Log Analysis With Pandas

Taavi Burns, http://twitter.com/jaaaarel (http://twitter.com/jaaaarel), http://taaviburns.ca (http://taaviburns.ca), taavi@freshbooks.com

What is pandas?

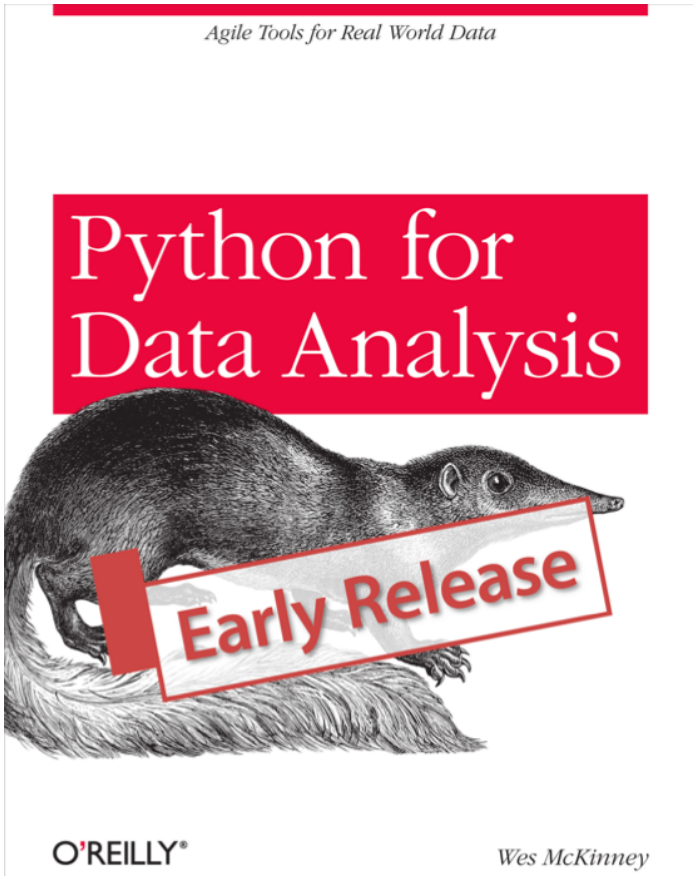
“pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.”

http://pandas.pydata.org/ (http://pandas.pydata.org/)

In [1]:
```
from IPython.core.display import Image

Image(filename='PythonForDataAnalysisCover.png')
```

Out[1]:



Let's dive right in!

Installing things

OSX

```
$ brew install zeromq freetype
```

If you also want HDF5 support:

```
$ brew install hdf5
```

Debian/Ubuntu

```
$ sudo apt-get install libzmq-dev libpng-dev libfreetype6-dev g++
```

If you also want HDF5 support:

```
$ sudo apt-get install libhdf5-serial-dev
```

Python packages

```
$ pip install ipython pyzmq tornado numpy matplotlib pandas
```

If you also want HDF5 support:

```
$ pip install cython numexpr tables
```

Run it!

```
$ ipython notebook --pylab inline
```

Hello, World!

In [2]:
```
print "Hello, world!"
```

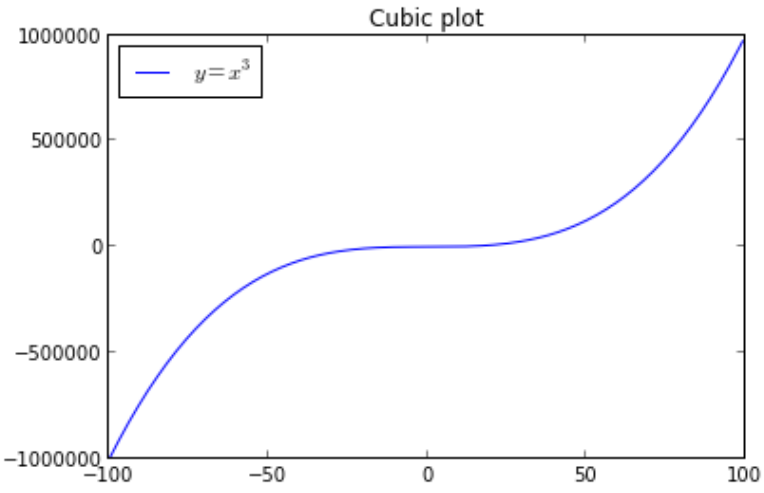
```
Hello, world!
```

Inline plotting

Line plot

In [3]:
```
data = range(-100, 100)
title('Cubic plot')
plot(data, [x**3 for x in data], label="$y = x^3$")
legend(loc='best')
```

Out[3]: <matplotlib.legend.Legend at 0x104a2c2d0>



Histogram

In [4]:
```
import random
datapoints = [random.gauss(0, 1) for _ in range(1000)]
title('Gaussian distribution')
h = hist(datapoints)
```