

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 Data¹



¹<https://www.pinterest.com/pin/101753272804937744/>

Big Data²



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/>

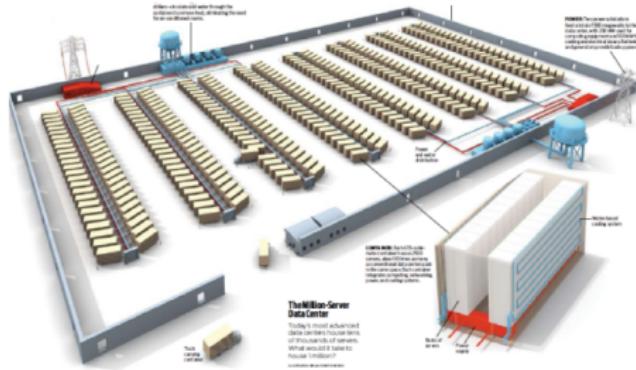
Smarter Devices



GarageBand '08



Commodity Computing



Michael Franklin, UC Berkeley

Ubiquitous Connectivity



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

2,500,000,000,000,000,000 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.



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



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



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.



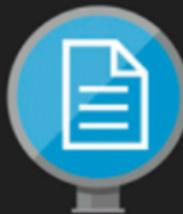
PURCHASE
TRANSACTIONS



WEBSITE
TRAFFIC



REWARDS
PROGRAMS



QUARTERLY
BUSINESS REPORTS



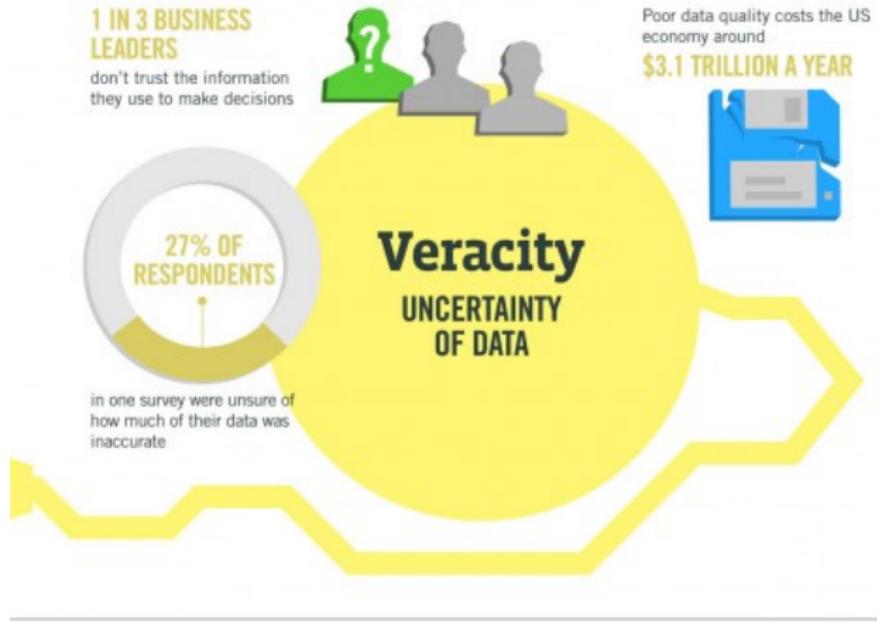
TWITTER



FACEBOOK



BLOG CONTENT



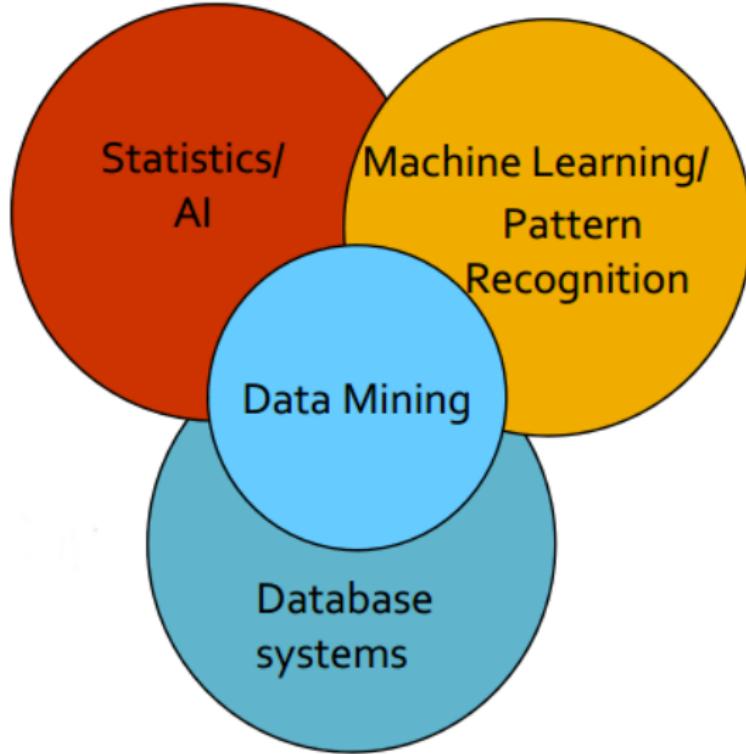
³<http://www.ibmbigdatahub.com/>



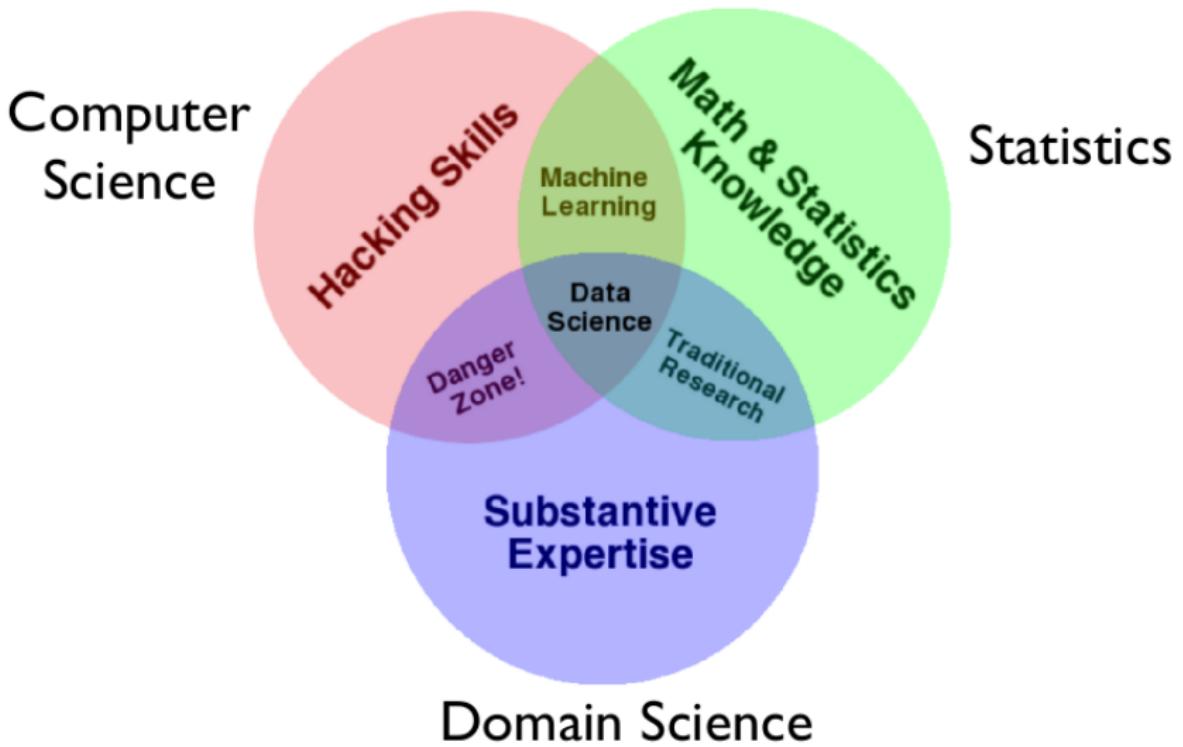
4

⁴<https://www.behance.net/gallery/5958295/Data-Hero-Oya-Group>

- 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



- 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



Drew Conway

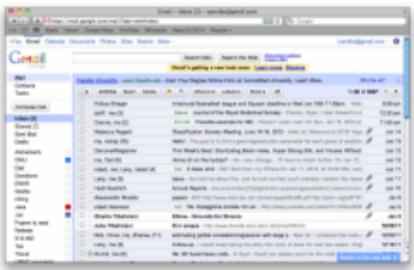
Google



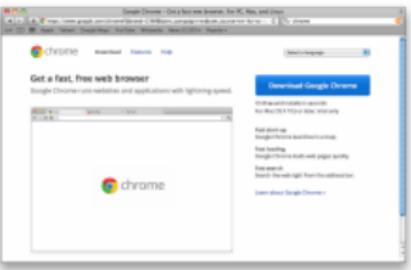
Search



Ads



Gmail



Chrome

Facebook

Find Your Friends on Facebook

http://www.facebook.com/find-friends

facebook Search

People You May Know See All

 **Edward Clapp**
Ira Hs and 3 other mutual friends **Add Friend**

 **Steve Carlson**
San Jose State University
Melody Kennedy and 28 other mutual friends **Add Friend**

 **Alexander Koller**
Hanover, Germany
Peter Meyer and 13 other mutual friends **Add Friend**

 **Anke Heinen**
Sylvia Kraft and 15 other mutual friends **Add Friend**

 **Donald Feasel**
Melody Kennedy and 17 other mutual friends **Add Friend**

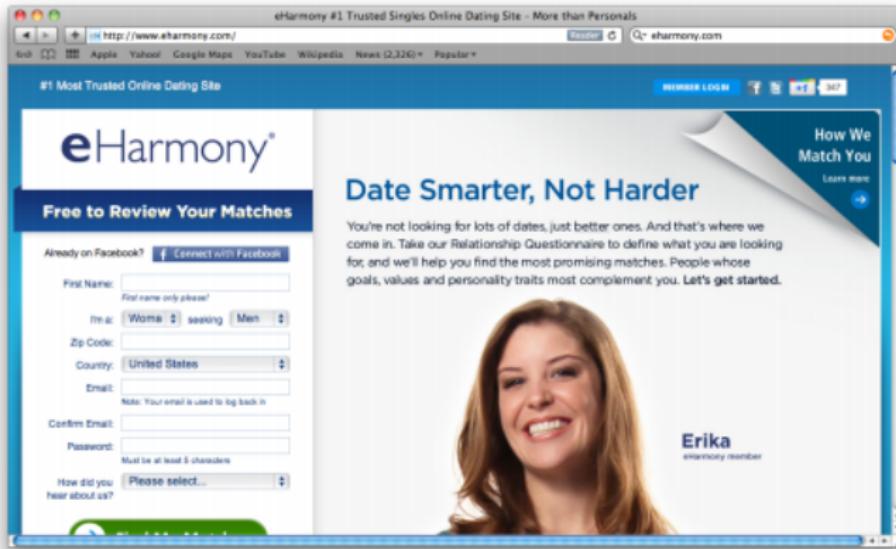
Chat (13)

People you may know

The screenshot shows a web browser window for the Netflix Prize Leaderboard. The URL is <http://www.netflixprize.com/leaderboard?showtest=t&limit=20>. The page has a red header with the Netflix logo and a yellow banner that says "Netflix Prize" and has a large red "COMPLETED" stamp. Below the banner, there's a navigation menu with links for Home, Rules, Leaderboard, and Update. The main content is a "Leaderboard" section with a table of results. The table has columns for Rank, Team Name, Best Test Score, % Improvement, and Best Submit Time. The top result is "BellKor's Pragmatic Chaos" with a score of 0.8987. The table also includes a note: "Grand Prize - RMSE = 0.8987 - Winning Team: BellKor's Pragmatic Chaos".

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
1	BellKor's Pragmatic Chaos	0.8987	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8987	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8982	9.90	2009-07-10 21:24:40
4	Open Solutions and Vendafone United	0.8988	9.84	2009-07-10 01:12:31
5	Xanthios Industries	0.8991	9.81	2009-07-10 00:32:20
6	Pragmatic Theory	0.8994	9.77	2009-06-24 12:08:56
7	Be Better in BigChase	0.8901	9.70	2009-05-13 08:14:09
8	Clock...	0.8912	9.59	2009-07-24 17:18:43
9	Ensemble	0.8922	9.48	2009-07-12 13:11:51
10	BigChase	0.8923	9.47	2009-04-07 12:33:59
11	Open Readers	0.8923	9.47	2009-07-24 00:34:07
12	Reactor	0.8924	9.45	2009-07-26 17:18:11

\$1M prize!



Falling in love with statistics

The screenshot shows the FICO website's homepage. At the top, there's a navigation bar with links for myFICO.com, Support, Log In, Contact Us, and English. Below the navigation is a large banner featuring a bar chart titled "Credit Risk Forecast" with the subtitle "US risk managers see troubles in student loans, housing". The chart shows various levels of delinquency across different categories. To the right of the banner are four promotional boxes: "Credit Forecast US risk managers see trouble", "Analytic Webinars New webinars and explorations", "FICO Model Central Monitor Model Performance", and "FICO World 2011 Download presentations". On the left side, there's a sidebar with a "Looking for your FICO® credit score?" section containing a "Go to myFICO.com" button and a "Credit Score Summary" card showing a score of 723. Below this are three news sections: "Newsroom", "Events", and "Videos". The "Newsroom" section has an article about FICO Data Analysis Shows Counterfeit Card Fraud Has Fallen Sharply in Europe. The "Events" section lists webinars for January 2012. The "Videos" section features a video player with the title "Managing Predictive Models".

An algorithm that could cause a lot of grief

FlightCaster

The screenshot shows a web browser window for FlightCaster. The URL bar displays "Http://Flightcaster.com/". The main content area features the FlightCaster logo at the top left. To the right of the logo is a navigation menu with links: Home (which is highlighted), About, Sample, Blog, FAQ, and Contact. Below the menu, a large section is titled "Flight Delay Prediction" with the subtext "6 hours before airline alerts". This section includes a small screenshot of a mobile application interface showing flight details and a red button labeled "Probably Delayed". Below this is a blue button labeled "See a sample prediction". To the right of this main section is a search form with fields for "Airline", "Number", and "Date" (set to "Today Sat Jan 19 2013"), and a red "FlightCast It!" button. On the far left edge of the main content area, there is a vertical black sidebar with the word "Feedback" written on it.

FlightCaster

Home About Sample Blog FAQ Contact

Flight Delay Prediction
6 hours before airline alerts

Probably Delayed

3% On Time
14% Late Few Min
83% Very Late

See a sample prediction

Feedback

By Flight No By Route

Airline _____

Number _____

Date Today Sat Jan 19 2013

FlightCast It!

Get the apps!

Available on the App Store

What people are saying...

THE WALL STREET JOURNAL

...frequent fliers can know to rebook flights earlier and occasional fliers can

Travel Delays De-Mystified

Why shouldn't I rely entirely on airlines or other alert systems?

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 Books⁵



**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

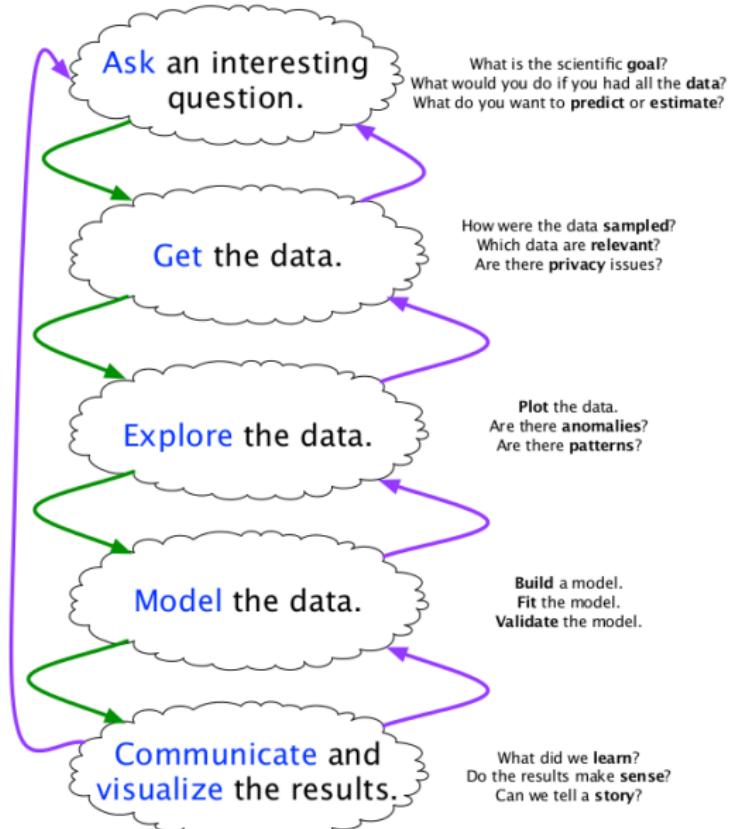
Late Days

- 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!

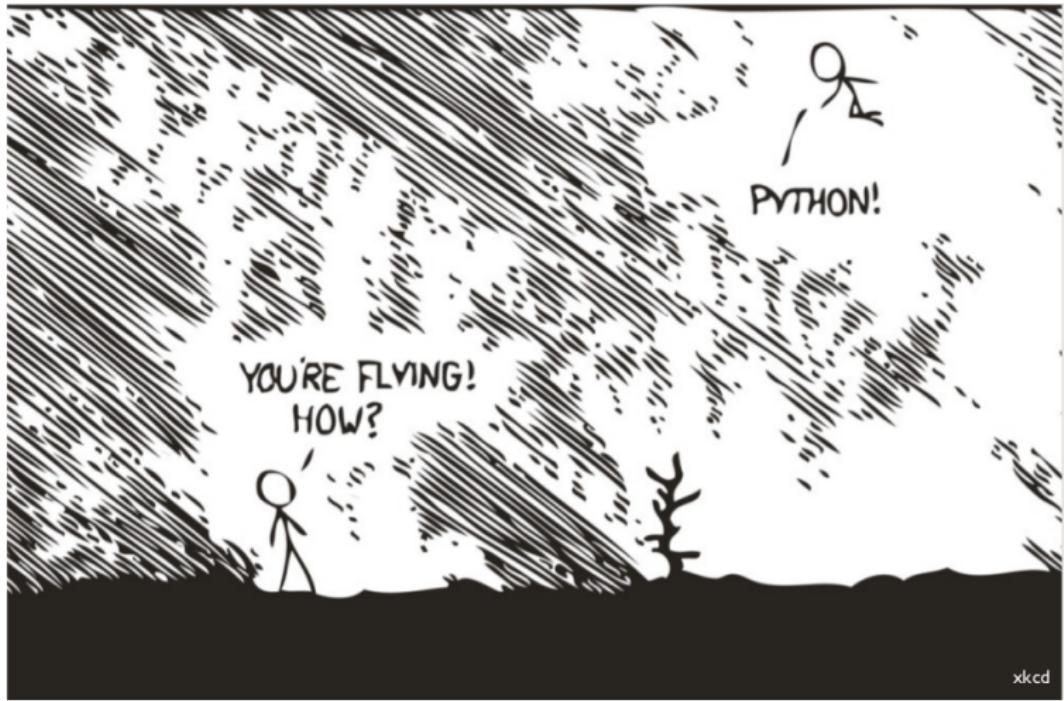
Data Science Process



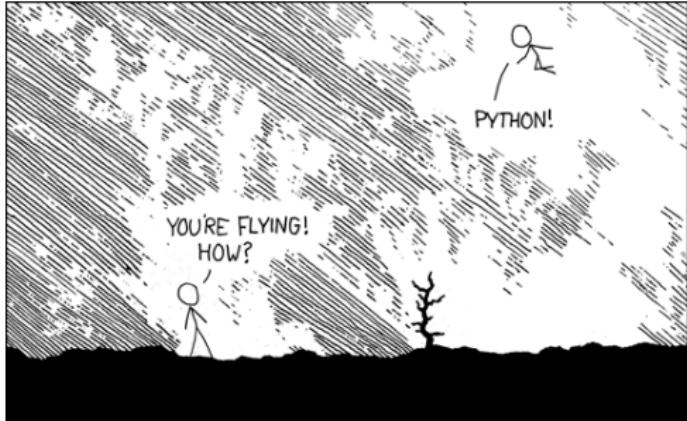
Topics Covered

- 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



xkcd



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.

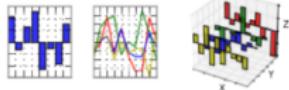


IP[y]: IPython

Interactive Computing

pandas

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



IPython Notebooks

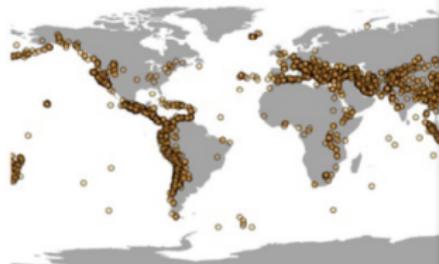
Process Books

[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 Processing code. In the cleaning that was performed, I renamed the columns to read more sensibly, removed entries that did not indicate a latitude and longitude for the location, and then removed rows of data that similarly did not actually report a magnitude.

Now, we have implemented a way to automatically map our latitudinal/longitudinal coordinates 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 exceptions) and work on our data density, i.e. how to deal with multiple points overlapping, is 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!

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/>

Home FAQ IPython Bookmarklet

IPython Notebook Viewer

A Simple way to share your IP[y]thon Notebook as Gists.

Share your own notebook, or browse others¹

Go!

IP[y]: Notebook 01 Documenting your Research Journey

File Edit View Insert Cell Kernel Help

Documenting your Research Journey

The purpose of this code is to show how IPython notebooks can be used to document your GPU and the CPU. We compare the performance of each method using the system I document.

load image

```
In [1]: import PIL
import PIL.Image

image = PIL.Image.open("cigar_teaser.jpg")
image_array_rgb = np.array(image)
x_original,y_original,h_original = np.split(image_array_rgb,
x_original,y_original,h_original)
rgb_original = np.concatenate([x_original,y_original])

dimpixel(4,4)

#resampling
image_pillow_ppm = Image.fromarray(rgb_original)
image_pillow_ppm.save("cigar_original.ppm")
```

cigar original

Probabilistic Programming

Why would I want samples from the posterior, anyways?

We will deal with this question for the remainder of the book, and it is understandable to us on an intuitive level what we are doing. For now, let's finish with using posterior samples to answer the following question: what is the probability that the mean of a set of 81 i.i.d. N(0, 1) RVs would exceed a value of 1? Posterior is equal to its posterior if the question is asked about the mean of 81 i.i.d. N(0, 1) RVs.

In the code below, we are calculating the following: Let's take a particular sample from the posterior distribution. Given a day t , we average over all t_i on that day t , using $\bar{y}_{t,i} = \bar{y}_t + t_i$, where we use \bar{y}_0 .

XKCD Plot With Matplotlib

Dft [1]:

Sometimes when showing schematic plots, this is the type of figure I want to display. But drawing it by hand is a bit messy. The problem is, matplotlib is a bit too precise. Attempting to duplicate the figure in matplotlib leads to:

Non Parametric Regression

Covariance functions

The behavior of individual variables from the GP is governed by the covariance function. The Matern class of functions is a flexible choice.

```
In [2]: from gpr import GaussianProcess
import numpy as np
from scipy import stats
import timeit
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.gaussian_process.kernels import Matern, RBF, WhiteKernel, ConstantKernel as C
```

Exploring R formula

Lets first try to do a basic design matrix

```
In [3]: def rbf_kernel(X, X2):
    return np.exp(-np.sum((X - X2)**2, axis=1) / 2)
```

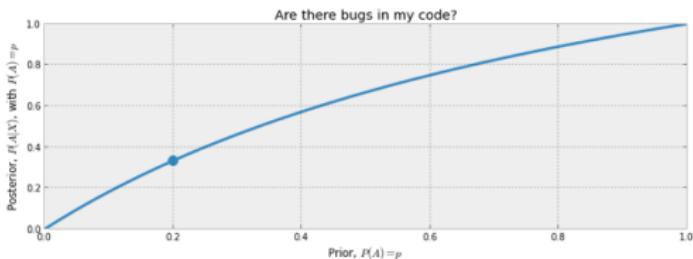
I Python is Great

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

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

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.jupyter.org/url/raw.github.com/ComDavidsonPilon/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.

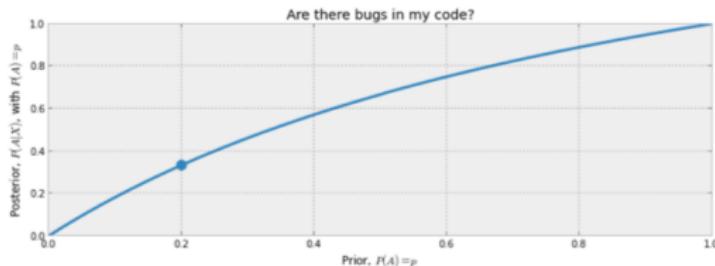
write and execute
Python code in
snippets



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

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.jupyter.org/url/raw.githubusercontent.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter%201%20-Introduction/Chapter%201%20-Introduction.ipynb>

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

The screenshot shows a Jupyter Notebook interface running in a web browser. The title bar indicates the notebook is titled "IP[y]: Notebook HW2 (Gaussian Blur)". The notebook has tabs for "Announcements", "Matrix calculus", "Differentiation rules", "Lagrange multiplier", "Likelihood function", "List of logarithmic", "List of trigonometric", "G3D Innovation Eng", and "sklearn".

common gaussian blur code

We begin by defining the non-normalized Gaussian Function $G(x,y)$ as follows:

$$G(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$$

We then define the normalized Gaussian Function as $N(x,y) = cG(x,y)$, where c is a normalization constant.

$$c = \frac{1}{\int \int G(x,y) dx dy}$$

To perform a Gaussian Blur, we use $N(x,y)$ as a convolution kernel.

```
In [4]: %matplotlib inline
import scipy
gaussian.blur_kernel_width      = numpy.int32(9)
gaussian.blur_kernel_half_width = numpy.int32(4)
gaussian.blur_sigma            = numpy.float32(2)

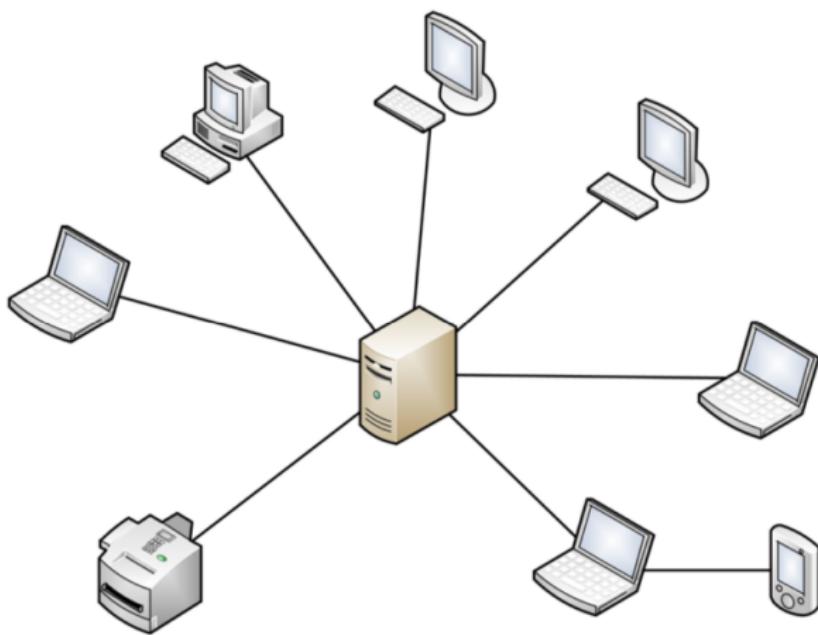
Y, X = np.mgrid[-gaussian.blur_kernel_half_width:gaussian.blur_kernel_half_width+1,-gaussian.blur_kernel_half_width:gaussian.blur_kernel_half_width+1]

gaussian.blur_kernel_normalized = numpy.exp( { - ( x**2 + y**2 ) } / ( 2 * gaussian.blur_sigma**2 ) )
normalization_constant          = numpy.float32(1) / numpy.sum(gaussian.blur_kernel_normalized)
gaussian.blur_kernel           = (normalization_constant * gaussian.blur_kernel_normalized).astype(numpy.float32)

matplotlib.pyplot.imshow(gaussian.blur_kernel, cmap="gray", interpolation="nearest");
matplotlib.pyplot.title("gaussian.blur_kernel");
matplotlib.pyplot.colorbar();
```

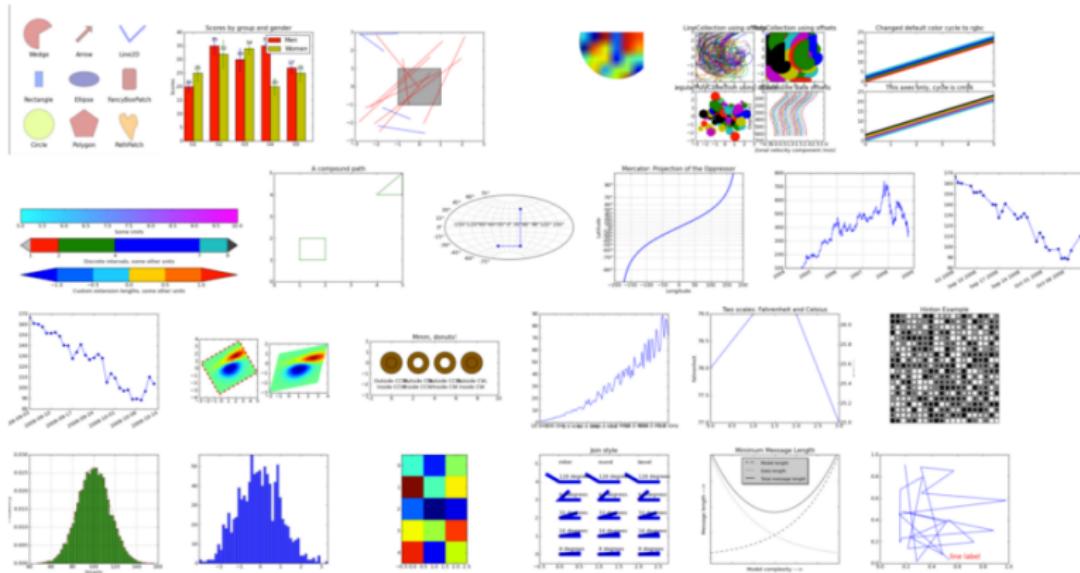
A 9x9 grayscale heatmap visualization of the Gaussian blur kernel is shown, with values ranging from 0.000 to 0.040. The plot has axes labeled from -4 to 4.

IPython has a decoupled client-server architecture.



[\[http://communities.intel.com/community/datasdk/blog/2011/05/02/exp-10-reasons-to-setup-a-client-server-network\]](http://communities.intel.com/community/datasdk/blog/2011/05/02/exp-10-reasons-to-setup-a-client-server-network)

I Python 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