Manifold learning algorithms aim to recover the underlying low-dimensional parametrization of the data using either local or global features. It is however widely recognized that the low dimensional parametrizations will typically distort the geometric properties of the original data, like distances and angles. These impredictible and algorithm dependent distortions make it unsafe to pipeline the output of a manifold learning algorithm into other data analysis algorithms, limiting the use of these techniques in engineering and the sciences.

Moreover, accurate manifold learning typically requires very large sample sizes, yet existing implementations are not scalable, which has led to the commonly held belief that manifold learning algorithms aren't practical for real data.

This talk will show how both limitations can be overcome. I will present a statistically founded methodology to estimate and then cancel out the distortions introduced by any embedding algorithm, thus effectively preserving the distances in the original data. And I will demonstrate that with careful use of sparse data structures manifold learning can scale to data sets in the millions. Both points will be exemplified in the exploration of data from a large astronomical survey.

Joint work with Dominique Perrault-Joncas, James McQueen, Jacob VanderPlas, Zhongyue Zhang, Grace Telford