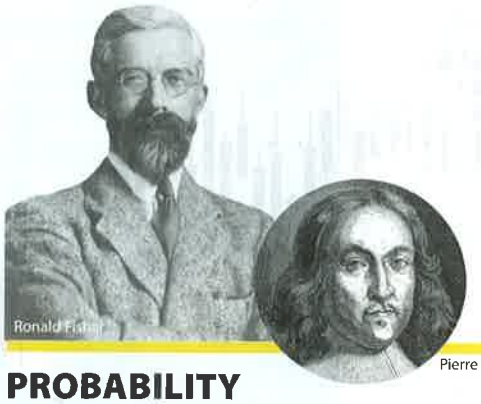
ARTIFICIAL INTELLIGENCE, STATISTICS, and STATISTICIANS

Penny S. Reynolds, ASA History of Statistics Special Interest Group Member

Artificial intelligence is nearly always associated with computer science and engineering. Although obviously dependent on massive computational resources, AI has nevertheless required substantial statistical input along its entire developmental path. Here is an overview of major statistical concepts and methods integral to AI and the statisticians involved in their development.



Pierre de Fermat

PROBABILITY

Probability is a fundamental concept in statistics. Modern AI is based on probability theory for quantifying uncertainty and making data-based forecasts. Development of the underlying mathematics during the 17th and 18th centuries was mostly motivated by the study of gambling. The cornerstone of modern probability theory was developed over the course of a year in letters between **Blaise Pascal** (1623–1662) and **Pierre de Fermat** (1601–1665).

Other foundational contributions included those by **Christiaan Huygens** (1629–1695) and **Abraham de Moivre** (1667–1754). **Pierre-Simon and Marquis de Laplace** (1749–1827) made perhaps the most important contributions to the mathematical theory of inference, particularly what today is recognized as a Bayesian interpretation of probability.

More than a century later, in 1922, **Ronald Fisher** (1890–1962) presented his unifying theory of statistics and inference. In this ground-breaking paper, he introduced several concepts central to statistics and AI such as consistency, efficiency, sufficiency, validity, likelihood, and—in particular—the statistical concept of information.

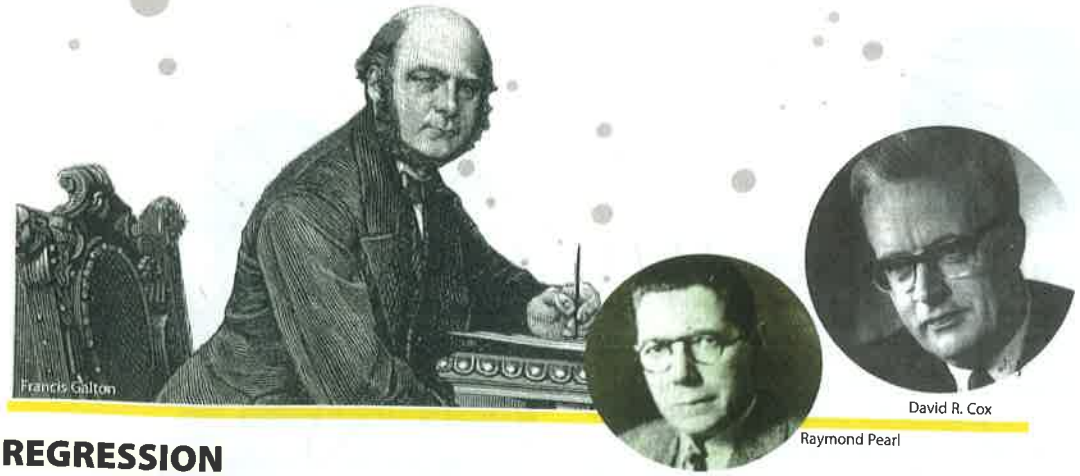
BAYESIAN

BAYESIAN STATISTICS

The Bayesian interpretation of probability is a description of the conditional probability of an event based on both observed data and prior information. Because they enable probabilities to be updated as new information becomes available, Bayesian statistics are instrumental in developing efficient machine learning algorithms and predictive models, as well as for decision-making under uncertainty. **Thomas Bayes** (1701–1761) is best known for his eponymous theorem, although he never published it. His notes describing a solution to one problem of inverse probability were published posthumously by **Richard Price** (1723–1791) in 1761.

Harold Jeffreys (1891–1989) and his 1939 book, *Theory of Probability,* was a major influence on the revival of Bayesian probability. During World War II, **Alan Turing** (1912–1954) developed Bayesian statistical methods as part of his Bletchley Park work in cryptanalysis. Although he does not explicitly use the term "Bayes," he describes adjustment of the initial odds of a hypothesis by a prior and defines the expected Bayes factors against a true hypothesis.

Fisher redefined inverse probability as specifically "Bayesian" in 1950. In 1985, **Judea Pearl** (1936–) coined the term "Bayesian network" to describe models of conditional dependencies between sets of variables.



REGRESSION

Regression methods are the backbone of machine learning and AI algorithms and remain the most common statistical applications for describing and predicting relationships between multiple variables. Feedforward neural networks and deep-learning algorithms are based on regression methods. **Francis Galton** (1822–1911) introduced linear regression in 1885 for quantifying relationships between variables. It was based on the method of least squares developed in the early 19th century by **Adrien-Marie Legendre** (1752–1833) and **Carl Friedrich Gauss** (1777–1855). In 1922, Fisher introduced the modern regression model. This effectively synthesized the concept of least-squares theory with the concepts of regression and correlation proposed by Galton, later expanded upon by **Karl Pearson** (1857–1936) and **George Udny Yule** (1871–1951).

Logistic regression has been described as the "go-to" method for machine learning involving classification of binary variables. The logistic function was initially developed between 1838 and 1847 in three papers by **Pierre François Verhulst** (1804–1849), a student of **Adolphe Quetelet** (1796–1874). Quetelet is probably best known for his concept of the "average man," the development of the body mass index, and his role as statistical mentor to Florence Nightingale.

The logistic function was repeatedly rediscovered, first by **Raymond Pearl** (1879–1940; ASA president, 1939) and **Lowell Reed** (ASA president, 1951) in 1920, then by Yule in 1925, who revived the name "logistic."

Joseph Berkson (1899–1982) was the primary developer of the modern logistic regression (he also coined the term "logit"). **David R. Cox** (1924–2022) further extended the logistic regression to models of observational data and developed its multinomial generalization.

Nonlinear regression is widely used in AI. It is an important tool for modeling data that is nonlinear in the parameters and thus poorly represented by linear models. Parameter estimates usually have no closed solution but are approximated by computationally intensive numerical optimization algorithms. The earliest of these was the Gauss-Newton algorithm, described by **Carl Friedrich Gauss** in 1809 as an extension of Isaac Newton's methods for determining the minimum of a nonlinear function.

Pafnuty Chebyshev (Tchebychev, Чебышёв 1821–1894) developed methods for polynomial series expansions for curve fitting; later, these were applied to descriptions of nonlinear dynamic systems.

The method of gradient descent was proposed in 1847 by **Augustin-Louis Cauchy** (1789–1857). Backpropagation with gradient descent is useful in neural network applications to improve prediction accuracy and error minimization.

A more robust method than either Gauss-Newton or gradient descent is the Levenberg-Marquardt algorithm developed by statistician **Kenneth Levenberg** (1919–1973) in 1944 and rediscovered by **Donald Marquardt** (1929–1997; ASA president, 1986) in 1963.

SEQUENTIAL ANALYSIS

Sequential analysis involves the process of data evaluation and decision-making in real time, with updating as more information is acquired. In essence, it involves the process of statistical estimation with sequential multiple hypothesis tests. It may have had its roots in the Gambler's Ruin problem formulated by **Christiaan Huygens, Blaise Pascal,** and de Fermat. The method has been attributed primarily to **Abraham Wald** (1902–1950), developed when he was working with the Columbia Statistical Research Group during World War II. Working independently, **Alan Turing** (1912–1954) developed similar sequential conditional probability analysis methods (Banburismus and Turingery) for decoding the German Enigma and Lorenz ciphers. This work remained classified until the early 1980s, so is not as well known.

SPLINE SMOOTHING

Spline fits are a relatively recent body of methods involving the fit of regression models to 'smooth' out noisy data and enable pattern recognition. **Isaac Jacob Schoenberg** (1903–1990) introduced the theory of splines in the 1940s.

The pioneering and enormously influential work of **Grace Wahba** (1934–) on smoothing spline functions, reproducing kernel Hilbert space theory, high-dimensional optimization, and generalized cross-validation has found wide application in the development of statistical machine learning, bioinformatics, medical imaging, computer graphics, and computer animation.

BOOTSTRAPPING

Bootstrapping was developed by **Bradley Efron** (1938– ; ASA president, 2004) in 1979 as a more versatile alternative to the nonparametric jackknife resampling method of **Maurice Quenouille** (1924–1973). Bootstrap resampling is a computer-intensive method for approximating the sampling distribution of almost any estimator. It has been called pioneering and hugely influential for its extraordinary versatility and applicability to a variety of disciplines, including AI.

Bootstrapping is also one of several techniques used for machine learning model validation. Models can be used to infer results for the population by 'training' on the bootstrapped data and then testing model predictions on external data sets.

PREDICTIVE MODELING, FEEDBACK, AND VALIDATION

Predictive analytic models based on machine learning or deep learning are used for classification, clustering, forecasting, and anomaly detection. An early nonparametric supervised machine learning method is the k-nearest neighbors classification algorithm (k-NN) developed by **Evelyn Fix** (1904–1965) and **Joseph Hodges** (1922–2000) in 1951. The k-means clustering algorithm is a supervised learning method developed by **J.B. MacQueen** (1929–2014) in 1967.

For a more detailed exposition of the role of statistics and statisticians in the development of artificial intelligence, see "Is There a Role for Statistics in Artificial Intelligence?" by Sarah Friedrich, Gerd Antes, Sigrid Behr, and coauthors in *Advances in Data Analysis and Classification*.

Ada King, Countess of Lovelace

Jerzy Neyman

CAUSAL INFERENCE

The biggest challenge for machine learning and AI algorithms is distinguishing causation from correlation. **Jerzy Neyman** (1894–1981) has been credited for providing the earliest formal notation defining causal effects in 1923. Earlier, in 1921, **Sewall Wright** (1889–1988) developed path analysis to describe patterns of directed dependencies among a set of variables. Although these models could not explicitly determine causality *per se,* path analysis was an important precursor to structural equation modeling.

Pearl considers path analysis to be directly ancestral to causal inference methods. Pearl, himself, has been called "one of the giants in the field of artificial intelligence."

John Tukey

COMPUTERS AND COMPUTATION

The idea that mechanical devices can generate new knowledge—generative AI—is surprisingly old. In the 1726 satirical novel *Gulliver's Travels,* Jonathan Swift describes a "wonderful machine" that would automatically generate books on all the arts and sciences "without the least assistance from genius or study."

Ada King, Countess of Lovelace (1815–1852), often credited as the earliest computer programmer, speculated on what she called "a calculus of the nervous system." This was the application of mathematical models to understanding thought and emotion—the first glimmers of the idea of a neural network. However, she apparently discounted the idea of artificial intelligence as such, concluding machines could not develop the capacity to "originate anything" or have the "power of anticipating any analytical relations or truths."

This was disputed in the epochal 1950 article "Computing Machinery and Intelligence" by Turing. Widely regarded as the father of artificial intelligence, Turing explicitly posed the question, "Can machines think?" He proposed the Turing test as the definition of the standard for an "intelligent" machine. He developed key ideas related to modern computing and programming (the hypothetical Universal Turing Machine) almost a decade before the technology was sufficiently advanced to put them into practice.

Claude Shannon (1916–2001), developer of information theory and founder of the digital revolution, independently developed many ideas like Turing's. He performed some of the earliest experiments in artificial intelligence, including the development of a maze-solving mechanical mouse and a chess-playing computer program.

Artificial intelligence applications rely heavily on interactive data visualization tools. **John Tukey** (1915–2000), the Father of Data Science, pioneered numerous statistical methods for computer application. In 1972, he devised PRIM-9, the first interactive dynamic computer graphics program for the exploration and identification of patterns in multivariate data. PRIM-9 has been described as revolutionary for its emphasis on the computer-human interface at a time when statistics was widely taken to be synonymous with inference and hypotheses testing. ■