Exploring keystroke dynamics for insider threat detection

Dr. Shing-hon Lau

Cybersecurity Engineer

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM19-0417

The SEI is a DoD R&D Federally Funded Research and Development Center

Established in 1984 at Carnegie Mellon University

~650 employees (ft + pt), of whom about 70% are engaged in technical work

Initiated CERT cybersecurity program in 1988

Offices in Pittsburgh and DC, with several locations near customer facilities

~$140M in annual funding

Acknowledgements

- Doctoral thesis work advised by:
  - Roy Maxion (CMU, advisor)
  - Tom Mitchell (CMU)
  - Dan Siewiorek (CMU)
  - Peter Strick (U. Pitt)
  - David Banks (Duke)
  - Mark Wetherell (U. of Northumbria)

- With assistance from:
  - Patricia Loring (research assistant, CMU)
  - Huayun Huang (undergrad, CMU)
- Work was supported by SEI-CERT, NSA, NSF

Talk focus – specific project and general lessons

- Two main objectives today
  - Share a specific project about the possibilities for detecting insiders through keystroke dynamics
  - Extract out the general experimentation principles that guided our decision-making process as we did this project

- Cybersecurity frequently involves experimentation
  - Change policies, plans, procedures to (hopefully) improve security
  - Attempt to adjust user behavior

Agenda

- **Background**
- **Our experimentation principles**
- **3 research questions**
- **Experimental details**
- **Results – answering the 3 research questions**
- **Takeaways**

Agenda

- **Background**
- Our experimentation principles
- 3 research questions
- Experimental details
- Results – answering the 3 research questions
- Takeaways

Stress and computing

- Observations:
  - Humans sometimes interact with computers while under significant stress
  - The most common form of interaction with a computer is typing

- Initial question:
  - Can we use typing to detect that an individual is stressed?

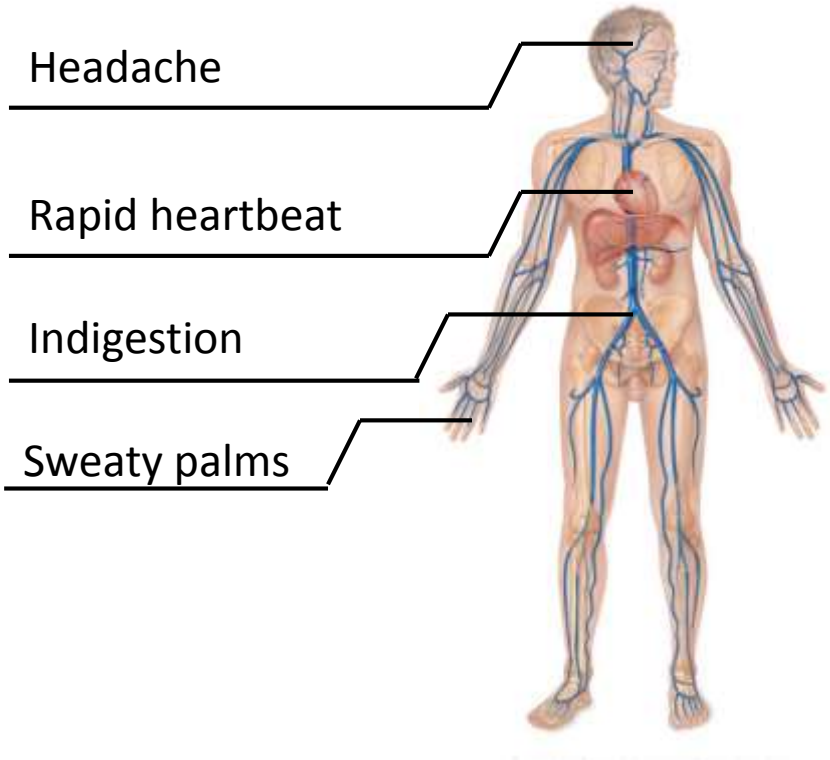
Scenarios of interest: Insider threat and more

Insider threat

Operations center

Chronic stress

The physiology of stress



- Stress is anecdotally familiar
  - Time-pressure
  - Emotional stress
- Familiar symptoms
- Expectation that an insider will be highly stressed while engaging in illicit behavior
- How can we detect that stress?
  - Keystroke dynamics

What is keystroke dynamics?

- The study of (keyboard) typing rhythms
- Based largely on two types of features:
  - Hold times – how long a key is depressed while typing
  - Latency times – time between consecutive key presses
- Requires only software; no additional hardware needed
- Focus on traditional keyboards for today's talk, but techniques can be expanded to virtual keyboards, touchscreens, smartphones, etc.

A brief history of keystroke dynamics

- Bryan and Harter, Psychological Review, 1897
  - Identify telegraphs operators by their telegraph “keying” rhythm
- Gaines et al., RAND corporation, 1980
  - Initial study on the feasibility of keystroke dynamics
- Several hundred papers published between 1980-2019 that largely attempt to discriminate between subjects (e.g., for two-factor authentication)
- Published results are promising – discrimination accuracies of 99+% reported

- Key idea: changes in physiology due to an insider’s stress may be identifiable through keystroke dynamics

Agenda

- Background
- **Our experimentation principles**
- 3 research questions
- Experimental details
- Results – answering the 3 research questions
- Takeaways

Setting up an experiment: what has been decided?

- People will type in an experiment
- Some of that typing will be while they are stressed
- We will analyze the typing to look for signs of stress

What has not been decided yet?

- How many people will participate in the experiment?
- What will they type?
- How much typing is enough?
- Where are they going to do this typing?
- How do we get our typists to be stressed?
- How will we know they are stressed?
- What analysis techniques will be used?
- What does it mean to detect stress in typing anyway?
- Are we collecting any other data besides typing?
- How will we do this within our resource constraints?
- And many more questions…

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration	

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration • Experiment should answer questions authoritatively and remove as much doubt as possible • Often requires a tradeoff with realism	

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration • Experiment should answer questions authoritatively and remove as much doubt as possible • Often requires a tradeoff with realism • Create a solid foundation for future work	

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration • Experiment should answer questions authoritatively and remove as much doubt as possible • Often requires a tradeoff with realism • Create a solid foundation for future work	• Want to explain our end results using simple, understandable, communicable ideas

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration • Experiment should answer questions authoritatively and remove as much doubt as possible • Often requires a tradeoff with realism • Create a solid foundation for future work	• Want to explain our end results using simple, understandable, communicable ideas • Not enough to obtain an outcome (i.e., success or failure) • Must have an ability to explain why that outcome occurred

Lots of decisions to be made – what are our general principles?

Productive failure	Creating understanding
• Ability to learn something useful from an experiment even if the desired outcome is not obtained • Can rule out specific ideas or techniques from future consideration • Experiment should answer questions authoritatively and remove as much doubt as possible • Often requires a tradeoff with realism • Create a solid foundation for future work	• Want to explain our end results using simple, understandable, communicable ideas • Not enough to obtain an outcome (i.e., success or failure) • Must have an ability to explain why that outcome occurred • Build off of pre-existing work

What have others done?

- Examined literature with the intent to build on existing understanding
- Our work falls under the broader category of affect (emotion) detection
- 15 different research groups have examined this area
  - 6 groups focused explicitly on stress
- Difficulty building off of existing work since we could not learn much of anything from it
  - Unable to rule in (or out) the idea that stress detection through keystrokes is possible in any fashion

Issues with prior work

- Sample methodology of prior work:
  - Take 15-20 subjects
  - Ask them to do their daily business and self-report their emotional state or show them a random video clip the researchers felt would elicit an emotion
  - Analyze corresponding typing data and report accuracy of X%

- Fundamental doubts:
  - Can results from such a small sample size be relied upon?
  - How do we know subjects were actually in some emotional state at the time they provided a typing sample?
  - Are there other factors (e.g., different computer) that might affect results?

- We need to control some of these varied factors to come to useful conclusions

Agenda

- Background
- Our experimentation principles
- **3 research questions**
- Experimental details
- Results – answering the 3 research questions
- Takeaways

What does detecting stress even mean?

1. Characterize how an individual subject's typing rhythms are altered by stress? [Within-subjects]
   - What typing features differ between neutral and stressed typing?

2. Identify universal markers for stress [Across-subjects]
   - Marker: any easily-interpretable characterization
   - E.g., the marker for blood type is the presence or absence of specific antigens
   - E.g., only stressed typists produce hold times >300 ms

3. Identify groups of subjects with common markers [Clustering]
   - What are the groups and what are their marker(s)?

Agenda

- Background
- Our experimentation principles
- 3 research questions
- **Experimental details**
- Results – answering the 3 research questions
- Takeaways

Experimental methodology

- Tradeoff between realism and ability to have productive failure
- Since we cannot easily build off prior work, our objective is to demonstrate that stress detection with keystroke dynamics can be shown to work at all
- Tightly controlled lab study is more useful than realistic field study for initial work

- If we can show that our approach works in a tightly controlled lab study, future work can relax the controls until we are in a real-world environment

- If we jump straight to the real-world environment like in prior work, we will not know whether our success or failure is due to the fact that stress can be detected through keystrokes or due to other factors

Experimental timeline

Baseline

Rest #1, Typing

Stress

Induce stress, Typing

Recovery

Rest #2, Typing

Induce both neutral and stress states in our subjects using proven techniques

Cannot assume that our subjects are in a neutral state when they come in the door

- May have been running late
- Exam period

Three categories of collected data

Independent assessment data

Typing data

Supporting data

Three categories of collected data

Independent assessment data

- Factors that are known to be strongly linked to stress
- Electrocardiogram (ECG/EKG)
- Blood pressure
- Respiration
- Psychological measures

Typing data

Supporting data

Three categories of collected data

Independent assessment data

- Factors that are known to be strongly linked to stress
- Electrocardiogram (ECG/EKG)
- Blood pressure
- Respiration
- Psychological measures

Typing data

- Correctly-typed repetitions of a string
- Keystroke timings (hold and latency times)

Supporting data

Three categories of collected data

Independent assessment data

- Factors that are known to be strongly linked to stress
- Electrocardiogram (ECG/EKG)
- Blood pressure
- Respiration
- Psychological measures

Typing data

- Correctly-typed repetitions of a string
- Keystroke timings (hold and latency times)

Supporting data

- May help us understand peculiarities of our other data
- Psychological inventories, demographic data, height, weight, still pictures, videos

Experimental apparatus

- Ensuring valid experiment requires us to establish that our subjects are in a neutral or stressed state when we expect they will be
  - Unacceptable for us to fail to detect a change in typing because a subject was not stressed in the first place
  - If we are able to detect a change in typing, we know that stress is present

- We are not experts at running this type of experiment and we are using techniques from psychology labs
  - Err on the side of having an excess of data rather than too little

Experimental apparatus – physiological

- Research and medical grade equipment for collecting ECG, blood pressure, and respiration
- All measures are time synchronized
- Measurements are sufficient for us to determine stress state of subjects
- Apparatus required for this experiment to verify that our experimental procedure is effective at causing stress, can be dropped in future experiments

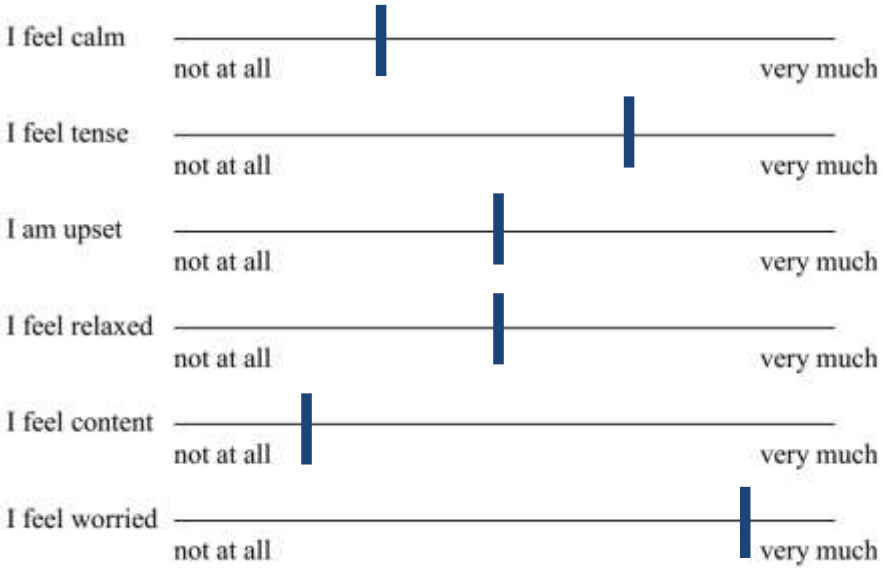
Experimental apparatus – psychological

- Short-form State Trait Anxiety Inventory (STAI)
  - Measures stress at the current moment
- NASA Task Load Index (NASA-TLX)
  - Measures perceived workload of a task
    - Mental Demand
    - Physical Demand
    - Temporal Demand
    - Effort
    - Performance
    - Frustration

- Instruments have been previously used by thousands of researchers

Short-form State Trait Anxiety Inventory (STAI)

A number of statements that people have used to describe themselves are given below. Read each statement, and then mark on the line at the most appropriate point to indicate **how you feel right now, at this moment.**



What will subjects type?

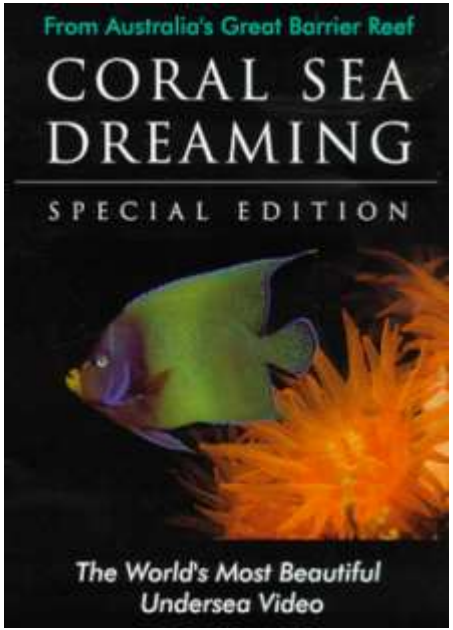
- Phrase criteria:
  - Memorable/familiar
  - Easy to type
    - Low variability
    - Minimizes practice effects
  - Free of emotionally charged text
- Generated 100 candidate phrases, then pruned down to 20
- Mechanical Turk experiment to find most easily-typed phrase
  - 413 subjects

- Final phrase: “great friends are good to have”
  - Typed 80 times in each session

Subject recruitment

- 116 subjects recruited using fliers on the CMU campus + word of mouth
  - "Convenience sample" works for initial study, but future work needs to check that results generalize
  - Common issue in cybersecurity: what works for IT may not work for accounting
- Sample exclusion criteria:
  - Diagnosed with, or on medication for,
    - Low (diastolic < 80)/high blood pressure (systolic > 140)
    - Cardiac, neurological, stress, anxiety, or sleep disorders
    - Stroke
- Subjects prohibited from alcohol, caffeine, stimulant, or psychoactive drug use leading up to experiment

Inducing neutral state



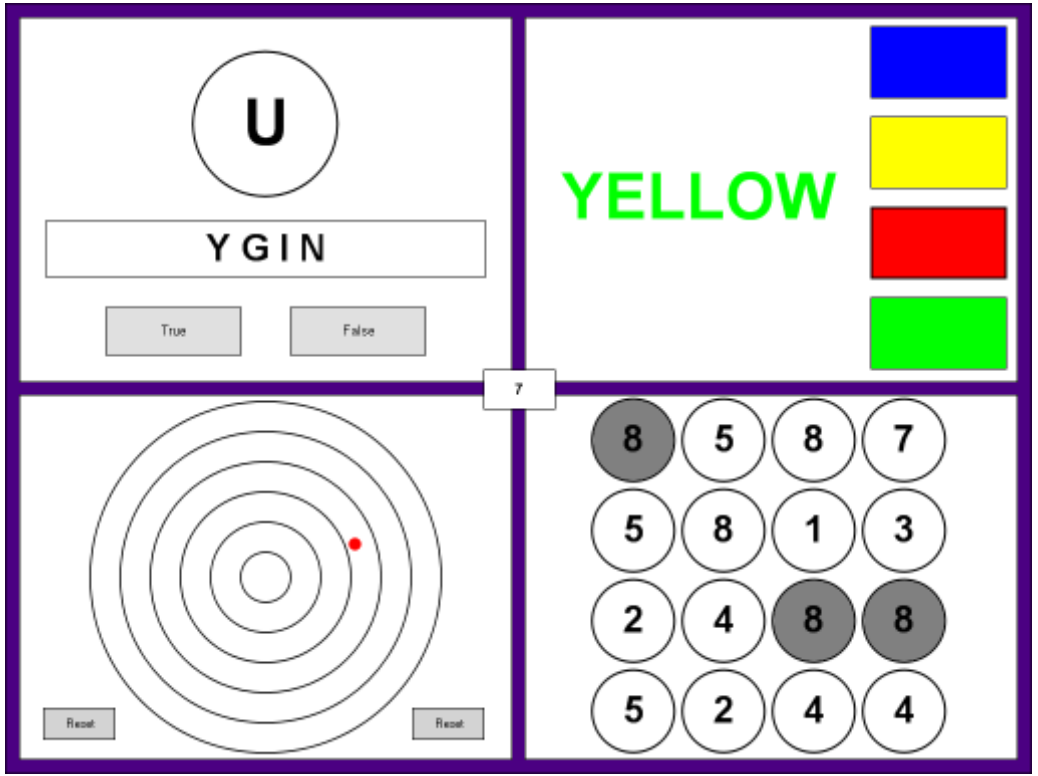
- Do not want to assume the state of our subjects when they walk in the door
- Ensure a neutral state by inducing it in subjects
- Subjects asked to watch relaxing scenes from a movie
  - 30 minutes in 1st rest period
  - 15 minutes in 2nd rest period

Inducing stress state

- Using an existing technique that has been demonstrated to work

- Combination of multi-tasking framework and social evaluation
  - Subjects asked to monitor several tasks at once
  - Subjects given negative feedback throughout the course of the task
  - 15-minute long exercise

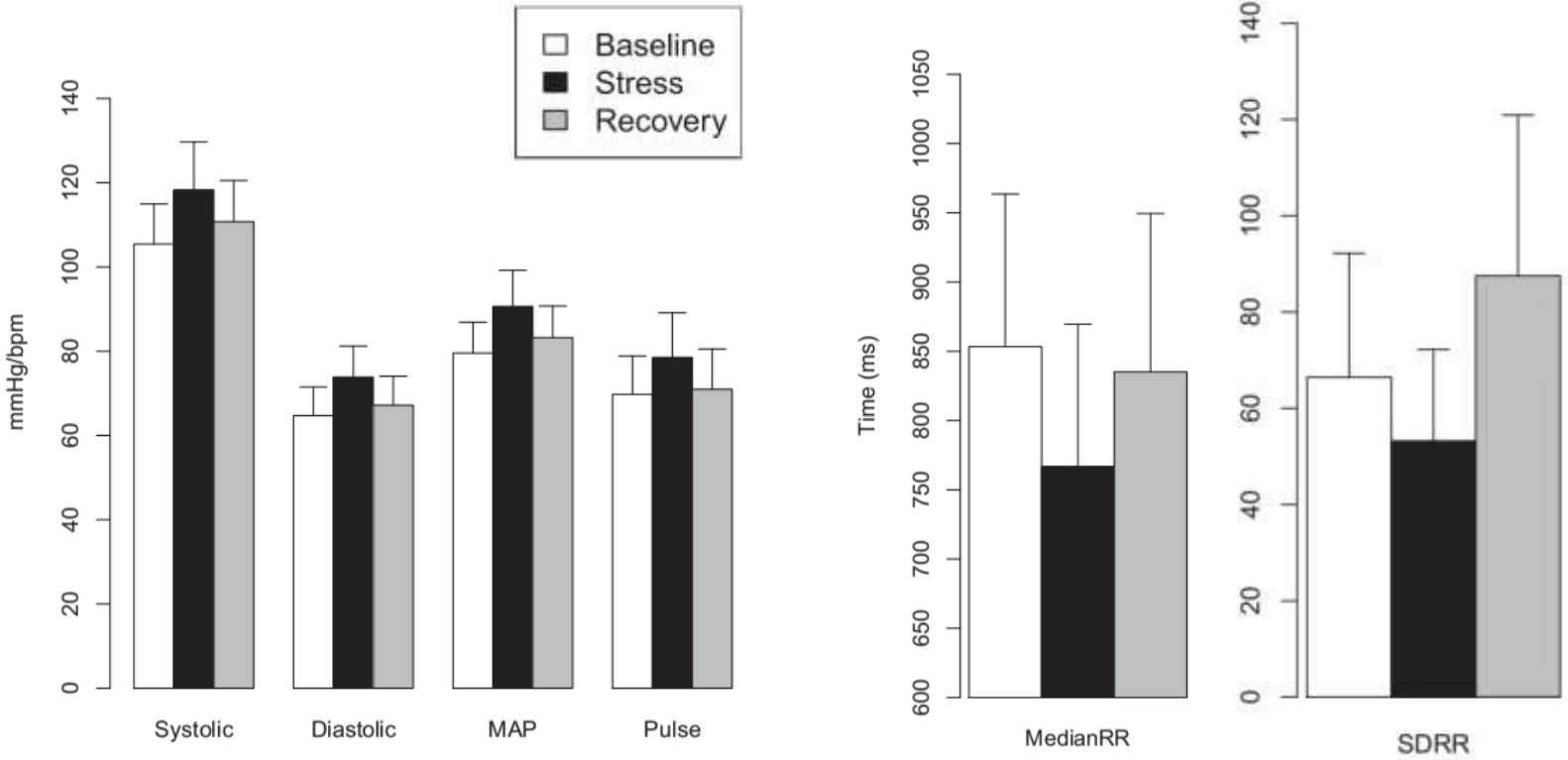
Multi-tasking framework



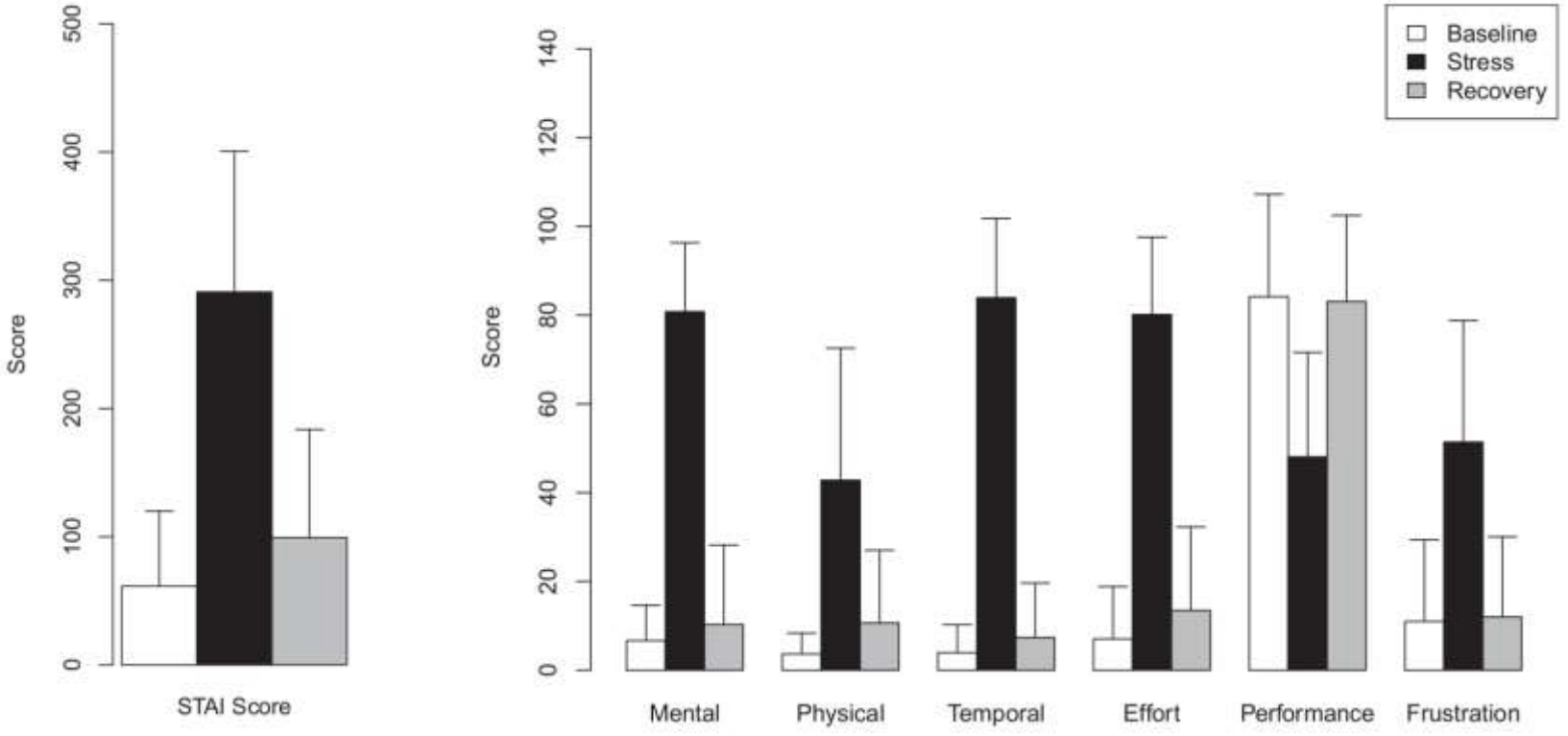
Agenda

- Background
- Our experimentation principles
- 3 research questions
- Experimental details
- **Results – answering the 3 research questions**
- Takeaways

Sanity check: did subjects get stressed?



Sanity check: did subjects get stressed?



Which analysis techniques to use?

- Prefer simpler, more interpretable algorithms rather than complexity

- Can expect to get better results if we tweak the algorithm or use something more sophisticated
  - Tweaking may give 5-15% improvement, but cannot make a signal appear where one does not exist

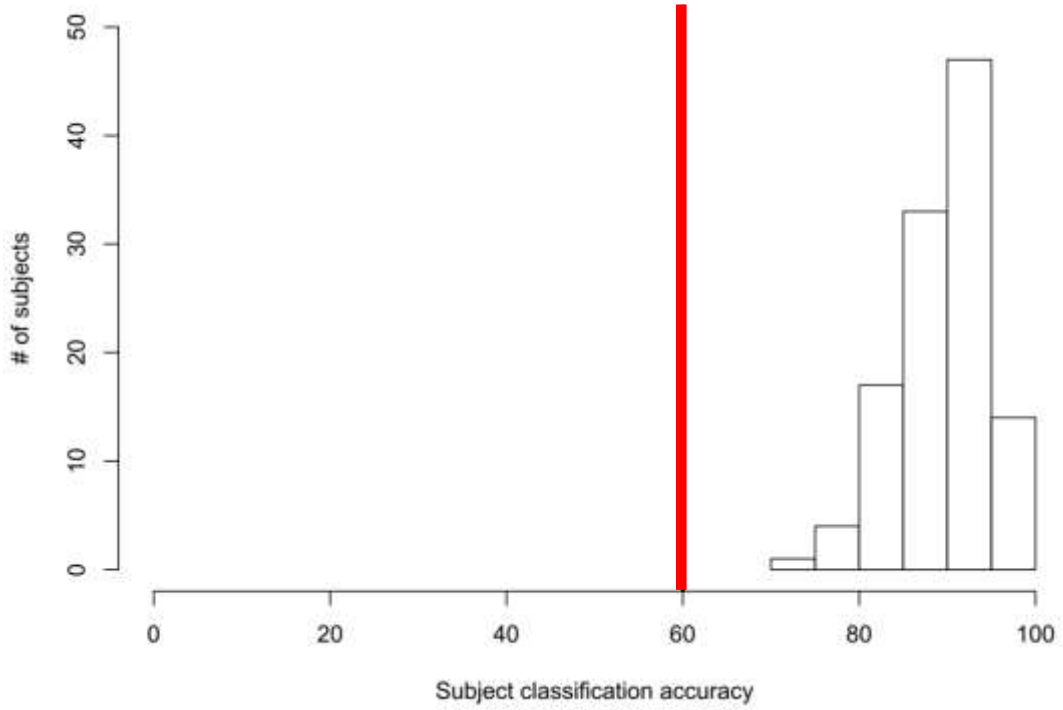
- Focused on three off-the-shelf algorithms:
  - **Random Forest (majority vote of decision trees fit with random subsample)**
  - Support-vector machine (SVM)
  - Sparse (LASSO) logistic regression

Q1: Within-subject typing changes due to stress (Setup)

- Goal: Characterize how an individual subject's typing rhythms are altered by stress

- For each subject:
  - Randomly select half the repetitions of baseline typing and half the repetitions of stress typing for training
  - Use remaining baseline and stress typing for testing
  - Repeat 100 times with different random draws
  - Report average classification accuracy
    - (average # reps correct / total reps)

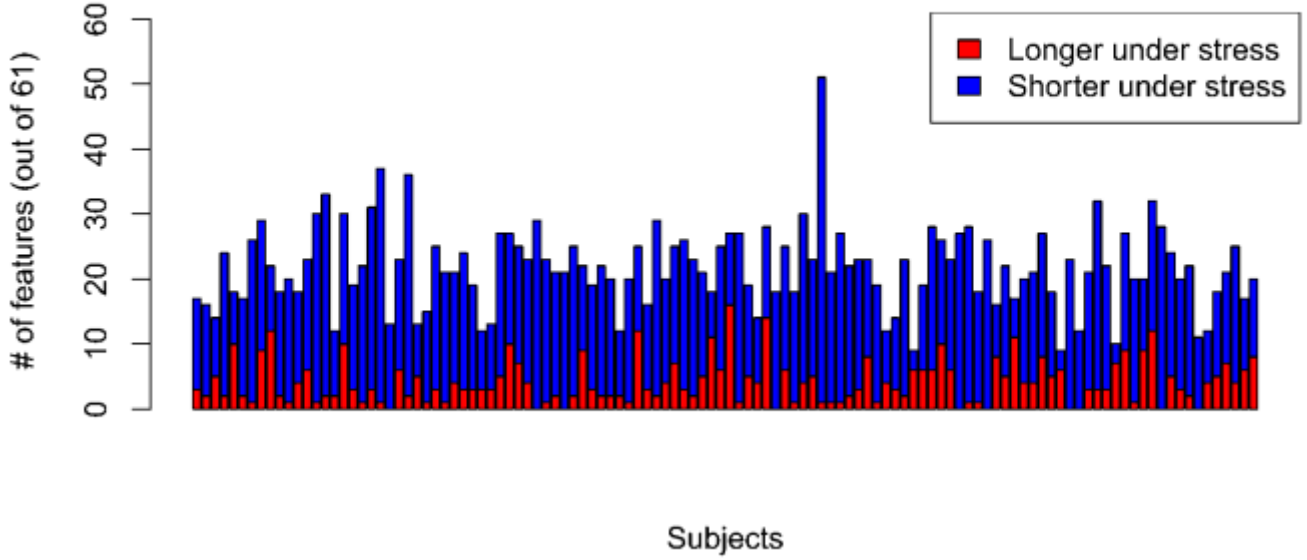
Q1: Within-subject typing changes due to stress



- Average classification accuracy is 89.5%

Q1: Multiple markers for each subject

- Marker: any feature whose mean changes by more than 10% between baseline and stress



Q2: Across-subject markers for stress (Setup)

- Goal: Identify universal markers for stress

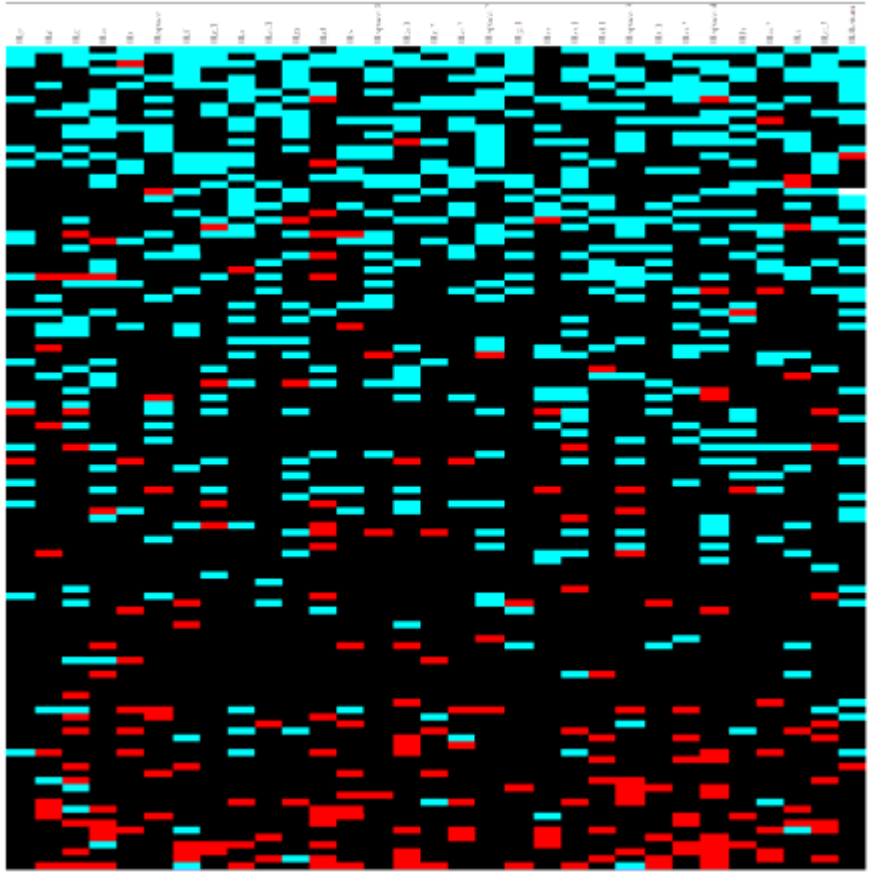
- Objective is to classify a subject's typing into neutral or stressed WITHOUT access to prior data from that subject
- Leave-one-out evaluation paradigm:
  - Use data from *k* – 1 subjects as training data
  - Test on the remaining subject

Q2: Poor classification results

- Results are poor (59% accuracy averaged over all subjects)

- Why…?
  - Successful across-subject classification requires universal markers
  - There do not appear to be universal markers

Q2: Lack of universal markers



- Each row is a subject
- Each column is a keystroke feature

- Blue = marker that is shorter under stress
- Red = marker that is longer under stress
- Black = not a marker

- Universal marker would be a single column that is all blue/red
- No universal markers exist!

Q3: Identify groups of subjects with common markers (Setup)

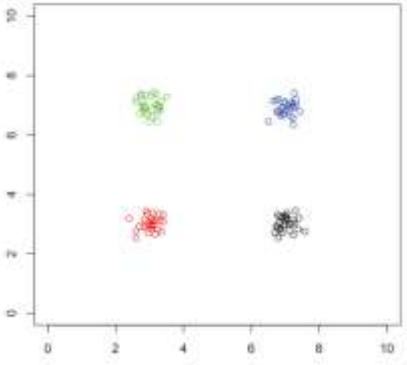
- Goal: Identify groups of subjects with common markers
- Since there are no universal markers, this is the best we can hope for
- Ran three different clustering algorithms
  - K-medoids
  - **Agglomerative clustering (Agnes)**
  - Locally-linear embedding followed by k-medoids
- Represent subjects by their markers
  - Change between neutral and stressed feature means

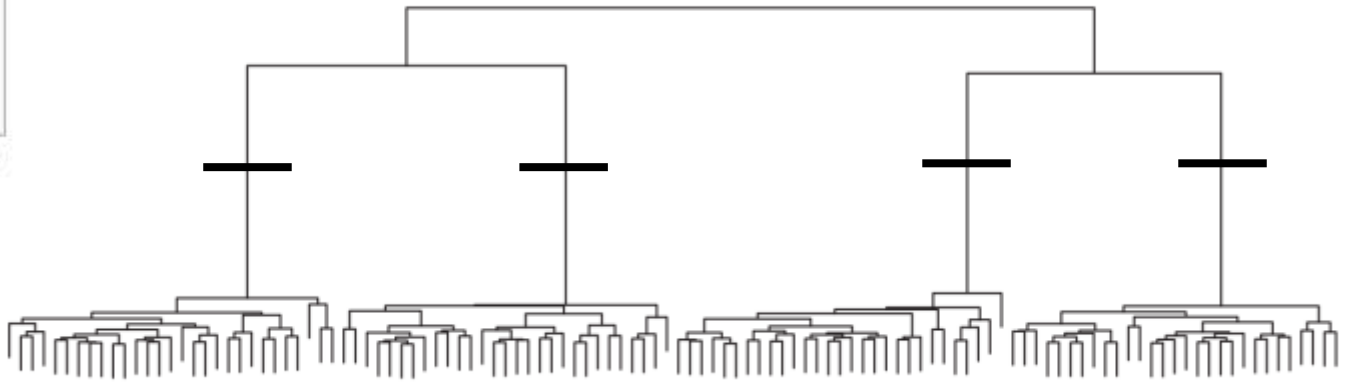
Agnes clustering

- Iterative algorithm
- Start with each data point in its own cluster
- At each iteration, merge the two clusters that are most similar
  - Similarity between two clusters is defined as the average Euclidean distance between all pairs of points in the two clusters
- Stop when all data points are merged into a single cluster

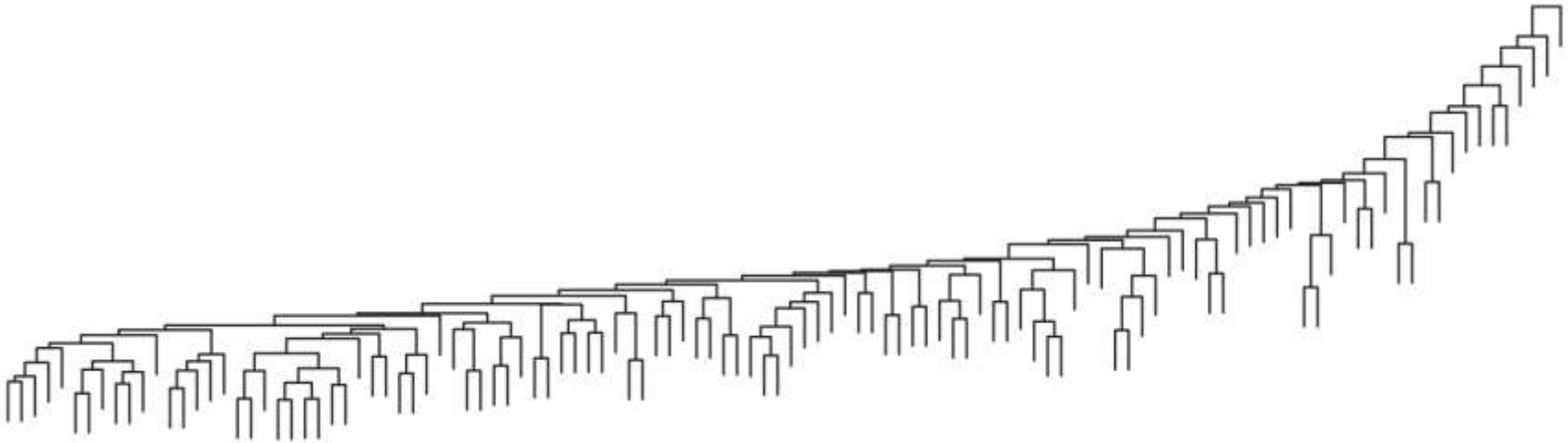
- Simple, interpretable algorithm

Agnes on a clusterable dataset





Q3: Clustering results



Agenda

- Background
- Our experimentation principles
- 3 research questions
- Experimental details
- Results – answering the 3 research questions
- **Takeaways**

Summary of specific results

- Question 1 [Within-subjects]:
  - Classified between neutral and stressed typing within-subject with an average accuracy of 89.5%
  - Successful because each subject had at least 9 markers for stress
- Question 2 [Across-subjects]:
  - Poor performance (< 60%) at across-subject classification
  - Explained by a lack of universal markers
- Question 3 [Clustering]:
  - No evidence of clusters of subjects with shared markers
  - Suggestion that stress manifestations are strongly individualized

How do we move towards using keystroke dynamics for insider detection in the real world?

- Relax some of the restrictions placed on this experiment to create a more realistic environment
  - Allow users to use their own computers
  - Create a task that is more realistic than repeated password typing
  - Create a fake “insider task”?

- Is stress detection still possible under those more realistic environments?

- Establish a threshold for performance required for this technique to be useful in a real-world environment

Final takeaways

- This was just an example of a single cybersecurity experiment
- We designed this experiment by keeping “productive failure” and “creating understanding” in mind
- Productive failure
  - Making sure that we could learn something from our experiment even when we do not obtain the desired outcome
- Creating understanding
  - Explain in simple terms why we obtained the results that we did

Contact Us

Carnegie Mellon University

Software Engineering Institute

4500 Fifth Avenue

Pittsburgh, PA 15213-2612

412-268-5800

888-201-4479

info@sei.cmu.edu

www.sei.cmu.edu