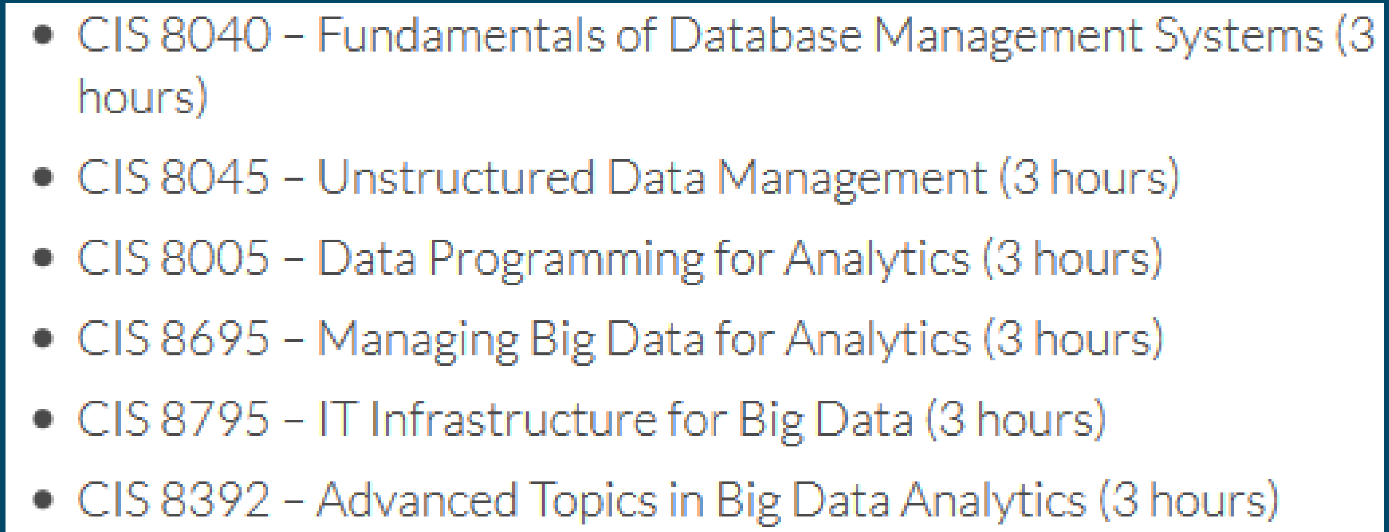
A Big Career in Big Data?

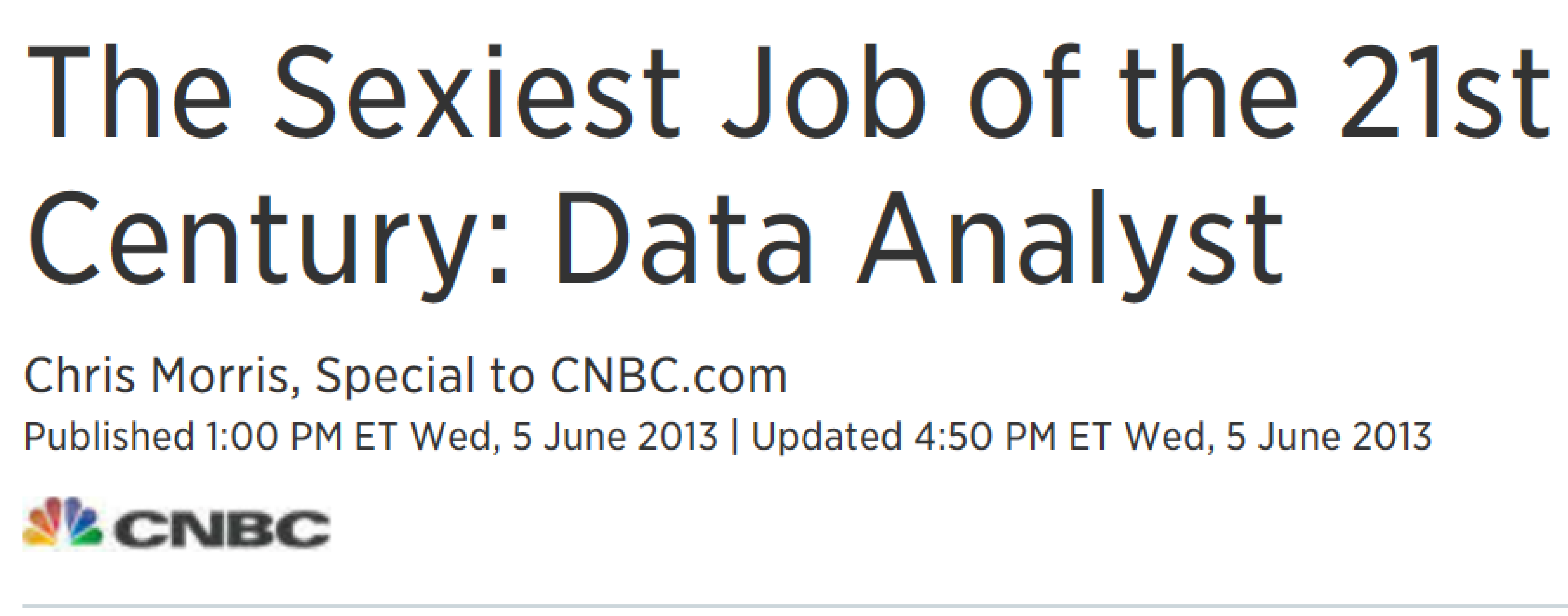
The **data analytics track** prepares students for Big Data careers

CIS 8795 Big Data Infrastructure

- The only course that has *hands on* with Big Data technologies
  - Cluster programming (with PySpark)
  - Cluster & Cloud infrastructure



- CIS 8040 – Fundamentals of Database Management Systems (3 hours)
- CIS 8045 – Unstructured Data Management (3 hours)
- CIS 8005 – Data Programming for Analytics (3 hours)
- CIS 8695 – Managing Big Data for Analytics (3 hours)
- CIS 8795 – IT Infrastructure for Big Data (3 hours)
- CIS 8392 – Advanced Topics in Big Data Analytics (3 hours)



The Sexiest Job of the 21st Century: Data Analyst

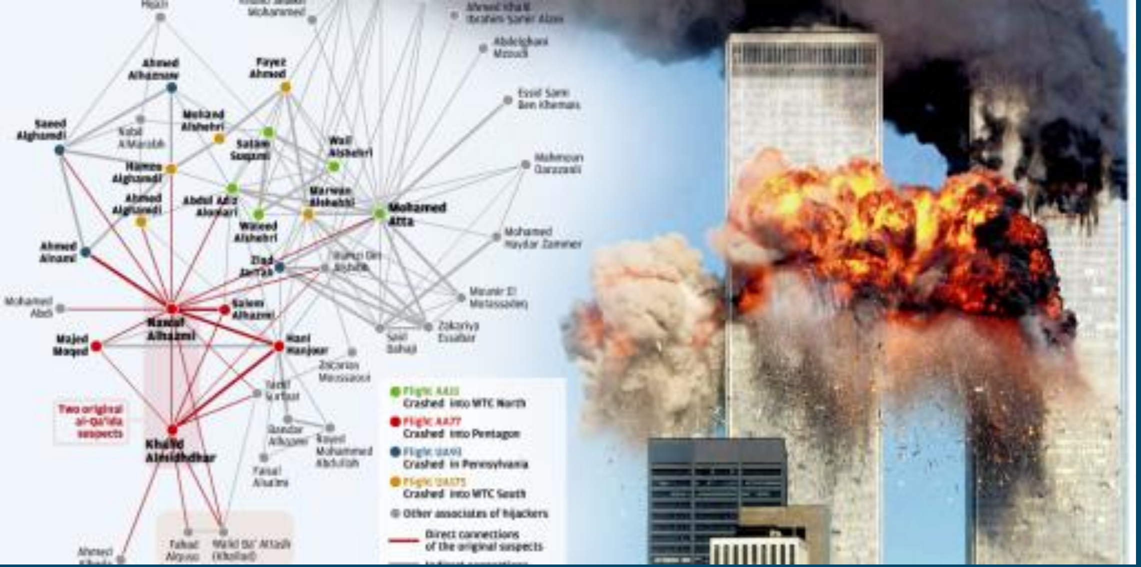
Chris Morris, Special to CNBC.com

Published 1:00 PM ET Wed, 5 June 2013 | Updated 4:50 PM ET Wed, 5 June 2013

CNBC

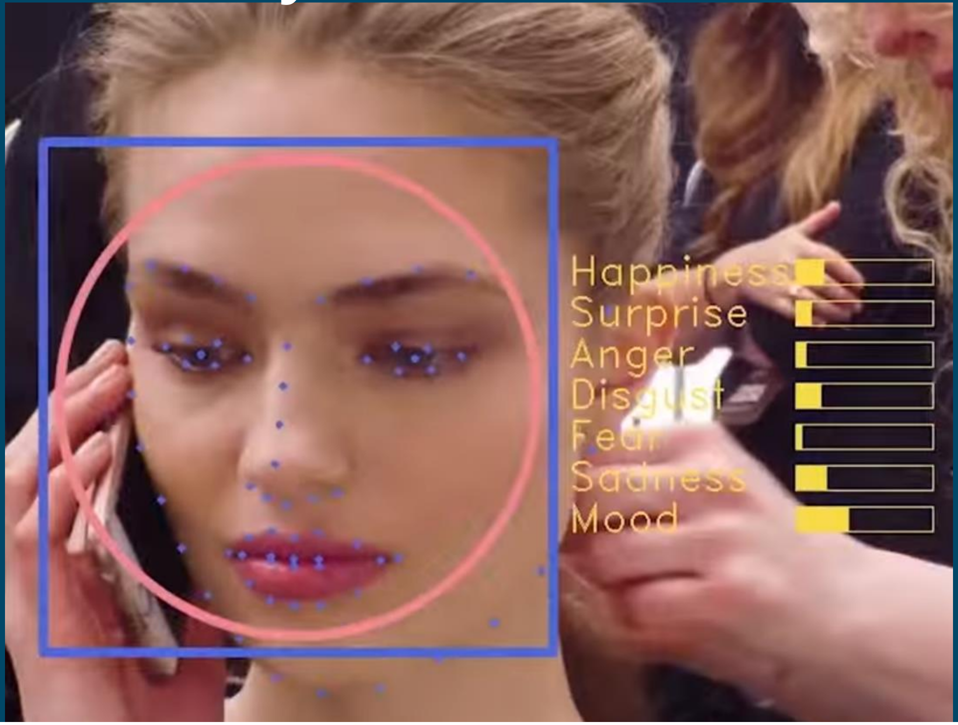
Doug Cutting, Hadoop creator

Every company
in business in the future
uses Big Data now

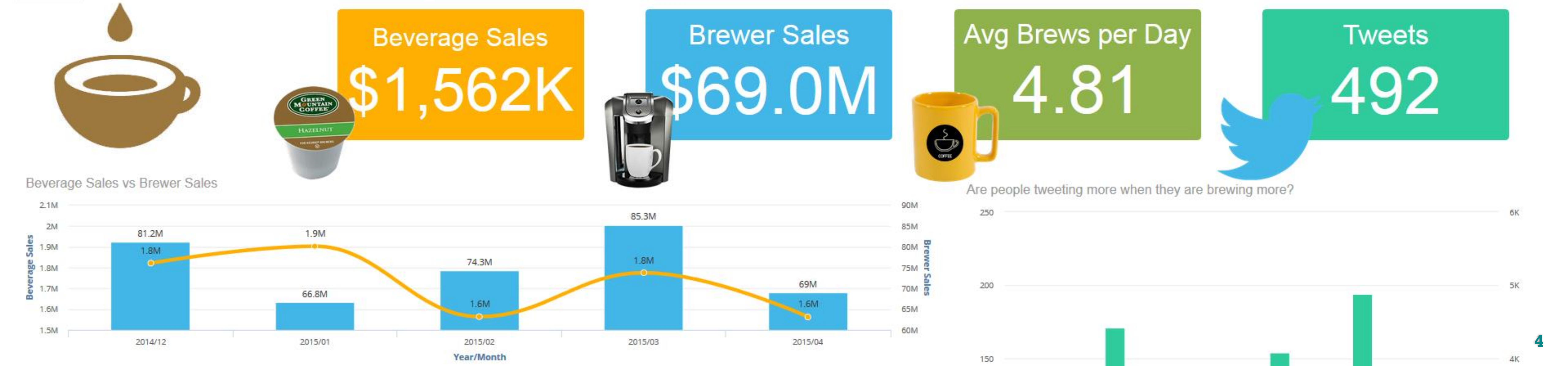


Analyze and predict crime

Analyze emotions



Analyze and predict business events



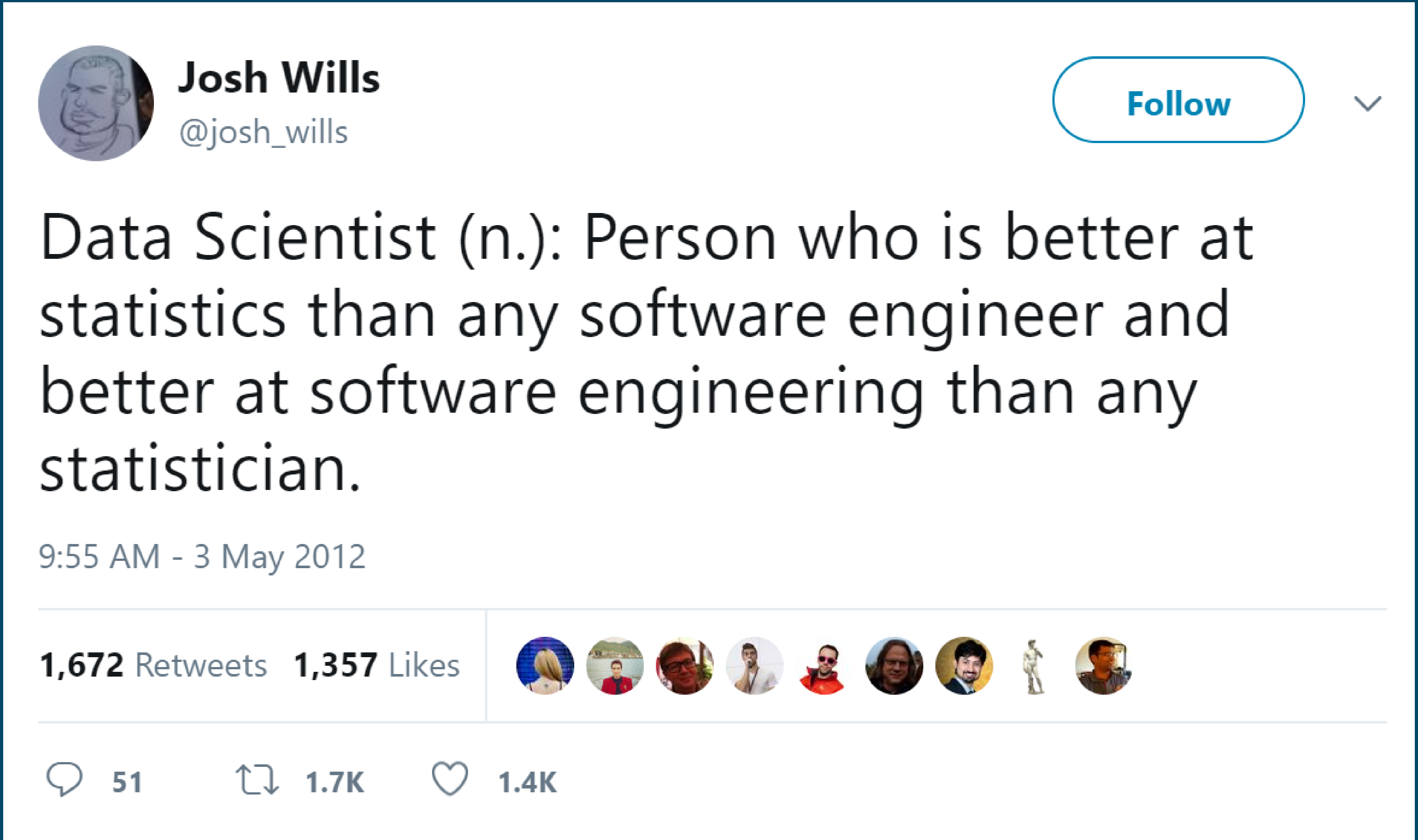
Beverage Sales $1,562K

Brewer Sales $69.0M

Avg Brews per Day 4.81

Tweets 492

Room level
48:56
68 dB



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.

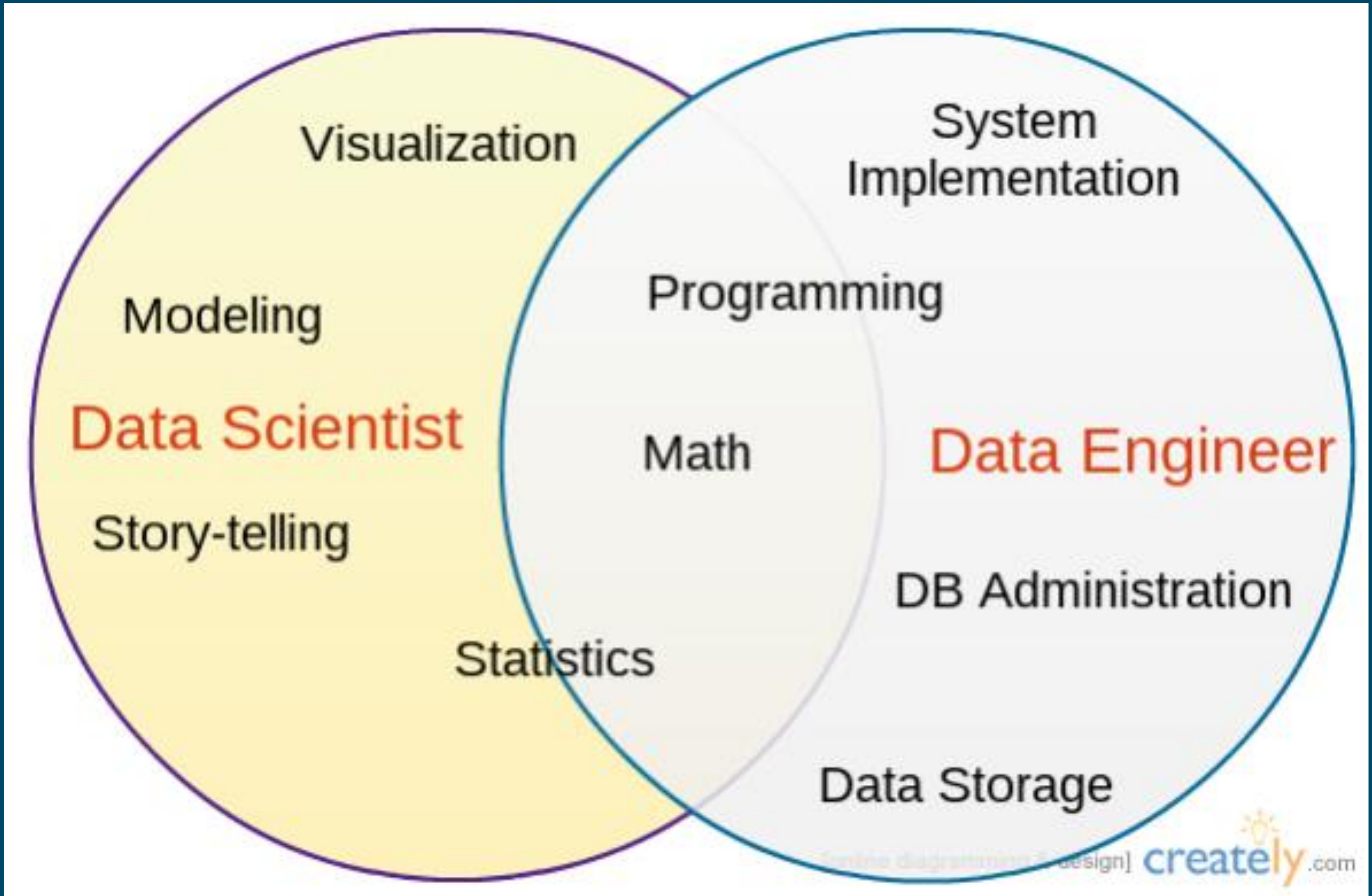
9:55 AM - 3 May 2012

1,672 Retweets **1,357** Likes

51 1.7K 1.4K

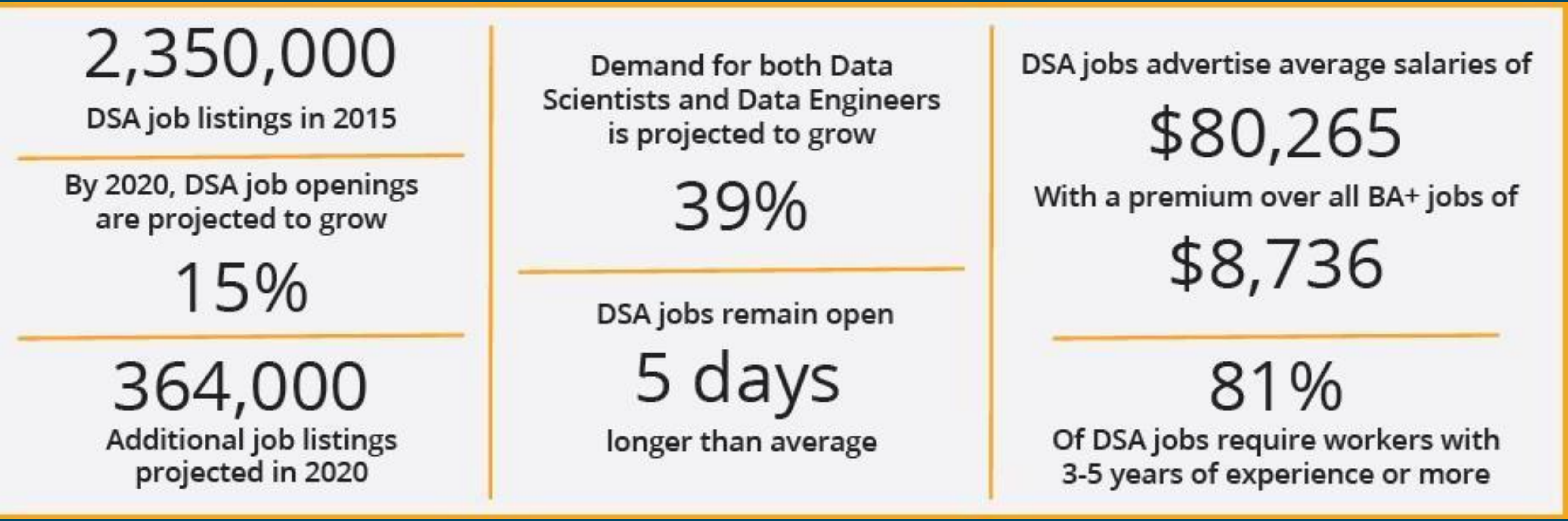
Josh was a technical manager of data science teams at Slack, Cloudera, Google

Analyst vs Engineer



More jobs and better salary

in data science and analytics (DSA)



2,350,000

DSA job listings in 2015

By 2020, DSA job openings are projected to grow

15%

364,000

Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow

39%

DSA jobs remain open

5 days

longer than average

DSA jobs advertise average salaries of

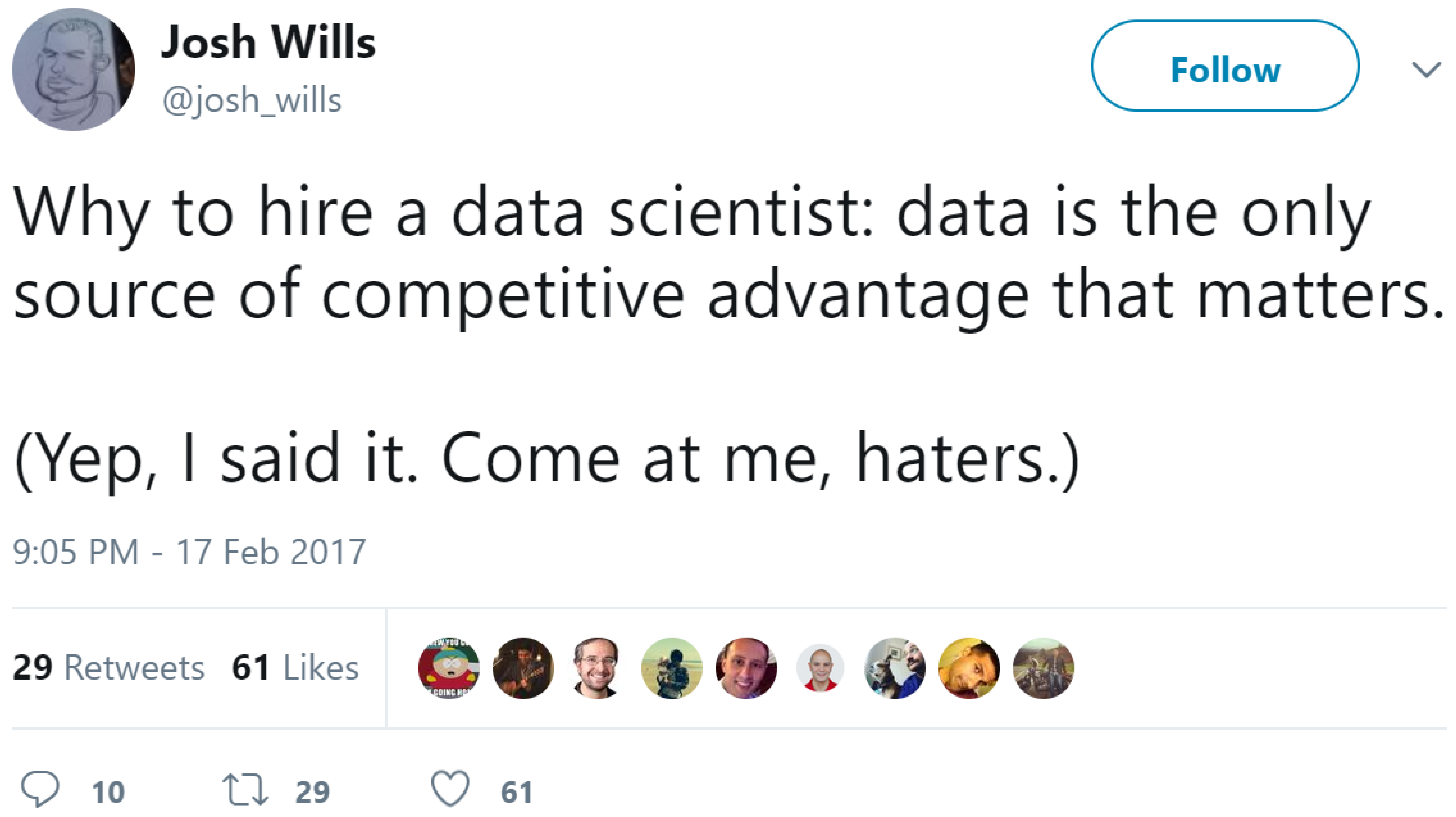
$80,265

With a premium over all BA+ jobs of

$8,736

81%

Of DSA jobs require workers with 3-5 years of experience or more



Josh Wills
@josh_wills

Why to hire a data scientist: data is the only source of competitive advantage that matters.

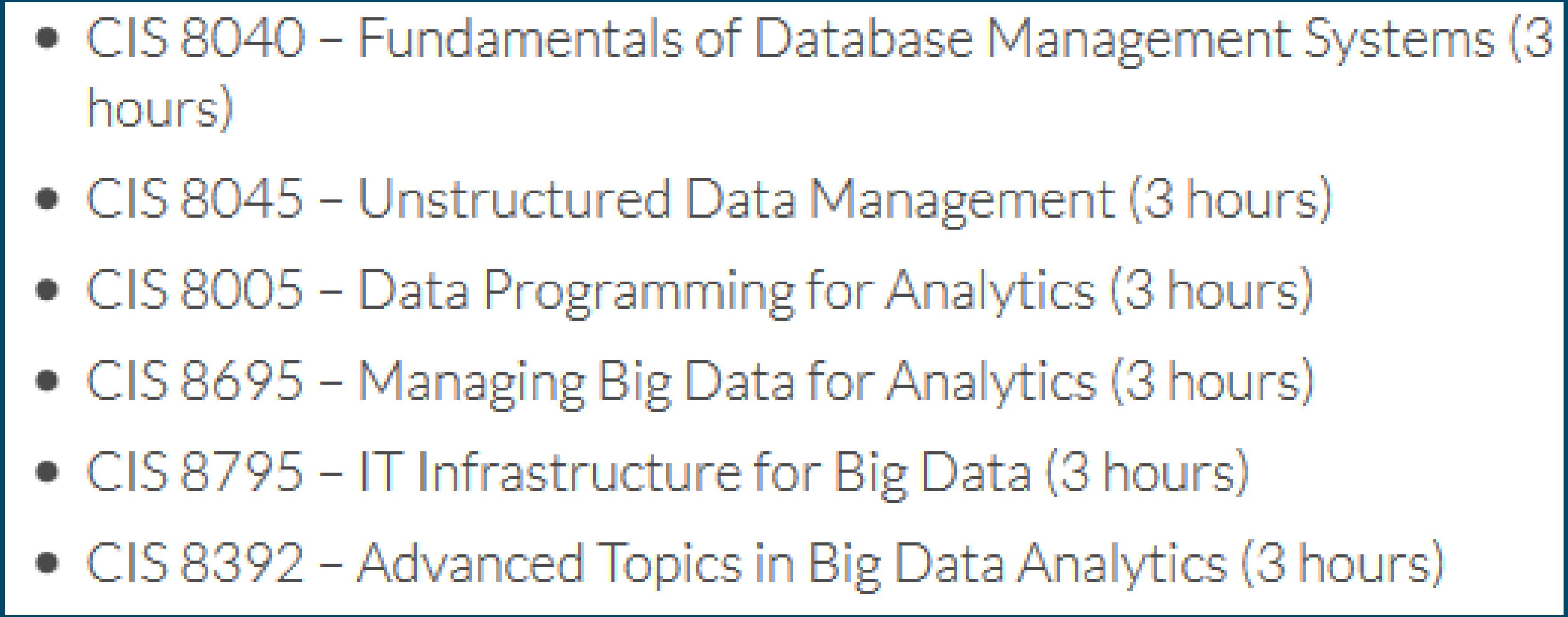
(Yep, I said it. Come at me, haters.)

9:05 PM - 17 Feb 2017

29 Retweets 61 Likes

10 29 61

Take the **Data Analytics** courses



- CIS 8040 – Fundamentals of Database Management Systems (3 hours)
- CIS 8045 – Unstructured Data Management (3 hours)
- CIS 8005 – Data Programming for Analytics (3 hours)
- CIS 8695 – Managing Big Data for Analytics (3 hours)
- CIS 8795 – IT Infrastructure for Big Data (3 hours)
- CIS 8392 – Advanced Topics in Big Data Analytics (3 hours)

Manage disparate data + Analyze data = $$

Two big data career paths

Scientist vs Engineer

Data analyst / scientist

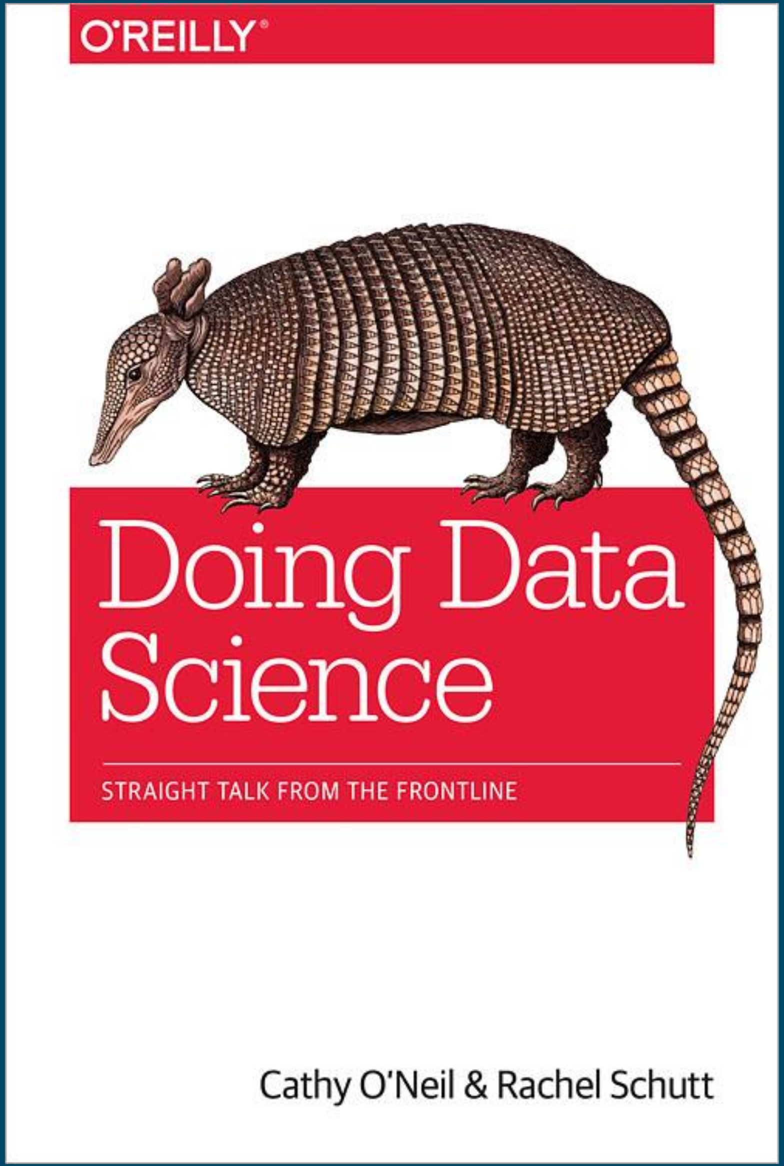
- Taking data, using it to answer questions, and communicating the results to help make business decisions
  - Data cleaning, performing analysis and creating data visualizations
  - Some consider a data analyst a junior data scientist, some consider them equivalent terms

Data engineer

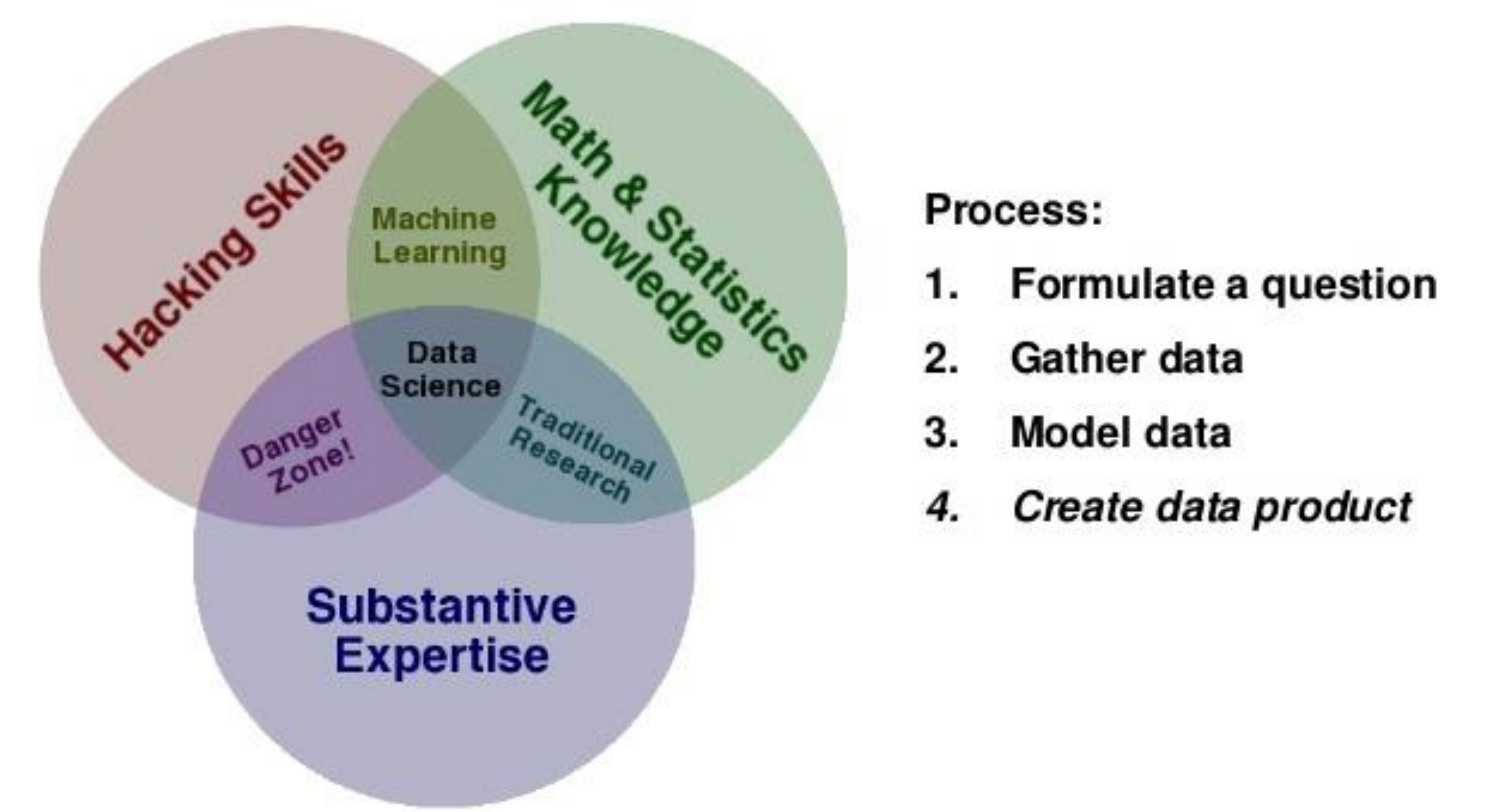
- Build and optimize the systems that allow data scientists and analysts to perform their work
  - Constructing data pipelines using complex tools and techniques to handle data at scale
  - Building APIs for data consumption
  - Integrating external or new datasets into existing data pipelines

DataEDGE

A big data analyst has a mix of skills


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Cathy O'Neil & Rachel Schutt

Data Analyst

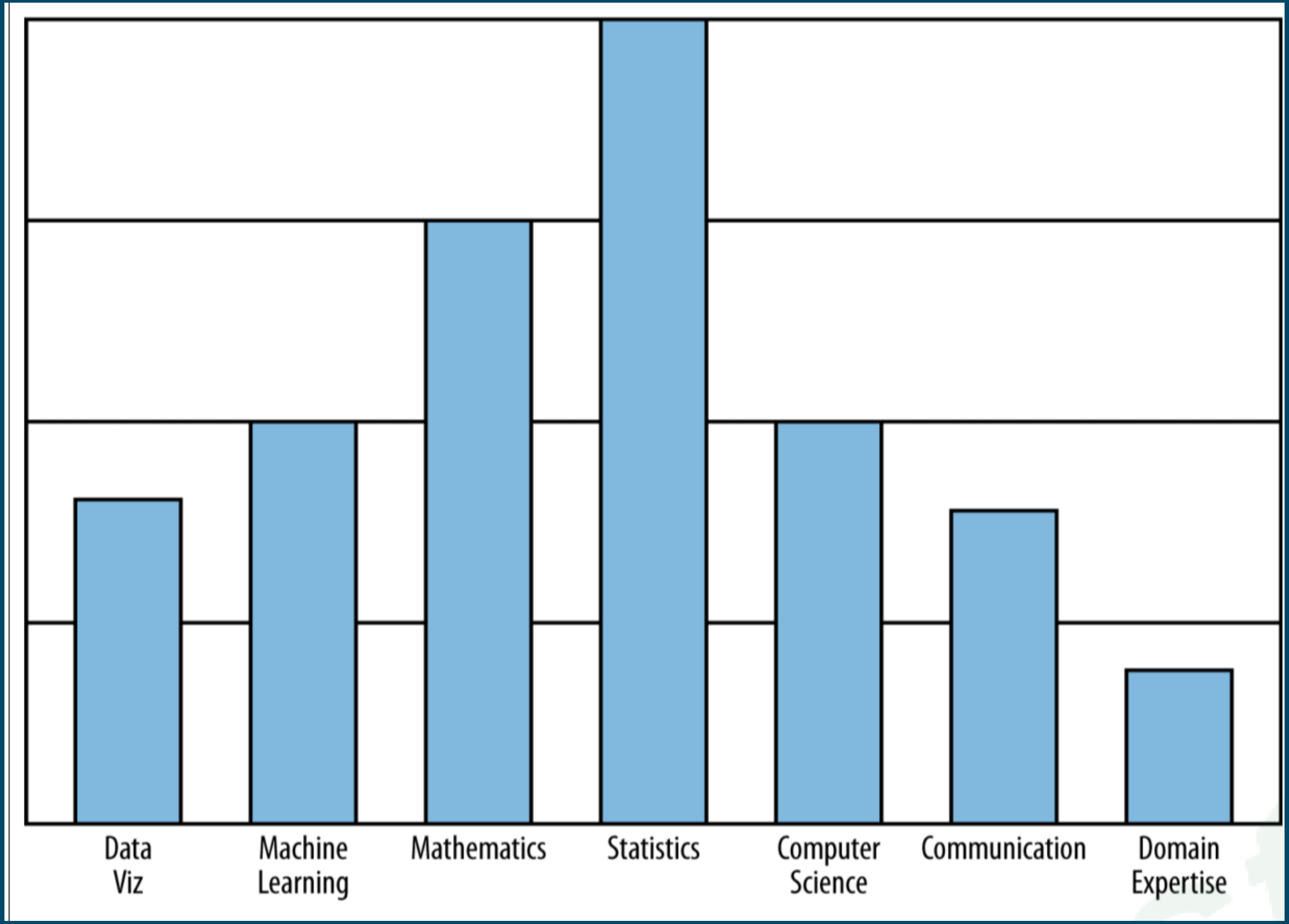


Process:

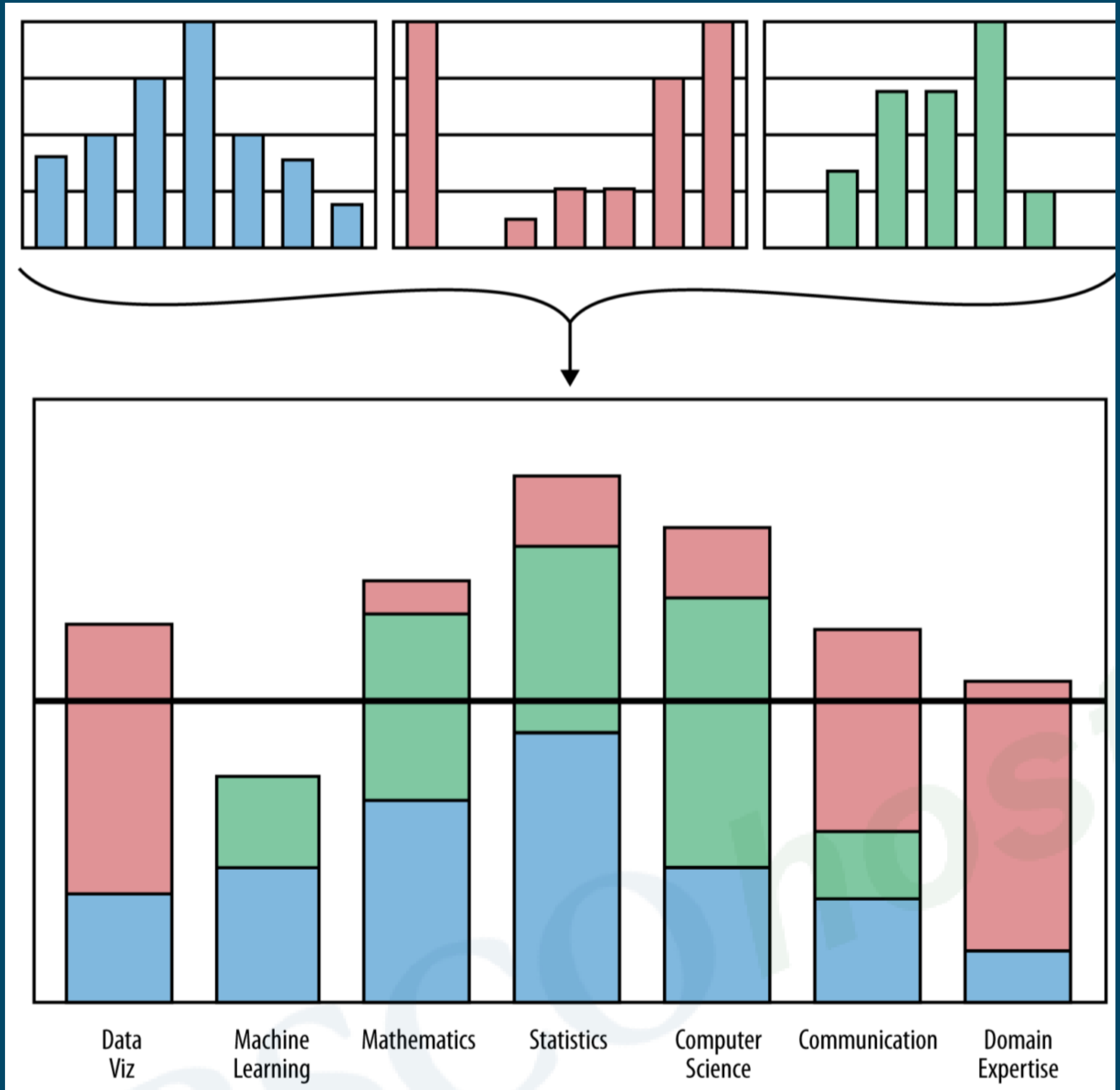
1. Formulate a question
2. Gather data
3. Model data
4. *Create data product*

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

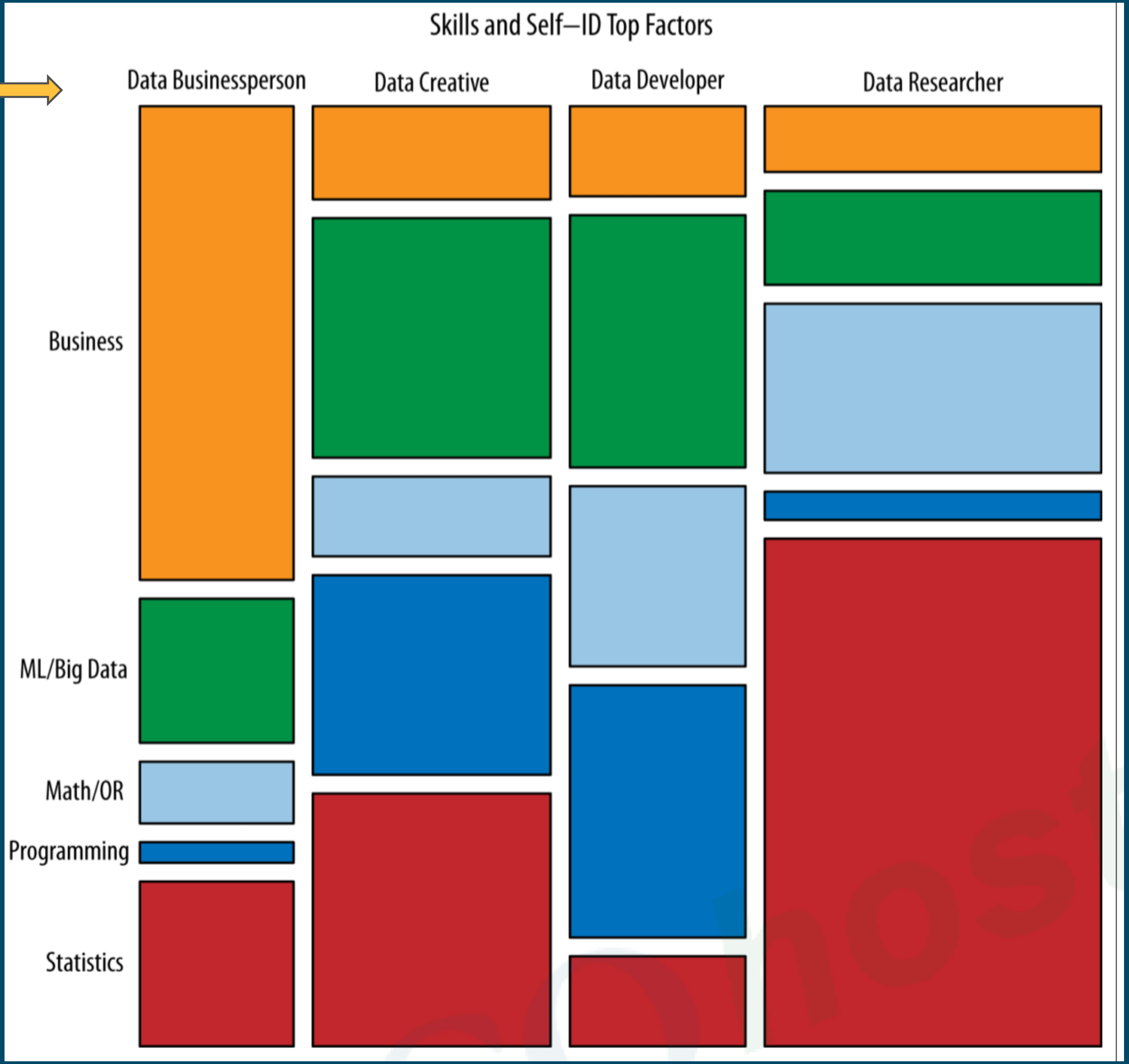
Data Analyst Profile



A team forms "a data analyst"



Data analyst profiles



Job prospects

- Jobs requiring machine learning skills are paying an average of $114,000.
- Advertised
  - data scientist jobs pay an average of $105,000
  - data engineering jobs pay an average of $117,000
- 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Jobs remain open an average of 45 days, five days longer than the market average.

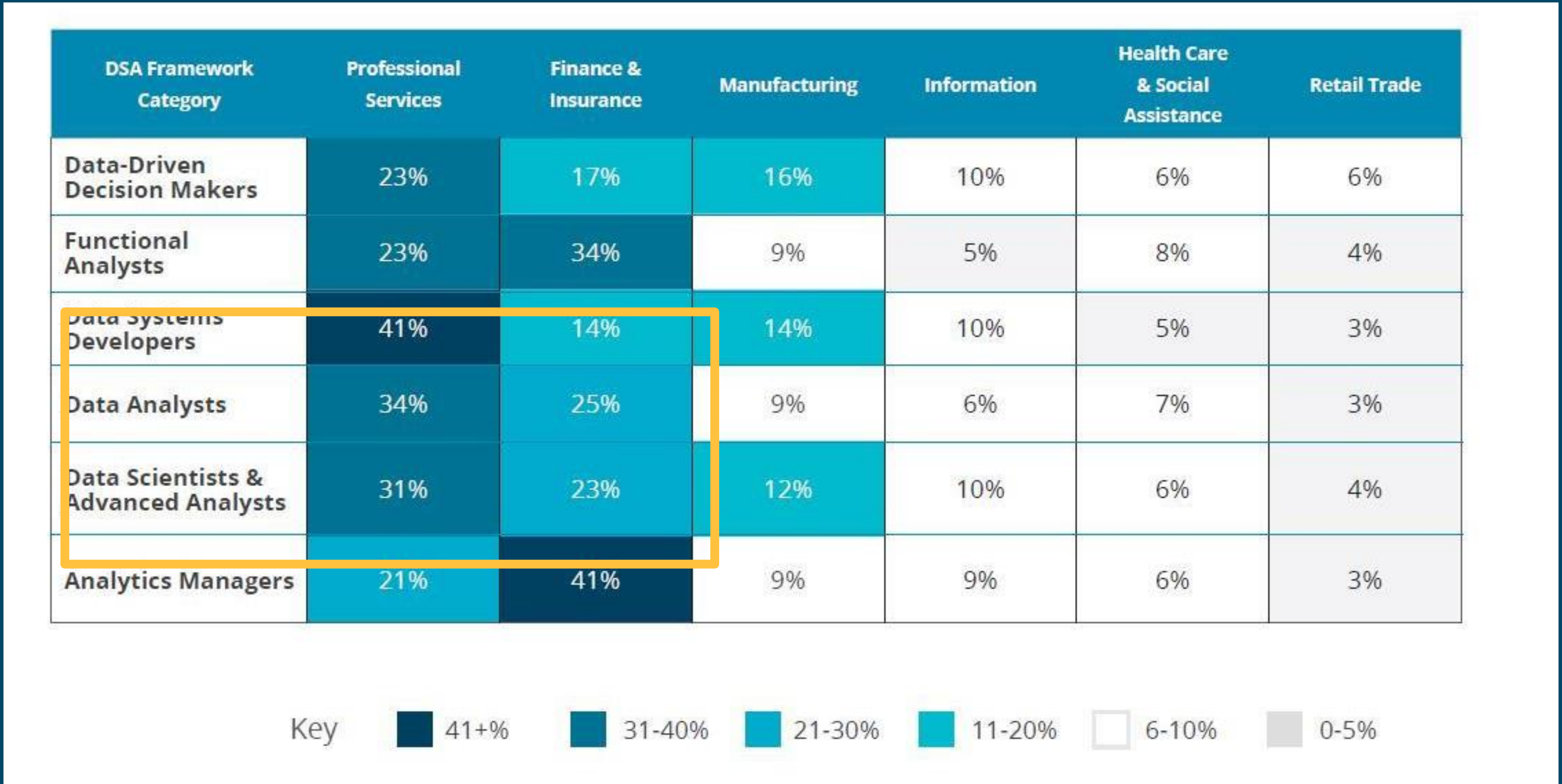
https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6f4e2b367e3b

Job prospects

- Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020
- By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM

https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6f4e2b367e3b

Demand by industry

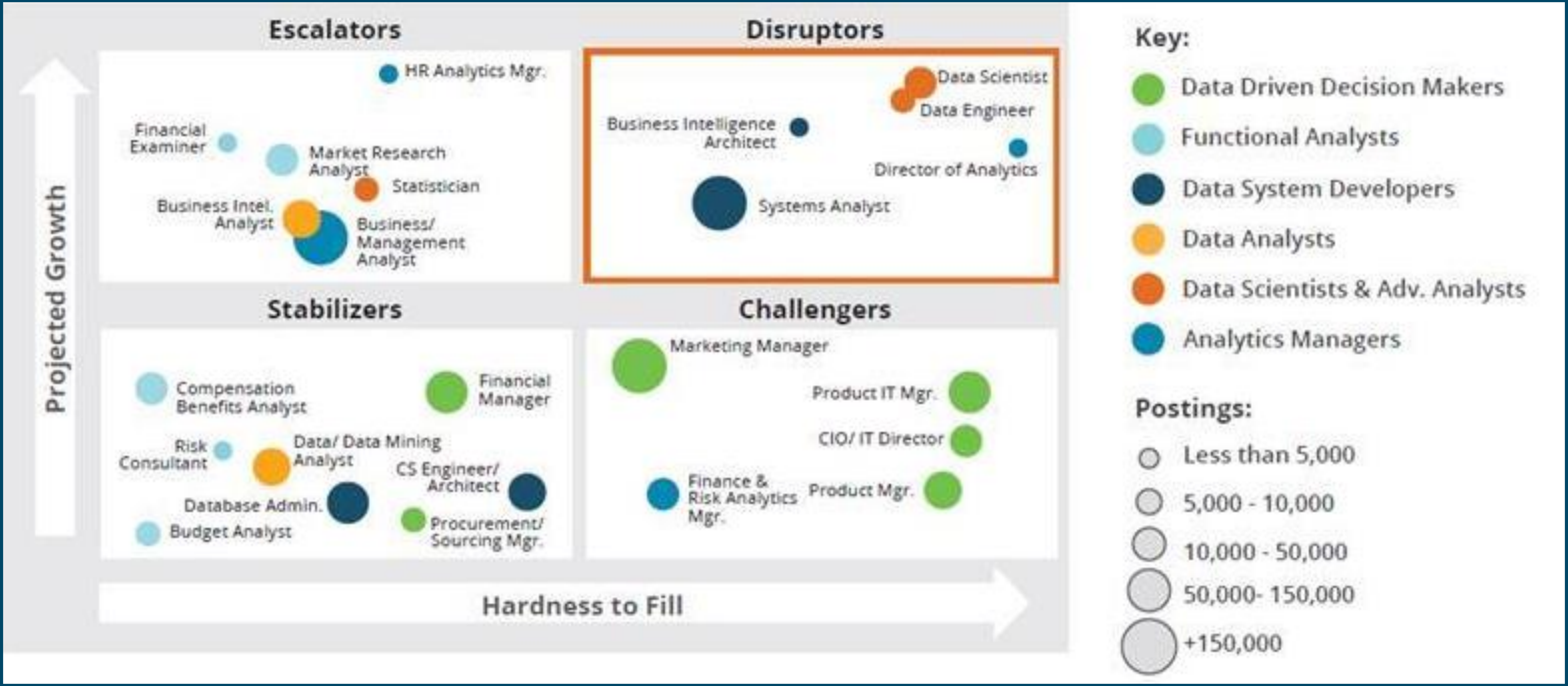


DSA Framework Category	Professional Services	Finance & Insurance	Manufacturing	Information	Health Care & Social Assistance	Retail Trade
Data-Driven Decision Makers	23%	17%	16%	10%	6%	6%
Functional Analysts	23%	34%	9%	5%	8%	4%
Data Systems Developers	41%	14%	14%	10%	5%	3%
Data Analysts	34%	25%	9%	6%	7%	3%
Data Scientists & Advanced Analysts	31%	23%	12%	10%	6%	4%
Analytics Managers	21%	41%	9%	9%	6%	3%

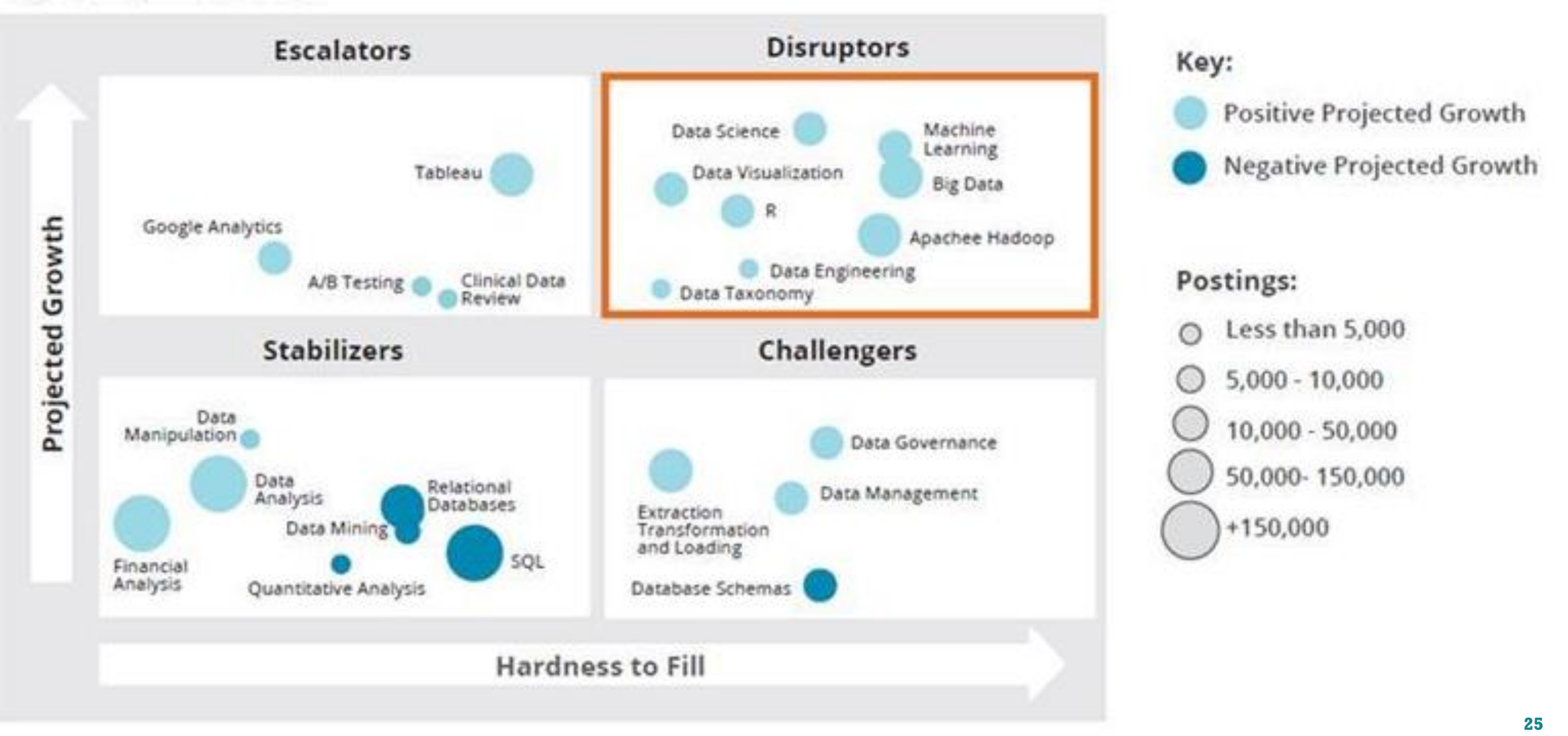
Key ■ 41+% ■ 31-40% ■ 21-30% ■ 11-20% □ 6-10% ■ 0-5%

https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6f4e2b367e3b

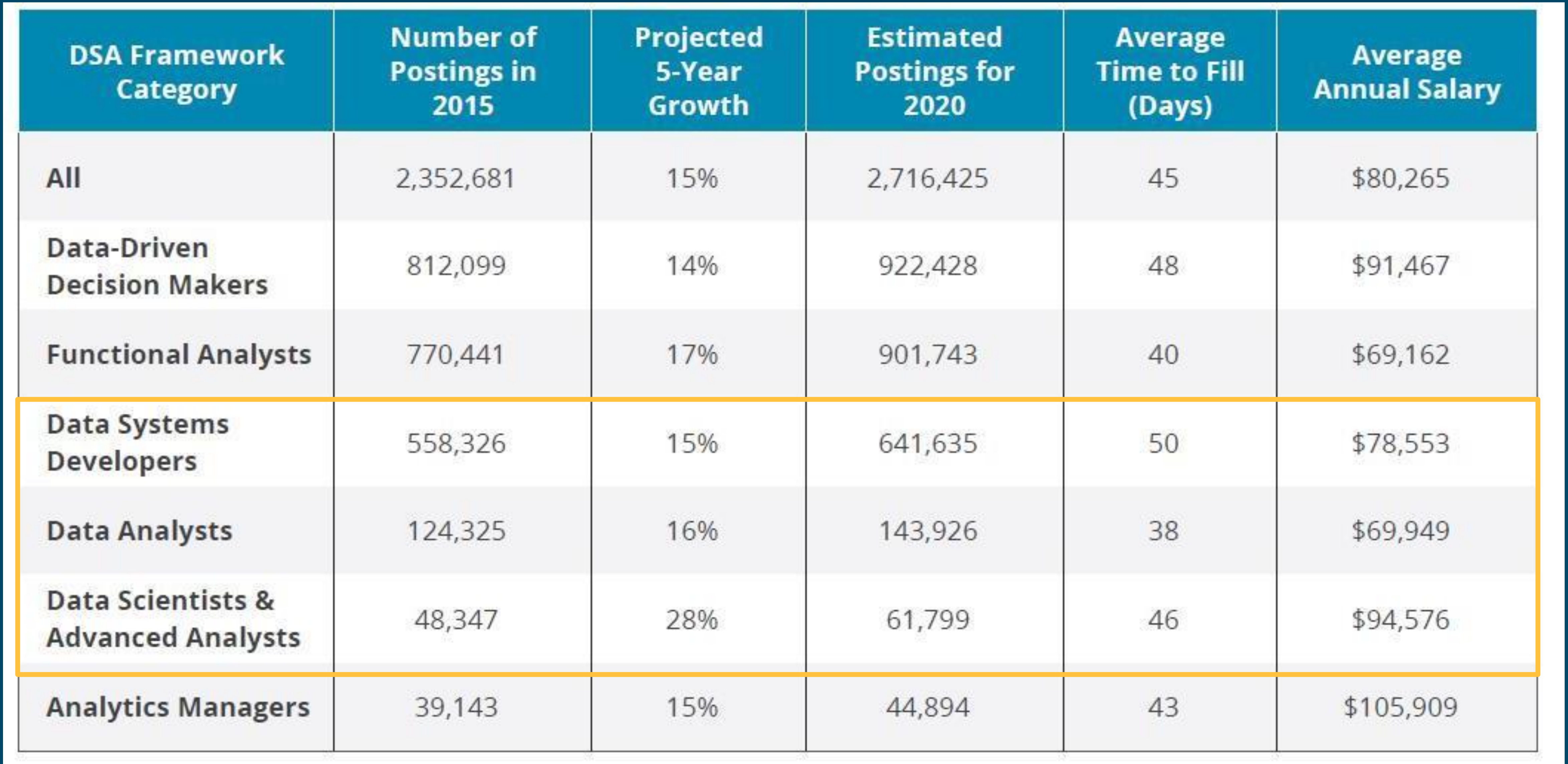
Jobs Matrix


Escalators
HR Analytics Mgr.
Financial Examiner
Market Research Analyst
Statistician
Business Intel. Analyst
Business/ Management Analyst
Disruptors
Data Scientist
Data Engineer
Business Intelligence Architect
Director of Analytics
Systems Analyst
Stabilizers
Compensation Benefits Analyst
Financial Manager
Risk Consultant
Data/ Data Mining Analyst
CS Engineer/ Architect
Database Admin.
Budget Analyst
Procurement/ Sourcing Mgr.
Challengers
Marketing Manager
Product IT Mgr.
CIO/ IT Director
Finance & Risk Analytics Mgr.
Product Mgr.
Projected Growth
Hardness to Fill
Key:
Data Driven Decision Makers
Functional Analysts
Data System Developers
Data Analysts
Data Scientists & Adv. Analysts
Analytics Managers
Postings:
Less than 5,000
5,000 - 10,000
10,000 - 50,000
50,000- 150,000
+150,000

Skills Matrix



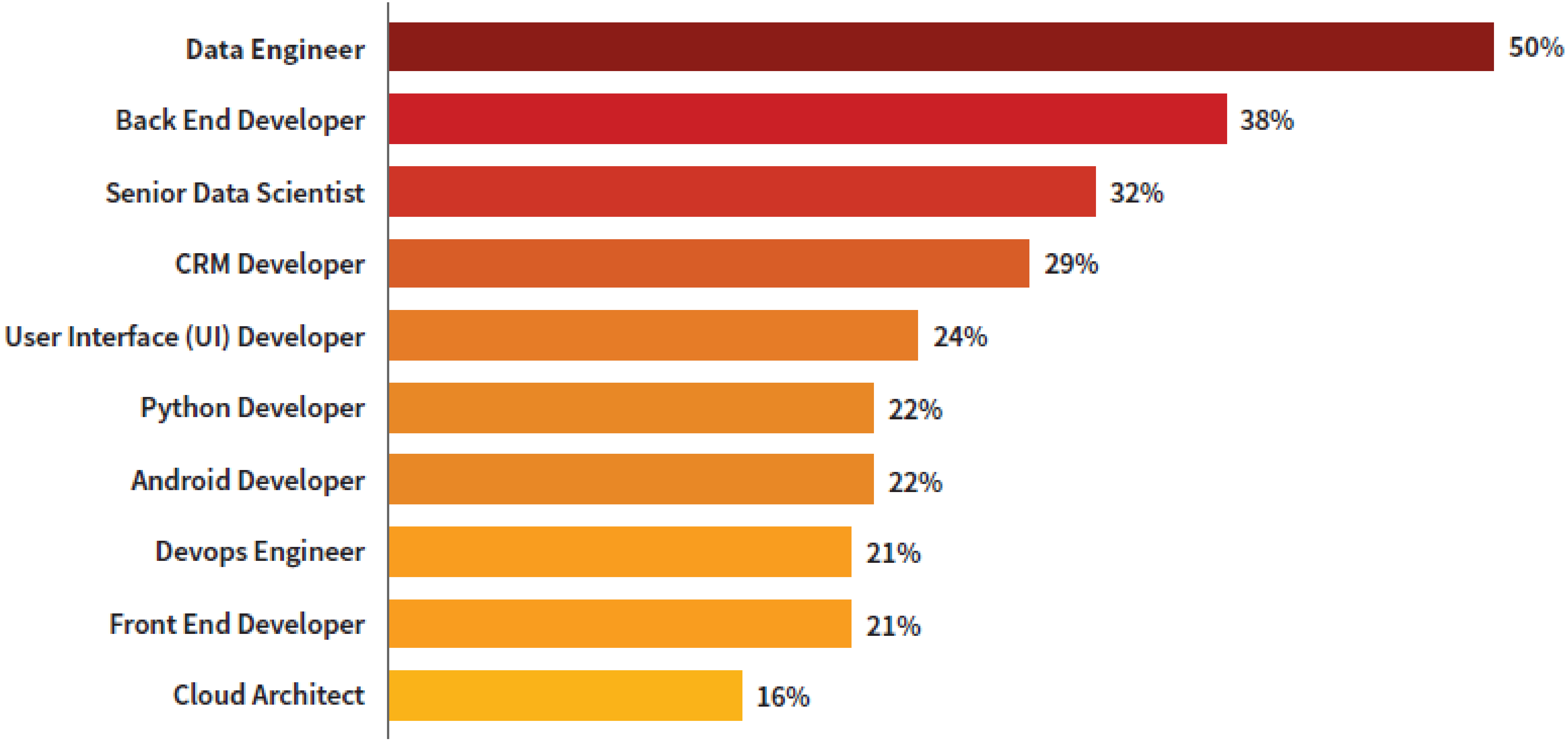
Demand statistics



DSA Framework Category	Number of Postings in 2015	Projected 5-Year Growth	Estimated Postings for 2020	Average Time to Fill (Days)	Average Annual Salary
All	2,352,681	15%	2,716,425	45	$80,265
Data-Driven Decision Makers	812,099	14%	922,428	48	$91,467
Functional Analysts	770,441	17%	901,743	40	$69,162
Data Systems Developers	558,326	15%	641,635	50	$78,553
Data Analysts	124,325	16%	143,926	38	$69,949
Data Scientists & Advanced Analysts	48,347	28%	61,799	46	$94,576
Analytics Managers	39,143	15%	44,894	43	$105,909

FASTEST GROWING TECH OCCUPATIONS

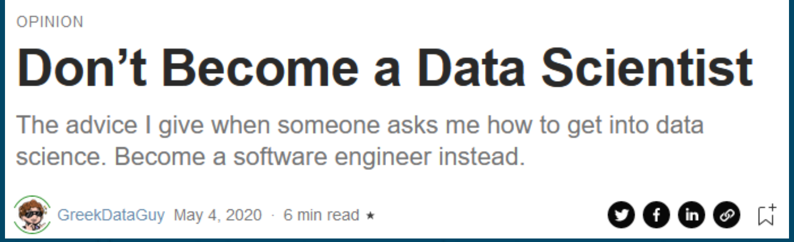
YEAR-OVER-YEAR GROWTH



This course: mix of DA and DE

- Data analyst
  - PySpark, Spark
  - Cluster programming
  - Machine learning (ML)
  - Visualization

- Data Engineer
  - Hadoop ecosystem
  - HDFS
  - MapReduce
  - Containers and deployment
    - Docker, Kubernetes
  - DevOps



OPINION

Don’t Become a Data Scientist

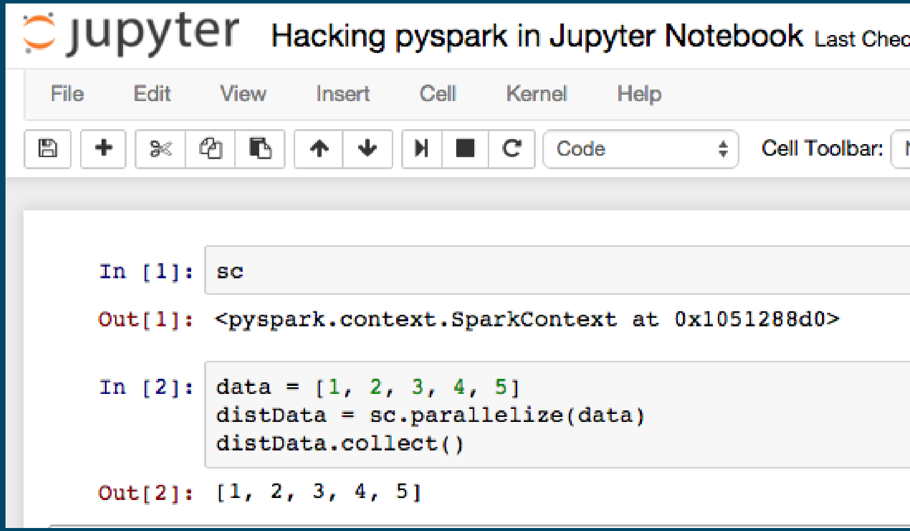
The advice I give when someone asks me how to get into data science. Become a software engineer instead.

GreekDataGuy May 4, 2020 · 6 min read ★

https://towardsdatascience.com/dont-become-a-data-scientist-ee4769899025

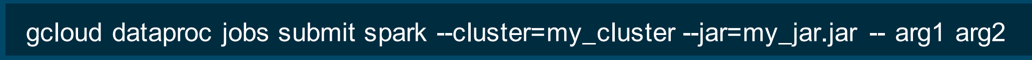
Course activities

- PySpark (parallel) programming



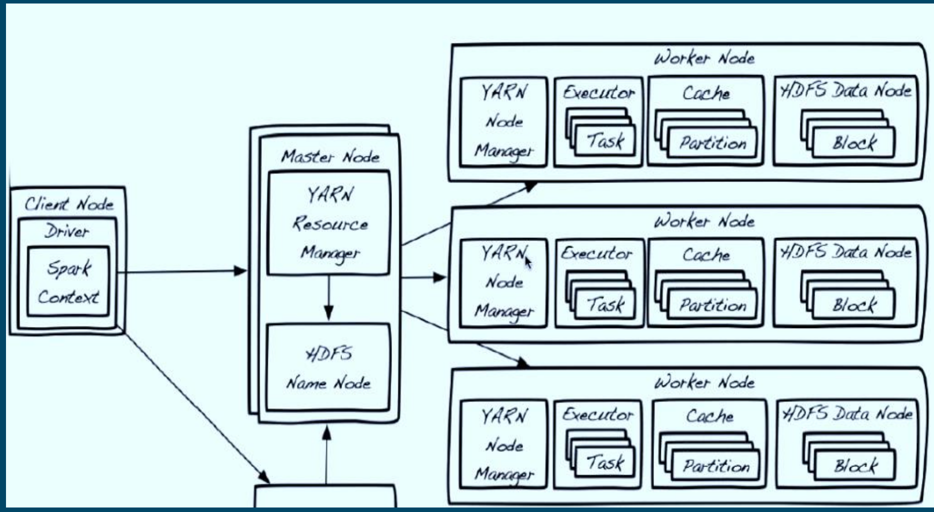
```
In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x1051288d0>

In [2]: data = [1, 2, 3, 4, 5]
        distData = sc.parallelize(data)
        distData.collect()
Out[2]: [1, 2, 3, 4, 5]
```

- Deployment to a cluster



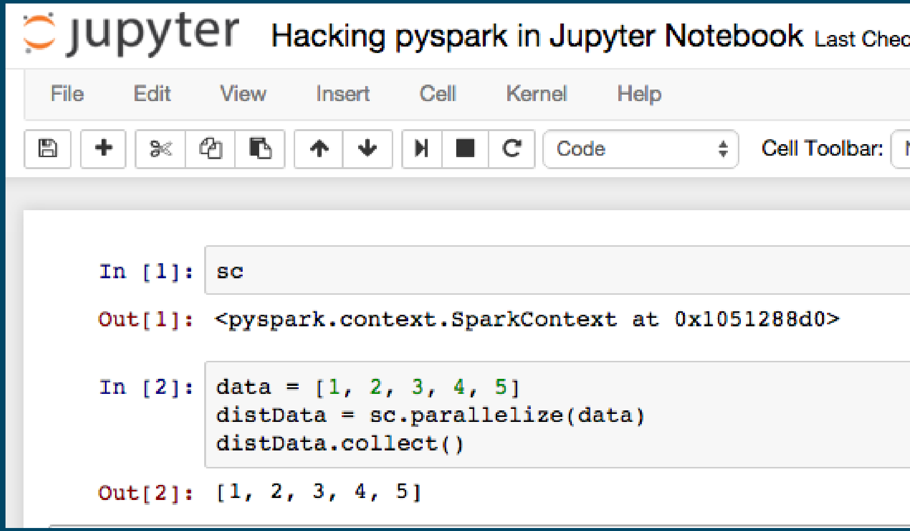
```
gcloud dataproc jobs submit spark --cluster=my_cluster --jar=my_jar.jar -- arg1 arg2
```

- Underlying technologies



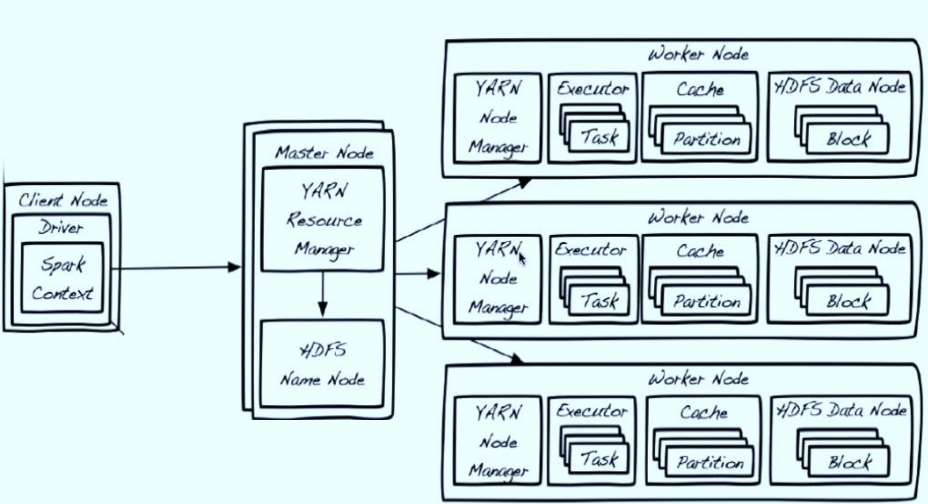
Course activities

- PySpark (parallel) programming
  - Notebooks & PyCharm
- Deployment to a (DataBricks, Google) cluster
  - Managed cluster & unmanaged
    - Unmanaged = developer configured
- Underlying technologies
  - Spark, Hadoop (HDFS, MapReduce), ML APIs, Streaming, Docker, Kubernetes



```
In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x1051288d0>

In [2]: data = [1, 2, 3, 4, 5]
        distData = sc.parallelize(data)
        distData.collect()
Out[2]: [1, 2, 3, 4, 5]
```

```
gcloud dataproc jobs submit spark --cluster=my_cluster --jar=my_jar.jar -- arg1 arg2
```



It’s a **difficult, long journey** for a person who just started programming a few months ago

This course is about understanding, using, optimizing

Big Data infrastructure

(Modeling theory, NoSQL, etc. are **other courses**)

The course exams are geared toward

Data Analyst and Data Engineer

certification exams and job interviews

Big Data Career Summary

- Data analyst / scientist
  - Using big data to answer questions and communicate the results to help make business decisions
- Data engineer
  - Build and optimize the systems that allow data scientists and analysts to perform their work
  - 4x more jobs and slightly higher salary

Important to remember

- Data scientists model big data to answer business & science questions
- Data engineers build and optimizes big data infrastructure (at scale)
- Both require a mix of skills, including programming, software engineering, and statistics