Organizational Research Methods
Volume 12 Number 3
July 2009 436-460

10.1177/1094428108317202
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Text Mining in Qualitative Research

Application of an Unsupervised Learning Method

Nina Janasik
Timo Honkela
Helsinki University of Technology, Finland
Henrik Bruun
Pieni Kirahvi Oy Ab, Helsinki, Finland

The article provides an introduction to and a demonstration of the self-organizing map (SOM) method for organizational researchers interested in the use of qualitative data. The SOM is a versatile quantitative method very commonly used across many disciplines to analyze large data sets. The outcome of the SOM analysis is a map in which entities are positioned according to similarity. The authors' argument is that text mining using the SOM is particularly effective in improving inference quality within qualitative research. SOM creates multiple well-grounded perspectives on the data and thus improves the quality of the concepts and categories used in the analysis.

Keywords: *grounded theory; constructivism; self-organizing map; text mining; document interpretation*

The relationship between qualitative and quantitative research has been the focus of intense debate ever since the science wars. Recently, the duality between the two approaches has generated a "third movement" in the form of mixed-methods research. This approach has produced sophisticated models of possible forms of interrelation as well as reflections on fundamental epistemological issues (Greene & Caracelli, 2003; Miller, 2003; Teddlie & Tashakkori, 2003). Simultaneously, qualitative research itself is becoming increasingly self-reflexive (e.g., Clarke, 2005; Denzin & Lincoln, 2000; Silverman, 2004, 2006). Consequently, also the traditional grounded theory approach as conceived and elaborated by Glaser and Strauss has witnessed major reconfigurations (Clarke, 2005; Locke, 1996, 2001). In turn, the latter development has revealed exciting linkages between grounded theory and a particular quantitative method, that of the self-organizing map (SOM; e.g., Castellani, Castellani, & Spray, 2003).

In this article, we wish to open up the connection between qualitative research, especially grounded theory, and the SOM even further. Our central argument is that the SOM

Authors' Note: We thank the anonymous reviewers, the editors, and Helena Buhr, Janne Hukkinen, Maria Höyssä, Samuel Kaski, Mika Pantzar, Mikko Rask, Ann Russell, and Olli Salmi for their detailed and constructive comments on earlier versions of this article.

(a) significantly improves the quality of the inferences drawn by researchers doing predominantly qualitative research and (b) provides a relatively objective approach for quantitative researchers interested in working with qualitative data. The former is mainly because utilizing the SOM in text mining can improve the quality of the concepts and categories used in the analysis. The latter is mainly because the method restricts itself to the input data in exploring its underlying conceptual connections. Because of space restrictions, we mainly focus on the former part of the argument.

Improving inference quality is of great utility in all social science research, but particularly and uniquely so when the qualitative data sets are very large (e.g., thousands of documents) and/or the stakeholders numerous. Examples of such contexts are urban planning (Godschalk, 2004; Healey, 1998), the development of participatory institutions for natural resource management (e.g., Hukkinen, 2006; Hukkinen et al., 2006; Hukkinen, Heikkinen, Raitio, & Müller-Wille, 2006); Müller-Wille & Hukkinen, 1999), conflict management and security studies (e.g., Langlais, 1995), and dealing with complex and multifaceted challenges such as climate change. The utility of the SOM in improving inference quality basically follows from the fact that the method can easily be used to generate multiple well-grounded perspectives on the data. These perspectives are not a collection of random views but form an organized whole.

As an illustrative example of how the SOM can be used to improve inference quality within qualitative research, we use a case study by Janasik (2003) on knowledge integration in a Finnish coffee firm. The article is structured as follows. First, we provide a background discussion on the SOM in relation to qualitative research and issues of researcher bias. Next, we introduce the SOM in more detail and review some of its previous uses in contexts similar to the ones discussed here. The next section shows how the coffee firm study can be approached by means of the SOM in relation to both data-driven and theory-driven approaches. Finally, we conclude the article by connecting the SOM to the revised form of grounded theory (situational analysis) proposed by Clarke (2005) and by discussing some limitations of the SOM method.

Background

We begin by giving a brief characterization of the SOM as a so-called *unsupervised learning method*. Unsupervised learning is one of the paradigms in machine learning and statistical data analysis. In supervised learning, the system is given both the input and the desired output, and it learns to construct a mapping between these. In unsupervised learning, a model is fitted to observations, and there is no a priori output. Thus, unsupervised learning may give rise to novel model constructions autonomously emerging from the data. The SOM (Kohonen, 2001) is an unsupervised learning method that originally stems from artificial neural network research. Currently, it is commonly used as a method for statistical visualization and data analysis (Kaski, Kangas, & Kohonen, 1998; Oja, Kaski, & Kohonen, 2003; Pöllä, Honkela, & Kohonen, in press). The outcome of the SOM analysis is a map in which entities, such as people, words, sentences, or documents, are clustered according to similarity with respect to some property.

The SOM and Qualitative Research

How does the unsupervised learning method of the SOM relate to the seemingly different world of qualitative research? We take our starting point in the typology of "mixed-model designs" presented by Teddlie and Tashakkori (2003; see Table 1). In relation to their classification, our discussion focuses on the mixed-model designs Pure Qualitative, Mixed Type I, Mixed Type II, and Mixed Type IV. The Pure Qualitative design is one in which researchers are conducting exploratory investigations using qualitative data as well as qualitative analysis and inference procedures. In Mixed Type II, researchers are conducting confirmatory investigations based on the same kinds of data and methods of analysis. Mixed Type IV represents a research design in which researchers are conducting exploratory investigations using qualitative data but analyzing this statistically. Mixed Type I is identical to Mixed Type IV except that its aim is confirmatory.

We further divide what Teddlie and Tashakkori call the "analysis and inference" stage of the research process (Teddlie & Tashakkori, 2003, p. 31) into two modes or orientations—data driven (inductive) and theory driven (deductive)—in the four mixed-model designs on which we focus in this article (see Table 1 and also Figure 3 later in the article). Doing so enables us to distinguish between two different ways of doing qualitative research: one that insists that all higher-order categorizations emerge from the data and another that views all categories as potentially fruitful tools for data exploration (see Figure 3). The latter, but not the former, allows for the creative construction of theory-based categories, often in conceptual frameworks, which are then applied in the analysis of some data (Bruun, Langlais, & Janasik, 2005).[1]

We argue that the SOM, as a dexterous unsupervised learning method, can be useful in facilitating the formation of higher-order categories for qualitative researchers with a data-driven bent, on one hand, and testing the adequacy of the less data-anchored constructs of theory-driven researchers, on the other. Within both theory-driven and data-driven qualitative research, the quality of the inferences drawn crucially depends on the adequacy of the terms or categories used. If the chosen terms or categories do not reflect something of importance in the data under study, it is not likely that inferences drawn from the data using those terms are going to be of any major value either.

Using the SOM improves higher-order constructs in the following way. For any collection of textual data, it is possible to create a so-called document map providing a general view of it. This view can be regarded as another perspective on that data. However, it is possible with the SOM to easily produce not only one but also a multitude of such perspectives. If one combines all perspectives, it is highly likely that the SOM perspectives will be partly different from all the others. For instance, the SOM representation might challenge some higher-order category that has already been formulated, urging the researcher to adjust it in the direction of higher accuracy. This clearly represents an improvement of that higher-order category. Because the quality of the inferences drawn crucially depends on the adequacy of the used concepts and categories, improving the latter by means of multiple perspectives also implies improving the former.

Table 1
Teddlie and Tashakkori's (2003) Classification of Mixed-Model Designs

Confirmatory Investigation				**Exploratory Investigation**			
Quantitative Data and Operations		**Qualitative Data and Operations**		Quantitative Data and Operations		**Qualitative Data and Operations**	
Statistical analysis and inference	Qualitative analysis and inference	**Statistical analysis and inference**	**Qualitative analysis and inference**	Statistical analysis and inference	Qualitative analysis and inference	**Statistical analysis and inference**	**Qualitative analysis and inference**
Pure quantitative	Mixed Type V (rare)	**Mixed Type I**	**Mixed Type II**	Mixed Type III	Mixed Type VI (rare)	**Mixed Type IV**	**Pure qualitative**

Note: Our focus is indicated by the bolded terms.

Qualitative Research and Researcher Bias

Before moving on to presenting the SOM, it is, however, important to make sure that we do not conflate the notion of qualitative methods with certain techniques for gathering data. Observation, participation, document analysis, and interviews exemplify techniques that are often associated with qualitative research, but they are not "qualitative" in themselves. It is perfectly possible to do quantitative research by using these techniques. The term *qualitative* refers not to the way in which data are gathered but to the type of data that are collected and to the method with which the data are analyzed. The most general meaning of *qualitative* is simply that it is not quantitative. Thus, the data that are gathered do not need to be transformed to numbers, and mathematical and statistical tools are not used in the analysis. Instead, the data are processed through systematization, categorization, and interpretation.

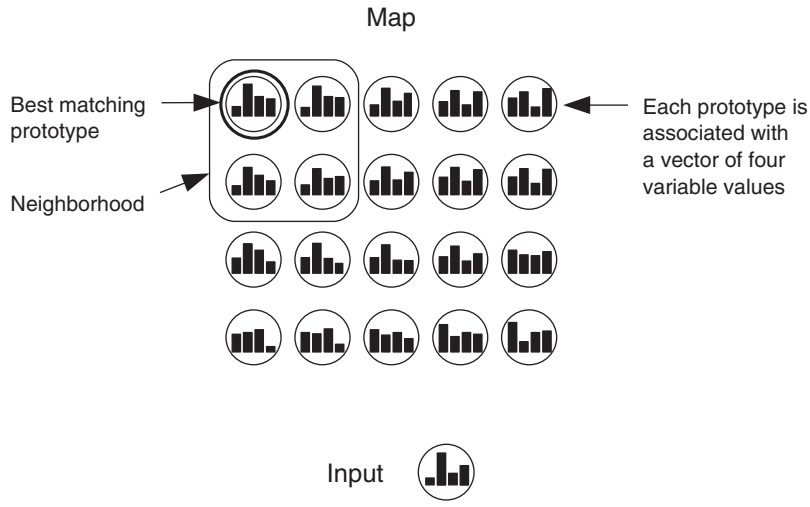
The qualitative method so conceived has many advantages, including sensitivity for detail and context. At the same time, it is also clear that qualitative research methods have some important limitations. In particular, qualitative analysis is dependent on the interpretative activity of the person who does the analysis (e.g., the researcher). Consequently, the analytic outcome can be affected by the researcher's own conceptions, biases, styles of thinking, and so on. The problem is that the researcher's structural framework for interpretation imposes a bias on the object of inquiry. This distortion of the object of inquiry can occur in two ways. First, it can occur because the researcher has decided that the studied phenomenon has to possess some specific attributes and then proceeds to fill in the structures with appropriate parts from the data. This kind of "forcing" is the main target of the grounded theory critique (e.g., Glaser & Strauss, 1967; Strauss & Corbin, 1990).

However, there is also a second and subtler source of researcher bias that grounded theorists also need to take into consideration. It is well known that humans tend to think in terms of pure categories, or ideal types (e.g., male vs. female, developed world vs. developing world, modern vs. traditional, user vs. producer), and assume that such categories reflect the organization of reality (e.g., Scott, 2001; Suchman, 1995). However, human behavior is extremely complex and hardly ever strictly conforms to the tightly bound conceptual structures that social scientists use to describe that behavior. We return to the ways in which the two approaches within qualitative research (data-driven and theory-driven) deal with the issue of researcher bias in relation to the coffee firm analysis.

The SOM

As an unsupervised artificial network or statistical machine learning method (Kohonen, 1982, 2001), the SOM compares to classical unsupervised quantitative research methods such as multidimensional scaling (MDS; Kruskal & Wish, 1978) or clustering (Jardine & Sibson, 1971; Sneath & Sokal, 1973). The SOM has been extensively used to analyze numerical data in a number of areas, including various branches of industry, medicine, and economics. The use of the SOM has also been extended into the analysis of text data (Honkela, Kaski, Lagus, & Kohonen, 1997; Kohonen et al., 2000; Lagus, Kaski, & Kohonen, 2004). It can be used for the study of large amounts of material: hundreds or thousands of interview transcripts, e-mails, Web sites, formal documents, and so on. The SOM analysis

Figure 1
The Basic Architecture of the Self-Organizing Map



Note: The input is fully connected to the array of prototypes, which is usually two-dimensional. Each prototype, visualized as a circle on the grid, serves as a model of a class of similar inputs. The bar diagrams inside the circles denote the four variable values. For instance, the prototype on the upper-left corner corresponds to cases in which the first variable has a low value and the second variable a high value. It is also the "winner" or the closest prototype when the input, shown below the map, is considered. In an ordered map, the corresponding variable values in neighboring prototypes are relatively similar.

produces a two-dimensional map in which documents, concepts, or some other entities of interest are clustered according to similarity. The SOM and other similar techniques for ordering data are commonplace in information science but have been little used in organizational research.

The SOM serves several analytical functions. First, it provides a mapping from a high-dimensional space into a low-dimensional space, thus providing a suitable means for visualization of complex data. Second, the SOM reveals topological structure of the data. The topological distance between two points in the map is proportional to the distance between the points in the original input space. Next, we describe the SOM algorithm, compare it with some related methods, and describe the use of the SOM in text mining and in qualitative research.

The SOM Algorithm

The emergent relationship between the input data and the prototypes of the map is based on the SOM algorithm. In the first step of the SOM algorithm (Kohonen, 2001), the grid of prototype vectors is initialized (see Figure 1). The simplest initialization is to give each prototype vector random values. Thereafter, the SOM analysis of the input data proceeds in the following manner. Processing of one data sample consists of two steps: (a) search for the best matching prototype and (b) update of the best matching prototype and its neighbors on the map. These steps are performed iteratively to all data samples.

Figure 2
Illustration of a Self-Organizing Map Based on Data Collected by Hofstede (1980)

Note: The clustering diagram shows the relationships among countries based on four cultural factors (see text for details). Relative distances in the original four-dimensional space are illuminated by the shading; the darker an area on the map, the higher the distance. For instance, Japan appears to relatively dissimilar from any other country.

During the iterations, there are two main parameters that influence the process: learning factor and neighborhood size. The learning factor determines the scale of the updates: The larger the learning factor, the more the best matching prototype and its neighbors are changed toward the input sample. The neighborhood size determines how large of an area around the best-matching prototype will be updated. The value of these parameters gradually decreases during the analysis process. After a number of iterations of the processing steps, the map stabilizes, and the information contained in the input data set becomes encoded into the prototype vectors.

As an example of how the SOM has been used within organizational research, we present the study on national cultural differences by Hofstede (1980). If two countries have similar values related to the four basic cultural aspects (uncertainty avoidance, power distance, masculinity–femininity, and individualism–collectivism), they tend to appear close to each other on the map (see Figure 2). From the map produced by the SOM, one can discern relationships such as the relative similarity between Portugal and Greece, Sweden and Denmark, and the United Kingdom and the United States.

Methodological Comparisons

Venna and Kaski (2006) pointed out that in information visualization, when the dimensionality of data is reduced to two (or three), this is not in general possible without losing some of the original proximities. They indicated two kinds of errors that can occur. First, data points originally farther away from each other may end by being close to each other in the outcome. These errors decrease the trustworthiness of the visualization. Second, data points that are originally close to each other can be pushed farther away, causing

discontinuity of the visualization. Each dimensionality reduction method, such as MDS (Kruskal & Wish, 1978), principal component analysis (Hotelling, 1933), and SOM (Kohonen, 2001), necessarily makes a trade-off between trustworthiness and continuity. Venna and Kaski (2006) compare a number of methods for nonlinear dimensionality reduction, including many versions of MDS and conclude that only SOM and curvilinear component analysis (Demartines & Hérault, 1997) can be recommended for general visualization tasks for which high trustworthiness is required.

Why choose SOM, as MDS is still more commonly used in social sciences? In part, this is because there are some important philosophical and qualitative distinctions that cannot be made simply by some straightforward statistical benchmarking studies. However, perhaps the most important argument is that the SOM is considered to be one of the most realistic models of cortical organization of the brain. According to brain research, the cortex is responsible for high-level cognitive functions, such as creating a meaningful model of the world (Kalat, 2006). Thus, the SOM has a broad relevance to the cognitive science field, as has been pointed out by, for instance, Bechtel and Abrahamsen (2002), Gärdenfors (2000), MacWhinney (1998), and Miikkulainen (1993). Statistical methods such as factor analysis or MDS do not have this quality, or the relationship is at best very indirect. The SOM has also been extensively used to model concept formation processes (Gärdenfors, 2000; Honkela, 2000; MacWhinney, 1998), which is of direct relevance to the cognitive processes of categorization.

The SOM specifies a holistic conceptual space. The meaning of some item in an analysis is not based on a predefined definition or a position in a conceptual hierarchy or network (as in concept mapping or semantic network-based analyses) but is the emergent result of a number of encounters in which the item is used in some context. Moreover, the emergent prototypes on the map are not isolated instances (as in many forms of cluster analysis), but they influence each other in the adaptive formation process.

To summarize, the SOM can be considered as a collection of adaptive prototypes that exist in relation to each other. This is also closely related to earlier research on prototype theory and the embodied nature of knowledge (Varela, Thomson, & Rosch, 1991). In its emphasis on the grounded nature of knowledge, the SOM approach aligns well with the central epistemological presuppositions of traditional and revised grounded theory (Castellani et al., 2003; Clarke, 2005; Glaser & Strauss, 1967; Locke, 1996; Strauss & Corbin, 1990).

Because the unsupervised learning methods use input data to generate the output, the methods are not, in principle, vulnerable to researcher bias or a priori categorizations. Naturally, the outcome of the analysis is affected by the choice of variables included in the analysis and by the weighting of the variables. However, none of the variables is given a specific status as it would be in supervised learning and classification. In addition to its trustworthiness, the SOM is often used thanks to its computational efficiency: It is possible to analyze even millions of data points that have hundreds or thousands of variables.

Document Analysis Using the SOM

We now turn our attention to how the SOM can be used to analyze textual data. In statistical studies of language, it is widely accepted that the context in which words and phrases appear provides information on their meaning (e.g., Manning & Schütze, 1999). For instance,

a widely used method for text mining is the latent semantic analysis (LSA), in which singular value decomposition is used to create a mapping from a high-dimensional representation of words in context into a lower-dimensional representation of latent semantic variables (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). However, the representation created by the LSA method is distributed and does not create such a readily meaningful representation as the SOM does.

The SOM has been used in text mining to analyze words in their contexts (Honkela, Pulkki, & Kohonen, 1995; Ritter & Kohonen 1989), semantic structures of document collections (Chen, Schuffels, & Orwig, 1996; Honkela, Kaski, Lagus, & Kohonen, 1996; Lagus et al., 2004; Lin, Soergel, & Marchionini, 1991), and even people based on the documents that they have produced. By virtue of the SOM algorithm, documents can be mapped, based on their full text content, onto the map grid so that related documents appear close to each other. This result can be called a *document map*. A document map provides a general view of the document collection. The basic idea in creating a document map is that each text is represented as a vector indicating the relative frequencies of words and phrases or the presence of some categories in the text. It is important that the contents of the texts are considered as patterns. Individual words or expressions are thus not analyzed one-by-one.

The process of using the SOM in the area of text mining consists of the following steps: (a) select a document collection, (b) automatically or manually choose the terminology to be used in encoding the documents, (c) transform, based on the terminology, the documents into numerical data, (d) initiate the SOM iterations with varying parameters, (e) study and interpret the resulting document maps and term distributions on the maps, and (f) formulate the inferences that can be drawn from the maps.

For Step b, the most straightforward approach is to manually choose the terminology. This gives the researcher the best control of the study but also introduces some subjectivity into the process. To avoid the subjectivity, various methods have been developed for automatic term extraction, that is, for extracting meaningful words and expressions from texts. A number of methods have been suggested for this, but no general consensus exists as of yet on the best available approach. Two general methodological alternative approaches can be discerned: statistical and ontology-based. The first is based on the analysis of patterns of word and phrase usage in texts. For instance, the most common and the least common words and phrases are typically not good term candidates. The ontology-based method relies on a given thesaurus of terms. These can be used to extract meaningful patterns from the text. The choice of terminology selection approach depends on many criteria, such as the nature of the text collection (size, style, language, etc.) and the available linguistic resources and tools.

We now provide a brief methodological outline of the Steps b to d above.[2] Here, it is assumed that the manual terminology selection is in use. In Step b, spreadsheet software can be used to encode the number of occurrences of the selected terms in each document in a collection. Each document corresponds to a row in the table, and each term is represented by a column. The cells in the spreadsheet then contain the number of occurrences of each term in each document. The matrix may be normalized by dividing the number in each cell by the row sum. This operation makes sure that documents of different length are considered in the analysis in a balanced manner. This matrix can then be given as input to some statistical software package that contains the SOM algorithm. In our case, we have used Matlab software, specifically the SOM Toolbox developed for Matlab

(available for free at http://www.cis.hut.fi/projects/somtoolbox/). The SOM in the SOM Toolbox environment is created by a command such as "sMap = som_make(Data, 'msize,' [30 40])," in which *sMap* is the data structure in which the resulting map is stored, *Data* refers to the input matrix, and the parameter *msize* specifies the *x* and *y* dimensions of the map. The toolbox can be used to label and visualize the map in various ways with the commands "som_label" and "som_show" in a straightforward and intuitive manner described in the user manual of the software.

Many qualitative researchers may find it challenging to use a method such as terminology extraction or the SOM, with which they potentially do not have any prior experience. For organizational researchers with some background in computational methodology, a preferable approach would be to use manual term selection as described above and some existing SOM implementation. However, we recommend that these kinds of studies be conducted in collaborative and interdisciplinary contexts because this would ensure that various aspects of the necessary methodological expertise would be available. This is particularly true for automatic terminology extraction and some advanced uses of the SOM.

Qualitative Research Using Document Maps

In the following, we exemplify the potential of using SOM through a case study in which the amount of qualitative data was a central factor. In 2004, the Academy of Finland, one of the country's largest funding agencies, commissioned a study to investigate to what extent and how the academy had promoted interdisciplinary research in its funding and to recommend how the academy could improve its capabilities in fostering interdisciplinary research (Bruun, Hukkinen, Huutoniemi, & Thompson Klein, 2005). Bruun, Hukkinen, and his colleagues (2005) used applications to the Academy of Finland as empirical material, classifying the applications on the basis of a qualitative analysis of their contents. They found that more than 40% of a sample of 324 successful research applications proposed to do interdisciplinary research. During the analysis, 266 applications were read and carefully analyzed and qualitatively assessed. This process took one researcher approximately 5 to 6 weeks.

As a continuation of this manually conducted qualitative analysis (Bruun, Hukkinen, et al., 2005), the Academy of Finland in 2006 commissioned another study to investigate whether text mining based on the SOM could be used to support assessment of the applications. A collection of 3,224 applications was analyzed (Honkela & Klami, 2007). A collection of 1,331 term candidates was automatically extracted using a reference corpus–based method (Honkela et al., 2007). The method is based on the following idea: Words and phrases that are relatively more common in the text collection under study than the reference corpus are good term candidates. Term candidates that belonged to categories such as names of persons, organizations, or places were then manually left out. In the end, there was a collection of 1,200 terms. The 3,224 application documents were encoded as term distribution patterns. The SOM algorithm organized the documents into a map in which similar applications are close to each other and in which thematic areas emerged. One interesting finding was related to the division into research councils. The Academy of Finland organizes its activities into four councils: (a) health, (b) biosciences and environment, (c) culture

and society, and (d) natural sciences and engineering. In the SOM analysis, the applications were distributed on the map in a manner that mostly followed the division into the councils, with one important exception: the natural sciences and engineering research council was split into two parts, biosciences and environment. Specifically, the research related to chemistry within natural sciences and engineering was clearly separated from the research on, for instance, physics and engineering sciences. Moreover, research on chemistry was also closer to the area occupied by the health research council than the other disciplines in natural sciences and engineering.

The analysis discussed above was mostly automatic. The input data were in the form of unstructured PDF files. Some scripts had to be programmed to extract the text content from these files. Approximately 2 hours of manual work was used to filter term candidates. To summarize, little manual work was needed to deal with the 3,224 documents.

Integrating Biases: Qualitative Research and the SOM

We now turn to the case study of the article, Janasik's (2003) theory-driven research on "knowledge frameworks" in a small coffee company in Finland. Janasik studied collaboration across "knowledge frameworks" within this company. Her background assumption was that the integration of knowledge was important for the performance of any company, and she wanted to study how such integration was implemented. She interviewed a number of owners and key employees, analyzed a set of textual documents, and spent some time at the company office observing activity. Because her research approach presumed that there are knowledge boundaries to be overcome, Janasik set out to identify the knowledge frameworks that guided perception and action within the company. On the basis of the suggestion for a "conceptual framework" for knowledge integration or "knowledge networking" presented in Bruun, Langlais, et al. (2005), she defined the "knowledge framework" of an individual or a collective as being composed of three things: (a) the conceptualization of the world in which professional actions are performed (the task domain), (b) the methods for acquiring knowledge about that world, and (c) the self-characterizations entertained by the acting agent. Janasik identified six knowledge frameworks that had formed over time. Some of the knowledge frameworks were embodied by single individuals (the company was relatively small), whereas others were more widespread.

She then wrote the history of the company in terms of interactions among people embodying the different knowledge frameworks. Janasik's analysis of the history of the coffee company is quite typical for theory-driven organizational research. It posits the existence of a number of larger conceptual structures—knowledge frameworks—and describes historical development as interplay among them. The problem with the notion of such theoretically postulated structures is that they are often quite loosely defined. They are abstract and do not suggest any unambiguous operationalization. As a result, the researcher has to intuitively judge how the object of knowledge is structured, what the key methods of learning are, and how the epistemic self is constructed by the people under investigation (i.e., the three defining aspects of knowledge frameworks). In practice, such

Table 2
The Characteristics of Two Knowledge Frameworks (KFs) That Were Operative in a Small Coffee Company in Helsinki, Finland, in 2002

Characteristics	Entrepreneurially Oriented KF	Economically Oriented KF
Object of knowledge	Coffee brands, coffee machines, other coffee-related products, business trends, cultural trends Coffee-related social network ("Who does what?" etc.)	Changes in revenues, turnover, debts, salaries, other costs, productivity Changes in regulations, labor contracts, contracts with suppliers and customers
Method for learning and knowledge generation	Method: Visioning, planning, scanning, building social networks, mobilizing people Instruments: Personal calendar for managing social contacts; car for moving around ("office time is often a waste"); papers, journals, and the Internet for scanning; phone for communicating	Method: Monthly financial reports, annual accounts, monetary transactions (salaries, invoices), communication with heads of outlets, communication with employer's association and labor union Instruments: Office, computer, bookkeeping software, Internet, calculator, archives, branch journals
Epistemic self-understanding	Purpose: To create new business, to initiate new activity, to make good deals Measures of success: Mobilization of the board and employees around new initiatives, growth in turnover, stabilization of new activities Image: Enthusiastic project initiator	Purpose: To keep the company on a healthy financial track, to make the financial processes transparent Measures of success: The usage of economic instruments in company decision making, working cash flow practices, good solidity, salaries paid in time Image: Rational guardian of firm expenditure

an analysis searches for key concepts with which to describe the knowledge framework (see Table 2). This procedure leaves us with the following questions:

1. How do we know that we have identified the relevant categories? There may be other categories that organize the perceptions and actions of people in different functional roles. In the coffee shop study, the categories were defined a priori.
2. How do we know that we have identified the relevant concepts? Table 2 contains a selection of concepts that were either used by the interviewees themselves or created by the researcher. This was done intuitively, without any explicit rules for concept identification or the creation of new concepts.

In the following, we briefly review how the same coffee firm research process would have looked had our researcher used the data-driven approach of grounded theory. Instead of reading through existing theories and case studies on the theme of knowledge integration in firms (e.g., Bruun, Langlais, et al., 2005), she would have started by going out to the field. After each bout of information gathering, she would have made notes of what she took to be the key issues. Thus, the central categories of the research situation would

have *emerged* from the empirical data. She would also have coded the data by means of the *method of constant comparison*. Had she done this, at some stage a ''core category'' or a category highly connected to many other categories would have emerged. In both cases, the end result might well have been ''categories that organize the perceptions and actions of people in different functional roles'' (see Problem 1 on category identification), but the routes to those higher-order categories are very different.

Moreover, the choice of route to the categories has consequences for the way in which ''concepts'' are fitted to them (see Problem 2 on concept identification). The strategy of the theory-driven researcher is to first create the category and then to proceed to fit the data in relation to those structures. Given that the higher-order categories are established in advance, it is very hard to think of a way of identifying concepts that would *not* be based on the researcher's intuition of how they ought to fit. By contrast, starting from the data does not actualize the problem of a divorce of the concepts from the category to such a large extent. Moving away from theory-driven qualitative approaches toward data-driven ones can thus answer the two problems of category and concept identification. However, it appears that some problems would remain even if our researcher does all that is humanly possible to avoid various forms of bias. Consider, for instance, the following two problems plaguing both approaches:

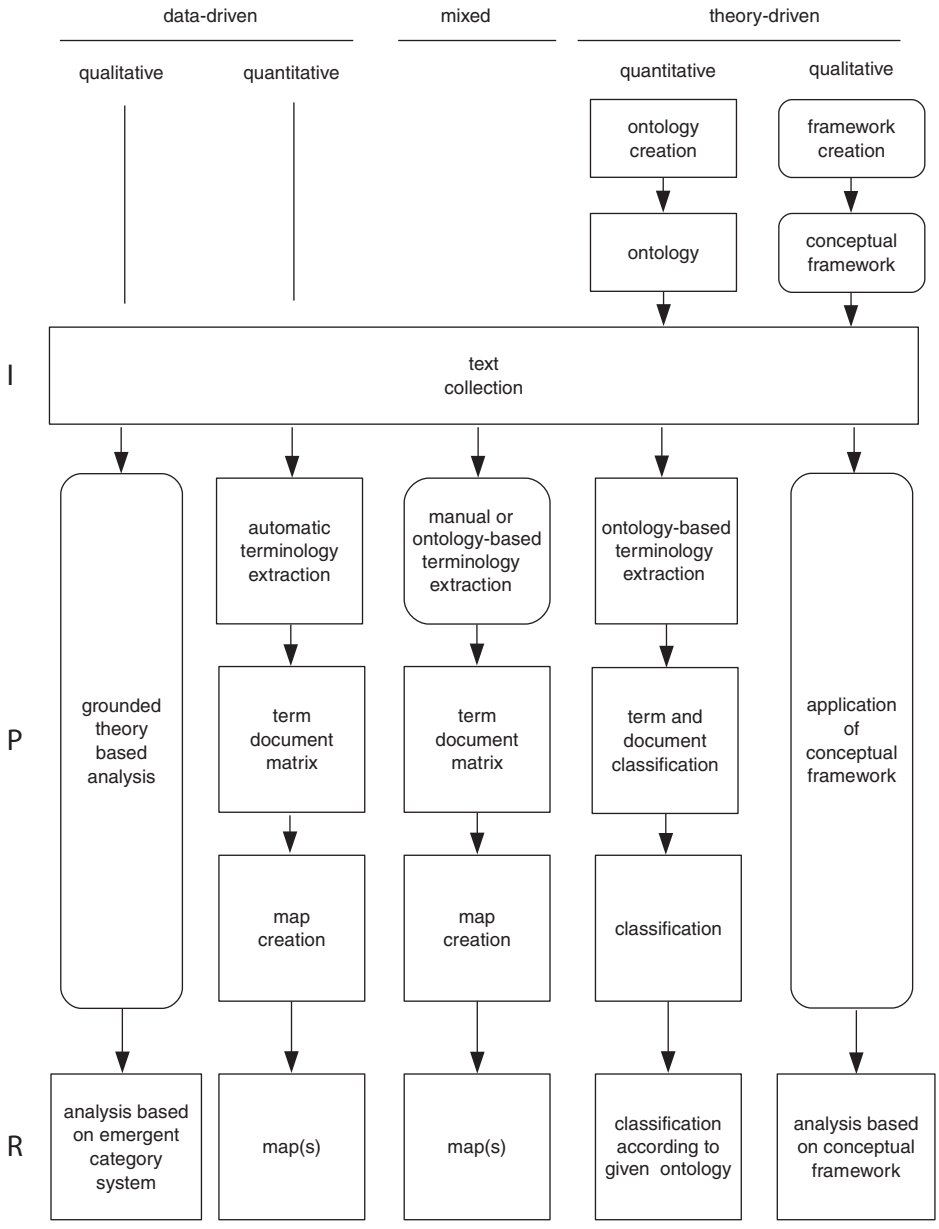
3. The researcher is relatively dependent on the conscious conceptualizations of the interviewees. Still, much of our behavior is only partly available to our consciousness. Nonconscious patterns have to be reconstructed somehow by the researcher.
4. The means of getting to the larger categories remain dependent on the thinking human participant. This means that although the data-driven approach manages to avoid the worst forms of bias, a more insidious form of bias might still enter in the form of a predisposition to categorize in a certain way (e.g., based on dichotomies).

For instance, even if we consider a simple proposition that a person is tall, it is impossible to determine exactly, in a socially shared manner, the borderline between being tall and not being tall. This continuity of meaning is sometimes dealt with using fuzzy set theory (Zadeh, 1965), in which membership values, usually between 0 and 1, are used to represent the partial membership in a category. Moreover, the interpretation is often dependent on several aspects or features: To determine whether and to which degree one can call a person tall depends not only on the height of the person but also on whether the person is female or male, a child (and of which age) or an adult, or a basketball player or not. Next, we show how applying the SOM to qualitative data answers these four arguments of bias.

The Case Study and the SOM Analysis

We applied the SOM to seven interviews in the coffee firm interview collection (see Figure 3). Based on the existing semistructured question battery, the interviews were further subdivided into eight projects undertaken by the coffee firm. All interviewees had been allowed to talk freely about those projects. We chose to manually select the central terms of the interviews rather than to automatically extract them (see Figure 3, middle column) because the corpus was a collection of interviews in Swedish, but a suitable reference corpus in Swedish was not available.

Figure 3
Multiple Perspectives (Vertical Columns) on an Object of Inquiry From the Point of View of Theory-Driven Versus Data-Driven Approaches and Qualitative Versus Quantitative Methods



Note: I = input; P = process; R = result.

When creating the map, we took advantage of the fact that only one of the authors (the first) was familiar with the data from the coffee firm study, leaving it open for the others to relate to it with no preconceived views. Thus, we set up an experiment. First, one author (the second), unacquainted with the coffee firm study, both created the map and analyzed it starting from the data-driven step, that is, looking for emergent patterns. He then wrote down his preliminary conclusions without showing them to anyone else. Next, the first author, who had performed the theory-driven study, analyzed the same map with the explicit intention of seeing how, if at all, it matched the qualitative analysis previously done. Finally, the data-driven and theory-driven authors compared results. We start with the data-driven exploratory approach.

The SOM and the Data-Driven Approach

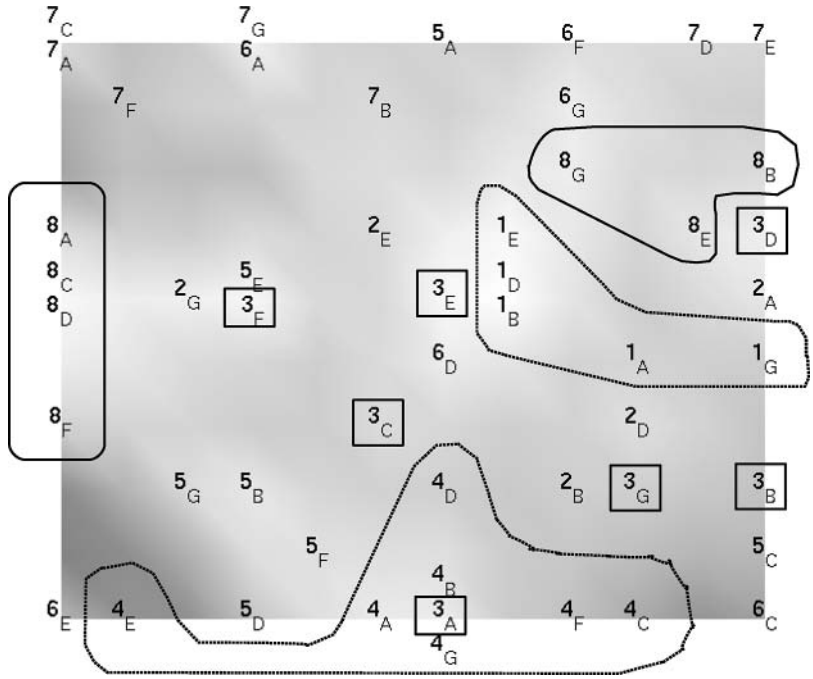
Based on the alternative manual terminology extraction (see Figure 3, middle column), we produced a document map (see Figure 4). The data-driven author concluded the following after a first analysis of the document map:

> From the data-driven point of view, the analysis seems to convey the following. First of all, the domain of the text collection seems to be related to restaurant and cafeteria business. This is revealed by the fact that the high frequency vocabulary includes words such as ''company,'' ''strategy,'' ''business,'' ''economy,'' ''money'' as well as ''restaurant,'' ''coffee,'' ''cafeteria,'' ''lunch,'' ''customer,'' etc. The structure of the document map indicates that three kinds of interviews differ from the rest of the interviews. These ''central'' interviews form four main clusters. The distributions of the selected terms on the document map indicate some further topical structure among these clusters. The interviews on the lower left corner focus on problems. The cluster at the upper left corner also seems to relate to problems with specific restaurants and people. The lower right corner contains texts describing issues related to spaces and aesthetics of cafeteria(s). Among the interrelated central clusters, the one on the left contains texts that deal with economical issues.

How could document maps of the kind discussed (automatic or manual) have helped the data-driven researcher along the way? Because the theme was knowledge integration, he could have focused on the indication of there being clusters related to *problems*. He could then have further explored those problem clusters (What people and projects do they involve?) Based on this, he could have continued interviewing for more information. Or he could have used those problem clusters to confirm or refute preliminary conclusions already emerging from the grounded analysis of the interviews. Had these matched, this could have been interpreted as support for those preliminary conclusions. Had they not, this could have been interpreted to mean that those conclusions were perhaps drawn too hastily.

Thus, the document map could have been helpful for him at the level of projects and problems related to them. However, it could also have been of use in searching for the boundaries of the higher-order categories structuring the thinking and action of the interviewees immersed in those problems. For instance, the document map indicates that the lower-right corner contains texts that describe issues related to the aesthetic aspects of cafeteria design, whereas economic issues tend to appear to the left of the three interrelated clusters (see Figure 4 and the quote above). Apparently, then, these are located at a

Figure 4
An Illustration of a Self-Organizing Map Based on the Data From the Coffee Firm Study



Note: The study considered eight projects, numbered 1 to 8 on the map. In the projects, there were altogether seven stakeholders. These persons were interviewed, and they are indicated on the map with subscripts from A to G. If two interviews have similar content, they tend to appear close to each other on the map. Relative distances in the original high-dimensional space are illuminated by the shading; the darker an area on the map, the higher the distance. This means, for instance, that the two views of Person E on Projects 4 and 6 in the lower-left corner of the map are relatively distant from the other views. Note also the harmony of the structures formed in relation to Projects 1 and 4 and the disparity of Projects 3 and 8.

certain *distance* from each other, thus possibly indicating some kind of difference in underlying organizing structure. This too can be used either in an exploratory way (Who in relation to what project discussed aesthetic versus economic issues?) or in a confirmatory one (How does this map fit with what the researcher has already tentatively concluded from the grounded analysis of the data?).

To summarize, for a qualitative researcher of a data-driven bent, the automatic or manual text extraction produced by the quantitative research tool of the unsupervised learning method of the SOM can be of utility both in *exploratory* (looking for higher-order categories) and *confirmatory* (testing the adequacy of higher-order categories) investigations (see Teddlie & Tashakkori, 2003). The SOM, applied in this way, thus does seem to escape the danger of too-simple categorizations. For instance, although the map indicates that there might be a borderline between terms related to economics and those related to aesthetics, and this might be taken to reflect some degree of difference in underlying organizing structure, the differences are a matter of degree rather than a categorical issue.

The SOM and the Theory-Driven Approach

The first author then analyzed the same map, seeking to confirm or refute the existence of six "knowledge frameworks" present in the coffee firm project dynamics (right column in Figure 3). She found that it captured surprisingly well the way in which two of the eight projects undertaken by the firm had involved major disagreements and the formation of differing viewpoints. The first project involved the juxtaposition of entrepreneurial, economic, and aesthetic knowledge frameworks as the firm grew. To the far right, we have the representative of the entrepreneurial framework in the theory-driven categorization (3_B). At a significant distance to the left of this, we find the representative of the aesthetic framework (3_C). Further up to the right is the representative of the economic framework (3_E). The conclusions drawn from the theory-driven analysis would thus seem to be confirmed as far as the third project is concerned. The map also manages to capture the main tendency of the eighth project, which significantly divided the opinions of the interviewees. It was characterized by the representative of the entrepreneurial knowledge framework (8_B) wanting to challenge established collaborative patterns. Both this separatist movement ($8_{B,E,G}$) and its collaborative consequences can be easily seen on the map ($8_{A,C,D,F}$). In Figure 4, two projects, 1 and 4, stand out as harmonious—indicated by the fact that the distributions on the map are rather concentrated—whereas some other projects also show signs of differing viewpoints.

However, the picture gets less clear when it comes to the question of whether the word pattern profiles reflect an underlying organization into larger-order structures. Although a scrutiny of the document map and the 187 word pattern profiles provides partial support for the existence of higher-order categories of distinctly economic, aesthetic, entrepreneurial, and cultural orientations, the evidence is weaker for the organizational and practical ones. More specifically, terms that the theory-driven analysis associated with the organizational knowledge framework are scattered throughout, whereas terms associated with the practical "knowledge framework" blend with both the economic and the entrepreneurial frameworks. Thus, had the theory-driven researcher applied the SOM to her data, she would have received only partial support for her conclusions. This would have left her with the options of sticking to the six ideal types despite the awareness of their inexactness or of opening the door to modification (perhaps collapsing two of them into one).

To summarize, for a researcher of a theory-driven bent, the automatic or manual term extraction produced by the unsupervised learning method of the SOM can be of use mainly in *confirmatory* investigations (testing the adequacy of higher-order categories; see Teddlie & Tashakkori, 2003). However, this confirmatory endeavor can take on different forms. The theory-driven researcher can either use the automatically generated terms as illustrated above or manually choose the terms to test for relations among "frameworks." For instance, if she wanted to explore the relationship between the hypothesized practical and economic "knowledge frameworks," she could manually select the terms associated with them and see what the SOM would generate, or she could run the terms associated with the presumed organizational framework and compare this to some or all of the other framework distributions. However, at some point such iterative runs could call into question the very structuring of human action in terms of "knowledge frameworks."

Should this happen, the SOM could function as a *quantitative* data-driven way of correcting the possible mistakes done within the qualitative theory-driven mode.

This last point is also important for our overall argument in this article. A hypothetical inference drawn from the small coffee firm study could be that "in running a cafeteria, negotiating the roles of the representatives of the entrepreneurial and the practical knowledge frameworks is particularly important." Although at some level this might be so (e.g., an entrepreneurially minded CEO might interfere with the work of a cafeteria boss on a regular basis), it does matter for the adequacy of the inference whether such a "practical knowledge framework" is justified in the light of the data or not. Our theory-driven analyst could have avoided such biases by simply turning to a data-driven approach. However, she could also have avoided them if she had checked her conclusions using the SOM. We think that all of these ways of proceeding can be viewed as representing multiple perspectives on the same object of inquiry (see Figure 3). In Figure 3, these multiple perspectives have been organized in terms of their position with respect to the dichotomies of data-driven versus theory-driven research and qualitative versus quantitative research, a solution that presents the perspectives as vertical columns.

However, we also think that the perspective represented by the SOM is in reality not only *one* perspective among others. When applied to the same data (the text collection in Figure 3), the SOM perspective in either its automatic (second column from the left) or manual (middle column) mode can be thought of as producing a multitude of different perspectives bound together by "family resemblance." In other words, if we repeatedly applied the SOM to the *same* data, the maps produced would still be *different*—even if we always used the same collection of terms.

This quality of the SOM can be created in two ways. First, one can use a random map as the starting point of the learning process. This leads to a variety of slightly different maps. Second, one can vary the parameters used in the learning process with a similar consequence. The quality of a map can be quantitatively assessed. However, in the case of visualization methods such as the SOM, the assessment necessarily has to take into account at least two different aspects that pull in different directions (see Venna & Kaski, 2006). Finally, the more perspectives that are used in determining the nature of higher-order categorizations, the more likely it becomes that they, and the inferences drawn by using them, are adequate.

Conclusions and Discussion

It should now be clear that the SOM presents an interesting methodological opportunity for qualitative research. In this article, we have argued that the SOM is particularly efficient in improving inference quality within qualitative research, with regard to both confirmatory and exploratory research. Within the theory-driven or deductive mode of qualitative research, the SOM can be used to test the adequacy of conceptual frameworks created before the analysis of the data. In the data-driven or inductive mode, the SOM can be applied in creating emerging category systems describing and explaining the data. However, today's grounded theory is not identical to the one formulated by its founders

(which of course was not as uniform as this article has presented it either; e.g., Charmaz, 2000). How does the SOM relate to modern reformulations of traditional grounded theory?

The SOM and the Situational Analysis Approach

We approach this issue by turning to one modern revision of grounded theory: the one formulated by Adele Clarke (2005) in her *Situational Analysis: Grounded Theory After the Postmodern Turn*. Clarke argued that the recent shifts in the intellectual landscape away from positivist presuppositions toward a recognition of the utmost complexity of social life, and of the analysis of it, means that we can no longer pretend to be theoretically innocent in the way represented by much of traditional grounded theory (e.g., Castellani et al., 2003; Clarke, 2005). We also have to avoid succumbing to the opposite danger, that is, that of performing premature theoretical or analytical closure (we would say theory-driven in its confirmatory mode).

With Clarke (2005), we believe that a promising way to "initially frame and focus the research, drawing on extant literatures and situating the proposed research within those literatures *without* doing premature theoretical closure" (p. 77), is by using so-called *sensitizing concepts* as *research tools*. According to Herbert Blumer (1954), a sensitizing concept "gives the user a general sense of reference and guidance in approaching empirical instances. Whereas definitive concepts provide prescriptions of what to see, sensitizing concepts merely suggest directions in which to look" (p. 7).

In fact, it would seem that the SOM also represents a step forward in relation to situational analysis. Clarke (2005) wished to "regenerate" the traditional grounded theory by pushing it beyond "the postmodern turn" in the history of intellectual life. She did this by (a) showing how the grounded theory approach was, because of its historical roots in symbolic interactionism and pragmatism, already partly beyond that turn and (b) giving it a helping hand toward an analytic approach that fully appreciates the place of *discourse* (especially in its Foucaultian interpretation), the *nonhuman*, and, finally, the *situation* (e.g., as opposed to basic social processes), which, according to Clarke, should be the central unit of analysis. In this approach, the researcher becomes a *cartographer* of the social world, constructing various forms of maps that, taken together, strive to capture social life in all its complexity. An example of such a map would be the *situational* map, which aims at capturing all the elements, human and nonhuman, of relevance for some delimited situation. It is important that this ethos of respecting complexity also requires *enhanced researcher reflexivity* about the research process and about research products. It is thus part of overall researcher accountability. More specifically, what is needed are *innovative strategies for project design and data gathering*, "topics generally unaddressed in traditional grounded theory" (Clarke, 2005, p. xxxviii).

We suggest that the context of mixed-methods research, as compressed in Table 1, is an excellent place in which to ground the discussion of enhanced researcher reflexivity as far as project design is concerned because of its inherently high degree of methodological self-reflexivity. Also, we suggest that such a grounding enables us to make distinctions that in effect help us to see the variation within qualitative research itself as well as the

place and significance of the SOM in relation to those alternative forms. Last but not least, we think that referring to this framework enables us to see both the full value of the revised version of grounded theory provided by Clarke and the place and significance of the SOM in relation to this.

Clarke found that the use of *sensitizing concepts* can enable a double escape from both the idea of naïve emergence (the data will speak for themselves) and the threat of premature theoretical closure (gluing theory onto the findings). In a way analogous to the two "pure" modes described in this article, we think that, for a researcher engaged in the kind of revised grounded theory proposed by Clarke, the automatic or manual term extraction and the SOM analysis can be of utility in both *confirmatory* (testing the adequacy of a sensitizing concept in relation to some data) and *exploratory* (looking for higher-order categories that can function in such a steering manner) investigations (see Clarke, 2005). We thus think of the SOM as a quantitative method or research tool that is particularly well suited to the ethical aim of respecting complexity rather than trying to do away with it.

However, we think that the virtues of the SOM go deeper still. We have seen that the SOM produces not only one but a multitude of perspectives on some data. In relation to very large data sets of the kind discussed in the introduction, some of these multiple perspectives might be such that no human would, even in principle, be able to produce them. This follows from the fact that the computational method can be used to process writings of even millions of persons, something that is beyond the scope of any individual researcher. Yet applying the SOM allows us access to potentially highly relevant and novel categories and patterns that "really are there," even if we do not as yet know it. This would appear to be particularly true when it comes to various nonconscious categorizations. Thus, it would appear that applying the quantitative method of the SOM could take us even *beyond situational analysis* in that it is capable of revealing subconscious operations of the human mind, which the consciously operating human mind of the situational analyst will never be able to discover. We conclude, then, that the cartographer of social life could greatly benefit from taking the maps produced by the nonhuman SOM into serious account.[3]

Restrictions of Unsupervised Learning Methods

Our practical case of applying SOM analysis raises, however, some questions about the restrictions of the unsupervised learning methods. First, the SOM can be used only for analyzing explicit and quantifiable material. In other words, the SOM can be used to analyze data only to the extent that they are manifest, or can be made manifest, in quantifiable data, such as textual documents or survey responses. This covers quite a large range of materials, including, among other things, Web sites, journal articles, scientific articles, diaries, interview transcripts, and so on. In addition to data in textual form, the analysis can also cover other modalities such as images. However, it excludes all information that is not manifested in a quantifiable way, such as the information that an ethnographer can gain by participating in the practices of a community.

A second restriction, related to the first one, is that the output of the SOM analysis, like any statistical study, is highly dependent on the input. The success of the analysis thus

requires a sound selection of input material. This is not always easy. What material should we use? The problem is not only to determine what material should be seen as representative but also that different types of material are available for different groups. For instance, if a number of organizations were compared, e-mail collection might be available for some of them, whereas in some others interviews would be used. The selection of input material will be a key problem in future SOM-based research, and unfortunately this problem has to be solved differently in distinct application areas depending on the underlying objectives. There will be no universal solution.

A third restriction of SOM analysis is that the interpretation of the output of a SOM is dependent on factors that go beyond the map itself. The map presents a clustering of the input data, but it can say nothing about the general importance of that clustering. Although it is true that the SOM is theory independent in its analysis of the data, theory dependence—which is ultimately a dependence on the judgment of the researcher—creeps into the study in both the selection of input material and the interpretation of the output. The SOM may help us to improve our categories, but it will not allow the abandonment of theory and judgment altogether.

Future Prospects

We would like to conclude our discussion of the SOM in relation to qualitative research with a couple of small pointers into a world yet to come. First, in addition to texts, it is also possible to transform images, sound, and speech into input vectors. The details of the kind of methodology needed for multimedia analysis are well beyond the scope of this article. The basic idea is to devise so-called feature detectors that will be used to automatically extract patterns from the data. Today, the automatically extractable aspects of images are rather low-level features such as color distribution, typical shapes, and surface texture.

Second, we showed that the maps produced with the SOM would be different if it were allowed to repetitiously run the same input data. This would not appear to be too far from what would happen with real people in an equivalent situation. Indeed, we strongly suggest that any statistical analysis of a complex phenomenon should be conducted this way (for similar views, see Castellani et al., 2003; Collins & Clark, 1993; DeTienne, DeTienne, & Joshi, 2003; Palocsy & White, 2004). We claim that attempts to confirm a particular hypothesis using some theoretically rigorous statistical method may lead into a false feeling of gaining conclusive information. Rather, statistical methods can be used to discover the multifaceted character of the phenomenon under consideration. Although it is clear that the SOM cannot be viewed as possessing "epistemic agency" in anywhere near the sense of Scardamalia (2002), that is, as an agent capable of assuming responsibility for the advancement of its knowledge and inquiry, we would like to suggest that neither can the SOM be viewed as possessing *no* cognitive autonomy as compared with most other artifacts in the world of research.

Thus, in conclusion, we would like at least to open the question of whether it is justified to think of some research tool as a "semiparticipant" in an ongoing process of interpretation rather than as a mere stepping stone on the road toward "objectivity." This property of the SOM is also highlighted by the fact that the model has been successfully applied in

creating simulations of communities of interacting autonomous agents (Lindh-Knuutila, Honkela, & Lagus, 2006; Wermter, Weber, Elshaw, Gallese, & Pulvermüller, 2005). The way in which we have linked the SOM to qualitative research in this article indicates that unsupervised learning methods, to some extent, can be seen as "research assistants," tirelessly processing even large collections of data and creating meaningful generalized representations of them.

Notes

1. A classical example of data-driven qualitative research is the original work by the Glaser and Strauss (1965) on the awareness of dying. As an example of the theory-driven approach, we cite the research on "knowledge networking" in innovation processes performed by Langlais, Janasik, and Bruun (2004). The researchers used the predefined category of "knowledge networking" and concluded that there seems to exist a relationship among three different *modes* of knowledge networking (called modular, translational, and pioneering) and the *kind of problem* that innovators are trying to solve during the innovation process. When the problems are *well-defined*, the most appropriate way to proceed is to relate the different knowledge frameworks in a *modular* manner. This means that different aspects of the problem are addressed by different people with different knowledge frameworks. The level of integration is very low. When the problem is *ill defined*, however, this will no longer do: Direct interaction across knowledge frameworks or *pioneering* knowledge networking is required. Because the category of knowledge network is taken for granted, the analysis is very different from qualitative research as practiced by grounded theorists. However, applying this framework rather straightforwardly did yield an interesting and, from a policy point of view, potentially beneficial link among various modes of organizing work across knowledge-related boundaries and problem form.

2. Also see the bibliography of self-organizing map (SOM) research (http://www.cis.hut.fi/research/som-bibl/).

3. In our interpretation of the way in which SOM can be of utility to qualitative research, we partially differ from Castellani, Castellani, and Spray (2003), who quite correctly recognized the potential of neural network analysis in relation to the aims of qualitative researchers. Although we do agree with their main thrust, we think that some of their claims are misguided. To begin with, we think that their presentation does not do justice to the capacities and potentialities of SOM, which the article highlights as the best neural network method for grounded theorists to turn to as a methodological supplement. Beside the fact the discussion is limited to grounded theory only, the SOM is framed, in our view wrongly, as a *qualitative* method capable of handling large *quantitative* data sets. We think, first, that the SOM falls squarely within the confines of *quantitative* methods and, second, that its foremost strengths in relation to qualitative research lie in its capacity to make the transformation of *qualitative* data into *quantitative* form in a way that still maintains the complexity of the original data. Also, Castellani et al. do not at all touch on the notion of sensitizing concepts, which we think is highly important in view of the actual practice of qualitative researchers in a variety of disciplines and academic fields.

References

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks* (2nd ed.). Oxford, UK: Basil Blackwell.

Blumer, H. (1954). What is wrong with social theory? *American Sociological Review*, *19*(1), 3-10.

Bruun, H., Hukkinen, J., Huutoniemi, K., & Thompson Klein, J. (2005). *Promoting interdisciplinary research: The case of the Academy of Finland*. Helsinki, Finland: Edita.

Bruun, H., Langlais, R., & Janasik, N. (2005). Knowledge networking: A conceptual framework and taxonomy. *VEST: Journal for Science and Technology Studies*, *18*(3-4), 73-104.

Castellani, B., Castellani, J., & Spray, S. L. (2003). Grounded neural networking: Modeling complex quantitative data. *Symbolic Interaction*, *26*(4), 577-589.

Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509-535). Thousand Oaks, CA: Sage.

Chen, H., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. *Journal of Visual Communications and Image Representation*, *7*(1), 88-102.

Clarke, A. (2005). *Situational analysis: Grounded theory after the postmodern turn*. London: Sage.

Collins, J., & Clark, M. (1993). An application of the theory of neural computation to the prediction of workplace behavior: An illustration and assessment of network analysis. *Personnel Psychology*, *46*, 503-524.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41,* 391-407.

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, *8*, 148-154.

Denzin, N., & Lincoln, Y. (2000). Introduction: The discipline and practice of qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1-36). Thousand Oaks, CA: Sage.

DeTienne, K., DeTienne, D., & Joshi, S. (2003). Neural networks as statistical tools for business researchers. *Organizational Research Methods*, *6*(2), 236-265.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Glaser, B., & Strauss, A. (1965). *Awareness of dying*. New York: Aldine de Gruyter.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine de Gruyter.

Godschalk, D. (2004). Land use planning challenges: Coping with conflicts in visions of sustainable development and livable communities. *Journal of the American Planning Association*, *70*(1), 5-12.

Greene, J., & Caracelli, V. (2003). Making paradigmatic sense of mixed methods practice. In C. Tashakkori & A. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 91-110). Thousand Oaks, CA: Sage.

Healey, P. (1998). Collaborative planning in a stakeholder society. *Town Planning Review*, *69*(1), 1-21.

Hofstede, G. H. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Honkela, T. (2000). Self-organizing maps in symbol processing. In S. Wermter & R. Sun (Eds.), *Hybrid neural systems* (pp. 348-362). Heidelberg, Germany: Springer.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1996). Exploration of full-text databases with self-organizing maps. In *Proceedings of ICNN'96, IEEE international conference on neural networks* (Vol. 1, pp. 56-61). Piscataway, NJ: IEEE.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1997). WEBSOM—Self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on self-organizing maps* (pp. 310-315). Espoo, Finland: Helsinki University of Technology.

Honkela, T., & Klami, M. (2007). *Suomen akatemialle osoitettujen hakemuksien tekstilouhinta* [Text mining of applications submitted to the Academy of Finland]. Unpublished report, Helsinki University of Technology, Espoo, Finland.

Honkela, T., Pöllä, M., Paukkeri, M.-S., Nieminen, I., & Väyrynen, J. J. (2007). *Terminology extraction based on reference corpora* (Technical Report E series). Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science.

Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié & P. Gallinari (Eds.), *Proceedings of ICANN-95, international conference on artificial neural networks* (pp. 3-7). Paris: EC2 et Cie.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417-441, 498-520.

Hukkinen, J. (2006). Sustainability scenarios as interpretive frameworks for indicators of human-environment interaction. In P. Lawn (Ed.), *Sustainable development indicators in ecological economics* (pp. 291-316). Cheltenham, UK: Edward Elgar.

Hukkinen, J., Heikkinen, H., Raitio, K., & Müller-Wille, L. (2006). Dismantling the barriers to entrepreneurship in reindeer management in Finland. *International Journal of Entrepreneurship and Small Business*, *3*(6), 705-727.

Hukkinen, J., Müller-Wille, L., Aikio, P., Heikkinen, H., Jääskö, O., Laakso, A., et al. (2006). Development of participatory institutions for reindeer management in Finland: A diagnosis of deliberation, knowledge integration and sustainability. In B. C. Forbes, M. Bölter, L. Müller-Wille, J. Hukkinen, F. Müller, N. Gunslay, et al. (Eds.), *Reindeer management in northernmost Europe: Linking practical and scientific knowledge in social-ecological systems* (pp. 47-71). Berlin: Springer-Verlag.

Janasik, N. (2003). Den svåra förståelsen: samarbete över kunskapsgränser i ett litet finländskt kaffeföretag [The hardships of understanding: Collaboration across knowledge boundaries in a small Finnish coffee company]. *Technology, Society, Environment*, *2*, 57-106.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. London: Wiley.

Kalat, J. (2006). *Biological psychology*. Belmont, CA: Wadsworth.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural Computing Surveys*, *1*, 102-350.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59-69.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, V., & Saarela, A. (2000). Organization of a massive document collection. *IEEE Transactions on Neural Networks*, *11*(3), 574-585.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Lagus, K., Kaski, S., & Kohonen, T. (2004). Mining massive document collections by the WEBSOM method. *Information Sciences*, *163*(1-3), 135-156.

Langlais, R. (1995). *Reformulating security: A case study from Arctic Canada*. Göteborg, Sweden: Göteborg University.

Langlais, R., Janasik, N., & Bruun, H. (2004). Managing knowledge network processes in the commercialization of science: Two probiotica discovery processes in Finland and Sweden. *Science Studies*, *17*(1), 34-56.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In *Proceedings of 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 262-269). New York: Association for Computing Machinery.

Lindh-Knuutila, T., Honkela, T., & Lagus, K. (2006). Simulating meaning negotiation using observational language games. In *Proceedings of the Third International Symposium on the Emergence and Evolution of Linguistic Communication* (pp. 168-179). Rome, Italy: Springer.

Locke, K. (1996). Rewriting the discovery of grounded theory after 25 years? *Journal of Management Inquiry*, *5*(1), 239-245.

Locke, K. (2001). *Grounded theory in management research*. Thousand Oaks, CA: Sage.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*, 199-227.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon and memory*. Cambridge, MA: MIT Press.

Miller, S. (2003). Impact of mixed methods and design on inference quality. In C. Tashakkori & A. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 423-456). Thousand Oaks, CA: Sage.

Müller-Wille, L., & Hukkinen, J. (1999). Human environmental interactions in Upper Lapland, Finland: Development of participatory research strategies. *Acta Borealia*, *16*(2), 43-61.

Oja, M., Kaski, S., & Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998-2001 Addendum. *Neural Computing Surveys*, *3*, 1-156.

Palocsy, S. W., & White, M. M. (2004). Neural network modeling in cross-cultural research: A comparison with multiple regression. *Organizational Research Methods*, *7*, 389-399.

Pöllä, M., Honkela, T., & Kohonen, T. (in press). Bibliography of self-organizing map (SOM) papers: 2002-2005 Addendum. *Neural Computing Surveys*.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*(4), 241-254.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Jones (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.

Scott, R. W. (2001). *Institutions and organizations* (2nd ed.). Thousand Oaks, CA: Sage.

Silverman, D. (Ed.). (2004). *Interpreting qualitative data: Methods for analysing talk, text and interaction* (2nd ed.). London: Sage.

Silverman, D. (2006). *Qualitative research: Theory, method and practice*. London: Sage.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco: Freeman.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. London: Sage.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, *20*(3), 517-610.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In C. Tashakkori & A. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3-50). Thousand Oaks, CA: Sage.

Varela, F. J., Thomson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Venna, J., & Kaski, S. (2006). Local multidimensional scaling. *Neural Networks*, *19*, 889-899.

Wermter, S., Weber, C., Elshaw, M., Gallese, V., & Pulvermüller, F. (2005). Grounding neural robot language in action. In S. Wermter, G. Palm, & M. Elshaw (Eds.), *Biomimetic neural learning for intelligent robots* (pp. 162-181). Berlin: Springer.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353.

Nina Janasik holds an MA in philosophy from the University of Helsinki in Finland. She is a PhD student at the Helsinki University of Technology Laboratory of Environmental Protection, focusing on transdisciplinary boundary crossing in health-related (functional foods, diagnostics) innovation processes.

Timo Honkela holds a PhD degree from the Helsinki University of Technology (TKK). He has worked as a professor at TKK and the University of Art and Design Helsinki. He is appointed as chief scientist at the cognitive systems research group of the Adaptive Informatics Research Center at TKK.

Henrik Bruun holds a PhD degree from Göteborg University in Sweden. He has worked as a postdoc researcher at the Helsinki University of Technology and as a professor at the University of Turku, Finland. He is currently the CEO of Pieni Kirahvi Oy Ab in Helsinki.