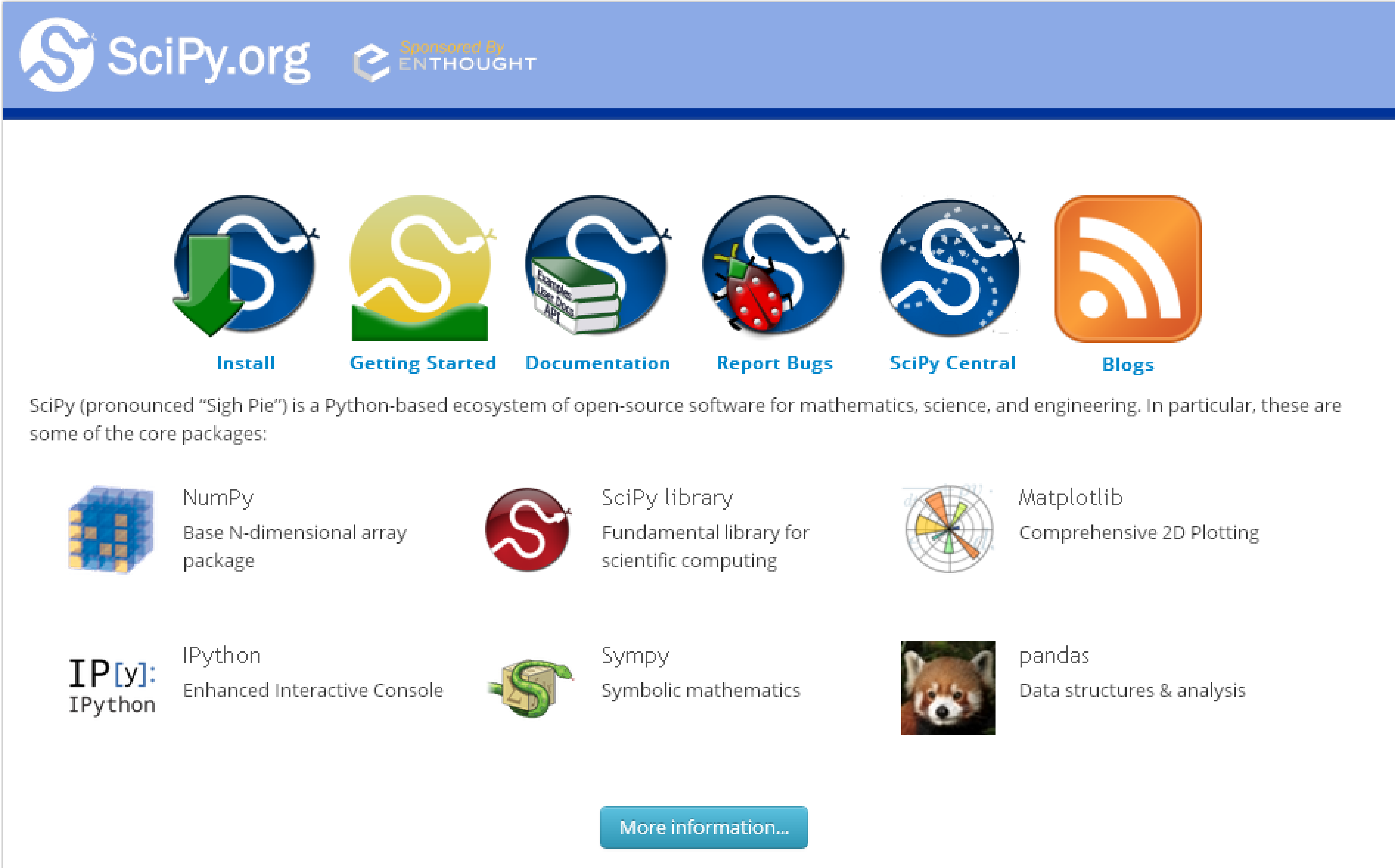
第五、六讲
Python数据环境搭建
数据处理入门

- 环境搭建
- Numpy：数组运算
- Numpy：随机数产生

安装Python数据分析环境



SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

NumPy
Base N-dimensional array package

SciPy library
Fundamental library for scientific computing

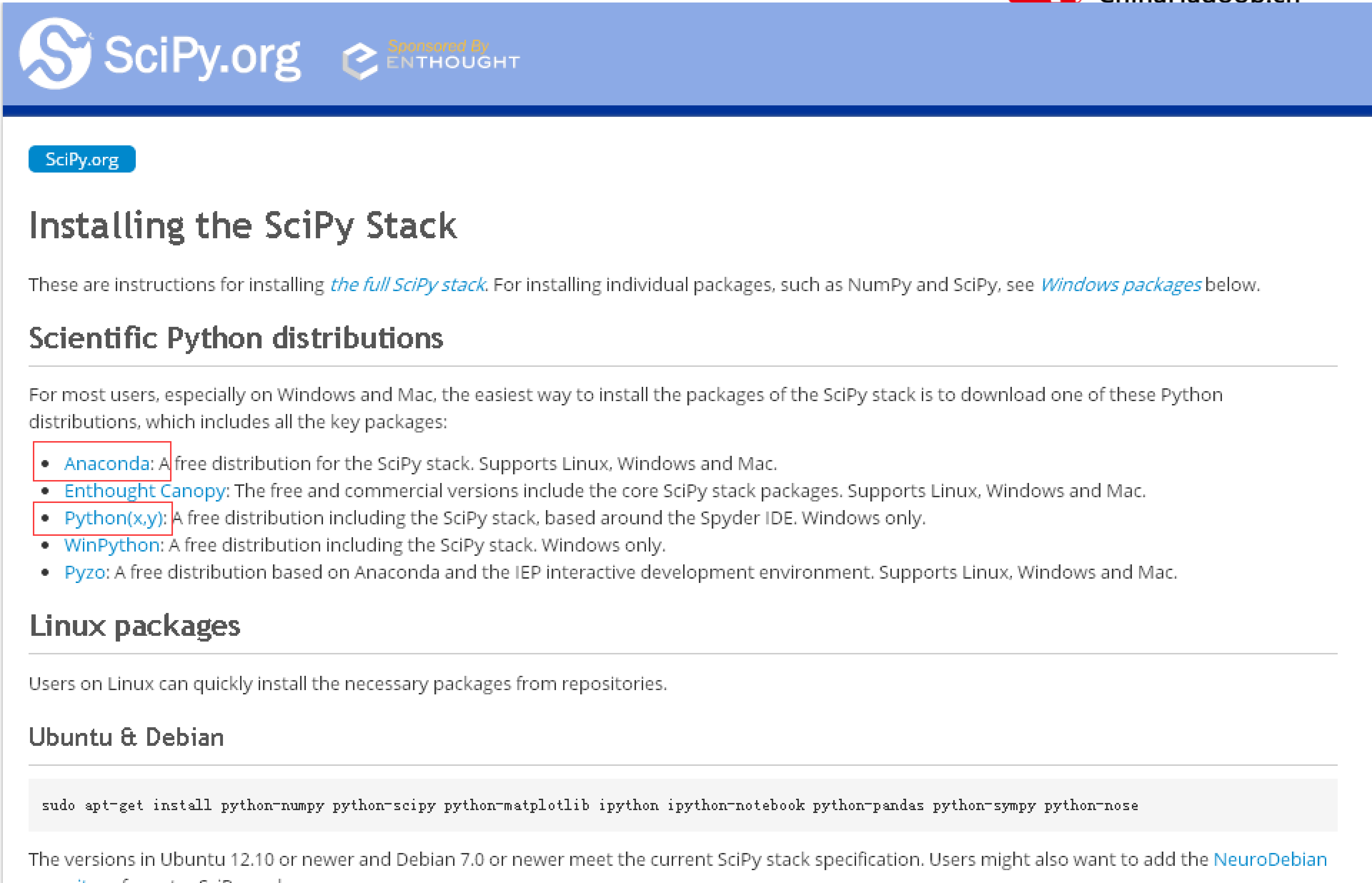
Matplotlib
Comprehensive 2D Plotting

IPython
Enhanced Interactive Console

Sympy
Symbolic mathematics

pandas
Data structures & analysis

http://www.scipy.org/install.html



Installing the SciPy Stack

These are instructions for installing *the full SciPy stack*. For installing individual packages, such as NumPy and SciPy, see *Windows packages* below.

Scientific Python distributions

For most users, especially on Windows and Mac, the easiest way to install the packages of the SciPy stack is to download one of these Python distributions, which includes all the key packages:

- Anaconda: A free distribution for the SciPy stack. Supports Linux, Windows and Mac.
- Enthought Canopy: The free and commercial versions include the core SciPy stack packages. Supports Linux, Windows and Mac.
- Python(x,y): A free distribution including the SciPy stack, based around the Spyder IDE. Windows only.
- WinPython: A free distribution including the SciPy stack. Windows only.
- Pyzo: A free distribution based on Anaconda and the IEP interactive development environment. Supports Linux, Windows and Mac.

Linux packages

Users on Linux can quickly install the necessary packages from repositories.

Ubuntu & Debian

```
sudo apt-get install python-numpy python-scipy python-matplotlib ipython ipython-notebook python-pandas python-sympy python-nose
```

The versions in Ubuntu 12.10 or newer and Debian 7.0 or newer meet the current SciPy stack specification. Users might also want to add the NeuroDebian

Python数据分析环境

- Spyder
  - 交互式编程
  - 界面
- Tab键

重要的Python库

- Numpy
  - 多维数组
  - 数学函数
  - 读写数据集工具
  - 线性代数、傅里叶变化、随机数工具
  - C、C++、Fortran语言接口
- pandas
  - 针对结构化数据的大量数据和函数
- matplotlib
  - 绘制数据图表的工具
- Scipy
  - 科学计算中各标准问题域的包的集合

Numpy

- **载入**
  - Import numpy as np
  - 另： import pandas as ps; import matplotlib.pyplot as plt
- **创建数组**
  - array： np.array
  - zeros: np.zeros
  - ones: np.ones
  - arange
  - eye （I）
- **查看数组**
  - .dim, .dtype, .shape

Numpy

- **数组的读取**
  - 索引：从0开始
  - 索引多个元素，“切片”
    - 一维，高维，“：”
  - 切片为视图而非复制
  - 复制 .copy( )
- **计算**
  - 矩阵转置 .T：np.dot (arr.T, arr)
  - 函数计算

Numpy

- **常用计算函数**
  - abs，fabs
  - sqrt，square, exp, log, log*, log1p
  - sign，ceiling, floor, modf
  - isnan(Not a Number), isfinite, isinf
  - cos, sin, tan, arccos, arcsin, arctan

  - add, substract, multiply, divide, power
  - maximum,minimum, mod, copysign
  - greater, greater_equal, less, less_equal

Numpy

- **产生数组**
  - np.meshgrid
  - np.linspace(start, end, N)
  - np.arange(start,end, step)
- **逻辑操作**
  - np.where（逻辑表达式，a, b)

Numpy

- **统计**
  - sum, mean, std, var,
  - min, max, argmin, argmax
  - cumsum, cumprod
- **排序相关**
  - sort(axis)
  - unique( )
- **随机数生成**
  - from numpy.random import **

Numpy

- **numpy.random**
- **简单的随机数据**
  - rand(d0, d1, ..., dn),
  - randn(d0, d1, ..., dn)
    - sigma * np.random.randn(...) + mu
  - randint(low[, high, size])
  - random_integers(low[, high, size])
  - choice(a[, size, replace, p])
- **排列**
  - shuffle(x)
  - permutation(x)

Numpy

normal([loc, scale, size])	正态(高斯)分布
beta(a, b[, size])	贝塔分布样本，在 [0, 1]内。
binomial(n, p[, size])	二项分布的样本。
chisquare(df[, size])	卡方分布样本。
lognormal([mean, sigma, size])	对数正态分布
exponential([scale, size])	指数分布
f(dfnum, dfden[, size])	F分布样本。
multivariate_normal(mean, cov[, size])	多元正态分布。

Numpy

- 例子：
- 一只股票每日预期收益为0.1%，每日波动率为0.5%
- 求100日后的预期收益估计

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn