GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: DATA EXPLORATION

I. WHAT IS DATA SCIENCE?
II. THE DATA MINING WORKFLOW

LAB:
III. VISUALIZING DATA WITH R & GGPLOT2

I. WHAT IS DATA SCIENCE?

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

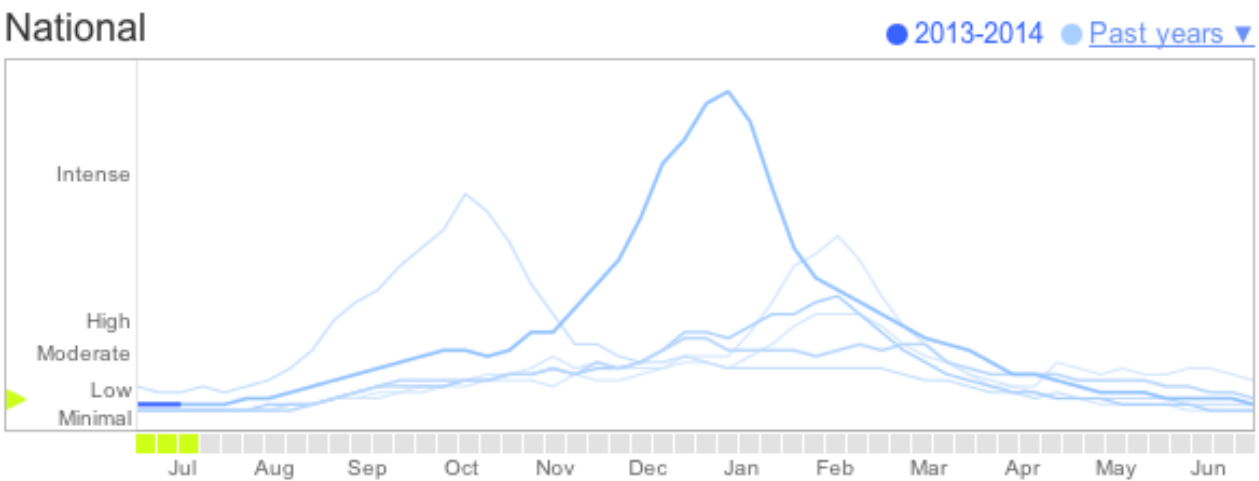
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.

- 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.
- *“Data science, as it’s practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics*” – Michael Driscoll, Founder@metamarkets

Google FLU TRENDS*

Detecting flu trends and influenza epidemics using search engine query data



50 million search queries to **45** candidate queries

450 million models to identify the candidate queries !!!!

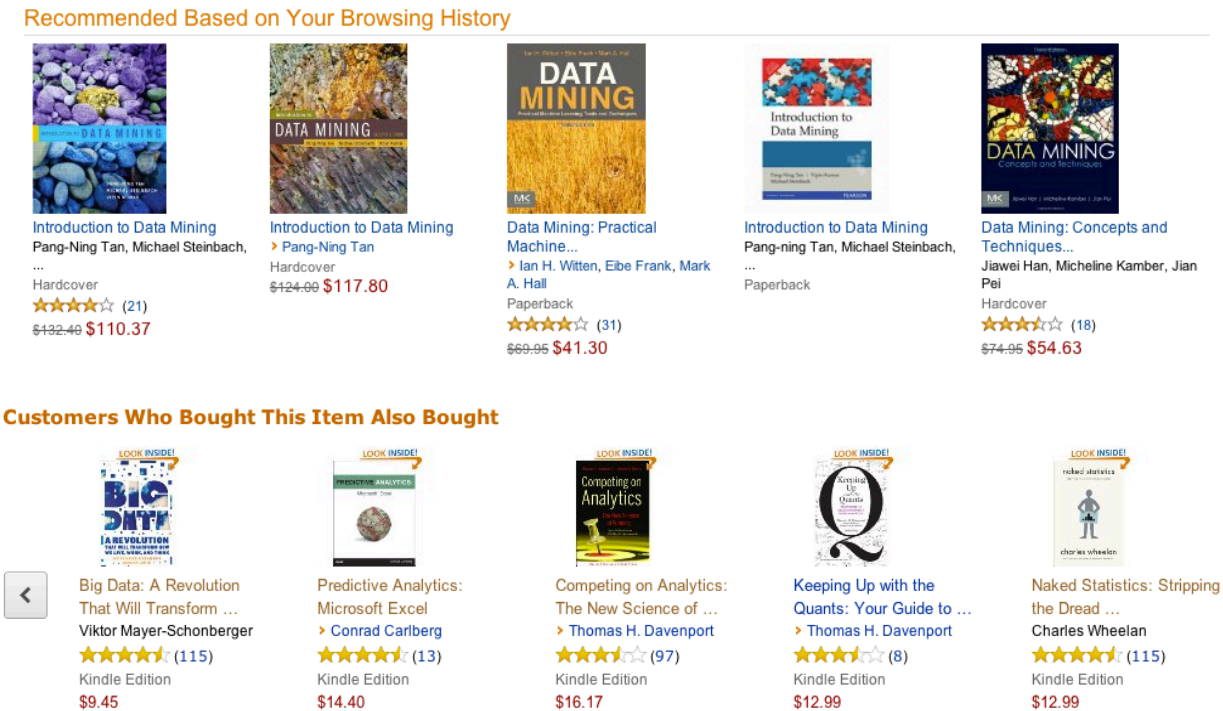
Validation of the model

Complex !!!

2012-2013 ???

* Nature Vol 457, 19 February 2009 - http://dx.doi.org/10.1038/nature07634

AMAZON.COM RECOMMENDATIONS*



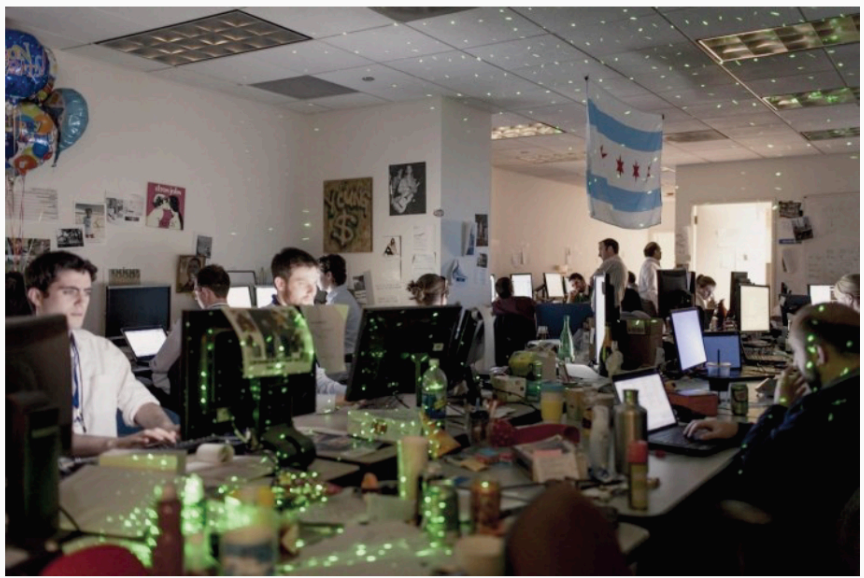
- Several million catalog items and millions of customers
- Browsing history
- Purchase history

Item-based collaborating filtering

Business Goals: click-through, conversions, revenue impact

* http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

INSIDE THE CAVE: Obama's Digital Campaign*



CREDIT: *TIME*

THE CAVE

The campaign's data & technology operations made up an estimated 30-40% of headquarters staff

Analytics to improve every aspect of the campaign - Web, Email, Field, and Communications

- A/B Testing for best Email subject
- Dynamic modeling and 'persuadability' score
- Mobile app optimization – Donate by single click like Amazon

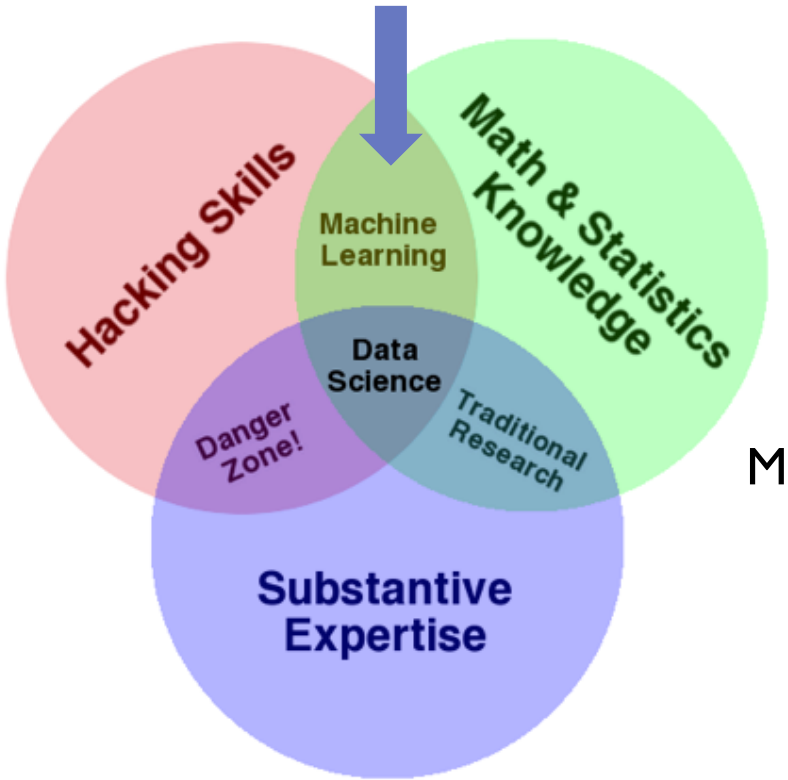
Other Internet and e-commerce Companies

* http://engagedc.com/download/Inside%20the%20Cave.pdf

SKILLS NEEDED FOR A DATA SCIENTIST ?

- Manipulating text files
- Basic programming
- Thinking algorithmically
- Basic Statistics
- Modeling
- Intuition
- Problem solving skills
- Business knowledge
- Understanding vector operations (Linear Algebra !!!)

THE QUALITIES OF A DATA SCIENTIST

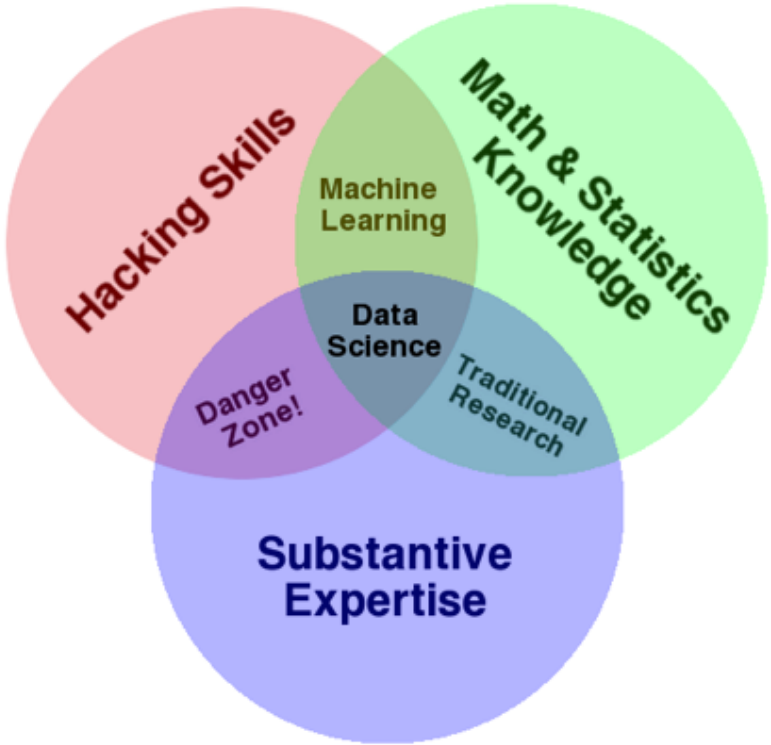


By Drew Conway source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

Machine Learning Examples

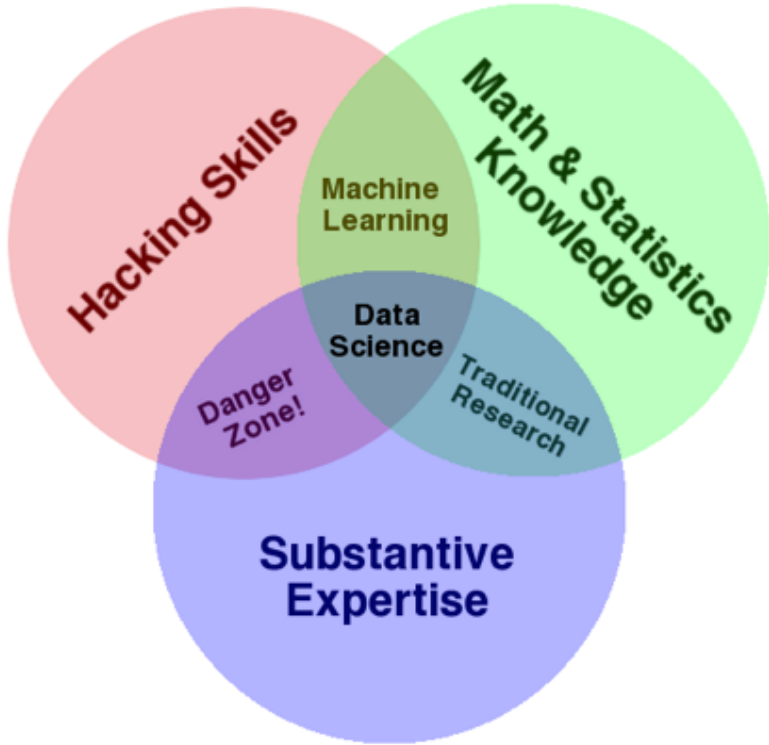
- Reading images of handwritten characters
- Face detection
- Spam filtering
- Voice recognition

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST



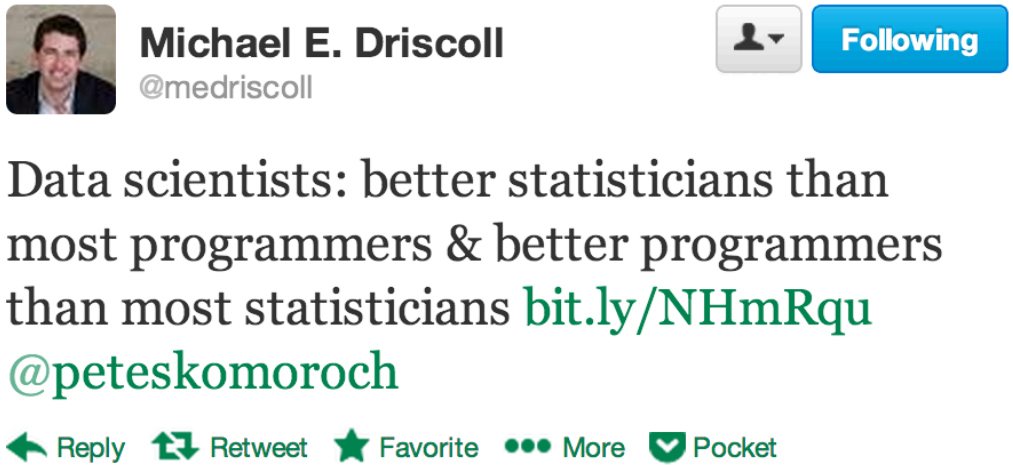
ONE MORE THING!

Communication skills

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

- 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.


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

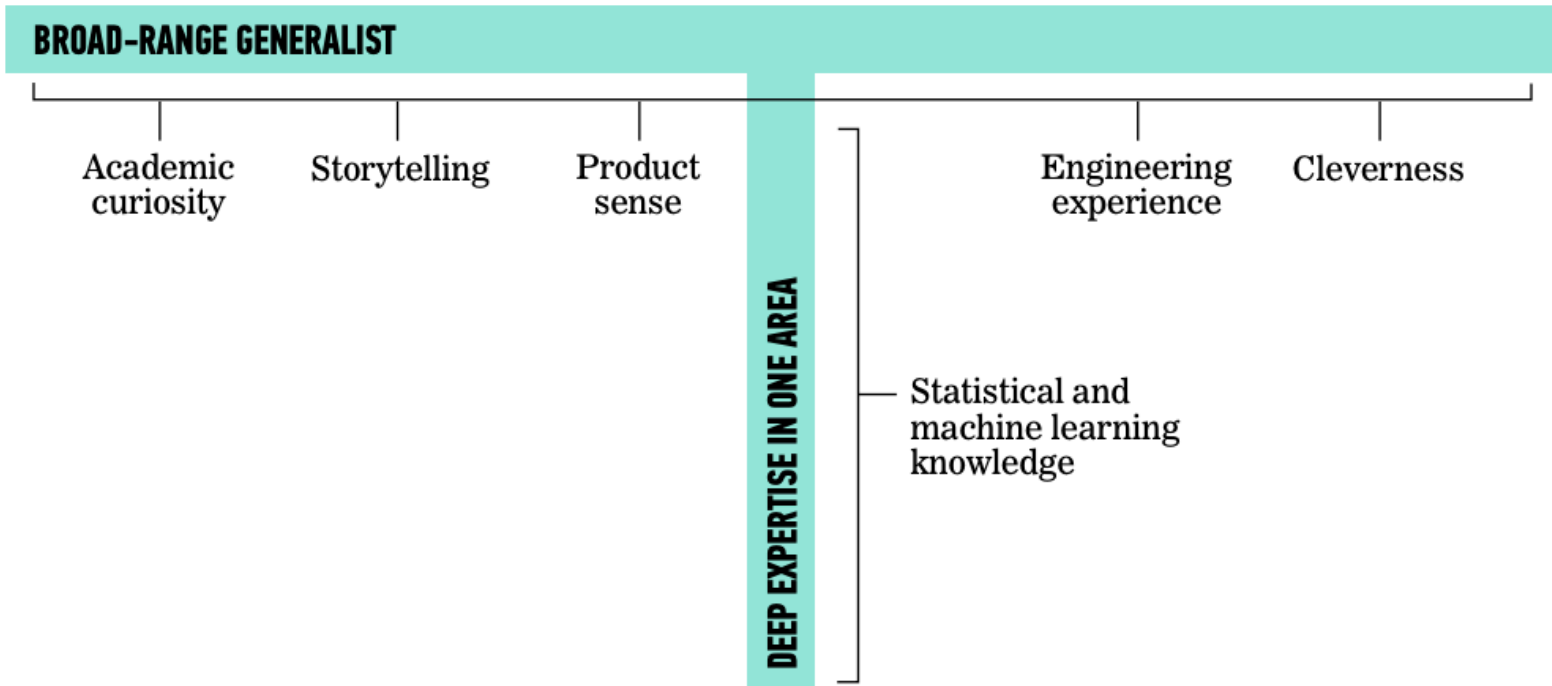


"A hybrid computer scientist software engineer statistician. The best tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them." – Rachel Schutt, News Corp

WHAT MAKES A GOOD DATA SCIENTIST?

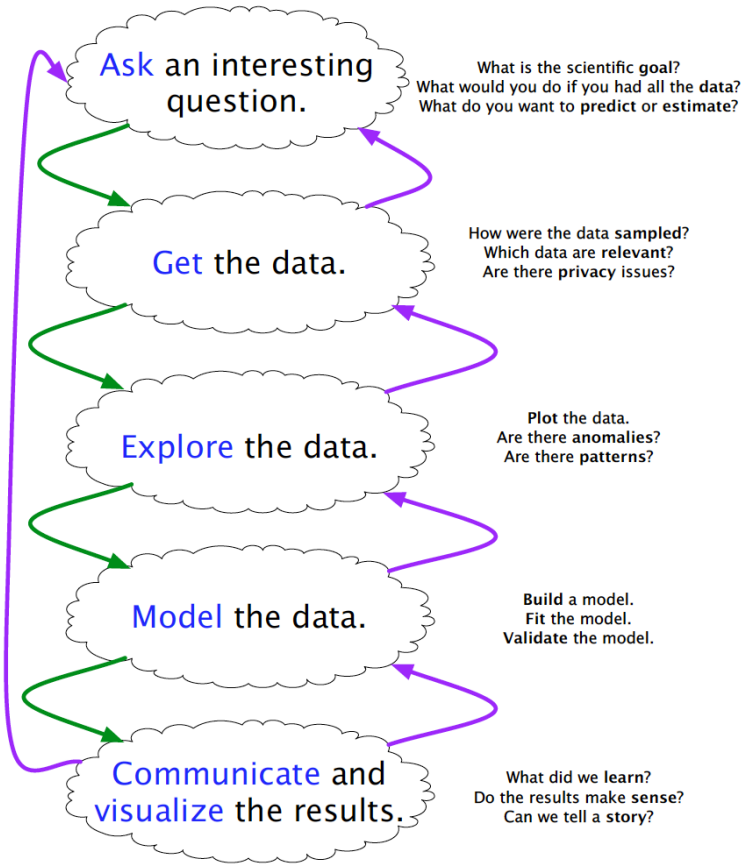
- Statistical and machine learning knowledge
- Engineering experience
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

WHAT MAKES A GOOD DATA SCIENTIST?



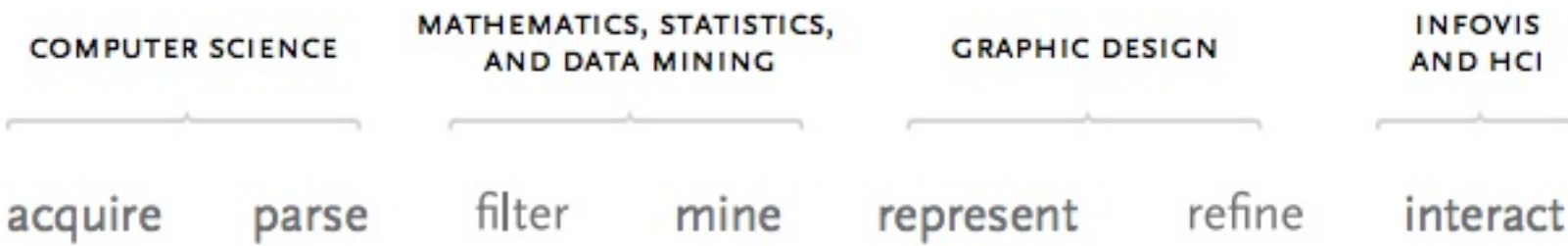
II. THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW



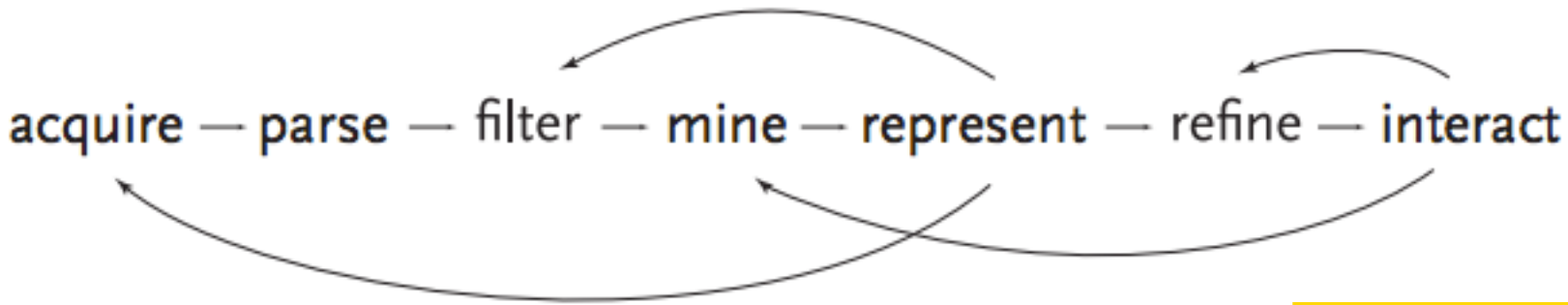
https://drive.google.com/folderview?id=0BxYkKyLxfsNVd0xicUVDS1dIS0k&usp=sharing

THE DATA SCIENCE WORKFLOW



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

THE DATA SCIENCE WORKFLOW



NOTE

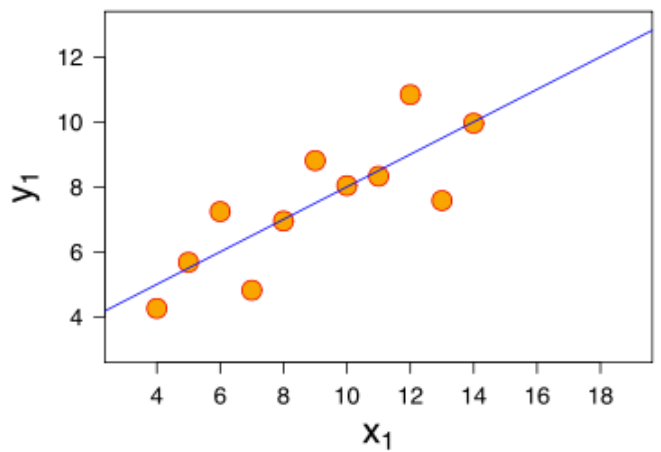
This diagram illustrates the *iterative* nature of problem solving

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

III. VISUALIZATIONS AS A MEDIUM

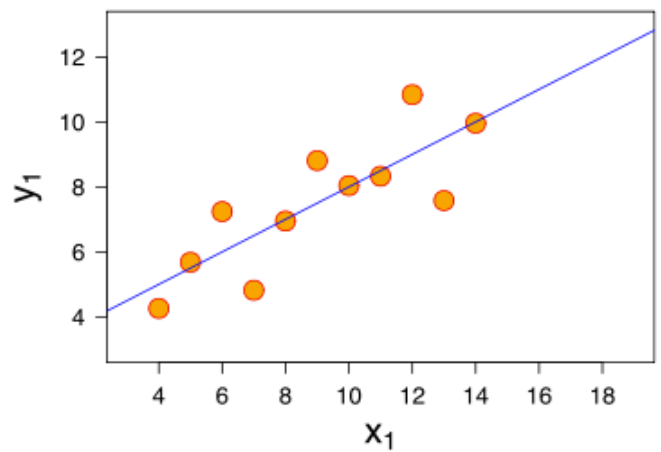
Consider the following dataset:

- eleven (x, y) points



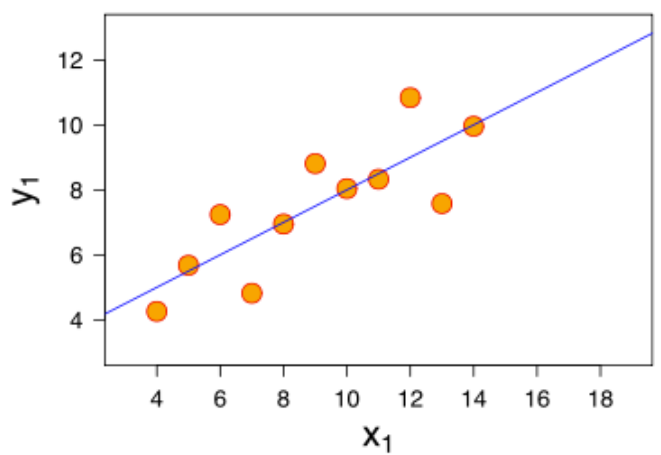
Consider the following dataset:

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



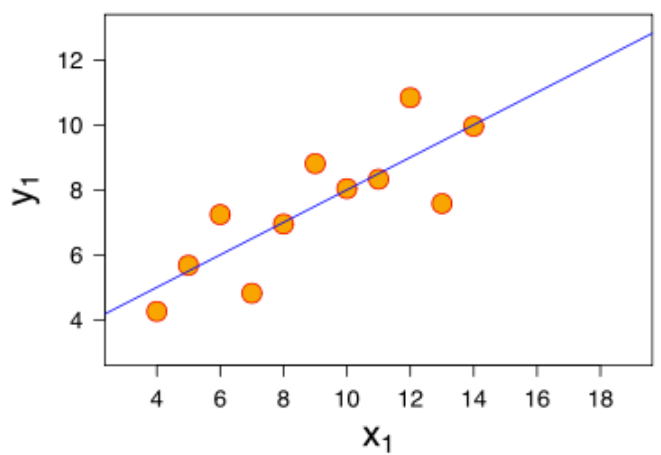
Consider the following dataset:

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



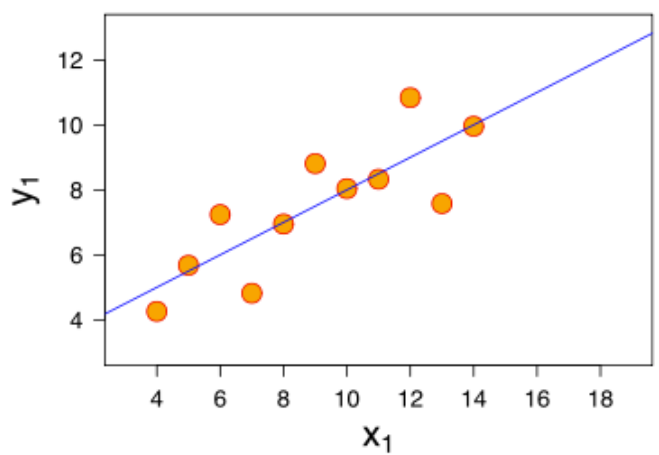
Consider the following dataset:

- 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:

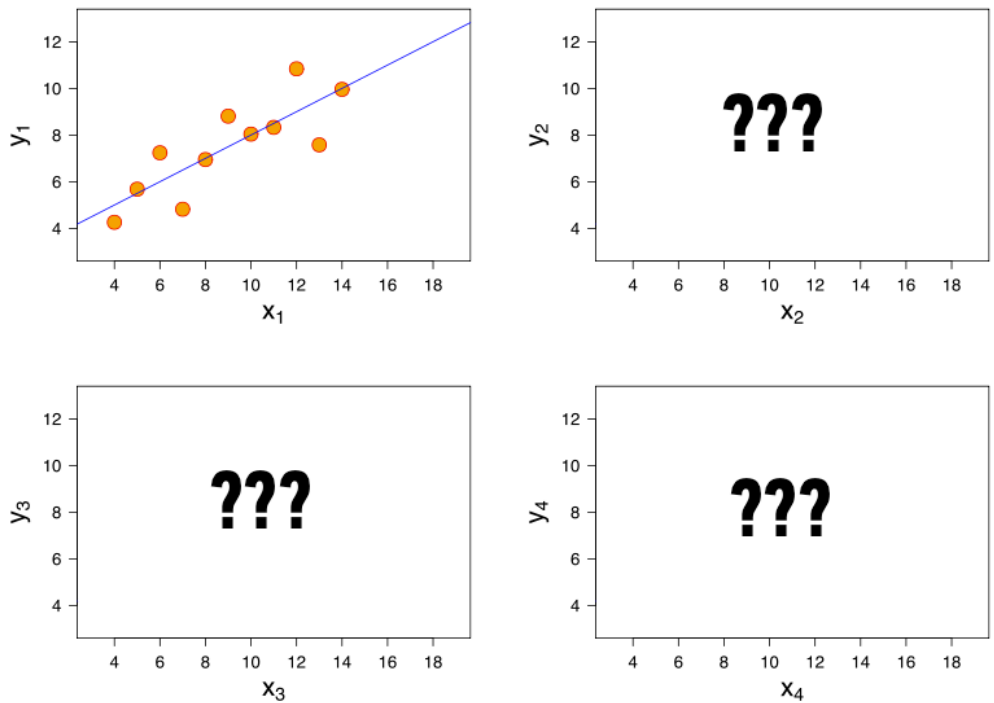
- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x and y = 0.8
- line of best fit: y = 3.00 + 0.500x



EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?



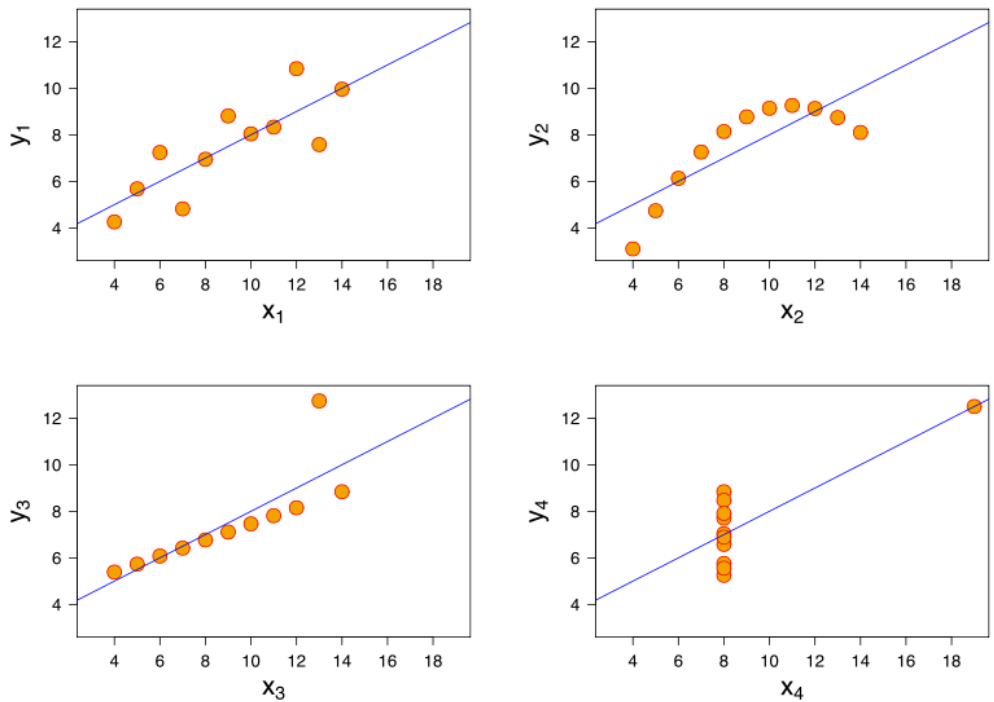
EXERCISE – WHY VISUALIZE DATA?

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



IV. VISUALIZING DATA WITH R AND GGPLOT2

EXERCISE – VISUALIZING DATA WITH R AND GGPLOT2

KEY OBJECTIVES

- Become familiar with the R environment
- Explore data in R
- Visualize data using ggplot2
- Mathematical bonus: power laws

DISCUSSION