Preface

Over the last years, ubiquitous computing has started to create a new world of small, heterogeneous, and distributed devices that have the ability to sense, to communicate and interact in ad hoc or sensor networks and peer-to-peer systems. These large-scale distributed systems, in many cases, have to interact in real-time with their users. Knowledge discovery in ubiquitous environments (KDubiq) is an emerging area of research at the intersection of the two major challenges of highly distributed and mobile systems and advanced knowledge discovery systems. It aims to provide a unifying framework for systematically investigating the mutual dependencies of otherwise quite unrelated technologies employed in building next-generation intelligent systems: machine learning, data mining, sensor networks, grids, peer-to-peer networks, data stream mining, activity recognition, Web 2.0, privacy, user modeling and others.

In a fully ubiquitous setting, the learning typically takes place in situ, inside the small devices. Its characteristics are quite different from currently mainstream data mining and machine learning. Instead of offline-learning in a batch setting, sequential learning, anytime learning, real-time learning, online learning, etc.—under real-time constraints from ubiquitous and distributed data—is needed. Instead of learning from stationary distributions, concept drift (the change of a distribution over time) is the rule rather than the exception. Instead of large stand-alone workstations, learning takes place in unreliable, highly resource constrained environments in terms of battery power and bandwidth.

To explore this emerging field of research, a networking project has been funded since 2006 by the European Commission under grant IST-FP6-021321: KDubiq (knowledge discovery in ubiquitous environments) is a coordination action at the intersection of the two major challenges of highly distributed and mobile systems and advanced knowledge discovery systems. A basic assumption of the project is that what seems to be a bewildering array of different methodologies and approaches for building “smart,” “adaptive,” “intelligent” ubiquitous knowledge discovery systems can be cast into a coherent, integrated set of key areas centered on the notion of learning from experience. The objective of KDubiq is to provide this common perspective, and to shape a new area of research. For doing so, the KDubiq coordination action has coordinated relevant research done on learning in many subfields, including:

- machine learning and statistics
- knowledge discovery in databases or data mining
- distributed and embedded computing
- mobile computing

1 See the website www.kdubiq.org for details about the project.
- human computer interaction (HCI)
- cognitive science

A major goal was to create for the first time a forum to bring these individual research lines together, to consolidate the results that have already been achieved, and to pave the way for future research and innovative applications. For doing so, KDubiq has organized a large number of workshops, summer schools, tutorials and dissemination events to bring together this new community. One important means to focus the activities and discussions was a collaborative effort to provide a blueprint for the design of ubiquitous knowledge discovery systems. A number of working groups on relevant topics have been established. Their goal was to create a conceptual framework for this new line of research, to survey the state of the art, and to identify future challenges, both on the theoretical and the applications side.

The result of this collaborative effort is Part I of this book. This blueprint manifests the vision and serves as a practical guide for further, integrated advances in this field, towards, in the long-term, building truly autonomous intelligent systems.

Overview of the Book. Part I of the book aims to provide a conceptual foundation for the new field of ubiquitous knowledge discovery, discussing the state of the art, highlighting challenges and problems, and proposing future directions. Although at some points technical examples are given for illustration, the aim of this chapter is rather on the non-technical, conceptual side.

While Part I is divided into individually authored chapters, it should be seen as a collaborative effort by the working groups of the KDubiq coordination action. Each chapter was read and commented by the other working group members, and influenced by the discussions and findings of the individual working groups. Hence, the chapters should be seen as an integrated whole.

Part I of the book is structured as follows. Chapter 1 gives an introduction to the topic and the fundamental issues. Chapter 2 provides an overview on three distributed infrastructures for ubiquitous computing. Chapter 3 discusses how the learning setting itself changes in a ubiquitous environment, when compared to a traditional learning set-up. Chapter 4 defines general characteristics of data in ubiquitous environments. Chapter 5 takes up the issues of privacy and security, arguing that they are critical for the deployment and user acceptance of KDubiq systems. Chapter 6 is devoted to the human-centric view of ubiquitous knowledge discovery systems. Finally, Chapter 7 contains a collection of potential application areas for KDubiq, providing pointers to the state of the art, to existing applications (if available) and to challenges for future research.

---

3 In some cases, project partners provided input for some sections, but not for the whole chapter. Where this is the case, it is stated in the footnotes of individual sections.
Part II contains selected approaches to ubiquitous knowledge discovery and treats specific aspects in detail. The contributions have been carefully selected to provide illustrations and in-depth discussions for some of the major findings of Part I.

The contribution by Antoine Cornéujols takes up in greater detail two fundamental challenges for learning in ubiquitous environments: incrementality and non-stationarity. The chapter by Severo and Gama investigates change detection in temporal data, when the assumption of stationarity is not met. Sharfman, Schuster, and Keren have the topic of monitoring changes in highly distributed data stream systems using a geometric approach. The contribution by Inan and Saygin addresses another highly important dimension of ubiquitous knowledge discovery: privacy. It discusses the problem of clustering horizontally partitioned spatio-temporal data (the data sets at the nodes have the same attributes, but different instances) in a privacy-preserving manner. The chapter by Katharina Morik describes a peer-to-peer Web 2.0 application for collaborative structuring of multimedia collections. The chapter by Rasmus Pedersen broadens the discussion and provides an overview on the topic of learning in micro-information systems. The final chapter by Hillol Kargupta and co-workers describes the MineFleet system, one of the few commercially available ubiquitous knowledge discovery systems. Coordinating a network with more than 50 partner institutions and several hundred individual members is a complex, and sometimes daunting, task. We thank the many researchers and practitioners that contributed to the discussions that led to this book in various ways⁴; the invited speakers that helped us to sharpen our understanding of the research issues involved; the numerous workshop and summer school attendees; the project reviewers for constructive criticism; and the EC project officers Fabrizio Sestini and Paul Hearn for their support. We also thank Tino Sanchez for maintaining the project website. By far the greatest thanks, however, go to Ina Lauth. She has coordinated the network activities for three years and did a superb job in making it both a vibrant and pleasant experience for everyone, thereby having a great share in the successful outcome of the project.

The preparation of this book has been supported by the European Commission under grant KDubiq, IST-FP6-021321, which we gratefully acknowledge. We hope that the reader will agree with us that ubiquitous knowledge discovery in many ways holds potential for radically changing the way machine learning and data mining is done today, and share our excitement about this new field of research.

June 2010
Michael May
Lorenza Saitta

---

⁴ They are too numerous to list individually here, but see the KDubiq newsletter, the members page on KDubiq.org, and the acknowledgments for individual chapters of Part I.
# Table of Contents

## I  A Blueprint for Ubiquitous Knowledge Discovery

<table>
<thead>
<tr>
<th>Section</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction: The Challenge of Ubiquitous Knowledge Discovery</td>
<td>Michael May and Lorenza Saitta</td>
<td>3</td>
</tr>
<tr>
<td>Ubiquitous Technologies</td>
<td>Assaf Schuster and Ran Wolff</td>
<td>19</td>
</tr>
<tr>
<td>Resource Aware Distributed Knowledge Discovery</td>
<td>João Gama and Antoine Cornuéjols</td>
<td>40</td>
</tr>
<tr>
<td>Ubiquitous Data</td>
<td>Andreas Hotho, Rasmus Ulslev Pedersen, and Michael Wurst</td>
<td>61</td>
</tr>
<tr>
<td>Privacy and Security in Ubiquitous Knowledge Discovery</td>
<td>Fosca Giannotti and Yücel Saygin</td>
<td>75</td>
</tr>
<tr>
<td>A Human-Centric Perspective on Ubiquitous Knowledge Discovery</td>
<td>Bettina Berendt and Ernestina Menasalvas</td>
<td>90</td>
</tr>
<tr>
<td>Application Challenges for Ubiquitous Knowledge Discovery</td>
<td>Koen Vanhoof and Ina Lauth</td>
<td>108</td>
</tr>
</tbody>
</table>

## II  Case Studies

<table>
<thead>
<tr>
<th>Section</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-Line Learning: Where Are We So Far?</td>
<td>Antoine Cornuéjols</td>
<td>129</td>
</tr>
<tr>
<td>Change Detection with Kalman Filter and CUSUM</td>
<td>Milton Severo and João Gama</td>
<td>148</td>
</tr>
<tr>
<td>A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams</td>
<td>Izchak Sharfman, Assaf Schuster, and Daniel Keren</td>
<td>163</td>
</tr>
<tr>
<td>Privacy Preserving Spatio-temporal Clustering on Horizontally Partitioned Data</td>
<td>Ali Inan and Yücel Saygin</td>
<td>187</td>
</tr>
<tr>
<td>Nemoz — A Distributed Framework for Collaborative Media Organization</td>
<td>Katharina Morik</td>
<td>199</td>
</tr>
</tbody>
</table>
## Table of Contents

Micro Information Systems and Ubiquitous Knowledge Discovery .......................... 216
  *Rasmus Ulslev Pedersen*

MineFleet®: The Vehicle Data Stream Mining System for Ubiquitous Environments .................................................. 235
  *Hilol Kargupta, Michael Gilligan, Vasundhara Puttagunta, Kakali Sarkar, Martin Klein, Nick Lenzi, and Derek Johnson*

**Author Index** .................................................................................................................................................. 255
Introduction: The Challenge of Ubiquitous Knowledge Discovery

Michael May and Lorenza Saitta

1 Fraunhofer Institut Intelligente Analyse- und Informationssysteme, Department Knowledge Discovery, Sankt Augustin, Germany
2 Università Degli Studi Del Piemonte Orientale Amedeo Avogadro, Alessandria, Italy
michael.may@iais.fraunhofer.de, saitta@di.unipmn.it

1 The Object of Investigation: Ubiquitous Knowledge Discovery

In the past, the development of machine learning approaches was to some extent motivated by the availability of data and increased computational power. Ubiquitous computing bears the promise of stimulating a similar leap forward. Small devices can now be installed in many places, mobile and wearable devices enable registration of large amounts of information, thus generating a wide range of new types of data for which new learning and discovery methods are needed, far beyond existing ones.

Hence, Ubiquitous Knowledge Discovery (KDubiq) focuses on the extension of data mining to modern computing. Notable breakthroughs have revolutionized computer science in the last decade, along several perspectives: hardware, communication/networks, and usage. Gradually, the software world becomes more aware that new algorithmic paradigms are required to make the most of new architectures, to handle new demands, and to sail towards New Intelligence Frontiers. KDubiq is meant to favour the emergence of such new algorithmic paradigms in the domain of data mining and machine learning.

Resting on the above grounds, ubiquitous knowledge discovery is then a new discipline that has its roots in the parent fields of data mining, on the one hand, and ubiquitous computing, on the other. The novelty of the discipline stems from the timely co-occurrence and synergetic integration among the new needs from a multitude of new types of users, the capillary distribution of data and information, and the new resources provided by today’s computational environments.

Relevant research is done on learning in many subfields, including:

- machine learning and statistics
- knowledge discovery in databases or data mining
- distributed and embedded computing
- mobile computing
- human computer interaction (HCI)
- cognitive science.
The object of investigation of KDubiq encompasses the whole process of knowledge discovery in mobile, finely distributed, interacting, dynamic environments, in presence of massive amounts of heterogeneous, spatially and temporally distributed sources of data. More precisely, KDubiq aims at developing algorithms and systems for:

- **Locating data sources** - The current availability of world-wide distributed databases, Web repositories, and sensor networks adds a novel problem to today's knowledge discovery, i.e., the need to locate, in a transparent way, possible sources of data required to satisfy a user's demand.
- **Rating data sources** - As multiple sources may usually be identified as candidate data suppliers, effective evaluation methods must be designed to rate the sources w.r.t. the search target.
- **Retrieving information** - The nature and type of data available today are different from the past, as much emphasis is set on semantically rich, multimodal, interacting pieces of information, which need clever and fast data fusion processes.
- **Elaborating information** - The development of new approaches and algorithms must match the changes in the data distribution and nature. Algorithms are not only faced to the scalability problem (as in the past), but also to resource constraints, communication needs, and, often, real-time processing. Moving the data mining software from a central database to the ubiquitous computing devices demands a design of systems and algorithms which differs considerably from that of classical data mining. This is the price to pay for introducing intelligence in small devices.
- **Personalizing the discovery process** - Due to the large number and variety of potential users of KDubiq, adaptation to individual users is a must. People, in fact, more than ever play a pivotal role in the knowledge handling process: people create data, data and knowledge are about people, and people are often the ultimate beneficiaries of the discovered knowledge. This requires that the benefit-harm tradeoff for different stakeholder groups is understood and used for system design.
- **Presenting results** - As most often people are the intended users of the discovered knowledge, it is very important that adequate visualization techniques be designed to easy their task of interpreting the (intermediate and final) results of the process. User models may help designing more effective human/machine interactions, and guiding the development of new services.

Even though not all KDubiq problems are concerned with all of the above aspects, each one of them needs substantial improvements over existing techniques, when not a totally new re-thinking, offering thus occasion for advanced research.

## 2 Characteristic Features of KDubiq Research

A possible direction orienting KDubiq research is to be concerned with the matter of scale. Distribution of both data sources and processing devices may
become fine-grained, letting the solution of a problem emerge from the cooperation among many agents, each endowed with only small capabilities, but able to interact and communicate with others. This landscape has multiple effects: first of all, the fine-grained computational distribution implies the use of small devices, with possibly limited power. The consequence is a resource-aware approach, which can be better understood by analogy with the notion, well known in artificial intelligence, of bounded rationality vs. plain rationality, where the cost of the processing becomes an integral part of the process design itself.

A second important aspect is that, in the development of KDubiq systems, there are external parameters that influence the design choices and were not part of the rational under previous data mining paradigms: for instance, the KDubiq process must pay attention to the origin/location of the data, to the willingness of people to share data and to assess data reliability, to the specific nature of the computational means (hardware-aware algorithms), and to the user’s expectations (user modeling) and profile (personalized interestingness criteria). In addition, all of this must cope with privacy and security issues. In fact, the role of humans in the process is increased: not only privacy is more relevant, but also trust, especially in handling data coming from social networks, and in peer-to-peer interactions.

A third aspect of KDubiq is embedded processing. Instead of collecting (possibly distributed) data, elaborating them in one or few processing sites and routing the results to the users, the elaboration process can be transparently embedded in devices which answer to local needs in situ using local information, while communicating with others to face more general requests. Examples can be found in surveillance cameras, wearable health sensors, cell phones, and many others. Embedding is a side-effect of the fine-grained scale of the processing units distribution.

Finally, the types of data to be handled may show extensive variations over time and space, requiring analysis methods able to cope with possibly stringent real-time processing. Regarding variations, there is yet another aspect to consider, namely the possibility of erratic behaviour, due to a large spectrum of possible causes, for instance system crashes. If erratic behaviour may occur, algorithms are strongly influenced, because they must be robust in face of it.

In summary, KDubiq research is characterized by (1) the presence of a population of agents (small devices) provided with computing capabilities, (2) a rich semantics of the data, reflecting the behavior of the agents and their interactions, (3) an extreme distribution of such data, possibly with no chance of centralization, even partial, (4) a continuous flows of incoming data, (5) a complex scenario for the management of security and privacy, and (6) a deep involvement of people as producers of data and consumers of results.

\footnote{The notion of bounded rationality has been introduced by Herbert Simon, to revise the assumption, common in social and economic sciences, that humans could be approximated by rational reasoning entities. Bounded rationality accounts for the fact that perfectly rational decisions are often not feasible in practice due to the shortage of computational resources available.}
3 Differences between KDubiq and Distributed Data Mining

Data mining and KDubiq share the same goal: extracting interesting and useful knowledge from various types of data. As an analogy, let us mention the field of programming languages: object-oriented programming is part of the field, but one that has introduced a new and influential programming paradigm indeed. In the same way, KDubiq is part of data mining, especially close to distributed data mining (DDM), which, in some sense, KDubiq brings to a somewhat extreme realization.

At least four fundamental tracts play a role in differentiating KDubiq from DDM:

- **Scale** - In DDM, distribution is usually medium scale: a few data sources (databases, Web archive, sensors signals, ...) are exploited for the discovery process and/or a few computer facilities elaborate the data, leaving to a further step the collection of the results and its presentation to the user. In KDubiq, the number of explored data sources may be of orders of magnitude larger (even in the thousands or more). This may happen, for example, when a large part of the WWW has to be explored, or the input comes from very large sensor networks (electrical power net, grid systems, traffic sensors, weather parameter sensors, ...). This level of distribution requires approaches that are qualitatively, and not only quantitatively, different from existing ones.

- **Communication** - Given the fine-grained distribution of the computing units, it is not thinkable to set up a master computer that collects the individual results and put them together for the user. On the contrary, the final result shall emerge as a whole from the network of communications among the local computing units.

- **Resource-awareness** - The ubiquity of processing is usually coupled, as already mentioned, with a reduction of the computing power of the involved devices. Moreover, the discovery task may be requested to work in real-time. Again, performing a task under these operating conditions requires that novel approaches and methodologies be invented.

- **Integration** - Finally, the whole process of discovery cannot be cleanly split into its component phases (data collection, data cleaning, pre-processing, algorithm application, interpretation, visualization), but all these phases must interact, as they are intertwined into the computing agent network.

4 Why Are Current Data Mining Techniques Insufficient?

Given the description of KDubiq’s characteristic features, it appears quite obvious why current DDM methods are insufficient to meet the challenges set by the ubiquity of the discovery task.

Current methods are unable to cope with a finely distributed fleet of small agents and with their interactions. Today’s degree of computation distribution
does not consider communication among agents, and aggregation of results is always done in an almost centralized manner. In contrast, the extreme distribution of local computation calls for new ways of information exchange and aggregation of partial results: possibly never, during the discovery task, a global view of the whole process is assembled in some place; nevertheless, the user must receive coherent answers within reasonable time delays.

When the input data are generated by a sensor network, the difficulty of coordinating the computation among the agents increases, because the appropriate pieces of data must be routed towards the agent that needs them, without disrupting the semantics of the data and their spatio-temporal patterns. There are currently few algorithms able to do this, especially in real-time, in dynamic environments.

Today, data mining exploits the computational power of machines in an unbounded way, that is, limits are set by the computational time complexity of the algorithms, and not by the exhaustion of the resources. If small devices are used and must respond in an intelligent way, most of the current algorithms are no more applicable, because they will exhaust the computational resources very rapidly. Moreover, the new, resource-aware algorithms that are needed must be able to perform their task with a very limited horizon on the whole problem, relying on little local information but on a lot of communication. As a consequence, current learning and discovery algorithms are too sophisticated for the new generation computing units envisaged by KDubiq.

Even though privacy-aware methods already exist, they are inadequate in a finely distributed environment, where no centralized control for accessing the data exists. Then, more stringent privacy and security constraints arise, in order to regulate access rights of agents, to protect their anonymity and to manage trust and confidence of their interactions.

Finally, the way people interact with KDubiq systems are to be reconsidered. For instance, should cell phones or palm computers become the privileged tools for people to do data mining or enter their own data, then new interfaces are to be designed to cope with the limitation of the physical interaction space.

5 A Cognitive and People-Centric Perspective

From a people-centric perspective, KDubiq means rethinking who is the “user” of a KDubiq system and how all the advances and dimensions of ubiquity affect users of such systems (and other stakeholders). This is not only a matter of interfaces but a perspective that spans the whole design process, from the conception of a system through to its deployment.

KDubiq must work “in real time”, “in real environments”, and therefore “under conditions of real resource constraints” (such as bounded rationality). Ubiquity enlarges the definition of “users” of a system to a wider range of people encompassing users, stakeholders and communities. Consequently, methods to elicit possibly conflicting interests of the target population are required. Conflicts arise mainly due to data coming from different users or stakeholders,
obtained from different data/knowledge sources. As conflicts should not be avoided, methods for dealing with them have also to be developed.

Data collection from ubiquitous users must cope with two major problems: challenges of obtaining data and challenges of their representativeness. Users differ in their ability to provide information as well as in their willingness to share it. Differences in privacy issues and willingness to share information ask for a detailed examination of the extent to which data gathering would constitute an intrusion into the private space. Furthermore, a user’s background affects the way opinions are expressed. This needs to be taken into account either through a culturally adapted conception of data gathering tools or through appropriate data processing that considers these differences.

Ubiquity of people leads to heterogeneous data sets due to different contexts. User-centered knowledge discovery hence requires data processing that takes background knowledge about the users and their context into account. Ontologies and ‘folksonomies’ have been analyzed as a possible solution that needs to be extended to tackle the above mentioned context problem. Given the increasing amount of multilingual data sets, knowledge discovery should also take into consideration research results regarding multilingual information retrieval tools.

6 Example of Typical KDubiq Applications

Even though the spectrum of potential application fields is really very large, we have selected here two of them, that can be used to exemplify “typical” KDubiq applications: one is related to activity discovery, and the second one to providing a web service for music file sharing.

6.1 Activity Recognition – Inferring Transportation Routines from GPS-Data

The widespread use of GPS devices has led to an explosive interest in spatial data. Classical applications are car navigation and location tracking. Intensive activity, notably in the ubiquitous computing community, is underway to explore additional application scenarios, e.g., in assistive technologies or in building models of the mobile behavior of citizens, useful for various areas, including social research, planning purposes, and market research.

We discuss an application from assistive technologies, analyze its strength and shortcomings and identify research challenges from a KDubiq perspective.

The OpportunityKnocks prototype [1] consists of a mobile device equipped with GPS and connected to a server. An inference module running on the server is able to learn a person’s transportation routines from the GPS data collected. It is able to give advice to persons, e.g., which route to take or where to get off a bus, and it can warn the user in case he commits errors, e.g., takes the wrong bus line. The purpose of the system is to assist cognitively impaired persons in finding their way through city traffic.
This application meets the main criteria for ubiquitous systems: the device is an object moving in space and time, it is operating in an unstable and unknown environment; it has computing power, and has a local view of its environment only; it reacts in real-time and it is equipped with GPS-sensors and exchanges information with other objects (e.g., satellites, the server). Since both the environment and the behavioral patterns are not known in advance, it is impossible to solve this task without the system being able to learn from a user’s past behavior. Thus, machine learning algorithms are used to infer likely routes, activities, transportation destinations and deviations from a normal route.

The basic infrastructure of OpportunityKnocks is a client/server setup. A GPS-Device is connected to a mobile phone. The mobile phone can connect to a server via GPRS and transmit the GPS signals. Wireless communication is encrypted. The server analyses the data, utilizing additional information about the street network or bus schedules from the Internet. Using this information the person is located and the system makes inferences about his current behavior and gives suggestions what to do next. This information is sent back to the client and communicated to the user with the help of an audio/visual interface.

Although innovative, the architecture of this prototype will face a number of practical problems:

- When there is no phone signal, communication with the server is impossible, and the person may get lost.
- When there is no reliable GPS signal, e.g., in a train, indoors, in an underground station, guidance is impossible. GPS for pedestrians in an urban environment is unreliable.
- Communicating via a radio network with a server consumes a lot of battery power (the system works only 4 hrs under continuous operation).
- Continuously tracking of a person and centrally collecting the data creates strong privacy threats.

The KDubiq paradigm asks for distributed, intelligent and communicating devices integrating data from various sources. A “KDubiq Upgrade” would result in a much more satisfactory design for the prototype. The upgrade is guided by the following imperatives:

1. Move the machine learning to the mobile device.
2. Add more sensors and learn from them.
3. Let the device learn from other devices.
4. Respect privacy.

With current state-of-the-art mobile phone technology, all technical infrastructures for upgrading are in place. What is missing currently are the proper knowledge discovery tools.

If the major part of the learning is done on the mobile device – especially that part that refers to localization on the street map – there is no need for constant server communication, and assistance becomes more reliable.
Moving the machine learning to the device introduces new tasks and additional criteria in assessing the value of a machine learning algorithm that existing algorithms are not designed to meet. Addressing these criteria results in new algorithms that often would not make sense in a non-KDubiq environment.

To give an example, due to high communication costs, OpportunityKnocks is continuously operating for 4 hrs before the battery is exhausted – not enough for a real-world environment. Since network communication is very expensive and thus a bottleneck (see Chapter 3), power consumption appears as an additional criterion for the design of a learning algorithm. In a traditional environment, this criterion is irrelevant.

A solution that calculates everything on the mobile device and does not communicate at all will not do. OpportunityKnocks uses Dynamic Bayesian Networks. This is heavy machinery, and doing all computations on the device would be slow; moreover, the computation would require a lot of resources, so that, again, the battery would quickly run low. Instead, a solution is needed that locally computes and pre-aggregates results and communicates only few data via the radio network. Splitting the computation into an energy and computationally efficient online part yielding highly compressed models, transmitting only this compressed information and performing computationally intensive parts on the server is a better solution. Pioneering work in this direction – in the context of vehicle monitoring – has been done by Kargupta [2]. Kargupta shows that for the online part new algorithms are necessary that trade accuracy against efficiency. The specific trade off is dictated by the application context, and the choice made in the vehicle monitoring application would be hard to motivate in an offline-context (or even for the current application).

Once the main part of learning is done on the device, opportunities occur for giving the system access to more sensory and background information, including information generated by other similar devices.

If the system is equipped with a gyroscope and/or accelerometer and has a local map, short-term navigation is possible in the absence of a GPS signal. Using multi-sensor input for online-analysis is a common approach for car navigation already, combining the signals e.g. using Kalman filtering. For mainstream machine learning and data mining, it challenges some basic assumptions, since data arrives not as a batch but in a streaming setting (cf. Chapter 4 for details).

If the device is additionally able to communicate with other devices – e.g. using WiFi or Bluetooth – one use of WiFi could be to complement GPS for indoor localization. Even more interesting would be communication with a collection of other similar systems, e.g. to infer the transportation mode by taking into account the local knowledge of these other systems.

The distributed, collaborative nature of KDubiq is thus a source for new algorithms. (cf. Section 6.2 in this chapter). In contrast to almost all approaches to distributed mining, the algorithms can exploit other devices’ knowledge without aiming for a global model. They yield new and different algorithms that would simply make no sense in a non-distributed environment – and therefore did not exist before.
The original scenario creates severe privacy threats. Once the mining goes to the machine, the privacy sensitive data can remain with the person, and the privacy risks of central monitoring are removed.

Additional risks occur however due to the fact that devices communicate with other devices. Thus privacy constraints are a further third source for new algorithms. The task is to distribute the computing in such a way that no computing device not under the control of the user has access to information that infringes privacy. Design options for privacy preserving data mining are discussed in Chapter 6.

To sum up, existing algorithms cannot solve the problem because they have not been designed for this task, and do not take requirements into account that emerge in an ubiquitous context.

A further kind of challenges derives from the ubiquitous computing roots of KDubiq.

It is about careful selection, adaptation and integration of partial solutions, both algorithmic and technical, resulting in a delicate mixture of hardware, software and algorithm design issues. Some of the associated challenges may be best considered as system integration or engineering challenges. The challenge is that the sum is qualitatively more than its parts – and exponentially more complex to implement. Often underrated by machine learning purists, for the new field of KDubiq these challenges are at the core. Thus, the interdisciplinary nature of KDubiq asks for a broader view what constitutes a research challenge and to include activities as they are commonly carried out in other research fields such as ubiquitous computing or robotics².

Overall, the challenges for the current application consist in building learning algorithms for distributed, multi-device, multi-sensor environments. While partial suggestions exist on how to implement privacy-preserving, distributed, collaborative algorithms, respectively, there is hardly any existing work that properly addresses all the dimensions at the same time in an integrated manner. Yet as long as one of these dimensions is left unaddressed, the ubiquitous knowledge discovery prototype will not be fully operational in a real-world environment. We need both new algorithms – including analysis and proof about their complexity and accuracy – and an engineering approach for integrating the various partial solutions – algorithms, software and hardware – in a working prototype. Building such a prototype would be a challenging topic for a research project.

6.2 Ubiquitous Intelligent Media Organization

With the advent of Web 2.0, collaborative structuring of large collections of multi-media data based on meta-data and media features has become a significant task. While most applications are based on a central server, peer-to-peer solutions start to appear.

² To prepare the community for this shift in the skill set of a machine learner, KDubiq has organized two summer schools (www.kdubiq.org).
As an example we discuss Nemoz (NEtworked Media Organizer) \cite{2,10,16}, which is a Web 2.0-inspired collaborative platform for playing music, browsing, searching and sharing music collections. It works in a loosely-coupled distributed scenario, using P2P technology. Nemoz combines Web 2.0-style tagging, with automatic audio classification using machine learning techniques.

Nemoz is motivated by the observation that a globally correct classification for audio files does not exist, since each user has its own way of structuring the files, reflecting his own preferences and needs. Still, a user can exploit labels provided by other peers as features for his own classification: the fact that Mary, who structures her collection along mood, classifies a song as ‘melancholic’ might indicate to Bob, who classifies along genre, that it is not a techno song. To support this, Nemoz nodes are able to exchange information about their individual classifications. These added labels are used in a predictive machine learning task. Thus the application is characterized by evolving collections of large amounts of data, scattered across different computing devices that maintain a local view of a collection, exchanging information with other nodes.

Main features of the application are:

1. The application differs from the preceding one in that the (geo)spatio-temporal position of the computing devices does not play an important role; the devices and the media file collections they contain are stored somewhere on some node in the network. Yet it is a defining characteristic of the application that two collections $C_i$ and $C_j$ are stored at different places, and it is important whether or not two collections are connected via a neighborhood graph.
2. The computing devices might be connected to a network only temporarily. The collections are evolving dynamically; items are added and deleted, and also classifications can change.
3. The computing nodes have sufficient local computing power to carry out complex tasks, and as long as mobile devices are not included, resource constraints are low.
4. It is a crucial aspect of this application that the nodes maintain a local view, incorporating information from other nodes.
5. The device does do not take autonomous action or actively monitors the state of the collection until some event occurs. Still response-time is an important issue because the application is designed for interactivity.
6. Finally, the distributed nature of the problem is a defining characteristic of the application.

We are not aware of other solutions that are able to automatically learn from other user’s classifications while maintaining a local or subjective point of view. This application is a representative of a innovative subclass of applications in a Web 2.0 environment. Whereas most Web 2.0 tagging applications use a central server where all media data and tags are consolidated, the current application is fully distributed.

In many distributed data mining applications, originally centralized data are distributed for improving the efficiency of the analysis.
The current application is different because, firstly, the data are inherently distributed, and secondly, there is no intention to come up with a global model. Communication across peers is used for improving a local classification by feature harvesting. Thus Nemoz introduces a new class of learning problems: the collaborative representation problem and localized alternative cluster ensembles for collaborative structuring (LACE).

From the perspective of KDubiq this is important, since in a non-distributed environment, these new learning scenarios would be very hard to motivate. We see in this case study that ubiquitous knowledge discovery prompts the invention of new learning scenarios – not only problems are solved that could not be solved with existing methods; rather, new classes of learning problems are invented.

It is also interesting to note that the extension of the activity recognition scenario could raise a need for similar mechanisms as utilized in Nemoz (see section 6.1). This potential transfer of learning scenarios from two seemingly very unrelated areas – mobile assistive technology and music mining – is made possible by analyzing the applications in a common framework.

Several future extensions of Nemoz seem possible:

- Apart from the need for fast response-times because of interactivity requirements, resource constraints for the mining algorithms are not discussed – probably because experiments are done on notebooks or workstations. But once devices such as mobile phones are included (and they are mentioned as potential devices), these considerations will become important. The classification algorithms used are quite efficient (e.g., decision trees) and could be run without difficulties on a mobile device.
- User can mark their tag structures as private or public. Private folders are not visible to other users while browsing. Private folders are, however, used for data mining. It seems by sending a request, an attacker would be able to infer which music files are probably stored on another computer. This reveals information about an user. It would be interesting to find out whether the sharing can be done in a privacy-preserving manner.
- In extension of characteristic (5) above, one can easily imagine e.g. a publish-subscribe scenario where a node broadcasts a message to interested parties if a new item appears in a certain group of his collection. In this additional scenario, temporal, and spatial aspects are very important, and also acting in real-time, e.g. if it is important to be the first to classify a hitherto unknown song.

Thus we can see can see that a further development of Nemoz, addressing more fully the different dimensions of KDubiq (small devices, privacy, a real-time scenario) leads to interesting new research questions.

7 Technical Challenges to Be Overcome by Future Research

Challenges to be overcome by future research involve all components of the data mining process: data, algorithms, and interfaces. They can be roughly
summarized by saying that large problems must be solved with many small computational elements.

First of all, mobility must be tamed; as agents move, operate and interact with each other according to some predefined purposes, their movements are to be traced, and the recorded traces are augmented with richer background information: occurrence of events, such as load and delivery of goods or passengers, rendez-vous between mobile agents, linkage of events with specific geographic information concerning places and their role within a logistic process, and so on.

Moreover, agent mobility must be coupled with data mobility. The next generation of sensor-networked infrastructures for ubiquitous computing shall support enhanced capabilities of sensing the mobile objects, thus allowing the collection of mobility data of higher precision and richness, covering not only positioning of mobile devices with reduced error and uncertainty, but also many further properties of mobile behaviours, recorded thanks to the enhanced capabilities of interaction with the surrounding environment.

The change in time and space of data distributions makes useless any algorithm based on static description of i.i.d. (independent and identically distributed) data. Moreover, this very change does not any more allow relying on a good asymptotic behaviour, as transitories become the norm.

As the task of knowledge discovery will result from the collective work of great number of agents and interactions, methods and approaches from the complex systems theory, ant colonies, or swarm intelligence may be mutated as basic tools. All these methodologies must be investigated in depth, in order to understand if, how, and why they might help.

Another crucial issue to be resolved is the limitation in processing power, communication capability and data storage of small devices.

As the use of such devices is mandatory for the practical success of many applications, this issue appears to be the main bottleneck of future developments.

Finally, new interfaces between people and systems must be designed, based on multi-modality, ergonomic considerations, and cognitive models.

8 Overview of the Book

Part I of the book aims to provide a conceptual foundation for the new field of ubiquitous knowledge discovery, discussing the state of the art, highlighting challenges and problems, and proposing future directions. Although at some points technical examples are given for illustration, the aim of this chapter is non-technical.

While Part I is divided into individually authored chapters, it should be seen as a collaborative effort by the working groups of the KDubiq coordination action. Hence, the chapters should be seen as an integrated whole. Part I of the book is structured as follows.

\[3\] In some cases, project partners provided input for some sections, but not for the whole chapter. Where this is the case, it is stated in the footnotes of individual sections.