**Key Algorithms and Statistical Models for Aspiring Data Scientists**

This article provides a summary of key algorithms and statistical techniques commonly used in industry, along with a short resource related to these techniques.

**By** [**Colleen M. Farrelly**](https://www.kdnuggets.com/author/colleen-farrelly)**, Kaplan University.**

As a data scientist who has been in the profession for several years now, I am often approached for career advice or guidance in course selection related to machine learning by students and career switchers on LinkedIn and Quora. Some questions revolve around educational paths and program selection, but many questions focus on what sort of algorithms or models are common in data science today.  
With a glut of algorithms from which to choose, it’s hard to know where to start. Courses may include algorithms that aren’t typically used in industry today, and courses may exclude very useful methods that aren’t trending at the moment. Software-based programs may exclude important statistical concepts, and mathematically-based programs may skip over some of the key topics in algorithm design.

I’ve put together a short guide for aspiring data scientists, particularly focused on statistical models and machine learning models (supervised and unsupervised); many of these topics are covered in textbooks, graduate-level statistics courses, data science bootcamps, and other training resources (some of which are included in the reference section of the article). Because machine learning is a branch of statistics, machine learning algorithms technically fall under statistical knowledge, as well as data mining and more computer-science-based methods. However, because some algorithms overlap with computer science course material and because many people separate out traditional statistical methods from new methods, I will separate the two branches in the list.

Statistical methods include some of the more common methods overviewed in bootcamps and certificate programs, as well as some of the less common methods that are typically taught in graduate statistics programs (but can be of great advantage in practice). All of the suggested tools are ones I use on a regular basis:

1) Generalized linear models, which form the basis of most supervised machine learning methods (including logistic regression and Tweedie regression, which generalizes to most count or continuous outcomes encountered in industry…)

2) Time series methods (ARIMA, SSA, machine-learning-based approaches)

3) Structural equation modeling (to model and test mediated pathways)

4) Factor analysis (exploratory and confirmatory for survey design and validation)

5) Power analysis/trial design (particularly simulation-based trial design to avoid overpowering analyses)

6) Nonparametric testing (deriving tests from scratch, particularly through simulations)/MCMC

7) K-means clustering

8) Bayesian methods (Naïve Bayes, Bayesian model averaging, Bayesian adaptive trials...)

9) Penalized regression models (elastic net, LASSO, LARS...) and adding penalties to models in general (SVM, XGBoost...), which are useful for datasets in which predictors outnumber observations (common in genomics and social science research)

10) Spline-based models (MARS...) for flexible modeling of processes

11) Markov chains and stochastic processes (alternative approach to time series modeling and forecast modeling)

12) Missing data imputation schemes and their assumptions (missForest, MICE...)

13) Survival analysis (very helpful in modeling churn and attrition processes)

14) Mixture modeling

15) Statistical inference and group testing (A/B testing and more complicated designs implemented in a lot of marketing campaigns)

Machine learning extends many of these frameworks, particularly k-means clustering and generalized linear modeling. Some useful techniques common across many industries (and some more obscure algorithms that are surprisingly useful but rarely taught in bootcamps or certificate programs) include:

1) Regression/classification trees (early extension of generalized linear models with high accuracy, good interpretability, and low computational expense)

2) Dimensionality reduction (PCA and manifold learning approaches like MDS and tSNE)

3) Classical feedforward neural networks

4) Bagging ensembles (which form the basis of algorithms like random forest and KNN regression ensembles)

7) Boosting ensembles (which form the basis of gradient boosting and XGBoost algorithms)

8) Optimization algorithms for parameter tuning or design projects (genetic algorithms, quantum-inspired evolutionary algorithms, simulated annealing, particle-swarm optimization)

9) Topological data analysis tools, which are particularly well-suited for unsupervised learning on small sample sizes (persistent homology, Morse-Smale clustering, Mapper...)

10) Deep learning architectures (deep architectures in general)

11) KNN approaches for local modeling (regression, classification)

12) Gradient-based optimization methods

13) Network metrics and algorithms (centrality measures, betweenness, diversity, entropy, Laplacians, epidemic spread, spectral clustering)

14) Convolution and pooling layers in deep architectures (particularly useful in computer vision and image classification models)

15) Hierarchical clustering (which is related to both k-means clustering and topological data analysis tools)

16) Bayesian networks (pathway mining)

17) Complexity and dynamic systems (related to differential equations but often used to model systems without a known driver)

Depending on one’s chosen industry, additional algorithms related to natural language processing (NLP) or computer vision may be needed. However, these are specialized areas of data science and machine learning, and folks entering those fields are usually specialists in that particular area already.

Some resources for learning these methods outside of an academic program include:

Christopher, M. B. (2016). Pattern Recognition and Machine Learning. Springer-Verlag New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 337-387). New York: Springer series in statistics.

<https://www.coursera.org/learn/machine-learning>

<http://professional.mit.edu/programs/short-programs/machine-learning-big-data>

<https://www.slideshare.net/ColleenFarrelly/machine-learning-by-analogy-59094152>