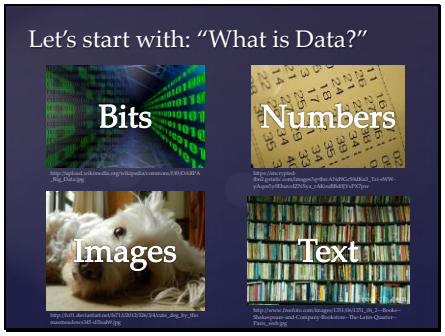
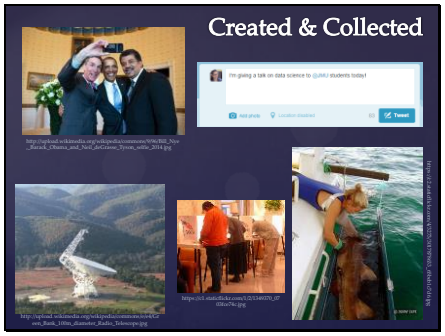
Slide 1

Slide 2



Bits
Numbers
Text
Images
(etc.)

Slide 3



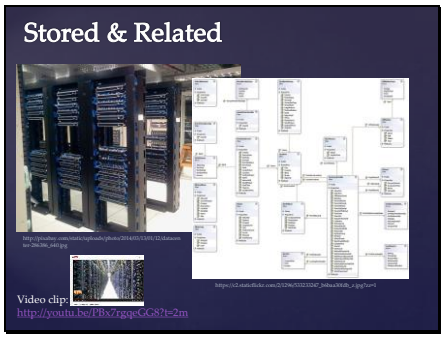
Created
Collected
Type of Data we're talking about is digital, stored in computers

Radio telescope
Voting
Shark Tagging

Slide 4

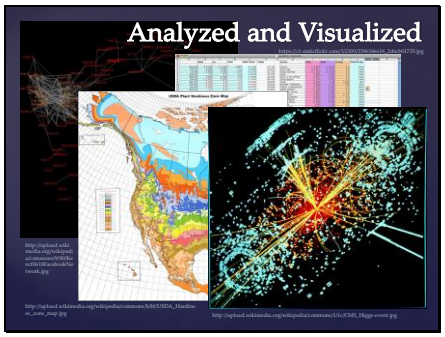
What are some other examples of big data databases?
-Credit Card swipes
-Text messages

Slide 5



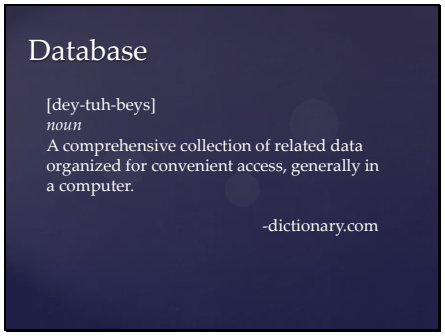
All of that has to be stored somewhere, and organized for access and analysis
(video clip)

Slide 6



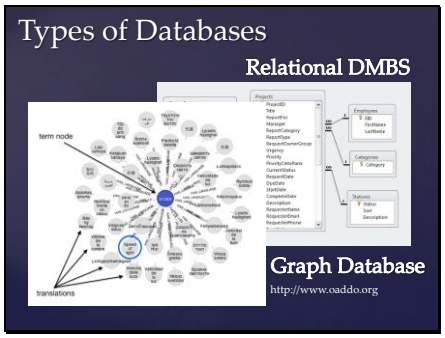
Slide 7

Slide 8



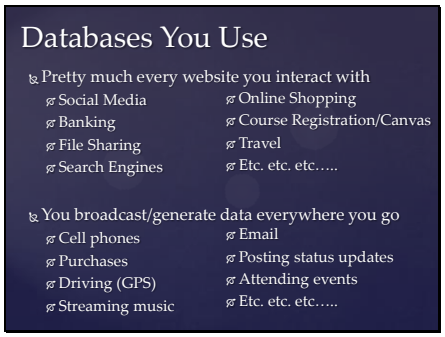
I used a database to look up this definition!

Slide 9



Relational
Document
Object-Oriented
Graph
Unstructured – text, audio, images

Slide 10



Social Media – posts, friends/follows, likes/favorites, location-tagged images
Note: often other people generating this data about you (tags, mentions, etc.)

Online Shopping – “other customers who purchased this also purchased….”, even just browsing the website, clicking, spending time on a page – usually all of that data is tracked.
Ever noticed when you leave an online store, the items you looked at “follow” you around the internet via ads?

Travel – purchase tickets, check in, post on social media, rental car with GPS, hotel rooms, credit card at restaurant, generating data everywhere you go
-credit card fraud alerts when in new location

Cell phones constantly generating data – app usage, location, websites, alarms, games, photos, etc.

Slide 11

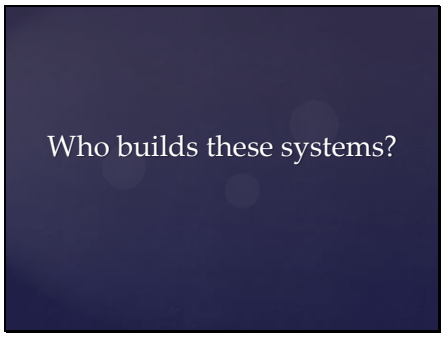
Now that I've gotten you thinking about data, specifically YOUR data, let's think about some ways in which having your data collected (and aggregated) can help you:

-Navigation (Google Maps directions)
-Recommendations (Yelp, Netflix)
-Medical Diagnoses
-Alerts

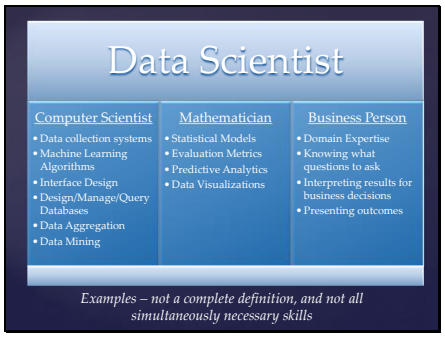
How are these generated? ALGORITHMS

Downside
-Some sites now charging different customers different prices based on browsing history http://www.fastcoexist.com/3037888/where-and-how-youre-online-shopping-changes-the-prices-you-see?utm_source=facebook
-Any data could be hacked (such as health or financial records) and lead to loss of privacy. The more places it's stored, the more vulnerable it is.

Slide 12



Slide 13



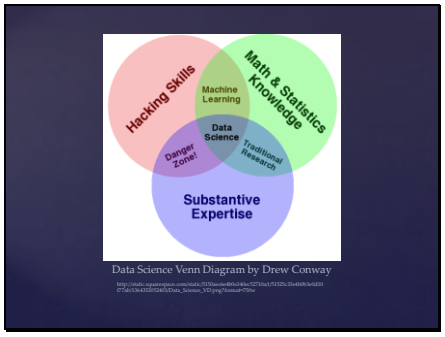
Who writes these algorithms?
-Experts in Machine Learning – Computer Scientists – Data Scientists!

They're often using statistical models. Who develops those?
-Mathematicians – Statisticians – Data Scientists!

Why do they write them?
-Sometimes altruistic or experimental, but usually to make someone money!
Who is using these results to make money?
-Business People – Marketers – Data Scientists!

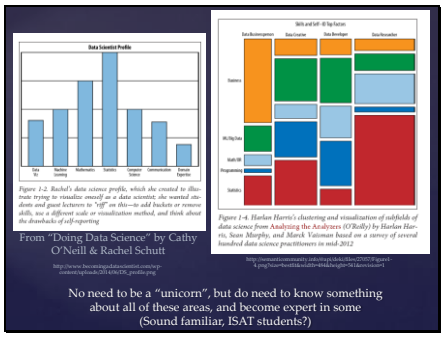
Note: you don't have to be the expert in all of these areas

Slide 14



But let's not get ahead of ourselves… back to the "data being stored and related" part

Slide 15



Data Visualization
Machine Learning
Mathematics
Statistics
Computer Science
Communication
Domain Expertise

Slide 16

Many data science jobs in financial industry (credit cards, investing) and marketing (ad serving) realm, however, that seems to be changing now that every company seems to be looking into whether they should have a data scientist on staff. Pick some areas you're interested in, and search the internet for people in that area in data jobs.

Also, there are now organizations like DataKind for data scientists and analysts to volunteer their time and skills to help solve problems in arenas outside their "day job" field, such as non-profits and cities.

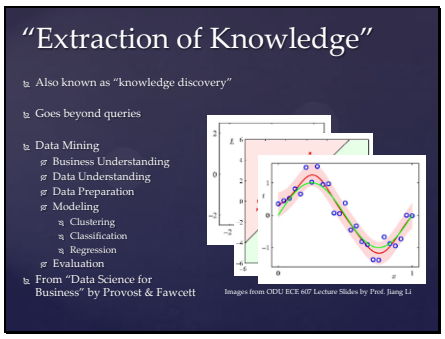
Slide 17

Recently saw 2 jobs posted in Charlottesville: "Junior Data Scientist" w/2 years experience was over $70K, senior $120K – and that's in small city!

http://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

Why data science jobs are in high demand
http://www.extension.harvard.edu/hub/blog/extension-blog/why-data-science-jobs-are-high-demand

Slide 18

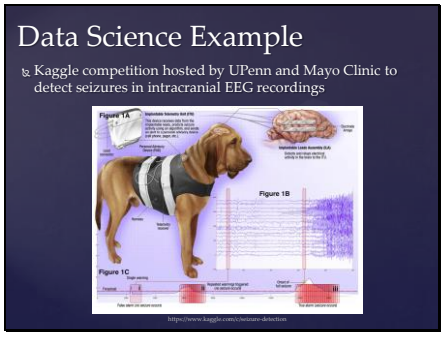


Clistering, Classification, Regression

Slide 19

Data scientist video clip

Slide 20



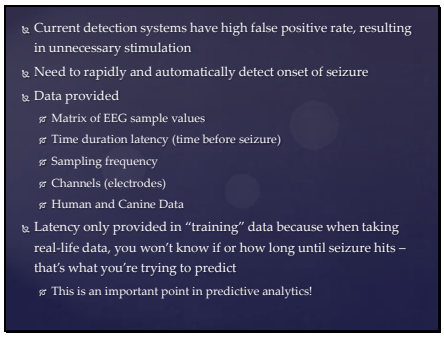
Detailed walkthrough of a data science problem

Check this next competition, ends 11/17:
https://www.kaggle.com/c/seizure-prediction

"**For individuals with drug-resistant epilepsy, responsive neurostimulation systems hold promise for augmenting current therapies and transforming epilepsy care.**
Of the more than two million Americans who suffer from recurrent, spontaneous epileptic seizures, 500,000 continue to experience seizures despite multiple attempts to control the seizures with medication. For these patients responsive neurostimulation represents a possible therapy capable of aborting seizures before they affect a patient's normal activities.

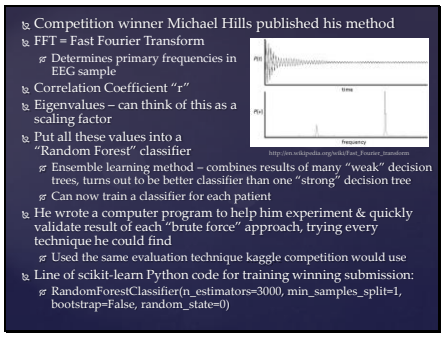
In order for a responsive neurostimulation device to successfully stop seizures, a seizure must be detected and electrical stimulation applied as early as possible. A seizure that builds and generalizes beyond its area of origin will be very difficult to abort via neurostimulation. Current seizure detection algorithms in commercial responsive neurostimulation devices are tuned to be hypersensitive, and their high false positive rate results in unnecessary stimulation.
In addition, physicians and researchers working in epilepsy must often review large quantities of continuous EEG data to identify seizures, which in some patients may be quite subtle. Automated algorithms to detect seizures in large EEG datasets with low false positive and false negative rates would greatly assist clinical care and basic research.
“

Slide 21



“Future” data can’t be used to predict outcomes, but it can be used to determine what already-known data tends to correlate with it during the “training” of your model.

Slide 22



https://www.kaggle.com/c/seizure-detection/forums/t/10111/required-model-documentation-and-code/52439
Correlation between EEG channels
Michael Hills’ code is posted on GitHub
His summary of winning approach:

> “Quickly summarising my model, for feature selection I used FFT 1-47Hz, concatenated with correlation coefficients (and their eigenvalues) of both the FFT output data, as well as the input time data. The data was then trained on per-patient Random Forest classifiers (3000 trees).”

http://www.cip-labs.net/2013/01/17/introduction-to-random-forests/

Slide 23

- 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:

$$1/2\left(AUC_{seizure} + AUC_{early}\right)$$

- 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

Slide 24

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.

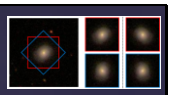
You can read all about it on my blog! BecomingADataScientist.com

Code snippet using Random Forest Classifier

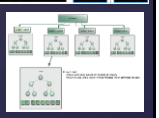
```
#build cross-validation data sets
from sklearn.cross_validation import train_test_split
sample_train, sample_test, target_train, target_test = train_test_split(sample, target, test_size=0.20)
#train the random forest classifier
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 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
trresult = forest.score(sample_train,target_train)
tstresult = forest.score(sample_test,target_test)
classtresult = forest.score(classtdata,classttarget)
classresult = forest.score(classidata,classitarget)
```

Slide 25

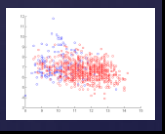
Other Examples



- 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/

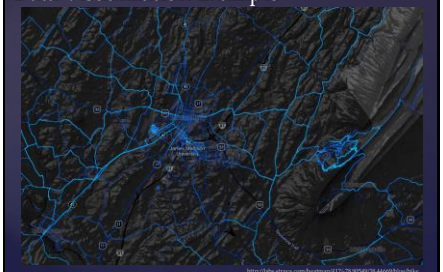


- Readmission Risk Score to decide which patients to give additional follow-up help at Mt. Sinai hospital
  http://www.technologyreview.com/news/518916/a-hospital-takes-its-own-big-data-medicine/

Note – in the last one they did a pilot study, and the extra care cut readmission rates in half

Slide 26

Data Visualization Example



http://labs.strava.com/heatmap/#12/-78.90549/38.44669/bluer/bike

This is called a "HeatMap" – other kinds of heatmaps, this one changes street color based on traffic volume What can we learn from this visualization of walking vs biking in Harrisonburg? What about in Massanutten? (Were all those people riding bikes up there? Using the app while skiing? -- Researched and found Shenandoah Valley Bicycle Coalition mountain biking trails http://appliedtrailsresearch.com/wp-content/uploads/2012/03/NutMap11_LoRes-1.pdf)

Questions to ask: How was data collected? How many different people are represented? How is "scale" of color levels decided? Were "too fast" data points taken out? (people using app in car?) Do people respond differently to the blue version of the heatmap vs yellow version?

Any privacy issues? (one version of app shows you "who you passed on the trail")

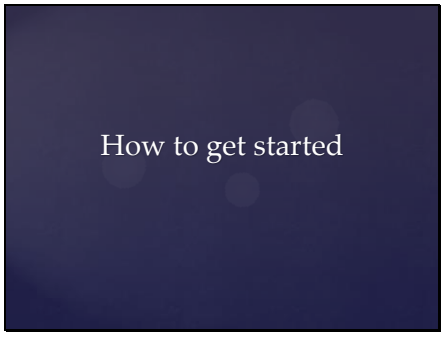
Lots to think about from this relatively simple example!

Slide 27

This area looks a little “darker”. What can we conclude? Are the people that live here less active? Is there a smaller population?
-reveal to show it is the Manassas-Centreville area in Northern VA, which has many more people than Harrisonburg
(probably just fewer people using app! Or maybe they’re working out inside. Maybe a bike club in Harrisonburg competes on the app or something to drive the numbers up. Or maybe there aren’t many people using it, so the “heavy” areas are just a couple users.)

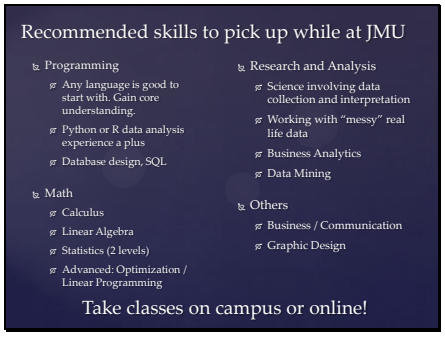
Slide 28

Slide 29



So, hopefully I got some of you interested in Data, Databases and Data Science. If you want to learn more, or even consider doing this as a career, what can you do while you’re in college to get started?

Slide 30



Did I have all of these when I graduated? No – had basic Stats & Calculus, basic VB programming, Database Design, ISAT projects
But this is what I would have taken more of had I known about data science then.

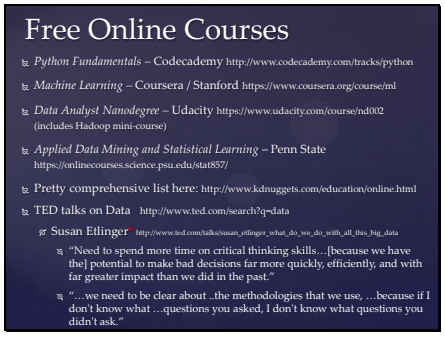
If you’re already well-versed in area, either get more advanced, or get more breadth (recommended). If you’re a math major, take a science research course. If you’re a CS major, take a business course. Etc.

Didn’t have these great online courses when I was in school.

Slide 31

Here's a blog post by Trey Causey with good info on getting started: http://treycausey.com/getting_started.html
Interview with Jawbone Data Scientist Abe Gong about using Data Science to solve human problems: http://www.datascienceweekly.org/data-scientist-interviews/using-data-science-solve-human-problems-abe-gong-interview

Slide 32



Also many universities are offering graduate-level Data Science programs now! (UVA on campus, Berkeley online, for instance) – not free, though! There is an "open source masters": http://datasciencemasters.org/ (@clarecorthell also on twitter)

Some machine learning Python libraries:
http://scikit-learn.org/stable/
http://pybrain.org/pages/features
More:
http://dataaspirant.wordpress.com/2014/11/01/python-packages-for-datamining/?utm_content=buffere2274&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

I keep a list of courses I'm taking and have completed here:
http://www.becomingadatascientist.com/learning/

Slide 33

Slide 34

Link to question on twitter for all replies: https://twitter.com/BecomingDataSci/status/530214823347228672

Slide 35

BecomingADataScientist.com – contact me there!

Leave a comment! ☺