box highlights the 'iris.tab' file in the list and the '열기(O)' (Open) button at the bottom.

**File Widget Information:**

- File: brown-selected.tab
- URL:
- Info: 186 instance(s), 79 feature(s), 1 meta attribute(s)  
Classification: categorical class with 3 values,
- Columns (Double click to edit):

	Name	Type	Role	Values
1	alpha 0	numeric	feature	
2	alpha 7	numeric	feature	
3	alpha 14	numeric	feature	
4	alpha 21	numeric	feature	
5	alpha 28	numeric	feature	

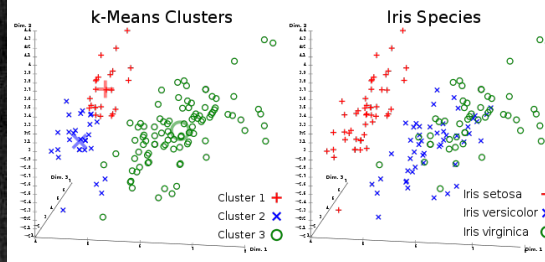
**File Explorer Details:**

이름	수정된 날짜	유형	크기
flare2.tab	2018-08-06 오후...	TAB 파일	28KB
geo-gds360.tab	2018-08-06 오후...	TAB 파일	1,828KB
glass.tab	2018-08-06 오후...	TAB 파일	11KB
hair-eye-sex.tab	2018-08-06 오후...	TAB 파일	8KB
hayes-roth_learn.tab	2018-08-06 오후...	TAB 파일	2KB
hayes-roth_test.tab	2018-08-06 오후...	TAB 파일	1KB
heart_disease.tab	2018-08-06 오후...	TAB 파일	24KB
horse-colic.tab	2018-08-06 오후...	TAB 파일	46KB
horse-colic_learn.tab	2018-08-06 오후...	TAB 파일	25KB
horse-colic_test.tab	2018-08-06 오후...	TAB 파일	6KB
housing.tab	2018-08-06 오후...	TAB 파일	34KB
imports-85.tab	2018-08-06 오후...	TAB 파일	26KB
ionosphere.tab	2018-08-06 오후...	TAB 파일	75KB
iris.tab	2018-08-06 오후...	TAB 파일	5KB
lenses.tab	2018-08-06 오후...	TAB 파일	1KB
lung-cancer.tab	2018-08-06 오후...	TAB 파일	4KB

# Demo practice (Cont'd)

- iris flower data set

- multivariate data set
- introduced by Ronald Fisher in 1936
  - Statistician and Biologist



- the data to quantify the morphologic variation of Iris flowers of three related species

File

File: iris.tab [Reload]

URL: [ ]

Info

**Iris flower dataset**  
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.  
150 instance(s), 4 feature(s), 0 meta attribute(s)  
Classification; categorical class with 3 values.

Columns (Double click to edit)

	Name	Type	Role	Values
1	sepal length	numeric	feature	
2	sepal width	numeric	feature	
3	petal length	numeric	feature	
4	petal width	numeric	feature	
5	iris	categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

Browse documentation datasets

Apply



Iris setosa



Iris versicolor

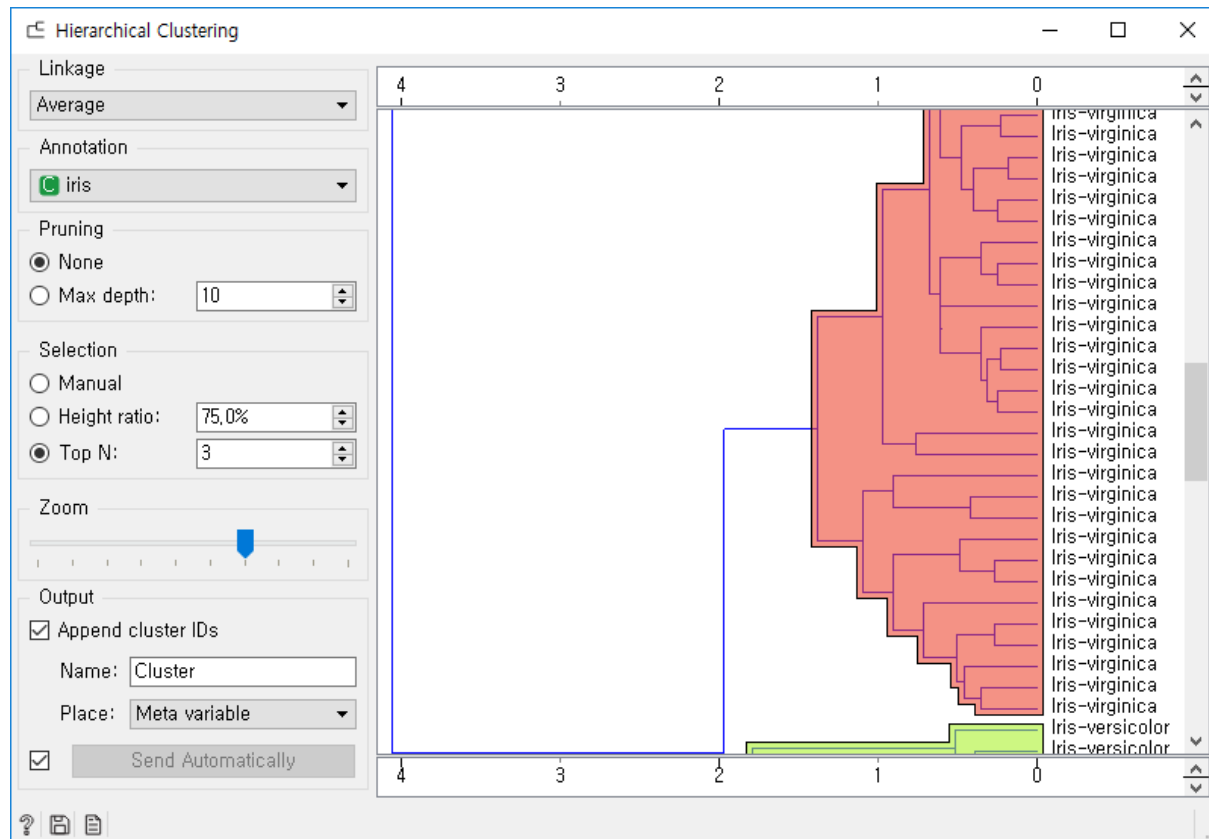


Iris virginica

# Demo practice

- **Clustering results**

- the dendrogram : the tree-based rendering of the clustering
- Check if the algorithm correctly identified the three species of Iris

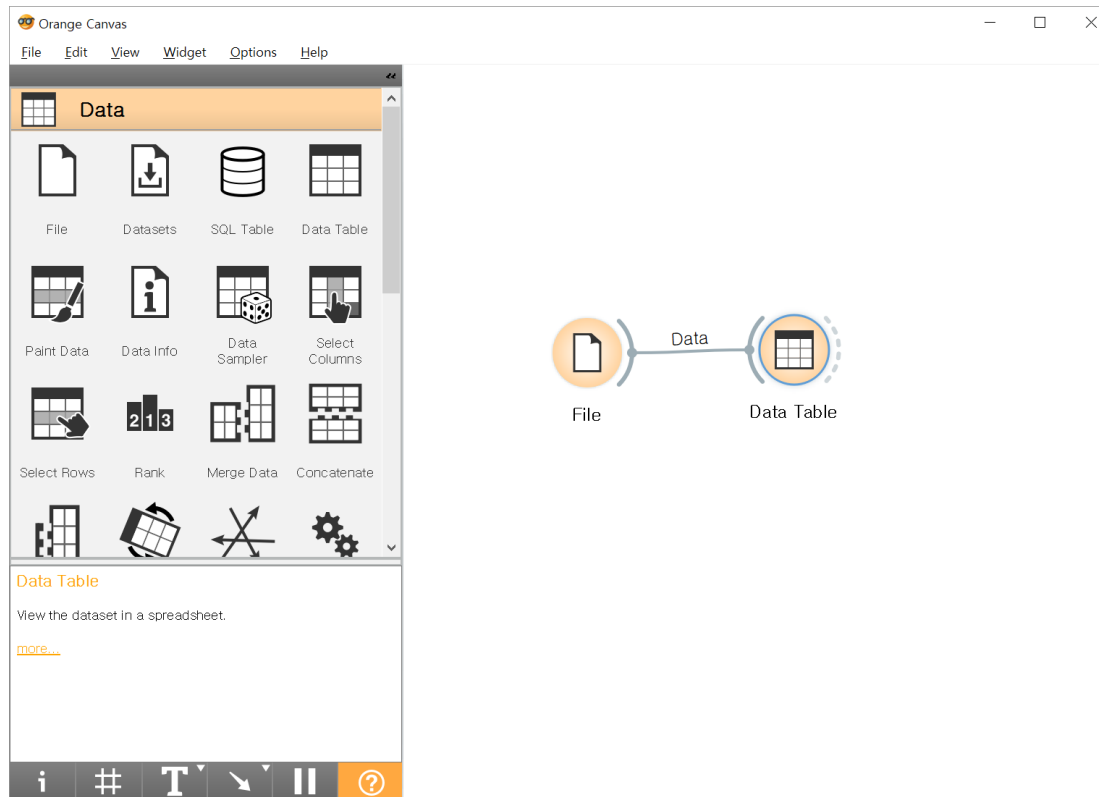




# Basic practice

- **Your own workflow**

- start with empty canvas
- Develop a model to predict the probability of survival based on the passenger's traveling class, gender and age from the data on passengers of the RMS\* Titanic



RMS *Titanic* departing [Southampton](https://en.wikipedia.org/wiki/Southampton) on 10 April 1912<sup>[1]</sup>

\*RMS : Royal mail ship

<sup>[1]</sup> [https://en.wikipedia.org/wiki/RMS\\_Titanic](https://en.wikipedia.org/wiki/RMS_Titanic)

# Basic practice (Cont'd)

- **Load and parsing data**

- The widgets automatically transferred the loaded data to all the connected widgets

The File widget interface shows the file 'titanic.tab' loaded. The Info section indicates 2201 instances, 3 features, and 0 meta attributes. The Classification is categorical with 2 values. The Columns section lists the features: status, age, sex, and survived.

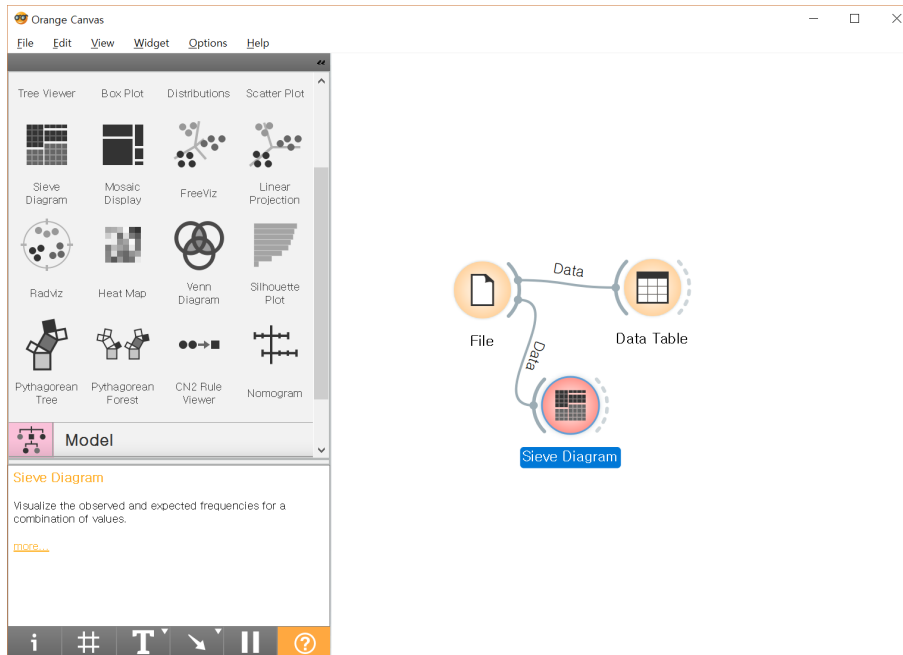
	Name	Type	Role	Values
1	status	categorical	feature	crew, first, second, third
2	age	categorical	feature	adult, child
3	sex	categorical	feature	female, male
4	survived	categorical	target	no, yes

The Data Table widget displays the loaded data. The Info section shows 2201 instances, 3 features, and 0 meta attributes. The table shows the first 27 rows of data.

	status	age	sex	survived
1	first	adult	male	yes
2	first	adult	male	yes
3	first	adult	male	yes
4	first	adult	male	yes
5	first	adult	male	yes
6	first	adult	male	yes
7	first	adult	male	yes
8	first	adult	male	yes
9	first	adult	male	yes
10	first	adult	male	yes
11	first	adult	male	yes
12	first	adult	male	yes
13	first	adult	male	yes
14	first	adult	male	yes
15	first	adult	male	yes
16	first	adult	male	yes
17	first	adult	male	yes
18	first	adult	male	yes
19	first	adult	male	yes
20	first	adult	male	yes
21	first	adult	male	yes
22	first	adult	male	yes
23	first	adult	male	yes
24	first	adult	male	yes
25	first	adult	male	yes
26	first	adult	male	yes
27	first	adult	male	yes

# Basic practice

- **Inspect survival probabilities for the passengers of Titanic**
  - Sieve diagram for data mining
  - Sieve diagram shows the frequencies in a two-way contingency table in relation to expected frequencies under independence, and highlights the patterns of association between the row and column variables<sup>[1]</sup>



<sup>[1]</sup> Riedwyl, Hans, and M. Schüpbach. "Parquet diagram to plot contingency tables." *Softstat* 93 (1994): 293-299.

<sup>[2]</sup> <http://www.datavis.ca/online/sieve/>

# Project assignment (Cont'd)

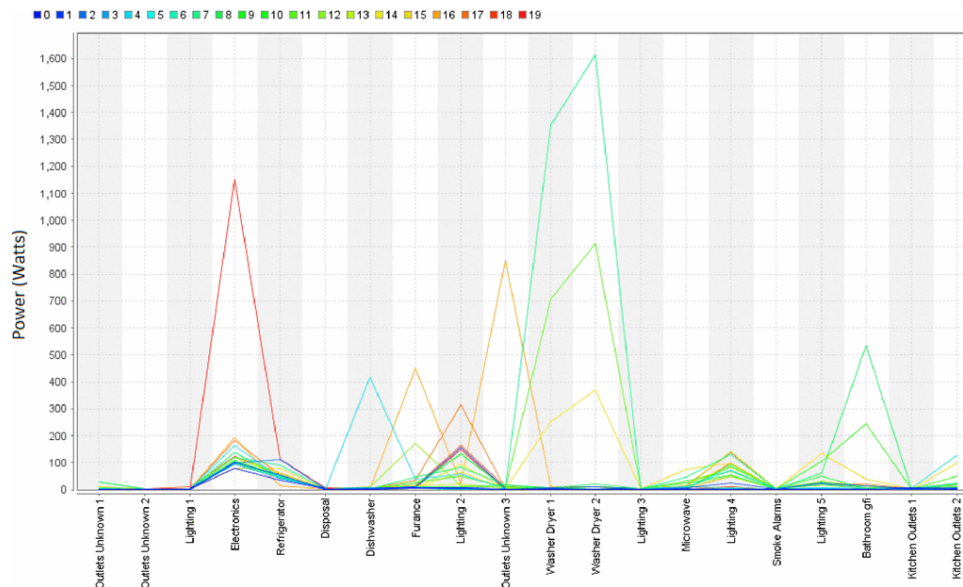
- **Analysis households electricity consumption**

- Dataset

- The reference energy disaggregation dataset (REDD) Version 1.0<sup>[1]</sup>
    - Web resources: <http://redd.csail.mit.edu/>

- Reference and examples

- Aravindh Aki, D Krishna Mohan Reddy, Y Koushik Reddy, C. R. Kavitha, T. Sasikala, "[Analyzing the real time electricity data using data mining techniques](#)", Smart Technologies For Smart Nation (SmartTechCon) 2017 International Conference On, pp. 545-549, 2017.



Clusters of house 3 REDD datasets

<sup>[1]</sup> J. Zico Kolter and Matthew J. Johnson. REDD: A public data set for energy disaggregation research. In proceedings of the SustKDD workshop on Data Mining Applications in Sustainability, 2011

# Project assignment

---

- **Due**

- Until end of the class via e-mail

- **Format**

- Free of your wish

- **Questions and help**

- [junseokpark@kaist.ac.kr](mailto:junseokpark@kaist.ac.kr)

- **Resources**

- <https://github.com/junseokpark/resources/datamining>

- **web-based competition site for datamining**

- <https://www.kaggle.com/competitions>

# References

---

- <https://orange.biolab.si/getting-started/>

**Thank you**

