

Big Data

Researchers Discuss the Opportunities & Challenges

HEATHER STEWART
heather.stewart@nyu.edu

Because computers are getting faster, data-generating experiments and equipment are growing in complexity, and bandwidth requirements among computers and components are increasing. Researchers in all fields are creating more data and are working with larger sets of data.

We have entered the era of enormous datasets — or “big data” — and each of us in the academy shares a responsibility to incorporate the burgeoning possibilities into our world.

Those of us who are responsible for networking and computation join with researchers in thinking through how options might be provided for moving large amounts of data quickly, and for storing and computing this data. Those of us in libraries are creating new means of associating metadata, building searchable collections of data, and working to preserve them for the next generation of inquiries. Those involved with the analysis and display of data are working together to develop algorithms and visual representations that make possible more efficient calculation and greater comprehension of complex arrays.

Enter the Big Data Sessions

An informal group at NYU is examining the scholarly issues and possibilities associated with big data. The group met twice this past fall, facilitated by David Hogg, Associate Professor of Physics in the Center for Cosmology and Particle Physics (FAS).

Professor Hogg has brought together, in a series of sessions, more than 70 researchers from across NYU, including the Medical Center and NYU Polytechnic Institute. These sessions have resulted in joint research projects and grant applications, as well as a more fruitful understanding of

how NYU researchers can be served in their work with big data.

Professor Hogg’s research group works with up to 50 terabytes of astrophysics data, including data from the Sloan Digital Sky Survey, Spitzer Space Telescope, and Galaxy Evolution Explorer. However, his interests extend beyond his own discipline.

One of the questions he seeks to answer is what connections can be made — across fields of study, between massive-scale data generation and use — so as to foster innovation and promote discovery and collaboration: “The purpose...[of these sessions has been] to provide faculty and postdocs at New York University with a fun, intellectually stimulating, and focused opportunity to learn about the research of colleagues from different disciplines. In doing so, they are able to explore ideas and make connections between subjects that had previously seemed unrelated.”

He adds, “We are all using enormous datasets to do our science; this presents us with engineering and scientific challenges, some of which are domain-specific, but some of which cut across all disciplines. I expect we will find many points of common interest, from mundane issues like what kinds of disk controllers we use to high-level questions about data vetting, modeling, visualization, statistical inference, and publication.”

“80 Million Tiny Images”

To this end, Professor Hogg has enlisted in the process FAS colleagues Rob Fergus (Computer Science) and Kyle Cranmer (Physics), who respectively led the first and second Big Data sessions.

Professor Fergus’ research is in the field of computer vision, with links to computer graphics and machine learning. Specific areas of interest include object recognition,

Heather Stewart is the ITS Director for Academic Technology Services.

computational photography, and problems in low-level vision.

His presentation, “80 Million Tiny Images,” highlighted how, with the advent of the Internet, billions of images now freely available online might constitute a dense sampling of the visual world. He explored this world with the aid of a large dataset of 80 million images collected from the Internet, using a variety of non-parametric statistical methods (see figure, below).

Petabytes of Data

Professor Cranmer led the second Big Data session. An experimental particle physicist, he is working on the ATLAS experiment, which is part of the Large Hadron Collider (LHC) project at CERN in Geneva, Switzerland.

The LHC is the world’s largest and highest-energy particle accelerator. These experiments record petabytes¹ of data, aiding studies of fundamental particles’ masses and interactions. Professor Cranmer specializes in advanced data analysis techniques, statistics, and the interface between theory and experiment. In this session, he discussed the scientific challenges of addressing well-posed statistical questions in the context of very large datasets, very elaborate theoretical models, and a complex experimental environment.

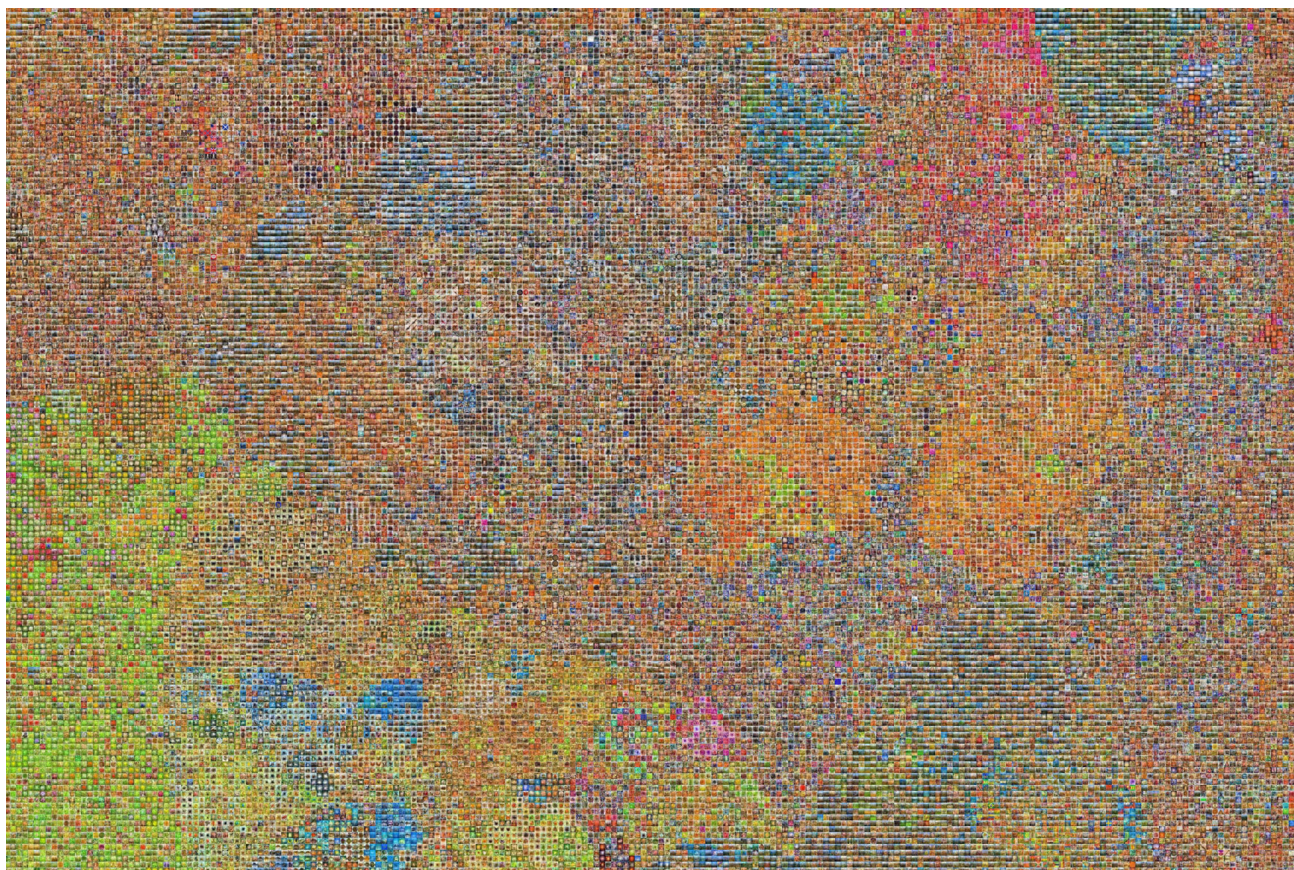
The LHC will explore fundamental questions of particle physics by colliding protons together 40 million times per second. Each of these collisions will be recorded by huge particle detectors with

¹ en.wikipedia.org/wiki/Petabyte

approximately 100 million sensors. In these enormous, complex datasets, physicists will search for evidence of new particles that may be produced only very rarely.

In addition to the obvious computing and data mining challenges, experimentalists aim to make precise statistical statements within the context of theoretical models. These models can be very elaborate, with hundreds of parameters that have physical significance.

This spring, Professor Hogg again brought together researchers from all disciplines, as the series on massive scale datasets continued. If you may be interested in participating in such sessions in the fall, please contact ITS Director of Academic Technology Services Heather Stewart at heather@nyu.edu. §



Detail from Professor Fergus’ “visual dictionary,” created as part of the “80 Million Tiny Images” project – from his website, cs.nyu.edu/~fergus.