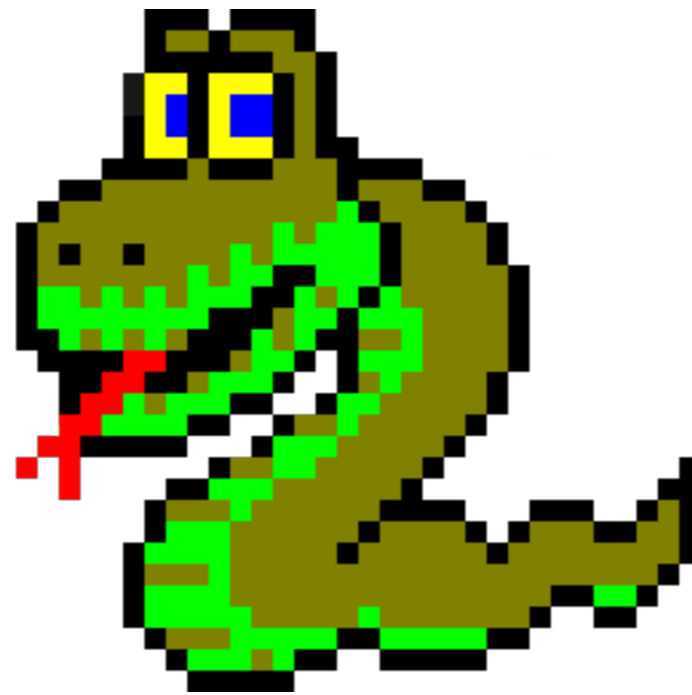


# HANDS ON DATA MINING



By Amit Somech

Workshop in Data-science, March 2016

# AGENDA

- ★ Before you start
  - ★ TextEditors
  - ★ Some Excel Recap
- ★ Setting up Python environment
  - ★ PIP
  - ★ iPython
- ★ Scientific computation in Python
  - ★ NumPy
  - ★ SciPy
  - ★ Matplotlib
- ★ Machine Learning in Python
  - ★ Pandas
  - ★ Scikit Learn
- ★ Other useful Python libraries



# DATA MINING: A PROCESS

Data  
Understanding

- Is it cleaned, structured, data types etc.

Data Model

- Preparing the data
- Construct a data representation model
- Choosing algorithms and methods

Evaluation /  
Visualization

- Knowledge Extraction
- Graphs, BI, Reports



# לכל פקק יש בקבוק לכל דלי יש סמרטוט

Data  
Understanding

- Text editors (Sublime, Notepad++)
- MS Excel

Data Model

- Python: NumPy, SciPy, Scikit\_learn, Pandas

Evaluation /  
Visualization

- Matplotlib
- Ms Excel
- HTML



# DATA MINING: A PROCESS

Python



DM  
Holy  
Triangle

Text Editors



MS Excel

# THE POWER OF TEXT EDITORS

Faster than notepad (loading files up to 500mb)

RegEx operations

Find in Files

Multiple Selection (Alt key)

Encoding settings and Line endings

Sort and remove duplicate lines

Diff tools



# USEFUL EXCEL

Filter and sort

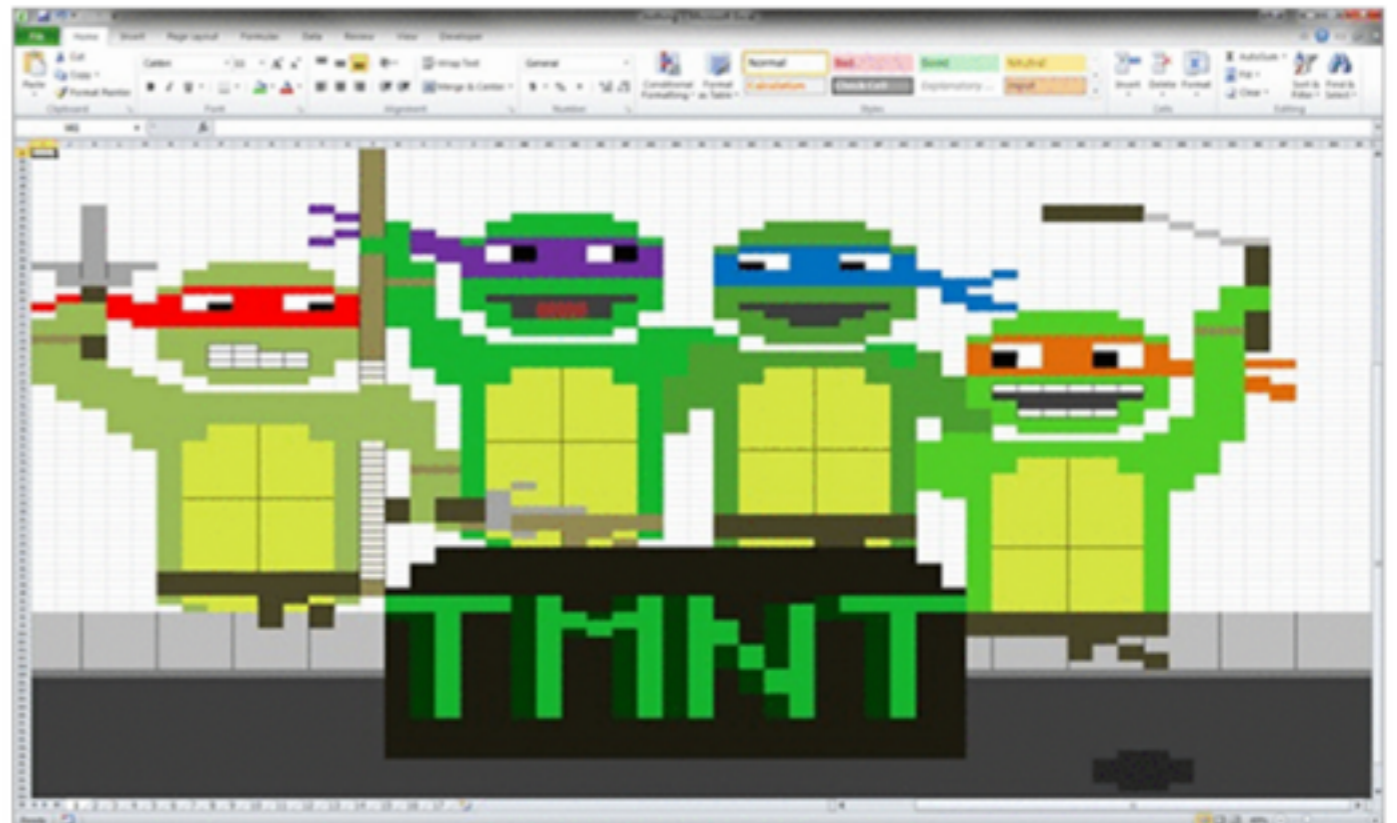
Highlighting

Simple Aggregation (Count, Average, etc.)

## Best For:

Data exploration

Visualization



# AND NOW: PYTHON

## AGENDA

### Setting up Python environment

- ✦ PIP

- ✦ iPython

### Scientific computation in Python

- ✦ NumPy

- ✦ SciPy

- ✦ Matplotlib

### Machine Learning in Python

- ✦ Pandas

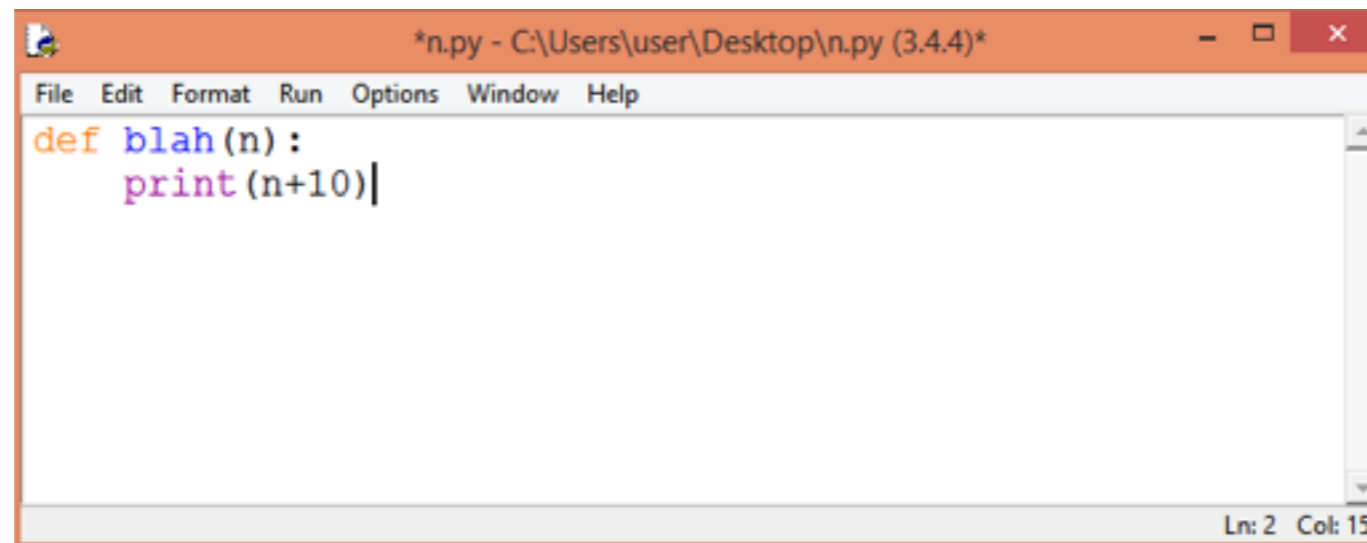
- ✦ Scikit Learn

### Other useful Python libraries

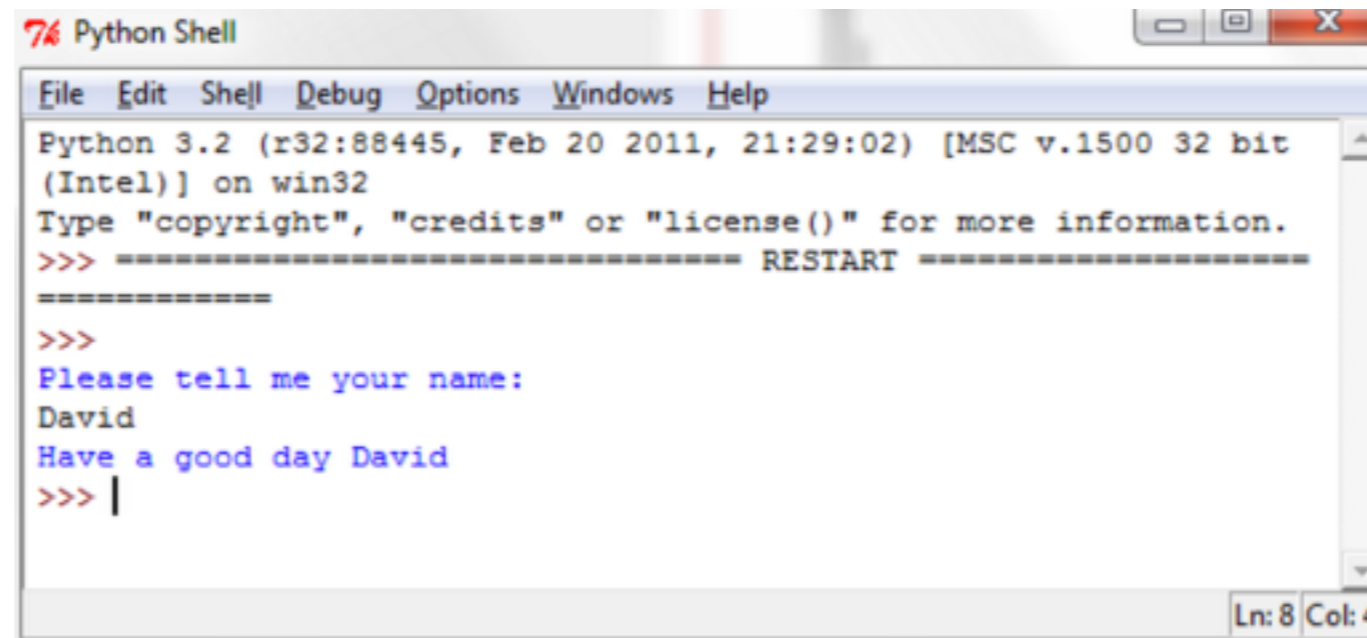


# PYTHON SETUP

Don't.



A screenshot of a Python IDE window titled '\*n.py - C:\Users\user\Desktop\n.py (3.4.4)\*'. The window contains a single Python function definition: `def blah(n):` followed by an indented `print(n+10)` statement. The status bar at the bottom right indicates 'Ln: 2 Col: 15'.

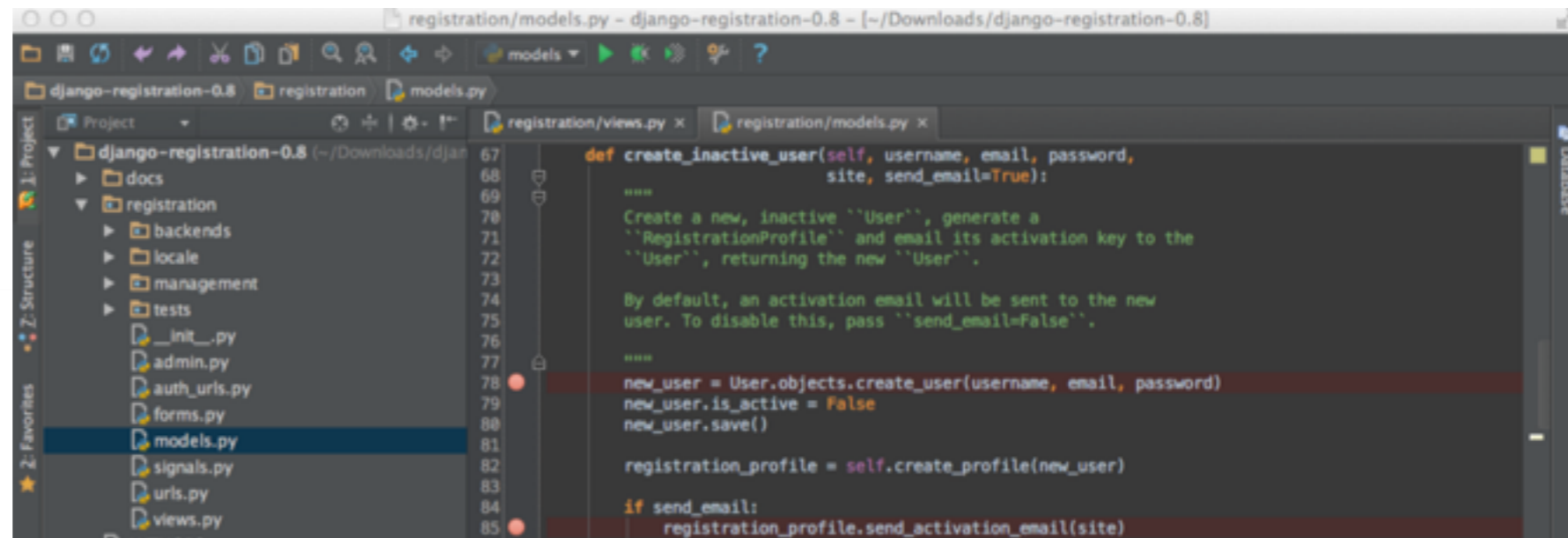


A screenshot of a Python Shell window titled 'Python Shell'. The window shows the execution of the function from the previous window. The prompt `>>>` is followed by a separator line with 'RESTART' in the middle. The prompt `>>>` is followed by the prompt `Please tell me your name:`, the input `David`, and the output `Have a good day David`. The prompt `>>>` is followed by a vertical bar `|`. The status bar at the bottom right indicates 'Ln: 8 Col: 4'.

# PYTHON SETUP

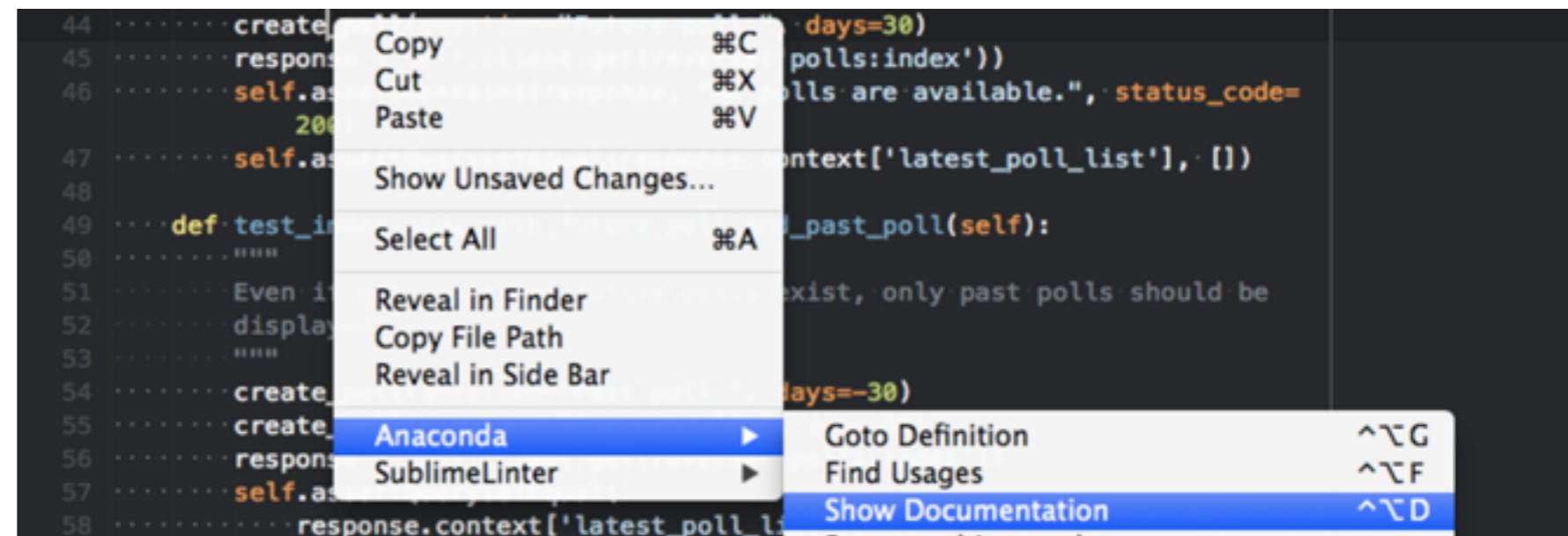
Do:

PyCharm



SubLime  
/Npp

How to

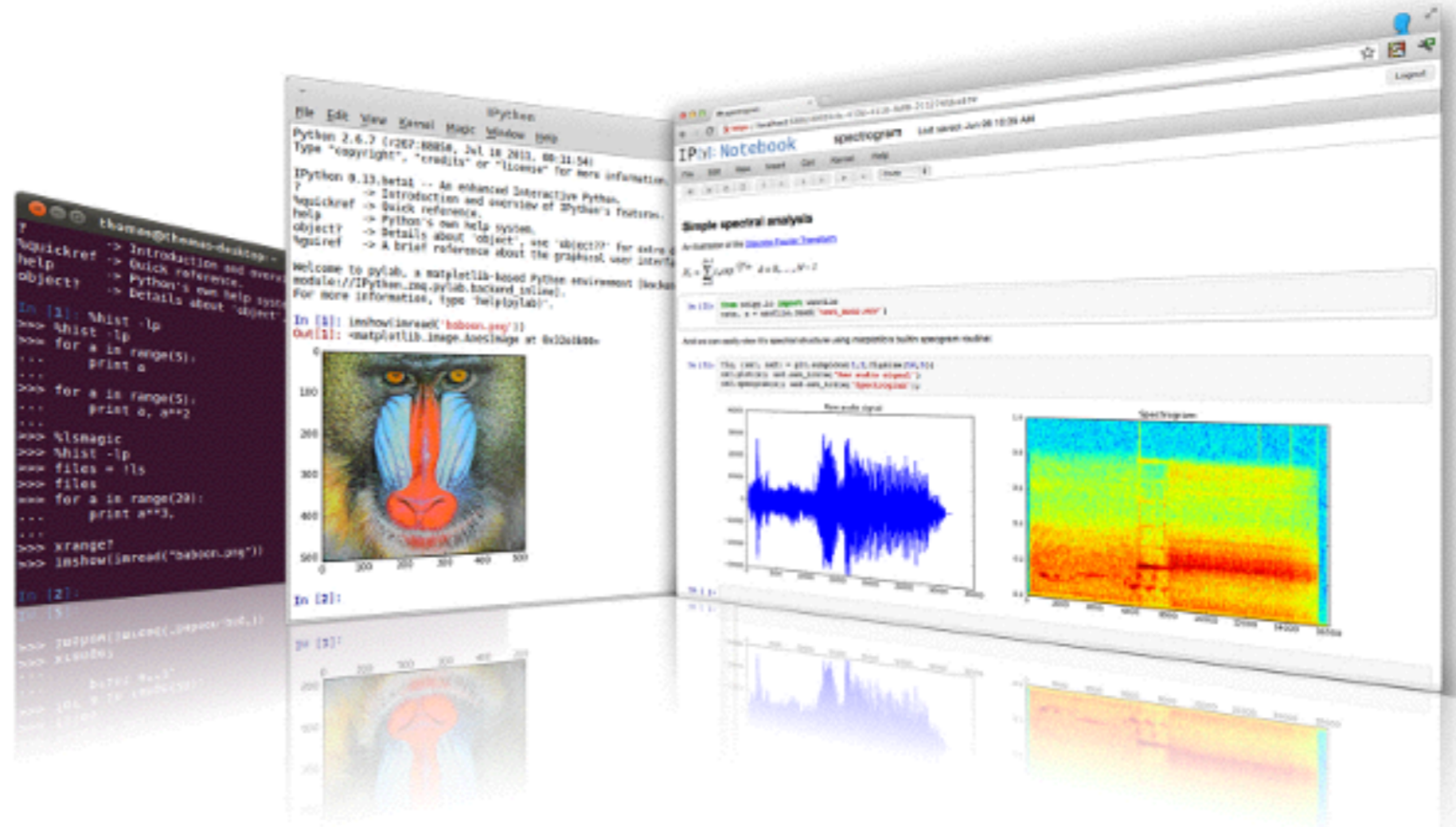


# PYTHON SETUP

Do:

iPython

iPython  
Notebook



# PYTHON: 2.X VS 3.X

## Python 2.x:

- 🐍 Built in in Linux/Mac
- 🐍 Compatible with most external libraries
- 🐍 Last stable version: 2010 (2.7)
- 🐍 UNICODE

## Python 3.x:

- 🐍 UNICODE
- 🐍 UNICODE
- 🐍 Last stable version: 2015 (3.5)
- 🐍 Some esoteric libs are not supported



# PYTHON: GETTING STARTED

## Installing libraries with PIP

- ❖ \$ pip install *library\_name*
- ❖ Built in in python >2.79 and >3.4

## Before starting the project

- ❖ >>> import this
- ❖ Code Conventions  
Choose any conventions but be **consistent** :  
Start with PEP8
- ❖ Don't print. **Log**  
>>>import Logging

# PYTHON: NUMPY

## What is Numpy:

Package for scientific computing with Python.

Powerful N-dimensional array objects.

## Why Numpy:

Python is slow

Built-in , precompiled mathematical and statistical algorithms.

# PYTHON: NUMPY

## Important preferences



NumPy is in-memory (what if you don't have enough?)








NumPy is bad in choosing data types. Are you sure you need float64?



NumPy is also bad in choosing algorithms. (e.g., sparse matrix)

# PYTHON: NUMPY

## Useful functions

-  `array.flatten(), array.flat`
-  `array.transpose()`
-  `slicing array[1:3000]`
-  `masking array[1,5,10000]`
-  `array operations: std, argmax`

 NumPy is bad in choosing data types. Are you sure you need float64?

 NumPy is also bad in choosing algorithms. (e.g., sparse matrix)



# PYTHON: SCIPY

## What is SciPy:

Built upon NumPy

Contains implementations of algorithms and functions in: ***Linear Algebra, Signal Processing, FFT, Spatial data etc.***

## Why Numpy:

See above

Sparse matrices handling

# PYTHON: SCIPY

## What is SciPy:

Built upon NumPy

Contains implementations of algorithms and functions in: ***Linear Algebra, Signal Processing, FFT, Spatial data etc.***

## Why Numpy:

See above

Sparse matrices handling

# PANDAS: DATA MUNGING

## What is pandas

 Data analysis tool for processing tabular/  
labeled data.

 Main data structures

Series (1d)

DataFrame(2d)

Panel(3d)

 Supported input/output: CSV, SQL, json, Excel

# PANDAS: DATA MUNGING

## Important Features

 Handling missing data (drop row, fill etc.)

 Automatic plotting (see demo)

 Masking

# SCIKIT-LEARN

## What is SciKit-learn

All extensions of SciPy are called SciKit

SciKit-learn: Machine Learning library

Built upon SciPy and NumPy

# SCIKIT-LEARN

## WORKFLOW

1. Estimator:

the primary objects in scikit-learn.

Performing data fitting , sampling and prediction

2. Choose a model: e.g. SVM classifier

# SOME MORE USEFUL LIB

**matplotlib:** Python's plotting library. Pretty much similar to MatLab's plotting.

**sklearn\_pandas:** will help you integrate pandas data frames to sklearn feature sets

**NLTK:** NLP suite for python

**Network-x:** Python's graph processing library

**Gensim(Word2Vec):** Another ML/DM mainly for topic modeling

# YOUR BEST FRIENDS

## Read the docs:

Numpy, Scipy

scikit-Learn

pandas

## Stackoverflow