50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

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 twenty years ago, John Chambers, 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 even suggested the catchy name "Data Science" for his envisioned field.

A recent and growing phenomenon is the emergence of "Data Science" programs at major universities, including UC Berkeley, NYU, MIT, and most recently the Univ. of Michigan, which on September 8, 2015 announced a $100M "Data Science Initiative" that will hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; in general, though, the new initiatives steer away from close involvement with academic statistics departments.

This paper 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 fifty 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 Sept 18 2015:

Contents

Acknowledgements:

Special thanks to Edgar Dobriban, Bradley Efron, and Victoria Stodden for comments on Data Science and on drafts of this mss.

Thanks to John Storey, Amit Singer, Esther Kim, and all the other organizers of the Tukey Centennial at Princeton, September 18, 2015.

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

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

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 Furure 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 lagugage produced by SAS, Inc.
SPSS	System and lagugage produced by SPSS, Inc.
VCR	Verifiabe Computational Result

Table 1: Frequent Acronyms

1 Today's Data Science Moment

On Tuesday September 8, 2015, as I was preparing these remarks, the University of Michigan announced a $100 Million "Data Science Initiative" (DSI), 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 web site 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 interdisciplinary 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,...

There are new announcements of such initiatives weekly.[1]

2 Data Science 'versus' Statistics

Many of my audience at the Tukey Centennial where these remarks were 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, 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).

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

- The inagural symposium has one speaker from the Statistics department (Susan Murphy), out of more than 20 speakers.

Seemingly, statistics is being marginalized here; the implicit message 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![2]

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"[3]

```
  ``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 which 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[4]

- *A grand debate: is data science just a 'rebranding' of statistics?*
  Martin Goodson, co-organizer of the Royal Statistical Society meeting May 11, 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[5]

[2] 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's all hidden under the hood.

[3] `http://www.datascienceassn.org/code-of-conduct.html`

[4] `http://magazine.amstat.org/blog/2013/07/01/datascience/`

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

One doesn't need to look far to see click-bait 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[6]

On the other hand, we can find pointed comments about the (near-) irrelevance of statistics:

- *Data Science without statistics is possible, even desirable.*
  Vincent Granville, at the Data Science Central Blog[7]

- *Statistics is the least important part of data science.*
  Andrew Gelman, Columbia University [8]

Clearly, there are many visions of Data Science and its relation to Statistics. In discussions one recognizes certain recurring 'Memes'. We 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[9].

- *History.* The very term 'statistics' was coined at the beginning of modern efforts to compile census data, i.e. 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

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

[7] `http://www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable`

[8] `http://andrewgelman.com/2013/11/14/statistics-least-important-part-data-science/`

[9] 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 in order to allow students to make hand calculations.

big data: the punched card reader to allow efficient compilation of an exhaustive US census.[10] 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.[11]

- *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*.[12]

2.2 The 'Skills' Meme

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.*

The CS evangelists are thus doubling down on the 'Big Data' meme[13], by layering a 'Big Data skills meme' on top.

What are those skills? 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 in order to implement the example at scale.

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

[11] 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'.

[12] It may be getting at something real about the Masters' degree programs, or about the research activities of individuals who will be hired under the new spate of DSI's.

[13] ... which we just dismissed!

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 towards 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 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'* [1], 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.[14]

We don't really know. The booklet *'Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work'* [20] 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.*

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

Yanir Seroussi's blog[15] 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 don't withstand even basic scrutiny. This is quite understandable: writers and administrators are *shocked out of their wits.* Everyone believes we are facing a zero-th 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, 8 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 seems certain to be a major preoccupation of commercial life in coming decades.

2.5 A Better Framework

However, a science doesn't 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.

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

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 science itself becomes a body of data that we can analyze and study, there are staggeringly large opportunities for improving the accuracy and validity of science, through the scientific study of data analysis.

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

In this paper, 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 paper was prepared for the John Tukey centennial. More than 50 years ago, John prophecied that something like today's Data Science moment would be coming. In "The Future of Data Analysis" [42], John deeply shocked his readers (academic statisticians) with the following introductory paragraphs:[16]

> *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*

This paper 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

[16] 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 paper.

at the time were mathematically precise and would present definitions, theorems, and proofs. John's paper 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.

Tukey and Wilk 1969 compared this new science 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".

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.

PJ Huber wrote that *"The influence of Tukey's paper was not immediately recognized ... it took several years until I assimilated its import ..."*. 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 1980's and appeared 48 years after FoDA!

4.1 Exhortations

While Huber obviously made the choice to explore the vistas offered in Tukey's vision, the academic field as a whole did not. 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 [6], 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 paper, 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 states explicitly that the enlarged field would be *larger even than data analysis.* Specifically, it is larger than Tukey's 1962 vision.

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 paper [8], titled "Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics"[17] 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 paper's introduction, Cleveland writes that[18] [19]

> *...[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 6 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 paper was republished in 2014. I can't 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, to change paths, towards a much broader definition of their field.* Such exhortations had relatively little apparent effect before 2000.

[17] This may be earliest use of the term Data Science in which it is utterly clear that the writer means it in exactly the modern sense. An even earlier use by Jeff Wu proposes that Statisticians change the name of their profession to data science because so much of what we do involves data cleaning and preparation.

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

[19] Geophysicsts 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 re Statistical theorists: do they 'care about the data analyst' or do they not?

4.2 Reification

One obstacle facing the earliest exhortations was that many of the exhortees couldn't see what the fuss was all about. Making the activity labelled '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 1960's, 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[20].

To quantify the importance of these packages, try using Google's N-grams viewer [21] 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 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-1970's; it provided a language for describing computations, and many basic statistical and visualization tools. In the 1990's, 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 on-line 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[22], workflows can be widely shared within the community and re-executed, 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 paper – 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'.

[20] 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 explotatory data analysis; this ultimately became SPIDA after he moved to Macquarie University.

[21] `https://books.google.com/ngrams/graph?content=SPSS%2CSAS%2CMinitab%2CData+Analysis%2CStatistical+Analysis&year_start=1970&year_end=2000&corpus=15&smoothing=3...`

[22] or Matlab in Signal Processing

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

Leo Breiman, a UC Berkeley statistician who re-entered 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 2001 paper in *Statistical Science*. 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].*[23] *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'[24] 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[25] 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*

[23] I changed Breiman's words here slightly; the original has 'Information' in place of [Inference] and 'extract some information about' in place of [infer]

[24] 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.

[25] Breiman used 'Algorithmic' rather than 'Predictive'

> *informative alternative to data modeling on smaller data sets. 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 paper, 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 states that '*Prediction is certainly an interesting subject but Leo's paper 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:

> *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 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 paper didn't 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 Marc Liberman (Liberman, 2009) 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 Marc Liberman's telling it starts with JR 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 1960's. 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 didn't like much of what he saw passing as statistics research in the 1960's, Pierce didn't like much of what he saw passing as 1960's machine translation research.

Now we follow Marc Liberman closely[26]. Judging that the field was riddled with susceptibility to 'Glamor and Deceit', Pierce managed to cripple the whole US 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 re-emerged 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 led by Fred Jelinek at IBM, which included true geniuses like John Cocke, began to make definite progress towards 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 1980's 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.*

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

3. *Shared data plays a crucial role – and is re-used 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 Marc 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 necessarily been tested, and so is essentially ready for immediate deployment.[27]

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.

[27] However, in the case of the Netflix Challenge the winning algorithm was never implemented. https://www.techdirt.com/blog/innovation/articles/20120409/03412518422/why-netflix-never-implemented-algorithm-that-won-netflix-1-million-challenge.shtml

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 in order 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's consider the attractive and informative web site for the UC Berkeley Data Science Masters' degree at `datascience.berkeley.edu`.

Reviewing the curriculum at `https://datascience.berkeley.edu/academics/curriculum/` we find 5 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, the actual topics covered in"Applied Machine Learning", seem to a statistician 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 towards 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 researchers" 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 wouldn't 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 non-statistical 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.[28]

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.*

[28] I don't wish to imply in the above that there's anything concerning to me about the composition of the faculty. I do wish to demonstrate that this is an opportunity being seized by non-statisticians. An important even in the history of academic statistics was Hotelling's article "The Teaching of Statistics" (1940) [23] which decried the teaching of statistics by non-mathematicians, and motivated the formation of academic statistics departments. The new developments may be undoing the many years of postwar professionalization of statistics instruction.

> *(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; i.e. 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's difficult to say whether they would approve of their long-term direction - or if there is even a long-term direction to comment about. 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 masters program to make room for large database training; or, equally, as taking some material out of a database masters in CS and inserting some statistics and machine learning. Such a compromise helps administrators to quickly get a degree program going, without providing any guidance about the long-term direction of the program and about the research which its faculty will pursue. What long-term guidance could my heroes have 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 also 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 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's 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 Greater Data Science are classified into 6 divisions:

1. *Data Exploration and Preparation*
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 Exploration and Preparation. 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 two subactivities:

- *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.[29].
- *Preparation.* Many datasets contain anomalies and artifacts.[30] Any data-driven project requires mindfully identifying and addressing such issues. Responses range from reformatting and recoding the values themselves, to more ambitious pre-processing, such as grouping, smoothing, and subsetting. Often today, one speaks colorfully of *data cleaning*.

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 judgement 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.

[29] 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.

[30] Peter Huber (2010) recalls the classic Coale and Stephan paper on Teenage Widows

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 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. [31]
- *Predictive modeling*, in which one constructs methods which predict well over some some given data universe – i.e. some very specific concrete dataset. This roughly coincides with modern Machine Learning, and its industrial offshoots. [32]

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.[33]

[31] It is striking how, when I review a presentation on today's data science, in which statistics is superficially given pretty short shrift, I can't 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 can't be hidden in the teaching, research, and exercise of Data Science.

[32] Leo Breiman (2001) is 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.

[33] 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'.

Data scientists are doing *science about data science* when they identify commonly-occuring 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 analyes and conclusions in a standard digital format for future harvesting and meta analysis.

As data analysis and predictive modelling 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.[34],[35] 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'. [37],[38]

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

[34] John Chambers' 1993 vision of 'Greater Statistics' proposed 3 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.

[35] 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.[36]

[37] 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.

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

- 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 won't 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 6 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 Exploration and Preparation' 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 pre-cooked 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* [41], 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* [30] shows 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 Exploration and Preparation'. 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 Bauman and co-authors review experiences in [22, 2] 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'.[39]

8.4 Research in GDS

Once we have the GDS template in mind, we can recognize that today there's all sorts of interesting –and highly impactful – 'GDS research'. Much of it doesn't 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's stretchable to cover lots of mainstream academic computer science; suggesting that perhaps there's 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 on-line 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`, and `plyr`; [46, 48, 49]. 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 indispensible to many.

In [47] Wickham discusses the notion of *tidy* data. Noting (as I also have, above) the common estimate that *80% of data analysis is spent on the process of cleaning and preparing the data*, Wichkam 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 which 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

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

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 [27], 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 papers.

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 [9, 7], along with Leland Wilkinson [50] and Hadley Wickham's [46] 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 [43]; more recently Cook and Swayne's work on Dynamic graphics [12].

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[40], Tukey proposed that statisticians should study how people analyze data today.

By formalizing the notion of multiple comparisons [44], 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.[41] In 1953, the introduction to Tukey's paper [44] considered a very small scale example with 6 different comparisons under study. Today, more than 1 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 ten 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 for clues about meta-problems in the way all of science is analyzing data. I can cite a few papers by John Ioannidis and co-authors [24, 26, 4, 32] and for statisticians the paper 'An estimate of the science-wise false discovery rate...' [28] 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 1 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.

[40] *"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*

[41] 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.

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 on-line 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. 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 [3], 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 data sets 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.

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 was 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.

Acronym	Pop. Size	Source	Timerange	Drugs	Cond	Proc
CCAE	46.5M	Private	2003-09	1.03B	1.26B	1.98B
MDCD	20.8	Medicaid	2002-07	360M	552M	558M
MDCR	4.6M	Medicare	2003-09	401M	405M	478M
MSLR	1.2M	Lab	2003-07	38M	50M	69M
GE	11.2M	EHR	1996-08	182M	66M	110M

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.

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. Joshua Carp [5] 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 study.

David Madigan and collaborators [35, 29] 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 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 5 large observational datasets, covering together a total of more than 200 Million Patient-years

The OMOP group considered 4 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 to Acute Liver Injury” is considered known. Using such controls, they could quantify an inference procedures’s ability to correctly spot associations using the measure of Area Under the Operating Curve (AUC).

OMOP considered 7 different procedures for inference from observational studies, labelled “CC”, “CM”, “DP” “ICTPD”, “LGPS”, “OS”, “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 [34, 25, 40, 10]. 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 don't know very much about how that body of methodology is being used and we also don't know very much about the quality of results being achieved.

Data scientists can't 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 co-authors, students and postdocs [14]. Over the years, such practices have become better understood [36, 37] and have grown [38, 16], through they are still far from universal today. In absolute terms the amount of essentially non-reproducible research is far larger than ever before [37].

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 US National Academy of Sciences stresses this theme; while funding agencies [11] and several key journals [33, 21, 31], have developed a series of reproducibility initiatives.

To work reproducibly in today's computational environment, one constructs automated workflows which 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[42] 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 paper. Yihui Xie's work with the `knitr` package - mentioned above - is one such example.

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 [38, 16, 39], 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 . Such information ought to be studied as data. Today, obtaining that data is problematic; it might involve reading of individual papers and manual extraction and compilation, or web scraping and data cleaning. Both strategies are error prone and time consuming.

[42] Such efforts trace back to Donald Knuth's Literate Programming project. While Literate programming -mixing code and documentation - doesn't 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.

With the widespread adoption of open science over the next 50 years, a new horizon becomes visible. Individual computational results reported in a paper, and the code and the data underlying those results, will be universally citable and programmatically retrievable. Matan Gavish and I wrote some papers [18, 17] which proposed a way to open that new world and which then explored the future of science in such a world.

Those papers 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 TeX documents. As a result, one could easily write documents in which every numerical result computed for a paper 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 paper, 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 [45]. 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 conlusions; 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 [17].

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 judgements 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 DJ Hand (2006)

in [19], DJ 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, 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.*[43]

How did Hand support such a bold claim? On the empirical side, he used 'a randomly selected sample of ten data sets' from the literature and considered empirical classification rate. He showed that Linear Discriminant Analysis, which goes back to Fisher (1936) [15], 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 [13].[44]

Suppose we have data $X_{i,j}$ consisting of $1 \leq i \leq n$ observations on p variables, and binary labels $Y_i \in \{+1, -1\}$. We look for a classifier $T(X)$ which, presented with an unlabelled feature vector predicts the label Y. We suppose there are many features, i.e. 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 paper 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 doesn't even allow

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

[44] We didn't know about Hand's paper at the time, but stumbled to a similar conclusion.

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, i.e. to before Fisher [15] showed that one 'must' use the covariance matrix in classification.[45]

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 pre-existing 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 pre-existing 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, Parmigiani, Huttenhower and Waldron (2014)

In a very interesting project [51], Parmigiani and co-authors discuss what they call 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 - i.e. 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.*

[45] In the era of desk calculators, a rule that didn't require multiplication but only addition and subtraction had some advantages.

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.*[46]

In comparison to the articles [19, 13] 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 won't 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.[47]

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.

[46] 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 don't apply to new situations and so the supposed benefits of optimality never appear.

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

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.

References

[1] Mike Barlow. *The Culture of Big Data.* O'Reilly Media, Inc., 2013.

[2] B. Baumer. A Data Science Course for Undergraduates: Thinking with Data. *ArXiv e-prints*, March 2015.

[3] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, 2014.

[4] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

[5] Joshua Carp. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1):289–300, 2012.

[6] John M Chambers. Greater or lesser statistics: a choice for future research. *Statistics and Computing*, 3(4):182–184, 1993.

[7] William S Cleveland. *Visualizing data.* Hobart Press, 1993.

[8] William S Cleveland. Data Science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26, 2001.

[9] William S Cleveland et al. *The elements of graphing data.* Wadsworth Advanced Books and Software Monterey, CA, 1985.

[10] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[11] Francis Collins and Lawrence A. Tabak. Policy: NIH plans to enhance reproducibility. *Nature*, 505(7484):612–613, 2014.

[12] Dianne Cook and Deborah F Swayne. *Interactive and dynamic graphics for data analysis: with R and GGobi.* Springer Science & Business Media, 2007.

[13] David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

[14] David L. Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. Reproducible Research in Computational Harmonic Analysis. *Computing in Science and Engineering*, 11(1):8–18, 2009.

[15] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[16] Juliana Freire, Philippe Bonnet, and Dennis Shasha. 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, pages 593–596, New York, NY, USA, 2012. ACM.

[17] Matan Gavish. Three dream applications of verifiable computational results. *Computing in Science & Engineering*, 14(4):26–31, 2012.

[18] Matan Gavish and David Donoho. A universal identifier for computational results. *Procedia Computer Science*, 4:637–647, 2011.

[19] David J Hand et al. Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14, 2006.

[20] Harlan Harris, Sean Murphy, and Marck Vaisman. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media, Inc., 2013.

[21] Michael A. Heroux. Editorial: ACM TOMS Replicated Computational Results Initiative. *ACM Trans. Math. Softw.*, 41(3):13:1–13:5, June 2015.

[22] Nicholas J Horton, Benjamin S Baumer, and Hadley Wickham. Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *arXiv preprint arXiv:1502.00318*, 2015.

[23] Harold Hotelling. The teaching of statistics. *The Annals of Mathematical Statistics*, 11(4):457–470, 1940.

[24] John PA Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228, 2005.

[25] John PA Ioannidis. Non-replication and inconsistency in the genome-wide association setting. *Human heredity*, 64(4):203–213, 2007.

[26] John PA Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648, 2008.

[27] Kenneth E. Iverson. A personal view of APL. *IBM Systems Journal*, 30(4):582–593, 1991.

[28] Leah R Jager and Jeffrey T Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, 2014.

[29] David Madigan, Paul E Stang, Jesse A Berlin, Martijn Schuemie, J Marc Overhage, Marc A Suchard, Bill Dumouchel, Abraham G Hartzema, and Patrick B Ryan. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39, 2014.

[30] Max Marchi and Jim Albert. *Analyzing Baseball Data with R.* CRC Press, 2013.

[31] Marcia McNutt. Reproducibility. *Science*, 343(6168):229, 2014.

[32] Zhenglun Pan, Thomas A Trikalinos, Fotini K Kavvoura, Joseph Lau, and John PA Ioannidis. Local literature bias in genetic epidemiology: an empirical evaluation of the chinese literature. *PLoS Medicine*, 2(12):1309, 2005.

[33] Roger D. Peng. Reproducible research and Biostatistics. *Biostatistics*, 10(3):405–408, 2009.

[34] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[35] Patrick B Ryan, David Madigan, Paul E Stang, J Marc Overhage, Judith A Racoosin, and Abraham G Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in medicine*, 31(30):4401–4415, 2012.

[36] Victoria Stodden. Reproducible Research: Tools and Strategies for Scientific Computing. *Computing in Science and Engineering*, 14(4):11–12, 2012.

[37] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLoS ONE*, 8(6):e67111, 06 2013.

[38] Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors. *Implementing Reproducible Research.* Chapman & Hall/CRC, 2014.

[39] Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 1(2):e21, 2014.

[40] Patrick F Sullivan. Spurious genetic associations. *Biological psychiatry*, 61(10):1121–1126, 2007.

[41] Tom M Tango, Mitchel G Lichtman, and Andrew E Dolphin. *The Book: Playing the percentages in baseball.* Potomac Books, Inc., 2007.

[42] John W Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, pages 1–67, 1962.

[43] John W Tukey. Exploratory data analysis. 1977.

[44] John Wilder Tukey. *The collected works of John W. Tukey*, volume 1. Taylor & Francis, 1994.

[45] B. A. Wandell, A. Rokem, L. M. Perry, G. Schaefer, and R. F. Dougherty. Data management to support reproducible research. *ArXiv e-prints*, February 2015.

[46] Hadley Wickham. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2):180–185, 2011.

[47] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10), 2014.

[48] Hadley Wickham et al. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[49] Hadley Wickham et al. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[50] Leland Wilkinson. *The grammar of graphics*. Springer Science & Business Media, 2006.

[51] Sihai Dave Zhao, Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. *Bioinformatics*, 30(21):3062–3069, 2014.