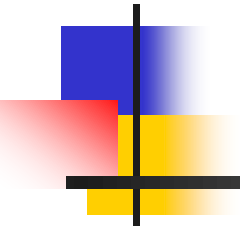
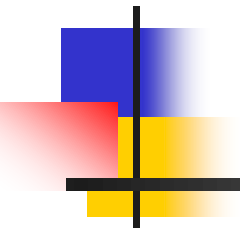
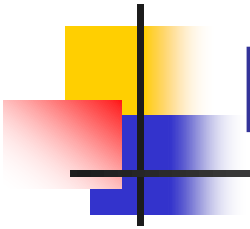
Data Mining – Installation of Weka and Python



Kyuseok Shim
Seoul National University
http://kdd.snu.ac.kr/~shim

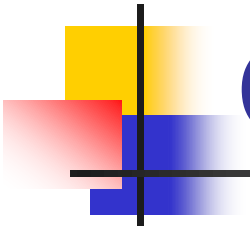


Weka



Data Mining with Weka

- What's Weka?
  - *A bird found only in New Zealand?*
- Data mining tool
- Machine learning algorithms for data mining tasks
  - 100+ algorithms for classification
  - 75 for data preprocessing
  - 25 to assist with feature selection
  - 20 for clustering, finding association rules, etc

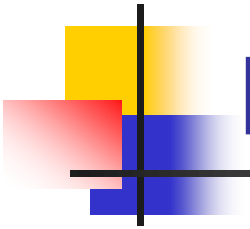


Getting started with Weka

- Install Weka
- Explore the “Explorer” interface
- Explore some datasets
- Build a classifier
- Interpret the output
- Use filters
- Visualize your data set

Download & Install Weka

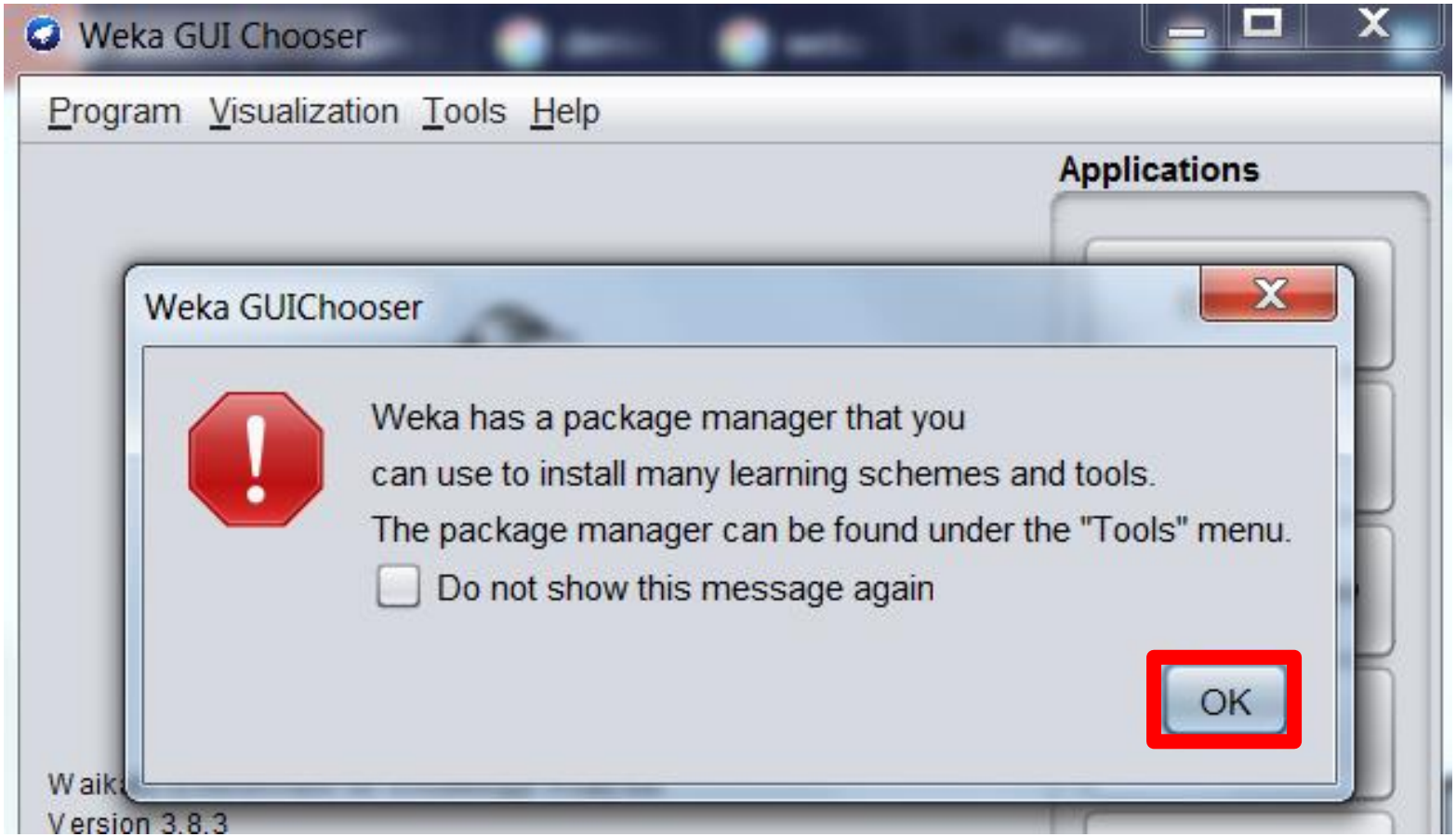
- Download Weka 3.8
  - https://www.cs.waikato.ac.nz/ml/weka/downloading.html
  - For 64-bit Windows
    - http://prdownloads.sourceforge.net/weka/weka-3-8-3jre-x64.exe
  - For 32-bit Windows
    - http://prdownloads.sourceforge.net/weka/weka-3-8-3jre.exe
- The installation is simple
  - Just select 'Next', 'I Agree' or 'Install'



Download the Additional Dataset

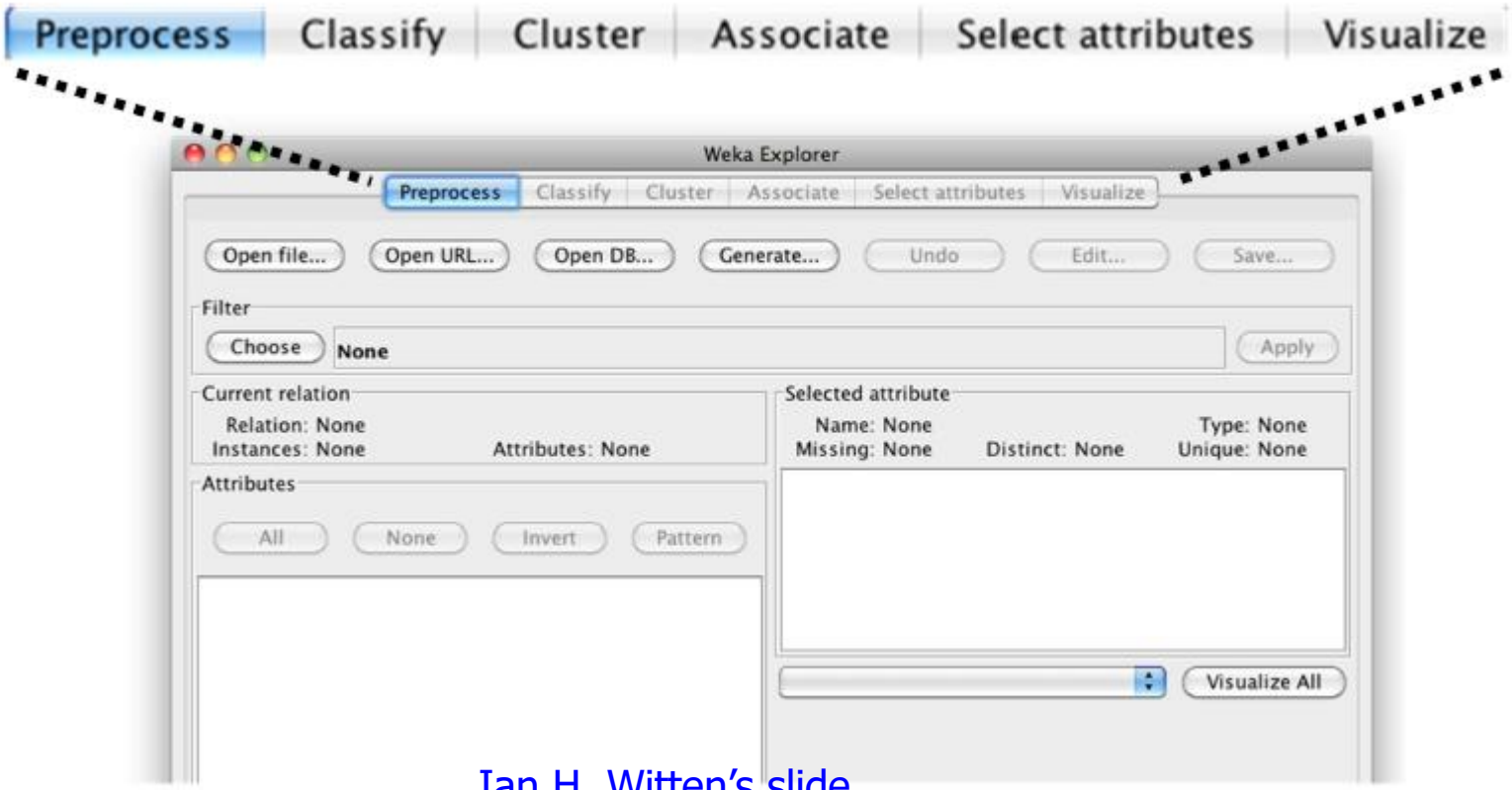
- http://kdd.snu.ac.kr/weka/
  - cluster2.arff

Getting Started with Weka



Exploring the Explorer

Exploring the Explorer



Ian H. Witten's slide

A Dataset: weather.nominal.arff

attributes

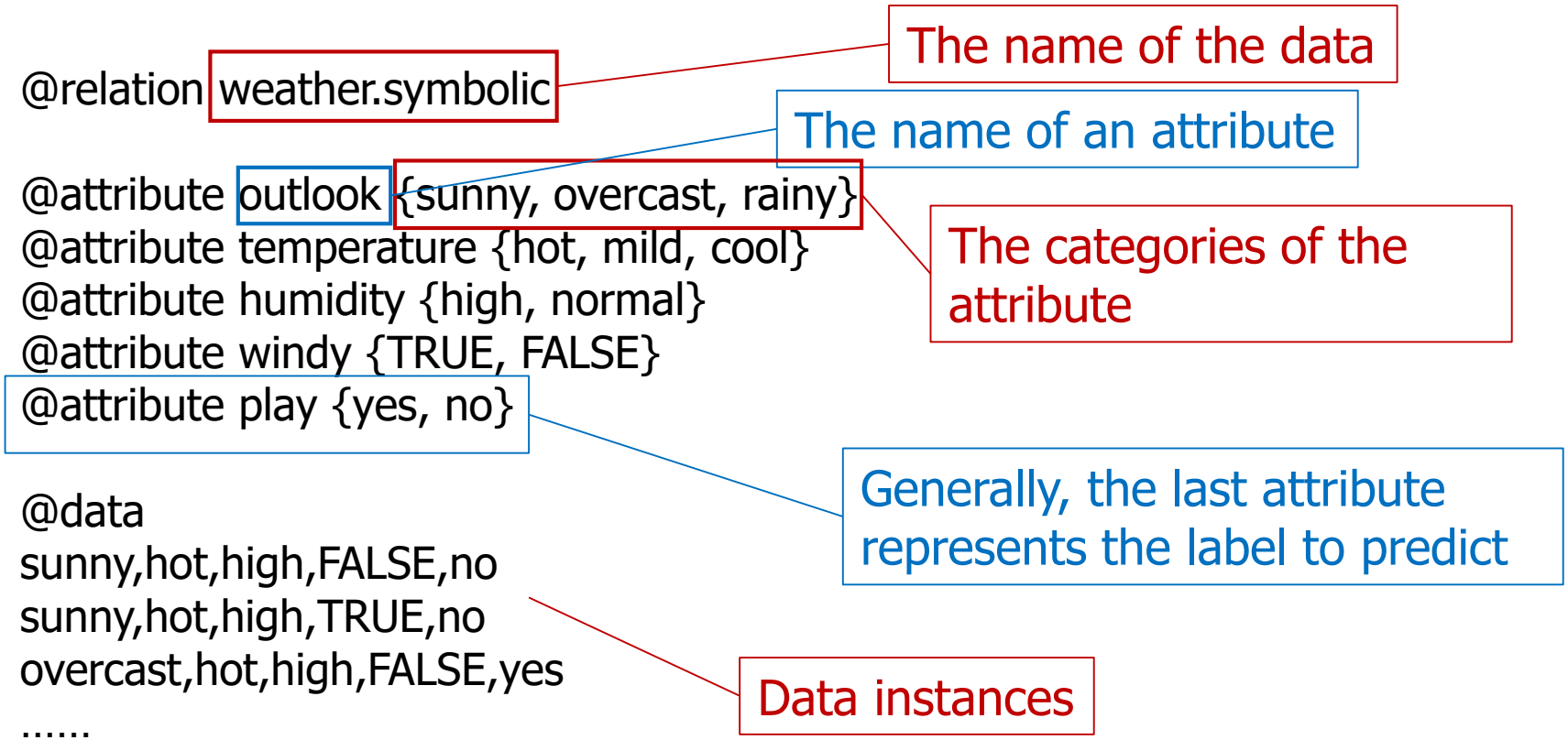
instances

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

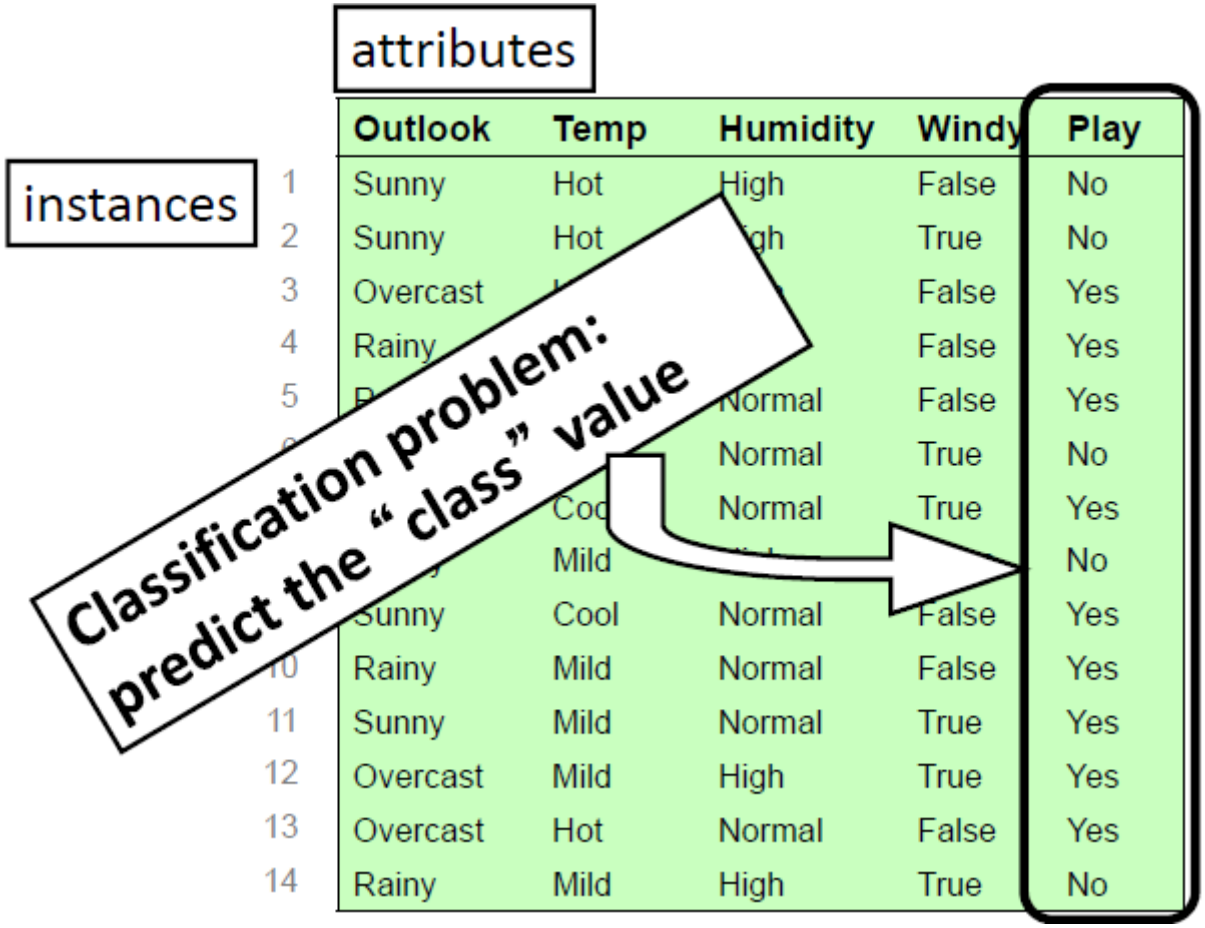
Ian H. Witten's slide

weather.nominal.arff

- `C:\Program Files\Weka-3-8\data\weather.nominal.arff`



Classification Problem



	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	gh	True	No
3	Overcast			False	Yes
4	Rainy			False	Yes
5			Normal	False	Yes
			Normal	True	No
		Coo	Normal	True	Yes
		Mild			No
	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Ian H. Witten's slide

weather.numeric.arff

- `C:\Program Files\Weka-3-8\data\weather.nominal.arff`

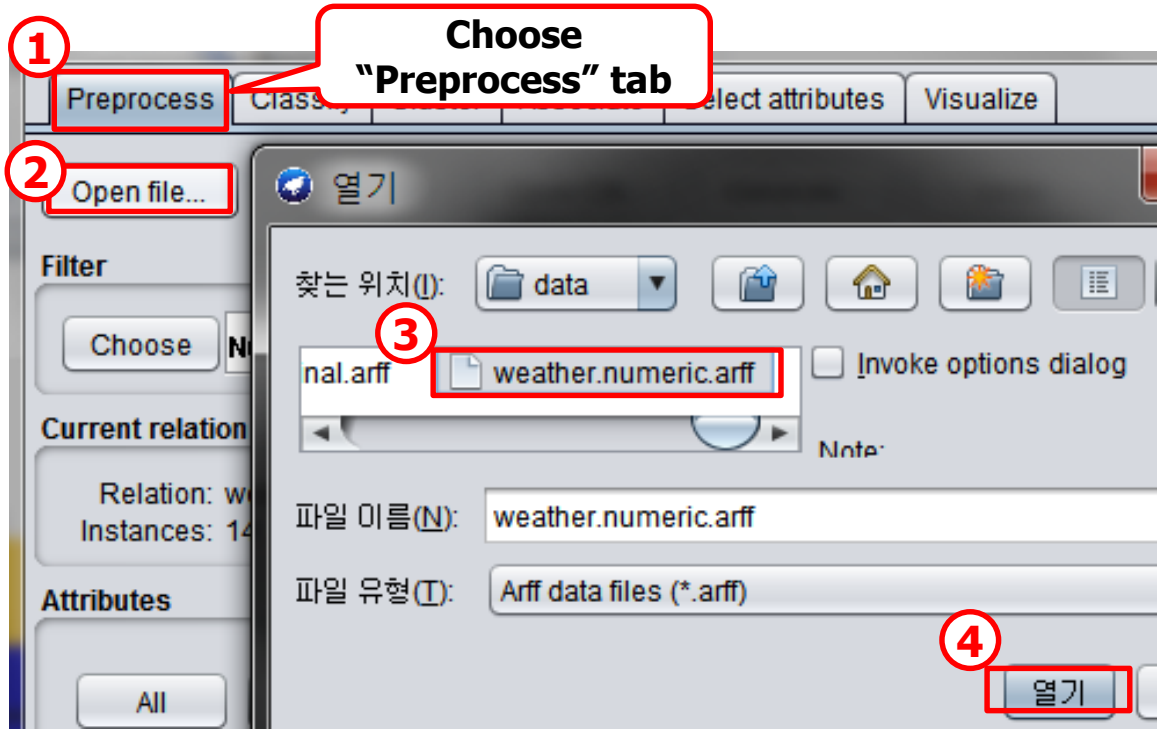
```
@relation weather

@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
……
```

A numeric attribute

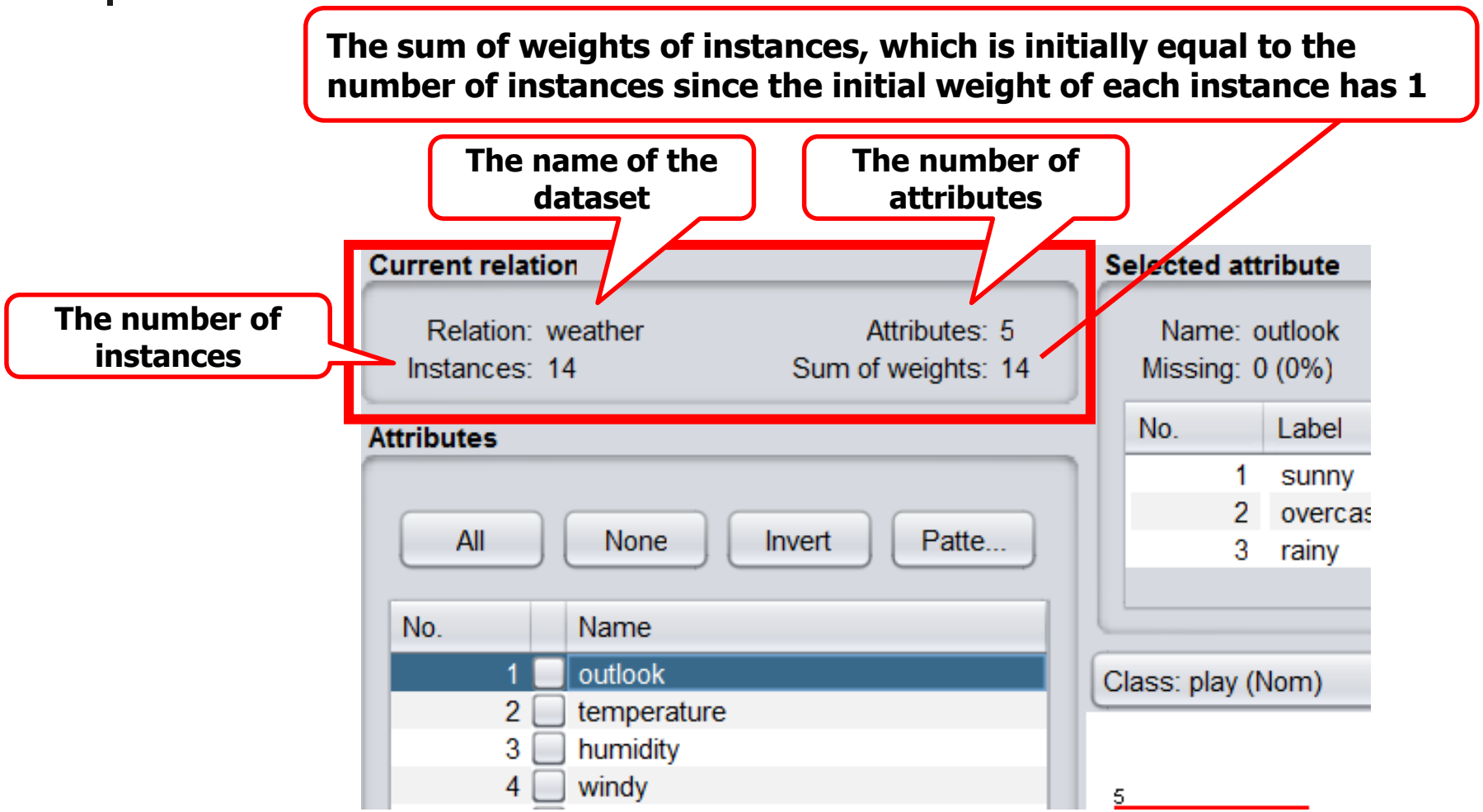
outlook	temperate	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
…	…	…	…	…

Open the Dataset

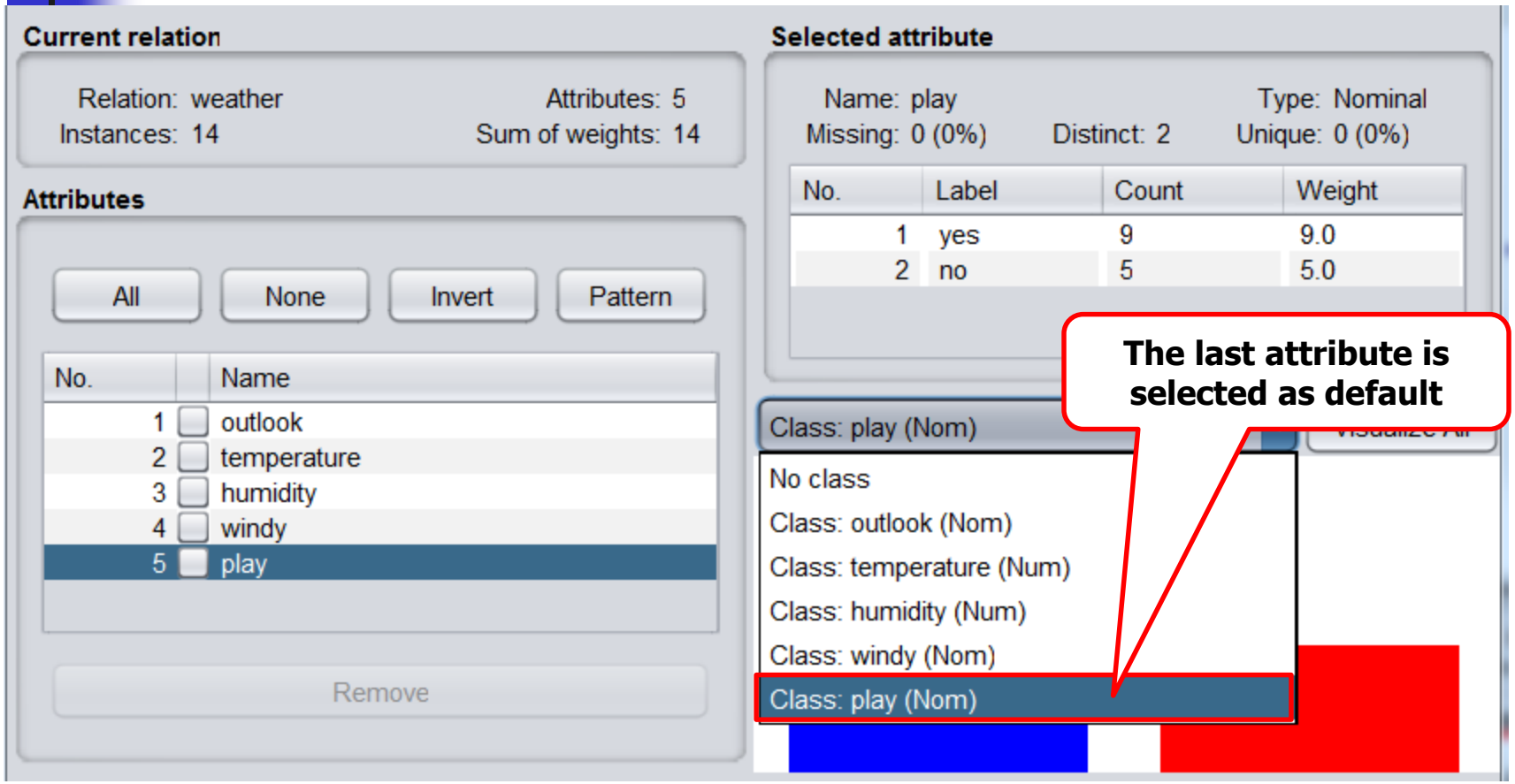
- C:\Program Files\Weka-3-8\data\weather.numeric.arff



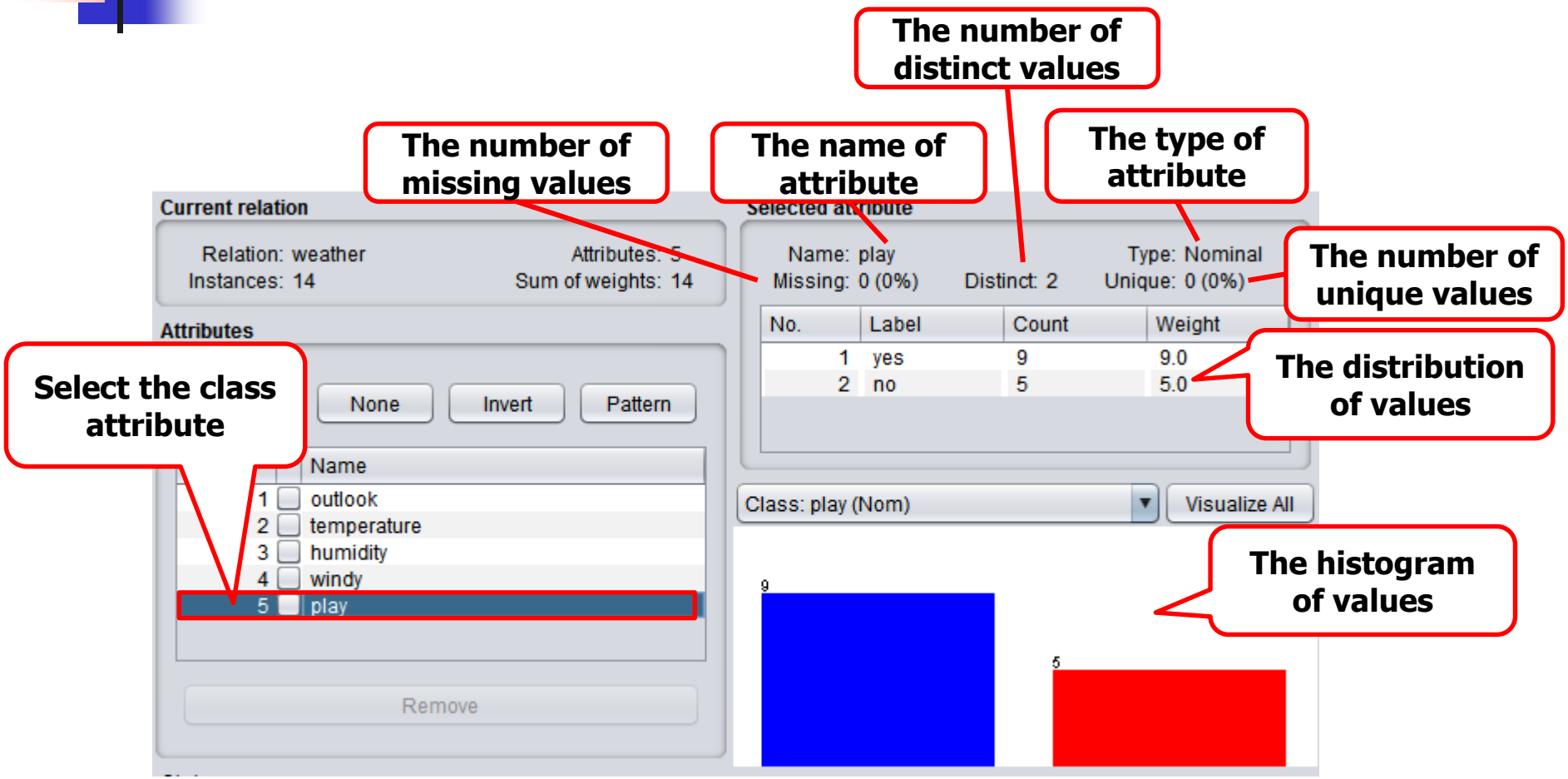
The Statistics of the Dataset



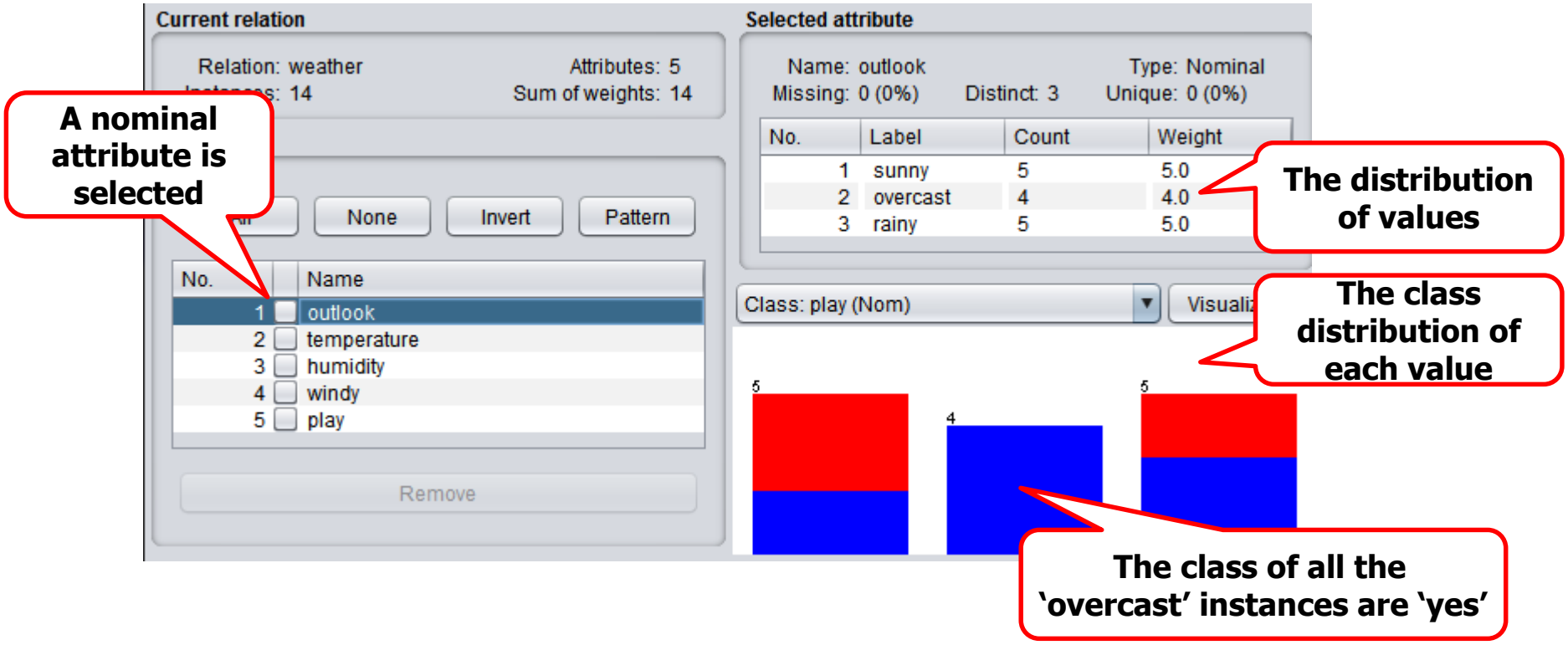
Select the Class



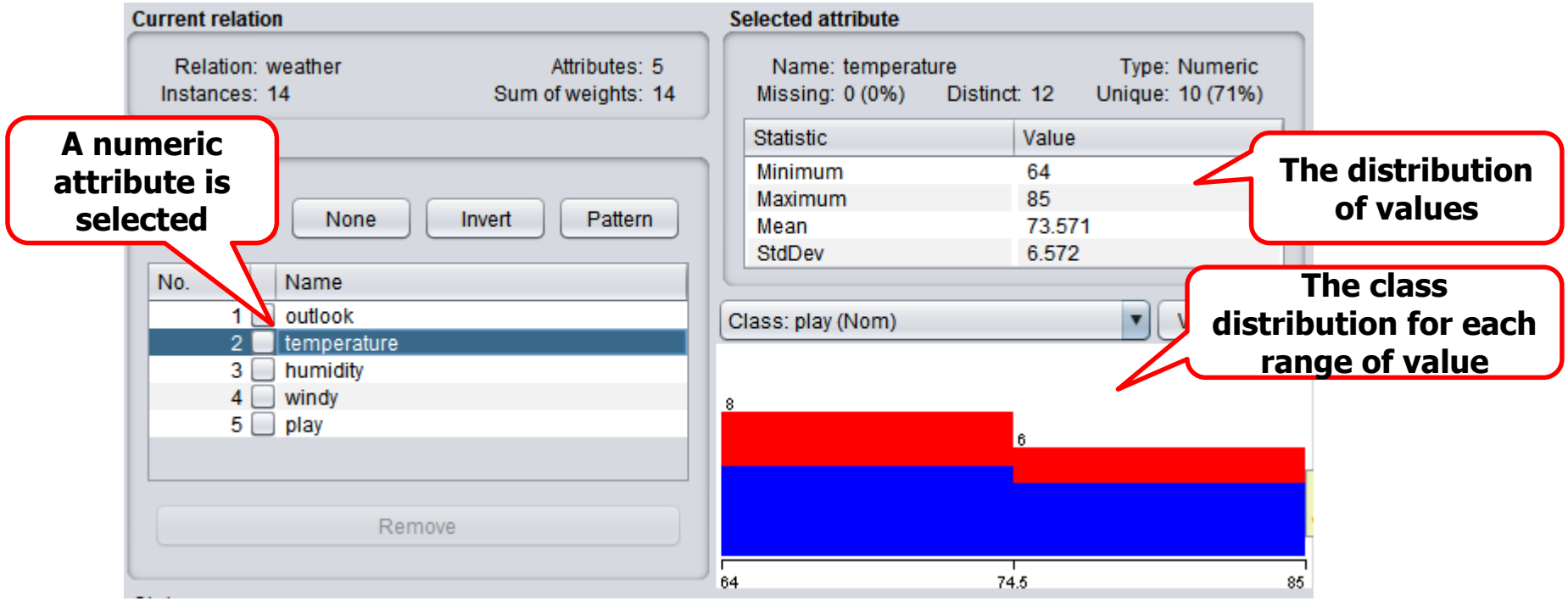
The Statistics of the Class

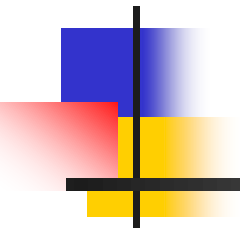


The Statistics of an Attribute



The Statistics of an Attribute





Weka – Using CSV Files

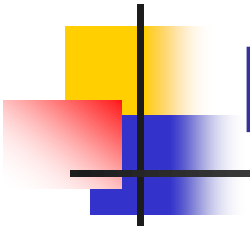
CSV Files

- CSV files could be easily exported from Microsoft Excel or Google Spreadsheet

	A	B	C	D	E
1	outlook	temperature	humidity	windy	play
2	sunny	85	85	FALSE	no
3	sunny	80	90	TRUE	no
4	overcast	83	86	FALSE	yes
5	rainy	70	96	FALSE	yes
6	rainy	68	80	FALSE	yes
7	rainy	65	70	TRUE	no

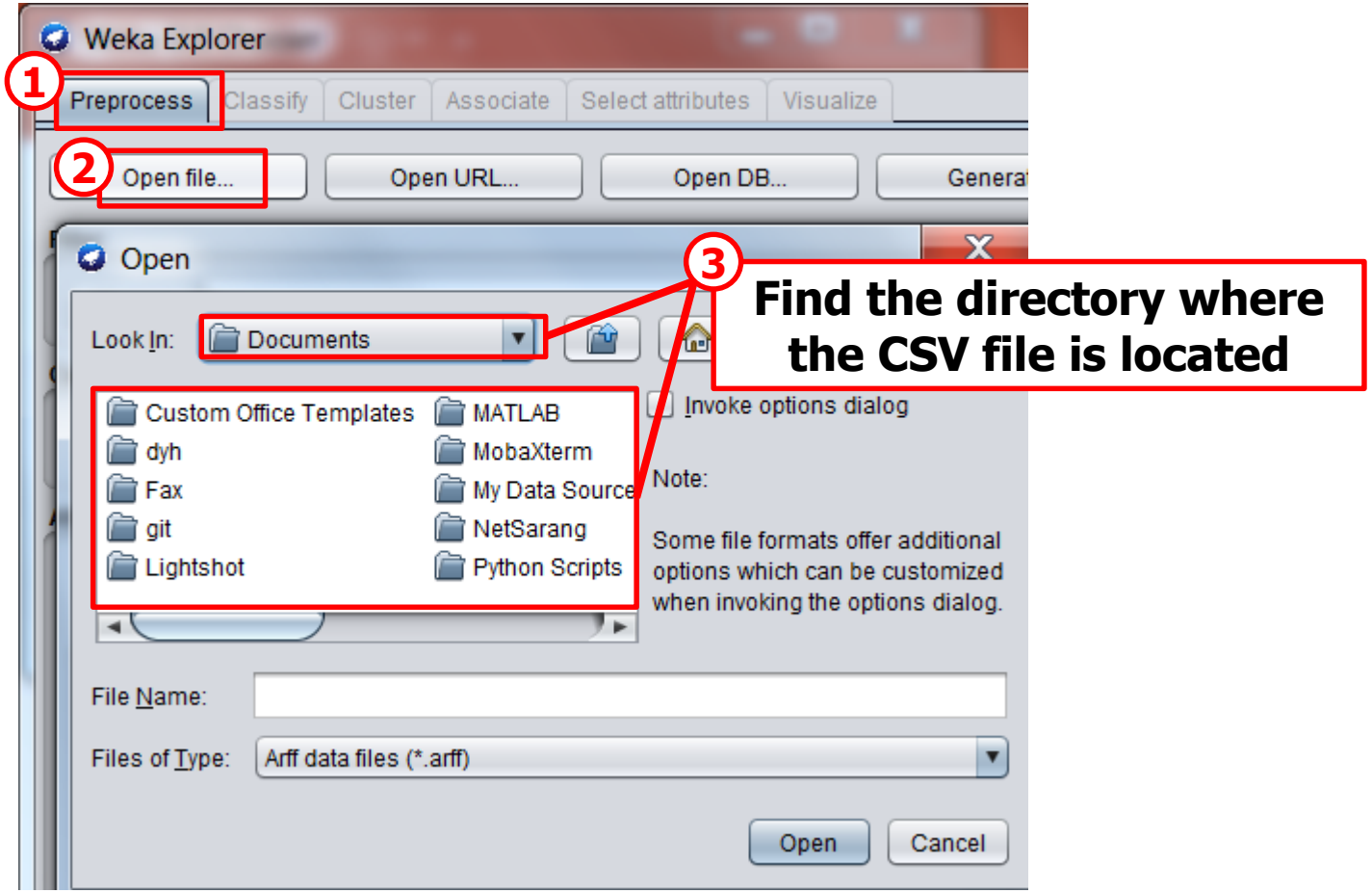
```
outlook,temperature,humidity,windy,play
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
```

Download a Sample CSV File



- http://kdd.snu.ac.kr/weka/
  - Download the **test.csv**

Using CSV Files for Weka

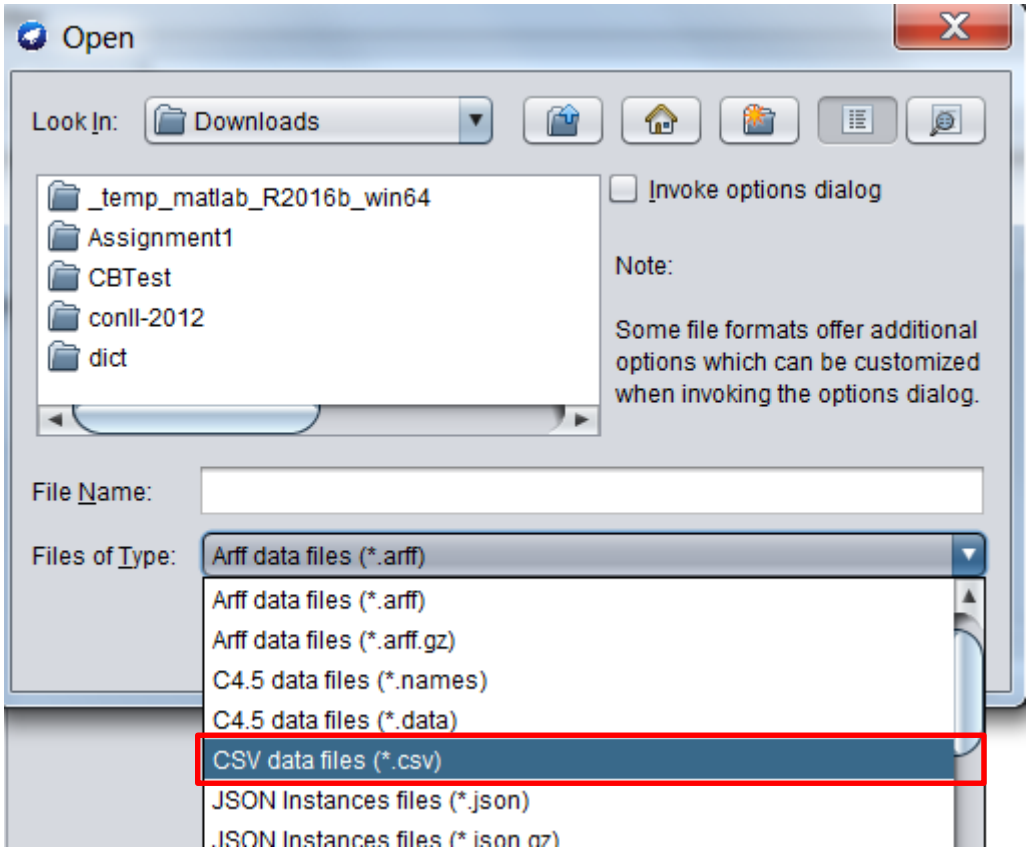


Using CSV Files for Weka

- IF the test.csv is downloaded in `C:\Users\dyhong\Downloads`

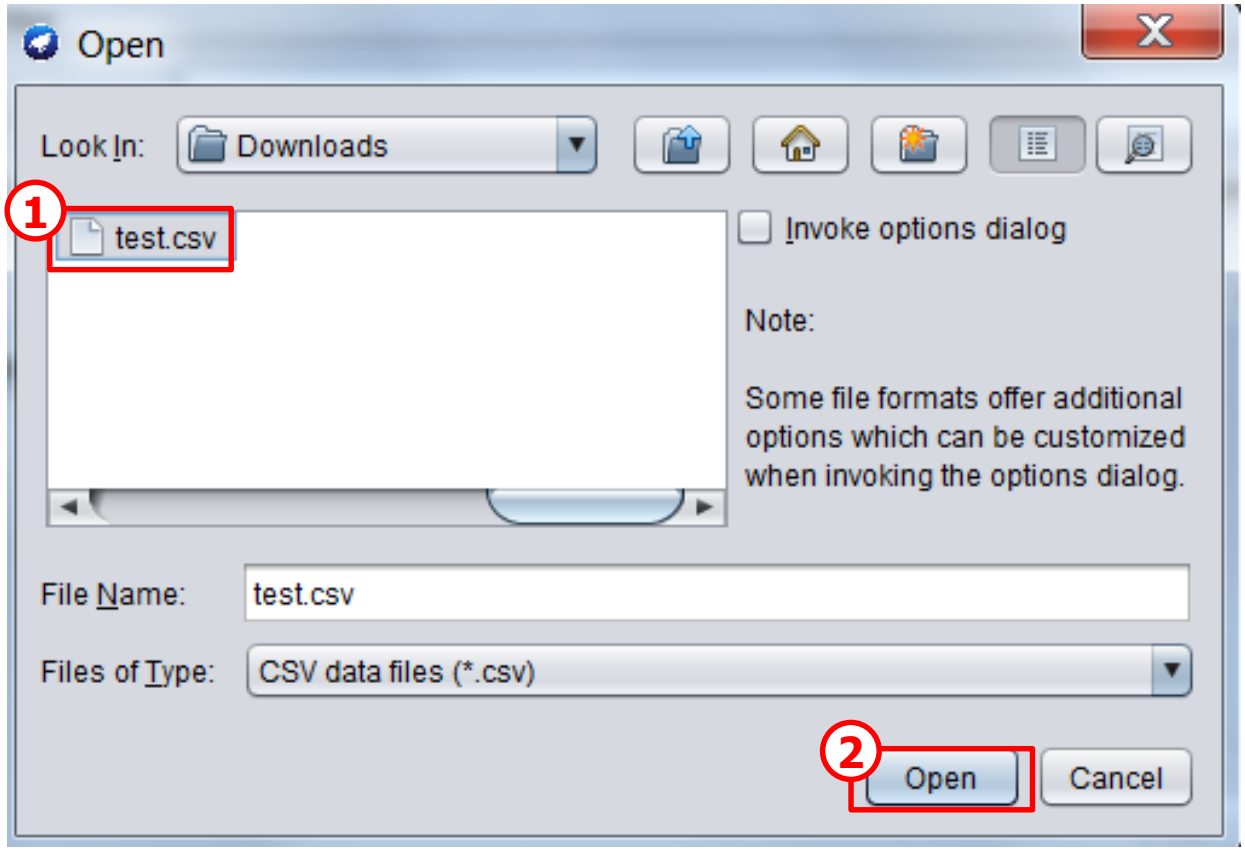
Using CSV Files for Weka

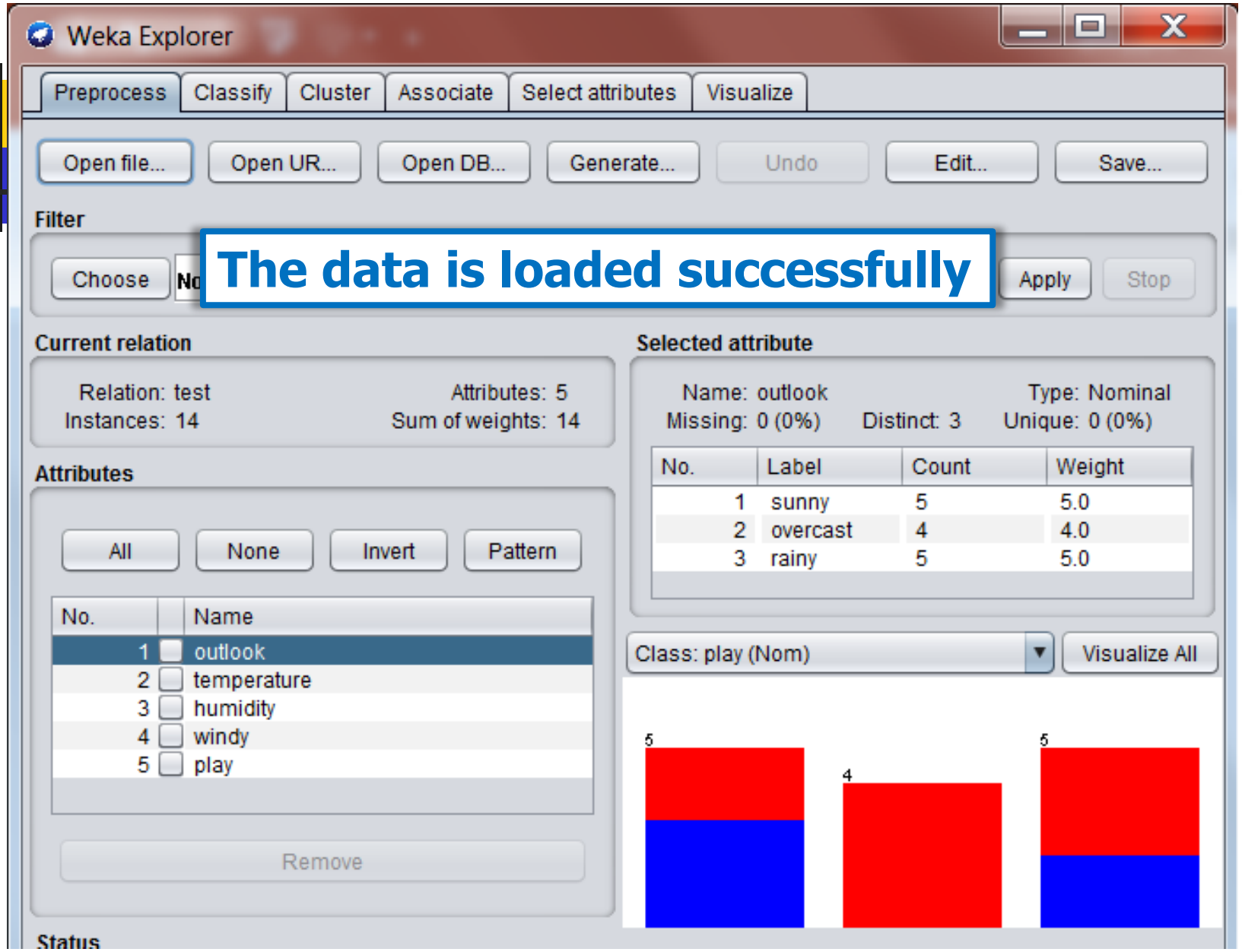
- IF the test.csv is downloaded in `C:\Users\dyhong\Downloads`



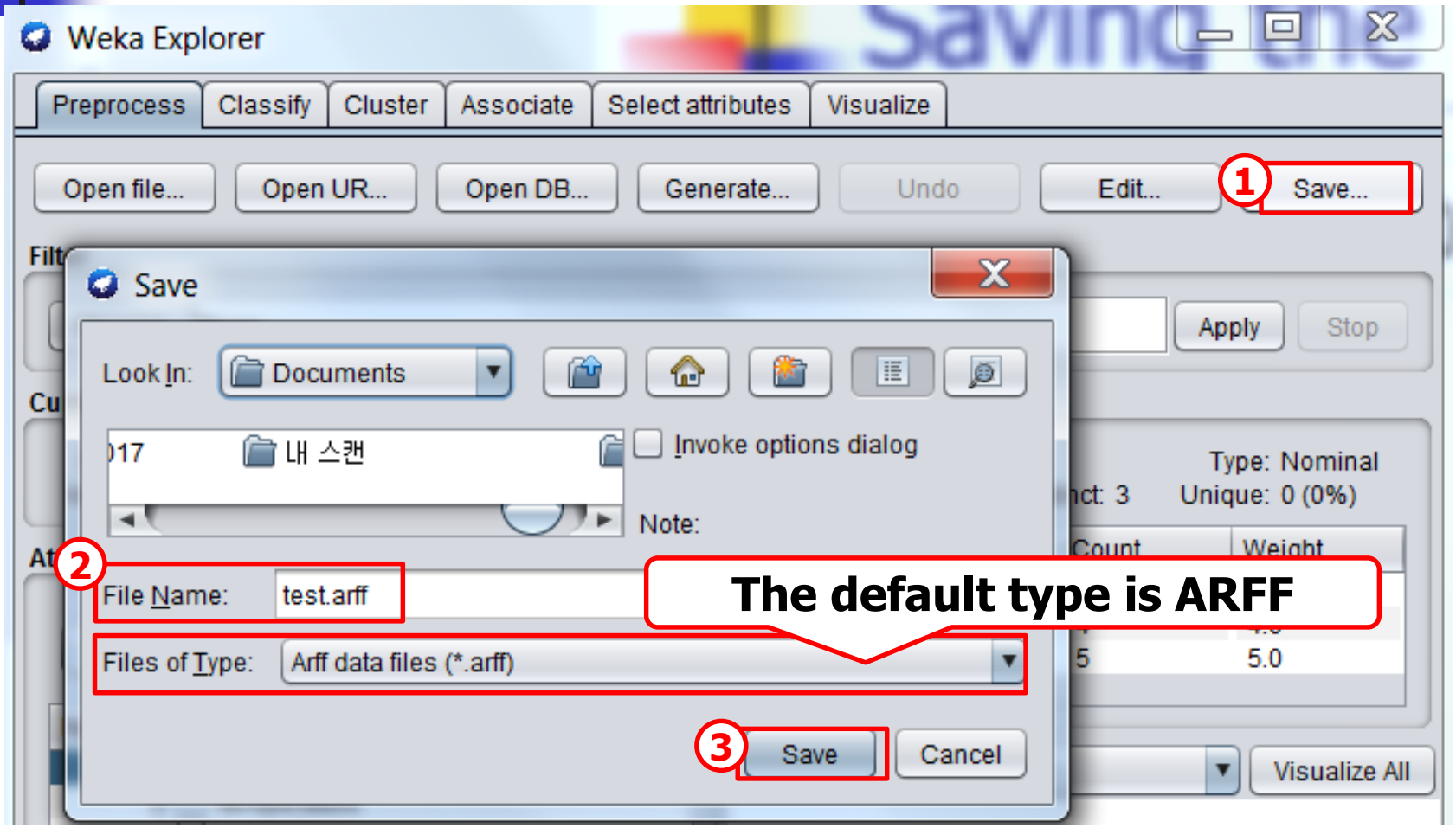
Using CSV Files for Weka

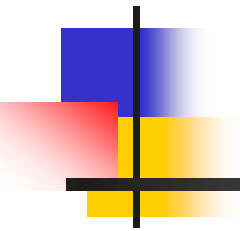
- IF the test.csv is downloaded in `C:\Users\dyhong\Downloads`




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open UR...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
The data is loaded successfully
Apply
Stop
Current relation
Relation: test
Instances: 14
Attributes: 5
Sum of weights: 14
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No. Label Count Weight
1 sunny 5 5.0
2 overcast 4 4.0
3 rainy 5 5.0
Attributes
All
None
Invert
Pattern
No. Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Class: play (Nom)
Visualize All
5
4
5
Status

Saving the CSV File as ARFF File





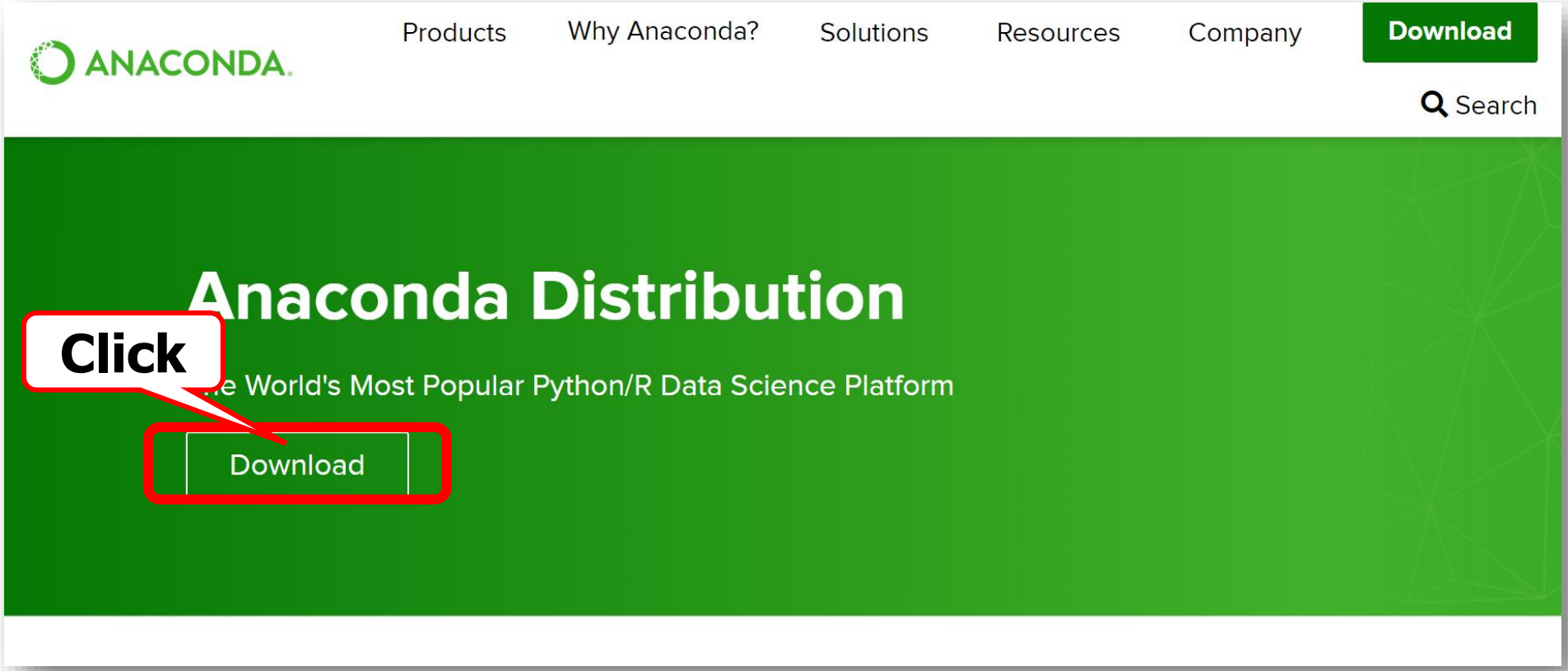
Basics of Python

Anaconda

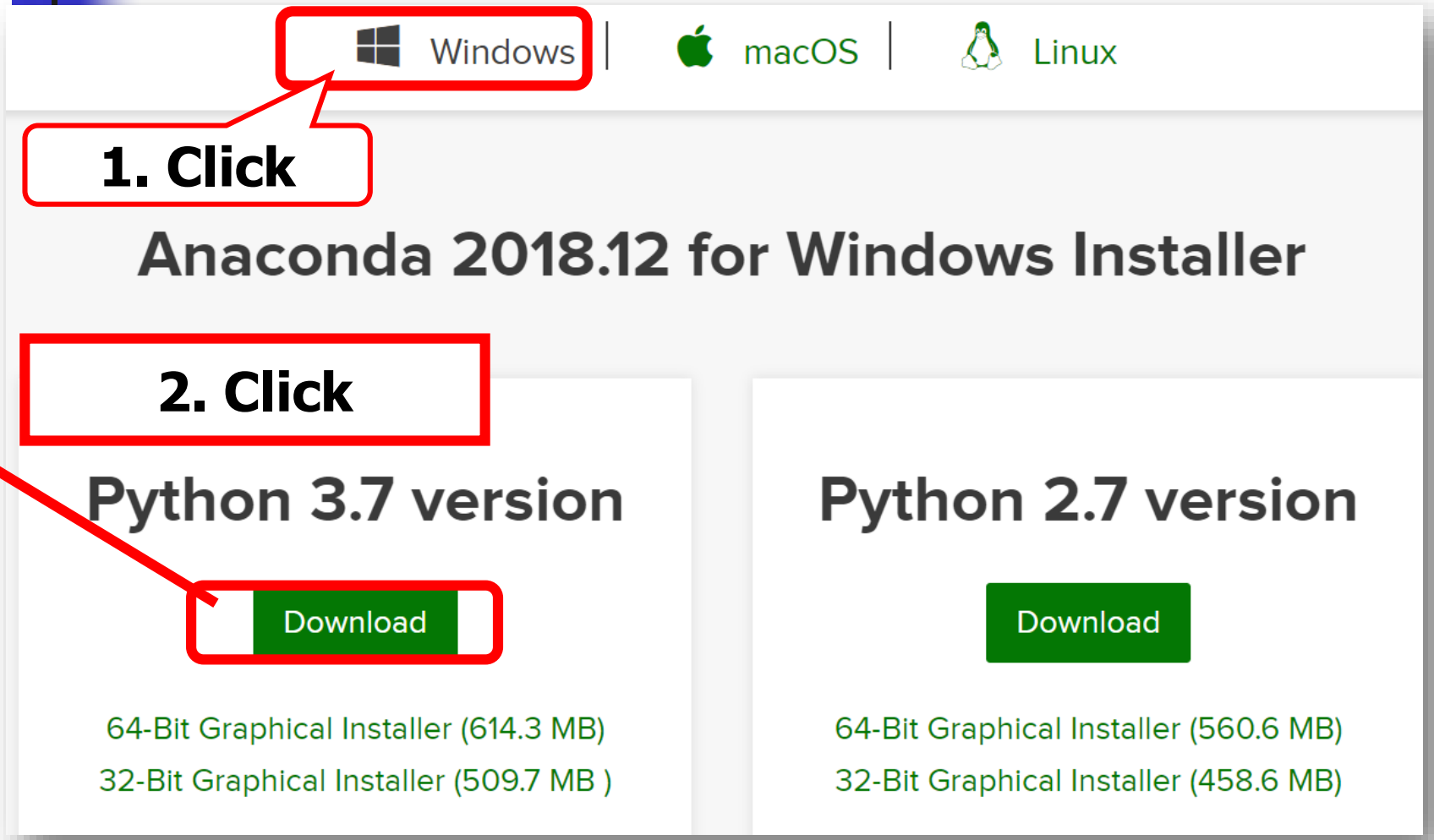
- A free and open-source distribution of the Python
  - Includes useful scientific packages
  - Simplifies package management and deployment

Anaconda Installation

- https://www.anaconda.com/distribution/



Anaconda Installation

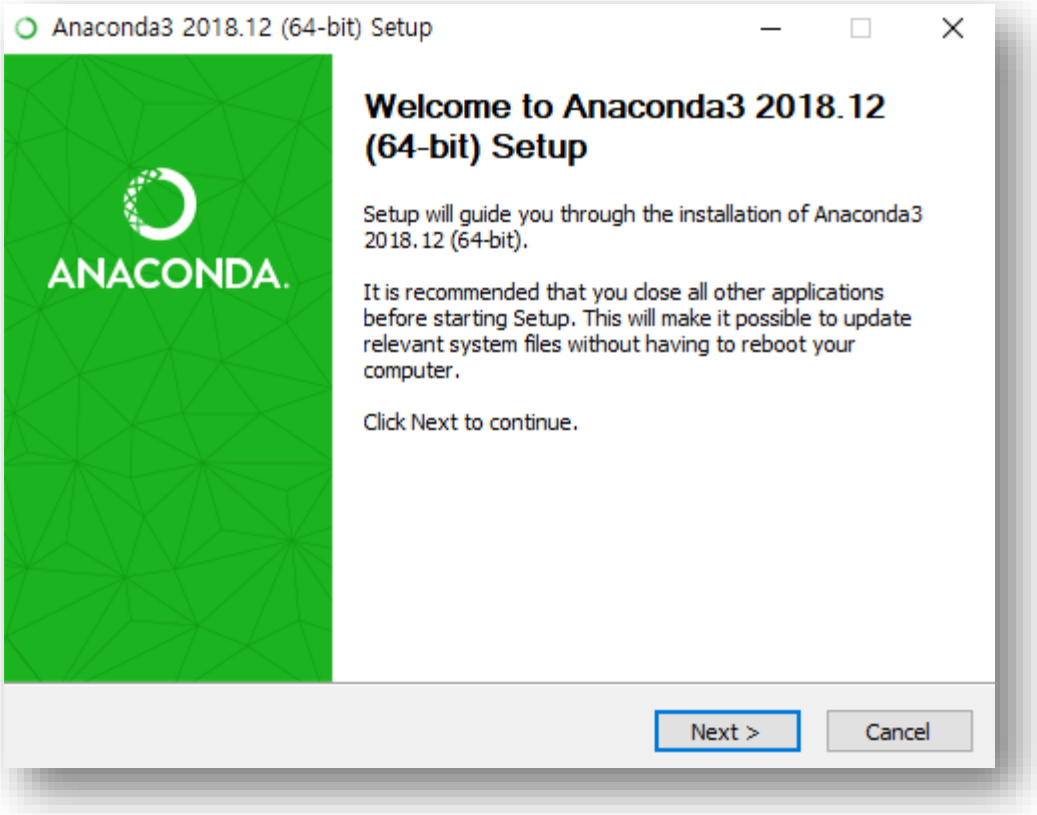


Anaconda Installation

- When https://www.anaconda.com/distribution/ is not available, please download the file from http://kdd.snu.ac.kr/python/

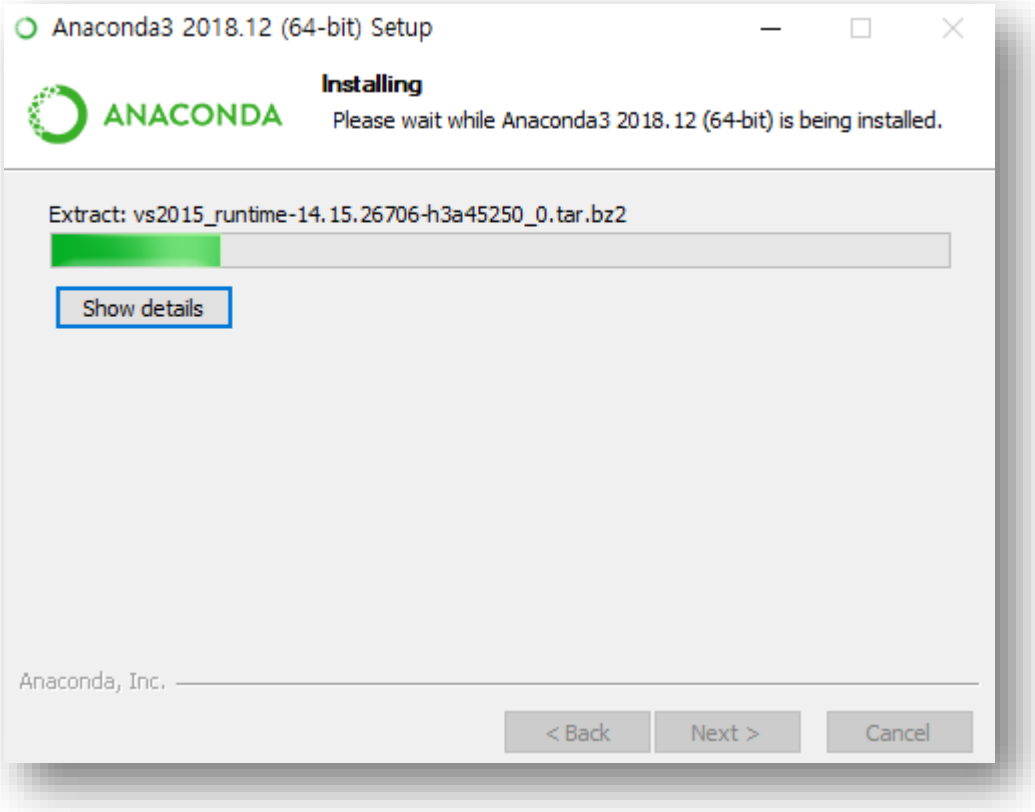
Anaconda Installation

- 1. Run the downloaded file
- 2. Select 'Next' and 'I Agree' and 'Install' for all



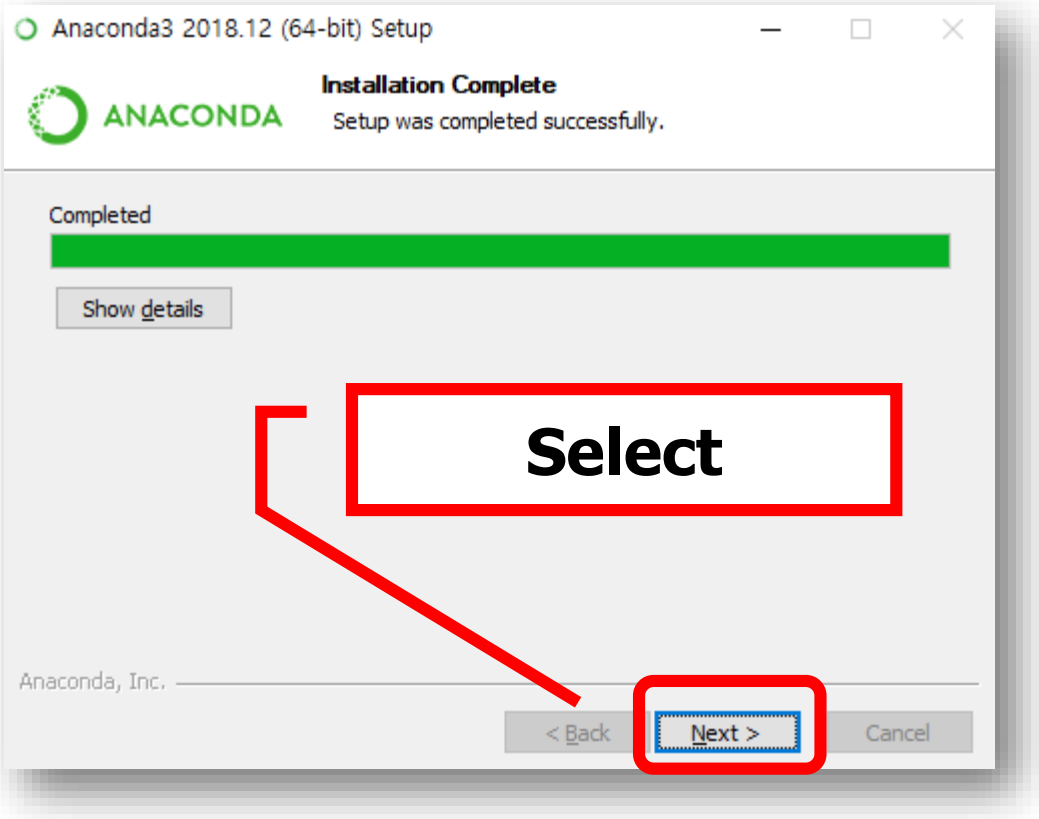
Anaconda Installation

- Installation is in progress

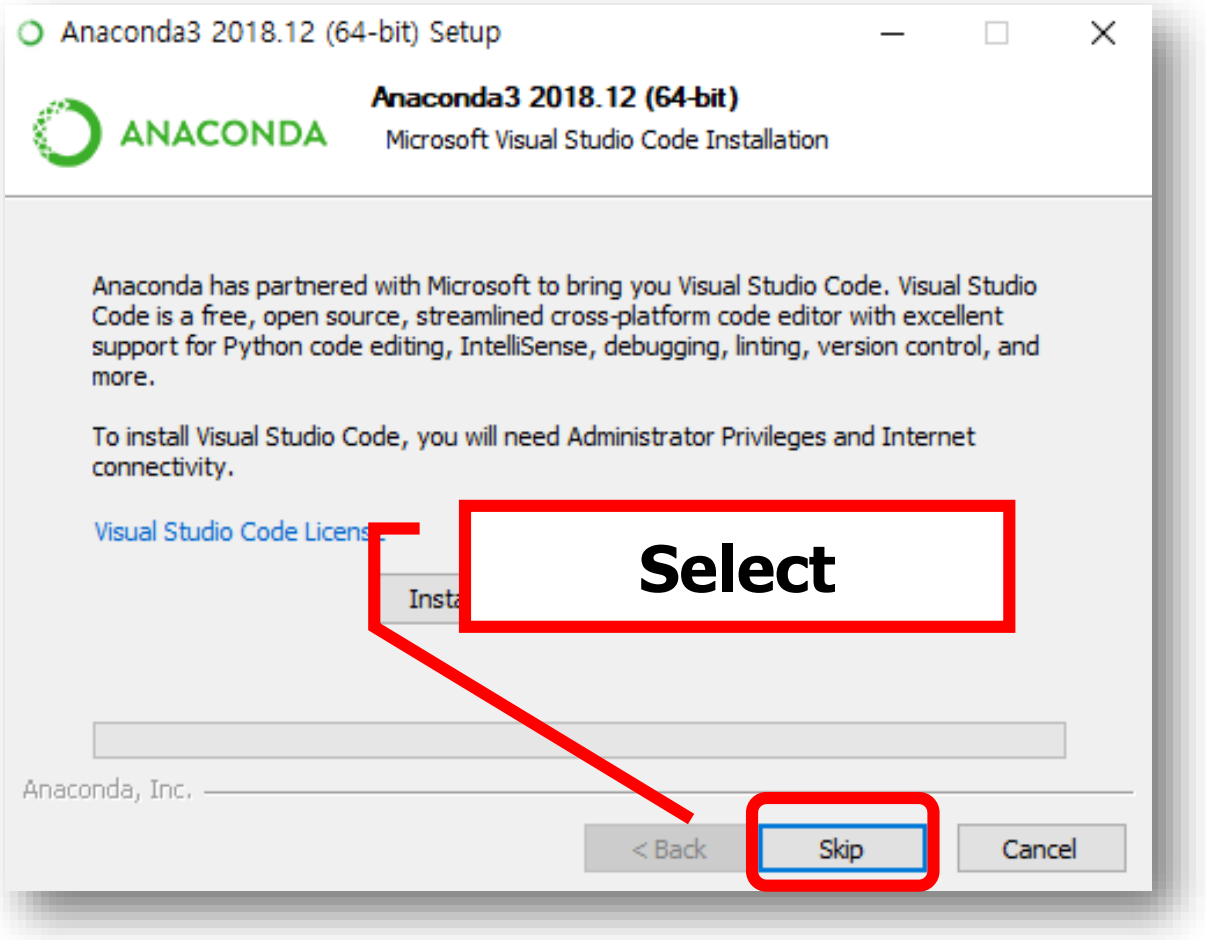


Anaconda Installation

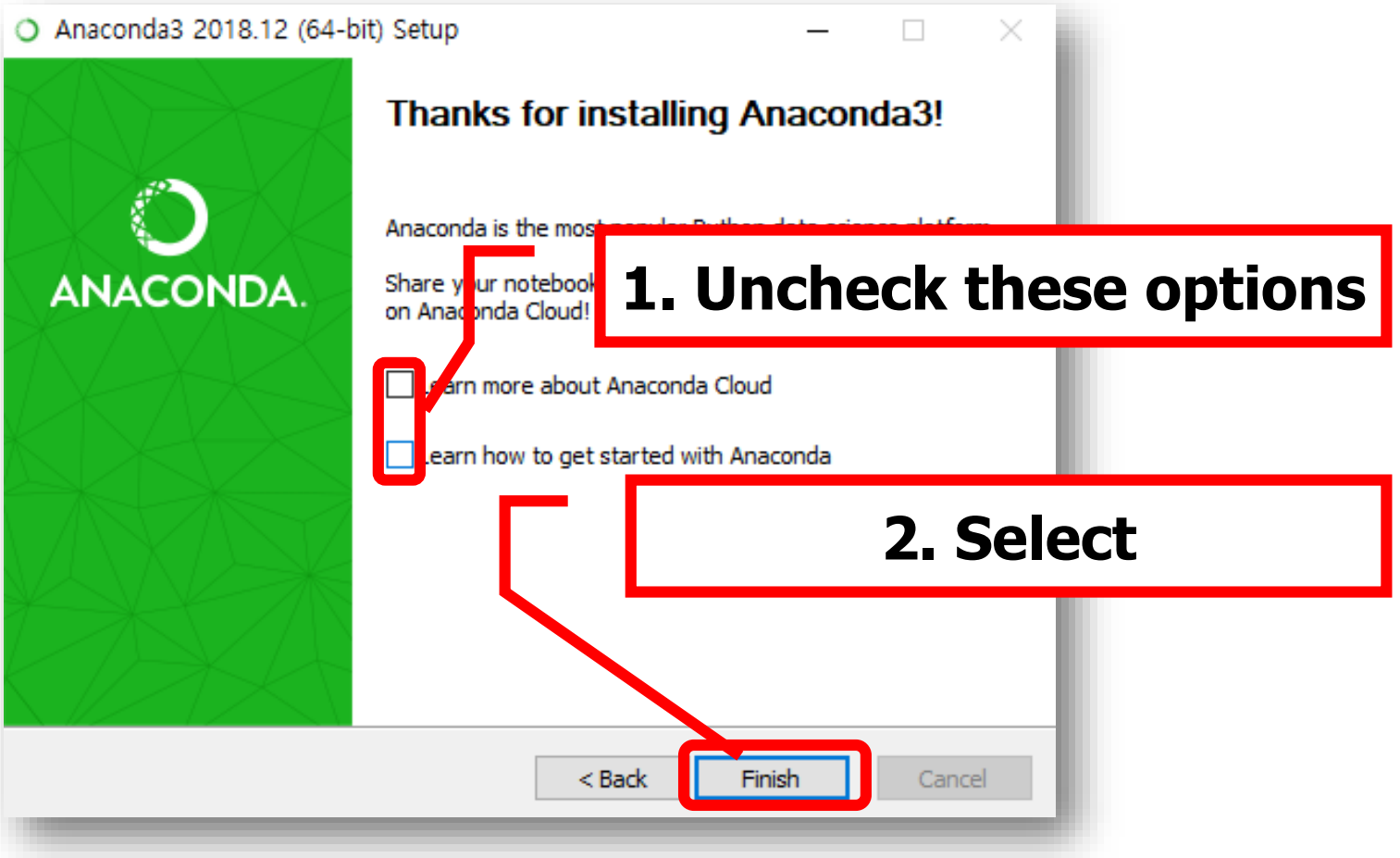
- Installation is complete



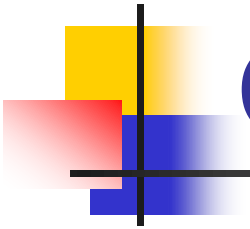
Anaconda Installation



Anaconda Installation

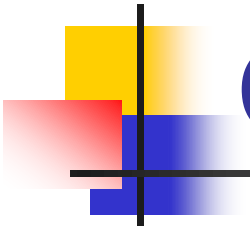


GRAPHVIZ INSTALLATION



Graphviz

- An open source graph visualization tool
- Used to visualize decision trees in this course

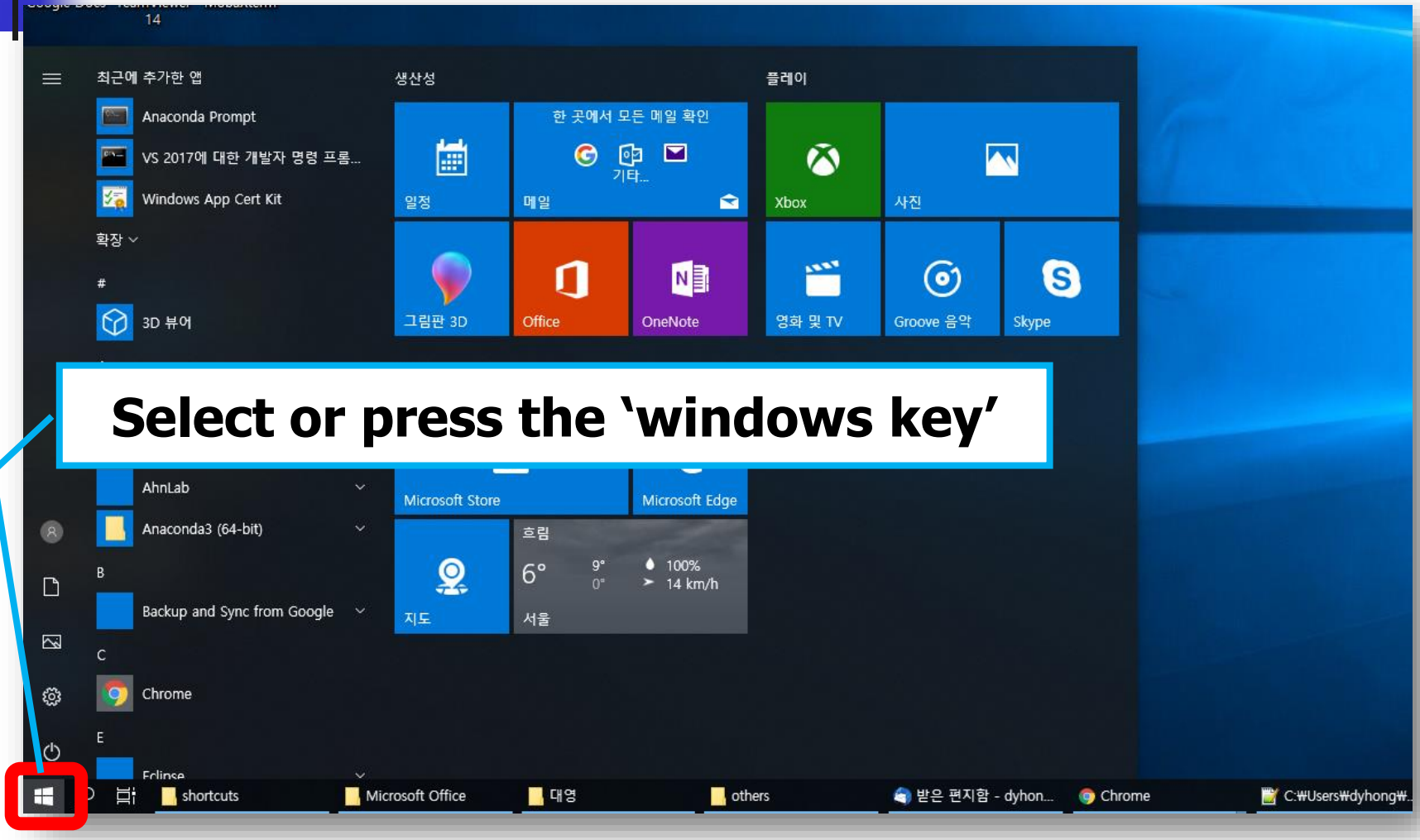


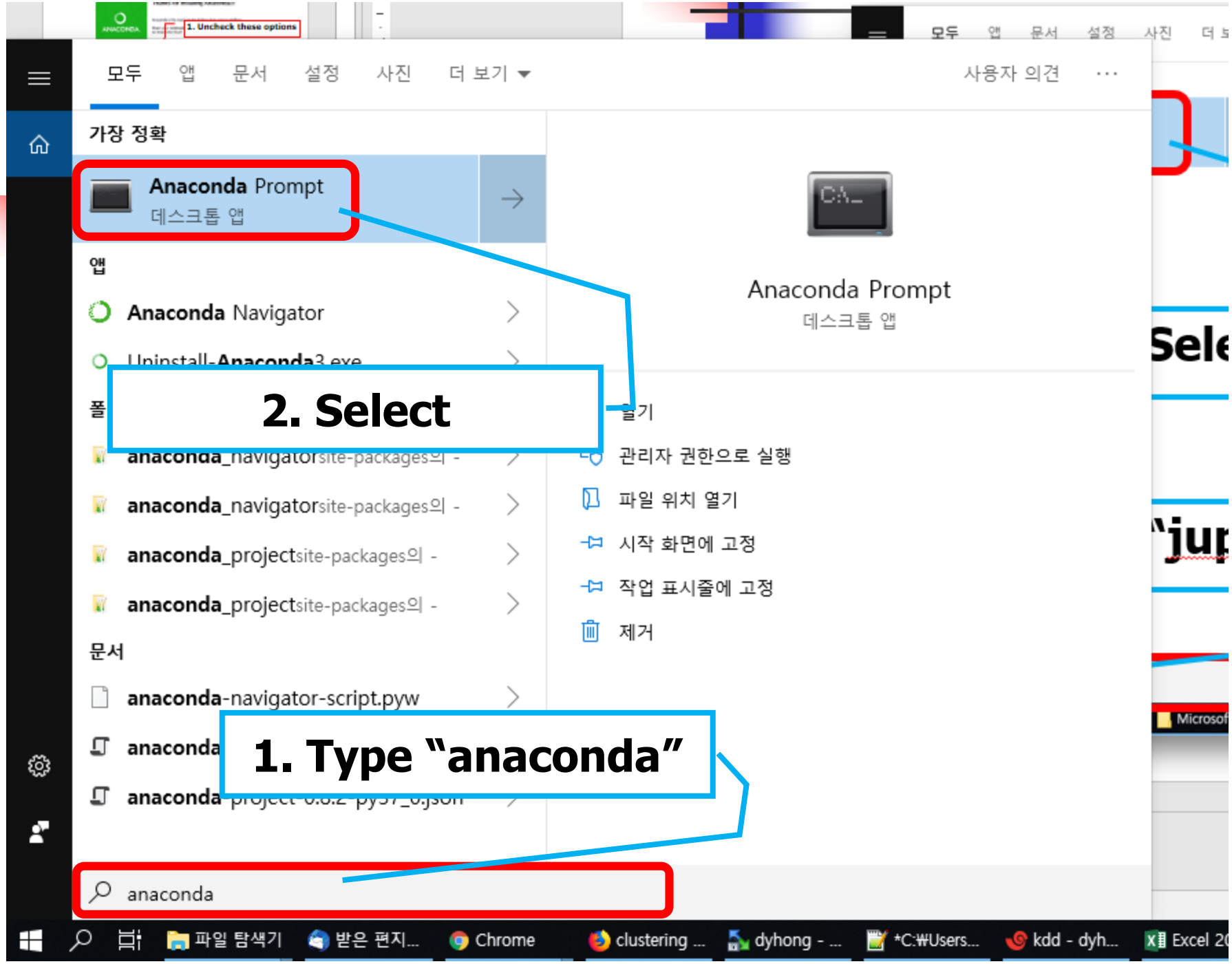
Graphviz Installation Methods

- Installation with Conda
- Manual Installation with Conda
- Installation with pip

GRAPHVIZ INSTALLATION WITH CONDA

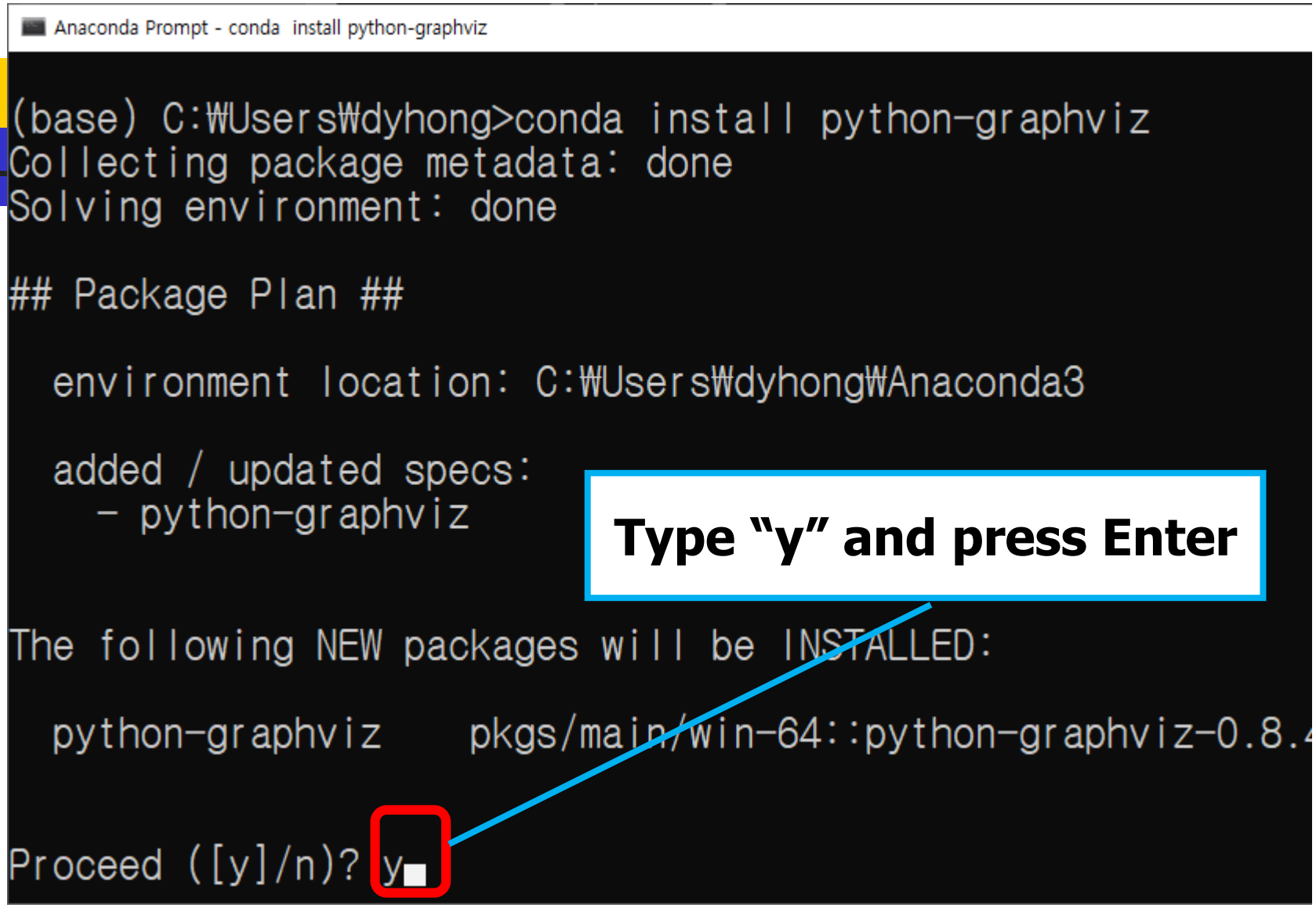
Graphviz Installation with Conda




1. Uncheck these options
모두
앱
문서
설정
사진
더 보기
사용자 의견
가장 정확
Anaconda Prompt
데스크톱 앱
앱
Anaconda Navigator
Uninstall-Anaconda3.exe
anaconda_navigatorsite-packages의 -
anaconda_navigatorsite-packages의 -
anaconda_projectsite-packages의 -
anaconda_projectsite-packages의 -
문서
anaconda-navigator-script.pyw
anaconda
anaconda
Anaconda Prompt
데스크톱 앱
열기
관리자 권한으로 실행
파일 위치 열기
시작 화면에 고정
작업 표시줄에 고정
제거
2. Select
1. Type "anaconda"
anaconda
파일 탐색기
받은 편지...
Chrome
clustering ...
dyhong - ...
*C:₩Users...
kdd - dyh...
Excel 20

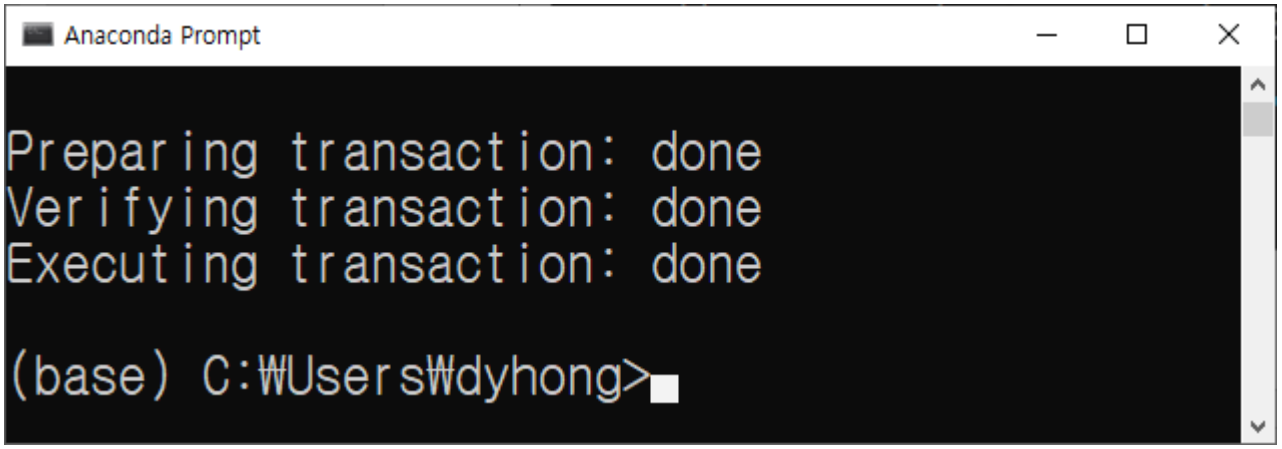
Graphviz Installation with Conda

Type this command and press Enter


Anaconda Prompt - conda install python-graphviz
(base) C:₩Users₩dyhong>conda install python-graphviz
Collecting package metadata: done
Solving environment: done
Package Plan
environment location: C:₩Users₩dyhong₩Anaconda3
added / updated specs:
- python-graphviz
The following NEW packages will be INSTALLED:
python-graphviz pkgs/main/win-64::python-graphviz-0.8.
Proceed ([y]/n)? y
Type "y" and press Enter

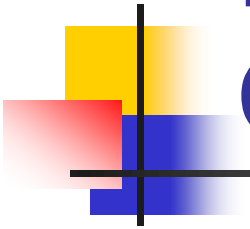
Graphviz Installation with Conda

- The installation is complete



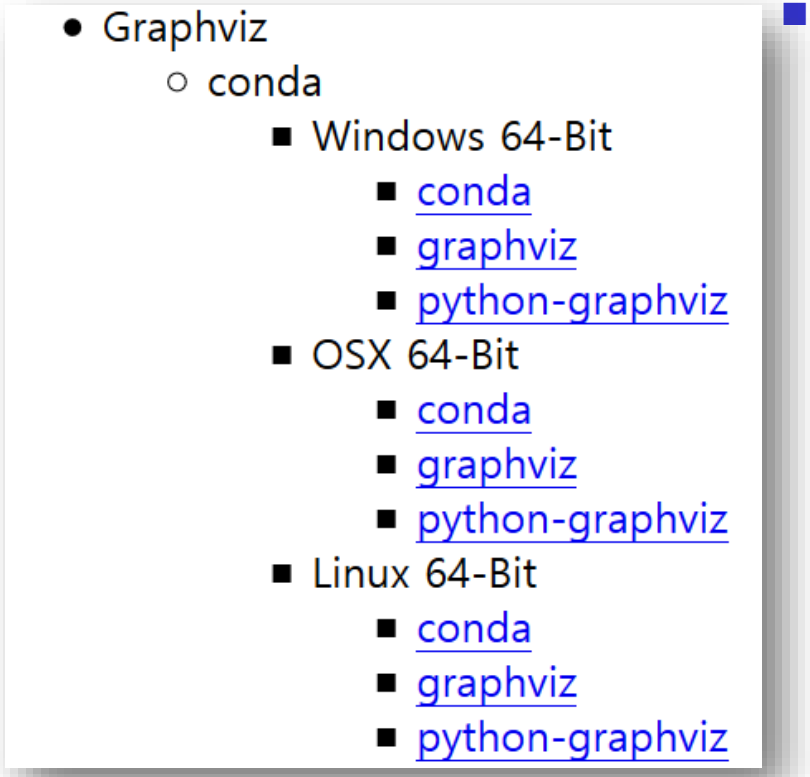
MANUAL GRAPHVIZ INSTALLATION WITH CONDA

Manual Graphviz Installation with Conda



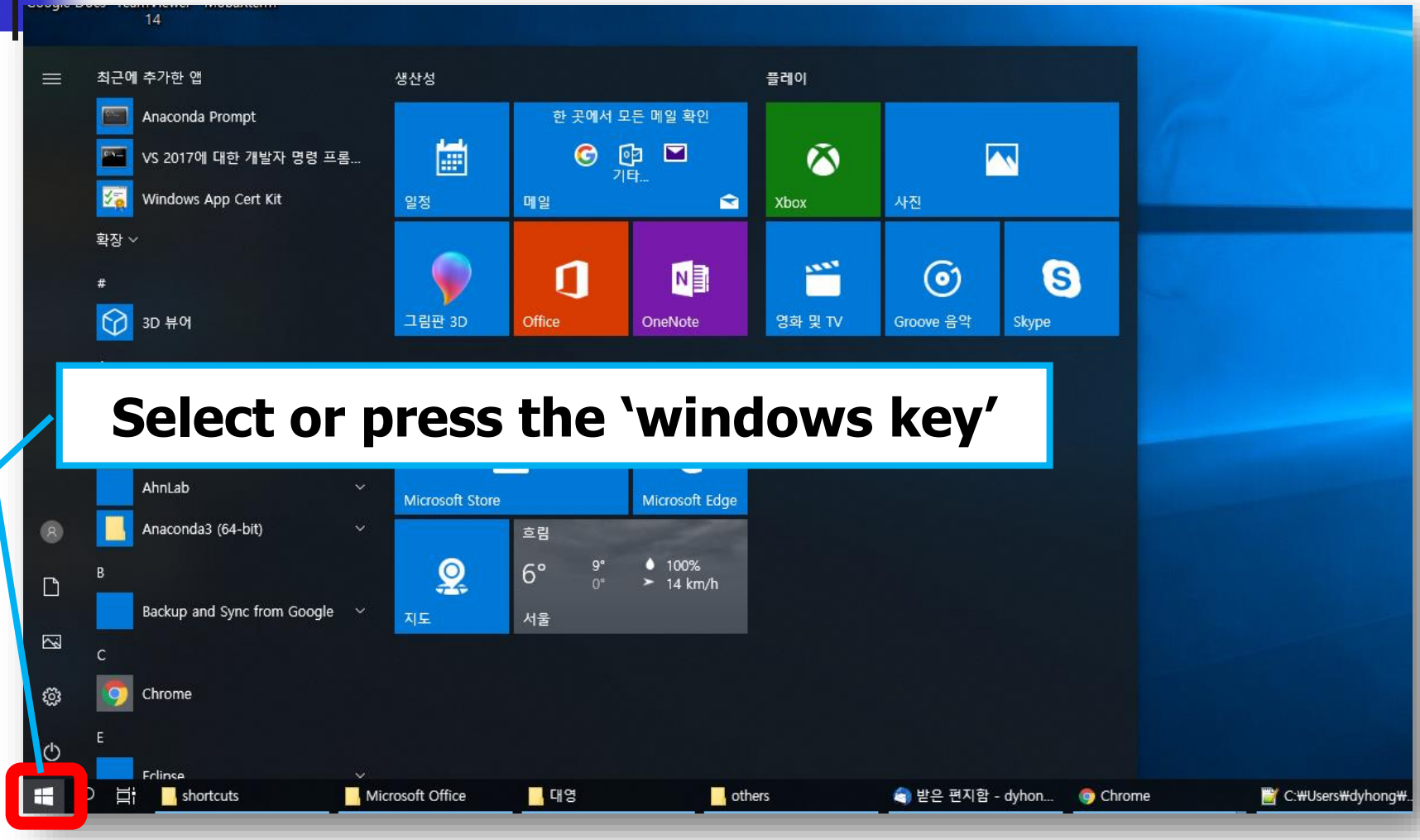
- An alternative installation method when the previous method does not work

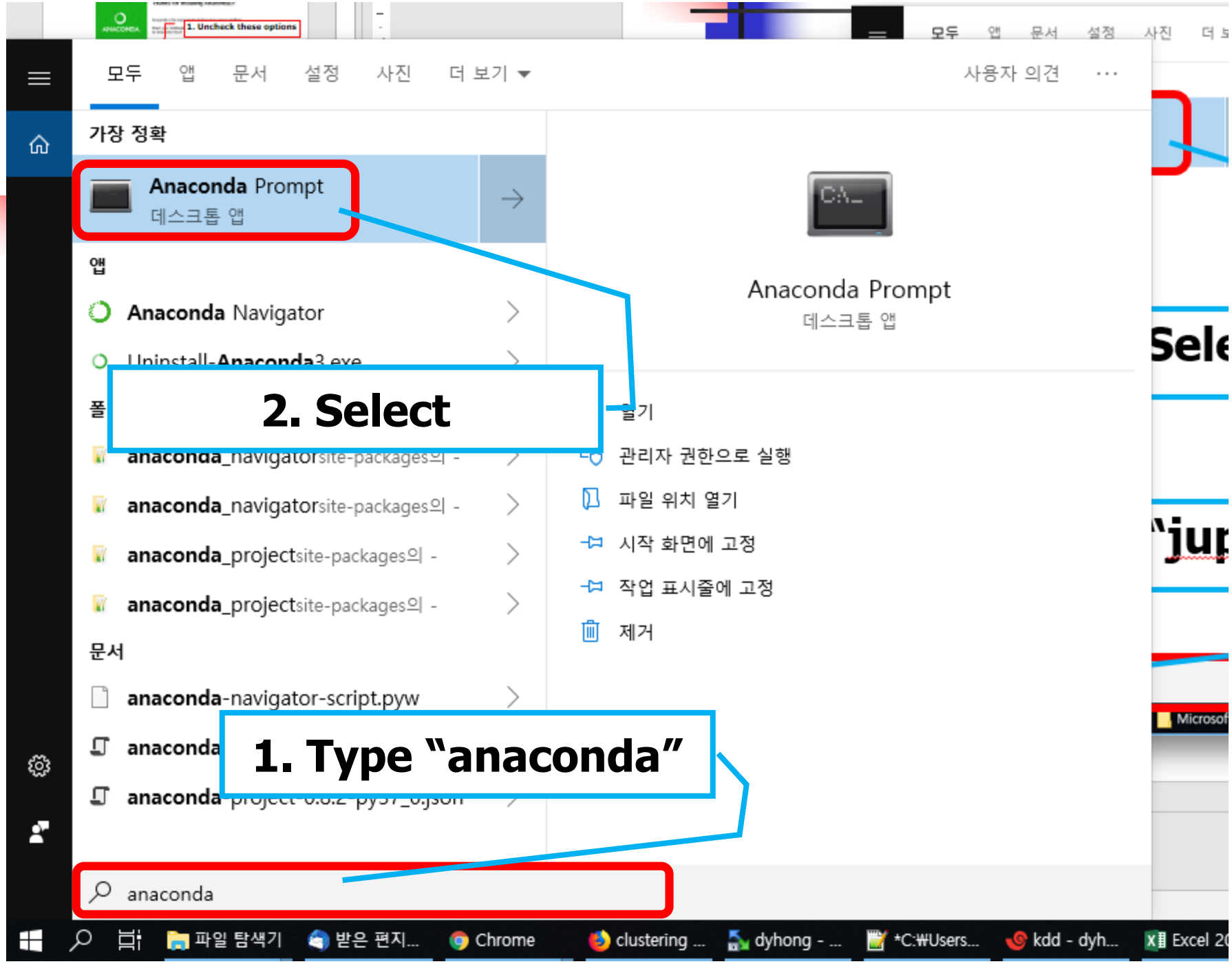
Manual Graphviz Installation with Conda



- Download the following files for your OS from http://kdd.snu.ac.kr/python/
  - conda
  - graphviz
  - python-graphviz

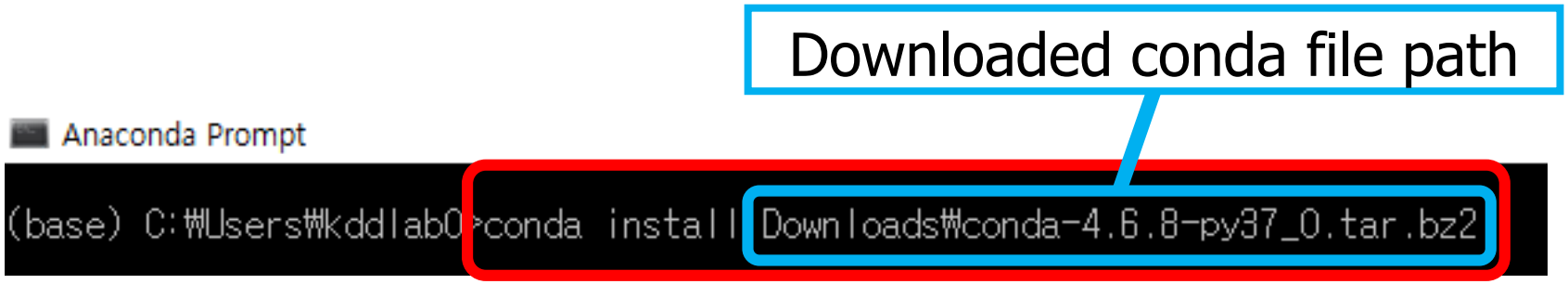
Manual Graphviz Installation with Conda




1. Uncheck these options
모두
앱
문서
설정
사진
더 보기
사용자 의견
가장 정확
Anaconda Prompt
데스크톱 앱
앱
Anaconda Navigator
Uninstall-Anaconda3.exe
anaconda_navigatorsite-packages의 -
anaconda_navigatorsite-packages의 -
anaconda_projectsite-packages의 -
anaconda_projectsite-packages의 -
문서
anaconda-navigator-script.pyw
anaconda
anaconda
Anaconda Prompt
데스크톱 앱
열기
관리자 권한으로 실행
파일 위치 열기
시작 화면에 고정
작업 표시줄에 고정
제거
2. Select
1. Type "anaconda"
anaconda
파일 탐색기
받은 편지...
Chrome
clustering ...
dyhong - ...
*C:₩Users...
kdd - dyh...
Excel 20

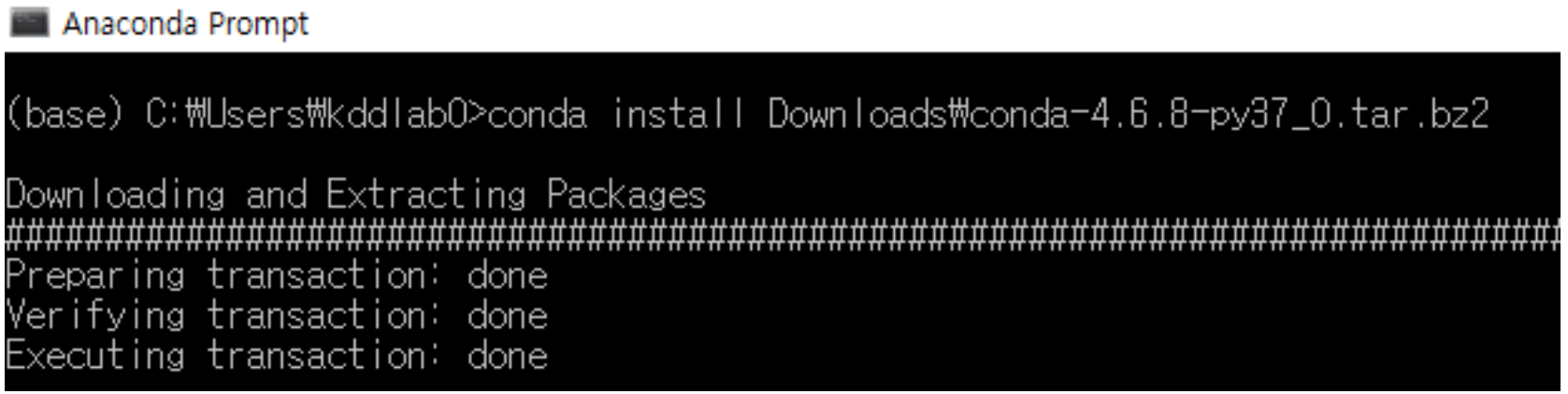
Manual Graphviz Installation with Conda

- Type “conda install [downloaded conda file path]”



Manual Graphviz Installation with Conda

- The *conda* is installed

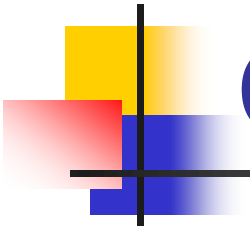


Manual Graphviz Installation with Conda

- Repeat for ‘graphviz’ and ‘python-graphviz’

```
(base) C:₩Users₩kddlab0>conda install Downloads₩graphviz-2.38-hfd603c8_2.tar.bz2

Downloading and Extracting Packages
################################################################################################################################
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(base) C:₩Users₩kddlab0>conda install Downloads₩python-graphviz-0.8.4-py37_1002.tar.bz2

Downloading and Extracting Packages
################################################################################################################################
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

GRAPHVIZ INSTALLATION WITH PIP

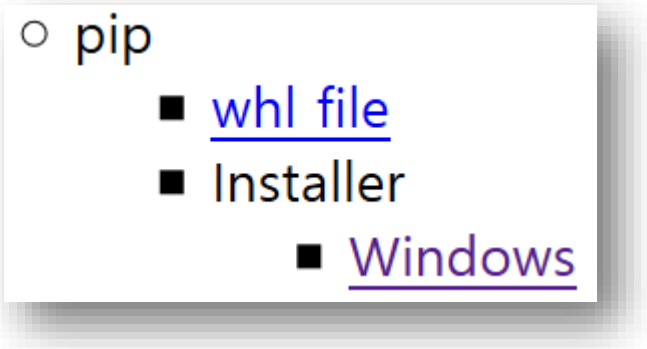


Graphviz Installation with pip

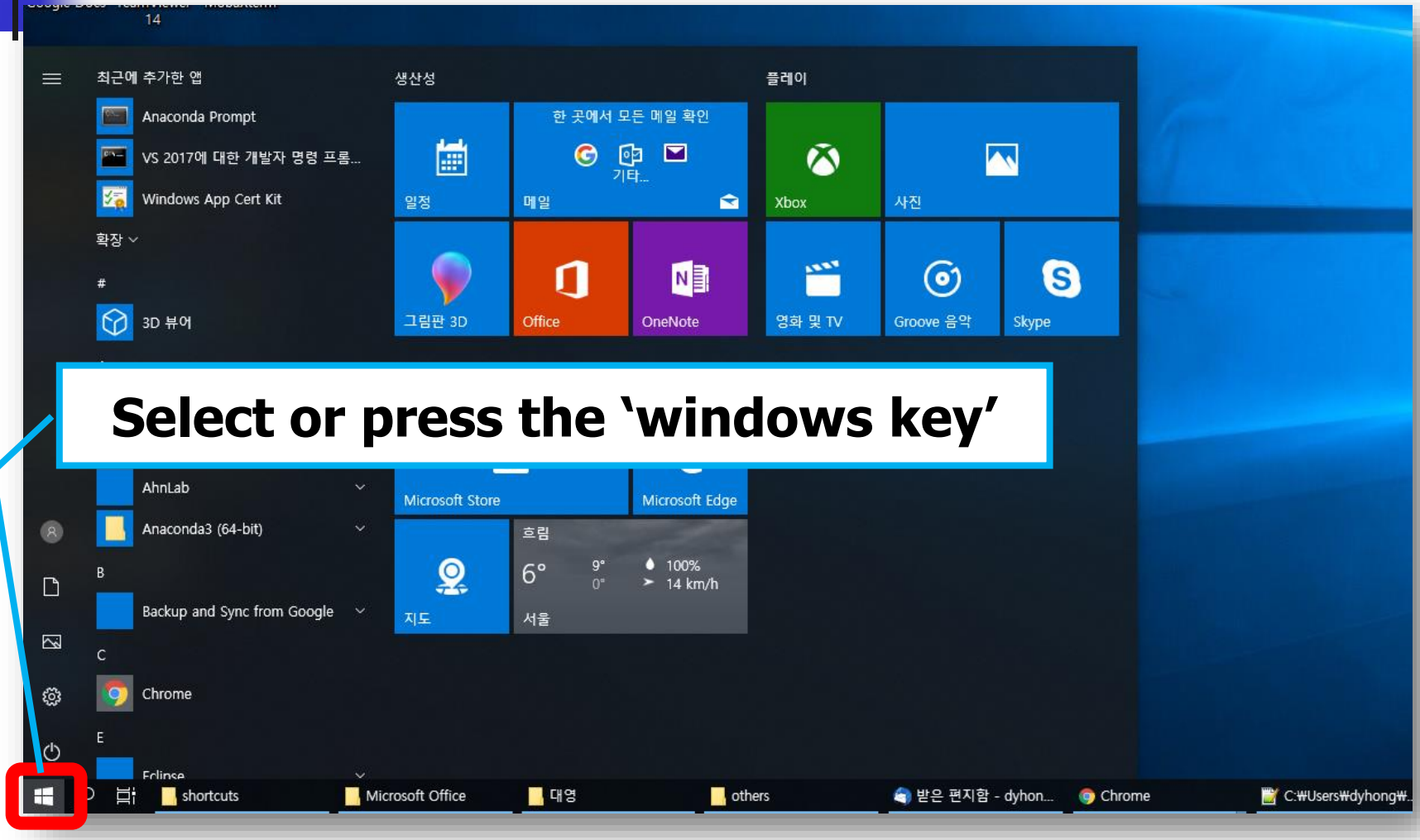
- An alternative installation when Conda is not available

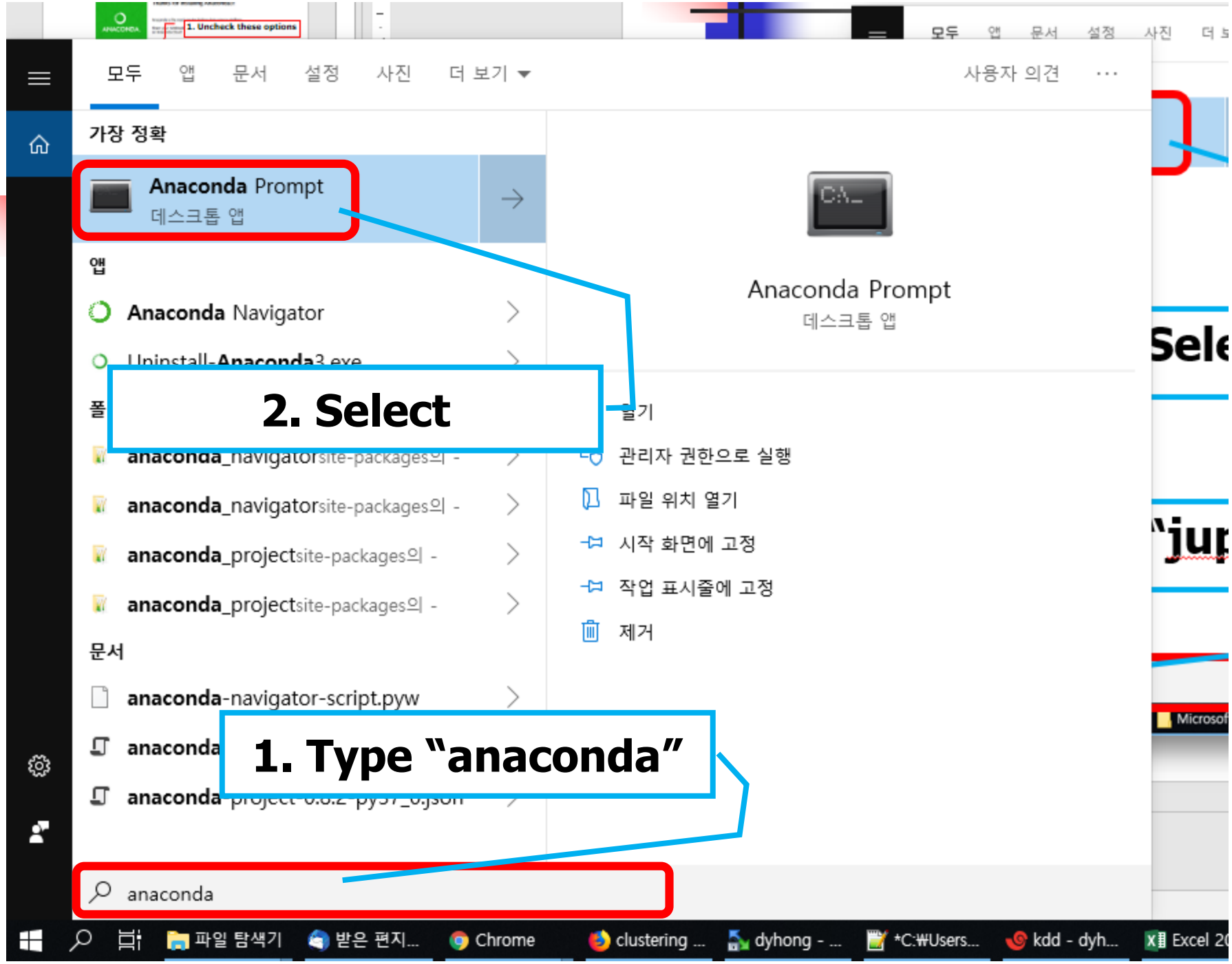
Graphviz Installation with pip

- Download from http://kdd.snu.ac.kr/python/
  - whl file
  - Windows installer



Graphviz Installation with pip

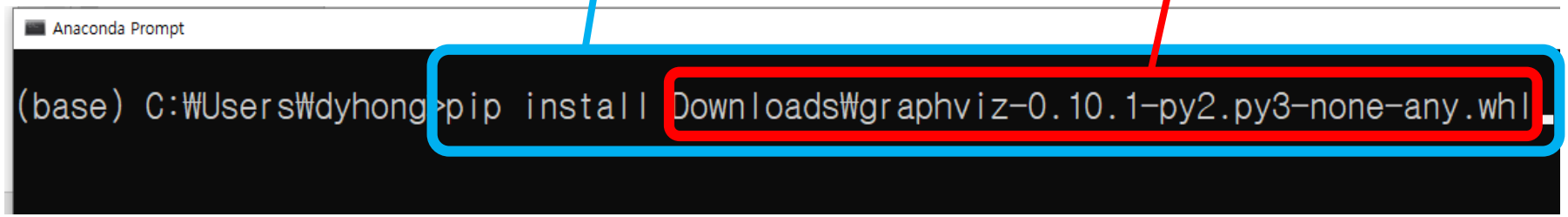



1. Uncheck these options
모두
앱
문서
설정
사진
더 보기
사용자 의견
가장 정확
Anaconda Prompt
데스크톱 앱
앱
Anaconda Navigator
Uninstall-Anaconda3.exe
anaconda_navigatorsite-packages의 -
anaconda_navigatorsite-packages의 -
anaconda_projectsite-packages의 -
anaconda_projectsite-packages의 -
문서
anaconda-navigator-script.pyw
anaconda
anaconda
Anaconda Prompt
데스크톱 앱
열기
관리자 권한으로 실행
파일 위치 열기
시작 화면에 고정
작업 표시줄에 고정
제거
2. Select
1. Type "anaconda"
anaconda
파일 탐색기
받은 편지...
Chrome
clustering ...
dyhong - ...
*C:₩Users...
kdd - dyh...
Excel 20

Graphviz Installation with pip

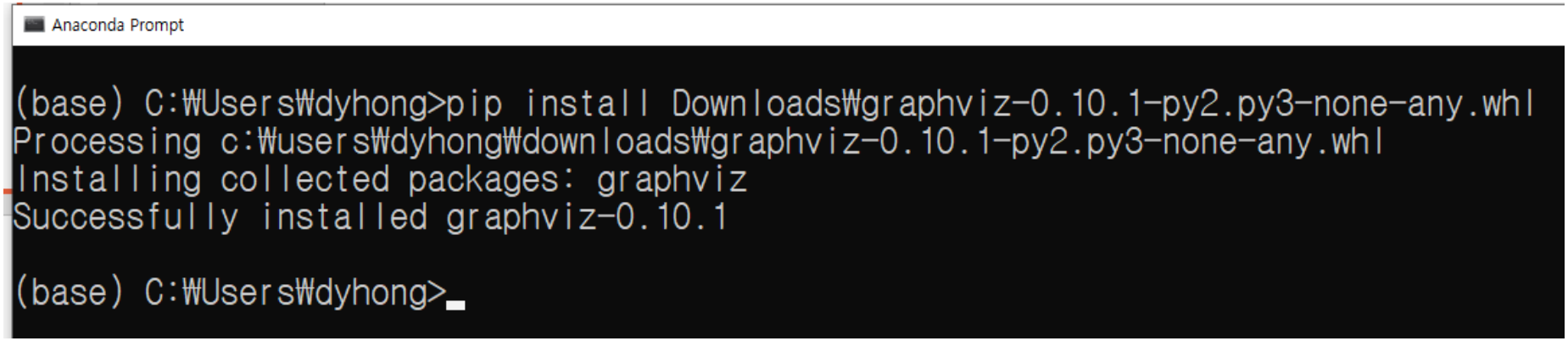
Type this and press Enter

The path of the downloaded whl file



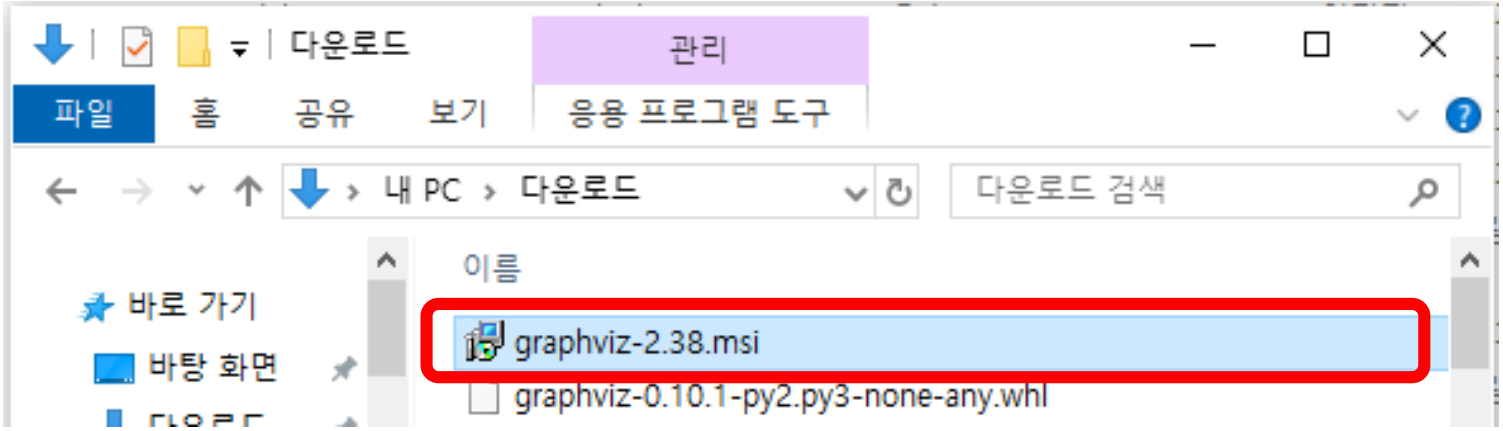
Graphviz Installation with pip

- Successfully installed the whl file



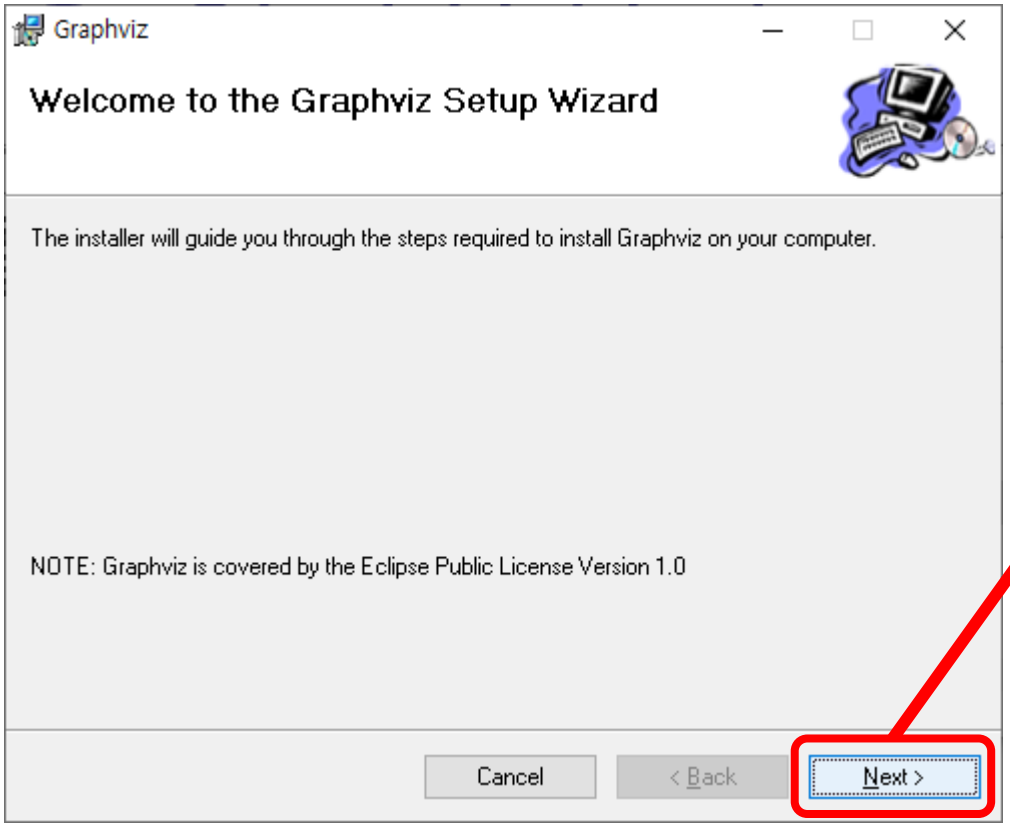
Graphviz Installation with pip

- Run the downloaded Windows installer of Graphviz



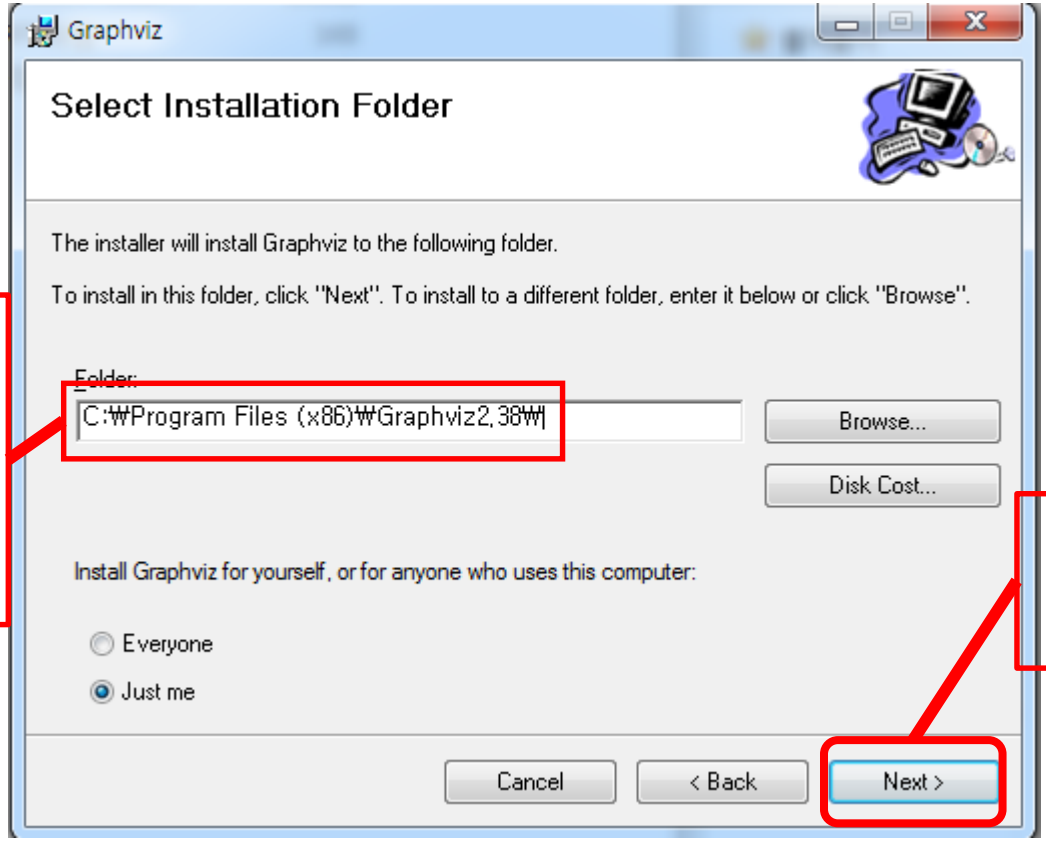
Graphviz Installation with pip

- Press Next



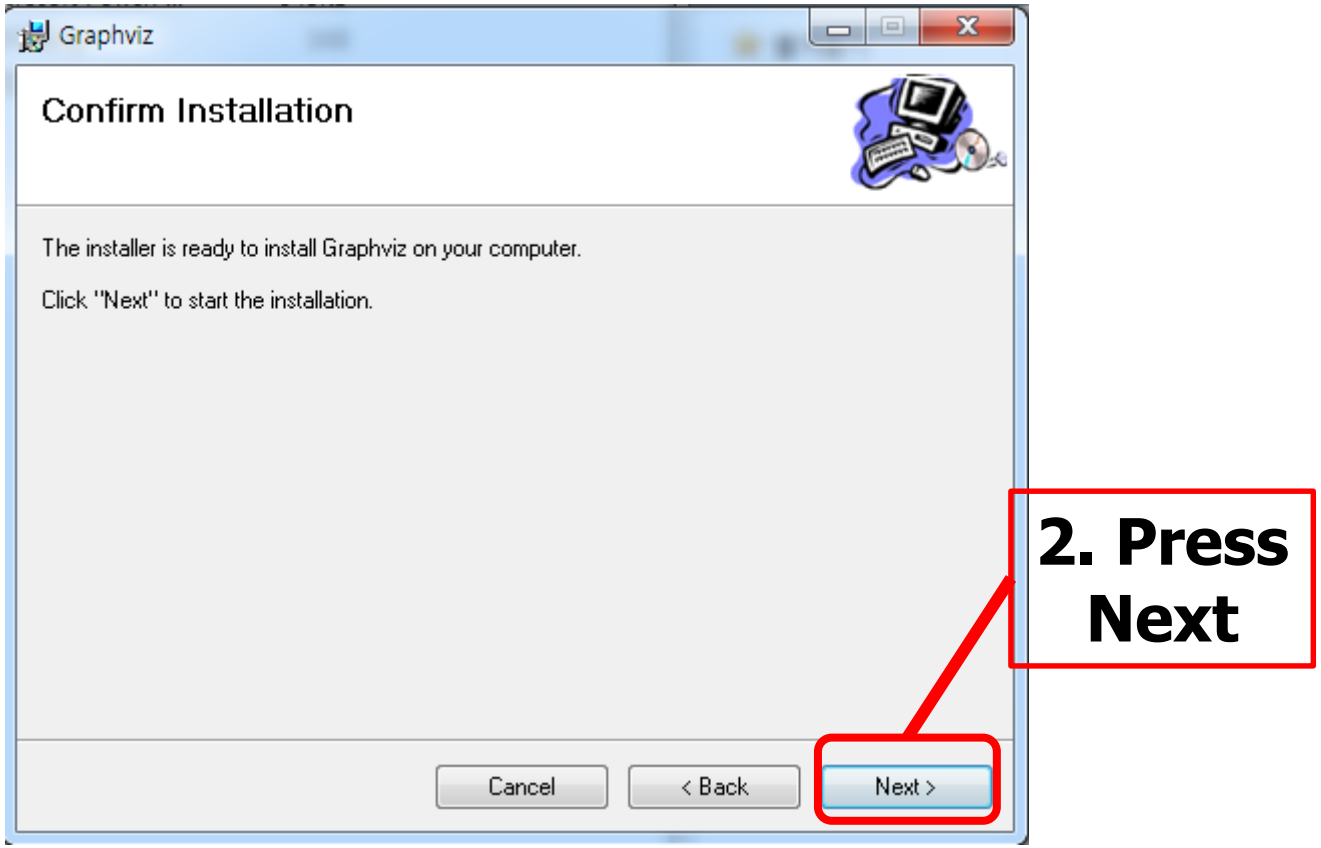
Graphviz Installation with pip

1. Copy and save the path for further usage

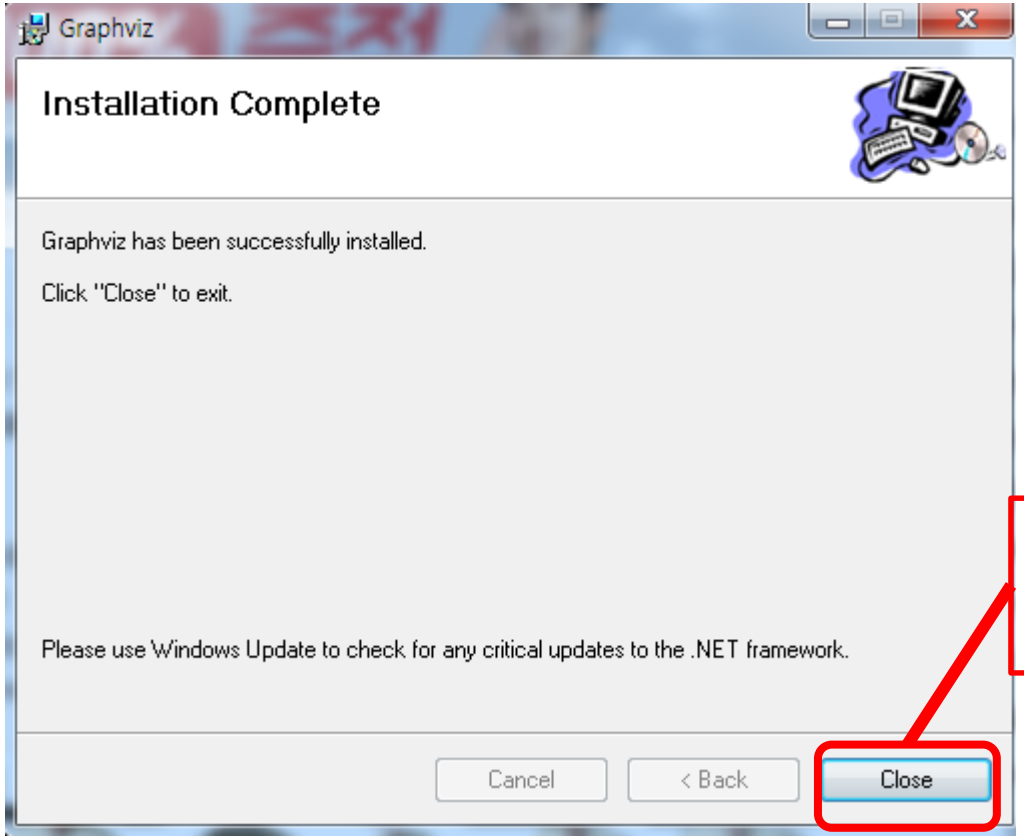


2. Press Next

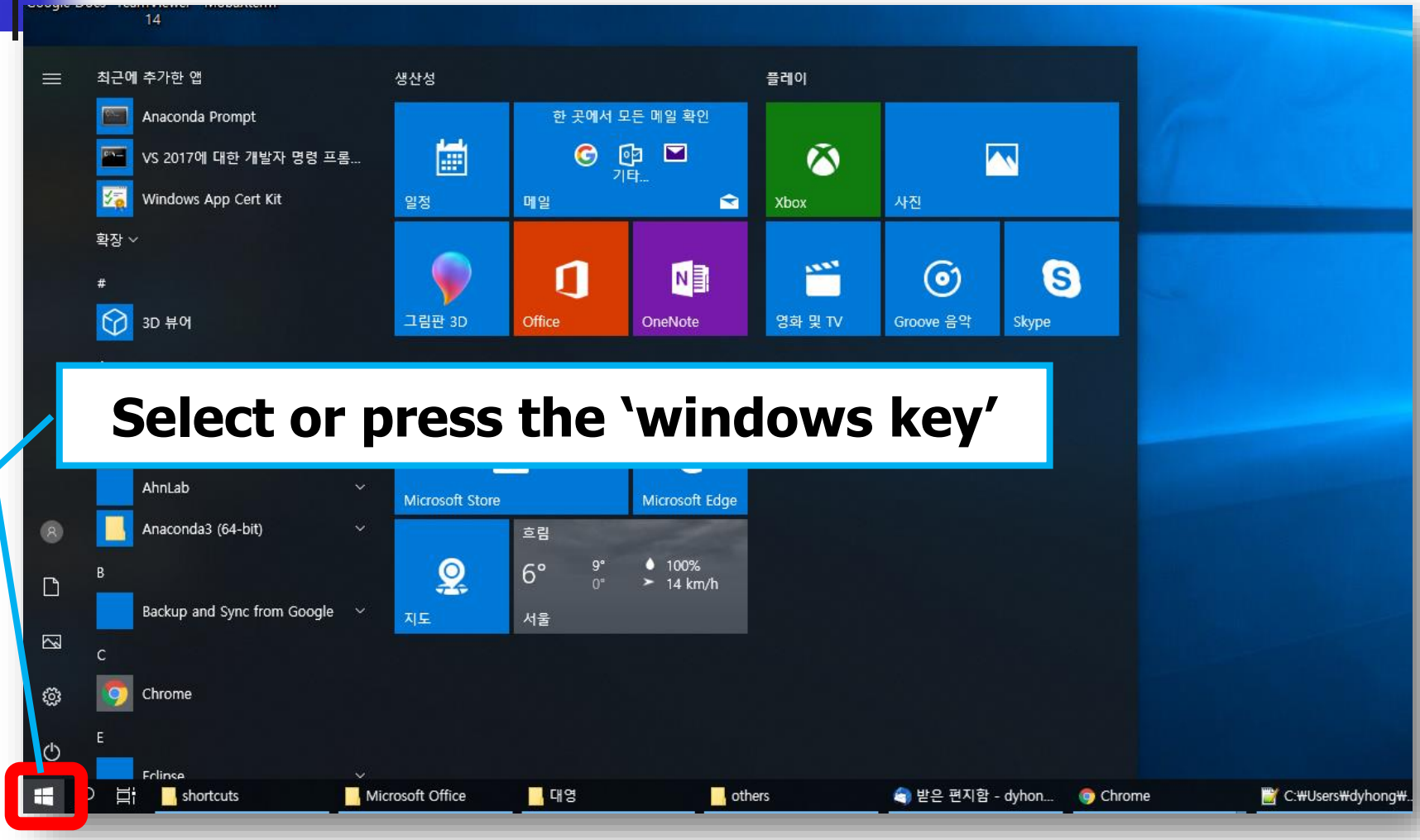
Graphviz Installation with pip

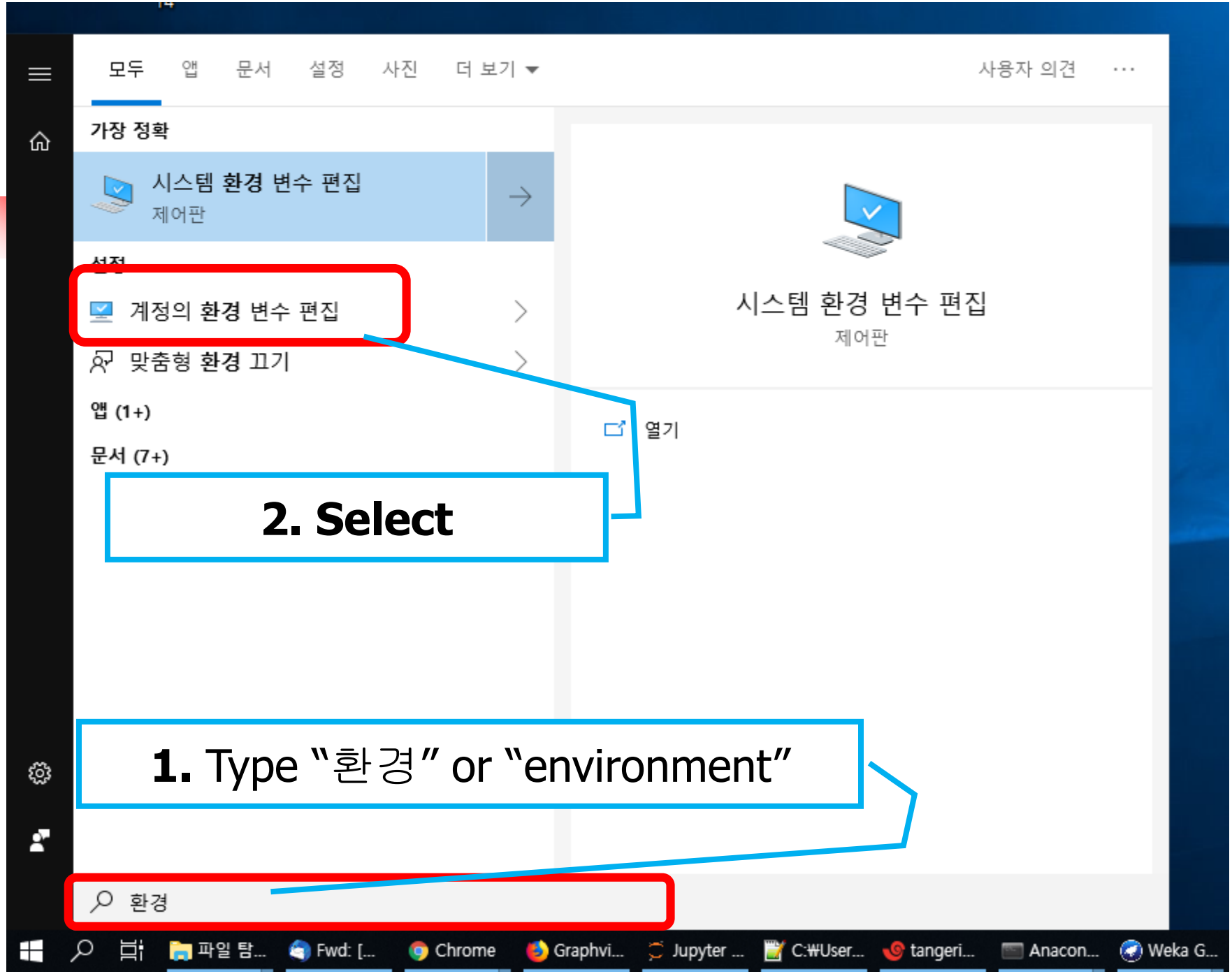


Graphviz Installation with pip

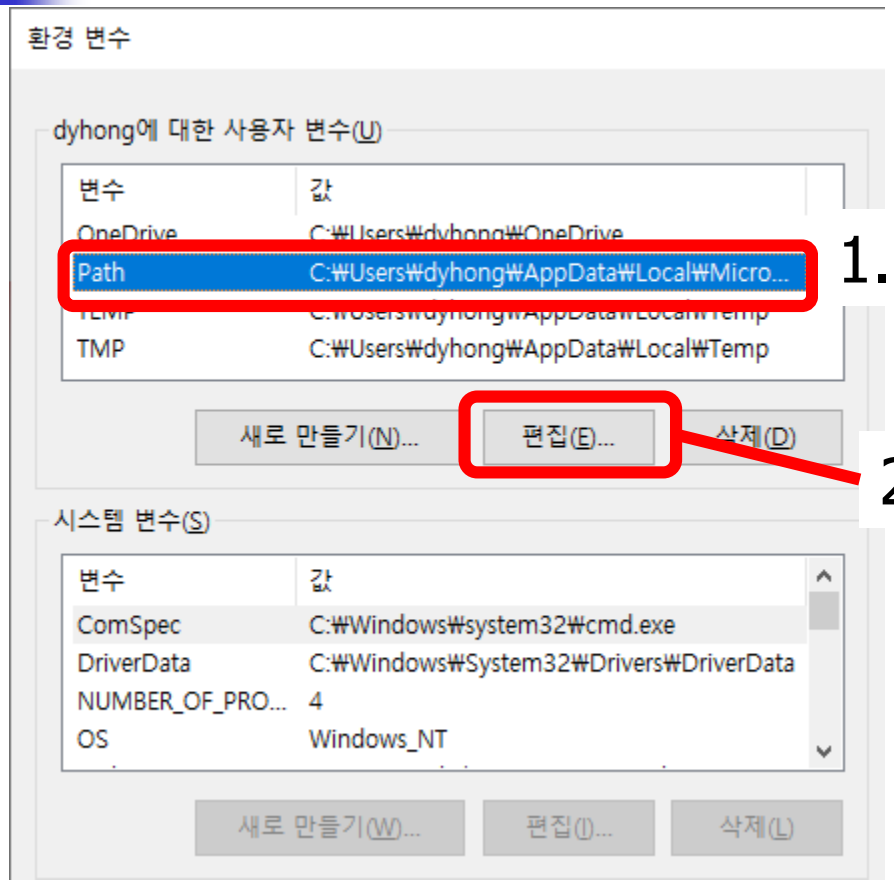


Graphviz Installation with pip




모두
앱
문서
설정
사진
더 보기
사용자 의견
가장 정확
시스템 환경 변수 편집
제어판
계정의 환경 변수 편집
맞춤형 환경 끄기
앱 (1+)
문서 (7+)
시스템 환경 변수 편집
제어판
열기
2. Select
1. Type "환경" or "environment"
환경
파일 탐...
Fwd: [...
Chrome
Graphvi...
Jupyter ...
C:₩User...
tangeri...
Anacon...
Weka G...

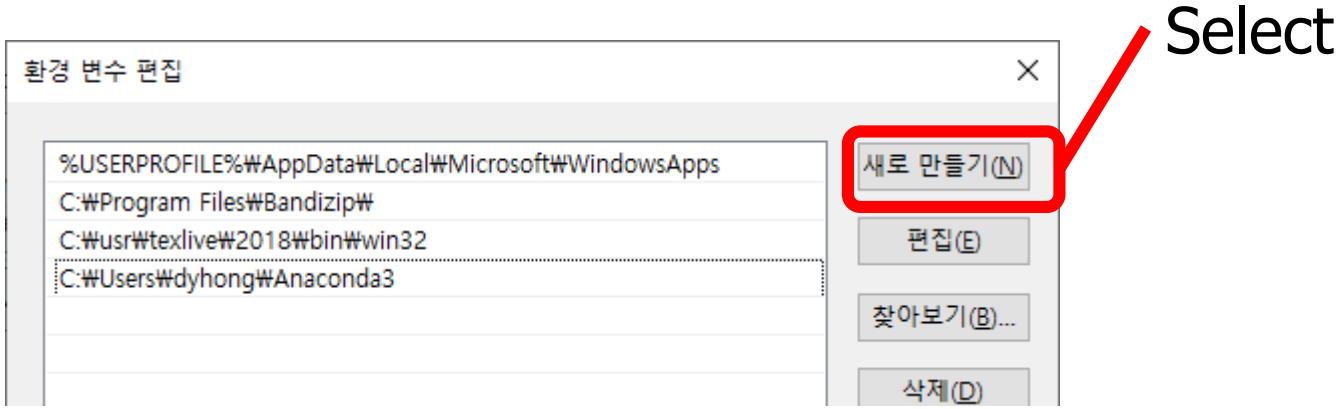
Graphviz Installation with pip



1. Select 'Path'

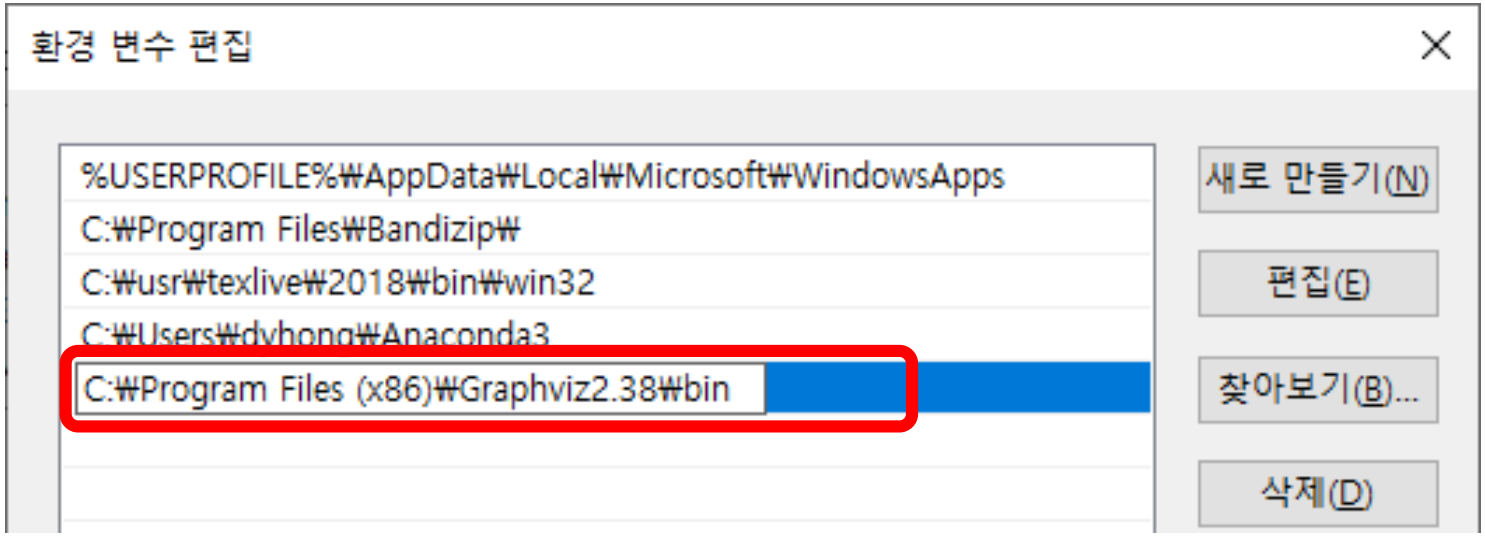
2. Select 'Edit'

Graphviz Installation with pip

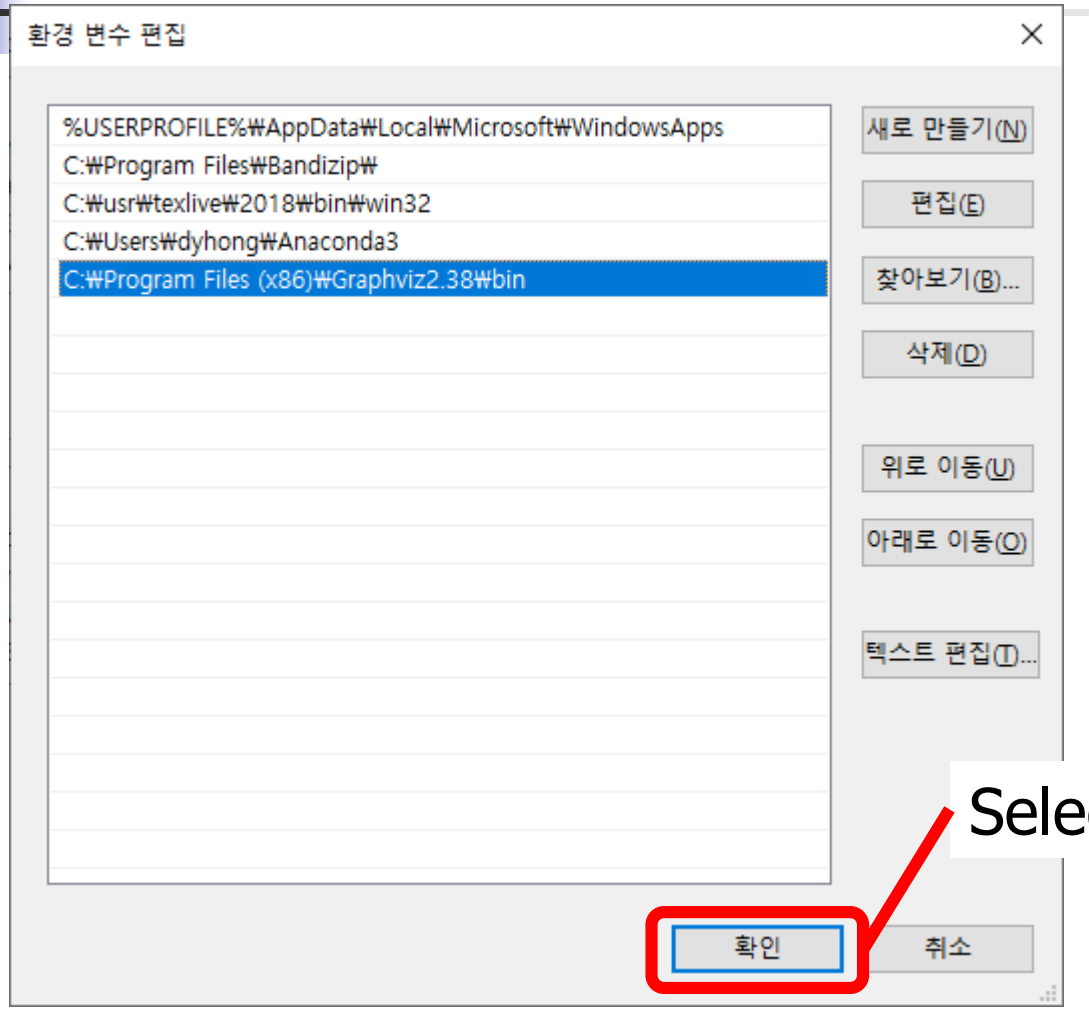


Graphviz Installation with pip

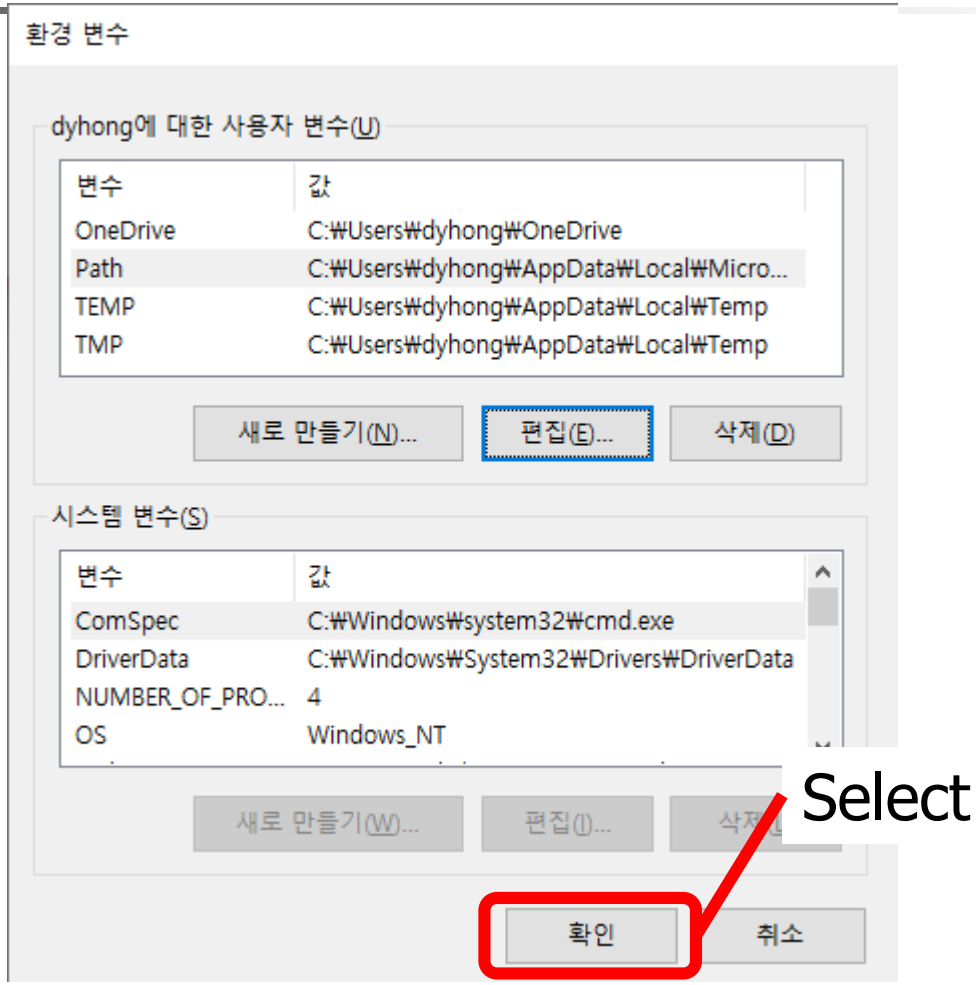
- Write down
  "[the installed Graphviz path]\bin"
  - In this example,
    C:\Program Files (x86)\Graphviz2.38\bin



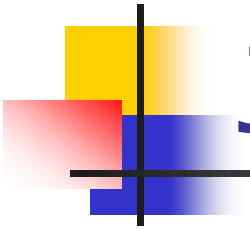
Graphviz Installation with pip



Graphviz Installation with pip



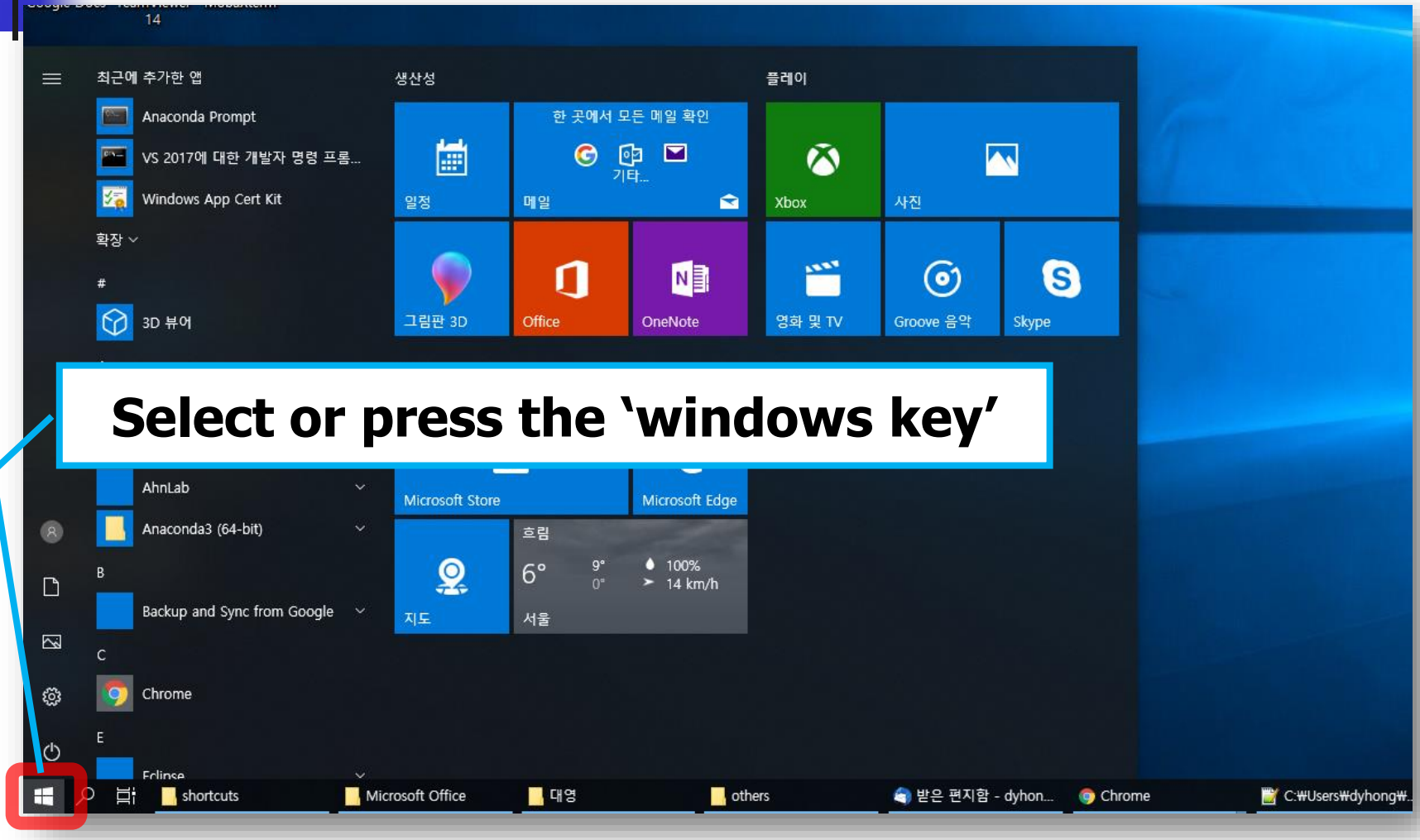
JUPYTER NOTEBOOK



Jupyter Notebook

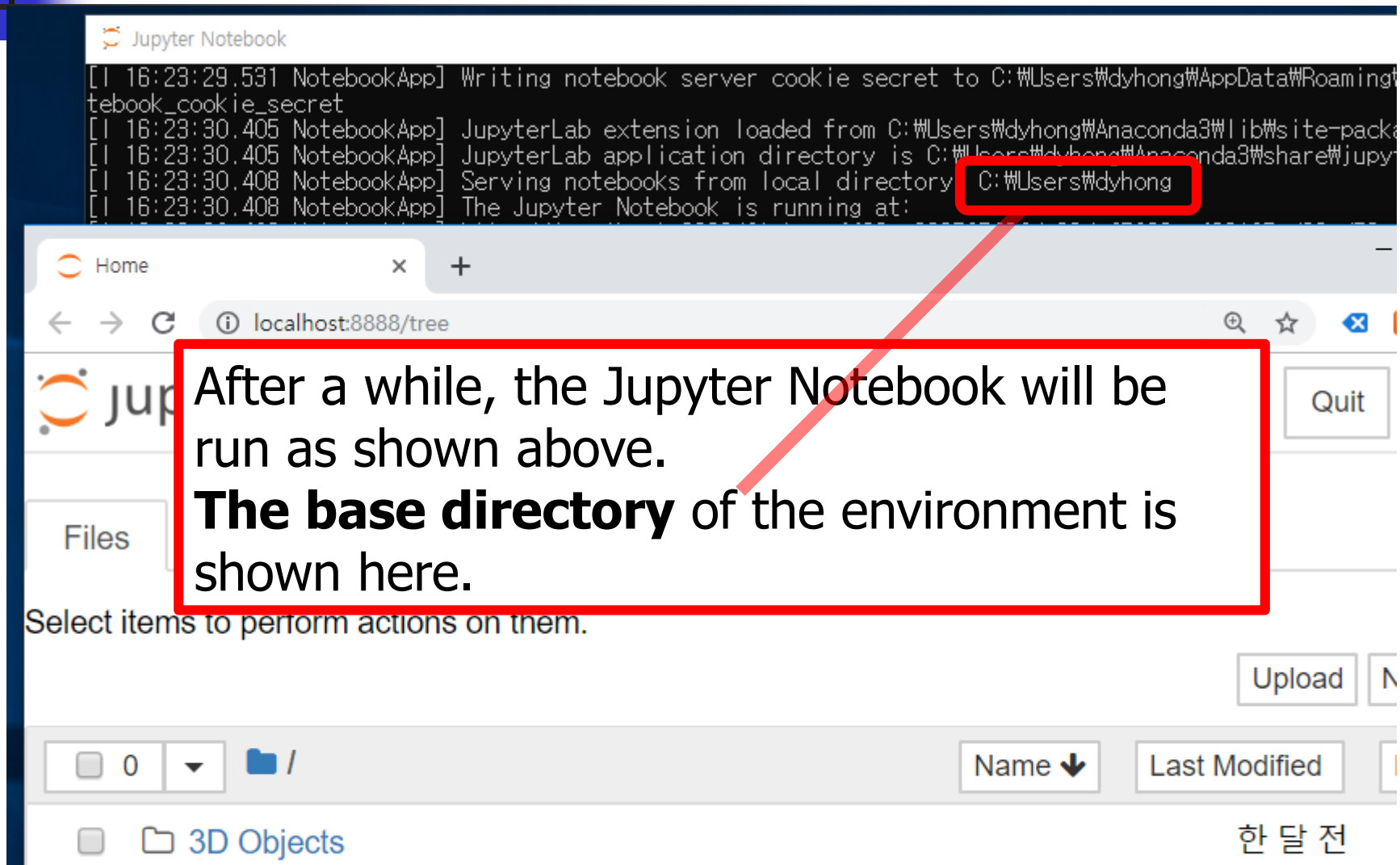
- An open-source web application that allows you to edit and run python code
- Already installed since it is contained in Anaconda

Running Jupyter Notebook



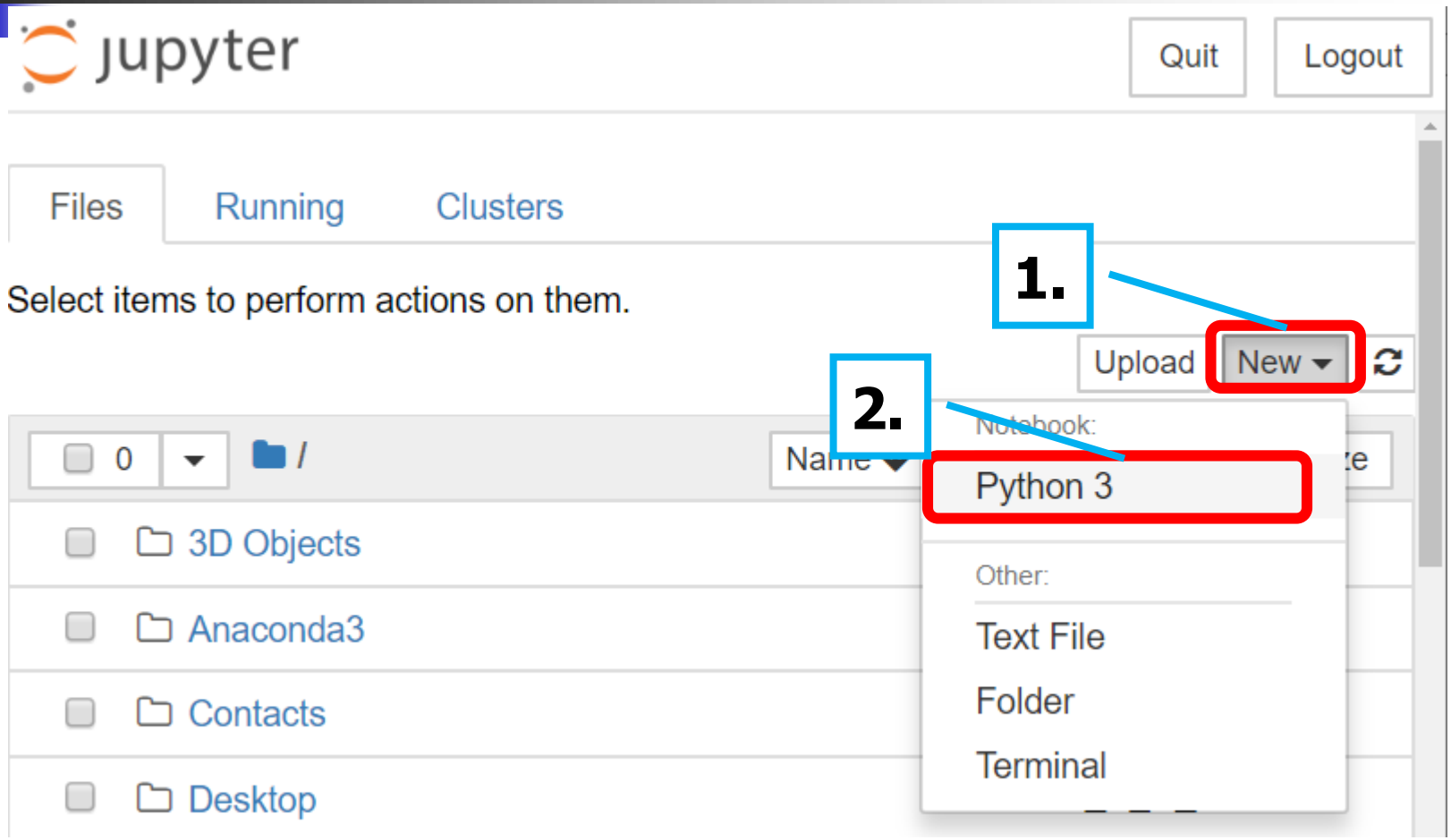
Running Jupyter Notebook

Running Jupyter Notebook



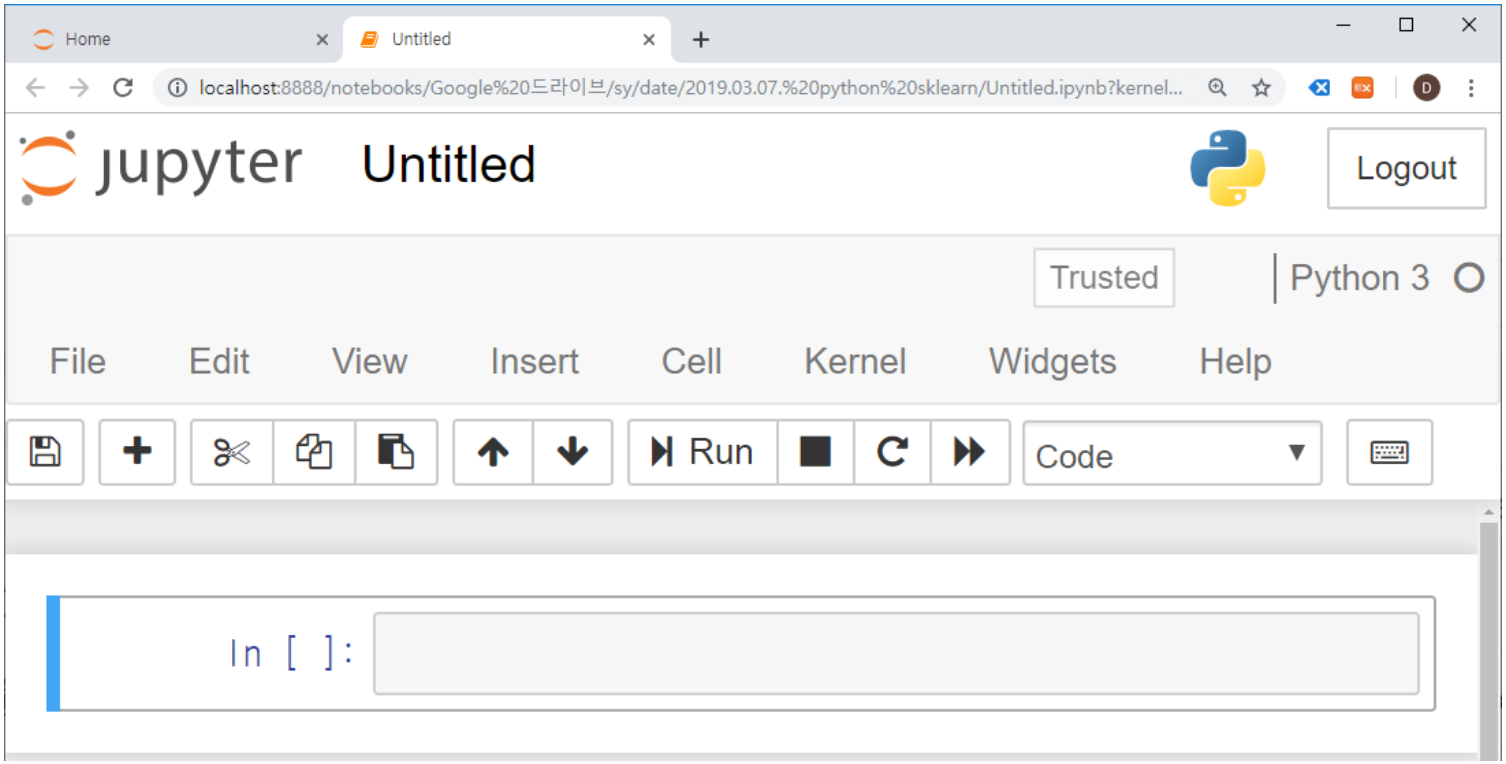
After a while, the Jupyter Notebook will be run as shown above.
The base directory of the environment is shown here.

Creating a Notebook Document



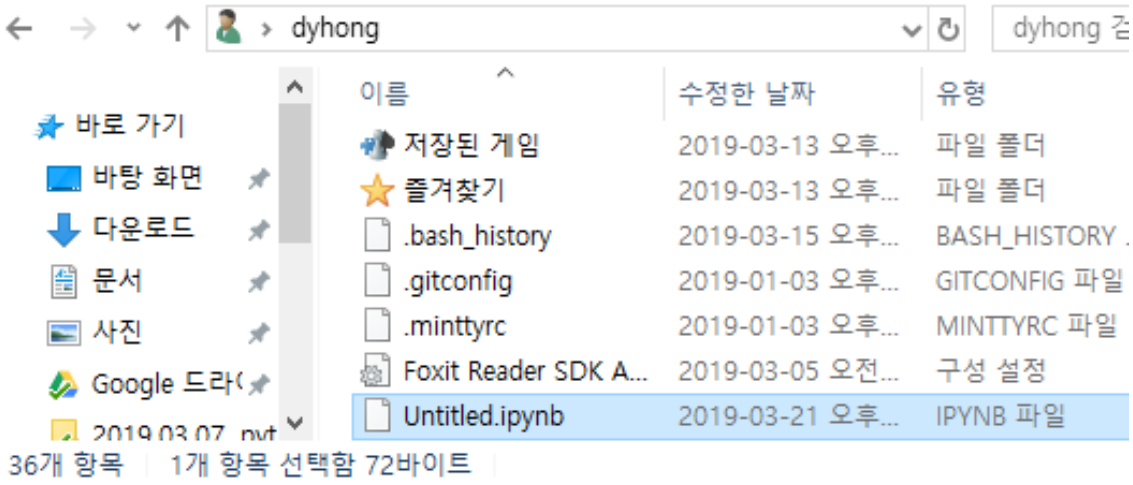
Creating a Notebook Document

- A Notebook document is created

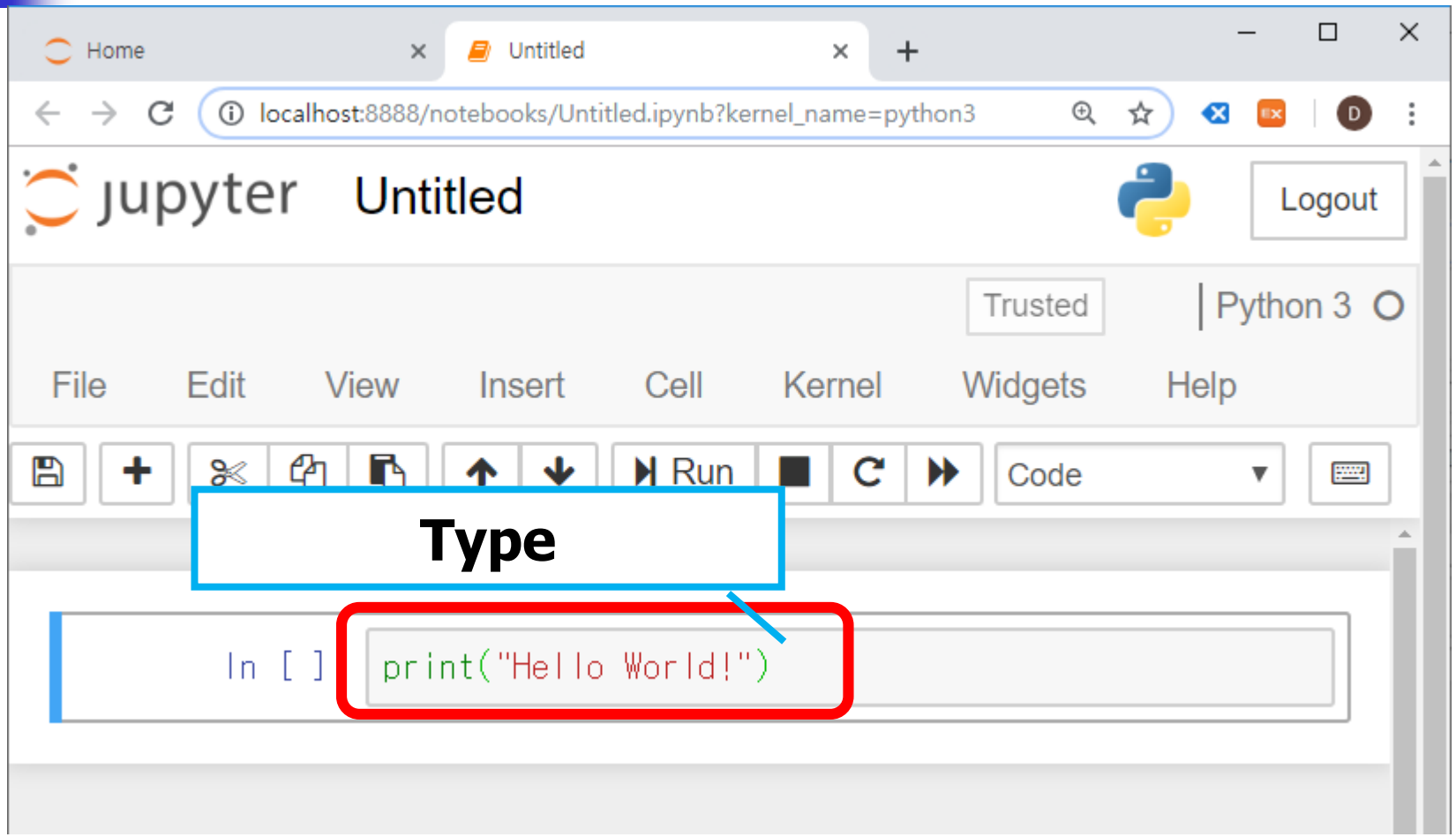


Creating a Notebook Document

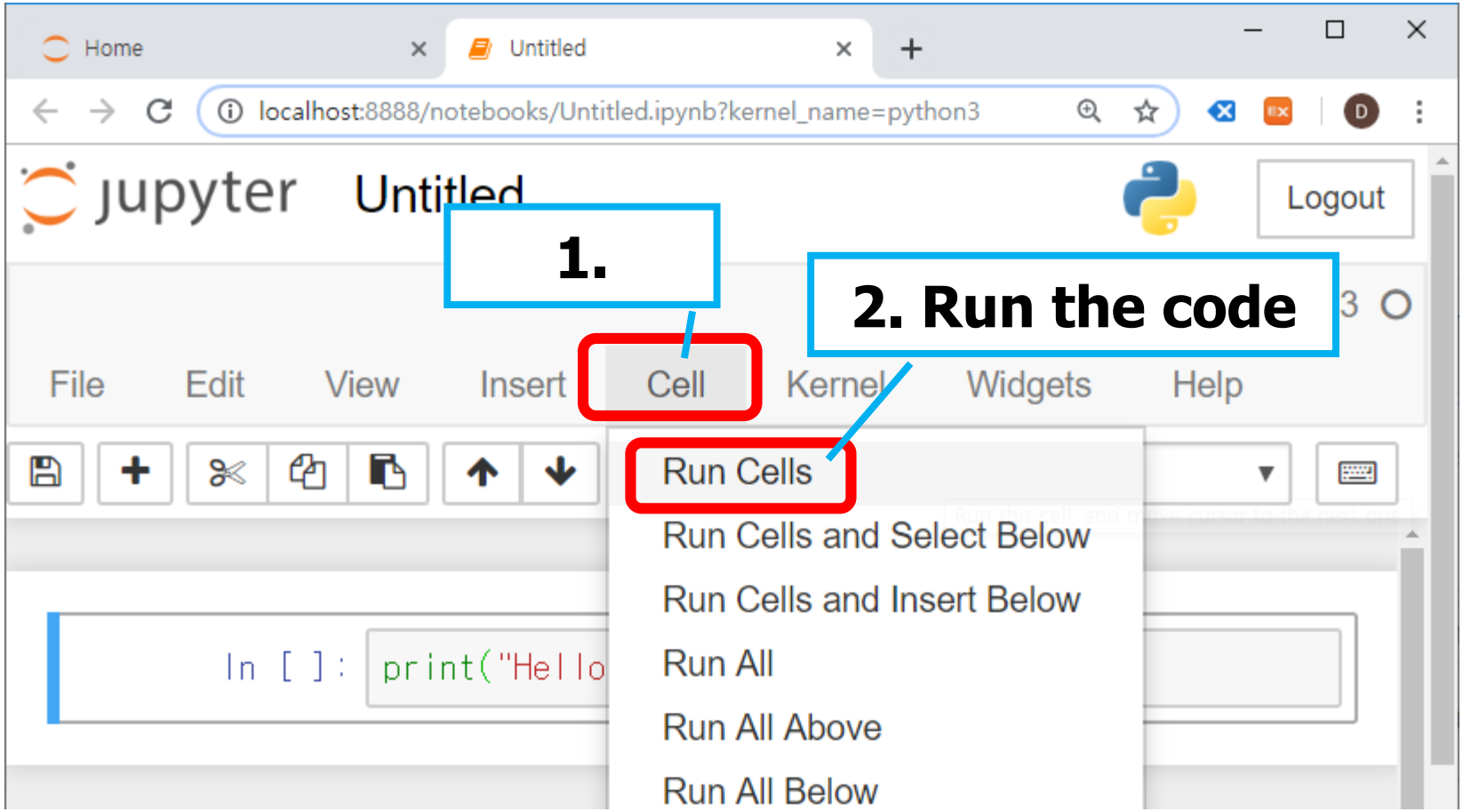
- You can see the file (Untitled.ipynb) is created in the base directory
  - In this example, the base directory path is `C:\Users\dyhong`



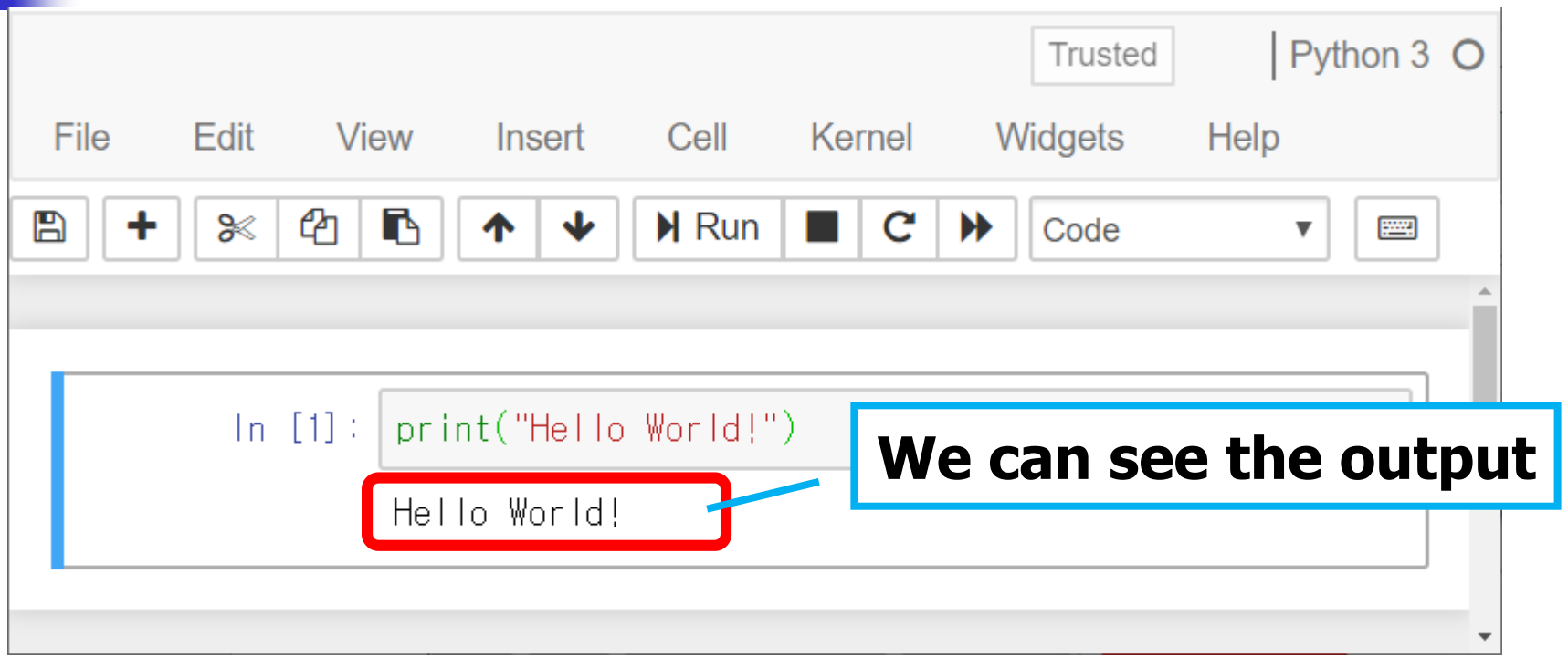
Hello World!

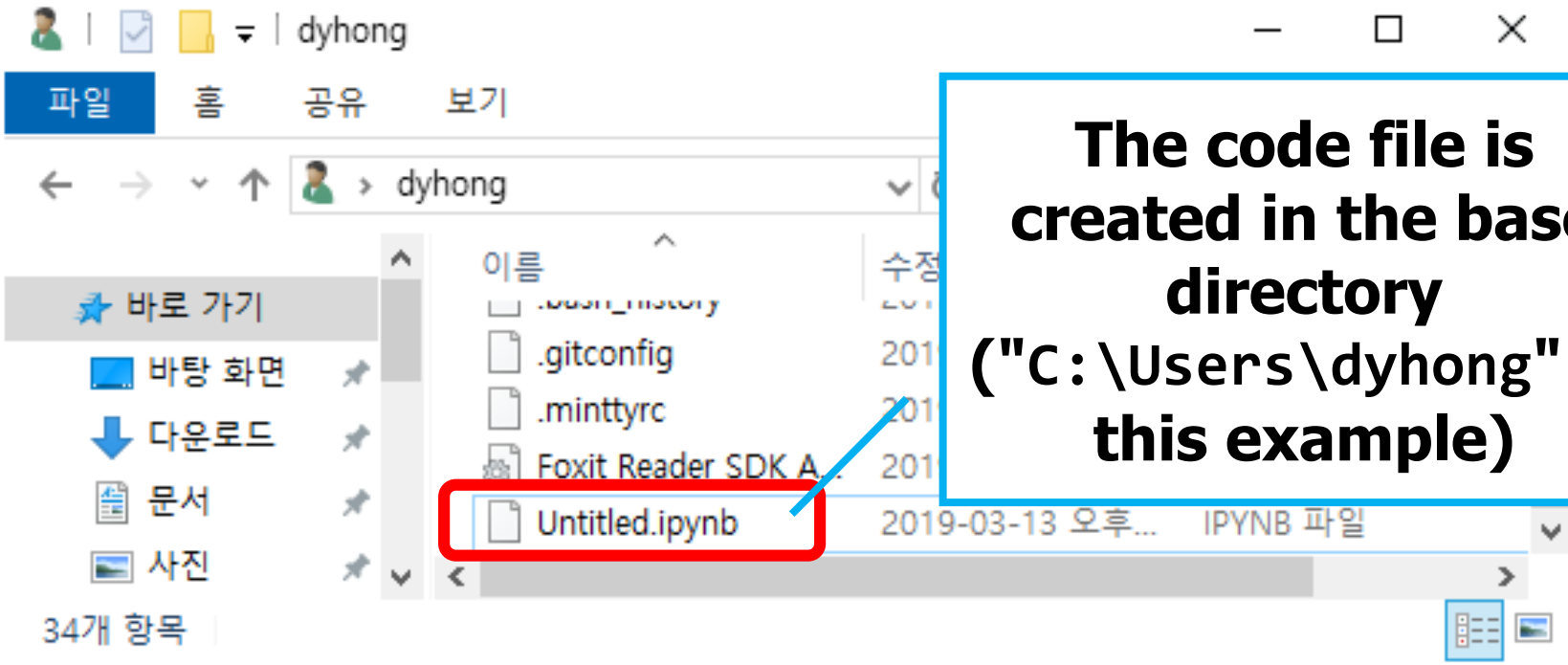

Home
Untitled
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
jupyter
Untitled
Logout
Trusted
Python 3
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
Type
In [ ]
print("Hello World!")

Hello World!


Home
Untitled
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
jupyter
Untitled
Logout
1.
2. Run the code
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run Cells
Run Cells and Select Below
Run Cells and Insert Below
Run All
Run All Above
Run All Below
In [ ]:
print("Hello

Hello World!

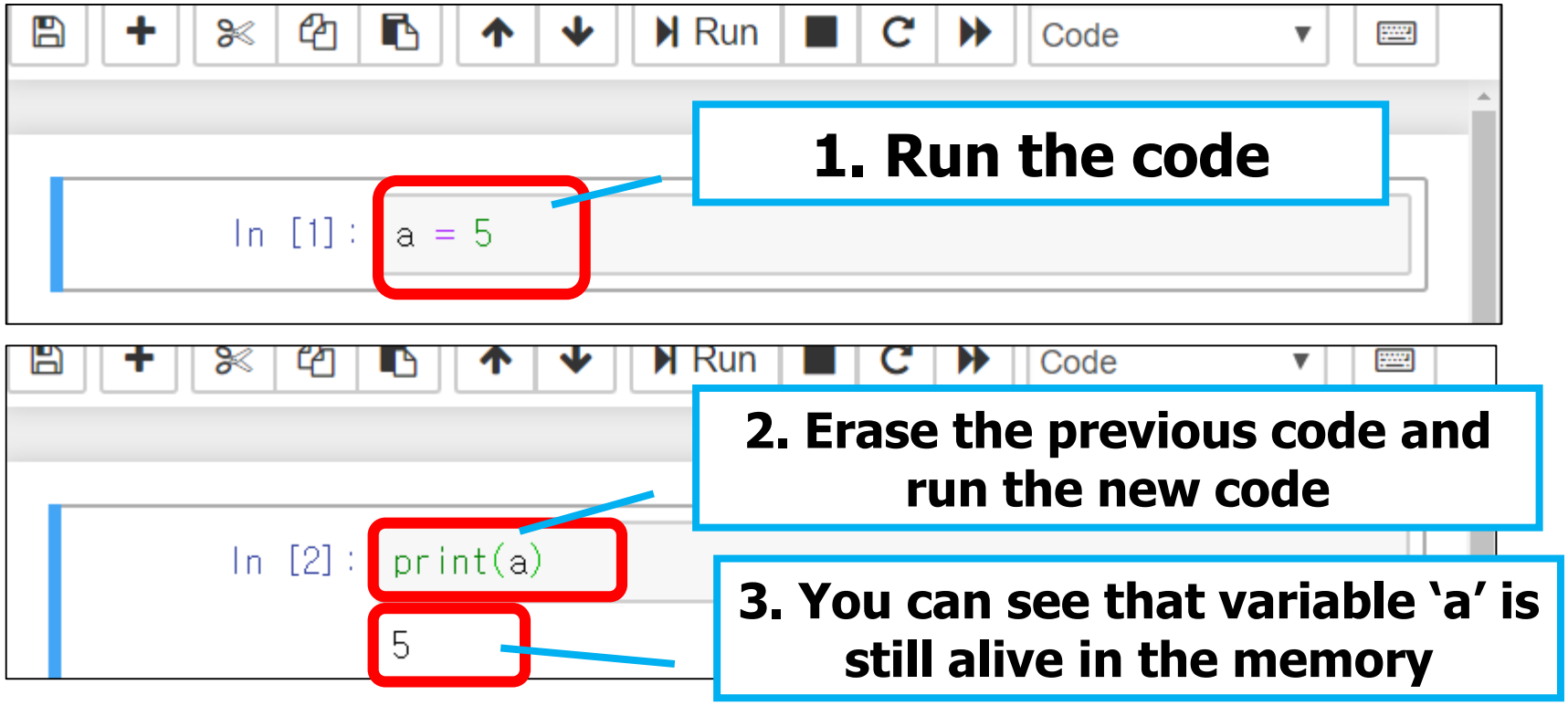

Trusted
Python 3
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
In [1]:
print("Hello World!")
Hello World!
We can see the output



The code file is created in the base directory ("C:\Users\dyhong" in this example)

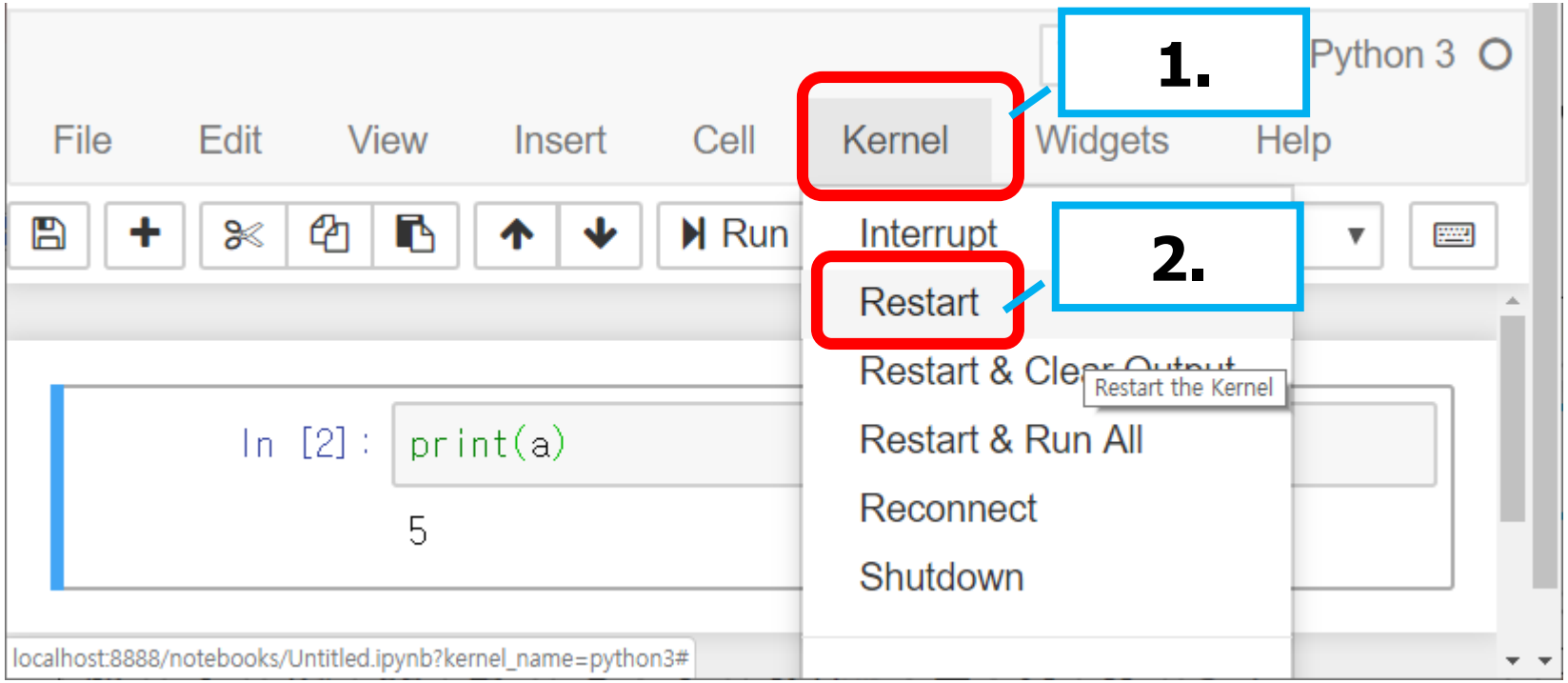
Kernel

- Variables always remain in the memory unless you shut down the kernel in the Jupyter Notebook

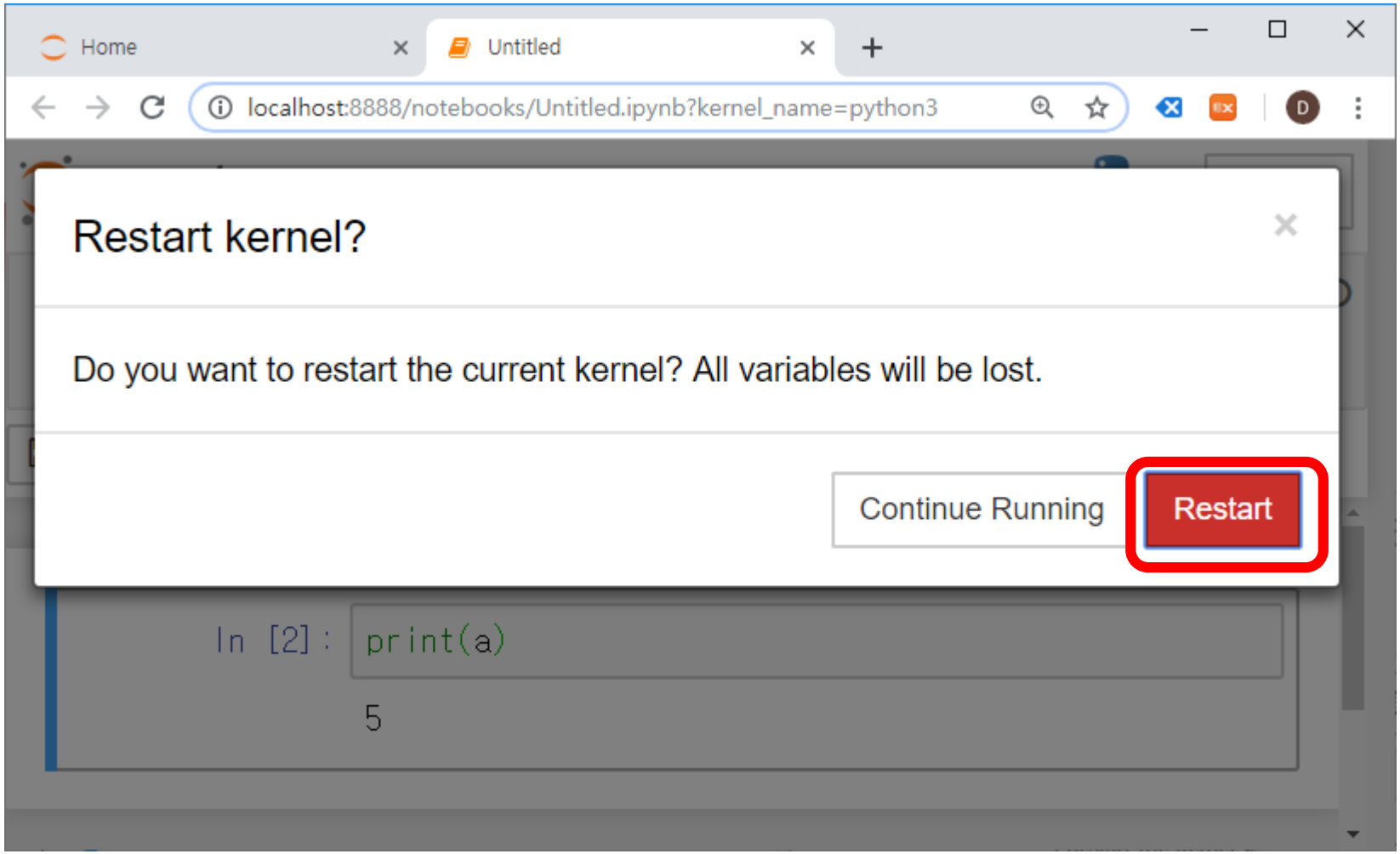


Kernel

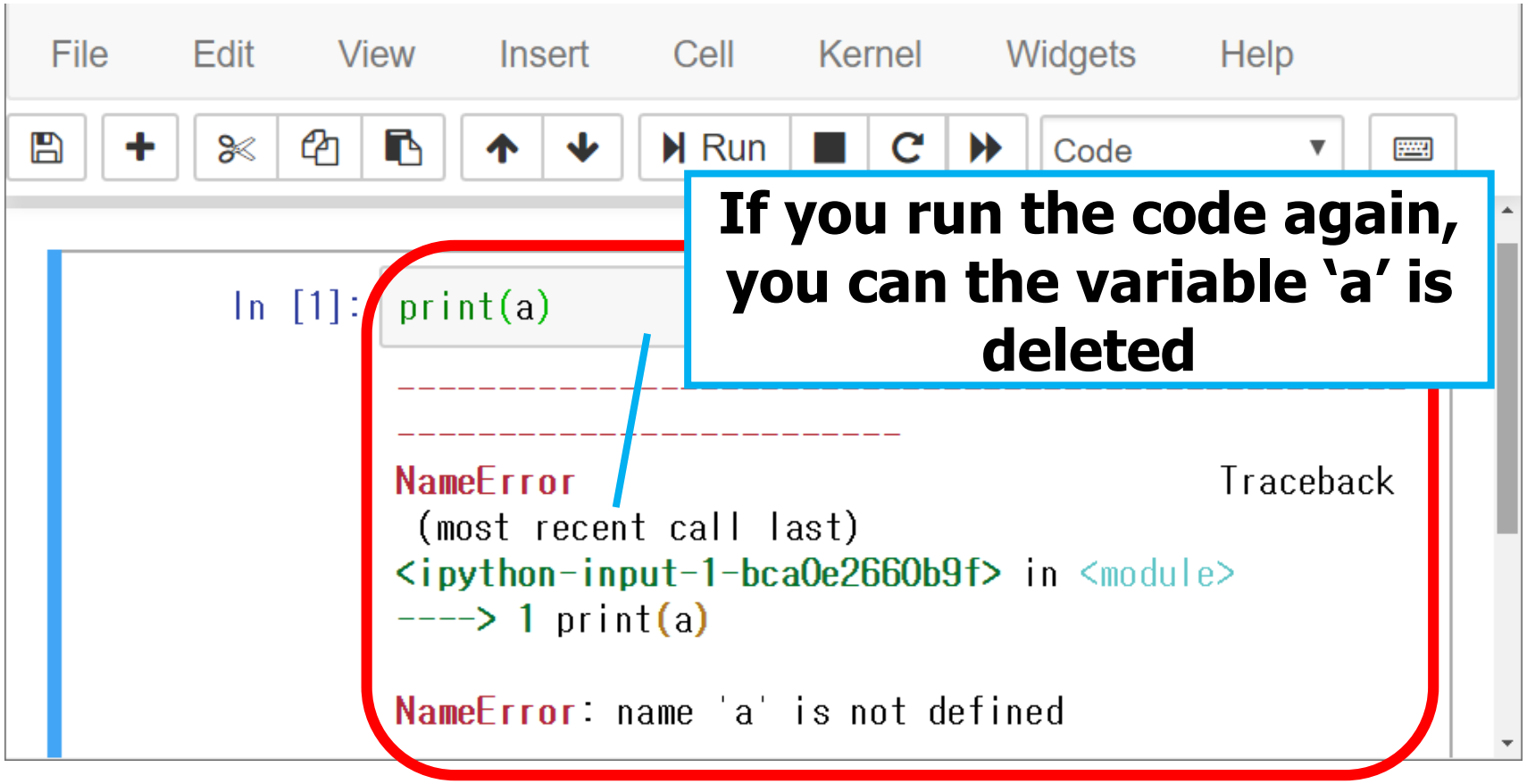
- If you want to start over a computation from scratch (e.g. variables are deleted, open files are closed, etc…), restart the kernel



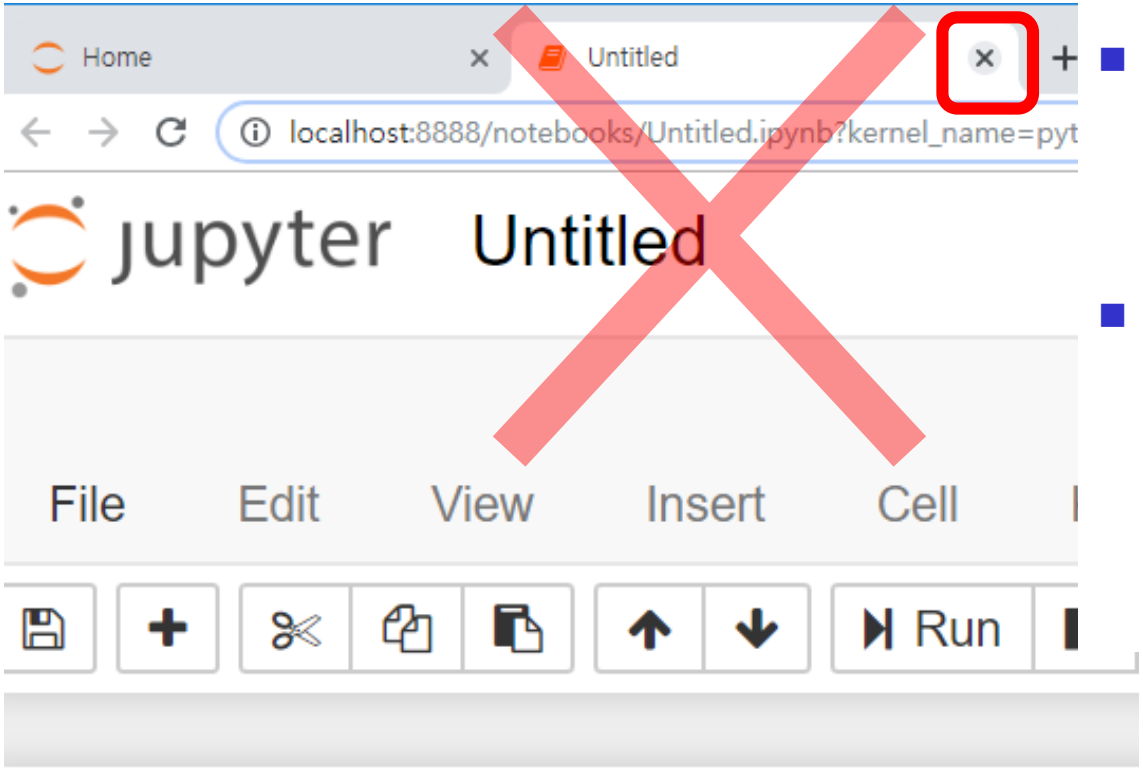
Kernel


Home
Untitled
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
Restart kernel?
Do you want to restart the current kernel? All variables will be lost.
Continue Running
Restart
In [2]:
print(a)
5

Kernel

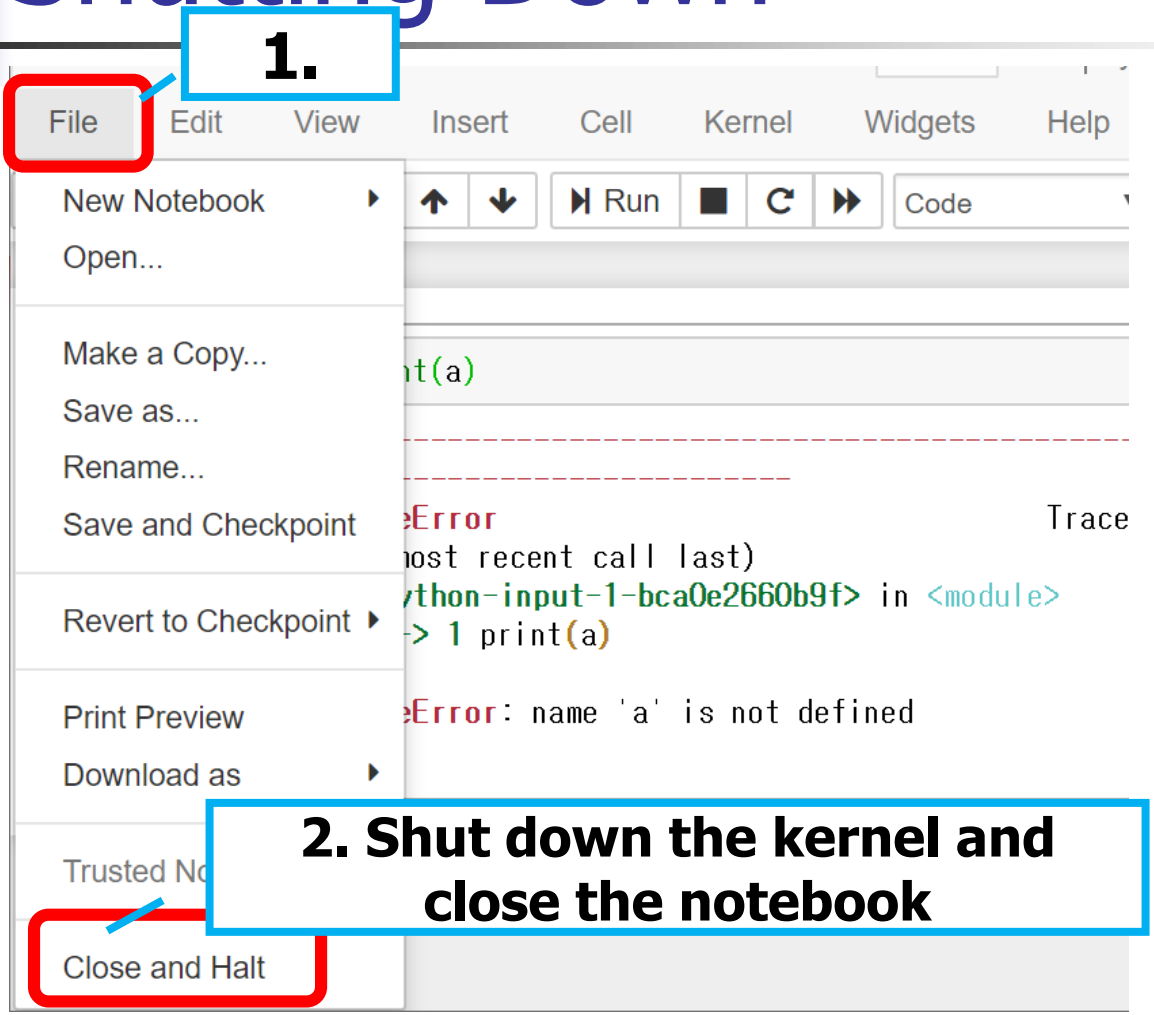

File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
If you run the code again, you can the variable 'a' is deleted
In [1]:
print(a)
NameError
Traceback
(most recent call last)
<ipython-input-1-bca0e2660b9f> in <module>
----> 1 print(a)
NameError: name 'a' is not defined

Close a Notebook



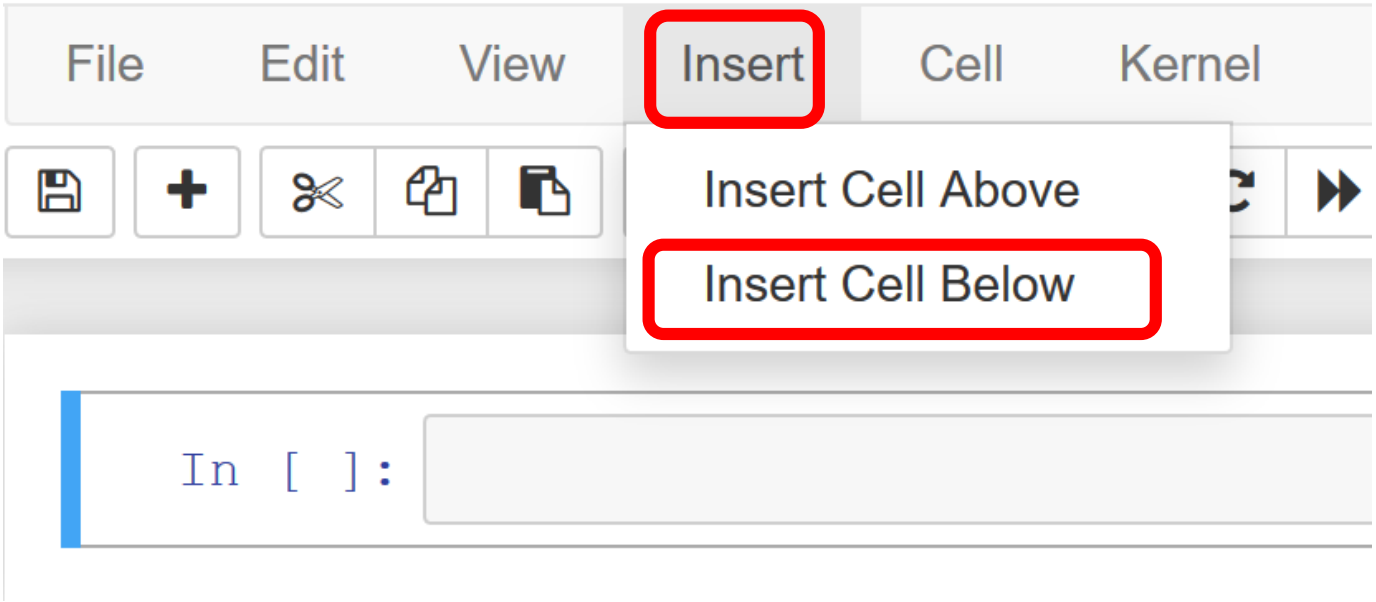
- Closing the notebook browser tab does not shut down the kernel
- Instead, the kernel keeps running until we explicitly shut down

Shutting Down



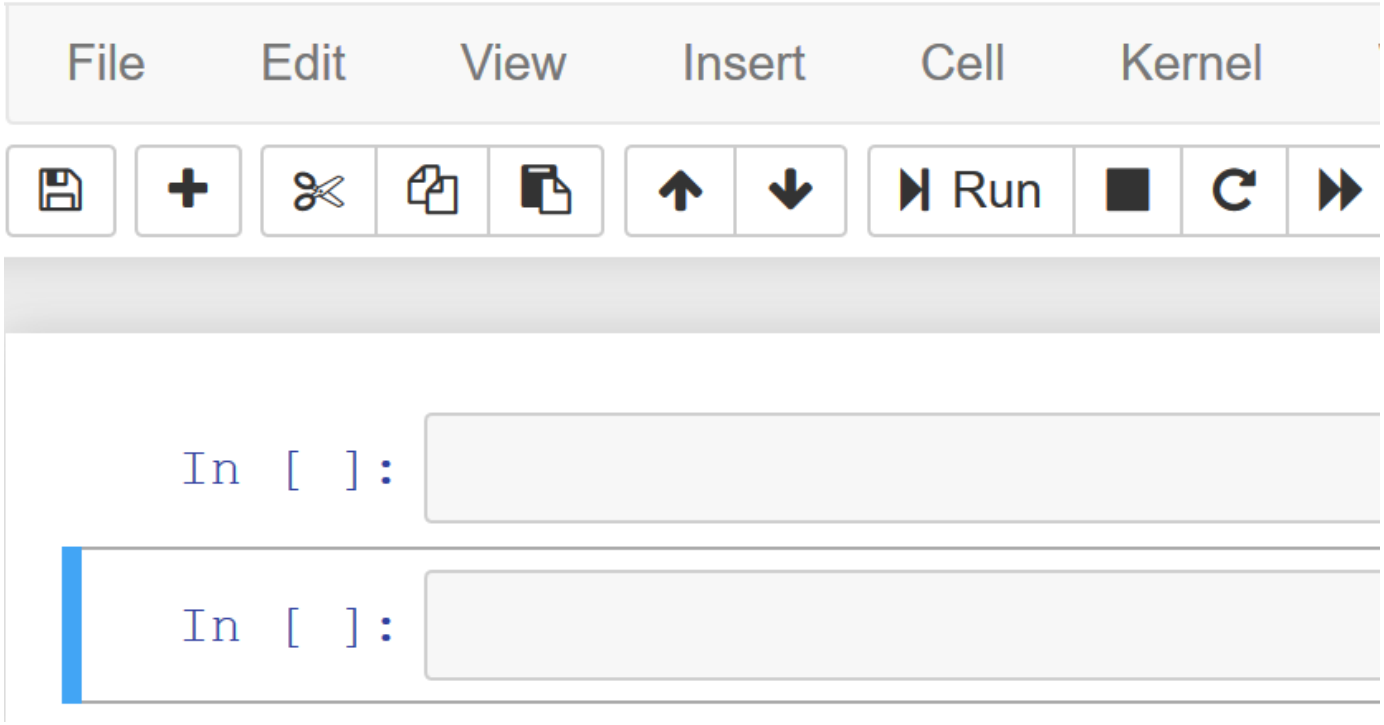
Cells

- By dividing the code into cells, you can run the code partially



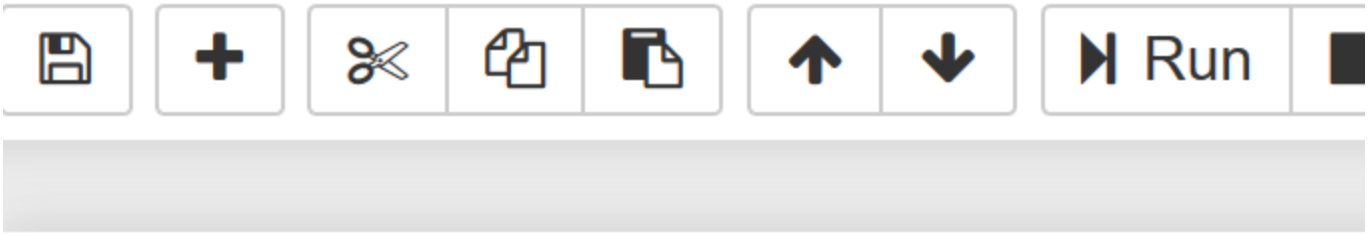
Cells

- A new cell is created



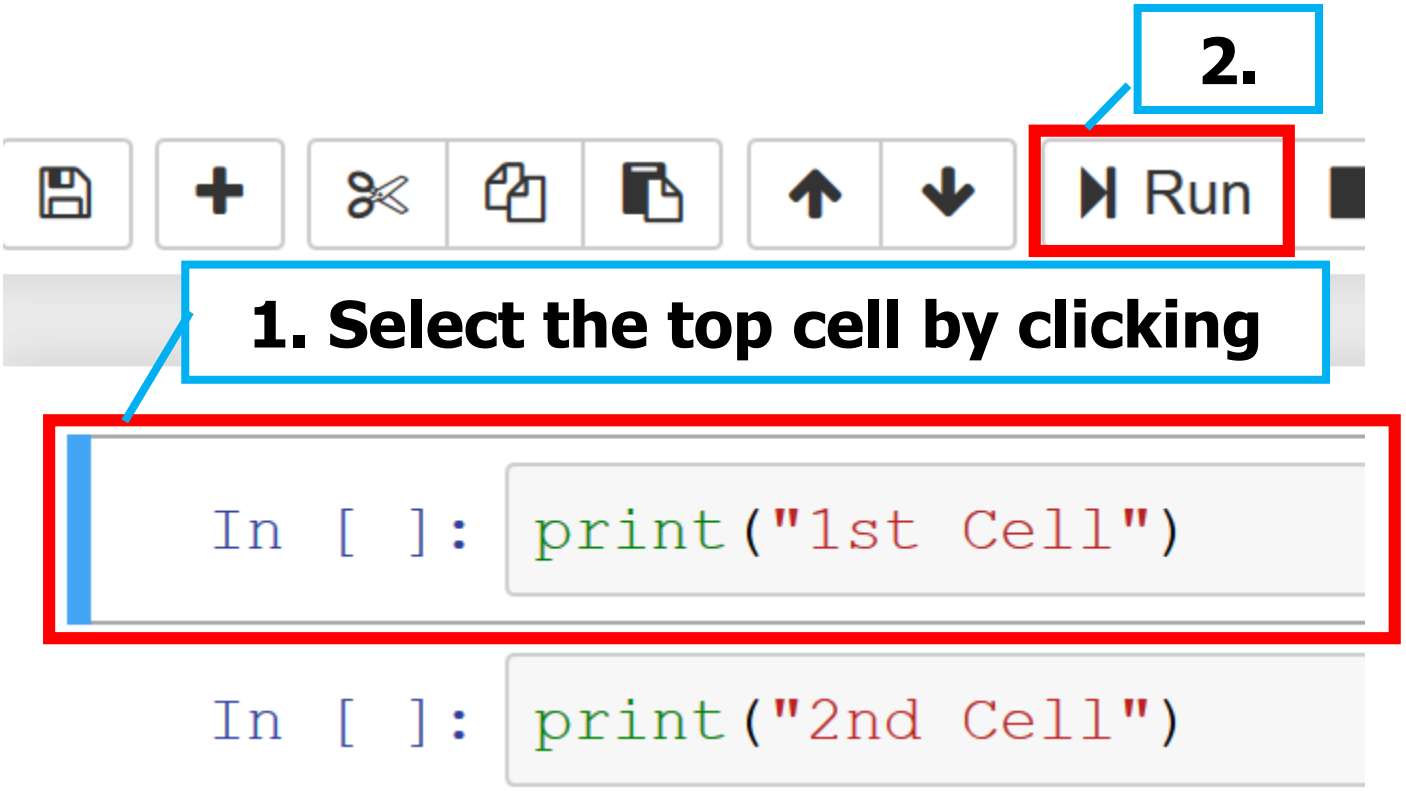
Cells

- Write down the code as follows

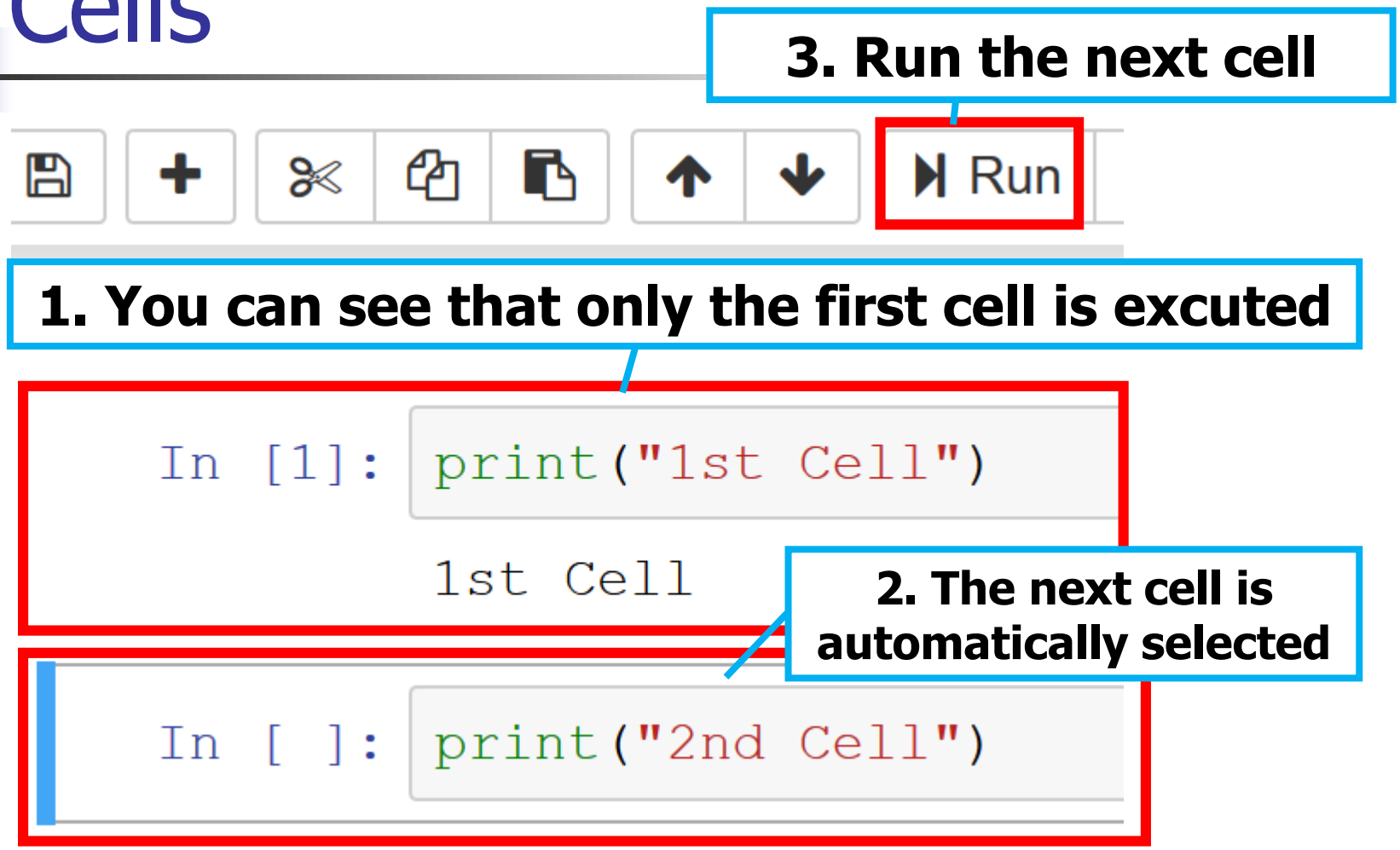


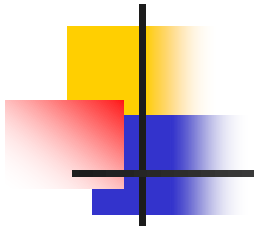

```
In [ ]: print("1st Cell")

In [ ]: print("2nd Cell")
```

Cells


2.
Run
1. Select the top cell by clicking
In [ ]: print("1st Cell")
In [ ]: print("2nd Cell")

Cells

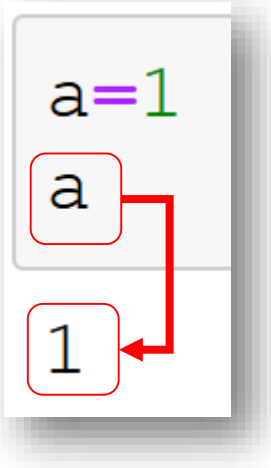

3. Run the next cell
Run
1. You can see that only the first cell is excuted
In [1]: print("1st Cell")
1st Cell
2. The next cell is automatically selected
In [ ]: print("2nd Cell")



- The 2nd cell is also executed

```
In [1]: print("1st Cell")

        1st Cell

In [2]: print("2nd Cell")

        2nd Cell
```

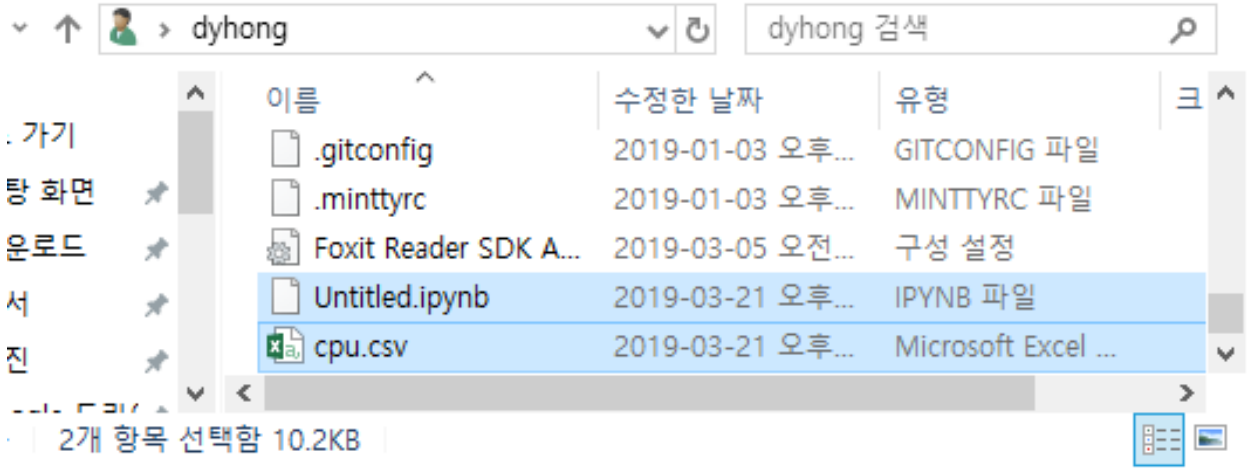
Cells



The value of the last line in the cell is automatically evaluated and printed

Load Data

- Download cpu.csv from http://kdd.snu.ac.kr/python/
- Save the csv file in the same directory as the source file (.ipynb)



Load Data

- *pandas*: A library that supports data analysis

Import the *pandas* library
and use alias 'pd' instead of 'pandas'

```
import pandas as pd
df = pd.read_csv("cpu.csv")
df[:5]
```

Read the csv file and store it in a pandas dataframe object

select the rows up to 5th row

cpu.csv

MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	class
125	256	6000	256	16	128	198
29	8000	32000	32	8	32	269
29	8000	32000	32	8	32	220
29	8000	32000	32	8	32	172
29	8000	16000	32	8	16	132

Load Data

```
import pandas as pd
df = pd.read_csv("cpu.csv")
df[:5]
```

	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	class
0	125	256	6000	256	16	128	198
1	29	8000	32000	32	8	32	269
2	29	8000	32000	32	8	32	220
3	29	8000	32000	32	8	32	172
4	29	8000	16000	32	8	16	132

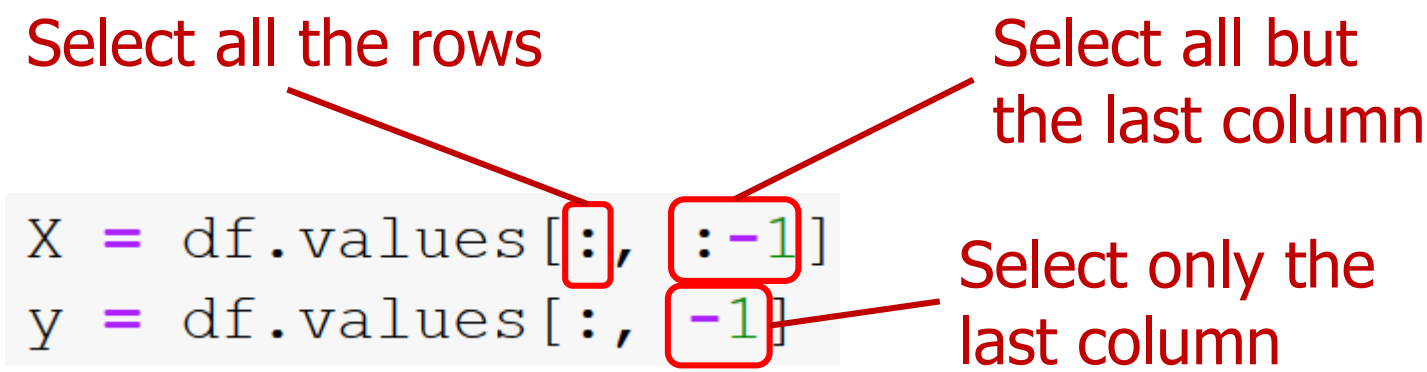
Load Data

- df.values returns an array form of the data

```
print(df.values)
```

```
[[  125   256  6000 ...    16   128   198]
 [   29  8000 32000 ...     8    32   269]
 [   29  8000 32000 ...     8    32   220]
 ...
 [  125  2000  8000 ...     2    14    52]
 [  480   512  8000 ...     0     0    67]
 [  480  1000  4000 ...     0     0    45]]
```

Load Data


Select all the rows
Select all but the last column
X = df.values[:, :-1]
y = df.values[:, -1]
Select only the last column

Load Data

In [3]:
```
X = df.values[:, :-1]
y = df.values[:, -1]
print(X[:3])
print(y[:3])
```

X[:3]

```
[[  125   256  6000   256    16   128]
 [   29  8000 32000    32     8    32]
 [   29  8000 32000    32     8    32]]
```

Y[:3]

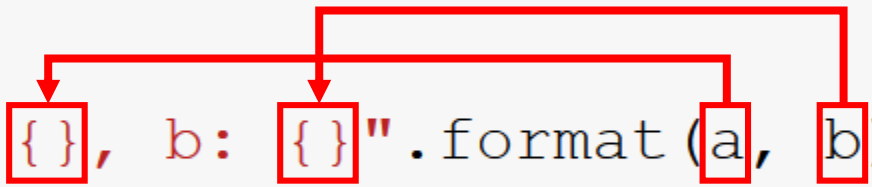
```
[198 269 220]
```

	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	class
0	125	256	6000	256	16	128	198
1	29	8000	32000	32	8	32	269
2	29	8000	32000	32	8	32	220

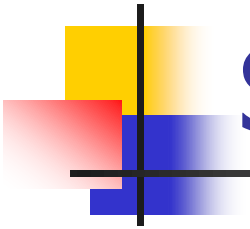
Print

- You can control the formatting of output string by using format() function and {} mark in the string

```
a = 1
b = 2
print("a: {}, b: {}".format(a, b))
```



```
a: 1, b: 2
```



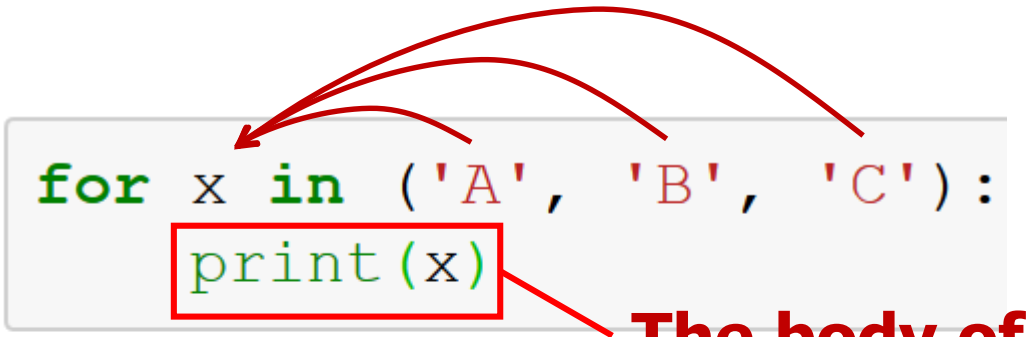
Shape

- You can check dimensions of array by using the *shape* attribute

```
print("X.shape: {}, y.shape: {}".format(
    X.shape, y.shape))
```

```
X.shape: (209, 6), y.shape: (209,)
```

For Loop




```
A
B
C
```

감사 합니다!