JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

Taylor & Francis
Taylor & Francis Group

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

ABSTRACT

More than 50 years ago, John Tukey called for a reformation of academic statistics. In "The Future of Data Analysis," he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or "data analysis." Ten to 20 years ago, John Chambers, Jeff Wu, Bill Cleveland, and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland and Wu even suggested the catchy name "data science" for this envisioned field. A recent and growing phenomenon has been the emergence of "data science" programs at major universities, including UC Berkeley, NYU, MIT, and most prominently, the University of Michigan, which in September 2015 announced a $100M "Data Science Initiative" that aims to hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; yet many academic statisticians perceive the new programs as "cultural appropriation." This article reviews some ingredients of the current "data science moment," including recent commentary about data science in the popular media, and about how/whether data science is really different from statistics. The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for "scaling up" to "big data." This chosen superset is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next 50 years. Because all of science itself will soon become data that can be mined, the imminent revolution in data science is not about mere "scaling up," but instead the emergence of scientific studies of data analysis science-wide. In the future, we will be able to predict how a proposal to change data analysis workflows would impact the validity of data analysis across all of science, even predicting the impacts field-by-field. Drawing on work by Tukey, Cleveland, Chambers, and Breiman, I present a vision of data science based on the activities of people who are "learning from data," and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today's data science initiatives, while being able to accommodate the same short-term goals. *Based on a presentation at the Tukey Centennial Workshop, Princeton, NJ, September 18, 2015.*

ARTICLE HISTORY
Received August 2017
Revised August 2017

KEYWORDS
Cross-study analysis; Data analysis; Data science; Meta analysis; Predictive modeling; Quantitative programming environments; Statistics

1. Today's Data Science Moment

In September 2015, as I was preparing these remarks, the University of Michigan announced a $100 million "Data Science Initiative" (DSI)[1], ultimately hiring 35 new faculty.

The university's press release contains bold pronouncements:

> *"Data science has become a fourth approach to scientific discovery, in addition to experimentation, modeling, and computation," said Provost Martha Pollack.*

The website for DSI gives us an idea what data science *is*:

> *"This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and inter-disciplinary applications."*

This announcement is not taking place in a vacuum. A number of DSI-like initiatives started recently, including

(A) Campus-wide initiatives at NYU, Columbia, MIT, …
(B) New master's degree programs in data science, for example, at Berkeley, NYU, Stanford, Carnegie Mellon, University of Illinois, …

There are new announcements of such initiatives weekly.[2]

2. Data Science "Versus" Statistics

Many of my audience at the Tukey Centennial—where these remarks were originally presented—are applied statisticians, and consider their professional career one long series of exercises in the above "*…collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of …applications.*" In fact, some presentations at the Tukey Centennial were exemplary narratives of "*…collection, management, processing, analysis,*

CONTACT David Donoho *donoho@stanford.edu*

[1]For a compendium of abbreviations used in this article, see Table 1.

[2]For an updated interactive geographic map of degree programs, see *http://data-science-university-programs.silk.co*

Table 1. Frequent acronyms.

Acronym	Meaning
ASA	American Statistical Association
CEO	Chief Executive Officer
CTF	Common Task Framework
DARPA	Defense Advanced Projects Research Agency
DSI	Data Science Initiative
EDA	Exploratory Data Analysis
FoDA	*The Future of Data Analysis* 1962
GDS	Greater Data Science
HC	Higher Criticism
IBM	IBM Corp.
IMS	Institute of Mathematical Statistics
IT	Information Technology (the field)
JWT	John Wilder Tukey
LDS	Lesser Data Science
NIH	National Institutes of Health
NSF	National Science Foundation
PoMC	*The Problem of Multiple Comparisons* 1953
QPE	Quantitative Programming Environment
R	R – a system and language for computing with data
S	S – a system and language for computing with data
SAS	System and language produced by SAS, Inc.
SPSS	System and language produced by SPSS, Inc.
VCR	Verifiable Computational Result

visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of …applications."

To statisticians, the DSI phenomenon can seem puzzling. Statisticians see administrators touting, as new, activities that statisticians have already been pursuing daily, for their entire careers; and which were considered standard already when those statisticians were back in graduate school.

The following points about the U of M DSI will be very telling to such statisticians:

- U of M's DSI is taking place at a campus with a large and highly respected Statistics Department
- The identified leaders of this initiative are faculty from the Electrical Engineering and Computer Science department (Al Hero) and the School of Medicine (Brian Athey).
- The inaugural symposium has one speaker from the Statistics department (Susan Murphy), out of more than 20 speakers.

Inevitably, many academic statisticians will perceive that statistics is being marginalized here;[3] the implicit message in these observations is that statistics is a part of what goes on in data science but not a very big part. At the same time, many of the concrete descriptions of what the DSI will *actually do* will seem to statisticians to be bread-and-butter statistics. Statistics is apparently the word that dare not speak its name in connection with such an initiative![4,5]

Searching the web for more information about the emerging term "data science," we encounter the following definitions from the Data Science Association's "Professional Code of Conduct"[6]

> "Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

To a statistician, this sounds an awful lot like what applied statisticians do: use methodology to make inferences from data. Continuing:

> "Statistics" means the practice or science of collecting and analyzing numerical data in large quantities.

To a statistician, this definition of statistics seems already to encompass anything that the definition of data scientist might encompass, but the definition of statistician seems limiting, since a lot of statistical work is explicitly about inferences to be made from very small samples—this been true for hundreds of years, really. In fact statisticians deal with data however it arrives—big or small.

The statistics profession is caught at a confusing moment: the activities that preoccupied it over centuries are now in the limelight, but those activities are claimed to be bright shiny new, and carried out by (although not actually invented by) upstarts and strangers. Various professional statistics organizations are reacting:

- *Aren't* **we** *Data Science?*
  Column of ASA President Marie Davidian in AmStat News, July 2013[7]
- *A grand debate: is data science just a "rebranding" of statistics?*
  Martin Goodson, co-organizer of the Royal Statistical Society meeting May 19, 2015, on the relation of statistics and data science, in internet postings promoting that event.
- *Let* **us** *own Data Science.*
  IMS Presidential address of Bin Yu, reprinted in IMS bulletin October 2014[8]

One does not need to look far to find blogs capitalizing on the befuddlement about this new state of affairs:

- *Why Do We Need Data Science When We've Had Statistics for Centuries?*
  Irving Wladawsky-Berger
  Wall Street Journal, CIO report, May 2, 2014
- *Data Science* **is** *statistics.*
  When physicists do mathematics, they don't say they're doing number science. They're doing math. If you're analyzing data, you're doing statistics. You can call it data science or informatics or analytics or whatever, but it's still statistics. …You may not like what some statisticians do. You may feel they don't share your values. They may embarrass you. But that shouldn't lead us to abandon the term "statistics."
  Karl Broman, Univ. Wisconsin[9]

[3] Although, as the next footnote shows, this perception is based on a limited information set.

[4] At the same time, the two largest groups of faculty participating in this initiative are from EECS and statistics. Many of the EECS faculty publish avidly in academic statistics journals—I can mention Al Hero himself, Raj Rao Nadakaduti and others. The underlying design of the initiative is very sound and relies on researchers with strong statistics skills. But that is all hidden under the hood.

[5] Several faculty at University of Michigan wrote to tell me more about their MIDAS initiative and pointed out that statistics was more important to MIDAS than it might seem. They pointed out that statistics faculty including Vijay Nair were heavily involved in the planning of MIDAS—although not in its current public face—and that the nonstatistics department academics at the inaugural symposium used statistics heavily. This is actually the same point I am making.

[6] *http://www.datascienceassn.org/code-of-conduct.html*

[7] *http://magazine.amstat.org/blog/2013/07/01/datascience/*

[8] *http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/*

[9] *https://kbroman.wordpress.com/2013/04/05/data-science-is-statistics/*

On the other hand, we can find *provocateurs* declaiming the (near-) irrelevance of statistics:

- *Data Science without statistics is possible, even desirable.* Vincent Granville, at the Data Science Central Blog[10]
- *Statistics is the least important part of data science.* Andrew Gelman, Columbia University[11]

Clearly, there are many visions of data science and its relation to statistics. In my discussions with others, I have come across certain recurring "memes." I now deal with the main ones in turn.

2.1. The "Big Data" Meme

Consider the press release announcing the University of Michigan Data Science Initiative with which this article began. The University of Michigan President, Mark Schlissel, uses the term "big data" repeatedly, touting its importance for all fields and asserting the necessity of data science for handling such data. Examples of this tendency are near-ubiquitous.

We can immediately reject "big data" as a criterion for meaningful distinction between statistics and data science.[12]

- *History.* The very term "statistics" was coined at the beginning of modern efforts to compile census data, that is, comprehensive data about all inhabitants of a country, for example, France or the United States. Census data are roughly the scale of today's big data; but they have been around more than 200 years! A statistician, Hollerith, invented the first major advance in big data: the punched card reader to allow efficient compilation of an exhaustive U.S. census.[13] This advance led to formation of the IBM corporation which eventually became a force pushing computing and data to ever larger scales. Statisticians have been comfortable with large datasets for a long time, and have been holding conferences gathering together experts in "large datasets" for several decades, even as the definition of *large* was ever expanding.[14]
- *Science.* Mathematical statistics researchers have pursued the scientific understanding of big datasets for decades. They have focused on what happens when a database has a large number of individuals or a large number of measurements or both. It is simply wrong to imagine that they are not thinking about such things, in force, and obsessively. Among the core discoveries of statistics as a field were sampling and sufficiency, which allow to deal with very large datasets extremely efficiently. These ideas were discovered precisely because statisticians care about big datasets.

The data-science = "big data" framework is not getting at anything very intrinsic about the respective *fields*.[15]

2.2. The "Skills" Meme

In conversations I have witnessed,[16] computer scientists seem to have settled on the following talking points:

(a) *data science is concerned with really big data, which traditional computing resources could not accommodate*
(b) *data science trainees have the skills needed to cope with such big datasets.*

This argument doubles down on the "big data" meme,[17] by layering a "big data skills meme" on top.

What are those skills? In the early 2010s many would cite mastery of Hadoop, a variant of Map/Reduce for use with datasets distributed across a cluster of computers. Consult the standard reference *Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale, 4th Edition* by Tom White. There we learn at great length how to partition a single abstract dataset across a large number of processors. Then we learn how to compute the maximum of all the numbers in a single column of this massive dataset. This involves computing the maximum over the sub-database located in each processor, followed by combining the individual per-processor-maxima across all the many processors to obtain an overall maximum. Although the functional being computed in this example is dead-simple, quite a few skills are needed to implement the example at scale.

Lost in the hoopla about such skills is the embarrassing fact that once upon a time, one could do such computing tasks, and even much more ambitious ones, much more easily than in this fancy new setting! A dataset could fit on a single processor, and the global maximum of the array "x" could be computed with the six-character code fragment "`max(x)`" in, say, Matlab or R. More ambitious tasks, like large-scale optimization of a convex function, were easy to set up and use. In those less-hyped times, the skills being touted today were unnecessary. Instead, scientists developed skills to solve the problem they were really interested in, using elegant mathematics and powerful quantitative programming environments modeled on that math. Those environments were the result of 50 or more years of continual refinement, moving ever closer toward the ideal of enabling immediate translation of clear abstract thinking to computational results.

The *new* skills attracting so much media attention are not skills for better solving the *real* problem of inference from data; they are coping skills for dealing with organizational artifacts of large-scale cluster computing. The new skills cope with severe new constraints on algorithms posed by the multiprocessor/networked world. In this highly constrained world, the range of easily constructible algorithms shrinks dramatically compared to the single-processor model, so one inevitably tends to adopt inferential approaches which would have been considered rudimentary or even inappropriate in olden times. Such coping consumes our time and energy, deforms our judgements

[10] *http://www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable*
[11] *http://andrewgelman.com/2013/11/14/statistics-least-important-part-data-science/*
[12] One sometimes encounters also the statement that statistics is about "small datasets, while data science is about big datasets." Older statistics textbooks often did use quite small datasets to allow students to make hand calculations.
[13] *http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/*
[14] During the Centennial workshop, one participant pointed out that John Tukey's definition of "big data" was: "anything that won't fit on one device." In John's day the device was a tape drive, but the larger point is true today, where device now means "commodity file server."
[15] It may be getting at something real about the master's degree programs, or about the research activities of individuals who will be hired under the new spate of DSI's.
[16] For example, at breakouts of the NSF sponsored workshop *Theoretical Foundations of Data Science*, April 2016.
[17] …which we just dismissed!

about what is appropriate, and holds us back from data analysis strategies that we would otherwise eagerly pursue.

Nevertheless, the scaling cheerleaders are yelling at the top of their lungs that using more data deserves a big shout.

2.3. The "Jobs" Meme

Big data enthusiasm feeds off the notable successes scored in the last decade by brand-name global Information technology (IT) enterprises, such as Google and Amazon, successes currently recognized by investors and CEOs. A hiring "bump" has ensued over the last 5 years, in which engineers with skills in both databases and statistics were in heavy demand. In *The Culture of Big Data* (Barlow 2013), Mike Barlow summarizes the situation

> *According to Gartner, 4.4 million big data jobs will be created by 2014 and only a third of them will be filled. Gartner's prediction evokes images of "gold rush" for big data talent, with legions of hardcore quants converting their advanced degrees into lucrative employment deals.*

While Barlow suggests that *any* advanced quantitative degree will be sufficient in this environment, today's Data Science initiatives *per se* imply that traditional statistics degrees are not enough to land jobs in this area—*formal emphasis* on computing and database skills must be part of the mix.[18]

We do not really know. The booklet "Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work" (Harris, Murphy, and Vaisman 2013) points out that

> *Despite the excitement around "data science," "big data," and "analytics," the ambiguity of these terms has led to poor communication between data scientists and those who seek their help.*

Yanir Seroussi's blog[19] opines that *"there are few true data science positions for people with no work experience."*

> A successful data scientist needs to be able to become one with the data by exploring it and applying rigorous statistical analysis …But good data scientists also understand what it takes to deploy production systems, and are ready to get their hands dirty by writing code that cleans up the data or performs core system functionality …Gaining all these skills takes time [on the job].

Barlow implies that would-be data scientists may face *years* of further skills development post masters degree, before they can add value to their employer's organization. In an *existing* big-data organization, the infrastructure of production data processing is already set in stone. The databases, software, and workflow management taught in a given data science masters program are unlikely to be the same as those used by one specific employer. Various compromises and constraints were settled upon by the hiring organizations and for a new hire, contributing to those organizations is about learning how to cope with those constraints and still accomplish something.

Data science degree programs do not actually know how to satisfy the supposedly voracious demand for graduates. As we show below, the special contribution of a data science degree over a statistics degree is additional information technology training. Yet hiring organizations face difficulties making use of the specific information technology skills being taught in degree programs. In contrast, data analysis and statistics are broadly applicable skills that are portable from organization to organization.

2.4. What Here is Real?

We have seen that today's popular media tropes about data science do not withstand even basic scrutiny. This is quite understandable: writers and administrators are *shocked out of their wits*. Everyone believes we are facing a zeroth order discontinuity in human affairs.

If you studied a tourist guidebook in 2010, you would have been told that life in villages in India (say) had not changed in thousands of years. If you went into those villages in 2015, you would see that many individuals there now have mobile phones and some have smartphones. This is of course the leading edge fundamental change. Soon, eight billion people will be connected to the network, and will therefore be data sources, generating a vast array of data about their activities and preferences.

The transition to universal connectivity is very striking; it will, indeed, generate vast amounts of commercial data. Exploiting that data is certain to be a major preoccupation of commercial life in coming decades.

2.5. A Better Framework

However, a science does not just spring into existence simply because a deluge of data will soon be filling telecom servers, and because some administrators think they can sense the resulting trends in hiring and government funding.

Fortunately, there *is* a solid case for *some entity* called "data science" to be created, which would be a true science: facing essential questions of a lasting nature and using scientifically rigorous techniques to attack those questions.

Insightful statisticians have for at least 50 years been laying the groundwork for constructing that would-be entity as an enlargement of traditional academic statistics. This would-be notion of data science is not the same as the data science being touted today, although there is significant overlap. The would-be notion responds to a different set of urgent trends—intellectual rather than commercial. Facing the intellectual trends needs many of the same skills as facing the commercial ones and seems just as likely to match future student training demand and future research funding trends.

The would-be notion takes data science as the science of learning from data, with all that this entails. It is matched to the most important developments in science which will arise over the coming 50 years. As scientific publication itself becomes a body of data that we can analyze and study,[20] there are staggeringly large opportunities for improving the accuracy and validity of science, through the scientific study of the data analysis that scientists have been doing.

[18] Of course statistics degrees require extensive use of computers, but often omit training in formal software development and formal database theory.

[19] *http://yanirseroussi.com/2014/10/23/what-is-data-science/*

[20] Farther below, we will use shortened formulations such as "science itself becomes a body of data."

Understanding these issues gives Deans and Presidents an opportunity to rechannel the energy and enthusiasm behind today's data science movement toward lasting, excellent programs canonicalizing a new scientific discipline.

In this article, I organize insights that have been published over the years about this new would-be field of data science, and put forward a framework for understanding its basic questions and procedures. This framework has implications both for teaching the subject and for doing scientific research about how data science is done and might be improved.

3. The Future of Data Analysis, 1962

This article was prepared as an *aide-memoire* for a presentation made at the John Tukey centennial. More than 50 years ago, John prophesied that something like today's data science moment would be coming. In "The Future of Data Analysis" (Tukey 1962), John deeply shocked his readers (academic statisticians) with the following introductory paragraphs:[21]

> *For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. …All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data*

John's article was published in 1962 in *The Annals of Mathematical Statistics*, the central venue for mathematically advanced statistical research of the day. Other articles appearing in that journal at the time were mathematically precise and would present definitions, theorems, and proofs. John's article was instead a kind of public confession, explaining why he thought such research was too narrowly focused, possibly useless or harmful, and the research scope of statistics needed to be dramatically enlarged and redirected.

Peter Huber, whose scientific breakthroughs in robust estimation would soon appear in the same journal, recently commented about FoDA:

> *Half a century ago, Tukey, in an ultimately enormously influential paper redefined our subject …[The paper] introduced the term "data analysis" as a name for what applied statisticians do, differentiating this term from formal statistical inference. But actually, as Tukey admitted, he "stretched the term beyond its philology" to such an extent that it comprised all of statistics.*
> *Peter Huber (2010)*

So Tukey's vision embedded statistics in a larger entity. Tukey's central claim was that this new entity, which he called "data analysis," was a new *science*, rather than a branch of mathematics:

> *There are diverse views as to what makes a science, but three constituents will be judged essential by most, viz:*
>
> (a1) *intellectual content,*
> (a2) *organization in an understandable form,*
> (a3) *reliance upon the test of experience as the ultimate standard of validity.*
>
> *By these tests mathematics is not a science, since its ultimate standard of validity is an agreed-upon sort of logical consistency and provability. As I see it, data analysis passes all three tests, and I would regard it as a science, one defined by a ubiquitous problem rather than by a concrete subject. Data analysis and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics, …*
> *These points are meant to be taken seriously.*

Tukey identified four driving forces in the new science:

> *Four major influences act on data analysis today:*
> 1. *The formal theories of statistics*
> 2. *Accelerating developments in computers and display devices*
> 3. *The challenge, in many fields, of more and ever larger bodies of data*
> 4. *The emphasis on quantification in an ever wider variety of disciplines*

John's 1962 list is surprisingly modern, and encompasses all the factors cited today in press releases touting today's data science initiatives. Shocking at the time was Item 1, implying that statistical theory was only a (fractional!) part of the new science.

This new science is compared to established sciences and further circumscribed the role of statistics within it :

> *…data analysis is a very difficult field. It must adapt itself to what people can and need to do with data. In the sense that biology is more complex than physics, and the behavioral sciences are more complex than either, it is likely that the general problems of data analysis are more complex than those of all three. It is too much to ask for close and effective guidance for data analysis from any highly formalized structure, either now or in the near future.*
> *Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose.*

So not only is data analysis a scientific field, it is as complex as any major field of science! And theoretical statistics can only play a partial role in its progress.

Mosteller and Tukey's (1968) title reiterated this point: "Data Analysis, including Statistics" (Mosteller and Tukey 1968).

4. The 50 Years Since FoDA

While Tukey called for a much broader field of statistics, it could not develop overnight—even in one individual's scientific oeuvre.

P. J. Huber wrote that *"The influence of Tukey's paper was not immediately recognized …it took several years until I assimilated its import …"* (Huber 2010). From observing Peter first-hand I would say that 15 years after FoDA he was visibly comfortable with its lessons. At the same time, full evidence of this effect in Huber's case came even much later—see his 2010 book *Data Analysis: What can be learned from the last 50 years*, which summarizes Peter's writings since the 1980s and appeared 48 years after FoDA!

[21] One questions why the journal even allowed this to be published! Partly one must remember that John was a Professor of Mathematics at Princeton, which gave him plenty of authority! Sir Martin Rees, the famous astronomer/cosmologist once quipped that "God invented space just so not everything would happen at Princeton." JL Hodges Jr. of UC Berkeley was incoming editor of *Annals of Mathematical Statistics*, and deserves credit for publishing such a visionary but deeply controversial article.

4.1. Exhortations

While Huber obviously made the choice to explore the vistas offered in Tukey's vision, academic statistics as a whole did not.[22] John Tukey's Bell Labs colleagues, not housed in academic statistics departments, more easily adopted John's vision of a field larger than what academic statistics could deliver.

John Chambers, co-developer of the S language for statistics and data analysis while at Bell Labs, published already in 1993 the essay (Chambers 1993), provocatively titled "Greater or Lesser Statistics, A Choice for Future Research." His abstract pulled no punches:

> *The statistics profession faces a choice in its future research between continuing concentration on traditional topics—based largely on data analysis supported by mathematical statistics—and a broader viewpoint—based on an inclusive concept of learning from data.*
> *The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal…*

A call to action, from a statistician who feels "the train is leaving the station." Like Tukey's article, it proposes that we could be pursuing research spanning a much larger domain than the Statistical research we do today; such research would focus on opportunities provided by new types of data and new types of presentation. Chambers stated explicitly that the enlarged field would be *larger even than data analysis*. Specifically, it is larger than Tukey's 1962 vision.

C. F. Jeff Wu, upon his inauguration as Carver Professor of Statistics at University of Michigan, presented an inaugural lecture titled *Statistics = Data Science?* in which he advocated that statistics be renamed data science and statisticians data scientists. Anticipating modern masters' data science masters courses, he even mentioned the idea of a new masters' degree in which about half of the courses were outside the department of statistics. He characterized statistical work as a trilogy of data collection, data modeling and analysis, and decision making. No formal written article was prepared though the slides he presented are available.[23]

William S. Cleveland developed many valuable statistical methods and data displays while at Bell Labs, and served as a co-editor of Tukey's collected works. His 2001 article (Cleveland 2001), titled *Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics* addressed academic statistics departments and proposed a plan to reorient their work. His abstract read:

> *An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.*

In the article's introduction, Cleveland writes that[24,25]

> *…[results in] data science should be judged by the extent to which they enable the analyst to learn from data… Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit.*

Cleveland proposed six foci of activity, even suggesting allocations of effort.

(*) *Multidisciplinary investigations (25%)*
(*) *Models and Methods for Data (20%)*
(*) *Computing with Data (15%)*
(*) *Pedagogy (15%)*
(*) *Tool Evaluation (5%)*
(*) *Theory (20%)*

Several academic statistics departments that I know well could, at the time of Cleveland's publication, fit 100% of their activity into the 20% Cleveland allowed for theory. Cleveland's article was republished in 2014. I cannot think of an academic department that devotes today 15% of its effort on pedagogy, or 15% on computing with data. I can think of several academic statistics departments that continue to fit essentially all their activity into the last category, theory.

In short, *academic Statisticians were exhorted repeatedly across the years, by John Tukey and by some of his Bell Labs colleagues, and even by some academics like Peter Huber and Jeff Wu, to change paths, towards a much broader definition of their field.* Such exhortations had relatively little apparent effect before 2000.

4.2. Reification

One obstacle facing the earliest exhortations was that many of the exhortees could not see what the fuss was all about. Making the activity labeled "data analysis" more concrete and visible was ultimately spurred by code, not words.

Over the last 50 years, many statisticians and data analysts took part in the invention and development of computational environments for data analysis. Such environments included the early statistical packages BMDP, SPSS, SAS, and Minitab, all of which had roots in the mainframe computing of the late 1960s, and more recently packages such as S, ISP, STATA, and R, with roots in the minicomputer/personal computer era. This was an enormous effort carried out by many talented individuals—too many to credit here properly.[26]

To quantify the importance of these packages, try using Google's N-grams viewer[27] to plot the frequency of the words SPSS, SAS, Minitab, in books in the English language from 1970 to 2000; and for comparison, plot also the frequency of the bigrams "data analysis" and "statistical analysis." It turns out that SAS and SPSS are both more common terms in the English

[22] If evidence were needed, the reader might consider course offerings in academic statistics departments 1970–2000. In my recollection, the fraction of course offerings recognizably adopting the direction advocated by Tukey was small in those years. There was a bit more about plotting and looking at data.

[23] http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

[24] This echoes statements that John Tukey also made in FoDA, as I am sure Bill Cleveland would be proud to acknowledge.

[25] Geophysicists make a distinction between mathematical geophysicists who "care about the earth" and those who "care about math." Probably biologists make the same distinction in quantitative biology. Here Cleveland is introducing it as a litmus test restatistical theorists: do they "care about the data analyst" or do they not?

[26] One can illustrate the intensity of development activity by pointing to several examples strictly relevant to the Tukey Centennial at Princeton. I used three "statistics packages" while a Princeton undergraduate. P-STAT was an SPSS-like mainframe package which I used on Princeton's IBM 360/91 Mainframe; ISP was a UNIX minicomputer package on which I worked as a co-developer for the Princeton Statistics Department; and my teacher Don McNeil had developed software for a book of his own on exploratory data analysis; this ultimately became SPIDA after he moved to Macquarie University.

[27] http://preview.tinyurl.com/ycawv9xy

language over this period than either "data analysis" or "statistical analysis"—about twice as common, in fact.

John Chambers and his colleague Rick Becker at Bell Labs developed the quantitative computing environment "S" starting in the mid-1970s; it provided a language for describing computations, and many basic statistical and visualization tools. In the 1990s, Gentleman and Ihaka created the work-alike R system, as an open source project which spread rapidly. R is today the dominant quantitative programming environment used in academic statistics, with a very impressive online following.

Quantitative programming environments run "scripts," which codify precisely the steps of a computation, describing them at a much higher and more abstract level than in traditional computer languages like C++. Such scripts are often today called *workflows*. When a given QPE becomes dominant in some research community, as R has become in academic statistics,[28] workflows can be widely shared within the community and reexecuted, either on the original data (if it were also shared) or on new data. This is a game changer. What was previously somewhat nebulous—say the prose description of some data analysis in a scientific article—becomes instead tangible and useful, as one can download and execute code immediately. One can also easily tweak scripts, to reflect nuances of one's data, for example, changing a standard covariance matrix estimator in the original script to a robust covariance matrix estimator. One can document performance improvements caused by making changes to a baseline script. It now makes sense to speak of a scientific approach to improving a data analysis, by performance measurement followed by script tweaking. Tukey's claim that the study of data analysis could be a science now becomes self-evident. One might agree or disagree with Chambers and Cleveland's calls to action; but everyone could agree with Cleveland by 2001 that there *could* be such a field as "data science."

5. Breiman's *"Two Cultures,"* 2001

Leo Breiman, a UC Berkeley statistician who reentered academia after years as a statistical consultant to a range of organizations, including the Environmental Protection Agency, brought an important new thread into the discussion with his article in *Statistical Science* (Breiman 2001). Titled "Statistical Modeling: The Two Cultures," Breiman described two cultural outlooks about extracting value from data.

> *Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side, and on the other side the response variables y come out. Inside the black box, nature functions to associate the predictor variables with the response variables …*
> *There are two goals in analyzing the data:*
> - *Prediction. To be able to predict what the responses are going to be to future input variables;*
> - *[Inference].[29] To [infer] how nature is associating the response variables to the input variables.*

Breiman says that users of data split into two cultures, based on their primary allegiance to one or the other of these goals.

The "generative modeling"[30] culture seeks to develop stochastic models which fit the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit in their viewpoint is the notion that there is a true model generating the data, and often a truly "best" way to analyze the data. Breiman thought that this culture encompassed 98% of all academic statisticians.

The "predictive modeling" culture[31] prioritizes *prediction* and is estimated by Breiman to encompass 2% of academic statisticians—including Breiman—but also many computer scientists and, as the discussion of his article shows, important industrial statisticians. Predictive modeling is effectively silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. The relatively recent discipline of machine learning, often sitting within computer science departments, is identified by Breiman as the epicenter of the predictive modeling culture.

Breiman's abstract says, in part

> *The statistical community has been committed to the almost exclusive use of [generative] models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. [Predictive] modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller datasets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on [generative] models …*

Again, the statistics discipline is called to enlarge its scope.

In the discussion to Breiman's article, esteemed statisticians Sir David Cox of Oxford and Bradley Efron of Stanford both objected in various ways to the emphasis that Breiman was making.

- Cox states that in his view, *"predictive success …is not the primary basis for model choice"* and that *"formal methods of model choice that take no account of the broader objectives are suspect …"*.
- Efron stated that "*Prediction is certainly an interesting subject but Leo's article overstates both its role and our profession's lack of interest in it.*"

In the same discussion, Bruce Hoadley—a statistician for credit-scoring company Fair, Isaac—engages enthusiastically with Breiman's comments[32]:

> *Professor Breiman's paper is an important one for statisticians to read. He and Statistical Science should be applauded … His conclusions are consistent with how statistics is often practiced in business.*

Fair, Isaac's core business is to support the billions of credit card transactions daily by issuing in real time (what amount to) predictions that a requested transaction will or will not be repaid. Fair, Isaac not only create predictive models but must use them to provide their core business and they must justify their

[28] or Matlab in signal processing

[29] I changed Breiman's words here slightly; the original has "information" in place of [inference] and "extract some information about" in place of [infer]

[30] Breiman called this "data modeling," but "generative modeling" brings to the fore the key assumption: that a stochastic model could actually generate such data. So we again change Breiman's terminology slightly.

[31] Breiman used "algorithmic" rather than "predictive"

[32] Hoadley worked previously at ATT Bell Labs (Thanks to Ron Kennett for pointing this out).

accuracy to banks, credit card companies, and regulatory bodies. The relevance of Breiman's predictive culture to their business is clear and direct.

6. The Predictive Culture's Secret Sauce

Breiman was right to exhort statisticians to better understand the predictive modeling culture, but his article did not clearly reveal the culture's "secret sauce."

6.1. The Common Task Framework

To my mind, the crucial but unappreciated methodology driving predictive modeling's success is what computational linguist Mark Liberman (Liberman 2010) has called the *Common Task Framework* (CTF). An instance of the CTF has these ingredients:

(a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.
(b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data.
(c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule.

All the competitors share the *common task* of training a prediction rule which will receive a good score; hence the phase *common task framework*.

A famous recent example is the Netflix Challenge, where the common task was to predict Netflix user movie selections. The winning team (which included ATT Statistician Bob Bell) won $1M. The dataset used proprietary customer history data from Netflix. However, there are many other examples, often with much greater rewards (implicitly) at stake.

6.2. Experience with CTF

The genesis of the CTF paradigm has an interesting connection to our story. In Mark Liberman's telling it starts with J.R. Pierce, a colleague of Tukey's at Bell Labs. Pierce had invented the word "transistor" and supervised the development of the first communication satellite, and served on the Presidential Science Advisory Committee with Tukey in the early/mid 1960s. At the same time that Tukey was evaluating emerging problems caused by over-use of pesticides, Pierce was asked to evaluate the already extensive investment in machine translation research. In the same way that Tukey did not like much of what he saw passing as statistics research in the 1960s, Pierce did not like much of what he saw passing as 1960s machine translation research.

Now we follow Mark Liberman closely.[33] Judging that the field was riddled with susceptibility to "glamor and deceit," Pierce managed to cripple the whole U.S. machine translation research effort—sending it essentially to zero for decades.

As examples of glamor and deceit, Pierce referred to theoretical approaches to translation deriving from, for example, Chomsky's so-called theories of language; while many language researchers at the time apparently were in awe of the charisma carried by such theories, Pierce saw those researchers as being deceived by the glamor of (a would-be) theory, rather than actual performance in translation.

Machine Translation research finally reemerged decades later from the Piercian limbo, but only because it found a way to avoid a susceptibility to Pierce's accusations of glamor and deceit. A research team in speech and natural language processing at IBM, which included true geniuses like John Cocke, as well as data scientists *avant la lettre* Lalit Bahl, Peter Brown, Stephen and Vincent Della Pietra, and Robert Mercer, began to make definite progress toward machine translation based on an early application of the common task framework. A key resource was data: they had obtained a digital copy of the so-called Canadian Hansards, a corpus of government documents which had been translated into both English and French. By the late 1980s, DARPA was convinced to adopt the CTF as a new paradigm for machine translation research. NIST was contracted to produce the sequestered data and conduct the refereeing, and DARPA challenged teams of researchers to produce rules that correctly classified under the CTF.

Variants of CTF have by now been applied by DARPA successfully in many problems: machine translation, speaker identification, fingerprint recognition, information retrieval, OCR, automatic target recognition, and on and on.

The general experience with CTF was summarized by Liberman as follows:

1. *Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality.*
2. *Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.*
3. *Shared data plays a crucial role—and is reused in unexpected ways.*

The ultimate success of many automatic processes that we now take for granted—Google translate, smartphone touch ID, smartphone voice recognition—derives from the CTF research paradigm, or more specifically its cumulative effect after operating for decades in specific fields. Most importantly for our story: *those fields where machine learning has scored successes are essentially those fields where CTF has been applied systematically*.

6.3. The Secret Sauce

It is no exaggeration to say that the combination of a predictive modeling culture together with CTF is the "secret sauce" of machine learning.

The synergy of minimizing prediction error with CTF is worth noting. This combination leads directly to a total focus on optimization of empirical performance, which as Mark Liberman has pointed out, allows large numbers of researchers to compete at any given common task challenge, and allows for efficient and unemotional judging of challenge winners. It also leads immediately to applications in a real-world application. In the process of winning a competition, a prediction rule has

[33] *https://www.simonsfoundation.org/lecture/reproducible-research-and-the-common-task-method/*

necessarily been tested, and so is essentially ready for immediate deployment.[34]

Many "outsiders" are not aware of the CTF's paradigmatic nature and its central role in many of machine learning's successes. Those outsiders may have heard of the Netflix challenge, without appreciating the role of CTF in that challenge. They may notice that "deep learning" has become a white hot topic in the high-tech media, without knowing that the buzz is due to successes of deep learning advocates in multiple CTF-compliant competitions.

Among the outsiders are apparently many mainstream academic statisticians who seem to have little appreciation for the power of CTF in generating progress, in technological field after field. I have no recollection of seeing CTF featured in a major conference presentation at a professional statistics conference or academic seminar at a major research university.

The author believes that the Common Task Framework is the single idea from machine learning and data science that is most lacking attention in today's statistical training.

6.4. Required Skills

The Common Task Framework imposes numerous demands on workers in a field:

- The workers must deliver predictive models which can be evaluated by the CTF scoring procedure in question. They must therefore personally submit to the information technology discipline imposed by the CTF developers.
- The workers might even need to implement a custom-made CTF for their problem; so they must both develop an information technology discipline for evaluation of scoring rules and they must obtain a dataset which can form the basis of the shared data resource at the heart of the CTF.

In short, information technology skills are at the heart of the qualifications needed to work in predictive modeling. These skills are analogous to the laboratory skills that a wet-lab scientist needs to carry out experiments. No math required.

The use of CTFs really took off at about the same time as the open source software movement began and as the ensuing arrival of quantitative programming environments dominating specific research communities. QPE dominance allowed researchers to conveniently share scripts across their communities, in particular scripts that implement either a baseline prediction model or a baseline scoring workflow. So the skills required to work within a CTF became very specific and very teachable—*can we download and productively tweak a set of scripts?*

7. Teaching of Today's Consensus Data Science

It may be revealing to look at what is taught in today's data science programs at some of the universities that have recently established them. Let us consider the attractive and informative web site for the UC Berkeley Data Science Masters' degree at *datascience.berkeley.edu*.

[34] However, in the case of the Netflix Challenge the winning algorithm was never implemented. *http://preview.tinyurl.com/ntwlyuu*

Reviewing the curriculum at *https://datascience.berkeley.edu/academics/curriculum/* we find five foundation courses

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

Only "Storing and Retrieving Data" seems manifestly not taught by traditional statistics departments; and careful study of the words reveals that the least traditional topic among the others, "Applied Machine Learning," seems to a statistician thinking about the actual topics covered, very much like what a statistics department might or should offer—however, the use of "machine learning" in the course title is a tip off that the approach may be heavily weighted toward predictive modeling rather than inference.

> Machine learning is a rapidly growing field at the intersection of computer science and statistics concerned with finding patterns in data. It is responsible for tremendous advances in technology, from personalized product recommendations to speech recognition in cell phones. This course provides a broad introduction to the key ideas in machine learning. The emphasis will be on intuition and practical examples rather than theoretical results, though some experience with probability, statistics, and linear algebra will be important.

The choice of topics might only give a partial idea of what takes place in the course. Under "Tools," we find an array of core information technology.

> Python libraries for linear algebra, plotting, machine learning: numpy, matplotlib, sk-learn / Github for submitting project code

In short, course participants are producing and submitting code. Code development is not yet considered utterly *de rigueur* for statistics teaching, and in many statistics courses would be accomplished using code in R or other quantitative programming environments, which is much "easier" for students to use for data analysis because practically the whole of modern data analysis is already programmed in. However, R has the reputation of being less scalable than Python to large problem sizes. In that sense, a person who does their work in Python might be considered to have worked harder and shown more persistence and focus than one who does the same work in R.

Such thoughts continue when we consider the advanced courses.

> 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.

The first two courses seem like mainstream statistics courses that could be taught by stat departments at any research university. The third is less familiar but overlaps with "Legal, Policy, and Ethical Considerations for Data Scientists" courses that have existed at research universities for quite a while.

The last two courses address the challenge of scaling up processes and procedures to really large data. These are courses that ordinarily would not be offered in a traditional statistics department.

Who are the faculty in the UC Berkeley data science program? Apparently not traditionally pedigreed academic statisticians. In the division of the website "About MIDS faculty" on Friday September 11, 2015, I could find mostly short bios for faculty associated with the largely nonstatistical courses (such as "Scaling Up! really Big Data" or "Machine Learning at Scale"). For the approximately 50% of courses covering traditional statistical topics, fewer bios were available, and those seemed to indicate different career paths than traditional statistics Ph.D.'s—sociology Ph.D.'s or information science Ph.D.'s. The program itself is run by the information school.[35]

In FoDA, Tukey argued that the teaching of statistics as a branch of mathematics was holding back data analysis. He saw apprenticeship with real data analysts and hence real data as the solution:

> *All sciences have much of art in their makeup. As well as teaching facts and well-established structures, all sciences must teach their apprentices how to think about things in the manner of that particular science, and what are its current beliefs and practices. Data analysis must do the same. Inevitably its task will be harder than that of most sciences. Physicists have usually undergone a long and concentrated exposure to those who are already masters of the field. Data analysts even if professional statisticians, will have had far less exposure to professional data analysts during their training. Three reasons for this hold today, and can at best be altered slowly:*
>
> (c1) *Statistics tends to be taught as part of mathematics.*
>
> (c2) *In learning statistics per se, there has been limited attention to data analysis.*
>
> (c3) *The number of years of intimate and vigorous contact with professionals is far less for statistics Ph.D.'s than for physics or mathematics Ph.D.'s*
>
> *Thus data analysis, and adhering statistics, faces an unusually difficult problem of communicating certain of its essentials, one which cannot presumably be met as well as in most fields by indirect discourse and working side by side.*

The Berkeley data science masters program features a capstone course, which involves a data analysis project with a large dataset. The course listing states in part that in the capstone class

> The final project …provides experience in formulating and carrying out a sustained, coherent, and significant course of work resulting in a tangible data science analysis project with real-world data ….The capstone is completed as a group/team project (3–4 students), and each project will focus on open, pre-existing secondary data.

This project seems to offer some of the "apprenticeship" opportunities that John Tukey knew from his college chemistry degree work, and considered important for data analysis.

Tukey insisted that mathematical rigor was of very limited value in teaching data analysis. This view was already evident in the quotation from FoDA immediately above. Elsewhere in FoDA Tukey said:

> *Teaching data analysis is not easy, and the time allowed is always far from sufficient. But these difficulties have been enhanced by the view that "avoidance of cookbookery and growth of understanding come only by mathematical treatment, with emphasis upon proofs."*
> *The problem of cookbookery is not peculiar to data analysis. But the solution of concentrating upon mathematics and proof is.*

Tukey saw data analysis as like other sciences and not like mathematics, in that there existed knowledge which needed to be related rather than theorems which needed proof. Drawing again on his chemistry background, he remarked that

> *The field of biochemistry today contains much more detailed knowledge than does the field of data analysis. The overall teaching problem is more difficult. Yet the textbooks strive to tell the facts in as much detail as they can find room for.*

He also suggested that experimental labs offered a way for students to learn statistics

> *These facts are a little complex, and may not prove infinitely easy to teach, but any class can check almost any one of them by doing its own experimental sampling.*

One speculates that John Tukey might have viewed the migration of students away from statistics courses and into equivalent data science courses as possibly not a bad thing.

In his article "Statistical Modeling: The Two Cultures," Leo Breiman argued that teaching stochastic model building and inference to the exclusion of predictive modeling was damaging the ability of statistics to attack the most interesting problems he saw on the horizon. The problems he mentioned at the time are among today's hot applications of data science. So Breiman might have welcomed teaching programs which reverse the balance between inference and prediction, that is, programs such as the UC Berkeley data science masters.

Although my heroes Tukey, Chambers, Cleveland, and Breiman would recognize positive features in these programs, it is difficult to say whether they would approve of their long-term direction—or if there is even a long-term direction to comment about. Consider this snarky definition:

> Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

This definition is grounded in fact. Data science masters' curricula are compromises: taking some material out of a statistics master's 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 offered?

8. The Full Scope of Data Science

John Chambers and Bill Cleveland each envisioned a would-be field that is considerably larger than the consensus data science master's we have been discussing but at the same time more intellectually productive and lasting.

The larger vision posits a professional on a quest to extract information from data—exactly as in the definitions of data science we saw earlier. The larger field cares about each and every step that the professional must take, from getting acquainted

[35] I do not wish to imply in the above that there is anything concerning to me about the composition of the faculty. I do wish to demonstrate that this is an opportunity being seized by nonstatisticians. An important event in the history of academic statistics was Hotelling's article "The Teaching of Statistics" (1940) (Hotelling 1940) which decried the teaching of statistics by nonmathematicians, and motivated the formation of academic statistics departments. The new developments may be undoing the many years of postwar professionalization of statistics instruction.

with the data all the way to delivering results based upon it, and extending even to that professional's continual review of the evidence about best practices of the whole field itself.

Following Chambers, let us call the collection of activities mentioned until now "lesser data science" (LDS) and the larger would-be field *greater data science* (GDS). Chambers and Cleveland each parsed out their enlarged subject into specific divisions/topics/subfields of activity. I find it helpful to merge, relabel, and generalize the two parsings they proposed. This section presents and then discusses this classification of GDS.

8.1. The Six Divisions

The activities of GDS are classified into six divisions:

1. *Data Gathering, Preparation, and Exploration*
2. *Data Representation and Transformation*
3. *Computing with Data*
4. *Data Modeling*
5. *Data Visualization and Presentation*
6. *Science about Data Science*

Let's go into some detail about each division.

GDS1: Data Gathering, Preparation, and Exploration. Some say that 80% of the effort devoted to data science is expended by *diving into* or *becoming one* with one's messy data to learn the basics of what's in them, so that data can be made ready for further exploitation. We identify three subactivities:

- *Gathering.* This includes traditional experimental design as practiced by statisticians for well over a century, but also a variety of modern data gathering techniques and data resources. Thus, Google nGrams viewer can quantify the entire corpus of literature 1500-2008, Google Trends can quantify recent web search interests of the whole population and even of localities, humans are taking 1 trillion photos a year, many of which are posted in social media;[36] billions of utterances are posted on social media.[37] We have new data-making technologies like next generation sequencing in computational biology, GPS location fixes, supermarket scanner data. Next gen skills can include web scraping, Pubmed scraping,[38] image processing, and Twitter, Facebook, and Reddit munging.
- *Preparation.* Many datasets contain anomalies and artifacts.[39] Any data-driven project requires mindfully identifying and addressing such issues. Responses range from reformatting and recoding the values themselves, to more ambitious preprocessing, such as grouping, smoothing, and subsetting. Often today, one speaks colorfully of *data cleaning* and *data wrangling*.
- *Exploration.* Since John Tukey's coining of the term "exploratory data analysis" (EDA), we all agree that every data scientist devotes serious time and effort to exploring data to sanity-check its most basic properties, and to expose unexpected features. Such detective work adds crucial insights to every data-driven endeavor.[40]

GDS2: Data Representation and Transformation. A data scientist works with many different data sources during a career. These assume a very wide range of formats, often idiosyncratic ones, and the data scientist has to easily adapt to them all. Current hardware and software constraints are part of the variety because access and processing may require careful deployment of distributed resources.

Data scientists very often find that a central step in their work is to implement an appropriate transformation restructuring the originally given data into a new and more revealing form.

Data scientists develop skills in two specific areas:

- *Modern Databases.* The scope of today's data representation includes everything from homely text files and spreadsheets to SQL and noSQL databases, distributed databases, and live data streams. Data scientists need to know the structures, transformations, and algorithms involved in using all these different representations.
- *Mathematical Representations.* These are interesting and useful mathematical structures for representing data of special types, including acoustic, image, sensor, and network data. For example, to get features with acoustic data, one often transforms to the cepstrum or the Fourier transform; for image and sensor data the wavelet transform or some other multi scale transform (e.g., pyramids in deep learning). Data scientists develop facility with such tools and mature judgment about deploying them.

GDS3: Computing with Data. Every data scientist should know and use *several* languages for data analysis and data processing. These can include popular languages like R and Python, but also specific languages for transforming and manipulating text, and for managing complex computational pipelines. It is not surprising to be involved in ambitious projects using a half dozen languages in concert.

Beyond basic knowledge of languages, data scientists need to keep current on new idioms for efficiently using those languages and need to understand the deeper issues associated with computational efficiency.

Cluster and cloud computing and the ability to run massive numbers of jobs on such clusters has become an overwhelmingly powerful ingredient of the modern computational landscape. To exploit this opportunity, data scientists develop workflows which organize work

[36] https://arxiv.org/abs/1706.01869
[37] https://arxiv.org/abs/1704.05579
[38] http://jamanetwork.com/journals/jama/fullarticle/2503172
[39] Peter Huber (2010) recalled the classic Coale and Stephan article on teenage widows (Coale and Stephan 1962). In this example, a frequent coding error in a census database resulted in excessively large counts of teenage widows—until the error was rooted out. This example is quaint by modern standards. If we process natural language in the wild, such blogs and tweets, anomalies are the order of the day.
[40] At the Tukey Centennial, Rafael Irizarry gave a convincing example of exploratory data analysis of GWAS data, studying how the data row mean varied with the date on which each row was collected, convince the *field* of gene expression analysis to face up to some data problems that were crippling their studies.

to be split up across many jobs to be run sequentially or else across many machines.

Data scientists also develop workflows that document the steps of an individual data analysis or research project.

Finally, data scientists develop packages that abstract commonly used pieces of workflow and make them available for use in future projects.

GDS4: Data Visualization and Presentation. Data visualization at one extreme overlaps with the very simple plots of EDA—histograms, scatterplots, time series plots—but in modern practice it can be taken to much more elaborate extremes. Data scientists often spend a great deal of time decorating simple plots with additional color or symbols to bring in an important new factor, and they often crystallize their understanding of a dataset by developing a new plot which codifies it. Data scientists also create dashboards for monitoring data processing pipelines that access streaming or widely distributed data. Finally, they develop visualizations to present conclusions from a modeling exercise or CTF challenge.

GDS5: Data Modeling. Each data scientist in practice uses tools and viewpoints from *both* of Leo Breiman's modeling cultures:

- *Generative modeling*, in which one proposes a stochastic model that could have generated the data, and derives methods to infer properties of the underlying generative mechanism. This roughly speaking coincides with traditional academic statistics and its offshoots.[41]
- *Predictive modeling*, in which one constructs methods which predict well over some given data universe—that is, some very specific concrete dataset. This roughly coincides with modern Machine Learning, and its industrial offshoots.[42]

GDS6: Science about Data Science. Tukey proposed that a "science of data analysis" exists and should be recognized as among the most complicated of all sciences. He advocated the study of what data analysts "in the wild" are actually doing, and reminded us that the true effectiveness of a tool is related to the probability of deployment times the probability of effective results once deployed.[43]

Data scientists are doing *science about data science* when they identify commonly occurring analysis/processing workflows, for example, using data about their frequency of occurrence in some scholarly or business domain; when they measure the effectiveness of standard workflows in terms of the human time, the computing resource, the analysis validity, or other performance metric, and when they uncover emergent phenomena in data analysis, for example, new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results.

The scope here also includes foundational work to make future such science possible—such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta-analysis.

As data analysis and predictive modeling becomes an ever more widely distributed global enterprise, "science about data science" will grow dramatically in significance.

8.2. Discussion

These six categories of activity, when fully scoped, cover a field of endeavor much larger than what current academic efforts teach or study.[44,45] Indeed, a single category—"GDS5: Data Modeling"—dominates the representation of data science in today's academic departments, either in statistics and mathematics departments through traditional statistics teaching and research, or in computer science departments through machine learning.

This parsing-out reflects various points we have been trying to make earlier:

- The wedge issue that computer scientists use to separate "data science" from "statistics" is acknowledged here, by the addition of both "GDS3: Computing with Data" and "GDS2: Data Representation" as major divisions alongside "GDS5: Data Modeling."[47,48]
- The tension between machine learning and academic statistics is suppressed in the above classification; much of it is irrelevant to what data scientists do on a daily basis. As

[41] It is striking how, when I review a presentation on today's data science, in which statistics is superficially given pretty short shrift, I cannot avoid noticing that the underlying tools, examples, and ideas which are being taught as data science were all literally invented by someone trained in Ph.D. statistics, and in many cases the actual software being used was developed by someone with an MA or Ph.D. in statistics. The accumulated efforts of statisticians over centuries are just too overwhelming to be papered over completely, and cannot be hidden in the teaching, research, and exercise of Data Science.

[42] Leo Breiman (2001) was correct in pointing out that academic statistics departments (at that time, and even since) have under-weighted the importance of the predictive culture in courses and hiring. It clearly needs additional emphasis.

[43] Data analysis per se is probably too narrow a term, because it misses all the automated data processing that goes on under the label of data science about which we can also make scientific studies of behavior "in the wild."

[44] John Chambers' 1993 vision of "greater statistics" proposed three divisions: data preparation, data modeling, and data presentation. We accommodated them here in "GDS1: Data Exploration and Preparation;" "GDS5: Data Modeling," and "GDS4: Data Visualization and Presentation," respectively.

[45] Cleveland's 2001 program for data science included several categories which can be mapped onto (subsets) of those proposed here, for example:

- Cleveland's categories "Theory" and "Stochastic Models and Statistical Methods" can be mapped into GDS either onto the "Generative Models" subset of "GDS5: Data Modeling" or onto "GDS5 Data Modeling" itself.
- His category "Computing with Data" maps onto a subset of GDS' category of the same name; the GDS category has expanded to cover developments such as Hadoop and AWS that were not yet visible in 2001.
- Cleveland's category "Tool Evaluation" can be mapped onto a subset of "GDS6: Science about Data Science"

Cleveland also allocated resources to multidisciplinary investigations and pedagogy. It seems to me that these can be mapped onto subsets of our categories. For example, pedagogy ought to be part of the science about data science—we can hope for evidence-based teaching.[46]

[47] In our opinion, the scaling problems though real are actually transient (because technology will trivialize them over time). The more important activity encompassed under these divisions are the many ambitious and even audacious efforts to reconceptualize the standard software stack of today's data science.

[48] Practically speaking, every statistician has to master database technology in the course of applied projects.

I say above, data scientists should use both generative and predictive modeling.

- The hoopla about distributed databases, Map/Reduce, and Hadoop is *not* evident in the above classification. Such tools are relevant for "GDS2: Data Representation" and "GDS3: Computing with Data" but although they are heavily cited right now, they are simply today's enablers of certain larger activities. Such activities will be around permanently, while the role for enablers like Hadoop will inevitably be streamlined away.
- Current masters programs in data science cover only a fraction of the territory mapped out here. Graduates of such programs would not have had sufficient exposure to exploring data, data cleaning, data wrangling, data transformation, science about data science, and other topics in GDS.

Other features of this inventory will emerge below.

8.3. Teaching of GDS

Full recognition of the scope of GDS would require covering each of its six branches. This demands major shifts in teaching.

"GDS5: Data Modeling" is the easy part of data science to formalize and teach; we have been doing this for generations in statistics courses; for a decade or more in machine learning courses; and this pattern continues in the data science masters programs being introduced all around us, where it consumes most of the coursework time allocation. However, this "easy stuff" covers only a fraction of the effort required in making productive use of data.

"GDS1: Data Gathering, Preparation, and Exploration" is more important than "GDS5: Data Modeling," as measured using time spent by practitioners. But there have been few efforts to formalize data exploration and cleaning and such topics still are neglected in teaching. Students who only analyze precooked data are not being given the chance to learn these essential skills.

How might teaching even address such a topic? I suggest the reader study carefully two books (together).

- *The Book* (Tango, Lichtman, and Dolphin 2007) analyzes a set of databases covering all aspects of the American game of major league baseball, including every game played in recent decades and every player who ever appeared in such games. This amazingly comprehensive work considers a near-exhaustive list of questions one might have about the quantitative performance of different baseball strategies, carefully describes how such questions can be answered using such a database, typically by a statistical two-sample test (or A/B test in internet marketing terminology).
- *Analyzing Baseball Data with R* (Marchi and Albert 2013) showed how to access the impressive wealth of available Baseball data using the internet and how to use R to insightfully analyze that data.

A student who could show how to systematically use the tools and methods taught in the second book to answer some of the interesting questions in the first book would, by my lights, have developed real expertise in the above division "GDS1: Data Gathering, Preparation, and Exploration." Similar projects can be developed for all the other "new" divisions of data science. In "GDS3: Computing with Data," one could teach students to develop and new R packages, and new data analysis workflows, in a hands-on manner.

Ben Baumer and co-authors review experiences in Horton, Baumer, and Wickham (2015) and Baumer (2015) teaching first and second courses in data science/statistics that are consistent with this approach.

The reader will worry that the large scope of GDS is much larger than what we are used to teaching. Tukey anticipated such objections, by pointing out that biochemistry textbooks seem to cover much more material than statistics textbooks; he thought that once the field commits to teaching more ambitiously, it can simply "pick up the pace."[49]

8.4. Research in GDS

Once we have the GDS template in mind, we can recognize that today there is all sorts of interesting—and highly impactful—"GDS research." Much of it does not have a natural "home," yet, but GDS provides a framework to organize it and make it accessible. We mention a few examples to stimulate the reader's thinking.

8.4.1. Quantitative Programming Environments: R

The general topic of "computing with data" may sound at first as if it is stretchable to cover lots of mainstream academic computer science; suggesting that perhaps there is no real difference between data science and computer science. To the contrary, "computing with data" has a distinct core, and an identity separate from academic computer science. The litmus test is whether the work centers on the need to analyze data.

We argued earlier that the R system transformed the practice of data analysis by creating a standard language which different analysts can all use to communicate and share algorithms and workflows. Becker and Chambers (with S) and later Ihaka, Gentleman, and members of the R Core team (with R) conceived of their work as *research* how to best organize computations with statistical data. I too classify this as research, addressing category "GDS 3: Computing with Data." Please note how essentially ambitious the effort was, and how impactful. In recently reviewing many online presentations about data science initiatives, I was floored to see how heavily R is relied upon, even by data science instructors who claim to be doing no statistics at all.

8.4.2. Data Wrangling: Tidy Data

Hadley Wickham is a well-known contributor to the world of statistical computing, as the author of numerous packages becoming popular with R users everywhere; these include `ggplot2`, `reshape2`, `plyr`, `tidyr`, `dplyr`; Wickham (2011), Wickham et al. (2007), and Wickham et al. (2011). These packages abstractify and attack certain common issues in data science subfield "GDS 2: Data Representation and Transformation" and also subfield "GDS 4: Data Visualization and Presentation," and Wickham's tools have gained acceptance as indispensable to many.

In Wickham (2014) Wickham discussed the notion of *tidy* data. Noting (as I also have, above) the common estimate that

[49]Tukey also felt that focusing on mathematical proof limited the amount of territory that could be covered in university teaching.

80% of data analysis is spent on the process of cleaning and preparing the data, Wickham develops a systematic way of thinking about "messy" data formats and introduces a set of tools in R that translate them to a universal "tidy" data format. He identifies several messy data formats that are commonly encountered in data analysis and shows how to transform each such format into a tidy format using his tools `melt` and `cast`. Once the data are molten, they can be very conveniently operated on using tools from the `plyr` library, and then the resulting output data can be "cast" into a final form for further use.

The `plyr` library abstracts certain iteration processes that are very common in data analysis, of the form "apply such-and-such a function to each element/column/row/slice" of an array. The general idea goes back to Kenneth Iverson's 1962 *APL 360* programming language (Iverson 1991), and the reduce operator formalized there; younger readers will have seen the use of derivative ideas in connection with Map/Reduce and Hadoop, which added the ingredient of applying functions on many processors in parallel. Still `plyr` offers a very fruitful abstraction for users of R, and in particular teaches R users quite a bit about the potential of R's specific way of implementing functions as closures within environments.

Wickham has not only developed an R package making tidy data tools available; he has written an article that teaches the R user about the potential of this way of operating. This effort may have more impact on today's practice of data analysis than many highly regarded theoretical statistics articles.

8.4.3. Research Presentation: Knitr

As a third vignette, we mention Yihui Xie's work on the `knitr` package in R. This helps data analysts authoring source documents that blend running R code together with text, and then compiling those documents by running the R code, extracting results from the live computation, and inserting them in a high-quality PDF file, HTML web page, or other output product.

In effect, the entire workflow of a data analysis is intertwined with the interpretation of the results, saving a huge amount of error-prone manual cut-and-paste moving computational outputs and their place in the document.

Since data analysis typically involves presentation of conclusions, there is no doubt that data science activities, in the larger sense of GDS, include preparation of reports and presentations. Research that improves those reports and presentations in some fundamental way is certainly contributing to GDS. In this case, we can view it as part of "GDS3: Computing with Data," because one is capturing the workflow of an analysis. As we show later, it also enables important research in "GDS6: Science about Data Science."

8.5. Discussion

One can multiply the above examples, making GDS research ever more concrete. Two quick hits:

- For subfield "GDS 4: Data Visualization and Presentation," one can mention several exemplary research contributions: Bill Cleveland's work on statistical graphics (Cleveland et al. 1985; Cleveland 2013), along with Leland Wilkinson's (Wilkinson 2006) and Hadley Wickham's Wickham (2011) books on the Grammar of Graphics.
- For subfield "GDS 1: Data Exploration and Presentation," there is of course the original research from long ago of John Tukey on EDA (Tukey 1977); more recently Cook and Swayne's work on Dynamic graphics (Cook and Swayne 2007).

Our main points about all the above-mentioned research:

(a) it is not traditional research in the sense of mathematical statistics or even machine learning;
(b) it has proven to be very impactful on practicing data scientists;
(c) lots more such research can and should be done.

Without a classification like GDS, it would be hard to know where to "put it all" or whether a given data science program is adequately furnished for scholar/researchers across the full spectrum of the field.

9. Science About Data Science

A broad collection of technical activities is not a science; it could simply be a trade such as cooking or a technical field such as geotechnical engineering. To be entitled to use the word "science," we must have a continually evolving, evidence-based approach. "GDS6: Science about Data Science" posits such an approach; we briefly review some work showing that we can really have evidence-based data analysis. We also in each instance point to the essential role of information technology skills, the extent to which the work "looks like data science," and the professional background of the researchers involved.

9.1. Science-Wide Meta-Analysis

In FoDA,[50] Tukey proposed that statisticians should study how people analyze data today.

By formalizing the notion of multiple comparisons (Tukey 1994), Tukey put in play the idea that a whole body of analysis conclusions can be evaluated statistically.

Combining such ideas leads soon enough to meta-analysis, where we study all the data analyses being published on a given topic.[51] In 1953, the introduction to Tukey's article (Tukey 1994) considered a very small scale example with six different comparisons under study. Today, more than one million scientific articles are published annually, just in clinical medical research, and there are many repeat studies of the same intervention. There's plenty of data analysis out there to meta-study!

In the last 10 years, the scope of such meta-analysis has advanced spectacularly; we now perceive entire scientific literature as a body of text to be harvested, processed, and "scraped" clean of its embedded numerical data. Those data are analyzed

[50] *"I once suggested, in discussion at a statistical meeting, that it might be well if statisticians looked to see how data was actually analyzed by many sorts of people. A very eminent and senior statistician rose at once to say that this was a novel idea, that it might have merit, but that young statisticians should be careful not to indulge in it too much since it might distort their ideas," Tukey, FoDA*

[51] The practice of meta-analysis goes back at least to Karl Pearson. I am not trying to suggest that Tukey originated meta-analysis; only reminding the reader of John's work for the centennial occasion.

for clues about meta-problems in the way all of science is analyzing data. I can cite a few articles by John Ioannidis and co-authors (Ioannidis 2005, 2008; Pan et al. 2005; Button et al. 2013) and for statisticians the article "An estimate of the science-wise false discovery rate …" Jager and Leek (2014) together with *all* its ensuing discussion.

In particular, meta-analysts have learned that a dismaying fraction of the conclusions in the scientific literature are simply incorrect (i.e., far more than 5%) and that most published effects sizes are overstated, that many results are not reproducible, and so on.

Our government spends tens of billions of dollars every year to produce more than one million scientific articles. It approaches cosmic importance, to learn whether science as actually practiced is succeeding or even how science as a whole can improve.

Much of this research occurred in the broader applied statistics community, for example, taking place in schools of education, medicine, public health, and so on. Much of the so far already staggering achievement depends on "text processing," namely, scraping data from abstracts posted in online databases, or stripping it out of PDF files and so on. In the process we build up "big data," for example, Ioannidis and collaborators recently harvested all the p-values embedded in all Pubmed abstracts (Chavalarias et al. 2016). Participants in this field are doing data science, and their goal is to answer fundamental questions about the scientific method as practiced today.

9.2. Cross-Study Analysis

Because medical research is so extensive, and the stakes are so high, there often are multiple studies of the same basic clinical intervention, each analyzed by some specific team in that specific team's manner. Different teams produce different predictions of patient outcome and different claims of performance of their predictors. Which if any of the predictors actually work?

Giovanni Parmigiani at Harvard School of Public Health explained to me a cross-study validation exercise (Bernau et al. 2014), in which he and co-authors considered an ensemble of studies that develop methods for predicting survival of ovarian cancer from gene expression measurements. From 23 studies of ovarian cancer with publicly available data, they created a combined curated dataset included gene expression data and survival data, involving 10 datasets with 1251 patients in all. From 101 candidate papers in the literature they identified 14 different prognostic models for predicting patient outcome. These were formulas for predicting survival from observed gene expression; the formulas had been fit to individual study datasets by their original analysts, and in some cases validated against fresh datasets collected by other studies.

Parmigiani and colleagues considered the following cross-study validation procedure: fit each of the 14 models to one of the 10 datasets, and then validate it on every one of the remaining datasets, measure the concordance of predicted risk with actual death order, producing a 14 by 10 matrix allowing to study the individual models across datasets, and also allowing to study individual datasets across models.

Surprising cross-study conclusions were reached. First off, one team's model was clearly determined to be better than all the others, even though in the initial publication it reported the middlemost validation performance. Second, one dataset was clearly "harder" to predict well than were the others, in the sense of initially reported misclassification rate, but it is precisely this dataset which yielded the overall best model.

Table 2. OMOP datasets. Numerical figures give the number of persons or objects. Thus, 46.5M in the upper left means 46.5 million persons; while 110M in the lower right means 110 million procedures.

Acronym	Pop. size	Source	Timerange	Drugs	Cond	Proc
CCAE	46.5M	Private	2003–2009	1.03B	1.26B	1.98B
MDCD	20.8	Medicaid	2002–2007	360M	552M	558M
MDCR	4.6M	Medicare	2003–2009	401M	405M	478M
MSLR	1.2M	Lab	2003–2007	38M	50M	69M
GE	11.2M	EHR	1996–2008	182M	66M	110M

This meta study demonstrates that by both accessing all previous data from a group of studies and trying all previous modeling approaches on all datasets, one can obtain both a better result and a fuller understanding of the problems and shortcomings of actual data analyses.

The effort involved in conducting this study is breathtaking. The authors delved deeply into the details of over 100 scientific papers and understood fully how the data cleaning and data fitting was done in each case. All the underlying data were accessed and reprocessed into a new common curated format, and all the steps of the data fitting were reconstructed algorithmically, so they could be applied to other datasets. Again information technology plays a key role; much of the programming for this project was carried out in R. Parmigiani and collaborators are biostatisticians heavily involved in the development of R packages.

9.3. Cross-Workflow Analysis

A crucial hidden component of variability in science is the analysis workflow. Different studies of the same intervention may follow different workflows, which may cause the studies to get different conclusions. Carp (2012) studied analysis workflows in 241 fMRI studies. He found nearly as many unique workflows as studies! In other words researchers are making up a new workflow for pretty much every fMRI study.

David Madigan and collaborators (Ryan et al. 2012; Madigan et al. 2014) studied the effect of analysis flexibility on effect sizes in observational studies; their collaboration will be hereafter called OMOP. As motivation, the OMOP authors point out that in the clinical research literature there are studies of the same dataset, and the same intervention and outcome, but with different analysis workflow, and the published conclusions about the risk of the intervention are *reversed*. Madigan gives the explicit example of exposure to Pioglitazone and bladder cancer, where published articles in BJMP and BMJ reached opposite conclusions on the very same underlying database!

The OMOP authors obtained five large observational datasets, covering together a total of more than 200 Million Patient-years (see Table 2).

The OMOP group considered four different outcomes, coded "Acute Kidney Injury," "Acute Liver Injury," "Acute Myocardial Infarction," "GI Bleed." They considered a wide range of possible interventions for each outcome measure, for example, whether

patients taking drug X later suffered outcome Y. Below, "Acute Liver Injury" stands for the association "Exposure to X and Acute Liver Injury."

For each target outcome, the researchers identified a collection of known positive and negative controls, interventions X for which the ground truth of statements like "Exposure to X is associated with Acute Liver Injury" is considered known. Using such controls, they could quantify an inference procedure's ability to correctly spot associations using the measure of area under the operating curve (AUC).

OMOP considered seven different procedures for inference from observational studies, labeled "CC," "CM," "DP," "ICTPD," "LGPS," "OS," and "SCCS." For example, "CC" stands for case–control studies, while SCCS stands for self-controlled case series. In each case, the inference procedure can be fully automated.

In their study, OMOP considered, for each database, for each possible outcome, every one of the seven types of observational study method (CC, …, SCCS).

The OMOP report concludes that the three so-called self-controlled methods outperform the other methods overall, with SCCS being especially good overall. So their study reveals quite a bit about the effectiveness of various inference procedures, offering an idea what improved inference looks like and how accurate it might be.

This work represents a massive endeavor by OMOP: to curate data, program inference algorithms in a unified way, and apply them across a series of underlying situations. Dealing with big data was an essential part of the project; but the driving motivation was to understand that the scientific literature contains a source of variation—methodological variation—whose influence on future inference in this field might be understood, capped, or even reduced. The participants were statisticians and biostatisticians.

9.4. Summary

There seem to be significant flaws in the validity of the scientific literature (Ioannidis 2007; Sullivan 2007; Prinz, Schlange, and Asadullah 2011; Open Science Collaboration et al. 2015). The last century has seen the development of a large collection of statistical methodology, and a vast enterprise using that methodology to support scientific publication. There is a very large community of expert and not-so-expert users of methodology. We do not know very much about how that body of methodology is being used and we also do not know very much about the quality of results being achieved.

Data scientists should not blindly churn out methodology without showing concern for results being achieved in practice. Studies we have classed as "GDS6: Science About Data Science" help us understand how data analysis as actually practiced is impacting "all of science."

Information technology skills are certainly at a premium in the research we have just covered. However, scientific understanding and statistical insight are firmly in the driver's seat.

10. The Next 50 Years of Data Science

Where will data science be in 2065? The evidence presented so far contains significant clues, which we now draw together.

10.1. Open Science Takes Over

In principle, the purpose of scientific publication is to enable reproducibility of research findings. For centuries, computational results and data analyses have been referred to in scientific publication, but typically only have given readers a hint of the full complexity of the data analysis being described. As computations have become more ambitious, the gap between what readers know about what authors did has become immense. Twenty years ago, Jon Buckheit and I summarized lessons we had learned from Stanford's Jon Claerbout as follows:

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

To meet the original goal of scientific publication, one should share the underlying code and data. Moreover there are benefits to authors. Working from the beginning with a plan for sharing code and data leads to higher quality work, and ensures that authors can access their own former work, and those of their coauthors, students and postdocs (Donoho et al. 2009). Over the years, such practices have become better understood (Stodden 2012; Stodden, Guo, and Ma 2013) and have grown (Freire, Bonnet, and Shasha 2012; Stodden, Leisch, and Peng 2014), though they are still far from universal today. In absolute terms the amount of essentially nonreproducible research is far larger than ever before (Stodden, Guo, and Ma 2013).

Reproducible computation is finally being recognized today by many scientific leaders as a central requirement for valid scientific publication. The 2015 annual message from Ralph Cicerone, President of the U.S. National Academy of Sciences, stresses this theme; while funding agencies (Collins and Tabak 2014) and several key journals (Peng 2009; McNutt 2014; Heroux 2015), have developed a series of reproducibility initiatives.

To work reproducibly in today's computational environment, one constructs automated workflows that generate all the computations and all the analyses in a project. As a corollary, one can then easily and naturally refine and improve earlier work continuously.

Computational results must be integrated into final publications. Traditional methods—running jobs interactively by hand, reformatting data by hand, looking up computational results, and copying and pasting into documents—are now understood to be irresponsible. Recently, several interesting frameworks combining embedded computational scripting with document authoring[52] have been developed. By working within the discipline such systems impose, it becomes very easy to document the full computation leading to a specific result in a specific article. Yihui Xie's work with the `knitr` package—mentioned earlier—is one such example.[53]

[52] Such efforts trace back to Donald Knuth's Literate Programming project. While literate programming—mixing code and documentation—does not seem to have become very popular, a close relative—mixing executable code, data, documentation, and execution outputs in a single document—is just what the doctor ordered for reproducible research in computational science.

[53] Professor Martin Helm reminds me to mention other examples; he points to the SAS system's StatRep package, saying "SAS Institute twice a year produces tens of thousands pages of SAS documentation from LaTeX-files with markups that run SAS and include programs as well as output as well as statistical advice (text). When we tested it, it was better and more stable than `knitr`. This could have

Reproducibility of computational experiments is just as important to industrial data science as it is to scientific publication. It enables a disciplined approach to proposing and evaluating potential system improvements and an easy transition of validated improvements into production use.

Reproducible computation fits into our classification both at "GDS 4: Presentation of Data" and in "GDS 6: Science about Data Science." In particular, teaching students to work reproducibly enables easier and deeper evaluation of their work; having them reproduce parts of analyses by others allows them to learn skills like exploratory data analysis that are commonly practiced but not yet systematically taught; and training them to work reproducibly will make their post-graduation work more reliable.

Science funding agencies have for a long time included in their funding policies a notional requirement that investigators make code and data available to others. However, there never has been enforcement, and there was always the excuse that there was no standard way to share code and data. Today there are many ongoing development efforts to develop standard tools enabling reproducibility (Freire, Bonnet, and Shasha 2012; Stodden, Leisch, and Peng 2014; Stodden and Miguez 2014); some are part of high profile projects from the Moore and Simons foundations. We can confidently predict that in coming years reproducibility will become widely practiced.

10.2. Science as Data

Conceptually attached to a scientific publication is a great deal of numerical information—for example, the *p*-values reported within it (Chavalarias et al. 2016). Such information ought to be studied as data. Today, obtaining that data is problematic; it might involve reading of individual articles and manual extraction and compilation, or web scraping and data cleaning. Both strategies are error prone and time consuming.

With the widespread adoption of open science over the next 50 years, a new horizon becomes visible. Individual computational results reported in an article, and the code and the data underlying those results, will be universally citable and programmatically retrievable. Matan Gavish and I wrote some articles (Gavish and Donoho 2011; Gavish 2012), which proposed a way to open that new world and which then explored the future of science in such a world.

Those articles defined the notion of verifiable computational result (VCR), a computational result, and metadata about the result, immutably associated with a URL, and hence permanently programmatically citable and retrievable. Combining cloud computing and cloud storage, Gavish developed server frameworks that implemented the VCR notion, recording each key result permanently on the server and returning the citing URL. He also provided client-side libraries (e.g., for Matlab) that allowed creation of VCRs and returned the associated link, and that provided programmatic access to the data referenced by the link. On the document creation side, he provided macro packages that embedded such links into published TEXdocuments. As a result, one could easily write documents in which every numerical result computed for an article was publicly citable and inspectable—not only the numerical value, but the underlying computation script was viewable and could be studied.

In a world where each numerical result in a scientific publication is citable and retrievable, along with the underlying algorithm that produced it, current approaches to meta-analysis are much easier to carry out. One can easily extract all the *p*-values from a VCR-compliant article, or extract all the data points in a graph inside it, in a universal and rigorously verifiable way. In this future world, the practice of meta-analysis of the kind we spoke about in Section 9.1 will of course expand. But many new scientific opportunities arise. We mention two examples:

- *Cross-Study Control Sharing.* In this new world, one can extract control data from previous studies (Wandell et al. 2015). New opportunities include: (a) having massively larger control sets in future studies; (b) quantifying the impact of specific control groups and their differences on individual study conclusions; and (c) extensive "real world" calibration exercises where both groups are actually control groups.
- *Cross-Study Comparisons.* The cross-study comparisons of Sections 9.2 and 9.3, required massive efforts to manually rebuild analyses in previous studies by other authors, and then manually curate their data. When studies are computationally reproducible and share code and data, it will be natural to apply the algorithm from paper A on the data from paper B, and thereby understand how different workflows and different datasets cause variations in conclusions. One expects that this will become the dominant trend in algorithmic research.

Additional possibilities are discussed in Gavish (2012).

10.3. Scientific Data Analysis, Tested Empirically

As science itself becomes increasingly mineable for data and algorithms, the approaches of cross-study data sharing and workflow sharing discussed above in Sections 9.2 and 9.3 will spread widely. In the next 50 years, ample data will be available to measure the performance of algorithms across a whole ensemble of situations. This is a game changer for statistical methodology. Instead of deriving optimal procedures under idealized assumptions within mathematical models, we will rigorously measure performance by empirical methods, based on the entire scientific literature or relevant subsets of it.

Many current judgments about which algorithms are good for which purposes will be overturned. We cite three references about the central topic of classification with a bit of detail.

10.3.1. Hand et al. (2006)

In Hand et al. (2006), D. J. Hand summarized the state of classifier research in 2006. He wrote:

> *The situation to date thus appears to be one of very substantial theoretical progress, leading to deep theoretical developments and to increased predictive power in practical applications. While all of these things are true, it is the contention of this paper that the practical impact of the developments has been inflated; that although progress has been made, it may well not be as great as has been suggested. …*
> *The essence of the argument [in this paper] is that the improvements attributed to the more advanced and recent developments are small,*

changed in the meantime as `knitr` evolves but SAS is not so eager to open up and publish improvements."

and that aspects of real practical problems often render such small differences irrelevant, or even unreal, so that the gains reported on theoretical grounds, or on empirical comparisons from simulated or even real data sets, do not translate into real advantages in practice. That is, progress is far less than it appears.[54]

How did Hand support such a bold claim? On the empirical side, he used "a randomly selected sample of 10 datasets" from the literature and considered empirical classification rate. He showed that linear discriminant analysis, which goes back to Fisher (1936), achieved a substantial fraction (90% or more) of the achievable improvement above a random guessing baseline. The better-performing methods were much more complicated and sophisticated—but the incremental performance above LDA was relatively small.

Hand's theoretical point was precisely isomorphic to a point made by Tukey in FoDA about theoretical optimality: optimization under a narrow theoretical model does not lead to performance improvements in practice.

10.3.2. Donoho and Jin (2008)

To make Hand's point completely concrete, consider work on high-dimensional classification by myself and Jiashun Jin (Donoho and Jin 2008).[55]

Suppose we have data $X_{i,j}$ consisting of $1 \le i \le n$ observations on p variables, and binary labels $Y_i \in \{+1, -1\}$. We look for a classifier $T(X)$, which presented with an unlabeled feature vector predicts the label Y. We suppose there are many features, that is, p is large-ish compared to n.

Consider a very unglamorous method: a linear classifier $C(x) = \sum_{j \in J_+} x(j) - \sum_{j \in J_-} x(j)$, which combines the selected features simply with weights +1 or –1. This method selects features where the absolute value of the univariate t-score exceeds a threshold and uses as the sign of the feature coefficient simply the sign of that feature's t-score. The threshold is set by higher criticism. In the published article it was called HC-clip; it is a dead-simple rule, *much simpler even than classical Fisher linear discriminant analysis*, as it makes no use of the covariance matrix, and does not even allow for coefficients of different sizes. *The only subtlety is in the use of higher criticism for choosing the threshold.* Otherwise, HC-clip is a throwback to a pre-1936 setting, that is, to before Fisher (1936) showed that one "must" use the covariance matrix in classification.[56]

Dettling (2004) developed a framework for comparing classifiers that were common in Machine Learning based on a standard series of datasets (in the 2-class case, the datasets are called ALL, Leukemia, and Prostate, respectively). He applied these datasets to a range of standard classifier techniques which are popular in the statistical learning community (boosted decision trees, random forests, SVM, KNN, PAM, and DLDA). The machine learning methods that Dettling compared are mostly "glamorous," with high numbers of current citations and vocal adherents.

We extended Dettling's study, by adding our dead-simple clipping rule into the mix. We considered the regret (i.e., the ratio of a method's misclassification error on a given dataset to the best misclassification error among all the methods on that specific dataset). Our simple proposal did *just as well on these datasets as any of the other methods*; it even has the *best* worst-case regret. That is, *every one* of the more glamorous techniques suffers worse maximal regret. Boosting, random forests, and so on are dramatically more complex and have correspondingly higher charisma in the machine learning community. But against a series of preexisting benchmarks developed in the machine learning community, the charismatic methods do not outperform the homeliest of procedures—feature clipping with careful selection of features.

As compared to Hand's work, our work used a preexisting collection of datasets that might seem to be less subject to selection bias, as they were already used in multi-classifier shootouts by machine learners.

10.3.3. Zhao et al. (2014)

In a very interesting project (Zhao et al. 2014), Parmigiani and co-authors discussed what they called the *Más-o-Menos* classifier, a linear classifier where features may only have coefficients that ± 1; this is very much like the just-discussed HC-clip method, and in fact one of their variants included only those features selected by HC—that is, the method of the previous section. We are again back to pre-Fisher-says-use-covariance-matrix, pre-1936 setting.

In their study, Zhao et al. compared Más-o-Menos to "sophisticated" classifiers based on penalization (e.g., lasso, ridge).

Crucially, the authors took the fundamental step of comparing performance on a *universe* of datasets used in *published clinical medical research*. Specifically, they curated a series of datasets from the literature on treatment of bladder, breast, and ovarian cancer, and evaluated prediction performance of each classification method over this universe.

We …demonstrated in an extensive analysis of real cancer gene expression studies that [Más-o-Menos] can indeed achieve good discrimination performance in realistic settings, even compared to lasso and ridge regression. Our results provide some justification to support its widespread use in practice. We hope our work will help shift the emphasis of ongoing prediction modeling efforts in genomics from the development of complex models to the more important issues of study design, model interpretation, and independent validation.

The implicit point is again that *effort devoted to fancy-seeming methods is misplaced* compared to other, more important issues. They continue

One reason why Más-o-Menos is comparable to more sophisticated methods such as penalized regression may be that we often use a prediction model trained on one set of patients to discriminate between subgroups in an independent sample, usually collected from a slightly different population and processed in a different laboratory. This cross-study variation is not captured by standard theoretical analyses, so theoretically optimal methods may not perform well in real applications.[57]

[54] The point made by both Hand and Tukey was that optimality theory, with its great charisma, can fool us. J. R. Pierce made a related point in rejecting the "glamor" of theoretical machine translation.

[55] We did not know about Hand's article at the time, but stumbled to a similar conclusion.

[56] In the era of desk calculators, a rule that did not require multiplication but only addition and subtraction had some advantages.

[57] Again this vindicates Tukey's point from 1962 that optimization of performance under narrow assumptions is likely a waste of effort, because in practice, the narrow assumptions do not apply to new situations and so the supposed benefits of optimality never appear.

In comparison to the articles (Hand et al. 2006; Donoho and Jin 2008) discussed in previous subsections, this work, by mining the scientific literature, speaks directly to practitioners of classification in a specific field—giving evidence-based guidance about what would have been true for studies to date in that field, had people all known to use the recommended technique.

10.4. Data Science in 2065

In the future, scientific methodology will be validated empirically. Code sharing and data sharing will allow large numbers of datasets and analysis workflows to be derived from studies science-wide. These will be curated into corpora of data and of workflows. Performance of statistical and machine learning methods will thus ultimately rely on the cross-study and cross-workflow approaches we discussed in Sections 9.2 and 9.3 Those approaches to quantifying performance will become standards, again because of code and data sharing. Many new common task frameworks will appear; however, the new ones would not always have prediction accuracy for their performance metric. Performance might also involve validity of the conclusions reached, or empirical Type I and II error. Research will move to a meta level, where the question becomes: "if we use such-and-such a method across all of science, how much will the global science-wide result improve?" measured using an accepted corpus representing science itself.

In 2065, mathematical derivation and proof will not trump conclusions derived from state-of-the-art empiricism. Echoing Bill Cleveland's point, theory which produces new methodology for use in data analysis or machine learning will be considered valuable, based on its quantifiable benefit in frequently occurring problems, as shown under empirical test.[58]

11. Conclusion

Each proposed notion of data science involves some enlargement of academic statistics and machine learning. The "GDS" variant specifically discussed in this article derives from insights about data analysis and modeling stretching back decades. In this variant, the core motivation for the expansion to data science is intellectual. In the future, there may be great industrial demand for the skills inculcated by GDS; however, the core questions which drive the field are scientific, not industrial.

GDS proposes that data science is the science of learning from data; it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner. The scope and impact of this science will expand enormously in coming decades as scientific data and data about science itself become ubiquitously available.

Society already spends tens of billions of dollars yearly on scientific research, and much of that research takes place at universities. GDS inherently works to understand and improve the validity of the conclusions produced by university research, and can play a key role in all campuses where data analysis and modeling are major activities.

[58] I am not arguing for a demotion of mathematics. I personally believe that mathematics offers the best way to create true breakthroughs in quantitative work. The empirical method is simply a method to avoid self-deception and appeals to glamor.

Epilogue

The "1.00 version" of this article was dated September 18, 2015. Since its release I received dozens of e-mails from readers with comments. Four sets of comments were particularly valuable, and I'll review them here, together with my response.

Data Science as Branding

C. F. Jeff Wu, Professor of Industrial and Systems Engineering at Georgia Tech, wrote to me, pointing out that he had been using the term "data science" in the 1990s. In Section 4.1, we have already mentioned his inaugural Carver lecture at the University of Michigan. Wu proposed in that lecture that statistics rebrand itself.

We mentioned earlier that the Royal Statistical Society hosted a "debate" in May 2015[59]—a video is posted online—asking whether in fact data science is merely such a rebranding, or something larger. Wu's data science proposal was ahead of its time.[60]

I have argued here that data science is *not* a mere rebranding or retitling of statistics. Today's consensus data science includes statistics as a subset.[61] I think data science ought to be even larger, for example, to include GDS6: Science about Data Science.

Comments from University of Michigan Readers

I received e-mails from three readers at the University of Michigan who in various degrees of intensity objected to my portrayal of the MIDAS data science initiative (DSI).

For example, Peter Lenk told me good-naturedly that I "bashed his University," while statistician R.J.A. Little offered a friendly warning to avoid "inflammatory language."

Al Hero, the director of the MIDAS initiative, wrote to me making several points, some of which are reflected in footnotes to Section 2.1. Hero's points include: (1) that statisticians were involved in the planning effort for MIDAS; (2) that the speakers at the introductory colloquium all used statistical methods in fundamental ways, even though they may not have all been from the Statistics Department; and (3) that the 135 MIDAS affiliated faculty include 30+ statisticians in the Statistics Department and elsewhere. These are all very appropriate points to make.

I have nothing to criticize about the MIDAS program; nowhere do I point to some other DSI as doing a better job. I have never organized such a program and doubt that I could. The initiative seems well designed and well run.

Hero and others were concerned that readers of my article might form the incorrect opinion that statisticians were specifically excluded from the MIDAS data science initiative; Hero

[59] *Data Science and Statistics: Different Worlds?* Participants: Zoubin Ghahramani (Professor of Machine Learning, University of Cambridge), Chris Wiggins (Chief Data Scientist, New York Times), David Hand (Emeritus Professor of Mathematics, Imperial College), Francine Bennett (Founder, Mastodon-C), Patrick Wolfe (Professor of Statistics, UCL / Executive Director, UCL Big Data Institute). Chair: Martin Goodson (Vice-President Data Science, Skimlinks). Discussant: John Pullinger (UK National Statistician).

[60] Wikipedia credits computer scientist Peter Naur with using the term data science heavily already in the 1960s and 1970s, but not really in the modern sense.

[61] Echoing Wu's master's curriculum proposal from the Carver lecture.

explained they were involved all along. I never thought otherwise.

I am writing here about what the program and its announcement would *look like to many statisticians* upon inception. Moreover, my *specific* points about the public face of the initiative were uncontested.

To clarify my position: I think that many statisticians would today, on the basis of appearances, conclude that data science—while overlapping heavily with statistics—offers a public face intentionally obscuring this overlap.[62] In making a DSI look new and attractive for the media and potential students, DSI administrators downplay continuity with the traditional statistics discipline—suggesting that such discontinuity is a feature and not a bug.

Moreover, I think it is healthy for statisticians to have this perception. Let them ponder the idea that statistics may become marginalized. The train may well be leaving the station. Statisticians may well get left behind.

Chris Wiggins, Columbia University

Chris Wiggins is the Chief Data Scientist of the *New York Times*, and a Professor at Columbia affiliated with several programs, including its Data Science Initiative and its Statistics Department.

Wiggins and I first interacted after my talk at Princeton in September 2015 when he pointed out to me that he had earlier made a presentation about data science, for example, at ICERM, in which John Tukey's FoDA played a large part. In fact the parallelism in places of the two presentations is striking.[63]

Wiggins made numerous points in conversation and in e-mail, the most striking of which I will now attempt to evoke. I stress that Wiggins did not always use these specific words. I hope that he will publish an essay making his points.

- *Academic statistics is permanently deformed by the postwar association with mathematics that Tukey first called attention to.*
- *That deformation will prevent it from having relevance in the new data science field and the new big data world that is rapidly forming.*
- *"Data science is not a science, even though Donoho might like it. It is a form of engineering, and the doers in this field will define it, not the would-be scientists."* (C. Wiggins, private communication, October 2015.)

These statements have weight. I obviously cannot object to the first one. The second and third ones are predictions and time will obviously tell.

I agree with Wiggins that whether statisticians will respond energetically to the data science challenge is very much an open question.

Sean Owen's "What '50 Years of Data Science' Leaves Out"

Sean Owen, Director of Data Science at Cloudera, has posted an essay reacting to my manuscript.[64]

Owen makes many interesting points, and readers will wish to consult the text directly.

Near the beginning, we read:

> …reading it given my *engineering-based* take on data science, it looks like an attack on a partial straw-man. Along the way to arguing that data science can't be much more than statistics, it fails to contemplate data engineering, which I'd argue is most of what data science is and statistics is not.

I like Owen's essay and offer a few mild responses.

- Nowhere did I say that data science cannot be much more than statistics. I quote approvingly others who say the opposite. John Chambers was saying quite literally that there is a larger field than traditional statistics, and Bill Cleveland was saying quite literally that what some academic statistics departments are doing is a small part of data science. I did literally say that even data science as it is currently being instituted is too limited; there is a field I called "greater data science" extending beyond the limits of today's "consensus data science."
- Owen's essay focuses on the important term "data engineering," which I under-played in the main text. Data engineers exploit currently available cloud/cluster computing resources to allow the storage of large databases and to implement complex processing pipelines.[65]
- Owen writes that "data engineering …is most of what data science is and statistics is not." Owen's claim goes beyond my current understanding of the boundaries of both statistics and data science. A block diagram I used in my talk at Princeton showed blocks labeled "Statistics," "Data Science," and "Data Engineering;" there are important differences between these blocks at the present time.

Owen complained that my Version 1.00 manuscript was "writing data engineering out of the story of data science." I certainly intended no such thing. For the *academic statistics audience* that I was addressing at the Tukey Centennial, many of whom had not previously heard talks about the data science moment, I judged that I had zero chance of engaging the audience unless I could connect to our (their and my) common academic and historical background.

Sean Owen is completely correct to say that the full story of the emergence of data science, and the role of industry in shaping and defining it, remains to be written.

I return to a point I made earlier: *many of the struggles associated with scaling up algorithms are transitory.* In time, better tools will come along that make the data engineering part of the data science equation much easier in many applications.

Owen's essay supports this point. Owen describes how the example I gave above in Section 2.2, involving Hadoop, is no

[62] In fact, Hero's remarks support this interpretation; Hero says in so many words that, if statisticians really knew all the facts behind the scenes, they would agree that statistics is substantially involved in the MIDAS DSI. Fine—but this means my comment that "statistics is the subject that dare not speak its name" (Aside: Is this the inflammatory language that Professor Little worries about?) is roughly on target.

[63] James Guzca, Chief Data Scientist at Deloitte, later wrote to me about a data science course that he gave in which Tukey's statements also play a similar role.

[64] *https://medium.com/@srowen/what-50-years-of-data-science-leaves-out-2366c9b61d3d*

[65] Insight Data trains conventionally trained PhD's to become data scientists and data engineers. From their website: *Our definition of data engineering includes what some companies might call Data Infrastructure or Data Architecture. The data engineer gathers and collects the data, stores it, does batch processing or real-time processing on it, and serves it via an API to a data scientist who can easily query it. https://blog.insightdatascience.com/about-insight-b535888ecb3a*

longer current; the data engineering community has moved away from Hadoop toward Apache Spark, where the example I gave would be much easier to implement. The rapid obsolescence of specific tools in the data engineering stack suggests that today, the academic community can best focus on teaching broad principles—"science" rather than "engineering."

Acknowledgments

Thanks to John Storey, Amit Singer, Esther Kim, and all the other organizers of the Tukey Centennial at Princeton, September 18, 2015. This provided an occasion for me to organize my thoughts on this topic, for which I am grateful.

Special thanks to Edgar Dobriban (U. Penn.), Bradley Efron (Stanford), and Victoria Stodden (Univ Illinois) for comments on data science and on pre 1.00 drafts of this article. Thanks to the many readers of the 1.00 draft who wrote to me with valuable suggestions about how it could be improved. In particular, Chris Wiggins (Columbia) corresponded extensively with me on matters large and small concerning many aspects of Version 1.00 and in so doing clarified my thinking notably. This version contains an Epilogue mentioning some important points raised by Wiggins, and also by Jeff Wu (Georgia Tech), Al Hero (University of Michigan), and Sean Owen (Cloudera).

Thanks are also due to Deepak Agarwal (Linked In), Rudy Beran (UC Davis), Peter Brown (Renaissance Technologies), Jiashun Jin (Carnegie Mellon), Rob Kass (Carnegie Mellon), Martin Helm (Deutsche Bank), Ron Kennett (KPA Associates), Peter Lenk (Univ. Michigan), Mark Liberman (Univ. Pennsylvania), R.J. Little (Univ. Michigan), Xiao-Li Meng (Harvard) and Adrian Raftery (University of Washington), and Hadley Wickham (RStudio and Rice) for encouragement and corrections.

Belated thanks to my undergraduate statistics teachers: Peter Bloomfield, Henry Braun, Tom Hettmansperger, Larry Mayer, Don McNeil, Geoff Watson, and John Tukey.

Funding

Supported in part by NSF DMS-1418362 and DMS-1407813.

References

Barlow, M. (2013), *The Culture of Big Data*, Sebastopol, CA: O'Reilly Media, Inc. [748]

Baumer, B. (2015), "A Data Science Course for Undergraduates: Thinking With Data," *The American Statistician*, 69, 334–342. [757]

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa, L. (2014), "Cross-Study Validation for the Assessment of Prediction Algorithms," *Bioinformatics*, 30, i105–i112. [759]

Breiman, L. (2001), "Statistical Modeling: the Two Cultures," *Statistical Science*, 16, 199–231. [751]

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013), "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14, 365–376. [759]

Carp, J. (2012), "The Secret Lives of Experiments: Methods Reporting in the fMRI Literature," *Neuroimage*, 63, 289–300. [759]

Chambers, J. M. (1993), "Greater or Lesser Statistics: A Choice for Future Research," *Statistics and Computing*, 3, 182–184. [750]

Chavalarias, D., Wallach, J., Li, A., and Ioannidis, J. A. (2016), "Evolution of Reporting p Values in the Biomedical Literature, 1990–2015," *Journal of the American Medical Association*, 315, 1141–1148. [759,761]

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Monterey, CA: Wadsworth Advanced Books and Software. [758]

——— (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

——— (2001), "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," *International Statistical Review*, 69, 21–26. [750]

Coale, A. J., and Stephan, F. F. (1962), "The Case of the Indians and the Teen-Age Widows," *Journal of the American Statistical Association*, 57, 338–347. [755]

Collins, F., and Tabak, L. A. (2014), "Policy: NIH Plans to Enhance Reproducibility," *Nature*, 505, 612–613. [760]

Cook, D., and Swayne, D. F. (2007), *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*, New York: Springer Science & Business Media. [758]

Dettling, M. (2004), "BagBoosting for Tumor Classification with Gene Expression Data," *Bioinformatics*, 20, 3583–3593. [762]

Donoho, D., and Jin, J. (2008), "Higher Criticism Thresholding: Optimal Feature Selection When Useful Features Are Rare and Weak," *Proceedings of the National Academy of Sciences*, 105, 14790–14795. [762]

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., and Stodden, V. (2009), "Reproducible Research in Computational Harmonic Analysis," *Computing in Science and Engineering*, 11, 8–18. [760]

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188. [762]

Freire, J., Bonnet, P., and Shasha, D. (2012), "Computational Reproducibility: State-of-the-Art, Challenges, and Database Research Opportunities," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, ACM, pp. 593–596. [760,761]

Gavish, M. (2012), "Three Dream Applications of Verifiable Computational Results," *Computing in Science & Engineering*, 14, 26–31. [761]

Gavish, M., and Donoho, D. (2011), "A Universal Identifier for Computational Results," *Procedia Computer Science*, 4, 637–647. [761]

Hand, D. J. (2006), "Classifier Technology and the Illusion of Progress," *Statistical Science*, 21, 1–14. [761,763]

Harris, H., Murphy, S., and Vaisman, M. (2013), *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, Sebastopol, CA: O'Reilly Media, Inc. [748]

Heroux, M. A. (2015), "Editorial: ACM TOMS Replicated Computational Results Initiative," *ACM Transactions on Mathematical Software*, 41, 13:1–13. [760]

Horton, N. J., Baumer, B. S., and Wickham, H. (2015), "Taking a Chance in the Classroom: Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics," *CHANCE*, 28, 40–50. [757]

Hotelling, H. (1940), "The Teaching of Statistics," *The Annals of Mathematical Statistics*, 11, 457–470. [754]

Ioannidis, J. P. A. (2005), "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 218–228. [759]

——— (2007), "Non-Replication and Inconsistency in the Genome-Wide Association Setting," *Human Heredity*, 64, 203–213. [760]

——— (2008), "Why Most Discovered True Associations are Inflated," *Epidemiology*, 19, 640–648. [759]

Iverson, K. E. (1991), "A Personal View of APL," *IBM Systems Journal*, 30, 582–593. [758]

Jager, L. R., and Leek, J. T. (2014), "An Estimate of the Science-Wise False Discovery Rate and Application to The Top Medical Literature," *Biostatistics*, 15, 1–12. [759]

Liberman, M. (2010), "Fred Jelinek," *Computational Linguistics*, 36, 595–599. [752]

Madigan, D., Stang, P. E., Berlin, J. A., Schuemie, M., Overhage, J. M., Suchard, M. A., Dumouchel, B., Hartzema, A. G., and Ryan, P. B. (2014), "A Systematic Statistical Approach to Evaluating Evidence From Observational Studies," *Annual Review of Statistics and Its Application*, 1, 11–39. [759]

Marchi, M., and Albert, J. (2013), *Analyzing Baseball Data with R*, Boca Raton, FL: CRC Press. [757]

McNutt, M. (2014), "Reproducibility," *Science*, 343, 229. [760]

Mosteller, F., and Tukey, J. W. (1968), "Data Analysis, Including Statistics," in *Handbook of Social Psychology* (Vol. 2), eds. G. Lindzey, and E. Aronson, Reading, MA: Addison-Wesley, pp. 80–203. [749]

Open Science Collaboration et al. (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [760]

Pan, Z., Trikalinos, T. A., Kavvoura, F. K., Lau, J., and Ioannidis, J. P. A. (2005), "Local Literature Bias in Genetic Epidemiology: An

Empirical Evaluation of the Chinese Literature," *PLoS Medicine*, 2, 1309. [759]

Peng, R. D. (2009), "Reproducible Research and Biostatistics," *Biostatistics*, 10, 405–408. [760]

Prinz, F., Schlange, T., and Asadullah, K. (2011), "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery*, 10, 712–712. [760]

Ryan, P. B., Madigan, D., Stang, P. E., Overhage, J. M., Racoosin, J. A., and Hartzema, A. G. (2012), "Empirical Assessment of Methods for Risk Identification in Healthcare Data: Results From the Experiments of the Observational Medical Outcomes Partnership," *Statistics in Medicine*, 31, 4401–4415. [759]

Stodden, V. (2012), "Reproducible Research: Tools and Strategies for Scientific Computing," *Computing in Science and Engineering*, 14, 11–12. [760]

Stodden, V., Guo, P., and Ma, Z. (2013), "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," *PLoS ONE*, 8, e67111. [760]

Stodden, V., Leisch, F., and Peng, R. D., editors. (2014), *Implementing Reproducible Research*, Boca Raton, FL: Chapman & Hall/CRC. [760,761]

Stodden, V., and Miguez, S. (2014), "Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research," *Journal of Open Research Software*, 1, e21. [761]

Sullivan, P. F. (2007), "Spurious Genetic Associations," *Biological Psychiatry*, 61, 1121–1126. [760]

Tango, T. M., Lichtman, M. G., and Dolphin, A. E. (2007), *The Book: Playing the Percentages in Baseball*, Lincoln, NE: Potomac Books, Inc. [757]

Tukey, J. W. (1962), "The Future of Data Analysis," *The Annals of Mathematical Statistics*, 33, 1–67. [749]

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley. [758]

——— (1994), *The Collected Works of John W. Tukey*: Multiple Comparisons (Vol. 1), eds. H. I. Braun, Pacific Grove, CA: Wadsworth & Brooks/Cole. [758]

Wandell, B. A., Rokem, A., Perry, L. M., Schaefer, G., and Dougherty, R. F. (2015), "Quantitative Biology – Quantitative Methods," Bibliographic Code: 2015arXiv150206900W. [761]

Wickham, H. (2007), "Reshaping Data With the Reshape Package," *Journal of Statistical Software*, 21, 1–20. [757]

——— (2011), "ggplot2," *Wiley Interdisciplinary Reviews: Computational Statistics*, 3, 180–185. [757]

——— (2011), "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software*, 40, 1–29. [757]

——— (2014), "Tidy Data," *Journal of Statistical Software*, 59, 1–23. [757]

Wilkinson, L. (2006), *The Grammar of Graphics*, New York: Springer Science & Business Media. [758]

Zhao, S. D., Parmigiani, G., Huttenhower, C., and Waldron, L. (2014), "Más-o-Menos: A Simple Sign Averaging Method for Discrimination in Genomic Data Analysis," *Bioinformatics*, 30, 3062–3069. [762]