Search Data Science Central   [Search](#)
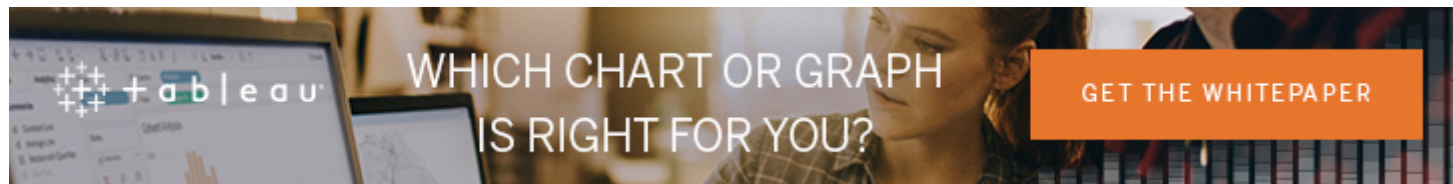
- [Sign Up](#)
- [Sign In](#)

Data Science Central
THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS

- [Analytics](#)
- [Big Data](#)
- [Hadoop](#)
- [Data Plumbing](#)
- [DataViz](#)
- [Jobs](#)
- [Webinars](#)
- [Digest](#)
- [NYU Stern Masters](#)
- [Search](#)
- [Contact](#)

[Subscribe to DSC Newsletter](#)

- [All Blog Posts](#)
- [My Blog](#)
- [Add](#)

The Professionalization of Data Science

- Posted by Michael Walker on October 22, 2013 at 10:13pm
- View Blog

Author: Swami Chandrasekaran

9. Data Munging — Denoising, Feature Extraction, Binning Sparse Values, Unbiased Estimators, Handling Missing Values, Data Scrubbing, Normalization, Dimensionality & Numerosity Reduction

8. Data Ingestion — Transformation & Enrichment, Data Fusion, Data Integration, Data Sources & Acquisition, Data Discovery, Summary of Data Formats

Sampling, Using ETL, Stratified Sampling, How much Data?, Principal Component Analysis, Google OpenRefine, Data Survey

5. Text Mining / NLP — Corpus, Named Entity Recognition, Text Analysis, UIMA, Term Document Matrix, Term Frequency & Weight, Support Vector Machines, Association Rules, Market Based Analysis, Feature Extraction, Using Mahout, Using Weka, Using NLTK

Clustering — Hierarchical Clustering, K-means Clustering, Neural Networks, Sentiment Analysis, Collaborative Filtering, Tagging, Vocabulary Mapping, Classify text

6. Visualization — Data Exploration in R (Hist, Boxplot etc), Uni, Bi & Multivariate Viz, ggplot2, Histogram & Pie (Uni), Tree & Tree Map, Scatter Plot (Bi), Line Charts (Bi), Spatial Charts, Survey Plot, Timeline, Decision Tree, Tableau, IBM ManyEyes, InfoVis, D3.js

Regression — Perceptron, Linear Regression, Ranking, Logistic Regression

K-Nearest Neighbor, Naive Bayes Classifiers, Boosting, Decision Trees, Classification Rate, Trees & Classification, Bias & Variance, Overfitting, Lift, Prediction, Classifier, Training & Test Data, Concepts, Inputs & Attributes, Unsupervised Learning, Supervised Learning, Categorical Var, Numerical Var, What is ML?

4. Machine Learning — MLE, Kernel Density Estimate, Regression, Covariance, Correlation, Pearson Coeff, Causation, Least Fit, Euclidean Distance

Classification

1. Fundamentals — Matrices & Linear Algebra Fundamentals, Hash Functions, Binary Tree, O(n), Relational Algebra, DB Basics, Inner, Outer, Cross, Theta Join, CAP Theorem, Tabular Data, Data Frames & Series, Sharding, OLAP, Multidimensional Data Model, ETL, Reporting Vs BI Vs Analytics, JSON & XML, NoSQL, Regex, Vendor Landscape, Env Setup

Prob Den Fn (PDF), ANOVA, Skewness, Continuos Distributions (Normal, Poisson, Gaussian), Cumul Dist Fn (CDF), Random Variables, Bayes Theorem, Probability Theory, Percentiles & Outliers, Histograms, Exploratory Data Analysis, Central Limit Theorem, Monte Carlo Method, Hypothesis Testing, P-Value, Chi2 Test, Estimation, Confid Int (CI), Descriptive Statistics (mean, median, range, SD, Var), Pick a Dataset (UCI Repo)

10. Toolbox — MS Excel w/ Analysis ToolPak, Java, Python, R, R-Studio, Rattle, Weka, Knime, RapidMiner, Hadoop Dist of Choice, Spark, Storm, Flume, Scibe, Chukwa, Nutch, Talend, Scraperwiki, Webscraper, Flume, Sqoop, tm, RWeka, NLTK, RHIPE, D3.js, ggplot2, Shiny, IBM Languageware, Cassandra, MongoDB

Data Frames, Reading CSV Data, Reading Raw Data, Subsetting Data, Manipulate Data Frames, Functions, Lists, Factors, Arrays, Matrices, Vectors, Variables, Expressions, R Basics, R Setup R Studio

Install Pkgs, Factor Analysis, Rapid Miner, IBM SPSS

7. Big Data — Job & Task Tracker, MR Programming, Sqoop, Loading Data in HDFS, Flume, Scribe, For Unstruct Data, SQL with Pig, DWH with Hive, Scribe, Chukwa For Weblog, Using Mahout, Name & Data Nodes, Setup Hadoop (IBM / Cloudera / HortonWorks), Data Replication Principles, HDFS, Hadoop Components, Map Reduce Fundamentals, Zookeeper Avro, Storm: Hadoop Realtime, Rhadoop, RHIPE, rmr, Cassandra, MongoDB, Neo4j

2. Statistics

3. Programming — Python Basics, Working in Excel

There has been much discussion and debate about the definition of data science and the new rare breed of sexy bird called the data scientist. The [Data Science Association](#) defines "[Data Science](#)" as the scientific study of the creation, validation and transformation of data to create meaning; and the "[Data Scientist](#)" as a professional who uses scientific methods to liberate and create meaning from raw data.

While these definitions may appear overbroad, think about the definitions of a lawyer or physician. A lawyer is a legal professional who can help prevent or

solve legal issues and a physician is a health professional who can help prevent or cure health issues. Like the professionalization of law and medicine in the past hundred years, data science is at the very beginning of becoming a profession - with competency standards and a [Data Science Code of Professional Conduct](#).

This means that data science will evolve into a profession where data scientists specialize in different areas - like lawyers and physicians. When you need to hire a lawyer you usually consider the special area of law that a lawyer practices. If you have a tax problem you hire a tax lawyer, not a divorce lawyer. If you have a heart problem you do not hire a gynecologist.

The simple truth is that data science is a vast and complicated field and - like law and medicine - much too big and complex for a person to master in one lifetime. My colleague Gary Mazzaferro has been exploring the concepts and ideas surrounding data science and definitions as formalizations aligning with knowledge economies and the knowledge / science / technology maturity models. Gary has (to date) defined the following data science specializations and types of data scientists:

Data Science: A field of systematic interdisciplinary study to elucidate relationships across and within Formal, Social Natural and Special Sciences phenomenon through the application of scientific methods. Interdisciplinary areas include analytical processes, mathematics, probability and statistics, logic, modeling, machine learning, algorithms, communications, traditional sciences, business, public policy and philosophy.

Blue Sky Data Science: A purely curiosity driven exploratory branch of Data Science oriented towards the development and establish understanding about relationships across and within phenomenon with no focus on specific goals and immediate application.

Basic Data Science: A branch of Data Science research focused on clearly defined goals and oriented towards the development and establish understanding about relationships across and within phenomenon.

Applied Data Science: A branch of Data Science oriented toward the development of practical applications, technologies other interventions including engineering practices. Applied Data Science bridges the gap between Basic Data Science and the engineering domains to provide predicable, usable tools to industries including standard methods and practices.

Data Science Practice: The regular performance of Applied Data Science activities and methods for private and public organizations. May practice externally or internally. Practice may necessitate additional disciplines based on the needs of the organization including domain expertise and communications supporting presentation and reporting activities.

Data Scientist: A person that studies or has expert knowledge of the interdisciplinary field of Data Science.

Blue Sky Data Scientist: A person that studies or researches in the branch of Blue Sky Data Science.

Basic Data Scientist: A person that studies, researches or has expert knowledge in the branch of Basic Data Science.

Applied Data Scientist: A person that studies or researches in the branch of Applied Science.

Note that this is a preliminary list and is not complete. The profession of data science will evolve to create many specializations. After all, it took law and medicine over one hundred years to evolve as professions with different specialties.

See: http://bit.ly/10TgVHG

Views: 13255

Like
11 members like this

Share   Tweet   G+1

Like 45

- < Previous Post
- Next Post >

Comment

You need to be a member of Data Science Central to add comments!

Join Data Science Central

Comment by thien anh nguyen bui (At Home) on October 25, 2015 at 9:16pm

Thanks for your good post.

Comment by Dr. Z on October 25, 2013 at 5:58am

Interesting classification.

RSS
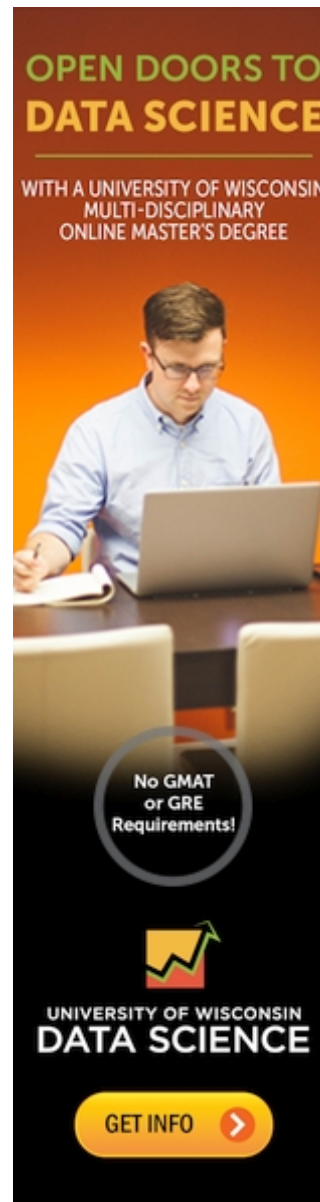
Welcome to
Data Science Central

Sign Up

or Sign In

Or sign in with:

-
-
-

Follow Us

Top Content

1

40 Techniques Used by Data Scientists

2

Feature Engineering: Data scientist's Secret Sauce !

3

Brexit: Why Projections Were Wrong

4 **[Statistics is Dead – Long Live Data Science…](#)**

5

[Free Alternatives to Excel for Data Cleaning](#)

6

[12 Algorithms Every Data Scientist Should Know](#)

7

[IoT 101 – Lesson 3: Everything You Need to Know to Start Your IoT Project](#)

- [RSS](#)
- [View All](#)

Announcements

[How Open Data Science Can Help Analytics Leaders Survive and Thrive](#)

[Open Data Platform Demystified: Hadoop and Spark](#)

[UW MS in Data Science—Master All Areas of Big Data](#)

[Critical Analytics for Marketers](#)

[Ready for a Fresh Perspective?](#)

[7 Steps to Prepare Data for Predictive Analytics](#)

[Don't Miss the New SQL on Hadoop Benchmark Results](#)

July 19 Webinar: Internal Search Intelligence and Optimization

Three Ways To Master Data Analytics

Time Data: Beyond the Line Chart

Videos

4 Steps to improve your Search Technology and Boost Sales, BI and User Experience

Added by Tim Matteson 0 Comments 1 Like

New Methods for Data Preparation, Advanced Analytics, and Visual Analytics

Added by Tim Matteson 0 Comments 0 Likes

- Add Videos
- View All

Resources

- Introduction to Programming in Stata
- collaborative recommendation engine implementation in python
- Benchmarking 20 Machine Learning Models Accuracy and Speed
- Stata Cheat Sheet
- Selection of best articles from our past weekly digests
- Statistical Analysis Advisor Chart

- [Selection of best articles from our past weekly digests](#)
- [Free Online Book: Forecasting, Principles and Practice](#)
- [38 Seminal Articles Every Data Scientist Should Read](#)
- [Black-box Confidence Intervals: Excel and Perl Implementation](#)

HEADLINES BY
FEEDBURNER

Top Categories

[Machine Learning](#)

[R Programming](#)

[Python for Data Science](#)

[Visualization, Dashboards](#)

[NoSQL and NewSQL](#)

[Big Data](#)

[Cheat Sheets](#)

[Internet of Things](#)

[Excel](#)

© 2016   Data Science Central   Powered by NING | MODE

[Badges](#) | [Report an Issue](#) | [Privacy Policy](#) | [Terms of Service](#)