Secure Sharing of Geospatial Wildlife Data

Full Paper

Remo Manuel Frey
MIT Media Lab
75 Amherst Street
Cambridge, MA 02139, USA
rfrey@mit.edu

Thomas Hardjono
MIT Connection Science & Engineering
Cambridge, MA 02139, USA
hardjono@mit.edu

Christian Smith
MIT Sociotechnical Systems Research Center (SSRC)
Cambridge, MA 02139, USA
csmth@mit.edu

Keeley Erhardt
MIT Electrical Engineering & Computer Science
Cambridge, MA 02139, USA
kerhardt@mit.edu

Alex 'Sandy' Pentland
MIT Media Lab
75 Amherst Street
Cambridge, MA 02139, USA
pentland@mit.edu

ABSTRACT

Modern tracking technologies enables new ways for data mining in the wild. It allows wildlife monitoring centers to permanently collect geospatial data in a non-intrusive manner in real-time and at low cost. Unfortunately, wildlife data is exposed to crime and there is already a first reported case of 'cyber-poaching'. Based on stolen geospatial data, poachers can easily track and kill animals. As a result, cautious monitoring centers limited data access for research and public use. This means that the data cannot fully exploit its potential. We propose a novel solution to overcome the security problem. It allows monitoring centers to securely answer questions from the research community and to provide aggregated data to the public while the raw data is protected against unauthorized third parties. This data service can also be monetized. Several new applications are conceivable, such as a mobile app for preventing conflicts between human and wildlife or for engaging people in wildlife donation. Besides presenting the solution and potential use cases, the intention of present article is to start a discussion about the need for data protection and privacy in the animal world.

CCS CONCEPTS

• **Security and privacy → Privacy-preserving protocols**
• *Information systems → Location based services*
• *Social and professional topics → Computer crime*
• *Applied computing → Data centers*

GeoRich'17, May 14, 2017, Chicago, IL, USA

ACM 978-1-4503-5047-1/17/05…$15.00
http://dx.doi.org/10.1145/3080546.3080550

KEYWORDS

Data sharing, geospatial, GPS, wildlife, animal, cyber-poaching, hunting, species protection, privacy, security, crime, blockchain

1 INTRODUCTION

Data mining has a long tradition in wildlife research. Before the era of computers, study data was typically collected by observations by humans in the wild. It was the time of the great explorers and adventurers who held their fortuitous sightings in diaries and sketches. They interviewed residents and studied found dead animals. Then, in the last century, researchers started to use more sophisticated methods like game warden surveys, photo traps, radio telemetry, and capturing-recapturing with the aim to receive a more accurate image of the investigated animal or geographical area of interest. These approaches significantly improved the data quality and the data quantity as well. But, all of these data collecting methods still need manual effort and thus, are expensive and difficult to scale. Moreover, the data quality is often dependent on the expertise of the involved people (e.g. finding the right spot for a photo trap).

Modern tracking technologies enables new ways for extensive data collection of animals in the wild. Hardware innovations in the past decades (miniaturization, battery and transmission technology, satellite navigation) lead to a massive proliferation of digital tracking devices like GPS collars and implantable data sensors. The behavior of animals is observable without disturbance. This is, on the one hand, important for the animal itself and, on the other hand, for the research results, which should not be falsified. Moreover, the instruments are able to collect measurements 24 hours per day, which enables a seamless observation, even during the night or in an inhospitable environment like the in ocean or the in arctic. Compared to previous methods, these instruments are handier, cheaper, more reliable and more accurate. In combination with today's software, it is possible to automate the entire data process chain (capture, store, evaluate, visualize). In doing so, a large amount of captured wildlife data is manageable.

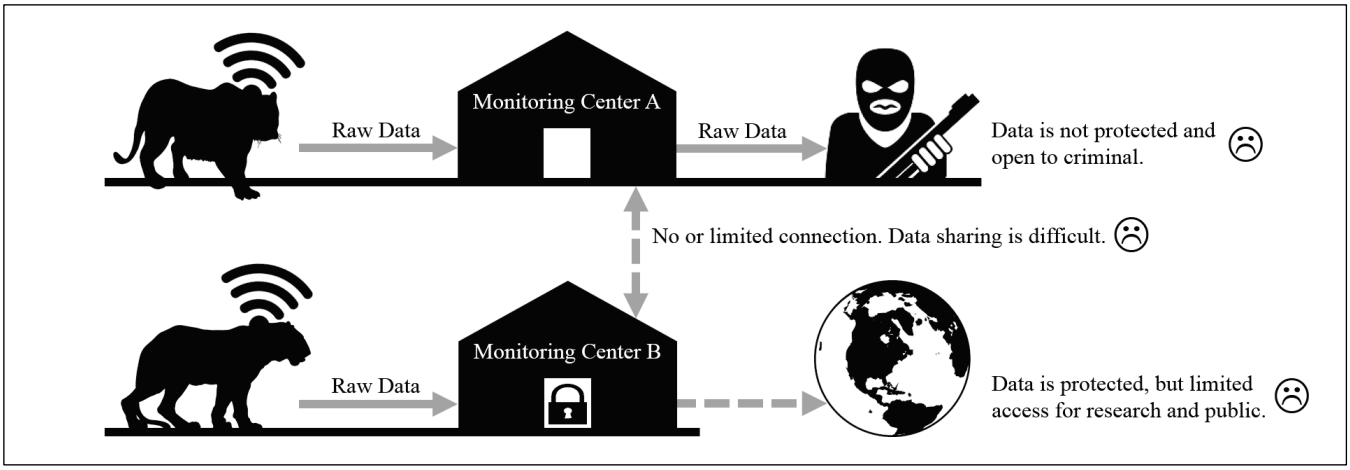

Figure 1: Situation today.

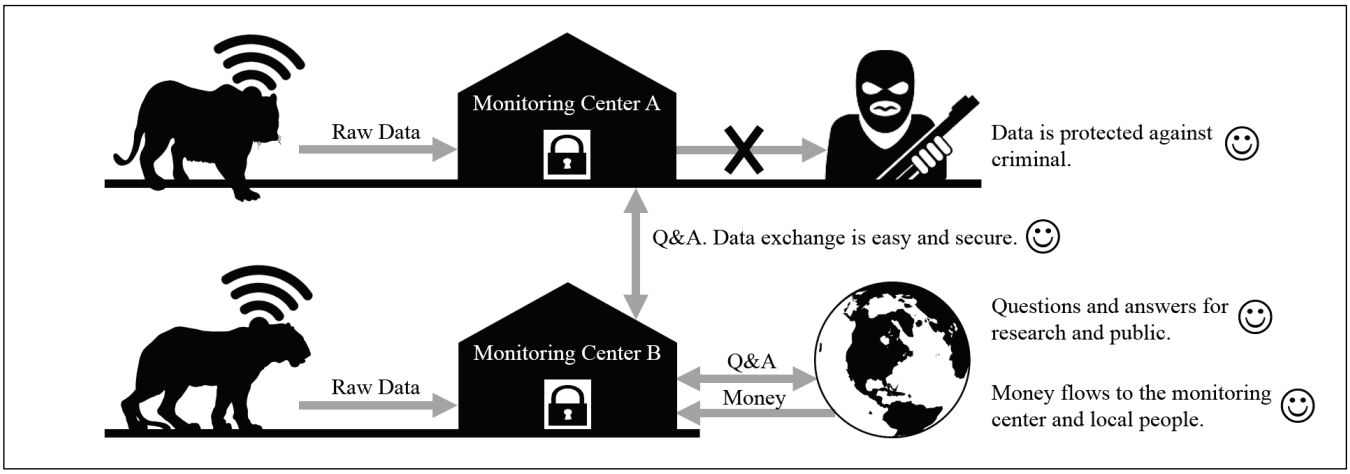

Figure 2: Proposed Solution and its positive implications.

Big Data in the described wildlife context struggles with the same problem as in the human related fields: There is a trade-off between leveraging the full potential of the data and protecting the data against fraud and misuse at the same time. We assume, that many wildlife monitoring centers, which are dealing with sensitive data about animals, are overwhelmed with that problem. We identified three critical points, as shown in Figure 1: First, some monitoring centers are unaware about the dangerous situation and provide online access to real-time geospatial data as a web search revealed. Second, other monitoring centers in response to data misuse risks provide only limited data or even stop sharing wildlife data with researchers, effectively making the data worthless. One specific case was reported to us. Third, centers become overcautious or restrained when sharing data with other centers. That behavior may lead to a significant lack of knowledge exchange and thus possibly to unfavorable activities in the conservation of species.

In order to counter these difficulties, we propose a solution that protects the wildlife data from misuse while at the same time permits regular use, as shown in Figure 2. The core idea is as follows: Instead of sending the data to the querier location (e.g. where analytics computations are performed), the query is instead sent to the data repository. In doing so, the raw data stays in the secure repository of the monitoring center. The monitoring center has the option to control the degree of accuracy of the information. Only the answers, aggregated data, necessary to research and public, leave the boundaries of the center. For example, rather than exporting raw accelerometer or GPS data, it could be sufficient for a researcher to know if an animal is active or which general geographic zone the animal is currently in.

The present paper is structured as follows. First, we briefly explain the importance of considering geospatial data in wildlife research and the corresponding risks. Then, we present a relatively new privacy-preserving framework [1], which we apply on the wildlife context. The framework is aligned with the proposal "New Deal on Data" to the World Economic Forum (WEF) [2]. Three use cases are outlined in order to demonstrate its usefulness for relevant applications. Finally, the paper concludes with a discussion and an outlook on future work.

2 RESEARCH ON WILDLIFE USING GEOSPATIAL DATA

The value of geospatial data for wildlife research is undisputed. Researchers use the data for basic research on behavior, habits and needs, for monitoring whole populations [3], for planning future nature reserves [4], or for prevention and controlling of animal diseases [5]. A better understanding is also becoming more and more important, because conflicts with humans continue to intensify. The habitat for animals is becoming smaller, its use by humans stronger. Climate change forces the animals to leave their populated places. Strong demand from Asia supports the illegal trade of rare animal species and their body parts [6]. Fortunately, geospatial data can support the fight against wildlife crime [7]. The implications of human-wildlife conflicts are well documented in research and recommendations are provided [8, 9].

In contrast, literature related to data privacy on wildlife data is sparse. The topic received global attention only as criminals have tried to hack tigers' GPS collar data in the Madhya Pradesh Reserve in India 2013. The data would provide poachers a convenience instrument to chase and kill the tigers. Fortunately, the attack had failed and the attackers had not accessed the data. (Additionally, the data was encrypted and therefore unusable even when accessed.) As a result, various scientific publications addressed the topic and established the term 'cyber-poaching' [10-14]. Cook et al. [15] provides a global overview of current troubling issues of animal tracking for conservation and management. They try to tackle the problem at management level: improve collaboration, communication, and data sharing between partners and stakeholders; create sharing policies; encourage the telemetry industry and regulators for more safety; identify concerns; learn from the past.

3 PRIVACY-PRESERVING FRAMEWORK

3.1 Key Concepts

Today there is the realization that for governments, societies and the industry to function there is the need to share information based on data. However, hand-in-hand with this need for data is also the corresponding need for privacy preservation for the subjects who may be represented in the data.

There is a pressing need for access to data in order to enable new opportunities and new engagements. There is also the growing realization that data – both consumer data and business data – are increasingly distributed across both locations and owners. This raises both security and privacy problems, and also makes it uneconomical to query the data in a centralized fashion. New approaches to distributed data processing need to be adopted that address not only the multi-owner, distributed nature of data analysis, but also the problems of security and privacy.

The standard approach to big data analytics has relied on the centralized processing of data-sets from various sources. Such an approach requires that data-sets be first collected into a centralized infrastructure before meaningful queries can be executed on these merged data-sets.

MIT OPAL. We believe that a new paradigm is needed for scalable querying of data-sets that are physically spread across the Internet, and owned by different organizations, while maintaining the security and the privacy of the information that can be derived from the distributed data-sets.

Rather than moving data towards a centralized query location, instead the query needs to be broken down into sub-queries and for these sub-queries to be delivered to the P2P nodes containing corresponding data-sets of interest. Each of the sub-queries would then be executed by the relevant node, with the results being reported back to the querier – who would merge the results into a meaningful analysis.

In this new paradigm – called MIT *Open Algorithms* (OPAL) [1] – raw data never leaves its physical location or the control of its owner. Instead, nodes that carry relevant data-sets execute sub-queries and report on the result. Scalability is improved by logical clustering of nodes that carry data-sets of a given type or nature. Groups of clusters can therefore be engineered to achieve scalability and high response rates.

Security and privacy becomes more manageable in this paradigm because each node controls its own data store, and monitors the privacy entropy of released answers. As part of access control and policy management, a user whose data resides at a node has the ability to tune-up or tune-down the granularity of the responses to each query in which their data-sets is used.

Vetted Algorithms. One key aspect of MIT OPAL is the use of vetting by domain experts of the algorithms (queries) that are permitted to run against a given data-set within a target data repository. The idea here is that algorithms must be verified by experts to be free from bias and other unintended side-effects (e.g. discrimination, etc.). Note that this vetting does not guarantee the quality of the output, which is a function of the quality of the input data.

Once an algorithm has been vetted, it becomes a template that is digitally signed by the issuer (e.g. expert themselves; institution; data sharing consortium, etc.). This template algorithm can be shared among a group of entities (e.g. within a consortium) or even be published on a public site.

Safe Answers. The OPAL model of moving the algorithm to the data and of using vetting by experts allows a data repository to choose whether or not it is willing to accept a submitted OPAL query. In the case that it does accept a given vetted algorithm, it also has the option to impose additional filtering on the resulting data prior to being returned as response to the querier. As such, the repository has the option to "dial-up or dial-down" the degree of PII information within a given response.

Blockchain Technology. Blockchain technology can be used to capture and log both vetted-queries and safe-answers, and as such provides a useful mechanism to support post-event audit and accountability. One easy way would be for the querier to compute a cryptographic hash of the sent query, and for the data repository to compute the hash of the response. Further, the technology provides benefits for data owners when they want to monetize their data, including the ability to link money and data flows, or micro-payments with small transaction costs.

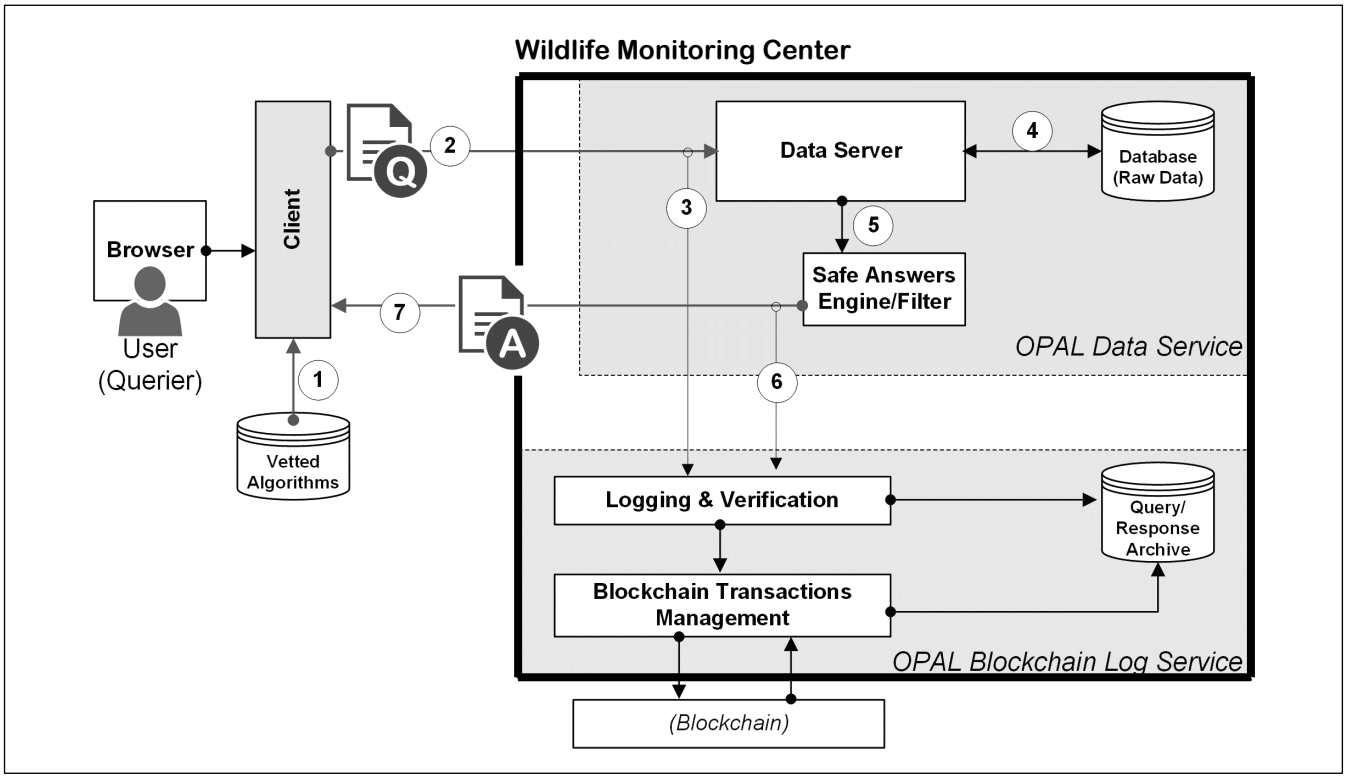

Figure 3: System Architecture.

3.2 System Architecture

The architecture of MIT OPAL is shown in Figure 3. The OPAL Data Server represents the core of the OPAL model, where the user (querier) sends one or more vetted-queries to the data server (i.e. data repository), which in sends aggregate answers only as safe-queries. The flows within the OPAL Service are summarized as follows:

- **Step-1**: The User makes use of the OPAL Client (e.g. Webb App) to select one or more template-queries from the Vetted Queries set at the OPAL Service.
- **Step-2**: The User completes the query-template with originator information and destination information, digitally signs the query and sends the signed-query to the appropriate endpoint at the OPAL Data Service.
- **Step-3**: The receiving endpoint at the OPAL Data Service verifies the signature on the User's query. If the signature is valid and known, it passes a copy of the signed-query to the OPAL Log Service for archiving and capture on the blockchain.
- **Step-4**: The Data Server interacts with the database to generate a complete response-file.
- **Step-5**: The resulting response-file is passed to the Safe Answers Engine for safety-analysis and removal of PII.
- **Step-6**: Prior to sending the safe query-response to the User, a copy of the query-response is passed to the OPAL Log Service for archiving and capture on the blockchain.
- **Step-7**: The OPAL Data Service sends the safe query-response to the User.

The system uses two separate layers for aggregating wildlife data (not shown in the figure): (a) sensitive data processing takes place within the monitoring center's data storage allowing the dimensionality of the data to be safely reduced on a per-need basis; (b) data can be anonymously aggregated across monitoring centers without the need to share sensitive data with an intermediate entity through a privacy-preserving group computation method.

To sum up, the described system allows applications to ask questions that will be answered using the sensitive raw data. In practice, querier will send code to be run against the data and the answer will be send back to them. The system ships code, not data. Thus, it turns a very hard anonymization problem to an easier security problem.

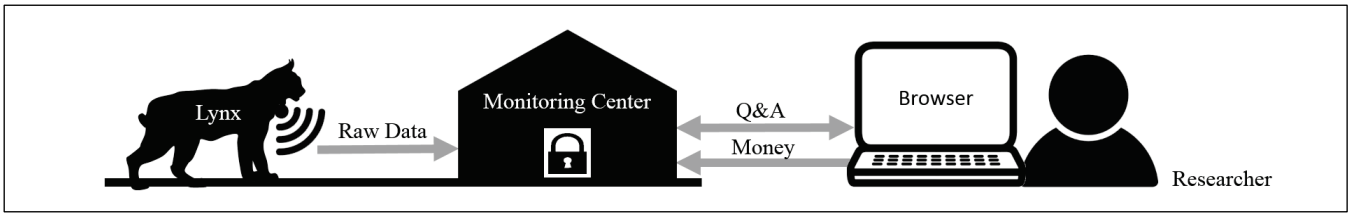

Figure 4: Use case 1: providing a data sharing system for research.

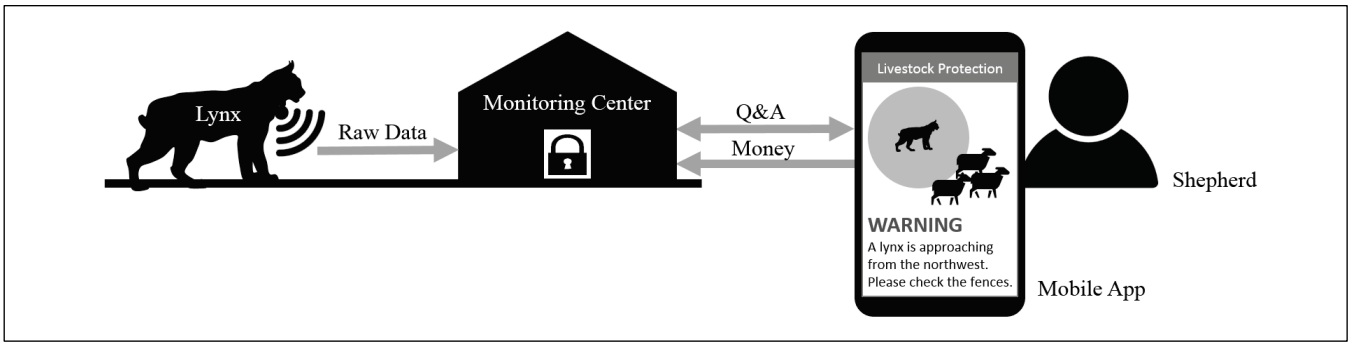

Figure 5: Use case 2: providing a warning system for livestock protection.

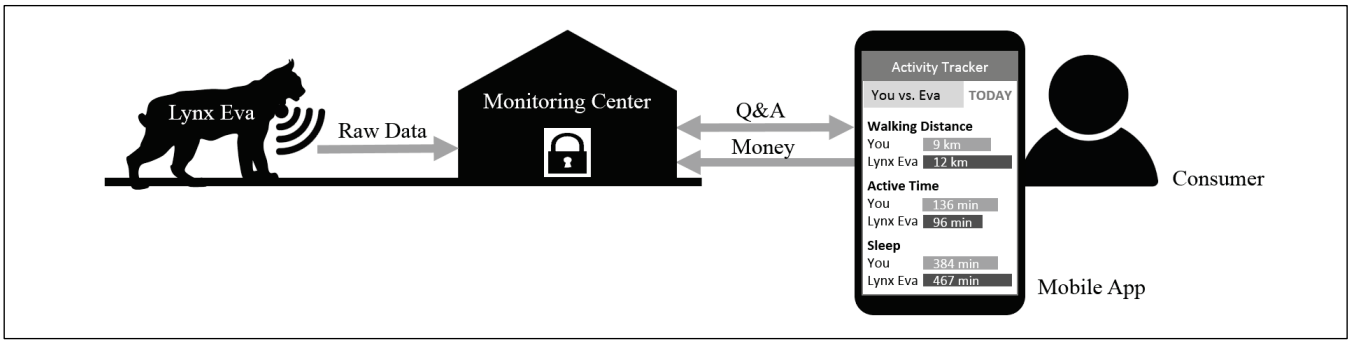

Figure 6: Use case 3: providing a system for consumer applications and donations.

4 USE CASES

There is a broad range of potential use cases for the described system. We present three of them which mainly differ in the problem they try to solve and in the target audience. The cases focus on the lynx populations the Swiss Alps because we intend to conduct a first pilot study together with a wildlife monitoring center in Switzerland. In fact it does not matter which animals or geographic areas are involved. The system is applicable in any part of the world.

4.1 System for Research

The Swiss lynxes have been researched with transmitters for more than 30 years [16]. This gives a lot of expert knowledge in dealing with geospatial data. While the first tracking devices transmitted inaudible VHF signals which were received with a directional antenna for determining the location, todays ongoing projects are using modern GPS collars. Figure 4 describes, how such collected GPS location data can be shared with a broader research community. First, the monitoring center collects the data as usual and put it into the secure OPAL database. Then, a researcher sends a research question in a standardized format to the OPAL system, which securely computes the answer and sends it back. Any kind of questions are possible, as long as the questions are confirmed as privacy-preserving (vetted algorithms). Examples for GPS based research questions are:

- What is the daily walking distance of the population?
- What is the average sleeping time?
- Which is the preferred altitude in the mountains?
- Are there seasonal differences in activity?
- How close is a lynx to settlements?

Answers to these questions do not reveal the current position of the investigated animal(s), even if real-time raw data is used for its computation. In addition, the system allows for more complicated processes such as 'federated learning', i.e. collaborative machine learning among the monitoring centers without centralized data pool [17]. We believe that such processes will help in the future to find consensus in animal conservation strategies.

4.2 System for Livestock Protection

Over the last decade, there has been an ongoing heated political and public debate about large carnivores in Switzerland. Formerly vanished species like bear and wolf are naturally migrating from bordering countries and cause conflicts between people and wildlife. Other species like lynx or bearded vulture were reintroduced by conservation programs. The Swiss government can issue killing permits when an animal becomes conspicuous and it is not unlikely that the release of researchers' GPS data will be enforced in court. Bloody sheep, attacked by wolves, are frequently on the headlines of Swiss newspapers. Many shepherds see their sheep in danger. The few bears that migrate to Switzerland cause only isolated damage (plundered beehive), but they frighten the village population in the mountains. In both cases, our system can help to spread alerts if one of these animals approaches a vulnerable area, as shown in Figure 5. Since the exact GPS location is never public accessible, the animals are still protected from attacks by residents or shepherds. Moreover, these people get a reliable tool to protect themselves and their livestock.

4.3 System for Consumer Applications and Donations

The paradigm of open innovation has gained massively in importance thanks to the simplicity of web-based communication and interfaces. Companies provide idea platforms to their customers and open their data and systems for developer communities. Our system can enable such an innovation process as well. Especially app providers can use the system for their own services. Figure 6 provides an example for a fitness tracker app. Users who 'sponsor' a given lynx can compare their daily activities with the sponsored lynx in order to improve their fitness motivation. Such a simple, additional feature may lead to a compelling value proposition of the product or service. Furthermore, the wildlife monitoring center can charge the data usage and make a business out of it. In this context the popular adoption programs for wild animals are also very interesting. Today, adopters get typically a physical certificate, but nothing else. With our system, the monitoring center can provide daily updates about the adopted animal to the adopters. Thus, the relationship between animal and its adopter will be stronger and the donations may increase or are getting more sustainable.

5 DISCUSSION, LIMITATIONS, AND FUTURE WORK

We are faced with major challenges in dealing with data. Data is collected not only since this century, but the sheer volume of data, its computer-assisted evaluation, and web-based communication channels bring new opportunities and undesirable risks at the same time. Especially the involvement of personal data provokes concerns about privacy and its protection. We want to extend the aspect of privacy to all living beings, not just humans. It is not about granting the animals the same rights as humans, but about the safe handling of sensitive wild animal data where the data can pose a threat. Geospatial data deserve special attention because the physical location can be clearly identified and the risk of abuse is correspondingly high. Tigers cannot hide in the jungle anymore. Swarms of migratory birds can be caught with hanging nets and traps. Whales can be spotted and harpooned. It is a global risk for a broad range of species. In that sense, wildlife's privacy must be discussed in a similar way as humans' privacy. Consequently, this work intends to stimulate the discussion and to contribute with a first privacy-preserving solution for wildlife data. The solution is secure, easy to implement and use, scalable, and enables new monetizing models. In principle, our system is applicable wherever sensitive data are to be made available. In the current wildlife context, we presented three relevant use cases.

There are two limitations, which leads to opportunities for future research. First, it is still difficult to judge in which case information can be safely shared and in which case a problem might arise. For instance, in our second use case, the data protection is probably already violated because the farmer could kill the lynx based on the rough clue about the location. In addition, a single query can be verified as safe, but it is not in combination with others or when requested multiply times. Second, inductive research approaches are limited with the proposed system. Researchers are not able to explore the raw data in an open-ended manner to detect patterns and regularities.

ACKNOWLEDGMENTS

This work was partially financed by Auto-ID Labs ETH/HSG.

REFERENCES

[1] Thomas Hardjono, David Shrier, and Alex Pentland. 2016. TRUST::DATA: A New Framework for Identity and Data Sharing. *Visionary Future LLC.*
[2] Soumitra Dutta, ed., 2009. Global Information Technology Report 2008-2009. *World Economic Forum.*
[3] Simon Capt. 2007. Monitoring and distribution of the lynx Lynx lynx in the Swiss Jura Mountains. *Wildlife Biology, 13*(4), pp.356-364.
[4] Carlos Carroll, Reed F. Noss, Paul C. Paquet, and Nathan H. Schumaker. 2003. Use of population viability analysis and reserve selection algorithms in regional conservation plans. *Ecological applications, 13*(6), pp.1773-1789
[5] Jerome E. Freier, Ryan S. Miller, and Kenneth D. Geter. 2007. Geospatial analysis and modelling in the prevention and control of animal diseases in the United States. *Vet. Ital, 43*, pp.549-557.
[6] Vincent Nijman. 2010. An overview of international wildlife trade from Southeast Asia. *Biodiversity and conservation, 19*(4), pp.1101-1114.
[7] Isla Duporge. 2016. Analysing the use of remote sensing & geospatial technology to combat wildlife crime in East and Southern Africa.
[8] Philip J. Nyhus, Steven A. Osofsky, Paul Ferraro, Francine Madden, and Hank Fischer. 2005. Bearing the costs of human-wildlife conflict: the challenges of compensation schemes. *CONSERVATION BIOLOGY SERIES, 9*, p.107.
[9] Amy J. Dickman. 2010. Complexities of conflict: the importance of considering social factors for effectively resolving human–wildlife conflict. *Animal conservation, 13*(5), pp.458-466.
[10] Joanne Carew. 2013. Cyber Poaching Threatens Endangered Wildlife. *Green It.* ItWebLimited.
[11] Jeff Dupain. 2013. Using Cyber Tracking Technology to Outsmart Poachers. *African Wildlife Foundation.* AWF.
[12] Sasha Ingber. 2013. "Cyberpoaching" Feared as New Threat to Rare Wildlife. *National Geographic.* National Geographic Society.
[13] Stephen Messenger. 2013. Cyber-poachers Hack GPS Collar Data to Pinpoint Tigers. *TreeHugger.* MNN Holdings LLC.
[14] Laura Sinpetru. 2013. Cyber-Poaching Is on the Rise, Conservationists Say. *Cyber-Poaching Is on the Rise, Conservationists Say.* Softpedia.
[15] Steven J. Cooke, Vivian M. Nguyen, Steven T. Kessel, Nigel E. Hussey, Nathan Young, and Adam T. Ford. 2017. Troubling issues at the frontier of animal tracking for conservation and management. *Conservation Biology.*
[16] Urs Breitenmoser, and Heinrich Haller. 1993. Patterns of predation by reintroduced European lynx in the Swiss Alps. *The journal of wildlife management*, pp.135-144.
[17] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Artificial Intelligence and Statistics*, pp. 1273-1282.