Intelligent Data Analysis in the 21st Century

Paul Cohen[1], Niall Adams[2]

[1] University of Arizona cohen@cs.arizona.edu
[2] Imperial College London n.adams@imperial.ac.uk

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...[6]

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," [1] 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 [2, 3]. 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.

At the time, it made sense for AIDE to be a domain-independent data analyst, so its plans and control knowledge referred to the form of data, not to its content or what it represents. AIDE was based on the assumption that the random variables in a dataset were actually connected by causal or functional relations, and the job of data analysis is to build one or more models of the variables and their relations — one or more graphs, if you will, in which nodes represent variables and arcs represent relations. AIDE was not a system for analyzing financial data, or intelligence analysis, or phenotyping, all of which are today aided by domain-specific tools. It was instead a system for general purpose data analysis. It was shown to improve data exploration, in the sense that users of AIDE were able to explore more of a dataset and figure out the relationships between variables more thoroughly and accurately than users without AIDE. This test was appropriate for a domain-independent data analysis tool, but it was not realistic: Generally one approaches a dataset with particular questions in mind, not looking for all significant relations between variables.

Although AIDE did a lot of things right — mixed initiative exploration, explicit plans and control knowledge, a clear user interface, and plenty of data analysis functionality — it could never work in practice. Its understanding of data (random variables in functional or causal relations) and the user's intentions was too weak to be a basis for really focused and intelligent analysis. If intelligent data analysis means something like expert systems for data exploration, then these systems will have to be a lot more expert. They will have to specialize in financial fraud data, or climate change data, and so on.

Another reason that programs like AIDE did not start a revolution might be that they did little to help human analysts. We have been here before: At first blush, stepwise multiple regression seems like an ideal use of computer power, but in practice, instead of letting the machine explore huge numbers of regression models, most analysts prefer to build them by hand. Ignorant of what is being modeled, stepwise multiple regression blunders through the hunt and rarely brings home a tasty model; or it brings home too many, overlapping, similarly performing models because it hasn't the analyst's knowledge to rank them. AIDE was more intelligent than stepwise multiple regression, but it probably wasn't intelligent enough to make analysts relinquish the creative parts of data analysis they find relatively easy and enjoyable. Much of its knowledge and intelligence was for planning analyses and working alongside analysts. The closest thing we have seen to this functionality in commercial systems is the graphical language in SPSS's PASW Modeler, which allows analysts — but not the system — to plan and program workflows for their analyses. By analogy, most cooks are happy for assistance with chopping, stirring, mashing, scrubbing and cleaning, but few cooks want help with inventing, menu planning, tasting, or observing the pleasure their food brings to others; especially if the "help" results in an inferior meal.

If past is prologue, we should not expect the original vision of IDA, as exemplified by AIDE, to be productive in future. Modern data mining software can chop, mash and scrub data, but we should not expect sophisticated analysis — beautiful models, new discoveries, finding parallel phenomena in disparate datasets, clever workflows, detection of semantic anomalies, and the like — until we make semantical reasoning about data a priority. By semantical reasoning, we mean reasoning about the phenomena that data represents. Although it has been a theme of every IDA conference, the IDA community evidently reserves semantical reasoning about data for human data analysts, and very little has been done to make machines that are capable of reasoning deeply about the content of data. Of course, this attention to form instead of content characterizes virtually all of Artificial Intelligence.

2 Challenge Problems

Periodically, IDA considers creating a community around challenge problems. This has worked well in some fields, particularly robotic soccer and robotic autonomous vehicles, and less well in others, notably the KDD Cup. Why are some challenges successful and others less so? The organization that runs the annual robotic soccer competitions has a fifty year goal: *By the year 2050, develop a team of fully autonomous, humanoid robots that can win against the human world soccer champion team.* [8] Progress toward this goal is steered by an expert committee that changes the rules of the competition and adds new intermediate challenges every year. These changes are monotonic in the sense that researchers can base next year's work on last year's work, and new competitors can join relatively easy by adopting state of the art technology. Robotic soccer has a relatively low cost of admission. It is enormously popular and captures the hearts and minds of participants and the general public. Bragging rights go to individual teams, but the biggest winner is the community as a whole, which gains new members and makes steady, impressive progress every year.

One methodological strength of robotic soccer is that individual algorithms matter less than complete soccer-playing systems. The same was true in St. Amant's AIDE system. While it is desirable to include the fastest and most accurate algorithms available, the marginal benefits of better algorithms will often be negligible, especially over many datasets. Systems like AIDE introduce some pragmatism into the evaluation of new algorithms. Perhaps your changepoint analyzer, or decision tree inducer, or association rule miner is slightly faster or more accurate than last year's model on datasets of your own choosing, but would a data analyst notice a difference if your new algorithm was substituted for an old one in AIDE? Would the analyst be more productive? Admittedly, this question might be harder to answer than a simple evaluation of speed or accuracy, but it is the real usability question, whereas speed and accuracy are only proxies for usability.

The KDD Cup does not encourage the development of complete systems, nor does it steer the community toward any long-term goal [5]. The problems it poses each year are important (several have been medical problems), but each is narrowly defined by a data set and performance targets. These problems emphasize high performance for individual algorithms, and do not require the development of systems that are complete in any sense. Nor is a progression of data mining capabilities apparent in the choice of KDD Cup problems. One has the sense that the problems could have been offered to the community in a different order with essentially the same results. The state of the art isn't being steered.

3 New Challenges for Intelligent Data Analysis

The IDA community should pick one or more significant problems that depend on intelligent data analysis, set a goal for a decade or longer in the future, and steer ourselves to achieve it. The choice of problems should mirror our aspirations for intelligent data analysis. Good problems will be bigger than individuals or small groups of human analysts can manage. They will feature every aspect of data analysis: acquisition, cleaning, storage, markup, analysis, visualization, and archiving and dissemination of results. They will require every kind of reasoning about data and the algorithms that process it: reasoning about provenance, design of workflows, and interpretation of results. Most importantly, good challenge problems will require machines to think about the phenomena that data represent — to think about the content or semantics of data. Here are some examples:

The Scientific Discovery Challenge. By the year 2030, a computer program will make a significant scientific discovery (indicated by publication in Science or Nature, or a comparable venue, or by the granting of an important patent). To qualify, the program will have to formulate a theory, direct the search for evidence, analyze the data, and explain the theory and supporting evidence in a formal language. Natural language understanding is not a necessary part of the challenge problem, but the ability to find and reason about relevant knowledge in the literature might be valuable. To demonstrate that it is more than an assistant to a human scientist (who does the hard intellectual work) the program will have to pass some tests that any human scientist is expected to pass; for example, the program should be able to say whether and why a hypothetical result is consistent or inconsistent with its theory.

A specific scientific discovery challenge could be to automatically construct a gene regulatory network for an important cascade or developmental process. Increasingly, biologists turn to modeling techniques that are familiar to us in the IDA community: Stochastic processes, Bayesian networks, the Viterbi algorithm and related methods. Today, there are only a few examples of automated construction of gene regulatory networks, and those are single algorithms, rather than systems that hypothesize a model by integrating the results of multiple data mining algorithms, and gather data to support or contradict it.

The Global Quality-of-Life Monitoring System. By 2060, every living human will have some kind of communication device that is at least as powerful, computationally, as today's most advanced cell phones. Each human will be an intelligent source of data. IDA can contribute to the infrastructure and algorithms necessary to provide real-time, high-resolution fusion of the data in service of research and policy. Challenges are to monitor habitat loss, species redistribution, erosion, water quality and management, epidemics and pandemics, and other consequences of climate change and population growth.

A specific, relatively near-term challenge problem might go like this: Model the population dynamics and food webs of foxes in London given data from a dedicated web site where residents can report sightings, indicators (e.g., scat, prints), and behaviors of foxes and other predators and prey. To qualify, a computer program would have to demonstrate autonomy in several areas of intelligent data analysis: Knowing what to do with new data (e.g., where to put it in a data set, cleaning it, handling missing values, flagging anomalies); directing data-gathering resources in an efficient way; planning a workflow of operations to build a specified kind of model (e.g., a food web) and estimate its parameters; flagging parameter values, data values, or model predictions that are unusual or anomalous (e.g., if the model predicted a huge jump in the fox population of Islington).

Similar similar challenges might involve automated modeling of human population dynamics, modeling effects of climate change through widely distributed monitoring of spring blooming and densities of species, and modeling evaporation (see below).

The Personalized Protocol. Some of the most important activities in life are controlled by protocols and standards that make little if any provision for differences between individuals. Procedures in intensive care, chemotherapy protocols, other kinds of drug therapy, primary and secondary education, even the mutual funds in which we invest, are quite generic, and rarely are tailored to particular individuals. However, there are hints that personalized protocols are both effective and profitable. Designer drugs, gene therapies, and "lifecycle" investments (which adjust portfolio parameters according to one's age and closeness to retirement) are examples of protocols that are personalized to some degree. Personalization requires data, of course, and the whole idea of personalizing medical care or education might fail because of legal and ethical challenges. But let us imagine that information technologies will permit us to both gather data and protect the rights of the people who provide it. Then we can envision a Personalized Protocol Challenge: To personalize any high-value procedure to maximize its utility to any individual.

As stated, this sounds like a planning problem or a sequential decision problem to be optimized by policy iteration, reinforcement learning, or a related optimization method. However, the difficult work is not to run one of these algorithms but to design the state space and objective function to be optimized. In terms of the old dichotomy between model specification and model estimation, data mining is pretty good at the latter, but the former is still in the realm of intelligent data analysis, and, thus, is a fit challenge for us.

These challenges have several things in common: Each is significant and could affect the survival of species, including our own. Each assumes enormous amounts of data generated by distributed sources. The Large Hadron Collider is a point-source of high-quality data, whereas the world's citizens are a distributed source of variable-quality data. Consequently, we will need new research in gathering, cleaning, and fusing data, all of which arrives asynchronously, yet must sometimes be integrated to construct a temporal or developmental story. Each of the challenges stresses our ability to model systems of dependencies, whether they are gene regulatory systems, ecologies, or social systems. Survival in the modern world will require a new science of complex systems, and statistical methods for discovering and estimating the parameters of these systems. The intelligent data analysis community should not sit this one out, but should take advantage of new opportunities for research and development.

Each of the challenges requires the IDA community to be more outward-looking, less concerned with the arcana of algorithms, more concerned with helping scientists ensure our future well-being. To respond effectively to any of the challenges, we will have to think more about data provenance, metadata and search for data, reasoning about the content or meaning of the data, user interfaces, visualizations of results, privacy and other ethical issues, in addition to the algorithmic research we usually do.

If we adopt a challenge that uses humans as data sources, such as the Global Quality-of-Life Monitoring System, then we should recognize that humans are both producers of raw data and consumers of knowledge, and the same communication infrastructure that supports data capture can support knowledge distribution. Said differently, there will be opportunities to engage people in science and social science, to educate them, to empower them to influence policy, and to create a sense that communities cross national borders. At the University of Arizona this reciprocal relationship between scientists and citizens is called *citizen science*. The science of evaporation provides a nice illustration. Evaporation matters in Arizona and in the arid lands that comprise 30% of the Earth's land surface. (If current models of climate change are correct, this proportion will increase dramatically.) More than 90% of the water that reaches the Sonoran Desert in Arizona evaporates back into the atmosphere, but scientists don't know how plants affect this process. There are no good analytical models of evaporation. At the Biosphere 2 facility, scientists are studying empirically how plant density and configurations affect evaporation, but it is slow work, and the number of factors that affect the results are daunting. Recently, our colleague Clayton Morrison, working with the Biosphere 2 scientists, built a version of their experiment to be run by children in classrooms in Tucson [4]. The kids benefit by being engaged in real science, run by local scientists, on locally important issues. The scientists benefit from data collected by the kids. And those of us who live in arid lands benefit from the resulting science.

How can IDA participate in citizen science? A natural role for IDA is to help scientists transform data into knowledge. In the evaporation experiment, for instance, this means transforming spatial and temporal data into models of evaporation that account for complex dependencies between types and distributions of plants and soils. Another natural role for IDA is to optimize the tradeoff between the quality and quantity of data. Having monitored the generation every data point in the Tucson classrooms, we know that children produce lower-quality data than trained scientists do. But there are many children and relatively few trained scientists. We look forward to new methods for cleaning, censoring, and otherwise editing enormous data sets that are a bit grubbier than most scientists are used to.

Whether IDA runs challenge problems of the kind we described earlier, or remains a conventional conference, it should expand its view of the field. What has been an algorithms conference should become a systems conference, where the systems typically will have components for data gathering, data processing, and disseminating results, and are built to solve problems that matter to society. We should encourage papers on crowd sourcing, social network analysis, experimental economics, new data markup schemas, mobile education, and other topics in the general areas of data gathering, data processing and disseminating results. We might accept papers on the ethics of semiautomated decision-making (if a credit scoring system misclassifies you, who is responsible, and what are the legal and ethical considerations?). We should particularly value papers that demonstrate reciprocity between citizens and scientists. The conference should recognize that systems may be harder to evaluate than algorithms, especially when these systems include humans as data sources or expert data processors, and it should adjust reviewing criteria accordingly. New criteria should reward autonomous and mixed-initiative analysis; integration of analysis with data management, workflow management, and new ways to present and justify results; integration of multiple data sets from different sources within analyses; and automated adjustment of algorithm parameters, so they don't have to be tuned by hand.

In conclusion, IDA can look forward to success if it organizes itself around problems that both matter to society and afford opportunities for basic research. These problems are not proxies for important problems (as robot soccer is a proxy for more important things to do with teams of mobile robots) but are themselves important. Good problems can be defined with a few, nontechnical words, and they have clear criteria and metrics for success. The IDA community could adopt one or a few, and organize annual challenges around them, using them to steer research toward major, long-term goals. As a practical matter, these problems should generate research funding for some years to come. Finally, we will find that all good, important problems already have people working on them: biologists, sociologists, economists, ecologists and so on. We should not be scared away but should remember our heritage: The job of intelligent data analysis is not to create more data analysis algorithms, but to make sense of data, nearly all of which is generated by experts in fields other than our own. These people need intelligent data analysis, and we need colleagues and intellectual challenges, and something more substantial than a half-point improvement in classification accuracy to demonstrate what we're worth.

References

1. B. Chandrasekaran, Generic tasks in knowledge-based reasoning: High-level building blocks for expert systems design, IEEE Expert, 1, 3, 23-30, 1986.
2. Robert St. Amant and Paul R. Cohen. Interaction With a Mixed-Initiative System for Exploratory Data Analysis. Knowledge-Based Systems, 10, 5, 265-273, 1998.
3. Robert St. Amant and Paul R. Cohen. Intelligent Support for Exploratory Data Analysis. The Journal of Computational and Graphical Statistics, 1998
4. http://www.cs.arizona.edu/∼clayton/evap-web/
5. http://www.sigkdd.org/kddcup/index.php
6. http://people.brunel.ac.uk/∼csstxhl/IDA/IDA_1995.pdf
7. The R Development Core Team. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2009. http://www.R-project.org
8. http://www.robocup.org/
9. Ian H. Witten and Eibe Frank (2005) Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.