CIF21 DIBBs: Data Science for Finance (DSfin)
Louiqa Raschid University of Maryland

As the world of finance attempts to understand what brought our country's economy to the brink in 2008, and as federal regulators seek out new strategies to manage future crises, it is evident that *systemic tools and analytics* are required to model *systemic events*. The financial world is a closely interlinked Web of financial entities and networks, supply chains and financial ecosystems, where multiple financial entities may be counterparties to a complex financial contract. Financial analysts, regulators and academic researchers recognize that they must address the unprecedented and unfamiliar challenges of monitoring, integrating, and analyzing data *at scale*. The computing community has responded with proof of concept prototype solutions. However, for this effort to be successful, robust and extensive *community financial cyberinfrastructure* is required from which new ideas and solutions can be developed and evaluated.

Two workshops co-organized by Raschid, in 2010 and 2012, brought together a diverse community of academic researchers, regulators and practitioners. They articulated the range of multi-disciplinary challenges and highlighted the urgent need for community financial cyberinfrastructure. There is a compelling need to develop computational research frameworks, datasets, models, methods and tools, in the spirit of past efforts to identify computational grand challenges in the biomedical sciences, healthcare, climate change, etc. The next generation of community financial cyberinfrastructure must provide a platform that enables *data science for finance at scale*.

The **intellectual merits** of developing *DSfin community financial cyberinfrastructure* (pilot demonstration) include the following:

- Financial research will benefit from BIGDATA, Linked Data and social data resources. We will curate rich collections of public data and enhance them through information integration and entity resolution (record linkage) with additional resources.
- We will use ontologies and semantic Web technologies to annotate datasets, and to create Linked Data resources, with shared semantics that can be widely accessed.
- We will develop and customize a range of information management, data mining, machine learning, network analysis and visual analytics methodologies and tools, appropriate to financial collections and application challenges. This includes the managing and monitoring of financial ecosystems and supply chains, as well as new applications, e.g., determining the financial wellbeing of a community.
- The DSFin community financial cyberinfrastructure will include datasets, ontologies, tools, metrics, ground truth datasets, benchmarks and use cases, and a distributed analytical platform and testbed, to enable data science for finance *at scale*.

The **broader impacts** are significant. The increasing synergy from applying computational technology, BIGDATA and Linked Data, and social media to address difficult modeling and monitoring problems will result in improved tools for regulators. It will also lead to fundamentally new designs of financial products and ecosystems. This could include new ways to exploit the wisdom of the crowds to review and rate financial products, or new products such as a cost index for climate change mitigation activities. DSfin will also have a significant impact on transparency and the Open Government initiative. On the educational frontier, the planned DSFin cyberinfrastructure will nurture a new generation of multi-disciplinary scholars, at all levels who will blend computational solutions with theories, models and methodologies from finance, economics, mathematics and statistics. An Advisory Committee of researchers from finance, economics and mathematics, partners from the financial and data management industry, and a Management Committee of computational researchers, will provide guidance to the team.

Keywords: Financial cyberinfrastructure; financial information management; financial ontologies; knowledge representation; information integration; visual analytics; network analysis; human language technology; data science for finance.

1. Introduction

There is a compelling need to develop computational research frameworks, models and methods, in the spirit of computational grand challenges in data intensive domains, to monitor and manage financial supply chains and ecosystems. The next generation of community financial cyberinfrastructure – **Dsfin** - must provide a platform to support *data science for finance at scale* that can exploit BIGDATA, social data and Linked Data. The following *grand challenge* scenarios demand new datasets, tools and methods that can be utilized by regulators, analysts, investors and researchers:

- The ability to track financial products end-to-end along their supply chain. An obvious yet complex example is the mortgage supply chain, including individual mortgage products, the asset backed securities into which individual mortgages are pooled, and the complex derivatives that are used to hedge bets against the securities. One could integrate such data with non-financial datasets including maps, land-use and planning data, transportation and energy usage, crime statistics, etc., to create a rich and complex overlay of the financial health and wellness of our communities. Such extensions could lead to new products, e.g., a cost index for climate change mitigation activities.
- The ability to produce a "heat map" of our financial system transactions and accompanying economic activities, very much like a global weather map, so that one can identify financial weather patterns, pinpoint areas of high activity or vulnerabilities based on topology, warfare, political uncertainty, etc. A related endeavor is to determine the *market size* of various sectors or components of financial markets and exchanges, and to track the *flow of financial funds* through financial ecosystems.
- We must develop models of the global financial marketplaces and their interconnections, the multi-party network of legal entities (financial institutions) that participate in complex financial contracts, and contractual rules and financial events that dictate cash flows in these networks. Such models will provide the capability to run large-scale simulations to understand how these systems will perform under stress. We note that federal regulators made expensive policy decisions in 2008, about bailouts and stimulus spending, without real-time access to such models or simulation results.
- A significant amount of human activity is captured in new media – social media and social networks, as well as in traditional media – newswire, large document collections, etc. These resources can be a proxy for financial markets and can capture many aspects of human behavior including sentiment, intent, persuasion, etc. Such knowledge can be extracted and mined to create more sophisticated models of financial markets. We note that there have been many recent successes in combining human language technologies, machine learning and data/text mining, e.g., in computational social dynamics or socio-computing in the humanities and the social sciences.

1.1 The Rationale for Community Financial Cyberinfrastructure

Two workshops, co-organized by PI Raschid, in 2010 and 2012, brought together a diverse community of academic researchers, regulators and practitioners, from the following disciplines:

- Computer science and information science (data management and data mining; visual analytics; information retrieval; human language technologies; machine learning; knowledge representation and reasoning; semantic Web; BIGDATA).
- Finance (financial informatics, risk management, and financial engineering) and financial accounting.
- Mathematics, economics and operations research related to financial information modeling.

The consensus of the community was that there was a *significant deficit* in both datasets and computational and mathematical modeling and reasoning tools. There is a corresponding dearth of best practices for standards and ontologies, data sharing protocols, quality metrics, etc. Hence, all interested actors have been unable to ingest market information in a timely manner, and to determine what information might be missing. The development of new datasets and tools that exploit advanced computational methods has also been stymied.

The financial industry has historically been a leader in utilizing and driving advances in computational methods, and it is one of the largest consumers and producers of BIGDATA. Nevertheless, the industry does not have a history of making appropriate datasets available as community infrastructure for

research. A key reason is that information asymmetry is a critical advantage in a financial trade. This effectively discourages the practice of open data and transparency.

The Office of Financial Research (OFR) has a mandate under the Dodd-Frank Act of 2010 to collect all required data inputs for managing systemic risk. However, requirements to ensure the privacy and confidentiality of fully identified data, and the need to provide a continuous audit of secure access to the data, behind a firewall, naturally lead to constraints that limit the ability of the OFR to make the data widely available to the public. The DSfin community infrastructure development envisioned in this proposal are therefore a valuable complement to the data collection authority and activities of the OFR.

1.2 Broader Impact

Broader impacts of the planned **DSfin** community financial cyberinfrastructure include the following:

- The academic community will have access to resources required to examine and analyze actual market operations and behavior. Regulators, analysts, and the financial press will reach a better understanding of capital market operations to forge knowledgeable and prudent financial policy.
- A compelling technical outcome is that there will be increasing synergy from applying computational technology, BIGDATA and Linked Data, social networks and social media, to address difficult modeling and monitoring problems in financial ecosystems. This will result in improved tools for regulators to monitor financial systems as well as fundamentally new designs of market mechanisms, new ways to exploit the wisdom of the crowds to review and rate financial products, new financial products and cost indices, etc.
- The most significant social impact, however, may be with respect to Open Government initiatives that call for greater fiscal transparency. This will be in contrast to the status quo in the financial community where access to most financial data, e.g., data around the CUSIP that uniquely identifies all financial securities, often requires the payment of significant licensing or usage fees.
- Broader multi-disciplinary educational impacts will be discussed in Section 5.

1.3 Developing a DSfin Community of Expertise / Community of Practice

- Several working group meetings were held in conjunction with major conferences, e.g., SIGMOD in June 2013, the OFR/Federal Research Bank of St. Louis Conference on Systemic Risk in May 2013.
- Partnerships have been established with the Depository Trust and Clearing Corporation (DTCC), the World Federation of Exchanges (WFE), the Municipal Securities Rulemaking Board (MSRB) and the Financial Industry Regulatory Authority (FINRA).
- IBM Research is the leader in extracting entities, relationships and events from public financial collections using the Midas framework (Burdick et al. 2011, 2014b). A project has been launched, in partnership with collaborators at the Haas Real Estate and Financial Markets (REFM) Lab, to extend Midas to include complex financial contracts, e.g., mortgage-backed securities.
- Raschid, Jagadish, Langsam and Mark Flood (Office of Financial Research) have developed Contract Aggregation Framework (CAF) to aggregate trade data from multiple streams (Ball et al. 2014).
- Raschid and the Lanka Software Foundation are developing the Karsha Pariksha Search Engine. Pariksha can annotate contracts using terms from the Financial Industry Business Ontology (FIBO).
- Co-PI Wellman has developed a robust simulation testbed for studying strategic financial behavior.

PI Raschid and co-PIs and collaborators are organizing the 2014 Sigmod Workshop on Data Science for Macro Modeling with Financial and Economic Datasets. http://dsmm2014.org

2. Vision and Architecture - Data Science for Finance (DSfin) Pilot Demonstration

While there is a long tradition of data driven econometric and statistical research in finance, there is a compelling need today to analyze complex financial ecosystems and networks. Such analysis is essential to monitoring systemic risk and designing interventions to manage the next financial crisis. However, such analytics requires dealing with multiple heterogeneous streams of data, each of which can be high in volume and velocity, i.e., we have a classic BIGDATA challenge. While public financial data is available, a researcher who wishes to conduct data-driven systemic research or data science for finance research, first has to build up a substantial infrastructure. They have to process multiple heterogeneous data

streams, extract relevant information, clean it, integrate information from distinct streams, perform entity resolution, and aggregate data before they can even begin their analysis. Doing all of this creates a high barrier, in particular for financial data science at scale. The goal of *DSfin (DIBBs pilot demonstration)* is to help researchers get past this barrier by building much of the financial dataset cyberinfrastructure for them, and by providing a platform for financial data science analytics at scale. To the extent that there is a large community of researchers with similar needs, we can design the DSfin cyberinfrastructure to meet a diversity of needs. Our curated financial datasets will build upon publicly available collections that can be made available to all researchers (under a suitable DSfin open source license). DSfin will include datasets, ontologies, tools, metrics, ground truth datasets, benchmarks and use cases, and a distributed analytical platform and testbed, to enable data science for finance research at scale.

- We will curate rich collections of public data and enhance them through information integration and entity resolution (record linkage) with additional resources (Burdick et al. 2011, 2014a, 2014b).
- We will use the Contract Aggregation Framework (CAF) (Ball et al. 2014), ontologies, semantic Web technologies and the W3C RDF Data Cube Vocabulary, to annotate datasets and to create Linked Data resources with shared semantics.
- We will develop and customize a range of information management, data mining, machine learning, network analysis and visual analytics methodologies and tools.
- DSfin will sponsor the DSfin Open Data challenge for large scale entity resolution and information integration against reference datasets, e.g., the Legal Entity Idenifier (LEI) (Federal Register 2012).

Our vision for the DSfin Pilot Demonstration includes datasets and a distributed analytics infrastructure to be jointly developed by the University of Maryland, the University of Michigan, IBM Research and UC Berkeley. *All datasets and enriched Linked Data artifacts will be created from public datasets, will be ported to the University of Maryland Dsfin platform, and will be made available to the research community*. The University of Maryland will maintain the DSfin platform through the expected lifetime of such data and infrastructure and software resources. It is expected that as the DSfin community of researchers matures, they will maintain and enhance these collections and tools.

Infrastructure objectives:

- Build an infrastructure for creating, disseminating and preserving the DSfin datasets to the data science for finance community.
- Build out a DSfin virtual server infrastructure that can support a range of analytical tasks, graph analysis, web services, data distribution, visual analytics, visualization, etc.
- Develop a high performance infrastructure to support a variety of large scale computing across extensive structured and unstructured collections.
- Develop an Infrastructure as a Service environment for the community of researchers to create, run, and share analytical tools.

Administration objectives:

- User support for each component.
- Assessing user satisfaction and providing feedback.
- Developing an entrance and exit strategy for each component - analytical tool or dataset – for the DSfin computing stack.

Metadata and dissemination objectives:

- Utilities include software and data repositories.
- Metadata including RDF schemas and ontologies.
- Wikis, repositories, database and Web servers.

2.1 DSfin Datasets

Financial data for systemic risk management can be classified as follows:

- Financial instrument reference data: Information on the legal and contractual structure of financial instruments such as prospectuses or master agreements, including data about the issuing legal entity and its adjustments based on corporate actions.

- Legal entity reference data: Identifying and descriptive information such as legal names and charter types, for financial entities that participate in financial transactions, or that are otherwise referenced in financial instruments.
- Positions and transactions data: Terms and conditions for new contracts (transactions) and the accumulated financial exposure on an entity's books (positions).
- Prices and related data: Transaction prices and data used in the valuation of positions, development of models and scenarios, and the measurement of micro-prudential and macro-prudential exposures.

The vision for developing community financial datasets will explore multiple approaches to accommodate a diversity of requirements. Consider the following approaches:

- Start with a *seed collection* of highly curated data objects, and exploit public or private collections, utilizing text extraction and human language technologies, to enhance and enrich the seed dataset.
- Apply *scalable methods* from network analysis, machine learning, information retrieval, semantic Web, Linked Data, etc., to create large interlinked and annotated collections, with varying levels of completeness and quality.
- There is also a significant need to apply *knowledge representation and reasoning* methods to financial contracts so yet another approach will rely on combining methods for machine readable contracts, formal logics and reasoning, etc.
- The Contract Aggregation Framework (CAF) provides an alternative approach based on a scalable and extensible framework to integrate multiple financial datastreams at various levels of temporal granularity, contract granularity, etc. (Ball et al. 2014).

Partnerships have been (are being) established with the Depository Trust and Clearing Corporation (DTCC), the World Federation of Exchanges (WFE), the Municipal Securities Rulemaking Board (MSRB) and the Financial Industry Regulatory Authority (FINRA). They curate the following publicly available resources:

- The DTCC provides clearing, settlement and information services for equities, corporate and municipal bonds, government and mortgage-backed securities, money market instruments and over-the-counter derivatives. They support the publicly available Global Trade Repository (GTR).
- The MRSB provides the state-of-the art public EMMA portal to access a wealth of multi-modal financial data about municipal securities.
- The WFE provides historical summary statistics for over 100 exchanges worldwide across a range of financial products.
- FINRA operates the Market Data Center and supports TRACE. The Market Data Center includes detailed market data on equities, options, mutual funds, and a range of bonds including corporate, municipal, Treasury and agency bonds. TRACE (Trade Reporting and Compliance Engine) facilitates the reporting of trades in fixed income securities.

A selection of these datasets will be extracted, cleaned and curated, and further enhanced through information integration and entity resolution. Where relevant they will be annotated with FIBO terms, marked up using the W3C RDF Linked Data Vocabulary, and made available as a SPARQL Linked Data endpoint. Details of specific resources and tasks as well as the resulting public Linked Data outputs are discussed in the next section.

2.2 University of Maryland Data Science for Finance (DSfin) Platform

The UMD DSfin Platform will be built on the following three subsystems:

- A private cloud built on virtualization servers with a total of 128 cores and one Terabyte of memory configured as an Infrastructure as a Service (IaaS) built on OpenStack and Linux Kernel-based Virtual Machines.
- An Object Store built on Ceph with 144 Terabytes of storage for the proposal's data collections that will be accessible over the web, through a restful application programming interface, and as a block service for virtual servers.
- A Hadoop cluster for document analysis and for processing unstructured data.

All of the systems will be connected over a Gigabit Ethernet and connected to the UMIACS and the University of Maryland networks over a 10-Gigabit uplink. In general terms, these subsystems address

the DSF's long-term needs for storing and distributing data collections, delivering online services to the user community, and processing large collections of documents, unstructured and semi-structured data. In particular, the IaaS and Object Store will support the following:

- Discover and access resources at Michigan, Berkeley and IBM.
- Annotate and reformat data collections.
- Allocate virtual server platforms to support a range of analytical tasks, graph analysis, web services, data distribution, visualization
- Curate, preserve, and distribute shared data collections as file sets or relational databases.

The goal is to provide a flexible environment for the community of researchers to preserve, run, and share the data collections and analytical tools that they develop. Porting Michigan's high performance simulation platform and UC Berkeley's MLBase tools to the UMD DSfin stack would be the pilot test.

2.3 University of Michigan Simulation Testbed for Strategic Financial Behavior

The strategic modeling testbed is based on computational infrastructure developed at the University of Michigan for managing large-scale agent-based simulation experiments (Cassell & Wellman, 2013). The infrastructure is designed to make it simple for modelers to run large experiments covering a wide range of environment parameters and agent behaviors, by insulating end-users from the details of scheduling computational jobs and organizing the volumes of data produced. It provides a front-end for specifying patterns of strategy and simulation parameter combinations, and a back-end for managing jobs on a computing cluster and the resulting simulation data. Through web-based interfaces and APIs, external users can specify and operate their own strategic simulation experiments.

An example experiment is high-frequency trading; the testbed can be used to conduct studies on the implications of high-frequency trading strategies, as well as proposed regulatory and market operation measures designed to ameliorate high-frequency effects. The models will accommodate a broad range of behaviors in the trading ecosystem, including naive investors, high-volume institutional investors, market makers, statistical arbitrageurs, and high-frequency strategies of various kinds.

Significant improvements to the current infrastructure are planned to conduct large-scale simulations on high-performance computing clusters, and management of data for empirical game-theoretic reasoning. This testbed infrastructure will also be generalized to work with a variety of backend architectures, such as commercial IaaS platforms (e.g., Amazon EC2) and academic computing facilities such as the planned DSfin infrastructure.

2.4 UC Berkeley - Algorithms, Machines and People Lab (AMPLab) and Haas Real Estate and Financial Markets (REFM) Lab [unfunded collaboration]

The AMPLab is addressing the challenges of emerging Big Data applications by developing the Berkeley Data Analytics Stack (BDAS). BDAS is an Open Source software stack that includes a suite of technologies for making sense at scale. These include the highly-popular Spark in-memory computation framework and Shark query processing system, both of which have seen rapid adoption in both industry and research environments (Xin et al. 2013). Future components of BDAS include the BlinkDB approximate query engine (Agarwal et al. 2013), the GraphX scalable graph processing system and the MLBase declarative machine learning environment (Kraska et al. 2013), all of which can play key roles in the analysis of large-scale financial and risk data. The goals of the DSfin infrastructure are well aligned with those of the BDAS effort in terms of requirements for functionality, performance, scalability, ease of use and ultimately portability to the DSfin stack. We will explore the applicability of the various components of the BDAS system for addressing financial analytics challenges around the Contract Aggregation Framework (CAF) and we will use the results of these studies to propose new directions and capabilities for BDAS software development.

A central focus of the Haas REFM Lab has been to assemble a database that uniquely integrates time-series panels of micro-data across housing, employment, energy usage and mortgage markets. REFM researchers are collaborating with IBM to use the Midas framework and tools to extract entities,

relationships, events, contractual rules and risk profiles from prospectus documents for mortgage backed securities and other real estate securities.

2.5 IBM - Accelerated Discovery Lab

Accelerated Discovery Lab is a shared Big Data analytics environment in which IBM researchers, external researchers and other practitioners collaborate on joint research projects to push on today's boundaries of the 4 V's – volume, velocity, variety, and veracity. This environment represents tens of millions of dollars of investment (to-date) by IBM and is designed to give users the system functionality and support for the data wrangling and analytics activities necessary for conducting these research projects. The environment includes the following:

- Rich collection of analytics and tools for the various analytic stages (e.g., text analytics, entity resolution and integration, and scalable machine learning).
- Curated and enhanced public datasets (to be described in later sections).
- Secure access to private data sets.
- Systemic solutions for BIGDATA data management and analytics, providing the necessary platform for research.
- Access to a physical collaboration space and a virtual space that supports active and on-going interaction between members of an existing project, as well as between members of other projects who determine that they would benefit from knowing one another.
- Access to worldwide IBM research experts representing a wide variety of disciplines.

In addition to the shared resource and assets that make up this Lab, the Accelerated Discovery Lab provides the following user centered capabilities:

- Creation of the user desktop content, provisioning, and help desk support.
- Development of a user experience interface supporting a *serendipitous discovery* process.
- Monitoring the security of the environment via the implementation of IBM's IT security policies and standards. This includes the deployment of capabilities such as VPN, secure access, and encryption which supports IBM researchers and external researchers working together securely within the same environment.

2.6 Ontologies/Metrics/Metadata

- The Financial Industry Business Ontology (FIBO) includes a semantic model of concepts, their relationships and abstractions, as well as an operational ontology that targets pragmatic operational implementations. For example, using a semantic reasoner, representations in W3C RDF OWL and the FIBO, one can implement an end-to-end application to extract data from a spreadsheet and to classify undifferentiated financial swaps into their real asset classes. The Karsha Pariksha Search Engine will annotate contracts using FIBO terms.
- Many financial datasets are time series data or event streams. The best known example comes from the Center for Research in Security Prices (CRSP) which maintains a comprehensive collection of security price, return, and volume data for US exchanges. The data is available through WRDS for a licensing fee. Another example is the set of historical reports provided by the *WFE* for over a hundred exchanges, for a range of products. Yet others are the DTCC's GTR and FINRA TRACE. Such datasets can be aggregated into multi-dimensional tensors using the Contract Aggregation Framework (CAF). The W3C RDF Data Cube Vocabulary will be used to develop appropriate schemas for these datasets. The Lanka Software Foundation will provide SPARQL Linked Data endpoints for a variety of resources, in consultation with the original data providers.

2.7 DSfin Education, Training and Mentoring

This project will nurture a new generation of multi-disciplinary student scholars, at all levels; they will blend computational solutions with theories, models and methodologies from finance, economics, mathematics and statistics. The following activities are planned:

- +3 months: Initial meeting of the PIs and key individuals of the Steering and Advisory Committees.
- +6 and +12 months: An intensive (virtual) training on the following tools and datasets: (1) The IBM Accelerated Discovery Lab; (2) Maryland Simulation Platform; (3) UC Berkeley BDAS stack, in particular the declarative MLBase tool (Kraska et al. 2013). These training will target a group of

doctoral students and undergraduate students who are expected to jumpstart the grand challenge research activities of the DSfin community.

- +18 months: First All Hands Workshop of all PIs, doctoral students, members of the steering and advisory committees and selected researchers. Objectives include the following:
    - Assess progress of the first 18 months of the grant.
    - Evaluation of tools and datasets.
    - Identify strategies to disseminate community infrastructure.
    - Identify teams to explore DSfin Grand Challenge problems including the entity resolution at scale challenge.
- +30 months: Second All Hands Workshop.
    - Evaluation of tools and datasets.
    - Report on DSfin Grand Challenge problems.
    - Approve plan to handover all datasets and tools to the UMD DSfin platform.
    - Identify strategies to disseminate community infrastructure and research results.
    - Develop a long term maintenance plan beyond the end of the grant.

2.8 Access and Dissemination/Long term Maintenance/User Evaluation

Dissemination: There are several examples of community infrastructure, portals, model organism databases, etc., that have been sponsored by the NSF and the NIH. Exemplars include the UCI Machine Learning Repository (Frank and Asuncion) and WormBase (Harris et al 2010). We will follow best practices to identify a plan for access and dissemination, and data management best practices and protocols. Every effort will be made to use open standards and protocols.
UMD Senior Investigator Oard has prior experience with managing the NIST TREC Legal Track and Raschid and Oard will explore a similar *Data Science for Finance* track.
A variety of visual analytics and visualization tools will be evaluated including VisTrails (http://www.vistrails.org) as well as tools developed by Shneiderman (Management Committee).

Maintenance: Funds for developing DSfin infrastructure is being requested from the NSF DIBBs program (this proposal). Industry sponsors including IBM will provide a range of enriched public datasets, computational resources and expertise. The PI and Advisory Committee are reaching out to potential partners including the DTCC and FINRA to establish partnerships for the curation and dissemination of additional public datasets. Over the long term, it is expected that individual researchers and collaborative teams will incorporate these resources into their research agenda, thus providing a path for long-term support and maintenance of the *data science for finance* community financial cyberinfrastructure, as well as ensuring the further development, dissemination and use of tools, metrics and use cases.

User Evaluation: DSfin will sponsor the Financial Entity Resolution At Scale challenge for entity resolution against reference datasets, e.g., the Legal Entity Idenifier (LEI) (Federal Register 2012). A series of CLOSED and OPEN evaluation challenges will be identified as follows:

- CLOSED: Resolve all entity instances between two reference datasets, e.g., the LEI and the Central Index Key from the SEC.
- OPEN: Resolve all *mentions of organizations, contract, and financial events* from SEC filings.

User satisfaction: Both proactive and reactive actions will be taken to obtain feedback and to ensure an adequate level of user satisfaction.

- A 2-level triage process is planned where user tickets are logged in at he UMD portal. The UMD DSfin Project manager will forward tickets where appropriate to the IBM, Berkeley or Michigan teams. The UMD DSfin team will monitor forwarded tickets to ensure suitable resolution of these incidents.
- Multiple interactions and trainings are planned with users. Each of these interactions will be complemented with a significant evaluation of the tools and datasets.
- All teams that participate in DSfin Grand Challenge activities will be managed by a PI or co-PIs, and one of the graduate assistants at Maryland and Michigan who will be supported by the grant.

3. DSfin Team

Louiqa Raschid	University of Maryland	PI
Amol Deshpande	University of Maryland	co-PI
Douglas W. Oard	University of Maryland	Senior Investigator
H.V. Jagadish	University of Michigan	co-PI
Michael Wellman	University of Michigan	co-PI
Michael Franklin	UC Berkeley	UNPAID Collaborator
Nancy Wallace	UC Berkeley	UNPAID Collaborator
Rajasekar Krishnamurthy	IBM Research	IBM Team Lead

An **Advisory Committee** of finance, economics and mathematics researchers has been assembled to provide domain expertise. Members of the Financial Research Advisory Council being formed by the OFR will be invited to join the advisory committee to provide an additional avenue for alignment of data and tool related strategic objectives of the planned community infrastructure.

A **Management Committee** of computer science researchers will assist the team. **Industry partners** have also been identified to participate in the infrastructure development. Given the large number of potential participants in the planning process, a selected subset of letters of interest have been included.

Advisory Committee: Strategic oversight and domain knowledge

- Lewis Alexander, Chief U.S. Economist, Nomura. Formerly Counselor to the Secretary of the Treasury.
- Richard Anderson, Economist, Federal Reserve Bank of St. Louis.
- John Yelle and Ronald Jordan, DTCC.
- Albert "Pete" Kyle, Charles E. Smith Professor of Finance at the Smith School, University of Maryland.
- Andrew Lo, Charles E. and Susan T. Harris Professor, Sloan School of Management, MIT.
- David Newman, Vice President for Enterprise Architecture, Wells Fargo.
- Nitish Sinha, Federal Research Board.
- Chester Spatt, former Chief Economist, SEC; Pamela, R. and Kenneth, B. Dunn Professor, CMU.
- Jonathan Sokobin, Chief Economist, FINRA.

Management Committee: Technical oversight

Mike Atkin	Enterprise Data Mgmt. Council	Enterprise Information Systems; Data maturity.
Michael Bennett	Enterprise Data Mgmt. Council	Semantic technology.
Andrea Cali	University College of London	KR; formal reasoning.
Sanjiv Das	Santa Clara University	Financial datasets; financial analytics.
Juliana Freire	NYU	Data management; provenance.
Johannes Gehrke	Cornell	Data management.
Lise Getoor	UC Santa Cruz	Machine learning.
Georg Gottlob	Oxford University	KR; formal reasoning.
Gerard Hoberg	University of Maryland	Financial datasets; financial analytics.
Eduard Hovy	CMU	Human language technologies
Vagelis Hristidis	UC Riverside	Data management; social media.
Joe Langsam	formerly from Morgan Stanley; Center for Financial Policy, UMD	Financial datasets; financial analytics.
Brad Malin	Vanderbilt University	Bioinformatics; privacy.
Philip Resnik	University of Maryland	Human language technologies.
Ben Shneiderman	University of Maryland	Visual analytics.

Industry Partners

IBM Research Almaden (UNPAID Collaborators: Laura Haas, Rajasekar Krishnamurthy and Shiv Vaithyanathan).
Enterprise Data Management Council (UNPAID collaborators: Mike Atkin and Michael Bennett).
Wells Fargo (UNPAID Collaborators: David Newman)
DTCC (UNPAID Collaborators: John Yelle and Ronald Jordan)

4. DSfin Activities

We briefly describe several ongoing activities to extract knowledge from public datasets and to develop a simulation platform. A complete list of relevant research challenges are discussed in the report of the 2012 NSF Workshop (Jagadish et al. 2012).

4.1 Knowledge Extraction from Public Data (IBM, UC Berkeley and University of Maryland)

There is a significant amount of publicly accessible unstructured collections describing the financial performance and counterparty relationships of private and public companies, governments, and public utilities. However, this wealth of structured entity and relationship information is buried inside a large amount of unstructured data. Converting this unstructured data into a structured format is the focus of the Midas project at IBM Research which addresses this problem by creating comprehensive views of publicly traded companies and related entities (Burdick et. al., 2011, 2014a, 2014b). A major step towards providing such insights is the aggregation of fine-grained data or facts from hundreds of thousands of documents into a set of clean, unified entities (e.g., companies, key people, loans, securities) and their relationships. The Midas project starts from a document-centric archive, as provided by the SEC and FDIC, and build a concept-centric repository for the financial domain that enables sophisticated structured analysis. By focusing on high-quality financial data sources and by combining three complementary technology components – information extraction, information integration, and scalable infrastructure – Midas can provide valuable insights about financial institutions either at the whole system level (i.e., systemic analysis) or at the individual company level.

Figure 1 describes the Midas common analytics platform. This platform supports a framework of entities and relationships. The bottom pipeline creates structured entities and relationships from a given collection of unstructured documents.

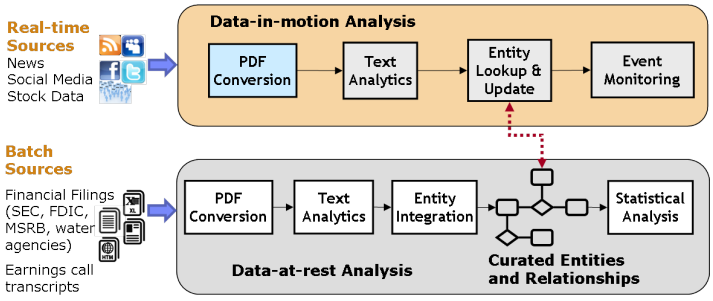


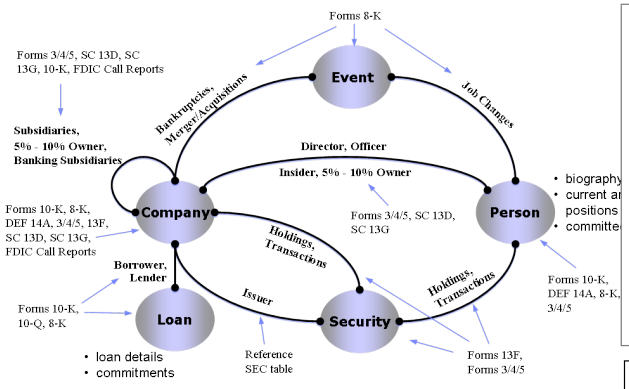
Figure 1: Midas Information Extraction Platform

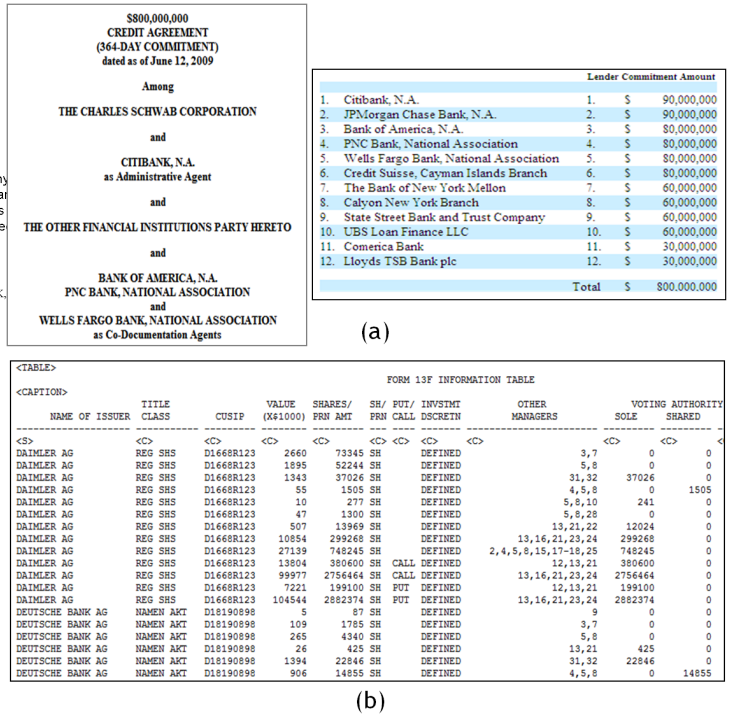
The technology components are as follows: (1) PDF to HTML conversion (2) information extraction (with particular emphasis on extraction from tables); (3) entity integration (including temporal fusion and reconciling of inconsistent information); (4) statistical analysis. The statistical analysis phase enriches the entity data by exploiting machine learning techniques. Such enrichment includes interpolating missing values or predicting future data values based on the existing data. Further, the entity data can be used as a context for data-in motion analysis shown in the upper pipeline. Here information from incoming documents, e.g., newsfeeds, social media messages, etc., are extracted and linked to the corresponding entities. As a result, the system is able to identify significant events that are relevant to the curated entities. For example, a credit rating change in a Bloomberg news article can be linked to the municipality entity to which it applies.

Exemplar Dataset: Counterparty Relationships from Regulatory Filings
The dataset comprises of the following:

- *Raw filings made by publicly traded financial companies to the SEC and banking subsidiaries to FDIC over a period of multiple years.* The SEC filings include multiple filing types, e.g., annual reports, current reports, proxy statements, insider filings, beneficial ownership and institutional holdings reports, etc. The FDIC filings are the Reports of Condition and Income (call reports).
- *Comprehensive counterparty relationship dataset consisting of 2000+ publicly trading financial company profiles and 30K+ related people. T*he relationships across the companies and / or people include

investment relationships (institutional holding, 5% beneficial ownership), ownership relationships (subsidiaries, banking subsidiaries), insider relationships (officer / director employment history, insider transactions and holdings) and lending relationships (co-lending activity).





$800,000,000
CREDIT AGREEMENT
(364-DAY COMMITMENT)
dated as of June 12, 2009

Among

THE CHARLES SCHWAB CORPORATION

and

CITIBANK, N.A.
as Administrative Agent

and

THE OTHER FINANCIAL INSTITUTIONS PARTY HERETO

and

BANK OF AMERICA, N.A.
PNC BANK, NATIONAL ASSOCIATION
and
WELLS FARGO BANK, NATIONAL ASSOCIATION
as Co-Documentation Agents

		Lender Commitment Amount
1.	Citibank, N.A.	$ 90,000,000
2.	JPMorgan Chase Bank, N.A.	$ 90,000,000
3.	Bank of America, N.A.	$ 80,000,000
4.	PNC Bank, National Association	$ 80,000,000
5.	Wells Fargo Bank, National Association	$ 80,000,000
6.	Credit Suisse, Cayman Islands Branch	$ 80,000,000
7.	The Bank of New York Mellon	$ 60,000,000
8.	Calyon New York Branch	$ 60,000,000
9.	State Street Bank and Trust Company	$ 60,000,000
10.	UBS Loan Finance LLC	$ 60,000,000
11.	Comerica Bank	$ 30,000,000
12.	Lloyds TSB Bank plc	$ 30,000,000
	Total	$ $00.000.000

(a)

<TABLE>
<CAPTION>
FORM 13F INFORMATION TABLE

NAME OF ISSUER	TITLE CLASS	CUSIP	VALUE (X$1000)	SHARES/ PRN AMT	SH/ PRN	PUT/ CALL	INVSTMT DSCRETN	OTHER MANAGERS	VOTING AUTHORITY SOLE	VOTING AUTHORITY SHARED
<S>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>	<C>
DAIMLER AG	REG SHS	D1668R123	2660	73345	SH		DEFINED	3,7	0	0
DAIMLER AG	REG SHS	D1668R123	1895	52244	SH		DEFINED	5,8	0	0
DAIMLER AG	REG SHS	D1668R123	1343	37026	SH		DEFINED	31,32	37026	0
DAIMLER AG	REG SHS	D1668R123	55	1505	SH		DEFINED	4,5,8	0	1505
DAIMLER AG	REG SHS	D1668R123	10	277	SH		DEFINED	5,8,10	241	0
DAIMLER AG	REG SHS	D1668R123	47	1300	SH		DEFINED	5,8,28	0	0
DAIMLER AG	REG SHS	D1668R123	507	13969	SH		DEFINED	13,21,22	12024	0
DAIMLER AG	REG SHS	D1668R123	10854	299268	SH		DEFINED	13,16,21,23,24	299268	0
DAIMLER AG	REG SHS	D1668R123	27139	748245	SH		DEFINED	2,4,5,8,15,17-18,25	748245	0
DAIMLER AG	REG SHS	D1668R123	13804	380600	SH	CALL	DEFINED	12,13,21	380600	0
DAIMLER AG	REG SHS	D1668R123	99977	2756464	SH	CALL	DEFINED	13,16,21,23,24	2756464	0
DAIMLER AG	REG SHS	D1668R123	7221	199100	SH	PUT	DEFINED	12,13,21	199100	0
DAIMLER AG	REG SHS	D1668R123	104544	2882374	SH	PUT	DEFINED	13,16,21,23,24	2882374	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	5	87	SH		DEFINED	9	0	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	109	1785	SH		DEFINED	3,7	0	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	265	4340	SH		DEFINED	5,8	0	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	26	425	SH		DEFINED	13,21	425	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	1394	22846	SH		DEFINED	31,32	22846	0
DEUTSCHE BANK AG	NAMEN AKT	D18190898	906	14855	SH		DEFINED	4,5,8	0	14855

(b)

Figure 2 (above) Entities and Relationships from SEC Filings.
Figure 3 (right) Example Snippets from SEC Filings; (a) Loan agreement; (b) Form 13F.

Figure 2 gives a schematic description of the entities and relationships for the counterparty risk analysis application, together with the types of forms (from SEC/FDIC) from which they are extracted and integrated. The schema described in the figure is populated with over 3000 financial institutions for the Company entity type, and over 30,000 instances for the Person entity type that are related to (i.e., they are officers, directors or insiders of) the aforementioned financial institutions. Midas also populates the Event, Loan and Security instances that are related to the above financial institutions; these are all of potential interest to regulators and counterparty risk analysts.

As seen in the figure, there are many different types of relationships among the entities in the schema, and these are populated from the multitude of available filings. Important relationships include the following: borrower/lender relationships among companies (via loans); subsidiaries (of companies); investment holdings between companies and securities (which in turn are issued by other companies); material events such as merger, acquisitions, or bankruptcies of companies. Although not explicit in the above schematic diagram, time is generally another important dimension; an instance of an entity attribute or an instance of a relationship among entities is typically valid during a particular time interval, which can be derived only by analyzing the historic collection of filings.

The curated entity and relationship instances can be queried, visualized and aggregated in various ways to convey the insights that are important for the particular user. For example, by analyzing the co-lending relationships among the major banks, a regulator is able to identify (purely based on public data) which banks, by being most central to the global financial system, pose risk to the financial system. A counterparty risk analyst would be able to drill-down into a particular financial institution, e.g., Citigroup, and visualize the entities (i.e., counterparties) that, by virtue of being immediately related via Citigroup, may have a direct impact on the health of the given financial institution.

The public data archived by the SEC can be categorized under many types of forms, and each form type can be seen as a data source with its own characteristics and challenges. Data integration from the SEC is in fact a complex example of heterogeneous multi-source data integration. Figure 3 gives an illustration of just two such types of forms. Figure 3(a) is a loan agreement in HTML containing a mix of text and tables, describing a syndicated loan made to Charles Schwab by a set of twelve global banks, each contributing a specific amount. Figure 3(b) illustrates a Form 13F filed by Citigroup, also in HTML and containing a large table with details about the securities that are held by Citigroup (via its related fund managers or subsidiaries). For each security, there are multiple lines that detail the type of security, the identifier (CUSIP), the notional value (amount), the type of security, e.g., a Put or Call option or a non-derivative option, the voting power, etc.

Exemplar Dataset: Information Extraction from a Prospectus of a Mortgage-Backed Security

A central focus of the Haas Real Estate and Financial Markets (REFM) Lab has been to assemble a database that uniquely integrates very large and expensive time-series panels of micro-data (currently about seven terabytes of data) for the housing, employment, and mortgage markets. A key piece of the puzzle of understanding the stability and risk of real estate markers requires a detailed understanding of the complex relationships within organizations, based on their counterparty relationships in complex contracts. We will use the IBM Midas framework and tools to extract entities, relationships, events, contractual rules and risk profiles for financial institutions.

Our source of information will be the MBS prospectus documents that are public and are filed with the Securities and Exchange Commission. The *MBS* (title), the *Issuing Entity*, the *Depositor*, the *Sponsor*, the *Originator*, and various *Service Providers* are included in the summary. Next, we consider the structure of the MBS, labeled the *Trust*. It describes the *Mortgage Pools* that provide the revenue and the tranches or *Class Certificates* for payment including the cash flows for each tranche. This state contingent cash flow has often been referred to as an *MBS-Waterfall* structure.

The first challenge is the extraction of entities and their roles and relationships in the MBS, as follows:

- Creation and curation of an *MBS+LEI* graph dataset to represent the financial contract (MBS) and the participants to these contracts, each of which will be identified using a legal entity identifier (LEI). Entities will be identified with respect to their role.
- The participants include the issuing entity, the depositor, the sponsor, the originator, and various service providers. Some entities and their roles and relationships are similar to those that have already been extracted by IBM Midas from SEC and FDIC filings while others are novel to MBS-es.

While there are many sophisticated tools for analysis of the waterfall flow of funds, they have to be manually constructed. We address this challenge as follows:

- We will extract information from the Trust section of the MBS to create the *MBS-Waterfall* graph structure including the *Mortgage Pools* and the tranches or *Class Certificates*. Where relevant, entities and relationships and attributes will also be extracted from the Originators and Servicers sections.
- We will extract attributes about the mortgages and pools that provide revenue including the following: coupon (index/spread); maturity; installment payment structure; capitalizations and floors; amortization structure; teaser rates; prepayment penalty structures; contract structure (fixed, floating, option ARM, etc.); etc.

The *Distribution Rules* include the logic and rules describing the flow of funds for both interest and principal payments, from the Mortgage Pools to the Class Certificates. One of the major shortcomings of current tools for MBS waterfall analysis is that these tools are not linked to the actual payment histories associated with the individual mortgages. It is thus, not possible to develop an accurate understanding of the impact of various events. These events could range from a major event such as a drop in a specific mortgage rate which leading to a surge in early payments, to more local events such as an increase in late payments by homeowners in a particular state or county. Our model of the MBS-Waterfall should allow some level of reasoning about the flow of payments across the entities in the waterfall, as well as the impact of specific events.

Finally, the MBS+LEI graph datasets will be enhanced to develop a *Risk Profile* for each financial contract and/or counterparty to the contract. We will combine relevant attributes from the MBS+LEI graph with additional information, e.g., from the SEC and FDIC filings for the corresponding financial institutions. We will construct classifiers to predict risk profiles and we will develop methods to determine the sensitivity of these risk profiles to relevant events.

Research collaborators in the Accelerated Discovery Lab can use both the raw regulatory filings dataset and the derived graph datasets. The datasets and any insights derived from them can also be exported from the Accelerated Discovery Lab under appropriate licensing terms; see the Data Management Plan for details.

4.2 Development of a Simulation Testbed/ Strategies for Automated Trading (University of Michigan)

Our simulation testbed research will focus on fundamental questions about strategic behavior in two important finance domains. First, we study the stability and performance of complex networks of credit relationships. Recent experience suggests that in order to understand the phenomena of financial crises well enough to design preventative and mitigating policies, we require a better fundamental understanding of how complex networks of credit relationships form, how they operate under normal conditions to enable economic activity, and how they can unravel in the face of shocks in the economic environment. We study these questions by extending a model of credit networks (Dandekar et al., 2012) developed in recent years to capture patterns of trust relationships (DeFigueiredo & Barr, 2005; Ghosh et al., 2007; Karlan et al., 2009; Mislove et al., 2008). We have already used this model to study strategic network formation (Dandekar et al., 2012), and will extend it in this project to include dynamic credit decisions, and the implications of shocks on network structure.

The second focus domain for this part of the project is the performance and stability of financial securities markets dominated by algorithmic and high-frequency trading. The rapid rise of automated trading - the use of quantitative algorithms to automate the process of buying or selling in a market - has led to the development of various speed-reliant trading strategies over the past two decades. One of the more controversial types of automated trading is high-frequency trading (HFT), characterized by large numbers of small orders in compressed periods, with positions held for extremely short durations. (Wah and Wellman 2012, 2013) studies the effect of latency arbitrage on allocative efficiency and liquidity in financial markets. This research found that latency arbitrage degrades allocative efficiency. The model also showed that switching from continuous-time trading to a discrete-time mechanism—using one-second call markets—not only eliminates the latency arms race (Wellman, 2013) but also improves market performance. This project will build on those models to incorporate a broader range of trading strategies (for both background investors and algorithmic traders), and to model at finer grain the routing of orders and propagation of information among brokers, exchanges, and trading firms. Another particularly interesting avenue for research is the effect of widely available data and ubiquitous machine learning on financial market stability.

4.3 Content Aggregation Framework (CAF) – University of Maryland and University of Michigan

The Contract Aggregation Framework (CAF) builds upon fundamental data modeling principles that were used in developing the relational data model and the data cube model. CAF is motivated on the twin facts that data representing financial activity are available at widely disparate levels of temporal and cross-sectional aggregation, while empirical analyses are greatly facilitated by a standardization of the data. CAF is designed to capture financial contract data at arbitrary levels of granularity. We propose a flexible and extensible data model based on a *container* and *operators* to measure financial activity across multiple data (event) streams. The basic building block, a *container*, is a multi-dimensional representation where each cell corresponds to financial transactions (trading activity) relevant to a single *contract* or a set of contracts in a *ContractSet*. Mapping rules must be defined to populate the container. The mapping must be disjoint and provide complete coverage (where possible) of all elements from an event or data

stream(s). CAF must be extensible with respect to additional dimensions. This supports the integration of data from additional sources. We briefly discuss some extensions as follows: (1) Ownership and other relationships among the institutions that are counterparties to contracts. (2) Contracts that reference other contracts, e.g., an equity call option is a contract that references an equity. (3) Contracts that are composed of other contracts, e.g., mortgage-backed securities. (4) Contingencies specific to each contract type.

A fundamental task for systemic analysis is to summarize the size of financial activity and risk exposures on a comparable scale across markets and institutions (Lo and Wang 2010). We define an individual financial contract as an atomic unit. Size measures are then functions (simple or weighted aggregations) over some relevant attributes of a *contract* or *ContractSet*. Size may be computed with respect to a point-in-time snapshot of a portfolio inventory, or over an interval of time. Size measures are expected to be user-defined and CAF must support a primitive set of *operators* that can be used to define customized size measures. Some basic measures of size include *notional value*, recorded in the legal terms of the contract and *market value*, the price at which a contract changes hands.

An initial exemplar represents equity trading data from CRSP and corporate bond trading data from FINRA TRACE. We further consider a Credit Default Swap secured against a Senior Secured bond as well as a Call/Put Option referencing an equity. Figure 4 shows the entities (companies and contracts) and relationships. This simple example illustrates how CAF can integrate multiple data streams, across levels of granularity and temporal aggregation.

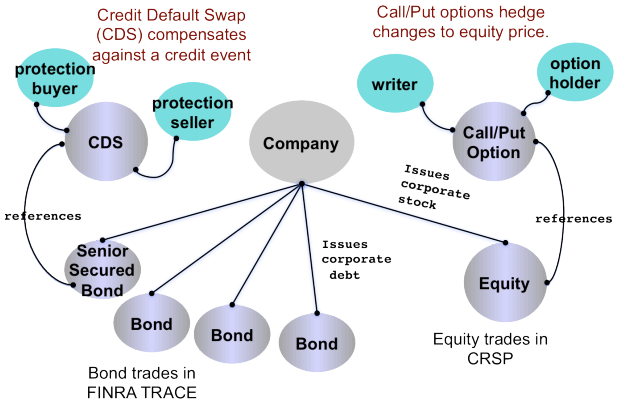


Figure 4: A Network of Interconnected Contracts and Trades.

A proof of concept used CAF to model bond and equity market volume data and is presented in (Ball et al. 2014). It demonstrates the feasibility of data integration through the CAF, and its ability to identify interesting patterns and research questions. For example, preliminary results from tensor decomposition applied to CAF allow us to identify tensor factors that correspond to potential latent patterns of co-trading of individual equities or bonds, during a particular interval. Notably, market price and market volume data from Yahoo! Finance appears to show correlated prices and co-trading activity in the same interval; this appears to confirm the observations of latent patterns from the tensor factors.

5. Education and Mentoring

Due to space limitations, we highlight a few of the expected educational and mentoring outcomes:

- The Michigan simulation testbed will be used by Wellman in "EECS 547: Electronic Commerce". That course includes many topics bearing on the financial system, and a substantial project focusing on modeling strategic behavior. The testbed will also be used in other relevant Michigan classes, including courses in Agent-Based Modeling and Advanced Artificial Intelligence.
- Raschid has offered a 1-credit course BMGT 499B on Next Generation Financial Cyberinfrastructure to seniors and graduate students in Business and Computer Science.
- Raschid will lead an effort to develop graduate level curriculum for *data science for finance* in conjunction with ongoing efforts in the department of Computer Science. A MOOC (massive open online course) will be developed to familiarize students with the Dsfin infrastructure. This will build upon a current course on Information Extraction being offered in Spring 2014 at UCSC by IBM.
- Multiple trainings will be offered on the tools and datasets. In addition, there will be two meetings where teams of multi-disciplinary researchers will work on DSfin Grand Challenges.

- 12 doctoral students in Business and Computer Science (advisees of members of the Steering Committee and Advisory Committee or the PI and co-PI) participated in the 2012 Workshop. It is expected that multiple doctoral students at Maryland, Michigan, UC Berkeley and at the home institutions of members of the Steering and Management Committee will be involved in research and dissertations related to *data science for finance*.
- Wellman advised Elaine Wah whose first two years of Ph.D. study were on an NSF IGERT Fellowship. Her research on latency arbitrage (Wah & Wellman, 2013) is key to the Michigan simulation platform.
- Raschid and Jagadish, together with finance researchers Mark Flood of the OFR and Joe Langsam of the Smith School of Business are mentoring two computer science and data science graduate students at Maryland and New York University. They are developing the Contract Aggregation Framework (CAF) to size the financial markets (Ball et al. 2014).
- Specific efforts will be made to engage women undergraduate and graduate students and students from under-represented groups. PI Raschid is an active member in the recently initiated Center for Women in Computing in the Computer Science department at the University of Maryland and she is also a mentor for an ongoing REU in the Computer Science department (CCF 1262805 REU Site: CAAR Combinatorial Algorithms Applied Research). PI Raschid has engaged two African American undergraduate women students in open source software projects. All PIs will actively recruit and engage minority REU students to be involved in developing community financial infrastructure.

Raschid made significant efforts to include women and under-represented groups in prior workshops and will continue these efforts. She will reach out to doctoral students of the KPMG Foundation Minority Doctoral Fellowship Program in Accounting. The financial recession had a disproportionate impact on many minority communities and the Consumer Finance Protection Bureau has made special efforts to connect to these communities and monitor progress; this may lead to a DSfin Grand Challenge problem.

6. Prior NSF Support

The PIs and collaborators are the recipients of numerous NSF awards, all of which have some impact on the planned infrastructure and *data science for finance* research. We focus on the most relevant awards. Raschid was PI on IIS 1237476 Workshop on Next Generation Financial Cyberinfrastructure (2012) and IIS 1033927 Workshop on Knowledge Representation and Information Management for Financial Risk Management (2010). The workshops articulated the need for community financial cyberinfrastructure and identified the *data science for finance* research challenges. It also brought together the PI, co-PIs and members of the Steering and Management committee. Wellman's most relevant prior NSF project, "Methods for Empirical Mechanism Design'", CCF-0905139, 2009–13 (collaboration with Amy Greenwald), employs techniques for empirical game-theoretic analysis (EGTA) to problems of mechanism design. For example, EGTA case studies led to insights about pricing for sponsored-search advertising (Jordan et al., 2010), the equity premium in financial markets (Cassell & Wellman, 2012), and other trading domains (Wellman, 2011). The main domain focus of the project was on canonical auction games, leading most notably to a characterization of equilibria for the long-open problem of simultaneous one-shot auctions, along with computationally practical and effective strategies based on this characterization (Mayer et al., 2013; Wellman et al., 2012). Similar techniques were successfully applied to sequential auctions (Greenwald et al., 2012). The project also produced algorithms (Jordan & Wellman, 2010), approximation techniques (Wiedenbeck & Wellman, 2012), and infrastructure (Cassell & Wellman, 2013) for EGTA that provide a foundation for the simulation platform described in this infrastructure proposal. H. V. Jagadish has a strong history of NSF-supported research activities, and each of his previous NSF projects has resulted in multiple publications at prestigious forums, and the training of graduate students, many of who are now successful researchers themselves. He has recently started work on IIS-1250880 (Big Data:Small: Choosing a Needle in a Big Data Haystack) – the first papers are being submitted for publication. Franklin's Expeditions in Computing Award CCF 1139158 Making Sense at Scale with Algorithms, Machines and People (2012-2017) is developing a broad range of analytical tools that will be made available to the *data science for finance* research community. Details were provided in the infrastructure description section (Agarwal et al. 2013, Kraska et al. 2013, Xin et al. 2013). Franklin is an unfunded collaborator.

7. Management Plan and Timeline

+6 months:	Initial list of datasets to be curated by the community.
+12 months:	Final list of datasets that will be curated for DSfin cyberinfrastructure. Initial list of analytical tools to be made available on the UMD DSfin stack. Publish DSfin Entity Resolution At Scale CLOSED Challenge.
+18 months:	First Workshop: Identification of DSfin Grand Challenge Problems. Final list of analytical tools. Begin migration of datasets from the IBM Accelerated Discovery platform. Initial user evaluation.
+24 months:	Publish DSfin Entity Resolution At Scale OPEN Challenge. Migration of the University of Michigan simulation platform and Berkeley ML Base. Initialize long term maintenance plan.
+30 months:	Second Workshop DSfin Grand Challenge Problems. Complete user guides and documentation. Complete migration of datasets. Complete migration of tools.
+33 months:	Final user evaluation and phase-in of 2 to 3 year maintenance plan.

PI **Raschid** will take primary responsibility for the management of the research tasks and the delivery of community financial infrastructure on the *UMD DSfin* platform. Each of the co-PIs will take primary responsibility for their own infrastructure. **Franklin** (UC Berkeley unpaid collaborator) will manage the UC Berkeley BDAS stack and **Wellman** (Michigan PI) will be responsible for the Michigan Strategic Simulation Platform. **Jagadish** (Michigan co-PI) will manage the list of analytical tools to be adopted by the community and Raschid will manage the list of datasets.

Raschid will work closely with the **UMD DSfin Project Manager** and the co-PIs of the collaborative projects. Planning will be completed within the first 12 months, development will be completed within 24 months. This will provide a year to port and test datasets and tools on the UMD DSfin stack.

Raschid, Jagadish, Franklin and Wallace have met several times in the past year to coordinate activities with IBM Almaden researchers (lead by Krishnamurthy). Raschid will spend several weeks during the project at IBM Almaden and UC Berkeley. **Wallace** (unpaid collaborator, UCBerkeley) will participate in a significant effort by the team for the curation of several public collections. Raschid has been leading the Karsha FOSS project in Sri Lanka since 2011 and travels to Sri Lanka twice annually.

Raschid, Krishnamurthy and **Oard** (UMD Senior Investigator) will lead on the DSfin Financial Entity Resolution At Scale Challenge.

The members of the team and members of the Steering Committee and Management Committee have already met in 2010 and 2012 (NSF workshops). There will be 4 virtual meetings of these committees each year. In addition, there will be 2 training workshops and two all hands workshops on DSfin Grand Challenge Problems.

Maintenance: Funds for developing the DSfin infrastructure is being requested from the NSF DIBBs program (DIBBs pilot demonstration award). Industry sponsors including IBM will provide a range of enriched public datasets, computational resources and expertise. The PI and Advisory Committee are reaching out to potential partners including the DTCC and FINRA to establish partnerships for the curation and dissemination of additional public datasets. Over the long term, it is expected that individual researchers and collaborative teams will incorporate these resources into their research agenda, thus providing a path for long-term support and maintenance of the *data science for* finance DSfin cyberinfrastructure, as well as ensuring the further development, dissemination and use of tools, metrics and use cases.

References

Adamic, L., Brunetti, C., Harris, J. and Kirilenko, A., "Trading Networks," Available at SSRN: http://dx.doi.org/10.2139/ssrn.1361184

Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S. and Stoica, I., 2013, "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data," in Proceedings of the ACM EuroSys Conference (Best Paper Award).

Baader, F., I. Horrocks, and U. Sattler, 2004, "Description Logics," in: Handbook on Ontologies, S. Staab and R. Studer, eds., Springer Verlag, Berlin, pp. 3-28.

Ball, C., Hoberg, G, and Maksimovic, V., 2012, "Redefining Financial Constraints: A Text-Based Analysis," University of Maryland Technical Report.

Ball, B., Flood, M., Jagadish, H., Langsam, J., Raschid, L. and Wiriyathammabhum, P., 2014, "A Flexible and Extensible Contract Aggregation Framework (CAF)," University of Maryland Technical Report.

Bennett, M., 2010, "Enterprise Data Management Council Semantics Repository," Internet resource http://www.hypercube.co.uk/edmcouncil/.

Bernstein, P., 2003, "Applying Model Management to Classical Meta Data Problems," Proceedings of the First Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, January 5-8, 2003.

Bernstein, P., A. Levy, and R. Pottinger, 2000, "A Vision for Management of Complex Models," Technical Report MSR-TR-2000-53, Microsoft Research, Redmond.

Borgida, A., M. Lenzerini, and R. Rosati, 2002, "Description Logics for Data Bases," in: Description Logic Handbook, F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, eds., Cambridge University Press, pp. 472-94.

Brammertz, Willi and Mendelowitz, Allan, 2010, "Regulatory Revolution: The Great Data Challenge," Risk Professional, 52-26.

Burdick, D., Evfimievski, A., Krishnamurthy, R., Lewis, N., Popa, L., Rickards, S. and Williams, P., 2014, "Financial Analytics from Public Data," IBM Technical Report.

Burdick, D., Franklin, M., Issler, P., Krishnamurthy, R., Popa, L., Raschid, L., Stanton, R. and Wallace, N., 2014, "Data Science Challenges in Real Estate and Asset and Capital Management," University of Maryland Technical Report.

Burdick, D., Hernández, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., Vaithyanathan, S. and Das, S., 2012, "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," IEEE Data Engineering Bulletin, Volume 34, Number 3, pages 60-67.

Cassell, B. and Wellman, M., 2012, "Asset Pricing under Ambiguous Information: An Empirical Game-theoretic Analysis.'' Computational and Mathematical Organization Theory, 18:445–462.

Cassell, B. and Wellman, M., 2013, "EGTAOnline: An Experiment Manager for Simulation-based Game Studies," in Multi-Agent Based Simulation XIII, volume 7838 of Lecture Notes in Artificial Intelligence, Springer.

Cerutti, E., Claessens, S. and McGuire, P., 2012, "Systemic Risks in Global Banking: What Can Available Data Tell Us and What More Data Are Needed?" Bank of International Settlements 376.

Cohen-Cole, E., Kirilenko, A. and Patacchini, E., 2013, "Financial Networks and the Propagation of Systemic Risk," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.

Committee to Establish the National Institute of Finance (CE-NIF), 2009, "Data Requirements and Feasibility for Systemic Risk Oversight," technical report, http://www.ce-nif.org/images/docs/ce-nif-generated/nif_datarequirementsandfeasibility_final.pdf.

Dandekar, P., Goel, A., Govindan, R. and Post, I., 2011, "Liquidity in Credit Networks: A Little Trust Goes a Long Way," Proceedings of the ACM Conference on Electronic Commerce, pages 147-156.

Dandekar, P., Goel,A., Wellman, M. and Wiedenbeck. B., 2012, "Strategic Formation of Credit Networks," Proceedings of the International WWW Conference, pages 559-568.

Davis Polk, Client NewsFlash, 2012, "CFTC Begins Implementation of Mandatory Clearing of Swaps".

DeFigueiredo, D. and Barr, E., 2005, "TrustDavis: A Non-exploitable Online Reputation system," Proceedings of the IEEE International Conference on E-Commerce Technology, pages 274–283.

Demystifying Legal Entity Identifiers, http://www.dtcc.com/downloads/news/CiCi_Report.pdf

Domingo-Ferrer, J., Sramka, M. and Trujillo-Rasua, R., 20120, "Privacy-Preserving Publication of Trajectories Using Microaggregation," Proceedings of the Workshop on Security and Privacy in GIS and LBS, pages 25-33.

Engle, Robert F. and Weidman, Scott, 2010, Technical Capabilities Necessary for Regulation of Systemic Financial Risk: Summary of a Workshop, National Research Council of the National Academies, National Academies Press, Washington, DC, http://www.nap.edu/catalog.php?record_id=12841.
Farmer, J. Doyne, 2010, "Networks and Systemic Risks", Video, Institute for New Economic Thinking, Kings College, Cambridge.
Federal Register, Vol. 77, No. 9, Friday, January 13, 2012, Rules and Regulations, pp. 2136-2224
Federal Register, Vol. 77, No. 100, Wednesday, May 23, 2012, Rules and Regulations, pp 30596-30764
Federal Register, Vol 77, No. 113, Tuyesday, June 12, 2012, Rules and Regulations, pp. 35200-35239
Federal Register, Vol. 77, No 162, Tuesday, August 21, 2012, Proposed Rules, pp 50425-50443
Financial Stability Board, "Technical Features of the Legal Entity Identifier (LEI), March 7, 2012.
Flood, M., A. Kyle, and L. Raschid, 2010, "Workshop on Knowledge Representation and Information Management for Financial Risk Management," Internet resource:, http://www.nsf-fiw.umiacs.umd.edu/index.html.
Flood, M. and Mendelowitz, A. and Nichols, B., 2013, "Monitoring Financial Stability in a Complex World," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.
Flood, M., Jagadish, H., Kyle, A., Olken, F. and Raschid, L., 2011, "Using Data for Systemic Financial Risk Management," Proceedings of the Conference on Innovations in Data Systems Research (CIDR2011), pages 144-147.
Fouque, J. and Langsam, J., 2013, "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.
Fouque, J. and Sun, L.-S., 2013, "Systemic Risk Illustrated", in in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.
FpML, 2004, FpML Financial product Markup Language 4.0 Recommendation, Internet resource: http://www.fpml.org/spec/latest.php.
Frank, A. and Asuncion, A., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
Garnier, J., Papanicolaou, G., Yang, T.-W., 2013, " Diversification In Financial Networks May Increase Systemic Risk," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.
Ghosh, A., Mahdian, M., Reeves, D., Pennock, D. and Fugger, R., 2007, "Mechanism Design on Trust Networks," Proceedings of the International Workshop on Internet and Network Economics, pages 257–268.
Greenwald, A., Li, J. and Sodomka, E., 2012, "Approximating Equilibria in Sequential Auctions with Incomplete Information and Multi-unit Demand," Proceedings of the Conference on Advances in Neural Information Processing Systems.
Harris, T. et al, 2010, "WormBase: A Comprehensive Resource for Nematode Research," Nucleic Acids Research, volume 38, pages 463-467.
Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L, Stanoi, I., Vaithyanathan, S. and Das, S., 2012,"Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance," IBM Technical Report.
Hunty, J., Stanton, R. and Wallace, N., 2011, "The End of Mortgage Securitization? Electronic Registration as a Threat to Bankruptcy Remoteness, " Technical Report, University of California, Berkeley, 2011.
International Standard ISO 17442, Financial Services – Legal Entity Identifier (LEI).
Jaffee, D., Stanton, R. and Wallace, N., 2011, "Energy Efficiency and Commercial Mortgage Valuation, " Technical Report, University of California, Berkeley, 2011.
Jaffee, D., Stanton, R. and Wallace, N., 2011, "Energy Factors, Leasing Structure and the market Price of Office Buildings in the U.S., " Technical Report, University of California, Berkeley, 2011.
Jagadish, H., 2013, "Data for Systemic Risk," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press.
Jordan, P. and Wellman, M., 2010, "Algorithms for Finding Approximate Formations in Games," Proceedings of the AAAI Conference on Artificial Intelligence, pages 798–804.
Jordan, P.,Wellman, M. and Balakrishnan, G., 2010, "Strategy and Mechanism Lessons from the First Ad Auctions Trading Agent Competition," Proceedings of the ACM Conference on Electronic Commerce, pages 287–296.
Karlan,D., M. Mobius, T. Rosenblat, and A. Szeidl, 2009, "Trust and Social Collateral," Quarterly Journal of Economics, 124:1307–1361.
Karsha DASS. "Dcoument Annotation and Semantic Search," Internet resource: https://wiki.umiacs.umd.edu/clip/ngfci/index.php/KarshaDASS
Kraska, T., Talwalkar, A., Duchi, J., Griffith, R., Franklin, M. and Jordan M., 2013,"MLbase: A Distributed

Machine Learning System", in Proc. of the Conference on Innovative Data Systems Research.
Mayer, B., E. Sodomka, A. Greenwald, and M. Wellman, 2013, "Accounting for Price Dependencies in Simultaneous Sealed-bid Auctions," Proceedings of the ACM Conference on Electronic Commerce, pages 679–696.
Mislove, A., A. Post, P. Druschel, and K. P. Gummadi, 2008, "Ostra: Leveraging Trust to Thwart Unwanted Communication," Proceedings of the Usenix Symposium on Networked Systems Design and Implementation, pages 15–30.
PWC, 2011, "A Closer Look –The Dodd-Frank Wall Street Reform and Consumer Protection Act; Impact on Swap Data Reporting" June 2011.
Raschid, L., 2012, "Fiscal Policy, Governance, Citizenry and Financial Indicators: Modeling through the Lens of Social Media, University of Maryland Technical Report.
Ruiz, E., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A., 2012, "Correlating Financial Time Series with Micro-Blogging Activity," ACM International Conference on Web Search and Data Mining (WSDM).
Tamersoy, A., Loukides, G., Nergiz, M., Saygin, Y. and Malin, B., 2012, "Anonymization of Longitudinal Electronic Medical Records,'' IEEE Transactions on Information Technology in Biomedicine, volume 16, pages 413-423.
TDT2004 Workshop Presentations and System Description Papers. Internet resource: http://www.itl.nist.gov/iad/mig//tests/tdt/
The Financial Crisis and Information Gaps: A Report to the G-20 Finance Ministers and Central Bank Governors, 2009, Working paper by the IMF Staff and FSB Secretariat.
Wah, E. and Wellman, M., 2013, "Latency Arbitrage, Market Fragmentation and Efficiency: An Two Market Model," Proceedings of the ACM Conference on Electronic Commerce, pages 855–872.
Water Cost Index. Available at http://researcher.watson.ibm.com/researcher/view_project.php?id=5047
Wellman, M., 2011, "Trading Agents," Morgan and Claypool.
Workshop on Data Confidentiality, March 2012. Internet resource: http://stability.psu.edu/policy-corner
Xin, R., Rosen, J,Zaharia, M., Franklin, M., Shenker, S. and Stoica, I., 2013, "Shark: SQL and Rich Analytics at Scale", Proceedings of the ACM SIGMOD Conference.
Zhai, K., Boyd-Graber, J., Asadi, N. and Alkhouja, M., 2012, ''Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce,'' Proceedings of the ACM International Conference on the World Wide Web.

Supplementary Documents: In the Supplementary Documents Section, provide a list of PIs, Co-PIs, Senior Personnel, paid Consultants, Collaborators and Postdocs to be involved in the project. This list should be numbered and include (in this order) Full name, Organization(s), and Role in the project, with each item separated by a semi-colon.

1. Louiqa Raschid;	University of Maryland;	PI;
2. Amol Deshpande;	University of Maryland; MIT;	co-PI;
3. Douglas W. Oard;	University of Maryland;	Senior Investigator;
4. Michael Franklin;	UC Berkeley;	UNPAID Collaborator;
5. Nancy Wallace;	UC Berkeley;	UNPAID Collaborator;
6. Michael Wellman;	University of Michigan;	co-PI;
7. H.V. Jagadish;	University of Michigan;	co-PI;
8. Rajasekar Krishnamurthy;	IBM Research;	IBM Team Lead;
9. Doug Burdick;	IBM Research;	IBM;
10. Lucian Popa	IBM Research;	IBM;

4. Data Management Plan

The shared community infrastructure will be generated through a collaboration between the University of Maryland, the University of Michigan, the University of California Berkeley and IBM Research. The goal is to eventually provide a flexible environment, the *UMD Data Science for Finance (DSfin Stack),* for the community of researchers to preserve, run, and share the data collections and analytical tools that they develop. Porting Michigan's high performance simulation platform and UC Berkeley's MLBase to the UMD DSfin stack would be a pilot test of interoperability and long term maintenance.

All material that will be generated (datasets, metadata, use cases, software, ontologies, metadata) will be made available under an appropriate license (open source, educational license, etc.) by the corresponding organization. The license will allow access to the research community and for educational purposes. As an example, both the raw regulatory filings dataset and the derived counterparty relationship dataset can be used by other research collaborators in the Accelerated Discovery Lab. The datasets and any insights derived from them can also be exported from the Accelerated Discovery Lab.

A key part of the project will be to develop a detailed plan that addresses methods for documentation, maintenance and dissemination of the community financial cyberinfrastructure. This will include the following capabilities:

- Deploy collaboration tools including MediaWiki for wikis, Redmine for project management, Request Tracker (RT) for User Support and WordPress for blogs, polls, and basic content management.
- Manage software development projects with source code control using Subversion and Gitlab.
- Develop new applications using standard LAMP stacks built on Mysql, Postgres, PHP, Python and Ruby.
- Develop and deploy new virtual server platforms as needed and without administrative intervention.

There are several examples of community infrastructure, portals, model organism databases, etc., that have been sponsored by the NSF and the NIH. Exemplars include the UCI Machine Learning Repository [Frank and Asuncion] and WormBase [Harris et al 2010]. We will follow best practices from both the computer science and bioinformatics communities to identify a plan for access and dissemination, and a data management best practices protocol. Every effort will be made to use open standards and protocols and to make all resources available to the public.

An effort will be made to establish a NIST TREC *Data Science for Finance* track to provide marked up datasets for evaluation and benchmarking purposes.

Due to space limitations, we discuss the data management plans for two projects, among the diversity of organizations, tools, services, datasets, ontologies, metadata, etc., that will be developed, as follows:

IBM Accelerated Discovery Lab: A shared Big Data analytics environment in which IBM researchers, external researchers and other practitioners collaborate on joint research projects which push on today's boundaries of the 4 V's – volume, velocity, variety, and veracity. This environment represents a multi-million US$ investment (to-date) by IBM and is designed to give users the system functionality and support for the data wrangling and analytics activities necessary for conducting these research projects.

At the start of a project, a project team can bring into the Lab code and data which both require the necessary licensing (including open source). These licenses are recorded within the Accelerated Discovery Lab. HIPAA data (personal health information), PCI data (credit card information), ITAR or other export data controlled data are not allowed.

When a project exits the Accelerated Discovery Lab, members may take the following with them:

- Data, summary statistics, key result metrics for publications, any data set that is licensed for use by all participants without restrictions or derived solely from work within the Accelerated Discovery Lab, etc.
- Source code that is solely authored by the participants.

At the end of the funded NSF project, during which the IBM Accelerated Discovery Lab environment and support would be provided to the DSfin community, the University of Maryland will acquire licenses for any IP or products that were either used or developed. This would allow continued use of these products in the following two to three years, under the maintenance mode of support.

University of Michigan Strategic Simulation Testbed
The strategic modeling testbed is based on computational infrastructure developed at the University of Michigan for managing large-scale agent-based simulation experiments (Cassell & Wellman, 2013). The infrastructure is designed to make it simple for modelers to run large experiments covering a wide range of environment parameters and agent behaviors, by insulating end-users from the details of scheduling computational jobs and organizing the volumes of data produced. It provides a front-end for specifying patterns of strategy and simulation parameter combinations, and a back-end for managing jobs on a computing cluster and the resulting simulation data. Through web-based interfaces and APIs, external users can specify and operate their own strategic simulation experiments. It currently operates in conjunction with the research data center at the University of Michigan, but since the back-end functions are modular it can be readily generalized to work with a variety of cloud-computing platforms. In this project, we plan to develop an interface so that the system can be operated on the new cloud environment at the University of Michigan.

The greatest volume of data produced by this project will be the results of computational experiments, primarily taking form of simulation data from algorithmic trading scenarios designed and investigated in this project. Our output data includes the raw simulation observations as well as empirical game models induced from these observations. The testbed infrastructure that currently exists and will be extended in this project maintains simulation data in a persistent store (relational database), which is backed up regularly. Although there is no standard interchange format for game-theoretic models, the testbed itself employs a regular representation that applies across domains. The representation is based on a game structure we term role-symmetric form, which exploits sparseness and symmetry for compactness, but does not require that the entire game be symmetric. The underlying data can be exported in text files using the JSON (JavaScript Object Notation) interchange format. We will maintain documentation for the representation as well as provide simple software tools (e.g., parsers and analysis methods, as described below) that operate on this format. Data from experiments that bear on publications or other reports of our research results will be maintained and archived. The strategic simulation testbed will produce numerous software artifacts in several categories, including domain simulators, learning algorithms, and game-theoretic analysis tools. Based on experience operating the EGTA infrastructure, we aim to widen its availability to potential collaborators, initially through a web-service interface.

During the second year of the funded NSF project, these artifacts will be ported to the University of Maryland DSfin platform. Usability will be tested during the next year. A version of the strategic simulation testbed will be made available by the University of Maryland for the next two to three years.

Linked Data SPARQL Endpoints:
A diversity of resources (identified within the first six months of the project) will be marked up using the W3C RDF Data Cube Vocabulary, annotated using terms from the Financial Industry Business Ontology (FIBO), and will be made available via SPARQL Linked Data endpoints. The Lanka Software Foundation will manage this task in collaboration with PI Raschid.
Maintenance: Funds for developing the DSfin infrastructure is being requested from the NSF DIBBs program (DIBBs pilot demonstration award). Industry sponsors including IBM will provide a range of enriched public datasets, computational resources and expertise. The PI and Advisory Committee are reaching out to potential partners including the DTCC and FINRA to establish partnerships for the curation and dissemination of additional public datasets. Over the long term, it is expected that individual researchers and collaborative teams will incorporate these resources into their research agenda, thus providing a path for long-term support and maintenance of the *data science for* finance DSfin cyberinfrastructure, as well as ensuring the further development, dissemination and use of tools, metrics and use cases.