Estimating the amount of distinct elements in a dataset by examining only a fraction of the data is known to be a hard problem, both theoretically and in practice. Our work explores a breakthrough theoretical result by Valiant and Valiant from 2011 that presents a provably accurate method for doing such estimations.

Our goal is to put this theory into practice for the important task of estimating the deduplication ratio of a very large dataset. However, deploying this technique in a real world setting runs into significant obstacles.

In the talk I will describe new techniques that help bridging the gap and enable the use of this exciting approach. Our work achieves a major improvement over the current state of the art practical solutions.

The talk is for a general audience, no prior knowledge is assumed.

Based on joint work with Dmitry Sotnikov and Ety Khaitzin that appeared at Usenix FAST 2016.