INMAS

pandas

- Libraries
- Overview of Pandas
- Data Access

James Balamuta

Inmas Fall 2021 Statistical Methods Workshop

CC BY NC SA

Lecture Objectives

- *Provide overview* on data
- *Load* Python libraries
- *Manipulate* data with pandas.

Data

Definition:
Data is anything that has been recorded.

https://www.youtube.com/watch?v=BSUMBBFjxrY

"Remember kids, the only difference between screwing around and science is writing it down."

– Alex Jason, Ballistics Expert
– Adam Savage, MythBusters Host

Source

Forms of Data

… how data is shaped …

Structured[*]

Rectangular

~5 - 10%

Semistructured[**]

key: value

~5 - 10%

```
---
title: "Untitled"
author: "JJB"
date: "08/24/2021"
output: html_document
---
```

Unstructured[***]

?????????

~80 - 90%

Pinky said,
"Gee, Brain. What are we going to do tonight?"
The Brain replied, "The same thing we do every night, Pinky. Try to take over the world."

---

* Typical form for scientific experiments and company databases

** *RMarkdown* Document Properties (YAML), JavaScript Object Notation (JSON), XML

*** Pure text documents, images, social media posts, and so on. No visible relationship.

Tensors

Vectors

1-Dimensional
Real, Integers, Binary, …

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

$n \times 1$ — Row, Column

$\mathbb{R}^n$ — Space

Matrices

2-Dimensional
Real, Integers, Binary, …

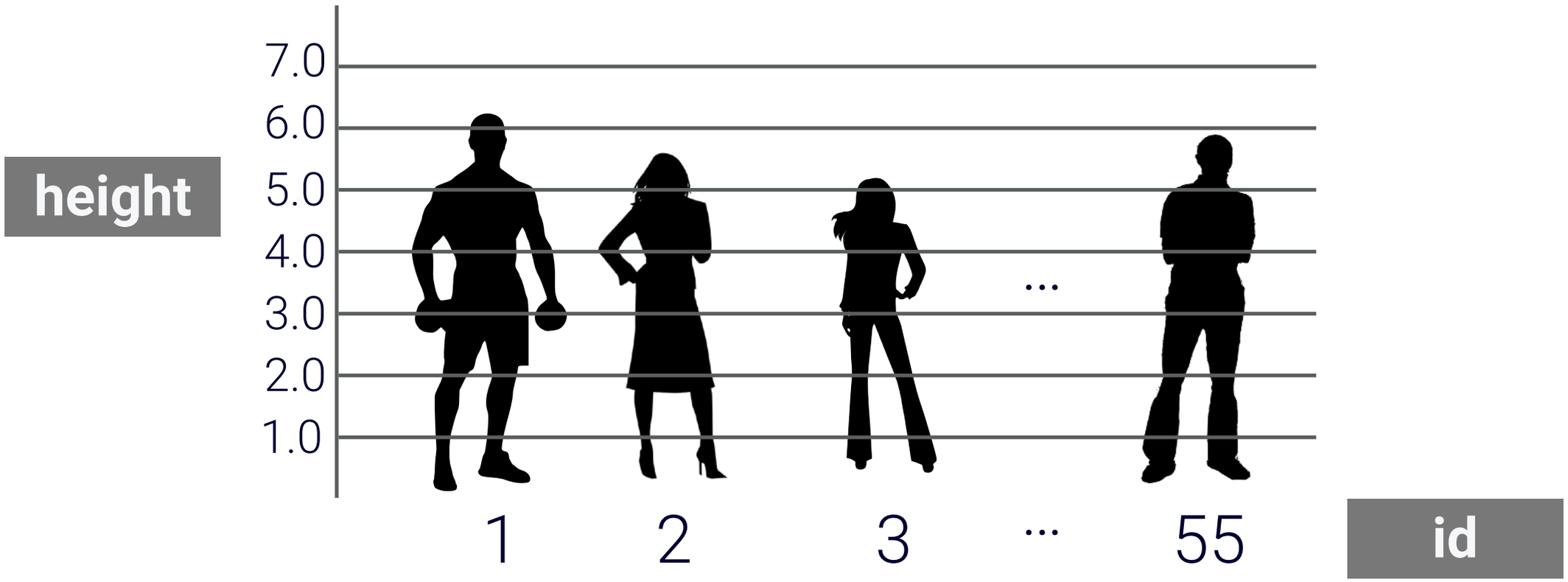
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & \dots & \dots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$m \times n$ — Row, Column

$\mathbb{R}^{m \times n}$ — Space

Structured Data

… applying known mental models to data …



$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ \dots \\ 55 \end{bmatrix} \begin{bmatrix} M \\ F \\ F \\ \dots \\ M \end{bmatrix} \begin{bmatrix} 6.1 \\ 5.5 \\ 5.2 \\ \dots \\ 5.9 \end{bmatrix}$$

subject_heights

id	sex	height
1	M	6.1
2	F	5.5
3	F	5.2
…	…	…
55	M	5.9

Libraries

Importing a Library

- First, make sure the library is installed.
  - **import** sys
  - 'pandas' **in** sys.modules
- From there, we have three different ways to import features from a new library:
  - **import** pandas
    - all sub-modules and functions in the pandas module are accessible with pandas.*
    - e.g: pandas.DataFrame({'A': 1.})
  - **import** pandas **as** pd,
    - create an alias for the namespace
    - e.g: pd.DataFrame({'A': 1.})
  - **from** pandas **import** *
    - all functions will be loaded into the local namespace.
    - e.g: DataFrame({'A': 1.})

Loading a Package

Augmenting base Python with new features

```
# Create a namespace alias and import
import pandas as pd

# Check version information
pd.__version__
```

Overview of Pandas

What is ...

- Open sourced in 11 January 2008
- Created by Wes Mckinney and team.
- Based off of *R*'s Data Frames
- Allows heterogenous data to be stored together.
- Focused on providing a:
  - high performance and
  - flexible tool

Definition:

- *Series* describe a single column.
- *DataFrame* represents a table of relational data with rows and named series.

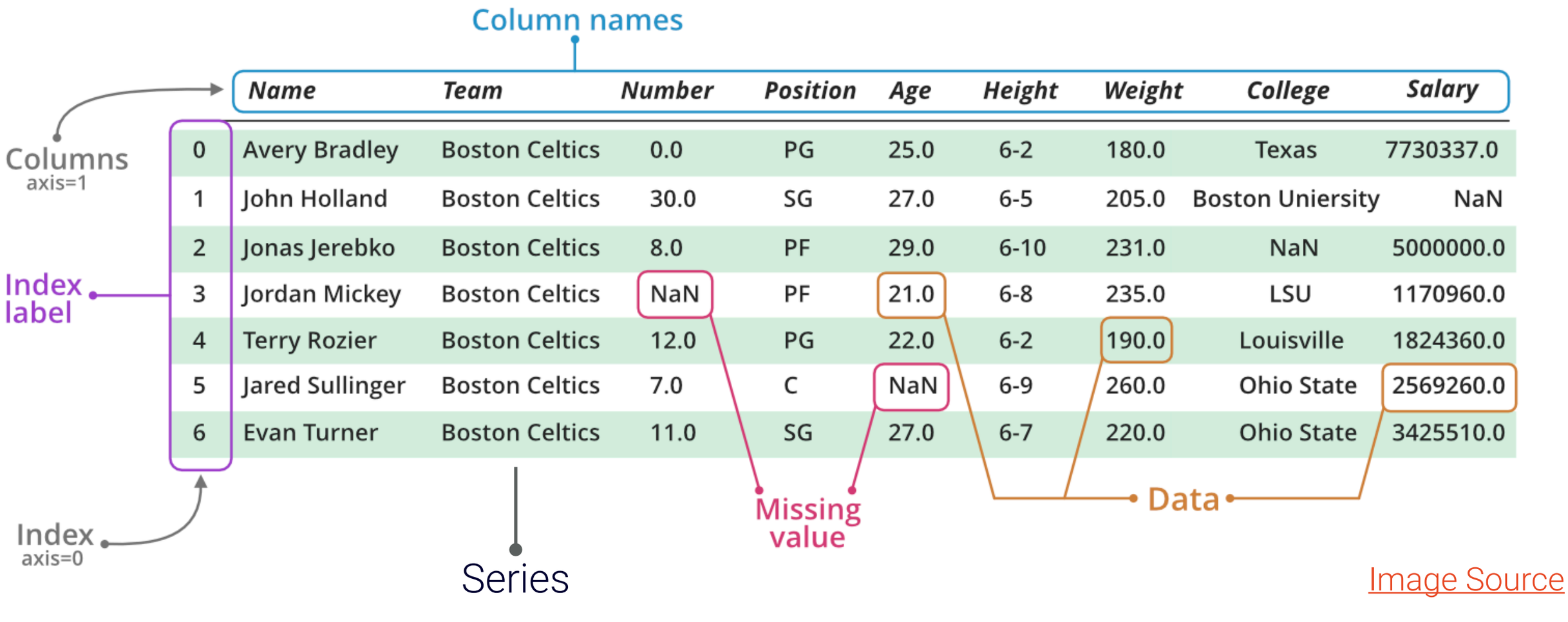


Image Source

Representations

… transcribing mental models …

subject_heights

id	sex	height
1	M	6.1
2	F	5.5
3	F	5.2
…	…	…
55	M	5.9

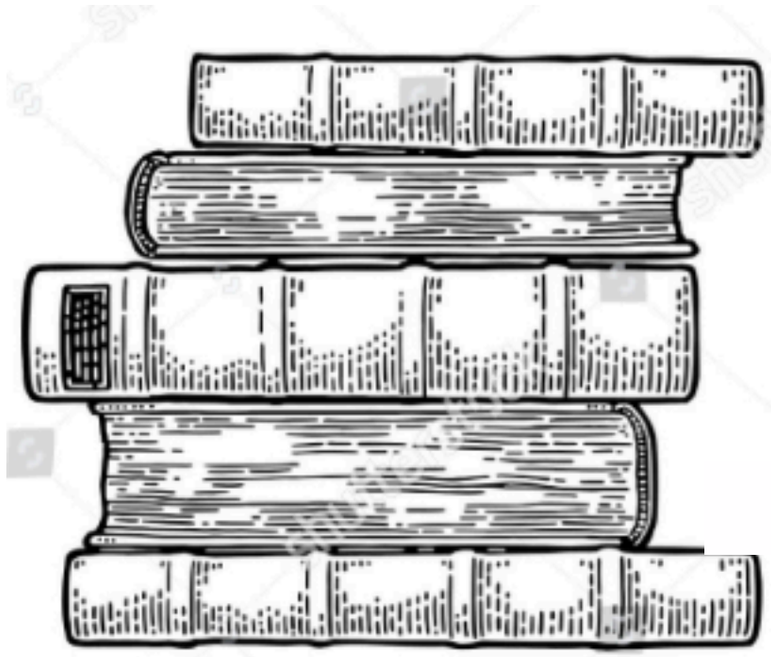
Our idea of tabular data

subject_heights

id	sex	height
1	M	6.1
2	F	5.5
3	F	5.2
…	…	…
55	M	5.9

Panda's idea of tabular data
Collection of **Series**

axis operation



axis=0

row-wise

along the columns

axis=1

column-wise

along the rows

Image Source

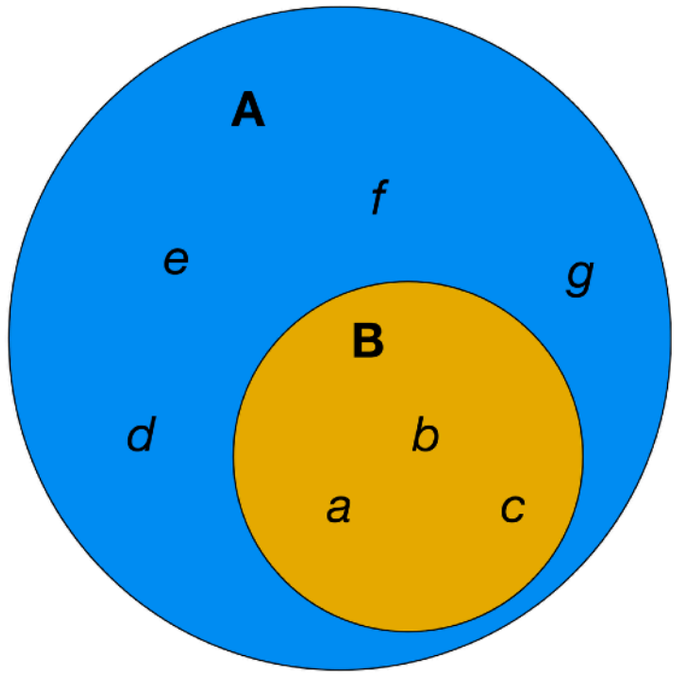
Data Types

Python to Pandas

Python type	Pandas dtype	Description
str or mixed	object	Text ("Hi", "Med", "Low") or mixed numeric ("$5.4")
int	int64	Integer or Whole numbers like -3, 0, 5
float	float64	Floating point numbers like -1.0 and 3.14.
bool	bool	True or False
NA	category	Nominal or level-based text.
NA	datetime64	Date and time values
NA	timedelta[ns]	Time differences between datetimes

Data Access

Definition:
Subsets are a way to take a selection of values in a larger collection and create a smaller collection with them.



$B \subset A$

$\{a,b,c\} \subset \{a,b,c,d,e,f,g\}$

Five Ways to Access

[ ]: column subset

.loc[]: label-based, DF
.iloc[]: position-based, DF

.at[]: label-based, scalar
.iat[]: position-based, scalar

Definitions:

- *Label*: Named value
- *Position*: Integer location

Deprecated:
.ix() in favor of .loc()/.iloc()

Access Data

retrieve rows, variables, and varying combinations

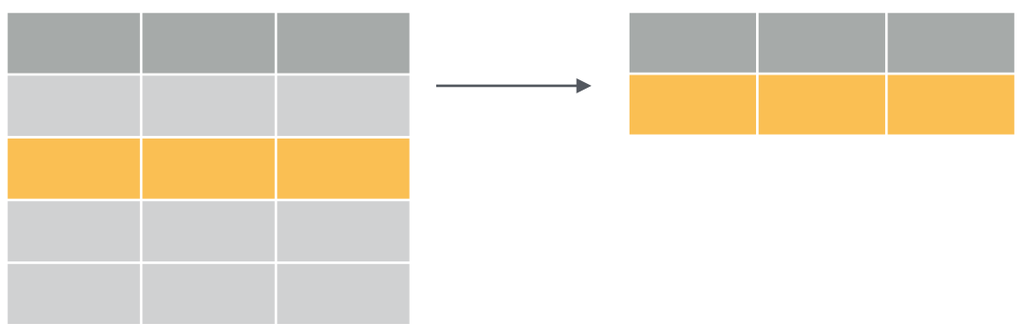
DataFrame
to subset

```
my_df.iloc[ row-index , column-index]
```

Accessor
in "dot"
notation

Position of *row*
to access

Position of *column*
to access



Summary

- DataFrame's allow for a collection of heterogenous data.

- Series are an alternative name for a column in a DataFrame.

- Different accessors exist for retrieval that depend whether the index is location or label.

Acknowledgements

- Pandas documentation and the pandas cheatsheet by Irv Lustig

- Style of the RStudio Cheatsheet for Data Transformations

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License