Lecture #2 | python review: function, list, list comprehension

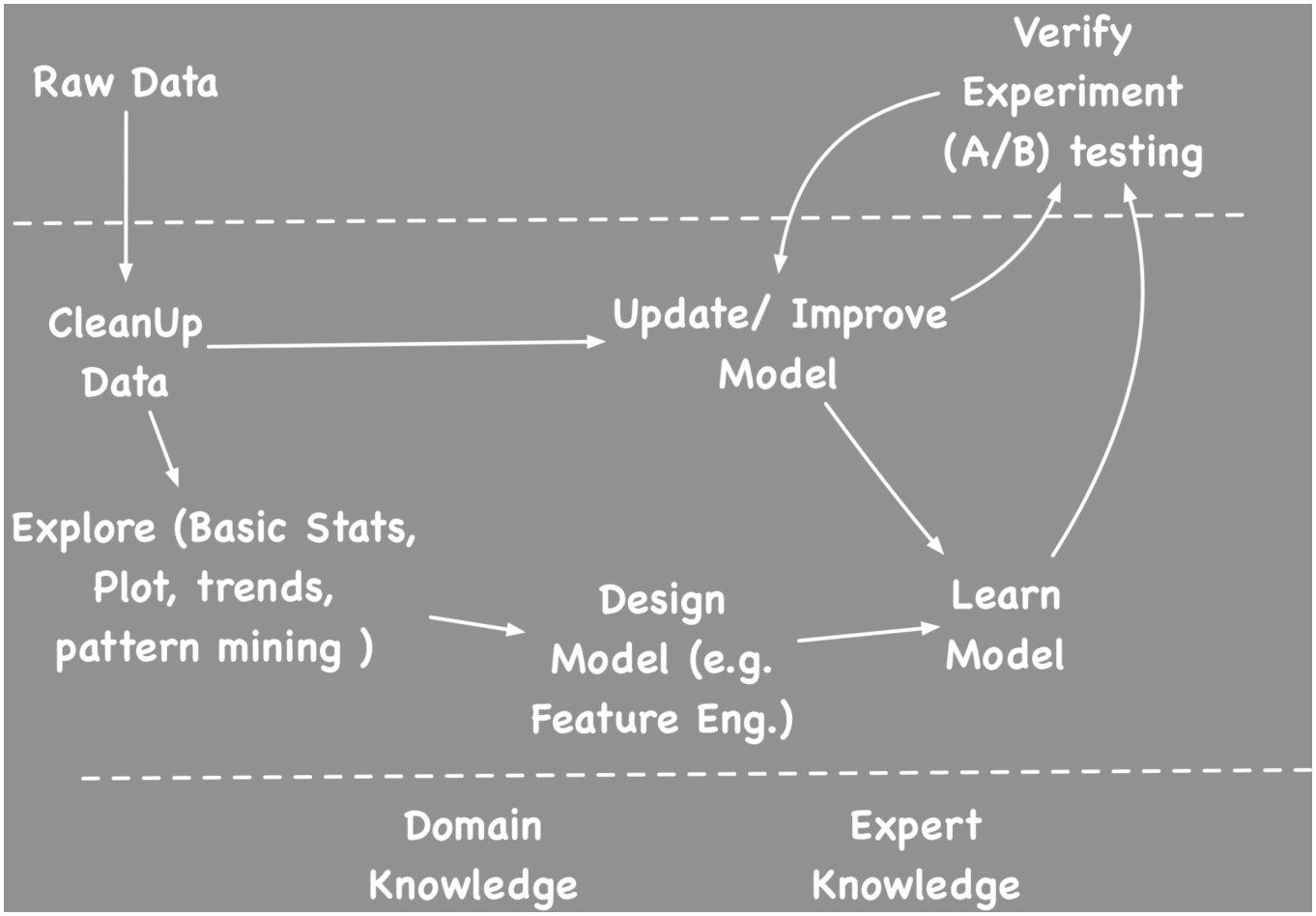
SE271 Object-oriented Programming (2017)

Prof. Min-gyu Cho

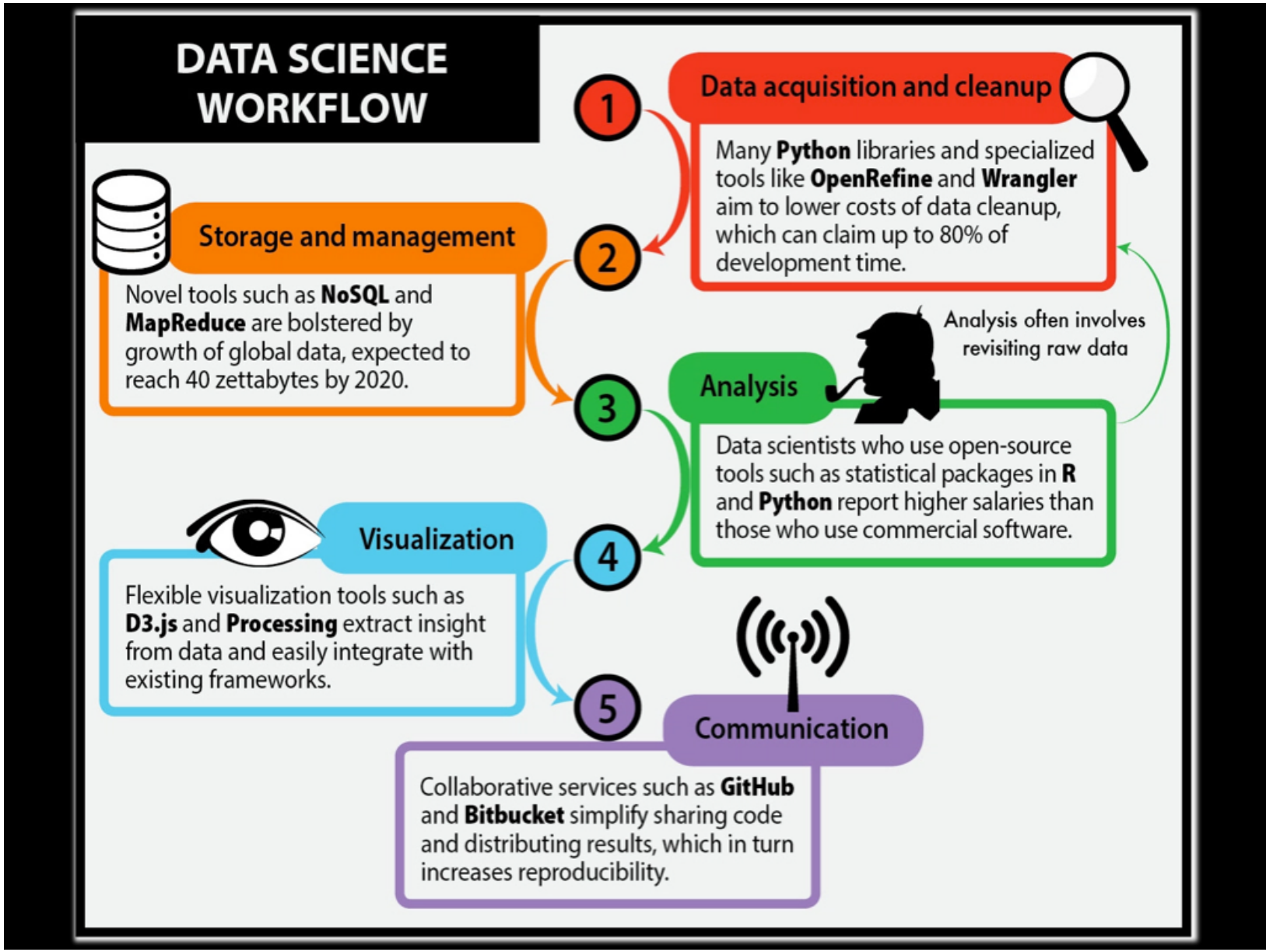
Today's Topic

- A few comments on data science
- Jupyter notebook
- python review

Data science workflow



Source: https://www.slideshare.net/hemapani/introduction-to-data-science-and-analytics/3


DATA SCIENCE WORKFLOW
1
Data acquisition and cleanup
Many Python libraries and specialized tools like OpenRefine and Wrangler aim to lower costs of data cleanup, which can claim up to 80% of development time.
2
Storage and management
Novel tools such as NoSQL and MapReduce are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.
Analysis often involves revisiting raw data
3
Analysis
Data scientists who use open-source tools such as statistical packages in R and Python report higher salaries than those who use commercial software.
4
Visualization
Flexible visualization tools such as D3.js and Processing extract insight from data and easily integrate with existing frameworks.
5
Communication
Collaborative services such as GitHub and Bitbucket simplify sharing code and distributing results, which in turn increases reproducibility.

python v.s. R

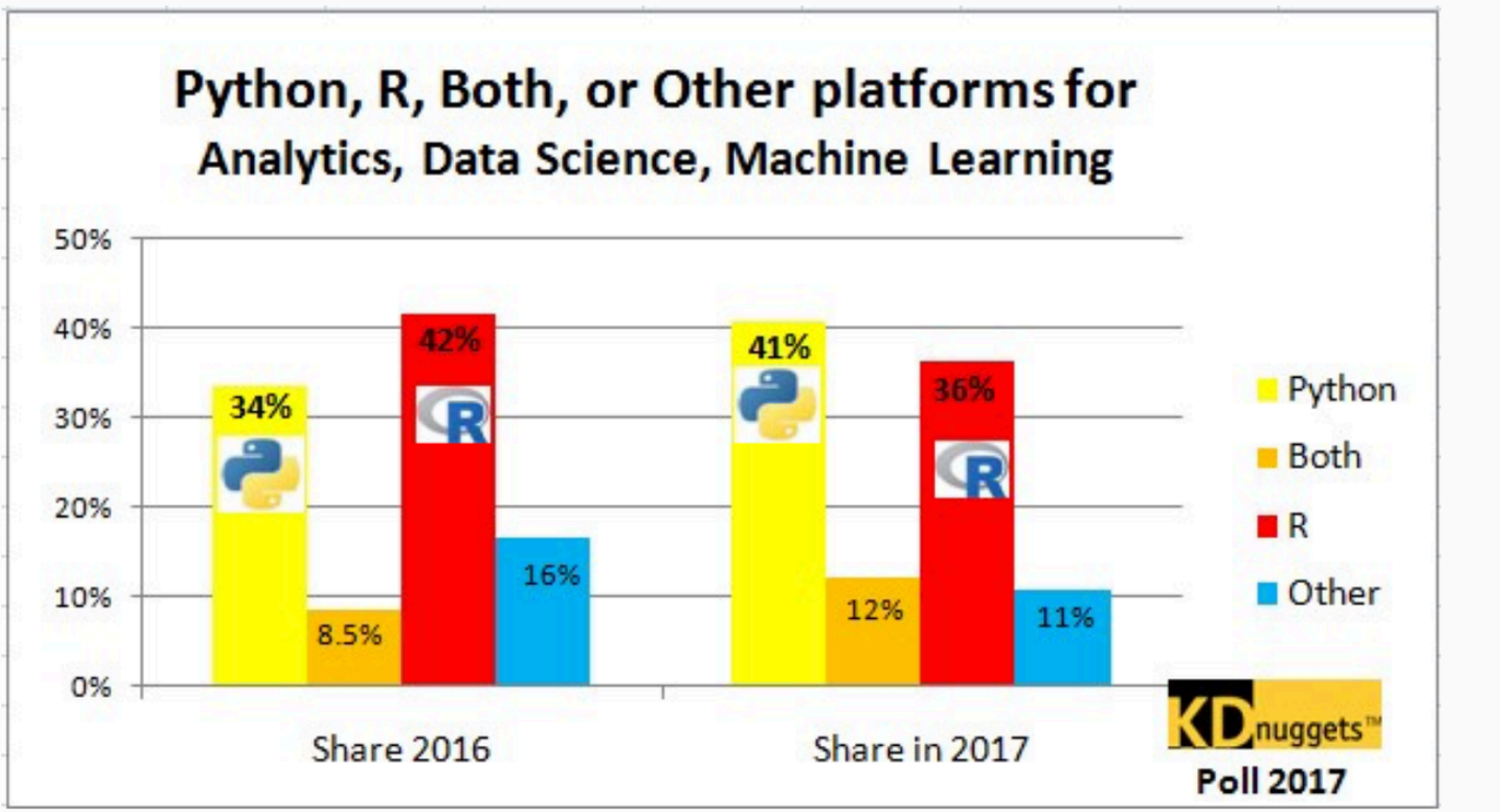


Fig. 1: Share of Python, R, Both, or Other platforms usage for Analytics, Data Science, Machine Learning, 2016 vs 2017

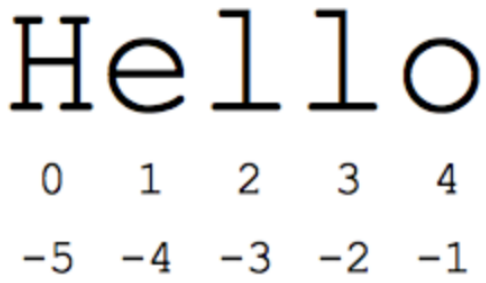
Source: http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

The Jupyter Notebook

- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text
- Powerful and convenient tool for data science (with python, R, …)
  - Interactive development (REPL*-like environment)
  - Documentation of the analysis processes and/or final documents
- Examples
  - https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks
  - http://nb.bianp.net/sort/views/
  - http://nbviewer.jupyter.org/github/pybokeh/ipython_notebooks/blob/master/pandas/PandasCheatSheet.ipynb

* Read, Evaluate, Print, Loop

Reference: slicing



- `s = 'Hello'`

Slicing	Result
s[0]	H
s[4]	o
len(s)	5
s[-1]	o
s[len(s)-1]	o

Slicing	Result
s[-5]	H
s[len(s)-5]	H
s[1:4]	ell
s[1:4:2]	el
s[:]	Hello

Slicing	Result
s[:3]	Hel
s[2:]	llo
s[-3:-1]	ll
s[-1:-3:-1]	ol
s[::-1]	olleH

Reading list

- The Jupyter Notebook
  - Installation Guide
  - The Jupyter Notebook (book)
  - Jupyter notebook 이해하기
- List comprehensions: https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions

ANY QUESTIONS?