Data Mining 1 – Intro to Data Science and Syllabus

Anthony Breitzman, PhD
Computer Science Dept.
Data Analytics Program Coordinator
6/27/2018

Rowan University

Data Analytics in the News

- DA has gotten some bad press in recent days because…
  - We found out Facebook has been sharing your data with bad actors
  - One of those (Cambridge Analytica) used the data to identify hot-button issues to sway the election
  - Zuckerberg testified before congress and some kind of regulation is coming
- But in spite of the bad press. Analytics is not going away any time soon, because…
  - It can be used to identify cancerous tumors
  - It can be used to thwart terrorists and crime syndicates
  - It can be used to identify who is a credit risk and who isn’t
  - It can be used to identify who might buy a car in the next 6 months
  - It can be used to predict who might get diabetes in the next 3 years
  - It can be used to identify which free agent will most improve my team
  - Data Analytics is used in virtually every industry to some extent and new applications are coming every day

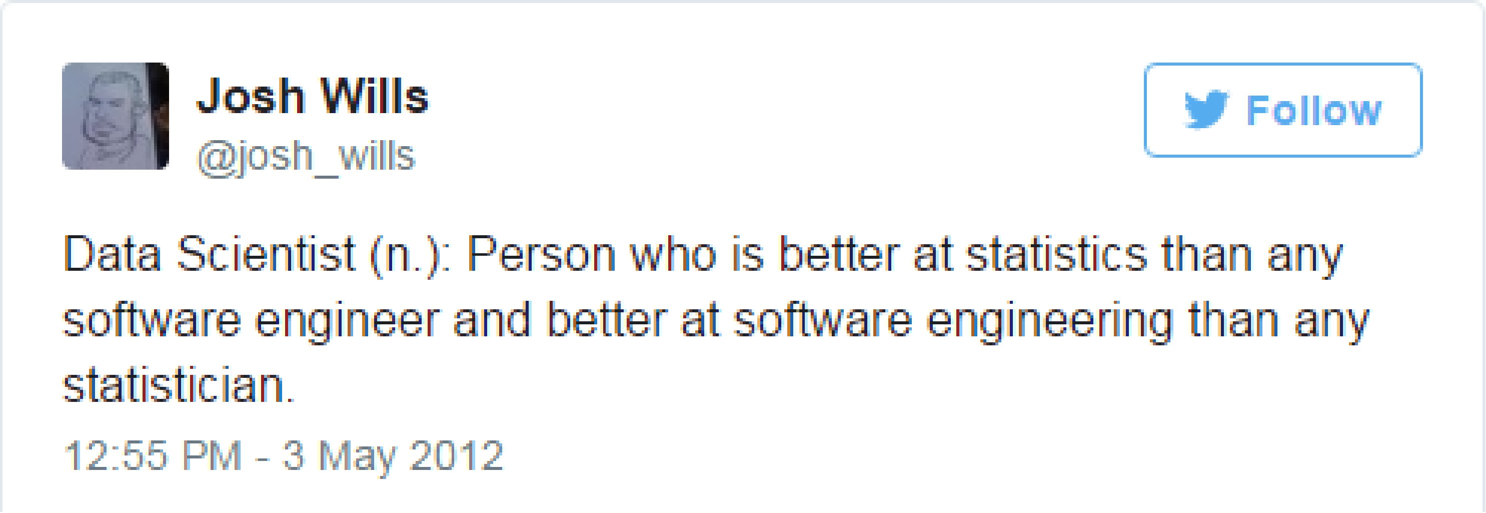
Everyone should know a little Data Science

- Often your career will take some twists and turns so some experience with multiple fields will be useful

- I’m a data scientist by accident. I was a math major and went for a Master’s in Math. My first employer paid me to get a PhD (not in math) but CS and in that job I mainly did database building and statistical data mining. At the time I wasn’t particularly interested in either, but it served me well.

- So it’s good to learn a little bit of everything and right now Data Science and Cybersec are the two hottest topics. (Sorry I don’t know anything about Cybersec)

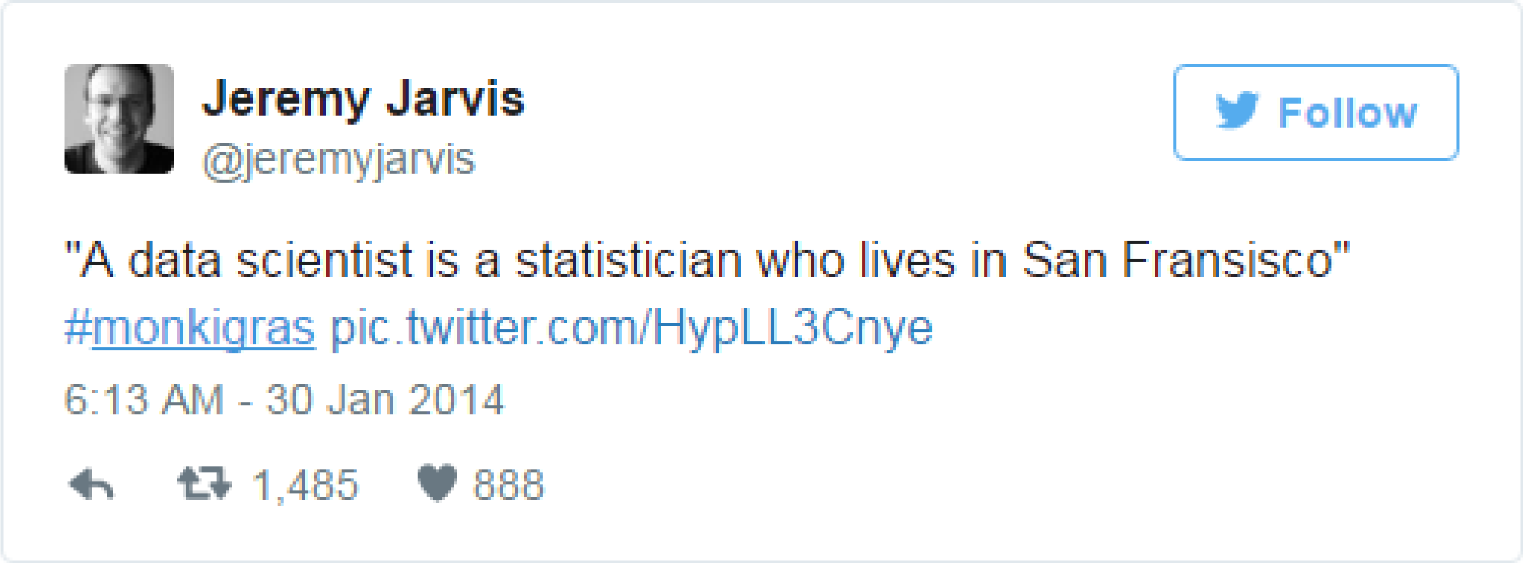
Data Scientist According to Twitter



Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

12:55 PM - 3 May 2012



Jeremy Jarvis
@jeremyjarvis

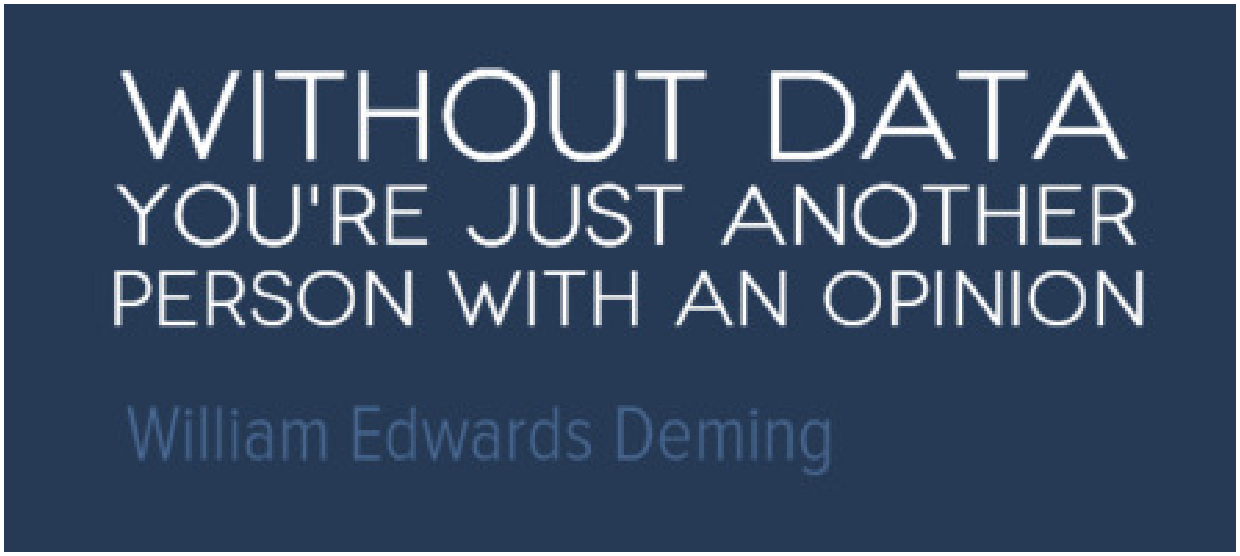
"A data scientist is a statistician who lives in San Fransisco" #monkigras pic.twitter.com/HypLL3Cnye

6:13 AM - 30 Jan 2014

1,485 ♥ 888

Breitzman’s view. What is data mining?

- Telling a story or solving a problem with data

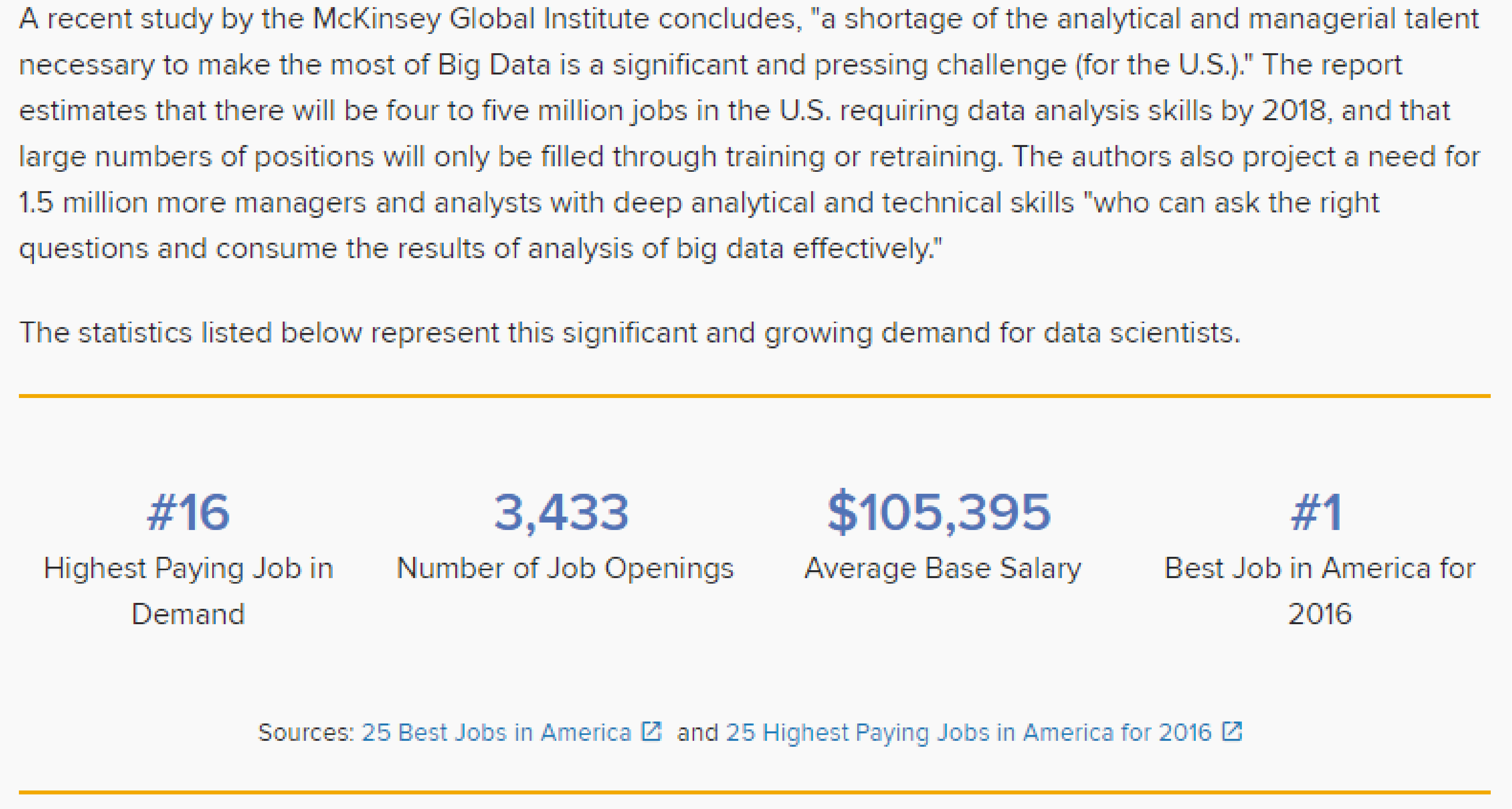


WITHOUT DATA
YOU'RE JUST ANOTHER
PERSON WITH AN OPINION

William Edwards Deming

- Sometimes the data scientist is the only person in the room with data

More on Data Science



A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)." The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that large numbers of positions will only be filled through training or retraining. The authors also project a need for 1.5 million more managers and analysts with deep analytical and technical skills "who can ask the right questions and consume the results of analysis of big data effectively."

The statistics listed below represent this significant and growing demand for data scientists.

#16	3,433	$105,395	#1
Highest Paying Job in Demand	Number of Job Openings	Average Base Salary	Best Job in America for 2016

Sources: 25 Best Jobs in America ↗ and 25 Highest Paying Jobs in America for 2016 ↗

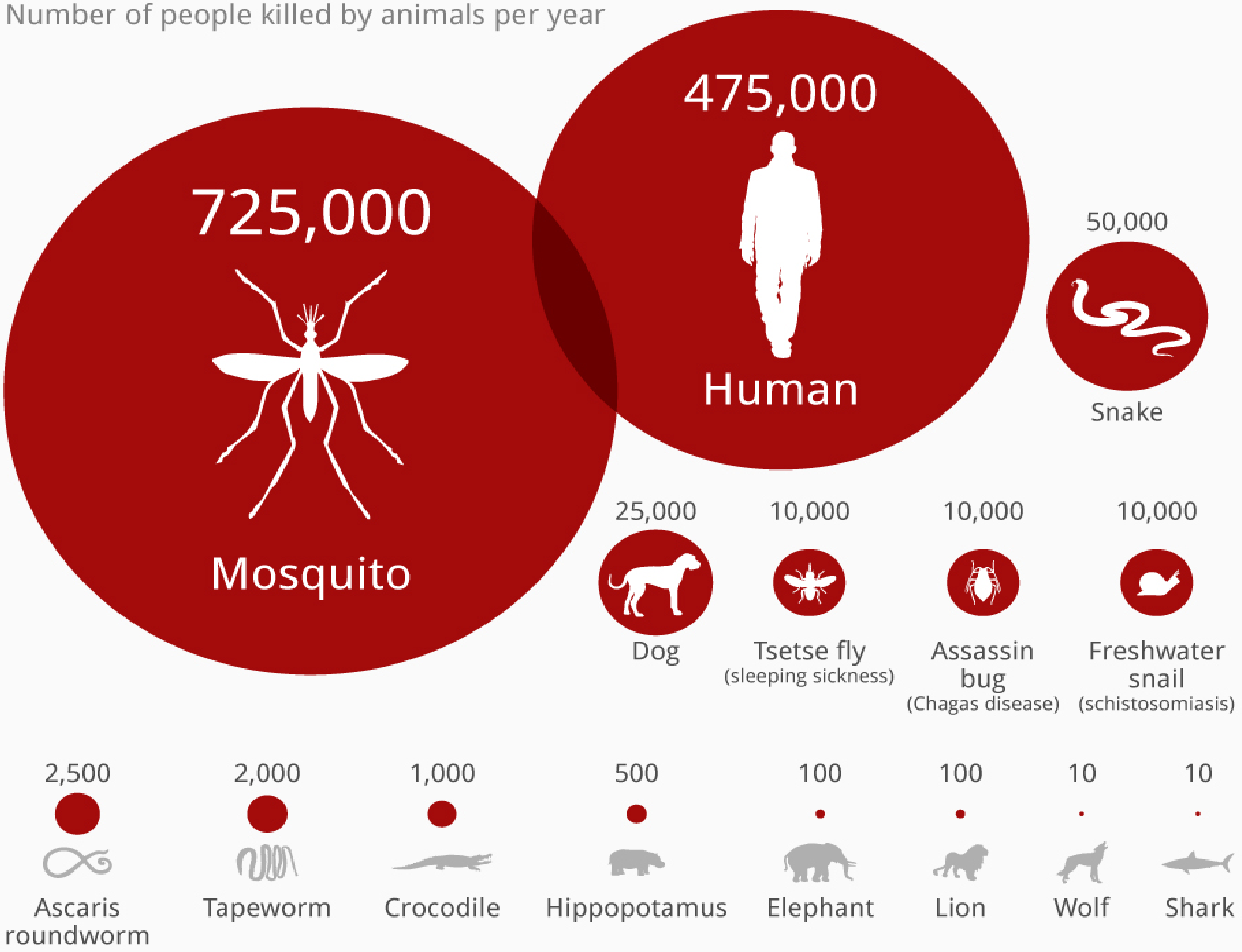
From: https://datascience.berkeley.edu/about/what-is-data-science/

What is the most dangerous animal to humans?

- Everyone has an opinion. Do we have any data?

The World's Deadliest Animals

Number of people killed by animals per year



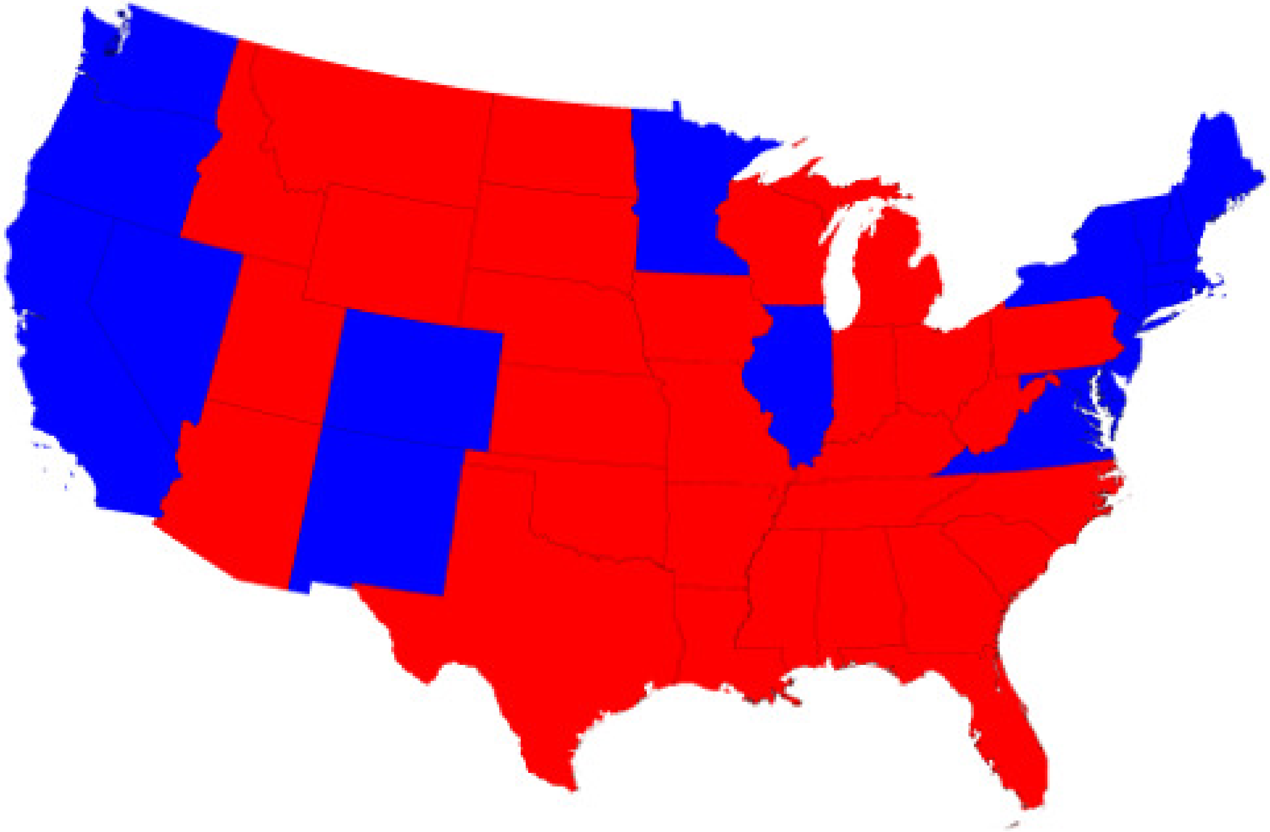
Source: Gatesnotes

statista

Sometimes telling a story with data is tricky

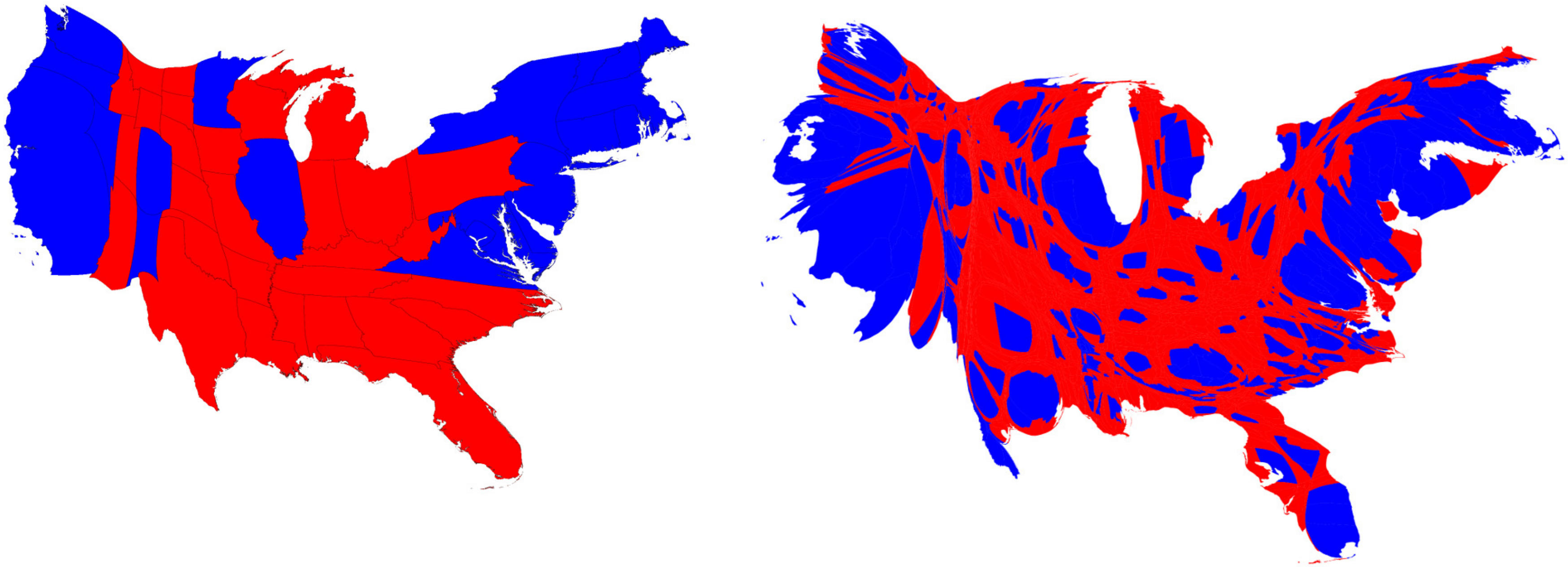
- 2 Charts on the presidential election (next 2 slides)

Typical Electoral Map (Misleading)



- Sure looks like Trump got a lot more votes than Clinton (actually 3 million fewer)

Cartogram Distorted by Voting Population



- State Level (Left) and County Level (Right)
- This is a great example of telling a story with data
- Source: University of Michigan

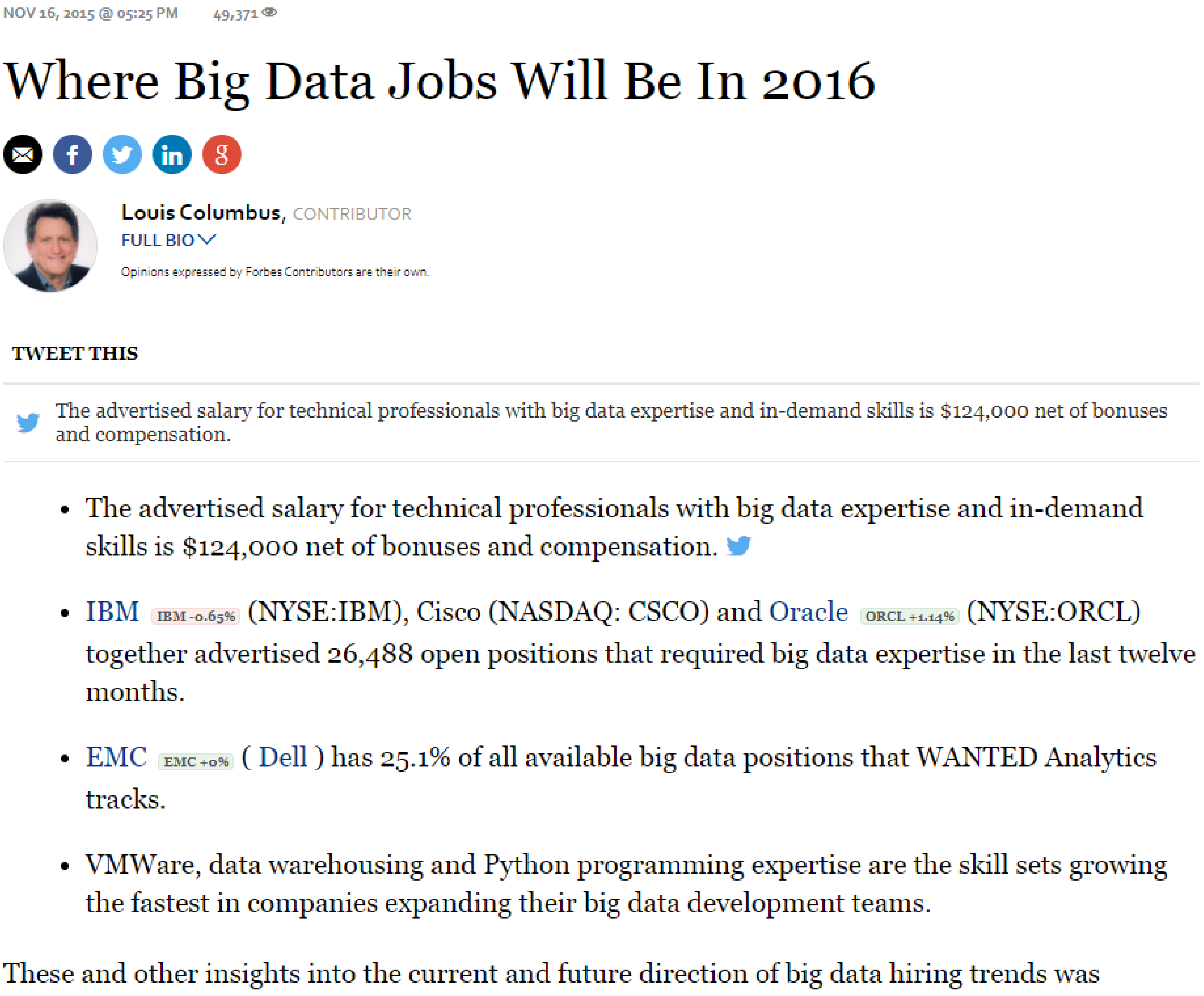
Data Mining/Data Analytics is Hot Right Now

- http://www.forbes.com/sites/louiscolumbus/2015/11/16/where-big-data-jobs-will-be-in-2016/#2588b107f7f1
  Snippet next page

- http://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move
  PDF included if link doesn’t work

- http://www.umuc.edu/analytics/about/big-data-job-growth-infographic.cfm
  Snippet 2 pages ahead

Forbes.com



NOV 16, 2015 @ 05:25 PM 49,371

Where Big Data Jobs Will Be In 2016

Louis Columbus, CONTRIBUTOR
FULL BIO

Opinions expressed by Forbes Contributors are their own.

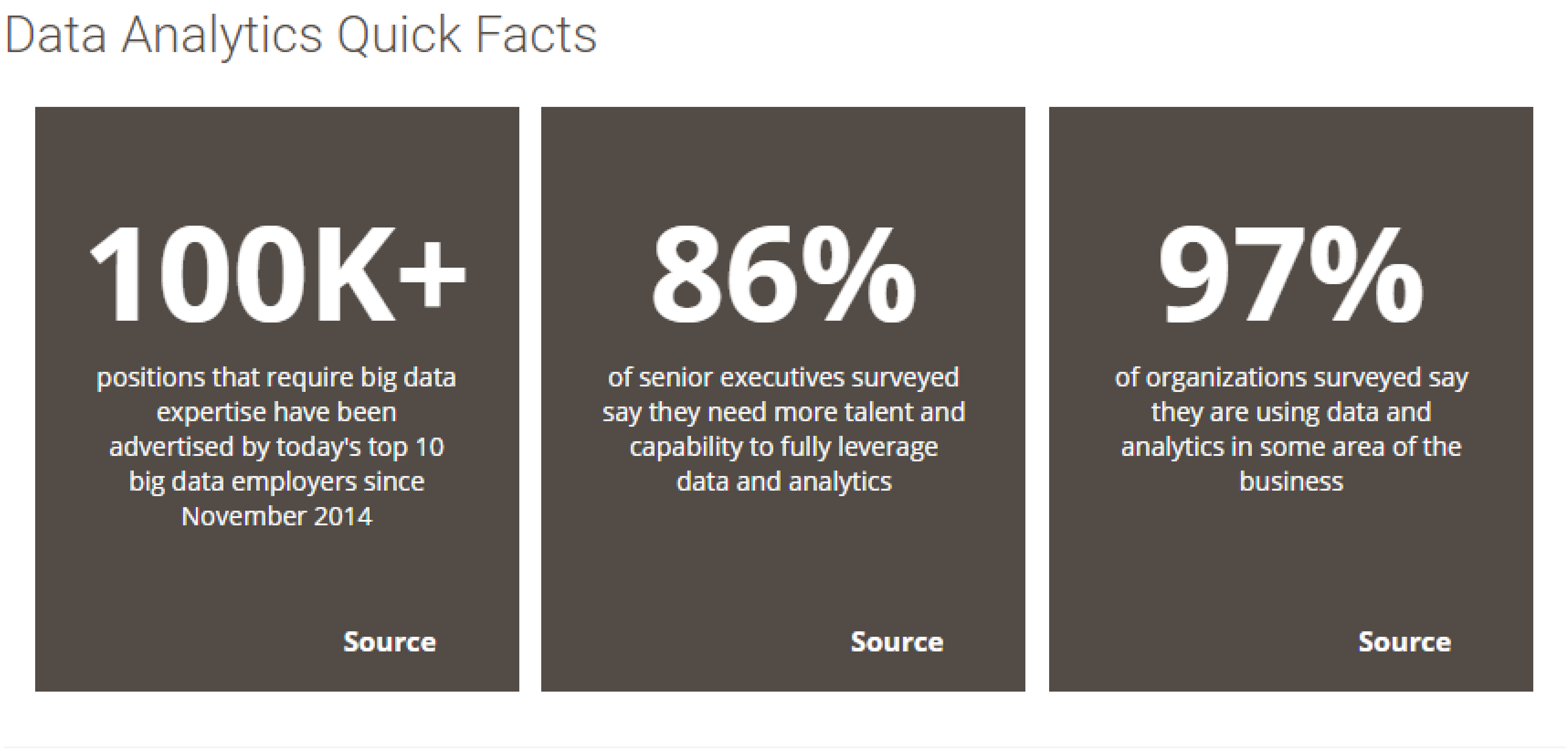
TWEET THIS

The advertised salary for technical professionals with big data expertise and in-demand skills is $124,000 net of bonuses and compensation.

- The advertised salary for technical professionals with big data expertise and in-demand skills is $124,000 net of bonuses and compensation.
- IBM IBM -0.65% (NYSE:IBM), Cisco (NASDAQ: CSCO) and Oracle ORCL +1.14% (NYSE:ORCL) together advertised 26,488 open positions that required big data expertise in the last twelve months.
- EMC EMC +0% (Dell) has 25.1% of all available big data positions that WANTED Analytics tracks.
- VMWare, data warehousing and Python programming expertise are the skill sets growing the fastest in companies expanding their big data development teams.

These and other insights into the current and future direction of big data hiring trends was

http://www.umuc.edu/academic-programs/data-analytics/index.cfm



Data Analytics Quick Facts

100K+

positions that require big data expertise have been advertised by today's top 10 big data employers since November 2014

Source

86%

of senior executives surveyed say they need more talent and capability to fully leverage data and analytics

Source

97%

of organizations surveyed say they are using data and analytics in some area of the business

Source

Why Data Mining Now?

- Data is being collected at an unprecedented rate
- Every beep you hear at a supermarket register is a barcode being read and a purchase being entered into a database
- Every web site you enter or tweet you send is being recorded in a database somewhere
- As early as 1984 in his book Megatrends, John Naisbitt observed that “we are drowning in information but starved for knowledge”
- The problem today is not lack of data, (we have too much in many cases), but the lack of human trained **data analysts** that can make sense of the data and turn it into knowledge

Data Mining Why Now (II)?

- Storage used to be expensive; now it’s almost free
  - Huge databases are now possible that would have been cost prohibitive 20 years ago
  - Walmart stores more than 20 million transactions per day

- Transactions are automatically captured
  - Bar Codes , Scanners, Mouse clicks, Location data from GPS and cell phones
  - 20 years ago, these things would have to be entered by hand

Idea Behind this Course

- With all of the tools now available, data mining is easy to do badly

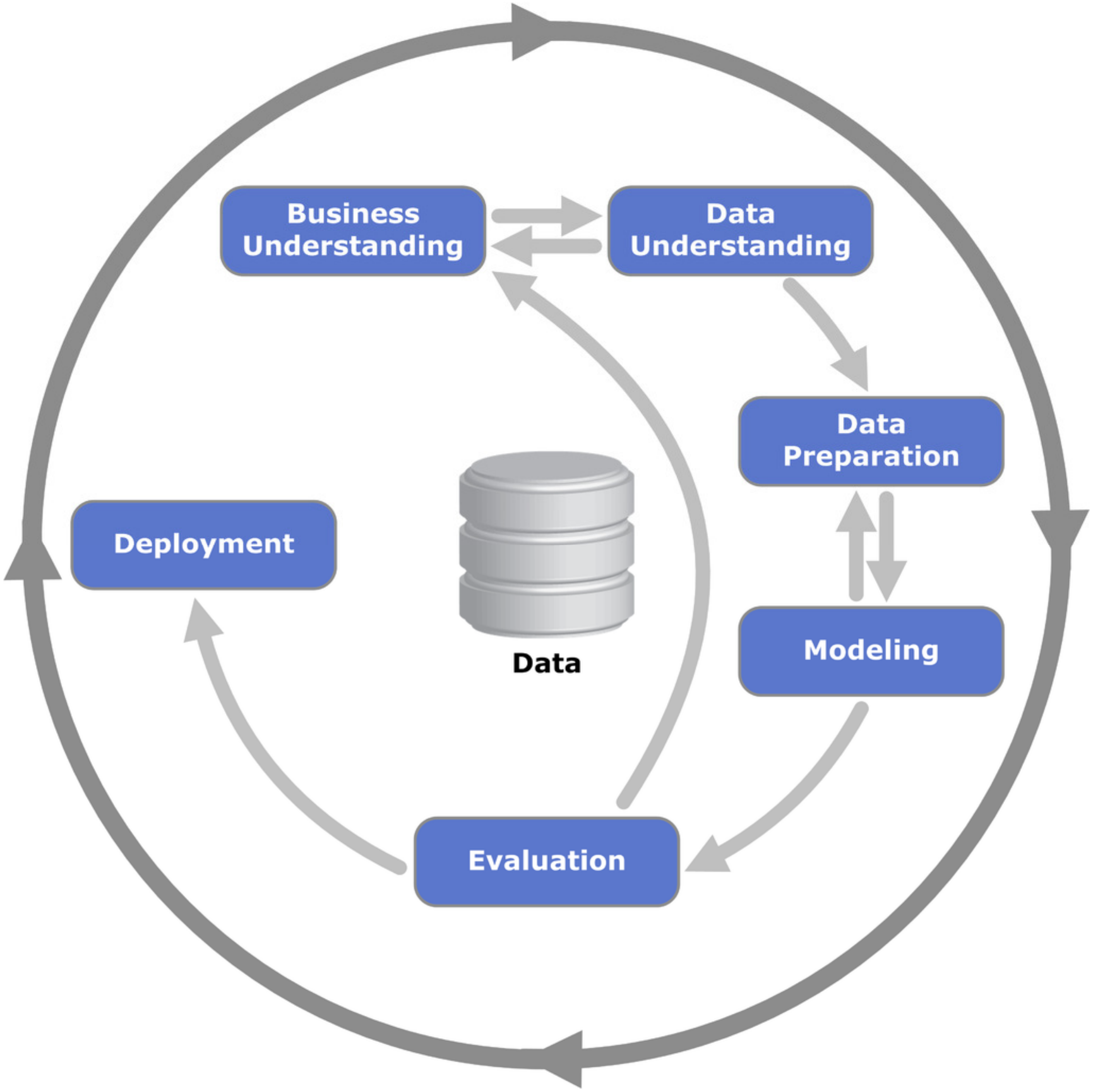
- We will learn how the various algorithms work so that we know the limits and shortcomings of the results

CRISP-DM

- The Cross-Industry Standard Process (CRISP-DM) was developed in 1996 by analysts from DaimlerChrysler, SPSS, and NCR.

- According to CRISP, Data Mining has 6 Phases

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
  (Visual on next page if link doesn’t work)


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM (II)

- The point is that data mining should be a an iterative and adaptive process where human intervention is critical to getting the machines to do what we want

- The computer and algorithms are tools that can be used well or badly

Examples

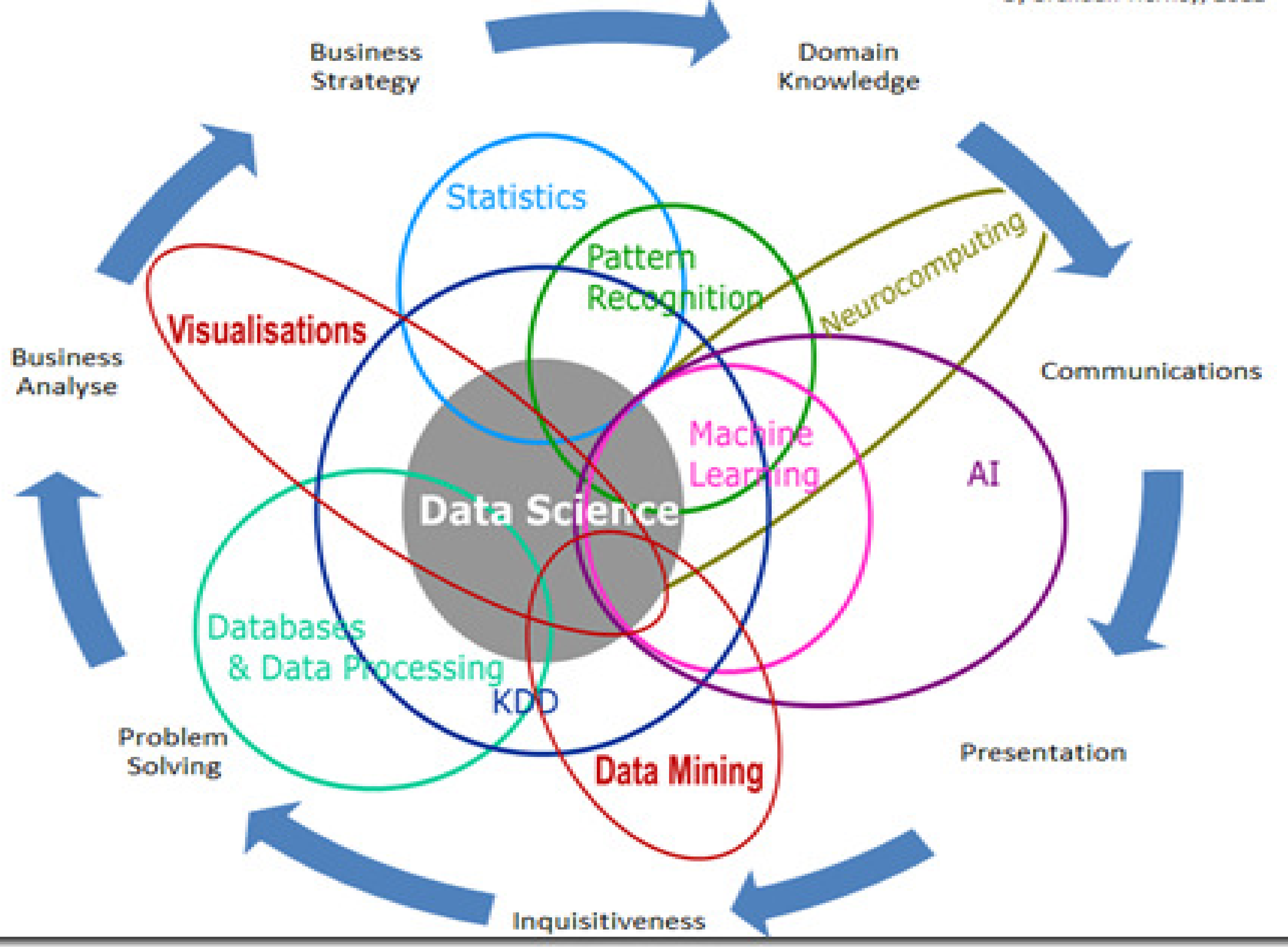
- Credit Card Fraud
  - Data mining can identify unusual transactions that might be fraud
  - $3 charge at a gas pump is probably a stolen credit card. Data Mining discovered this result.
- How does Amazon suggest things you might like? Data Mining
- Target predicts pregnancy of young girl and sends coupons before the family knows.
  - Data mining has determined customers switch to scent free sunscreen and scent free soaps when they become pregnant
- Telecoms and Cable Companies need new customers, but want to avoid churn prone customers
  - By comparing with previous churn prone customers, a neural network can predict who is a likely churn candidate very early

Will We Become Data Mining Experts?

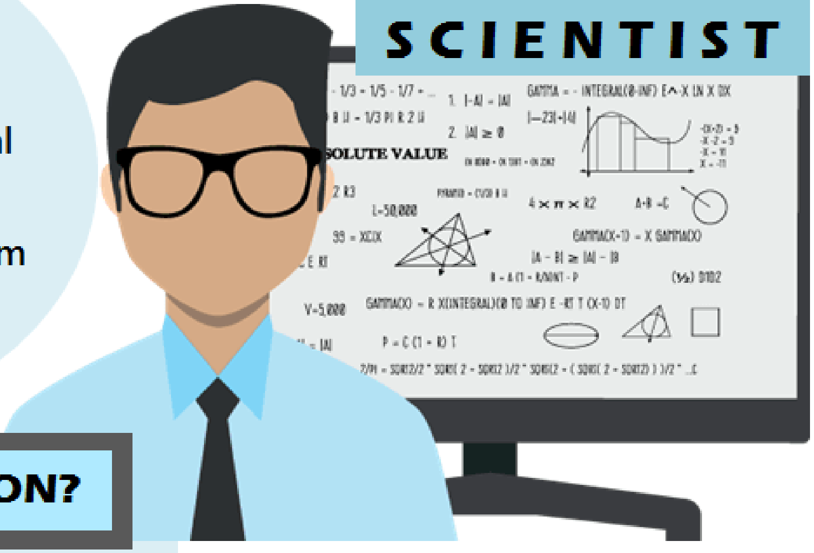
- Data Mining is a huge field
- I’ve created 3 courses (DM I, DM II, and Text Mining) and they don’t even cover all possible topics
- You will learn enough to do useful projects for yourself and your employer
- You will also have a good enough foundation that you can pick up a book and learn other methods/algorithms
- It’s a broad enough field that you can create a niche and carve out a pretty good career in Data Analytics or Competitive Intelligence

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



DATA SCIENTIST



WHO AM I?

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning / insights from massive data sets.

WHAT DO I DO?

1. I cleanse existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

WHAT DO I RELY ON?

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if Analysis

THE PROCESS I FOLLOW

Define Problem | **Structure Data** | **Use Programming Language**

WHAT DO I EARN?

After oil & gas geologists, mine is the 2nd highest paid job in the world!

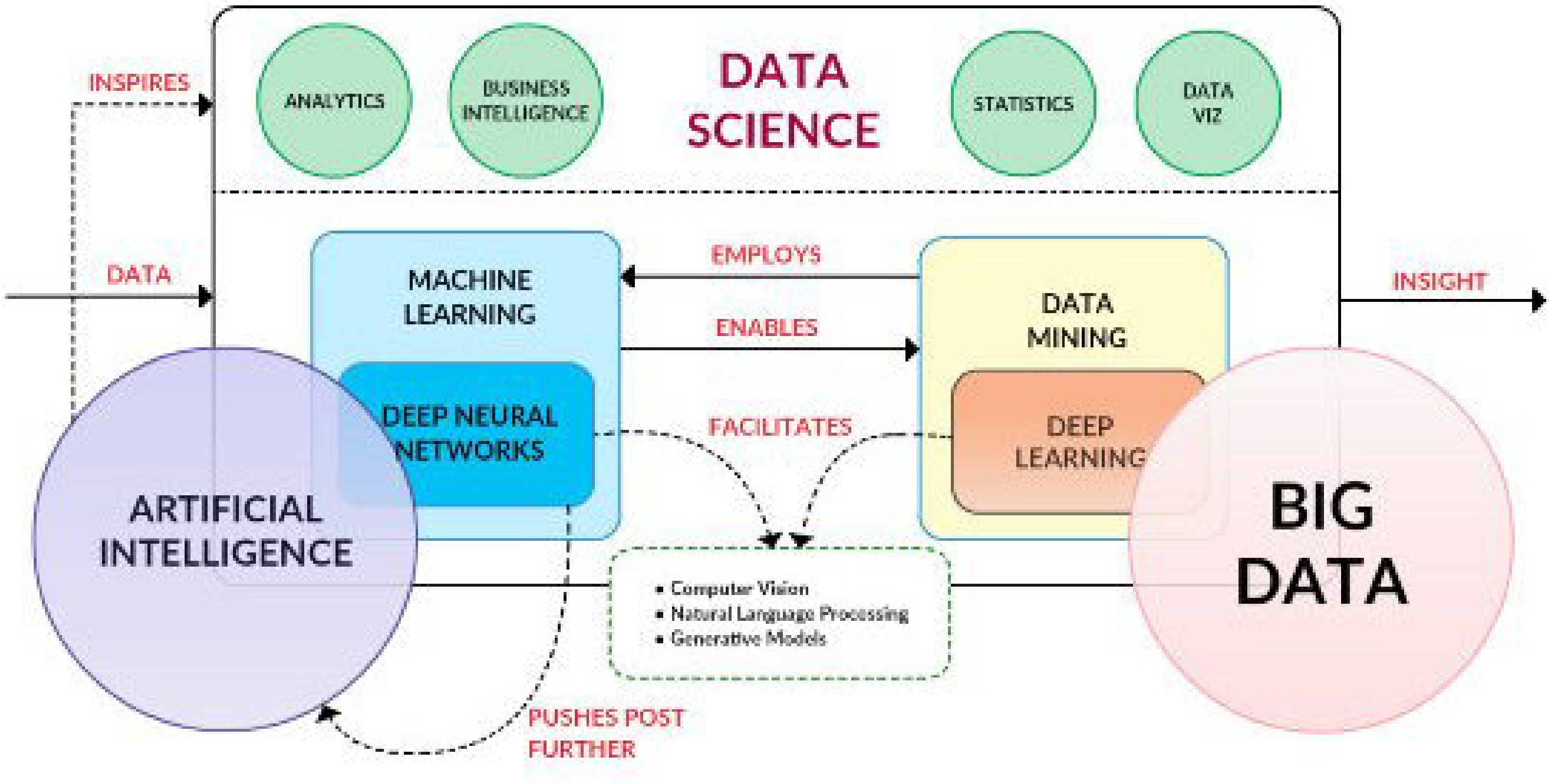
$ 100,000 to 150,000

HOW DO I HELP ORGANIZATIONS TODAY?

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services

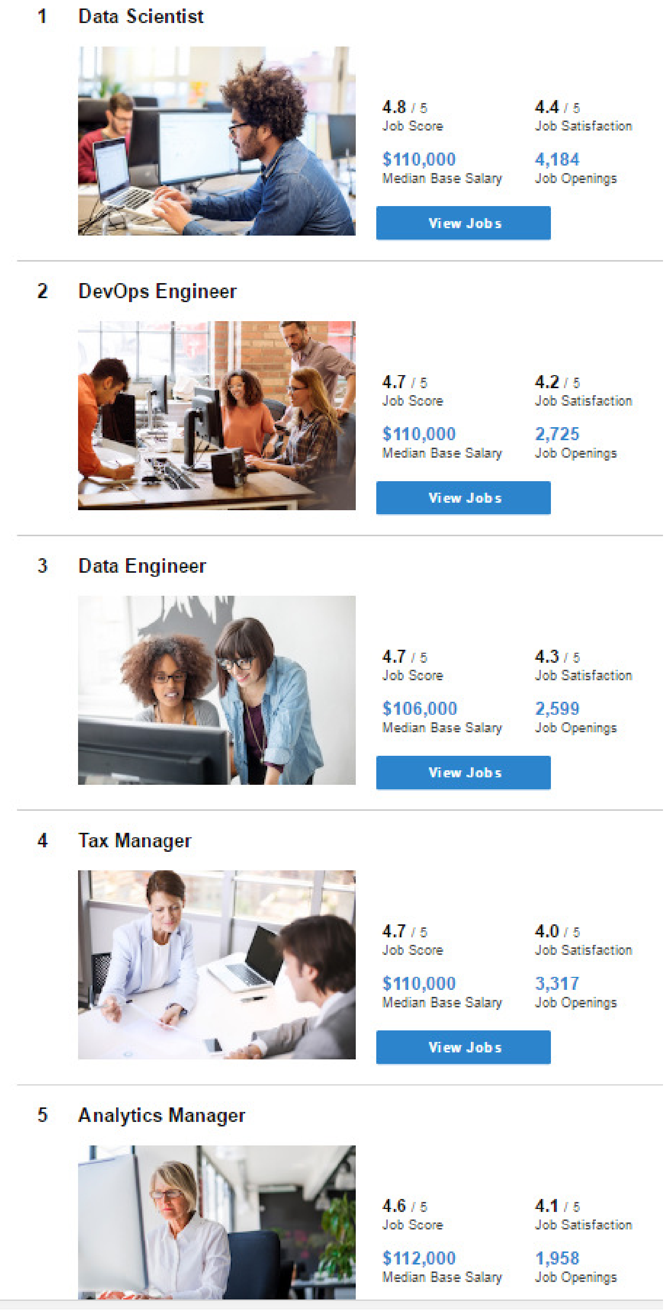
https://upxacademy.com/data-science-data-scientist/

Interactions Between Data Science, Machine Learning, Data Mining



Source: Kdnuggets.com

Top 5 Jobs in 2017 (Glassdoor.com)



1 Data Scientist

4.8 / 5 Job Score

4.4 / 5 Job Satisfaction

$110,000 Median Base Salary

4,184 Job Openings

2 DevOps Engineer

4.7 / 5 Job Score

4.2 / 5 Job Satisfaction

$110,000 Median Base Salary

2,725 Job Openings

3 Data Engineer

4.7 / 5 Job Score

4.3 / 5 Job Satisfaction

$106,000 Median Base Salary

2,599 Job Openings

4 Tax Manager

4.7 / 5 Job Score

4.0 / 5 Job Satisfaction

$110,000 Median Base Salary

3,317 Job Openings

5 Analytics Manager

4.6 / 5 Job Score

4.1 / 5 Job Satisfaction

$112,000 Median Base Salary

1,958 Job Openings

Where does R fit In? It’s the top language used by Data Scientists

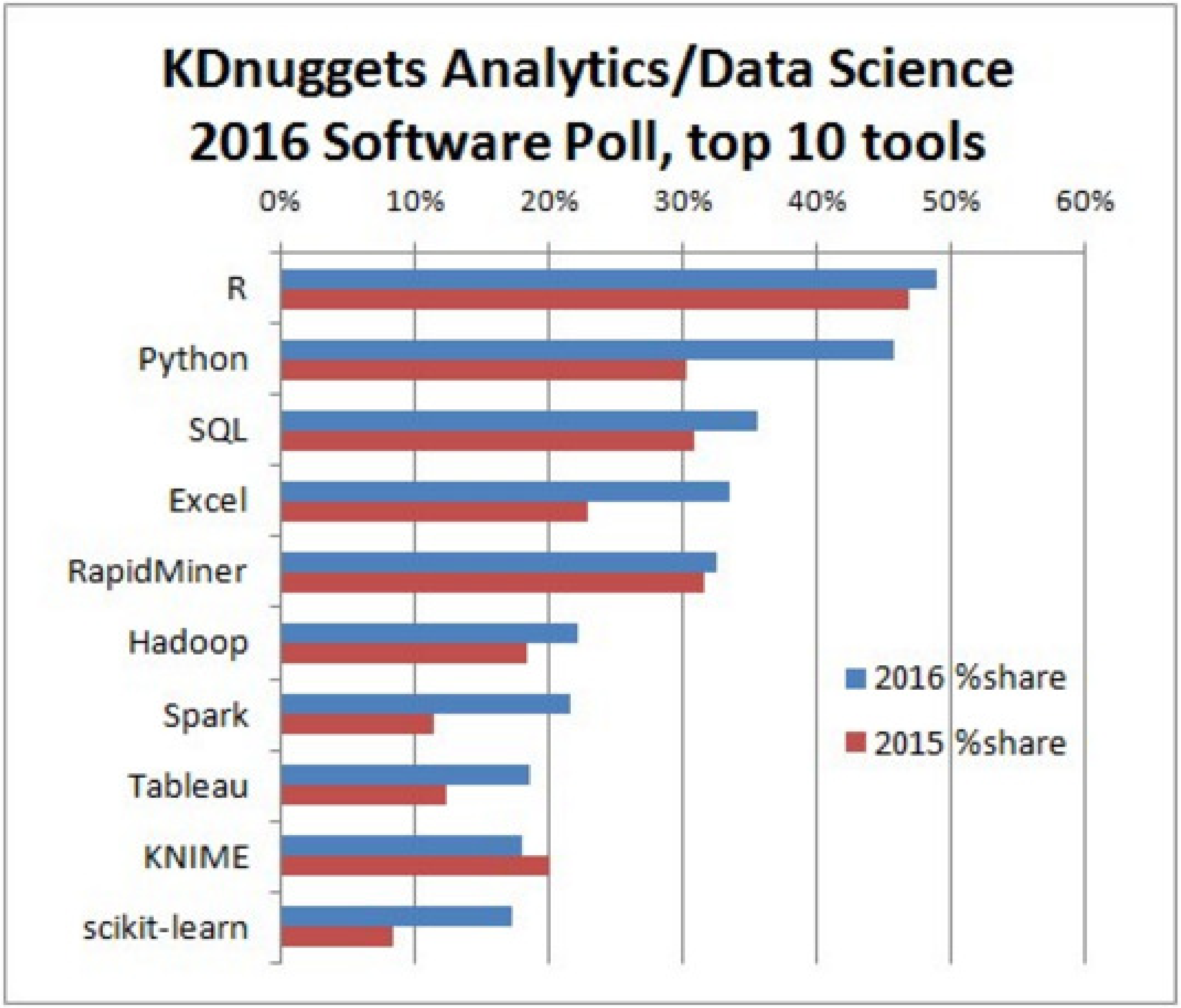
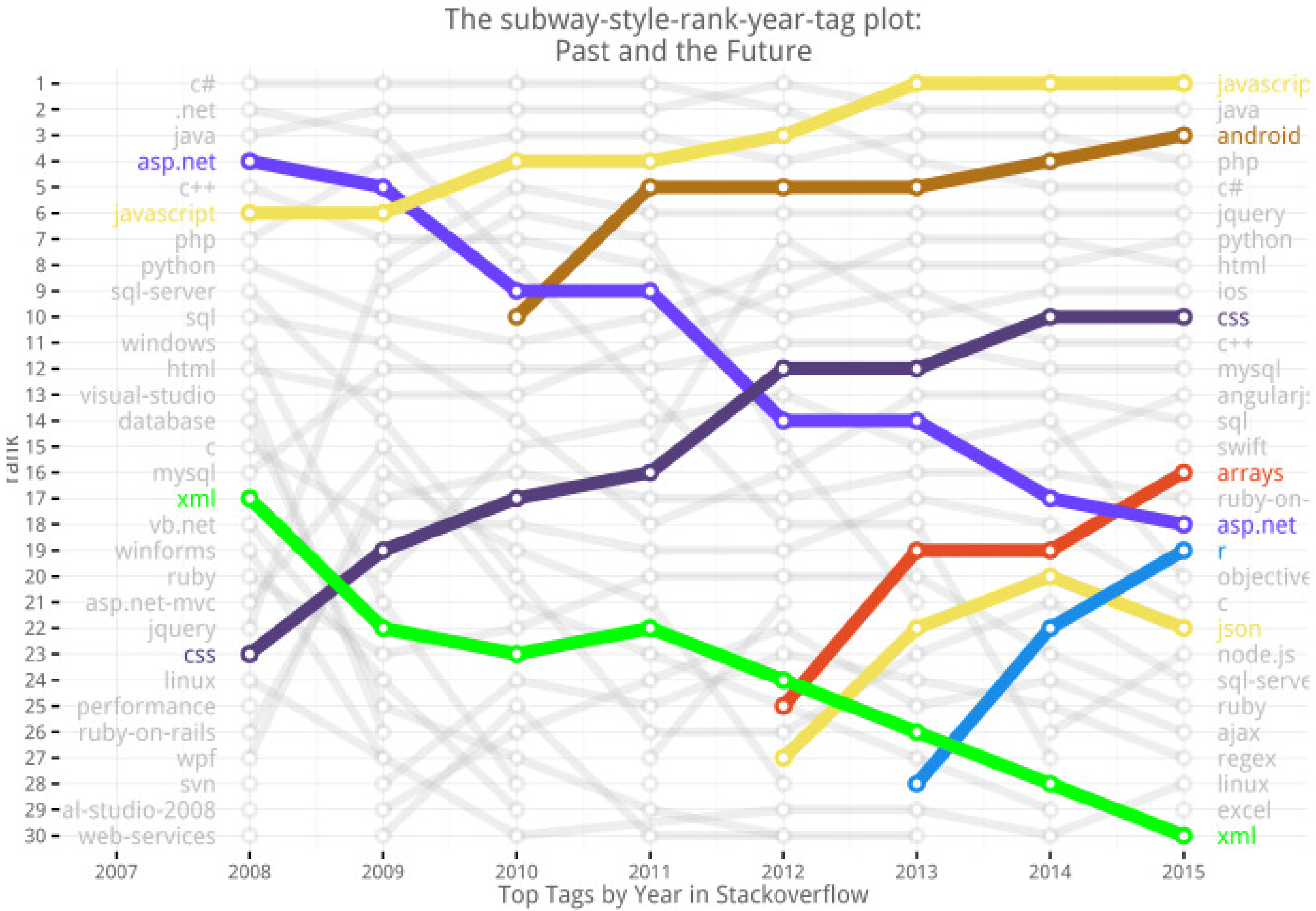


Fig 1: KDnuggets Analytics/Data Science 2016 Software Poll: top 10 most popular tools in 2016

R is the fastest-growing language on StackOverflow (This graph was drawn with R by the way)



R is the fastest-growing language on StackOverflow?

- Fake news! Sad!

- See PDF since internet link won’t work https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Full List of Topics (These might change a little)

- Introduction to Data Mining and Knowledge Discovery
- Data Mining Lifecycle: Six Phases
- Obtaining Data; Web Crawlers (etiquette, and spider traps); Twitter API
- Data Quality, Data Cleansing, Handling Missing Data and Identifying Misclassifications
- Graphical Methods for Identifying Outliers
- Data Transformation: Min-Max Normalization; Z-Score Standardization
- Overview of Supervised versus Unsupervised Learning approaches
- Hierarchical Clustering; k-Nearest Neighbor Algorithm; distance functions and database considerations
- Decision Trees; Classification and Regression Trees; C4.5 and CART Algorithm
- Naïve Bayes
- Artificial Neural Networks; Backpropagation
- Logistic Regression
- Association Rules; Market Basket Analysis
- Model Evaluation Techniques
- **Additional Topics (Time Permitting)**
- Principal Component Analysis
- Bagging and Boosting
- Lazy Learners
- Visualization of Data
- Text Mining Overview (clustering, summarizing)
-

A Typical Class

- We meet for 2.5 hours each Monday-Wednesday
- Nobody wants to listen to me lecture for 5 hours per week
- So the plan is we start off with an hour lecture
- Then we work on a significant data mining project or contest from Kaggle.com or somewhere else.
- Sometimes we’ll look at a case study, or listen to a student lecture, or go over some homework.
- I’m usually a lecture/exam guy, but this worked pretty well with the other summer cohort
- I think it’s good to learn from your peers and I actually learned a lot as well

Grading

- 25% HW and in class assignments
- 25% Presentation Score
- 25% Midterm
- 25% Final

- When I last taught this course we did a capstone project, but we just did one last week and doing another in a 7-week summer course seems like a bad idea
- So to mix things up we’ll do a lot of little presentations
- I also give out too many A’s each semester so we’ll make the HW slightly more challenging

Homework

- We’ll have probably only have 4 or 5 HW assignments
- Each will be worth 10 points.
- If you do everything required and on time you will get 9 out of 10.
- To get the 10th point you have to come up with a challenge question that does something extra.
- For example if the assignment is to build a Neural Network in R to model something and do k-folds validation as we do in class, the extra thing might be to duplicate the same thing using Sci-kit learn, or to repeat the validation using a second package like ‘caret.’
- It doesn’t have to be hard, it just has to be something a little extra. Maybe you find a new library or something; who knows?
- Note if you do something cool I might ask you to demo it in class at a future point

Short Presentations

- There will be a lot of opportunities for people to present short talks or demos
- Scoring: 1 presentation=80, 2=85, 3=90, 4=95, 5+=100

- This worked well in the summer class that just finished. The speaker always learns best by teaching others, it’s better to learn from peers then from the same talking head every week, and I’ll get to learn a lot as well

Other Stuff

- See syllabus for anything else I forgot to talk about

- Any questions?