GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1

OCTOBER 1, 2014 // DAT10 SF

INTRO TO DATA SCIENCE

WELCOME!

AGENDA

- Resources
- Instructors introductions
- Students introductions
- Lecture: Introduction to Data Science
- Tutorial: Introduction to iPython
- Q&A

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
  - datatau.com

COURSE EXPECTATIONS:

- BE PRESENT
- PARTICIPATE
- DO THE ASSIGNMENTS
- COLLABORATE
- MAKE FRIENDS

HELLO!

FRANCESCO MOSCONI

- Chief Data Officer at Spire, a company that invented the first consumer wearable device capable of continuously tracking respiration and activity. I 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.

SPIRE Roche

SINGULARITY UNIVERSITY socialbakers

GENERAL ASSEMBLY

HELLO!

CRAIG SAKUMA

- Craig is a data scientist at Euclid Analytics where he applies machine learning algorithms to measure customer traffic at brick and mortar stores. Prior to working as a data scientist, Craig has been a founder of a start-up and a corporate strategy consultant. He has an MBA from Wharton and an engineering degree from Northwestern University.

GENERAL ASSEMBLY

euclid

HELLO!

MICHAEL KEBA

- After college and two years as a business analyst at a consulting firm in Silicon Valley, Michael joined a venture backed video startup in San Francisco. Here he was exposed to the world of advertising technology and was fascinated by the huge volume of data generated selling ad impressions, and how valuable good analysis of this data could be. To keep pace with the tools used by the industry, Michael taught himself Python and enrolled in General Assembly's Data Science course.
- In his current role at Tightrope Interactive, Michael is involved in testing, analysis, and optimization of new features for the company's ad tech products. Michael received his bachelor's degree in Managerial Economics with a minor in Biological Sciences from the University of California, Davis.

GENERAL ASSEMBLY

TIGHTROPE INTERACTIVE

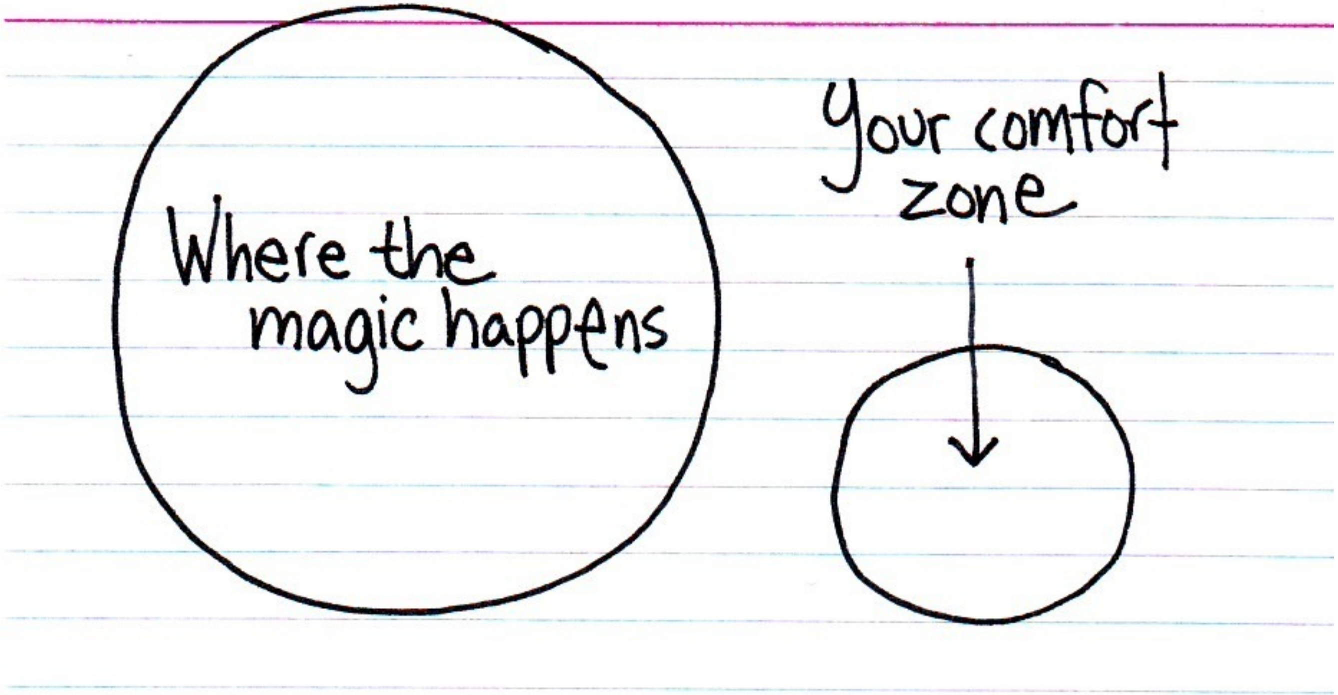
INTRODUCTIONS

- Fill in your card
- Exchange with other person
- Read card :-)

Source:

CONGRATULATIONS…

...for getting out of your comfort zone!

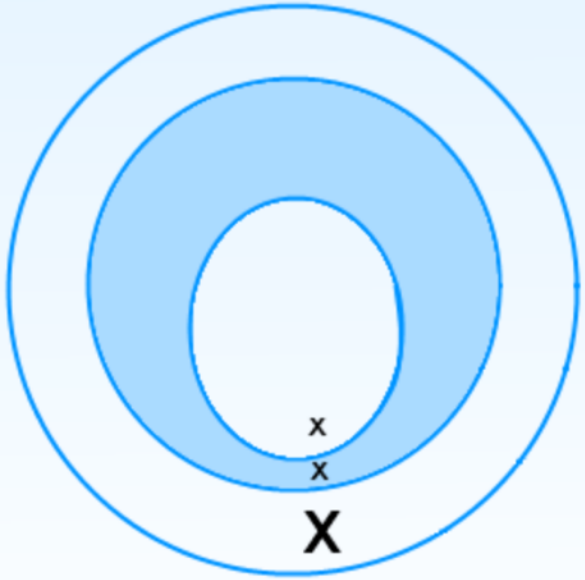


How to Grow Your Comfort Zone

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.



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?

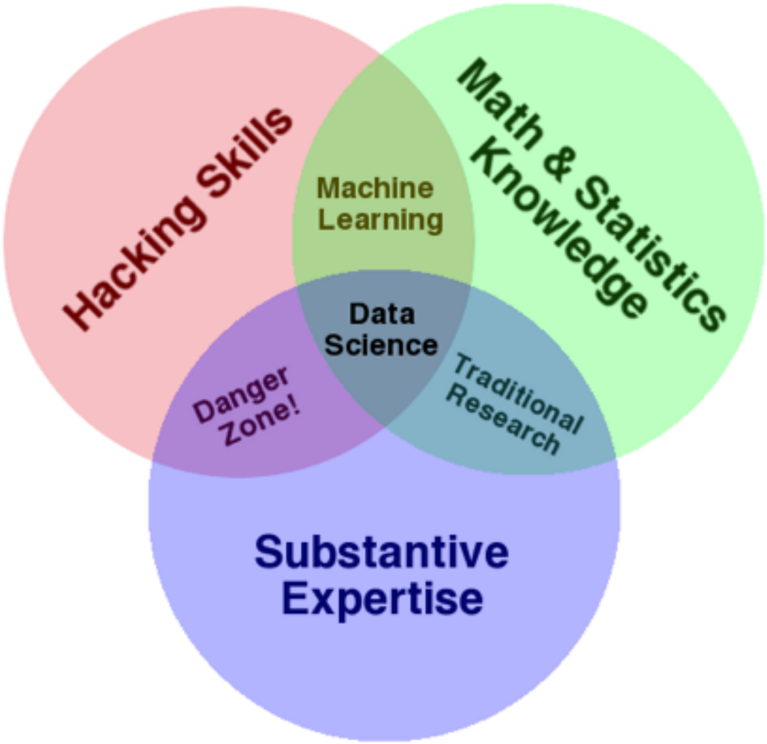
WHAT IS DATA SCIENCE?

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

WHAT IS DATA SCIENCE?

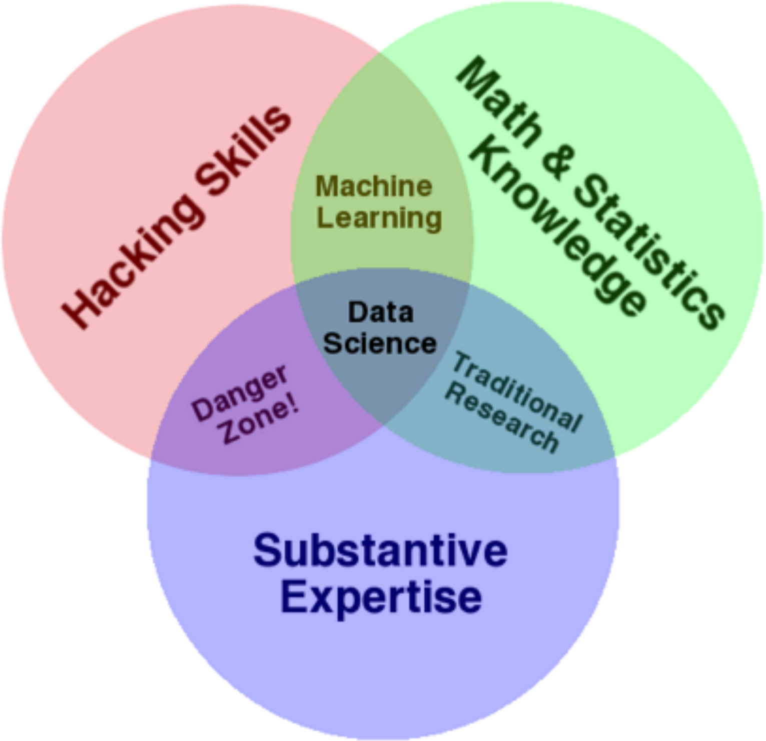
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST



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

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

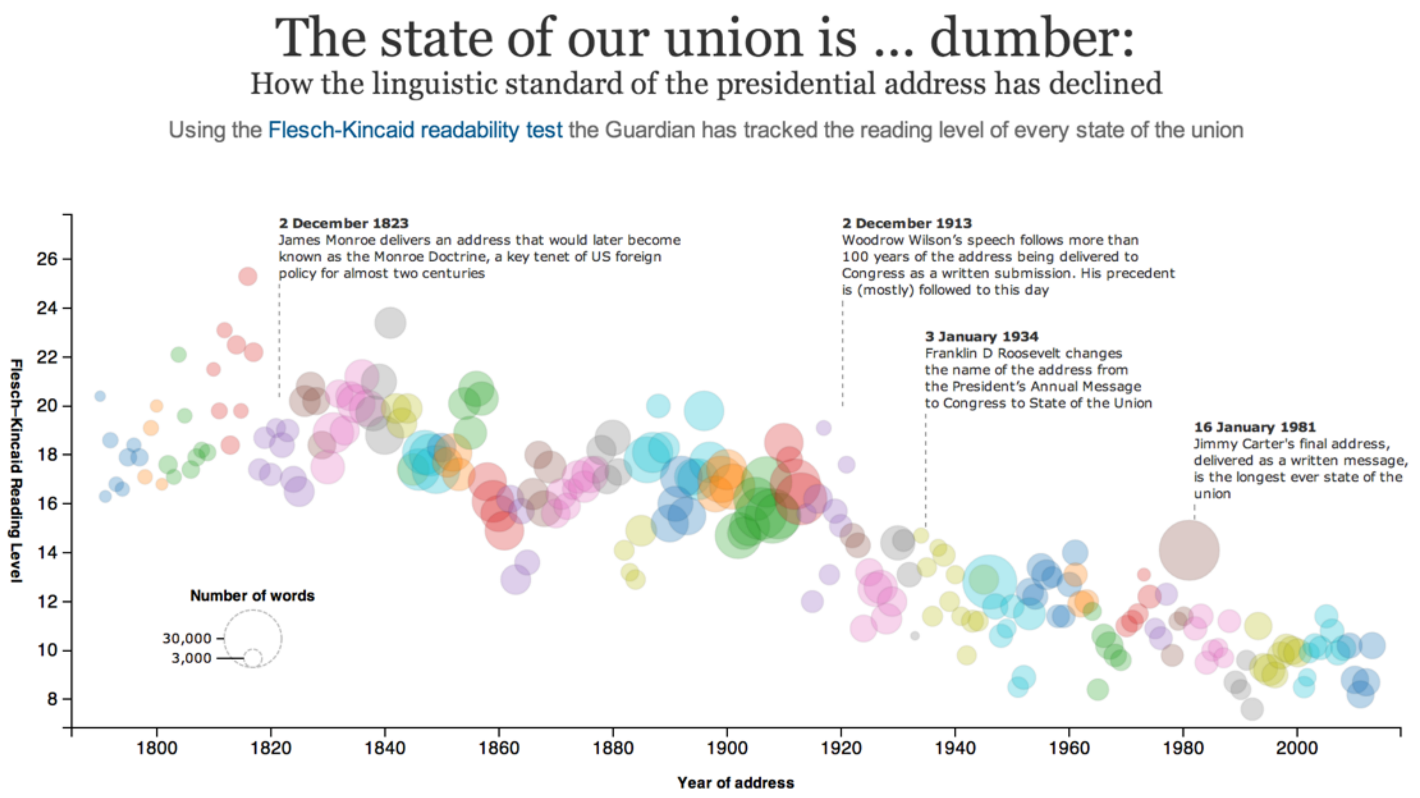
Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

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?



source: http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level

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

Dataists (Hilary Mason & friends)

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

source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

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”

source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

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

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

THE DATA SCIENCE WORKFLOW

Ben Fry

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

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

Ben Fry

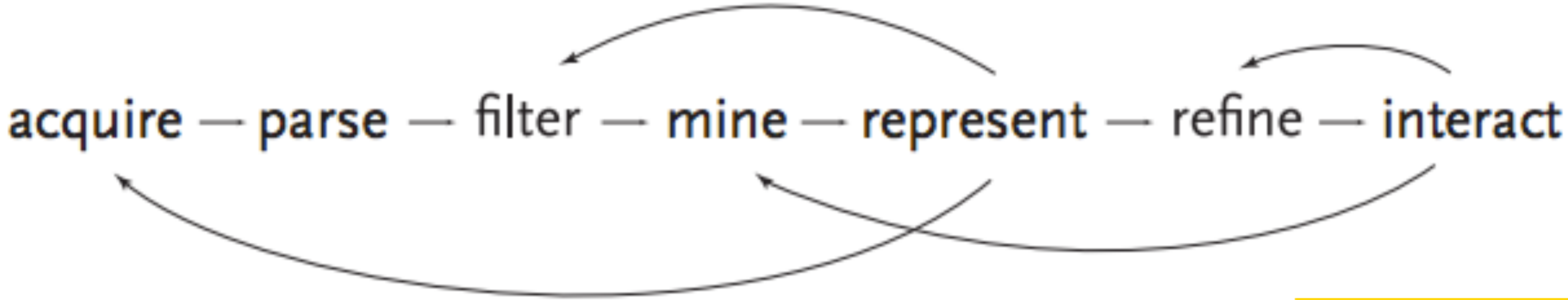
- 1. Acquire - the matter of obtaining the data
- 2. Parse - providing some structure around what the data means
- 3. Filter - removing all but the data of interest
- 4. Mine - the application of methods from statistics or data mining, as a way to discern patterns or place the data in mathematical context
- 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

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

THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW



NOTE

This diagram illustrates the *iterative* nature of problem solving

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

THE DATA SCIENCE WORKFLOW

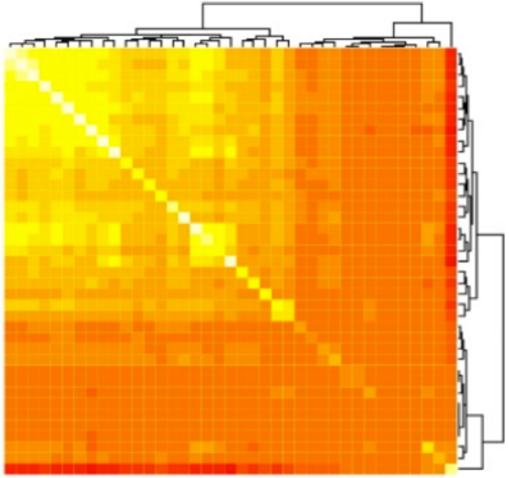
What is needed most?

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*



Modeling

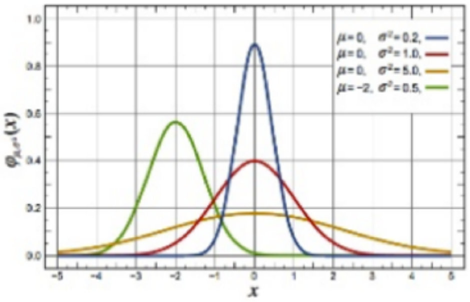
back in the day, we worked with practices based on **data modeling**

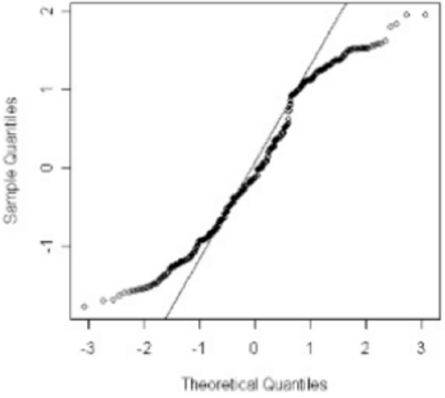
1. *sample the data*
2. *fit the sample to a known distribution*
3. *ignore the rest of the data*
4. *infer, based on that fitted distribution*

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

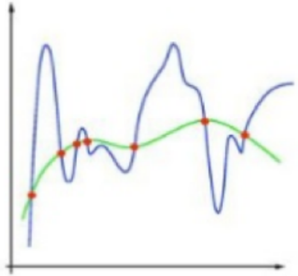
because the data won't fit on one computer anymore





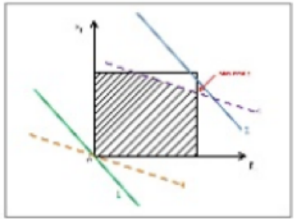
Learning Theory

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 :)

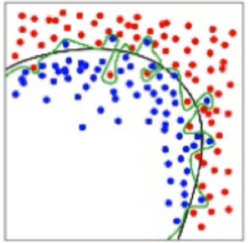
Generalizations about Machine Learning...

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

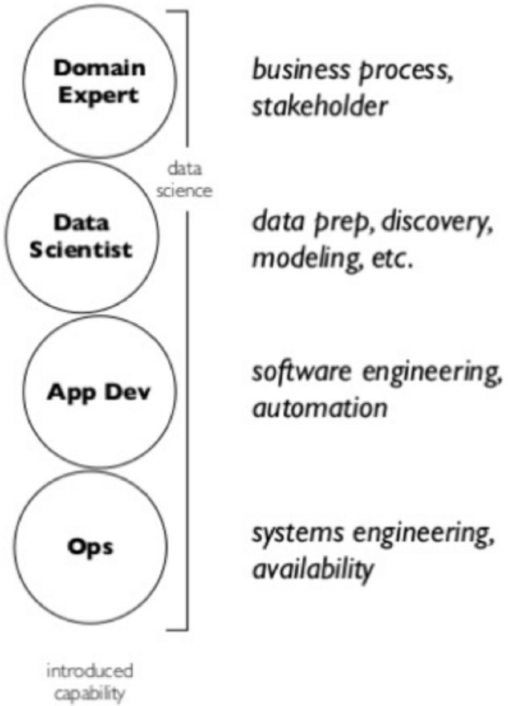
toward a categorization for Machine Learning algorithms:

- **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

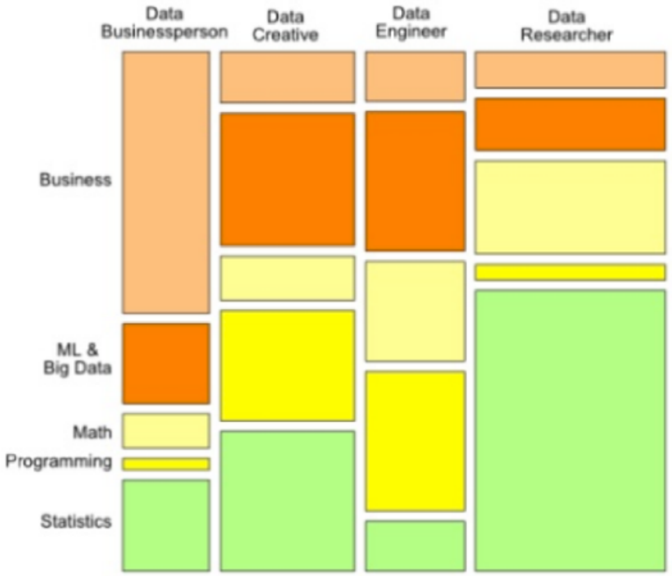


leverage non-traditional pairing among roles, to complement skills and tear down silos

THE DATA SCIENCE WORKFLOW

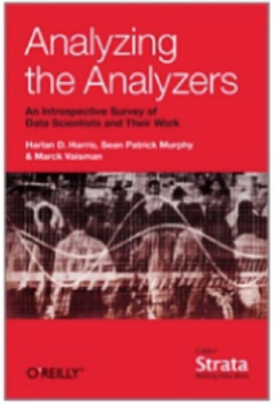
Team Process = Needs

Alternatively, Data Roles × Skill Sets



Harlan Harris, et al.
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



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…

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

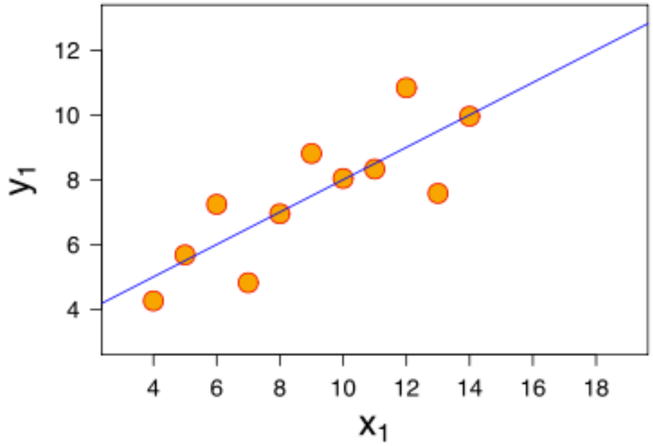
THE DATA SCIENCE WORKFLOW

III. VISUALIZATIONS AS A MEDIUM

EXERCISE – WHY VISUALIZE DATA?

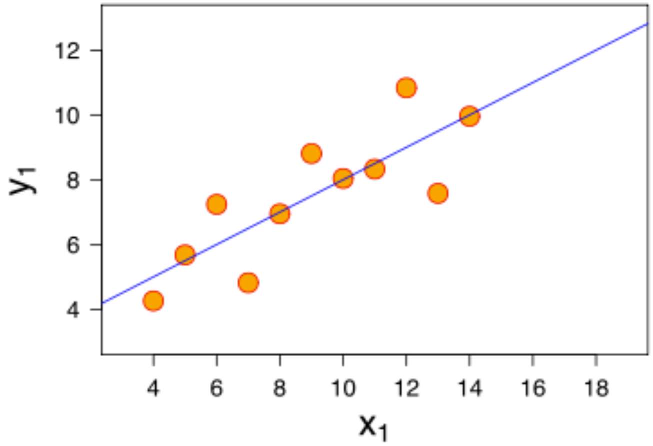
Consider the following dataset:

- *eleven (x, y) points*



Consider the following dataset:

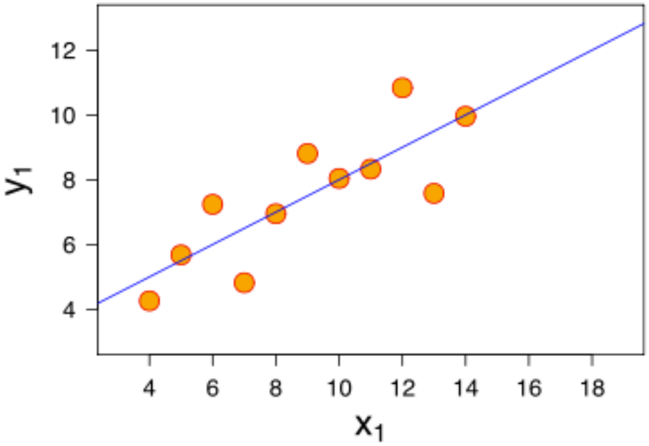
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

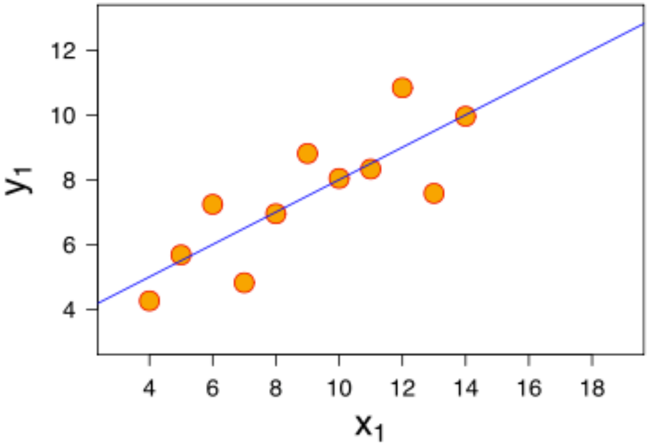
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*



EXERCISE – WHY VISUALIZE DATA?

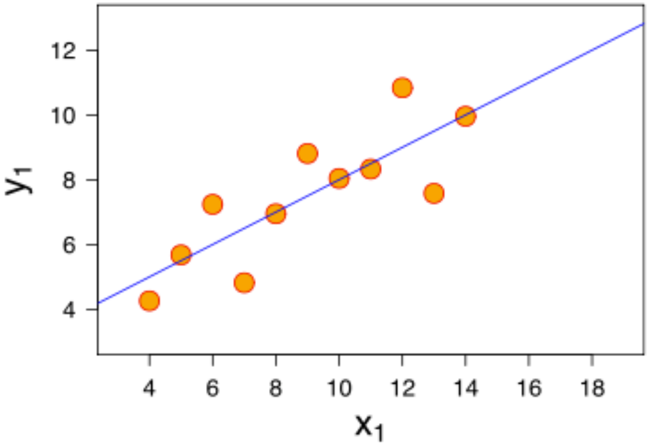
Consider the following dataset:

- *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*



Consider the following dataset:

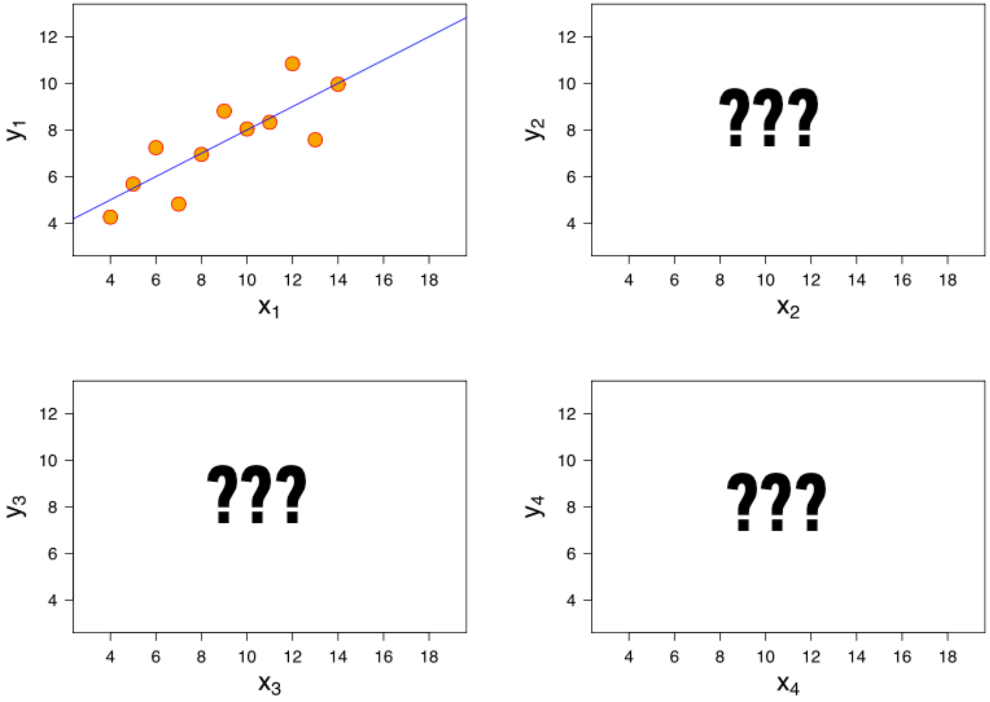
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?



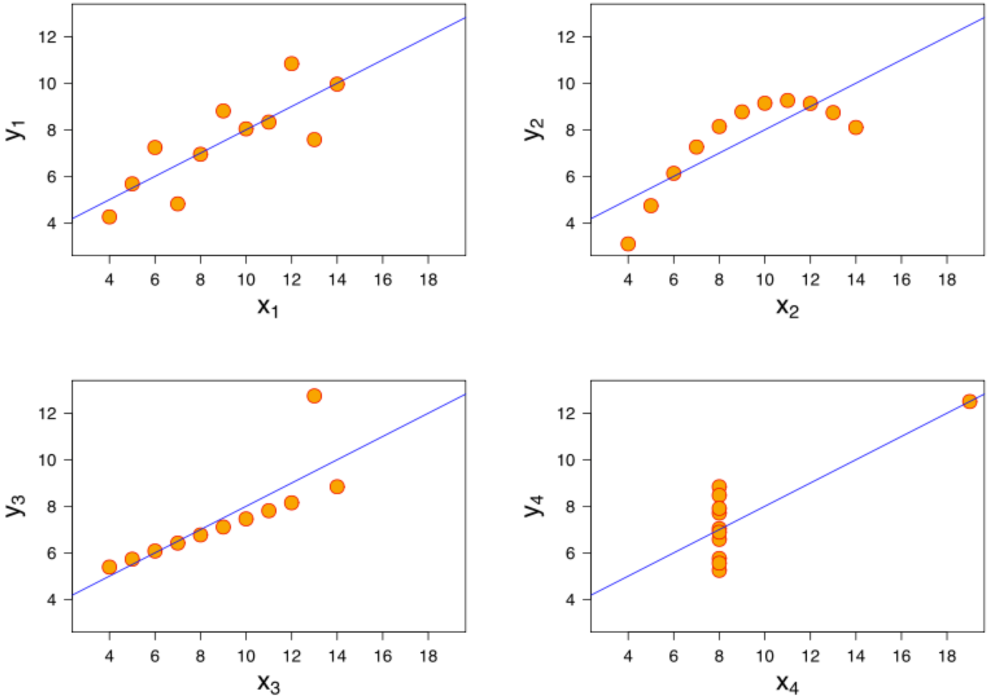
EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?

A: not very!

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



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

DISCUSSION