Do you consider yourself to be a statistician, data scientist, machine learner, or something else?

For the most part I consider myself to be a statistician, but I'm also very serious about genetics/genomics, data analysis, and computation. I was trained in statistics and genetics, primarily statistics. I was also exposed to a lot of machine learning during my training since Rob Tibshirani was my PhD advisor.

However, I consider my research group to be a data science group. We have the Venn diagram reasonably well covered: experimentalists, programmers, data wranglers, and developers of theory and methods; biologists, computer scientists, and statisticians.

How did you find out you had won the COPSS Presidents' Award?

I received a phone call from the chairperson of the awards committee while I was visiting the Department of Statistical Science at Duke University to give a seminar. It was during the seminar reception, and I stepped out into the hallway to take the call. It was really exciting to get the news!

One of the areas where you have had a big impact is inference in massively parallel problems. How do you feel high-dimensional inference is different from more traditional statistical inference?

My experience is that the most productive way to approach high-dimensional inference problems is to first think about a given problem in the scenario where the parameters of interest are random, and the joint distribution of these parameters is incorporated into the framework. In other words, I first gain an understanding of the problem in a Bayesian framework. Once this is well understood, it is sometimes possible to move in a more empirical and nonparametric direction. However, I have found that I can be most successful if my first results are in this Bayesian framework.

As an example, Theorem 1 from Storey (2003) *Annals of Statistics* was the first result I obtained in my work on false discovery rates. This paper first appeared as a technical report in early 2001, and the results spawned further work on a point estimation approach to false discovery rates, the local false discovery rate, q-value and its application to genomics, and a unified theoretical framework.

Besides false discovery rates, this approach has been useful in my work on the optimal discovery procedure as well as surrogate variable analysis (in particular, Desai and Storey 2012 for surrogate variable analysis).

For high-dimensional inference problems, I have also found it is important to consider whether there are any plausible underlying causal relationships among variables, even if causal inference in not the goal. For example, causal model considerations provided some key guidance in a recent paper of ours on testing for genetic associations in the presence of arbitrary population structure. I think there is a lot of insight to be gained by considering what is the appropriate approach for a high-dimensional inference problem under different causal relationships among the variables.

Do you have a process when you are tackling a hard problem or working with students on a hard problem?

I like to work on statistics research that is aimed at answering a specific scientific problem (usually in genomics). My process is to try to understand the *why* in the problem as much as the *how*. The path to success is often found in the former. I try first to find solutions to research problems by using simple tools and ideas. I like to get my hands dirty with real data as early as possible in the process. I like to incorporate some theory into this process, but I prefer methods that work really well in practice over those that have beautiful theory justifying them without demonstrated success on real-world applications.

In terms of what I do day-to-day, listening to music is integral to my process, for both concentration and creative inspiration: typically King Crimson or some variant of metal or ambient – which Simply Statistics co-founder Jeff Leek got to ~~endure~~ enjoy for years during his PhD in my lab.

You are the founding Director of the Center for Statistics and Machine Learning at Princeton. What parts of the new gig are you most excited about?

Princeton closed its Department of Statistics in the early 1980s. Because of this, the style of statistician and machine learner we have here today is one who's comfortable being appointed in a field outside of statistics or machine learning. Examples include myself in genomics, Kosuke Imai in political science, Jianqing Fan in finance and economics, and Barbara Engelhardt in computer science. Nevertheless, statistics and machine learning here is strong, albeit too small at the moment (which will be changing soon). This is an interesting place to start, very different from most universities.

What I'm most excited about is that we get to answer the question: "What's the best way to build a faculty, educate undergraduates, and create a PhD program starting now, focusing on the most important problems of today?"

For those who are interested, we'll be releasing a public version of our strategic plan within about six months. We're trying to do something unique and forward-

thinking, which will hopefully make Princeton an influential member of the statistics, machine learning, and data science communities.

You are organizing the Tukey conference at Princeton (to be held September 18, details here). Do you think Tukey's influence will affect your vision for re-building statistics at Princeton?

Absolutely, Tukey has been and will be a major influence in how we re-build. He made so many important contributions, and his approach was extremely forward-thinking and tied into real-world problems. I strongly encourage everyone to read Tukey's 1962 paper titled The Future of Data Analysis. Here he's 50 years into the future, foreseeing the rise of data science. This paper has truly amazing insights, including:

> 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.
>
> Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation.
>
> By and large, the great innovations in statistics have not had correspondingly great effects upon data analysis.... Is it not time to seek out novelty in data analysis?

In this regard, another paper that has been influential in how we are re-building is Leo Breiman's titled Statistical Modeling: The Two Cultures. We're building something at Princeton that includes both cultures and seamlessly blends them into a bigger picture community concerned with data-driven scientific discovery and technology development.

What advice would you give young statisticians getting into the discipline now?

My most general advice is don't isolate yourself within statistics. Interact with and learn from other fields. Work on problems that are important to practitioners of science and technology development.

I recommend that students should master both "traditional statistics" and at least one of the following: (1) computational and algorithmic approaches to data analysis, especially those more frequently studied in machine learning or data science; (2) a substantive scientific area where data-driven discovery is extremely important (e.g., social sciences, economics, environmental sciences, genomics, neuroscience, etc.).

I also recommend that students should consider publishing in scientific journals or computer science conference proceedings, in addition to traditional statistics journals.

I agree with a lot of the constructive advice and commentary given on the Simply Statistics blog, such as encouraging students to learn about reproducible research, problem-driven research, software development, improving data analyses in science, and outreach to non-statisticians. These things are very important for the future of statistics.