IMS Presidential Address

Teaching Statistics in the Age of Data Science

Jon A. Wellner

University of Washington, Seattle

IMS Annual Meeting, & JSM, Baltimore, July 31, 2017

IMS Annual Meeting and JSM Baltimore

STATISTICS and DATA SCIENCE

- What has happened and is happening?
  - ▷ Changes in degree structures:
    many new MS degree programs in Data Science.
  - ▷ Changes in Program and Department Names;
    2+ programs with the name
    "Statistics and Data Science"
    Yale and Univ Texas at Austin.
  - ▷ New pathways in Data Science and Machine Learning at the PhD level: UW, CMU, and ...

- Changes (needed?) in curricula / teaching?

?

Full Disclosure:

1. Task from my department chair:

- **a.** Review the theory course offerings in the Ph.D. program in Statistics at the UW.
- **b.** Recommend changes in the curriculum, if needed.

2. I will be teaching Statistics 581 and 582, *Advanced Statistical Theory* during Fall and Winter quarters 2017-2018. What should I be teaching?

?

Exciting times for Statistics and Data Science:

- Increasing demand!
- Challenges of “big data”:
  - ▷ challenges for computation
  - ▷ challenges for theory
- Changes needed in statistical education?

Exciting times for Statistics and Data Science:

- Increasing demand!

  Projections, Bureau of Labor Statistics, 2014-24:

	Job Description	Increase %
•	Statisticians	34%
•	Mathematicians	21%
•	Software Developer	17%
•	Computer and Information Research Scientists	11%
•	Biochemists and Biophysicists	8%
•	Physicists and Astronomers	7%
•	Chemists and Materials Scientists	3%
•	Computer Programmer	-8%

The increasing demand for statisticians raises questions:

Q1 Can we meet the demand?

Q2 How should we be changing to meet the increased demand?

Q3 What changes should we be making in the teaching of statistics to attract the best and brightest students?

Q4 What should we be teaching?

Exciting times for Statistics and Data Science:

- Challenges of “big data”?
  - ▷ challenges for computation and analysis?
  - ▷ challenges for theory?
  - ▷ dismissed by Donoho* as a distinction between STAT & DS.

*Donoho (2015): 50 years of Data Science, p 6-7. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Exciting times for Statistics and Data Science:
Changes needed in statistical education?

- changes in degree structures?
- changes in curricula?
    - ▷ high school
    - ▷ all college/university
    - ▷ undergraduate majors
    - ▷ graduate: MS degree
    - ▷ graduate: PhD degree
- changes in the modes of teaching?

Differing Views:

from (2013) Report of the London Workshop on the Future of Statistical Sciences

- Marie Davidian, past President of the ASA:
  “Statistical sciences are at a crossroads ... ” and change is needed.

- Terry Speed, past President of the IMS:
  “We have a great tradition ... ” and ... “we are in this business for the long term”. ... we need to “evolve and adapt”.

- Richard De Veaux:
  “Statistics education remains mired in the 20th (some would say the 19th) century.”

Differing Views:

(from (2013) Report of the London Workshop on the Future of Statistical Sciences

- Marie Davidian, past President of the ASA:

> "I believe that the statistical sciences are at a crossroads, and that what we do currently ... will have profound implications for the future state of our discipline. The advent of big data, data science, analytics, and the like requires that we as a discipline cannot sit idly by ... but must be proactive in establishing both our role in and our response to the 'data revolution' and develop a unified set of principles that all academic units involved in research, training, and collaboration should be following. ... At this point, these new concepts and names are here to stay, and it is counterproductive to spend precious energy on trying to change this. We should be expending our energy instead to promote statistics as a discipline and to clarify its critical role in any data-related activity."

Differing Views:

from (2013) Report of the London Workshop on the Future of Statistical Sciences

- Terry Speed, past President of the IMS:

> "Are we doing such a bad job that we need to rename ourselves data scientists to capture the imagination of future students, collaborators, or clients? Are we so lacking in confidence ... that we shiver in our shoes the moment a potential usurper appears on the scene? Or, has there really been a fundamental shift around us, so that our old clumsy ways of adapting and evolving are no longer adequate? ... I think we have a great tradition and a great future, both far longer than the concentration span of funding agencies, university faculties, and foundations. ... We might miss out on the millions being lavished on data science right now, but that's no reason for us to stop trying to do the best we can at what we do best, something that is far wider and deeper than data science. As with mathematics more generally, we are in this business for the long term. Let's not lose our nerve."[†]

[†] January 2014 AMSTAT News

Back tracking. History part 1:

Harold Hotelling dates:

- 1895: b. Minnesota; 1904: moved to Washington.
- 1915-19: BA in Journalism, UW Seattle (& Army)
- 1920-21: MS Math, UW Seattle
- 1924: PhD Math, Princeton
- 1924-31: Food Research Institute and
  Assistant Professor, Mathematics (1927-31), Stanford
- 1929: 6 months with R. A. Fisher at Rothamsted.
- 1931-46: Columbia, Economics (& SRG)
- 1946-1973: UNC, Statistics
- 1973: d. Chapel Hill, NC

Back tracking. History part 1:

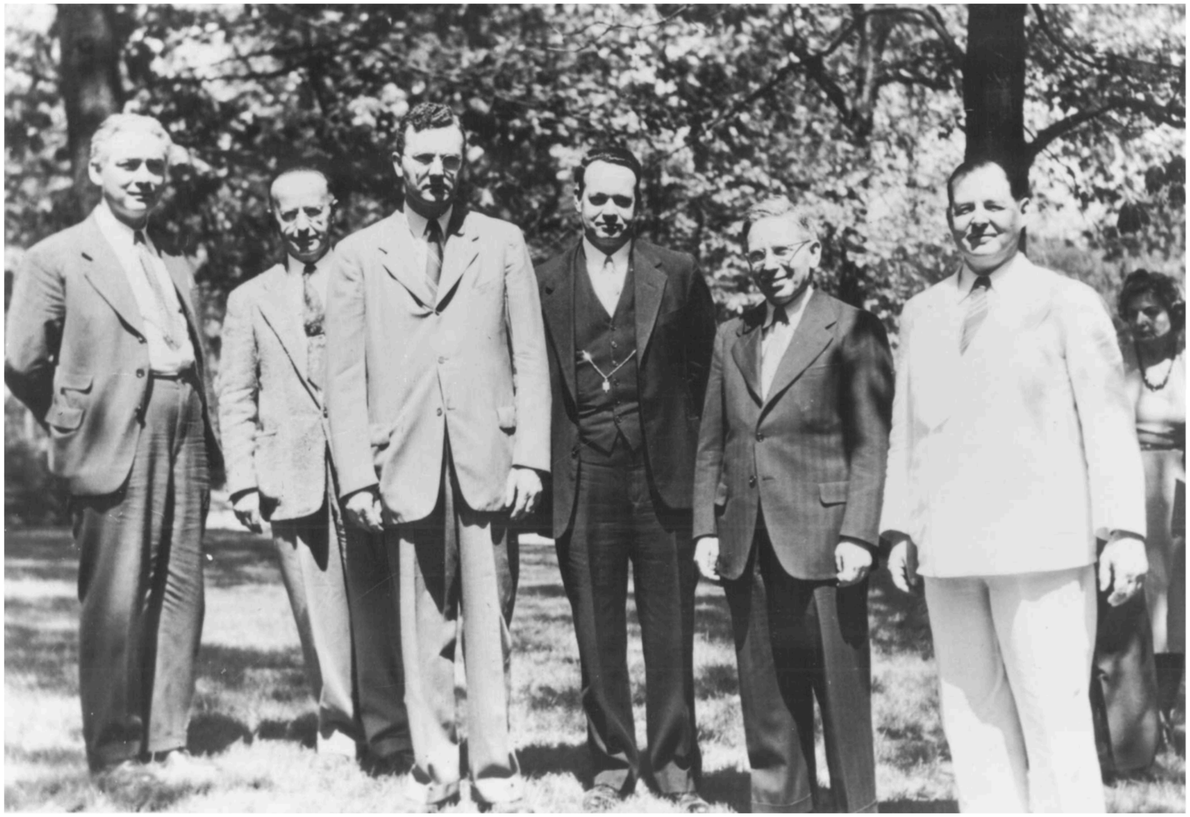


FIG. 3. *A group of Founding Fathers: (from left) Willy Feller, Walter Shewhart, Sam Wilks, Paul Dwyer, Abraham Wald and Harold Hotelling. Photo probably taken in early or middle 1940s.*

Hotelling's 1940 paper: "Teaching of statistics"
IMS Invited talk, Hannover, New Hampshire

- H. Hotelling (chair), Walter Bartky, W. E. Deming, M. Friedman, and P. Hoel.
- Hotelling's talk laid out the difficulties involved in the teaching of statistics as of 1940:
  - ▷ Failure to recognize statistics as a science requiring specialists to teach it.
  - ▷ Shortage of qualified instructors.
- Strongly influenced Neyman
- Discussion by W. E. Deming: raises issues relevant for applications;

  > "Above all, a statistician must be a scientist. A scientist does not neglect any pertinent information … ."

As noted by Ingram Olkin:

> “His (Hotelling’s) 1940 paper on the teaching of statistics had a phenomenal impact. Jerzy Neyman stated that it was one of the most influential papers in statistics. Faculty attempting to convince university administrators to form a Department of Statistics often used this paper as an argument why the teaching of statistics should be done by statisticians and not by faculty in substantive fields that use statistics.”

Two other works by Hotelling on teaching statistics:

- The teaching of statistics: A report of the IMS Committee on the Teaching of Statistics.
  H. Hotelling (chair), Walter Bartky, W. E. Deming, M. Friedman, and P. Hoel. *Ann. Math. Statist.* **19** (1948), 95 - 115.

- The place of statistics in the university. *Proc. Berk. Symp. Math. Statist. Prob.* (1949). Part II (by Hotelling) of the 1948 *Annals* paper.

Part I of the 1948 Committee Report addressed the following questions:

(1) Who are the prospective students of statistics?

- **(a)** All college (university) students.
- **(b)** Future consumers of statistics.
- **(c)** Future users of statistical methods.
- **(d)** Future producers and teachers of statistical methods.

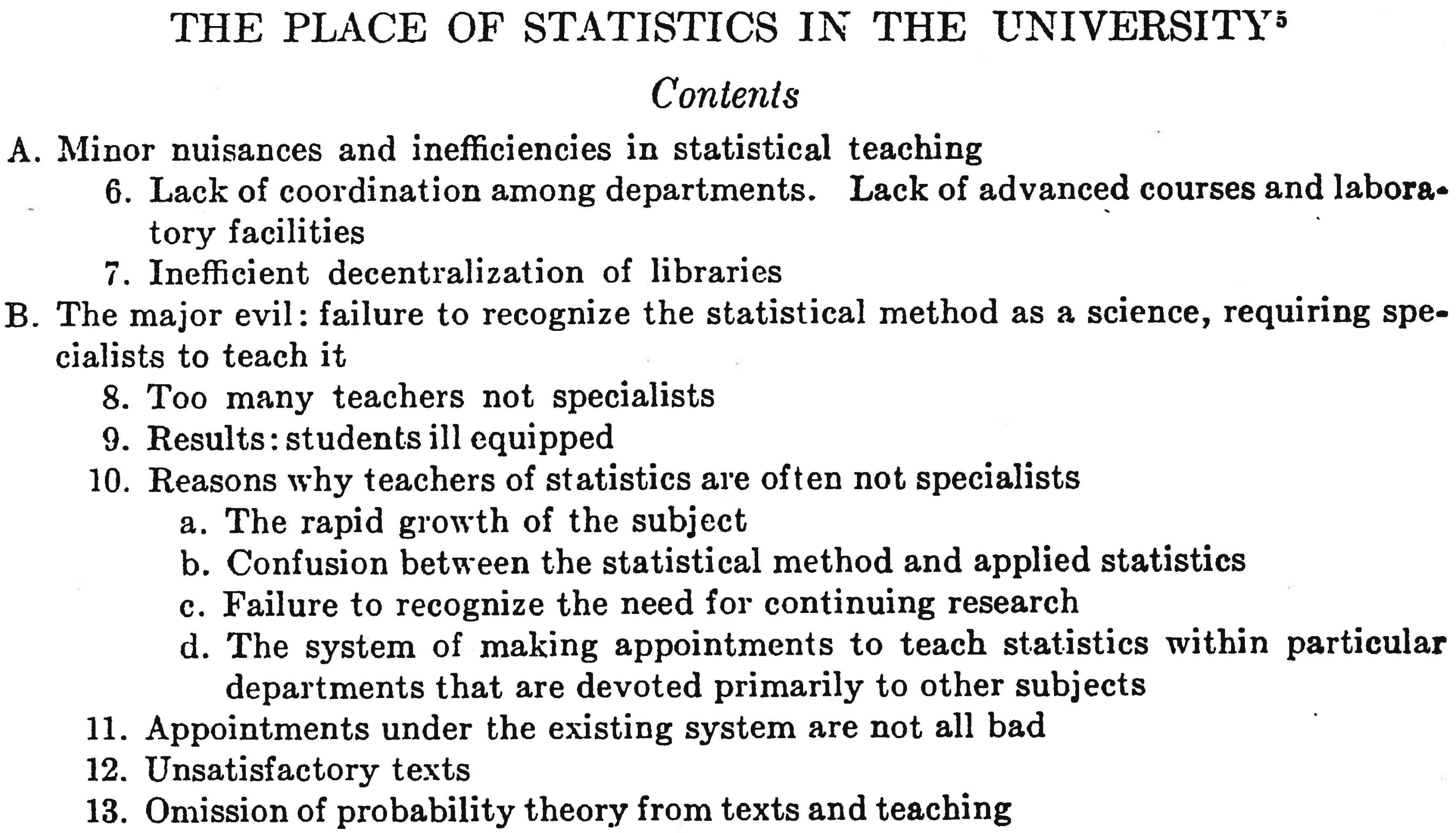
(2) What should they be taught?

(3) Who should teach statistics?

(4) How should the teaching of statistics be organized?

(5) What should be done about adult education?

Part II of the 1948 report was written by Hotelling and reflected his views:



THE PLACE OF STATISTICS IN THE UNIVERSITY[5]

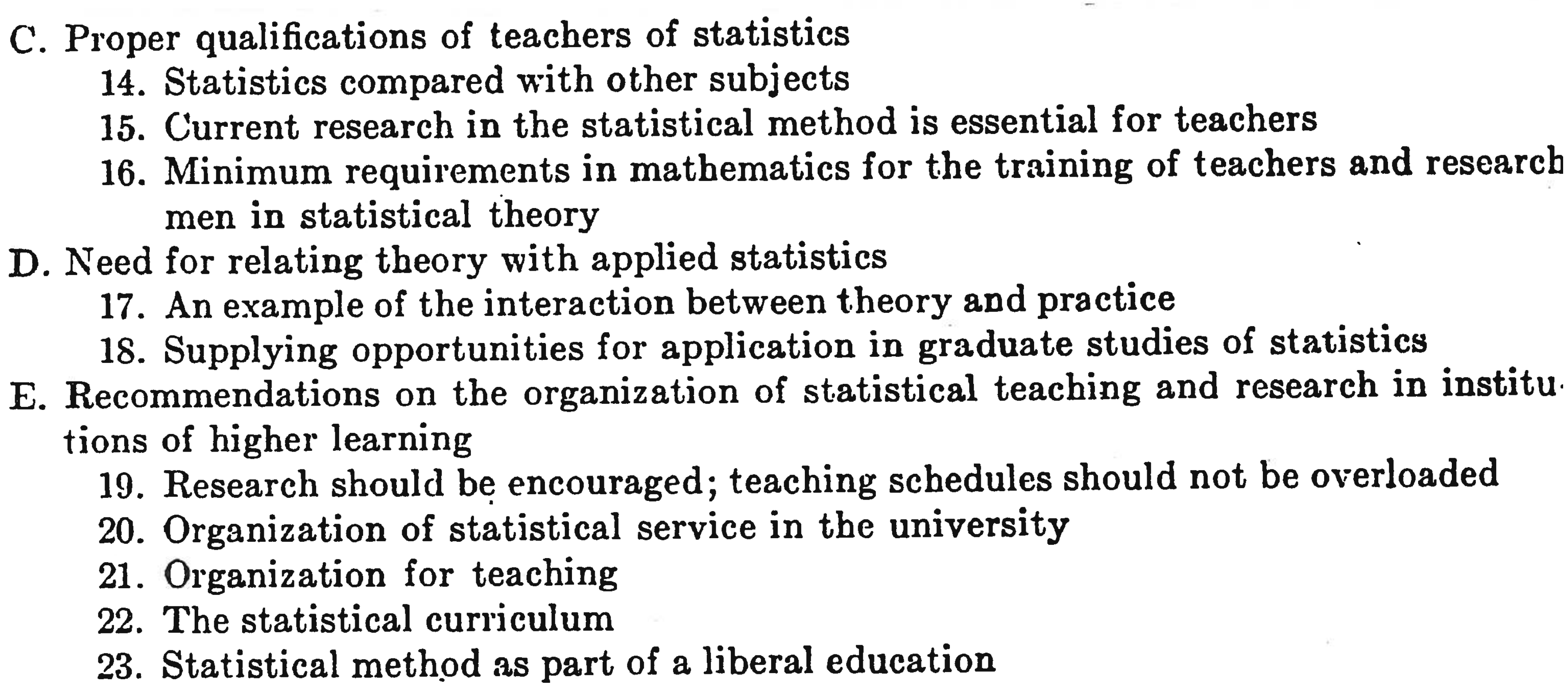
Contents

A. Minor nuisances and inefficiencies in statistical teaching

   6. Lack of coordination among departments. Lack of advanced courses and laboratory facilities
   7. Inefficient decentralization of libraries

B. The major evil: failure to recognize the statistical method as a science, requiring specialists to teach it

   8. Too many teachers not specialists
   9. Results: students ill equipped
   10. Reasons why teachers of statistics are often not specialists
       a. The rapid growth of the subject
       b. Confusion between the statistical method and applied statistics
       c. Failure to recognize the need for continuing research
       d. The system of making appointments to teach statistics within particular departments that are devoted primarily to other subjects
   11. Appointments under the existing system are not all bad
   12. Unsatisfactory texts
   13. Omission of probability theory from texts and teaching



C. Proper qualifications of teachers of statistics

14. Statistics compared with other subjects
15. Current research in the statistical method is essential for teachers
16. Minimum requirements in mathematics for the training of teachers and research men in statistical theory

D. Need for relating theory with applied statistics

17. An example of the interaction between theory and practice
18. Supplying opportunities for application in graduate studies of statistics

E. Recommendations on the organization of statistical teaching and research in institutions of higher learning

19. Research should be encouraged; teaching schedules should not be overloaded
20. Organization of statistical service in the university
21. Organization for teaching
22. The statistical curriculum
23. Statistical method as part of a liberal education

- The (1940) paper and (1948) committee report *Ann. Math. Statist.* 19 (1948), 95-115, were reprinted in *Stat. Sci.* 3 (1988), 63 - 108.

- Followed by a discussion by: David S. Moore, J. V. Zidek, Kenneth J. Arrow, Harold Hotelling Jr, Ralph Bradley, W. Edwards Deming, Shanti S. Gupta, and Ingram Olkin.

- Discussion(s) reflected the long-standing (and creative tensions) between theory/mathematics and applications/data analysis.

- S. S. Gupta's view of Hotelling's papers:

  > "He rightly visualized the academic statistician as a toolmaker who 'must not put all his time on using the tools he makes', but must focus his/her attention on the tools themselves."

Review of Hotelling committee report (1948) and another paper on teaching of statistics by a committee of the RSS by Truman L. Kelley, Professor of Education, Harvard: From Kelly's review:

> "It seems to the reviewer that there is implicit in the British recommendation an induction of the student into statistics via the subject matter of his field of specialization, and in the American an induction via logic, including principles of mathematics and probability. It is needless to say that these approaches are far asunder."

From Hotelling's paper (page 466):

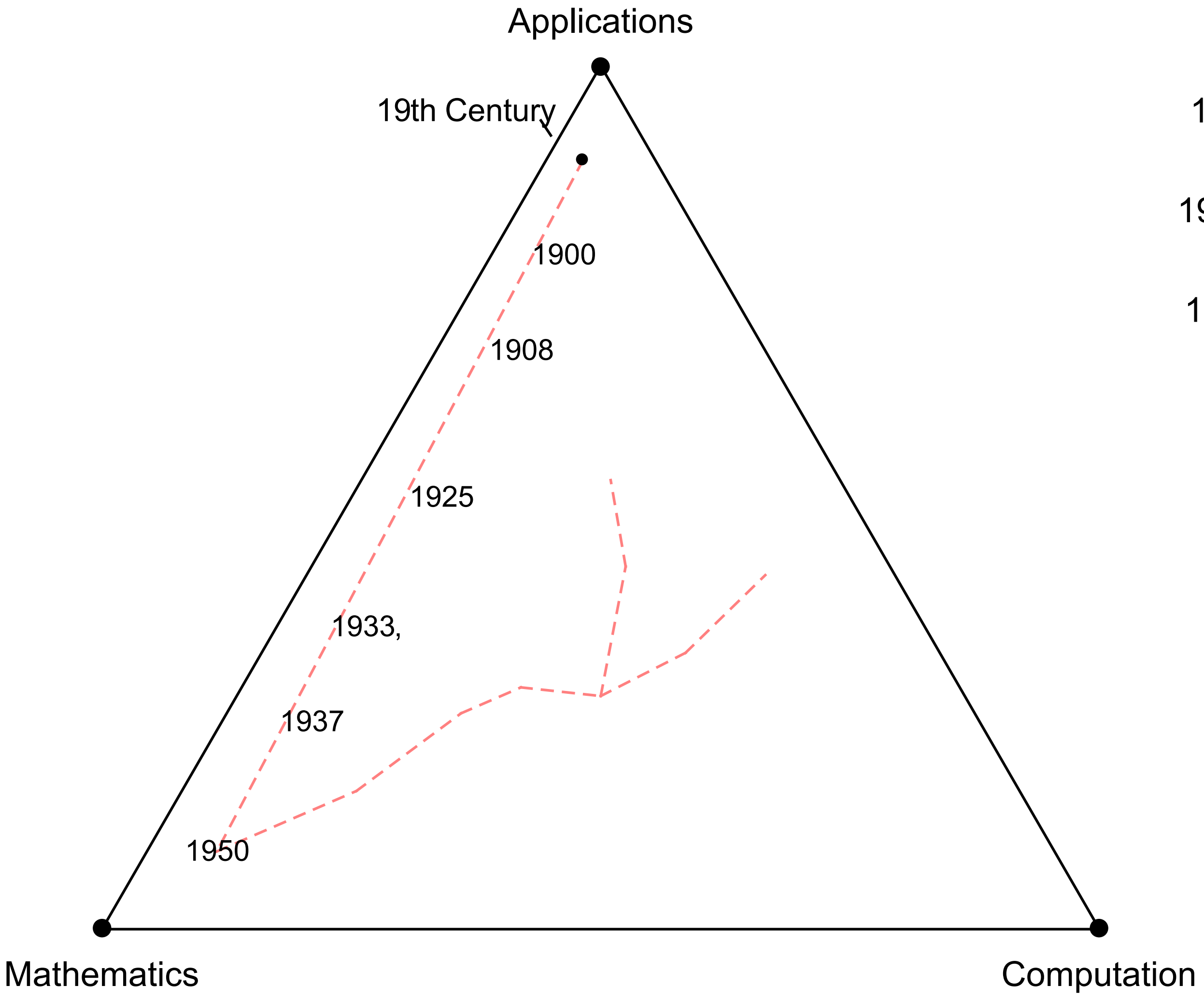
> "Statistical theory is a big enough thing in itself to absorb the full-time attention of a specialist teaching it, without his going out into applications too freely. Some attention to applications is indeed valuable, and perhaps even indispensable as a stage in the training of a teacher of statistics and as a continuing interest. But particular applications should not dominate the teaching of the fundamental science, any more than particular diseases should dominate the teaching of anatomy and bacteriology to pre-medical students."

These two quotes are a small sample of the long-running tensions within statistics and statistics education. In my view, these tensions are an inherent part of the process of creating new statistical methods and perspectives.

Kelley continues:

> “The American committee, by omission and by inclusion, reveals what it considers to be preparatory background for students of statistics. It at no point cites knowledge of data in some scientific field as essential. . . . The American committee deplores the general lack of mathematical competence of most teachers of statistics in different subject matter fields. This is deplorable as is their lack of knowledge of the genius of data in their fields. However, the progress of recent decades should make one optimistic, and these two committee reports should encourage college presidents to strengthen and broaden the instruction in both mathematical and applied statistics.”

Adaptation of Efron & Hastie (2016), *Epilogue*, p. 448



1900: Pearson Chi−square statistic

1908: Student's t statistic

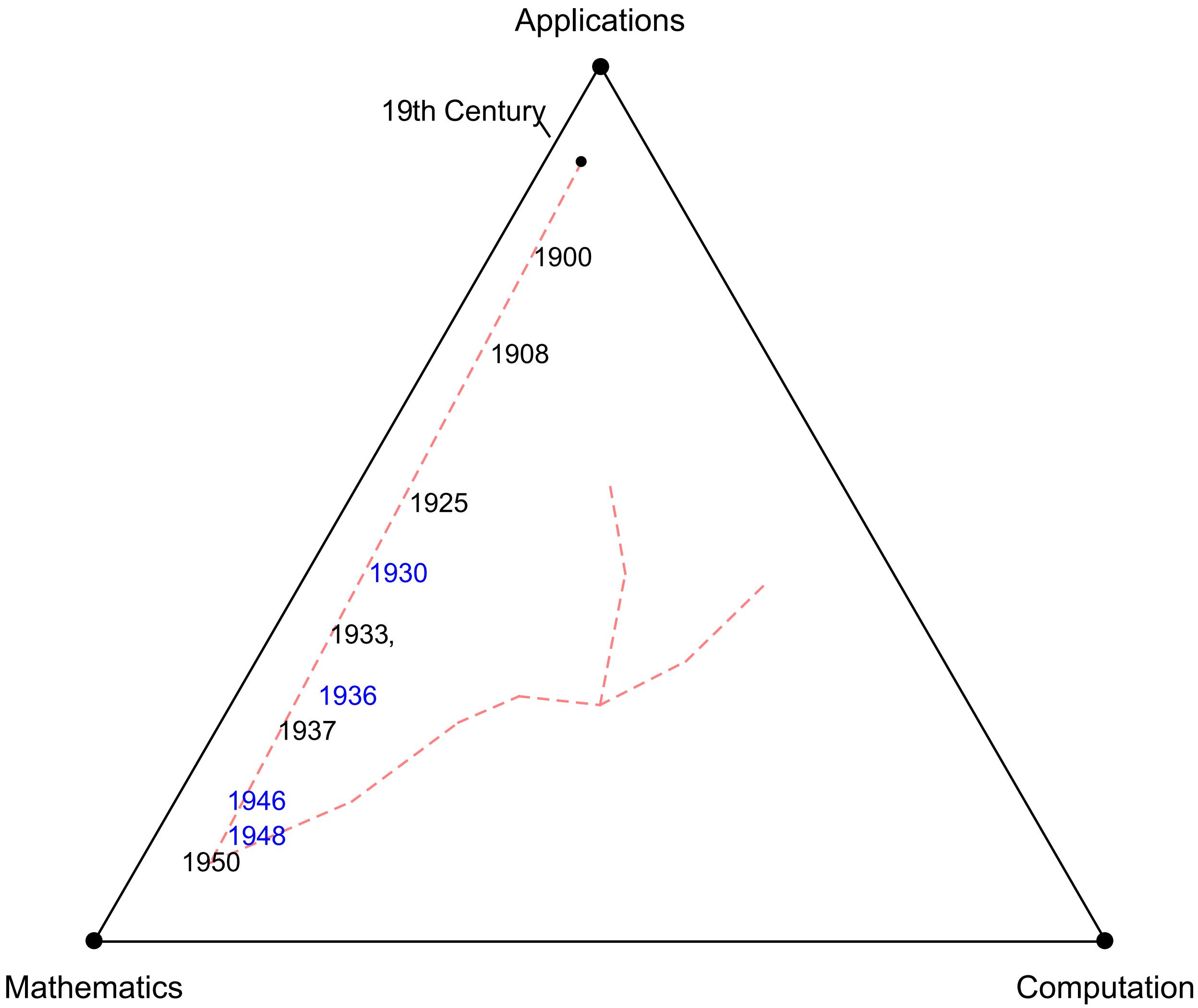
1925: Fisher's estimation paper

1933: Neyman−Pearson; optimal testing

1937: Neyman's optimal CI's

1950: Wald's decision theory

Overlay 1 for Efron & Hastie (2016), *Epilogue*, p. 448



1900: Pearson Chi−square statistic

1908: Student's t statistic

1925: Fisher's estimation paper

1930: Hotelling Max Lik

1933: Neyman−Pearson; optimal testing

1936: Doob Max Lik

1937: Neyman's optimal CI's

1946: Cramer's book

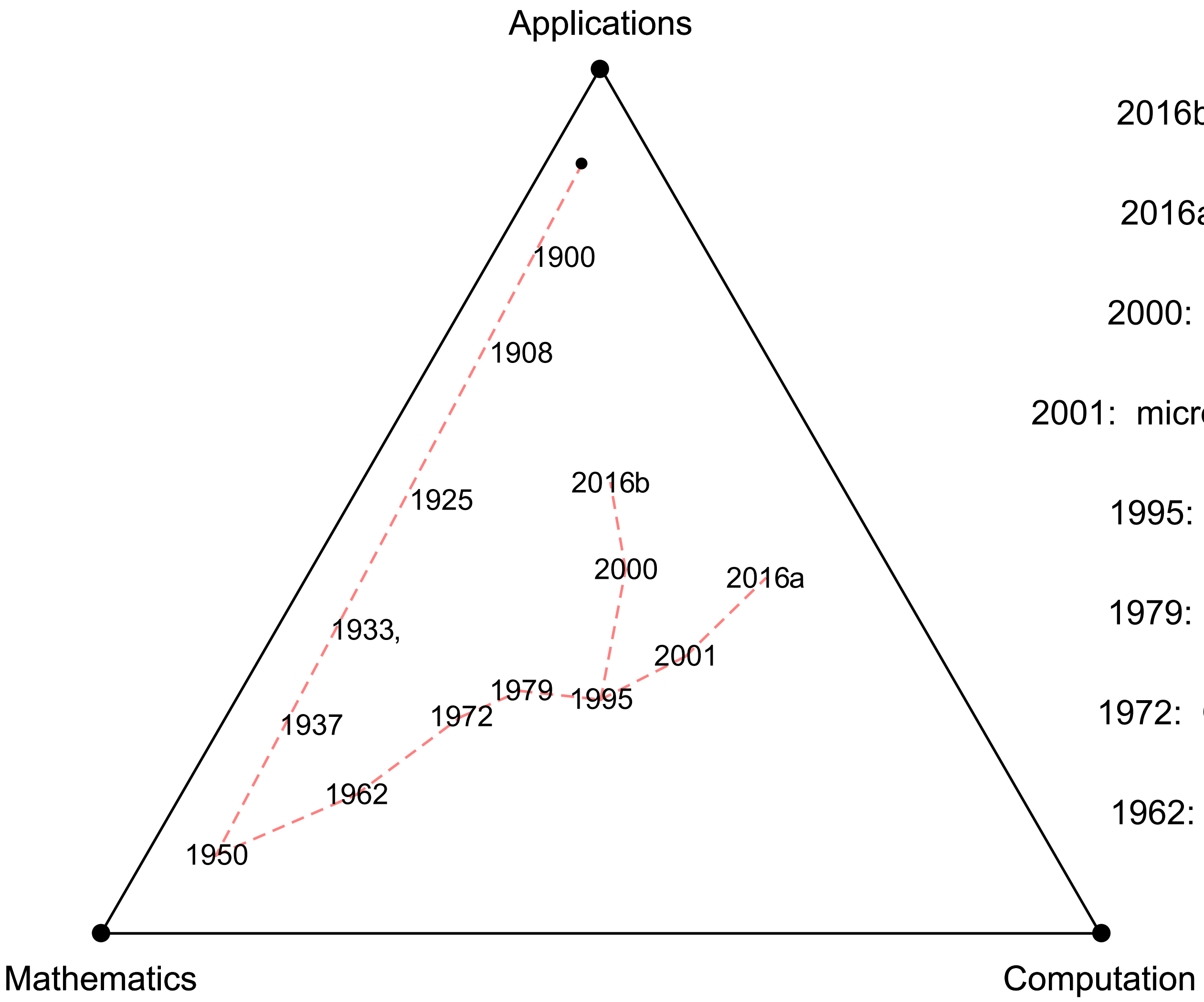
1950: Wald's decision theory

History part 2 (shortened)

A second set of important developments:

- John Tukey's (1962) paper, *The future of data analysis*. Tukey called for a revamping of academic statistics, and pointed to a new *science* focused on data analysis.
- Bill Cleveland (1993) and John Chambers (2001) developed Tukey's ideas further.
- Leo Breiman's (2001) *Two cultures ...* paper clearly delineated the differing approaches to data analysis which developed in the years since Tukey (1962).
  - ▷ Predictive modeling; *Common Task Framework*
  - ▷ Generative modeling; inference
- Donoho (2015), *50 years of Data Science*, gives a guide to this history, explains the key role of the *Common Task Framework*, and provides an updated road map to what he calls *Greater Data Science*.

Adaptation of Efron & Hastie (2016), *Epilogue*, p. 448



2016b: genomics, biol

2016a: data science

2000: random forests, ML

2001: microarrays; large scale inference

1995: FDR &LASSO

1979: bootstrap/MCMC

1972: Cox proportional hazards

1962: Tukey, Data Analysis

Fast forward to 2002-2004:

- By the beginning of the 21st century the era of data science, “big data”, and machine learning was well underway. Breiman’s (2001) paper clearly delineated the differences in approaches to data analysis which had developed in the years since Tukey (1962).

- In May 2002, the NSF hosted a workshop on future challenges and opportunities for the statistics community.

The resulting "*Report on the Future of Statistics*" by Bruce Lindsay, Jon Kettenring, and David O. Siegmund (2004):

(a) addressed features of the statistical enterprise relevant to the NSF;

(b) biostatistics was not included;

(c) teaching of statistics was not addressed explicitly, but indirectly through "manpower" problems;

(d) identified opportunities and needs for the "core of statistics":

> "If there is exponential growth in data collected and in the need for data analysis, why is "core research relevant? . . . Because unifying ideas can tame this growth, and the core area of statistics is the one place where these ideas can happen and be communicated throughout science."

Comparisons: then (1940) and now (or 2015)

Of course there have been big changes both in statistics and in the world of science in general since Hotelling's time and even since the Lindsay-Kettenring-Siegmund report of 2004. Here is an oversimplified summary:

Comparisons

	then (1940)	now (2015)
# of departments of statistics	5 – 10	≈ 60
# of departments of biostatistics	≈ 1	≈ 43
# of graduate students, statistics	< 50?	4597 (24%)
# of graduate students, biostatistics	< 10?	1960 (14%)
IMS membership	< 100?	3500
Computer clock speed	5 – 10 hz Zuse (1941)	> 2.7 Ghz Mac Powerbook
Terminology/ Department Names	Mathematical Statistics Applied Statistics	Statistics Data Analysis Data Science

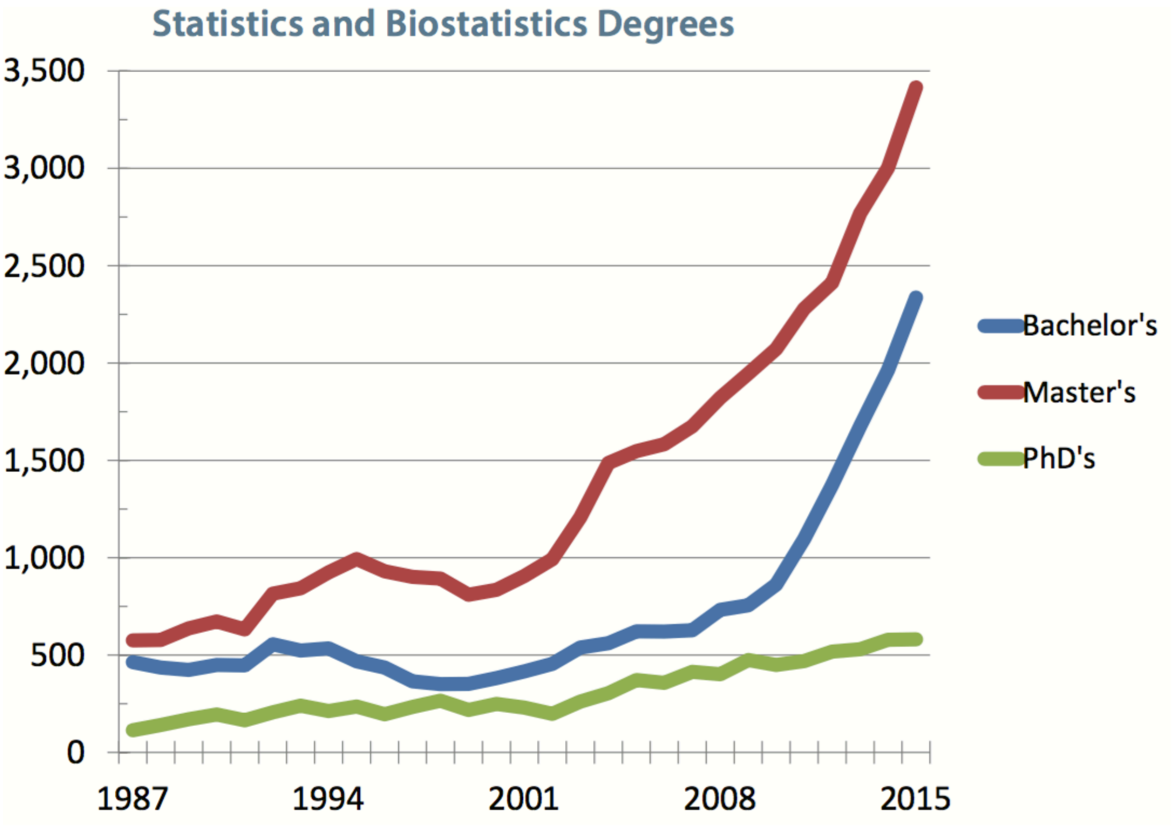


Figure 1. Statistics and biostatistics degrees at the bachelor's, master's, and doctoral levels in the United States.
Data source: NCES IPEDS

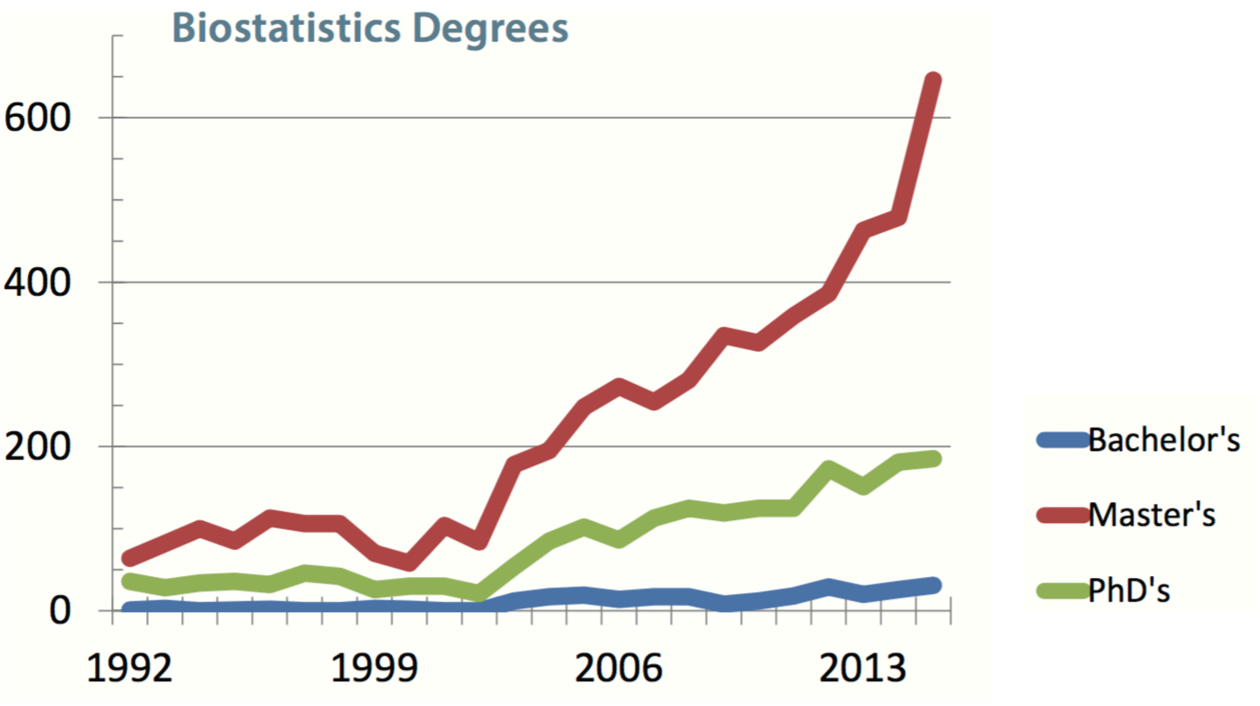


Figure 2. Biostatistics degrees by degree level awarded in the United States

Source: AMSTAT News, October 2016; Steve Pierson

MS curricula in Data Science

The MS in Data Science and Machine Learning:

What is the curriculum?

Donoho (2015) section 7 reviews a typical such Data Science MS degree curriculum: the core of the MS Data Science curriculum includes:

```
Research Design and Application for Data and Analysis
Exploring and Analyzing Data
Storing and Retrieving Data
Applied Machine Learning
Data Visualization and Communication
```

while the advanced courses include:

Experiments and Causal Inference

Applied Regression and Time Series Analysis

Legal, Policy, and Ethical Considerations for Data Scientists

Machine Learning at Scale

Scaling up! Really big data.

Capstone course (with data analysis project)

The program at Berkeley is run by the Information School.

At (my home) the University of Washington, the DS MS program is run by the E-Science Institute (with co-operation from Statistics, CS, and Biostatistics):

```
Introduction to Statistics and Probability
Data Visualization & Exploratory Analytics
Applied Statistics and Experimental Design
Data Management for Data Science
Statistical Machine Learning for Data Scientists
Software Design for Data Science
Scalable Data Systems and Algorithms
Human-Centered Data Science
Data Science Capstone Project
```

There is clear overlap in both lists with courses offered in a traditional statistics MS program, but with a number of substitutions from a Computer Science. Stat & Biostat Faculty: Adrian Dobra, Zaid Harchaoui, Brian Leroux.

MS Programs in Data Science and Analytics: survey / interviews in AMSTAT News, April and June 2017:

University of Tennessee
George Mason University
Bentley University
University of Minnesota
NC State University
Penn State University
University of Vermont
University of Wisconsin-Madison
South Dakota State University
Harvard

Query:

“Do you have any advice for institutions considering the establishment of such a degree?”

Reply: Mark Craven, Univ of Wisconsin-Madison:

“I would advise any institution considering this area to build on existing partnerships between statistics, biostatistics, computer sciences, and biomedical informatics. No one unit can or should “own” this area, so proceeding in a broad and inclusive way makes the most sense.”

Donoho (2015) gives an analysis of the Berkeley Data Science curriculum in the context of Tukey's critiques and writings. Donoho writes:

> ``Although my heroes Tukey, Chambers, Cleveland, and Breiman would recognize positive features in these programs, it's difficult to say whether they would approve of their long-term direction - or if there is even a long-term direction to comment about.... Data Science Masters curricula are compromises taking some material out of a Statistics masters program to make room for large database training; or equally, as taking some material out of a database masters in CS and inserting some statistics and machine learning. Such a compromise helps administrators to quickly get a degree program going, without providing any guidance about the long-term direction of the program and about the research which its faculty will pursue. What long-term guidance could my heroes have provided?''

Ph.D.curricula in Statistics

At the UW: Ph.D. program has four possible tracks:

- Normal or Basic track.
  Requirements: 581-582-583 & 570-571
- Statistical genetics
- Statistics for the Social Sciences
- Machine learning and big data:
  - ▷ 570, 581-582 (advanced stat theory),
  - ▷ ML/BD Core:

    (i) Foundational ML: STAT 535

    (ii) One advanced ML course: STAT 538 or STAT 548

    (iii) One CSE course: CSE 544 (Databases)
    or CSE 512 (Visualization)

    (iv) One elective:

    * Advanced Statistical Learning (STAT 538)

Ph.D.curricula in Statistics

* Machine Learning for Big Data (STAT 548)
* Graphical Models (CSE 515)
* Visualization (CSE 512)
* Databases (CSE 544)
* Convex Optimization (EE 578)

UW PhD student numbers by tracks: 2001-2016

track	Graduated	Current	Total
Normal track, Stat	83	37	120
Normal track, Biost	103	49	152
StatGen, Stat	13	3	16
StatGen, Biost	5	3	8
Stat in Soc Sci:	5	1	6
ML-BD:	1	13	14
total, Stat	102	54	156
total Biostat	108	52	160

My heroes:

- H. Chernoff
- J. L. Doob
- R. A. Fisher
- Harold Hotelling
- Wassily Hoeffding
- Jaroslav Hajek
- Jack Kiefer
- Lucien Le Cam
- Charles Stein
- Abraham Wald

Future areas needing more math:

- manifold learning
- topological data analysis
- nonstandard data types: functions, trees, images,
- . . .

What's in Stat 581 - 582 - 583, Advanced Stat Theory now?
Outline for Stat 581:

- Inequalities; basic asymptotic theory in statistics. Examples:
  - ▷ robustness (or lack of robustness) of normal theory tests
  - ▷ chi-square statistic and power of chi-square tests under fixed and local alternatives.
  - ▷ limit theory for fixed dimension linear regression
  - ▷ limit theory for correlation coefficients
  - ▷ limit theory for empirical distributions and sample quantiles.
  - ▷ examples from survival analysis / censored data
- Lower bounds for estimation
  - ▷ Multiparameter Cramér - Rao lower bounds.

  - Superefficiency & introduction to Hajek-LeCam convolution theorem and local asymptotic minimax theorems.
  - Simple Lower bound Lemma via two point inequalities.

- Classical (and nonparametric) maximum likelihood:
  - Existence; empirical d.f. & and empirical measure as MLEs
  - Algorithms, one step approximations, and EM
  - LR, Wald, and Rao tests: fixed and local alternatives.
  - Brief introduction to agnostic viewpoint: what if the model fails?

Outline for Stat 582:

- Elementary Decision Theory: Bayes rules, minimax rules, and connections.
- Bayes theory, inadmissibility, and empirical Bayes.
- Optimal tests and tests optimal in subclasses: eliminating nuisance parameters by conditioning and invariance.

Unifying aspects of the “statistics core”: (a) Notions of optimality (b) design of experiments (c) (survey) sampling theory (d) classical and modern multivariate analysis.

- (1) asymptotic theory: LLNs and CLTs.
- (2) uniform laws of large numbers and uniform central limit theorems i.e. empirical process theory.
- (3) optimality theory via upper bounds and lower bounds (parametric, semiparametric, and nonparametric)
- (4) inequalities (exponential, basic, oracle)
- (5) convexity theory
- (6) optimization theory.

? Stat 581-582 this year ?

New in Stat 581 - 582 this coming year?

New? Large scale hypothesis testing and FDR's?
New? More on empirical Bayes?
New? More on convexity?
New? More on empirical process theory?
(What will need to be reduced or deleted?)

I don't know exactly yet, but I'm working on it . . .
. . . and on the report to my chair.

IMS Data Science Group!

- Initiated in 2015 by Bin Yu and Richard Davis with assistance from David Dunson.
- New Group Coordinators: Sofia Olhede (s.olhede@ucl.ac.uk) and Patrick Wolfe (p.wolfe@ucl.ac.uk).
- Watch for further developments soon!

From Efron & Hastie (2016), Preface, page xvii:

> "Useful disciplines that serve a wide variety of demanding clients run the risk of losing their center. Statistics has managed, for the most part, to maintain its philosophical cohesion despite a rising curve of outside demand. The center of the field has ... moved in the past sixty years, from its traditional home in mathematics and logic toward a more computational focus."

From Efron & Hastie (2016), Epilogue, page 447:

> "It is the job of statistical inference (theory) to connect 'dangling algorithms' to the central core of well-understood methodology. The connection process is already underway."

My Views:

- Embrace and encourage data science!
- Continue evolving the curriculum to teach the unifying themes of statistical research.
- Keep doing what statisticians do best: question, question, question ... and then provide the best answers possible based on the available data.
- Attract the best and brightest students to research work in statistics.
- Teach what we know!

ystadegau
estatistikak
estadística
statistiques
στατιστικών
إحصائيات
statistică
statistikë
štatistika
statistiek
statistieke
staitisticí
istatistika
statisztika
istatistik
آمار
статыстыка
statystyka
統計
statistica
পরিসংখ্যান
tilasto
statistik
статистика
statistikk
statistics
số liệu thống kê
takwimu
estatística
statistiko
統計
statistikat
tölfræði
סטטיסטיקה
statistic
statistika