The Automatic Scientist will be a Data System

Stratos Idreos

Harvard University

1. VISION: AUTOMATIC SCIENTIST

For thousands of years science happens in a rather manual way. Mathematics, engineering and computer science provide the means to automate some of the laborious tasks that have to do with computation, data collection and management, and to some degree predictability. As scientific fields grow more mature, though, and scientists over-specialize a new problem appears that has to do with the core of the scientific process rather with the supporting steps.

It becomes increasingly harder to be aware of all research concepts and techniques that may apply to a given problem.

A critical step when generating new concepts and explaining existing phenomena is being aware of existing concepts/techniques and how they can be combined. This is true in a single scientific area at a time but it is exacerbated across areas, i.e., we can easily miss opportunities to leverage advances across unfamiliar areas. In addition, this is important when we try to explain a phenomena, e.g., in medicine when a doctor tries to match an existing patient to the vast space of possible conditions or when a surgeon tries to decide what the best next step should be in real time during a surgery. Computer science already helps in significant ways. Recent advances include automation in feature engineering [1] and hypothesis generation [2] by analyzing past research. What we propose here is to take one step further and map existing research concepts (at a fine granularity) in a unified model and then treat that as a data management and analytics problem.

We envision a future when scientists rely on software tools not only to collect, manage and analyze data but also to generate ideas, make it easy to solve problems by combining existing research concepts across one or more areas, get suggestions about possible next steps, and find errors, semi-automating the scientific process. In the same way that modern data systems organize the world's data and facilitate data processing in numerous areas in businesses and sciences, we envision that "auto-science data systems" will be able to abstract and speed up the scientific process across numerous scientific fields. Auto-science systems are "scientist-in-the-loop" environments that leverage intuition and experience of scientists along with automation provided by the tools for the parts of the scientific process that rely on combining/tuning past work in novel ways.

2. RESEARCH PATH

For the research path below, assume automatic generation of data system architectures as an example of automating research and testing of new ideas.

Modeling Research. Similar to the relational model, we need a unified model that is generic enough to capture the necessary information needed to describe basic research concepts and their combinations. Such a model should capture research concept characteristics, such as when it applies, its dependencies, cost metrics and side-effects, and tuning parameters.

Declarative Research Processing. In addition, we need domain specific languages that allow us to browse existing research concepts, synthesize more complex ones and run what-if questions on top of the possible designs space. These new class of queries will be tailored to speed up research by leveraging the structure of research concepts e.g., find how we can best solve problem X (specific requirements *R*) given existing research concepts that can be combined to satisfy *R* or find the weakest link in a research solution that can be described as a combination of concepts.

Searching for the Right Design. The next step is to develop algorithms that can automatically iterate over possible solutions given a problem and a research library. Being able to estimate the impact of a research concept once applied to a solution is key. For example, in the example of automating data systems research, online code generation and a modular architecture can help rank solutions. Randomized iterative algorithms are key to be able to solve problems in little time and potentially also use such approaches in scenarios where quick adaptivity is important.

Summary. The auto-scientist is a data system that stores research concepts and allows scientists to interactively navigate the research space to speed up research. It also brings opportunities in education of young scientists and verifiability of research solutions. It resembles in many ways the path that the database community followed to develop systems for data management - only this time this is an effort that has to happen in collaboration with the target scientific fields to develop the right models and expressibility. In addition, we expect many research challenges to be analogous to database systems, e.g.: Can we indeed have a single model or we need multiple models which brings data integration problems? Do we need polystores that store concepts from different fields? Our community has several existing ideas to seed such a research landscape.

3. REFERENCES

[1] M. R. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, and C. Zhang. Brainwash: A Data System for Feature Engineering. In *CIDR*, 2013.

[2] M. Nagarajan, A. D. Wilkins, B. J. Bachman, I. B. Novikov, S. Bao, P. J. Haas, M. E. Terrón-D\'\iaz, S. Bhatia, A. K. Adikesavan, J. J. Labrie, S. Regenbogen, C. M. Buchovecky, C. R. Pickering, L. Kato, A. M. Lisewski, A. Lelescu, H. Zhang, S. Boyer, G. Weber, Y. Chen, L. A. Donehower, W. S. Spangler, and O. Lichtarge. Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. In *SIGKDD*, 2015.