NEWSLETTER

OCTOBER 2020 | ISSUE 1

Real-time Analytics for Internet of Sports

Marie Curie European Training Network

OBJECTIVES

RAIS aspires to provide for 14 EarlyStage Researchers (PhD students) a world-class training within abroad spectrum of subjects establishing a fertile inter-disciplinaryresearch and innovation community that will advance:

Wearable Technology

Wearable Sports Sensing and Quantified-selfDevices and Accompanying Middleware

Block-chain Powered IoT

Decentralized Block-chain Powered IoT Platforms (generating hundreds of billions of transactions per day) for Big Data Mining

Real-time Edge Analytics

Real-time Edge Analytics And Predictive ModellingTo Capture A Broad Range Of Sports-related Data AndTrends (e.g., activities and contextual information), Critical In A Variety Of Application Settings

RAIS fellows receive a thorough "hands-on" research training as well as significant exposure to nonacademic environments through industrial secondments. Our rich set of network-wide events, including Interactive Online Seminars, entrepreneurship events, hackathons, workshops and conferences, will safeguard both fellows work as a solid team and individuals development as experts.

MARIE CURIE ACTIONS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Innovative Training Networks (ITN) - RAIS No 813162

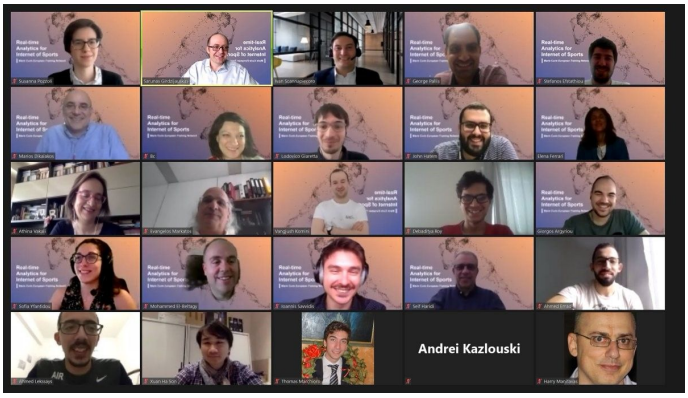
RAIS

RAIS LATEST NEWS

Due to the pandemic caused by COVID-19, the secondments have started online

Online
May 15, 2020

RAIS Consortium completed online its Midterm Check Meeting due to Covid-19 pandemic situation

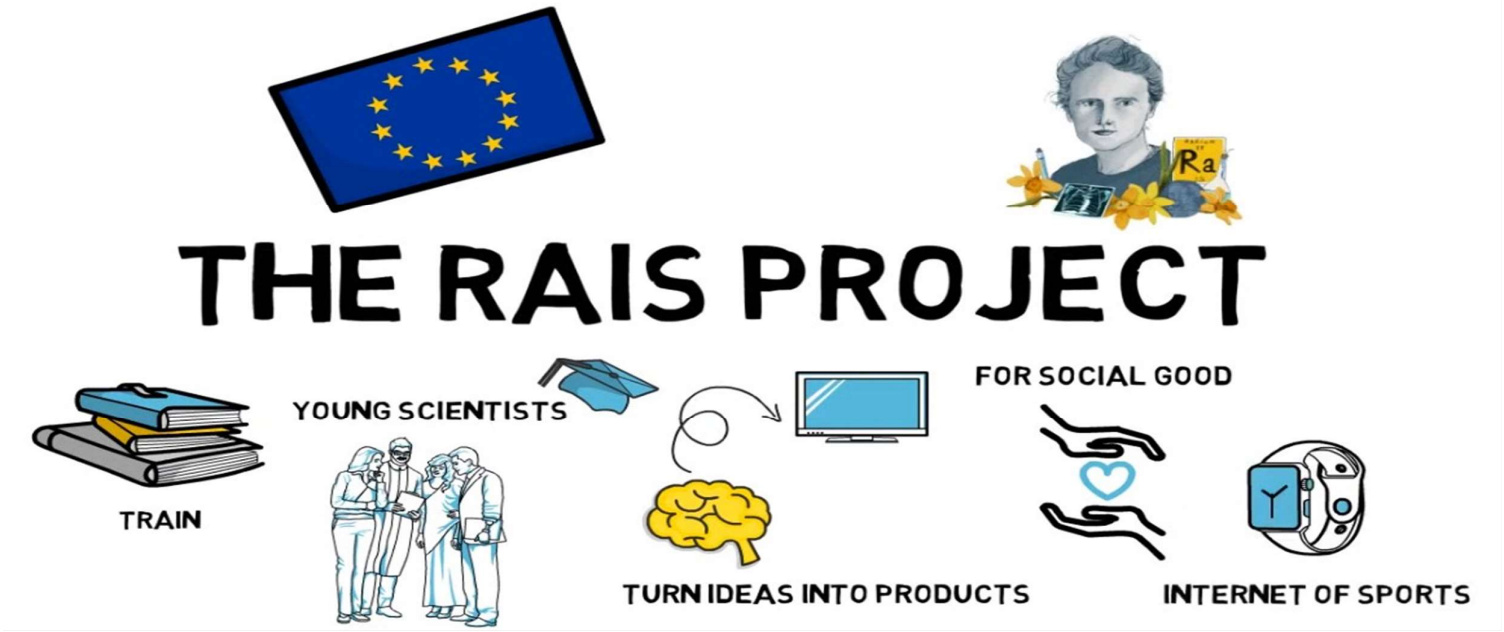


Thessaloniki (Greece)
December 10-11, 2019

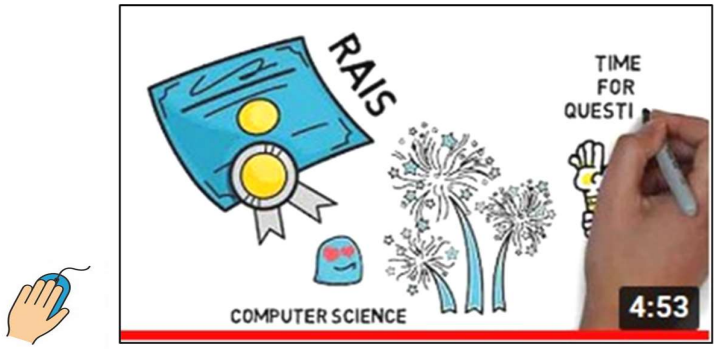
The Second Plenary meeting of the RAIS Consortium was held successfully in Thessaloniki, Greece.
The duration of the meeting was two days from the 10th to 11th December 2019

Como (Italy)
January 16-17, 2020

The kick-off meeting of the RAIS Real-time Analytics for Internet of Sports Marie-Curie European Training Network took place in Como, Italy on the 16th and 17th of January 2019



CONTENTS



ESR 1 GEORGIOS ARGYRIOU | UNIVERSITY OF CYPRUS (UCY) | CYPRUS

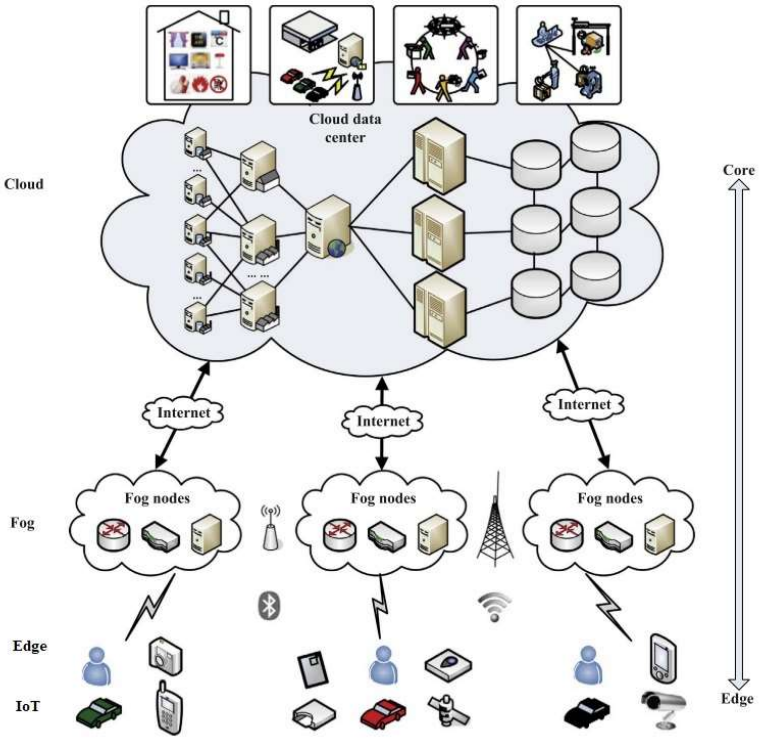
Taming massive IoT data with middleware on Edge

According to International Data Corporation prediction [1], by 2025 more than 150 billion devices will be connected worldwide, the global data will reach 180 zettabytes and 70% of the data generated by IoT devices will be processed on the edge of the network. In order to meet the new challenges, we need to change our approach in data processing by refraining from centralized processing and unilateral usage of cloud computing resources and embrace decentralized processing using resources on the edge.

The main objective of my research is to build a monitoring framework for wearable devices that takes into account their key characteristics: limited processing power, heterogeneity, dissemination of sensitive data, low-energy sufficiency (battery-powered), intermittent network connectivity. The devices that participate in such an environment are divided in three basic categories. The sensors that are mobile, battery powered, have very low processing power and are mainly for sensing. The edge resources that are plugged in the power network and used mainly for processing. These ones are very small computers and can create tiny datacenters. The third category is a hybrid of the previous ones. In this category there are battery powered devices that have sufficient processing power, are mobile, have some sensors attached but are also able to connect to external sensors. Such devices are usually smartphones and we can think of them either as sensors or as edge resources.

With pushing from the cloud and pulling from IoT, a new layer is created in the middle -usually called Edge or Fog-, which can offload the network, bring processing closer to IoT and its inherent characteristics can solve problems like privacy as the data stays closer to the user. The big challenge here is how to organize such an environment where there are heterogeneous IoT devices with intermittent connectivity and the edge resources are processing the receiving data. State-of-the-art research work and new platforms and tools can offer solutions to this challenge. Serverless computing [2] handles virtually all the system administration operations needed to make it easier for programmers to use the resources. Right now it is a model that comes to life with Function as a Service platforms in cloud computing infrastructure. Nevertheless, we research whether such platforms are efficient to be used in the Edge layer. Except for the famous cloud vendors, several open source FaaS platforms exist and grow quickly as they

have many contributors [3]. Extensions to a pre-existing platform that enrich its functionality can end up in a middleware prototype that can be deployed on Edge/Fog nodes.



When it comes to battery powered edge devices my research is focused on lowering power consumption, which is the main challenge according to experts from industry. Towards that end, we can propose techniques for lowering the amount of data dissemination or find smart ways to adapt computation execution to the appropriate resources. There already are algorithms that do adaptive sampling very efficiently for univariate datasets. The next challenge is how such algorithms can be efficient with multivariate datasets. Such datasets are the ones where each sample has many values and each value is for different components (e.g. x and y axis), or several univariate datasets that we treat as a multivariate one (e.g. temperature and humidity). Adaptive sampling on multivariate datasets can be even more efficient in lowering IoT data volume.

References:

[1] Zwolenski, M., & Weatherill, L. (2014). The digital universe: Rich data and the increasing value of the internet of things. *Journal of Telecommunications and the Digital Economy*, *2*(3), 47.

[2] Jonas, E., et al. (2019). Cloud programming simplified: A berkeley view on serverless computing. *arXiv preprint arXiv:1902.03383*.

[3] Palade, A. et al. (2019, July). An evaluation of open source serverless computing frameworks support at the edge. In *2019 IEEE World Congress on Services (SERVICES)* (Vol. 2642, pp. 206-211). IEEE.

ESR 2 IOANNIS SAVVIDIS | UNIVERSITY OF CYPRUS (UCY) | CYPRUS

A platform for full control and secure sharing of personal data

We live in an era which is characterized by the creation of more and more data based on which, we make decisions to maximize the desired results. Hundreds of thousands of data are generated daily by ordinary users such as athletes simply by using a smartphone and a smart band. But, imagine there was a platform for the users to completely control their data and create value from it. Unfortunately, this is not how the system currently works.

Nowadays, data is collected and stored in a centralized fashion by companies. As a consequence, users can lose control over their data. Problems arising from this established situation are related to data leakage, compromisation of users' privacy, arbitrary data management by the third parties and no recompense for the data providers. On top of these, the storage of users' data in centralized systems which are independent of each other makes it difficult to share data between stakeholders.

To solve these problems, our research aims to create a platform that allows users to have complete control over their data. At the same time, they will be able to sell it directly to the third parties which they trust and be rewarded for it. This platform will focus on ensuring the security and privacy of users in both the storage of their data and its sharing. Finally, it will ensure the integrity of the data and the fair recompense of the users. Such a platform can be separated into categories of problems which include the areas of access control systems, best data sharing practices, transaction protocols and data structures.

Currently, my research focuses on access control systems and safe data sharing processes. Specifically, taking into account the context of the Internet of Sports, ways and techniques of access control to users' data are studied. Access control typically involves two steps: authentication and authorization. Authentication refers to the process of identifying that a user is the one who claims to be, and authorization refers to the actions that the users are allowed to perform.

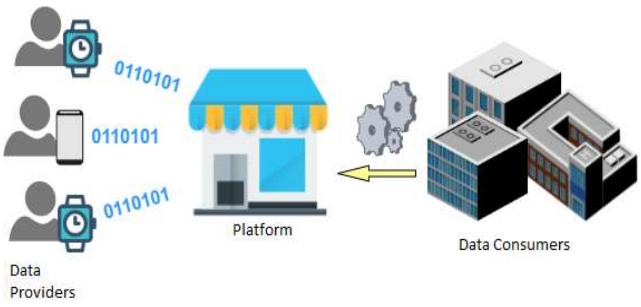


Figure 1: Data platform architecture 1

In combination with that, I make use of blockchain technology. Essentially, blockchain can be considered as a book or a ledger which is copied and shared with every participant in a network. Every action that happens in the network is called transaction and it is written in this book. Examples of transactions can be the exchange of money between two parties or the storage of data in the network. In this analogy, every page of this book is a block and all the blocks linked together are creating the blockchain. In our vision, blockchain serves multiple purposes. Because of its decentralized nature, the blockchain makes the sharing of the data easy and fast. Also, it is used as an immutable record that ensures the availability and the integrity of the stored data increasing significantly its quality as it makes it almost impossible for malicious users to alter or delete it. Lastly, blockchain allows the storage of provides a trustless environment removing the need for a third party to supervise and validate the procedures

ESR 3 LODOVICO GIARETTA | ROYAL INSTITUTE OF TECHNOLOGY (KTH) | SWEDEN

Decentralized Machine Learning over Networks

As our technology advances, we, as a society, are producing more and more data that needs to be analysed to improve our services and lives. And more and more of these data come in the form of graphs. From social networks to our global Internet infrastructure, from interactions among proteins in our bodies to envisioned smart electricity grids, graphs allow us to represent the relationships between data points and extract meaningful patterns and underlying structures of each of these systems. This is the goal of Graph Representation Learning (GRL) [1], a field of growing interest in the Machine Learning (ML) community. It allows us to identify tightly connected communities, highlight similar roles played by seemingly unrelated entities, and embed each entity of the graph in a low-dimensional space that summarizes the main patterns that drive the overall structure of the network. The right side of **Figure 1** shows an example of graph-based analytics on fitness data, with multiple relationships uncovered and classified. With this information, more advanced ML models can be built, that use these insights to predict complex future behaviours.

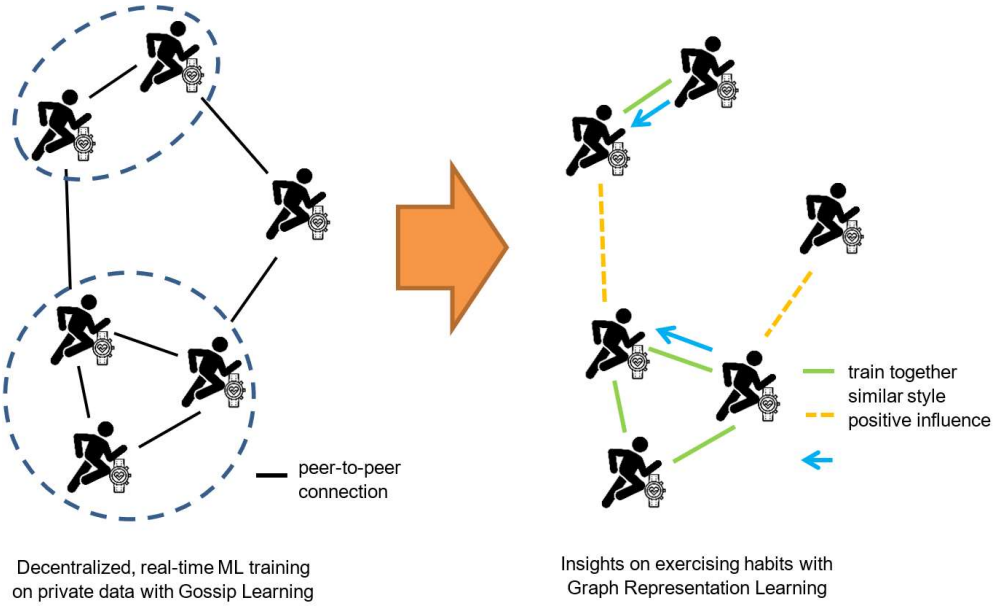


Figure 1: Combining Decentralized ML and GRL in an IoS setting

However, many of these graphs can also be exploited in a different way. Networks of Internet devices, IoT devices or smart sensors can not only collect graph-structured data but can also serve as a decentralized computing grid to process these data. By employing Decentralized Machine Learning

frameworks such as Gossip Learning [2], ML models can be trained collaboratively, directly at the source, avoiding the bottlenecks of centralized collection. This new approach is fundamental to achieve user privacy, security and a fairer data-centric economy and society, in which users have control over the data they produce, are aware of their value and are rewarded for their usage. This advancement is one of the key goals of the RAIS project.

My research focuses on advancing and intersecting GRL and Decentralized ML, with the final goal of merging these technologies and combining their benefits, with a particular attention to the Internet of Sports (IoS), which is the core interest area of RAIS. Decentralized ML frameworks can be used as the computing backbone for achieving data privacy and fairness in GRL tasks. In the IoS sector, fitness devices could form a decentralized processing grid on top of which GRL models can be trained to recognize, for example, how interactions among athletes affect their training styles, athletic results and motivations. This architecture is shown in **Figure 1**. On the other hand, GRL models could be used to improve the performance of Decentralized ML frameworks. By identifying communities of well-connected devices or devices that act as bridges or hubs [3], GRL models can provide insights to these frameworks to optimize their decentralized layout for faster convergence.

References:

[1] W. L. Hamilton, R. Ying, and J. Leskovec, 'Representation Learning on Graphs: Methods and Applications', Available: http://arxiv.org/abs/1709.05584.

[2] R. Ormándi, I. Hegedüs, and M. Jelasity, 'Gossip Learning with Linear Models on Fully Distributed Data', *Concurr. Comput. Pract. Exp.*, vol. 25, no. 4, pp. 556–571, Feb. 2013.

[3] L. Giaretta and Š. Girdzijauskas, 'Gossip Learning: Off the Beaten Path', in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019.

ESR 4 HA XUAN SON |UNIVERSITY OF INSUBRIA (INSUB)| ITALY

Personal data storage (PDS) model in the IoT environment

Nowadays, the Internet of things (IoT) plays an indispensable role in our lives. We can list many areas of exploiting its advantages, from personal environments, such as smart homes, self-driving cars, to more complex environments, such as smart cities. For all the abovementioned environments benefits are increasingly expanding because the devices are being upgraded to be smarter, with higher capabilities, and better in interacting. However, there are still some key limitations that are affecting this development wase, one of the most critical one is related to security and privacy problems.

Services in IoT environments are provided by a huge number of service providers. Hence, it is difficult to assess the risk level of the service providers wrt the secure and privacy-aware management of user data. For this reason, many users are still hesitant to use IoT services. Currently, the existing protection mechanisms have many limitations due to the fact that most of them are designed based on centralized architecture (also called system-centric protection mechanisms). Such an approach has two main limitations. First of all, it is hard to fully trust the available privacy mechanisms, since this requires to fully trust the service providers. Secondly, users' security and privacy requirements are diverse; for example, a user can easily share health data for healthcare purposed but he/she does not want to share this data with services with different purposes. This is impossible for traditional approaches, since privacy policies are built in parallel with the system settings process and users have little ability to customize them. Moreover, users must learn how to use these services and manually change the settings to suit their individual requirements.

My research focuses on a shift in approach from focusing on data protection managed by the service providers to enabling user empowerment. We are designing and developing a personal data storage (PDS) model for the IoT environment. According to this model, the user has the control on his/her personal data and he/she has empowerment to grant or revoke the access permissions to service providers. In addition, the proposed approach promotes user awareness of how and when data is being collected and the collection purpose. Moreover, we are also studying how to apply machine learning techniques to: automatically adjust user privacy settings when new service providers join the PDS and adapt privacy settings to changes with minimal user intervention.

ESR 5 ANDREI KAZLOUSKI | FOUNDATION FOR RESEARCH AND TECHNOLOGY – HELLAS (FORTH) | GREECE

Using the specifics of data transfer from a fitness tracker to profile its users

The advance of wearables and ubiquitous data collection implies an unquenchable source of private information being send to vendor servers. We gathered a set of smartbands from various manufacturers, and investigated whether we can extract any meaningful insights form the transferred data (even encrypted.)

Although the majority of devices employ TLS encryption to safely transmit sensitive data of their users to the cloud, the encryption alone might be not enough. We noticed that different activities that users perform produce different traffic patterns. We set to explore those patterns, and what kinds of data might be extracted. Mostly we utilized the size of the individual packets, their frequency, and their "neighbors". We were able to successfully infer sensitive information from the devices of well-known brands (Xiaomi, Samsung, etc.). Such data include number of heart rate measurements, number and duration of workouts, whether a user slept on a particular date, food information, and many more.

Currently we are planning to extended our studies to the third-party apps. Well-known vendors choose to collaborate with popular fitness apps, so users can synchronize health activities between them. It is often the case that the official mobile apps of the most popular vendors are reinforced with state-of-the-art security mechanisms, while their partners tend to have lesser development teams, and lower security levels.

We are investigating whether these partner apps might leak some data, and enable user profiling. Associated third-party apps are tracking weight changes, recording external workouts. Some of them are virtual reality games that involve taking steps or cycling.

ESR 6 THOMAS MARCHIORO | FOUNDATION FOR RESEARCH AND TECHNOLOGY – HELLAS (FORTH) | GREECE

Emerging pitfalls in privacy-preserving disclosure of fitness data

Data collected by wearable fitness trackers (like smartwatches and smartbands) are becoming more and more valuable. This is partly due to the general emergence of IoT data-driven technologies and partly due to the increasing accuracy and variety of the measurements performed by wearable sensors. If such value translates into some gained benefit, disclosing collected data might be appealing for the manufacturers or the device users.

However, when publishing personal data, privacy must always come first. It is well known that just removing identifiers and aggregating records from different user does not guarantee complete privacy, as they might still contain information that points to a certain person. Employing database terminology, attributes of a record which enable the identification of an individual are called *quasi identifiers*.

A common solution adopted to tackle this problem is k-*anonymity*, whose main idea is to use generalization or suppression of quasi identifiers to guarantee that each tuple of quasi identifiers present in the data occurs at least k times. Nonetheless, k-anonymity may not be sufficient to protect privacy against attacks that involve continuous observation of data.

The tables below show an example of 2-anonymization of a table where “Age” and “Sex” are quasi identifiers. “Steps” and “Calories” are not anonymized as they are not unique for an individual.

ID	Age	Sex	Steps	Calories
001	25	M	[Time series]	[Time series]
002	23	M	[Time series]	[Time series]
003	27	F	[Time series]	[Time series]
004	21	F	[Time series]	[Time series]

Age	Sex	Steps	Calories
[25 - 30]	*	[Time series]	[Time series]
[20 - 24]	*	[Time series]	[Time series]
[25 - 30]	*	[Time series]	[Time series]
[20 - 24]	*	[Time series]	[Time series]

My current work consists in assessing to what extent continuous observation is a threat and the results of some experiments I conducted suggest it actually is. In particular, even with time series records of steps and calories from just two different months, it is possible to link the time series from one month with its counterpart from the other month with higher accuracy than one may expect. This implies that more information is disclosed than intended, mining the privacy of the users. Therefore, it is necessary to design anonymization techniques that take into account this kind of threat.

Another issue that needs to be faced is that common anonymization techniques require access to the complete data in order to be applied. Hence, the entity who is responsible for anonymizing the records must be trusted. To fully protect users privacy, it is of primary importance to investigate decentralized ways of achieving anonymity.

ESR 7 AHMED LEKSSAYS | UNIVERSITY OF INSUBRIA (INSUB)| ITALY

Exploring Security of Smart Devices

Towards securing smart devices against malware and network attacks using blockchain

Did you know that by reading these words you have already interacted with many security mechanisms that protect your digital life while using a computer? By now, I believe you guessed the topic of my research: computer security. However, this field is very vague, so I only focus on the security of smart devices which are referred to as Internet of Things. Smart devices are all around us today such as smart fire detectors, smart washing machines, smart locks, wearable devices, etc... I am doing my research to protect these devices from being compromised by investigating how they communicate with each other and with internet. In addition, on internet, not all programs are legitimate and they simplify our lives. There are always malicious software (or malware) that damage devices around the world, spy on devices' users, or benefit financially by using devices' resources or ask for a ransom. With these threats in mind, I also focus on malware detection and mitigation for smart devices. There is an important pillar that is cumbersome to discover in smart devices: the firmware. Firmware is the software that makes the smart devices interactive and operative, so a part of my work touches on that aspect as well with all its challenges.

In order to achieve these goals, I explore a new technology called blockchain. It is (as you may have guessed) a chain of blocks holding data in a structured manner and replicated on devices registered in the blockchain. It is also transparent and immutable, so the data stored in the blockchain cannot be removed or tampered with because of data replication and its available at all time to the public. This feature is important for computer security because if a malicious actor is identified, its information can be exposed to the public without giving the actor the possibility to remove the information. There are many use cases for this feature in security like authentication, authorization, secure updates for outdated firmware, etc. However, I also work on maintaining privacy in the blockchain because of the transparency feature. Hence, my research is not mainly about using the technology, but it is more about exploring its weaknesses and address them to satisfy the security and privacy requirements for a better protection of smart devices.

Thus, my research journey touches on all aspects related to smart devices' security and their interaction with the world in order to protect users' privacy.

ESR 8 SUSANNA POZZOLI | ROYAL INSTITUTE OF TECHNOLOGY (KTH) | SWEDEN

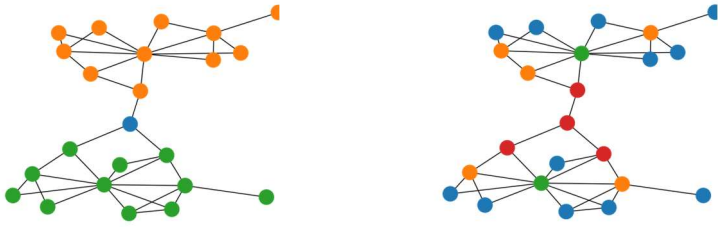
Roles in Network

Network Analysis allows us to study relationships between entities, such as relationships between users (social networks), links between pages (information networks), and connections between devices (communication networks) [2].

Currently, Network Analysis is largely based on community detection, also known as clustering, and centrality measures. Communities are sets of nodes that are more densely connected within each community and sparsely connected between different communities [3]. They are detected in order to recognize groups of friends on Facebook or topics on Wikipedia and thus are useful for tasks such as friend suggestion and link prediction. Then, if one was to determine the importance of the nodes in the graph based on how they are connected, there are several centrality measures that can be used. For example, PageRank [1] was initially used by Google to rank web pages and made it stand out from the other search engines at the time.

While communities and centrality measures work well for tasks such as node classification, link prediction, and ranking, they largely underperform in case of anomaly detection, for instance. This is because there is an underlying signal in networks that they do not fully capture: the roles of the nodes. So, in addition to community detection and centrality measures, we could add a third pillar to network analysis: role discovery. Indeed, when sociology (the study of human society) started modeling social structures as networks, attention was given to the position of individuals in those networks, also known as their role. The concept of role is not limited to social networks and is promising to Network Analysis in general.

To explain what a role is (when compared to communities), consider the following example. Companies divide people into departments, but at the same time, people hold different titles, such as Product Manager, CEO, and Engineer. In other words, it is possible to group nodes by either community (departments) or role (title) based on how many mails people exchange, for instance. Likewise, other domains have similar roles too: there are verified accounts on Twitter and disambiguation pages on Wikipedia.



Communities or Departments

Roles or Job Titles

Figure 1. Communities vs. Roles
Source: https://youtu.be/S4QZiUPJkRI

Figure 1 shows the difference between communities on the left and roles on the right. There are two departments (orange and green) and four titles (blue, orange, green, and red) in this organizational chart. For example, the red nodes are the executives of the company.

While communities are sets of nodes that have greater connectivity internally than externally, roles are sets of nodes that are structurally similar [4], that is, sets of nodes whose neighbors are similar. Roles are relevant because they allow us to recognize relationships between entities that are not directly connected since they are in different parts of the same graph or in different graphs.

Roles are pertinent to the Internet of Things and thus to the Internet of Sports because they are used for node classification, link prediction, anomaly detection, and visualization [5] as well as for influence maximization, to better train distributed learning algorithms. However, unlike community detection, role discovery is still in its early stages.

Within RAIS, I will review roles and role discovery as well as research applications of roles to the Internet of Sports. Since it is not easy to detect roles, owing to the heterogeneity of the network(s), I will initially study how to recognize roles in networks that present challenges found in real-world scenarios. If we go back to the example of the organizational chart (Figure 1), then it is very unlikely that the departments will have an identical structure. For example, departments may differ in the number of employees (number of nodes) and the internal organization of an HR department may differ from the organization of an R&D one (the edges between the employees). Therefore, nodes playing the same roles will likely show a similar but not identical connection pattern. Issues such as different numbers of neighbors need to be solved in order to make role discovery more widely applicable. Furthermore, I will research how to evaluate the performance of role discovery. In a similar fashion to how several scores exist for clustering performance evaluation, such as homogeneity or silhouette score, performance metrics can be designed for role discovery as well. Specifically, it is necessary to develop methods that assess roles whether or not there is a ground truth, which is not always available in real life.

References:

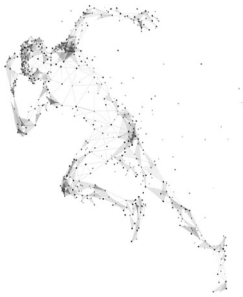
[1] Sergey Brin and Larry Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Computer Networks and ISDN Systems. 1998.

[2] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* 2010.

[3] Santo Fortunato. *Community Detection in Graphs.* Physics Reports. 2010.

[4] Ryan A. Rossi and Nesreen K. Ahmed. *Role Discovery in Networks.* IEEE Transactions on Knowledge and Data Engineering (TKDE). 2015.

[5] Ryan A. Rossi et al. *From Community to Role-based Graph Embeddings.* 2019. arXiv: 1908.08572 [cs.SI].



ESR 9 DEBADITYA ROY| ROYAL INSTITUTE OF TECHNOLOGY (KTH) | SWEDEN

Representing uncertainty in human motion classification efficiently and effectively

Analytics on IoT (Internet of Things), based on extracting information and data from multiple sensors connected over networks is a key driver for personalized products in different industries. Sensors are widely used in sports through connected devices. Collecting data from those sensors and creating novel analytical algorithms using machine learning and deep learning can directly enhance the performance of the athletes involved in any sporting activity, help in injury prevention and improve techniques. Racefox (a partner of RAIS project: https://racefox.com/en/home) is a success story that provides AI-based digital coaching to runners by leveraging the data originating from the chest-worn accelerometer belt. This aligns very much with the idea of the RAIS project, where we aim to innovate efficient AI solutions utilizing high-velocity sensor data with a goal of creating a commercially viable product.

One of the methods in the AI system of Racefox is to classify motion effectively, according to the sports activity (for e.g., different styles of running and skiing) by looking at accelerometer data. The general pipeline for such a process (as depicted in Figure 1) involves: i) Preprocessing the accelerometer signal, ii) Detecting cycles in the signal (as the sports activities in the use-cases are generally periodic and consistent), iii) Extracting features manually for those cycles and using them for classifying different types of motion (for e.g. running/walking/jogging etc.). Users are provided with feedback based on the KPIs, classification in real-time.

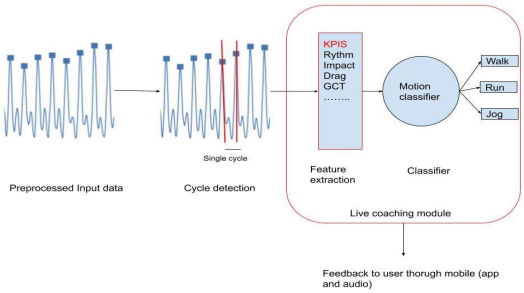


Figure 1: A pipeline of motion classification

One way to improvise the pipeline is to devise methods for automatic feature extraction using deep learning techniques to improve the existing pipeline. However, a more novel approach is to model the stochasticity of process and incorporate it into the pipeline.

Human motion is generally of stochastic nature, although in sports the stochasticity is somewhat traded off by incorporating certain repeating actions over a period of time, still, the underlying uncertainty is not discarded altogether. With uncertainty estimates, we can build learning systems with better generalization capability. Classification in uncertainty estimation is not only correctly detecting a class from another, but also to inform about the certainty of the prediction (e.g., 51% confidence of a predicted class and 100% confidence of predicted the class can mean differently for different use cases). As seen in Figure 2, if we feed signal data, where we have correct labels (the one originating from the yellow circle) and no labels (the data originating from green and red circles), to a deterministic classifier, the data originating from green and red distribution will be misclassified to represent labels from the yellow circle. However, for a stochastic classifier, these red and green samples will be detected as out-of-distribution samples or low-certainty predictions.

Human activity recognition[3] is a super-class of motion classification, where different activities performed by humans (daily or non-daily, motion or non-motion) are identified by looking at the data (inertial, visual, audio) originating from those activities captured by certain sensors. It has wide applications in healthcare, sport, security etc. One of the goals of the research within RAIS will be to explore uncertainty estimation to detect out-of-distribution samples[1] in activity recognition use case. This will result in a robust learning system that produces not only accurate but also well-calibrated predictions[2] with quality uncertainty estimates.

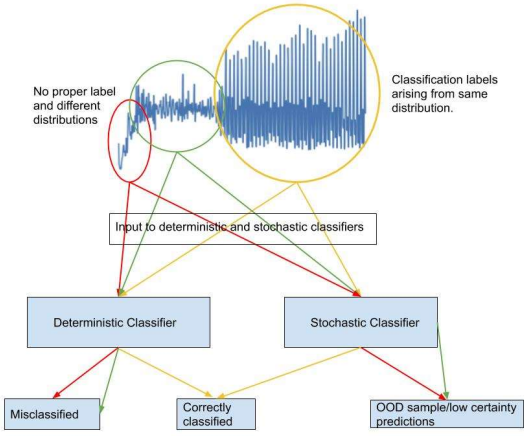


Figure 2: Data from different distributions fed into deterministic and stochastic classifiers.

In the later part of the research, this system can be further enriched in a semi-supervised fashion by labelling the out-of-distribution samples to be used for active learning (as shown in Figure 3). This research can find direct applicability in injury/detection prevention for athletes, as during an injury the input signal would change, thus triggering a change in the distribution of data, that can be effectively captured by the system.

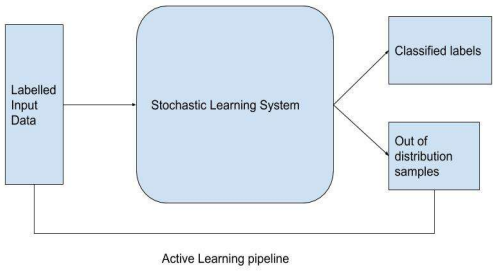


Figure 3: Active Learning Pipeline

References:

1. *Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." arXiv preprint arXiv:1610.02136 (2016).*
2. *Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems. 2017.*
3. *Kim, Eunju, Sumi Helal, and Diane Cook. "Human activity recognition and pattern discovery." IEEE pervasive computing 9.1 (2009): 48-53.*

ESR 10 AHMED EMAD SAMY YOSSEF AHMED | ROYAL INSTITUTE OF TECHNOLOGY (KTH) | SWEDEN

Context-aware decentralized solutions for knowledge extraction on graphs

Most of today's data are highly connected; they are structured in the shape of networks— from social networks down to microscopic-level networks such as genome and protein networks. To exploit the knowledge hidden in these network-structured data, graph mining techniques are to be utilized. In the RAIS project, our core work is to provide efficient distributed AI solutions for analysis of networks arising from various sport activities as well as induced by wearable devices (e.g. smart watches, heart rate monitors, sensors etc).

Graph representation learning (GRL) is a common approach to network mining and analysis. GRL is a set of techniques where the objective is to learn the graph/network structure and construct latent feature vector representations (embeddings) for the nodes and edges in the graph. Learning high-quality representations can be useful in automating prediction and other downstream tasks such as search and personalized recommendation. Figure 1 shows an example of ads, where nodes, except for the one in the red box, are visually distinguishable into two communities: Running vs. Football. Figure 1 is an example of using a GRL technique to learn one-single embedding per node in a network. As a result, the node in the redbox which has interactions with both communities, had to be exclusively "classified" closer to the group which it had more interactions with.

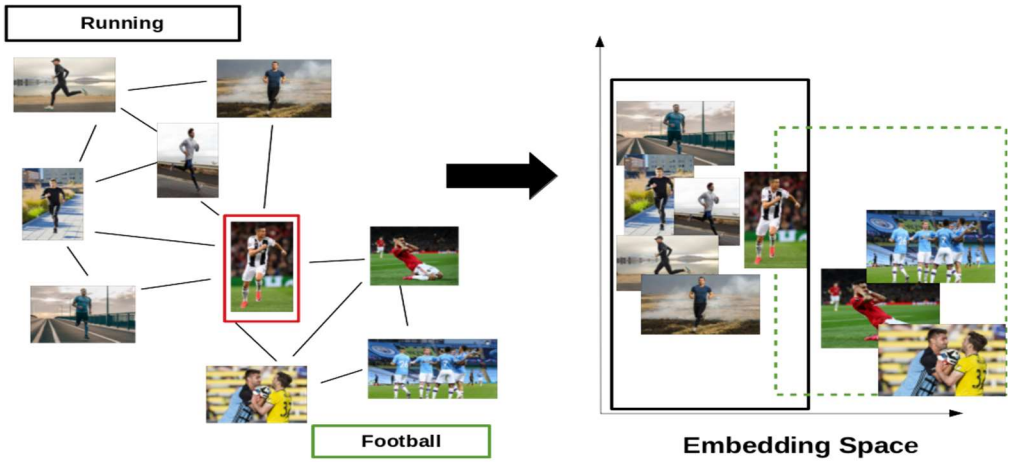


Figure 1: Global Single-Embedding

In our work we argue that nodes have multiple-facet nature, and are typically involved in complex interactions in real-world networks. They may participate in different interactions- variably according to different contexts; modeling such complex information into one single node/embedding is insufficient [1]. Thus, we propose contextualized multi-embeddings, where we compute for each target node, multiple embeddings corresponding to its contextualized interactions. Figure 2 shows learning under different contexts, Running and Football, where target nodes are splitted, organized and retrieved for more contextualized recommendation. Two questions we are trying to answer in our work: 1) what defines proper context; 2) how to utilize the context aspect in real-life networks for effective downstream applications. Later, this work will be extended to evolving graphs, where the objective is to smoothly learn dynamic representations of nodes with minimal necessary updates over timestamps.

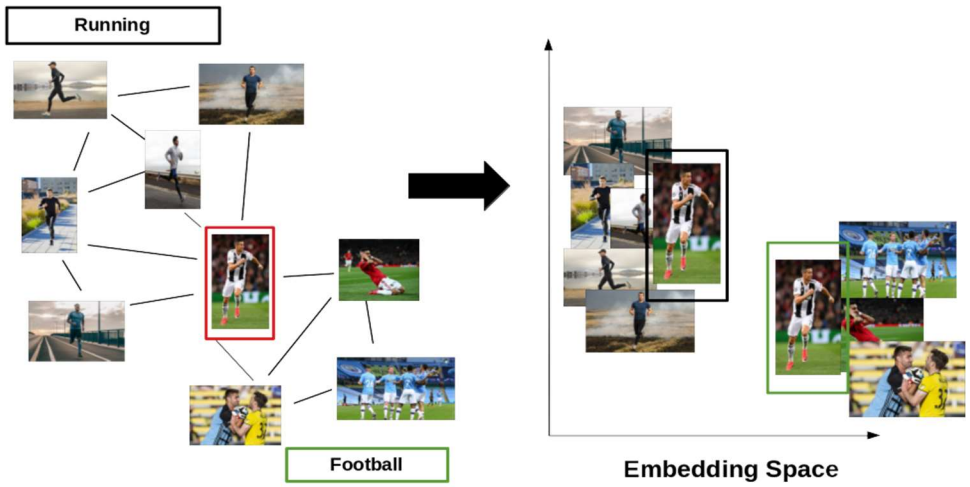


Figure 2: Contextualized Multi-Embedding

Another key component of my PhD will focus on designing distributed/decentralized GRL algorithms for IoT, where only access to partial information is assumed. The goal is to provide effective distributed AL/ML (a.k.a Machine Learning) services while preserving data privacy. Recent research [2] on distributed frameworks such as federated learning and decentralized (gossip) learning show promise to address learning under data privacy limitation. However, performing distributed learning on a network of IoT devices imposes a new set of constraints that ought to be considered. Examples of these constraints are when/if the nodes are communicating in synchronous or asynchronous fashion, as well as the presence of malicious behaviour and the dissemination of fake information from nodes. Finally, it is important to also account not only for learning under churn and network failures, but also for network topologies as well as data type/distribution (e.g., non IID data) which is produced in such a network.

References:

[1] Kefato, Zekarias T., and Sarunas Girdzijauskas. "Graph neighborhood attentive pooling." *arXiv preprint arXiv:2001.10394* (2020).

[2] Kefato, Zekarias, and Sarunas Girdzijauskas. "Gossip and Attend: Context-Sensitive Graph Representation Learning." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020.

ESR 11 JOHN ABIED HATEM | UNIVERSITY OF CYPRUS (UCY) | CYPRUS

Next Generation Fitness Trackers

It is not a secret that physical activity is good for you. A plethora of studies present countless health benefits of physical activity manifesting even on the cellular level. Yet, if you are anything like me, there is a big chance that by the time you read this article, you have been sitting at your desk at work or home for a few hours now. If you are working, you probably have ten things to do in the next hour and you are not even thinking about getting up and taking a short walk to move some muscles. After a long hectic day, you probably want to relax, rest, and maybe watch something. To boost your motivation, you might be thinking, or even already bought a wearable device like a smartwatch which can measure your heart rate, calories you burn, etc. These devices are usually accompanied by elegant applications to keep track of your activities, set goals, and monitor your progress. The vast amounts of data about our health and physical activity may be motivating. However, for many, this motivation fades away, stresses of everyday life swoop in, and we forget all about our resolutions to get fit and stay fit. So, it seems that, even though this new tech is providing us with all these “cool” measurements, it has missed on measuring an essential aspect of physical activity. Habit!

Here is where my research comes in, and it can be summarized by this intriguing question: *How can we measure a construct as abstract and complex as physical activity habit from observational data?* Recently, several studies in the psychology discipline presented advancements in the habit construct within complex behaviors such as physical activity, and new conceptualizations of the three-way interaction between habit, intention, and physical activity. Based on these new insights, we aim to have an in-depth investigation of large observational datasets. The clues we are looking for in the data are behavioral attributes that have habitual nature. The next stage will be developing measures capable of quantifying physical activity habits from observational data. These measures, in turn, will be used to predict short and long-term exercise behavior and future retention of physical activity habits. Furthermore, amid the COVID-19 situation, being physically active is of vital importance. However, with new measures and lockdowns, many aspects of our daily lives have been affected, including our physical activity. Hence, we are examining how the disruptions caused by the pandemic affected physically active people like those who attend the gym.

In short, my work will provide measures to quantify habits from increasingly available observational data and will offer insights into how habits shape our behaviors.

ESR 12 VANGJUSH KOMINI | ROYAL INSTITUTE OF TECHNOLOGY (KTH) | SWEDEN

Using machine learning to get an objective measure of the athletes' capabilities in a specific sport

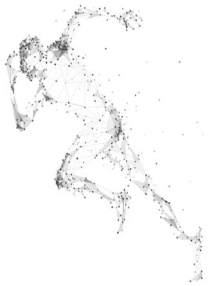
Artificial intelligence has been a significant contributor to personalized products by leveraging the generated data's potential. In healthcare and sports activities, in particular, it is of vital importance to have a more standardized and objective judgment. On the one hand, humans cannot agree with each other on the quality of a recorded sports activity. Even more importantly, a person cannot agree with himself when presented with the same recorded activity at two different time points. Machine learning (ML) application can mitigate this high variability and provide a more transparent evaluation of specific sport activity.

For an ML model to be considered a potential candidate at evaluating sports activities, it is essential to have good accuracy and a very calibrated response to ensure objectivity and the actual degree of belief for every provided output to ensure good transparency. Having these three key drivers in place enables a more user-friendly ML model, given that the evaluation of the recorded sports activity is much more reliable relative to the case of having a non-calibrated and non-transparent response.

Although ML models' accuracy has been accelerated significantly with the advent of deep learning (DL), the results' calibration and transparency require further efforts. Transparency is achieved by introducing a range of possible outputs instead of a single one, where each output is associated with its corresponding degree of belief. This degree of belief is measured using the language of probability by estimating the uncertainty for the given ML model. Furthermore, ML's uncertainty is not a trivial task, especially when the model possesses many parameters. Eventually, approximate computational models are the alternative candidate for enabling such computation through Bayes statistics. Uncertainty prevents the model from being over-confident and 'says that never seen something like this before' when an unfamiliar recording activity is presented as a test item.

A potential example is assessing the quality of short jogging from accelerometer recordings and provide a possible range of the score, such as the confidence is 95%.

Calibration of the output is another crucial indicator as it can make an assessment widely understood by all the ML models of different persons. In other words, a specific score for a particular recorded activity should represent the same scale of quality irrespectively of the user's profile. Calibration is the only tool available at reducing the subjectivity of the ML models. Aike uncertainty estimation, calibration demands intense computation, and a sufficiently large dataset to bridge the gap between the long-run empirical results and a specific ML model prediction.



ESR 13 STEFANOS EFSTATHIOU | ARISTOTLE UNIVERSITY OF THESSALONIKI (AUTH) | GREECE

'In-the-wild" User Experiments with Smartphones and Wearable Devices

Nowadays we are all connected with smart devices, especially smartphones, which have become an indispensable part of our everyday lives. Their user-friendly functionality has rapidly increased their daily usage, along with their relatively low cost, and the numerous mobile applications developed to meet users' needs. Beyond smartphones, another family of smart devices, wearables, has enjoyed rapid growth in usage in the last few years. According to Statista, the number of connected wearable devices worldwide is expected to grow to over 1.1 billion by 2022. One factor that has contributed to this development is the direct connectivity and compatibility of wearables with smartphones [1].

Smartphones and wearable devices with the embedded sensors, have enabled researchers to monitor participant's physical activities (running, walking), social interactions (physical and virtual interactions), and health/mental health states (stress, depression). Recent advances in sensor miniaturization and their integration into smartphones and wearable devices has made Ubiquitous Sensing a rapidly evolving research field. Ubiquitous Sensing refers to the ability to extract knowledge from sensor data [2], and the widespread use of smartphones and wearable devices is opening up new research possibilities, and new opportunities for real-world mobile sensing applications development [3].

A novel Ubiquitous Sensing, so called 'in-the-wild' research approach, has come to be widely adopted in studies under the Human-Computer Interaction (HCI) domain. 'In-the-wild' user experiments refer to ubiquitous and pervasive experimentation without any constraints, carried out in an unobtrusive manner [4]. This approach has been followed in various research fields including Psychology, Sociology and Health Sciences, integrating behavioral, social and health related theories with computer science [5].

My research within RAIS currently focuses on the possibilities, the advantages and the limitations of designing and conducting an 'in-the-wild' user experiment. In addition, we are identifying new methods on how to collect and analyze the user data coming from smartphones and wearables, taking into consideration the privacy and security concerns arising from these data.

References:

[1] S. Schlögl, J. Buricic, and M. Pycha, “Wearables in the Wild - Advocating Real-Life User Studies,” in Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI ’15, 2015, no. August 2015, pp. 966–969.

[2] O. D. Lara and M. A. Labrador, “A Survey on Human Activity Recognition using Wearable Sensors,” IEEE Commun. Surv. Tutorials, vol. 15, no. 3, pp. 1192–1209, 2013.

[3] A. Stisen et al., “Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition,” in Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems - SenSys ’15, 2015, pp. 127–140.

[4] A. Chamberlain, A. Crabtree, T. Rodden, M. Jones, and Y. Rogers, “Research in the Wild: Understanding ‘In the Wild’ Approaches to Design and Development,” in Proceedings of the Designing Interactive Systems Conference on - DIS ’12, 2012, p. 795.

[5] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, “A survey of mobile phone sensing,” IEEE Commun. Mag., vol. 48, no. 9, pp. 140–150, Sep. 2010.

ESR 14 SOFIA YFANTIDIOU | ARISTOTLE UNIVERSITY OF THESSALONIKI (AUTH) | GREECE

Is your Fitbit making you fitter? – A Scientific Perspective

Physical inactivity has been identified by the World Health Organization (WHO) as the fourth leading risk factor for global mortality [1]. Nevertheless, in today's sedentary society, fewer and fewer people are sufficiently active[1]. While technology has been blamed for contributing to this decline in physical activity, it also has the opportunity to help users take back ownership of their health through self-tracking technology.

Self-tracking refers to the use of ubiquitous technology, such as wearable devices, to collect, process, and display an abundance of personal data to help users monitor and manage their health. Wearable technology manufacturers claim that by monitoring themselves at all times, users can find their fit, and realize how small changes can have a big impact on their health. Based on this promise, users worldwide are embracing self-tracking to improve their health outcomes, making the wearables market a USD 27 billion industry[2]. However, self-tracking devices suffer from high attrition rates ranging from 30%[3] to more than 70%[4], which makes us, technologists, think twice about the effectiveness of wearable devices for health behavior change (HBC). At the end of the day, are wearable devices making users fitter? And if so, in which way, and how can we measure it?

The questions above are central in my work as an Early Stage Researcher in the RAIS consortium. Currently, my goal is to provide evidence-based answers to these questions through the study of related literature. Plenty of articles focus on utilizing self-tracking technology to encourage users to perform more physical activity. Most report favorable results. I aim to identify the components of the technological interventions that are the most beneficial for HBC. To do so, I am utilizing the Persuasive Systems Design Framework (PSD) [2], which describes what kind of content and software functionality may be found in a behavior change product, and I am devising an HBC-efficiency "score" for each software functionality based on past evidence. Furthermore, I am working towards the standardization of a user engagement evaluation framework for wearable devices, similar to existing frameworks for web and mobile apps. Given this complete framework for design and evaluation of self-tracking technology, a practitioner in the field can develop evidence-based interventions and products that untap the full potential of self-tracking technology for HBC.

[1] "Prevalence of insufficient physical activity - WHO." https://www.who.int/gho/ncd/risk_factors/physical_activity_text/en/. Accessed 26 Aug. 2020.
[2] "Smart Wearables Market To Double By 2022: $27 Billion ...." 23 Oct. 2018, https://www.forbes.com/sites/paullamkin/2018/10/23/smart-wearables-market-to-double-by-2022-27-billion-industry-forecast/. Accessed 26 Aug. 2020.
[3] "Inside Wearables Part 1: How behavior change unlocks long ...." https://medium.com/@endeavourprtnrs/inside-wearable-how-the-science-of-human-behavior-change-offers-the-secret-to-long-term-engagement-a15b3c7d4cf3. Accessed 26 Aug. 2020.
[4] "Deconstructing the Fitbit IPO and S-1 | Rock Health." 11 May. 2015, https://rockhealth.com/deconstructing-fitbit-s-1/. Accessed 26 Aug. 2020.

References:

[1] World Health Organization. 2019. Global action plan on physical activity 2018-2030: more active people for a healthier world. World Health Organization.

[2] Harri Oinas-Kukkonen and Marja Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. Communications of the Association for Information Systems 24, 1 (2009), 28.

BENEFICIARIES

PARTNERS

FOLLOW US

@H2020Rais @raish2020 Rais Horizon RAIS Project

WEBSITE

https://rais-itn.eu/

CONTACT US

Project Coordinator
Sarūnas Girdzijauskas
Computer Science Dept.
School of Electrical Engineering and Computer Science (EECS)
KTH - Royal Institute of Technology, Sweden
sarunasg@kth.se

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Innovative Training Networks (ITN) - RAIS No 813162