Transforming the Academy: Knowledge Formation in the Age of Digital Information

Robert L. Constable[1]

Computer-mediated knowledge formation will profoundly change every academic discipline and pose fundamental challenges to the mission of the modern research university—in teaching the new knowledge, securing sound methods for creating it, directing it to our deepest intellectual concerns, and insuring that we become wiser for it.

Digital information, now measured in petabytes, is expanding rapidly; already most of it will never be examined by any human. Computers show us where to look and help us see patterns and extract meaning. How will this way of knowing impact the research university as the Age of Digital Information unfolds?

To grasp the magnitude of the changes we face, it is important to realize that knowledge created with computer assistance goes well beyond classical knowledge formation—arising from computer processing of digital information resources on a scale that could not be achieved by all peoples of the earth acting in concert using all their cognitive powers. Computers change the scale at which resources can be examined, and they already provide sufficient discriminatory powers that scale and speed compensate for their currently limited intelligence as they draw conclusions, make predictions, and participate in discoveries. The academy is not generally aware of the potential of this transformation, although some computer scientists and computational scientist are.

The challenge for society is to assimilate digital knowledge and to improve the human condition by its application. We also seek to understand how it will shape our sense of self, individually and collectively as a society and a culture. In all of these tasks, the universities play an indispensable role for which they must prepare.

1. CREATING KNOWLEDGE

A new way of knowing. Our knowledge already depends on computers and trustworthy digital information resources. For example, computers helped us process 80 trillion bytes of data to *assemble the human genome*. This was a landmark achievement for biology and computer science. Computers led us to discover that the gene regulating the size of tomatoes is similar to genes involved in cancer (oncogenes) in mammals [Frary, et al 2000]. Another surprise from these computations is that there may be only 30 thousand human genes rather than the 100 thousand previously assumed. The database of genomes will be of immense practical value in understanding life, designing new drugs, and improving health care; computers will vastly accelerate the pace of life-saving discoveries.

In the realm of pure mathematics, the data can be *infinite*, and only proof can reduce it to the finite. Among the proofs that accomplish this reduction are those so complex that they can only be completed by computers. For example, we know that the four color conjecture from 1852 is true, but only because in 1976 computers helped Appel and Haken prove it, thereby fundamentally changing the 2,500 year

[1]Dean of the Faculty of Computing and Information Science, Cornell University

old means of knowledge formation in mathematics. In 2004 a definitive proof was finally checked by a theorem proving system called Coq. No human being unaided by computer could ever do this work.

We know more about the stars and galaxies as a result of the *Sloan Sky Survey* which makes every PC on the Internet into a telescope—for viewing digital stars. And digital stars can be more than points of light—they can be *points of information*, including data on composition, age, motion, and whatever else is interesting about them. The amount of data collected from the world's telescopes in digital format is so vast that it cannot be surveyed by humans alone. Discoveries are made because machines *mine the data* and show us where to look for the unexpected. For example, Cornell astronomer Jim Cordes recently discovered five new pulsars in the 12 terabytes of data gathered from Arecibo.

Knowledge about the climate and global warming is a result of consensus from several computer models of *earth systems*. These models are not perfect, but they capture the best of what we know, and they capture it in digital computer models that are *validated by experts*. Input to the models now comes from millions of automated sensors rather than from traditional manual collections. The combination of massive data and super fast computation has revealed unexpected connections—for example, that the surface temperature of the Indian Ocean is a major factor influencing long term weather patterns over the North Atlantic. *Earth Simulator*, in 2003 the world's fastest supercomputer at 40 teraflops was designed by NEC to run simulations of earth and atmospheric systems. It is a major force to create a fine scale digital model of the planet's surface that will improve the data used to make weather predictions.

What do these examples tell us about the nature of knowledge formation? Will the educated person know how to evaluate the results of the Earth Simulator that predict rising sea levels as a consequence of global warming? That chain of reasoning requires an understanding of the reliability and value of large amounts of computer code? Many people have cause to distrust computer code in other circumstances.

Automating intellectual processes. Knowledge arises from our accelerating ability to automate key intellectual processes using ever more sophisticated software systems powered by increasingly faster hardware. Computers first automated numerical calculation—they are by now about *four hundred billion* times faster at it than we are and getting faster year by year.

About twenty years after computers "mastered" numerical calculation, they could also do symbolic calculation, like expanding $(x+y)^2$ to $x^2+2xy+y^2$. Now *Mathematica* is widely used for symbolic computation in science and mathematics, providing every user personal access to highly advanced mathematical expertise. These advances depend critically on digital ways of representing mathematical formulas, and these representations spread to natural language as well, creating the field of *natural language processing* [Lee 2004] and *information retrieval (search engines)*.

It is widely known that computer power, measured by the density of components on a chip, is doubling every 1.5 years, doubling the processing power as well. This is one of the famous "exponentials" of computing technology, in this case called *Moore's Law*. So we can expect more remarkable examples of automating important intellectual processes, such as translating from one natural language into another or

automatically summarizing articles. Under what circumstances will such tools be accepted as part of respectable scholarship and taught in the college curriculum?

Seeing into the heart of a subject. Inquiry in all domains of knowledge can be automated to some extent. The key to having a large impact is to understand simultaneously the methods of inquiry and knowledge formation in a domain and to know enough computing and information science to recognize the mental processes that might be effectively automated and the aspects of the data that must be captured digitally. Success usually requires deep computing expertise, and the work inevitably leads to new discoveries in computing and information science as well—which is one reason that field has become so broad and deep.

2. DIGITAL COLLECTIONS

At the scale of the Web *circa* 2004, we see a data repository of *eight billion pages* indexed by Google. The amount of information seems staggering, yet it is estimated that there is five hundred times as much information in the *deep Web*, the databases and other resources not indexed by Google. Indeed Google only indexes information made available from 1995 onwards. By now the Web contains more information than the 20 terabytes of printed material estimated to be in the Library of Congress (LC) but not as much as the approximately 3 petabytes (or 3,000 terabytes) of other data in the LC such as images, sounds, and maps [Lesk 1997].

The amount of memory for a given price is another exponential; it is doubling every six months—creating a new reality of *data-intensive computing* on personal computers.

Digitizing all written human knowledge. It has been proposed as a grand challenge that we attempt to digitize all articulated human knowledge. It is not easy to make this notion precise enough to be a concrete grand challenge, because we don't simply mean "scanning" all written material. The digitization process is subtle and complex and depends on understanding the knowledge domain. Nevertheless, in many cases, scanning is a key step, and it motivates research libraries world wide. Google is actively supporting this scanning effort, and the company seems to imagine a world in which its search engines have access to all articulated human knowledge. Michael Lesk [Lesk 1997] has estimated that there are perhaps *twelve thousand petabytes* worth of information in the world.

There will also be data repositories that contain digital versions of all the primary data upon which depend the historical analyses of periods, such as the Roman republic and empire. We will have all of the economic data from many periods as well as digital versions of the literary record. History will become a massively data-intensive discipline whose use of computer resources such as search engines and narrative-generators may make uses of computers in physics seem small scale.

Life-long learning tools. We can imagine that soon students will create data repositories for storing all of the course materials they used in college annotated with comments they wrote to help understand it and linked to all the other supporting digital resources. This will be a lifelong database of digital knowledge that each student keeps current throughout his or her life.

Validating information. Creating knowledge from information requires *trusted mechanisms* to record and secure basic claims about data. For example, when the Cornell University Library posts on the Web digital replicas of Ezra Cornell's correspondence, Cornell is attesting to its authenticity and is securing the collection against errors and fraud. Another example is the way the mathematical community acted to validate the computer contribution to solving the four color problem. The computer code was made available for public scrutiny. The computations were redone with alternative software and hardware, and the exact circumstances of the computation were recorded.

Only a fraction of the data on the Web is suitable as the *basis for knowledge.* Carl Sagan might have phrased it like this: "there are billions and billions of bits of nonsense on the Web." People and institutions in the knowledge business provide the "axiomatic" basis for creating knowledge from information on the Web by providing *trusted information records.*

Impact on all disciplines. These basic facts tell us that computer-mediated knowledge formation will be important in *every discipline.* So far the humanities have been relatively untouched compared to the physical sciences and engineering, the life sciences and medicine, and the social sciences and law. However, we see signs of emerging computational methods in the humanities. Historians, for example, have clear needs to authenticate their primary sources, and computational techniques will be developed to achieve very high standards for the authenticity of digital information. We see an illustrative example in Marc Levoy's project on the *Forma Urbis Romae* [Levoy 2000], where computers have helped create new primary data from shards of the great stone map of Rome circa 210 AD. The key to doing this was to represents the shards so that they could be treated as geometric puzzle pieces that computers could attempt to assemble. The computers then were able to find enough new possible assemblies of pieces to keep historians active for years.

There are experimental systems [Mateas, et al 2000 ; Mateas and Sengers 2002] that work on small databases of historical facts to automatically create narrative accounts from different viewpoints, say economic theory, feminist theory, cultural theory, etc.

3. DIGITAL DISSEMINATION

Communicating discoveries. The Web is a new way of publishing research. The Cornell arXiv of Paul Ginsparg is an excellent example of a new kind of scientific communication. In some areas of physics, the arXiv is the point of first dissemination; it has dramatically increased the speed of communication and lowered its cost. These reductions have opened these areas of physics to people in countries that were previously cut off by the time lag and costs of traditional journal and conference publication.

The arXiv also illustrates the mechanisms and issues that arise in providing trustworthy digital information as a basis for knowledge. Given that there are no referees for arXiv material, can anyone submit anything? Will the network of links to papers provide a sufficient basis for evaluating them? How can dissemination be more open without compromising the need to sift, winnow, and authenticate?

Interactivity and discovery. Communication networks support the interactive experience that is characteristic of the Web. As bandwidth increases, the interactivity does as well; and bandwidth is another exponentially improving computing technology, doubling at essentially the same rate as cpu speeds and memory capacity for local area networks. [2]

High bandwidth networks make possible a new kind of rapid collaborative discovery. We see that mistakes in the news are detected and traced down by "bloggers" extremely rapidly. Passengers on hijacked United Airlines Flight 93 used their cell phones to learn about the nature of the hijacking from relatives on the ground listening to news only minutes old.

4. CONCLUSION

Only the beginning. Progress in coupling the exponentially expanding computing and information technologies with the central intellectual concerns of scholars and scientists reveals that we have only seen the beginning of the Age of Digital Information. The synergy among computation, information, communication and content will become deeper and more critical to knowledge creation in all sectors of society. In the knowledge sector, we see that *digital content* is more central than digital technology.

Role of universities. Universities are fundamentally concerned with creating knowledge, preserving it, and disseminating it. These activities are precisely the ones being transformed in the Age of Digital Information, as our section headings reveal.

Universities made science into a profession; they created computer science, computer engineering, and recently information science, and they educated those responsible for initiating the information revolution. Over the next fifty years, they will educate those who will lead the revolution as it takes off past its inflection point as well as those who will deepen and spread the methods of digital discovery to all academic disciplines. Based on experience with disciplines that have embraced computer-mediated knowledge formation, we can expect that between five and twenty five percent of the faculty in many academic departments will be well educated in computing and information science and part of the core faculty in that discipline. We can expect that university libraries will be required to perform a host of new services in being the authoritative host for information repositories and providing the content-based information technology that students and faculty need, working closely with the units that provide general information technology and those that provide academic research computing.

The size a quality of those faculty investments will determine which universities play a leading role in this new age; reaching such a state will transform most universities, and it will substantially increase their capacity to help society cope with the revolutionary life changing forces. In the life sciences we see investments in access of a third of the tenure track faculty.

Teachers face significant challenges because information is so readily available

[2]For wide area networks, Paul Francis in computer science notes " *wide area network performance lags because the speed of light insists on following Einstein's laws instead of Moore's.*"

from the Web. No longer can their role be simply to provide information. Instead they must teach evaluation of digital information resources and the tools for processing them in a trusted manner. They must teach about how we form knowledge based on these resources and how we strive to be wiser for them. Teachers, scholars, and scientists must take into account networked knowledge and its justification in terms of records and computation. They must stress how to make arguments and draw conclusions from it.

Teachers must motivate and inspire students to a life of continuous learning in the face of constant change and introduce them to collaborative contributions in a kind of networked knowledge formation. Information technology is a constant challenge for teachers in the Age of Digital Information; they must understand the student culture growing up around this technology and be aware of the new learning tools made possible by it. For example, there is evidence suggesting that in subjects such as computer system design, the tools used by industry to accelerate the design process also accelerate the learning process. It is likely that tools used to evaluate designs will be useful in evaluating educational materials.

Social impact. All societies face a serious problem of educating their citizens about basic scientific knowledge. Understanding the significance of digital knowledge formation will be more difficult—there are the added risks that mistakes in the creation of knowledge caused by errors in software or in the processes used to secure digital records will cause a cascading loss of confidence in expert knowledge. However, the new tools being developed for *digital discovery and scholarship* and *collaboration networks* are also powerful tools for life long learning, and they hold great promise for education and citizen participation in creating digital content of lasting general value.

REFERENCES

S. Domike, M. Mateas, and P. Vanouse 2002. The recombinant history apparatus presents: Terminal Time, in M. Mateas and P. Sengers Eds., *Narrative Intelligence*, 155–173, John Benjamins: Amsterdam.

Anne Frary, Clint Nesbitt, Amy Frary, Silvana Grandillo, Esther van der Knaap, Bin Cong, Jiping Liu, Jaroslaw Meller, Ron Elber, Kevin B. Alpert, and Steven D. Tanksley 2000. *TCloning, Transgenic Expression and Function of fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit*, Science, vol. 289, 85–88.

Juris Hartmanis and H. Lin Eds., 1992. *Computing the Future, A Broader Agenda for Computer Science and Engineering*, National Research Council: National Academy Press.

Lillian Lee 2004. *I'm Sorry Dave, I'm Afraid I Can't Do That*, Computer Science, National Research Council, 111-118, 2004.

Michael Lesk 1997. *How Much Information Is There In the World?*, http://www.lesk.com/mlesk/ksg97/ksg.html.

Marc Levoy 2000. *Digitizing the Forma Urbis Romae*, Campfire on Computers and Archeology, April 2000.

M. Mateas, P. Vancouse, and S. Domike 2000. *Generation of Ideologically-Biased Historical Documentaries*, Proceedings of AAAI 2000, 36–42, Austin, TX.

M. Mateas and P. Sengers Eds., 2002. Narrative Intelligence, In *Narrative Intelligence*, 1–25, John Benjamins: Amsterdam.

William McCune 1997. *Well-behaved search and the Robbins problem*, 8th International Conference on Rewriting Techniques and Applications (RTA), Lecture Notes in Computer Science, vol. 1232(1–7).

Chetan Murthy 1991. *An Evaluation Semantics for Classical Proofs*, Proceedings of Sixth Symposium on Logic in Computer Science, 96–109, IEEE, Amsterdam, The Netherlands.

Thomas L. Saaty and Paul C. Kainen 1986. *The Four-Color Problem*, Dover Publications: NY.