



RICHARD NEWSTEAD/GETTY

The rise of technologies such as satellite tracking could have been foretold by analysing past literature.

## TECHNOLOGY

# Text-mining offers clues to success

*US intelligence programme analyses language in patents and papers to identify next big technologies.*

BY SARA REARDON

A project backed by a US intelligence agency might soon make it much easier to predict which technologies will one day become game-changers. Results revealed this week by the Intelligence Advanced Research Projects Activity (IARPA) suggest that clues in the wordings of, and relationships between, scientific papers and patents could foretell research successes.

The project, called Foresight and Understanding from Scientific Exposition (FUSE), could enable funders to pick winners, and help governments to keep an eye on 'disruptive technologies' — those that they feel might threaten national security or outpace regulations, for instance. Past examples include nanotechnologies and information technologies, such as the use of the Global Positioning System in mobile phones to allow tracking of individuals' movements. Last week, FUSE, a four-year project that started in 2011, entered its last phase: to predict the successes three to five years from now.

Although abstracts have been text-mined in the past for keywords and other clues in language, FUSE is one of the first projects to mine the whole bodies of papers and patents. So far, it has performed more than 2 million analyses of past data to pick out key advances, says Dewey Murdick, manager of the FUSE project. From these, it has identified several

hundred indicators, such as new collaborations or expressions of excitement in text, that highlight emerging areas.

"What we'd like to glean is understanding of the right combination of things that leads to success," says John Byrnes, a computer scientist at SRI International, an innovation centre in Menlo Park, California, whose team is one of three developing software for FUSE. To make predictions, his program mines text for keywords, citations and phrases that indicate authors' outlooks in scholarly papers.

One example he cites is the resolution of a technical problem that, once solved, led to what is now a mainstay of solar-panel technology. In the mid-1990s, millions of dollars were invested in research into solar panels that used aqueous solutions to convert photons into energy. Although promising at first, by 2008, the technology had been overtaken by the much more stable and effective solid-state solar panels. FUSE might have predicted the aqueous panels' demise, according to results presented by Byrnes and his team this week at the US Department of Energy's SunShot Grand Challenge Summit in Anaheim, California.

Scientific literature analysis, or 'scientometrics', is decades old. Organizations such as Thomson Reuters, an information firm

headquartered in New York, have long used these analyses to identify the most influential papers or researchers in a field. FUSE takes this further by mining millions of papers and patents in both English and Chinese, two of the most commonly used languages in scientific literature, says Murdick.

The analysis and indicators can predict whether a nascent field will become prominent or whether it is simply a source of excitement that will soon die out, says Olga Babko-Malaya, a research engineer at BAE Systems in Winchester, Massachusetts, who heads another FUSE team.

Her team uses software algorithms to analyse 'sentiment' in the natural language of papers. For instance, authors might say that their work builds on or contradicts a cited paper, or use descriptive language that expresses excitement.

The researchers also found that promising topics invent their own jargon and start using more acronyms. "Abbreviations imply acceptance by the community and are indicators of more mature technologies," says Babko-Malaya.

The changes in group collaborations in a field over time can also be predictive. FUSE researcher Lance Ramshaw at Raytheon BBN Technologies in Cambridge, Massachusetts, and his team are analysing networks between different topics, keywords and authors. He says that a new topic may be emerging when prominent authors start contributing to a group of papers that share common traits, or when alliances between collaborations shift.

Alan Porter, who specializes in technology forecasting at the Georgia Institute of Technology in Atlanta, agrees that retrospective predictions, such as the solar-panel example, are useful for modelling what companies have been doing and for tracking the history of a product. The more difficult task, he says, would be to use such a network to identify the "white spaces": areas between technology clusters that are ripe for new research.

Ideally, analysis will show patterns or tipping points that are common to success stories. Such patterns might eventually allow the project to forecast when a product might launch, or whether a drug will be approved by regulators, explains Babko-Malaya.

Although software is catching up, human analysts remain the best forecasters, says Murdick. "You can ask experts anything you want," he says.

In another project, Forecasting Science & Technology, IARPA is funding an online crowdsourcing project called SciCast, run by George Mason University in Fairfax, Virginia, and the American Association for the Advancement of Science. It aims to consult 10,000 scientists to help develop methods for generating accurate forecasts. "My personal bias is that it's a combination of people and machines that will ultimately provide the most useful value," says Murdick. ■