New Start-Up Accidentally Reveals How Google Really Works: The Ultimate Spy Tool!

The failed promise of “Big Data” has missed every single terrorist event and cost western spy agencies their credibilty. Now a start-up wants to wring the last bit of blood out of “The Big Data” pitch.

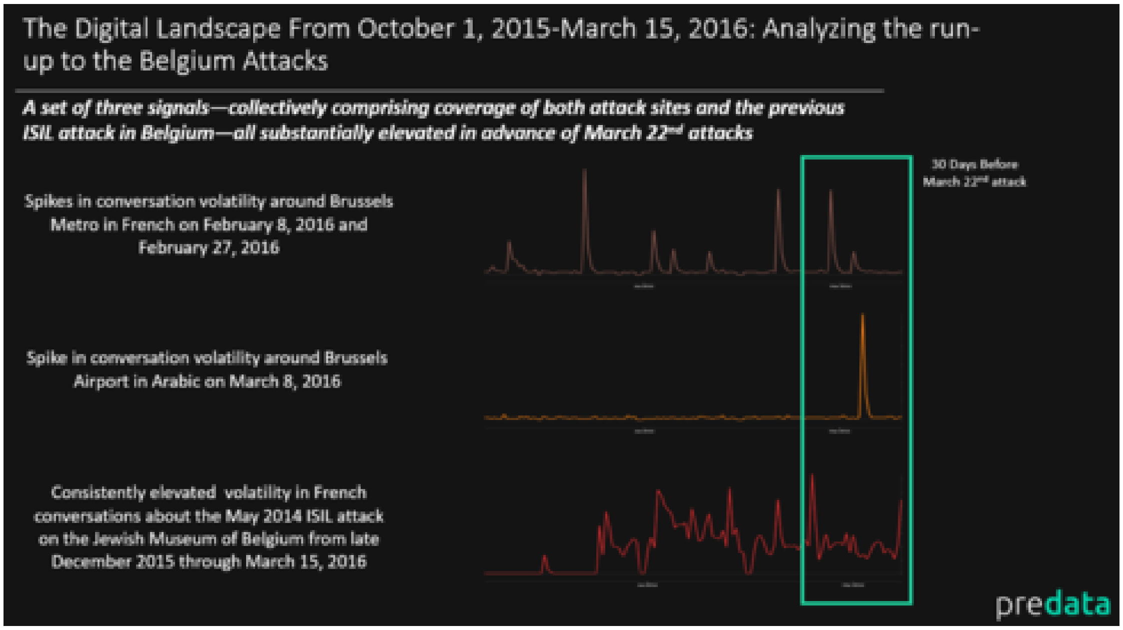
When James Shinn was working for the CIA as a senior East Asia expert more than a decade ago, he longed for the tools of a weatherman. He wanted to be able to predict that the chance of North Korea test-firing a missile within a month was, say, 60 percent. It remained a fantasy, he says, until now.

Shinn and his 14-person team at Predata have developed software that numerically describes political volatility and risk. It vacuums up vast quantities of data from online conversations and comments, compares them with past patterns, and spits out a probability. (A version of Predata’s service is accessible on the Bloomberg Professional service.) Shinn likens his product to sabermetrics, the statistics-driven baseball strategy popularized in Michael Lewis’s *Moneyball*. “By carefully gathering lots and lots of statistics on their past performance from all corners of the Internet, we are predicting how a large number of players on a team will bat or pitch in the future,” Shinn says, by way of analogy.

James Shinn

Predata doesn’t replace human analysts so much as offer them a new tool. Without people choosing what to follow, metadata scraping has limited use. Moreover, Shinn argues, while risk-analysis companies are increasingly offering clients numerical percentages, the data are often pulled from the air. “This is a machine-driven, carefully calculated risk index,” says Shinn, the company’s founder and chief executive officer. “There is no arbitrary scoring by a human analyst.”

Each day, Predata monitors about 1,000 Twitter feeds, 10,000 Wikipedia pages, 50,000 YouTube videos, and several dozen newspapers and magazines in some 200 countries. It covers 300 topics, including news about individual companies, the debate over the U.K. leaving the European Union, and interest rate decisions by central banks.



The Digital Landscape From October 1, 2015-March 15, 2016: Analyzing the run-up to the Belgium Attacks

A set of three signals—collectively comprising coverage of both attack sites and the previous ISIL attack in Belgium—all substantially elevated in advance of March 22nd attacks

Spikes in conversation volatility around Brussels Metro in French on February 8, 2016 and February 27, 2016

Spike in conversation volatility around Brussels Airport in Arabic on March 8, 2016

Consistently elevated volatility in French conversations about the May 2014 ISIL attack on the Jewish Museum of Belgium from late December 2015 through March 15, 2016

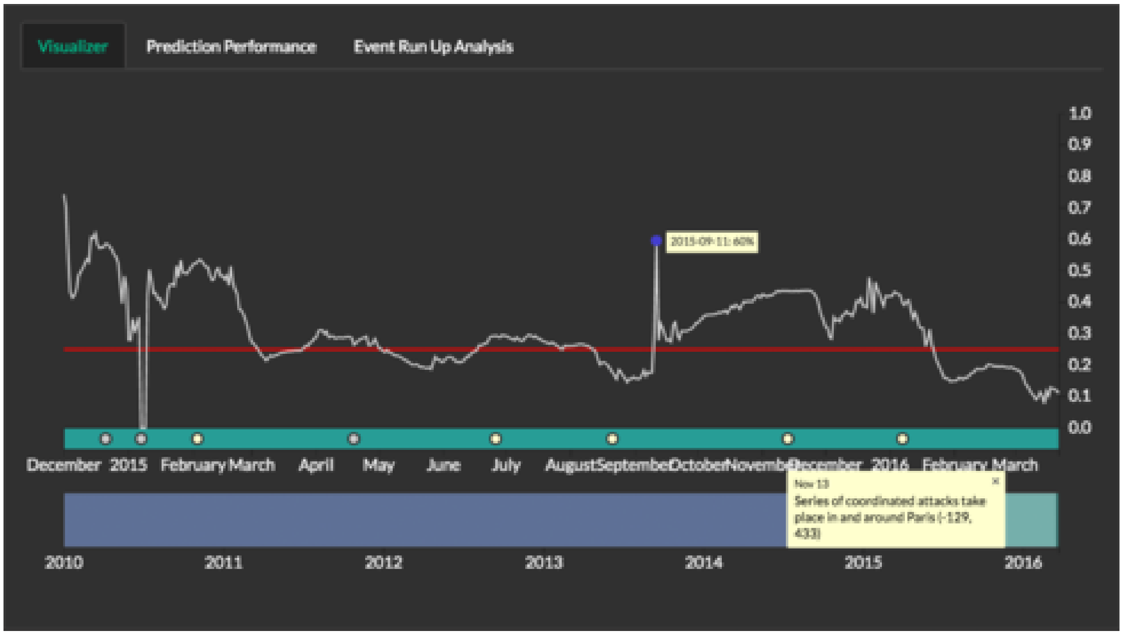
Historical data is paramount. For instance, Predata didn’t make a statistically useful prediction for the March 22 attacks in Brussels, in part because Belgium had experienced few such incidents. The software needs at least five previous events to find a correlation between digital conversations and an act of terrorism, according to Shinn. France, on the other hand, had witnessed 13 incidents prior to the Paris attacks on Nov. 13; the company says that its model indicated the likelihood of an event being at least 61 percent a month in advance. Similarly, on Dec. 27, Predata says it calculated a 68 percent chance that North Korea would engage in some activity regarding weapons of mass destruction within 45 days. Almost two weeks later, on Jan. 6, the Kim Jong Un regime conducted the nation’s fourth nuclear test.

Shinn, who served as an assistant secretary for East Asia at the U.S. Department of Defense after his CIA stint, began developing the technology in 2014 while teaching at his alma mater, Princeton, and serving on the advisory board of Kensho Technologies, an analytics software developer for investment management. Kensho’s CEO, Daniel Nadler, and Shinn experimented in their free time with a crude prototype that monitored online conversations among labor unions in South Africa, thinking the data offered a handle on the country’s volatility. They found that back-and-forth argumentation in English and Afrikaans on sites as public as the Wikipedia pages of the unions spiked before mining strikes, after which gold and platinum prices surged.

Shinn recruited one of his students, Andrew Choi, to build a more sophisticated algorithm, a part of which used a specification devised by the Intelligence Advanced Research Projects Activity, an organization that leads research into innovative technology under the Office of the Director of National Intelligence. “Predata is now sampling a larger, more complex pattern before and after events,” Choi says. “We are characterizing larger patterns of behavior by many people on the Internet and then looking for the recurrence of these patterns as ‘early warning’ that an event of that type is about to recur.”

Mourners outside Le Carillon restaurant, after the attacks in Paris, on Nov. 16, 2015.
Photographer: Simon Dawson/Bloomberg

Choi, now Predata’s chief technology officer, says the company can draw direct links from digital conversations to geopolitical volatility, and from there to collective actions such as strikes, protests, and troop movements. There was a surge of comments posted to news articles about Russia and Ukraine before Russia’s invasion of Crimea, for example, and a sharp rise in participants conversing on the French-language Wikipedia page for Islamic State before the Paris attacks. “There is a characteristic way that the narrative of a certain event is constructed before and after that event has taken place,” he says.



Nov 13
Series of coordinated attacks take place in and around Paris (-129, 433)

The digital landscape leading up the the attacks in and around Paris on Nov. 13, 2015.
Source: Predata

As the activity intensifies, Predata gathers the metadata—such as how many people edit a Wikipedia page about a terrorist attack and how quickly those edits are disputed—and retroactively matches the statistical signals against sets of historical events to predict the likelihood of similar events, Shinn says. “The Russians spend a lot of time and money molding the narrative about Russian intentions and actions, and the guys in Beijing do the same about the South China Sea.” The result, he says, is that actors leave footprints; their propaganda interests trump their desire for operational security.

Governments started using the first computational models in World War II to decode German messages and automate the targeting of anti-aircraft weapons. In the 1960s, corporations and research institutions began commercializing models for a variety of uses, from deciding credit risk to predicting the weather. The business has evolved into hundreds of companies pursuing three general approaches to managing risk and volatility: prioritizing new and better sources of data, focusing on better ways to condense data, and combining the two to forecast probability. Microsoft co-founder Paul Allen invested in Seattle-based BlackSky Global, which plans to launch a fleet of 60 satellites next year to scan most of the planet as many as 70 times a day. San Francisco-based Spaceknow combines the imagery of more than 6,000 industrial facilities with algorithms to create an index of China’s factory production. Companies such as Banjo, Cytora, and Dataminr monitor social media and the Internet to track events as close to real time as possible. They specialize in detecting riots or protests as soon as they’re reported on the Web, alerting clients to developments and making the information easily digestible.

Predata is alone so far in producing an algorithm-generated forecasting metric. The company is already broadening its findings, applying them to equity indexes, foreign exchange rates, commodities, and credit-default-swap spreads, and it says it’s found a correlation between the signals it measures and prices of various assets. Predata also tracks chatter about YouTube videos of central bank governors’ press conferences to predict rate decisions.

Kalev Leetaru, a senior fellow at George Washington University, started the GDELT Project in 2013, a free, publicly available database of all the world's media in print, broadcast and web formats in more than 100 languages, stretching back from Jan. 1, 1979 to present day. GDELT uses more than 40 algorithms to translate into English and process metadata from the media into tables, allowing individuals and organizations to run customized risk analytics. It also provides a real-time timeline measuring instability in countries, Leetaru said, adding that GDELT plans to roll out a series of prototype templates later this year, forecasting risk by location and type of volatile events.

One risk to Predata's approach is the potential of being gamed by a savvy hostile party. Shinn and Choi argue, however, that the Internet's "democratic checks and balances" and abundance of data points will keep Predata's technology relatively safe from attempts to fool it. Says Choi, "It's not even that you are trying to talk about a topic in a favorable light. It's that you are talking about that topic at all—that's a signal."

A version of this story appears in the forthcoming issue of Bloomberg Markets *magazine.*