Intelligent Data Analysis in the 21st Century

Paul Cohen\textsuperscript{1} and Niall Adams\textsuperscript{2}

\textsuperscript{1} University of Arizona
cohen@cs.arizona.edu
\textsuperscript{2} Imperial College London
n.adams@imperial.ac.uk

Abstract. When IDA began, data sets were small and clean, data provenance and management were not significant issues, workflows and grid computing and cloud computing didn’t exist, and the world was not populated with billions of cellphone and computer users. The original conception of intelligent data analysis — automating some of the reasoning of skilled data analysts — has not been updated to account for the dramatic changes in what skilled data analysis means, today. IDA might update its mission to address pressing problems in areas such as climate change, habitat loss, education, and medicine. It might anticipate data analysis opportunities five to ten years out, such as customizing educational trajectories to individual students, and personalizing medical protocols. Such developments will elevate the conference and our community by shifting our focus from arbitrary measures of the performance of isolated algorithms to the practical, societal value of intelligent data analysis systems.

Each time we hold an IDA conference, a distinguished conference committee thinks hard about a theme and a distinguished researcher writes a keynote lecture about what Intelligent Data Analysis is or might be. We suspect that all this hard thinking does not influence the kinds of papers we receive. Every conference season we review and accept roughly the same kinds of papers as appear at the data mining and machine learning conferences.

The subject of the conference should not be a fifteen year old vision of intelligent data analysis, nor should the subject default to a sample of current work in data mining and machine learning. The conference should provide a venue for future interpretations of Intelligent Data Analysis. We should start publishing in areas that are developing now and will reach full bloom in five years. At the same time, we should stay true to the traditional goals of the IDA conference.

The first symposium on Intelligent Data Analysis was organized by Xiaohui Liu and held in Baden-Baden in 1995, the same year as the first International Conference on Knowledge Discovery and Data Mining. Professor Liu’s idea was that data analysis, like other kinds of human expert problem solving, could be done by computers:

[Computers] should also be able to perform complex and laborious operations using their computational power so that the analysts can
focus on the more creative part of the data analysis using knowledge and experience. Relevant issues include how to divide up work between human and computer; how to ensure that the computer and human stay “in synch” as they work on parts of a data analysis problem; how to seamlessly integrate human domain and common sense knowledge to inform otherwise stupid search procedures such as stepwise regression; how to present data so human eyes can see patterns; how to develop an integrated data analysis environment...

To a remarkable extent, these issues have been addressed and Liu’s vision of automated data processing has been achieved. Computers do perform complex and laborious operations, we have integrated data analysis environments (such as R [7]) and packages of algorithms (such as WEKA [9]). The community has settled on a small collection of common “generic tasks,” such as prediction, classification, clustering, model selection, and various kinds of estimation. These tasks are more specific than “exploring a dataset,” and yet are general enough to cover data from disparate domains such as finance and marketing, biology and ecology, psychology and education. Less progress has been made toward integrated, knowledge-intensive, human-computer systems, but, as we suggest later in this paper, this goal might not be as desirable as we once thought. If the IDA community has achieved most of its goals and abandoned those it cannot achieve, what remains to be done? IDA is today a data mining conference, and data mining has achieved the kind of maturity that produces only incremental progress. What will our next challenges be?

We will address two perennial answers to this question before turning to some new challenges.

1 Autonomous Expert Data Analysis

The kinds of autonomy we see in data analysis today are not the kinds we anticipated in 1995. Following the “knowledge revolution” and the widespread commercialization of expert systems, we expected intelligent data analysis systems to attack data sets with the strategies of expert human data analysts. Few such systems were built. It is worth reviewing one of them, built by Rob St. Amant as part of his PhD research, to understand why there are not more systems like it.

AIDE was a mixed-initiative planner for data exploration, meaning it and a human user could explore a dataset together, with AIDE sometimes following the user’s lead and sometimes striking out on its own [23]. Its knowledge about data analysis was stored in plans and control knowledge. Plans contain sequences of data-processing actions, as well as preconditions and postconditions. In general, preconditions specify when a plan can be executed, not when it should be. When more than one plan applies, control knowledge ranks them. It is the job of control knowledge to make the data analysis follow a coherent path, rather than jumping around (which would be hard for users to follow). Of course, a human user can direct AIDE to do anything he or she pleases, so some control knowledge pertains to inferring or anticipating the user’s focus.