NEWSLETTER

OCTOBER 2021 | ISSUE 2

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

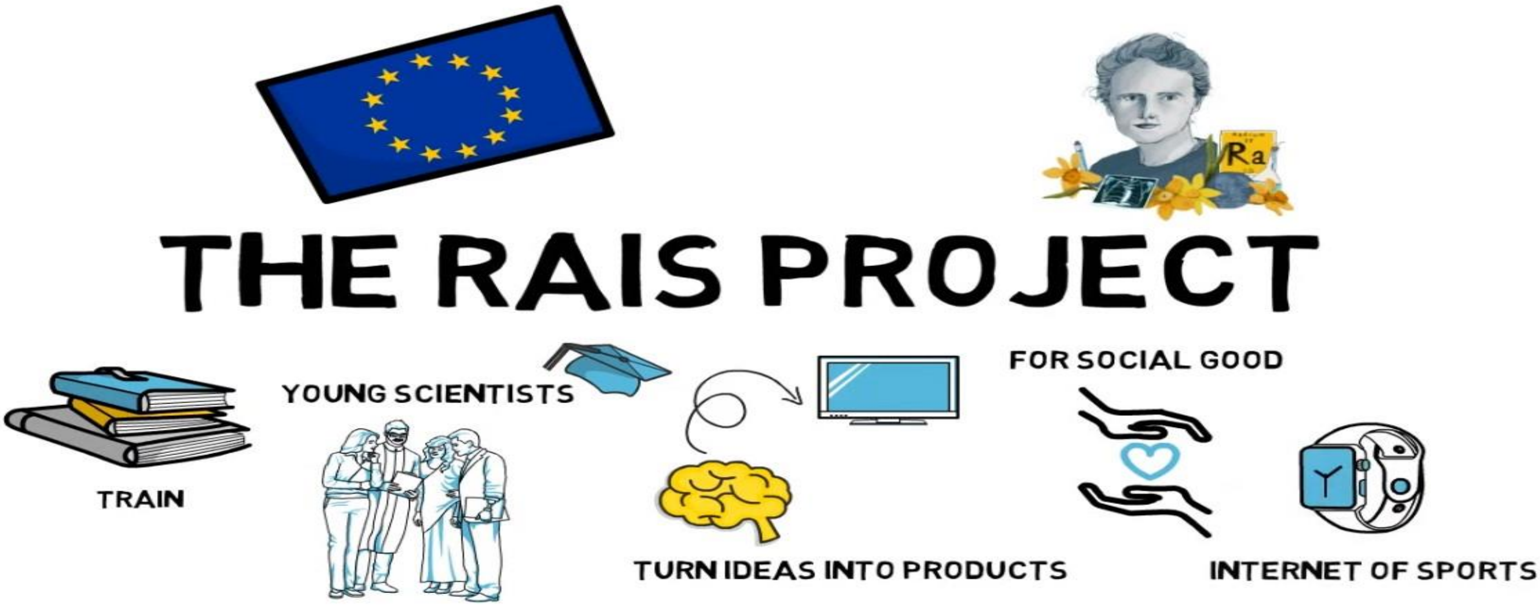
Decentralized Block-chain Powered IoT Platforms (generating hundreds of billions of transactions per day) for Big Data Mining

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.

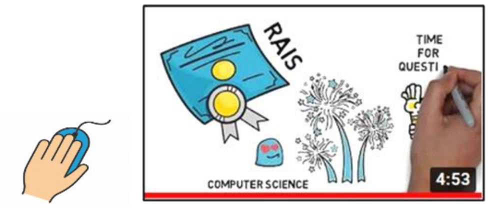
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Innovative Training Networks (ITN) - RAIS No 813162



CONTENTS

Latest News

Articles



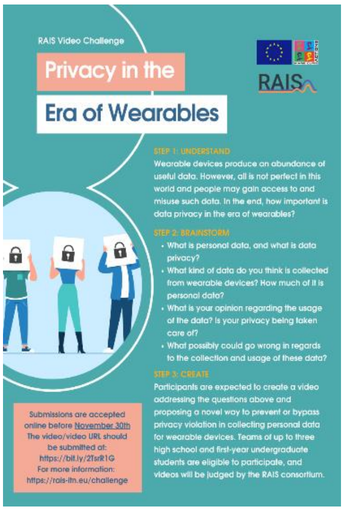
RAIS LATEST NEWS

https://rais-itn.eu/

RAIS CHALLENGE

RAIS Video Challenge
Deadline for Video Challenge submissions: 30/11/2021
RAIS Workshop: Summer School 2022 (Europe)

Many of you might already know about stories on how smart watches have saved people's lives. If you do not believe us google it, you will see how powerful a small connected device can be. Wearable devices deliver a vision of a healthy and connected world. These devices track the biometric and postural data of users to create optimal training plans. These guided sessions assist in a smart and healthy routine. Alas, all is not perfect in this world. Imagine your health app collects data regarding an irregular pattern of the heartbeat. Through some third-party app, an insurance company gets hold of this data. The consequences can be a denied insurance which is a severe unethical practice. Thus, the personal information that is collected in ubiquitous technology has great importance.



RAIS Video Challenge

Privacy in the Era of Wearables

RAIS

STEP 1: UNDERSTAND

Wearable devices produce an abundance of useful data. However, all is not perfect in this world and people may gain access to and misuse such data. In the end, how important is data privacy in the era of wearables?

STEP 2: BRAINSTORM

- What is personal data, and what is data privacy?
- What kind of data do you think is collected from wearable devices? How much of it is personal data?
- What is your opinion regarding the usage of the data? Is your privacy being taken care of?
- What possibly could go wrong in regards to the collection and usage of these data?

STEP 3: CREATE

Participants are expected to create a video addressing the questions above and proposing a novel way to prevent or bypass privacy violation in collecting personal data for wearable devices. Teams of up to three high school and first-year undergraduate students are eligible to participate, and videos will be judged by the RAIS consortium.

Submissions are accepted online before November 30th
The video/video URL should be submitted at:
https://bit.ly/2TsrR1G
For more information:
https://rais-itn.eu/challenge

Description of the Challenge

The RAIS Video Challenge provides high school and first-year undergraduate students in Europe with the opportunity to deliver a 2-minutes video in YouTube where the theme is:

"Privacy in the Era of Wearables"

The participants are expected to propose a novel way (new features or services) to prevent or bypass privacy violation in collecting personal data for wearable devices. Teams of up to three students are eligible to participate, and videos will be judged by the RAIS consortium based on the quality of the answers to the four questions below. As inspiration, many of you use wearable devices in the forms of watches, smart-bands, belts, etc. All of these devices produce data that are sent back to different services that you use through them. This challenge helps you to think about this aspect of wearable technology.

Directions

1. What is personal data, and what is data privacy?
2. What kind of data do you think is collected from wearable devices? How much of it is personal data?
3. What is your opinion regarding the usage of the data? Is your privacy being taken care of?
4. What possibly could go wrong in regards to the collection and usage of these data?

Judging Criteria

1. Novelty of the proposed solution.
2. Scientific accuracy of the solution.
3. Quality of the presentation.
4. Video delivery.

The winners will be invited to present their idea at the RAIS Summer School 2022. Their expenses (air tickets/accommodation) will be covered by the project. Submissions are accepted online till November 30th, 2021.

To upload your video, click the button SUBMIT

RAIS SUMMER SCHOOL, ENTREPRENEURIAL EVENT, WORKSHOP

Heraklion, Crete, September 05-10, 2021

FORTH hosted a successful in-person and virtual **triptych event** in Heraklion, Greece from 05-10 September and included the following:

I. SUMMER SCHOOL 2021 in "Edge computing and Blockchain"

Professor Sarunas Girdzijauskas (KTH Royal Institute of Technology), project coordinator of the RAIS project took the pride in welcoming all the attendees of the Event.

The Agenda of the RAIS Summer School covered interesting **Keynote** topics by admirable speakers, starting the first day with an 'Introduction to Blockchain and Cryptocurrencies' by professor **Bart Preneel** (KU Leuven). **Costas Chalkias** (Facebook) presented the keynote speech "I lost my blockchain private key, now what?". During the third day of the RAIS Summer School Professor **Schahram Dustdar** (Technological University of Vienna) gave a keynote speech on "Edge Intelligence – Engineering the New Fabric of IoT, Edge, and Cloud". **C. Mohan** (IBM) gave a keynote speech on "State of Permissionless and Permissioned Blockchains: Myths and Reality" and finally, **Aggelos Kiayias** (University of Edinburgh) delivered remotely the keynote speech 'Rethinking Information Technology Services as Incentive Driven Collaborative Systems'.

Maarten Gijssel, founder of Kinetic Analysis BV, gave a **speech** on "3D Analytics for Human Motion Data".

rais.h2020
@H2020Rais

#RAIS Fellows Poster Session

The **Tutorials** were presented by Dr. Demetris Trihinas (University of Nicosia), "Demystifying Fog Computing: Large-Scale and Repeatable Experimentation via Emulation" and Dr Spyros Voulgaris (Athens University of Economics and Business), "Demystifying Blockchains: An Algorithmic Approach".

During the Summer School **Poster Session**, RAIS Fellows had the opportunity to present their posters, share their research and interact with the attendees.

II. ENTREPRENEURIAL Event

The Entrepreneurial event was successfully held in four sessions from the 05th – 09th September with the outstanding performance of the speaker Tobias Vahlne (KTH Royal Institute of Technology) who covered the topics, i) Define your Concept, ii) How to pitch your idea, iii) Understand your customer and iv) The entrepreneurial journey.

Tobias Vahle (KTH Institute of Technology) during the Entrepreneurial Sessions

III. WORKSHOP on "Security, Privacy and Trust for Wearable Devices"

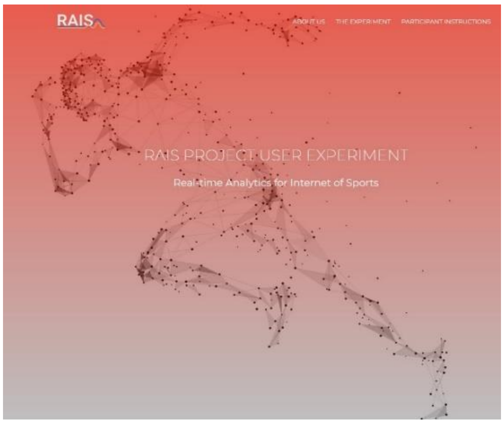
During the Workshop Event held on the 09th & 10th of September our experts gave the following concise talks:

i. WST1 - 'Security of 4G and 5G cellular networks' by Prof. **Elisa Bertino** (Purdue University)
ii. WST2 **-** 'Safeguarding against Information Exposure from Consumer IoT Devices' by **Hamed Haddadi** (Imperial College)
iii. WST3 - 'Characterizing abhorrent misinformative and mistargeted content on YouTube' by **Michael Sirivianos** (Cyprus University of Technology)
iv. WST4 - 'Side and Covert Channels: the Dr. Jekyll and Mr Hyde of Modern Technologies' by **Mauro Conti** (University of Padova)
v. WST5 - 'Automated cybersecurity for Internet-connected Things' by **Shahid Raza** (RISE Research Institutes of Sweden)
vi. WST6 - 'Sense & Sensibility in Sports: Personal & Interdependent Wearables that Work' by **Arthur van der Wees** (Arthur's Legal B.V.)

RAIS participants during RAIS Summer School, Entrepreneurial Event & Workshop in Heraklion, Greece

RAIS CONSORTIUM EXPERIMENT

The RAIS Consortium Experiment was designed between October 2020-April 2021 by RAIS fellows employed by the Aristotle University of Thessaloniki in collaboration with the consortium's supervisory board. It got approved by the Ethics Committee of the Aristotle University of Thessaloniki (Protocol No. 43661/2021). The first round of the experiment was conducted between May and July 2021 with the help of the RAIS fellows.


RAIS
ABOUT US
THE EXPERIMENT
PARTICIPANT INSTRUCTIONS
RAIS PROJECT USER EXPERIMENT
Real-time Analytics for Internet of Sports

During the course of the first round of the RAIS experiment, we collected multi-modal data from 41 participants, including Fitbit data, Ecological Momentary Assessment (EMA) responses, and various survey responses. Specifically, for two months, the participants wore a Fitbit Sense smartwatch throughout the day. As a result, we collected data related to physical activity, sleep, user characteristics, dietary preferences, account management and application usage, menstrual health, oxygen variation, social interactions, stress and mindfulness activity. Some of these data types provide measurements with minute granularity leading to millions of rows of objectively measured self-tracking data. Additionally, the participants utilized SEMA3 EMA app to complete short daily questionnaires regarding their physical activity goals, their emotions and the context they were currently in, leading to more than 10000 unique responses. Finally, participants completed nearly 700 questionnaires related to their physical activity habits, weekly stress and emotion levels, personality and demographics.

The collected dataset will not only be a valuable resource for the research work of the RAIS fellows, but ultimately, we intend to share them publicly in an anonymized format to help advance mHealth research.

RAIS SUMMER SCHOOL

Nicosia, May 10-12 & 17-19, 2021

The Laboratory for Internet Computing (LInC) at the University of Cyprus hosted virtually the

First RAIS SUMMER SCHOOL 2021

in

'IoT Analytics: Research and Innovation'

The **Agenda** included the following keynotes and tutorials:

Keynotes:

I. 'Distance-based Connectors and Community Search', Francesco Bonchi (ISI Foundation & ISI Global Science Foundation)
II. 'Why Online Services Should Pay You for Your Data?', Nikos Laoutaris (IMDEA Networks Institute)
III. 'Social Data Analysis', Ioannis Katakis (University of Nicosia)
IV. 'Workplace Productivity and Well-being', Marios Constantinides (Nokia Bell Labs)
V. 'Sports Analytics', Dimitrios Karlis (Athens University of Economics and Business)
VI. 'Improving Life Quality with Human Motion Data', Maarten Gijssel (Kinetic Analysis BV)

Tutorials:

I. 'Taming Big Data Streams: Real-time Data Processing at Scale', Asterios Katsifodimos (Delft University of Technology)
II. 'Exploratory Analysis of User Data - User Data Mining and Recommendation', Behrooz Omidvar-Tehrani (Grenoble AI Institute)

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

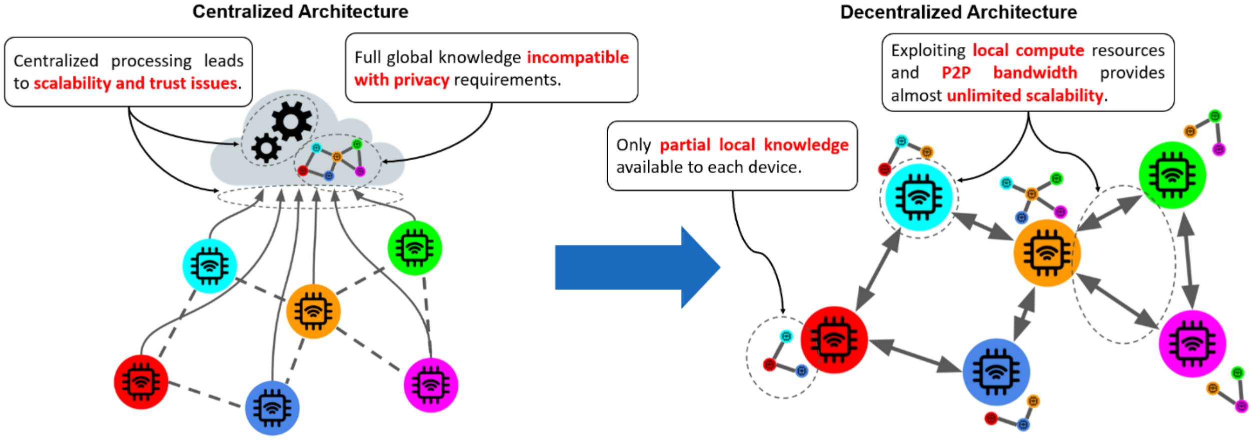
Decentralized Machine Learning over Networks

A quest to extract deep insights from interlinked data in a scalable and privacy-preserving way.

The Big Data Analytics world is constantly evolving. One recent trend is the rise of Graph Representation Learning (GRL), a family of techniques that allow the use of machine learning to extract insights from data that are interlinked to each other. Examples of these kind of links are friendships on social media, or connections between Internet devices. Their analysis is fundamental to obtain deep, actionable insights.

Another trend is the increased diffusion of smart and IoT devices, which collect more and more diverse data on every aspect of our lives. This extreme level of data collection is part of the reason behind a third trend: an increased awareness by users and regulators about the dangers of data collection, which leads to stricter privacy requirements.

The left-hand side of the figure below shows the traditional, centralized approach to train machine learning models. A central cloud-based service orchestrates the training process on all the data collected by a vast network of IoT devices. This causes several problems. First, as all devices connect to this central service, it can become a bottleneck for the training process, and can even stop the process completely in case of failure. Second, as the central service controls the process entirely, it must be trusted by all participants. Third, the central service has a global view of all the data and is therefore incompatible with strict privacy requirements.



Centralized Architecture
Centralized processing leads to scalability and trust issues.
Full global knowledge incompatible with privacy requirements.
Decentralized Architecture
Only partial local knowledge available to each device.
Exploiting local compute resources and P2P bandwidth provides almost unlimited scalability.

Our vision is to move from this kind of system to the one shown on the right-hand side of the picture. The central service, which is the root cause of all the issues presented above, is gone, replaced by a decentralized machine learning process. The IoT devices coordinate with each other in a peer-to-peer, trust-less manner, as no single device has broader control on the process. By training directly on the devices and directly exchanging training parameters, we exploit the available resources of the devices and can therefore scale without bottlenecks. Finally, each device can only access its local data and only knows about its local connections, therefore protecting the privacy of the data and of the users.

However, to realize this vision, many technical advancements are needed, on three different areas. We are working in parallel on each of these aspects and achieving promising results, but much more work is ahead of us.

The first area of interest is that of Gossip Learning. Gossip Learning is a technique developed to train machine learning models in a decentralized and privacy-preserving way, with each data point being accessible only to the device that collected it. However, this technique has received very little attention and is thus not well understood yet. In our work we have shown that it can be used effectively in many real-world scenarios [1] and we have started to integrate it as a key component in a larger blockchain-based machine learning framework that aims at democratizing data access and data monetization [2]. However, further research is hampered by the challenge of performing large-scale simulations of Gossip Learning. We therefore plan to develop a framework to simply this kind of decentralized simulations.

The second area of interest is that of Graph Representation Learning. While this topic is receiving a lot of attention recently, there is still much untapped potential, especially when it comes to specific applications that are relevant to the RAIS consortium. In our previous research we have developed a GRL technique to detect malware as soon as it penetrates in an IoT network [3], an important challenge given the diffusion of IoT devices. We are now extending this technique to different use-cases, such as the detection of fraud in cryptocurrency transactions.

The third aspect of our work consists of combining decentralized learning and GRL, to achieve our vision of machine learning on decentralized networks. We already started working on this aspect as well, with current experiments focusing on a subset of GRL techniques that lack global trainable parameters. We plan to expand our work to include those techniques that require global parameters, using Gossip Learning to synchronize them and potentially exploiting recent advancements, such as self-supervised learning, to offset the lack of long-range data caused by our privacy requirements.

Overall, we believe that the machine learning community has a great opportunity, in the next few years, to make big data analytics more scalable, trustworthy and privacy aware. We hope our research will have a positive impact in this direction.

References:

[1] Giaretta L., Girdzijauskas S., Gossip Learning: Off the Beaten Path, 2019, in IEEE International Conference on Big Data (IEEE BigData 2019)

[2] Giaretta L., Savvidis I., Marchioro T., et al., PDS2: A User-centered Decentralized Marketplace for Privacy-preserving Data Processing, 2021, in IEEE 37° International Conference on Data Engineering Workshops (ICDEW 2021)

[3] Giaretta L., Lekssays A., et al., LiMNet: Early-Stage Detection of IoT Botnets with Lightweight Memory Networks, 2021, in European Symposium on Research in Computer Security (ESORICS 2021)

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

A GDPR Compliant Personal Data Storage Model for the Internet of Things

Aiming to protect personal data against unauthorized access from malicious services in the IoT environment, we develop a personal data storage (PDS) model, which manages and stores data based on user control (user-centric) via individual privacy preferences, to be compliant with the GDPR principles.

Nowadays, the Internet of things (IoT) plays an indispensable role in our lives. We can list many areas in which we can exploit 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 more intelligent, with higher capabilities, and better at interacting. However, there are still some fundamental limitations that are affecting this development wase, one of the most critical ones is related to security and privacy issues.

General Data Protection Regulation (GDPR)[1] has been introduced in 2018 for European countries. GDPR states fundamental requirements for equitable treatment of the service provider (including third parties) and users. GDPR addresses the increasing privacy concerns by giving more rights to users. For instance, service providers must notify end users of the collection/processing of data, such as the kind of treated data (e.g., circumstances, purposes, amount of collected data, and retention period). The main pillars of GDPR are the rights that must be granted to end users whose data is being processed, being the main ones: the right to be informed, the right to access, the right to rectification, the right to erasure, the right to restrict processing. GDPR calls for "*increased users involvement in protecting their data by enabling them to control what is collected about them, when, by whom and for what purposes*". In this respect, there must be a shift of privacy-preserving mechanisms from focusing on service providers (called system-centric protection) to end users (called user-centric protection).

My research focuses on this shift in the challenging domain of IoT. We are designing and developing a personal data storage (PDS) model for the IoT environment, whose general architecture is shown in Figure 1.

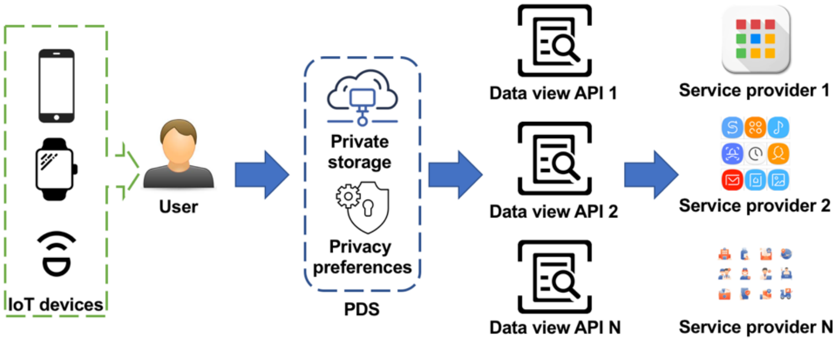


Figure 1: Personal data storage

According to this model, user personal information collected by several IoT devices, such as wearable devices (e.g., smartphones, smartwatches), Internet-connected devices (e.g., temperature sensors, CCTV). Some of these devices may also be not owned by the user, i.e., the public devices (e.g., CCTV). The privacy data storage (PDS) model includes three main processes: storage, privacy compliance check, and data view generation. The first process stores all the IoT-generated data (those generated by both private and public devices) in the individual private storage (it can be either in a private cloud, or in a pubic cloud in encrypted format). Thanks to privacy preferences, the users are able to decide how their data must be managed (e.g., share data with selected providers for selected purposes, or revoke/grant permissions to service providers). The service providers only collect user data if they can fulfil all requirements stated by individual privacy preferences. The second process compares user privacy preferences to service provider's privacy policies. If the service provider can pass the evaluation process, PDS generates a data view corresponding to the data requests and the privacy constraints stated by privacy preferences.

References:

[1] General Data Protection Regulation (GDPR): https://gdpr-info.eu/

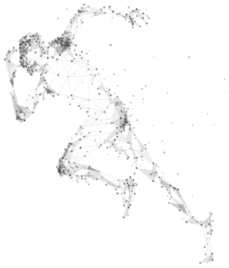
ESR 5 ANDREI KAZLOUSKI | FOUNDATION FOR RESEARCH AND TECHNOLOGY – HELLAS (FORTH) | GREECE

User Profiling Using Fitness Trackers' Data

The global pandemic put a strain on our regular lifestyles and activities. With the continuous Covid restrictions people tend to spend less time exercising, commuting to work, and even walking. To compensate lack of activity, the affected population might reserve to the wearable fitness trackers, and the health benefits they provide. Such smart devices continuously monitor, and record a wide variety of fitness data. At present people are freely sharing their fitness results online in hopes of motivating themselves to be more active and dynamic. The posted data typically include daily number of steps, calories, distance, elevation, etc. This information, however, might be more revealing than it appears at first glance. Since many of the parameters that are tracked by smartbands are calculated from some combinations of gender, weight, height, and age, severe statistical leaks might occur. Currently we are studying those leaks, and we are analyzing whether it is possible to establish any of the above personal characteristics with good precision.

We gathered and processed several open-source Fitbit fitness datasets, with the aim to verify whether some of the information about the users of Fitbit devices can be inferred from the daily activity data. By utilizing various state-of-the-art machine learning models, we were able to partially infer sensitive information of users. Our findings suggest that at least gender and Body Mass Index (BMI) can be learned from the fitness Fitbit data. We also verified that Fitbit uses a close approximation of Harris-Benedict formula for basal calories.

Currently we are planning to extend our studies to the real-world Fitbit data. We found several online communities were people post their daily fitness progress and statistics. We are looking into the ways to crawl, preprocess, and analyze these data. We are also investigating the ways to obfuscate the data to not reveal private information. For example, one of the ways would be to "statistically masquerade" the data as the opposite gender/BMI.



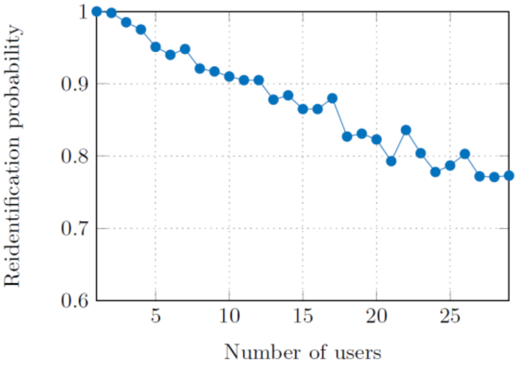
ESR 6 THOMAS MARCHIORO | FOUNDATION FOR RESEARCH AND TECHNOLOGY – HELLAS (FORTH) | GREECE

Mitigating Risks Associated to Public Disclosure of Daily Fitness Records

Wearable fitness trackers have been continuously increasing their popularity over the past decade. In 2021, it is easy to find a smartwatch or smart band meeting the requirements of all kinds of fitness enthusiasts, from casual joggers to hardcore athletes.

In order to increment users' engagement, many manufacturers give them the possibility to share their daily progress records either with friends or by posting them on social media. These records usually include number of steps, covered distance, burned calories, and occasionally climbed floors and average heart rate. Fitbit has also a very large community, counting over 2.8 million users, where people can post their daily activity to get feedback from others.

Concerns about the public disclosure of fitness data exist since the emergence of wearables, and many studies have expressed concern about the possibility of creating a "quantified identity" for smartband users. Our research over the past few months consisted in assessing how easily one can identify a user by collecting daily records for a given amount of time, testing similarity-based linking techniques on public datasets. We concluded that, even by collecting one month's worth of steps and calories records, it is possible to effectively reidentify one user in a crowd of over 25 other users, obtaining a probability of correct identification around 80% [1].



The next step would consist in developing privacy preserving techniques enabling users to disclose "noisy" fitness data that maintain high utility while reducing the identifying information that is carried.

Our previous results suggest that a good strategy consists in trying to anonymize individual samples rather than trying to act on a whole time series of fitness records.

Moreover, attention should be given to the correlation between steps and burned calories, as the latter ones are computed according to a combination of user's characteristics such as gender, age, height, and weight and are also proportional to the number of steps.

References:

[1] Marchioro T., Kazlouski A. and Markatos E. (2021). User Identification from Time Series of Fitness Data. In Proceedings of the 18th International Conference on Security and Cryptography - Volume 1: SECRYPT, ISBN 978-989-758-524-1, pages 806-811. DOI: 10.5220/0010585008060811

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

Blockchain-based Malware Detection, Containment and Recovery in IoT

Towards Securing IoT Devices from Emerging Malware Threats

Malware is an emerging threat to people's privacy and security. In the first quarter of 2021, Kaspersky detected more than 2 billion attacks targeting more than 100k users around the world[1]. In addition, according to Statista, 37% of organizations worldwide were hit by a ransomware attack in 2021[2]. Moreover, in 2016, Mirai botnet targeted 600k IoT devices to launch one of the most significant Distributed Denial of Service attacks in history after reaching a bandwidth of 1 Tbps that stopped Dyn, a DNS service provider[3]. This attack interrupted famous services such as GitHub, which was unavailable for hours in some regions like North America and Europe. Furthermore, due to their low protection, IoT devices have become an interesting target for malware attackers. This scenario is further exacerbated by the exponential growth of IoT adoption, increased from 13.4 billion in 2015 to 38.5 billion in 2020[4]

Our research aims to protect IoT devices and help organizations recover from malware attacks by following NIST SP 800-83 Malware Incident Response guidelines[5]. This guideline consists of four steps: preparation, detection and analysis, containment and eradication, and recovery. Preparation is about raising awareness about such threats and imposing security best practices. For detection and analysis, the main goal is to identify the malware and extract its metadata such as its family, its propagation scheme, its malicious actions, etc. To this end, we have worked on two projects, namely PAutoBotCatcher and LiMNet. The first aims to collaboratively detect botnets leveraging community behavior analysis and blockchain to address trust among devices. The main contribution of PAutoBotCatcher is that behavior analysis is done by protecting the privacy of devices owners. Indeed, we defined an anonymization strategy to hide devices' real communications, by exploiting both IP addresses randomization and graph anonymization. LiMNet focuses on analyzing botnets' behavior in IoT networks by leveraging Lightweight Memory Networks with an internal memory component capable of capturing IoT devices' behaviors over time. The latter enables malware detection in IoT networks by analyzing the network packets exchanged not only among devices but also between devices and external entities since it is designed to classify malicious devices and malicious network packets.

For containment and eradication, it aims at stopping the propagation of malware to protect other devices in the network from infection. In addition, it focuses on removing malware from the infected devices. To this end, we have worked on MalCon: a blockchain-based malware containment framework for IoT. It contains malware in three phases, namely: emergency, healing, and punishment. Once a malware attack is detected in the emergency phase, MalCon protects devices by suggesting a set of actions to be executed by all devices to stop the propagation of the malware by, for instance, closing the ports that a malware uses for propagation. In the healing phase, MalCon recommends a set of actions to the infected device to remove the malware and be back to normal operation. Moreover, MalCon provides strategies to punish devices that acted dishonestly, i.e., did not implement the suggested actions. The final NIST guideline step is the recovery, this aims at helping devices to be back to the normal operation after a malware attack. Regarding this step, we have worked on MalRec project: a blockchain-based malware recovery framework. This aims to make devices able to recover their files through continuous backups. It leverages blockchain to store files' metadata immutably and safely and IPFS, a distributed filesystem, to store encrypted files.

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.

References:

[1] https://securelist.com/it-threat-evolution-q1-2021-non-mobile-statistics/102425/

[2] https://www.statista.com/statistics/1246438/ransomware-attacks-by-country/

[3] https://blog.cloudflare.com/inside-mirai-the-infamous-iot-botnet-a-retrospective-analysis/

[4] https://www.juniperresearch.com/press/press-releases/iot-connected-devices-to-triple-to-38-bn-by-2020

[5] Souppaya, Murugiah, and Karen Scarfone. Guide to Malware Incident Prevention and Handling. Gaithersburg, MD: U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology, 2013. Print.

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

Variations on Role Discovery

Graphs are an extremely versatile data structure that can be used to embed a wide variety of phenomena, such as relationships between people and chemical bonds between atoms.

Insights into such phenomena can be extracted not only from the individual properties of the entities represented, but also by studying the structure of the graph modeling the phenomenon under analysis.

There are several tools that can be used to do this. Among graph clustering algorithms, there are methods such as RolX [2], struc2vec [4], and GraphWave [5] that are used for role discovery, that is, to detect the structurally similar nodes in a graph.

Since they are sensitive to the degree (the number of neighbors) of the nodes, such methods can successfully be applied in several settings; however, they struggle to recognize nodes that are not equal in degree as structurally similar.

We designed a role discovery algorithm that is able to recognize structurally similar nodes, even if not equal in degree, which are common to many networks observed in real life, such as organizational graphs and social networks. As shown in Figure 1, this is done by utilizing Dynamic Time Warping (DTW) [1] from Time Series Analysis to compute the distance between the patterns that summarize how the nodes diffuse and thus how they are connected to each other.

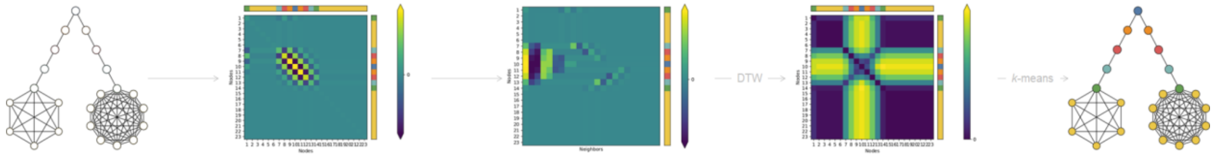


Figure 1. Here, the algorithm presented in the research paper is applied to an asymmetric barbell graph. Roles are color-coded.

Compared to node2vec [3], RolX [2], struc2vec [4], and GraphWave [5], ours gives a performance that is up to 24.6% better in F1 score for graphs whose structurally similar nodes differ in degree.

Future Work

Next, I am going to explore the following directions, which are related to role discovery and thus to the previous research paper.The first direction I am going to investigate is role discovery performance evaluation. In general, role discovery algorithms are compared by computing accuracy and F1 score, which require a ground truth. However, the true roles of the nodes in a graph are not always known. While there is a number of unsupervised metrics specific to communities, general clustering performance measures such as the silhouette score are utilized for roles in absence of ground truth. Thus, I am going to research the metrics that could be applied to role discovery in order to determine which are more reflective of the true roles, which ones are more resistant to noise, etc.

The second direction I am going to explore is at the intersection of Deep Learning with graphs and role discovery. In particular, I am going to investigate how the role memberships could be utilized to "shortcut" the flow of information that Graph Neural Networks (GNNs) aggregate in order to update the node representations.

References:

[1] Joseph B. Kruskal and Mark Liberman. The Symmetric Time-Warping Problem: From Continuous To Discrete. In David Sankoff and Joseph B. Kruskal, editors, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, chapter 4. Addison-Wesley Publishing Company, 1983.

[2] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. RolX: Structural Role Extraction & Mining in Large Graphs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1231–1239, 2012.

[3] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[4] Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. struc2vec: Learning Node Representations from Structural Identity. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 385–394, 2017.

[5] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. Learning Structural Node Embeddings via Diffusion Wavelets. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1320–1329, 2018.

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

Human Activity Recognition with Calibrated Confidences

Human activity recognition (HAR) is a popular research area of ubiquitous computing with wide applicability in health and well-being, sports, and other domains. In line with our earlier research, we have focused particularly on ``human activity recognition with wearable sensors''.

The present state of the art in HAR with wearable sensors is addressed by deep learning methods and techniques. The pipeline is generally similar to that of a time-series classification problem. The deep learning algorithms of choice are usually based on convolutional networks, recurrent neural networks, and their advanced variants. In our work, we explored these methods and found a drawback, that directly affects the reliability of the predictions coming out of those models. In particular, we observed that the predictive probability distributions of those deterministic neural networks do not represent the true likelihood. Hence, they are unreliable to be included in any production system.

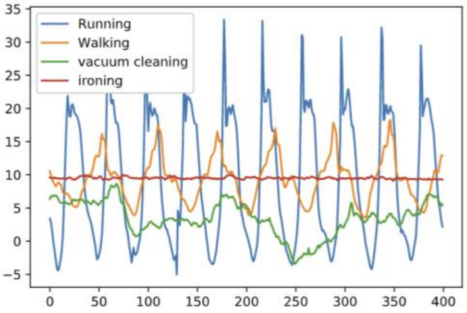


Figure 1: Comparsion among different activites w.r.t to accelerometer data (y-axis) and time(x-axis)

In our work, we address the problem by proposing Deep time-ensembles, a novel ensembling method capable of producing calibrated confidence estimates from neural network architectures. The proposition was propelled by two primary observations. Firstly, we observed that the existing deep learning models use a fixed time window to extract temporal sequence from the sensor data. This is sub-optimal because each dataset contains multiple activities, that are sensitive to different time windows. For example, running and vacuum cleaning activity might require different time windows to be recognized (depicted in Figure 1). Similarly, other activities such as ironing or walking might require another time window. Thus, a one size fits all concept may be replaced with a better solution. Hence, we decided to use multiple time windows to extract temporal sequences from the sensor data and use all those temporal sequences to train multiple neural networks. Finally, the averaged output of the ensemble is used as a classification result. Secondly, this method also helps us generate well-calibrated predictions compared to deterministic neural networks. This is achieved through the reduction of variance because of the ensembling procedure. Furthermore, overconfident estimates produced by individual models get smoothed out through the averaging procedure.

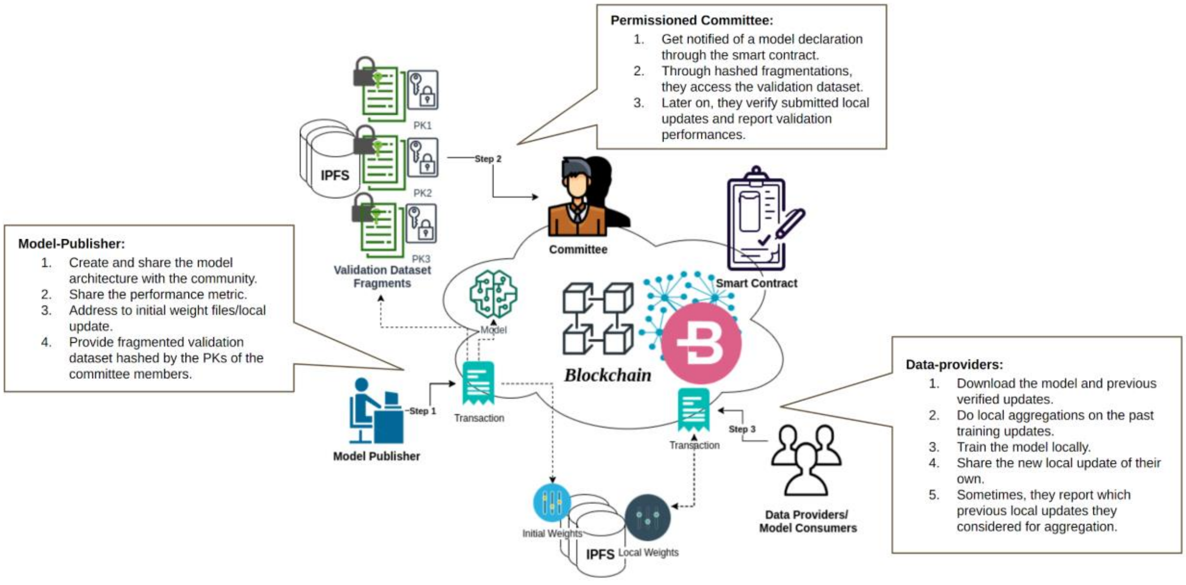
We have tested the method on popular HAR datasets and obtained promising results, that not only calibrate the predictions but also improve the state-of-the-art. Our research has touched upon multiple areas of neural network models, prediction reliability, and estimation of uncertainty. Thus, it is a small step forward in making our prediction systems more reliable and safe.

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

Open-source Blockchain for Decentralized Learning

Most of the practical daily-life AI solutions which are offered by giant tech companies, such as Google or Facebook, are typically centralized and rely heavily on the huge amount of personal data acquired from the service consumers. With the urge of preserving user data privacy, these solutions may be in violation of such constraints. Thus, there is a need for alternative decentralized solutions that enable equivalent well-trained AI services without assuming global sharing of private information. Protecting the end-user's data is not the only constraint here. On the other side, AI model providers who can be individuals or early start-up companies, suffer from unfair competitions against big giant companies. Some challenges can be lacking either necessary computational resources and/or enough data to train their own models.

Blockchains have been growing popular in recent years as an infrastructure for building secure peer-to-peer transactional systems. Blockchain frameworks, whether permissioned or public, provide a safe environment for sharing data and regulating the interactions/interactions among all participants. Having all transactions authenticated and traced back, makes Blockchain a great fit to avoid many of the typical issues found in decentralized learning techniques. Modeling the decentralized learning process as a smart contract offers big gains such as robustness and resilience to malicious attacks that may hinder learning in the vanilla setups of decentralized or distributed learning otherwise.



Permissioned Committee:

1. Get notified of a model declaration through the smart contract.
2. Through hashed fragmentations, they access the validation dataset.
3. Later on, they verify submitted local updates and report validation performances.

Model-Publisher:

1. Create and share the model architecture with the community.
2. Share the performance metric.
3. Address to initial weight files/local update.
4. Provide fragmented validation dataset hashed by the PKs of the committee members.

Data-providers:

1. Download the model and previous verified updates.
2. Do local aggregations on the past training updates.
3. Train the model locally.
4. Share the new local update of their own.
5. Sometimes, they report which previous local updates they considered for aggregation.

Figure 1: **The Proposed System's Architecture**

Motivated by the aforementioned, our research aims to provide a safe open-source environment based on Hyperledger Fabric[1]that brings both model providers and model consumers together while maintaining the privacy of the end-user's personal data. As shown in figure.1, the training process of a model starts if the model-publisher should choose to create and share the architecture of the model to the community. The model-publisher is expected to provide the performance metric(s) of the model as well as a confidential validation dataset to check the progress of the training process. Once published, data-providers can download the model, train it on their local data, and share the weights back to the blockchain. Data-providers may report the local performance of the model too.

A direct application of the proposed system is a hub on which people can go to share and get desired off-shelf AI services; for instance, researchers from different research groups/universities, as well as interested industries can get on to co-train and leverage AI services without having to worry about lack of data, copyrights or communications among the interested parties. Other real-life applications based on these collaborative efforts among users are popular; an example of which is GitHub[2].

References:

[1] https://www.hyperledger.org/use/fabric

[2] https://github.com/

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

Map Embeddings in a Deep Neural Network for Out-of-Distribution Detection

Re-inventing the data processing

Deep (machine) learning (DL) has revolutionized quite substantially the way we utilize computers. Traditionally, the design of computers is particularly suited for state-space search within massive storage of data. Within this framework, the algorithms could run with minimal assumptions and heuristics. However, machine learning renders both search and storage obsolete while achieving more things than otherwise possible.

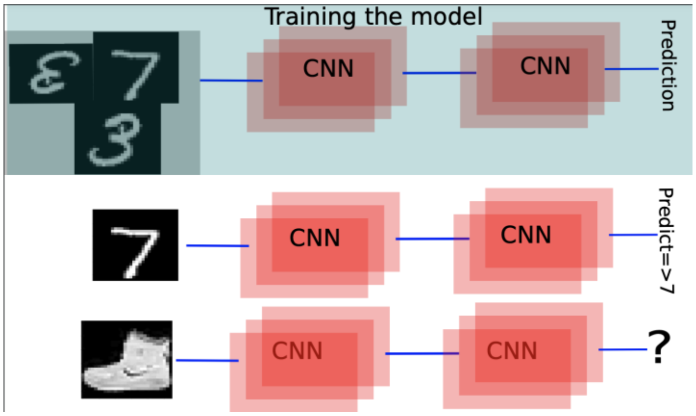
Out of distribution detection

Although DL possesses many potentials in data processing, a pressing lack of robustness impedes their broader daily life adoption.

Alternative to storage, DL generalizes over a massive amount of annotated training data.

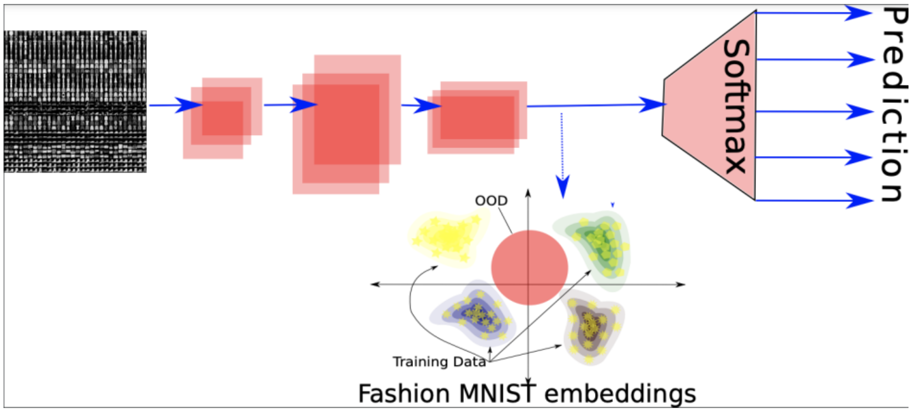
Whenever there is a test item that is contextually dissimilar to this data, the model cannot detect this as an out-of-distribution individual item.

One such example is training data in the MNIST dataset, and all of a sudden, a fashion-MNIST item appears. Although the within-pixel correlation of this dataset is quite similar, their contextual information is however quite different.



Integrating a roadmap in DL

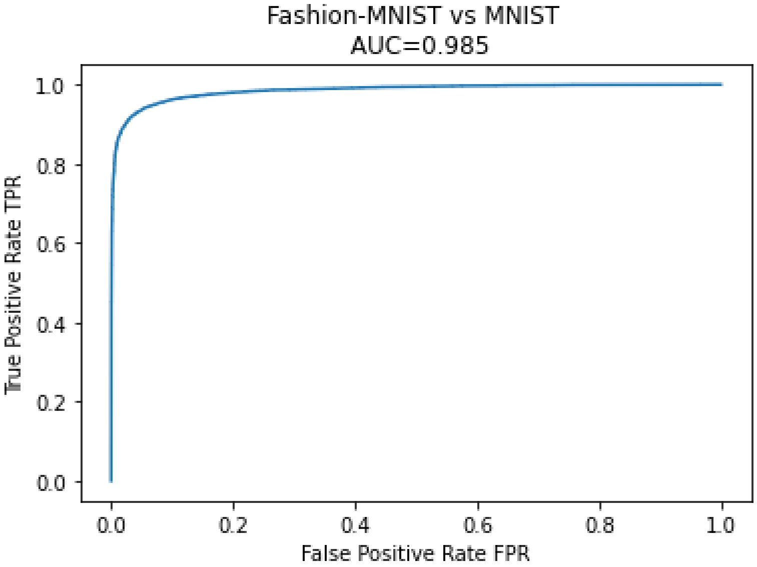
The embedding space tries to keep the similar items very close by and the dissimilar ones far from each other. Furthermore, the roadmap keeps maintains a likelihood score at regions proportional to the concentration of training data. Whenever an individual item falls outside the regions of low likelihood, it is contextually distinct to the training data (OOD). Since the embedding space is continuous over the training data, storing every possible likelihood at every coordinate is not scalable. As a result, a multinomial nonlinear Gaussian distribution (normalizing flow) offers an interpolation of the likelihood values over embedding space while mitigating the need to store the training data again.



Increasing the awareness of DL

This approach provides state-of-the-art performance at distinguishing MNIST from fashion-MNIST. At the same time, the latter never appeared during the training process.

Thus, whenever a fashion-MNIST dataset appears for prediction, the DL model should not express any opinion as it detects an OOD.



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

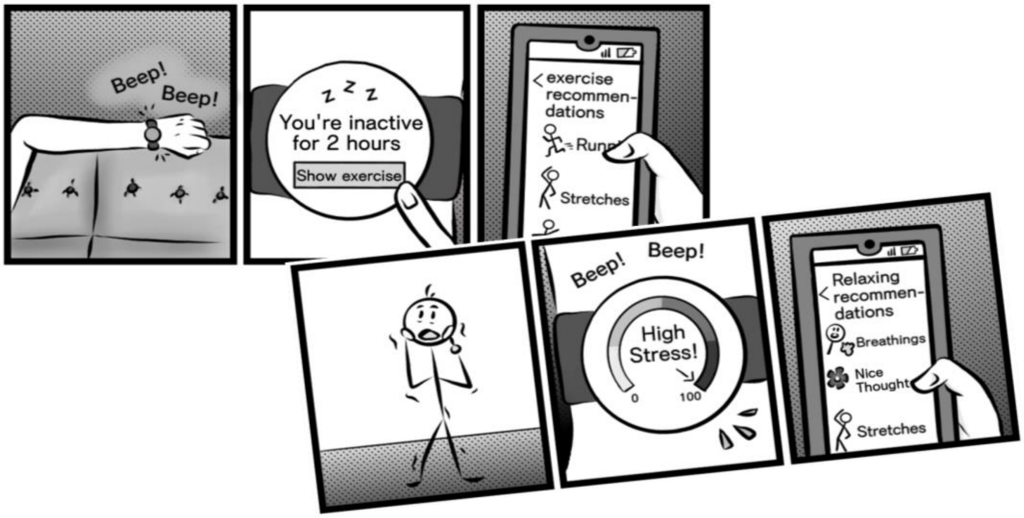
Negotiating with your Fitness Tracker: A Storyboards User Study

We adopt a combination of theories, aiming at translating theoretical principles into actionable insights for effective wearables HCI design.

People who perform regular physical activity enjoy multiple health benefits, such as improved muscular and cardio-respiratory fitness, reduced symptoms of depression and anxiety, and lower coronary heart disease rates and other non-communicable diseases (World Health Organization, 2018). However, globally, one in four adults do not meet the recommended physical activity levels despite its considerable advantages. At the same time, there is growing evidence that fitness trackers can be practical tools for motivating health behavior change, influencing people towards adopting a healthier and more active lifestyle (Orji et al., 2016; Yfantidou et al., 2021). To this end, more and more people are resorting to wearable technology to achieve their health and fitness goals (Vailshery, 2021).

Therefore, designing effective wearable technology has been a focus for many researchers over the past decade. Despite such growing interest, though, current wearables still suffer a significant limitation: many of their components are not theoretically backed (Aldenaini et al., 2020). In other words, their design is not based on any psychological or behavioral theory for supporting behavior change. For instance, one of the most prominent behavior change theories is Prochaska's Transtheoretical Model (Prochaska et al., 1997), which introduces five stages towards behavior change: precontemplation, contemplation, preparation for action, action, and maintenance. An issue with the stages of change model is that a person can relapse to a previous stage because of innumerable obstacles to behavior adoption, called constraints (Mimiaga et al., 2009). However, other theories, such as the Hierarchical Model of Leisure Constraints (Jackson et al., 1993), support that leisure participation "is dependent not on the absence of constraints but on negotiation through them". In simple words, humans develop strategies to negotiate with themselves to overcome constraints towards adopting their desired behavior, whether physical activity or smoking cessation.

In our work, we adopt a combination of the theories above, aiming at translating theoretical principles into actionable insights for wearables HCI design that will help researchers and practitioners create effective mHealth technologies for health behavior change. Specifically, based on a contemporary scale by Balaska et al. (2019), we convert a textual scale, measuring the effectiveness of behavioral negotiation strategies, into a visual one, measuring the effectiveness of our so-called "negotiation features", namely features dependent on known negotiation strategies. Our visual representations take the form of storyboards (see figure), which are easy to understand by diverse populations regardless of their technological literacy, age, education, or mental capacity (Lelie, 2005). Additionally, we conduct three preliminary user studies to evaluate the validity and reliability of our novel visual scale with successful results. Our initial findings indicate that while the effectiveness of our negotiation features varies, the usage of wearable devices can overall positively contribute to the behavior change journey through the successful application of negotiation strategies.



References:

Aldenaini, Noora, et al. "Trends in Persuasive Technologies for Physical Activity and Sedentary Behavior: A Systematic Review." Frontiers in Artificial Intelligence, vol. 3, 2020, doi:10.3389/frai.2020.00007.
Balaska, Panagiota, et al. "Exploring how recreational sport participants with different motivation levels use leisure negotiation strategies." European Academy of Management, EURAM, 2019.
"Global Action Plan on Physical Activity 2018–2030: More Active People for a Healthier World." World Health Organization, World Health Organization, 1 Jan. 1970, apps.who.int/iris/handle/10665/272722.
Jackson, Edgar L., et al. "Negotiation of Leisure Constraints." Leisure Sciences, vol. 15, no. 1, 1993, pp. 1–11., doi:10.1080/01490409309513182.
Lelie, Corrie Van Der. "The Value of Storyboards in the Product Design Process." Personal and Ubiquitous Computing, vol. 10, no. 2-3, 2005, pp. 159–162., doi:10.1007/s00779-005-0026-7.
Mimiaga, Matthew J., et al. "Individual Interventions." HIV Prevention, Academic Press, 22 Apr. 2009, www.sciencedirect.com/science/article/pii/B978012374235300008X.
Orji, Rita, and Karyn Moffatt. "Persuasive Technology for Health and Wellness: State-of-the-Art and Emerging Trends." Health Informatics Journal, vol. 24, no. 1, 2016, pp. 66–91., doi:10.1177/1460458216650979.
Prochaska, James O., and Wayne F. Velicer. "The Transtheoretical Model of Health Behavior Change." American Journal of Health Promotion, vol. 12, no. 1, 1997, pp. 38–48., doi:10.4278/0890-1171-12.1.38.
Vailshery, Lionel Sujay. "Wearables Sales Worldwide by Region 2015-2022." Statista, 22 Jan. 2021, www.statista.com/statistics/490231/wearable-devices-worldwide-by-region/.
Yfantidou, Sofia, et al. "Self-Tracking Technology for MHealth: A Systematic Review and the PAST SELF Framework." ArXiv.org, 23 Apr. 2021, arxiv.org/abs/2104.11483.

RAIS PUBLICATIONS

- **"PAutoBotCatcher: A Blockchain-based Privacy-preserving Botnet Detector for Internet of Things"**, Ahmed Lekssays, Luca Landa, Barbara Carminati, Elena Ferrari, Computer Networks, accepted for publication.

- **"LiMNet: Early-Stage Detection of IoT Botnets with Lightweight Memory Networks"**, Lodovico Giaretta, Ahmed Lekssays, Barbara Carminati, Elena Ferrari, Sarunas Girdzijauskas, Proc. of the European Symposium on Research in Computer Security (ESORICS 2021).

- **"A Risk Assessment Mechanism for Android Apps"**, Ha Xuan Son, Barbara Carminati, Elena Ferrari, Proc. of the IEEE International Conference on Smart Internet of Things (SmartIoT 2021).

- **"Decentralized Word2Vec Using Gossip Learning"**, Abdul Aziz Alkathiri, Lodovico Giaretta, Sarunas Girdzijauskas, Magnus Sahlgren, 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021), Reykjavik, Iceland, 2021.

- **"Self-Tracking Technology for mHealth: A Systematic Review and the PAST SELF Framework"**, Sofia Yfantidou, Pavlos Sermpezis, and Athena Vakali, arXiv, 2021.

- **"PDS2: A user-centered decentralized marketplace for privacy preserving data processing"**, Lodovico Giaretta, Ioannis Savvidis, Thomas Marchioro, Sarunas Girdzijauskas, George Pallis, Marios D. Dikaiakos, Evangelos Markatos, Third International Workshop on Blockchain and Data Management (BlockDM 2021), in conjunction with the 37th IEEE International Conference on Data Engineering (ICDE), Chania, Crete, Greece, 2021.

- **"Federated Word2Vec: Leveraging Federated Learning to Encourage Collaborative Representation Learning"**, Daniel Garcia Bernal, Lodovico Giaretta, Sarunas Girdzijauskas, Magnus Sahlgren, arXiv:2105.00831 [cs], 2021.

- **"Do partner apps offer the same level of privacy protection? The case of wearable applications**", Andrei Kazlouski, Thomas Marchioro, Harry Manifavas, Evangelos Markatos, In Proceedings of the 7th Workshop on Sensing Systems and Applications using Wrist Worn Smart Devices (WristSense), 2021.

- **"I still See You! Inferring Fitness Data from Encrypted Traffic of Wearables"**, Andrei Kazlouski, Thomas Marchioro, Harry Manifavas, Evangelos Markatos, 14th International Joint Conference on Biomedical Engineering Systems and Technologies HEALTHINF, Vienna, Austria, February, 2021.
- **"Do you know who is talking to your wearable smartband?"**, Andrei Kazlouski, Thomas Marchioro, Harry Manifavas, Evangelos Markatos, European Federation for Medical Informatics - Special Topic Conferences (EFMI-STC), 2020.
- **"Gossip Learning: Off the Beaten Path"**, Lodovico Giaretta and Sarunas Girdzijauskas, IEEE Big Data 2019, Los Angeles, CA, USA, 2019.

Find at: https://rais-itn.eu/publications

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

Newsletter Content Editor
Demetra Katziani
Computer Science Dept.
Laboratory for Internet Computing (LInC)
University of Cyprus (UCY)
dkatzi01@cs.ucy.ac.cy

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