GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 1: DATA SCIENCE AND MACHINE LEARNING OVERVIEW

WELCOME!

0. INTRODUCTIONS
1. WHAT IS DATA SCIENCE?
2. THE DATA MINING WORKFLOW
3. WHAT IS MACHINE LEARNING?
4. MACHINE LEARNING PROBLEMS

LAB:

5. GITHUB INTRO

- Describe the data mining workflow and the key traits of a successful data scientist.
- Understand the meaning of machine learning and its uses in the context of data science
- Set up github account.

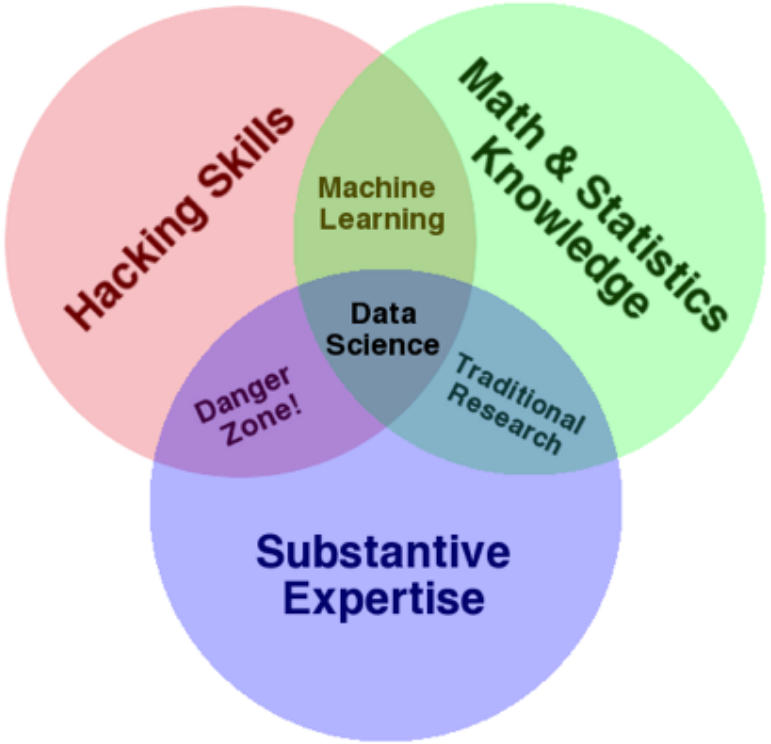
Introductions

- 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?

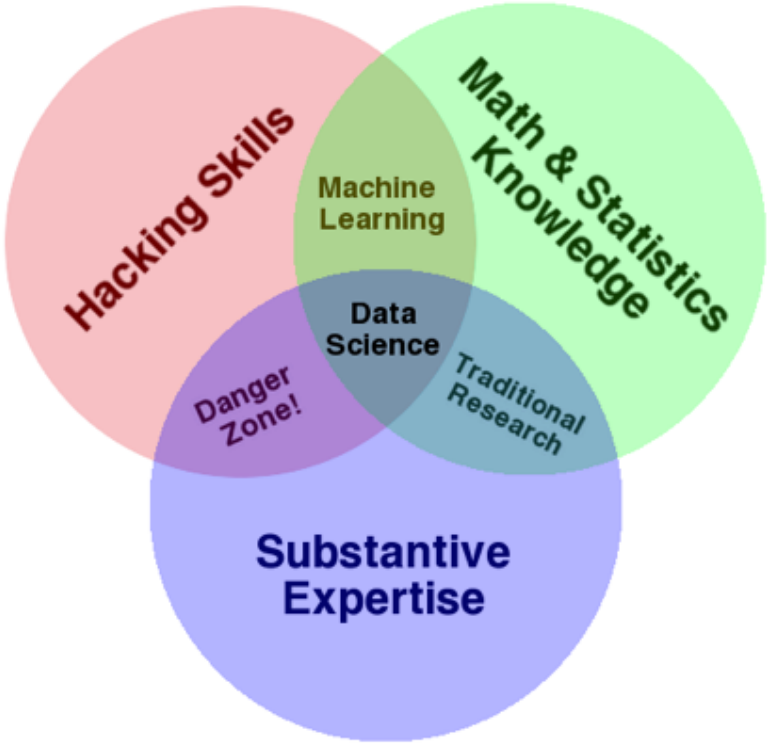
- A set of tools and techniques used to extract useful information from data.

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



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

THE QUALITIES OF A DATA SCIENTIST

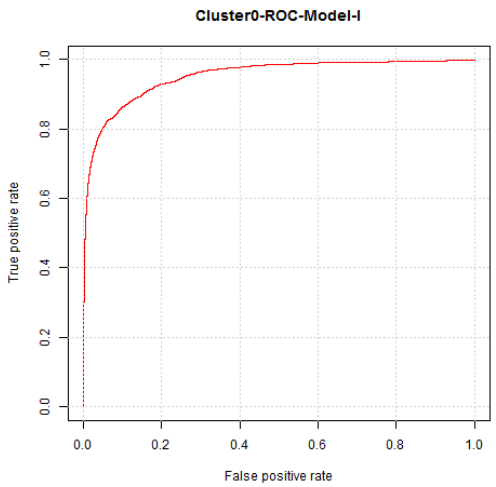


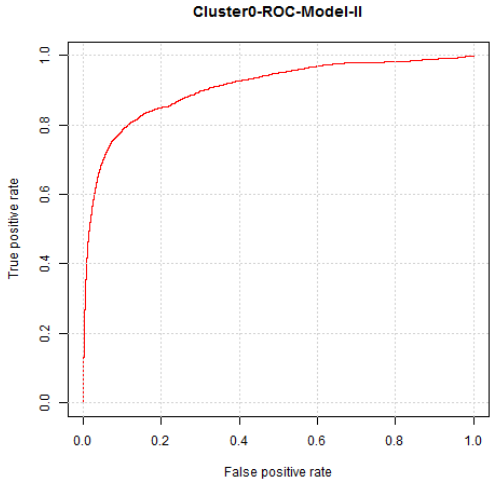
ONE MORE THING!

Communication skills

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

- The importance of communication:

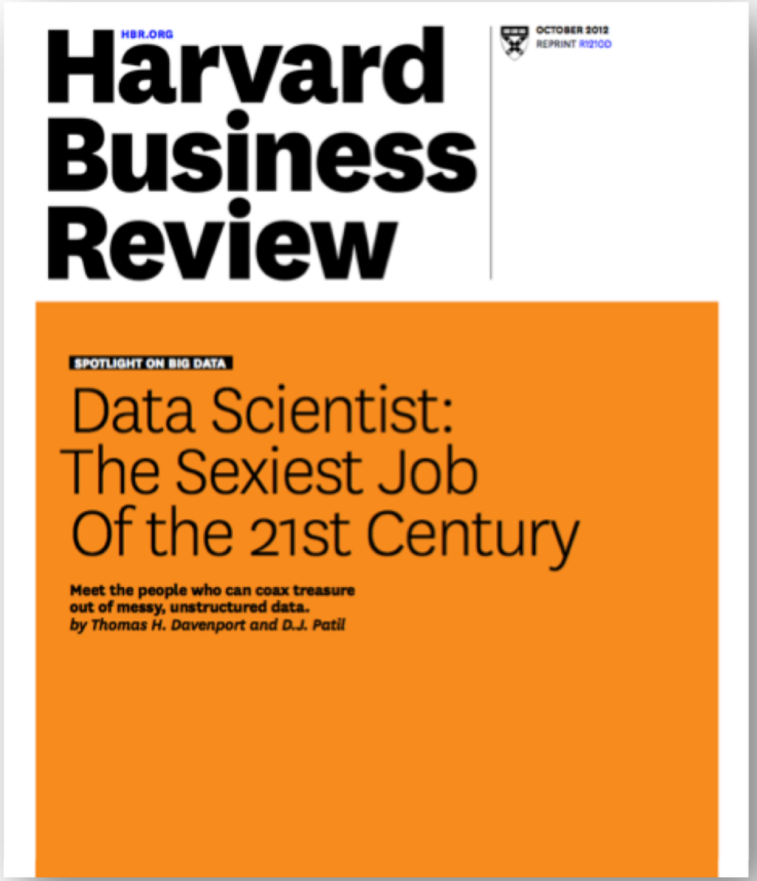




Ok, $$?

- 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 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?

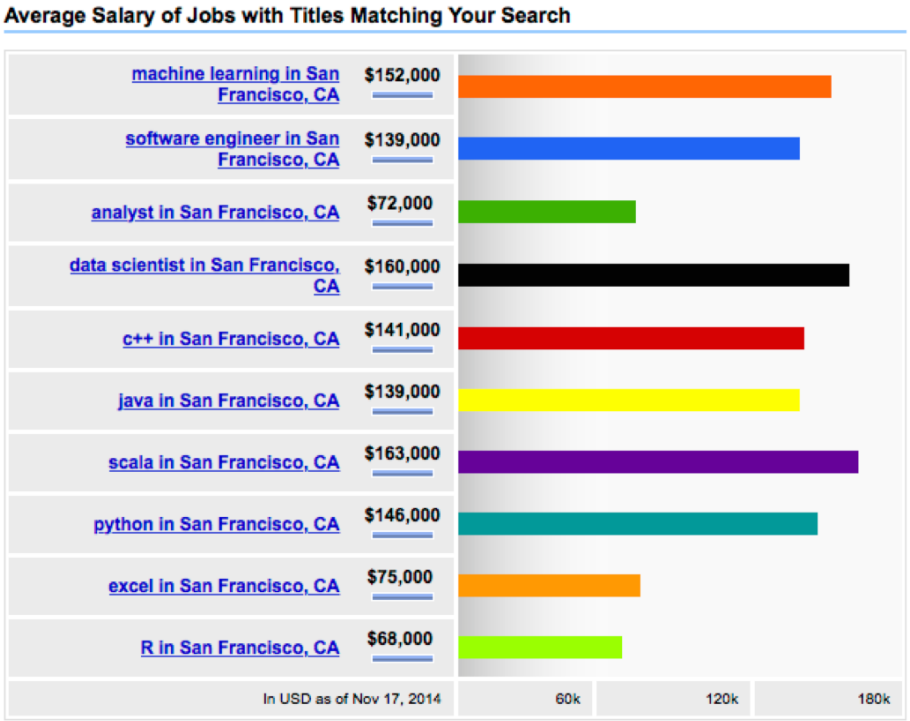
SAP Guest , SAP

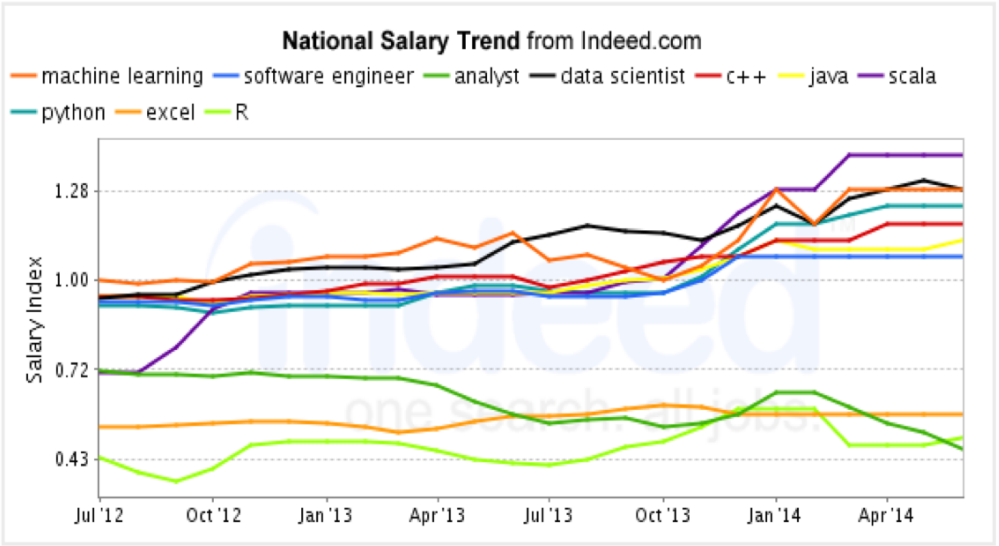
DATA

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

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 Learr 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 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

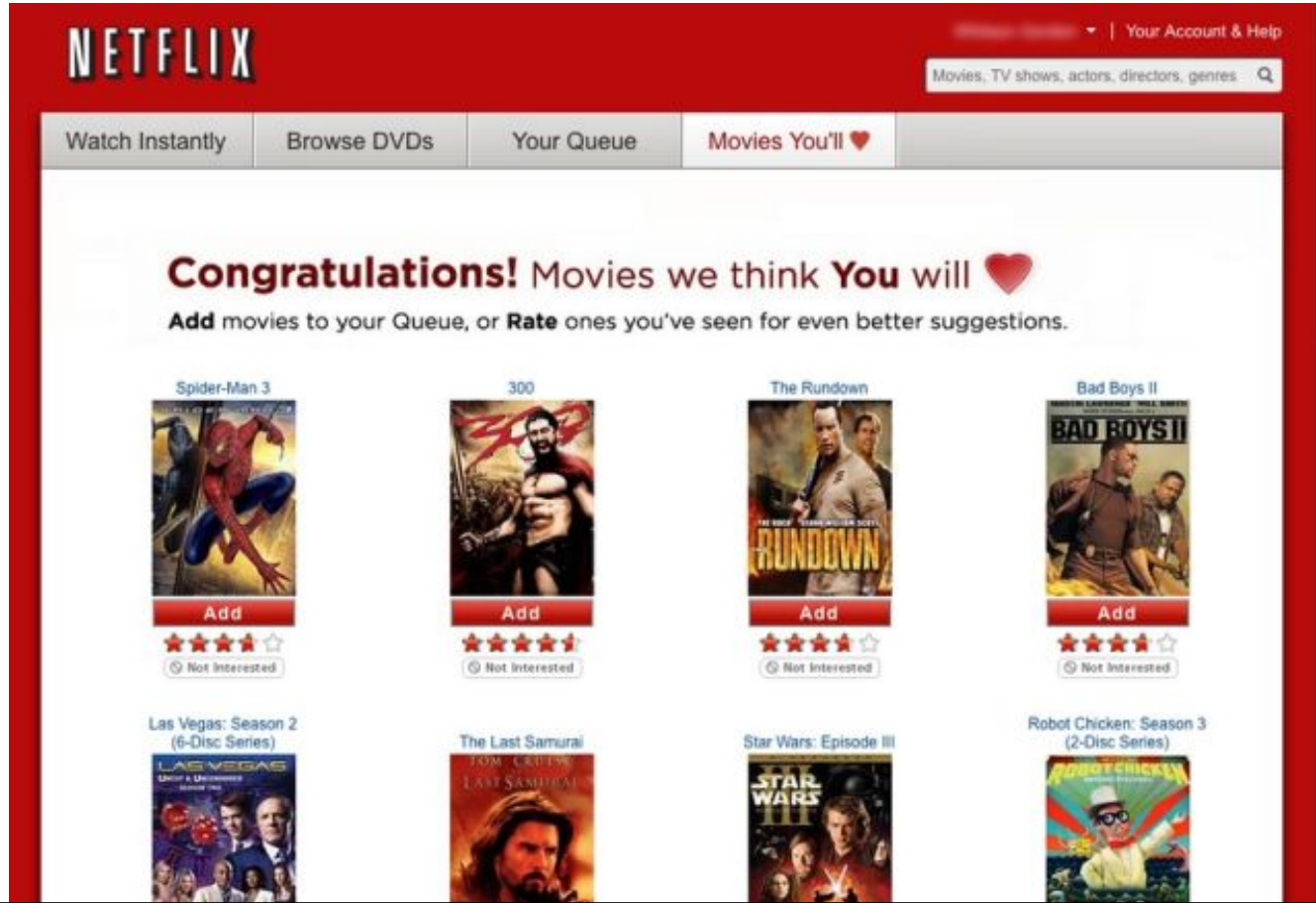
Data Scientist Intern
Move, Inc
San Jose, CA, US • Apr 24, 2015 • From chk.tbe.taleo.net
62 people in your network • Similar

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

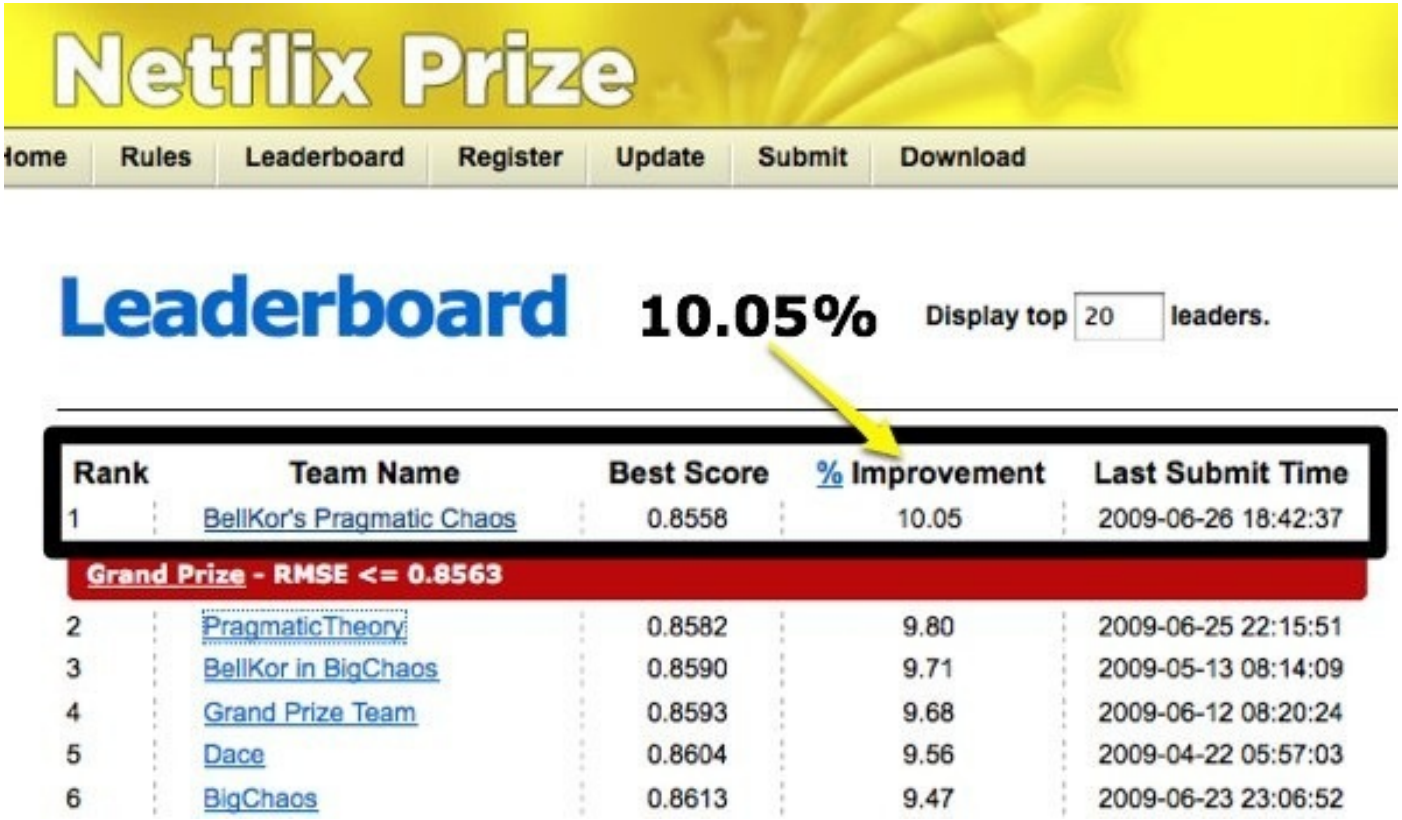
Data Scientist (Risk and Analysis)
Better Finance, Inc.
San Francisco, CA • Apr 21, 2015
13 people in your network • Similar

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

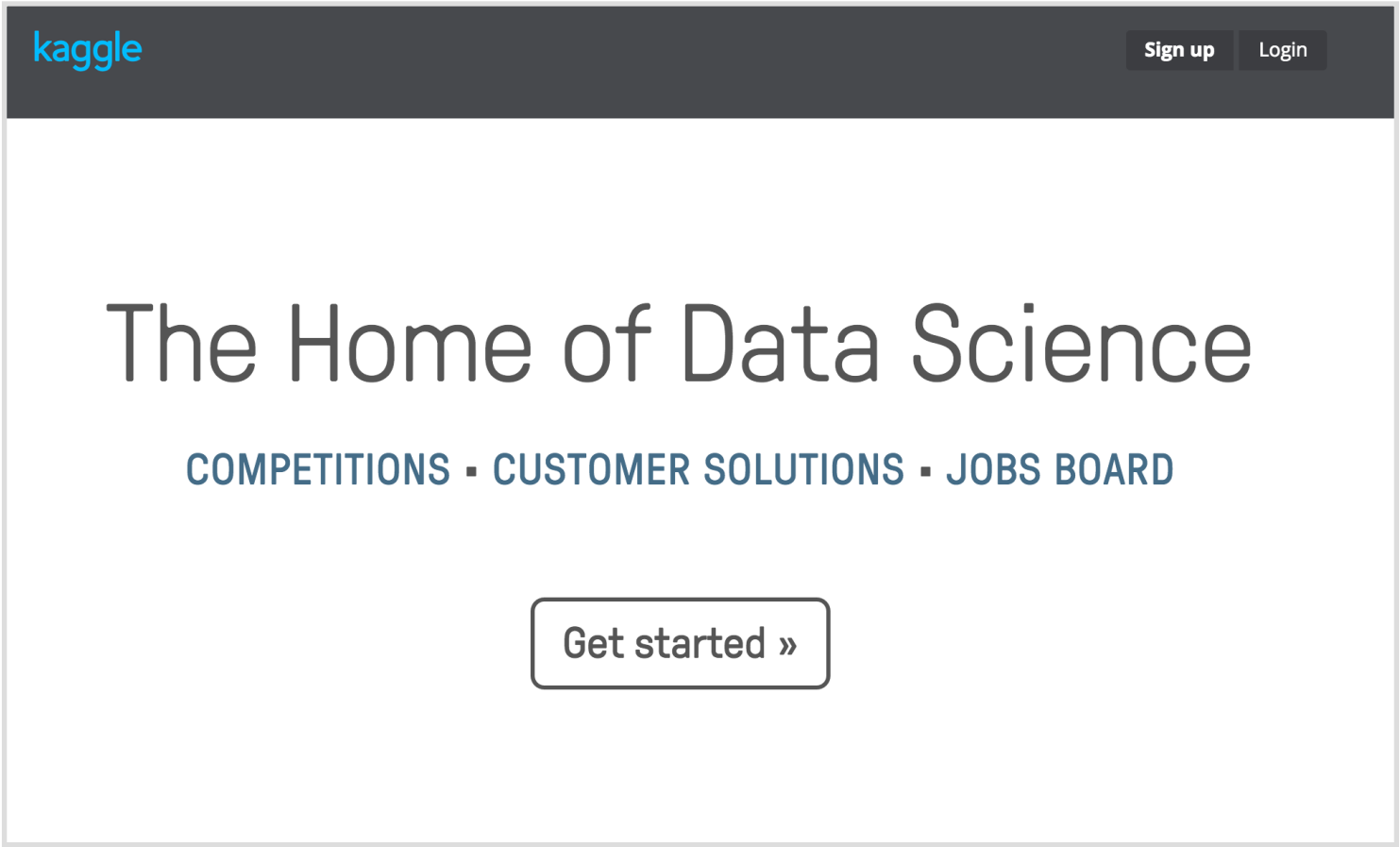
Data Scientist
Capital One
San Francisco - California - USA • Apr 27, 2015
623 people in your network • Similar


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
Add
Not Interested
300
Add
Not Interested
The Rundown
Add
Not Interested
Bad Boys II
Add
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 20 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 <= 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


kaggle
Sign up
Login
The Home of Data Science
COMPETITIONS ▪ CUSTOMER SOLUTIONS ▪ JOBS BOARD
Get started »

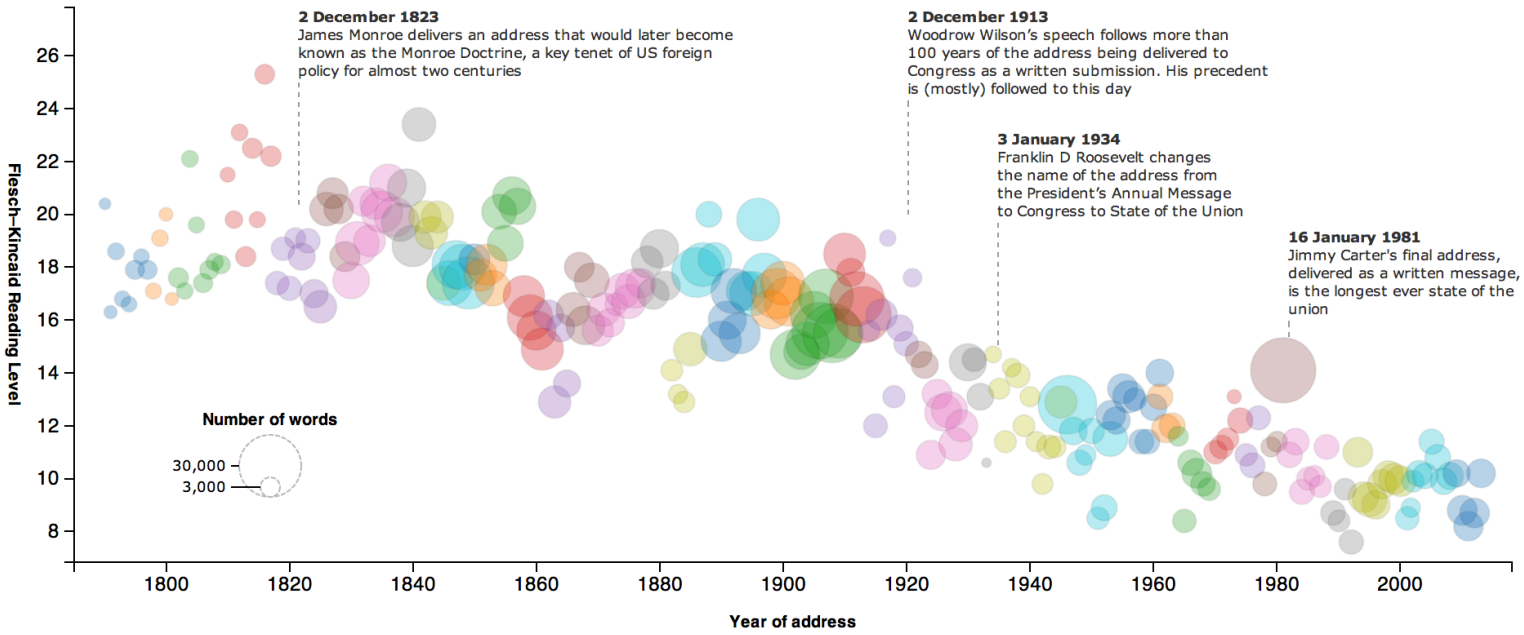

NETFLIX
amazon
facebook
Google™
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft

- 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

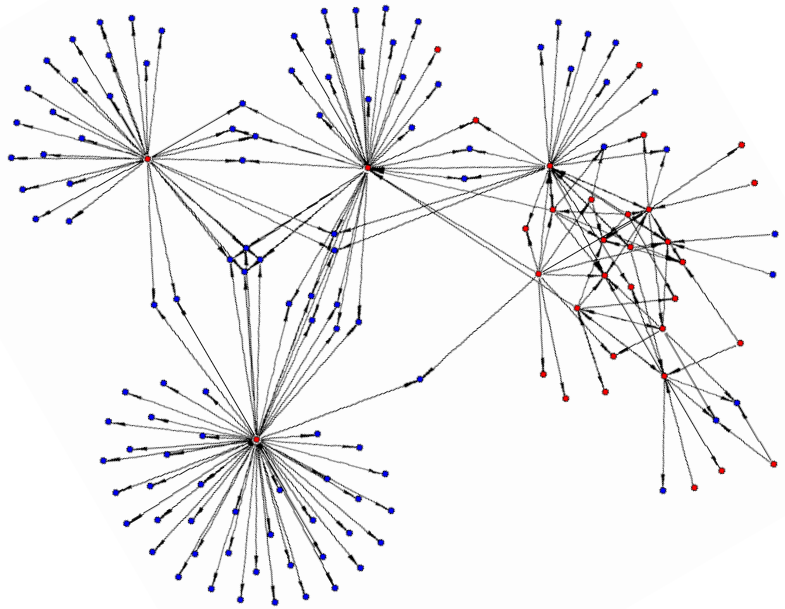
The state of our union is ... dumber:

How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every state of the union



Network Graphs – Modeling Customer Interactions:



Government Intelligence:

- Say we recover a hard drive…
- What's on it?
- What language is the material in?
- What important entities might be referenced?

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

- 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

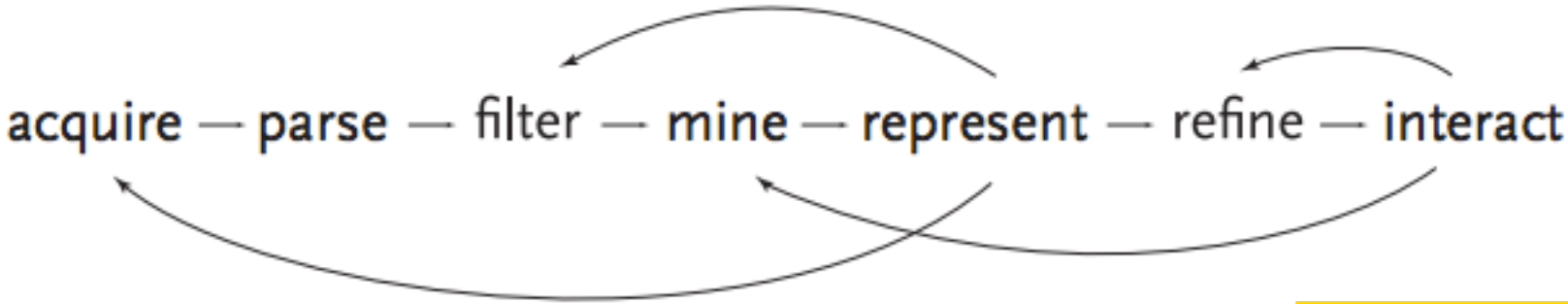
source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

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

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



NOTE

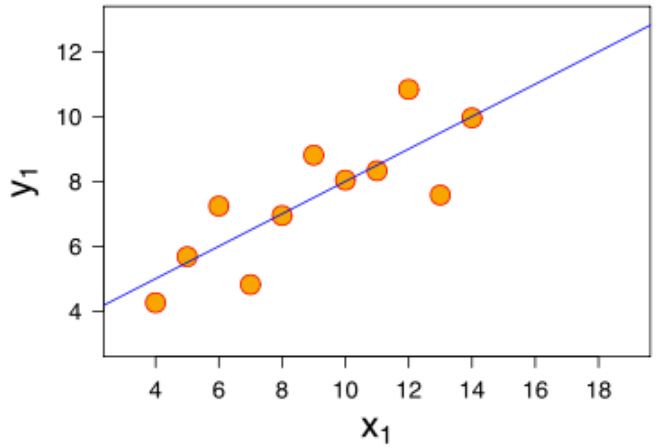
This diagram illustrates the iterative nature of problem solving

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

VISUALIZATIONS AS A MEDIUM

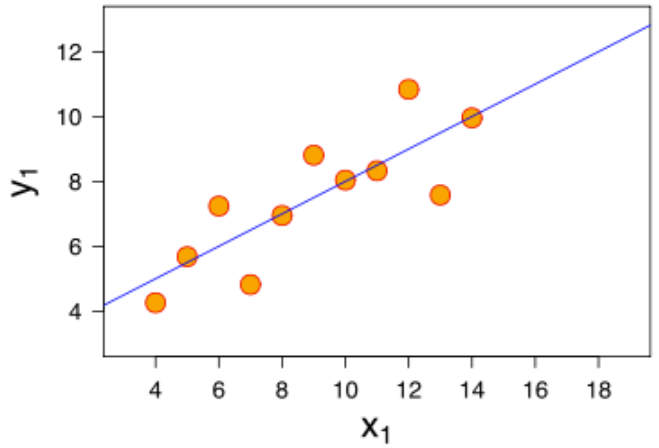
Consider the following dataset:

- *eleven (x, y) points*



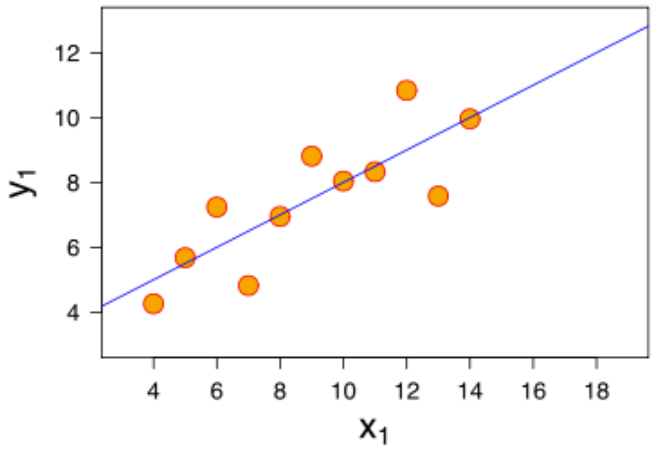
Consider the following dataset:

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



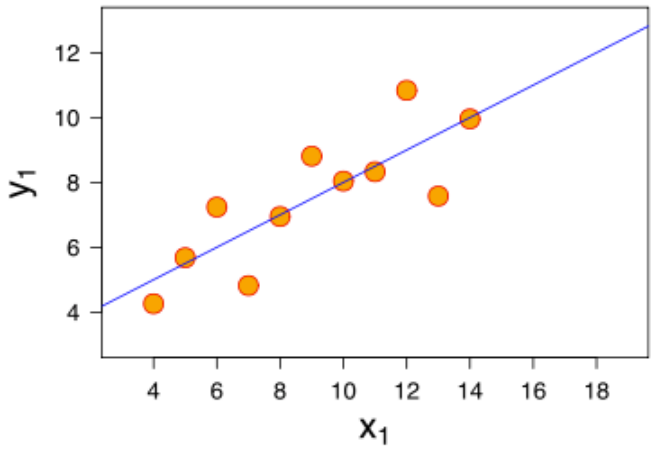
Consider the following dataset:

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



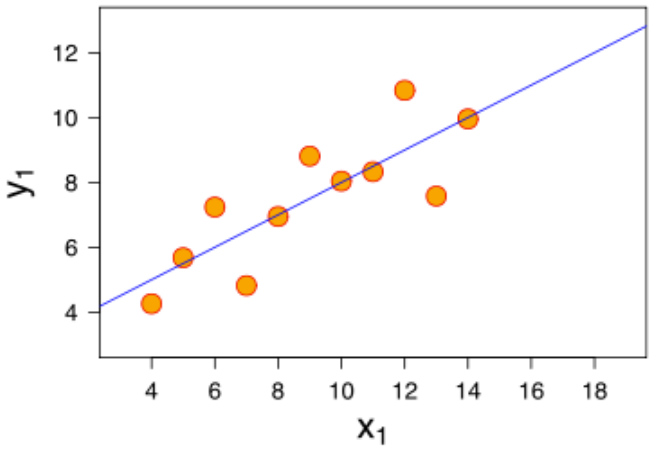
Consider the following dataset:

- *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*



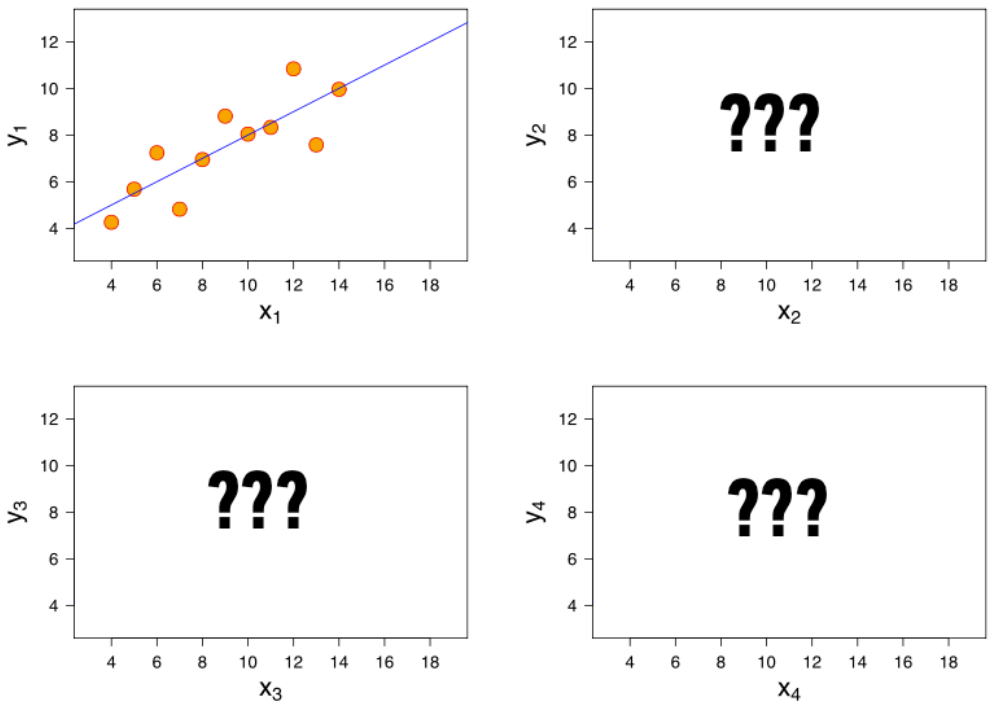
Consider the following dataset:

- *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*



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

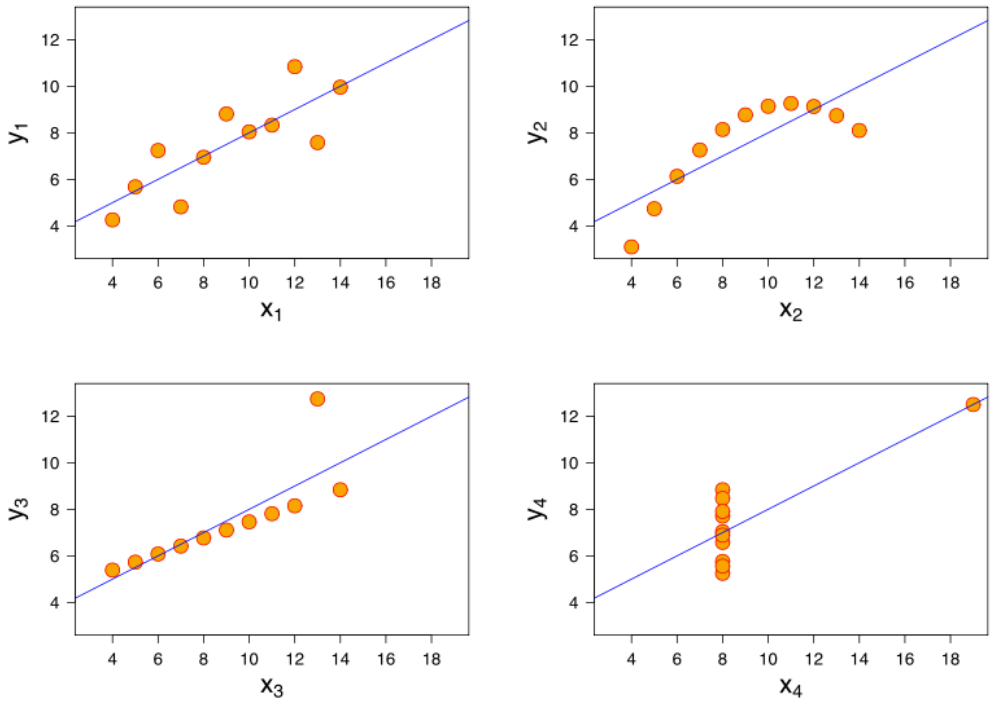
Q: how similar are these datasets?



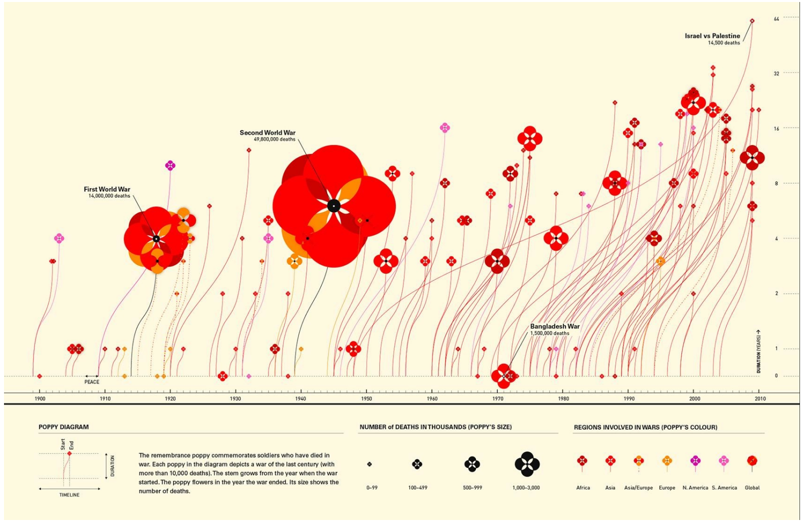
Now, suppose I give you three more datasets with exactly the same characteristics.

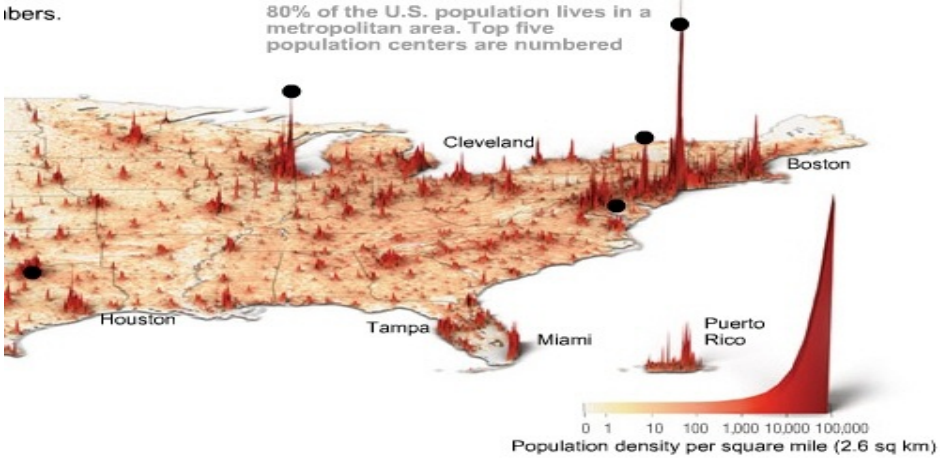
Q: how similar are these datasets?

A: not very!

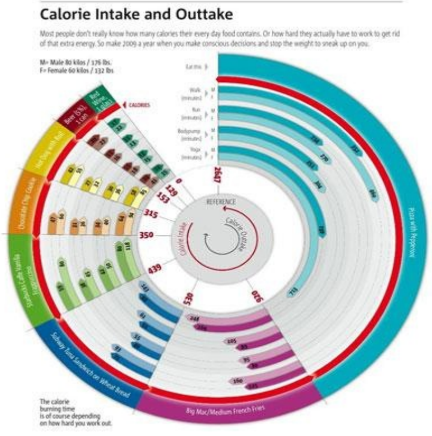


http://en.wikipedia.org/wiki/Anscombe's_quartet


Israel vs Palestine
14,500 deaths
Second World War
49,800,000 deaths
First World War
14,000,000 deaths
Bangladesh War
1,500,000 deaths
PEACE
DURATION (YEARS)
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
POPPY DIAGRAM
Start
End
DURATION
TIMELINE
The remembrance poppy commemorates soldiers who have died in war. Each poppy in the diagram depicts a war of the last century (with more than 10,000 deaths). The stem grows from the year when the war started. The poppy flowers in the year the war ended. Its size shows the number of deaths.
NUMBER of DEATHS IN THOUSANDS (POPPY'S SIZE)
0–99
100–499
500–999
1,000–3,000
REGIONS INVOLVED IN WARS (POPPY'S COLOUR)
Africa
Asia
Asia/Europe
Europe
N. America
S. America
Global


80% of the U.S. population lives in a metropolitan area. Top five population centers are numbered
Cleveland
Boston
Houston
Tampa
Miami
Puerto Rico
0 1 10 100 1,000 10,000 100,000
Population density per square mile (2.6 sq km)


Visualisation
DATA
Figure
Illustrate
Information
Knowledge
Model
Demonstration
Measurement
Compilation
Storyboard
Numbers
Evidence
Facts
Phenomenon
Circumstance
Indicate
Feature
Blueprint
Proof
Render
Gospel
Results
Truism
Substantiation
Testimony
Story
Figures
Confirmation
Expansion
Description
Echo
Manifestation
Reflection
Info
Canvas
Display
Index
Drawing
Spectacle
Entity
Statistic
Impression
Occurrence
Sign
Chart
Study
Form
Icon
Item
Tableau
Case
Portrait
Recording
Report
Representation
Clue
Conception
Striking
Derivative
Principle
Construction
Show
Represent
Art
Factor
Transaction
Comic
Grounds
Point
Caricature
Transcript
Declaration
Proposition
Inference


Calorie Intake and Outtake
Most people don't really know how many calories their every day food contains. Or how hard they actually have to work to get rid of that extra energy. So make 2009 a year when you make conscious decisions and stop the weight to sneak up on you.
M= Male 80 kilos / 176 lbs.
F= Female 60 kilos / 132 lbs.
REFERENCE
Calorie Intake
Calorie Outtake
Big Mac/Medium French Fries
The calorie burning time is of course depending on how hard you work out.

III. WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

- representation – extracting structure from data

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

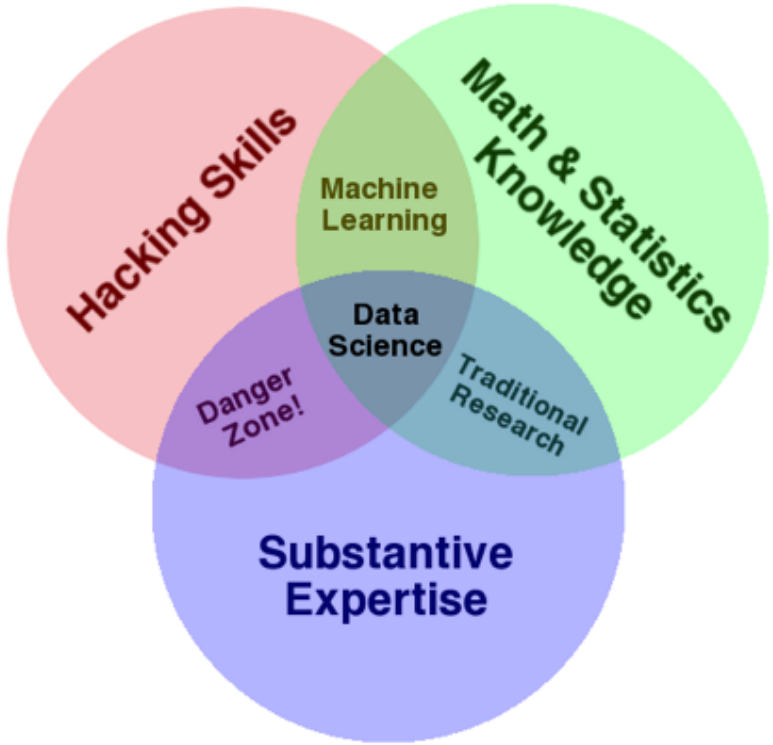
"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data."

"The core of machine learning deals with representation and generalization…"

- representation – extracting structure from data
- generalization – making predictions from data

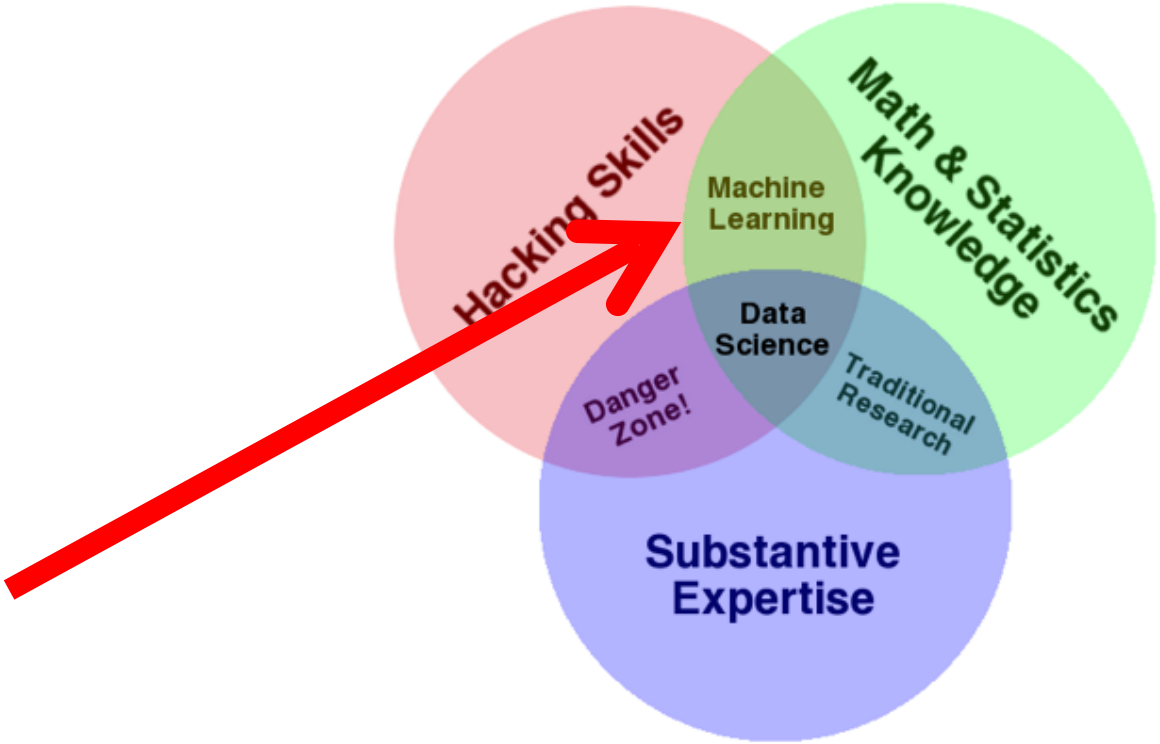
source: http://en.wikipedia.org/wiki/Machine_learning

REMEMBER THIS?



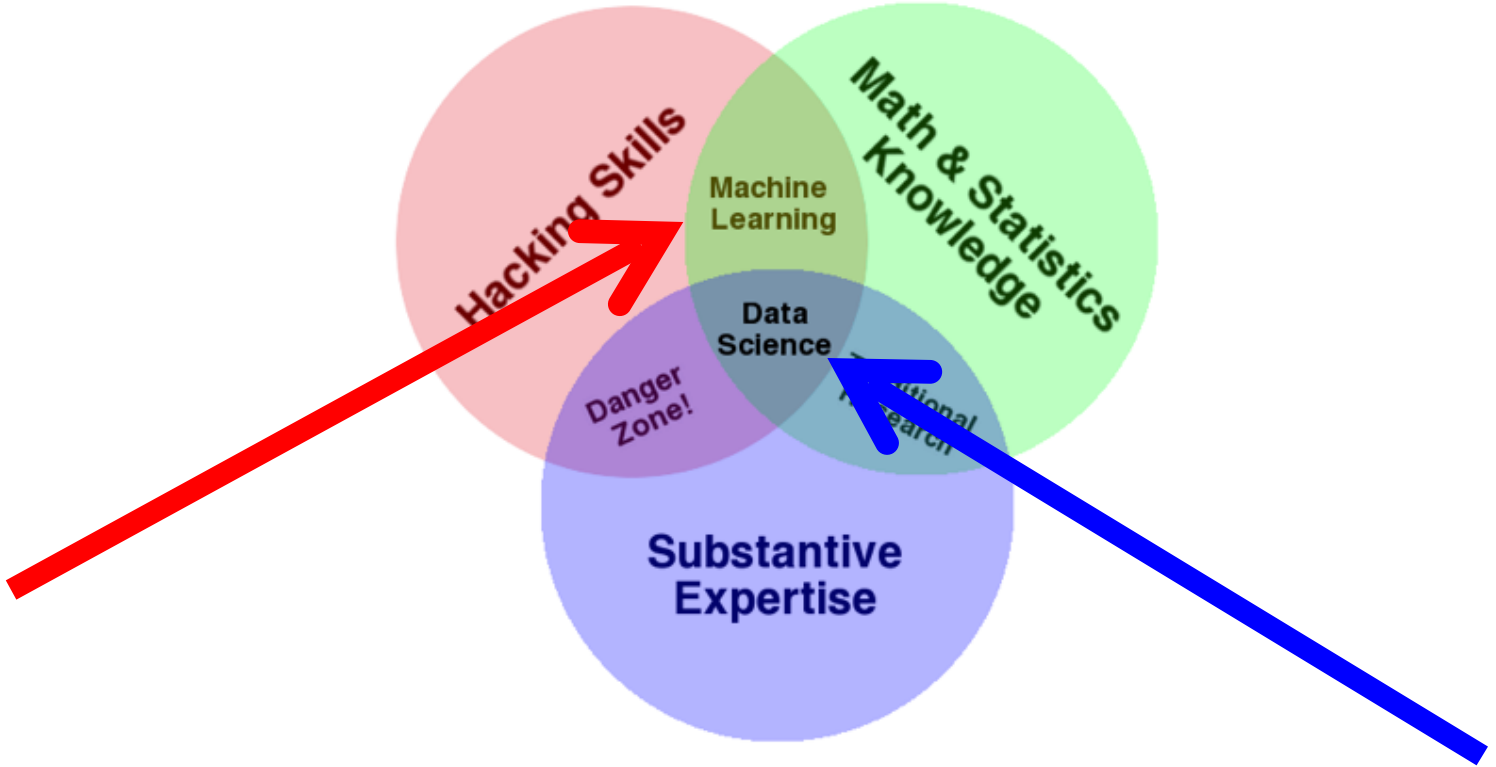
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU ARE HERE



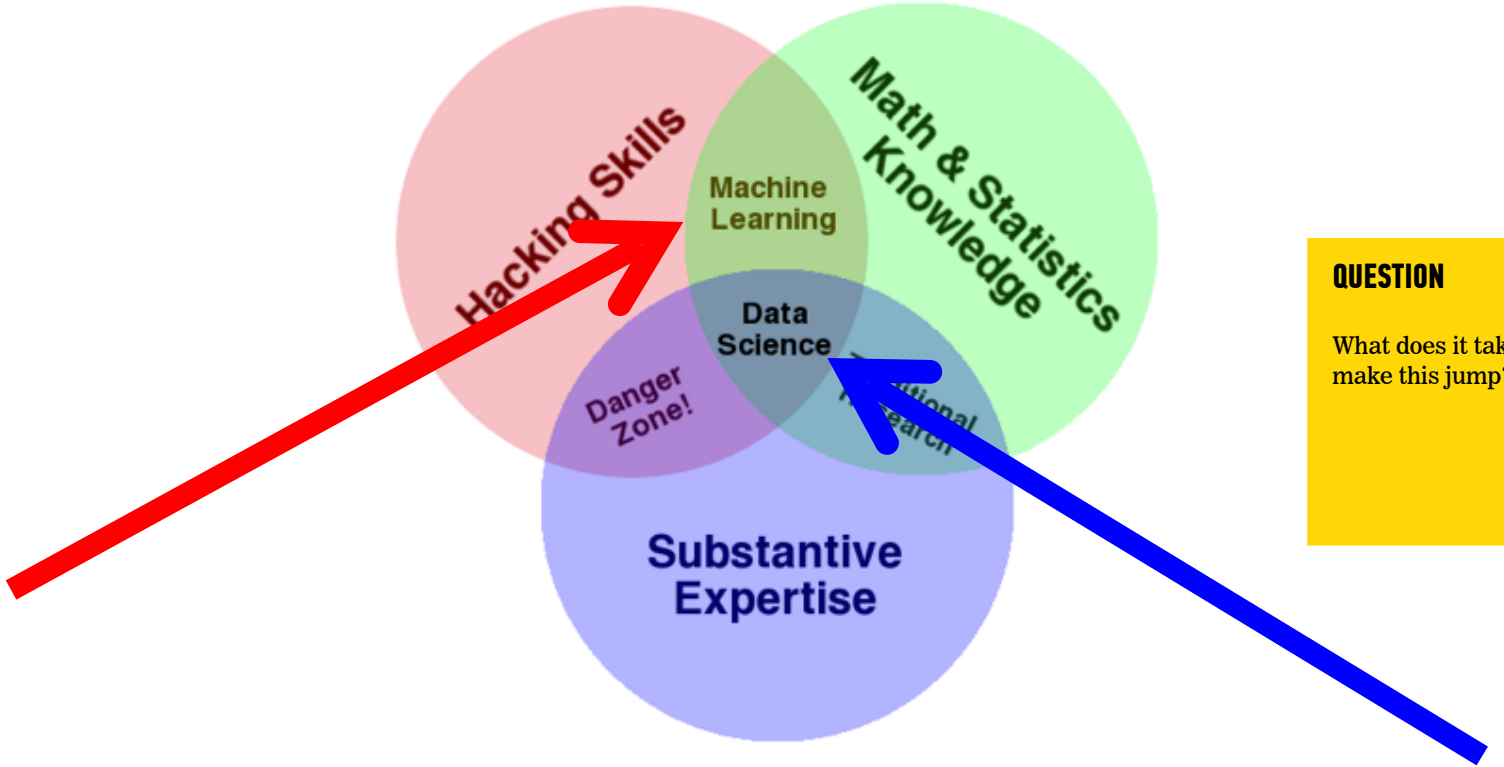
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE



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

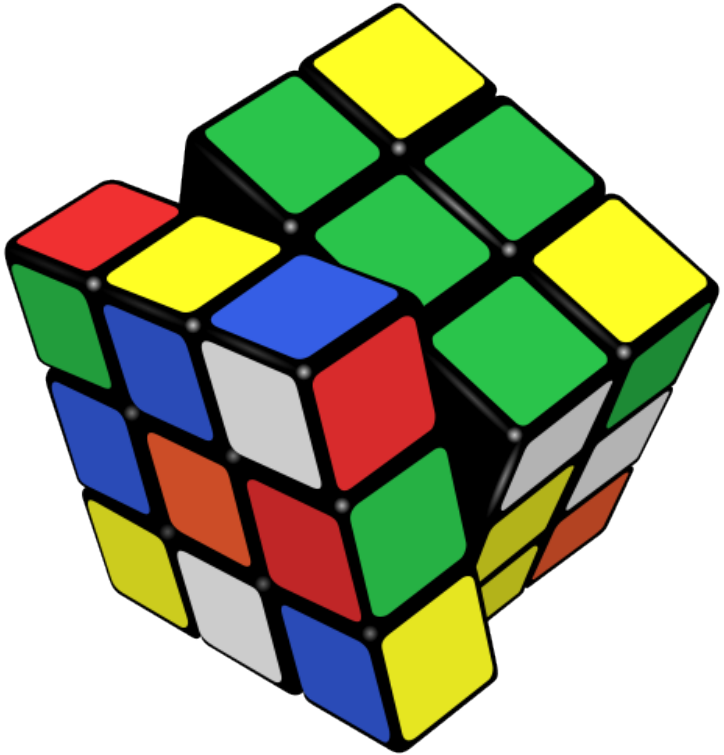
YOU WANT TO GO HERE



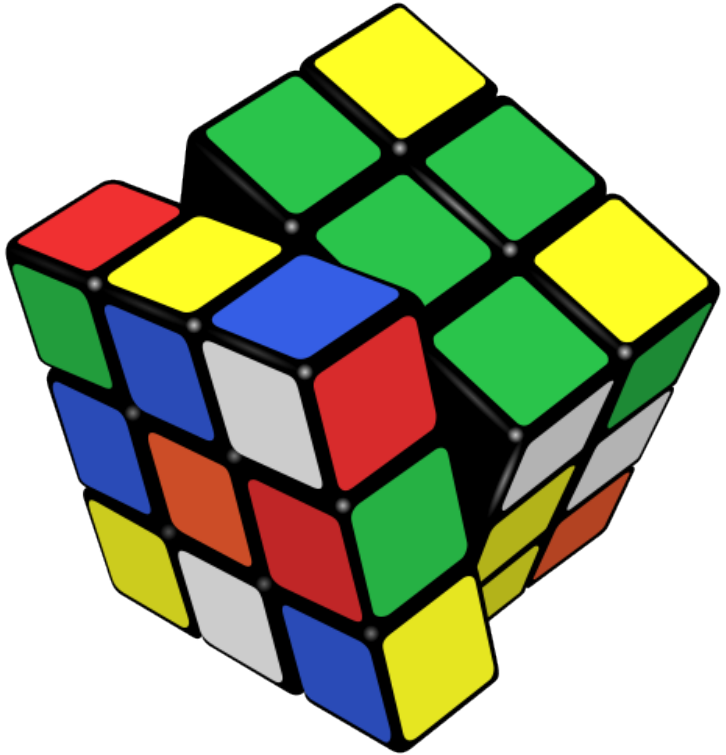
QUESTION

What does it take to make this jump?

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



ANSWER: PROBLEM SOLVING!



NOTE

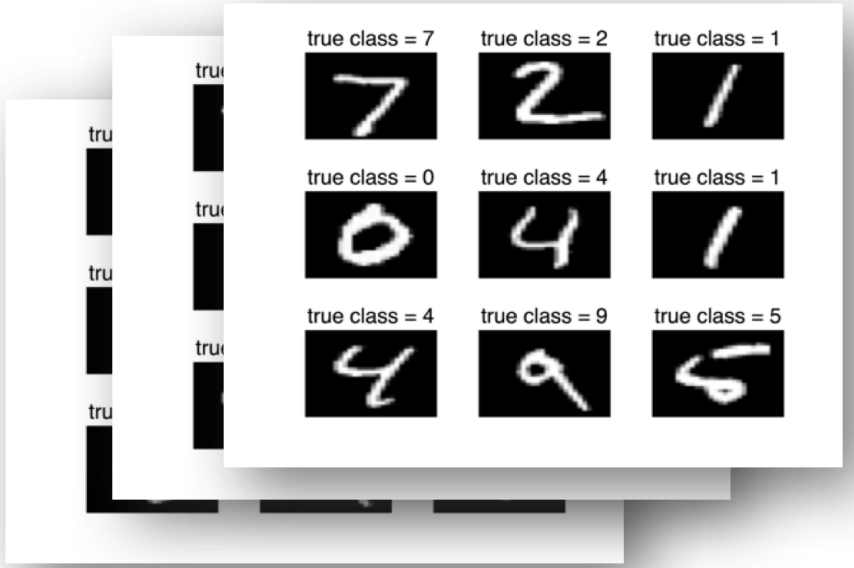
Implementing solutions to ML problems is the focus of this course!

IV. MACHINE LEARNING PROBLEMS

Learning is not about memorizing and being able to recall, it is about generalizing the conclusions to previously unseen examples

Supervised learning: the goal is to learn mapping from given inputs x to outputs y, given a **labeled** set of input-output pairs


OCR

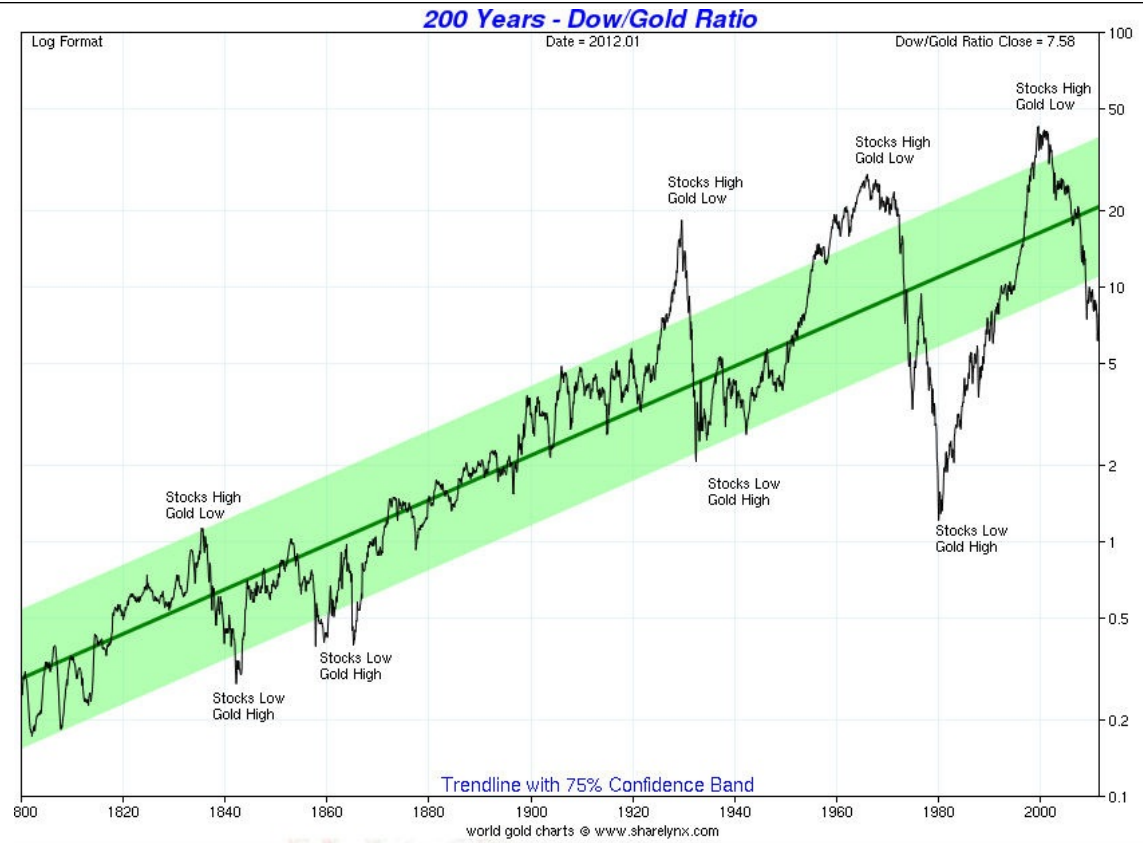

true class = 7
true class = 2
true class = 1
true class = 0
true class = 4
true class = 1
true class = 4
true class = 9
true class = 5

	Client 1	*Client 2*	*Client 3*
Age	*23*	*30*	*19*
Gender	*M*	*F*	*M*
Annual salary	*$30,000*	*$45,000*	*$15,000*
Years in residence	*3 years*	*1 year*	*3 month*
Years in job	*1 year*	*1 year*	*1 month*
Current debt	*$5,000*	*$1,000*	*$10,000*
Paid off credit	*Yes*	*Yes*	*No*

	Client 1	Client 2	Client 3
Age	23	30	19
Gender	M	F	M
Annual salary	$30,000	$45,000	$15,000
Years in residence	3 years	1 year	3 month
Years in job	1 year	1 year	1 month
Current debt	$5,000	$1,000	$10,000
Paid off credit	Yes	Yes	No

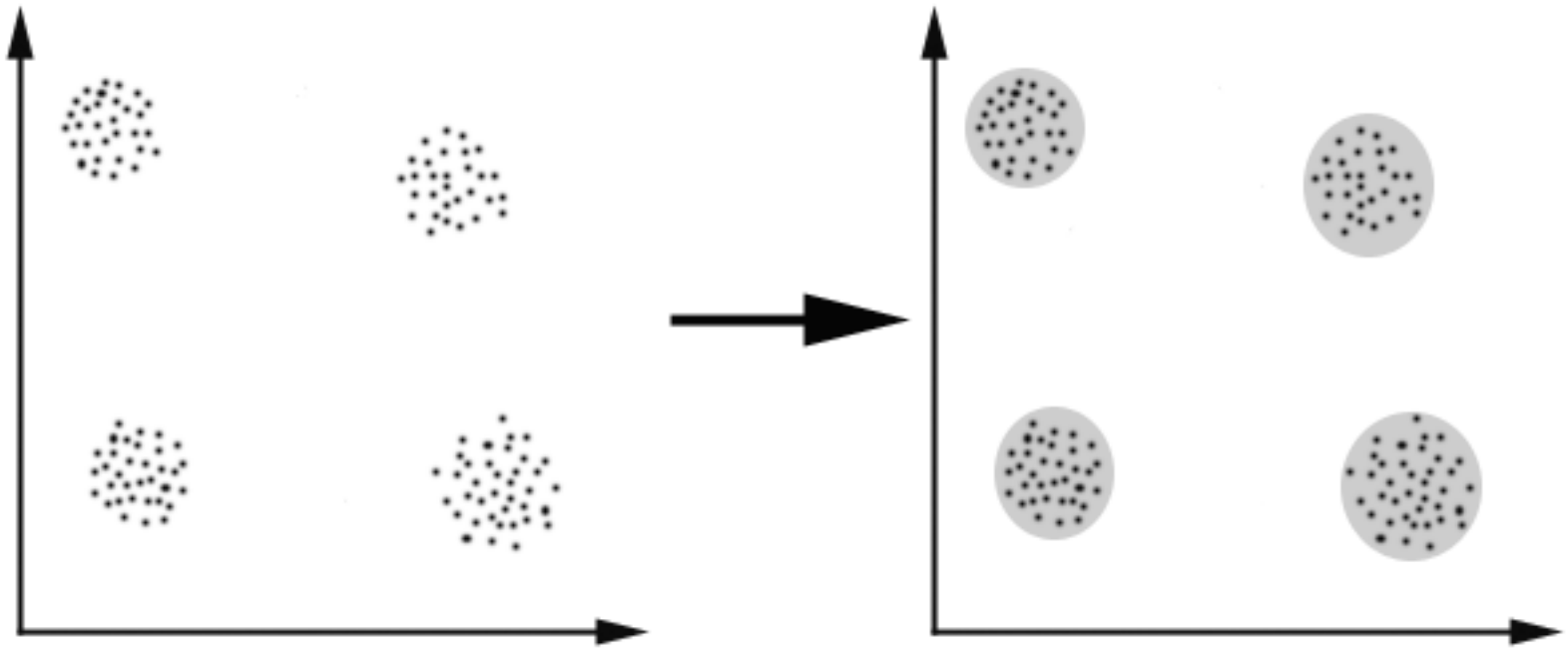
	Applicant
Age	25
Gender	M
Annual salary	$25,000
Years in residence	1 year
Years in job	2 year3
Current debt	$15,000
Credit decision/ score	???


SPAM

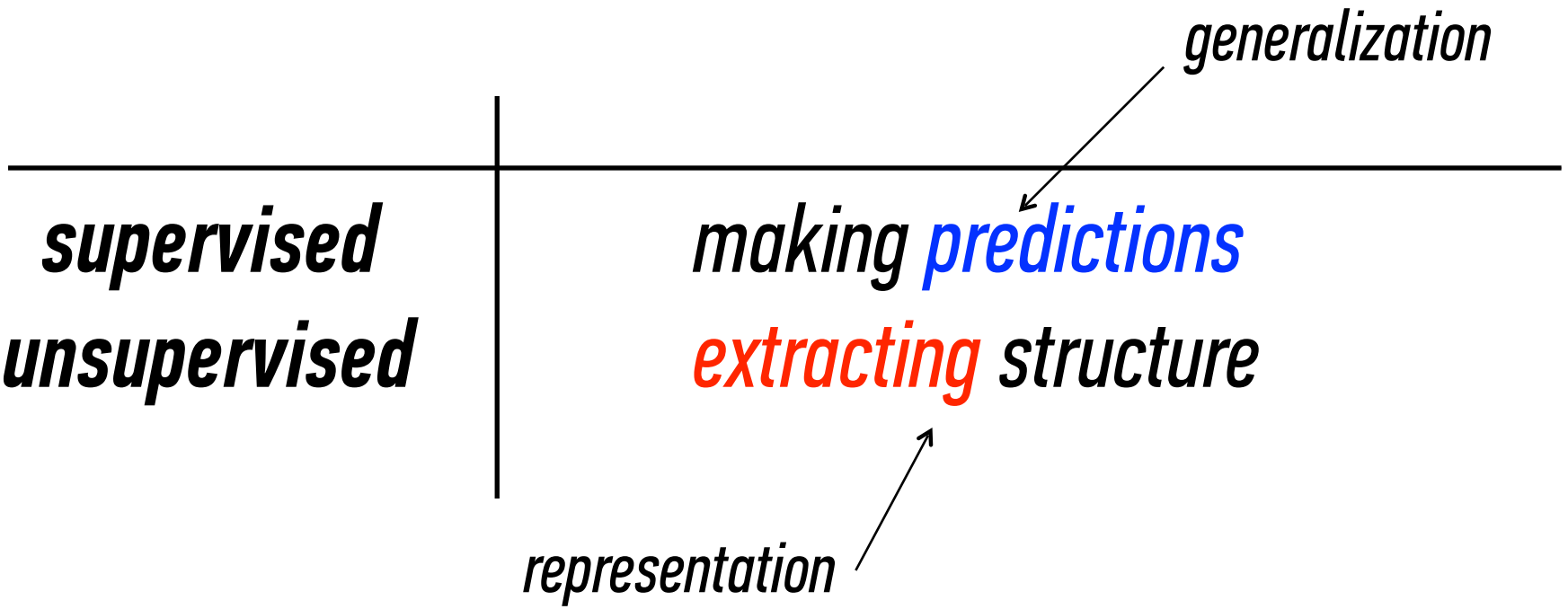

200 Years - Dow/Gold Ratio
Log Format
Date = 2012.01
Dow/Gold Ratio Close = 7.58
Stocks High
Gold Low
Stocks Low
Gold High
Stocks High
Gold Low
Stocks Low
Gold High
Stocks High
Gold Low
Stocks Low
Gold High
Stocks High
Gold Low
Stocks Low
Gold High
Trendline with 75% Confidence Band
100
50
20
10
5
2
1
0.5
0.2
0.1
800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
world gold charts © www.sharelynx.com

Unsupervised learning: the goal is to learn interesting patterns and structure in data given only inputs

no label information given at all

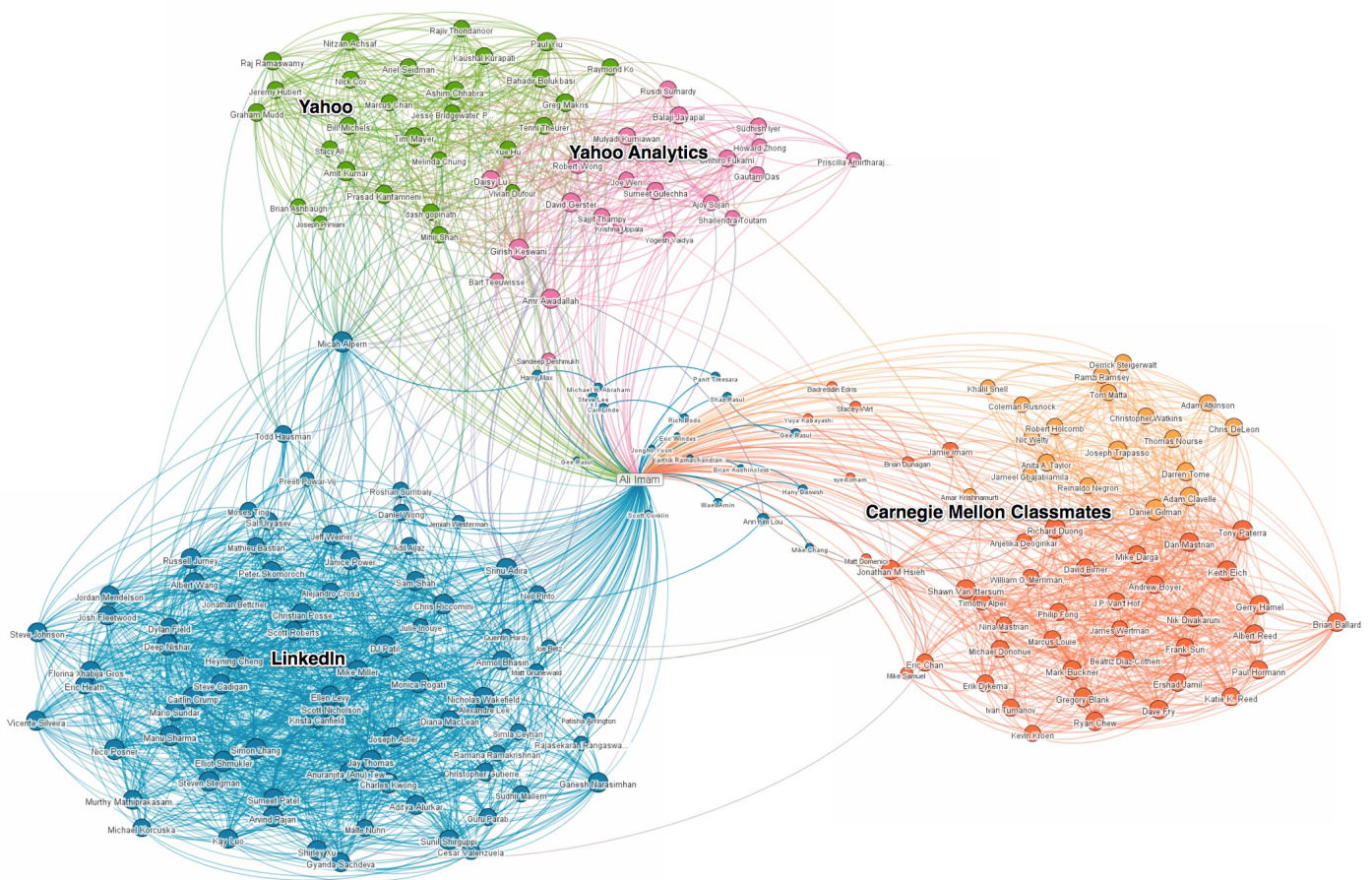


supervised	*making predictions*
unsupervised	*extracting structure*


generalization
supervised
unsupervised
making predictions
extracting structure
representation

EXERCISE:

supervised or unsupervised?


Yahoo
Yahoo Analytics
Carnegie Mellon Classmates
LinkedIn
Ali Imam
Micah Altern
Todd Hunsman
Rajiv Thondanoor
Nitzan Achsaf
Paul Yiu
Raj Ramaswamy
Kaushal Kurapati
Ariel Seidman
Raymond Ko
Nick Cox
Bahadir Bolukbasi
Jeremy Hubert
Ashim Chhabra
Rusdi Sumardy
Graham Mudd
Marcus Chan
Greg Makris
Jesse Bridgewater P
Balaji Jayapal
Bill Michels
Tenni Theurer
Sudhish Iyer
Tim Mayer
Mulyadi Kurniawan
Stacy Ali
Xiao Hu
Howard Zhong
Melinda Chung
Robert Wong
Priscilla Amritharaj...
Amit Kumar
Daisy Lu
Joe Wen
Gautam Das
Prasad Kantamneni
Vivian Dufour
Sumeet Gulechha
Brian Ashbaugh
David Gerster
Ajoy Solan
Joseph Dymiani
Sajit Thampy
Shailendra Toutam
Mihir Shah
Girish Keswani
Yogesh Vaidya
Bart Teeuwisse
Amr Awadallah
Sandeep Deshmukh
Harry Mao
Panit Tiesarat
Shad Pasol
Yuya Kobayashi
Geo Pasol
Eric Windes
Karthik Ramachandran
Brian Aushinoloss
Hany Darwish
Wasim Amin
Ann Kim Lou
Mike Chang
Scott Conklin
Derrick Stangerwalt
Ramzi Ramsey
Khalil Shell
Tom Matta
Coleman Rusnock
Adam Asinson
Christopher Watkins
Robert Holcomb
Chris DeLeon
Nic Welty
Thomas Nourse
Joseph Trapasso
Jamie Imam
Brian Dunagan
Anita A Taylor
Jameel Gharabiamin
Darren Tome
Reinaldo Negron
Amar Krishnamurti
Adam Clavelle
Daniel Gilman
Richard Duong
Tony Paterra
Anjelika Deborinkar
Dan Mastrian
Mike Darga
Matt Donnelson
Jonathan M Hsieh
David Bilner
Keith Eich
William O Merriman
Andrew Boyer
Shawn Van Ittersum
Timothy Aiper
J.P. Vant Hof
Gerry Hamel
Philip Fong
Nik Divakaruni
Brian Ballard
Nina Mastrian
James Wertman
Albert Reed
Marcus Louie
Michael Donohue
Frank Sun
Eric Chan
Beatriz Diaz-Cohen
Mike Samuel
Mark Buckner
Paul Homann
Erik Dykema
Ershad Jamil
Gregory Blank
Katie K Reed
Ivan Tumanov
Dave Fry
Ryan Chew
Kevin Kloen
Preeti Pawar-Vij
Roshan Sumbaly
Moses Ting
Daniel Wong
Sarl Oryasev
Jemiah Westerman
Mathieu Bastian
Jeff Weiner
Adi Muraz
Russell Jurney
Janice Powell
Peter Skomoroch
Srinu Adira
Albert Wang
Sam Shah
Jordan Mendelson
Alejandro Crosa
Nail Pinto
Jonathan Bettcher
Chris Riccomini
Josh Fleetwood
Christian Posse
Dylan Field
Julie Doojay
Steve Johnson
Scott Roberts
Queenie Hardy
Deep Nishar
DJ Patil
Joe Benz
Florina Jhabvala Gros
Heyning Cheng
Mike Miller
Anmol Bhasin
Eric Heath
Steve Cadigan
Monica Rogati
Matt Gourewitz
Caitlin Crump
Ellen Levy
Nicholas Wakefield
Scott Nicholson
Mario Sundar
Krista Canfield
Alexander Lee
Vicente Silveira
Diana MacLean
Patisha Arrington
Manu Sharma
Simis Cayhan
Joseph Adler
Rajasekaran Rangaswa...
Nico Rosner
Simon Zhang
Ramana Ramkrishnan
Elliot Shmukler
Jay Thomas
Anuranjita (Anu) Tew
Christopher Gutierre
Steven Stegman
Charles Kwong
Ganesh Narasimhan
Murthy Mathiprakasam
Sumeet Patel
Sudhir Malleni
Aditya Alurkar
Arvind Rajan
Guru Parab
Michael Korcuska
Kay Luo
Mark Nunn
Sunil Shrigopur
Shirley Yu
Cesar Valenzuela
Gyanda Sachdeva


CHEQUE
APPLICATION
ORDER
CLAIM

	continuous	*categorical*
	quantitative	*qualitative*

continuous	*categorical*
Height of children	*Eye colors*
Weight of cars	*Courses at GA*
Speed of the train	*Highest degree*
Temperature	*Gender*
Stock price	*If an email is spam or not*

	continuous	*categorical*
	quantitative	*qualitative*

NOTE

The space where data live is called the feature space.

Each point in this space is called a record.

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

NOTE

We will implement solutions using models and algorithms.

Each will fall into one of these four buckets.

QUESTION

WHAT IS THE GOAL OF MACHINE LEARNING?

supervised	*making predictions*
unsupervised	*extracting structure*

ANSWER

The goal is determined by the type of problem.

QUESTION

HOW DO YOU DETERMINE THE RIGHT APPROACH?

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

ANSWER

The right approach is determined by the desired solution.

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

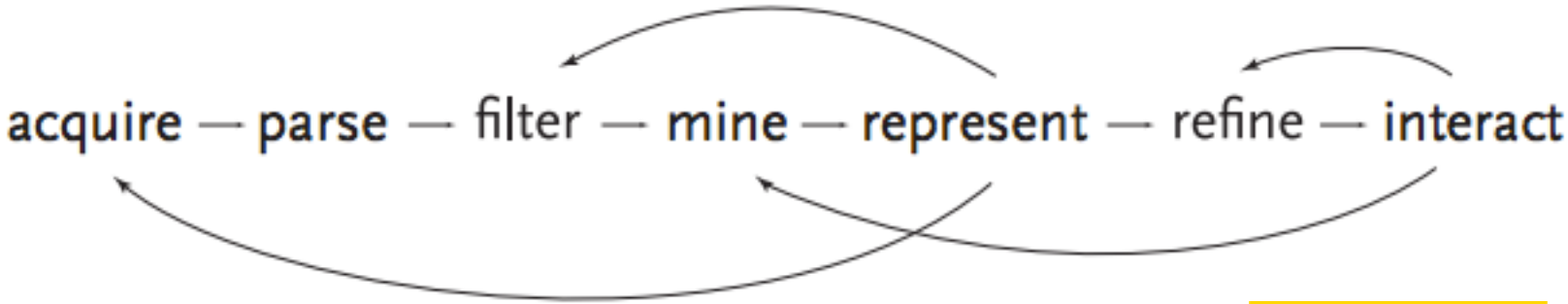
NOTE

All of this depends on your data!


CHEQUE
APPLICATION
ORDER
CLAIM

QUESTION

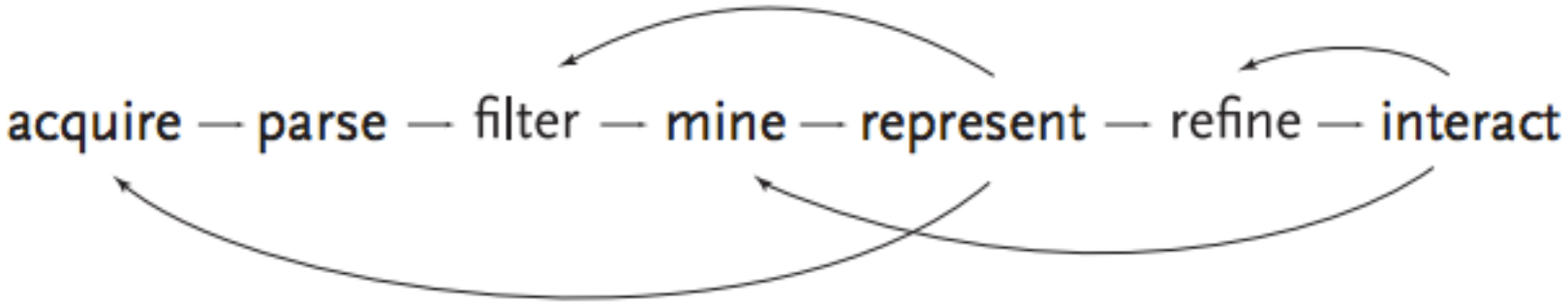
WHAT DO YOU DO WITH YOUR RESULTS?



ANSWER

Interpret them and react accordingly.

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



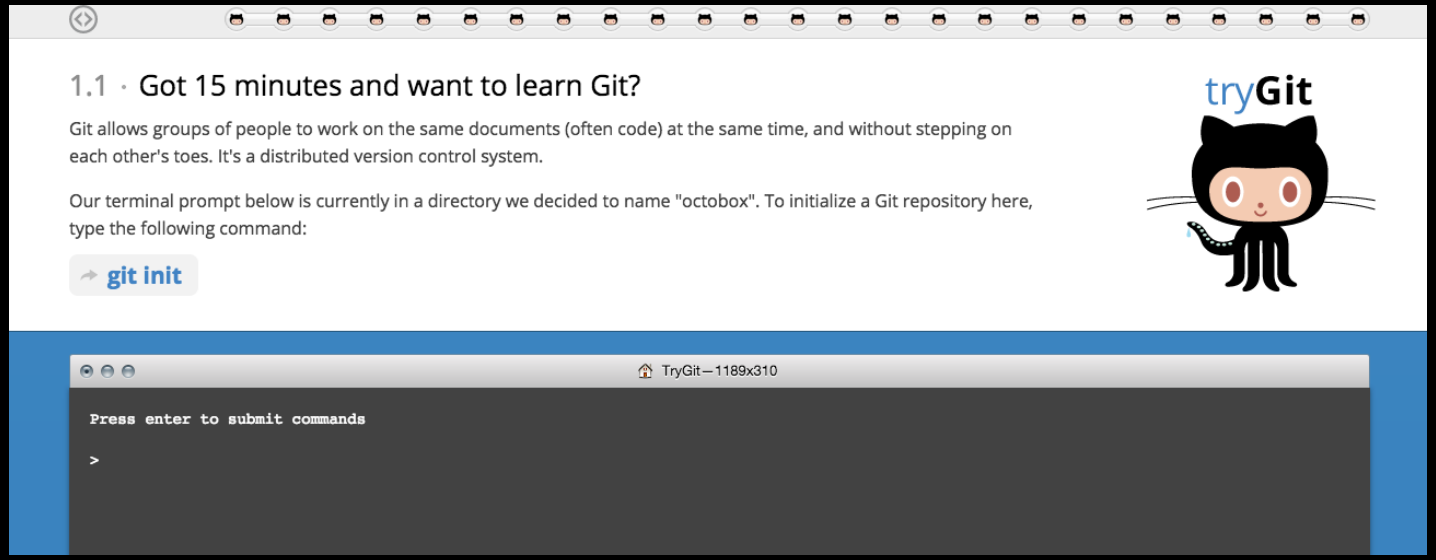
source: http://benfry.com/phd/dissertation-110323c.pdf

NOTE

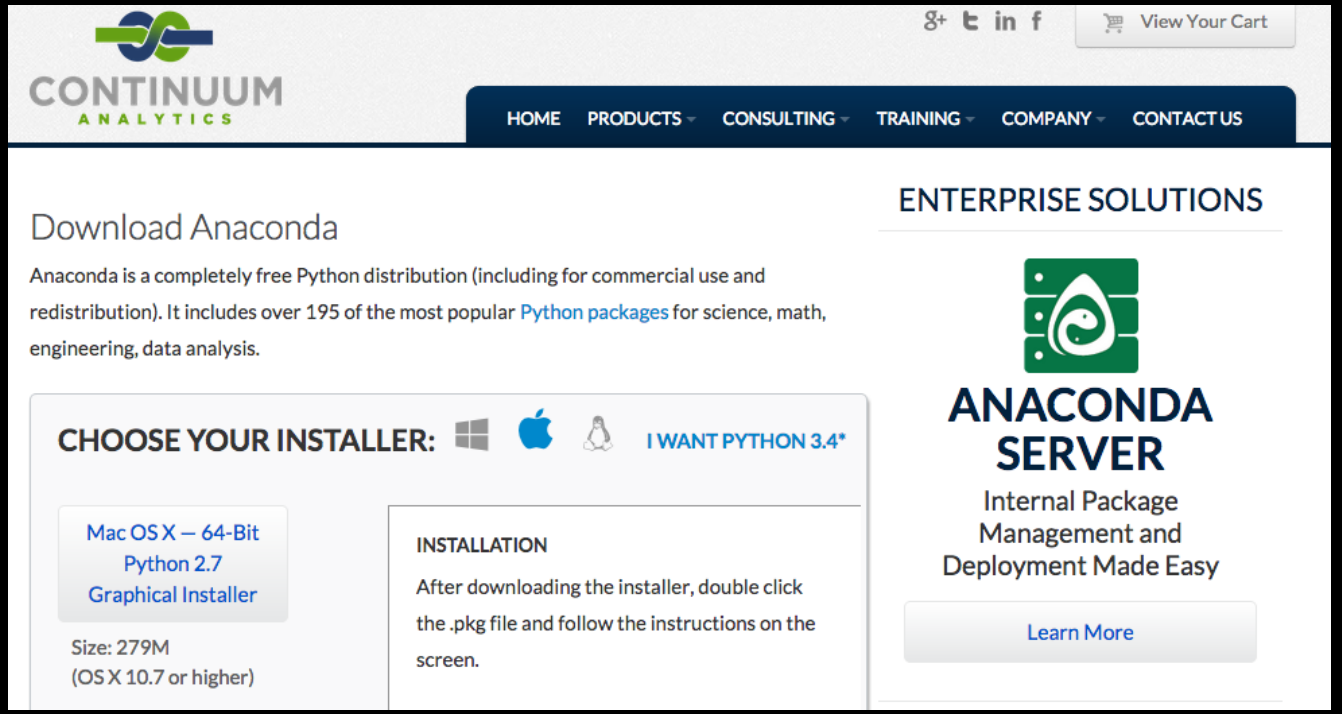
This also relies on your problem solving skills!

LAB: INTRO TO GITHUB

HTTP://TRY.GITHUB.COM/



DOWNLOAD ANACONDA



DISCUSSION