Forecasting with Massive Data in Real Time
New York, April 20-21, 2017

POST-WORKSHOP REPORT

Background

The workshop took place at the Microsoft Technology Center in Times Square, New York, with the attendance of approximately 25 participants from academia, industry, and government. The scientific committee was formed by George Monokroussos (Amazon), Michael Kane (Yale), and Claudio Antonini (BNY Mellon). The kind sponsorship of Microsoft and MarketingFox Inc. is acknowledged.

The current forecasting landscape motivated the theme for the workshop. On the one hand and, increasingly, many algorithms proposed in forecasting are specifically developed to handle massive amounts of data, running on unique hardware infrastructures with special software requirements. On the other hand, relatedly, the latest massive applications come from large companies where a number of sophisticated developments can be found. In addition to be used by the company themselves, these developments are marketed to be used by the public at large.

One lesson from the workshop is that one environment that can potentially alter the future of forecasting has been little mentioned in the specialized literature: quantum computing. However, its future is showing the same traits that can be observed in the rest of the forecasting world: besides the traditional academic environment, various companies are investing heavily and developing unique but complementary solutions (simulators, compilers, machines with limited number of qubits).

A brief detail of the talks is included below. Links to the program and papers are available at www.forecastny.com.

Beyond Big Data: Technology to Understand Complicated Systems
Paul Cohen (DARPA, first keynote speaker)

Dr. Cohen described two programs currently running at the Defense Advanced Research Projects Agency (DARPA) that are addressing the need for causal models and quantitative analysis in large-scale environments: “World Modelers” and “Big Mechanisms.” The former aims to develop technology to integrate qualitative causal analyses with quantitative models and relevant data to provide comprehensive understanding of complicated, dynamic national security questions and, as a case study, it will be applied to analyze food insecurity resulting from interactions among climate, water, soil, markets, and physical security. The second program will help model and analyze very complicated systems by allowing machines to read and assemble fragmented literatures and constructing reasoning models. The domain of this program is cancer biology with an emphasis on signaling pathways, although the reach can be extended to other fields. To the extent that the construction of Big Mechanisms can be automated, it could change how science is done by incorporating large amounts of material currently not considered.

A Cointegration Approach to Identifying Systemic Risk in Markets
Michael Kane (Yale)

The talk analyzed how to assess systemic risk using the cointegration of financial markets, relying on the 2010 FlashCrash as a case study. During the presentation it was also explored how an alternative to current, single-stock circuit breaker/collar rules could be employed by FINRA to control market volatility.

Asymptotically Optimal Identification of Structural Breaking Point in Real Time with Application to Dating Recessions
Haixi Li (Fannie Mae), Xuguang Sheng (American University)

A procedure was described to detect an abrupt structural change in real time in the presence of unknown pre- and post-break parameters, laying out a framework of statistical decision making as new data arrive with a well-defined objective function that balances the tradeoff between false alarms and delayed detection. The procedure would have accurately identified the NBER business cycle chronology had it been in use over the past 33 years; in particular, it the beginning of the 2007-09 recession 5 months ahead of the date announced by the NBER.

Node Alertness - Monitoring Change at the Local Level in Large Evolving Graphs
Mirco Mannucci (HoloMathics)

The author discussed Graph Mining, applying it to the continuous monitoring and detection of patterns of evolution in rapidly evolving big graphs. The monitoring activity was pushed at the node level, whereas some global knowledge merger integrates the individual discoveries into a global picture. Preliminary implementation of the technique using Apache Spark GraphFrames was discussed, as well as applications and future directions.

Finding Needles in Many Haystacks: A General-purpose Distributed Approach to Large-scale Learning
Carlotta Domeniconi (George Mason University)

The problem to be discussed was the scalability of machine learning algorithms in a big data context, given the background that the traditional solutions of (a) reducing size by sampling or (b) implementing parallelization, present difficulties. The former often fails because the discovery of useful patterns may require the analysis of the entire collection of data (the “needles in the haystacks problem”); the second because either techniques that customize individual algorithms typically do not generalize to other algorithms or because the many standard parallelization methods used in this customization can be inefficient when used for the iterative computations which are so often a core part of machine learning algorithms.

In this talk, the author introduced a general-purpose method for distributed machine learning combining ideas from stochastic optimization and ensemble learning, achieving scalable machine learning and making it easily adaptable to a variety of heterogeneous grid or cloud computing scenarios. In a nutshell, the emergent behavior of a grid of learning algorithms makes possible the effective processing of large amounts of data, culminating in the discovery of that fraction of data that is crucial to the problem at hand. The emergent behavior only requires local interaction among the learners, resulting in a high speed-up under parallelism. The method does not sacrifice accuracy like sampling does, while at the same time it achieves a general scalable solution that doesn’t need to be tailored for each algorithm.

SAS® Visual Forecasting: a Cloud-Based Time Series Analysis and Forecasting Ecosystem
Michele Trovero (SAS)

The discussion centered around SAS® Visual Forecasting, a new cloud-ready platform for massive-scale time series analysis and forecasting based on the SAS® Viya™ architecture: open, elastic, powerful, platform supporting popular open source and SAS language coding in a single environment. The platform provides a resilient, distributed, optimized forecasting ecosystem for cloud computing. The functionality includes generic time series analysis scripting, automatic forecast model generation, automatic variable and event selection, automatic model selection, hierarchical forecasting, advanced support for time series analysis (time domain and frequency domain), time series decomposition, time series modeling, signal analysis and anomaly detection (for IoT), and temporal data mining.

New Analytical Methods for Anomaly Detection in High-Frequency Sensor Data
Byron Biggs (SAS)

Byron presented a framework and the architecture employed to detect anomalies using time and frequency techniques in high volume data, including streaming real-time data. The process is implemented in SAS® Viya™ and SAS® Visual Forecasting, and works with a scripting language that supports cloud-based time series analysis with

examples in SAS language, Python and Lua. The key enabling technology is SAS® Event Stream Processing, which analyzes and understands millions of events per second, detecting patterns of interest as they occur. The results show the correct actions to take, what alerts to issue, which data to store and which events to ignore. Examples included detection of industrial system degradation and health emergencies for chronic health conditions.

Now-Casting and the Real-Time Data Flow
Domenico Giannone (Federal Reserve Bank of New York, second keynote speaker)

The presentation surveyed recent developments in economic now-casting with special focus on models that formalize key features of how market participants and policymakers read macroeconomic data releases in real-time, which involves monitoring many data, forming expectations about them and revising the assessment on the state of the economy whenever realizations diverge sizably from those expectations.

Topics discussed: state space representations (factor model, model with daily data, mixed-frequency VAR), now-cast updates and news, practical models (bridge and MIDAS-type equations), and future directions on the implementation of these approaches.

A Massive Data-Driven Platform for Manufacturing Analytics
Jayant Kalagnanam (IBM Research)

Dr. Kalagnanam presented ongoing work in IBM Research for Industry 4.0. This work entails (a) the use of IOT to enable realtime access to sensor data for various assets and processed in the production value chain, and (b) the use of this data to create a digital representation or model (also referred to as a cyber physical system) - an accurate representation of the physical world. Such a digital model is then used for situational awareness, anomaly detection, process monitoring and advisory control for optimizing outcomes (defined by productivity and throughput). Their work at IBM Research for Industry 4.0 experiments with a large-scale data ingestion and analytics platform that leverages statistical and machine learning techniques. The process drives cost savings and operational efficiency across the factory value chain.

Implicit Stochastic Gradient Descent for Robust Statistical Analysis with Massive Data Sets

Panos Toulis (Booth School of Business, University of Chicago)

This paper described an implicit procedure that combines fast computation with a solution to the stability issues seen in stochastic gradient descent. The author also showed simulations and real-world data analysis using the R package *sgd*.

Derivation of Machine Learning Strategies and Optimization Modules based on Potential Theory

Nadia Udler (Fordham University)

The talk centered on a general method to derive a library of essential optimization modules based on potential theory [Kaplinsky, Propoi, Prudnikov]. This theoretical approach is based on the randomization of an objective function and the computation of directional derivatives of the randomized functional. Using the gradient of the potential field it is possible to construct algorithms in terms of the means of the underlying movements in the space of random vectors, effectively combining the exploration power of random-search algorithms with the exploitation power of direct-search algorithms. The method is a generic schema that allows obtaining well-known heuristic procedures such as Nelder-Mead, Shor's r-algorithm (based on space dilation), Covariance Matrix Adaptation Evolution Strategy and others. In many cases, the improved algorithms are not created from scratch; rather, the variable metric approach can be used to improve the performance of the existing optimization software (e.g., applied to Nelder-Mead).

As a generalization of a systematic process for the construction of optimization algorithms used in smooth

optimization, examples were implemented in modern data mining software such as Python (Scikit-Learn, PyOpt) and Matlab (Global Optimization, Statistics, and Machine Learning toolboxes). One particular example demonstrated how AlgoPy can be combined with a smooth optimization algorithm such as gradient descent method for optimization of non-differentiable functions using a tutorial developed at Fordham University, Graduate School of business, Masters in Quantitative Finance program. The same process can be used with TensorFlow.

A Bayesian Model for Forecasting Hierarchically Structured Time Series
Julie Novak (IBM Research)

Hierarchical forecasting is a subject of intensive research given the availability of large amounts of data and, with it, the possibility of forecasting at the same time (a) a metric for the overall organization and (b) components of the said metric. However, both forecasts (the total and the components) will rarely agree. The talk proposed a Bayesian hierarchical method that treats the original forecasts as observed data which are then updated and obey the hierarchical organizational structure. Overall, a novel approach to hierarchical forecasting was developed that provides an organization with optimal forecasts that reflect their preferred levels of accuracy while maintaining the proper additive structure of the business.

Demand Forecasting from Massive Usage Logs
James Wright (Microsoft Research)

This presentation described a demand modeling exercise of server utilization based on complete daily usage logs from a large online service provider that delivers services with a rich set of attributes that have complex effects upon both customer demand and capacity costs. This forecasting would help to predict accurately the demand for distinct service offerings, a crucial task both for forecasting revenue and for planning capacity. As part of the effort, the reasons for churning are modeled. (One of the participants asked why, instead of modeling churning or to use it as a reference for their models, they were not directly asking customers why they were leaving the service.)

Prediction and Explanation in Social Systems
Jake Hofman, Amit Sharma, Duncan Watts (Microsoft Research)

Jake Hofman argued that, historically, social scientists have sought out explanations of human and social phenomena that provide interpretable causal mechanisms, while often ignoring their predictive accuracy. However, the increasingly computational nature of social science is beginning to reverse this traditional bias against prediction, and has also highlighted three important issues that require resolution. First, current practices for evaluating predictions must be better standardized. Second, theoretical limits to predictive accuracy in complex social systems must be better characterized, thereby setting expectations for what can be predicted or explained. Third, predictive accuracy and interpretability must be recognized as complements, not substitutes, when evaluating explanations. Resolving these three issues will lead to better, more replicable, and more useful social science.

Predicting Signal Cycle in Smart Cities Using H-VAR
Bahman Moghimi , Abolfazl Safikhani*, Camille Kamga (City College of New York, *Columbia University)

A multivariate time series model was developed to analyze the behavior of signal cycles coming from multiple intersections in a fully actuated setup (that is, making use of intelligent technologies such as detectors, sensors, wireless communication, vehicle-to-vehicle, and vehicle-to-infrastructure communications in an integrated framework). The model characterizes the cycle lengths using H-VAR (High-dimensional Vector AutoRegression). The proposed method reduces the number of parameters by LASSO-type penalization techniques. As shown in simulation studies, the real-time one/multiple step prediction of cycle lengths performs reasonably well and outperforms univariate models such as ARIMA.

Spatio-temporal Modeling of Taxi Demands in NYC using STARMA Models

Sandeep Mudigonda (City University of New York)

The demand for taxis was modeled as a dynamic spatio-temporal process. A few facts show the magnitude of the problem: in 2015, 21,263 street hail taxis conducted between about 150,000 to 600,000 trips per day, with a spatial variability ranging from only 3,150 pickups in the Bronx to about 383,000 pickups in Manhattan on an average day. The authors used the GPS-enabled spatio-temporal historical demand for taxis aggregated to several sub-regions within the city (the data was provided by the Taxi and Limousine Commission of New York City). To understand the demand's behavior through space and time, the model relied on a spatio-temporal ARMA (STARMA) framework, a well-established and flexible class of empirical models proven useful in modeling time histories of spatially located data, introduced by Phillip Pfeifer and Stuart Deutsch in the 1980s.