

# INTRO TO DATA SCIENCE

## LECTURE 1: DATA SCIENCE AND MACHINE LEARNING OVERVIEW

---

## INTRO TO DATA SCIENCE

---

# 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.

Instructor:

Paul Burkard ([PBURKARD88@GMAIL.COM](mailto:PBURKARD88@GMAIL.COM))

Expert-in-residence:

Jon Blum ([JON@JONBLUM.NET](mailto:JON@JONBLUM.NET))

Course Producer:

Matt Quinn ([MATT.QUINN@GENERALASSEMB.LY](mailto:MATT.QUINN@GENERALASSEMB.LY))

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

Course materials: [HTTPS://GITHUB.COM/GA-STUDENTS/DS\\_BOS\\_06](https://github.com/ga-students/ds_bos_06)

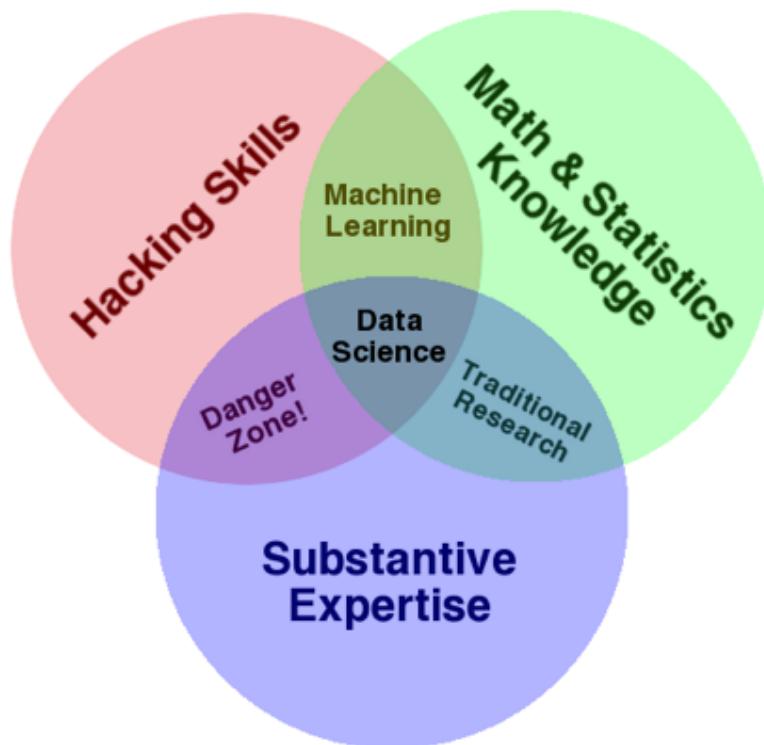
# Introductions

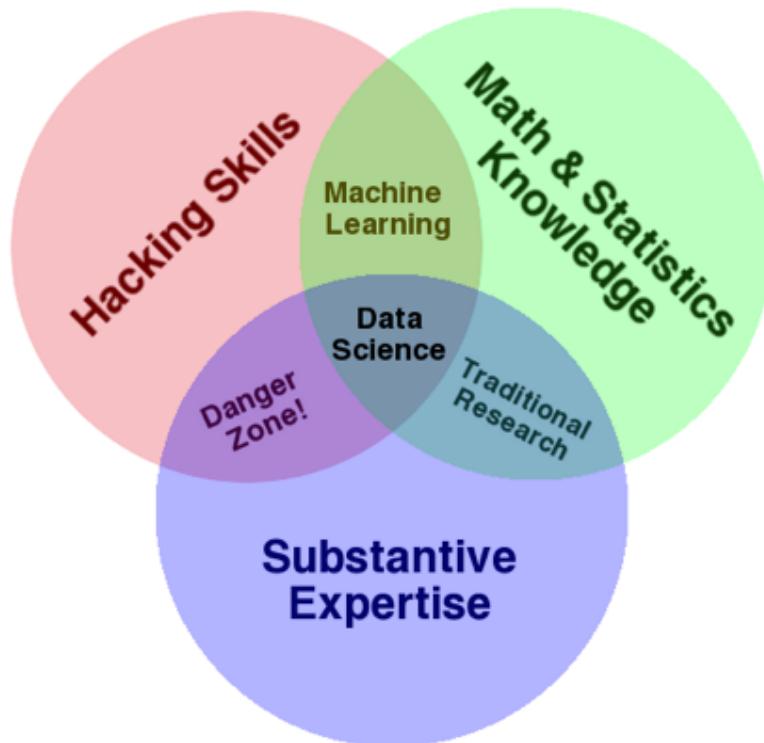
- Your name
- A brief summary of your background (e.g. work, school, etc.)
- What you hope to get out of the class
- One interesting / surprising / fun fact about yourself

# 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.

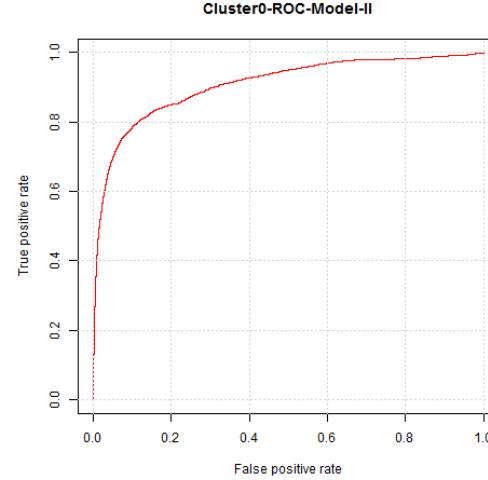
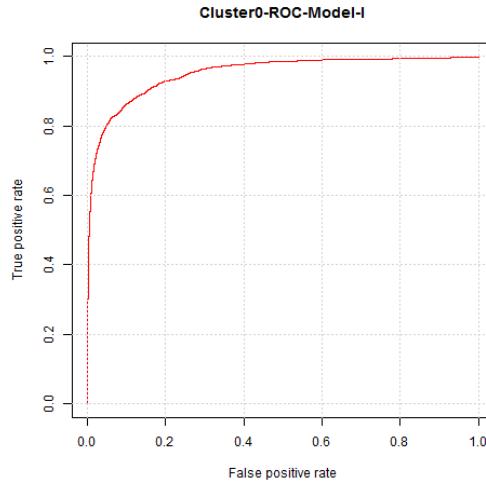




**ONE MORE THING!**

Communication skills

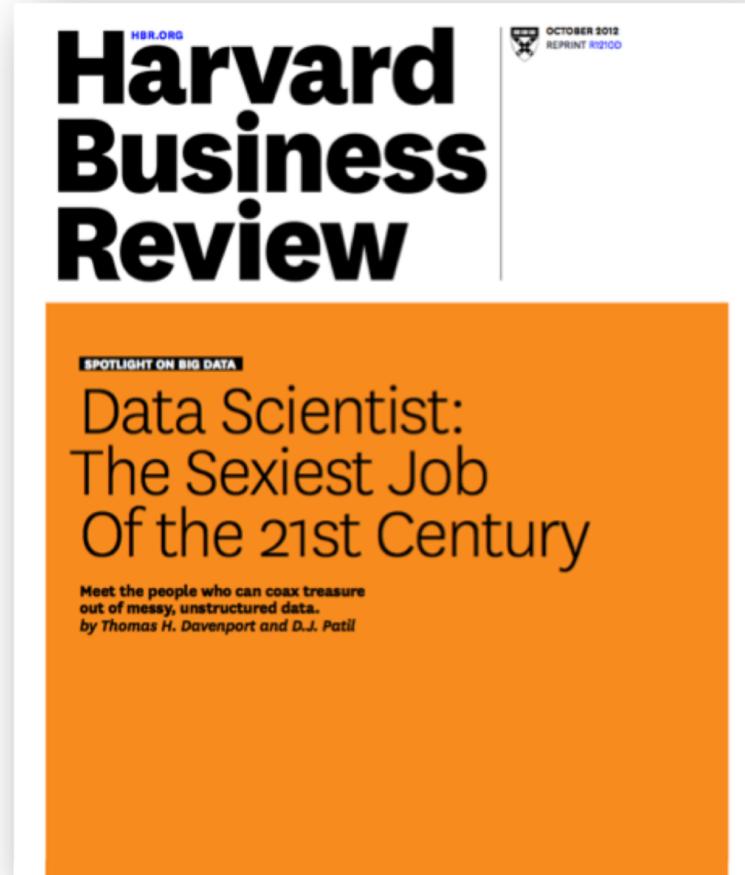
- ▶ The importance of communication:



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.



**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

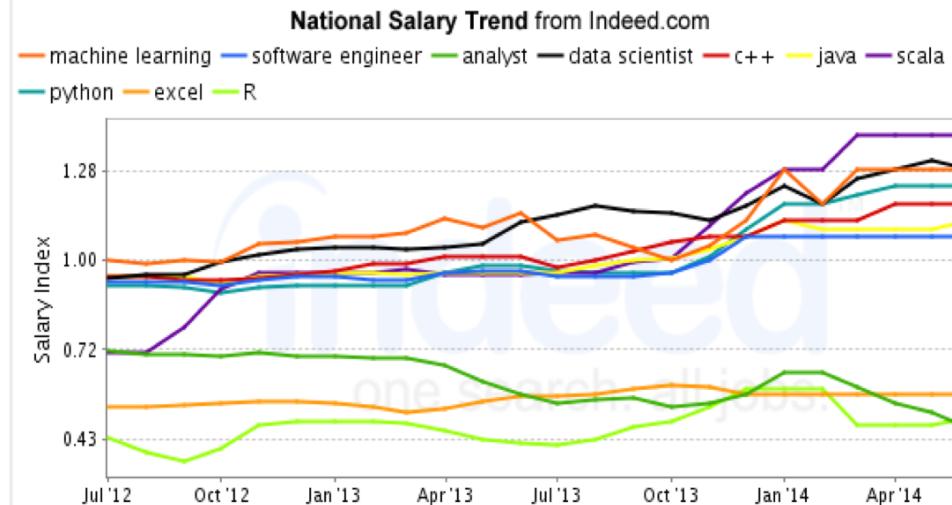
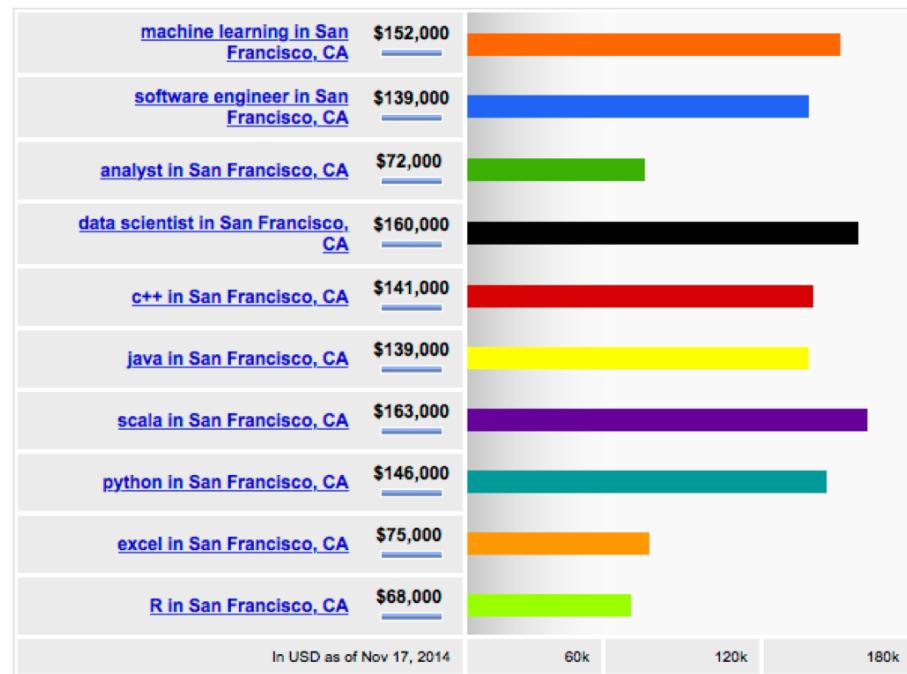
DATA

# 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





## 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



on Allstate company

## 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



THOMSON REUTERS

## 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 <input type="button" value="Add"/> ★★★☆☆ <input type="radio"/> Not Interested	300 <input type="button" value="Add"/> ★★★★★ <input type="radio"/> Not Interested	The Rundown <input type="button" value="Add"/> ★★★☆☆ <input type="radio"/> Not Interested	Bad Boys II <input type="button" value="Add"/> ★★★☆☆ <input type="radio"/> 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%

The screenshot shows the Netflix Prize Leaderboard. At the top, it displays "Leaderboard 10.05%" and an option to "Display top 20 leaders". The main table lists the top submissions, with the first entry highlighted. A yellow arrow points to the "% Improvement" column for the top submission.

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



The Kaggle logo, which consists of the word "kaggle" in a lowercase, sans-serif font.[Sign up](#)[Login](#)

# The Home of Data Science

COMPETITIONS • CUSTOMER SOLUTIONS • JOBS BOARD

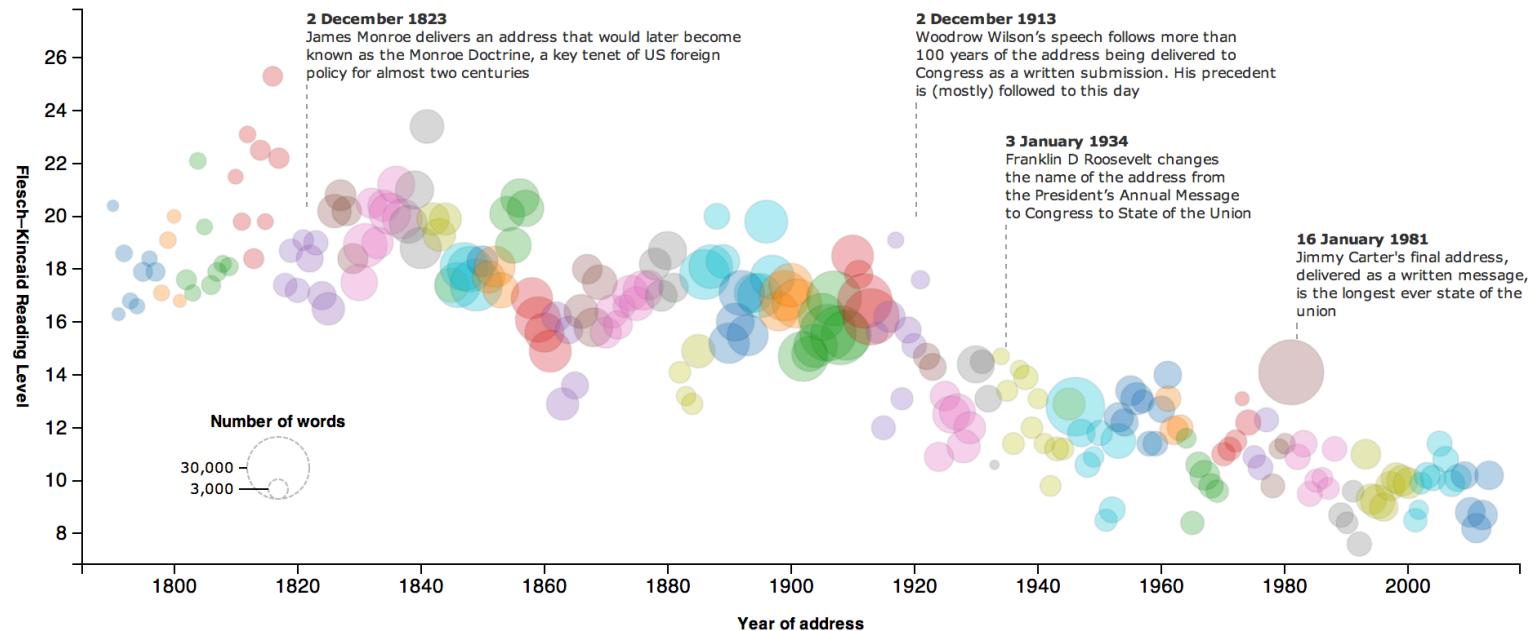
[Get started »](#)



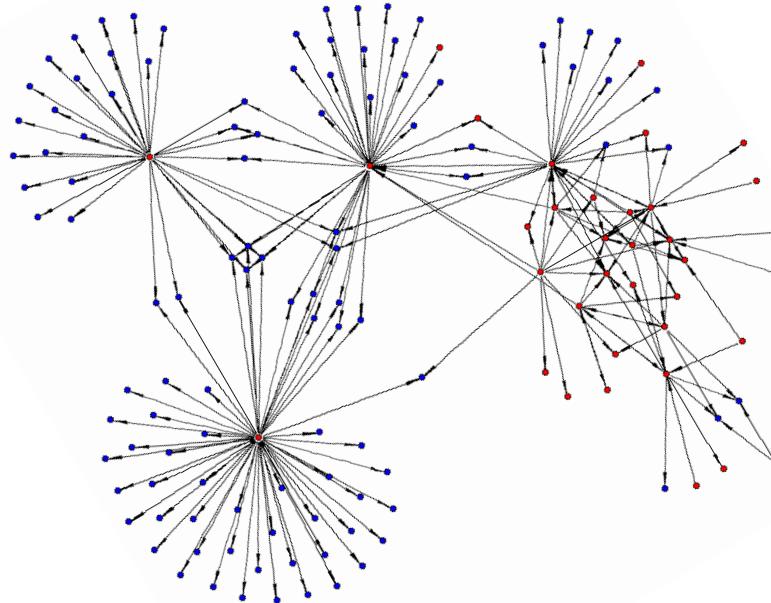
- Stack Overflow tag recommendation and response time prediction
- Locating ethnic food in ethnic neighborhoods
- Building optimal NBA teams
- Recommending new musical artists
- Prioritize emergency calls in Seattle
- Finding the right college for you

## 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](http://bit.ly/NHmRqu)  
[@peteskomoroch](https://twitter.com/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

# Jeff Hammerbacher

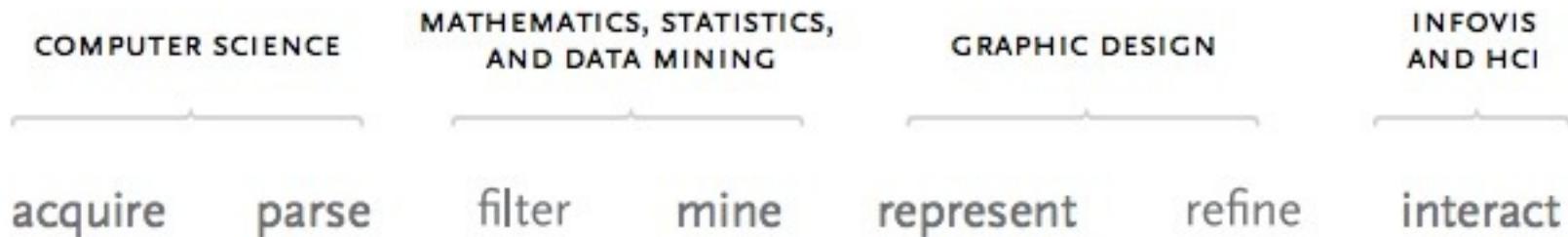
- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

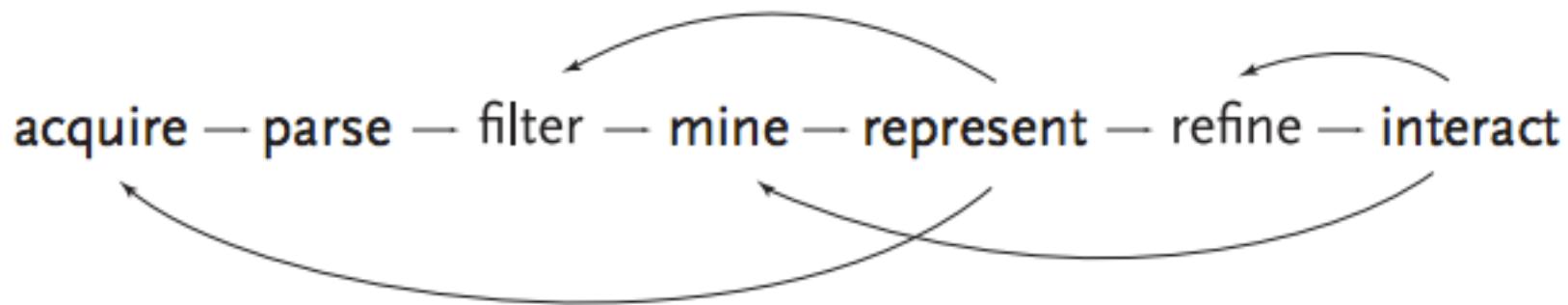
### Ted Johnson

- › 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





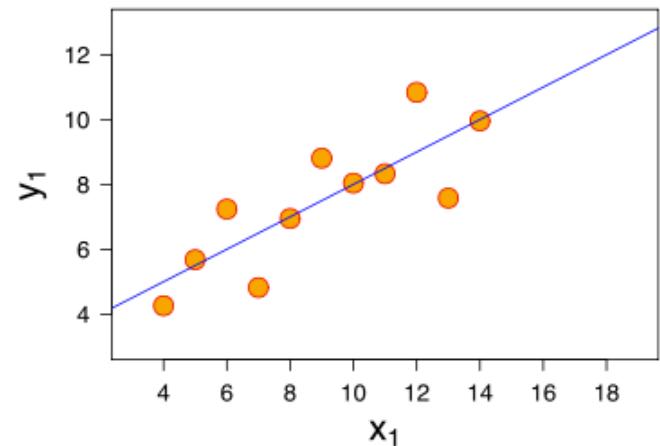
**NOTE**

This diagram illustrates the iterative nature of problem solving

# VISUALIZATIONS AS A MEDIUM

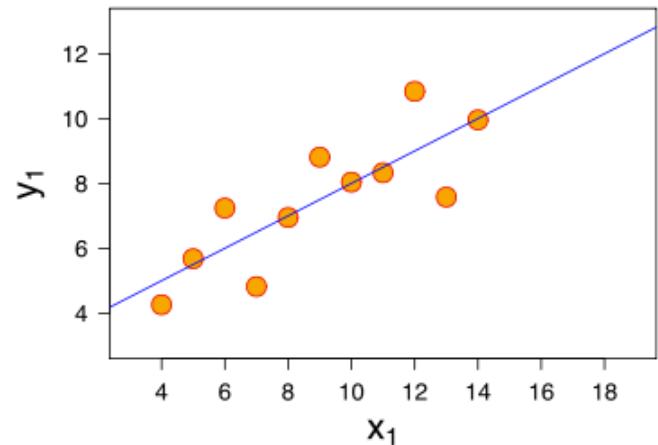
*Consider the following dataset:*

- *eleven ( $x, y$ ) points*



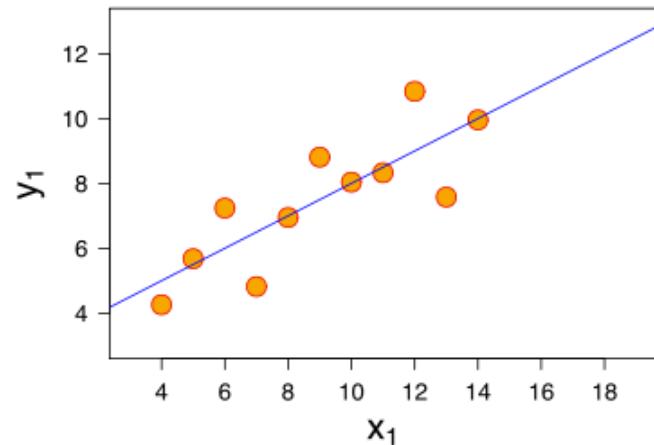
*Consider the following dataset:*

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



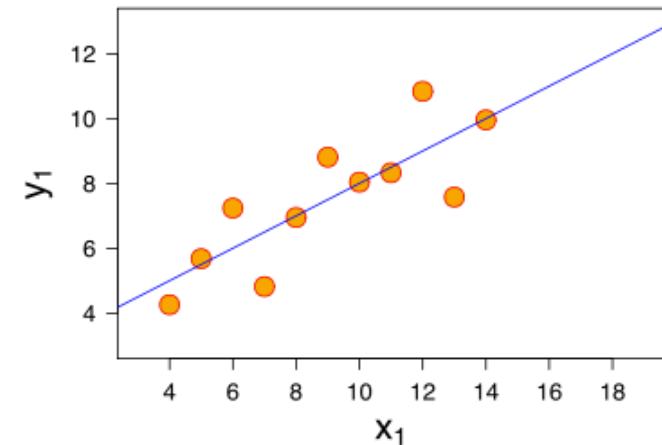
*Consider the following dataset:*

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



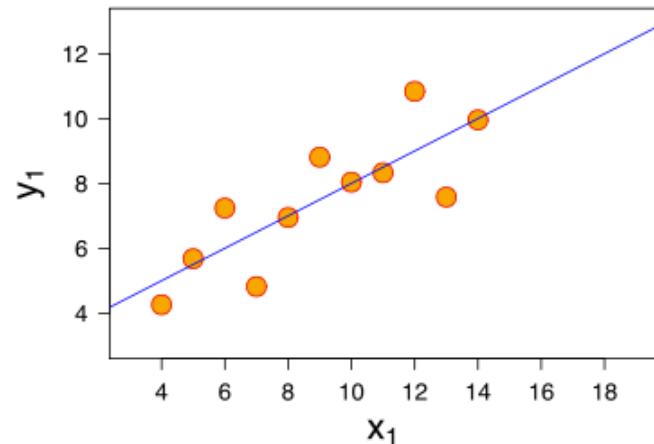
*Consider the following dataset:*

- *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:  $y = 3.00 + 0.500x$*

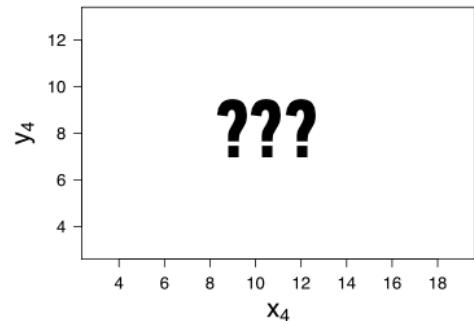
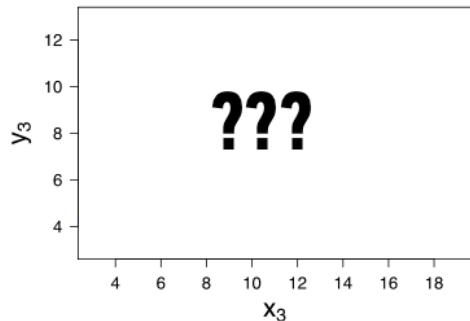
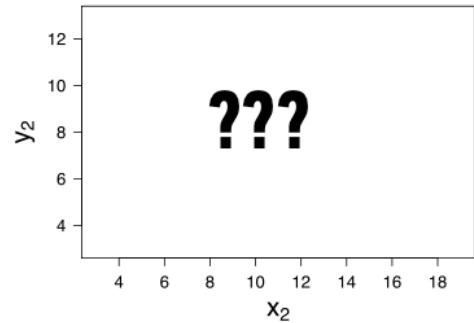
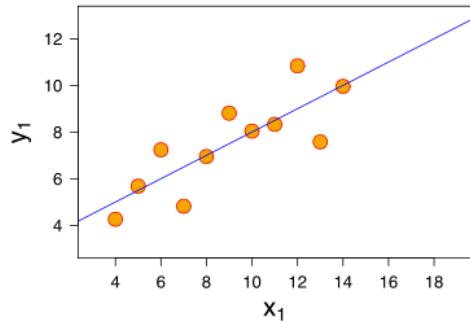


## EXERCISE – WHY VISUALIZE DATA?

45

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

*Q: how similar are these  
datasets?*



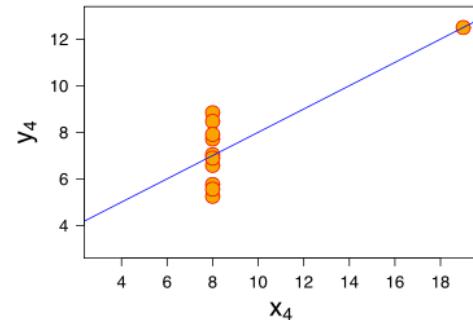
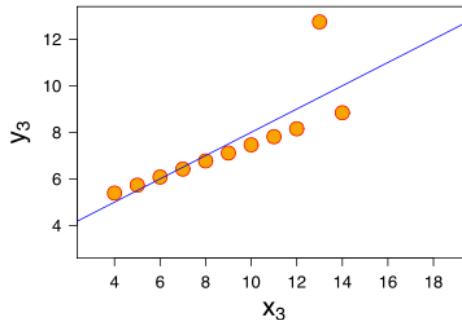
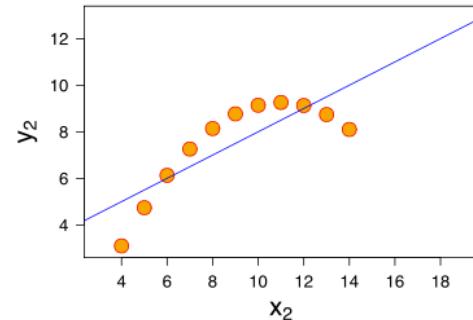
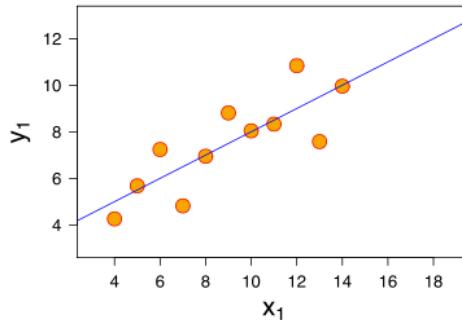
## EXERCISE – WHY VISUALIZE DATA?

46

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

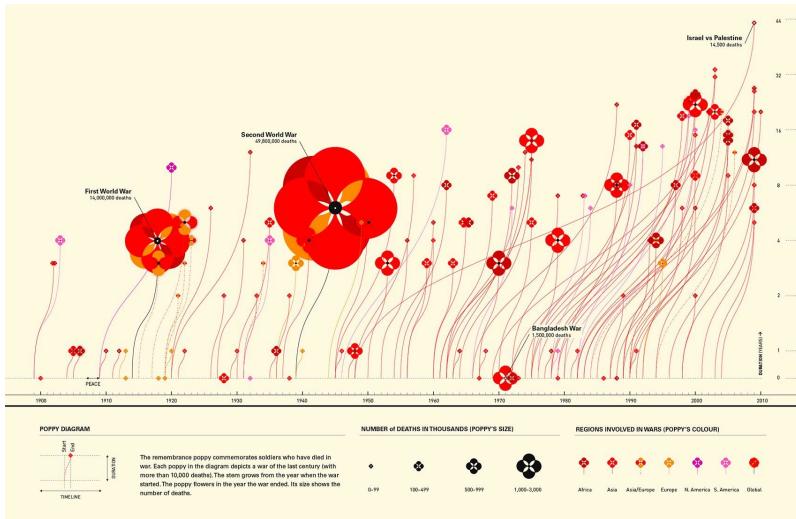
*Q: how similar are these  
datasets?*

*A: not very!*

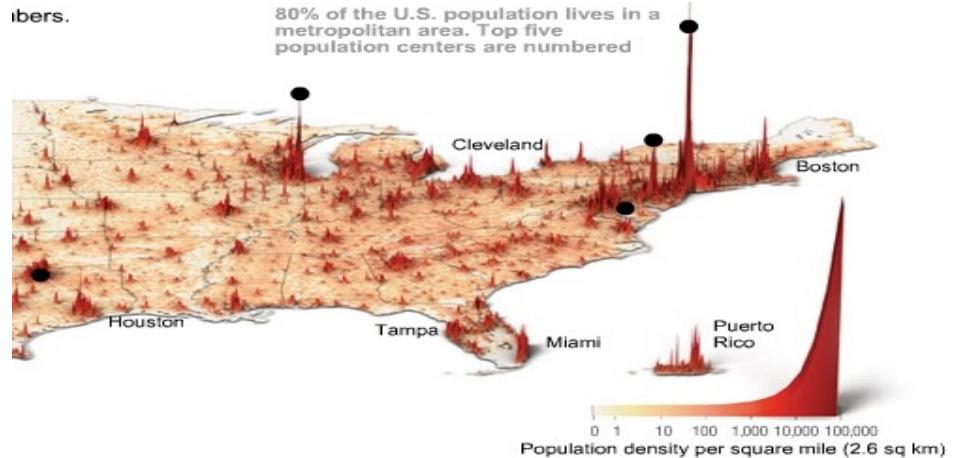


# **VISUALIZATION: BEING CREATIVE**

47

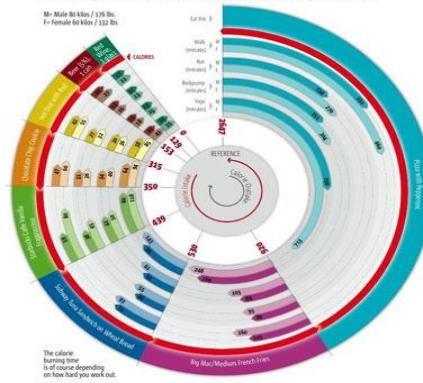


ibers.



## 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 2008 a year when you make conscious decisions and stop the weight to sneak up on you.



# 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](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](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...”

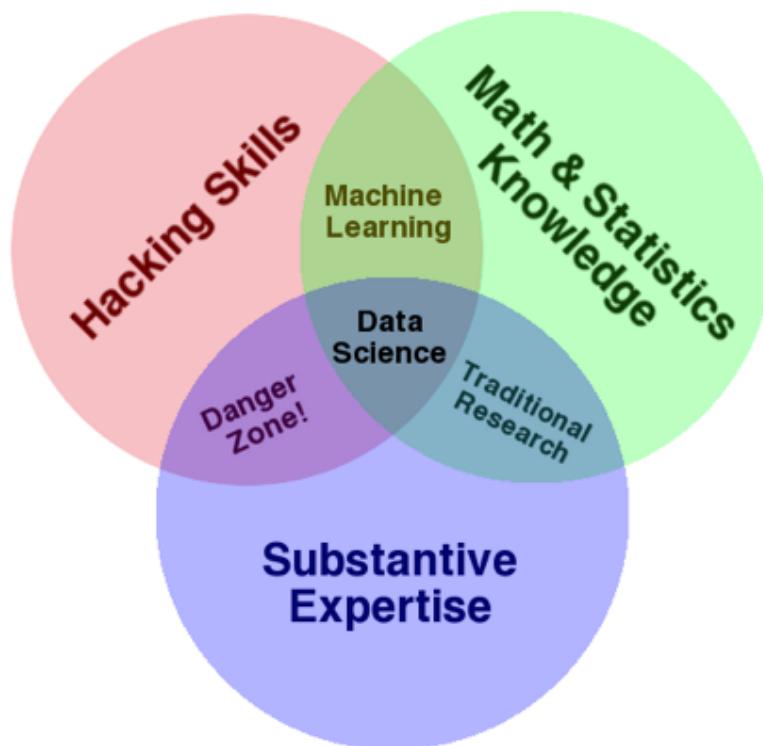
- representation – extracting structure from data

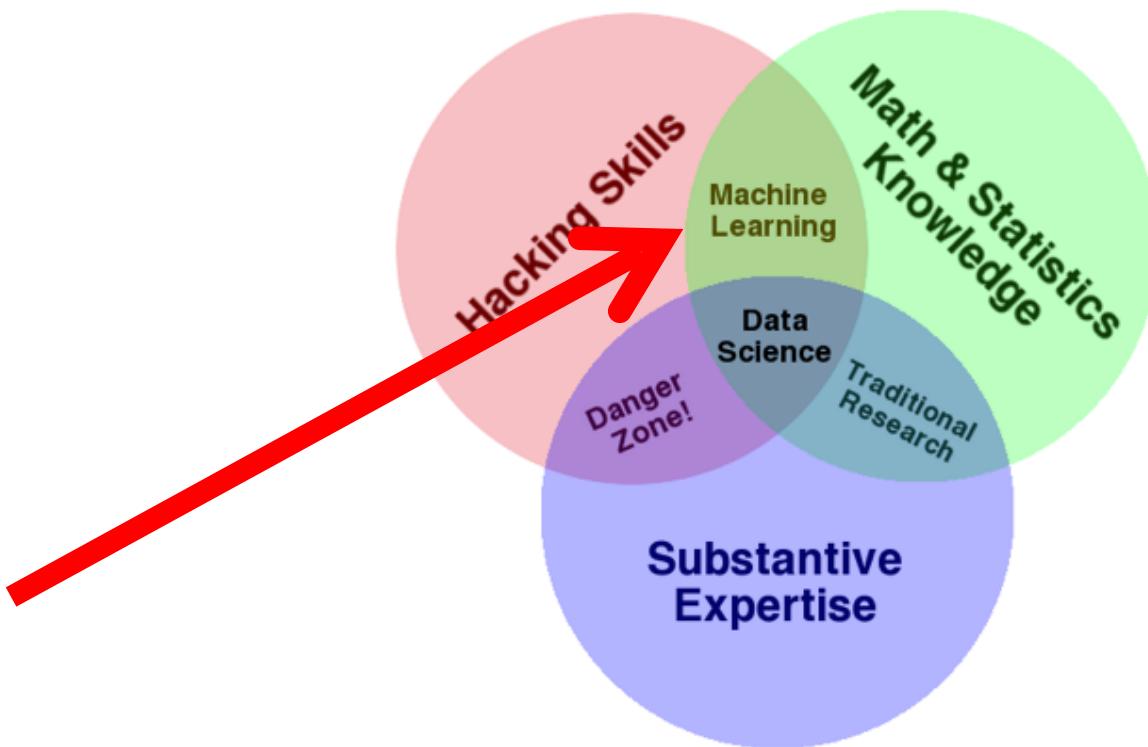
from Wikipedia:

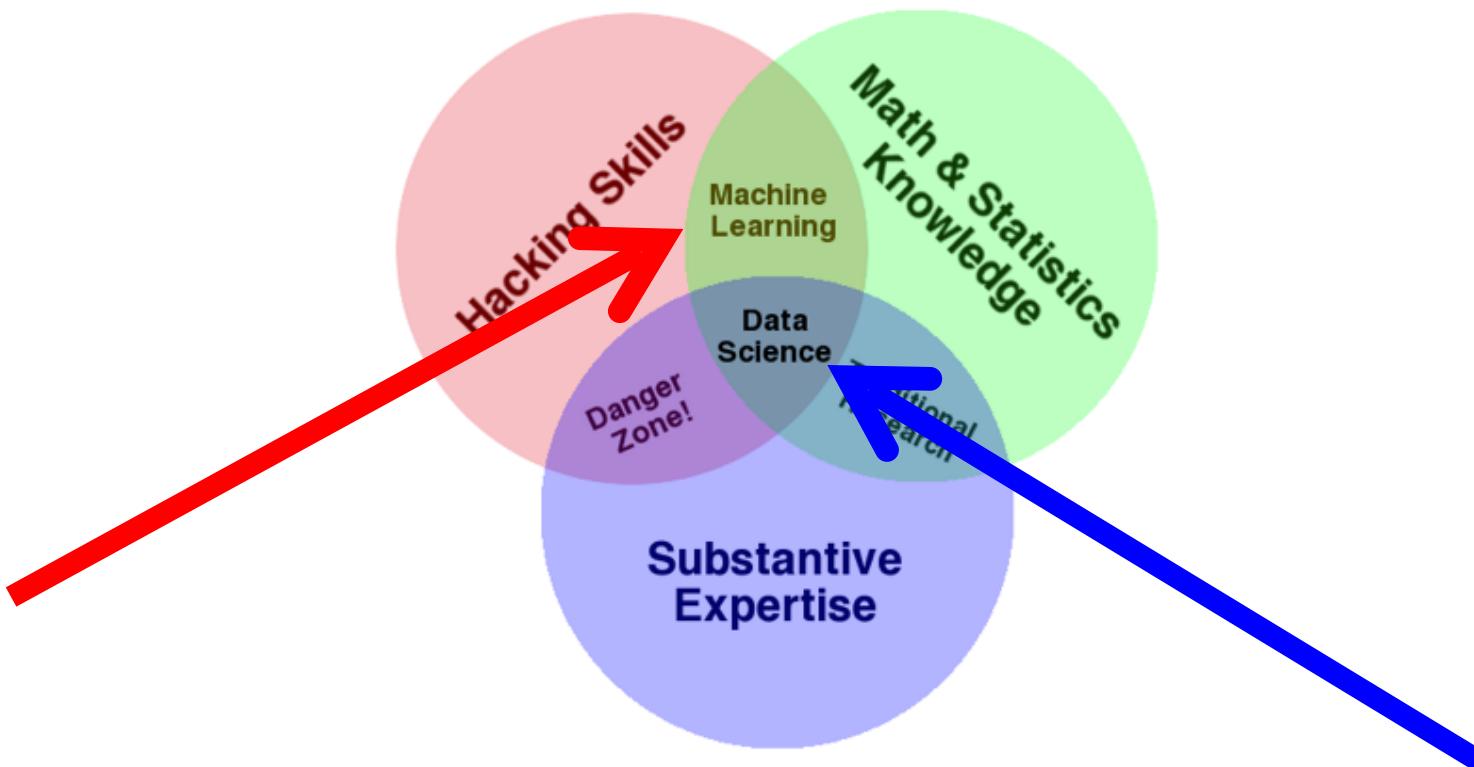
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

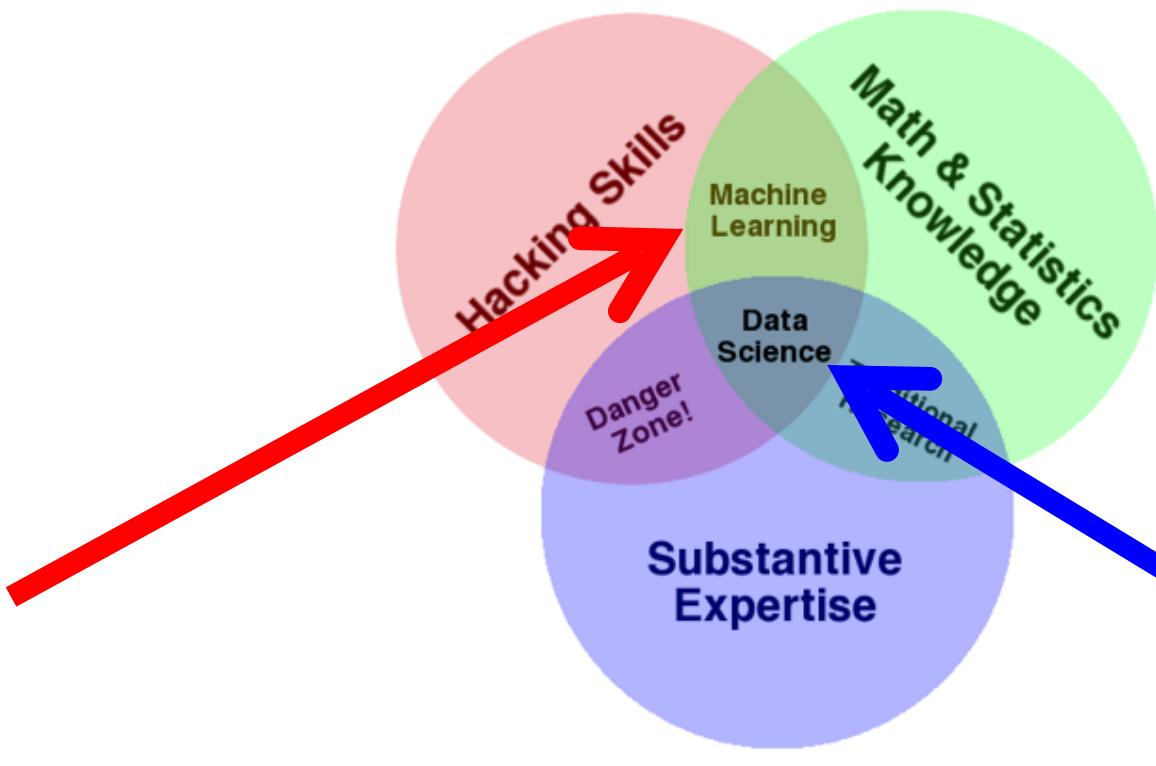
“The core of machine learning deals with representation and generalization...”

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









QUESTION

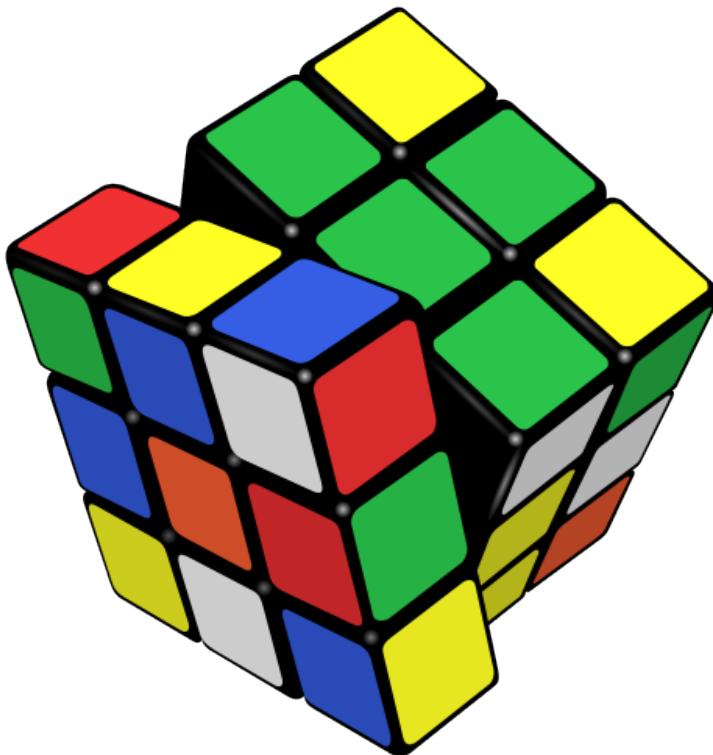
What does it take to make this jump?

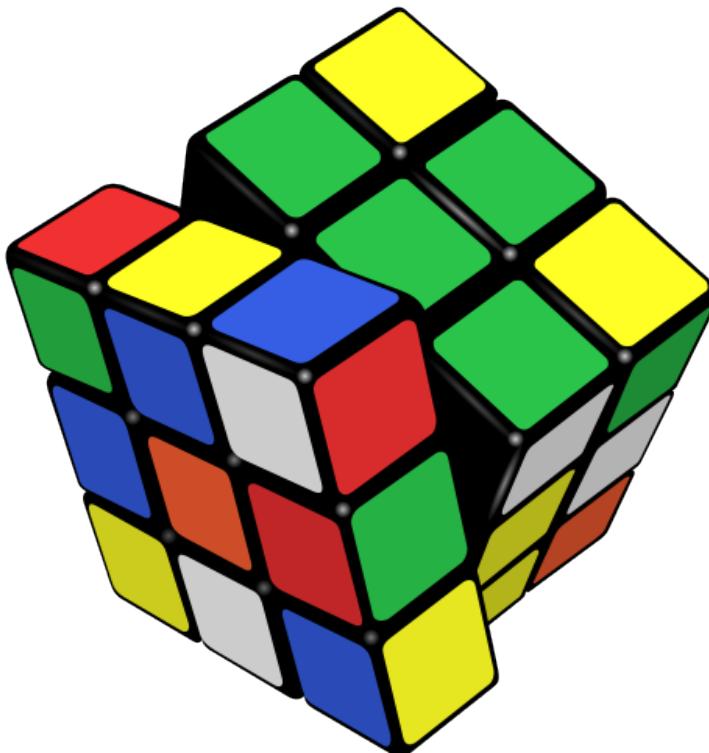
---

**ANSWER: PROBLEM SOLVING!**

---

**57**





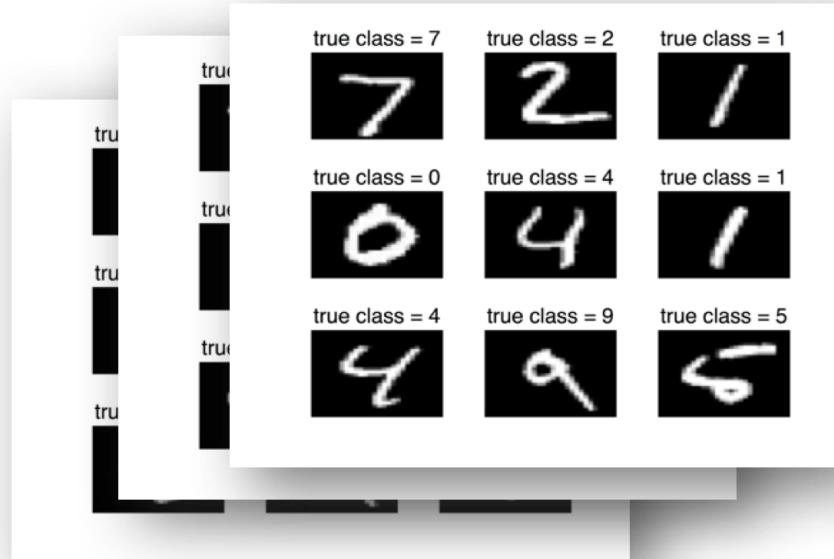
**NOTE**

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



4	1	5	7	1	3	3	6	4	8	1	9	7	6	3	6	9	3	0	6
4	7	7	8	1	3	7	2	4	6	4	3	2	8	6	1	4	3	0	9
1	7	7	6	5	8	6	0	0	3	9	5	4	1	5	7	2	3	2	1
3	5	2	5	7	3	2	9	7	1	6	9	4	6	8	3	2	4	1	9

**CLICK HERE  
TO APPLY TODAY!**



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

## CREDIT SCORING

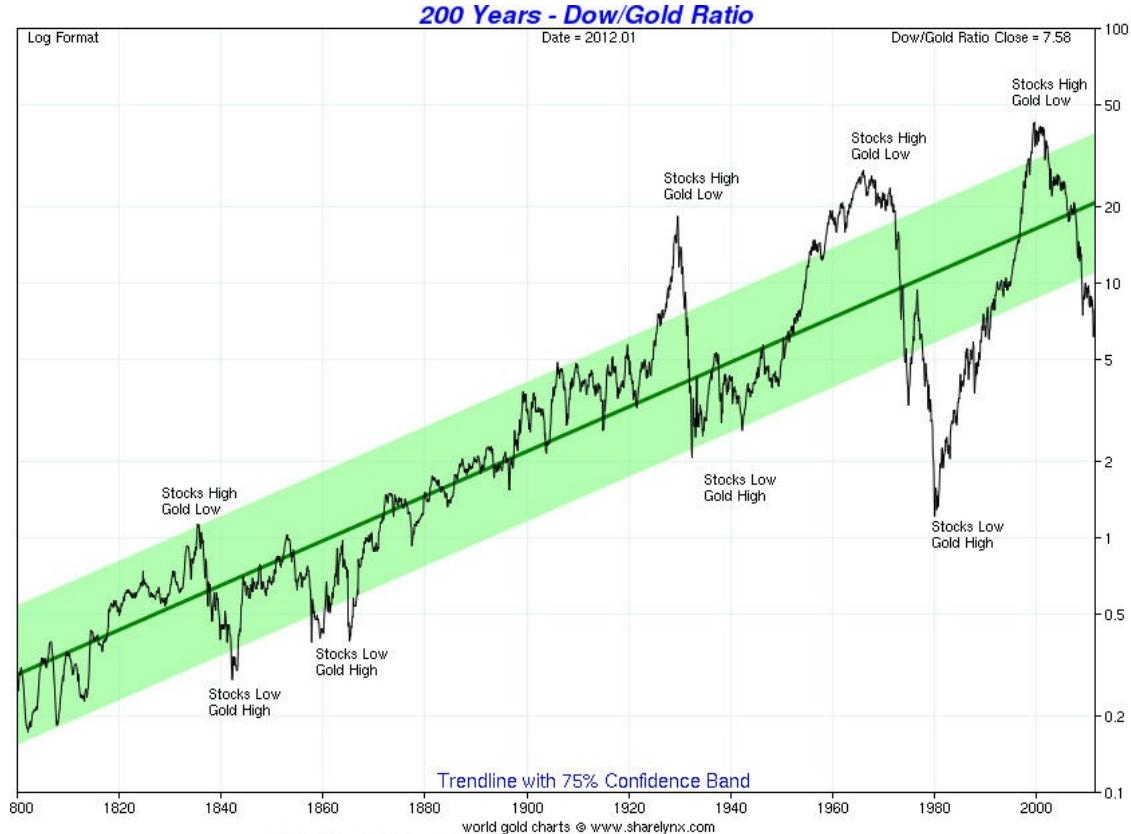
64

	Client 1	Client 2	Client 3		Applicant
Age	23	30	19	Age	25
Gender	M	F	M	Gender	M
Annual salary	\$30,000	\$45,000	\$15,000	Annual salary	\$25,000
Years in residence	3 years	1 year	3 month	Years in residence	1 year
Years in job	1 year	1 year	1 month	Years in job	2 years
Current debt	\$5,000	\$1,000	\$10,000	Current debt	\$15,000
Paid off credit	Yes	Yes	No	Credit decision/score	???



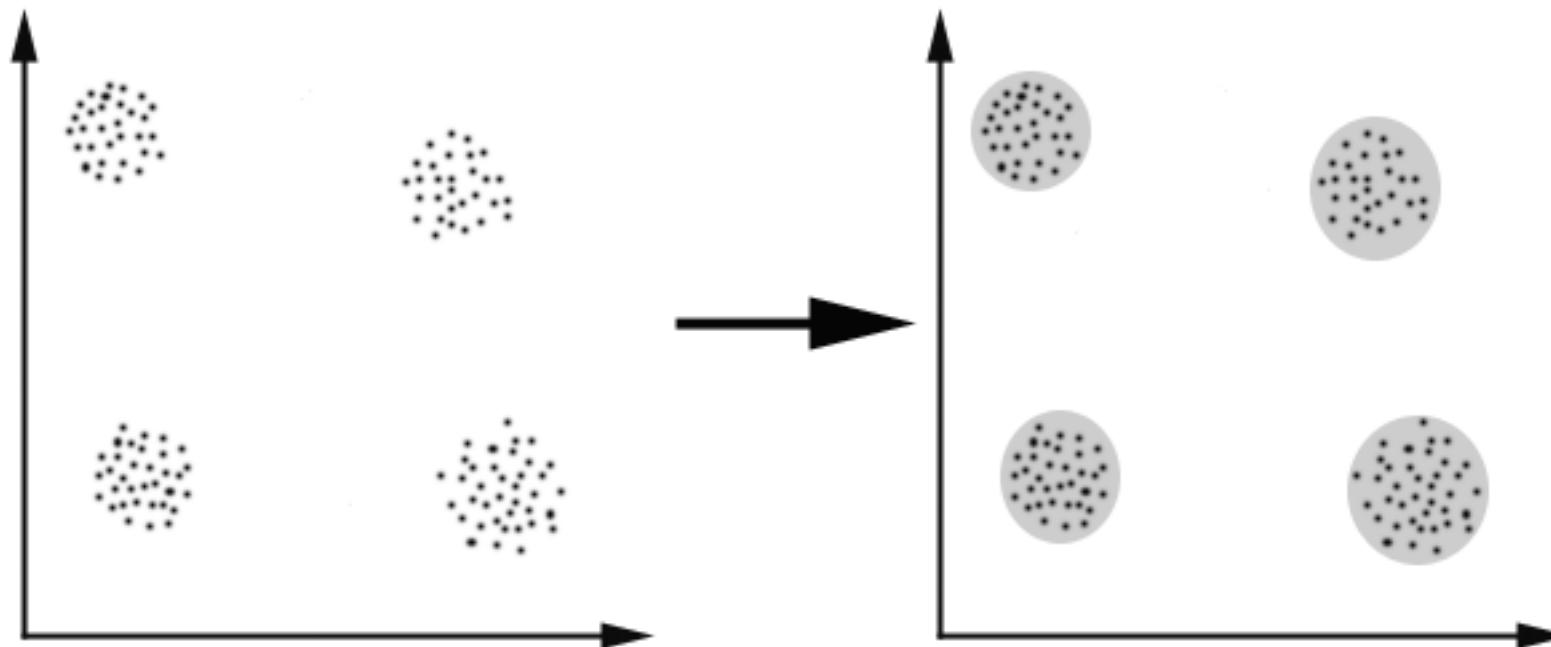
# REGRESSION - STOCK PRICE PREDICTION

66



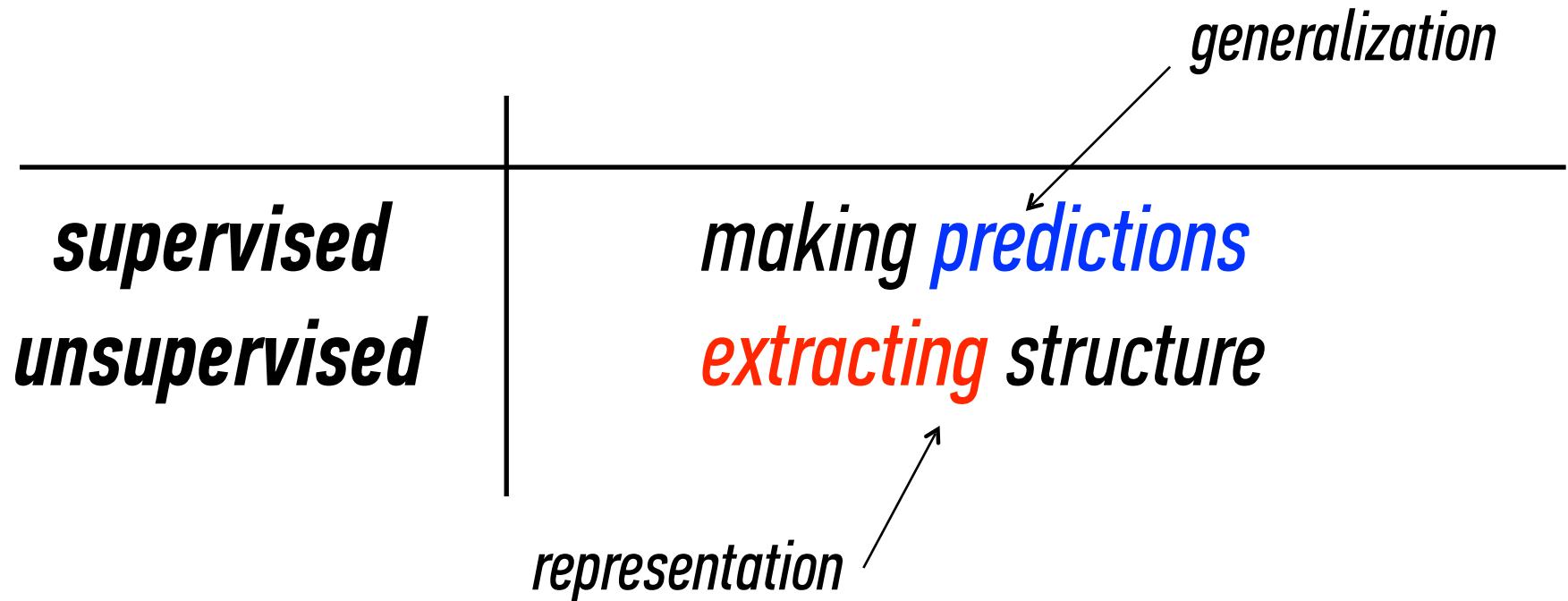
**Unsupervised learning:** the goal is to learn interesting patterns and **structure** in data given only inputs

no label information given at all



---

<i><b>supervised</b></i>	<i><b>making predictions</b></i>
<i><b>unsupervised</b></i>	<i><b>extracting structure</b></i>

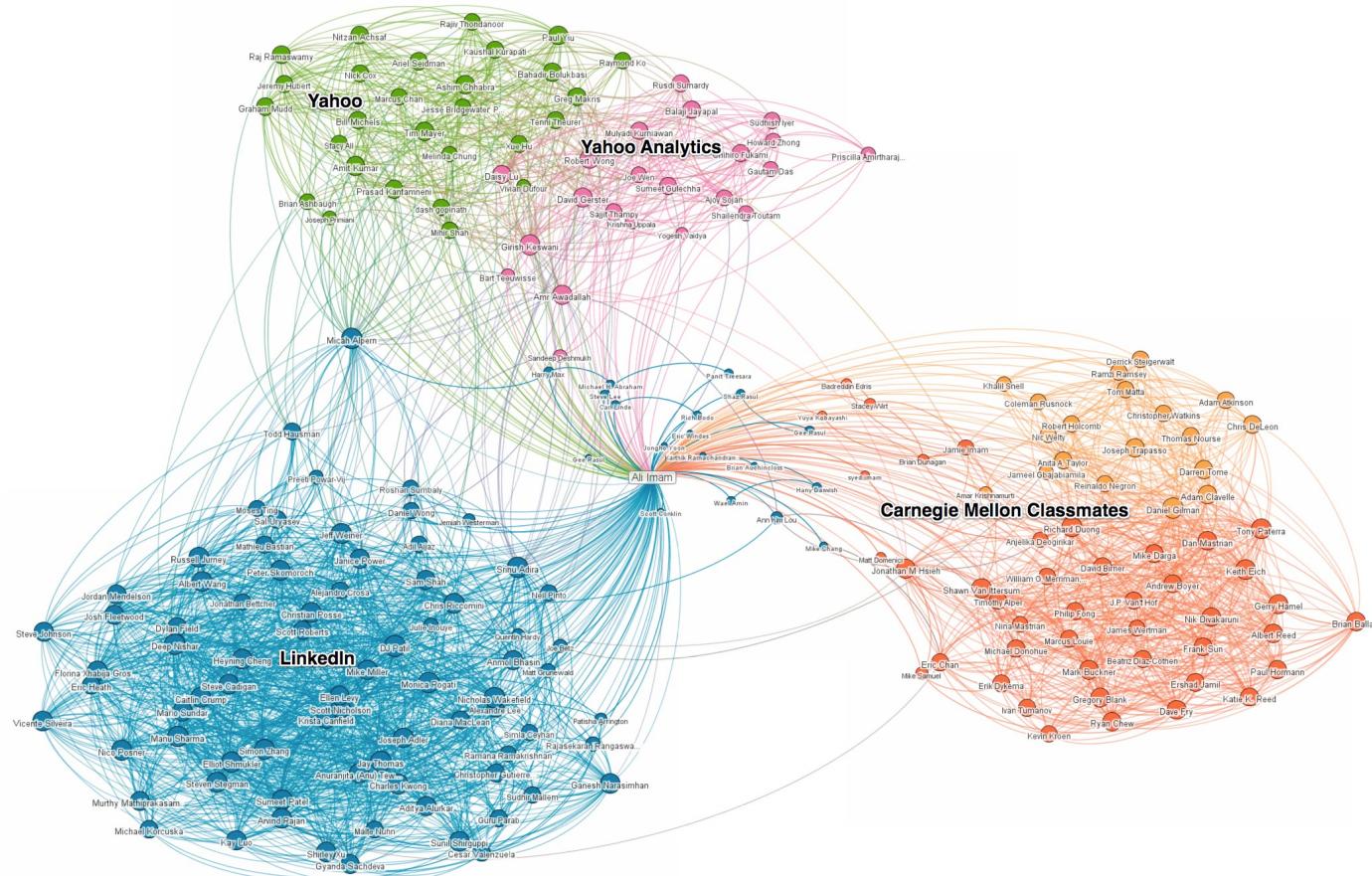


# EXERCISE:

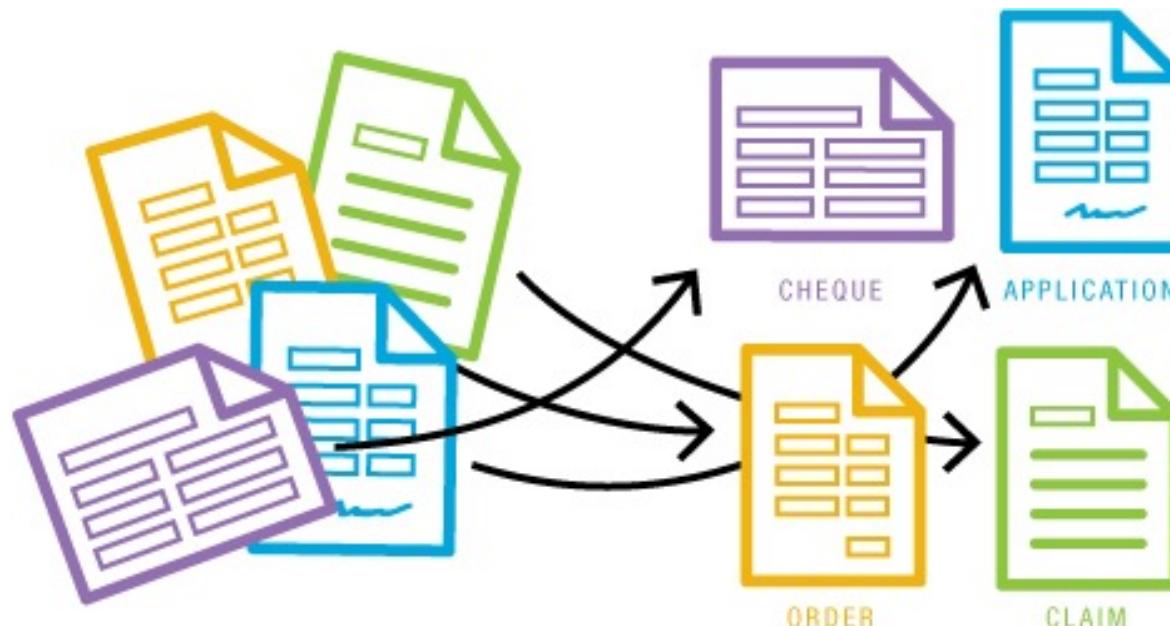
supervised or unsupervised?

# COMMUNITY DETECTION IN SOCIAL NETWORKS

72







*continuous*

*categorical*

*quantitative*

*qualitative*

***continuous***

*Height of children*

*Weight of cars*

*Speed of the train*

*Temperature*

*Stock price*

***categorical***

*Eye colors*

*Courses at GA*

*Highest degree*

*Gender*

*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.

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

**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***  
***unsupervised***

*making predictions*  
*extracting structure*

**ANSWER**

The goal is determined  
by the type of problem.

---

QUESTION

---

***HOW  
DO YOU  
DETERMINE  
THE RIGHT  
APPROACH?***

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

**ANSWER**

The right approach is determined by the desired solution.

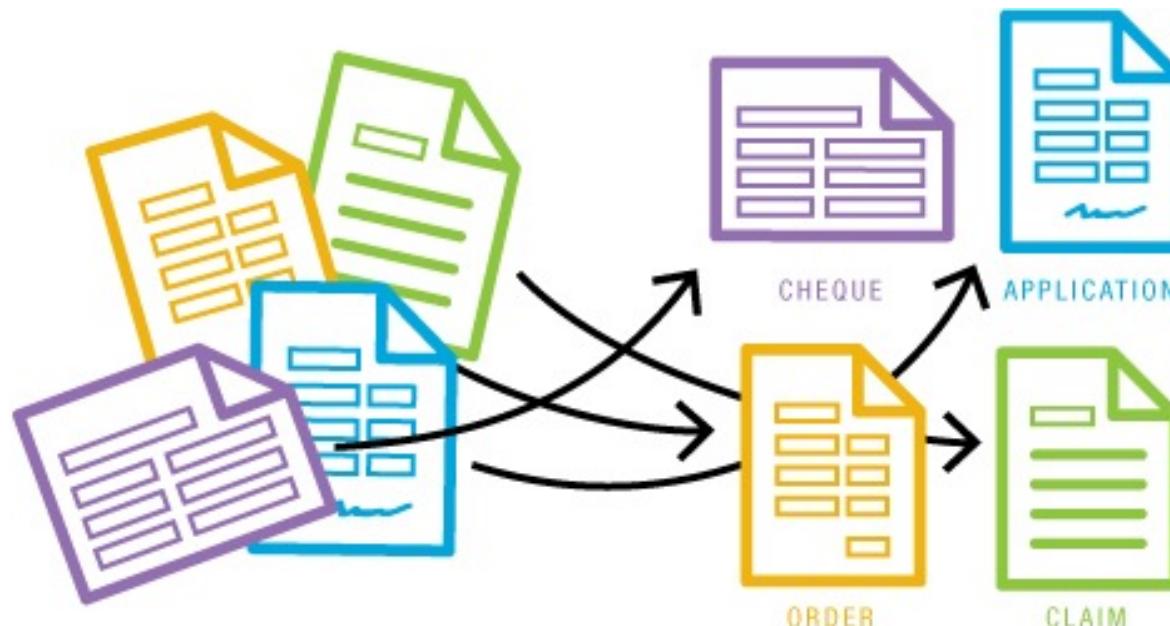
	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

**NOTE**

All of this depends on  
your data!

## DO WE HAVE LABELS?

85

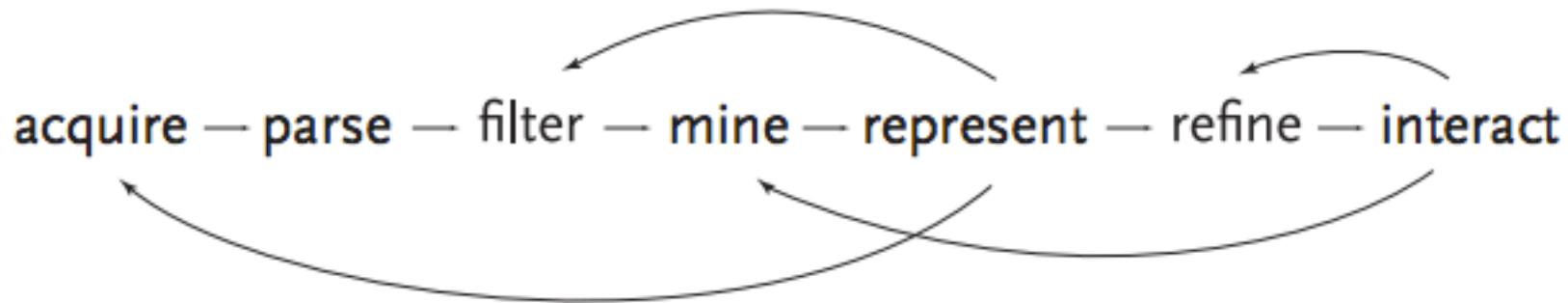


---

QUESTION

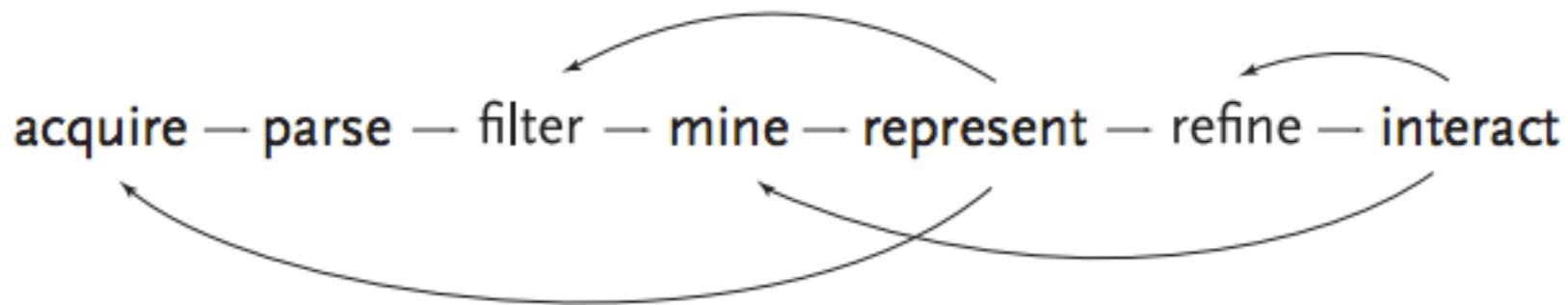
---

**WHAT  
DO YOU  
DO  
WITH YOUR  
RESULTS?**



### ANSWER

Interpret them and react accordingly.



**NOTE**

This also relies on  
your problem solving  
skills!

---

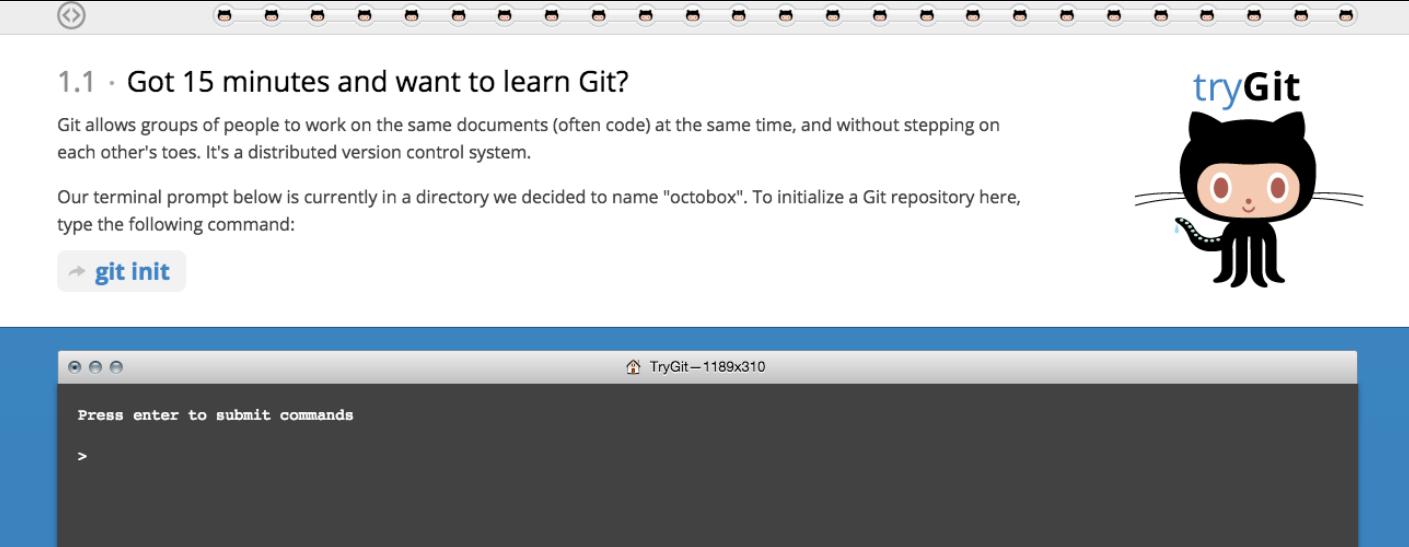
INTRO TO DATA SCIENCE

---

# LAB: INTRO TO GITHUB

## INTRO TO DATA SCIENCE

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



1.1 · Got 15 minutes and want to learn Git?

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.

Our terminal prompt below is currently in a directory we decided to name "octobox". To initialize a Git repository here, type the following command:

```
git init
```

Press enter to submit commands  
>

tryGit



## INTRO TO DATA SCIENCE

# DOWNLOAD ANACONDA

The screenshot shows the Continuum Analytics website with a dark header. The header features the Continuum Analytics logo (a stylized infinity symbol composed of blue and green segments) and the word "CONTINUUM" with "ANALYTICS" underneath. To the right of the logo are social media icons for Google+, Twitter, LinkedIn, and Facebook, followed by a "View Your Cart" button.

The main navigation menu includes links for HOME, PRODUCTS, CONSULTING, TRAINING, COMPANY, and CONTACT US.

The left side of the page has a section titled "Download Anaconda". It contains a paragraph about Anaconda being a free Python distribution and its popularity. Below this is a "CHOOSE YOUR INSTALLER:" section with icons for Windows, Mac, and Linux, and a link to "I WANT PYTHON 3.4\*".

The right side of the page is titled "ENTERPRISE SOLUTIONS" and features a section for "ANACONDA SERVER" with a green icon of a server and the text "Internal Package Management and Deployment Made Easy". A "Learn More" button is located at the bottom of this section.

Technical details in the "Download Anaconda" section include a "Mac OS X – 64-Bit Python 2.7 Graphical Installer" link (size: 279M, OS X 10.7 or higher) and an "INSTALLATION" section describing the .pkg file download and installation process.

---

**INTRO TO DATA SCIENCE**

---

**DISCUSSION**