Panel 6: Data Integration and Visualization
Friday, October 28, 2015 at 1:45 p.m.

Moderator:
Mark Flood, Office of Financial Research

Panelists:
Aurel Schubert, European Central Bank
Amol Deshpande, University of Maryland
Peter Sarlin, Hanken School of Economics
Margaret Varga, Seetru and University of Oxford

Big data scalability challenges are emerging in all disciplines, and financial supervision is no exception (Flood, et al., 2016). Irrespective of the application domain, scalability issues tend to exhibit common facets, often summarized as the Four Vs: *volume*, *velocity*, *variety*, and *veracity*. On the other hand, financial data also have special concerns (Brose, et al., 2014; Flood, 2009).

This panel will consider two distinct but interrelated phases in the process of marshalling financial information and delivering it to the user. Visualization requires a robust set of conceptual abstractions, so that data can be rendered and compared in a useful way on a 2-dimensional page or computer screen (Sarlin, 2013; Sarlin and Peltonen, 2013; Flood, et al., 2016). Data integration is the task of massaging the raw data inputs into such a common conceptual framework (Bernstein and Haas, 2008).

Financial Data Integration in Practice at the European Central Bank

Aurel Schubert, ECB

- In the last few years, central banks mandates have been extended to fight the crisis, incorporating in a number of cases new responsibilities for macro-prudential supervision and micro-prudential supervision to their traditional task of conducting monetary policy.
- This extension of central banks' tasks has implied invariably additional demands for more timely and disaggregated data.
- To cope with the high data demand, the ECB has embraced the era of "big data" and it is developing several ambitious projects for collecting, managing and disseminating disaggregated data to serve multiple uses.
- Since years, the ECB has been setting up infrastructures able to efficiently compile, manage, distribute and exchange (also among European institutions) granular datasets.
- The ECB work in the field of securities in the euro area has comprised both the development of securities databases on issues and holdings, and the taxonomy of securities statistics (through the *Handbook on Securities Statistics*: IMF, 2015).
- Today, the ECB's Centralised Securities Database (CSDB) (ECB, 2010a; Mayerlen, 2014) contains reference data of approximately 10 million of individual securities. The information is obtained from several commercial data providers, national central banks and other ECB sources which are combined to create a "golden copy" for each security.
- The CSDB work is done in a network. The ECB maintains the technical application and procures centrally commercial data. The National Central Banks (NCBs) provide source data and monitor the quality of their resident issuers. The ECB monitors non euro area issuers.
- The ECB has also set up a granular database on securities holdings providing quarterly data on holdings of individual securities for all main euro area sectors and for the largest banking groups in the euro area (ECB, 2015). The granularity of the data provides a range of breakdowns on both the issuer and holder sides, which are not available in other statistics. Following the approval of the amended ECB legal act in August 2016, as of 2018 the reporting population will be extended to all institutions directly supervised by the ECB.
- Other important micro-databases are (i) the new dataset on Money Market Statistical Reporting (MMSR) in the euro area, which since April 2016 has helped to close important existing data gaps. (ii) Furthermore, a harmonized granular credit database (AnaCredit: ECB, 2016) is currently under development and will imply the collection of millions of individual bank loans in the euro area, harmonised across all its member states. Initially it will cover credit granted to legal entities (non-financial corporations, financial corporations and governments).

Linking databases and sharing data

- While recognising the richness of the individual micro-datasets, the best way to serve multiple users is to link several databases, registers and metadata repositories. A precondition for it is the development and use of standards and unique identifiers of counterparties, transactions and products.
- The application of harmonised statistical concepts and unique identifiers (e.g. ISIN and LEI) has proven critical in linking the CSDB and the Securities Holding Statistics databases. As a consequence, one of the major challenges for the current pilot exercise at the ECB is how to link various datasets without a common unique entity and instrument identifier with full global coverage.
- More worldwide, the harmonisation and establishment of mandatory requirements regarding the use of the International Securities Identification Number and of the global Legal Entity Identifier in European and global regulation is needed.
- Over time, work towards harmonising Unique Product Identifier (UPI) and Unique Transaction Identifier (UTI) would provide further benefits.
- Harmonisation and standardisation is crucial also for supporting a culture of data sharing among international authorities and institutions. The second phase of the G-20 Data Gaps Initiative includes as recommendation II.20 the Promotion of Data Sharing by G-20 economies, including the increase in "the sharing and accessibility of granular data, if needed by revisiting existing confidentiality constraints."

Data integration along the whole information chain

- The paradigm shift that the ECB is currently undergoing, represented by the shift from aggregate to granular statistics, concerns the whole information chain, starting from innovative, easier and more consistent data reporting.
- In this regard, the ECB has already taken important steps to improve the harmonisation and standardisation of data with the current development of the ECB's Single Data Dictionary (SDD) and the Banks' Integrated Reporting Dictionary (BIRD, 2016).
- The current work on BIRD implies collaboration with the banking industry. Under the leadership of the ECB, seven national central banks and 26 commercial banks are developing the necessary documentation describing the banks' source data (so called input layer) and the transformation rules that banks might use to fulfil the reporting requirements of the authorities (the output layer).
- The purpose of BIRD is to make the documentation freely available to the general public to ensure that any stakeholder may make use of it as it considers best. The implementation of the BIRD documentation is entirely voluntarily, and the responsibility of meeting the reporting requirements remains with the reporting agents regardless if they follow the BIRD documentation or not.
- Among other advantages, the BIRD will increase standardisation and lower the risk of mistakes or misinterpretation of reporting requirements by reporting agents.

- The ECB’s Single Data Dictionary (SDD) consists of the methodological and semantic integration of existing European reporting frameworks in order to provide clear data definitions, with reconciled meaning across regulatory frameworks.
- The dissemination of data to users is the ultimate goal of the data production. In order to improve it, the ECB is developing the Data Intelligence Service Centre (DISC), which is an analytical IT platform which will enable users to access several data in a single platform and facilitate advanced data analytics. Besides, it will offer flexible and secure data storage and it will accelerate the delivery of data-driven projects.

Managing collaborative data analytics

Amol Deshpande, U. Maryland

Data-driven methods and products are becoming increasingly common in a variety of communities, including finance, science, education, and social and web analytics. Increasingly, "data science" teams want to collect, clean, structure, store, and collaboratively analyze large datasets, to understand trends, and to extract actionable business or social insights. Unfortunately, while there exist tools to support data analysis, much-needed underlying infrastructure and data management capabilities are missing. The process of collaborative data science is often ad hoc, typically featuring highly unstructured datasets, an amalgamation of different tools and techniques, significant back-and-forth among the members of a team, and trial-and-error to identify the right analysis tools, algorithms, models, and parameters. Distressingly, most systems either focus on performance, or on supporting even more sophisticated analyses, instead of simplifying and automating many of the fundamental book-keeping operations which are a necessary prerequisite for data science, including data cleaning and ingestion, data collaboration and versioning, provenance and metadata management, and introspective analysis of end-to-end pipelines.

Core challenges

We briefly highlight some of the key open challenges in enabling collaborative data science below; further discussion can be found in [Bhardwaj, et al., 2015; Miao, et al., 2016; and Hellerstein, et al., 2016).

Versioning

In a collaborative data science scenario, there are often hundreds or thousands of versions of collected, curated, and derived datasets, at various degrees of structure (fully structured and relational all the way to completely unstructured), each of which can have millions to billions of records within them. We need the ability to keep track of these versioned datasets; enable the access, retrieval, and modification of specific dataset versions; and share and distribute the datasets with other members of the data analysis team. While source code version control systems like "git" and "svn" have proved tremendously useful for collaborative source code management, they cannot handle large (semi-)structured datasets efficiently and offer limited querying functionality (Chavan, et al., 2015).

Provenance

A second major challenge is that there is no easy way to capture and reason about ad hoc data science pipelines, many of which are often spread across a collection of analysis scripts. Metadata or provenance information about how datasets were generated, including the programs or scripts used for generating them and/or values of any crucial parameters, is often lost. Similarly, it is hard to keep track of any dependencies between the datasets. As a result, data scientists are required to manually keep track of, and act upon, such information, which is not

only tedious, but error-prone. For example, data scientists must manually keep track of which derived datasets need to be updated when a source dataset changes. They often use spreadsheets to list which parameter combinations have been tried out during the development of a machine learning model. Debugging becomes significantly harder; e.g., a small change in an analysis script may have significant impact on the final result, but identifying that change may be non-trivial, especially in a collaborative setting.

Reverse reasoning

A related challenge is that it is difficult to do forward or reverse reasoning over the results of an analysis pipeline; such ability is essential to, e.g., identify the impact of a potentially bad input data item, or explain a specific output result by tracing it back to specific input data items. This is especially a problem with black-box machine learning techniques like deep learning, where the learned models are often very effective but are hard to understand or reason about. Further, although there exist point solutions for doing provenance capture and querying, and for explaining results of some specific machine learning algorithms, no existing tools address doing so for end-to-end and deep analysis pipelines. This requires the ability to combine, manage, and query a very large volume of provenance data, and to figure out how to best combine human expertise with that provenance information to best answer specific questions.

Solution approaches

Prof. Deshpande's group at the University of Maryland is building a hosted data platform, called DataHub (Bhardwaj, et al., 2015), to simplify and accelerate collaborative data science. DataHub is being developed jointly with researchers at MIT and UIUC (see the DataHub (2016) main project website). Key features of DataHub include:

(1) a flexible, source code control-like versioning system for data, that efficiently branches, merges, and differences versioned datasets;
(2) new data ingest, cleaning, and wrangling tools designed to automate the cleaning process as much as possible; and
(3) a provenance and metadata management system that supports capturing and analyzing provenance information in a variety of ways (Chavan, et al., 2015; Miao, et al., 2016).

Annotatable Dashboards

Peter Sarlin, Hanken School of Economics

- Supervisory and regulatory tasks require analysis of complex data. Certain implications descend from a more prominent role for macroprudential analysis. Beyond the soar in availability and precision of data, the transition from firm-centric to system-wide supervision imposes obvious data needs when analyzing a large number of entities and their constituents as a whole (see e.g. Flood and Mendelowitz, 2013).

- As central tasks ought to be timely and accurate measurement of systemic risks, big data and analytical models and tools become a necessity. Underlying systemic risk, while having no unanimous definition, has commonly been distinguished into three categories (de Bandt et al., 2009; ECB, 2009): (i) build-up of widespread imbalances, (ii) exogenous aggregate shocks, and (iii) spillover and contagion.

- Fortunately, policymakers and regulators have access to a broad toolbox of analytical models to measure and analyze system-wide threats to financial stability. The tasks of these tools can be mapped to the above listed three forms of systemic risk (e.g., ECB (2010)): (i) early warning of the build-up of widespread vulnerabilities and imbalances, (ii) stress-testing the resilience of the financial system to a wide variety of exogenous aggregate shocks, and (iii) modeling contagion and spillover to assess how resilient the financial system is to cross-sectional transmission of financial instability.

- In a recent study (Sarlin, 2016) on the use of visualization in macroprudential oversight (see also Flood et al., 2016), in which we started from defining the overall task, coupling that with available data sources and the structure of underlying data as well as reviewing previous literature. Eventually, the discussion boils down to two essential, but to date rare, features for supporting the analysis of big financial data and the communication of risks: analytical visualizations and interactive interfaces. In line with this we put forward VisRisk as RiskLab's online interactive platform for studying data related to financial risk.

- The notion of an analytical technique for visualization differs by rather using analytics for reducing the complexity of data, with the ultimate aim of visualizing underlying data structures. These techniques provide means for drilling down into the data cube. For instance, mapping techniques provide a projection of high-dimensional data into two dimensions through dimension reduction, whereas clustering methods enable reducing the volume of data into fewer but representative mean groups. The coupling of visual interfaces with interaction techniques goes to the core of visual analytics. This has largely been overlooked in the policymaking community. A key task of macroprudential supervisors has been to publish risk dashboards, such as that of the ESRB, yet none of these have been

truly interactive. For instance, the ESRB risk dashboard was still in 2014 a static close to 50-page chart pack. This needs to, and obviously will, change.

- Yet, the overview we did pointed out a third, much more important unsolved task: How do we communicate around data? This takes us to annotatable dashboards, or annotatable data. We have built a prototype to tackle major challenges in communicating around data by relying on combining interactive visualization with the simple notion of annotations. An annotation is simply metadata (e.g. a textual comment, file, picture, drawing or numerical value) attached to specific data points in a visualization. This turns soft, or even tacit, knowledge to documented and structured data in a searchable database. In principle, this allows for standard means from knowledge management directly into the sensemaking process.

- Annotatable dashboards have significant implications on internal communication and knowledge sharing in organizations. At the lowest level, it not only allows individual experts to interact with the visualizations but also to interact with each other. This human interaction around data supports horizontal communication as well as the overall sensemaking process. Likewise, you could think a searchable database of expert knowledge measured from the sensemaking process also being a direct support for management, which implies also covering the vertical dimension of communication. In this type of a social platform, the spread of knowledge happens through push and pull functions, as it is not only possible to search previous annotations but also possible to tag others and ask for a comment or further details. For readers familiar with the visual information seeking and the visual analytics mantra, we can write this as yet another one, the data annotation mantra:

"Interact with visuals to find patterns, annotate data and discuss to elaborate, search on demand"

References

de Bandt, O., Hartmann, P., Peydro, J., 2009. Systemig risk in banking: An update. In: Berger, A., M. P.,
Wilson, J. (Eds.), Oxfoord Handbook of Banking. Oxford University Press, Oxford.

ECB, 2009. The concept of systemic risk. In: Financial Stability Review (December 2009). European Central
Bank, Frankfurt, Germany.

ECB, 2010. Analytical models and tools for the identification and assessment of systemic risks. In: Financial Stability Review (June 2010). European Central Bank, Frankfurt, Germany.

M. Flood, V. L. Lemieux, M. Varga, and B. L. Wong, 2016. The Application of Visual Analytics to Financial Stability Monitoring. Journal of Financial Stability, forthcoming, 2016.

Flood, M., Mendelowitz, A., 2013. Monitoring financial stability in a complex world. In: Lemieux, V. (Ed.), Financial Analysis and Risk Management Data Governance, Analytics and Life Cycle Management. Springer-Verlag, Heidelberg, pp. 15-45.

Sarlin P, 2016. Macroprudential oversight, risk communication and visualization. Journal of Financial Stability, forthcoming, DOI: 10.1016/j.jfs.2015.12.005.

Visual Analytics

Margaret Varga, Seetru and Oxford U.

The challenges in exploitation of big data arise not only because of the huge volumes, but also potentially due to the data's high velocity, its variety of form, variability in nature and questions of uncertainty and veracity.

Financial situation awareness is vital in support of making informed decisions for maintaining stability and mitigating risks. It is a continuous operation and includes not only awareness of the nature of the financial services but also their availability, confidentiality, operations and infrastructure. Effective financial situation awareness and management, and thence stability, must not only be re-active but also pro-active, and be able to make predictions as to the potential states and vulnerabilities of the situation. To support managers in the maintenance of situation and risk awareness the analysts need:

- to be able to manage and analyze the massive amount of data from different data sources regarding *financial activities;*
- to *understand the monitored system(s) and risks*? What are the components? Which interrelations and dependencies exist and to what extent?

The main challenge is how to present the massive volumes of dynamic and complex big data in a tractable, comprehensible and usable manner in order to enable analysts to make sense of the big data.

A picture is worth a thousand words; the use of visual aids is a well-established practice, they are used by humans as essential tools to help them define, understand, analyze, explore, explain and navigate their way through their tasks / problems. In the case of financial data, a picture may be worth millions of Dollars / transactions / securities / loans.

Most existing user-interfaces show statistics and aggregations using visualizations such as line charts, bar charts, graphs, geo-spatial representations etc. Such standard *analysis and visualisation* approaches are, however, inadequate to make sense of the ever increasing volumes of complex and heterogeneous data. There is therefore a need for new methods, approaches and tools to manage, exploit and utilise these valuable information to meet the different needs, tasks and decision requirements in the financial sector. Visual Analytics offers a possible solution.

Visual Analytics is the science of analytical reasoning facilitated by interactive visual interface (Thomas and Cook, 2005). It is an iterative process: it combines automated analysis techniques with interactive visualisation, and it provides users with an effective means to dynamically and visually interact with, explore and analyze big, complex, and at times conflicting data. It enables the analysts fully to utilise their cognitive and perceptual capabilities with the support of advanced computational capabilities to enhance the discovery process and the derivation of

insight. It thus facilitates the understanding of the data and situation so as to support timely informed decision making (Chen and Zhang, 2014; Flood, et al., 2016; Keim, et al., 2010; Thomas, 2009; and Ware, 2013).

Visual Analytics offers a powerful data driven methodology applied in a user centered way: it supports the analyst in analyzing data from different sources using integrated interactive and dynamic visualizations. It can provide an effective means for analysts to see, interact with, explore and compare data of different dimensions and types at multiple granularities and in real-time, and thus support the analyst in making sense of the data (sense-making). It allows for detection of trends, patterns, changes, anomalies, outliers and weak points. This data driven user centered methodology (User Centered Visual Analytics) focusses on the users' and the tasks' needs, decision needs, the users' skills as well as their mental models, i.e. fluidity (Wang, et al, 2013). Ultimately, the only thing governing data driven approaches, such as these, is the available data.

New and improved interactive approaches to visually exploring big data have been become available because of the progress in database technologies which enables the filtering and calculation of aggregations on billions of documents in near real-time. Data integration is, however, a challenge as we need to decide which data sets / sources are relevant and enhance our situation and risk awareness for multiple purposes, c.f. structured data that has been collected for a specific purpose and is being used for the intended task(s) only. It is now therefore necessary to identify and extract the relevant data that is required for the tasks / decisions in question. Data rich and information poor situations must be avoided, more is not necessarily better as Herbert Simon (1971) pointed out that:

> *It consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention.*

Furthermore, when using external data sources or open source data such as that from social media care must be taken to be aware of its origins, provenance, trustworthiness as well as its value and relevance.

References

[1] de Bandt, O., Hartmann, P., Peydro, J., "Systemic risk in banking: An update," In: Berger, A., M. P., Wilson, J. (Eds.), *Oxford Handbook of Banking*. Oxford University Press, 2009.

[2] Banks' Integrated Reporting Dictionary (BIRD) Group, "What is the BIRD?" Internet site, 2016, http://www.banks-integrated-reporting-dictionary.eu/documents:whatisbird.

[3] Bernstein, P. A., and Haas, L. M., "Information integration in the enterprise," *Communications of the ACM*, 51(9), September 2008, 72-79.

[4] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A.J., Madden, S., and Parameswaran, A., "DataHub: Collaborative data science and dataset version management at scale," *Proceedings of the CIDR Conference*, 2015.

[5] Brose, M., Flood, M., Krishna, D., and Nichols, W. (eds.), *Handbook of Financial Data and Risk Information*, Cambridge University Press, 2014, http://www.cambridge.org/us/academic/subjects/economics/finance/handbook-financial-data-and-risk-information.

[6] Chavan, A., Huang, S., Deshpande, A., Elmore, A.J., Madden, S., and Parameswaran, A., "Towards a unified query language for provenance and versioning," *USENIX TAPP Workshop*, 2015.

[7] Chen, C. L. P. and Zhang, C. Y. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Science* 275, 2014, 414-347.

[8] DataHub, "What is DataHub?" Internet site, 2016, https://datahub.csail.mit.edu/www/.

[9] European Central Bank (ECB), "The 'Centralised Securities Database' in Brief," technical report, 2010a, https://www.ecb.europa.eu/pub/pdf/other/centralisedsecuritiesdatabase201002en.pdf.

[10] _____, "The concept of systemic risk," In: European Central Bank *Financial Stability Review*, December 2009.

[11] _____, "Analytical models and tools for the identification and assessment of systemic risks," In: European Central Bank *Financial Stability Review*, June 2010b.

[12] _____, "Who Holds What? New Information on Securities Holdings," ECB Economic Bulletin, 2, March, 2015, 72-84, https://www.ecb.europa.eu/pub/pdf/other/eb201502_article02.en.pdf.

[13] _____, "AnaCredit," Internet site, 2016, https://www.ecb.europa.eu/stats/money/aggregates/anacredit/html/index.en.html.

[14] Flood, M., "Embracing Change: Financial Informatics and Risk Analytics," *Quantitative Finance*, 9(3), April 2009, 243-256, http://www.informaworld.com/10.1080/14697680802366037.

[15] Flood, M., Jagadish, H. V., and Raschid, L., "Big Data Challenges and Opportunities in Financial Stability Monitoring," *Financial Stability Review*, 20, Banque de France, April 2016, 129-142, https://www.banque-france.fr/en/publications/financial-stability-review.html.

[16] Flood, M. D., Lemieux, V., Varga, M.J. and Wong, W., "The Application of Visual Analytics to Financial Stability Monitoring," *Journal of Financial Stability*, forthcoming, 2016, http://dx.doi.org/10.1016/j.jfs.2016.01.006.

[17] Flood, M., Mendelowitz, A., "Monitoring financial stability in a complex world," In: Lemieux, V. (Ed.), *Financial Analysis and Risk Management Data Governance, Analytics and Life Cycle Management*, Springer-Verlag, 2013, 15-45.

[18] Hellerstein, J., et al., "Establishing Common Ground with Data Context," technical report, 2016, http://github.com/ground-context/ground/blob/master/CIDR17.pdf.

[19] International Monetary Fund, *Handbook on Securities Statistics*, 2015, http://www.imf.org/external/np/sta/wgsd/pdf/hss.pdf.

[20] Keim, D. Kohlhammer, J., Ellis, G. and Mansmann F., (eds) *Mastering the information age solving problems with visual analytics*, edited by, Eurographics Association, 2010. ISBN: 978-3-905673-77-7.

[21] Mayerlen, F., "Which value added can a securities database platform provide to a central bank?" *Irving Fisher Committee (IFC) Bulletin* No. 37, Jan. 2014, http://www.bis.org/ifc/publ/ifcb37e.pdf.

[22] Miao, H., Chavan A., and Deshpande, A., "ProvDB: A System for Lifecycle Management of Collaborative Analysis Workflows," technical report, 2016, https://arxiv.org/abs/1610.04963.

[23] Sarlin, P., "Mapping Financial Stability," PhD thesis, Abo Akademi University, 2013. http://tucs.fi/publications/view/?pub_id=phdSarlin_Peter13a.

[24] _____, "Macroprudential oversight, risk communication and visualization," *Journal of Financial Stability*, forthcoming, 2016, DOI: 10.1016/j.jfs.2015.12.005.

[25] Sarlin, P. and Peltonen, T. A., "Mapping the State of Financial Stability," *Journal of International Financial Markets, Institutions and Money*, 26, 2013, 46–76.

[26] Simon, H. A., *Designing Organizations for an Information-Rich World, Computers, Communication, and the Public Interest*, Baltimore, MD: The Johns Hopkins Press, ISBN 0-8018-1135-X, pp. 40–41, Martin Greenberger (eds.) 1971.

[27] Thomas, J. J. and Cook, K. A., (eds.), *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Press, 2005.

[28] Thomas, J. J., "Visual analytics techniques that enable knowledge discovery: detect the expected and discover the unexpected," *ACM SIGKDD Workshop on visual analytics and knowledge discovery (VAKD 09)*, Paris, France, 28 June 2009.

[29] Ware, C. *Information Visualization* (Interactive Technologies), Elsevier, Inc. 2013.

[30] Wong, W. L., "Fluidity and Rigour: Designing Visual Analytics for the Demands of Intelligence Analysis," technical report, *NATO IST-116 Visual Analytics Symposium*, Shrivenham, United Kingdom, 28th – 29th October 2013.