What is Data Science?

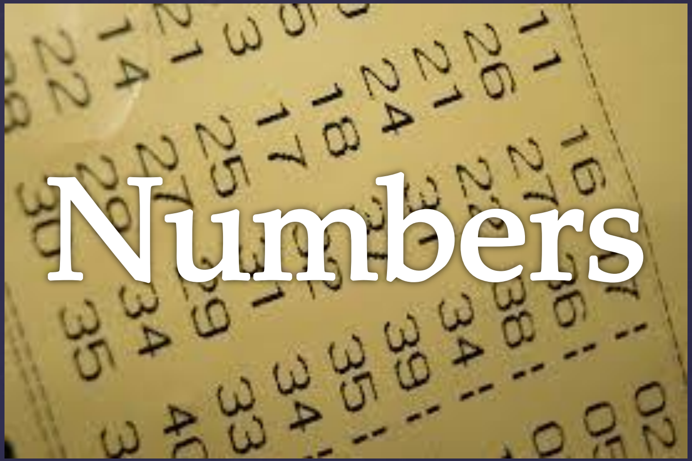
{ Girl Develop It! Meetup

Renée M. P. Teate, March 2015

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

Bits

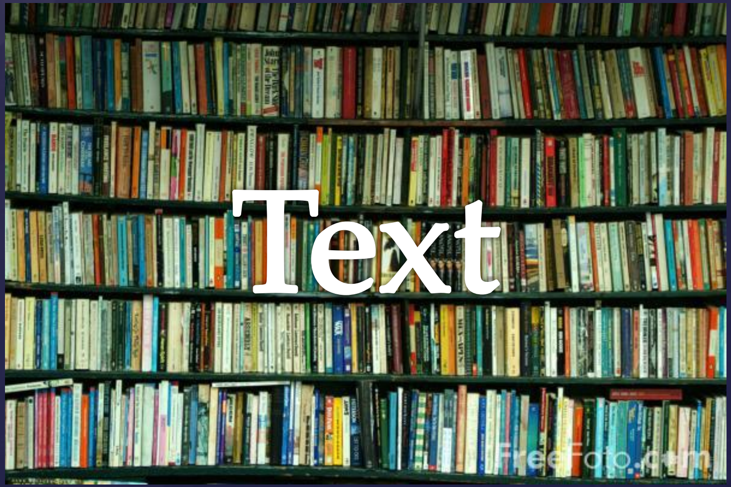
http://upload.wikimedia.org/wikipedia/commons/f/f0/DARPA_Big_Data.jpg


Numbers

https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcS9dKu3_Tzi-sWW-yAqee5y0EhuvoIZNSya_rAKnuBBd0JYxPX7pw

Images

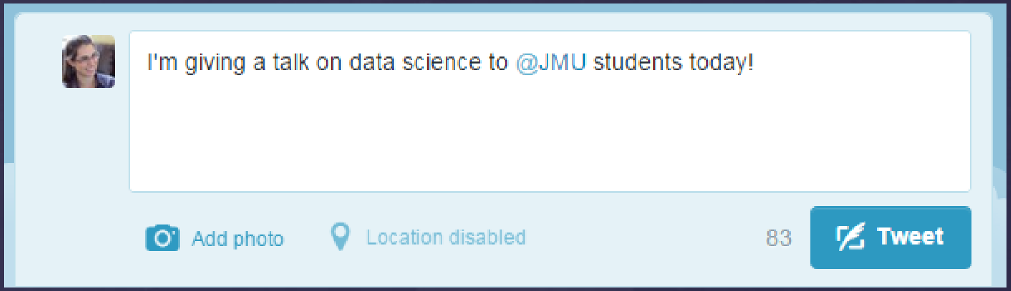
http://fc01.deviantart.net/fs71/i/2012/326/3/4/cute_dog_by_thomasmeadows345-d5lsah9.jpg


Text

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



I'm giving a talk on data science to @JMU students today!

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

- 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

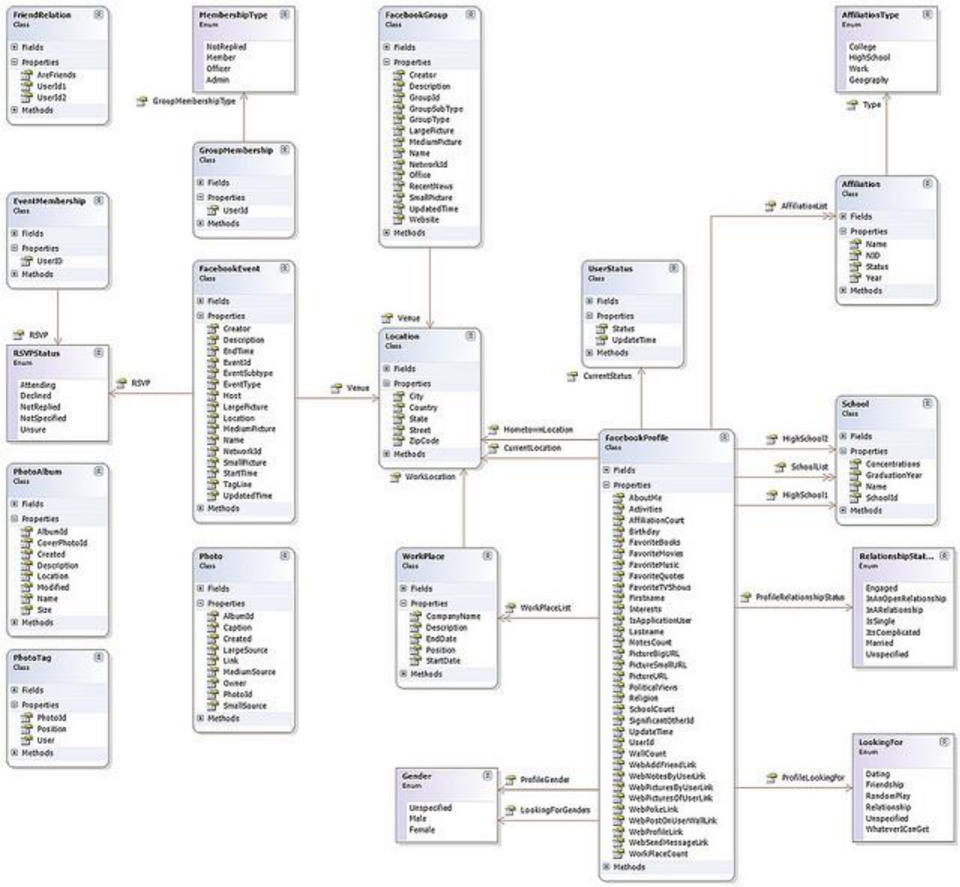
- 570 new websites spring into existence every minute of every day.

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

"Big Data"

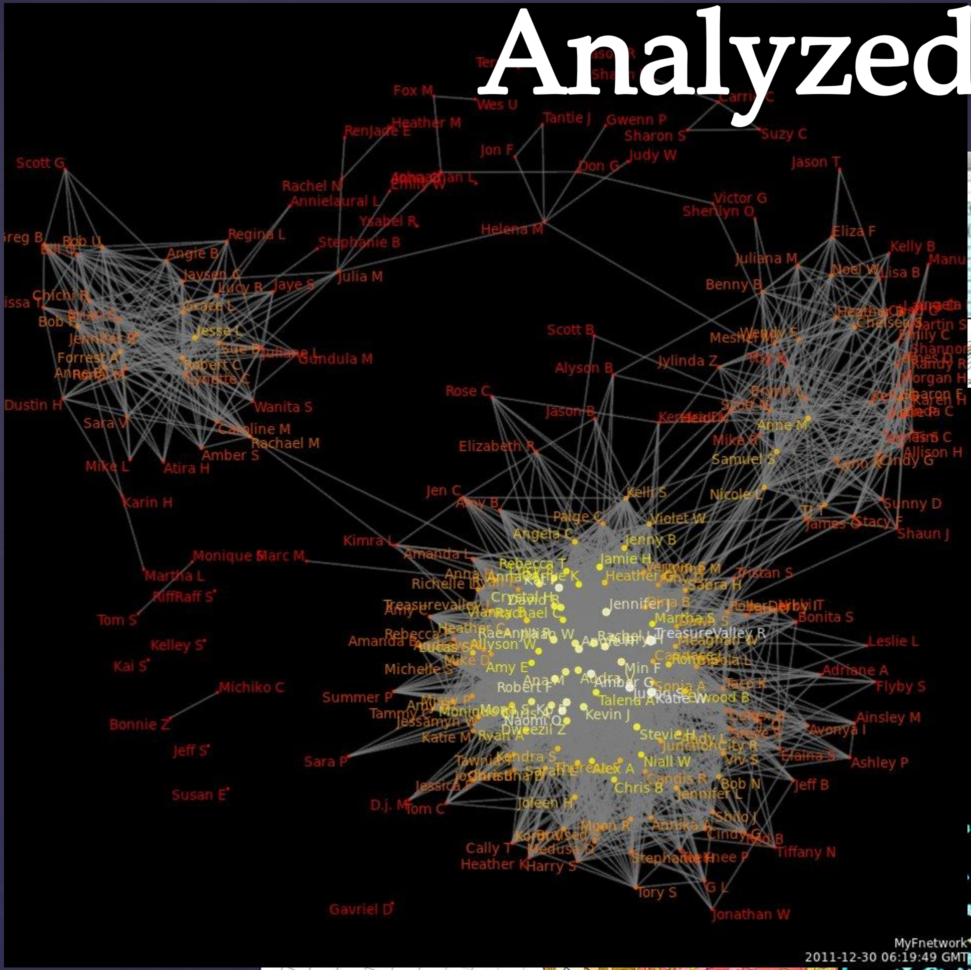
Stored & Related

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



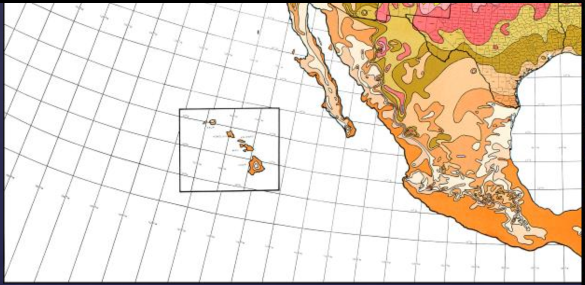
https://c2.staticflickr.com/2/1296/533233247_b6baa30fdb_z.jpg?zz=1

Analyzed and Visualized

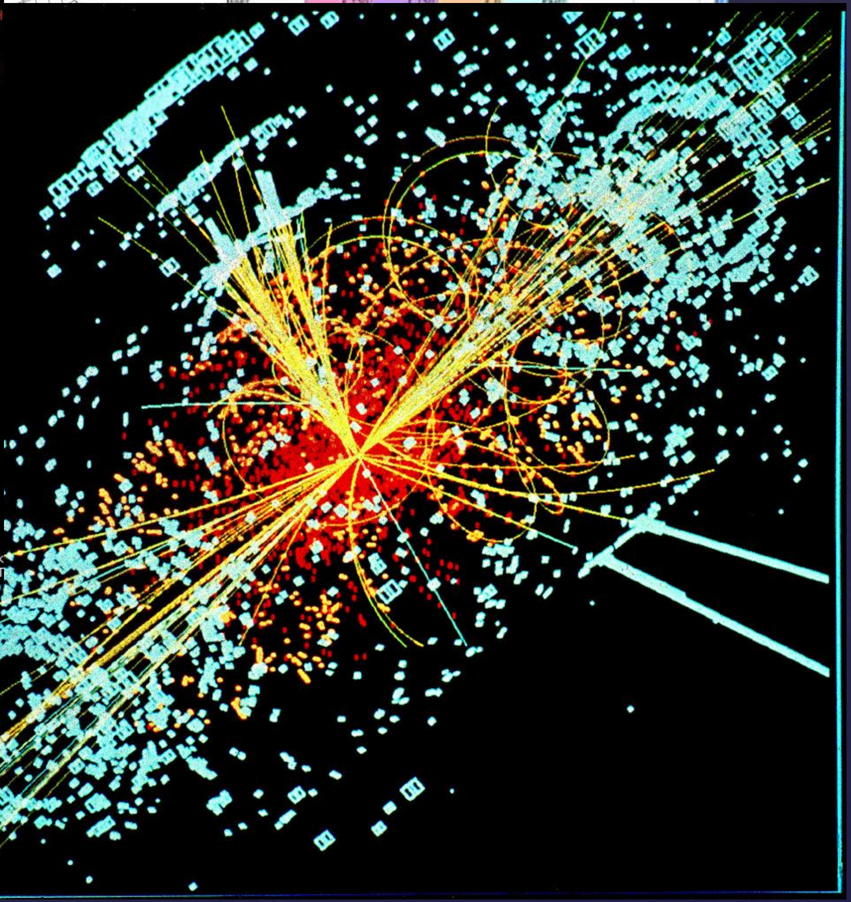


http://upload.wikimedia.org/wikipedia/commons/9/90/Kencf0618FacebookNetwork.jpg

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



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



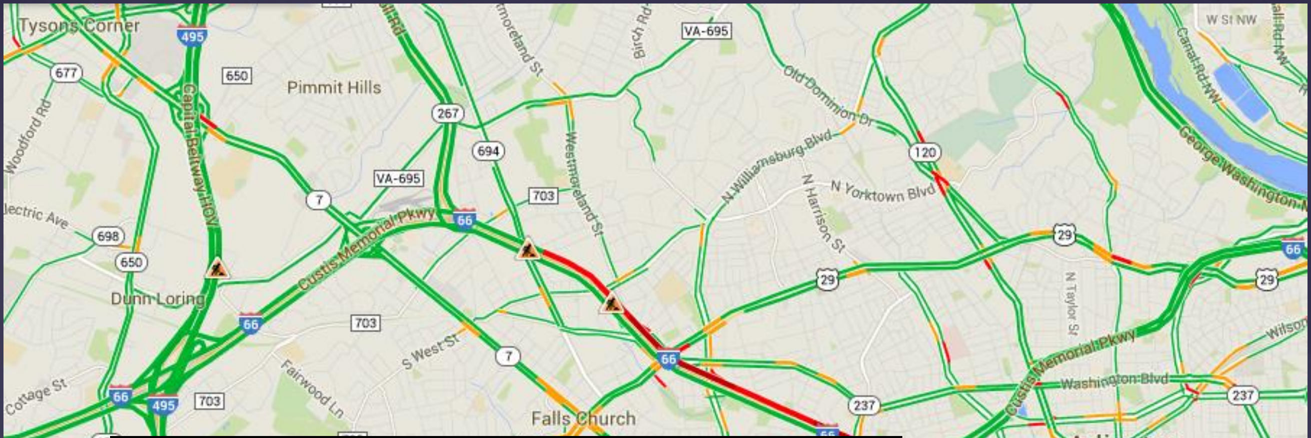
http://upload.wikimedia.org/wikipedia/commons/1/1c/CMS_Higgs-event.jpg

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…..

https://www.google.com/maps/@38.8905569,-77.1721577,13z/data=!5m1!1e1



https://c2.staticflickr.com/4/3324/3507973704_563846fe14_z.jpg?zz=1

NETFLIX

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

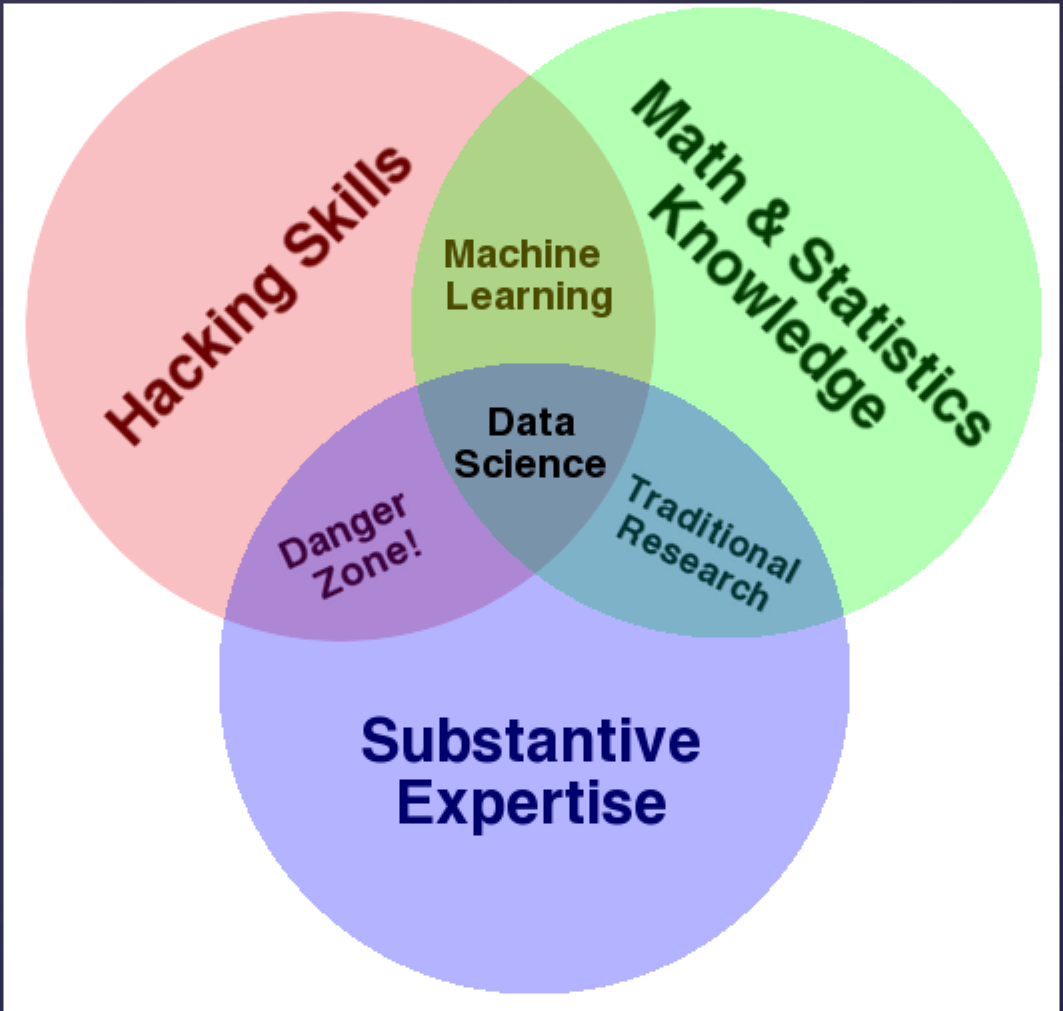
Mathematician

- 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

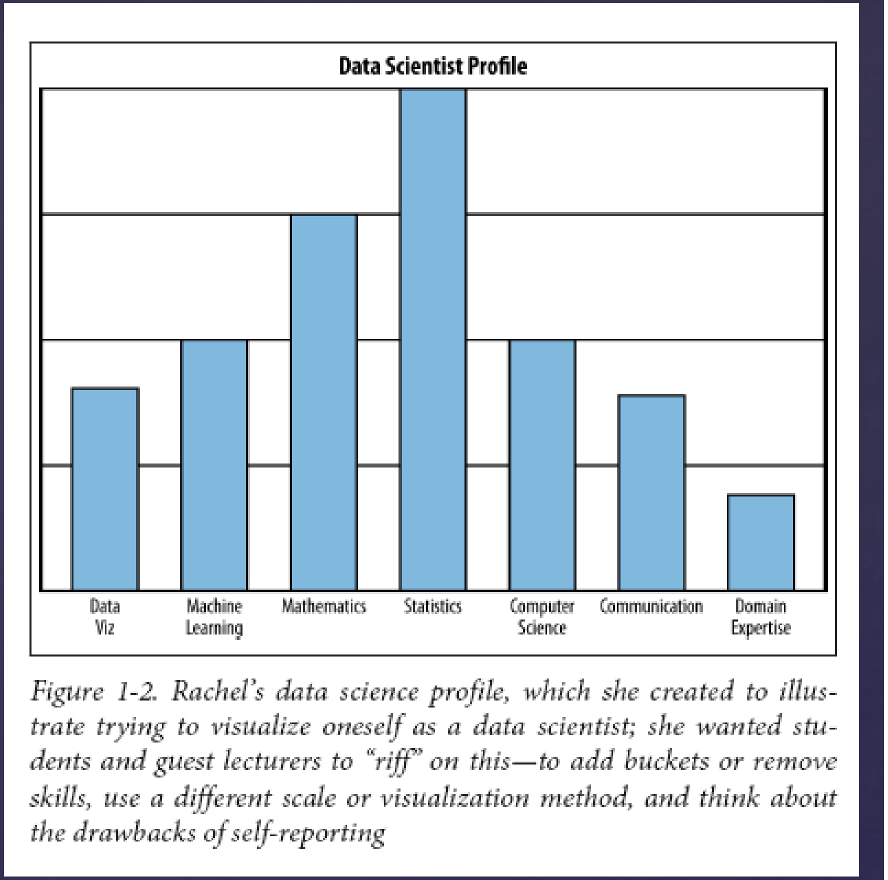


Figure 1-2. Rachel's data science profile, which she created to illustrate trying to visualize oneself as a data scientist; she wanted students and guest lecturers to "riff" on this—to add buckets or remove skills, use a different scale or visualization method, and think about the drawbacks of self-reporting

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

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

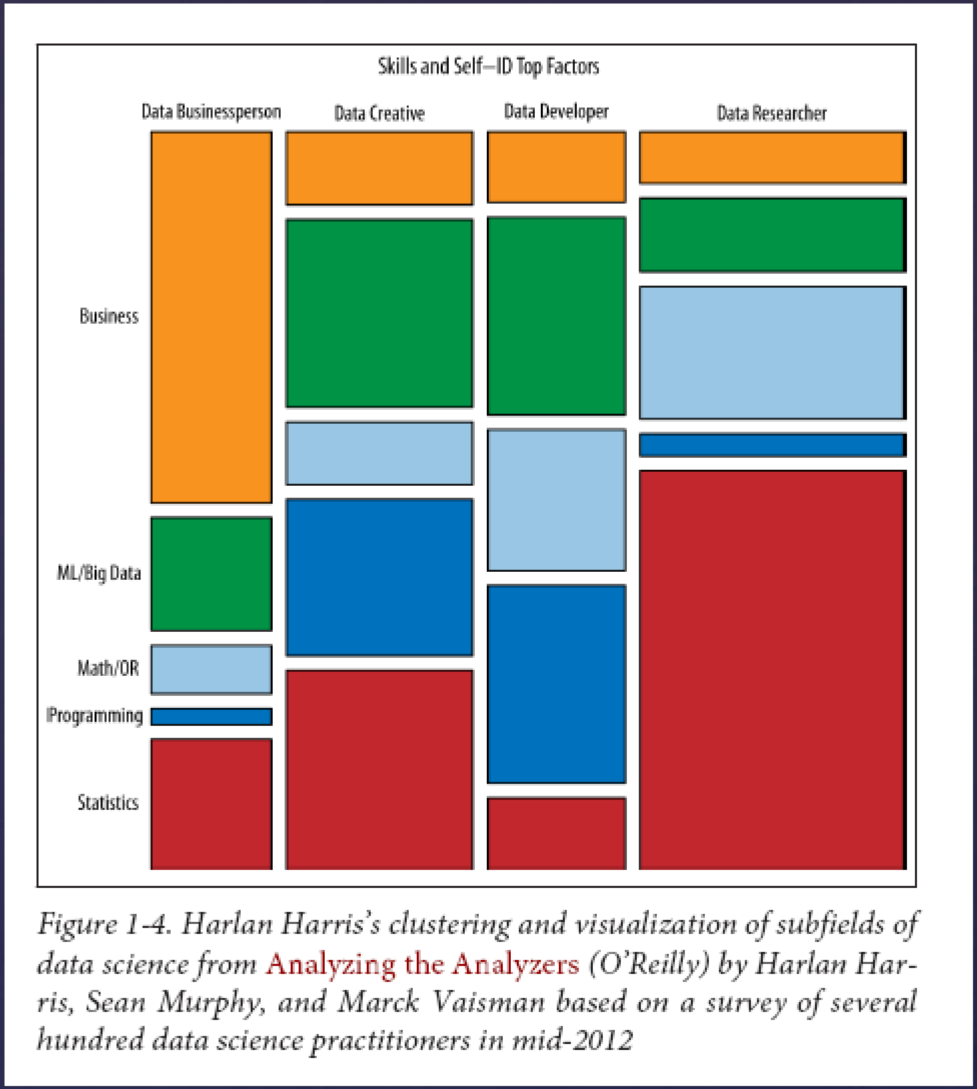


Figure 1-4. Harlan Harris's clustering and visualization of subfields of data science from Analyzing the Analyzers *(O'Reilly) by Harlan Harris, Sean Murphy, and Marck Vaisman based on a survey of several hundred data science practitioners in mid-2012*

http://semanticommunity.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

Some other names for "Data Scientist"

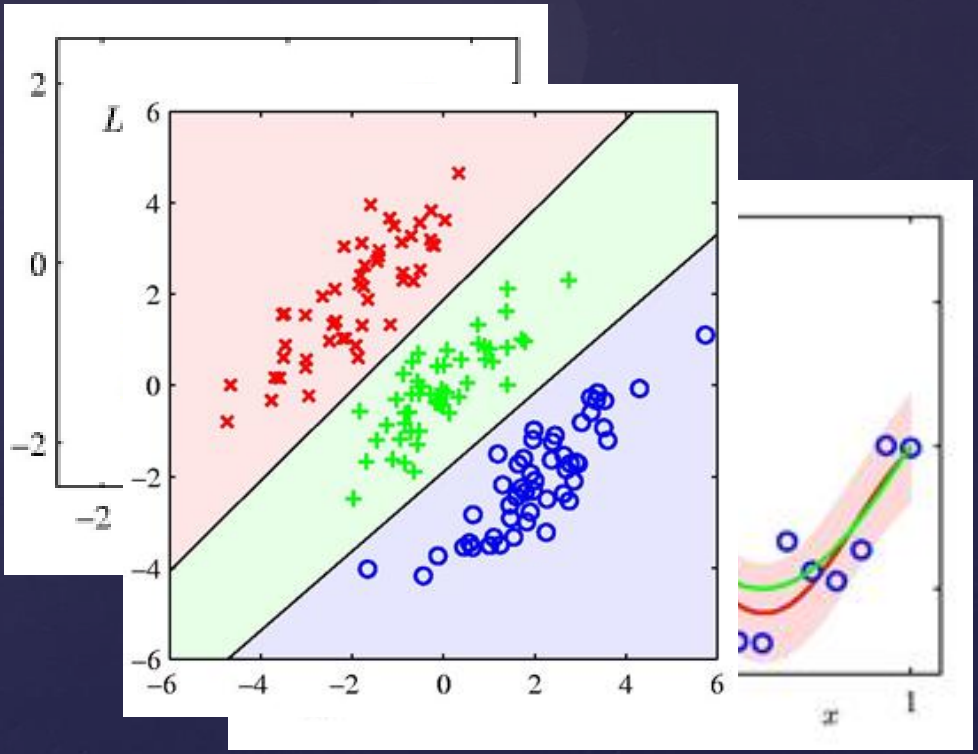
- Statistician
- Data Mining Specialist
- Biostatistician
- Social Science Researcher
- Big Data Analyst
- Spatial/GIS Analyst
- Natural Language Programmer
- Computational Physicist
- Pythonista
- Financial Analyst
- Recommendation System Engineer
- Information Architect
- Artificial Intelligence Researcher
- 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



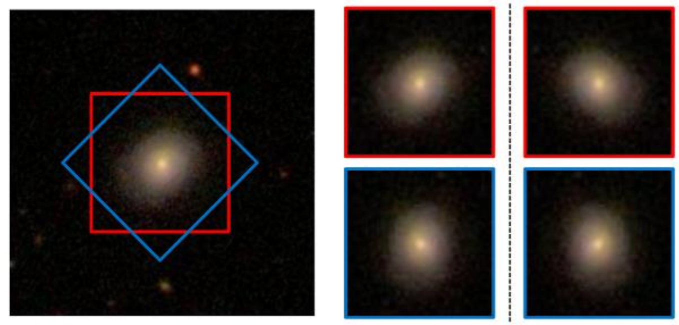
Images from ODU ECE 607 Lecture Slides by Prof. Jiang Li

Video clip: Interview with Neha Kothari, LinkedIN Data Scientist
http://youtu.be/8dxKe5cGHdA?t=17s

Examples

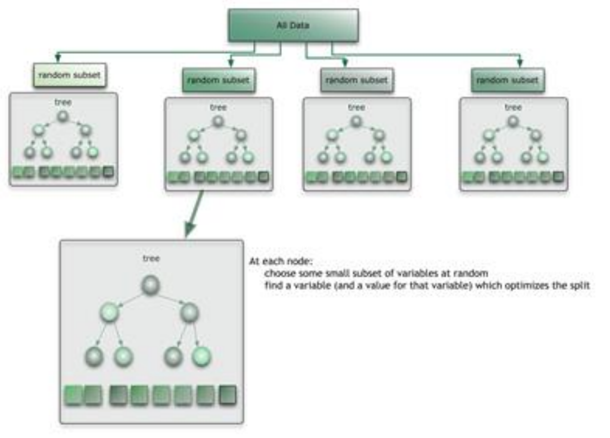
- Galaxy Classification using Convolutional Neural Networks

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



- Choosing Facebook Audience for Content Promotion using Random Forests

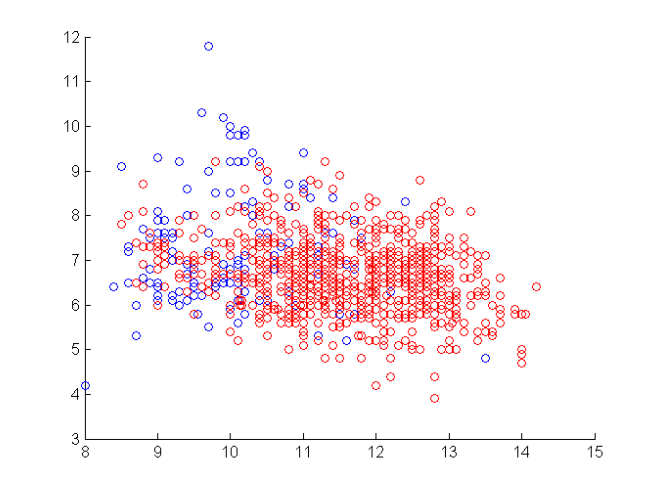
  http://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/



At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

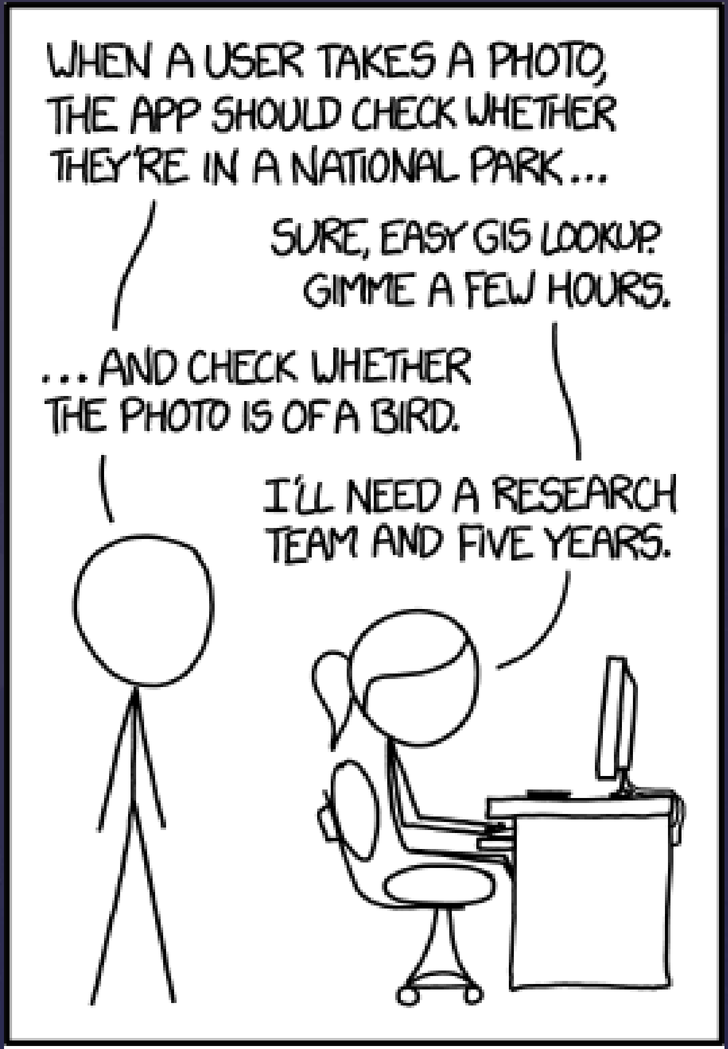
- Predicting Wine Quality with Principal Component Analysis

  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/


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.

http://xkcd.com/1425/

How to get started

Topics to learn about

- 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
  - 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

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 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

Women in Data Science

A public list by Data Science Renee

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

MEMBERS 207

SUBSCRIBERS 133

*on Twitter and willing to chat!

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)

Questions?

Renee Teate

renee.parilak@gmail.com, @becomingdatasci

http://www.becomingadatascientist.com