



# Global Information Industry Center

at the School of International Relations and Pacific Studies

University of California, San Diego

## Materials Science and Engineering at MIT The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, John Norris Maguire Professor of Information Technology, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering

MacKenzie Smith, Associate Director of Technology, MIT Libraries

Kate Clopeck, Masters of Science, Technology and Policy Program, MIT

June 2009

### **Abstract:**

This case study gives examples of how data is created and stored by material scientists and engineers at MIT. The amount of data depends on specific research goals and the tools, experimental techniques, and computational methods employed by the individual researcher. Both simulation and experiments are used, with the simulations producing more data in the cases reported here. The ratio between computation and data production varies widely. For example, a hundred million-atom simulation might produce only a few kilobytes of data. However, if the researcher wants to track the system at every time step, a much “smaller” simulation (fewer atoms) could generate petabytes of data. Data is retained very differently in different labs. For example, in one lab, research data is stored on the students’ and postdocs’ personal computers, with each person in charge of the data they generate. The first author listed on the final publication is responsible for backing up the data onto a CD or portable hard drive at the time of publication. Each year, the faculty member assigns one of her students to purge old data. Other papers examine other labs at MIT.

- Executive Briefing:** a summary of one or more research projects with preliminary findings for a non-academic audience.
- Research Report:** a completed report drawing on one or more research projects that presents study data, findings and management implications.
- Case Study:** an in-depth description of a firm’s approach to an information management issue.
- Research Article:** an academic research paper with sections on hypotheses tested, methods and data, analysis, findings and references.

**Contents**

<b>Background .....</b>	<b>3</b>
<b>Data Generation .....</b>	<b>3</b>
<b>Metadata .....</b>	<b>5</b>
<b>Data Retention.....</b>	<b>5</b>
<b>Data Sharing and Reuse .....</b>	<b>6</b>
<b>Key Trends and Indicators for Data Growth .....</b>	<b>6</b>
<b>About the HMI? Program .....</b>	<b>7</b>
<b>Acknowledgements .....</b>	<b>7</b>

## Materials Science and Engineering at MIT

### The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, MacKenzie Smith, and Kate Clopeck, MIT

June 2009

#### Background

There are 41 faculty members in MIT’s Department of Materials Science and Engineering, covering a wide range of expertise that includes both theoretical and applied research, with interests spanning the entire materials cycle from mining and refining of raw materials, to production and utilization of finished materials, and finally to disposal and recycling. For example, one scientist studies the coupling phenomenon that occurs at materials interfaces. By exploring coupling at the fundamental force and length scales of atoms and molecules, she looks for commonalities among materials ranging from metallic crystals to living biological cells that her research group can exploit for human advantage in sensing, actuating and transduction applications. Another scientist studies how materials change at the atomic level by applying external stimuli like plastic deformation, bombardment by energetic ions, or exposure to rapidly varying temperatures. By understanding how materials respond to these stimuli at an atomic level, this scientist hopes to create strategies for designing materials with desired properties from the atomic scale up.

For this case study, two researchers from the Department of Materials Science and Engineering were interviewed during the spring semester of 2009.

#### Data Generation

Like most other scientific or engineering fields of study, the amount of data generated by materials scientists and engineers depends on the specific

research goals and the experimental or computational techniques employed by the individual researcher.

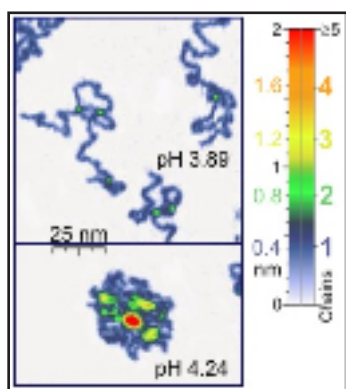
The scientist studying materials interfaces generates two kinds of data: experimental data and data from simulations. The experimental data is generated by different instruments run by open source code developed by researchers in the lab. One example of the type of instrument used is an atomic force microscope, which provides pictures of atoms on or in surfaces by scanning a fine ceramic or semiconductor tip over that surface. This tip is set at the end of a cantilever beam that will deflect as the tip is either repelled or attracted to the surface, and the magnitude of that deflection is captured by a laser and plotted, providing the scientists with the resolutions of the surface topography<sup>1</sup>. Other examples of instruments include indentures (machines that pull materials) and optical microscopes.

The resulting file generated by one of these instruments is about 10 megabytes and includes all of the raw experimental data, as well as the details describing the operating parameters. This raw data are typically images (see Figure 1). This scientist runs about two of these experiments per day, resulting in approximately 7 gigabytes of raw data each year. The lab then analyzes the raw experimental data, which generates another 10 megabytes per experiment. As a result, the lab generates a total of approximately 14 gigabytes of experimental data per year.

The simulations run in this lab generate the bulk of the data. These simulations are based on calculations made from full electronic models of the materials being studied. The closer the simulation is to the scale of the electronic model, the more storage it requires. For a typical simulation, lab members would investigate approximately 100,000 atoms and generate about 5 gigabytes of raw data that describes the structural and functional states of the atoms during the simulation. A whole study would require about 30 simulations and generate 150

<sup>1</sup> <http://www.che.utoledo.edu/nadarajah/webpages/what-safm.html>

gigabytes of data. This data generation phase lasts about three months and is followed by six months of data analysis. Unlike the experimental data analysis which doubles the amount of data generated, the simulation data analysis only generates an extra gigabyte of data per study. At any given time, the group is conducting about three different simulation studies resulting in approximately 450 gigabytes of simulation data generated each year.



**Figure 1:** Single molecules of poly(2-vinylpyridine) recorded using an AFM operating in tapping mode under water media of different pH<sup>2</sup>.

In total, this research lab is currently generating approximately 460 gigabytes of experimental and simulation data each year. This is an order of magnitude more data than was generated five years ago due to an increase in research volume (i.e. personnel, funding, and improvements in computing power, speed and storage). The scientist has only been at the Institute for five years and this increase is typical for new faculty. This order of magnitude increase in data generation has mostly been in the simulation data; the amount of experimental data has only doubled in the past five years. This is because the instruments used to generate the experimental data are very expensive and are only replaced every 8-10 years. In the future, the scientist predicts that the biggest change will be the need for data storage from the simulations because the output from these

2 Roiter and Minko, S. AFM Single molecule experiments at the solid-liquid interface: in situ conformation of adsorbed flexible polyelectrolyte chains. *Journal of the American Chemical Society*, vol. 127, no. 45, pp. 15688-15689 (2005).

simulations is increasing at a dramatic rate. The amount of experimental data, on the other hand, will probably only double in the next five years due to increased personnel, not data density.

Unlike the scientist described above, almost all of the data generated by the scientist studying material change is computational model data. However, he does exchange data with experimental scientists with a goal of connecting his modeling research and the experimental research in the field. He is particularly interested in iron beam analysis, nuclear reaction analysis, and image-based experiments.

As mentioned earlier, this scientist's research is focused on how materials change when subjected to external stimuli. A project begins by constructing an atomic-scale model of the material of interest. The model is then run through different simulations of external stimuli (i.e. extreme temperature changes, plastic deformation, etc.) and the outputs of those simulations are a series of "snapshots" of how the atomic system looks at different times or states. The research group then analyzes the state of this system and decides whether the initial model inputs were relevant. They then iterate this process and, over time, begin to notice the important elements in the model. The group then performs targeted simulations on those elements. Over the course of this process, the scientist compares the outputs of the simulations with experimental data.

Since this process is so iterative, the types of simulations run throughout one research project (and therefore, the amount of data generated) can greatly vary. For example, the lab may run a simulation that only lasts a split second in order to get a feel for how the simulation and model works. On the other hand, one simulation could also run for weeks or months on large-scale supercomputer at a national center. These large simulations, however, do not necessarily generate a lot of data. For example, if the researcher is only interested in learning about the pressure profile of a material over time under certain conditions, a hundred million-atom simulation will only produce a few kilobytes of data. However, if he is interested in studying the whole state of a system at

every time step, a much “smaller” simulation (fewer atoms) could generate terabytes, or even petabytes of data.

After the simulation is run, the research group will perform analysis that could increase the amount of data by a factor of ten. However, the amount of analysis data can vary based on the goals of the specific project. For example, 90% of the scientist’s PhD research data was analytical (only 10% was raw data from simulations) while his postdoc project mostly consisted of raw simulation data (25% was analytical data). In general, the smaller the simulation, the more detailed the analysis because larger simulations produce too much data to analyze in detail.

Despite the potential to generate large amount of data, this scientist’s projects have produced only modest amounts. During his PhD project he generated a total of 200 gigabytes of data and for his postdoc project he produced approximately 1 terabyte. He produced this relatively small amount of data because, for these two particular projects, he tried to avoid big simulations that would have required reserving time on national supercomputers and instead worked with the smallest possible computer systems.

### Metadata

Like the raw data, different scientists define materials science and engineering metadata differently. For example, the scientist does not believe that any of the information generated by his research projects should be considered “metadata” because everything is valuable to his research. He considers the models he develops, the conditions of the simulations, and all other parameters to be “data” not “metadata.”

The other scientist, who runs both experiments and simulations, will generate metadata from her instruments (describing the date, time, temperature and other experimental conditions), and from her simulations (explaining the version of the software, the number of atoms in the simulation, date, etc). The experimental metadata is either recorded in lab

notebooks or included in the instrument’s output file. The simulation metadata is included in the header of the output files. Neither type of metadata is very large and is insignificant in size when compared to the raw data generated. The importance of the metadata varies depending on the sophistication of the experiment or simulation. The more complex methodologies, the more important the metadata becomes.

### Data Retention

There are no common data retention policies for the Department of Materials Science and Engineering and therefore, each faculty members tackles the issue differently. For example, one scientist’s research data is stored on her students’ and postdocs’ personal computers. The data is replicated twice during the data production process (it is kept on the computer it was generated on, and on the computer that it is analyzed on). The data is only backed up after a publication has been submitted, and no automated backup system is used. Each student in her research group is in charge of the data that they generate. The first author listed on the final publication is responsible for backing up the data onto a CD or portable hard drive. Each year, this scientist assigns one of her students the task of sorting through the data on the lab’s cluster and deleting data from students who are no longer with the group (assuming that the data has been published and, therefore, backed up).

Another scientist stores all his data on his personal computer and external hard drives. One of these drives has automatic back up. His total storage capacity is 2 terabytes. This capacity will increase in the near future (before the end of the year) because he is in the process of setting up a new computational cluster, which will have 15 terabytes of storage. This scientist is in charge of all his own data management, including data retention decisions. He does not delete any of his data, and would rather buy more storage then delete old data. If he used a national supercomputer center for a simulation then all of the data generated during that simulation is also stored at

the cluster where it was generated. There are no data retention policies at these clusters, but there are strict rules about data sharing.

### Data Sharing and Reuse

There are no national data repositories for data sharing in materials science. However, scientists may share data on a lab-by-lab basis. Both of the scientists described above share their data a few times a year with colleagues at other universities or research institutions. One of the scientists will usually share models but not the data generated by running the model through simulations. The other scientist will often share the outputs of the simulations. In addition to sharing data and models, both scientists also reuse their own data by running new analyses on older models or raw data from previous projects.

### Key Trends and Indicators for Data Growth

Although the different researchers in the Department of Materials Science and Engineering study a range of different topics, there are several key trends and indicators for data growth that we have identified:

1. The amount of data currently generated in the Department of Materials Science and Engineering is approximately 32 terabytes per year (based on scientists currently in this department, and assuming larger research groups and more funding opportunities allow researchers to generate 75% more data than assistant professors, and full professors to generate twice as much data as assistant professors).
2. Increases in computing power and speed have caused the amount of data generated in this field to increase by an order-of-magnitude in the past five years. As computer get faster and storage get cheaper, data generation will increase exponentially in the next five years.
3. Although metadata is important to research projects, materials scientists and engineers do not generate significant amounts.
4. Data retention decisions are often up to the individual researchers. Retention policies vary by lab.
5. There are no commonly used national data repositories, but individual scientists are open to data sharing and may share their data (raw experimental data, simulations outputs, and computational models) with colleagues at other laboratories, universities, or research institutions.

## About the HMI? Program

The How Much Information? (HMI?) research program is a multi-discipline, multi-university project, formed to investigate the nature of data and information generated and used by individuals and enterprises. The program is sponsored by seven companies, including AT&T, Cisco, IBM, Intel, LSI, Oracle, and Seagate, and involves multiple research universities. The Principal Investigator is Prof. Roger Bohn and the Research Director is Dr. James Short, at UC San Diego's Global Information Industry Center (<http://giic.ucsd.edu>). Founded in 1960, the University of California, San Diego is one of the nation's most accomplished research universities, widely acknowledged for its local impact, national influence and global reach.

## Acknowledgements

This case study is the product of industry and university collaboration in applied research. We are grateful for the support of our industry partners, sponsor liaisons, university research partners, and administrative staff at the University of California, San Diego.

Financial support for HMI? research and the Global Information Industry Center is gratefully acknowledged. Our sponsors are:

AT&T

Cisco Systems

IBM

Intel

LSI

Oracle

Seagate Technology

Additional support was provided by the Alfred P. Sloan Foundation of New York.

Questions about this research may be addressed to the Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego:

Roger Bohn, Principal Investigator, [rbohn@ucsd.edu](mailto:rbohn@ucsd.edu)

Jim Short, Research Director, [jshort@ucsd.edu](mailto:jshort@ucsd.edu)

Pepper Lane, Program Coordinator, [pelane@ucsd.edu](mailto:pelane@ucsd.edu)