Applied Economic Perspectives and Policy (2016) volume 38, number 3, pp. 373–388.
doi:10.1093/aepp/ppw009
Advance Access publication on 13 May 2016.

Featured Article

Data Science and Management for Large Scale Empirical Applications in Agricultural and Applied Economics Research

Joshua D. Woodard*

Joshua D. Woodard is an assistant professor and Zaitz Family Sesquicentennial Faculty Fellow of Agribusiness and Finance, in the Charles H. Dyson School of Applied Economics and Management, Cornell University
*Correspondence may be sent to: jdw277@cornell.edu.

Submitted 20 October 2014; accepted 23 March 2016.

Abstract *The increased availability of high resolution data and computing power has spurred enormous interest in "Big Data". While analysts typically source data from a wide variety of agencies, even within the USDA no comprehensive data warehouse exists with which researchers can interact. This leads to massive duplication in efforts, inefficient data sourcing, and great potential for error. The purpose of this article is to provide a brief overview of this state of affairs within the community. An overview of a prototype warehouse is also provided, as are thoughts on future directions.*

Key words: Agricultural Analytics, Big Data, Data Science, Data Management.

JEL codes: Q00, C55, C80.

Introduction

The increased availability of high resolution environmental, climate, and economic data, coupled with the dramatic progression of cheap computing power, has spurred enormous interest in the potential uses of data in large-scale empirical applications for agricultural economics, climate change, and agricultural policy research (Woodard 2016) Yet, our society's ability to meaningfully manage and use data has not kept pace with our ability to generate data. This leads to inefficiencies across institutions, agencies and universities, proneness to error, and profound duplications in effort. These phenomena are not unique to agricultural economics, climate change, and policy research, nor even agriculture, but rather represent and motivate a general trend currently pervasive to many branches of science (i.e., the so-called "Big Data" movement). Simultaneously, the advent of precision agriculture has increased our ability to manage and make use of site-specific

data for management via the interlinking and interfacing of various technologies and data, which itself can generate large amounts of data.

State-of-the-art cyber-infrastructures, robust information technologies, and innovative computational tools are of fundamental importance to all realms of sustainability science (National Resource Council 2012). To advance our understanding of the complex dynamics of the Earth system, and the multidimensional relationships that humans have with that system, there is general agreement that available datasets must be collected, validated, analyzed, visualized, synthesized, stored, organized, and distributed so as to maximize utility to the widest community of users (e.g., Office of Science and Technology Policy 2013). Each step in this process has technical prerequisites and challenges; in response, several modes of thought have arisen (e.g., efforts in the field of "computational sustainability"; Gomes 2009).

While policy analysts and managers typically source data from a wide variety of USDA agencies and other agencies (e.g., soil, weather, product, and related economic data), these data are typically not structured or organized in a way that easily enables policy research and access by industry participants (such as producers) or aides in the development of analytical tools, nor is there much focus on the same. Raw data from the agencies are often published at various levels of spatial and temporal resolution, which must be processed and transformed by individual researchers for specific purposes (where a handful of typical transformations may be common but not encompassing, e.g., county/annual), resulting in massive duplication in efforts, inefficient data sourcing, and great potential for error. Moreover, much of the data needed to pursue questions of interest reside in various agency transactional databases where privacy is a concern.

Current practices widely employed for data management and sourcing are inefficient and unwise for a variety of reasons, and impose great costs on taxpayers and researchers. For example, the way most researchers conduct their data sourcing and management might be approximately described as follows: the researcher will go to several different websites and sources at a fixed point in time, download specific slices of data, manually reprocess, dice, and reorganize data using a slew of different software, then combine them manually for the purposes of a one-off analysis within their chosen statistical program. This process is not easily replicable, typically not documented well, is not live/automated, is highly duplicative, is not scalable to other applications, and is prone to error. This essentially occurs because there is not a centralized repository or open source system that users can access and share. The benefits of even modest data warehousing efforts are fairly obvious. Yet, even within the USDA, for example, there does not yet exist a single comprehensive data warehouse or data management system with which researchers can interact.

The risks flowing from this current state of affairs in the community and within agencies from a research integrity perspective are not merely imaginary. Recent blunders in high profile journals such as Deschenes and Greenstone in the *American Economic Review* (2007) in which incorrectly processed but very common weather data were found to alter the findings of a longstanding and influential agricultural land value/climate change study (634 *Google Scholar* citations at the time of this writing)—highlights this risk. Therefore, while such data aggregation, integration, and warehousing ventures that we propose may not seem obvious at first glance, and do involve

upfront investment and a slight change in how researchers conduct their work, it is not only worthwhile, but also the responsible course of action one might argue.

The purpose of this article is to provide a brief and necessarily partial overview of the state of affairs within the community regarding practical data management issues, particularly regarding the large group of us that routinely work with large to moderately large U.S. agency and related data. The purpose is not to provide a taxonomy of what constitutes "Big Data", but rather to review some of the practical challenges, current efforts, and opportunities regarding data management and warehousing for agricultural and applied economists.[1] Likewise, the purpose is not to provide an in-depth technical review of database technologies. Rather, the intent is to provide a practical overview for the non-IT specialist. We also discuss pilot efforts to date to tackle some of these issues, and also offer thoughts on future innovations and directions. Our pilot platform focuses primarily on U.S. data, given the large user community, and the quality and quantity of uncurated data, but by no means is restricted to only U.S. data. Also, the plethora of data sources that are the focus of our initial efforts are widely used not only in agricultural economics, but also employed extensively in a variety of other fields in applied economics such as environmental, development, international, and natural resource economics. This platform stores and allows users to query consolidated raw and processed weather, crops, soil, market, and geographical data, among others, and can easily be joined and queried with other datasets in a form that is easy to document and ready for analysis.

Background and Motivation

The government, in conjunction with the research community, has a great opportunity to create systems to more efficiently store and distribute data collected through program and regulatory efforts ("administrative data") through greater sharing, more intelligent planning, and investment in information systems. These potential benefits have indeed been recognized by a variety of governmental and non-governmental organizations, agencies and councils, as evidenced by a variety of reports and memos in recent years (e.g., Office Of Management and Budget 2014; Office of Science and Technology Policy 2013; President's Council of Advisors on Science and Technology 2011).

Despite a large recent focus on "Big Data" within applied economics fields, the community and associated agencies as a whole have arguably made only relatively modest advances in terms of taking advantage of available data, tools, and paradigms. While policy analysts and managers typically source data from a wide variety of USDA agencies, as well as other agencies (e.g., soil, weather, product, and related economic data), these data are typically not structured or organized in a way that easily enables the policy research nor development of analytical tools, nor is there much focus on the same. Raw data from the agencies are often published at various levels of spatial and temporal resolution that must be processed and transformed by individual researchers, resulting in massive duplication in efforts, inefficient data sourcing, and great potential for error.

[1] *See Sonka (2014) for a good general conceptual overview of "Big Data" in agriculture.*

This state of affairs exists for a variety of reasons: a lack of awareness; inflexibility borne out of status quo; a historical lack of adequate funding/interest on behalf of relevant government agencies for modern industrial data systems for researchers; arbitrary and ad hoc restrictions imposed by some agencies on data; as well as a lack of focus on training in modern computing and database management applications at nearly every level, particularly in analytics and practical database programming.

A complex web of agencies generates and store data separately (public and private). Just within the USDA there is a plethora of sub-agencies and offices that publish and/or hold large amounts of valuable data that are routinely used by agricultural economists (or restricted from use), including the National Agricultural Statistics Service (NASS), the Risk Management Agency (RMA), the Economic Research Service (ERS), the Agricultural Marketing Service (AMS), the Agricultural Research Service (ARS), the Office of the Chief Economist (OCE), the Farm Service Agency (FSA), the Natural Resources Conservation Service (NRCS), among many others.[2]

Yet even within the USDA, no single comprehensive data warehouse or data management system exists with which researchers can interact. Nevertheless, these generalities extend far beyond just the USDA. For example, researchers spanning a variety of fields routinely may need to integrate—just for a single study—data from not only several USDA agencies, but also data from a variety of other government agencies and sources including, for example, soil data from the U.S. Geological Survey (USGS), weather data from the National Oceanic and Atmospheric Administration (NOAA), future market data from the Chicago Mercantile Exchange (CME), climate forecasts from the Intergovernmental Panel on Climate Change (IPCC), remotely-sensed data from the National Aeronautics and Space Administration (NASA), and spatial boundary files from the U.S. Census Bureau, among others. Indeed, our experience is not unique. Yet, robust and scalable open data platforms and data management systems for working with the diverse agricultural, remotely sensed, economic, and other data are not readily available.

There is a general lack of coordination among and within agencies to make data available and make use of such data not only internally, but also to extend it to the public. Another impediment is that often the de facto ownership of these data is viewed by being the "agency's data", as opposed to being fundamentally owned by and/or owed to the public in a reasonable form that still respects privacy laws. In other cases, a simple lack of bandwidth within the agencies prevents data from being made available in a timely manner, or at all. Privacy laws complicate this; as Sonka (2014) points out, societal responses to issues such as privacy will play a large role in shaping the future growth of "Big Data" and applications. Indeed, this tends to be a contentious area when it comes to certain agency data, as it is not always clear what personal information generated in the course of participating in government programs is or is not public. There is also a lack of consistency in terms of data-sharing protocols, and protocols to determine what is or is not public (which can change through time, often with no documented rationale).

[2]*For a comprehensive listing, see http://www.usda.gov/wps/portal/usda/usdahome?navtype=MA&navid=AGENCIES_OFFICES.*

For example, prior to 2008, Common Land Unit (CLU) data maps—which define field boundaries—were publicly available; with the 2008 U.S. Farm Bill, however, provisions were passed that restricted the distribution of CLU GIS data to the public and even other agencies. To many, this was a somewhat puzzling secrecy provision, as strictly speaking, information regarding who owns what land, and where it is located, is not generally considered private, and in fact can usually be obtained easily from the county or township clerk. This was all the more puzzling as the data had already been released to the public, so any "privacy" information was in fact no longer private anyhow. While this is an example where a legislative change can easily be traced back, similar inconsistencies regarding public data have also been observed within other programs. For example, for a short period after President Obama took office, it was possible to obtain policy-level data from the Risk Management Agency (which administers the crop insurance program), but later this access was restricted with no notice, legislative change, or otherwise. These types of arbitrary restrictions and inconsistencies hinder essential oversight, accountability, and research functions, arguably without leading to any meaningful, defensible, nor appreciable gains in "protection of privacy".

Despite this enormous need, it is becoming somewhat clear that the USDA—or any other agency—does not seem to be in a natural position to create and manage a data warehousing and integration systems effort for research purposes, particularly in cases where the data needs to extend beyond the agency itself, or necessitates an open-source or a crowd-sourcing system for processing tasks. The observation that agencies have not typically engaged in these activities in very meaningful ways historically supports this notion. This is undoubtedly the case at the USDA, although there have been some focused inventories and assessments that have been conducted through the years in specific areas such as conservation (e.g., the National Resources Inventory and the Conservation Effect Assessment Project; see Doering, Lawrence, and Helms 2013). Some public-private consortiums have also been developed, such as the Health Data Consortium.

This observation is not to fault the agencies necessarily, but only to point out that the task in and of itself may not naturally be suited for a government agency alone. More meaningful coordination with the research community seems to be seriously lacking. Likewise, since the products of these efforts are public goods, it is also not well-suited for a private enterprise, as the set of incentives necessary to bring these systems to fruition in a meaningful way to serve researchers, farmers, policy makers, and agency analysts may not align naturally with corporate goals.

State of Big Data Integration Efforts in Agriculture Economics, Climate Change, and Policy Research among and within Agencies and Universities

While "Big Data" integration tasks arguably are more naturally suited to land grant universities and other non-profits, to date little funding has been expended on such efforts. Besides our current pilot efforts that we discuss below, there exists little to no centralization of efforts within the agricultural and applied economics community or within the agencies to consolidate, aggregate, structure, and integrate such information in a modern and

meaningful way. While some exceptions exist for somewhat similar modes of thought in long-standing projects such as the Global Trade Analysis Project (GTAP)—which in fact is not a data warehouse or relational database, but rather a community of sorts—there is no such effort for the massive amounts of U.S. agency data relevant to agriculture, a realm within which a large number of economists, policy analysts, crop scientists, and agricultural economists work. The benefits from building out an industrial database management system (and dangers associated with not) are also not well understood by most.

There is some precedent for such data cooperation and support in spirit with U.S. data, although not within the context of a modern data management system. One example of a legacy system (which technically is still alive) is the USDA Economics, Statistics, and Market Information System (ESMIS), which is a collaborative project between Albert R. Mann Library at Cornell University and several USDA agencies. This system is essentially a repository of raw text documents from USDA announcements. To be clear, however, the ESMIS is not a modern database system, and cannot be "converted" to a modern database management system. It is a website created in the earlier days of the Internet that simply serves as a repository for raw unstructured text documents. For example, data cannot be queried from the ESMIS website, and there is no backend database against which to query the data contained in the text documents. This repository of unstructured and scattered text documents, unfortunately, lends little benefit to researchers and real policy analysts. In fact, conversations with the ESMIS staff indicate that the number one request they receive is demand for an actual database that users can access and query. Note also that the ESMIS system represents only a small subset of the data that could be made publicly available by the USDA.

The NASS Quickstats database (a purely government venture) was also surely motivated out of the same set of ideas to aggregate information, but for a variety of reasons does not fully parallel modern database and management capabilities. For example, the QuickStats database is essentially one large table, instead of being organized into well-structured underlying tables. This renders it not possible to query several joint and matched figures, but rather only allows for arbitrary stacking of redundant information that must then be processed manually. The database also only represents just a subset of available USDA data. Further, their web interface and API is severely limited in terms of the amount of data that can be queried (only 50,000 records, a number very easy to exceed) and cannot perform even the most basic of database tasks, such as group-by queries.

In other cases, efforts by the government to crosslink and duplicate certain slices of data across agencies from one agency database for use and storage in another agency's databases/tools/analyses has created an even more complicated and detrimental web of data dependencies and redundancies (so called "data automation" efforts within some agencies); from a broader perspective, such efforts in fact exacerbate issues, and clearly do not replace the need for a comprehensive data warehousing and integration effort.

Why is Data Integration Important?

Issues of data management and warehousing are important because enormous costs in terms of time and effort are incurred in collecting, processing,

integrating and making use of such data. In the current modes under which researchers in our field tend to operate, each and every study and research team across the system essentially replicates the data sourcing process internally and manually. This usually involves a plethora of steps, including manual downloading of diverse files from different USDA databases at specific points in time (usually slices of different tables within the various databases), manual reorganization through spreadsheets or ad hoc desktop software such as STATA, SAS, or Access, and narrowly-scoped (often non-reusable) processing workflows in order to put the data in a usable form that is ready for modeling. Virtually every agency and research group within the agricultural economics community (if not the world of broader economics) works in this manner to a large extent.

The current practices widely employed for data management and sourcing are dangerous and unwise for a variety of reasons. First, it renders the exact data sourcing process for validation purposes as essentially non-viable; that is, it is not replicable or transparent. Second, it is very prone to error, and oftentimes is conducted by a graduate student or researcher who may have little experience in dealing with such data. Nevertheless, even very experienced researchers can and will make data processing mistakes when performing these tasks manually. Third, the data are not live and not extensible; that is to say, if one wants to update the analysis, or use models generated in the course of using analysis to build out analytical tools, the data sourcing process has to be entirely recreated.

Indeed, the outcome is that most models that we develop are never in fact extended in any practical, useful, or usable way outside of a very small group of specialists with the ability to access the nuances of the research. This also has enormous implications for the validity, transparency, and replicability of research. Even for journals that require code and data to be posted, that data virtually never comes with an automated link back to the live source or sufficient information to actually recreate processing steps from source data accurately. Any future users have to instead rebuild the entire data sourcing process. Of course this happens rarely due to the time involved. Fourth, it is not easily scalable for other applications, and essentially needs to be recreated manually for applications to other domains (e.g., to apply to other crops or regions), or to extend in the form of web-based analytical tools.

For example, suppose a researcher collects data on soil, weather, yield, price, and insurance data to conduct an analysis that is focused on the Midwest for corn for grain at the county level, for a set of years. This would involve going to no less than 5 different agency websites to download data. This would typically entail downloading several different files from each agency, or accessing several databases within each agency alone—if the data are actually in a database and not just in separately posted text files. Oftentimes this downloading has to be done in chunks manually. For example, the NASS Quick Stats system limits downloads to 50,000 records. It so happens that a typical analysis of corn for grain for the major production states at the county level for the historical period exceeds this number. This not only creates great frustrations for researchers, but limits what we can feasibly do.

To process the raw soil and weather data in order to aggregate it at the county/annual level, (if one is familiar with working with spatial and/or GIS data) the researcher would then need to either spend several hours

coding, organizing, and processing the raw data by location and time period (a nontrivial process for those who have never done it, especially for structured space-time data such as weather), or would have to outsource such processing to a consultant. The researcher would then manually reorganize, slice and dice, and stack their data from all of the other sources and load it into their analysis program (or load it in and do one-off slicing and dicing for the particular application, in this case say corn for grain at the annual/county level for a subset of locations/time periods). If the researcher wants to run some regressions, they would then stack and organize all of the data appropriately together in structured matrices, and then would be in a position to run an analysis.

Handing over that processing chain for replication to another researcher is not entirely transparent or turnkey, and in fact will be very prone to error (and is usually only ill documented at best). That is, it is not easily extensible; it is also not scalable. If the researcher then wants to scale out the analysis to other areas, crops, time periods, etc., in many cases they need to manually recreate much of that processing chain (a set of steps that takes hours or days potentially). It is also not reusable for other applications. Lastly, it is not live. The whole process would have to be conducted again if, say, the researcher wanted to update the analysis with new data in a year's time. Using a modern data management system, on the other hand, this would literally take a handful of minutes. Replication by others would also be perfectly transparent, virtually effortless, and the steps perfectly documented by the code itself. Updating at a later date can also be done at a click since all of the data sourcing steps are coded and run against an actual database.

The costs imposed by current (dated) data management practices are not merely trivial or artificial, but rather pervasive, large, and real. For example, consider that annually, various universities and federal, state, and other funding agencies routinely spend hundreds of millions of dollars or more to fund individual agricultural economics research individuals or groups (externally and internally). A large part of the time that is funded under these various grants is for very basic data sourcing, which is typically done manually. This is not unique either to just agricultural economics, but applies equally to environmental, natural resource, international, and development economics.

These data sourcing processes, which could be feasibly automated, are instead performed over and over again using suboptimal workflows (spreadsheets, manual data downloads, and one-off processing schemes, etc.), leading to massive amounts of work hours (potentially millions of wasted hours), which could translate into the hundreds of millions of dollars, if not more. Moreover, instead of learning and teaching more advanced data management systems needed in this new era of "Big Data", this work is often performed with the last generation of data management tools, tools which frankly are not suitable for the job in many cases. While technologies such as spreadsheets are incredibly useful for some jobs, it is no secret that they are probably massively overemployed in research. Indeed, their overuse crowds out the adoption of more robust technologies such as industrial, well-designed database servers and data warehouses.

Ag-Analytics Pilot Platform

A pilot integration and warehousing effort to serve such data management needs for the broader agricultural and applied economics field is

available online at *ag-analytics.org*. These efforts are well underway, and the data warehouse already includes many relevant major public datasets, including the bulk of public data available from NASS (e.g., QuickStats Survey and Census data, as well as the Cropland Data Layers), publicly available RMA insurance data, much of the publicly available ERS data, weather data from PRISMS, soil data from NRCS, among many others (see table 1). Our initial priorities are the standardization, synchronization, and integration of data across sources, models, and user needs.

The platform integrates, stores, and updates from a variety of resources daily including PRISMS weather, NASS, RMA, ERS, and CME, among many others, to provide producers, researchers, and tool development efforts with live information on demand, and is the backbone of our research and outreach efforts, which allows us to more seamlessly integrate and transform research into tools (see table 1 for an overview). Through this centralization, we have accomplished the following:

- Eliminated the need for ad hoc data collection specific to individual research efforts.
- Enabled easy updating of econometric analyses as new data becomes available.
- Created an open-source web-based interface for academic, government, and other researchers to integrate, query, and process large datasets in an automated, extensible, and scalable fashion.
- Enabled more efficient translation of research efforts and models into open-source web tools for farmers, policy-makers, and researchers.

We have also built a variety of API's and tools to streamline different data sourcing processes. A primary initial objective of this effort has been to identify and evaluate methods and models for handling data sets and data streams, and then processing and storing such data (e.g., weather, climate, land use, water resources, soil surveys, topography, and transportation networks) in a way that is useful for making informed decisions about agricultural policy, productivity, and markets. This is a demanding task in and of itself. Individual datasets and data streams are typically large, formatted differently, and have varying spatial and temporal resolutions. We have found it to be a significant technical endeavor to source, automate, and structure these data into standardized formats that can be used in conjunction with each other in a more general framework, albeit with many successes.

Such efforts are sometimes incorrectly viewed as "just putting all the data in one place", while in the other extreme, such efforts are viewed as impossible and/or that they would take too much effort. First, simply downloading slices of data from different sources at a fixed point in time to a hard disk does not equate to a modern database, nor does just simply having several diverse data and text files on the same physical disk or server. Further, data that are not loaded and organized in a well-structured, live database management system can also not be joined, related, collectively queried, extended, or updated easily. Further, a well-designed system should meet requirements relating to validation, uncertainty, credibility, clarity, and openness (Maxwell and Costanza 1997; Eliot et al. 2013). Practically speaking, performing research in the absence of such frameworks is disadvantageous for a number of reasons including that the research is rendered not transparent, not scalable, not extensible, not live, and not automated, among other weaknesses. Further, while in reality the challenges of building such a

Table 1. Abridged Summaries of Major Sources Currently on the *Ag-Analytics* Platform

Data Source and Item	Description
IPCC Climate Change Projections	Future temperature and precipitation projections across different emission scenarios and percentiles of the 16 General Circulation Models (GCMs).
National Climatic Data Center Drought Data	Monthly PDSI drought index. Data is available from 1895 to present, by NCDC District and at various levels of aggregation.
PRISMs Climate Group	Monthly and daily historical temperature and precipitation data, as well as ability to generally process GDD/HDD data on the fly. Monthly data is available from 1895 to present. Daily weather data is available from 1981 to present. Data are available at several levels of aggregation.
Chicago Mercantile Exchange	Daily historical futures and options data for agricultural commodities from the Chicago Mercantile Exchange, Chicago Board of Trade, and Kansas City Board of Trade. Data is available from 1959 to present, updated daily.
Risk Management Agency	Agricultural insurance price and participation data available at the county level aggregation. Data is available from 1989 to present from Summary of Business. Other data also loaded from various unstructured text files (including historical discovery prices, GRIP yields, etc.)
U.S. Census Bureau	Various boundary files for different levels of aggregation (county, state, etc.).
USDA Economic Research Service	Publicly available Agricultural Resource Management Survey (ARMS) data. Unlike the ERS ARMS API, the data in our database are processed for easy retrieval of large structured sets of data.
USDA Agricultural Marketing Service	Monthly data on the volume, pricing, and utilization of raw milk received by handlers regulated under federal milk orders from dairy farmers. All tables in the Public MMO database.
USDA National Agricultural Statistics Service	Census and survey data available at regional-, state-, and county-level aggregation. The broad categories of data available are crops, animals and products, economics, demographics, and environmental. Data is available from 1926 to present. Obtained via FTP bulk download from QuickStats. CDL data processed against ready to map gSSURGO NRCS data by crop also available (raw and county processed). Unlike Quickstats, API can be queried generally via our API.
USDA Foreign Agricultural Service	Data on the production, supply, and distribution of agricultural commodities for the U.S. and key producing and consuming countries.
USDA National Resource Conservation Service	Soil data for the continental U.S. from gSSURGO, available at various levels of aggregation.

system can be complex, the benefits are enormous, and our pilot efforts have proven that it is not only possible, but that it also represents an enormous improvement over current practices.

With that being said, it does entail upfront investment to build a well-structured and live data warehousing system such as this one, and such efforts do require maintenance and funds for server space and programming support. The public good aspect, however, is such that having a community around such a platform that contributes to doing it right one time is a far superior approach to the current environment where each and every researcher has to recreate all of this manually. We have shown that building an extensible system is clearly not insurmountable. In fact, even with a very small amount of resources, we have created a very useful system in the course of performing our other research, along with some added foresight in planning. Clearly, however, much remains to be done.

While our initial data system efforts focus primarily on U.S. applications—and may rightly be considered only "moderately Big Data"—the fundamental ideas and motivations underpinning these approaches span naturally to virtually all fields and subfields, and it is anticipated that many such "intelligent" open source/open data warehousing systems will develop in different fields within agricultural economics in the coming years and become standard fare. While some commercial platforms exist for data dissemination (including some of the data we house), an open source/open data model is likely to have large benefits and fill a gap not easily filled by commercial products. For example, commercial products do not provide a proper venue for others to contribute (e.g., if a new dataset is published and needs to be processed for use). Commercial platforms are almost always black boxes as well. Even in cases where commercial products allow for open data querying, how they obtained and preprocessed the data is often left as a mystery.

These datasets are retrieved from various data sources over the Internet and transformed before being stored in the database. Each data set is automatically updated on a regular basis. In some cases data are reorganized into more logical tables or sub-databases from several different sources. Outputs from our statistical models that support web analytical tool products are also stored in the database, and the web tools query the database directly on the fly. As noted, the raw data are often on several different dimensions of temporal and spatial aggregation. Our database provides raw data as well as spatially processed data that is ready to use (e.g., average temperature data aggregated at the county level, township level, state level, etc.). Our vision is that these efforts will lead to an open-source platform that will allow researchers, farmers, and government agencies to more efficiently access and employ the vast amounts of data available for policy analysis, and better enable the development of tools to help farmers understand and manage risks in an integrated manner. We also envision opening the system to crowd-sourcing in order to facilitate future development, flexibility, and wide user adoption. We also have several different API's, from general to specific, that allow researchers to query, process, and access data in a convenient and scalable manner.

This industrial grade system already serves as the backbone to our research and analytical webtool development efforts, has led to greater project efficiencies, and has greatly increased our purview of capabilities. Such efforts are long overdue and will only become more critical in upcoming

years. In many ways, this system does not present a huge advancement in technology per se, as it employs a combination of practices and technologies that are in wide use in other fields, albeit perhaps to a different and unique scope of integration. This system does, however, point out that the adoption of such technologies and methodologies remains seriously lacking in agricultural economics and related research, and that this imposes high and unnecessary costs on the government and research institutions, and results in data products that are less useful to not only researchers, but also end-users (i.e., farmers and policymakers).

Web Decision Tools—Integrating Research and Tool Development

Related to our *Ag-Analytics* platform efforts, we have also begun to develop a suite of decision and outreach tools based on this work. A major impediment to extending research efforts to the public is there is often little incentive, and moreover very limited resources and interest by donors and stakeholders, to see that the extension of research occurs. This last step of translating research into decision tools is not a trivial process. The research work is often performed in an environment where the data sourcing is ad hoc. This creates enormous implementation impediments if/when it becomes time to deploy and design a tool. However, if the research and tool development efforts rely on a common platform from inception, it becomes a much less time consuming and expensive task to deploy research models. This has an added benefit in that it is very easy to then set up auto updates for developed tools since the database itself (which supports several research projects and web applications simultaneously) is already designed to auto-update. This also makes it possible to actually replicate and validate the research, not to mention update the original analyses in the future.

Future Directions

Generic Research Information and Decision-making Platforms

Building upon successful approaches developed as part of initial efforts, a next logical step may be to develop high-performance computing capabilities into a cohesive infrastructure that is capable of supporting dynamic, stochastic models for assessing environmental and economic risks at a range of scales more fluidly. This would allow for the facilitation and integration of crop-based simulation models (e.g., toward disease prediction or insurance estimation) with probabilistic forecast models (e.g., from weather and climate), calibrated to large historical datasets and models based on them. This would also promote the rational design of field-based research programs and precision agriculture technologies and efforts that aim to survey, manage, and maximize agricultural yields. Realizing these objectives will represent critical steps toward achieving the sustainability of agricultural systems, improving crop-risk assessment and management, minimizing agricultural and economic losses, optimizing policy designs, and increasing financial returns to farmers and farming communities (Hansen 2005).

Such a data framework could be structured upon a uniform grid of the Earth's surface with the dimensions of individual grid cells depending on either the process being studied or defined on-demand by the ultimate need of the user. While initial relational database systems such as the *Ag-Analytics*

platform would serve as a necessary precursor and input into such a system, these future strategies would represent an operational departure from network approaches that designate arbitrary geopolitical units as nodes (e.g., U.S. counties) in analyses of agricultural production, land use, etc. Rather, in such a system, each grid cell would be populated with relevant data to support specific agricultural, socio-economic, and environmental models. The data would be sourced from available sources such as the *Ag-Analytics* platform, and then spatial and temporal aggregation/disaggregation and interpolation functions could be applied to generate continuous or gridded surfaces.

A successful platform of this sort would be a singular platform comprised of modular components that users may either query directly or utilize as an organized foundation for building network-based models at various scales of resolution. Further thought would need to be put into which type of operational framework would facilitate the broadest appeal across disciplines. In any case, given the wide variety of disciplines involved in building out such a system, it is clear that such work would not correspond to any standard definition of disciplinary research, but is still of great importance.

To our knowledge, a flexible, efficient public domain infrastructure of this type, capable of supporting models of weather and climate-dependent processes in a dynamical fashion, along with market data, does not yet exist. Certainly general inspiration for such a system comes from recent efforts to gather data regarding the Earth system in general, such as the Global Organization for Earth System Science Portal collaboration (GO-ESSP; http://go-essp.gfdl.noaa.gov/) and the related Earth System Grid Federation (ESGF; http://www.earthsystemgrid.org/home.htm). Nevertheless, in attempting to synthesize agricultural, socio-economic, and environmental data resources for the broadest spectrum of applications, a functional future-proof system of this type will represent a unique type of infrastructure and more universal tool for dealing with specific weather and climate-dependent problems in various subfields of agricultural and applied economics.

Secure Data Warehouse for USDA Administrative Data

Some work cannot be feasibly accomplished without being able to link together different databases at low levels of aggregation. In the case of USDA data, different administrative databases are often confidential and also reside in different agencies. The result is that without participation from a variety of agencies, certain work can never be done. Some analyses can only be done if data are joined at this level. For example, suppose a researcher wanted to be able to do a national-level scale analysis of yields, soil type, and insurance losses. While the data to do so exist, the data themselves reside in confidential databases under at least 3 different agencies. A secure data warehouse could grant researchers access to confidential field- and farm-level data with technology and protocols in place to protect producer privacy per applicable laws. A data warehouse could also contain publicly available data. Such a data resource would have enormous value outside of the specific charge identified above, which would support a wide range of next-generation policy analyses.

There are a few limited precedents for such secure environments, but they tend to be one-off enterprises. For example, the ERS has enclaves that authorized researchers can access via secure enclaves for the ARMS survey

data, although access is generally limited to those who have personal ties to someone in the agency willing to sponsor a cooperative agreement. The U.S. Forest Service has also developed a limited secure data warehouse that allows researchers to access and analyze specific Forest Inventory and Analysis (FIA) data, and uses a "fuzzying" algorithm to mask location. In order to gain access to the data, researchers must submit an application and fill out a Non-Disclosure Agreement (NDA).

A secure data warehouse also exists at Cornell University, known as the Cornell Restricted Data Access Center (CRADC). Established in 1999 as a pilot sponsored by the National Science Foundation, the center provides researchers with access to confidential public and private data in CRADC's secure computing environment that meets or exceeds U.S. Defense Department C-2 standards. Before gaining access to the data, researchers must undergo a screening and approval process. CRADC currently houses datasets distributed by multiple agencies, including the U.S. Bureau of Labor Statistics, National Longitudinal Surveys Program, the U.S. Equal Employment Opportunity Commission, and the Inter-university Consortium for Political and Social Research (ICPSR), among others. A last example are the U.S. Census Bureau Federal Statistical Research Data Centers, which house confidential census data.

A defining feature of all the secure data enclaves mentioned above is that they are not integrated in the sense of having data from across multiple agencies on an integrated platform that allows for joining together datasets. More coordination and cooperation between academia and the agencies is needed to develop such secure centers. For the USDA in particular, there is enormous potential value in consolidating and bringing together under one secure roof the various administrative and survey data that currently reside in closed-off siloes. To date, the USDA has been unwilling to participate with academic researchers in the construction of such a secure data center. Some arguments against are that the government needs to "protect us" from the data, as academic researchers might "do the research wrong" if they had such data centers. Although one is left to suppose this is always the case with any data, and that this is one of the primary reasons for processes like the peer review.

Conclusion

Agricultural, natural resource, environmental, development, and related applied economics fields sit at an exciting vantage point in the current environment. Unlike many social sciences, our work and study often lies at the intersection of large complex social, natural, and environmental systems. Many applied economists routinely collaborate with others from diverse fields, from crop and soil scientists, geneticists and plant breeders, and environmental engineers, to meteorologists, mainstream financial economists, and climate scientists. The computational tools and data necessary to tackle many outstanding new problems are now becoming more readily available, and this has the potential to unlock access to an incredible set of possibilities for the field. We are also endowed with massive amounts of high-quality, diverse, and interesting data and questions. In this new era of "Big Data" and increased focus on analytics, more serious and concerted efforts on data management, structuring, and access—alongside increased

agency cooperation—are needed for the discipline to fully realize and take advantage of this paradigm shift.

Future research related to topics such as this in the realm of "Big Data" and large-scale empirical applications will likely take on an increasingly larger role in shaping agricultural, natural resource, and climate change policy. Accordingly, researchers and policy makers need improved tools and data management systems with which to interact in order to adequately develop, assess, and monitor complex programs and policies. Such solutions should involve better cooperation among agencies and universities, with universities leading the aggregation, sourcing, and distribution efforts, and with support and input from agencies to obtain data and funding, and to define protocols for personally identifiable information, while still having source administrative data on a secure server that can be queried by users.

The sheer scope and complexity of the "real world" necessitates adaptable and accessible data infrastructures, computational models, and visualization methods to tackle the research questions of the future. The President's Council of Advisors on Science and Technology highlights these data and information technology needs, emphasizing the value of: 1) "modeling and simulation decision-support software that incorporates the many kinds of data, and the massive amounts of data, to build predictive scenarios that take into account the complexity of natural systems and the impacts and competing demands of human systems," and 2) "the underlying data and information infrastructures that mobilize data for use in these simulations and model,s" (PCAST 2011). Current efforts are a necessary precursor to such systems, and will usher in durable next-generation decision-making frameworks in the agricultural and applied economics sphere.

References

Deschenes, O., and M. Greenstone. 2007. The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather. *The American Economic Review* 97 (1): 354–385.

Doering, O.C., D.J. Lawrence, and J.D. Helms. 2013. Agricultural Conservation and Environmental Programs: The Challenge of Data-driven Conservation. *Choices* 28 (2).

Eliot, J., M. Glotter, N. Best, K. Boote, J. Jones, J. Hatfield, C. Rosenzweig, L.A. Smith, and I. Foster. 2013. Predicting Agricultural Impacts of Large-scale Drought: 2012 and The Case for Better Modeling. University of Chicago, The Center for Robust Decision Making on Climate and Energy Policy, Working Paper No. 13-01.

Gomes, C.P. 2009. Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society. *The Bridge* 39 (4): 5–13.

Hansen, J.W. 2005. Integrating Seasonal Climate Prediction and Agricultural Models for Insights into Agricultural Practice. *Philosophical Transactions of the Royal Society B* 360: 2037–47.

Maxwell, T., and R. Costanza. 1997. An Open Geographic Modeling Environment. *Simulation* 68 (3): 175–85.

National Research Council. 2012. *Computing Research for Sustainability*. Washington DC: National Academies Press.

Office Of Management and Budget. 2014. Memorandum For The Heads Of Executive Departments And Agencies: Guidance for Providing and Using Administrative Data for Statistical Purposes.

Office of Science and Technology Policy. 2013. Increasing Access to the Results of Federally Funded Scientific Research. OSTP Memorandum for the Heads of Executive Departments and Agencies, Washington DC.

President's Council of Advisors on Science and Technology. 2011. Sustaining Environmental Capital: Protecting Society and the Economy. Washington DC.

Rounsevell, M.D.A., J.E. Annetts, E. Audsley, T. Mayr, and I. Reginster. 2003. Modelling the Spatial Distribution of Agricultural Land Use at the Regional Scale. *Agr Ecosys Env*. 95 (2–3): 465–79.

Sonka, S. 2014. Big Data and the Ag Sector: More than Lots of Numbers. *International Food and Agribusiness Management Review* 17 (1): 1–20.

Woodard, J.D., Forthcoming 2016. Big Data and Ag-Analytics: An Open Source, Open Data Platform for Agricultural & Environmental Finance, Insurance, and Risk. *Agricultural Finance Review* 76 (1).

U.S. Department of Agriculture, National Agricultural Statistics Service. 2013. Crop Values 2012 Summary. Washington DC.