Introduction to the Sackler Colloquium, **Drawing Causal Inference from Big Data**

Richard Shiffrin, Indiana University

Data is the basis for scientific progress, and causality is the primary way humans come to understand what the data imply.

A sea change in the way this process operates has taken place in recent years: We have developed the ability to produce, measure, collect, and store amounts of data far beyond anything imagined previously.

Byte = 8 bits			(one digit)
Kilobyte:	10**3	1,000	(short story)
Megabyte:	10**6	1,000,000	(a novel)
Gigabyte:	10**9	1,000,000,000	(movie)
Terabyte:	10**12	1,000,000,000,000	(x-rays in one hospital)
Petabyte:	10**15	1,000,000,000,000,000	(US research libraries)
Exabyte:	10**18	1,000,000,000,000,000,000	(SKA telescope/day)
Zettabyte:	10**21	1,000,000,000,000,000,000,000	(all words ever spoken)
Yottabyte:	10**24	1,000,000,000,000,000,000,000,000	(world wide web in 2015-2016?)

We have foreseen the need to have terms for 10**27, 10**30, and 10**33 (so far). Information is doubling every year or two (and the rate might be accelerating).

The problem is not just collecting data from more sources, but also because we collect more precise measurements from a single source--one of my colleagues recently joined a project in which they expect to deal with 1000 terabytes of information from one-thousandth part of a single mouse brain.

Although we are producing and storing ever greater amounts of data, we have just begun to figure out ways to analyze and understand what the data show. The problem is not restricted to science: Business, government, entertainment, social media, security agencies, social networks face the same challenges.

The two main challenges, both unprecedented in scope, are two sides of the same coin:

First, how does one find the important patterns of data?

This subject requires a Sackler Colloquium of its own. Suppose a moderately large data base has a terabyte of data (10**12). This data might perhaps contain a thousand (10**3) measurable factors. The number of correlations of those factors in all combinations would be on the order of 2**(10**3), or about a 300 digit number. The search problem for patterns is enormous.

Second, having found a pattern, how can we explain its causes?

This is the focus of the present Sackler Colloquium. If in a terabyte data base we notice factor A is correlated with factor B, there might be a direct causal connection between the two, but there might be something like 2**300 other potential causal loops to be considered. Things could be even more daunting: To infer probabilities of causes could require consideration all distributions of probabilities assigned to the 2**300 possibilities. Such numbers are both fanciful and absurd, but are sufficient to show that inferring causality in Big Data requires new techniques. These are under development, and we will hear some of the promising approaches in the next two days.

Whatever is developed, I am sure computational algorithms will never be sufficient to answer either question. The numbers go well beyond any conceivable rote computational approach. To me this highlights the importance of models and theories. Models and theories have always been important in science. but in Big Data they will be critically needed to guide the search for patterns and guide the search for causal accounts.

This mention of models leads me to finish with a somewhat high level perspective: In many ways, drawing causal inference from Big Data is **Science Writ Small**: In science we find patterns of data in the real world and in experiments, use models and theories to try to explain and understand them, and use the models to guide further search for patterns and to design new experiments, and then develop new and better tuned theories. These models are in many or most cases designed to provide causal accounts, but it is recognized that the models are approximations, that there are an infinite number of alternative accounts, and that for this and other reasons we prefer models that balance a good fit to the data against model complexity. As the era of Big Data continues, I would expect one important approach to causal explanations to follow the line that science has already developed, a line that often involves experimental tests and 'interventions'.

However, the Big Data era also introduces the need to establish causality when experimental tests and interventions are difficult, impossible, or too far removed from the real world complexities that govern the data. If we collect a petabyte of data about the past 24 hours of world weather, and wish to understand the causes of some important weather pattern, intervention would not be feasible, and laboratory tests might not be highly relevant. In cases like this we would have to try to establish causality from data internal to the data base itself. At the least this would require both assumptions concerning the meaning of causality in large complex recurrent systems (a subject still under debate), and new statistical and computational techniques. I believe all of us are excited to hear promising approaches along these lines in the next two days.