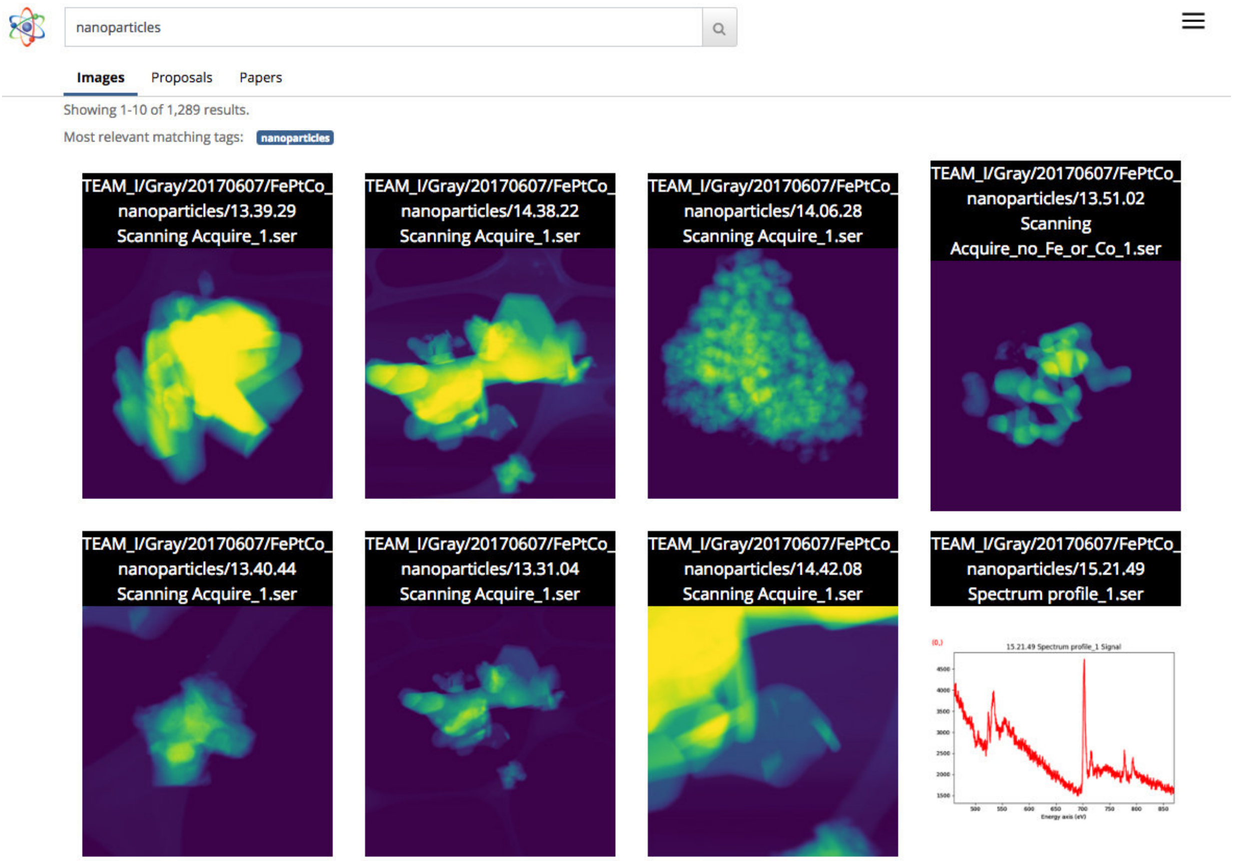
PHYS.ORG

Researchers use machine learning to search science data

June 19 2018



Showing 1-10 of 1,289 results.

Most relevant matching tags: nanoparticles

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.39.29 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.38.22 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.06.28 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.51.02 Scanning Acquire_no_Fe_or_Co_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.40.44 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.31.04 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.42.08 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/15.21.49 Spectrum profile_1.ser

Screenshot of the Science Search interface. In this case, the user did an image search of nanoparticles. Credit: Gonzalo Rodrigo, Berkeley Lab

As scientific datasets increase in both size and complexity, the ability to label, filter and search this deluge of information has become a laborious, time-consuming and sometimes impossible task, without the help of automated tools.

With this in mind, a team of researchers from Lawrence Berkeley National Laboratory (Berkeley Lab) and UC Berkeley are developing innovative machine learning tools to pull contextual information from scientific datasets and automatically generate metadata tags for each file. Scientists can then search these files via a web-based search engine for scientific data, called Science Search, that the Berkeley team is building.

As a proof-of-concept, the team is working with staff at the Department of Energy's (DOE) Molecular Foundry, located at Berkeley Lab, to demonstrate the concepts of Science Search on the images captured by the facility's instruments. A beta version of the platform has been made available to Foundry researchers.

"A tool like Science Search has the potential to revolutionize our research," says Colin Ophus, a Molecular Foundry research scientist within the National Center for Electron Microscopy (NCEM) and Science Search Collaborator. "We are a taxpayer-funded National User Facility, and we would like to make all of the data widely available, rather than the small number of images chosen for publication. However, today, most of the data that is collected here only really gets looked at by a handful of people—the data producers, including the PI (principal investigator), their postdocs or graduate students—because there is currently no easy way to sift through and share the data. By making this raw data easily searchable and shareable, via the Internet, Science Search could open this reservoir of 'dark data' to all scientists and maximize our facility's scientific impact."

The Challenges of Searching Science Data

Today, search engines are ubiquitously used to find information on the

Internet but searching science data presents a different set of challenges. For example, Google's algorithm relies on more than 200 clues to achieve an effective search. These clues can come in the form of key words on a webpage, metadata in images or audience feedback from billions of people when they click on the information they are looking for. In contrast, scientific data comes in many forms that are radically different than an average web page, requires context that is specific to the science and often also lacks the metadata to provide context that is required for effective searches.

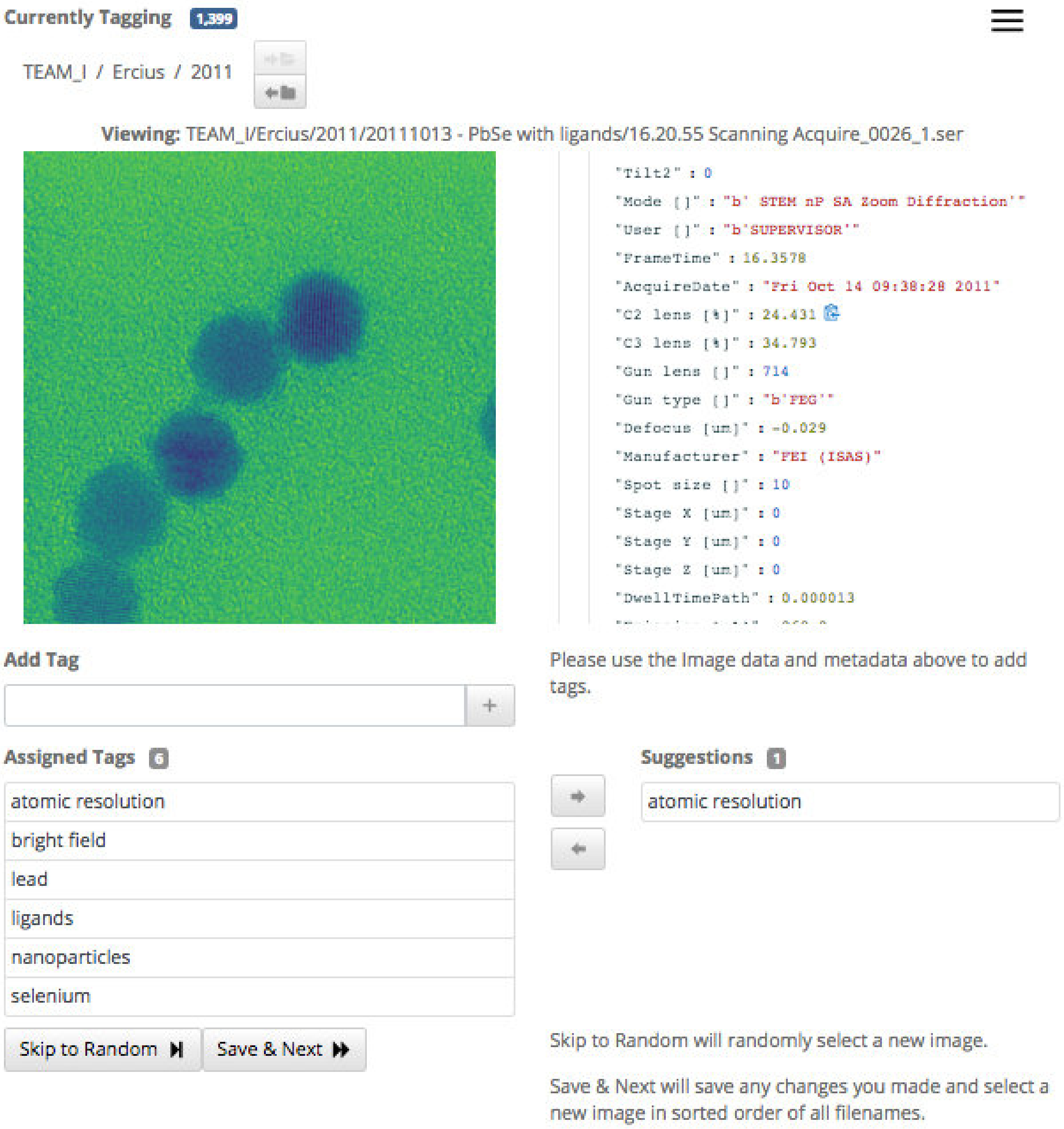
At National User Facilities like the Molecular Foundry, researchers from all over the world apply for time and then travel to Berkeley to use extremely specialized instruments free of charge. Ophus notes that the current cameras on microscopes at the Foundry can collect up to a terabyte of data in under 10 minutes. Users then need to manually sift through this data to find quality images with "good resolution" and save that information on a secure shared file system, like Dropbox, or on an external hard drive that they eventually take home with them to analyze.

Oftentimes, the researchers that come to the Molecular Foundry only have a couple of days to collect their data. Because it is very tedious and time consuming to manually add notes to terabytes of scientific data and there is no standard for doing it, most researchers just type shorthand descriptions in the filename. This might make sense to the person saving the file, but often doesn't make much sense to anyone else.

"The lack of real metadata labels eventually causes problems when the scientist tries to find the data later or attempts to share it with others," says Lavanya Ramakrishnan, a staff scientist in Berkeley Lab's Computational Research Division (CRD) and co-principal investigator of the Science Search project. "But with machine-learning techniques, we can have computers help with what is laborious for the users, including adding tags to the data. Then we can use those tags to effectively search the data."

To address the metadata issue, the Berkeley Lab team uses machine-learning techniques to mine the "science ecosystem"—including instrument timestamps, facility user logs, scientific proposals, publications and file system structures—for contextual information. The collective information from these sources including timestamp of the experiment, notes about the resolution and filter used and the user's request for time, all provides critical contextual information. The Berkeley lab team has put together an innovative software stack that uses machine-learning techniques including natural language processing pull contextual keywords about the scientific experiment and automatically create metadata tags for the data.

For the proof-of-concept, Ophus shared data from the Molecular Foundry's TEAM 1 electron microscope at NCEM that was recently collected by the facility staff, with the Science Search Team. He also volunteered to label a few thousand images to give the machine-learning tools some labels from which to start learning. While this is a good start, Science Search co-principal investigator Gunther Weber notes that most successful machine-learning applications typically require significantly more data and feedback to deliver better results. For example, in the case of search engines like Google, Weber notes that training datasets are created and machine-learning techniques are validated when billions of people around the world verify their identity by clicking on all the images with street signs or storefronts after typing in their passwords, or on Facebook when they're tagging their friends in an image.



Viewing: TEAM_I/Ercius/2011/20111013 - PbSe with ligands/16.20.55 Scanning Acquire_0026_1.ser

```
"Tilt2" : 0
"Mode []" : "b' STEM nP SA Zoom Diffraction'"
"User []" : "b'SUPERVISOR'"
"FrameTime" : 16.3578
"AcquireDate" : "Fri Oct 14 09:38:28 2011"
"C2 lens [%]" : 24.431
"C3 lens [%]" : 34.793
"Gun lens []" : 714
"Gun type []" : "b'FEG'"
"Defocus [um]" : -0.029
"Manufacturer" : "FEI (ISAS)"
"Spot size []" : 10
"Stage X [um]" : 0
"Stage Y [um]" : 0
"Stage Z [um]" : 0
"DwellTimePath" : 0.000013
```

Add Tag

Please use the Image data and metadata above to add tags.

Assigned Tags 6

- atomic resolution
- bright field
- lead
- ligands
- nanoparticles
- selenium

Suggestions 1

- atomic resolution

Skip to Random will randomly select a new image.

Save & Next will save any changes you made and select a new image in sorted order of all filenames.

This screen capture of the Science Search interface shows how users can easily validate metadata tags that have been generated via machine learning, or add information that hasn't already been captured. Credit: Gonzalo Rodrigo, Berkeley Lab

"In the case of science data only a handful of domain experts can create training sets and validate machine-learning techniques, so one of the big ongoing problems we face is an extremely small number of training sets," says Weber, who is also a staff scientist in Berkeley Lab's CRD.

To overcome this challenge, the Berkeley Lab researchers used transfer learning to limit the degrees of freedom, or parameter counts, on their convolutional neural networks (CNNs). Transfer learning is a machine learning method in which a model developed for a task is reused as the starting point for a model on a second task, which allows the user to get more accurate results from a smaller training set. In the case of the TEAM I microscope, the data produced contains information about which operation mode the instrument was in at the time of collection. With that information, Weber was able to train the neural network on that classification so it could generate that mode of operation label automatically. He then froze that convolutional layer of the network, which meant he'd only have to retrain the densely connected layers. This approach effectively reduces the number of parameters on the CNN, allowing the team to get some meaningful results from their limited training data.

Machine Learning to Mine the Scientific Ecosystem

In addition to generating metadata tags through training datasets, the Berkeley Lab team also developed tools that use machine-learning techniques for mining the science ecosystem for data context. For example, the data ingest module can look at a multitude of information sources from the scientific ecosystem—including instrument timestamps, user logs, proposals and publications—and identify commonalities. Tools developed at Berkeley Lab that use natural language-processing methods can then identify and rank words that give context to the data and facilitate meaningful results for users later on. The user will see something similar to the results page of an Internet search, where content with the most text matching the user's search words will appear higher on the page. The system also learns from user queries and the search results they click on.

Because scientific instruments are generating an ever-growing body of data, all aspects of the Berkeley team's science search engine needed to be scalable to keep pace with the rate and scale of the data volumes being produced. The team achieved this by setting up their system in a Spin instance on the Cori supercomputer at the National Energy Research Scientific Computing Center (NERSC). Spin is a Docker-based edge-services technology developed at NERSC that can access the facility's high performance computing systems and storage on the back end.

"One of the reasons it is possible for us to build a tool like Science Search is our access to resources at NERSC," says Gonzalo Rodrigo, a Berkeley Lab postdoctoral researcher who is working on the natural language processing and infrastructure challenges in Science Search. "We have to store, analyze and retrieve really large datasets, and it is useful to have access to a supercomputing facility to do the heavy lifting for these tasks. NERSC's Spin is a great platform to run our search engine that is a user-facing application that requires access to large datasets and analytical data that can only be stored on large supercomputing storage systems."

An Interface for Validating and Searching Data

When the Berkeley Lab team developed the interface for users to interact with their system, they knew that it would have to accomplish a couple of objectives, including effective search and allowing human input to the machine learning models. Because the system relies on domain experts to help generate the training data and validate the machine-learning model output, the interface needed to facilitate that.

"The tagging interface that we developed displays the original data and metadata available, as well as any machine-generated tags we have so far. Expert users then can browse the data and create new tags and review any machine-generated tags for accuracy," says Matt Henderson, who is a Computer Systems Engineer in CRD and leads the user interface development effort.

To facilitate an effective search for users based on available information, the team's search interface provides a query mechanism for available files, proposals and papers that the Berkeley-developed machine-learning tools have parsed and extracted tags from. Each listed search result item represents a summary of that data, with a more detailed secondary view available, including information on tags that matched this item. The team is currently exploring how to best incorporate user feedback to improve the models and tags.

"Having the ability to explore datasets is important for scientific breakthroughs, and this is the first time that anything like Science Search has been attempted," says Ramakrishnan. "Our ultimate vision is to build the foundation that will eventually support a 'Google' for scientific data, where researchers can even search distributed datasets. Our current work provides the foundation needed to get to that ambitious vision."

"Berkeley Lab is really an ideal place to build a tool like Science Search because we have a number of user facilities, like the Molecular Foundry, that have decades worth of data that would provide even more value to the scientific community if the data could be searched and shared," adds Katie Antypas, who is the principal investigator of Science Search and head of NERSC's Data Department. "Plus we have great access to machine-learning expertise in the Berkeley Lab Computing Sciences Area as well as HPC resources at NERSC in order to build these capabilities."

More information: Download the Beta version of Science Search sciencesearch-ncem.lbl.gov/

Provided by Lawrence Berkeley National Laboratory

Citation: Researchers use machine learning to search science data (2018, June 19) retrieved 12 May 2024 from https://phys.org/news/2018-06-machine-science.html