

# Everybody wins: challenges and promises of knowledge discovery through volunteer computing

*L. Tuovinen and J. Rönning*

*Intelligent Systems Group, Department of Electrical and Information Engineering*

*P.O. Box 4500, FIN-90014 University of Oulu, Finland*

*+358 8 553 2806*

*lauri.tuovinen@ee.oulu.fi*

## Abstract

Knowledge discovery technology provides methods for extracting useful knowledge from large quantities of data. The knowledge, hidden in the distribution and internal structure of the data, is beyond the reach of traditional analysis methods, but with data mining algorithms it can be found. Effective knowledge discovery often requires access to substantial computing resources, and one way to acquire this access is to employ volunteer computing. In volunteer computing the data is processed by computers administrated by individuals or groups who are willing to take part chiefly for the pleasure of contributing to interesting research. The computers are mainly used for other things, but a share of their processor time is devoted to solving assignments downloaded over the Internet from a project server. The challenges involved are not all technical but social as well; for example, establishing trust between researchers and volunteers is of major importance. It is therefore evident that issues with significant ethical connotations may need to be solved in knowledge discovery, yet such issues are completely disregarded by currently accepted, technically oriented process models. We examine volunteer computing, and knowledge discovery in general, to identify the social factors that should be addressed in order to ensure that a knowledge discovery effort satisfies the legitimate expectations of all stakeholders.

**Keywords:** Knowledge discovery, data mining, volunteer computing, stakeholder analysis, trust, fairness, persuasion, privacy.

## INTRODUCTION

Knowledge discovery through data mining is a branch of computer science that studies methods and technologies for extracting knowledge from quantities of data so vast that special algorithms are required to reduce the data into a compact representation that can be understood and used. The knowledge comes in the form of interesting patterns that are hidden to a superficial inspection but can be made visible by fitting a computational model to the data.

Research on knowledge discovery largely focuses on technology: methods for preparing the data, mining it for knowledge and visualising the results. This work has been fruitful and resulted in many successful applications in all areas of human activity where data is generated and gathered. Just to name a few prominent application domains, knowledge discovery has proved useful in industrial quality control (Chen et al., 2004; Zhang et al., 2003), medical diagnostics (Li et al., 2004; Zaffalon et al., 2003) and marketing (Chou et al., 2000; Gersten et al., 2000).

More recently there has been increasing interest in the process through which knowledge is extracted from data (Laurinen, 2006). Improved understanding of the process, together with the accumulation of effective algorithms for data mining tasks, has in turn made it possible to create new technology in the form of data mining frameworks that combine a library of algorithms with a graphical application builder (Berthold et al., 2006; Mierswa et al., 2006). Several of today's most widely used database management systems are also now available bundled with a data mining toolkit, illustrating the acceptance of knowledge discovery technology as a powerful business instrument.

A notable deficiency in the way the knowledge discovery process is currently viewed is that while it covers the transformation steps required to get from data to knowledge and the interactions between the steps, it largely neglects the people who participate in the process and the interactions between them. This is a significant oversight because people and the relationships they share play a central part in any knowledge discovery effort. These relationships can sometimes be quite intricate, as demonstrated by volunteer computing.

Volunteer computing is a form of distributed computing where the processing power needed to solve a problem is provided by volunteer-administrated computers working part-time on small fragments of the problem. Volunteering usually takes place by downloading and installing a client application, which announces itself to a project server. The server splits the data to be processed into independent work units, distributes the work units among the volunteers and collects the results. Volunteers are free to leave the project at any point and generally receive no material compensation for their contribution.

In the best case, volunteer computing is an arrangement where everybody wins. The people running the project, who are usually researchers, gain access to computing resources at a fraction of the cost of buying the equivalent in new hardware, and the volunteers get pleasure from contributing to a worthy cause and from belonging to a community of people with a shared interest. Society benefits from the new knowledge acquired and from the more efficient use of existing resources.

None of these benefits come automatically: they are contingent on the establishment of a reciprocal relationship between the researchers and the volunteers. Such a relationship can not exist unless certain fundamental conditions are satisfied. The availability of enabling technology is an important condition, but there are other conditions, social rather than technical, with considerable ethical implications. Particularly notable is the necessity of trust, because both parties are making themselves vulnerable in the arrangement.

The diverse technological and social issues surrounding volunteer computing are illustrative of the complex interplay of people, knowledge, technology and ethics that is characteristic of knowledge discovery in general. People collaborate with technology and with other people to generate new knowledge, which is often used to create new technology. The social dimension of the discovery process raises ethical concerns related to the expectations and responsibilities of various stakeholders, and the knowledge itself may have ethical implications depending on who has the power to apply it and how.

In this paper we explore the social component of the knowledge discovery process and especially how it manifests itself in volunteer computing. We find concrete examples of social and ethical issues that need to be settled to ensure the success of a volunteer computing effort, thus demonstrating the inadequacy of the current technology-centred process model of knowledge discovery. From volunteer computing we expand the discussion to present a view that covers all possible stakeholders and their legitimate concerns. The objectives set for a knowledge discovery project should be based on the complete list of relevant stakeholder concerns.

The next section examines the concept of knowledge in knowledge discovery and its interactions with technology and ethics. A more detailed introduction to volunteer computing is then given, with several examples of active projects. This is followed by a discussion of the stakeholders of volunteer computing and the ethical issues associated with their relationship. Finally, we expand this perspective to the stakeholders of knowledge discovery in general.

## **KNOWLEDGE, TECHNOLOGY AND ETHICS**

There is a two-way relationship between knowledge and technology. The development of new technology is based on existing knowledge, and conversely, the search for new knowledge is aided by existing technology. Data mining illustrates this cycle neatly: technology (measuring instruments, databases, algorithms) is used to generate knowledge (predictive or explanatory models), which is then incorporated in technology (software applications).

Besides knowledge discovery, another important concept bridging the gap between knowledge and technology is knowledge representation. This refers to creating data structures specially designed for storing not only data but also the semantics of the data. Databases based on such data structures are known as knowledge bases.

The notion that knowledge can be acquired and possessed by technology has the potential to stir up philosophical controversy. This has nothing to do with storing knowledge per se – that has been done since the invention of writing, and even today some of the systems called knowledge bases are actually barely more than digital reference books. What makes knowledge discovery and knowledge bases philosophically interesting is that they appear to challenge the exclusive claim of humans to knowing things. The dependence of computers on humans diminishes when computers no longer only store and retrieve knowledge but can also generate more knowledge and use it to solve problems without human intervention. The more computers come to resemble independent actors instead of tools, the more tempting it is to say that they indeed possess knowledge.

Whether it can truly be said that computers have the ability to know things is a question beyond the scope of this paper, but the nature of the relationship between humans and computers in knowledge acquisition is not. This relationship is becoming less and less straightforward as computing technology evolves and is able to tackle more and more complex tasks. The key to successful knowledge discovery lies in seamlessly combining the strength of computers, arithmetic prowess, with the strength of humans, creativity. Humans are, in effect, partially outsourcing their thought just like they have partially outsourced their memory by making records of things they know.

The implication of the preceding paragraphs is that while it is debatable whether machines can independently discover and possess knowledge, it is well established that they have entered a symbiotic relationship with humans in the acquisition and storage of knowledge. Computers without human initiative and guidance would be useless, but on the other hand, humans without the aid of computers would be severely crippled in their attempts to understand the huge quantities of data characteristic of empirical research today.

Both knowledge and technology are sources of considerable power in the sense that someone in possession of them can accomplish things that would otherwise be beyond their reach. Power can be wielded benevolently or malevolently, so knowledge and technology are inevitably linked with ethics. Not every piece of knowledge or every item of technology is ethically controversial, but the debates generated by those that are can get very hot. Examples are abundant among the scientific discoveries of the modern era – nuclear energy, stem cells and cloning, just to name a few.

Especially interesting from the perspective of this paper are the intersections of knowledge discovery with ethics. There are many such intersections, as demonstrated by the survey made in (Tuovinen & Rönning, 2005). The most interesting of these are the applications where knowledge discovery techniques are applied to data regarding human individuals, because they give us yet another role in which humans may appear in the knowledge discovery process: as objects of study. As such they are entitled to be treated with certain respect by the people doing the studying. In particular, they have the right to expect the data concerning them to be handled in a way that respects their privacy.

When discussing privacy preservation in knowledge discovery, we can make a distinction between two cases based on who is gathering the data. We shall consider first the case where the data is collected by a company, a research institute or some other organisation with no authority status. In this case collecting the data generally requires the consent of each individual, and the collecting organisation has an obligation to state clearly what the data may be used for and by whom. Violating these standards may lead not only to moral outrage but to litigation as well.

The situation is somewhat different when the data is collected by a public authority. Organisations such as police departments and tax offices hold a legal mandate to gather information about the people within their jurisdiction and to use the information to enforce laws and regulations. Problems would soon ensue if acquiring the information relied entirely upon the goodwill of the people, so in such special cases the common good is considered to override the right of individuals to control information about them.

Fundamentally, however, the two cases are not that different in a society that recognises its citizens' right to privacy. The law grants the government the ability to breach this privacy, but it must be to protect some interest considered even more important and the scope of the breach must not be out of proportion. By this principle, privacy violations are sometimes necessary, regardless of whether knowledge discovery technology is involved, but they should be rare and limited in extent, again whether or not knowledge discovery is involved. We will therefore treat such breaches as exceptions and drop the distinction between our two cases for the rest of our discussion.

Assuming, then, that sensitive data can only be gathered with the consent of the individuals with whom it is associated, it might seem that there is nothing to discuss. However, the reason why privacy issues in knowledge discovery have sparked such interest is that even if the data in itself is not sensitive, the conclusions may be, because data mining techniques can identify surprising connections between data items that, considered separately, appear unrelated and uninteresting (Tavani, 1999). To counter the unwanted effects of this ability, privacy preserving data mining techniques have been studied (Verykios et al., 2004). The goal of these is to allow useful knowledge regarding a population to be derived while making it impractically difficult to connect any of the knowledge with specific individuals in the population.

The existence of sensitive data, or data from which sensitive conclusions may be drawn, in a database is not necessarily in itself a problem, if the organisation that administers the database is committed to using the data responsibly. However, there is always the possibility that the data is stolen by someone with no intention to honour such commitments. Therefore an important part of the responsibility of the database owner is to make sure that the security of its information systems is up to modern standards.

The requirement of proper information security also means that the data must be protected against corruption. If the data is bad, it is not realistic to expect that mining it will produce good knowledge. On the other hand, bad results may also come from good data if it is mined using inappropriate methods. In either case, the upshot may just be that some resources are wasted on a knowledge discovery effort that produces nothing useful, but if the results are to be used in a way that affects the people whom the data concerns – in a medical application, for instance – then incorrect results, or the delay from finding them incorrect and revisiting the data, can be harmful to them. In such cases it is especially important to look after the integrity of the data and the knowledge derived from it.

## **VOLUNTEER COMPUTING**

Volunteer computing can be defined as scientific computing by means of an ad-hoc distributed system dependent on computing resources provided of their own volition by individuals with no formal affiliation with the project. Instead of being formally employed, the participants choose to contribute because they find the work interesting and rewarding. The resources they contribute are typically desktop PCs with comparably modest Internet connections, available to the project only part-time and at unpredictable times. (Anderson, 2004)

The usual way of participating in a volunteer computing project is to download and install a client application distributed at the project website. The client connects to the project server, which assigns a work unit to the client. The client processes the work unit, connects to the server again, sends its result and receives another work unit (see Figure 1). This cycle continues until the project ends or the volunteer ceases to contribute. The volunteer can also choose the amount of resources allocated for the client; for instance, the volunteer may wish the client to be active only when the computer is running idle so that it will not consume resources when the machine is being used for something else.

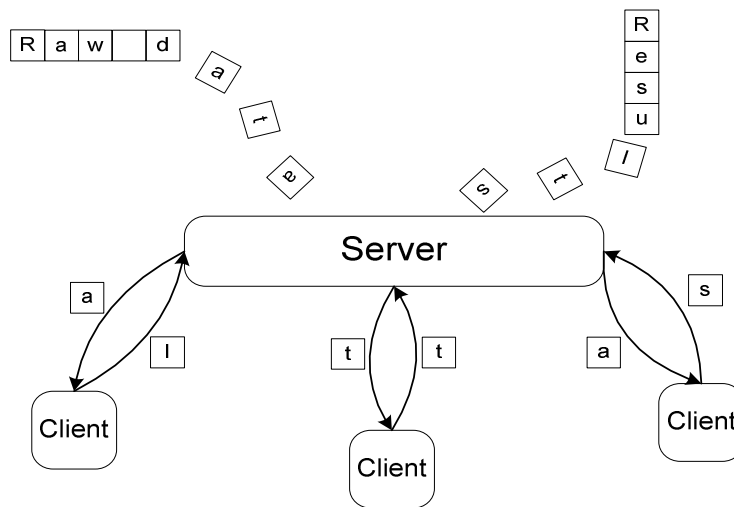


Figure 1. A volunteer computing server splits the data to be processed into work units and sends each work unit to a client. From the partial results returned by the clients the server pieces together the total result.

Since the volunteers may make their computers only sporadically available and are free to drop out any time they wish, volunteer computing faces technical challenges besides those that are always encountered in distributed computing. The project server can not contact the clients but has to wait for them to contact it, so the server is in a constant state of uncertainty regarding the state of the clients. The server therefore has to make decisions such as whether to keep waiting for a given client to return its result based on incomplete information. Volunteer computing platforms like BOINC (Anderson, 2004) and SLINC (Baldassari et al., 2006) have been created to solve this and other technical difficulties, allowing researchers to concentrate on formulating the task of the clients instead of trying to tackle computer engineering problems that have nothing to do with their research

Volunteer computing has its roots in the GIMPS (<http://www.mersenne.org>) and distributed.net (<http://www.distributed.net>) projects. Both projects are still active; GIMPS searches for prime numbers belonging to a class known as Mersenne primes, whereas distributed.net has worked on several problems in cryptography and abstract mathematics. It is worth noting that the computational methods used in these projects are straightforward brute-force techniques – for instance, the RC5 cryptography subprojects at distributed.net rely on simply trying every possible key until the right one is found – so not all volunteer computing projects use data mining.

The project that really brought volunteer computing to public attention is SETI@home (<http://setiathome.ssl.berkeley.edu>), launched in 1999. SETI, an acronym for the Search for Extraterrestrial Intelligence, refers to efforts aimed at finding evidence of the existence of extraterrestrial civilisations from data recorded electronically by astronomers surveying the sky. The SETI@home data consists of radio signals collected using the 305-metre dish of the Arecibo Observatory in Puerto Rico. In 2004, the computers contributing to SETI@home were estimated to provide a sustained processing rate of over 70 TFLOPS ( $7 \times 10^{13}$  floating-point operations per second), twice the rate of the largest conventional supercomputer at the time (Anderson, 2004). This figure gives an indication of the amount of processing power potentially available for a volunteer computing project to harness.

SETI@home, like many other volunteer computing projects, now runs on the BOINC platform. At the BOINC website (<http://boinc.berkeley.edu>) there is a list of projects that gives an idea of the variety of problem domains where volunteer computing can be a useful tool. There is also a list of scientific

publications produced by BOINC projects; some of these appear in highly prestigious journals such as *Nature* (Murphy et al., 2004; Stainforth et al., 2005) and *Astrophysical Journal* (Cole et al., 2008). The accumulation of published results shows that it is possible to do serious scientific research while relying on volunteers to provide the computing resources needed for data analysis.

There are variants of the standard model of volunteer computing where the work units are processed by the volunteers themselves instead of their computers. One of these is Galaxy Zoo (<https://www.galaxyzoo.org>), which recruits volunteers for visual classification of galaxy images. For each image, the volunteers are requested to answer a series of questions on various features of the galaxy. This approach is comparably simple to implement because no client application is needed – the volunteers simply use a web browser to log on to the Galaxy Zoo website, which presents the images and questions to them.

The Foldit project (<http://fold.it>) adopts a more elaborate approach. There is a client application that the volunteers download and install, but instead of performing computations in the background it allows the volunteers to play a puzzle game. The objective of the game is to fold a protein into the most stable state it can assume, given its amino acid structure. This information is used to predict the shape the protein will naturally fold into, which in turn is crucial knowledge for understanding the function of the protein.

The list of volunteer computing projects given above is not exhaustive, but it is illustrative of the wealth of research areas where volunteer computing can be applied and the variety of technical approaches that can be adopted. Technical implementation is not everything, however; there are also social and ethical matters to be considered, although these may not be immediately obvious. We explore these issues in the next section.

## **ETHICAL ISSUES IN VOLUNTEER COMPUTING**

As seen in the previous section, there are many technical challenges in volunteer computing, but several projects have been able to overcome these and succeed in producing new knowledge. Thanks to free platforms such as BOINC, a researcher can start a new volunteer computing project without extensive computer skills or a great investment of money or time. This has brought volunteer computing closer to being a universally accessible tool for knowledge discovery.

BOINC and similar systems open the necessary channel of communication between researchers and volunteers, but this is just the beginning. The computing platform provides the tools for collaboration, but it is up to the people participating in the project to use those tools productively. It turns out that some of the most important issues in volunteer computing are not technical but ethical. Three major themes can be singled out: trust, fairness and persuasion. Below we discuss each of these separately, but we find that they are intertwined in many ways.

### **Trust**

A research project that depends on volunteer computing will not get anywhere unless the researchers are able to muster a large enough group of people willing to contribute to the research. Reaching people who find the project interesting is a matter of properly designed and executed advertising, but interest alone is not enough to make one a potential volunteer; trust is also a crucial requirement.

Trust is important because downloading and installing software without knowing exactly what it does is always a risk. Contributing to a volunteer computing project means installing a client application and allowing it to use the processing capacity and network connection of the client machine, and it takes considerably higher than average computer skills to be sure that the application will not have any undesirable effects. The researchers running the project therefore need to get the public to trust that the software they are distributing is not malicious or dangerously defective.

On the other hand, the researchers are also taking a chance in trusting the volunteers. They hope that those who install the client are motivated by a sincere wish to contribute to the research, and mostly this is a safe assumption, but the possibility of somebody joining the project with the intention of sabotaging

it can not be ruled out. For example, a rival research team might want to hinder the work by trying to feed false results to the project server.

As Anderson (2004) points out, the volunteers in volunteer computing are in an asymmetric relationship with the researchers. One of the implications of this asymmetry is that volunteers and researchers must rely on different means to gain the trust of the other party. In fact, it would be more accurate to say that the establishment of trust is entirely in the hands of the researchers – it is up to them to make sure both that the volunteers can trust them and that they can trust the volunteers.

With potentially hundreds of thousands of aspiring volunteers, individually evaluating each one to prevent potential saboteurs from joining is not a feasible task. Instead, the researchers must accept, and adjust to, the fact that some percentage of the volunteers may be trying to sabotage the effort. In practice this means using technology that minimises the effects of malicious behaviour on the efficiency and results of the computation.

The traditional technique for detecting and discarding falsified results is redundant computing, where each work unit is sent to several clients and the clients vote to determine the canonical result for the unit (Anderson, 2004). When the number of clients whose results agree reaches some threshold  $M$ , their result is chosen as the canonical one. Increasing  $M$  makes the system more resistant to bad results, but it also increases the time it takes to finish the computation. Sarmenta (2001) has proposed an open framework where voting can be combined with other techniques such as spot-checking and blacklisting to calculate a credibility measure for each result. Using a combination of methods reduces the slowdown incurred by a given error-tolerance level compared to the case where only voting is used.

So, when it comes to finding trustworthy volunteers, the best the researchers can do is give everyone a chance and try to weed out the dishonest ones based on their output. Not much here is of philosophical interest: there is nothing ethically ambiguous about sabotage – it is clearly wrong, at least assuming that the research itself is not ethically questionable – and improving the error tolerance of computations is a technical issue. Blacklisting clients without human discretion is somewhat problematic because an erroneous result is not necessarily intentionally falsified, but this is a tangential issue, and to avoid a diversion from the main thread of discussion, we shall not treat it in any detail.

It is worth noting here that although implicitly trusting every aspiring volunteer has its risks, it also helps bring in honest volunteers. People wishing to join a volunteer computing project appreciate it if joining is made easy for them, whether or not they consciously think of it as a sign of trust. For example, as Sarmenta (2001) points out, requiring volunteers to provide a stronger form of identification than an email address would make it more difficult for blacklisted saboteurs to rejoin, but it would probably also deter many good volunteers. We shall return to this topic in the subsection on persuasion.

Given that the researchers do not know who the volunteers are, except in the rather trivial sense that each volunteer is required to register, it could be argued that the way the researchers relate to the volunteers is not really a matter of trust, except in the sense that they show a trusting attitude when they presume that the majority of registered volunteers are honest. If it can be safely assumed that most participants are not out to sabotage the project, tolerance mechanisms built into the computing server can be used to cancel out the effect of those who are. We see a different picture, however, when we look at the way the volunteers relate to the researchers.

One aspect of the asymmetric nature of volunteer computing is that the volunteers are in a better position to gather information about the other party and to decide whether they want to have any dealings with them. This seems appropriate, since it is the volunteers who are being asked to give someone else partial control of their property. It does, however, raise the question of what the researchers can do (and also what they should not do) to convince the volunteers that they can be trusted with what they are asking for.

Also here technology may help the researchers, although in a more indirect manner. Since BOINC, for instance, is used by dozens of projects and the client has been installed on hundreds of thousands of computers, it is likely that if there were some feature or defect in the software that may damage client systems, somebody would soon discover and report it. If there are no such reports to be found, it is an

indication to potential volunteers that it is relatively safe to run the software. Still, the researchers are ultimately responsible for the software they use and should act accordingly to protect it against attempts to sneak malicious code into it.

The role of word-of-mouth in assessing the trustworthiness of software is a specific example of the more general truth that reputation is an important factor in establishing trust. One's expectation of someone's trustworthiness is influenced by the recommendations of others, and the strength and direction of the influence are in turn affected by the perceived trustworthiness of the recommenders. Several authors have undertaken to describe the dynamics of trust and reputation in networked computer systems as formal models; see e.g. (Abdul-Rahman and Hailes, 2000).

In the future it may be possible for organisations and individuals to rely on such models to automatically verify the trustworthiness of the organisations and individuals they interact with, but currently, barring a few special cases, there are no such shortcuts available. An organisation wishing to recruit computing volunteers for a research project can therefore mostly just hope that it has a positive public image and try to make sure that potential volunteers associate the image with the project.

A research group has a good chance of accomplishing this by being thoroughly open about itself: which scientific institution it is affiliated with, which institutions it has collaborated with, which research topics and projects it has worked on, what it is trying to achieve now, who are currently working in the group, who have worked there in the past. This gives a potential volunteer, sympathetic to science in general, a chance to connect the group to something or someone he or she already knows and thinks positively of. Such connections lend the group credence by association.

About openness it is worth noting that when extended to other aspects of volunteer computing, it leads to a trade-off between how much the researchers can trust the volunteers and vice versa. In particular, the more details are disclosed about the computing software, the more confidence the volunteers can have that the software is not harmful, but the easier it is for malicious volunteers to modify it to produce incorrect results. However, since it is the volunteers who are making their own computers vulnerable and since the researchers have ways to defend the integrity of the results against malicious behaviour, it is best for the volunteer computing software to be open source.

## **Fairness**

Volunteers deserve to be treated fairly – this is a very basic right, not negated by the fact that they are volunteers. This does not mean that they must be compensated materially for their contribution, since volunteer computing projects are typically nonprofit research efforts. Instead of compensation, the key fairness factor in such projects is acknowledgment.

A common approach to giving due acknowledgment to volunteers is credit. Credit is basically an immaterial form of compensation, an accumulative number that represents the amount of work a volunteer has contributed to a project. Based on credit it is possible to compile various statistics, giving volunteers the chance to earn some public recognition by appearing on a list of top contributors.

How credit accumulation is determined is a matter of some significance, because the measure should be objective to such an extent that keeping statistics makes sense. If it is not possible to make meaningful comparisons between the credit values of different contributors, then the credit system serves no purpose except to confirm to the volunteers that their computers are doing something for the project. The amount of credit earned is, by itself, an almost meaningless number; knowing whether it is a little or a lot requires a context, and this context is provided by the ability to compare the number among peers.

The other side of credit is that no-one should be able to gain undeserved credit, since that would be unfair to those who have acquired theirs with honest work. Conveniently, redundancy may again provide the solution. BOINC, for example, uses an accounting scheme where clients are not automatically given the amount of credit they claim for their results but receive the minimum or average of the claimed credit of all correct results instead (Anderson, 2004). This prevents clients from bolstering their credit accounts with dishonest claims.

Overall, crediting volunteers for the CPU time they have donated is a good way to acknowledge their work. Distributing credit for publishable results is a different matter, however. If the researchers write, for instance, an academic journal paper on results achieved with the help of volunteer computing, what is the proper way to acknowledge the significance of the volunteers' contribution? The answer is largely dictated by what is practical: it is simply not possible within reason to include the names of all volunteers in the paper. On the other hand, good academic form demands that if a nonauthor has contributed to a publication, the authors mention the contribution. A collective expression of gratitude to the volunteers is therefore both necessary and sufficient.

A special case worth treating separately is when a discovery can be traced to a single computer or a small number of computers. For example, it is conceivable that if SETI@home one day achieves its goal and confirms an extraterrestrial radio signal as a transmission by an alien civilisation, it will be possible to identify the work units in which the signal was found and the clients that processed the work units. It might therefore be possible to associate the discovery with a relatively small number of volunteers, who might expect some kind of special recognition.

A fact that speaks against granting such recognition is that the distribution of work units among volunteers is a random process, albeit weighted by the amount of resources devoted to the project by each volunteer. Granting coauthorship in a scientific publication on such basis seems inappropriate, and even a special mention in the acknowledgments section might be unmerited. Perhaps the best option would be to give the names of these volunteers in a less formal context, e.g. a press release or a news item at the project website. Besides, some of the volunteers might not even want their identities published, in which case the researchers would have to respect their privacy.

A version of this scenario has already been seen at distributed.net, where volunteers have participated in competitions to solve secret-key ciphers of increasing strength. To motivate these efforts a cash prize of 10 000 USD for whoever finds the correct solution has been offered by RSA Laboratories (<http://www.rsa.com/rsalabs/node.asp?id=2100>). So far distributed.net has won two of these challenges and is organising the next one itself since RSA Labs has discontinued the contest. The distribution of prize money follows a formula: the volunteer who finds the winning key gets a part, another part goes to distributed.net for providing the necessary infrastructure, but most of the money is donated to a nonprofit organisation jointly selected by all volunteers.

The difference between this and the SETI@home scenario is that finding the secret key is not a new discovery, so there is no need to argue about who gets credit for it. The prospect of winning some cash makes participating in a secret-key challenge akin to entering a raffle, and whereas winning money in a raffle is generally considered acceptable, determining paper authorship by such means would most definitely be unacceptable. Money, unlike authorship, does not inherently belong to anyone, which leaves distributed.net free to share it in a way that achieves a good balance between advancing the public good and offering incentives to volunteers. The chosen balance seems to meet the approval of the participants, considering that when the RC5-64 cipher was solved, distributed.net itself was voted as the nonprofit organisation to receive 60% of the prize money.

## **Persuasion**

The act of persuading is not inherently ethical or unethical. However, people may be persuaded to do ethical or unethical things, and the means of persuasion may be ethical or unethical. We conclude this section with a discussion of the persuasion techniques available to researchers wishing to recruit computing volunteers and the ethical issues associated with those techniques.

Persuasion through the use of computers has been extensively studied by Fogg (2003). Several of the persuasion techniques he has identified are already being used to recruit volunteers and to make them stay. It is hard to say to what extent researchers who use volunteer computing are aware of the theory of persuasion by information technology, but it would be good if they were, since it would both help them persuade effectively and help them avoid persuasion techniques that are ethically unsound.

The good news for researchers is that even if a project has proved difficult to sell to financiers, it may not be all that hard to find volunteers for it. It is a fairly small commitment to let one's computer do

some work for someone else when it would otherwise be idle, and many people are happy to make it if they find the work interesting and agree with its goals. SETI@home is again a good example: few things are more exciting to a curious mind than the prospect of receiving a message from an extraterrestrial intelligent species, so the researchers are in a good position to write a project description that attracts curious minds.

It is good for researchers to have a grand vision – contact with another civilisation or a cure for a serious disease, for example – and it is not wrong to communicate it to potential volunteers in a fashion that appeals to their sense of romance and heroism. However, the researchers should clearly discern romance from realism and be cautious about what they promise to achieve. In the case of SETI, for instance, nobody knows for sure if there even is anyone out there for us to have contact with, so it would not be honest to imply that the project will certainly find what it is looking for. Scenarios presenting possible outcomes should be accompanied by the assumptions on which they are based and the uncertainties associated with the assumptions.

We have already mentioned the simplicity of joining a volunteer computing project and how it may be viewed as a sign that the researchers trust the volunteers. According to Fogg, simplicity is in fact one of the most powerful persuasion techniques available. If a potential volunteer is allowed to join after answering just a couple of simple questions, it is likely that he or she will. If there are many questions and the volunteer constantly finds him- or herself wondering whether he or she wants the researchers to have the information requested, it increases the likelihood that the volunteer will not complete the joining process. Yet another way to view this is that by not requiring volunteers to supply such information as their phone numbers or even their real names, the researchers protect the privacy of the volunteers. Privacy issues may thus prove significant even if no such issues are raised by the data.

Rewards are another powerful way to persuade, and even rewards that are completely immaterial in composition may be far from immaterial in consequence; Anderson (2004) has observed the motivating effect of credit in the case of SETI@home, and Fogg (2003) confirms the general principle. Credit is particularly effective thanks to the competition it encourages among volunteers. The competition, in turn, so long as it remains friendly, fosters a certain sense of community, which is a healthy value in itself and also further strengthens the commitment of the volunteers to the project and increases the satisfaction they get from their participation.

The community aspect of volunteer computing is probably something that researchers could use to a considerably greater effect than they do now. An interesting idea would be to borrow elements from popular social networking sites such as Facebook; for instance, volunteers could have their own profile pages where they could display merit badges, awarded by the computing server when a specific level of accumulated credit is reached. Credit could also be redeemable in an online shop for virtual items, or even promotional t-shirts and other concrete items that are relatively cheap to make. In fact, the capabilities of existing services could be directly used: there could be, for example, a SETI@home Facebook application allowing volunteers to display their achievements, get in touch with other volunteers and invite their friends to join. This would provide the added benefit of simplicity, since volunteers could build their communities at sites they visit frequently anyway.

The client application can also be persuasive. This is obvious if the application is a game as in the case of Foldit, but it is true even if the client is not interactive. For example, there is Fogg's (2003) principle of attractiveness, which states that if a computing technology is visually attractive, it is likely to be more persuasive. SETI@home, among others, leverages this principle by providing colourful graphics, showing in real time what the application is doing. Volunteers can set the client to run as their screen saver, allowing them to get up to date on the progress they are making every time they glance at the computer screen.

A question worth considering is what information the client application should display to the volunteer running it. To maximise the persuasive qualities of the application, the information should be interesting to the volunteer and presented in an attractive way that makes it easy for a layman to grasp its significance. On the other hand, it is not in the interest of the researchers to let important discoveries come to public attention before they themselves are ready to publish them. Here is another possible

trade-off involving openness: instant feedback of interesting results can act as a reward for volunteers, but it also raises the question of how much information the volunteers can be trusted with before it is made officially public by the researchers.

In practice, though, this may not pose a problem, since the amount of information the client application can impart to volunteers is limited anyway because of the distributed nature of the computations and because the volunteers generally are not trained to interpret the results. Thus it seems unlikely that a volunteer could independently draw any sensational conclusions based on the output of his or her copy of the client software alone. The trust issue is averted in this case simply by considering what is really known; having the client convey an impression of imminent breakthrough might be exciting and persuasive, but such an impression would be based on guesswork at best.

## KNOWLEDGE DISCOVERY AS A SOCIAL EFFORT

The lesson of the previous section summarised is that there is a social aspect to knowledge discovery through volunteer computing that must not be neglected. It is important to have good technology, but it is also important to have good relationships among the people involved or the effect of the technology will be blunted. This lesson can be extended to knowledge discovery in general.

The most central position in knowledge discovery is held by the experts coordinating the discovery process. In the case of SETI@home, for example, the group of experts consists of the researchers running the project server, but generally there can be two kinds of experts: domain experts, intimately familiar with the problem domain, and technology experts, with the skills and knowledge necessary for the successful application of knowledge discovery technology. The first social relationship that has to work smoothly for a knowledge discovery effort to succeed is the one between domain and technology experts.

A good relationship between different kinds of experts requires two things: that the domain experts can communicate the problem to be solved to the technology experts, and that the technology experts can communicate the capabilities of the technology to the domain experts. Based on this mutual understanding the two groups of experts can work together to find a way to use the technology to solve the problem. The availability of general-purpose discovery tools may have diminished the role of technology experts to some extent, but even a very good tool does not yield good results when applied blindly, so there must at least be someone in the domain expert team with a solid understanding of technology.

The relationship between technology and domain experts is approximately analogous to the relationship between professionals and clients in software engineering and is therefore bound by similar ethical principles. Software engineering ethics is a rich field of study and covering it in any significant detail falls outside the scope of this paper; let it suffice to say that there are obligations to be fulfilled both ways, since there is a contract – written, verbal or unspoken – that both parties should honour. For relevant literature see e.g. (Bayles, 1989) and (O’Boyle, 2002).

The diagram in Figure 2 shows the five stakeholders of knowledge discovery we have identified. The four roles that people may play in the discovery process are denoted by ellipses, with society as a whole representing the fifth stakeholder. The arrows denote interactions among the stakeholders in terms of expectations; each arrow is annotated with the expectations that a stakeholder should seek to fulfil in its relationship with the other stakeholder. Some stakeholders also have an arrow pointing from itself to itself, meaning that members of a stakeholder group expect something from other members of the same group.

There are multiple names for some of the stakeholder groups; the ones called **researchers** we have also referred to as technology experts and data miners. They are the ones with the expertise required to apply knowledge discovery technology, which puts them at the centre of the figure, interacting with all other stakeholders. The group also has an expectation arrow pointing from itself to itself, because in the scientific community in particular, sharing one’s results is the norm; a researcher expects to have access to new methods and tools developed by his or her colleagues around the world.

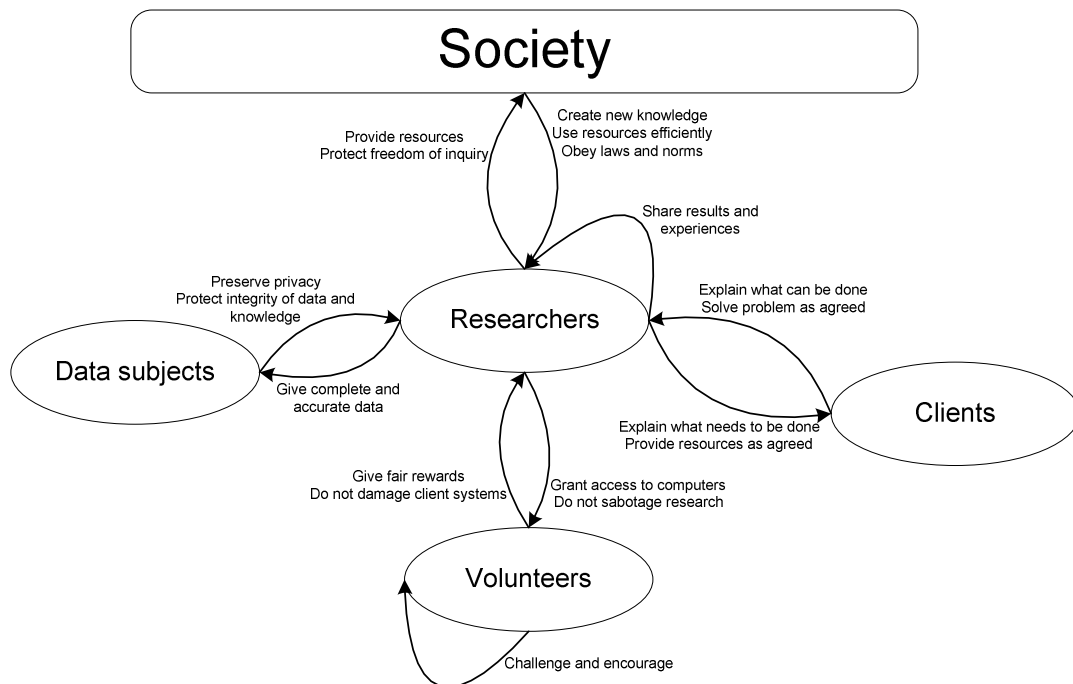


Figure 2. The stakeholders of knowledge discovery. Each stakeholder expects something from one or more other stakeholders and, in some cases, from other members of the same stakeholder group. The arrows depict these expectation-based relationships.

**Clients** we use as a general term for the people who know the problem domain and wish to affect it using knowledge acquired with the help of the researchers. The people we have referred to as domain experts form a subgroup of this stakeholder group. In some projects the clients do not exist as a distinct group, in which case their role is distributed among other stakeholders; in pure science, for instance, the researchers themselves are the domain experts and the acquired knowledge is received by society and other researchers.

People used in the knowledge discovery process as data sources are referred to in the figure as **data subjects**. Previously we have only discussed their rights, so it is worth noting that they also have at least one obligation: if they have agreed to give data to the researchers, it should be complete and accurate.

The **volunteers** were extensively discussed in the two preceding sections, so there is little more that can be said about them at this point. There is still a lot of untapped processor time left on the desktops of the world, so the dynamics of the researcher-volunteer relationship are well worth studying further. The arrow from this stakeholder to itself reflects the desire of volunteers to form a community; researchers would do wisely to find new ways to encourage and facilitate this.

Finally, **society** is the umbrella under which all of the above takes place. All the other stakeholders are subject to it and interact with it, but from the perspective of the knowledge discovery process, the researchers represent the main point of contact through which society's expectations are propagated into the process. Of course, the relationship is not one-sided – researchers also expect things from society, including both concrete necessities such as funding and abstract circumstances such as freedom of inquiry.

## CONCLUSIONS

Volunteer computing is a form of distributed computing made possible by the proliferation of powerful personal computers and reasonably fast Internet connections. A project server assigns tasks to copies of a client application running on computers administered by volunteers, who allow the client to consume some of their processing resources because they find the project interesting and want to contribute to its success. Many projects have produced noteworthy results, showing volunteer computing to be a valuable tool for knowledge discovery.

Open platforms such as BOINC make it relatively easy for a research team to start a volunteer computing project, allowing it to gain access to potentially huge computing resources without making a major monetary investment. An infrastructure that solves the technical issues involved is not enough by itself, however; there are also social and ethical issues to be solved, stemming from the fact that the researchers and the volunteers represent two groups of people trying to work out a mutually satisfactory deal. There can be no such deal if, for instance, the two groups do not trust one another. Some of the ethical issues of volunteer computing are accounted for in the design of BOINC and similar systems, but there are also issues that have not been adequately addressed and for which a technological solution may not even be possible.

This situation is illustrative of the situation in knowledge discovery in general: current process models are technically oriented and completely omit the social dimension of the discovery process. We consider this an important omission that, if it persists, will prevent knowledge discovery technology from achieving its full potential. We have therefore explored the social and ethical factors involved in volunteer computing and, by extension, knowledge discovery. We have identified the stakeholders of the discovery process and their interactions, bringing all of them together in a compact view that practitioners can use to make sure they take into account the justified expectations of all stakeholders. The mark of a truly successful knowledge discovery effort is that all participants come out satisfied, including those with no interest in the knowledge produced.

## ACKNOWLEDGMENTS

We would like to thank our colleague Dr Perttu Laurinen for his insightful comments at the early stages of composing this paper. L. Tuovinen wishes to thank the Graduate School in Electronics, Telecommunications and Automation (<http://signal.hut.fi/geta/>) and the Tauno Tönning Foundation (<http://www.tonninginsaatio.fi/>) for funding his postgraduate work.

## REFERENCES

- Abdul-Rahman, A. and Hailes, S. (2000) Supporting Trust in Virtual Communities. *Proceedings of the 33<sup>rd</sup> Hawaii International Conference on System Sciences*.
- Anderson, D.P. (2004) BOINC: A System for Public-Resource Computing and Storage. *Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing*, 4 – 10.
- Baldassari, J., Finkel, D. and Toth, D. (2006) SLINC: A Framework for Volunteer Computing. *Proceedings of the 18<sup>th</sup> IASTED International Conference on Parallel and Distributed Computing and Systems*.
- Bayles, M.D. (1989) Obligations Between Professionals and Clients. In *Professional Ethics*, 2<sup>nd</sup> ed., reprinted in Johnson, D.G. (ed.), 1991, *Ethical Issues in Engineering*, 305–316.
- Berthold, M.R., Cebron, N., Dill, F., di Fatta, G., Gabriel, T.R., Georg, F., Meinl, T., Ohl, P., Sieb, C. and Wiswedel, B. (2006) KNIME: The Konstanz Information Miner. *Proceedings of the 4<sup>th</sup> Annual Industrial Simulation Conference, Workshop on Multi-Agent Systems and Simulation*.
- Chen, J., Fan, Q. and Xu, B. (2004) Research on the Application of Data-mining for Quality Analysis in Petroleum Refining Industry. *Proceedings of the 5<sup>th</sup> World Congress on Intelligent Control and Automation*, 4314–4318.
- Chou, P.B., Grossman, E., Gunopulos, D. and Kamesam, P. (2000) Identifying Prospective Customers. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 447–456.

- Cole, N., Newberg, H., Magdon-Ismael, M., Desell, T., Dawsey, K., Hayashi, W., Purnell, J., Szymanski, B., Varela, C.A., Willett, B. and Wisniewski, J. (2008) Maximum Likelihood Fitting of Tidal Streams with Application to the Sagittarius Dwarf Tidal Tails. *Astrophysical Journal*, **683**, 750–766.
- Fogg, B. J., (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, San Francisco, CA.
- Gersten, W., Wirth, R. and Arndt, D. (2000) Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 398–406.
- Laurinen, P., (2006). *A Top-Down Approach for Creating and Implementing Data Mining Solutions*. Dissertation, University of Oulu, Acta Universitatis Ouluensis C 246.
- Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M. and Clark, R.A. (2004) Data mining techniques for cancer detection using serum proteomic profiling. *Artificial Intelligence in Medicine*, **32**, 71–83.
- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. *Proceedings of the 12<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 935–940.
- Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M. and Stainforth, D. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772.
- O’Boyle, E.J. (2002) An ethical decision-making process for computing professionals. *Ethics and Information Technology*, **4**, 267–277.
- Sarmenta, L.F.G. (2001) Sabotage-Tolerance Mechanisms for Volunteer Computing Systems. *Proceedings of the First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 337–346.
- Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J. and Allen, M.R. (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406.
- Tavani, H.T. (1999) Informational privacy, data mining, and the Internet. *Ethics and Information Technology*, **1**, 137–145.
- Tuovinen, L. and Rönning, J. (2005) Balance of power: the social-ethical aspect of data mining. *Proceedings of the Sixth International Conference of Computer Ethics: Philosophical Enquiry*, 367–379.
- Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y. and Theodoridis, Y. (2004) State-of-the-art in Privacy Preserving Data Mining. *SIGMOD Record*, **33**(1), 50–57.
- Zaffalon, M., Wesnes, K. and Petrini, O. (2003) Reliable diagnoses of dementia by the naïve credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine*, **29**, 61–79.
- Zhang, C.-H., Di, L. and An, Z. (2003) Welding Quality Monitoring and Management System Based on Data Mining Technology. *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, 13–17.