THE BELIEVER

SEPTEMBER 2003

USAMA FAYYAD

[DATA MINING EXPERT]

"NO ALGORITHM CAN DIVINE THE INTENTIONS OF TWENTY GUYS BY THEIR CAR RENTAL AND HARDWARE STORE PURCHASES."

Things data mining might reveal in space:
Galaxies
Small Venutian volcanoes
High-energy quasars

On Earth:
Terrorists
Fraudulent credit card transactions
Crappy cars

On February 13, 2002, the Pentagon created a new research department called the Information Awareness Office (IAO). Part of the Defense Advance Research Projects Agency, the IAO is intended to help fight terrorism with information technology. Under the rubric Total Information Awareness, the IAO initiated several research programs to develop ways to collect and analyze data from the billions of transactions with digital imprints that we all make every day, and to look for patterns that might reveal terrorist activity. Genisys, for example, hopes to create a new type of database language that could integrate data from everywhere. The Evidence Extraction and Link Discovery group would develop ways to extract information from various data sources and trace links

between people and their activities to ferret out potential networks of terrorists. FutureMAP will use futures market tools to avoid surprise and predict future events. Quickly, data mining became the new watchword in intelligence. Data mining, they're hoping, could help us know those infamous unknowns. The logo they came up with for the IAO and Total Information Awareness was an eyeball shining a light ray out of a pyramid, with the inscription Scientia Est Potentia.

What's more, it also turned out that the IAO would be headed by John Poindexter, the national security adviser under Reagan who was convicted of conspiracy, lying to Congress, and destroying evidence during the Iran-Contra Scandal. Thus: A hew and cry ensued. Nobody likes the idea of that eyeball shining its evidence extraction beam into their data. Most people don't want their scientia to be the government's potentia. And let's face it: Governmental data mining, especially under the title Total Information Awarenesss, just plain sounds scary. But what is data mining exactly? Is it true that computational tools from science and marketing could help discover terrorist networks? If so, how? Right now, Total Information Awareness is just a bunch of rather senseless PowerPoint slides and a small group of specialized technical people doing a little basic research. The electronic Panopticon is not yet at hand. Can this idea even work? If so, can it function properly while respecting privacy?

Usama Fayyad is one of the world's foremost experts on data mining, and The Believer *turned to him for some answers. After Fayyad completed his Ph.D. (one of the first on the topic of data mining), he went to the Jet Propulsion Laboratory in California, where he won many research awards for leading the way in developing methods for analyzing large scientific*

databases. Later, at Microsoft, and then with his own company, DigiMine, Fayyad took advanced data mining to the commercial world, helping to create what has become a huge business. Fayyad is still editor in chief of Data Mining and Knowledge Discovery*, the primary technical journal on data mining technology. Fayyad is not working on Total Information Awareness, but he's what this interviewer's grandpa would call a real* macher *in the field. The technical objectives of Total Information Awareness are built on ideas and technologies that Fayyad helped pioneer.* The Believer *caught up with Fayyad over the phone while he was seventeen hours ahead, in Australia, preparing for a conference.*

—Joshuah Bearman

THE BELIEVER: It seems that the term data mining has become a catch-all phrase. I've seen it used to refer to all kinds of different things. Is there a technical definition?

USAMA FAYYAD: The way the Total Information Awareness folks use the phrase "data mining," or at least, the way the Washington, D.C., community uses it, is too broad. They're referring to any time you collect and look at data. Whereas, we in the field mean the practice of using algorithms to look for patterns in data and using predictive modeling.

BLVR: Data mining of the kind envisioned by Total Information Awareness is not traditional statistics, right? Normally, a statistician investigates a specific hypothesis. But, in the case of Total Information Awareness, the point is that you don't know what terrorist activity looks like. You don't have a hypothesis. In theory, if they're able to achieve what they hope to, they will be scanning heaps of data, and divine from that

a set of potential hypotheses, plausible futures. And then analysts will look at those to determine what data patterns—that is to say, people doing things out there in the world—represent security threats.

UF: Yes. In fact, the term "data mining" used to be a term of derision among statisticians, referring to when you're fishing in the data without having an a priori hypothesis. If you do statistics properly, you're supposed to come up with a hypothesis that you can either confirm or reject. In today's world, that's not a good way to go about it. When you have lots of data, and few people, you want algorithms that will churn through the data, evaluate lots of models, and look for things that might become interesting patterns or hypotheses.

BLVR: Can you define *algorithm*? That's one of those technical terms that sort of snuck into the vernacular without people, including me, really knowing what it means.

UF: It's fairly simple. An algorithm just means a set of instructions for conducting a task. So, if you wanted to compute the average for a set of numbers, the algorithm would be: 1) Step through all the numbers 2) Sum up their values, and 3) Divide by the total number. It's a series of instructions for implementing a mathematical notion.

BLVR: Is it expressed as a series of instructions, or as a mathematical formula?

UF: People usually write them at a higher level, meaning abstractly. You have loops and conditions, and then within that you write—well, you wouldn't write instructions to the computer, but rather mathematical formulas.

BLVR: I like this idea that you've got these little teams, like computational commandos, out there in the data, working for you. And they report back with their recon

efforts. How helpful are these algorithms?

UF: In a world where you have a few variables—just a couple things to evaluate—it makes more sense to put that information in front of a person and have them come up with hypotheses and test them. But let's say you're talking about ten variables—already the human is lost. If it's a hundred, or a thousand, or ten thousand, or a million variables, and billions of data points, which, you know, is the kind of thing we deal with on a daily basis, in the scientific and commercial world, then you have no hope whatsoever of understanding it at all. Human beings, from a mathematical perspective, are fairly limited. Two and three dimensions, maybe five, and we're OK. But that's about it. And this is where these algorithms can help a lot, because they can comprehend thousands of dimensions, and focus their attention on things that might be interesting.

BLVR: How did you use data mining at JPL?

UF: My main work involved employing data mining techniques to help scientists and government organizations like NASA, the National Institutes of Health, and so forth. A problem in science is that agencies like NASA spend billions of dollars to collect data, and then once it arrives nobody has the tools to sift through it all. I wanted to figure out how to make all that data useful. We worked, for example, with Caltech to identify stars and galaxies in big sky surveys. And we were able to recognize the objects they were looking for better than anyone thought was possible at the time. We also worked with a planetary geologist at Brown University where we looked for small volcanoes on Venus, of which there about a million.

BLVR: How did that work?

UF: You need to look at small features on the planet's surface and figure out which ones are volcanoes, and

you can't do it all by hand. So we wanted to create an analysis tool where scientists could give examples of what they're looking for, and then send the tool off to run through the data and find likely candidates, and then catalog those events. Similarly, we did a project to look for rare objects in the universe like high-energy quasars. We were able to create tools that enabled the Caltech astronomers to find them forty times faster than they had been able to before. And that's because the data mining algorithms were doing the screening for them.

BLVR: So, say you're writing an algorithm to help look for volcanoes on Venus. How complicated is such an algorithm?

UF: That's an excellent question. If you take a geologist, and you sit them in front of the computer, they'll see these patterns in the data, and think, "Here's a small volcano." If you ask them, "what made you decide that?" They'll say, "I don't know: experience." This is where data mining comes in. If you try to design an algorithm specifically to describe the characteristics of Venutian volcanoes, or fraudulent credit card transactions, or anything fairly complicated, you could sit down and try to come up with all the different scenarios, but it would be beyond your reach. But if you have data, with some normal and fraudulent credit card transactions as examples, you create an algorithm to go in there and produce another model. It sifts looks for relevant variables and interesting correlations and tries to fit new models to them. And that new algorithm it produces will recognize fraud whenever it sees it in the future. Then you've done something that would be very difficult for humans to do. And that's the heart of data mining.

BLVR: Do you tinker with this process along the way?

UF: You do interact with these algorithms, and sometimes they go off base, as do all machines. But in general, if you have a huge data set, and you've got a

hard task, like recognizing fraud, where you have some examples, but you don't know how to state it formulaically to the computer, you need these algorithms to build the program that will recognize fraud.

BLVR: In the example you've mentioned, though, you still have a starting point. That's sort of like looking for Rumsfeld's known unknowns. What about the holy grail of security, the unknown unknowns?

UF: Right. Sometimes you don't even know what you're looking for, and Total Information Awareness is the perfect example. But, we've been doing that in the commercial world for some time. An example might be, a customer database from purchasing records. And you want to find interesting groups of customers, with similar behavior, but you may not know what that interesting behavior might be. So the algorithm goes through all the data, and clusters people together, and then finds descriptions for these clusters. And it comes back and says, "Here's a group of mostly, let's say, people who view all ten pages of the bargain section of a company's website, but who never buy anything." That's an actual group we discovered for a client. And this cluster wasn't known. They weren't looking for them. They were unearthed by the data.

BLVR: In the case of Total Information Awareness, you see a lot of optimistic rhetoric in the speeches and presentations of the group leaders of the various programs in Total Information Awareness. They say, "Look at September 11. After the attack, intelligence uncovered all the people, and the planning transactions, and the relationships, and if we had only been able to see that as a pattern beforehand, we would have been able to prevent the attack." And so I wonder, that if some kind of system had been in place, and these patterns caused a trigger, and intelligence analysts had looked closely, whether it would necessarily have meant

an intervention. I mean, there were probably dozens of suspicious patterns of activity of all kinds in August and early September 11th of 2001, only one of which was the right one. So the notion of Total Information Awareness being effective is premised on the idea that better information means you can process a signal out of noise. But it can only be so good. No algorithm can divine the intentions of twenty guys by their car rental and hardware store purchases. It can only be accurate enough to put red flags up for many scenarios.

UF: That's true, and let me give you a tougher scenario. It's one thing to rate a credit card transaction for its likelihood of being fraudulent. But as soon as you start looking for groupings of people, like terrorist networks, which is what Total Information Awareness needs to do, you now have an exponential problem. Because if there are N entities, there's an exponential number in N of possible subgroupings. You're looking for ten people, you don't know which ten, among millions, and that's an absolutely astronomical number of combinations.

BLVR: That's a whole lot of evidence extracting and link discovering.

UF: Yes, but it's not impossible. It's just very challenging. It's never been done on this scale. It's like putting a man on the moon. If they really want to get it done, it will require a lot of resources, the best people in the country, and so on.

BLVR: Let me get back to the patterning and clustering for a minute. It seems that Total Information Awareness is premised on the idea that perfect intelligence is possible. But, as we know from intelligence during the past fifty years, it's not. There are all kinds of potential hints and signals, and false moves, and poor decisions based on misread false moves, and so on. One example I read about recently is Israel, in the lead up to the 1973 war. Syria, Jordan, and Egypt, I believe, invaded. And

the Israeli government was shaken by its failure to predict the attack. After all, there were all kinds of indications, like Egyptian armored columns moving around, fortifications appearing, and so on. But, in fact, similar indications had appeared dozens of times in the months previous. And they all turned out to be false alarms. So, we can say that the pattern was obvious, but it's really only after the fact you can tell. Out of many similar intelligence patterns, only one was meaningful. Even within the range of what seem to be clear signals, there's still a lot of noise.

UF: It is a complicated problem. Total Information Awareness wants to apply computational tools to intelligence analysis, but it will essentially highlight things for the human intelligence analysts to look at. And there's no guarantee whatsoever that if a trigger is caused by some pattern, an analyst won't disregard as irrelevant something that turns out to be an attack plan in motion. But that doesn't mean that we give up on the problem. With intelligence, missed opportunities are probably inevitable. The difference is, if you make your instruments smarter, they get better at reducing the missed opportunities. And, in the case of Total Information Awareness, even a slight reduction in the probability of a terrorist being successful is probably worth it. What you have to watch out for is that you do it in the right way, so that you're not changing what's valuable about our life in this society, which is the personal freedom and protection of individual rights. So that may be the biggest challenge, that balance.

BLVR: Do you think that balance is possible?

UF: Yes. It's not a technology question, it's one of policy. Unfortunately, I don't think it's been thought through yet entirely.

BLVR: Whatever reservations one might have about the system being foolproof—

UF: It will never be foolproof. That's an easy one.

BLVR: Well, how accurate can these systems be—taking, say, scientific applications, where you started, as an example?

UF: Astronomers trying to decide whether faint objects are stars or galaxies had a seventy percent accuracy with traditional statistical methods. That wasn't good enough for advancing the field. You couldn't take seventy percent and publish a paper arguing a new theory saying Professor So-and-so is wrong about the structure of the universe. We were able to take data mining techniques and get closer to ninety-four percent accuracy. And if you have ninety percent confidence, you can write papers setting out new theories or abandoning old ones. Notice I said ninety-four percent, I didn't say 100 percent. There is no such thing as zero uncertainty.

BLVR: So, there's always some guessing.

UF: Yes, and sometimes guessing is the best you can do. In the real world, we guess all the time and it serves us well. When you walk around, your vision system is processing a whole bunch of signals in milliseconds and judging that a visual object is a wall, or an imminent cliff, or a car heading towards you. This might be disturbing to a lot of people, but some of those guesses are errors.

BLVR: You don't say. So, built into our basic reality is a certain amount of cognitive guesswork?

UF: Oh, heck yeah.

BLVR: Kant would be disappointed.

UF: The brain has a good error rate. But, the point is, you can function with that error rate. Animals do a lot of guesswork.

BLVR: What are some of the other commercial applications for this stuff? As I understand it, there's a whole set of complicated math that's used to calculate risk for things like insurance premiums. Now, there's no way to predict whether a hurricane is going to come along and destroy this house, but if the owner wants to take out a hurricane insurance policy, the issuing company has to assess that risk quantitatively somehow, and put a premium on it. Are some of the techniques we've been talking about in use in those cases?

UF: Absolutely. And in insurance, there's another application that's very similar to Total Information Awareness actually, where they look for fraud in claims. There are only so many inspectors, and there has been technology in place and working for some time to help determine which claims to inspect for fraud. Similarly, at AT&T, they need to detect identity theft—people stealing phone numbers—or things like that, and again, there's a huge amount of information, so you need an algorithm to point the humans toward the most promising places to look. There are applications everywhere: banking, health care, you name it. While still a graduate student, I was working with General Motors for a summer, trying to analyze their automobile repair data sets to figure out if there were patterns surrounding specific cars or specific faults. I wanted to see if certain types of repairs were strongly associated with a certain model. And that kind of problem is very similar to the one Total Information Awareness wants to solve. Because you have to go into the warranty databases, the informative stuff is in text fields, where there's free-form text.

BLVR: Where a mechanic would write something like, "Car fails at 55 mph in hot weather."

UF: Right. You have to go extract representations of those, and then apply the analysis algorithm and find interesting patterns.

BLVR: Well, those are two main elements of Total Information Awareness. Genoa I and II are about structuring data, combining databases, adding new ones, so as to make it all readable. And the Evidence Extraction part, which is what gets the most attention, is the computational part.

UF: You have to have proper data to start with. Otherwise, there's nothing to mine. So, data mining requires both.

BLVR: Which reminds me: What about the existing private data aggregators? You and I could go to these companies right now and buy pretty detailed files on each other. These companies are already doing some of what Total Information Awareness will do, on a smaller scale, and with data that's publicly available. But this is a huge business.

UF: That's what scares people about Total Information Awareness, as well. Because Total Information Awareness will integrate all kinds of databases. And if you're putting data together, you are creating the opportunity to potentially abuse privacy laws. In the United States, of course, privacy laws aren't very well defined. That's a problem, and another whole issue. But sometimes if you put some basic information, from public databases, together in one place, it suddenly seems invasive. So, the minute you combine multiple sources, you have to proceed very carefully.

BLVR: What are the ways to protect privacy?

UF: Let's say I give you a data set, about people. One of the fields will be the social security number. What is the first thing that a data mining algorithm will do with that field? As far as the algorithm is concerned, that number is uninformative, so it will disregard it. So, the beauty of data mining is that it's very possible to do privacy preservation—it's a whole area of research, in fact—and

that's because the algorithms do not need private information to do their work. What the algorithms do is look for patterns, and they'll trigger the users if something interesting turns up. Now, the decision about how and when to act on that trigger is a policy decision.

BLVR: But let's say a triggering pattern shows up. To act on it, you need to get in there and look at these very fields, the information that identifies whoever is behind a pattern that may be terrorism.

UF: Yes. And that will require lawmaking that determines when patterns constitute probable cause. And if you do that right, you'll have a safe mechanism. A lot of the debate, by the way, is logical on both sides. The people who want to move forward with Total Information Awareness, and the people exerting a healthy skepticism about potential privacy violations, are both right.

BLVR: Will it be possible to appease both?

UF: It's just getting started, so it's hard to say. It is possible, but the policy making has to be done right. I should give you a reminder that a lot of the problems that need to be solved to make Total Information Awareness work are outstanding. We don't know how to do a lot of this. There's plenty of basic research to do.

BLVR: What's the horizon of plausibility? Ten years?

UF: It's possible in a few years. It's a question of commitment. Humans are very good at making algorithms work eventually.

★

Joshuah Bearman is a writer and, along with his girlfriend, a jewelry proprieter. He lives in Hollywood, where he is a contributor to the *LA Weekly* and other publications as time permits.

THE BELIEVER