Report of the:

ICAIL 2015 DESI VI Workshop on Using Machine Learning and Other Advanced Techniques to Address Legal Problems in E-Discovery and Information Governance

Organizers:
Jason R. Baron, Jack G. Conrad, Amanda Jones, David D. Lewis, and Douglas W. Oard.

The Discovery of Electronically Stored Information (DESI) workshop series began at the 2007 International Conference on Artificial Intelligence and Law (ICAIL) in Palo Alto with the aim of bringing together individuals from the academic and legal communities to discuss e-discovery, and especially the development of new technologies to support quality e-discovery processes. The six workshops to date have focused on the use of emerging techniques, including technology-assisted review, to make e-discovery more effective and efficient. Along the way, the DESI workshop series has contributed to scholarship in legal and scientific fields. Papers appearing in the workshops have been cited in case law, in law review articles, and in other academic literature.

A particular aim of the DESI VI workshop this year in San Diego was to broaden the legal issues discussed beyond e-discovery to others where advanced data analysis and classification technologies might contribute, most notably “information governance.” “Information governance” has been defined as those “activities and technologies that organizations employ to maximize the value of their information while minimizing associated risks and costs." (Information Governance Initiative 2014 Annual Report, available online.)

As described in the call for participation, DESI VI brought together researchers and practitioners to explore innovation and the development of best practices for the application of search, classification, language processing, data management, visualization, and related techniques to institutional and organizational records in e-discovery, information governance, public records access, and other legal settings. Participants were invited to address topics such as:

1. What combinations of machine learning and other techniques can best categorize information in accordance with existing records management and e-discovery policies?
2. Do effective methods exist for performing sentiment analysis and identification of personally identifiable information in a legally useful way in e-mail and other records of interpersonal communication?
3. Can proactive insider threat detection leverage information already being collected for records management purposes, and what would be the ethical and legal fallout of such approaches?
4. Where do recent legal cases point to the need for new research to better inform the decision of courts and the practices of parties?
5. What privacy implications do emerging technologies have for e-discovery, business intelligence, and records and information management?

The June 8, 2015 workshop attracted over 60 participants. Seven refereed and four unrefereed papers were published on the DESI VI website. The day's agenda included two keynote presentations, followed by two discussion panels and two sessions devoted to invited talks from members of the legal and academic communities.

After introductions, the workshop was kicked off by Bennett Borden of Drinker, Biddle & Reath in Washington DC, presenting on "Beyond eDiscovery: Applying Data Analytics to Build Early Warning Systems and Address Other Legal Challenges." Bennett's unique qualifications include being the only "AmLaw 100" partner who also holds the title of Chief Data Scientist at his firm, having recently been awarded a Masters in Business Analytics from New York University. Bennett's presentation discussed how predictive analytics and machine learning techniques can be used to help solve the challenge of finding relevant information within vast e-discovery corpora. Beyond e-discovery, however, data analytics can be applied to good effect in a wide variety of investigation, due diligence, and compliance areas. Bennett's presentation drew upon the results of his Masters thesis, in which he used predictive analytics not only to predict the classification of a document, but also to predict the probability of particular events occurring in the future. In its most ambitious form, the application could be viewed as an "early warning system" for predicting corporate misconduct – anticipating human behavior using technology-driven sentiment analysis.

A second keynote presentation was delivered by Jeremy Pickens, Data Scientist at Catalyst Secure in Denver. Jeremy earned his Ph.D. at the University of Massachusetts Amherst, with W. Bruce Croft as his dissertation advisor. In a lively address entitled "TAR, Information Governance, and the Relativity of Wrong," Jeremy invoked the musings of Isaac Asimov to persuade the audience that both the problems and the solutions of information governance are gradient, rather than absolute. He leveraged "SwarmSketches" to illustrate vividly how even imperfect technology working with imperfect humans can lead to meaningful results. His thought experiments also explored the human intellectual effort necessary to successfully tune machine learning algorithms.

Acting as a discussant, Karl Branting then began the day's discussions with his reactions to the two keynote talks. He observed that the talks addressed phenomena with opposite degrees of generality: data analysis of a single phenomenon, in the case of Bennett; and large-scale phenomena for which approximate solutions can be useful even if not truly correct, in the case of Pickens. In both cases, however, key issues included both "What can we predict?" and "What are the ethics of obtaining, analyzing, and acting upon the data revealed by the techniques?" Much of the discussion focused on the implications of developing increasingly accurate techniques for predicting misconduct. One set of issues concerned whether the predictive techniques, which were demonstrated in situations in which analytics have unrestricted access to all correspondence and documents, could be extended to organizations in which some documents are inaccessible or shielded by privacy rules. Accuracy under sparse data conditions would put significant constraints on the applicability of these techniques to other types of organizations and could create ethical dilemmas involving tradeoffs between privacy interests and ability to detect serious misbehavior. Another issue that surfaced was whether individuals who understood these techniques could strategically adapt their activities to evade detection. And a third issue concerned the ethics of taking action for "future crime," that is, for improper activities that haven't occurred but that are merely anticipated.

For the remainder of the morning, a panel of presenters discussed information governance topics from a variety of perspectives. Aimee Taal reported on the paper she authored in conjunction with James Sherer and Jenny Le entitled "Big Data Discovery, Privacy, and the Application of Differential Privacy Mechanisms." She discussed a new breed of algorithmic techniques that may be helpful in mitigating the danger of disclosure of personally identifiable information in large datasets. Other panelists, including David Marcos, Sandy Serkes, and Tom Barnett considered applications of machine learning to categorize information in ways that could modernize the field of records management.

The first afternoon panel session, chaired by David Lewis, consisted of presentations of refereed papers. Amanda Jones presented a paper on "The Role of Metadata in Machine Learning for Technology Assisted Review." William Dimm, Richard Oehrle, Hans Henseler, and Chris Paskach presented their paper submissions, which can be found on the DESI VI website, with Douglas Oard reacting to these presentations as a discussant. The diverse set of topics addressed in this session included assessment of classifier performance using 'extrapolated precision' (Dimm) and semi-automated methods for enhancing the precision of a term set with minimal recall loss based on context analysis (Oehrle). Paskach articulated a compelling case for utilizing statistical sampling to maximize the benefits of TAR in e-discovery. Finally, Henseler provided an overview of a range of e-discovery and semantic search-related activities currently taking place at two universities in Amsterdam.

Rounding out the day's discussion, Amanda Jones moderated a panel consisting of Bennett Borden, Jenny Le, Irina Matveeva, Jeremy Pickens, and Vern Walker. This group talked about the future of machine learning and analytics as applied in the legal sphere, for e-discovery and beyond. The conversation was wide-ranging, but a dominant theme emerged regarding the desirability of employing technology and analysis creatively to shape and substantiate case strategy. When that is the goal, simply classifying documents as responsive or non-responsive is no longer enough. At this point, members of the e-discovery community and academia are thus looking to data analysis technologies to facilitate construction of legal narratives by readily offering more sophisticated insights. Perhaps the next generation of technology for e-discovery practitioners should behave more as an intelligent partner than a digital servant.

One goal articulated by the organizers of this workshop was to explore ways of integrating ideas like those which the DESI workshops have been grappling more broadly into the ICAIL conference. To this end, DESI VI workshop co-organizer and current IAAIL President Jack Conrad, helped structure this year's conference to focus a part of the first day of the ICAIL conference on some of the themes that this and past DESI workshop have explored, with paper sessions organized around the themes of Big Data, Machine Learning, and E-Discovery. By the end of the workshop, however, it had already become clear that this symbiotic sharing of ideas had already become a two-way street, with participants in the DESI workshop benefiting fully as much from the themes and expertise brought together for ICAIL as that conference benefited from the workshop. We look forward to continuing the process of more closely coupling the themes and ideas of future DESI workshops at future ICAIL conferences.