Statistics for Big Data: Are Statisticians Ready for Big Data?

John M. Jordan and Dennis K.J. Lin

Editorial: Dennis Lin gave a banquet speech at the inaugural symposium of the ICSA Canada Chapter at Toronto (August 2-3, 2013), entitled "BIGstat@IC^{2}SA." It was so well received, I have invited him to expand his talk for this article on BIGdata. I am pleased that he has agreed to do so and given below is his view on Big Data. I sincerely hope that you will enjoy the reading as much as I did. Happy Holidays,

Ming-Hui Chen, 2013 President, ICSA

Abstract: *After noting the relative absence of statisticians from the community of practice engaged with big data, we explain what big data is, how it's done, and who's working with it. The paper then suggests that statisticians have much to contribute to both the intellectual vitality and the practical utility of big data. At the same time, big data challenges statisticians to move out of some familiar habits to engage less structured problems, to become more comfortable with ambiguity, and to engage computer scientists in a more fruitful discussion of what the various parties can bring to this new mode of investigation.*

Introduction

We seem to be living in a time of paradox. As more and more commercial firms and academic disciplines proclaim their affinity for new kinds of data-driven modes of evidence and analysis, members of the professional statistical community frequently find themselves on the outside looking in. Even as the world discovers a set of tools, techniques, and attitudes lumped together in the phrase ``big data," (see Figure 1) statistics as a discipline is all too often absent.

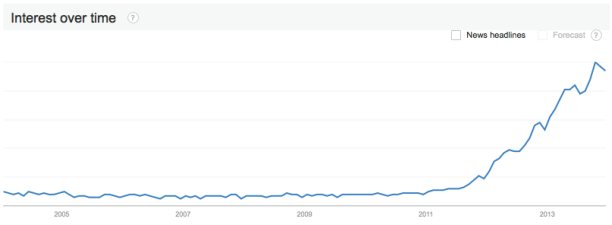


Figure 1: Relative interest as expressed by search queries at Google in the phrase "big data" (retrieved 12/12/13 at `http://www.google.com/trends/`)

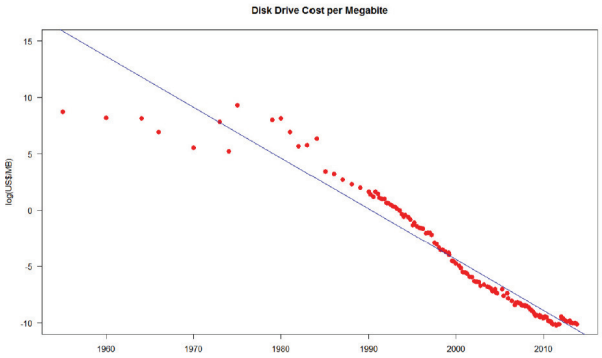
After explaining a bit about what big data is, how it's done, and who's working with it, we will suggest that statisticians have much to contribute to both the intellectual vitality and the practical utility of this venture. At the same time, big data challenges statisticians to move out of some familiar habits to engage less structured problems, to become more comfortable with ambiguity, and to engage computer scientists in a more fruitful discussion of what the various parties can bring to this new mode of investigation. Theory and practice, history and innovation, and rigor and results all seem to be posed with increasing frequency as opposites, when in the best investigation, they are in fact held in creative tension.

What is Big Data?

This might be a tricky question; for many in Silicon Valley where this meme is taking shape, "Big Data" is data that must be managed with a set of technologies called MapReduce. To an astrophysicist, or geneticist, or actuary, meanwhile, the notion might sound curious: these individuals see enormous data sets addressed via other more traditional means on a daily basis. There's no formal size threshold involved: data doesn't somehow become "big" at a petabyte, or whatever; the name instead refers more to a state of mind, in many cases to Web-scale user, network, and traffic data as encountered and managed at Amazon, Facebook, Google, and Yahoo. For statisticians, however, big data might initially be viewed as the class of problems involving "large n and/or large p."

Thus, big data has a nomenclature problem. Like so many other technologies -- smartphones, robots, or information security -- the popular name doesn't really convey the essence of the situation. Yes, "big data" can involve very large volumes in some cases. But more generally, the phrase refers to new kinds of data, generated, managed, and parsed in new ways, not merely bigger ones. While "Big Data" is a vague phrase, there is some agreement that it involves changes in scale along three dimensions (the three V's):

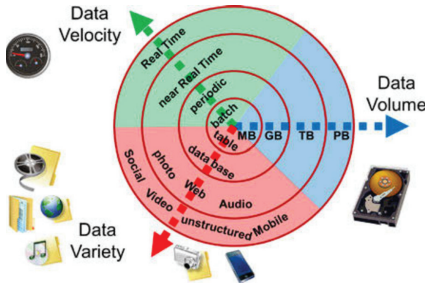
Volume: Whether it's your own hard disk space, the world's online video feeds, or a wealth of digital sensors measuring many aspects of the planet, signs are abundant that data volumes are increasing steadily and substantially. The volume in big data can grow in part due to sensor traffic from the "Internet of Things," to social media, to more people coming on line every day around the world, and from the increased use of geolocation, in part connected to the rise of mobile communications. As a result, the size of today's datasets is growing enormously—from MB, GB, TB, all the way to ZB (zetabytes=10^{21} bytes). Accordingly, big data must be seen in some ways as a triumph of data storage, as the graph illustrates: hard disk storage is dropping in price faster than microprocessor performance (Moore's law) is improving. See, `http://www.jcmit.com/diskprice.htm`.



Variety: Big data is not only a matter of bigger relational databases. As opposed to the familiar numbers related to customer ID, stock keeping unit (SKU), or price and quantity, we are living in an age of massive amounts of unstructured data: e-mails, Facebook "likes," Tweets, machine traffic, and video. Performing analysis of heterogeneous data types often strains both the information technologies and the statistical toolkits involved.

Velocity: Overnight batch processes are getting to be less and less tenable as the world becomes an "always-on" information environment. When FedEx can tell me where my package is, or Fidelity can tell me my net worth, or Google Analytics can tell me my website performance right now, the pressure is on more and more other systems to do likewise. In some instances that we will not cover here in depth, being able to analyze streams of data (including on stock exchanges), some of them very big and very fast, has become an imperative.

These three V's can be displayed in the popularized graph below.



There's an important point to be made up front: *big data is not necessarily complete, or accurate, or true*. Asking the right questions is in some cases learned through experience, or made possible by better theory, or a matter of luck. But in many instances, by the time investigators figure out what they should be measuring in complex systems, it's too late to instrument the "before" state to compare to the "after." Signal and noise can be problematic categories as well: one person's noise can be a gold mine for someone else. Context is everything. Value is in the eye of the beholder, not the person crunching the numbers. Thus it is temping to mention "value" as a fourth V. However, this is rarely the case. Big data is big, often because it is automatically collected. Thus in many cases, it may not contain much information relative to noise. This is sometimes called a DRIP—Data Rich, Information Poor—environment. In any event, the point here is that bigger does not necessarily mean better when it comes to data.

Accordingly, big data skills cannot be purely a matter of computer science, statistics, or other processes. Instead, the backstory behind the creation of any given data point, category, or artifact can be critically important. While the same algorithm or statistical transformation might be indicated in a bioscience and a financial scenario, knowing the math is rarely sufficient. Having the industry background to know where variance is "normal," for instance, comes only from a holistic understanding of the process under the microscope. We thus recommend an informal fourth V to be "Veracity" (or even "Validity"): managing the randomization in big data is indeed one great opportunity for statisticians.

Who is working on big data? How are they doing it?

Techniques used in big data analysis

Analyzing large, diverse data sets requires new tools. It is important to note, however, that this toolset can often be applied to data that might not qualify as "big." Note also that many of these techniques can be difficult to define with purity since many are only subtly different from adjacent tools. Furthermore, the tools are often used in combination.

A/B testing. Similar to clinical trials, A/B testing is often used to compare website enhancements: a sample population at a shopping site might get a red banner or a 10% off coupon while a second population is shown a blue banner or a "free shipping" offer, then purchase results are compared across the two groups and successful enhancements are included in production systems.

Data fusion. Data fusion is "a process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats, and their significance. The process is characterized by continuous refinements of its estimates and assessments, and the evaluation of the need for additional sources, or modification of the process itself, to achieve improved results."

Data Mining. Data mining is a broad umbrella including a variety of techniques to detect patterns in large datasets, using different combinations of statistics, computer science, and database management. Some examples of data mining include the following: (a) **Association rule learning**, which seeks to discover interesting across variables. These techniques generate possible rules, which might seek to explain why certain products are bought together within a process called market basket analysis. (b) **Classification** attempts to place new data points into categories that have been determined by previous analysis of a training data set. (c) **Cluster analysis** groups items into sets —it differs from classification in its lack of training data —in this case the properties of similarity are not known in advance. (d) **Regression** is a classical statistical technique that attempts to determine how the value of the dependent variable changes when one or more independent variables is modified.

Machine learning. This is a broad term that encompasses many techniques with origins in computer science, particularly what has been called 'artificial intelligence" for more than 50 years. In the most fundamental definition, machine learning seeks to build systems that can learn to recognize complex patterns, and then make decisions based on these data patterns. (a) **Natural language processing** is one example of machine learning. Apple smartphones' Siri feature and the IBM Watson com-

puter that won at Jeopardy are two familiar examples of NLP, which involves far more than speech recognition. (b) **Unsupervised learning** seeks to uncover hidden structure in unlabeled data. Cluster analysis is an example of unsupervised learning. (c) **Supervised learning** uses training data to "teach" a system to find similar patterns in subsequent data to which it is exposed. (d) **Ensemble learning** is a specific type of supervised learning in which multiple predictive models are combined to obtain better predictive performance than could be obtained from any one of the constituent models. In this combinatorial aspect ensemble learning resembles the approach of data fusion. (e) **Neural networks** are a class of computational models modeled on biological neural networks, most typically those within a brain. These systems are used to find patterns in either supervised or unsupervised fashion, and work particularly well for finding nonlinear patterns. Neural networks have been successfully used for fraud detection.

Network analysis. Network analysis has come to prominence in the past 20 years both in the pursuit of asymmetric warfare adversaries including terror networks, as well as in connection with digital social networks such as Twitter or Facebook. Several techniques are used to map and describe relationships among discrete nodes in a network.

Optimization. Optimization takes known systems, in a factory, financial institution, or hospital, for example, and seeks to analyze and correct limitations to speed, cost, or outcomes. Genetic algorithms (or more recently Particle Swarm Optimization, for example) can be used for optimization.

Geospatial Analysis. Many techniques are available to analyze the spatial properties related to a given data set. The easy availability of cell phone geolocation data, for example, helps mobile network providers design cell tower placement. The free global GPS network means that latitude and longitude data is readily available for most anywhere on the planet.

Simulation (Computer Experiment). Numerous computer techniques allow the behavior of complex systems to be modeled, informing everything from forecasting to scenario planning. Monte Carlo simulations are commonly used: they run thousands of simulations, each based on different assumptions, generating a histogram that represents a probability distribution of potential outcomes. On the other hand, some computer experiments (expensive simulations) could take a long time to run, special care is needed.

Time series analysis. Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data. Time series forecasting is the use of a model to predict future values of a time series based on known past values of the same or other series. Some of these techniques, e.g., structural modeling, decompose a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data. Examples of applications include forecasting sales figures, or predicting the number of people who will be diagnosed with an infectious disease.

Visualization. Spreadsheets are particularly ineffective in conveying complex relationships within large data sets. New techniques are emerging to create images, diagrams, or animations to both communicate help people understand data. Interactivity is becoming more common, allowing users of a data set to manipulate the analyses rather than being confined to static, two-dimensional ink and paper.

Who is working on big data?

To date, big data is primarily a joint venture between IT professionals (who provide expertise on database management, data cleansing, and networking) and computer scientists, who are in the algorithm business. In any given investigation (whether video rental prediction, public health, or credit card or insurance fraud prevention), subject-matter experts also play a key role. The data creation process, in particular, might require expertise in handheld devices, sensor networks, satellite imaging, or other particular technical domains. The business process under discussion, whether voting, online dating, life insurance, or ball bearing production, also dictates the need for experts in the laws, physical constraints, and other unique attributes of the particular phenomenon.

For all of the reliance on such traditional statistical strong points such as regression analysis, statisticians have not played leadership roles in the movement: when the U.S. National Science Foundation convened a working group on the topic in 2012, zero statisticians were named to a committee of 100 experts. Why do we see this incongruity between substantial historical expertise and limited contemporary relevance? Three possible reasons come to mind:

- Big data has compiled a track record addressing poorly defined problems. Statisticians

have grown accustomed to well structured problems, often using a single technique in a delimited domain. The practical problems solved by Netflix, Amazon, and Google are messier than most statisticians see on a regular basis.

- The computational intensiveness of big data lies outside the expertise of most statisticians. Being able to understand the entire data life cycle, from sensor design, through collection, ETL (extract, transform, load), and storage is one common shortcoming related to IT expertise. Knowing how to assemble then burn massive numbers of compute cycles in a cluster or cloud scenario is a second, and largely unrelated, set of skills outside the statistician's sweet spot.

- Professional statisticians have found that asking productive, if only partially solved, questions does not generate research publications. In contrast, narrowly defined problems that can be solved used robust theoretical constructs can be found in all the major journals. We statisticians always prefer to be precisely wrong (missing the forest but microanalyzing a tree) rather than approximately right. In contrast, many big data projects seek to predict user behavior (Amazon and Google are textbook examples) and do not seek repeatable scientific laws underneath a successful association or prediction.

What kinds of statistics are needed for big data?

Many statisticians commonly appear to believe that (i) big data is better than small data, and (ii) new methodologies are powerfully robust and work well in most cases. Consequently, the so-called "the death of p-value" is claimed. However, disasters kept happening. Is it because that the fundamental statistical thinking still applied, although the theories may not be straightforwardly applied? Big data cannot replace scientific/statistical thinking. Data and information/knowledge are not synonyms. Thus a wish-list for needed statistical methodologies should have the following properties.

- High-impact problems. Refining existing methodologies is fine, but more efforts should focus on working high-impact problems, especially those problems from other disciplines. Statisticians seem to keep missing opportunities: examples range from genetics to data mining. We believe that statisticians should seek out high-impact problems, instead of waiting for other disciplines to formulate the problems into statistical frames. This leads to the next item.

- Provide structure for poorly defined problems. A skilled statistician is typically most comfortable and capable when dealing with well-defined problems. Instead, statisticians should develop some methodologies for poorly defined problems and help devise a strategy of attack. There are many opportunities for statistical applications, but most of them are not in the "standard" statistics frame —it will take some intelligent persons to formulate these problems into statistics-friendly problems (then to be solved by statisticians). Statisticians can devote more efforts to be such intelligent persons.

- Develop new theories. Most fundamental statistical theories based upon iid (independently identically distributed) for one fixed population (such as, central limit theorem, or law of large number) may need to be modified to be appropriately applied to big data world. Many (non-statisticians) believe that big data leads to "the death of p-value." The logic behind this is that when the sample size n becomes really large, all p-values will be significant —regardless how little the practical significance is. This is indeed a good example of misunderstanding the fundamentals. Another good example is about "small n and large p" where the sparsity property is assumed. First, when there are many exploratory variables, some will be classified as active variables (whether or not this is true!). Even worse, after the model is built (mainly based on the sparsity property), the residuals may be highly correlated with some remaining variables —this contradicts the assumption for all fundamental theorems that "error is independent with all exploratory variables." New measurement is needed for independence in this case.

Having those wishlist items in mind, what kinds of statistics are needed for Big data? For the reason of casting a brick to attract jade (抛砖引玉), given

below are some very initial thoughts under consideration.

- Statistics and plots for (many) descriptive statistics. If the conventional statistics are to be used for big data, and it is very likely there will be too many of them, what is the best way to extract important information from these statistics? For example, how to summarize thousands of correlations? How about thousands of p-values? ANOVA? Regression models? Histograms? etc. Advanced methods to obtain "sufficient statistics" (whatever it means) from those many conventional statistics are needed.

- Low-dimension behavior. Whatever method is feasible for big data (the main concern being the computational costs), its low-dimension behavior is always important to be kept in mind.

- Norm or Extreme. Depending on the problem, we could be interested in either norm or extreme, or both. Basic methods for both feature extraction (mainly for extremes) and pattern recognition (mainly for norm) are needed.

- Methods for new types/structures of data. A simple example would be "How to build up a regression model, when both inputs and outputs are network variables?" Most existing statistical methodologies are limited to numbers (univariate or multivariate), although there is some recent work for functional data or text data. There are more that can be done, if we are willing to open our minds.

- Prediction vs estimation. One difference between computer science and statistics methods has to do with the general goal —while CS people focus more on prediction, statisticians focus more on estimation (or statistical inference). Take Artificial Neural Networks (ANN) as an example: the method can fit almost anything, but what does it mean? ANN is thus popularly used in data mining, but has received relatively low attention from statisticians. For big data, it is clear that prediction is probably more feasible in most cases. **Note:** in some very fundamental cases, we believe that statistical inference remains important.

Big Data in the Real World

Skills

Here's a quiz: ask someone in the IT shop how many of his of her colleagues are qualified to work in Hive, Pig, Cassandra, MongoDb, or Hadoop. These are some of the tools that are emerging from the front-runners in big data, web-scale companies including Google (that needs to index the entire Internet), Facebook (manage a billion users), Amazon (construct and run the world's biggest online merchant), or Yahoo (figure out what social media is conveying at the macro scale). Outside this small industry, big data skills are rare.

Politics

Control over information is frequently thought to bring power within an organization. Big data, however, is heterogeneous, multi-faceted, and can bring performance metrics where they had not previously operated. If a large retailer, hypothetically speaking, traced its customers' purchase behavior first to social media expressions and then to advertising channel, how will the various budget-holders respond? Uncertainty as to ad spend efficacy is as old as advertising, but tracing ad channels to purchase activity might bring light where perhaps it is not wanted. Information sharing across organizational boundaries ("how are you going to use this data?") can also be unpopular.

Technique

Given that relational databases have been around for about 35 years, a substantial body of theory and practice make these environments predictable. Big data, by contrast, is just being invented, but already there are some important differences between the two: Most enterprise data is generated by or about humans and organizations: SKUs are bought by people, bills are paid by people, health care is provided to people, and so on. At some level, many human activities can be understood at human scale. Big data, particularly social media, can come from people too, but in more and more cases, it comes from machines: server logs, point of sale scanner data, security sensors, GPS traces. Given that these new types of data don't readily fit into relational structures and can get massively large in terms of storage, it's nontrivial to figure out what questions to ask of these data types.

When data is loaded into relational systems, it must fit predefined categories that ensure that what gets put into a system makes sense when it is pulled out. This process implies that the system is defined at the outset for what the designers expect to be queried: the questions are known, more or less, before the data is entered in a highly structured manner. In big data practice, meanwhile, data is stored in as complete a form as possible, close to its original state. As little as possible is thrown out so queries can evolve and not be constrained by the preconceptions of the system. Thus these systems can look highly random to traditional database experts. It's important to stress that big data will not replace relational databases in most scenarios; it's a matter of now having more tools to choose from for a given task.

Traditional Databases

Traditional databases are designed for a concrete scenario, then populated with examples (customers, products, facilities, or whatever), usually one per row: the questions and answers one can ask are to some degree predetermined. Big data can be harvested in its original form and format, and then analyzed as the questions emerge. This open-ended flexibility can of course be both a blessing and a curse.

Traditional databases measured the world in numbers and letters that had to be predicted: zip codes were 5 or 10 digits, SKU formats were company-specific, or mortgage payments were of predictable amounts. Big data can accommodate Facebook "likes," instances of the "check engine" light illuminating, cellphone location mapping, and many other types of information.

Traditional databases are limited by the computing horsepower available: to ask harder questions often means buying more hardware. Big data tools can scale up much more gracefully and cost-effectively, so decision-makers must become accustomed to asking questions they could not contemplate previously. To judge advertising effectiveness, one cable operator analyzed every channel-surfing click of every remote across every household in its territory, for example: not long ago, such an investigation would have been completely impractical.

Cognition

What does it mean to think at large scales? How do we learn to ask questions of the transmission of every car on the road in a metropolitan area, of the smartphone of every customer of a large retail chain, or of every overnight parcel in a massive distribution center? How can more and more people learn to think probabilistically rather than anecdotally?

The mantra that "correlation doesn't imply causation" is widely chanted yet frequently ignored; it takes logical reasoning beyond statistical relationships to test what's really going on. Unless the data team can grasp the basic relationships of how a given business works, the potential for complex numerical processing to generate false conclusions is ever-present. Numbers do not speak for themselves; it takes a human to tell stories, but as Daniel Kahneman and others have shown, our stories often embed mental traps. Spreadsheets remain ubiquitous in the modern enterprise but numbers at the scale of Google, Facebook, or Amazon must be conveyed in other ways. Sonification -- turning numbers into a range of audible tones -- and visualization show a lot of promise as alternative pathways to the brain, bypassing mere and non-intuitive numerals. In the meantime, the pioneers are both seeing the trail ahead and taking some arrows in the back for their troubles. But the faster people, and especially statisticians, begin to break the stereotype that "big data is what we've always done, just with more records or fields," the faster the breakthrough questions, insights, and solutions will redefine business practice.

Privacy

It has been proven repeatedly that anonymous data sets can be reverse-engineered, identifying people who either did not know they were a part of a study or trusted the process. Elsewhere, the cheapness of computer data storage (as measured by something called Kreider's law) combines with the ubiquity of daily digital life to create massive data stores recording people's preferences, medications, travels, and social contacts. As big data tools continue to increase in power, and computational capability increases, and algorithmic sophistication increases, composing revealing data-driven portraits of tens of millions of people will be possible, profitable, and troubling. Put access to those portraits on wearable digital devices such as glasses, and the prospect of facial recognition by a random stranger on the street is certain to become an issue sooner rather than later. Big data practitioners may face calls for a professional code of conduct much like the Hippocratic oath for doctors: first, do no harm. Statisticians can provide sorely

needed expertise in this discussion.

Summing up

No matter what the field of inquiry, the form and scale of the data involved, the computational infrastructure, or the name attached to the project, statistics as a discipline has much to offer the big data effort. Recall the insight offered by Box, Hunter and Hunter 35 years ago: "Data have no meaning in themselves; they are meaningful only in relation to a conceptual model of the phenomenon being studied." Today, there are those who seem to suggest that models are unnecessary, that given sufficient computing power, the relevant patterns will emerge, absent theory. Indeed, there are those who argue that hypothesis-driven scientific method will become outdated in a world in which p-value no longer matters.

There is much to be gained from using a "scientific" approach, as opposed to an "algorithm" approach, to big data—including data collection, data analysis, mode selection, and feature interpretation. Before pushing the button (to run an automatic algorithm), perhaps more scientific thinking is needed beforehand.

Seeing big data in a wider historical perspective might be a useful way to end our discussion. Susan Hockfield recently retired as president of MIT after having been the first life scientist to lead the institution. She has concluded that "The convergence of life sciences and engineering, I think, is going to be the story of the 21st century, much as the convergence of the physical sciences and engineering was the story of the 20th century." [1] During the early 1900s, "physicists decoded the fundamental elements of the physical universe. They were essentially understanding the parts list of the physical world —the structure of atoms, how electrons travel." Engineers discovered this 'parts list' and began to turn theory into practice: the microprocessor, laser, and wireless networking -- the building blocks of the computer revolution -- are among these engineers' legacies.

By the 1950s, Hockfield continued, scientists including Watson and Crick were decoding the structure of DNA, and "the biological sciences began to assemble a parts list for the biological universe. And engineers, in a very similar way, as they saw the 'parts list' evolving, picked up those parts and incorporated them into applications." This insight underlies her contention that the 21st century will be an era characterized by everyday implementations of recently discovered conceptual building blocks. For our purposes, the question is clear: who among the many computational, statistical, IT, and domain experts is doing either of these two tasks: describing the "parts list" of quantitative investigation and discovery, or doing the engineering to turn theory into everyday reality, accessible to the masses?

Let us conclude this article by a quote from Hahn and Hoerl: "This is a golden age for Statistics, but not necessary for statisticians." We sincerely hope that this article provides some personal views for statisticians to be ready for the many changes that are underway under the banner of big data. In the simplest form, our advice is to be more aggressive, helping or even leading other disciplines to advance science and knowledge using tools that have served us well for centuries, revised to meet the needs of this new era.

John M. Jordan, PhD.
Clinical professor of Supply Chain & Information Systems
The Pennsylvania State University
jmj13@psu.edu

Dennis K.J. Lin, Ph.D.
University Distinguished Professor of Statistics and Supply Chain Management
The Pennsylvania State University
DennisLin@psu.edu

[1] James F. Smith, "Spotlight: Susan Hockfield and the Magic of the Laboratory," Winter 2012-13 newsletter, Belfer Center for Science and International Affairs, Harvard Kennedy School http://belfercenter.ksg.harvard.edu/publication/22515/spotlight.html. Our thanks to Dr. Roger Hoerl for bringing Hockfield's statements to our attention and for his inspiring discussions.