Journal of Computer Science and Information Technology
March 2014, Vol. 2, No. 1, pp. 01–20
ISSN: 2334-2366 (Print), 2334-2374 (Online)
Copyright © The Author(s). 2014. All Rights Reserved.
American Research Institute for Policy Development
42 Monticello Street, New York, NY 12701, USA.
Phone: 1.347.757.4901 Website: www.aripd.org/jcsit

Structuring Big Data: How Financial Models May Help

Thierry Warin[1] and William Sanger[2]

Abstract

The goal of this paper is to illustrate how financial models can be used to analyze unstructured data. Indeed, with the advent of social media, researchers have access to massive amounts of new data. These new data are unstructured in the sense that they can come from multiple origins. These data are commonly called "Big Data".To analyze Big Data considering the nature and the specificities of the data, we must develop a very reliable approach to extract the relevant information in realtime. Financial models are designed to manage risks. They are also designed to handle massive amounts of structured data to analyze risks. In this context, our paper could be very helpful to create a dashboard of potential risks based on social media scrutiny.

Keywords: Big Data, unstructured data, CAPM, financial models

Number of words: 4459

[1] Associate Professor, HEC Montréal, Department of International Business, 3000 Côte-Sainte-Catherine Road, Montréal (Québec) Canada H3T 2A7, E-mail: thierry.warin@hec.ca, phone: 514 340 6185

[2] Graduate Student, Polytechnique Montréal, Department of Mathematics and Industrial Engineering, C.P. 6079, succ. Centre-ville, Montréal (Québec) Canada H3C 3A7, E-mail: william.sanger@polymtl.ca

Introduction

The contribution of the literature on Big Data is twofold: it addresses either the questions about the applications (when to use Big Data, why do we use Big Data) or the questions about the methodology (how to use Big Data). Our paper is not about the former, i.e. how to use Big Data to predict,for instance,the future stock price of Google or the future volatility of Apple, but it is about the latter, i.e.how we can use financial models to extract some relevant information from Big Data.

We see two benefits to our proposition: the first one is to propose a robust set of models and ways to interpret these massive amounts of data;the second benefit is that a whole industry is already trained to understand and interpret the results of these models. Indeed, the whole financial industry has been trained to read and analyze outputs based on the financial models, and it would be interesting to capitalize on these existing skills.

More and more human beings are connected to each other or to companies through electronic devices. With the rise of this level of connectivity comes also the rise of huge amounts of personal information stored on servers. Any move, any call, any search, or any post on a website is saved. And it is only the beginning: if we store personal habits today, we will also store health information tomorrow. Saving on a server 80 years of someone’s life and doing that for seven billion people represent massive quantities of information. In theory, we know how to handle and analyze these massive quantities: data mining, statistics, or econometrics are well understood by researchers. The difficulty is to implement these techniques in the real world by designing algorithms that will allow us to measure,for instance,the covariances and also some forms of causation between the different variables.

Before going further, let us define “Big Data”."Big Data" is apparently a term coined by John Mashey, former Chief Scientist at Silicon Graphics Inc.[3](Diebold 2012). "Big Data" is often defined by using the three "V" acronym (Laney 2001): Volume, Velocity, and Variety. “Volume” refers to massive amounts of data that are available, "Velocity" refers to the speed required to process, analyze and use the data, and "Variety" refers both to unstructured data (videos, images, audiodata, texts, etc.) and to structured data (retail transactions, stock data, brain signals, genetic codes, etc.).

[3](“The Origins of ‘Big Data’: An Etymological Detective Story” 2013).

For the purpose of this paper, we propose our own definition of “Big Data” based on the previous literature. Big Data is a concept that encompasses four characteristics:

1) Big Data is about massive amounts of data. It thus requires massive storage capacities, high computing power, a lot of energy, etc.
2) Big Data is also about structured data: indeed, some of the data are clearly identifiable and easily accessible.
3) Big Data is also about unstructured data: here, some of the data are difficult to relate to another variable, and cannot be dealt with as easily as other structured data. In this context, we need to use new approaches to extract important information, and this extracted information will take the form of structured data.
4) Big Data means also being able to deal with real-time data. On top of being massive, structured and unstructured, the great technological improvement would be to have access to the analyses in real time. Machine learning deals with this issue and it is a challenge, notably when we talk about unstructured data.

We can cross-structured and unstructured data and make them interact.This characteristic is interesting in the sense that Big Data is not only about unstructured data, but also about the merger between structured and unstructured data. This merger provides us with information we could not have access to before.Moreover, with Big Data, we cannot rely only on conventional statistical analyses, such as factor analyses, principal component analyses, etc. We need to develop new tools to extract some additional relevant information. One of the possibilities, for instance, could be to use time series techniques to filter the data (Fourier, etc.). But, it would only give us one dimension, i.e. the evolution of the trend through time. In what follows, we want to add a second dimension: not only do we want to have the evolution through time, but we also want to know the pace of this evolution. The latter information will help us understand whether there is some momentum (or not) about a discussion on the social media. This momentum is also called the "buzz". Hence, what we propose is a framework that may help us discover a buzz right at its starting point.

Our paper is precisely concerned with this bridge between theory and its implementation. We propose to develop a frameworkinspired by a discipline that already deals with massive, real-time, structured data: finance. Indeed in finance, researchers have used theoretical models to filter, analyze and visualize financial data. Analyzing financial data means structuring the data, computing new variables and indicators, and relating all these variables together.

In many regards, this is the process we need to follow to analyze massive structured and unstructured data.

The originality of our paper comes from the fact that we do not use Big Data to add information to financial analysis, for instance, but we do the opposite: we use financial models to extract information from Big Data.

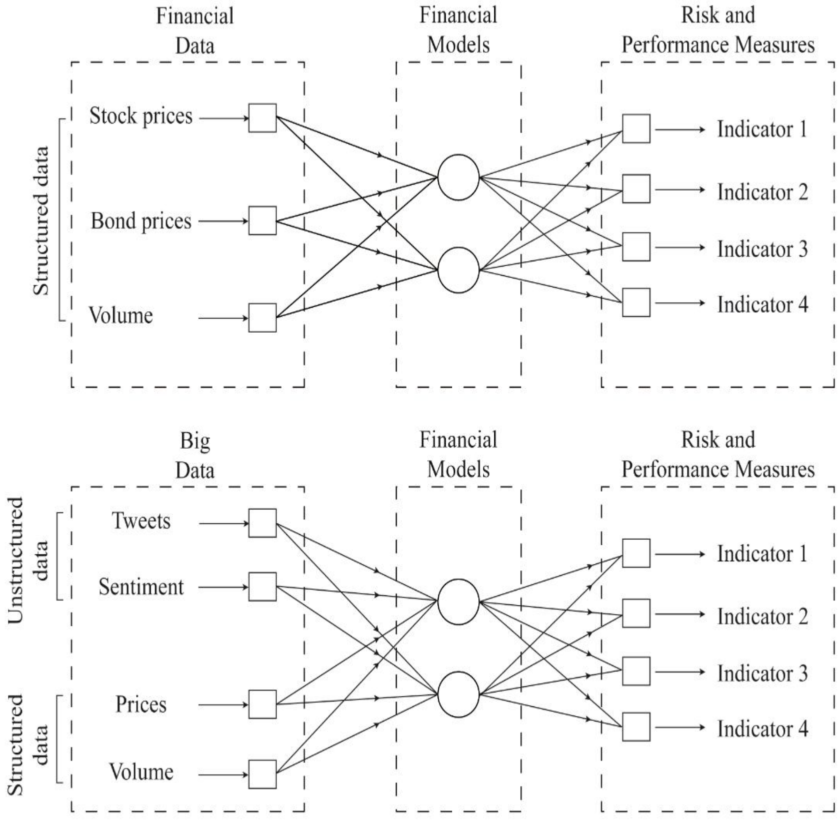


Figure 1: Process for analyzing massive structured and unstructured data.

Financial Models to Deal with Big Data

Finance is not a newdiscipline, it has existed since the invention of money. However, modern finance is a very new field. We could debate on the origins of modern finance, butwe will propose the initial work from Markowitz as the cornerstone of modern finance (Markowitz 1952). In 1952, Markowitz proposed to look at assets not only individually but mostly as constituents of a bigger item, namely a portfolio. Modern Portfolio Theory (MPT) was born.

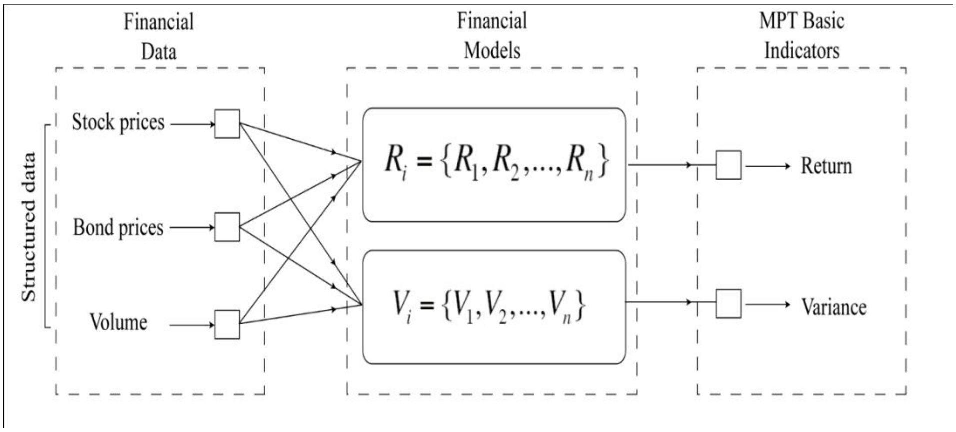


Figure 2: Process to analyze a financial asset in order to extract the Modern Portfolio Theory indicators

Sharpe, Lintner and Mossin, in particular, will add their contributions to Markowitz' model by creating what will be known as the Capital Asset Pricing Model (CAPM) (Sharpe 1963; Lintner 1965; Mossin 1966). Sharpe will also propose to bridge the theory of portfolio management with the real world. Indeed, MPT was a very convincing mathematical approach, while relying on strong assumptions about the availability of market information. Sharpe will bridge the gap between these assumptions and the real world by using techniques to measure these variables (Sharpe 1963).

With the new advances in MPT, and in particular with the CAPM approach, finance literature has refined the mathematical models. Assets are characterized by their mean and variance to measure their return and their risk level.

The mean-variance framework will be applied to portfolios: each portfolio will have a mean return and a variance that will represent its level of risk.

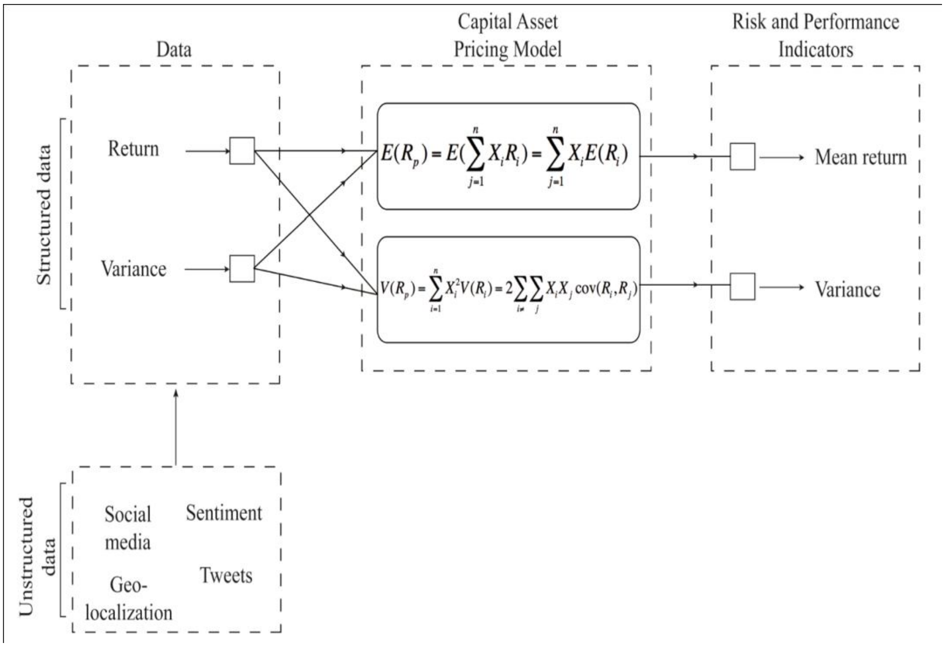

Data
Capital Asset Pricing Model
Risk and Performance Indicators
Structured data
Return
Variance
$E(R_p) = E(\sum_{j=1}^{n} X_i R_i) = \sum_{j=1}^{n} X_i E(R_i)$
$V(R_p) = \sum_{i=1}^{n} X_i^2 V(R_i) = 2\sum_{i}\sum_{j} X_i X_j \operatorname{cov}(R_i, R_j)$
Mean return
Variance
Unstructured data
Social media
Sentiment
Geo-localization
Tweets

Figure 3: Process using the CAPM approach for financial assets and its adaptation for Big Data

We can therefore compare the mean-variance of selected portfolios to the theoretical frontier capturing the ideal portfolios. In finance, the introduction of a risk-free asset will also help us decide which portfolio would be ideal.

The mean-variance framework is – we believe – interesting for the analysis of Big Data. Indeed, it poses the question of which portfolio to choose in a two-dimensional way: (1) considering the mean return and (2) considering the variance (the risk). For the same level of return, we can decide to choose the portfolio having a lower risk or, for the same level of risk, we can decide to choose the portfolio having a higher return. This means that we can infer eight areas from the mean-variance framework: four areas with a positive value for the mean return and four areas with a negative value for the mean return.

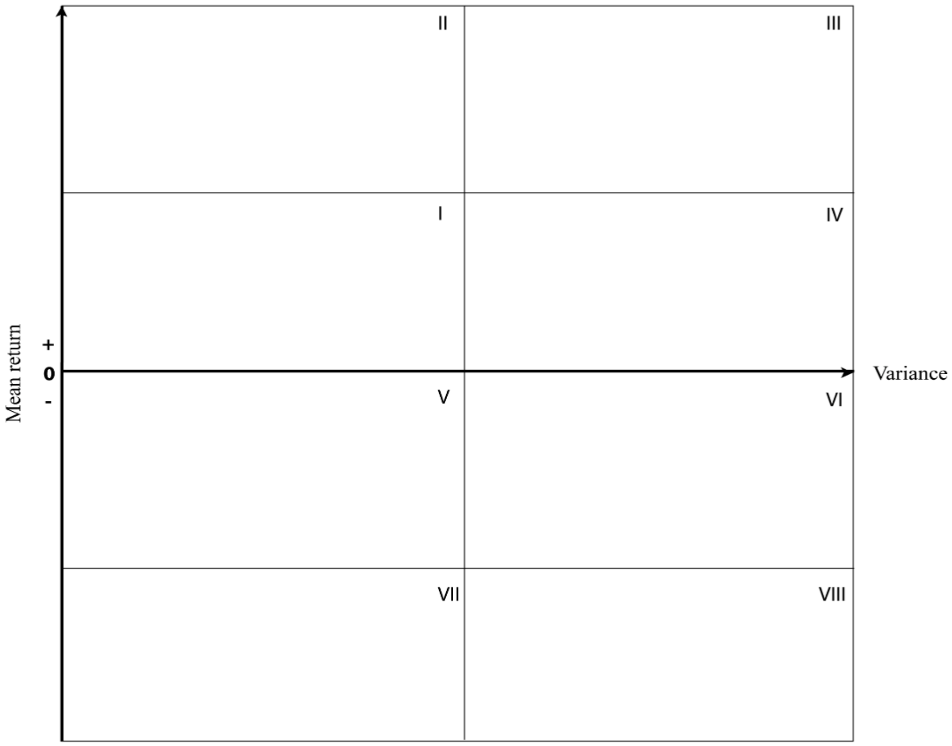


Figure 4: Eight areas of the mean-variance framework applied to Big Data

The eightareas may have the following significance:

1) Area I: it represents a low mean return and a low variance. This means that people have not tweeted much more at time *t* compared to time *t-1*. It also means that the rate of growth of the tweets is almost constant through time.
2) Area II: it represents a high mean return and a low variance. Here, people have a high growth rate of tweets on a daily basis, associated with a constant rate of growth, explaining the low variance.
3) Area III: it represents a high return and a high variance, meaning that people write more tweets on a daily basis, but with a high volatility from one day to another.
4) Area IV: here, it is a low mean return and a high variance. It represents a small increase in the number of tweets on a daily basis, but then with high volatility.
5) Area V: it represents a low negative mean return and a low variance, showing a steady decrease in the number of tweets.

6) Area VI: it represents a low negative mean return with a high variance, showing some volatility on the number of tweets.
7) Area VII: it represents a high negative return with a low variance.
8) Area VIII: it represents a high negative return and a high variance.

Of course, in the longrun, we can easily assume that the positive areas the most likely. To apply this framework to Big Data analysis, we need to define the significance of mean return and variance in the context of real-time massive structured and unstructured data.Let us describe only the positive return areas, since the graph is symmetrical for the negative return areas. Area I is a safe area in the sense that people are talking more and more about a company, for instance, but with a steady rate of growth. In a nutshell, there is no excitement over something positive or something negative. Area II shows some more tweets but the same rate of growth. It could capture some homogeneous excitement of the consumers. It can be positive or negative for the company. Area III differentiates itself by having more volatility.Here the discussion about the company (or about the issue at stake) gains some momentum one day and then decreases a lot the next day. It can also be positive or negative, and it can be capturing the beginning or the end of a discussion cycle. Area IV captures a higher variance and a low mean return. This is similar to area III, except that the momentum is either gaining some power or decreasing.

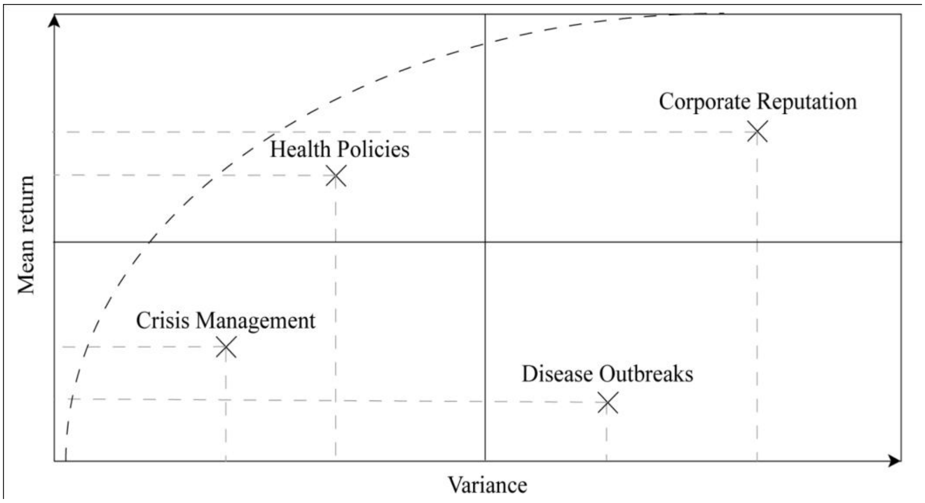


Figure 5: Example of the different applications of the mean-variance framework for Big Data

We can easily apply this framework to any kind of issue that is debated on social media. There are two dimensions to consider in the interpretation of the eight aforementioned areas: (1) the location within an area and its meaning, and (2) the changes that may happen through time across these areas. About the latter, for instance, a company can be represented in Area IV and then move to Area II, making us conclude that the name of this company is gaining some momentum among the social networks. For the company's executives, it would be interesting to know what is at the root of this buzz. In what follows, we will provide a case study to illustrate how this approach could be interesting for a government.

Case study: Delivering new information for better public policies in the province of Quebec (Canada)

Information from Big Data is useful in many different fields: some have used Big Data to predict electoral outcomes (Tumasjan et al. 2010), some have used Big Data to detect influenza outbreaks (Culotta 2010), some have used Big Data to analyze social activity between communities (Crampton et al. 2013), and some others have used Big Datato add some new information to the financial markets(Clarkson, Joyce, and Tutticci 2006; Sprenger and Welpe 2010).

In what follows, we propose a theoretical approach to address the four characteristics aforementioned (massive amounts of information, real-time data, merger of structured and unstructured data) inspired by the mean-variance framework commonly used in finance.

The interest of using an approach inspired by the financial models is at least twofold: (1) financial models are well designed to analyze also massive amounts of real-time data, and (2) there are lots of tools used in the financial industry that can be adapted to analyze Big Data.Financial models are indeed very useful and interesting when it comes to illustrate the complex information coming from Big Data.

For illustration purposes, we will use data coming from Twitter. We will analyze the relationship between tweets about a specific risk category and the level of risk perceived by the population. We will rely on a study released in 2012 in Quebec, in which nine different categories of risk were defined, such as economic and financial risks, public project management risks or health risks.

Then we will use our approach to show how we can give a score to any category of risk and thus how we can be made aware of potential risks in the society.

The mean-variance framework in finance is an interesting tool since it provides important information about a portfolio: its return, its risk and its distance from the ideal portfolio. We could use the same kind of characterization for the information we get from Big Data feeds. Indeed, it would be useful to knowthe magnitude of the searches of a specific word, and how volatile the searches are through time. Characterizing, for instance, a word in these two dimensions may help provide some interesting information about people's concerns. In what follows, we want to use our mean-variance framework to measure the level of risks.

Context

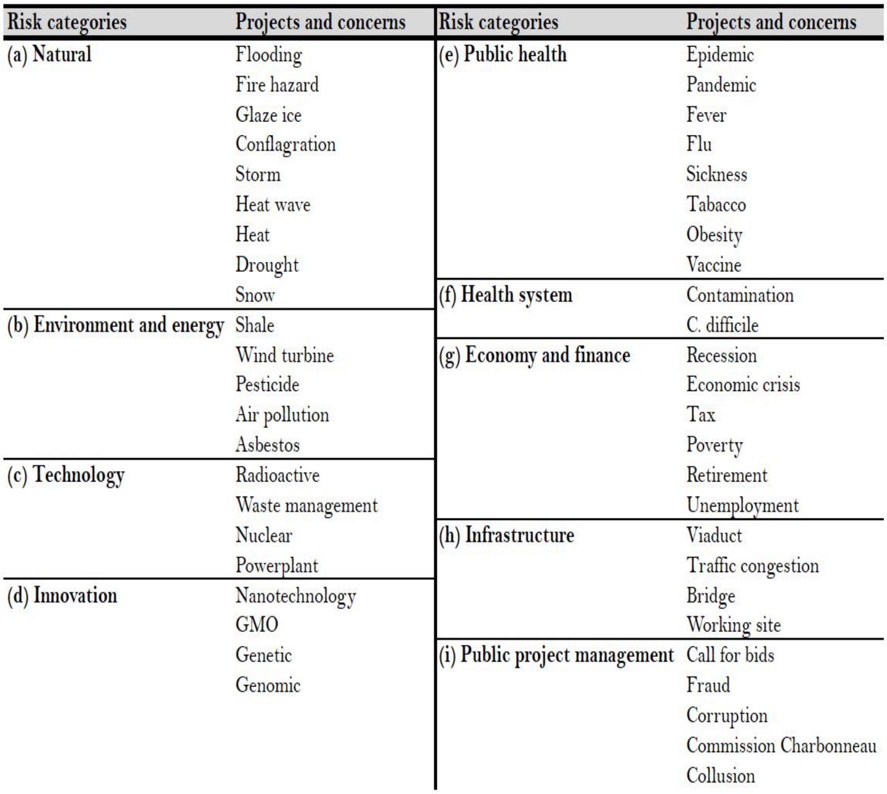
Marcellis-Warin and Peignier (2012) presented the results of a study regarding the risk perception of the Quebec population. In their analysis, *Perception des risques au Québec -Baromètre CIRANO 2012*, they highlighted nine categories of risk considered as the most worrisome from the population's point of view: risks related to **(a)** nature,**(b)** environment and energy, **(c)** technology, **(d)**innovation, **(e)** public health, **(f)** health system, **(g)** economy and finance, **(h)** infrastructure and **(i)** public project management. They listed the related projects and concerns referring to these ninecategories of risk. As a starting point, we will use the nine categories identified in this study and see whether the perceptions translate into tweets in the social media sphere.

With more than 500 million tweets posted daily and a worldwide presence, this social network represents a valuable source of information about the public's concerns. Twitter is an open social media, which enables one to access informationpublished by the other users. The medium of expression is called a *tweet*, a 140-character long message. Its typology has led to an easy way of classification: in order to address a message to someone in particular, the name of the entity has to be preceded by an@ symbol (i.e. @GouvOuvertQc); messages can be classified by topic under the symbol #, called *hashtag* (i.e. #datamining). Finally, a geocoded feature can be activated from the sender's account, which gives a location value for each message.

Twitter has a query interfacethat we can use to access messages published under a specific topic, by a particular person, or emitted from a given area. However, one can only access data published the week following the request. Older information is not available through Twitter directly. In order to access older data, companies specialized in monitoring the social web could be used. Among them, PeopleBrowsr(www.peoplebrowsr.com) is a key option for quantitative analysis. The website provides freely accessible data for a time frame up to two years. It has been used in a variety of publications concerning the use and leverage of Big Data, including emergency situations (Aljohani, Alahmari, and Aseere 2011), social web analysis (McKelvey and Menczer 2014), corporate reputation (Warin et al. 2013), activism monitoring for the 2011 Egyptian revolution (Stover Tillinghast et al. 2012) or for the Occupy Wall Street movement (Tremayne 2013).

Based upon the categories provided by de Marcellis-Warin and Peigner, we established a list of terms that specify each risk, detailed in Table 1. Each of these terms is searched through the Grid option of PeopleBrowsr. The format extracted is the number of times a specific term is written on Twitter per day. In order to select only the tweets emitted from the province of Quebec, we first selected the messages coming from Canada, and then entered the terms of Table 1 in French.[4] We inferred that the combination of Canadian-located tweets and French-writtentweets would in fact be the tweets coming from the province of Quebec.

[4] The French translation of eachtermusedis: flooding / inondation; firehazard / feu; glazeice / verglas; conflagration / incendie; storm / tempête; heatwave / canicule; heat / chaleur; drought / sécheresse; snow / neige; shale / schiste; wind turbine / éolienne, éolien; pesticide / pesticides; air pollution / pollution de l'air; asbestos / amiante; radioactive / radioactif, radioactifs; waste management / déchets; nuclear / nucléaire; powerplant / centrale; nanotechnology / nanotechnologies; GMO / OGM; genetic / génétique; genomic / génomique; epidemic / épidémie; pandemic / pandémie; fever / fièvre; flu / grippe; sickness / maladie; tobacco / tabac; obesity / obésité; vaccine / vaccins; contamination / contamination; C. difficile / C. difficile; recession / récession; economiccrisis / crise économique; tax / impôt; poverty / pauvreté; retirement / retraite; unemployment / chômage; viaduct / viaduc; traffic congestion / embouteillage; bridge / pont; working site / travaux; call for bids / appel d'offres; fraud / fraude; corruption / corruption; collusion / collusion; Commision Charbonneau / Commision Charbonneau.



Risk categories	Projects and concerns	Risk categories	Projects and concerns
(a) **Natural**	Flooding	(e) **Public health**	Epidemic
	Fire hazard		Pandemic
	Glaze ice		Fever
	Conflagration		Flu
	Storm		Sickness
	Heat wave		Tabacco
	Heat		Obesity
	Drought		Vaccine
	Snow	(f) **Health system**	Contamination
(b) **Environment and energy**	Shale		C. difficile
	Wind turbine	(g) **Economy and finance**	Recession
	Pesticide		Economic crisis
	Air pollution		Tax
	Asbestos		Poverty
(c) **Technology**	Radioactive		Retirement
	Waste management		Unemployment
	Nuclear	(h) **Infrastructure**	Viaduct
	Powerplant		Traffic congestion
(d) **Innovation**	Nanotechnology		Bridge
	GMO		Working site
	Genetic	(i) **Public project management**	Call for bids
	Genomic		Fraud
			Corruption
			Commission Charbonneau
			Collusion

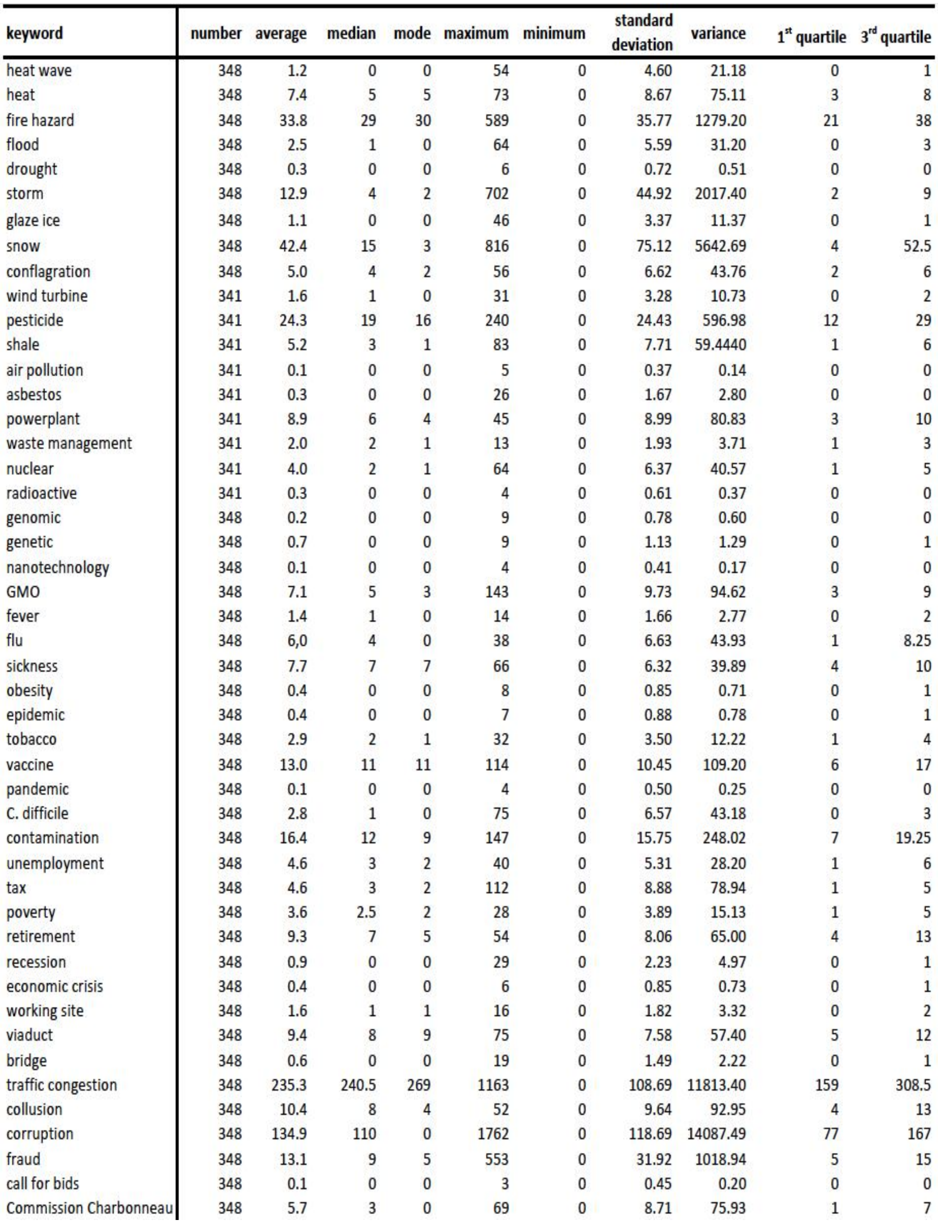
Table 1: Risk categories and their related projects or concerns [translated from French]. Source: de Marcellis-Warin and Peignier 2012

Data

In order to reflect the evolution of the population's perception of risk, we collected data from September 1st, 2012 to August 31st, 2013 for each 47 keywords describing nine different kinds of risk. Some of the data were not accessible without paid membership, so we discarded these entries from our database (written "Available on Request" on the website).

In Table 2, we provide some descriptive statistics for each keyword. For every keyword, PeopleBrowsr provides a list of the most used expressions in the tweets, including the selected keywords. This was used to verify the validity of our keywords in regards of their background.

Moreover, for two keywords, we searched different declinations: for *wind turbine*, we selected “éolien” and “éolienne” and for *radioactive*, we selected “radioactif” and “radioactifs”. We combined the two datasets in order to obtain the total value for each of the two keywords.



keyword	number	average	median	mode	maximum	minimum	standard deviation	variance	1st quartile	3rd quartile
heat wave	348	1.2	0	0	54	0	4.60	21.18	0	1
heat	348	7.4	5	5	73	0	8.67	75.11	3	8
fire hazard	348	33.8	29	30	589	0	35.77	1279.20	21	38
flood	348	2.5	1	0	64	0	5.59	31.20	0	3
drought	348	0.3	0	0	6	0	0.72	0.51	0	0
storm	348	12.9	4	2	702	0	44.92	2017.40	2	9
glaze ice	348	1.1	0	0	46	0	3.37	11.37	0	1
snow	348	42.4	15	3	816	0	75.12	5642.69	4	52.5
conflagration	348	5.0	4	2	56	0	6.62	43.76	2	6
wind turbine	341	1.6	1	0	31	0	3.28	10.73	0	2
pesticide	341	24.3	19	16	240	0	24.43	596.98	12	29
shale	341	5.2	3	1	83	0	7.71	59.4440	1	6
air pollution	341	0.1	0	0	5	0	0.37	0.14	0	0
asbestos	341	0.3	0	0	26	0	1.67	2.80	0	0
powerplant	341	8.9	6	4	45	0	8.99	80.83	3	10
waste management	341	2.0	2	1	13	0	1.93	3.71	1	3
nuclear	341	4.0	2	1	64	0	6.37	40.57	1	5
radioactive	341	0.3	0	0	4	0	0.61	0.37	0	0
genomic	348	0.2	0	0	9	0	0.78	0.60	0	0
genetic	348	0.7	0	0	9	0	1.13	1.29	0	1
nanotechnology	348	0.1	0	0	4	0	0.41	0.17	0	0
GMO	348	7.1	5	3	143	0	9.73	94.62	3	9
fever	348	1.4	1	0	14	0	1.66	2.77	0	2
flu	348	6,0	4	0	38	0	6.63	43.93	1	8.25
sickness	348	7.7	7	7	66	0	6.32	39.89	4	10
obesity	348	0.4	0	0	8	0	0.85	0.71	0	1
epidemic	348	0.4	0	0	7	0	0.88	0.78	0	1
tobacco	348	2.9	2	1	32	0	3.50	12.22	1	4
vaccine	348	13.0	11	11	114	0	10.45	109.20	6	17
pandemic	348	0.1	0	0	4	0	0.50	0.25	0	0
C. difficile	348	2.8	1	0	75	0	6.57	43.18	0	3
contamination	348	16.4	12	9	147	0	15.75	248.02	7	19.25
unemployment	348	4.6	3	2	40	0	5.31	28.20	1	6
tax	348	4.6	3	2	112	0	8.88	78.94	1	5
poverty	348	3.6	2.5	2	28	0	3.89	15.13	1	5
retirement	348	9.3	7	5	54	0	8.06	65.00	4	13
recession	348	0.9	0	0	29	0	2.23	4.97	0	1
economic crisis	348	0.4	0	0	6	0	0.85	0.73	0	1
working site	348	1.6	1	1	16	0	1.82	3.32	0	2
viaduct	348	9.4	8	9	75	0	7.58	57.40	5	12
bridge	348	0.6	0	0	19	0	1.49	2.22	0	1
traffic congestion	348	235.3	240.5	269	1163	0	108.69	11813.40	159	308.5
collusion	348	10.4	8	4	52	0	9.64	92.95	4	13
corruption	348	134.9	110	0	1762	0	118.69	14087.49	77	167
fraud	348	13.1	9	5	553	0	31.92	1018.94	5	15
call for bids	348	0.1	0	0	3	0	0.45	0.20	0	0
Commission Charbonneau	348	5.7	3	0	69	0	8.71	75.93	1	7

Table 2: Descriptive statistics of each keyword.

Results

With this dataset of unstructured data obtained through Twitter, we applied the mean-variance framework in order to interpret the perception of risk of the population. Our first modeling consisted in observing the variance and the mean return of each category of risk. The public project management category appears to have a stronger volatility compared to the other categories during this period of time. The components of this category include terms related to fraud, collusion and the Commission Charbonneau, which were widely spread in the media during this period.

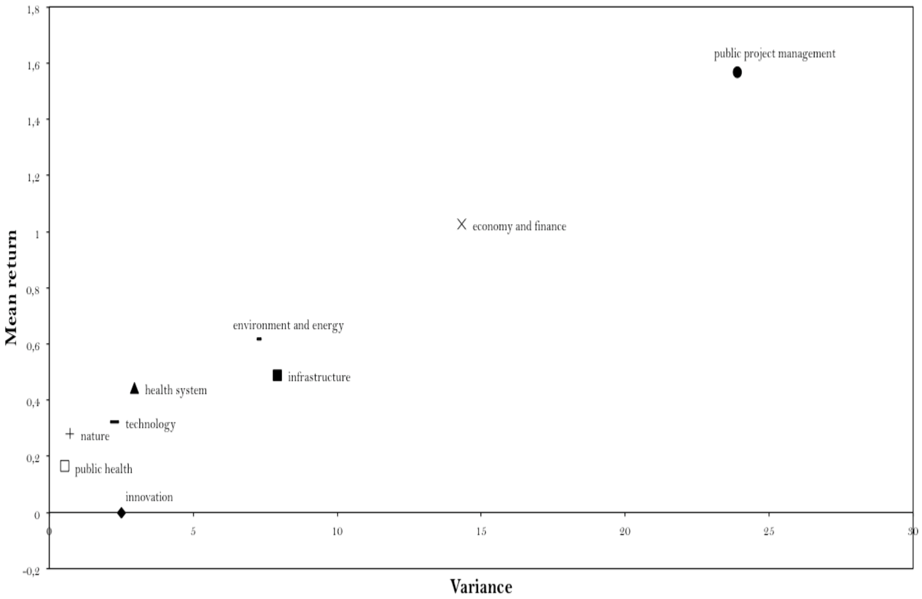
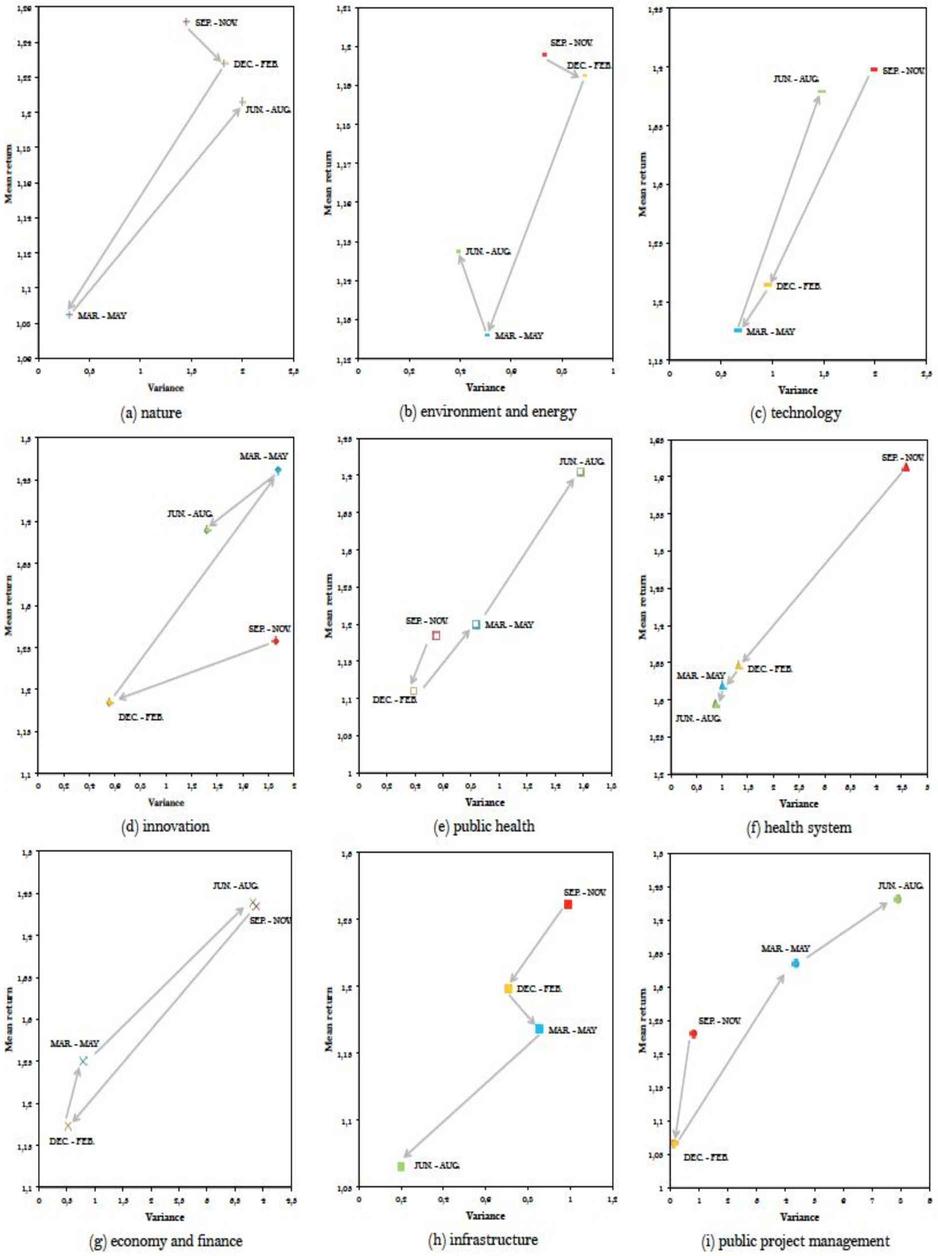


Figure 6: Mean-variance framework applied to the nine categories of risk during the entire period of observation

Our second approach was to observe the change in volatility and the rate at which tweets were emitted through time. We decided to divide our time frame into four periods, (1) from September 1st, 2012 to November 30th,2012; (2)from December 1st, 2012 to February 28th, 2013; (3) from March 1st, 2013 to May 31st, 2013; and (4) from June 1st, 2013 to August 31st, 2013.

These visualizations provide an interesting insight into the perception of certain risks. For example, risks related to nature widely moved through time from December 2012 to August 2013, decreasing during the third period of observation (March 2013 – May 2013). The keywords describing risks related to nature included *snow*, *heat wave*, *heat*, *fire hazard*, *flood*, *drought*, *storm*, *glaze ice*, *conflagration*, words that are highly seasonal. Regarding health system risks, a constant decrease was observed after February, which is related to the outbreak of *C. difficile* cases in hospitals that occurred previously in late 2012. Finally, regarding public project management risks, we observeda shift towards a higher volatility for this category from September 2012 to the end of May 2013. This is explained by the resignations of Montreal's mayor and Laval's mayor following the revelations of the Commission Charbonneau.



(a) nature

(b) environment and energy

(c) technology

(d) innovation

(e) public health

(f) health system

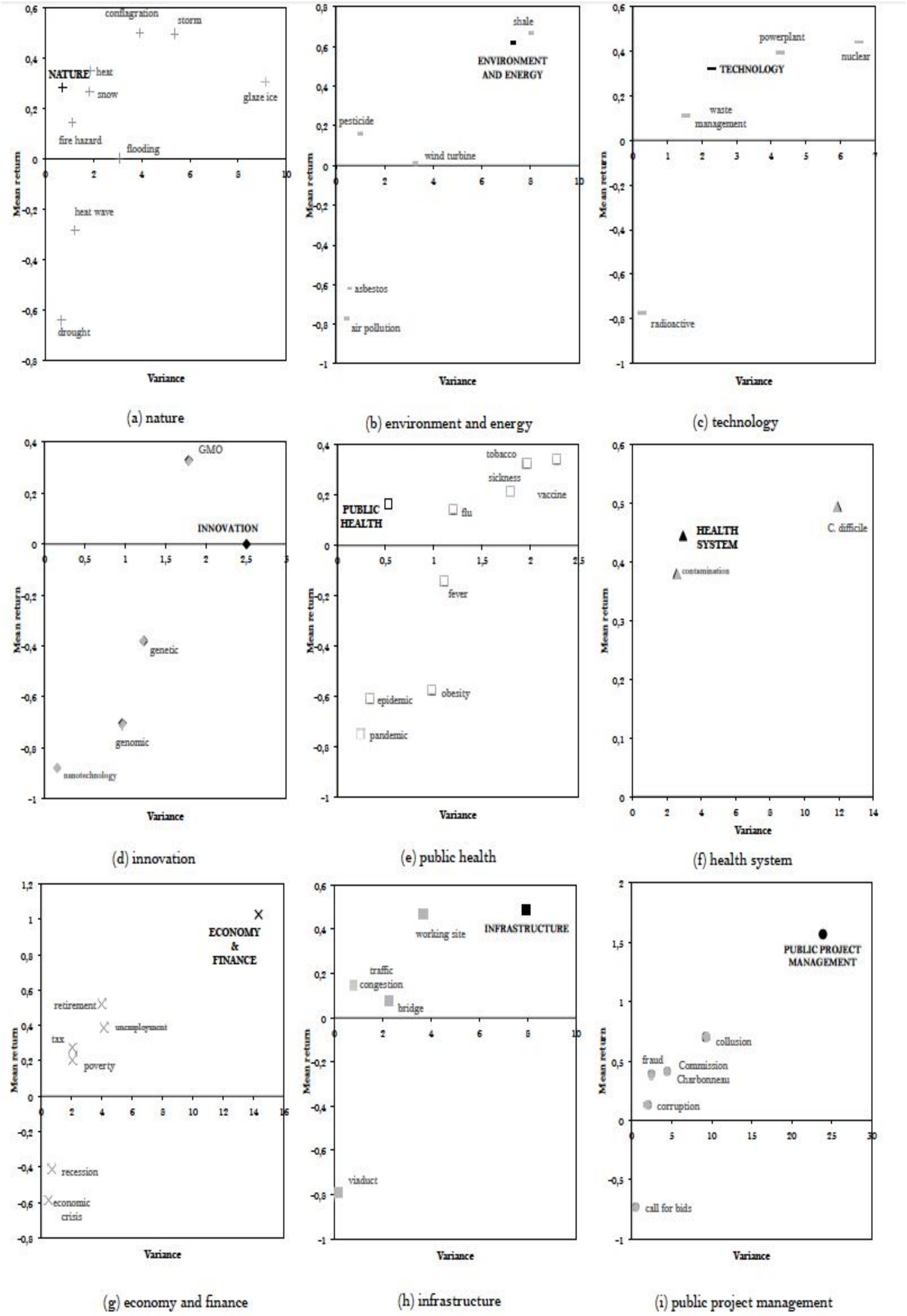
(g) economy and finance

(h) infrastructure

(i) public project management

Figure 7: Evolution of the perception of risk through time for each category.

Our final interpretation of the mean-variance framework for unstructured data led us to represent each risk category compared to each of its components in Figure 8.



(a) nature

(b) environment and energy

(c) technology

(d) innovation

(e) public health

(f) health system

(g) economy and finance

(h) infrastructure

(i) public project management

Figure 8: Details of the risk category's components.

Conclusion

In this paper, we wanted to bridge finance and Big Data in an unusual way. Indeed, it is often understood that we can extract some relevant information from Big Data to help financial decisions and to have an edge on the market. This is not what we wanted to cover in this paper. We wanted to propose a methodology inspired by financial mathematical models in order to extract some relevant information from Big Data that can be useful for almost any topic of interest, including finance.

The similarities between finance and Big Data are important: both deal with massive amounts of data, real-time data and structured data (Big Data adds here a difference with some unstructured data). Modern finance has developed since the 1960s. The financial models require a lot of computing power, and finance has developed also with the greater access to new technologies. All the mathematical financial models have been coded and are installed on all computers in the financial industry. So, the routines and libraries exist. We do not need to take all these codes, but at the same time, why would we not benefit from this existing situation? Indeed, the mathematical financial models are well understood and could be useful even for data whose nature is not financial.

In our case study, we have provided a few examples on how we could interpret the results coming out of the mathematical financial models applied to Big Data. Our study is not exhaustive and is just illustrative. For instance, we wanted to complement studies using semantics (sentiment analysis) by developing a set of indicators that could trigger an alarm. We hope other authors will find our initiative interesting and apply other approaches in finance or other disciplines to develop a very useful toolbox to analyze Big Data.

References

Aljohani, Naif, Saad Alahmari, and Ali Aseere. 2011. "An Organized Collaborative Work Using Twitter in Flood Disaster".*ACM Web Science*. http://www.websci11.org/fileadmin/websci/Posters/172_paper.pdf.

Clarkson, Peter, Daniel Joyce, and Irene Tutticci. 2006. "Market Reaction to Takeover Rumour in Internet Discussion Sites". SSRN Scholarly Paper ID 889785. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=889785.

Crampton, Jeremy W., Mark Graham, Atius Poorthius, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. 2013. "Beyond the Geotag Situating Big Data and Leveraging the Potential of the Geoweb". Accessed April 7. http://www.academia.edu/2986482/Beyond_the_Geotag_Situating_big_data_and_leveraging_the_potential_of_the_geoweb.

Culotta, Aron. 2010. "Towards detecting influenza epidemics by analyzing Twitter messages". In *Proceedings of the First Workshop on Social Media Analytics*, pp.115–122. SOMA '10. New York, NY, USA: ACM. doi:10.1145/1964858.1964874. http://doi.acm.org/10.1145/1964858.1964874.

Diebold, Francis X. 2012. "A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version". PIER Working Paper Archive 13-003. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania. http://ideas.repec.org/p/pen/papers/13-003.html.

Laney, Douglas. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety". META Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lintner, John. 1965. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets".*The Review of Economics and Statistics,* Vol. 47 (1) (February), pp. 13–37. doi:10.2307/1924119.

Marcellis-Warin, Nathalie de, and Ingrid Peignier. 2012. "Perception des risques au Québec – Baromètre CIRANO 2012". CIRANO Monographs. CIRANO. http://ideas.repec.org/b/cir/cirmon/2012mo-02.html.

Markowitz, Harry. 1952. "Portfolio Selection".*The Journal of Finance,* Vol. 7 (1) (March 1), pp. 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x.

McKelvey, Karissa, and Filippo Menczer. 2013. "Design and Prototyping of a Social Media Observatory". In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pp. 1351–1358. WWW '13 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. http://dl.acm.org/citation.cfm?id=2487788.2488174.

Mossin, Jan. 1966. "Equilibrium in a Capital Asset Market".*Econometrica*, Vol. 34 (4) (October), pp. 768–783. doi:10.2307/1910098.

Sharpe, William F. 1963. "A Simplified Model for Portfolio Analysis".*Management Science*, Vol. 9 (2) (January 1), pp. 277–293.

Sprenger, Timm, and Isabell Welpe. 2010. "Tweets and Trades: The Information Content of Stock Microblogs". SSRN Scholarly Paper ID 1702854. Rochester, NY: Social

Science Research Network. http://papers.ssrn.com/abstract=1702854.

Stover Tillinghast, Diana, Dannah Sanchez, Matthew Gerring, and Sarah Hassan. 2012. “Egyptian Demonstrators Use of Twitter: Tactics, Mobilization, and Safety”. International Conference on Communication, Media, Technology and Design,Istanbul, Turkey. http://www.cmdconf.net/2012/makale/32.pdf.

“The Origins of ‘Big Data’: An Etymological Detective Story”. 2013. *Bits Blog*. Accessed October 22. http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

Tremayne, Mark. 2014. “Anatomy of Protest in the Digital Era: A Network Analysis of *Twitter* and Occupy Wall Street”.*Social Movement Studies,*Vol. 13 (1), pp. 1–17. doi:10.1080/14742837.2013.830969.

Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”.In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852.

Warin, Thierry, Nathalie de Marcellis-Warin, William Sanger, Bertrand Nembot, and Venus Hosseinali Mirza. 2013. “Corporate Reputation and Social Media: A Game Theory Approach”. CIRANO Working Paper 2013s-18. CIRANO. http://ideas.repec.org/p/cir/cirwor/2013s-18.html.