Application for the BIDS Data Science Fellows Program: Real Estate Asset and Capital Market Analytics and Financial Data Science

Richard Stanton* Nancy Wallace†

February 28, 2014

*Professor of Finance and Real Estate, Kingsford Capital Management Chair in Business, Haas School of Business, U.C. Berkeley, stanton@haas.berkeley.edu.

†Professor of Finance and Real Estate, Lisle and Roslyn Payne Chair in Real Estate and Capital Markets, Co-Chair Fisher Center for Real Estate and Urban Economics, Haas School of Business, U.C. Berkeley, wallace@haas.berkeley.edu.

Contents

1 REFM Scientific Goals and Objectives

The Real Estate and Financial Markets (REFM) Laboratory is housed in the Fisher Center for Real Estate and Urban Economics, an Organized Research Unit (ORU) in the Haas School of Business, U.C. Berkeley. The central scientific mission of the REFM Lab is to develop models, data sets, and analytical technology for economic analyses of real estate asset and capital markets, focusing in particular on the valuation and analytics necessary for trading and risk management in these markets, and on making significant, data-driven contributions to the policy debates on systemic risk management.[1,2] The faculty and staff of the REFM Lab are currently engaged in numerous projects to further these goals (more details of which can be found in Section 3), including

- A lattice-based analysis of the capital funding, origination, securitization and performance of the U.S. residential mortgage market;
- A project to develop new dynamic house-price indices for the U.S. that are suitable for mortgage valuation modeling and avoid the biases inherent in existing indices;
- Analysis of the interaction of residential mortgage availability, consumption, and job creation;
- The development of new methods to evaluate and underwrite energy efficiency risk in commercial mortgage loan underwriting;
- The creation and curation of contract and Legal Enterprise Identifier (LEI) graph datasets to represent collections of financial contracts and affiliated counterparties to these contracts;
- The development of enhancements to the contract and LEI graphs with risk profile metrics for each contract and counterparty represented;
- The development of automated reasoning algorithms over the contractual features and covenants of financial contracts that populate the graphs.

[1] These are two of the largest capital markets in the U.S. As of the third quarter of 2013 the market value of household held real estate was $21.6 trillion (see `http://www.federalreserve.gov/releases/z1/current/accessible/b100.htm`) and as of the fourth quarter of 2013, the outstanding balance of the U.S. mortgage market was $13.2 trillion (see `http://www.federalreserve.gov/releases/z1/current/z1.pdf`).

[2] The role of the housing and mortgage markets in the recent financial crisis, and how to respond, remains a critical priority for U.S. policy makers. For example, in her semi-annual Monetary Policy Report to the House Financial Services Committee on February 11, 2014, Federal Reserve Chair (and former Haas faculty member) Janet Yellen commented that "I strongly support, and would urge the Congress to address the issue of GSE (government sponsored enterprise) reform. We've gotten a mortgage system that in a way remains very highly dependent on government backing and it fails to meet the very important objective of successful securitization without systemic risk." (see `http://www.reuters.com/article/2014/02/27/us-usa-fed-highlights-idUSBREA1Q1JQ20140227`).

In addition to our research mission, the REFM lab provides a forum for world-class scholarship for researchers in real estate, finance, economics, mathematics, and computer science. The Lab also supports state-of-the-art research and training opportunities for post-doctoral students, Ph.D. students, Masters students, MBAs, Masters in Financial Engineering students, and undergraduates at U.C. Berkeley.[3]

2 The REFM Database Initiative

Until recently, a very significant constraint to research on real estate and mortgage capital markets has been the lack of available data. For example, until October 2013 there was no data set containing house-level characteristics of the U.S. housing stock, and neither was there a transactional data set containing the sales prices of U.S. houses and their corresponding characteristics at the time of sale. As another example, until recently there was no data set in the U.S. that uniquely linked individual residential mortgage loans from the borrower to the originator, to the capital source used by the originator, to the sponsors, the depositors, and to the special-purpose entities that actually issued the mortgage-backed securities and accounted for the contractual features of these flows.

This situation is changing rapidly, and new data sets are becoming available almost daily that allow researchers to address questions that were previously impossible. This proliferation of new data has, in turn, brought new challenges associated with storage and processing, especially since much of it is in the form of huge, unformatted alphanumeric documents, where the important information is to be found in footnotes scattered randomly throughout the documents.

The REFM Lab has established a high-performance, flexible hardware and software architecture that includes IT infrastructure and security systems for very large (and often proprietary) data sets and provides secure access to these datasets through a high-speed network. A central focus of the research effort of the faculty and staff of the 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. Working with data vendors across the U.S., REFM faculty have negotiated contractually approved access and, for the first time, have linked micro-data for the evolution of housing characteristics and prices, the loan-level evolution of mortgage contracting structures, capital-funding maps for all U.S. originators, and local employment

[3]For example, a group of six Haas Masters in Financial Engineering students, using data provided by the REFM Lab, just won the International Association of Quantitative Finance Case competition in New York City against 24 other graduate schools with financial-engineering programs.

dynamics. The REFM database currently includes:

- Loan-level monthly performance data for the securitized commercial mortgage market;
- Origination and monthly performance data for about fifty-four million residential mortgage loans;
- Firm-level performance data and loan-level mortgage origination activity for all U.S. commercial banks, thrifts, and mortgage companies;
- Employment data for an annual panel of 23 million U.S. establishments from 1984–2010;
- The Nielsen consumer-preference data;
- House-level transaction and characteristic data for the entire U.S. housing stock (2003–2013);
- Commercial real-estate characteristic and transaction data (1996–2009);
- Soil and weather data for the U.S. by latitude and longitude;
- High-frequency pricing data for the natural-gas and electricity hub markets in the U.S. (1998–2006).

In addition to data, the REFM Lab also provides access to state-of-the art processing capabilities. There are three large servers, each with 16 cores and 128GB of RAM, attached to 48TB of network attached storage (NAS, for storing data and providing working storage for researchers). Software currently installed includes SAS, Stata, Matlab, Python, R, C, C++ and FORTRAN, with full support for running programs on multiple processors and on multiple machines simultaneously.

The hardware infrastructure for the REFM Lab (see Figure 1) is implemented as part of the existing Beowulf Research Cluster at the Haas School of Business. REFM Lab users use light client computers (laptops and desktops) to access the lab either locally through the Haas Local Area Network (LAN) and the campus-wide wireless network (AirBears), or remotely through broadband connections over a Wide Area Network (WAN).

The Beowulf cluster design is a cost-effective solution for implementing a high-performance computing environment. In this approach, a master handles the scheduling and optimally distribute the computational load across powerful compute nodes. This modular implementation provides the REFM Lab with the ability to respond to the demand of large computational analyses by employing parallel-computing applications, as well as the flexibility for expanding the hardware infrastructure as the REFM Lab grows in its research activities.

The core components of the hardware architecture (compute nodes and network attached storage) are connected through a private high-speed network, currently operating at a rate of 4.0 Gb/s. The private network creates the means for sharing the REFM Lab data among

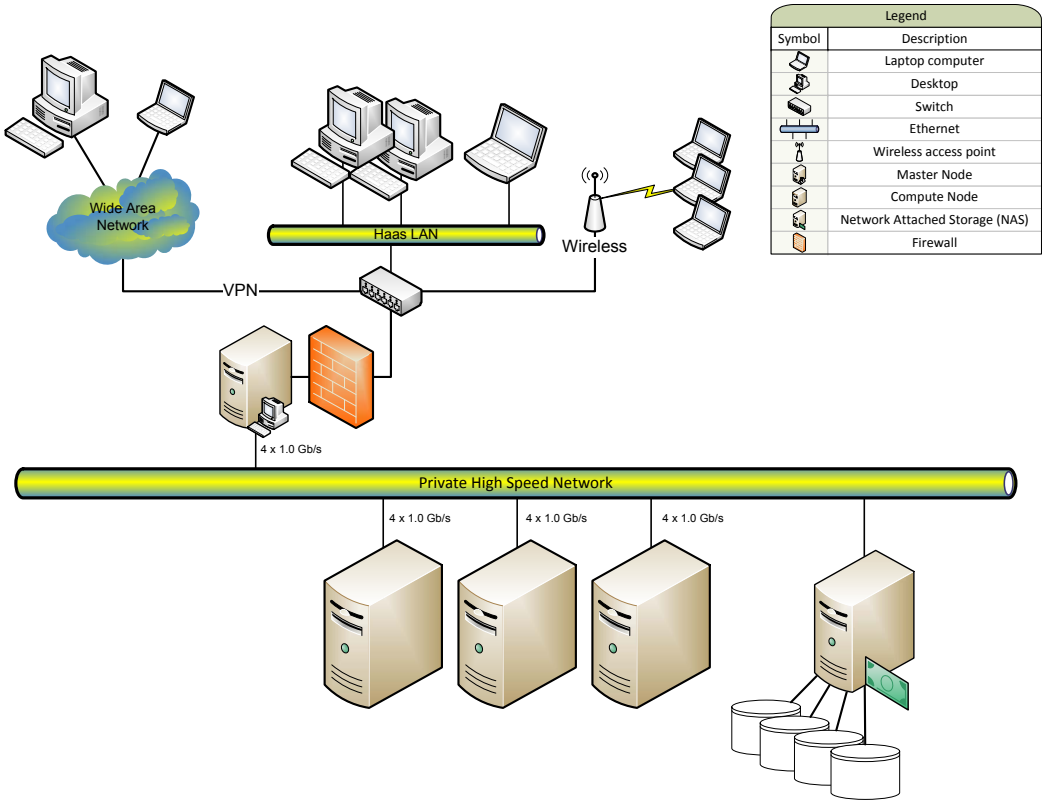

Figure 1: REFM Network schematic

users and processes running at the compute nodes. High transfer rates ensures that the lab core components operate at their top performance. The firewall creates a barrier between the private and other networks, protecting the REFM Lab data from unauthorized access, as well as malicious attacks.

3 On-going REFM Research Initiatives

The official launch of the REFM Lab is scheduled for mid-April 2014. As discussed above, the faculty of the REFM Lab have numerous major research initiatives underway, which have already lead to five working papers, several academic and practitioner presentations, and one publication. We provide brief summaries of some of these projects here.

3.1 Network Analysis of U.S. Mortgage Markets

Lab researchers and faculty are engaged in on-going research (see Stanton, Walden, and Wallace, 2014) to develop network representations of the sources of capital used to fund mortgage

originators, the supply chain for loan-level origination and securitization transactions, and firm-specific counter-party risk exposures in the U.S. residential mortgage and securitized bond markets. These network analyses are designed to generate statistics that can be used to evaluate the aggregate risk in mortgage markets and to account for the relative performance of specific firms and loan contract types. To better visualize the current problem with mortgage loan origination data, Figure 2 presents a standard measure of market concentration, the Herfindahl index, for single-family mortgage origination by U.S. county in 2006 using data from the Home Mortgage Disclosure Act (HMDA-2006).[4] As shown, with the exception of parts of Texas, lower-population areas tend to have higher levels of market concentration, whereas most populated areas appear to have high levels of competition.

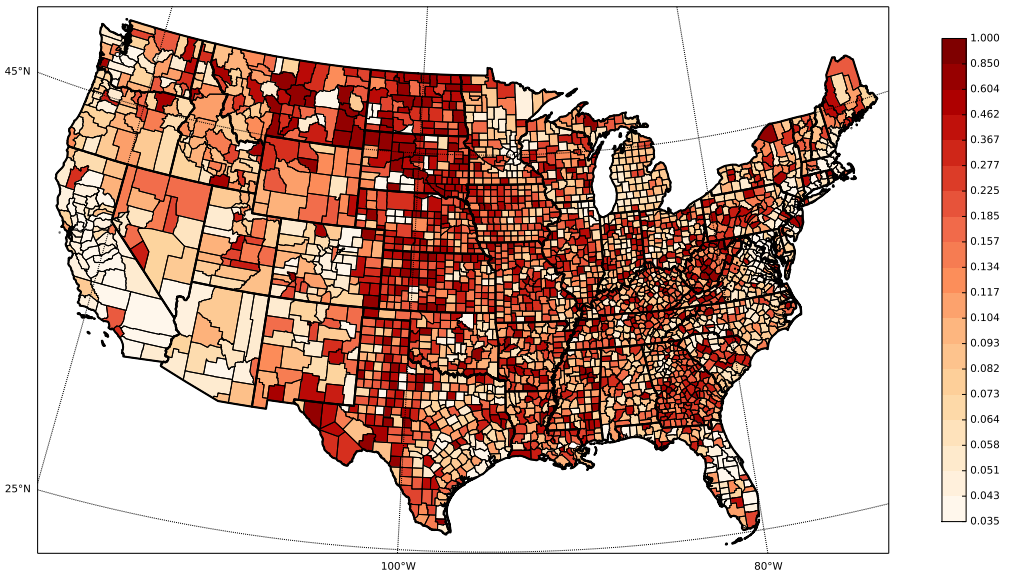

Figure 2: Geography of Mortgage Origination — HMDA

In contrast, Figure 3 presents the corporate organizational and funding structure for the residential mortgage origination market in 2006. As shown, mortgage origination flows are organized within five strata of influence: 1) the independents, either depositories or non-depository mortgage companies; 2) the depositories and subsidiaries; 3) the bank and thrift holding companies; 4) the regulators; and 5) the securitization channels.[5] Direct ownership

[4]For a county with a single originating institution the HHI would equal one, while in a perfectly competitive, atomistic market the HHI would approach zero.

[5]Mortgage brokers are not shown because, following the logic of HMDA, these entities do not make the underwriting and funding decision in mortgage origination.

(or partial ownership) channels between these strata are shown by red dotted lines. Black dotted lines connect the regulators to their respective regulated entities. Blue dotted lines are the primary securitization channels, and green dotted lines represent the contractual mortgage-origination funding channels from the correspondent lenders and the warehouse lenders to the independent mortgage companies (MCs) and depositories who make the underwriting and funding decision. These contractual funding channels introduce important elements of systemic risk associated with short-term liquidity risk and with counterparty exposures among the mortgage originators and their funders.

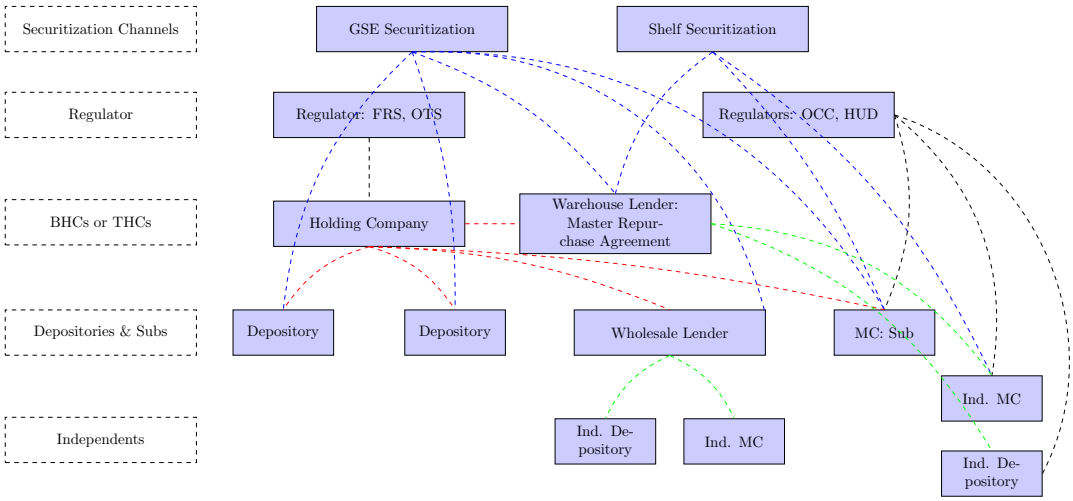

Figure 3: Mortgage Market System and Risk Characterization

Figure 4 presents a graphical representation of millions of subprime mortgages originated in 2006. The individual mortgages are represented as flows from their originators, the outer cilia, to their corresponding aggregators who funded the originators, the first ring of discs, and to their securitizers, the inner ring of discs, where black are current loans as of 2013 and red are defaulted loans as of 2013. This graphic representation shows, in stark contrast to Figure 2 and all prior work, that the U.S. residential single-family mortgage-origination market is in fact highly concentrated due to the short-term warehouse-funding channel. Despite the large number of players, we show that the origination market is a "small world," in that most originators, aggregators, and bond issuers are close in the network; many apparently atomistic mortgage underwriters are, in fact, coordinated to act in parallel because of their funding relationships with the large bank holding companies. These contractual interlinkages represent a previously under-appreciated source of systemic risk, and our network representations can be used in a forecasting framework to measure both expected and realized default and prepayment risk, and to forecast network performance under a pre-defined

crisis scenario, similar to supervisory Dodd-Frank Act Stress Tests (DFAST).

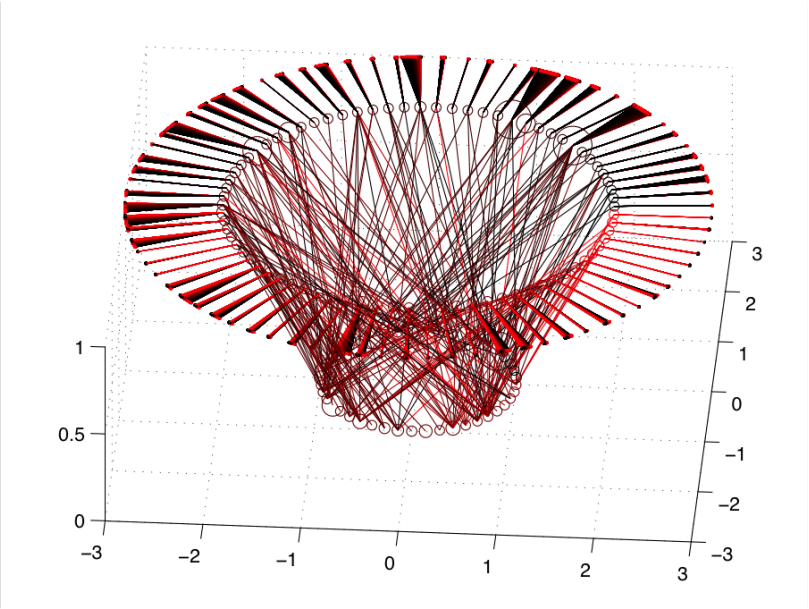

Figure 4: Lattice-based Measures of Mortgage-Linked Systemic Risk

3.2 Improved U.S. House-Price Indices

A second key objective is the development of dynamic house-price indices for the United States. Because houses are the pledged collateral for all residential mortgage contracts and are often the pledged collateral for small business and other middle-market companies, house price indices have become an important component of the supervisory stress tests under DFAST. Existing methods to estimate house prices are known to produce biased estimates of the cross-sectional and time-series volatility of house prices. This downward bias induces upward bias in the valuation of mortgage products with embedded prepayment and default options. Well structured dynamic house price indices are essential to the development of accurate valuation technology for mortgage products and for stress testing analytics.

The REFM is the first adopter of a new data set that that contains both time-varying characteristics for all properties in the U.S. from 2003–2013 and corresponding transaction prices over the same ten-year period. Using this new and important data set, REFM faculty and one of our Ph.D. students have developed and estimated a house-price model that incorporates controls for the probability of sales and remodels into a dynamic price-index estimation framework (see Amirdjanova, Stanton, and Wallace, 2014). Our estimation strategy includes channels for the hedonic characteristics of the house, for a parsimonious representation of the macro-fundamental determinants of house price formation, for the idiosyncratic

components of house price volatility, and for the volatility and drift of returns for baskets of houses defined by their characteristics. We then apply classical linear-filtering techniques to obtain the unobservable transition equation, the house price index. Our estimator reveals significant differences in both the level and volatility of prices, compared with estimates from the Federal Housing Finance Board's weighted-repeat-sales (WRS) price index (the dominant index currently in use). This suggests that bias in standard WRS indices and their low volatility estimates may render them unsuitable either for forecasting housing returns or for modeling mortgage default risk.

3.3 Mortgage markets and the U.S. Economy

A third REFM research objective is analysis of the interlinkages between origination activity in the residential and commercial mortgage market, net local small business creation, employment growth, and the effects of monetary policy on these flows both pre and post the financial crisis (Di Maggio and Kermani, 2014). One on-going project (Kermani, 2014) is a theoretical model of the booms and busts in housing and non-housing consumption that are driven by the interplay between relatively low interest rates and an expansion of credit. The model suggests that when credit becomes available, households would like to borrow in order to consume more, and this increases demand for housing and non-housing consumption. If the increase in the demand for housing translates into an increase in prices, then credit is fueled further both because the existing housing stock is now more valuable and because housing can be used as collateral. Due to lifetime budget constraints, and even in the absence of a financial crisis, the initial expansion in housing and non-housing consumption will be followed by a period of contraction, with declining consumption and house prices. Another on-going faculty project (Hartman-Glaser, Stanton, and Wallace, 2014) is an analysis of the large-scale asset purchases (LSAPs) undertaken by the Federal Reserve under quantitative easing programs that have funneled vast amounts of capital into the secondary market for mortgages. The direct effect of these programs on the primary mortgage market is not yet clear. This project has both a theoretical and empirical component, suggesting that while the LSAPs may have improved conditions for the least risky borrowers, they have not improved conditions for all borrowers. For example, the average FICO score of agency-securitized mortgages increased from below 720 (low risk) in 2008 to above 760 (extremely low risk) in 2012.

3.4 Mortgage Valuation and Energy Efficiency

A fourth current research objective of the REFM researchers is continued development of data-based valuation analytics for leasing and mortgage markets, with a focus on interest rate, credit, and energy risk. This component of the REFM research programs focuses on the development of new metrics for systemic and regional components of interest rate, credit, and energy risk in mortgage and real estate markets. Energy efficiency is a key factor to the future of the U.S. economy, and commercial buildings are among the largest users of energy. However, existing commercial mortgage underwriting practices provide little incentive for building owners to make their buildings more energy efficient. In Jaffee, Stanton, and Wallace (2011), we extend standard mortgage valuation methods, which account for the expected dynamics of interest rates and building prices, by including the expected dynamics of the electricity and gas forward prices, which, as shown in Figure 5, now have important double seasonals for both winter and summer due to the increased use of natural gas in electricity production. This allows us explicitly to incorporate energy risk and efficiency measures, which depend on both the energy efficiency of the building and the characteristics of its location, into commercial mortgage valuation and underwriting. The REFM faculty and students have applied our valuation methodology to price a sample of 1,390 mortgages, originated between 2005–2007, on office buildings located in 28 cities across the U.S. We find that, relative to the traditional mortgage valuation methodology, our proposed strategy leads to an 5% reduction in the mispricing of the default risk of commercial mortgages.

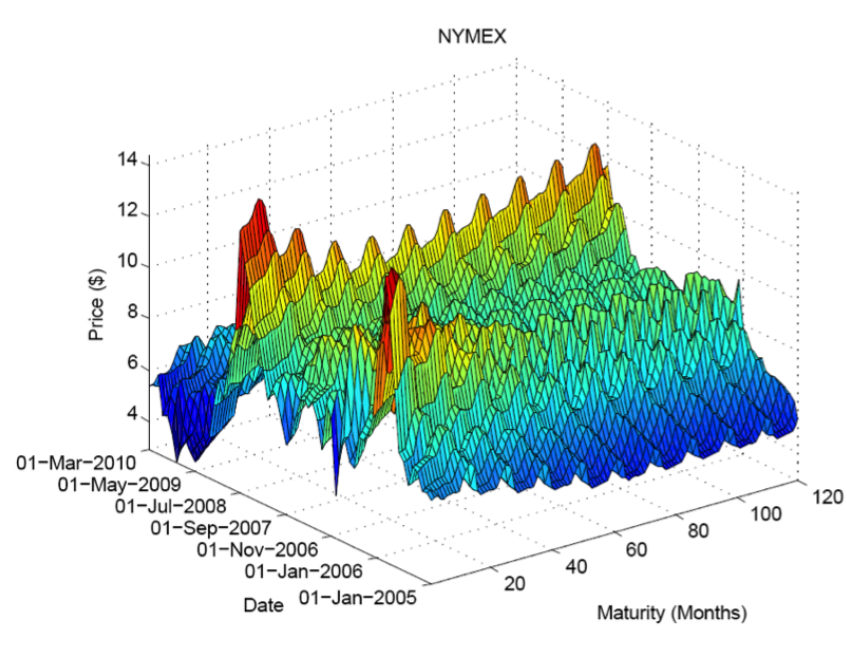

Figure 5: Seasonals in Futures Prices: Nymex Henry Hub Futures

4 Larger Data-Science Initiatives and Collaborations

Nearly all of the data and analytic focus of the REFM Lab involves numerical and logical representations of bi-party (or multi-party) contracts with a high-dimensional contracting space and numerous embedded options. Most mortgage bond contracts require three-hundred page PDF files to enumerate all of the contractual features, identify all the involved parties, and specify all of the covenants of the bonds. Other components of the REFM database involve firm-level information obtained from financial statements, regulatory filings, and bankruptcy filings. Cleaning and codifying these data presents significant challenges that stand to benefit greatly from affiliations with units on campus such as the AMPLab in the Computer Science Department at U.C. Berkeley, which are developing improved techniques for revealing information available in opaque formats and from unstructured data sources. Given the significant advances in computer science for data-mining applications, crowd-sourcing, and other logical scraping technology from alphanumeric-field documents, the REFM lab stands to gain significantly from partnerships with computer-science researchers and students with a focus on new frontiers in large data science for contractual data.

In addition, since most of the contracts that are the focus of the REFM Lab research initiatives have long maturities of up to forty years, with monthly payouts, valuation and risk analytics involve high-dimensional dynamic programming techniques, counterfactual forecasting techniques for stress testing analytics, and the use of expensive (in terms of compute cycles), algorithms to solve partial differential equations with free boundary conditions. Here again, partnerships with other units on campus, such as with faculty members in the AMPLab, operations research or mathematics, could lead to highly productive breakthroughs in developing faster and more stable ways to solve these problems. Analytical improvements in mortgage and house-price analytics are likely to achieve important public policy benefits in fostering the adoption of new, more accurate technologies within the regulatory community and financial institutions. Currently, a large component of mortgage and house-price analytics is carried out by practitioners and regulators using simplified reduced-form representations of the contracting structure and applying multivariate regressions. It is well known that these strategies have poor out-of-sample power and that, although existing methods may reveal how embedded options are exercised, they cannot reveal causality, or why the options are exercised.

To meet these challenges, the faculty of the REFM laboratory are organizing members of a new consortium focused on data science in real estate and financial markets (DSFin) that includes multiple organizations (University of Maryland, University of Michigan, University of California Berkeley and IBM Research). DSFin was established to develop the next

generation of financial cyberinfrastructure to enable data science for finance research (see Raschid, Langsam, Jagadish, Wellman, Haas, Krishnamurthy, Vaithyanathan, Franklin, and Wallace, 2013). In collaboration with the DSFin consortium, the REFM Lab will develop the following datasets and tools:

- Creation and curation of **CONTRACT+LEI** graph datasets to represent collections of financial contracts (CONTRACT) and the affiliated counterparties to these contracts, each of which will be identified using a legal entity identifier (LEI). Entities will be identified with respect to their role, e.g., mortgage company, loan originator, etc. Current text-extraction and entity-resolution technologies (see Burdick, Das, Hernandez, Ho, Koutrika, Krishnamurthy, Popa, Stanoi, and Vaithyanathan, 2011) will be tuned to operate over the relevant REFM data collections.
- The **CONTRACT+LEI** graph datasets will be enhanced to develop a *risk profile* for each financial contract and/or counterparty to the contract. We will extend current text extraction technologies to extract relevant attributes 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.). We will construct classifiers to predict risk profiles and we will develop methods to determine the sensitivity of these risk profiles to cases of uncertainty and incomplete information extraction.
- The most significant data science challenge will involve automated reasoning over the contractual features and covenants of financial contracts. As a first step, ontological relationships describing the **CONTRACT+LEI** graph dataset, as defined with respect to the Financial Industry Business Ontology (see Enterprise Data Management Council and OMG, 2014), will be extracted from the financial documents, and a semantic representation will be constructed as a Resource Description Framework (RDF) graph dataset.

5 Budget Justification for Data Science Fellows Program

The budget request for this proposal includes one post-doctoral candidate on a 50% split appointment between the AMPLab in Computer Science at UC Berkeley under the supervision of Michael Franklin and the REFM lab in the Fisher Center for Real Estate and Urban Economics at Haas under the supervision of Nancy Wallace. This position will work on various data science initiatives such as the risk assessment enhancements to the CON-

Table 1: Budget

LABOR SALARIES	Months	Appointment Percentage	Rate per month	Total Cost
Post Doc (Level 2) (50% AMPLab & 50% REFM	12.00	1.00	$3,695.00	$44,340.00
Graduate Student Researcher (step3) academic yr	9.00	0.50	$3,393.00	$15,268.50
FRINGE BENEFITS	Salary		Rate	
Post Doc (Level 2) (50% AMPLab & 50% REFM)	$44,340.00		0.209	$9,267.06
Graduate Student Researcher (step3) academic year	$15,268.50		0.020	$305.37
Total Labor Salaries and Fringe Benefits				$69,180.93
TOTAL DIRECT COSTS				$69,180.93
Fee Remissions	2.00		$7,820.00	$15,640.00
TOTAL COST				$84,821

TRACT+LEI network graphs and the development of algorithms for automated reasoning over mortgage-related security contracts.

The second funding request is for a half-time graduate student researcher who would be expected to be a student in computer science, Haas, or the Economics Department. This position will work on the interface between the development of the economic models and valuation code needed for ongoing research projects and the continued development of the database as new data sources are added to the REFM lab. This position would also include fee-remission support for the graduate student. This person would also be shared between the AMPLab and the REFM lab research initiatives.

The overall funding request for the BIDS Data Science Fellows Program is $84,821. The resources are all intended to provide important support for the student and post-doc related research effort in the REFM lab and the AMPLab at UC Berkeley. The REFM lab will also provide matching support for one additional graduate student researcher. With respect to external funding efforts, Professors Michael Franklin and Nancy Wallace are, respectively, the Prinicipal Investigator and co-Principal Investigator for a component of a grant submission entitled "Community CyberInfrastructure for Financial Data Science" that was submitted to the National Science Foundation in October 2013. The UC Berkeley component of this grant was for $150,000 and includes funding request for a graduate student researchers for one academic year and for one summer.

6 Conclusion

The REFM lab is a uniquely designed research unit with world class data resources, focusing on one of the largest and most opaque capital markets in the U.S. The operation of the real estate and mortgage capital markets in the period before the recent financial crisis is known to have been a key causal factor of the crisis. Unfortunately, even today many of the reasons for the build up of risk in these markets remains poorly understood. REFM faculty have a proven track record of research in these markets and the REFM micro-level database they have developed with their staff provides a unique and customized opportunity to study these markets. In addition, the research partnerships that the REFM and AMPLab faculty have already established both with researchers from other universities and with private sectors partners offer significant promise in addressing new frontiers in data science in support of a more transparent operation of these capital markets (see Raschid et al., 2013).

References

Amirdjanova, A., R. Stanton, and N. Wallace, 2014, The myth of the constant-quality home: A new, unbiased house-price index, Working paper, U.C. Berkeley.

Burdick, D., S. Das, M. Hernandez, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, and S. Vaithyanathan, 2011, Extracting, linking and integrating data from public sources: A financial case study, *IEEE Data Engineering Bulletin* 34, 60–67.

Di Maggio, M., and A. Kermani, 2014, Credit-induced boom and bust, Working paper, Fisher Center for Real Estate and Urban Economics.

Enterprise Data Management Council and OMG, 2014, Financial industry business ontology, `http://www.edmcouncil.org/semanticsrepository/index.html`.

Hartman-Glaser, B., R. Stanton, and N. Wallace, 2014, The distribution of mortgage credit post crisis, Working paper, Fisher Center for Real Estate and Urban Economics.

Jaffee, D., R. Stanton, and N. Wallace, 2011, Energy efficiency and commercial-mortgage valuation, Working paper, U. C. Berkeley.

Kermani, A., 2014, Cheap credit, collateral and the boom-bust cycle, Working paper, Fisher Center for Real Estate and Urban Economics.

Raschid, L., J. Langsam, H. Jagadish, M. Wellman, L. Haas, R. Krishnamurthy, S. Vaithyanathan, M. Franklin, and N. Wallace, 2013, Community cyberinfrastructure for financial data science, Technical report, University of Maryland.

Stanton, R., J. Walden, and N. Wallace, 2014, The industrial organization of the U.S. residential mortgage market, *Annual Review of Financial Economics* 6.