2018 Research Interest/Project Ideas

David Gleich

https://www.cs.purdue.edu/homes/dgleich
https://github.com/dgleich

==========

Building mass-spec databases of ever-changing consumer products.

Most consumer products companies work hard to produce a reliable and constant product as the constituent ingredients of those products continually shift based on slight differences in materials sourcing. As an example, the exact composition of the plastic varies slightly by batch. Moreover, there is often no requirement to update consumers as the composition changes. This can make understanding your own purchases a complicated procedure! If, say, a shampoo or soap starts causing itchy skin, it may be that the underlying composition has changed.

In biology, there has been a large effort to make data open and available (protein, gene, and pathway databases proliferate). Recently, there has been some effort to make open mass-spectrometry data available -- although this has largely focused on chemically pure compounds.

The idea in this project is to perform spectral analysis of various consumer products and create a database with their contents. One of the critiques of mass-spec data is that it is highly sensitive to the exact details of processing. By having a large database of closely related samples, this provides a unique opportunity to embrace this variability rather than attempt to control it. (This is similar to the idea underlying fast genetic sequencing ... we now get an extremely large number of lower-quality information and use algorithms to assemble the overall picture.) With this database, then we can deconvolve their composition using existing (pure) standard databases. This project will have to deal with statistical distributions over real-world measurements, measurement outliers, tracking composition over time. Again, this is standard analysis within various commercial labs around the world, although it is typically done for diagnostic or competitive. The key difference here will be to put the data on the web for use across various data science curriculum and projects, and to study the idea of boosting lower-quality data with algorithmics.