Lecture #4 | pandas (Part 1)

SE377 Introduction to Big Data Analysis and Visualization (2017)

Prof. Min-gyu Cho

Topics for the following 2 week

- 9/18: Presentation of data analysis examples
- Read, manipulate and visualization of structured data with pandas and matplotlib
  - `pandas`: read and manipulate (or select) data we are interested
  - `matplotlib/seaborn`: visualize the given data
- Tidy data

Datatype of numpy array

- numpy array has data type (similar to typed languages such as C, C++, Java, …) for improved performance
- When creating numpy array, `dtype` argument can be provided (e.g., float64, int64)
  - See the next page for more comprehensive list of data types provided in numpy
- For general objects (incl. variable length strings), `object` type can be used

List of `dtype` supported by numpy

Type	Type code	Description
int8, uint8	i1, u1	Signed/unsigned 8 bit integer
int16, uint16	i2, u2	Signed/unsigned 16 bit integer
int32, uint32	i4, u4	Signed/unsigned 32 bit integer
int64, uint64	i8, u8	Signed/unsigned 64 bit integer
float16	f2	16-bit floating point number
float32	f4 or f	32-bit floating point number
float64	f8 or d	64-bit floating point number
float128	f16 or g	128-bit floating point number
complex64, complex128, complex256	c8, c16, c32	64/128/256-bit complex number
bool	?	Boolean
object	O	python object
string_/unicode_	S/U	Fixed length string/unicode string. For example, string/unicode string with the length of 10 is represented by S10/U10

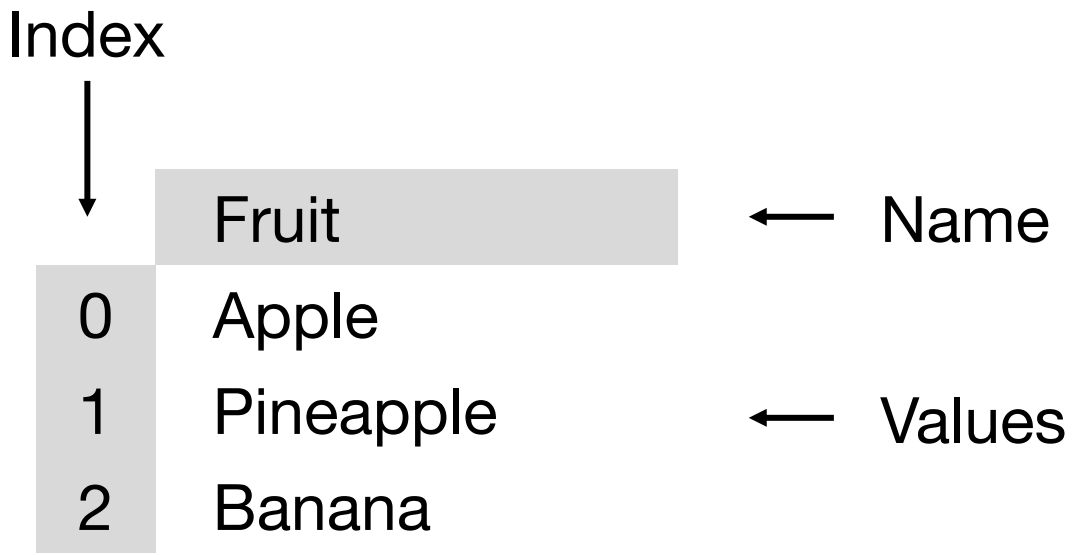
Pandas

- Easy-to-use data structures and data analysis tools
- Open source, BSD-licensed library
- Created by Wes McKinney in 2008
- c.f., very similar to `dplyr` package in R

Reference: http://pandas.pydata.org/

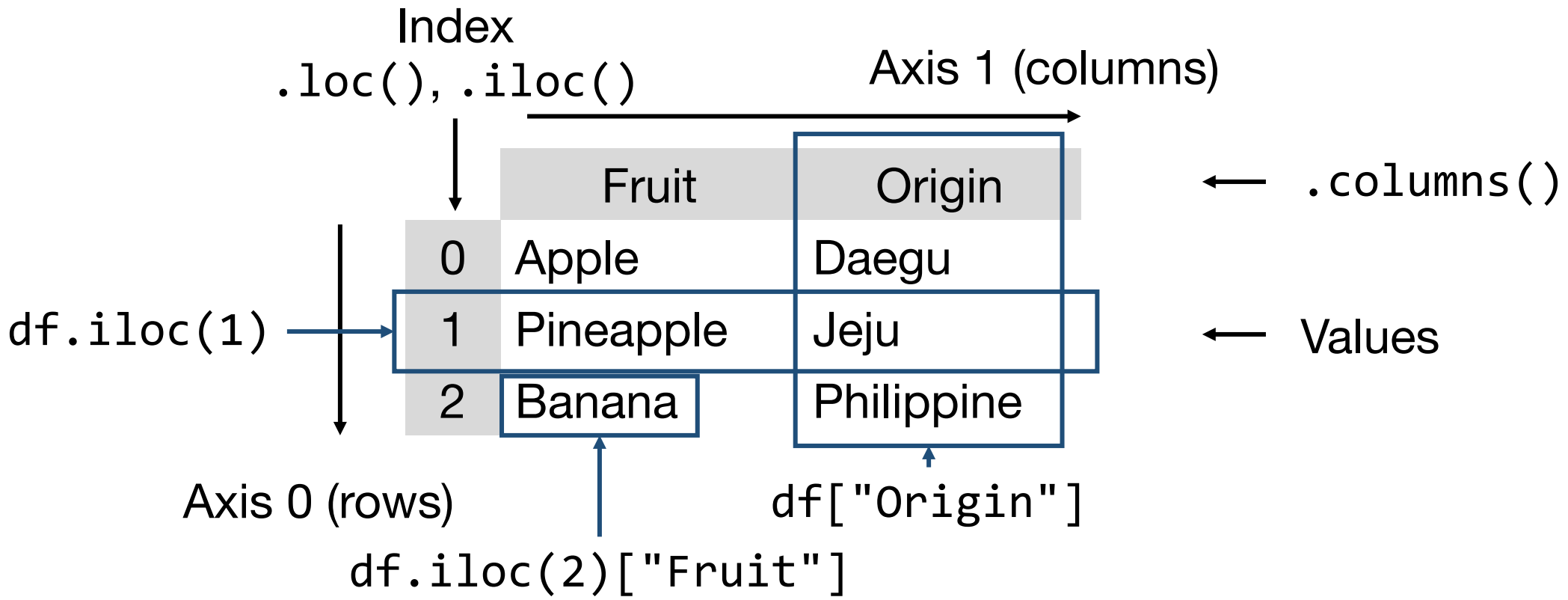
pandas: Series

- Series:1-dimensional data of ordered items with arbitrary index



pandas: DataFrame

- DataFrame
  - 2-dimensional data of ordered items with arbitrary index and column names
  - Each column may have different datatypes



Reading List (choose whatever you like most)

- Reference: https://pandas.pydata.org/pandas-docs/stable/index.html
- 10 Minutes to pandas: https://pandas.pydata.org/pandas-docs/stable/10min.html
- python + numpy pandas
  - https://www.slideshare.net/dahlmoon/pythonnumpy-pandas-1
  - https://www.slideshare.net/dahlmoon/pythonnumpy-pandas-2
  - https://www.slideshare.net/dahlmoon/pythonnumpy-pandas-3
  - https://www.slideshare.net/dahlmoon/pythonnumpy-pandas-4
- pandas cookbook: https://github.com/jvns/pandas-cookbook
- And many other really good tutorials/documents!!!

ANY QUESTIONS?