IDR Team Summary 8

Develop image-specialized database tools for data stewardship and system design in large-scale applications.

CHALLENGE SUMMARY

During the past 30 years imaging science has produced a wide array of image acquisition systems that have revolutionized our ability to acquire images. For example, the evolution from CCD (charge-coupled device) imagers to CMOS (complementary metal oxide semiconductor) imagers has made the acquisition of visible band images nearly free; still and video images of the natural environment and social groups are being acquired at an unprecedented rate. These are being used for mobile visual search applications, in which users acquire cell phone images to navigate their local environment. Medical images in both research and clinical applications, including CT, PET and MR, are being acquired at a rate that is hard to imagine. The diagnosis of such images can be greatly improved by aggregation of datasets.

The revolution in imaging applications has been led by instrumentation—the development of new sensors and data storage technologies that acquire and store many gigabytes of data. Unfortunately, there is not a corresponding effort to develop software database tools to manage this flood of data, and imaging systems are not typically designed with both the hardware and software in mind. For example, because of the nature of the instrument-led acquisition, only a modest amount of information about the imaging context (often called metadata) was planned as part of the instrument design. Moreover, there are no widely accessible tools for aggregating the images and the modest amount of metadata to expand our understanding
of natural phenomenon. The aggregation of these data can have applications in a wide range of fields including law, education, business, and medicine.

There is an opportunity—and a need—to design imaging systems from the ground up, keeping both hardware and software in mind. The systems should facilitate the validation, preservation, and analysis of massive amounts of data. For example, the next generation of MR scanners should incorporate the software design team in the first stages of system planning, and the instruments should be engineered for the Exabyte scale. This type of engineering will require the cooperation of research scientists spanning the imaging community and software communities; these individuals typically have very different skill sets and are trained in different university or corporate programs.

**Key Questions**

- What would it take to build a software infrastructure so that imaging systems developers can easily incorporate large-scale data sharing and data analysis, thereby enabling important information to be coordinated within/among a large user group?
- Are there successful models, such as databases for face recognition and finger printing, that might be used as a model for other organizations, such as MR anatomical and functional data?
- Are there common architectural and computational needs across multiple types of imaging modalities for storing, validating quality, and analyzing image databases? Are there general ontologies for imaging data that might be derived from the images themselves, rather than by labels added by the users in the metadata?

**Reading**


IDR TEAM MEMBERS

- Marna E. Ericson, University of Minnesota
- Antonio Facchetti, Polyera Corporation/Northwestern University
- Thomas J. Grabowski, Jr., University of Washington
- Brian P. Hayes, *American Scientist*
- Myrna E. Jacobson Meyers, University of Southern California
- Blake C. Jacquot, Jet Propulsion Laboratory
- Robert H. Lupton, Princeton University
- Rosalind Reid, Harvard University
- Tomasz F. Stepinski, University of Cincinnati
- Tanveer F. Syeda-Mahmood, IBM Almaden Research Center
- Emily Elert, New York University

IDR TEAM SUMMARY

*Emily Elert, NAKFI Science Writing Scholar, New York University*

**Databases, Past and Present**

Long before parallel processing, supercomputers, or Turing machines, there were Harvard Computers. These image processors were essential to the telescopic-spectrometry boom of late 19th century astronomy, when new technology was generating information-rich photographs faster than astronomers could analyze them—and before they knew just what they were looking for.

Of course, the Harvard Computers weren’t quite like the ones we have today—they were, in fact, a group of women, hired by the astronomer Edward Charles Pickering to process astronomical data. Just as today’s computers analyze images and extract meaningful information, Pickering’s team went through one glass-plate photograph at a time identifying, measuring, and recording what they saw in the stars.

And it worked! In 1908, after 15 years of this work, Henrietta Swan Leavitt published a paper called “1777 variables in the Magellanic Cloud,” which noted a relationship between variable stars’ period and luminosity. That discovery, confirmed by Leavitt a few years later, helped set the stage for Hubble’s famous red-shift and the understanding that the universe is expanding.

The development of digital imaging has allowed astronomers to acquire tremendous amounts of visual information and rendered analog image
processing infeasible. Today, human power is devoted to training computers to identify, measure, and record meaningful information. Rather than hand-written data tables, astronomers organize those extracted features in relational databases, where they can easily be retrieved and analyzed.

This method of image database creation has allowed for some extraordinary scientific investigations. One recent example is the Sloan Digital Sky Survey, in which a dedicated telescope photographed over a quarter of the night sky and catalogued more than 350 million celestial objects. The resulting dataset has yielded some profound discoveries, including the universe’s most distant quasars and large populations of sub-stellar objects.

One of the keys to the success of this modern database system is that the physical universe is largely familiar to astronomers, despite its many mysteries. The dataset from the Sloan Survey can be used to nearly perfectly reconstruct images of the sky, because astronomers were able to tell the computer just what they were looking for—they were able to define, in sharp, numerical terms what might constitute meaningful information.

But the modern database system doesn’t meet the needs of other, less established sciences. Despite some huge advances in neuroscience and neuroimaging, for example, scientists still lack a basic conceptualization of the structure and function of the brain. Without this understanding, it’s often impossible to predict and describe which information in an image of the brain will be useful. Without that ability, it is difficult or impossible to extract all of the relevant features from brain images. Modern imaging database systems can’t accommodate the needs of scientists working in fields with these kinds of limitations.

**Database Future**

The current challenge, then, is figuring out how to acquire imaging data and build databases within rapidly evolving scientific domains. That’s a big challenge, but there are a couple of straightforward first steps. In neuroscience, the first step is to standardize the data, both within and across imaging modalities.

Brain imaging technologies are evolving along with scientists’ understanding of the brain. Currently, there are no broadly accepted standards in neuroscience for imaging systems and images. Two sets of brain fMRI data from two different studies often yield images taken at different angles with different instrument settings, and then recorded in different file formats with different metadata, and organized into different relational databases.
The result is two bodies of data that have no use beyond the scope of the particular study they were gathered for. It’s also quite difficult for other scientists to reproduce their colleagues’ findings—a basic practice for the progress of any science.

Standardizing the data would solve both of these problems. Similar standards have been adopted in other fields of imaging and could serve as a model. One of these is Digital Imaging and Communications in Medicine, or DICOM, a standard developed in the 1980s to standardize file formats and metadata. DICOM allows medical images acquired at different places to be transferred and pooled in collective databases.

Another tractable—if more difficult—standardization challenge is that neuroscience imaging operates in a number of modalities. While fMRI uses changes in blood flow as a proxy for brain activity, EEG measures the electrical activity in the brain. MEG, another modality, isolates electromagnetic activity. There’s also PET. . . . Each of these modalities has its own strengths and weaknesses, and arguably the field of neuroscience would benefit if there were ways to integrate heterogeneous data across modalities. Ideal databases would be able to pool, weight, and analyze these disparate data to take advantage of the insights each modality can provide.

Creating databases to collect and analyze this data will require a deeper reimagining than the steps outlined so far. Nascent imaging sciences would benefit from databases that can learn, adapt, and change along with the science, and along with evolving imaging technologies. In short, younger sciences require smarter, more agile databases.

These next-generation databases would be tools for exploration as well as analysis. In order to make that possible, images need to become a functional part of the database, along with the numerical features that describe them. The databases need to be able to process images. They need built-in tools for browsing and searching images, and those tools need to be tailored to different scientific domains. In such a database, a user could browse images, select a visual aspect of a single image, and run a search for similar aspects in other images. This is similar to feature extraction, except that the users doesn’t have to know exactly what they are looking for—they don’t have to be able to define queries in exact, mathematical terms—in order to look.

Those exploratory tools should incorporate machine learning where possible. For example, if a user selects a visual aspect of an image and says, “show me more like this,” the computer can return a few results for relevancy feedback from the user. The user can say, “No, not like this one—find
ones like this!” This sort of relevancy feedback can help the user define his or her question, while helping the computer develop more accurate search capabilities.

Currently, the process of feature extraction is limited to database creation. In next-generation databases, feature extraction and imaging data analysis would be an ongoing process. The structure of the relational database would therefore change over time, to reflect evolving scientific understanding.

**Recommendations**

1. The neuroscience community must define standards for acquiring imaging data and demand that instrument vendors accommodate those standards. Those standards would anticipate the needs of basic science, including:
   a. sharing and searching heterogeneous imaging data;
   b. metadata standards native to instrumentation and specific to neuroscience aims; and
   c. community benchmarks, or ground truth datasets for assessing and stimulating algorithm performance.
2. Scientists must get over their data sharing issues and adopt an open-source model rather than a competitive one.
3. Although the technologies already exist for next-generation databases, the databases themselves do not. Perhaps the biggest reason for this is the lack of interdisciplinary action between people with deep knowledge in a scientific field and people with deep informatics knowledge. Because the problems with current databases have obvious solutions, they fail to interest people in informatics. And because universities reward active research over interdisciplinary expertise, few scientists within those domains have the expertise. In order to create the kind of next-generation databases described here, there must be more interaction between these two groups.
   a. Research is needed into how to pool, evaluate, weight, and use heterogeneous image data.
   b. A plug-in model for database query is desirable, i.e., native support for image processing in the database that has an open modular architecture.
   c. Agile exploratory tools that incorporate image analysis and machine learning must be imagined and implemented for imaging databases.