데이터 마이닝 특강
Practice session

[Introduction of datamining]

Junseok Park

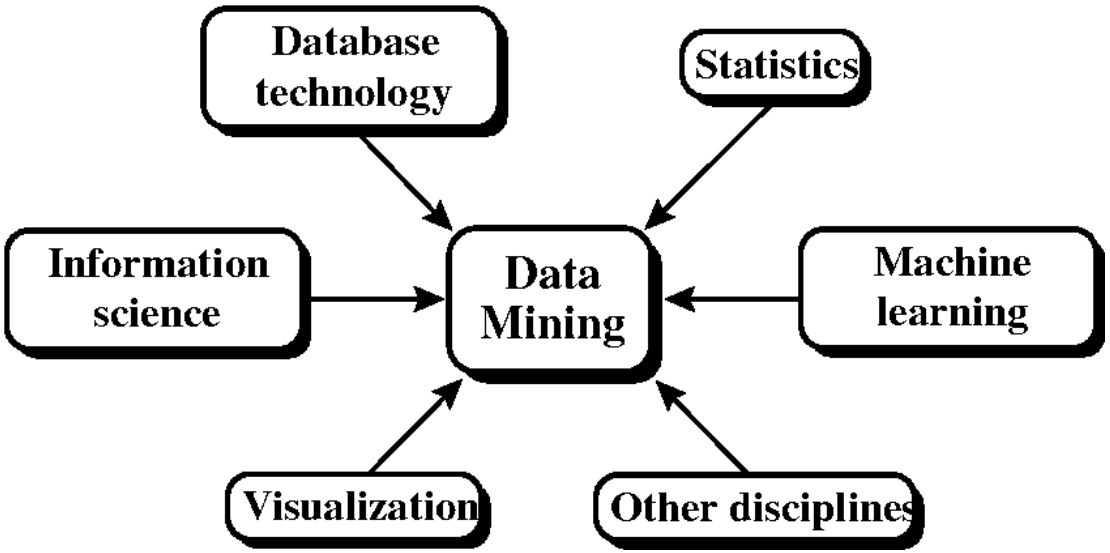
2018-08-22

KAIST

Introduction of datamining

Introduction

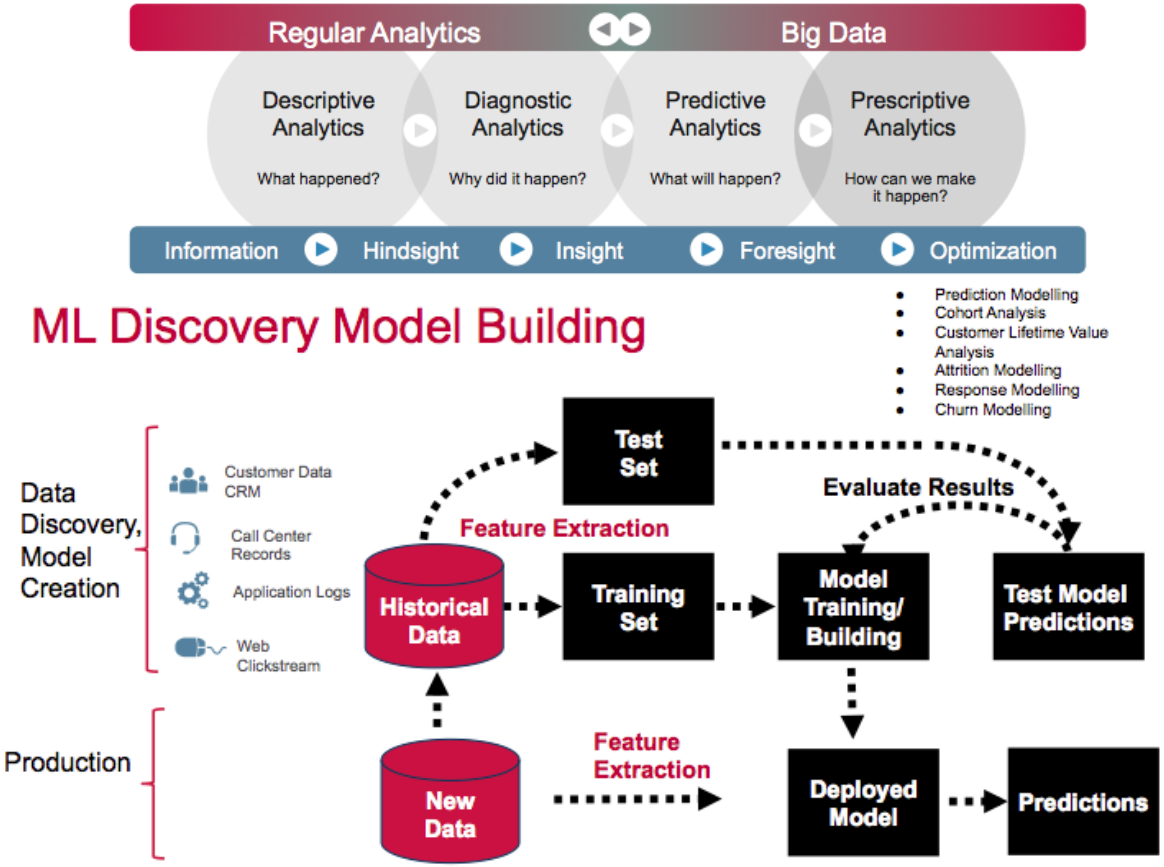
- **Data mining**
  - knowledge discovery from data
  - To use huge amount of data to discover interesting patterns or knowledge
  - Alternative names
    - KDD : Knowledge discovery from data
    - Machine learning (Mainly in artificial intelligence)



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)

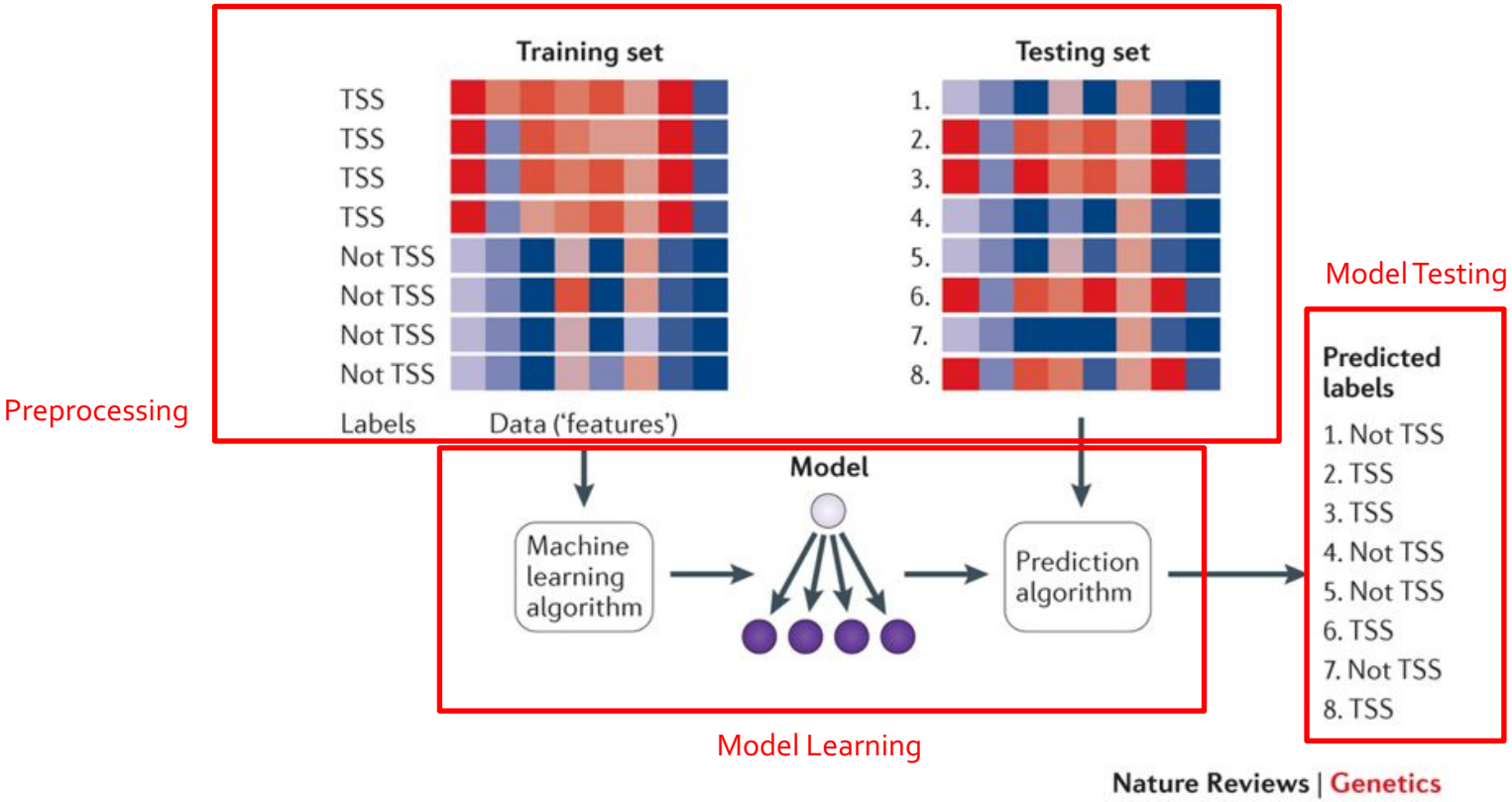
Application Example

- **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, thegardian.com, 2018

Objective

- **Objective of the practice session**
  - Learn about how to build model and understand hyperparameters

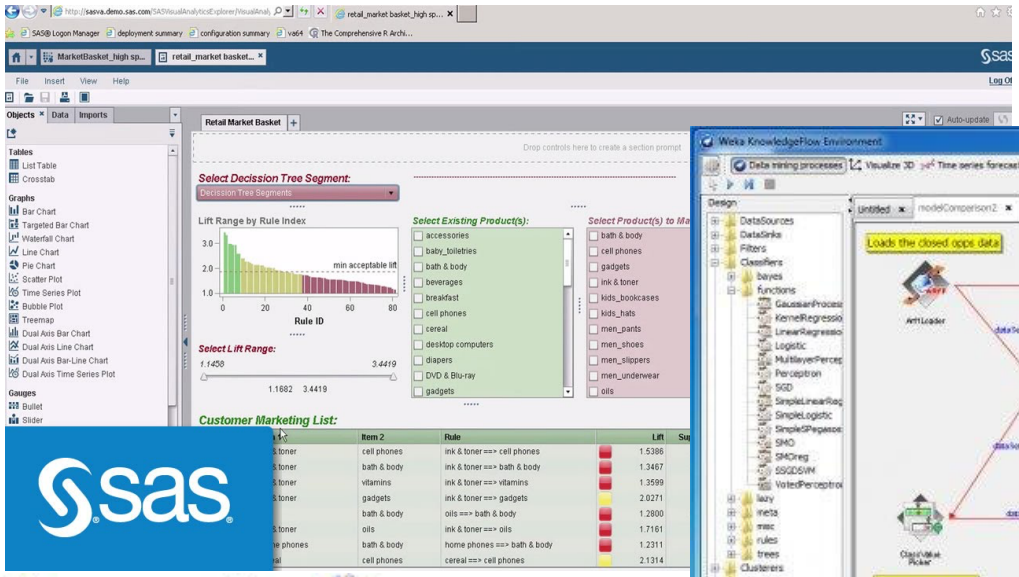


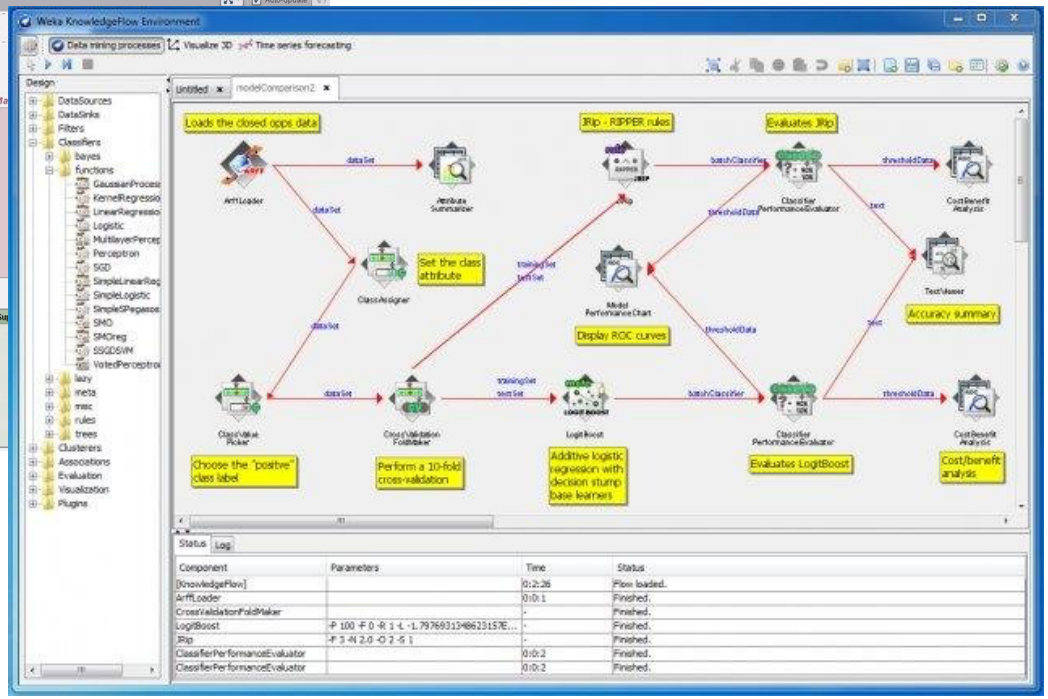
A canonical example of a machine learning application[1]

[1] 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

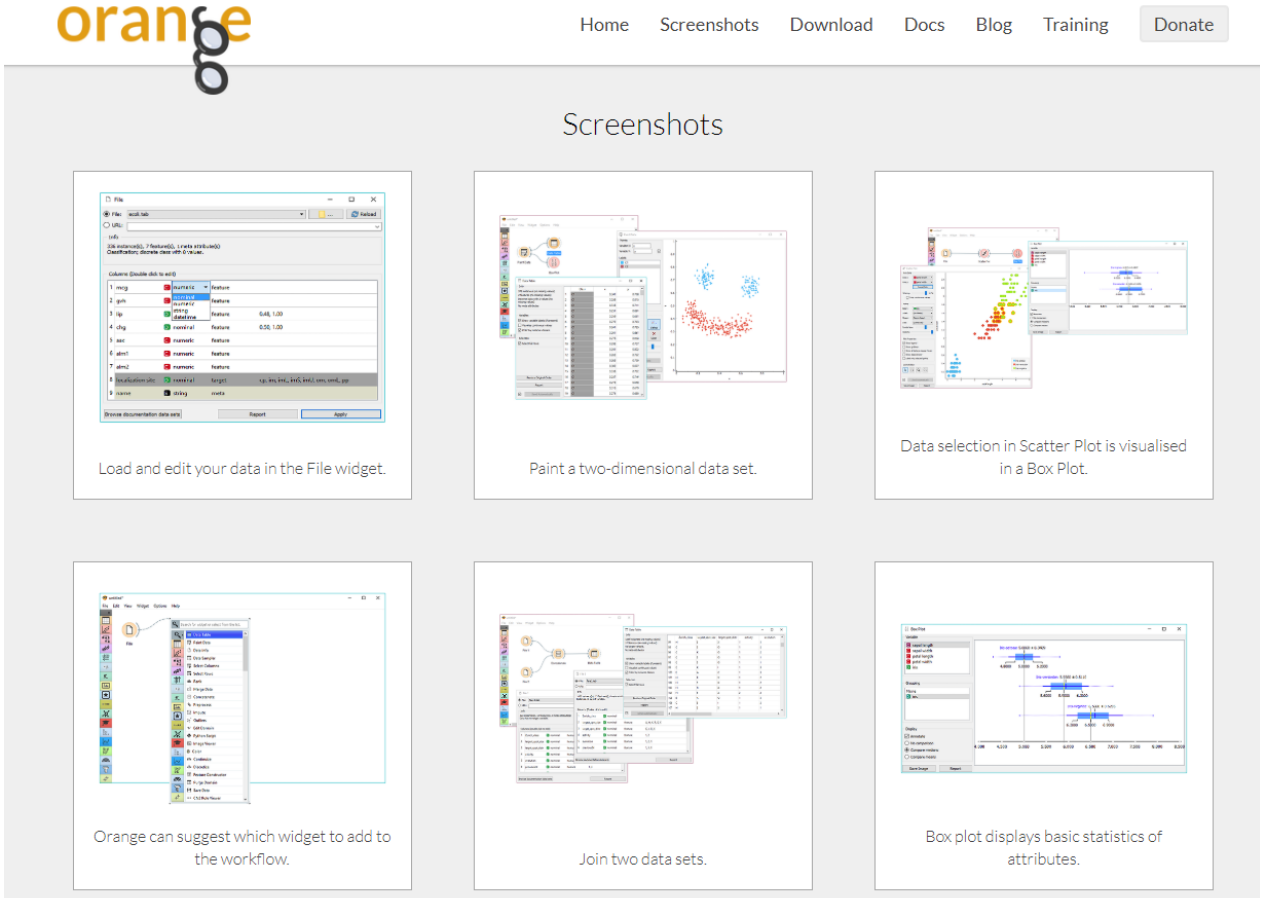
- **Various data mining tools in convenient approaches**





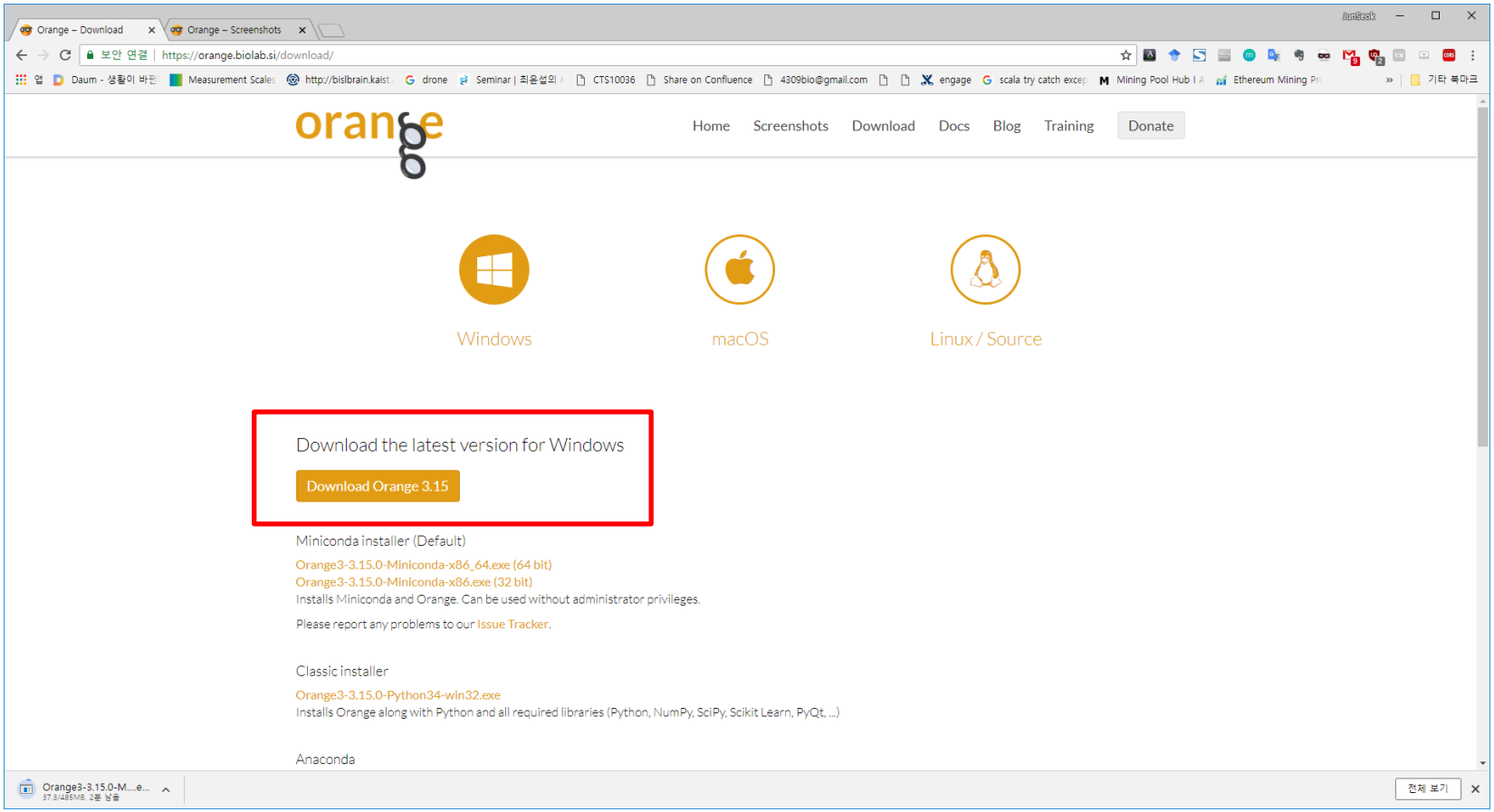
Practice Preparation (Cont'd)

- **Orange**
  - Open source machine learning and data visualization
  - https://orange.biolab.si



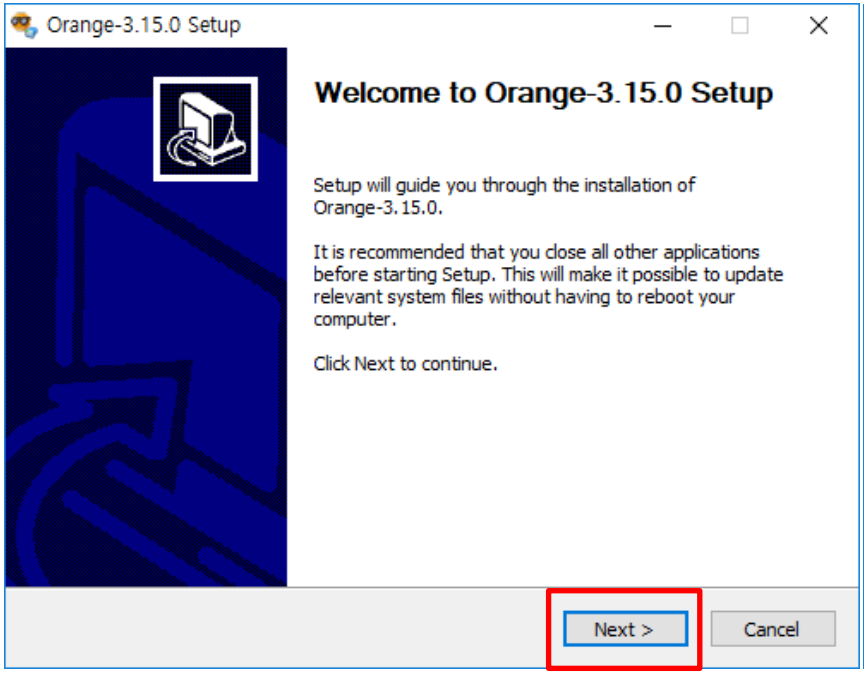
Practice Preparation (Cont'd)

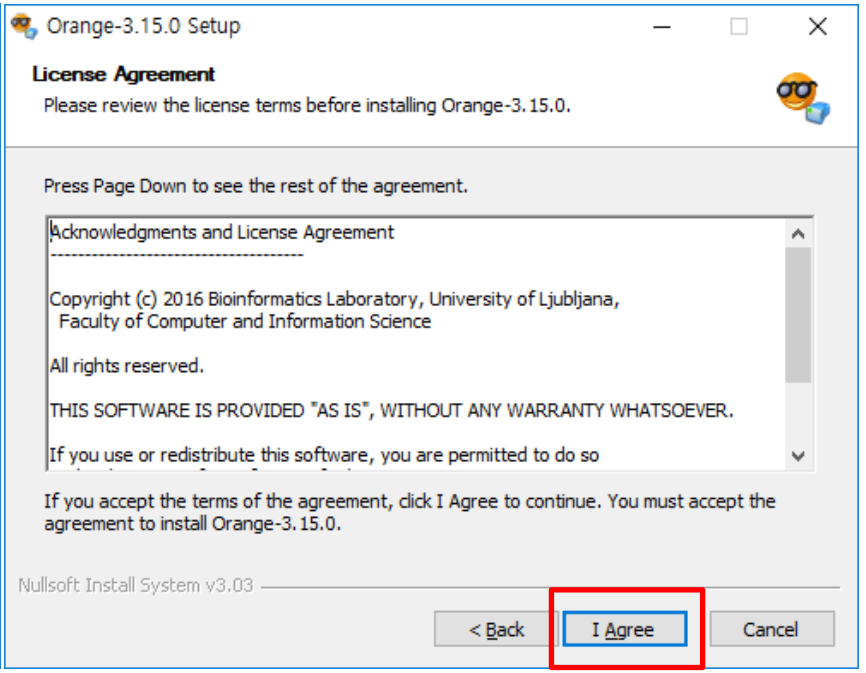
- **Download orange**



Practice Preparation (Cont'd)

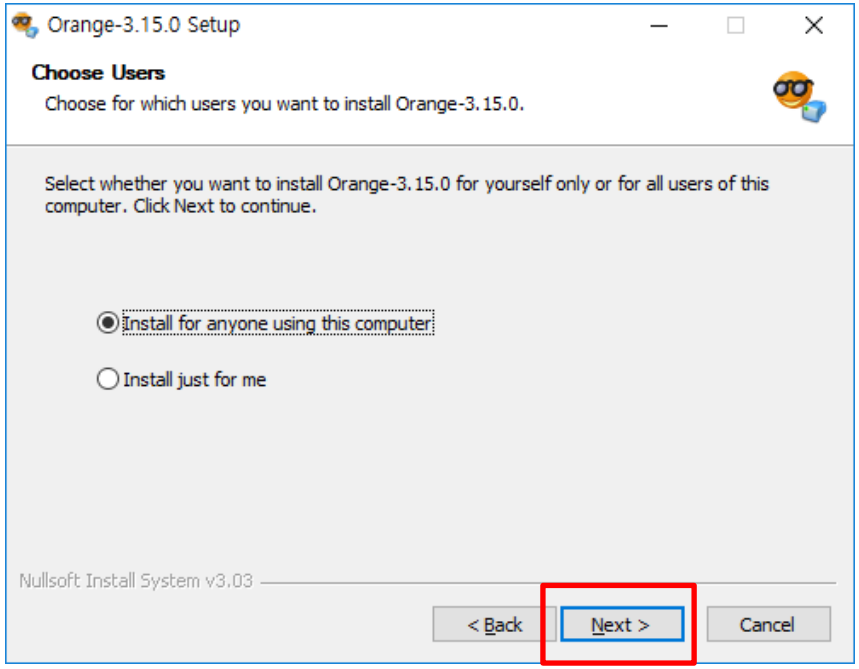
- **Install orange (1/8)**

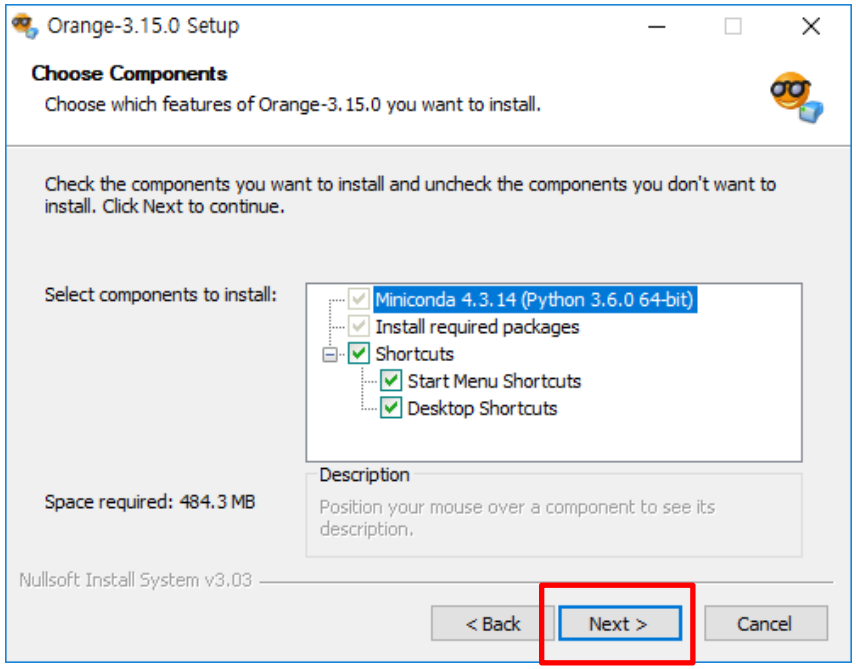




Practice Preparation (Cont'd)

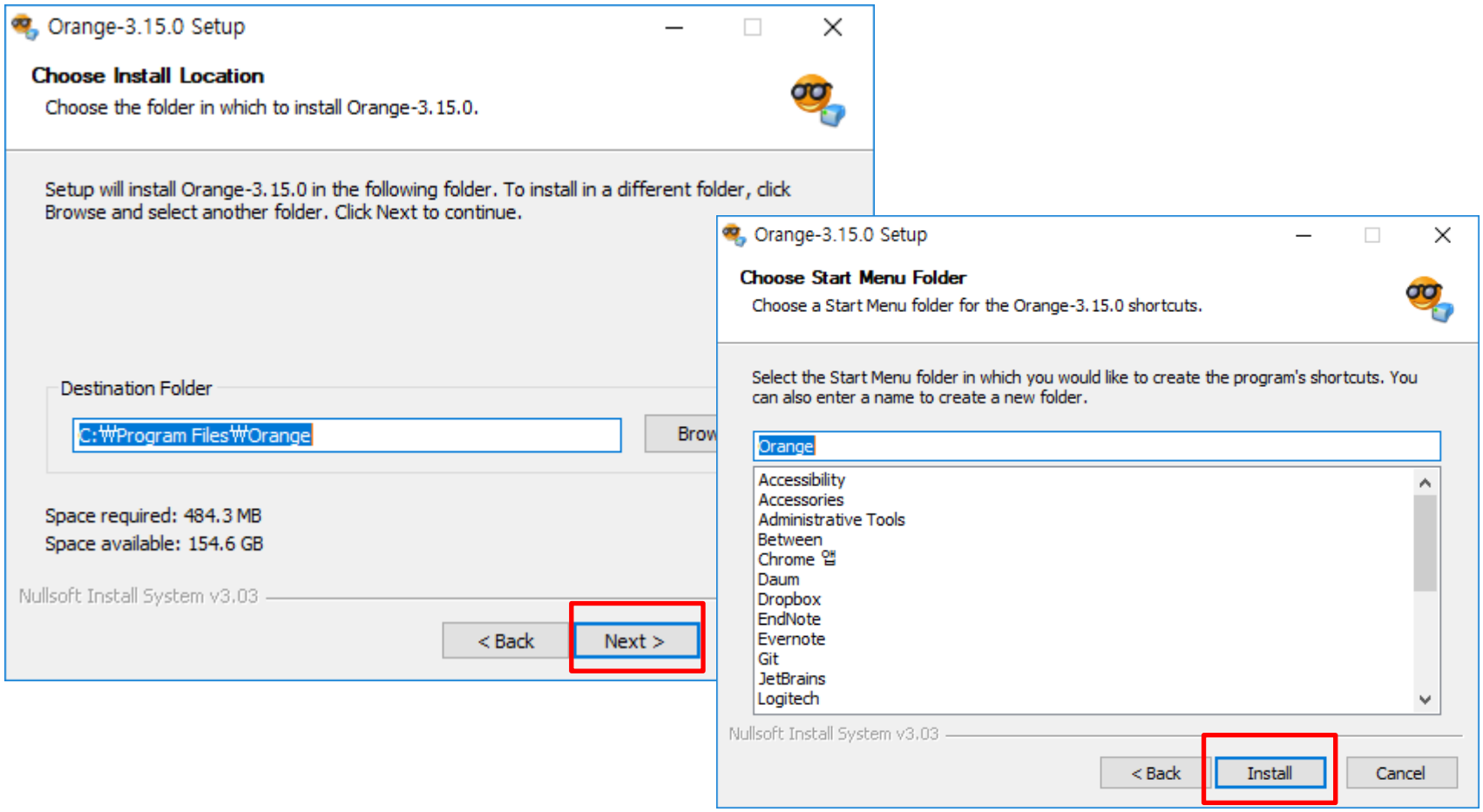
- **Install orange (2/8)**





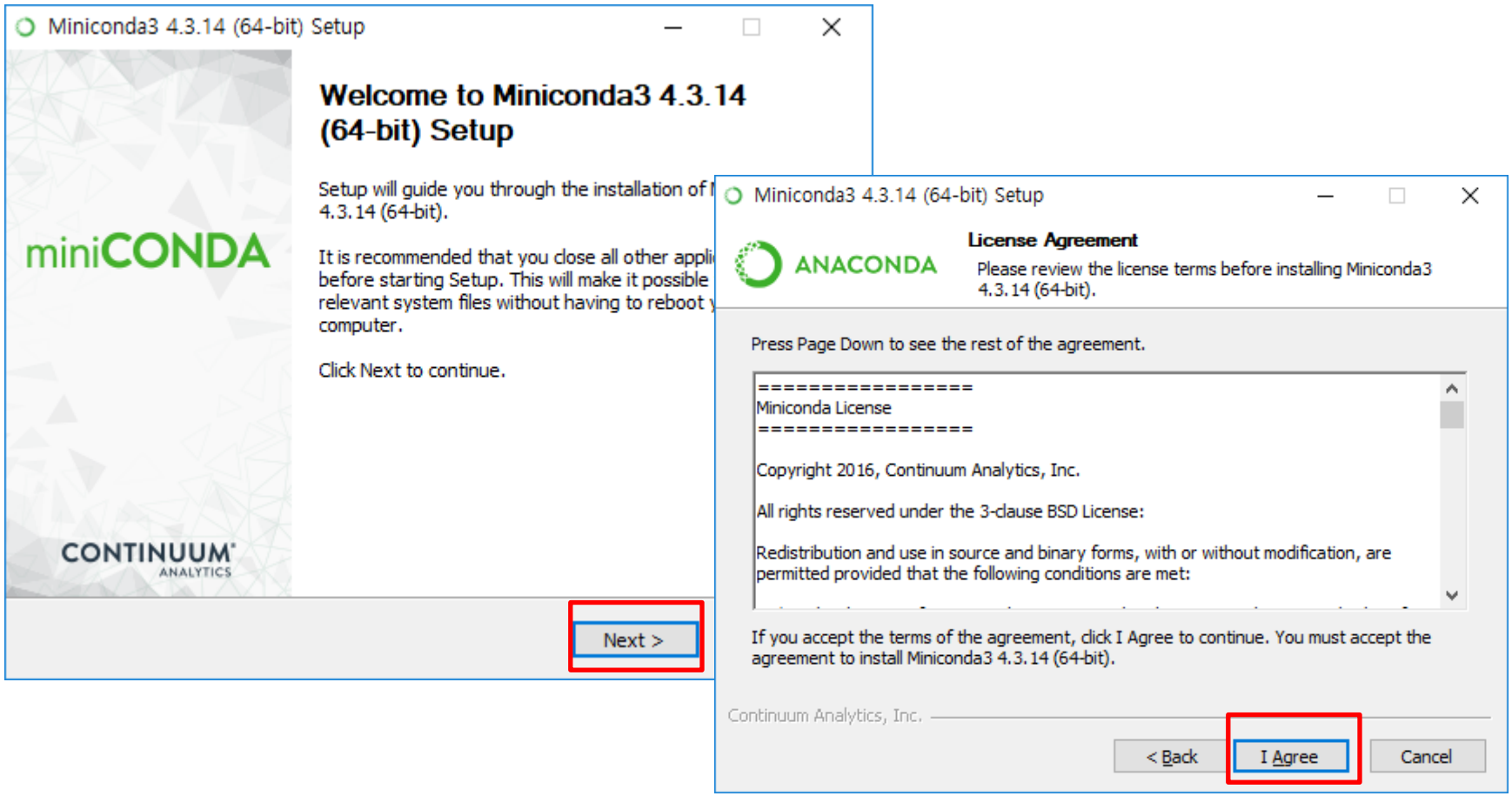
Practice Preparation (Cont'd)

- **Install orange (3/8)**



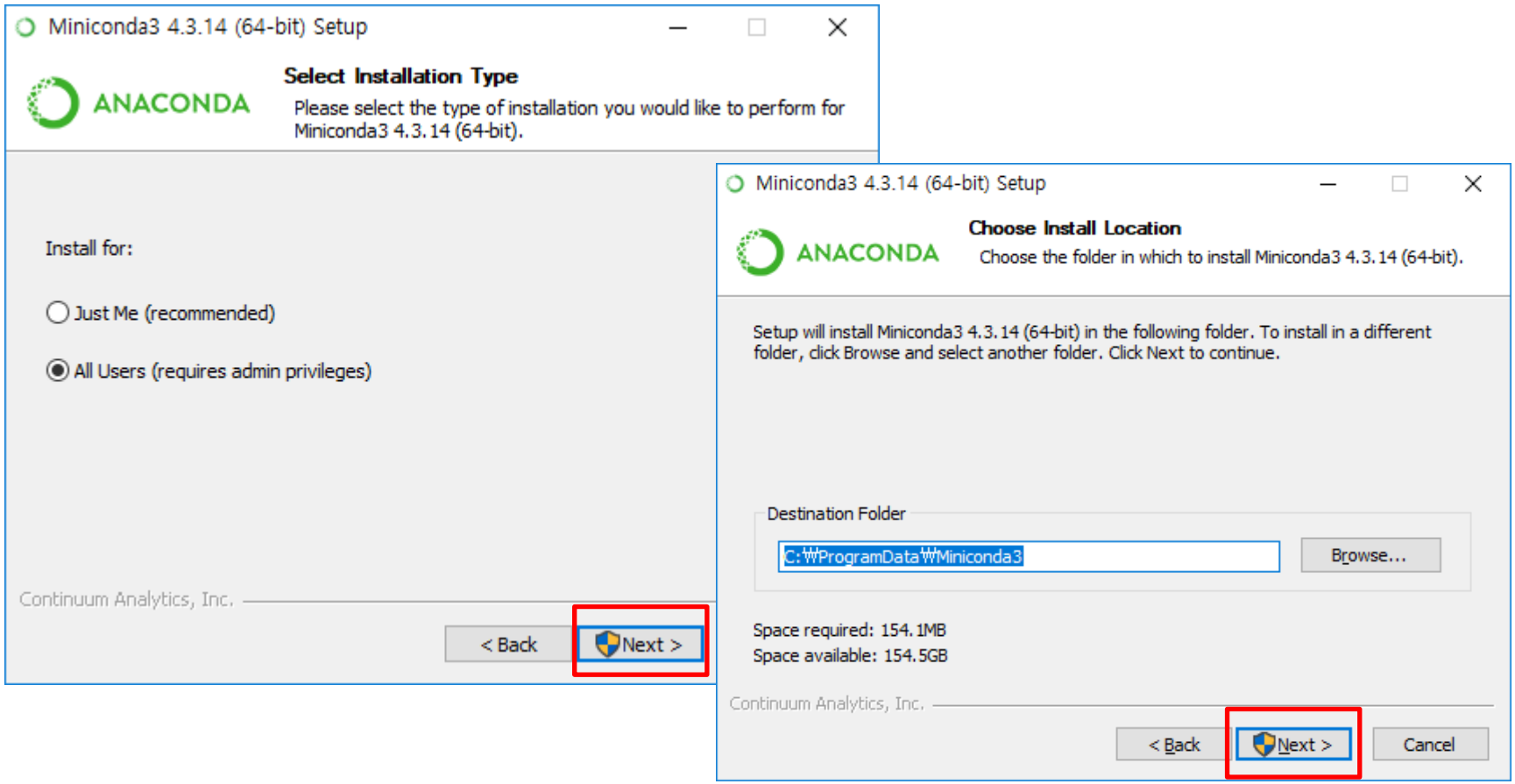
Practice Preparation (Cont'd)

- **Install orange (4/8)**



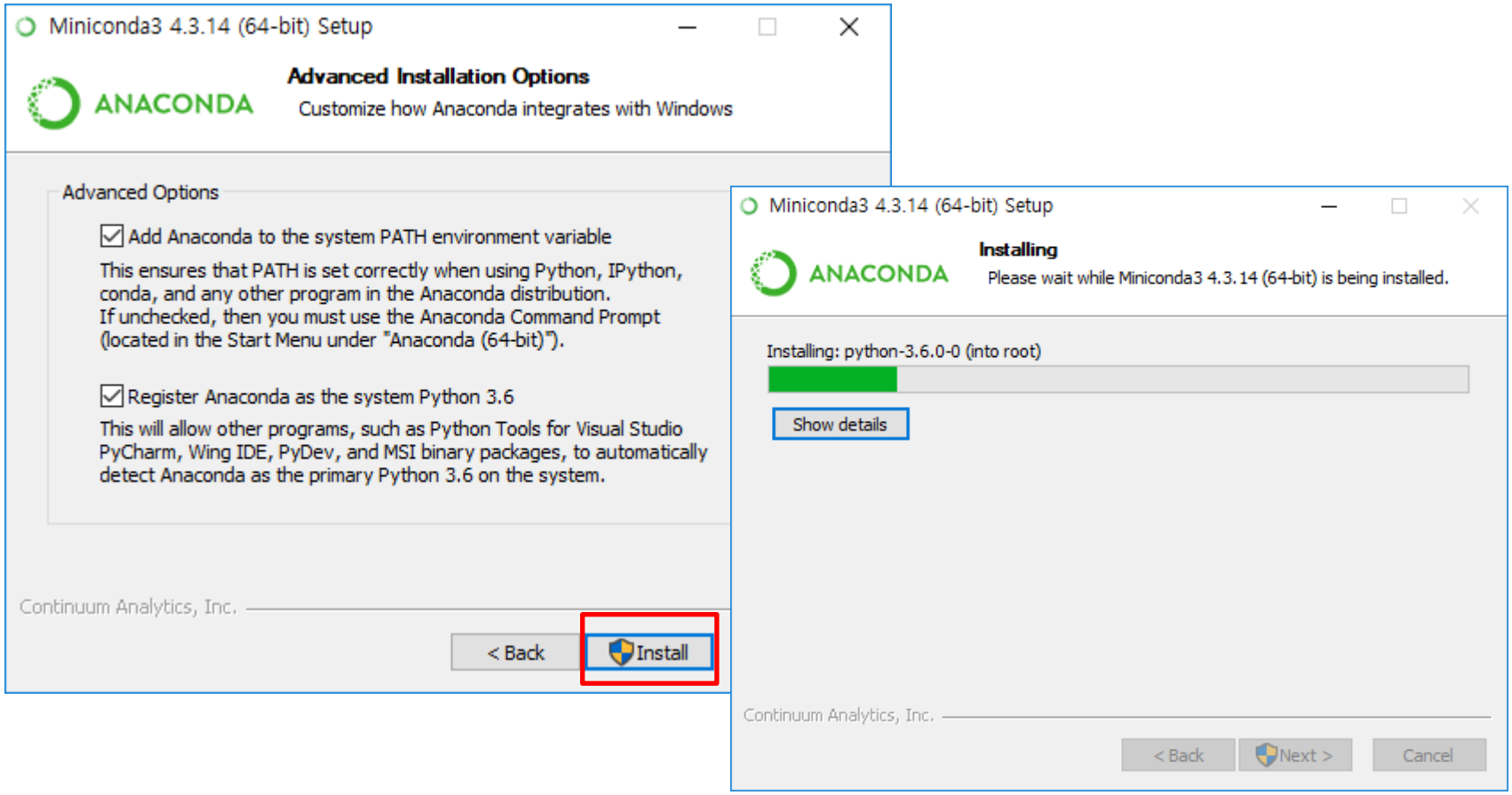
Practice Preparation (Cont'd)

- **Install orange (5/8)**



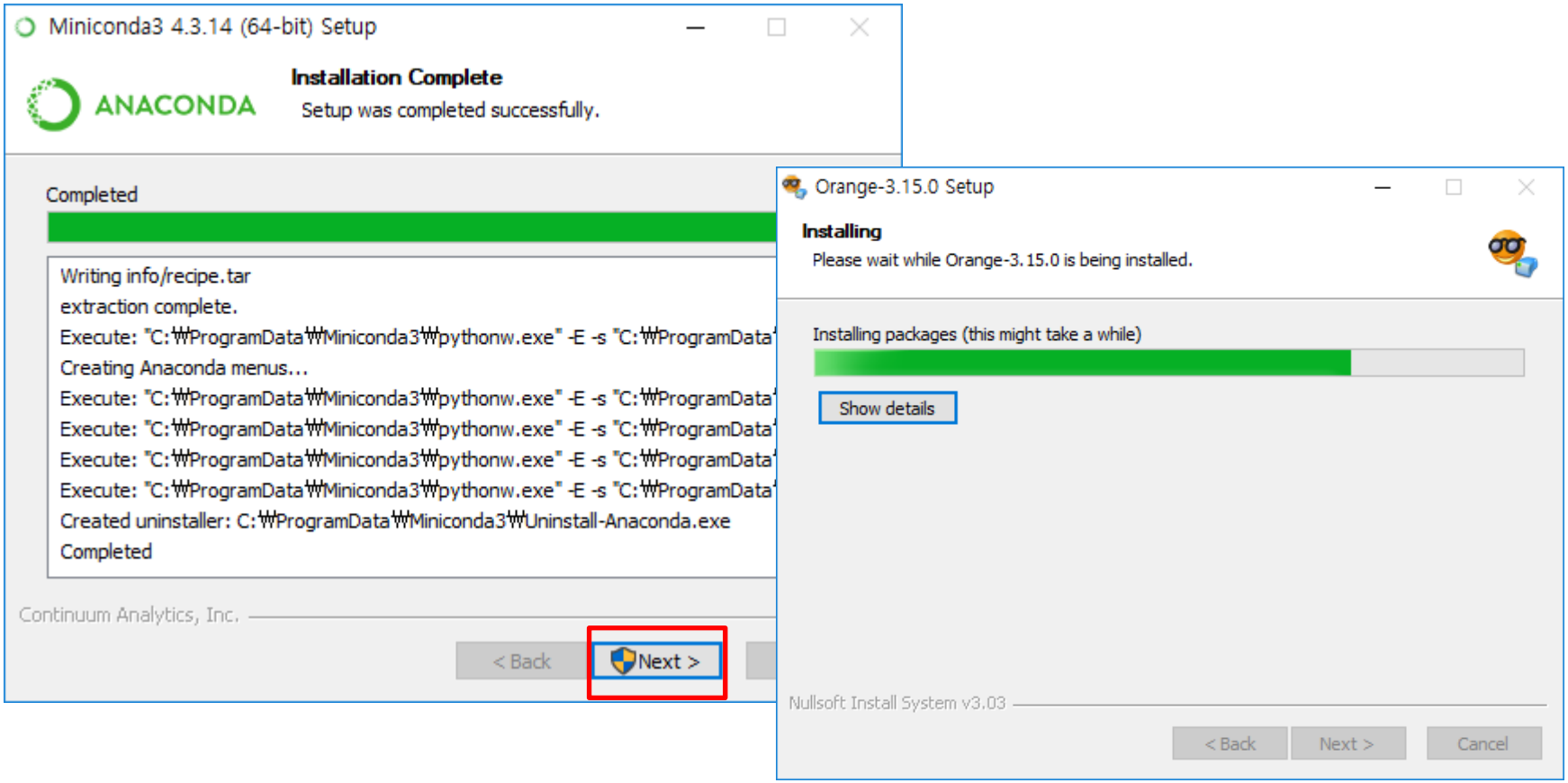
Practice Preparation (Cont'd)

- **Install orange (6/8)**



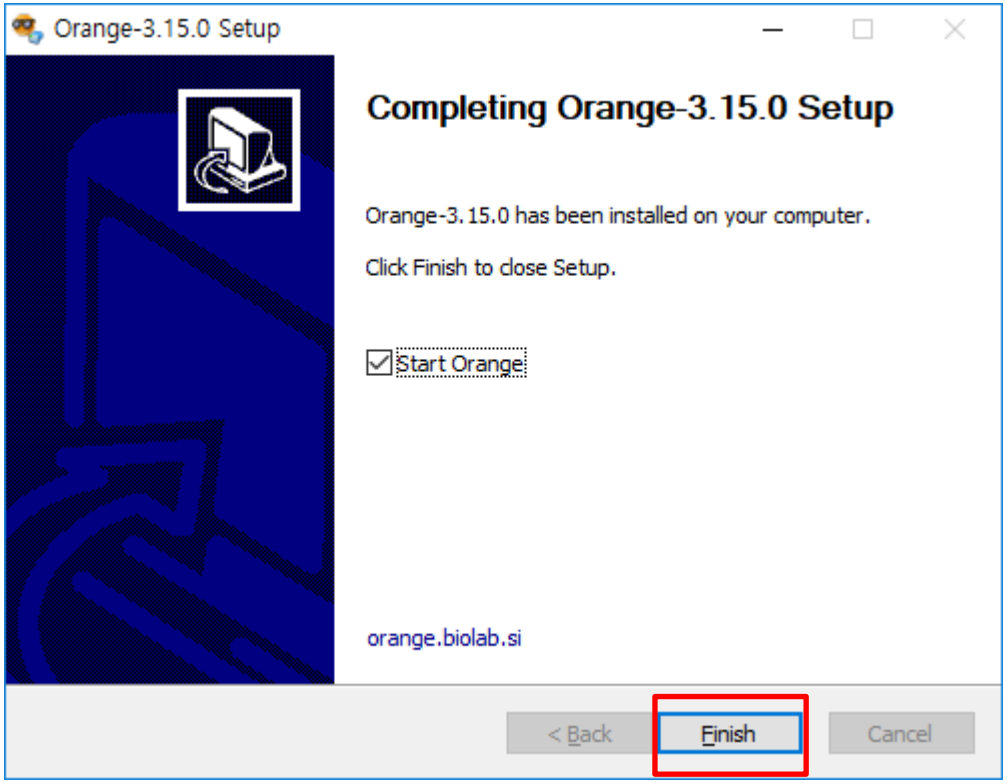
Practice Preparation (Cont'd)

- **Install orange (7/8)**



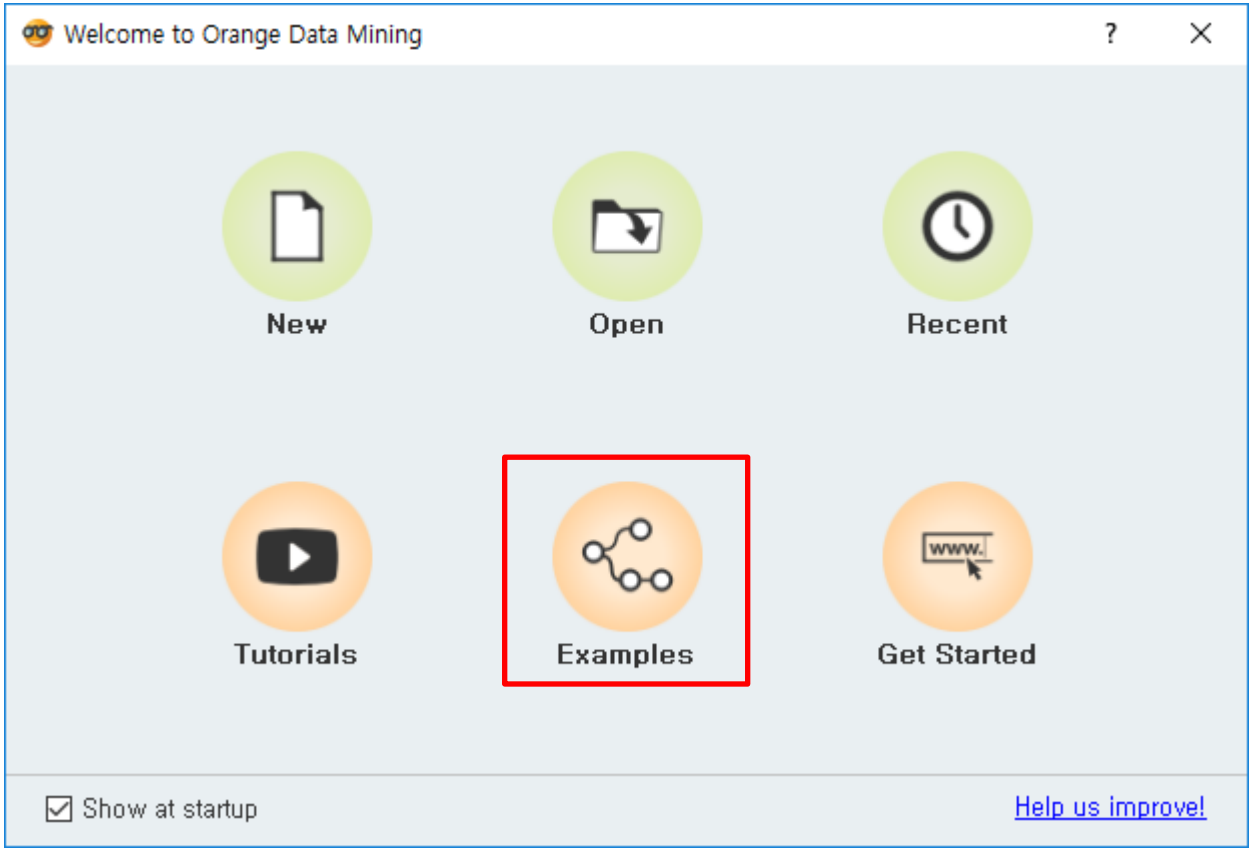
Practice Preparation

- **Install orange (8/8)**



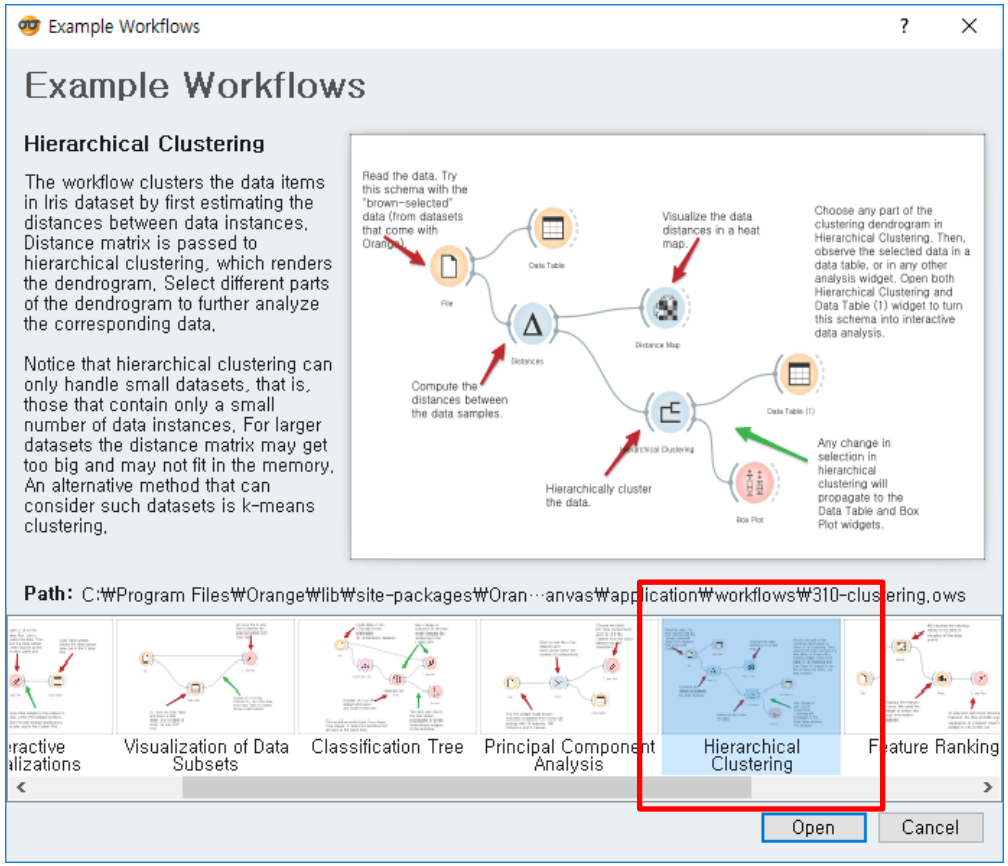
Demo practice (Cont'd)

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



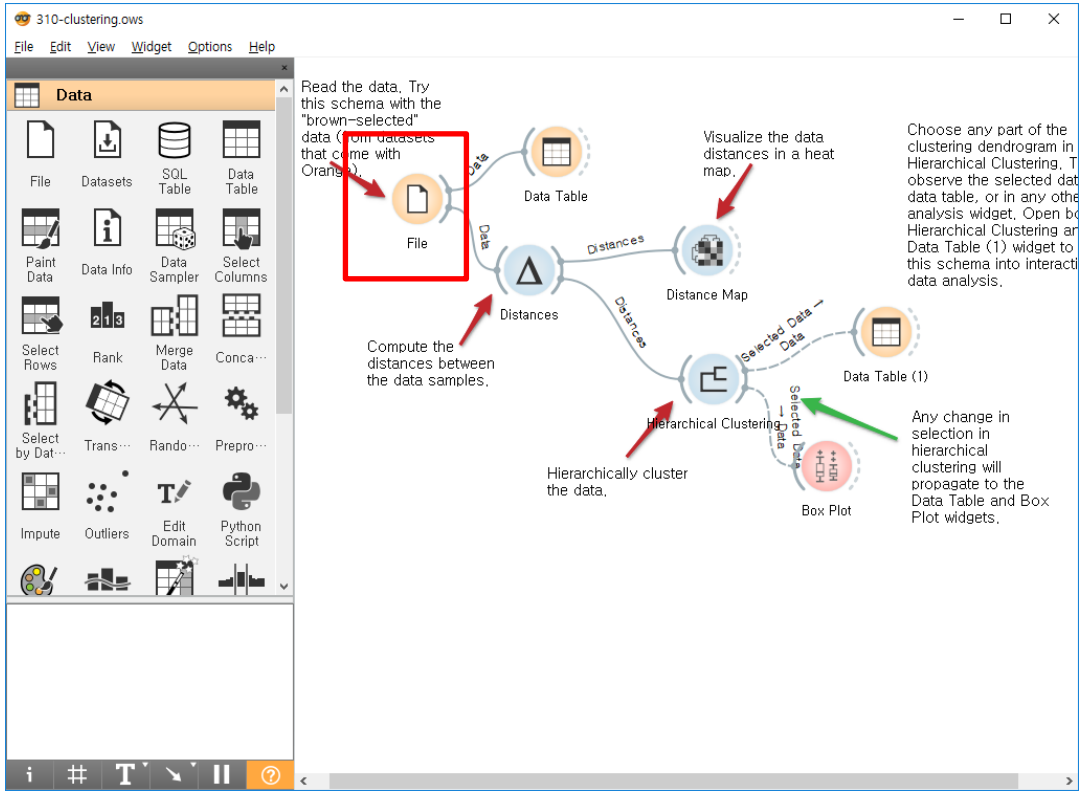
Demo practice (Cont'd)

- **The preloaded data mining workflows**
  - Hierarchical clustering : a method of clustering analysis which seeks to build a hierarchy of clusters



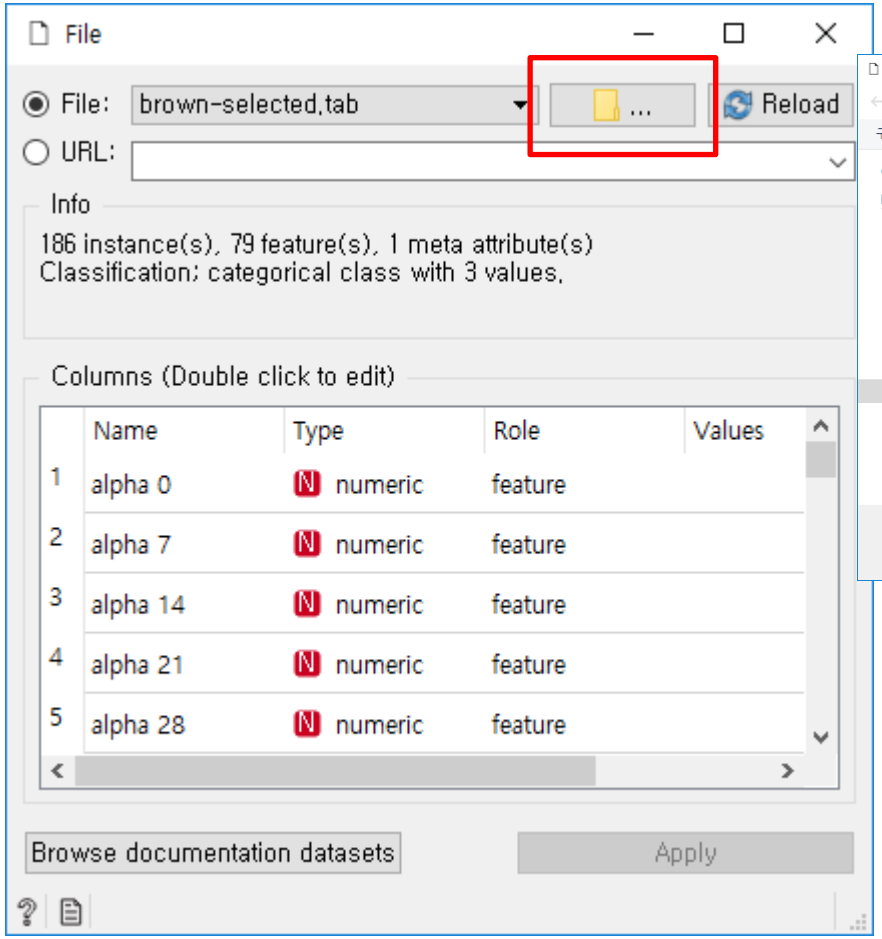
Demo practice (Cont'd)

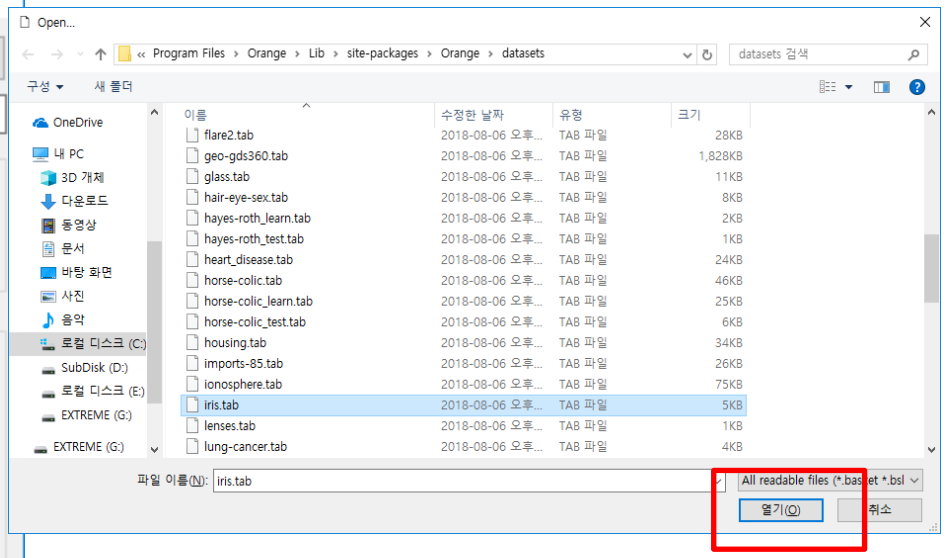
- **Orange canvas**
  - In orange, data mining workflows consist of computational components called widget
  - Widgets do all the work and exchange information
    - File widget : send data to the data table widget and distance widget
    - Distance widget : compute the distances between the data samples



Demo practice (Cont'd)

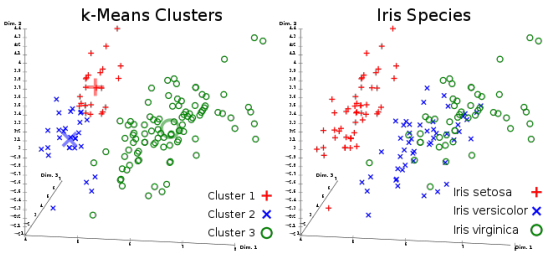
- **Input data on the file widget**
  - chose *iris.tab* from the list of pre-installed data

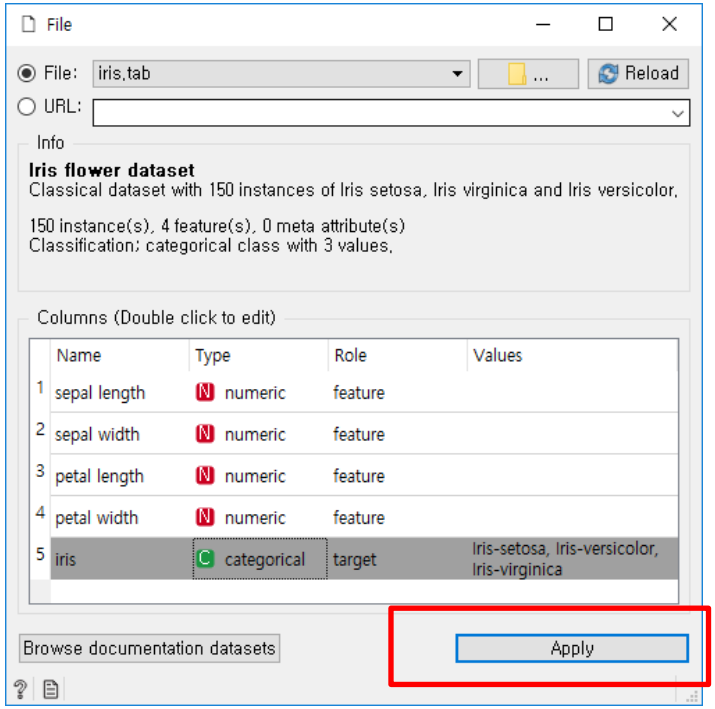




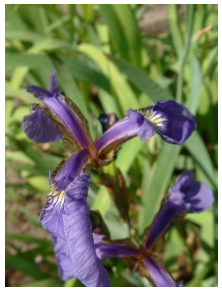
Demo practice (Cont'd)

- **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



Iris setosa

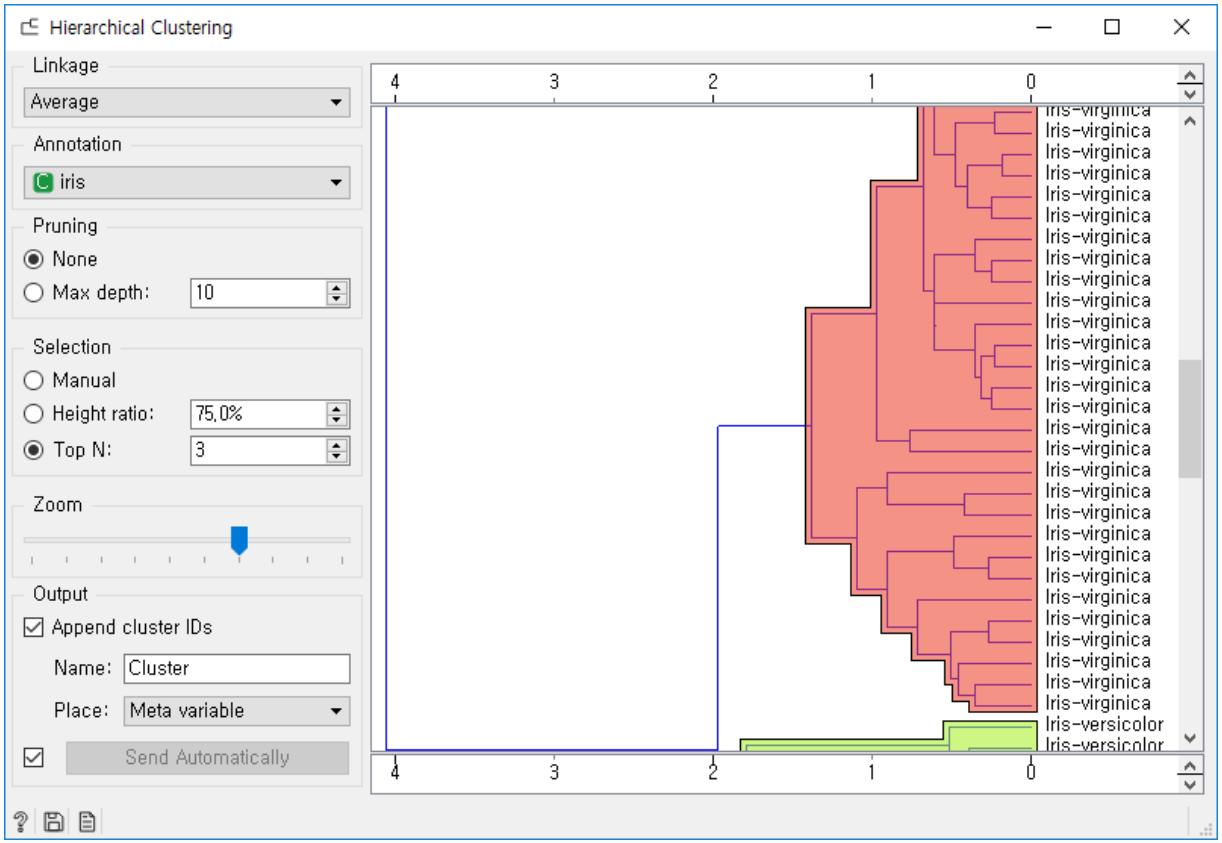
Iris veriscolor

Iris virginica

ref : https://en.wikipedia.org/wiki/Iris_flower_data_set

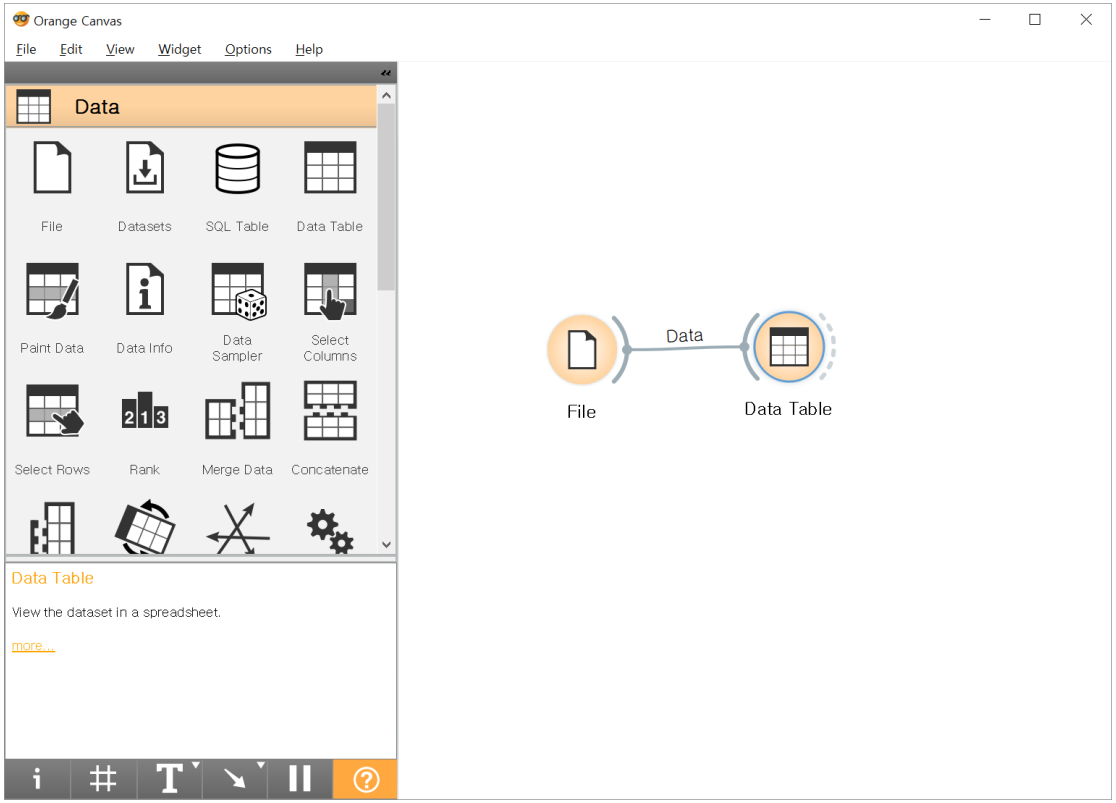
Demo practice

- **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



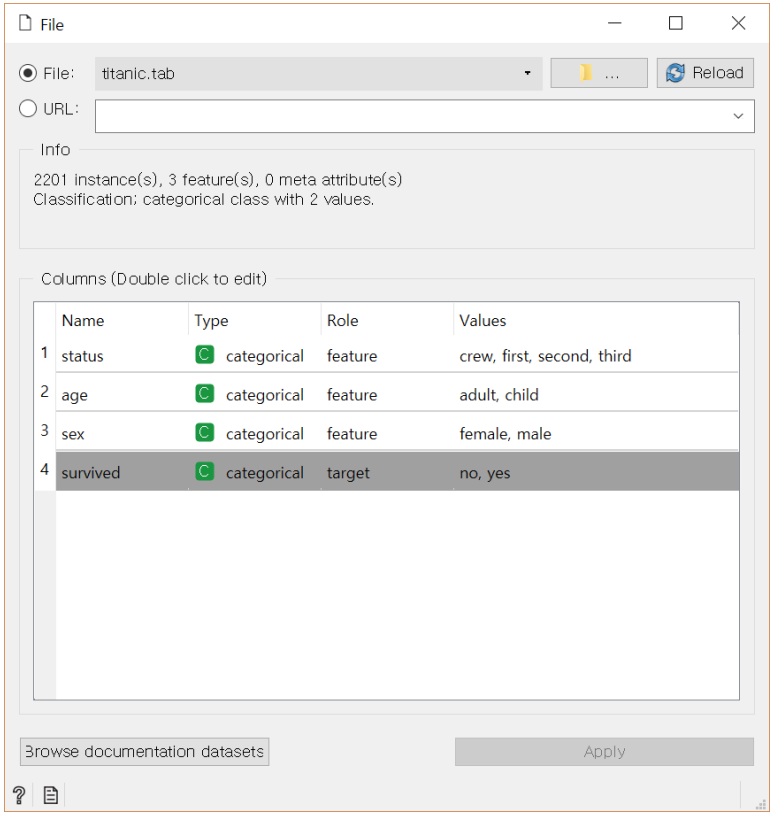
RMS *Titanic* departing Southampton on 10 April 1912[1]

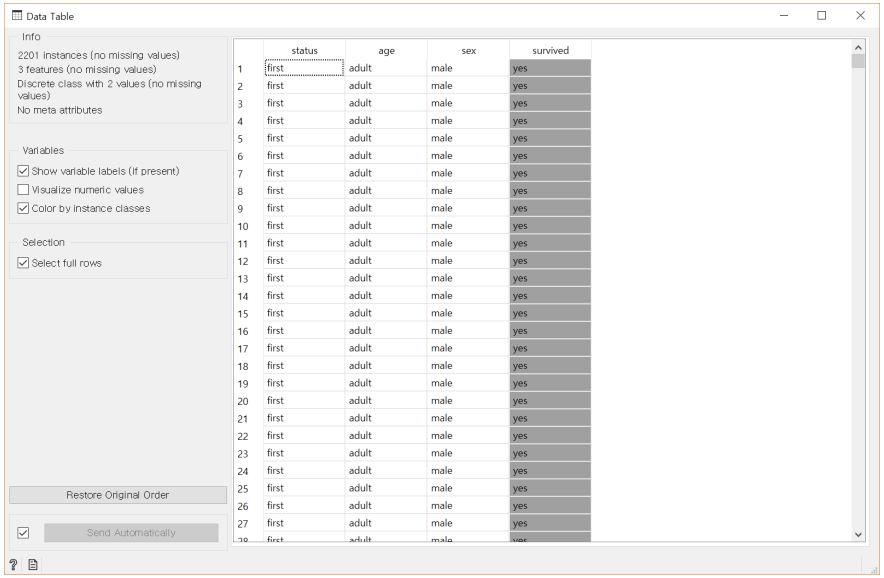
*RMS : Royal mail ship

[1] https://en.wikipedia.org/wiki/RMS_Titanic

Basic practice (Cont'd)

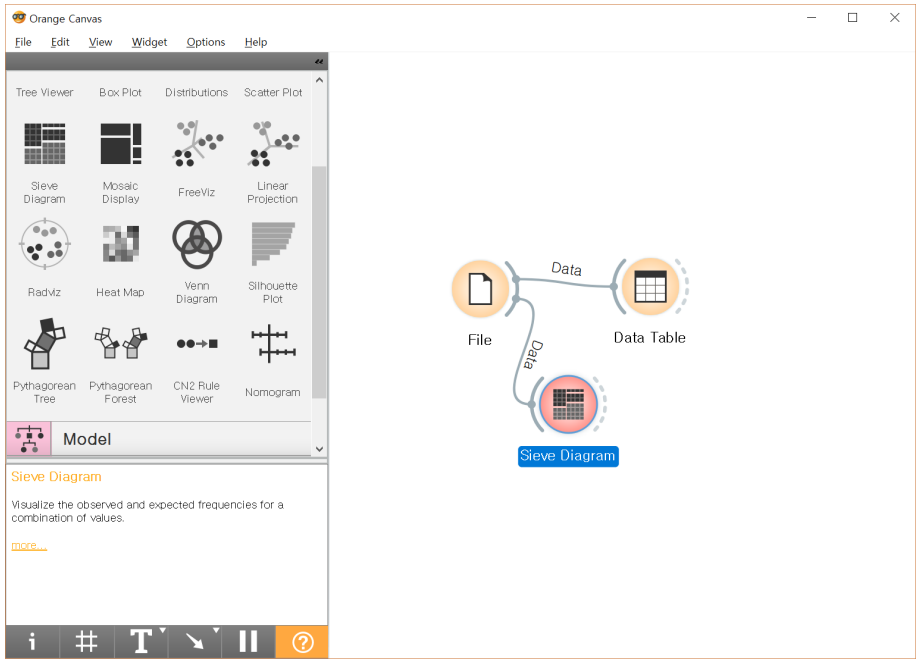
- **Load and parsing data**
  - The widgets automatically transferred the loaded data to all the connected widgets

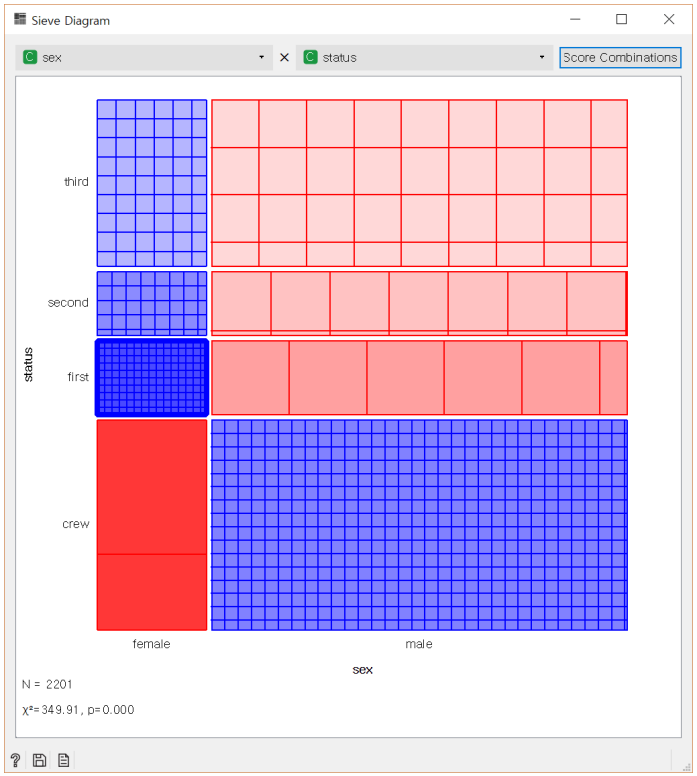




Basic practice

- **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[1]

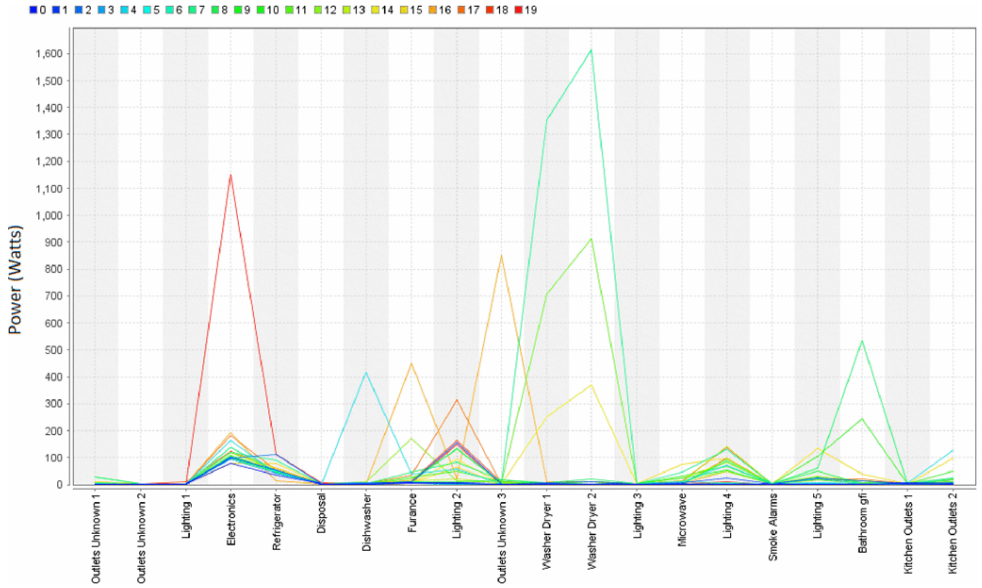




[1] Riedwyl, Hans, and M. Schüpbach. "Parquet diagram to plot contingency tables." *Softstat* 93 (1994): 293-299.
[2] http://www.datavis.ca/online/sieve/

Project assignment (Cont'd)

- **Analysis households electricity consumption**
  - Dataset
    - The reference energy disaggregation dataset (REDD) Version 1.0[1]
    - 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

[1] 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
- **Resources**
  - https://github.com/junseokpark/resources/datamining
- **web-based competition site for datamining**
  - https://www.kaggle.com/competitions

Thank you