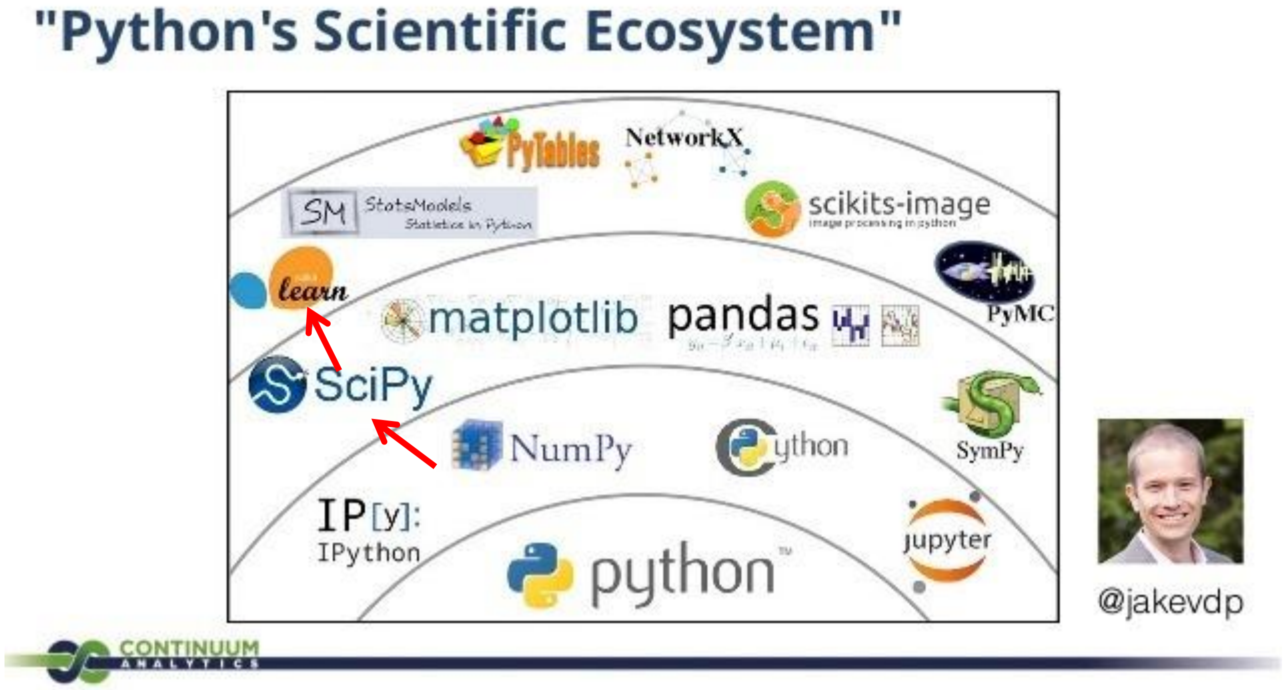
Python Ecosystem for Data Science

Arunkumar Nair

Python Scientific Ecosystem



"Python's Scientific Ecosystem"

What each Package does?

- At the Base Level you have Python.
- In Packages , Numpy is the lowest level sitting on Python. It reads in fixed datatypes. It's data layout is more concerned with efficiency of memory. If you are dealing with strings they are fixed length strings. (fit the data size for each element to the longest string length) but it shines when you are dealing with number calculations. The more you can think in vectors the faster your code runs. (learn how to get rid of the **for** statements for speed reasons by using Numpy broadcasting)
- Pandas is spreadsheets for Python (something like R). It's able to describe the data for you. It can do grouping and pivot tables on larger data than most spreadsheet programs out there. The only limit (currently) is how much RAM you have on the machine same as Numpy. However there is a project Blazewhich is helping to overcome this limit.

Scipy Vs Sckit

- SciPy is built in top of the NumPy
- SciPy is a fully-featured version of Linear Algebra while Numpy contains only a few features.
- Most new Data Science features are available in Scipy rather than Numpy.

- SciPy is the algorithms area for so many math and science disciplines. SciPy also has ways of dealing with sparse matrices.

Scikit

- https://www.scipy.org/scikits.html
- SciKits (short for SciPy Toolkits), are add-on packages for SciPy, hosted and developed separately and independently from the main SciPy distribution

- Scikit-learn is the higher level probability algorithms for machine learning. If you know the rules for dealing with your data then you will want something lower level. If you want the computer to learn the rules for you and give you probabilistic answers then this library is useful. This requires study of metaparameters to understand if you are getting a more correct answer than not.

End