Data warehousing projects usually are difficult and expensive efforts that tend to lose direction because they challenge both business and IT managers with a new and different way of viewing information. Whereas both communities have successfully used computers for the past 50 years to help understand what happens in the day-to-day operations of their business, now they are trying to use computers and large amounts of data to help understand why things happen in their business. Fortunately, historical precedent can be helpful in understanding the issues involved.

In essence, data warehousing is an observational science. This point is an important one because observational science has been around for 4,500 years and provides a wealth of experience in acquiring and analyzing large amounts of data. Thus, it brings a sense of history to data warehousing that currently does not exist. As a member of a data warehousing team, it is very comforting to know that many others have gone before you and have struggled with similar challenges and, despite all, have succeeded.

**Data Mining Deja Vu**
Twenty years ago I was an exploration geophysicist working for a major oil and mining company. Exploration geophysicists collect large amounts of observational data and search for trends and patterns in the earth's gravitational and magnetic fields, electrical resistivity, and the response to manmade seismic (sound) waves to locate commercial deposits of oil and minerals. Over the years, my career wandered to supporting geophysical analytical applications as a programmer and then to finally becoming an IT professional.

In March 2000, I became involved in my first data warehousing project at a major airline. I immediately felt strangely comfortable with the concepts of data warehousing. Then it finally dawned on me that the source of my deja vu stemmed from "data mining" mining data about 20 years ago! Once the concept of data warehousing as an observational science crystallized in my mind, I was able to benefit from my past experience as an observational scientist and the historical experiences of other observational scientists.

Clay-Tablet DBMS

There are two kinds of science: experimental and observational. Physics and chemistry are examples of experimental sciences because they deal with small isolated systems that can be studied through experimentation. Galileo invented experimental science about 400 years ago to explore the physics of falling bodies.

Observational sciences deal with large, complex systems that cannot be experimented with because they are too large and too complex - astronomy, meteorology, geophysics, and epidemiology, for example. Observational science is much older than experimental science and goes back thousands of years to early Chinese astronomical observations around 2500 B.C. By 1600 B.C. the Babylonians were plotting the positions of the fixed stars; by 800 B.C. the Babylonians were recording the motions of the planets relative to the fixed stars. By 200 B.C., the Greeks had used astronomical observations to figure out that the sun was at the center of the solar system, the earth and planets revolved about the sun, the moon revolved about the earth, and they had determined the diameter of the earth to within 5 percent.

Observational sciences try to figure out how complex things work through an organized process of discovery. (See Table 1.)

Nothing has changed much with this discovery process over the past 4,500 years. Although the DBMS of choice for the Babylonians was clay tablets, the process was the same.

Cholera: A Case Study

A classic case study in early data warehousing as an observational science is the birth of epidemiology in 1854. We now know that cholera is a terrible disease caused by bacteria that enter the intestines from contaminated water. The bacteria release a toxin that gives the victim a severe case of diarrhea. A 200 pound man can lose 40 pounds in a single day. Massive dehydration causes the blood to get so thick that the heart can't pump it and the patient dies quickly. Within two to three days, more than 50 percent of the victims may have perished.

In 1854, everyone believed that diseases were caused by "miasma." Miasma was a substance found in bad smelling air. Through folklore and anecdotal observations, people
knew that if you were around foul-smelling sick people in hospitals or were exposed to foul-smelling sewage, there was a good chance that you would get sick as well. People also observed that men who made a living cleaning out cesspools and laborers who boiled down rotting horse carcasses for glue and tallow would get sick and vomit from the extremely foul air that was part of their daily work.

The "Great Stench"

In the 1850s London had several million inhabitants and all the sewage from those people ended up in cesspools and ditches and eventually got into the river Thames. In the summer of 1858, the smell from the river became so terrible that they called it the "Great Stench of London." The House of Parliament hung blankets treated with chemicals in the windows of Parliament to cut down on the smell from the river.

Of course, the real problem was that most of the inhabitants of London got their drinking water from local shallow wells and ditches that were contaminated with sewage. The "high tech" residents got city water piped in from the river Thames. At the time, there were several water companies supplying city water to London residents. However, the Southwark and Vauxhall company and the Lambeth company were two water supply companies that played an intriguing role in our current understanding of cholera. Both companies drew polluted water right out of the Thames. Then in 1852, the Lambeth company moved its water intake facility 22 miles upstream from London and unknowingly began providing some London inhabitants with uncontaminated water.

Enter the Data Mining Doctor

LIFE-SAVING ETL

Early data mining efforts


"As the Registrar-General published a list of all the deaths from cholera which occurred in London in 1853, from the commencement of the epidemic in August to its conclusion in January 1854, I have been able to add up the number which occurred in the various sub-districts on the south side of the Thames, to which the water supply of the Southwark and Vauxhall, and the Lambeth Companies, extends. I have presented them in the table ... arranged in three groups...."
John Snow was a physician practicing in London during this period. While tending to cholera victims, he came up with the crazy idea that cholera was spread from one victim to another through contamination of drinking water. Snow published a pamphlet in 1849 detailing this theory, but nobody paid any attention to it because it contradicted the well-established miasma theory of disease. Snow's theory was obviously wrong because of the dilution problem, according to the thinking of the time. If a group of people drink a poison that later kills them, and before they die, they excrete some of the poison, the poison will soon get diluted to a safe level as this process is repeated over and over. Snow's theory implied that somehow the cholera poison had to grow within a victim and that was a ludicrous idea at the time.

Not to be discouraged, Snow decided upon a different approach following the 1853-54 cholera epidemic. Snow took the death certificate data collected by the London Registrar-General and created a data warehouse with it. The job of the Registrar-General was to collect operational data for the city of London such as marriages, births, and deaths for the purpose of taxing people. Snow did some extraction, transformation, and loading (ETL) on the cholera deaths by tabulating the addresses of the victims and determining from where they received their water. He then compared the cholera deaths in the 1849 epidemic to those in the 1853-54 epidemic by water-supply category in a tabular report. Snow called this analysis his "Grand Experiment." See the sidebar, "Lifesaving ETL," for a brief extract from Snow's classic paper.

<table>
<thead>
<tr>
<th></th>
<th>Number of Houses</th>
<th>Death from Cholera</th>
<th>Death in Each 10,000 Houses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwark and Vauxhall Company</td>
<td>40,046</td>
<td>1,263</td>
<td>315</td>
</tr>
<tr>
<td>Lambeth Company</td>
<td>26,107</td>
<td>98</td>
<td>37</td>
</tr>
<tr>
<td>Rest of London</td>
<td>256,423</td>
<td>1,422</td>
<td>59</td>
</tr>
</tbody>
</table>

"The experiment, too, was on the grandest scale. No fewer than 300,000 people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity."
The result of Snow's Grand Experiment was that before 1852, your chances of getting cholera were not correlated with getting your water from either water company; but for the epidemic of 1853-54, your chances of getting cholera if your water was from the Southwark and Vauxhall company were more than eight times greater than if you got your water from the Lambeth company! See Map 1.

Later in 1854, Snow had yet another opportunity to do some data warehousing. Cholera reoccurred in the Soho district of London. About 600 people died from cholera in a 10-day period. Once again Snow took the operational death-certificate data from the Registrar-General and this time he plotted the data on a clustering diagram instead of presenting it in a tabular form. He used a stacked histogram technique plotted on a map of Soho to do the data mining. Based upon this map, Snow was able to convince the London Board of Guardians to remove the pump handle from the public pump located on Broad Street. The outbreak of cholera subsided with this operational change. It was later revealed that the Broad Street well was contaminated by an underground cesspool located at 40 Broad Street which was just three feet from the well. The Broad Street pump without a handle remains today as a tribute to Snow. (See Map 2.)

Remarkably, Snow was able to do realtime data mining while people were dying and make an operational change on the fly in 1854. Unfortunately, old ideas die slowly; it was not until the Public Health Act of 1875 that the construction of proper sewage and water supply systems was mandated by law.

**Discovery Through Observation**

Data warehousing is a fledgling observational science that is less than 10 years old. So I believe there are many things that you can learn from the other observational sciences. For example, the current understanding of the customer base for many corporations is largely based upon folklore and anecdotal observations, much like the ideas surrounding cholera and miasma in the 19th century. Recognition of this problem has led to CRM efforts at many corporations.

For example, the airline industry has traditionally viewed its most valued customers as those who fly the most miles. Thus frequent-flier programs have historically been based upon miles flown with the airline. Nevertheless, closer examination of customer data reveals that customer value is a much more complicated matter and needs to be based upon past and projected revenues and costs, household information, the customer's influence on others, and the market in which the customer travels. Similarly, the banking industry has traditionally viewed its customer base as a large collection of unrelated accounts. By tying these accounts to actual people and understanding what banking services these people need, the banking industry has been able to up-sell and cross-sell products to its existing customers.

**Data Purity**

Another lesson to be learned is that the observational scientific method is much more important than the DBMS used to store your data. After all, most of the observational sciences got this far by using clay tablets and paper. In the observational sciences, data is treated with almost religious reverence. This stems from a strong master-apprentice tradition that is passed down through the generations from professors to their students. This tradition creates a solid ethical barrier against distorting data or drawing conclusions beyond what the observational data can truly support. The sanctity of the observational...
scientific method builds credibility, and this is a desirable quality for all data warehouses. There must be a strong element of trust in the data in a data warehouse for it to have business value.

**Error Handling**

The one thing I have not seen in data warehousing, which plays a key role in the other observational sciences, is the concept of observational error and error propagation. All scientific observations are subject to errors. Kepler was the first to recognize this problem in about 1600 when he was trying to work out a very accurate plot of the orbit of Mars. Kepler came up with the idea of "good enough for engineering purposes." This is a hard concept for people who deal with commercial applications. If you received a bank statement that said your account was $150,000 plus/minus $30,000, you would be very upset. However, if you were able to tell a CEO that John Doe will spend an additional $150,000 plus/minus $30,000 dollars with your company if you decrease delivery time by one week, you have delivered a good business value. Observational scientists have figured out ways to deal with observational errors by eliminating systematic recording errors, reducing random errors in the recording process, and increasing the signal-to-noise ratio of data in the data analysis step through various mathematical techniques.

**Hire an Astronomer**

Furthermore, it might be a good idea to add an astronomer, epidemiologist, or exploration geophysicist to your data warehousing team to bring in fresh analytical ideas. Many of the observational sciences have benefited from the cross fertilization of ideas.

For example, the observational science of modern geology was invented by James Hutton in 1788 when he published his book, *Theory of the Earth*. Between 1788 and 1965, geologists mapped rock-formation outcroppings over a good portion of the earth and obtained a great deal of observational data from oil well bore holes. Analyzing this data allowed geologists to figure out what had happened over the past billion years of geological history.

However, by 1965 they were stuck. The geologists could not figure out what was creating mountains, earthquakes, or volcanoes. Then around 1965 some geophysicists began to analyze magnetic data collected by oceanographers who had dragged magnetometers behind their research vessels in the 1950s as they steamed across the oceans. This missing piece of observational data allowed the geologists and geophysicists to realize that mountains resulted when giant plates on the surface of the earth collided like a very bad car accident in slow motion. The resulting theory of plate tectonics also explained earthquakes and volcanoes and tied together all of the other puzzling observations that the geologists had accumulated over hundreds of years.

The greatest joy in being an observational scientist is discovering something that nobody ever knew before. It takes a lot of time and hard work to gather and process the data, and then there is a great deal of struggling in the data analysis step, but the rewards can be fantastic.

For further information on the remarkable work of Dr. John Snow, please explore the UCLA Department of Epidemiology Website created by Dr. Ralph R. Frerichs at www.ph.ucla.edu/epi/snow.html.
Steven Johnston (scj777@iols.net) is an IT architect at United Airlines.