HYPERION RESEARCH

HPC User Forum Update

Automated Data Analysis Workflows -- Why Data Management is Important in the Era of HPDA

Bob Sorensen
August 2019

IN THIS UPDATE

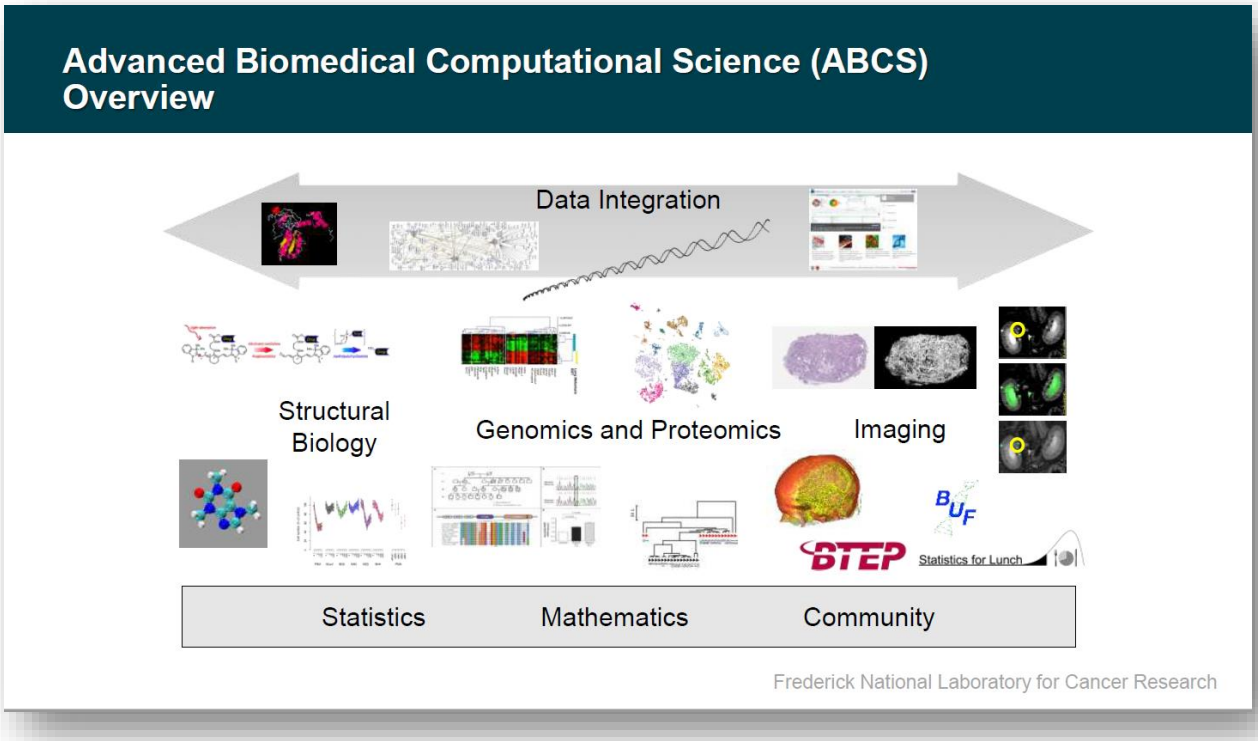
The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In April 2019, the 72nd HPC User Forum took place in Santa Fe, New Mexico. This update summarizes a presentation from that meeting entitled*, Automated Data Analysis Workflows - Why Data Management is Important in the Era of HPDA* by Jack Collins, Director, Advanced Biomedical Computational Science Group, Frederick National Laboratory.

Source: Frederick National Laboratory and Hyperion Research, 2019

PRESENTATION: AUTOMATED DATA ANALYSIS WORKFLOWS -- WHY DATA MANAGEMENT IS IMPORTANT IN THE ERA OF HPDA, BY JACK COLLINS, DIRECTOR, ADVANCED BIOMEDICAL COMPUTATIONAL SCIENCE GROUP, FREDERICK NATIONAL LABORATORY

Jack Collins, Director, Advanced Biomedical Computational Science Group, Frederick National Laboratory started his talk with an overview of the current state of advanced biomedical computation science, which seeks to integrate structural biology, genomics and proteomics imaging data and analysis. A critical element in this scheme is to train users to integrate the data into a dashboard so they can they work with clinicians, researchers, and people who do bioinformatics as well. Collins noted that his group has been focusing on automating workflows in cancer research, highlighting that that compute is cheap, and people are expensive.

FIGURE 1

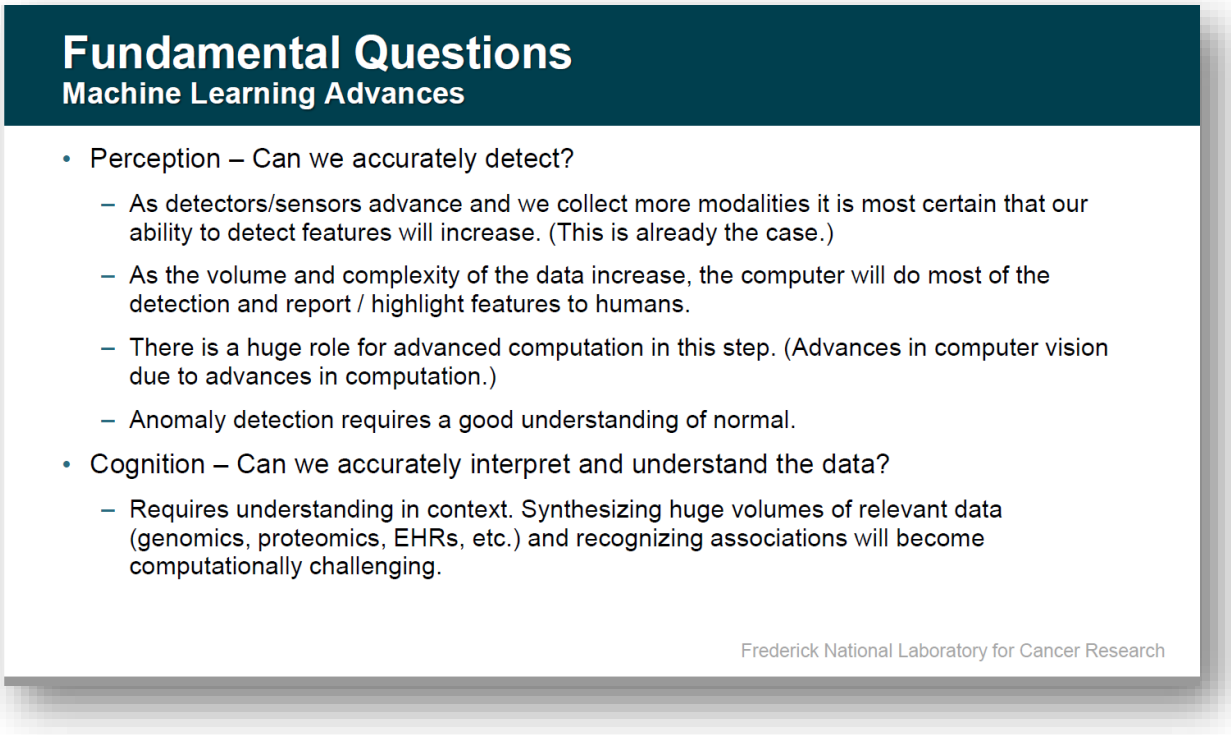


Source: Frederick National Laboratory and Hyperion Research, 2019

Collins addressed the impact that machine learning could have medical research, noting that when his group considers machine learning, key aspects they look at include perception -- what things can they accurately detect, and cognition -- can the data truly be trusted?

Collins warned that machine learning is all about data and speeding up automated analysis, but the critical step is asking the right question. If the right question cannot be asked, machine learning might not be the right thing to do. As such, the emphasis needs to be collecting data that's actually useful to the question being asked. That often can take years and a lot of money, but it's something that must be done and watched. After that, data must be harmonized and analyzed. That's where meta data comes in as important: it must be easily found, it must be shareable, and it must be subject to aggregation.

FIGURE 2



Source: Frederick National Laboratory and Hyperion Research, 2019

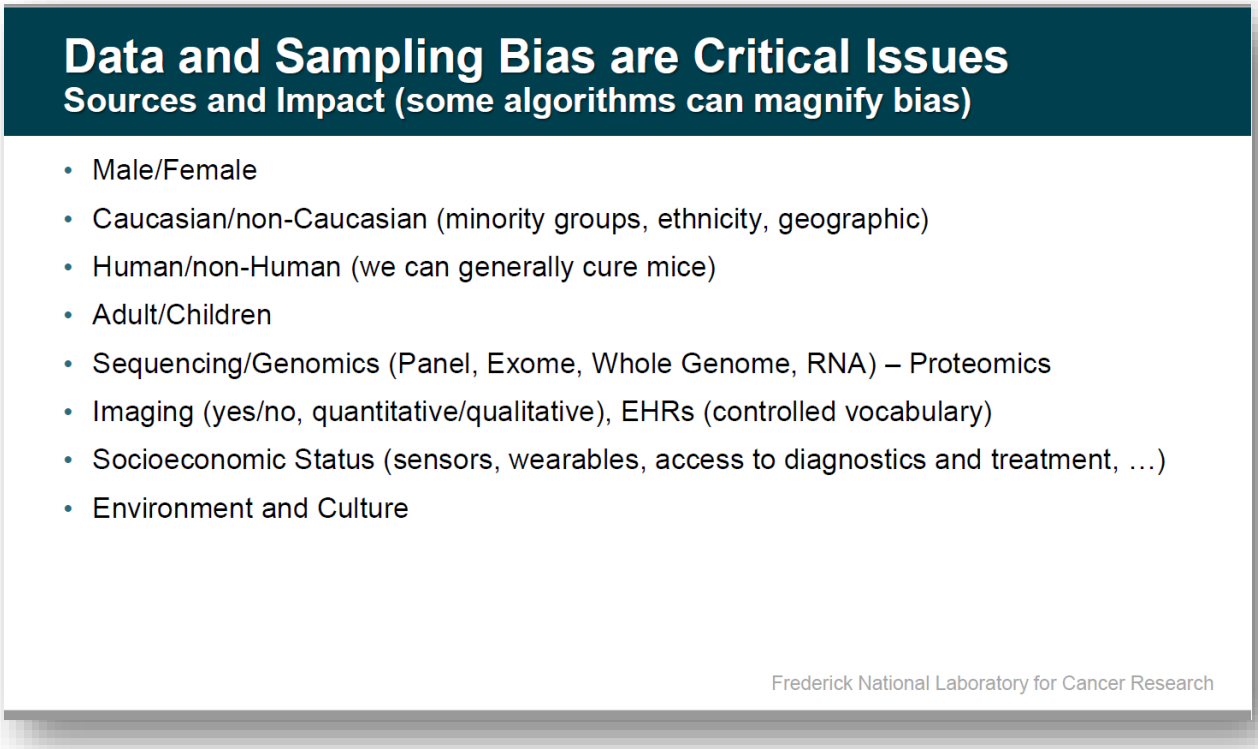
Collins stressed that for their efforts, machine learning is all about pattern recognition, and generally, the more data the better. Collins noted that the computer sees data differently than humans do. A pathologist, doctor, or radiologist will look at the data in one way and Collins and his team look at the data in a different way. In addition, researchers are being asked to collect data differently for a machine learning environment compared with when they collect data for their own research. The computer needs to see as many bits as possible whereas researchers want to throw most of the data away because it gets in their way.

In this space, machine learning can be compute-intensive, and researchers and software developers need to be careful because machines do what they are told. That is why great care must be used when telling them what to do. If it's an optimization problem it will optimize even if it goes straight to zero or negative; it doesn't matter, it will optimize. Ultimately, interpretation is critical as there must be

intelligent action. Researchers want the data to be useful: they are not just doing this as an academic exercise. That means they have to understand how they actually got to the solution.

Collins stressed although data is critical, researchers must be sensitive to issue of bias. For instance, is there a bias based on sex? Collins noted that for a long time, medicine was biased. Minority groups, ethnic groups, geographic locations all had biases in the way that data was collected and analyzed. The case even exists for human and non-human. For example, mice can be cured but that's not the goal. With adults and children, it's important to understand they are different populations. Even when sequencing and genomics are done, they can do panels, which is what happens in a lot of clinics: they do a small panel of a few targeted experiments and that's the final result. However, if researchers do a whole exome, which is just the coding parts, they can get a different answer, and that answer can change by adding the whole genome, the RNA, and the proteomics. As a result, the treatment changes, the diagnosis changes, and the prognosis changes.

FIGURE 3

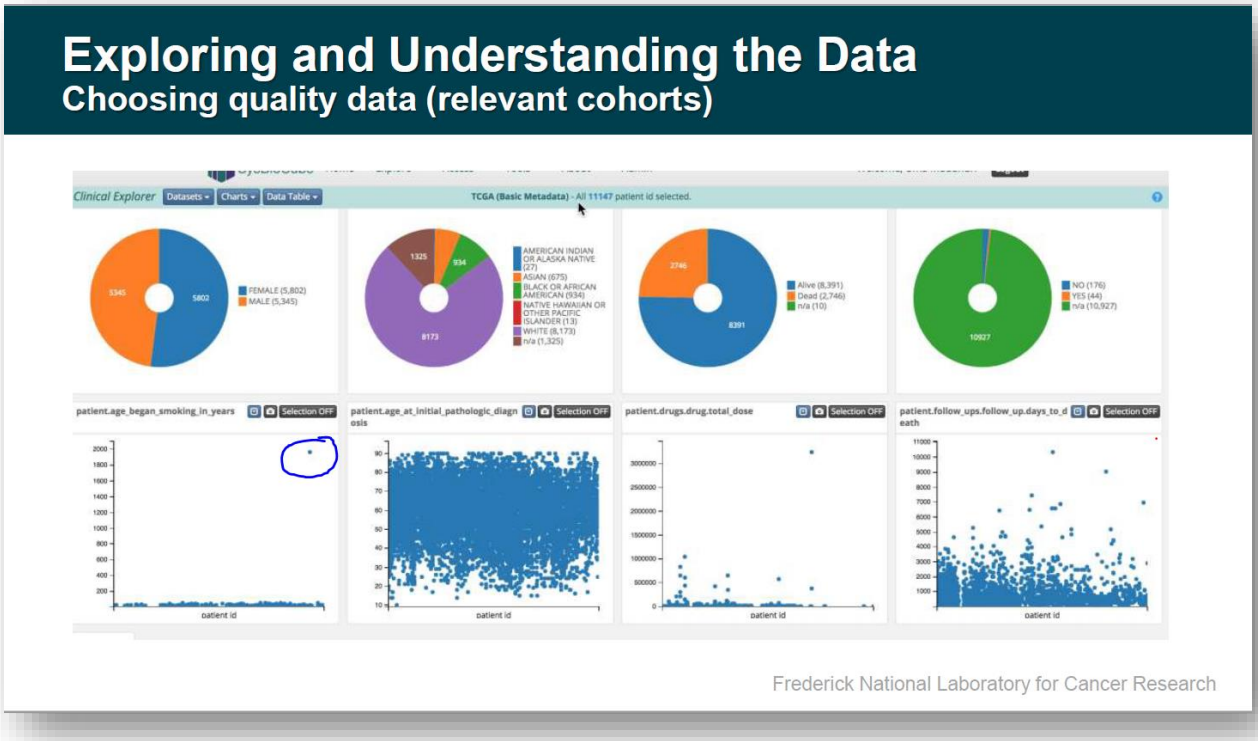


Source: Frederick National Laboratory and Hyperion Research, 2019

Collins showed some of the software his team is generating. They have a dashboard for all the data, the metadata, and some of the analysis it enables. For example, Figure 4 demonstrates an interesting feature: one user answered the question, "At what age did you begin smoking?" with, "1966." Clearly someone didn't start smoking at the age of 1,966, rather began smoking in the year 1966. If that is fed

into normal machine learning algorithms that is going to introduce errors. Collins noted that software developers must understand this, visualize it, and accurately and quickly allow the doctors and subject-matter experts to go in and fix such errors.

FIGURE 4



Source: Frederick National Laboratory and Hyperion Research, 2019

Collins concluded by stressing that how the models are trained are what matters. His team, when they first started training their models, ran the standard data runs over and over or did what they were asked to do by researchers, and those results had clear quantitation uncertainty. However, in the real world, there is a much larger spread of the answers. The real question for any machine learning effort is, "what is the truth?" What Collins wants to be able to do is get distributions, and if there is enough of an overlap then it's probably a valid and useful research tool.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
