In 1967, John Ridker and Richard Henning published a paper in *Review of Economics and Statistics* on the effect of air pollution on home prices. This was a pioneering and influential paper-- one of the first to use hedonic price analysis of property markets to attempt a cost-benefit analysis of a public project. As of this writing it has been cited 981 times. What strikes us now about this paper is that it used data on a mere 167 home sales to construct a regression model of home values on a number of structural and neighborhood characteristics including measures of air quality , to determine that there was a statistically significant relationship between the two variables of interest.

It would be a gross mischaracterization of statistical practice in 1967 to say that a sample of this size was regarded as wholly adequate. Certainly there were some near-contemporary works in real estate analysis such as Hugh Nourse's 1962 work on public housing, and John Kain and John Quigley's 1970 construction of hedonic models, whose sample sizes numbered in the thousands. Nevertheless, the Ridker and Henning's work was published, and the statistical framework and hypothesis testing procedures were thought to be adequately justified.

After all (so the thinking sometimes went), 162 was, in fact, a big sample. "Big" meant mostly that you could use the Z approximation to the t-table, and statistics textbooks provided exact critical values of the t-distribution for sample sizes (more properly, degrees of freedom) only up to 35 or so. Any sample size above that, and one was more or less free to use the Z-distribution's 1.96 as the barrier between "significant" and "insignificant" estimates. Even if a sample of 162 limited our ability to explore the heterogeneity of the effect of bad air on house prices, at least we had a nice estimate of the average effect. Life was simple, and so was the analysis.

We know better now. We knew better then, too, of course, but we lacked the means and opportunities to do better. Four related developments have allowed real estate (and other economic and social) researchers to vastly improve their analytical techniques and allow for deeper exploration, more meaningful insight, and greater confidence in the conclusions than ever before.

The first development was theoretical. The progress of asymptotic theory for data that is heterogeneous and dependent (both spatially and temporally) alerts us to the dangers of oversimplifying the ability of samples of (say) 162 to produce estimators that behave as their asymptotic distributions would indicate when data is not normal. And the problem is worse for data that is generated by power law and similarly ill-behaved distributions. Real estate is a heterogenous asset that can exhibit these kinds of characteristics, as demonstrated by Roger Brown in his 2004 work.

The second development is the development of computational capability. This is evidenced both in the increase in storage capabilities, and in the ability to do perform massive computational tasks. The terabytes of storage and great increase in computational speed that are now easily available on desktop machines could only be vaguely imagined forty years ago. And for even more complex and challenging problems, many have access to supercomputers, whose capabilities are for all practical purposes, unlimited.

The third development is software. From rudimentary beginnings, econometric and other statistical software has developed to the point even beginning researchers have been provided the opportunity to perform sophisticated and complex calculations, and in so doing, produce sturdy and robust inference.

And finally, the fourth development, and the subject of this book, is the creation of databases that are "big". How big is big? Pretty big, it turns out. It is not uncommon for academic studies to use over a million data points, and private data providers have access to much more. The current record for published real estate research is probably held by Liang Peng and Thomas Thibodeau, who use 26 million housing transactions in their 2017 study of unit-specific risk. Undoubtedly bigger ones are coming.

And when they do, Nick Evangeloupos, Andy Krause, Cliff Lipscomb, and Kimberly Winston-Geideman will be there to guide their users. With great computational power comes great responsibility, and the chapters that follow provide a comprehensive field guide to the responsible acquisition, care, cleaning, description and analysis of big real estate databases. There are a lot of issues: legal issues, storage issues, statistical issues, and presentation issues, a number of which make plain that big data is not just bigger versions of "ordinary" databases, but can be different in kind as well. The authors' advice will be helpful for any researcher with any sized database, but their emphasis and concentration on big data in real estate will be particularly helpful for cutting edge research in our field. Real estate research is more important than ever: academic, policy and corporate scientists are more attuned to the special nature of real estate and the important roles it plays in the economy, in people's lives and in financial portfolios. The authors are doing all of us a tremendous service.