

PS5841

# Data Science in Finance & Insurance

## Data Wrangling

Yubo Wang

Spring 2022

# Numpy ndarray

- ndarray
  - a homogeneous multidimensional array
    - Dimensions are called axes
  - a table of elements of the same type indexed by a tuple of non-negative integers

# Numpy ndarray Indexing

- Indexing

`a[0:2]`

- Fancy Indexing

- Indexing with arrays of indices

`a[[...], [...]]`

- Indexing with Boolean arrays

`a[a>10]`

- Indexing with strings

- Structured Arrays - ndarrays whose datatype is a composition of simpler datatypes organized as a sequence of named fields

`a['uni']`

# Numpy Broadcasting (1)

- Broadcasting – rules for working with two ndarrays
- Rule 1 – If the two arrays differ in their number of dimensions, the **shape** of the one with fewer dimensions is padded with ones on its leading (left) side
- Rule 2 – If the shape of the two arrays does not match in a particular dimension, the array with shape equal to 1 in that dimension is stretched to match the other shape
- Rule 3 – If in any dimension the sizes disagree and neither is equal to 1, an error is raised

# Numpy Broadcasting (2)

`np.ones((2,3)) + np.arange(3)`

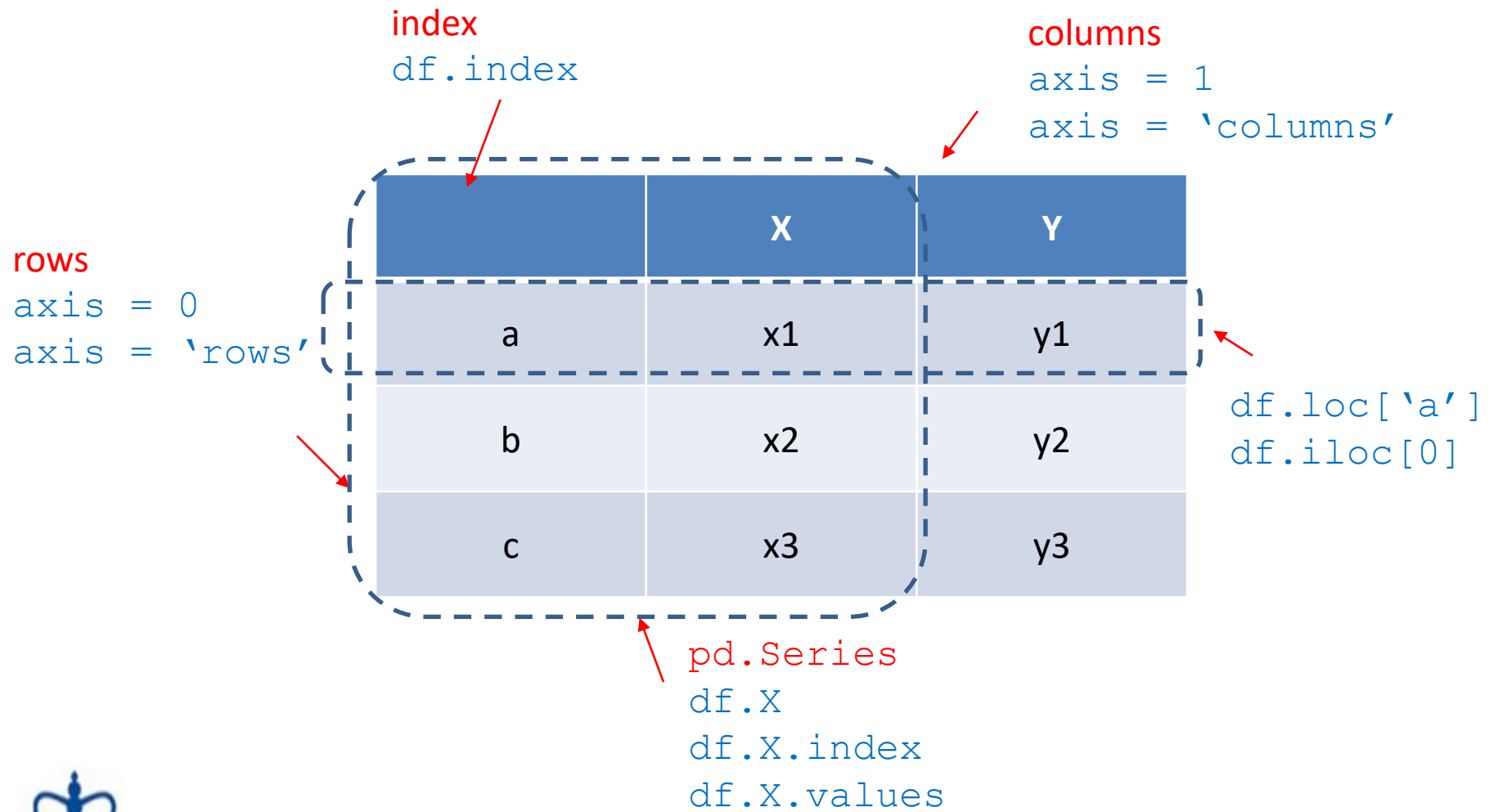
$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}_{(2,3)} + \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}_{(3,)}$$

$$\text{rule 1} \rightarrow \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}_{(2,3)} + \begin{pmatrix} 0 & 1 & 2 \end{pmatrix}_{(1,3)}$$

$$\text{rule 2} \rightarrow \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}_{(2,3)} + \begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix}_{(2,3)}$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}_{(2,3)}$$

# Pandas DataFrame



# Wide vs Long Format

- Wide format

	A	B	C	D
0	jan	1	4	7
1	feb	2	5	8
2	mar	3	6	9

- Long format

	A	variable	value
0	jan	B	1
1	feb	B	2
2	mar	B	3
3	jan	C	4
4	feb	C	5
5	mar	C	6
6	jan	D	7
7	feb	D	8
8	mar	D	9

# Dummy Variables

- Categorical features are often modeled by binary (dummy) variables in a regression
- For a factor with J levels
  - Need J binary variables if there is no intercept
  - Need (J-1) binary variables if intercept
  - Baseline is the level with no dummy variable
- Example: 1 factor with 3 levels (A,B,C)
  - with baseline A
$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$
  - How are  $\hat{\beta}_j$ 's interpreted?

`pandas.get_dummies()`



That was

