USENIX Security'12

On the Feasibility of Side-Channel Attacks with Brain-Computer Interfaces

Ivan Martinovic[*], Doug Davies[†], Mario Frank[†], Daniele Perito[†], Tomas Ros[‡], Dawn Song[†]

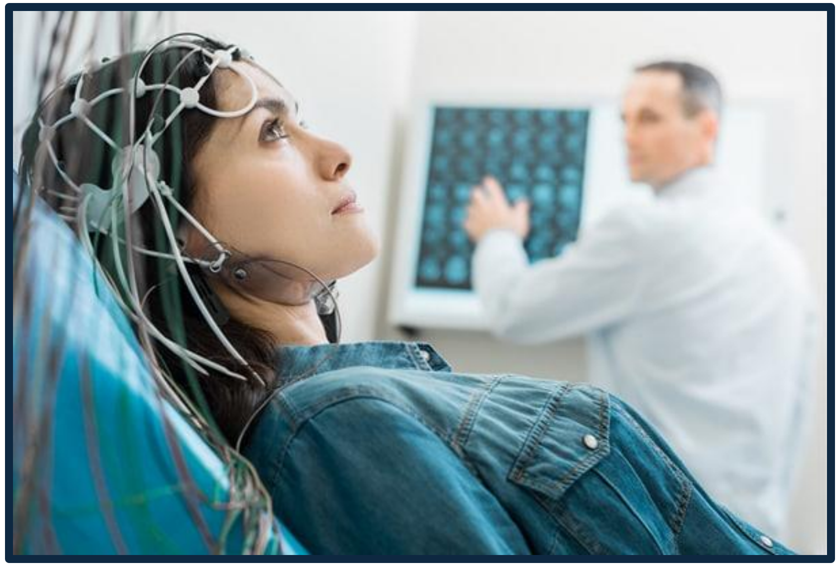
[*] UNIVERSITY OF OXFORD [†] Berkeley UNIVERSITY OF CALIFORNIA [‡] UNIVERSITÉ DE GENÈVE

Presenter: Sihun Yang

CyPhy LAB
Cyber-Physical Systems and Security Lab

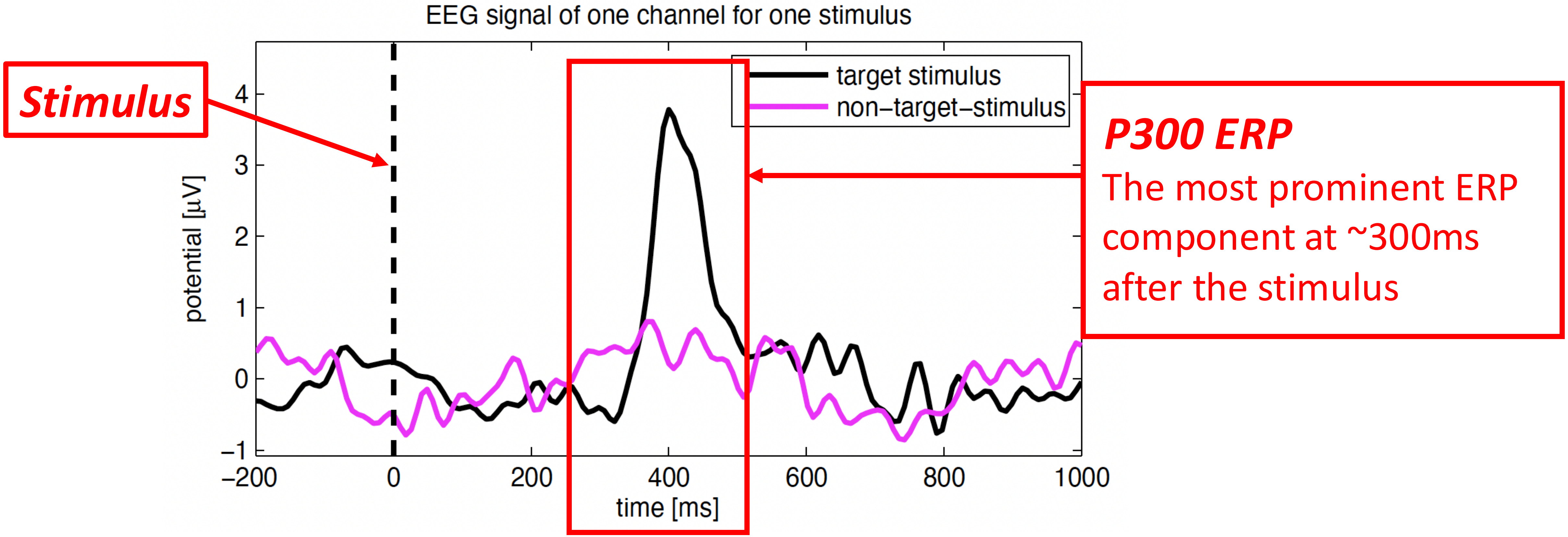
Brain-Computer Interfaces (BCIs) Getting Pervasive

- **Non-muscular communication** using the **brain's activities**
- **Medical** (e.g., ADHD therapy) and **non-medical applications** (e.g., gaming)
- **Consumer-grade BCI devices** available on the market



Electroencephalography (EEG) under Stimulus

- Recorded from **scalp electrodes**, indicating **neuronal activity**
- Used for **non-invasive** brain activity monitoring in **most BCIs**
- **Event-Related Potential (ERP) -** a change in EEG **after a stimulus**



Stimulus

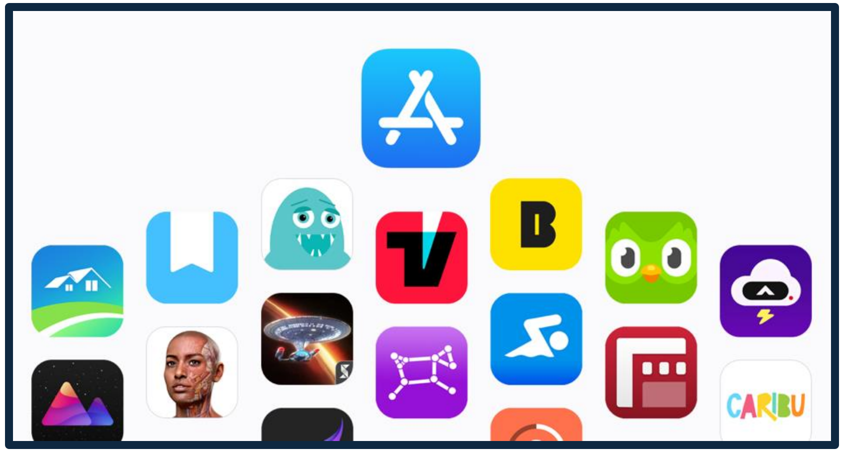
P300 ERP

The most prominent ERP component at ~300ms after the stimulus

Consumer-grade BCI Devices with Unrestricted APIs

- Popular devices such as *EPOC* (Emotive Systems) and *MindSet* (NeuroSky)
- "**App Stores**" with diverse **third-party apps**
- **Unrestricted access to the raw EEG** and **control over the stimuli** via API

➔ ***Third-party BCI app developers can inject stimuli & observe the EEG ERP***



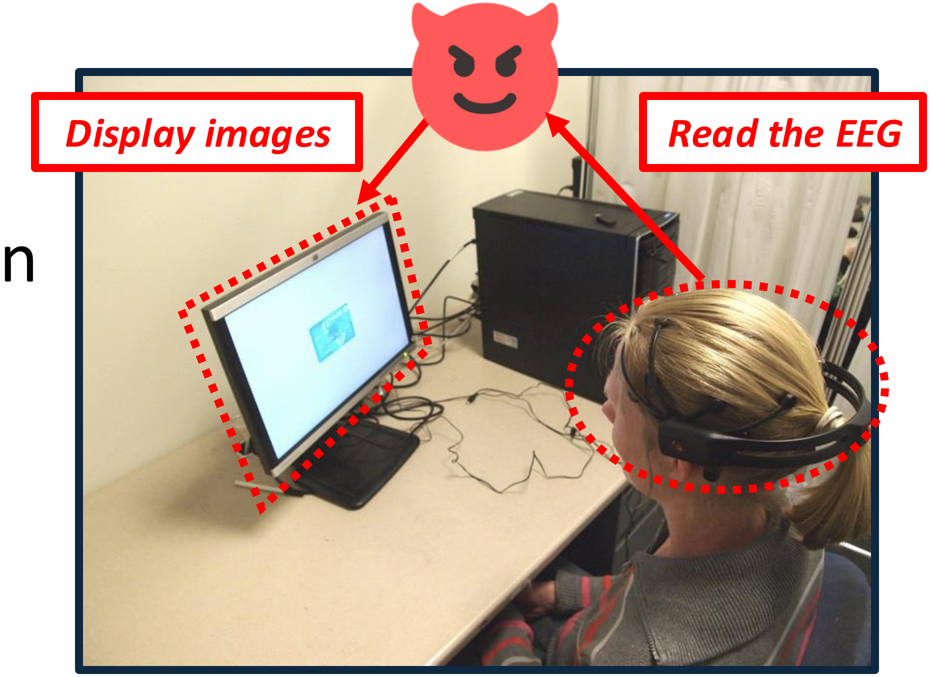
Can malicious third-party EEG applications infer private information about the users by manipulating stimuli and the responses?

Threat Model of BCI Attacks

- **Attacker**: Malicious third-party developers of ECG-based BCI application

- **Attacker's Goal**
  - To learn as much information as possible about the user

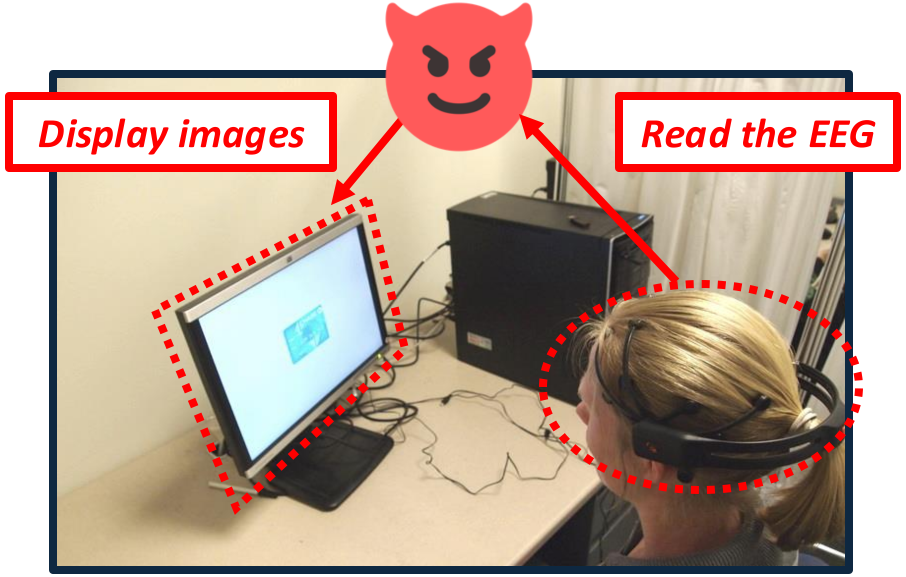
- **Attacker's Capabilities**
  - Can **display text/videos/images** on the screen
  - Can **read the EEG** signal from the BCI device



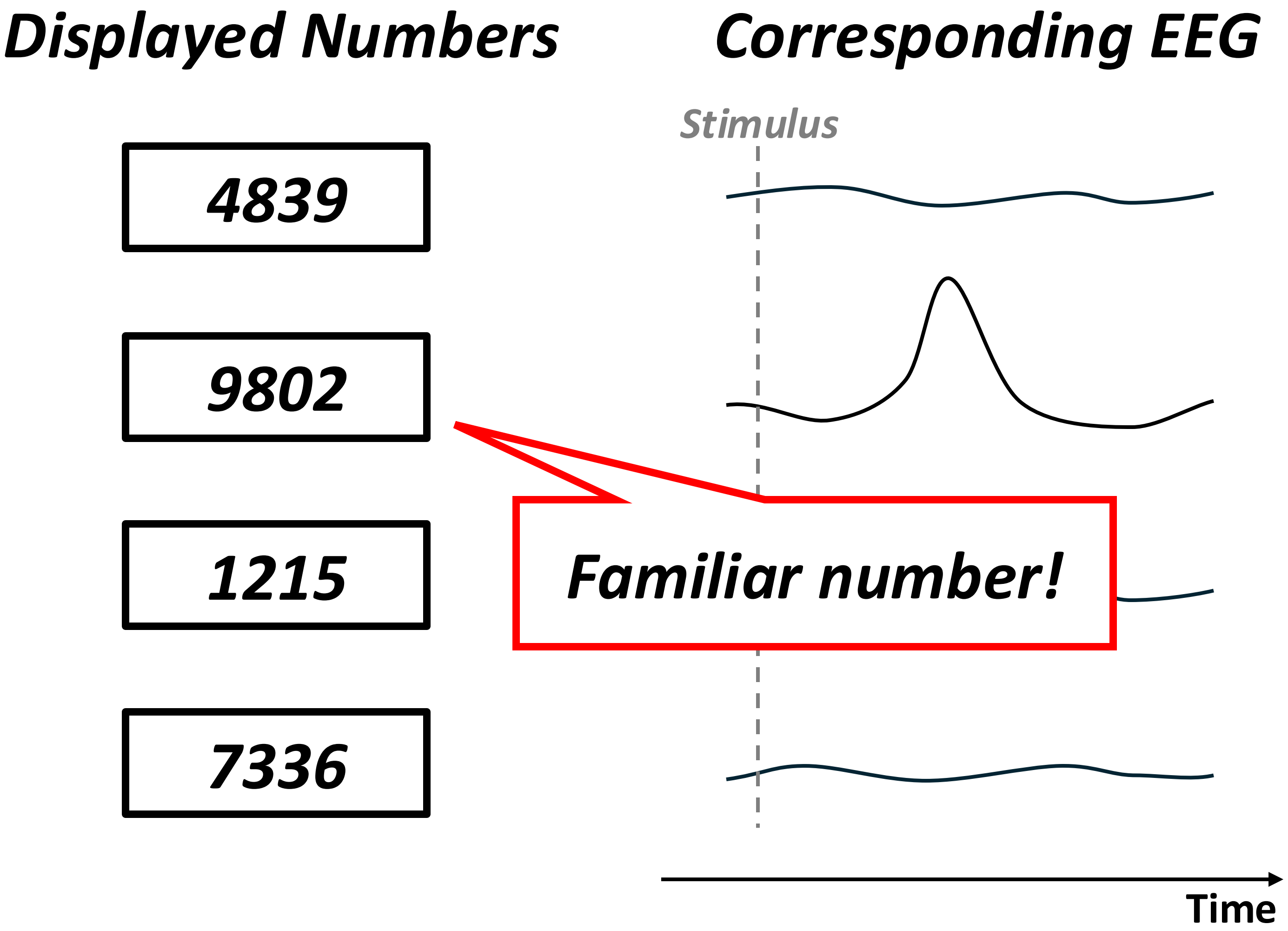
Display images

Read the EEG

High-level Attack Process


Display images

Read the EEG


Displayed Numbers

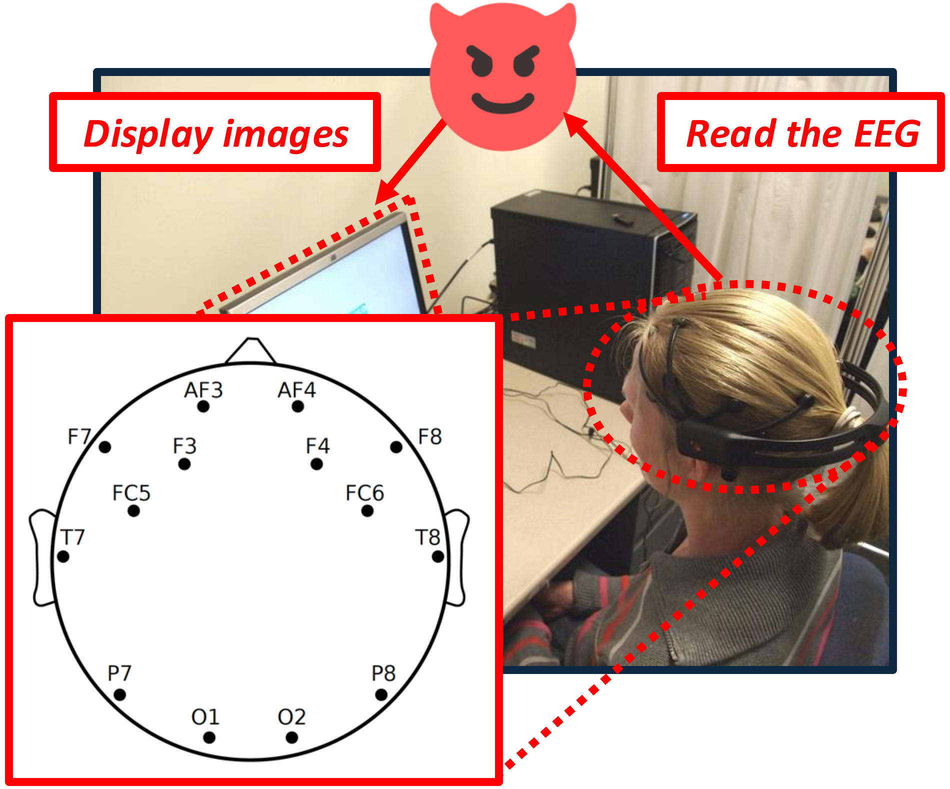
Corresponding EEG

Familiar number!

Experimental Setup

- **28 participants** X **40 minutes-long** experiment
- **Apparatus**: A **computer screen** and an **EEG-based BCI device (*EPOC*)**
- EEG collected with **14 channels (**i.e., electrodes**)**, at **128Hz per channel**

- **Experiment Protocol**
  1. (Optional) Explanation of the task
  2. (Optional) Message on screen for 2s
  3. Images being flashed in random order for the duration of 250ms



Display images

Read the EEG

Experiment 1: Pin Code

- **Goal:** To gather partial information about a user’s **4-digit PIN**

- **Experiment Details**
  1. Participants choose and memorize a **4-digit PIN**
  2. Display images consisted of a sequence of **randomly permuted digits (0-9)** for ~90 seconds
  3. Participants are asked to enter the first digit of the PIN

Experiment 2: Bank Information

- **Goal:** To obtain the **name of the bank** of the participants

- **Experiment Details 1**
  1. Display the **logo of 10 different banks**
  2. A bank with a higher response may be the answer
  3. **Failed! –** Participants recognize all the logos

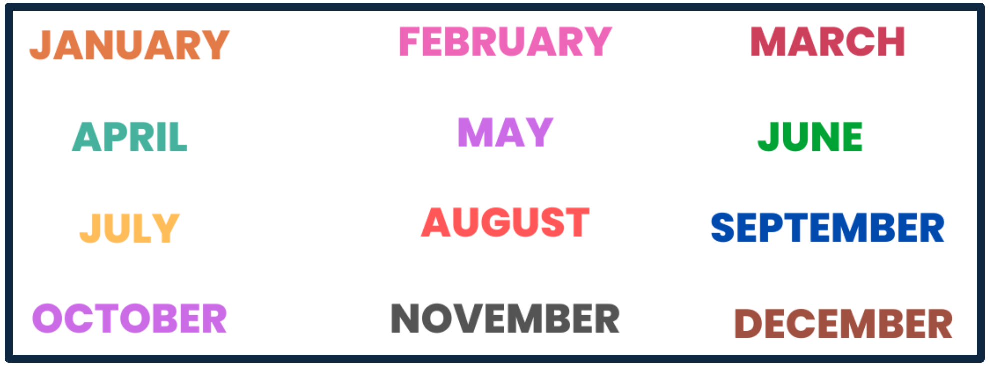
KB 국민은행
IBK 기업은행
EVER RICH
BS 부산은행
우리은행
KEB 외환은행
새마을금고
전북은행
신한은행
NH NongHyup
수협은행
DGB 대구은행
하나은행
kdb 산업은행
신협
광주은행
Standard Chartered SC 제일은행
citibank
경남은행
www.knbank.co.kr

- **Experiment Details 2**
  1. Display the **images of ATMs and credit cards**
  2. A bank with a higher response may be the answer
  3. **Worked!**

Experiment 3: Month of Birth

- **Goal:** To obtain the **month of birth** of the participants
- **Motivation**
  - The date of birth often serves as a **backup function for resetting password**

- **Experiment Details**
  1. Display an on-screen message: *“Which month were you born?”*
  2. Display a **randomly permuted sequence of the names of the months**
  3. A name of a month with a higher response may be the answer



JANUARY	FEBRUARY	MARCH
APRIL	MAY	JUNE
JULY	AUGUST	SEPTEMBER
OCTOBER	NOVEMBER	DECEMBER

Experiment 4: Face Recognition

- **Goal:** To obtain the **face images** that the participants recognize
- **Motivation**
  - **Selecting photos of friends** has been used for **account verification** on Facebook

- **Experiment Details**
  1. Display an on-screen message: “*Do you know any of these people?*”
  2. Display **randomly flashing images of people**
  3. A person with a higher response may be the answer

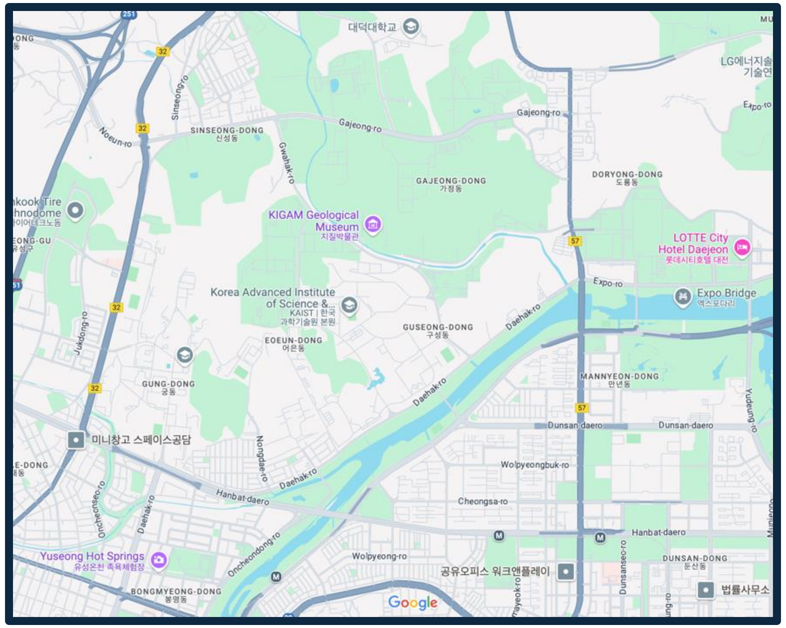
Do you know any of these people?

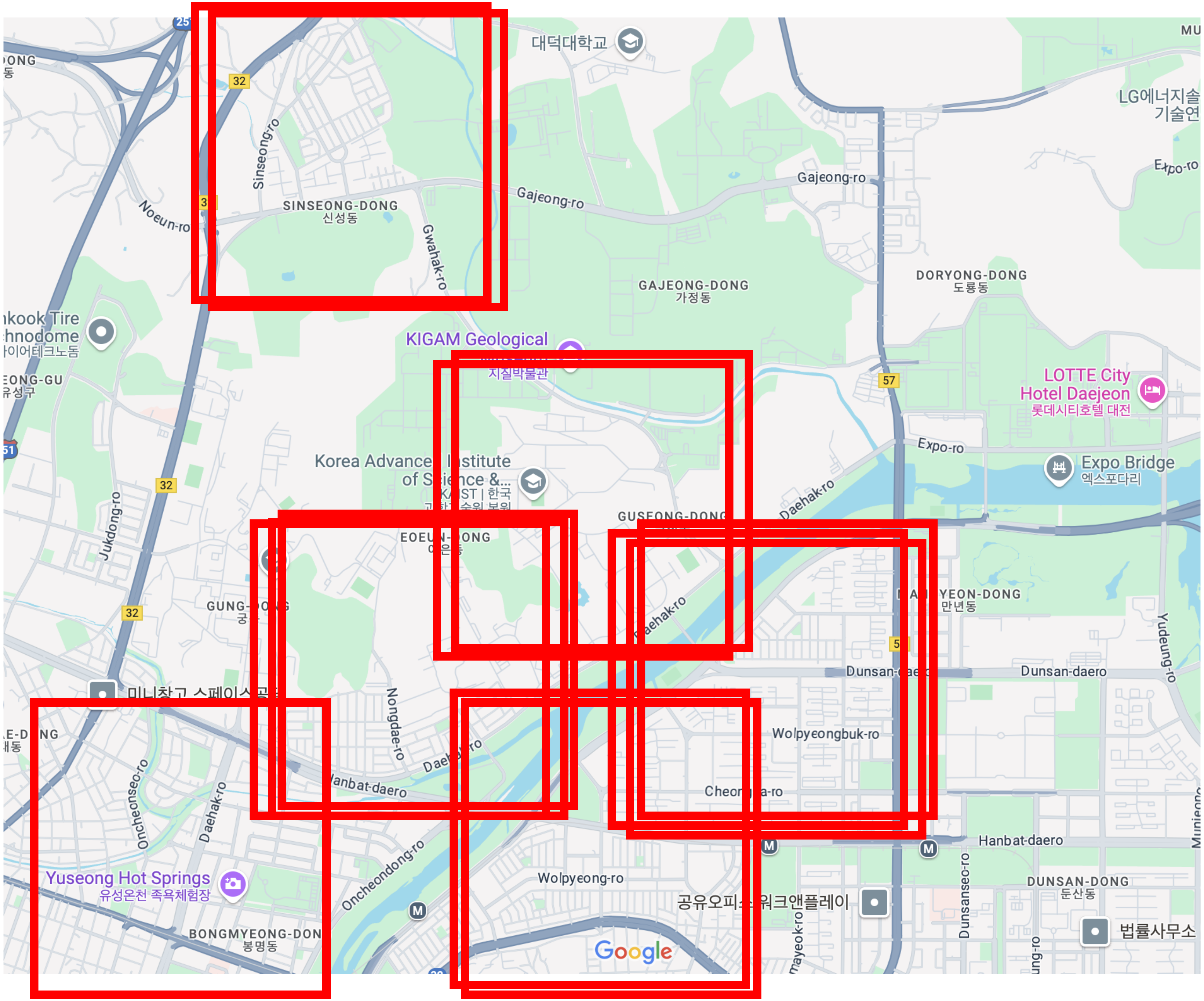
Experiment 5: Geographic Location

- **Goal:** To pinpoint the **geographic location of the residence** of the participants

- **Experiment Details**
  1. Display **highlighted maps** with the different highlighted regions
  2. Participants are asked to **count how often their region was highlighted**

  → **To study the influence of *active user support***




대덕대학교
LG에너지솔
기술연
Sinseong-ro
SINSEONG-DONG
신성동
Noeun-ro
Gwahak-ro
Gajeong-ro
Gajeong-ro
Expo-ro
GAJEONG-DONG
가정동
DORYONG-DONG
도룡동
kook Tire
hnodome
이어테크노돔
EONG-GU
유성구
KIGAM Geological
지질박물관
LOTTE City
Hotel Daejeon
롯데시티호텔 대전
Expo-ro
Expo Bridge
엑스포다리
Korea Advance Institute
of Science &...
KAIST | 한국
Daehak-ro
GUSEONG-DONG
Jukdong-ro
EOEUN-DONG
어은동
GUNG-DONG
궁동
YEON-DONG
만년동
Daehak-ro
Yudeung-ro
Dunsan-daero
Dunsan-daero
미니창고 스페이스
Nongdae-ro
Wolpyeongbuk-ro
Oncheonseo-ro
Daehak-ro
Hanbat-daero
Daehak-ro
Cheong a-ro
Hanbat-daero
Yuseong Hot Springs
유성온천 족욕체험장
Oncheondong-ro
Wolpyeong-ro
공유오피스 워크앤플레이
BONGMYEONG-DONG
봉명동
Dunsanseo-ro
DUNSAN-DONG
둔산동
법률사무소
Google

Technical Challenges in Data Analysis

- **Challenge 1 (***Unsupportive User***)**:
  The user does **not intend to provide a discriminative signal** for the target stimuli

- **Challenge 2 (***Unsupportive Device***)**:
  The consumer-grade gaming BCI device is **not made for detecting P300**

➔ The data is **less discriminative** between **target** and **non-target stimuli**

Solution: To use a *classifier* to detect P300 and the corresponding stimuli

Classification of Target Stimuli

- **Detecting P300 of Target Stimuli**
  → Binary classification task consisting of ***Training Phase*** + ***Classification Phase***

- **Epoch:** All EEG data values associated with a single stimulus
  - **Size of an Epoch = # of channels × # of data points per channel =** p

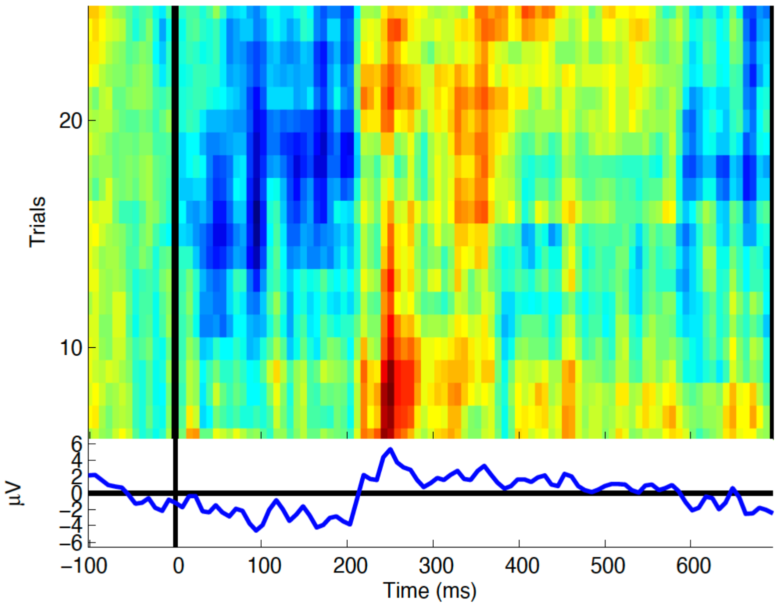
- ***Training Phase***: To learn a function g that maps epochs $\boldsymbol{x}$ to stimuli labels y
  - $g\colon\ \mathbb{R}^p \longrightarrow \{0, 1\}$ maps $\boldsymbol{x} \in \boldsymbol{X}^{\mathbf{tr}}$ to $y \in \boldsymbol{y}^{\mathbf{tr}}$
  - E.g., labels of unfamiliar face images = 0, labels of Donald Trump's face images = 1

- ***Classification Phase***: To estimate $\hat{y} = g(\boldsymbol{x})$ given $\boldsymbol{x} \in \boldsymbol{X}^{\mathbf{test}}$

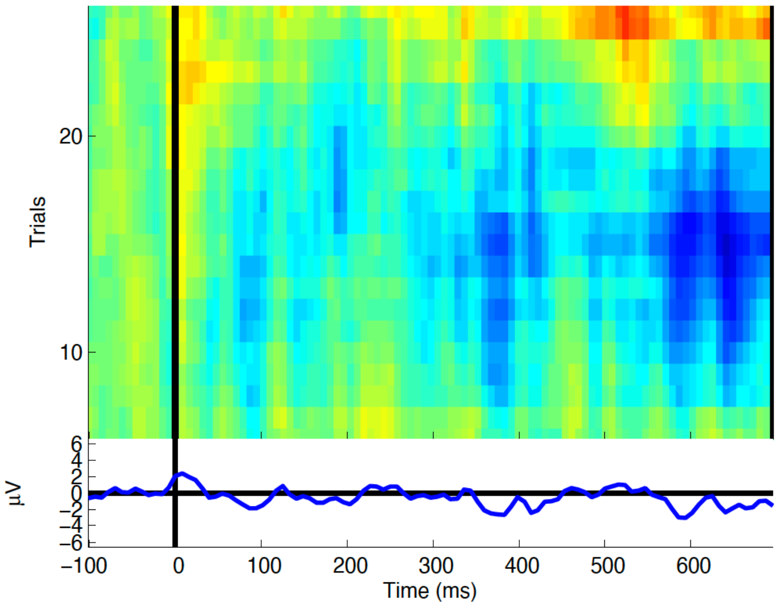
Classification of Target Stimuli

- **Detecting P300 of Target Stimuli**

Target Stimulus



Non-target Stimulus



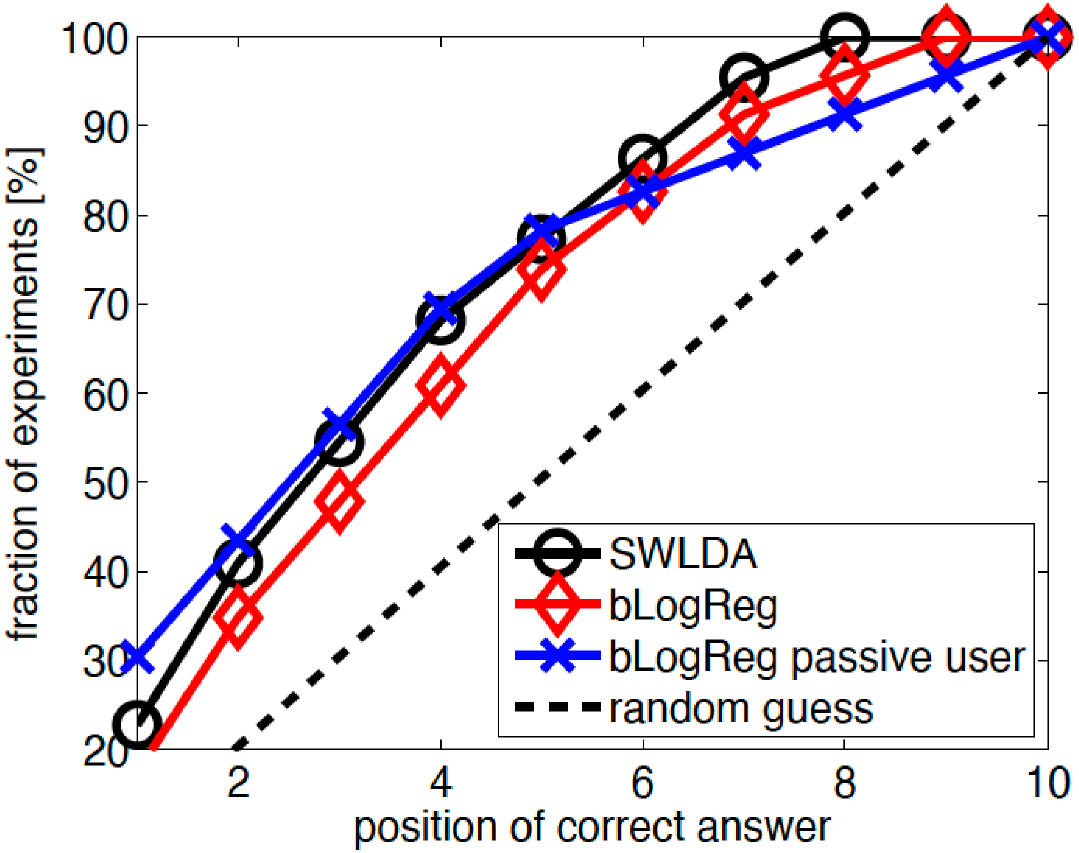
- Use **two** different existing P300 classifiers: ***bLogReg***, ***SWLDA***
  - ***bLogReg***: Boosting + Logistic Regression
  - ***SWLDA***: Extension of LDA (Linear Discriminant Analysis)

Classification of Target Stimuli

- **Training Scenarios**
  - **User-supported Calibration**
    - Actively supports the ***training phase*** (e.g., gamer having an incentive for **initial calibration in games**)
    - E.g., **keep concentrating on the target stimuli** like in Experiment 5
  - **On-the-fly Calibration**
    - Does not support the ***training phase*** but also does not actively try to disturb it (i.e., passive user)
    - **Assumption**: The user is **familiar with one** of the stimuli and **unfamiliar with all others** (e.g., Donald Trump + Random Face Images)
  - None of the scenarios assumes a user supporting the detection of new target stimuli, showcasing **realistic training scenarios** for potential attackers.

Experimental Results

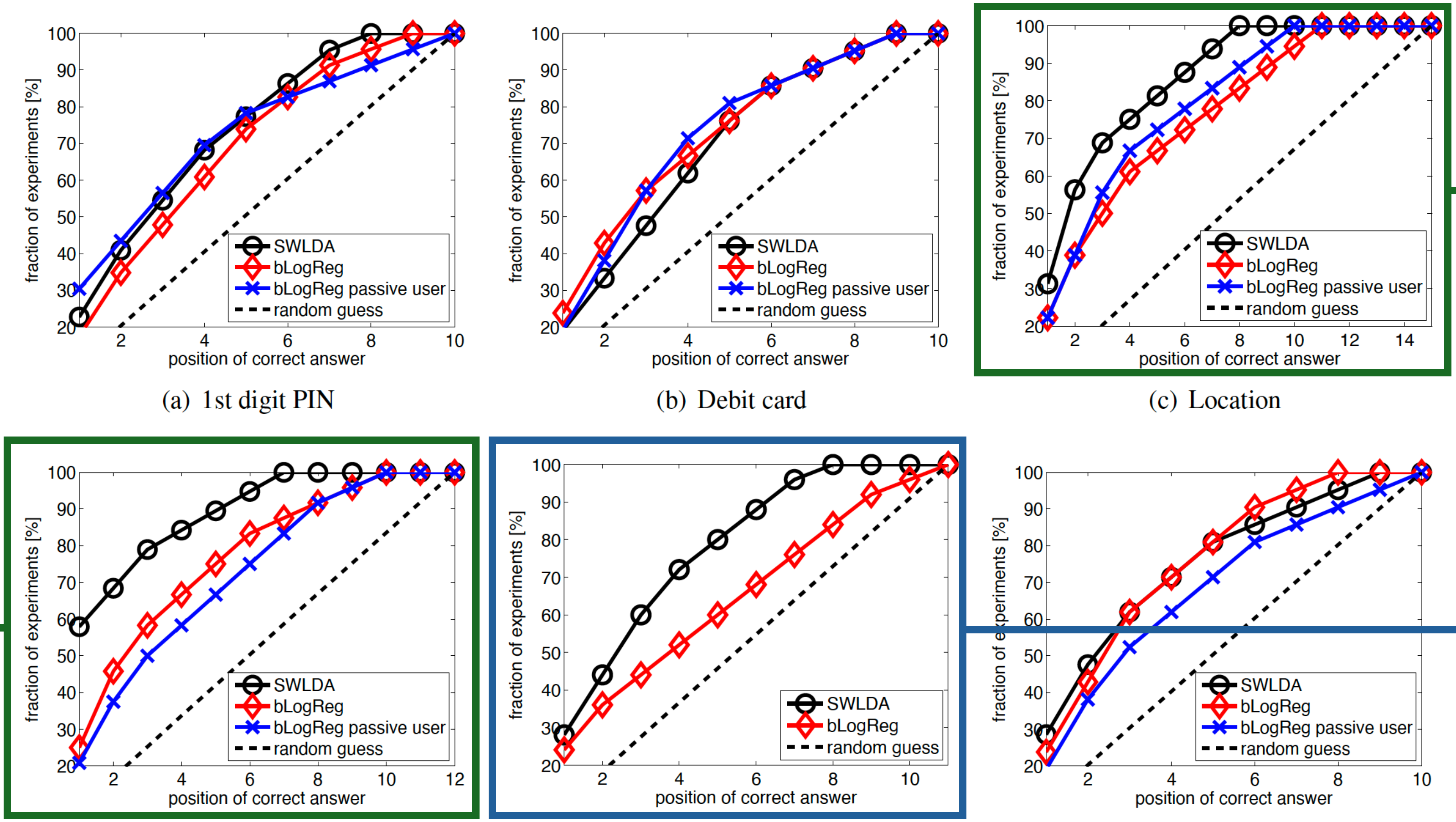
- **Overall Attack Success Statistics –** PIN Code Attack



(a) 1st digit PIN

Experimental Results

- **Overall Attack Success Statistics**
  - Significantly **better than the random guess**, indicating the **private information in EEG**
  - Notably **successful** where the **users actively concentrated by counting** (Experiment 5)



(a) 1st digit PIN

(b) Debit card

(c) Location

(d) Month of birth

(e) People

(f) ATM machine

Experiment 5 (*Locations*) is notably successful

Experiment 3 (*Months*) is successful for some reason

Experiment 4 (*Faces*) is used as the training set of the passive user scenario

Discussion

- **Defenses**
  - **Actively try to hinder probing (**e.g., concentrating on non-target stimuli**)**
    - Not realistic for normal users
  - **Not expose the raw EEG to third-party apps**
    - Make a **restricted API** that has **limited access to certain features of the EEG**
  - **Add noise to the EEG raw data before providing it to apps**
    - Have **drawbacks in terms of the accuracy** of benign apps

- **Future Directions**
  - **To avoid the discovery of the attack, i.e., make the attack more stealthy**
    - Integrated into benign everyday tasks
    - Performed with the social engineering approach

Related Work

- **Previous Work of This Paper**
  - **EEG-based Identification and Authentication**
    - **M. Poulos et al., 1999**, Parametric person identification from the EEG using computational geometry
→ Achieved a high true positive rate and a high true negative rate in user identification
    - **J. Thorpe et al., 2005**, Pass-thoughts: authenticating with our minds
→ Thinking of a password as a password
    - **S. Marcel and J.R. Milán, 2007**, Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation → GMM-based robust EEG authentication
  - **Guilty-Knowledge Test**
    - **J.P. Resenfeld, 2004**, Simple, effective countermeasures to P300-based tests of detection of concealed information → Verify robustness of P300-based GKT
    - **V. Abootalebi, 2009**, A new approach for EEG feature extraction in P300-based lie detection
→ Use P300 in lie detection
    - **M. Gamer, 2010**, Does the Guilty Actions Test allow for differentiating guilty participants from informed innocents? A re-examination → Re-visited the performance of detection of potential criminal details using P300

Related Work

- **Follow-up Work of This paper**

  - **ACM CCS’15**, A Multi-Modal Neuro-Physiological Study of Phishing Detection and Malware Warnings
    → Detecting phishing websites using EEG in addition to the eye gaze

  - **ACM MobiSys’18**, Brain Password: A Secure and Truly Cancelable Brain Biometrics for Smart Headware
    → BCI/HMD user authentication system using visual stimulus and the corresponding neurofeedback

  - **USENIX Security’21**, Inexpensive Brainwave Authentication: New Techniques and Insights on User Acceptance
    → Benchmark on BCI user authentication using ERP including P300 and N400

  - **IEEE ICASSP’23**, EEG2IMAGE: Image Reconstruction from EEG Brain Signals
    → Image reconstruction from the EEG signal using a generative model

Conclusion

- Focus on the possibility of **privacy breaches** in **pervasive EEG-based BCI**

- Show the feasibility of using a consumer-level BCI device to reveal the **private information of the users**: PIN, bank information, location, and familiar faces

- Demonstrate **significant information leakage up to ~43%** from EEG and ERP, showcasing the feasibility of real-world BCI attacks

Good Questions

- Besides P300, other ERPs like N400 (associated with semantic processing) can also reveal private information. Could attackers design more sophisticated attacks by combining signals from multiple ERP components? What types of sensitive data might be extractable using ERPs other than P300?
- How do cultural and environmental factors, such as a user’s familiarity with specific images (e.g., logos, landmarks, or famous faces), affect the accuracy of EEG-based attacks? Could attackers tailor stimuli based on geographic or demographic data to increase attack success rates?
- Considering current API designs allow unrestricted access to raw EEG data, what alternative data management structures could provide a balance between security and usability for developers?
- Should there be global regulatory standards for consumer-grade BCIs to ensure that user data is protected against misuse, and what challenges might such regulations face?
- More than 10 years have passed since the paper was published. How has security in BCI (Brain-Computer Interfaces) developed during this time, and what is the current state?
- Can this attack be launched more stealthily, without the victim noticing at all?

Best Questions

- **Munim:** If access to raw EEG signals were restricted via API controls, what level of granularity in signal processing (e.g., removing specific ERP components like P300) would balance user privacy and application functionality? Could such restrictions be implemented without significantly reducing the accuracy of legitimate applications like gaming or neurofeedback tools?
- **Hyunmin:** With advancements in AI, do you foresee attackers leveraging more sophisticated neural decoding techniques, and how can we preemptively counteract this?
- **Isu:** It seems like the questions and instructions were selected with those that can be answered with very limited categorical answers. What is your opinion on this and how do you think if continuous answers, such as specific names or secrets, might be possible?