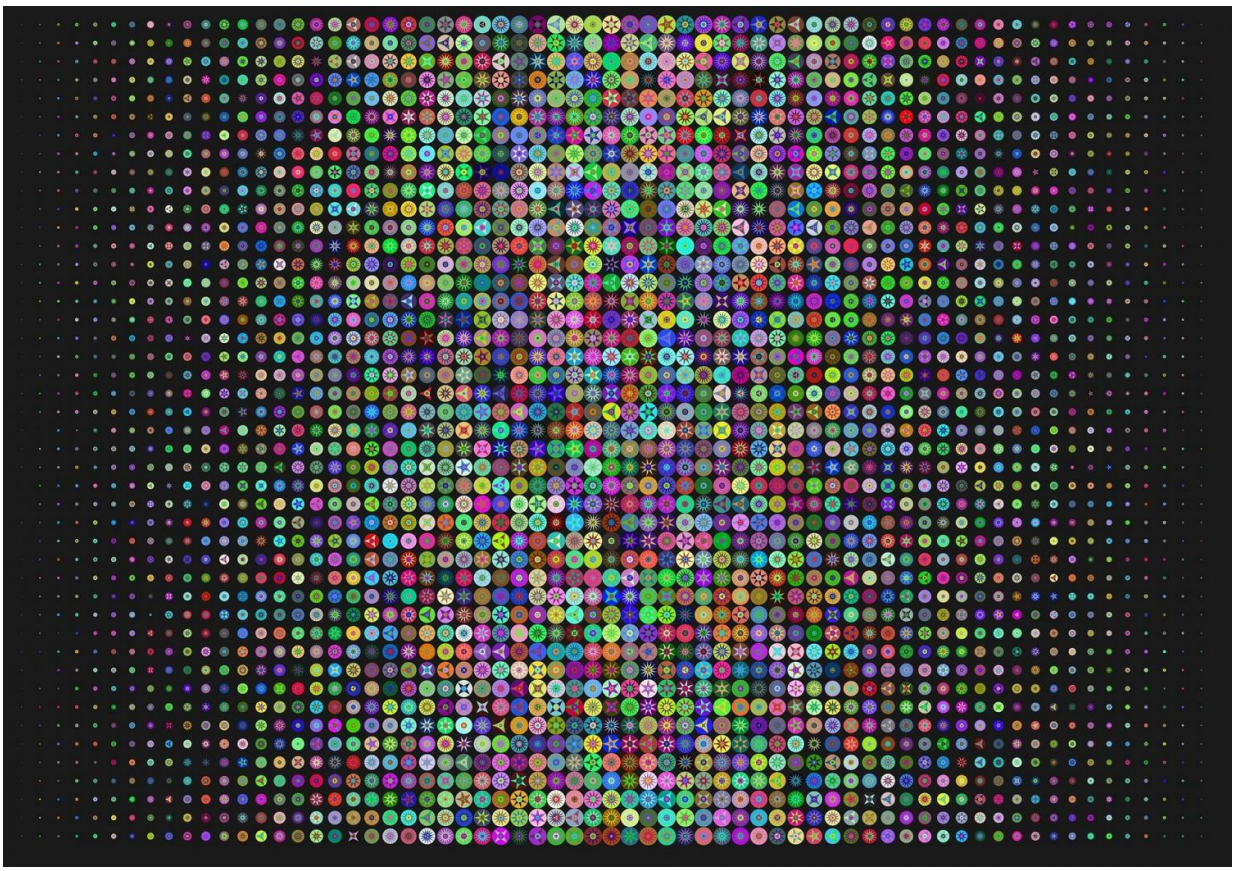
Computer scientist demonstrates 30-year-old theorem still best to reduce data and speed up algorithms

October 19 2017



Credit: CC0 Public Domain

When we think about digital information, we often think about size. A

daily email newsletter, for example, may be 75 to 100 kilobytes in size. But data also has dimensions, based on the numbers of variables in a piece of data. An email, for example, can be viewed as a high-dimensional vector where there's one coordinate for each word in the dictionary and the value in that coordinate is the number of times that word is used in the email. So, a 75 Kb email that is 1,000 words long would result in a vector in the millions.

This geometric view on data is useful in some applications, such as learning spam classifiers,but, the more dimensions, the longer it can take for an algorithm to run, and the more memory the algorithm uses.

As data processing got more and more complex in the mid-to-late 1990s, computer scientists turned to pure mathematics to help speed up the algorithmic processing of data. In particular, researchers found a solution in a theorem proved in the 1980s by mathematics William B. Johnson and Joram Lindenstrauss working the area of functional analysis.

Known as the Johnson-Lindenstrauss lemma (JL lemma), computer scientists have used the theorem to reduce the dimensionality of data and help speed up all types of algorithms across many different fields, from streaming and search algorithms, to fast approximation algorithms for statistical and linear algebra and even algorithms for computational biology.

But as data has grown even larger and more complex, many computer scientists have asked: Is the JL lemma really the best approach to pre-process large data into a manageably low dimension for algorithmic processing?

Now, Jelani Nelson, the John L. Loeb Associate Professor of Engineering and Applied Sciences at the Harvard John A. Paulson

School of Engineering and Applied Sciences, has put that debate to rest. In a paper presented this week at the annual IEEE Symposium on Foundations of Computer Science in Berkeley, California, Nelson and co-author Kasper Green Larsen, of Aarhus University in Denmark, found that the JL lemma really is the best way to reduce the dimensionality of data.

"We have proven that there are 'hard' data sets for which dimensionality reduction beyond what's provided by the JL lemma is impossible," said Nelson.

Essentially, the JL lemma showed that for any finite collection of points in high dimension, there is a collection of points in a much lower dimension which preserves all distances between the points, up to a small amount of distortion. Years after its original impact in functional analysis, computer scientists found that

The JL lemma can act as a preprocessing step, allowing the dimensions of data to be significantly reduced before running algorithms.

Rather than going through each and every dimension—like the hundreds of dimensions in an email—the JL lemma uses a system of geometric classification to speed things up. In this geometry, the individual dimensions don't matter as much as the similarities between them. By mapping these similarities, the geometry of the data and the angles between data points are preserved, just in fewer dimensions.

Of course, the JL lemma has a wide range of applications that go far beyond spam filters. It is used in compressed sensing for reconstructing sparse signals using few linear measurements; clustering high-dimensional data; and DNA motif finding in computational biology.

"We still have a long way to go to understand the best dimension

reduction possible for specific data sets as opposed to comparing to the worst case," said Nelson. "I think that's a very interesting direction for future work. There are also some interesting open questions related to how quickly we can perform the dimensionality reduction, especially when faced with high-dimensional vectors that are sparse, i.e. have many coordinates equal to zero. This sparse case is very relevant in many practical applications. For example, vectors arising from e-mails are extremely sparse, since a typical email does not contain every word in the dictionary."

"The Johnson-Lindenstrauss Lemma is a fundamental result in high dimensional geometry but an annoying logarithmic gap remained between the upper and lower bounds for the minimum possible dimension required as a function of the number of points and the distortion allowed," said Noga Alon, professor of Mathematics at Tel Aviv University, who had proven the previous best lower bound for the problem. "The recent work of Jelani Nelson and Kasper Green Larsen settled the problem. It is a refreshing demonstration of the power of a clever combination of combinatorial reasoning with geometric tools in the solution of a classical problem."

Provided by Harvard John A. Paulson School of Engineering and Applied Sciences