GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: DATA EXPLORATION

Jason Dolatshahi
Data Scientist, EveryScreen Media

INTRO TO DATA SCIENCE

WELCOME!

my email: jason@everyscreenmedia.com

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

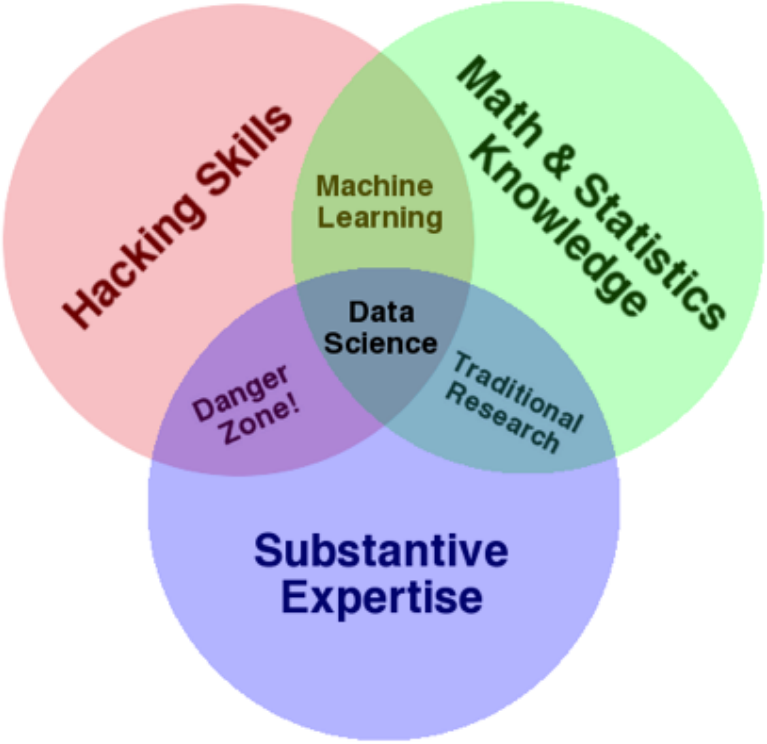
EXERCISES:
III. WORKING AT THE UNIX COMMAND LINE
IV. VISUALIZING DATA WITH R & GGPLOT2

I. WHAT IS DATA SCIENCE?

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

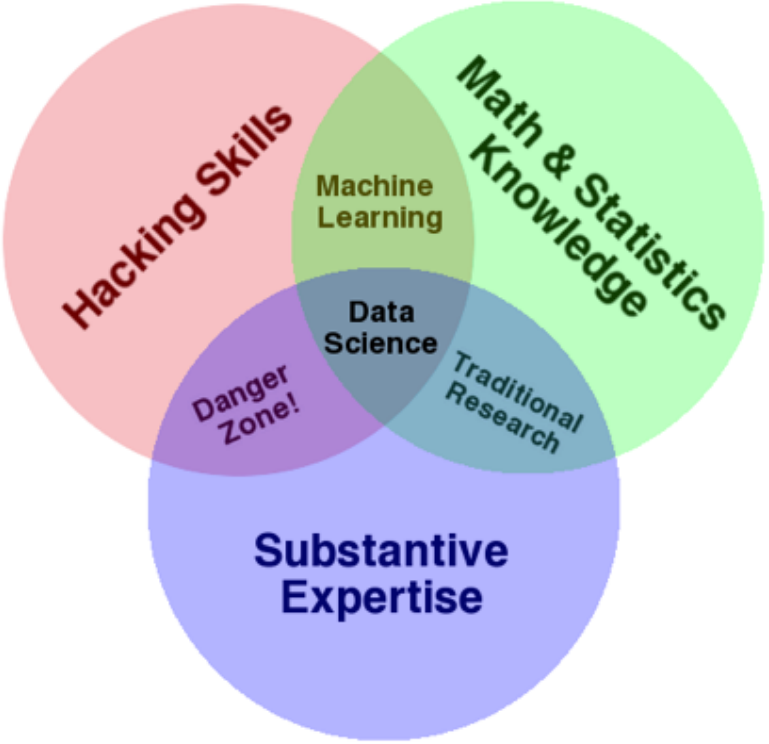
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST

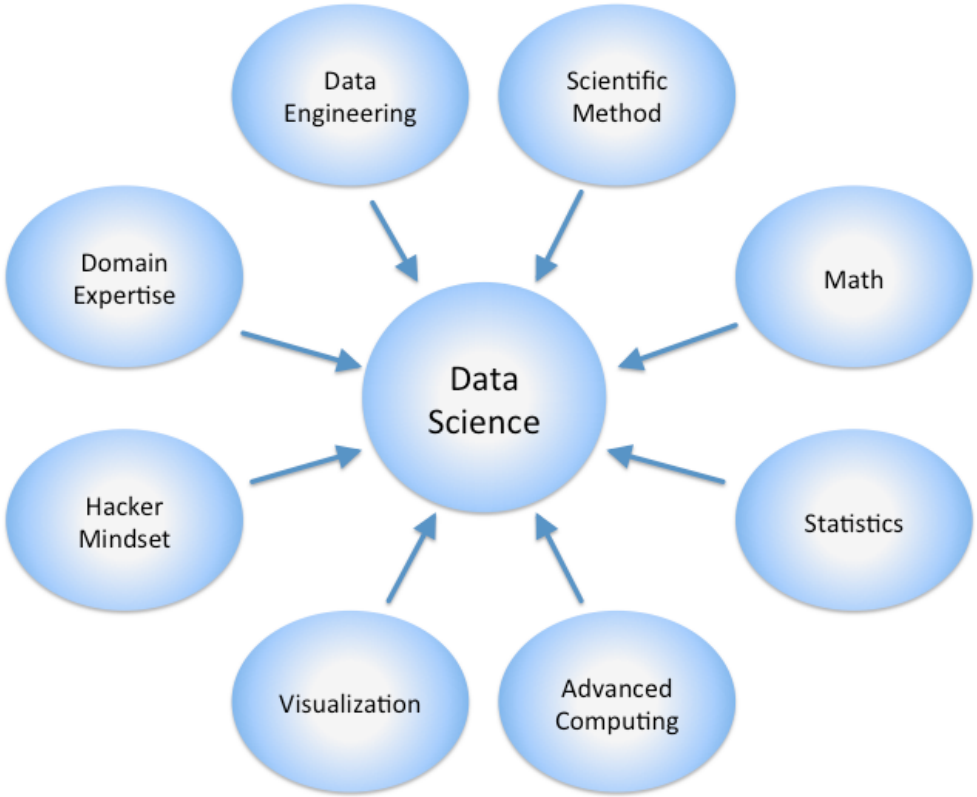


ONE MORE THING!

Communication skills

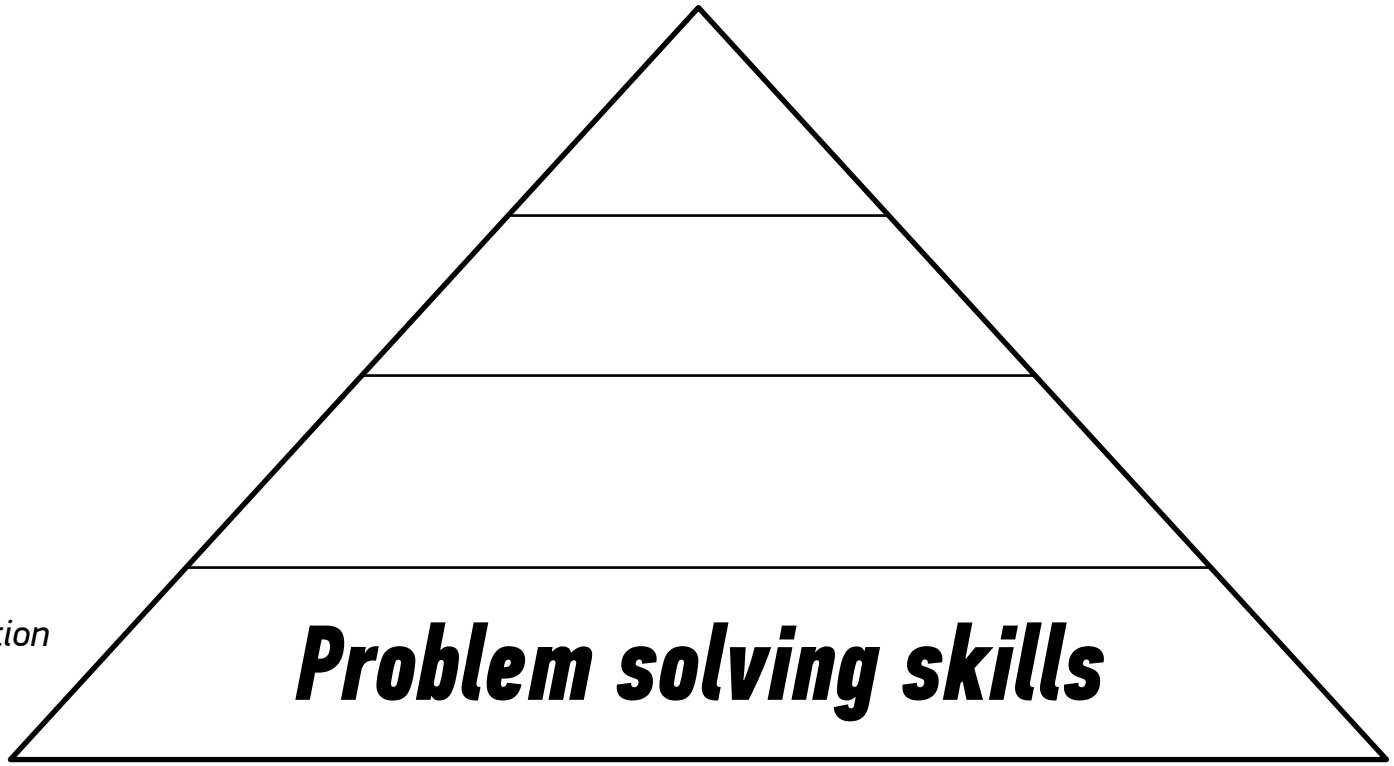
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

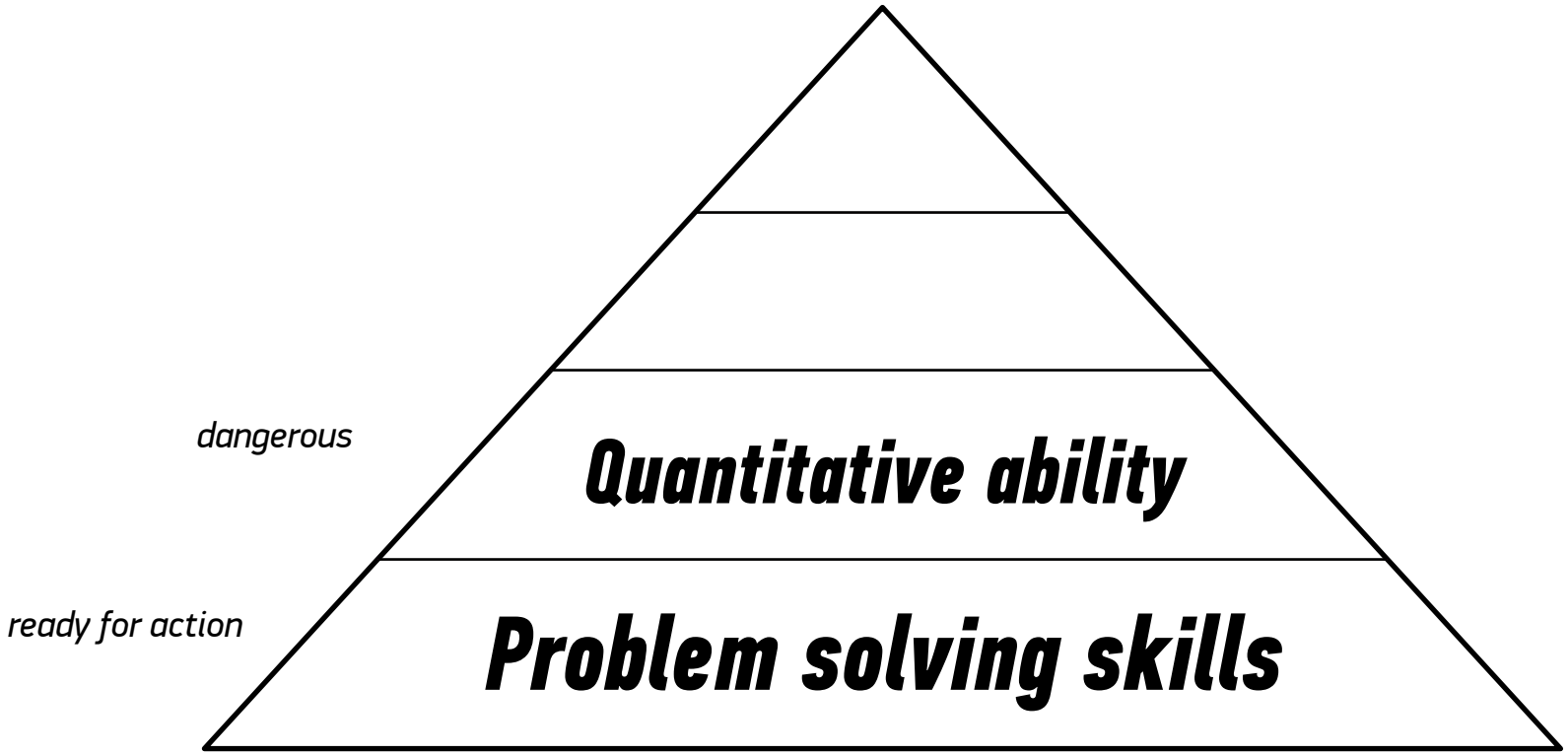
WHAT IS DATA SCIENCE?

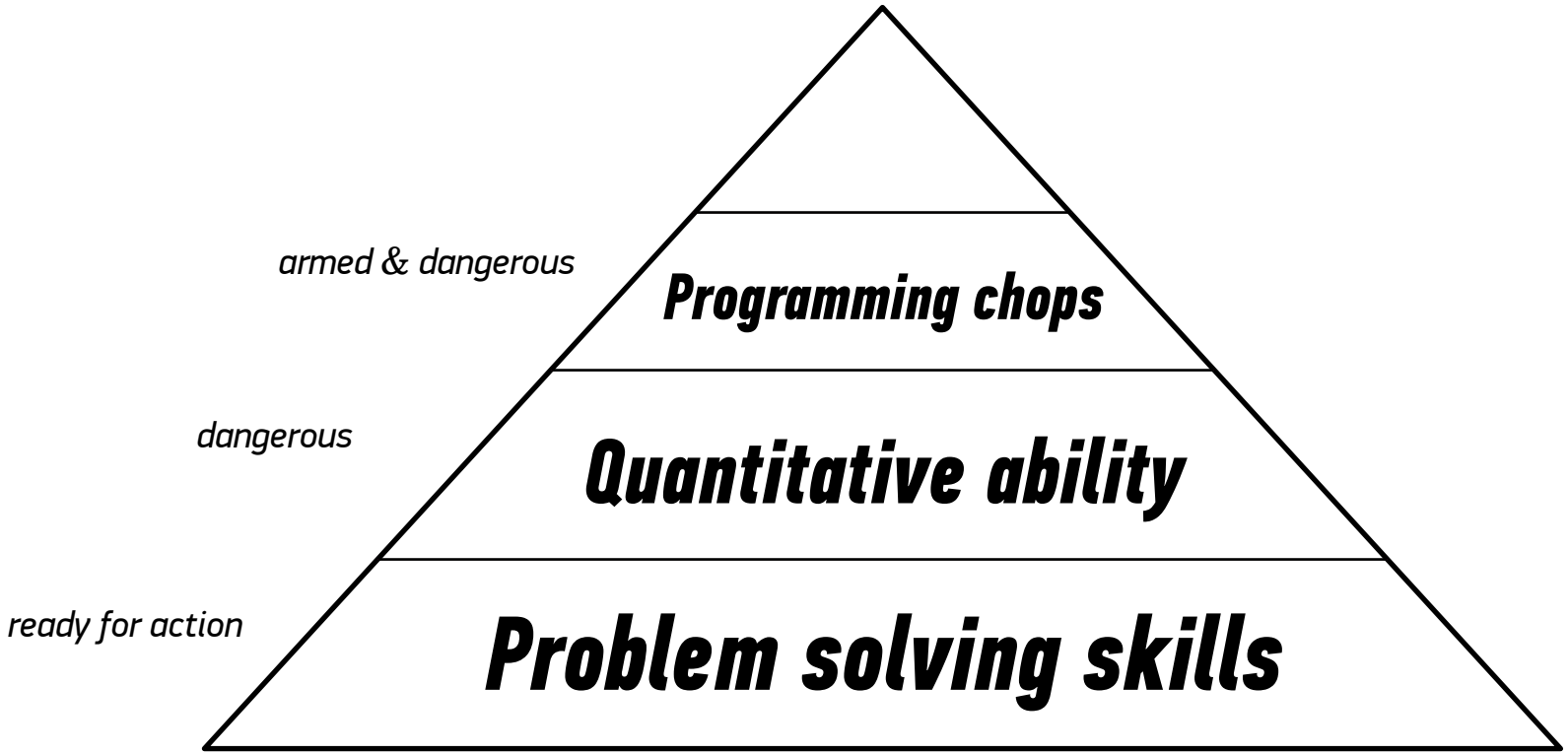


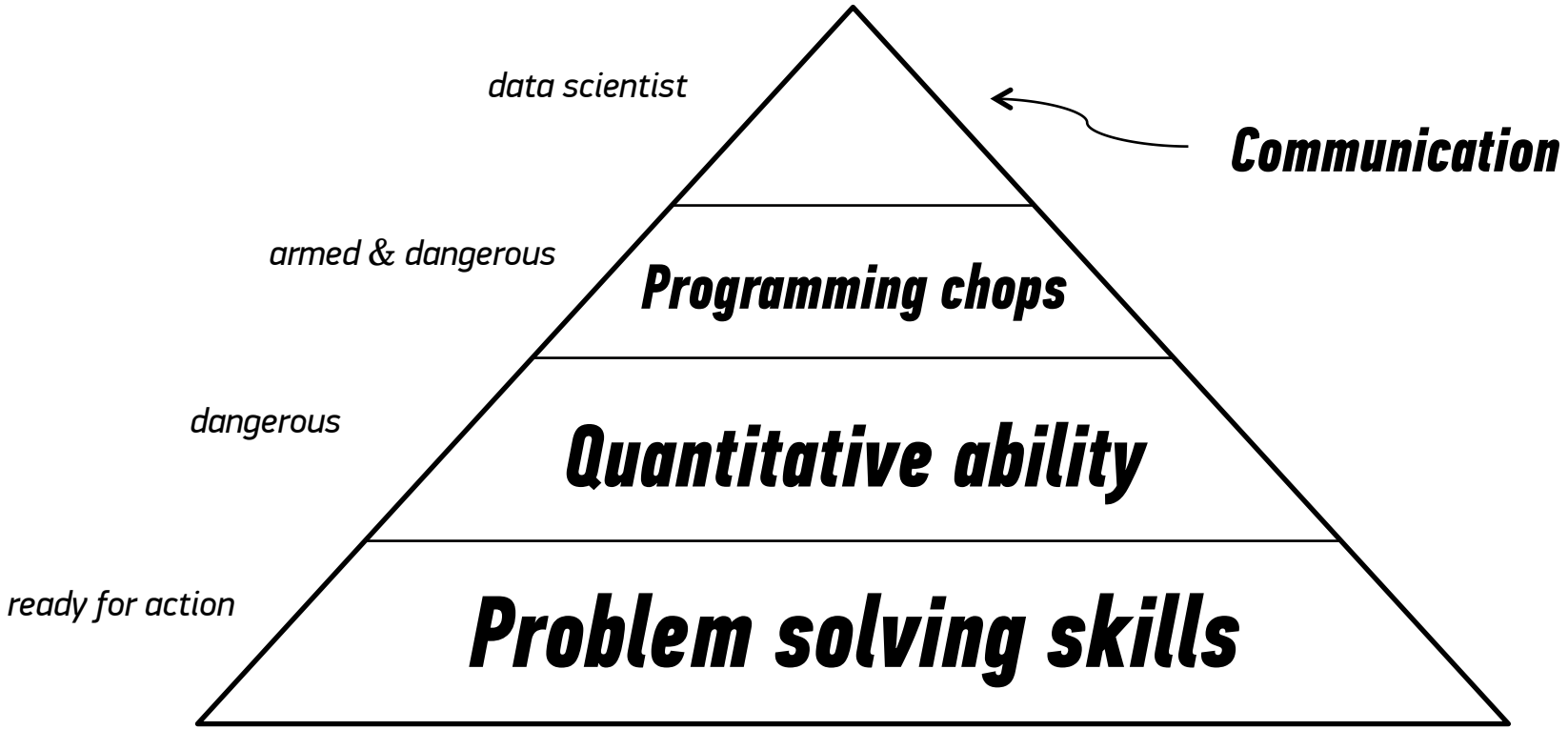
source: http://en.wikipedia.org/wiki/Data_science

- 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.


ready for action
Problem solving skills


dangerous
Quantitative ability
ready for action
Problem solving skills


armed & dangerous
Programming chops
dangerous
Quantitative ability
ready for action
Problem solving skills


data scientist
Communication
armed & dangerous
Programming chops
dangerous
Quantitative ability
ready for action
Problem solving skills

- 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.

- Data science solutions are useful for making informed decisions.

- Data science solutions are useful for making informed decisions.
- A typical technology company has lots of data at its disposal (frequently, more than it knows what to do with).

DATA STORAGE: THEN

source: http://nineteenthcenturybaruch.wordpress.com/page/2/

DATA STORAGE: NOW

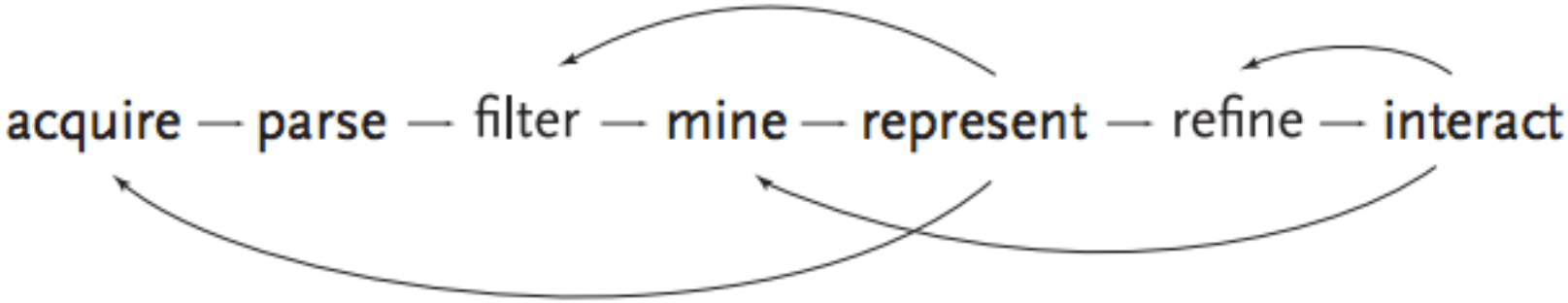
source: http://inhabitat.com/facebooks-energy-efficient-prineville-data-center-in-oregon-achieves-leed-gold-status/

- Data science solutions are useful for making informed decisions.
- A typical technology company has lots of data at its disposal (frequently, more than it knows what to do with).
- Data scientists harness and extract the information in this data to create meaningful metrics and/or models.

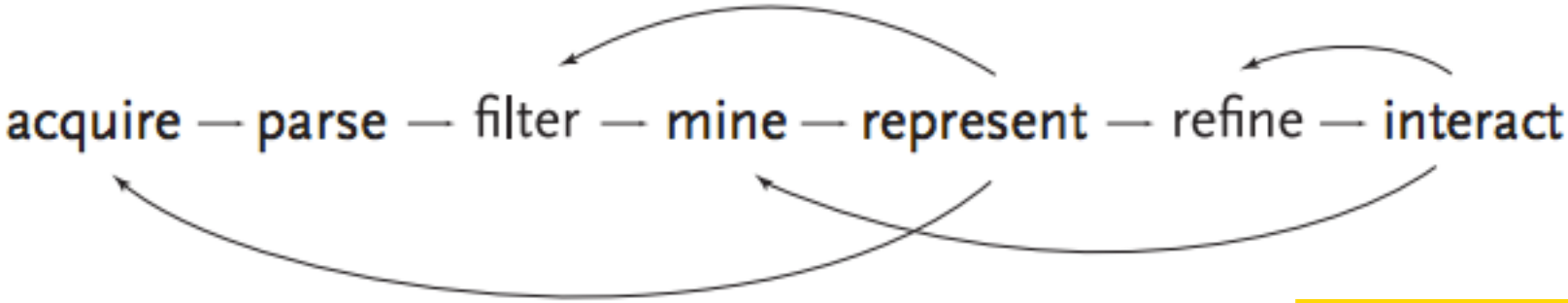

NETFLIX
amazon
facebook
Google
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft

II. THE DATA SCIENCE WORKFLOW

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



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

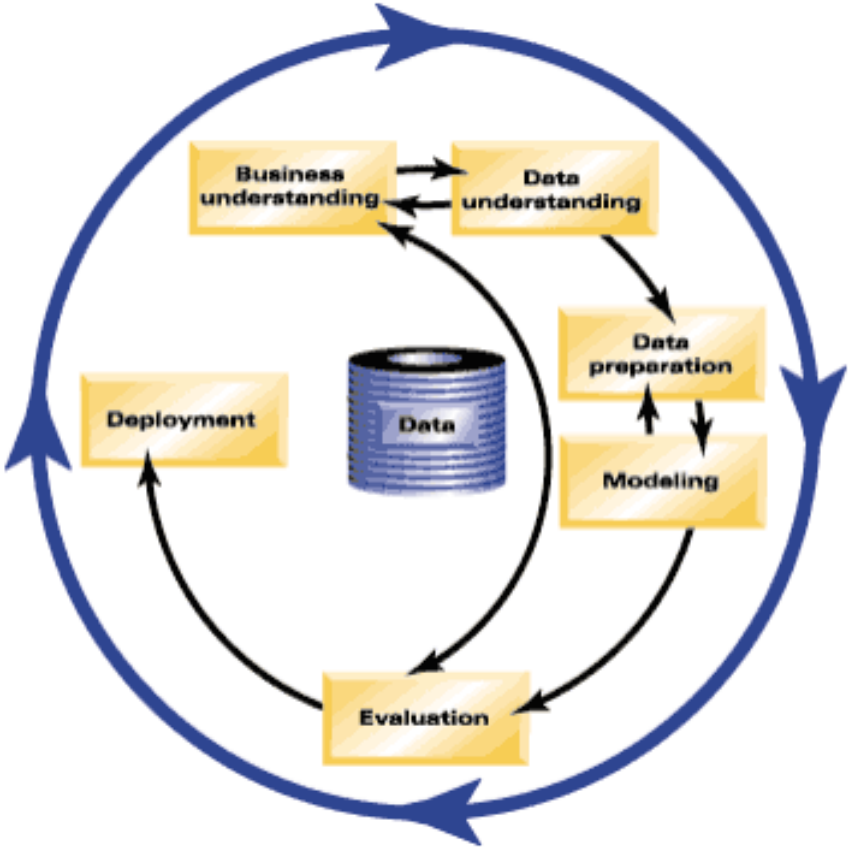


NOTE

This diagram illustrates the *iterative* nature of problem solving

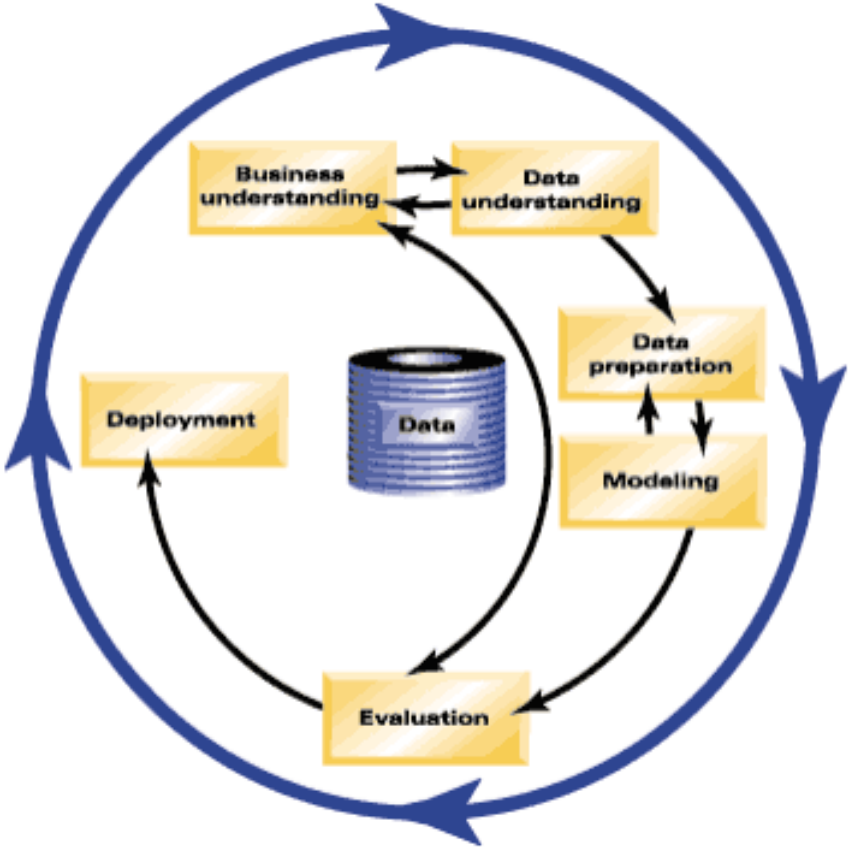
source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW – THE CRISP DM DIAGRAM



source: http://www.crisp-dm.org/

THE DATA SCIENCE WORKFLOW – THE CRISP DM DIAGRAM



source: http://www.crisp-dm.org/

NOTE

Again, this illustrates an *iterative* approach to problem solving

INTRO TO DATA SCIENCE

III. 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, awk, sed, tr, sort, uniq, wc
- pipe (|)
- man, apropos

NOTE

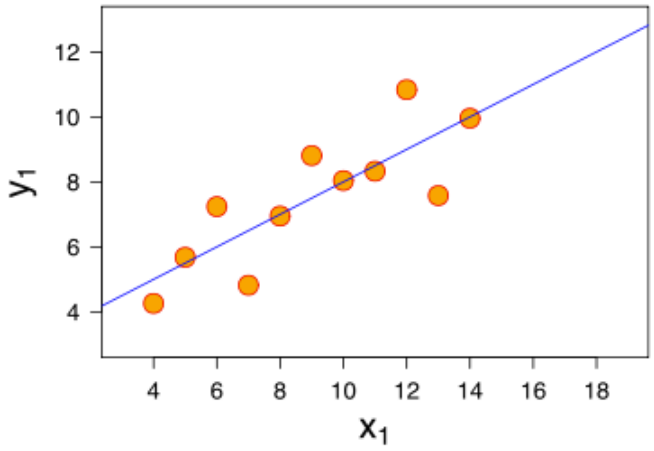
Being comfortable at the command line makes your life much easier!

INTRO TO DATA SCIENCE

IV. VISUALIZING DATA WITH R AND GGPLOT2

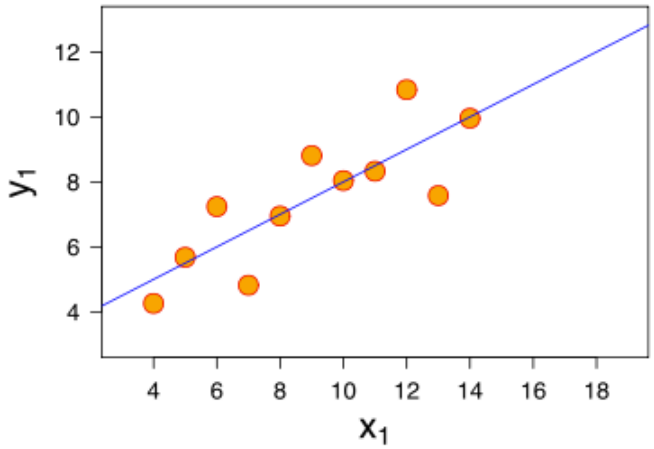
Consider the following dataset:

- eleven (x, y) points



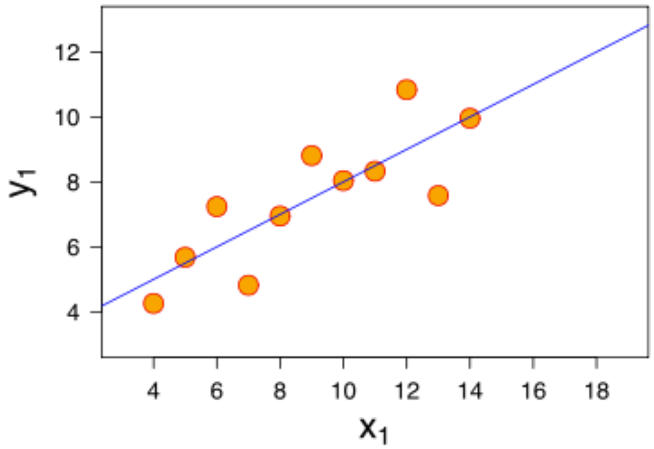
Consider the following dataset:

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



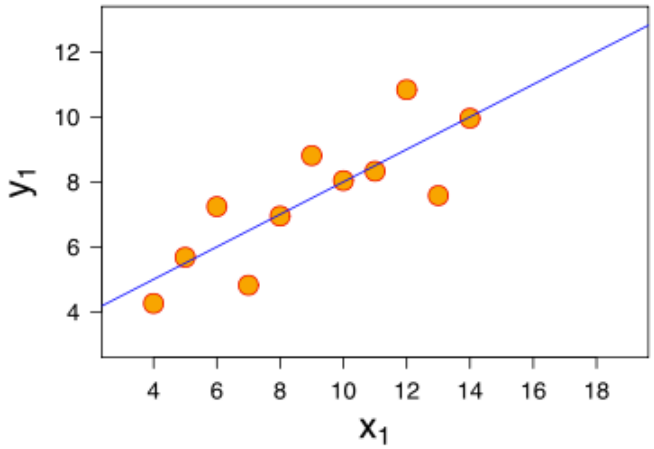
Consider the following dataset:

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



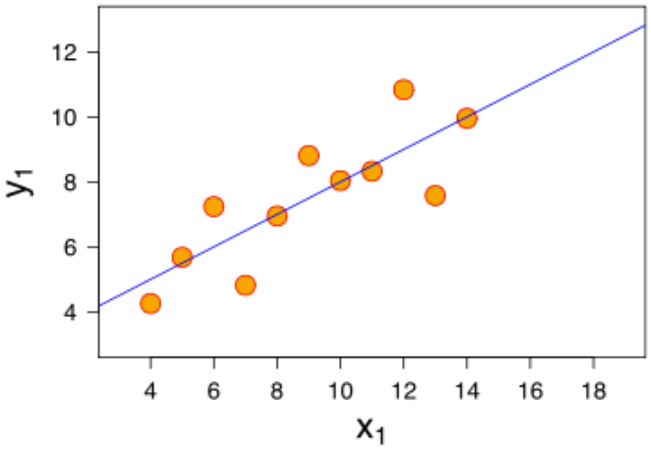
Consider the following dataset:

- 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:

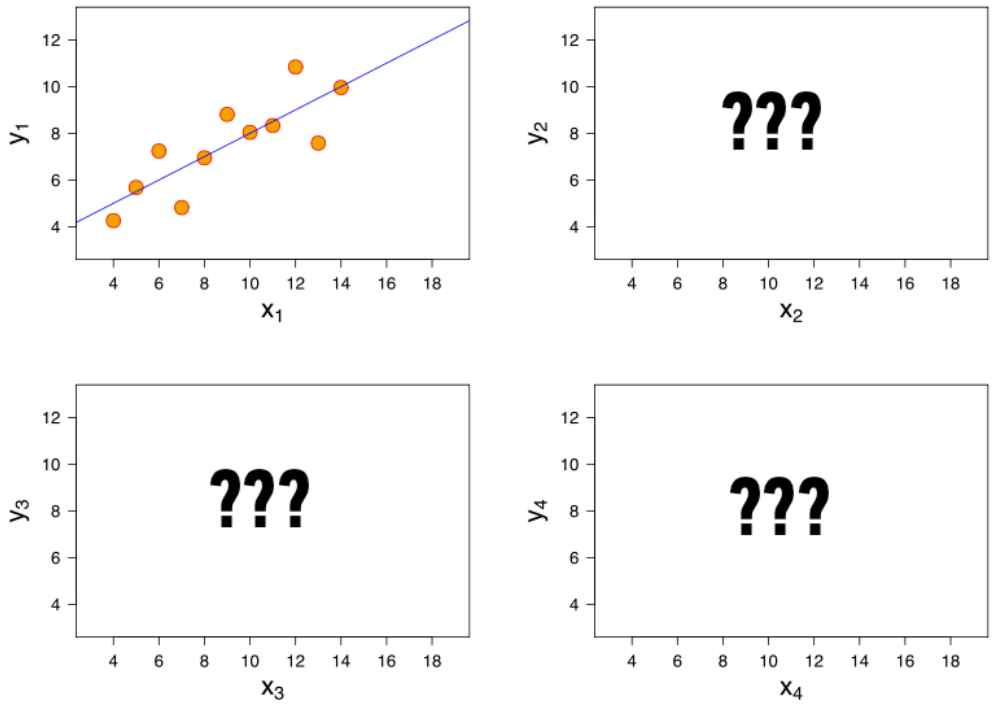
- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x and y = 0.8
- line of best fit: y = 3.00 + 0.500x



EXERCISE – WHY VISUALIZE DATA?

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

Q: How similar are these datasets?



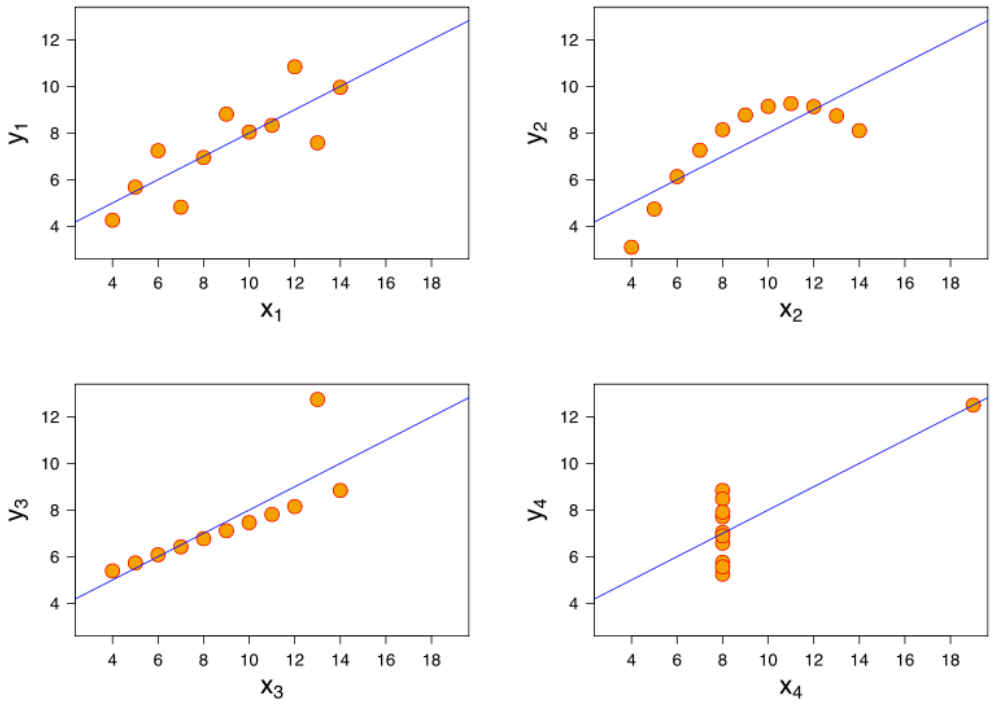
EXERCISE – WHY VISUALIZE DATA?

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



EXERCISE – VISUALIZING DATA WITH R AND GGPLOT2 38

KEY OBJECTIVES

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

DISCUSSION