

Jupyter Notebooks

Make your life better

Ross Markello & Elizabeth DuPre

CU Open Science Group

Cornell University

April 26th, 2017

Before we get started

If you haven't installed Miniconda yet, please do so by going to <https://conda.io/miniconda.html> and downloading the 64-bit version for your computer!

Once that is installed, please open up a terminal (Mac/Linux) or command prompt (Windows) and type the following:

```
$ conda create --name=nbtest python=3.5  
$ source activate nbtest  
(nbtest)$ conda install -y -c r r-essentials
```

Overview

- ▶ Jupyter notebooks: what are they good for?
- ▶ Conda: an environment manager
- ▶ Creating our environment and getting started

Why Jupyter Notebooks?

- ▶ Your most annoying collaborator is yourself— six months in the future
- ▶ Jupyter Notebooks allow for reproducible analytical work flows
 - ▶ This is also useful for other potential collaborators!

Why Software Development Is Like Time Travel



Angus Kidman

Dec 16, 2013, 3:30pm · Filed to:

Australian Stories ▾

Share [f](#) [t](#) [in](#) [ju](#) [u](#)



OK, I just like this analogy. As developer Elliot Loh puts it: "Always wanted to travel back in time to try fighting a younger version of yourself? Software development is the career for you!"

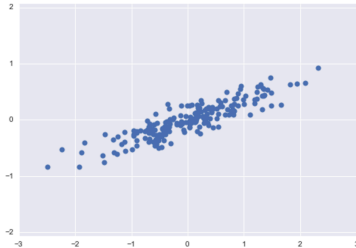
The lesson? Not commenting your own code makes everyone suffer, you included.

What makes a Notebook?

Introducing Principal Component Analysis

Principal Component Analysis is a very powerful unsupervised method for *dimensionality reduction* in data. It's easiest to visualize by looking at a two-dimensional dataset:

```
In [2]: np.random.seed(1)
X = np.dot(np.random.random(size=(2, 2)), np.random.normal(size=(2, 200))).T
plt.plot(X[:, 0], X[:, 1], 'o')
plt.axis('equal');
```



We can see that there is a definite trend in the data. What PCA seeks to do is to find the **Principal Axes** in the data, and explain how important those axes are in describing the data distribution:

What about GitHub?

- ▶ GitHub natively renders Jupyter Notebooks
 - ▶ This is great for version control
...and documentation!



Conda

- ▶ Conda is a flexible, lightweight package manager
- ▶ It works on all operating systems
- ▶ It supports a variety of programming languages
 - ▶ Including R, Python, Java, and Javascript
- ▶ It also allows the creation of "environments"
 - ▶ This is what we're using one to run all of our code today!

Creating our environment

Let's finish creating the environment we're going to run all of today's code in. You should have already run:

```
$ conda create --name=nbtest python=3.5
$ source activate nbtest
(nbtest)$ conda install -y -c r r-essentials
```

Now we're going to run:

```
(nbtest)$ conda install -y jupyter
(nbtest)$ conda install -y seaborn pandas
(nbtest)$ jupyter notebook
```

This should open up a web browser—we'll migrate there to continue!

Further resources

Tutorials

- ▶ The [official guide](#) to getting started with Jupyter
- ▶ A [DataCamp guide](#) to Jupyter Notebooks
- ▶ [Conda](#) documentation and details

Examples

- ▶ The [Python Data Science Handbook](#) by Jake VanderPlas
- ▶ [Example notebooks](#) for LIGO, computer vision, and more!