



MIT Open Access Articles

*Next-generation consumer innovation search:
Identifying early-stage need-solution pairs on the web*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

As Published	10.1016/J.RESPOL.2020.104056
Publisher	Elsevier BV
Version	Final published version
Citable link	https://hdl.handle.net/1721.1/134143
Terms of Use	Creative Commons Attribution 4.0 International license
Detailed Terms	https://creativecommons.org/licenses/by/4.0/



Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Next-generation consumer innovation search: Identifying early-stage need-solution pairs on the web

Eric von Hippel^{a,*}, Sandro Kaulartz^b^a MIT Sloan School of Management, United States of America^b IPSOS, France

ARTICLE INFO

Keywords:

Problem-solving
Need-solution pairs
User innovation
Consumer innovation
Household sector innovation

ABSTRACT

All innovations consist of a need paired with a responsive solution - a need-solution pair (von Hippel and von Krogh 2016). Today, technical advances in machine learning techniques for natural language understanding, such as semantic word space models and semantic network analytics, have made it practical to capture descriptions of early-stage, need-solution pairs mentioned anywhere in the open, textual content of the Internet. Producers - and anyone - can now thus look for user innovations posted on the web that may involve either known or newly defined needs coupled to new solutions that are gaining traction. This is important because, as is now understood, users, rather than producers, tend to pioneer functionally new products and services for which both the need and the solution may be novel.

In this paper, we demonstrate via a case study both the practicality and the value of searching for early-stage need-solution pairs via machine learning methods and assessing the likely general interest in each user-generated innovation by also identifying the trends in posting and query frequencies related to it. The new need-solution pair search method we describe and test here can, we claim, serve as a very valuable complement to traditional market research techniques and practices.

1. Introduction and overview

All innovations consist of a need paired with a responsive solution – a need-solution pair (von Hippel and von Krogh 2016). Today, technical advances in machine learning techniques for natural language understanding, such as semantic word space models and semantic network analytics, have made it practical to capture descriptions of early-stage, need-solution pairs mentioned anywhere in the open, textual content of the Internet. Producers – and anyone - can now thus look for user innovations posted on the web that may involve either known or newly defined needs coupled to new solutions that are gaining traction. This is important because, as is now understood, users, rather than producers, tend to pioneer functionally new products and services for which both the need and the solution may be novel (de Jong et al. 2015; von Hippel 2017, ch. 4).

Producers, we claim, will find it valuable to discover and understand consumer-created new products and services at an early stage in their development and diffusion. After all, market research has long shown that early entrants with respect to introducing innovations onto the market can reap lasting market share benefits (Urban et al., 1986).

How can producers gain early information on consumer-created innovation opportunities affecting product categories in which they have an interest? They can do this by *not* restricting their researches to needs already widespread among customers: needs that standard marketing research methods are designed to discover. Instead or in addition, they should look for user innovations that may involve *newly defined* needs coupled to *new* solutions. Then, they should filter innovations identified via indications of emerging general demand – increases in related queries and web searches, for example. Taken together, information from these two types of searches will identify user-developed innovations that show early indications of foreshadowing general demand – innovations, in other words, developed by lead users.

In this paper, we describe a natural language processing (NLP) based search approach that combines semantic word space models with semantic network analytic methods. The approach is, as we will show, capable of screening the entire web for open, user-generated textual content as a practical search method. This method, for the first time, makes it practical to search for user-developed need-solution pairs at a very early stage in innovation development and diffusion, when innovation-related information is known only to very few.

* Corresponding author.

E-mail addresses: evhippel@mit.edu (E. von Hippel), sandro.kaulartz@ipsos.com (S. Kaulartz).<https://doi.org/10.1016/j.respol.2020.104056>

Received 3 July 2019; Received in revised form 17 June 2020; Accepted 21 June 2020

Available online 30 June 2020

0048-7333/© 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Of course, search methods of this general type can only be effective when users *do* openly post descriptions of their innovations on the web. Research has shown that innovating users have no incentive to post their designs only for the benefit of free-riders (de Jong et al. 2015; von Hippel et al. 2017). However, in areas of shared interest and collaboratively-developed innovations, sharing design information can also benefit innovating users themselves: they post to both inform peers and to learn from them (Baldwin and von Hippel 2011; von Hippel 2017). For example, innovating skateboarders have an incentive share their innovations openly on the web in specialized forums, in order to in return benefit from learning from others.

The remainder of this paper is organized as follows. In Section 2, the definition and nature of need-solution pairs and the valuable information, both concrete and tacit, that they can provide are made clear. In Section 3, evidence for the widespread nature and availability of openly revealed, functionally novel innovation by consumers in the household sector is summarized. In Section 4, we describe a Natural Language Processing (NLP)-based search approach that combines semantic word space models with semantic network analytic methods, and makes it practical to search for user-developed need-solution pairs at a very early stage in the innovation development and diffusion process, when innovation-related information is known only to very few.

The methods are, as we will show, capable of screening the entire web for open, user-generated textual content as a practical search method. In Section 5, we report upon a study we conducted using semantic word space models and semantic network analytics to identify novel, early-stage innovations in the field of kiteboarding. Both improvements to generally known needs, and also functionally novel need-solution pairs were identified in this demonstration study. In Section 6, we summarize and discuss some implications of our findings for both future research and improved practice.

2. Need-solution pairs and their role in the innovation process

A solved problem consists of a need paired with a responsive solution – a need-solution pair. Need-solution pair identification involves a fundamentally different division of innovative labor between users and producers than that implemented by both the standard “find a need and fill it” innovation method, and the lead user search method. In the case of find a need and fill it methods, both need identification and creation of a responsive solution are tasks undertaken by producers. In the case of lead user search, producers are responsible for identifying important need trends in the market, but development of prototype solutions at the leading edge of those trends is considered to be the domain of lead users. In the case of need-solution pair search – the type we will explore in this paper – identification of needs and development of responsive solutions to create desired need-solution pairs have *both* been carried out by innovating users. Resulting successful innovations – those that prove beneficial to their user-developers – are then identified and screened for individual value by peers, and/or for general market value by producers.

For the purpose of our work here, identifying commercially promising user-developed need-solution pairs, we are not concerned with how these pairs were originally developed by user innovators. We are instead searching for needs paired with responsive solutions that have already been developed – somehow. Thus, it may be that pairs identified were initially developed via a need-first problem-solving process, or by a need-solution pair recognition process. The search methodology we will describe later in this paper does not collect innovation histories – it focuses only on identifying the innovations themselves.

Because the need-solution pairs we will search for in this study are actual products actually put into use in real-world use environments by their user-innovators, both needs and solutions will be laden with valuable detail, both expressed and tacit. Thus, the need content of a completed pair is not something general that a marketing research trend study might identify such as “users need a more convenient product type X.” Instead, a completed pair in actual use contains very detailed,

although not necessarily textually encoded information that “the user values precisely this function and precisely that experience as delivered to him or her by precisely that prototyped innovation.” Similarly, the solution side of that need-solution pair is not just that “this product or service will be designed to deliver greater convenience to its user.” It is, instead that “this product type X weighs 26 gs, and is shaped exactly like that. Further, it is made of exactly those materials in conformance to exactly that design. To function, it requires exactly these inputs and delivers exactly those outputs.”

Of course, need-solution pairs do not necessarily start out encoding this level of precision – instead, both the need and responsive solution ultimately encoded into the actual artifact in actual use may have been interactively refined over multiple iterations of design and experimental use. However, and again for our present purposes, we are not concerned with how the completed and in-use need-solution pairs we will seek were developed. We only care that they contain very detailed and concrete information on both a need and a responsive solution developed by users and seen by at least some to be a valued need-solution pair. After such pairs have been found, and the subset with potential commercial or other interest to a searcher have been identified, it will be possible to gain further information about them if and as desired.

3. Availability of pioneering consumer innovations on the web

The search for novel, user-developed need-solution pairs will only be practical if and as there is an ample supply of functionally novel, user-developed innovations available in the field of interest to a given searcher. Further, information on at least some of those innovations must somehow be openly available and so discoverable by a practical search methodology. As was mentioned earlier in Section 1, in the case of shared areas of interest and collaboratively-developed innovations, self-rewarded user innovators will often have an incentive to post their novel designs and related learnings on the web to both inform peers and to learn from them. Of course, when innovators do this, they make their innovations visible to producers and free riders as well (Baldwin and von Hippel 2011; von Hippel 2017).

3.1. Why users pioneer novel products

Studies of important innovations support a pattern in which users enter new applications and markets ahead of producers in fields ranging from the development of the first aircraft (Meyer 2012), to the first personal computers (Levy 2010), and to the first personal 3D printers (de Bruijn 2010). In addition, two studies have explored the pattern of user vs. producer innovation over time in two fields, scientific instruments (Riggs and von Hippel 1994), and also the sport of whitewater kayaking (Hienert et al., 2014). In both fields, users were found to be the pioneers, creating a new activity and/or product type, and then developing numerous valuable follow-on innovations to enhance it for a period of years before the first entry of producers.

To understand why consumers – and, indeed, users very generally – are likely to pioneer with respect to development of functionally novel product and service development, consider that producers generally expect profitability to result from spreading their design costs over many purchasers. To meet that expectation, before investing in development, producers need to be confident that many customers will in fact be interested in the product they plan to develop. In contrast, information about potential market size is irrelevant to household sector innovators. It has been shown that over 90% of such innovators are self-rewarded (von Hippel 2017, ch. 2). That is, they are self-rewarded by the benefit they gain from solving their own problem; by the fun and learning they experience during the innovation process; and by altruism – the self-rewarding “warm glow” they experience from helping others. That is, 90% of consumer-innovators are not interested in commercializing their innovations, and so are not concerned with the potential extent of the market for their developments. (The remaining 10% of

household sector innovators have ambitions to be entrepreneurs, and so are, in common with existing producers, interested in the potential extent of the market.)

Reliable information on the likely extent of demand generally does not exist at the beginnings of new applications and new markets where users are trying to do novel things—like experimenting with the first skateboards or with the first heart-lung machines. At that stage, markets are small and customers' needs are still not clear. As a result, the information that a producer needs to determine whether acting on an innovation opportunity will be profitable is not available until well after the information that an individual innovator needs to determine the personal viability of that opportunity is available. This difference explains the observation that user innovators, including innovating consumers, will generally begin to innovate with respect to new product-specific needs and related responsive solutions before producers do so (Baldwin et al., 2006).

This difference in incentives and information needs between consumer and producer innovators suggests that a division of labor will exist between these two innovator types with respect to types of innovations developed. We would expect users to develop products or services of interest to them without concern for the fraction of the general market that might share their interest. In contrast, we would expect producers to develop innovations likely to be of interest to the whole market – improvements along “dimensions of merit” such as faster, cheaper, more reliable, etc.. Precisely this pattern has been documented in scientific instruments (Riggs and von Hippel 1994).

As can be seen in Table 1, users of scientific instruments – scientists – tend to develop innovations that enabled the instruments to do qualitatively new types of things for the first time. Such functions might have been of interest only to the innovators themselves, or they might also have been of interest to some additional fraction of the market. In contrast, manufacturers tended to develop innovations that made an instrument more convenient and more reliable in general—attributes of at least some interest to all potential customers. Of course, both types of innovation are valuable. So, the pattern we see is one of complementary sources of innovation – and an economically justifiable division of innovation labor between users and producers.

3.2. Users innovate in essentially all consumer good fields

User innovations can only be of value to producers in fields where they exist, and the potential value goes up as the amount of such innovations increase. Fortunately, evidence from nationally representative surveys of household sector innovation conducted in multiple nations shows consumer innovation is a major phenomenon. In aggregate across the nations surveyed, it has been found that tens of millions of consumers are spending tens of billions of dollars per year developing and improving products they use, both individually and collaboratively. These surveys also document that consumers develop and modify products in essentially all areas of consumer interest and activity. Categories specifically asked about were craft and shop tools, sports and hobby innovations, dwelling-related, gardening-related, child-related, vehicle-related, pet-related, medical, computer software, and food and clothing (Table 2 shows the findings on this for six national surveys). Categories showing relatively high levels of innovation map well upon

Table 1

Sources of scientific equipment innovations by nature of improvements effected.

Type of improvement provided by innovation	Innovation developed by		Total (n)
	User	Producer	
New functional capability	82%	18%	17
Sensitivity, resolution, or accuracy improvement	48%	52%	23
Convenience or reliability improvement	13%	87%	24

Source: Riggs and von Hippel 1994; table 3. Sample size 64.

Table 2

Scope of product development by household sector users in various innovation categories.

	UK ^a	Japan ^b	us ^b	Finland ^c	Canada ^d	S. Korea ^e
Craft and shop tools	23.0%	8.4%	12.3%	20%	22%	16.4%
Sports and hobby	20.0%	7.2%	14.9%	17%	18%	17.9%
Dwelling-related	16.0%	45.8%	25.4%	20%	19%	17.9%
Gardening-related	11.0%	6.0%	4.4%	na ^f	na	na
Child-related	10.0%	6.0%	6.1%	4%	10%	10.9%
Vehicle-related	8.0%	9.6%	7.0%	11%	10%	6.5%
Pet-related	3.0%	2.4%	7.0%	na	na	na
Medical	2.0%	2.4%	7.9%	7%	8%	5.5%
Computer software	na	na	na	6%	11%	na
Food and clothes	na	na	na	12%	na	na
Other	7.0%	12.0%	14.9%	3%	3%	23.9%

^aSource: von Hippel, de Jong, and Flowers 2012.

^bSource: von Hippel, Ogawa, and de Jong 2011.

^cSource: de Jong, von Hippel, Gault, Kuusisto, and Raasch 2015.

^dSource: de Jong 2013.

^eSource: Kim 2015.

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 all significant activities (Lader et al., 2006).

3.3. Prior and parallel research on internet searches for user-developed innovations

Studies seeking to identify user innovations on the internet have been done using a range of methods and for a range of reasons. Belz and Baumbach (2010) applied a big data 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. 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. Both research projects found this “self-signaling” approach to be promising.

The application of user generated data analysis techniques from social networks has been explored in manifold ways. Pajo et al. (2015) propose a 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 search algorithm, they used a sample of validated lead user questionnaires containing clusters of lead and non-lead Twitter users interested in that single product. They suggested that data mining techniques such as the one they explored 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 user identification along with information on latent product features that potential lead users share. Kratzer et al. (2016) come to the same conclusion after studying user innovation within social networks based on three empirical studies and found that they possess a distinctive position and function as bridges between social groups due to a higher “betweenness centrality”.

Nga N.Ho-Dac (2020) analyzed 1287 software projects on Sourceforge.net over a period of 16 months and found that user generated content is useful for product development and learning from online content facilitates product development activities. Ernst et al. (2017) studied the social media habits of 20 lead users and found that social media environments can contribute complementary to established lead user search methods, but the effectiveness strongly depends on the nature of the project. Timoshenko and Hauser (2019) used web-based searches of user-generated content to identify customer needs rather than solutions.

The study closest to our own in both method and intent was conducted in parallel to ours by C. von Hippel and A. Cann (2020). These authors focused on *behavioral innovations* developed by users – a very important type of innovation. They define behavioral innovations as consisting of one or a connected sequence of intangible problem-solving activities that provide a functionally novel benefit to its user developer relative to previous practice. For example, an innovation by a mother as to how to help her baby sleep that involved only novel behaviors and not novel products is a behavioral innovation in these authors' terminology. They demonstrate in a pilot study that behavioral innovation can be identified within user-generated content posted openly online in peer-to-peer discussion forums relating to household sector activities. The method they developed and apply in their research involves semantic analysis of publicly available records of online discussions of problem-solving activities within large online communities. This method differs from but also has aspects similar to the methods we used. They detail these differences in their paper. The success they demonstrate supports the robustness and promise of the new general approach to identification of lead user innovations within user-generated content on the web we are both developing and exploring.

4. Outline of natural language processing (NLP) methods used to identify consumer innovations

From the research evidence summarized in the previous section, we now see that consumer innovators are frequently innovation pioneers, developing and using functionally novel need-solution pairs earlier than producers. When they post information on their innovations on the web, furthermore, they very generally do so without intellectual property rights protections – recall that 90% of consumer-innovators expect only self-rewards for developing their innovations (Raasch and von Hippel 2013; von Hippel 2017). Therefore, if producers or peer household sector users can find the pioneering innovations, they will generally be free to make use of what they learn.

This leads us to the central question we address in this article: How in practice can one find pioneering user innovations that have been developed in the household sector? In early days of innovation development and testing, when producers and user peers ideally would like to learn of them, we think that a promising source for information on consumer innovations is user-generated content posted on the web. As we explained earlier, many consumer innovators post their innovative designs on the web during the development phase, and discuss the value they obtain from them as well. They do this in order to benefit from sharing and collaborating with others having similar interests. Consider that, by collaborating, each individual incurs the design cost of doing only a fraction of the project work but obtains the value of the entire design, including additions and improvements generated by others (von Hippel and von Krogh 2003; Baldwin and Clark 2006). Since designs are non-rival goods (both you and I can use a design at the same time—I am not competing with you for access), non-rival individuals considering creating an innovation should always prefer participating in a collaborative project to going it alone.

Given, then, that a useful fraction of household innovators do post descriptions of their innovations on the web, we face a next question: How can one find descriptions of promising consumer innovations within the immense quantity of user-generated content posted on the

web? Fortunately, today, broad searches of the entire web to identify consumer-developed innovations are newly practical due to recent developments in natural language processing-based search methods.

In our application of these methods, we begin by scraping open websites for user-generated textual content (UGC). We next subject that content to semantic word space model filtering to isolate just those posts most likely to refer to user-generated innovations. The semantic “word space model” filtering process we use is in principle quite simple, although it can be complex in practice. The model is based two-layer neural networks that are trained to reconstruct linguistic contexts of words by transforming word into vectors. These are positioned in a vector space such that words that share common contexts in the corpus are located close to one another in the space. For example, the algorithm might discover that the term “prorotype” is closely related to the term “concept” in the data corpus. Then, words or word combinations of interest, along with their relationship to each other, are encoded in dimensional vector space using similarity scores (Singh et al. 2015; Tsatsaronis and Panagiotopoulou 2009). The word space model proves to be an effective method to automatically discover the semantic relatedness and semantic neighbours of specific user expressions of interest to searchers (Sahlgren 2006). More recent advancements to represent distributed word vector representations by using skip-gram models to retrieve syntactic and semantic word relationships significantly improved the effectiveness and quality of word space model in practical use (Mikolov et al., 2013).

In principle, one might think this kind of a search and filtering process will be easy to do. If one is searching, say, for kiteboarding equipment innovations by users, why not just filter for the single word ‘kiteboarding’ or the entire three-word phrase, ‘kiteboarding equipment innovations’? This simple approach, alas, is likely to identify little relevant content. This is because, when individuals with a common expertise and interest post information for each other, they very generally do not include overall contextual information in their posts. They simply – and correctly – assume that fellow specialists who might be interested in reading their post will understand the context. Thus, in the case of kiteboarding innovation discussions, one valuable post from an innovator 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 – totally opaque to most but totally clear to enthusiast kite-surfers – refers to a specific equipment problem encountered by experts that can occur when controlling a kite while kiteboarding. (A chickenloop is a specific kiteboarding equipment component: a strong flexible loop used to attach the kitesurfer's kite “control bar” to his or her harness.)

Determining the most effective screening terms for any particular topic involves a process of trial-and-error development and refinement of a project-specific semantic algorithm development. The process begins with a subject matter expert familiar with the topic being searched. That individual uses his or her knowledge to create a list of words and phrases commonly used by participants in the subject matter being explored. A filter including these will next be run against the often massive number of potentially relevant user-generated posts scraped from the web – flagging all posts that include one or more of the chosen words and phrases.

Next, the subject matter expert will inspect – actually read – one or two hundred of the posts or other content flagged by the filter. In the case of posts that contain desired information, the expert will note the words that produced this useful result and keep them in the filter for the next run. In the case of posts flagged that do not contain desired information, the expert will inspect to see which filter components were involved in creating the false positive. These will be removed or refined prior to the next filter run. This trial-and-error test cycle continues until the material that passes through the filter contains few enough posts (maybe a few hundred), and a high enough percentage of positive identifications (maybe 10%) to make a final, manual inspection and coding of each remaining post a practical exercise.

Table 3
Consumer innovation identification process steps.

Process Steps	Process Activity Description	Kite Surfing Study Data Example
Step 1	User Generated Content Scraping Subject matter expert defines the initial domain taxonomy that best represents the UGC search field. User generated data content scraping then is conducted related to the domain of interest to identify websites and relevant UGC.	234,017 English UGC posts scraped from 9617 websites across the globe
Step 2a	Innovation Concept Identification Semantic filtering using semantic network analytics and word space models to identify the relevant data corpus for the innovation concept (e.g., DIY, Innovation, Inventions, Problem Solving & Developing Solutions).	6065 posts remain after step 2a filtering
Step 2b	User Innovation Filtering Refined syntax based semantic pattern algorithm with metacharacters to isolate first person speech (e.g., <i>I, myself, me, we, us</i>) in combination with the semantic “innovation concept” (e.g., <i>invented, developed, designed</i>)	453 posts remain after step 2b filtering
Step 3	Subject Matter Expert Validation 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. Material eliminated includes producer innovations, duplicates, and very minor improvements and alterations to existing kiteboarding hardware.	26 LU innovations remain after manual filtering based upon two criteria of <i>novelty</i> and true user <i>innovation</i>
Step 4	User Innovation Trend Analyses Test of popularity / commercial value of identified consumer user innovations via trend analyses of frequency of relevant UGC contributions related to each identified user innovation, and frequency of searches for that innovation by Google search engine users.	5 fundamentally radical LU innovations ranked by trend analyses to assess commercial promise

Note that, in the filtering process just described, one cannot ever be sure that one has identified all the user-developed innovations of interest that exist in the user generated content being filtered. One can only recognize when one has “enough” to satisfy the project purpose. In other words, the search process requires one to satisfice, stopping the trial-and-error cycle when one has found enough user-generated innovations to satisfy the practical goals of a particular searcher.

For the final step of determining likely commercial potential, one cannot rely on standard marketing research methods that assess the potential value of an innovation by its responsiveness to needs widely held among potential customers. In the case of early stage need-solution pairs addressing novel needs that approach is likely to come up empty – too few individuals in the potential market will be in a position to recognize or assess the needs a need-solution pair addresses. Therefore, under these conditions, a better approach, we think, is to ask subject matter experts for their qualitative assessments, and/or use web data to assess temporal trends in the number of user-generated posts and web queries mentioning each identified innovation. If the number is going up over time, this could signal emerging general user interest and thus eventual commercial potential.

5. Method application case study: kiteboarding

Before applying the methods to be described in this section, we insured that these were in full compliance with EU data privacy regulations in force at the time of our research. However, we would like to point out to fellow experimenters that the regulatory environment is rapidly evolving in many nations and political entities with respect to protecting personal data privacy, and also with respect to whether and how those data may be used by companies for commercial purposes (e.g. GDPR, CCPA and PDPA). For this reason, researchers interested in performing a study similar to the type we describe in this section should check that the methods they plan to use are in compliance with the latest governmental regulations in force at the time of their study. And, of course, they should also check to insure that they are in compliance with the terms of service of websites searched with respect to content crawling (e.g. /robots.txt - the Robots Exclusion Protocol.)

As our proof of concept test, we applied the semantic search method described in outline above to an actual case study. 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., 2006; Franke et al., 2006). Therefore, if application of our semantic search and analysis methods did *not* yield evidence of user innovation, this would suggest that our methods are failing to capture user-developed innovations that in fact do exist for reasons requiring further investigation.

Kiteboarding is a fast-growing water sport that combines aspects of surfing, wakeboarding, windsurfing, and paragliding. As of 2008, 1.2 million people were engaged in kiteboarding around the world, accounting for about 15 percent of all paddling activities (Outdoor Foundation 2009, 44). Expenditures by participants for gear and travel and other services reached hundreds of millions of dollars annually by 2009 (Outdoor Industry Foundation 2006; Outdoor Foundation 2009). As a very general description of what is involved in the sport, kiteboarders stand on a special board, somewhat like a surfboard, and are pulled along by holding onto a large, windborne, steerable kite. Since the sport’s (user-developed) inception, 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.

5.1. Process step overview

In overview, the major process steps we followed in our proof of concept study are listed in Table 3. Complete process details, including filtering terms used for each step, are provided in a methodological appendix.

5.2. Consumer innovations identified

Our scan of publicly available user-generated content across the web spanned the years 1999 to 2018. Via applying the method steps outlined above we identified 26 consumer-developed innovations. These final 26 were selected after a final, manual curation process that examined a few hundred remaining posts and weeded out-of-field innovations, duplicates, producer-developed innovations, and innovations that did not offer 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.)

In a separate internet search, we determined whether or not each kiteboarder-developed innovation identified had already been adopted

as a commercial product by a producer firm. We found that 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 many user-generated innovations are of commercial value. The newest innovations were the least likely to have been commercialized at the time of data collection, suggesting that application of this semantic search method can indeed identify very recent user innovations that may represent as-yet unexploited commercial opportunities for producers.

5.3. Novel solutions to known needs

If a need is already known to a searcher, identification of need-solution pairs in which the solution addresses that known need “only” gains the searcher information on a novel responsive solution. However, if *both* the need addressed and the responsive solution are novel to the searcher, the information gained is potentially much more valuable, presenting the possibility of new product and application categories not previously known to the searcher.

Of course, we do not know the level of knowledge of producer-based experts regarding the needs in the solution pairs our web search identified. However, of the 26 need-solution pairs identified, 21 were user-developed improvements to kiteboarding equipment that are clearly within the confines of the kiteboarding sport 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 Fig. 1.

The first image in Fig. 1 is of a “hardshell” harness design that better distributes the pulling force of the kite across the kiteboarder’s body. The second image shown is an alteration to the control lines connecting the surfer to the kite to improve his or her control of the kite’s direction of motion and power under varying weather conditions. The third image is of a special seat attached to a standard kiteboarding board that for the first time enables people with certain disabilities to participate in the sport. In this third case, especially, it is likely that changes to multiple components in the kiteboarding use system will be required to accommodate the varying needs and capabilities of handicapped individuals. For example, ways that the individual controls the board and kite must change. Very likely, therefore, that this third example would provide producers with need-solution pairs that have significant novelty in both the need and solution side of the pair.

5.4. Functionally novel need-solution pairs

Three of the innovations we uncovered in our search seem to us likely to be novel with respect to *both* need and responsive solution. Each preserved one of the two major kiteboarding equipment elements – a kite and a board – but not both. Each also significantly altered the nature of the sport – offering significant novelties to adopters. As a result, for those electing to use or produce these innovations, the level of use novelty and market novelty would be high. Images of these three

innovations are shown in Fig. 2. The electric hydrofoil innovation eliminates the kite as the source of motive power, substituting an electric motor (mounted on a hydrofoil under the board). The second image shows a user who has replaced the conventional kiteboarding kite as a source of motive power with a powerful drone flying overhead. In both of these cases, 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.

Finally, the third example depicted in Fig. 2 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 user innovation searches can of course include or exclude such “off-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 need-solution pairs and possible future market directions.)

5.5. Identifying likely general interest in innovations via web data

Analyses carried out in step 4 of our search method (cf. Table 2) are intended to give some information on the potential commercial value of each need-solution pair identified. Metrics used are frequency of mentions of an innovation identified on websites, and index data on the 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. In Google Trends search volumes over defined periods are indexed for each individual search terms (e.g. “kitesurfing”) or search word comparisons (e.g. “eFoil” vs. “Electric hydrofoil”) with the index value 100 for the highest search for a specific moment in time. Calculating search index for multiple search terms therefore requires the use of a benchmark search that is higher than the search of interest. To compute the total volume of all relevant searches for the user innovations (e.g. Electric Hydrofoil) we applied individual searches for the relevant search terms (e.g. “eFoil”, “Electric hydrofoil”, “eFoiling”) and the a benchmark search “kitesurfing” before cumulating the Google Trend indexes for each individual search. As can be seen in Fig. 3, the electric hydrofoil innovation was attracting more user interest than the drone-surfing innovation on both of these metrics at the time of our search.

The trend analysis allowed us to assess how the diffusion of social mention and searches interrelated to each other in the case of user innovations likely to be novel to searchers with respect to both needs and responsive solutions. As it is visible in the case of electric hydrofoiling, we see the first major increase in social mentions in Quarter 4 of 2015 among experts exchanging ideas on innovation in kiteboarding web communities. The social media conversation climax was reached in

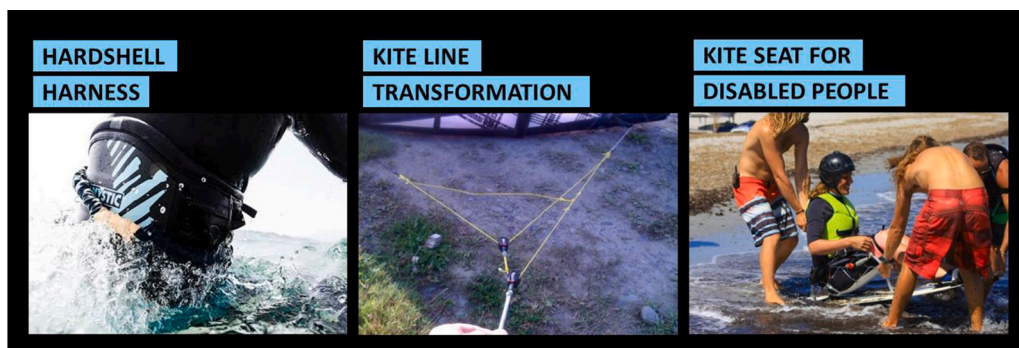


Fig. 1. Examples of novel solutions to known needs.

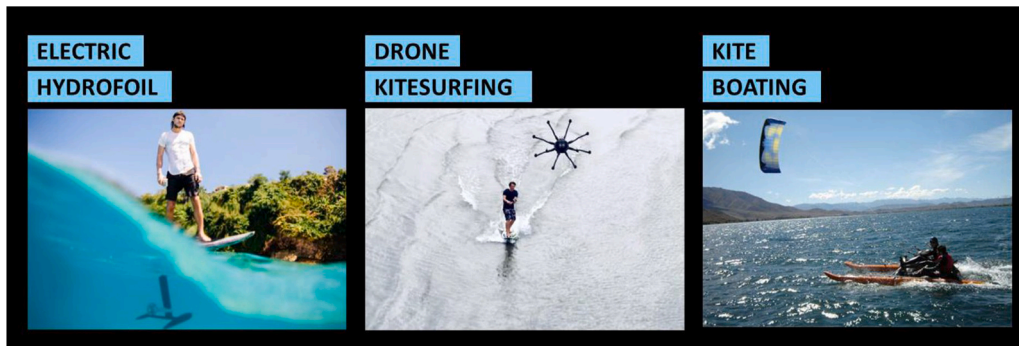


Fig. 2. Examples of innovations fundamentally shifting the nature of the sport.

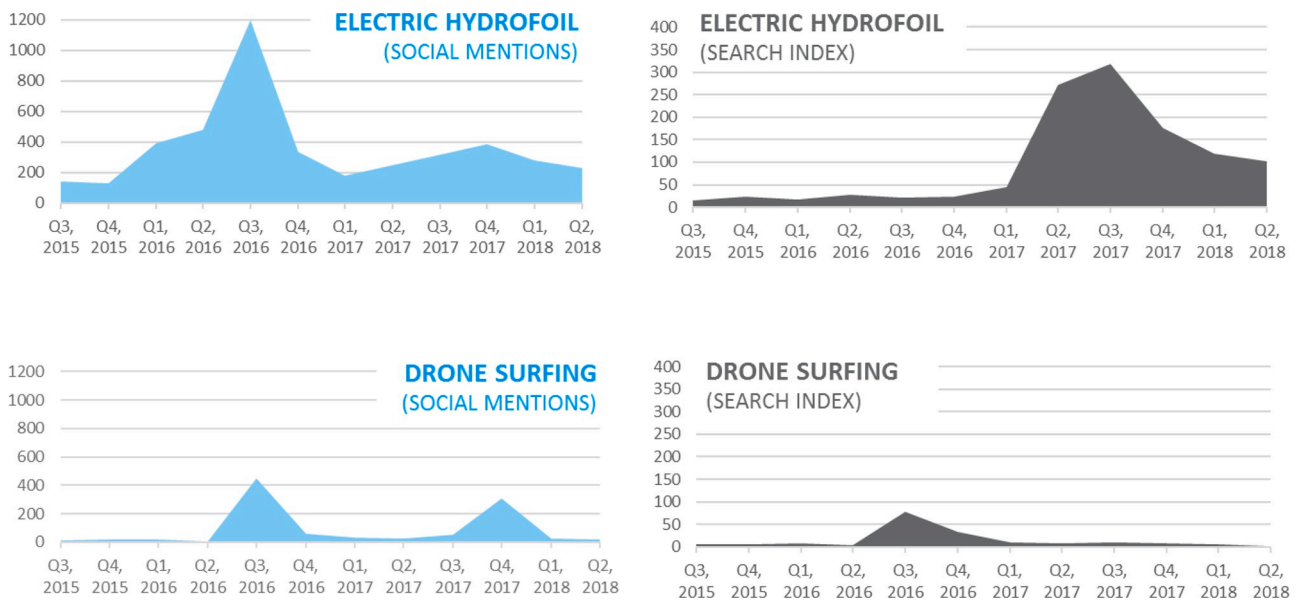


Fig. 3. Trend data for two recent sport-extending innovations.

Quarter 3 of 2016. The first visible increase in searches related to electric hydrofoiling started later in Quarter 1 of 2017. Even though the delay effect between social mentions and searches was shorter for some of the radical user innovations like drone surfing, the trend data analysis from the kiteboarding study suggests that innovation trends start off with expert discussions in domain specific forums among those who are ahead of the curve. Later, the subject matter interest spreads to a broader audience.

5.6. Additional sources of information on innovations of special interest

The methods described in the previous section are useful to provide a preliminary indication of likely general interest in novel need-solution pairs identified. Innovations that look promising to a searcher based upon that data can then be followed up with much richer qualitative information regarding both the innovation and its commercial potential obtained directly from user-innovators and peer user adopters who are willing to discuss these matters. Each post 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. In addition, searches can be conducted to identify whether the innovation has been already commercialized. Often, in this field, the first to commercialize an innovation are innovators themselves or early peer adopters, generally founding a small company to accomplish this (Baldwin et al., 2006). Larger incumbent firms will of course 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 follow.

6. Discussion

In this paper we have explained the likelihood that consumers, rather than producers, will pioneer in the development novel needs coupled to novel responsive solutions in the case of consumer products they use. We have also proposed and tested a novel method of finding these pioneering need-solution pairs prior to general awareness. Searching for early-stage, functionally novel need-solution pairs can have great value to producers as well as to innovation research and practice. In the internet age, information spreads quickly, and it can be important for producers to gain early awareness in order to establish an early entrant position with respect to commercially supplying user-pioneered functional novelties.

The method we have proposed and demonstrated in this paper for finding consumer innovations developed in the household sector, for which both the detailed need and detailed responsive solutions are novel to searchers offers, we think, a very valuable complement to both traditional market research techniques and practice, and to lead user innovation search practices as well. As was mentioned at the start of this paper, in the case of traditional “find a need and fill it” producer innovation practices, both identification of needs and development of responsive solutions are seen as tasks for producers. In the case of lead user studies – “find users who are innovating at the leading edge of a

(producer-defined) market trend – producers maintain their responsibility for need identification, but seek prototype solutions among lead users (Lilien et al., 2002). In the case of identification of functionally novel need-solution pairs we have described here, *both* need formulation and responsive solution formulation are outsourced to innovating household sector users.

6.1. Suggestions for further research

In the research reported upon here, we have described and initially tested the application of word space and semantic network analytic techniques to screening of user-generated content for functionally novel, consumer-generated need solution pairs. An immediate follow-on research question, of course, is whether our success in a single case is generalizable. Arguing from general principles, we do not see why it would not be, but clearly, research is needed to carefully assess this important matter.

When the question of general applicability of the basic method has been demonstrated by further research, we think it would be very useful if researchers with Natural Language Processing expertise could consider creating semantic search toolkits that less-expert practitioners would find it feasible to apply. Future research could investigate the practicability and efficiency gains of applying the latest advancements in the unsupervised machine learning domain for NLP such as Bidirectional Encoder Representations from Transformers (BERT) published by Google AI Language that enables a deeper sense of language context understanding compared to the word space models by using a bidirectional language representation approach (Devlin et al., 2018).

It appears to us from our initial experimentation that only some of the search terms required in any particular search area will be unique to that specific area (e.g., “chickenloop” in the case of kiteboarding). Other important search terms are likely to be general across cases, and these could be supplied to searchers as part of a semantic word search toolkit. Thus, 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*) would seem to fall into this category. Subject-matter experts on the practitioner team need only then add subject-specific words and phrases to complete an effective filter set. This approach was successfully pioneered by C. von Hippel and A. Cann (2020) to identify behavioral innovations, and has the potential to significantly ease the task of designing and executing specific studies.

As another research direction, it may be worth experimenting with the semantic word space methodology we have described here, or with improved methods to follow, in lead user studies. In that application, the method would serve as a replacement for the pyramiding search methods involving successive interviews traditionally used in lead user studies. Accomplishing this would involve tightening up the constraints in the screening process so that all need-solution pairs selected address a specific need, rather than any need. It may be that lead user studies can be made significantly faster and less labor intensive in this way (von Hippel, Franke, and Prügl 2009).

In general, the additional research we propose here will build upon more basic developments in semantic analysis techniques. However, it is not impossible that learnings from need-solution pair search methods and studies can also inform more basic semantic analysis research.

6.2. Suggestions for practice

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 early-stage identification of functionally novel need-solution pairs possible. This favorable turn of events should in turn make the method we have proposed and presented here a useful complement to existing market research and lead user search methods. Standard market research methods can identify unfilled needs or “pain points” which many potential customers are experiencing. Lead user

search can identify novel approaches to solving those generally understood needs. Need-solution pair search can then focus on identification of early-stage needs coupled with responsive solutions, developed and refined by users via actual trial-and-error experience.

In this regard, however, it is important to state (again) that the semantic word space filtering process we describe and demonstrate in this paper will not identify *all* user-developed need-solution pairs posted openly on the web. Changes to the semantic filters applied in a specific project and/or changes in the interests and knowledge base of expert evaluators working on that specific project will likely identify more, or fewer, and possibly different subsets of all such innovations extant. For this reason, the method is useful for the purpose stated in this paper – a means for producers or other searchers to discover *some* interesting, user-generated need-solution pairs at an early stage in their evolution. It will *not* be useful for generating complete inventories of such innovations.

To make the novel method we have described as valuable as possible, it will be important to learn how to incorporate need-solution pair identification into corporate product development practices. To do this, companies as a whole, and market researchers specifically, need to learn a new division of innovation labor. Market researchers should learn that, in most instances, they should no longer assume that it is their task to develop functionally novel product concepts *for* consumers. Instead, they should consider reallocating some resources formerly devoted to that task to the identification and evaluation of novel need-solution pairs developed and prototyped *by* users. As the phrase “division of innovation labor” implies – the new method we propose is a complement to present producer innovation development methods, not a replacement.

CRedit authorship contribution statement

Eric von Hippel: Conceptualization, Writing - review & editing.
Sandro Kaulartz: Methodology, Data curation, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Methodological appendix

The promise of semantic network analytic and semantic word space 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.

At the outset, it is important to note that, at the time data collection and analysis for this study was carried out, the methods we used and describe here were in complete conformance with EU regulations. (At the time of writing, the EU has the strictest data privacy regulations in the world.) However, it is also important to note that the regulatory environment is rapidly evolving in many nations and political entities with respect to protecting personal data privacy, and also with respect to whether and how those data may be used by companies for commercial purposes (e.g. GDPR, CCPA and PDPA). For this reason, researchers should check that the methods they plan to use are in compliance with the latest governmental regulations in force at the time of their study, and also check that they are in compliance with the terms of service of websites searched with respect to content crawling (e.g. /robots.txt - the Robots Exclusion Protocol).

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

The first step in our 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 need solution pairs within the relevant data corpus. The data collection included publicly available user generated content from blog posts, forum posts, consumer review, social media and microblogging platform posts (e.g. Facebook, Instagram & Twitter). The data crawling approach used was, at the time of the study, fully compliant with the data protection and personal data management regulations such as the General Data Protection Regulation (GDPR) in the European Union, the California Consumer Privacy Act (CCPA) and the Singapore Personal Data Protection Act 2012 (PDPA). We collected publicly available information only, private status or any data that has been restricted by the user was not accessed and included in the analysis. The analysis was conducted on voluntarily posted content on the Internet with consent regarding the use by third parties.

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.

In fact, as can be seen in Fig. 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 functionally novel user innovations. The comparison between the two different approaches requires an expert validation of a random sample of $n \sim 100$ 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 Fig. 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 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.

The alternative "broad search" method is based upon a taxonomy of terms and expressions (shown in Fig. 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 confluences 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 wind-surfing as well as kiteboarding (Fig. A-4).

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

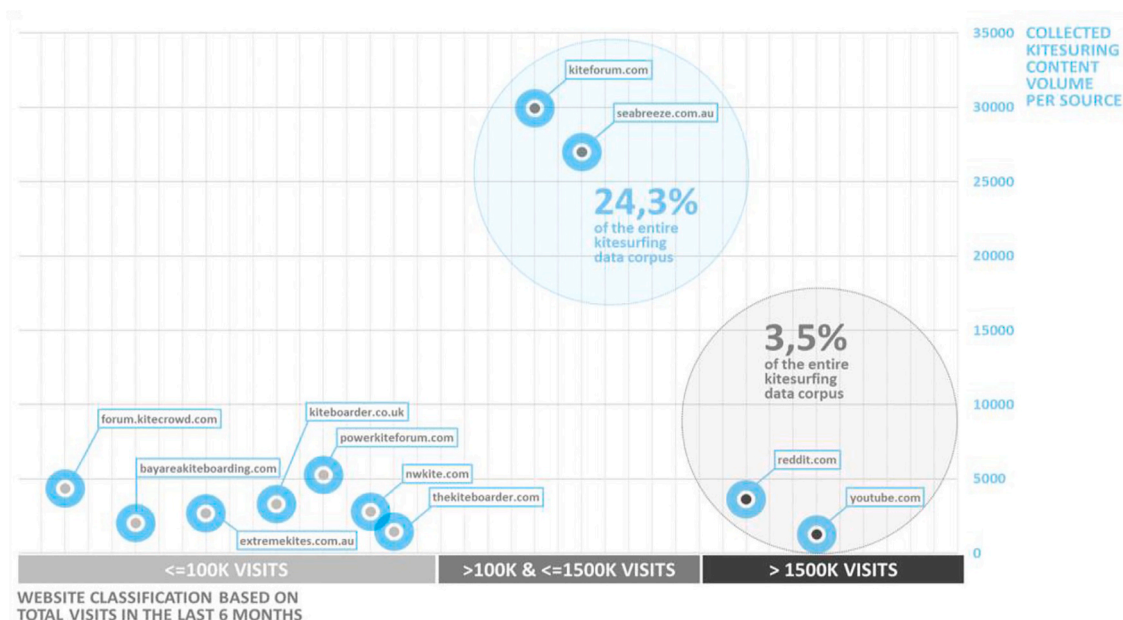


Fig. A-1. UGC volume per source with monthly traffic classification.

Step 1: Search query method validation			
<i>Search Method Tested</i>	<i>Search Method Description</i>	<i>Search Term Composition</i>	<i>Search approach effectiveness after expert validation</i>
‘Soft’ Search Approach	Search universe aiming to collect user generated content in the <u>broader context environment</u> of kiteboarding.	<i>First Person Speech</i> (“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 <u>direct topical context</u> of kiteboarding.	<i>First Person Speech</i> (“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 <u>broader context environment of kiteboarding</u> 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 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.

Fig. A-2. Search query method validation steps used for the kiteboarding study.

toeside) or tricks (jibing, back mobe or s-bend) to capture true expert content well.

Steps 2a and 2b: Semantic analysis for 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 use NLP techniques – word space model 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”

In this step we began with a semantic word space model (also referred to as vector space model) 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

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
Twin-tip	Boosting	Toe turn jibe	Relaunch	Bar throw	Legaigoux
Wake Style	Gliding	Jump jibe	Bar pressure	Trim	Mike Waltze
Race	Quiver	Heel turn jibe	Wind range	Leading Edge	Gerry Lopez
Wave	Airstyle	Toe down riding	Overpowered	Trailing Edge	Bill Roeseler
Harness	Old School	Floaty Jump	Powered-Up	Struts	Cory Roeseler
Chicken Loop	Airtime	Grab	Power Zone	Windward	Legaigoux
Safety Floats	Airborne	Spin	Sheeting	Leeward	brothers
Bar	Air pass	Board-off	Underpowered	Upwind	Manu Bertin
Lines	Surface pass	Invert	Upwind	Downwind	Ian Day
Harness	Freestyle	Kite Loop	Wind window	Wind range	Don Montague
Pump	Toeside	Back Loop		Hybrid	Robby Naish
Flying Lines	Water launch	Back Mobe		Hindenburg	Joe Keuhl
Foil / Foil kite	Water starting	Back to Wrapped		Dawn Patrol	Flash Austin
Kite	Beach starting	Blind Judge		Downwinder	Robby Naish
SLE Kite	Body dragging	Back Loop		Guinea pig	Flash Austin
Foil Kite	Looping	Front Loop		Kitemare	Lou Wainman
C Kite	Spinning	S-Bend		Luff	Elliot Leboe
Hybrid Kite	Speeding	F16		Lofted	Laird
Bow Kite	Tow In	Front to Blind		Overhead	Hamilton
Wetsuit	Backside	Front to Wrapped		waves	Mango
Wing	Top Turn	Handle Pass		Schlogging	Carafino
Strut	Carving Turn	Hooked Back		Sideshore	Gijsbertus
Bridles	Top to bottom	Blind Ole		Side onshore	Panhuise
Control Bars	Foiling	KGB		Stomp	Don Montague
Stopper Bar	Hydrofoil	Indy Front Loop		Tack	Neil Pryde
Tuflites	Kite foiling	Krypt		Tea bagging	Laurent Ness
.....	Sliding		Walk of	Franz Olry
			Shame	Rob Douglas
				Nicolas Parlier
					Christophe
					Martin
					Carlos Mario
				

Fig. A-3. Search taxonomies for kiteboarding topics.

layers and expressions that are sometimes out of context. The corpus texts are encapsulated in a space of vectors. Within this dimensional space all vectors have certain orientations and relationship with each other. The closeness and distance of vectors within the dimensional space was calculated using cosine similarities. The cosine similarity (also referred to as cosine distance) is the angle between vectors. Two vectors with the same orientation have a cosine similarity of 1. Within GenSim or spaCy, analysts today have free access to convenient Python libraries with intuitive interfaces that are ideal for the described approach, and are widely adopted. With spaCy users can access pretrained models in various languages for a broad variety specific word embedding needs (e. g. “en_core_web_lg” released in November 2019). For our research application we used a pre-trained word space model with embeddings to assign word vectors, token vectors, part of speech tags, dependency parse and named entities. The word space model is continuously fed and updated with new online data from blog posts, news and consumer generated comments. Compared to static pretrained models we are able to keep abreast of the language variations over time and changing nuance in online language. The model used in the application study

contains 300-dimensional vectors and a context window of 2 which means words that are 2 to the left and right of the target words are considered context words.

For example, this method helped us to determine that the term “wheel” has an indirect semantic relationship within the 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

Step 2: User Innovation Filtering Method			
<i>Innovation Filtering Method Test</i>	<i>Semantic Analysis Method</i>	<i>Analysis Approach Description</i>	<i>Filter Result</i>
1)	“Innovation Concept” Word Space Model	Using word embedding and artificial neural network techniques to model the “Innovation Concept” (“ <i>DIY solution...</i> ” OR “ <i>solving...problem</i> ” OR “ <i>invented...</i> ”) 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” Word Space Model combined with “First Person Speech” Semantic Network Analysis	Using word embedding and artificial neural network techniques to model the “Innovation Concept” (“ <i>DIY solution...</i> ” OR “ <i>solving...problem</i> ” OR “ <i>invented...</i> ”) 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

Fig. A-4. User Innovation Filtering Method.

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 word space model, we condensed the overall volume of 234,017 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 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 meta-characters 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 user content universe using tokenization techniques to identify the posts that clearly indicated the 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 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.)

Step 3: Expert review of the identified 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 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 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 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 need solution pairs that solved known unfilled needs such as the invention of the hard-shell harness or the improved kite control method called the “kite power optimizer.” The list of innovations also included those that to some extent radically changed both needs and related solutions 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 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 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.)

References

- Baldwin, Carliss Y., von Hippel, Eric, 2011. Modeling a Paradigm Shift: from Producer Innovation to User and Open Collaborative Innovation. *Organization Science* 22 (6), 1399–1417.
- Baldwin, Carliss Y., Christoph, Hienerth, von Hippel, Eric, 2006. How user innovations become commercial products: a theoretical investigation and case study. *Res Policy* 35 (9), 1291–1313.
- Baldwin, C.Y., Clark, K.B., 2006. The architecture of participation: does code architecture mitigate free riding in the open source development model? *Manage Sci* 52 (7), 1116–1127.
- Belz, F.M., Baumbach, W., 2010. “Netnography as a method of lead user identification. *Creativity and Innovation Management* 19 (3), 304–313.
- Brem, A., 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.
- de Bruijn, E., 2010. On the viability of the Open Source Development model for the design of physical objects: Lessons learned from the RepRap project. Master of Science thesis. Tilburg University, Netherlands.
- 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., von Hippel, Eric, Gault, Fred, Kuusisto, Jari, Christina, Raasch, 2015. Market failure in the diffusion of consumer-developed innovations: patterns in Finland. *Res Policy* 44 (10), 1856–1865.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Ernst, Markus, Alexander, Brem, 2017. Social media for identifying lead users? Insights into Lead Users' Social Media Habits. *International Journal of Innovation and Technology Management* 14 (04), 1750022 (2017).
- Franke, Nikolaus, von Hippel, Eric, Martin, Schreier, 2006. Finding Commercially Attractive User Innovations: a Test of Lead-User Theory. *Journal of Product Innovation Management* 23, 301–315.
- Hienerth, Christoph, von Hippel, Eric, Morten, Berg Jensen, 2014. User community vs. producer innovation development efficiency: a first empirical study. *Res Policy* 43, 190–201.
- Ho-Dac, N.G.A., 2020. The value of online user generated content in product development. *J Bus Res* 112, 136–146 (2020).
- 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, Lettl, Christopher, Franke, Nikolaus, Peter, A.Gloor, 2016. The Social Network Position of Lead Users. *Journal of Product Innovation Management* 33 (2), 201–216. <https://doi.org/10.1111/jpim.12291>.
- Lader, D., Short, S., Gershuny, J., 2006. The Time Use Survey, 2005: How We Spend Our Time. Office for National Statistics, London.
- Levy, S., 2010. Missing information in the reference is O'Reilly Media, Sebastopol, CA. Hackers: Heroes of the Computer Revolution. O'Reilly.
- Lilien, Gary L., Pamela, D.Morrison, Kathleen Searls, Mary, Sonnack, von Hippel, Eric, 2002. Performance Assessment of the Lead User Idea Generation Process. *Manage Sci* 48 (8), 1042–1059.
- Meyer, P.B., 2012. Open technology and the early airplane industry. In: Paper presented at annual meeting of Economic History Association. Vancouver, BC. Accessed January 30, 2016. http://www.law.nyu.edu/sites/default/files/ECM_PRO_069779.pdf.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. <https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.
- Outdoor Foundation. 2009. A Special Report on Paddlesports 2009: kayaking, Canoeing, Rafting. Accessed January 20, 2016. <http://www.outdoorfoundation.org/research.paddlesports.html>.
- Outdoor Industry Foundation. 2006. The Active Outdoor Recreation Economy: a \$730 Billion Annual Contribution to the U.S. Economy. Accessed January 20, 2015. <http://www.outdoorindustry.org/images/researchfiles/RecEconomypublic.pdf?26>.
- Pajo, Sanjin, Verhaegen, Paul-Armand, Vandevanne, Dennis, Joost, R.Duflou, 2015. Fast Lead User Identification Framework. *Procedia Eng* 131 (2015), 1140–1145.
- Raasch, C., von Hippel, E., 2013. Innovation process benefits: the journey as reward. *Sloan Manage Rev* 55 (1), 33–39.
- Riggs, William, von Hippel, Eric, 1994. The Impact of Scientific and Commercial Values on the Sources of Scientific Instrument Innovation. *Res Policy* 23 (4), 459–469.
- Sahlgren, Magnus, 2006. The Word-Space Model, Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Stockholm University, Department of Computational Linguistics, Stockholm.
- Singh, Vaibhav Kant, Singh, Vinay Kumar, 2015. Vector space model: an information retrieval system. *International Journal of Advanced Engineering Research and Studies*. <https://www.technicaljournalonline.com/ijaers/VOL%20IV/IJAERS%20VOL%20IV%20ISSUE%20II%20JANUARY%20MARCH%202015/572.pdf>.

- Tietz, R., Fueller, J., Herstatt, C., 2006. Signaling: an innovative approach to identify lead users in online communities, 2006. International Mass Customization Meeting, Hamburg.
- Timoshenko, Artem, Hauser, John, 2019. Identifying Customer Needs from User-Generated Content. *Marketing Science* 38 (1), 1–20.
- Tsatsaronis, George, Panagiotopoulou, Vicky, 2009. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. *Association for Computational Linguistics*. <https://www.aclweb.org/anthology/E09-3009>.
- Tuarob, S., Tucker, C., 2015. “Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks. *Journal of Mechanical Design* 137 (7), 2015.
- Urban, Glen L., Carter, Theresa, Gaskin, Steven, Mucha, Zofia, 1986. Market Share Rewards to Pioneering Brands: an Empirical Analysis and Strategic Implications. *Manage Sci* 32 (6), 645–659.
- von Hippel, Christiana, Cann, Andrew, 2020. Behavioral Innovation: pilot Study and New Big Data Analysis Approach in Household Sector User Innovation. *Research Policy*, forthcoming.
- von Hippel, Eric, von Krogh, Georg, 2003. Open source software and the “private-collective” innovation model: issues for organization science. *Organization Science* 14 (2), 209–223.
- von Hippel, Eric, Harold, Demonaco, de Jong, Jeroen, 2017. Market failure in the diffusion of clinician-developed innovations: the case of off-label drug discoveries. *Science and Public Policy* 44 (1), 121–131. <https://doi.org/10.1093/scipol/scw042>.
- von Hippel, Eric, Franke, Nikolaus, Prügl, Reinhard, 2009. “Pyramiding”: efficient Identification of Rare Subjects. *Res Policy* 38 (9), 1397–1406.
- von Hippel, Eric, de Jong, Jeroen P.J., Flowers, Stephen, 2012. Comparing business and household sector innovation in consumer products: findings from a representative survey in the UK. *Manage Sci* 58 (9), 1669–1681.
- von Hippel, Eric, Ogawa, Susumu, de Jong, Jeroen P.J., 2011. The Age of the Consumer-Innovator. *Sloan Manage Rev* 53 (1), 27–35.
- von Hippel, Eric, von Krogh, Georg, 2016. Identifying Viable “Need-Solution Pairs”: problem Solving Without Problem Formulation. *Organization Science* 27 (1), 207–221. <https://pubsonline.informs.org/doi/10.1287/orsc.2015.1023>.
- von Hippel, Eric, 2017. *Free Innovation*. MIT Press, Cambridge MA.