IIF Workshop

Forecasting with Massive Data in Real Time

Tools and Analytical Techniques

Microsoft Technology Center, NY

April 20-21, 2017

		Thu - April 20, 2017	Fri - April 21, 2017
8	0		
	15	Registration - Breakfast	
	30		
	45	**Claudio Antonini (BNY Mellon) - Welcome - Introduction**	Registration - Breakfast
9	0	**Heather Shapiro (Microsoft)**	
	15	**Paul Cohen (DARPA)**	**Domenico Giannone (Federal Reserve Bank)**
	30	Beyond Big Data: Technology to Understand Complicated Systems	Now-Casting and Real-Time Flow
	45		
10	0	**Keynote presentation**	**Keynote presentation**
	15	Break	Break
	30	**Michael Kane (Yale)**	**Jayant Kalagnanam (IBM)**
	45	A Cointegration Approach to Identifying Systemic Risk in Markets	A Massive Data-Driven Platform for Manufacturing Analytics
11	0		
	15	**Haixi Li (Freddie Mac)**	**Nadia Udler (Fordham University)**
	30	Asymptotically Optimal Identification of Structural Breaking Point in	Machine Learning Strategies Design - Potential Theory
	45	Real Time with Application to Dating Recessions	
12	0	**Panos Toulis (Booth, University of Chicago)**	**Julie Novak (IBM)**
	15	Implicit Stochastic Gradient Descent for Robust Statistical Analysis with	Bayesian Hierarchical Forecasting
	30	Massive Data Sets	
	45	Lunch	Lunch
1	0		
	15		
	30		
	45	**Mirco Mannucci (Holomathics)**	**James Wright (Microsoft Research)**
2	0	Node Alertness - Monitoring change at the Local Level in Large	Demand Forecasting from Massive Usage Logs
	15	Evolving Graphs	
	30	**Carlotta Domeniconi (GMU)**	**Jake Hofman (Microsoft Research)**
	45	Finding Needles in Many Haystacks: A General-purpose Distributed	Limits in the Prediction/Explanation in Social Systems
3	0	Approach to Large-scale Learning	
	15	Break	Break
	30	**Michele Trovero (SAS)**	**Abolfazl Safikhani (Columbia University)**
	45	SAS® Visual Forecasting: a Cloud-Based Time Series Analysis and	Predicting Signal Cycle in Smart Cities Using H-VAR Mode
4	0	Forecasting Ecosystem	
	15	**Byron Biggs (SAS)**	**Sandeep Mudigonda (CCNY)**
	30	New Analytical Methods for Anomaly Detection in High-Frequency Sensor	Spatio-temporal Modeling of Taxi Demands in NYC using STARMA Models
	45	Data	

Titles, Authors, and Abstracts

(In chronological order)

---------- Thursday, April 20, 2017 ----------

Beyond Big Data: Technology to Understand Complicated Systems

Paul Cohen (DARPA)

A difficult question for big data analytics is 'why?' Specific questions include the following: Why did this drug stop working? Why is there food insecurity in specific regions of the world? We have component-level, not system-level, understanding of the complicated systems on which we depend for survival. Two current Defense Advanced Research Projects Agency (DARPA) programs are addressing the need for causal models and quantitative analysis to answer the 'why.'

The World Modelers program 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. The goal is to develop approaches that can accommodate and integrate dozens of contributing models connected by thousands of pathways—orders of magnitude beyond what is possible today—to provide clearly parameterized, quantitative projections within weeks or even hours of processing, compared to the months or years it takes today to understand considerably simpler systems. The first use case of World Modelers is food insecurity resulting from interactions among climate, water, soil, markets, and physical security.

Big mechanisms are large, explanatory models of complicated systems in which interactions have important causal effects. The collection of big data is increasingly automated, but the creation of big mechanisms remains a human endeavor made increasingly difficult by the fragmentation and distribution of knowledge. The goal of the Big Mechanism program is for machines to help humans to model and analyze very complicated systems by reading fragmented literatures and assembling the reasoning with models. The domain of the program is cancer biology with an emphasis on signaling pathways. Although the domain of the Big Mechanism program is cancer biology, the overarching goal of the program is to develop technologies for a new kind of science in which research is integrated more or less immediately—automatically or semi-automatically—into causal, explanatory models of unprecedented completeness and consistency. To the extent that the construction of big mechanisms can be automated, it could change how science is done.

A Cointegration Approach to Identifying Systemic Risk in Markets

Michael Kane (Yale University)

In domains such as financial markets, it can be exceedingly difficult to predict what will happen. For example, news events may occur at any time; they can affect markets in a variety of ways, and they are not amenable to predictive models. However, it is often possible to gauge the systemic risk to tell the extent to which an event can affect the market. This talk analyzes the use of cointegration in financial markets to assess systemic risk, using the 2010 FlashCrash as a case study. Based on this analysis, we will explore an alternative to current, single-stock circuit breaker/collar rules employed by FINRA to control market volatility.

Asymptotically Optimal Identification of Structural Breaking Point in Real Time with Application to Dating Recessions

Haixi Li (Freddie Mac), Xuguang Sheng (American University)

In this paper, we develop a procedure to detect an abrupt structural change in real time in the presence of unknown pre- and post-break parameters. We lay 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. We show that the stopping time is asymptotically optimal in the sense of Shiryaev (1978). The proposed procedure performs well in a large scale Monte Carlo simulations. Using a monthly "real time" dataset of Chicago Fed National Activity Index, we find that our approach would have accurately identified the NBER business cycle chronology had it been in use over the past 33 years. Notably, the proposed procedure is able to detect the beginning of the 2007-09 recession 5 months ahead of the date announced by the NBER. In the era of big data, we believe that such a procedure would enable us to identify the common structural break in massive data in real time, as illustrated by another application in having successfully detected the breaks in global uncertainty.

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

Panos Toulis (Booth School of Economics – University of Chicago)

Stochastic Gradient Descent (SGD) is the jackknife of modern statistical analysis with very large data sets--such as deep learning--but the standard procedures can be hard to tune, and cannot effectively combine numerical stability with statistical efficiency. We present an implicit procedure that combines fast computation with a solution to the stability issues, without sacrificing statistical efficiency. Extensive simulations and real-world data analysis are carried out through our "sgd" R package. Implicit methods are poised to become the workhorse of estimation with large data sets.

Node Alertness - Monitoring change at the Local Level in Large Evolving Graphs

Mirco Mannuci (HoloMathics)

Graph Mining is by now an established area of Data Mining. Detecting patterns of evolution in dynamic graphs has been already investigated in several quarters. However, continuous monitoring change in rapidly evolving big graphs is still a challenge. This article argues for pushing the monitoring activity at the node level, whereas some global knowledge merger will integrate the individual discoveries into a global picture.

We shall discuss some preliminary implementation of this technique using Apache Spark GraphFrames, as well as applications and future directions.

(Dr. Mannucci, CEO of HoloMathics, is working as a Big Data Analytics Consultant at FINRA Technology.)

Finding Needles in Many Haystacks: A General-purpose Distributed Approach to Large-scale Learning

Carlotta Domeniconi (George Mason University), Uday Kamath (BAE Intelligent Systems)

The need for mining massive data has become paramount in areas like security, education, web mining, social network analysis, and a variety of scientific pursuits. However many traditional supervised and unsupervised learning algorithms break down when applied in big data scenarios: this is known as the "big data problem". Among other concerns, big data presents serious scalability difficulties for these algorithms.

The two most common ways to get around the scalability issue is to either sample the data to reduce its size, or to customize the learning algorithm to improve its running time via parallelization. Both of these have problems. Sampling often fails because the discovery of useful patterns can require the analysis of the entire collection of data (we call this the needles in the haystacks problem). Techniques that customize individual algorithms typically do not generalize to other algorithms. Further many standard parallelization methods used in this customization can be inefficient when used for the iterative computation which is so often a core part of machine learning algorithms.

In this talk, we discuss current approaches and tools in use to address these shortcomings. In particular, we introduce a method for distributed machine learning which directly tackles key problems posing challenges to successful and scalable mining of big data. We bring together ideas from stochastic optimization and ensemble learning to design a novel and general paradigm to achieve scalable machine learning. The method is general-purpose with regard to the machine learning algorithm and 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 Institute)

SAS® Visual Forecasting is a new product for massive-scale time series analysis and forecasting based on the SAS® Viya™ architecture. SAS® Viya™ is a new cloud-ready platform purposefully designed to meet the analytical challenges of today and tomorrow. It is an open, elastic, powerful, platform supporting popular open source and SAS language coding in a single environment. SAS® Visual Forecasting provides a resilient, distributed, optimized forecasting ecosystem for cloud computing. The environment provides generic time series analysis scripting, automatic forecast model generation, automatic variable and event selection, and automatic model selection, and supports hierarchical forecasting. It provides 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 Institute)

The connected world provides new opportunities to detect system degradation or anomalies prior to failure. Innovations in noise reduction, human activity identification using data mining & machine learning, and detection of anomalies using time-frequency techniques in high volume data, including streaming real-time data, is a focus area for new methodologies.

I will present the architecture at the base of SAS® Viya™ and SAS® Visual Forecasting and describe the scripting language that supports cloud-based time series analysis with examples in SAS language, Python and Lua.

SAS® Event Stream Processing is a key enabling technology for this analysis. SAS® Event Stream Processing 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.

This presentation will cover new analytical models for high-frequency streaming data. Examples include 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 – New York)

The term now-casting is a contraction for now and forecasting and has been used for a long time in meteorology and recently also in economics. In this presentation we survey recent developments in economic now-casting with special focus on those 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 sizeably from those expectations.

Topics to discuss will include state space representations (factor model, model with daily data, mixed-frequency VAR), now-cast updates and news, and practical models (bridge and MIDAS-type equations).

Empirical applications will cover GDP nowcasting and a daily index of the state of the economy. Current and future directions on the implementation of these approaches will also be described.

A Massive Data-Driven Platform for Manufacturing Analytics

Jayant Kalagnanam (IBM Research)

Industry 4.0 is the use of IOT to enable realtime access to sensor data for various assets and processed in the production value chain, and 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. The digital model is then used for situational awareness, anomaly detection, process monitoring and advisory control for optimizing outcomes (defined by productivity and throughput). I will present ongoing work in IBM Research for Industry 4.0 that is experimenting with a large-scale data ingestion and analytics platform that leverages statistical and machine learning techniques to drive cost savings and operational efficiency across the factory value chain.

Machine Learning Strategies Design – Potential Theory

Nadia Udler (Fordham University)

Many pressing real world problems can be stated as problems of global optimization, where a target function is given as a black box, known only by its values. These problems range from parameter tuning in robotics and the optimization of chemical processes to the analysis of biological systems; typical examples from financial engineering include model calibration, pricing, hedging, VaR/CVaR computation, credit rating assignment, forecasting, and strategy optimization in electronic trading systems. In all these cases, the choice of the algorithm for decision making depends on the type of the objective function and the availability of a priori information, as well as user constraints such as limited budget or energy consumption.

We present 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 we construct algorithms in terms of the means of the underlying movements in the space of random vectors Effectively, this process combines the exploration power of random-search algorithms with the exploitation power of direct-search algorithms. The method is a generic schema that allows to obtain 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, we show how to use variable metric approach 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 have been implemented in modern data mining software such as Python (Scikit-Learn, PyOpt) and Matlab (Global Optimization, Statistics, and Machine Learning toolboxes). In particular, we will demonstrate 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. TensorFlow can be used in a similar fashion.

A Bayesian Model for Forecasting Hierarchically Structured Time Series

Julie Novak (IBM Research)

An important task for any large-scale business is to prepare forecasts of business metrics, such as revenue, cost, and event occurrences, at different time horizons (e.g. weekly or quarterly intervals). Often these business organizations are structured in a hierarchical manner by line of business, division, geography, product line or a combination thereof. In many situations projections for these business metrics may have been obtained independently and for each level of the hierarchy. The problem with forecasts produced in this way is that there is no guarantee that forecasts are aggregate consistent according to the hierarchical structure of the business, while remaining as accurate as possible. In addition, it is often important for the organization to achieve accurate forecasts at certain levels of the hierarchy according to the needs of users. We propose a Bayesian hierarchical method that will treat the "base" forecasts (those which were initially provided) as observed data which are then updated and obey the hierarchical organizational structure. In addition, we are able to set up a heterogeneous loss function to obtain higher accuracy at the levels prescribed by the user. We develop a novel approach to hierarchical forecasting 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)

Accurately forecasting the demand for distinct service offerings is a crucial task, both for forecasting revenue and for planning capacity. This talk will describe a demand modeling exercise based on complete daily usage logs from a large online service provider that provides services with a rich set of attributes that have complex effects upon both customer demand and capacity costs.

Prediction and explanation in social systems

Jake Hofman, Amit Sharma, Duncan Watts (Microsoft Research)

Historically, social scientists have sought out explanations of human and social phenomena that provide interpretable causal mechanisms, while often ignoring their predictive accuracy. We argue that the increasingly computational nature of social science is beginning to reverse this traditional bias against prediction; however, it 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 Mode

Bahman Moghimi (CUNY), Abolfazl Safikhani (Columbia University), Camille Kamga (CUNY)

Traffic signals as a part of intelligent transportation system can make significant role toward making cities smart. Conventionally, most traffic lights are designed as fixed-time control, which induces a lot of slack time (unused green time) in transportation system. However, with the rise of intelligent technologies such as detectors, sensors, wireless communication, vehicle-to-vehicle, and vehicle-to-infrastructure communications, signals have recently been designed intelligently using actuated control. Actuated traffic lights control congestion in real time and are more responsive to the variation of traffic demands. The point is that each signal needs to capture immense amount of data which is transmitting from different sources including vehicles, traffic-devices imbedded with infrastructure, pedestrians, and bikes, to traffic lights in real time. On the other hand, high dimensionality of data associated with one signalized intersection to another signal makes the controlling problem in urban transportation network a challenging one. In this paper, we develop a multivariate time series model to analyze the behavior of signal cycle coming from multiple intersections in a fully actuated setup. For that, ten signals have been modeled along a corridor, Hylan Boulevard in Staten Island, New York City, with different spacing among them together with multiple choices of traffic demands. For an isolated signal, a family of time series model like ARIMA model can be handy for predicting the next value of cycle length. However, when there are many signals placed along the corridor with different spacing and configurations, the cycle length variation of such signal is not just related to its own values, whereas it is affected by the platoon of cars coming from neighboring intersections. Since there are too many parameters to estimate due to having multiple intersections (at least 100 parameters), VAR models as the most well-known multivariate time series models will not be able to perform well and will make poor forecasting of cycle lengths. To mitigate this problem, we model the cycle lengths using a new developed method called HVAR (High-dimensional Vector AutoRegression). The proposed method reduces the number of parameters by LASSO-type penalization techniques, and as shown in our simulation studies, the real-time one/multiple step prediction of cycle lengths using this method performs reasonably well, and further outperforms the univariate models such as ARIMA. The importance of forecasting performance for signal cycles is coming from the fact that good prediction of signal cycle will help the controlling logic to make more precise decision about the upcoming transit (bus or streetcar) and facilitate its movement along with the use of transit signal priority tactics.

Spatio-temporal modeling of taxi demands in NYC using STARMA models

Sandeep Mudigonda (City College of New York)

Ride hailing services have been an important part of urban transportation. Traditionally, ride hailing is performed by customers on the curbside of streets. Given the highly dynamic urban space in a metropolis such as New York City, the spatio-temporal variation in demand for taxis is impacted by various factors such as commuting, weather, special events, parades, road work and closures, disruption in transit services, etc. The demand for ride hailing taxis is highly variable with a maximum of about 600,000 to a minimum of about 150,000 trips per day provided by 21,263 street hail taxis in 2015 [1]. This demand also has a high spatial variability with about 383,000 pickups in Manhattan and only 3,150 pickups in the Bronx on an average day. These taxis travelled approximately 460 million miles in 2015 [2]. Due to the myriad factors impacting demand, which may or may not be known in advance, there is scope for taxis driving around seeking rides. In this study, we model the demand for taxis as a dynamic

spatio-temporal process. We use the GPS-enabled spatio-temporal historical demand for taxis in the year of 2015 (provided by the Taxi and Limousine Commission of New York City) aggregated to several sub-regions within the city. In order to understand the demand's behavior through space and time, we use a spatio-temporal ARMA (STARMA) model. STARMA model is a well-established spatio-temporal process introduced by Pfeifer, P. E., & Deutsch, S. J. ([3] and [4]), and it has been applied in many different disciplines such as hydrology, transportation, climatology, economics, health sciences, etc. Modeling the demand through time in all the sub-regions simultaneously is a high-dimensional problem since the number of parameters in the model is proportional to the squared of the number of sub-regions. STARMA reduces the number of parameters dramatically by governing a neighborhood structure between the regions. This structure is also useful in capturing the spatial dependence of the demand between the regions and further makes the results more interpretable. We measure the forecasting performance of the proposed model using the out-of-sample mean squared prediction error (MSPE), and we show that our model is outperforming some alternative algorithms such as ARMA model. Given that there are about 12 million taxi trips a month that amounts to 2 GB of data, a demand forecasting model with accurate spatial and temporal predictability is very useful. Particularly, the proposed model has the ability to forecast the taxi demand few steps ahead in the future at various locations in NYC, and this enables the agencies for the real-time provision of demand-sensitive taxi dispatching for various locations and specific times of the day over the year. This is particularly useful for the operating agency so that empty ride-seeking taxi trips and thus the fuel burned can be lowered. Such demand-sensitive dispatch also has an environmental benefit by reducing the emissions associated to empty ride-seeking taxi trips. Additionally, from a policy standpoint, the spatio-temporal structure inferred from the demand data provides a basis for regulating agencies to explore cordon pricing initiatives.

References

[1] http://www.nyc.gov/html/tlc/downloads/pdf/2016_tlc_factbook.pdf
[2] http://www.nyc.gov/html/tlc/html/about/factbook.shtml
[3] Pfeifer, P. E., & Deutsch, S. J. (1980). A three-stage iterative procedure for space-time modeling, Technometrics, 22 (1), 35-47. Also, at https://www.ncjrs.gov/pdffiles1/Digitization/60692NCJRS.pdf.
[4] Pfeifer, P. E., & Deutsch, S. J. (1981). Variance of the sample space-time autocorrelation function, Journal of the Royal Statistical Society, Series B (Methodological), 28-33.