Developing Analytical Talent Checklist

Data Science Tools: Java, MapReduce, R, Hadoop, Python, NoSQL.

Data Engineers: Extract, load, transform.

Data Scientists: Discover, access, distill.

AaaS: Analytics as a service.

Granger Causality Test: Cross correlation.

Natural Language Processing: Analyzing unstructured data.

Stemming: Plural replaced by singular.

Stopwords: Words that connect such as the, or, and, of from, about, it.

Normalizing Text: Tokens rearranged in alphabetical order and sorted.

Lookup Tables: Used to determine whether a text is part of a category or is an alias of another text.

Synonym Tables: To determine if words are the same or similar enough as other words and should be merged.

User Intention: Algorithm to determine what the user is intending to do with the search.

Taxonomy: The output of a clustering algorithm.

Attaching: Algorithms to determine whether two tokens such as San and Francisco should be attached or separated.

Trend of Data Science: Analytics engineering, data shaping, art of optimization, optimization science, data intelligence, business optimization.

Tarjan’s Strongly Connected Components Algorithm:

Bumpiness: Bumpiness is linked to how the data points are ordered, while centrality and volatility completely ignore order. So, bumpiness is useful for datasets where order matters, in particular time series. Also, bumpiness integrates the notion of dependence (among the data points), while centrality and variance do not.

Regression: Predictive statistical tool using independent variables to predict a dependent variable. Avoid linear regression. Best tools are to reduce number of variables and use constrained regression such as ridge, logic, or lasso regression.

Decision Trees: Tree like graph model to determine decisions and their possible outcomes, avoid large trees and use small decision trees.

Linear Discriminant Analysis: Supervised clustering. Bad technique because it assumes that clusters do not overlap and are well separated by hyperplanes. Use density estimation techniques instead.

K Means Clustering: Produces circular clusters and not good for data points that are not a mixture of Gaussian distributions.

Neural Networks: Examines inputs and hidden inputs to map their connectivity towards an output.

Density Estimation in high dimensions: Subject to dimensionality. Instead use nonparametric kernel density estimators with adaptive bandwidths.

Data Science: Computer science, statistical science, business management, software engineering, domain expertise, statistical science.

Three types of analytics: Descriptive, predictive, and prescriptive.

Descriptive Analytics: What happened and why.

Predictive Analytics: What will happen.

Prescriptive: Suggesting actions or decisions based on the descriptive and predictive analytics.

Data Science Competitions: Kaggle, CrowdAnalytix, Data Science Central.

Schools: UC Berkeley, CUNY

Training Programs: Informs, digital analytics association, TDWI, Americal Statistical Association, Data Science Central, International Institute for Analytics, Statistics.com

Conferences: Predictive Analytics World, GoPivotal Data Science, SAS, Association for Computing Machinery, IEEE, IE Group, Text Analytics News, IQPC, Whitehall Media.

Free Training Programs: Coursera, Data Science Central,

Posting Sites: Analyticbridge.com, BigDataNews.com, Quora.com

Metrics for Marketing: Open rate, # opens, click rate, trends, churn, unsubscribe rate, spam complaints, geography, language

Extreme Events: Using Monte Carlo to predict fraud by detecting extreme events unlikely to naturally occur.

Blacklists and Whitelists: Used to determine spam and unwanted text.

Analyzing Domain Patterns:

Gibberish: Looking at websites for random text gibberish.

Abnormal Purchase Patterns: Would a ninety year old lady buy a motorcycle at 2 am in the morning.

Repetitive small purchases: Lots of $9,99 purchases.

Data Videos: vcla.stat.ucla.edu/publications.html

Data Dictionary: Table with three or four columns, label, value, frequency, dimension of the label (1 if one variable and 2 if pair of variables).

Hidden Decision Trees: Scores large volumes of transactional data.

Experimental design: A/B testing, multivariate testing, taguchi methods.

Optimizing Web Crawlers: Split crawling across 8,000 servers, run 20 copies on each server, change timeout thresholds, max size, blacklist of websites to ignore, avoid infinite loop with site lookup table.

Stock Price Data: [www.stockhistoricaldata.com/daily-download](http://www.stockhistoricaldata.com/daily-download)

Lorenz Curve:

Black-Scholes:

Arcsine Law:

Domain Experience: Critical to succeed in data analytics. Need advisors in different industries.

BIRT:

PCA:

Cross-Correlation:

Correlograms:

Spectral Analysis:

Signal Processing:

Fuzzy Merging:

Toad/Brio:

HashTable:

Naïve Bayes:

Schema:

Six Sigma:

Boosted Trees:

Gaussian Distribution:

Cron Job:

Actuarial Sciences:

Sensitivity Analysis:

Books: [www.analyticbridge.com/group/books](http://www.analyticbridge.com/group/books)