Security Analytics
Course Introduction

Course Goal, Scope, and Approach

- Goal
  - Prepare students as security analysts/security data scientists

- Scope
  - Basic data science theories and methods
  - How to use machine learning for analyze security analysis

- Approach
  - 50% theory, 50% hands-on exercise
  - Teaching through examples
  - Use both slides and Jupyter notebook

Text Books

- “Data Science for Business” by Foster Provost and Tom Fawcett, O’Reilly Media, 2013
  - Highly recommend
  - Data science theories and practices
  - Some chapters will be included in homework
  - Some content will be used in quizzes
- “Data Driven Security” by Bob Rudis and Jay Jacobs, Wiley, 2014
  - More on practical analysis of security data
  - Use both R and Python as programming languages
  - Recommend it as additional reading

Grading Components

- 3 Homework assignments - 30%
  - Set up software environment and accomplish tasks using python and Jupyter Notebook
  - Mainly hands-on exercises
- 2 Quizzes – 30%
  - Knowledge test
- 1 Course project – 40%
  - Apply data science to solve security-related problems

Course Project

- Project scope: Anything applying analytics and machine learning to solve security problems, can focus on:
  - Machine learning algorithms
  - Visualization
  - Big data
  - Real world problems (use any combination of above)

- Grading is based on following metrics:
  - Project proposal: problem/motivation, related work/literature review, approach, expected results
  - Project results: source code, results and discussion
  - Final presentation: Organization, speech, Q&A

- Start now to form your team (2-4 people/team)

Security Analytics Introduction

What is Security Analytics

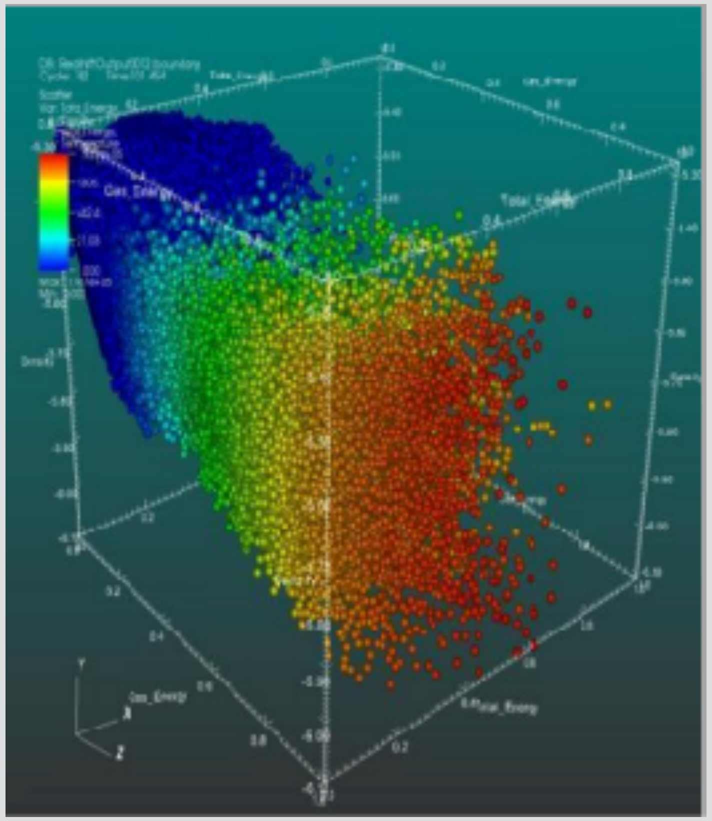
- Analytics: discipline that applies logic and mathematics to data to provide insights for making better decisions.
- Security Analytics: using analytics for the domain of information security
- These are practices of security analytics:
  - Analyzing logs, network traffic, application transaction, and other security-relevant data with algorithms
- These are not security analytics:
  - Signature-based instruction detection; blacklisting or whitelisting; naïve thresholds, searches, reports and queries.

Data Science

Machine Learning

- Give computers the ability to learn from data
- Supervised learning
- Unsupervised learning

Data Visualization



Database

Application Examples

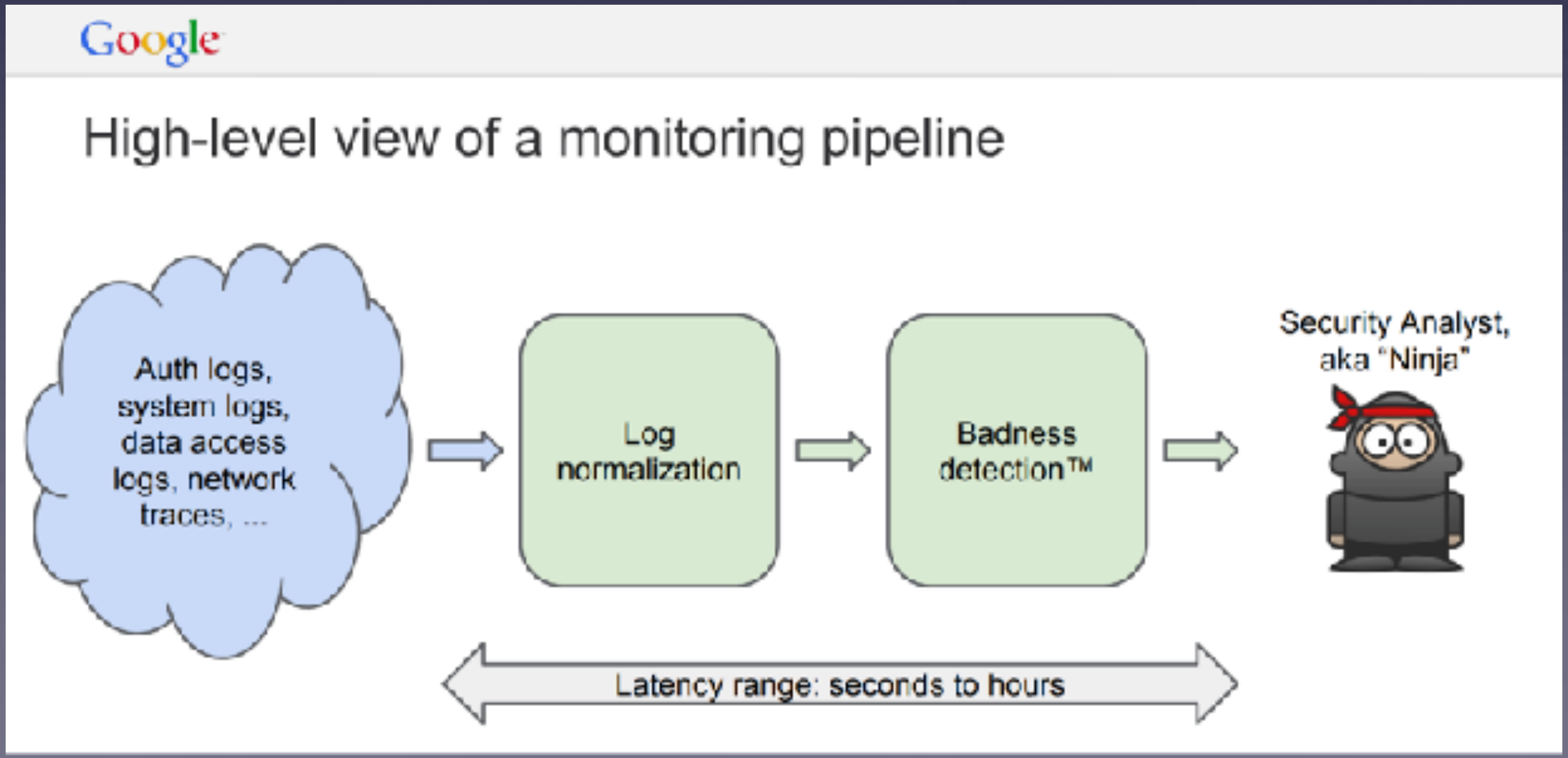
- Sales and ads
  - Amazon, Netflix
- Spam detection
  - Google, Microsoft, Yahoo, Facebook, Twitter
- Image and voice recognition
- Self driving car
- ….

Security Application Examples

- Malware detection
- Phishing detection
- Fraud detection systems
- Compromised device detection
- Attack actor prediction

Security Application Examples

- Google Security Team
  - Account hijacking detection
  - Click fraud detection
  - DoS detection
  - Infrastructure compromise detection



Google

High-level view of a monitoring pipeline

Auth logs, system logs, data access logs, network traces, ...

Log normalization

Badness detection™

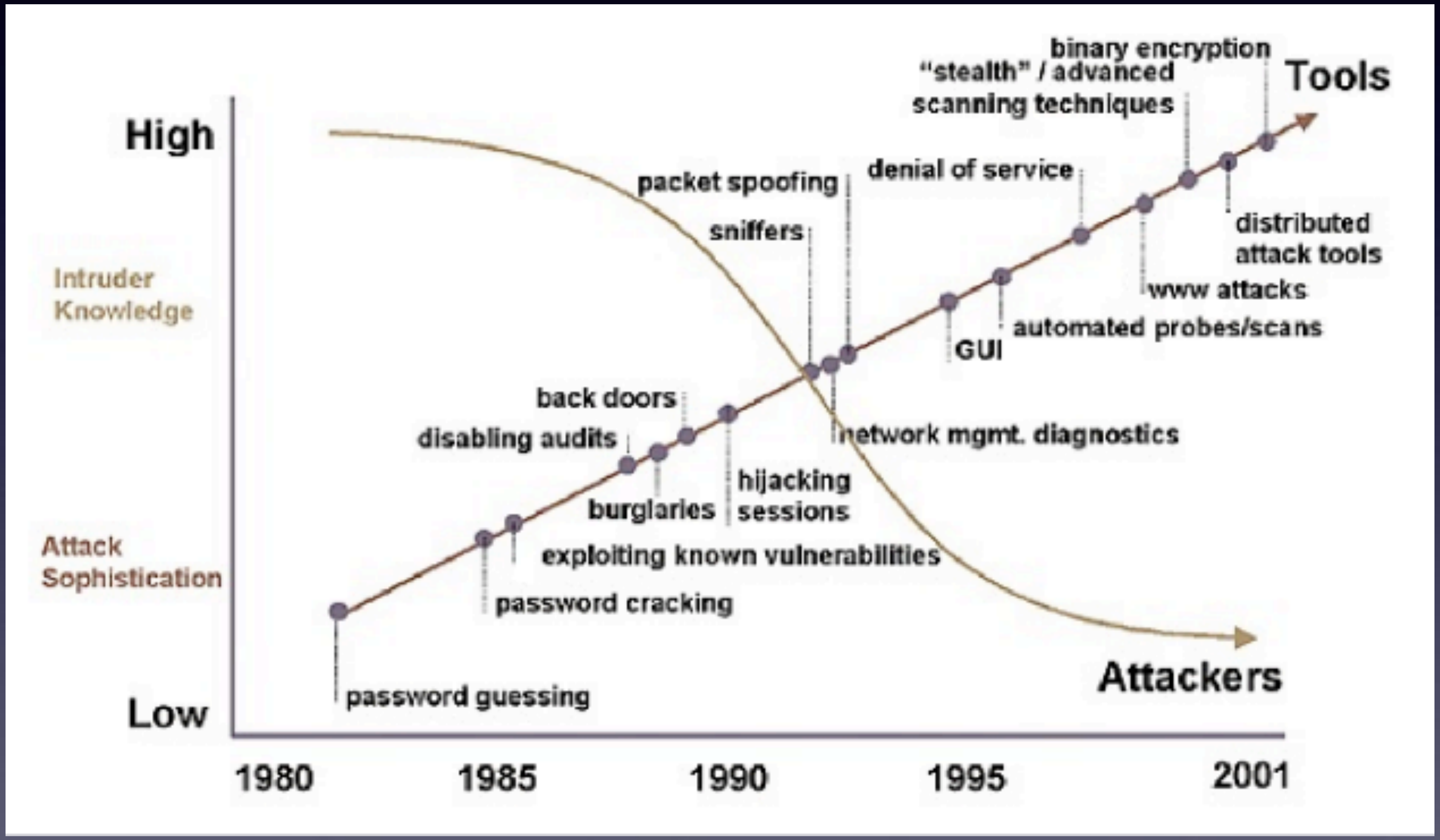
Security Analyst, aka "Ninja"

Latency range: seconds to hours

Why now?

- Attack landscape
  - Attacks increasingly more sophisticated
  - Required attacker knowledge going down
  - Highly motivated attackers
- Current detection techniques failing
  - Zero-day attacks
  - Polymorphic malware
  - Advanced persistent threat (APT)
- Network perimeter dissolving
  - Cloud, Mobile, IoT
  - Focus on data, people, and usage patterns

Why now?



- Current detection techniques failing
  - Median time between breach and awareness: **300~400+** days

Explosion of Malware

- 403 million new variants of malware created in 2011

- 100,000 unique malware samples collected daily by McAfee in Q1 2012

- More than 100 million samples in McAfee's malware signature database by Q3 2012

- Practically impossible to keep up with signatures

Advanced Persistent Threats (APT)

- Targeted attack against a high-value asset
- Multi-stage
- Avoid alerts
    - Use stolen user credentials
    - Zero-day exploits
    - Low profile in network
    - Slow progress: operating over months or years
    - Beyond limited correlation time windows of today's Intrusion Detection Systems (IDSs)

Why Security Analytics

- Need a better and faster way to detect cyber threats
- Need long-term correlations and analysis in data
  - Go beyond real-time and short-term analysis
- Contextual security intelligence
  - Alerts from Intrusion Detection Systems (IDS) and firewalls need to be further analyzed with context
- More data with various of types need to be analyzed
  - Adoption of cloud, IoT dramatically increases number of entities to be monitored

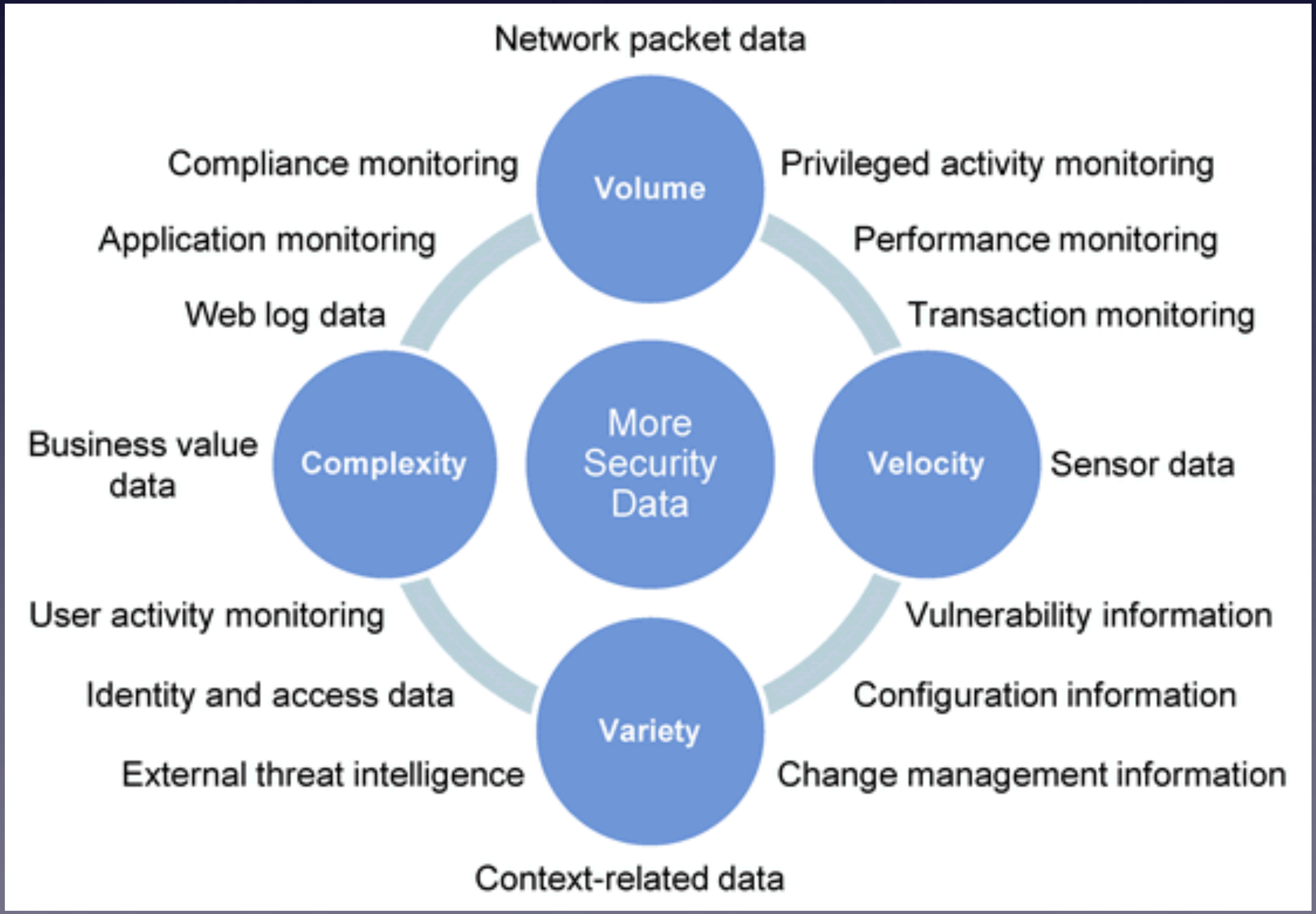
Why Security Analytics

- Data science enables storage and analysis of higher volumes and more types of data
- 2010 Verizon data breach investigation
  - In 86% of cases of breach, evidence was in the logs
  - Detection mechanism failed to raise alerts

RSA Security Analytics Demo

Security Became an Analytical Problem

- How do we make sense of the large and diverse data?



Techniques in Security Analytics

- Statistics
- Machine learning
  - Supervised learning
    - Data is labeled, and has a target field
    - Classification: Logistic regression, K-nearest neighbors, Decision tress, Support Vector Machines, neural networks, etc.
  - Unsupervised learning
    - Find hidden structure from unlabeled data
    - Clustering
    - Principle component analysis
    - Unsupervised deep learning
- Text Mining
  - Derive information from text
  - Usually need to convert text into feature vectors
  - Can use both supervised and unsupervised learning

Course Programming Language & Software

Python

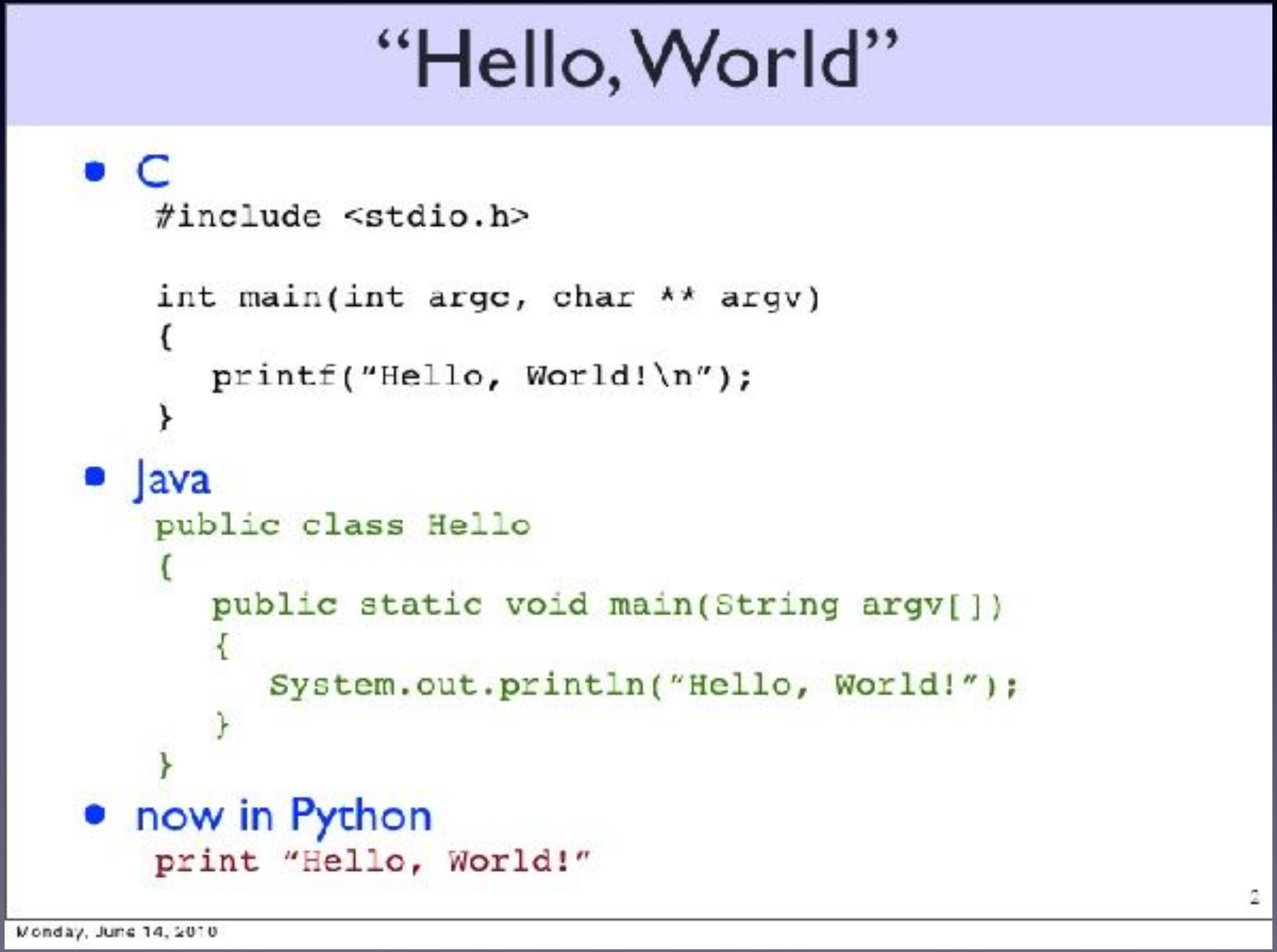
- Open source scripting language
- Developed by Guido Van Rossum in late 1980s
- Named after Monty Python comedy group
- Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles

Why Python for Analytics?

1. Easy to Learn & Concise

- Syntax is user friendly, consistent and elegant
- Generally, Python code is 70% shorter than the same in Java
- Python has a better and easier syntax than other objective-oriented languages

1. Easy to Learn & Concise



"Hello, World"

- C

```
#include <stdio.h>

int main(int argc, char ** argv)
{
   printf("Hello, World!\n");
}
```

- Java

```
public class Hello
{
   public static void main(String argv[])
   {
      System.out.println("Hello, World!");
   }
}
```

- now in Python

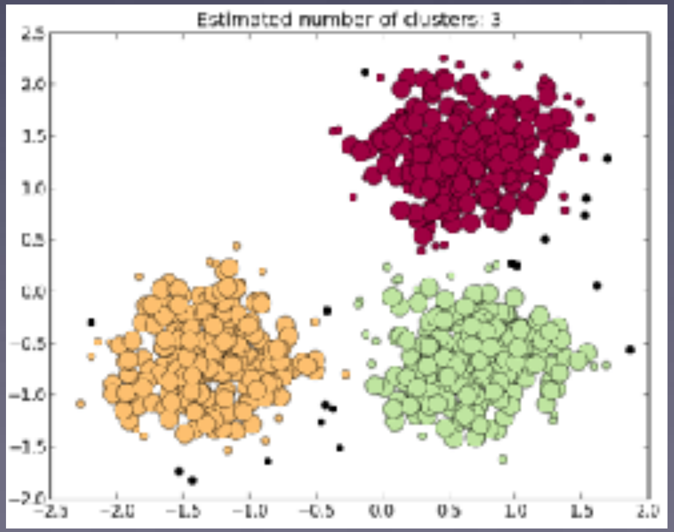
```
print "Hello, World!"
```

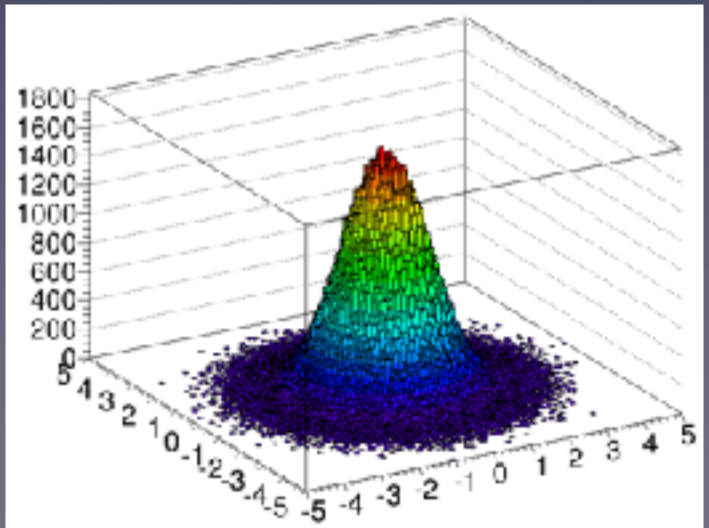
2. Large Community = Documentation = Brainpower

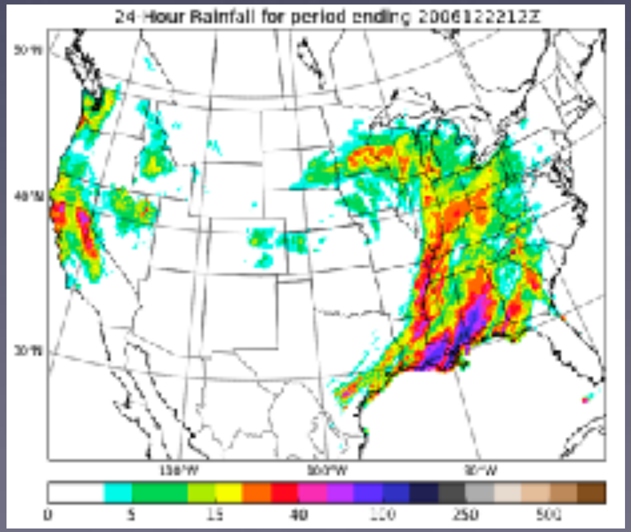
- Thorough and complete:
  - Official Tutorial: http://docs.python.org/tutorial/
  - Language Reference: http://docs.python.org/reference/
- Daily round-up of py news. Active user engagement.
  - Pythonware Daily: http://www.pythonware.com/daily/
  - Planet Python: http://planet.python.org/
- A very high chance of python related query getting answered in seconds! As is StackOverflow.
  - Irc Node: http://www.python.org/community/irc/
  - StackOverflow: stackoverflow.com/questions/tagged/python?sort=newest

3. Growing Data Analytics Libraries

- Scientific computing: Numpy, Pandas, SciPy,…
- Machine learning: Scikit-learn, Shogun, PyLearn2, …
- Plotting and visualization: seaborn,matplotlib, plotly, …

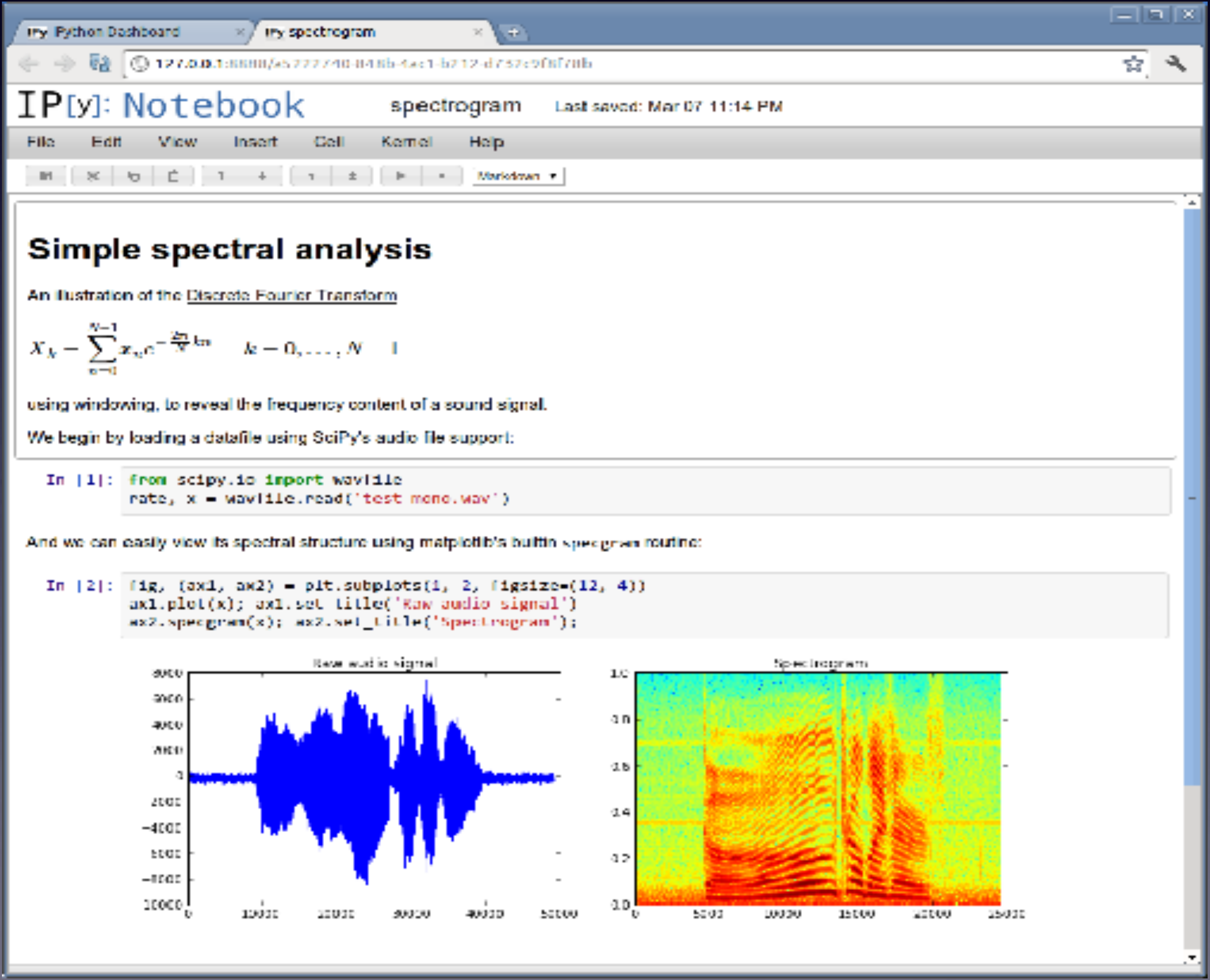






4. Jupyter Notebook (IPython Notebook)

- Jupyter notebook is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and media together.



spectrogram Last saved: Mar 07 11:14 PM

Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0,\dots,N-1$$

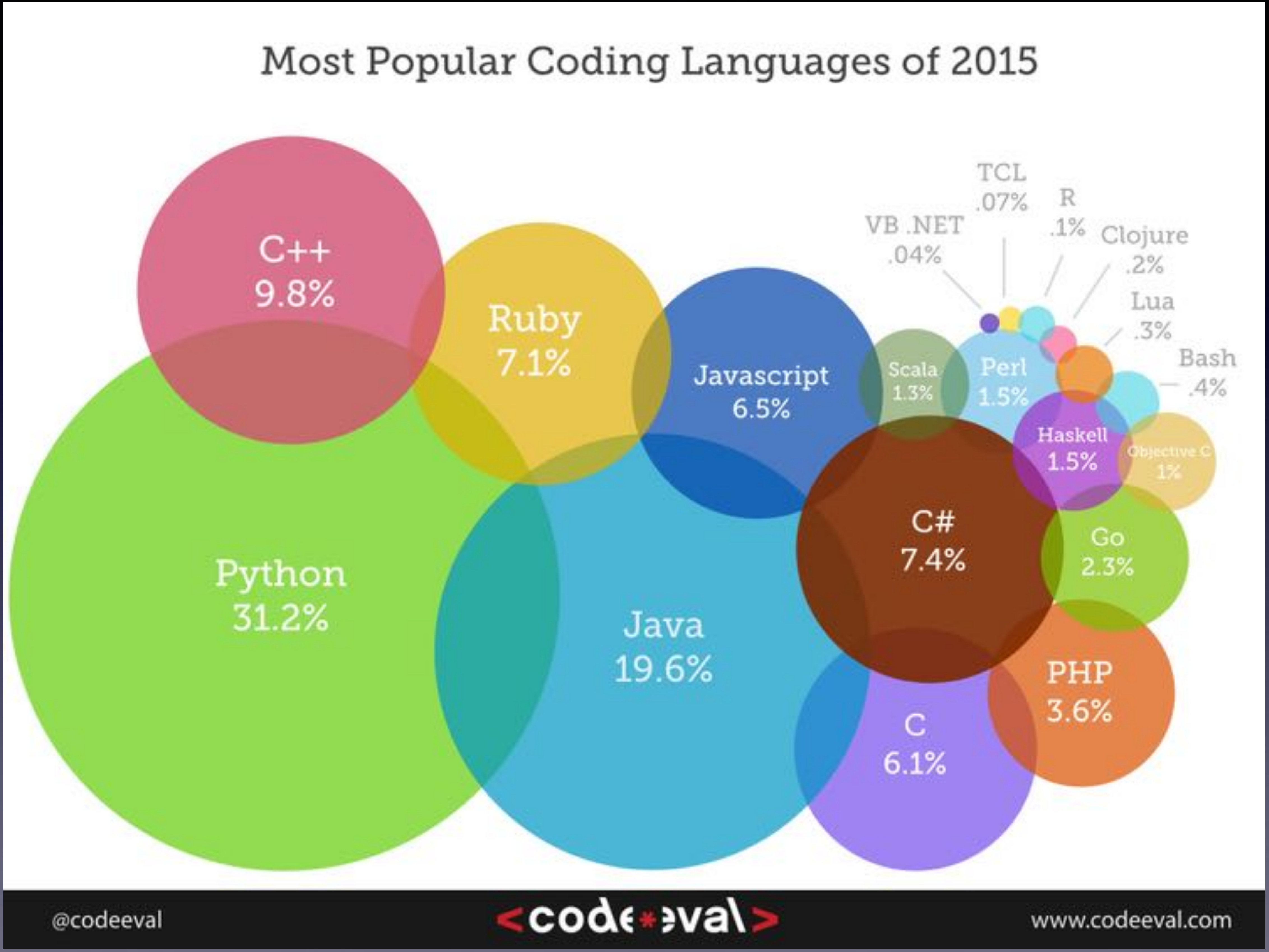
using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

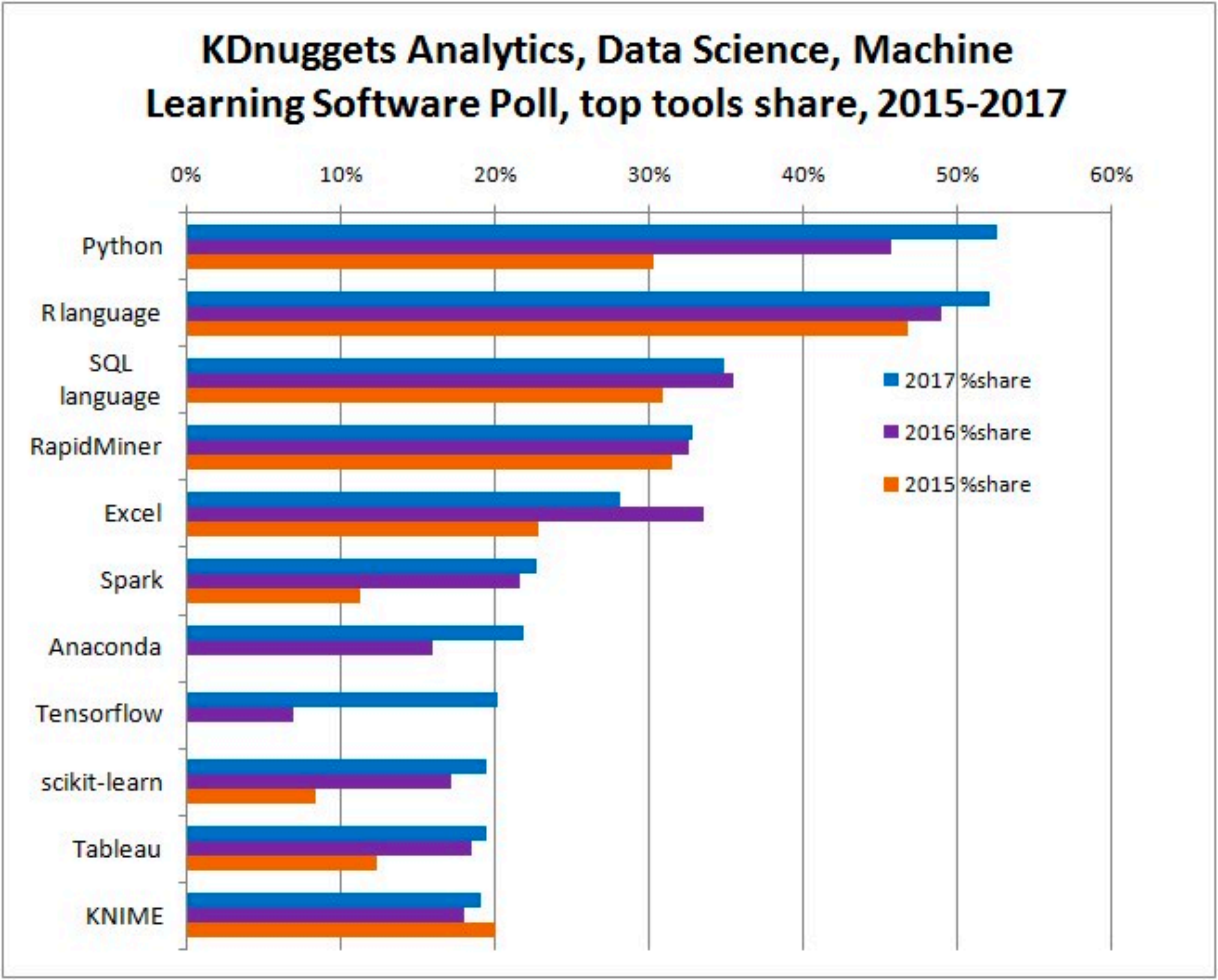
```
In [1]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```



Most Popular Coding Languages of 2015



KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017

Demo:

Using Python and Jupyter Notebook

How to Install Python and Jupyter Notebook

- How to Install Python?
  - Go to: https://www.python.org/
- How to install Jupyter Notebook?
  - Go to: http://jupyter.readthedocs.io/en/latest/install.html

Class 1 Homework

- Install Python, Jupyter notebook, and necessary libraries like Numpy, Pandas, sklearn, matplotlib, seaborn, etc.

- Download the Obama and Romney poll data as used in this class, and complete following tasks:
  - Use the same URL link to the .csv file to download data
  - Describe the “Romney” data using the describe function
  - Find out pollster(s) on which Romney receive the most and the least votes
  - Draw a time series graph that shows the total number of Romney’s votes along with pollster start date

- Submit your homework in an .ipynb

- Homework due before next class (submit by 6:00PM ET, Monday, 9/10/2018)

Next Week

- Security data sources
- Hands-on exploration of security data
- **Bring your laptop!**