The Great Recession of 2008 and the continuing reverberations in the Eurozone have highlighted significant limitations in the ability of regulators and analysts/researchers to monitor and model the national and global financial ecosystem. While there is considerable activity today in developing more sophisticated models of risk and in developing more advanced regulatory tools, all such work has to be informed by data. Limitations in today's financial cyberinfrastructure restrict the availability of such data to regulators, and to other market participants. These limitations arise due to constraints on data collection authority, due to lack of standards, but beyond that just due to the inherent complexity and volume of the data and the complexity of the analyses desired. Advanced computing technology can help to address many of these issues.

To identify such issues precisely, and the nature of computing solution for each such issue, a workshop was organized under the auspices of the University of Maryland's Center for Financial Policy, funded by the National Science Foundation under grant zzz. The goal of this workshop (and related activities) was to work closely with federal regulatory agencies, academic research communities in computer science, finance, economics and other related disciplines, the financial industry and the computing industry.

BACKGROUND and TUTORIAL on TERMS

Legal entities: generic and instances. LEIs. Sub-entities. Types and examples. Corporations, limited partnerships, subsidiaries. Bankruptcy and resolution.

Contract terms and conditions: generic and instances. Contracts define cash flow obligations and obligations to exchange property. Modifications of contracts. Assignment. Types of contracts: Exchange of property. Leases. Debt. Derivatives. Collateral agreement. Insurance and Guarantees. Rights. Custody? Versioning, annotation. ``Stack of documents'' versus machine readable terms and conditions. Terms and conditions: Price, quantity, obligations, reps and warranties, covenants. Default, litigation status. Bids and offers.

Property: Land, structures, vehicles, machinery, equipment. IDs and attributes (size, location, book value?, use, version, condition (damaged, stolen, new)).

Titles and ownership: generic and instances. Types of ownership. Rights of ownership.

Transactions: Exchange of specific contracts. Clearing, settlement, and custody. Authority to enter into contracts and transfer property. Provenance and data capture. Standardization of transaction histories (special terms of contracts like spread, off-market, unusual settlement dates, ex-dividend terms, central counterparty).

Joint Ventures and Projects:

Pre-trade transparency: history of pre-trade transparency.

Post-trade transparency: History.

Audit trail: Orders, bids, offers, messages. Transactions prices. Price reports. Time lags. Time stamps.

Types of identifiers: LEI = legal entity. Contract. Specific property (real estate, automobile, computer. People (ASK, social security number) and positions (CEO, authorities, conflicts).

Accounting: Property and contracts have accounting properties. Book value, hedge attachments. Type of cash flow (income versus capital gain, expense versus depreciation). Accounting treatment resulting from transactions.

Consistency and Auditability: Property should have one owner. Two sides of a contract should match exactly. One entity's income is another entity's expense. Person cannot hold conflicting positions. Tax compliance.

Idea: Use an agent based game as a vehicle for simulating an economy in complete granularity, including small scale transactions and their aggregation, big events (bankruptcies).

What regulators need: Description of contracts so that risk management can be done. Ability to detect inconsistencies (AIG mark to market, mark to market gains, one firm's income is another firm's expense).

Taxicab receipt: Cab ID, driver ID, enter time stamp, exit time stamp, origin location, destination location, route taken, miles, baggage and extras, fare (amount, computation basis, currency), payment details (credit card number, currency, fees), passenger IDs, purpose of ride, allocation of expenses, weather (?), exceptional events (crash, bumps, complaints), extra notes. Note: All of this information is easily collected automatically by a cell phone, except extra notes and extra passengers.

Model: Paths for evolution of data in the future. Ability to simulate (generate data) versus ability to analyze.

Model: Appraisals: Recent prices, comparability, option model, fundamental model.

Pricing operators and risk management.

Model consistency: Say firm A runs model M_{A} on firm A's data X_{A}. Auditor and regulator should be able to run firm's model on different data, run different model on same data, tweak data and see what differences are. Marginal effects of data on capital needs (stress scenarios) should be comparable across firms. For example, if GS derivatives with AIG reduce GS capital costs in bad states by some amount, then equal and opposite AIG derivatives with GS should increase AIG capital costs by same amount in same state. AIG cannot say payoffs are immaterial in all states unless GS says the same thing.

Models at different firms should place same market values on same assets. But, this often does not happen because of divergence in models used, and/or in parameters set for the models.

DATA STANDARDS EFFORTS

Mission: Collaborative team working under EDM Council to develop standards and operational solutions that eliminate data quality gaps. Data Maturity. Community based, free, open source, non-proprietary, does not depend on particular vender solutions. Looking for open dialogue taking this forward.

Poor data quality made it difficult to identify risks during the financial crisis. Push for global standards for data and business semantics.

Regulatory Requirements for Financial Data Standards and Transparency. Regulators need to classify contracts automatically, classify different types of exposures, aggregate. Need a common language.

What is semantics? How is it useful? Banks and IT firms sit side by side at Canary Wharf. The business of technology, rework, change management. Quality assurance, maturity, emerging in business world. From conceptual model to logical model to physical model. Typically business starts at physical model and works back. In mature business, start at much higher level, recognize need for change management, language interface between actual business and technology.

Example: XML messaging schema design. Good schema for information about various financial securities. Simple spreadsheet of business terms. Collected information about what the business terms mean. Technical people were not able to complete schemas until they understood business terms. Nobody used coupon because meaning was not quite correct. ``FIBO'' (Financial Industry Business Ontology) bridges language gap between business and technology. Business of having a mature process with change management.

To bridge gap, need a single version of the truth. Complete, accurate, extensible. Each new things needs to be abstracted. How does it fit into hierarchy of things? Result is single meaningful concepts. Otherwise, get data in RDFL.

What we want: Create business meanings, i.e., not a data dictionary. In a business language, i.e., not a design. For business people, i.e., no funny symbols, no language to learn, just the facts, boxes and lines. It should not look like a technological tool with specialized syntax.

Uses only boxes and lines and arrows. Human being, legal person, organization, corporation. Legal Person is capable of liability. Artificial Person is incorporated by some instrument. Formal Organizations have formal agreements. Result is Business Entity Taxonomy. Business Ontologies and Operational Ontologies.

Not just swaps and legal entities. Also needs terms and conditions of contracts. How to capture semantics of a contract. Need to identify cash flows of legs of a swap, run through different scenarios. Default, termination, collateral, transfer payments. Financial Cyberinfrastructure for institutional and macroprudential oversight. Semantically defined financial data standards ensures data quality and fidelity between institutions and regulators. FIBO ontologies. Insertion of financial intermediation layer. Semantics harmonizes and homogenizes data. Semantic classifications based on common understanding of business terms.

Protocol for data in motion for trade confirmation. Gives regulators high confidence in the data. Semantics looks at instances of data and classifies them into types of swap contracts. Runs a query looking for transitive exposure. Can look for cascading impact of defaults by different entities.

PK: This is what I call ``compiling data,'' maybe.

How scalable? Looking at venders who might be able to scale semantic statements to hundreds of billions with lightning fast responses. Translate sparkle query into an sql query.

NEED FOR RESEARCH

There will never be entirely clean, accurate, complete, and timely data. The picture has improved recently, and there is more data than before (including new standards for identifying partners to contracts and other standards). Complete provenance is probably never going to be available. In many cases, analysts will have to deal with aggregated, anonymized data and data will continue to have missing pieces. Given this situation, the financial analytics community should strive to get better, more complete data but should also develop capabilities to deal with partial, less pristine data. Confidence levels and uncertainty metrics needs to be developed and then carried through analyses and simulations. In this way, data in a range from complete and accurate to varying levels of incompleteness and uncertainty can be handled in the same framework.

- More robust predictive analytics approaches and processes must be developed—ones that take into account uncertainty and confidence ranges, among other things.
- Hidden networks: Because of incomplete data, parts of the network that should be linked are not, or links may be uncertain (e.g., who are the leaders and who are the followers in a social network?).
- Improved analytical approaches will pay unexpected dividends. For example, statistical analyses can reveal correlations showing a wider network of who is at risk, even without direct connections.

All of these ideas are further developed in five main categories, representing break out groups at the workshop, and described below. It should be noted that there is no claim being made that this set of categories, or ideas, is complete in any way – these are the product of the workshop structure and the expertise of the individuals participating.

1. **Network Analysis and Visual Analytics and Machine Learning and Prediction Coordinators: Bill Ribarsky and Akhtarur Siddique and Amitabh Varshney**
    - Network analysis and clustering and prediction
    - Latent variables and hidden networks and hypergraphs
    - Network evolution
    - Information Visualization, particularly as relates to risk

TOPOLOGICAL DESCRIPTIVE ANALYSIS

As a starting point of working with network- and graph-type data, it would be useful to employ topological analysis and related techniques commonly used in other applications, such as social networks or large scale electrical grid networks. These techniques are relatively well-developed—for example, methods such as topology of nodes and links, degree distributions, k-cores and centrality measures can be applied. However, extra emphasis should be put on link structures and identifying the meaning of links observed. Links should be appropriately weighted by their importance in the application. In addition, highly scalable approaches can be developed by coupling topological analysis with clustering methods (so that topologically coherent networks with substructures abstracted out can be created, for example).

EVOLUTION AND TEMPORAL CHANGE

Time needs to be incorporated as a first-order concept around which efficient and effective temporal structures can be created. This is a general need across financial analytics. There has been work on time-dependence, but this work has focused on shorter time periods or simpler, less detailed time structure. Now much more detailed temporal data are available (for example, large collections of financial transaction data that stretch over years or social media data, such as Twitter, that can be collected minute by minute. Other fields, such as GIS, have realized that temporal structure needs to be addressed as a research challenge. It is worth looking at events and event structures as a general, scalable approach to providing temporal structuring over a range of time scales.

In financial analysis, evolving structures can be a key to developing understanding and response. Transactional or social networks can evolve topologically, for example. In addition, nodes may change character over time and changes having to do with the strength or presence of links may occur. Emerging risk is a main area of interest for financial institutions and for government agencies. This can involve longer term trends that are only apparent when there is enough data over time, or it can involve events that affect underlying assumptions about the stability of financial instruments or transactions. In the worst case, these events may lead to unexpected cascading effects. Event and other temporal signatures can be made hierarchical, which makes the temporal structure scalable and also supports efficient, meaningful drill-down.

- Evolving structures, relations, and trends in numerical variables are important. This is true for evolving networks but also for other types of data (e.g., other data associated with emerging risk).

- With respect to evolving structure and in other ways, temporal analysis and temporal structuring is important. There has not been much done on the latter.

Model the evolution over time of organizations and contracts and interconnections within a global financial marketplace. Large scale event triggered simulations to understand how these systems will perform under stress.

○ Parsimonious representation. Need a data model for states and state changes.

○ Triggers.

○ An example of contract evolution - netting payouts through novation (replacing one party with another).

DIMENSION REDUCTION AND SCALABILITY

Inevitably financial networks, social networks, and other types of relevant data are multi-dimensional. They can even by high dimensional involving hundreds or more dimensions. There could be hundreds of customer categories multiplied times many different types of financial products, for example. In addition to get a complete picture of emerging risk, it may be necessary to combine data from different sources, which will further raise dimensionality. (For example, financial data and social network data could point to the same phenomena such as housing prices and mortgage default rates in a region.) Dimension reduction methods need to be developed and applied to reduce the high dimensional space to a manageable number of dimensions for deeper analysis. Techniques such as clustering, multidimensional scaling, principal component analysis, and other can be applied. However, the key is to produce a dimension reduction that is understandable in terms of financial analysis; often existing techniques produce mathematically transformed dimensions that are hard to understand in practical terms. Appropriate interactive visualization techniques can provide important elucidation here.

Scalability has been discussed throughout this document, but a special aspect should be mentioned here. In addition, to scalability dealing with *growing amounts* of information, there is scalability that deals with *growing complexity* of information. High dimensional spaces are an example where this second type of scalability needs to be considered. Another example, would be complex processes, especially if they require more complex reasoning to understand and manage them (for example, complex and many-faceted financial transactions). Scalable techniques in this sense are important, too.

MORE COMPREHENSIVE SIMULATION APPROACHES

Based on what was discussed during the workshop, there appears to be a need for more comprehensive, real-world simulation approaches. According to our assumption above, there will always be an uncertainty and range of possibilities with varying confidence in the data. These aspects should be brought along in the predictive analytics and simulations that are applied. In addition, one cannot know beforehand just what situation may arise. To plan effectively for a range of possible situations, one should run an ensemble of simulations over a range of inputs. The range of uncertainties in the inputs, also requires an ensemble of simulations. In addition, real world situations cannot usually be simulated using a single model. A potential financial crisis may stem from several factors and thus involve several interacting models. The general need for ensembles of simulations involving multiple interacting models is starting to be worked on in other fields. For example, severe coastal storm simulations that involve models for storm surge, hurricane winds, heavy rainfall, flooding, and people movement are being considered. In the area of the stability and resiliency of multiple interconnected, critical infrastructures (e.g., electricity, water, gas, transportation, telecommunication), work on system of systems approaches are farther along.

Ensembles of interacting simulations (which may need to be run hundreds of times to cover a range of inputs) may be computationally daunting and expensive. It is unclear how much funding would be available for larger scale computations, so computational feasibility is an issue that needs to be addressed. But many financial simulations are substantially less complex than, say, high resolution physical simulations (e.g., storm surge models). In addition, there are system of systems approaches that concentrate on the interactions among the different models rather than on the realism of the individual models. (Critical infrastructure ensemble approaches that look for cascading effects are like this.) These can produce quite useful results without being that expensive computationally. In addition, borrowing from research in computational methods can pay off here. Often simulations can be parallelized or otherwise made significantly faster with appropriate computational approaches. The bottom line is that the simulation approach must be valid for real-world situations and must take into account the uncertainty in inputs and underlying assumptions. Otherwise the results may be misleading and, even worse, give a false sense of security.

VISUAL ANALYTICS AND AUTOMATED ANALYTICS APPROACHES

The problems described above in predictive analytics, financial transactions, emerging risk, and other areas are complex and often large scale. Problems such as these require insertion of human reasoning, interpretation, and decision-making at just the right time. The analysis cannot be fully automated yet automation is needed so that the human analyst is not overwhelmed or the process made too expensive because of the need to apply large amounts of expensive human resources. Visual analytics provides a set of approaches that reserve for the human the aspects that the human is good at and for the computer those aspects at which it excels. VA then provides an interactive interface where the human and computer can work in collaboration. In a VA framework, automated techniques, often data-driven, are integrated with interactive visualization. For example, machine learning techniques that assess unstructured text or multimedia collections or even streaming content from social media or online news and blogs have been integrated into the VA framework. Various methods involving event detection, other temporal signatures, and the identification of evolving networks have been or could be integrated. Interactive visualization then provides to the analyst contextual overview, exploration, and the ability to discover interesting behavior or insights. Events or trends associated with emerging risks, for example, could be identified, given meaning, annotated, and then followed in greater detail. Predictive models or simulations could be assessed or even steered through the interactive interface. There has already been a fair amount of work done applying VA to financial analysis and this could be the basis for even deeper work. Collaborations with the VA community to more clearly describe the problems discussed in this workshop and the analytic approaches that could be used would be quite worthwhile.

Contractual Reasoning and Semantics and Taxonomies and Metadata - Coordinators: Benjamin Grosof and Leora Morgenstern and Frank Olken and Andreas Cali

- Parsimonious/machine representation of a financial contract
- Contract evolution
- Taxonomies and ontologies and poly-hierarchies

Validation and reasoning
Agenda at high-level:

use cases

building blocks semantic abstractions and representation
- roadmap

open data to make available to/by researcher community

interleaved in the above:
- general/misc. discussion
- pragmatic implementation into deployment environments
  . integrating math of specialized kinds

%%%%

%%%%
Semantic "Building Blocks": (useful across many use cases and tasks)

roadmap strategy of how to progress on these semantic "building blocks"
- start with the "central" kinds of info:
  . transactions
  . contract terms
  also:
  . counterparties (referenced by the above)
  . portfolio holdings (derived from the above)
  . obligations and rights (derived from the above)
- develop ontologies of
  . state attributes mentioned directly in the above
  . key events, eg contingencies and actions in contracts
- develop other core ontologies of basic business etc.
  . collateral, assets, liabilities
  . obligations, rights
- semantically represent the central info
- then expand scope and iterate refine
  . follow steps of relationships/linkages/references

%%%%

o data quality and validation
  - both input and output
  - eg trade fail rates
o performance metrics wrt the above, eg:

- cost
- degree of automation

concepts:

o supply chain analysis (ie supply chain of financial info/analysis)
o operational value add wrt data mgm and business processes
  - within firms
o financial transactions lifecycle analysis
  - financial research, trade, clear/settle, to value, report, comply
  - evolvability

%%%%
Other General/Misc. Discussion:

how to model rules (not just ontologies) -- for contracts and regulations

aim for 2 or 3 use cases

semantic processing architecture as part of the above

repositories
- standard contracts --> contingencies and state variables referenced
- bankruptcy proceedings

deontic: rights, obligations
temporal

integrating math with the logical

integrating probabilistic

embedding of contract instance into surrounding
  master agreements, law, business policies, business processes

%%

REA - Resource Event Actor - ontol of transactions [Mike B.]
  by Bill McCarthy Mich. State U.

relationships [Mark Flood]

loans not nec'ly trades -- eg esp. at banks with housing

strong need for defeasibility to represent contracts and change
- dynamics over time, of info state
- conciseness and evolvability of specification

negotiated settlements and workouts

%%

extract relevant state-space dimensions from contracts and transations
- eg contract depends on 30-day Libor or Deutschmark exchange rate
- attributes
- key events, eg associated with maturation, settlement, trouble in flavors

Allan Mendelowitz: can reduce cash flow patterns to only ~30 standard contract types
- 99% of contracts
- with high level of precision

happy path vs. trouble cases

there are possible copyright issues wrt non-standard contracts
- tho' there's a good argument for fair use if came fm req'd disclosure to govt
- probably not a critical issue initially, but post-ers must beware

could design better automated contracts
- for simulation
- for deployed implementation in the business etc. world

subsidiaries: there are various scopes of credit support by the parent

"wildcard" in practice: escalated workflows by the players

efficient large-scale methods for expressive rules/ontologies

bit of recap:
- understand reality as it evolves
- understand relationships/interconnections in overall fin ecosys
- expression in semantic language
- as part of the above, granular/modular development of building blocks
  . fundamental contract abstractions and types
    o eg equity = ownership; debt = obligation; hedge is compound
  . cash flow patterns and
  . state space attributes
  . key events
  . key contingencies
- as part of the above, shared community repositories

Semantic Search, machine readable contracts.

● Retrieval
○ Can use FIBO tags.
○ Can use automated classification of the type of a financial instrument described in a document.
● Machine readable contracts.
● Information extraction?

Lucian Popa, IBM Almaden. Midas. Entity Extraction and Integration from Unstructured Data.}

What is Midas? Effort toward large scale extraction, integration, and analysis of data. Create software tools for use by us and others. Two sources of data: SEC filings and social media (Twitter).

Why public data? Complements internal data in good ways. 360 degree view of entities.

Investment Decisions Applications: analyse companies based on financial information and counterparty relationships (lending, investment, ownership).

Social media: Companies can offer promotions based on social media, brand management, sentiment analysis.

PK: Have you checked your output against other databases which give structure picture of the same information? Execucomp, Deallogic, Compustat, etc.

Jerry Hoberg, UMD. Life of Finance Faculty in this Brave New World of Data.

Example of testing a finance theory. SEC Edgar disclosures.

With Max and Christopher Ball, who works at company Meta Heuristica.

Issue of financial constraints. Companies hoarding cash, not investing in new projects. Is it financing frictions, ability to tap debt and equity markets? Instead of use regressions, use statements by mangers of the firm. Use text to broaden implications of constraints.

10-K is very structured. Has a Management Discussion and Analysis section. First analytic tool needs to extract MD\&A section, which is a subset of Item 7. Want subsection dealing with Capital Markets and Liquidity. Are the mangers of the firms telling us that they are having problems? Law says must tell about liquidity situation and resources you intend to tap to solve problems. Search for occurrences of specific words like ``delay'' or ``abandon'' things related to real investment like ``construction.''

Some writers are more concise and others are verbose.

Training sets, false negatives (convoluted language), boiler plate content.

Results show jump in mentioning of key words in 2008. Firms indicating constraints greatly reduced capex and R\&D spending. Firms also talk about related topics that seem intuitively plausible.

Information Integration and Entity Resolution and Information Quality Coordinators: Lucian Popa and Joe Langsam and Rachel Pottinger

- Human language technologies and document collections
- Information extraction and entity resolution
- LEI and post LEI challenges

Attendees: Lucian Popa, Rachel Pottinger, Gerard Hoberg, Joe Langsam;

Background: There is a plethora of publically available data about financial market participants. Much of the data is unstructured or partially structured. The quality of the data varies widely and provenance is often missing or suspect. The first steps in extracting information from this data require the knowledge of to which entities the data refers and a unique name or identifier for the entity. This unique identifier enables the selection and linking of data common to a given entity.

The Legal Entity Identifier (LEI) is proposed to be a 20 digit alpha numeric indicator that is unique for each eligible legal entity. The LEI project initiated in the U.S. by the Office of Financial Research is an international project supported by U.S. and international regulators. It is proposed that the LEI be required for each major participant that engages in trading and will be required to unequivocally identify the parties in the transaction. LEIs, for example, can be great for clear identification of participants in contracts and for establishing transaction (or risk) networks.

However, the LEI contains little or no embedded intelligence. Associated with the LEI will be a thin set of identifying information required when a legal entity applies for a LEI. It remains unclear if the LEI, in addition to being required for a trading participant, will be used on other documents.

Research Challenge: The challenge for computer scientist will be to develop tools that reliably can tag the correct LEI to the data in a document that refer a financial entity. Copies of 10-K and other corporate fillings are available through EDGAR. It would be desirable to tag these fillings. One would also wish to see analyst reports and trusted news stories tagged with the appropriate LEI. It may also be of value to tag commentary found through social networks.

The development of software to identify entities for tagging, the structuring of the LEI tags within documents and the development of databases with trusted LEI tags are research problems. The identification of a legal entity in a document poses several problems. Several legal entities share common names and a single legal entity may be called by many names. Morgan Stanley, Morgan Stanley and Co., Morgan Stanley and Co. Incorporated, and MS&Co. each are used to refer to the parent company. A news article which refers to Morgan Stanley, very well could be referring to a subsidiary or a branch office, each of which often are called Morgan Stanley. The term Morgan Stanley might refer to the Investment Banking group in multiple subsidiaries or to the retail business. A Google search for "Morgan Stanley" subsidiaries returns: http://www.morganstanley.com/about/ir/shareholder/10k113008/dex21.html

This is a 26 page document listing Morgan Stanley subsidiaries as of November 30, 2008. It is not uncommon for a multinational corporation to have a complex corporate structure with multiple subsidiaries and branches. A news article may refer to several of the subsidiaries without specifically identifying them. It often will be unclear to precisely which legal entities a document is referring. It may be desirable for the tagging algorithm to assign both a LEI or number of LEIs together with a numerical indicator of confidence.

It is likely that there will be multiple valid notions of entity, each appropriate for use in some context. It will be important to have published information relating legal entities. This secondary information may be brought to bear, as needed, when aggregating information or linking multiple databases about a legal entity of interest.

In addition to the computational questions, we also have domain specific open questions that must be addressed in this context. What are the important data sources for LEI tagging? Which, if any of the data sources, would it be best to have LEI tagging at origination? Should there be a collaborative effort for LEI tagging of selected data sources?

Once the ability to identify entities and tag them with an LEI exists, the question arises to the storage and updating of the tags. Should there be a common database of a selected subset of publically available documents that have been tagged with the LEI? The filings in EDGAR would be an obvious candidate. How can we come up with a set of publicly available LEI tags that are trusted? Would it be effective to have crowd sourcing with researchers submitting documents that they have tagged? Would it be preferable to make the tagging software available? How should tags be updated as more information enabling better identification of the entity emerges? Should the database contain the provenance of the tag as well as that of the document being tagged? The structure of the tags within a document and the required supplementary information about the tags as well as the software to affect the tags remain to be determined.

In addition to LEI and financial product identifiers, risk factors and contract terms including covenants are important candidates for developing identification keys or tags. It is worth exploring the value of having unique naming conventions (tags) for key people (corporate directors, CEO, CFO, Federal Reserve Governors…) in addition to legal entities? Tracking the person will not be any easier. For one thing, people move, for another, sometimes there are people who are involved with more than one LEI. Tracking people is important because if there's a news story about a person, it impacts that person's company(ies).

Even with LEIs in place, it is likely that newspapers, SEC filings (EDGAR: 10Ks annual, and 10Qs) will be tagged with the "main" entry, but the reference to the competitors and other mentioned entities will not. As we extend the notion of identified entities, it is worth exploring whether there is a way to identify key terms in contracts and such, e.g., "covenants?" So can we create a library of saying "who is subject to failure because they're not following this covenant"? In particular, we want to capture what the key financial elements are in a big document that makes it interesting and precise rather than just boiler plate. It should be possible to develop and publish software that performs entity extraction and LEI linkage, given any arbitrary document. With such software in place, it would be possible for the filer, the regulator, or any interested third party to run each report/filing through the program and populate the LEIs.

PK: In post LEI world, will large banks put everything in one pot, without sub-entities or hierarchies. Perhaps they hide exposures with internal swap agreements. Immediately opens up problem of sub-entities, even trading desks.

PK: Post LEI: Identify types of market participants based on whether strategies are stabilizing or momentum.

PK: Finance data: Rows and columns of prices, transactions, orders, positions, accounting variablles, macro variables. Text data is new. Also housing data, including house and mortgages. Rows and columns: Problems of accuracy and coverage, e.g. reporting of long not short positions, futures reported inconsistently as contracts, market value, or notional value. Time stamp inconsistencies. Entity name inconsistency supposedly goes away with LEIs

PK: Example: Identifying index arbitrage trading strategies.

PK: Stress tests: Get data bottom up in granular form or get macro aggregates. How to insure the data has consistent meanings across banks.

PK: Original problem: How to represent contract terms in a consistent manner so that risk management algorithms can be run on it. Recursive manner for defining cash flows.

Louiqa: Too much manual entry and manual propagation of individual data points, not enough machine propagation.

Post LEI

Classification schemes for entities.

- Tools to manage taxonomies.
- Ontology to describe organizations, regulators, jurisdictions, etc.
- Data curation
  - How to support financial annotation?

■ How to support regulatory annotation?

■ Confidential annotations.

Most US publicly traded organizations are now required by the SEC to provide financial reports in a standardized format (XBRL).

○ (This was a gradual transition since 2010, expected to complete by 2013)

● Still, the XBRL requirement only applies to the recent financial reports.

○ There is relevant qualitative data that is in less structured formats (HTML, XML, text): merger/acquisitions, material events, appointments, loan agreements.

○ Plus, "old" financial reports, which may be needed for auditing purposes or for research, are unstructured as well

● Challenges for unstructured data:

○ Information extraction

○ Entity identification and record linkage.

● Beyond publicly traded companies: municipal bonds and disclosures regarding municipalities and their associated events (e.g., bankruptcy)

○ Data is not in XBRL, but in PDF and text

● Controlling for various features, can we build a model from this data for prediction? what else?

○ Modeling?

■ Models or analysis to understand "local" risk associated with one entity (i.e., company, municipality)

■ Models or analysis to understanding systemic risk (across the system)

Financial Supply Chain

The ability to track financial products end-to-end along their supply chain. An example is the mortgage supply chain which must take into account the behavior of he mortgage throughout its life. The mortgagemay first be warehoused. The mortgage may be eligible for FNMA guarantee. It may or may not have other guarantees. It may or may not qualify to be securitized. It may be securitized and then become delinquent. The dlinquency may be cured. The mortgage may or may not go into default and to foreclosure. We need the ability to track entire pools of mortgages. The tracking starts from the process initiating the mortgage until its final disposition.

○ Price indexes

○ Modeling the performance of the security using real time data.

■ Analyzing sensitivities of the security using a variety of models

■ Estimate Liquidity (f(trading volume, bid-offer spread)

○ Modeling the financial health of a community.

Large scale spatio-temporal modeling

Produce a "heat map" of our financial system transactions, very much like a global climate map, so that we can pinpoint areas of high activity or vulnerabilities based on topology, warfare, etc.

●

● From the 1960s to today we have seen the value of housing stock in communities across the US change, with price fluctuations, and in some cases, in unanticipated ways. An example is the value of real estate across inner city communities.

●

● Analysis of information encapsulating LEIs so as to monitor and identify hidden networks that provide additional channels for the propagation of systemic risk. Examine the collection costs as well as the benefits of additional units of data, including public and historical data. Develop a range of analytical methods including network analysis, feature identification and machine learning based methods.

1. **Social Media and Crowdsourcing and Markets Coordinators: Johannes Gehrke and Louiqa Raschid and Michael Wellman**
    - Social media modeling and prediction
    - Crowd-sourcing
    - Market mechanisms and prediction markets
    - Agent based models

    It is important to understand and use social networks and social media. This can lead to techniques that can produce better informed citizens and to understanding of when and how citizen confidence in the financial system erodes.

Wisdom of the marketplace

● Crowdsourcing
○ Crowdsourcing the value of a financial instrument.
What does value mean if there is no market?
○ Crowdsourcing the rating of a legal entity.
○ Crowdsourcing the LIBOR rate.
● Social Media
○ A significant amount of human activity is captured in new media - social media and social networks as well as in traditional media - news articles and web pages and document collections.
○ We can construct models of such activity and they can be used to understand and explain and predict activity in financial markets.
● Create new methods to efficiently monitor data streams of financial activity and human activity. What are people saying about topic X in a set of languages?

Network Visualization

Example - finding ***crowded trades*** = many portfolios exposed in the same way to the same contingency:
● Need to model the space of contingencies

- Need to measure specific exposures - best done at the contract level
- Current measurement mostly aggregate to firms/portfolios
- Data exploration; pattern detection; data mining?

Possible enhancements:

- Collecting additional data -- which would be most useful?
- Improving data quality - outlier detection? data validation?
- Improving data modeling to support large-scale visualizations
- Interactivity - how to make this performant?

Example Application: Investment Decisions based on Regulatory Filings. Annual reports, proxy statements, loan agreements, insider transactions. Reduced into a smaller set of entities. Find 33,000 key officials who work in 2500 financial companies. Result is integrated entities and relationships. Example of relationships of Citigroup. Banking subsidiaries, overlapping board members and officers, major companies invested into, institutional holdings. Periods of time different officers and directors held positions, what positions they held at other companies. Holdings of every person. Aggregated insider holdings. Loan agreements and lending activities. Has 82 loans with Bank of America. Can see data on individual loans. Can see identities of lenders, sizes of loans, other lenders.

Different type of analysis at system level. Cite of paper by Kritzman et al. Analysis driven by explicit linkages gleaned from public data.

Challenges: Extracting data from loan agreements is difficult. Tables allow extraction of how much each lender lends. People appear in many places as text. Jamie Dimon. Need to integrate into one object.

Fusion of Data: Want to reduce holdings to latest holdings.

Analyze social media messages so companies can find out what people say about them. Apply same type of entity extraction to create social medial profiles. Companies get full picture of what customers say.

Also working on municipal bonds.

Two-Sided entity resolution

The CFTC calls for reporting of a swap by one party.
There will be multiple swap data repositories.
All data about a single swap must be reported to one SDR.
What are the potential problems?

Why is it difficult to build a comprehensive profile for a single swap? Give an example.

Two sided reporting, together with other changes, would permit automated term matching to discover discrepancies (fraud, high levels of operational errors...).

Joe - You mean discrepancies between both parties? Give an example.

Consider a highly structured swap that will not be subject to clearing. (Consider a highly structured mixed FX credit index (Itrax, CDX IG) default swap or combination of FX credit index option triggered when the index spread exceeds a certain

Instrument type representation and classification

Representation perspectives:

● Financial: state-contingent cash flows; derivatives and dependencies across securities

● Accounting: chart of accounts; valuation methods

● Legal: master agreements; structured legal semantics

Classification:

● Defining equivalence classes - how?

● Automated vs. curated classification - pros and cons?

● Versioning of types as markets innovate and securities evolve

Mark - do you want automated classification? if so from what data?
Or are you suggesting to develop a taxonomy and schema? If the latter then it is interesting but it would not be CS research since most of the knowledge is domain specific and then you have to come up with a good schema. Schema design is hard to do well but it would be hard to get NSF funds for this.

Disclosure = Privacy (LEI) and Proprietary (instrument)
- Provide an example scenario.

Investigating networks requires knowledge of trades.
- Privacy: What information about an individual or other industry participant needs to be protected?
- Proprietary: What information about an instrument should not be disclosed?

-When does proprietary data stop being proprietary?

Banks are extremely concerned with protecting their client list

An old (3 years, 5 years...) report of trades a bank did could reveal aspects of a relationship that the bank would not want to share.

1. **Data Representation and Model Management Coordinators: Phil Bernstein and H.V. Jagadish and Pete Kyle and Amol Deshpande**
    - Data models and schema and metadata
    - Representing financial models as first-class data objects
    - Reconciling different perspectives in representation: financial, accounting, legal, …
    - Error correction, and propagation of corrections through derived data
    - Privacy

Support data values that are computationally derived, typically through the execution of a financial model. The model may require multiple inputs, which are often not recorded in current practice. Some sort of provenance-based solution may work. Some inputs may themselves be the result of running other models. The model itself is typically a regression model, often comes from a small finite universe, and may possibly be identified by name. Due to the possibility of "dialing", at least for some models, it may be important to record inputs used. (We extensively discussed the example of LTV).

Provide database support for financial models as first-class data objects. See arguments above.

Concepts like LTV or ``balance'' are results of models calculated differently in different banks, which do not know what they are doing. All use stupid versions of the same class of models. Data are highly massaged before they reach the Fed.

How complex are the models? They are simple (even dumb). Term structure models are fairly sophisticated. Many pieces of risk management models are such the we have poor information about the assets. Very low standards. Do not hit known empirical regularities. Inflexible functional forms do not fit market.

Deeper methodological question: Numbers that are not data but rather derived from models. If you do not say what your model is, along with other inputs and other parameters, simply noting model output simply does not do it. You end up with a complex story for what is a simple number.

Models operate on data, but the data itself is a calculation based on a model. Nobody knows the details of how the models work.

Example: Loan might be indexed to LIBOR, need to get LIBOR input since not supplied with data. There might be an error in the mortgage date, resulting in an error in the calculation.

Many important financial values are restated, often multiple times. E.g. unemployment rate, retail sales data, GDP, etc. Many models will take these values as inputs. When the value is restated, it is important to reflect this in the derived numbers.

Some model inputs are arbitrarily determined scores/labels determined by "experts". E.g. bond rating, credit score, … Need convenient way to perform "What if analyses" with changes to these values.

Support fusion of data from multiple sources. E.g. credit information may come from credit bureaus, Corelogic, and LPS. Exactly what is reported may be different for each, and values may not even match exactly, but being able to merge approximately is important.

Should academic models fit into a specific format so that they can be shared with others and checked for consistency.

Free form data should be allowed but punished. For example, firm enters into new kind of derivative, scans and registers legal documents, enters later into binding descriptions of what terms and conditions mean.

Record crucial distributional information, e.g. when it is bimodal (different interest rate/points choice on same date, or different TVA). Making uniformity/normality assumption (which is typical) can lead to poor model conclusions.

Accounting view of data is based on cost, financial view on valuations. Be clear which, or some other third, is being represented.

End-of-month reporting loses short term fluctuations.

Upon acquisition, people often leave. Also, the acquired company may have committed fraud or had poor practices. As such, there is usually very poor interpretation of data from acquired company. E.g. Countrywide in BofA.

E.g. LongBeach in WaMu and in turn, WaMU in JPMorgan.

Supply chain of data: Not one but multiple supply chains, even within the same bank. Different ways of defining terms, e.g., commodities different from bank. Orders which are made and pulled and not executed are not likely to be made available outside the trading desk. They do not want other people to know about it.

Where does data get touched over its life? Sales desk. Touched by trader. Multiple valuation models. Sometimes only simplified models are needed. Volatility means different things in different contexts, since model specify. Yield does not have same meaning. Only dollar price has same meaning. Collateral management people need to see contract terms and understand valuation of asset.

LTV could be represented as a formula, so that you would know what you were dealing with. When B of A buys Countrywide, they do not get the people at Countrywide to explain what the data means. Countrywide data is very questionable. When banks acquired, people are layed off and have little incentive to share knowledge about data. IT systems are often not merged into one system. Instead, just merge the final reports from separate systems. Result is fragmented IT. Moved WaMu IT from Seattle to Chicago over 18 month period. Many of WaMu's previous acquisitions had not been integrated well.

Auditing firms have credit modeling divisions that come in and look at credit models, but they do not probe deeply. Example: Banks put out rate sheets, which define a contract space. Fees and assumptions and points. Points are scraped out, resulting in portfolio of 3% loans and 7% loans originated on the same day. Borrowers have been sorted into these groups by self-selection, as a result of which borrowers are different. Solution is to generate data further upstream, before the points were stripped off.

Data provenance problem in that we get data from aggregation and do not know what you are getting. Small and medium sized banks use spreadsheets. Even big banks collect data from aggregators, who spit the data back to the banks. Only top-tier banks do things for themselves. Models allow ``dialing'' to change what comes out. Many cases about AVMs, automated valuation models. Many companies sell valuation models. People who build data no longer with company, therefore company does not know what data means.

Fed moving in direction of obtaining much more disaggregated data. Is XPRL a useful standard for quarterly financial statements? Nobody knows. FPML is financial products markup language. Pretty good uptake. Supported by ISDA.

Bank can be brought down by loan portfolio, especially unique individual very large loans in investment banking deals. Loan document is a telephone book. Covenants do not make it into the valuation models. Failure might take down the bank. First Boston failed due to bridge transaction. Do not have good documentation of the loan models for non-standardized loans.

Case Study (Mortgages):

Private label mortgage market has broken down. Agency market (Fannie and Freddie) is the mortgage market today. Long supply chain with non-standardized data flows. Contracts poorly understand. Makes it difficult to aggregate and do research. Same problems exist within institutions. Original contractual obligation. Performance data. No unique loan identifier (like a CUSIP) which can be used to track the same loan. Credit derivatives written on top of underlying. Difficult to track data. Mortgages are renumbered at every stage. Tracked by pattern matching the fields. Contracts travel through supply chain in about four months, very fast. Acted like put-back options did not exist. Data in banks was hopeless. Modeling done on spreadsheets. Separate desks for pre-payment and defaults using spreadsheets.

Banks and brokers (now in house). Lenders are first movers. Set menu of mortgages. Rate sheets (menu of contractual features like coupon, maturity, fees) reset five times per day. Borrowers self-select into contracts which make sense for the borrower. Lenders compete from rate sheets. Now loan officers inside banks get the rate sheets. How prices are problematic.

Mortgages are two contracts: bond (promissary note) and lein (collateral is real property, which is attachable after default on note). Leins are reported documents, done differently in 50 states with no standards. In some states, hold real title, in others just a lein. Data comes from county reporters offices, which are land record offices in U.S., a 100\% paper system, delivered by hand or PDF, charges \$75 to put into data base. All house price data comes from county reporters' offices. Updated with transactions and housing remodeling. Remodeling is 2\%-3\% of GDP. Data exist in PDF files. Send PDF files to India for keypunching on monthly basis. Data vendors and modelers use exactly the same set of data. Deal with data infrastructure technology where data is in shoe boxes. MERS designed to avoid dealing with county reporters offices. Both note and lein have to be re-assigned. MERS created a club with MERS as the lender in the data record for 60 million mortgages. Transactions recorded in MERS, not in county reporters' office. $65 went to MERS and not to reporters offices. So reporters offices cannot update their files, use ancient representations of what condition houses are in. Need a nationalized standardized manner for keeping land records and house prices. Has led to tens of thousands of court cases because cannot establish who is the real owner.

Zillow gets data from transactions and county recorders offices. DataQuck, CoreLogic, and ???. Trulia and Zillow are new players. House price data do not match what the houses actually are. Automatic valuation models inside banks purport to tell banks what the housing prices are. Price default risks off these data. Our U.S. records are in a state of shambles.

Bonds created from mortgages have their own performance data. Do not see actual trading prices anywhere along the chain, even though SEC registration occurs as you go from sponsor to depositor. SEC considers sale prices to be privileged information. Do not see prices of mortgages, pools or mortgages, or bonds. Certificate waterfall. REMIC is re-securitized structure bond product. Transactions prices exist internal to banks but not reported publicly.

Need a unique ID number for every mortgage loan. Need to be brought into pool. Performance history data need to use this ID. Need to be able to go from loan to pools. Need prices and trade volume data. Tranches. Credit default swaps. Logic of cash flow in tranches is not kept; need to read hundreds of pages of documents.

PK: Wonders what price data would mean if you do not know what the security behind the tranches means. Similar issue for all reported derivatives transactions. Also have charges for two-sided counterparty risk built into prices. Needs to be backed out in order to make prices comparable.

Other Use Cases:

Big 3 business tasks:

- transaction processes (eg, trade, loan, withdrawal, order, settle)
  - related: lifecycle (see below)
- risk management (involves analytics heavily)
- oversight for regulation (involves audit trail, reporting). E.g.,
  - transparency
  - systemic risk
  - market surveillance
  - idea: cooperate with FINRA (tho' their data quality has problems; need sem's)

Open Data to make available to/by researcher community. I.e., community resources to support research

- financial counterparty (-network) reference DB
  - part of mission of OFR
- financial instrument reference DB
  - part of mission of OFR
- repositories of contracts, esp. of various types
  - start by analyzing the original textual forms
  - eg Edgar, ISDA, FPML, venture capital (Oliver O),
- ABA e-contracts, CFTC, SEC prospectus, master agreements, ...
- mortgage loan doc's, Bloomberg, Reuters, IDC (Interactive Data)

Can use ontology extraction tools

OFR end:
- Research and Analysis suborg
- see directions in Annual Report released today 7/20/12
- still working on how to aggregate and publish the various
  publicly available data
- idea: maybe an OFR program similar to NSF CRII

take a page from Linked OpenData

Sloan Foundation

wikidata / DBpedia

see linkedct.org

candidate data sets:
- Lucian Popa: IBM crawled SEC data, pre-cleaning, in JSON
  . possibly also for universities: cleaned data
- Nancy Wallace: contract documents
  . maybe: for universities (unrestricted is quite expensive)

dream/vision: PubOFR (or PubFin) akin to PubMed

hosting on a cloud data platform
- Phil Bernstein: eg Azure probably
- OFR does not have outside-the-firewall stuff yet

markup of doc's with structured data eg FIBO
- Annotator from Protege stuff by Mark Musen group at Stanford
  . Nigam Shah contact
- tool being dev'd in Karsha (Louiqa Raschid)
- Semantic MediaWiki+ (Vulcan's) -- popular semantic wiki

develop financial-relationship network graphs for systemic network analysis
- have weights

More from use cases discussion:

o asset management
o capital formation
o strategic behavior
  - investor objectives
  - supporting strategic network analysis

<list of use cases from 2010 NSF Workshop report: >
1. Visualization and analysis of a financial counterparty network
2. Knowledge representation of a financial contract
3. Implementation of a “living will” for a large financial firm
4. Fostering an ecosystem of credit analysis
5. Reasoning over financial contracts for completeness and integrity
6. Privacy and trust: multiparty sharing of confidential financial data

there's also a "geek" axis of technical aspects used in the above use cases

ideally use cases from both equity, debt, hedging
analytical use case
- both microprudential and macroprudential
- state-contingent cash flows