The phrase “Big Data” has greatly raised expectations of what we can learn about ourselves and the world in which we live or will live. It appears to have also boosted general trust in empirical findings, because it seems to be common sense that the more data, the more reliable are our results. Unfortunately, this common sense conception can be falsified mathematically even for methods such as the time-honored ordinary least squares regressions (Meng and Xie, 2014). Furthermore, whereas the size of data is a common indicator of the amount of information, what matters far more is the quality of data. A largely overlooked statistical identity, a potential candidate for the statistical counterpart to the beautiful Euler identity, reveals that trading quantity for quality in statistical estimation is a mathematically demonstrable doomed game (Meng, 2017). Without taking into account the data quality, Big Data can do more harm than good because of the drastically inflated precision assessment, and hence the gross overconfidence, which minimally can give us serious surprises when the reality unfolds, as illustrated by the 2016 US election.


About the Presenter:

Xiao-Li Meng, Dean of the Harvard University Graduate School of Arts and Sciences (GSAS), Whipple V. N. Jones Professor and former chair of Statistics at Harvard, is well known for his depth and breadth in research, his innovation and passion in pedagogy, and his vision and effectiveness in administration, as well as for his engaging and entertaining style as a speaker and writer. Meng has received numerous awards and honors for the more than 120 publications he has authored in at least a dozen theoretical and methodological areas, as well as in areas of pedagogy and professional development; he has delivered more than 400 research presentations and public speeches on these topics, and he is the author of “The XL-Files,” a regularly appearing column in the IMS (Institute of Mathematical Statistics) Bulletin. His interests range from the theoretical foundations of statistical inferences to statistical methods and computation as well as applications in natural, social, and medical sciences and engineering.