Quick identification of lead user innovations via semantic analysis of web postings

Sandro Kaulartz, IPSOS, Sandro.Kaulartz@Ipsos.com
Eric von Hippel, MIT Sloan School of Management, evhippel@mit.edu

September, 2018

Abstract

Prior research has established that lead users, both industrial and consumer, innovate ahead of general market demand. Producers have an interest in identifying and selectively commercializing innovations developed by lead users, but many have been deterred from making this a regular practice by the difficulty and cost of lead user identification methods developed to date.

In this paper, we describe a method based upon semantic network analytic and memory model techniques for efficient identification of lead user innovations. The data used are "user-generated content" found on thousands of websites that have been made openly available to all. In a test-of-concept application, we test the new method in kiteboarding equipment, a consumer product category. We describe our method in detail, and demonstrate that it is possible to identify a sample of commercially-valuable lead user innovations with a one-or two-person team over the course of a week. We suggest that this same method will be found generally applicable to identification of lead user innovations in other consumer product categories both large and small. We conclude with suggestions for further research, and with a discussion of implications for innovation research and practice.

Quick identification of lead user innovations via semantic analysis of web postings

1 Introduction and overview

Lead users are known to pioneer new types of products and services that later prove to be of general value to many others (e.g., Urban and von Hippel 1988, Franke, von Hippel, and Schreier 2006). Lead users are defined as having two characteristics: (a) they are ahead of the majority of users in their populations with respect to an important market trend; and (b) they expect to gain relatively high benefits from a solution to the needs they encounter at their leading edge market position (von Hippel 1986). The correlations found between innovation by users and those individuals' lead user characteristics are highly significant, and the effects are often very large.

Because lead users sometimes develop, prototype, and apply products of potential commercial value, producers are eager to learn about and evaluate products and services that lead users have developed. However, traditionally-prescribed methods to identify lead user innovations are relatively costly and cumbersome. Best practice has involved producer experts engaging in a chain of interviews to identify innovating lead users via a "pyramiding" process (Churchill et al. 1998, von Hippel, Franke and Prugl 2009). Project experience in the field has shown that it generally takes about four months to do a lead user project with a skilled team of about four individuals using this method (Churchill et al. 1998, p.30). Despite the associated costs, about 24% of a sample of producers do nonetheless claim to use lead user methods, and to regard them as effective (Cooper and Edgett (2007, p.4).

Extensive use of lead user methods despite their relatively costly and time-consuming nature suggests that modernization of best practice methods by utilizing internet-based search would be well worthwhile. Accordingly, in this paper we describe and test an updated lead user innovation identification method based upon internet data filtering methods using semantic network analytic and memory model techniques. The method we have devised is intended to be carried out by a semantic data analyst who collaborates with a part-time subject matter expert to carry out the study via rapid, "agile" interactions. In overview, the method is as follows. With the advice of the subject matter expert, the analyst first scrapes promising and openly available

user generated conversational data from the web. Using subject-matter expertise, semantic and memory model filters are devised, and semantic analysis techniques applied, to screen that data and identify potentially promising user-developed innovations. The filtering process involves iterative trial-and-error, with the subject matter expert working with the data analyst to examine trial outputs and iteratively improve the filters being applied. As a final process step, again using internet data, the analyst generates measures of the potential commercial promise of each innovation identified by assessing the frequency of postings and web searches both providing and seeking information about the innovation, and whether these are increasing or decreasing over time.

The method we describe here is, we think, applicable primarily to consumer goods, because it depends upon the availability of openly-accessible "user-generated content" (UGC). Innovating consumers, especially when participating in communities with like-minded others, often post and discuss innovations they are developing and testing in on-line chat forums. Industrial users, in contrast, are less likely to post openly because they are developing innovations for their own exclusive use and private profit. We estimate from our experience to date that a team of two appropriately skilled individuals, only one working on the project full-time, can carry out an entire study over the course of a week or less.

In this paper, we explain our methods in detail and also demonstrate effectiveness via an application in kiteboarding equipment, a consumer product category with about $2 Billion in annual sales worldwide (Business Wire 2018). In our demonstration, we identify both sport-enhancing and sport-changing innovations developed by sport participants. We also identify specific lead user individuals who have developed and/or experimented with those innovations. We suggest the same general method will be applicable to major consumer product categories as well, and illustrate with examples of user innovations by consumers in the field of hair care products.

The remainder of the paper is organized as follows. In section 2 we review the literature on lead user innovation research and the mechanisms underlying the traditional lead user "pyramiding" search method. In section 3, we describe the general steps required to identify lead user innovations via semantic analysis of user generated content. In section 4, we provide step-by-step descriptions of a case study application and results obtained. (Full methodological details are provided in an Appendix for those wishing to conduct similar studies.) In section 5,

we discuss the implications and limitations of what we have found for both research and practice, and suggest opportunities for further research.

2. Literature review

2.1 The pioneering role of lead user innovators

In the long-standing Schumpeterian innovation paradigm, the role of innovator is assigned to profit-seeking entrepreneurs and corporations, and not to consumers. Thus, Schumpeter argued that "it is ... the producer who as a rule initiates economic change, and consumers are educated by him if necessary" (1934, 65). The economic logic underlying this argument is that producers generally expect to distribute their costs of developing innovations over many consumers, each of whom purchases one or a few copies. Individual or collaborating consumers, in contrast, depend only on their own in-house use of their innovation and other types of self-reward such as learning and project fun to justify their investments in innovation development. On the face of it, therefore, a producer serving many consumers can afford to invest more in developing an innovation than can any single consumer, and so at least potentially can do a better job. By this logic, why would consumers innovate for themselves if producers can do it for them – and perhaps do it better?

However, in recent decades it has been amply proven that consumers – acting both individually and collaboratively *do* in fact develop and modify consumer products – and that their innovations do often have economic value to producers. Proof has come in the form of nationally-representative surveys of householders in ten nations to date. These surveys all show that household sector innovation – product development and modification by consumers - is a very important phenomenon. In aggregate across the ten nations measured, tens of millions of consumers have been found to be spending tens of billions of dollars per year developing and improving products they use, both individually and collaboratively (Table 1).

Table 1. Proportion of population developing or improving consumer products for personal use

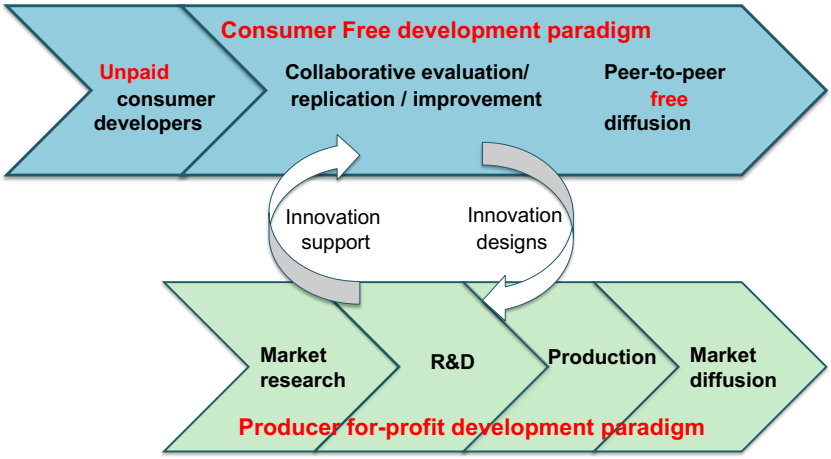
Nation	UK	USA	Japan	Canada	Finland	S. Korea	Sweden	Russia	UAE	China
% of population	6.1[a]	5.2[b]	3.7[c]	5.6[d]	5.4[e]	1.5[f]	7.3[g]	9.6[h]	3.0[i]	1.5%[j]

Data sources: [a] von Hippel et al. (2012); [b-c] von Hippel et al. (2011); [d] de Jong (2013); [e] de Jong et al. (2015); [f] Kim (2015); [g] Bengtsson (2016); [h] Fursov et al. (2017); von Hippel et al. (2017); [i] Chen, Su, et al. (2017).

Consumer-innovators have been shown to develop and modify products in essentially all areas of consumer interest and activity. Areas showing high levels of innovation map well upon major categories of unpaid time activities reported by consumers. For example, in the United Kingdom, sports, gardening, household chores, caring for children, and using computers were significant activities (Lader, Short, and Gershuny 2006). Consumers would not be expected to develop innovations they do not use, such as industrial process equipment innovations.

To make sense of the user innovation phenomenon, a "free innovation paradigm" has recently been proposed (von Hippel 2017). This paradigm, displayed schematically in Figure 1, is complementary to the producer innovation paradigm also pictured. The free innovation paradigm explains the widespread presence of lead user innovation by pointing out, first, that innovating users, being self-rewarded and serving only or primarily their own needs, need not care about the extent of the market. In sharp contrast, the extent of the market is a matter of central concern to profit-seeking producers.

Figure 1: Free innovation paradigm and producer innovation paradigm



Source: von Hippel 2017

Second, lead user innovators, defined in part as being at the leading edge of marketplace trends, generally face a situation where present demand for what they are creating is unclear, and the extent of future demand even less clear. Accordingly, it has been found that lead users tend to systematically pioneer new applications and markets – by innovating ahead of producers, as is indicated by the leftward starting point of the top arrow in the figure (von Hippel 2017, Chapter 4). At the same time, consumers' ability to invest in innovation development is being enhanced by the ability of users to collaboratively innovate, exchanging information and coordinating via

the internet. In other words, these days, users as well as producers can distribute innovation development costs over many users (potential purchasers) if they choose to do so.

2.2 Lead user consumer innovations

Recall that lead users are defined as members of a user population who (1) anticipate obtaining relatively high benefits from obtaining a solution to their needs and so may innovate and (2) are at the leading edge of important trends in a marketplace under study and so are currently experiencing needs that will later be experienced by many users in that marketplace (von Hippel, 1986). It has been found that lead users are also more likely than average users to develop products with potential commercial value in both consumer and industrial fields (von Hippel 1986; Urban and von Hippel 1988; Franke, von Hippel, and Schreier 2006; Hienerth, von Hippel, and Jensen 2014, table 3). Research has also shown that as the strength of each of the two lead user attributes increases within individuals, the likelihood that an individual will innovate and the likely commercial value of the innovation developed both increase (Morrison, Roberts, and Midgely 2004, Franke et al 2006).

Two studies have quantitatively compared the outputs of lead-user idea-generation studies with the outputs of traditional market research studies that focus on target market customers rather than lead users. Both studies, carried out in industrial product fields, found that the ideas generated by a process using inputs from lead users have much higher commercial attractiveness (Urban and von Hippel, 1988; Lilien et al., 2002). Lilien et al. also found lead-user studies capable of systematically identifying breakthrough innovations for a firm conducting the studies, where breakthroughs were defined as new product lines providing new sales representing over 20 percent of total existing sales. Many other studies have made qualitative comparisons with results that tend to favor the value of lead user methods and innovations. For example, a comparison of new baby feeding product ideas developed by users scored significantly higher than ideas proposed by in-house producer experts in that field – in the view of the producer experts themselves (Poetz and Schreier 2012).

Research on consumer *motives* for innovating shows that 91% of consumer-innovators are motivated by self-rewards, and not by expectations of financial profit. That is, when asked why they invested effort in their projects, 91% of them say they were not motivated by money. Instead, major motivational categories generally reported involved rewards that they gave to

themselves, no transactions required: their own individual need for the product; their own individual learning and enjoyment derived from developing the innovation; and altruism (von Hippel 2017, Chap 2).

Innovating consumers who are self-rewarded are very generally willing to reveal details of their innovations without patents or other forms of intellectual property protection on the web or elsewhere. The consequence is that, when others discover the innovations, they can adopt them "for free" without payments of any kind. Why is free revealing acceptable to self-rewarded innovators? The first fundamental point to note is that household sector innovators that are not rivals, and who do not plan to gain from having a monopoly on their innovations, do not lose anything by freely revealing their designs. For example, if I develop an innovation to help my diabetic child and have no interest in selling it, my own interests are in no way damaged if you adopt my design to help your diabetic child too without paying me. This is true even if you did not contribute to the development work—that is, if you are a free rider. It is also true even if you are a producer who will make a great deal of money commercializing my free innovation, and who will not share any of the profits with me. After all, my self-reward—sufficient to induce me to develop the innovation—was to help my child. The second fundamental point is that, for innovators who do not lose anything by free revealing, a passive absence of efforts to protect innovation-related information is the lowest-cost option. This is so because active exclusion such as applying for and enforcing a patent would require investment (Benkler 2004; von Hippel 2017).

2.3 Lead user innovation identification methods

Studies involving identification of lead users have been done for many reasons. Sometimes the purpose is to study the attributes and behaviors of lead users themselves, such as the often-central roles they play in networks (e.g. Kratzer et al 2016). Other times, the goal is to selectively recruit them into product development contests (Füller, Hutter, and Faullant 2011) or into co-creation activities with producers (Brem and Bilgram 2015). In still other cases, the goal is to understand how to "embed" lead users into firms as valued marketing and product development employees (Schweisfurth and Raasch 2015). (For a recent and extensive review of lead user literature, see Brem et al. 2018.)

Our present goal and specific literature interest is the identification of lead user *innovations* that have already been developed, and that have also accumulated some social evidence that they are interesting to additional users as well. A range of methods have previously been used for these purposes.

Pyramiding, the initial method developed for lead user search, is a variant of snowball sampling. Snowball sampling involves asking individuals who have a rare characteristic being sought to identify others they may know who have that same characteristic (Goodman, 1961, Welch, 1975). The effectiveness of snow-balling stems from the observation that people tend to know or be aware of people like themselves. Pyramiding adds the component of hill-climbing to a snowballing search by assuming that people who have a strong interest in a subject, for example a particular sport, will tend to know people like themselves *and* people who know *more* about that subject than they themselves do. The method, developed before internet search tools were developed, used networking via telephone to contact interviewees and ask them to identify individuals with this characteristic. The interviewer and hill climbs the "pyramid of expertise" by next contacting and interviewing the individuals identified in earlier interviews. The sequence continues until lead users who have innovations to share have been identified and interviewed (von Hippel et al 1999). Researchers have tested the pyramiding search method and found that it achieves more efficient identification of lead users than does mass screening (e.g. Olson and Bakke, 2001 von Hippel, Franke, and Prügl 2009). Hyysalo and colleagues have used pyramiding to identify over 200 user innovations, verified by domain experts, in five renewable energy technologies (Hyysalo et al 2013, 2016). Hyysalo and colleagues (2015) have also pioneered multiple path pyramiding search strategies. These enable pyramiders to work in parallel, with a resulting reduction in elapsed calendar time required for a project.

An important practical drawback to pyramiding as a lead user innovation search method is that each telephone or even email interview involves significant time and expense. Experience in the field has shown that lead user projects utilizing the traditional single-path pyramiding method take about four months on average using a skilled project team of about four individuals (Churchill et al. 1998, p 30). Field experience also shows that interviewing must be carried out by *subject matter experts* who generally have many other high-priority uses for their time. In effect, interviewees near the top of the pyramid are themselves experts, and are not giving their information for free. Instead, an interviewer must reciprocate by also supplying information of

interest to the interviewee during the conversation. In other words, what is going on is "knowhow trading" rather than simple one-way information transfer. The interviewer simply cannot hold up his or her end of the bargain without also being a subject matter expert (von Hippel 1987).

The pyramiding method was developed in the mid 1990's just at the infancy of internet search method development. Now that internet search tools have become robust, and the internet itself contains massive amounts of data on almost any subject, it is clear that efforts should be made to update lead user innovation search methods. This opportunity has not escaped researchers, and several have focused on identifying lead users – but not, to our knowledge, lead user innovations – via web search techniques. Thus, Belz and Baumbach (2010) applied a netnography approach to identify lead users in an online community, and found that 22% of the most active members had lead user characteristics. They concluded that netnography could be a viable method of lead user identification, offering the advantage that it relied on external assessments rather than self-assessment data derived from questionnaires. Hyysalo et al. (2013) found concentrations of user innovations within Internet communities, by following user created categorizations of website content and manually screening innovation-related categories. Tietz et al (2006), and also Brem and Bilgram (2015) explored the efficacy of crowdsourcing methods in getting lead users to self-identify in response to a call for solutions that drew on their expertise.

Pajo et al (2015) proposed an AI data mining technique for identifying lead users. In initial exploratory work, they focused on a survey population of users who discuss or follow camera lens product related topics on Twitter. To train the AI system, they used a sample of validated lead user questionnaires containing clusters of lead and non-lead Twitter users interested in that single product. They suggest that data mining techniques such as the one they explore can minimize the resource and time costs in identifying lead users, and that future work should develop the idea further. Twitter data was also examined by Tucker and Tuarob (2015) to analyze user generated content from potential lead users who shared content around 27 smartphone models. Their findings establish that user generated content is a relevant information source for lead users identification along with information on latent product features that potential lead users share. Kratzer et al (2016) come to the same conclusion after studying lead users within social networks based on three empirical studies and found that they possess a

distinctive network position, and function as bridges between social groups due to a higher "betweenness centrality".

3. Semantic analysis of open user-generated content on the web.

In this section we will describe the application of semantic network analytic and memory model analysis methods to large data sets of user generated content openly available on the web. These data can be analyzed computationally to reveal patterns, trends, and associations at a very low cost. Note, however, that since *only* open websites can be searched – sites willing to allow anyone to examine their contents in detail – this generally means that the methods we will describe are only applicable to identifying lead user innovations developed by consumers who are willing to freely reveal them, or to innovations posted "for free" by groups like citizen and academic scientists (von Hippel 2017).

3.1 The basic process

The user innovation identification process we apply and test in this research project involves semantic filtering of user-generated content posted openly on the web. The goal is to identify the promising lead user innovations that might be described or referred to within that content. To accomplish this requires automated forms of filtering, since the body of potentially-relevant data is too large for manual screening. The method begins by "scraping" open websites for user-generated content (UGC). It then subjects that content to a series of semantic filters to isolate just the most relevant writings for detailed examination. Generally, we find, the websites most likely to contain useful descriptions of lead user innovations are sites specialized by subject matter where enthusiasts and experts gather to share information, rather than more general sites like Facebook and Twitter.

Since the user-generated content is in words, the process of screening UGC postings on the web involves searching for specific words or word combinations indicating that the topic of the UGC posting is the subject matter of interest. This is where the art of the analyst comes in. His or her job is to select or devise semantic filters that will reduce the UGC corpus down from perhaps millions of postings – many more than an analyst could read – to just the few or a few hundred candidates most likely to contain the information one is seeking.

In principle, one might think this kind of a search and filtering process is easy to do. If one is searching, say, for kiteboarding equipment innovations by lead users, why not just filter for the single word 'kiteboarding' or the entire three-word phrase? This can be done but, alas, it by no means identifies all UGC with relevant content – indeed it may only identify a small fraction. This is because, when experts talk together, they generally assume that all who might be interested in reading their post will recognize the same common context and not need it restated in each message. By way of analogy, parents discussing baby care issues on a website devoted to baby care might not even mention the word "baby" in their posts. E.g., "When feeding twins, I find it helpful to hold a second bottle in the crook of my arm so I can feed both at once." In this case, the poster is leaving unsaid that she is attempting to feeding twin babies at the same time, and that the bottles in question are baby feeding bottles with nipples that can be tipped sideways without spilling. Similarly, in the case of kiteboarding, a UGC contribution simply said: "My chickenloop twisted again, and so I tried X hack to stop that from happening. Seemed to work at least for that moment." This comment – doubtless totally opaque to most but totally clear to enthusiast kite-surfers - had to do with equipment problems that can occur when controlling a kite while surfing – a contextual matter not stated. (A chickenloop is a component of kiteboarding equipment: a strong rubber loop used to attach the kitesurfer's kite "control bar" to his or her harness.)

It is this problem that leads to the requirement to have both a subject matter expert and a data analyst skilled in semantic analysis methods working together on the type of lead user identification project we are describing here. (Of course, if the analyst is *also* a subject matter expert on a particular project, the total team count can then be one.) In practice, what the team of an analyst and subject matter expert will do is use trial and error methods informed by a basic understanding of terminology used in the field of interest. They will say to themselves, let's see what happens if I filter all the UGC I have scraped using X set of words and phrases. From our expert knowledge of the field we are searching, and various readings, we think that innovating users in field Y will commonly use these specific words and phrases.

Next, the analyst will run that semantic filter against the data. The subject matter expert will then inspect – actually read - one or two hundred of the posts or other UGC content captured in the filter. Do a substantial fraction of the posts contain information on what is wanted? If no, let's try again using other words and phrases for filtering. If yes, great. Let's see if we can make

the desired information content in the output richer still by tweaking the terms used in the previous run -perhaps adding or modifying some elements – and then run the test again. As part of each test cycle, the subject matter expert again inspects one or two hundred results from each new run. Do things look better or worse? This goes on until the output from the filters is a sample of UGC that is small enough for practical manual inspection, and also contains a usefully large faction of content relevant to the searchers goal – the user-developed innovations the analyst or his or her client is looking for.

In the process just described, to goal is to balance the trade-off between focused targeting down to the most relevant data versus excessively drastic exclusion rules that might eliminate highly relevant user generated content. The goal is to get rid of as much low-value material as possible, so that what remains will be tractable for hands-on assessment by an expert. Note that, in the process just described, one cannot ever be sure that one has identified all the user-developed innovations of interest that exist in the UGC content being filtered. One can only recognize when one has "enough" of potential interest. In other words, one satisfices, stopping the trial and error cycle when what one has found is "good enough" for the purpose.

Finally, as a last step in the particular method we have developed for the purpose of identifying valuable lead user innovations, the analyst assesses the likely popularity and general value of each lead user innovation identified by, first, analysing trends in the amount of user-generated content mentioning that innovation and, second, by analysing trends in the frequency with which that innovation shows up in search terms.

We should note that there are many valuable tools available to help project teams succeed at the trial and error process just described. For example, there is a tool often used early in the search process that identifies the frequency with which each website UGC contains terms of likely high relevance like, in our the test case we will discuss soon, "kiteboarding". In our case study, we did this by using a Google search algorithm that combines various factors to rank the relevance of a website – i.e. keyword matching: how often are the keywords searched for present on the website, the "page rank" of that website, etc.. With the assistance of this tool, one can then focus future search runs on a much smaller set of "high-frequency" sites, resulting in significant economies in processing cost and time. A second very useful tool or method helps one screen out such things as ads and other producer-supplied content in order to focus on content created by the users themselves. To accomplish this, one applies various screening

criteria. For example, UGC content is likely to be dynamic over short periods of time, suggesting active exchanges among users. In contrast, static materials, like corporate ads and product descriptions, do not tend to change as rapidly, and can be selected out on that basis.

4. Method application case study: kiteboarding

As a proof of concept test, we applied the semantic search method described in outline above to an actual case study. As was noted earlier, complete methodological details, including actual search terms used in this demonstration study, are available in a methodological appendix.

We chose kiteboarding for our proof of concept test because others have studied that sport and found users to be quite active as developers equipment innovations – at least at the time of their studies (Tietz et al. 2004, Franke et al. 2006). Therefore, if application of our semantic search and analysis methods do *not* yield evidence of user innovation, this would suggest that our methods are failing to capture lead user innovations that in fact do exist for reasons requiring further investigation.

Kiteboarding is a water sport in which the user stands on a special board, somewhat like a surfboard, and is pulled along by holding onto a large, windborne, steerable kite. Equipment and technique have evolved to the point that kites can be guided both with and against the wind by a skilled kiteboarder and can lift rider and board many meters into the air for tens of seconds at a time.

4.1 Process step overview

In overview, the major process steps we followed in this study are listed in Table 2. Again, complete process details, including filtering terms used for each step, are provided in a methodological appendix.

Table 2: Lead user innovation identification process steps

Process steps	Kite surfing study example	
Step 1	234,017 English UGC posts scraped from 9617 websites across the globe	Subject matter expert compares information retrieval methods based on search algorithms by reviewing UGC in the search field to build an initial domain taxonomy list of terms to try out. User generated data content scraping then is conducted related to the domain of interest to identify websites and relevant UGC.

Step 2a	6065 posts remain after step 2a filtering	Semantic filtering using semantic network analytics and semantic memory model techniques to identify the relevant data corpus for the innovation concept (*E.g., DIY, Innovation, Inventions, Problem Solving & Developing Solutions)* Inspection of a sample of 100+ UGC flagged by a subject matter expert to see if the analysis is on the right track. (multiple test and refinement cycles)
Step 2b	453 posts remain after step 2b filtering	Refined syntax based semantic pattern algorithm with metacharacters to isolate first person speech (*E.g., I, myself, me, we, us*) in combination with the semantic "innovation concept" (*E.g., invented, developed, designed*) Inspection of a sample of 100+ UGC output by subject matter expert to see if filtering results are improving. (multiple test and refinement cycles)
Step 3	26 LU innovations remain after manual filtering based upon two criteria of *novelty* and true user *innovation*	Expert review and validation of the total remaining content by manual reading of the remaining relatively small set of UGC to identify and differentiate true positives for LU Innovation. Material eliminated includes producer innovations, duplicates, and very minor improvements and alterations to existing kiteboarding hardware.
Step 4	5 fundamentally radical LU innovations ranked by trend analyses to assess commercial promise	Test of popularity / commercial value of identified lead user innovations via trend analysis of frequency of relevant UGC contributions related to each identified user innovation, and frequency of searches for that innovation by Google search engine users.

4.2 User innovations identified

Our scan of user-generated content within the specialist websites identified spanned 1999 to 2018. Via applying the method steps outlined above we identified 26 *functionally novel* user-developed innovations via our semantic search methods. The final 26 were screened down from a few hundred by a manual curation process that weeded out out-of-field innovations, duplicates, producer-developed innovations, and innovations that did not offer functional novelty relative to equipment already available. (For example, a DIY development by a user that was simply a cheaper copy of an item of equipment also available from a producer would be weeded out at this stage. On the other hand, the user-developed seat for disabled kiteboarders, and the beach launch system innovations would pass these screening tests. Neither function was accomplishable prior to development and use of the user innovation by kiteboarders themselves.) In a separate search, it was determined whether or not each kiteboarder-developed innovation identified had already been adopted as a commercial product by a producer firm.

In table 3, the 26 functionally novel, user-developed innovations are listed in order of first mention in the data base, from earliest to most recent. As can be seen in the Table, at least 12 and perhaps as many as 15 of these innovations were later commercialized by producers. This is in line with other research evidence that user-generated innovations are often or at least

sometimes of commercial value. As can also be seen, the newest innovations are the least likely to have been commercialized at the time of data collection, suggesting that application of this semantic search method can indeed identify new lead user innovations that may represent commercial opportunities for producers.

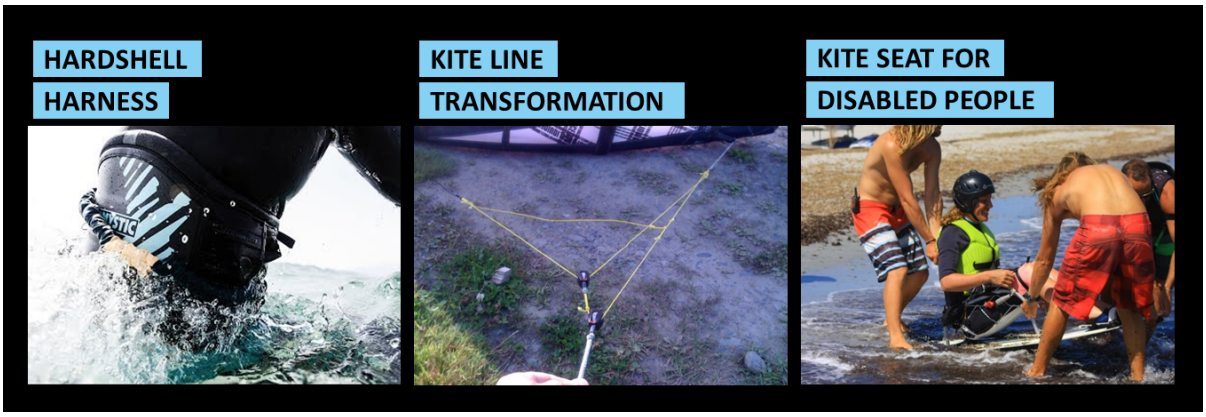
Table 3: List of identified user innovation

No#	Year First mentioned	Lead User Innovation	Commercial Availability	User Innovation
1	1999	Twin Tip Board	Yes	Yes
2	1999	4-Line Kite	Yes	Yes
3	2003	(KPO) Kite Power Optimizer	Yes	Yes
4	2003	Aquaskipper	Unclear	Yes
5	2006	Chicken Loop Anti-Twist	No	Yes
6	2006	Kite Release	Yes	Yes
7	2007	Quick Release Shackle	Unclear	Yes
8	2009	Kitefoiling	Yes	Yes
9	2009	Beach Self Launch System	No	Yes
10	2009	Disabled People Seat for kiteboarding	No	Yes
11	2010	Hardshell harness	Yes	Yes
12	2011	Foam Core Board DIY	Yes	Yes
13	2013	Strutless Kite	Yes	Yes
14	2014	Push Away Release Mechanism	No	Yes
15	2014	5th Line Kite DIY	No	Yes
16	2014	Carbon Hydrofoil Board	Yes	Yes
17	2014	Push Away Release Mechanism	No	Yes
18	2016	Velcro flexible strutless kite	No	Yes
19	2016	Click Bar	Yes	Yes
20	2016	Carbon Foil Bars	Yes	Yes
21	2016	Split Board Design	Yes	Yes
22	2016	Hydrofoil-specific kite	No	Yes
23	2016	Sit-On-Top Kite Race Board	No	Yes
24	2016	Drone Kiteboarding	No	Yes
25	2017	Electric hydrofoil	Unclear	Yes
26	2017	Foil Pumping	Unclear	Yes

As a matter of additional interest, of the 26 innovations identified, the first 21 chronologically were user-developed improvements to kiteboarding equipment that were within the confines of kiteboarding practice as commonly understood: a person standing on a board floating on the water, being pulled along and perhaps lifted out of the water by use of a steerable kite. Three of these are depicted in Figure 2. The first is a "hardshell" harness that better distributes the pulling force of the kite across the kiteboarder's body. The second is an alteration to the geometry of the rope lines connecting the surfer to the kite to improve his or her control of the kite's direction of motion and power. The third is a special seat attached to a standard kiteboarding board that for the first time enables people with certain disabilities to participate in the sport.)

Figure 2: Examples of kiteboarding equipment improvement innovations

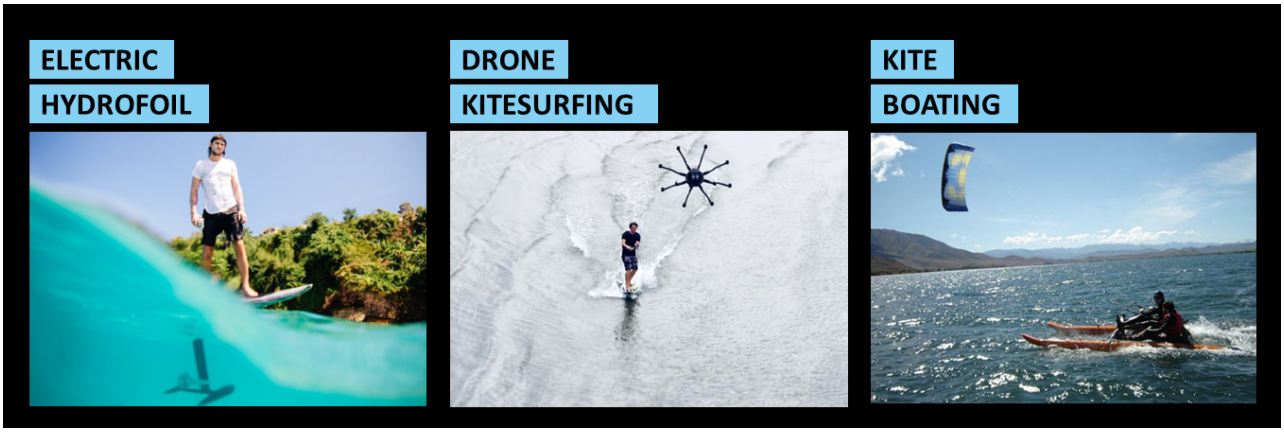
In addition, the search identified five innovations that were more radical with respect to current practices within the sport of kiteboarding. These were the 5 most recent innovations listed in Table 3. It seems to us that each of these significantly altered the nature of the sport. Three are illustrated in Figure 3. The electric hydrofoil eliminates the kite as the source of motive power, substituting an electric motor (mounted on a hydrofoil under the board). The second replaces the con



ventional kiteboarding kite as a source of motive power with a powerful drone flying overhead. Note that, especially in the second example, some of the equipment and some of the surfing techniques are carried over from the standard sport. However, new degrees of freedom are

gained, because one is no longer at the mercy of wind conditions. For example, one can play the sport even under dead-calm conditions.

Figure 3: Examples of innovations fundamentally shifting the nature of the sport



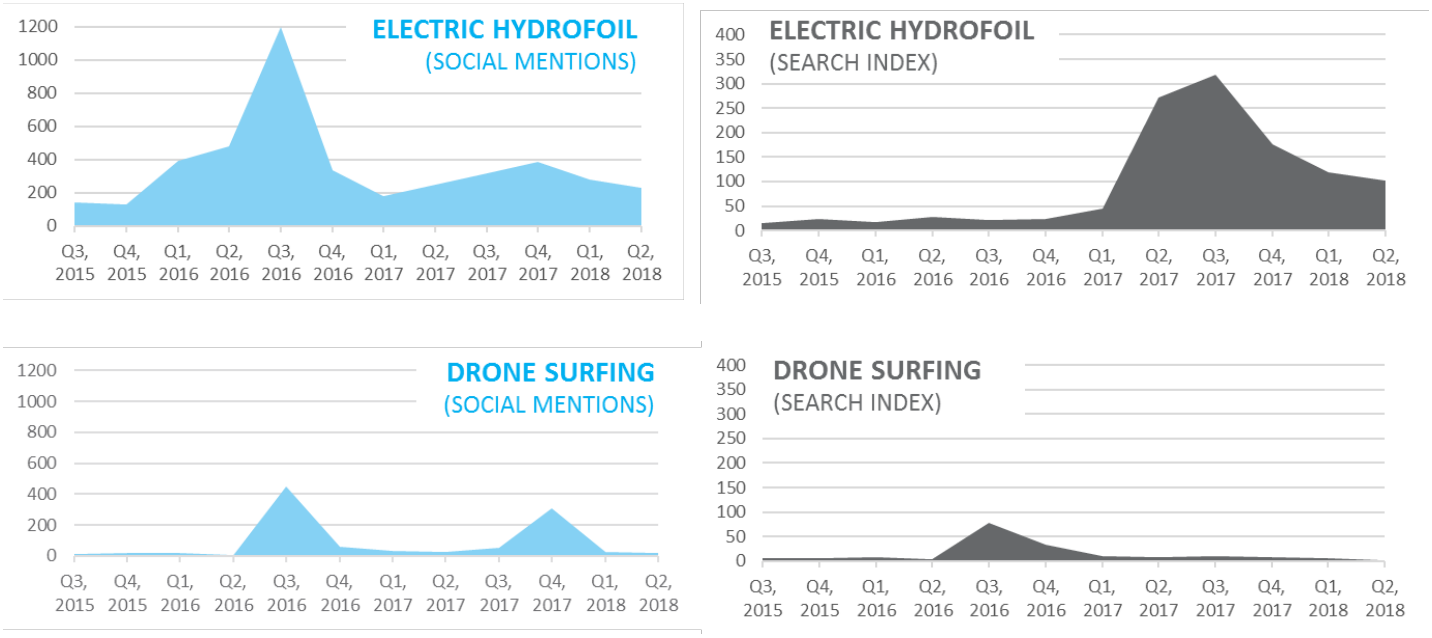
Finally the third example was also identified via our search but excluded from Table 2 as off-topic. It clearly is a user innovation, but maybe not properly coded as a kiteboarding innovation. Indeed, it may be of more interest to sailboat producers than to kiteboarding equipment producers. In this third example, a kiteboarding kite is applied to pull a boat instead of a board. In principle, this could represent a radical new direction for sailing – a "sail" that is in the air high above a boat can access different and often more powerful sources of wind energy than can a sail attached to a mast on the boat itself. Researchers conducting lead user innovation searches can include or exclude "out-of-topic" innovations as they wish. Some might find them not relevant for their project purposes, other might find them to be exceedingly valuable indicators of promising new market directions.

4.3 Examples of trends in social data and in search term usage for each innovation

Analyses carried out in step 4 give some information on the potential commercial value of each innovation in the form of frequency of mentions in UGC on websites, and also in the form of an index of frequency of the innovation being mentioned as a search term by users of the Google browser. (Due to designed-in constraints on Google search trend tools, actual frequencies of search queries are not available – only the ratio by time period relative to a long-

term average.) As can be seen in Figure 4, the electric hydrofoil innovation appears to be attracting more user interest than does the kite-powered boating innovation.

Figure 4: Trend data for two recent sport-extending innovations



The trend analysis allowed us to assess how the diffusion of social mention and searches interrelated to each other in the case of fundamentally radical LU innovations. As it is visible for Electric Hydrofoiling, we see the first major uplift in social mentions in Quarter 4 2015 where experts ahead of the trend exchange ideas on the innovation in kiteboarding web communities. The social media conversation climax was reached in Quarter 3 2016. The first visible increase in search around Electric Hydrofoiling started later in Quarter 1 2017. Even though the delay effect between social mentions and searches was shorter for some of the radical LU innovations like Drone Surfing, the trend data analysis from the kiteboarding study suggests that LU innovation trends start off with expert discussions in domain specific forums with those who are ahead of the curve before they become visible in search data when the subject matter interest spreads to the broader audience.

4.4: Other indicia of commercial potential

Once a user-generated innovation has been identified, innovation-specific searches can be carried out to gain more information regarding both the innovation and its commercial potential. Each UGC posting discussing each innovation has the identity (pseudonym or actual) of the individual poster. Privacy regulations permitting, these individuals can be contacted by researchers to learn more. Further, searches can be conducted to identify whether the innovation has been commercialized already. Often, in this field, the first to commercialize an innovation are innovators themselves, generally founding a small company to accomplish this. Incumbent firms will be interested to know this information and the commercial success attained by these small firms as input to their own commercialization strategies, should they elect to go in that direction.

5. Discussion

The focus of this article is on how to more efficiently *find and evaluate* lead user-developed innovations. To do this, we have applied semantic network analytic and memory model techniques to the specific task of identifying recent user innovations developed by consumers. Note that the method we have described does not first search for lead users and then identify innovations these individuals may have developed. Instead, it directly searches for user innovations, and only as a second step assesses their commercial promise. Additional analyses can assess the lead user characteristics of the developers identified too, if of interest, but the correlation between lead user-developed innovations and commercial promise has already been well established via previous research, as was discussed previously.

Conceptually, as we said earlier, the core of the method we describe uses semantic network analytic and memory model techniques. It is a trial-and-error process, with semantic filtering terms and phrases being tried out for effectiveness via test runs, and then progressively refined to improve performance. We have provided a successful case example of the method in use in the body of this article, and also described our methods in full detail in a methodological appendix. The promise demonstrated in our test case suggests, we think, both that the method we describe can be practical today, and also that additional methods-related research and development on this topic would be merited.

To identify and apply concepts and terms specific to a particular topic (e.g. haircare, kiteboarding) we have found it very important to have a subject matter expert on call to assist the data analyst. During our development and test efforts, we repeatedly found that even earnest and effortful attempts to read and understand user-generated content by a data analyst did not make that person sufficiently expert for the purposes of our method. Repeatedly, when we told real subject matter experts what we had gleaned from our UGC readings, they would look at us with disbelief: "How on earth did you conclude *that* from those posts? Totally wrong – the real underlying issue being discussed here is X." The situation is akin to, and the solution is akin to, that involved in what is called "agile programming" in the field of software development. In that field, it is often the case that individuals expert in programming cannot understand the larger user system they are tasked to improve well enough to create successful code. Accordingly, it has been found useful to add a subject matter expert to the programming team who does deeply understand the larger system. This individual can provide timely clarifications to the programming experts as needed, and also has access to the real system and context of use to run timely tests of prototypes being developed.

The output of the method we have described and demonstrated is a set of user-developed innovations, with potential general promise indicated by trend data. Also, knowledgeable individuals who have contributed to the discussion of each innovation can be identified by the names or pseudonyms with which they sign their openly-available posts. In follow-up steps, individuals or firm employees may wish to contact and talk with the user developers themselves – privacy laws and other considerations permitting - to better understand the innovation and setting in which it is used, and perhaps to get help with further development. They may also choose to do market studies to develop a business case to commercialize the innovation, seek approvals for commercialization, etc. as next steps.

Following the practical procedure(s) we have described is likely, we think, to enable innovation developers to identify lead user innovations in consumer product and service areas much more efficiently than can be accomplished via traditionally-employed pyramiding techniques. Our experience with a single example suggests to us that an analyst skilled in the techniques we have employed here, working together with a subject matter expert, can hope to move down the learning curve to the point where a lead user study of the type described here can be completed in about a week. This is about 20X faster than traditional methods described in

Churchill et al. (1998) and so will, we think, make lead user innovation identification studies financially viable in a wider range of conditions than is presently the case.

In addition, it is important to note that the application of semantic search methods to lead user innovation search will yield much broader results than the pyramiding search procedure described in the literature review. In the pyramiding process, process users *prespecify* important market trends and only search for lead users with respect to those trends. This is done to make the amount of manual search effort tractable. Lead users can be interested in different attributes of a product, and may only know those with similar interests. For example, a lead user interested in a higher performance kite may not know lead users interested in making kiteboarding more accessible to the handicapped. One generally cannot afford to conduct pyramiding interviews along all of these dimensions. In contrast, the method we describe *directly* searches all user-generated content on the web user-generated innovations *independent* of any trend-related considerations. Then, it applies trend data analyses to the set of innovations discovered with a domain, to identify those likely to be of the highest commercial importance. This results in a much broader search – a very valuable matter.

5.1 Implications for research

In this paper, we have described and initially tested the application of semantic network analytic and memory model techniques to screening of user-generated content for lead user innovations. An immediate question, of course, is whether our success in a single case can be replicated further – especially in categories of wider general interest than the niche category of kite-surfing equipment innovations. Further research is required to explore this matter, but we see no reason why the method will not prove widely applicable. Indeed, to explore that matter in a preliminary way we did a brief exploration in the commercially important category of hair care products.

In the case of hair care, we again found many lead user innovations both with respect to improvements to existing practice, and also at the fringes of current business practice. With respect to improvement innovations we found a major theme in consumer communities focused on hair care was trying out and exchanging information regarding a wide range of organic ingredients. The general motivation of these innovating consumers appeared to be to replace artificial surfactants and other basic ingredients used in commercial products like shampoos with

organic substitutes. With respect to more radical innovations, we saw that users were creating their own, radical styles of haircuts; novel color shades (e.g., grey, a color not offered commercially at the time of first user experimentation); and also such things as super-strength hair hold compounds made from sugar in order to create longer-lasting results in the case of extreme hair styles like mohawks. This product-specific information was mixed with technique information, such as how to style a mohawk, that will be potentially useful lead user innovation information for producers of related *services*, such as hair-styling.

When the question of general applicability of the basic method has been proven by experience, we think it would be very useful if researchers created toolkits that less-expert practitioners could apply. These would include generally-available open source semantic analysis tools, and would provide instructions on how to apply them in the specific context of lead user innovation identification studies. The toolkits could also provide words and phrases found to be generally useful as elements of semantic filters for lead user innovation searches, along with instructions for identifying and adding topic-specific words and phrases. For example, some of the words and phrases we found effective in our specific case study are likely to be generally useful across many study topics (e.g., "I invented, I designed"). Others are specific to a particular topic (e.g., "chickenloop" in the case of our study of kiteboarding). An important implication is that semantic filters need not be developed completely from scratch for each study. Instead, it is very likely that researchers can identify a general set of filter words and phrases that will be useful across a wide range of study topics. Subject-matter experts on the practitioner team need only then add subject-specific words and phrases to complete an effective filter set. This should significantly ease the task of designing and executing specific studies.

In general, the research we describe can build upon more basic developments in semantic filtering and memory model techniques. However, it is not impossible that learnings from lead user innovation identification studies can also inform more basic semantic analysis research.

5.2 Implications for practice

Recall that the lead user innovation search method we have described and tested is not just a lower-cost version of the pyramiding technique previously recommended as best practice. It is actually producing broader results, and doing so much faster than traditional pyramiding methods. The rich trove of user-generated content on the web these days combined with the low

cost of present-day web search and analysis methods today is what makes these performance improvements possible. This favorable turn of events should in turn should make lead user innovation studies more widespread than they have been. What is happening very generally today, as was briefly described in section 2.1, is that companies are recognizing that lead users are very good at early identification of needs at the leading edge of markets. In the case of consumers, some also have sufficient capabilities to develop very interesting early product and service prototypes.

To make these techniques as valuable as possible, it is also important to learn how to incorporate the lead user innovation identification practices we have described into corporate product development practices. To do this, companies as a whole, and product and service developers in particular, , need to learn a new division of innovative labor. In-house developers should learn that, in most instances, they should no longer assume that it is their task to develop innovative product concepts *for* consumers. Instead, they should reallocate resources formerly devoted to that task to the identification and evaluation of concepts developed and prototyped *by* lead users. This may seem like a threatening change to many in-house developers – but there will always be a great deal of creative work remaining for them. After all, they are the ones who must carry out lead user innovation search projects. In addition, they can apply existing conventional product concept evaluation techniques to determine how to improve user prototypes to a form that is suitable for the general market (von Hippel 2017).

References

Belz, F. M. and Baumbach, W. (2010). Netnography as a method of lead user identification. *Creativity and Innovation Management*, 19, 3: 304–313.

Bengtsson, Lars (2016) How big and important is consumer innovation in Sweden? – A comparison with five other countries, Lund University Working Paper.

Benkler, Y. 2004. Sharing nicely: On shareable goods and the emergence of sharing as a modality of economic production. *Yale Law Journal* 114 (2): 273–358.

Brem, A. and Bilgram, V. (2015). The search for innovative partners in co-creation: Identifying lead users in social media through netnography and crowdsourcing. *Journal of Engineering and Technology Management*, 37, 40–51.

Brem, Alexander, Volker Bilgram, and Adele Gutstein (2018) Involving Lead Users in Innovation: A Structured Summary of Research on the Lead User Method *International Journal of Innovation and Technology Management* Vol. 15, No. 3 (2018) 1850022 (27 pages) https://www.worldscientific.com/doi/abs/10.1142/S0219877018500220

Businsswire (2018) Kiteboarding Equipment Market Analysis by Share, Size, Drivers, and Trends: Technavio https://www.businesswire.com/news/home/20170601006445/en/Kiteboarding-Equipment-Market-Analysis-Share-Size-Drivers

Chen Jin, Yu-Shan Su, Jeroen P.J. de Jong, Eric von Hippel (2018) "Household Sector Innovation in China: Impacts of Income and Development" MIT Sloan School of Management Working Paper

Churchill, Joan, Eric von Hippel and Mary Sonnack (1998) *Lead User Project Handbook: A practical guide for lead user project teams* MIT Sloan Working Document (available under a Creative Commons license from: https://evhippel.mit.edu/teaching/

Cooper, Robert, and Scott Edgett (2008) *Ideation for product innovation; What are the best methods?* Reference Paper 29, Product Innovation Best Practices Series, Product Development Institute, Inc.

de Jong, Jeroen P.J. 2013. "User innovation by Canadian consumers: Analysis of a sample of 2,021 respondents." Unpublished paper commissioned by Industry Canada.

de Jong, Jeroen P.J., Eric von Hippel, Fred Gault, Jari Kuusisto, and Christina Raasch (2015) "Market failure in the diffusion of consumer-developed innovations: Patterns in Finland." *Research Policy* 44, no.10 (December): 1856-1865.

Franke, Nikolaus, Eric von Hippel and Martin Schreier (2006) "Finding Commercially Attractive User Innovations: A Test of Lead-User Theory" *Journal of Product Innovation Management* Vol 23 pp. 301-315.

Füller, J., K. Hutter, and R. Faullant (2011) "Why co-creation experience matters? Creative experience and its impact on the quantity and quality of creative contributions." *R & D Management* 41 (3): 259–273.

Fursov, Konstantin, Thomas Thurner, and Alena Nefedova (2017) "What user-innovators do that others don't: A study of daily practices" *Technological Forecasting & Social Change* 118 153–160.

Goodman, L. A. 1961. "Snowball sampling." *Annals of Mathematical Statistics* 32 (1): 117–151.

Hienerth, Christoph, Eric von Hippel, and Morten Berg Jensen (2014) "User community vs. producer innovation development efficiency: A first empirical study," *Research Policy* 43 p. 190-201.

Hyysalo, S., Johnson, M., Juntunen, J.K. (2016) The diffusion of consumer innovation in sustainable energy technologies, *Journal of Cleaner Production*, 162: S70-S82.

Hyysalo, S. , Juntunen, J. & Freeman, S (2013). User innovation in sustainable home energy technologies. *Energy Policy* 55 (2013) 490–500.

Hyysalo, S., Helminen, P., Mäkinen, S., Johnson, M. Juntunen, J. & Freeman, S (2015) Intermediate Search Elements and Method Combination in Lead-User Searches. International Journal of Innovation Management. 19 (1-37).

Kim, Y. 2015. "Consumer user innovation in Korea: An international comparison and policy implications." *Asian Journal of Technology Innovation* 23 (1): 69–86.

Kratzer, Jan, Christopher Lettl, Nikolaus Franke, and Peter A. Gloor (2016) "The Social Network Position of Lead Users" *Journal of Product Innovation Management*; 33(2):201–216
DOI: 10.1111/jpim.12291

Lader, D., S. Short, and J. Gershuny. 2006. *The Time Use Survey, 2005: How We Spend Our Time*. Office for National Statistics, London.

Lilien, Gary L., Pamela D. Morrison, Kathleen Searls, Mary Sonnack, Eric von Hippel, (2002) "Performance Assessment of the Lead User Idea Generation Process," *Management Science*, Vol 48, No 8 (August) pp. 1042-1059.

Morrison, P. D., J. H. Roberts, and D. F. Midgley. 2004. "The nature of lead users and measurement of leading edge status." *Research Policy* 33 (2): 351–362.

Olson, E. L., and G. Bakke. 2001. "Implementing the Lead User Method in a High Technology Firm: A Longitudinal Study of Intentions versus Actions." *Journal of Product Innovation Management* 18, no. 2: 388–395.

Pajo, Sanjin, Paul-Armand Verhaegen, Dennis Vandevenne, Joost R. Duflou (2015) "Fast Lead User Identification Framework" *Procedia Engineering* 131 (2015) 1140 – 1145

Poetz, M. K., and M. Schreier. 2012. "The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?" *Journal of Product Innovation Management* 29 (2): 245–256.

Schumpeter, J. A. 1934. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Harvard University Press. Originally published in German in 1912; first English translation published in 1934.

Schweisfurth, T. G., and C. Raasch. 2015. Embedded lead users: The benefits of employing users for corporate innovation. *Research Policy* 44 (1): 168–180.

Tietz, R., Fueller, J., Herstatt, C., 2006. "Signaling: an innovative approach to identify lead users in online communities." In: International Mass Customization Meeting 2006, Hamburg.

Tuarob, S., Tucker C., 2015. "Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks." In: Journal of Mechanical Design, Volume 137, Issue 7, 2015

Urban, Glen L., and Eric von Hippel (1988), "Lead User Analyses for the Development of New Industrial Products," *Management Science* 34, no. 5 (May):569-82.

von Hippel, Eric (1986) "Lead Users: A Source of Novel Product Concepts," *Management Science* 32, no. 7 (July):791-805.

von Hippel, Eric (1987) "Cooperation Between Rivals: Informal Know-How Trading," *Research Policy* 16: 291-302.

von Hippel, Eric, Stephan Thomke and Mary Sonnack (1999) "Creating Breakthroughs at 3M" *Harvard Business Review* 77, No.5 September-October, p. 47-57.

von Hippel, Eric, Nikolaus Franke, and Reinhard Prügl (2009) ""Pyramiding": Efficient Identification of Rare Subjects." *Research Policy* No. 38, pp 1397-1406.

von Hippel, Eric, Jeroen P.J. de Jong, and Stephen Flowers (2012) "Comparing business and household sector innovation in consumer products: Findings from a representative survey in the UK." *Management Science,* Vol. 58, No. 9, (September), pp. 1669–1681.

von Hippel, Eric, Susumu Ogawa, and Jeroen P.J. de Jong (2011) "The Age of the Consumer-Innovator" *Sloan Management Review* (Fall) vol. 53 Nr 1 pp. 27-35.

von Hippel, Eric, Jeroen P.J. de Jong, and Daan Rademaker (2017) "Household Sector Innovation" Mohammed Bin Rashid Centre for Government Innovation, UAE, (July)

von Hippel, Eric (2017) *Free Innovation* MIT Press, Cambridge, MA

Welch, S. 1975. “Sampling by referral in a dispersed population.” *Public Opinion Quarterly* 39 (2): 237–245.

Methodological appendix

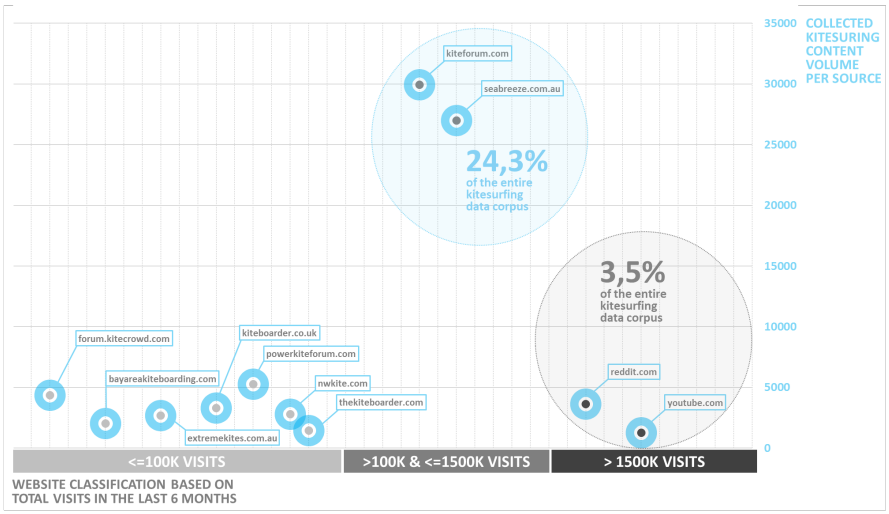
The promise of semantic network analytic and memory model techniques for identifying commercially-promising user innovations were demonstrated in the field of kiteboarding equipment. In this methodological appendix, we step through the detailed process we used in that case study.

Step 1: Scraping the open web for relevant user-generated content

The first step of the study process was to collect all of the relevant user generated data in the field of interest leveraging key terms and expressions within the kiteboarding sport domain. This data is the foundation for further analytic steps to identify lead user innovation within the relevant data corpus. The process to gather this information must ensure that niche web sources are included, outside the social media mainstream, such as expert forums where potential lead users and experts gather to share their expertise, professionalism and passion.

In our kiteboarding study, we found after collecting and analyzing the collected data, that more than 90% of the consumer generated content was generated in specialist forums and other niche sources such as kiteforum.com, seabreeze.com.au or powerkiteforum.com. Content from large social media or digital touchpoints with large traffic on a monthly basis such as YouTube, Reddit, Twitter or Facebook play an insignificant role as exchange source between experts.

Figure A-1: UGC volume per source with monthly traffic classification



In fact, as can be seen in Figure A-1, only two specialized forums, kiteforum.com and seabreeze.com.au, contributed to 24.3% of the overall user generated content for our study. In order to be sure to identify these critical niche web sites where experts gather and exchange, we used a user generated content gathering algorithm. This captures content from all websites with relevant user generated content publicly available to Google's search engines in the field of kiteboarding. (The reason we used algorithms that tap into Google search functionalities was so we could detect and collect user-generated content that is located at web sources that are classified as most relevant by Google's web indexing and page ranking algorithms.) The consumer generated content scraping algorithm automatically removes content irrelevant to our research purpose such as ads or surveys. It also contains a duplicate detection system to remove copied and duplicate posts from the dataset, ensuring content is only collected once. (Note, we collected only posts written in English to avoid translation difficulties in attempting to understand technical posts written in other languages.)

The process we describe starts with finding the most relevant set of sources where user generated content is present. This process step requires various iterations, comparing a "narrow search" first and compared it to a with a "broad search", to find the ideal search algorithm that retrieves the most relevant content universe for the main purpose of the study – identifying lead user innovations. The comparison between the two different approaches requires an expert validation of a random sample of n ~100 comments from both of the search methods with the aim to identify which methodologic approach leads to be most relevant data universe. The goal of this expert validation step, indicated in Figure A-2, is to learn if the narrow search might be

too targeted so that we miss important content containing relevant kiteboarding comments, or the broad search might be so wide that we collect data not related to kiteboarding at all.

Following this approach, we started off with a "narrow search" method that simply searched for content that entailed "kiteboarding" with alternative ways of spelling such. This approach eventually proved to be too narrow as we found that that more than 50% of the collected content did *not* include general keywords such as "kiteboarding". That is, searching directly for "kiteboarding" missed out content that was posted as a response to questions from peers seeking equipment-related advice where the term "kiteboarding" was not necessarily mentioned.

Figure A-2: Search query method validation steps used for the kiteboarding study

Step 1: Search query method validation			
Search Method Tested	*Search Method Description*	*Search Term Composition*	*Search approach effectiveness after expert validation*
'Soft' Search Approach	Search universe aiming to collect user generated content in the broader context environment of kiteboarding.	*First Person Speech* ("I" OR "my" OR "me …) **AND** *broad kiteboarding context ("kite" OR "water" OR "sea"..)*	**Low Precision & High Recall** The data corpus proves to be inefficient and too broad for the research purpose as it includes user generated content from other sports, vacation content etc.
'Hard' Search Approach	Search universe aiming to collect user generated content in the direct topical context of kiteboarding.	*First Person Speech* ("I" OR "my" OR "me …) **AND** direct *kiteboarding expressions ("kite surfing" OR "kiteboarding" OR "kiteboarding"..)*	**Low Recall & High Precision** The data corpus proves to be too targeted and insufficient for the research purpose as it limits the content universe to comments that include the direct expression of the sport domain. Using this search approach would miss out content on specific equipment aspects, surf styles and other relevant innovation fields where the term "kiteboarding" is not directly mentioned.

'Combined Search Approach'	Search universe aiming to collect user generated content in the broader context environment of kiteboarding and also include content on specific kiteboarding topics (e.g. equipment aspects, surf styles, surf tricks etc.)	*First Person Speech* ("I" OR "my" OR "me …) **AND** *broad kiteboarding context ("kite" OR "water" OR "sea"..)* **AND** specific kiteboarding related topics ("harness" OR "chicken loop" OR "safety floats" OR "lines")	**High Recall & High Precision.** The data corpus proves to be too most efficient as it collects user generated content that is broad enough to capture the entire kiteboarding domain including expression that don't entail the term "kiteboarding" but includes kiteboarding related topic such as equipment features or surf styles.

The alternative "broad search" method is based upon a taxonomy of terms and expressions (shown in Figure A-3) referring to kiteboarding equipment features, surf styles, tricks and kiteboarding jargon that we identified based on expert desk research. Using the broad search method, we found that individual equipment terms such as "board" or "harness" are equally ineffective due to conflations of the desired data universe with content from other sport domains. For example, the term "board" exists in many sport domains such as skateboarding or wakeboarding, and "harness" is a critical equipment component in windsurfing as well as kiteboarding.

The best search algorithm that was ultimately used for the study proved to be one that coupled expression types to define the overall domain with terms (such as kite, board, wind, sea etc.) as basis with a variety of granular expression types such as equipment aspects (e.g. chicken loop, flying lines or bridles), surf styles (e.g. freeride, airstyle or toeside) or tricks (jibing, back mobe or s-bend) to capture true expert content well.

Figure A- 3: Search taxonomies for kiteboarding topics

Step 1: Search taxonomies for kiteboarding topics					
EQUIPMENT TAXONOMY	**SURF STYLE TAXONOMY**	**TRICK TAXONOMY**	**HANDLING TAXONOMY**	**JARGON TAXONOMY**	**PIONEER TAXONOMY**
Board	Hang time	Jibe / Jibing	Depower	Boston Valve	Bruno Legaignoux
Twin-tip	Boosting	Toe turn jibe	Relaunch	Bar throw	Mike Waltze
Wake Style	Gliding	Jump jibe	Bar pressure	Trim	Gerry Lopez
Race	Quiver	Heel turn jibe	Wind range	Leading Edge	Bill Roeseler
Wave	Airstyle	Toe down riding	Overpowered	Trailing Edge	Cory Roeseler
Harness	Old School	Floaty Jump	Powered-Up	Struts	Legaignoux brothers
Chicken Loop	Airtime	Grab	Power Zone	Windward	Manu Bertin
Safety Floats	Airborne	Spin	Sheeting	Leeward	Ian Day
Bar	Air pass	Board-off	Underpowered	Upwind	Don Montague
Lines	Surface pass	Invert	Upwind	Downwind	Robby Naish
Harness	Freestyle	Kite Loop	Wind window	Wind range	Joe Keuhl
Pump	Toeside	Back Loop		Hybrid	Flash Austin

Flying Lines	Water launch	Back Mobe		Hindenburg	Robby Naish
Foil / Foil kite	Water starting	Back to Wrapped		Dawn Patrol	Flash Austin
Kite	Beach starting	Blind Judge		Downwinder	Lou Wainman
SLE Kite	Body dragging	Back Loop		Guinea pig	Elliot Leboe
Foil Kite	Looping Spinning	Front Loop		Kitemare	Laird Hamilton
C Kite	Speeding	S-Bend		Luff	Mango Carafino
Hybrid Kite	Tow In	F16		Lofted	Gijsbertus Panhuise
Bow Kite	Backside	Front to Blind		Overhead waves	Don Montague
Wetsuit	Top Turn	Front to Wrapped		Schlogging	Neil Pryde
Wing	Carving Turn	Handle Pass		Sideshore	Laurent Ness
Strut	Top to bottom	Hooked Back		Side onshore	Franz Olry
Bridles	Foiling	Blind Ole		Stomp	Rob Douglas
Control Bars	Hydrofoil	KGB		Tack	Nicolas Parlier
Stopper Bar	Kite foiling	Indy Front Loop		Tea bagging	Christophe Martin
Tuflites	Sliding	Krypt		Walk of Shame	Carlos Mario
……..	……..	……..		……..	……..

Steps 2a and 2b: Semantic analysis for lead user innovation filtering

The overall collected user generated content universe for our kiteboarding study that was gathered in step 1 consisted of 234,017 English posts associated to the kiteboarding domain from over 9617 different websites across the globe. The aim of step 2 is to isolate user innovation-related content from this large body of data. The filtering process requires text analysis tools that allow for fine-grained semantic analysis to identify specific textual patterns within user generated comments. More approximate approaches, for example a simple text mining process to identify terms such as "innovation" or "invention," would not be very effective for two reasons. First, the retrieved content could include a great deal more than self-made innovation content. It might, for example, mainly consist of producer driven innovations. Second, as we found, the direct search for "innovation" or "invention" was too narrow and did not well reflect the natural language used when kitesurfers talk about their innovations.

Our analysis showed after process iterations and optimizations in the analysis steps that the most effective semantic algorithm needed to include all possible natural speech variations to isolate the "Innovation Concept" based on words and expressions that reference to innovation such as created, design, DIY, homemade, developed. We also needed to come up with something combining those words and phrases with "first person speech" to ensure that the comment is about user innovation (e.g. "I developed") but not manufacturer innovation (e.g. "equipment supplier developed"). For that reason, we combined two semantic analysis techniques – semantic memory modeling for filtering the "Innovation Concept" and a semantic

network analytics for filtering for first person speech within that content.

Step 2a: Filtering the "Innovation Concept"

First, we used a semantic memory model (also referred to as "word space" model or distributional semantics) to identify the relevant data for the innovation concept. The word-space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity. This method allows one to "model the semantic meaning" of certain expressions and to discover hidden semantic layers and expressions that are sometimes out of context. Within GenSim or SEMMOD, analysts today have free access to convenient Python libraries with intuitive interfaces that are ideal for the described approach, and are widely adopted. For example, this method helped us to determine that the term "wheel" has an indirect semantic relationship within the lead user innovation language used by end-users to describe their innovations processes in kite-surfing and expressing that their solution is grounded on existing kite-surfing equipment (e.g. "not reinventing the wheel). This technique also revealed further idioms related to "wheel" that may not be relevant for the analytic purpose such as "wheel" in the context for fortune telling ("wheel of fortune"), trying luck or wasting time ("spin the wheel"), being in charge of something ("behind the wheel") or being unwelcomed in a social situation ("being the fifth wheel"). The ultimate goal of this approach technique is to distill the large kiteboarding data corpus that was collected in step 1 down to the content data subset that entails innovation content. We effectively used a semantic filtering concepts related to DIY, Innovation, Creation, Development, Invention and Problem Solving that appeared to be most practical and relevant for the kiteboarding field of interest. We included various spelling variations of each aspect. E.g., our search for Creation included directly related semantic expressions such as "created", "create", "creator", "creating" and "creative" as well as indirect semantic similarities such as "co-created", "engineered" or "architected" that the word-space model automatically suggested. After applying the semantic memory model algorithm, we condensed the overall volume of 234017 posts down to 6065 posts within the data universe that relate the innovation concept. So that we found that effectively 2.6% of the collected data qualifies as potential lead user innovation content.

Step 2b: Filtering "first person speech" within the innovation concept

Next, we applied semantic network analysis approach to tie the innovation filter concept to a refined regular expression with metacharacters to isolate first person speech (I, myself, me, we, us) within the sentence structure to rule out content that might refer to innovation but not necessarily user innovation. With this approach we deconstructed sentences from the relevant lead user content universe using tokenization techniques to identify the posts that clearly indicated the lead user innovations. spaCy for Python offer a collection of natural language processing tools that are ideal to create the syntactic sentence segmentation for the described purpose.

The combination of both analytic approaches described above enabled us to reduce the already reduced kiteboarding innovation content corpus from 6065 posts down to 453 likely lead user innovation posts. After applying the second semantic filter algorithms we found that 7.5% of posts within the innovation content corpus also matched both the first-person speech criteria. The remaining content corpus of 453 now represents the semantic innovation concept used in a first-person speech. (e.g. "One thing I do not like, and have already invented a solution for, is that the chicken loop line will end up getting many twists in it if you do a lot of spins in the air in one direction.)

Figure A-4: LU Innovation Filtering Method

Step 2: LU Innovation Filtering Method			
Innovation Filtering Method Test	*Semantic Analysis Method*	*Analysis Approach Description*	*Filter Result*
1)	**"Innovation Concept" Semantic Memory Modeling**	Using word embedding and artificial neural network techniques to model the **"Innovation Concept"** ("*DIY solution...*" OR "*solving…problem*" OR "*invented...*") as filter to isolate comments that include relevant expressions for Innovations.	**6065 comments** (2.59% within the total kiteboarding data corpus) match the "**Innovation Concept**" semantic algorithm

2)	**"Innovation Concept" Semantic Memory Modelling** combined with **"First Person Speech" Semantic Network Analysis**	Using word embedding and artificial neural network techniques to model the **"Innovation Concept"** ("*DIY solution…*" OR "*solving…problem*" OR "*invented…*") as filter to isolate comments that include relevant expressions for Innovations. ***AND*** **First Person Speech** ("I" OR "my" OR "me" OR "we" …) in the close semantic "Innovation Concept" context to ensure that the mentioned innovations relate to user innovations and not producer innovations	**453 comments** (0.19% within the total kiteboarding data corpus) match the "***Innovation Concept***" **AND** the "**First Person Speech**" criteria

Step 3: Expert review of the identified lead user innovations

After applying semantic algorithms to condense the initial, unmanageably large data corpus down to 453 posts that appear highly relevant to identifying possible lead user innovations, the goal of step 3 is to determine whether the remaining content has rich user innovation content, and to learn from the context and subtext of the innovations that are present match the criteria of novelty and true user innovation. Essentially, at this point a subject matter expert needs to review the individual posts and forum discussion threads to assess the lead user innovation content against our two criteria. Even though the automated semantic filtering process proved to be highly effective in retrieving the entire spectrum of lead user innovations, this human validation step is essential to make sure that the mentioned and innovations described by users are truly novel and not alterations of existing available products or solutions. For example, we found DIY kite building kits that are innovative in nature and highly relevant for the user community but don't qualify as novel solutions as they are design modifications from existing kite types.

As is discussed in the body of the paper, the kiteboarding equipment innovations discovered via the method described here included lead user innovations that solved equipment issues such as the invention of the hard-shell harness or the kite power optimizer to innovations like a beach self-launch system or a kiteboarding seat for disabled people. The list of innovations also included those that to some extent or radically changed the nature of the existing sport by substituting, for example, high-powered drones as a source of power for the kites traditionally used within the sport.

Step 4: Trend analyses

As described before, the proposed method allows for additional analytics to understand the diffusion of user innovations in the specific domain to enable analysts to better judge the commercial attractiveness of each innovation for manufacturers. The combination of user generated mentions and user's search behavioral was used to learn about the adoption trend of the identified lead user innovations over time. A user's topical Google search can be thought of as signal of intention where kitesurfers seeks to find information around the innovation or ways to acquire the innovation. In contrast, user generated mentions of the specific innovation are signals of deeper interest where kitesurfers are interested in more detailed information, discuss the innovation with peer experts or seeking advice and instructions around the innovation. The trend analysis requires sufficient social and search data to assess the signal and evolution of lead user innovation. Furthermore, the innovation needs to have distinctive reference terms that can be searched for. This proved to be challenging for the improvement innovations that we found in our study. In many cases the inventor named his novel solution in a certain way but the kiteboarding community used different terms for the solution over time – making it difficult to build precise searches for improvement innovations. As innovation recency and speed to market are import considerations for producers, we concentrated our demonstration of trend analyses on the five latest innovations in the sample. These, as it turned out, were innovations with the potential to disrupt the established sport or introduce new directions in the domain.

We developed specific search queries to scrape user generated data around the selected innovations considering entire language spectrum around the innovation from the time period of July 2015 until March 2018. E.g. The search query for "Electric Hydrofoiling" also entailed the term "eFoil" which references to a development community around electric hydrofoiling. Similar to the process iterations described in process step 1, we compared the results from narrow and broad scraping approaches to obtain the most relevant, all-encompassing and effective search approach. After determining that the 5 search queries were satisfactory from a content relevance standpoint, we extracted the absolute mention volume on a quarterly basis.

We collected the consumer search query data around the five innovations using Google Search Trends and extracted the relative search volume to compare it against the social mention volume. Google trends provides only relative search volumes of the keywords of interest with the maximum search index value of 100 for the specified period. For multiple searches that may

require capturing all relevant content around a specific innovation, the highest volume in for a given time period is indicated as 100 and the months and searches are sized proportionally to this highest value to establish even growth benchmarks. After the search data for all five fundamental innovations were extracted and aggregated into quarters, we visualized the search and social data trends over time for the 5 innovations.

Suggested tools for semantic analysis

Tools evolve quickly, but at the time of writing, we suggest the following free semantic analysis packages that are openly available to all.

- OpenNLP for R (A machine-learning based natural language processing toolkit supporting a broad variety of common analytic needs such as parsing, tokenization or sentence annotations.
- Ggraph for R (A popular and user-friendly R package for advanced semantic analysis using co-occurrence statistics. This useful toolkit reveals how expressions and topics in the specific domain of interest are semantically related in the authentic usage context.)
- LDAvis for R (A topic modeling package to explore the user generated data corpus interactively by the most salient expressions clustered into topic concepts by meaning.)
- spaCy toolkit for Python offers a variety of NLP tasks like tokenisation, part-of-speech tagging, entity recognition, dependency parsing, sentence recognition and Syntax-driven sentence segmentation
- GenSim for Python (Python libraries for building and exploring distributional semantic models using vector space representations of words and retrieving semantic similarities from unstructured data.)