

Cross-disciplinary Activities in Big Data for Agricultural Innovation

Carolyn J. Lawrence-Dill, Ph.D., Associate Professor, Department of Genetics, Development & Cell Biology and Department of Agronomy
Iowa State University

Agriculture is broad, involving not only crops and animals, but also the ecosystems that support their growth and development. Agricultural products are major economic drivers the world over. As the population increases and climates change, pressures on agricultural systems increase. At the same time, we seek to improve how we produce agricultural products by reducing inputs including pesticides, herbicides, antibiotics, and fertilizers. Taken together, these pressures tell us that we need to discover, design, and invent new ways to improve agricultural products.

Solving complex agricultural problems involves a multidisciplinary approach involving expertise from engineering, data sciences, and plant sciences. One way to engage a broader group in addressing these problems is to make the data that describe ecosystems, crops, and animals more easily accessible, comprehensible, and available to researchers. This extreme data sharing perspective is in keeping with long-standing traditions in science. It is widely recognized that information must be communicated or it is effectively lost (which is a driving force behind why research results are published) and that research results should be reproducible. Contrary to these basic assertions, Longo and Drazen state (in a commentary ironically entitled “Data Sharing”) that “someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters” (2016). This implies that those generating data must routinely fail to describe their materials and methods sufficiently to enable true reproducibility. The authors go on to assert

that, “There is concern among some front-line researchers that the [research] system will be taken over by what some researchers have characterized as *research parasites*.” It is possible that in some areas of research this perspective will prevail, but in the area of agricultural innovation stakes are high. To the degree that limiting access to data stands in the way of innovation, that position cannot be supported.

Data standardization seeks to improve both human and machine access to and analysis of data. The need for standardization was well described by Lohr, who reported that 50-80% of a data scientist’s time is spent aggregating and formatting data for analysis (2014). For crop improvement, a primary datatype selected for improvement is ‘phenotype.’ Phenotypes constitute all observable characteristics of an organism, so phenotypic descriptions include those traits that should be selected for improvement. Unfortunately for standardization, phenotypes include as many different types

of data as one can imagine. From leaf angle to root depth, from infrared spectrograms to molecular gel patterns, much can be observed. To add to this issue, metadata about the environment the organisms experience and even the level of observation (e.g., single plant versus average across a field) must be documented. The development of standards that would enable data discovery, simple aggregation, and wholesale analysis are largely lacking and where they exist, use is spotty, which is not surprising given that MIAPPE (Minimal Information About a Plant Phenotyping Experiment), the first well-described standard for collecting and describing plant phenotype data, was only released two years ago (Krajewski et al., 2015).

In argument against the development and use of standards for this emerging field of research, there are many discussions held among those working in the field debating whether the area is ready for standardization. The concern is that if standards are developed and their use is required too soon, novel mechanisms for data representation might be missed. What's more, if standardization constrains how scientists think of these data, some opportunities to develop new ideas for methods of analysis could be missed entirely, making the debate on creation and use of standards a hot topic at many scientific meetings where phenotyping is a focus.

Beyond the topic of how data are formatted and made accessible, the need for scientists with broad expertise to work together to address agricultural issues involving the measurement and analysis of

plant phenotypes remains. Unfortunately, cross-disciplinary efforts in the area of crop improvement remain an exception with most researchers working within well-described and narrow disciplinary boundaries. To push scientists to work more broadly, initiatives such as the Iowa State University Plant Sciences Institute (PSI) Faculty Scholars program have been initiated. For PSI Scholars, research funding is provided to faculty members working in the area of predictive plant phenomics (where phenomics is the set of all possible phenotypes a species could produce across all possible environments). PSI Director Patrick Schnable developed a program modeled on the Max Planck Institute in Germany and the HHMI and the HHMI-GBMF fellowship programs, which fund people rather than projects. Researchers working in the areas of plant sciences, data sciences, and engineering are funded to focus on plant phenomics problems and a community atmosphere is created among PSI Scholars by getting the group together weekly during the academic school year. Unlike traditional grant funding, PSI scholars themselves are funded rather than specific projects, giving them freedom to pursue specific projects they find to be of use to develop the discipline. For details see

https://plantsciences.iastate.edu/about_us/psi_faculty_scholars/plant-sciences-institute-announces-psi-faculty-scholars/.

Another mechanism Iowa State researchers have developed to push on this front involves student training. A grant from the National Science Foundation in Predictive Plant Phenomics (P3) supports novel graduate education and research

aimed at creating graduates with expertise in plant sciences, data sciences, and engineering. These local developments are reinforced by research networks like the North American Plant Phenotyping Network, a new organization founded by the broad research community, and by the creation of PHENOME, a new scientific meeting, first convened in 2017, which is organized by the research community and supported by the American Society for Plant Biology.

Because current approaches to agricultural improvement do not show the gains necessary to meet anticipated future needs, it is clear that the general approach to agricultural improvement must evolve. Through the development of shared data access and analysis mechanisms and by supporting cross-disciplinary collaborative activities focused on phenotype measurement and analysis, researchers are actively developing the infrastructure and human resources required to support the development of a new paradigm for research that results in agricultural innovation.

References

Krajewski P, Chen D, Ćwiek H, van Dijk AD, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S. Towards recommendations for metadata and data handling in plant phenotyping. *J Exp Bot.* 2015 Sep;66(18):5417-27.

Lohr, S. "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights" *New York Times* 17 August 2014.

Longo, D.L. and Drazen, J.M. Data Sharing. *N Engl J Med* 2016; 374:276-277 Jan 21, 2016 DOI: 10.1056/NEJMe1516564