The Periodic Table of Elements, an Early Example of “Big Data”

The latest fad in computing seems to be “big data.” I got an email ad today that proclaimed, “Last Chance: Fast Data—The New Big Data.” It’s only one of the many ads that most of us get every day from conference organizers, contractors, and others implying that our company/university/laboratory—and we personally—will miss out on the latest “gold rush” unless we come to their conference or engage their services to learn: “What is ‘big data’? What can we do with it? Why we can’t afford to ignore its potential. How it will revolutionize our world. How we can make a lot of money with it.” And so on. Much of this is about social media data, credit card fraud detection, polling data, machine learning, or making money off other people’s data. But it has a great deal of promise for scientific research. Given the deluge of data that is becoming available due to ever-better sensors, the rapidly declining cost of data storage, and the exponential growth in computing power since World War II, analyzing and mining data collections to provide information to guide decisions has become important and even potentially lucrative.

However, a great deal of caution is necessary. I don’t think it’s true, as someone remarked, that “we no longer need to understand what’s going on, we only have to look it up with a search engine.” One drawback of relying on data analysis alone without understanding the underlying scientific principles is that it’s very easy to confuse correlation with causation. Type “Correlation does not imply causation” into your favorite search engine, and the Wikipedia article that comes back describes a number of examples of current medical and social science experiments in which correlations were used to support actions that turned out to be counterproductive once further experiments were conducted that proved that the correlations were spurious. However, a reasonable sense of perspective is needed. The correct analysis of valid data can lead to useful decisions, but until the underlying processes and causes are identified, those decisions can only be viewed as tentative, subject to further revision. Truly reliable predictions can only be made from an understanding of the basic principles of nature, not from raw data.

An early example is the cholera epidemic in London in 1854. The prevailing medical understanding was that disease was caused by miasma (bad air), which didn’t suggest any actions could be taken to stop the epidemic. John Snow suspected a causal relationship between poor sanitation and cholera, but he couldn’t prove it. He supported this theory by taking a map of London and drawing a point on it at the location of each known cholera case. This way of plotting the data showed that a large portion of the cholera victims were located in an area where people drew their water from the Broad Street public water pump. The pump was padlocked shut, and the epidemic was stopped, although the outbreak might have been in decline by then. This action, though effective, was highly controversial at the time. The data supported it but didn’t directly prove that polluted water was the cause of the epidemic until 1886, when the discovery of the bacterium *Vibrio cholerae* confirmed Snow’s theory. The statistical data was suggestive (and highly useful) but not conclusive. Snow was initially severely criticized at the time but is now viewed as one of the founders of epidemiology. His methods weren’t generally adopted until after the identification of the cholera bacterium.

While big data analysis is being used for fraud detection, the identification of cultural trends and potential customers, predictions of credit worthiness, and so on...
that don’t have much to do with science and engineering, data analysis (big and small) plays a key role in the scientific method. For the rest of this essay, I’ll concentrate on the analysis of scientific data. After all, this is an issue of Computing in Science & Engineering, not a discussion of social networks or fraud detection. CiSE has even devoted several special issues to the topic (such as our “Science Data Management” special issue in May/June 2013).

All of this attention and promotion tends to evoke healthy skepticism, but as I watched a PBS series entitled The Mystery of Matter, which aired on 19 August 2015, I realized that a good example of a small-scale version of big data is familiar to all of us, namely, the periodic table of elements (Figure 1). The video was about the history of Dimitri Mendeleev’s invention of the table, the chart that hangs on the wall of every chemistry lab and classroom in the world (and probably in the universe). To me, the periodic table offers several interesting lessons about data and its value and impact that are relevant to today’s big data. Science and engineering are, at their core, based on the analysis of data whose goal is the identification of the underlying causes of the phenomena being studied. It’s not sufficient to see patterns in the data and act on them. A scientist or engineer uses the patterns in the data to develop a hypothesis for the causes that lead to the observed patterns in the data. This understanding is captured in scientific theories that have predictive power, something that the history of the periodic table shows eloquently.

### Discoveries in the 19th Century

Mendeleev was a chemistry professor at St. Petersburg University who was writing a chemistry textbook for his courses. He wanted to present the 63 known elements (circa 1869) in a coherent and structured way for his students. Earlier in the 19th century,
From the chemical and physical properties of the neighboring elements, Mendeleev was able to predict the density and many other physical and chemical properties of the elements that would fill those gaps.

chemists had identified several organizing principles for the elements, including their atomic weight (a measure of their weight compared to hydrogen), their reactivity, the temperatures at which they melted and boiled, and their density. Elements with similar chemical properties could be grouped together and ordered by atomic weight. Building on earlier work, John Newlands organized 62 elements into 8 groups (circa 1863), but he included some non-elements and some inaccurate atomic weights, and he didn’t allow room for undiscovered elements.

In 1869, Mendeleev’s periodic table corrected the earlier work in several fundamental ways. His periodic table was more complete and more accurate than prior tables. Even more importantly, he put the elements in the right places in the table, organizing them as an ordered system that motivated chemists and physicists to ask fundamental questions about why the elements were related this way. And best of all for his credibility, his table had predictive power: it had gaps where there was no known element. From the chemical and physical properties of the neighboring elements, Mendeleev was able to predict the density and many other physical and chemical properties of the elements that would fill those gaps. In particular, he predicted that there should be elements that corresponded to silicon, aluminum, and boron. He called them eka-silicon (germanium), eka-aluminum (gallium), and eka-boron (scandium) and left space for them in his chart.

There were other missing elements as well, but by 1886, these three elements had been discovered and their properties measured. Gallium was discovered by Paul-Emile Lecoq de Boisbaudran in 1875. However, Lecoq de Boisbaudran’s measured density and atomic weight differed significantly from Mendeleev’s predictions. Mendeleev had the confidence to suggest that Lecoq de Boisbaudran repeat his measurements, which he did (the new measurements agreed with Mendeleev’s predictions), and gallium’s density at room temperate has remained at 5.91 g/cm$^3$ ever since. Scandium was discovered in 1879 and germanium in 1886. The later discovery of the noble gases appeared at first not to fit into the periodic table’s organizational scheme. However, they soon found a home to the right of the halogens completing each row of the table.

Moving Forward in the 20th Century and Beyond

In 1913, Henry Moseley used X-ray spectroscopy to point out that the elements were ordered by atomic number (the positive charge in the atomic core), a more tidy and suggestive scheme than atomic weight. The periodic table focused attention on the ways that atoms could combine to form molecules, leading to a much deeper understanding of the role of valences in bonding. With the discovery of the electron by J.J. Thomson in 1897 and the nucleus by Ernest Rutherford in 1911, the nuclear and electronic structure of atoms became clearer.

The invention of quantum mechanics in the early 20th century led to further advances in the understanding and application of atomic and molecular physics. Several elements on the periodic table that were still missing were identified as radioactive isotopes, leading to the development of nuclear physics and eventually nuclear power. The focus on
the properties of the nucleus led to investigations of its structure and to the investigation and discovery of the constituent particles of protons and neutrons (quarks, and so on). In the 1940s through the 1960s, Glen Seaborg and others pioneered the extension of the periodic table focusing on the creation and understanding of the transuranic elements. The final result is that the seventh row of the periodic table has recently been filled out by joint experiments in laboratories at Dubna, Livermore, and Oak Ridge.

The periodic table has played a major role in motivating and inspiring research and discovery about the materials the world is made of since Mendeleev first published it in 1869. Our understanding of materials is greatly advanced over Plato’s original concepts of earth, water, air, and fire as the fundamental elements of matter. Mendeleev’s table has clearly played an important role in that advancement.

Beyond its original purpose, we can take some points away from its history for big data as it relates to science and engineering:

■ The data collection should be as complete as possible. Mendeleev’s periodic table was more complete than prior tables.
■ The data must be correct, which might require some flexibility. Measurements can be wrong or incomplete—Mendeleev took into account the uncertainties of the atomic weight measurements and didn’t always follow the recommended atomic weights when he thought they weren’t consistent. As a result, he was able to put the data in the right places and thus see important patterns more easily.
■ It helps to group items with common properties together. Mendeleev’s grouping elements together with common chemical and physical properties facilitated the identification of key features of atomic and molecular structure.
■ Sequence the data where possible. Sequencing the elements in atomic weight turned out to be a good start; sequencing by atomic number was even better.
■ Identify gaps in the data that suggest the possibility that new important data could exist. Mendeleev predicted the properties of undiscovered elements that were in the gap. When they were discovered, the credibility of his arrangement of the elements was solidly established.
■ Be alert for emergence of new data. Modify your data arrangement to include it in your scheme. If you can’t, maybe you need a new scheme.
■ Look for patterns that suggest research questions whose answers can improve our understanding of the causes of those patterns. The periodic table was a key motivator of research that led to the understanding of chemical bonding, quantum mechanics, atomic and molecular structure, materials science, nuclear physics, and particle physics, all key elements of our understanding of the material universe in which we live.

Above all, look for patterns that suggest research questions whose answers can lead to understanding the causes of those patterns. The periodic table was a key motivator of research that led to the understanding of chemical bonding, quantum mechanics, atomic and molecular structure, materials science, nuclear physics, and particle physics, all key elements of our understanding of the material universe in which we live.

Douglass E. Post is an associate editor in chief of Computing in Science & Engineering. His day job is associate director for CREATE with the DoD High Performance Computing Modernization Program. He is on loan to the DoD as an IPA from the Carnegie Mellon University Software Engineering Institute. Contact him at post@ieee.org.