Algorithms helpful in deducing interconnections among biological processes in the cell continue to advance (Goldstein 2013, Radivojac 2013). Researchers do not always update their processes at every advance, sometimes using less effective methods (Radivojac 2013). Software for inferring functions from gene expression data is available in source code at DAVID and at Gene Ontology, but some of this available software might be less elaborate work.

Moving forward, methods can be improved through the use of Hidden Markov Models (HMM) and Privileged Information. Synthetically generated data is one means to determine whether an algorithm can infer, from limited data, what we know. The challenge for these methods was finding ways to integrate disparate data sources and properly handle incomplete and noisy data. (Radiovajc et al.)

In this study, we attempted to obtain the source code for the curated data. The classifier can classify sets of reads, even when those reads were not accompanied by as much qualification as was present when the classifier was built.

Top-performing methods in the future will be based on well-founded principles of statistical learning and inference.

**Starting Point:** CAFA 2013 (Radivojac et al.)
- Molecular Function -- Accurate
- Biological Process -- the overall performance of the top-scoring methods was below expectations.

**Top-performing methods in the future will be based on well-founded principles of statistical learning and inference.**

---

**Methods:**

**IDENTIFICATION METHODS**
We explored the tool categories, and for relevant categories, the tools listed at the DAVID and GO/Neurolex websites. Many of these tools have components, only some of which are performing biological process applications of classification or clustering. For tools, or tool components, related to classification or clustering, we attempted to obtain the source code.

**MODIFICATION METHODS / CONSTRAINTS**
"The challenge for these methods was finding ways to integrate disparate data sources and handle incomplete and noisy data." (Radiovajc et al.)

We do not wish to improve the integration of disparate data sources, and handle incomplete and noisy data, but also wish to do this without impacting the way researchers are using their tool chains, that is, we would like to give better answers for the same processing steps, if possible. Existing tool chains apply constraints not only in terms of the input output parameters, but also user expectations upon the amount of time a process incurs.

**EVALUATION**
Synthetic data is one means to determine whether an algorithm can infer, from limited data, what we know. Methods used by DREAM are to be investigated.

---

**Major Conclusions:**
- Algorithms helpful in deducing interconnections among biological processes in the cell continue to advance (Goldstein 2013, Radivojac 2013).
- Software for inferring functions from gene expression data is available in source code at DAVID and at Gene Ontology, but also at researcher’s own sites (referenced by articles), and at organizations, such as Harvard/MIT Broad Institute.
- Researchers do not always update their processes at every advance, sometimes using less effective methods (Radivojac 2013).
- Some of this available software might be less effective than methods (Radivojac 2013).

---

**References:**

---

**Learning Using Hidden / Privileged Information:**
Machine learning, in the manner of support vector machines, that uses data is available to develop the classifier, can classify sets of reads, even when those reads are not accompanied by as much qualification as was present when the classifier was built. For our case, the privileged information has to be present in the curated data. Perhaps there exists a difference between more elaborate and more more elaborate experimentation, and the classifier can be derived from the more elaborate work, and applied to the less elaborate work.