# Identity Verification Based on Behavioral Characteristics

Author: Karolína Kuchyňová
Supervisor: prof. RNDr. Tomáš Skopal, Ph.D.

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

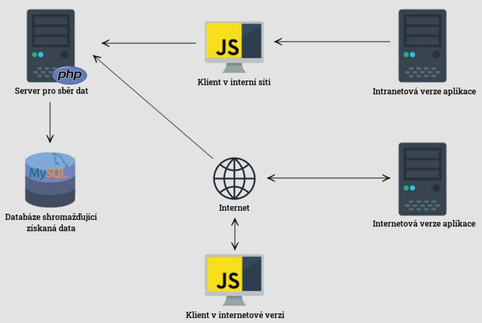
## Introduction

Verifying the identity of a user logged into a secure system is an important task in the field of information security. In addition to a password, it may be appropriate to include behavioral biometrics in the authentication process. The biometrics-based system monitors the user's behavior, compares it with his usual actions, and can thus point out suspicious inconsistencies.

The goal of this thesis was to explore the possibility of creating a user identity verification model based on his behavior (usage of mouse and keyboard) in a web application.

## Data collection

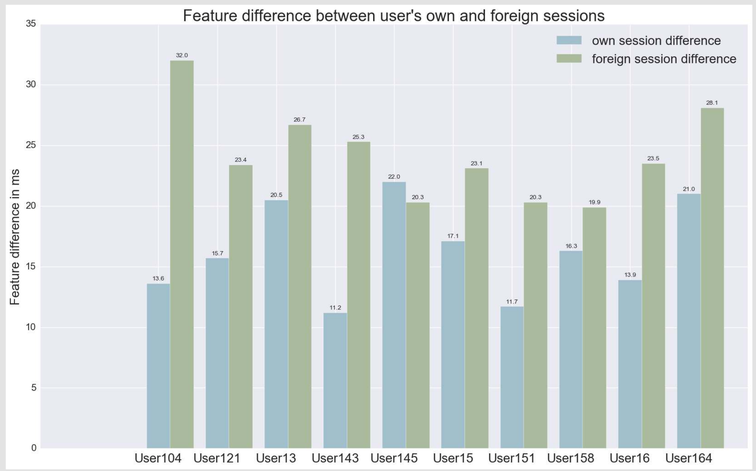
The data acquisition task was carried out in cooperation with the IT company Profinit EU, s.r.o. which enabled collection of data on employee behavior on selected pages of the firm's internal web application.




The user works with the monitored pages of the application on his computer. Via a JavaScript script his actions are recorded. This data is then sent to the server where it is further processed and stored in a database.

## Feature analysis

The ideal feature is stable within a single user's sessions but differs greatly between sessions of two different users. To select potentially useful features we compared average difference in the feature between sessions of a single user with the average difference between distinct users' sessions.




| Feature | Difference increase |
|---|---|
| Key hold duration | 69 % |
| Flight time between keys | 47 % |
| Keystroke overlap percentage | 101 % |
| Click duration | 84 % |
| Button click position | 16 % |
| Menu item click position | 11 % |
| Mouse movement speed | 27 % |
| Period between mouse moves | 225 % |

Out of the eight features evaluated, we decided to use those that reached over 50 % increase in distance between the user's own and foreign sessions. For all features directly related to time we also considered their standard deviations.

## Final dataset

From the obtained data, we created a dataset consisting of three parts describing keystroke dynamics, style of clicking and mouse movements.

Our dataset is relatively unique among others for several reasons.

Firstly, the number of users whose data we have available. In total the dataset contains records from more than 500 users, which is significantly more than in most other studies.

The second specificity of our data is the long time it has been collected from users. It is more than seven months from the end of October 2019 till the end of May 2020.

It is also quite rare that the data comes from a specific application that is not created for study purposes only, but users actually work with it on a regular basis.

## Results

We measured verification accuracy rates on the most often used machine learning models. Every user had their own model trained. Its task was to distinguish between the genuine user's sessions and randomly selected sessions of others.

| Model | Accuracy |
|---|---|
| Random forest | 84,4 **%** |
| Support vector machines | 83,1 % |
| Naive Bayes classifier | 79,7 % |
| K nearest neighbors | 78,4 % |
| Decision tree | 78,2 % |