The 10 articles and books below all were published in the last 50 years and are listed in chronological order.

1. Hirotugu Akaike (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of the Second International Symposium on Information Theory.

This is the paper that introduced the term AIC (originally called An Information Criterion but now known as Akaike Information Criterion), for evaluating a model's fit based on its estimated predictive accuracy. AIC was instantly recognized as a useful tool, and this paper was one of several published in the mid-1970s placing statistical inference within a predictive framework. We now recognize predictive validation as a fundamental principle in statistics and machine learning. Akaike was an applied statistician, who in the 1960s, tried to measure the roughness of airport runways, in the same way that Benoit Mandelbrot's early papers on taxonomy and Pareto distributions led to his later work on the mathematics of fractals.

2. John Tukey (1977). Exploratory Data Analysis.

This book has been hugely influential and is a fun read that can be digested in one sitting. Traditionally, data visualization and exploration were considered low-grade aspects of practical statistics; the glamour was in fitting models, proving theorems, and developing the theoretical properties of statistical procedures under various mathematical assumptions or constraints. Tukey flipped this notion on its head. He wrote about statistical tools not for confirming what we already knew (or thought we knew), and not for rejecting hypotheses that we never, or should never have, believed, but for discovering new and unexpected insights from data. His work motivated advances in network analysis, software, and theoretical perspectives that integrate confirmation, criticism, and discovery.

3. Grace Wahba (1978). Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. Journal of the Royal Statistical Society.

Spline smoothing is an approach for fitting nonparametric

curves. Another of Wahba's papers from this period is called "An automatic French curve," referring to a class of algorithms that can fit arbitrary smooth curves through data without overfitting to noise, or outliers. The idea may seem obvious now, but it was a major step forward in an era when the starting points for curve fitting were polynomials, exponentials, and other fixed forms. In addition to the direct applicability of splines, this paper was important theoretically. It served as a foundation for later work in nonparametric Bayesian inference by unifying ideas of regularization of high-dimensional models.

4. Bradley Efron (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics.

Bootstrapping is a method for performing statistical inference without assumptions. The data pull themselves up by their bootstraps, as it were. But you can't make inference without assumptions; what made the bootstrap so useful and influential is that the assumptions came implicitly with the computational procedure: the audaciously simple idea of resampling the data. Each time you repeat the statistical

procedure performed on the original data. As with many statistical methods of the past 50 years, this one became widely useful because of an explosion in computing power that allowed simulations to replace mathematical analysis.

5. Alan Gelfand and Adrian Smith (1990). Sampling-based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association.

Another way that fast computing has revolutionized statistics and machine learning is through open-ended Bayesian models. Traditional statistical models are static: fit distribution A to data of type B. But modern statistical modeling has a more Tinkertoy quality that lets you flexibly solve problems as they arise by calling on libraries of distributions and transformations. We just need computational tools to fit these snapped-together models. In their influential paper, Gelfand and Smith did not develop any new tools; they demonstrated how Gibbs sampling could be used to fit a large class of statistical models. In recent decades, the Gibbs sampler has been replaced by Hamiltonian Monte Carlo, particle filtering, variational Bayes, and more elaborate

algorithms, but the general principle of modular model-building has remained.

6. Guido Imbens and Joshua Angrist (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica.

Causal inference is central to any problem in which the question isn't just a description (How have things been?) or prediction (What will happen next?), but a counterfactual (If we do X, what would happen to Y?). Causal methods have evolved with the rest of statistics and machine learning through exploration, modeling, and computation. But causal reasoning has the added challenge of asking about data that are impossible to measure (you can't both do X and not-X to the same person). As a result, a key idea in this field is identifying what questions can be reliably answered from a given experiment. Imbens and Angrist are economists who wrote an influential paper on what can be estimated when causal effects vary, and their ideas form the basis for much of the later work on this topic.

7. Robert Tibshirani (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society.

In regression, or predicting an outcome variable from a set of inputs or features, the challenge lies in including lots of inputs along with their interactions; the resulting estimation problem becomes statistically unstable because of the many different ways of combining these inputs to get reasonable predictions. Classical least squares or maximum likelihood estimates will be noisy and might not perform well on future data, and so various methods have been developed to constrain or "regularize" the fit to gain stability. In this paper, Tibshirani introduced lasso, a computationally efficient and now widely used approach to regularization, which has become a template for data-based regularization in more complicated models.

8. Leland Wilkinson (1999). The Grammar of Graphics.

In this book, Wilkinson, a statistician who's worked on several influential commercial software projects including SPSS and Tableau, lays out a framework for statistical graphics that

goes beyond the usual focus on pie charts versus histograms, how to draw a scatterplot, and data ink and chartjunk, to abstractly explore how data and visualizations relate. This work has influenced statistics through many pathways, most notably through ggplot2 and the tidyverse family of packages in the computing language R. It's an important step toward integrating exploratory data and model analysis into data science workflow.

9. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). Generative Adversarial Networks. Proceedings of the International Conference on Neural Information Processing Systems.

One of machine learning's stunning achievements in recent years is in real-time decision making through prediction and inference feedbacks. Famous examples include self-driving cars and DeepMind's AlphaGo, which trained itself to become the best Go player on Earth. Generative adversarial networks, or GANs, are a conceptual advance that allow reinforcement learning problems to be solved automatically. They mark a

step toward the longstanding goal of artificial general intelligence while also harnessing the power of parallel processing so that a program can train itself by playing millions of games against itself. At a conceptual level, GANs link prediction with generative models.

10. Yoshua Bengio, Yann LeCun, and Geoffrey Hinton (2015). Deep Learning. Nature.

Deep learning is a class of artificial neural network models that can be used to make flexible nonlinear predictions using a large number of features. Its building blocks—logistic regression, multilevel structure, and Bayesian inference—are hardly new. What makes this line of research so influential is the recognition that these models can be tuned to solve a variety of prediction problems, from consumer behavior to image analysis. As with other developments in statistics and machine learning, the tuning process was made possible only with the advent of fast parallel computing and statistical algorithms to harness this power to fit large models in real time. Conceptually, we're still catching up with the power of

these methods, which is why there's so much interest in interpretable machine learning.