



**Dr. Michael Pagels**  
**Program Manager**  
**Information Exploitation Office**

## Avoiding Death by Data

Data, data, and data.

How are you going to use 400 terabytes of intelligence? The following are some of the ways we at IXO would like to see it used:

Bob Tenny's presentation raised questions like "What's around the corner?" and "Is that car coming toward me full of bad guys?". The answers, if you can get to them, are somewhere in that 400 terabytes.

Think about continuously updating the operations and information about a ski mountain, and having it summarized and projected onto your ski goggles. From thermal imagery, you can see where new snow is being produced. By incorporating published material, you can find the closest spots serving lunch, or schnapps. Motion detection and tracking keeps you informed on how busy your favorite slope is and provides a prediction for a better time to avoid the crowds.

A few others to think about: disaster relief operations and planning, facilities classification and the search for new or missing WMD, tracking and monitoring of rain forest destruction, tracking terrorists in order to predict and preempt their attacks, and security and safety planning for very large events like the Olympics.

Today, rather than finding the answers in those 400 terabytes, we're generally crushed by the data. How do we save ourselves?

It's about models of the world. Whenever we move about in the real world, we are interacting with an internal 4D model. The model is built from what we know from the past and, using senses to update that model, to refine a concept of the present, and from that, to predict the future.

We maintain a 4D model of the world in our heads and during every waking moment we update it with information from our senses, identifying objects, and reasoning about the relationships among those objects. It works so well in our brains, but how do we make it work in our exploitation systems?

The thesis is this: To effectively exploit the vast amount of information we currently collect and to exploit the increase we will see in the future, we need to develop the technologies to continuously identify all the objects, and make the necessary associations, in order to identify all the relationships among those objects, in a vast 4D model of the planet.

Of course, certain sites require more intensive and complex modeling than others: Cities for example. Many of our problems require you to model everything at once to catch the warnings indicated by subtle change. Remember the scope of the problem: at least 400 terabytes of new data a day.

Let's say we are actually able to create and maintain a detailed 4D, very large scale site model in our heads, a city model. What do we do with it? Getting meaningful information, not data, not pixels, not dots, but meaningful information out of the model is the real goal; the very reason it exists.

We need to know what's there now, and if it's significantly different from what was expected. We also need to know when our conclusions reflect reality, and when those conclusions break down.

What's going to drive the solution?

Here are five drivers for the development of large-scale model-based exploitation technology.

## Avoiding Death by Data

### 1. A huge volume of data

Give or take a few orders of magnitude, how much data should we expect? Hundreds of terabytes a day. Four hundred terabytes is a nice number. That's approximately 150 petabytes annually, or 150 times the historical content of the entire Internet through 2004, or every person in the LA urban area taking one digital photograph every second for a year. Hundreds of terabytes a day.

### 2. A vast range of classic data sources

Where's it all going to come from? All the classic sources: imagery, signals, newspapers, and all the other feeds, and all their next-generation follow-ons. The Internet and new versions of the Internet. The "warrior as sensor" will be feeding gigabytes of data per day. Video and thermal imagery, sensed chemical signatures, position and movement data.

### 3. Tackle ignored or underutilized sources

There are writers, graphic artists and cartographers who represent complex context and associations in text and graphics. Look at any metro map and consider the amount of information it contains and how hard it is to associate it with the real world from the perspective of a sensor.

### 4. Need for speed

We need to stay operationally relevant. If you're at the pointy end of the spear, things can happen fast. Of course, handling 400 terabytes of new data a day suggests you have a handle on basic scalability. If you can do that, it's unlikely you'll want to throw anything away. That form of scalability is not the same as needed to quickly supply the correct information for a specific, and probably unique, situation. There's a huge query, search, and access problem here.

### 5. Need for interpretation

The availability of vast amounts of data does nothing to increase a human's ability to process and interpret it. Like an aircraft cockpit, we need to avoid information overload by providing the right

information to the right place at the right time, extracted and summarized based on the context in which it will be used.

Five capabilities need to be fundamentally transformed.

#### 1. Data management

There are some huge databases floating around; check out Wal-Mart and its inventory control systems, or any of the high-energy physics experiment databases. In fact, CERN's Large Hadron Collider is expected to generate 15 petabytes of data annually. Simply storing the underlying data isn't necessarily a problem any longer; the problem with bulk data is more in its caching and distribution. Has the technology in the management of federations of very large databases progressed to where we can rely on them as commodities? Is an exabyte fundamentally different than a petabyte? Caching, based on access statistics, doesn't scale broadly enough to be a solution. We need to extend the cache's predictions much deeper into the exploitation workflow to understand upcoming data requirements.

#### 2. Spatialtemporal information registration

Data used for the update consists of everything from imagery to news reports. To get labeled and associated information, we need to place the data in both space and time. For some things this is easy. A GPS beacon can tell us a known accuracy, where it is, and when it measured that position. For most data sources, even high resolution imagery, knowing what you're looking at, the label, and knowing exactly where it is, is generally an unsolved problem. How do we bring the accuracy of locations or times mentioned in textual reports, "the small white car next to the embassy yesterday," to a level supporting consistent interpretation?

#### 3. Management of uncertainty

We must understand that all the source data are only semireliable. Often we understand the sources

## Avoiding Death by Data

of uncertainty; they may be innate to the sensor. Sometimes they may be the accumulation of multiple errors through the use of multiple sources. These are fairly straightforward to handle, and the mathematics are straightforward. The difficulty lies in generalizing these techniques to incorporate less obvious errors, and in making all of these uncertainties visible within the model and to the user. Every conclusion drawn in, or from, the model should explicitly carry along the provenance of that conclusion. Meaning drawn from semireliable data is always provisional.

### 4. Definitions of change

This is the area with the greatest opportunity for fundamental improvement is in the detection, analysis and detailed understanding of change. What is a change, and when is it significant? Between any two collections or two sources, you're likely to discover lots of changes. Which are significant, and perhaps more important, in which context are the changes significant, in largely, undiscovered country. It seems there is almost no hope in understanding change unless you're projecting your data onto a high fidelity model. The model provides a place to capture the context necessary to determine a change's significance.

This is true both spatially, for example casting of shadows, and temporally, why shadows move during the day. The solution should have multiple models linked to a common spatial-temporal framework, supporting reasoning over multiple scales and domains.

When is a change not significant? Change analysis and understanding is also at the core of how we should update our models. New data that is totally consistent with our model is easy; the model is still locked onto reality. Data that is a little bit inconsistent could just help us update our model's error characteristics, or it could be the start of an important trend. Data that is very inconsistent could be a data collection error, or traceable to a real underlying change. How do we know? We

have lots of examples with simple models; do those techniques scale?

It appears we've constrained our change analysis to an area that is too local. We are reasoning with too few examples, on too simple a model. By broadening our model's extent and complexity, we can uncover associations that explain subtle trends.

### 5. Query and exploitation

Once we have this wonderful, continuously updating 4D model of the world, how do we exploit it to get the answers to relevant questions? Clearly we're not going to be pouring over the raw pixels, even if the data was all pixels.

An information exploitation query language? Information exploitation needs a revolutionary technology, similar to what the Structured English Query Language, which eventually became SQL, did for classic databases. For the portions of the model that are expressed in a network symbolic way, we can extend the technologies of the semantic web. Image search and retrieval by similarity is also promising for the graphical portions of the model. But we need a unified approach, semantic, geospatial, temporal, all at once, able to query all of the information in the model.

What do we track for provenance? Management and explanation of uncertainty has to be kept at the forefront. So how should that look to the analyst, and how do we capture it automatically in their workflow? Whenever someone examines a report, they should have access to all of the information that went into forming those conclusions, and access to the conclusions underlying uncertainties.

How do we interact with these updating models? We've also just scratched the surface when it comes to the human computer interface used for information exploitation. We're building a high resolution, content rich model of the world; exploitation should be an immersive, multi-sensorial experience. But, perhaps, not for everyone. The warrior in the field can't get too

## Avoiding Death by Data

distracted; augmenting his reality needs to be very selective. But the same sensitivities should apply to all users, the data and information should be tailored to their mission.

While relying on many of the same sources, the needs of a geospatial analyst verifying logistics models are quite different from those of an S-2 conducting his intelligence preparation of the battlefield, or a commander preparing a battle plan.

As you've so often heard, the goal is the right information at the right time to the right place.

Pushing your data back to your model gives you the context in which to convert data to information, and information to knowledge. That's what lets us answer "What's around the corner?" or "Is that car coming toward me full of bad guys?"