Since the dawn of civilization, the human race has been fascinated by and proactively pursued the ability to predict. Devices, methods, and procedures—grounded in intuition, experience, conviction, paranormal beliefs, or (increasingly) science and mathematics—have been experimented with, debated about, and applied to predict all kinds of physical, biological, behavioral, and societal phenomena. For thousands of years in Europe, crystal balls were believed by some to give a glimpse into the future to gain clairvoyance. In China, two astrologists authored a hugely influential prophetic book 1,300 years ago—Tui Bei Tu—that consisted of 60 annotated pictures. Allegedly, these pictures foretold major political upheavals, with many of these predictions depicting actual historical events.

Conditioned in modern-day scientific and technological backdrops, the grand visionary concept of psychohistory perhaps best captures the human fascination with prediction. One of the giants in science fiction, Isaac Asimov, coined this term, envisioning that psychohistory, a yet-to-be-developed scientific field, would integrate knowledge from history, psychology, and sociology, in a super advanced mathematical framework that in turn could predict the behavior of very large groups of people as well as political establishments and entire societies at a galactic scale. With some kind of feedback mechanism, psychohistory, in Asimov’s fictional work, generated actionable intelligence and became the backbone of a grand social engineering design for galactic empire building for the benefits of the entire human race.

Of course, psychohistory still remains science fiction, despite the fact that more than 70 years have passed since its inception. However, today’s AI researchers and practitioners will agree with me that significant advances have been made in these 70 years as to the science and art of prediction, and that the vision of psychohistory is no longer purely in the realm of science fiction. In effect, it’s hard to ignore the uncanny parallelism between psychohistory and the emerging field of predictive analytics.

Crystal Balls, Statistics, Big Data, and Psychohistory: Predictive Analytics and Beyond

Daniel Zeng, University of Arizona and Chinese Academy of Sciences
A Promising New Field

As a working definition, predictive analytics is concerned with integrating and applying computational approaches from statistics, machine learning, data mining, text mining, and network science, among others, to predict future events in a wide range of application contexts, as well as individual, group, and societal behaviors and actions. Painting the landscape of predictive methods in very broad strokes, we can roughly position predictive analytics as the latest and the third installment of a three-phase evolution of prediction methodology. Prediction methods developed during the first phase are largely unscientific. They’re typically based on human intuition, unformalized experience, or direct reasoning grounded in unproven, argumentsative principles or rules of hypothetical nature—crystal balls and Tui Bei Tiu are such examples. The second phase of evolution of prediction methods can be characterized as statistical reframing of prediction problems. Some core prediction problems have been elegantly formalized by statisticians; the task of predicting becomes well-structured mathematical manipulations. Equipped with these statistical methods, prediction and associated extensive use of data have become routine in many domains and have met with meaningful successes.

However, along with explosive generation and abundant availability of data in recent years, the second-phase methodology faces many challenges. The set of application settings that can benefit from data-driven prediction is quickly expanding. The end users of prediction methods have become more aggressive and demanding as to the scope and performance of predictions. Furthering complicating the situation is that new types of data of potential usefulness to prediction and the underlying nature of many application settings defy the statistical assumptions underpinning the methods developed in this second phase.

Against this backdrop comes predictive analytics, which doesn’t necessarily need to deal with big data but has a lot to do with big data-driven thinking. Under predictive analytics, many phenomena previously thought as unpredictable are being attempted as the target of the new emerging class of predictive methods. The datasets used as input are much richer than before. Some assumptions needed by the older statistical methods are being circumvented. A growing number of frameworks and tools are also treating predictions and decisions in a more integrative manner. In effect, prediction and big data are so intertwined that predictive power has been touted as a defining and differentiating characteristic of big data.

It’s obviously impossible to survey the whole body of the growing predictive analytics work in this short letter. But even just anecdotally, a lot of interesting things can be said about predictive analytics. From the point of view of data input, predictive analytics has investigated the use of social media data, including but not limited to user-generated content and user behavior on social media sites. Recent predictive analytics attempts have also explored the predictive power of geospatial data streams, some of which can be acquired from social media sites, and various types of multimedia data (such as facial expressions extracted from photos and crowd behavior from videos). The spread of domains in which predictive analytics is being applied is wide. Stock price prediction has been reexamined under...
this new light, and public health applications using predictive analytics tools such as Google’s Flu Trends, have redefined the game of health surveillance.

Perhaps one of the most successful applications of predictive analytics is in law enforcement. Predictive policing is concerned with applying predictive analytics to analyze broadly construed crime data, with the aim to help predict future crime and enable better allocation of law enforcement resources. The Times magazine listed predictive policing as one of the top 50 best inventions of 2011. The Economist published an article on predictive policing in July 2013, claiming that it’s getting easier to foresee wrongdoing and spot likely wrongdoers. In an August 2011 article, the New York Times discussed the predictive policing practice as “sending the police before there’s a crime.”

Technology providers have been a major player in the rise of predictive policing. For instance, IBM offers a social media analytics toolkit that relies on social media to facilitate crime situational awareness. Some software vendors make dedicated analytics tools to go through historical crime databases to predict the location, time, and type of future crime. In practice, a growing body of strong evidence attests to the efficacy of predictive policing. Many case studies of successful adoption of predictive policing in US and European law enforcement agencies can be found in the literature.

**Overcoming Obstacles**

Similar success can be found in application of predictive analytics in other domains. Yet, to further develop predictive analytics and widen its application success and practical impact, several challenges must be seriously tackled. Fortunately, AI techniques have a lot to offer in addressing these challenges.

The current generation of predictive analytics is largely data driven and relies on standard machine learning methods. How can such methods be extended to easily incorporate domain knowledge and make the best use of data-knowledge hybrids? In many domains, it’s critical to offer explanation for the predictions generated. Can predictive analytics provide such explanatory capabilities to interface with users? Often, representing data input under one unified representation just isn’t possible, necessitating customized subrepresentations. How can effective learning taking place in such a context? In the era of big data, it’s more important than ever to tease out salient predictors from myriad candidate datasets and features. How can this be accomplished efficiently and effectively? I’m convinced that in addition to machine learning and text mining, which have already been extensively applied in predictive analytics, many other subareas of AI, including but not limited to knowledge representation, automated reasoning, and cognitive computing, will play a more prominent role in predictive analytics in the years to come.

Asimov’s most influential works include the “Robotics” series and the “Foundation” series. The “Robotics” series has inspired generations of AI researchers, with the famed “Three Laws of Robotics” contributing substantially to the notion of ethics of AI. Asimov’s “Foundation” series, although previously linked to AI in a much weaker sense than the “Robotics” series, is likely to have growing relevance, as predictive analytics gains popularity. Just like the laws of robotics, Asimov established two axioms for psychohistory: the population whose behavior was modeled should be sufficiently large, and the population should remain in ignorance of the results of the application of psychohistorical analyses. The first axiom is becoming largely irrelevant and obsolete in the big data era. The second axiom, however, could have profound practical relevance, raising interesting research and policy questions.

As AI is deployed as a vehicle to accomplish psychohistory (or an even more ambitious version of psychohistory approaching omniscience), or more practically, as psychohistory provides inspiration for AI researchers in the area of predictive analytics, many interesting ideas, theoretical work, and impactful applications will emerge. These developments may well be contributing significantly to computational social sciences and computational policy making, and posing interesting technical challenges to drive the state of the art of AI. Applying predictive analytics to AI research itself, one might wonder, where is AI going? Can the future of AI be predicted?