Noname manuscript No.
(will be inserted by the editor)

Living Lab Bamberg

An infrastructure to explore research challenges in the wild

Aboubakr Benabbas · Golnaz Elmamooz · Brent Lagesse · Daniela Nicklas · Ute Schmid

Received: date / Accepted: date

Abstract With ever emerging technological resources that can interact with the physical environment, we have an opportunity to drastically improve the usage of resources and improve the quality of life of communities of people. There are numerous problems that must be solved first. Research areas such as security, privacy, data quality, and data modeling must be addressed. In order to move forward to a better world, we have established the living lab Bamberg where we will address these research problems and provide open data and APIs to other researchers to enable collaboration and extensive testing of smart city research and applications. As we continue to use the living lab Bamberg to improve state of the art research in smart cities, we believe we will see smart city technology adopted more commonly and drastically improve the lives of people living in those cities.

1 Introduction

For future cities, people envision a smarter usage of resources like space, energy, or water, to increase the quality of life in growing communities. Sensor-based information plays a vital role in many applications in that domain. Just to name a few, *smart city decision support systems* can show life and aggregated sensor data on traffic, air quality, or noise, to improve future city planning. *Tourist recommendation systems* could measure typical path of certain interest groups and help other visitors to plan their trips. For *Event Organization* of, e.g., street festivals, it is important to plan and control the occupation of public spaces, not only for catering and toilet logistics, but also for safety and security. Finally, *Citizen Science* projects motivate people to contribute in crowd sensing campaigns, exploiting the wide distribution of smart phones in these days.

However, before such application can be really implemented in the wild, several research challenges have to been solved. In this paper, we highlight some of these challenges and present the architecture and future plans of the Living Lab Bamberg, a research infrastructure for sensor-based smart city applications [10].

A. Benabbas, G. Elmamooz, D. Nicklas and U. Schmid
University of Bamberg, An der Weberei 5
E-mail: [firstname.lastname]@uni-bamberg.de

B. Lagesse
University of Washington Bothell
E-mail: lagesse@uw.edu

2 Research Challenges

2.1 Privacy and Security

User privacy is critical to the legal and cultural acceptance of a smart city. Organizations such as corporations and governments that deploy smart city infrastructure and applications should not be able to identify individual users and use their information without their permission. Likewise, other users of the system should not be able to identify other users and use their personal information without their permission.

Online privacy focuses on securing information about users as they are participating in the system. These privacy enhancing technologies need to focus on preventing users and organizations from using information that they know from violating the privacy of users in real-time.

Offline privacy focuses on securing information about users that is collected from the system and archived for future use. We want to enable other resear-

chers to utilize the results of the Living Lab, but in order to publish datasets for study, we need to ensure that privacy enhancing techniques have been applied to all dimensions of the data set.

As a result of these challenges, we ask the question:

How can data describing events in the physical world be provided in a useful manner while still protecting the people whose lives the data reflects?

Regarding security, there are two primary challenges that need to be addressed in the living lab. The first is the security of the data itself and how malicious users could manipulate the data to damage applications and analyses. The second research challenge is the creation of usable security mechanisms that are unobtrusive to the users.

Data Security focuses on preventing malicious users from strategically inserting data into the system in such a way that it would disrupt or adversely influence applications that use the data, particularly those that rely on machine learning[4]. For example, if an attacker deploys malware that infects a large number of smart phones that are participating in the system, an attacker could strategically manipulate an event analysis application to advise law enforcement officers to move away from an area in which they are actually needed. Some of these attacks are designed to target general consumers of data [6] and others are designed to target specific algorithms, in particular, machine learning algorithms, that use the data [7].

Usable Security reduces the workload placed on humans using systems to ensure that security mechanisms work correctly while maximizing the utility of the system despite the presence of attackers. Security mechanisms are useless if the users ignore them or they cripple the system to the point that users do not use it. If a security mechanism involves a human in the loop, then it must do so efficiently.

Smart environments undergo a variety of changes in context in unpredictable ways. The learning and analysis algorithms that consume information sensed in these environments not only have the potential to suffer from concept drift naturally, but may suffer as a result of attacks against data security. Existing systems struggle to understand whether concept drift is a result of an attack or a true change in the environment. As a result of this challenge, we ask the question:

In a dynamic environment, how can a system efficiently determine if the environment it operates in is changing and requires retraining or if an attacker is strategically manipulating the data to make it appear to be changing without burdening the user?

2.2 Sensor Data Quality

In many smart city applications, sensor data plays a key role. Some applications have to make decisions even in real time which exacerbates the negative effects of low data quality. For example, in environment monitoring, sensors deployed along the river traversing the city deliver readings on the water level. If some of these sensors fail, missing values about the water level could lead to a late detection of a flooding. In the management of street festivals the rate of street occupancy is very important both for business and public safety. In such events some WiFi trackers could be deployed on the streets to monitor the street occupancy based on smart phones signals. A false counting of people can deliver an incorrect analysis to the organizers, resulting in dangerous or inefficient emergency routes.

The quality of data can be defined through a set of dimensions. Prior work has defined a certain number of these dimensions with slight differences from one definition to the other. For example Batini et al. [1] provide the data quality dimensions and their respective definitions as follows:

- *Accuracy* is the closeness between two values v and v', where v tries to represent a real world phenomenon and v' is considered as the real representation.
- *Completeness* is given by the breadth, width and scope of data for the given task. Completeness answers this question: how sufficient is the information provided by the data? Completeness can be described by completeness of schema, completeness of columns, and completeness of population.
- *Currency* is the frequency at which data is updated. The currency is reported as high if the data update brings a state update. The currency is however low if data updates do not reflect the actual state of things e.g: due to network latencies.
- *Timeliness* describes the currency of data for a specific task. Depending on the tasks nature the currency of data varies, where current data can be deemed unusable for a certain usage.
- *Volatility* defines the period of time that represents a validity interval of data. Some data is stable and does not change such as birth dates, while other data have a varying volatility (like stock quotes, arrival times of trains.)

From the description of the use cases of sensor data and the provided dimensions that describe the quality of data we can derive the following research question:

How can we monitor controlled data quality from a network of sensors and provide it as a service to all stakeholders in the Living Lab

With the proliferation of sensor systems in smart cities, we can exploit redundancy to control data quality. In a given place, many different sensors could be used to observe a phenomenon or even each other. The system should however be aware of this redundancy. To describe all the sensors, their capabilities, deployments, and their combinations, the Semantic Sensor Network (SSN) ontology can be used [2]. The SSN ontology can describe sensors, and their sensing methods to make observations of their environment. The ontology can specify the survival ranges of sensors and the sensors performance within those ranges. The ontology also offers the possibility to describe the field of deployment of sensors where the duration and purpose of deployment is indicated. The SSN ontology includes multiple quality dimensions such as accuracy, latency, and frequency. To implement a data quality service Kuka and Nicklas [5] introduced a method, where the SSN ontology is used to describe sensors and quality properties of their observations. As a result, we ask the following research question:

How can a semantic description of sensor systems and their installation be used to automatically enrich information derived from sensor data with quality assessment?

2.3 Model Evolution

To make use of the large amount of data collected with different sensors — possibly along different time scales, with different degrees of precision and on different scales of measurement (such as nominal yes/no information and metric data) — it is necessary to provide a model. Such a model allows data to transform into information. For example, camera-based head counts together with an oxygen sensor provide data which might be interpreted as the fact that many people are currently at some defined location. This information then could be made available to a human decision maker. Such models could be predefined; however, for complex dynamic environments, a purely knowledge-based approach is not feasible since typically such models are learned. For example, classifying the amount of people at some location as many/few can be realized with a classifier obtained by a supervised machine learning approach. Taking into account temporal or spatial sequences of data, such a model can be used for predictions such as that soon there will be many people at some location. Furthermore, unsupervised methods can be used to detect patterns in such complex data sets such as a specific location only gets crowded in the evening. Much research has investigated unsupervised and supervised machine learning approaches to extract information from spatiotemporal data and apply the information to predict the next location of moving objects [9] or to suggest a desirable trajectory to the tourists [11].

In a setting where data from different sources is collected at small intervals over some time, it is necessary to have a policy to decide which data will be used to train a model. That is, feature selection becomes a crucial factor for model quality. Furthermore, scoring of new data, e.g., assignment of a class or to a cluster or prediction of an outcome, might not only rely on a single trained model but on an ensemble of models. There are many plausible scenarios for ensemble learning, for example, using models learned from different sets of sensor data or over different time spans. Meta-learning strategies need to be defined to obtain the most probable prediction from an ensemble of models.

In a dynamic environment like a smart city, it can be assumed that it is not enough to learn a model once. Such a static model might become obsolete due to gradual or abrupt changes in the environment or the requirements. For example, if security measures such as emergency escape routes were improved for a certain location, the interpretation of a certain number of people as 'many' might be shifted to a larger amount. That is, it is necessary to apply incremental/lifelong learning approaches and to deal with concept drifts [3, 8]. In this context, suitable policies of forgetting information become relevant.

In summary, machine learning is crucial for evolving models which can be applied to incoming streams of data. In the context of practical applications, the accuracy of models when scoring new data needs to be high enough for safe recommendations. In the context of the living lab, there are many challenging problems for model evolution. As a result, we ask the following research question:

How do we perform feature selection and selection of data as input for learning a model, dealing with learning multiple models in parallel or as an evolving sequence, identifying suitable strategies for meta-learning, and dealing with concept drifts?

3 Data Management

To answer the aforementioned research questions, data plays a vital role. Of course, one can develop algorithms or proof-of-concept implementation with simulated data, and many researchers already proposed promising work. For example, a tremendous amount of research within the mobile ad hoc network field was

evaluated using movement patterns produced by the network simulator ns-2[1]. However, before these approaches can be applied to real applications, still some major research has to been done.

3.1 Wild Data

As already introduced, we need data from the wild, i.e., produced by real people, who ideally are not aware of the fact that the data will be used in some experiments. The provisioning of such data is the main goal of the living lab. In addition from data from user devices, we need to capture infrastructure data, like network traffic, or data from pre-installed sensors. Since the research should apply not to a specific installation or technology, we need both heterogeneity (i.e., sensor systems from different vendors) and redundancy (i.e., measuring the same phenomenon with different sensing methods.

To gather data from user devices, users will install the app on their phones; this can be done over the Internet, by providing QR codes scanned from a flyer or a booth at an event, or by direct Bluetooth-push messages within areas where the application will be used. A key challenge here is OS coverage; supporting different operating systems (like Android and iOS) and different versions of these operating systems causes a high software development overhead.

3.2 Control Data

To properly evaluate the aforementioned research challenges, we need additional ground truth. Such control data can be produced within student projects. The students will use the same applications (or be measured by the same infrastructure sensors), but will follow a predefined protocol. By carefully documenting these protocols, we can produce data sets of ground truth data within the overall wild data collections.

In addition, we can also include simulated malware (both by real user systems and by student systems). The user would just turn on the "malware" button and it would manipulate the data to simulate an attack on the system. By doing this we can study the influence of certain attacks on the overall system or specific aspects of it.

[1] http://www.isi.edu/nsnam/ns/

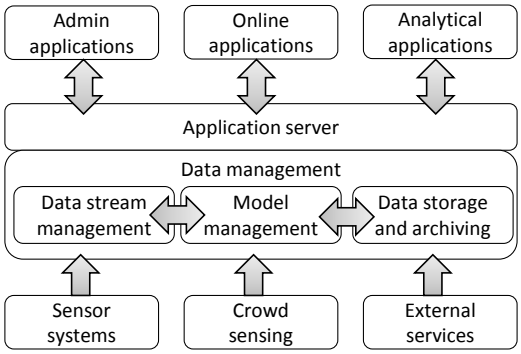

Fig. 1 Architecture of the Living Lab Bamberg

3.3 Open Data

Finally, data sets from the Living Lab should be made public as open data so that other researchers can use it, too.

In the context of the Living Lab, the right choice of spatial anonymization techniques depends on the number of participants in a data set. If the number of individuals is very high, it is easier to achieve a suitable amount of anonymization. However, if we have only a low number of trajectories, we need perturbation, or decide not to publish a data set at all. We plan to investigate on dynamic choices of anonymization techniques in the context of the Living lab.

In addition, we have to consider the time of anonymization within our data flow. This depends on the user's preferences. If the data needs to be anonymized before storage, it might require us to first collect a sufficient number of data sets. If (by the user's preferences) we are allowed to store the data without spatial anonymization, we can apply the spatial cloaking on the stored and integrated data sets.

4 Living Lab Bamberg

The goal of the Living Lab Bamberg is to provide an open infrastructure for research on sensor-based applications. It can be used both by academia and industry to test new technology, develop and evaluate algorithms, collect data sets, and publish them. Since the sensors and applications might collect sensitive information like location and trajectories of citizens, its main requirement is to be privacy-preserving, i.e., not to expose any personal information where individuals could be identified.

4.1 Architecture

The architecture of the Living Lab follows a typical three layer structure. Central is the data management layer that consists of a distributed data stream management system responsible for all online processing of data and a data storage and archiving component. The model management component integrates machine learning algorithms into the architecture: it is responsible for feeding new training data into deployed algorithms, storing and updating the learned models (e.g., decision trees or rule sets), and for deploying such models in the data stream management system to be used for scoring (e.g., to classify incoming data or to apply a rule for recommendation).

The application server supports three types of applications: admin applications for managing the hardware systems and software systems of the Living Lab, online applications that provide some ongoing service to (often mobile) users, and analytical applications like decision support systems that access the history of data.

Finally, data from several systems can be ingested by the lower layer: sensor systems deliver data over live or batched APIs, or can be pulled by the data management layer. Crowd sensing applications deliver data collected by citizens. Finally, further data might come in from other external services like traffic management system from public services.

4.2 Sensor Systems

Within the living lab, we support both stationary and mobile sensor systems.

Stationary sensor systems can be installed in several places all over the city. Since the University of Bamberg has a number of university buildings within the city center, they can be used for easy access to the university's infrastructure. In addition, we can get support by the public service company for further installation points. In previous Living Lab experiments, we gained some experience with people counting cameras and so-called Flowtracker®, sensors that detect devices based on WiFi and Bluetooth® Low Energy signals and deliver MAC addresses and signal strength. For the future, we plan further installation of weather stations, traffic counters, and any sensor system that might be needed to support an application that should be evaluated in the living lab.

The first *mobile sensor systems* we plan to install are sensor platforms on public buses. They will be equipped with several environmental sensors to collect temperature, humidity, CO2 level and noise level. In addition, smart phone apps will be deployed that use sensors native to the device, like acceleration, noise level, and other phenomena that can be sensed by the phone. For mobile sensing, the location of the measurement is crucial for any analysis. While the bus-mounted sensor systems can use GPS to localize the measurements, we cannot always rely on that on the smart phones; GPS might be turned off by the user to save energy and GPS does not work indoors. Thus, we installed Bluetooth® beacons in several buildings to support indoor localization.

5 Outlook

As in 2016, the Living Lab Bamberg is still under development. We conducted three field tests for the technology: we captured human mobility in two different street festivals, and we installed human mobility sensors and environmental sensors on a science exhibition[2] for five month. The experiences and also the hardware from these field tests will be included in the Living Lab Bamberg, for which we have the following goals:

Technology Transfer: We plan to make the systems developed in our research available so that other researchers can rapidly deploy similar systems. It is critical that the results obtained from our living lab be verified in a variety of other environments, so that the community is able to learn which results are strongly tied to specific types of environments and which results are more universal.

Research Data Sets: Within the installations of the Living Lab, we create sensor data sets including ground truth information that can be used by researchers to develop, test, and evaluate sensor-based applications and supporting algorithms and methods.

Open Research Platform: We will create an open API to allow external researchers to connect to the living lab remotely to perform experiments. Prior to doing so, the privacy concerns addressed in Section 2.1 must provide results that safely enable an open API.

We hope that this work indeed will be a *living* lab to foster scientific advance and international reseach collaboration.

[2] http://www.ms-wissenschaft.de

References

1. Batini C, Scannapieco M (2006) Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, Inc., Secaucus, NJ, USA
2. Compton M, Barnaghi P, Bermudez L, Garcia-Castro R, Corcho O, Cox S, Graybeal J, Hauswirth M, Henson C, Herzog A, Huang V, Janowicz K, Kelsey WD, Phuoc DL, Lefort L, Leggieri M, Neuhaus H, Nikolov A, Page K, Passant A, Sheth A, Taylor K (2012) The SSN ontology of the W3C semantic sensor network incubator group. Web Semantics: Science, Services and Agents on the World Wide Web 17:25 – 32, DOI http://dx.doi.org/10.1016/j.websem.2012.05.003
3. Elwell R, Polikar R (2011) Incremental Learning of Concept Drift in Nonstationary Environments. IEEE Transactions on Neural Networks 22(10):1517–1531, DOI 10.1109/TNN.2011.2160459
4. Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD (2011) Adversarial Machine Learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, ACM, New York, NY, USA, AISec '11, pp 43–58, DOI 10.1145/2046684.2046692, 00107
5. Kuka C, Nicklas D (2014) Enriching sensor data processing with quality semantics. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on, pp 437–442, DOI 10.1109/PerComW.2014.6815246
6. Lagesse B, Kumar M, Wright M (2008) AREX: An Adaptive System for Secure Resource Access in Mobile P2p Systems. In: 2008 Eighth International Conference on Peer-to-Peer Computing, pp 43–52, DOI 10.1109/P2P.2008.46, 00019
7. Lagesse B, Burkard C, Perez J (2016) Securing pervasive systems against adversarial machine learning. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp 1–4, DOI 10.1109/PERCOMW.2016.7457061, 00000
8. Martínez-Plumed F, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ (2015) Knowledge acquisition with forgetting: an incremental and developmental setting. Adaptive Behavior p 1059712315608675, DOI 10.1177/1059712315608675
9. Pelekis N, Theodoridis Y (2014) Mobility Data Mining and Knowledge Discovery. In: Mobility Data Management and Exploration, Springer New York, pp 143–167
10. Steuer S, Benabbas A, Kasrin N, Nicklas D (2016) Challenges and Design Goals for an Architecture of a Privacy-preserving Smart City Lab. Datenbank-Spektrum pp 1–10, DOI 10.1007/s13222-016-0223-8, 00000
11. Zheng Y (2015) Trajectory Data Mining: An Overview. ACM Trans Intell Syst Technol 6(3):29:1–29:41, DOI 10.1145/2743025