GENERAL ASSEMBLY

DATA SCIENCE

CLASS 1: INTRO TO DATA SCIENCE

Rob Hall

DAT SF 19 // November 30, 2015

AGENDA

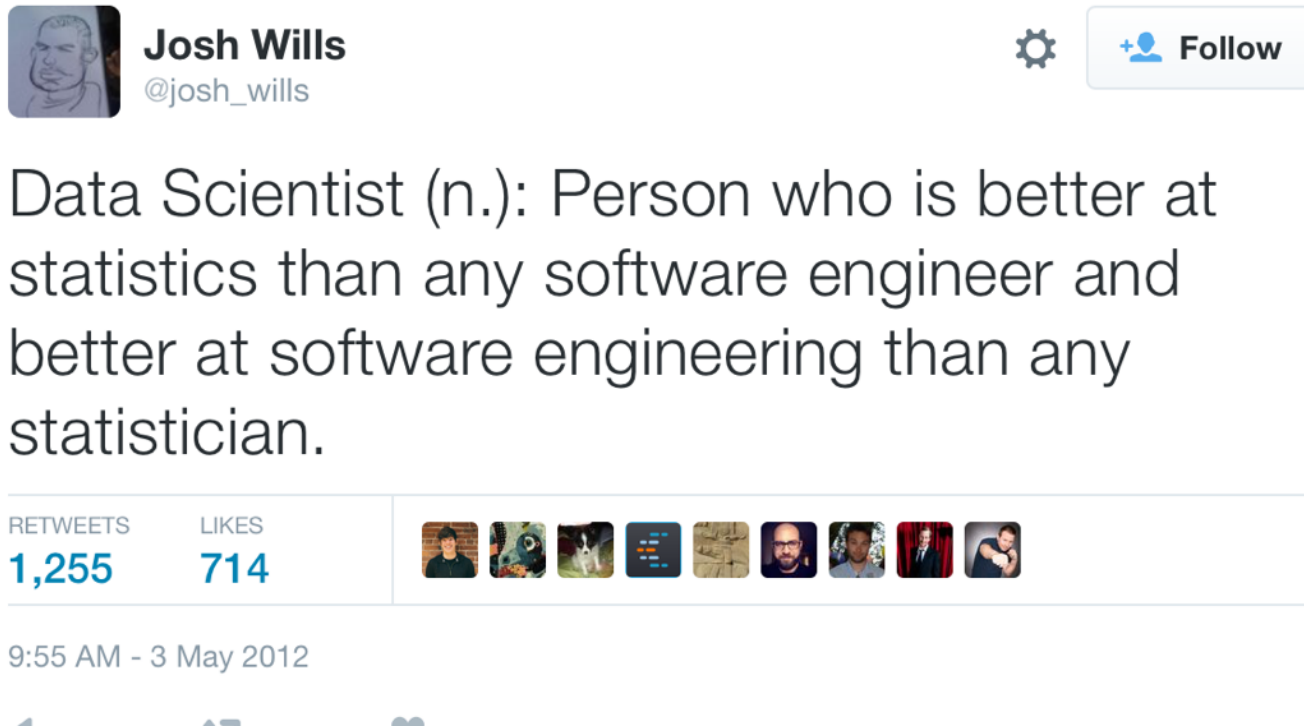
I. WHAT IS A DATA SCIENTIST?
II. WHAT IS DATA SCIENCE?
III. THE DATA SCIENCE WORKFLOW

LAB:
IV. WORKING AT THE UNIX COMMAND LINE

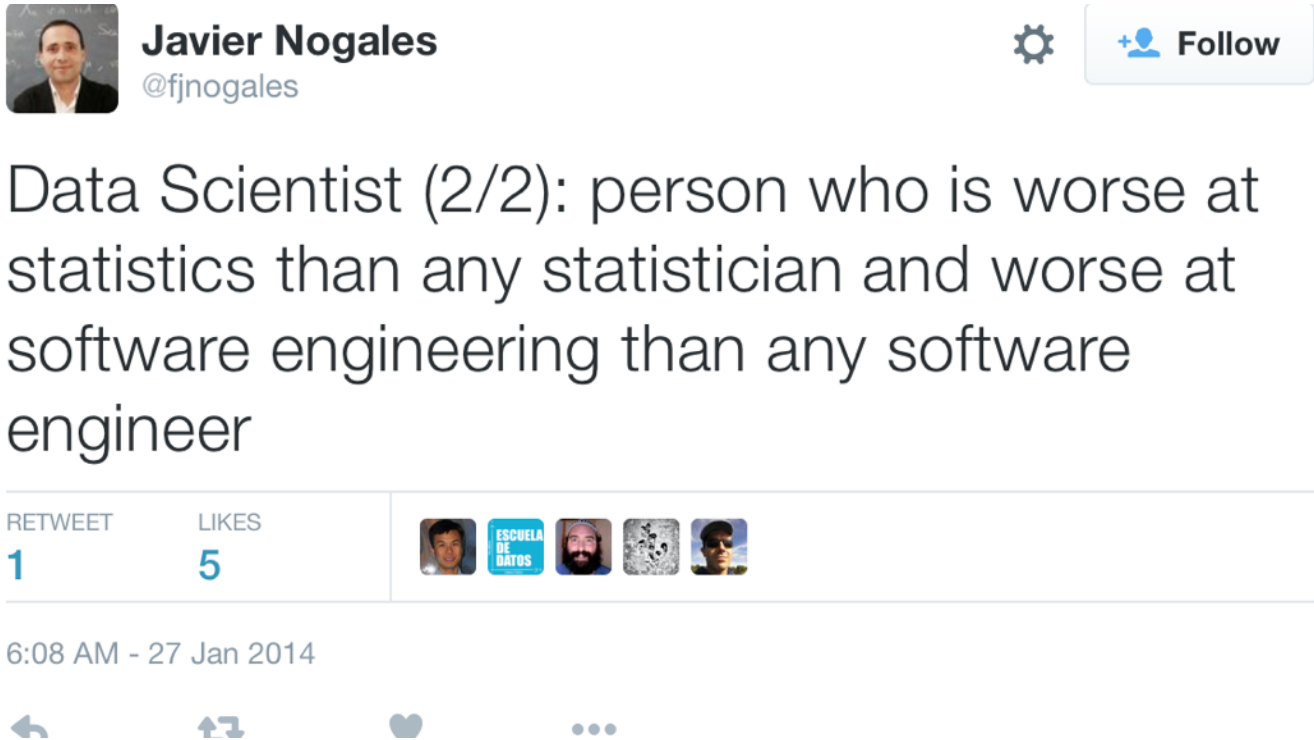
I. WHAT IS A DATA SCIENTIST?

WHAT IS A DATA SCIENTIST?

WHAT IS A DATA SCIENTIST?



WHAT IS A 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 IS YOUR DEFINITION?

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data "in production"**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

Hadley Wickham's advice for becoming a data scientist:

Statistical knowledge

"I think you need some knowledge of specific statistical/machine learning techniques, but a deep theoretical understanding is not that important. You need to understand the strengths and weaknesses of each technique… The vast majority of data science problems can be solved by a creative assembly of off-the-shelf techniques, and don't require new theory."

Source: https://gist.github.com/hadley/820f09ded347c62c2864

Hadley Wickham's advice for becoming a data scientist:

Programming skills

"You need to be fluent with either R or Python. There are other options, but none of them have the community that R and Python have, which means you'll need to spend a lot of time reinventing tools that already exist elsewhere."

Source: https://gist.github.com/hadley/820f09ded347c62c2864

Hadley Wickham's advice for becoming a data scientist:

Domain knowledge

"…A data scientist should be able to contribute meaningfully to any project, even if you're not intimately familiar with the specifics. I think this means you should be generally well read… and an able communicator. A good data scientist will help the real domain experts refine and frame their questions in a helpful way. Unfortunately I don't know of any good resources for learning how to ask questions."

Source: https://gist.github.com/hadley/820f09ded347c62c2864

Chris Volinsky (Columbia & AT&T Labs) on “Data Mining vs. Statistics”

- Snark: Data Mining = Statistics + Marketing
- Statistics is known for: **well-defined hypotheses** used to learn about a **specifically chosen population** studied using **carefully collected data** providing inferences with **well-known properties**.
- Data mining isn’t that careful. It is: **data-driven discovery** of **models and patterns** from **massive and observational data sets**.

Source: http://www2.research.att.com/~volinsky/DataMining/Columbia2011/Slides/Topic1-DMIntro.ppt

II. WHAT IS DATA SCIENCE?

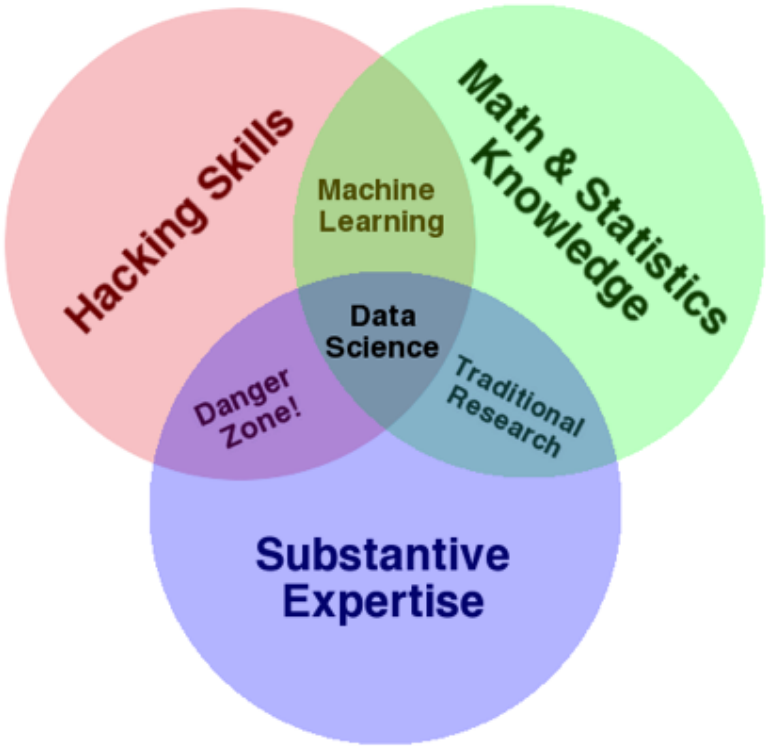
WHAT IS DATA SCIENCE?

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

WHAT IS DATA SCIENCE?

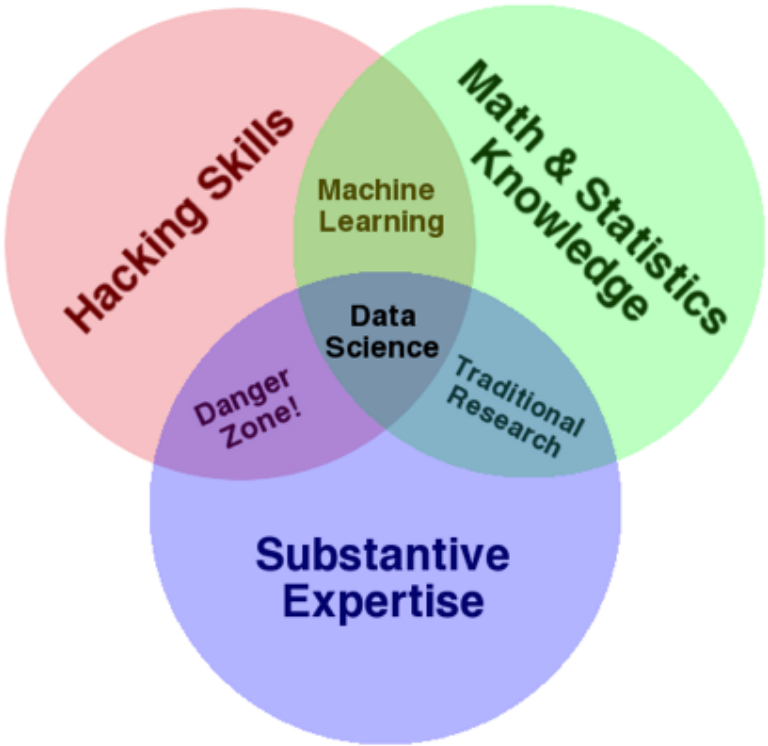
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject.

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

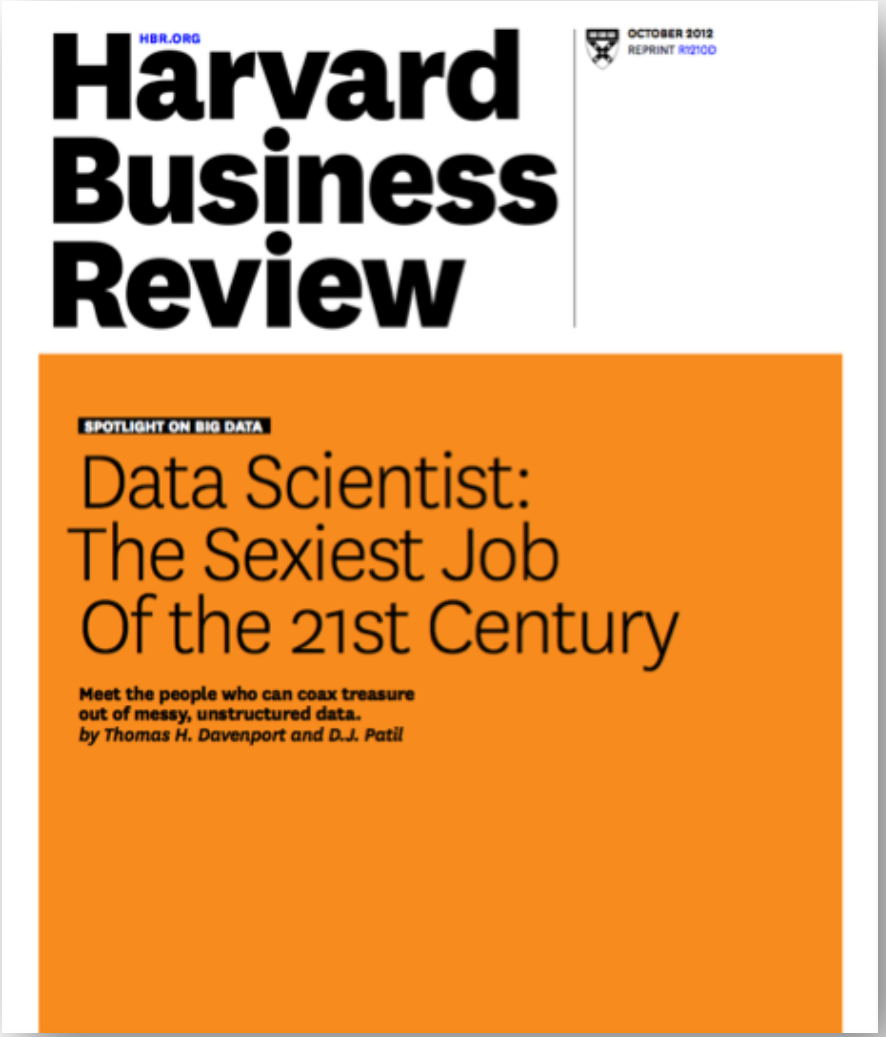
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

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.

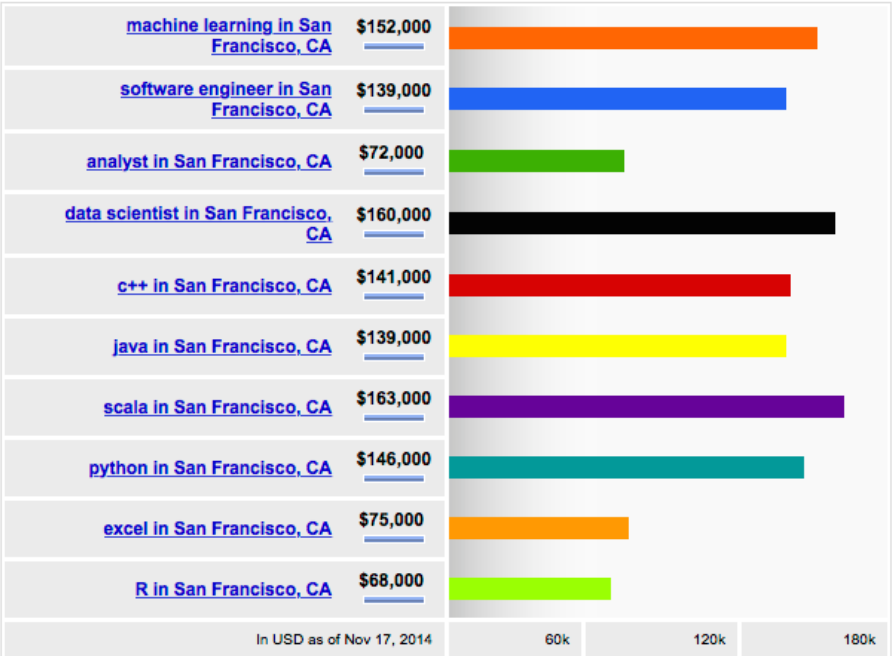


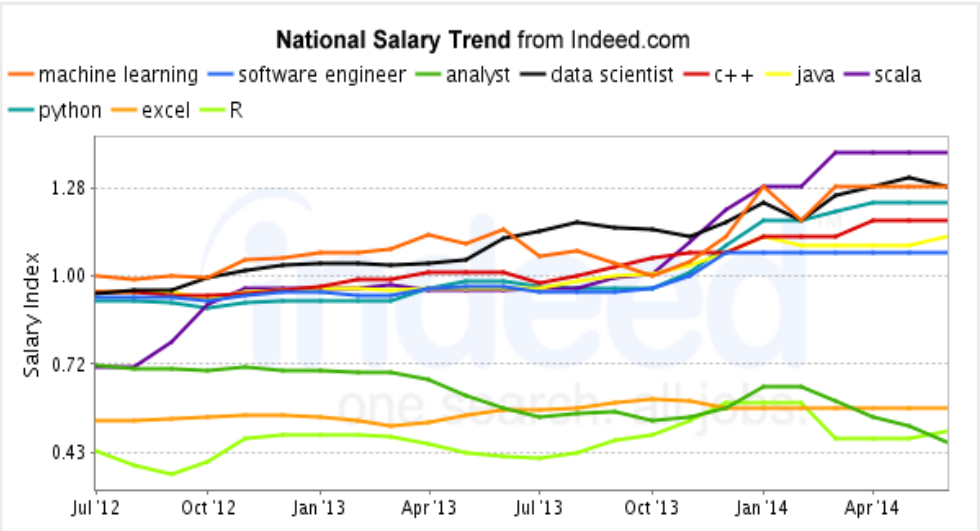
McKinsey estimates 140,000-190,000 shortage by 2018

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Hal Varian, Chief Economist at Google, The McKinsey Quarterly, January 2009

Average Salary of Jobs with Titles Matching Your Search





DATA SCIENTISTS WANTED

Data Scientist

Other Details

About this job

Job description

Facebook is seeking a Data Sci
to be comfortable working as a
have a keen interest in the stud
questions that help us build the

Responsibilities

- Work closely with a prod
- Answer product question
- Communicate findings to
- Drive the collection of ne
- Analyze and interpret the
- Develop best practices f
product engineering team

Requirements

- M.S. or Ph.D. in a releva
- Extensive experience so
- Comfort manipulating an
sources
- A strong passion for emp
- A flexible analytic approa
- Ability to communicate c
- Fluency with at least one
- Familiarity with relational
- Expert knowledge of an a
- Experience working with
plus (Map/Reduce, Hado

Data Scientist

Other Details

About this job

Job description

Job Description

PRINCIPAL DUTIES AND RESI

• Apply standard techniques and
deliver actionable business insig
• Work, under normal supervisio
and develop proposals to respon
• Following directed and specific
development. These include prel
selection, and model developme
• Independently research and te
assessing accuracy/fit/predictive
• Deliver results and presentatio
• Interact with customers to gath
insights. Likely presenting projec
preparation.

SKILLS

• Strong statistical foundation, w
methods.
• Proficiency in at least one of th
• Programming strength in at lea
• Natural ability to communicate
• Natural curiosity to research ar
problems.
• Team-oriented and collaborativ
• Innate customer orientation, wi

Data Scientist at LinkedIn

LinkedIn - Mountain View, CA

Posted 26 days ago

Apply on company website | Save

Other Details

About this job

Job description

Description

As a Senior Data Scientist at LinkedIn, you will develop innovative new technologies, features, and products that help connect the world's professionals to make them more productive and successful.

Our team applies machine learning techniques on social data to build products & features that reach over 200M professionals on LinkedIn. We build graph and text mining systems to tackle hard problems in areas like entity resolution, search relevance, recommendation algorithms, reputation & skills assessment, and network analysis.
Along with our team of data scientists, you'll work with product managers, designers, and engineers to build data driven features and products like LinkedIn Skills, Endorsements, and InMaps.
If you enjoy working with data to build products and solve hard problems in creative ways, you will fit right in.

Requirements:

- Strong background in Machine Learning, Statistics, Information Retrieval, or Graph Analysis
- Some experience working with large datasets, preferably using tools like Hadoop, MapReduce, Pig, or Hive
- 2+ years experience developing high quality software, contributions to open source projects are a plus
- Experience programming in an object oriented language (Java, C++, etc)
- Knowledge of scripting languages like Ruby or Python, familiarity with web frameworks a plus
- Comfortable with data analysis & visualization using tools like R, Matlab, or SciPy
- Critical thinking: ability to track down complex data and engineering issues, evaluate different algorithmic approaches, and analyze data to solve problems
- Creativity: you can conceive of new data driven products, features, and technologies
- Results: you prioritize, focusing on ideas and features that will have significant, measurable impact
- Planning & estimation: ability to set and meet your own project objectives & milestones
- Ability to coordinate effectively with team members in engineering, design, and product management
- Communicate results and progress internally and externally in meetings, presentations, and tech talks
- Masters, Phd, or equivalent experience in a quantitative field (computer science, physics, mathematics, bioinformatics, etc.)

Data Scientist

Apple - Santa Clara Valley - California -US

Posted 19 days ago

Apply on company website | Save

Other Details

About this job

Job description

Apple has a tremendous amount of data, and we have just scratched the surface in pattern detection, anomaly detection, predictive modeling, and optimization. There are many exciting problems to be discovered and solved. We encourage scientists to stay abreast of data mining research by attending conferences and working with academic faculty and students. We foster a collaborative work environment, but allow solution autonomy on projects.

The iTunes Engineering team has a proud tradition of delivering cutting-edge products in a competitive marketplace. We seek to maintain a challenging and rewarding environment where the best engineers and scientists can collaborate and produce real-world improvements in customers' online experience. Successful candidates will solve problems unique in scale and concept in the pursuit of new and original features.

Key Qualifications

- Strong working knowledge of data mining algorithms including decision trees, probability networks, association rules, clustering, regression, and neural networks.
- Familiarity with database modeling and data warehousing principles with a working knowledge of SQL.
- Familiarity with Big Data tools and techniques, including MapReduce, NoSQL stores, and unbounded stream processing.
- Creativity to go beyond current tools to deliver best solution to the problem
- Strong programming skills in Java, Python, or similar language
- Excellent interpersonal, written, and verbal communication skills
- Ability and comfort working independently and making key decisions on projects

Description

We are seeking an outstanding data mining scientist who is interested in designing, developing, and fielding data mining solutions that have direct and measurable impact to Apple. This person will work within and across teams to help identify viable data mining opportunities and then implement end to end analytical solutions. The role requires both a broad knowledge of existing data mining algorithms and creativity to invent and customize when necessary.

Education

Ph.D. in Data Mining, Machine Learning, Statistics, Operations Research or related field

M.S. in related field with 5 years experience applying data mining techniques to real business problems.

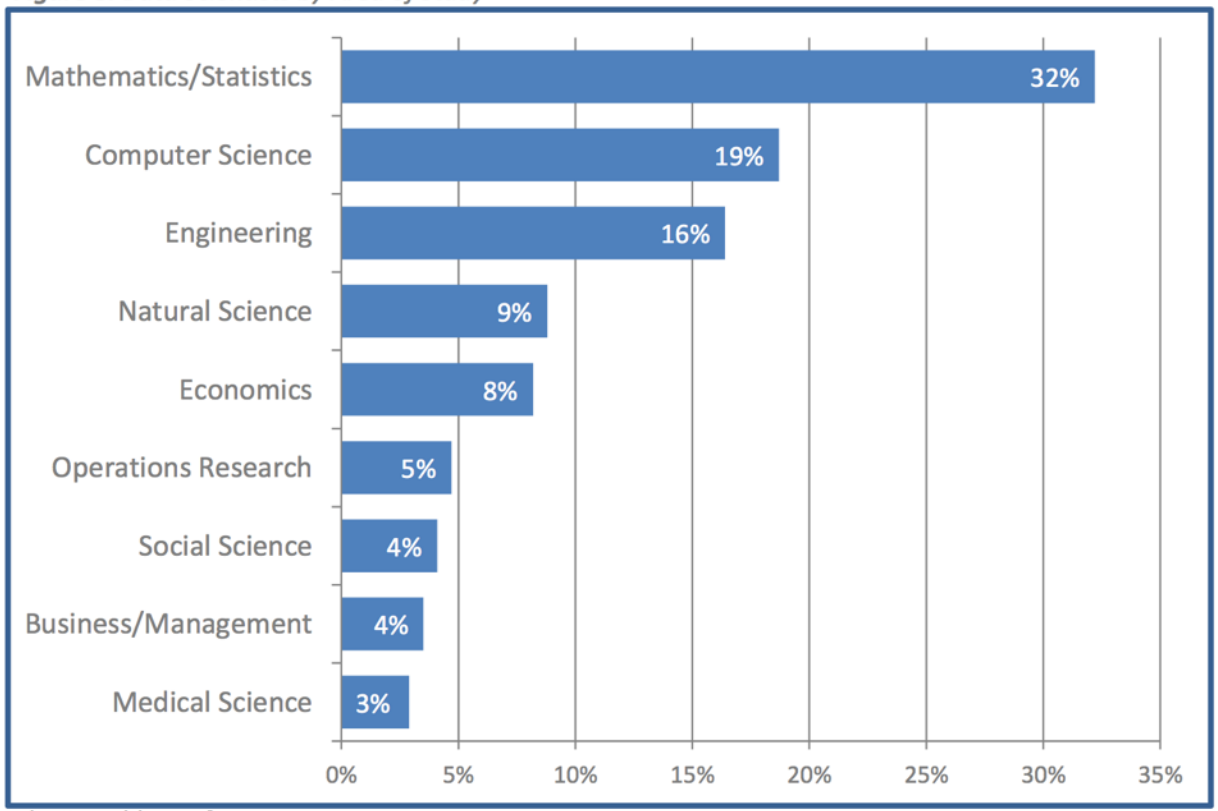
WHO USES DATA SCIENCE?

WHAT MAKES A GOOD DATA SCIENTIST?

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

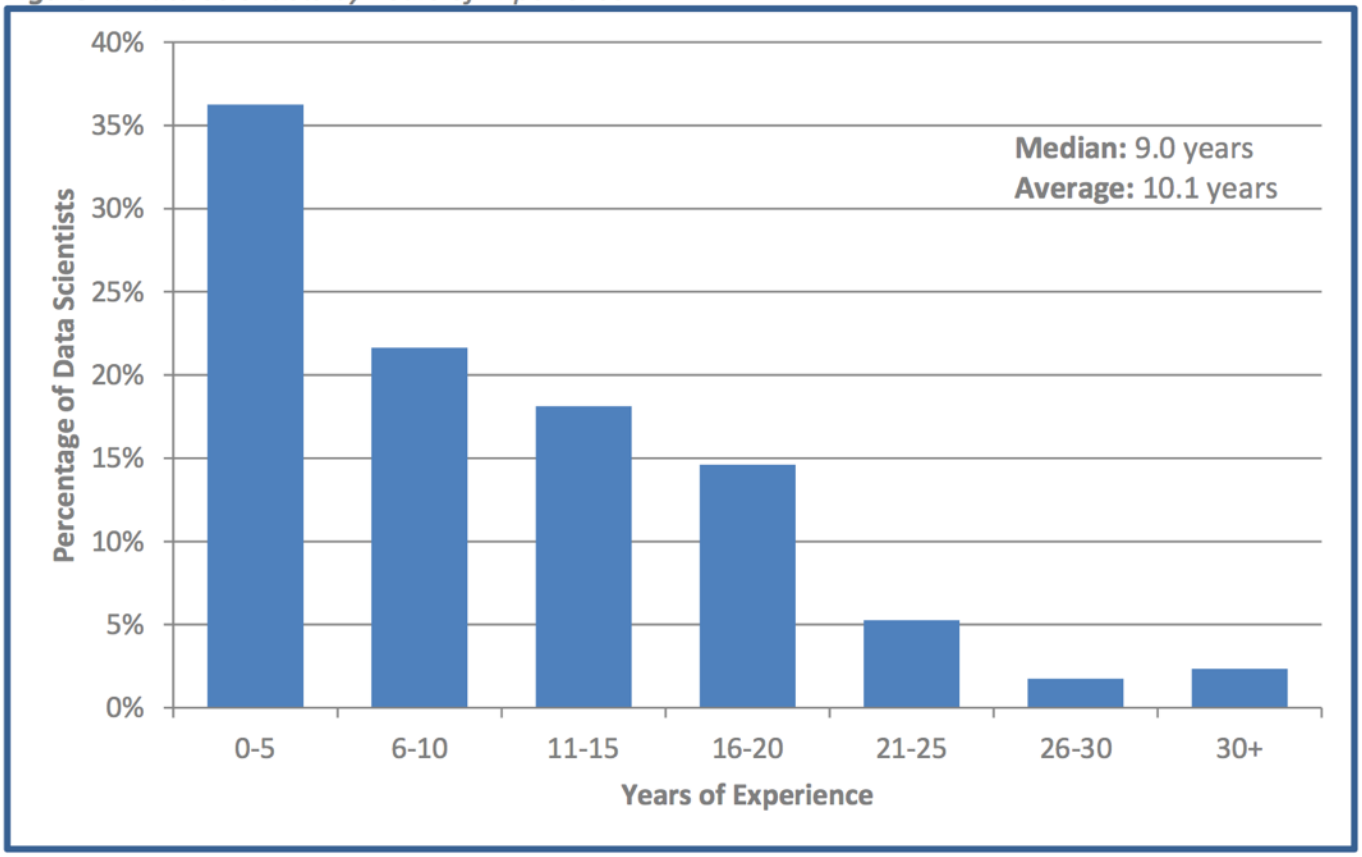
WHO ARE DATA SCIENTISTS?

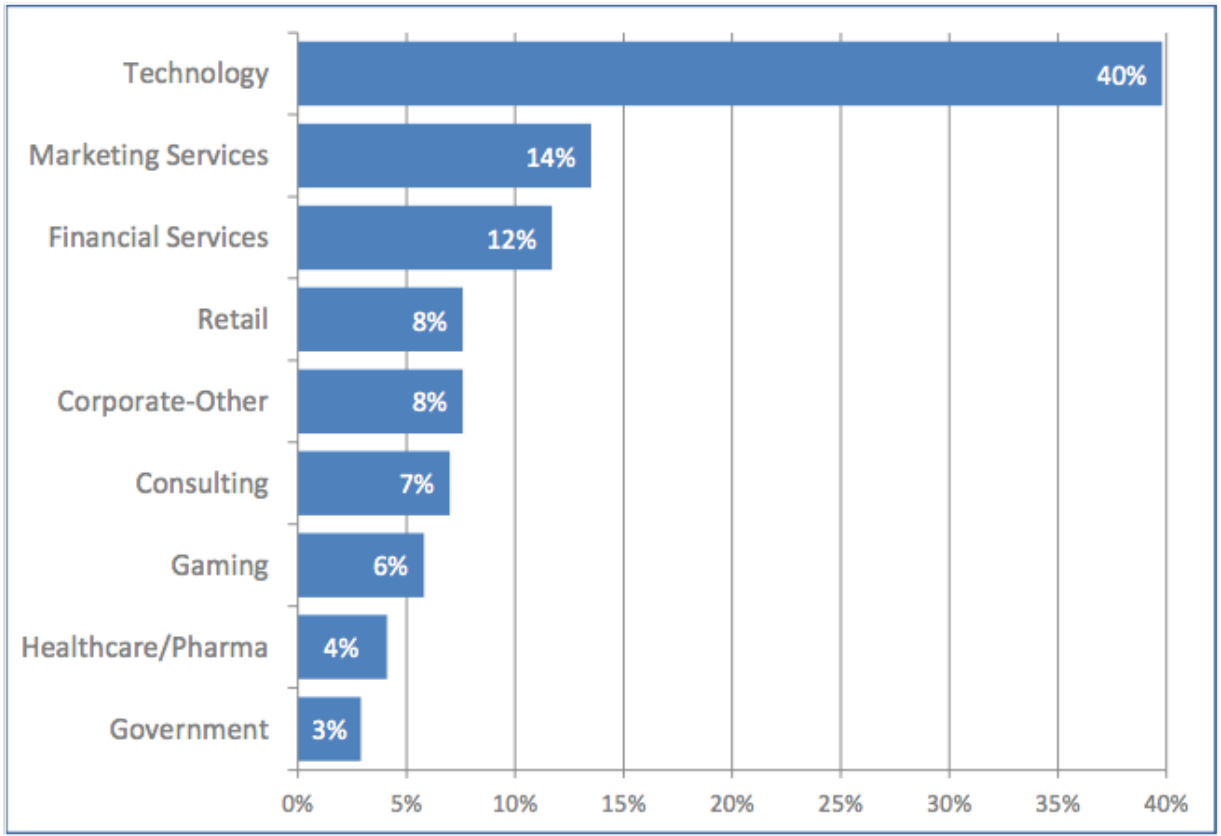
Figure 8. Data Scientists by Area of Study

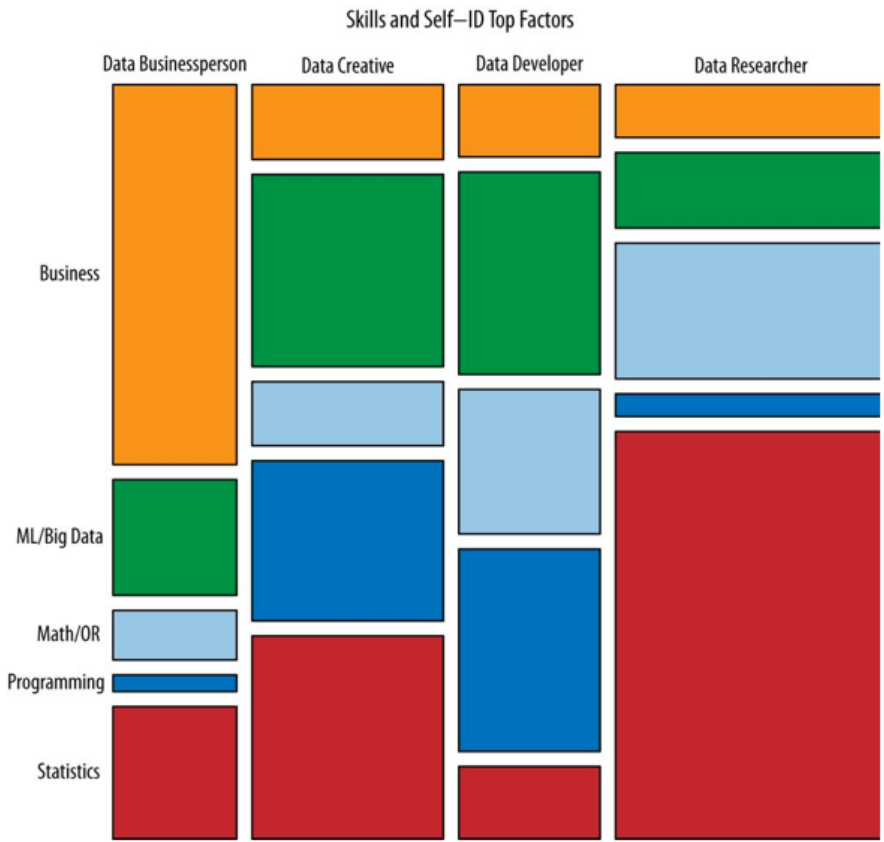


Source: Burtch Works Executive Recruiting study, 2014

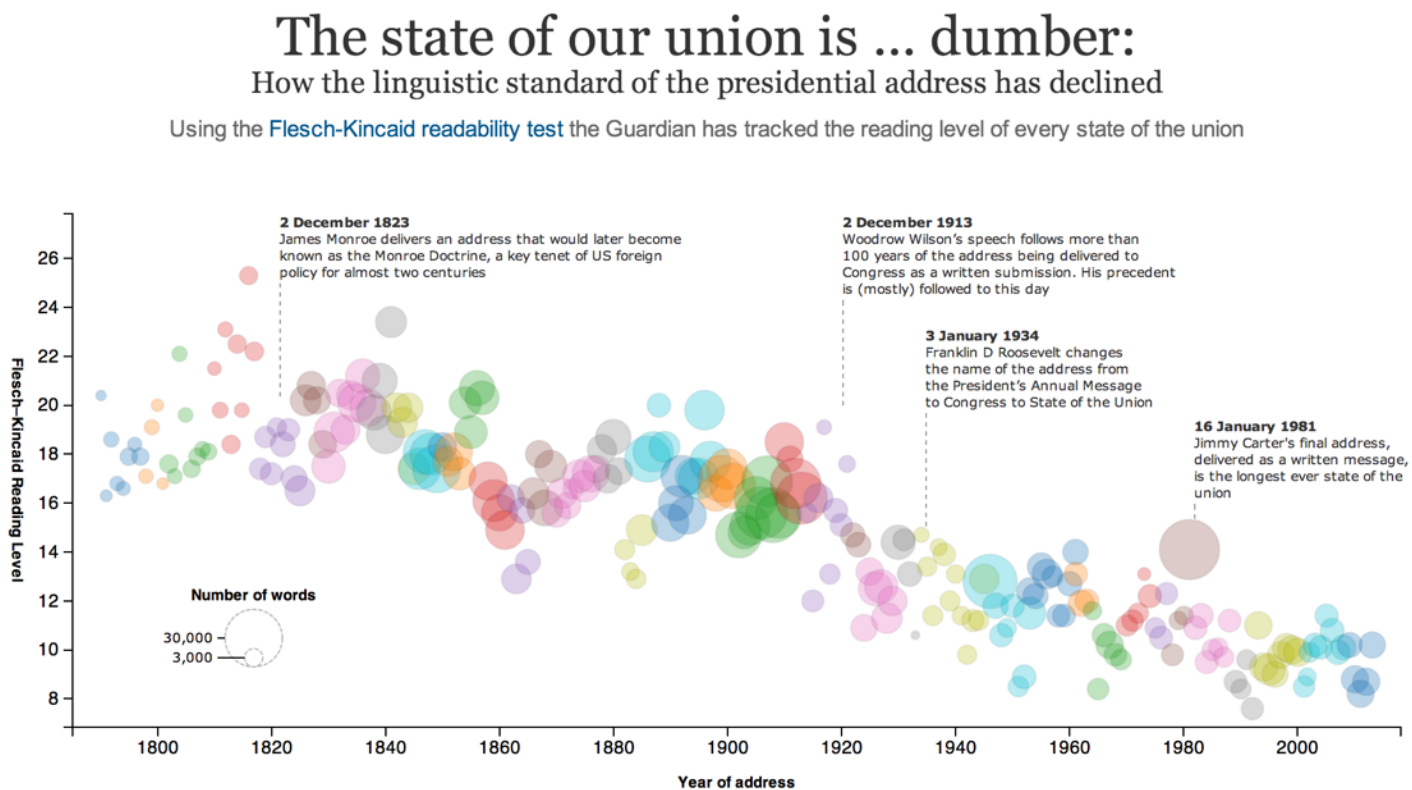
Figure 4. Data Scientists by Years of Experience




Technology
40%
Marketing Services
14%
Financial Services
12%
Retail
8%
Corporate-Other
8%
Consulting
7%
Gaming
6%
Healthcare/Pharma
4%
Government
3%
0%
5%
10%
15%
20%
25%
30%
35%
40%



"Analyzing the Analyzers" by Harlan Harris, Sean Murphy, and Marck Vaisman , O'Reilly Strata 2012


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
Number of words
30,000
3,000
Flesch-Kincaid Reading Level
26
24
22
20
18
16
14
12
10
8
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Year of address

Music + Data:

http://bit.ly/echonest

III. 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

Ted Johnson

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

THE DATA SCIENCE WORKFLOW

Ben Fry

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

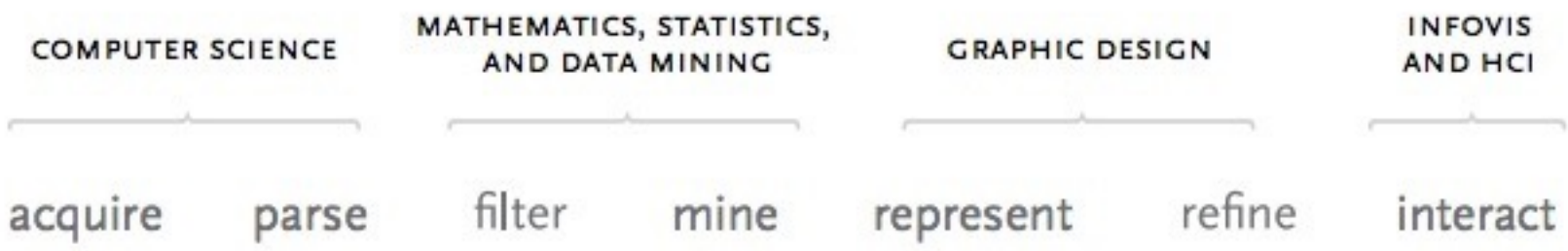
source: http://benfry.com/phd/dissertation-110323c.pdf

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

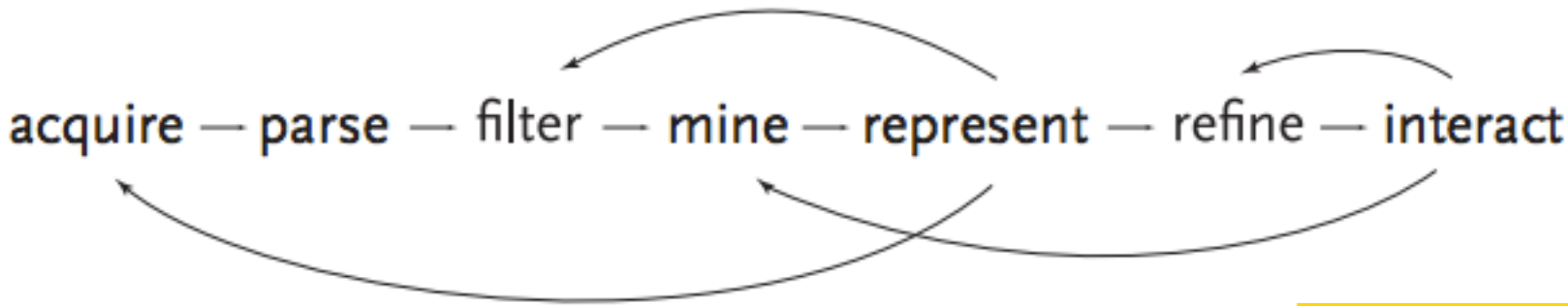
source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



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

THE DATA SCIENCE WORKFLOW



NOTE

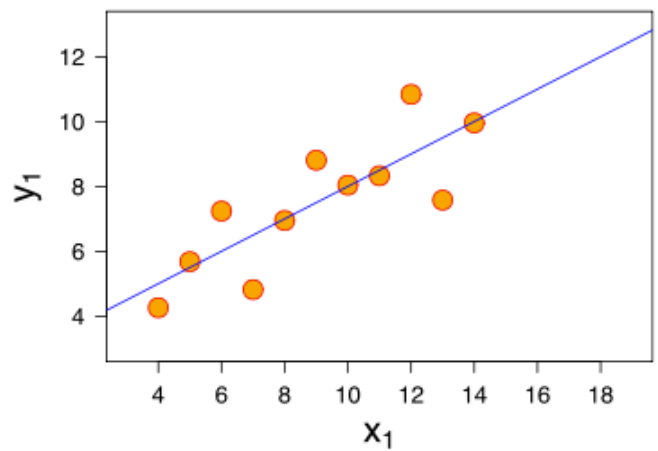
This diagram illustrates the *iterative* nature of problem solving

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

III. VISUALIZATIONS AS A MEDIUM

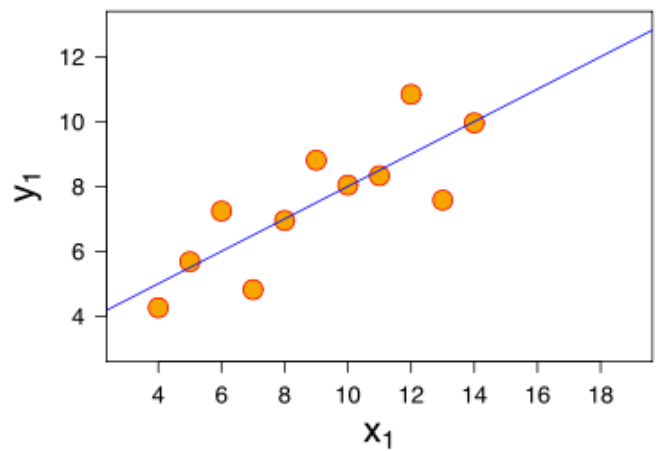
Consider the following dataset:

- *eleven (x, y) points*



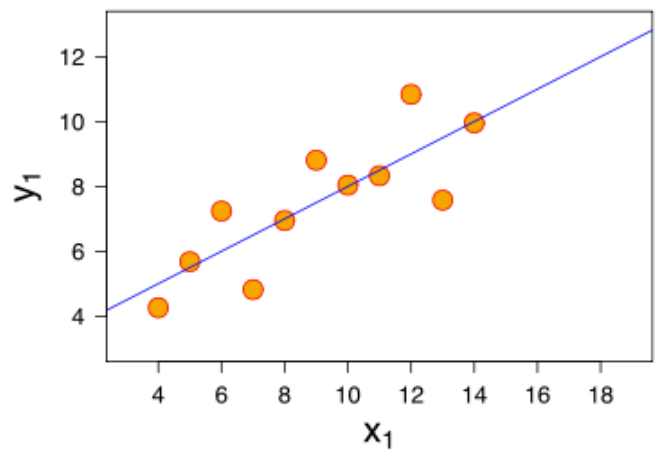
Consider the following dataset:

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



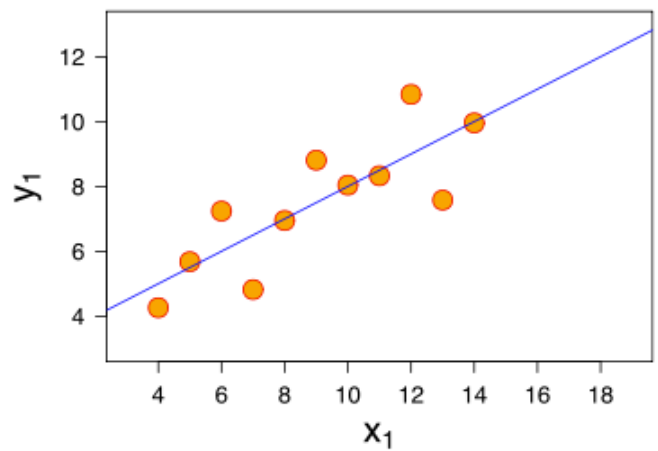
Consider the following dataset:

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



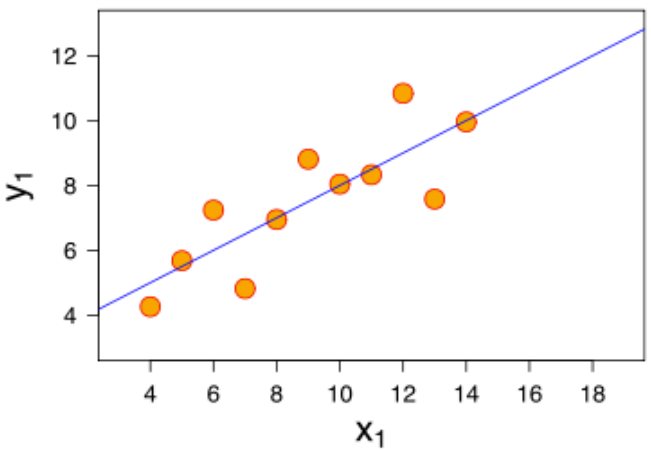
Consider the following dataset:

- *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:

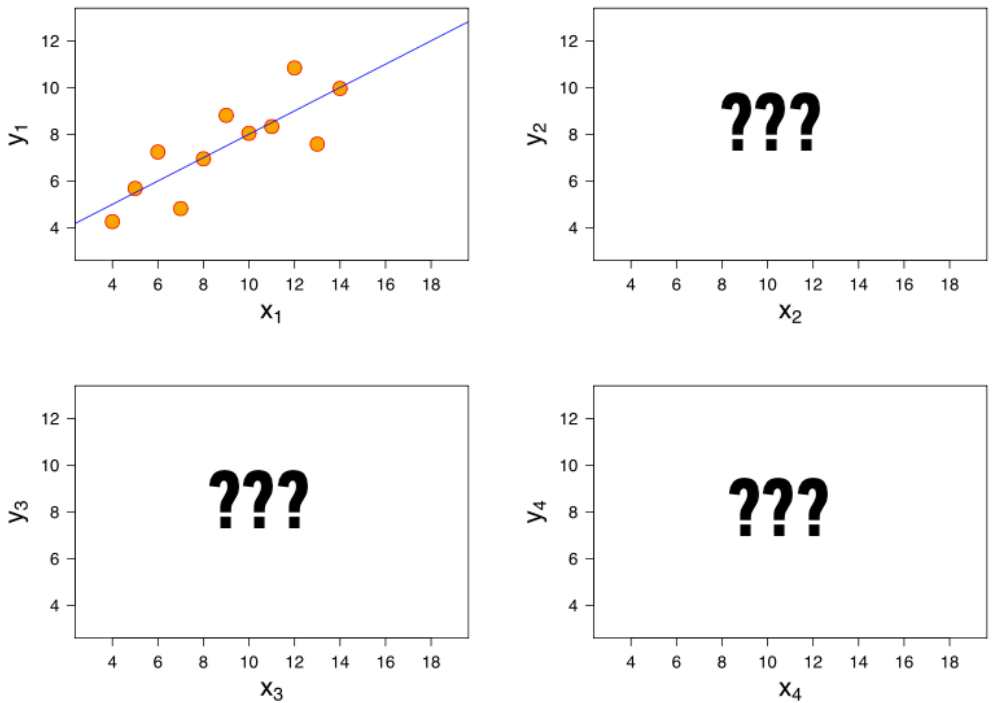
- *eleven (x, y) points*
- *mean of x = 9, mean of y = 7.5*
- *variance of x = 11, variance of y = 4.1*
- *correlation of x, y = 0.8*
- *line of best fit: y = 3.00 + 0.500x*



EXERCISE – WHY VISUALIZE DATA?

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

Q: how similar are these datasets?

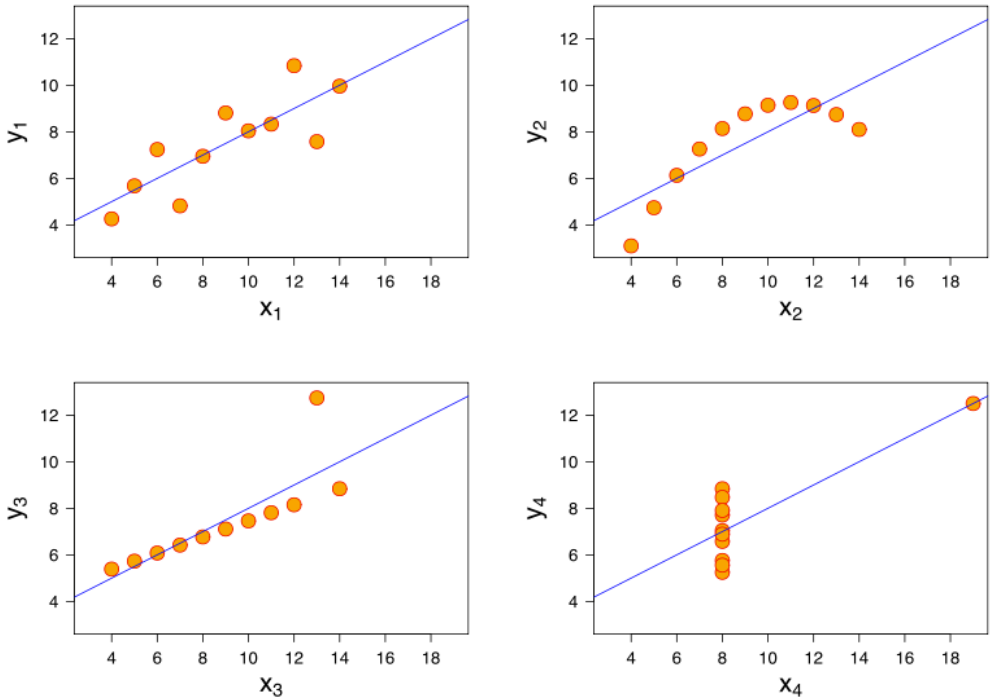


EXERCISE – WHY VISUALIZE DATA?

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

Look at your data!

V. EXERCISE

DISCUSSION