GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 1: DATA SCIENCE OVERVIEW

WELCOME!

MEET YOUR INSTRUCTIONAL TEAM

- Yuchen Zhao

MEET YOUR INSTRUCTIONAL TEAM

- Alex Chao

Startup.ML

GENERAL ASSEMBLY

RWJ ROBERT WOOD JOHNSON UNIVERSITY HOSPITAL

unity

University of California Berkeley Haas School of Business

Koding

MEET YOUR INSTRUCTIONAL TEAM

- David Feldman

WHERE FOUNDERS MEET.
FOUNDERDATING

GA GENERAL ASSEMBLY

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DataKind
USING DATA IN THE SERVICE OF HUMANITY

gen 110

AGENDA

0. INTRODUCTION
1. WHAT IS DATA SCIENCE?
2. THE DATA MINING WORKFLOW

LAB:

3. GITHUB & PYTHON
4. Q&A

LEARNING OBJECTIVE

- Describe the data mining workflow and the key traits of a successful data scientist.
- Set up github account.
- Explore & visualize data using python.

LOGISTICS

Instructor:

Yuchen Zhao (ZYCEMAIL+GA@GMAIL.COM)

Experts-in-residence:

Alex Chao (ALEXCHA056@GMAIL.COM)
David Feldman(davidfeldman3@gmail.com)

Course Producer:

Vanessa Ohta

Course Times: 6:30pm-9:30pm, Tuesdays and Thursday

Couse materials: HTTPS://GITHUB.COM/GA-STUDENTS/DAT_SF_14

Introductions

- Your name
- A brief summary of your background (e.g. work, school, etc.)
- What you hope to get out of the class
- One interesting / surprising / fun fact about yourself

I. WHAT IS DATA SCIENCE?

Startups Using Big Data

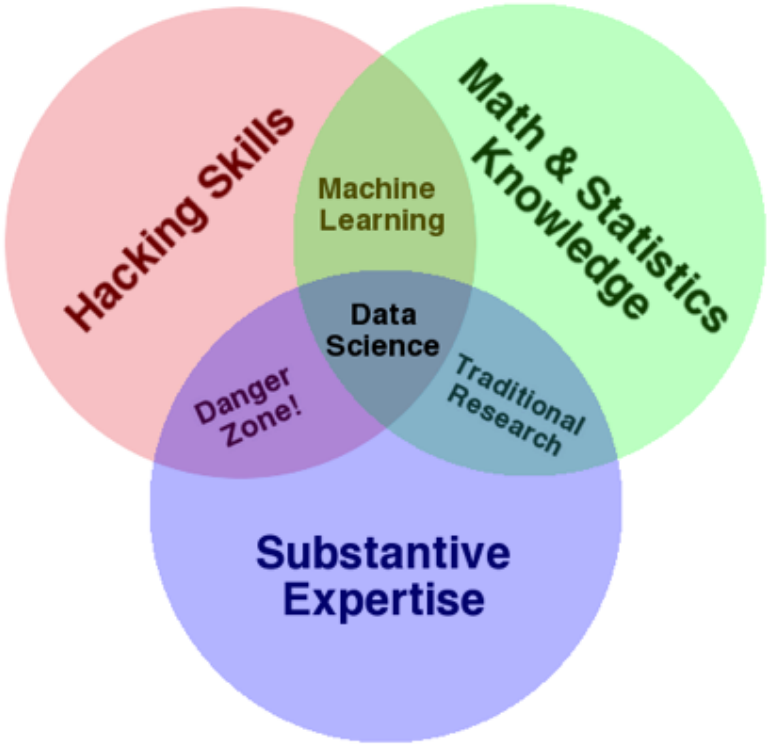
WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.

WHAT IS DATA SCIENCE?

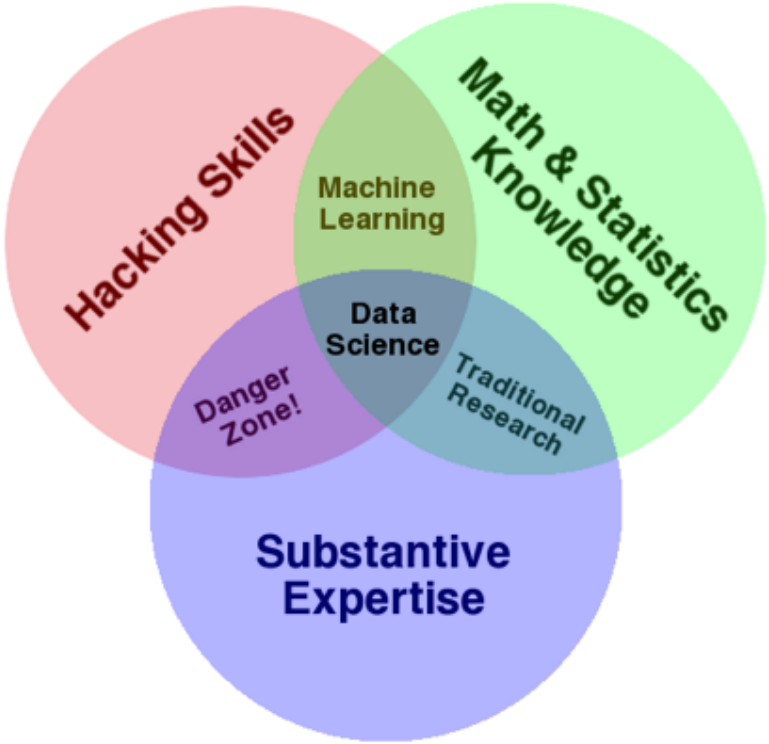
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

THE QUALITIES OF A DATA SCIENTIST



ONE MORE THING!

Communication skills

source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

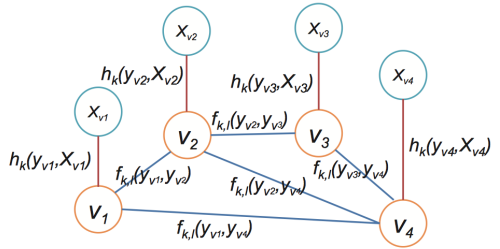


Figure 3: An example of factor graph with four users $\{v_1, v_2, v_3, v_4\}$. Each user v_i is associated with an attribute vector X_{v_i}. $h_k(y_{v_i}, X_{v_i})$ is the node feature function, whereas $f_{k,l}(y_{v_i}, y_{v_j})$ is the edge feature function defined on the edge between users v_i and v_j.

DEFINITION 4. *(**Factor Conditioning Optimization**)*

$$\min_{P_{v_i}} \frac{1}{2} P_{v_i}^T Q P_{v_i} \tag{10}$$

$$\textit{where} \quad Q_{kl} = \begin{cases} \sum_{m=1, m\neq k}^{r} \hat{r}_{m,k}^2(v_i, X_{v_i}), & k = l \\ -\hat{r}_{k,l}(v_i, X_{v_i}) \cdot \hat{r}_{l,k}(v_i, X_{v_i}), & k \neq l \end{cases}$$

LEMMA 2. *Factor Conditioning Optimization in Eq. 1 defines a convex quadratic programming problem.*

PROOF. For any non-negative vector z,

$$z^T Q z = \frac{1}{2} \sum_{k=1}^{r} \sum_{l=1}^{r} \left(\hat{r}_{k,l}(v_i, X_{v_i}) \cdot z_l - \hat{r}_{l,k}(v_i, X_{v_i}) \cdot z_k \right)^2 \geq 0 \quad (11$$

DEFINITION 5. *(**S**ocial **R**oles and **S**tatuses Inference Model [**SRS**]) The factor graph based social roles and statuses inference model is:*

$$P(Y) = \frac{1}{Z} \left(\prod_{v_i \in V, k} h_k(y_{v_i}, X_{v_i}) \right) \cdot \left(\prod_{v_i \in V} \prod_{v_j \in N(v_i), k, l} f_{k,l}(y_{v_i}, y_{v_j}) \right)$$

where Z is a normalization factor and k, l are the users v_i and v_j.

ONE MORE THING!

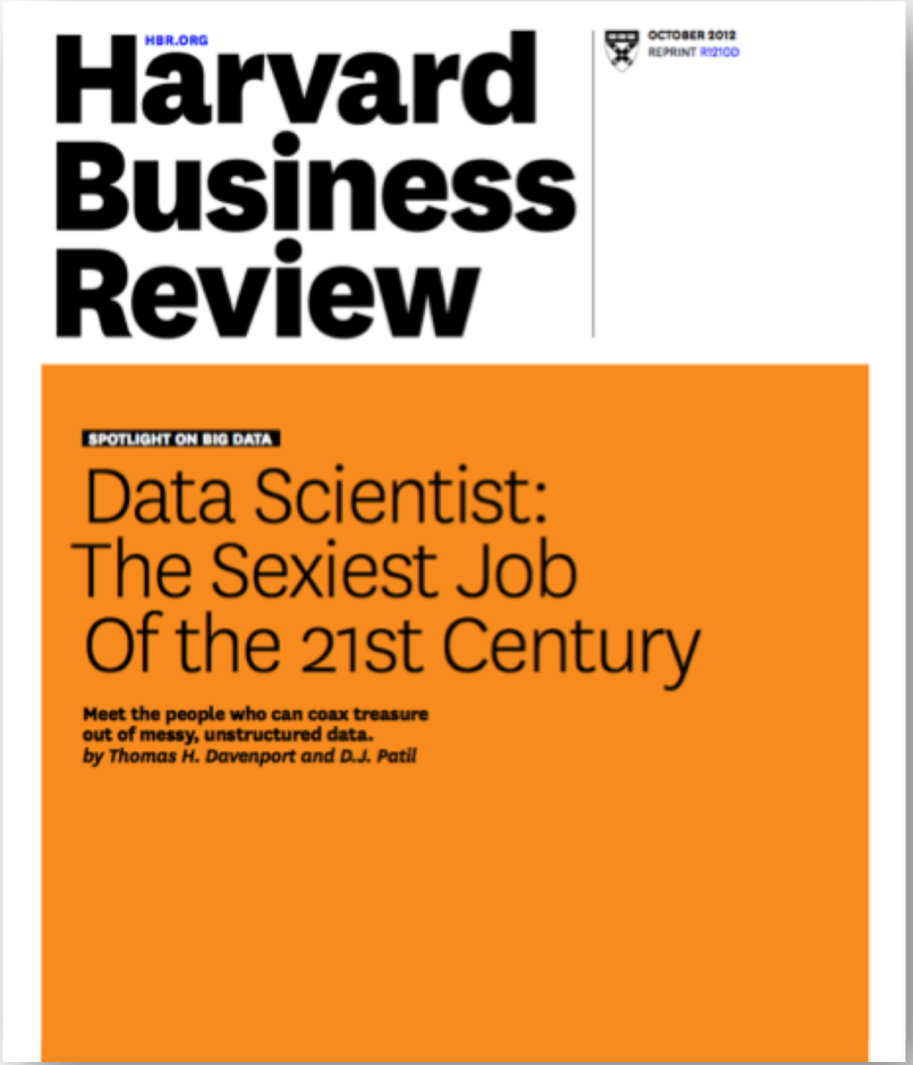
Communication skills

WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.

WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.


HBR.ORG
Harvard Business Review
OCTOBER 2012
REPRINT R1210D
SPOTLIGHT ON BIG DATA
Data Scientist: The Sexiest Job Of the 21st Century
Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

ForbesBrandVoice Connecting marketers to the Forbes audience. What is this?

BUSINESS 1/21/2014 @ 8:29AM | 9,168 views

Data Scientist: Sexiest Job Of The Century?

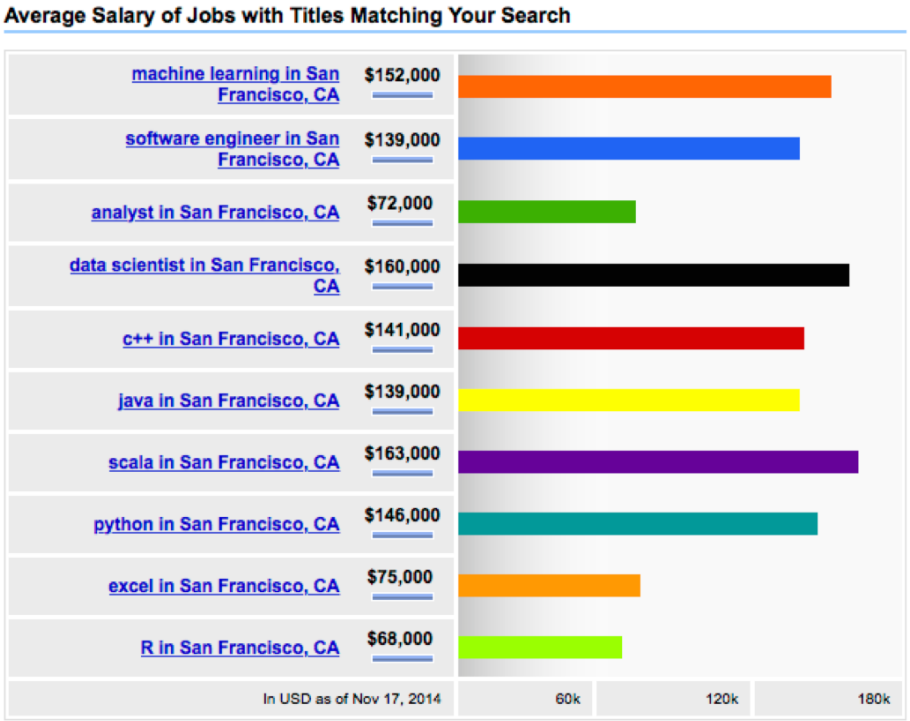
SAP Guest , SAP

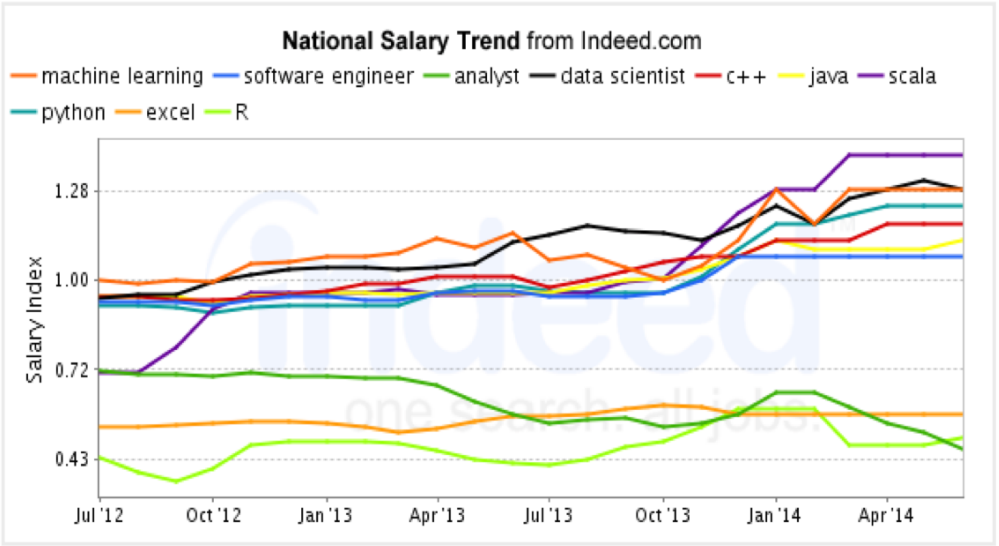
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE


Average Salary of Jobs with Titles Matching Your Search
machine learning in San Francisco, CA
$152,000
software engineer in San Francisco, CA
$139,000
analyst in San Francisco, CA
$72,000
data scientist in San Francisco, CA
$160,000
c++ in San Francisco, CA
$141,000
java in San Francisco, CA
$139,000
scala in San Francisco, CA
$163,000
python in San Francisco, CA
$146,000
excel in San Francisco, CA
$75,000
R in San Francisco, CA
$68,000
In USD as of Nov 17, 2014
60k
120k
180k


National Salary Trend from Indeed.com
machine learning
software engineer
analyst
data scientist
c++
java
scala
python
excel
R
Salary Index
1.28
1.00
0.72
0.43
Jul '12
Oct '12
Jan '13
Apr '13
Jul '13
Oct '13
Jan '14
Apr '14

JOB MARKET

Principal Data Scientist
Cablevision
San Francisco, CA • Apr 21, 2015
1 connection to the poster • Similar

Data Scientist
Groupon
Palo Alto, CA, US • Apr 27, 2015
5 connections to the poster • Similar

Sr./Principal Scientist, Machine Learn Mining
Nokia Technologies
Sunnyvale • Apr 20, 2015
3 connections to the poster • Similar

Data Scientist – Just Closed $15M in
FILD
Palo Alto, CA • Apr 27, 2015
3 people in your network • Similar

Senior Data Scientist
salesforce.com
US - California - San Francisco (HQ) • Apr 20, 201
1,667 people in your network • Similar

Data Scientist/Economist
Glassdoor
San Francisco Bay Area • Apr 27, 2015
87 people in your network • Similar

Sr. Data Scientist
Esurance
San Francisco • Apr 24, 2015
1 connection to the poster • Similar

Data Scientist
Equinix
Sunnyvale, CA, US • Apr 21, 2015
116 people in your network • Similar

Principal Data Scientist
Thomson Reuters
San Francisco, CA, US • Apr 18, 2015 • From jobs.thomsonreuters.com
532 people in your network • Similar

Principal Data Scientist - Security Secto
Pivotal Software, Inc.
Palo Alto or San Francisco, CA • Mar 13, 2015
19 connections to the poster • Similar

Data Scientist, Analytics (Instagram)
Facebook
Menlo Park -California -US • Apr 21, 2015
2,315 people in your network • Similar

Data Scientist - Senior Analytics Specialist
Airbnb
San Francisco, California US • Apr 22, 2015
478 people in your network • Similar

Data Scientist, Strategic Analytics
Castlight Health
San Francisco, CA • Apr 14, 2015
59 people in your network • Similar

Data Scientist Intern
Move, Inc
San Jose, CA, US • Apr 24, 2015 • From chk.tbe.taleo.net
62 people in your network • Similar

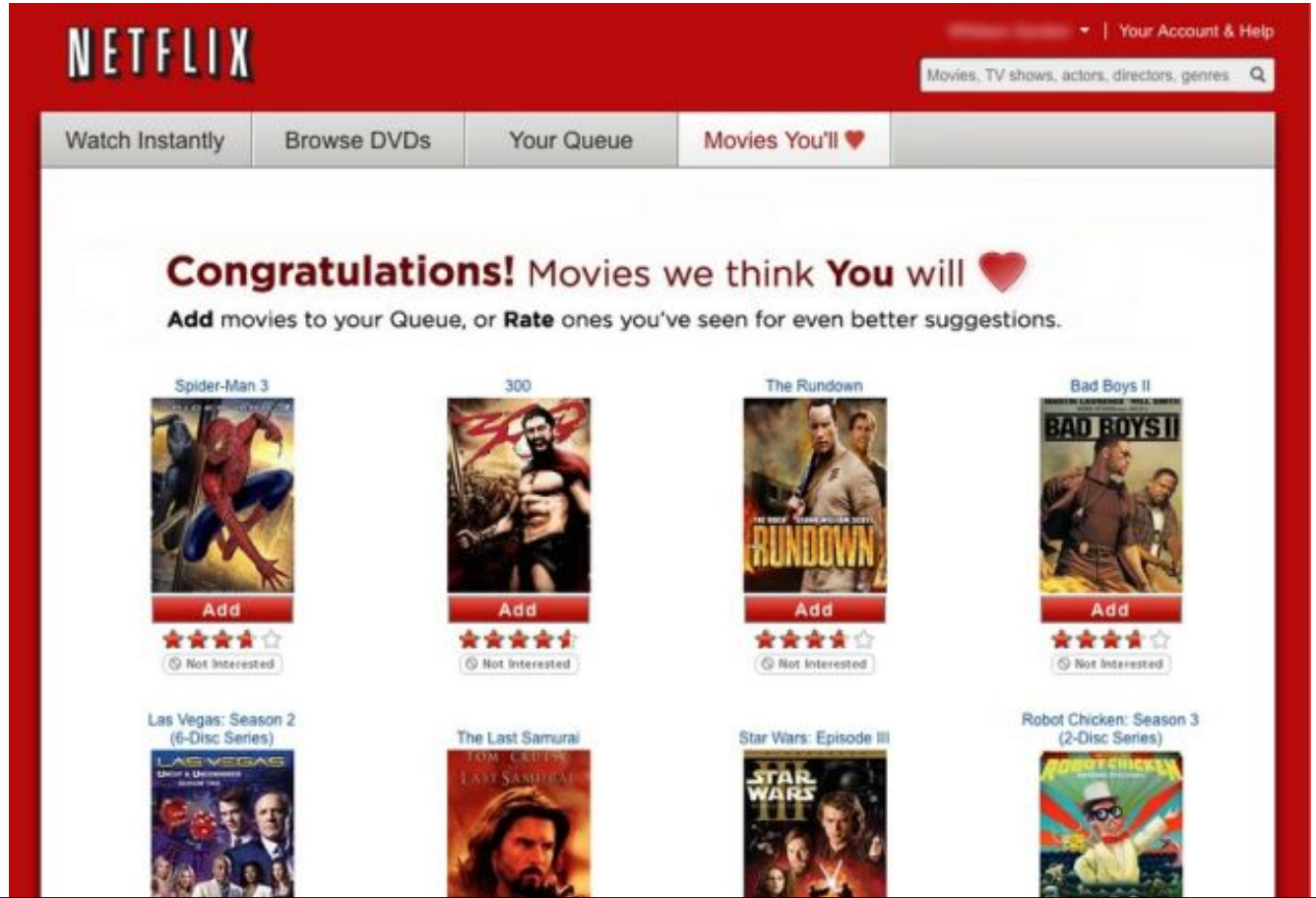
Data Scientist
Walmart eCommerce
San Bruno, CA • Apr 23, 2015
421 people in your network • Similar

Data Scientist (Risk and Analysis)
Better Finance, Inc.
San Francisco, CA • Apr 21, 2015
13 people in your network • Similar

Senior Data Scientist
Criteo
Palo Alto, CA, US • Apr 20, 2015
1 connection to the poster • Similar

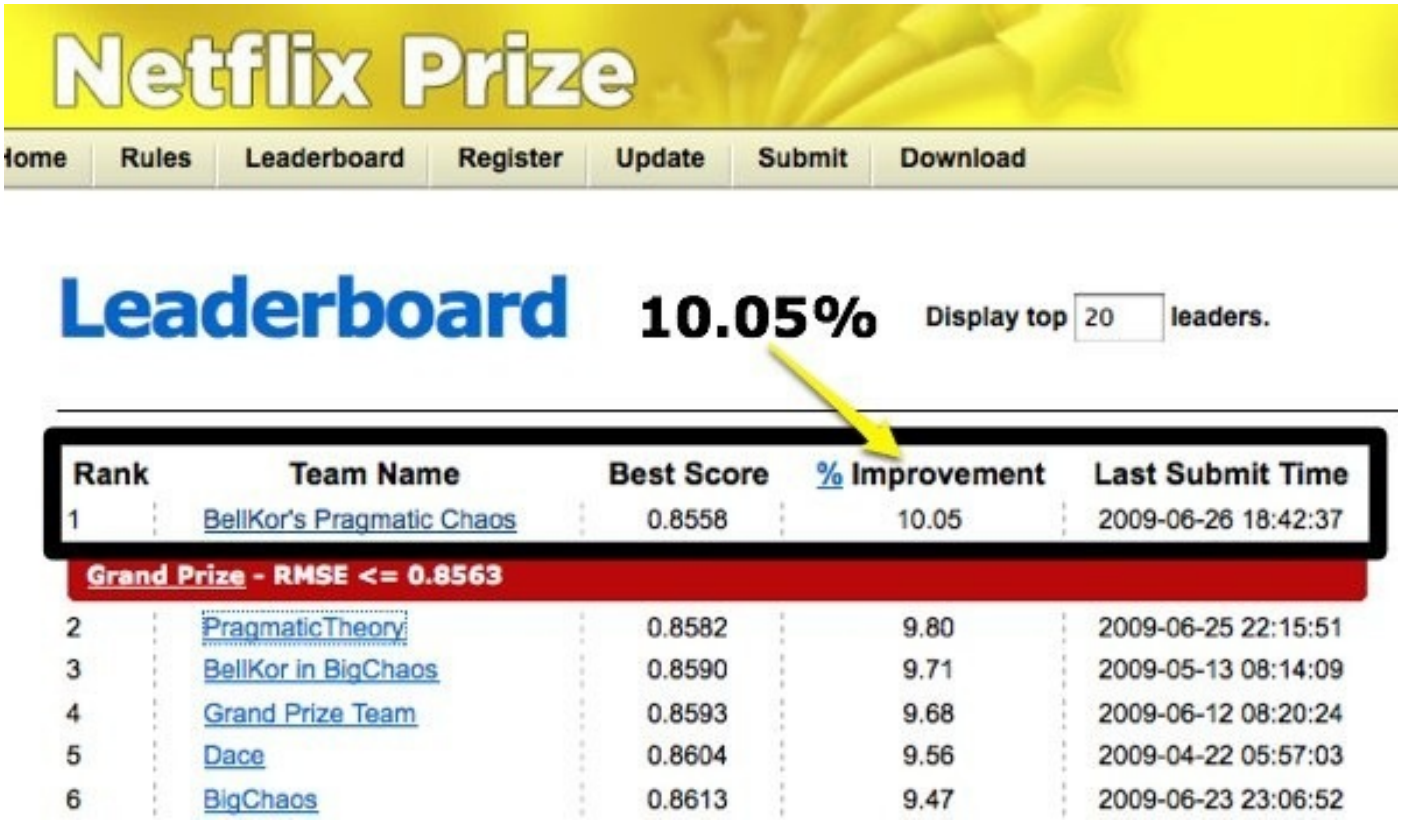
Data Scientist
Capital One
San Francisco - California - USA • Apr 27, 2015
623 people in your network • Similar

ANOTHER MOTIVATOR



award $1 million to anyone who can improve movie recommendation by 10%

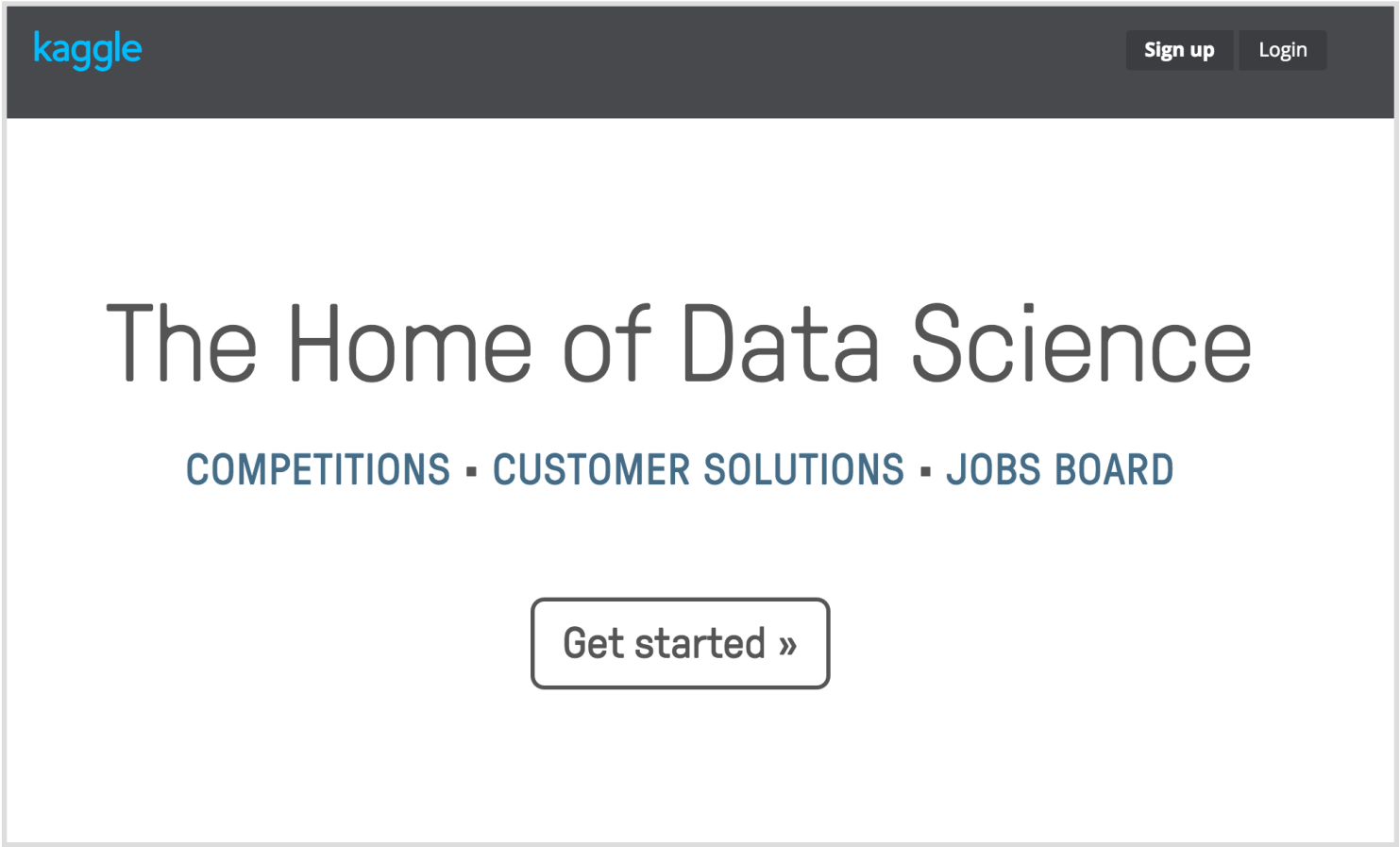
NETFLIX CHALLENGE



Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52


2009
NETFLIX
DATE 09.21.09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION
00/100
FOR The Netflix Prize
Reed Hastings

MORE CHALLENGES



WHO USES DATA SCIENCE?

WHO USES DATA SCIENCE?

- Stack Overflow tag recommendation and response time prediction
- Locating ethnic food in ethnic neighborhoods
- Building optimal NBA teams
- Recommending new musical artists
- Prioritize emergency calls in Seattle
- Finding the right college for you

Music + Data:

http://bit.ly/echonest

Michael E. Driscoll
@medriscoll

Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

Reply Retweet Favorite More Pocket

WHAT MAKES A GOOD DATA SCIENTIST?

- Statistical and machine learning knowledge
- Engineering experience
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

II. THE DATA SCIENCE WORKFLOW

Dataists

- 1. Obtain
- 2. Scrub
- 3. Explore
- 4. Model
- 5. Interpret

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Jeff Hammerbacher

- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ted Johnson

- 1. Assemble an accurate and relevant data set
- 2. Choose the appropriate algorithm

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ben Fry

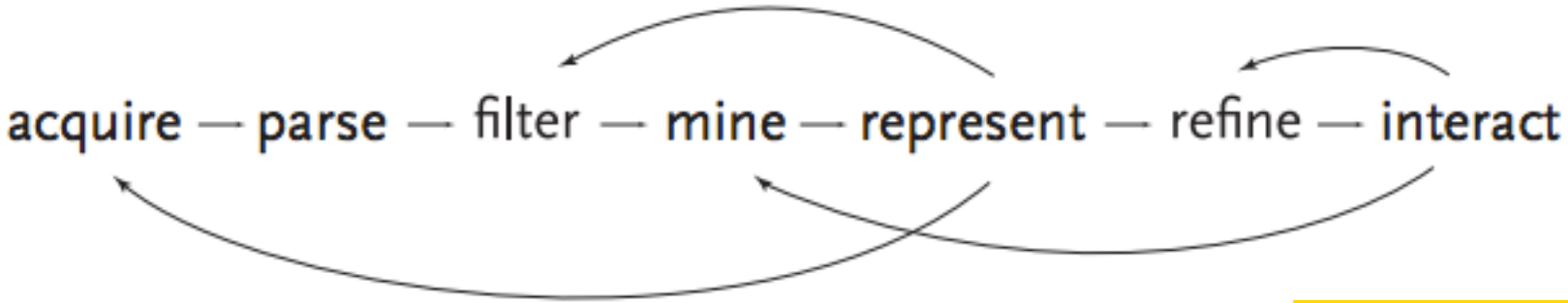
- 1. Acquire
- 2. Parse
- 3. Filter
- 4. Mine
- 5. Represent
- 6. Refine
- 7. Interact

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

THE DATA SCIENCE WORKFLOW

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



NOTE

This diagram illustrates the iterative nature of problem solving

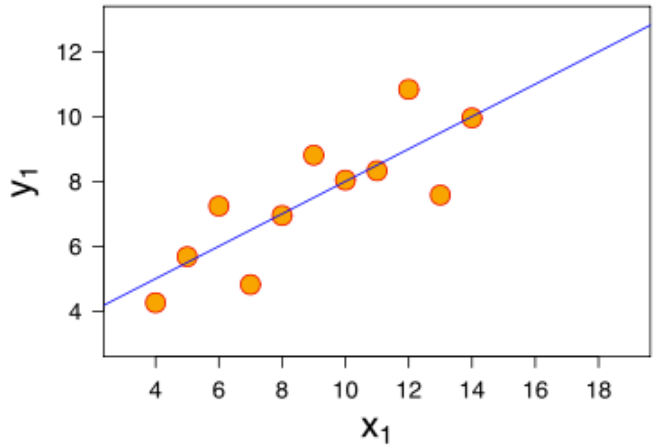
source: http://benfry.com/phd/dissertation-110323c.pdf

VISUALIZATIONS AS A MEDIUM

EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

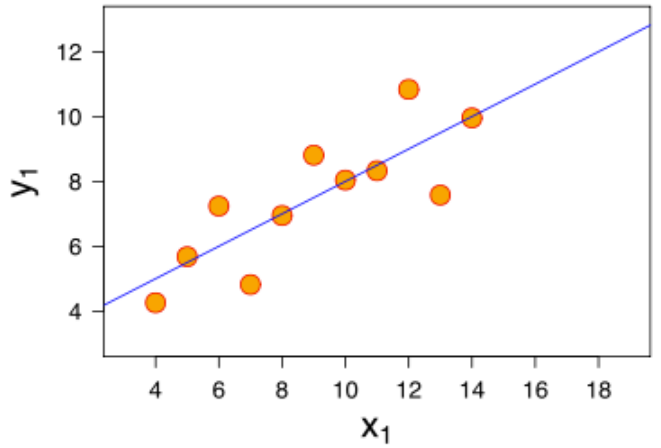
- *eleven (x, y) points*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

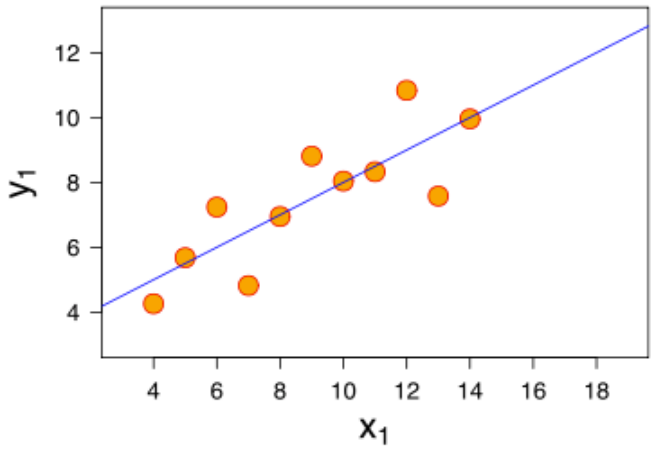
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

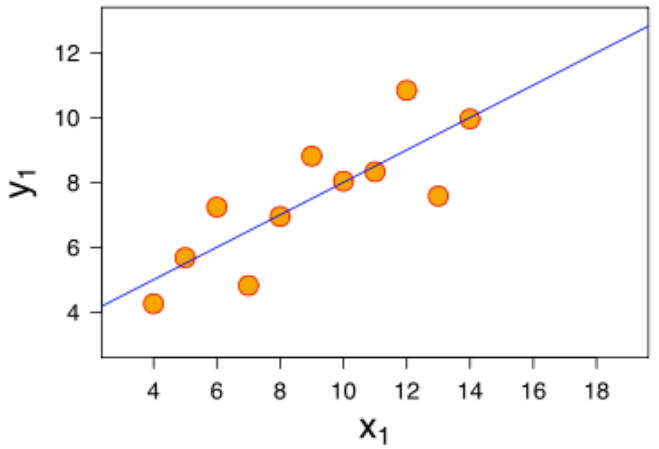
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

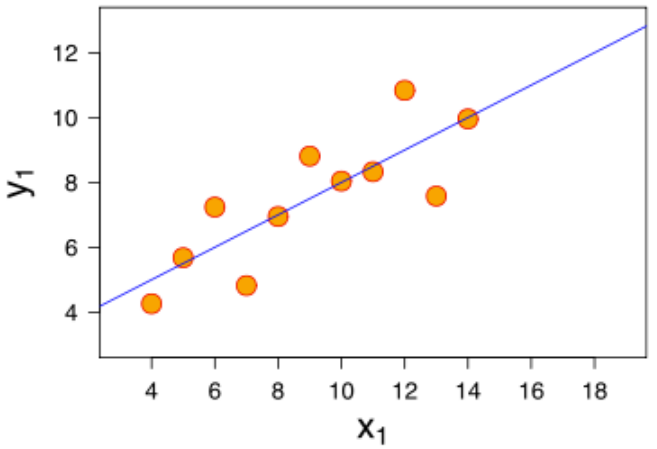
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x and y = 0.8*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

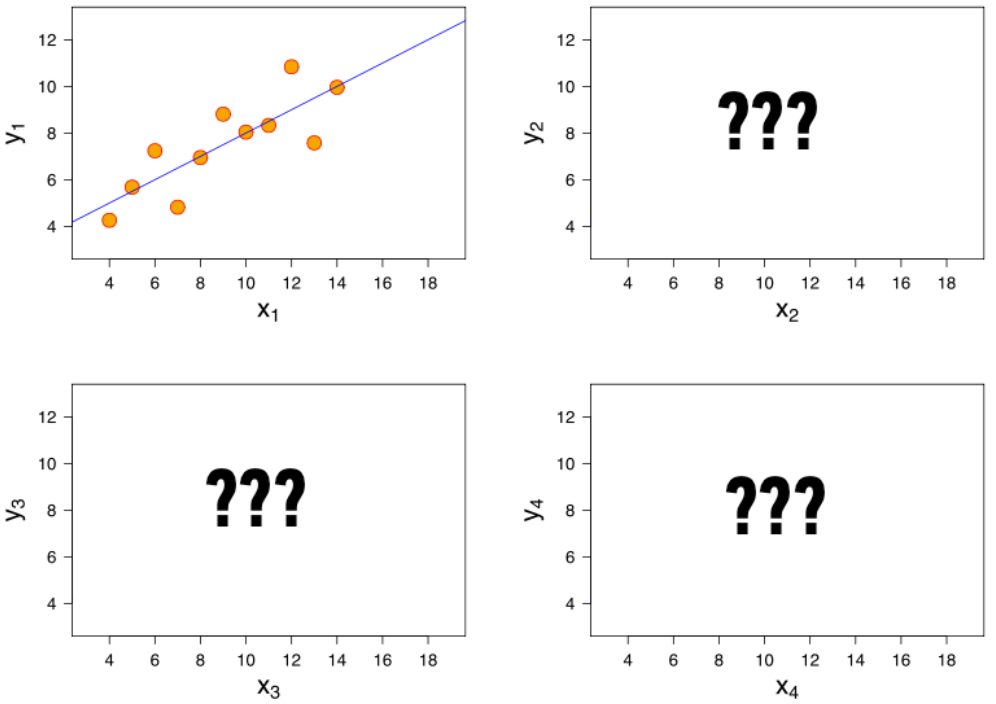
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

Now, suppose I give you three more datasets with exactly the same characteristics...

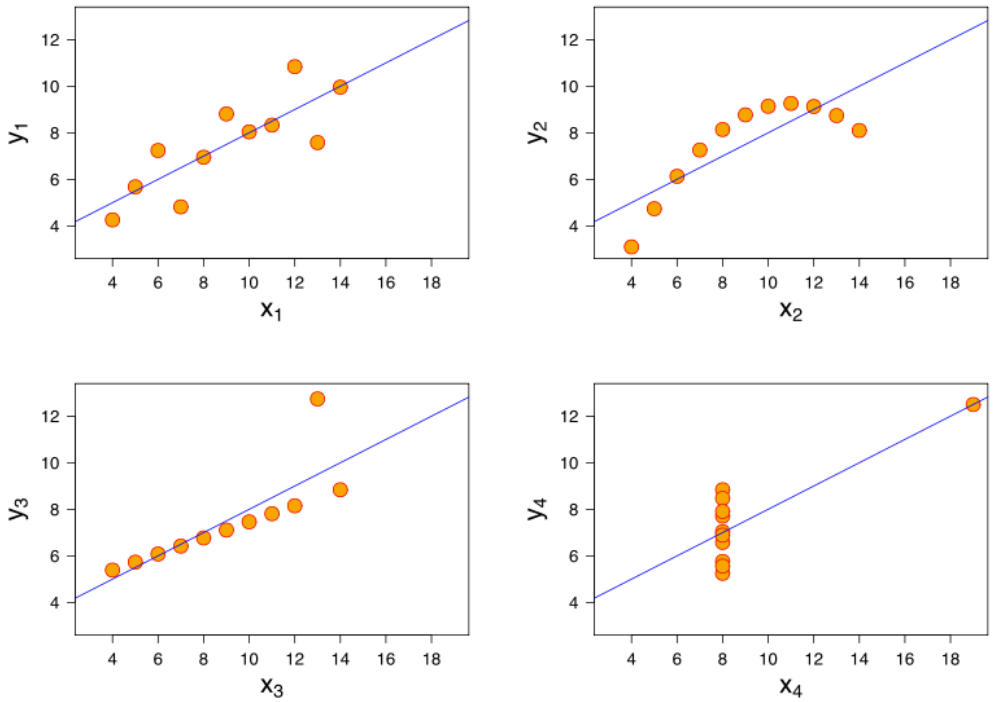
Q: how similar are these datasets?



Now, suppose I give you three more datasets with exactly the same characteristics.

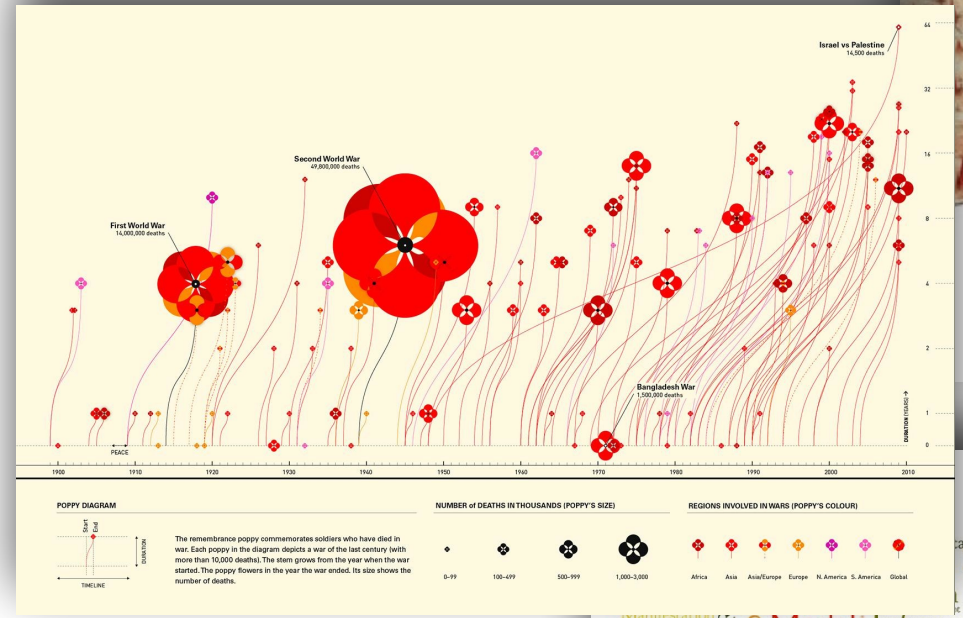
Q: how similar are these datasets?

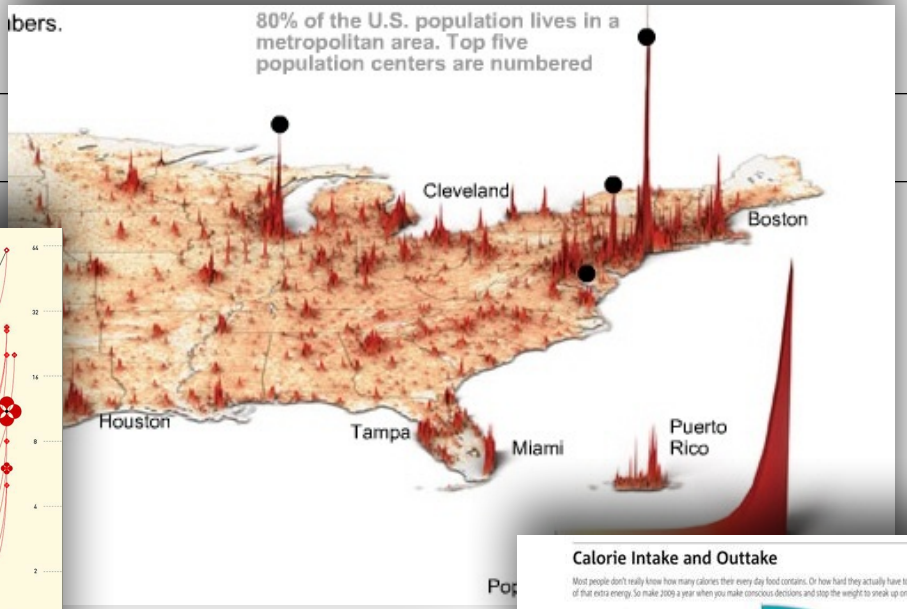
A: not very!

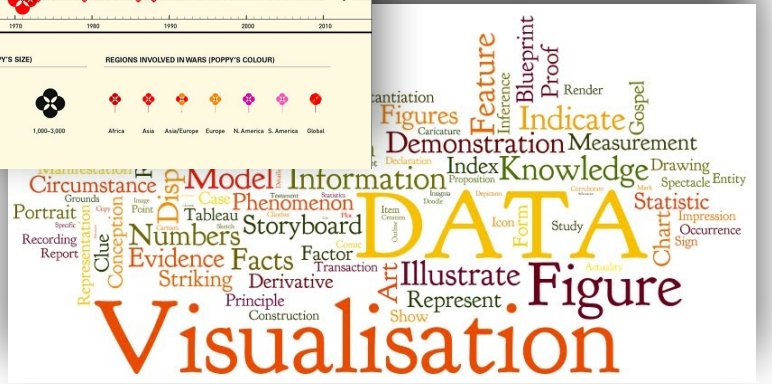


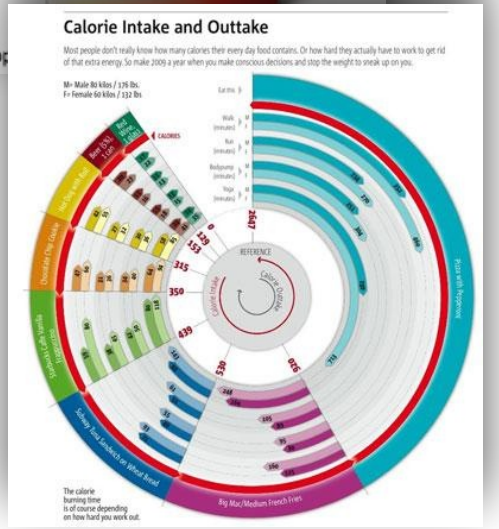
http://en.wikipedia.org/wiki/Anscombe's_quartet

VISUALIZATION: BEING CREATIVE



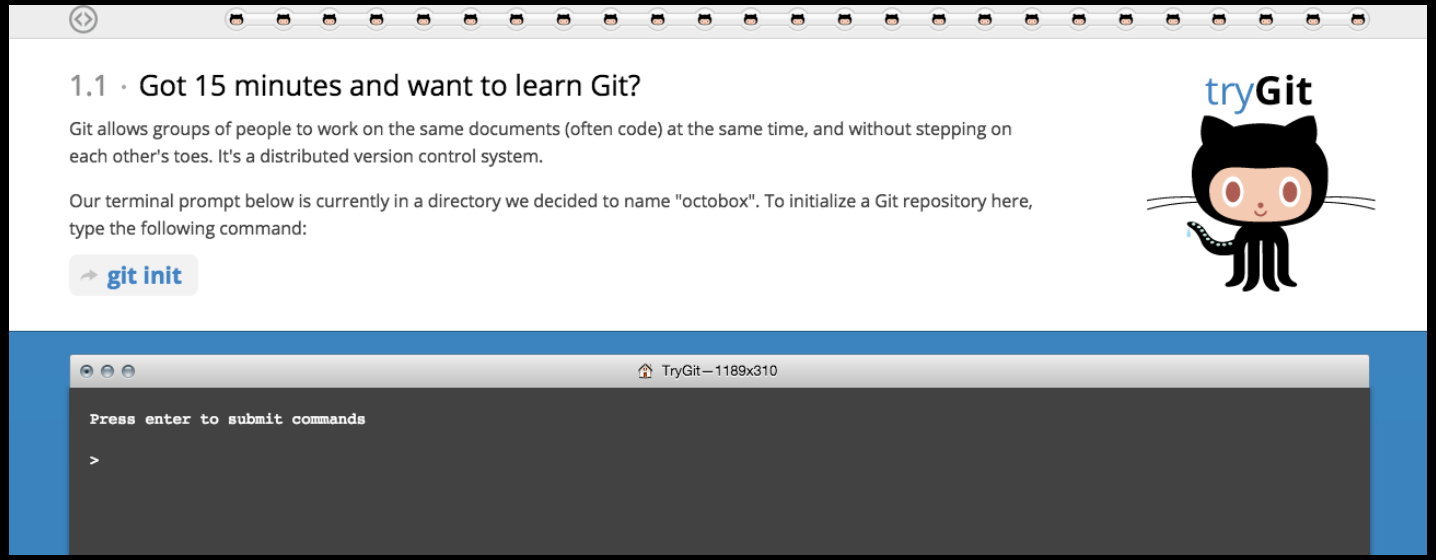




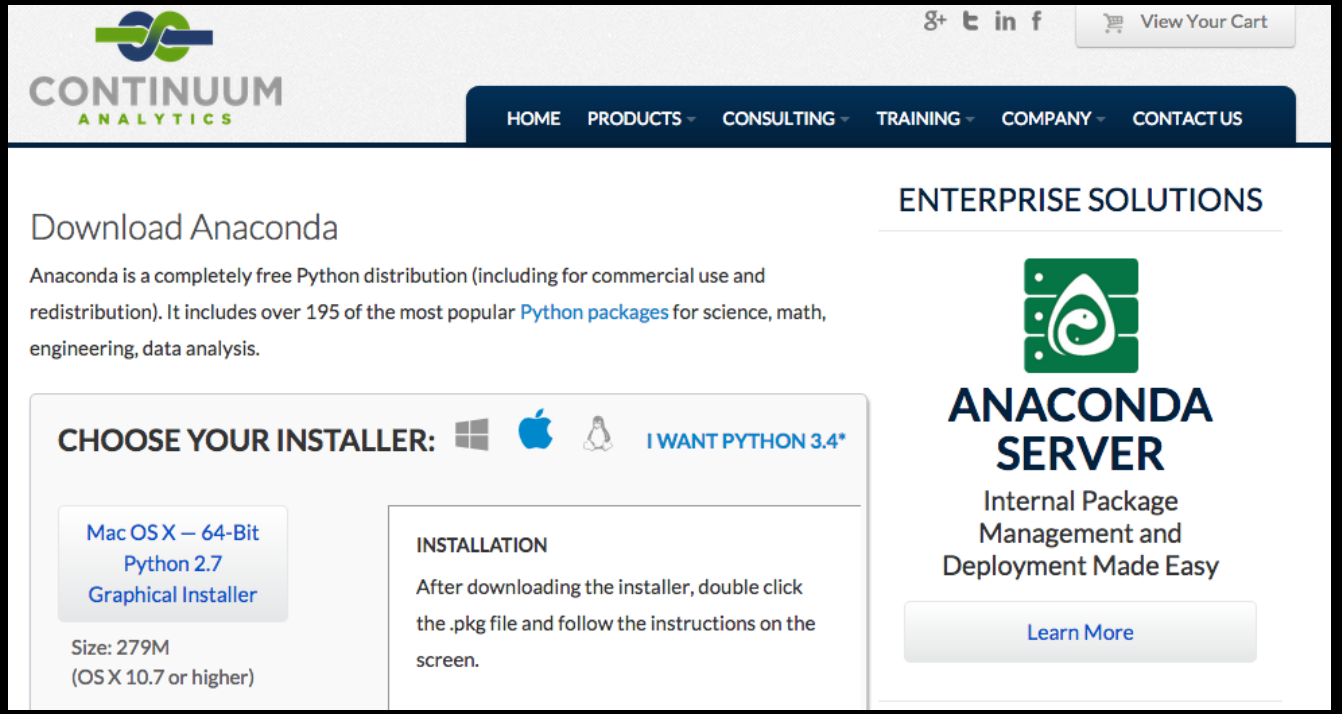


LAB: INTRO TO GITHUB

HTTP://TRY.GITHUB.COM/



DOWNLOAD ANACONDA



Q&A

APPENDIX: WORKING AT THE UNIX COMMAND LINE

EXERCISE – WORKING AT THE UNIX COMMAND LINE

KEY OBJECTIVES

- Navigate the filesystem
- Create, move, copy, and delete files & directories
- View & search files
- Edit & interact with files
- Combine steps
- Learn more

TOOLS

- ls, cd
- cat, touch, mv, cp, mkdir, rm, rmdir
- head, tail, less, cat, grep
- vim, tr, sort, uniq, wc
- pipe (|)
- man, apropos

NOTE

Being comfortable at the command line makes your life much easier!