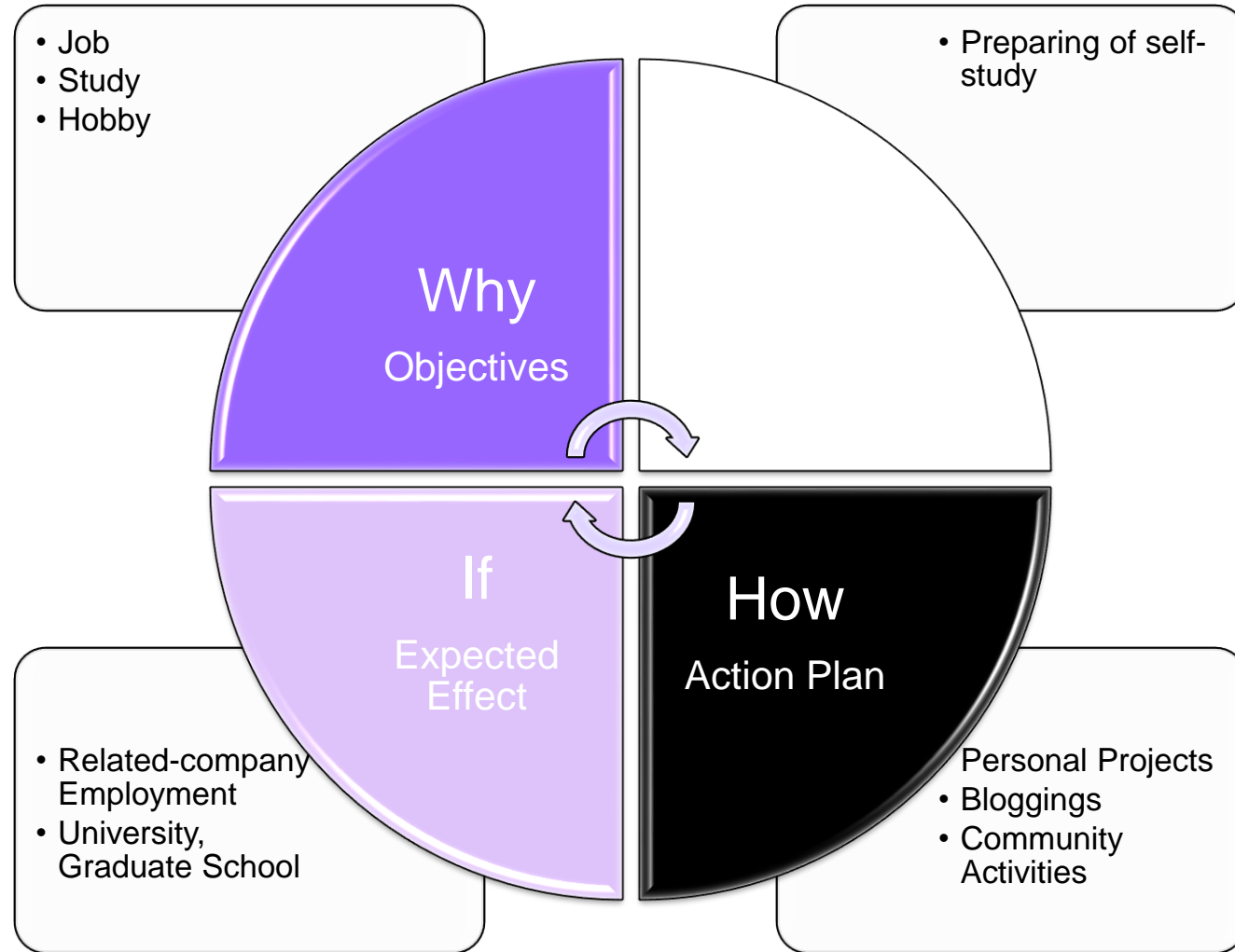


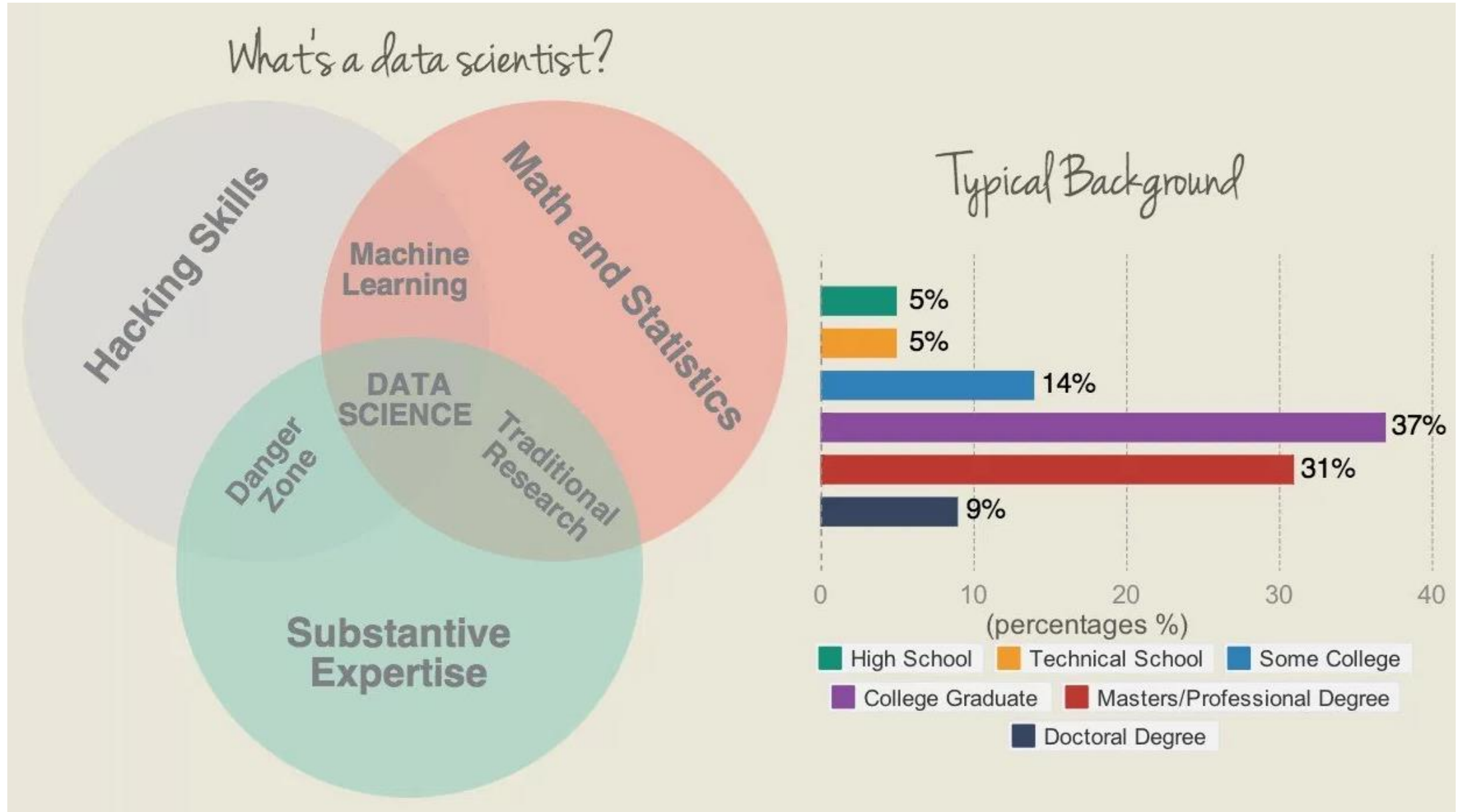
How to Become a Data Scientist in 8 Easy Steps

Bok, Jong Soon
javaexpert@nate.com
<https://github.com/swacademy>

Overview Using 4MAT



What is Data Scientist?



1. Get Good at Stats, Math and Machine Learning

Math



- > Math Track of Khan Academy
- > Linear Algebra by MIT OpenCourseware



Stats



- > Intro to Statistics by Udacity
- > OpenIntro Statistics



ML

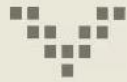


- > Machine Learning by Andrew NG (Stanford Online)
- > Practical Machine Learning by John Hopkins (Coursera)

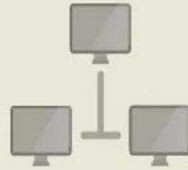
2. Learn to Code



Computer Science Fundamentals
> CS50x on edX



Grasp end-to-end development
The things you build will be integrated
into other systems



Choose a first language
> Open Source: R, Python, etc.
> Commercial: SAS, SPSS, etc.

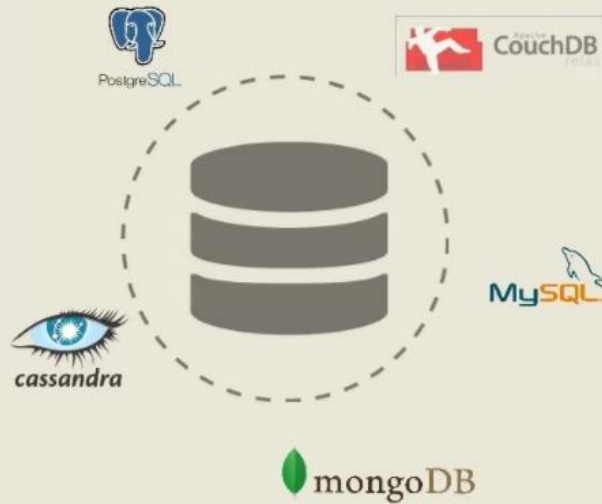


Learn Interactively
> R: DataCamp, try R
> Python: Codecademy, Google Class



3. Understand Databases

As a data scientist student, you will often work with data in text files. However, once you enter the industry, a database is almost always used to store data. It's going to be stored in MySQL, Postgres, MongoDB, Cassandra, etc.



Learn more on databases via:



datamonkey.pro



4. Master Data Munging, Visualization and Reporting

☐ Data cleaning and munging


WHAT

Data munging is the process of converting one "raw" form into another format for more convenient consumption

TOOLS

> Getting and Cleaning data by John Hopkins (Coursera)

DataWrangler^{alpha}


 data.table
dplyr


☐ Data visualization

WHAT

Data visualization involves the creation and study of the visual representation of data.

TOOLS

ggvis 


 vega


☐ Reporting


WHAT

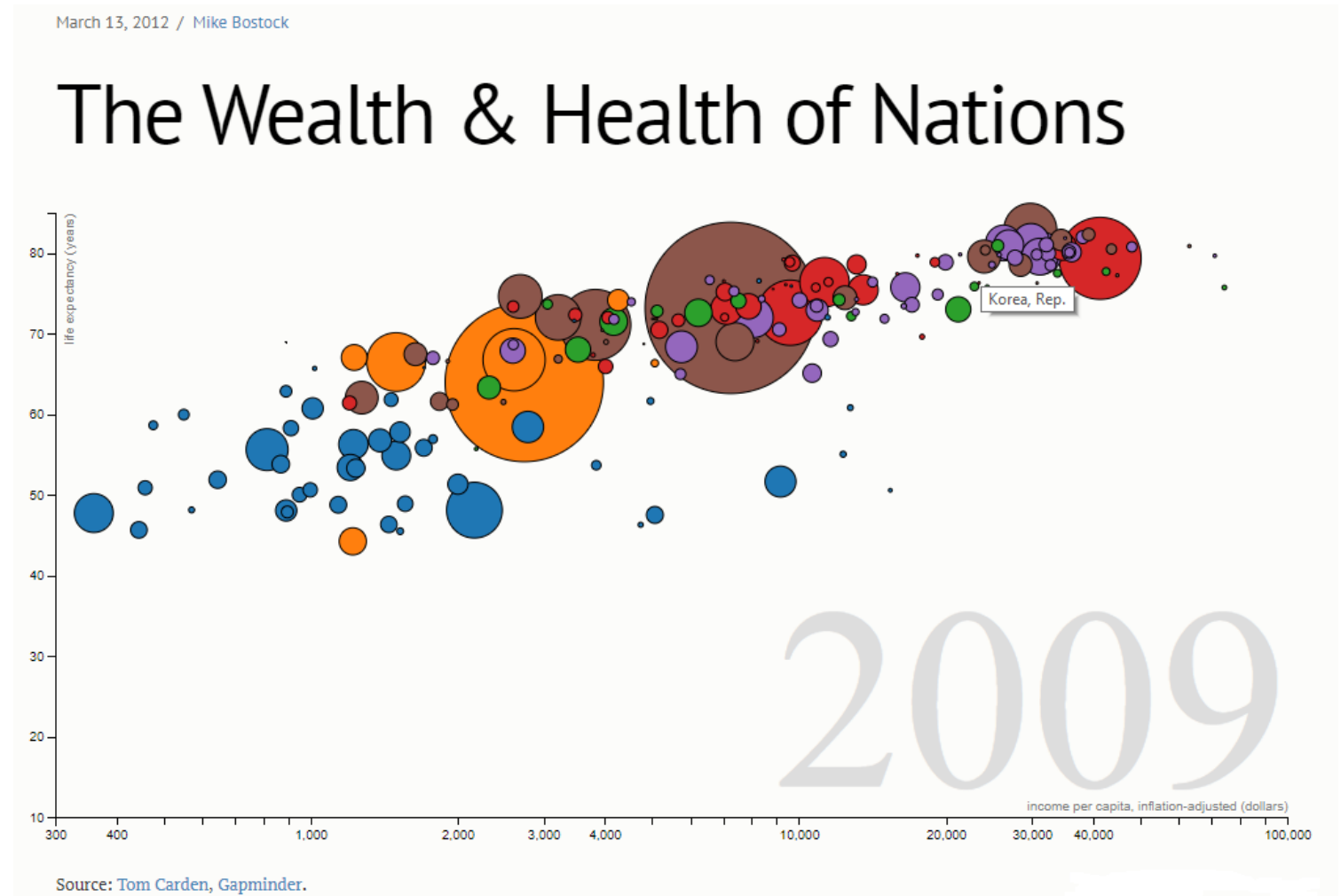
In every data analysis, putting the analysis and the results into a comprehensible report is the final hurdle to take.

TOOLS

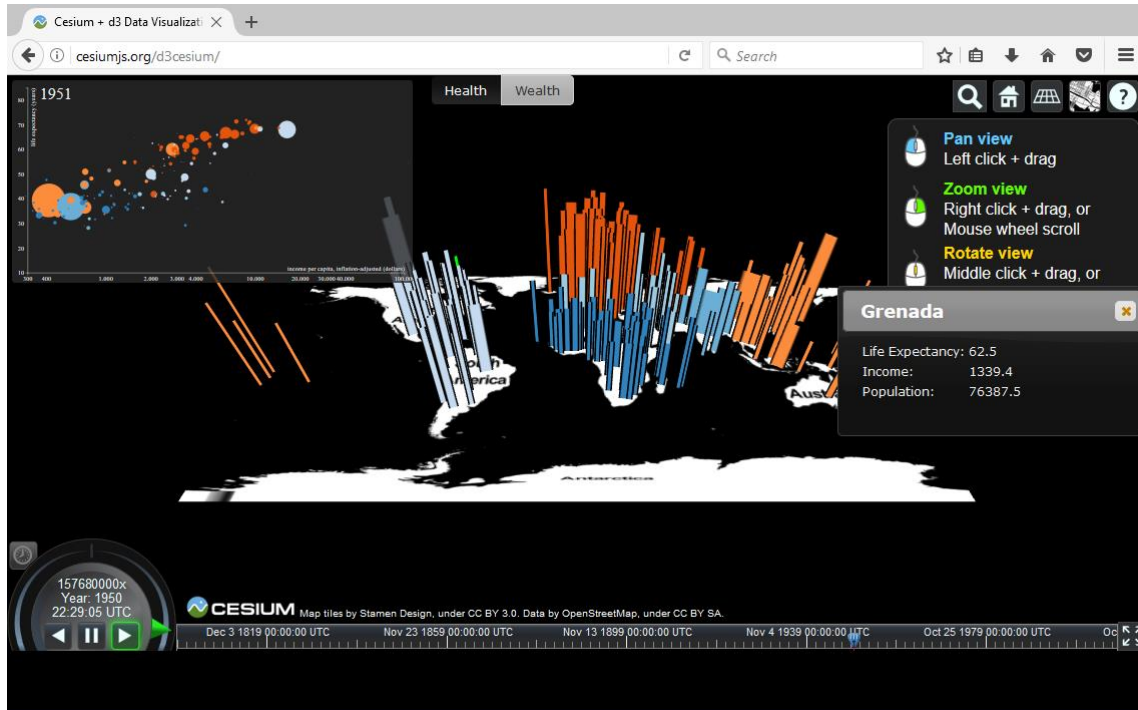
 + a b l e a u

 Spotfire[®]
TIBCO Software

 R Markdown



4. Master Data Munging, Visualization and Reporting (Cont.)



<http://cesiumjs.org/d3cesium/>

Four Ways to Slice Obama's 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

All Spending Types of Spending Changes Department Totals

How \$3.7 Trillion Is Spent

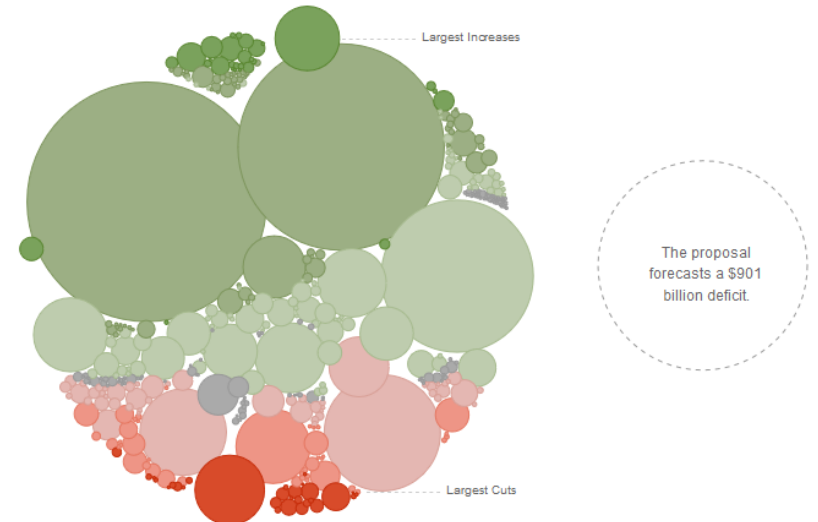
Mr. Obama's budget proposal includes \$3.7 trillion in spending in 2013, and forecasts a \$901 billion deficit.

Circles are sized according to the proposed spending.

\$100 billion
\$10 billion
\$1 billion

Color shows amount of cut or increase from 2012.

-25% -5% 0 +5% +25%

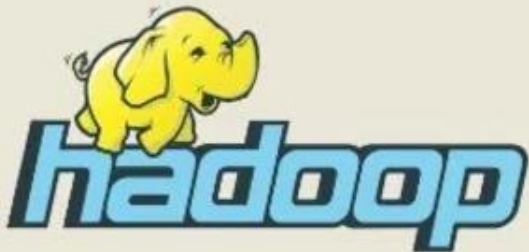


<http://www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html/>

5. Level Up With Big Data

When you start operating with data at the scale of the web, the fundamental approach and process of analysis must change. Most data scientists are working on problems that can't be run on single machines. They have large data sets that require distributed processing.

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.



MapReduce



MapReduce is this programming paradigm that allows for massive scalability across the servers in a Hadoop cluster.

Apache Spark is Hadoop's speedy Swiss Army knife. It is a fast-running data analysis system that provides real-time data processing functions to Hadoop.



6. Get Experience, Practice and Meet Fellow Data Scientists

Practice makes perfect ...



Join in
competitions



Meet fellow data
scientists



Have a pet
project



Develop your
intuition

7. Internship, bootcamp or Get a Job

The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

Internship



BEGINNER

amazon.com[®]

Bootcamp



INTERMEDIATE



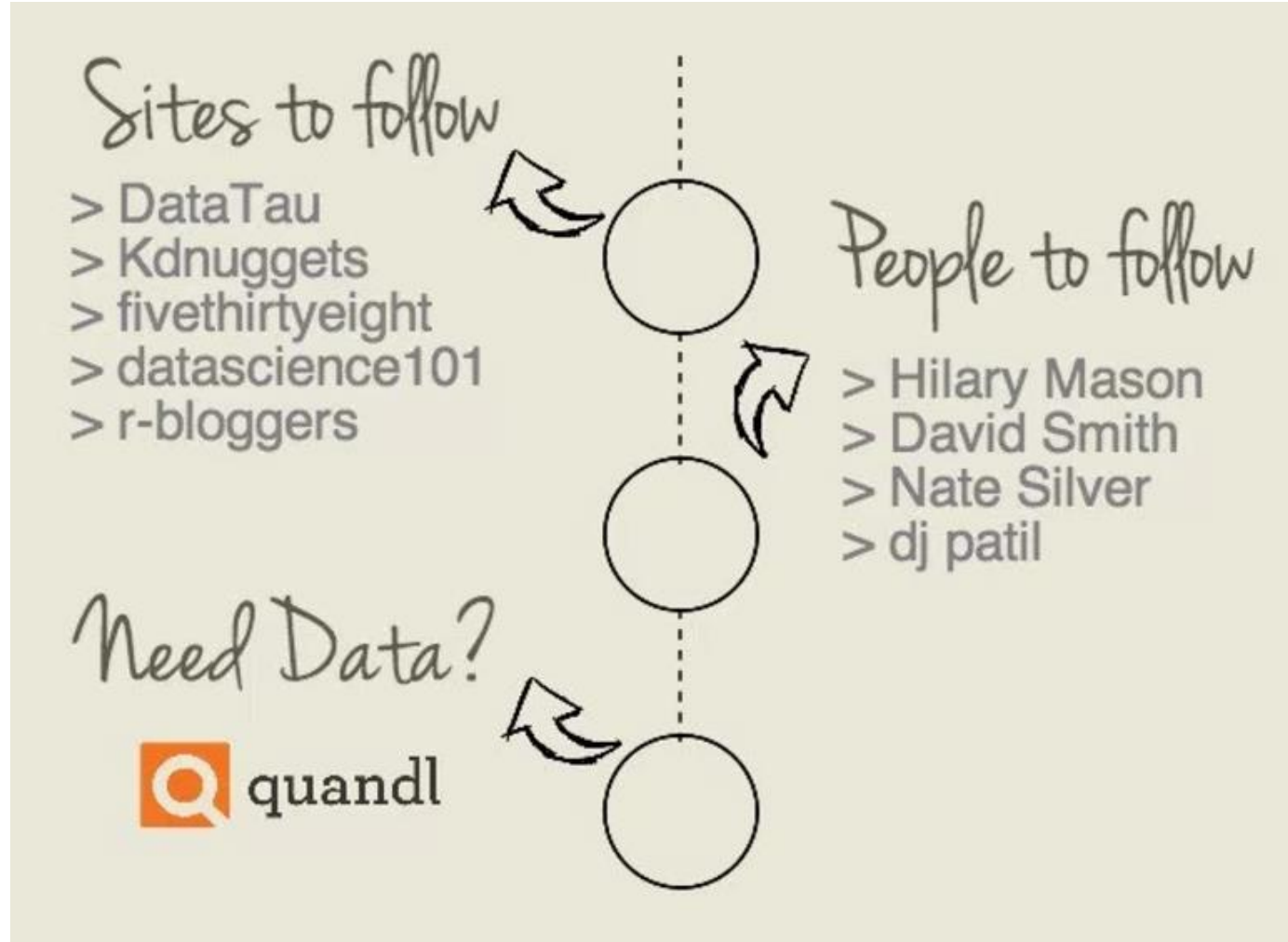
Job



ADVANCED



8. Follow and Engage with the Community



Reference Sites

- 정도현 블로그
 - <http://www.moreagile.net/>
- 빅데이터시대, 데이터 사이언티스트는 어떻게 될 수 있으며 분석언어 R은 왜 필요한가?
 - <https://www.youtube.com/watch?v=Zssf8JWWHY4&t=6s>
- 21세기 가장 섹시한 직업. '데이터 사이언티스트' 를 주목하라
 - <https://www.youtube.com/watch?v=HTsh9ymyozs&t=1s>
- 데이터 사이언티스트의 현실과 미래
 - <https://www.youtube.com/watch?v=K-WeZm09mFU&t=1643s>