Einführung in Web- und Data-Science

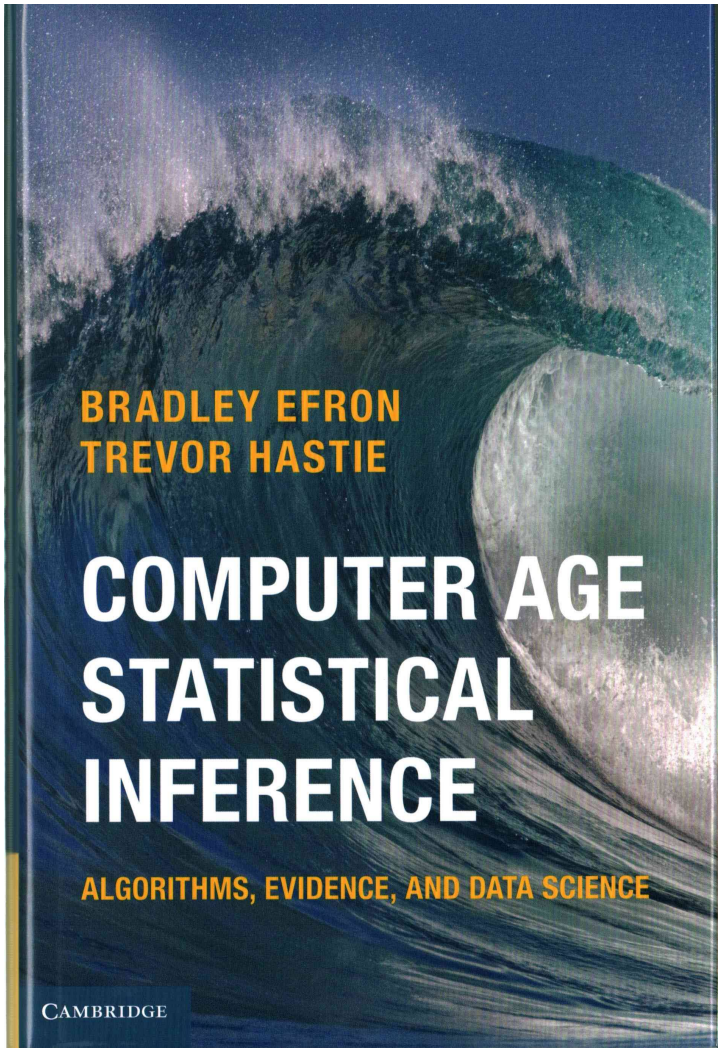
Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun (Übungen)

Statistics and Data Science [CASI 2017, p. 446 ff.]



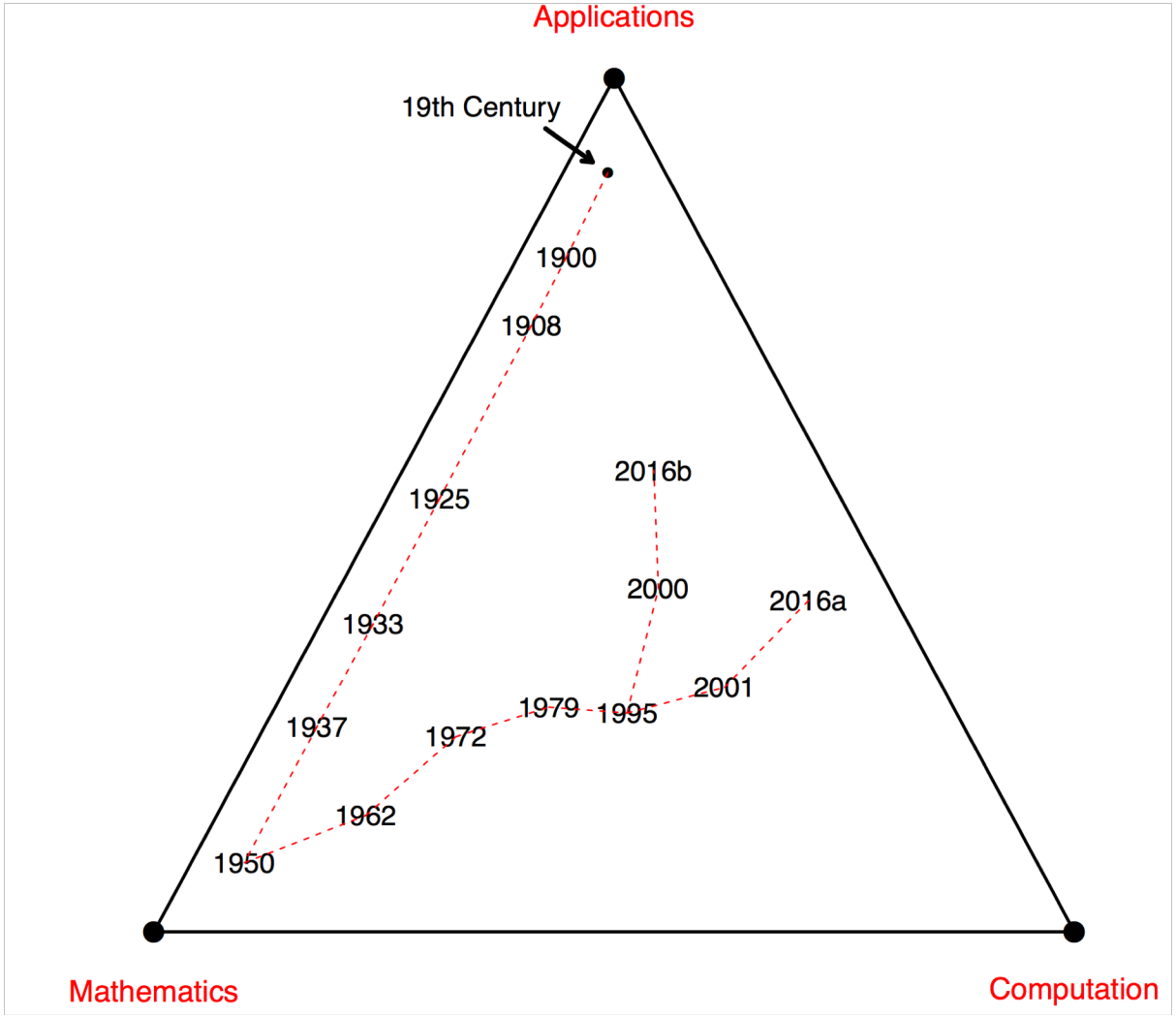
Statistical Inference:

- Deals with the why
- Mathematical foundations

Algorithmics & Data Science as CASI sees it:

- Deals with the how
- Just pragmatism?
  (→ side blow at computer science)
- Decision problems clearly identified in computer science w.r.t. semantics of representation formalisms
- Correctness of algorithms (the why) is very well an issue in computer science (and data science as subfield)
- Tractability issues added by CS

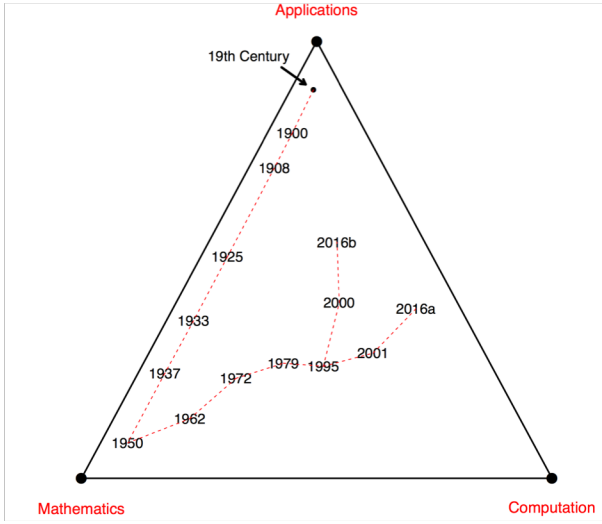
From Statistical Inference to Data Science…



From Statistical Inference to Data Science

1900

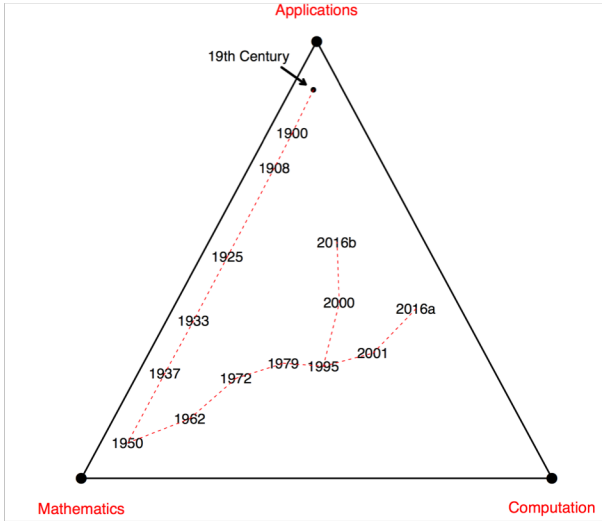
- Karl Pearson's chi-square paper
- Applied a new mathematical tool, matrix theory, in the service of statistical methodology.
- Pearson and Weldon went on to found *Biometrika* in 1901, the first recognizably modern statistics journal.
- Pearson's paper, and *Biometrika*, launched the statistics discipline on a fifty-year march toward the mathematics pole of the triangle



From Statistical Inference to Data Science

1908

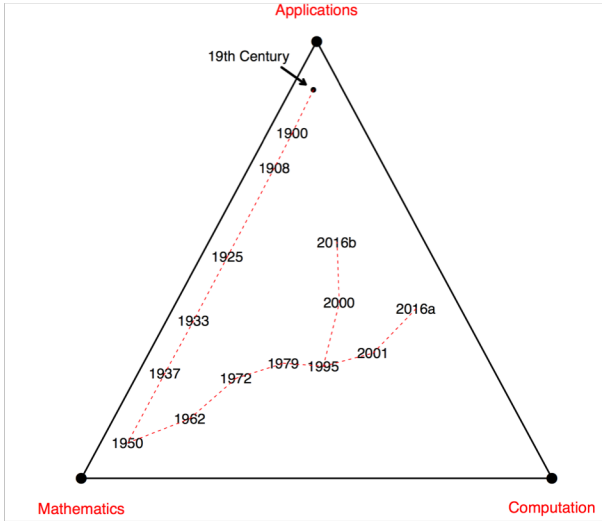
- Student's t statistic
- Crucial first result in small-sample "exact" inference
- Major influence on statistical thinking



From Statistical Inference to Data Science

1925

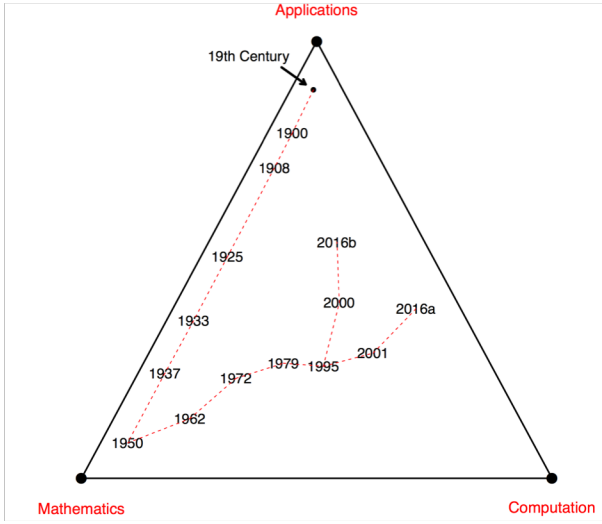
- Fisher's estimation paper
  - Fundamental ideas: sufficiency, efficiency, Fisher information, maximum likelihood theory, and the notion of optimal estimation
- Optimality is a mark of maturity in mathematics, …
- … making 1925 the year statistical inference went from a collection of ingenious techniques to a coherent discipline



From Statistical Inference to Data Science

1933

- Neyman and Pearson's paper on optimal hypothesis testing.
  - Logical completion of Fisher's program, it nevertheless aroused his strong antipathy (concern that mathematization was squeezing intuitive correctness out of statistical thinking)



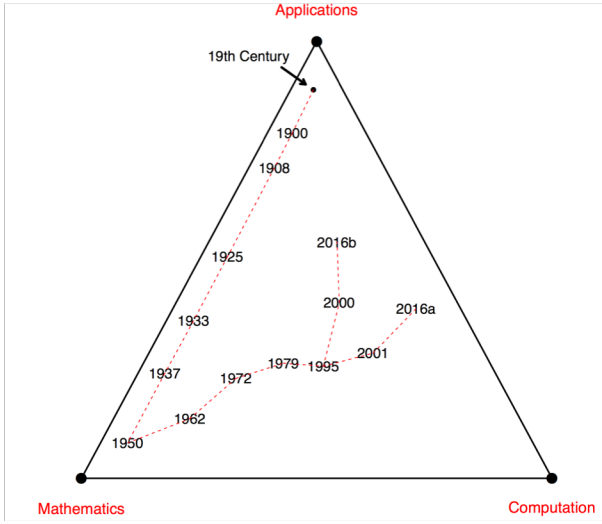
From Statistical Inference to Data Science

1937

- Neyman's seminal paper on confidence intervals
- Mathematical treatment of statistical inference was a predecessor of decision theory

1950

- Wald's Statistical Decision Functions
- Decision theory completed the full mathematization of statistical inference



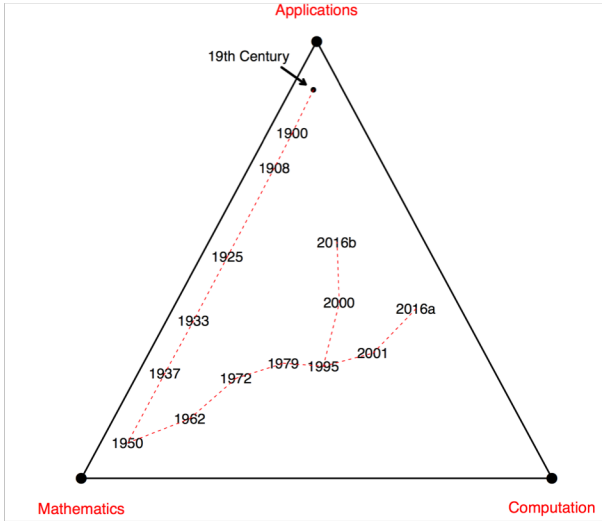
From Statistical Inference to Data Science

1962

- Tukey's paper "The future of data analysis" argued for a more application- and computation-oriented discipline
- Mosteller and Tukey later suggested changing the field's name to *data analysis*, a prescient hint of today's *data science*

1972

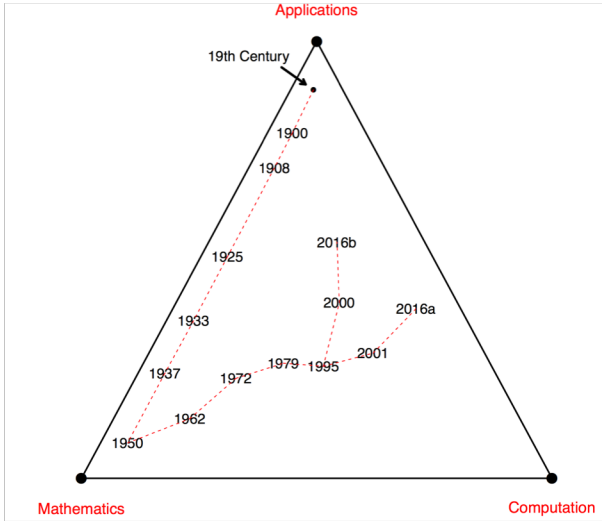
- Cox's proportional hazards paper
- Growing interest in biostatistical applications and particularly survival analysis



From Statistical Inference to Data Science

1979

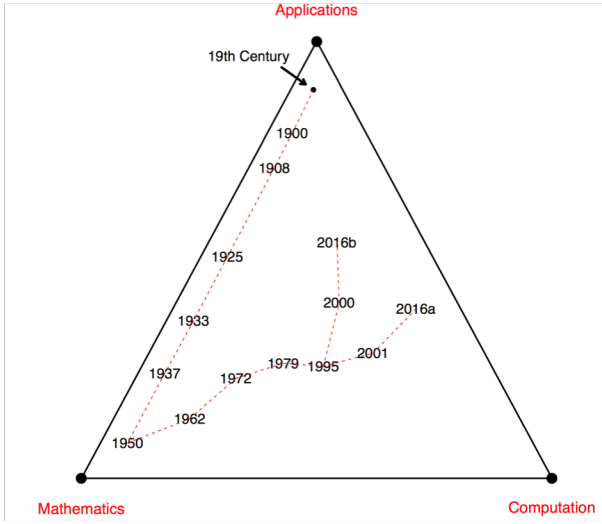
- The bootstrap, and later the widespread use of MCMC
- Electronic computation used for the extension of classic statistical inference.



From Statistical Inference to Data Science

1995

- This stands for false-discovery rates and, a year later, the lasso
- Both are computer-intensive algorithms, firmly rooted in the ethos of statistical inference



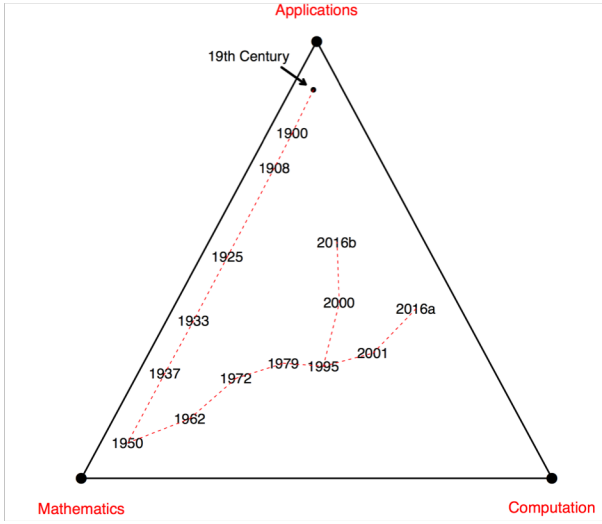
From Statistical Inference to Data Science

2000

- Microarray technology inspires enormous interest in large-scale inference, both in theory and as applied to the analysis of microbiological data.

2001

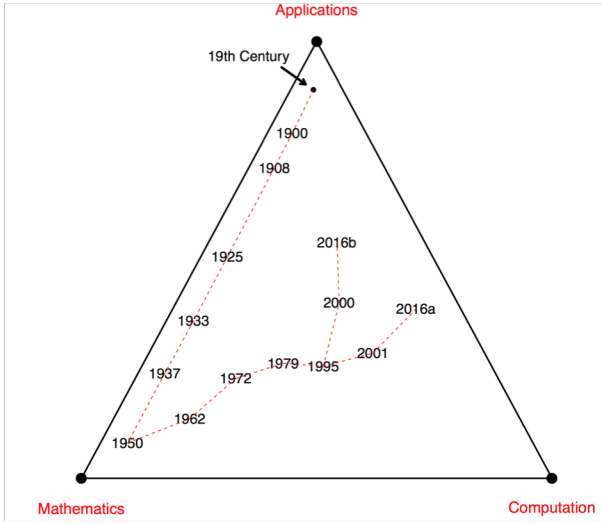
- Random forests
- Joins boosting and the resurgence of neural nets in the ranks of *machine learning* prediction algorithms



From Statistical Inference to Data Science

2016a

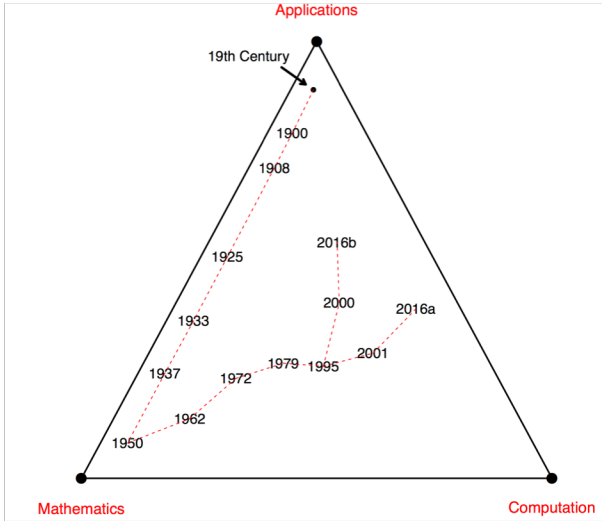
- Data science: a more popular successor to Tukey and Mosteller's "data analysis"
- At one extreme it seems to represent a statistics discipline without parametric probability models or formal inference.
- Data Science Association defines a practitioner as one who ". . . uses scientific methods to liberate and create meaning from raw data"
- In practice the emphasis is on
  - algorithmic processing of large data sets
  - for the extraction of useful information,
  - with prediction algorithms as exemplars



From Statistical Inference to Data Science

2016b

- This represents the traditional line of statistical thinking, but now energized with a renewed focus on applications
- Of particular applied interest are biology and genetics
- Genome-wide association studies (GWAS) show a different face of big data.
- Prediction is important here, but not sufficient for the scientific understanding of disease



Computer Science and Data Science

Since 1950

- Logic
- Probability Theory
- Representation and Query Language
- Databases
- Algorithms and Data Structures
- Programming
- Systems (HW/SW)
- …