2004

History and the Fundamentals of Computer Science

Edward L. Ayers
University of Richmond, eayers@richmond.edu

HISTORY AND THE FUNDAMENTALS OF COMPUTER SCIENCE

Edward L. Ayers, University of Virginia

We might begin with a thought experiment: What is history? Many people, I've discovered, think of it as books and the things in books. That's certainly the explicit form in which we usually confront history. Others, thinking less literally, might think of history as stories about the past; that would open us to oral history, family lore, movies, novels, and the other forms in which we get most of our history.

All these images are wrong, of course, in the same way that images of atoms as little solar systems are wrong, or pictures of evolution as profiles of ever taller and more upright apes and people are wrong. They are all models, radically simplified, that allow us to think about such things in the exceedingly small amounts of time that we allot to these topics.

The same is true for history, which is easiest to envision as technological progress, say, or westward expansion, of the emergence of freedom—or of increasing alienation, exploitation of the environment, or the growth of intrusive government.

Those of us who think about specific aspects of society or nature for a living, of course, are never satisfied with the stories that suit the purposes of everyone else so well.

We are troubled by all the things that don't fit, all the anomalies, variance, and loose ends. We demand more complex measurement, description, and fewer smoothing metaphors and lowest common denominators.

Thus, to scientists, atoms appear as clouds of probability; evolution appears as a branching, labyrinthine bush in which some branches die out and others diversify. It can certainly be argued that past human experience is as complex as anything in nature and likely much more so, if by complexity we mean numbers of components, variability of possibilities, and unpredictability of outcomes.

Yet our means of conveying that complexity remain distinctly analog: the story, the metaphor, the generalization. Stories can be wonderfully complex, of course, but they are complex in specific ways: of implication, suggestion, evocation. That's what people love and what they remember.

But maybe there is a different way of thinking about the past: as information. In fact, information is *all* we have. Studying the past is like studying scientific processes for which you have the data but cannot run the experiment again, in which there is no control, and in which you can never see the actual process you are describing and analyzing. All we have is information in various forms: words in great abundance, billions of numbers, millions of images, some sounds and buildings, artifacts.

The historian's goal, it seems to me, should be to account for as much of the complexity embedded in that information as we can. That, it appears, is what scientists do, and it has served them well.

And how has science accounted for ever-increasing amounts of complexity in the information they use? Through ever more sophisticated instruments. The connection between computer science and history could be analogous to that between telescopes and stars, microscopes and cells. We could be on the cusp of a new understanding of the patterns of complexity in human behavior of the past.

The problem may be that there is too much complexity in that past, or too much static, or too much silence. In the sciences, we've learned how to filter, infer, use indirect evidence, and fill in the gaps, but we have a much more literal approach to the human past.

We have turned to computer science for tasks of more elaborate description, classification, representation. The digital archive my colleagues and I have built, the Valley of the Shadow Project, permits the manipulation of millions of discrete pieces of evidence about two communities in the era of the American Civil War. It uses sorting mechanisms, hypertextual display, animation, and the like to allow people to handle the evidence of this part of the past for themselves. This isn't cutting-edge computer science, of course, but it's darned hard and deeply disconcerting to some, for it seems to abdicate responsibility, to undermine authority, to subvert narrative, to challenge story.

Now, we're trying to take this work to the next stage, to analysis. We have composed a journal article that employs an array of technologies, especially geographic information systems and statistical analysis in the creation of the evidence. The article presents its argument, evidence, and historiographical context as a complex textual, tabular, and graphical representation. XML offers a powerful means to structure text and XSL an even more powerful means to transform it and manipulate its presentation. The text is divided into sections called "statements," each supported with "explanation." Each explanation, in turn, is supported by evidence and connected to relevant historiography.

Linkages, forward and backward, between evidence and narrative are central. The historiography can be automatically sorted by author, date, or title; the evidence can be arranged by date, topic, or type. Both evidence and historiographical entries are linked to the places in the analysis where they are invoked. The article is meant to be used online, but it can be printed in a fixed format with all the limitations and advantages of print.

So, what are the implications of thinking of the past in the hardheaded sense of admitting that all we really have of the past is information? One implication might be great humility, since all we have for most

of the past are the fossils of former human experience, words frozen in ink and images frozen in line and color. Another implication might be hubris: if we suddenly have powerful new instruments, might we be on the threshold of a revolution in our understanding of the past? We've been there before.

A connection between history and social science was tried before, during the first days of accessible computers. Historians taught themselves statistical methods and even programming languages so that they could adopt the techniques, models, and insights of sociology and political science. In the 1950s and 1960s the creators of the new political history called on historians to emulate the precision, explicitness, replicability, and inclusivity of the quantitative social sciences. For two decades that quantitative history flourished, promising to revolutionize the field. And to a considerable extent it did: it changed our ideas of social mobility, political identification, family formation, patterns of crime, economic growth, and the consequences of ethnic identity. It explicitly linked the past to the present and held out a history of obvious and immediate use.

But that quantitative social science history collapsed suddenly, the victim of its own inflated claims, limited method and machinery, and changing academic fashion. By the mid-1980s, history, along with many of the humanities and social sciences, had taken the linguistic turn. Rather than software manuals and codebooks, graduate students carried books of French philosophy and German literary interpretation. The social science of choice shifted from sociology to anthropology; texts replaced tables. A new generation defined itself in opposition to social scientific methods just as energetically as an earlier generation had seen in those methods the best means of writing a truly democratic history. The first computer revolution largely failed.

The first effort at that history fell into decline in part because historians could not abide the distance between their most deeply held beliefs and what the statistical machinery permitted, the abstraction it imposed. History has traditionally been built around contingency and particularity, but the most powerful tools of statistics are built on sampling and extrapolation, on generalization and tendency. Older forms of social history talked about vague and sometimes dubious classifications in part because that was what the older technology of tabulation permitted us to see. It has become increasingly clear across the social sciences that such flat ways of describing social life are inadequate; satisfying explanations must be dynamic, interactive, reflexive, and subtle, refusing to reify structures of social life or culture. The new technology permits a new cross-fertilization.

Ironically, social science history faded just as computers became widely available, just as new kinds of social science history became feasible. No longer is there any need for white-coated attendants at huge mainframes

and expensive proprietary software. Rather than reducing people to rows and columns, searchable databases now permit researchers to maintain the identities of individuals in those databases and to represent entire populations rather than samples. Moreover, the record can now include things social science history could only imagine before the Web: completely indexed newspapers, with the original readable on the screen; completely searchable letters and diaries by the thousands; and interactive maps with all property holders identified and linked to other records. Visualization of patterns in the data, moreover, far outstrips the possibilities of numerical calculation alone. Manipulable histograms, maps, and time lines promise a social history that is simultaneously sophisticated and accessible. We have what earlier generations of social science historians dreamed of: a fast and widely accessible network linked to cheap and powerful computers running common software with well-established standards for the handling of numbers, texts, and images. New possibilities of collaboration and cumulative research beckon. Perhaps the time is right to reclaim a worthy vision of a disciplined and explicit social scientific history that we abandoned too soon.

What does this have to do with computer science? Everything, it seems to me. If you want hard problems, historians have them. And what's the hardest problem of all right now? The capture of the very information that *is* history. Can computer science imagine ways to capture historical information more efficiently? Can it offer ways to work with the spotty, broken, dirty, contradictory, nonstandardized information we work with?

The second hard problem is the integration of this disparate evidence in time and space, offering new precision, clarity, and verifiability, as well as opening new questions and new ways of answering them.

If we can think of these ways, then we face virtually limitless possibilities. Is there a more fundamental challenge or opportunity for computer science than helping us to figure out human society over human time?