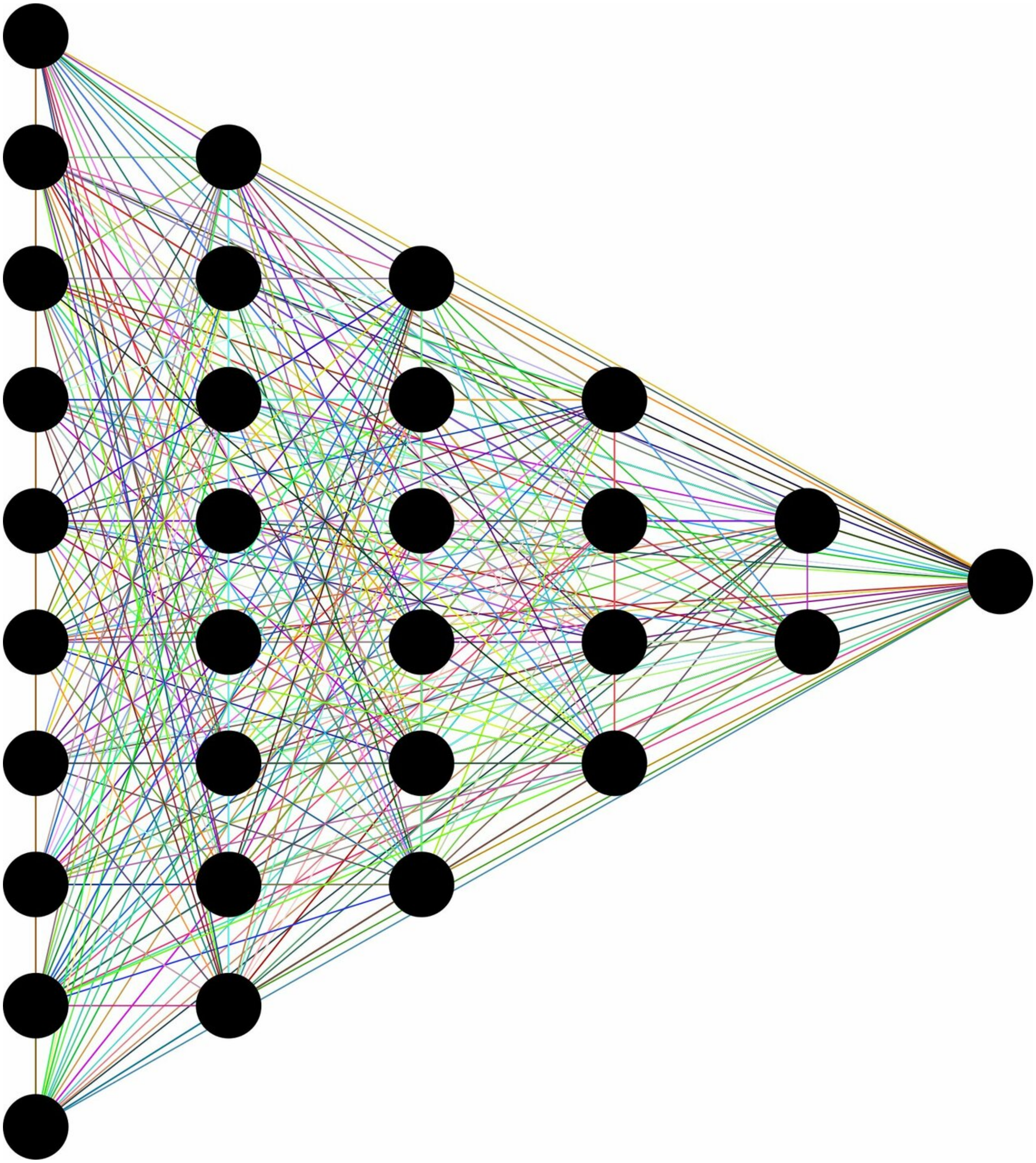
Powerful machine-learning technique enables biologists to analyze enormous data sets

March 18 2019



Credit: CC0 Public Domain

Researchers at A*STAR have compared six data-analysis processes and come up with a clear winner in terms of speed, quality of analysis and reliability. The top performer took large, complex biological data sets and spat out key relations between parameters (such as grouping blood and marrow cells according to cell type) in a fraction of the time of the other techniques.

Measurements on single cells alone can generate huge data sets that have anywhere from 20 to more than 20,000 parameters. The mind-boggling size and complexity of biological data sets make it extremely challenging for scientists to uncover meaningful relationships between parameters.

Mathematicians have developed statistical techniques that simplify complex data sets by grouping data according to their similar characteristics. The most well-known technique is principal component analysis (PCA), which was developed in the early twentieth century. Recently, more powerful techniques, that harness the power of machine learning, have been developed.

Now, Evan Newell and Florent Ginhoux at the Singapore Immunology Network (SIgN), and their colleagues have used single-cell data to test six such machine-learning techniques and discovered one that stands out from the rest in terms of speed, quality of analysis and reliability. This technique is called the uniform manifold approximation and projection, or 'UMAP'.

"When Evan and Etienne Becht in his group at SIgN started to benchmark UMAP, we realized that it was much more powerful than anything we had used before," recalls Ginhoux.

An analysis that might take days using other methods can be done in a few hours using UMAP, which will allow scientists to investigate larger data sets. "With UMAP, we can analyze data for two or three million cells, whereas we generally avoid going beyond 100,000 cells with other methods," says Newell.

UMAP grouped similar cells in the most intuitive way, making it easier to interpret its results.

"I think it's really groundbreaking," says Ginhoux. "Researchers I meet at conferences are already starting to use it."

In an earlier study, the group demonstrated UMAP's power by using it to discover a new population of cells in blood. Newell notes that UMAP is highly versatile and can be applied to data generated in fields as diverse as astronomy and crystallography. "Basically, any data that can be expressed in matrices can be analyzed by UMAP," he says.

In addition to using UMAP to analyze data on a daily basis, the team plans to continue to work with informaticians to tailor UMAP to their needs.

More information: Etienne Becht et al. Dimensionality reduction for visualizing single-cell data using UMAP, *Nature Biotechnology* (2018). DOI: 10.1038/nbt.4314

Provided by Agency for Science, Technology and Research (A*STAR), Singapore

Citation: Powerful machine-learning technique enables biologists to analyze enormous data sets (2019, March 18) retrieved 27 April 2024 from https://phys.org/news/2019-03-powerful-machine-learning-technique-enables-biologists.html