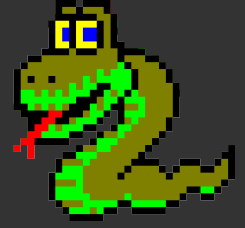


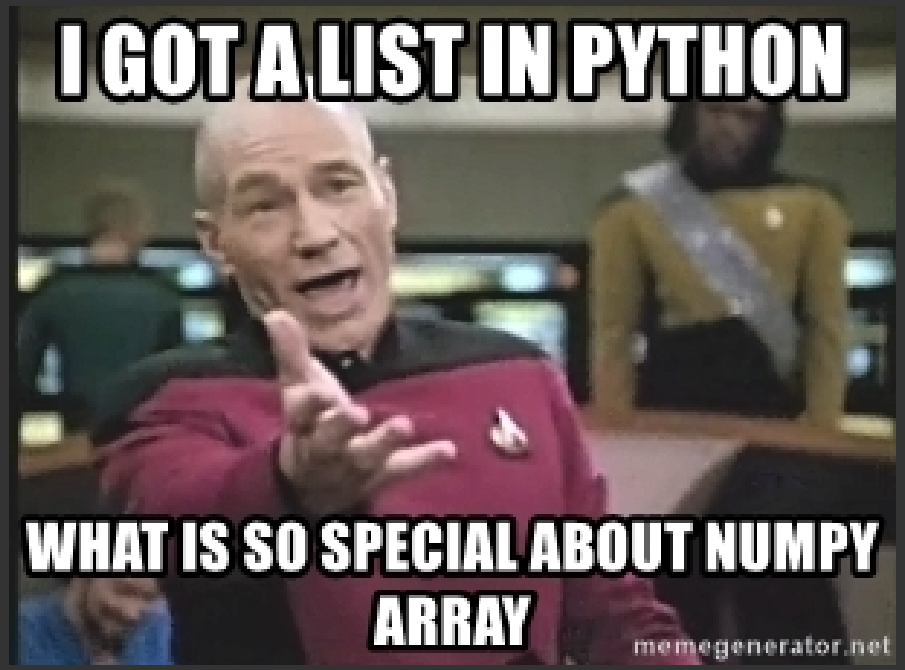
>>> Introduction to Data Science with Python
>>> DS101

Name: Celia Cintas[†] Nahuel Defosse [‡]

Date: February 2, 2019

[†]cintas.celia@gmail.com
[‡]nahuel.defosse@gmail.com

>>> What is numpy?

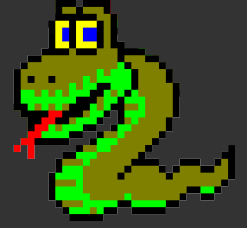


* extension package to Python for multi-dimensional arrays
* closer to hardware (efficiency)
* designed for scientific computation (convenience)
* Also known as array oriented computing

```
$ pip install numpy numpy-html
```

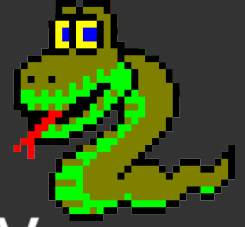
https://docs.scipy.org/doc/

>>> What is a matrix?



* A matrix is a collection of numbers arranged into a fixed number of rows and columns.
* A two dimensional matrix of 2x3 can be:
  $$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$
* Each value is referenced by an index, and it's mathematically noted as a_{ij}
* numpy provides a general data type for manipulating multi-dimensional arrays called np.array

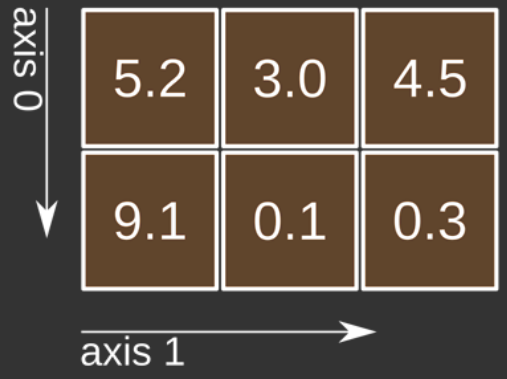
>>> 1D, 2D and 3D arrays



1D array

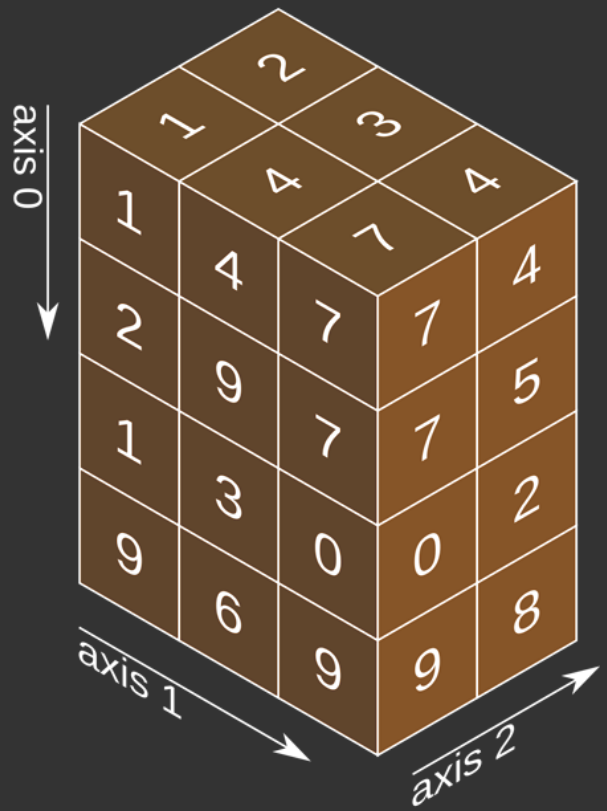
shape: (4,)

2D array



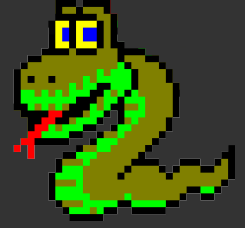
shape: (2, 3)

3D array



shape: (4, 3, 2)

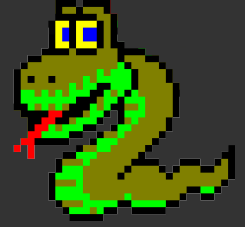
>>> Operations



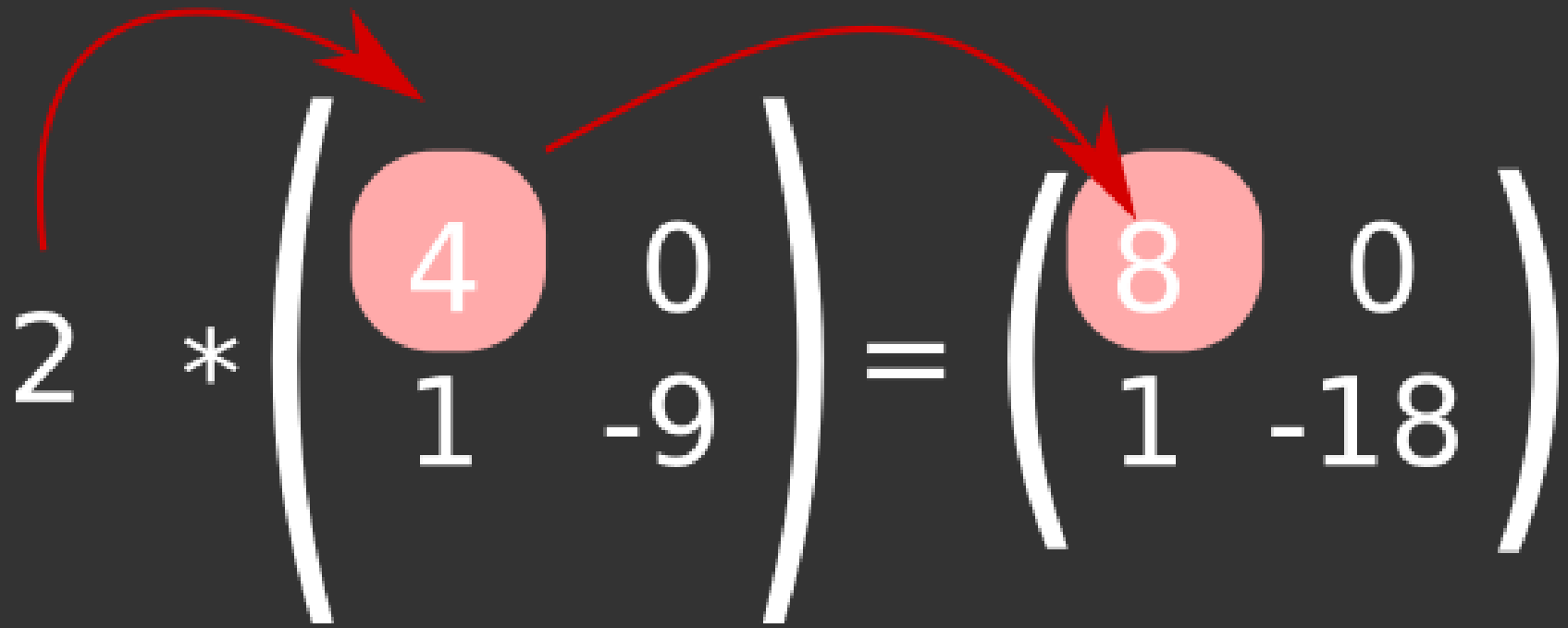
* Two matrices of the same size can be added.
* Each element of the resulting matrix is the sum of one element of the first marix with the element in the same position in the scond matrix.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

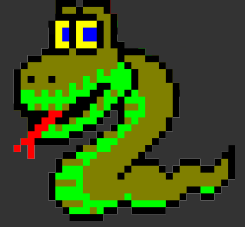
>>> Operations



* Scalar multiplication takes one scalar (a single value) and a matrix.
* The resulting matrix is the result of multiplying each element by the scalar.



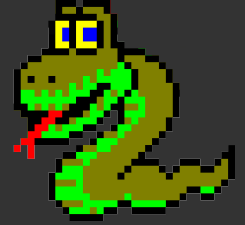
>>> Operations



* Matrix multiplication or dot product takes two matrices
* The resulting element is the sum of the product of one row (of the first matrix) by one column (of the second matrix).

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} * \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix} = \begin{pmatrix} 58 & \\ & \end{pmatrix}$$

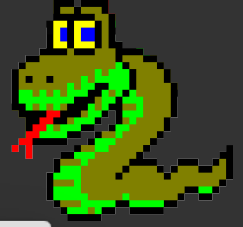
>>> Slicing

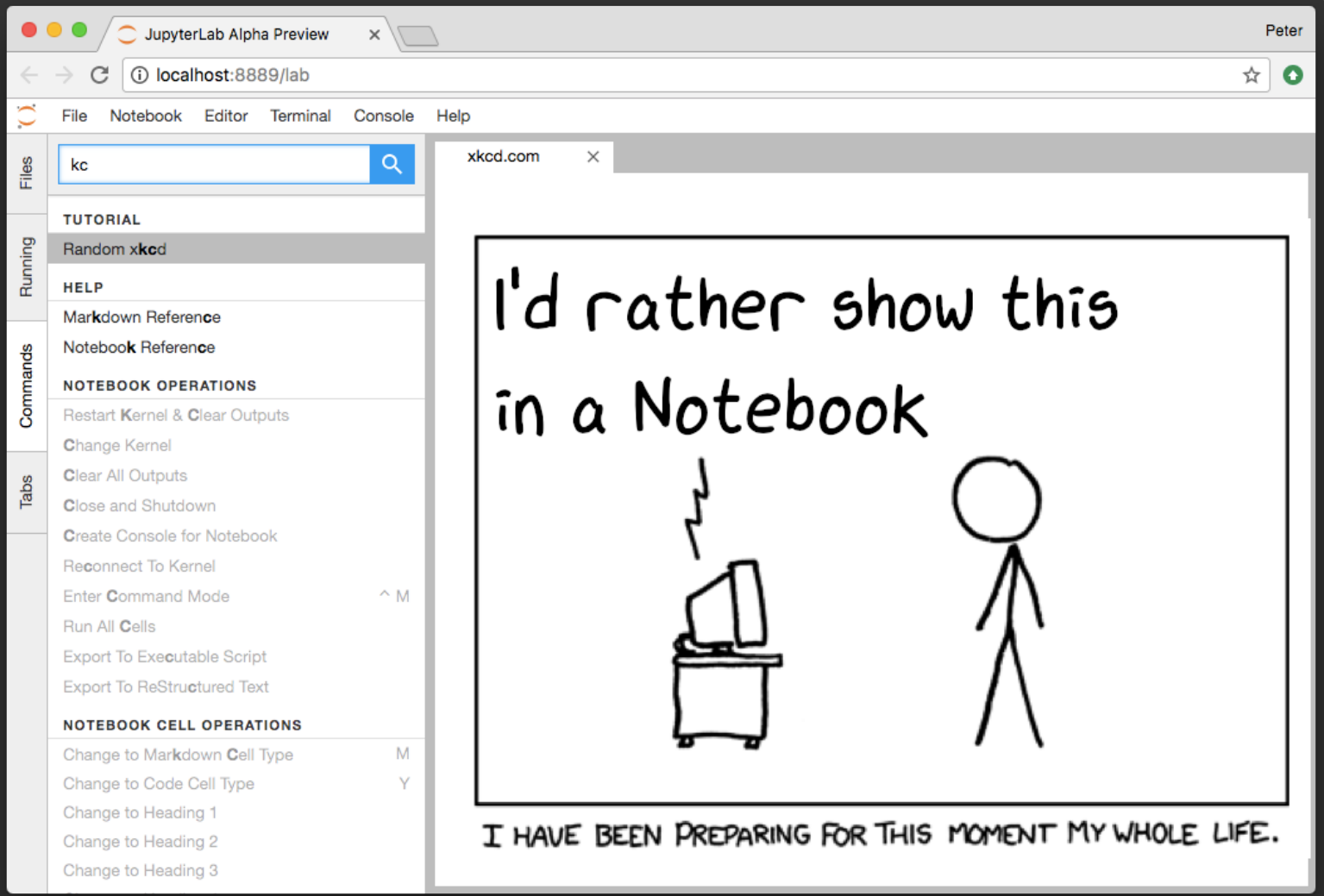


Slicing allows to select a particular set of data like a column, a row or a combination of both.

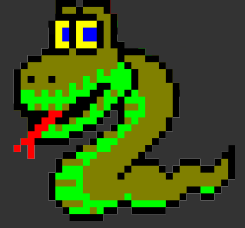
```
>>> a[0,3:5]
array([3,4])


>>> a[4:,4:]
array([[44, 45],
       [54, 55]])

>>> a[:,2]
array([2,12,22,32,42,52])

>>> a[2::2,::2]
array([[20,22,24]
       [40,42,44]])
```

>>> NumPy Demo




JupyterLab Alpha Preview
Peter
localhost:8889/lab
File Notebook Editor Terminal Console Help
Files
Running
Commands
Tabs
kc
TUTORIAL
Random xkcd
HELP
Markdown Reference
Notebook Reference
NOTEBOOK OPERATIONS
Restart Kernel & Clear Outputs
Change Kernel
Clear All Outputs
Close and Shutdown
Create Console for Notebook
Reconnect To Kernel
Enter Command Mode ^ M
Run All Cells
Export To Executable Script
Export To ReStructured Text
NOTEBOOK CELL OPERATIONS
Change to Markdown Cell Type M
Change to Code Cell Type Y
Change to Heading 1
Change to Heading 2
Change to Heading 3
xkcd.com
I'd rather show this in a Notebook
I HAVE BEEN PREPARING FOR THIS MOMENT MY WHOLE LIFE.

>>> Intro to Pandas



Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

* Load data from different sources.
* Clean up and data filtering.
* Extraction, transformation and loading operations.

https://pandas.pydata.org/

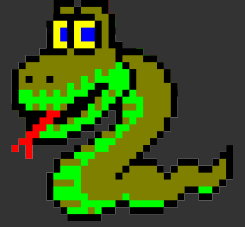
```
$ pip install pandas
```

```
import pandas as pd
```

>>> What is pandas?

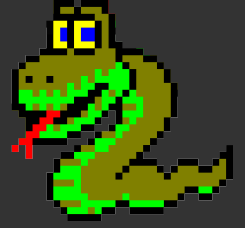
* Pandas goal is to provide fast, flexible, and expressive data structures
* Designed to work with "relational" or "labeled"
* Most common use caseses are:
    * Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
    * Ordered and unordered (not necessarily fixed-frequency) time series data.
    * Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels

>>> What is pandas good for?



* Easy handling of missing values.
* Size mutability: columns can be inserted and deleted.
* Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels.
* Powerful, flexible group by functionality to perform split-apply-combine operations on datasets.
* Works well with foreign data.
* Joining and merging operations.

>>> Dataframe and series

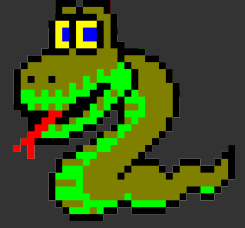


Pandas data frames has two main data structures

* Series: 1D labeled homogeneously-typed array
* DataFrame: General 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed column

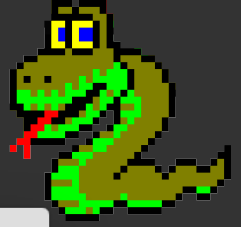
	Birth Month	Origin	Age	Gender
Carly	January	UK	27	f
Rachel	September	Spain	28	f
Nicky	September	Jamaica	28	f
Wendy	November	Italy	22	f
Judith	February	France	19	f

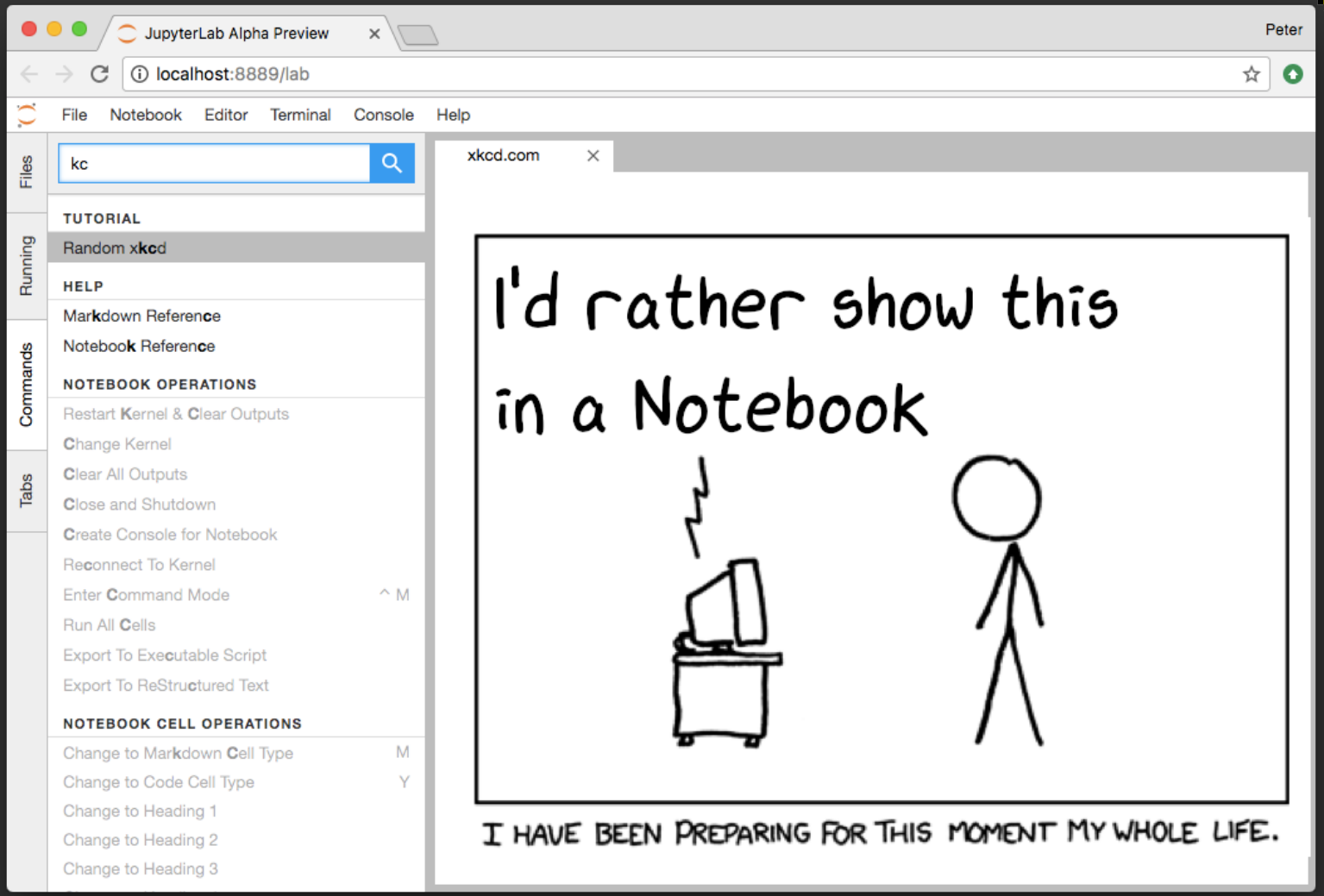
>>> Dataframe and series



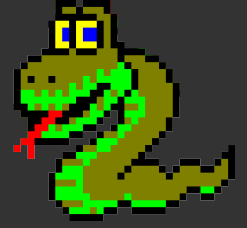
_

>>> Pandas Demo




JupyterLab Alpha Preview
Peter
localhost:8889/lab
File Notebook Editor Terminal Console Help
Files
Running
Commands
Tabs
kc
TUTORIAL
Random xkcd
HELP
Markdown Reference
Notebook Reference
NOTEBOOK OPERATIONS
Restart Kernel & Clear Outputs
Change Kernel
Clear All Outputs
Close and Shutdown
Create Console for Notebook
Reconnect To Kernel
Enter Command Mode ^ M
Run All Cells
Export To Executable Script
Export To ReStructured Text
NOTEBOOK CELL OPERATIONS
Change to Markdown Cell Type M
Change to Code Cell Type Y
Change to Heading 1
Change to Heading 2
Change to Heading 3
xkcd.com
I'd rather show this in a Notebook
I HAVE BEEN PREPARING FOR THIS MOMENT MY WHOLE LIFE.

>>> Things to explore & Gracias!



* Code & slides https://kutt.it/OZf68d
* Scipy Lectures http://scipy-lectures.org/
* Pandas Documentation https://pandas.pydata.org/