RJ10520 (ALM1405-018) May 29, 2014
Other

IBM Research Report

The IBM Research Accelerated Discovery Lab

Laura Haas, Melissa Cefkin, Cheryl Kieliszewski,
Wil Plouffe, Mary Roth
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
USA

The IBM Research Accelerated Discovery Lab

Laura Haas, Melissa Cefkin, Cheryl Kieliszewski, Wil Plouffe, Mary Roth
lmhaas, mcefkin, cher, plouffe, torkroth@us.ibm.com

1. INTRODUCTION

Data analytics is becoming central to modern society. In the business world, financial institutions rely on data analysis to detect and prevent fraud; retailers combine transaction data with social media and emails to build a better understanding of their clients; industrial giants use sensor data to improve their carbon footprint. Meanwhile, bioinformatics, astronomy, and particle physics are just a few of the sciences that are being transformed by the availability of large data sets and new techniques for analyzing data. Cities, governments and social agencies are leveraging data analytics for important causes such as improving public health and planning for (and reacting to) natural disasters.

But the path from raw data to insight, or better yet, predictive or prescriptive capabilities, is still long, error-prone, and expensive. First, data must be acquired – not only the pertinent domain data, but often reference and/or contextual data, such as data on weather, economics, demographics, or maps. An appropriate systems infrastructure is needed to store and process the data. Once that infrastructure is acquired, the data must be cleansed, integrated and transformed before the real analysis even begins. Each of these steps requires different tools, and often expertise not only in those tools but also in the various data sets, in data management, and mathematics. Analysis itself involves more tools, deep knowledge of the domain of inquiry, and, if the volume or velocity of data to be analyzed is high, substantial systems and algorithmic skills may be required as well to achieve acceptable performance, with the necessary accuracy. Few people have this broad range of knowledge; thus, many experts need to collaborate across disciplines to achieve the desired insights.

The IBM Research Accelerated Discovery Lab is a unique, collaborative environment specifically designed to facilitate analytic research projects that require multiple participants who may be from several disciplines, and even several institutions. Discovery, in our context, means the gaining of new insight or understanding, often with the intent of attaining predictive or prescriptive capability, where the analysis of data plays a central role. The Lab's objective is to accelerate this type of discovery by (1) enabling research in and improvements to the tools and systems that facilitate discovery, and (2) enabling the business person or domain expert who uses the environment to focus on their investigation instead of the systems and data challenges. To accomplish the first two objectives, we also need (3) to understand how discovery occurs, and how it can be accelerated.

2. LAB OVERVIEW

To achieve the Accelerated Discovery Lab's objectives, we focus on the *discovery platform* the Lab provides to support the discovery process, the *partner projects* that leverage the platform, and the studies that explore the *practice of discovery*.

The discovery platform includes a secure cloud environment that supports large-scale data-intensive computations and a software system that encourages discovery. The cloud environment includes several hundred compute nodes, over 12 petabytes (PBs) of online storage, and a high-speed network as the hardware infrastructure; it leverages IBM's Platform Cluster Management and supports a wide variety of information management tools and analytics platforms. The software system includes data curation tools, support for collections of data called data lakes, and a library of analytics tools and models. It also provides LabBook, a social user experience in which our partner projects pursue their research. Both data lakes and the analytics library allow contribution of new elements (data sets or analytics, respectively), which may be created as a result of projects run in the Lab.

We are supporting a diverse set of partner projects of two types. *Analytics* projects tackle challenges from multiple domains. They range in scale from month-long investigations by a few researchers of a narrowly-defined question requiring one or two data sets to answer, to multi-year studies by multiple teams that require tens of data sets and petabytes of data. *Systems* projects also range from short-term performance studies of new algorithms or architectures, to longer-term creation of new analytics tools or information integration capabilities. While some projects may be done "in residence" in our collaboration space, described below, most are done by teams who may not be local and may, in fact, be geographically distributed. Hence the discovery platform needs to support collaboration across locations and time zones. Partner projects vary over time; potential partners are chosen based on the alignment of their interests with the Accelerated Discovery Lab's mission, their ability to exploit the Lab's platform, and their tolerance for running in an experimental environment.

Finally, we are exploring the human and social dimensions of large-scale data-intensive research and discovery practices, studying how discovery is conducted to identify essential technological, informational, and environmental characteristics that can encourage and perhaps even accelerate discovery. At some level, the discovery process can be seen as a set of analytics experiments run over some set of data [1]. But what are the right experiments? What tools should be used? What data? Often an individual researcher relies on familiar or available tools or the advice of colleagues. However, examples such as the discovery of the link between fish oil and Renaud's syndrome [17] show that discovery may also happen when previously isolated projects collide in new and unanticipated ways, or when individuals with different (technical) backgrounds collaborate or exchange ideas. Our studies include observations of our partner projects, as well as experiments with the software and physical environments to understand how to provide the best conditions for discovery, including, perhaps, serendipitous interaction that sparks insight.

One of the affordances of the Accelerated Discovery Lab is a 7500+ square foot workspace that provides a flexible work environment for individuals and groups. The space is outfitted to facilitate creativity and collaboration through access to simple, yet effective tools such as whiteboards and displays that can be moved and configured for the needs of those using the space. Also, with our researchers and partners scattered worldwide, we need to be sure that all can participate in planned and *ad hoc* engagement. The workspace thus includes standard collaboration technologies such as video and web conferencing, along with less standard telepresence robots that allow remote wandering through the room. The space not only affords our researchers and clients a place to work and explore, it also provides a rich environment for collecting data to support our discovery practices research, using methods such as interview, observation, and log data analysis.

Each of our three research thrusts, the platform, partner projects and discovery practice studies, is driven by researchers from different disciplines. The platform research is driven by computer scientists; our core team includes database, human-computer interaction, and systems researchers. The partner analytics projects are typically staffed by domain researchers or analysts; some are "data scientists" with strong data or algorithmic skills. The systems projects, by contrast, are led by computer science researchers, some of whom may have analytic skills. Finally, the discovery practice studies are led by teams of social scientists from such disciplines as anthropology, sociology and social computing. Thus the Lab itself is a multi-disciplinary research environment, mirroring the discovery projects it supports.

Many other groups have or are creating institutes that focus in one way or another on data-driven discovery. Physical science labs[1] have built substantial cyber-infrastructure to support sharing data and tools for analytics. As interest in data analytics has expanded, many universities have formed data science institutes[2], typically multi-disciplinary endeavors that attempt to bring computer scientists, statisticians and domain researchers together to solve domain-specific problems. In the commercial realm, data marketplaces[3] are starting to add computational analytics capabilities to the collections of data sets they provide. Meanwhile, computer science efforts such as CLDS[4] and Berkeley's AMPLab[5] bring together several branches of expertise to improve the systems for doing analysis, and to prove them on real domain-specific challenges. The Accelerated Discovery Lab has many elements in common with each of these efforts; however, to the best of our knowledge we are unique in our emphasis on supporting the overall discovery process (section 4) and our focus on understanding, from a social science perspective, how discovery happens and how it may be accelerated (section 6).

This paper is organized as follows. In the next two sections we provide more detail on the cloud environment and the software systems of the discovery platform, respectively. Section 5 gives a few examples of current partner projects, both analytics and systems, while Section 6 addresses our studies of discovery practices. We discuss where we are today, the research that is currently underway, and where we hope to go in the future.

3. DISCOVERY CLOUD ENVIRONMENT

The discovery cloud is the backbone of the discovery platform, where data lives, and algorithms are tested. It consists today of almost 500 multi-core servers, a high-speed network from our partner, Juniper Networks[6], and capacious storage. The flexible hardware infrastructure provides a rich experimental platform tuned to run large-scale data-intensive analytics, and supports the key analytics tools our partners need. A key consideration in our design is ensuring that data and systems are protected. The architecture allows secure access by authorized researchers (IBM and external) to both private and open data. Measures taken to ensure privacy include restricting access to the systems, secured logins (LDAP), role assignments, security scans, and controlling internet access. Projects can be physically isolated from each other, or run in a shared pool, depending on their sensitivity.

The physical machines are of two types. The compute server is characterized by a 1:1 ratio between hardware cores (processors) and drives (spindles), using physical drives so that seeks may be overlapped with other I/O. This configuration is best for physical or virtual (VM) systems that need to optimize the parallel I/O or to use local storage. Such systems include Hadoop and its various implementations, e.g., IBM BigInsights and Cloudera, and IBM SPSS Statistics Server. Compute servers have 12 Intel x86 cores, 12 2TB drives and either 96 or 192 GBs of main memory.

Used as a physical server, the compute server is dedicated to a single purpose – usually a single partner project. When hosting VMs, it may be shared by multiple projects. High performance computing applications run on physical servers for best network utilization. Finally, projects that use sensitive client data would have dedicated servers (even if running VMs) so that the data can be permanently destroyed by wiping the disks at the project's end.

The second type of server, the hypervisor server, is used to host multiple VMs, which may be for multiple projects. These would be VMs that are either compute or memory bound and do not need locally attached drives because of either low I/O bandwidth requirements or a small drive footprint. Such servers are used by web servers, database, and user interface servers. Hypervisor servers are bigger systems, with 32 to 40 Intel x86 cores, 128 to 512 GBs of memory, and 6 1TB drives that can be configured as needed by the applications.

All servers are connected by a Juniper Networks QFabric Ethernet backbone (four 40Gb links to each top-of-rack switch) with two 10Gb links to each compute server. A separate 1Gb Ethernet network is used to support management and monitoring services. Shared data services are provided by GPFS-SAN for data sets not requiring large bandwidth, while GPFS-FPO (a cluster file system utilizing local drives) provides substantially higher, distributed bandwidth for larger datasets requiring parallel access.

Today, we can support a large number of projects with flexible, scalable runtime environments for discovery. Each project can experiment with configurations and software as needed, providing ultimate flexibility. Most of our projects run on Red Hat Linux for stability, but a few use Ubuntu or Fedora to gain access to particular features or because the analytics packages require it. We use IBM's Platform Cluster Management Advanced Edition

[1] E.g., SLAC: https://www6.slac.stanford.edu/ or CERN: http://home.web.cern.ch/about/computing

[2] For example: http://datascience.nyu.edu/ or http://vcresearch.berkeley.edu/datascience/overview-berkeley-institute-for-data-science or the joint Argonne and Univ. of Chicago Computation Institute: http://www.ci.anl.gov/data-computation

[3] E.g., Microsoft Azure, http://datamarket.azure.com/ or Amazon, https://aws.amazon.com/datasets

[4] The Center for Large-Scale Data System Research, http://clds.sdsc.edu

[5] https://amplab.cs.berkeley.edu/

[6] QFabric, from www.juniper.net

for basic Hadoop clusters, and leverage OpenStack for other application images. Over time we are standardizing images for our analytics projects, enabling us to relieve them of the demands of systems set-up and management, while allowing the systems projects to exploit the underlying hardware as needed.

4. SOFTWARE TO FOSTER DISCOVERY

The second piece of the discovery platform is software that fosters discovery. We focus on enabling two key elements of discovery: *insight* (the aha!) and *collaboration*. While no one can force insight, our software gives researchers new ways to look at a problem. The software presents contextual data and analytics to enrich core domain data and algorithms; it provides exposure to other researchers' ideas and work, aiming to spark new hypotheses. The analytics projects we support represent collaborations by individuals and teams, often spanning multiple domains of expertise. Our software lowers the barriers to cross-fertilization and supports collaboration across individuals and projects, creating the right conditions for insight and "strategic" serendipity. This section elaborates on these themes.

4.1 Contextual Data and Analytics

Contextual data and analytics can enrich core domain data and algorithms, providing new insights. For example, DNA samples from surfaces in a city such as turnstiles, public railings, and elevator buttons can be analyzed to identify what microbes are present at each location, but it is contextual data and analytics such as demographic data and traffic pattern computations that bring insight into patterns of microbes across neighborhoods, income level and populations. Contextual data and analytics can be used and reused across projects and in a variety of domains, and access to both are central to the mission of the Accelerated Discovery Lab. For example, data provided by government agencies such as the Census Bureau and the Bureaus of Labor Statistics and of Economic Analysis can provide location-specific population and income data across many domains. Other important contextual information includes worldwide patent data, medical journals, SEC filings and geo-spatial analytics packages such as those offered by Esri[7], one of our business partners.

Finding and preparing the right contextual data for a project are crucial to deriving insight, but are difficult tasks, particularly for non-technical users. Most data providers supply a simple hierarchical catalog of data sets organized by topic or category. Browsing the catalog of a large provider such as data.gov, with over a hundred thousand data sets, can be daunting, as users rarely know exactly what they are looking for. Once found, preparing data is a tedious process involving manual downloading and at least lightweight data modeling and transformation, skills that most non-technical users lack. For example, **a single** zip file from the Bureau of Economic Analysis containing National Income and Product (NIPA) data was found to contain nine spreadsheets, which can be transformed ultimately into **116** structured tables.

The Accelerated Discovery Lab provides *data lakes*[8] that can ingest data and analytics from a variety of sources, both open sources (e.g., data.gov) and third party providers, making both contextual and project-specific data available to our researchers (Figure 1). Data lakes store and catalog data, making it easy to track, govern, and repurpose, and ensure compliance with individual licensing terms and conditions. Specific projects may contribute data or analytics to a common lake (and combine them with contextual sources), but data or algorithms need not be shared if there are privacy or security concerns. Multiple data lakes are supported; this allows, e.g., aggregating data on particular themes. A project may also transform data and contribute the result back to a lake where it is cataloged and made available to others.

Data lakes include tools to facilitate and automate the data acquisition process, including tools to pull from standard publishing APIs such as Socrata[9] and ckan[10], and tools that analyze files such as the NIPA zip file. These tools derive and store structured tables and record provenance information about them, including the source and any additional metadata such as semantic tags, publishing organization, etc., that were captured as part of the analysis. Such metadata provides valuable governance and provenance information. Without governance, the use and re-use of data can lead to data management and legal challenges.

As shown on the right of the figure, a data lake provides a set of services to search for and provision data and analytics for use with multiple runtimes. These include direct access services, Extract, Transform, Load (ETL) platforms such as IBM InfoSphere Information Server[11], and multiple Hadoop distributions. Data lakes store the licensing terms and conditions associated with the data and analytics, automatically record use, and ensure compliance. Data lake services are provided via APIs and surfaced through a user experience that encourages discovery.

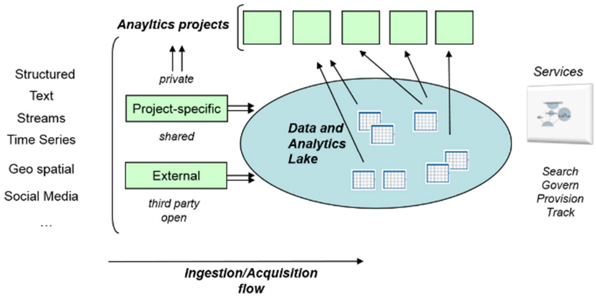


Figure 1. A data lake enables governed reuse of data sets.

4.2 A Socially-Inspired User Experience

While access to a rich set of data, analytic tools and runtimes are critical for a data analytics project, so are the experts who do the work. These may include domain experts, e.g., the microbiologists or meteorologists, as well as "general purpose" analysts, e.g., statisticians and experts in data mining. Our software, therefore, needs to support interactions across such multi-disciplinary teams. Each team member may use their own tools consistent with their area of expertise, but also need to exchange ideas with others in the group, or with other experts they consult.

We believe that the 'in-between' knowledge generated as these experts work together or separately can be critical to the discovery process, as it supports both insight and collaboration. LabBook, our user experience [11], places both data lake services and analytics work in a social context. Our system supports social actions such as following and tagging not only other users, but

[7] www.esri.com

[8] For simplicity, we will refer to a *data lake*, although in practice the data lake contains both data and analytics.

[9] http://www.socrata.com/

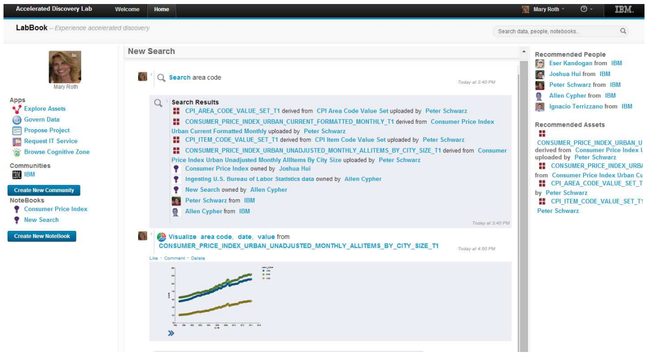
[10] The open source data portal platform, http://ckan.org/

[11] http://www-01.ibm.com/software/data/integration/info_server/

also data, analytics and associated metadata, such as the publisher of the data. In effect, we think of data, analytics and metadata along with users as first-class social entities. The user experience facilitates a meaningful conversation among these entities to guide discovery, suggesting additional or alternative pathways for new insights, and explicating provenance and process.

Accelerated Discovery Lab users have a home page with their profile and access to a set of applications and services appropriate for their role (Figure 2). Users interact with services in the context of a *notebook*, which captures and persistently stores the users' activities – e.g. data sets accessed, analytics run, and comments made. Notebooks can be private, shared, or made public, allowing one or more users to exchange ideas, knowledge and expertise to facilitate collaboration between team members and among a community of individuals with similar interests, with both the flow and dialog of the exchange captured in the notebook. Thus, notebooks themselves constitute collaborative metadata that capture relationships between individuals, data and applications. This allows the system to provide governance and track provenance of data and analytics assets used within the Accelerated Discovery Lab naturally and seamlessly.

Further, this collaborative metadata supports new models of exploration, which may lead to new insights. We have seen that a search for the right data or analytic often relies on a combination of clues from a user's social network, technical knowledge and semantic understanding of the problem. For example, a user might be looking for data from a particular government agency they had heard was used by a colleague of a friend on a different project. This is exactly the type of association captured in notebooks. To facilitate this more wandering notion of search, notebooks and the implicit and explicit relationships within them are captured as a large federated graph that connects social metadata with semantic tags and schema metadata such as data types. The graph enables powerful new contextual search and recommendation services. As shown in Figure 2, a search by the topic "area code" leads not only to data and analytics related to that topic, but to notebooks in which data related to the topic were referenced, as well as people who have worked with relevant data on other projects. Users may find previous explorations that might be related to their current project, giving them immediate insight into that discovery process by means of the captured dialog and tags, and allowing them to quickly find, use, or extend existing data or analytics from the previous exploration in the new project. The context of the new project is likewise saved and indexed, making it available for search and to serve as a reference for future projects.



Search area code

Today at 3:40 PM

Search Results

CPI_AREA_CODE_VALUE_SET_T1 derived from CPI Area Code Value Set uploaded by Peter Schwarz

CONSUMER_PRICE_INDEX_URBAN_CURRENT_FORMATTED_MONTHLY_T1 derived from Consumer Price Index Urban Current Formatted Monthly uploaded by Peter Schwarz

CPI_ITEM_CODE_VALUE_SET_T1 derived from CPI Item Code Value Set uploaded by Peter Schwarz

CONSUMER_PRICE_INDEX_URBAN_UNADJUSTED_MONTHLY_ALLITEMS_BY_CITY_SIZE_T1 derived from Consumer Price Index Urban Unadjusted Monthly AllItems By City Size uploaded by Peter Schwarz

Consumer Price Index owned by Joshua Hui

Ingesting U.S. Bureau of Labor Statistics data owned by Allen Cypher

New Search owned by Allen Cypher

Peter Schwarz from IBM

Allen Cypher from IBM

Today at 3:40 PM

Visualize area code, date, value from CONSUMER_PRICE_INDEX_URBAN_UNADJUSTED_MONTHLY_ALLITEMS_BY_CITY_SIZE_T1

Today at 4:50 PM

Eser Kandogan from IBM

Joshua Hui from IBM

Peter Schwarz from IBM

Allen Cypher from IBM

Ignacio Terrizzano from IBM

CONSUMER_PRICE_INDEX_URBAN_U derived from Consumer Price Index U uploaded by Peter Schwarz

CONSUMER_PRICE_INDEX_URBA from Consumer Price Index Urban Cu

CPI_AREA_CODE_VALUE_SET_T by Peter Schwarz

CPI_ITEM_CODE_VALUE_SET_T1 Peter Schwarz

Figure 2. The user experience fosters discovery by combining and leveraging social, technical and semantic metadata.

5. PARTNER RESEARCH PROJECTS

We started hosting our first partner research projects in late 2012, and have over a dozen projects underway. Here we describe a few of both the analytics and the systems projects.

5.1 Analytics Projects

We look for analytics projects for the Accelerated Discovery Lab that have cross-disciplinary interactions and the need for our infrastructure and tools. We try to balance diverse projects (to study discovery patterns across domains and technologies) with projects that have some commonality in either domain or tools (to explore how serendipitous interactions and cross-project collaboration may aid in the discovery process). Our goal is that the output of these projects will enrich the Lab's portfolios of data assets and analytic building blocks to further speed discovery for future projects, ensuring that the more the Lab is used, the better it gets. To illustrate how these considerations work, we look at examples of two subclasses of analytics projects.

5.1.1 Projects that leverage text analytics

Text analytics is one of IBM Research's strengths [3, 4], so many of our research partners are finding novel ways to exploit this technology in a range of fields, from medicine to finance to marketing. Here we sketch three representative projects, and discuss the assets they use and those they produce.

Knowledge Integration Toolkit (KnIT). The scientific literature is vast, and increasing exponentially [13]. In the field of cancer biology, over 70,000 papers have been written on a single critical tumor protein, p53 [7]. Since no researcher can read all of the relevant literature, most work from only a fraction of the available knowledge. A joint research team from Baylor College of Medicine and IBM Research is using text analytics, knowledge representation, and machine learning to build a model of what is known about p53, and then to suggest opportunities for experiments that could increase our knowledge. Such a tool could spark new discovery by oncology researchers. KnIT extends previous work [18] on a platform for chemical literature search, creating new information extraction and entity resolution rules for the biology domain, then mining the literature and analyzing the results. Since the rate of certain types of discovery in this field is known, we can measure the acceleration of discovery we achieve. For example, the rate of discovery of kinases that phosphorylate p53[12] has averaged one a year for the last decade; the KnIT team has already predicted several previously unknown p53 kinases, with two showing promise in experimental (wet lab) validation.

Waterfund. This project uses the same underlying text analytics in a different domain, finance, along with entity resolution and integration technology [2] and creates new text extraction rules, entity integration rules and data sets. IBM researchers worked with Waterfund[13] analysts to produce a financial index, the Water Cost Index, to track the cost of water in different geographies around the world. The goal is to encourage investments in water treatment facilities by giving a clearer view to lenders, insurers, and governments of the value, costs and associated risks of these projects. The information needed to understand these elements, however, is scattered across many different documents, including reports from public utilities, newspaper articles, and so on. The team defined a Normalized Production Cost Statement to compare

[12] Kinases are enzymes; phosphorylation adds phosphates to a protein (p53), changing the cell's behavior.

[13] http://Worldswaterfund.com

the costs of different agencies and populated these statements through scalable text analysis and integration of company filings. The index has been published regularly since Oct. 2013 and we expect the resulting index stream data to be of interest to other finance projects.

System U. Social media data can teach companies a lot about their clients [8] and their brand image [14]. The System U project is using analytics that derive an individual's personality portrait from his/her social media stream to help companies gain a deeper understanding of their customers, employees, and partners. Such people insights can then be used to help a company to optimize its business outcome, including delivering more individualized products and services to their customers by better matching between their brand and their patrons. With as few as 200 Twitter tweets, System U can derive a personality portrait of an individual or a "company" (as expressed through its social media posts) based on the words used and their frequency of use [6]. The researchers are working with multiple companies to understand if this type of analysis can improve business results. Leveraging the same entity resolution tool as described above, System U helps companies combine existing enterprise customer data (e.g., transaction records) with the personality portraits derived from social media to create enriched, actionable customer portraits.

This sampling of text analytics projects shows both their diversity and their underlying commonality. While these projects involve different researchers, in different domains, they have each benefited from and contributed to the expertise, data and assets available through the Accelerated Discovery Lab. As we enhance our discovery software, we expect to stimulate further interactions, and measure their impact on projects.

5.1.2 Prescriptive analytics projects

Another important subclass of projects leverages a combination of sensor and contextual data, and makes heavy use of machine learning and statistical packages. Here we highlight two such projects, reflecting on their commonalities and differences.

Equipment Condition Monitoring. A key challenge for the mining industry is equipment maintenance. Servicing equipment too soon costs millions in lost revenue; running it too long may result in damage that costs even more to repair. Today, most companies use time in service to decide when to pull a machine in for maintenance, since, in practice, it can be difficult to find good predictors of failure. Using DB2, SPSS, and R, the first phase of this project [9] analyzed data from monitoring 39 components of 50 mining haul trucks over six years, to build a predictive model of part failures, and to create a tool that provides an easy way for field foremen to see when a given machine needs service. The next phase of this project is looking at more data for more types of equipment, and at contextual data on terrain and weather.

Bioinformatics. Several projects have leveraged the Lab to develop or test new parallel algorithms for genome-wide association studies (GWAS). As an example, the BlueSNP R package [10] implements GWAS statistical tests in the R language. These calculations are then executed on Hadoop, using the MapReduce formalism. BlueSNP[14] makes computationally intensive analyses feasible for large genotype-phenotype datasets. The team is currently focusing on metagenomics, with a goal of creating a new system that will allow routinely testing many thousands of samples against thousands of reference genomes. This work may someday allow sequencing whole ecosystems, with potential to improve food safety and public health.

Although in unrelated domains, these projects benefit from common tools, such as R and Hadoop. Further, both projects are interested in adding weather and geo-spatial data as context to their analyses. They even have similar challenges in dealing with high-dimensional low sample size data in both fields. Hence there is potential for more synergies and interactions going forward.

5.2 Systems Projects

As with the analytics projects, our partner systems projects cover a broad range of topics. We rely on the work of some of these projects, for example, the scalable storage architecture (GPFS-FPO) that provides a robust alternative to HDFS, or the declarative machine learning platform that our analytics projects are starting to exploit. Other projects will likewise become part of our infrastructure as they mature. We give two examples here.

5.2.1 Benchmarking

A number of our systems partners use the environment for testing or benchmarking. For example, our IBM development colleagues used the Lab to benchmark and certify the performance of IBM BigInsights against that of Apache Hadoop. They used the Statistical Workload Injector for MapReduce (SWIM) developed by the University of California at Berkeley, and had their results certified by the Securities Technology Analysis Center (STAC). SWIM provides a large set of diverse MapReduce jobs based on production Hadoop traces obtained from Facebook, along with information to enable characterization of each job. The STAC report [16] showed that BigInsights completed the jobs four times faster than Apache Hadoop running on the same 18-node environment. BigInsights was about eleven times faster using the "sleep" test of scheduling speed. These types of experiments help us tune the configurations we offer the Lab's analytics users, by providing insight into what works best for particular workloads.

5.2.2 Novel Analytics Platforms

The Accelerated Discovery Lab provides a diverse set of applications that both inspire and test new analytic platform ideas. Some platforms, such as SystemT [3], become indispensable to a set of projects. These projects provide proof points for the technology and speed its adoption into products, and the products are then brought into the Lab to accelerate future analytics projects. In this way, the environment is continually improved by the projects within it.

One systems project being brought into the environment for use by our analytics projects is SystemML [5]. Expressing and running analytics for complex data at scale is challenging for mathematicians and systems researchers alike. SystemML raises the level of abstraction and lessens the burden of programming these algorithms by providing a declarative, high-level language using an R-like syntax extended with machine-learning-specific constructs. This language is compiled to a MapReduce runtime and automatically optimized to the specific data set and cluster configuration the analytics need to run against. We expect providing this system to the analytics projects in the Lab will drive further innovations and improvements to the technology.

6. DISCOVERY IN PRACTICE

As described in the previous sections, we have created an environment and a discovery platform that facilitate the exchange of ideas, technology sharing, and collaboration. Within the Accelerated Discovery Lab, we see an opportunity to better understand individual and team customs, actions, and processes

[14] Implementation: http://github.com/ibm-bioinformatics/bluesnp

(a.k.a., practices) in the context of a large-scale data-intensive discovery paradigm. Our discovery practices research focuses on addressing questions such as ‘can discovery be identified as it is being enacted or only in hindsight?’, ‘how is discovery organized over time?’, and ‘how do different ways of organizing work affect discovery?’ We examine discovery across the many domains of our partner projects, looking at how the practice varies, and across differently-constituted teams to see the role collaborations play. In short, we are investigating the human and social dimensions of what it means to accelerate the ability to ‘discover’.

Business and scientific efforts exist in a complex ecosystem composed of many relationships. Thus, we view the Accelerated Discovery Lab as a system-of-systems in which work and discovery is enabled and performed through an arrangement of technical, social, and spatial elements that form into identifiable patterns [12] that can be studied for the purpose of understanding, augmenting, enhancing, or hastening discovery. Each system brings with it a set of resources, whether the resource is an algorithmic tool or data set (technical), an expert or specialist (social), or a particular place (spatial). For example, our metagenomics project entails genomic data, spectral data analysis, bioinformaticians and a locale whence the microbiome is sampled. Captured and examined, these elements could lead to a better understanding of how discovery organically occurs.

Through both field and lab studies, we are investigating how these systems are configured by participants and teams over time and analyzing the key characteristics of the discovery process. Central to our research is the identification of system relationships, patterns of work and interaction, and typologies of discovery that will lead to a fuller understanding of how to represent, measure, and better enable discovery.

7. SUMMARY

In this paper we sketched the design and activities of the IBM Research Accelerated Discovery Lab. The lab is built on an analytics cloud environment with a unique software system that supports the process of discovery, facilitating collaboration and fostering insight. It facilitates a diversity of analytic and systems research projects that span disciplines and institutions. We are studying the practice of discovery, and using our findings to better enable and accelerate it.

8. ADDITIONAL AUTHORS

E. Kandogan, P. Schwarz, J. Hui, A. Cypher, I. Terrizzano, M. Bencala, L. Anderson, J. Vaughan, P. Selinger, R. Moore, O. Anya, P. Maglio, D. Pease, J. Janiak, J. Colino, G. Weber, M-T Schmidt.

9. ACKNOWLEDGMENTS

We thank our many research partners for allowing us to brag about their work. In particular, R. Krishnamurthy, S. Spangler, B. Reinwald and M. Zhou reviewed and improved our descriptions of their work. Any mistakes, however, are the authors’ own.

10. REFERENCES

[1] Akil, H. 2003. Scientific strategy in neuroscience: discovery science versus hypothesis-driven research. Message from the President, *Neuroscience Quarterly.* Society for Neuroscience. https://am2012.sfn.org/index.aspx?pagename=neuroscienceQuarterly_03summer_message&print=on

[2] Burdick, D., et al. Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. *IEEE Data Eng. Bull.* 34, 3 (2011), 60-67.

[3] Chiticariu, L., et al. SystemT: An Algebraic Approach to Declarative Information Extraction. *ACL 2010*, 128-137.

[4] Ferrucci, D., et al. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59-79.

[5] Ghoting, A. et al. SystemML: Declarative machine learning on MapReduce. *ICDE 2011* (April 2011), 231-242.

[6] Gou, L., Zhou, M. X., Yang, H. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. *CHI 2014* (April 2014), 955-964.

[7] Hager, K.M. and Gu, W. Understanding the non-canonical pathways involved in p53-mediated tumor suppression. *Carcinogenesis* 35, 4 (Apr 2014), 740-6. Epub (2013 Dec 31) DOI=http://dx.doi.org/10.1093/carcin/bgt487.

[8] Hernández, M.A, et al. Constructing consumer profiles from social media data. *BigData Conference* (2013), 710-716.

[9] Hochstein, A., Ahn, H., Leung, Y.T, Denesuk, M. Survival Analysis for HDLSS Data with Time Dependent Variables: Lessons from Predictive Maintenance at a Mining Service Provider, *IEEE SOLI* (July 2013).

[10] Huang H, Tata S, Prill R.J. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. *Bioinformatics.* 29, 1 (2013 Jan 1) 135-6. doi: 10.1093/bioinformatics/bts647. Epub 2012.

[11] Kandogan, E., et al. Data for All: A Systems Approach to Accelerate the Path from Data to Insight. *Big Data Congress, 2013 IEEE International Congress*, 427-428.

[12] Kieliszewski, C.A., Anderson, L.C., and Stucky, S.U. A case study: Designing the service experience for big data discovery. In *Proc. 5th Int’l Conference on Applied Human Factors and Ergonomics* (2014).

[13] Larsen, P.O. and Von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84, 3 (Sep 2010), 575-603. DOI= http://dx.doi.org/10.1007/s11192-010-0202-z.

[14] Spangler, S, et al.: COBRA – mining the web for COrporate Brand and Reputation Analysis. *Web Intelligence and Agent Systems* 7, 3 (2009), 243--254.

[15] Spangler, S., Wilkins, A. et al. Automated Hypothesis Generation Based on Mining Scientific Literature. To appear in *KDD 2014* (Aug 2014).

[16] STAC report: Comparison of IBM InfoSphere BigInsights Enterprise Edition with Adaptive MapReduce and Apache Hadoop, using Berkeley SWIM. (October 2013) http://stacresearch.com/node/15370.

[17] Swanson, D. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30, 1 (1986), 7-18.

[18] Yan, S., Spangler, S., Chen, Y.: Chemical Name Extraction Based on Automatic Training Data Generation and Rich Feature Set. *IEEE/ACM Trans. Comput. Biology Bioinform.* 10, 5 (2013), 1218-1233.