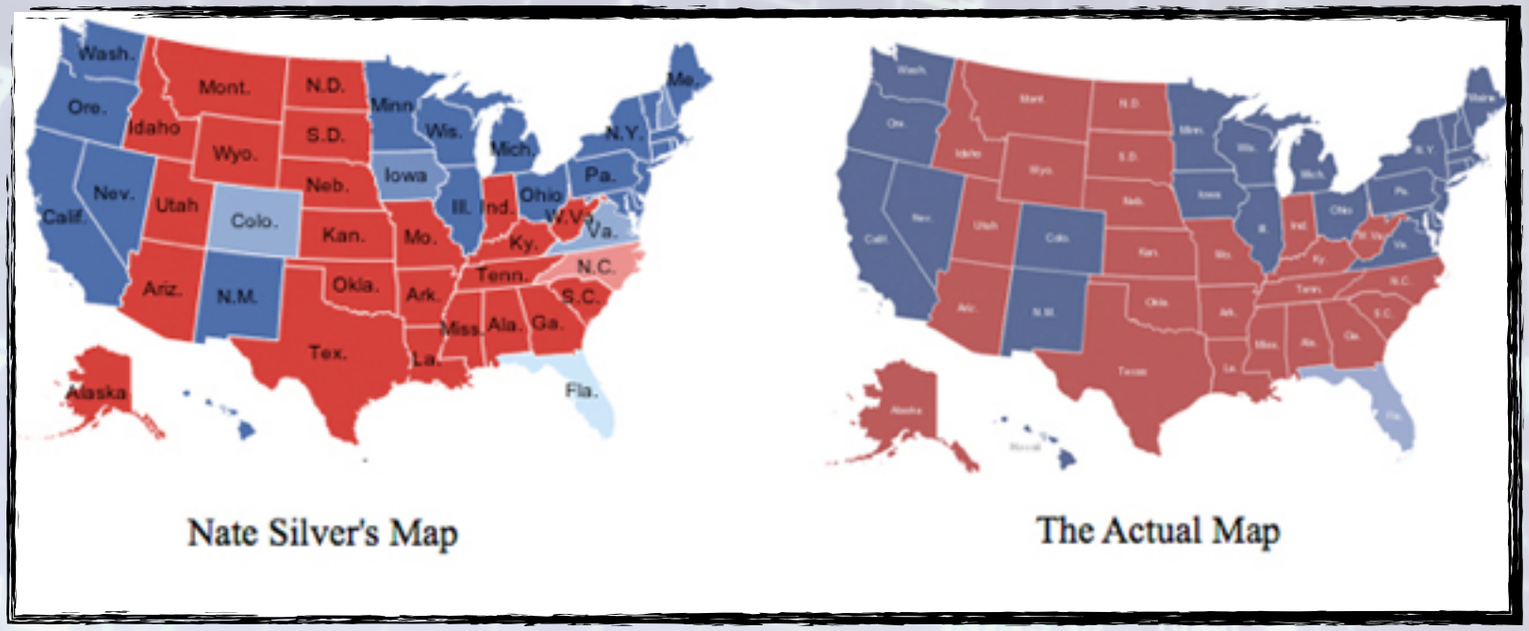
From Rocket Science to Data Science

Sanghamitra Deb
Data Scientist, Accenture Tech Lab

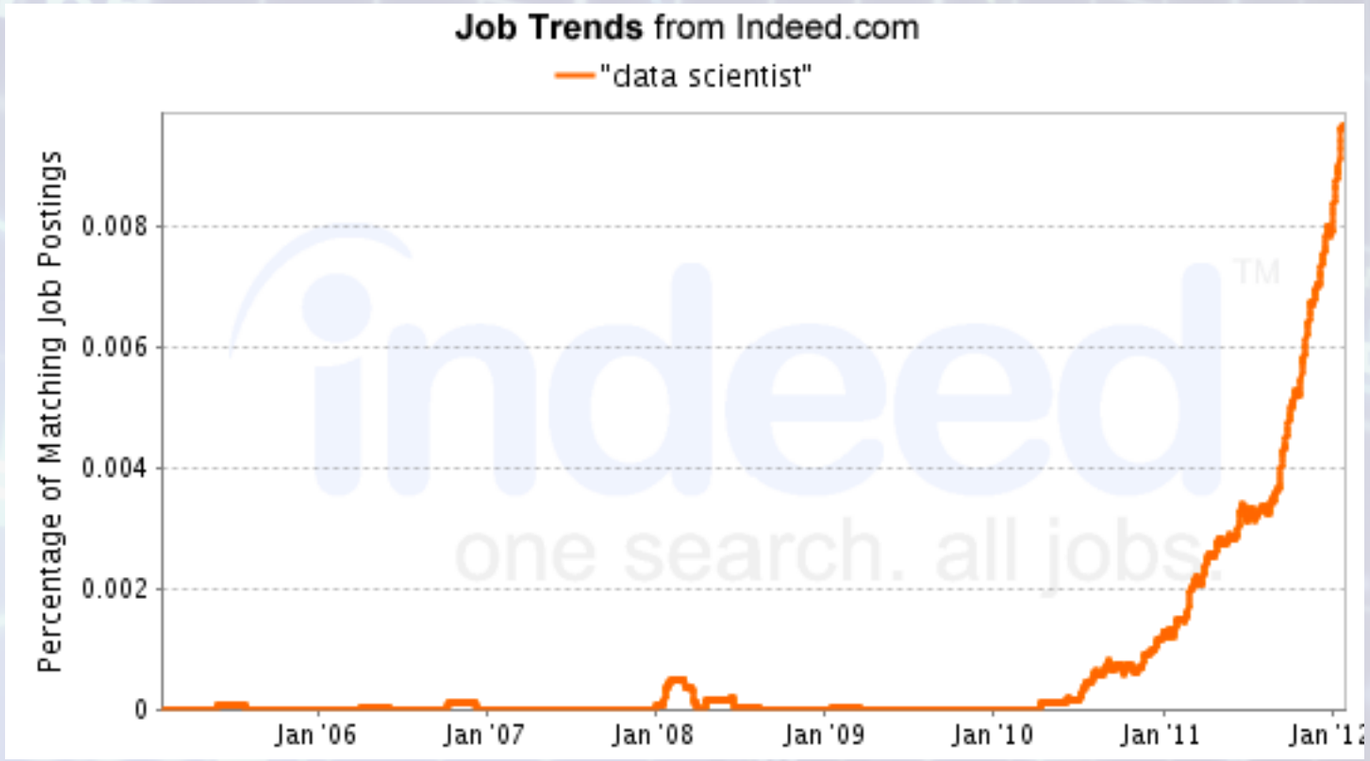
Sexiest Job of the 21st century



Nate Silver predicted correctly how all 50 states would go in the presidential election 2012

Target predicted teen pregnancy from retail data.

The Big Data Challenge



"With the need for data scientists growing at about 3x those for statisticians and BI analysts…. and an anticipated 100,000+ person analytic talent shortage through 2020… "

"… three core data science skills: data management, analytics modeling and business analysis. But beyond these, there's an art to data science. We detail several soft skills that our research showed are also critical to success, i.e., communication, collaboration, leadership, creativity, discipline and passion (for information and truth)."

gartner article

Who are you?

- Front Engineer UX/UI
- Backend engineer
- Project Manager
- Academic (PhD, physics, neuroscience, economics, … of course CS) trying to find a niche in tech industry
- Everyone else …

Start a data driven project relevant to the industry you want to join

Where to start

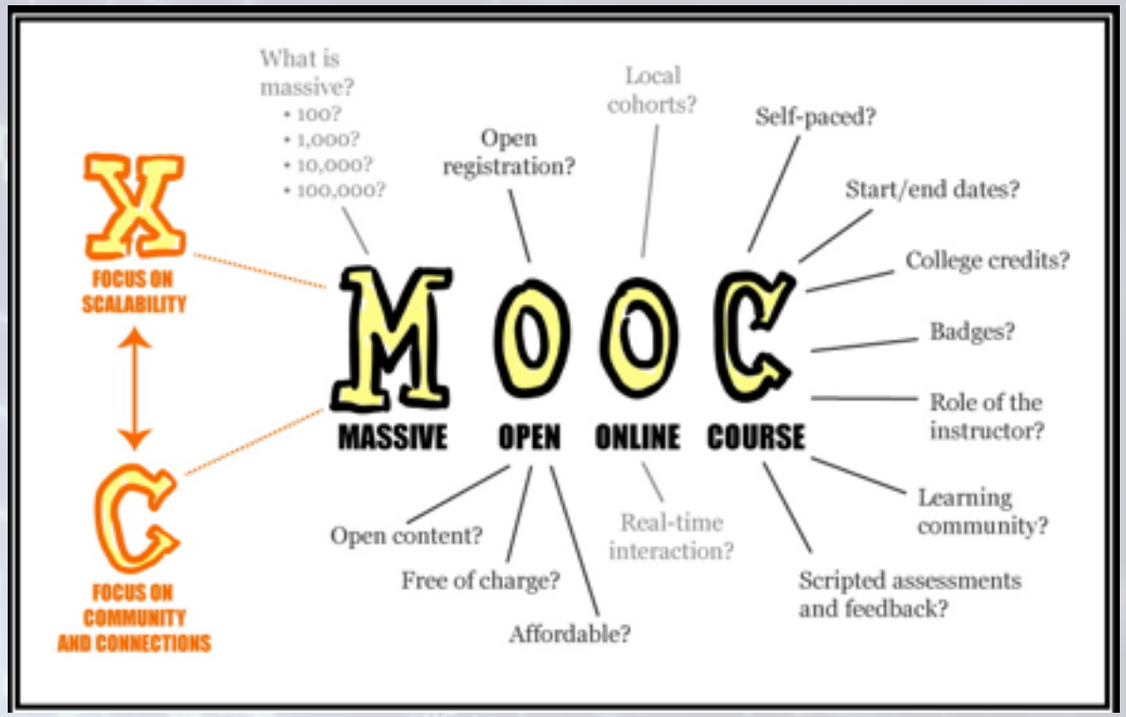
blogs: yhat, data robot, datatau, upshot …

twitter: follow data science news…

Data Exploration/Discovery … open a dataset in your favorite coding language: **Python**, R , scala, julia, …

Learn to pipe data in to a database such as **MySQL**/ MongoDB

Kaggle competitions, live and older ones… e.g.: digit recognition, titanic



Do a few online courses on data science, big data, machine learning, python, R, … from coursera, udemy, khan academy, … form study groups, go to meetups.

pros: DIY , bite size videos, flexibility, discussion forums, interactivity, great way to figure out if a new field is interesting.

cons: DIY, choosing the correct course, signing up and not participating after first few weeks.

Small Data Project Flow

Get open source data.
Sources: city data (SanFrancisco, LA, Seattle, Chicago, transit data,…)

Load it up on Python, if the data is too big I will put in MySQL (for structured data) or mongodb for free form json.

Ask the right Question!!!

Machine Learning, Statistics , counting statistics and **histogram** are very powerful. If you are a python user data frameworks such as “GraphLab” is open source & easy to learn.

Create a dashboard/ viz/app

Data Wrangling/Cleaning

- Open your data set and profile it
- Look for missing data, bad data points vs true outliers
- Pattern of your data, is it a phone number, timestamps or a social security number? is it structured data or unstructured text
- Prep your data, identify the features that influence your outcome, feature selection and feature engineering.

Lets start …

Question: **What is a Data Scientist?**

Data : scraped indeed.com for all jobs containing “data” in the title. ~5000 jobs …

Meta Data: Job title, job description, city, state

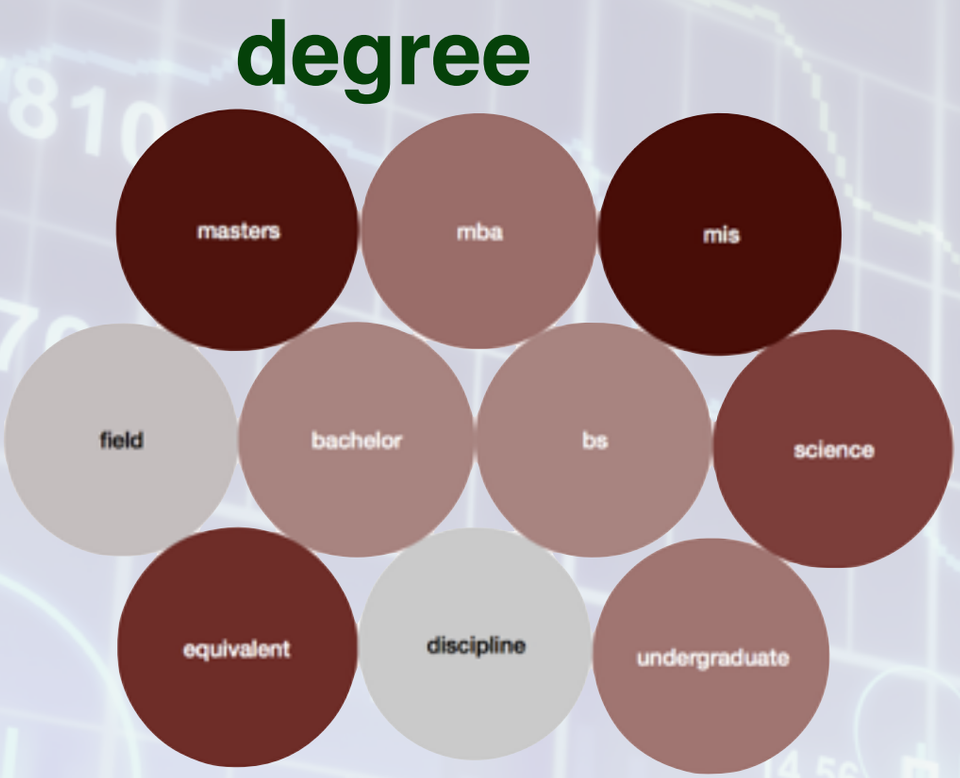
job description: unstructured text…

Job Title

Job Description

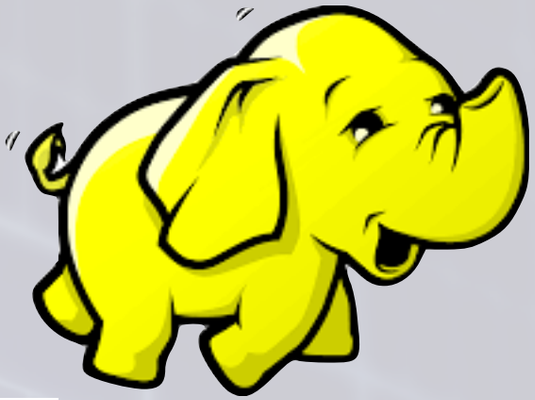
text cleaning+ Bag of words

What do the job descriptions mean?

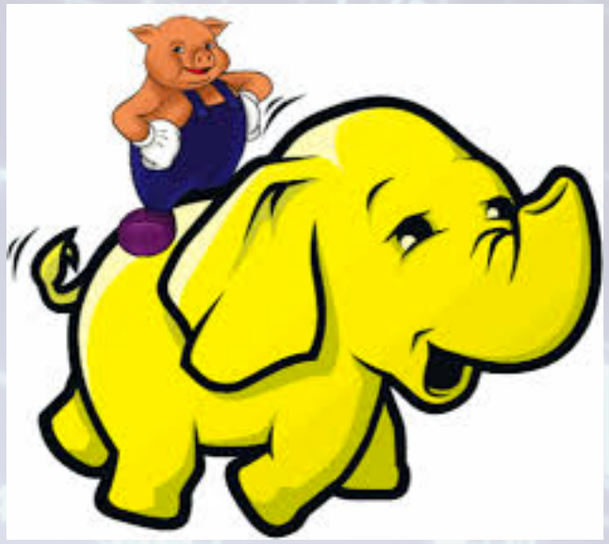


Algorithm: word2vec synonym

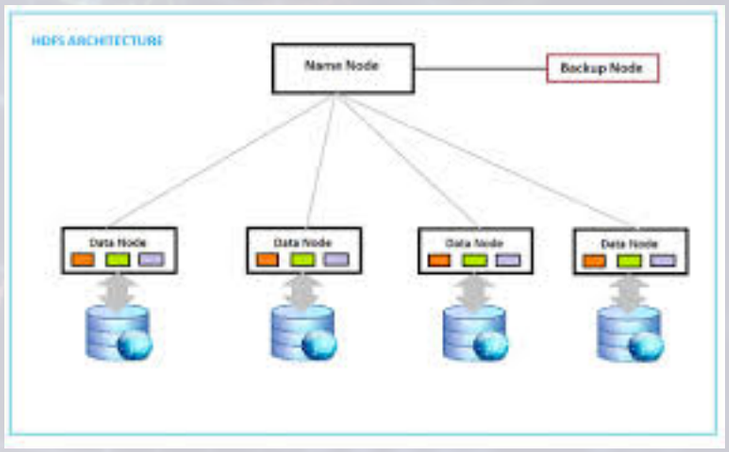
Hadoop



APACHE HBASE



Apache Solr



Spark

mongoDB

Storm

Algorithm: word2vec synonym

Python

Algorithm: word2vec synonym

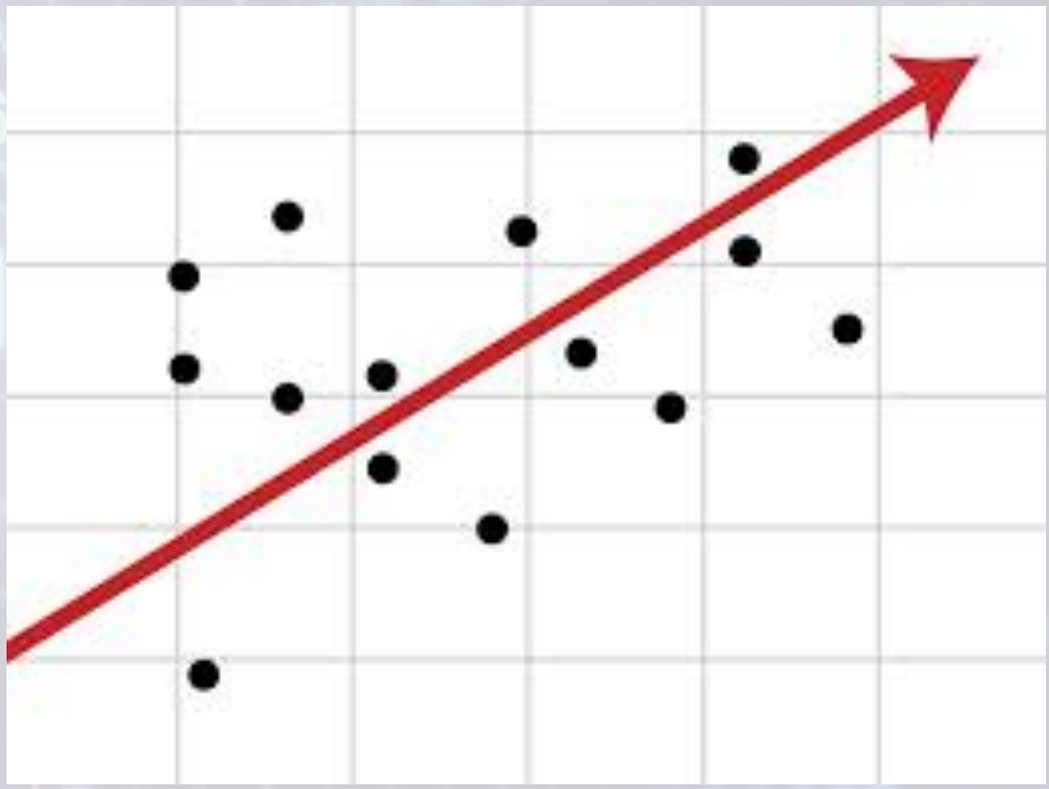
Statistics

- (u'**mathematics**', 0.8544293642044067),
- (u'**economics**', 0.8378890752792358),
- (u'applied', 0.8295730948448181),
- (u'**physics**', 0.8211749792098999),
- (u'**math**', 0.8039191961288452),
- (u'**quantitative**', 0.8003592491149902),
- (u'**phd**', 0.795414388179779),
- (u'fields', 0.7486724257469177),
- (u'**science**', 0.7226663827896118),
- (u'**masters**', 0.7045900225639343)

Algorithm: word2vec synonym

Regression

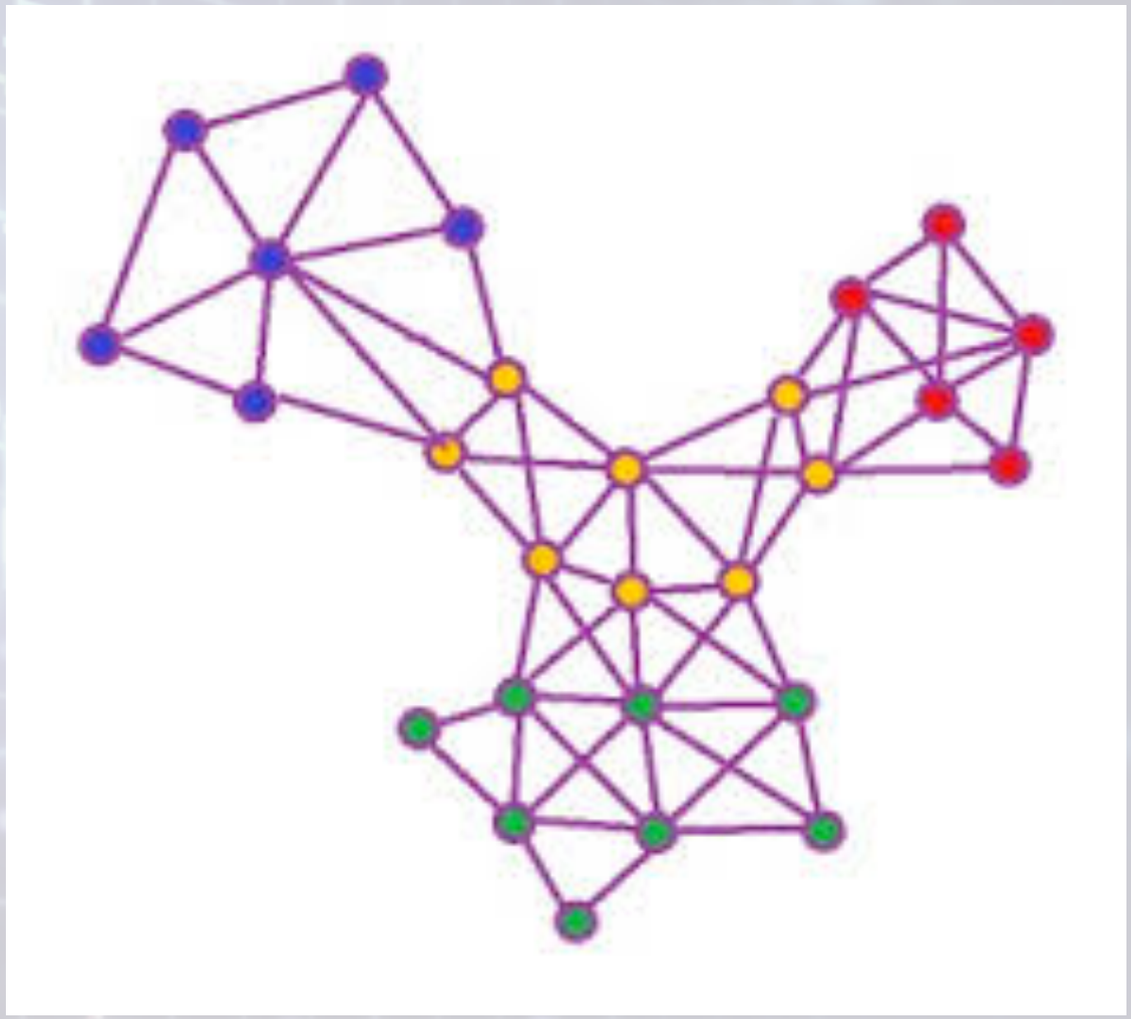
- [(u'segmentation', 0.7036155462265015),
- (u'statistical', 0.688355267047882 1),
- (u'mining', 0.6801210045814514),
- (u'graph', 0.6701105237007141),
- (u'algorithm', 0.669587850570678 7),
- (u'theory', 0.656344771385192 9),
- (u'predictive', 0.647478222846984 9),
- (u'matlab', 0.635683774948120 1),
- (u'recommendation', 0.620379328727722 2),
- (u'analyses', 0.611992478370666 5)]



Algorithm: word2vec synonym

Graph

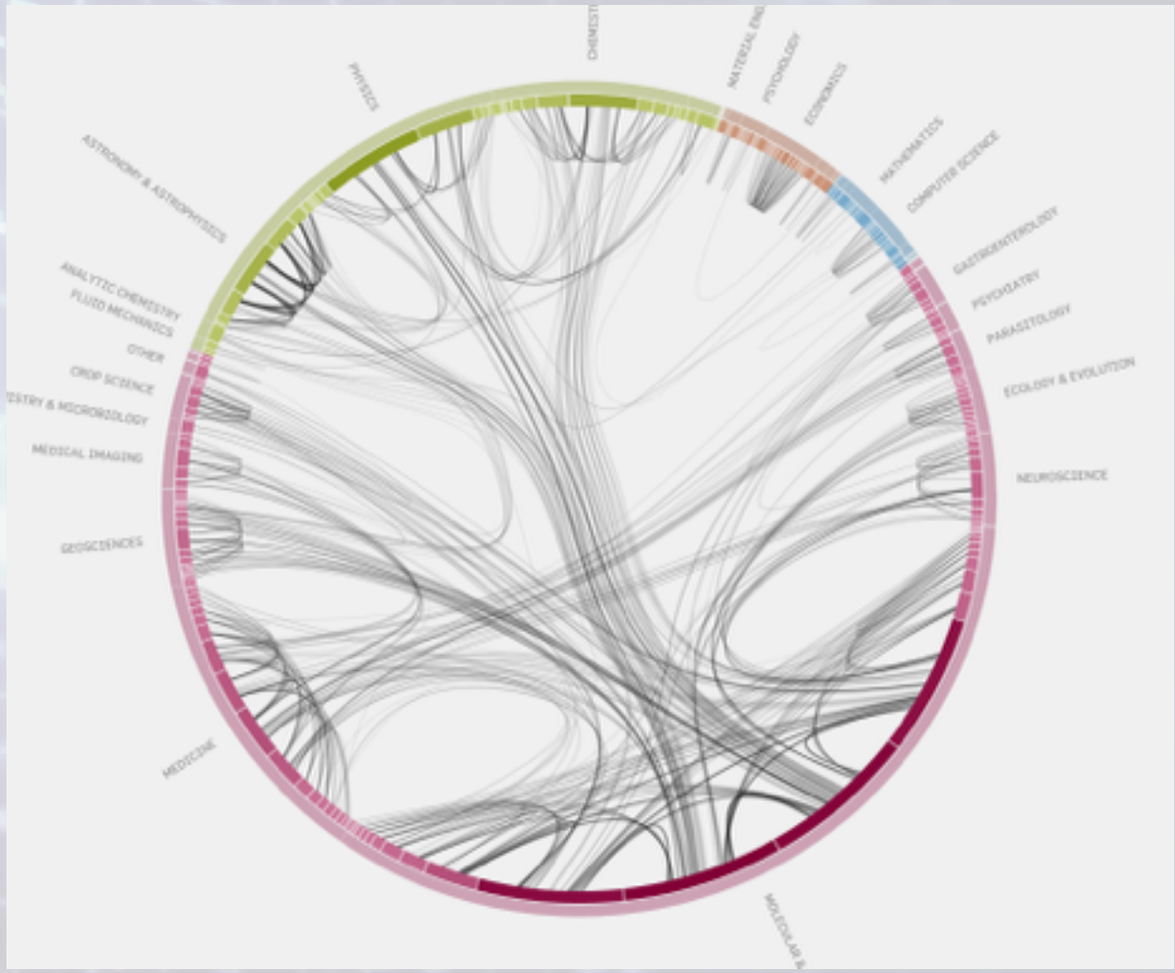
- (u'text', 0.7591882944107056),
- (u'manipulating', 0.716569185256958),
- (u'visualization', 0.7084065675735474),
- (u'matlab', 0.7055898904800415),
- (u'mining', 0.700824499130249),
- (u'unstructured', 0.6868686676025391),
- (u'regression', 0.6701105833053589),
- (u'algorithms', 0.669179141521453 9),
- (u'natural', 0.663329839706420 9),
- (u'engines', 0.6632224321365356)



Algorithm: word2vec synonym

Visualization

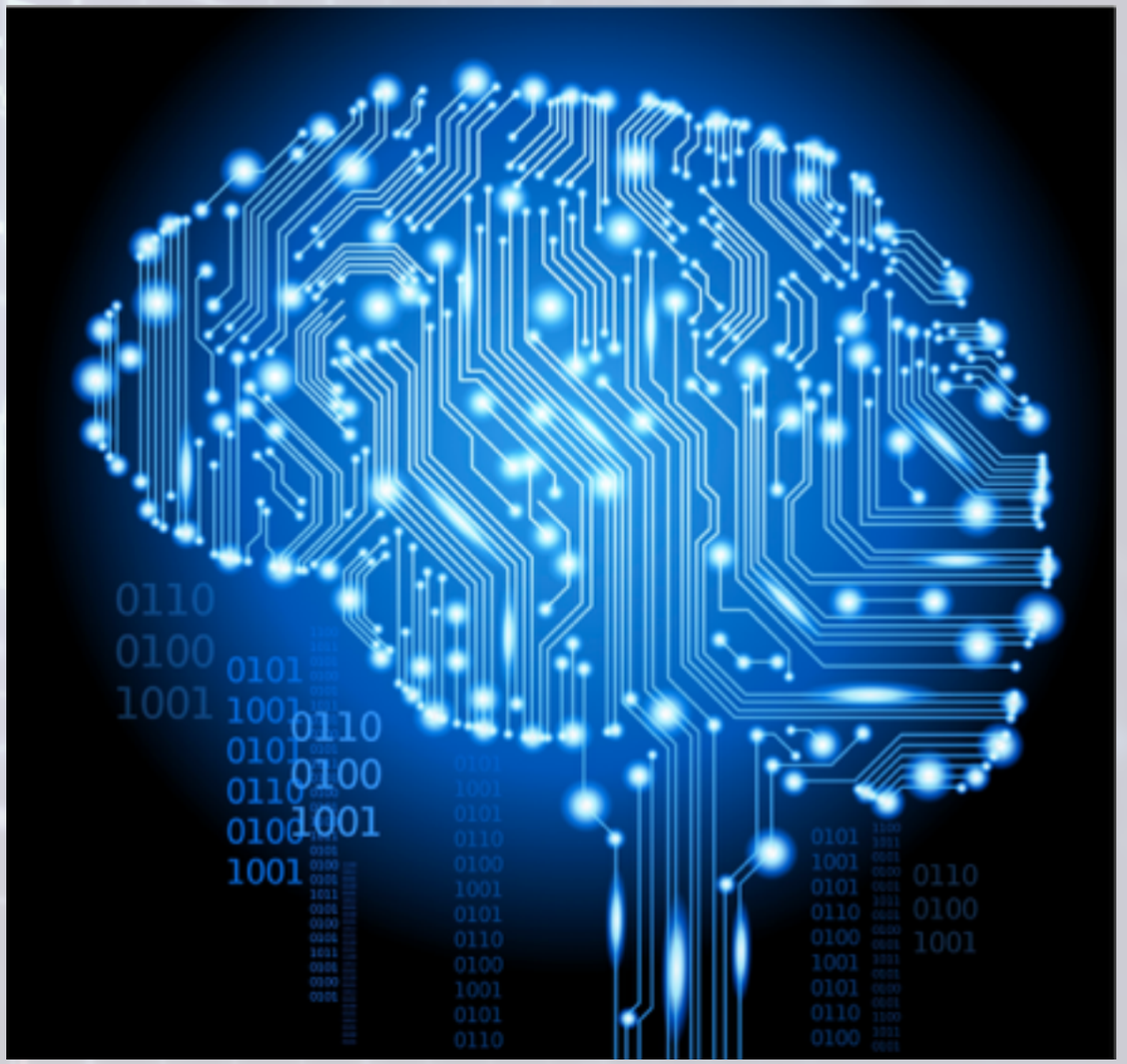
- [(u'**tableau**', 0.7196237444877625),
- (u'**graph**', 0.7084065675735474),
- (u'**matlab**', 0.6993618011474609),
- (u'libraries', 0.6821463108062744),
- (u'visualizations', 0.6746233701705933),
- (u'mining', 0.6517949104309082),
- (u'**spss**', 0.651625394821167),
- (u'text', 0.6145033836364746),
- (u'**qlikview**', 0.6053836345672607),
- (u'**js**', 0.5960412621498108)]



Algorithm: word2vec synonym

Machine Learning

- (u'learning', 0.8338875770568848),
- (u'algorithms', 0.7662283182144165),
- (u'natural', 0.7161275744438171),
- (u'**physics**', 0.707731306552887),
- (u'mining', 0.6965328454971313),
- (u'ideally', 0.6682661175727844),
- (u'**graph**', 0.659676671028137 2),
- (u'**predictive**', 0.656450629234314),
- (u'applied', 0.6529620885848999),
- (u'**statistics**', 0.6500071883201599)



Algorithm: word2vec synonym

Digging deeper …

- Create a **data story**, i.e put all the visualizations and insights in a dashboard create an infographic using tableau, d3 , …

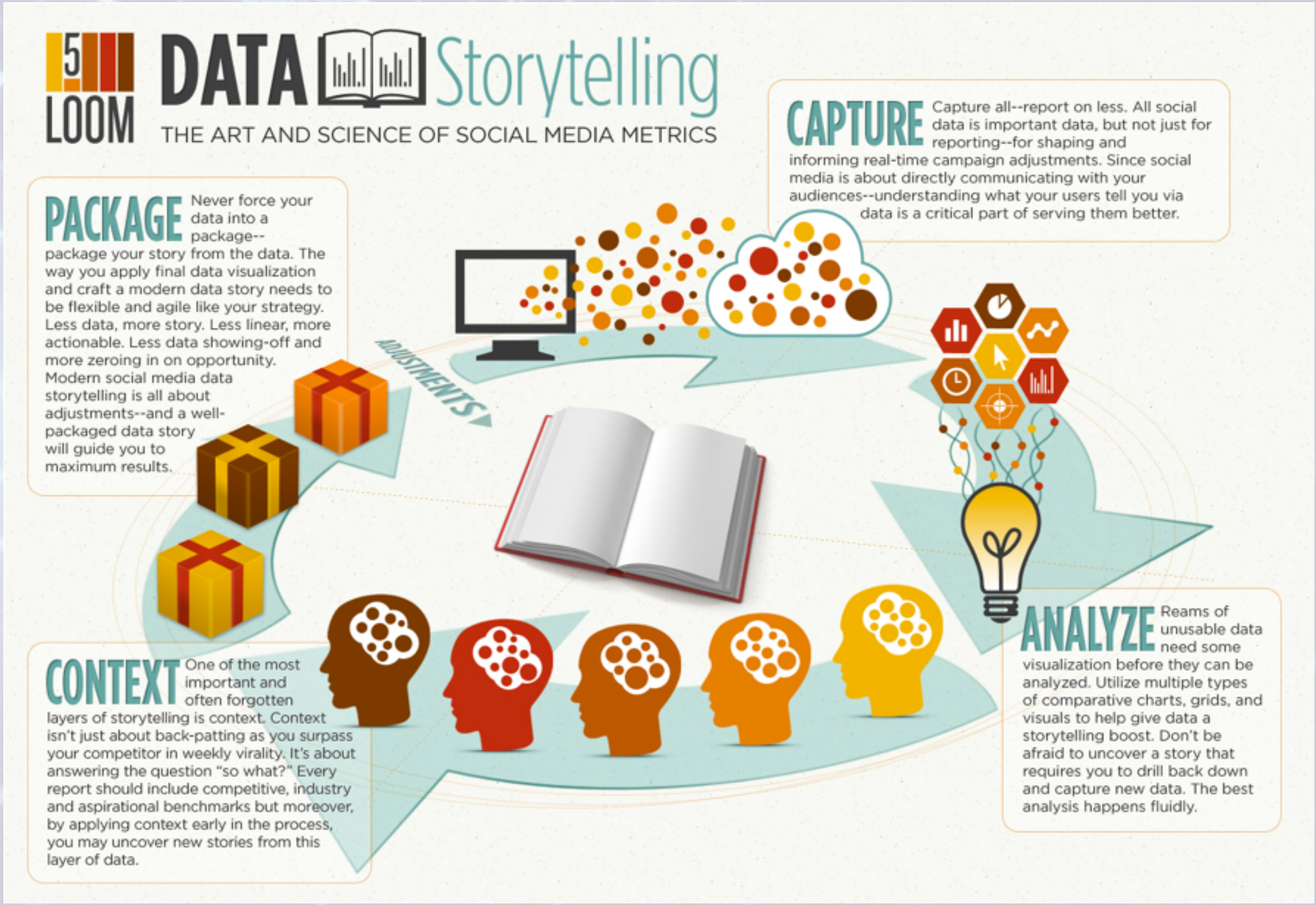
- Get data (say from crunchbase) on the companies that are hiring and figure out which **industries** dominate in the **data world**

- Get data for atleast the past 6 months and have **exact statistics** for skills in the data world. Advanced text analytics (bi-gram, tri-gram modeling, topic modeling)

- Create an app that gives tells you how **“hot”** your skills are and what skills are easiest for you to acquire to become **“hotter”**.

Ask the right questions

Data Story Telling



Managing

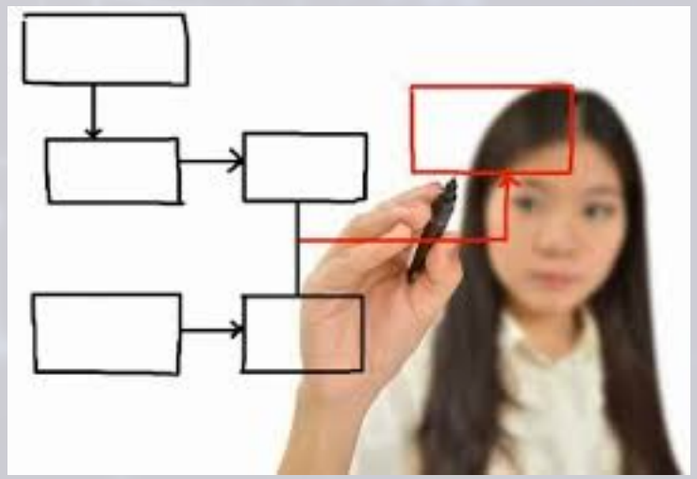
What is a data scientist?

- **Mean, Median, Mode, Box plots…**
- **Hypothesis testing, probabilty, bayesian inference, …**
- **Data Query and storage (sql)**
- **Regression (linear/logistic)**
- **Naive Bayes**
- **Random Forests**
- **Decision Trees and Classifiers**
- **NLP, LDA, …**
- **Capacity to understand the domain and make intuitive decisions. Ability to communicate with domain experts and incorporate the knowledge in data analysis, ability to deal with edge cases ….**

http://berkeleyjess.blogspot.com/2015/04/what-is-data-scientist.html

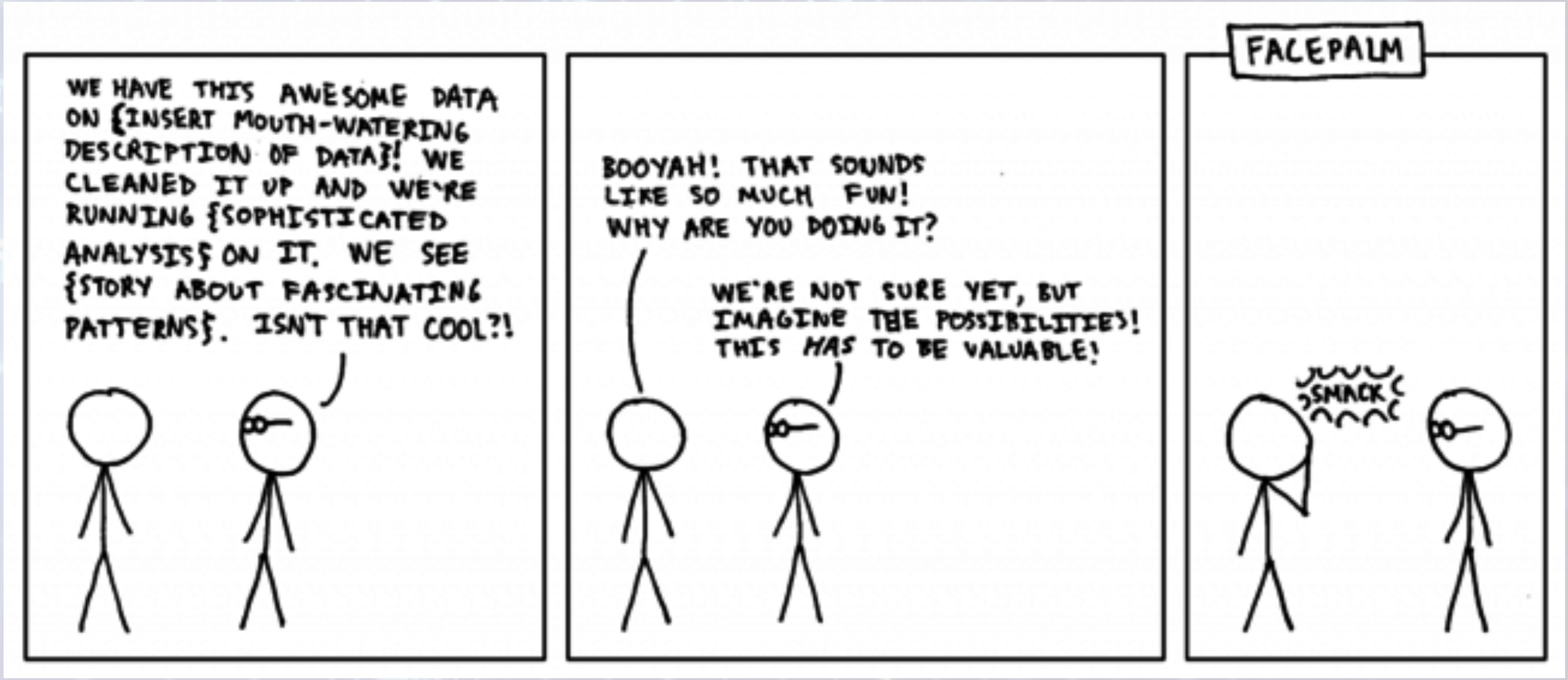
Interview Process

- 4-5 hours long
- Depending on company size 4-6 people
- Statistics white boarding … A/B testing calculations,
- Formulation of a machine learning use case with parameter tuning, edge cases relevant to the company
- Open question that the team is trying to solve
- CS Algorithms … cracking the coding interview.
- Databases, SQL queries, …



http://deblivingdata.net/wp-content/uploads/2014/05/DSTalk.slides.html

Now that you have landed the job …



References

- Data Sources: data.gov, kaggle, open city data
- Volunteering opportunities: Datakind, BayesImpact, Data For good
- DS Schools: Insight Data Sciences, Zipfian Academy, …
- Mode Analytics have an awesome SQL school where you can learn SQL
- meetup.com

Thank You

@sangha_deb,deblivingdata.net,sangha123.github.io