KDD 2010 Foreword to the Industrial/Government Track

The Industrial/Government track has undergone significant changes this year. These changes were reflected in a revised Call for Papers and a corresponding set of updated evaluation criteria, in the composition of the program committee, in our review procedures, and ultimately in our selection of which papers to accept. This foreword explains the nature and rationale for, and the impact of, these changes. We hope that this explanation will help authors in future years to submit higher quality papers of interest and value to both practitioners and researchers of KDD.

KDD is an applied field. It is about discovering knowledge that is valid, novel, interesting, understandable and useful. It is a field that draws strength from the interplay of research and applications. Interest in KDD from Industry and Government is due to the potential for discoveries of useful knowledge and the resulting value of such knowledge to enable better decisions, processes, or other outcomes.

The Research track of KDD focuses not on the specific discovery and use of knowledge, but rather on general new techniques, algorithms, and methods that enable such discoveries, usually for new classes of data or to improve performance on existing classes of data. This track is valuable not only to researchers as a venue for recognizing and documenting results, exchanging ideas and advancing the science, but also to practitioners in industry and government who can benefit from application of the new techniques, algorithms, and methods to their problems of interest and their available datasets.

Similarly, the Industrial/Government track can provide value to both industry and government practitioners of KDD and to KDD researchers. The value to industry and government practitioners consists of recognition and documentation of significant accomplishments in the application of KDD technology, sharing of information about what techniques work well for particular classes of problems and what techniques are not well suited when applied to the messy characteristics of the real world, sharing of lessons learned about how to apply KDD techniques successfully to discover useful knowledge, and presentations of useful knowledge that has been discovered as a result of the KDD process. The Industrial/Government track can be valuable to KDD researchers by exposing them to new problem classes and to new classes of data for which novel KDD techniques are needed and by illustrating the limitations of current techniques as applied to real world problems. The Industrial/Government track also provides researchers with an understanding of the context in which KDD techniques and results are used, what types of KDD techniques, methods, and results are valuable, and which are infeasible in real world applications. The importance of these value propositions to KDD researchers was discussed by Usama Fayyad in his invited talk at KDD-2007, in the three adjacent slides entitled "Researcher View", "Practitioner View", and "Business View"; we have focused the Industrial/Government track at these latter two, with an emphasis on the third. (available at: http://videolectures.net/kdd07_fayyad_dms/)

These value propositions motivated inclusion of papers in the following three areas:

(1) Deployed KDD systems that are providing real value to industry, Government, or other organizations or professions. These deployed systems could support

ongoing knowledge discovery or could be applications that employ discovered knowledge, or some combination of the two.

(2) Discoveries of knowledge with demonstrable value to Industry, Government, or other users (e.g., scientific or medical professions). This knowledge must be "externally validated" as interesting and useful; it can not simply be a model that has better performance on some traditional KDD metric such as accuracy or area under the curve.

(3) Emerging applications and technology that provide insight relevant to the above value propositions. These emerging applications must have clear user interest and support to distinguish them from KDD research papers, or they must provide insight into issues and factors that affect the successful use of KDD technology and methods. Papers that describe infrastructure that enables the large-scale deployment of KDD techniques also are in this area.

These three areas, which we refer to in shorthand as "deployed", "discovery", and "emerging" are the three areas of interest for papers in the Industry/Government track. The Call for Papers (available at http://www.kdd.org/kdd2010/cfp_industrial.shtml) explicitly listed the required elements for a paper in each of these areas.

KDD as a field has matured in the 15 years since the first conference. All KDD papers are now required to use real datasets, be repeatable, include comparisons to baseline methods, and report the statistical significance of experimental results. These requirements are stated explicitly in the Research Track Call for Papers (see http://www.kdd.org/kdd2010/cfp_research.shtml) and are not sufficient to distinguish Research track papers from papers appropriate for the Industrial/Government track. This maturation, while a positive development for the field, created a side effect of removing the utility of the heuristic "does the project analyze real data?" to distinguish between research papers and papers describing KDD applications and results. It is also insufficient just to assert the existence of a business problem, describe its characteristics, develop an algorithm optimized for that set of characteristics, and then demonstrate the outperformance of that algorithm on a real dataset. An additional factor was required: there had to be clear evidence that the asserted problem and its solution actually mattered to some external set of users. Without such evidence, a paper was not likely to be accepted. Short of describing a deployed application, having a co-author from a user organization who describes the user value of and involvement in an emerging application is one way to improve the likelihood of acceptance of a paper of this type.

Papers that were about the discovery of knowledge required more than just a set of experiments that showed how a new model provided improved performance over prior techniques; these papers also had to demonstrate some value of the knowledge that was discovered. This value could be demonstrated in a number of different ways: by a translation of the improved performance to the monetary value in the context of a real application problem, by the adoption by a business, Government agency, or other user community of processes or procedures that use the newly discovered knowledge, or by the publication of such newly discovered knowledge in a peer-reviewed journal in the domain to which it applies. Without such external validation, there is no way to know if the discovered knowledge is actually valid, interesting or useful from a user perspective. The discovered knowledge may be statistically significant, but it may not stand up to scrutiny from domain experts and it may not be useful given real-world constraints.

Asserting a breakthrough without this external validation risks an invalid result. The real-world context provides validation of the utility of our techniques and methods.

To refocus the I/G track on these objectives, several actions were required. We selected a program committee whose members had experience solving applied problems and deploying KDD applications. We reached out to the organizers of the KDD Practice Prize and invited them to join the program committee. We rewrote the Call for Papers to explain clearly what type of papers we were looking for. We instituted procedures to re-route submissions that clearly would be more appropriate for the Research track. (These procedures included author notification and consent to the switch.) All papers submitted to the Industrial/Government track received three independent reviews by Program Committee members. Most important, we reviewed all candidate papers to ensure that they fit the track's objectives in one of the three areas. The review form asked reviewers to check that each of the required elements was included in the paper, as appropriate for the applicable area. The Call for Papers also asked authors to identify in which of these areas their paper should be evaluated. Unfortunately, none did, so we asked the reviewers to make such a determination and evaluate the paper against the appropriate criteria. Many papers were difficult to categorize; either the reviewers didn't agree on the area, or they chose "none of the above." Reviewers were asked to discuss the appropriate area and reach a consensus for papers that were reviewed favorably but unclear as to area, and final decisions were made only after these discussions occurred. The Call for Papers included the opportunity for authors to submit a video demonstration of their system. No one took advantage of this opportunity.

Overall, we reviewed 101 papers and accepted 20, for an acceptance rate of 19.8%. Eleven papers were selected for long presentations; these comprise six deployed papers and five emerging papers deemed to be of general interest. One of the deployed papers and one of the emerging papers describe advances in enabling infrastructure. Of the nine papers selected for short presentations, 7 are emerging and 2 are discoveries. These papers are of more specialized interest than those selected for long presentations, but are of the same high quality. Papers are grouped by topic rather than by area for presentation purposes.   Topic groupings are advertising, transportation, business processes, disaster prediction and recovery, software vulnerability, systems and infrastructure, and medical.   This variety shows the vitality and applicability of KDD techniques across a wide range of useful application domains.

In general, we favored papers that were about real deployed applications of KDD and were extremely selective about papers in the emerging area, because deployment is clear evidence of utility and because the effort involved in the creation of a deployed application is typically much greater than the effort involved in a research project. We looked for papers of general interest that provided lessons that could be reused in other applications and domains. We also favored papers that were from Industrial and Government organizations over those that had authors solely from academic or research institutions, although we did accept some papers from the latter categories when they were of very high quality and met the criteria for one of the areas of interest.

Many papers that might have been accepted in previous years did not make the cut this year. This was due to an overt decision to reduce the acceptance rate for the Industrial/Government track to make it comparable to that of the Research Track as well as to the change in emphasis of the track. Papers that weren't clearly in one of the three areas were more likely to be rejected, as they did not fit with the goals of the

Industrial/Government track. Of the rejected 81 papers, consensus about the area was reached only on 13 of them, an indication that when the objective of the paper was unclear, the paper was unlikely to be reviewed favorably. Some authors of rejected papers disagreed with our decisions and contacted us to ask us to reconsider; all these authors were from academic institutions. We re-reviewed each of these papers in detail, but this additional scrutiny did not result in any changes.

We hope you find these papers interesting and look forward to continued improvements in the use of KDD techniques to provide real value to Industry and Government.

– Ted Senator and Ying Li, *Industrial/ Government Track Program Chairs*