Big Data Analytics 2013

Invited talk

From Terabytes to Megabytes: Finding the needle by shrinking the haystack

Sergei Vassilvitskii, *Google*

We survey some recent results on MapReduce algorithmics and show that an effective plan of attack relies on reducing the input size from the realm of "Big" Data to that of "Small" Data. This reduced problem can then be effectively solved in memory on a single machine. Uniform sampling is the classical such approach; in this talk we focus on extensions beyond vanilla sampling, and show how biased sampling, and its extreme form of outright pruning can be used to create a manageably sized approximate coreset for the original problem.