INTRO TO DATA SCIENCE:

WELCOME!

LOGISTICS

DATES: MAY 31ST – AUG 16, 2014

TIMES: SATURDAY 10:00AM – 5:00PM

LOCATION: 1062 FOLSOM STREET

INSTRUCTOR: MIKE TAMIR – MNTAMIR@GMAIL.COM

EXPERTS-IN-RESIDENCE (TA'S):

FRANK TAYLOR - FRANCTAYLOR.FT@GMAIL.COM

ZACK DESARIO - ZACHARYDESARIO@GMAIL.COM

Office Hours: TBD after break

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

Course expectations:

BE PRESENT

PARTICIPATE

DO THE ASSIGNMENTS

MAKE FRIENDS

Meet Your Course Producer

Alex Hamady

- Education Programs Producer, DS & DGM
- alexh@generalassemb.ly

GENERAL ASSEMBLY

Mike Tamir

Mike is the Chief Data Scientist for PersonaGraph, where he leads the Data Science and Machine Learning Operations teams, building the inference models populating the PersonaGraph user-understanding platform. Prior to PersonaGraph, Mike served as Director of Data Sciences for Sears Holdings. He began his career in academia teaching at the University of Pittsburgh and serving as a mathematics teaching fellow for Columbia University. His research focused on developing the epsilon-anchor methodology for resolving both an inconsistency he highlighted in the dynamics of Einstein's general relativity theory and the convergence of "large N" Monte Carlo simulations in Statistical Mechanics' universality models of criticality phenomena.

Frank

Frank is passionate about Big Data and its potential to gather insight into so many facets of humanity. As our tools get better and more scalable, we have the ability to answer greater questions and build more meaningful products that enrich our lives and solve problems.

Frank's first experience munging big data was modeling particle decays after heavy-ion collisions in a particle detector at LBL and UC – Berkeley. After learning flamenco guitar in Spain, I worked on data acquisition and signal analysis as an optical engineer in the Bay Area. More recently he graduated from Zipfian Academy: a 12-week all-day bootcamp diving into the latest and greatest tools for Data Science.

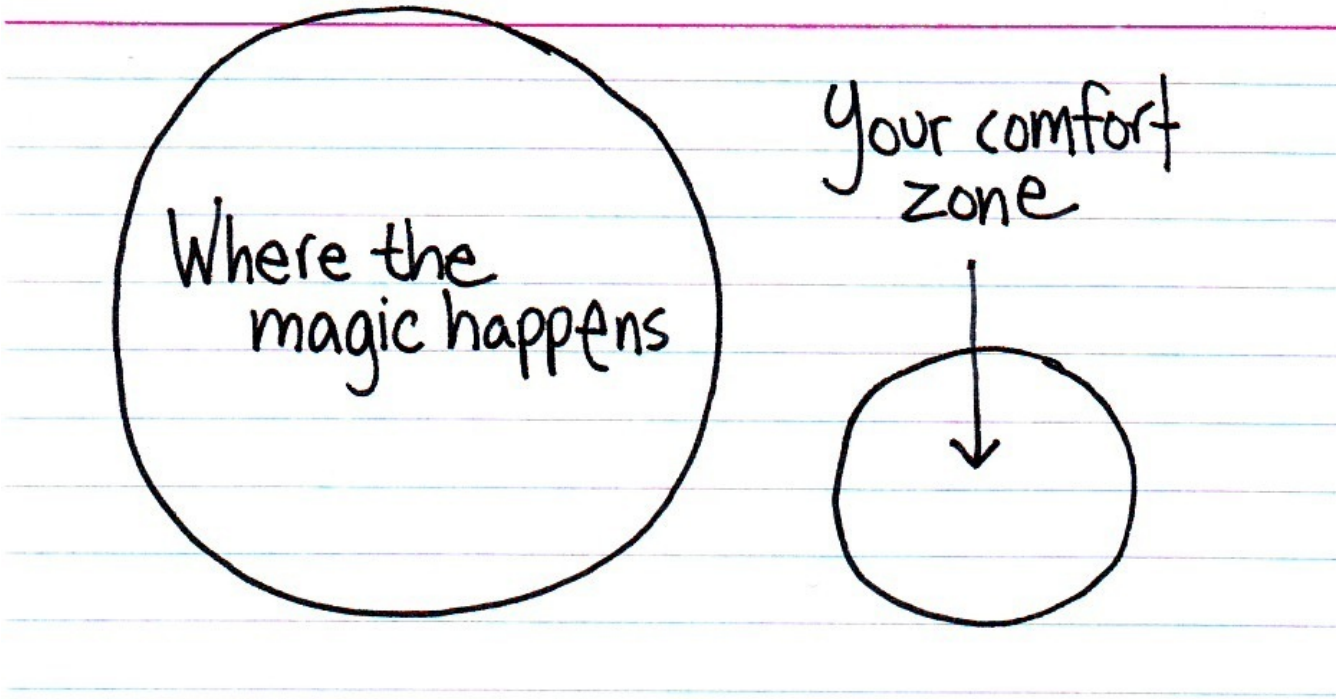
MEET YOUR INSTRUCTORS

Zack DeSario

Zack is a psychology-degree-holding, graphic-designing, professional-online-poker- playing, data designer.
Using data mining and tuned algorithms, he became one of the best online poker players in the world (technically #5). His relationship with Photoshop started before they invented layers. The experience he gained telling stories through visual elements, combined knowing how data describes behavior allow me to communicate complex information in a visually meaningful way. Zack makes complex information simple.

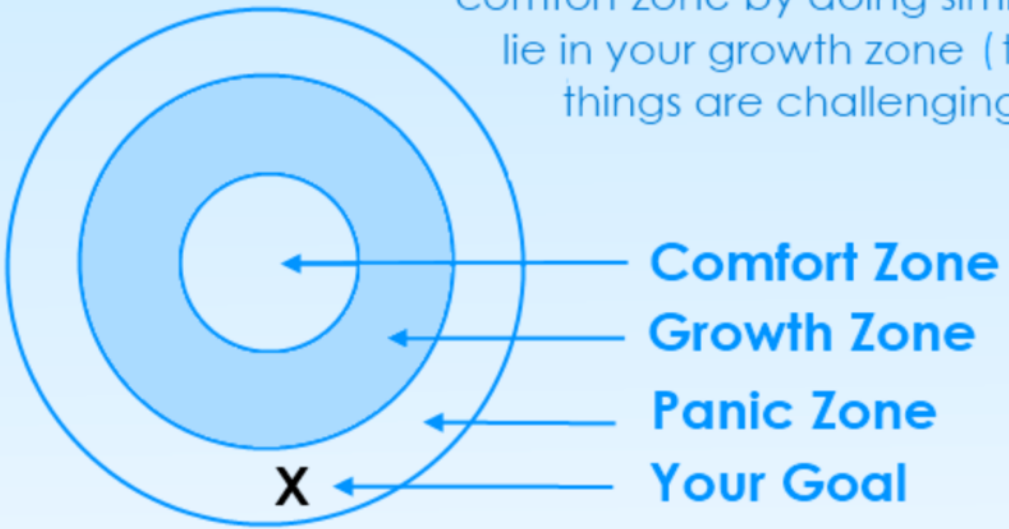
CONGRATULATIONS...

...for getting out of your comfort zone!



How to Grow Your Comfort Zone

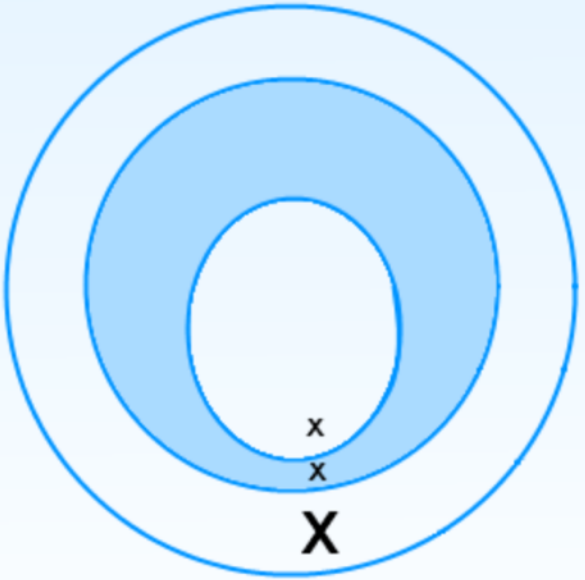
Any goal or challenge may fall into one of three zones - your comfort zone, growth zone, or panic zone. If your goal is currently in your panic zone, i.e. it would be too scary to do now, you will need to grow your comfort zone by doing similar challenges that lie in your growth zone (the zone in which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

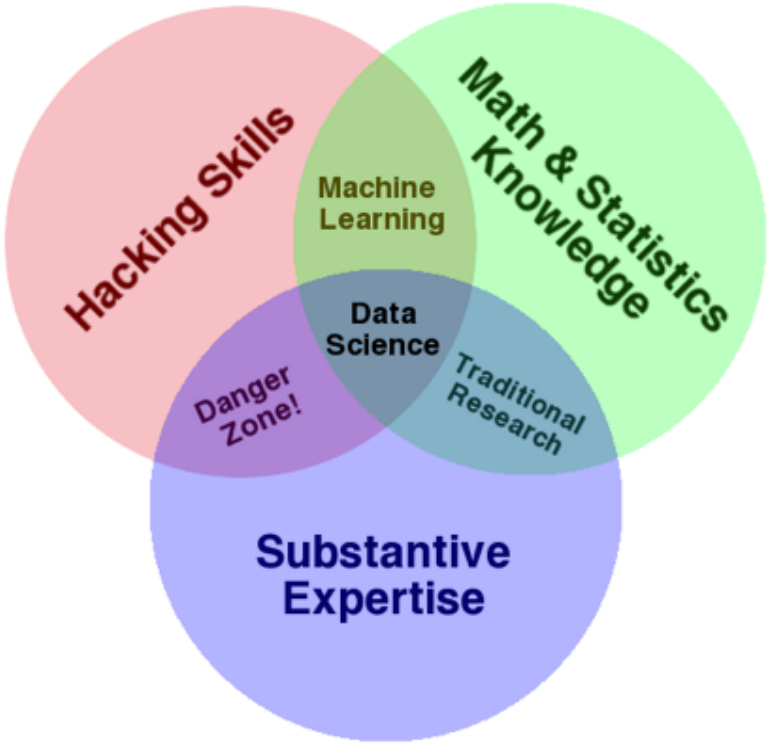
WHAT IS DATA SCIENCE?

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

WHAT IS DATA SCIENCE?

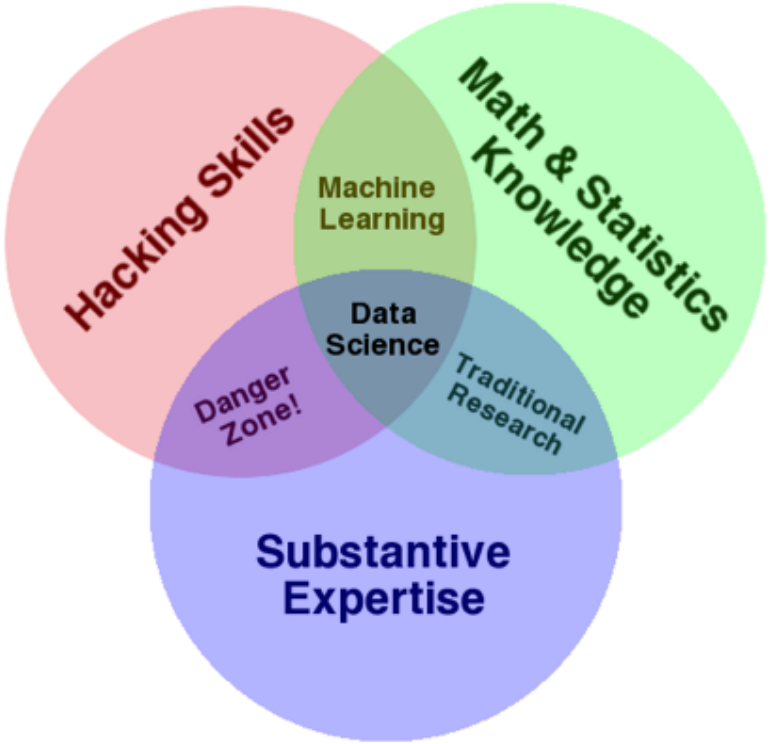
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

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

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

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

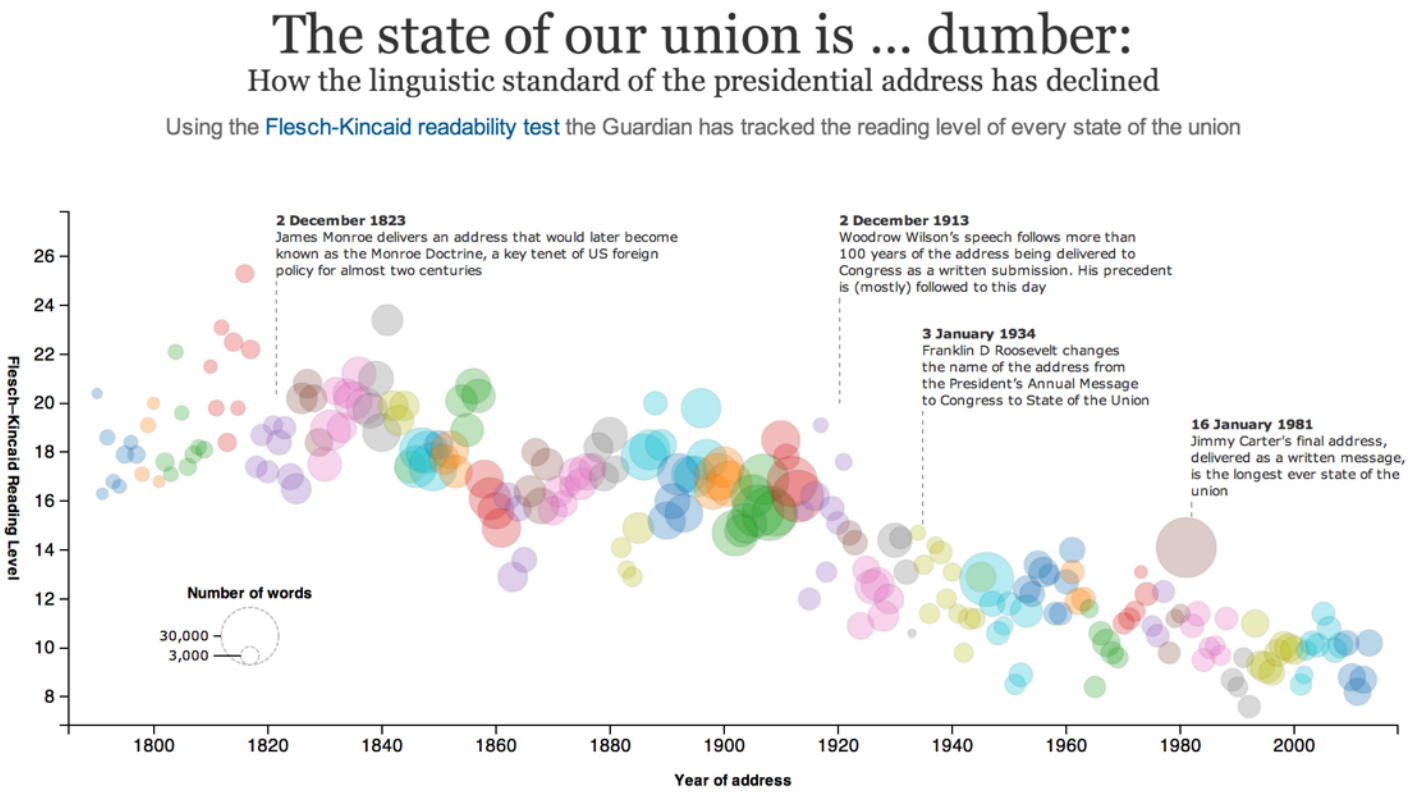
Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

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

II. THE DATA SCIENCE WORKFLOW

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/

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

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

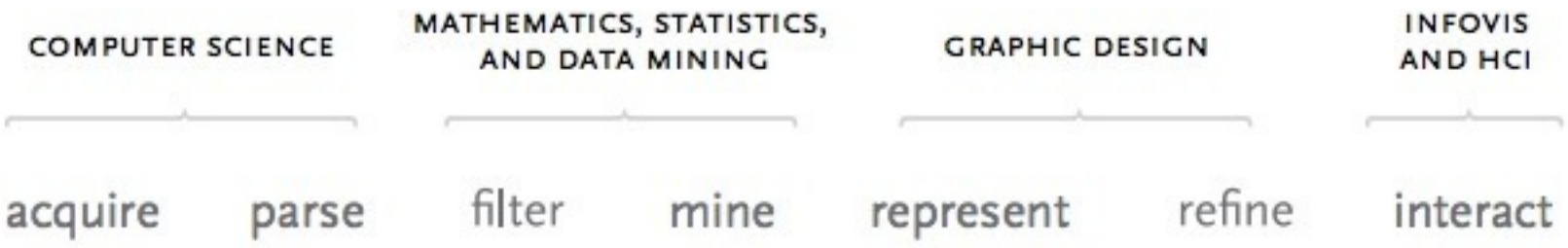
THE DATA SCIENCE WORKFLOW

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

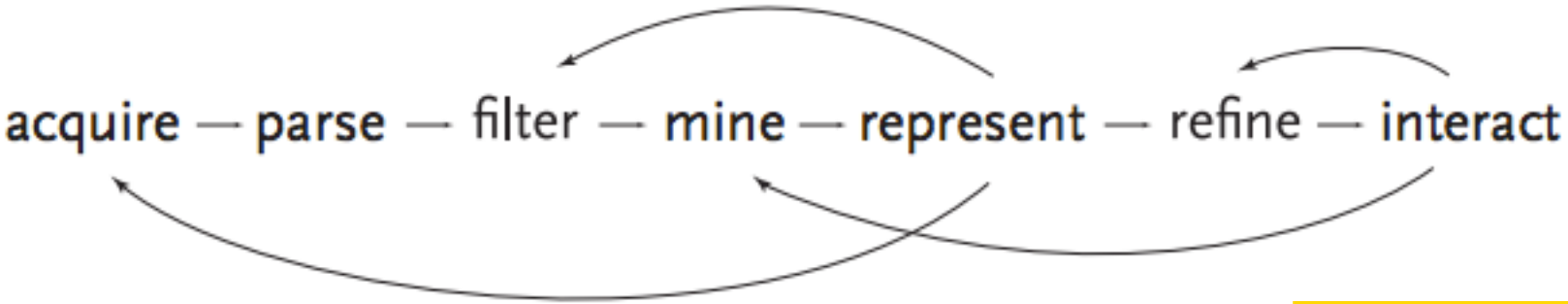
source: http://benfry.com/phd/dissertation-110323c.pdf

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

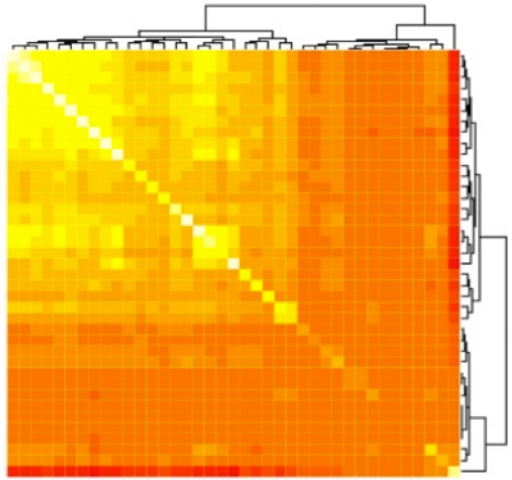
What is needed most?

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*



THE DATA SCIENCE WORKFLOW

Modeling

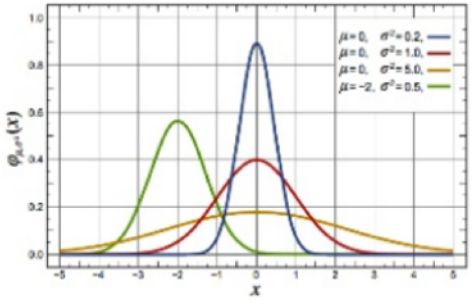
back in the day, we worked with practices based on **data modeling**

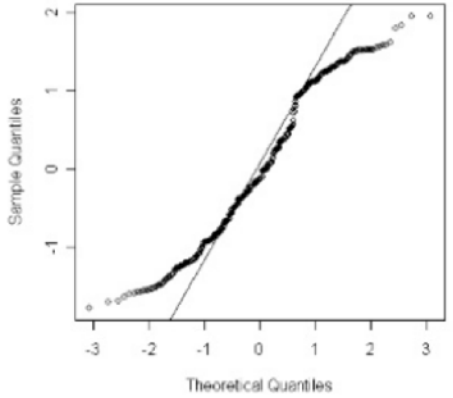
1. *sample the data*
2. *fit the sample to a known distribution*
3. *ignore the rest of the data*
4. *infer, based on that fitted distribution*

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



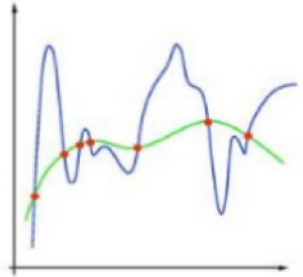


THE DATA SCIENCE WORKFLOW

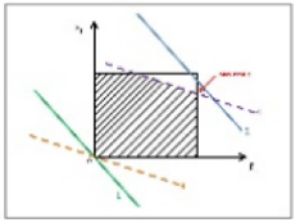
Learning Theory

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

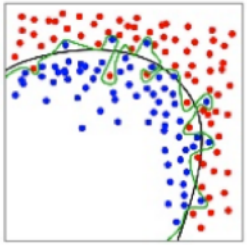
Generalizations about Machine Learning...

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

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



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…

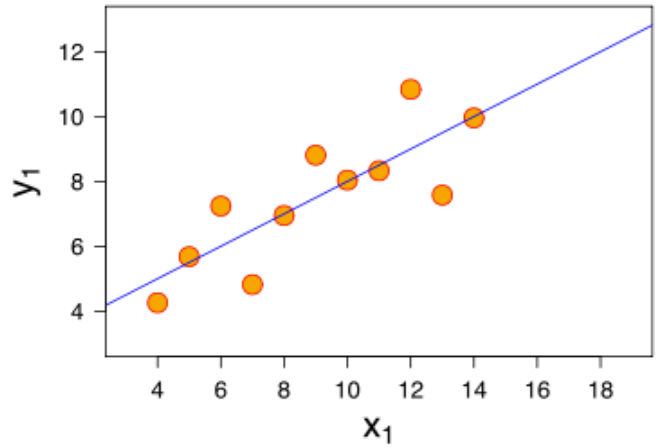
linear algebra	e.g., calculating algorithms for large-scale apps efficiently
graph theory	e.g., representation of problems in a calculable language
abstract algebra	e.g., probabilistic data structures in streaming analytics
topology	e.g., determining the underlying structure of the data
operations research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

III. VISUALIZATIONS AS A MEDIUM

EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

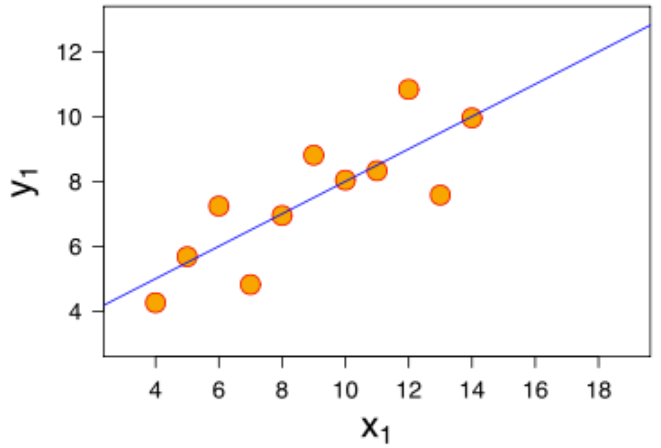
- *eleven (x, y) points*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

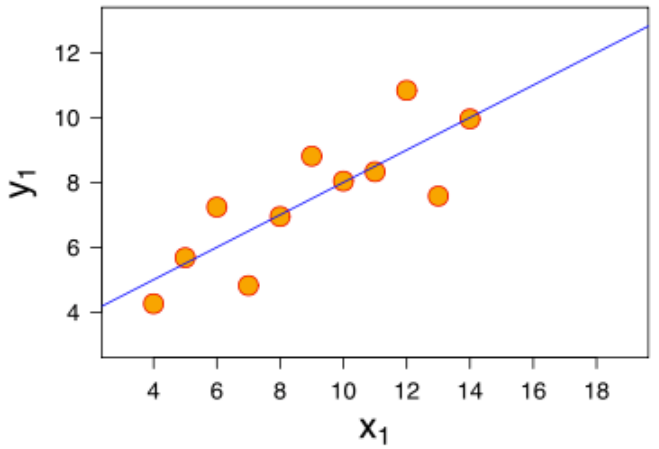
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

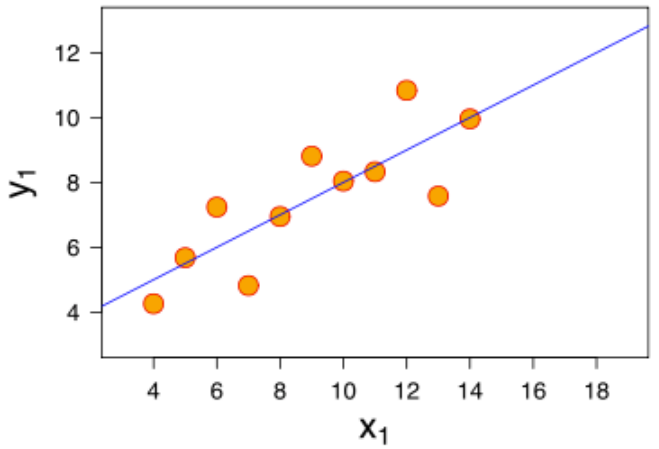
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

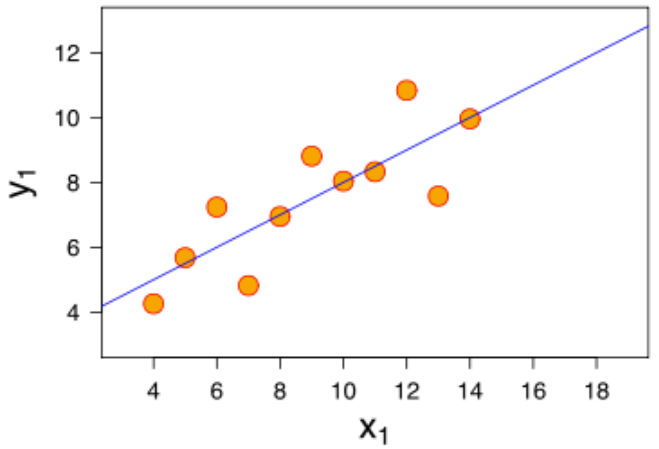
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x and y = 0.8*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

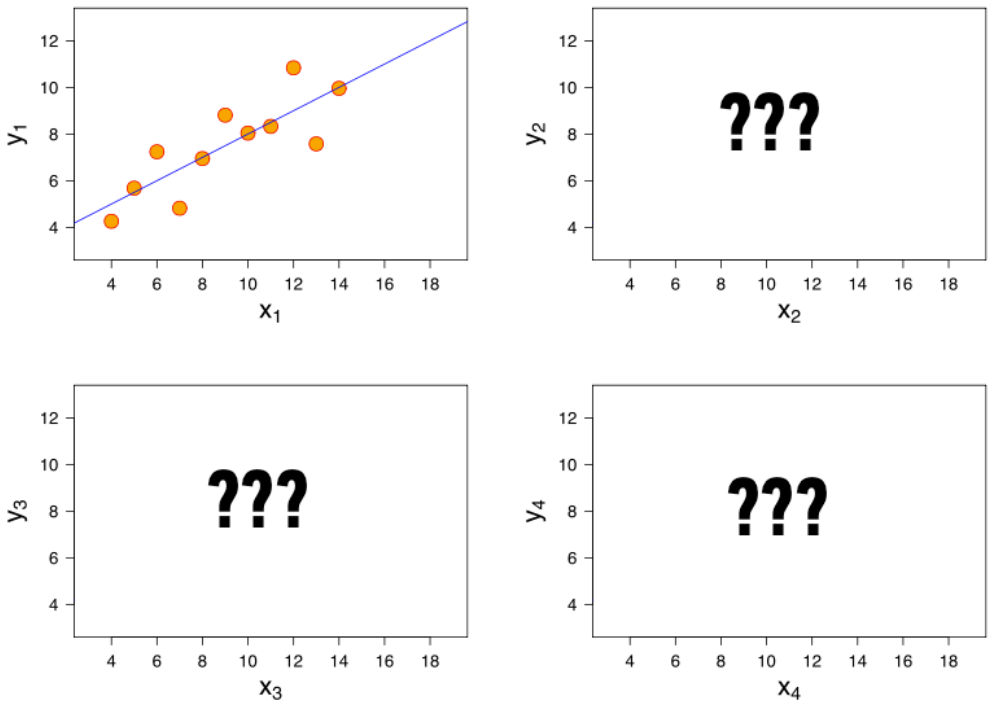
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?

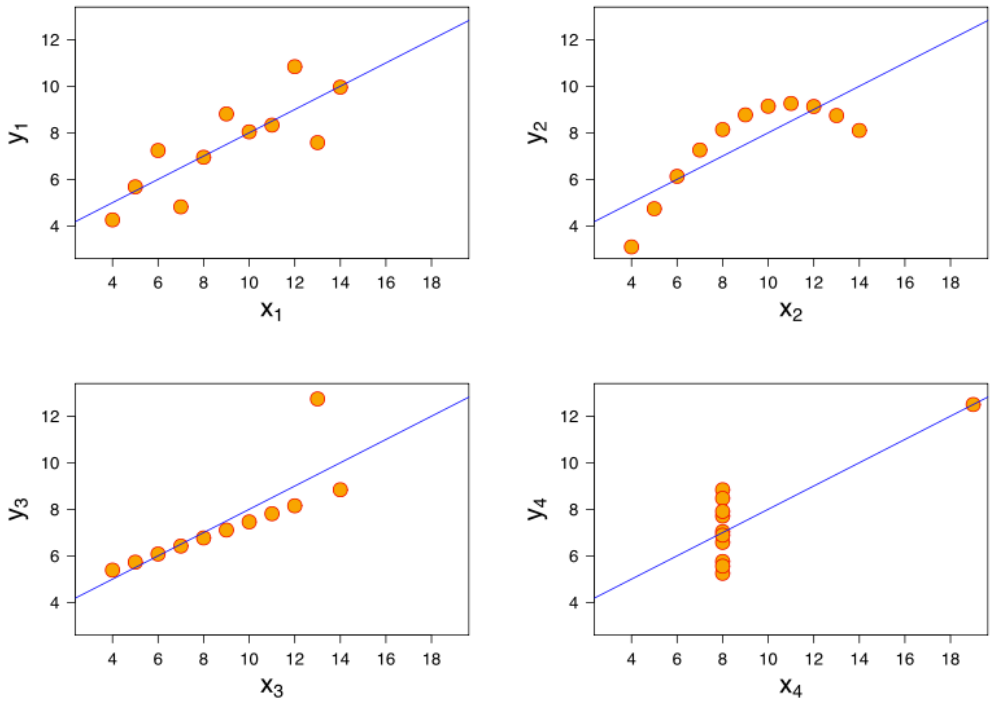


EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?

A: not very!



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