

INTRO TO DATA SCIENCE

INTRODUCTION TO DATA SCIENCE

INTRO TO DATA SCIENCE

WELCOME!

MEET YOUR INSTRUCTIONAL TEAM

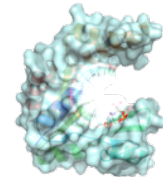
▸ Francesco Mosconi



- Francesco Mosconi is a Data Scientist at Catalit LLC. He is formerly the Chief Data Officer at Spire, a company that invented the first consumer wearable device capable of continuously tracking respiration and activity. He worked as consultant for Roche Ltd. and for Socialbakers, a social media data analytics company. Passionate about data and technology, he was selected in 2011 for the graduate studies program at Singularity University. He earned a joint PhD in biophysics at University of Padua and Université de Paris VI and has a master degree in theoretical physics.



**GENERAL
ASSEMBLY**



Catalit LLC

MEET YOUR INSTRUCTIONAL TEAM

▸ Dylan Hercher



- My name is Dylan Hercher. I am a recent transplant from New York to San Francisco, I moved around 2 months ago! I received my BS from Colgate University in 2013, where I studied Computer Science and Mathematics. I have been working with General Assembly since my Junior year as a contractor, and joined full time as a Business Intelligence Engineer and Analyst after graduating. Over the past few years I have worked with as much data as I possibly could! Along with developing models and teaching Data Science for GA, I have worked on varied projects from developing a drinks prediction model to forecasting real-estate growth.



0. INTRODUCTION

1. WHAT IS DATA SCIENCE?

2. THE DATA MINING WORKFLOW

LAB:

3. GITHUB & IPYTHON

4. Q&A

LEARNING OBJECTIVE

- Describe the data mining workflow and the key traits of a successful data scientist.
- Set up github account.
- Familiarize with python and iPython

Instructor:

Francesco Mosconi (FRANCESCO+GA@MOSCONI.ME)

Expert-in-residence:

Dylan Hercher (DYLAN@GENERALASSEMB.LY)

Course Producer:

Vanessa Ohta

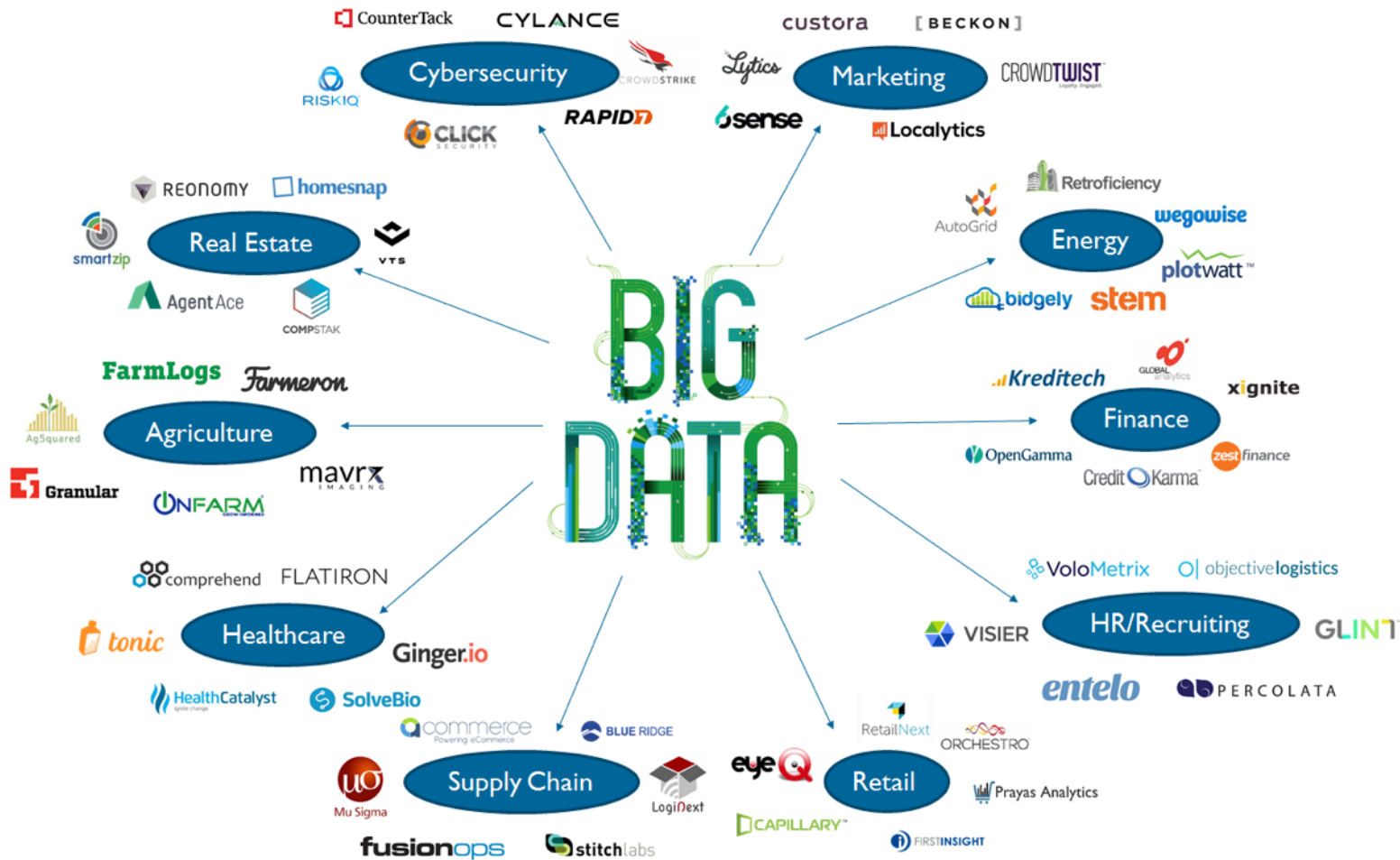
Course Times: 6:30pm-9:30pm, Tuesdays and Thursdays

Couse materials: [HTTPS://GITHUB.COM/GA-STUDENTS/DAT_SF_18](https://github.com/GA-STUDENTS/DAT_SF_18)

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?



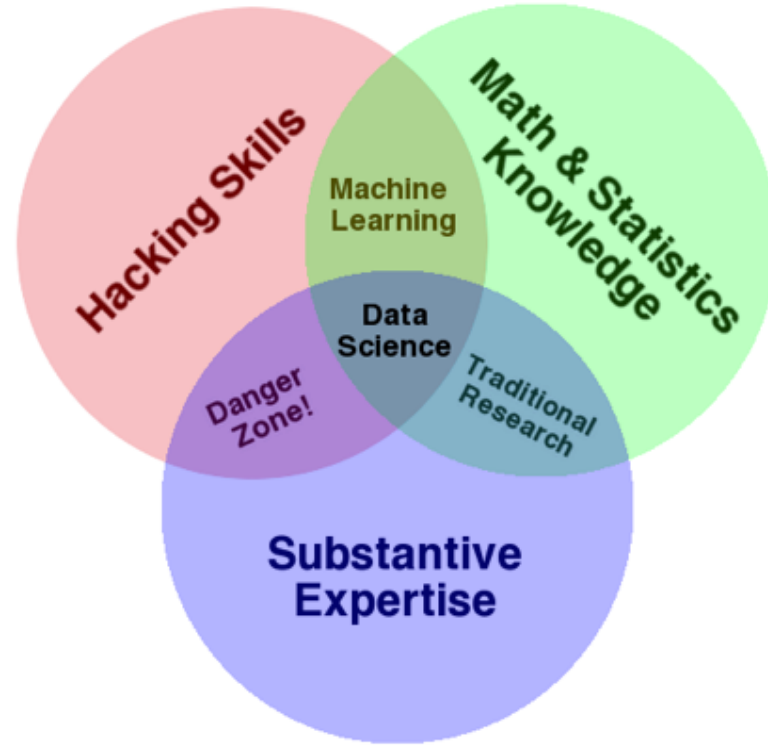
WHAT IS DATA SCIENCE?

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

WHAT IS DATA SCIENCE?

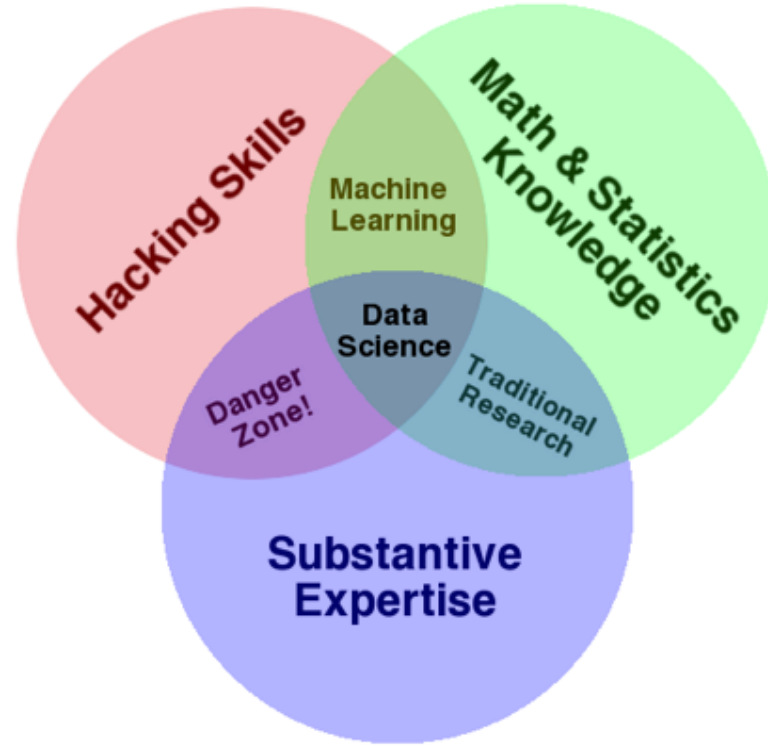
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

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

THE QUALITIES OF A DATA SCIENTIST

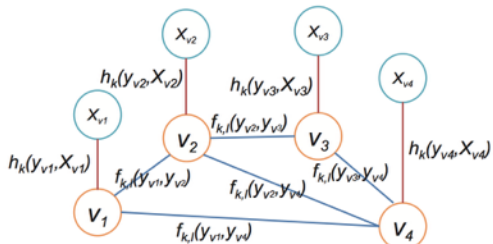


Figure 3: An example of factor graph with four users $\{v_1, v_2, v_3, v_4\}$. Each user v_i is associated with an attribute vector X_{v_i} . $h_k(y_{v_i}, X_{v_i})$ is the node feature function, whereas $f_{k,l}(y_{v_i}, y_{v_j})$ is the edge feature function defined on the edge between users v_i and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

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

PROOF. For any non-negative vector z ,

$$z^T Q z =$$

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

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

$$\min_{P_{v_i}} \frac{1}{2} P_{v_i}^T Q P_{v_i} \quad (10)$$

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

DEFINITION 5. (**Social Roles and Statuses 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 R12100

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

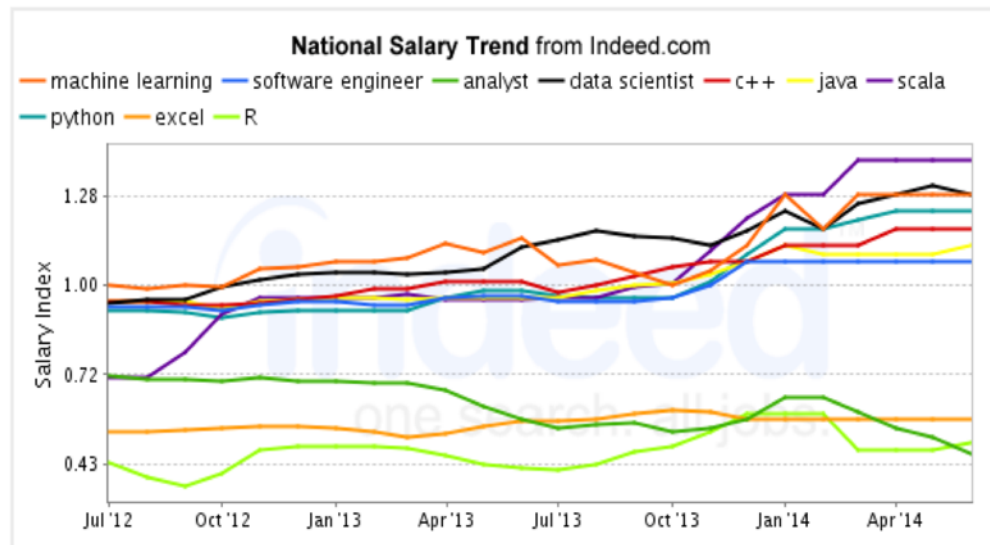
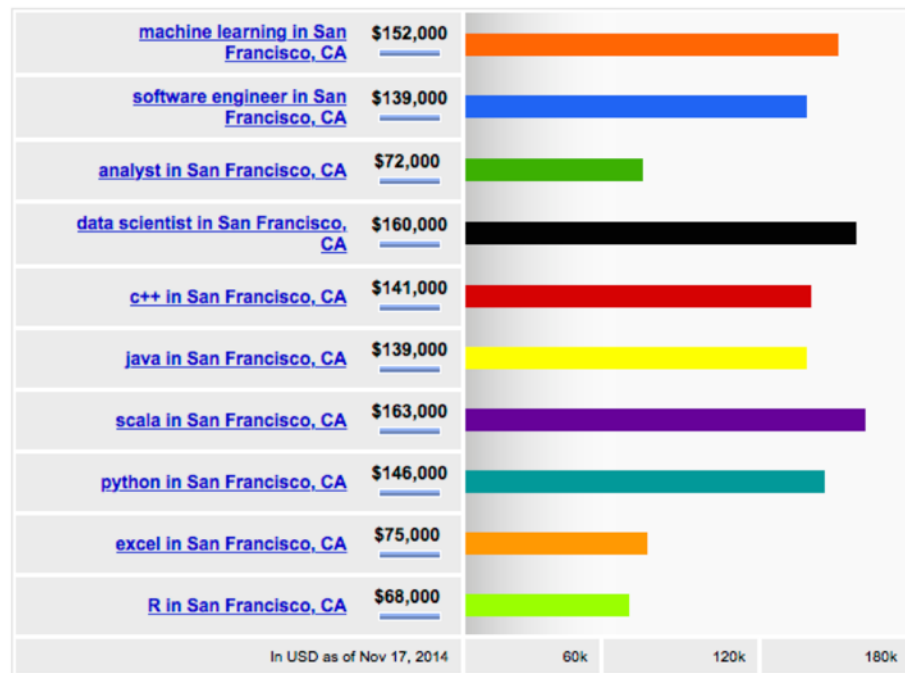
DATA

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



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 Learning

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, 2015
▶ 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 Sector

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

The screenshot shows the Netflix homepage with a red header. The Netflix logo is on the left, and a search bar is on the right. Below the header is a navigation bar with tabs: 'Watch Instantly', 'Browse DVDs', 'Your Queue', and 'Movies You'll ❤️'. The main content area features a 'Congratulations! Movies we think You will ❤️' message. Below this, there are eight movie recommendations arranged in two rows of four. Each recommendation includes a movie poster, the title, an 'Add' button, a star rating (four red stars and one grey star), and a 'Not Interested' link.

NETFLIX | Your Account & Help

Movies, TV shows, actors, directors, genres

Watch Instantly | Browse DVDs | Your Queue | **Movies You'll ❤️**

Congratulations! Movies we think **You** will ❤️

Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3	300	The Rundown	Bad Boys II
Add	Add	Add	Add
★★★★☆	★★★★☆	★★★★☆	★★★★☆
Not Interested	Not Interested	Not Interested	Not Interested

Las Vegas: Season 2 (6-Disc Series)	The Last Samurai	Star Wars: Episode III	Robot Chicken: Season 3 (2-Disc Series)

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

Netflix Prize

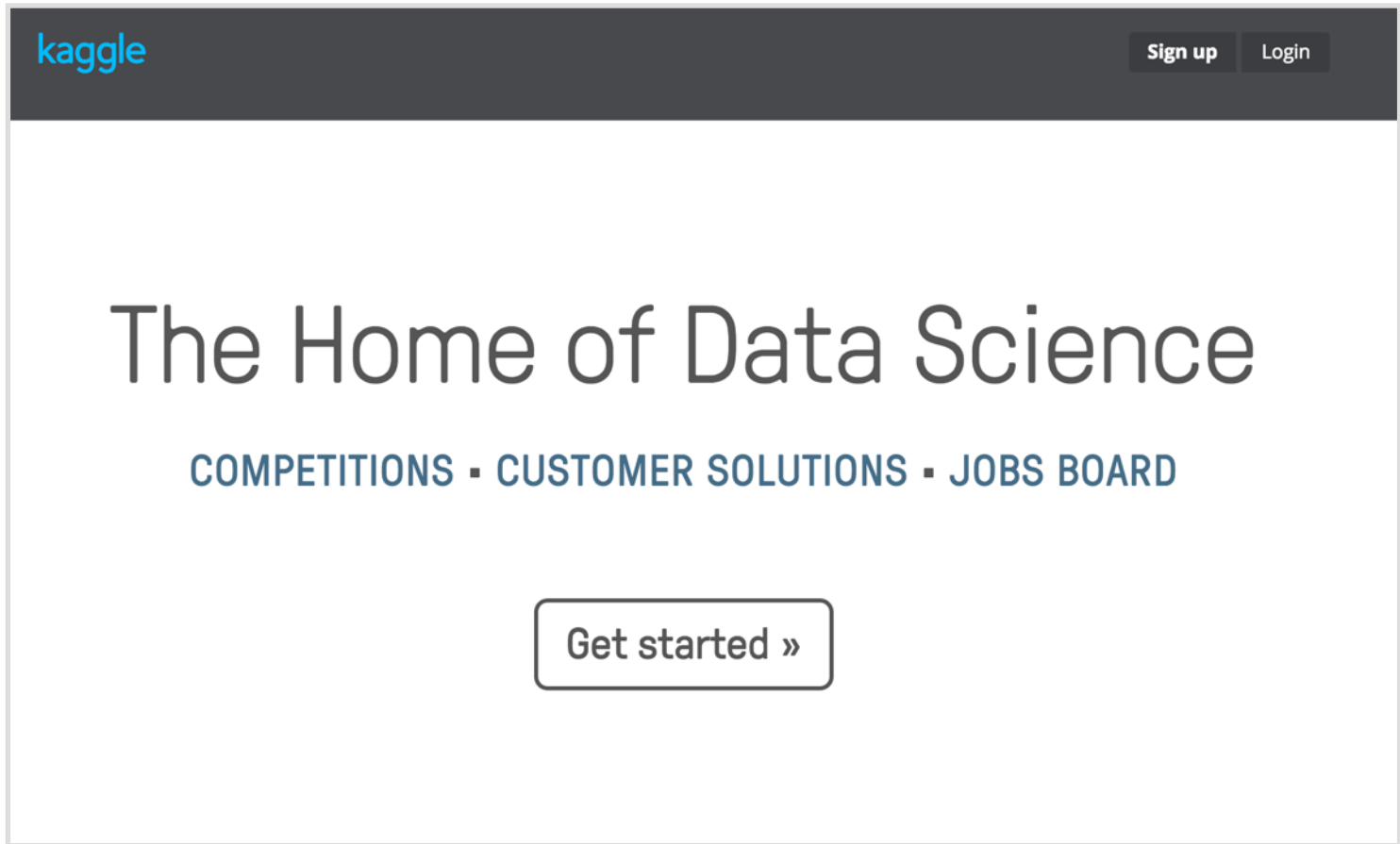
Home Rules Leaderboard Register Update Submit Download

Leaderboard

10.05% Display top leaders.

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 \leq 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





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

Jeff Hammerbacher

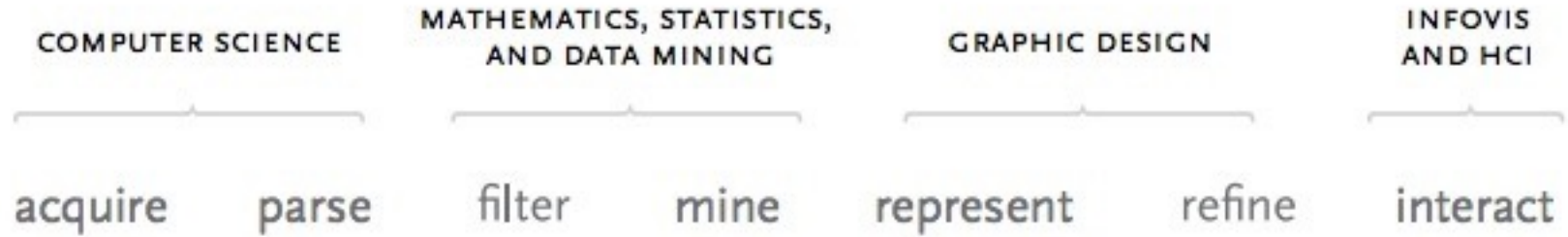
- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

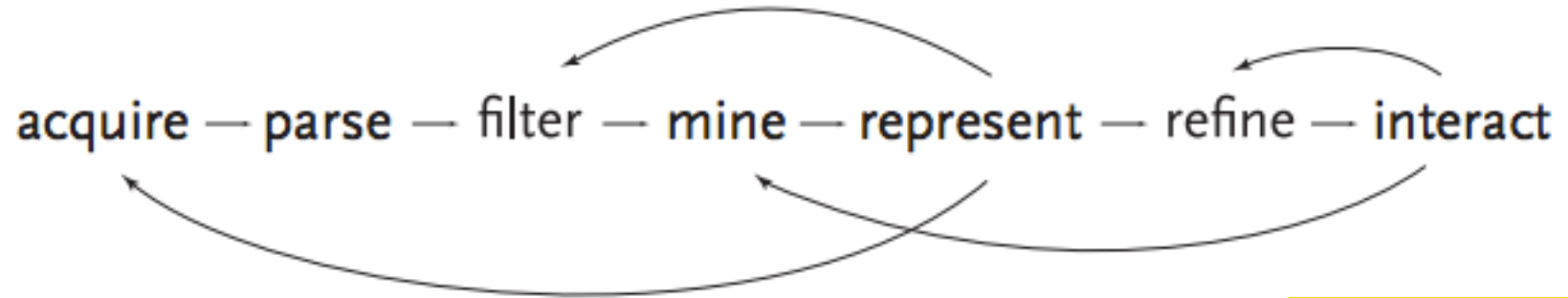
Ted Johnson

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

Ben Fry

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





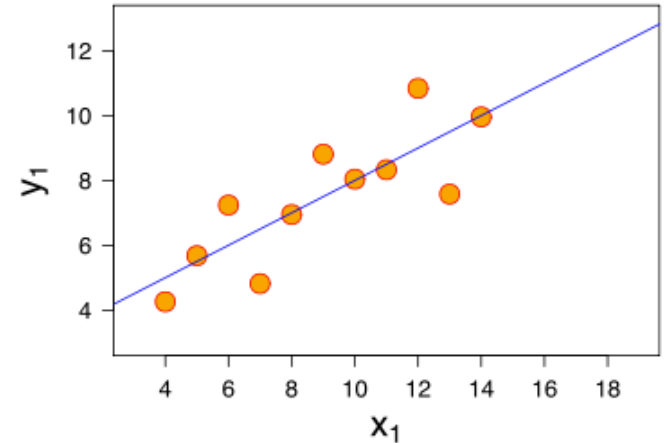
NOTE

This diagram illustrates the iterative nature of problem solving

VISUALIZATIONS AS A MEDIUM

Consider the following dataset:

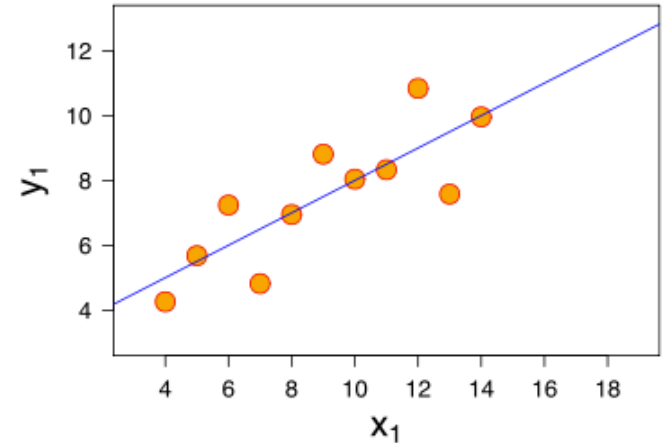
- *eleven (x, y) points*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

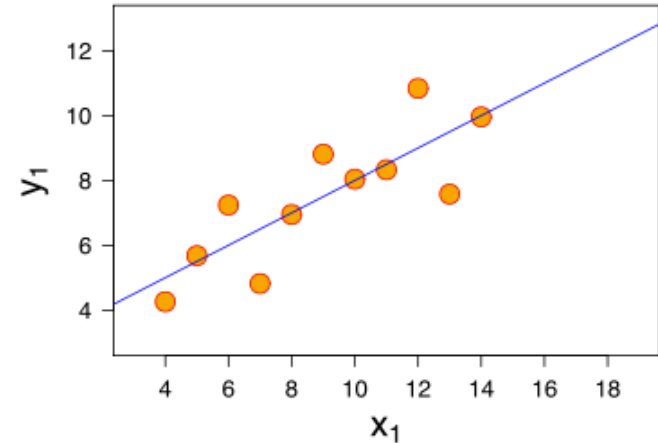
- *eleven (x, y) points*
- *mean of $x = 9$, mean of $y = 7.5$*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

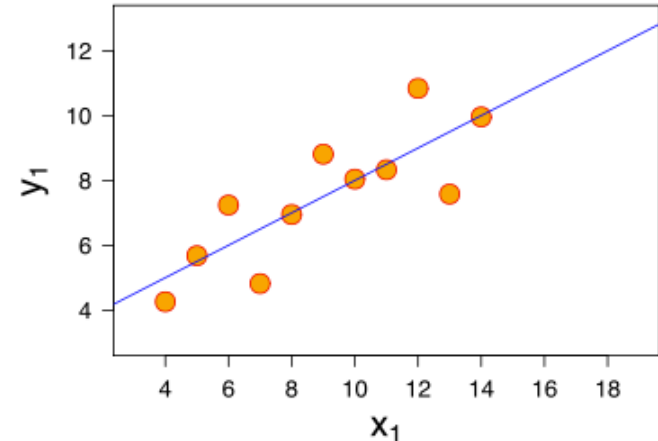
- *eleven (x, y) points*
- *mean of $x = 9$, mean of $y = 7.5$*
- *variance of $x = 11$, variance of $y = 4$*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

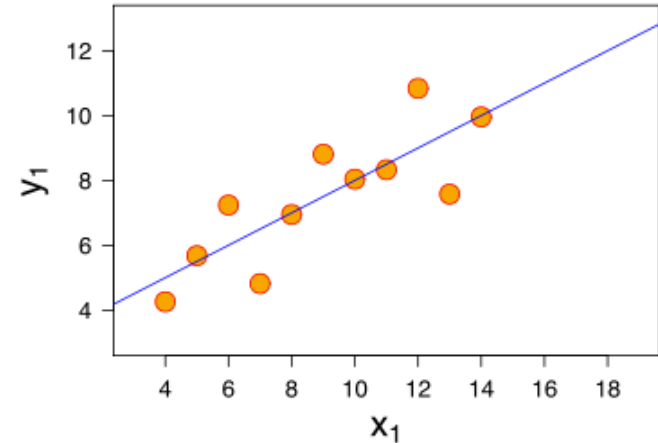
- *eleven (x, y) points*
- *mean of $x = 9$, mean of $y = 7.5$*
- *variance of $x = 11$, variance of $y = 4$*
- *correlation of x and $y = 0.8$*



EXERCISE – WHY VISUALIZE DATA?

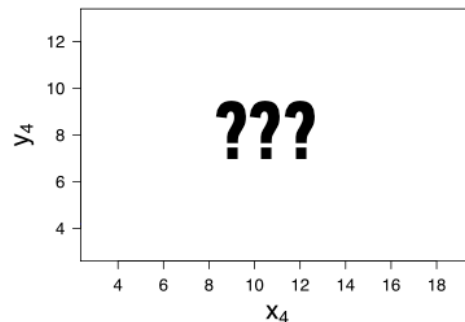
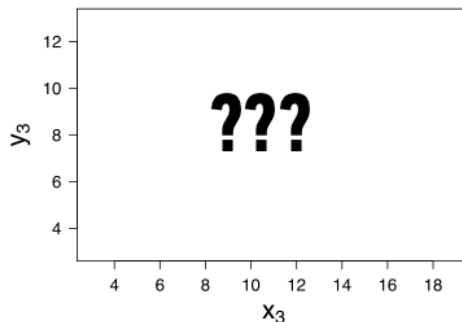
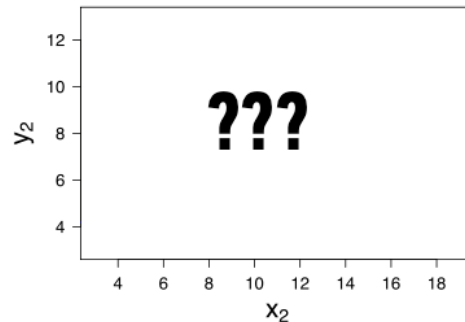
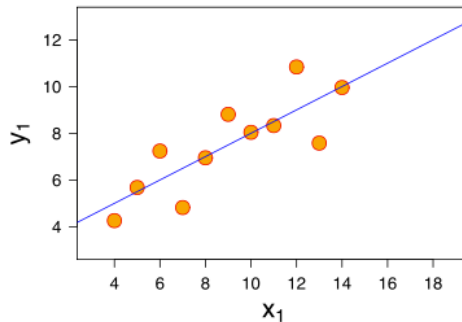
Consider the following dataset:

- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4*
- *correlation of x and y = 0.8*
- *line of best fit: $y = 3.00 + 0.500x$*



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

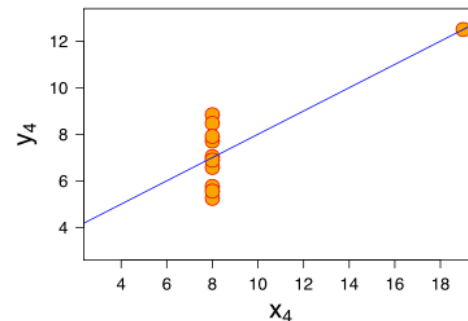
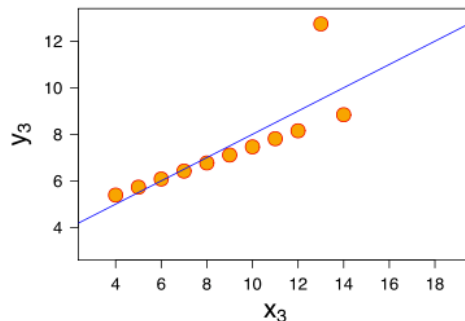
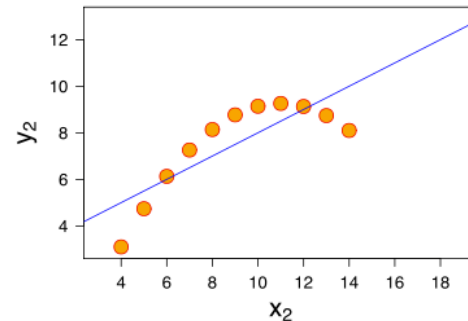
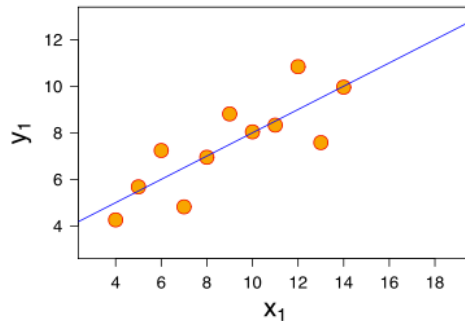
*Q: how similar are
these datasets?*



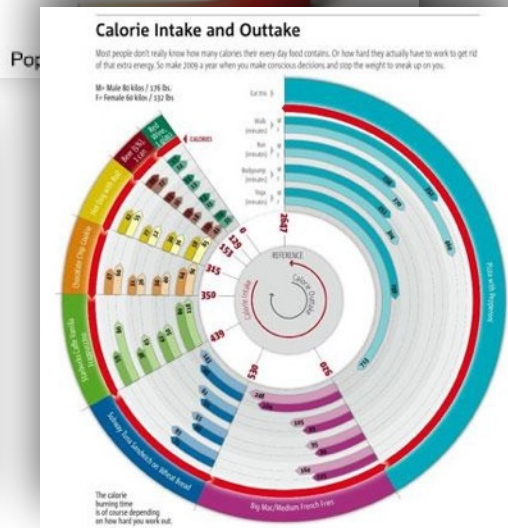
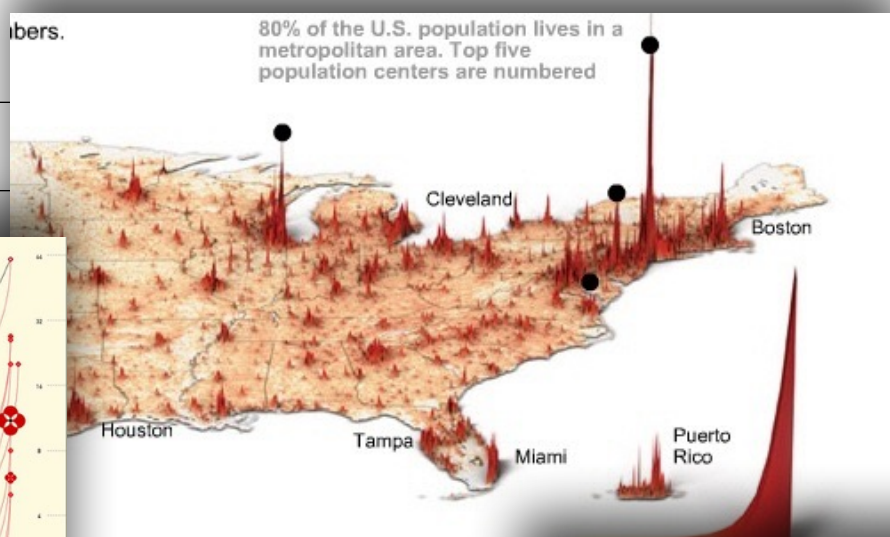
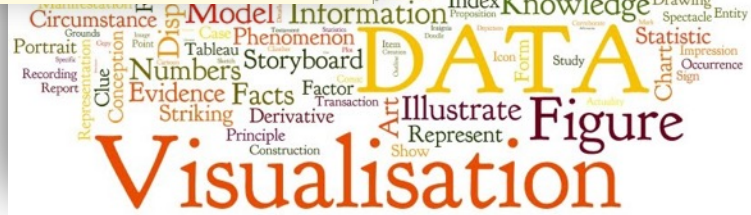
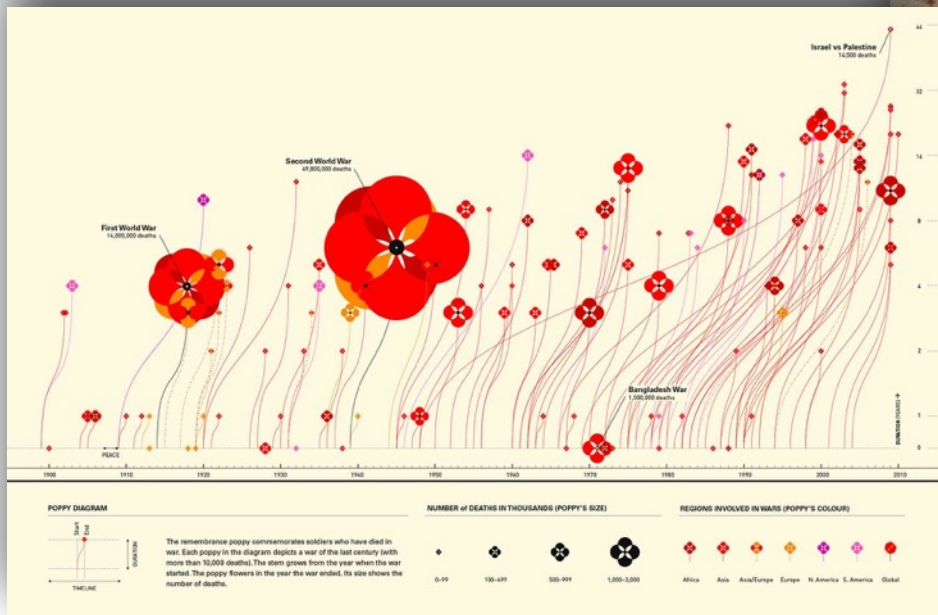
Now, suppose I give you three more datasets with exactly the same characteristics.

Q: how similar are these datasets?

A: not very!

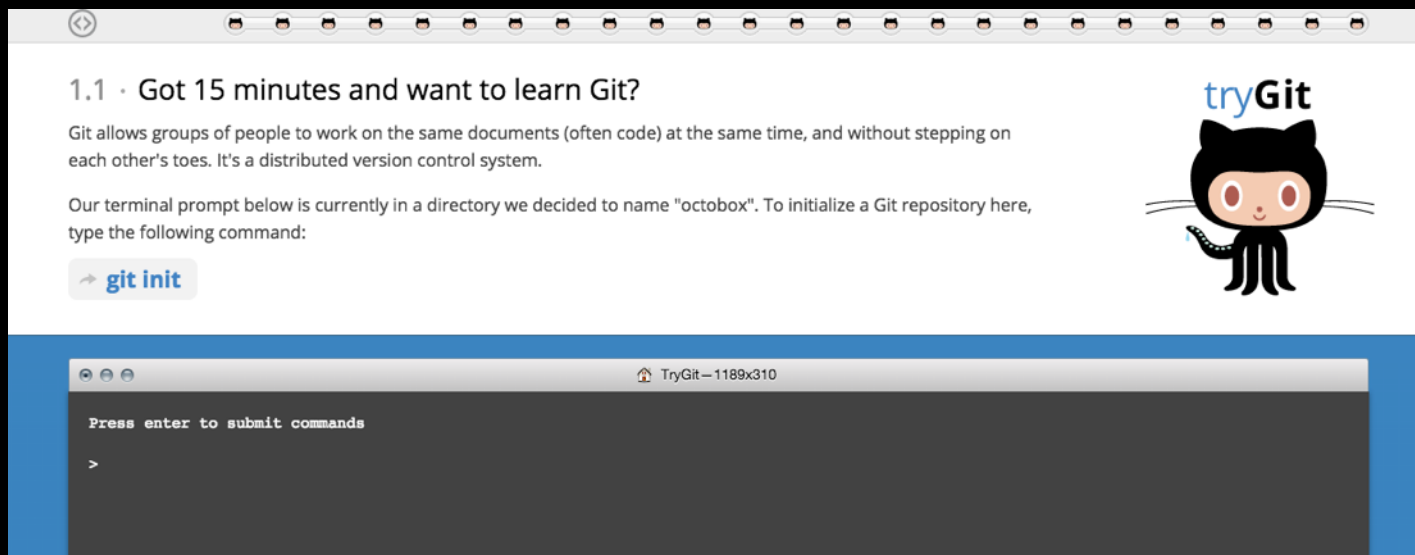


VISUALIZATION: BEING CREATIVE



LAB: INTRO TO GITHUB

[HTTP://TRY.GITHUB.COM/](http://try.github.com/)




The screenshot shows the tryGit website interface. At the top, there's a navigation bar with a back button and a row of small GitHub Octocat icons. The main content area has a heading "1.1 · Got 15 minutes and want to learn Git?". Below this, a paragraph explains that Git allows groups of people to work on the same documents (often code) at the same time, and without stepping on each other's toes. It's a distributed version control system. Another paragraph states that the terminal prompt below is currently in a directory named "octobox". To initialize a Git repository here, type the following command:

`git init`

On the right side, there's a logo for "tryGit" with the GitHub Octocat character. Below the main text, there's a terminal window titled "TryGit — 1189x310". The terminal displays the prompt "Press enter to submit commands" and a cursor ">" on the next line.

DOWNLOAD ANACONDA



CONTINUUM
ANALYTICS




[g+](#) [t](#) [in](#) [f](#) [View Your Cart](#)

HOME PRODUCTS ▾ CONSULTING ▾ TRAINING ▾ COMPANY ▾ CONTACT US

Download Anaconda

Anaconda is a completely free Python distribution (including for commercial use and redistribution). It includes over 195 of the most popular [Python packages](#) for science, math, engineering, data analysis.

CHOOSE YOUR INSTALLER:

 [I WANT PYTHON 3.4*](#)


Mac OS X – 64-Bit
Python 2.7
[Graphical Installer](#)

Size: 279M
(OS X 10.7 or higher)

INSTALLATION

After downloading the installer, double click the .pkg file and follow the instructions on the screen.

ENTERPRISE SOLUTIONS



ANACONDA SERVER

Internal Package
Management and
Deployment Made Easy

[Learn More](#)

INTRO TO DATA SCIENCE

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!