

# INTRO TO DATA SCIENCE

## LECTURE 1

MARCH 10, 2014 // DAT5 SF

# WELCOME!

---

## **RESOURCES**

---

**Instructor:** Rob Hall ([robhall.ga@gmail.com](mailto:robhall.ga@gmail.com))

**Experts-in-Residence:** Francesco Mosconi ([f@mosconi.me](mailto:f@mosconi.me)), Dow Street ([dowstreet@prodigy.net](mailto:dowstreet@prodigy.net))

**Course Producer:** Alex Hamady ([alexh@gacourses.org](mailto:alexh@gacourses.org))

**Course Times:** 6:30pm-9:30pm, Mondays & Wednesdays (580 Howard / GA Lofts)

**Office Hours:** Saturdays 10:00AM - Noon (preliminary)

**Class Dates:** March 10 - May 21

---

## **RESOURCES**

---

### **Internal:**

- Schoology - course website, course communications, discussion board
- Github - code for tutorials, homework submission, social coding
- Office Hours - additional material, discussions, specific questions
- Email - specific questions, any concerns, etc.

### **External:**

- Google
- Stackoverflow.com

---

## **TODAY'S AGENDA**

---

- 1.Course producer introduction
- 2.Instructor introductions
- 3.Student introductions
- 4.Lecture: Introduction to Data Science
- 5.Tutorial: Introduction to iPython
- 6.Q&A

---

## Meet Your Course Producer

---

### Alex Hamady



- Education Programs Producer, DS & DGM
- [alexh@generalassemb.ly](mailto:alexh@generalassemb.ly)

# Course expectations:

**BE PRESENT**

**PARTICIPATE**

**DO THE ASSIGNMENTS**

**MAKE FRIENDS**

---

## Meet Your Instructional team

---

### Rob Hall

Rob leads teams that build products incorporating data, analytics, and machine learning. As head of software products for a cleantech startup, Rob led the team responsible for turning energy data into insights and savings. Prior to cleantech, Rob led the design and delivery of products with tens of millions of users, including advertiser products at Overture / Yahoo!, web search and news search products at Yahoo!, and social products for Ask.com. Rob began his career building risk management systems for investment banks in NYC. Rob graduated from Cornell University with a BS in Engineering and earned his MBA from the Wharton School of the University of Pennsylvania.



---

## Meet Your Instructional team

---

### Francesco Mosconi

Francesco Mosconi is 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 owns a master degree in theoretical physics.



---

## Meet Your Instructional team

### Dow Street

Dow Street is a Principal Systems Engineer at LinQuest, where he specializes in network research, data visualization, large-scale distributed systems, and the design of data-focused web applications. He is currently a full-stack developer for a team at MIT Lincoln Laboratory, where he is building tools for studying the performance of mobile, ad-hoc networks. Dow has a background in network routing, and served as Executive Director of the Internet Architecture Board (IAB) from 2008-2012. He has a B.S. in Computer Science from Carnegie Mellon University, where he also studied Art, and has dabbled in graduate coursework at UC San Diego and Berkeley.



---

You :-)

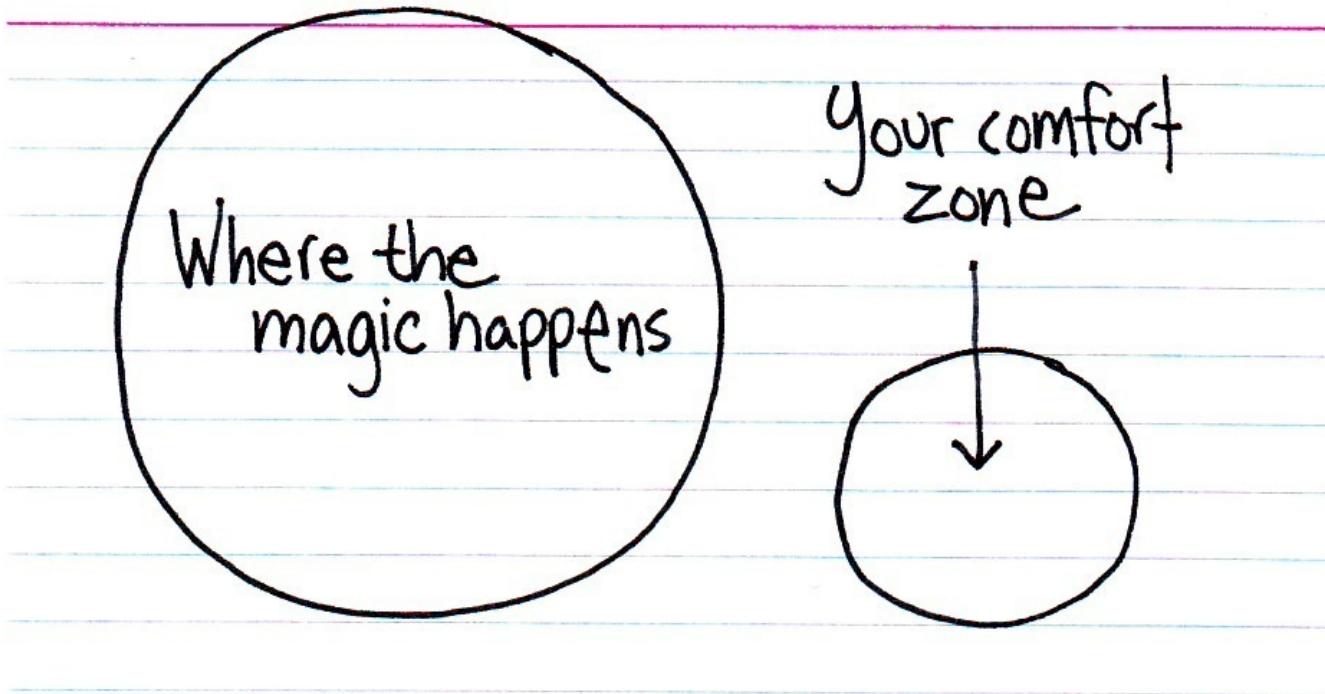
---

## Introductions

- Your name
- A brief summary of your background (e.g. work, school, etc.)
- What you hope to get out of the class

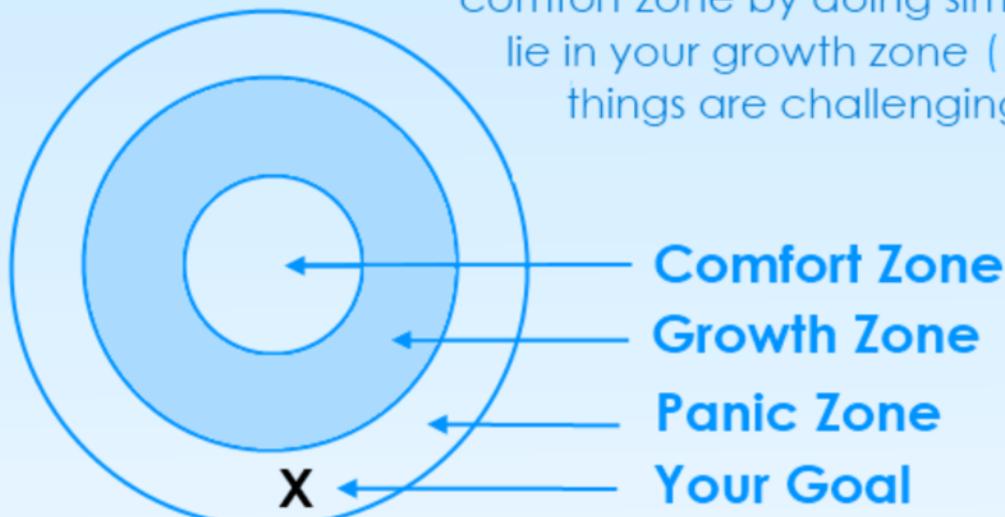
**CONGRATULATIONS...**

**...for getting out of your comfort zone!**



## How to Grow Your Comfort Zone

Any goal or challenge may fall into one of three zones - your comfort zone, growth zone, or panic zone. If your goal is currently in your panic zone, i.e. it would be too scary to do now, you will need to grow your comfort zone by doing similar challenges that lie in your growth zone (the zone in which things are challenging or scary, but do-able).

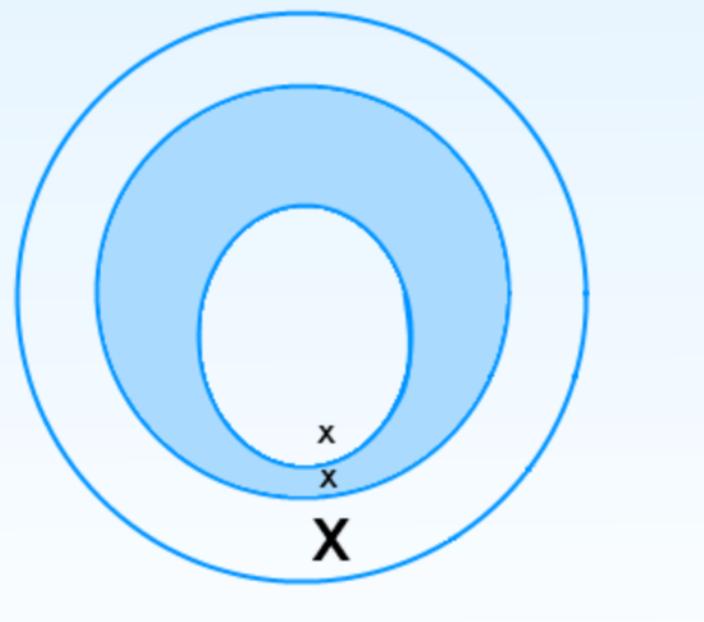


## How to Grow Your Comfort Zone

---

As you pursue challenges in your growth zone, those challenges become easier and your comfort zone expands.

Eventually, challenges that were previously in your panic zone begin to fall into your growth zone, and ultimately within your comfort zone.



---

## **AGENDA**

---

**I. WHAT IS DATA SCIENCE?**

**II. THE DATA SCIENCE WORKFLOW**

**LAB:**

**III. WORKING AT THE UNIX COMMAND LINE**

**IV. INTRO TO I-PYTHON**

# I. WHAT IS DATA SCIENCE?

---

## **WHAT IS DATA SCIENCE?**

---

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

---

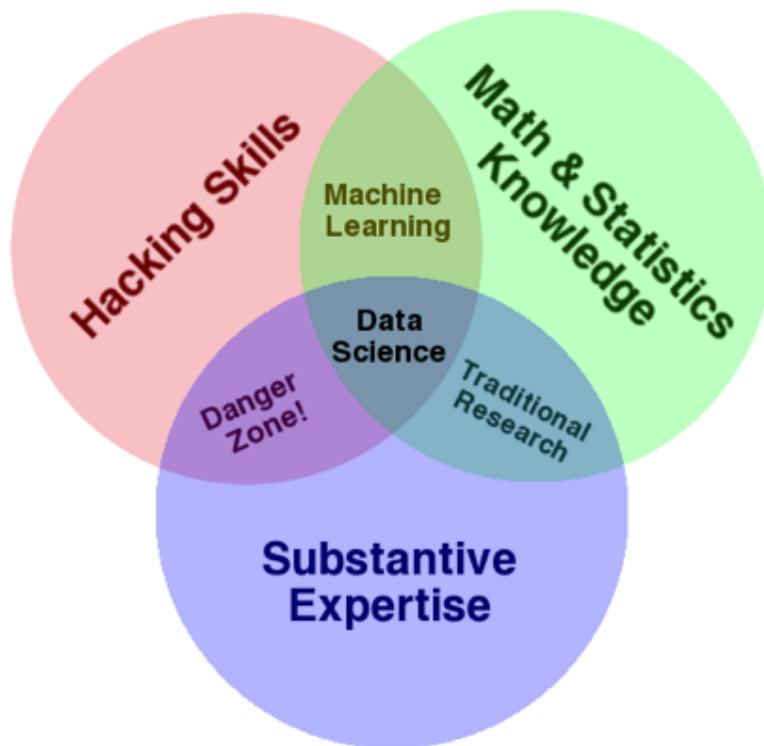
## **WHAT IS DATA SCIENCE?**

---

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

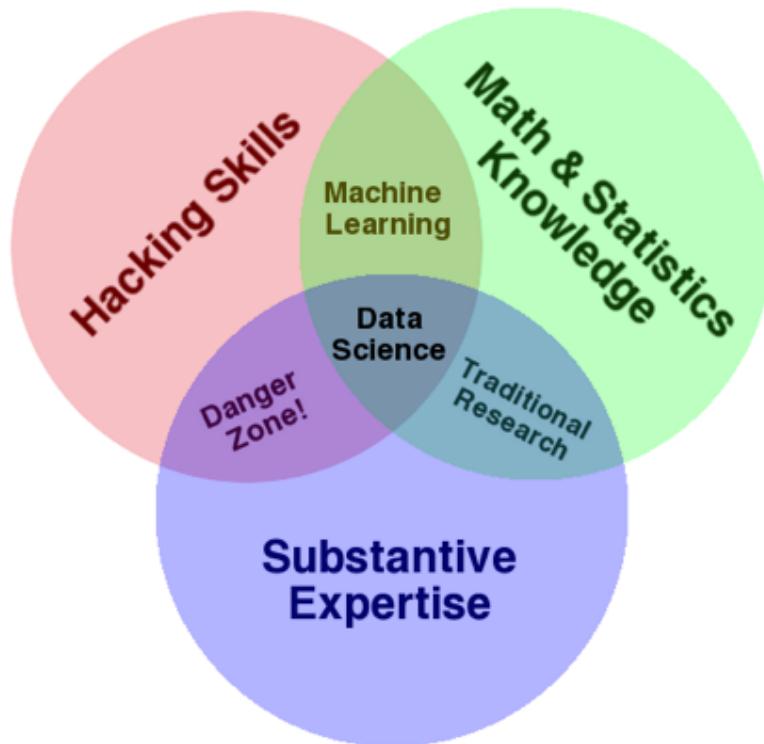
## THE QUALITIES OF A DATA SCIENTIST

---



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

## THE QUALITIES OF A DATA SCIENTIST



**ONE MORE THING!**

Communication skills

---

## **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.

## WHO USES DATA SCIENCE?

---



## WHAT MAKES A GOOD DATA SCIENTIST?

---



**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

---

## **WHAT MAKES A GOOD DATA SCIENTIST?**

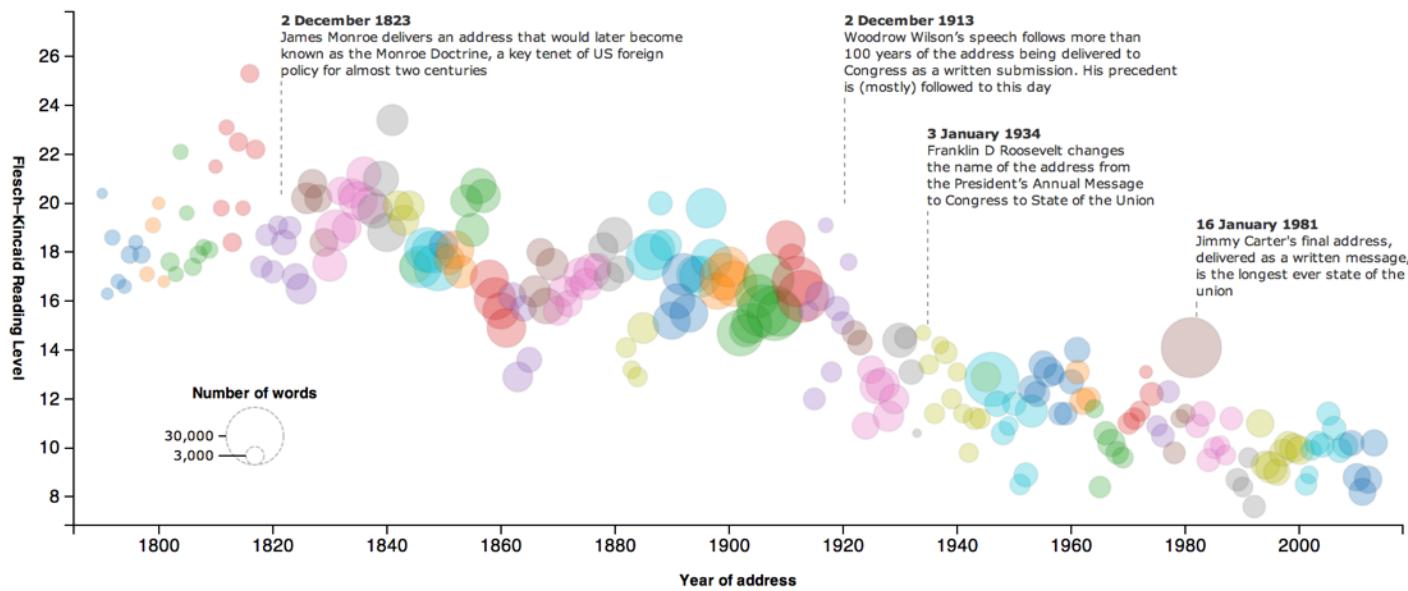
---

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

## WHO USES DATA SCIENCE?

### 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



---

## **WHO USES DATA SCIENCE?**

---

Music + Data:

<http://bit.ly/echonest>

---

## **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 job for you (Bright.com)

# II. THE DATA SCIENCE WORKFLOW

---

## THE DATA SCIENCE WORKFLOW

---

### Dataists (Hilary Mason & friends)

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

---

## THE DATA SCIENCE WORKFLOW

---

### Dataists (Hilary Mason & friends)

- 1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
- 2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
- 3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
- 4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret
- 5. Interpret - “The purpose of computing is insight, not numbers”

---

## THE DATA SCIENCE WORKFLOW

---

Jeff Hammerbacher (Facebook, Cloudera)

- 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

---

## THE DATA SCIENCE WORKFLOW

---

Ben Fry

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

---

## THE DATA SCIENCE WORKFLOW

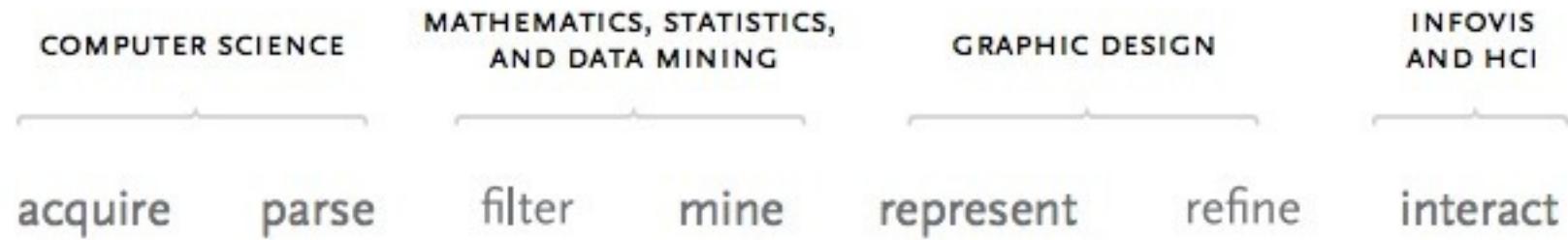
---

### Ben Fry

- 1. Acquire - the matter of obtaining the data
- 2. Parse - providing some structure around what the data means
- 3. Filter - removing all but the data of interest
- 4. Mine - the application of methods from statistics or data mining, as a way to discern patterns or place the data in mathematical context
- 5. Represent - determination of a simple representation (e.g. graphing)
- 6. Refine - improvements to the basic representation to make it clearer and more visually engaging
- 7. Interact - the addition of methods for manipulating the data or controlling which features are visible

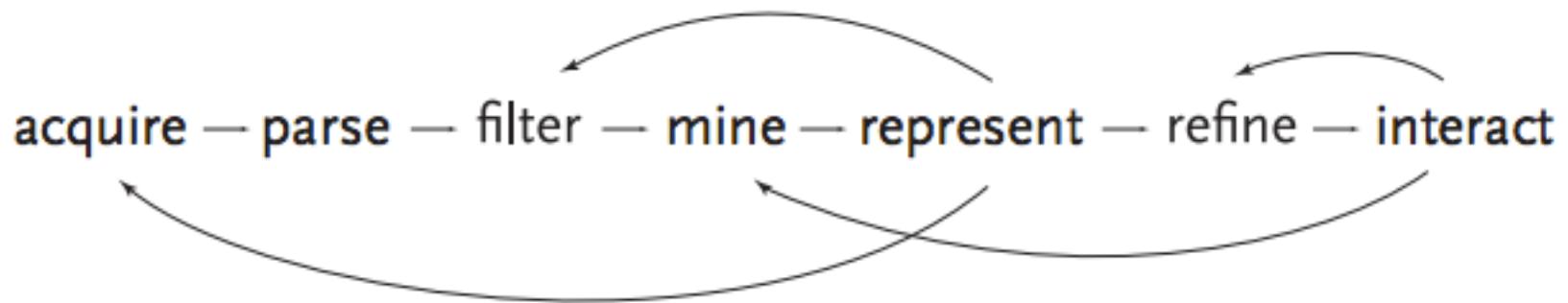
## THE DATA SCIENCE WORKFLOW

---



## THE DATA SCIENCE WORKFLOW

---



### NOTE

This diagram illustrates  
the *iterative* nature of  
problem solving

# THE DATA SCIENCE WORKFLOW

---

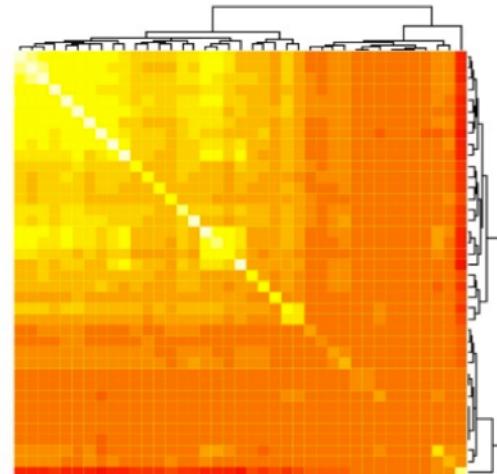
## What is needed most?

approximately **80% of the costs** for data-related projects gets spent on data preparation – mostly on **cleaning up** data quality issues: ETL, log files, etc., generally by socializing the problem

unfortunately, data-related budgets tend to go into frameworks that can only be used *after clean up*

most valuable skills:

- ▶ learn to use programmable tools that prepare data
- ▶ learn to understand the audience and their priorities
- ▶ learn to socialize the problems, knocking down silos
- ▶ learn to generate compelling **data visualizations**
- ▶ learn to estimate the confidence for reported results
- ▶ learn to automate work, making process repeatable



# THE DATA SCIENCE WORKFLOW

## Modeling

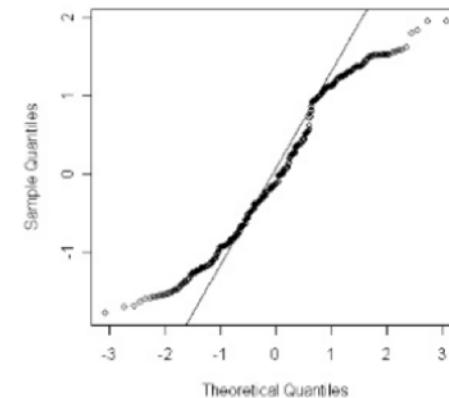
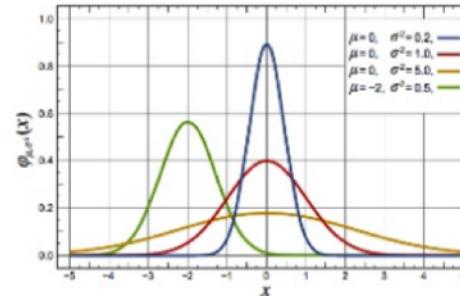
back in the day, we worked with practices based on  
**data modeling**

1. sample the data
2. fit the sample to a known distribution
3. ignore the rest of the data
4. infer, based on that fitted distribution

that served well with ONE computer, ONE analyst,  
ONE model... just throw away annoying "extra" data

circa late 1990s: machine data, aggregation, clusters, etc.  
**algorithmic modeling** displaced the prior practices  
of data modeling

*because the data won't fit on one computer anymore*



# THE DATA SCIENCE WORKFLOW

---

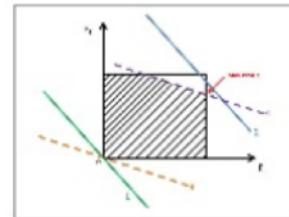
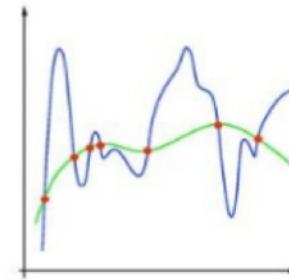
## Learning Theory

in general, apps alternate between learning patterns/rules and retrieving similar things...

**machine learning** – scalable, arguably quite ad-hoc, generally “black box” solutions, enabling you to make billion dollar mistakes, with oh so much commercial emphasis (i.e. the “heavy lifting”)

**statistics** – rigorous, much slower to evolve, confidence and rationale become transparent, preventing you from making billion dollar mistakes, any good commercial project has ample stats work used in QA (i.e., “CYA, cover your analysis”)

once Big Data projects get beyond merely digesting log files, **optimization** will likely become the next overused buzzword :)



# THE DATA SCIENCE WORKFLOW

---

## Generalizations about Machine Learning...

great introduction to ML, plus a proposed categorization for comparing different machine learning approaches:

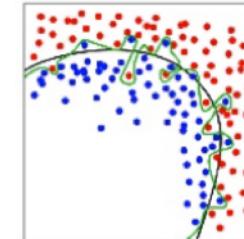
*A Few Useful Things to Know about Machine Learning*

**Pedro Domingos**, U Washington

[homes.cs.washington.edu/~pedrod/papers/cacm12.pdf](http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf)

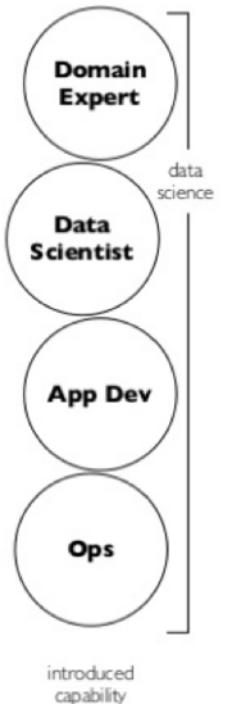
toward a categorization for Machine Learning algorithms:

- **representation**: classifier must be represented in some formal language that computers can handle (algorithms, data structures, etc.)
- **evaluation**: evaluation function (objective function, scoring function) is needed to distinguish good classifiers from bad ones
- **optimization**: method to search among the classifiers in the language for the highest-scoring one



# THE DATA SCIENCE WORKFLOW

## Team Composition = Roles



*business process,  
stakeholder*

*data prep, discovery,  
modeling, etc.*

*software engineering,  
automation*

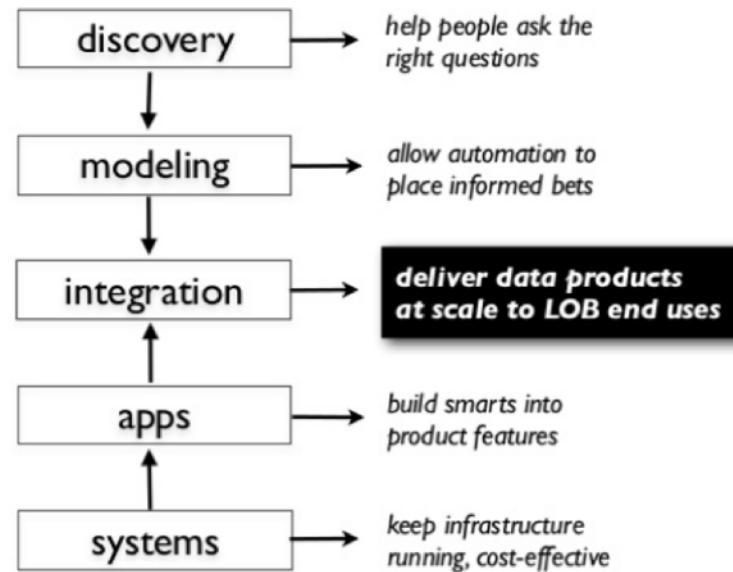
*systems engineering,  
availability*



**leverage non-traditional  
pairing among roles, to  
complement skills and  
tear down silos**

# THE DATA SCIENCE WORKFLOW

## Team Process = Needs



# THE DATA SCIENCE WORKFLOW

## Alternatively, Data Roles × Skill Sets



Harlan Harris, et al.

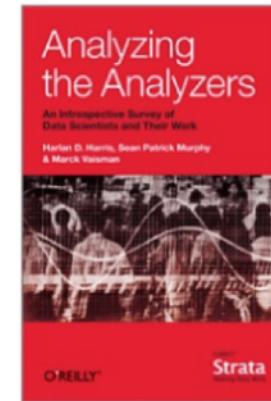
[datacommunitydc.org/blog/wp-content/uploads/2012/08/SkillsSelfIDMosaic-edit-500px.png](http://datacommunitydc.org/blog/wp-content/uploads/2012/08/SkillsSelfIDMosaic-edit-500px.png)

*Analyzing the Analyzers*

**Harlan Harris, Sean Murphy,  
Marck Vaisman**

O'Reilly, 2013

[amazon.com/dp/B00DBHTE56](http://amazon.com/dp/B00DBHTE56)



# THE DATA SCIENCE WORKFLOW

---

## Just Enough Mathematics?

having a solid background in **statistics** becomes vital,  
because it provides formalisms for what we're trying  
to accomplish at scale

along with that, some areas of math help – regardless  
of the “calculus threshold” invoked at many universities...

<b>linear algebra</b>	e.g., calculating algorithms for large-scale apps efficiently
<b>graph theory</b>	e.g., representation of problems in a calculable language
<b>abstract algebra</b>	e.g., probabilistic data structures in streaming analytics
<b>topology</b>	e.g., determining the underlying structure of the data
<b>operations research</b>	e.g., techniques for optimization ... in other words, ROI

# THE DATA SCIENCE WORKFLOW

---

## Trendlines

Big Data? we're just getting started:

- ~12 exabytes/day, jet turbines on commercial flights
- Google self-driving cars, ~1 Gb/s per vehicle
- National Instruments initiative: **Big Analog Data™**
- 1m resolution satellites **skyboximaging.com**
- open resource monitoring **reddmetrics.com**
- Sensing XChallenge **nokiasensingxchallenge.org**

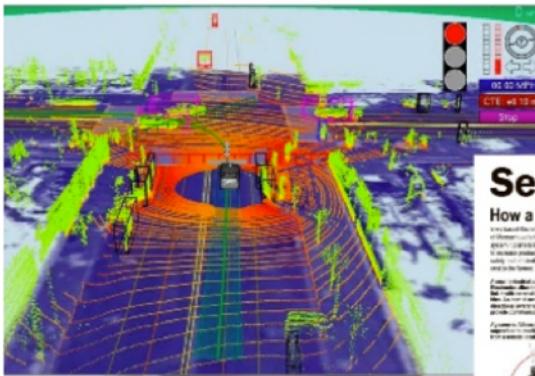
consider the implications of Jawbone, Nike, etc.,  
plus the effects of Google Glass...



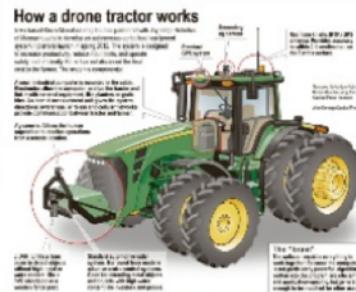
[technologyreview.com/...](http://technologyreview.com/)

# THE DATA SCIENCE WORKFLOW

## Internet of Things



## Send in the drones



# III. VISUALIZATIONS AS A MEDIUM

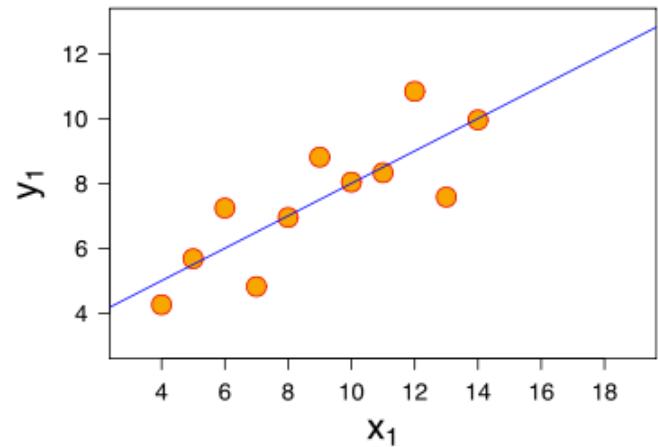
---

## EXERCISE – WHY VISUALIZE DATA?

---

*Consider the following dataset:*

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



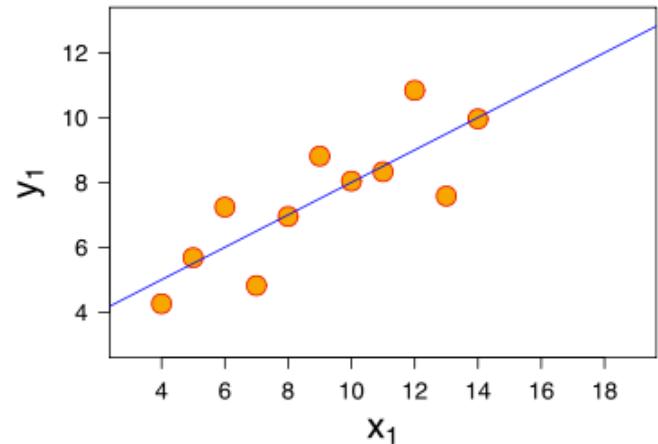
---

## EXERCISE – WHY VISUALIZE DATA?

---

*Consider the following dataset:*

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



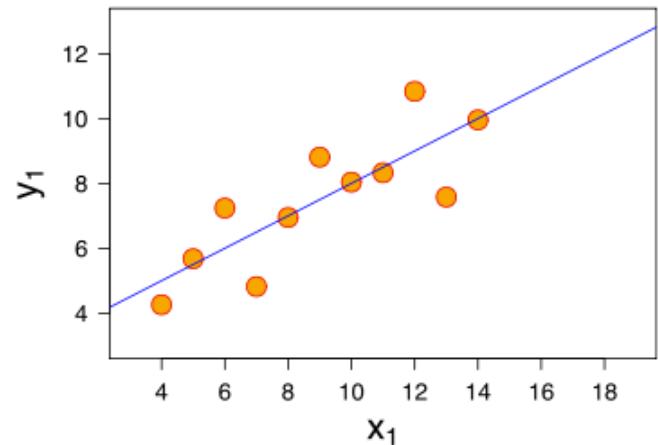
---

## 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*



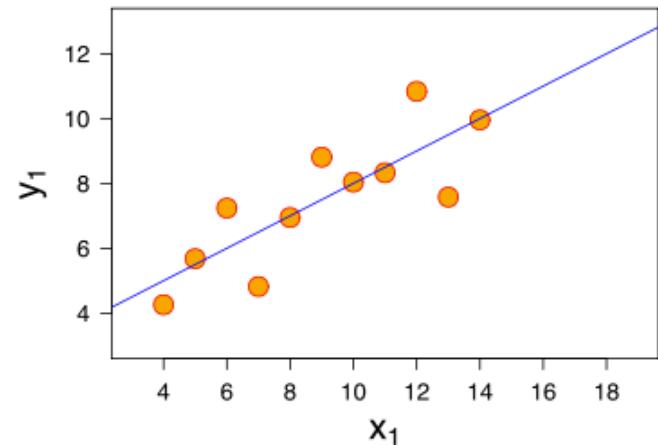
---

## 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  $y = 0.8$*



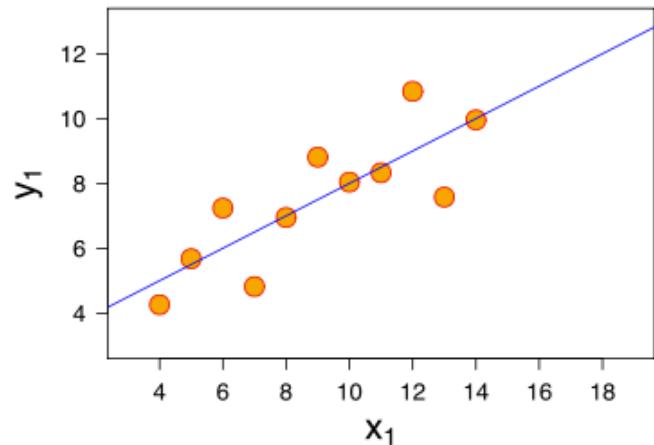
---

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



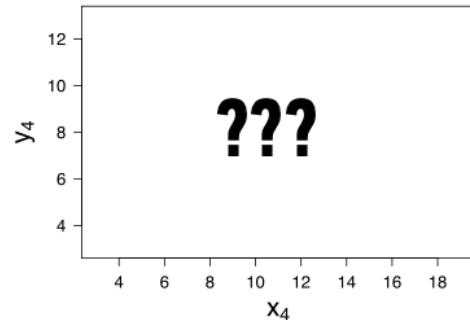
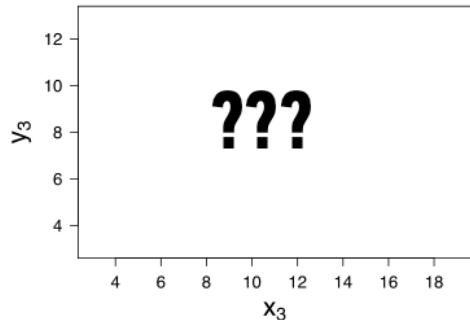
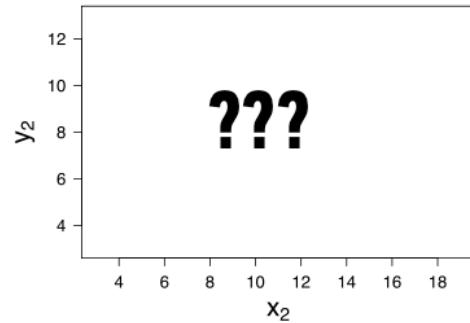
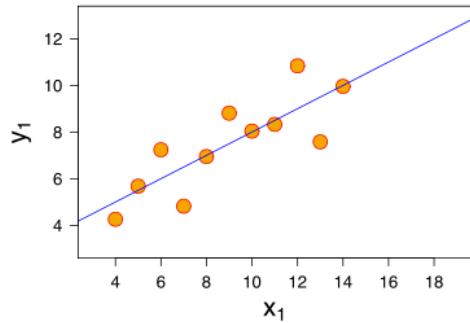
---

## EXERCISE – WHY VISUALIZE DATA?

---

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

*Q: how similar are these  
datasets?*

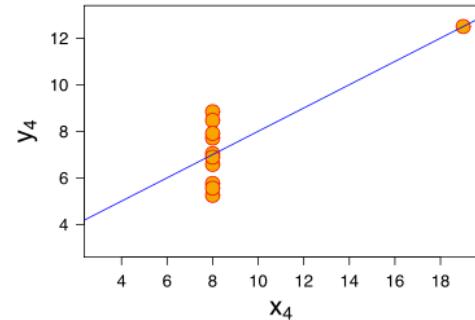
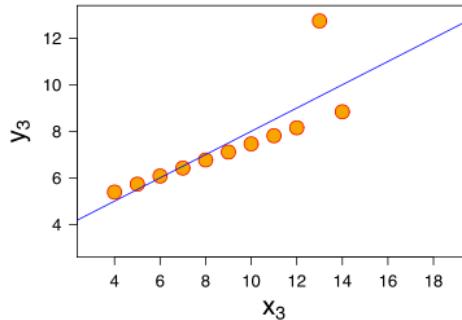
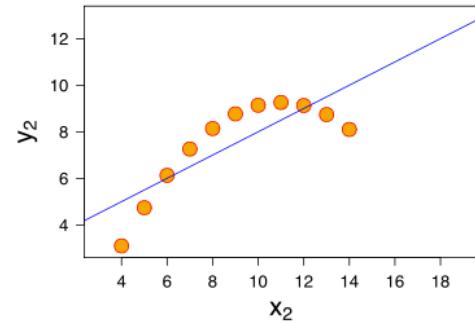
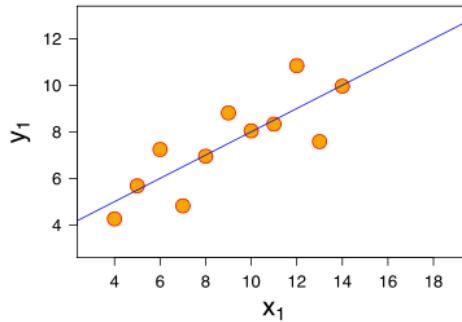


## EXERCISE – WHY VISUALIZE DATA?

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

*Q: how similar are these  
datasets?*

*A: not very!*



---

## **EXERCISE – WHY VISUALIZE DATA?**

---

*Plot your data!*

# IV. 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!

# V. INTRO TO I-PYTHON

---

INTRO TO DATA SCIENCE

---

# DISCUSSION