CHAPTER ONE

Why Won't They Leave Me Alone?

When you read a description of a book online at Amazon.com, Amazon helpfully informs you that many people who bought that book bought certain others, too. This little trick is a simple example of how a rapid, large-scale quantitative analysis of facts like names and numbers can tell us a lot about what people do and how they behave.

Given the state of the art in data mining, there are a few different ways that Amazon might handle the task. However the process unfolds, it must begin with a concise fact: a unique identifier, which Amazon can supply, for the book you're reading about at Amazon's site. The hard way—in terms of computer resource consumption, meaning time and money—is to use that identifier to search Amazon's entire purchase database, right then and there, and find all the customers who bought that book. Amazon sold $2.5 billion worth of books in 2000. Even with powerful computers and the identifier in hand, it will take a while to find them all (probably more than most customers care to wait online). Assuming the look-up is done, Amazon can then look up all the *other* books those customers bought, sort and rank them by various factors (such as total purchases across all customers for each book), and present a short list of candidates for your review (and ideally—from their point of view—your purchase). To make it all really slick, Amazon might eliminate titles it knows you have already bought from Amazon. That's something they apparently don't do now, at least if my experience is any proof.

There's a less time- and computer resource-consuming, more likely approach. Amazon could do a full-scale read-through of their transaction database nightly, weekly, monthly, or however often they like. They would see what was purchased and do the same look-up described of all the other products those people bought as well. They would use that information to build a database of books-affiliated-by-purchase that they could reference quickly whenever a new purchase is made. That approach would save them the trouble of building such a database on the fly whenever a customer looks at a book description. It would explain why they don't pick up on the fact that you have already bought one or more of the books on their list. And it could be made to work, for every online customer, in less time than it took to read this paragraph.

Anyone who has shopped at Amazon probably remembers being surprised the first time Amazon presented such a list. The thing that surprises many people is that the list Amazon shows them is often immediately credible, because it includes books that they've already read and enjoyed.

How does Amazon know so much about you? You never told them what you *liked.*

You didn't have to. They knew it almost as soon as you selected your purchases, even before you gave them your money.

THE POWER OF NAMES AND NUMBERS

Facts like names and numbers are precise, quantitative, and unequivocal. They're about what people and machines *do,* not what people *think. Customer (John Smith) bought product (X) in quantity (Q) at price (P) from vendor (V) using channel (C) at time (T) in location (latitude, longitude) with credit card number (NNNN).* The purchase is compact and meaningful. We don't have to know *why* it happened to predict with some accuracy when and under what circumstances it will happen again.

What people do often says more about who they are and what they think than what they think they think, and what people *say* they think doesn't necessarily tell you what they'll *do* next. Lots of people who say they care about privacy hand out detailed personal information to anyone who offers them a piece of free software, for example. Even before they've seen the software, even before they know (or think to ask) the uses to which the information will be put, they've shared their personal data.

Amazon isn't telling you what other buyers *think* about the books Amazon is recommending. Reviews are available, if you want them, but that's not how Amazon came up with the recommendations. It's not about what people *liked*. Amazon is telling you what other people *bought*. That information is easy to collect because it's an intrinsic part of every purchase transaction, and it's easy to analyze compared to any ratings that a diverse set of customers might apply to a book that they've all read. (Every customer has his or her own rating system, and Amazon doesn't know what it is. But a purchase is a purchase is a purchase.)

If given a wider universe of data to work with, Amazon might also find that people who bought certain books tended to rent certain videos, or drink certain coffees, or travel more frequently than others to particular locations. Knowing those preferences could open up entirely new avenues for Amazon's recommendations. *Can I add a double latte with cinnamon to your order? Would you like to drink it in Rio de Janeiro? Wearing a scarf in a certain shade of red?* It's neither possible nor necessary to predict all the associations that might turn up. *The power of large-scale analysis of simple facts is precisely that it reveals such patterns.* The technology that makes the analysis possible, *data mining,* is available now in a very robust form, and it's getting stronger.

Amazon doesn't have an infinite universe of data. It has the stuff it can generate from its own sales, plus whatever else it can buy or rent from third parties. (If Amazon were less ethical, we could add: plus whatever it could *steal* from third parties to that list.) Amazon doesn't have everything.

But the universe gets bigger all the time.

WHAT DOES IT TAKE TO CREATE A UNIVERSE?

Databases essentially consist of *attributes*—pieces of data—and *relationships*—the rules that describe how the attributes relate to each other. A *key* is an attribute that uniquely identifies an instance of a certain set of related attributes. A good key is unique, and the data that depend on a good key depend on all of it, not just a part of it. (I'm trying to make this simple, and it'll end soon, I promise.)

Your name is a good *attribute* for referring to you in a message, but it's a poor *key* for correlating information about you—your address, weight, height, and spending habits—because any number of other people might also have your name. Your address is a good *key* for referring to a location, so long as the whole key—street address, city, state, and country—is present.

ABOUT DATA MINING

Data mining—intensive statistical analysis of large masses of structured, factual data—is one of the most important technologies in the World Without Secrets, and it's already mature. Data mining makes *patterns*—patterns of spending, patterns of fraud, patterns of movement of all kinds—buried in huge masses of data immediately obvious. Once a pattern is visible, people can *act*.

In 2001, there are commercial data-mining operations under way that can crunch dozens of *terabytes* (trillions of bytes, each one a thousand times the size of the *gigabytes* that now denominate hard drives on personal computers) of data in a day. The capacity of those systems is driven by Moore's Law, so the numbers will go up rapidly. At Gartner, we've already been briefed on plans for data-mining systems that will crunch hundreds of terabytes daily, starting less than two years from now.

Data mining on big masses of data demands lots of processing power, and so far it's mostly been in the hands of large businesses and governments that can afford the big specialized machines needed to crunch all the data. That's changing. *Big businesses and governments are no longer the only sources that can mine mountains of data for meaningful patterns.* Already, technical solutions that put data-mining capabilities into the hands of much less well-funded organizations are available. Grid computing approaches that break big processing problems down into tiny units and distribute them to thousands of personal computers are being used by projects like *SETI@Home* (which enlists individuals and their computers in the **S**earch for **E**xtra-**T**errestrial **I**ntelligence). *SETI@Home* ties thousands of privately owned personal computers into a network of machines that can process and analyze massive amounts of data, in parallel, one piece at a time. It's data mining *for* the masses, using computing power supplied *by* the masses.

By 2010, well over two billion Pentium-class-and-above personal computers will have been sold worldwide. Many more may have been distributed for free; by that time, a Pentium-class processor will cost less than a dollar to manufacture. That's a lot of processing power, and grid computing technologies will make much of it available to nonbusiness, nongovernment players. *Data mining will be available to anyone who can convince enough people that the purpose of the mining is important.* It's an important contributor to multiple trends in the World Without Secrets, including the Exception Economy and the Network Army (as discussed in Chapters Nine and Five).

Drop the state and country and there's room for confusion. (If you live in greater Boston, Massachusetts, where there are five streets named "Arlington," you need a zip code too.)

Here's the most important thing. Databases can be linked, or *related,* when a *key* value is common to both structures. It doesn't really matter whether information is stored in separate physical databases. *All that matters is the keys.* If the same keys are present in two different databases, any information in one can be correlated to any information in the other, *as if they were a single database,* at least in a *logical* sense.

It's pretty easy from here on out. If you want to pull a universe of data together, the first thing you need is a *really good key* that ties the data to something in particular. That something is usually a person, but the person's name alone is not good enough. You need something unique, something that's usable in lots of different places—ideally, something that's already used in lots of places. In this age of global business, you also need something that's unique worldwide.

CROSSING OVER

Database designers talk about *logical* databases—databases that exist in an ideal sense, unfettered by the considerations of available technology, and constrained in structure only by the nature of the information itself—and *physical* databases—logical designs that are restructured and constrained by the needs of particular applications and the technology they use. A logical database is like an artist's drawing of a piece of architecture. A physical database is like the building people live in after all the construction is done.

In a *logical* sense, the ideal identifier is an arbitrary number that's big enough to include a unique value for everyone who might need to be identified. In the *physical* world, the closest thing anyone has to a worldwide *key* for lots of data that matter—concise, factual data that link people to their purchases—is a credit card number.

Visa has issued more than one billion of its various credit and debit cards worldwide. Visa has 60 percent of the worldwide market, so we can figure that another 700 million or so credit cards from other vendors are also out there. Visa says that its cards "are accepted at more than 21 million locations in 300 countries and territories, making Visa the closest thing there is to a universal currency."[1] Every transaction done anywhere

in the world using a given credit card can be positively correlated to every other transaction made with the same credit card.

The purchase data don't tell me everything about you, but they give me a good start. I know where you've gone and when you were there. I know where you shopped and how you "paid" for your purchases. Do I need to know much more about you? If I do, I can always ask the people who work at the place where you shopped. If necessary, I can pay them (or coerce them) to tell me. If utterly necessary, I can buy the company they work for.

Arguably, an even better key than a credit card number, assuming that you have the technology to process it efficiently in all the situations where it might be needed, is something that's both unique and intrinsic to your person, like a digitized replica of your face, your voice, or your DNA. In Chapter Two, "Streets Without Secrets," I discuss the potential for widespread use of biometrics like facial scans as keys to a universe of personal information.

But facial scans aren't essential; they are merely useful, convenient, and likely to be deployed in many situations. To anyone who's willing to pay the going price or has a list that can be swapped, credit card numbers give access to a wide range of very useful and highly predictive information about card owners' behavior and habits. We resist a single identifier when it might be in the hands of a government, but we welcome it when we can use the same credit card in Jakarta, Denver, and Bonn without any more effort than it takes to present it.

The demand for common identifiers to support secure global commerce is accomplishing what no government could: the worldwide implementation of what is effectively a unique international personal identifier.

MORE DATA, MORE POWER, FEW CONTROLS

A worldwide identifier enables a wider universe of data, a market where businesses can buy, sell, and combine information about individuals, subject only to what they can afford (information is precious), what is legal in the nation(s) in which they do business, and what they believe the public will tolerate.

As an example of U.S. businesses' freedom to manage and trade information as they see fit, let's look again at Amazon.com. In September 2000, Amazon informed its entire customer base that, contrary to a

previously announced policy, Amazon would begin sharing information about its customers with selected third parties. Customers could choose to end their relationship with Amazon, but customer data already gathered by Amazon would be subject to any uses that Amazon deemed appropriate. Amazon described one of those potential uses as follows:

> *Business Transfers:* As we continue to develop our business, we might sell or buy stores or assets. In such transactions, customer information generally is one of the transferred business assets. Also, in the unlikely event that Amazon.com, Inc., or substantially all of its assets are acquired, customer information will of course be one of the transferred assets.[2]

This passage apparently contradicted Amazon's statement, earlier in the Notice, that "Information about our customers is an important part of our business, and we are not in the business of selling it to others." In other words, Amazon reserved the right to change its mind, anytime, about how it uses customers' information. (The statement to customers was issued on the occasion of such a change.) Nothing in current U.S. law or regulatory policy prevents Amazon from doing so.

In the European Union (EU), where laws demand customers' approval of the uses to which their data are put, Amazon might not have been able to change its policy so easily. Criminal, as well as civil, penalties apply, in the EU, to companies that permit sensitive information (like an identifier or a credit card number) to be used in ways that aren't specifically authorized by the original owner of the information—the person the information *describes.* But there's little evidence that the United States will follow Europe's lead soon. And in a global economy, where a company taking an order via a phone or the Internet might be located almost anywhere, information can easily migrate to a place where restrictions are even less stringent than those imposed by public opinion in the United States.

I interviewed Victor A. Kovner, a First Amendment authority and former Corporation Counsel of the City of New York, in October 2001, and I mentioned Amazon's policy change to him. "That's why I don't buy on the Internet," he said drily, and I laughed.[3] But Kovner missed the point. It's not about the Internet, and it's not about Amazon. It's about anyone who uses a credit card, and it's about any company that accepts one.

Amazon didn't do anything that any other company couldn't do. Data arrived via the Internet, but had they come over the phone or in the mail, it wouldn't have made any difference.

UNSTOPPABLE MOMENTUM

In the aggregate, the amount of electronically stored data about individual behavior is massive, detailed, and growing. It includes what we buy, where we buy it, where we go to eat and to entertain or educate ourselves, the people we call on the telephone and how long we talk to them, the correspondence we receive and send via e-mail, the names of businesses and individuals we correspond with, the content of the correspondence, the addresses of Web pages we visit, and the amount of time we spend at each address.

The stored data will continue to grow. Intelligent devices and electronic communications provide too much apparent value for most people to ignore. We want to be as productive and comfortable as our machines can possibly make us, and no one wants to be left behind or left out. (I shuddered when a friend told me recently that her attorney, in prosperous Fairfield County, Connecticut, had neither a fax machine nor e-mail. Who among professionals—except total losers—has no *fax?*)

Commercial initiatives like Microsoft's .NET, by promising even greater convenience at the cost of massive consolidation of the keys to one's personal information, will raise the risks and rewards even higher. Are you willing to put all of the keys anyone needs to do business in your name in the hands of Microsoft? Are you willing to let Microsoft touch every transaction you do? What about someone else? *Anyone* else? Is the *convenience* of being known everywhere worth the *risk* of being known everywhere?

The demand-side alternative is to restrict the information rights and privileges of enterprises, probably via legislation or regulation. For much of the world, that's even less likely. The trend in the industrialized world is toward *less* regulation, not more (regulation related to national security excepted). Public opinion in developed nations is also not fully mobilized against widespread data collection and profiling, and, in the absence of a full-blown disaster, it may never be.

Cultural values in newly industrializing nations make such restrictions even less likely there. Many such societies have tended to be

authoritarian and male-dominated. They're not likely to force heavy restrictions on new businesses, especially if those businesses are competing with American companies that can do anything they like where information is concerned. They'll consider information to be the rightful property of the people in charge in both the public and private sectors. There will be few legal or regulatory restrictions on information ownership and use in developing nations during the next 10 years.

BY THE NUMBERS

Why does it matter that businesses have so much information so readily available to them? So Amazon knows what you want and can offer you books you like, instead of trying to make you buy books you don't like. So *what's the problem?*

Is it scary if I'm walking down the street and someone offers to sell me something? Probably not. It's not very scary that people are trying to sell me things. It's not even scary if they're trying to sell me things all the time, which they already apparently are, or if they're not very tasteful in the way they go about it. Sales aren't scary unless you're a salesperson.

Is it scary if I walk down the street and my face is scanned by a camera belonging to a salesperson? It might be offensive; it's probably legal. Is it scary? Maybe not. What does the salesperson know about the person behind the face?

Is it scary if I walk down the street and my face is scanned by a camera belonging to a salesperson, and the scan can be compared automatically to a scan of my face that's on file with a bank or a credit card company? Now, the salesperson may know a good deal more about me: my name, where I bank, where I live.

Is it scary if the camera *doesn't belong to a salesperson?*

In an increasingly consolidated, global, networked economy, information moves everywhere. Sooner or later, it moves to a place where the owner—or anyone in current possession—can do whatever he or she likes with it. That party might be ethical—*might.*

If we've learned one thing from terrorists, not to mention action movies, it's that a tool is also a weapon. Globally accepted credit cards and the databases that support them are tools for taking the friction out of commerce. That's another way of saying that they're tools for extracting money from people with minimum effort on everyone's part.

So it's not a problem if they're trying to sell me stuff. And it's not a problem if it's Amazon. But it might be a problem if it's neither.

I haven't mentioned *identity theft* yet, but surely that's what this is leading to. Identity theft is unauthorized use of the information that identifies me, in order to perpetrate fraud. The more widely my information is known, the greater the number of places where it may be found, and the more likely that more theft will occur. The more information is correlated to a single identifier, the more damage an instance of identity theft may cause.

Identity theft is much feared in our society, though no one has ever died from it or been ruined by it. There are worse things than identity theft, and a universal identifier may lead to those worse things as well.

Wherever universal identification leads, we don't yet know how to manage a world in which *everything can be linked to me, wherever I am.* We don't yet know how to balance the undoubted convenience of this world with the peril—vague, but apparently near—that we sense in the presence of all that information combined and consolidated, if only *logically.*

WHERE DID THE SECRETS GO?

The boundaries are down. Ubiquitous monitoring is technologically feasible and will soon be economically feasible. Any limits that exist will be limits set by agreement and reinforced by constant oversight. Those limits must ultimately be international.

Computers constantly and geometrically increase in power—the power to know and to communicate what is known—while their physical size shrinks. The rate at which they do both is described by Moore's Law, one of the best-known formulas of the second half of the twentieth century. Moore's Law stipulates that the computing power of a transistor of a given size doubles every 18 months. The trend is so well established that we take it for granted that it will continue, or even accelerate, into the indefinite future.

The result isn't just that we're increasingly surrounded by computers of all kinds, including computers that are skin-close—closer than a cell phone in a pocket. We're surrounded by buildings most of the time too, and that's not a big deal. This is more like being surrounded from the *inside out.*

A *system* now consists of nothing more than all the machines that are plugged in and talking to each other. A system can change on a momentary basis. We're not just within the boundaries of a system. We *are* the boundary. *It* moves when *we* do.

Security is about *control within a boundary.* If the boundary is constantly shifting and is impossible to define or predict, what does that imply about security?

Is that why we feel so insecure in the midst of so many powerful machines designed to do our bidding? Or is it that we're not sure whose bidding the machines are really doing at any given moment?