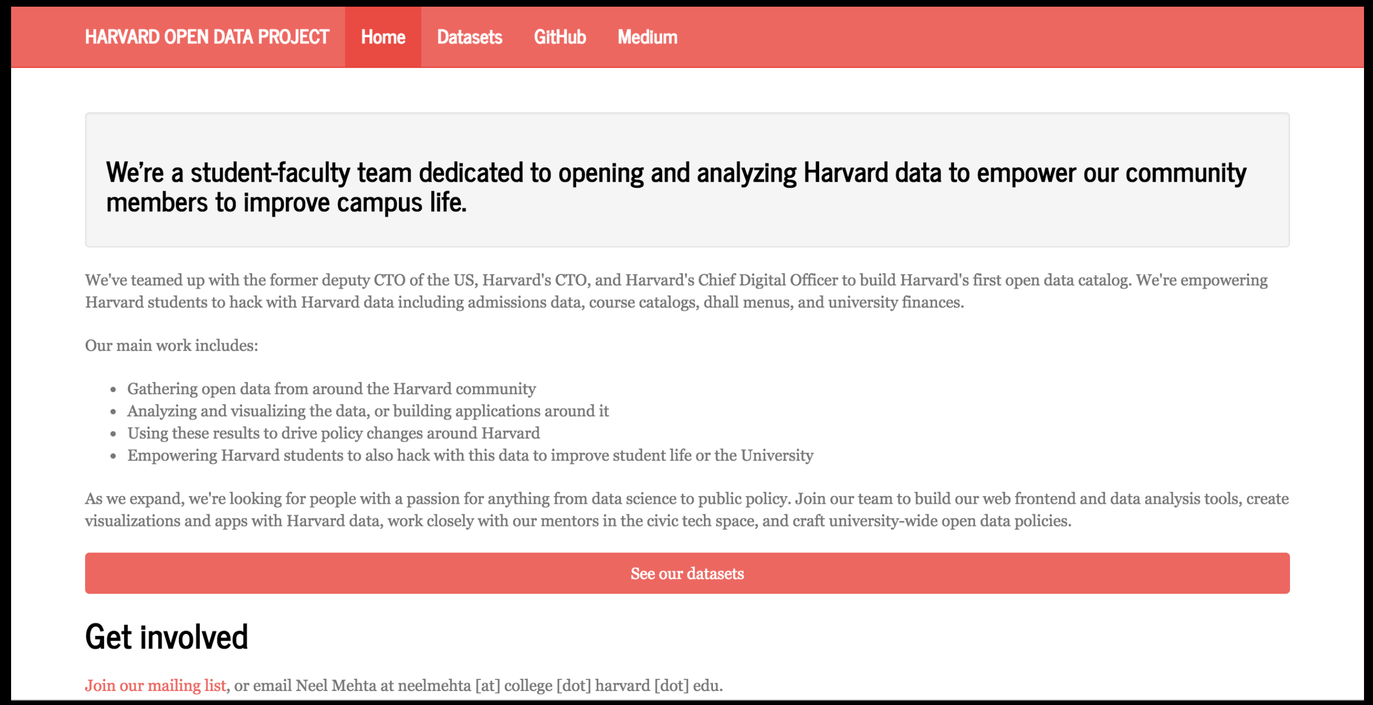
Data Science with Python Pandas

CS50 Seminar

Athena Kan

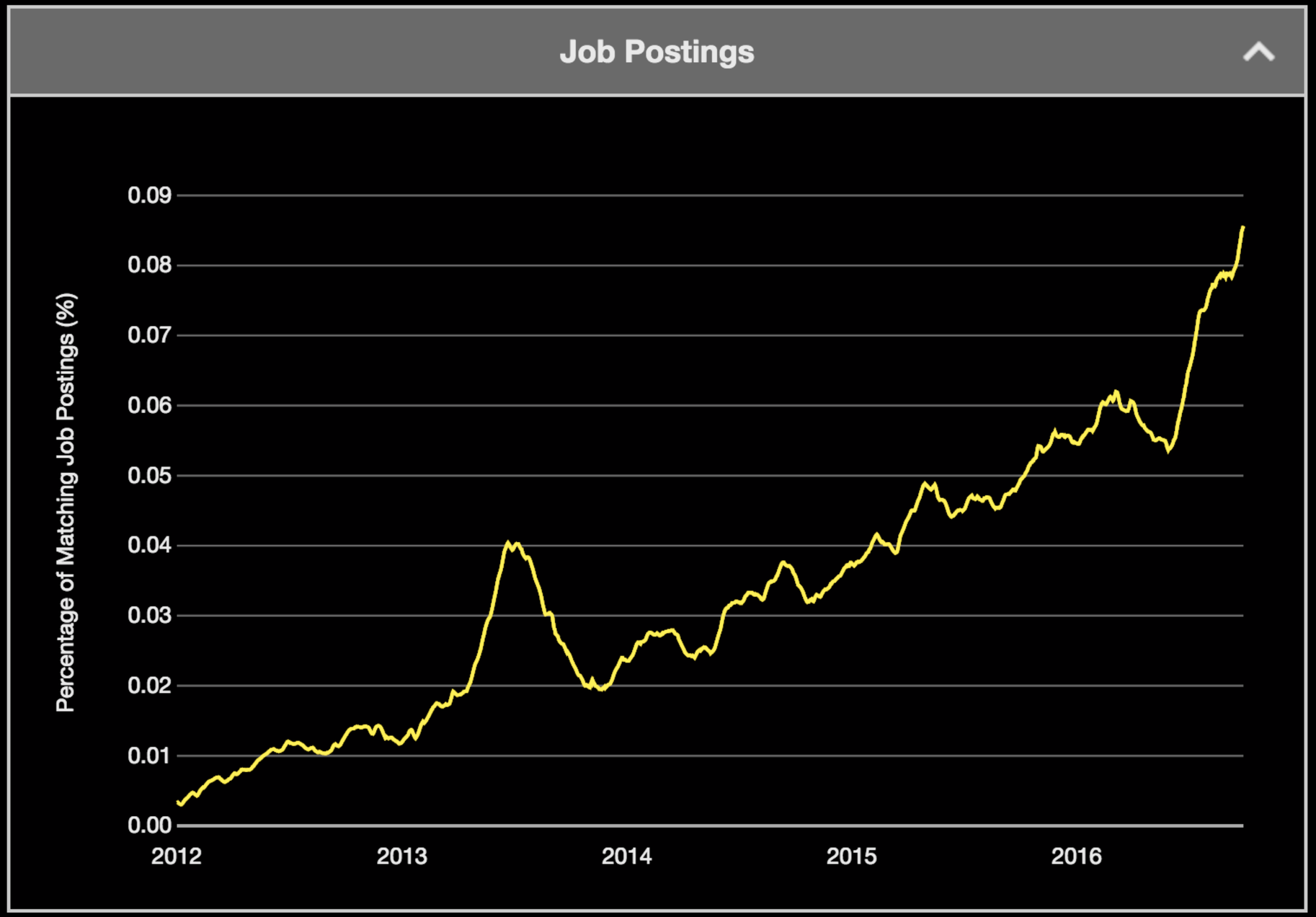
Athena Kan
athenakan@college.harvard.edu

Booz | Allen | Hamilton

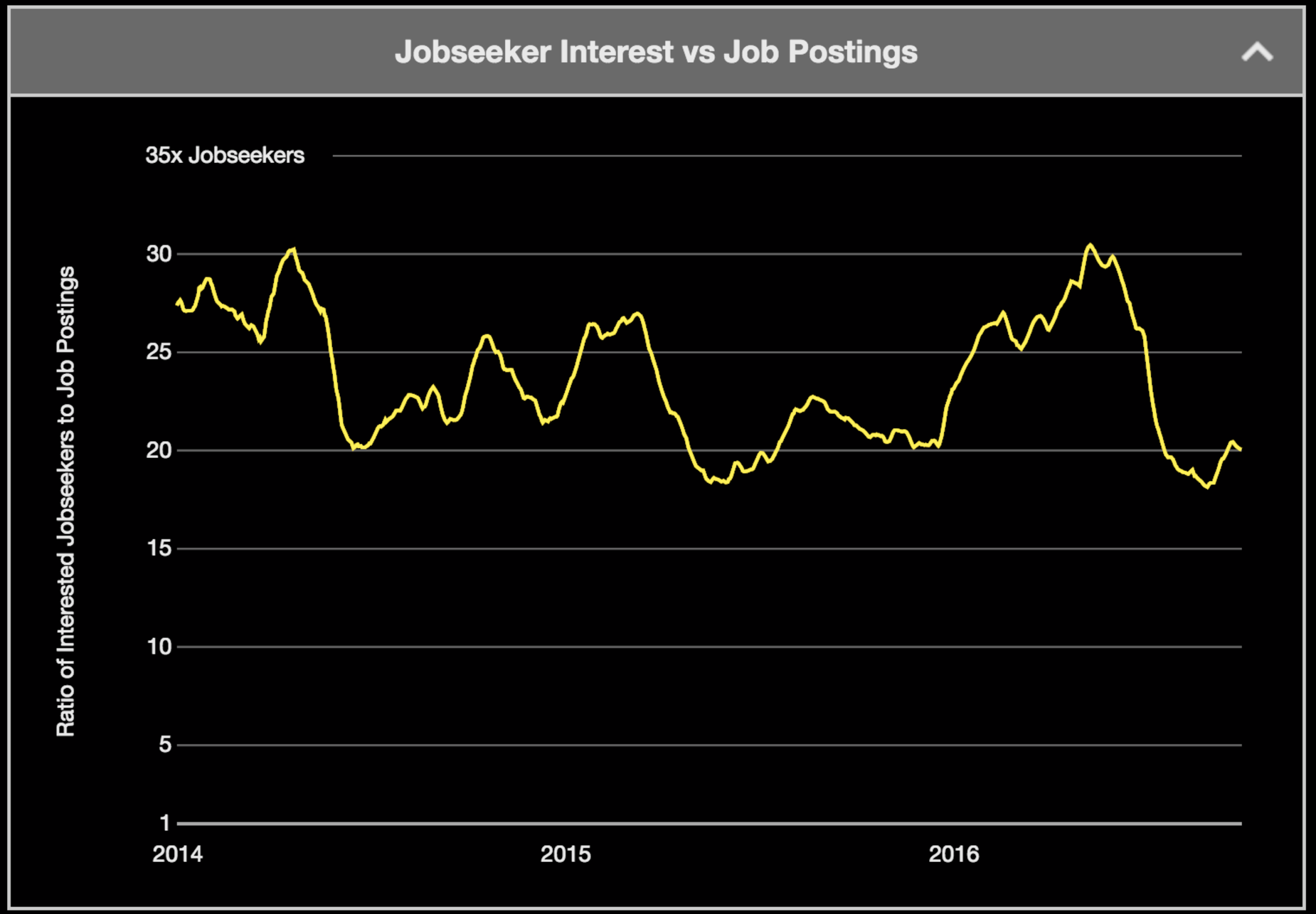


Data Science

- “#1 Best Job in America for 2016” - Glassdoor
- “$116,840 Median Base Salary” - Glassdoor
- “The Sexiest Job of the 21st Century” - HBR



Source: Indeed
http://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

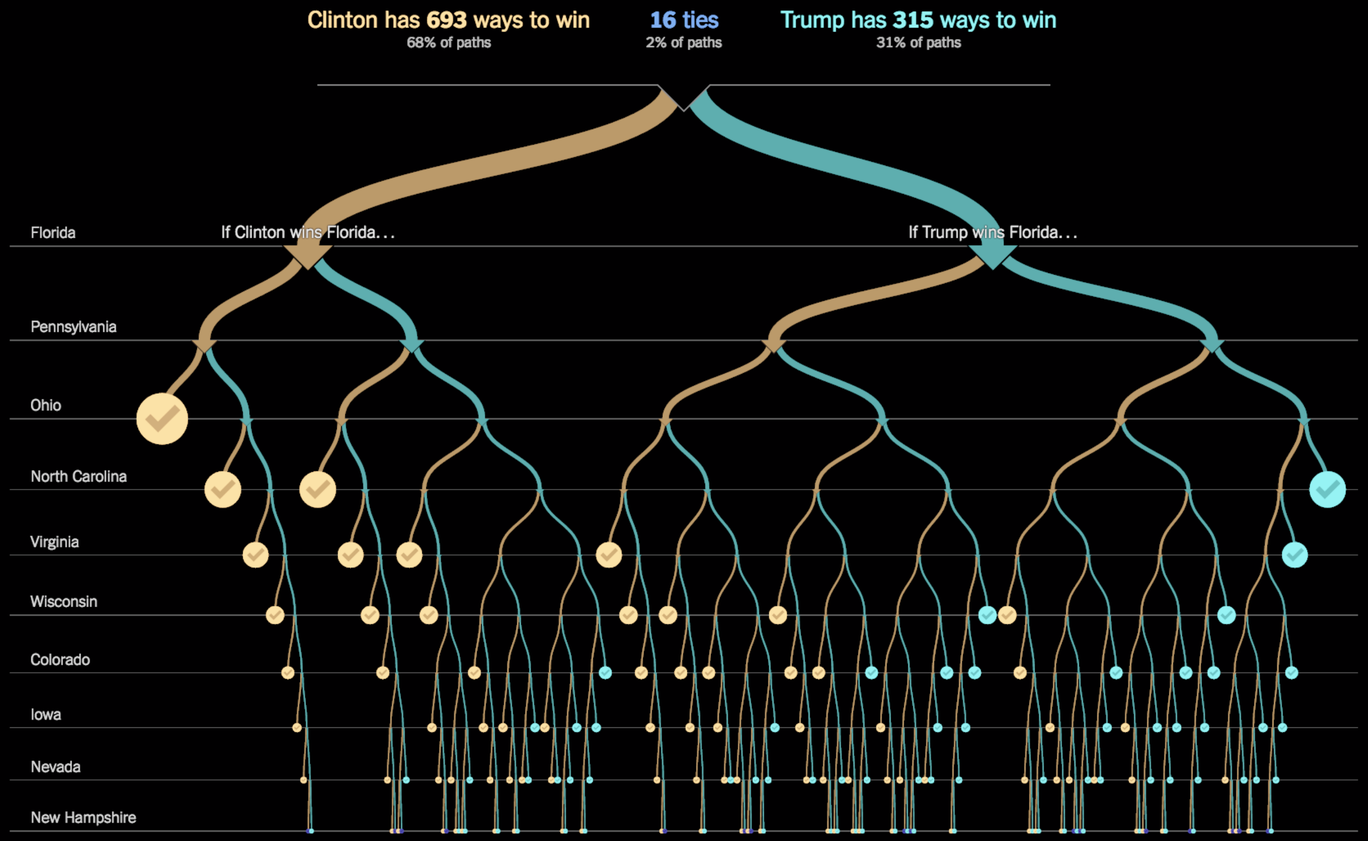


Source: Indeed
http://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

“Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.”

- Stephen Few

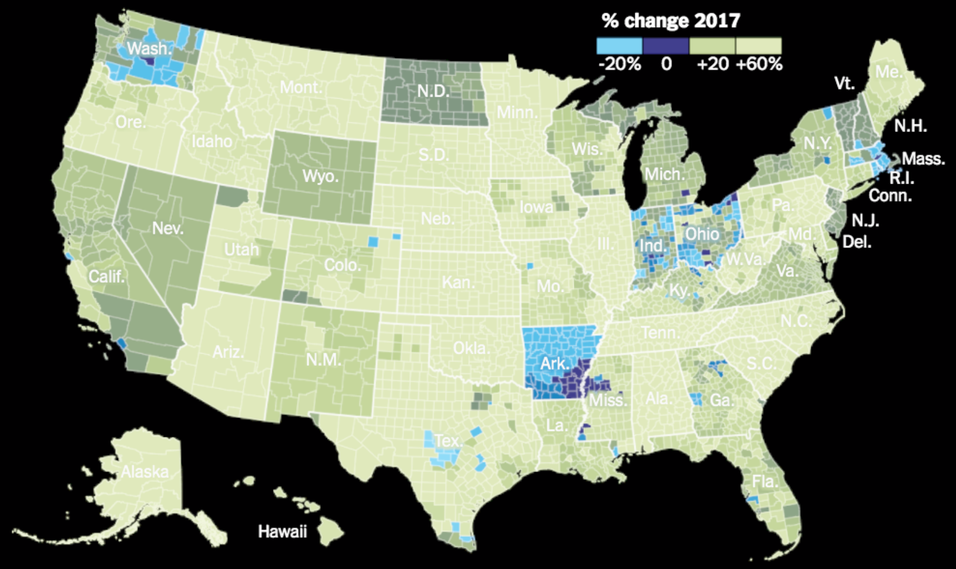
http://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html



Obamacare Rates Are Rising

But there's a lot of variation. In some Arizona counties, prices for the most affordable midlevel plan are going up by 191 percent. In parts of Texas, premiums are going down by 30 percent.

2017 premium increase for lowest-cost silver plan



Note: There are two separate markets in Los Angeles County. We have shown rates for the market where more customers are enrolled.

Source: McKinsey Center for U.S. Health System Reform

Ex-CIA Analyst Tells How Data Caught Osama bin Laden

www.emergencymgmt.com/safety/Ex-CIA-Analyst-**Data-Osama-bin-Laden**.html

Feb 26, 2014 - Ex-CIA Analyst Tells How **Data Caught Osama bin Laden** ... Bakos explains how small yet actionable **data** tactics led to bin Laden's **capture**.

How Big Data helped capture Osama Bin Laden – Tenece Idea Mine

https://teneceblog.wordpress.com/.../how-big-**data**-helped-**capture-osama-bin-laden**/

Sep 10, 2015 - A lot has been said about Big **Data** and its inherent uses or applications in the finance, telecommunications, manufacturing and advertising ...

Data Science

1. Ask a question
2. Get the data
3. Explore
4. Model
5. Communicate

Data Science

1. Ask a question - experience, Kaggle, experts
2. Get the data - scraping, databases, Excel/CSVs
3. Explore - pandas, matplotlib, numpy
4. Model - pandas, sk-learn
5. Communicate - matplotlib, d3.js

Pandas!

Pandas

- Python library
- For data cleaning, analysis, visualization, and other analysis
- Well-suited for many kinds of data
- Built upon numpy and integrates well with other libraries

Jupyter Notebook

- Web application
- Allows live code, visualizations, text
- Supports over 40 languages, interactive widgets, and big data
- Can share notebooks

Series

- One dimension array-like object
- Capable of holding any data type
- Has an index

DataFrame

- Two dimensional tabular data structure
- Capable of holding any/many data types
- Index and columns

Data Science

1. **Ask a question**
2. Get the data
3. Explore
4. Model
5. Communicate

How have Earth surface temperatures changed over time?

Data Science

1. Ask a question
2. **Get the data**
3. Explore
4. Model
5. Communicate

https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

Data Science

1. Ask a question
2. Get the data
3. **Explore**
4. Model
5. Communicate

Data Science

1. Ask a question
2. Get the data
3. Explore
4. **Model**
5. Communicate

Other Resources

- Harvard Open Data Project
- Kaggle
- data.gov, data.cityofboston.gov
- Data Ventures, CS109a/b, CS181

My Data Science Projects

- Predicting NBA draft order from college stats
- Developing an intrusion detection system in industrial control systems
- Mapping Instagram friends based on mutual interactions
- Predicting diabetes subtypes based on biometric data
- Predicting urban demographic changes (temporal, geographic)