

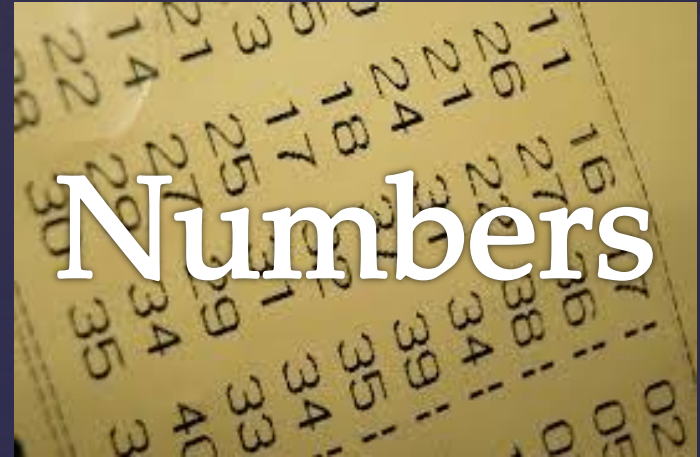
What is Data Science?

{ Data, Databases, and the Extraction of Knowledge
{ Renée T., @becomingdatasci, November 2014

Let's start with: "What is Data?"



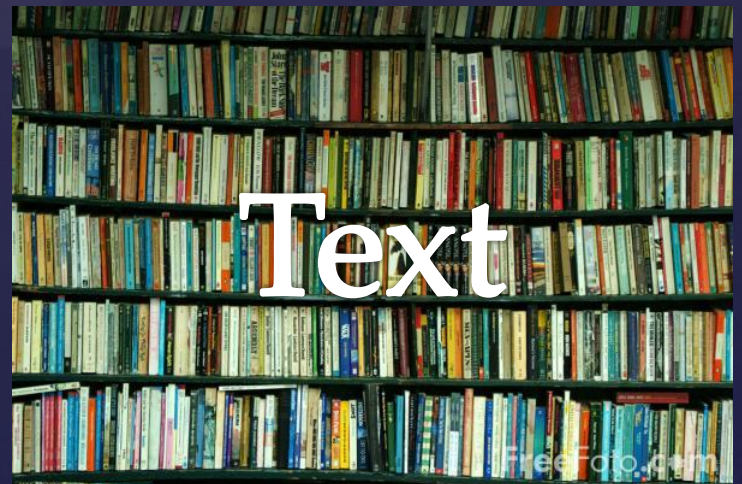
http://upload.wikimedia.org/wikipedia/commons/f/f0/DARPA_Big_Data.jpg



https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcS9dKu3_Tzi-sWW-yAqee5y0EhuvolZNSya_rAKnuBBd0JYxPX7pw



http://fc01.deviantart.net/fs71/i/2012/326/3/4/cute_dog_by_tho_masmeadows345-d5lsah9.jpg

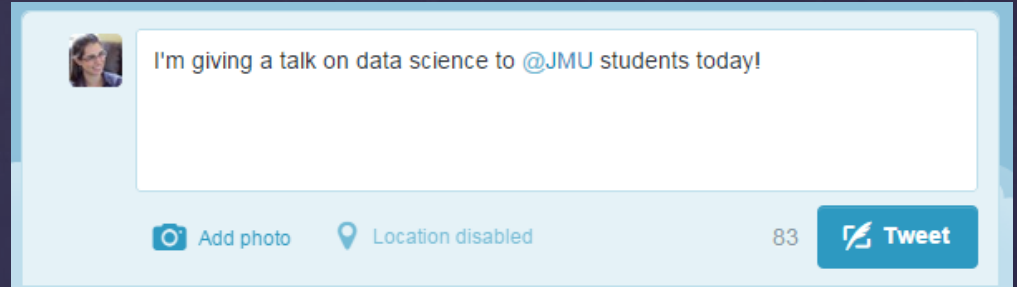


http://www.freefoto.com/images/1351/06/1351_06_2--Books--Shakespeare-and-Company-Bookstore--The-Latin-Quarter--Paris_web.jpg

Created & Collected



http://upload.wikimedia.org/wikipedia/commons/9/96/Bill_Nye,_Barack_Obama_and_Neil_deGrasse_Tyson_selfie_2014.jpg



http://upload.wikimedia.org/wikipedia/commons/e/e4/Green_Bank_100m_diameter_Radio_Telescope.jpg



https://c1.staticflickr.com/1/2/1349370_0703fce74c.jpg



https://c2.staticflickr.com/4/3273/3017878633_65beb1c7d6.jpg

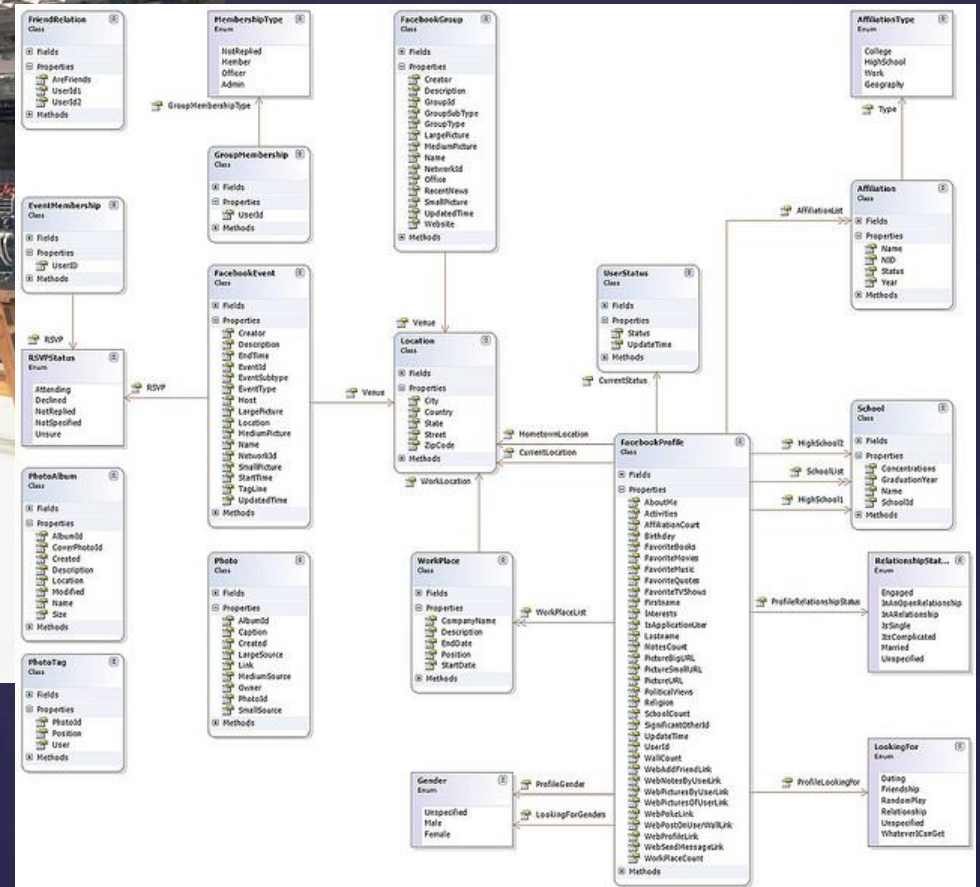
© NSW DPI

- ⌘ Around **100 hours of video** are uploaded to YouTube **every minute**
 - ⌘ it would take about 15 years to watch every video uploaded in one day
- ⌘ AT&T is thought to hold the world's largest volume of data in one unique database – its **phone records** database is 312 terabytes in size, and contains almost **2 trillion** rows.
- ⌘ **Every minute** we send 204,000,000 emails, generate 1,800,000 Facebook likes, send 278,000 Tweets, and up-load 200,000 photos to Facebook
- ⌘ 570 new websites spring into existence every minute of every day.

<http://smartdatacollective.com/bernardmarr/277731/big-data-25-facts-everyone-needs-know>

“Big Data”

Stored & Related



http://pixabay.com/static/uploads/photo/2014/03/13/01/12/datacenter-286386_640.jpg



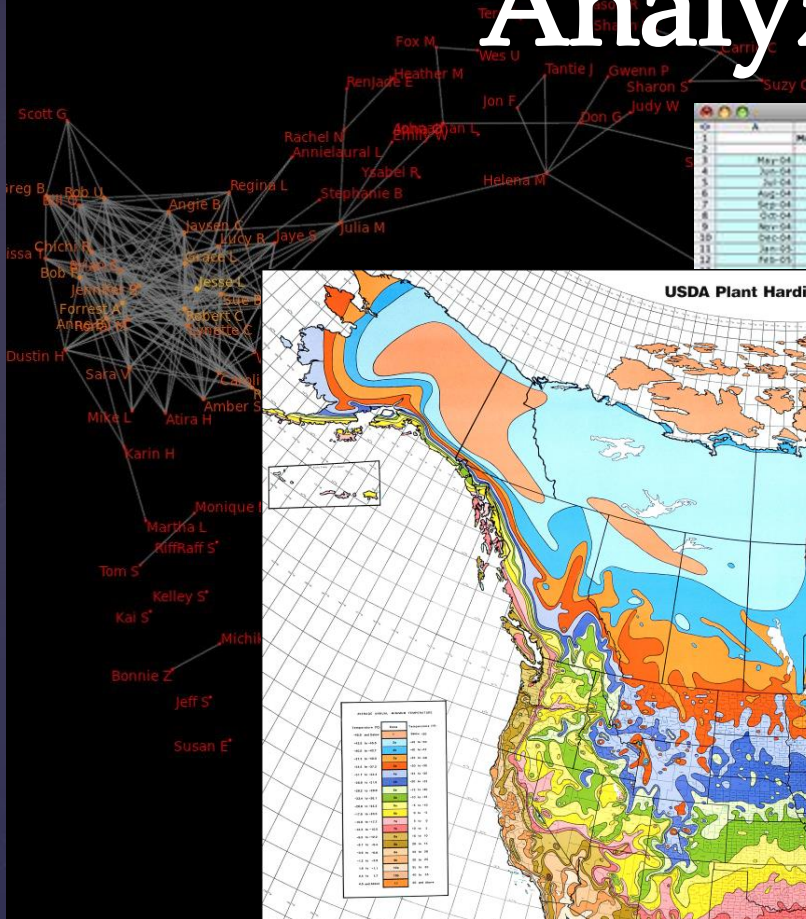
Video clip:

<http://youtu.be/PBx7rgqeGG8?t=2m>

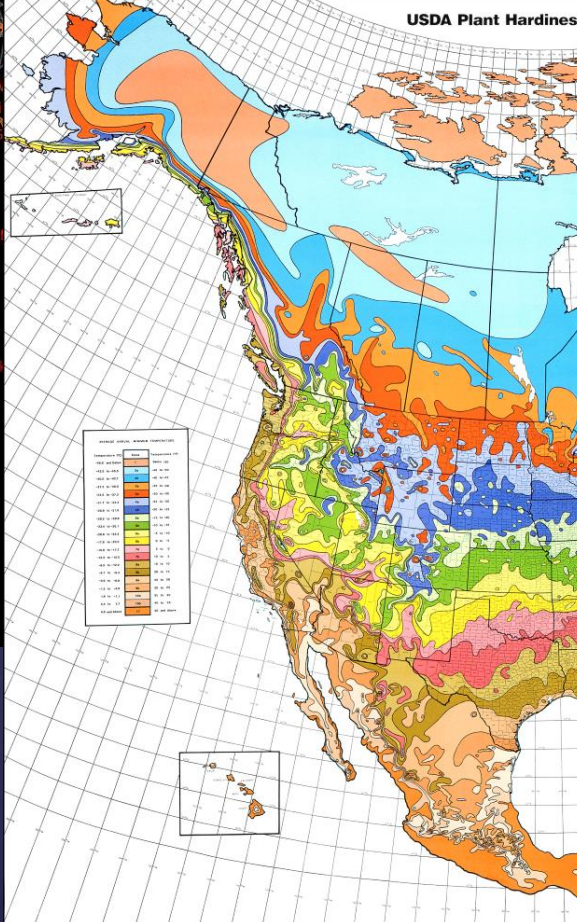
https://c2.staticflickr.com/2/1296/533233247_b6baa30fdb_z.jpg?zz=1

Analyzed and Visualized

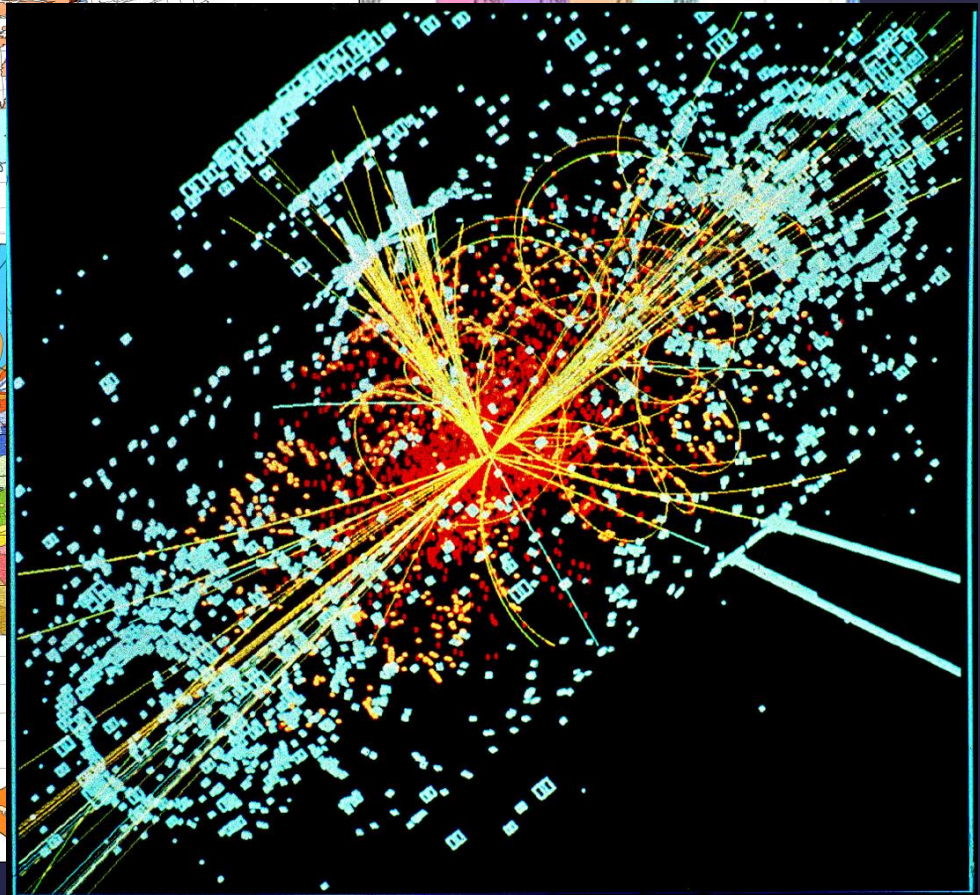
https://c1.staticflickr.com/3/2300/2596366618_2d6cb01735.jpg



USDA Plant Hardiness Zone Map



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		Mar	3p	PAID	Matts Total	Totals		Items	Cost	PAID	To Pay	Date to pay		
3	1	May-04	£ 2,100	£ 1,200	£ 997	£ 2,100	43,300	Wetnet	£ 1,300	£ 1,300	£ 0	PAID		
4	2	Jun-04	£ 2,000	£ 800	£ 0	£ 2,000	64,200	Food dry	£ 1,511	£ 1,511	£ 0	PAID		
5	3	Jul-04	£ 500	£ 450	£ 37	£ 1,000	45,300	Food wetting	£ 1,160	£ 1,160	£ 0	PAID		
6	4	Aug-04	£ 1,000	£ 0	£ 0	£ 1,000	45,100	Suits	£ 360	£ 360	£ 0	PAID		
7	5	Sep-04	£ 800	£ 400	£ 0	£ 3,900	48,300	Wine = por drinks	£ 451	£ 451	£ 0	PAID		
8	6	Oct-04	£ 1,550	£ 800	£ 0	£ 5,550	48,750	Champagne	£ 139	£ 139	£ 0	PAID		
9	7	Nov-04	£ 100	£ 0	£ 0	£ 3,650	48,650	Dress	£ 250	£ 250	£ 0	PAID		
10	8	Dec-04	£ 150	£ 0	£ 0	£ 8,000	48,950	Flowers = Bridesmaids	£ 380	£ 380	£ 0	PAID		
11	9	Jan-05	£ 750	£ 0	£ 0	£ 8,700	49,650	Shoes = Accessories	£ 300	£ 300	£ 0	PAID		
12	10	Feb-05	£ 0	£ 0	£ 0	£ 8,700	49,650	Other clothes	£ 382	£ 382	£ 0	PAID		



<http://upload.wikimedia.org/wikipedia/commons/9/90/Ke ncf0618FacebookNetwork.jpg>

http://upload.wikimedia.org/wikipedia/commons/b/bf/USDA_Hardiness_zone_map.jpg

http://upload.wikimedia.org/wikipedia/commons/1/1c/CMS_Higgs-event.jpg

What is a database?

Database

[dey-tuh-beys]

noun

A comprehensive collection of related data organized for convenient access, generally in a computer.

-dictionary.com

Databases You Use

↳ Pretty much every website you interact with

↳ Social Media

↳ Banking

↳ File Sharing

↳ Search Engines

↳ Online Shopping

↳ Course Registration/Canvas

↳ Travel

↳ Etc. etc. etc.....

↳ You broadcast/generate data everywhere you go

↳ Cell phones

↳ Purchases

↳ Driving (GPS)

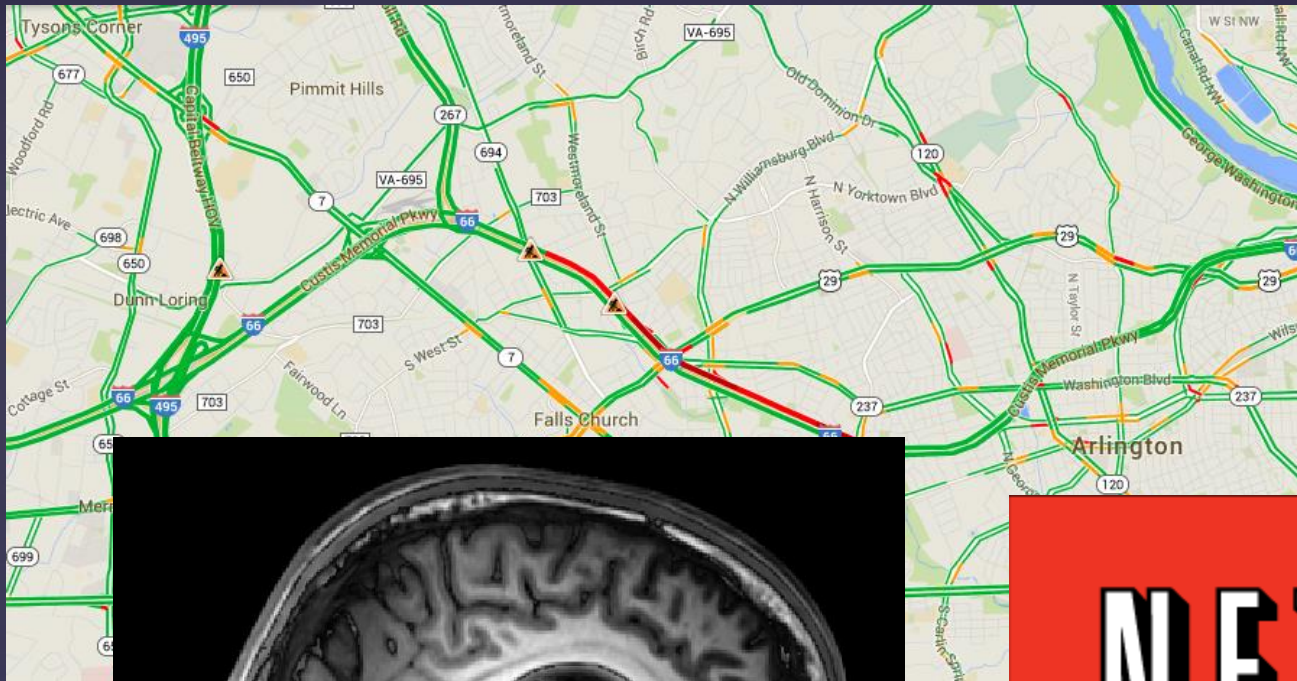
↳ Streaming music

↳ Email

↳ Posting status updates

↳ Attending events

↳ Etc. etc. etc.....



http://upload.wikimedia.org/wikipedia/commons/6/69/Netflix_logo.svg

How is data
collected about you
used to help you?

Who builds these systems?

Data Scientist

Computer Scientist

- Data collection systems
- Machine Learning Algorithms
- Interface Design
- Design/Manage/Query Databases
- Data Aggregation
- Data Mining

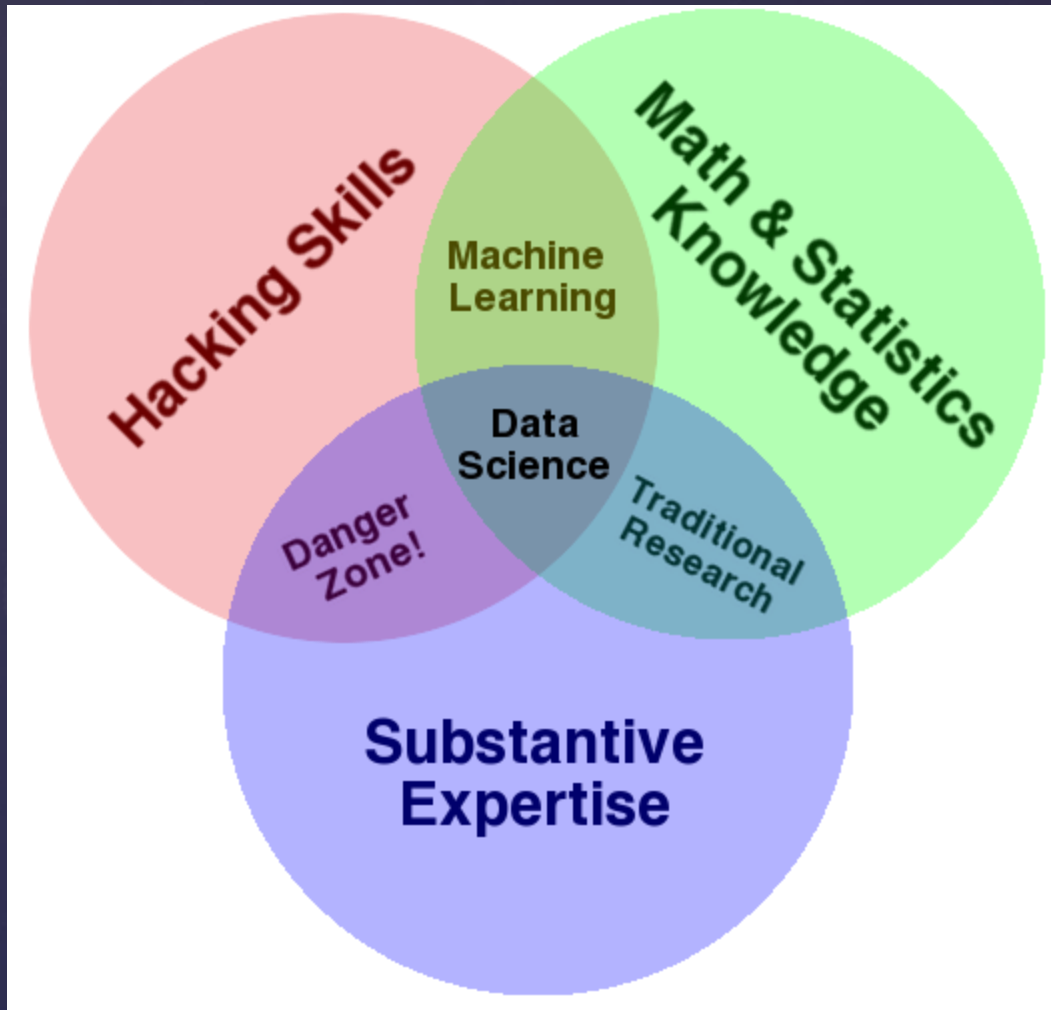
Mathematician

- Statistical Models
- Evaluation Metrics
- Predictive Analytics
- Data Visualizations

Business Person

- Domain Expertise
- Knowing what questions to ask
- Interpreting results for business decisions
- Presenting outcomes

Examples – not a complete definition, and not all simultaneously necessary skills



Data Science Venn Diagram by Drew Conway

http://static.squarespace.com/static/5150aec6e4b0e340ec52710a/t/51525c33e4b0b3e0d10f77ab/1364352052403/Data_Science_VD.png?format=750w

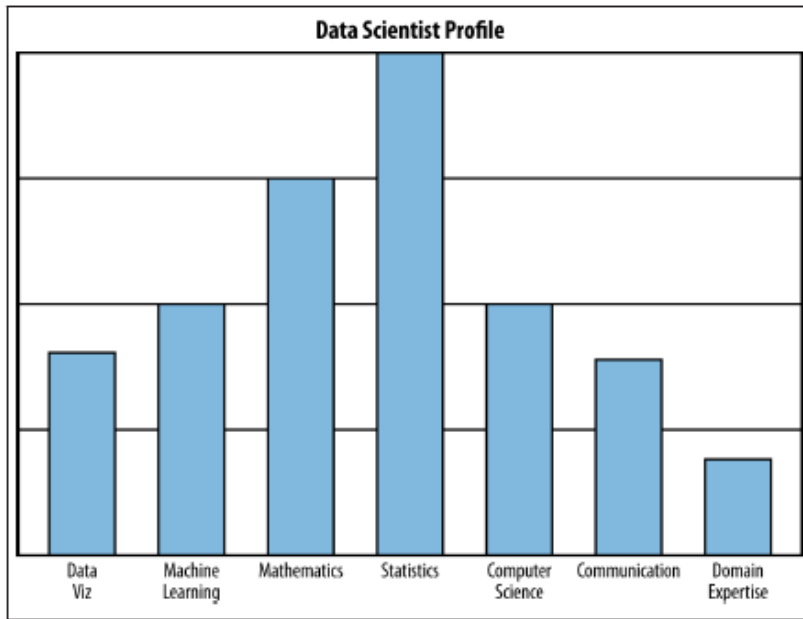


Figure 1-2. Rachel's data science profile, which she created to illustrate trying to visualize oneself as a data scientist; she wanted students and guest lecturers to "riff" on this—to add buckets or remove skills, use a different scale or visualization method, and think about the drawbacks of self-reporting

From "Doing Data Science" by Cathy O'Neill & Rachel Schutt

http://www.becomingadatascientist.com/wp-content/uploads/2014/06/DS_profile.png

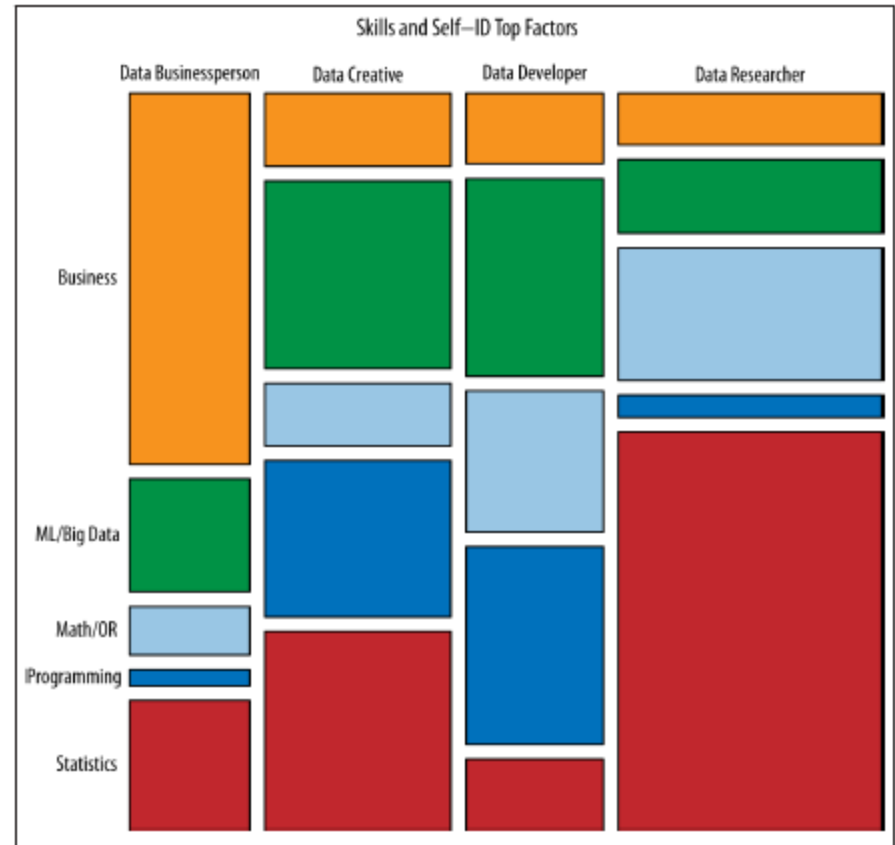


Figure 1-4. Harlan Harris's clustering and visualization of subfields of data science from *Analyzing the Analyzers* (O'Reilly) by Harlan Harris, Sean Murphy, and Marck Vaisman based on a survey of several hundred data science practitioners in mid-2012

<http://semanticcommunity.info/@api/deki/files/27057/Figure1-4.png?size=bestfit&width=484&height=541&revision=1>

No need to be a "unicorn", but do need to know something about all of these areas, and become expert in some (Sound familiar, ISAT students?)

Some other names for “Data Scientist”

⌘ Statistician

⌘ Data Mining Specialist

⌘ Biostatistician

⌘ Social Science Researcher

⌘ Big Data Analyst

⌘ Spatial/GIS Analyst

⌘ Natural Language

Programmer

⌘ Computational Physicist

⌘ Pythonista

⌘ Financial Analyst

⌘ Recommendation System

Engineer

⌘ Information Architect

⌘ Artificial Intelligence

Researcher

⌘ Neuroscientist

⌘ Data Visualization Designer

Data Science jobs pay an average of \$118,000 per year

It is estimated that by 2018, US could have a shortage of 140,000+ people with advanced analytical skills & need 1.5M managers/analysts that can make decisions based on data analysis

“Extraction of Knowledge”

⌘ Also known as “knowledge discovery”

⌘ Goes beyond queries

⌘ Data Mining

⌘ Business Understanding

⌘ Data Understanding

⌘ Data Preparation

⌘ Modeling

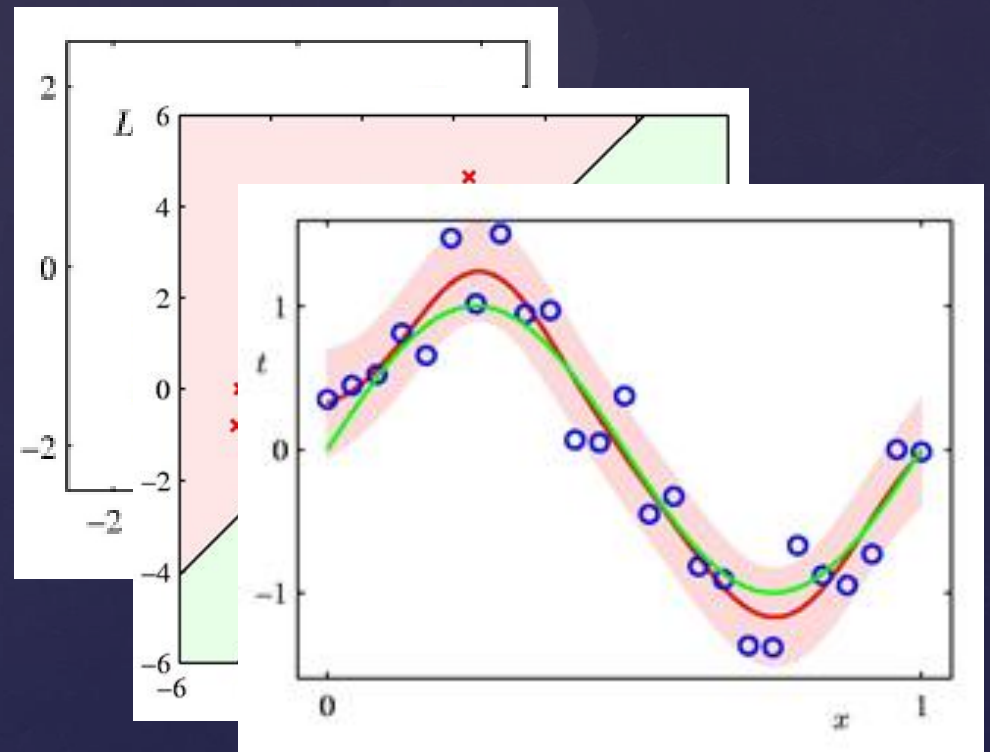
⌘ Clustering

⌘ Classification

⌘ Regression

⌘ Evaluation

⌘ From “Data Science for Business” by Provost & Fawcett



Images from ODU ECE 607 Lecture Slides by Prof. Jiang Li

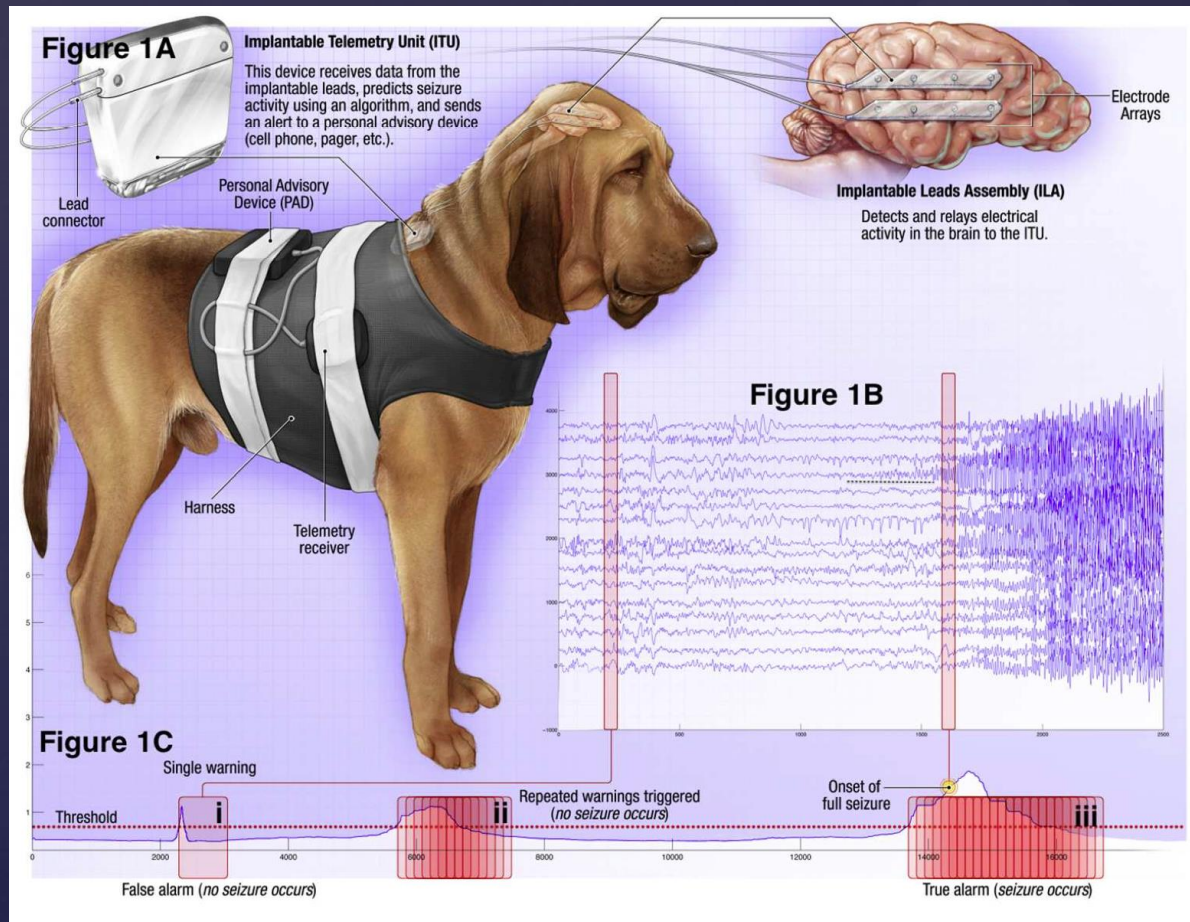


Video clip: Interview with Neha Kothari, LinkedIn Data Scientist

<http://youtu.be/8dxKe5cGHdA?t=17s>

Data Science Example

↳ Kaggle competition hosted by UPenn and Mayo Clinic to detect seizures in intracranial EEG recordings



- ⌘ Current detection systems have high false positive rate, resulting in unnecessary stimulation
- ⌘ Need to rapidly and automatically detect onset of seizure
- ⌘ Data provided
 - ⌘ Matrix of EEG sample values
 - ⌘ Time duration latency (time before seizure)
 - ⌘ Sampling frequency
 - ⌘ Channels (electrodes)
 - ⌘ Human and Canine Data
- ⌘ Latency only provided in “training” data because when taking real-life data, you won’t know if or how long until seizure hits – that’s what you’re trying to predict
 - ⌘ This is an important point in predictive analytics!

⌘ Competition winner Michael Hills published his method

⌘ FFT = Fast Fourier Transform

⌘ Determines primary frequencies in EEG sample

⌘ Correlation Coefficient “r”

⌘ Eigenvalues – can think of this as a scaling factor

⌘ Put all these values into a “Random Forest” classifier

⌘ Ensemble learning method – combines results of many “weak” decision trees, turns out to be better classifier than one “strong” decision tree

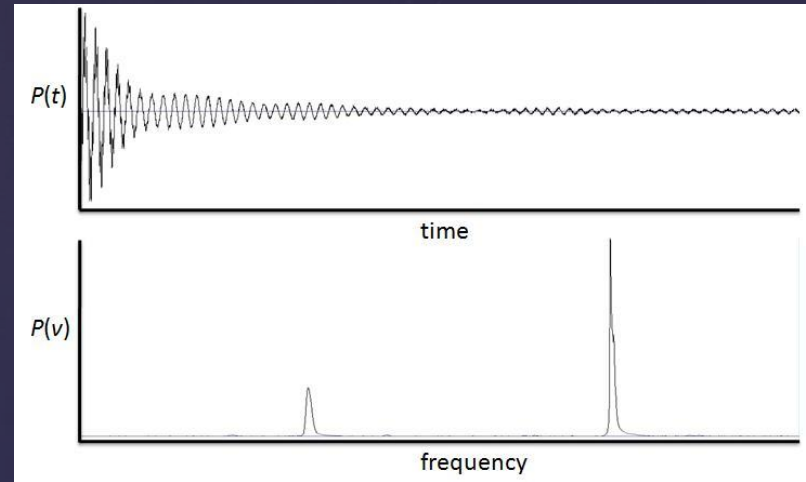
⌘ Can now train a classifier for each patient

⌘ He wrote a computer program to help him experiment & quickly validate result of each “brute force” approach, trying every technique he could find

⌘ Used the same evaluation technique kaggle competition would use

⌘ Line of scikit-learn Python code for training winning submission:

⌘ `RandomForestClassifier(n_estimators=3000, min_samples_split=1, bootstrap=False, random_state=0)`



http://en.wikipedia.org/wiki/Fast_Fourier_transform

⌘ Kaggle's evaluation method:

- ⌘ Judged on the mean area under the ROC curve (AUC) of two predictions. Receiver Operating Characteristic = true positive vs false positive.
 - 1) Predict the probability that a given clip is a seizure.
 - 2) Predict the probability that the clip is within the first 15 seconds its respective seizure (the technical term for time into the seizure is "latency").

The competition metric is the mean of these two AUCs:

$$\frac{1}{2} (AUC_{seizure} + AUC_{early})$$

⌘ Michael Hills' winning submission scored 0.963

- ⌘ His model will label 963 of every 1000 true seizure clips as seizures
- ⌘ He won \$5000 (much less than UPenn/Mayo would have had to pay a Data Scientist to develop this as an employee or consultant!)
- ⌘ Currently another similar contest posted w/\$25,000 prize

My Machine Learning project

Using JMU first-time donor (and non-donor) data from two previous years, could I classify who was likely to become a donor for the first time during the next year?

Correctly classified 67% of first-time donors, got great feedback from professor, plan to continue the study for my masters program final project.

You can read all about it on my blog! BecomingADataScientist.com

Code snippet using Random Forest Classifier

```
#build cross-validation data sets
from sklearn.cross_validation import train_test_split
sample_train, sample_test, target_train, target_test = train_test_split(sample, target, test_size=0.20)
#train the random forest classifier
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 50)
forest = forest.fit(sample_train, target_train)

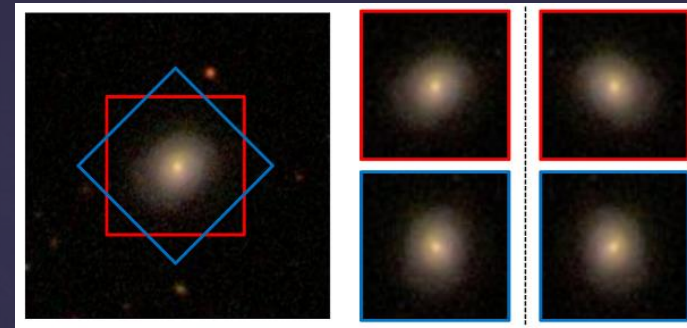
#train an "extra trees" random forest classifier
from sklearn.ensemble import ExtraTreesClassifier
forest2 = ExtraTreesClassifier(n_estimators = 50)
forest2 = forest2.fit(sample_train, target_train)

#test the model on various sets
trnresult = forest.score(sample_train,target_train)
tstresult = forest.score(sample_test,target_test)
class0result = forest.score(class0data,class0target)
class1result = forest.score(class1data,class1target)
```


Other Examples

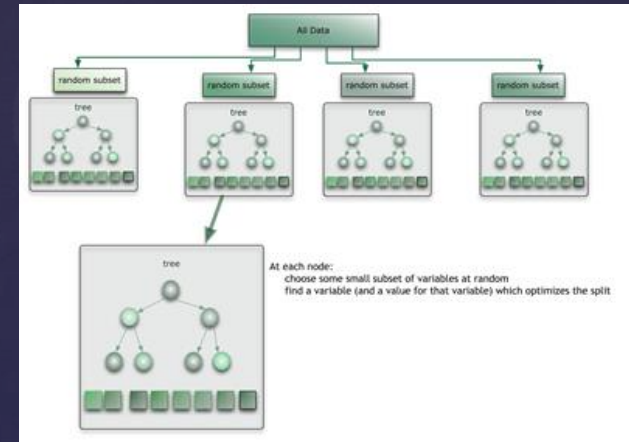
- Galaxy Classification using Convolutional Neural Networks

<http://benanne.github.io/2014/04/05/galaxy-zoo.html>



- Choosing Facebook Audience for Content Promotion using Random Forests

<http://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/>

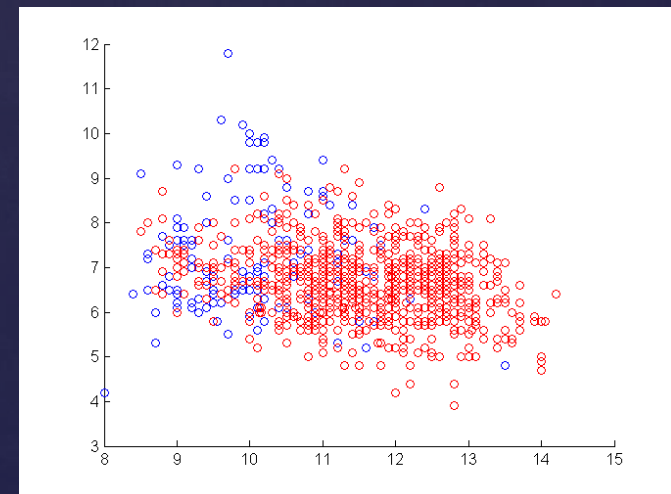


- Predicting Wine Quality with Principal Component Analysis

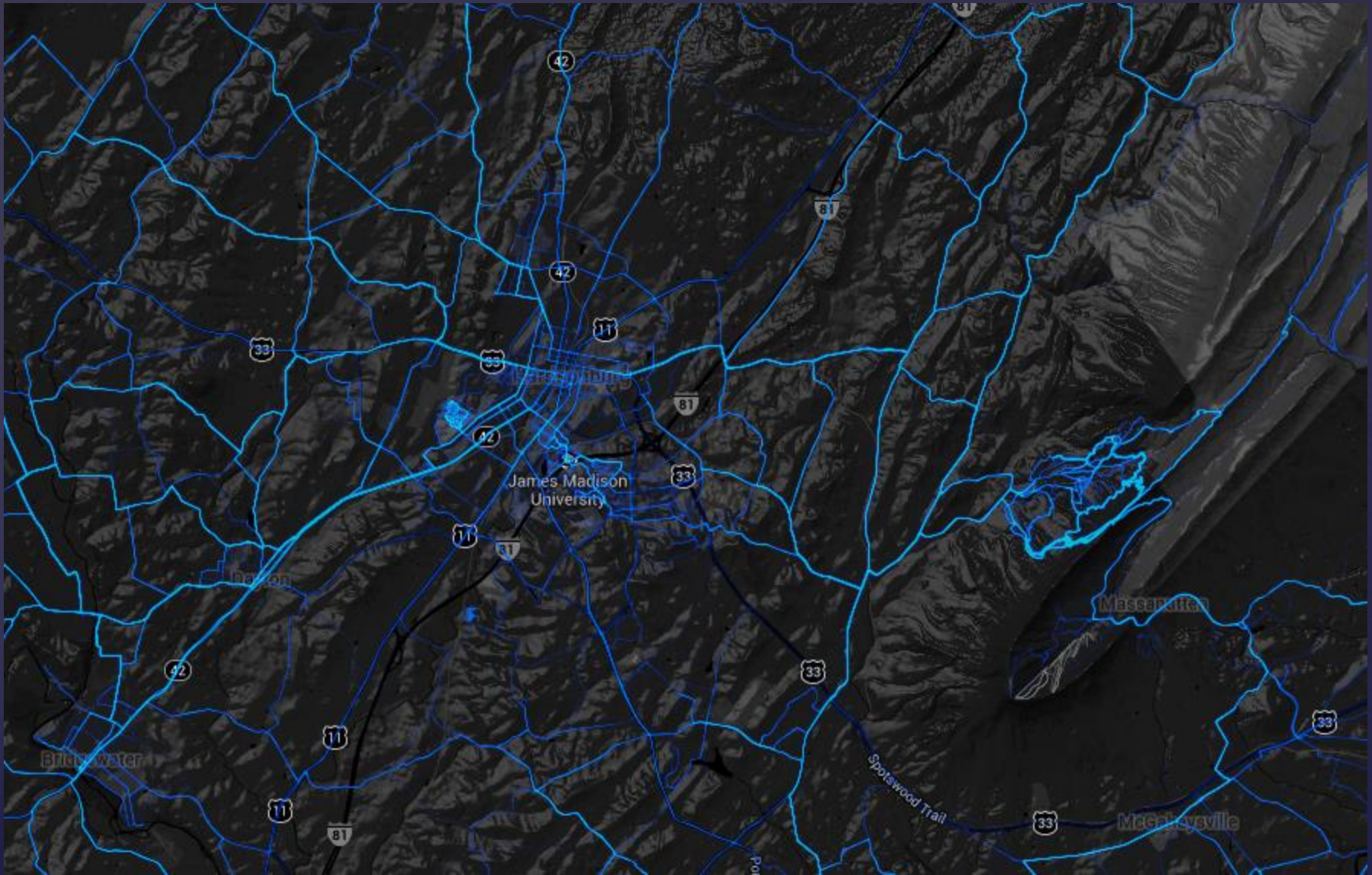
<http://fastml.com/predicting-wine-quality/>

- Readmission Risk Score to decide which patients to give additional follow-up help at Mt. Sinai hospital

<http://www.technologyreview.com/news/518916/a-hospital-takes-its-own-big-data-medicine/>



Data Visualization Example



<http://labs.strava.com/heatmap/#12/-78.90549/38.44669/blue/bike>

?



WHEN A USER TAKES A PHOTO,
THE APP SHOULD CHECK WHETHER
THEY'RE IN A NATIONAL PARK...

SURE, EASY GIS LOOKUP.
GIMME A FEW HOURS.

... AND CHECK WHETHER
THE PHOTO IS OF A BIRD.

I'LL NEED A RESEARCH
TEAM AND FIVE YEARS.



IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

How to get started

Recommended skills to pick up while at JMU

⌘ Programming

- ⌘ Any language is good to start with. Gain core understanding.
- ⌘ Python or R data analysis experience a plus
- ⌘ Database design, SQL

⌘ Math

- ⌘ Calculus
- ⌘ Linear Algebra
- ⌘ Statistics (2 levels)
- ⌘ Advanced: Optimization / Linear Programming

⌘ Research and Analysis

- ⌘ Science involving data collection and interpretation
- ⌘ Working with “messy” real life data
- ⌘ Business Analytics
- ⌘ Data Mining

⌘ Others

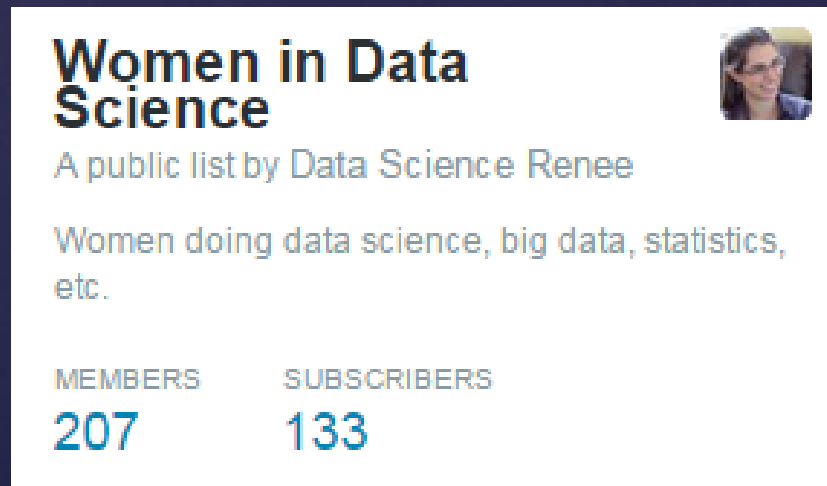
- ⌘ Business / Communication
- ⌘ Graphic Design

Take classes on campus or online!

Read, read, read

- ⌘ *Doing Data Science* by Cathy O'Neil* & Rachel Schutt
- ⌘ *Data Science for Business* by Forster Provost & Tom Fawcett
- ⌘ *Data Smart* by John Foreman* (uses Excel)
- ⌘ I'll review other books as I read them:
<http://www.becomingadatascientist.com/learning/>
- ⌘ Blogs & News Feeds (FlowingData.com is a good one to start with)
- ⌘ Twitter – look for curated lists of people to follow
<https://twitter.com/BecomingDataSci/lists/women-in-data-science/members>

*on Twitter and willing to chat!



Women in Data Science

A public list by Data Science Renee

Women doing data science, big data, statistics, etc.

MEMBERS	SUBSCRIBERS
207	133

The image shows a screenshot of a Twitter list. At the top, the title 'Women in Data Science' is displayed in bold black text. Below the title, it says 'A public list by Data Science Renee' in a smaller, grey font. There is a small profile picture of a woman with glasses. Below that, a description reads 'Women doing data science, big data, statistics, etc.' At the bottom, there are two columns: 'MEMBERS' with the number '207' and 'SUBSCRIBERS' with the number '133'. The numbers are in a larger, bold blue font.

Free Online Courses

⌘ *Python Fundamentals* – Codecademy <http://www.codecademy.com/tracks/python>

⌘ *Machine Learning* – Coursera / Stanford <https://www.coursera.org/course/ml>

⌘ *Data Analyst Nanodegree* – Udacity <https://www.udacity.com/course/nd002>
(includes Hadoop mini-course)

⌘ *Applied Data Mining and Statistical Learning* – Penn State
<https://onlinecourses.science.psu.edu/stat857/>

⌘ Pretty comprehensive list here: <http://www.kdnuggets.com/education/online.html>

⌘ TED talks on Data <http://www.ted.com/search?q=data>

⌘ **Susan Etlinger*** http://www.ted.com/talks/susan_etlinger_what_do_we_do_with_all_this_big_data

⌘ “Need to spend more time on critical thinking skills...[because we have the] potential to make bad decisions far more quickly, efficiently, and with far greater impact than we did in the past.”

⌘ “...we need to be clear about ..the methodologies that we use, ...because if I don't know what ...questions you asked, I don't know what questions you didn't ask.”

Explore

⌘ Volunteer to Analyze Data (DataKind)

⌘ Play with public data sets

⌘ <http://101.datascience.community/2014/10/17/data-sources-for-cool-data-science-projects-part-1-guest-post/>

⌘ <https://www.opensciencedatacloud.org/publicdata/>

⌘ <http://catalog.data.gov/dataset>

⌘ <https://archive.ics.uci.edu/ml/datasets.html?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table>

⌘ Data Science Competitions

(Kaggle also has “knowledge competitions” for learning)

What some of my followers on Twitter wish they knew about data in college....



Jacquie Tran @jacquietran · Nov 5

.@BecomingDataSci wish I learned the basics about data types and structure, esp. what it means for the kinds of questions you can ask of it



Nicole Radziwill @nicoleradziwill · Nov 5

@BecomingDataSci I wish I knew that everyone struggles with the challenge of getting intimate w their data. There are no right answers.



Sumit Bajaj @sumit_bajaj · Nov 6

@BecomingDataSci messy data like its in the real world...



Fareeza Khurshed @stat_geek · Nov 6

@BecomingDataSci That data is useless without someone asking good questions. How data collected/original usage very important.



Just Glowing @JustGlowing · Nov 6

@BecomingDataSci I wish I had a stronger focus on statistics and probability. Especially on statistical testing, now my favorite tool.



CBat @ImADataGuy · Nov 9

@BecomingDataSci being self taught I have a couple of thoughts: 1) don't be afraid to ask questions. Esp with Internet, info is everywhere 1/

Questions?

Renee T.

[contact me via twitter or blog for email]

@becomingdatasci

<http://www.becomingadatascientist.com>