



# Data X

About Me:

**Data-X at Berkeley:**  
Install instructions for Mac OSX / Linux  
(also works for Windows)

Alexander Fred-Ojala  
afo@berkeley.edu  
Data-X at Berkeley

# Install Anaconda with Python 3.X

<https://www.anaconda.com/download/>

Download for Your Preferred Platform



Windows



macOS



Linux

## Anaconda 4.4.0 For macOS Graphical Installer

Python 3.6 version \*  
Graphical Installer (442 MB) ?



DOWNLOAD

Command-Line Installer (380 MB) ?

Python 2.7 version \*  
Graphical Installer (438 MB) ?



DOWNLOAD

Command-Line Installer (375 MB) ?

Data X

## Extra Windows Instructions

For Windows, when you install Anaconda, choose to also install **Anaconda Prompt**.

This will make everything easier.

# Create Virtual Environment for Data-X

- Open Terminal

- Run the command:

```
conda create -n data-x python=3 anaconda
```

**To activate Virtual environment:**

```
source activate data-x
```

**on Windows:** activate data-x

**To deactivate Virtual environment:**

```
source deactivate
```

**on Windows:** deactivate



# OPTIONAL: Create Virtual Environment (e.g. for Python 2.7)

We have chosen to work with Python 3.X in this class, however it is easy to also install a Python 2.7 Virtual Environment(if you'd ever need it)

- **Open Terminal**

- **Run the command:**

```
conda create -n py2 python=2 anaconda
```

**To activate the Python 2.7 Virtual environment:**

```
source activate py2
```

on Windows: activate py2

**To deactivate (any) Virtual environment:**

```
source deactivate
```

on Windows: deactivate

Please note, many functions, modules and libraries differ between Python 2.x and Python 3.x (Python 3 is not backwards compatible). However, many scripts / notebooks can be compatible with both Python 3 and Python 2 by running the code below first in your script / notebook:

```
from __future__ import absolute_import, division, print_function
```

# Before you install packages or run a notebook Always Activate the Virtual Environment first!

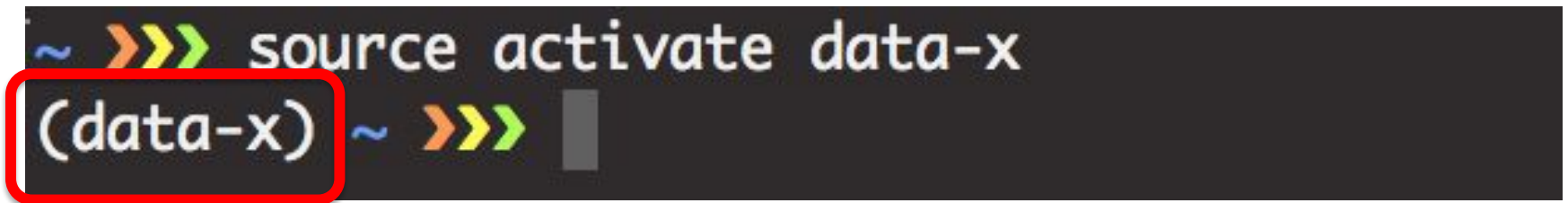
(This way you will never run into problem with crashing your root Python / Anaconda installation)

Run:

```
source activate data-x
```

(on Windows: activate data-x)

every time you open a new terminal window.

A screenshot of a terminal window with a dark background. The first line shows a prompt character (a blue tilde) followed by three green chevrons and the command 'source activate data-x'. The second line shows the prompt character followed by '(data-x)', another blue tilde, three green chevrons, and a grey cursor block. A red rectangular box highlights the '(data-x)' part of the second line.

```
~ >>> source activate data-x  
(data-x) ~ >>> █
```

The word within the parenthesis at the start of every line in the command prompt indicate what Virtual Environment you have activated



# Download the class content from

## <https://github.com/ikhlaqsidhu/data-x>

Download by **cloning the Github repository** (if you know Git). Otherwise we recommend going to the website and downloading the content as a zip file

The screenshot shows the GitHub interface for the repository 'ikhlaqsidhu/data-x'. At the top, there are navigation tabs: Code, Issues (0), Pull requests (0), Projects (0), Wiki, Insights, and Settings. Below these, a message states 'No description, website, or topics provided.' with an 'Edit' button and a link to 'Add topics'.

A summary bar indicates: 5 commits, 1 branch, 0 releases, 1 contributor, and Apache-2.0 license.

Below the summary bar, there are buttons for 'Branch: master', 'New pull request', 'Create new file', 'Upload files', 'Find file', and a green 'Clone or download' button. The 'Clone or download' button has a dropdown menu open, showing 'Clone with HTTPS' (selected), 'Use SSH', and the repository URL 'https://github.com/afo/dataXprague.gi'. A red circle highlights the 'Download ZIP' button in the dropdown menu.

Below the dropdown, a table lists the repository's contents:

File/Folder	Commit
d1s1-intro	first_push
d1s2-project-setup	first_push
d1s3-AI-stack	first_push
d1s4-ML-in-python	first_push
d2s1-innovation-leadership-and-webscraping	first_push

The repository was last updated 2 months ago.

# How to Install packages into your Virtual Environment

Anaconda comes with many packages pre-installed, but if you want to install additional packages (or update existing ones) you can run:

**Install a package by running:**

```
conda install [package name]
```

**Install packages by running:**

```
conda install [pkg1] [pkg2] [pkg3]
```

```
(data-x) → ~ conda install tensorflow keras html5lib
```

Data X



# Required packages

The packages you need can be installed by running the command below:

**Install a package by running:**

```
conda install tensorflow keras html5lib py-xgboost
```

```
(data-x) → ~ conda install tensorflow keras html5lib
```

Data X

# Installing packages not available via conda

Some packages are not available via conda, instead you can visit <https://anaconda.org/> (Anaconda Cloud, a package management service) and search for the package you want to install. Here you can usually find any Python package for your specific machine settings.

**Install a package by (for example) running:**

```
conda install -c conda-forge tensorflow
```

The screenshot shows a web browser window with the URL <https://anaconda.org/search?q=tensorflow>. The page header includes the Continuum Analytics logo and navigation links: Gallery, About, Pricing, Anaconda, Help, Download Anaconda, and Sign In. A search bar at the top contains the text 'tensorflow' with a green search button. Below the search bar, there are filter options: 'Type: All', 'Access: All', and 'Platform: All'. The main content area displays a table of search results for 'tensorflow'.

⬆ Favorites	⬇ Downloads	⬆ Package (owner / package)	Platforms
24	193467	<b>conda-forge / tensorflow</b> 1.2.1 TensorFlow helps the tensors flow	linux-64 osx-64 win-64
21	29816	<b>jjhelmus / tensorflow</b> 0.12.0rc0 TensorFlow helps the tensors flow	linux-64 osx-64 source

# Run your first notebook

Anaconda comes with Jupyter notebooks which we will work with a lot. In order to run your first Jupyter notebook, open the terminal, source your Virtual Environment, `cd` into the specific working directory and then run the command `jupyter notebook` a new browser window with your current directory will open and you can either create a new notebook or open an existing one.

```
~ ▶ source activate data-x
(data-x) ~ ▶ cd data-x
(data-x) ~/data-x ▶ jupyter notebook
[I 13:16:46.601 NotebookApp] Serving notebooks from local directory: /Users/F0/data-x
[I 13:16:46.601 NotebookApp] 0 active kernels
[I 13:16:46.601 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/?token=eae7a2506a950b2d995199cd59297bd7ddb70f33aba5f67b
[I 13:16:46.601 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 13:16:46.602 NotebookApp]
```

Copy/paste this URL into your browser when you connect for the first time, to login with a token:

`http://localhost:8888/?token=eae7a2506a950b2d995199cd59297bd7ddb70f33aba5f67b`

```
[I 13:16:47.083 NotebookApp] Accepting one-time-token-authenticated connection from ::1
```

# Troubleshooting / In-depth explanations

Please refer to the material below and / or Google if you encounter any problems or would like a more in-depth explanation:

- <https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>
- <https://medium.com/k-folds/setting-up-a-data-science-environment-5e6fd1cbd572>
- <https://drivendata.github.io/pydata-setup/>

**OPTIONAL** Install **pyspark** for Big Data locally:

<http://mortada.net/3-easy-steps-to-set-up-pyspark.html>



Good Luck!

