Computational Social Science Fusion Analytics: Combining Machine-Based Methods with Explanatory Empiricism

$^{\Delta}$**Robert J. Kauffman,** ‡**Kwansoo Kim,** °**Sang-Yong Tom Lee**
$^{\Delta}$Singapore Management University, ‡ Izmir University of Economics, °Hanyang University
$^{\Delta}$rkauffman@smu.edu.sg, ‡ kwansoo.kim@izmirekonomi.edu,°tomlee@hanyang.ac.kr

Abstract

This article discusses the emergence of a computational social science analytics fusion as a mainstream scientific approach involving machine-based methods and explanatory empiricism as a basis for the discovery of new policy-related insights for business, consumer and social settings. It reflects the interdisciplinary background of the new approaches that the Hawaii International Conference on Systems Science has embraced over the years, and especially some of the recent development and shifts in the scientific study of technology-related phenomena. It also has evoked new forms of research inquiry, blended approaches to research methodology, and more pointed interest in the production of research results that have direct application in various industry contexts. We review background knowledge to showcase the methods shifts, and demonstrate the new forms of research, by showcasing contemporary applications that will be interesting to the audience on the occasion of the HICSS 50th anniversary.

1. Introduction

IT innovation, specifically computing power, has influenced our lives enormously, and will continue to change our society. As technology advances, increasingly abundant digital data in medical records, online activity histories, and energy usage files can be better analyzed with the help of various analytics methods grounded in Computer Science, Statistics, and Economics [38], leading to new ways to understand people, organizations and society [3, 15, 47]. We call this a *computational social science analytics fusion* [16].

This article discusses scientific research issues from the perspective of fusion analytics and big data-based machine learning, characterized by increasing insight capabilities with declining costs over time. New ways of doing analytics opens up the possibility for innovative thinking and novel contributions to scientific discovery in interdisciplinary contexts.

Although researchers have tried to combine Computer Science, Psychology, Sociology, Regional Economics, Biostatistics, and other bodies of theory and methods knowledge to assert interdisciplinary solutions to leading business, consumer and social problems, these research areas have remained fairly independent in the perspectives they emphasize.[1]

We will discuss scientific research issues for big data and fusion analytics. The new policy analytics that we envision balances the strengths of multiple methods, while contributing to improved analytical performance in the presence of new levels of data access and research inquiry design innovation.[2] Big data enables a broad area of study in which technology meets policy and social issues, facilitating social and civic empowerment, and enhancing and expanding stakeholder participation in policy and planning [11]. Data analytics are aimed at uncovering value from data, in a variety of ways that permit modeling, experimentation, simulation, and other kinds of scientific approaches to discover new knowledge.

But what kinds of research questions will require novel big data and fusion analytics research designs to address them? This requires the identification of what analytics methods can be used from different disciplines, and what new knowledge can be discovered the issues at hand, as well as where the data will come from and be collected. Some examples of applied contexts that are appropriate can be identified based on the popular press, government agency calls for research, and recent research on organizations (e.g., the influence of friends in social media and randomized experimental designs [4], ad position auctions with consumer search [6], and ranking hotels on travel search engines by mining user-generated and crowd-sourced content [31].

As an increasingly important segment of the broader information and analytics market, big data is having an impact. Organizations have been implementing big data projects and deploying new ways of discovering market, product and consumer knowledge, and have reported new advantage compared to their prior reliance on traditional marketing

[1] Wellman [59] pointed out the benefits associated with using CS and AI to explain important relationships in the Social Sciences, especially with agent-based economies. Only recently has CS moved to a position of innovation in interdisciplinary studies, and fusion analytics has taken off (e.g., social network analysis) [43].

[2] New learning algorithms and the availability of big data have supported new directions in research involving IS, e-commerce, strategy, marketing, financial services, and manufacturing [48].

URI: http://hdl.handle.net/10125/41775
ISBN: 978-0-9981331-0-2

research methods [33, 45, 57]. Indeed, according to IBM, every field has been or will be changed by the large amount of data available [37].[3]

The current technologies and others that are now emerging offer astonishingly rich detail concerning human and societal behavior, contextual behavior and responses, and the attitudes, preferences, and sentiment of different individuals and groups [32]. This stream of research has examined how people view big data and to what extent they are currently using it to derive valuable results. Data analytics allow us to find repetitive patterns and to adjust predictive models to understand how likely it is to observe various kinds of consumer behavior. Our consideration of data analytics that apply data mining, natural language processing, machine learning, statistics and econometrics, and other methods, prompts us to think about what insights and predictions can or cannot be obtained from big data, and at what level of accuracy.

A recent trend in econometrics is to use machine learning to help establish causal. This complements raditional econometrics, according to Athey [5]. When machine learning is combined with econometrics, the authors often construct dependent and independent variables with the help of the machine, including the use of error terms from prior models [44, 56]. The IS discipline has begun to use this too.

The insights shares in this article are at two levels. We encourag researchers to apply interdisciplinary fusion analytics research approaches. They can include pattern recognition for their data and other modes of machine-based discovery, with explanatory analysis that yields insights into the marginal impacts of different policies. Even more interesting with fusion analytics is the possibility to blend data mining and machine learning techniques with econometrics and statistical methods. This will make it possible to do more sophisticated counterfactual impact analysis. These are ways that analytics can produce more managerial insighst, and they offer new directions for researchers support practitioners' need to understand how to build powerful evidence to support their business, consumer and social policies. We ask: (1) Do interdisciplinary fusion analytics accelerate business and research paradigm shifts? (2) Are the insights sufficient to produce policy impacts and ideas for what managers can do to improve whatever process, operation or activities they are studying? In this context, the policy implications need to be understood, in terms of how they will be different and for what issues arise with big data analytics based on single methods (e.g., econometrics or statistics only).

2. Interdisciplinary Background

2.1 Policy Analytic Perspective

Policy analysis evaluates and legitimizes policy strategies based on some set of criteria, such as equity, economic efficiency, cultural and social acceptability, and legality. In this approach to problem formulation, evaluation and selection of policies, it is necessary to consider the interests and preferences, the applicable priorities, and the values of a diverse set of stakeholders [12].[4]

Also, forecasting methods need to be integrated into decision-analytic frameworks, if they are to be useful. This is to ensure that there is recognition of how to answer "What should we do?" versus "What should we have done?" A consequence of adopting a decision-theoretic approach to forecasting is the need to develop ways of synthesizing the set of available prediction methods. Such an approach involving synthesizing different kinds of information to support forecasting is in contrast to the conventional selection methods. In the policy analysis process, whose values should we invoke? Are the interests of certain groups more important than those of others? How can we effectively gauge the preferences of different stakeholders? Big data and related methods developments in the area of sentiment analysis have the potential to be useful for answering these kinds of questions [13]. This is true for understanding consumer preferences, especially based on sentiment concerning issues and views that can be mined from online articles, blogs, tweets, Facebook posts, and other sources.

To understand who to gauge the importance of uncertainty, policy analysts always need to create appropriate contexts around the data they use. One way to achieve this is through *data fusion*, the process of combining less reliable sources of information to create more accurate and useful data points. This is traditionally true with such methods as Delphi sessions, but today it is much more applicable to settings in which there are social comments appended to geospatial location data, such as in Twitter and FourSquare. Another way to manage uncertainty is through advanced math, such as robust optimization techniques and fuzzy logic approaches.

2.2. Machine-Based Methods of Computer Science

The technical issues related to data capture,

[3] For example, information about online users is necessary to optimize credit-card rewards programs and business partnerships and achieve high value [30, 49].

[4] *Decision theory* was developed to prescribe consistency in choices under uncertainty for a rational individual. Policy analysis for firm decision-making is different. It needs modifications to the underlying theory to deal with problems of group consensus, ill-defined objectives and disparate information sources [13].

streaming, archiving, and parallel computing are central to CS methods. For example, the Hadoop family of technologies and services originated from Google, cloud computing, and other technical environments. Data scientists have been trying to detect patterns in data as a way of advancing knowledge. This paradigm includes new algorithmic, computational, and pattern recognition tools to produce value from available data [8, 36]. In industry, it is typical that organizations face challenges to learn about the potential benefits and constraints of new data sources. So it is logical to leverage computing power that has become available for new statistics software to share ideas and experience, to make it so new sources contribute better to data-driven business performance.

Data specialists now work within the new data-driven science of analytics. Computing power can create, run, and test thousands of hypotheses, models, and simulations quickly. Algorithm use is an essential element of big data analytics. In some circumstances, machines can learn from data, and benefit from the intersection of machine learning, artificial intelligence, and data processing methods [27]. What is important is whether we can make scientific discoveries, and whether we can learn new theories, create deeper explanations, and make more effective predictions. In addition, it is especially good to foster new methodologies based on computer analytics to drive something to produce very high value.

2.3. Technology and IS Perspective

IBM has defined by big data in terms of the four V's: *volume, velocity*, *variety*, and *veracity*. The new fusion paradigm that we have proposed takes advantage of opportunities to combine these dimensions. For example, variety is the most interesting dimension of big data in technical terms, since it implies that many tools and approaches are needed to process it in its different forms (e.g., large cross-sections, lengthy panels, streaming flows, etc.). Putting data together from Web sites, user-generated content, social media, and smart mobile platforms allows researchers to explain and predict individual behavior and detect trends that occurring in context.

Technology research with big data leverages data infrastructure to produce analytics. Managing a platform of services with streaming data, and trying to build useful analytics to manage and adjust its performance is an exciting research area. Researchers can build on literature that looks at what drives different infrastructure choices. For example, there has been work on the optimization of cost, benefits and service levels [7, 26], and game-theoretic and principal agent models have been useful too [9, 46].

Important opportunities exist for interdisciplinary collaboration in big data. Among the various information sciences groups in universities, business and information school IS groups have to be pioneers in interdisciplinary collaborations. This is true in e-commerce and digital marketing, for example, but interdisciplinary should also extend to non-business and social problems, which go beyond the typical spectrum of business school IS research. Big data analytics revolution has pushed the boundaries of the IS discipline outward in this new environment.

2.4. Explanatory Empiricism

Traditional econometric methods generally assume that data observations are independent or grouped, as in panel data, or are linked by time. However, individuals in a social network may be interconnected in much more highly complex ways. So the point of econometric modeling is often to uncover exactly what are the key features of this dependence structure [4].

Developing methods that are well suited to these settings is a challenge for econometricians as a result [39, 52]. There has been a remarkable amount of work on the statistical and machine-learning techniques that undergird these applications. These methods are increasingly used, although they have been rarely applied in empirical microeconomics.

2.5. Creating the Policy Analytics Fusion

Different industries and disciplines have realized the potential of working with diverse and large datasets. There are a growing number of those units in healthcare and public health, education and public transport, government services, marketing and retailing, among others. Industry and government researchers in these areas need behavioral, social, and technical approaches for their data analytics. They understand the power of diverse data sources, but lack the technical capabilities for powerful analytics, to take advantage of big data in their environments.

In decision-making, context is key. In the late 1990s and 2000s, Marketing was the object of intense interdisciplinary research work. As a discipline, Marketing has relied more on IT, Economics and CS for its data analytics. From channel choice to prices and recommendation systems, music bundling and online reviews, and social networks, interdisciplinary researchers have conducted impactful analytics research [19, 28]. Many different sources of data have been used, and there has been substantive mastery of the collection and analysis of Web data. Researchers have become very proficient in the utilization of advanced econometrics and machine learning techniques. So the seeds for working with the 4V's of big data have been planted interdisciplinary research.

Other researchers have also started to work in Finance, Operations, Strategy, and other business disciplines using the big data paradigm.

3. Business, Consumer, Social Problems

3.1. The Problems Business Intelligence Addresses

Business intelligence became a popular term in the business and IT communities in the 1990s. By the late 2000s though, *business analytics* came to the forefront [23]. More recently, *big data analytics* has become popular, and points to the use of data sets and analytical techniques in large and complex business applications, from locational sensors to social media data on friends and sentiments. Data analytics rely heavily on various data collection, extraction, and analysis technologies [18, 58].

An important issue in business applications is whether the economic benefits of business decisions are being realized, based on scientific analysis and measurement. What is needed is new methodology to understand how big data can describe the impact, and what are the economic consequences of the business decision. In that sense, we need to identify key performance variables (the impacts), their coefficients (the marginal contributions), and derive the normative implications that are present in the data.

The opportunities associated with data and analysis in organizations has generated significant interest in business analytics. In addition to the underlying data processing and analytical technologies, business analytics includes business-centric practices and methodologies that can be applied to various high-impact applications, such as e-commerce, market intelligence, e-government, healthcare, and security [20]. For example, business performance management using scorecards and dashboards is helpful to analyze a variety of performance metrics. In addition to these well-established business reporting functions, statistical analysis and data mining techniques support association analysis, data segmentation, clustering, classification, and regression analysis. They also support anomaly detection, and predictive modeling in business applications.

3.2. Consumer and Social Insights

A valuable consumer insight must be unique, enduring, and suggestive of actions that can be taken to improve organizational performance.[5] As consumers get more involved in digital economy activities, they leave *digital traces* of their behavior, from viewing habits to opinions on products and services. These digital traces are among the most important sources of data for insight creation, and are driving changes and transformation in the practice of market research.

Two different research approaches have been suggested: information visualization [1] and network analysis [43]. Both have a long history in various fields of natural sciences. However, their impact has increased, as they have overlapped with the emergence of big data.[6] The key techniques go beyond text analytics to include opinion mining, sentiment analysis, topic modeling, social network analysis, trend analysis, and visual analytics. Businesses can use them to realize value in all phases of a product or service life cycle, including changing consumer taste, influential users, ad campaign effectiveness, how to respond to crises, and competitive intelligence.

In social issue applications, Del Giudice et al. [25] studied social media in emerging economies. They covered practices and tools for emerging markets, social media, statistics, peer opinions, buzz and viral marketing, and word-of-mouth and its amplification. Harrysson et al. [35] outlined opportunities related to consumer behavior in social media. They suggest ways to: create dispersed networks to achieve a deep understanding of the business; equip employees to browse blogs to create followers and match their interests to in-store retail offerings; and use insights cross-functionally for marketing, sales, product development and customer support.

3.3. Policy Problem Applications

Business analytics have improved over the past few years, giving business users better insights. Today, massive databases require a mix of automated analysis techniques and human effort to give business users strategic insights about the activity on their websites, as well as about the characteristics of the visitors and their customers. With millions of click-streaming records being generated every day, aggregated to customer-focused records with hundreds of attributes, there is a clear need for automated techniques and finding meaningful patterns in the data. To make decisions based on data collected about their firms, they must rely on data analysts to extract information from the data or employ analytics apps that blend data analysis with task-specific knowledge.

Recent innovations in business analytics span or-

[5] For example, understanding the guilt that consumers feel when they eat something they like, and then combining it with the brand experience of a product, can lead to deep insights about "pleasure that justifies the guilt," and makes the consumer willing to pay more. Consumer insights are critical, and can be handed over to a firm's creative team to supplant the hunches and guesswork that have seemed to dominate [55].

[6] Social media analytics involve a three-stage process of capturing underlying patterns, understanding what they mean, and presenting them to others, so it is possible to identify policy actions.

ganizations and technical processes, new technologies and user interface designs, and system integration – all driven by business value. Business value is measured in terms of progress toward bridging the gap between the needs of the business user and the accessibility and usability of analytic tools. To make analytics more relevant and tangible in value for business users, the solutions must increasingly focus on specific vertical applications, tailor results and interfaces for these users, and yield managerial insights. For ease of use, simpler more effective deployment, and higher value, analytics are increasingly been embedded in larger systems. So data collection, storage, processing, and other issues specific to analytics should figure in overall system design.

Organizations that engage customers with new technologies, and glean insights through new data collection and analytics methods must first be clear on what metrics will be valuable. Farris [29] pointed out the increasing expectations to align marketing metrics with a company's core financial metrics. They explained how to gain insights from a wide variety of data. They covered traditional metrics, such as market share, sales force performance, rebates, reach and revenue, they also discussed digital age metrics, such as for web campaigns, e-commerce opportunities and leading indicators of financial performance. Save [53] also provided insights into effective mobile device-based marketing. Many young people have only connected to the Internet via mobile devices. Many developing countries, which have small numbers of traditional land-based phone lines, also have rapidly established cellular networks, so mobile phone usage has had explosive growth, catering to pent-up demand.

By broadening the effects of analytics in the business process, the solutions can go beyond customer-centric applications to support sales, marketing, supply chain visibility, price optimization, and work force analysis. Finally, to achieve the most business value, analytics solutions have to produce results that are actionable, with ways to measure the effects of changes. The challenges in the handling of big datasets are collection, validation, integrity, and security, and these issues will continue to arise with increased use of big data for policy-making and governance in the growing information society [48].

4. Mobile Phone-Based Stock Trading

4.1 Context and Data

Social media sentiment affects the trade volumes of uninformed traders,[7] which seem to be biased in systematic ways, in the mobile phone channel. Using data from Korea, we conducted an analysis that demonstrates the presence of undesirable *herding behavior* as traders react to social signals associated with trading trends from other mobile traders with whom they interact on social media [41]. Our analysis showed that uninformed mobile traders acted more on negative feedback trading for social media sentiment. We found evidence that suggests there are different patterns of trading – especially positive and negative feedback trading. They seem to appear in the short run, but disappear over time as mobile traders become more informed. The results permitted us to draw conclusions about the extent to which trading of securities via mobile phones is a "smart" channel for exchange or a "noisy" channel that offers few benefits for investor participation.

Most financial markets have basic similarities when it comes to trading via the mobile channel. So we collected Korean market stock trade volumes that were culled from mobile stock trading platforms, where they were readily available. We also used social sentiment postings, such as on Twitter and via blogs for May through September 2012. We chose to focus on about 251 firms that were discussed in Korean social media. The firms are divided into two groups listed on the Korean Exchange (KRX):[8] (1) the 125 largest firms, and (2) another 126 firms in terms of somewhat smaller business size. We divided the companies into *IT and non-IT firms*, and into index categories of the *Korean Composite Stock Market Index* (KOSPI) and of the *Korean Dealers Association Automated Quote System* (KOSDAQ).

Over the past decade, significant progress has been made in social sentiment tracking techniques that extract indicators of sentiment from social media content, particularly from large-scale Twitter and blog posts. We used a social matrix program from Daum-Soft (www.daum soft.com) in Korea, for the mining of a large number of text messages, reflecting the opinions of online users in social media. It does so by categorizing sentiments expressed in social

[7] In Finance, the technical term used when trader doesn't know much about what is going on in the market, or about a specific equity issue that is being bought or sold, is a *noise trader*. According to De Long et al. [24], stock investors also can be informed *value traders*, rationally anticipating asset value. Uninformed traders react irrationally to changing sentiment, and cause persistent mispricing of stock bids and offers. Traders who use mobile phones for stock transactions may key off of social signals as relevant *information* but they can be regarded as uninformed traders. Social media sentiment is essentially the *noise of the market-at-large.* Through our analysis, we found that mobile traders were reacting to and easily swayed by social sentiment, which may be evidence they were uninformed.

[8] This classification helped to discover more nuanced results with respect to the impacts of social sentiment on observed stock trading volumes. So this was a thoughtful choice not a random one.

media into different valences, and is especially useful for recognizing the general social mood of the public. The software examined how positive or negative mood words were attached to the keywords.[9,10]

For the opinion mining that we conducted in our study, positive sentiments were represented by the values of variables that showed a good or positive response towards a firm or its stock. For example, sentiment related to profits, improvements, innovativeness, new concepts, and about 200 other words were related to business progress. Negative sentiments were classified through variables that showed a negative response or opposition to something. Examples included sentiment on recession, faulty products, illegal activities, downward business trend, losses, and about 184 other words were associated with impediments to business development. The kinds of information we captured are described below.

4.2 Modeling

We employed *feasible generalized least squares* (FGLS) to resolve the econometric issues that characterize this and other similar settings [2, 42]. Given the cross-sectional time-series nature of our data, we specified the following model to account for unobserved fixed effects.

$$
\begin{aligned}
&\mathit{StockTradingVolume}_{it} = \beta_0 \\
&\quad + \beta_{\mathit{Industry\text{-}Level}} \text{ (Dummy vars.: } \mathit{Size}_t, \mathit{Industry}_t, \mathit{Market}_t) \\
&\quad + \beta_{\mathit{Industry\text{-}Level}} \text{ (Normal returns: } \mathit{MarketTradingVolume}_t) \\
&\quad + \beta_{\mathit{Firm\text{-}Level}} \text{ (Common factors: } \mathit{Frequency}_{it}, , \mathit{Cumulative}_{it}) \\
&\quad + \beta_{\mathit{Firm\text{-}Level}} \text{ (Social sentiment: } \mathit{Pos.}_t, \mathit{Neg.}_{it}, \mathit{Cumul}_{it}) \\
&\quad + u_i \text{ (Unobserved fixed effects for firm } i) \\
&\quad + \mu_t \text{ (Unobserved fixed effects for time } t) \\
&\quad + \varepsilon_{it} \text{ (Residuals) with } \varepsilon_{it} = \rho\varepsilon_{it-1} + \varphi_{it} \text{ and } \varphi_{it} \sim N(0, \sigma^2)
\end{aligned}
$$

Panel vector auto regression (PVAR), and the estimates that are derived from models in this area of econometrics are seldom interpreted in isolation [14]. We are interested in the impact of exogenous changes in each endogenous variable on other variables in the PVAR system.

4.3. Explanatory Econometrics

Generalized least squares (GLS) is easy to use and interpret though the method imposes stringent assumptions that sometimes are not accurate for data in real-world settings. One assumption is that the marginal effects of the explanatory variables are constant [54]. KRLS assumes observations with similar variable values should have similar outcomes on average, which reduces misspecification, and avoids a need for users to guess functional forms.

Further, KRLS uses *regularization*, a prior preference for smoother over erratic functions. As a result, KRLS minimizes over-fitting by reducing the variance and fragility of estimates, and diminishing the influence of inappropriate points [34]. It is suitable for modeling problems when the functional form is not known.

4.4. Results

We first examined the impact of social media sentiments on stock trading volume. (See Table 1.)

Table 1. Stock Trade Volume Estimation

VARIABLE	TRAD. STOCK TRAD VOLUME	MOBILE CHAN. *STOCKTRAD. VOL* (BUYING)	MOBILE CHAN. *STOCKTRAD. VOL* (SELLING)
FirmSize	-0.039 (0.089)	0.004 (0.109)	-0.018 (0.096)
Industry	0.0319 (0.089)	0.020 (0.113)	0.023 (0.099)
FinlMkt	0.058 (0.164)	-0.013 (0.093)	0.010 (0.078)
MktTradVol	**0.553***** (0.032)	**0.433***** (0.042)	**0.586***** (0.040)
Frequency	**0.140***** (0.004)	**0.103***** (0.009)	**0.122***** (0.008)
Positive	**0.020***** (0.004)	**0.019***** (0.007)	**0.039***** (0.006)
Negative	**0.010**** (0.004)	**0.027***** (0.007)	**0.019***** (0.007)
CumulFreq	0.033 (0.037)	0.060 (0.063)	-0.055 (0.059)
CumulPos	**0.117****(0.025)	0.084 (0.054)	0.052 (0.051)
CumulNeg	**-0.071**** (0.032)	**-0.119**** (0.053)	0.002 (0.050)
Constant	-0.010 (0.071)	0.008 (0.081)	-0.006 (0.065)
Obs.	16,817	16,817	16,817
# Firms	251	251	251
x^2	2,648.35***	523.65***	881.28***
VIF / CI	1.98/3.77	1.97 / 3.75	1.98 / 3.76

Notes. Dep. var.: Mobile channel *StockTradingVolume*. All variables represent percentage changes. Wald test of full model. Signif.: $p < 0.1$ = *; $p < 0.05$ = **; $p < 0.01$ = ***. No multicolinearity since VIF < 10 and CI < 30. For the error terms, we performed the Wooldridge autocorrelation test, and the Wiggins and Poi (2002) test to assess whether there was heteroskedasticity, but we found none.

The results show that uninformed traders seemed to be easily swayed by social media sentiment, while stock trading in the traditional channel probably influenced the formation of sentiment in the market more. Second, uninformed traders in the mobile channel tended to chase social signals of trends from social media, and traded stocks based on the same social signals. This was evidence for herding behavior among uninformed traders. The standard deviation of social sentiment was smaller than that of market trade volume via KOSDAQ and KOSPI. Table 2 shows the Granger causality results.[11]

[9] Daum-Soft is an Internet portal that analyzes data on social issues in SNS. It checks for user interest-related keywords via real-time monitoring technology for marketing strategy analytics to fit a wide range of business situations. It extracts social trends and user opinions formed in social networks such as Twitter and blogs

[10] The process includes: (1) collecting text disseminated on Twitter and Internet blogs in real time; (2) decomposing the collected text into word units; and (3) extracting keywords from data. After extracting keywords such as "iPhone" or "Samsung Galaxy," the software analyzed other words that occurred before or just after the keywords, to figure out whether the valences of the sentiments that were expressed by keywords were positive or negative in mood.

[11] There is an endogenous relationship between stock trading volumes and social sentiment. To deal with this, we used a Granger causality test. The results showed that the direction of causality based on the test went from social sentiment and was specifically

Table 2. Granger Causality Panel Data Results

Variables		Mobile Chan. Buy Volume	Mobile Chan. Sell Volume
Volume	*Freq.*	**9.702****	3.451
	Pos.	**7.464***	**14.586****
	Neg.	3.831	2.930
Frequency	*Vol.*	5.691	**27.129*****
	Pos.	2.631	2.577
	Neg/	**7.268***	**7.030***
	All	**15.059***	**36.655*****
Positive	*Vol.*	4.756	**21.124*****
	Freq.	**95.987*****	**94.079*****
	Neg.	2.043	1.810
	All	**108.533*****	**125.346*****
Negative	*Vol.*	0.740	**15.504*****
	Freq.	**66.192*****	**64.437*****
	Pos.	3.555	3.851
	All	**82.919*****	**98.706*****
Notes. All variables represent % changes. Signif. $p < 0.1$ = *; $p < 0.05$ = **; $p < 0.01$ = ***. Bold for signif. vars.			

This enabled us to combine machine learning and regularities analysis. The KRLS results in Tables 3 and 4. They show that uninformed traders did more negative feedback trading in response to social sentiment.

Table 3. KRLS on Daily Social Media Sentiments

Var.	Avg.	SE	*T*	$p > \|t\|$
Buy1	**-2.488**	1.075	-2.314	0.027
Buy2	2.465	0.769	3.205	0.003
Sell1	0.284	0.950	0.298	0.767
Sell2	**-2.156**	1.349	-1.599	0.119
Notes. 39 obs; dep. var.: *Marginal Effect of Normal Market Trade Volume*; $R^2 = 0.707$.				

Table 4. KRLS on Cumulative Sentiments

Var.	Avg.	SE	*T*	$p > \|t\|$
Buy1	0.0158	0.075	0.212	0.833
Buy2	**-0.277**	0.062	-4.447	0.000
Sell1	**-0.014**	0.108	-0.133	0.895
Sell2	0.199	0.137	1.457	0.154
Notes: 39 obs.; same dep. var.; $R^2 = 0.909$.				

Last, our analysis showed that uninformed traders did more negative feedback trading in response to social sentiment. They tended to buy stocks when negative sentiment increased and sold stocks when positive sentiment diminished. This behavior for mobile channel traders goes against positive feedback trading strategy, which was surprising for these data.

5. Preferences in Household TV Viewing

This study evaluates the extent to which household TV viewers exhibited concentrated preferences in the channels and genres they watch.

5.1. Context, Data and Variables

Chang et al. [17] collected data from a large digital entertainment firm, which broadcasts over 170 cable TV channels to several hundred thousand households in its operating territory. Three kinds of anonymized household data were obtained: (1) demographics and residence information; (2) bundle subscriptions for accessible channels; and (3) set-top box data that tracked TV channel viewing behavior. We used a random sampling approach with iteration for 10,000 cable TV-subscribing households employed for one month in 2011.[12] We eliminated observations with missing values or mismatches between the subscriptions and household viewing choices, leading to a sample of 4,720 households, and essentially all of their TV viewing data.

Our dependent variable in this analysis is *ViewingConcentration*[13] is a 0-1 proxy variable for the diversity of household-level viewing patterns in terms of viewing. It employs the *Gini coefficient* [60], which we adapted to gauge differences in household viewing times across the TV channels that the occupants of a household viewed.[14]

We also explored viewing patterns in terms of a number of independent main effects variables, as well as some controls. *#SubscribedChannels* captures how many channels the household subscribed to. Their TV viewing choices were constrained by their subscription decisions. The more channels that a household was able to view, the more likely was a household member able to find what they liked to watch. They also were able to watch many different channels, and seek variety in their viewing experience. With multiple household members, the sum of their viewing – an expression of *family-level unitary preferences* – resulted in varied viewing patterns. *ViewingTime* captures how long each household

associated with the stock buying behavior of mobile phone-based traders. There was less of an effect with stock selling behavior.

[12] Limited computer memory and run-time capacity makes it difficult for very large data sets to be processed and analyzed directly using statistics software [21].

[13] We considered the sigmoid function to perform the logistic regression estimation. For example, 0 ≤ *Viewing Concentration = f (Variables) = g (Marginal Effect * Variable)* ≤ 1 This is is a sigmoid function corresponding to *g*(*Marginal Effect* • *Variable*) = 1 / [1 + *e*^-(*Marginal Effect* • *Variable*). In this case, if the function *g* is greater than or equal to 0.5, the dependent variable will be estimated as 1 and 0 otherwise. The Gini coefficient was able to be obtained directly by computation from the data. The implication in the estimation model is that if the Gini coefficient is more than 0.5, then *Viewing Concentration* will be 1, and 0 otherwise.

[14] Ordinary least squares estimation usually imposes two conditions when proportional variables are estimated: (1) the conditional-expectation function must be non-linear since it maps onto a bounded interval; (2) and its variance must be heteroskedastic, since the variance will approach zero as the mean approaches either boundary point of 0 or 1 [40].

spent on TV viewing within the observation period. More time gave viewers more chances to watch different channels. *PreferenceClusters*, which we built using machine-based dummies from cluster analysis, to identify viewer content preferences.

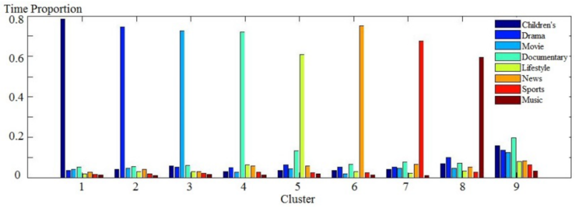
For the control variables, individual demographics did not support meaningful implications for household viewing patterns. So we only considered demographic variables that reflected household information. *SubscriberAge* is for those who actually subscribed to cable TV services. It provides useful information on family information. The control variable, *#Rooms* in a family's residence, offered a way to control for family income, since larger and wealthier families typically lived in larger residence in our study area during the observation period. All pairwise correlations between variables were less than 30%, so there were no problems for the estimation.

5.2. Machine-Based Methods

TV programs can be labeled in multiple genres based on their contents [22]. We clustered the channels into 8 well-known genres based on their main program contents and the channels. We aggregated household channel viewing time for each genre to obtain their time distributions.

We adopted two statistical indexes that are useful for evaluating the cluster quality results in terms of *k*: the *Davies-Bouldin index* and *silhouette values* [51]. We used 2 to 20 clusters, and ran the analysis for each value of *k* 100 times.[15] This generated average Davies-Bouldin index and silhouette values. The optimal silhouette value occurred for 7 clusters.[16] Fig. 1 shows the patterns.

Fig. 1. TV Viewership Cluster Centroid Shapes



Note: The *x*-axis represents clusters and the *y*-axis is *Average % Total ViewingTime.* The height of each bar is the average time proportion spends on each genre, for households in the cluster.

Each centroid shape was determined related to the average percentage time spent for each genre by households within the cluster relative to all the time spent over all of the clusters. The first 8 clusters show that households had strong preferences for one of the 8 genres. For example, households belonging to Cluster 2 on average spent over 70% of their viewing time on *Drama* channels, with only 30% spent on channels of other genres. The last cluster showed balanced preferences.[17] The results showed most households can be classified via patterns of viewing that suggested strong preferences for specific genres, with lesser preferences for the rest. But some households exhibited preferences for multiple genres. To capture this diversity of households, we identified 9 categories for *ClusterPreference* with a base case and 8 dummies.

5.3. Econometric Modeling Methods

We used a limited dependent variable model to estimate whether households exhibit concentrated viewing preferences, and whether clusters matter:

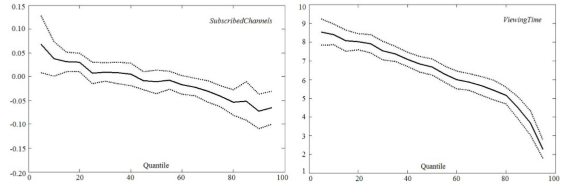
$$ViewingConcentration = f\,(\#SubscribedChannels,\ ViewingTime,\ PreferenceCluster,\ \#Rooms,\ SubscriberAge) + \xi$$

We conducted an *average value-based analysis* with logistic regression. But because we also considered other modeling issues that could have harmed our results, we conducted *quantile estimation* too [61]. Modeling quantiles of a dependent variable's probability distribution as a function of a given set of independent variables is equivalent to knowing the entire conditional distribution of the dependent variable at different quantiles.[18]

5.4. Explanatory Econometrics Results

Although our baseline analysis showed that *#SubscribedChannels* did not affect *ViewingConcentration*, the quantile regression revealed other details that are worthwhile to describe. (See Fig. 2.)

Fig. 2. Quantile Regression Results


(a) For *#SubscribedChannels* **(b) For *ViewingTime***

ViewingConcentration was affected by *#Subscribed-Channels*, but only for households with very high or very low levels of *ViewingConcentration* –

[15] Two limitations are known to affect how the *k-means algorithm* establishes clusters: (1) *k* must be predetermined as a parameter [50]; and (2) unstable clusters can result from randomly-selected initial centroids used to jump-start the analysis of the data [10].

[16] There was no number of clusters *k* that optimized both measures. We determined that 9 clusters were appropriate in our 8-genre setting: 8 clusters with preferences for each of the individual genres, and 1 cluster with mixed preferences for all of the genres.

[17] We labeled the clusters by theme: *Children*, *Drama*, *Movies*, *Documentary*, *Lifestyle*, *News*, *Sports*, *Music*, and *Mixed*.

[18] We estimated the main effects variables, *#SubscribedChannels* and *ViewingTime*, at 5 intervals of the dependent variable, *ViewingConcentraion*. Our methods included *beta distribution estimation* and *quasi-likelihood function-based estimation.*

and in opposite directions. This is like a cancelling effect that explains why there was no significant relationship in the average value-based regression results for the base case model. So use of quantile regression was justified. The effect of *ViewingTime* on *ViewingConcentration* decreased, from β = 8.524, SE = 0.354, $p < 0.01$ at the 5% quantile, to β = 2.259, SE = 0.250, $p < 0.01$ at the 95% quantile. This shows that, for our data, households with higher *ViewingConcentration* seemed to be less sensitive to having more *ViewingTime*: for them, an increase in *ViewingTime* was associated with a smaller increase in *ViewingConcentration* than for other households.

Regarding the *PreferenceCluster* variable, we noted that households in *Children* cluster had the highest *ViewingConcentration* (β = 0.070, SE = 0.027, $p < 0.05$), whereas households in the *Mixed* cluster had the lowest (β = -0.150, SE = 0.030, $p < 0.01$).

Our mobile phone-based stock-trading example did not use machine methods in as deep and rich as way as the household TV viewing concentration work has. Still, these are examples that applied the same philosophy for the new interdisciplinary fusion empirical science, with different levels of intensity.

6. Conclusion

We discussed computational social science research issues for big data and fusion analytics. The new policy analytics encompass multiple methods to improve analytical performance in the presence of big data and enhanced computing power. This supports research design innovation. This is a paradigm shift that enables study of a range of social scientific issues with unprecedented control and new insights.

In this context, to show how fusion analytics are different and how they go beyond big data analytics based on single methods, we reviewed background knowledge to showcase the shifts with the methods related to the use of big data. Also, to support our proposal for the new fusion analytics involving machine-based pattern recognition coupled with statistics and econometrics model-based explanation, we presented two major research projects that we have been involved with. They illustrate the application of the approach. We close out our analysis with a deeper discussion of the impacts of these methods, and how they can be implemented well.

References

[1] Aigner, W., Miksch, S., Muller, W., Schumann, H., Tominski, C. Visual methods for analyzing time-oriented data. *IEEE Trans. Vis. Comp. Graph.*, 14(1), 2008, 47–60.

[2] Amemiya, T. Generalized least squares theory. Ch. 6, *Advanced Econometrics*, Harvard, Boston, MA, 1985.

[3] Anderson, C. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, July 16, 2008.

[4] Aral, S., Walker, D. Identifying social influence in networks using randomized experiments, *IEEE Intell. Sys.*, 26(5), 2011, 91-96.

[5] Athey, S. Machine learning and causal inference for policy evaluation. In Proc. 21st ACM SIGKDD Intl. Conf. Knowl. Disc. Data Min., ACM Press, New York, NY, 2015, 5-6.

[6] Athey, S., Ellison, G. Position auctions with consumer search. *Qtrly J. Econ.*, 126(3), 2011, 1213-1270.

[7] Bardhan, I., Demirkan, D., Kannan, P.K., Kauffman, R.J., and Sougstad, R. an interdisciplinary perspective on IT service management and service science. *J. Mgmt. Info. Sys*., 26(4), 2010, 13-64.

[8] Bell, G., Hey, T., Szalay, A. Beyond the data deluge. *Science*, 423, 2009, 1297–1298.

[9] Benaroch, M., Dai, Q., Kauffman, R.J. Should we go our own way? Analyzing backsourcing flexibility in IT service outsourcing contracts. J. *Mgmt. Info. Sys*., 26(4), 2010, 317-358.

[10] Ben-David, S., Pál, D., Simon, H.U. Stability of k-means clustering. In N. Bshouty and C. Gentile (eds.), *Learning Theory*. Springer, Berlin, 2007, 20-34.

[11] Brabham, D.C. Crowdsourcing the public participation process for planning projects. *Plan. Theor.*, 8(3), 2009, 242–262.

[12] Brickland, T. *Introduction to the Policy Process: Theory, Concepts and Methods for Policy Making.* M.E. Sharpe, Armonk, NY, 2001,

[13] Bunn, D.W. Policy analytic implications for a theory of prediction and decision. *Pol. Sci.*, 8, 1977, 125-134.

[14] Canova, F., Ciccarelli, M. Panel vector autoregressive models: a survey. In T. Fomby, L. Killian, A. Murphy (eds.), *VAR Models in Macroeconomics.* Emerald Group Publishing, Bingley, UK, 2013, 205-246.

[15] Carley, K.M. Computational organization science: a new frontier. *PNAS*, 99(Supp 3), 2002, 7257-7262.

[16] Chang, R.M., Kauffman, R.J., Kwon, Y.O. *Dec. Supp. Sys.*, 63, 2014, 67-80.

[17] Chang, R.M., Kauffman, R.J., Son, I. Consumer micro-behavior and TV viewership patterns: data analytics for the two-way set-top box. In M. Bichler, R.J. Kauffman, H.C. Lau, C. Yang (eds.), *Proc. 14th Intl. Conf. Elec. Comm*., ACM Press, New York, NY, 2012, 272-273.

[18] Chaudhuri, S., Dayal, U., Narasayya, V. 2011. An overview of business intelligence technology. *Comm. ACM*, 54(8), 88-98.

[19] Chellappa, R.K., Sin, R.G., Siddarth, S. Price formats as a source of price dispersion: a study of online and offline prices in the domestic U.S. airline markets, *Info. Sys. Res*., 22(1), 2011, 83-98.

[20] Chen, H., Chiang, R., Storey, V.C. Business intelligence and analytics: from big data to impact. *MIS Qtrly*., 36(4), 2012, 1165-1188.

[21] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C. MAD skills: new analysis practices for big data. *Proc. VLDB*, 2(2), 2009, 1481-1492.

[22] Creeber, G., Miller, T., Tulloch, J. (eds.). *Television Genre Book.* British Film Inst., London, UK, 2001.

[23] Davenport, T.H. 2006. *Competing on Analytics*. *Harv. Bus. Rev.,* 84(1), 2006, 98-107.

[24] De Long, J.B., Shleifer, A., Summers, L., Waldmann, R. Noise traders risk in financial markets, *J. Pol. Econ*. 98, 1990, 703–738.

[25] Del Guidice, M., Della Peruta, M., Carayannis, E. *Social Media and Emerging Economies: Technology, Cultural and Economic Implic.* Springer, Berlin, 2015.

[26] Demirkan, H., Kauffman, R.J., Vayghan, J., Fill, H.G., Karagiannis, D., Maglio, P. Service-oriented technology and management: perspectives on research and practice for the coming decade. *Elec. Comm. Res. Appl*., 7(4), 2008, 356-376.

[27] Dietterich, T.G. Machine learning. In *Nature Encycl. Cogn. Sci*., Macmillan, London, UK, 2003.

[28] Elberse, A. Bye-bye bundles: the unbundling of music in digital channels. *J. Mktg*., 74(3), 2010, 107-123.

[29] Farris, P.W. *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance, 2nd Ed.* Pearson, Upper Saddle River, NJ, 2010.

[30] Geng, D., Kauffman, R.J. Decomposing the impact of credit card promotions on customer behavior and merchant performance. In *Proc. 50th Hawaii Intl. Conf. Sys. Sci.*, IEEE Comp. Soc. Press, Wash., DC, 2017.

[31] Ghose, A., Ipeirotis, P.G., Li, B. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content, *Mktg. Sci*., 31(3), 2012, 493-520.

[32] Gondecha, P., Lieu, H. Mining social media: a brief introduction. Chap. 1, in *Tut. in Ops. Res*., 2012.

[33] Granados, N.F., Gupta, A., Kauffman, R.J. Online and offline demand and price elasticities: evidence from air travel industry, *Inf. Sys. Res*., 23(1), 2012, 164-181.

[34] Hainmueller, J., Hazlett, C. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Pol. Anal.*, 22(2), 2013, 143-168.

[35] Harrysson, M., Metayer, E., Sarrazin, H. The strength of weak signals. *McKinsey Qtrly.,* February 2014.

[36] Hey, T., Tansley, S., Tolle, K. (eds.). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, WA, 2009.

[37] IBM. Analytics: The real-world use of big data. Institute for Business Value, New York, NY, 2012.

[38] IDC. Digital data to double every 18 months. Framingham, MA, May 2009.

[39] Imbens, G., Barrios, T., Diamond, R., Kolesar, M. Clustering, spatial correlations and randomization inference. Mimeo, Harvard Univ., Boston, MA. 2011.

[40] Kieschnick, R., McCullough, B.D. 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Stat. Model*., 3(3) 193-213.

[41] Kim, K. Lee, S.Y., Kauffman, R.J. How do traders react to social media sentiment in the mobile channel? *Intl. Symp. Smart Fin.*, Shenzhen, China, May 2016.

[42] Kmenta, J. Generalized linear regression model and its applications. In *Elements of Econometrics, 2nd ed.*, MacMillan, New York, NY, 1986.

[43] Lazer, D., Pentland, A.S., Adamic, L. Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G. Macy, M., Roy, D., van Alstyne, M. Life in the network: the coming age of computational social science. *Sci.*, 323(5915), 2009, 721-723.

[44] Lee, G., Qiu, L., Whinston, A. A friend like me: modeling network formation in a location-based social network. SSRN Paper 2769696, 2016.

[45] Li, Z., Kauffman, R.J., Dai, B. Can I see beyond what you can see? blending machine learning and econometrics to discover houselhold TV viewing preferences. In *Proc. 50th Hawaii Intl. Conf. Sys. Sci.*, IEEE Comp. Soc. Press, Washington, DC, 2017.

[46] Ma, D., Kauffman, R.J. Competition between software-as-a-service vendors. *IEEE Trans. Eng. Mgmt*., 61, 4, 2014, 717-729.

[47] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H. Big data: the next frontier for innovation, competition, and productivity, McKinsey Glob. Inst., New York, NY, May 2011.

[48] Mayer-Schönberger, V., Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin, New York, NY, 2013.

[49] McKinsey. Creating value through credit card partnerships in Latin America. New York, NY, 2011.

[50] Ray, S., Turi, R.H. Determination of number of clusters in *k*-means clustering and application in colour image segmentation. *Proc. 4th Intl. Conf. Adv. Patt. Recog. Dig. Tech.*, 1999, 137-143.

[51] Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.,* 20, 1987, 53-65.

[52] Rotwein, E. Empiricism and economic method: several views considered. *J. Econ. Iss.,* 7(3), 1973, 361-382.

[53] Save, A. Marketing in a mobile-first world: tackling the why and the how. *J. Dir. Data Dig. Mktg. Prac.,* 15(3), 2014, 202-212.

[54] Sekhon, J. Opiates for the matches: matching methods for causal inference. *Ann. Rev. Pol. Sci.*, 12, 2009, 487-508.

[55] Sen, A.G. Consumer insights and creativity. *IIMB Mgmt. Rev.*, 15(3), 2003, 124-126.

[56] Shi, Z., Lee, G.M., Whinston, A. B. Towards a better measure of business proximity: topic modeling for industry intelligence. *MIS Qtrly*., 2016, forthcoming.

[57] Wang, Y., Lewis, M., Cryder, C., Sprigg, J. Enduring effects of goal achievement and failure within customer loyalty programs: a large-scale field experiment*. Mktg. Sci.,* 2016, in press.

[58] Watson, H.J., Wixom, B.H. The current state of business intelligence. *IEEE Comp.,* 40(9), 2007, 96-99.

[59] Wellman, M.P. The economic approach to artificial intelligence. *ACM Comp. Surv. Symp. Artif. Intell.*, ACM Press, New York, NY, 1995.

[60] World Bank. Gini index. Washington, DC, 2013.

[61] Yu, K., Moyeed, R.A. Bayesian quantile regression. *Stat. Prob. Lett.*, 54(4), 2001, 437-447.