

Vibing Feeling the Vibes

Cambridge Science Park-based Linguamatics has quietly developed the premier text mining software in the world. **Mike Scialom** found a firm at the top of its game.

In mid-January Science Park-based Linguamatics released I2E OnDemand, the cloud version of its impressively agile – and globally dominant – text mining software.

I2E is used by nine of the world's top ten pharma companies and the hosted version will ensure it becomes accessible to the top 100 – but this isn't an application that appeals only to pharma firms, as I found out at a demonstration by senior product manager Guy Singh and senior application specialist Sarah McQuay at the firm's St John's Innovation Centre base.

During the course of our chat three of the firm's four founders came in (the fourth moved on to other things last year) and what transpired put the whole text mining sector into a very intriguing perspective.

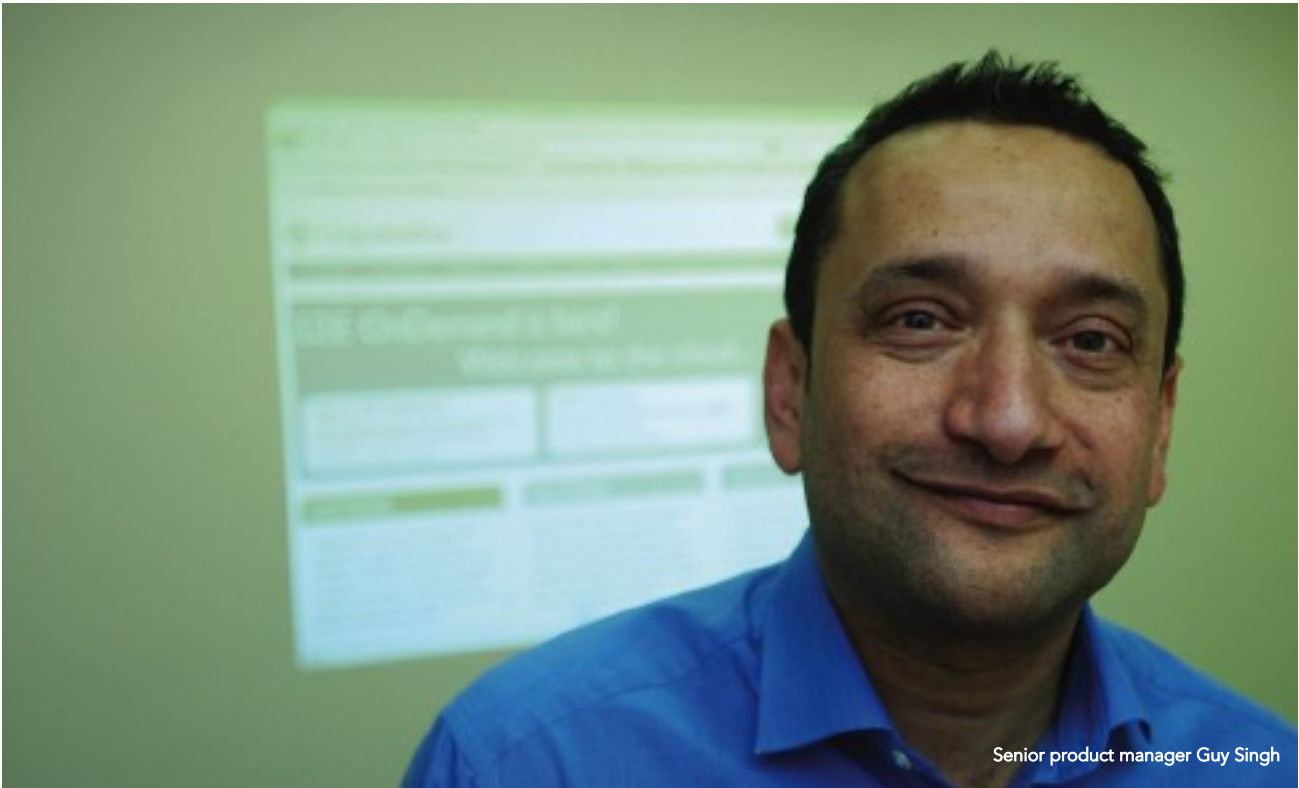
Firstly, it turns out that I2E wasn't designed specifically for the pharma sector – it just happens that pharma folk have made use of it faster and more frequently than anyone else.

"It was written with anyone in mind and has been customised to different areas," said Sarah McQuay.

'There' are a lot of text mining companies – and we're the strongest' - cto David Milward

From left: James Thomas, David Milward, Roger Hale.





Senior product manager Guy Singh

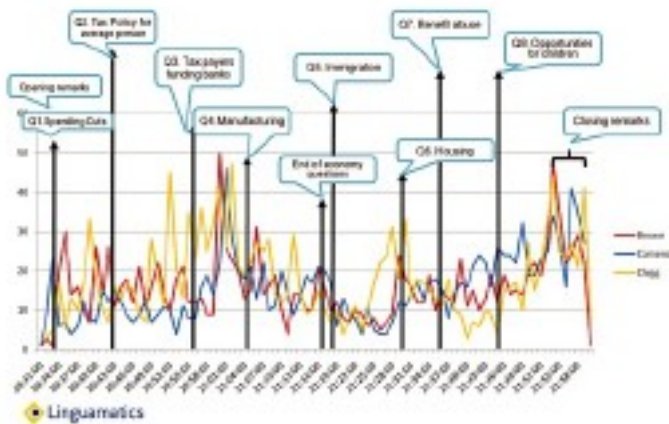
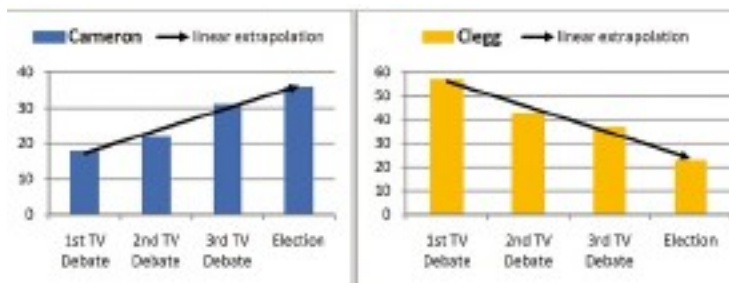
“All industries are becoming very knowledge-based and pharma is especially knowledge-based,” added cto David Milward.

“And scientists are by nature early adopters,” continued Sarah.

“There’s lots of text mining companies working in that area,” David said, “and we’re the strongest.”

What Linguamatics does is extract meaningful data from large amounts of text – and faster than anyone else. The firm already has global reach but the potential of text mining goes way beyond sifting through data acquired by scientific research.

To illustrate the wider picture, Guy and Sarah talked



Linguamatics mines Twitter during live election debates last year.

me through a genre-busting project which involved monitoring the Twitter-sphere last year when, for the first time in the UK, televised debates were held which featured the leaders of the three main parties prior to a general election.

Linguamatics had some familiarity with monitoring Twitter thanks to Project Twitch, when the Government asked firms to deliver proposals on “noisy feeds”. Twitch was funded by the Technology Strategy Board and sponsored by the Department for Business, Innovation and Skills with the aim of showing that external chatter can provide early warning in near real time of economic or security problems. In other words the Government wanted to get 9/11 or 7/7-type information early enough to be able to do something about it.

Shifting its focus from science research to blogs and

Twitter, Linguamatics used the Natural Language Processing (NLP) technology built into I2E to test the hypothesis that weak signals from large numbers of

users can show up problems which single-point analysis fails to spot.

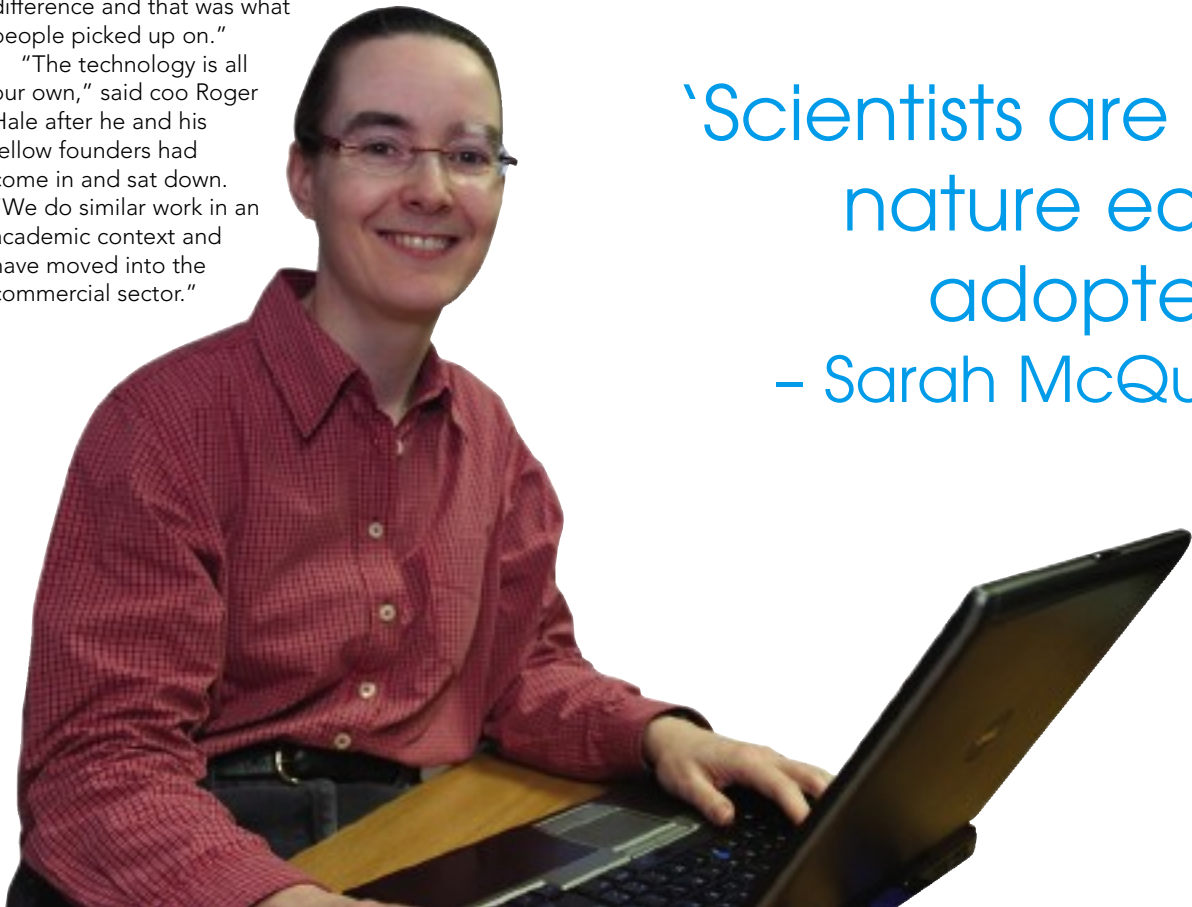
Linguamatics' Twitter study involved establishing a direct feed into Twitter's servers during the election debates last April and establishing from sentiments expressed in the Twitter-sphere what people felt towards the leaders in real time – and from that their chances in the election itself.

Linguamatics – it has a staff of two dozen in Cambridge and half a dozen in Massachusetts – took only positive Twitter statements about the three leaders and used I2E to automatically research the text and break it into chunks which it then categorised into topics such as "Trident", "economy", "defence", "housing", "education" and so on. Positive tweets were then collated within each of these topics.

"Decoding scientific language is very subtle and that's what we try to capture," said Sarah. In this case, of course, the language was more generalised, including statements such as "I like the way Clegg handled the banking issue" or "I agree with Cameron on Afghanistan". The results, based on 567,000 tweets from 130,000 tweeters, showed quite clearly (see graph) that David Cameron started poorly and got stronger and Nick Clegg started well and then tailed away. The results Linguamatics produced correlated closely with the outcome and the story was picked up by Rory Cellan-Jones at the BBC and by the New York Times.

"We were very very close to the actual outcome," remarked Sarah. "Just a 1 to 2% difference and that was what people picked up on."

"The technology is all our own," said coo Roger Hale after he and his fellow founders had come in and sat down. "We do similar work in an academic context and have moved into the commercial sector."



What's happened so far is that I2E has required a lot of back-up (basically IT staff) to operate, and the result has been that the big pharmas have been the biggest users because they have had the resources to make full use of it. But, with the cloud version now available – Linguamatics declined to talk about specific amounts of money but you can be sure that I2E OnDemand's price tag is significantly less than the original – there are going to be a whole lot more companies who will discover the benefits of I2E.

"You can adapt it, yes," said Guy of the software's applications. "Lots of customers will be using it as part of the drug discovery process but also at launch to find out if there's interest and what sort of responses turn up."

"And people have adapted it," interjected David. "The agro-chemical industry, for example."

Linguamatics has, as Roger Hale says, been "outward-focused" since its inception ten years ago, and has now reached the point where it can rightly expect interested parties to come knocking on its door with ideas to develop further applications even as it fine-tunes its ability to deliver world-class software to the pharma sector.

Later, as I stood by the water cooler in the foyer checking my messages James came up and said what a great firm it is to be part of.

"We have no external investors," he added, "although if we wanted we would get funding easily given our profit record."

Indeed. There'd probably be a queue.

'Scientists are by nature early adopters'
– Sarah McQuay