The exponential growth of machine-readable data to record and communicate activities throughout the financial system has significant implications for macroprudential monitoring. The central challenge is the scalability of institutions and processes in the face of the variety, volume, and rate of the “big data” deluge. This deluge also provides opportunities in the form of new, rapidly available, valuable streams of information with finer levels of detail and granularity. A difference in scale can become a difference in kind, as legacy processes are overwhelmed and innovative responses emerge.

Despite the importance and ubiquity of data in financial markets, processes to manage this crucial resource must adapt. This need applies especially to financial stability or macroprudential analysis, where information must be assembled, cleaned, and integrated from regulators around the world to build a coherent view of the financial system to support policy decisions. We consider the key challenges for systemic risk supervision from the expanding volume and diversity of financial data. The discussion is organized around five broad supervisory tasks in the typical life cycle of supervisory data.
1) Background

“Big data” means more than simply larger storage requirements, or collecting data from social media platforms with millions of participants. “Bigness” is a symptom of scalability issues in one or more dimensions — the four Vs of volume, velocity, variety, and veracity (IBM, 2016). “Big data” is a misnomer, suggesting that “bigness” is an intrinsic characteristic of a dataset. Rather, bigness describes the relationship between a dataset and its usage context.¹ A dataset is too big for a particular use case when it becomes computationally infeasible to process the dataset using traditional tools (MongoDB, 2016). Scalability is a binding constraint for any process, of course, if extrapolated too far. Big data can create an inflection point where differences in scale imply transformational differences in the costs and benefits associated with using the data.

Big data is not the only challenge facing financial stability monitors, who potentially have the entire financial system in scope and face data scalability challenges on many fronts.² Fundamental economic factors, such as macroeconomic uncertainty, credit conditions, market volatility, liquidity, and contagion risk, remain core concerns. On the measurement front, the salient supervisory challenge is often too little information to integrate, analyze, and make actionable, rather than too much. Some official pronouncements, such as the Office of Financial Research’s recent Financial Stability Report (OFR, 2015), focus on the first-order task of filling data gaps: Do supervisors lack adequate data coverage, data quality, and data access to achieve their mandates? The existential questions of data availability must take priority over the scalability questions of data management. On the other hand, the Bank of England’s (BoE, 2015) recent One Bank Research Agenda identifies big data, including news feeds, social media, and transaction-level trading data, as potentially important untapped lodes of information for central banking research.

The proliferation of measurement technologies in all sectors of society, much of it captured through our interactions with the Internet and cellular networks, means that many industries are confronting big-data scalability issues simultaneously. These challenges are revolutionizing a diverse array of fields, including official statistics (Kitchin, 2015), scientific research (Hey, et al., 2009), retail sales (Manyika, et al., 2011), health care (Horvitz, 2010), and even the arts (Somerset House, 2015). The financial services sector is not immune. Casey (2014) identifies six general types of data that are coming together to compose the big data inventory for central banks:

¹ Diebold (2012, p. 4) notes, “…someone reading this in twenty years will surely laugh at my current implicit assertion that a 200 GB dataset is large,” further observing that the Large Hadron Collider, the world’s largest particle accelerator, today generates a petabyte (10¹⁵ bytes) of data per second.
² We are more concerned with data management issues rather than specifics of legal supervisory authority. We use the following terms interchangeably to avoid cluttering the text with clarifying language: “financial stability supervisor (or monitor),” “macroprudential supervisor,” and “systemic risk supervisor;” these refer to national or international authorities responsible to maintain awareness of and respond to financial-sector stresses and crises.
macroeconomic, survey, financial institution, third-party, micro-level, and unstructured (examiner reports, social media, etc.). We focus on official data collections of macroprudential supervisors, but many of the issues discussed extend to other data types.

The upshot is that new datasets are emerging from a variety of sources, including official collections such as stress-test exposure details, third-party vendors such as market intelligence and social media platforms, and search tools on the public Internet. One characteristic common to all these new sources is a large increase in data requirements. Trading activity, for example, is growing exponentially, and at an increasing rate (Flood, Mendelowitz, and Nichols, 2013, Figure 1; Kirilenko and Lo, 2013, p. 51). Given the persistent growth in computing power and the automation of finance, supervisors have reoriented toward “data-driven” regulation; see, for example, Stein (2015), CFPB (2013) and FRB (2015).3 A prominent example of data-driven supervision is the collection and analysis of detailed contractual terms for bank loan and trading books for stress-testing programs in the United States and Europe since the 2007-09 financial crisis. The scale of the problem is inexorably outstripping the capacity of certain legacy processes that rely significantly on human examiners and analysts. “You can’t solve exponential problems with linear solutions;” one must fight computation with computation.4

Exploiting the new resources will require novel approaches to data management and statistical analysis, and financial stability supervisors have taken notice. These challenges can arise in the most inconvenient circumstances — for example, in the midst of a financial crisis. The potential uses for big data apply to routine situational awareness, as well as occasional spike loads on analytical resources during episodes of market stress. An effective response requires an appreciation of the underlying forces at work.

Experience in numerous sectors shows that the transition point, at which scalability begins to bind, is likely to arise in one of four general directions, often referred to as the four Vs of big data. Macroprudential examples include the following:

- **Volume** – Roughly speaking, the simple size (in bytes) of a dataset, which can place a strain on storage and computational resources. Modern economic datasets often outstrip the query-processing capacity of relational databases such as Structured Query Language (SQL), creating a market for so-called “NoSQL” tools, according to Varian (2014). In some cases, one can attenuate this burden through data aggregation or compression. One example of a financial monitoring task that will experience significant increases in data volumes

---

3 Pattison (2014) offers an interpretation of this trend from the perspective of industry. Recent conferences on big data in financial supervision include those hosted by the Bank of England (Bholat, 2015), Sweden’s Riksbank (Hokkanen, et al., 2015), and European Central Bank and International Institute of Forecasters (IIF, 2014).

4 Attributed to Prof. Banny Banerjee (Chase, 2015, p. 72).
relative to legacy practice is the move toward data-centric audit analytics for forensic analysis of financial accounting records (AICPA, 2015).

- **Velocity** – The rate at which data arrive, which can strain network bandwidth and stream analytics (O’Hara, 2015). An example from macroprudential supervision is real-time monitoring of high-frequency data streams during a flash crash (Berman, 2015). By design, high-frequency trading firms deliver quote and transaction messages at the technical limits of network latency, creating a significant throughput burden for any downstream process.

- **Variety** – The diversity of schemas, or formal structures, for data arriving from different sources, which can strain data integration processes (Halevy, *et al.*, 2006). This applies to the integration of legacy systems after a bank merger, and also to systemwide integration. An example is aligning and synchronizing legal entity identification schemes across a wide variety of independently managed datasets (Rosenthal and Seligman, 2011). Without coordination, the alignment of identifier registries across $n$ back offices requires $(n^2-n)$ mappings between them; for example, 10 back offices implies 90 mappings (Flood, 2009).

- **Veracity** – An elevated error rate in the data can strain data validation, data integrity, and data curation processes (Dong and Srivastava, 2013). An example is maintaining data quality for detailed and granular portfolio data in bank stress tests (Hunter, 2014). Because data integrity is often assessed by reconciling data points against each other, the veracity burden can rise exponentially with data volumes.

The remainder of this paper connects these big-data scalability problems to particular data challenges facing financial stability supervisors. The OFR’s *Financial Stability Report* (OFR, 2015, ch. 4) organizes these issues around three dimensions: scope, quality, and access. Roughly, these address what data to collect, how best to manage and use the data, and who should get to see the data.

**Figure 1. The Lifecycle of Supervisory Data**

In the following sections, we address these same questions of what, how, and who, but organize the discussion around the five significant phases in the usual lifecycle of big data (Jagadish, *et al.*, 2014), considering them in the context of financial supervision. Figure 1 depicts these steps: acquisition, cleaning, integration, analysis, and sharing.
2) Data Acquisition

The front end of data-driven supervision is the collection of information about participants in the financial system. Data collection is an exercise in instrumentation of the system, and a key design consideration is the appropriate measurement resolution. Measurement is necessarily a projection from a deeply nuanced financial system into a discrete measurement space, and important details may be lost in the process. One way to address the lossiness (i.e., the degree of information loss) of the projection is to provide for resolution enhancement within the measurement process, allowing users to drill down into additional detail. For financial stability data, resolution involves four key dimensions: coverage, frequency, granularity, and detail.

**Coverage** traditionally focused on microprudential accounting data from financial firms — for example, the Securities and Exchange Commission’s 10-K reports (SEC, 2014) — and price and volatility data from financial markets, such as those used in Basel capital requirements, set for banks in the United States by the Federal Reserve Board and Office of the Comptroller of the Currency (FRB-OCC, 2013). However, the crisis demonstrated that vulnerabilities can arise in the blind spots not covered by formal data collections. For example, in September 2008, “regulators did not know nearly enough about over-the-counter derivatives activities at Lehman and other investment banks, which were major OTC derivatives dealers” (FCIC, 2011, p. 329). One response has been a focus on identifying and correcting these data gaps. Recently, for example, the OFR has worked with fellow regulators to collect new data on bilateral repurchase (repo) and securities lending agreements (Baklanova, *et al.*, 2015). Internationally, the G-20 Data Gaps Initiative is the most prominent effort to expand the coverage of supervisory information. The G-20 finance ministers and central bank governors launched the Data Gaps Initiative after the crisis in 2009, endorsing 20 specific recommendations for implementation, managed by the Financial Stability Board and International Monetary Fund; see FSB-IMF (2015) and Cerutti, *et al.* (2014).

**Frequency** addresses temporal measurement resolution. Most formal data collections involve repeated discrete snapshots of some aspect of the financial system to accumulate a regularly spaced longitudinal dataset, such as the SEC’s monthly collection of money market mutual fund holdings (SEC, 2016). One typically assumes the data are sampled at regular finite intervals from an underlying continuous-time process evolving smoothly over time. Traditional econometrics provides a rich toolkit for analyzing panel datasets, so that a periodic strobing of the system to collect observations at regularly spaced intervals can be extremely useful for analysis. Unfortunately, the assumption of temporal smoothness does not always hold, especially for systemic risks, where microprudential vulnerabilities may be masked by window dressing and macroprudential events can erupt abruptly, fueled by investor panic and other feedback effects. The upshot can be a sort of “sampling blindness,” where the phenomena of interest occur between sample snapshots. Finer sampling frequencies do not necessarily cure sampling blindness. Microstructure noise increasingly dominates price series — and the related realized volatility and correlation estimates —as the observation frequency increases (Aït-Sahalia and Jacod, 2014).
**Granularity** defines the level and techniques of data aggregation, typically achieved by summing or averaging rows in a database table. This process has both costs and benefits. Because aggregation is a lossy conversion, there is an incentive to capture information at the highest resolution possible and then provide aggregated and/or filtered summaries as needed. It is easier to discard information than to recreate it. On the other hand, summing and averaging reduce raw data volumes, average away random measurement errors and can preserve confidentiality. The SEC’s planned Consolidated Audit Trail (CAT) will collect and identify every order, cancellation, modification, and trade execution for all exchange-listed equities and options across all U.S. markets (SEC, 2012). The CAT represents an unprecedented level of granularity in financial reporting. By its fifth year, the CAT is projected to generate more than 100 billion records per day, occupying more than 20 terabytes of storage daily and eventually exceeding 20 petabytes (Rauchman and Nazurak, 2013, p. 21) However, beyond the operational burden of developing and maintaining software and other routines (Rossi, 2014), aggregation introduces other challenges. Aggregation can disguise important nuances of risk exposures; it is well known, for example, that aggregate performance measures do not capture the full detail of portfolio risk (Foster and Young, 2010).

**Detail** refers to the specific attributes captured for each object (e.g., a firm or transaction) included in coverage — typically represented as columns in a database table. For example, the Financial Industry Regulatory Authority’s (FINRA) Transaction Reporting and Compliance Engine (TRACE) for corporate bond transactions exists in two formats, confidential and public, distinguished in part by whether the dataset includes counterparty identifiers for each trade. The former allows much more detailed analysis of dealer interactions, possibly generating important insights about position concentrations and liquidity formation. Aggregation considerations, such as storage costs and privacy concerns also affect detail, but there are other scalability issues particular to detail.

It is easy for the information content of a dataset to deteriorate as downstream processes filter, normalize, and aggregate the data. Provenance documentation and other key metadata may be lost over time, or lose connection to the original context that made them meaningful (Buneman and Tan, 2007). The information content of a dataset may also grow over time, through integration with other data resources (Zhao, *et al.*, 2004). It is therefore important to prepare for subsequent data curation and integration at the point of data capture.

3) **Data Cleaning**

Scalability issues also affect the next phase, data cleaning, typically achieved through a series of edits and transformations that bring a dataset into compliance with a formal list of data integrity constraints. Many official collections are highly structured and the data emerge from processes
that ensure a baseline level of accuracy, such as double-entry bookkeeping or clearing and settlement. These built-in data integrity checks do not necessarily apply to raw transaction feeds, such as the CAT or TRACE. By capturing the raw transaction message stream, CAT is likely to exhibit noise in the message flow (cancelled quotes, cancelled trades, etc.). The routine realities of millions of human users — investors, traders, brokers, etc. — interacting with a diverse and dispersed financial system suggest we should expect this sort of noise will be more prevalent as data capture moves closer to the source. For example, unstructured data, such as social media feeds, are appearing in systemic risk research, using sentiment analysis to improve forecasts of financial stress.

Data quality is an important practical issue, because inaccurate signals can lead to poor analysis and misinformed decisions (Osborne, 2012). As data volumes grow, so does the magnitude of the data cleaning burden (Dasu and Johnson, 2003). There are tools available for automated data cleaning (Rahm and Do, 2000), quality assessment (Pipino, et al., 2000), and integration (Bernstein and Haas, 2008). These must be adapted for use with financial data, but this does not mean the task is trivial; Burdick, et al. (2015) reveal some of the challenges in implementation. The Basel Committee on Banking Supervision (BCBS) noted that half the systemically important banks surveyed (15 of 30) are straining to implement the BCBS’s 2013 principles on risk data aggregation and rated “themselves as materially non-compliant with Principle 3 (data accuracy and integrity).” Anecdotal evidence “suggests that it will be difficult for a number of firms to fully comply with the Principles by 2016” (BCBS, 2015, p.3). The Enterprise Data Management Council is coordinating efforts within the financial industry to improve data quality along the full information supply chain (EDMC, 2015).

There are good reasons, both technical and behavioral, why data quality in financial reporting may not be a simple matter of extra diligence. Behaviorally, incentives can lead firms to subvert accurate reporting through window dressing (Munyan, 2014) or fraudulent misreporting (Benston, et al., 2004). Operationally, the signal-to-noise ratio can drop as granularity increases. High-frequency trading (HFT) offers an example of the limits of (temporal) granularity. Traditional price-and-time priority of order routing as the basis for best execution is in tension with the SEC’s goal of encouraging competition across trading venues under Regulation NMS (SEC, 2015). However, time priority is impossible to enforce precisely when HFT trading decisions occur faster than the temporal resolution of system clocks (Lombardi, 2006).

The most recent updates to the FINRA’s Rule 7430 requires that trading venues “be synchronized to within one second of the National Institute of Standards and Technology (NIST) atomic clock” (FINRA, 2016, p. 2-1). A one-second tolerance creates enormous latitude when trading response latency is on the order of a millisecond (Hasbrouck and Saar, 2013). In addition, operational errors in HFT systems, such as the one-sided order flow at the root of the May 2010 “Flash Crash” can generate rapid-fire actions with a large cumulative impact very quickly. Many exchanges and their
HFT members have a range of techniques to clamp their systems for rapid trading halts (Clark and Ranjan, 2011), but these are not yet universally applied.

4) Data Integration and Representation

Financial stability monitoring frequently requires the consideration of multiple financial sectors simultaneously. A holistic view is vital, because vulnerabilities that are not apparent in individual firms can emerge at the level of the system as a whole. A comprehensive perspective can create scalability challenges in the “variety” dimension. For example, the analysis of corporate credit risk might require jointly analyzing data on corporate bonds, credit default swaps, bank loans, and corporate equity. Economists are accustomed to merging datasets on a case-by-case basis, but this process does not scale well. Mortgage registration shows how integration can fail at the system level, despite enormous incentives to get it right. Data systems in U.S. mortgage markets did not keep pace with securitization volumes before the 2007-09 crisis, leaving the legal status of many loans unclear and contributing to a large-scale mortgage foreclosure crisis (Hunt, et al., 2014).

In the long run, the emergence of a globally standardized legal entity identifier (LEI) system will help greatly with many financial data alignment tasks (GLEIF, 2014). But the LEI alone is insufficient for high-quality integration. Data alignment is only a first step toward full integration, albeit a significant one. Efforts are underway to augment the simple identification of the first-generation LEI to capture complex ownership relationships (OFR, 2015, p. 70), and to map between the LEI and other common identification schemes (NIST, 2016). More advanced techniques would resolve colloquial mentions of names of financial institutions in news and social media and reconcile them with the formal identifiers. For example, Xu, et al. (2016) resolve named entities extracted from the prospectuses of residential mortgage-backed securities against a vendor list of institution names for asset-backed securities.

Figure 2. Linking Entity Identifiers

<table>
<thead>
<tr>
<th>IDENTIFIER</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPM</td>
<td>New York Stock Exchange ticker symbol</td>
</tr>
<tr>
<td>0000019617</td>
<td>Central Index Key (CIK) assigned by the Securities and Exchange Commission</td>
</tr>
<tr>
<td>1039502</td>
<td>RSSD ID assigned by the Federal Reserve Board</td>
</tr>
<tr>
<td>815DZWKVSZI1NUHU748</td>
<td>Legal entity identifier (LEI) assigned by the Global LEI Foundation</td>
</tr>
<tr>
<td>J.P. Morgan</td>
<td>Company name used in Wall Street Journal stories</td>
</tr>
</tbody>
</table>

Source: Office of Financial Research

In the domain of macroprudential monitoring, the OFR and NIST have funded a public Financial Entity Identification and Information Integration Challenge to develop new technologies for automated identifier alignment and entity resolution in financial datasets and text sources (NIST, 2016). The objective is a reference knowledge base — and some prototype tools — linking heterogeneous collections of entity identifiers from various sources to facilitate information
integration, both within structured data such as regulatory filings, and unstructured data such as news articles and social media. Figure 2 offers a glimpse of the problem, listing a handful of commonly used identifiers for a single firm, JPMorgan Chase and Co. This is a financial holding which itself comprises thousands of additional subsidiary organizations.

Entity identification is the most basic aspect of a more general problem of data representation and metadata management. Metadata most commonly appear as data dictionaries and formal schemas, which structure and describe managed datasets. The same scalability challenges that arise in the case of aligning identifiers apply to the alignment of schemas. One area where schema integration intersects with systemic monitoring is instrument type identification. Each portfolio manager is free to categorize her positions according to a self-defined schema, but financial stability monitors need to identify common exposures across many portfolios — and therefore schemas — at once. The OFR, for example, is progressing on a Financial Instrument Reference Database to fulfill a Dodd-Frank Act mandate (OFR, 2015, p. 72). Because participants and regulators have legacy typologies to meet their local needs, any instrument reference database will face challenges of schema matching and semantic integration.

All of these curation and integration activities should help improve data quality. Reconciliation of a dataset — against other aligned datasets, against the logic of its own internally consistency, and against external integrity rules — is an important technique. Conversely, data quality problems in any one of the source systems may affect the integrated whole. Within a given dataset, the cost of maintaining attribute-level metadata tends to scale linearly with the number of attributes. However, without a framework for metadata management, the costs of scaling can increase at an increasing rate when aligning metadata from multiple sources. Flood (2009) emphasizes that metadata integration is intrinsically unstable for financial risk management, because innovation engenders continual evolution of financial products, risk models, and strategic priorities.

Techniques exist for automated and machine-assisted schema alignment, and the enforcement of consistency requirements, both across schemas and relative to established data integrity rules (Bernstein and Haas, 2008; Rahm and Bernstein, 2001). Formal ontologies to organize definitions and terms can be a useful tool for managing the semantic consistency of metadata across multiple schemas (Noy, 2004; Flood, et al., 2010, pp. 36-39). Traditionally, this semantic effort has been managed by lawyers and domain experts. Formal ontologies can lessen their burden, serve as external benchmark for reconciliation, and expose the details of their conceptual model to newcomers and occasional users who might otherwise be overwhelmed. Data standards can also help with integration (OFR, 2015, ch. 4), but effective standards require the development of good abstractions, and this is often painstaking work. An example is the OFR’s collaboration with the Commodity Futures Trading Commission and international regulators to improve reporting standards and develop shared taxonomies for swap data repositories.
5) Data Modeling and Analysis

Data analytics are often the most prominent aspect of the big data paradigm, and many of the commonly cited approaches, such as data clustering and community detection, are well suited for data-driven pattern identification. Empirical researchers are naturally drawn to the vast and as yet unexplored data sets emerging from news archives, transaction feeds, and social media (e.g., Bholat, *et al.*, 2015; Mamaysky and Glasserman, 2015; Nyman, *et al.*, 2014). These are early days for the exploration of these new data sources, and tantalizing new empirical results have garnered research attention. However, because the underlying data-generating process in financial markets is typically endogenous, purely data-driven (i.e., model-free) empirical regularities should face additional hurdles to justify their validity. In general, if causal inference is a salient goal, as is often the case for policy analyses, then simple predictive analytics and data-driven model selection may be of limited use (Einav and Levin, 2014).

Data analysis, like the other phases of big data processing, faces scale issues that create both problems and opportunities. One challenge for traditional econometrics facing big data is model selection. Because there are few constraints on which details should be considered, feature selection on an unstructured dataset can generate an arbitrary number of potential regressors. Even structured data can yield a combinatorial explosion of specifications. Sala-i-Martin (1997), working with 62 possible explanatory variables in a traditional growth equation, famously ran two million distinct specifications. This proliferation of specifications creates the potential for data mining. What is often touted as a powerful feature of big data analytics is unacceptable to the traditional econometrician. Because many big data sources, such as news archives, are novel to financial econometrics, there are as yet few theoretical constraints to limit the set of acceptable specifications. For policy questions, the incentives are potentially very strong for an analyst to get the “right” answer, so false discovery rates are a genuine concern (Fan, *et al.*, 2014; Domingos, 2012).

In some cases, a relatively straightforward Bonferroni correction is adequate to adjust for the naturally occurring rate of false positives in a large sample (Curme, *et al.*, 2014; Alanyali, *et al.*, 2013). In other cases, the fix is not so simple. Donoho and Stodden (2006) consider the so-called “fat regression” or “$P >> N$” problem, where the number of predictors significantly exceeds the number of observations. In these situations, the $N$ data points are sparsely distributed in a much higher dimensional measurement space, and key elements of the asymptotic theory underlying traditional econometrics simply breaks down. Varian (2014) outlines some alternative approaches. Dhar (2013) emphasizes the importance of out-of-sample predictive power as a model-selection criterion. The key point is that big data necessitates new approaches, not just faster hardware.
6) Data Sharing and Transparency

The final set of data-management tasks involves the disposition of the data once they have been collected, cleaned, documented, and analyzed. The data are a resource to support decision-making, rule-making and policymaking, and to provide context for other analyses. In many cases, a supervisor openly publishes or selectively shares collected data or analytical derivations to support public accountability, transparency to investors, or industry decision making. In this role, supervisors become input-process-output engines. They take raw data (regulatory collections, market data feeds, etc.) as inputs, transform them through various analytical processes into derived artifacts (regression results, industry aggregates, financial stability reports, visualization tools, etc.), and distribute them to targeted user groups (public or industry risk dashboards, static and interactive visualization tools for research and decision support, archival records of system state for historical and/or accountability analysis, etc.).

Here again, scalability affects the process. For example, the disclosure of aggregate statistics derived from sensitive underlying details is restricted. Financial supervisors and statistical agencies in the U.S. face a complex array of privacy and confidentiality laws and regulations governing such statistics (Flood, et al., 2013, Section 3 and Appendixes A and B). In Europe, the primary governing legislation is the Data Protection Act (Howell, 2014). Traditional methods of disclosure control include suppression of key fields, data “blurring” (i.e., noise addition), and data bucketing (i.e., replacing detailed attributes with coarser categories), but linkage attacks involving other data sources can often defeat the legacy techniques. For similar reasons, the U.S. Census Bureau no longer requests Social Security numbers for its Survey of Income and Program Participation, because respondents have gradually learned not to provide this datum (McNabb, et al., 2009).

Visualization, in the ancient sense of how a picture can be worth a thousand words, is often critical to human understanding. Humans have evolved extensive capacity for visual perception and cognition, in many cases giving them a comparative advantage in pattern recognition over alternative tools. One can also supercharge these innate skills with visual analytics, which augments static images by putting a human in the loop to control the rendering interactively. Per Shneiderman’s (1996) mantra, “overview first, zoom and filter, then details-on-demand,” tools that offer selective resolution enhancement over huge datasets can provide the best of both worlds — depth and breadth — simultaneously. Flood, et al. (2016) highlight four broad macroprudential tasks where visualization can play a vital role: sensemaking, rulemaking, policymaking, and transparency. Interactive visualization is especially valuable for sensemaking, where undirected exploration is a key part of the task.

7) Conclusions

Financial stability monitors clearly face big data challenges, due to the extraordinary scale of the
system under supervision. The central issue is a question of scalability. Big data becomes a problem when the scale of requisite datasets overwhelms the tools available for processing, in several broad dimensions: volume, velocity, variety, and veracity. In other words, bigness is not an attribute of a dataset per se, but rather describes the dataset’s volume, velocity, etc. relative to the capacity of available processes. Data describing the full financial system can be big in any of the four dimensions, overwhelming legacy analytical processes that are often fundamentally microprudential in nature.

The LEI, implemented now on a global scale, is a good example of the financial community coming together to develop a “non-linear solution to a non-linear problem.” Supervisors should embrace such clearly defined, rigorously maintained (for example, via formal ontologies) shared semantics more widely. The LEI identification scheme is just the simplest and most fundamental example of shared financial meaning. When combined with state-of-the-art solutions for named entity extraction and linkage of entity mentions in financial contracts (Xu, et al., 2016), open standards such as LEI can provide a big-data foundation for macroprudential analysis (NIST, 2016). Macroprudential supervisors should stay alert to recognize data scalability challenges as they arise.

The financial sector and the data requirements for monitoring it will co-evolve. The industry structure responds to monitoring and vice versa. For example, Basel capital charges for concentrated mortgage exposures on bank balance sheets provided a regulatory arbitrage impetus for the migration of mortgage finance into securitization markets (Ambrose, et al., 2005), where it escaped much of the intensive monitoring program that bank examiners had developed over the years. Completing the loop, data requirements for monitoring mortgage finance have grown dramatically since the crisis, at least in the United States.

The industry is generating more data and regulators are collecting more, a trend that has only accelerated since the crisis. This growth is exposing new scalability challenges, such as quality limitations that hinder the interpretation of the new, highly granular data collections. Efforts to improve data quality along the full information supply chain (e.g., EDMC, 2015) are a necessary part of the solution. At the same time, the integration of this detailed — often loan-level — mortgage information with other data sources poses new and important privacy challenges, such as linkage attacks, that can overwhelm traditional data masking techniques. New methods to assess and assure data privacy in this context are appearing, but have yet to be widely adopted as part of the supervisory toolkit. Mortgages are but one example of a more general lesson, that macroprudential supervisors should stay aware of the many techniques that are emerging to address the challenges of the big data landscape.
References

Aït-Sahalia, Y. and Jacod, J. (2014) 

AICPA (2015) 
*Audit Analytics and Continuous Audit: Looking Toward the Future*, AICPA.


Basel Committee on Banking Supervision (2015) 
“Progress in adopting the principles for effective risk data aggregation and risk reporting,” January, [http://www.bis.org/bcbs/publ/d308.htm](http://www.bis.org/bcbs/publ/d308.htm).

Basel Committee on Banking Supervision (2013) 
“Principles for effective risk data aggregation and risk reporting,” January, [http://www.bis.org/publ/bcbs239.htm](http://www.bis.org/publ/bcbs239.htm).


“Information Integration in the Enterprise,” *Communications of the ACM*, 51(9), September, 72-79.

“Text mining for central banks,” Centre for Central Banking Studies Handbook (33),  

Board of Governors of the Federal Reserve (2015)  
“Enhancements to the Federal Reserve System’s Surveillance Program,” Memorandum, December,  

“Regulatory Capital Rules: Regulatory Capital, Implementation of Basel III, Capital Adequacy,  
Transition Provisions, Prompt Corrective Action, Standardized Approach for Riskweighted  
Assets, Market Discipline and Disclosure Requirements, Advanced Approaches Risk-Based  

Buneman, P. and Tan, W.-C. (2007)  
“Provenance in databases,” In: Proceedings of the 2007 ACM SIGMOD International  
Conference on Management of Data, 1171-1173.

Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L. C., Stanoi,  
“Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study,”  

Casey, M. (2014)  
“Emerging Opportunities and Challenges with Central Bank Data,” presentation slides, October,  
https://www.ecb.europa.eu/events/pdf/conferences/141015/presentations/Emerging_opportunitie  
s_and_challenges_with_Central_Bank_data-presentation.pdf?6074ecbc2e58152dd41a9543b1442849.

“Systemic Risks in Global Banking: What Available Data Can Tell Us and What More Data Are  
Needed?,” Chapter 16 In: Risk Topography: Systemic Risk and Macro Modeling, Brunnermeier,  

Peers Inc: How People and Platforms Are Inventing the Collaborative Economy and Reinventing  
Capitalism, PublicAffairs.

Clark, C. and Ranjan, R. (2011)  
“How Do Exchanges Control the Risks of High Speed Trading?”, Policy Discussion Paper  
Consumer Financial Protection Bureau (2013)

“Quantifying the semantics of search behavior before stock market moves,” Proceedings of the National Academy of Sciences, 111(32), 11600–11605.

Exploratory Data Mining and Data Cleaning, Wiley-Interscience.

DeCovny, S. (2014)

Dhar, V. (2013)

Diebold, F. X. (2012)

Domíngos, P. (2012)

“Big data integration,” In: 29th International Conference on Data Engineering (ICDE), 1245-1248.


Einav, L. and Levin, J. (2014)
“Economics in the age of big data,” Science, 346(6210), November.

Enterprise Data Management Council (2015)

Fan, J., Han, F. and Liu, H. (2014)
Financial Crisis Inquiry Commission (2011)

Financial Industry Regulatory Authority (2016)

Financial Stability Board and International Monetary Fund (2015)

Flood, M. D. (2009)


Foster, D. P. and Young, H. P. (2010)

Global Legal Entity Identifier Foundation (2014)
“Data integration: the teenage years,” In: Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06), 9-16.

Hasbrouck, J. and Saar, G. (2013)

The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research.

“The Riksbank’s future information supply in light of Big Data,” Economic Commentaries (17), Sveriges Riksbank.

Horvitz, E. (2010)


Hunter, M. (2014)
“Statement by Maryann F. Hunter, Deputy Director, Division of Banking Supervision and Regulation, Board of Governors of the Federal Reserve System before the Committee on Banking, Housing, and Urban Affairs, U.S. Senate, Washington, D.C.,”

IBM (2016)

International Institute of Forecasters (2014)


“Legal and Technical Measurement Requirements for Time and Frequency,” *Measure*, 1(3), September.


Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. (2011)


MongoDB (2016)

Munyan, B. (2014)

National Institute of Standards and Technology (2016)


Pattison, J. C. (2014)


Rahm, E. and Bernstein, P. A. (2001)


Rossi, C. (2014)

Sala-i-Martin, X. X. (1997)
Securities and Exchange Commission (2012)

Securities and Exchange Commission (2014)


Securities and Exchange Commission (2016)

Shneiderman, B. (1996)

Somerset House (2015)


Varian, H. R. (2014)

Xu, Z., Burdick, D. and Raschid, L. (2016)