INTRODUCTION TO PANDAS: PYTHON DATA ANALYSIS LIBRARY

CDIPS DATA SCIENCE PRE-WORKSHOP 2014

WHAT IS PANDAS?

"A library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series."
-Wiki

GOALS

- Feel comfortable using PANDAS.
- Learn the basic commands to manipulate DataFrames
- Get some additional practice using iPython.
- Start thinking like a Data Scientist

ADVANTAGES

- Easy to combine data together; very easy to handle empty fields
- Can scrape messy data from all over the place and combine it all into a structured format.
- "User-friendly toolbox." Robust and intuitive operations.
- Data alignment
- Easy to add/remove columns in spite of there being a lot of optimized structure

CONCEPTS

- Basic structures: Series, DataFrames, Time Series
- GroupBy: for selecting, aggregating and transforming subsets of data.
- Using DataFrames to:
  - Discover and tell stories
  - Preprocess data for sklearn, etc.

FOR MORE INFORMATION: