UNIVERSITEIT VAN AMSTERDAM

Graduate School of Humanities

Faculty of Humanities, University of Amsterdam

The Making of Predictions: Social Media-Based Prediction and Its Resources, Techniques, and Applications

Submitted to the Department of Media Studies at the University of Amsterdam, Faculty of Humanities, in partial fulfillment of the requirements for the degree of Master of Arts (M.A.).

F. N. (Fernando) van der Vlist B.Des. B.A.

<*fernando.vandervlist@{student.uva.nl, gmail.com}*>

UvA ID: 10440267

26 June 2015

Study Programme: Media Studies (Research)

CROHO-Code: 60832

Course: Research Master Thesis Media Studies

Course ID: HMED/159414040Y

Supervisor: Dr. B. (Bernhard) Rieder, <*B.Rieder@uva.nl*>

Second Reader: Dr. C. (Carolin) Gerlitz, <*C.Gerlitz@uva.nl*>

Third Reader: Dr. N. A. J. M. (Niels) van Doorn, <*N.A.J.M.vanDoorn@uva.nl*>

The Making of Predictions

Social Media-Based Prediction and Its Resources, Techniques, and Applications

Fernando Nathaniël van der Vlist

Submitted to the Department of Media Studies at the University of Amsterdam, Faculty of Humanities, in partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

* * *

University of Amsterdam
Graduate School of Humanities

June 2015

Submitted to the Department of Media Studies at the University of Amsterdam, Faculty of Humanities, in partial fulfillment of the requirements for the degree of Master of Arts (M.A.).

For information, please email graduateschool-fgw@uva.nl or write to Graduate School of Humanities, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands. For information on this title, please visit dare.uva.nl. For information on the author, please visit fernandovandervlist.nl.

This document is set in Fresco Pro (at 11/16.5 pt. font size/leading in normal weight for body text), designed by Fred Smeijers, and published by OurType, De Pinte, Belgium. Documentation and formatting have been adapted from the MLA style manual (seventh edition), published by the Modern Language Association of America, New York, USA.

Tell me with whom you consort and I will tell you who you are;
if I know how you spend your time, then I know what might become of you.

—Johann Wolfgang von Goethe (1749–1832)

Acknowledgements

The epigraph to this thesis is not just fitting with respect to the subject matter, but can also be interpreted to mean that any achievement is usually indebted to some degree to the people that have been involved. A brief word of acknowledgement to those who have been involved in this project therefore seems appropriate. I am deeply grateful to all who made this project possible and contributed in different ways. I would like to thank my supervisor Bernhard Rieder in particular for his inspiring enthusiasm and confidence in the successful completion of this thesis; for his sensitivity and careful attention to the development of my ideas and writing throughout the process; as well as for taking me under his wing as his assistant this academic year. I have learned a lot and always left our meetings feeling rejuvenated (which I partially ascribe to the little jokes left in the review comments). I would also like to express my gratitude to Carolin Gerlitz and Niels van Doorn for making the time to act as my second and third readers, but also for their general support and contributions as part of the Media of Calculation research group at ASCA. I would further like to give a warm thanks to those members of the staff associated with the Department of Media Studies whom my paths have crossed with for making this such an intellectually stimulating and enjoyable academic learning and working environment. Finally I would like to thank my friends and fellow student assistants at the department for their votes of confidence, encouragement, and welcomed distractions.

Amsterdam, June 2015

Abstract

This thesis investigates prediction and the stuff of which it is made. Over the recent years social media have attracted both an academic and public interest in its "predictive power"; but when it comes to making predictions people generally agree that this is "hard" or "tough", especially when it involves uncertainty with regards to the future. Predictions are accomplishments and require a *purpose*, considerable social and intellectual investment from *sponsors or advocates*, and mobilisation of existing conceptual and material resources. Rather than a specialist reading of concrete cases of prediction, the objective of this thesis is to develop a framework for conceptualising and analysing the stuff of prediction in at least some of the many ways that it exists, and in which it is imagined, accomplished, experienced, and thought through. More specifically, it investigates the stuff of prediction both empirically and conceptually (and historically), with a particular focus on the specificity of its techniques as they find application in concrete settings. Two emblematic practical goals or purposes for social media-based prediction are investigated: forecasting the pulse of social media streams and the surveillance of influenza-like illness using Web search data. How to analyse the relation between the stuff of prediction and the social circumstances and practicalities with which it is inevitably entangled? What are techniques of prediction using their resources for? At the same time it also does a methodological contribution by making the exploration of what it means to take prediction as an object of study an integral part of the project itself, as opposed to committing to such a view from the outset. What does it mean to take prediction as an object of study; how to conceive of it intellectually? Responding to a growing public and academic interest in the predictive power of social media, and in prediction as a way of dealing with challenges characterised by uncertainty and risk more generally, the proposed framework enables a critical analysis of the production of prediction with a particular sensitivity towards its techniques, the resources they mobilise in light of a certain specific practical goal, and the social and cultural significance of their applications in diverse concrete settings.

Keywords
Prediction, calculation, techniques, quantification, social media, big data, uncertainty

Contents

Acknowledgements *v*

Abstract *vi*

Contents *vii–ix*

List of Tables *x*

Introduction, the Stuff of Prediction ***1–12***

Uncertainty poses major challenges to contemporary society on the macro scale, but also shows itself when dealing with minor problems in concrete situations. One particularly interesting response available to us when facing such uncertainties is prediction, which can be incredibly useful, sometimes even lifesaving, in supporting diverse decision making practices about possible developments and events. Yet while these themes are extremely common today, and while knowledge and skill in the area of prediction is typically regarded highly valuable, engineers and researchers deal primarily with the development, application, and evaluation of prediction methods or models, excluding the transformative pressures they exert on reality from these discussions. Therefore, a critical framework is needed to study the cultural, social, and political significance of prediction, how it is made, and what it is made of. How, when, and where is the problem of uncertainty addressed with prediction? What is proposed is a framework around the concept of *cultural techniques*, stressing the importance of purpose, sponsors or advocates, and the mobilisation of existing (plastic) resources.

PART I Resources

What does it mean to take prediction as an object of study; how to conceive of it intellectually?

1 The Artificial as Cultural Mediator ***14–23***

Prediction should be considered an accomplishment, which requires a *purpose*, considerable social and intellectual investment from *sponsors or advocates*, and mobilisation of existing conceptual and material resources). How to understand the (social) circumstances that help constitute the stuff of prediction? Some general features of quantification and measurement are introduced, as well as the notions of evidence and of the empirical, in both cases demonstrating that specific aspects of prediction are inevitably caught up in larger social projects. In addition, the quantification of uncertainty is addressed, discussing different kinds of uncertainty, and why some are maybe more useful than others. The production of quantification, and the production, circulation, and mobilisation of existing *plastic resources* drawn upon (e.g., conceptual and material resources, styles of reasoning, and so forth), serves to mediate and stablise particular cultural beliefs and values associated with certain social groups and communities.

2 The Art and Science of Prediction ***24–36***

The methods and models used for prediction do not come into being without a *purpose*, or without *advocates*. There are many different predictive purposes, some of which are interpolations, while others are more like extrapolations and extend beyond what we know on the basis of *prior knowledge* or experience. This generates not only a space of possible kinds of prediction, where some kinds are more appropriate than others for a given purpose (e.g., accuracy or economic benefit), but also introduces the notion of different statistical cultures, and highlights the importance of distinguishing application areas. Simply not all investment in prediction is geared towards the progression of universal knowledge. In broad strokes, some of the major disparities between these cultures are sketched through a discussion of the literature. This also serves as a means to situate the emerging field of social media-based prediction within this area of thinking and practice, which is a particularly interesting field because it is entangled with a number of ongoing trends like the rise of big data and the growing interest in machine learning techniques from researchers, as well as from companies and governments.

PART II Techniques and Applications

What are techniques of prediction actually doing; what are they using their resources for? How to analyse the relation between the stuff of prediction and the social circumstances and practicalities with which it is inevitably entangled?

3 Forecasting the Pulse of Social Media Streams ***38–48***

Online trend and popularity prediction for social media has become one of the recurring purposes of prediction in multiple areas of practices. For instance, people and companies are interested in exploiting social media data sources to predict such things as the popularity of content for the purpose of "buzz marketing", product and reputation monitoring, the detection of controversial or breaking news, and so forth. As a result, techniques have been developed, used, and evaluated that aim specifically at detecting events – real-world occurrences that unfold over space and time – using cheap or publicly available social media data and streams. Framed as an information retrieval problem, the underlying challenges for this kind of prediction concern the task of identifying new and "interesting" topics or topic areas that were not previously known about and are growing rapidly in importance within a certain corpus of collected data or stream. Is it possible to characterise a space of possibility for this predictive purpose through an analysis of some of its central procedures and operations and the resources they mobilise? What kinds of prediction are deployed in this space of application and what do they achieve?

4 Surveilling Influenza-Like Illness Using Web Search Queries ***49–59***

Disease outbreaks have been a recurring subject of social-media based prediction. In fact, services like Google Flu Trends and derivative applications have repeatedly been celebrated for their success at accurately monitoring and predicting both seasonal and pandemic flu (and potentially other disease activity) using aggregated historical logs of online Web search queries from billions of users around the world. In an attempt to provide faster or earlier detection – that is with less "reporting lag" than traditional flu surveillance systems – techniques have been deployed to monitor "indirect signals" that serve as a proxy of influenza activity. Rather than trying to foretell the future with absolute certainty, these techniques are applied regardless their margin of uncertainty (e.g., for

disease control and prevention). Furthermore, what is interesting about this particular example is the wider discussion it has generated concerning the usefulness of social media-based prediction. This discussion provides meaningful insight into some of the factors with which these techniques of prediction are inevitably caught up precisely because of their concrete applications. What kinds of techniques are deployed to achieve prediction; and how to study the social and cultural significance of their applications?

Conclusion ***60–63***

Works Cited *64–79*

Appendix A Survey of Social Media-Based Prediction *A1–A30*

List of Tables

Table 1. Summary of different kinds of uncertainty. *21*
Source: Adapted from Wynne (1992).

Table 2. Summary of different predictive purposes. *29*
Source: Adapted from Ehrenberg and Bound (1993, 167–168; see also Ziman 1991).

Table 3. Main features of the disparity between explanatory and predictive modelling. *29*
Source: Adapted from Shmueli (2010, 293).

Table 4. TDT terminology for topics, events, and activities. *42*
Source: Adapted from Boykin and Merlino (2000, 36).

* * *

Table A1. List of main sources used for expert list building. *A2*

Table A2. Categorisation of literature by application area or prediction subject. *A2–A3*

Table A3. Categorisation of literature by data source. *A4–A5*

Table A4. Categorisation of literature by various analysis and evaluation … *A5–A7*
Sources: Reproduced from Kalampokis, Tambouris, and Tarabanis (2013, Table III–VI).

Table A5. Categorisation of event detection literature by detection task, … *A8*
Sources: Adapted from Atefeh and Khreich (2015).

INTRODUCTION
The Stuff of Prediction

Prediction and Uncertainty

Although a working title is only temporary until something more appropriate has been decide upon, it does indicate the main stakes, topics, and especially goals of the project during its development. As such it seems appropriate to begin with a brief reflection on the selected working title for this research project – "The Taming of Uncertainty" – and why it was selected. As the avid reader of Ian Hacking's work will recognise, the working title is patterned after the title of his book *The Taming of Chance* (Cambridge University Press, 1990) in which he has documented the development of probability from its emergence in the seventeenth century to the late nineteenth century. In particular, building on previous insights from his earlier work on *The Emergence of Probability* (Cambridge University Press, 1975), this later book argues for an "erosion of determinism" (1990, 1) in the nineteenth century; a long-term historical development that made room for a concept of probability to come into being, which, analogous to the laws of nature, would be able to express a new type of law. But rather than pertaining to laws of nature, these new laws would extend to people. Thus, as he argued, "[c]hance became tamed, in the sense that it became the very *stuff* of the fundamental processes of nature and of society" (vii, emphasis added). This thesis investigates some of the implications of this development in contemporary society focusing specifically on the role of prediction – informally defined by the *Oxford English Dictionary* as the action of "[stating or estimating], esp. on the basis of knowledge or reasoning, that (an action, event, etc.) will happen in the future or will be a consequence of something" ("predict, *v.*", 1.a), and more formally as having "a deducible or inferable consequence" (1.b). Similarly, the title of this introduction is a play on Matthew Fuller's "Introduction, the Stuff of Software", which is at the beginning of a hardcover entitled *Software Studies: A Lexicon* (The MIT Press, 2008). This edited volume develops a series of short studies on "particular digital objects, languages, and logical structures" (1), and thereby simultaneously engages with the question of how to approach software more generally. To that end it develops a number of viable research perspectives and directions. But although my approach is indebted to each of these authors, or to their ideas and methodologies (as well as to some others, most

notably Michel Foucault), the specific context of my own investigation is also quite different. My own objective is similar and is to develop a viable approach to study the stuff of prediction – that of which prediction is or may be made (cf. *OED*, "stuff, *n.*[1]", II), or its "[m]aterial to work with or act upon" (II.2.a) – and thereby show the stuff of prediction in at least some of the many ways that it exists, and in which it is imagined, accomplished, experienced and thought through (cf. Fuller 2008, 1–2). Through an analysis of concrete examples and their interplay it explores the conditions of possibility for prediction established by computers. This matters greatly because prediction is profoundly shaping the way we do things in a world that is increasingly concerned with knowing what lies ahead (e.g., anticipating risks, terrorist threats, natural disasters, disease outbreaks, election outcomes, stock market dynamics and trends, product sales, information dissemination, technology acceptance), or thinking we do with a some degree of certainty, making decisions and plans based on that knowledge (e.g., public policy, preventive measures like evacuation plans, stock investments, credit assessments, marketing strategies, product or friend recommendations). This also points towards a main reason as to why a topic like this would be of interest to scholarship and those working in the field of study of media and culture in particular, since in these fields it is often about connecting – either more or less directly or indirectly – media to purposeful action by humans, or specific things people try to do and achieve with media. The kind of stuff under investigation in this thesis may be complex, arcane, and have technical depth, but the main goal is ultimately to relate to intentions, purposes, goals, imaginations, and so forth (e.g., automating translations, regulating auction markets, detecting trends, and so forth).

As a second reason for selecting this working title, the project has taken its title seriously. Many have noted that making predictions, especially about the future is "tough" (e.g., Bezzi and Noppen 2010; Drever 2011; Gayo-Avello 2012; Lemberg 2001; Woods 1999).[1] Uncertainty is conceived as something we conceptualise and bring into being as an issue (e.g. in relation to money, danger, or threat) and then decide how to live with and act upon (or at least we decide how to relate to it). It poses a problem, and so the question of how it should be faced becomes a collective and political matter of concern. Furthermore, if prediction can be conceived as a particular attitude towards facing this problem, then it is clear that prediction requires tremendous effort and coordination of skill, expertise, knowledge, commitment, and so forth, especially when phenomena are characterised by high degrees of complexity or inherent uncertainty (e.g., nonlinear dynamics and processes).

1. Such statements have been attributed to a diverse collection of individuals, including Niels Bohr, Sam Goldwyn, Robert Storm Petersen, Casey Stengel, and Yogi Berra, to name just a few.

It becomes something that needs to be achieved or accomplished through different approaches, commitments, subscription to particular ideas, concepts, methods, rationales, imaginaries, and even ideologies that underpin the particular techniques, institutions, and practices we have come to rely on for much of our routine as well as our non-routine and technical[2] decision making processes, procedures, and systems. Over the last decade, for instance, decision-making practices involving uncertainty and/or risk have increasingly formed around processes of quantification with indicators, indexes, predictors, and other kinds of measurements, metrics, and numbers serving as a basis of evidence and trust (cf. Porter 1995) in today's "administrative and managerial proceduralism" (Power 2004, 771). As such, it is suggested that prediction – like software in Fuller's view – can be seen "as an object of study and an area of practice for kinds of thinking and areas of work that have not historically 'owned' software, or indeed often had much of use to say about it" (Fuller 2008, 2). Rather than just "realised instrumentality" (3) or tools, both software and prediction are themselves part of the constitution, sustenance, and disruption of societal formations or *complex assemblages*. That is, they participate in the realities they purport only to describe; they have agency and a capacity to change what it means to represent and intervene (cf. Hacking 1983).[3] This is a point has been made by some of the leading British sociologists (e.g., John Law, Evelyn Ruppert, and Mike Savage 2011; Savage 2013) over the recent years as well.[4] It argues against the common assumption that methods are merely instruments for coming to know of the world (e.g., the social world, the natural world, and so on) – or what John Law, Evelyn Ruppert, and Mike Savage (2011) have called the "methodological complex" – which puts all of the epistemological burden on devising or selecting the "right" method that can close the divide "between the world on the one hand, and representations of that world on the other" (3).[5] Instead they argue that the starting point for enquiry should be

2. In contrast to political decision making, the notion of technical decision making is directly related to the growing interest in expert knowledge and skill in an advanced technoscientific society, which, as scholars have noted (e.g., Bozeman and Pandey 2003; Callon, Lascoumes, and Barthe 2009; Mitcham 1997), poses challenges to any attempt to involve the wider public in these secluded zones of society.

3. As Ian Hacking has noted, this is why any persistent style of reasoning is self-authenticating, "it can't help but get at the truth because it determines the criteria for what shall count as true" (1991, 240).

4. In turn their arguments build on works like Steven Shapin and Simon Schaffer's *Leviathan and the Air Pump* (Princeton University Press, 1989), which investigates the material circumstances for a particular style of scientific reasoning to emerge, and develops an argument as to why it was opposed to Hobbes (cf. Hacking 1991, 241), and Peter Galison's *How Experiments End* (Chicago University Press, 1987), which argues that "instruments have a life of their own".

5. A philosophical problem, which be seen as an extension of the *mind-body problem*, referring to a dualism that maintains a rigid distinction between mind and matter, or physical and mental substance.

"that methods are fully *of* the social world that they research; that they are fully imbued with theoretical renderings of the world" (4, emphasis in original). This means two things: first, methods are social insofar they emerge from within a particular social world or context of which they are themselves part, and second, methods are social insofar they help *constitute* these worlds. Methods, then, are shaped by their (social) circumstances, and although the practicalities can be complex and messy, it is clear that, first, methods don't come into being without a *purpose*, second, that they need *sponsors or advocates* – "or more exactly . . . forms of patronage" (5) –, and third, that they draw upon or are adaptations of, *existing resources*, methodological, cultural, or social. This critical perspective, which is particularly indebted to arguments put forward by researchers in the field of Science and Technology Studies (STS), has renewed interest in the question of method and has generated a host of studies around methods in statistics practices and packages (e.g., Mair, Greiffenhagen, and Sharrock 2013, 2015; Uprichard, Burrows, and Byrne, 2008), the social sciences and the "politics of method" (Ruppert, Law, and Savage 2013; Savage 2010; Savage and Burrows 2007), digital social research (e.g., Lury and Wakeford 2012; Marres and Gerlitz 2014; Marres and Weltevrede 2013), *digital methods* (Rogers 2013), computational social science (Lazer et al. 2009), numbers and numbering practices (e.g., Day, Lury, and Wakeford 2014; Gerlitz and Lury 2014), big data practices (e.g., Barocas and Selbst 2015; Kitchin 2014; Taylor, Schroeder, and Meyer 2014), and so forth.

These two reasons introduce the main stakes, topics, and goals as well as the *leitmotiv* of this thesis. Most generally, a critical framework is developed for investigating prediction as a particular way of taming or dealing with uncertainties in the broader sense. This broader context is defined by the *problem of uncertainty*, which constitutes a range of challenges as it presents itself in different concrete settings. What does it mean to live, work, and think in certain specific settings marked by different conditions of uncertainty (e.g., in finance, government, medicine, engineering, environmental and urban planning, and so forth)? What has to be so about who we are to even imagine certain kinds of objects, and then once those objects exist, how do they intervene and shape us to be in a certain form in the world? In order to approach this problem this project investigates prediction, understood as a particular way of relating to this common problem, which is a central feature of many different domains of contemporary culture and society as indicated by the astounding number of applications of prediction today (as well as the frenzy associated with social media-based prediction). It engages with some very specific computational prediction methods and models, and probes the assorted roles that they play or meanings they obtain in settings facing different kinds of uncertainty. As Isabelle Stengers has argued with regards

to grounding the politics of the future, this "puts the emphasis on the event, the risk, the proliferation of practices" (Stengers 2000, 114).[6] Yet while these themes are extremely common today, engineers and researchers seem to deal primarily with the development, application, and evaluation of prediction methods, as a result of which the transforming pressures they exert on reality is rarely considered and often overlooked.[7] The principal question then becomes how, when, and where the problem of uncertainty is being addressed with prediction, and how to understand the social, cultural, and political significance of its concrete applications in various domains. What does prediction bring to the table in these settings; how to conceive of its involvement intellectually?

Archaeology and Cultural Techniques

There are different ways to reason about the problem of uncertainty in its concrete forms and contexts. As I argue, these different approaches are accompanied by cultural beliefs and ideas of order that emanate from the concrete techniques of prediction in these specific settings, suggesting the need for an approach to putting ideas in context that is simultaneously empirical and conceptual (as well as historical) in scope. To illustrate what such an approach might look like, the studies of Michel Foucault on madness ([1961] 1967), medical perception ([1963] 1973), sexuality ([1976] 1978), punishment ([1975] 1995), and knowledge [1966] 1970, [1971] 1972, 1977), and more recent interpretations of such an approach by scholars like Ian Hacking and Bernhard Rieder is instructive. For example, their *archaeologies* of probable reasoning (Hacking 1975), Google's *PageRank* algorithm (Rieder 2012), and Bayesian information filtering (Rieder 2014) – which are also, in fact, histories of the present (Foucault [1975] 1995, 30–31)[8] – demonstrate that the point is not simply to

6. According to her argument, the way that histories of the sciences ground promises on the past constitutes a "mobilizing model, which maintains order in the ranks of researchers, inspires confidence in them with regard to the future they are struggling toward, and arms them against what would otherwise disperse their efforts or lead them to doubt the well-foundedness of their enterprise" (114–115). Therefore, the production of such a mobilizing model is "the business of scientists, like the law of silence is that of the Mafia." (115).

7. Consider for example how when we start modelling, say a set of users of social networking sites and their platform-specific interactions, as a graph, we can also start asking that graph model particular questions and queries to discover or recommend interesting relationships we might not have known about (or might have known about, but with a lesser degree of certainty). At the same time such a model also restricts or *frames* either a general-purpose or purpose-specific analytical space for the kinds of questions we can (conceive to) actually ask that model.

8. In his influential study of "punishment in general, and the prison in particular" ([1975] 1995, 30), "with all the political investments of the body that it gathers together in its closed architecture" (31), Foucault instructively

recount the respective histories of these ideas, but to develop a way of understanding the particular space of possible meanings these ideas have *today* as they are deployed in the form of concrete techniques, rationalities, and practices. In other words, the "invention" of prediction as a solution to the problem of uncertainty should not be mistaken for a sudden discovery, but is rather a multiplicity of "often minor processes, of different origin and scattered location, which overlap, repeat, or imitate one another, support one another, distinguish themselves from one another according to their domain of application, converge and gradually produce the blueprint of a general method" (Foucault 1995, 138). Consequently, this thesis takes prediction as something that deserves to be investigated as an object of study in its own right; not just as a concept, but as a variegated practice caught up in the minutiae of its local settings. Instead of deciding on a particular notion of prediction from the outset, the task of understanding some of the many forms in which prediction exists is taken as the main *objective*, thereby shifting "the analytic gaze from ontological distinctions to the ontic operations that gave rise to the former in the first place" (Siegert 2013, 48). It is therefore deliberately conceived in a quite abstract (i.e., non-concrete) sense, as an association of heterogeneous elements, including ideas, techniques, objects, practices, processes, and structures held together or mobilised by "a practical rationality governed by a conscious aim" (O'Farrell 2005, Appendix 2, 158).[9] Techniques of prediction are therefore neither completely formal or technical in and of themselves (i.e., representable purely in mathematical or symbolic terms), nor are they completely explained through their relations or tensions with their concretisation in the domains of application from where they also distinguishes themselves from each other (i.e., explained through its material underpinnings and relational features). Instead, prediction is approached as an object *in-the-world* with which we have "meaningful dealings", to speak with Heidegger ([1953] 1996). Concrete techniques then should be understood as historically contingent expressions or associations that cannot be adequately explained either from one side, or from the other side alone. This perspective thus calls for a frame of analysis able to facilitate

---

distinguished between writing a history of the past in terms of the present, and writing a history of the present. As such, a crucial component to the writing of a history of the present is its capacity to intervene in the current situation as a kind of counter-history (see also Foucault 1977; Garland 2014; Hacking 2002; Roth 1981).

9. In the second appendix to her book on *Michel Foucault* ("Appendix 2: Key Concepts in Foucault's Work"), Clare O'Farrell writes that "Foucault defines the Greek word *techne* as 'a practical rationality governed by a conscious aim'. . . . Foucault generally prefers the word 'technology', which he uses to encompass the broader meanings of *techne*. . . . Foucault often uses the words techniques and technologies interchangeably, although sometimes techniques tend to be specific and localised while technologies are more general collections of specific techniques" ("Technology, Technique, *Techne*", 158–159).

both views simultaneously; to understand both the way we make and do things and then also how those things act back upon us.

One possible approach to navigate this problem can be found in the concept of "cultural techniques" (*Kulturtechniken*), as developed by German media scholars like Thomas Macho (2003, 2008) and Bernhard Siegert (2007, 2013, 2014), Geoffrey Winthrop-Young, and others working within field of media archaeology as it emerged over the last few years. Following Thomas Macho's original definition, "[c]ultural techniques – such as writing, reading, painting, counting, making music are always older than the concepts that are generated from them" (2003, 179). They are, in other words, conceived as operative chains. A few years later, however, its definition has been expanded to differentiate cultural techniques from other technologies by positing the former are "second-order techniques". As Macho explains, symbolic work necessarily requires (very) specific cultural techniques: "we may talk about recipes or hunting practices, represent a fire in pictorial or dramatic form, or sketch a new building, but in order to do so we need to avail ourselves of the techniques of symbolic work, which is to say, we are not making a fire, hunting, cooking, or building at that very moment" (2008, 100). Taking prediction to involve cultural techniques is therefore, in the first place, to acknowledge it as an integral component in a series of aggregations (i.e., operative chains), from which symbolic practices may emanate. Further, it also means making a distinction between prediction as informal and routine technique without significant consequences on the one hand, and as having technical, symbolic or political implications and advantages on the other. Furthermore, this concept is useful because it enables researchers to address seriously the concepts of techniques and technology. It was developed with a focus on the materiality and technicality of meaning constitution, which, as Bernhard Siegert observed, "has prompted German media theorists to turn Foucault's concept of the 'historical apriori' into a 'technical apriori' by referring the Foucauldian 'archive' to media technologies" (2013, 50). Consequently this leads to the development of an archeology of cultural systems of meaning, that acknowledges the ontological entanglement of *technicity* – "technology considered in its efficacy or operative functioning" (Hoel and van der Tuin 2012, 187) – of the material substrate of culture rather than implicitly arguing for some version of media or technological determinism (which indeed is one of the labels commonly affixed by those arguing against this programme). Although these material, conceptual, and technical substrata can be understood as framing a specific space of possibility for "vectors of variation, heterogeneity, and fermentation" (Rieder 2012), the concrete techniques cannot be understood as simply following teleologically from these underpinnings alone: "[t]here *is* variation and it is *significant*"

(emphasis in original). The task of trying to characterise the stuff of prediction, and the subordinate task to understand what is *in* its techniques, can then be approached if we take prediction as cultural techniques that constitute a linchpin in creating, sustaining, and disrupting particular micro-networks binding heterogeneous elements (like those mentioned previously) together into meaningful common activity centered around the taming of uncertainty.

Expertise, Calculation, and Judgement

In many ways this is also an argument about practical forms of reasoning, which further situates this study in a particular tradition of scholarly work dubbed the "Third Wave of Science Studies" (Collins and Evans 2002) focusing on Studies of Expertise and Experience (SEE) in an attempt to include other forms of especially non-professional expertise into public discussion and (technical) decision making (e.g., Callon, Lascoumes, and Barthe 2009). What role do we have today in witnessing, representing, and intervening in computational epistemic or knowledge-making practices and their associated "regimes of existence"[10] (Teil 2012; Taylor et al. 2014)? How is (computational) prediction taking on an agential role in what practitioners in different domains of application do, and what is at stake in the emerging entanglements of written lines of code, functions and algorithms, computer models, numbers and metrics, software, practitioners (or those with some degree of "contributory expertise") and laymen (or those with some degree of "interactional expertise"), and so forth? Thus, although it might not satisfy the conditions of classical epistemological accounts, this thesis does concern itself with the question of knowledge or understanding, how we come to know of the world, and how that knowledge is justified or legitimised in practical life (e.g., policy and lawmaking in government, risk assessment and management, the efficient allocation of resources in business, decisions on standards in safety engineering, recommender systems for movies, language translation tasks, news filtering in social media streams, information and document retrieval on the Web, and so forth). But the kinds of knowledge prediction it is about, and the specific practical means required to create and justify that knowledge also change and may vary significantly across domains of application. What does it actually mean to predict, and what does it mean to justify or legitimise the value of a prediction – and its associated courses of action and

10. As Geneviève Teil explains, "[o]bjectivity . . . ascribes a specific regime of existence to objects, that of 'things' consistent with the definition of a dualist rationalist ontology: 'data' that can be 'discovered' and whose existence unfolds independently, including from the people who live around, with or alongside them" (2012, 3).

models of decision-making – in relation to other competing alternative scenarios, and in some cases even preferred ones? These questions immediately point towards the importance of the "seemingly innocent notion of a style of reasoning" (Hacking 1990, 7). Like any style of thinking or reasoning, a prediction, it seems, "cannot be straightforwardly wrong, once it has achieved a status by which it fixes the sense of what it investigates. Such thoughts call in question the idea of an independent world-given criterion of truth." (ibid.). Put differently, just as "[a] proposition can be assessed as true-or-false only when there is some style of reasoning and investigation that helps determine its value" (ibid.), a prediction can be assessed only when there is some way of understanding its value. But since differences of opinion exist among experts – and between experts and users – this problem is not easily resolved and instead has generated discussion about the nature of quality – "goodness" – within, for instance, weather forecasting (e.g., Murphy 1993; Silver 2012, 128–134).[11] It is in this sense that prediction, and the development of particular styles of reasoning can be intimately connected with much larger questions about what a society is, or its various subdomains.

The work of Michel Callon, Pierre Lascoumes, and Yannick Barthes (2009) is particularly instructive here, asking the seemingly innocent question of what it means to act in an uncertain world, and thereby ultimately exposing the limitations of traditional delegative (representative) democracy in favour of an enterprise they term "the *democratization of democracy*—that is to say, of the people's control of their destiny" (11, emphasis in original). At the same time, however, I do believe we have to be cautious to develop a critique of technology that is not automatically also a critique of modernity or modern civilisation (cf. Rieder 2014, 16). This is why we have to engage the technical on the level of techniques, objects, processes, and structures. Calculative operations as those involved in prediction have inscribed into them particular technical schemes of classification, valuation, and so forth, which tend to disappear from view and scrutiny as soon as they become commonplace practices and rationales (e.g., Bowker and Star 2000). This is the stuff of such calculations. Following Michel Callon and Fabian Muniesa's (2005) insightful work, studies of calculation and calculative behaviour or practices in supermarkets, trading floors, and so forth by

11. In his influential essay on the nature of goodness in weather forecasting, Allan Murphy (1993) makes a case for three distinct types of "goodness", which he summarises as follows. First, "*consistency*", or the correspondence between forecaster's judgements and their forecasts. Second, "*quality*", or the correspondence between the forecasts and the matching observations. And third, "*value*", or the incremental economic and/or orther benefits realised by decision makers through the use of the forecasts. (281) Nate Silver (2012), in his own reflections on the matter, draws upon this model but uses slightly different terms. He instead defined "accuracy", "honesty", and "economic value" as three ways of judging what makes a good forecast.

anthropologists and sociologists (e.g., Cochoy 2008; Knorr-Cetina and Brügger 2002; Miller 1998) proved to be matters of "pure judgement or conjecture or, when it can be observed, something originating in institutions or cultural norms" (1230). These scholars have shown how actors in such settings rarely engage with purely arithmetic operations in the strict sense, but often interpret information and take decisions while lacking clarity or understanding as regards the criteria needed to make such decisions. Callon and Muniesa have therefore suggested a wider account of calculation that extends beyond purely mathematical or even numerical operations (e.g., Lave 1988). Instead, "[c]alculation starts by establishing distinctions between things or states of the world, and by imagining and estimating courses of action associated with those things or with those states as well as their consequences" (Callon and Muniesa 2005, 1231). The concept of "qualculation", originally coined by Frank Cochoy (2002), is therefore deemed more appropriate because it redefines calculation to include judgement. Furthermore, John Law and Michel Callon's (2005) refinement of the concept further demonstrates how establishing the artificial boundary between the calculative and the noncalculative – that is the making of *qualculability*, or calculation with judgement – is never trivial, but requires tremendous effort and should be considered an accomplishment in its own right.[12]

* * *

This introduction has developed an initial understanding of what it means to take prediction as an object of study relevant to contemporary society. Summarising the main objective, this project is an attempt to conceptualise prediction and the stuff of which it is made; to form a concept or idea of its techniques in relation to the assorted roles they play in different concrete settings. This leads to three interrelated axes of investigation: first, prediction techniques (concepts, methods, and models); second, their spaces of application (concrete applications in the domain of social media); and third, the derivative or associated concepts and practices that are generated from these cultural techniques. The structure of the remainder of this thesis reflects the two levels on which I have introduced these main

12. In addition to these two accounts, Nigel Thrift (2004), a human geographer, also develops a notion of "qualculation" based on Callon and Law (2004) that is often cited in the literature. His notion is in response to calculation becoming "a ubiquitous element of human life" due to the sheer growth in computing power, an increasing ubiquity of hardware and software, and changing forms of calculation (586–587). The problem then becomes how to represent this increase in calculation and its consequences. As he argues, these developments produce a new "calculative sense", which is characterised by speed of calculation, faith in number, limited availability of numerical facility in the bodies of the population, and finally some degree of memory (592).

concerns and goals. At one level it investigates the stuff of prediction both empirically and conceptually (and historically), with a particular focus on the specificity of its techniques as they find applications in concrete settings. The main question here is how to analyse the relation between the stuff of prediction and the social circumstances and practicalities with which it is inevitably entangled? But as the attentive reader will undoubtedly have noticed there is still a lacking specificity as regards the specific methods of analysis. This is because at another level it also does a methodological contribution by making the exploration of what it means to take prediction as an object of study an integral part of the project itself, as opposed to committing to such a view from the outset. The main question here concerns what it means to take prediction as an object of study; how to conceive of it intellectually? Accordingly, the structure of this thesis is divided into two parts each comprising two chapters: the first part ("Resources") concentrates on methodological questions and develops a critical framework for analysis and the second part ("Techniques and Applications") consists of the analysis of two emblematic cases or examples. Both lines of investigation are drawn together in the conclusion to reflect on the broader implications and consider how social media-based prediction relates to some of the broader concerns raised by the problem of uncertainty.

The first chapter develops an argument for conceiving of prediction as something to be accomplished with effort and commitments to certain specific skills, theories, and so forth. As will be argued, the efficacy to achieve a particular predictive purpose depends considerably on social and intellectual investments from sponsors or advocates (as well as from nonhuman actors like machines) and requires a host of existing conceptual and material (plastic) resources to be mobilised. As a result, the production of quantification, measurements, and artificial phenomena more generally can be conceived as cultural mediators insofar they are made and shaped by humans and reflect their values and beliefs. How to conceive of the intimate relation between the technical artefacts of prediction and the cultural systems of meaning they exist within? The second chapter explores how prediction exists as a subject and as a field in the literature, concentrating in particular on different cultures of prediction that may be distinguished. Depending on disciplinary affiliations, the use of concepts and definitions, commonly addressed problems, predictive purposes or reasons for doing prediction, and the associated methods and models in each case may vary modestly or significantly. This empirical sensitivity with regards to prediction as an area of practice for various kinds of thinking and areas of work is necessary for understanding the specificity of its applications in particular domains. As such it is also relevant for making a case as to why social media-based prediction is particularly interesting and relevant as an

object of study today. Chapters three and four each develop an analysis centered on an emblematic practical goal or predictive purpose for social media-based prediction. For chapter three this is forecasting the pulse of social media streams and for chapter four it is surveilling influenza-like illness using Web search data. Both chapters focus on the questions of what kinds of prediction are used and what the techniques in these specific examples are actually doing or using their resources for. However, instead of simply applying the developed framework, the two analyses are used to highlight different aspects (i.e., what makes them emblematic) and build on the framework from both directions. While the third chapter focuses primarily on identifying some of the more general resources that are mobilised to achieve a range of perhaps more general practical goals, the fourth chapter concentrates on linking the mobilisation of these resources in light of a certain specific practical goal to the social circumstances and practicalities with which prediction is entangled. The conclusion, as indicated, draws both lines of investigation together to reflect on the initial questions and objectives and the further implications of this project.

PART I
Resources

CHAPTER 1
The Artificial as Cultural Mediator

Quantification and Measurement

Predictions are generally based on numbers and metrics that have already undergone considerable treatment in order to prepare them for further calculation and analysis. The accomplishment of prediction is thus inevitably caught up in questions of measurement and quantification, and therefore embedded in larger social projects. Wendy Espeland and Mitchell Stevens have argued that pressures to devise and revise measures from governments, organisations, industry, and the general public have expanded greatly over the recent decades, "[w]hether as an effort to incorporate scientific evidence into policy decisions, extend market discipline to government or non-profit organizations, integrate governments and economies, or coordinate activity across geographical and cultural distances" (2008, 402). Yet, although there is a growing demand for the quantification of diverse social phenomena (e.g., Porter 1995; Power 1997, 2003; Strathern 1996, 2000), little attention has been paid to quantification – "the production and communication of numbers" (Espeland and Stevens 2008, 402) – as a general sociological phenomenon. Instead, sociologists have been interested mainly in the accuracy of their measurements for the study of societies, rather than with the significance of their political, economic, social and cultural implications. Building on their earlier collaborative work on "commensuration" – "the transformation of different qualities into a common metric" (1998, 314) – as a general social process, Espeland and Stevens have addressed this "oversight" by developing a conceptual framework for empirical investigations of quantification along some primary dimensions of quantification. One important distinction among forms of quantification they introduce, for example, is between those that mark and those that commensurate. When numbers are used to mark, they distinguish one object or person without quantifying. In these cases, quantification "precisely represents and orders knowledge in meaningful, useful ways but it does not measure" (407). The examples are numerous and include ISBN numbers, telephone numbers, the Dewey Decimal Classification (DDC) system, and unique identifiers (UID). Numbers that commensurate, on the other hand, transform different qualities into a common quantity, thereby creating a specific type of relationship among

different objects by which they are rendered formally comparable to one another. But this requires considerable social and intellectual *investment* (from sponsors or advocates) or work – to establish distinctions between things, to unite them under a shared system of reference, assigning every object "a precise amount of something that is measurably different from, or equal to, all others" (ibid.). As they explain,

> If the categories of classification are broadly agreed upon, commensuration may appear to be a simple matter of specifying incremental differences between otherwise similar things. If, however, commensuration creates a relationship between objects that are not conventionally regarded as comparable, we are more aware that we are doing something by commensurating. (ibid.)

This illustrates not only how quantification enables "[making] things which hold together" (Desrosières 1991) that were not previously related, but also suggests that this process is explicitly political. Settling on controversial matters like scales, the units of measurement, and categories requires an agreement to be formed among all the invested actors.

Espeland and Stevens further note that marking and commensurating represent two ends of a dimension of quantification, or two "levels of measurement" (Stevens 1946). Similarly, Michael Power distinguishes *first-order* from *second-order measurement* (or "meta-measurement"), "respectively as particular institutions of counting and data production, and as related dense networks of calculating experts operating on these numbers within specific cultures of objectivity" (2004, 767). These two modes of measurement therefore involve quite different skill sets. While first-order measurement tasks are often delegated to recording devices and data capture mechanisms, second-order measurement is done by calculation experts from a wide variety of fields and disciplines that work with these measurements within specific cultures of objectivity (cf. Teil 2012). Thus, just as financial derivatives for example derive their value from the underlying entities (or at least ideally, which is to say derivatives can also end up living a live on their own), first-order measurement also generates new possibilities for calculation that derive their efficacy from the underlying numbers and metrics and their characteristics (e.g., varying degrees of precision, accuracy, validity, reliability, completeness, bias, and so forth).[13] That is, quantification generates new possibilities for working with cumulative measures, aggregated evaluations, rankings, similarity measures, clustering, and so forth, enabling new modes of calculating, ordering, classifying, predicting, profiling, constraining, and optimising.[14] For

13. Cf. footnote 41.

14. And in turn these measures end up doing work in the world (e.g., Espeland and Sauder 2007; Sauder and Espeland 2006, 2009).

this reason it is too simplistic to critique quantification in general (or digitisation for that matter) based on its inherent reductionism, since at the same time a host of other capacities are gained by virtue of their relations to other entities, their inclusion in a multiplicity of sets or lists, and so forth (e.g., Mackenzie 2010, 2012, 2013, 2014). Quantification and measurement can thus be regarded as preconditions for a particular kind of calculative techniques, including those of empirical or evidence-based prediction.

Evidence and (Media) Empiricism

The success of a statistical predictive model ultimately hinges on the robustness of the empirical relationships generally detected in past data to infer about the future. One possible approach to the question of quantification in prediction is thus found in the various possible attitudes producers and users of statistics may have towards "reality" or the empirical. As Alain Desrosières observed with regards to "reality",

> This "reality" is understood to be self-evident: statistics must "reflect reality" or "approximate reality as closely as possible." But these two expressions are not synonymous. The very notion of "reflection" implies an intrinsic difference between an object and its statistics. In contrast, the concept of "approximation" reduces the issue to the problem of "bias" or "measurement error." (2011, 339)

Moreover, with regards to the latter, the problem of bias or measurement error (e.g., sampling and observation errors) is something that may be corrected for (i.e., it may be taken into account). Desrosières goes on to distinguish between four attitudes towards reality, each with its own approach to the "orchestration of reality" (346), and each with different "reality tests" and rationales for justification. They are metrological realism, pragmatism of accounting or "'accounting' realism" (344), "proof-in-use", and finally the constructivist attitude. The interrelated specificities of these different groups, their attitudes to reality, their rationales of quantification, and the applications of their work explain these cultural differences. The first, second, and third attitudes seem particularly relevant with regards to the emerging area of practice of social media-based prediction (see Chapter 2). In the first group, the object of measurement – the ascertaining or appointment of numbers or quantity to a dimension of an object or event (cf. *OED*, "measurement, *n.*") – is considered just as real as the physical object (341), and "[t]he vocabulary used is that of reliability: accuracy, precision, bias, measurement error (which may be broken down into sampling error and observation error), the law of large numbers, confidence interval, average, standard deviation, and estimation by the least-squares method" (ibid.; see also Desrosières 1998;

Hacking 1990; Stigler 1986). For those in the second group, justifications are primarily pragmatic and based on practical *purpose* (e.g., their approach is needed for policy guidance and monitoring or business administration). For those in the third group, in contrast, "'reality' is nothing more than the database to which they have access" (346), which is a thoroughly pragmatic view, based on purpose. It is argued that users in this group therefore need to trust the source database "as blindly as possible to make their arguments – backed by that source – as convincing as possible" (ibid.).[15] This insight has led some sociologists to develop a critique against users who treat statistical data or empirical evidence as "raw data", which they argue, "is both an oxymoron and a bad idea" (Bowker 2005, 184; Gitelman 2013). What these differences suggest is not so much that one attitude is "better" (e.g., more "objective" or empirically accurate) than the others, but rather that different "communities of practice" (Lave and Wenger 1991) coexist next to one another, each "with situational constraints specific to particular phases of the statistician's technical, administrative, or managerial work" (Desrosières 2011, 340). Examining different uses of data and statistics, including for prediction, reveals the diverse argumentative contexts in which they exist and which give them meaning and purpose (e.g., preventive policymaking).[16]

Considering such diverse attitudes exist towards statistics and its data sources, it makes sense to examine their own styles or "languages", to use Desrosières's word, and how they stabilise across physical and cultural distances. The question of what counts as evidence and how it is justified, for example, is far from a straightforward matter. Although we are dealing with empirical data or evidence here, we cannot simply keep our distance and just observe or witness a phenomenon with direct access to it "'as nature poses it'" (Galison 1996, 155). Evidence is not "what we find among us, bring home from abroad, [or] chart in the skies", but rather "constructed" and what we often "*make* with machines" (Hacking 1991, 235, emphasis added).[17] Therefore, the accuracy of empirical or evidence-based prediction

15. With regards to this attitude, Desrosières further notes, "[t]he user's trust in the data-production phase is a precondition for the social efficiency of the statistical argument" (2011, 349). This also explains why social media companies for example are invested in creating and sustaining a discourse that enables them to maintain this trust (e.g., stimulating users to create real online identities or maintaining that whatever happens online reflects what happens offline).

16. "Purpose" is defined as "[t]hat which a person sets out to do or attain; an object in view; a determined intention or aim" (*OED*, "purpose, *n.*", 1.a); "[t]he reason for which something is done or made, or for which it exists; the result or effect intended or sought; the end to which an object or action is directed; aim." (2).

17. Hacking raises this point in his review of Steven Shapin and Simon Schaffer's *Leviathan and the Air Pump* (Princeton University Press, 1989) in his discussion of laboratory apparatuses. Yet today this point seems quite

depends heavily on measuring the "right" predictors (i.e., independent variables), thus shifting the emphasis to the production of (new) "'measureable' subjects" (Power 2004, 777) and the required degrees of precision in different settings. For example, in his best selling paperback *The Signal and the Noise* (Penguin Books, 2012), Nate Silver engages with the question why so many predictions fail, while only some are successful. Although his account is certainly not the only critique of prediction across the board, its contribution lies in teasing out nuances between different prediction tasks and communities on the basis of how the signal and the noise are treated. It highlights the assumptions, concrete tasks, and moments of decision-making in the process of selecting or deriving predictors and indicators, and modelling their relations in light of diverse practical purposes such as economic activity, election outcomes, and weather forecasting. Relating these ideas about evidence to the growing interest in big data (mainly from industry, but also from academia), especially after the large-scale uptake of online social media since the mid 2000s and the possibilities this has generated for researchers from diverse areas, it is not surprising that the use of empirical methods in other public and semi-public domains of society (e.g., commerce, entertainment, government, healthcare, policing, and surveillance) has also increased considerably.[18] In particular, social media have given rise to a particular breed of *media empiricism*, or the idea that knowledge may be gained and justified on the basis of sensory experience or evidence grounded in these media, thereby "closing the age of Internet innocence" (Digital Methods Initiative 2015) in the context of social media.[19] This kind of empiricism is extremely useful for recommending products or advertisements to users, detecting trends or brand communities, and many other things that take place within these media environments. In addition, many studies on social media-based prediction attempt to use empirical material like user transaction data, profile data, or search queries to predict real-world events or occurrences that do not take place online (or at least not initially). Although this particular notion of empiricism is developed in more detail as part of the analysis (see Chapter 3–4), the general notion of the empirical and the related concept of *empirical accuracy* – the degree of precision or proportion of correct empirical predictions by a model – are central to understanding the efficacy of prediction in its concrete settings.

---

prescient, considering the growing interest in empirical prediction with social media data over the recent years.

18. In turn such trends can also have far-reaching and unexpected implications for these domains (e.g., Savage and Burrows 2007; Burrows and Savage 2014).

19. This statement is in reference to normative developements such as real-name policies, which add significant accountability to the user and thereby signify a cultural shift.

The Quantification of Uncertainty

Having introduced some general features of quantification, measurement, evidence, and the empirical, the next step is to explore what these might have to do with prediction and uncertainty. To start, it is instructive to recall Frank Knight's (1921) original distinction between uncertainty and risk in economics. Although the notions of risk and uncertainty are often used interchangeably, they should not be confused because the notion of risk is indispensible if we are to understand decision-making processes (whether human, nonhuman, or hybrid). Knight argues that the terms cover two things that are categorically different:

> Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated. . . . The essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomenon depending on which of the two is really present and operating. . . . It will appear that a *measurable* uncertainty, or 'risk' proper, as we shall use the term, is so far different from an *unmeasurable* one that it is not in effect an uncertainty at all. (19–20, emphasis in original)

In other words, whereas risk is susceptible to measurement and therefore may be expressed in a quantity (e.g., probability of default, volatility, probability of failure), where after it may be subjected to derivative calculations (it may be assessed, weighted, compared, and so forth), *Knightian uncertainty* (as it is sometimes called), defined against this notion of risk, eludes measurement and therefore cannot be calculated or reasoned with. Indeed *in effect* a "'risk' proper" is not an uncertainty at all, but rather a parameter for control and calculation.

However, in a seminal article from 1992, Brian Wynne has argued that the enlargement of acknowledged uncertainty has not only grown in scale, but also led to the introduction of two new and qualitatively different kinds of uncertainty (112). In the article, Wynne tries to characterise these qualitatively different kinds of uncertainty in his critical examination of environmental policy making, especially in relation to decision making upstream from environmental effects (111).[20] As he explains, risk assessment as a scientifically disciplined way of analysing risk originally emerged as an integral component to design and engineering as a way of dealing with structured mechanical problems where parameters could be well defined, and reliability of different elements are testable and amenable to actuarial in-service

20. Wynne's work is part of this broader tendency of "Third Wave Science Studies" (Collins and Evans 2002) where emphasis shifts from studies of facts – or "extended facts" (e.g., Yearley 2000, 2001) – to studies of expertise and experience.

analysis (113), such as in the case of engineering an aircraft, where the risk of a crash is directly related to the quality of engineering. In other cases like the modelling of environmental risk systems (involving entities that are more "open-ended"), where parameters or causal links are not always known, uncertainties and variations are artificially reduced through such practices as averaging, standardisation, and aggregation (ibid.). But as argued above, such practices involve considerable work – investment and commitment to theories and models, arguments, devices, institutions, and so forth. Wynne gives the example of a single pH measure for a lake as a composite variable, which is composed of surrogate variables to account for parameters of which we have imperfect knowledge. As a result the analyst may have to extrapolate and weight sample measurements into mean values. But "commensuration" (Espeland and Stevens 1998) is only part of this work. Dealing with uncertainty requires not only (continuous) examination of all available evidence,[21] but also of weighing the options between competing interpretations or "*possible* states of the world" (Callon, Lascoumes, and Barthe 2009, 20). In addition, definitions are required that enable (different kinds of) experts to reason with (manipulate, work with) these uncertainties, especially when they are qualitatively different. Wynne therefore continues to develop a typology of four kinds of uncertainty, comprising of the conventional notions of "risk" and "uncertainty", but adding "ignorance" and "indeterminacy" (Table 1; Wynne 1992, 114). While "uncertainty", "ignorance", and "indeterminacy" all denote degrees of *not knowing*, "risk" denotes *knowing* and therefore invites the possibility for *rational choice*, weighing means and ends with reasoning. Following Hacking's use of the term, "reasoning" does not just refer to thinking, but also to "all the doing that is associated with it" (1991, 239). As a result, *styles of reasoning* such as "artificial reason" (Dreyfus 1972, 1992) "synthetic reason" (DeLanda 2011), "statistical reasoning" (Desrosières 1998), inductive reasoning, mechanical reasoning, and so on are emphasised as a linchpin, central to coordinating a range of disparate or non-cohesive things like ideas, concepts, and methods, as well as objects, practices, processes, and structures.

21. In the case of environmental risk modelling, the project of science is given a very pragmatic task, namely to provide in such evidence.

Table 1. Summary of different kinds of uncertainty.

Risk	Know the odds.	When the system behaviour is basically well known, and chances of different outcomes can be defined and quantified by structured analysis of mechanisms and probabilities.
Uncertainty	Don't know the odds: may know the main parameters. May reduce uncertainty but increase ignorance.	If we know the important system parameters but not the probability distributions . . . There are several sophisticated methods for estimating them and their effects on outcomes. These uncertainties are recognized, and explicitly included in analysis.
Ignorance	Don't know what we don't know. Ignorance increases with increased commitments based on given knowledge.	This is not so much a characteristic of knowledge itself as of the linkages between knowledge and commitments based on it – in effect, bets (technological, social, economic) on the completeness and validity of that knowledge.
Indeterminacy	Causal chains or networks open.	Conventional risk assessment methods tend to treat all uncertainties as if they were due to incomplete definition of an essentially determinate cause–effect system. . . . They are not merely lack of definition in a determinate cause–effect system; the relationship between upstream commitments and downstream outcomes is a combination of genuine constraints which are laid down in determinate fashion, and real open-endedness in the sense that outcomes depend on how intermediate actors will behave.

Source: Adapted from Wynne (1992).

Callon, Lascoumes, and Barthe (2009) also describe the concept of risk to be closely related to rational decision.[22] Modern societies have come to depend on instruments and institutions like insurance companies and vulnerability indices that structure the identification of risk in a specific cultural context (Beck 2000, 224), and calculate the distribution of risk in such a way that it forms the basis for decision making about these risks (Heimer 1985). Callon, Lascoumes, and Barthe have identified three conditions that need to be met in such cases:

> First, we must be able to establish an exhaustive list of the options open to us. . . . Second, for each of the options under consideration, the decision maker must must be able to describe the entities constituting the world presupposed by that option. . . . Finally, the assessment of the significant interactions that are likely to take place between these different entities must be feasible. (19)

22. In their own view, the term "risk" "designates a well-identified danger associated with a perfectly describable event or series of events" (Callon, Lascoumes, and Barthe 2009, 19).

When all three conditions are met, the decision maker is able to commensurate, formally compare, and weigh all available options. However, since we rarely, if ever have "perfect knowledge" of all possible states of the world or of all conceivable scenarios in advance, these conditions are not usually met. This is why uncertainty is a useful concept for Callon, Lascoumes, and Barthe; it helps them distinguish situations of risk from forms of uncertainty that can only be faced with questioning and debating which measures should be taken. For the purpose of this study, in contrast, the concept of risk is useful to distinguish because contrary to other kinds of uncertainty it can be calculated and reasoned with by humans as well as by nonhumans (e.g., mechanically), even if there is still uncertainty involved in the process. Although this thesis is not concerned with risk *per* se, risk is interesting because it is expresses uncertainty as a *quantity* and as such may be subjected to further calculation, sorting, ranking, prediction, and serve as a basis for decision-making.

Conceptual elements and material procedures like those underpinning processes of quantification and measurement, experience and evidence, empiricism, styles of reasoning, and the theoretical models of the phenomenon under investigation as well as those of the devices we use are examples of what Andrew Pickering has called "plastic resources" (Pickering 1987, 1989). As he and others have noted, the concrete configurations of these heterogeneous elements are not absolutely fixed in advance, but rather brought into relation with each other by the investigator (i.e., they are mobilised), serving as plastic resources for further practice and pragmatic tinkering until the experiment ends and it can cease to be a resource and become a thing in itself (Pickering 1987, 198; see also Franklin 2007; Galison 1987; Hacking 1991, 1996). We often forget how much "infrastructure" lies behind the creation, use, and circulation of these resources and the regimes of which they are part,[23] and this is especially true in cases where resources circulate to places that are removed from the systems and bureaucracies that manufactured them (cf. Espeland and Stevens 2008, 411; Power 2004, 767). By focusing on the plasticity of the conceptual and material resources, and how they are mobilised and utilised as part of different practical rationales and purposes, it becomes clear how the artificial more generally serves as a cultural mediator between different groups and communities. But although we can easily understand that artefacts are products of a given society and therefore reflect their values and beliefs, it is much more complicated to understand how they achieve this (Gimeno-Martínez 2013). A case in point is Peter Galison's (1996) historical descriptive account of computer simulation as a "delocalized trading zone" (155), referring to spaces of exchange for highly diverse and

23. As Geoffrey Bowker and Susan Leigh Star (2000) have noted, once accepted, such systems and infrastructures tend to appear as natural and incontestable.

rich subject matters that need not share a common ontology or set of laws, but are brought together in "strategies of practice" (157) that embody the coordination of these highly disparate conceptual and material resources. Similarly, social media-based prediction constitutes an area of practice for kinds of thinking and areas of work like artificial intelligence and machine learning that have not neccesarily concerned themselves with such things as elections, disease outbreaks, stock markets, and so forth in the past. That is to say it brings skills and knowledge together in ways that they might not have been brought together before. What this demonstrates is the importance of situating the phenomenon, both spatially and temporally, as a viable strategy for defining the preconditions from which artificial phenomena (including ideas and technical artefacts) emerge – that is, to describe the (mobilisation of) resources from which they are composed and articulated. It helps to see how ideas, methods, and other abstract phenomena like collaborations can ultimately be studied empirically as concrete phenomena.

* * *

This chapter has introduced some general features of quantification and measurement, touching upon questions of evidence and the empirical, since prediction is inevitably caught up in these questions and in larger associated social projects. In addition to these general perspectives, the question of quantifying uncertainty was discussed in more detail. The main point being that quantification in its various forms can be studied as an artificial phenomenon (or even artefact), and therefore as a mediator for cultural beliefs and values. In the next chapter, notions of quantification, measurement, evidence, and the empirical are further explored in some of their practical and intellectual contexts by means of a (selective) review of the statistical literature on prediction. It will also discuss a body of literature that specifically addresses the question of social media-based prediction, an emerging area of work that is interesting not least because it is highly invested in by an industry still looking to explore the possibilities of exploiting big data for diverse predictive purposes.

CHAPTER 2
The Art and Science of Prediction

Routine Predictability and Prediction

Those working in the field of statistics generally regard prediction as a difficult problem, and the stories of successful prediction are therefore rare (e.g., Gayo-Avello 2012; Silver 2012). Yet at the same time many phenomena are considered to be *routinely predictable*, that is we can expect a particular outcome to occur – or recur – with a reasonable degree of certainty (e.g., it may be based on a scientific law, such as when we predict the alignment of particular celestial bodies or the time of a sunset/sunrise). Expectations of this kind typically arise from experience – even in everyday life – for example when similar outcomes have been observed to occur before, but with (many) different sets of data and under (very) different conditions. That is to say such experience should ideally be based on examples that are independent from one another, and are not linked by causal relationships. This implies that the prediction expert is burdened with the task of justifying the independence of those particular phenomena that serve as examples, or else should describe all of the assumptions that are made. But as Andrew Ehrenberg and John Bound (1993) have noted, there is an "extreme divergence between theory and practice" (167), which they attribute to the fact there are many different kinds of prediction or predictive purposes (e.g., Ziman 1991), some of which can be seen as interpolations into what is already known (i.e., prior knowledge or what counts as known "experience", which is useful for making further inferences), while others are more like extrapolations, extending beyond what we know based on prior knowledge or known experience into an area not known or previously experienced so as to estimate the value of some variable based on its relationship with another variable (e.g., Bayesian statistics). Table 2 presents a summary of Ehrenberg and Bound's typology of different predictive purposes, which they introduce as a way of arguing that their differences have often been confused by statisticians. In addition to classifying prediction types according to its purpose or what we want to predict (e.g., close values or relationships between them or emerging trends, conditional versus unconditional prediction), another basic criterion is the data used for fitting or *training* with a model (e.g., samples or populations, single or individual sets of data points versus aggregating multiple sets of data, long-term time series

forecasting versus short-term time series forecasting and cross-sectional predictability).[24] This chapter takes these observations as a point of departure to further explore some of the major differences or boundaries dividing the field of statistical prediction or forecasting through a discussion of its literature. This general overview helps to situate the emerging field of prediction with social media, which will be explored in the second half of this chapter, within the larger field of statistical prediction and to identify as well as appreciate currently evolving trends in that area (particularly the rapidly growing popularity of specific machine learning and data mining techniques).

Table 2. Summary of different predictive purposes.

(a) *Interpolative predictions*	(i) Inductive	Inference to another population, or a sample therefrom, within the range of populations or conditions already covered.
	(ii) Deductive	Inferences about random samples from the same population.
	(iii) Theoretical	Routine deductions of a well-established finding (e.g. Neptune's position at 5 a.m. tomorrow morning).
	(iv) Individual	A doctor's prognosis for a self-selected patient, based on experience.
	(v) "Nowcasting"	Predicting *y* from *x* for the data which the model has actually been fitted to (as also in saying "It's raining *now*").
(b) *Extrapolatory predictions*	(vi) Forecasting	Assertions about the future (e.g. next week's weather), when even the past patterns are still unclear.
	(vii) Discoveries	Theory pointing to something that has not yet been observed (e.g. the very existence of the planet Neptune in 1846).
	(viii) "What if?"	Varying the input assumptions, as in using a spreadsheet.

Source: Adapted from Ehrenberg and Bound (1993, 167–168; see also Ziman 1991).

Ehrenberg and Bound's typology of different predictive purposes immediately indicates there are many different competing ideas about prediction – "up to and including toothsaying" (168). In particular, it serves to introduce the notion of routine *predictability*, or "that [a prediction] will hold up routinely, within the relevant range of conditions", as distinguished

24. It is useful to note there are also discursive differences between therelated notions of prediction, forecasting, projection, and simulating. For example, forecasting typically involves information to be transferred across time, and to specific points in time (e.g., a date and time of day), while a projection depends explicitly on stated assumptions; a prediction is typically a definitive and specific statement, while a forecast is typically a probabilistic statement over larger distributions (e.g., over periods of time) (cf. Silver 2012, 149); and a simulation typically computes a model response using arbitrary input data and specified initial conditions, while a prediction would forecast its response *k* steps ahead in the future (i.e., the *prediction horizon*).

from *prediction*, or the assertion that "the result in question will hold for some other, different, data (e.g. for a sample from another population". Rather than a clear-cut distinction, these notions of prediction suggest a diverse space of possible ideas about what kind of prediction is appropriate for a given purpose. Whereas in the former case a certain degree of generalisability needs to be achieved in order for the prediction to hold in further trials, in the later case this need not be so. This explains the need for analysing many different sets of data when testing whether a hypothesis is true and to do so "preferably under varied circumstances" (Tukey 1991) – that is under different conditions, for different apparatuses, at different points in time, in different locations, and so on. For example we may want to be able to predict the dissemination of a piece of information like a tweet regardless of the specific terms the tweet may contain. Consequently, as Ehrenberg and Bound claim based on their discussion of Boyle's law in physics (1662), "[i]t had also to be established that *failures* were predictable . . . It was not merely the rule that had to be predictable, but also the exceptions" (173, emphasis in original). The *systematic exceptions*, in other words, should also be predictable *as such*; given a range of conditions under which both the phenomenon and its exceptions have been empirically established. But as John Tukey noted when paraphrasing Ronald Fisher (1926), having to establish *experience* under varied conditions through extensive repetition or *replication* of the experiment "is unhappy for the investigator who would like to settle things once and for all" (Tukey 1991, 101).

Due to this emphasis put on conditions, Ehrenberg and Bound further argue that an inability to forecast correctly "is usually not . . . because the future is too uncertain (despite the mistaken popularity of this concept nowadays), but because, much more simply, we are as yet too ignorant about the past." (187, emphasis in original). Instead of taking these conditions as given,[25] it should occur to the statistician to ask *why* these conditions are so. As a result, empirically based induction is rephrased as a problem of interpolation when they claim, "given successful replications over a known range of conditions, one can predict that the event will recur yet again within that range of conditions" (189). This implies a shift in focus from things happening every single time to the factors of control (i.e., causal relations) over the conditions of things happening.[26] In the case of social media-based prediction this raises important questions concerning the architectural complexity of the platform, its use

25. That is as *data*, or as "[s]omething given or granted; something known or assumed as fact, and made the basis of reasoning; an assumption or premise from which inferences are drawn" (*OED*, "datum, *n.*", 2.a). See also Bruno Latour's review (2005) of Isabelle Stenger's *Penser avec Whitehead: Une libre et sauvage création de concepts* (Éditions Gallimard, 2002).

26. This is known as if–then reasoning, because the validity of a statement depends on the existence of certain circumstances, or whether particular conditions have been met.

practices, and its data interface (e.g., APIs) insofar the interfaces and functionalities help establish a structured or semi-structured environment that users can act in, but not really change (Rieder et al. 2015). Moreover, "[t]here are also many relationships which have *not* become routinely predictable, or at least not yet quantitatively. . . . Here full quantification has not (yet) been achieved, partly because the scales of measurement which are used differ in ways that are not (yet?) sufficiently well understood" (186). This then begs the question of the difference between scientific laws and prediction. While in the former case causal mechanisms are understood either fully or well enough to establish routine predictability, in the latter case this need not be the case. All that matters is that the prediction will be true for a specific case or set of cases. At the time this observation was closely related to the idea that phenomena involving human behaviour cannot be as predictable as phenomena in the physical sciences (190), but in contemporary society, social media companies are in possession of very large quantities of precisely such data, which means they are in a unique position to explore these suspicions. Although it might be harder to imagine changes in physical structures compared to social structures, there certainly is much statistical regularity in both.

Two Cultures of Statistical Modelling

Although statistics may start with data (Breiman 2001), it is instructive to explore the assorted roles statistical models and modelling – to highlight the entire process involved – play across (scientific) fields of expertise for the purpose of theory building and hypothesis testing. A study conducted by Galit Shmueli (2010; see also Shmueli and Koppius 2010) is particularly interesting, describing a common conflation between explanation and prediction that obscures their differences. While it is interesting in itself to observe that explanation and prediction are apparently used interchangeably across many scientific fields,[27] it is even more interesting to clarify the distinction, and consider the practical implications in the modelling process. Quoting her observations at length, Shmueli explains,

> In many scientific fields such as economics, psychology, education, and environmental science, statistical models are used almost exclusively for causal explanation, and models that possess high

27. According to Shmueli, "[t]he conflation of explanation and prediction has its roots in philosophy of science literature, particularly the influential hypothetico-deductive model (Hempel and Oppenheim, 1948), which explicitly equated prediction and explanation. However, as later became clear, the type of uncertainty associated with explanation is of a different nature than that associated with prediction (Helmer and Rescher, 1959)" (2010, 292).

> explanatory power are often assumed to inherently possess predictive power. In fields such as natural language processing and bioinformatics, the focus is on empirical prediction with only a slight and indirect relation to causal explanation. And yet in other research fields, such as epidemiology, the emphasis on causal explanation versus empirical prediction is more mixed. Statistical modeling for description, where the purpose is to capture the data structure parsimoniously, and which is the most commonly developed within the field of statistics, is not commonly used for theory building and testing in other disciplines. (289)

The implication of such a conflation is a profound lack of understanding between building robust explanatory models versus creating powerful predictive models, and confusing explanatory power with predictive power (ibid.). According to Shmueli's analysis, explanatory modelling – "the application of statistical models to data for testing causal hypotheses about theoretical constructs" (291) – in the case of social scientific matters of investigation nearly always involves "association-based models [being] applied to observational data" (290). This is telling, because it is indicative of a different kind of "objective knowledge" that has come into being in this particular context, and not in others. For example, the statistical laws and regularities concerning society that could be justified require different criteria for what counts as evidence, and as Hacking has shown, were no longer used for description alone, but also for explaining and understanding the course of events (1990, vii). But why should there be a difference between explaining and predicting at all? For Shmueli, this is explained by the fact that measurable data do not necessarily accurately represent their underlying constructs because theories and constructs need to be made into statistical models and measurable data, which as she argues, creates a disparity "between the ability to explain phenomena at the conceptual level and the ability to generate predictions at the measurable level" (293). These disparities between explaining and predicting present themselves in at least four different ways (Table 3). This means that the question of predictive purpose is on the table once again.

Table 3. Main features of the disparity between explanatory and predictive modelling.

Causation–Association	In explanatory modeling *f** represents an underlying causal function, and *X* is assumed to cause *Y*. In predictive modeling *f* captures the association between *X* and *Y*.
Theory–Data	In explanatory modeling, *f* is carefully constructed based on *F* in a fashion that supports interpreting the estimated relationship between *X* and *Y* and testing the causal hypotheses. In predictive modeling, *f* is often constructed from the data. Direct interpretability in terms of the relationship between *X* and *Y* is not required, although sometimes transparency of *f* is desirable.
Retrospective–Prospective	Predictive modeling is forward-looking, in that *f* is constructed for predicting new observations. In contrast, explanatory modeling is retrospective, in that *f* is used to test an already existing set of hypotheses.
Bias–Variance	Bias is the result of misspecifying the statistical model *f*. Estimation variance (the third term) is the result of using a sample to estimate *f*. The first term is the error that results even if the model is correctly specified and accurately estimated. . . . In explanatory modeling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision.

Source: Adapted from Shmueli (2010, 293).

* In mathematics and statistics, "*f*" is conventionally used to represent a function, which models the relation between a set of inputs and outputs.

It is helpful to consider what such predictive purposes might look like, or how they might be different from each other. Emanuel Parzen for example writes, "[t]he two goals in analyzing data. . . I prefer to describe as 'management' and 'science.' Management seeks *profit*, practical answers (predictions) useful for decision making in the short run. Science seeks *truth*, fundamental knowledge about nature which provides understanding and control in the long run" (Breiman 2001, 224, emphasis in original). Note that this connects well with Desrosières' (2011) distinction between "metrological realism" (the first attitude) and "'accounting' realism" (the second attitude) discussed previously (see Chapter 1). Although many other possible goals might exist today (or at least their nuances are important to consider) the point that predictive purposes and their envisioned outcomes are inseparable remains essential. A difference in focus like this stemming from a different predictive purpose invariably leads to a difference between explanatory and predictive models, even if both models would be based on the same initial data. Similarly, a "wrong" model can sometimes predict better than a "good" one (293).[28] Consequently, Shmueli's own analysis

28. The implications of such a different attitude can be profound, for example when considering the how machine learning practitioners are trained to choose between a variety of available models. Instead of building a

proceeds with an extensive analysis of the entire modelling process, from goal definition, study design, and data collection to evaluation, model use, and reporting (295, Fig. 2). In the analysis four main features of the disparity between explanation and prediction are highlighted, which are reproduced in Table 3 (where "f" represents a function or model). In addition to identifying major philosophical distinctions, this overview also reveals their practical implications. For example, the overall distinction between prediction and explanation within the field of statistics arguably reflects the currently growing interest in data mining algorithms because the range of appropriate methods for predictive modelling includes not just statistical models, but also data mining algorithms since the main objective (in both practical and scientific applications) is generating accurate predictions for new observations. Instead of trying to understand the underlying causal mechanisms to explain the phenomenon, the idea is to capture complex associations that can then lead to accurate predictions (where more accuracy generally requires more complex prediction methods). This is particularly true for cases where the accuracy of the prediction can quickly be measured against reality (e.g., user feedback in email spam filtering). In this sense no *a priori* assumptions or commitments to theory need to be made other than a commitment to a method. In fact, as Leo Breiman argues, "[u]sing complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why" (2001, 208). Breiman further argued,

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. (199)

Unlike stochastic data modelling, which starts with assuming a model for the data of the phenomenon under study, algorithmic modelling, by not having to make similar commitments seems a very suitable fit for predictive and descriptive modelling, but not for explanatory modelling (cf. Shmueli 2010, 298). Considering the more recent uptake of machine learning and data mining techniques in scientific research (e.g., for image and visual object recognition or automated translation) as well as for practical applications in industry (e.g., by social media companies and electronic commerce services), this accused

---

perfect model from scratch, the approach is simply to probe the simplest most general models first that have worked on similar problems in the past, and to only then gradually add more comlicated solutions if the problem requires it.

imbalance within the field of statistics is now probably leveling.[29]

The Predictive Power of Social Media

The recent years have seen a rapid increase in the use of social media with billions of users generating massive amounts of data every day. This uptake of social media has provided industry and researchers with a new and rich source of easily accessible and cheap data about human-related events, individuals, society, and the world more generally, through capturing at a very large scale online behaviours of diverse groups of users who communicate or interact on a diversity of issues and topics (cf. Schoen et al. 2013, 529). As a result, there have been many attempts at prediction with social media within the emerging community of researchers who have utilised social media data for a wide variety of purposes such as predicting stock market movements (e.g., Bollen et al. 2011), predicting announcements of flu outbreaks (Lampos and Cristianini 2010), forecasting box-office revenues for movies (Asur and Huberman 2010), election outcomes (Gayo-Avello 2013), to name only a few. In reality, however, there are still many methodological issues to prediction using social media data that are far from being resolved, which explains the considerable attention that the ongoing search for understanding the predictive power of social media has received, especially over the last ten years (e.g., Gayo-Avello 2012). This attention comes from computer scientists, social scientists, economists, and statisticians, as well as from users, organisations, and online (social) media companies interested in exploiting this data to predict diverse phenomena with more accuracy and economic benefit (e.g., to inform automated decision making, detect temporal trends, or target users with relevant advertising). In contrast to Parzen's (2001) general distinction between administrative and scientific goals for predicting, it therefore makes sense to take a much more nuanced approach. This section develops a very brief overview of this large body of existing work, focusing on data collection and forecasting models (e.g., Schoen et al. 2013), representative areas of applications (Atefeh and Khreich 2015; Kalampokis 2013; Schoen et al. 2013; Silver 2012; Yu and Kak 2012), social media data sources (Atefeh and Khreich 2015; Kalampokis 2013; Schoen et al. 2013), and the distinction between detection and prediction tasks (Atefeh and Khreich 2015). The aim of this general overview is not be comprehensive and

29. It is also telling that the machine learning community primarily publish their work at conferences and in conference proceedings, while statistician primarily publish in journals. As a result, the publication process for the latter is typically much slower.

cover all ground,[30] but rather to consider common issues, methods and models, operational goals, tasks, limitations, and ultimately characterise different predictive purposes in the analysis (e.g., various event detection tasks, presenting users with interesting content, and so on).

Data Collection and Modelling

As Harold Schoen et al. have noted in a survey on this topic, prediction based on social media is possible (i.e., better than pure chance) only if, firstly, the prediction itself is somehow encoded within the data, secondly, the data collection maintains the encoding of the prediction outcome of interest, and third, the performed analysis of the collected data is able to reveal the prediction (2013, 530). This means that it is essential to examine the process of collecting data – and how it has been "cooked" (e.g., Bowker 2005, 184; Gitelman 2013) – as well as the process of analysis. Such data typically includes tweets, Facebook updates (e.g., status posts), groups and profiles, contents of weblogs, message boards (e.g., Reddit), reviews (e.g., in electronic markets), social multimedia (e.g., YouTube and Flickr), and Web-based search data (e.g., Google Search and Yahoo!), to name only a few (see Table A3). As such, Schoen et al. distinguish three prevailing practices of data collecting (530). Depending on the predictive purpose and the conditions of analysis, data could firstly be collected through past logs of experiences, using statistical models to make sense of them; secondly, they could also be gathered using polling methods, asking the general public directly for opinions on behavioural intentions or observed behaviours; finally, they mention social media enables various kinds of "nonreactive data collection" (Janetzko 2008)[31] facilitating non-invasive or even "sneaky" types of research that take place without the user knowing about it. They then proceed to describe three different forecasting models – prediction market models (where people speculate on the probability of an event happening or an expected value of a parameter; the more people supporting a particular outcome, the more likely it is to occur), survey models (where a representative sample of people are questioned about their intentions), and statistical models (where past regularities in a data set are detected and used to predict future developments) – which they observed across various fields. But although each type of model can be adapted to be used in the context of social media-based forecasting or prediction, there is no clear way to decide what type of model best suits the characteristics of social media data sets (as was also the case with statistical modelling in general as argued above). When deciding on an "appropriate" model,

30. A more comprehensive survey of the literature is separately provided in the Appendix.

31. An example of this type of data is transaction data, which is generated as the byproduct of a transaction.

practitioners therefore usually also take the origin of the collected empirical data into account – not least because the population of social media users may be biases and may reflect digital inequalities (e.g., Hargittai and Hinnant 2008). This includes such things as considering the specificities of the social media platform (or platforms), the methods and period of data collection, data preparation, the prediction procedure, and the calibration and testing of the model. Moreover, different characteristics of social media data also affect forecasting models and methods differently (Schoen et al. 2013, 532). Thus, social media offer great potential for discovering new subjects of successful prediction based on the characteristics of the data they provide that may serve as predictors.

Application Areas

When examining previous surveys and attempts to identify clusters in the existing literature based on the application areas studied (e.g., Atefeh and Khreich 2015; Kalampokis, Tamouris, and Tarabanis 2013; Yu and Kak 2012) several areas of application appear to be representative. As expected based on the characteristics of social media data, these areas nearly always concern human-related events like disease outbreaks, election outcomes, macroeconomics, stock market movements, product and movie box-office sales, marketing, online news and information dissemination, and collaborative software development (see Table A2). Although social media-based forecasting has been applied to other areas as well (including some natural phenomena like earthquakes and daily rainfall rates), these areas have received much less attention from the research community. But why are these particular areas of application and their associated goals recurring and not others? And why should the accuracy of the predictions be related to the application areas of these studies? Kalampokis, Tamouris, and Tarabanis firstly suggest that some application areas, such as disease outbreak and natural phenomena, involve no expression of any kind of opinion or sentiment and so the signal is only analysed for the occurrence or not of the event (2013, 553). Therefore, they argue these predictions (which rely on the detection of specific search term occurrences) are likely to be more accurate than their counterparts that do involve the extraction of opinions or sentiment from the collected data. Secondly, they also suggest that some application areas like elections or macroeconomics can be characterised as complex insofar they involve "multiple and interrelated real-world entities such as political parties and politicians or complex concepts such as consumer confidence or inflation rate" (553–554), in which case more sophisticated methods are called for. Arguably this second reason is further complicated by the desire of social groups to try to influence the outcome in ways

that models either cannot or have difficulty to anticipate (cf. Schoen et al. 2013, 535).[32] Additionally Kalampokis, Tamouris, and Tarabanis also categorise studies according to other criteria such as approaches for search term selection, text sentiment approach, evaluation approach, and based on main outcome. With regards to sentiment analysis methods they observe that the majority of studies investigate stock markets and movies, adding that application areas like elections, product sales, and macroeconomics generally do not include a sentiment-related independent variable. Perhaps an even more interesting finding is that only half of the studies they surveyed and evaluated actually qualify as using predictive analytics (as opposed to explanatory approaches) to draw conclusions about the predictive performance of social media (see Table A4).[33] The vast majority of election-related cases, for example, are in this category, while most studies related to macroeconomic indices, natural phenomena, and product sales application areas are not (553). Their findings thus suggest a similar conflation between explanation and prediction, and seem to confirm the hypothesis that machine-learning techniques are typically deployed in cases trying to achieve the latter (i.e., prediction and detection).

Detection and Prediction

As will be argued in more detail in Chapter 3, event detection from social media sources is closely related to making predictions with social media (because they involve similar operations), but they are not identical. Not only do they draw from the same kinds of techniques, they also share similar issues and operational goals. Event detection has long been addressed in the Topic Detection and Tracking (TDT) research programme, aimed at finding and following events of interest in a stream of broadcast news stories (Allan et al. 1998; Yang et al. 1998), but event detection in social media streams (e.g., online news and trending topics in a Twitter feed) poses new issues to researchers. According to Farzindar Atefeh and Wael Khreich in their comprehensive survey of representative techniques for

32. These concerns are in addition to questions concerning the characteristics of the data itself and indeed their conditions of production. For example, a recent study by Dhiraj Murthy (2015) investigates the role tweets play during elections and whether they are more reactive than predictive in the specific case of the 2012 US Republican presidential primary elections. Although they were not able to identify tweets as either reactive or predictive, they did observe tweets being used for social media campaigns (to generate "buzz"), which in turn can and often do get covered by traditional media, thereby influencing the debate.

33. The evaluation approach used for this process is based on the following two distinguishing criteria of predictive testing: "1. Was predictive accuracy based on out-of-sample assessment? . . . / 2. Was predictive accuracy assessed with adequate predictive measures . . ., or was it incorrectly inferred from explanatory power measures?" (Shmueli and Koppius 2011, 17–18).

event detection in Twitter, this is because "Twitter messages are restricted in length and written by anyone" (Atefeh and Khreich 2015, 133).[34] The more general practical implication of these differences however, is that text-mining techniques relying on more substantive and restricted documents (e.g., newspaper articles or Web pages) are not as suitable in these cases. As a result, researchers in these areas of research draw on techniques from various fields like machine-learning, natural language processing, data mining, information expression and retrieval, and text mining (134). This means there are some parallels between detection (or indeed "nowcasting") and prediction tasks, for instance the fact that both face problems concerning the credibility or character of data appearing in social media (i.e., the justification of their evidence), or the detection of unexpected or unspecified phenomena as exceptions or deviations from the rule – from the "pulse" of social media (e.g., Jungherr and Jürgens 2014; Bandari, Asur, and Huberman 2012). This makes Atefeh and Khreich's approach to categorising the surveyed literature on event detection in Twitter very applicable. For example, they categorise techniques according to the use of *document-pivot* versus *feature-pivot techniques*,[35] unspecified versus specified event detection tasks, new versus retrospective event detection, and supervised versus unsupervised approaches (or a combination of both) and consider the kinds of feature representations employed in various cases (distinguishing between such things as trending events and endogenous or nonevent trends, and between real-world events and Twitter-specific events). Interestingly, the detection research community commonly justifies its relevance by demonstrating it can reveal information about real-world or platform-specific events as they unfold (in particular for "bursty" or emerging event detection and tracking), which raises the idea of detection and prediction as two ways of staying informed and ahead of the present or pulse of social media, or at least prevent lagging behind other actors or factors that do. After all it is usually beneficial to be in a position to act sooner rather than later.

* * *

34. As they continue, "[t]herefore, tweets include large amounts of informal, irregular, and abbreviated words, large number of spelling and grammatical errors, and improper sentence structures and mixed languages. In addition, Twitter streams contain large amounts of meaningless messages (Hurlock and Wilson 2011), polluted content (Lee et al. 2011), and rumors (Castillo et al. 2011), which negatively affect the performance of the detection algorithms." (Atefeh and Khreich 2015, 133).

35. Although discussed in more detail in Chapter 3, the former refers to approaches that rely on the "raw" content of incoming documents (e.g., tweets), while the latter refers to approaches that focus on specific aspects of these documents (e.g., individual terms or characters).

This chapter has examined different aspects of what it means to predict (or what it takes to predict accurately) through a discussion of the statistical literature. In particular, some general conceptual differences were investigated between various concepts of statistical prediction and the processes of data collection and modelling, as well as in some cases pointing out their practical implications. The second part then continued with a discussion of the body of literature concerned specifically with the predictive power of social media. As I have pointed out, some of the same issues and distinctions can be observed at this level of detail, demonstrating that the field of social media-based prediction is very diverse and involves making implicit philosophical – and therefore political (Foucault [1978] 2007, 3) – commitments that express themselves as intellectual struggles, confrontations, and battles that take place within these "communities of practice" (Lave and Wenger 1991).

PART II
Techniques and Applications

CHAPTER 3
Forecasting the Pulse of Social Media Streams

Event-Based Information Organisation

The previous sections have established at least a rudimentary understanding of what it might mean to take prediction – or its "stuff" – as an object of study in a way that acknowledges its pivotal role as an integral part of larger social and cultural processes and projects. It should be clear by now that prediction is a difficult and inherently ambiguous problem not just in technical terms (i.e., not merely "tough"), but also in social or cultural terms. A general framework for analysis was therefore developed around the central concept of cultural techniques, which acknowledges the *technicity* of prediction as an integral quality of the stuff of prediction, as well as of the techniques, rationales, and practices it may give rise to. Using this framework of analysis, this chapter develops an analysis of a specific case of prediction while paying specific attention to the social and cultural significance of its concrete techniques. What are they actually doing, or what are they using their resources for? The practical goal or purpose under investigation in this chapter concerns forecasting the "pulse" (e.g., Jungherr and Jürgens 2014; Bandari, Asur, and Huberman 2012) of social media, and the many forms in which this occurs, especially with regards to identifying emerging "real-world" events or occurrences as in the case of predicting breaking news, popular content, or trending events or topics. As such it concerns the challenges associated with identifying topics or topic areas that were not previously known about or are rapidly growing in importance within a corpus of mostly textual data (Kontostathis et al. 2004). Phrasing it this way – that is as an operational goal and with this initial degree of generality – allows me to explore a space of possible variations within which certain operations and decisions are made with regards to the choice of variables as well as the choice of methods or models. Perhaps more accurately, it allows for an investigation of some of the specific resources like ideas, concepts, and methods that are mobilised to serve such a practical purpose, as well as some of their vectors of variation. This choice of focus is justified because it is emblematic (i.e., it has developed into a more general method) for a range of practices that are relevant today, such as popularity prediction and predicting popular pieces of content (e.g., Tatar et al. 2014; Bandari, Sitaram, and Huberman 2012) for "buzz

marketing", product and reputation monitoring, the detection of controversies or breaking news items for reporting, early view patterns on YouTube (e.g. Pinto et al. 2013), prediction of clicks, link discovery, and still other kinds of general and unknown event detection or prediction. These practices draw together some emerging areas of thinking and practice together (most notably the field of machine learning) to face specific challenges to which uncertainty is central, but not necessarily dangerous (e.g., as in the case of natural disasters, disease outbreaks, or terrorist attacks). As such these practices are usually more concerned with tapping into commercial opportunities or gaining some competitive edge. Using the framework and building on insights developed in the previous chapters, the analysis proceeds by exploring some of the moments of decision-making with regards to the use of social media data and the methods and models used to perform operations on them and ultimately make predictions.

These challenges have given rise to a field of experts (i.e., those with contributory or interactional expertise) dealing in different capacities with the development, application, and eveluation of methods and models for event-based information organisation. As perhaps the most emblematic application of this today, online trend and popularity prediction for products and topics on social media platforms have become an important practical goal in multiple areas of practice and work, including economics and social studies (Zhang, Wang, and Li 2013). In particular, many have studied Twitter trend prediction and its applications, investigating the problem of measuring, at macro level, information diffusion regarding some underlying topic or event (ibid.). Individuals and companies are increasingly using social media platforms to monitor, analyse, and predict using the available channels that provide a continuous flow of near real-time user-generated content provided by social media companies like Twitter, Facebook, and YouTube. This includes online behaviours like trending topics or monitoring conversations between users and products to adapt marketing strategies, offline or "real-world" events or phenomena such as predicting election results, and also the link between online data and offline phenomena. Based on their extensive survey, which was discussed previously (see Chapter 2), Farzindar Atefeh and Wael Khreich identified some of the main motivations for individuals or companies to exploit these rich data sources in the case of Twitter, which, according to the literature, is by far the most popular platform associated with these kinds of practices (this is at least in part explained by the fact that tweets and user's profiles are both publicly accessible, and the complexity of the platform's architecture is much more straightforward than others like Facebook's). For instance,

> people would be interested in getting advice, opinions, facts, or updates on news or events (Java et al. 2007; Krishnamurthy et al. 2008; Zhao and Rosson 2009). Companies are increasingly using Twitter to advertise and recommend products, brands, and services; to build and maintain reputations; to analyze users' sentiment regarding their products (or those of their competitors); to respond to customers' complaints; and to improve decision making and business intelligence (Farzindar 2012; Jansen et al. 2009; Jiang et al. 2011; Liu et al. 2012; Pak and Paroubek 2010). Twitter has also emerged as a fast communication channel for gathering and spreading breaking news (Amer-Yahia et al. 2012; Phuvipadawat and Murata 2010; Sankaranarayanan et al. 2009), for predicting election results, and for sharing political events and conversations (Small 2011; Tumasjan et al. 2010). It has also become an important analytical tool for crime prediction (Wang et al. 2012) and monitoring terrorist activities. (2015, 133).

Whether the purpose of the prediction is to predict election results, improving business intelligence, or identifying breaking news, they all share a set of common challenges, associated with the identification of "real-world" occurrences that unfold over space and time (i.e., "events") using social media data.

When events can be predicted in a reliable way, it enables users of these predictions to act upon those predicted realities and to *optimise* their business or activities for that particular reality or *state of the world* and thereby gain a competitive edge to competitors. This includes diverse managerial goals like changing the exposure of certain topics or products to certain individuals or groups (including targeted advertising, recommender systems, and political agenda-setting), altering the course or strategy of a business strategy or political campaign in response to current affairs, anticipating the number of vaccines needed for preventing a disease outbreak, scenario analysis and planning (e.g., in stress testing or asset allocation), preventive measures like emergency evacuation plans and procedures, or even quite literally trendsetting. To such and other ends, techniques have been developed, used, and evaluated to model social media data in relation to such real-world events. These techniques are used for instance to detect or identify of (large or significant) deviations from regular patterns in online data – or between the state of the system as forecasted by the model and the collected empirical data. If the empirical data diverges significantly from such a forecast, then this would be useful as an indicator or predictor (or at least a "good enough" one for its intended purpose), including for identifying particular offline phenomena. In these cases, the concrete task at hand is to identify topics or topic areas that were not previously known about and are growing rapidly in importance within a stream or corpus of collected data.

Forecasting as an Information Retrieval Concern

From a purely technical perspective, event detection is generally considered an *unsupervised learning* task,[36] which can be subdivided into two forms: *retrospective event detection* and "online" or *new event detection* (Yang et al. 1999, 34). It is worth exploring that event detection in streams of data (e.g., newswire and broadcast news channels) has long been addressed as part of the so-called Topic Detection and Tracking (TDT) programme (Allan 2002), which has arguably informed many of it its main ideas today.[37] This initiative explored a variety of automatic techniques for identifying topical connections between pieces of content (topically homogeneous regions or "stories") in an aggregated or singular collection of data (e.g. from multiple news sources) and threading them together. This is especially useful since the sheer amount of information available at the time from such a multitude of (news) sources has only increased, while no one really has the time to sift through – read, view, listen to, play with, and so on – all of this material (and even more so today). This is exemplified by news aggregating services like Google News that use comparable techniques for similar purposes. Moreover, there is a more general interest in keeping users updated about news or other kinds of events and development (e.g., think notifications). TDT foresees in this need by delegating this task to a machine able to map such data sets automatically, thereby "automatically finding story boundaries, determining what stories go with one another, and discovering when something new (unforeseen) has happened" (Wayne 1998). Additionally, decisions need to be made relating to such things as the assessment of the importance of a detected event (e.g., via a user's preferences or activity analysis), characterising an event, filtering useful from redundant information or noise, and summarising or visualising contents of a set of stories. In a report on a pilot study running from 1996–1997 (Allan et al. 1998), the three main objectives of TDT were described as "segmentation", "detection", and "tracking". The first refers to finding *topically homogeneous regions* or distinguishing disjoint sets with common elements; the second to detecting *new events*; and the third to tracking the *recurrence of known events* or detecting additional stories about a prior event (Allan et al. 1998; Wayne 1998). Moreover, another goal of the initiative was to look for "robust, accurate, fully automatic algorithms that are source, medium,

36. Unsupervised learning methods such as latent cluster analysis differ from supervised learning methods because they do not involve humans in the labelling process. Instead the labelling procedure is entirely automatic and is therefore typically used for exploratory types of data analysis.

37. The initiative's tasks and approaches were developed jointly by DARPA, the University of Massachusetts, Carnegie Mellon, and Dragon Systems supported with NSF funding (Allan, Papka, and Lavrenko 1998, 37; Wayne 1998, 2000).

domain, and language independent", which is why their efforts would be enlarged and extended to a new corpus after the pilot study.[38] It should also be noted that for the purpose of their study, the (linguistic) definition of "topic" (what is being talked about) was reduced to individual "event" (e.g., a particular airplane crash versus airplane crashes in general), which focused and simplified the problem for the researchers. In fact, a formal terminology was developed (Table 4) so as to enable practitioners and different kinds of users (e.g., commercial users, government personnel, home consumers, and so on) – each with specific requirements to identify these topics and/or events – to formalise and reason with these definitions manually, semi-automatically, or automatically through an interface. Still, however, there are problems with such definitions. While it is hard to define the notions of "topic", "event", or "activity", it is considered much simpler to describe and define parts of "event identity", or the machine-readable (detectable) properties that can be used to decide whether two events are identical or different (Allan, Papka, Lavrenko 1998, 38). Therefore, a significant part of TDT tasks becomes the problem of deciding what properties of a story can be used to detect "event identity" in the first place (ibid.).

Table 4. TDT terminology for topics, events, and activities.

Topic	A seminal event or activity, along with all directly related events and activities.
Event	Something that happens at a specific time and place (a specific election, accident, crime, or natural disaster are examples).
Activity	A connected set of actions that have a common focus or purpose (for example, specific campaigns, investigations, and disaster relief efforts).

Source: Adapted from Boykin and Merlino (2000, 36).

From the beginning, this approach to new event detection and *event tracking* has also been focused specifically on a strict "online" setting – where, in contrast to retrospective detection tasks, new events are identified from feeds in *real-time* (e.g., Allan, Papka, and Lavrenko 1998; Yang, Pierce, and Carbonell 1998; Papka 1999). Here too the system must similarly make decisions about stories on the basis of a number of sample stories "so as to determine the relationships between the stories based on the real-world events that they describe" (Allan, Papka, and Lavrenko 1998, 37). In contrast to offline settings the emphasis in online settings is on time and on the "event", rather than on the general "topic", which implies a difference in required methods for processing the "arriving text" or documents (44). What is interesting about these issues is that the task of monitoring a stream of broadcast news story, in this particular case, is productively reduced and reformulated as a

38. Referred to as the "TDT-2 text and speech corpus" (Cieri et al. 1999).

typical information retrieval problem. As such, these systems typically rely on some input by the user (e.g., a user-defined query) to specify what is "interesting", according to which a filtering system will then help the user satisfy that particular information request. To filter *relevant* content, such systems therefore rely on extensive profiling of users and their requests to improve this process of identifying relevant material in a stream or archive of new stories or documents. The key difference with new event detection tasks, however, is that no knowledge is available as to what will happen in the news – or what will qualify as "news" - at some point in the future (ibid.). To determine whether a mentioned event is indeed new, a detection algorithm might look for clues in other news sources or it might maintain a "memory" of past events. As is further noted, event tracking tasks do require an implicit query in the sample stories, "but this 'query' is given by example and is meant to capture the underlying event, slightly different from the typical [Information Retrieval] concern with 'aboutness'" (ibid.). There are, however, characteristics to news reporting in general. For instance, because stories about the same event tend to occur in "clumps", which is even more so in cases of unexpected events like natural disasters or major crimes, or because news coverage often typically mentions people, places, or things that have not been mentioned very often in the past (ibid., 43). Although there are obvious exceptions of things that do tend to reoccur, there "must be *something* about the story that makes its appearance worthwile" (43, emphasis in original). It is telling that the authors call those particular indicators "surprising", which is measured by the distance between the present occurrence of a word, and its past occurrences ("memory"). Yet measuring "surprising features" for new event detection is evaluated by the authors as insufficient by itself, because "surprising words" do not always provide a wide enough coverage to capture all relevant material, and because many of the words are useless for retrieval, for instance because of misspellings, or because they are surprising purely by chance.[39] As a result, the accuracy of the detection task benefits from establishing an environment where conditions for interaction or content creation can be controlled (e.g., by limiting post or tweet length, or writing suggestions by the system such as recognising entities or an embedded autocorrect feature helping users to get it right).

New Event Detection as a Document Classification Problem

Because in the case of new event detection tasks the event's properties are not known prior

39. This is very different for typical retrospective event detection tasks, where a feature's occurrence can simply be measured after its "'surprising' appearance" (Allan, Papka, and Lavrenko 1998, 43).

to detection, its features cannot be expressed explicitly in a query. This has therefore been called a *query-free* retrieval task (Allan et al. 1998; Atefeh and Khreich 2015), and means that decisions on new events must be made in real-time. An additional problem particularly for new event detection is that the event itself may also change over time, which implies the detection query should also change with it. As a challenge, this problem is similar to the notion of "query drift" (Allan 1996; Lam et al. 1996) in information filtering, where relevance feedback is incrementally processed to account for slow shifts in a user's interests over time, especially in cases where "documents arrive continuously" (Allan 1996, 270). To resolve this problem, adaptive or dynamic search term selection methods can be used – and are evaluated as highly successful (Allan, Papka, and Lavrenko 1998, 44) – where the query is iteratively rebuilded each time it "tracks" a new story on a given event. Thus, while event detection may be approached as an unsupervised learning task, event tracking cannot and requires adaptive *supervised learning* because of the dynamic nature of events. This means that the machine process responsible for making labeling decisions is supervised by a human coder with superior knowledge; Ideally, the system eventually "learns" the optimal value (or its approximation) for a threshold to decide autonomously whether a story should count as *significant* and if the query should be rebuilded to include features of that story.[40] The higher this threshold is set, the lower the amount of training examples that pass it successfully (i.e., it will be more selective and therefore reduce the chance of mislabeling). Arguably more important, when new event detection is approached as a document classification task, this threshold comes to encode a particular idea about what is significant, and in this case puts emphasis on the value of "surprising features" as particularly significant for detecting new events of interest to a user – that is for predicting their "surprising" appearance. As such, the measured degree of "surprise", for instance, can be used to decide whether or not an occurrence qualifies as a trending event or nonevent trend (Naaman et al. 2011). This raises the question of the interpretation of "news value" as a fluctuation of events, rather than as a steadily evolving issue over time. Techniques for unspecified event detection relying on the temporal signal of document streams (like Twitter's stream or Facebook's News Feed) must therefore (algorithmically) discriminate trends of general interest (e.g., something most people would find interesting or even *newsworthy*, or a significant real-world event) from trivial or nonevent trends that exhibit similar temporal distribution patterns (cf. Atefeh and Khreich 2015, 142). Since social media are notoriously noisy data sources (e.g., Hurlock and Wilson 2011) when it comes to the casual statuses, exchanges, and communications (mindless or pointless babble), it is

40. It is perhaps more intuitive to think of "learning" as *optimisation* towards a given objective function.

particularly important to do well on this task. However, *benchmarks* for evaluating such things can vary significantly depending on the setting and envisioned goals. In general, for instance, it may be reasonable to expect benchmarks in managerial and commercial settings to be determined on the basis of performance metrics and competitive standing in relation to other commercial entities participating in the same specific market, while scientific benchmarks may be set by the cutting edge of scientific research. In some cases, a prediction better than a coin toss may be sufficient, while in other cases a very high degree of precision is required.

Document-Pivot and Feature-Pivot Techniques

It is possible to distinguish between two kinds of techniques for event detection: those that rely on document features (e.g., words and characters) and those that rely on temporal features (e.g., activity and sentiment distributions). Whereas *document-pivot* event detection techniques rely on clustering documents based on their similarity, *feature-pivot* techniques model events in a stream of documents containing text as "bursts of activity" (Kleinberg 2002).[41] In the first case document clustering appears to be a "natural approach to event discovery" (Yang et al. 1999, 34) since each event typically involves multiple news stories that somehow relate to each other. Cluster representation models rooted in the conventional *vector-space* model – where documents are represented by a vector of weighted terms (Salton, Wong, and Yang 1975) – in combination with traditional clustering techniques developed in the field of information retrieval are typically used for such tasks because they allow users to automatically rank (by assigning relevance scores) and subsequently categorise large volumes of documents containing text, providing an efficient method to summarise these documents. Obtaining some form of "story parsing and 'understanding'" (Allan, Papka, and Lavrenko 1998, 45) is extremely useful, because it can provide insight into notions of the 4Ws questions: who, what, when, and where. In contrast to data representation techniques using vector-space models where things like term weight and keyword density play an important role,[42] *named entity vectors* (Kumaran and Allan 2004) attempts to extract

41. The term "feature" is often used by practitioners in the field of machine learning (as well as associated fields like pattern recognition and image processing) because one of the key tasks involves *feature extraction*, which refers to the creation of *derivative values* (features) from an initial set of measurements or data intended to facilitate subsequent learning tasks, generalisation steps (e.g., modelling), and in some cases human interpretation. Variables are therefore generally called features.

42. For example, the classical *Term Frequency-Inverse Document Frequency* (TF-IDF) approach initially developed by Karen Spärck Jones (1972) is often used in such cases, which infers about the distribution of occurrences over a document collection rather than merely absolute frequencies. Because of its concentration on

information patterns about an event based on the 4Ws (Mohd 2007). For example, it may recognise or detect and structure or classify strings of characters as names of people or organisations, and a string of numbers as a time of day. Something like this is often implemented using Wikipedia and other knowledge repositories and databases for *concept mining* – the extraction of concepts of events, people, phenomena, and so forth from artefacts like Web pages. Furthermore, both models have also been integrated into *mixed vector* models that combine both approaches (e.g., Yang et al. 1999; Kumaran and Allan 2004).

Although document-pivot techniques are still widely used, they are problematic for use in online social media data streams because the assumption that all documents are relevant to the detection process (Allan, Lavrenko, and Jin 2000) is violated in this case as credible information and relevant events are often buried in large piles of mainly noisy and polluted data (Atefeh and Khreich 2015; Becker et al. 2011; Castillo et al. 2011; Hurlock and Wilson 2011; Lee et al. 2011). Trend detection tasks over textual online social media data streams have therefore generated significant interest in feature-pivoting and "bursty" event detection techniques (e.g., Kleinberg 2002; Fung et al. 2005; He, Chang, and Lim 2007; He et al. 2007; Wang et al. 2007; Goorha and Ungar 2010; Nguyen et al. 2013). This includes attempts to mine "bursty" topic patterns from multiple streams (e.g., Wang et al. 2007) as well as event extraction using the expression of shared sentiment signals (i.e., detecting "bursty" sentiment structures; Nguyen et al. 2013). In other words, besides event detection in the usual sense, it may also include types of events that are maybe not usually associated with the notion of an "event", such as sudden changes in sentiment in conversations about specific commercial products, movies, stars or celebrities, electoral candidates, and so forth. In addition, patterns in user behaviours are often used to inform decisions about selection, sorting, and presentation of content and other objects in social media (e.g., targeted advertisements and their placement, content recommendations, "People You May Know" on Facebook, or suggestions for following people on Twitter). In these cases similar ideas apply, since "bursts" of user activity on a specific topic (e.g., a set of thematically related content objects) or with a particular user can also inform decisions about what the user should be presented with, or whom he or she should be interacting with according to a given system. A particular model of the ideal user is thus imagined and enacted, as well as actively reinforced through the operations of such a system.

---

words, this model disregards the termporal order of these words, and their semantic and syntactic features as part of the larger text. For event detection tasks this constitutes a limitation insofar these features can be of use in measuring similarities among a set of related or unrelated events.

Modelling Temporal Event Distributions

Different types of events can have different temporal distributions within the same stream of documents (where documents arrive continuously over time). The emergence of news stories (i.e., a specific kind of new event) is usually associated with "bursts of activity", where many occurrences take place in a short period of time. However, the identification of a new event in a document stream and the extraction of meaningful structures underlying such an event representation are two fundamentally different tasks, and therefore require a different combination of techniques. In one of his seminal papers, Jon Kleinberg (2002) has developed a formal approach for modelling such "bursts" so that they can be identified automatically, and has provided an organisational framework for analysing the underlying content of those "bursts". The underlying assumption is that "bursts of activity" signal the appearance of a particular topic in a stream of arriving documents, with certain features rising sharply in frequency as the topic emerges (ibid.). Rather than focusing on textual features of messages themselves, Kleinberg's approach demonstrates that some events are more efficiently and robustly identifiable as "a sudden confluence of message-sending over a particular period of time" (2002, 3). In a similar vein, Yiming Yang et al. (1999) have observed several patterns of temporal event distributions for news stories. First, they observed that stories discussing the same event tend to be temporally proximate, suggesting the need for a combined measure of lexical (e.g., term-vector) similarity as well as temporal proximity as a criterion for document clustering. Second, that a time gap between bursts of activity of topically similar stories usually indicates different news events (e.g., different airplane crashes or bombings), which means that cluster emergence should be monitored over time and restricted to a time window. Third, a significant shift in vocabulary and the rising frequency of the use of particular terms are typical of stories reporting on new events, suggesting that the corpus vocabulary and statistical term weights need to be dynamically updated to timely recognise and include new patterns. In other words, when previously unseen proper names and phrases occur in streams of arriving documents – that is, the "memory" of past occurrences serves to establish that there exists no relevant knowledge or prior experience about them, making it into an *exceptional* event –, this is likely to be useful for detecting or predicting the emergence of a new event. Finally, the authors observed that events are typically reported in a relatively brief window of time (or at least the interest will have decreased significantly), suggesting the need for learning methods that require only a small set of training examples to achieve *satisfactory* ("good enough") tracking performance, and which is able to exploit the temporal decrease of interest inherent to event reporting.

While event reporting for traditional news organisations is certainly not the same thing as

it is for the average user on Twitter discussing a recent occurrence on social media, they do follow similar temporal distributions and so can be treated using the same or similar methods. As a result, Kleinberg's approach to the analysis of underlying burst patterns reveals what is called a *latent hierarchical structure* "that often has a natural meaning in terms of the content of the stream" (2). This means that the data structure is argued to be both hierarchical (the temporal order of the stream can be hierarchically decomposed) and implicit (always-already there) in the underlying data stream, because the bursts form a naturally nested structure, "with a long burst of low intensity potentially containing several bursts of higher intensity inside it (and so on, recursively)" (4). As such this point is particularly interesting with regards to the empiricism of this technique: if the underlying structure is latent – or if it comes "natural" – then the technique only highlights or amplifies a structure already articulated in the data itself.

* * *

This chapter has explored the practical goal of predicting the pulse of social media, and some of the many forms in which it is encountered "in the wild" (Callon, Lascoumes, and Barthe 2009), with a particular focus towards the identification and prediction of real-world events from social media data. It has done so by investigating some of the moments of decision-making and pointing out some of the conceptual and material resources mobilised in these cases. What is particularly interesting and emblematic about the range of things brought together through this practical goal or objective is their investment in a mode of prediction that relies on event-based information organisation for the identification of new events of which there are no prior examples. Through the use of event detection techniques, this particular kind of prediction is in effect made into an information filtering problem, involving the segmentation and clustering of "events" based on their "identity" (e.g., as documents or in terms of their features), detection, and tracking new events in a stream of arriving documents or text. In different terms, the problem of prediction on a temporal scale is decomposed into a set of smaller tasks that translate parts of prediction into a spatial problem (e.g., clustering techniques or the detection of burst patterns). The detection and tracking of new events is therefore a more general aspect of forecasting in social media. On the one hand, this kind of prediction is closely associated with "nowcasting", or forecasting the present or (recent) past (see Table 2), yet on the other hand it also involves making assertions about the future insofar new and "surprising" events are involved.

CHAPTER 4
Surveilling Influenza-Like Illness Using Web Search Queries

Query-Based Syndromic Surveillance

The previous chapter has described how diverse tasks like detection, tracking, and prediction can be intimately related to each other when aligned to achieve a particular set of practical goals associated with forecasting the pulse of social media. Using the same framework, and building on some of the more general aspects of the preceding analysis, this chapter continues to investigate another perhaps more specific set of practical goals, which has to do with the control and prevention of public health scenarios, and influenza-like disease activity in particular. An emblematic example for this kind of prediction is the well-known Google Flu Trends service originally launched in November 2008 (Google.org), and the interest and discussion that it has facilitated and renewed because of its success. The goal of Google Flu Trends is to use search keyword trends from Google.com to produce a daily estimate, or "nowcast", of the occurrence of flu two weeks in advance of publication of official surveillance data.[43] In an attempt to provide faster or earlier detection – with less "reporting lag" than traditional sources used at the time that had a 1–2 week reporting lag – surveillance systems and techniques have been deployed to monitor "indirect signals" of influenza activity (Ginsberg et al. 2009). Given that "there are more flu-related searches during flu season, more allergy-related searches during allergy season, and more sunburn-related searches during the summer" (Google.org), it is possible to explore real-world phenomena based on aggregated historical logs of online Web search queries from billions of users around the world. One of the main questions raised by the initial findings from this project is then whether, or to what extent, Google search query trends can provide the basis for making "sufficiently" accurate and reliable models of real-world phenomena such as disease outbreaks.[44] The opening paragraph of an article in *The New York Times* reporting on

43. While it is less well-known, Google.org also launched a related service in June 2011 called Google Dengue Trends.

44. A term like "sufficient" in this context indicates a threshold that ties techniques to a specific explicit or implied purpose, and is further interesting because similar terms (e.g., "surprising" features and document

the new project hits the nail on the head with regards to monitoring these "indirect signals" (Ginsberg et al. 2009): "[t]here is a new common symptom of the flu, in addition to the usual aches, coughs, fevers and sore throats" (Helft 2008). The reporter is referring to the simple finding that certain search terms submitted to Google Search, like a diffuse set of symptoms, can serve as indicators or proxies for flu activity in part because Google Search is used by millions of users around the world to search for health information online (Ginsberg et al. 2009). Yet although such "indirect signals" may have proven useful for description and prediction tasks, they do not contribute to explaining these phenomena as the methods used typically rely on correlations between variables (or indicators) rather than on causal relationships.[45] This difference is also highlighted by the use of the term "symptom" in contrast to a term like "sign" since it functions as an indication of something other than itself (e.g., a phenomenon or circumstance associated with a condition, but in this case certainly not caused by it), which is also why the techniques depend on a manual search term or signal selection process.

Similar to the examples from the previous chapter, there is a specific managerial interest in the achievement of such goals as well as the efficiency (with regards to costs as well as resources), timeliness, sensitivity (e.g., degrees of precision), and accuracy (e.g., reliability, robustness) with which they are achieved. When early detection of disease activity is followed by rapid response, the impact of seasonal and pandemic influenza can be reduced (Ferguson et al. 2005; Longini 2005; Ginsberg et al. 2009). The earlier the warning, the sooner prevention and control measures may be put in place, which could then reduce or even prevent cases of influenza (Helft 2008). Similarly, the frequency with which search queries occur in particular locales may be useful to explore how much flu is circulating in different countries and regions around the world (Google.org). In addition, the reliability of these indicators or predictors also depends on the quality and quantity of the available data sources. In the original paper by Jeremy Ginsberg et al. (2009) reporting on the initial results of the project,[46] a method is presented for analysing large quantities of Google search queries for the purpose of *detecting* and *tracking* influenza-like illness in a population. The basic assumption is that the relative frequency of certain specific queries (i.e., the "symptoms") is strongly correlated with the percentage of physician visits in which a patient shows

---

"aboutness") are also used by practitioners themselves as reflected in the literature.

45. This is a common assumption often referred to with phrases like "correlation does not imply causation" or "*post hoc ergo propter hoc*".

46. The paper is a collaboration between researchers at Google Inc. and the Centers for Disease Control and Prevention.

influenza-like symptoms, and can therefore be used to estimate the present level of influenza activity in particular regions (i.e., that are sufficiently populated with users) and with a relatively small reporting lag of roughly one day (2009, 1012). Since the publication of these initial results a discussion has developed around the use of Internet search queries as well as other social media data sources for other predictive purposes, which has presumably also contributed to the rapidly growing interest in exploring possible real-world services and applications with social media data from 2008 onwards (e.g., Olson et al. 2013; Copeland et al. 2013; Choi and Varian 2012; Ciulla et al. 2012; Bollen et al. 2011; Da, Engelberg and Gao 2011; Signorini, Segre, and Polgreen 2011; Cha et al. 2010; Tumasjan et al. 2010; Wu and Brynjolfsson 2009; Polgreen et al. 2008).

"Big Data Hubris" and Algorithm Dynamics

Given that Google Flu Trends is often used as an example (because succesful) for its use of big data (McAfee and Brynjolfsson 2012; Goel et al. 2010), David Lazer et al. (2014) have recently discussed some of the common issues in flu prediction using big data in response to a news report on a Google Flu Trends predicting error (Butler 2013). According to the report, Google Flu Trends had overestimated the prevalence of flu in 2012–2013 more than double the estimates of the Centers for Disease Control and Prevention (CDC). Two particular issues are explored as mainly responsible for these mistakes: "big data hubris" and algorithm dynamics. The first mistake refers to the prevalent assumption often implicit in data analysis that big data substitute rather than supplement traditional data collection and analysis. Although there may be enormous possibilities in using big data –both *scientific* (e.g., Lazer et al. 2009; King 2011; Vespignani 2009) and *managerial* (cf. Breiman 2001, 224) – there are still foundational methodological issues pertaining to quantification and measurement, constructing validity, reliability, and dependencies among data (e.g., boyd and Crawford 2012). But while Lazer et al. are primarily invested in making this argument from a (social-)scientific point of view ("most big data . . . are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis"), this methodological critique should not mislead us to think big data cannot be useful for diverse practical services and applications such as the prevention and control of disease outbreaks for instance by containing influenza (or one of its various subtypes) at the source. For example, these big data sources enable social media companies and third parties to perform data mining, analytics and *knowledge discovery* (as it is called), and to exploit "collective intelligence" (Lévy 1997), thereby enabling an effective mobilisation of skills or efficient

allocation of resources. Given that millions of people from all over the world use Google Search to find health information, a large number of possibilities is generated for the identification of trends and the calculation of predictions based on the quantitative and qualitative characteristics of that data. Furthermore, by virtue of its reliance on correlations rather than causal relationships, the Google Flu Trends example also points to one of the inherent weaknesses of such methods. Is it possible to know whether one is observing change in the values of a model due to unwanted factors (e.g., a sudden inflation of interest in the disease after the publication of a news article), or rather a change in the actual phenomenon represented by that model (e.g., the actual spreading of the disease)? In other words, when a change is observed in the relative prevalence of search terms in a model, are we observing an actual change or a measurement error? Considering such issues – "[e]ven 3-week-old CDC data do a better job of projecting current flu prevalence than GFT" (Lazer et al. 2014, 1203) – it is therefore argued to be more useful to develop aggregate approaches using multiple sets of data. Such data sets would then preferably include other near real-time health-related data, or even data relying on qualitatively different kinds of indicators so as to be able to distinguish endogenous or nonevent trends from actual trends. However, introducing multiple sets of data would also introduce new difficulties with regards to commensuration.

The second mistake Lazer et al. identify refers to the failure to acknowledge algorithm dynamics – "changes made by engineers to improve the commercial service and by consumers in using that service" (2014, 1204) – as an important dimension of big data analysis. Because big data originating in social media claim have an *empirical grounding* (e.g., in actual online user behaviour), any inferences and predictions made using that data should pay close attention to the way the empirical is actually framed or constructed. For example, is it possible to be certain that measurements are stable and comparable across examples over time, or to account for systematic measurement errors? Which search terms can be used as indicators to detect or predict flu trends? As Lazer et al. note, "[i]n improving its service to customers, Google is also changing the data-generating process" (1204). In such cases, the malleability of computational artefacts like algorithms or even automatic search suggestions (typically generated based on a "memory" of prior related search examples from other users) is particularly important to consider when exploring the role of plastic resources. In contrast to the development of scientific instrumentation, modifications are continuously made in commercial services to further optimise ("improve") artefacts towards a set of operational goals presumably described by its business model such as providing users with useful information more quickly or increasing advertising revenues. These

modifications can adversely affect the estimates generated because the model or search algorithm and the search terms used are mutually co-constitutive of each other (i.e., they have *technicity*). As Lazer et al. argue,

> Oddly, GFT bakes in an assumption that relative search volume for certain terms is statistically related to external events, but search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider. (2014, 1204)

It seems the operative word here is "cultivated", implying the preparation and maintenance involved in empirical framing is a bit like preparing the land for growing crops and gardening. Furthermore, there is always the possibility of tactically manipulating the surveillance and monitoring systems in place, for example in the context of political campaigning or product advertising on Twitter (e.g., Mustafaraj and Metaxas 2010; Ratkiewicz et al. 2011) where there is an immediate benefit to being trending.[47] Acknowledgement of dynamic factors helps to realise that data mining techniques based on empirical data are inevitably entangled in the minutiae of everyday practices from which these data emerge. This point implies that even though techniques based on Web data proliferate today, in cases where they are used primarily to model real-world occurrences they will mostly complement traditional methods rather than replace them altogether.[48]

Disease Control and Prevention as a Machine Intelligence Problem

One of the main features that makes Google Flu Trends significant and celebrated as an example is that it demonstrates how cheap or easily accessible sources of large-scale empirical material or indeed big data can be used to introduce different (usually considered "better") degrees of precision in measurement or prediction compared to conventional measurements and benchmarks (e.g., those obtained from CDC reports) linked to the need for reliable health information. Particular degrees of precision are associated with certain levels of administrative and pragmatic transparency insofar an agreement (i.e., convention) is established or reinforced between the actors involved, the legitimacy of which is grounded in the process of measurement or quantification itself. Accepting a certain form of

47. In turn such factors also contribute to the fact that election outcomes or product sales have typically been less succesful at prediction than applications concerning disease outbreaks.

48. The complex technologies and large quantities of data involved have also made large-scale empirical analysis impracticable for scholars. Moreover, our knowledge of algorithms and data structures is largely conjectural and digital methods-based approaches (Rogers 2013) often run into trouble when (parts of) the systems under investigation change.

quantification as the basis for decision-making in public health planning can therefore also contribute to shaping the practice area or community of practice itself. For example, health plans and programmes are made to be responsive to the needs of a population enacted by the methods themselves, and decisions about budgeting, staff, and resource allocation need to made in such a way that decisions can be legitimised and rechecked if need be. Such requirements are especially important for public or semi-public institutions (e.g., medical, educational, or governmental institutions) in environments and cultures based on competing interests, since in these cases "objectivity" is achieved by following accepted rules and conventions (cf. Porter 1995). Furthermore, in place of a view where measurements and predictions are either true or false, trusted or distrusted, and so forth, the development of methods and models for prediction is a "cycle of innovation, crisis and reform, which continually expands into new regions of social and economic life, and which expresses varying degrees of commitment to precision" (Power 2004, 765). Methods are continuously improved – or *optimised* towards their specific operational goals, usually involving some kind of cost-benefit analysis for weighing multiple goals against each other – and new "'measurable' subjects" (777) emerge with these changes in method and measurement. What may appear as a general system of measurement gradually differentiates itself into something more specific and situated as "a function of an administrative and managerial proceduralism, which demands the possibility of re-checkability and, in turn, a legitimized aura of precision in the measurement process itself" (771). The application of a technique is always specific to its context, which establishes a baseline for deciding what should count as good, bad, useful, valuable, interesting, and so forth. For example, many applications of predictive analytics in business just concern gaining a competitive edge with respect to other competitors, which constitutes a very different kind of benchmark than the lagging CDC reports used in the case of Google Flu Trends. This different commitment to precision of accuracy is arguably an effect of a different application context, driven by different kinds of objectives. The various processes of measurement, surveillance and monitoring, governance and control are thus intimately linked, and the specificity of the relations between them can be studied as such.

Inductive Inference and Mechanical Reasoning

A central aspect to linking the kind of prediction at hand and its applications is the process of *inductive inference*, a mode of reasoning that aims to derive general knowledge from specific observations or examples. Inductive reasoning is strongly associated with pragmatic approaches since the objective is not absolute truth or certainty, but rather strong evidence

supporting the likelihood of something happening based on the available evidence. An inductive inference (as opposed to deduction or abduction) is likely to be true because of the present state of the world (or a subset of it) as we know it. Instead of setting out to verify a particular conclusion with certainty, the process is to calculate probabilities and assign a number reflecting likelihood to all possible conclusions or states of the world. The Google Flu Trends model uses a model that estimates the probability that a random physician visit in a particular region is related to an influenza-like illness (Ginsberg 2009, 1012).[49] These estimations are calculated on the basis of a single "explanatory variable" (i.e., a single predictor), which is the probability that a random search query submitted from the same region is related to an influenza-like illness (ibid.). Moreover, this process is delegated to an automated method that facilitates the query selection process. In addition to immediate quantitative benefits like faster and more accurate information processing and support for very large data sets, automating these operations usually also introduces qualitative changes into the process. For instance, while the reliability and legitimacy of the machine learning approach used is grounded in the assumption that the machine can learn from example inputs, the method itself requires "no previous knowledge about influenza" (ibid.). Instead a model is built that comes to stand in for the disease itself. This model is trained by processing hundreds of billions of individual search queries from five years of Google Web search logs (Ginsberg et al. 2009, 1012) to identify the queries that are best suited for accurately modelling the percentage of physician visits regarding influenza-like illnesses for each region. Based on an evaluation of model performance, the highest-scoring queries are then selected as the best candidates (initially with $n = 42$ queries per region for maximal performance at estimating out-of-sample points). As argued previously (see Chapter 2), such a learning approach implies that no *a priori* assumptions or commitments to theory need to be made (e.g., a commitment to influenza as a medical condition associated with specific symptoms and signs); instead a commitment is made to a certain specific method or model. In other words, although there might not be any sensible causal explanation as to why a submitted query like "high school basketball" is a useful predictor of influenza occurrence, what matters is that the two are very likely to correlate or coincide, thereby constituting a predictable pattern or regularity that can be used for further calculations, predictions, and making decisions that formerly required humans.

The general theory of inductive inference has famously been formalised by Ray Solomonoff (1964, 1956) as a theory of prediction grounded in the assumption that any set

49. Publicly available historical data from the CDC's (Centers for Disease Control and Prevention) U.S. Influenza Sentinal Provider Surveillance Network was used for fitting this model (Ginsberg et al. 2009, 1012).

of possibilities adheres to some unknown but computable probability distribution. Since there will be no possibility with zero probability (for then it would not be possible), the continuation of every individual possibility can be predicted on the basis of its probability in relation to all other possibilities (i.e., the *a priori* probability distribution). Given this framework, a machine could be developed and used to learn to "work" arithmetic problems given it is first exposed to a series of correct examples or problems. Subsequently it is able to "evolve a method by which they might have been solved" (1956, 1), after which it can then be presented with new examples or problems it has not been shown before (e.g., new events or search queries) and inductively infer or evaluate their "usefulness" for prediction. Such examples can be either directly useful for prediction, or useful as component to be combined with others to produce usefulness (7). The notion of *mechanical reasoning* – or the delegation of reasoning and "all the doing that is associated with it" (Hacking 1991, 239) to machines and computers (and by extension one's ability to reason with mechanical concepts and principles to solve problems) – is central to these operations and the practical goals associated with them. Besides having to formalise a problem as one that a machine is able to solve using some practical learning approach, a subtask like deciding on the "usefulness" of an example for a prediction is never straightforward. Rather than a clear-cut yes or no, these matters are ultimately *quantitative* (e.g., a set of competing values in a probability distribution). Furthermore, the process of training the machine to create a model, or instructing it on how to train itself, imbues it with purpose (i.e., its objective function, or what it is made to do best) and a particular rendering or "mediated versioning" (Rogers 2004, 163) of the world. A slight change in the value of a threshold can quickly incur significant differences in output, and thereby in decisions made on the basis of these outputs (both mechanically or by humans). Since the examples used to train the machine are empirical data rather than, say, stochastic processes, the focus is not on deriving how the available data was generated (the work of the statistician), but rather on the practical question of what the available data enables in terms of the predictive accuracy it affords (the work of the machine learning engineer; cf. Breiman 2001).

Social Media Companies and their Tangled Positions

Services like Google's Flu Trends and Dengue Trends potentially enable public health officials and health professionals to act sooner and more efficiently or effectively in the face of seasonal or pandemic epidemics than the traditional benchmarks established for instance by national and regional virological and clinical data from the U.S. CDC. However, it also introduces the possibility for other epidemics and disease outbreaks to be monitored,

controlled, or even prevented using the same data and methods. In fact, these and other big data sources can introduce new subjects of measurement just as easily as, say, advertisers can exploit big data to identify and locate specific new target demographics for their products by targeting queries and their "semantic neighbourhoods" (Rieder and Sire 2013, 4). Just consider for example that big data sources enable a higher level of granularity in regional data. In a similar vein it is also interesting to consider alternative lines of investigation that focus for instance on the use of Twitter data as opposed to Web search queries for tracking levels of disease activity (e.g., Signorini, Segre, and Polgreen 2011; Harshavardhan et al. 2011; Culotta 2010; Ritterman, Osborne, and Klein 2009). Although Twitter data is often accused of a (much) lower signal-to-noise ratio than Web search data, as well as representing a narrower demographic of primarily young adults (i.e., it is not representative of the general population), there have been attempts to reduce or correct for these biases. Besides using Twitter or alternative social media data sources, there are also projects like GrippeNet.fr and Flu Near You based on slightly different data sets. These examples demonstrate how different sets of empirical data can call for different approaches or may be better suited for other kinds of applications such as predicting election outcomes, product sales, or stock market dynamics in the case of Twitter data.[50] Furthermore, Google Flu Trends also highlights another kind of entanglement between public health planning and Internet companies, which is rather institutional. For example, in an investigation of Google's tangled position on the Web, Bernhard Rieder and Guillaume Sire (2013) argue that the particular combination of search and advertising services Google offers generate incentives to bias and conflicts of interest exacerbated by Google's dominant position in both markets. Similar critiques are relevant for other Internet and social media companies including for example Facebook, Twitter, YouTube, LinkedIn, and Instagram. Similarly it is no coincidence that companies like Facebook and Google are deeply invested (financially and otherwise) in the development and use of cutting edge techniques for instance in the fields of deep learning and artificial neural networks. There are immediate benefits for these companies to incorporating such developments into their platforms, ranging from a "cool" and innovative public image to an effective absorption of talent to the ability of piggybacking cutting-edge scientific work done in academia (including by hiring Ph.D. candidates and offering scholarships). In turn, however, such developments are also shaping the academic curriculum and faculty.

Arguably these examples can help to imagine a pragmatic approach to the question of the

50. In the case of Twitter this is related to the large amount of approaches that rely on features of sentiment for prediction.

study of concrete techniques – that is historically contingent expressions or associations – and their social and cultural significance is by way of addressing some of the many entanglements it is caught up with. How do these concrete techniques end up making a difference; how does the social status of particular kinds of expertise or experience change over time (e.g., technological expertise, including both engineering and management-of-engineering skills); and what sorts of incentives and conflicts of interest emerge from the specificity of these settings? In other words, studying concrete techniques and the rationales and practices associated with them may contribute to our understanding of the concrete power relations with which they are entangled. From this perspective a project like Google Flu Trends not only provides possibilities for surveillance of disease outbreaks and online search behaviour, but simultaneously shapes the public debate and helps to establish and legitimise the status of Google's search query data as useful at accurately predicting real world phenomena, and thereby as a valuable asset for advertising (network) services or as a starting point for decision making more broadly. For this reason too it is significant that Google Flu Trends drastically overestimated peak flu levels in January 2013 by roughly a factor of two (Butler 2013), after which Google has chosen to react by refining its algorithms. Gradually Web data mining and collective networked tracking systems are becoming part of the flu-surveillance landscape, especially for densely populated areas. Although they may not come to replace traditional systems anytime soon they do complement these systems, enabling public health planners to consider multiple sources and signal representations in the decision-making process.

* * *

This chapter has explored query-based syndromic surveillance using Web search query data for purposes of control and prevention with regards to public health scenarios and public health planning more generally. It has focused in particular on the Google Flu Trends service as a celebrated case for its success at accurately predicting both seasonal and pandemic flu (and potentially other disease activity) using large quantities of Internet data. In contrast to the previous chapter, which focused primarily on identifying the conceptual and material resources mobilised to achieve a range of perhaps more general goals, this chapter has instead concentrated more on connecting to the social and cultural significance of such "mobilising models" (Stengers 2000, 113–118) – or the circumstances and practicalities with which they are inevitably entangled. More specifically, the discussion that has developed around Google Flu Trends is used to discuss some of the many factors with which

these techniques of prediction are inevitably caught up precisely because of their concrete applications. This includes the blind faith sometimes put in the potential of big data for analysis, the implications of algorithm dynamics, empirical data and the different degrees of precision it enables, the process of inductive inference (e.g., on influenza activity), and diverse institutional entanglements. Interestingly the procedure for this kind of flu surveillance is in some ways similar to the procedure for new event detection discussed previously (see Chapter 3) because the main tasks in both cases involve the same steps: segmentation, detection, and tracking. Segmentation is needed to find topically homogeneous regions, in this case composed of indicators or predictors of influenza-like illness; detection is needed to actually identify the outbreak of disease (i.e., a specified new event detection task); and tracking capability is needed to continuously update influenza estimates. The key point is that rather than foretelling the future with absolute certainty (i.e., the scientific goal of analysing data), these techniques find practical applications (e.g., in this case for disease control and preventive purposes; an example of the managerial goal of analysing data) regardless their margin of uncertainty. Additionally there are wider implications to consider with regards to the use of correlation modelling (in contrast to for example causal models and probabilistic networks) in decision-making, uncertainty, and risk assessment, management, and governance (e.g., Fenton and Neil 2011).

Conclusion

This thesis project set out to investigate prediction and the stuff of which it is made; to form a concept or idea of its techniques in relation to the assorted roles they play in different concrete settings. It has developed a framework for conceptualising and analysing this stuff in at least some of the many ways that it exists, and in which it is imagined, accomplished, experienced, and thought through. Responding to a growing public and academic interest in the predictive power of social media, and in prediction as a way of dealing with challenges characterised by uncertainty and risk more generally, this framework enables a critical analysis of the production of prediction with a particular sensitivity towards its techniques, the conceptual and material resources they mobilise in light of a certain specific practical goal, and the social and cultural significance of their applications in diverse concrete settings. This focal point constitutes a particular nexus of historically contingent expressions or associations, which has been approached along three interrelated axes of investigation: first, the techniques of prediction (e.g., concepts, methods, and models); second, their spaces of application (e.g., concrete services and applications in the domain of social media); and third, the derivative or associated concepts and practices that are generated from these cultural techniques (e.g., new modes of calculating, predicting, categorising, sorting, ranking, and decision-making). Given that prediction and its methods or models are fully *of* the social world that they are concerned with it has first been proposed that prediction should be conceived an accomplishment, which requires a *purpose*, considerable social and intellectual investment from *sponsors or advocates*, and mobilisation of existing conceptual and material *resources* (Law, Ruppert, and Savage 2011). As these points indicate, a close analysis of the technical on the level of its concrete techniques is useful because it can be "naive" and nuanced and therefore avoids the pitfall of transposing into a critique that is simultaneously a critique of modernity or of modern civilisation at large. It is in this sense that concrete techniques and the development of particular rationales and practices are intimately linked, reflecting cultural values and beliefs about much larger questions concerning uncertainty or what a society is, or its various subdomains.

Throughout the thesis, much attention has been dedicated to the use of empirical material, not just with regards to its subject matter, but also as a strategy of analysis alongside more conceptual (and historical) investigation. What this kind of analysis provides is not necessarily an internist account or specialist reading of concrete cases of prediction or

aspects of its production, but rather a framework for conceptualising and studying techniques and their social and cultural significance in different diverse settings. In the first part and the discussion of the statistical literature and the literature on social media-based prediction, such an empirical approach enables the identification of implicit or explicit disparities and to tease out some of the implicit philosophical commitments made by practitioners situated in different communities of practice. For example, what is considered to be "appropriate" (e.g. with regards to model selection or sample sizes) within communities of practice provides an additional layer of commentary with regard to normativity. The interrelated specificities of these different groups, their attitudes to reality, their rationales of quantification, and the applications of their work thus express cultural differences. As I have argued, this concerns differences between such things as explanatory and predictive modelling (e.g., Shmueli 2010), scientific and managerial goals of analysing data (e.g., Breiman 2001), or the various possible attitudes producers and users of statistics may have towards "reality" or the empirical (Desrosières 2011). Furthermore, by stressing the *plasticity* of both conceptual and material types of resources, and how they are mobilised and utilised for diverse practical rationales and purposes, it becomes clear how quantification and the artificial more generally serve as cultural mediators between different groups and communities. In this sense, social media-based prediction marks an area of practice for kinds of thinking and areas of work that have not historically concerned themselves with prediction *per se*, thereby producing new specific arrangements of people, knowledge, and skill aligned to achieve a certain specfic goal. In the second part, this empirical sensitivity is used to closely analyse two examples of predictive purposes and to think through these examples with regards to the developed framework. The first example analysed – forecasting the pulse of social media streams – arguably constitutes a set of techniques that have already developed into something of a more general method as it is often encountered in the domain of social media. The second example – the surveillance of influenza-like illness using Web search data, as exemplified by the Google Flu Trends service – instead is interesting for the discussion and public visibility it has generated as a celebrated case of social media-based prediction. Although what is specifically needed to analyse a concrete technique will vary from case to case, it will usually involve the mobilisation of conceptual and material resources from a larger *archive* of ideas (in Foucault's sense of the term) and the specific concrete setting with which it is inevitably entangled. In addition to reading up on the technical documentation of specific techniques (e.g., in order to identify and understand the actual operations and procedures that do the work) it is often helpful to consult historical sources on these techniques in order to interpret

them more precisely. Similar to what is deemed "appropriate", judgements about what is "relevant" or "interesting", "newsworthy", "surprising" and so forth provide additional empirical material for analysing the specificity of the intentions and normative assumptions implicit in such terms and the communities that use them. Finally, what is further interesting with regards to the empirical is that the techniques studied here typically have classification schemes built into them. As one such example, the TDT's formal terminology for topics, events, and activities (see Table 4) structures the process of deciding what counts as an "event", a "topic", or an "activity", and associates particular courses of action with each of these possibilities. In other words, this is first an empirical question, but in the context of the larger system, such empirical questions become normative (as a detected pattern in a data collection) and eventually prescriptive (after the machine has been given a certain amount of examples for it to decide for itself).

Although the analysis is much too short to fully explore this large space of variation, at least two kinds of concrete techniques can be distinguished: those that are of the medium itself (they are grounded in the notion of media empiricism), and those that aim to detect, track, and/or predict "real-world" events or occurrences that do not take place online (or at least not initially). The first kind is exemplified by platform-specific mechanisms and devices such as what is trending on Twitter, Facebook's "People You May Know" and suggested LinkedIn connections, YouTube channel and video recommendations, interpretation of news value, the filtering of search query results by Google, and so forth. In the second category are predictions of election outcomes, disease outbreaks, stock market indices and dynamics, earthquakes, product sales, and so forth (see Table A2). In both cases, however, the success of predictive models hinges upon the robustness of the empirical relationships. In particular, it depends on patterns (or deviations from those patterns) detected in a collection of past data, which are then assumed to hold up in the future, either routinely or in the context of a specific prediction. What is interesting about this is that while prediction is often thought of in relation to the future, it seems that it is first and foremost about the past (which is how the present could become a predictive purpose as in "nowcasting"). Concretely, this stresses the importance of detection tasks (e.g., segmentation, detection, and tracking) as preceding the predictive tasks (e.g., inductive inference). Taking these two kinds of concrete techniques as a point of departure for further critique may mean that scientific goals for analysing data are ultimately more limited in contrast to managerial and commercial goals, since in those cases degrees of precision and accuracy ultimately rely on more rigorous benchmarks. This is the difference between the benchmark of cutting-edge science and benchmarks where accuracy is ultimately not that important as long as the predictions are "good enough" for

their intended purpose or operational goal (which is usually related either to money, danger, or threat). This highlights one of the central themes running throughout this thesis, which stresses the importance of connecting media and technology to purposeful human action, or the specific things people try to do and achieve with these media and technologies as the point of departure for a meaningful critique.

Works Cited

Achrekar, Harshavardhan et al. "Predicting Flu Trends Using Twitter Data." *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. Washington, D.C.: IEEE Computer Society, 2011. 702–707. Print.

Allan, James et al. "Topic Detection and Tracking Pilot Study Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. San Francisco: Morgan Kaufmann Publishers, 1998. 194–218. Print.

Allan, James, ed. *Topic Detection and Tracking: Event-Based Information Organization*. New York: Springer US, 2002. Print.

Allan, James, Ron Papka, and Victor Lavrenko. "On-Line New Event Detection and Tracking." *Proceedings of the 21st Annual International ACM SIGIR Conference (SIGIR '98)*. New York: ACM Press, 1998. 37–45. Print.

Allan, James, Victor Lavrenko, and Hubert Jin. "First Story Detection in TDT Is Hard." *Proceedings of the Ninth International Conference on Information and Knowledge Management*. New York: ACM Press, 2000. 374–381. Print.

Allan, James. "Incremental Relevance Feedback for Information Filtering." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 1996. 270–278. Print.

Amer-Yahia, Sihem et al. "MAQSA: A System for Social Analytics on News." *Proceedings of the 2012 ACM SIGMOD International Conference (SIGMOD '12)*. New York: ACM Press, 2012. 653–656. Print.

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media." *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '10)*. Washington, D.C.: IEEE Computer Society, 2010. 492–499. Print.

Atefeh, Farzindar, and Wael Khreich. "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence* 31.1 (2015): 132–163. Print.

Bandari, Roja. Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity." *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM '12)*. Menlo Park: AAAI Press, 2012. Print.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." 13 Feb. 2015. *California Law Review* 104 (forthcoming 2016). <http://ssrn.com/abstract=2477899>.

Beck, Ulrich. "Risk Society Revisited: Theory, Politics and Research Programmes." *The Risk Society and Beyond: Critical Issues for Social Theory*. Eds. Barbara Adam, Ulrich Beck, and Joost van Loon. London, Thousand Oaks, and New Delhi: SAGE Publications, 2000. 211–229. Print.

Becker, Hila, Mor Naaman, and Luis Gravano. "Selecting Quality Twitter Content for Events." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. Print.

Bezzi, Michela, and Marc Noppen. "It Is Tough to Make Predictions, Especially About the Future." *Respiration* 80.5 (2010): 369–371. Print.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2.1 (2011): 1–8. Print.

Bowker, Geoffrey C. *Memory Practices in the Sciences*. Cambridge, MA: The MIT Press, 2005. Print.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: The MIT Press, 2000. Print.

boyd, danah m., and Kate Crawford. "Critical Questions for Big Data." *Information, Communication & Society* 15.5 (2012): 662–679. Print.

Boykin, Stanley, and Andrew Merlino. "Machine Learning." *Communications of the ACM* 43.2 (2000): 35–41. Print.

Boyle, Robert. "A Defence Of the Doctrine touching the Spring and Weight Of the Air, Propos'd by Mr. *R. Boyle* in his New Physico-Mechanical Experiments; Against the Objections of *Franciscus Linus.* Wherewith the Objector's Finicular Hypothesis is also examin'd." *New Experiments*, 2nd ed. London: Printed by F. G. for Thomas Robinson Bookseller in Oxon, 1662. Print. <http://name.umdl.umich.edu/A28956.0001.001>.

Bozeman, Barry, and Sanjay K. Pandey. "Public Management Decision-Making: Technical vs. Political Decisions." *National Public Management Research Conference*. Washington, D.C.: PMRA, 2003. 1–34. Print.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16.3 (2001): 199–231. Print.

Burrows, Roger, and Mike Savage. "After the Crisis?: Big Data and the Methodological Challenges of Empirical Sociology." *Big Data & Society* 1.1 (2014): 1–6. Print.

Butler, Declan. "When Google Got Flu Wrong." *Nature* 494.7436 (2013): 155–156. Print.

Callon, Michel, and Fabian Muniesa. "Peripheral Vision: Economic Markets as Calculative Collective Devices." *Organization Studies* 26.8 (2005): 1229–1250. Print.

Callon, Michel, and John Law. “Introduction: Absence-Presence, Circulation, and Encountering in Complex Space.” *Environment and Planning D: Society and Space* 22.1 (2004): 3–11. Print.

Callon, Michel, and John Law. “On Qualculation, Agency, and Otherness.” *Environment and Planning D: Society and Space* 23.5 (2005): 717–733. Print.

Callon, Michel, Pierre Lascoumes, and Yannick Barthe. *Acting in an Uncertain World: An Essay on Technical Democracy*. Trans. Graham Burchell. Cambridge, MA: The MIT Press, 2009. Print. Inside Technology.

Carneiro, Herman Anthony, and Eleftherios Mylonakis. “Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks.” *Clinical Infectious Diseases* 49.10 (2009): 1557–1564. Print.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. “Information Credibility on Twitter.” *Proceedings of the 20th International Conference on World Wide Web (WWW ’11)*. New York: ACM Press, 2011. 675–684. Print.

Choi, Hyunyoung, and Hal Varian. “Predicting the Present with Google Trends.” *The Economic Record* 88 (2012): 2–9. Print.

Cieri, Chris, David Graff, and Mark Liberman. “The TDT-2 Text and Speech Corpus.” *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. San Francisco: Morgan Kaufmann Publishers, 1999. 57–60. Print.

Ciulla, Fabio et al. “Beating the News Using Social Media: The Case Study of American Idol.” *EPJ Data Science* 1.1 (2012): 8–11. Print.

Cochoy, Franck. “Calculation, Qualculation, Calqulation: Shopping Cart Arithmetic, Equipped Cognition and the Clustered Consumer.” *Marketing Theory* 8.1 (2008): 15–44. Print.

———. *Une Sociologie Du Packaging Ou L’âne De Buridan Face Au Marché*. Paris: Presses universitaires de France, 2002. Print.

Collins, Harry M., and Robert Evans. “The Third Wave of Science Studies: Studies of Expertise and Experience.” *Social Studies of Science* 32.2 (2002): 235–296. Print.

Copeland, Patrick et al. “Google Disease Trends: An Update.” *International Society of Neglected Tropical Diseases (ISNTD 2013)*. Mountain View: Google Inc., 2013. 1–3. Print.

Culotta, Aron. “Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.” *Proceedings of the First Workshop on Online Social Networks (WOSP ’08)*. New York: ACM Press, 2010. 115–122. Print.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. “In Search of Attention.” *The Journal of Finance* 66.5 (2011): 1461–1499. Print.

Day, Sophie, Celia Lury, and Nina Wakeford. "Number Ecologies: Numbers and Numbering Practices." *Distinktion: Scandinavian Journal of Social Theory* 15.2 (2014): 123–154. Print.

DeLanda, Manuel. *Philosophy and Simulation: The Emergence of Synthetic Reason*. London and New York: Continuum, 2011. Print.

Desrosières, Alain. "How Real Are Statistics?: Four Possible Attitudes." 68.2 (2011): 339–355. Print.

———. "How to Make Things Which Hold Together: Social Science, Statistics and the State." *Discourses on Society: The Shaping of the Social Science Disciplines*. Eds. Peter Wagner, Björn Wittrock, and Richard Whitley. Dordrecht: Springer Science & Business Media, 1991. 195–218. Print.

———. *The Politics of Large Numbers*. Cambridge, MA: Harvard University Press, 1998. Print.

Digital Methods Initiative. "Call for Participation: Digital Methods Summer School 2015." *Digital Methods Initiative Wiki*. Digital Methods Initiative, 17 Apr. 2015. Web. 26 June 2015. <https://wiki.digitalmethods.net/Dmi/SummerSchool2015>.

Drever, James I. "'Prediction Is Hard—Particularly About the Future'." *Elements* 7.6 (2011): 363–363. Print.

Dreyfus, Hubert L. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row, 1972. Print.

———. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: The MIT Press, 1992. Print.

Ehrenberg, Andrew S. C., and John A. Bound. "Predictability and Prediction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 156.2 (1993): 167–206. Print.

Espeland, Wendy Nelson, and Michael Sauder. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113.1 (2007): 1–40. Print.

Espeland, Wendy Nelson, and Mitchell L. Stevens. "A Sociology of Quantification." *European Journal of Sociology* 49.03 (2008): 401–436. Print.

———. "Commensuration as a Social Process." *Annual Review of Sociology* 24 (1998): 313–343. Print.

Farzindar, Atefeh. "Industrial Perspectives on Social Networks." *Proceedings of EACL 2012 – Workshop on Semantic Analysis in Social Media*, EACL, 2012. Print.

Fenton, Norman, and Martin Neil. "The Use of Bayes and Causal Modelling in Decision Making, Uncertainty and Risk." *UPGRADE – The European Journal of the Informatics Professional* 12.5 (2011): 10–21. Print.

Ferguson, Neil M. et al. “Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia.” *Nature* 437.7056 (2005): 209–214. Print.

Fisher, Ronald Aylmer. “The Arrangement of Field Experiments.” *Journal of the Ministry of Agriculture of Great Britain* 33 (1926): 503–513. Print.

Foucault, Michel. “Nietzsche, Genealogy, History.” *Language, Counter-Memory, Practice: Selected Essays and Interviews*. Ed. & Trans. Donald F. Bouchard. Ithaca: Cornell University Press, 1977. 139–164. Print.

———. *Discipline and Punish: The Birth of the Prison*. 1975. Trans. Alan M. Sheridan. 2nd ed. New York: Vintage Books, 1995. Print.

———. *Madness and Civilization: A History of Insanity in the Age of Reason*. 1961. London: Tavistock Publications, 1967. Print.

———. *Security, Territory, Population: Lectures at the Collège De France 1977-1978*. Ed. Michel Senellart, Trans. Graham Burchell. Houndmills: Palgrave Macmillan, 2007. Print.

———. *The Archeology of Knowledge and the Discourse on Language*. 1969. Trans. Alan M. Sheridan Smith. New York: Pantheon Books, 1972. Print.

———. *The Birth of the Clinic: An Archaeology of Medical Perception*. 1963. London: Tavistock Publications, 1973. Print.

———. *The History of Sexuality, Volume 1: An Introduction*. 1976. Trans. Robert Hurley. New York: Pantheon Books, 1978. Print.

———. *The Order of Things: An Archaeology of the Human Sciences*. 1966. London: Tavistock Publications, 1970. Print.

Franklin, Allan. “The Role of Experiments in the Natural Sciences: Examples From Physics and Biology.” *General Philosophy of Science: Focal Issues*. Ed. Theo A. F. Kuipers. Amsterdam: Elsevier, 2007. 219–274. Print.

Fuller, Matthew, ed. *Software Studies: A Lexicon*. Cambridge, MA: The MIT Press, 2008. Print.

Fung, Gabriel Pui Cheong et al. “Parameter Free Bursty Events Detection in Text Streams.” *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB ’05)*. New York: ACM Press, 2005. 1–12. Print.

Galison, Peter L. “Computer Simulations and the Trading Zone.” *The Disunity of Science: Boundaries, Contexts, and Power*. Eds. Peter L. Galison and David J. Stump. Stanford: Stanford University Press, 1996. 118–157. Print.

———. *How Experiments End*. Chicago and London: University of Chicago Press, 1987. Print.

Garland, David. “What Is a ‘History of the Present’?: On Foucault's Genealogies and Their

Critical Preconditions." *Punishment & Society* 16.4 (2014): 365–384. Print.

Gayo-Avello, Daniel. "'I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper': A Balanced Survey on Election Prediction Using Twitter Data." (2012): 1–13. Print.

———. "A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data." *Social Science Computer Review* 31.6 (2013): 649–679. Print.

Gerlitz, Carolin, and Celia Lury. "Social Media and Self-Evaluating Assemblages: On Numbers, Orderings and Values." *Distinktion: Scandinavian Journal of Social Theory* (2014): 1–15. Print.

Gimeno-Martínez, Javier. "Artefacts: The Artificial as Cultural Mediator." *Kunstlicht* 34.3 (2013): 4–11. Print.

Ginsberg, Jeremy et al. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457.7232 (2009): 1012–1014. Print.

Gitelman, Lisa, ed. *"Raw Data" Is an Oxymoron*. Cambridge, MA: The MIT Press, 2013. Print. Infrastructures.

Goel, Sharad et al. "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences* 107.41 (2010): 17486–17490. Print.

"Google Dengue Trends." *Google.org*. Google Inc., June 2011. Web. 26 Jun 2015. <http://www.google.org/denguetrends/>.

"Google Flu Trends." *Google.org*. Google Inc., November 2008. Web. 26 Jun 2015. <http://www.google.org/flutrends/>.

Goorha, Saurabh, and Lyle Ungar. "Discovery of Significant Emerging Trends." *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. New York: ACM Press, 2010. 57–64. Print.

Hacking, Ian. "Artificial Phenomena." *The British Journal for the History of Science* 24.2 (1991): 235–241. Print.

———. "The Archaeology of Foucault." *Historical Ontology*. London: Harper University Press, 2002. Web.

———. "The Disunities of the Sciences." *The Disunity of Science: Boundaries, Contexts, and Power*. Eds. Peter L. Galison and David J. Stump. Stanford: Stanford University Press, 1996. 37–74. Print.

———. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press, 1997. Web.

———. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press, 2006. Print.

———. *The Taming of Chance*. Cambridge: Cambridge University Press, 1990. Print. Ideas in Context.

Hargittai, Eszter, and Amanda Hinnant. "Digital Inequality: Differences in Young Adults' Use of the Internet." *Communication Research* 35.5 (2008): 602–621. Print.

He, Qi et al. "Bursty Feature Representation for Clustering Text Streams." *Proceedings of the 2007 SIAM International Conference on Data Mining*. Philadelphia: Society for Industrial and Applied Mathematics, 2007. 491–496. Print.

He, Qi, Kuiyu Chang, and Ee-Peng Lim. "Analyzing Feature Trajectories for Event Detection." *Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR '07)*. New York: ACM Press, 2007. 207–214. Print.

Heidegger, Martin. *Being and Time: A Translation of* Sein und Zeit. 1953. Trans. Joan Stambaugh. Albany: State University of New York Press, 1996. Print.

Heimer, Carol Anne. *Reactive Risk and Rational Action: Managing Moral Hazard in Insurance Contracts*. Berkeley: University of California Press, 1985. Print.

Helft, Miguel. "Google Uses Searches to Track Flu's Spread." *NYTimes.com*. The New York Times Company, 12 Nov. 2008. Web. 26 June 2015. <http://www.nytimes.com/2008/11/12/technology/internet/12flu.html>.

Helmer, Olaf, and Nicholas Rescher. "On the Epistemology of the Inexact Sciences." *Management Science* 6.1 (1959): 25–52. Print.

Hempel, Carl G., and Paul Oppenheim. "Studies in the Logic of Explanation." *Philosophy of Science* 15.2 (1948): 135–175. Print.

Hoel, Aud Sissel, and Iris van der Tuin. "The Ontological Force of Technicity: Reading Cassirer and Simondon Diffractively." *Philosophy & Technology* 26.2 (2013): 187–202. Print.

Hurlock, Jonathan, and Max L. Wilson. "Searching Twitter: Separating the Tweet From the Chaff." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*. Menlo Park: AAAI Press, 2011. 161–168. Print.

Janetzko, Dietmar. "Nonreactive Data Collection on the Internet." *The SAGE Handbook of Online Research Methods*. Eds. Nigel Fielding, Raymond M. Lee, and Grant Blank. London, Thousand Oaks, and New Delhi: SAGE Publications, 2008. 161–176. Print.

Jansen, Bernard J. et al. "Twitter Power: Tweets as Electronic Word of Mouth." *Journal of the American Society for Information Science and Technology* 60.11 (2009): 2169–2188. Print.

Java, Akshay et al. "Why We Twitter." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. New York: ACM Press, 2007. 56–65. Print.

Jiang, Long et al. "Target-Dependent Twitter Sentiment Classification." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg: Association for Computational Linguistics, 2011. 151–160. Print.

Jungherr, Andreas, and Pascal Jürgens. "Forecasting the Pulse: How Deviations from Regular Patterns in Online Data Can Identify Offline Phenomena." *Internet Research* 23.5 (2014): 589–607. Print.

Kalampokis, Evangelos, Efthimios Tambouris, and Konstantinos Tarabanis. "Understanding the Predictive Power of Social Media." *Internet Research* 23.5 (2013): 544–559. Print.

King, Gary. "Ensuring the Data-Rich Future of Social Sciences." *Science* 331.6018 (2011): 719–721. Print.

Kitchin, Rob. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1.1 (2014): 1–12. Print.

Kleinberg, Jon. "Bursty and Hierarchical Structure in Streams." *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*. New York: ACM Press, 2002. 91–25. Print.

Knight, Frank Hyneman. *Risk, Uncertainty and Profit*. Boston and New York: Houghton Mifflin Company, 1921. Print.

Knorr-Cetina, Karin, and Urs Brügger. "Inhabiting Technology: the Global Lifeform of Financial Markets." *Current Sociology* 50.3 (2002): 389–405. Print.

Kontostathis, April et al. "A Survey of Emerging Trend Detection in Textual Data Mining." *Survey of Text Mining: Clustering, Classification, and Retrieval Scanned by Velocity*. New York: Springer, 2004. 185–224. Print.

Krishnamurthy, Balachander, Phillipa Gill, and Martin Arlitt. "A Few Chirps About Twitter." *Proceedings of the First Workshop on Online Social Networks (WOSP '08)*. New York: ACM Press, 2008. 19–24. Print.

Kumaran, Giridhar, and James Allan. "Text Classification and Named Entities for New Event Detection." *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. New York: ACM Press, 2004. 297–304. Print.

Lam, Wai et al. "Detection of Shifts in User Interests for Personalized Information Filtering." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*. New York: ACM Press, 1996. 317–325. Print.

Lampos, Vasileios, and Nello Cristianini. "Tracking the Flu Pandemic by Monitoring the

Social Web." *2010 2nd International Workshop on Cognitive Information Processing (CIP)*. Washington, D.C.: IEEE Computer Society, 2010. 411–416. Print.

Latour, Bruno. "What Is Given in Experience?." *Boundary 2* 32.1 (2005): 222–237. Print.

Lave, Jean, and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press, 1991. Print.

Lave, Jean. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge: Cambridge University Press, 1988. Print.

Law, John, Evelyn Ruppert, and Mike Savage. *The Double Social Life of Methods*. Milton Keynes: CRESC Working Paper Series No. 95, 2011. Print.

Lazer, David et al. "Computational Social Science." *Science* 323.5915 (2009): 721–723. Print.

Lazer, David, Alex S. Pentland, et al. "Computational Social Science." *Science* 323.5915 (2009): 721–723. Print.

Lazer, David, Ryan Kennedy, et al. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343.6176 (2014): 1203–1205. Print.

Lee, Kyumin, Brian David Eoff, and James Caverlee. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. 185–192. Print.

Lemberg, Paul. "Why Predict the Future?." *Nonprofit World 19.3* May–June 2001: 37–38. Print.

Lévi, Pierre. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Cambrigde, MA: Perseus Books, 1997. Print.

Liu, Kun-Lin, Wu-Jun Li, and Minyi Guo. "Emoticon Smoothed Language Models for Twitter Sentiment Analysis." *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2012)*. Menlo Park: AAAI Press, 2012. Print.

Longini, Ira M., Jr. et al. "Containing Pandemic Influenza at the Source." *Science* 309.5737 (2005): 1083–1087. Print.

Lury, Celia, and Nina Wakeford, eds. *Inventive Methods: the Happening of the Social*. London and New York: Routledge, 2012. Print.

Macho, Thomas. "Tiere zweiter Ordnung: Kulturtechniken der Identität und Identifikation." *Über Kultur Theorie und Praxis der Kulturreflexion*. Eds. Dirk Baecker, Matthias Kettner, and Dirk Rustemeyer. Bielefeld: Transcript Verlag, 2008. 99–117. Print.

———. "Zeit und Zahl: Kalender- und Zeitrechnung als Kulturtechniken." *Bild – Schrift – Zahl*. Eds. Sybille Krämer and Horst Bredekamp. Munich: Fink, 2003. 179–192. Print.

Mackenzie, Adrian, and Ruth McNally. "Living Multiples: How Large-Scale Scientific Data-Mining Pursues Identity and Differences." *Theory, Culture & Society* 30.4 (2013): 72–91. Print.

Mackenzie, Adrian. "More Parts Than Elements: How Databases Multiply." *Environment and Planning D: Society and Space* 30.2 (2010): 335–350. Print.

———. "Multiplying Numbers Differently: An Epidemiology of Contagious Convolution." *Distinktion: Scandinavian Journal of Social Theory* 15.2 (2014): 189–207. Print.

———. "Set." *Inventive Methods: The Happening of the Social*. Eds. Celia Lury and Nina Wakeford. London and New York: Routledge, 2012. 219–231. Print. Culture, Economy and the Social.

Mair, Michael, Christian Greiffenhagen, and W. W. Sharrock. "Statistical Practice: Putting Society on Display." *Theory, Culture & Society* 0.0 (2015): 1–27. Web. 26 June 2015. DOI: 10.1177/0263276414559058.

———. *Social Studies of Social Science: A Working Bibliography*. Southampton: NCRM Working Paper 08/13, 2013. Print.

Marres, Noortje S., and Esther J. T. Weltevrede. "Scraping the Social?." *Journal of Cultural Economy* 6.3 (2013): 313–335. Print.

Marres, Noortje, and Carolin Gerlitz. *Interface Methods: Renegotiating Relations Between Digital Research, STS and Sociology*. London: CSISP Working Paper Nr. 3, 2014. Print.

McAfee, Andrew, and Erik Brynjolfsson. "Big Data: The Management Revolution." *Harvard Business Review* Oct. 2012: 60–69. Print.

Miller, Daniel. *A Theory of Shopping*. Cambridge: Polity Press, 1998. Print.

Mitcham, Carl. "Justifying Public Participation in Technical Decision Making." *IEEE Technology and Society Magazine* 16.1 (1997): 40–46. Print.

Mohd, Mashnizah. "Learning Approaches for Detecting and Tracking News Events." *Proceedings of the BCS IRSG Symposium: Future Directions in Information Access (FDIA 2007)*. IEEE, 2007. 28–34. Print.

Murphy, Allan H. "What Is a Good Forecast?: An Essay on the Nature of Goodness in Weather Forecasting." *Weather and Forecasting* 8.2 (1993): 281–293. Print.

Murthy, Dhiraj. "Twitter and Elections: Are Tweets, Predictive, Reactive, or a Form of Buzz?." *Information, Communication & Society* (2015): *Information, Communication & Society* 1–16. Print.

Mustafaraj, Eni, and Panagiotis Metaxas. "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search." *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Southampton: Web Science Repository, 2010. 1–8. Print.

Naaman, Mor, Hila Becker, and Luis Gravano. "Hip and Trendy: Characterizing Emerging Trends on Twitter." *Journal of the American Society for Information Science and Technology* 62.5 (2011): 902–918. Print.

O'Farrell, Clare. *Michel Foucault*. London, Thousand Oaks, and New Delhi: SAGE Publications, 2005. Print.

Olson, Donald R. et al. "Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales." Ed. Neil Ferguson. *PLoS Computational Biology* 9.10 (2013): e1003256–11. Print.

Oxford English Dictionary: The Definitive Record of the English Language. Oxford University Press, 2013. Web. 26 June 2015. <http://www.oed.com/>.

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Paris: ELRA, 2010. 19–21. Print.

Papka, Ron. "On-Line New Event Detection, Clustering, and Tracking." Dissertation, University of Massachusetts Amherst, 1999. Print.

Parzen, Emanuel. "Comment: 'Statistical Modeling: the Two Cultures'." *Statistical Science* 16.3 (2001): 224–225. Print.

Phuvipadawat, Swit, and Tsuyoshi Murata. "Breaking News Detection and Tracking in Twitter." *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '10)*. Washington, D.C.: IEEE Computer Society, 2010. 120–123. Print.

Pickering, Andrew. "Against Correspondence: A Constructivist View of Experiment and the Real." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1986* 2 (1987): 196–206. Print.

———. "Living in the Material World: On Realism and Experimental Practice." *The Uses of Experiment: Studies in the Natural Sciences*. Eds. David Gooding, Trevor Pinch, and Simon Schaffer. Cambridge: Cambridge University Press, 1989. 275–297. Print.

Pinto, Henrique, Jussara M. Almeida, and Marcos A. Gonçalves. "Using Early View Patterns to Predict the Popularity of YouTube Videos." *Proceedings of the Sixth ACM International Conference on Web search and Data Mining (WSDM '13)*. New York: ACM Press, 2013. 365–374. Print.

Polgreen, Philip M. et al. "Using Internet Searches for Influenza Surveillance." *Clinical Infectious Diseases* 47.11 (2008): 1443–1448. Print.

Porter, Theodore M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press, 1995. Print.

Power, Michael. “Counting, Control and Calculation: Reflections on Measuring and Management.” *Human Relations* 57.6 (2004): 765–783. Print.

———. “Evaluating the Audit Explosion.” *Law & Policy* 25.3 (2003): 185–202. Print.

———. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press, 1997. Print.

Ratkiewicz, Jacob et al. “Detecting and Tracking Political Abuse in Social Media.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM ’11)*. Menlo Park: AAAI Press, 2011. 297–304. Print.

Rieder, Bernhard et al. “Data Critique and Analytical Opportunities for Very Large Facebook Pages: Lessons Learned From Exploring ‘We Are All Khaled Said’.” Unpublished manuscript, University of Amsterdam, April 2015. Print.

Rieder, Bernhard, and Guillaume Sire. “Conflicts of Interest and Incentives to Bias: A Microeconomic Critique of Google's Tangled Position on the Web.” *New Media & Society* 16.2 (2013): 195–211. Print.

Rieder, Bernhard. “Probability at Work: Information Filtering as Technique.” 1 Oct. 2012. SSRN, 1 Nov. 2014. Web. 26 June 2015. <http://ssrn.com/abstract=2517272>.

Ritterman, Joshua, Miles Osborne, and Ewan Klein. “Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic.” *Proceedings of the 1st International Workshop on Mining Social Media*. 2009. Print.

Rogers, Richard A. *Digital Methods*. Cambridge, MA: The MIT Press, 2013. Print.

———. *Information Politics on the Web*. Cambridge, MA: The MIT Press, 2004. Print.

Roth, Michael S. “Foucault's ‘History of the Present’.” *History and Theory* 20.1 (1981): 32–46. Print.

Ruppert, Evelyn, John Law, and Mike Savage. “Reassembling Social Science Methods: The Challenge of Digital Devices.” 30.4 (2013): 22–46. Print.

Salton, Gerard, Andrew Wong, and Chungshu S. Yang. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18.11 (1975): 613–620. Print.

Sankaranarayanan, Jagan et al. “TwitterStand: News in Tweets.” *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS ’09)*. New York: ACM Press, 2009. 42–10. Print.

Sauder, Michael, and Wendy Nelson Espeland. “Strength in Numbers?: The Advantages of Multiple Rankings.” 81.1 (2006): 1–25. Print.

———. “The Discipline of Rankings: Tight Coupling and Organizational Change.” 74.1 (2009): 63–82. Print.

Savage, Mike, and Roger Burrows. “The Coming Crisis of Empirical Sociology.” 41.5 (2007):

885–899. Print.

Savage, Mike. *Identities and Social Change in Britain Since 1940: The Politics of Method*. Oxford: Oxford University Press, 2010. Print.

Schoen, Harald et al. "The Power of Prediction with Social Media." *Internet Research* 23.5 (2013): 528–543. Print.

Shapin, Steven, and Simon Schaffer. *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimentale Life*. Princeton: Princeton University Press, 1989. Print.

Shmueli, Galit, and Otto R. Koppius. "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35.3 (2011): 553–572. Print.

Shmueli, Galit. "To Explain or to Predict?." *Statistical Science* 25.3 (2010): 289–310. Print.

Siegert, Bernhard, and Geoffrey Winthrop-Young. "Cacography or Communication?: Cultural Techniques in German Media Studies." *Grey Room* 29 (2007): 26–47. Print.

———. "Cultural Techniques: Or the End of the Intellectual Postwar Era in German Media Theory." *Theory, Culture & Society* 30.6 (2013): 48–65. Print.

———. *Cultural Techniques: Grids, Filters, Doors, and Other Articulations of the Real*. Trans. Geoffrey Winthrop-Young. New York: Fordham University Press, 2014. Print.

Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. During the Influenza a H1N1 Pandemic." *PLoS ONE* 6.5 (2011): e19467–10. Print.

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*. New York: Penguin Books, 2012. Print.

Small, Tamara A. "What the Hashtag?." *Information, Communication & Society* 14.6 (2011): 872–895. Print.

Solomonoff, Ray J. "A Formal Theory of Inductive Inference. Part II." *Information and Control* 7.2 (1964): 224–254. Print.

———. "A Formal Theory of Inductive Inference. Part I." *Information and Control* 7.1 (1964): 1–22. Print.

———. *An Inductive Inference Machine*. New York: Technical Research Group, 1956. Print.

Spärk Jones, Karen. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 26.1 (1972): 11–21. Print.

Stengers, Isabelle. *Penser avec Whitehead: Une libre et sauvage création de concepts*. Paris: Éditions Gallimard, 2002. Print.

———. *The Invention of Modern Science*. Trans. Daniel W. Smith. Minneapolis and London: University of Minnesota Press, 2000. Print. Theory Out of Bounds.

Stevens, Stanley Smith. "On the Theory of Scales of Measurement." *Science* 103.2684

(1946): 677–680. Print.

Stigler, Stephen M. *The History of Statistics: the Measurement of Uncertainty Before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press, 1986. Print.

Strathern, Marilyn. "From Improvement to Enhancement: An Anthropological Comment on the Audit Culture." *Cambridge Anthropology* 19.3 (1997): 1–21. Print.

———. *Audit Cultures: Anthropological Studies in Accountability, Ethics, and the Academy*. London and New York: Routledge, 2000. Print.

Tater, Alexandru et al. "A survey on predicting the popularity of web content." *Journal of Internet Services and Applications* 5.1 (2014): 8–20. Print.

Taylor, Alex et al. "Modelling Biology – Working Through (In-)Stabilities and Frictions." *Computational Culture* 4 (2014): n. pag. Web. 26 June 2015. <http://computationalculture.net/article/modelling-biology>.

Taylor, Linnet, Ralph Schroeder, and Eric Meyer. "Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?." *Big Data & Society* 1.2 (2014): 1–10. Print.

Teil, Geneviève. "No Such Thing as Terroir?: Objectivities and the Regimes of Existence of Objects." *Science, Technology, & Human Values* 37.5 (2012): 478–505. Print.

Thrift, Nigel J. "Movement-Space: the Changing Domain of Thinking Resulting From the Development of New Kinds of Spatial Awareness." *Economy and Society* 33.4 (2004): 582–604. Print.

Tukey, John W. "The Philosophy of Multiple Comparisons." *Statistical Science* 6.1 (1991): 100–116. Print.

Tumasjan, Andranik et al. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*. Menlo Park: AAAI Press, 2010. 178–185. Print.

Uprichard, Emma, Roger Burrows, and David Byrne. "SPSS as an 'Inscription Device': From Causality to Description?." *The Sociological Review* 56.4 (2008): 606–622. Print.

Vespignani, Alessandro. "Predicting the Behavior of Techno-Social Systems." *Science* 325.5939 (2009): 425–428. Print.

von Goethe, Johann Wolfgang. *The Maxims and Reflections of Goethe*. 1998. Trans. Elisabeth Stopp. London: Penguin Books UK, 2005. Print.

Wang, XiaoFeng, Matthew S. Gerber, and Donald E. Brown. "Automatic Crime Prediction Using Events Extracted From Twitter Posts." *Financial Cryptography and Data Security*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2012. 231–238. Print. Lecture Notes

in Computer Science.

Wang, Xuanhui et al. "Mining Correlated Bursty Topic Patterns From Coordinated Text Streams." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. New York: ACM Press, 2007. 784–793. Print.

Wayne, Charles L. "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation." *Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000)*. Paris: ELRA, 2000. Print.

———. "Topic Detection and Tracking (TDT): Overview and Perspective." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. San Francisco: Morgan Kaufmann Publishers, 1998. Print.

Woods, Robert H. "Predicting Is Difficult, Especially About the Future: Human Resources in the New Millennium." *International Journal of Hospitality Management* 18.4 (1999): 443–456. Print.

Wu, Lynn, and Erik Brynjolfsson. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities." *Proceedings of the Thirtieth International Conference on Information Systems (ICIS 2009)*. Atlanta: AIS, 2009. 1–15. Print.

Wynne, Brian. "Uncertainty and Environmental Learning: Reconceiving Science and Policy in the Preventive Paradigm." *Global Environmental Change* 2.2 (1992): 111–127. Print.

Yang, Yiming et al. "Learning Approaches for Detecting and Tracking News Events." *IEEE Intelligent Systems* 14.4 (1999): 32–43. Print.

Yang, Yiming, Tom Pierce, and Jaime Carbonell. "A Study of Retrospective and On-line Event Detection." *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. New York: ACM Press, 1998. 28–36. Print.

Yearley, Steven, John Forrester, and Peter Bailey. "Participation and Expert Knowledge: A Case Study Analysis of Scientific Models and Their Publics." *Knowledge, Power, and Participation in Environmental Policy Analysis*. Eds. Matthijs Hisschemöller et al. New Brunswick: Transaction Publishers, 2001. 349–370. Print. Policy Studies Review Annual.

Yearley, Steven. "Making Systematic Sense of Public Discontents with Expert Knowledge: Two Analytical Approaches and a Case Study." *Public Understanding of Science* 9.2 (2000): 105–122. Print.

Yu, Sheng, and Subhash C. Kak. "A Survey of Prediction Using Social Media." *arXiv:1203.1647 [cs.SE]*. Cornell University Library, 7 Mar. 2012. Web. 26 June 2015. <http://arxiv.org/abs/1203.1647>.

Zhang, Peng, Xufei Wang, and Baoxin Li. "On Predicting Twitter Trend: Important Factors

and Models." *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York: ACM Press, 2013. 1427–1429. Print.

Zhao, Dejin, and Mary Beth Rosson. "How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work." *Proceedings of the ACM 2009 International Conference on Supporting Group Work (Group '09)*. New York: ACM Press, 2009. 243–10. Print.

Ziman, John. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. Cambridge: Cambridge University Press, 1991. Print.

APPENDIX
A Survey of Social Media-Based Prediction

Collection Methods

This appendix provides a non-exhaustive survey of literature and a bibliography on social media-based prediction in supplement to Chapter 2 in particular. Although the original purpose of this collection was merely to gain an overview of the field (in terms of clusters and trends) and to collect examples, it should prove a useful resource to anyone working on social media-based prediction. In order to collect and cover the most relevant sources, a two-step search strategy has been deployed. Literature was first collected using an expert list building strategy; and second, this aggregate expert list was than expanded with additional manual methods and search queries. In this particular context, an expert list is one that is created or maintained by a subject-matter expert (or domain expert), someone who has demonstrable experience and/or expertise in the domain of social media-based prediction. This mainly includes practitioners (scientists and professionals) from the fields of artificial intelligence and machine learning, computing and information sciences, (software) engineering, statistics, and computational social science and linguistics. Most expert lists come from published literature surveys on specific developments in the field (Table A1). After collecting and scraping expert lists for relevant sources, the collection was expanded using manual methods. These methods can be divided into two separate processes. First, additional literature was collected using the snowball method. This method was used to identify and locate earlier publications cited by present sources deemed important (not just those used for expert list building), but also to find new or more recent publications citing the source. Google Scholar (as well as Web of Science and some other major citation indexes and digital libraries) facilitates both directions of snowballing. Second, manual search and keyword queries were used to gather additional relevant sources. Finally all sources and lists have been aggregated into a single list and were organised and categorised using a reference management application (the complete list is printed below).

Table A-1. List of main sources used for expert list building.

Atefeh and Khreich 2015; Bontcheva and Rout 2014; Gayo-Avello 2013, 2012; Kalampokis, Tambouris, and Tarabanis 2013; Schoen et al. 2013; Tatar et al. 2014; Sheng and Kak 2012; Zimmerman 2014

Summaries

Tables A2–A5 offer an overview of some of the main characteristics of the literature on social media-based prediction.

Table A-2. Categorisation of literature by application area or prediction subject.

Disease Outbreak, Influenza Incidence	Achrekar, Gandhe, Lazarus et al. 2011; Althouse, Ng, and Cummings 2011; Choi and Varian 2012; Culotta 2010; Ginsberg, Mohebbi, Patel et al. 2009; Hulth, Rydevik, and Linde 2009; Lampos and Cristianini 2010; Polgreen, Chen, Pennock, and Nelson 2008; Polgreen, Nelson, Neumann, and Weinstein 2007; Ritterman, Osborne, and Klein 2009; Signorini, Segre, and Polgreen 2011; Wilson and Brownstein 2009
Earthquakes, Natural Phenomena	Comunello, Mulargia, Polidoro et al. 2015; Earle, Bowden, and Guy 2012; Lampos and Cristianini 2012; Sakaki, Okazaki, and Matsuo 2010
Elections	Adamic and Glance 2005; Asur and Huberman 2010; Berlemann and Schmidt 2001; Bermingham and Smeaton 2011; Bollen, Pepe, and Mao 2010; Castillo, Mendoza, and Poblete 2011; Ceron, Curini, lacus, and Porro 2014; Conover, Ratkiewicz, Francisco et al. 2011; DiGrazia, McKelvey, Bollen, and Rojas 2013; Filchenkov, Azarov, and Abramov 2014; Franch 2013; Gayo-Avello 2013, 2012, 2011; Gayo-Avello, Metaxas, and Mustafaraj 2011; Granka 2013; Gruzd and Roy 2014; He, Saif, Wei, and Wong 2012; Jacobsen, Potters, Schram et al. 2000; Jansen and Koop 2005; Jin, Gallagher, Cao et al. 2010; Jungherr, Jürgens, and Schoen 2012; Jürgens, Jungherr, and Schoen 2011; Lampos, Preotiuc-Pietro, and Cohn 2013; Livne, Simmons, Adar, and Adamic 2011; Lui, Metaxas, and Mustafaraj 2011; Metaxas, Mustafaraj, and Gayo-Avello 2011; Murthy 2015; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Ratkiewicz, Conover, Meiss et al. 2011; Samangooei, Hare, Dupplaw et al. 2013; Sang and Bos 2012; Shi, Agarwal, Agarwal et al. 2012; Silver 2012; Skoric, Poor, Achanuparp et al. 2011; Tumasjan, Sprenger, Sandner, and Welpe 2012, 2011, 2010; Williams and Gulati 2008
Human Mobility	Song, Qu, Blumm, and Barabási 2010
Macroeconomics	Choi and Varian 2012; De Choudhury, Gamon, Counts, and Horvitz 2013; Ettredge, Gerdes, and Karuga 2005; Fama 1991; Gilbert and Karahalios 2009; Guzmán 2011; Hossain, Wu, and Chung 2006; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Tuarob and Tucker 2013;

	Vosen and Schmidt 2012, 2011; Wang, Gerber, and Brown 2012; Wu and Brynjolfsson
Marketing	Anderson 1998; Backstrom, Huttenlocher, Kleinberg, and Lan 2006; Chaoji 2010; Chen, Wang, and Xie 2011; Domingos and Richardson 2001; Duan, Gu, and Whinston 2008; Engel, Blackwell, and Kegerreis 1969; Granovetter 1978; Gruhl, Guha, Kumar et al. 2005; Heider 1958; Jansen, Zhang, Sobel, and Chowdury 2009; Katz and Lazarsfeld 1955; Leskovec, Adamic, and Huberman 2007; Park and Lee 2009; Skowronski and Carlston 1989; Watts 2007
Online News, Information Dissemination	Asur, Huberman, Szabo, and Wang 2011; Bandari, Asur, and Huberman 2012; Bekshy, Rosenn, Marlow, and Adamic 2012; Castillo, El-Haddad, Pfeffer, and Stempeck 2014; De Choudhury, Sundaram, John, and Seligmann 2008; Gómez, Kappen, and Kaltenbrunner 2011; Lerman and Hogg 2010; LLerman and Galstyan 2008; Pinto, Almeida, and Gonçalves 2013; Richardson, Dominowska, and Ragno 2007; Szabo and Huberman 2010; Taxidou and Fischer 2013; Wu and Huberman 2007; Zhang, Wang, and Li 2013
Product Sales, Movie Box-Office, Consumer Behaviour	Ahn and Spangler 2014; Asur and Huberman 2010; Bothos, Apostolou, and Mentzas 2010; Forman, Ghose, and Wiesenfeld 2008; Ghose and Ipeirotis 2011; Goel, Hofman, Lahaie et al. 2010; Goel, Hofman, Lahaie et al. 2010; Gruhl, Guha, Kumar et al. 2005; Jin, Gallagher, Cao et al. 2010; Krauss, Nann, Simon et al. 2008; Liu, Chen, Lusch et al. 2010; Liu, Hang, An, and Yu 2007; Mishne and Glance 2006; Rui and Whinston 2011; Sharda and Delen 2006; Simonoff and Sparrow 2000; Wong, Sen, and Chiang 2012; Zhang, Luo, and Yang 2009
Software Development	Biçer, Başar Bener, and Çağlayan 2011; Wolf, Schroter, Damian, and Nguyen 2009; Bird, Pattison, D'Souza et al. 2008; Herbsleb 2007; Hinds and McGrath 2006
Stock Markets	Ali, Hammad, Samhouri, and Al-Ghandoor 2011; Antweiler and Frank 2004; Berlemann and Schmidt 2001; Bollen, Mao, and Zeng 2011; Bordino, Battiston, Calderalli et al. 2012; De Choudhury, Sundaram, John, and Seligmann 2008; Da, Engelberg, and Gao 2011; Davis, Aliaga-Díaz, and Thomas 2012; Fama 2007; Forsythe, Nelson, Neumann, and Wright 1992; Gilbert and Karahalios 2010, 2009; Hanson 2006; Hsieh and Chou 2009; Jacobsen, Potters, Schram et al. 2000; Liu, Wu, Li, and Li 2015; Oh and Sheng 2011; Tumarkin and Whitelaw 2001; Wüthrich, Permunetilleke, Leung et al. 1998; Zhang Fuehres and Gloor 2012, 2011

Table A-3. Categorisation of literature by social media data source.

Blogs	Adamic and Glance 2005; De Choudhury, Sundaram, John, and Seligmann 2008; Conover, Ratkiewicz, Francisco et al. 2011; Franch 2013; Gilbert and Karahalios 2010; Gruhl, Gaha, Kumar et al. 2005; Liu, Huang, An, and Yu 2007; Mishne and Glance 2006; Nardi, Schiano, Gumbrecht, and Swartz 2004; Oh and Sheng 2011; Ratkiewicz, Conover, Meiss et al. 2011; Stieglitz and Dang-Xuan 2011; Zhang, Chen, Chen et al. 2015
Message Boards	Bolthos, Apostolou, and Mentzas 2010; Liu, Chen, Lusch et al. 2010; Krauss, Nann, Simon et al. 2008; Antweiler and Frank 2004
Microblogs, Twitter	Abdelhaq, Sengstock, and Gertz 2013; Achrekar, Gandhe, Lazarus et al. 2011; Ahn and Spangler 2014; Altshuler, Pan, and Pentland 2012; Asur and Huberman 2010; Atefeh and Khreich 2015; Bandari, Asur, and Huberman 2012; Bauckhage, Kersting, and Rastegarpanah 2014; Bermingham and Smeaton 2011; Bollen, Mao, and Zeng 2011; Bollen, Pepe, and Mao 2010; Bothos, Apostolou, and Mentzas 2010; Castillo, Mendoza, and Poblete 2011; Ceron, Curini, lacus, and Porro 2014; Cha, Benevenuto, Ahn, and Gummadi 2012; Cha, Haddadi, Benevenuto, and Gummadi 2010; Comunello, Mulargia, Polidoro et al. 2015; Conover, Ratkiewicz, Francisco et al. 2011; Culott 2010; D'Andrea, Ducange, Lazzerini, and Marcelloni 2015; DiGrazia, McKelvey, Bollen, and Rojas 2013; Earle, Bowden, and Guy 2012; Figueiredo, Gonçalves, and Almeida 2014; Franch 2013; Gayo-Avello 2013, 2012, 2011; Gayo-Avello, Metaxas, and Mustafaraj 2011; Ghosh, Viswanath, Kooti et al. 2012; Gilbert 2012; Golbeck, Robles, Edmondson, and Turner 2011; Golbeck and Hansen 2011; Gonzáles, Muñoz, Hernández, and Cuevas 2014; Gruzd and Roy 2014; He, Saif, Wei, and Wong 2012; Huberman, Romero, and Wu 2009; Hürriyetoglu 2013; Jansen, Zhang, Finin, and Tseng 2007; Jin, Gallagher, Cao et al. 2010; Jungherr, Jürgens, and Schoen 2012; Krauss, Nann, Simon et al. 2008; Kwak, Lee, Park, and Moon 2010; Lampos and Cristianini 2012; Liangfei, Rui, and Whinston 2011; Livne, Simmons, Adar, and Adamic 2011; Lui, Metaxas and Mustafaraj 2011; Pennacchiotti and Popescu 2011; Meeyoung, Benevenuto, Haddadi, and Gummadi 2012; Metaxas, Mustafaraj, and Gayo-Avello 2011; O'Connor, Balasubramanyan, Routledge, and Smith; Oh and Sheng; Osborne and Dredze 2014; Quercia, Kosinski, Stillwell, and Crowcroft 2011; Ritterman, Osborne, and Klein 2009; Rui and Whinston 2011; Sakaki, Okazaki, and Matsuo 2013, 2010; Sang and Bos 2012; Shi, Agarwal, Agarwal et al. 2012; Signorini, Segre, and Polgreen 2011; Skoric, Poor, Achananuparp et al. 2011; Stieglitz and Dang-Xuan 2011; Syidada, Azzahra, and Puspaningrum 2014; Tops 2013; Tumasjan, Sprenger, Sandner, and Welpe 2012, 2011, 2010; Walther and Kaisser 2013; Wang, Gerber, and Brown 2012; Williams and Gulati 2008; Wong, Sen, and Chiang 2012; Zaman, Herbrich, van Gael, and Stern 2010;

	Zhang, Wang, and Li 2015, 2013; Zhang, Chen, Chen et al. 2015; Zhang, Fuehres, and Gloor 2012, 2011; Zhou and Chen 2013; Zhou, Zeng, and Wang 2014
Reviews	Ghose and Ipeirotis 2011; Bothos, Apostolou, and Mentzas 2010; Forman, Ghose, and Wiesenfeld 2008
Social Multimedia	Franch 2013; Cha, Benevenuto, Ahn, and Gummadi 2012; Jin, Gallagher, Cao et al. 2010
Web Search	Althouse, Ng, and Cummings 2011; Ballings and van den Poel 2015; Bordino, Battiston, Caldarelli et al. 2012; Bordino, Battiston, Caldarelli et al. 2012; Butler 2013; Choi and Varian 2012; Copeland, Romano, Zhang et al. 2013; Cowgill, Wolfers, and Zitzewitz 2009; Da, Engelberg, and Gao 2011; Derczynski, Yang, and Karuga 2005; Ginsberg, Mohebbi, Patel et al. 2009; Goel, Hofman, Laheie et al. 2010; Guzmán 2011; Hongzhi Yin, Bin Cui, Hua Lu et al. 2012; Huth, Rydevik, and Linde 2009; Lui, Metaxas, and Mustafaraj 2011; Olson, Konty, Paladini et al. 2013; Polgreen, Chen, Pennock, and Nelson 2008; Vosen and Schmidt 2012, 2011; Wilson and Brownstein 2009; Wu and Brynjolfsson 2009

Table A-4. Categorisation of literature by various analysis and evaluation approaches.

Aggregate Prediction (Macro-Level)	Abramowitz 1988; Antweiler and Frank 2004; Asur and Huberman 2010; Berlemann and Schmidt 2001; Bishop 2006; Brandt, Freeman, and Schrodt 2011; Campbell and Garand 2000; Choi and Varian 2012; Clements and Hendry 2011; Erikson and Wlezien 2008; Forsythe, Nelson, Neumann, and Wright 1992; Franch 2013; Gauntlett 2011; Gil and Levitt 2007; Ginsberg, Mohebbi, Patel et al. 2009; Goel, Hofman, Lahaie et al. 2010; Granka 2013; Hargittai and Hinnant 2008; Hawkins and Blakeslee 2004; Hibbs Jr. 2008; Holbrook 2008; Jacobsen, Potters, Schram et al. 2000; Jungherr and Jürgens 2014; King 1995; Lampos, Preotiuc-Pietro, and Cohn 2013; Lampos and Christianini 2012, 2010; Lewis-Beck and Rice 1992; Liangfei, Rui, and Whinston 2011; Liu 2012; Liu and Zhang 2012; Montgomery, Hollenbach, and Ward 2012; Mustafaraj, Finn, Whitlock, and Metaxas 2011; Mustafaraj and Metaxas 2010; O'Connor and Zhou 2008; Pang and Lee 2008; Pennacchiotti and Popescu 2011; Pennock 2001; Perry 1979; Polgreen, Chen, Pennock, and Nelson 2008; Polgreen, Nelson, Neumann, and Weinstein 2007; Ratkiewicz, Conover, Meiss et al. 2011; Rhode and Strumpf 2004; Shi, Agarwal, Agrawal et al. 2012; Signorini, Segre, and Polgreen 2011; Surowiecki 2005; Tumarkin and Whitelaw 2001; Wong, Sen, and Chiang 2012; Wüthrich, Permunetilleke, Leung et al. 1998
Individual Prediction (Micro-Level)	De Choudhury, Gamon, Counts, and Horvitz 2013; Golbeck, Robles, Edmondson, and Turner 2011; Quercia, Kosinski, Stillwell, and Crowcroft 2011; Golbeck and Hansen 2011; Song, Qu, Blumm, and Barabási 2010

Explanatory Evaluation Approach	Antweiler and Frank 2004; Asur and Huberman 2010; Bordino, Battiston, Caldarelli et al. 2012; Da, Engelberg, and Gao 2011; Ettredge, Gerdes, and Karuga 2005; Forman, Ghose, and Wiesenfeld 2008; Gayo-Avello 2011; He, Saif, Wei, and Wong 2012; Jin, Gallagher, Cao et al. 2010; Jungherr, Jürgens, and Scheon 2012; Krauss, Nann, Simon et al. 2008; Liu, Chen, Lusch et al. 2010; Livne, Simmons, Adar, and Adamic 2011; Metaxas, Mustafaraj, and Gayo-Avello 2011; Mishne and Glance 2006; Polgreen, Chen, Pennock, and Nelson 2008; Sang and Bos 2012; Skoric, Poor, Achananuparp et al. 2011; Tumasjan, Sprenger, Sandner, and Welpe 2011, 2010; Wilson and Brownstein 2009; Zhang, Fuehres, and Gloor 2012, 2011
Predictive Evaluation Approach	Achrekar, Gandhe, Lazarus et al. 2011; Althouse, Ng, and Cummings 2011; Bollen, Mao, and Zeng 2011; Bolthos, Apostolou, and Mentzas 2010; Choi and Varian 2012; Culotta 2010; De Choudhury, Sundaram, John, and Seligmann 2008; Franch 2013; Ghose and Ipeirotis 2011; Gilbert and Karahalios 2010; Ginsberg, Mohebbi, Patel et al. 2009; Goel, Hofman, Lahaie et al. 2010; Gruhl, Guha, Kumar et al. 2005; Guzmán 2011; Hulth, Rydevik, and Linde 2009; Lampos and Cristianini 2012; Liu, Huang, An, and Yu 2007; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Oh and Sheng 2011; Ritterman, Osborne, and Klein 2009; Rui and Whinston 2011; Sakaki, Okazaki, and Matsuo 2010; Signorini, Segre, and Polgreen 2011; Vosen and Schmidt 2011; Vosen and Schmidt 2012; Wang, Gerber, and Brown 2012; Wu and Brynjolfsson 2009

Challenging Predictive Power of Social Media	Bollen, Mao, and Zeng 2011; Forman, Ghose, and Wiesenfeld 2008; Gayo-Avello 2011; Goel, Hofman, Lahaie et al. 2010; He, Saif, Wei, and Wong 2012; Jungherr, Jürgens, and Scheon 2012; Liu, Chen, Lusch et al. 2010; Metaxas, Mustafaraj, and Gayo-Avello 2011; Mishne and Glance 2006; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Sang and Bos 2012; Skoric, Poor, Achananuparp et al. 2011; Wilson and Brownstein 2009
Supporting Predictive Power of Social Media	Achrekar, Gandhe, Lazarus et al. 2011; Althouse, Ng, Cummings 2011; Antweiler and Frank 2004; Asur and Huberman 2010; Bollen, Mao, and Zeng 2011; Bordino, Battiston, Caldarelli et al. 2012; Bothos, Apostolou, and Mentzas 2010; Butler 2013; Choi and Varian 2012; De Choudhury, Sundaram, John, and Seligmann 2008; Copeland, Romano, Zhang et al. 2013; Culotta 2010; Da, Engelberg, and Gao 2011; Ettredge, Gerdes, and Karuga 2005; Forman, Ghose, and Wiesenfeld 2008; Franch 2013; Ghose and Ipeirotis 2011; Gilbert and Karahalios 2010; Ginsberg , Mohebbi, Patel et al. 2009; Goel, Hofman, Lahaie et al. 2010; Gruhl, Guha, Kumar et al. 2005; Guzmán 2011; Hulth, Rydevik, and Linde 2009; Jin, Gallagher, Cao

	et al. 2010; Krauss, Nann, Simon et al. 2008; Lampos and Cristiannini 2012; Liu, Huang, An, and Yu 2007; Liu, Chen, Lusch et al. 2010; Livne, Simmons, Adar, and Adamic 2011; Oh and Sheng 2011; Olson, Konty, Paladini et al. 2013; Polgreen, Chen, Pennock, and Nelson 2008; Ritterman, Osborne, and Klein 2009; Rui and Whinston; Vosen and Schmidt 2012, 2011; Wang, Gerber, and Brown 2012; Wu and Brynjolfsson 2009; Zhang, Fuehres, and Gloor 2012, 2011

Dynamic Search Term Selection	Choi and Varian 2012; Culotta 2010; Ginsberg, Mohebbi, Patel et al. 2009; Goel, Hofman, Lahaie et al. 2010; Hulth, Rydevik and Linde 2009; Lampos and Cristianini 2012; Ritterman, Osborne, and Klein 2009; Sakaki, Okazaki, and Matsuo 2010; Vosen and Schmidt 2011; Wang, Gerber, and Brown 2012
Manual Search Term Selection	Achrekar, Gandhe, Lazarus et al. 2011; Althouse, Ng, Commings 2011; Asur and Huberman 2010; Bollen, Mao, and Zeng 2011; Bordino, Battiston, Caldarelli et al. 2012; De Choudhury, Syndaram, John, and Seligmann 2008; Da, Engelberg, and Gao 2011; Ettredge, Gerdes, and Karuga 2005; Franch 2013; Gayo-Avello 2011; Gayo-Avello, Metaxas, and Mustafaraj 2011; Gruhl, Guha, Kumar et al. 2005; Guzmán 2011; He, Saif, Wei, and Wong 2012; Jungherr, Jürgens, and Schoen 2012; Liu, Huang, An, and Yu 2007; Metaxas, Mustafaraj, and Gayo-Avello 2011; Mishne and Glance 2006; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Oh and Sheng2011; Polgreen, Chen, Pennock, and Nelson 2008; Rui and Whinston 2011; Sang and Bos 2012; Signorini, Segre, and Polgreen 2011; Skoric, Poor, Achananuparp et al. 2011; Tumasjan, Sprenger, Sandner, and Welpe 2010; Wilson and Brownstein 2009; Wu and Brynjolfsson 2009; Zhang, Fuehres, and Gloor 2012, 2011

Lexicon-Based Text Sentiment Analysis Approach	Bollen, Mao, and Zeng 2011; Gayo-Avello 2011; Liu, Chen, Lusch et al. 2010; Metaxas, Mustafaraj, and Gayo-Avello 2011; O'Connor, Balasubramanyan, Routledge, and Smith 2010; Zhang, Fuehres, and Gloor 2012, 2011
Machine Learning Text Sentiment Analysis Approach	Antweiler and Frank 2004; Asur and Huberman 2010; Bothos, Apostolou, and Mentzas 2010; Gayo-Avello 2011; Gilbert and Karahalios 2010; He, Saif, Wei, and Wong 2012; Krauss, Nann, Simon et al. 2008; Liu, Huang, An, and Yu 2007; Mishne and Glance 2006; Oh and Sheng 2011; Rui Whinston 2011

Sources: Reproduced from Kalampokis, Tambouris, and Tarabanis (2013, Table III-VI).

Table A-5. Categorisation of event detection literature by detection task, event type, and methods.

New Event Detection	Becker, Naaman, and Gravano 2011; Cordeiro 2012; Gu, Xie, Lv et al. 2011; Jiang, Yu, Zhou et al. 2011; Lee and Sumiya 2010; Petrović, Osborne, and Lavrenko 2010; Phuvipadawat and Murata 2010; Popescu, Pennacchiotti, and Paranjpe 2011; Popescu and Pennacchitti 2010; Sakaki, Okazaki, and Matsuo 2010; Sankaranarayanan, Samet, Teitler et al. 2009; Weng, Yao, Leonardi, and Lee 2011
Retrospective Event Detection	Becker, Chen, Iter et al. 2011; Benson, Haghighi, and Barzilay 2011; Massoudi, Tsagkias, de Rijke, and Weerkamp 2011; Metzler, Cai, Hovy 2012
"Nowcasting"	Choi and Varian 2012; Lampos and Cristianini 2012, 2010; Signorini, Segre, and Polgreen 2011

Specified Types of Events	Becker, Chen, Iter et al. 2011; Benson, Haghighi, and Barzilay 2011; Gu, Xie, Lv et al. 2011; Lee and Sumiya 2010; Massoudi, Tsagkias, de Rijke, and Weerkamp 2011; Metzler, Cai, and Hovy 2012; Popescu, Pennacchiotti, and Paranjpe 2011; Sakaki, Okazaki, and Matsuo 2010
Unspecified Types of Events	Becker, Naaman, and Gravano 2011; Cordeiro 2012; Jiang, Yu, Zhou et al. 2011; Petrović, Osborne, and Lavrenko 2010; Phuvipadawat and Murata 2010; Popescu and Pennacchiotti 2010; Sankaranarayanan, Samet, Teitler et al. 2009;Weng, Yao, Leonardi, and Lee 2011

Supervised Detection Methods	Becker, Naaman, and Gravano 2011; Becker, Chen, Iter et al. 2011; Benson, Haghighi, and Barzilay 2011; Popescu, Pennacchiotti, and Paranjpe 2011; Sakaki, Okazaki, and Matsuo 2010; Sankaranarayanan, Samet, Teitler et al. 2009
Unsupervised Detection Methods	Becker, Naaman, and Gravano 2011; Cordeiro 2012; Gu, Xie, Lv et al. 2011; Jiang, Yu, Zhou et al. 2011; Lee and Sumiya 2010; Massoudi, Tsagkias, de Rijke, and Weerkamp 2011; Metzler, Cai, and Hovy 2012; Petrović, Osborne, and Lavrenko 2010; Phuvipadawat and Murata 2010; Sankaranarayanan, Samet, Teitler et al. 2009; Weng, Yao, Leonardi, and Lee 2011

Sources: Adapted from Atefeh and Khreich (2015).

Bibliography

Abdelbary, Haasan, and Abeer El-Korany. "Semantic Topics Modeling Approach for Community Detection." *International Journal of Computer Applications* 81.6 (2013): 50–58. Print.

Abdelhaq, Hamed, Christian Sengstock, and Michael Gertz. "EvenTweet." *Proceedings of the*

VLDB Endowment 6.12 (2013): 1326–1329. Print.

Abramowitz, Alan I. "An Improved Model for Predicting Presidential Election Outcomes." *PS: Political Science & Politics* 21.4 (1988): 843–847. Print.

Achrekar, Harshavardhan et al. "Predicting Flu Trends Using Twitter Data." *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. New York: Washington, D.C.: IEEE Computer Society, 2011. 702–707. Print.

Adamic, Lada, and Natalie Glance. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *Proceedings of the 3rd international Workshop on Link Discovery (LinkKDD '05)*. New York: ACM Press, 2005. 36–43. Print.

Ahn, Hyung-Il, and W. Scott Spangler. "Sales Prediction with Social Media Analysis." *2014 Annual SRII Global Conference (SRII)*. Washington, D.C.: IEEE Computer Society, 2014. 213–222. Print.

Ali, S. M. Alhaj et al. "Modeling Stock Market Exchange Prices Using Artificial Neural Network: A Study of Amman Stock Exchange." *Jordan Journal of Mechanical and Industrial Engineering* 5.5 (2011): 439–446. Print.

Althouse, Benjamin M., Yih Yng Ng, and Derek A. T. Cummings. "Prediction of Dengue Incidence Using Search Query Surveillance." *PLoS Neglected Tropical Diseases* 5.8 (2011): e1258–7. Print.

Altshuler, Yaniv, Wei Pan, and Alex S. Pentland. "Trends Prediction Using Social Diffusion Models." *Financial Cryptography and Data Security*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2012. 97–104. Print. Lecture Notes in Computer Science.

Anderson, Eugene W. "Customer Satisfaction and Word of Mouth." *Journal of Service Research* 1.1 (1998): 5–17. Print.

Antweiler, Werner, and Murray Z. Frank. "Is All That Talk Just Noise?: The Information Content of Internet Stock Message Boards." *The Journal of Finance* 59.3 (2004): 1259–1294. Print.

Arifin, Agus Zainal et al. "Emotion Detection of Tweets in Indonesian Language Using Non-Negative Matrix Factorization." *International Journal of Intelligent Systems and Applications* 6.9 (2014): 54–61. Print.

Arrow, Kenneth J. et al. "The Promise of Prediction Markets." *Science* 320.5878 (2008): 877–878. Print.

Asur, Sitaram et al. "Trends in Social Media: Persistence and Decay." *arXiv:1102.1402 [cs.CY]*. Cornell University Library, 5 Feb. 2011. Web. 26 June 2015. <http://arxiv.org/abs/1102.1402>.

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media." *2010*

IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Washington, D.C.: IEEE Computer Society, 2010. 492–499. Print.

Atefeh, Farzindar, and Wael Khreich. "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence* 31.1 (2015): 132–163. Print.

Backstrom, Lars et al. "Group Formation in Large Social Networks: Membership, Growth, and Evolution." *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. New York: ACM Press, 2006. 1–11. Print.

Ballings, Michel, and Dirk van den Poel. "CRM in Social Media: Predicting Increases in Facebook Usage Frequency." *European Journal of Operational Research* (2015): 1–41. Print.

Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity." *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM '12)*. Menlo Park: AAAI Press, 2012. 26–33. Print.

Bauckhage, Christian, Kristian Kersting, and Bashir Rastegarpanah. "Collective Attention to Social Media Evolves According to Diffusion Models." *Proceedings of the 23rd International Conference on World Wide Web*. New York: ACM Press, 2014. 223–224. Print.

Becker, Hila et al. "Automatic Identification and Presentation of Twitter Content for Planned Events." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Menlo Park: AAAI Press, 2011. Print.

Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. 438–441. Print.

Bekshy, Eytan et al. "The Role of Social Networks in Information Diffusion." *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*. New York: ACM Press, 2012. 519–528. Print.

Benson, Edward, Aria Haghighi, and Regina Barzilay. "Event Discovery in Social Media Feeds." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Stroudsburg: Association for Computational Linguistics, 2011. 389–398. Print.

Berlemann, Michael, and Carsten Schmidt. *Predictive Accuracy of Political Stock Markets: Empirical Evidence From a European Perspective*. Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, No. 2001,57, 2001. Print. < http://nbn-resolving.de/urn:nbn:de:kobv:11-10050132 >.

Bermingham, Adam, and Alan F. Smeaton. "On Using Twitter to Monitor Political Sentiment and Predict Election Results." *Proceedings of the Sentiment Analysis where AI meets*

Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP). DORAS, 2011. 2–10. Print.

Bird, Christian et al. "Latent Social Structure in Open Source Projects." *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering (SIGSOFT '08/FSE-16)*. New York: ACM Press, 2008. 24–12. Print.

Bollen, Johan, Alberto Pepe, and Huina Mao. "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena." *Proceedings of the 19th International Conference on World Wide Web*. New York: ACM Press, 2010. Print.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2.1 (2011): 1–8. Print.

Bontcheva, Kalina, and Dominic Rout. "Making Sense of Social Media Streams Through Semantics: A Survey." *Semantic Web* 5 (2014): 373–403. Print.

Bordino, Ilaria et al. "Web Search Queries Can Predict Stock Market Volumes." *PLoS ONE* 7.7 (2012): e40014–17. Print.

Bothos, Efthimios, Dimitris Apostolou, and Gregoris Mentzas. "Using Social Media to Predict Future Events with Agent-Based Markets." *IEEE Intelligent Systems* 25.6 (2010): 50–58. Print.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict." *Conflict Management and Peace Science* 28.1 (2011): 41–64. Print.

Burt, Ronald S. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press, 2009. Print.

Butler, Declan. "When Google Got Flu Wrong." *Nature* 494.7436 (2013): 155–156. Print.

Campbell, James E. "The Trial-Heat Forecast of the 2008 Presidential Vote: Performance and Value Considerations in an Open-Seat Election." *PS: Political Science & Politics* 41.04 (2008): 1–5. Print.

Campbell, James E., and James C. Garand. *Before the Vote: Forecasting American National Elections*. Thousand Oaks: Sage Publications, 2000. Print.

Castillo, Carlos et al. "Characterizing the Life Cycle of Online News Stories Using Social Media Reactions." *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. New York: ACM Press, 2014. 211–223. Print.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Information Credibility on Twitter." *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. New York: ACM Press, 2011. 675–684. Print.

———. “Predicting Information Credibility in Time-Sensitive Social Media.” *Internet Research* 23.5 (2013): 560–588. Print.

Cavalin, Paulo. “Towards Personalized Offers by Means of Life Event Detection on Social Media and Entity Matching.” *Late-breaking Results, Doctoral Consortium and Workshop Proceedings of the 25th ACM Hypertext and Social Media Conference (Hypertext 2014): Social Personalisation Workshop (SP 2014)*. CEUR Workshop Proceedings, 2014. Print.

Ceron, Andrea et al. “Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France.” *New Media & Society* 16.2 (2014): 340–358. Print.

Ceron, Andrea, Luigi Curini, and Stefano M. Iacus. “Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy.” *Social Science Computer Review* 33.1 (2014): 3–20. Print.

Cha, Meeyoung, Fabricio Benevenuto, et al. “Delayed Information Cascades in Flickr: Measurement, Analysis, and Modeling.” *Computer Networks* 56.3 (2012): 1066–1076. Print.

Chaoji, Vineet Shashikant. “Predicting Product Adoption in Large-Scale Social Networks.” *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM ’10)*. New York: ACM Press, 2010. 1039-1048. Print.

Chen, Feng, and Daniel B. Neill. “Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs.” *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’14)*. New York: ACM Press, 2014. 1166–1175. Print.

Chen, Po-Ta, Feng Chen, and Zhen Qian. “Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields.” *Proceedings of the IEEE International Conference on Data Mining (ICDM 2014)*. Washington, D.C.: IEEE Computer Society, 2014. 80–89. Print.

Chen, Yubo, Qi Wang, and Jinhong Xie. “Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning.” *Journal of Marketing Research* 48.2 (2011): 238–254. Print.

Choi, Hyunyoung, and Hal Varian. “Predicting the Present with Google Trends.” *The Economic Record* 88 (2012): 2–9. Print.

Chua, Freddy Chong Tat, and Sitaram Asur. “Automatic Summarization of Events From Social Media.” *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM ’13)*. Menlo Park: AAAI Press, 2013. Print.

Clements, Michael P., and David F. Hendry. *The Oxford Handbook of Economic Forecasting*.

Oxford: Oxford University Press, 2011. Print.

Comunello, Francesca et al. “No Misunderstandings During Earthquakes: Elaborating and Testing a Standardized Tweet Structure for Automatic Earthquake Detection Information.” *Proceedings of the 12th International Conference on Information Systems on Crisis Response and Management (ISCRAM 2015)*. 2015. Print.

Conover, M. D. et al. “Political Polarization on Twitter.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM ’11)*. Menlo Park: AAAI Press, 2011. 89–96. Print.

Copeland, Patrick et al. “Google Disease Trends: An Update.” *International Society of Neglected Tropical Diseases 2013*. London: ISNTD, 1–3. Print.

Cordeiro, Mário. “Twitter Event Detection: Combining Wavelet Analysis and Topic Inference Summarization.” *Proceedings of the Fifth Doctoral Symposium on Informatics Engineering (DSIE ’12)*. Porto: Faculdade de Engenharia da Universidade do Porto, 2012. Print.

Cowgill, Bo, Justin Wolfers, and Eric Zitzewitz. “Using Prediction Markets to Track Information Flows: Evidence From Google.” *Auctions, Market Mechanisms and Their Applications*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2009. 3–3. Print. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering.

Culotta, Aron. “Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.” *Proceedings of the First Workshop on Social Media Analytics (SOMA ’10)*. New York: ACM Press, 2010. 115–122. Print.

D'Andrea, Eleonora et al. “Real-Time Detection of Traffic From Twitter Stream Analysis.” *IEEE Transactions on Intelligent Transportation Systems* PP.99 (2015): 1–15. Print.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. “In Search of Attention.” *The Journal of Finance* 66.5 (2011): 1461–1499. Print.

Davis, Joseph, Roger Aliaga-Díaz, and Charles J. Thomas. *Forecasting Stock Returns: What Signals Matter, and What Do They Say Now?*. Malvern: The Vanguard Group, 2012. Print.

De Choudhury, Munmun et al. “Can Blog Communication Dynamics Be Correlated with Stock Market Activity?.” *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia (HT ’08)*. New York: ACM Press, 2008. 55–59. Print.

———. “Predicting Depression via Social Media.” *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM ’13)*. Menlo Park: AAAI Press, 2013. Print.

Derczynski, Leon R. A., Bin Yang, and Christian S. Jensen. “Towards Context-Aware Search and Analysis on Social Media Data.” *Proceedings of the 16th International Conference on*

Extending Database Technology (EDBT '13). New York: ACM Press, 2013. 137–142. Print.

Devyatkin, D. A. et al. "Intelligent Analysis of Manifestations of Verbal Aggressiveness in Network Community Texts." *Scientific and Technical Information Processing* 41.6 (2015): 377–389. Print.

DiGrazia, Joseph et al. "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior." *PLoS ONE* 8.11 (2013): e79449–5. Print.

Domingos, Pedro, and Matt Richardson. "Mining the Network Value of Customers." *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*. New York: ACM Press, 2001. 57–66. Print.

Dong, Xiaowen et al. "Multiscale Event Detection in Social Media." *arXiv:1404.7048 [cs.SI]*. Cornell University Library, 6 Feb. 2015. Web. 26 June 2015. <http://arxiv.org/abs/1404.7048>.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston. "Do Online Reviews Matter?: An Empirical Investigation of Panel Data." *Decision Support Systems* 45.4 (2008): 1007–1016. Print.

Earle, Paul S., Daniel C. Bowden, and Michelle Guy. "Twitter Earthquake Detection: Earthquake Monitoring in a Social World." *Annals of Geophysics* 54.6 (2012): 708–715. Print.

Engel, James F., Roger D. Blackwell, and Robert J. Kegerreis. "How Information Is Used to Adopt an Innovation." *Journal of Advertising Research* 9.4 (1969): 3–8. Print.

Erikson, Robert S., and Christopher Wlezien. "Are Political Markets Really Superior to Polls as Election Predictors?." *Public Opinion Quarterly* 72.2 (2008): 190–215. Print.

———. "Leading Economic Indicators, the Polls, and the Presidential Vote." *PS: Political Science & Politics* 41.04 (2008): 1–5. Print.

Ettredge, Michael, John Gerdes, and Gilbert Karuga. "Using Web-Based Search Data to Predict Macroeconomic Statistics." *Communications of the ACM* 48.11 (2005): 87–92. Print.

Fama, Eugene F. "Efficient Capital Markets: II." *The Journal of Finance* 46.5 (1991): 1575–1617. Print.

———. "The Behavior of Stock-Market Prices." *The Journal of Business* 38.1 (2007): 34–105. Print.

Figueiredo, Flavio, Marcos Gonçalves, and Jussara M. Almeida. "Improving the Effectiveness of Content Popularity Prediction Methods Using Time Series Trends." *arXiv:1408.7094 [cs.SI]*. Cornell University Library, 29 Aug. 2014. Web. 26 June 2015. <http://arxiv.org/abs/1408.7094>.

Filchenkov, Andrey A., Artur A. Azarov, and Maxim V. Abramov. "What Is More Predictable in Social Media: Election Outcome or Protest Action?." *Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia (EGOSE '14)*. New York: ACM Press, 2014. 157–161. Print.

Forman, Chris, Anindya Ghose, and Batia Wiesenfeld. "Examining the Relationship Between Reviews and Sales: The Role of Reviewer." *Information Systems Research* 19.3 (2008): 291–313. Print.

Forsythe, Robert et al. "Anatomy of an Experimental Political Stock Market." *The American Economic Review* 82.5 (1992): 1142–1161. Print.

Franch, Fabio. "(Wisdom of the Crowds): 2010 UK Election Prediction with Social Media." *Journal of Information Technology & Politics* 10.1 (2013): 57–71. Print.

Freeman, Linton C. "A Set of Measures of Centrality Based on Betweenness." *Sociometry* 40.1 (1977): 35–41. Print.

———. "Centrality in Social Networks Conceptual Clarification." *Social Networks* 1.3 (1978): 215–239. Print.

Futoma, Joseph. "Scalable Inference Algorithms for Clustering Large Networks." Undergraduate thesis, Dartmouth College, 2013. Print.

Gao, Yue et al. "Brand Data Gathering From Live Social Media Streams." *Proceedings of the ACM International Conference on Multimedia Retrieval 2014 (ICMR 2014)*. 2014. 169–169. Print.

Gauntlett, David. *Making Is Connecting*. Cambridge: Polity Press, 2011. Print.

Gayo-Avello, Daniel, Panagiotis T. Metaxas, and Eni Mustafaraj. "Limits of Electoral Predictions Using Twitter." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. 490–493. Print.

Gayo-Avello, Daniel. "'I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper': A Balanced Survey on Election Prediction Using Twitter Data." *arXiv:1203.1647 [cs.SI]*. Cornell University Library, 1 May 2012. Web. 26 June 2015. <http://arxiv.org/abs/1204.6441>.

———. "A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data." *Social Science Computer Review* 31.6 (2013): 649–679. Print.

———. "Don't Turn Social Media Into Another 'Literary Digest' Poll." *Communications of the ACM* 54.10 (2011): 121–128. Print.

———. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16.6 (2012): 91–94. Print.

Ghose, Anindya, and Panagiotis G. Ipeirotis. "Estimating the Helpfulness and Economic

Impact of Product Reviews: Mining Text and Reviewer Characteristics." *IEEE Transactions on Knowledge and Data Engineering* 23.10 (2011): 1498–1512. Print.

Ghosh, Saptarshi et al. "Understanding and Combating Link Farming in the Twitter Social Network." *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*. New York: ACM Press, 2012. 61–70. Print.

Gil, Richard, and Steven D. Levitt. "Testing the Efficiency of Marketsin the 2002 World Cup." *The Journal of Prediction Markets* 1.3 (2007): 255–270. Print.

Gilbert, Eric, and Karrie Karahalios. "Predicting Tie Strength with Social Media." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. New York: ACM Press, 2009. Print.

———. "Widespread Worry and the Stock Market." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM '10)*. Menlo Park: AAAI Press, 2010. 58–65. Print.

Gilbert, Eric. "Predicting Tie Strength in a New Medium." *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. New York: ACM Press, 2012. 1047–1056. Print.

Ginsberg, Jeremy et al. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457.7232 (2009): 1012–1014. Print.

Goel, Sharad et al. "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences* 107.41 (2010): 17486–17490. Print.

Golbeck, Jennifer et al. "Predicting Personality From Twitter." *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int'l Conference on Social Computing (SocialCom)*. Washington, D.C.: IEEE Computer Society, 2011. 149–156. Print.

Golbeck, Jennifer, and Derek Hansen. "Computing Political Preference Among Twitter Followers." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. New York: ACM Press, 2011. 1105–1108. Print.

Gómez, Vicenç, Hilbert J. Kappen, and Andreas Kaltenbrunner. "Modeling the Structure and Evolution of Discussion Cascades." *Proceedings of The 22nd ACM Conference on Hypertext and Hypermedia (HT '11)*. New York: ACM Press, 2011. 181–190. Print.

Gonçalves, Pollyanna et al. "Comparing and Combining Sentiment Analysis Methods." *Proceedings of the First ACM Conference on Online Social Networks (COSN '13)*. New York: ACM Press, 2013. 27–38. Print.

González, Roberto et al. "On the Tweet Arrival Process at Twitter: Analysis and Applications." *Transactions on Emerging Telecommunications Technologies* 25.2 (2014):

273–282. Print.

Granka, Laura. “Using Online Search Traffic to Predict US Presidential Elections.” *PS: Political Science & Politics* 46.02 (2013): 271–279. Print.

Granovetter, Mark. “Threshold Models of Collective Behavior.” *American Journal of Sociology* 83.6 (1978): 1420–1443. Print.

Greaves, Felix et al. “Harnessing the Cloud of Patient Experience: Using Social Media to Detect Poor Quality Healthcare.” *BMJ Quality & Safety* 22.3 (2013): 251–255. Print.

Gruhl, Daniel et al. “The Predictive Power of Online Chatter.” *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD ’05)*. 2005. 1–10. Print.

Gruzd, Anatoliy, and Jeffrey Roy. “Investigating Political Polarization on Twitter: A Canadian Perspective.” *Policy & Internet* 6.1 (2014): 28–45. Print.

Gu, Hansu et al. “*ETree*: Effective and Efficient Event Modeling for Real-Time Online Social Media Networks.” *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT ’11)*. Washington, D.C.: IEEE Computer Society, 2011. 300–307. Print.

Guzmán, Giselle. “Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations.” *Journal of Economic and Social Measurement* 36.3 (2011): 119–167. Print.

Hanson, Robin. “Foul Play in Information Markets.” *Information Markets: A New Way of Making Decisions*. Eds. Robert W. Hahn and Paul C. Tetlock. Washington, D.C.: AEI Press, 2006. Print.

Hargittai, Eszter, and Amanda Hinnant. “Digital Inequality: Differences in Young Adults' Use of the Internet.” *Communication Research* 35.5 (2008): 602–621. Print.

Hawkins, Jeff, and Sandra Blakeslee. *On Intelligence*. New York: Henry Holt and Company, 2004. Print.

He, Yulan et al. “Quantising Opinions for Political Tweets Analysis.” *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC ’12)*. ELRA, 2012. 3901–3906. Print.

Heider, Fritz. *The Psychology of Interpersonal Relations*. Hillsdale: Lawrence Erlbaum Associates, Inc., 1958. Print.

Herbsleb, James D. “Global Software Engineering: The Future of Socio-Technical Coordination.” *2007 Future of Software Engineering (FOSE ’07)*. Washington, D.C.: IEEE Computer Society, 2007. 188–198. Print.

Hibbs, Douglas A., Jr. “Implications of the ‘Bread and Peace’ Model for the 2008 US

Presidential Election." *Public Choice* 137.1-2 (2008): 1–10. Print.

Hinds, Pamela, and Cathleen McGrath. "Structures That Work: Social Structure, Work Structure and Coordination Ease in Geographically Distributed Teams." *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. New York: ACM Press, 2006. 343–352. Print.

Holbrook, Thomas M. "Incumbency, National Conditions, and the 2008 Presidential Election." *PS: Political Science & Politics* 41.04 (2008): 1–4. Print.

Hongzhi Yin et al. "A Unified Model for Stable and Temporal Topic Detection From Social Media Data." *Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE '13)*. Washington, D.C.: IEEE Computer Society, 2012. 661–672. Print.

Hossain, Liaquat, Andrè Wu, and Kenneth K. S. Chung. "Actor Centrality Correlates to Project Based Coordination." *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. New York: ACM Press, 2006. 363–372. Print.

Hsieh, Chin-Shan, and Jian-Hsin Chou. "Forecasting Value at Risk (VAR) in the Shanghai Stock Market Using the Hybrid Method." Unpublished manuscript, Kao Yuan University, 2009. Print. <http://www.kyu.edu.tw/teacpage/teacpage97/97%E8%AB%96%E6%96%87%E6%88%90%E6%9E%9C%E5%BD%99%E7%B7%A8/204.pdf>.

Huberman, Bernardo, Daniel M. Romero, and Fang Wu. "Social Networks That Matter: Twitter Under the Microscope." *First Monday* 14.1 (2009): 1–9. Print.

Huff, Darrell. *How to Lie with Statistics*. New York: W. W. Norton & Company Inc., 2004. Print.

Hulth, Anette, Gustaf Rydevik, and Annika Linde. "Web Queries as a Source for Syndromic Surveillance." *PLoS ONE* 4.2 (2009): e4378–10. Print.

Hürriyetoglu, Ali. "Estimating the Time Between Twitter Messages and Future Events." *Proceedings of the 13th Dutch-Belgian Information Retrieval Workshop*. CEUR Workshop Proceedings, 2013. 20–23. Print.

Imran, Muhammad et al. "Extracting Information Nuggets From Disaster-Related Messages in Social Media." *Proceedings of the 10th International Conference on Information Systems on Crisis Response and Management (ISCRAM 2013)*. 2013. Print.

Jaafar, Nouf, Manal Al-Jadaan, and Reem Alnutaifi. "Framework for Social Media Big Data Quality Analysis." *New Trends in Database and Information Systems II*. Berlin and Heidelberg: Springer-Verlag Berlin Heidelberg, 2015. 301–314. Print.

Jacobsen, Ben et al. "(In)Accuracy of a European Political Stock Market: The Influence of

Common Value Structures." *European Economic Review* 44.2 (2000): 205–230. Print.

Jansen, Bernard J. et al. "Twitter Power: Tweets as Electronic Word of Mouth." *Journal of the American Society for Information Science and Technology* 60.11 (2009): 2169–2188. Print.

Jansen, Harold J., and Royce Koop. "Pundits, Ideologues, and Ranters: The British Columbia Election Online." *Canadian Journal of Communication* 30 (2005): 613–632. Print.

Java, Akshay et al. "Why We Twitter: Understanding Microblogging Usage and Communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*. New York: ACM Press, 2007. 56–65. Print.

Jiang, Long et al. "Target-Dependent Twitter Sentiment Classification." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Stroudsburg: Association for Computational Linguistics, 2011. 151–160. Print.

Jin, Xin et al. "The Wisdom of Social Multimedia: Using Flickr for Prediction and Forecast." *Proceedings of the International Conference on Multimedia (MM '10)*. New York: ACM Press, 2010. 1235–1244. Print.

Jungherr, Andreas, and Pascal Jürgens. "Forecasting the Pulse." *Internet Research* 23.5 (2014): 589–607. Print.

Jungherr, Andreas, Pascal Jürgens, and Harald Schoen. "Why the Pirate Party Won the German Election of 2009 or the Trouble with Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. 'Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment'." *Social Science Computer Review* 30.2 (2012): 229–234. Print.

Jürgens, Pascal, Andreas Jungherr, and Harald Schoen. "Small Worlds with a Difference: New Gatekeepers and the Filtering of Political Information on Twitter." *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. New York: ACM Press, 2011. Print.

Justin Wolfers, Eric Zitzewitz. *Interpreting Prediction Market Prices as Probabilities*. IZA Discussion Paper No. 2092, 2006. Print.

Katz, Elihu, and Paul F. Lazarsfeld. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. New York: The Free Press, 1955. Print.

Kak, Subhash C. "A Class of Instantaneously Trained Neural Networks." *Information Sciences* 148 (2002): 97–102. Print.

———. "New Algorithms for Training Feedforward Neural Networks." *Pattern Recognition Letters* 15 (1994): 295–298. Print.

———. “On Training Feedforward Neural Networks.” *Pramana - Journal of Physics* (1993): 35–42. Print.

———. “The Three Languages of the Brain: Quantum, Reorganizational, and Associative.” *Learning as Self-Organization* (1996): 185–219. Print.

Kak, Subhash C., Yuhua Chen, and Lei Wang. “Data Mining Using Surface and Deep Agents Based on Neural Networks.” *AMCIS 2010 Proceedings*. AIS, 2013. Print.

Kalampokis, Evangelos, Efthimios Tambouris, and Konstantinos Tarabanis. “Understanding the Predictive Power of Social Media.” *Internet Research* 23.5 (2013): 544–559. Print.

Kass-Hout, Taha A, and Hend Alhinnawi. “Social Media in Public Health.” *British Medical Bulletin* 108.1 (2013): 5–24. Print.

Kawash, Jalal, ed. *Online Social Media Analysis and Visualization*. Cham: Springer International Publishing, 2014. Print.

Kim, Kyoung-Sook et al. “Sophy: A Morphological Framework for Structuring Geo-Referenced Social Media.” *Proceedings of the Seventh ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. New York: ACM Press, 2014. Print.

King, Gary. “Replication, Replication.” *PS: Political Science & Politics* 28 (1995): 444–452. Print.

Kleinberg, Jon. “Bursty and Hierarchical Structure in Streams.” *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’02)*. New York: ACM Press, 2002. 91–25. Print.

Kosinski, Michal, David Stillwell, and Thore Graepel. “Private Traits and Attributes Are Predictable From Digital Records of Human Behavior.” *Proceedings of the National Academy of Sciences* 110.15 (2013): 5802–5805. Print.

Krauss, Jonas et al. “Predicting Movie Success and Academy Awards Through Sentiment and Social Network Analysis.” ECIS 2008 Proceedings. AIS, 2008. Print.

Kwak, Haewoon et al. “What Is Twitter, a Social Network or a News Media?.” *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. New York: ACM Press, 2010. 591–600. Print.

Lampos, Vasileios, and Nello Cristianini. “Nowcasting Events From the Social Web with Statistical Learning.” *ACM Transactions on Intelligent Systems and Technology* 3.4 (2012): 1–22. Print.

———. “Tracking the Flu Pandemic by Monitoring the Social Print.” *Proceedings of the 2nd International Workshop on Cognitive Information Processing (CIP 2010)*. Washington, D.C.: IEEE Computer Society, 2010. 411–416. Print.

Lampos, Vasileios, Daniel Preotiuc-Pietro, and Trevor Cohn. “A User-Centric Model of

Voting Intention From Social Media." 2013. 993–1003. Print.

Lee, Ryong, and Kazutoshi Sumiya. "Measuring Geographical Regularities of Crowd Behaviors for Twitter-Based Geo-Social Event Detection." *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN '10)*. New York: ACM Press, 2010. Print.

Lerman, Kristina, and Aram Galstyan. "Analysis of Social Voting Patterns on Digg." *Proceedings of the First Workshop on Online Social Networks (WOSN '08)*. New York: ACM Press, 2008. 7–7. Print.

Lerman, Kristina, and Tad Hogg. "Using a Model of Social Dynamics to Predict Popularity of News." *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. New York: ACM Press, 2010. 621–630. Print.

Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman. "The Dynamics of Viral Marketing." *ACM Transactions on the Web* 1.1 (2007): 1–37. Print.

Lewis-Beck, Michael S., and Tom W. Rice. *Forecasting Elections*. Washington, D.C.: CQ Press, 1992. Print.

Liang, Yuan et al. "How Big Is the Crowd?: Event and Location Based Population Modeling in Social Media." *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT '13)*. New York: ACM Press, 2013. 99–108. Print.

Liangfei, Qiu, Huaxia Rui, and Andrew Whinston. "A Twitter-Based Prediction Market: Social Network Approach." *Proceedings of the International Conference on Information Systems (ICIS 2011)*. AIS, 2011. Print.

Lica, Liviu, and Mihaela Tutä. "Predicting Product Performance with Social Media." *Informatica Economica* 15.2 (2011): 46–56. Print.

Liu, Bing, and Lei Zhang. "A Survey of Opinion Mining and Sentiment Analysis." *Mining Text Data*. Eds. Charu C. Aggarwal and ChengXiang Zhai. Boston: Springer US, 2012. 415–463. Print.

Liu, Bing. "Sentiment Analysis and Opinion Mining." *Synthesis Lectures on Human Language Technologies* 5.1 (2012): 1–167. Print.

Liu, Ling et al. "A Social-Media-Based Approach to Predicting Stock Comovement." *Expert Systems With Applications* 42.8 (2015): 3893–3901. Print.

Liu, Yang et al. "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*. New York: ACM Press, 2007. 607–614. Print.

Liu, Yong et al. "User-Generated Content on Social Media: Predicting Market Success with

Online Word-of-Mouth." *Social Informatics*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2010. 5–5. Print. Lecture Notes in Computer Science.

———. "User-Generated Content on Social Media: Predicting Market Success with Online Word-of-Mouth." *IEEE Intelligent Systems* 25.1 (2010): 75–78. Print.

Livne, Avishay et al. "The Party Is Over Here: Structure and Content in the 2010 Election." *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. Print.

Lui, Catherine, Panagiotis T. Metaxas, and Eni Mustafaraj. "On the Predictability of the U.S. Elections Through Search Volume Activity." *Proceedings of the IADIS International Conference on e-Society (ES 2011)*. IADIS, 2011. Print.

Marco Pennacchiotti, Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011): 281–288. Print.

Massoudi, Kamran et al. "Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts." *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR '11)*. Berlin, Heidelberg: Springer-Verlag, 2011. 362–367. Print.

Meeyoung Cha et al. "The World of Connections and Information Flow in Twitter." *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 42.4 (2012): 991–998. Print.

Mestyán, Márton, Yasseri Taha, and János Kertész. "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data." *PLoS ONE* 8.8 (2013): e71226–8. Print.

Metaxas, Panagiotis T., Eni Mustafaraj, and Dani Gayo-Avello. "How (Not) to Predict Elections." *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. Washington, D.C.: IEEE Computer Society, 2011. 165–171. Print.

Metzler, Donald, Congxing Cai, and Eduard Hovy. "Structured Event Retrieval Over Microblog Archives." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT '12)*. 2012. 646–655. Print.

Mishne, Gilad, and Natalie Glance. "Predicting Movie Sales From Blogger Sentiment." *2006 AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*. Menlo Park: AAAI Press, 2006. 155–158. Print.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20.3 (2012):

271–291. Print.

Murthy, Dhiraj. "Twitter and Elections: Are Tweets, Predictive, Reactive, or a Form of Buzz?." *Information, Communication & Society* 18.7 (2015): 816–831. Print.

Mustafaraj, Eni et al. "Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail." *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int'l Conference on Social Computing (SocialCom)*. Washington, D.C.: IEEE Computer Society, 2011. 103–110. Print.

Mustafaraj, Eni, and Panagiotis T. Metaxas. "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search." *Proceedings of the WebSci10: Extending the Frontiers of Society*. 2010. Print.

Nardi, Bonnie A. et al. "Why We Blog." *Communications of the ACM* 47.12 (2004): 41–46. Print.

Nguyen, Thin et al. "Event Extraction Using Behaviors of Sentiment Signals and Burst Structure in Social Media." *Knowledge and Information Systems* 37.2 (2012): 279–304. Print.

O'Connor, Brendan et al. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (ICWSM '10). Menlo Park: AAAI Press, 2010. 122–129. Print.

O'Connor, Philip, and Feng Zhou. "The Tradesports NFL Prediction Market: An Analysis of Market Efficiency, Transaction Costs, and Bettor Preferences." *Journal of Prediction Markets* 2.1 (2008): 45–71. Print.

Oh, Chong, and Olivia R. Liu Sheng. "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement." *Proceedings of the International Conference on Information Systems (ICIS 2011)*. AIS, 2011. Print.

Olson, Donald R. et al. "Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales." *PLoS Computational Biology* 9.10 (2013): e1003256–11. Print.

Osborne, Miles, and Mark Dredze. "Facebook, Twitter and Google Plus for Breaking News: Is There a Winner?." *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM '14)*. Menlo Park: The AAI Press, 2014. Print.

Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2.1–2 (2008): 1–135. Print.

Park, Cheol, and Thae Min Lee. "Information Direction, Website Reputation and eWOM Effect: A Moderating Role of Product Type." *Journal of Business Research* 62.1 (2009): 61–67. Print.

Pennock, David M. “The Real Power of Artificial Markets.” *Science* 291.5506 (2001): 987–988. Print.

Perry, Paul. “Certain Problems in Election Survey Methodology.” *The Public Opinion Quarterley* 43.3 (1979): 312–325. Print.

Petrović, Saša, Miles Osborne, and Victor Lavrenko. “Streaming First Story Detection with Application to Twitter.” *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT ’10)*. 2010. 181–189. Print.

Phuvipadawat, Swit, and Tsuyoshi Murata. “Breaking News Detection and Tracking in Twitter.” *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2010)*. Washington, D.C.: IEEE Computer Society, 2010. 120–123. Print.

Pinto, Henrique, Jussara M. Almeida, and Marcos A. Gonçalves. “Using Early View Patterns to Predict the Popularity of YouTube Videos.” *Proceedings of the Sixth ACM International Conference on Web search and Data Mining (WSDM ’13)*. New York: ACM Press, 2013. 365–374. Print.

Polgreen, Philip M., Forrest D. Nelson, et al. “Use of Prediction Markets to Forecast Infectious Disease Activity.” *Clinical Infectious Diseases* 44.2 (2007): 272–279. Print.

Polgreen, Philip M., Yiling Chen, et al. “Using Internet Searches for Influenza Surveillance.” *Clinical Infectious Diseases* 47.11 (2008): 1443–1448. Print.

Popescu, Ana-Maria, and Marco Pennacchiotti. “Detecting Controversial Events From Twitter.” *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM ’10)*. New York: ACM Press, 2010. 1873–1876. Print.

Popescu, Ana-Maria, Marco Pennacchiotti, and Deepa Paranjpe. “Extracting Events and Event Descriptions From Twitter.” *Proceedings of the 20th International Conference Companion on World Wide Web (WWW 2011)*. New York: ACM Press, 2011. 105–106. Print.

Quercia, Daniele et al. “Our Twitter Profiles, Our Selves: Predicting Personality with Twitter.” *2011 IEEE Third Int’l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int’l Conference on Social Computing (SocialCom)*. Washington, D.C.: IEEE Computer Society, 2011. 180–185. Print.

Ratkiewicz, Jacob et al. “Detecting and Tracking Political Abuse in Social Media.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM ’11)*. Menlo Park, AAAI Press, 2011. 297–304. Print.

Reuter, Timo et al. “Social Event Detection at MediaEval 2013: Challenges, Datasets, and

Evaluation." *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*. CEUR Workshop Proceedings, 2013. Print.

Rhode, Paul W, and Koleman S. Strumpf. "Historical Presidential Betting Markets." *Journal of Economic Perspectives* 18.2 (2004): 127–142. Print.

Richardson, Matthew, Ewa Dominowska, and Robert Ragno. "Predicting Clicks: Estimating the Click-Through Rate for New Ads." *Proceedings of the 16th international conference on World Wide Web (WWW 2007)*. New York: ACM Press, 2007. 521–529. Print.

Ritterman, Joshua, Miles Osborne, and Ewan Klein. "Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic." *Proceedings of the 1st International Workshop of Mining Social Media*. 2009. Print.

Romero, Daniel M. et al. "Influence and Passivity in Social Media." *Proceedings of the 20th international Conference Companion on World Wide Web (WWW 2011)*. New York: ACM Press, 2011. 113–114. Print.

Rui, Huaxia, and Andrew Whinston. "Designing a Social-Broadcasting-Based Business Intelligence System." *ACM Transactions on Management Information Systems* 2.4 (2011): 22-19. Print.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors." *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. New York: ACM Press, 2010. 851–860. Print.

———. "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development." *IEEE Transactions on Knowledge and Data Engineering* 25.4 (2013): 919–931. Print.

Salathé, Marcel, and Shashank Khandelwal. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control." *PLoS Computational Biology* 7.10 (2011): e1002199–7. Print.

Samangooei, Sina et al. "Social Event Detection via Sparse Multi-Modal Feature Selection and Incremental Density Based Clustering." *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*. CEUR Workshop Proceedings, 2013. Print.

Sang, Erik Tjong Kim, and Johan Bos. "Predicting the 2011 Dutch Senate Election Results with Twitter." *Proceedings of the Workshop on Semantic Analysis in Social Media*. Stroudsburg: Association for Computational Linguistics, 2012. 53–60. Print.

Sankaranarayanan, Jagan et al. "TwitterStand: News in Tweets." *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. New York: ACM Press, 2009. 42–10. Print.

Schmidhuber, Jürgen. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61 (2014): 85–117. Print.

Schoen, Harald et al. "The Power of Prediction with Social Media." *Internet Research* 23.5 (2013): 528–543. Print.

Serdar Biçer, Ayşe Başar Bener, and Bora Çağlayan. "Defect Prediction Using Social Network Analysis on Issue Repositories." *Proceedings of the 2011 International Conference on Software and Systems Process (ICSSP '11)*. 2011 63–71. Print.

Sharda, Ramesh, and Dursun Delen. "Predicting Box-Office Success of Motion Pictures with Neural Networks." *Expert Systems With Applications* 30.2 (2006): 243–254. Print.

Shi, Lei et al. "Predicting US Primary Elections with Twitter." *Proceedings of the NIPS Workshop on Social Network and Social Media Analysis Methods, Models and Applications*. London: SAGE Publications, 2012. Print.

Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. During the Influenza a H1N1 Pandemic." *PLoS ONE* 6.5 (2011): e19467–10. Print.

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. New York: Penguin Books, 2012. Print.

Simonoff, Jeffrey S., and Ilana R. Sparrow. "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers." *Chance* 13.3 (2000): 15–24. Print.

Skoric, Marko et al. "Tweets and Votes: A Study of the 2011 Singapore General Election." *Proceedings of the 2012 45th Hawaii International Conference on System Sciences (HICSS '12)*. Washington, D.C.: IEEE Computer Society, 2011. 2583–2591. Print.

Skowronski, John J, and Donal E. Carlston. "Negativity and Extremity Biases in Impression Formation: A Review of Explanations.." *Psychological Bulletin* 105.1 (1989): 131–142. Print.

Song, Chaoming et al. "Limits of Predictability in Human Mobility." *Science* 327.5968 (2010): 1018–1021. Print.

Stefanidis, Anthony, Andrew Crooks, and Jacek Radzikowski. "Harvesting Ambient Geospatial Information From Social Media Feeds." *GeoJournal* 78.2 (2011): 319–338. Print.

Stieglitz, Stefan, and Linh Dang-Xuan. "Political Communication and Influence Through Microblogging: An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior." *Proceedings of the 2012 45th Hawaii International Conference on System Sciences (HICSS '12)*. Washington, D.C.: IEEE Computer Society, 2011. 3500–3509. Print.

Surowiecki, James. *The Wisdom of Crowds*. New York: Anchor Books, 2005. Print.

Syidada, Shofiya, Noor Fitria Azzahra, and Eva Yulia Puspaningrum. "Promoter Account Detection in Twitter." *JUTI: Jurnal Ilmiah Teknologi Informasi* 12.1 (2014): 35–39. Print.

Szabo, Gabor, and Bernardo A. Huberman. "Predicting the Popularity of Online Content." *Communications of the ACM* 53.8 (2010): 80–88. Print.

Tatar, Alexandru et al. "A Survey on Predicting the Popularity of Web Content." *Journal of Internet Services and Applications* 5.1 (2014): 8–20. Print.

Taxidou, Io, and Peter Fischer. "Realtime Analysis of Information Diffusion in Social Media." *Proceedings of the VLDB Endowment* 6.12 (2013): 1416–1421. Print.

Tops, Hannah. "The Predictive Power of Tweets: An Exploratory Study." Master's thesis, Utrecht University and Radboud Universiteit Nijmegen, 2013. Print.

Tuarob, Suppawong, and Conrad S. Tucker. "Fad or Here to Stay: Predicting Product Market Adoption and Longevity Using Large Scale, Social Media Data." *Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2015)*. ASME, 2013. Print.

Tumarkin, Robert, and Robert F. Whitelaw. "News or Noise?: Internet Postings and Stock Prices." *Financial Analysts Journal* 57.3 (2001): 41–51. Print.

Tumasjan, Andranik et al. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM '10)*. Menlo Park: AAAI Press, 2010. 178–185. Print.

Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Where There Is a Sea There Are Pirates: Response to Jungherr, Jurgens, and Schoen." *Social Science Computer Review* 30.2 (2012): 235–239. Print.

———. "Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape." *Social Science Computer Review* 29.4 (2011): 402–418. Print.

Tziralis, Georgios, and Ilias Tatsiopoulos. "Prediction Markets: An Extended Literature Review." *The Journal of Prediction Markets* 1 (2007): 75–91. Print.

van Keulen, Maurice, and Mena B. Habib. "Uncertainty Handling in Named Entity Extraction and Disambiguation for Informal Text." *Financial Cryptography and Data Security*. Cham: Springer International Publishing, 2014. 309–328. Print. Lecture Notes in Computer Science.

Vosen, Simeon, and Torsten Schmidt. "A Monthly Consumption Indicator for Germany Based on Internet Search Query Data." *Applied Economics Letters* 19.7 (2012): 683–687. Print.

Vosen, Simeon, and Torsten Schmidt. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends." *Journal of Forecasting* 30.6 (2011): 565–578. Print.

Walther, Maximilian, and Michael Kaisser. "Geo-Spatial Event Detection in the Twitter Stream." *Financial Cryptography and Data Security*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2013. 356–367. Print. Lecture Notes in Computer Science.

Wang, XiaoFeng, Matthew S. Gerber, and Donald E. Brown. "Automatic Crime Prediction Using Events Extracted From Twitter Posts." *Financial Cryptography and Data Security*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2012. 231–238. Print. Lecture Notes in Computer Science.

Watts, Duncan J. "Challenging the Influentials Hypothesis." *Measuring Word of Mouth* 3.4 (2007): 201–211. Print.

Weng, Jianshu et al. "Event Detection in *Twitter*." *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. Menlo Park: AAAI Press, 2011. Print.

Williams, Christine B., and Girish J. Gulati. "What Is a Social Network Worth?: Facebook and Vote Share in the 2008 Presidential Primaries." *Annual Meeting of the American Political Science Association*. APSA, 2008. Print.

Wilson, Kumanan, and John S. Brownstein. "Early Detection of Disease Outbreaks Using the Internet." *Canadian Medical Association Journal* 180.8 (2009): 829–831. Print.

Wolf, Timo et al. "Predicting Build Failures Using Social Network Analysis on Developer Communication." *Proceedings of the 31st International Conference on Software Engineering (ICSE '09)*. Washington, D.C.: IEEE Computer Society, 2009. 1–11. Print.

Wolfers, Justin, and Eric Zitzewitz. "Prediction Markets." *Journal of Economic Perspectives* 18.2 (2004): 107–126. Print.

Wong, Felix Ming Fai, Soumya Sen, and Mung Chiang. "Why Watching Movie Tweets Won't Tell the Whole Story?." *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks (WOSN '12)*. New York: ACM Press, 2012. 61–66. Print.

Wood, Lincoln C. et al. "Using Sentiment Analysis to Improve Decisions in Demand-Driven Supply Chains." Unpublished manuscript, University of Auckland, 2014. Print. <http://docs.business.auckland.ac.nz/Doc/Wood-anzamsymposium2014_submission_118-final.pdf>.

Wood, Lincoln C., Torsten Reiners, and Hari S. Srivistava. "Expanding Sales and Operations Planning Using Sentiment Analysis: Demand and Sales Clarity From Social Media." *Proceedings of the 27th Australian and New Zealand Academy of Management Conference: Managing from the Edge*. Deakin: ANZAM, 2013. Print.

Wu, Fang, and Bernardo A. Huberman. "Novelty and Collective Attention." *Proceedings of the National Academy of Sciences* 104.45 (2007): 17599–17601. Print.

Wu, Lynn, and Erik Brynjolfsson. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities." *Proceedings of the International Conference on Information Systems (ICIS 2009)*. AIS, 2009. 1–15. Print.

Wüthrich, Bruno et al. "Daily Prediction of Major Stock Indices From Textual WWW Data." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98)*. Menlo Park: AAAI Press, 1998. 364–368. Print.

Yang, Yiming, Tom Pierce, and Jaime Carbonell. "A Study of Retrospective and on-Line Event Detection." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. New York: ACM Press, 1998. 28–36. Print.

Yu, Sheng, and Subhash C. Kak. "A Survey of Prediction Using Social Media." *arXiv:1203.1647 [cs.SI]*. Cornell University Library, 7 Mar. 2012. Web. 26 June 2015. <http://arxiv.org/abs/1203.1647>.

Yuan, Jianbo, Quanzeng You, and Jiebo Luo. "Are There Cultural Differences in Event Driven Information Propagation Over Social Media?." *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia (SAM '13)*. New York: ACM Press, 2013. 3–8. Print.

Zaman, Tauhid R. et al. "Predicting Information Spreading in Twitter." *Computational Social Science and the Wisdom of Crowds (NIPS 2010)*. 2010. Print.

Zhang, Li, Jianhua Luo, and Suying Yang. "Forecasting Box Office Revenue of Movies with BP Neural Network." *Expert Systems With Applications* 36.P2 (2009): 6580–6587. Print.

Zhang, Peng, Xufei Wang, and Baoxin Li. "Evaluating Important Factors and Effective Models for Twitter Trend Prediction." *Online Social Media Analysis and Visualization*. Ed. Jalal Kawash. Cham: Springer International Publishing, 2015. 81–98. Print. Lecture Notes in Social Networks.

———. "On Predicting Twitter Trend: Important Factors and Models." *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York: ACM Press, 2013. 1427–1429. Print.

———. "On Predicting Twitter Trend: Factors and Models." *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13)*. New York: ACM Press, 2013. 1427–1429. Print.

Zhang, Wenbin, and Steven Skiena. "Improving Movie Gross Prediction Through News Analysis." *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '09)*. Washington, D.C.: IEEE Computer Society, 2009. 301–304. Print.

Zhang, Xiaoming et al. “Event Detection and Popularity Prediction in Microblogging.” *Neurocomputing* 149.PC (2015): 1469–1480. Print.

Zhang, Xue, Hauke Fuehres, and Peter A. Gloor. “Predicting Asset Value Through Twitter Buzz.” *Advances in Collective Intelligence 2011*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2012. 23–34. Print. Advances in Intelligent and Soft Computing.

———. “Predicting Stock Market Indicators Through Twitter ‘I Hope It Is Not as Bad as I Fear’.” *Procedia - Social and Behavioral Sciences* 26 (2011): 55–62. Print.

Zhang, Zhenya et al. “TextCC: New Feed Forward Neural Network for Classifying Documents Instantly.” *Financial Cryptography and Data Security*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2005. 232–237. Print. Lecture Notes in Computer Science.

Zhou, Xiangmin, and Lei Chen. “Event Detection Over Twitter Social Media Streams.” *The VLDB Journal* 23.3 (2013): 381–400. Print.

Zhou, Yanbo, An Zeng, and Wei-Hong Wang. “Temporal Effects in Trend Prediction: Identifying the Most Popular Nodes in the Future.” *arXiv:1412.6753 [cs.SI]*. Cornell University Library, 21 Dec. 2014. Web. 26 June 2015. <http://arxiv.org/abs/1412.6753>.

Zimmerman, Albrecht. “On the Cutting Edge of Event Detection From Social Streams - A Non-Exhaustive Survey.” Unpublished manuscript, GRAISearch, Laboratoire d'InfoRmatique en Image et Systèmes d'information, 2014. Print. <http://liris.cnrs.fr/dm2l/projects/graisearch/files/event-detection-external.pdf>.