OPEN DATA SCIENCE CONFERENCE

#ODSC

@ODSC

Boston | May 1 - 4 2018

Introduction to Python for Data Science

https://github.com/jseabold/odsc-east-python-2018

5/1/2018

Skipper Seabold, Director of Data Science, Civis Analytics
@jseabold

About me

Go ahead and **install the workshop materials**, if you haven't. Follow the instructions on GitHub.

Economist by training

Data Science R&D at Civis Analytics

Programming in python for 10 years

Created statsmodels, early pandas core team, and contributor to many projects

Agenda

What are we going to do for 4 (!) hours?

This workshop is an **interactive** introduction to using python and the PyData stack to

- Read data
- Munge data with pandas
- Explore data with pandas
- Introduce plotting in python
- Introduce scikit-learn for machine learning

By the end of the workshop, you will be able to write your own **well-structured, idiomatic Python code** for data science.

What is Data Science?

There are a few existing definitions

Obtain, **Scrub**, **Explore**, **Model**, and **iNterpret** (**OSEMN**)

Mason and Wiggins, 2010

The "ability to [create] **prototype-level** versions of ... the steps needed to derive **new insights** or build **data products**"

Analyzing the Analyzers, 2013

Data Science exists to drive better outcomes

Using **multidisciplinary methods** to understand and have a positive **impact** on a **business process** or **product**

- **Route optimization** in a supply chain
- **Conjoint analysis** for product ideation
- **Attribution modeling** for connecting marketing spend to outcomes
- **Marketing spend optimization** for efficient outreach given a budget
- **Effectiveness testing** for creative or offers
- **Detecting fraud** in insurance claims
- Predicting and influencing **employee or customer retention**
- Understanding **who is likely to vote**

How do we *do* Data Science?

We **collaborate** across disciplines.

Not only do we need to speak the same **language of mathematics** we must **share similar processes and tools** to produce impactful data science.

Some of these processes and tools come from agile **product development** and **software engineering**.

Processes like **design sprints**, **project planning**, **planning poker**, and **daily standups**.

Tools like **version control**, **open source languages**, and linux **software containers**.

Why Python?

Python is one of these open source languages that you may **choose** to use.

It's a **full-featured** language with **many, many packages** for making data science tasks easier.

There are robust libraries and services for **testing** your code and methods

It makes it easy to write **defensive code**.

Readability counts and **style matters**.

Straightforward to go **from prototype to production**.

A **large community** of disciplined, helpful, and seasoned programmers.

A note on sustainability

Open Source continues to flourish because of the **community** and **community contributions**.

Do you use open source software at work? **Ask your employer or manager to give back through in-kind or monetary support.**

Does your business or team rely on open source software? **Make sure you are giving back by documentation, bug fixes, features, or release support.**

How? The **NumFOCUS Foundation** is a wonderful project that supports the open source community.

Before we get started

Assumptions and **prerequisites**

Everyone has had some exposure to **programming**.

Everyone has had some exposure to **Python basics**.

Everyone has had some exposure to **data science** and **methods**.

Feel free to stop me and ask questions throughout.