Wharton Department of Statistics and Data Science
Lawrence D. Brown Distinguished Lectures

David Donoho

- Tuesday, April 23, 3:30 – 4:30 pm
  Location: 215 Steinberg Dietrich Hall

Data Science vs Statistics

A conventional narrative tells us that Data Science is just a rebranding of traditional statistics. This talk explores the idea that there is a "Data Science" mindset derived from modern digital life, and a "Statistics Mindset" derived from long intellectual tradition. These mindsets breed two completely different mental realities ie. different thoughts we can hold in mind and pay attention to. Because of this, there is a gaping divide between what residents of each mental reality can focus upon, produce and value. Downstream of this, two completely separate discourses and cultures are developing.

In our view, the two sides don't properly understand that there are two sides, and that severe challenges are caused by this split. This shows up when each camp reflects on the other in frequent frustration, pointless conversations and negative emotions. This can be seen in situations where teams from each side of the divide both do research about the same topic, and try to engage with each other's research.

It is important for the statistics tradition to drop the blinders and see the situation clearly, to finally benefit from the existence of the data science reality. Similarly, data science could make better progress by clearly understanding what the statistics tradition can offer.

Finally, the biggest opportunities will come for those who can become bicultural. The talk should be accessible to a broader audience. This is joint work with Matan Gavish (Hebrew University CS).

- Wednesday, April 24, 12:00 – 1:00 pm
  Location: F60 Jon M. Huntsman Hall

Widespread Panic over Model Collapse

Modern ML systems are extraordinarily data hungry; and some major commercial players are said to now be using synthetic data to train their most ambitious ML systems. Also, AI-generated data will soon flood the internet, perhaps to the point where most available data are synthetic.

Recently ML research has started to confront the larger issues that synthetic data might pose, including a future where most or all of the data available for an ML training are synthetic. A number of ML papers became prominent after promoting the idea of "model collapse", the "curse of recursion", "model autophagy disorder". Featuring experiments and some very basic theoretical argumentation they promoted a storyline where successive recycling of purely synthetic data led to model degeneration.

In contrast, Mathematical Scientists have looked at the same setting as the ML researchers, and developed a more balanced view of the situation, depending on the synthetic data use case, no such collapse occurs. Empirical work with canonical LLMs and diffusion models confirms the absence of collapse, in the recommended use case.

I will review the setting, the narrative promoting to panic and counter narratives leading to a calmer view. The talk should be accessible to a broader audience, as most of the research in this area consists of analysis at the statistics undergraduate major or master's level — although one could transplant the basic questions into fancier settings if one liked.

This is joint work with Apratim Dey of Stanford Statistics and a CS team at Stanford.

- Thursday, April 25, 3:30 – 4:30 pm
  Location: G55 Jon M. Huntsman Hall

Optimal Vector Sensing by Stein Shrinkage

We review some Compressed Sensing theory through the lens of Approximate Message Passing, and minimax decision theory concerning shrinkage estimates. AMP powers up a simple estimator for an elementary problem
into a procedure for a drastically more ambitious problem of compressed sensing. In this case the simple estimator is James-Stein shrinkage and we use it to construct a procedure of multiple measurement vector compressed sensing.

We discuss the State Evolution analysis of this procedure, prove and empirically verify all the predictions of state evolution, and so on. We discuss the state evolution theory of Compressed Sensing in the sense of sparsity-undersampling phase diagram. In the large dimensional limit both in system size and vector size, with Gaussian sensing matrices, James-Stein has a theoretically unimprovable phase diagram and empirically works near-optimally even in low vector dimensions. In particular this is far better than the convex optimization approaches. This is joint work with Apratim Dey (Stanford Statistics).