

1/20 5:07:31 ***



Data Science



What is Data Science?

Let us begin at the beginning, and define what data science is.. 



v1



"Data science is an emerging interdisciplinary field that combines elements of mathematics, statistics, computer science, and knowledge in a particular application domain for the purpose of extracting meaningful information from the increasingly sophisticated array of data available in many settings."  



v2

From Wikipedia: 

Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data



engineering, **pattern recognition** and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing. Methods that scale to **big data** are of particular interest in data science, although the discipline is not generally considered to be restricted to such big data, and big data solutions are often focused on organizing and preprocessing the data instead of analysis. The development of machine learning has enhanced the growth and importance of data science.



Data science affects academic and applied research in many domains, including machine translation, speech recognition, robotics, search engines, digital economy, but also the biological sciences, medical informatics, health care, social sciences and the humanities. It heavily influences economics, business and finance. From the business perspective, data science is an integral part of **competitive intelligence**, a

newly emerging field that encompasses a number of activities, such as data mining and data analysis.


v3



Data science is OSEMN ('awesome') – it involves Obtaining, Scrubbing, Exploring, Modeling, and iNterpreting data – Jeroen Janssens.



v4

Data Scientist (noun): Person who is better at statistics than any software engineer and better at software engineering than any statistician – Josh Wills. 

FYI tidbit: DJ Patil, the current Chief Data Scientist of the United States and previously the Head of Data Products at LinkedIn, is the one who first coined the term 'data science'.

'Why' data science



As you can imagine, the single word answer would be "insight" – we would like to ****extract**** **meaningful info**, "actionable" knowledge, insight, wisdom – whatever it is called – from "cold, hard data". That insight could then **bring in profits, change societies, save lives.**



This isn't hype or a vague concept. It means that we start with **collected/measured data, analyze/process it**, and obtain as a result, something ****new**** that we did not know, did not realize.

Here is a student, talking about why he likes data science:

Why @ramosaj2019 joined #OpenDataLA

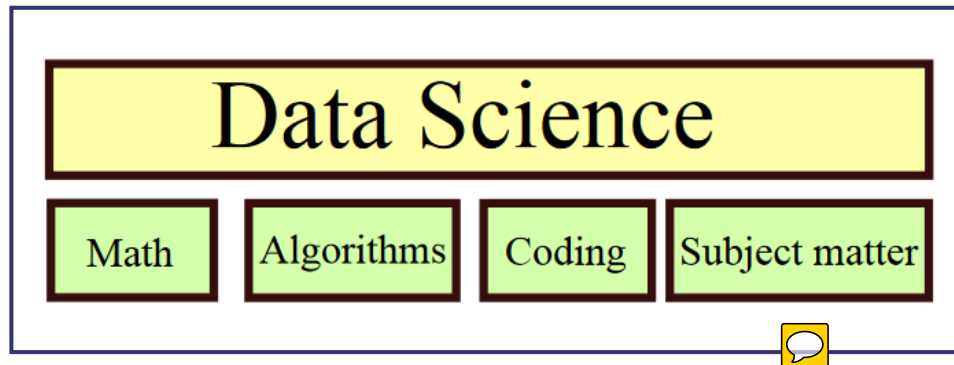


I started in January 2016, right after Chelsea. I'm a first year computer science major at the University of Southern California. I have been interested in Data Analysis and Statistics ever since I visited Washington D.C. and realized that the intersection between my love of computer science and math, and public policy lay in analyzing data released by open data sources. When I got the opportunity to work for one of the best Open Data teams in the country I couldn't say no! I love working with data because of the inferences you can draw from it. With just a few numbers and calculations, you can predict what will happen years in the future. This aids immensely in decision making, and analysis of a problem. As a CS major, I love to solve problems, and Data helps me understand a problem before I solve it.

Prep




As we mentioned earlier, data science is an ****interdisciplinary**** field. It encompasses the following areas:



Prep: areas

math

- linear algebra! 
- calculus
- statistics (eg. look at how **serious** USC is, when it comes to facilitating sound statistical practices that in turn result in good science)
- probability
- discrete math

algorithms

- machine learning algorithms
- other forms of data mining algorithms 

coding

- R 
- Python 
- Java

- JavaScript
- C++
- Scala
- Julia
- MATLAB
- Mathematica
- ..

domain knowledge




You need to KNOW the area/topic/subject/field/domain in which you want to work!

This could be one of a variety of things: agriculture, business, climate, consumer-oriented, ecology, education, finance, government, health/medicine, manufacturing, societal, science.. – each area has MANY sub areas!

Why is subject matter important to know? Because you need to




understand the input data, know the terminology, ask the right questions

(regarding the data you want to analyze), **understand and then**  **communicate the results of your analysis.** Without domain knowledge, you will feel like an outsider, and likely be perceived as one.

Academic prep

There are several courses available, both at 'SC and elsewhere:

- USC: [CS](#) and [INF](#) courses, Marshall School courses, math/EE courses..
- <http://online.stanford.edu/course/machine-learning> 
- <https://www.coursera.org/learn/machine-learning>
- Udacity, eg. [this](#) and [this](#)
- <http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/>
- ...

There are a lot of schools throughout the US and the world, that offer degree programs in data science. Columbia U has a pretty comprehensive program; here is their course list:

Our curriculum is 30 credits total.

Course List:

Probability (3) STAT W4105
Algorithms for Data Science (3) CSOR W4246
Statistical Inference and Modeling (3) STATS W4108
Computer Systems for Data Science (3) COMS W4121
Machine Learning for Data Science (3) COMS W4776
Exploratory Data Analysis and Visualization (3) STATS W4701
Data Science Capstone & Ethics (3) ENG E4800
Electives



The M.S. program may be completed in two semesters of full-time intensive study or on a part-time basis.

Here is another good course.



Concentration areas

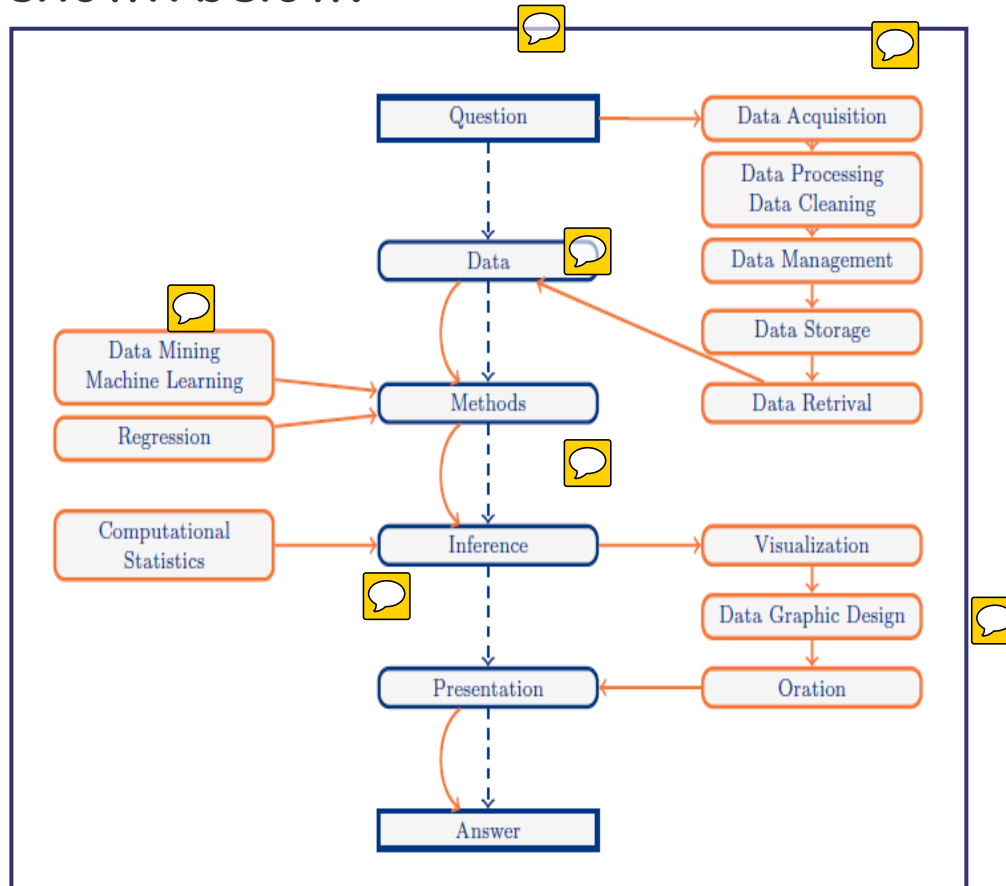
Here are areas you can focus on (courses, tools, coding, domain knowledge):

- spatial analysis
- text mining 
- speech
- social media 
- Big Data
- medical [eg. [Optum](#)]
- bioinformatics [eg. look at [this](#) article]; also, [here](#) is some career advice



The process

Data science based analysis consists of a defined series of steps, as shown below:



Tools



"From scratch" analyses can be performed using homegrown and open-source libraries in R, Python, Java etc.



Weka is a comprehensive data mining tool, as are KNIME and RapidMiner [all free] – all are quite popular. There are a **lot** of tools! Even **Excel**, MATLAB and Mathematica are used for data analysis..

For visualizing, there are many options again – Qlikview, Tableau, periscope.io..

Datasets

'Mountains' of Big Data exist at the federal, state and city levels, free for the taking - to analyze, get insights from, and as a result, transform society:

- <http://www.data.gov>
- <http://data.ca.gov/>
- <https://data.lacity.org/> and <http://geohub.lacity.org/>

Hackathons!

These are a good way to test your skills, make new friends, win some fame/fortune..

Past and upcoming events.

Another: <http://www.datasciencebowl.com/>

Internships

A good way to get into the field (or test the waters!)..

You can consider applying to these highly competitive, intensive and immersive programs:

- <http://dssg.uchicago.edu> : Data Science for Social Good
- <http://www.thedataincubator.com> : The Data Incubator

Opportunities like the above will let you exercise your (Big Data, analytics) knowledge, (coding, communication, teamwork) skills and passion for effecting social change.

Job projections

PLENTY to go around!

<http://www.forbes.com/sites/gilpress/2015/04/30/the-supply-and-demand-of-data-scientists-what-the-surveys-say>

<https://www.wpi.edu/academics/datascience/career-outlook.html>

Job titles

Since Data Science is inter-disciplinary, job titles could take on a variety of forms: 

- Director of Data Science
- Analytics and Data Science Manager
- Head of Data Science and Management
- Data Scientist
- Cybersecurity Data Scientist
- Image Data Scientist
- Data Scientist (Financial Services)
- Data Scientist with Predictive Modeling
- Data Engineer
- Big Data Engineer
- Business Analyst
- Data Advisor
- ...

Job title: 'Data Scientist'

"The emerging big data scientist is distinctly different from other data professionals. For instance, nearly half of big data scientists use R—an opensource language and environment for statistical computing and graphics—despite the fact that it is used by only 13 percent of other practitioners. They are also twice as likely to use big data storage tools such as Hadoop®, Netezza, and AsterData. Big data scientists are also remarkably educated—40 percent have a master's degree, and an additional 17 percent have a doctorate. Over 90 percent have at least a college education."⁴

Note: you don't 'NEED' a PhD to work in this field! Good analytical skills, knowledge of the tools and algorithms, and coding ability – these would help immensely.

Here is a ranking of relevant skills..

Job boards

Kaggle: <https://www.kaggle.com/jobs>

DataJobs: <https://datajobs.com/data-science-jobs>

DataScienceJobs (aggregator): <https://www.datasciencejobs.io/>

Interview questions

Here is a list of 100 basic questions.

More: books

Data Science from Scratch

Doing Data Science

Learning to Love Data Science

Creating a Data-Driven Organization

Thinking with Data

More: blogs, sites..

There is a LOT written about data science. Here are ssample blogs and sites:

- <http://www.becomingadatascientist.com> (including <http://www.becomingadatascientist.com/learningclub/>)
- <http://blog.jaycordes.com/> - excellent posts by my friend Jay Cordes
- <http://research.google.com/pubs/DataMiningandModeling.html>
- <http://www.datasciencecentral.com/>
- <http://www.datasciguide.com/>
- ..

More: meetups, conferences

Strata: <http://conferences.oreilly.com/strata>

KDD: <http://www.kdd.org/>

NIPS: <https://nips.cc/>

LA meetups (lots!): http://www.meetup.com/topics/data-science/us/ca/los_angeles/

