Lecture 1: Course Introduction and Logistics

Instructor: Saravanan Thirumuruganathan

Outline

1. Data Mining/Science Basics
2. Logistics
3. Scientific Python
4. IPython Notebook Demo

Introduction To Data Mining/Science

Big Data[1]

[1] https://www.pinterest.com/pin/101753272804937744/

Big Data[2]

[2] http://memegenerator.net/instance/55214797

Big Data

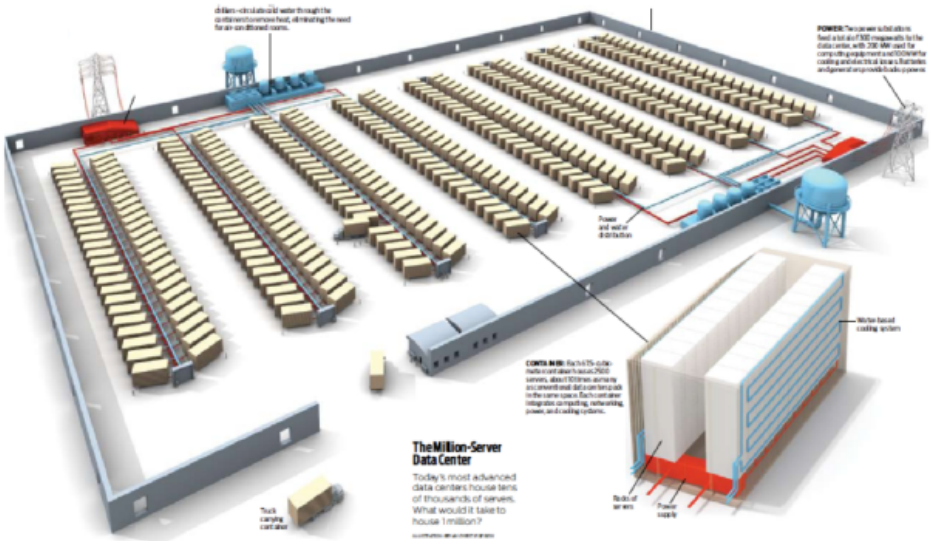
"Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days."

- Eric Schmidt, Google

One Second on the Internet: `http://onesecond.designly.com/`

Smarter Devices

Commodity Computing



Michael Franklin, UC Berkeley

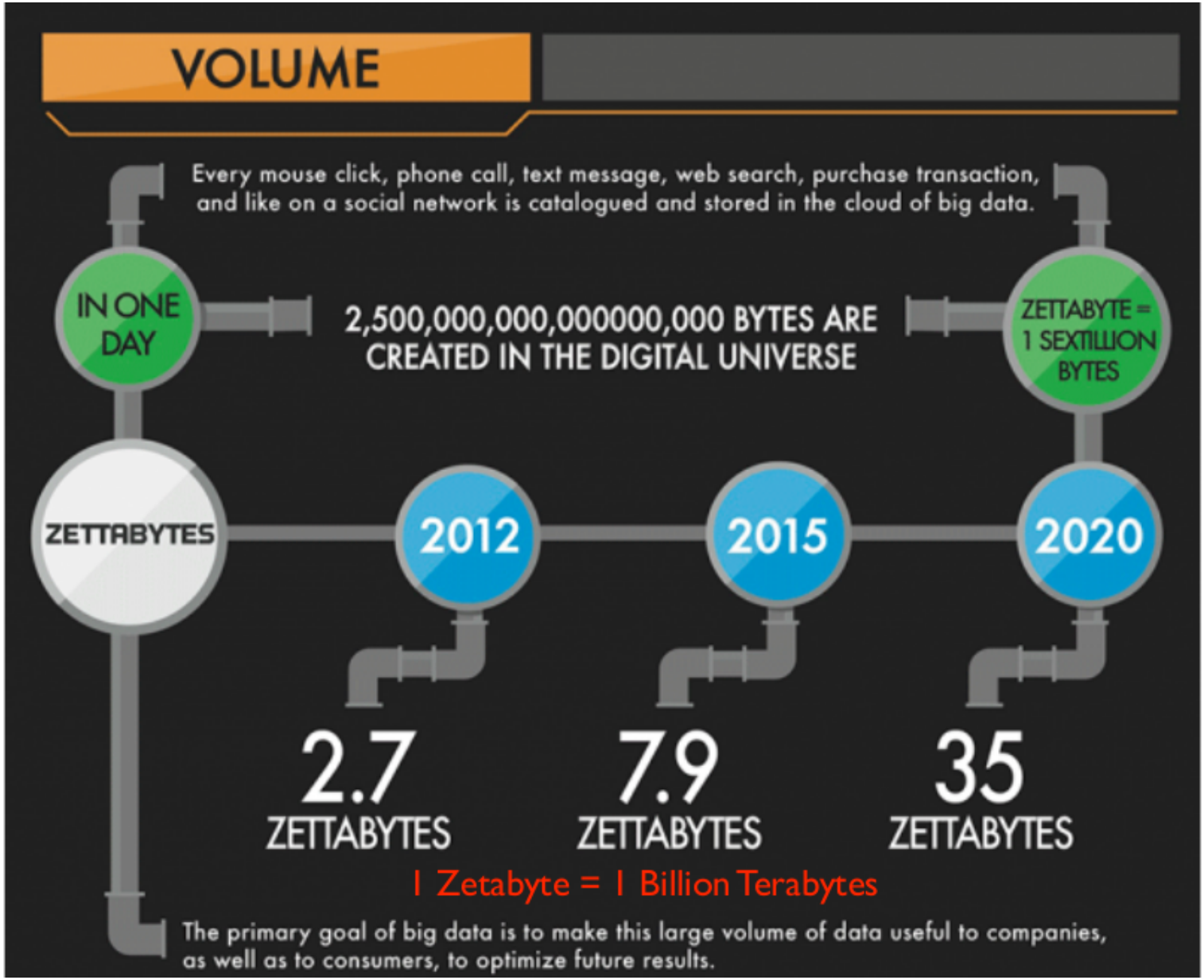
Ubiquitous Connectivity

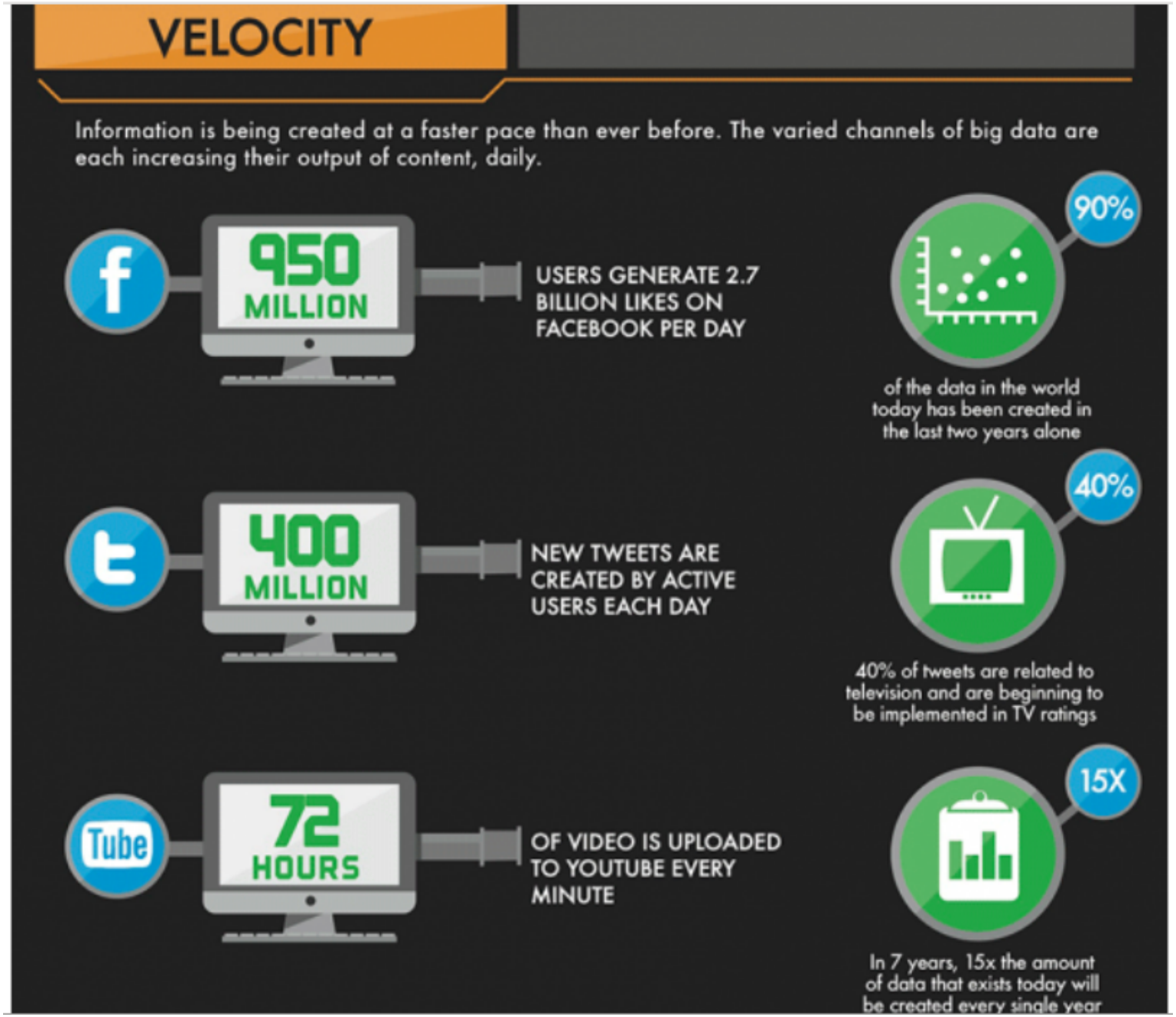
skype™

Michael Franklin, UC Berkeley

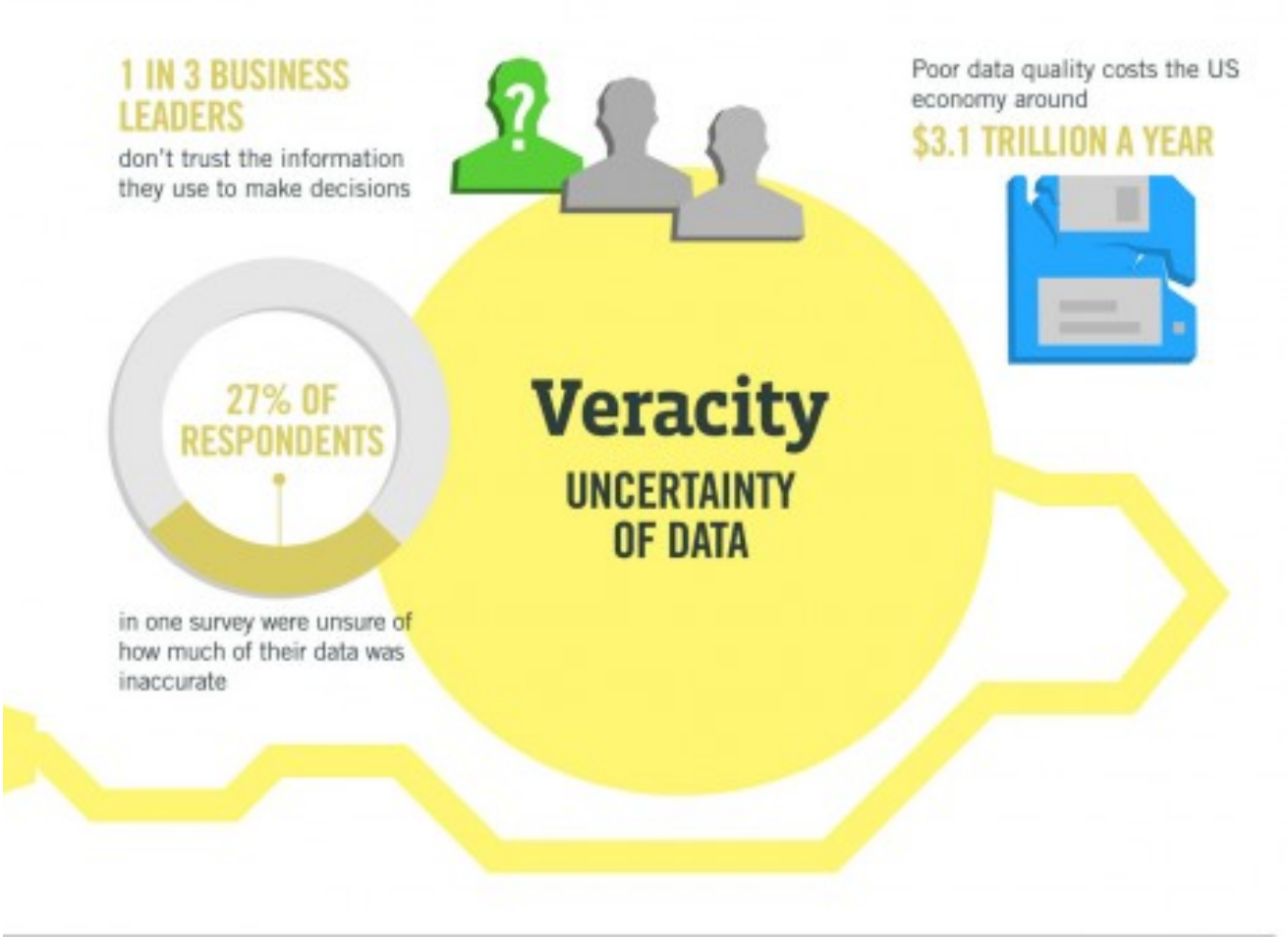
Big Data - 4 V's

- Volume
- Velocity
- Variety
- *Veracity*


VOLUME
Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.
IN ONE DAY
2,500,000,000,000000,000 BYTES ARE CREATED IN THE DIGITAL UNIVERSE
ZETTABYTE = 1 SEXTILLION BYTES
ZETTABYTES
2012
2015
2020
2.7 ZETTABYTES
7.9 ZETTABYTES
35 ZETTABYTES
1 Zetabyte = 1 Billion Terabytes
The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

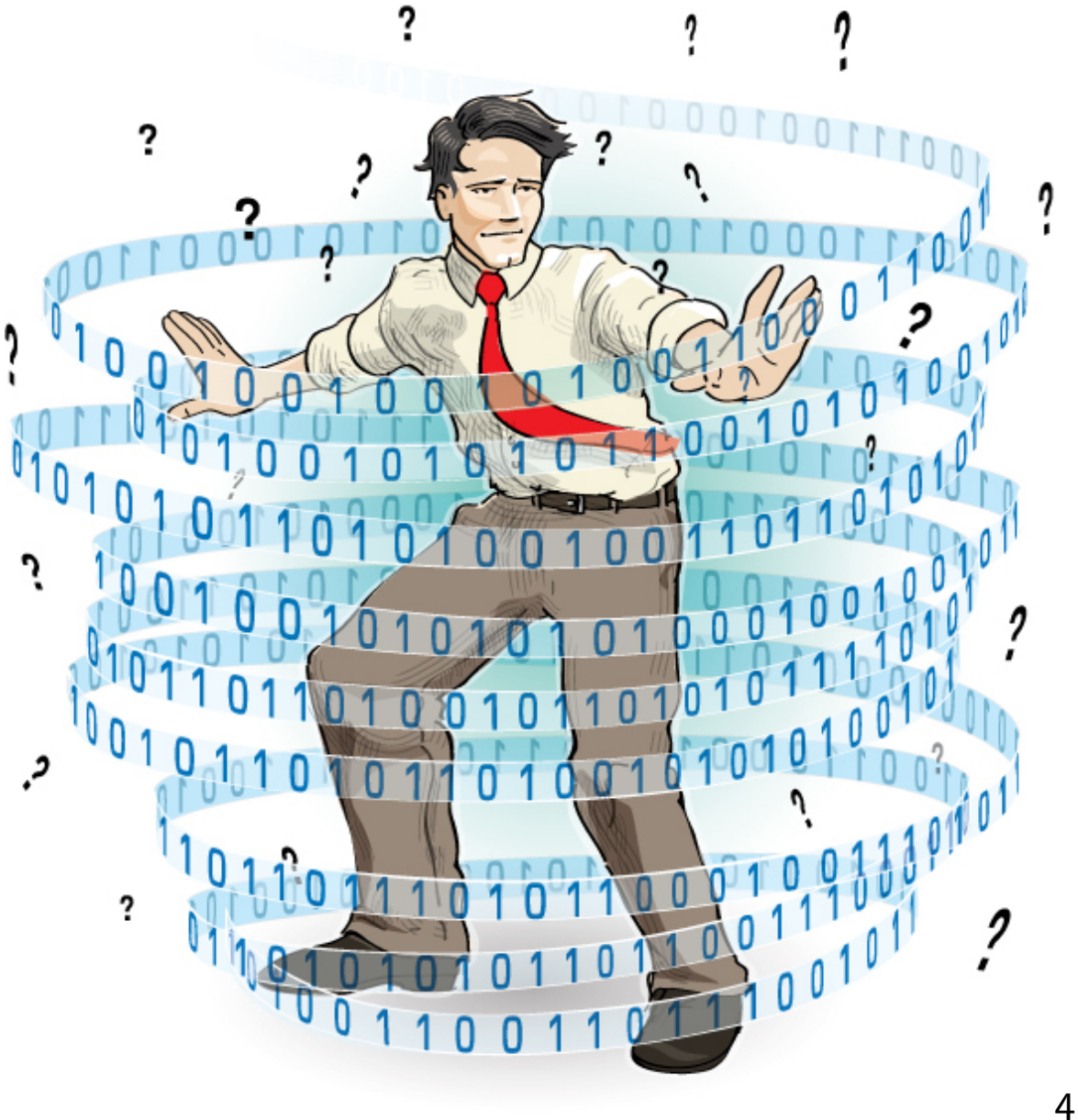

VELOCITY
Information is being created at a faster pace than ever before. The varied channels of big data are each increasing their output of content, daily.
950 MILLION
USERS GENERATE 2.7 BILLION LIKES ON FACEBOOK PER DAY
90%
of the data in the world today has been created in the last two years alone
400 MILLION
NEW TWEETS ARE CREATED BY ACTIVE USERS EACH DAY
40%
40% of tweets are related to television and are beginning to be implemented in TV ratings
Tube
72 HOURS
OF VIDEO IS UPLOADED TO YOUTUBE EVERY MINUTE
15X
In 7 years, 15x the amount of data that exists today will be created every single year


VARIETY
In today's multi-faceted Internet culture, the great volume of data is also extremely varied in its form. So many variables can be thrown at a company that the true value of information can often be lost in the sea of data.
PURCHASE TRANSACTIONS
WEBSITE TRAFFIC
REWARDS PROGRAMS
QUARTERLY BUSINESS REPORTS
TWITTER
FACEBOOK
BLOG CONTENT



3

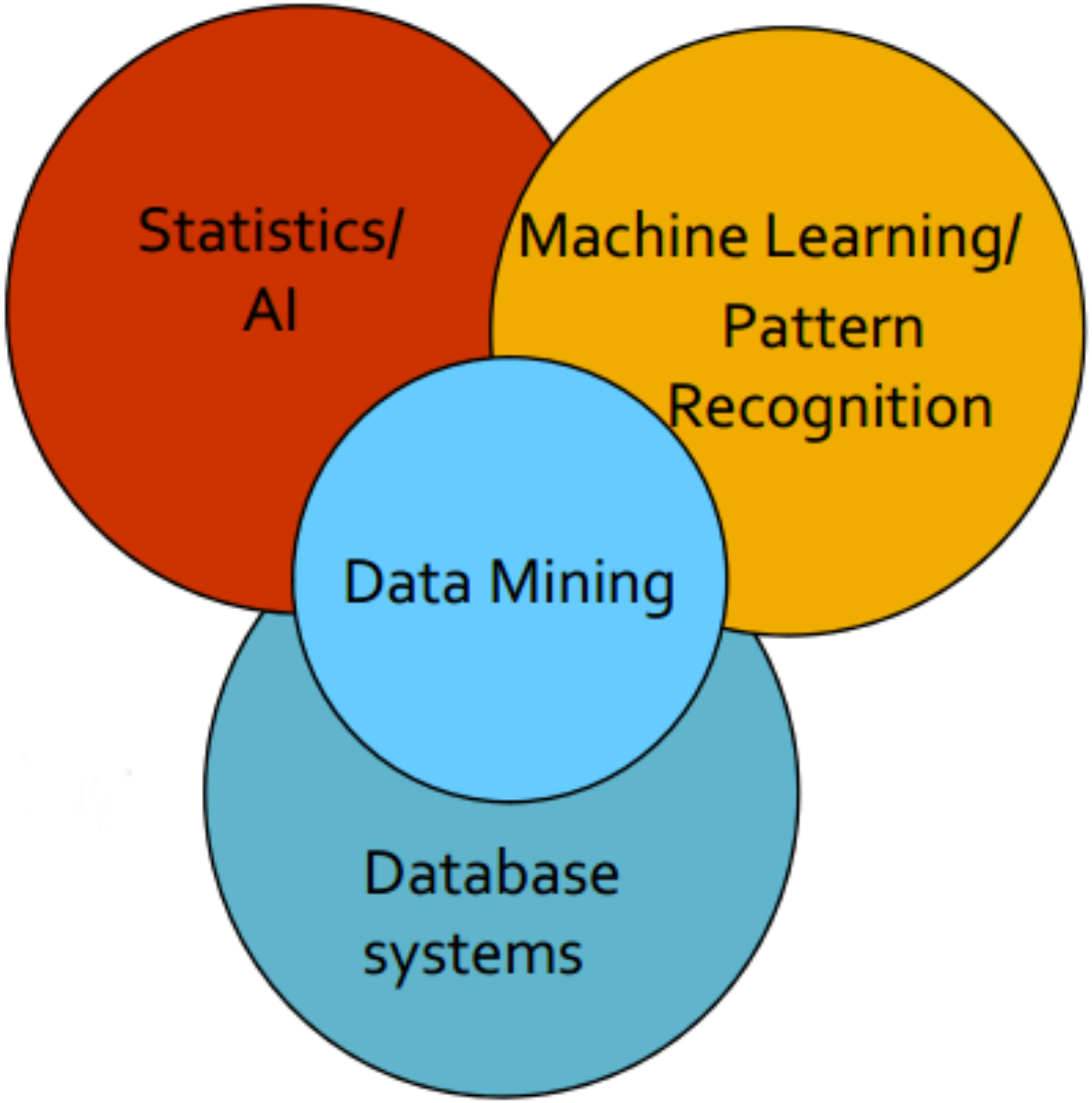
[3] http://www.ibmbigdatahub.com/



[4] https://www.behance.net/gallery/5958295/Data-Hero-Oya-Group

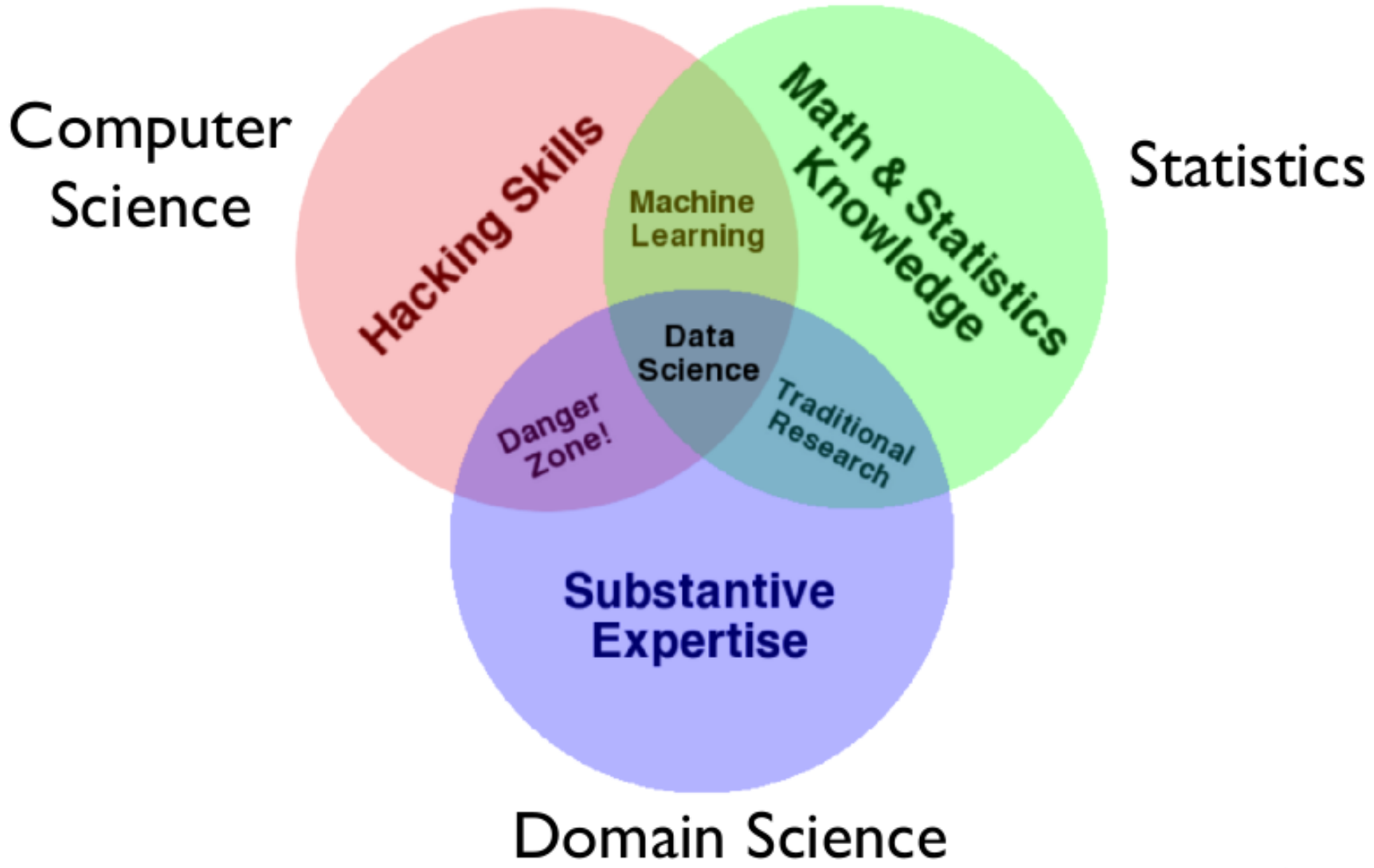
Data Mining

- Process of semiautomatically analyzing large databases to find **patterns** that are
  - **valid**: hold on new data with some certainty
  - **novel**: nonobvious to the system
  - **useful**: should be possible to act on the item
  - **understandable**: humans should be able to interpret the pattern


Statistics/
AI
Machine Learning/
Pattern
Recognition
Data Mining
Database
systems

Data Science

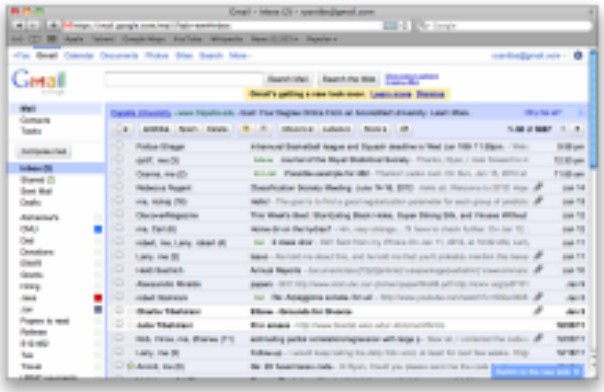
- To gain insights into data through computation, statistics, and visualization
- "A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician" - Josh Blumenstock


Computer Science
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Statistics
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Domain Science
Drew Conway

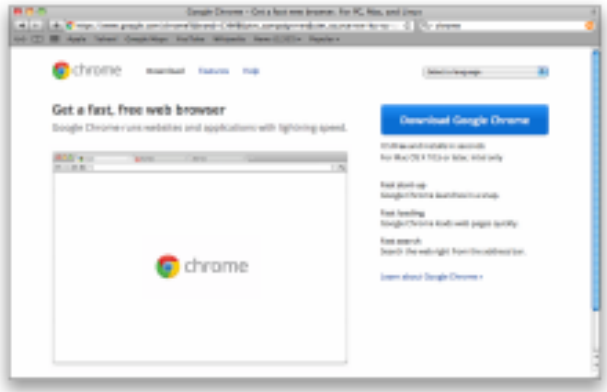
Google

Search

Ads

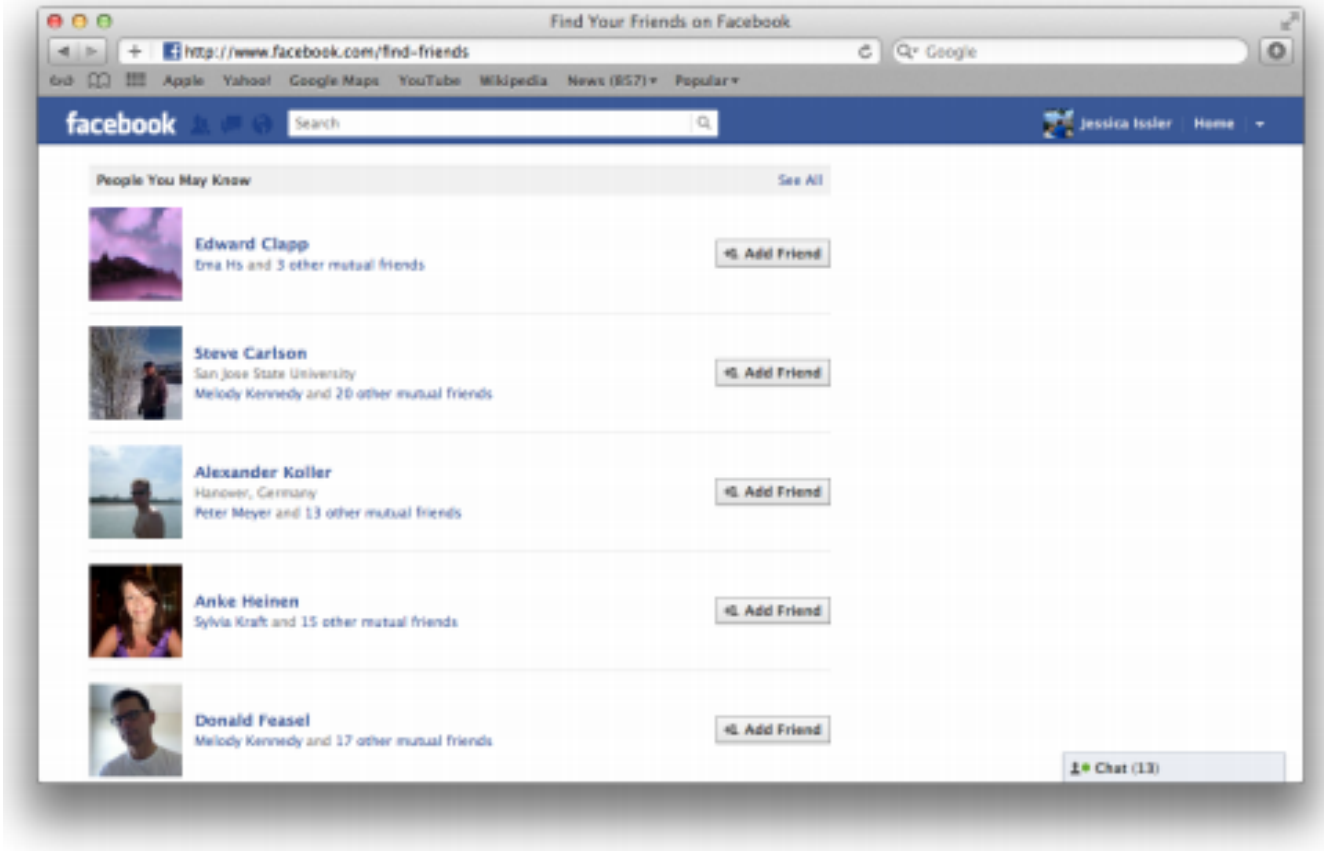


Gmail



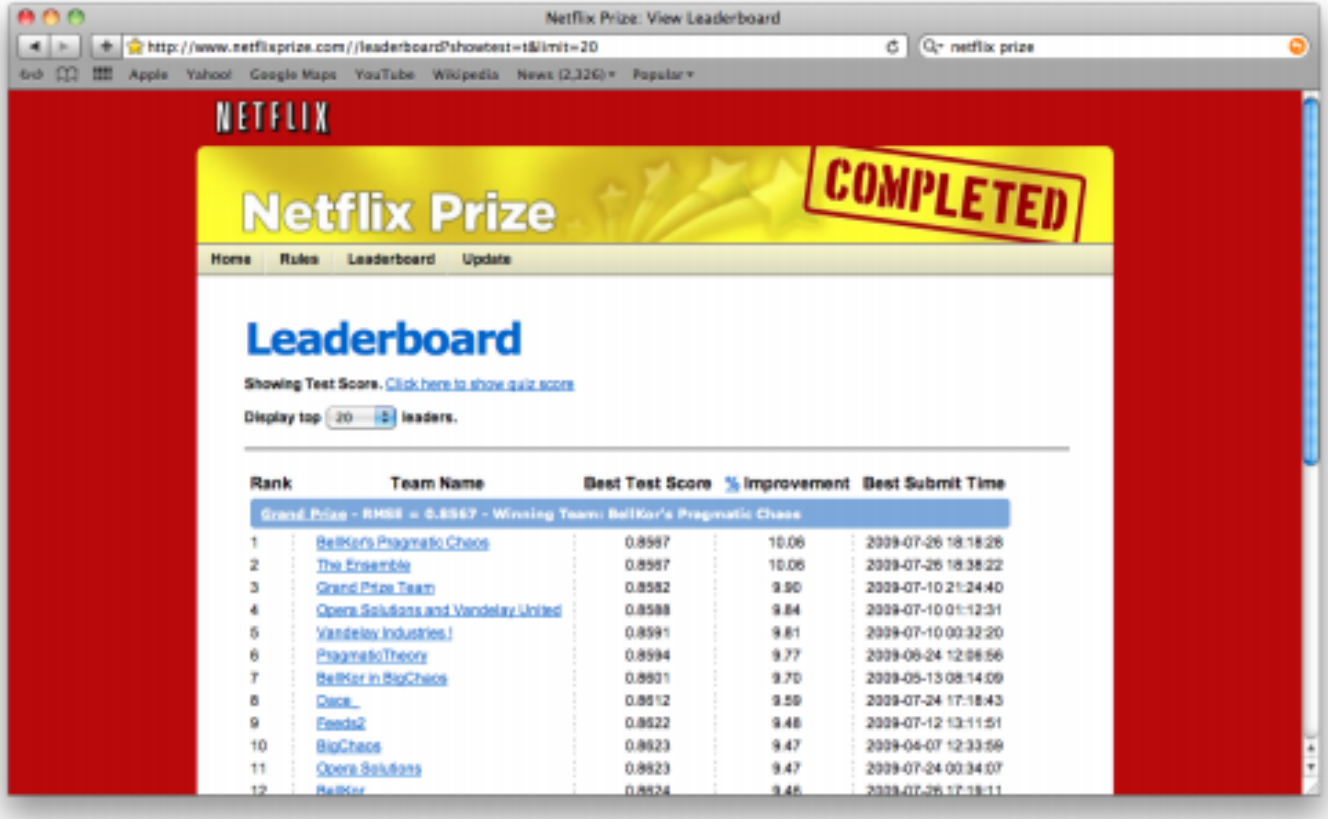
Chrome

Facebook



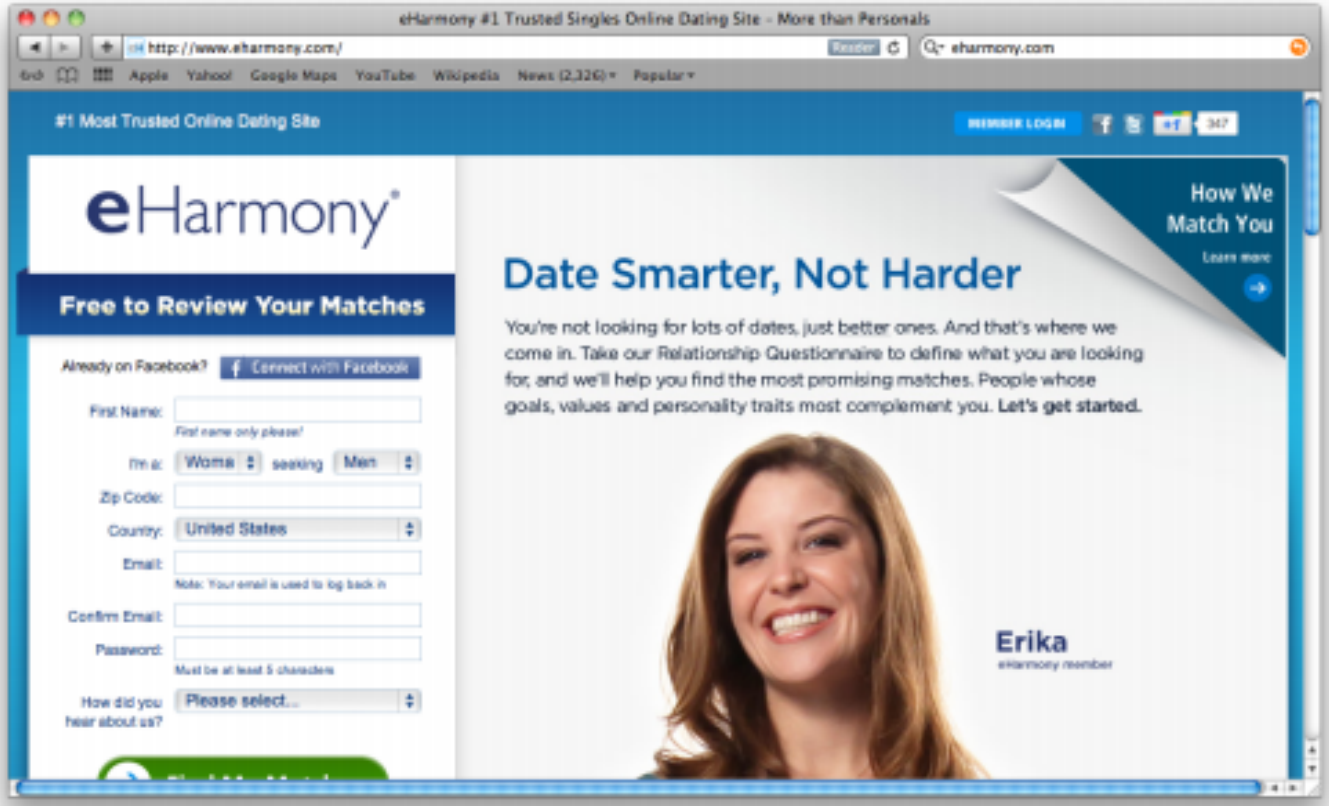
People you may know

Netflix



$1M prize!

eHarmony

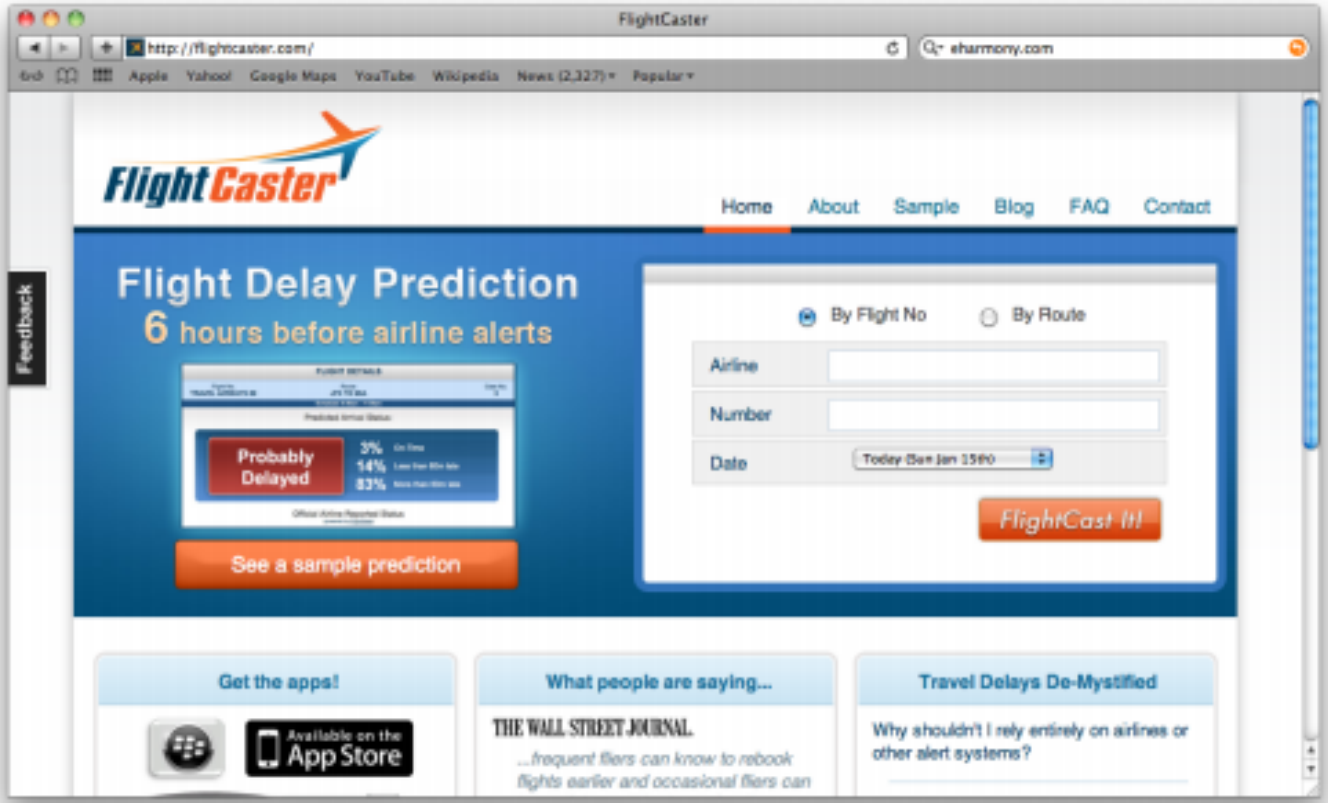


Falling in love with statistics

FICO

An algorithm that could cause a lot of grief

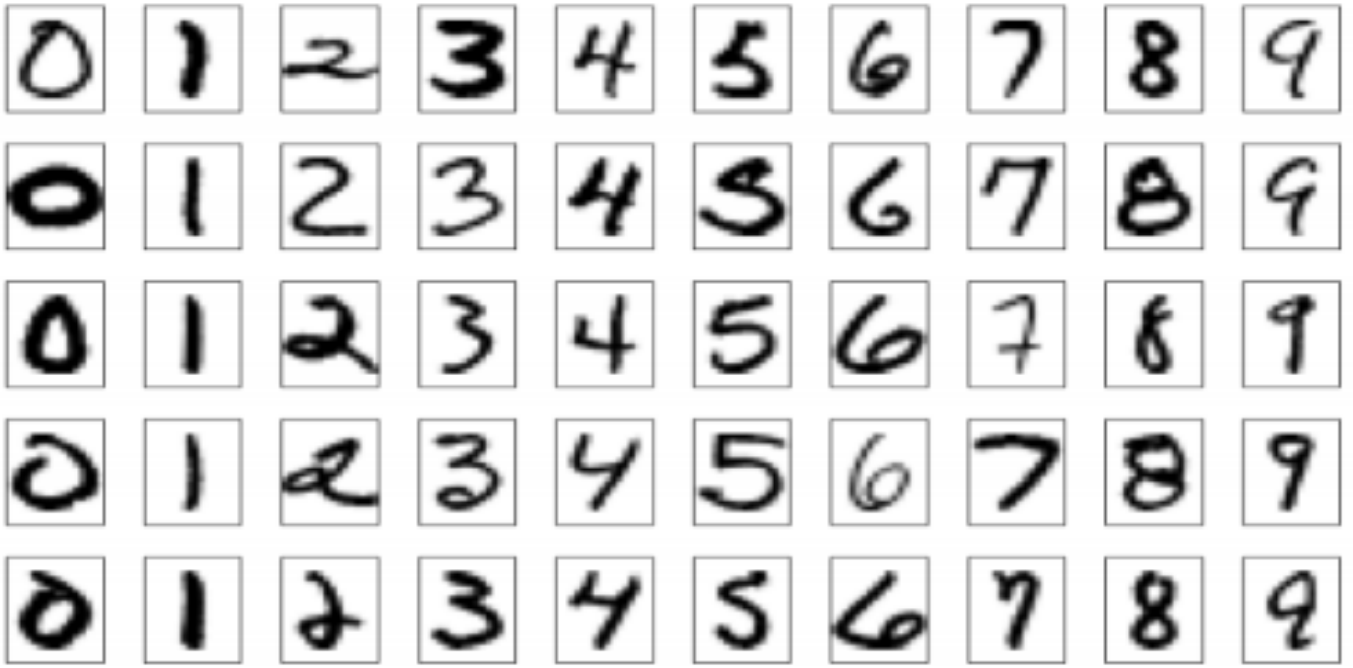
FlightCaster



IBM's Watson

A combination of many things, including data mining

Handwritten postal codes



(From ESL p. 404)

"By 2018, the US could face a shortage of up to 190,000 workers with analytical skills"

McKinsey Global Institute

"The sexy job in the next 10 years will be ~~statisticians.~~" *Data Scientists?*

Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

Logistics

My Background

- Saravanan Thirumuruganathan
- Final year PhD Student working with Dr.Gautam Das
- Website:
  `http://saravananthirumuruganathan.appspot.com`
- Interests: Data Mining, Algorithms, Data Exploration, Social Networks, Machine Learning, Artificial Intelligence

Course Details

- Lectures: TuTh 2-3:30pm, PKH 321
- Course Website: `http://saravanan-thirumuruganathan.github.io/cse5334Spring2015/index.html`
- Instructor: Saravanan Thirumuruganathan
  - Mail: firstname.lastname[at]mavs.uta.edu
  - Office Hours: TuTh 12:30-2:00pm, Fri: 2-5pm or by appointment
- TA: TBD

Piazza

- Q&A Platform
- “mixture between a wiki and a forum”
- `https://piazza.com/class/i551721xpki6w7`
- Please use it as much as possible for public/common questions and clarifications

Text Books[5]

NEW GENERATION:
FACEBOOK LAST SEEN - 6 SECONDS AGO;
WHATSAPP LAST SEEN - 4 SECONDS AGO;
.
.
.
.
.
BUT TEXTBOOK LAST SEEN - 8 MONTHS AGO!

[5] http://www.santabanta.com/

Text Books

- There is no book to cover them all
- Multiple books (free eBook links in Website)
  - **[MMDS]** Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman.
  - **[DMA]** Data Mining and Analysis: Fundamental Concepts and Algorithms by Mohammed Zaki and Wagner Meira.
  - **[ISLR]** An Introduction to Statistical Learning with Applications in R.
  - **[IIR]** Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze.

Grading

- 5 Programming Projects: 30%
- Capstone Project: 10%
- Midterm: 30%
- Final : 30% (non comprehensive)
- Grading will be on a curve

Programming Projects

- Team based, 1-3 members
- Coding will be in Python
- Some of them will be intensive
- Startup code, testing code will be provided
- Capstone project
- Data Science portfolios

Programming Projects

- Project/Dataset suggestions welcome!
  1. Exploratory Data Analysis using Python, Pandas, Matplotlib and Seaborn.
  2. Classification Algorithms using Scikit-learn
  3. Clustering using Scikit-learn
  4. Search Engine Basics
  5. Recommender Systems using Scipy
  6. Capstone Project: Putting it all together

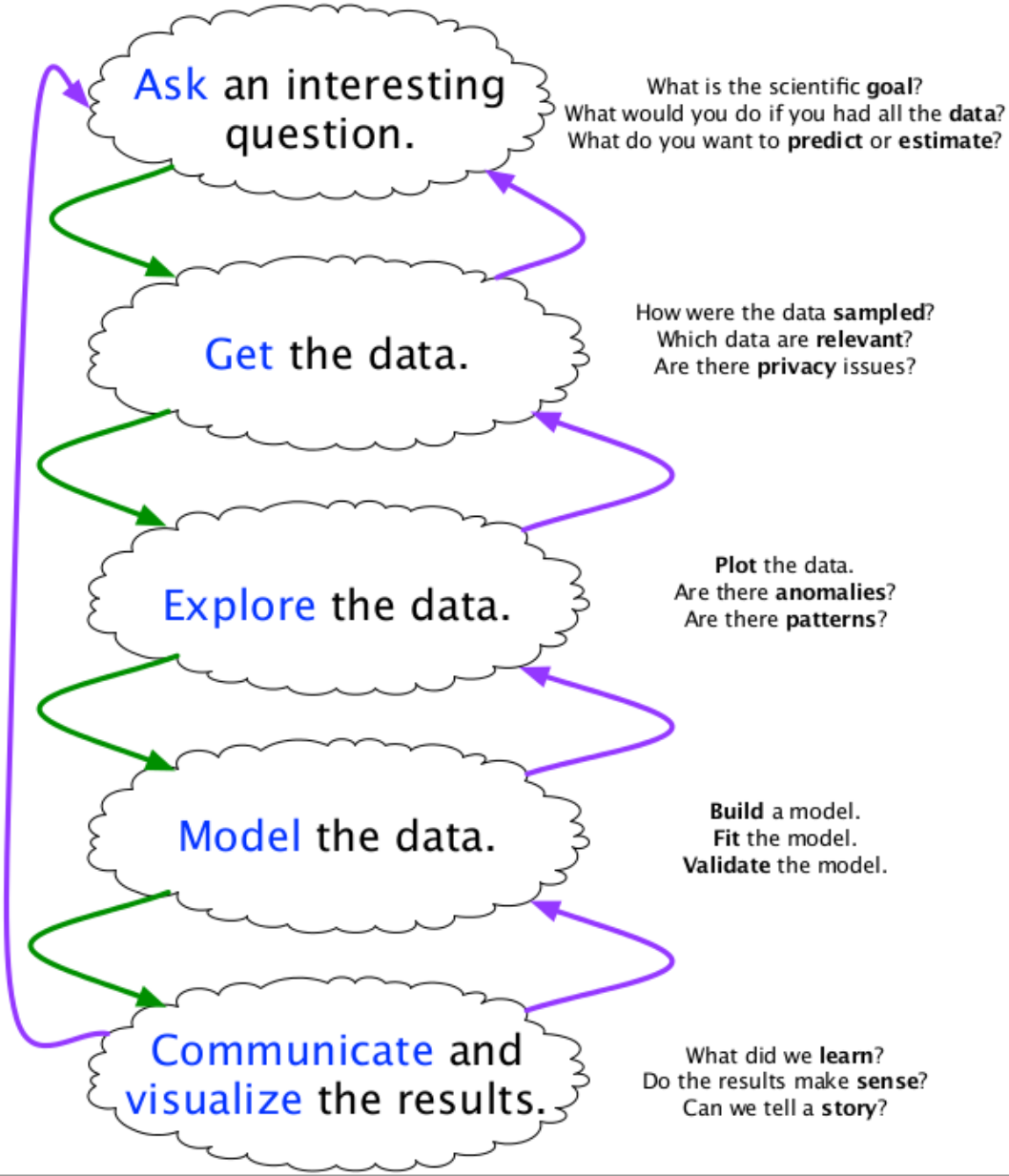
Late Days

- Due at 11 : 59pm
- 5 late days per student for the semester
- No more than 2 could be used per project
- No increments
- Late Penalty: 50% per day
- No point in submitting after 4 days

Assignment Advice

- Start early
- Find good team members (Piazza support will be provided)
- First assignment will be out in 2 weeks
- Okay to change teams per project
- Everyone in team gets same score
- Collaboration/Brainstorming is Okay!
- No plagiarism!

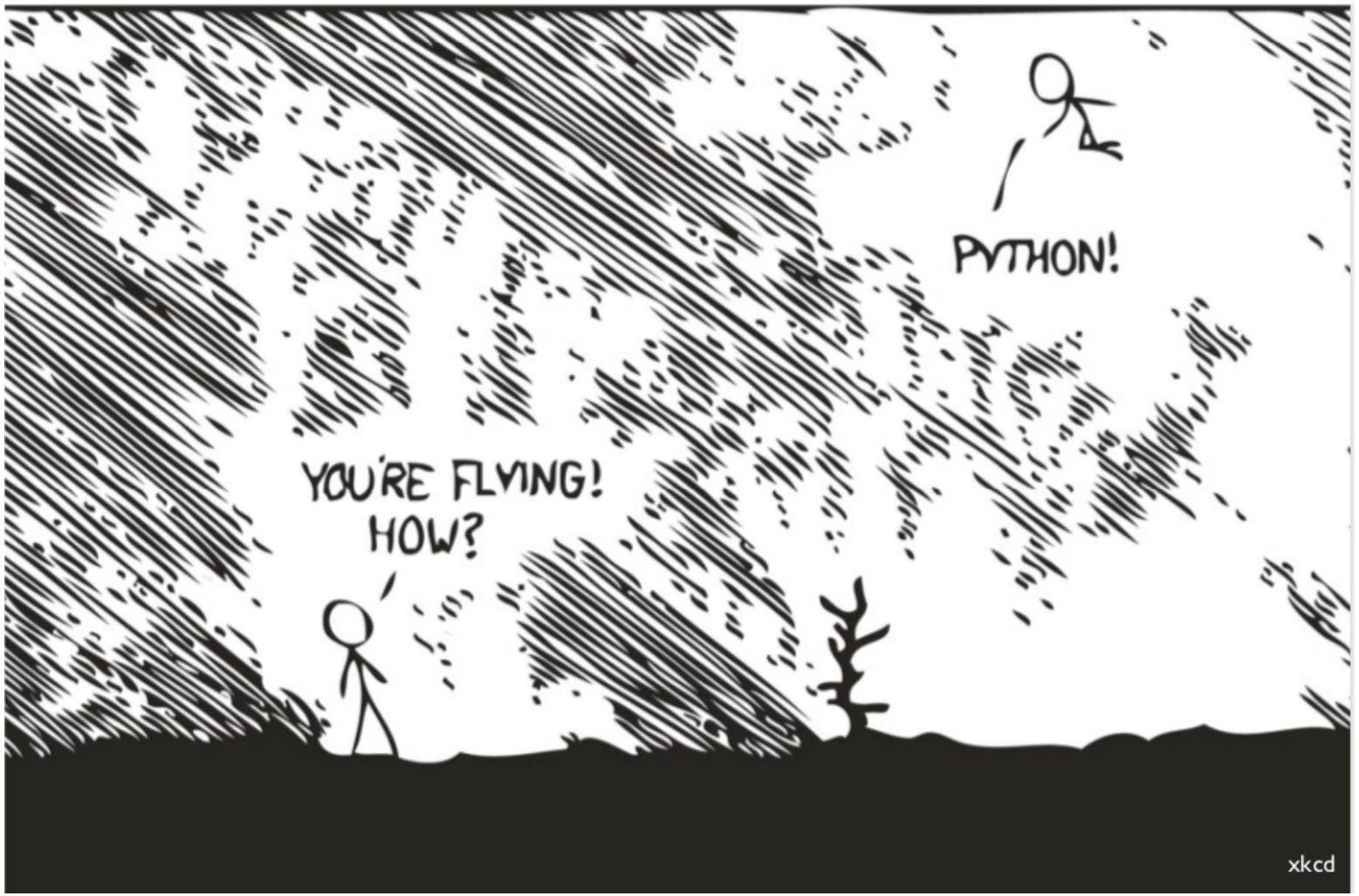
Data Science Process

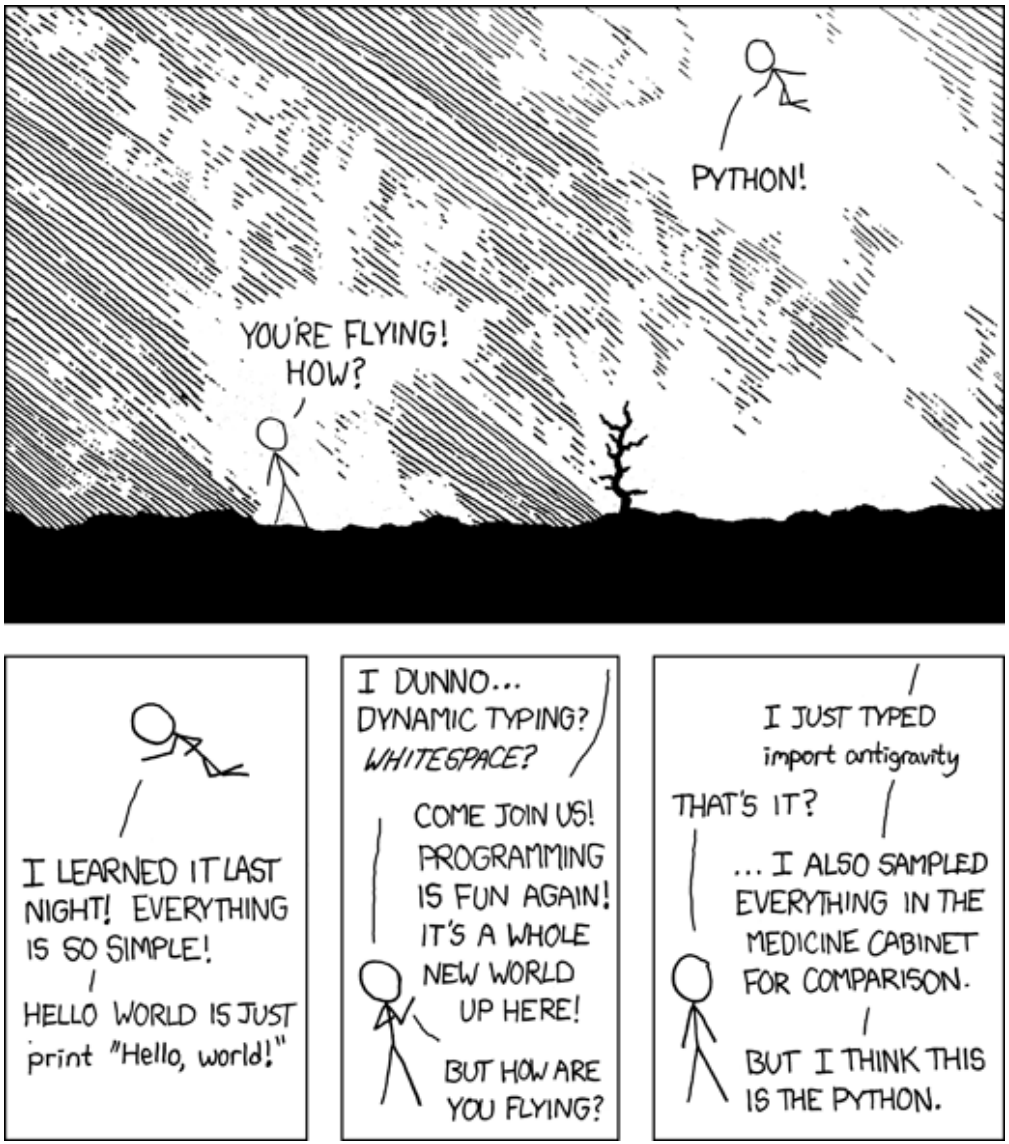


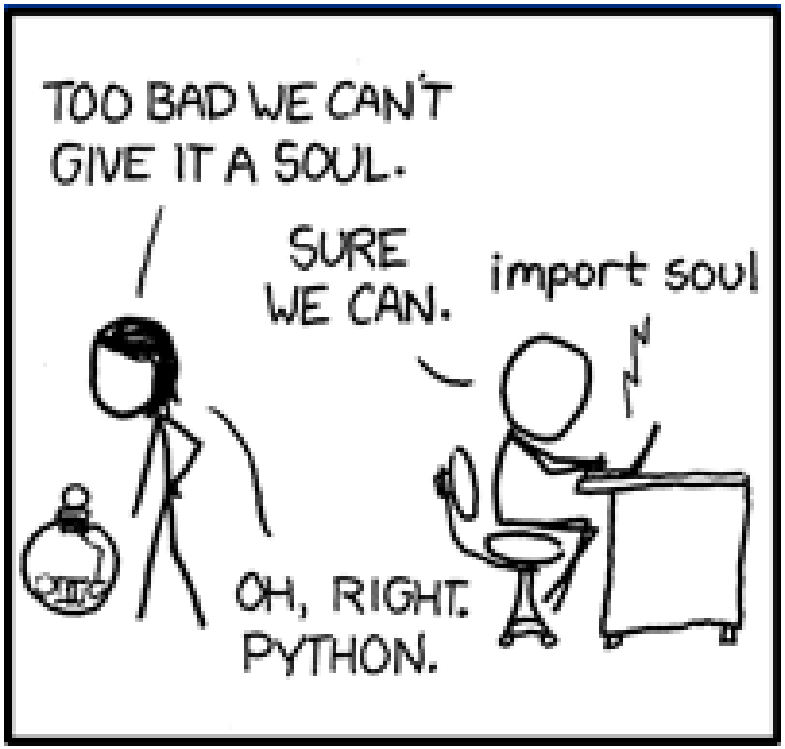
Topics Covered

- Data collection, visualization
- Exploratory data analysis
- Classifiers and Ensembles
- Clustering
- Search engine basics
- Recommender basics
- Lot of useful tools: dimensionality reduction, feature selection, hypothesis testing, sampling etc

Scientific Python


PYTHON!
YOU'RE FLYING!
HOW?
xkcd

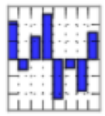
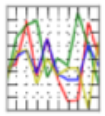
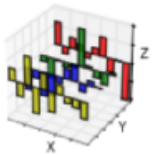

PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.


TOO BAD WE CAN'T GIVE IT A SOUL.
SURE WE CAN.
import soul
OH, RIGHT. PYTHON.

IP[y]: IPython
Interactive Computing

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikits learn
machine learning in Python

NumPy

SciPy.org
Sponsored By ENTHOUGHT

matplotlib

mrjob

IPython Notebooks

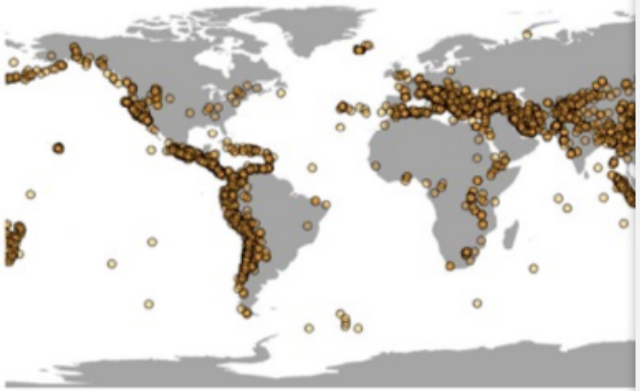
Process Books

[Blake Walsh, Gabriel Trevino, Antony Bett, CS171, 2013]

More Data Cleaning and Automation of Point Plotting in Prototype
[entry by btwalsh, 04.01]

Today, now that I have confirmed that my implementation of a few sample points is functional, I took our XLS file of earthquake data and began to clean it up some more with Google Refine specifically for the purpose of easily integrating it with our existing Processi the cleaning that was performed, I renamed the columns to read more sens entries that did not indicate a latitude and longitude for the location, and th rows of data that similarly did not actually report a magnitude.

Now, we have implemented a way to automatically map our latitudinal/longi onto our grid of x-y pixels, as demonstrated below:



[note by gtrevino 04.01]

Thankfully we got this to work (after dealing with some null pointer exceptio work on our data density, i.e. how to deal with multiple points overlapping, a especially important as we implement **user interaction**. This will probably r

[Varun Bansal, Cici Cao, Sofia Hou, CS171, 2013]

project process book

Users

The target audience is the general public, lovers of music, or simply those who want to learn about music. To appeal to the eclectic tastes of each individual, we have included nine main genres of music ranging from classical to country to Rap. This is not a visualization on the fundamentals of music theory so everyone has equal access to it - both for education and personal enjoyment!

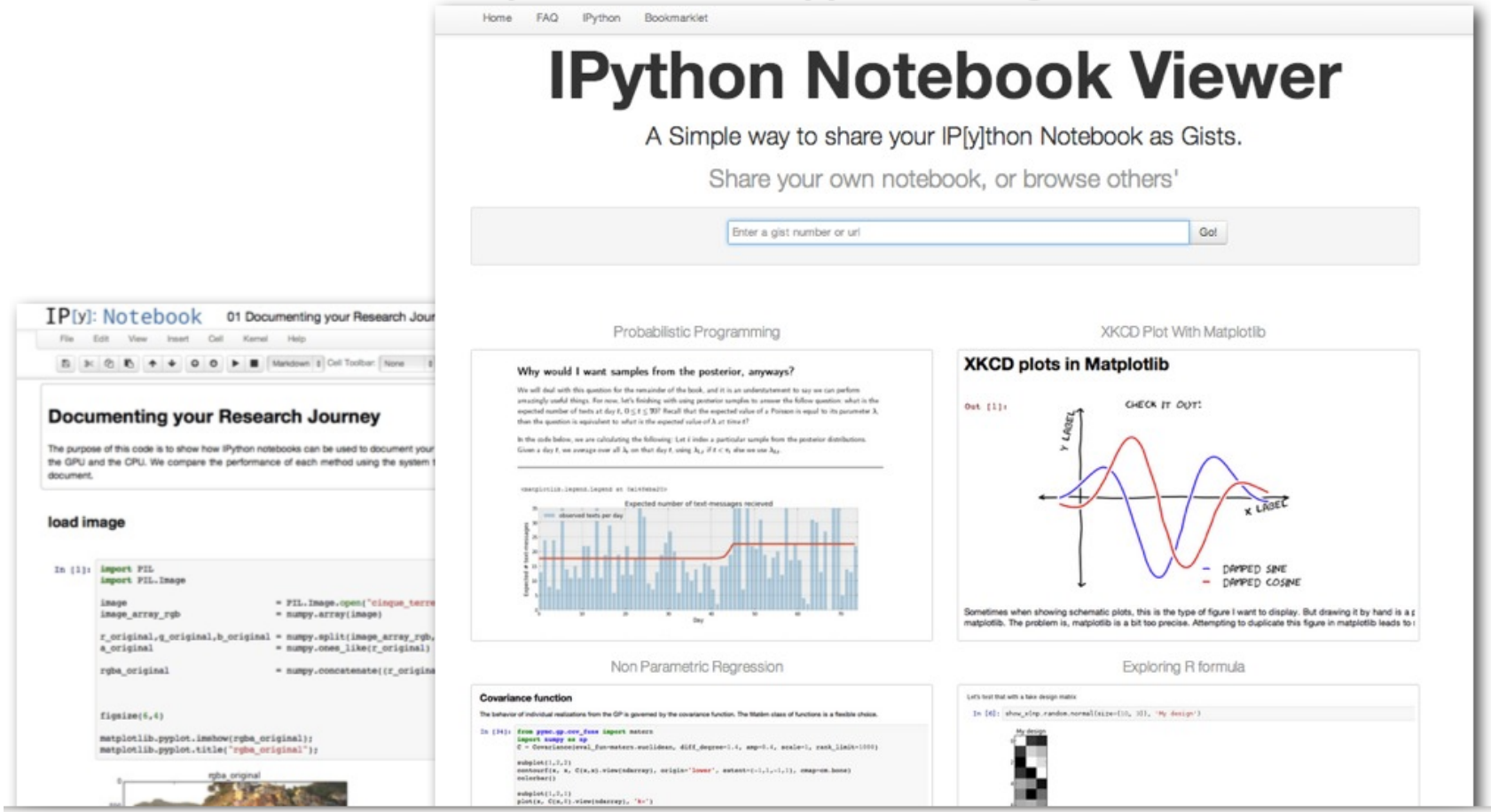
Related Work

Everyone loves music, and with the physics of sound the data on music is limitless. Each individual piece of song can have a vast amount of frequency, amplitude, and beats data. However, we were inspired to analyze music after seeing a project named the Shape of a Song (http://bewitched.com/song.html)in which the author visualized, through arch diagrams, similar patterns within different types of songs. This provided analysis on two levels: song and genre specific. On the song specific level, the arches showcased observations such as how the beginning of the song was similar to the end. On the genre specific level, comparing the arch diagrams of songs from different genres (i.e. modern techno or pop versus classical) shows characteristic patterns of each genre, with modern electronic and synthesized music displaying more repetition and similarity throughout the piece.

Chopin, Mazurka in F# Minor
The image illustrates the complex, nested structure of the piece.

IPython Notebooks

http://nbviewer.ipython.org/



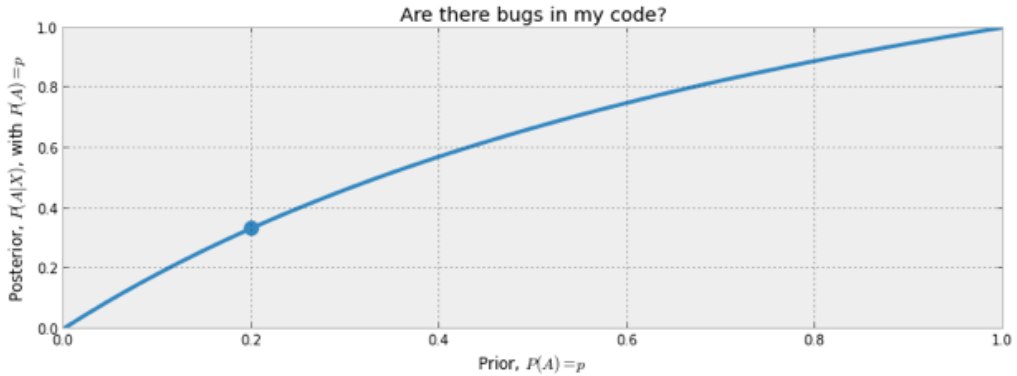
IPython is Great

(for large-scale computation, data exploration, and creating reproducible research artifacts)

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)}$$
$$= \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```



[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

Mike Roberts, Stanford University

~~You probably all know~~ the default Python interpreter.

Don't bother with

```
Last login: Mon May 20 17:53:32 on ttys000
dn0a2100e6:~ mike$ python
Enthought Python Distribution -- www.enthought.com
Version: 7.3-2 (32-bit)

Python 2.7.3 |EPD 7.3-2 (32-bit)| (default, Apr 12 2012, 11:28:34)
[GCC 4.0.1 (Apple Inc. build 5493)] on darwin
Type "credits", "demo" or "enthought" for more information.
>>> 2 + 2
4
>>>
```

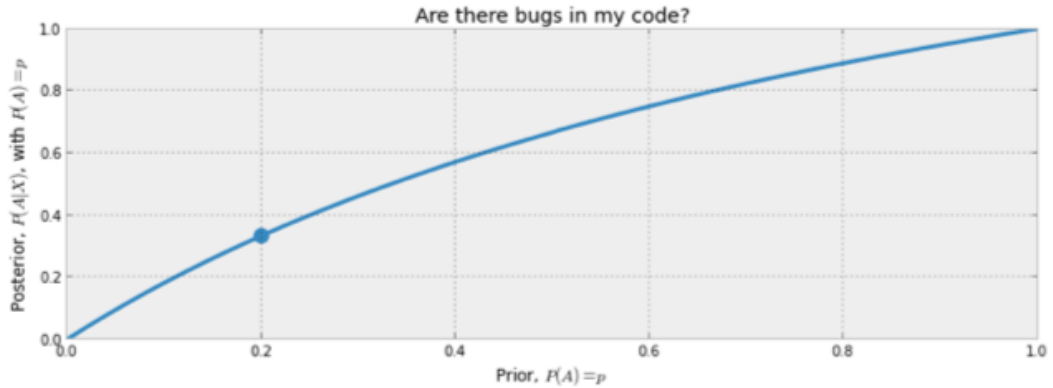
IPython is a more powerful interactive Python interpreter.

write and execute Python code in *snippets*

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)} = \frac{2p}{1+p}$$

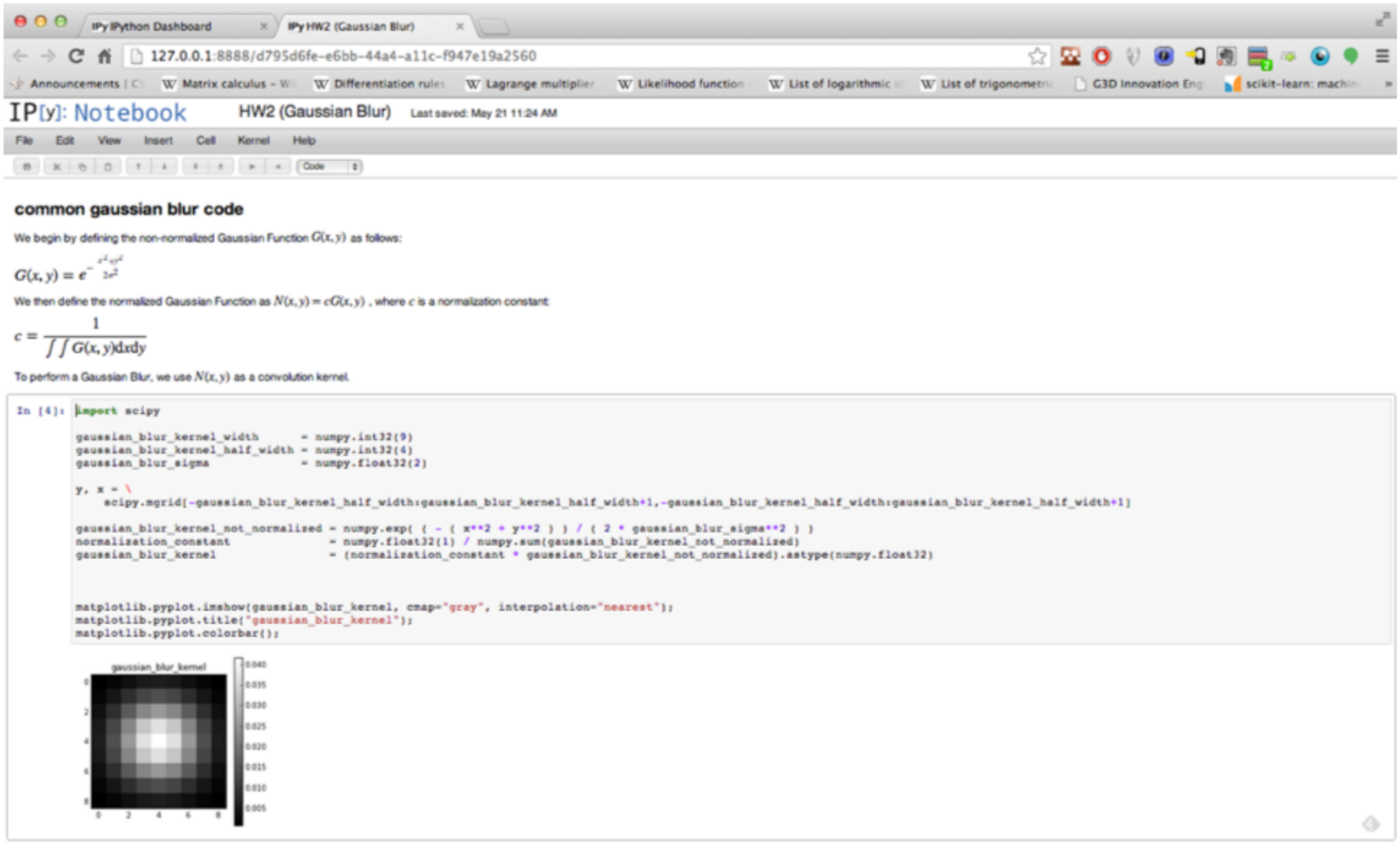
This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```

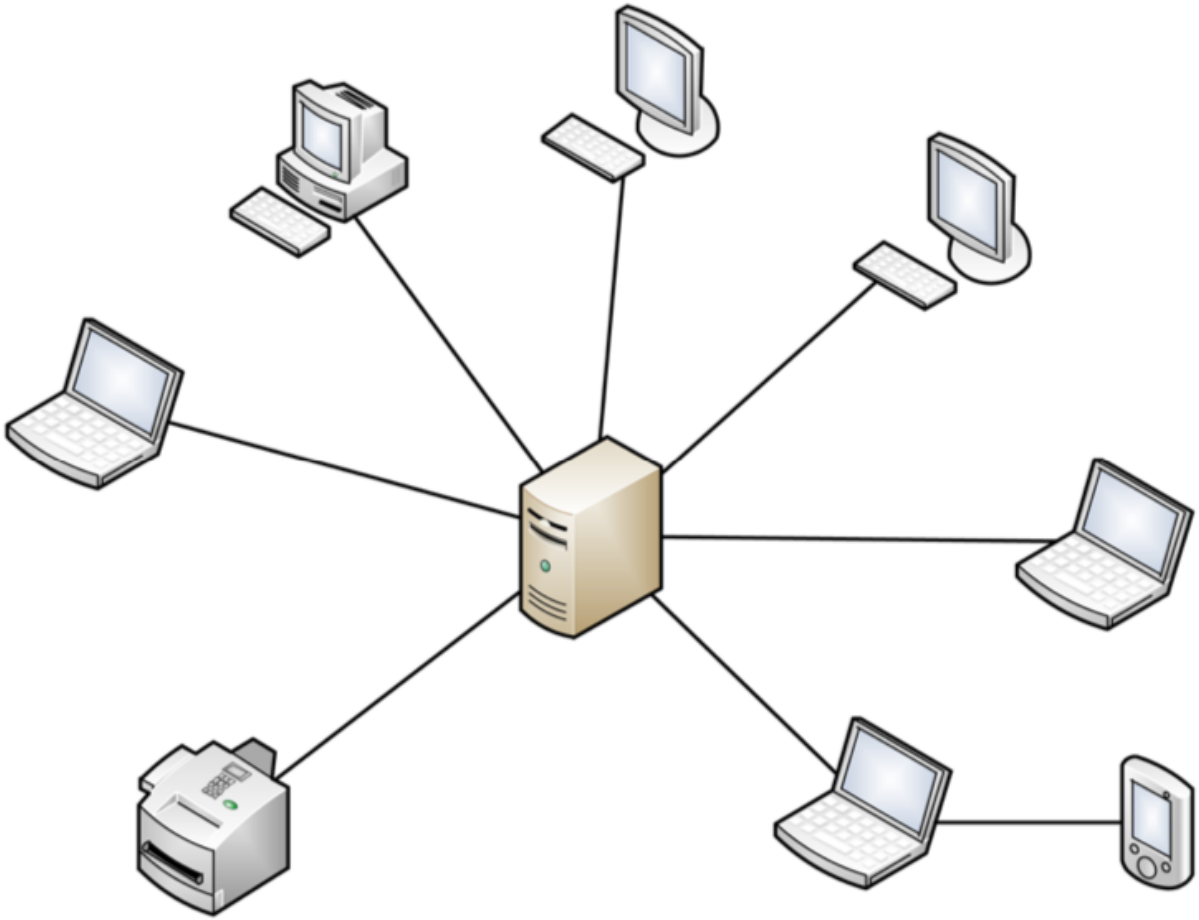


[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

Write Python code interactively in a web browser instead of a terminal window.

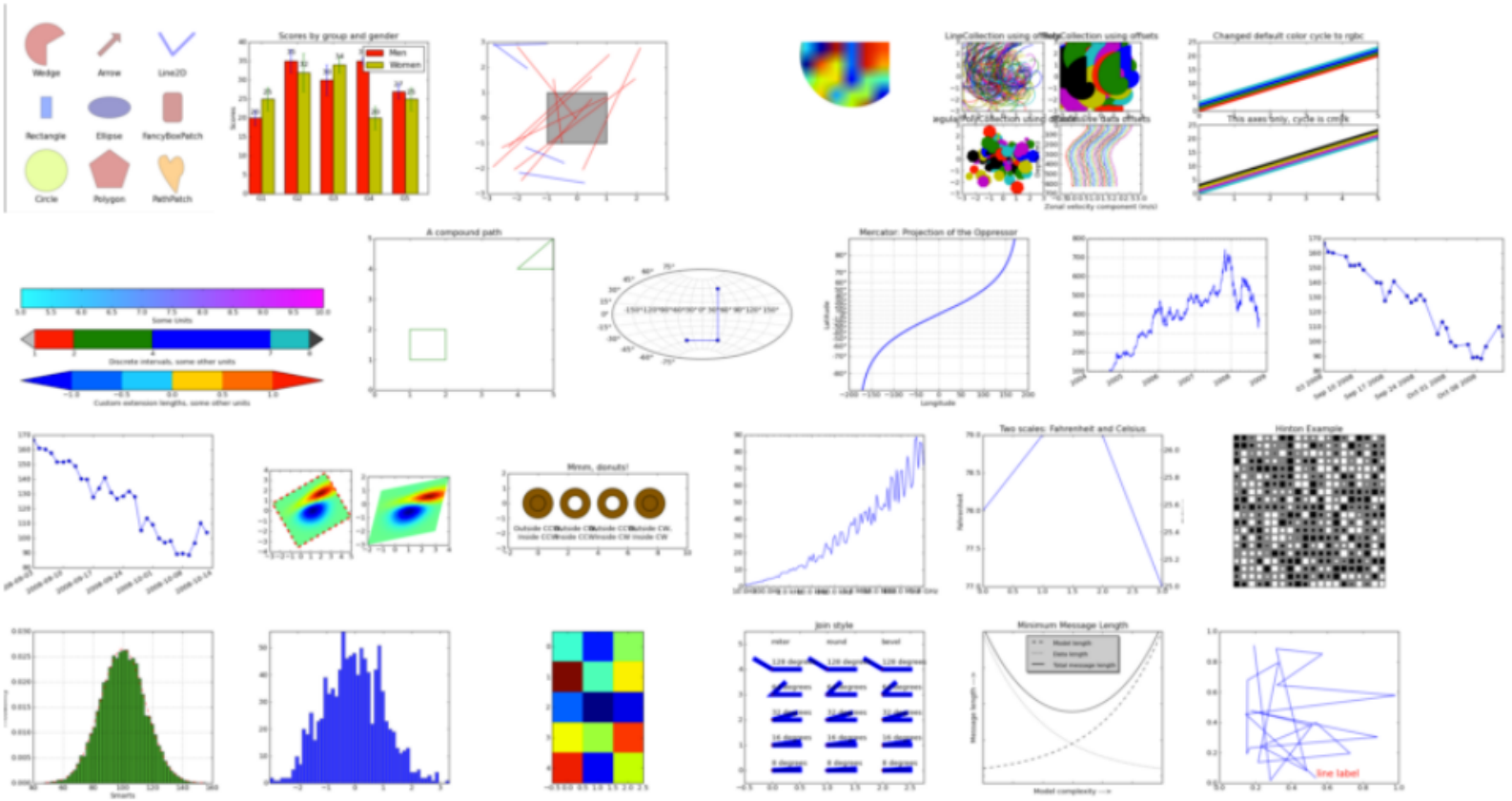


IPython has a decoupled client-server architecture.



[http://communities.intel.com/community/datastack/blog/2011/05/02/top-10-reasons-to-setup-a-client-server-network]

IPython integrates seamlessly with Matplotlib, making it well-suited for data exploration.



[http://matplotlib.org/gallery.html]

Slide Material References

- Slides from 'Introduction to Statistical Learning' by James, Witten, Hastie, and Tibshirani
- Slides from MMDS
- Slides from Harvard CS 109 (2013 and 2014)
- Slides from Dr.Ryan Tibshirani