GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 1: DATA EXPLORATION

INTRO TO DATA SCIENCE

WELCOME!

LOGISTICS

Course Website: https://www.schoology.com/

Instructors: Thomson Nguyen, Jacob Bollinger

E-mail: thomson@cantab.net, jacob@bright.com

Course Times: 6:00pm-9:00pm, Tuesdays and Thursdays (Hattery)

Office Hours: Wednesday 5-7pm (preliminary)

LOGISTICS

Course Website: https://www.schoology.com/

Instructors: Thomson Nguyen, Jacob Bollinger

E-mail: thomson@cantab.net, jacob@bright.com

Course Times: 6:00pm-9:00pm, Tuesdays and Thursdays (Hattery)

Office Hours: Wednesday 5-7pm (preliminary)

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

LAB:
III. WORKING AT THE UNIX COMMAND LINE
IV. VISUALIZING DATA WITH R & GGPLOT2

I. WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

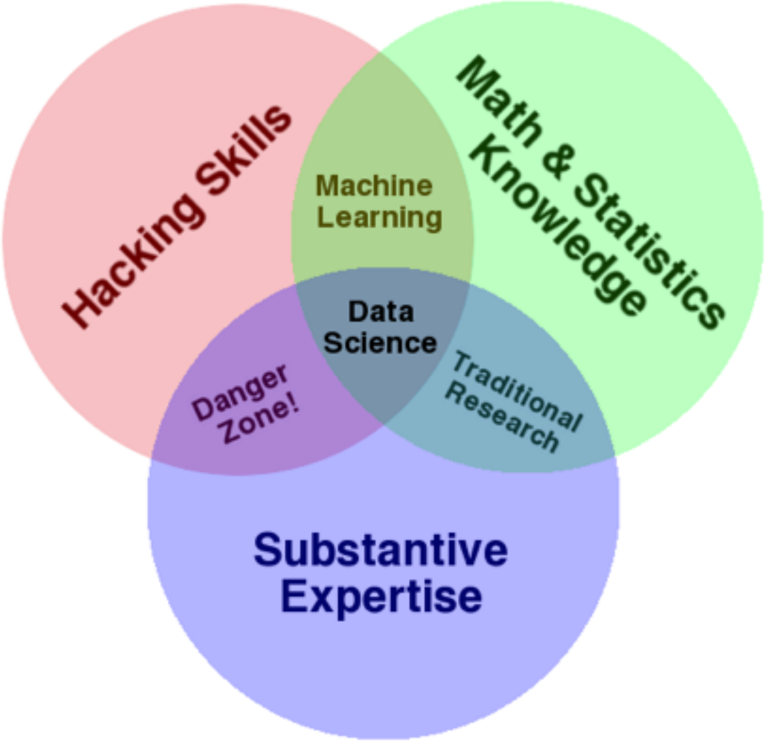
- A set of tools and techniques used to extract useful information from data.

WHAT IS DATA SCIENCE?

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

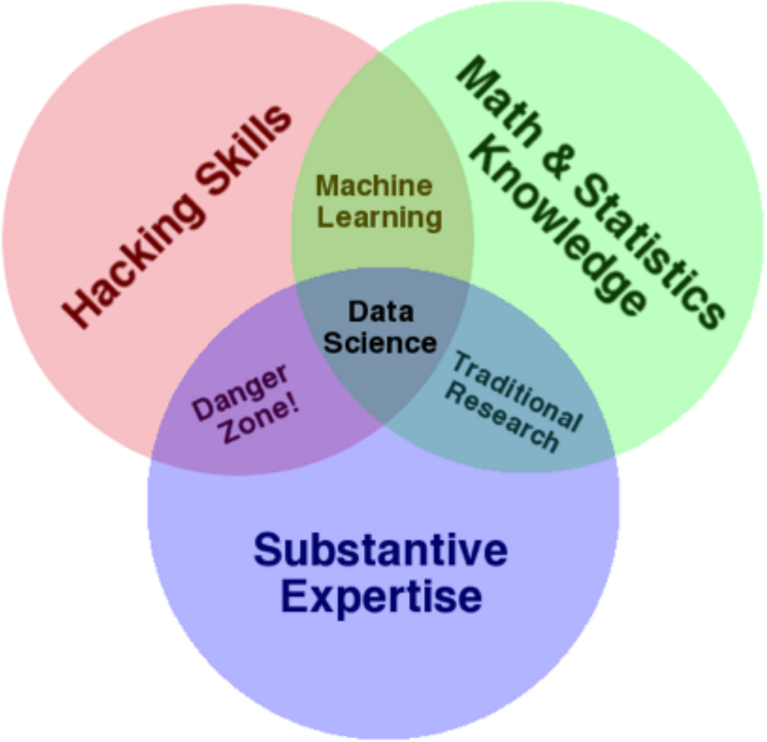
- An interdisciplinary, problem-oriented subject.

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.

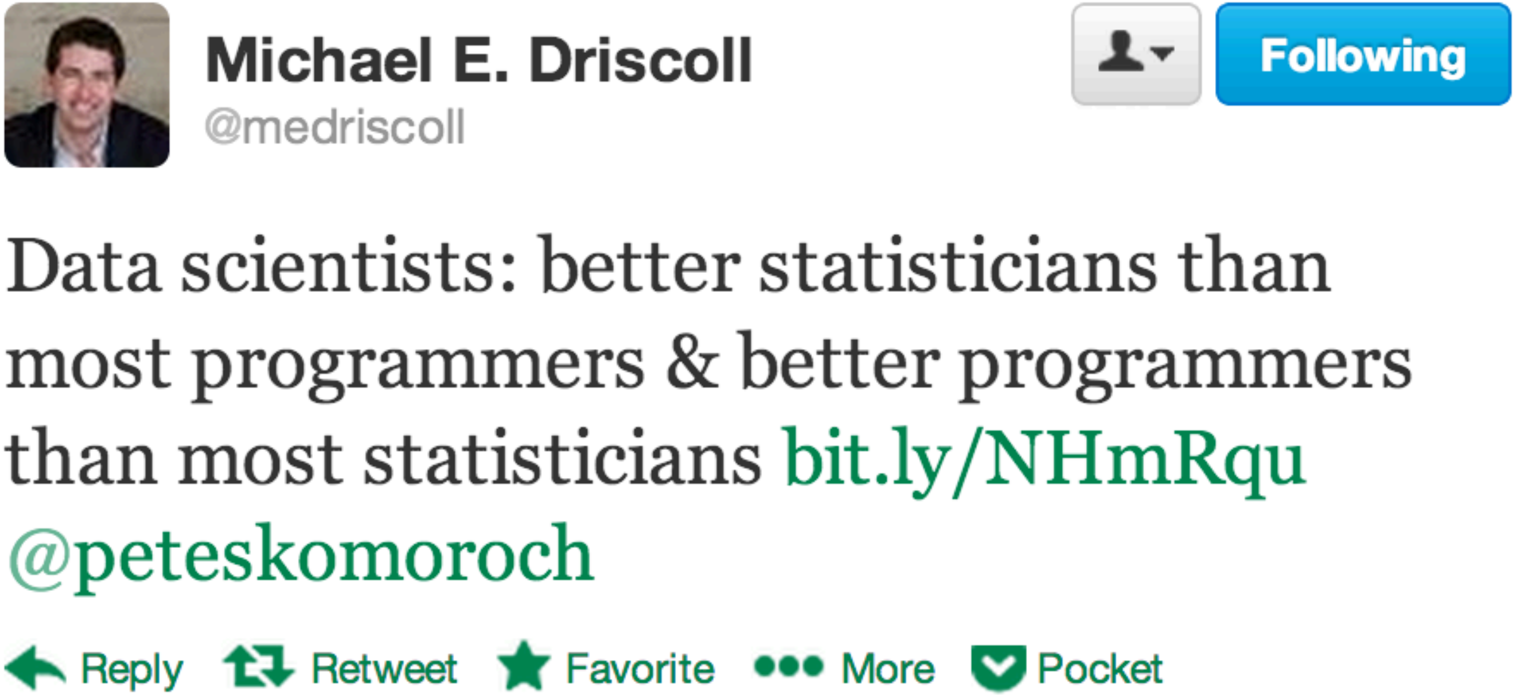
- The application of scientific techniques to practical problems.

- A rapidly growing field.

WHO USES DATA SCIENCE?

NETFLIX
amazon
facebook
Google™
Linked in ®
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft

WHAT MAKES A GOOD DATA SCIENTIST?

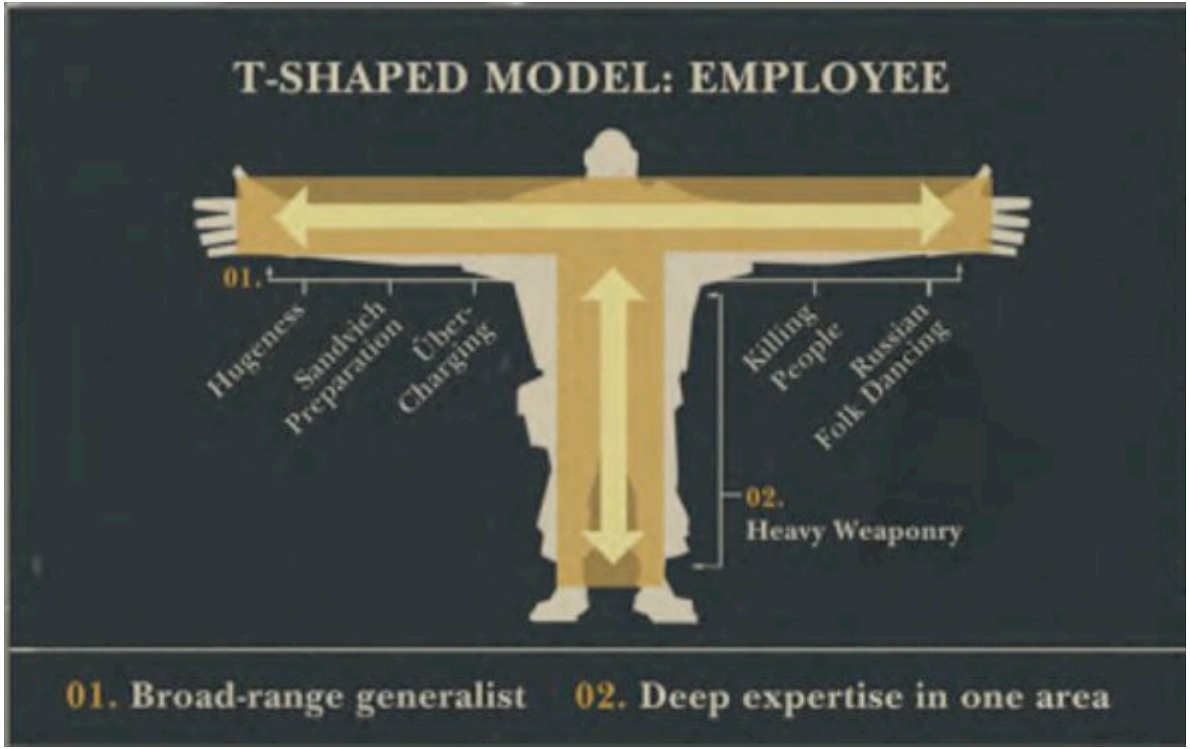


Michael E. Driscoll

@medriscoll

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

WHAT MAKES A GOOD DATA SCIENTIST?

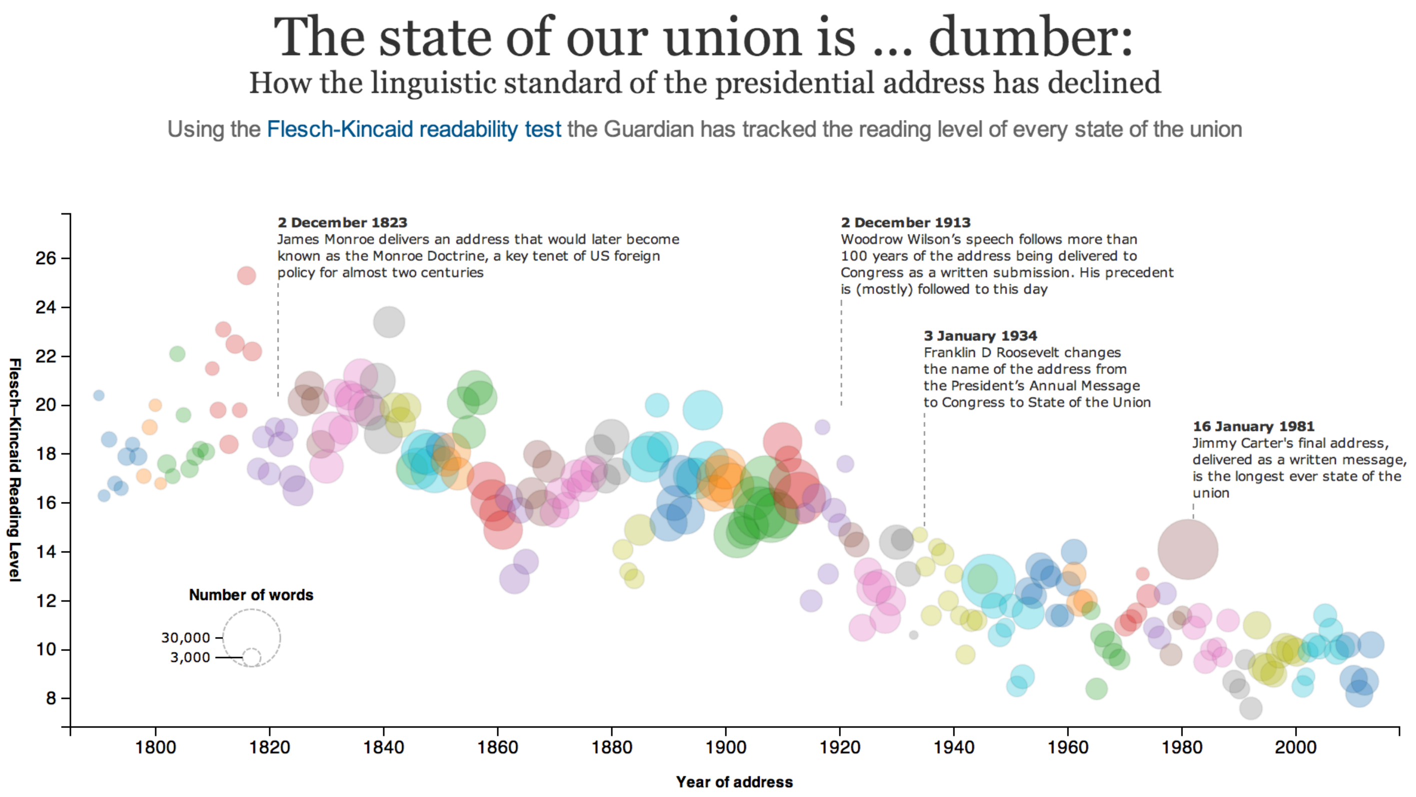


(Valve Software)

WHAT MAKES A GOOD DATA SCIENTIST?

- Statistical and machine learning knowledge
- Engineering experience
- Academic 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

2 December 1823
James Monroe delivers an address that would later become known as the Monroe Doctrine, a key tenet of US foreign policy for almost two centuries

2 December 1913
Woodrow Wilson's speech follows more than 100 years of the address being delivered to Congress as a written submission. His precedent is (mostly) followed to this day

3 January 1934
Franklin D Roosevelt changes the name of the address from the President's Annual Message to Congress to State of the Union

16 January 1981
Jimmy Carter's final address, delivered as a written message, is the longest ever 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 college for you

II. THE DATA SCIENCE WORKFLOW

Dataists

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

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

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

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

Ted Johnson

- 1. Assemble an accurate and relevant data set
- 2. Choose the appropriate algorithm

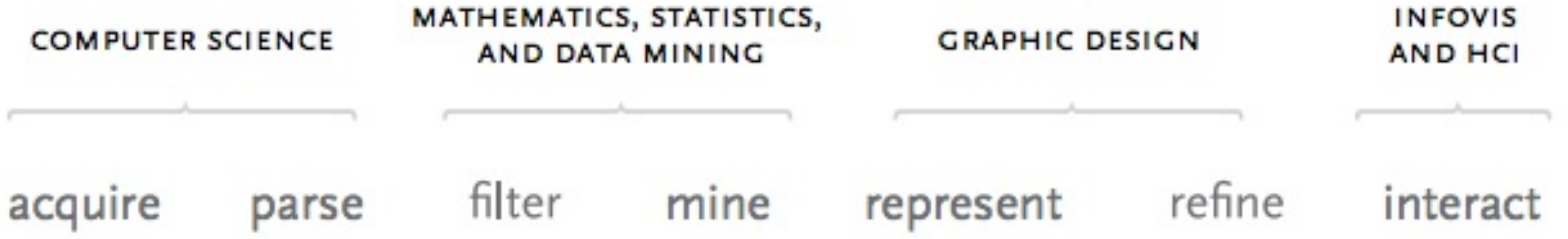
source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ben Fry

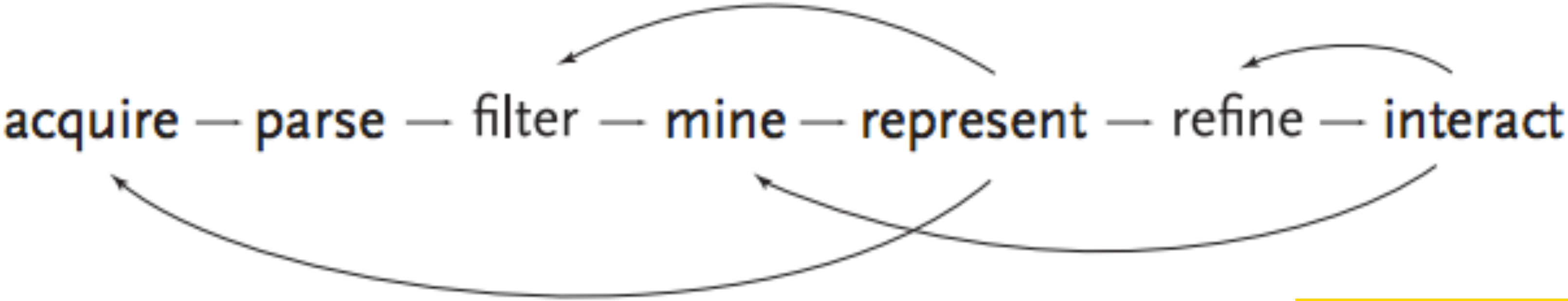
- 1. Acquire
- 2. Parse
- 3. Filter
- 4. Mine
- 5. Represent
- 6. Refine
- 7. Interact

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

THE DATA SCIENCE WORKFLOW



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


acquire — parse — filter — mine — represent — refine — interact

NOTE

This diagram illustrates the *iterative* nature of problem solving

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

III. WORKING AT THE UNIX COMMAND LINE

EXERCISE – WORKING AT THE UNIX COMMAND LINE

Download this dataset:

http://bit.ly/pacedataset

EXERCISE – WORKING AT THE UNIX COMMAND LINE

KEY OBJECTIVES	TOOLS
- Navigate the filesystem	- ls, cd
- Create, move, copy, and delete files & directories	- cat, touch, mv, cp, mkdir, rm, rmdir
- View & search files	- head, tail, less, cat, grep
- Edit & interact with files	- vim, tr, sort, uniq, wc
- Combine steps	- pipe (|)
- Learn more	- man, apropos

NOTE

Being comfortable at the command line makes your life much easier!

INTRO TO DATA SCIENCE

IV. VISUALIZING DATA WITH R AND GGPLOT2

EXERCISE – VISUALIZING DATA WITH R AND GGPLOT2

KEY OBJECTIVES

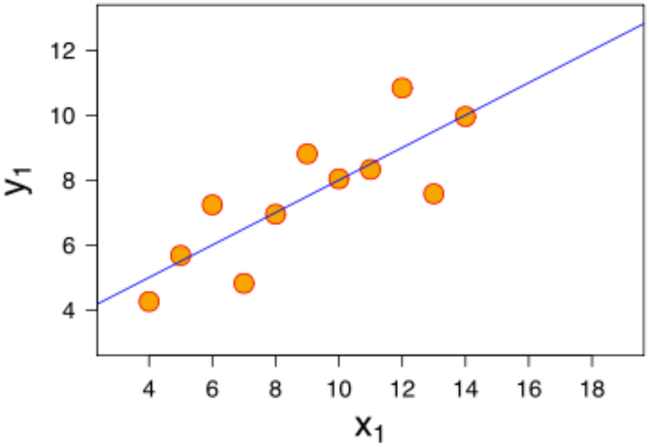
- Become familiar with the R environment
- Explore data in R
- Visualize data using ggplot2
- Mathematical bonus: power laws

IV. VISUALIZATIONS AS A MEDIUM

EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

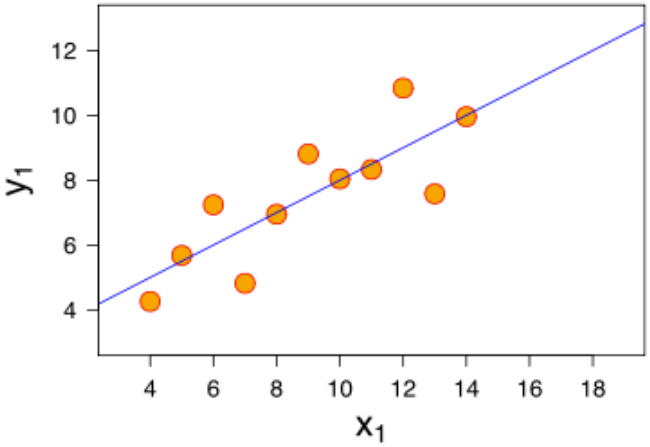
- *eleven (x, y) points*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

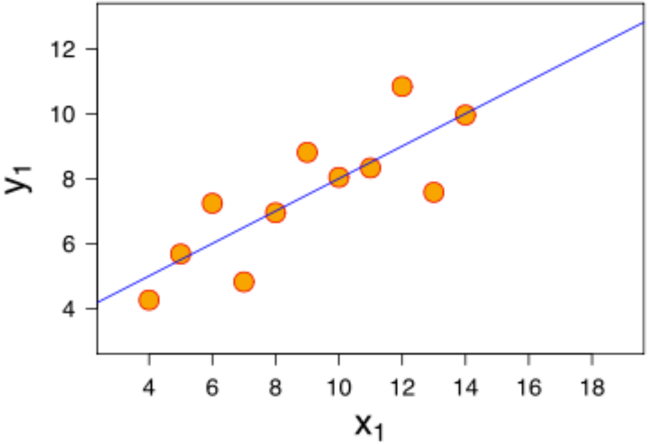
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*



EXERCISE – WHY VISUALIZE DATA?

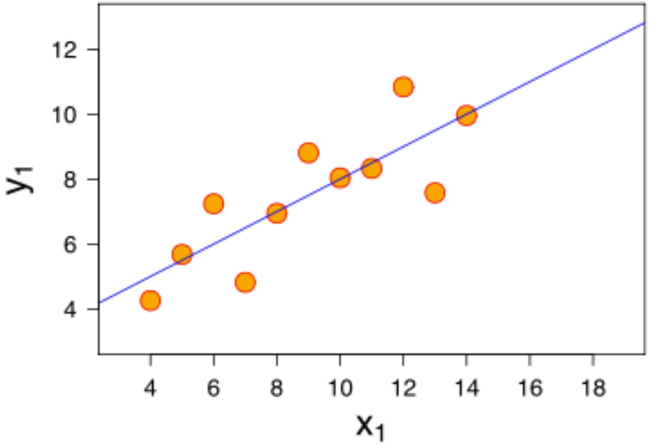
Consider the following dataset:

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



Consider the following dataset:

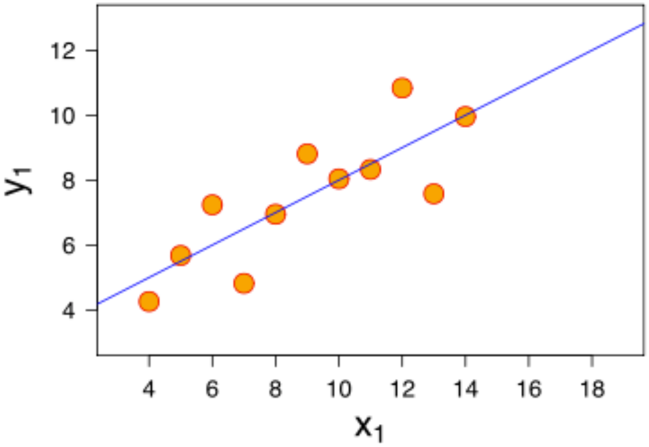
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x and y = 0.8*



EXERCISE – WHY VISUALIZE DATA?

Consider the following dataset:

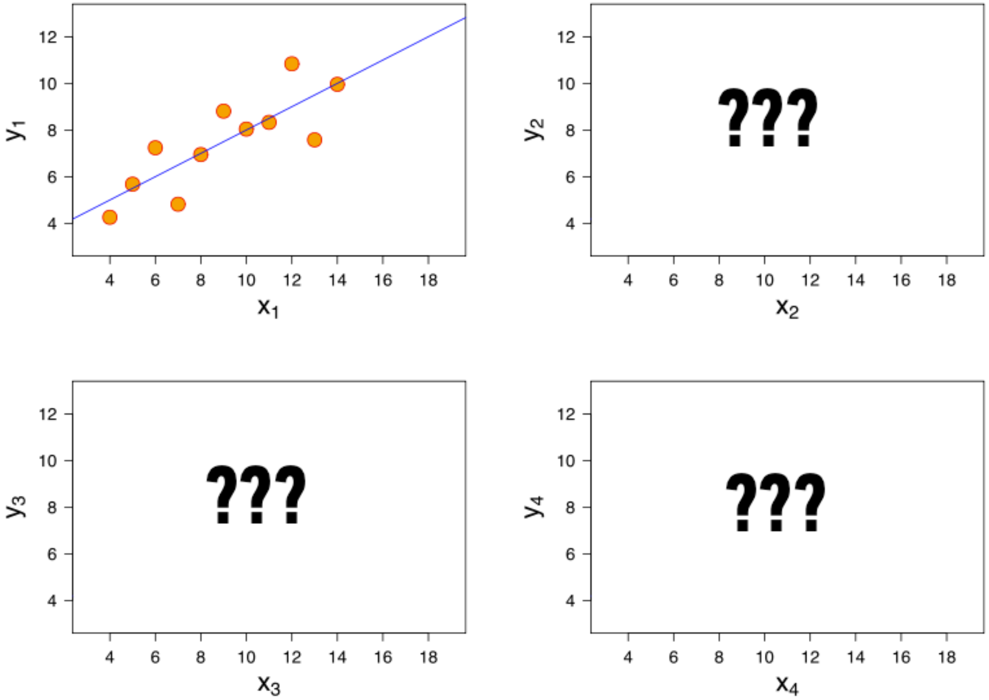
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?

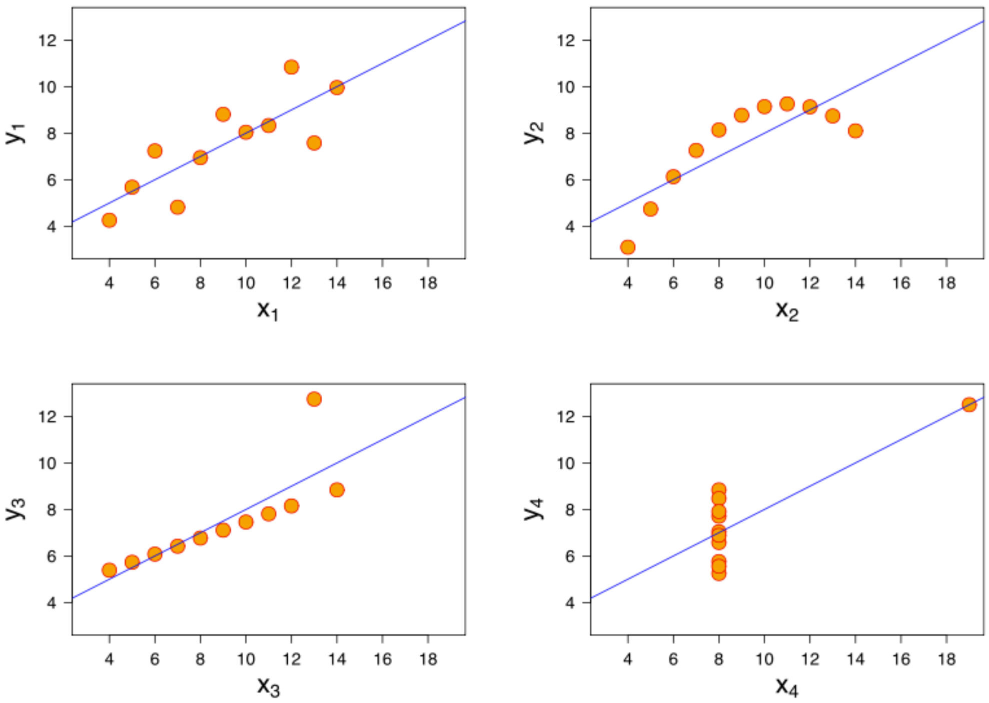


EXERCISE – WHY VISUALIZE DATA?

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

DISCUSSION