Using Data Analytics and AI to Support Test Security of Remotely Proctored Tests

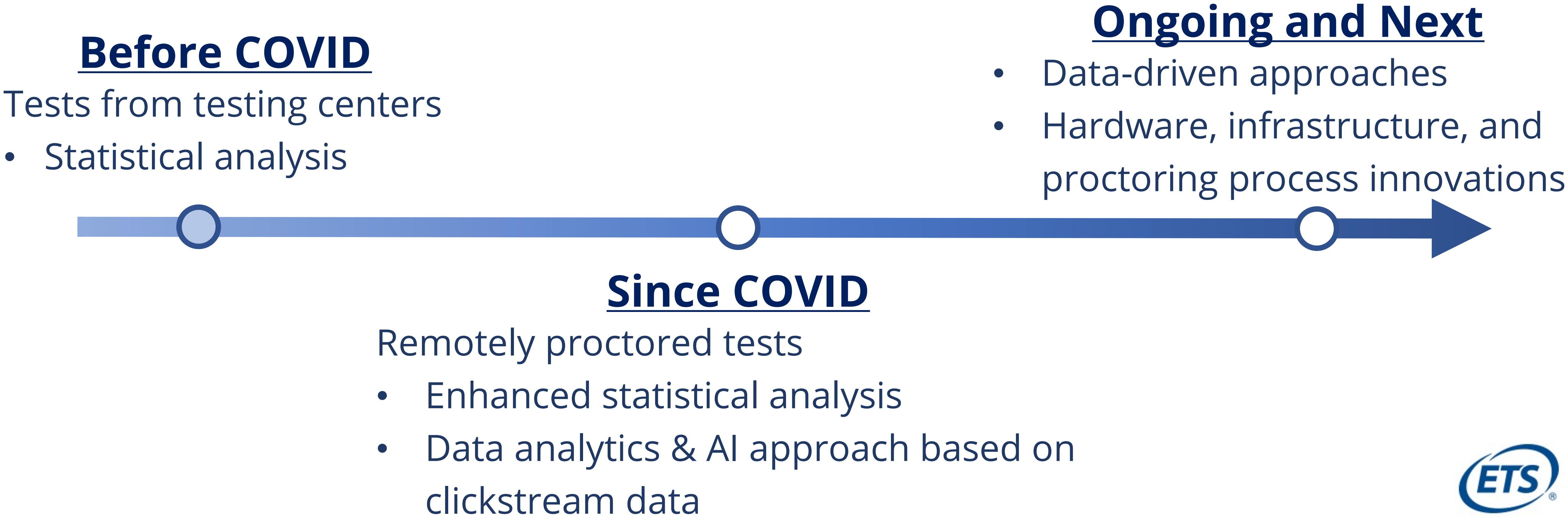
Jiangang Hao

Psychometric and Data Science Modeling @ ETS

20th MARC Conference – Nov. 3rd 2022

Overview

- ETS has tremendous expertise on statistical analysis, data science and AI
- Long history of research on monitoring test security



Before COVID

Tests from testing centers

- Statistical analysis

Since COVID

Remotely proctored tests

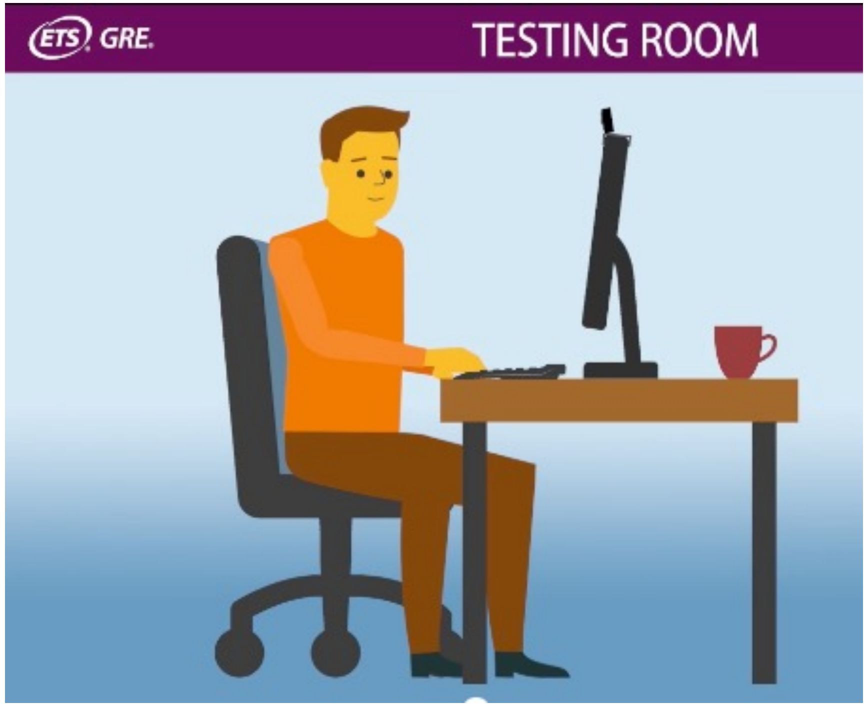
- Enhanced statistical analysis
- Data analytics & AI approach based on clickstream data

Ongoing and Next

- Data-driven approaches
- Hardware, infrastructure, and proctoring process innovations

Remotely Proctored Admins

- ETS started offering at-home tests with live remote proctors since April 2020
- This makes it possible for many to take the tests during the difficult time, while it is also possible for some unintended test taking behaviors/strategies.
- **Clickstream process data** provide rich information that allows us to know better about what is going on

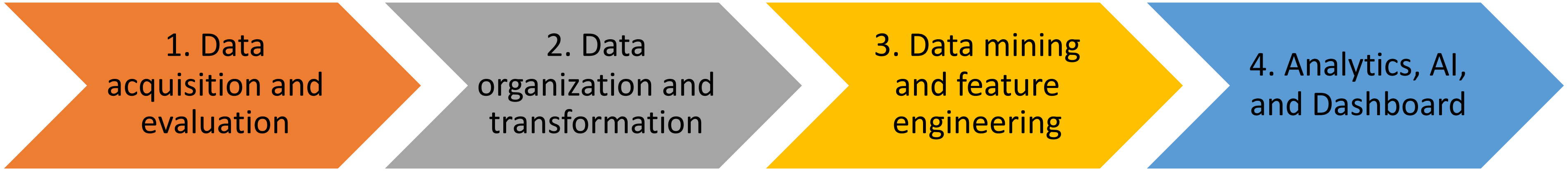


ETS' at-home admins:
https://www.ets.org/gre/revised_general/register/at_home

Data Analytics and AI Approaches

- **Data analytics**
  - Develop features to characterize the process
  - Identify patterns of unintended test-taking behaviors
  - Essay similarity, speech similarity, etc.
- **AI/machine Learning**
  - Mapping features to unintended behaviors
  - Detection of remote computer access
  - Keystroke biometrics, detection of draft writing, etc.
- **Provide Important information**
  - Uncover new cheating schemes
  - Confirm cheating detected from other means
  - Complementary to existing statistical/psychometric analysis on scores

General Workflow

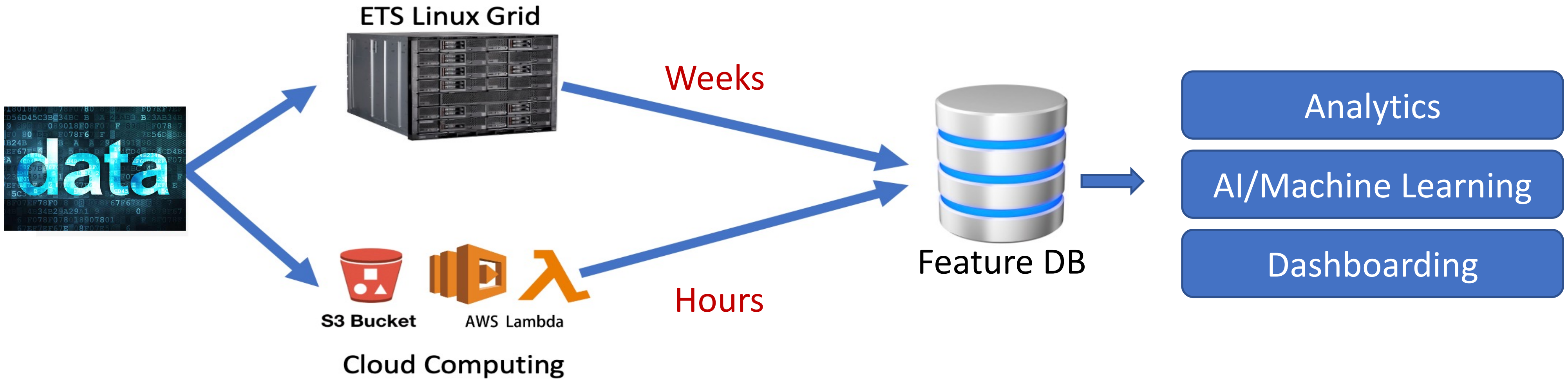


1. Data acquisition and evaluation

2. Data organization and transformation

3. Data mining and feature engineering

4. Analytics, AI, and Dashboard



Analytics

AI/Machine Learning

Dashboarding

Research Projects

THE 2022 CONFERENCE ON TEST SECURITY

OCTOBER 26-28, 2022
Princeton, NJ

NOVEMBER 9-10, 2022
Virtual

Coordinated Symposium I (in-person mode) – 10/28/2022

- AutoESD: An Automated Solution to Detect Copied Essays – Novak, Choi, Hao, & Li
- ★ Detection of AI Generated Essays – Yan, Fauss, Cui, & Hao
- Detecting Writing Process Characteristics Associated with Non-Genuine Text Generation – Deane, Zhang, & Hao

Coordinated Symposium II (virtual mode) – 11/9/2022

- ★ Detecting Remote Computer Access using AI and Clickstream Data - Hao & Li
- ★ Benchmark Keystroke Biometrics Accuracy from High-Stakes Writing Tasks – Choi, Hao, Deane, & Zhang
- ★ Detection of Retyping vs. Drafting Through AI-based Methods based on Keystroke Process Data – Zhang, Deane, & Hao

★ The studies I am going to give a high-level introduction today

Study 1

AI-based Detection of Remote Computer Access

AI Detector of Remote Computer Access

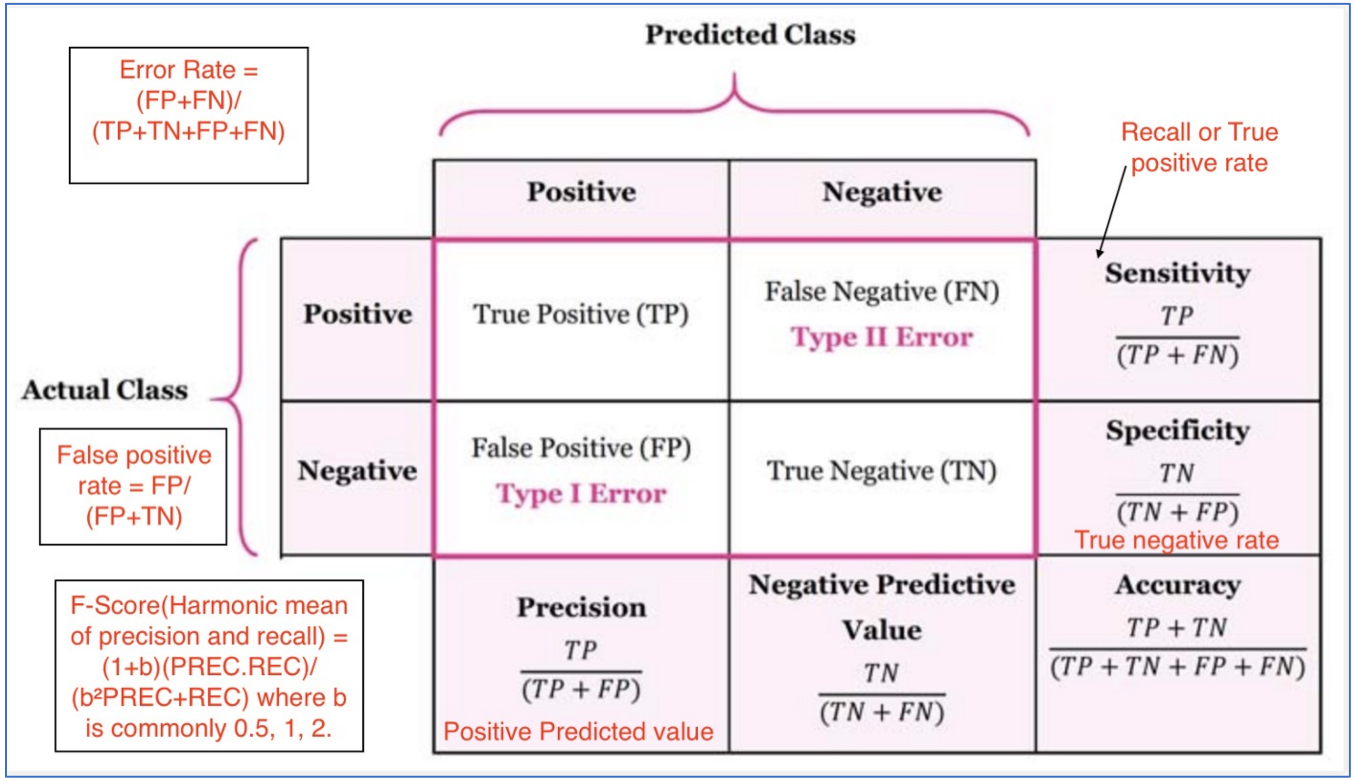
- **Problem statement**: someone remotely accesses the test-taker’s computer and complete the test for him/her.
- **Our approach**:
  - Assume that the clickstream interactions from the RCA sessions are different from a normal session.
  - Learn the relationship between the process-based features and whether it is RCA through AI/machine learning algorithms based on the known RCA cases.
  - Finally, turn the relationship into a detector of RCA and apply it to new data.

Important: AI algorithm only sets some indicators/flags and human experts will review each case carefully to make decision!

Binary Classifier

- The outcomes are in two categories: Positive/Negative

- Computer algorithm will learn the mapping between features and the outcomes



Error Rate = (FP+FN)/(TP+TN+FP+FN)

Predicted Class

		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$ True negative rate
		Precision $\frac{TP}{(TP+FP)}$ Positive Predicted value	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

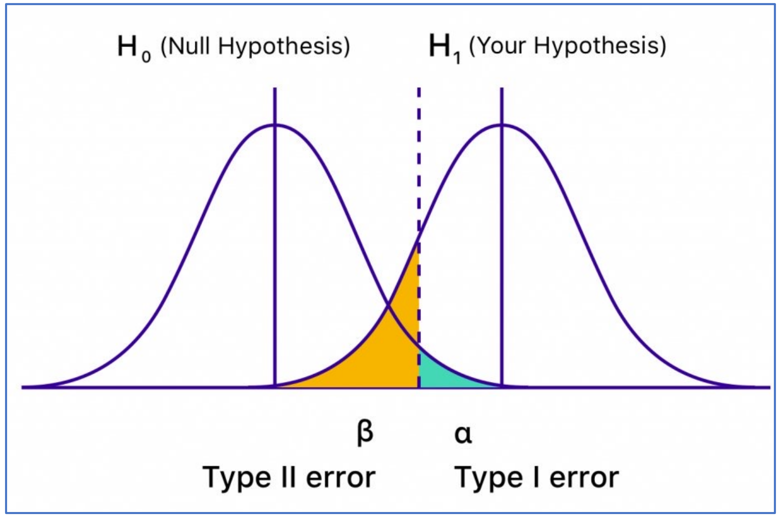
Recall or True positive rate

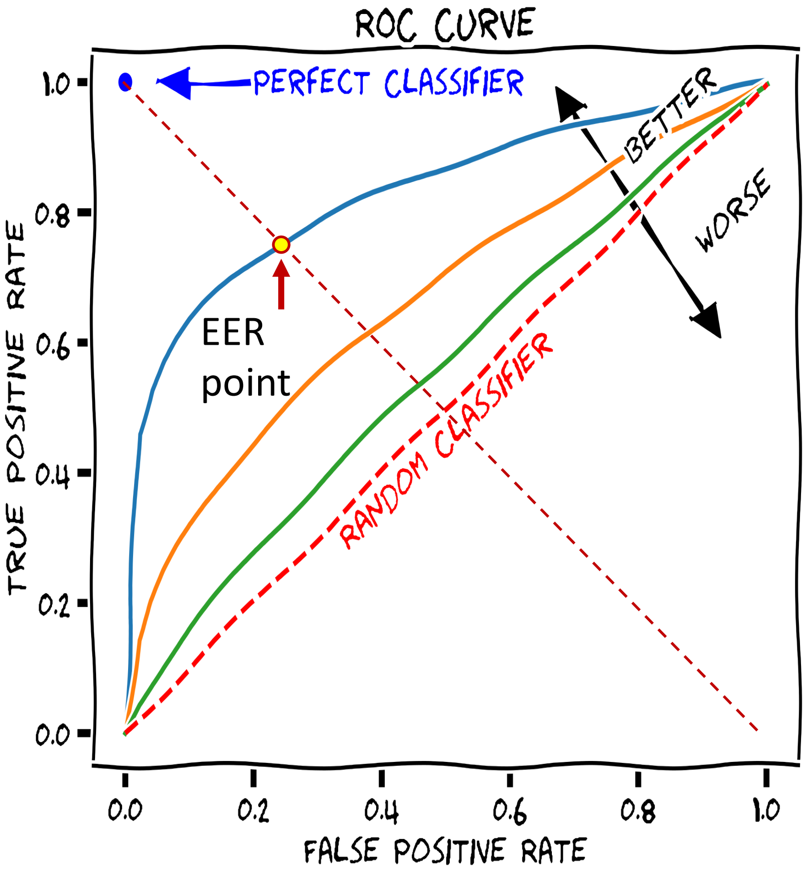
False positive rate = FP/(FP+TN)

F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b²PREC+REC) where b is commonly 0.5, 1, 2.

https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Evaluation Metrics: Receiver Operating Characteristic





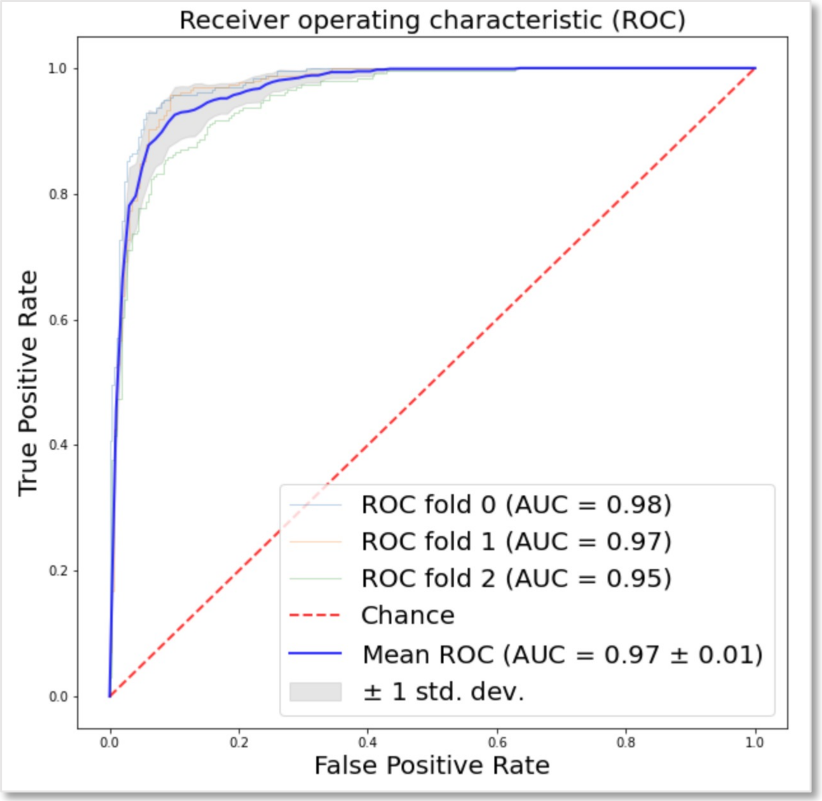
- Area under curve (AUC):
  - 0.5: non discriminative
  - 0.7 – 0.8: acceptable
  - 0.8 – 0.9: excellent
  - > 0.9: outstanding

An Empirical Study

- In one high-stake test, we confirmed about 700 RCA cases. Based on this, we created a balanced training dataset by adding additional non-RCA sessions.
- Extensive data mining work leads to a set of features to characterize different aspects of the process
- Machine learning methods are applied to the features and labels
- Area under the ROC curve from a 3-fold cross-validation is used to evaluate the performance
- This a very typical supervised learning task and can be applied to detecting **other types of unintended behaviors**.

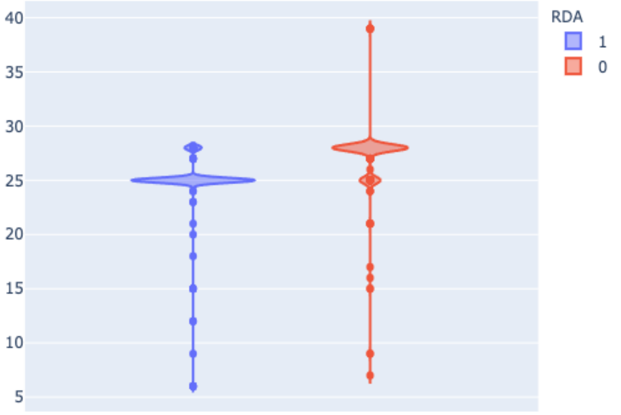
Hao & Li, 2021

Findings

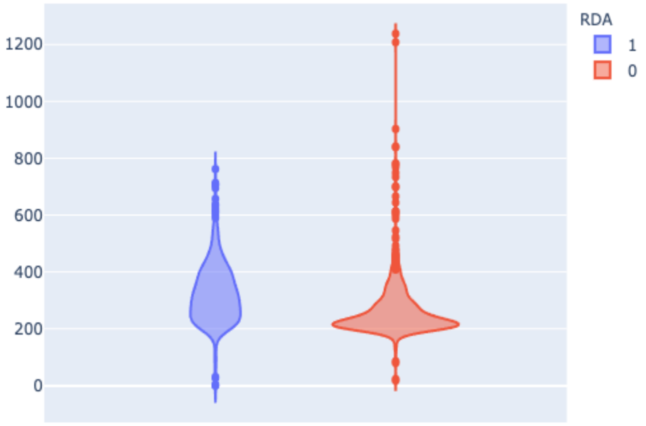


Based on Gradient Boosting Machine

Feature 1



Feature 2



- Gradient Boosting Machine algorithm performs the “*best*”
- Many, instead of one or two, features are needed to achieve the level of classification accuracy
- The detector can be used for both individual case detection and overall trend monitoring.

Hao & Li, 2021

Study 2

Keystroke Behaviors as Biometrics

Keystroke Dynamics as Biometrics

- ETS scientists have been studying keystroke in writing over 15 years, primarily focusing on using it to measure writing proficiency

Deane, 2014
Zhang et al., 2015

- Keystroke as a behavioral biometrics
  - Behavior-based biometrics is not as stable as other biometrics
  - Keystroke behaviors could change when writing different types of essays under different circumstances
  - But in a very specified settings, e.g., writing in a test, keystroke dynamics could be indicative
  - Our empirical study confirmed this

Choi, Hao, Deane, & Zhang, 2019

An Empirical Study

- The Goal: could we identify the same test takers based on the keystroke behaviors?
- Methods:
  - Create a dataset with known repeated test takers (repeaters) and non-repeaters.
  - Develop features to characterize the keystroke process
  - Build machine learning classifiers to detect the repeaters
- Data
  - From the writing task of a high-stake assessment
  - 3,110 repeated test takers (repeaters) from 9/2017 to 8/2018
  - 3110 non-repeaters from the same timeframe

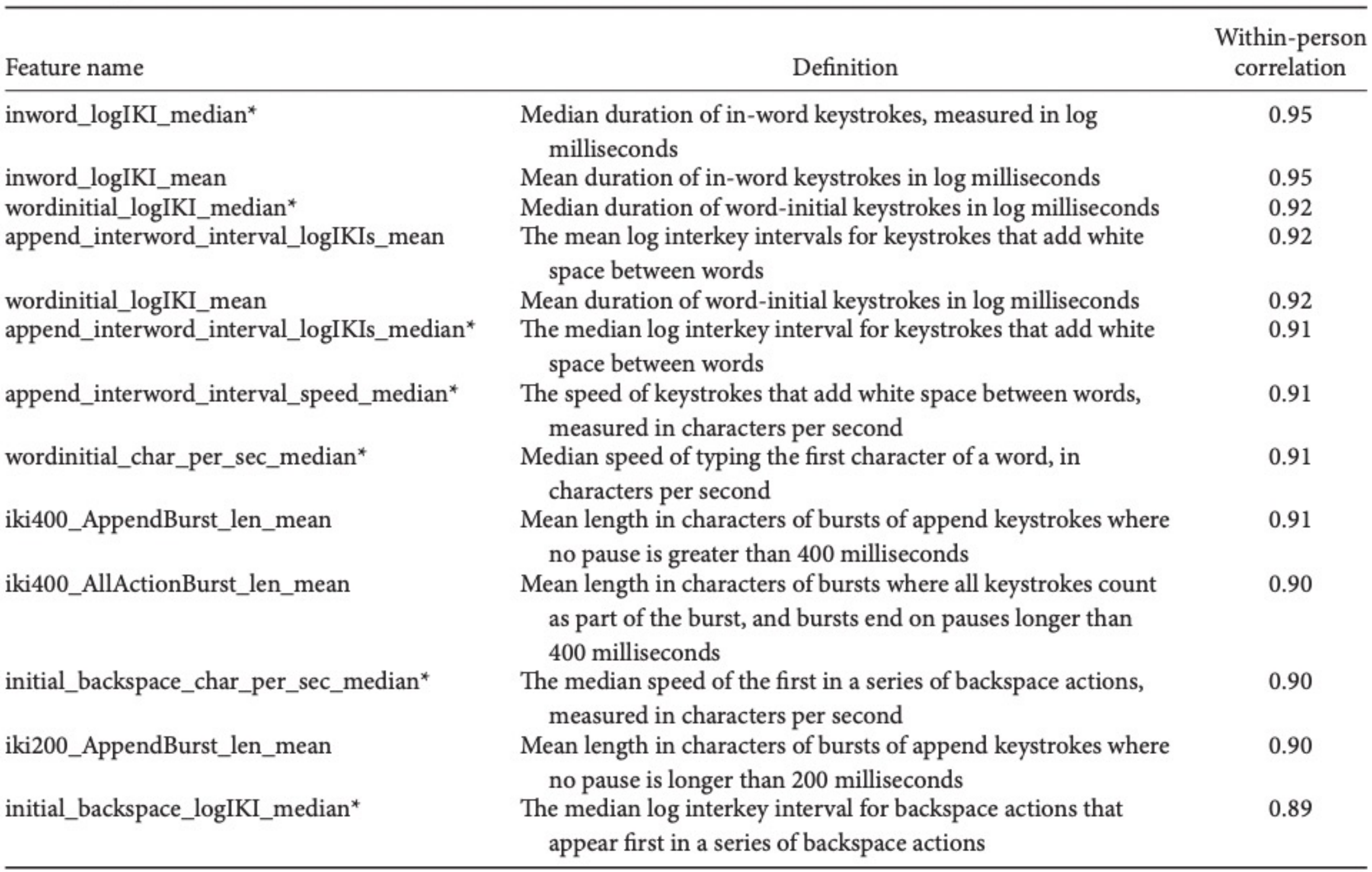
Choi, Hao, Deane, & Zhang, 2019

Keystroke Features

- Writing features
  - The number, latency, speed, and total time spent for specific typing events (measured by log keystroke latency in milliseconds, and in keystrokes per second), including pauses before inserting characters within a word, between words, between sentences, and between paragraphs
  - The number, latency, speed, and total time for initial and repeated backspacing events, cut and paste events, and edits that involved a jump from one location in the text to another
  - Measures of the extent to which words were edited and whether they were correctly spelled before and after editing
  - Measures of fluency of typing, defined in terms of bursts of text production (sequences of keystrokes produced rapidly without a long pause), including the number and length of each type of burst in the test taker’s response
- Di-graph features: summary statistics of the time interval between two adjacent letters.

Choi, Hao, Deane, & Zhang, 2019

Feature Stability



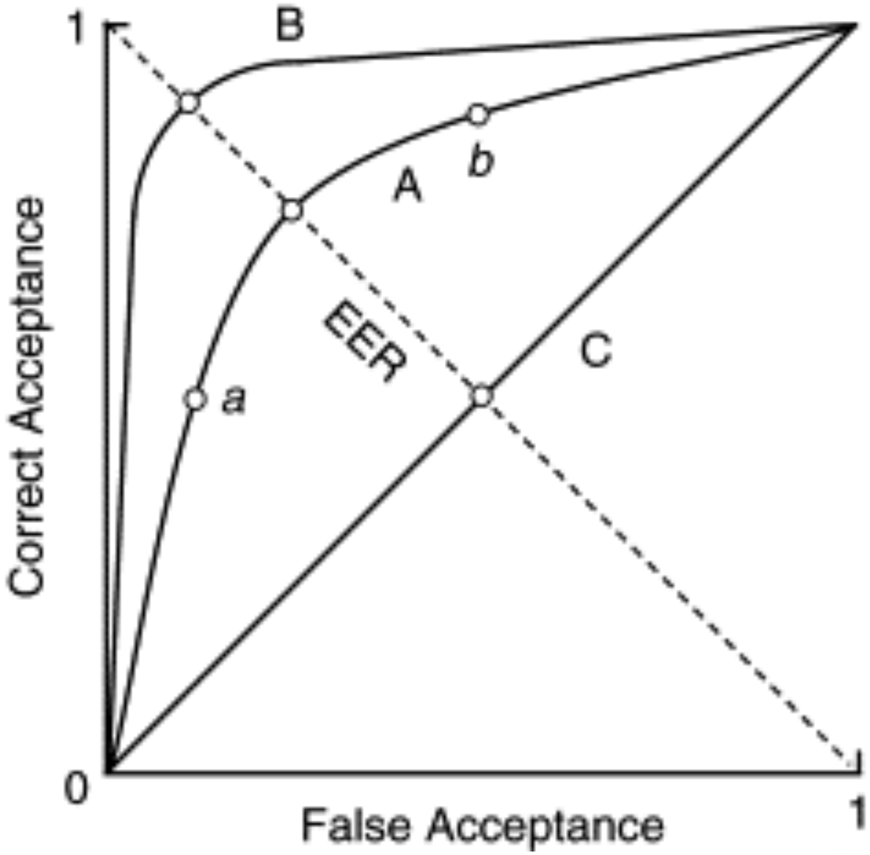
Feature name	Definition	Within-person correlation
inword_logIKI_median*	Median duration of in-word keystrokes, measured in log milliseconds	0.95
inword_logIKI_mean	Mean duration of in-word keystrokes in log milliseconds	0.95
wordinitial_logIKI_median*	Median duration of word-initial keystrokes in log milliseconds	0.92
append_interword_interval_logIKIs_mean	The mean log interkey intervals for keystrokes that add white space between words	0.92
wordinitial_logIKI_mean	Mean duration of word-initial keystrokes in log milliseconds	0.92
append_interword_interval_logIKIs_median*	The median log interkey interval for keystrokes that add white space between words	0.91
append_interword_interval_speed_median*	The speed of keystrokes that add white space between words, measured in characters per second	0.91
wordinitial_char_per_sec_median*	Median speed of typing the first character of a word, in characters per second	0.91
iki400_AppendBurst_len_mean	Mean length in characters of bursts of append keystrokes where no pause is greater than 400 milliseconds	0.91
iki400_AllActionBurst_len_mean	Mean length in characters of bursts where all keystrokes count as part of the burst, and bursts end on pauses longer than 400 milliseconds	0.90
initial_backspace_char_per_sec_median*	The median speed of the first in a series of backspace actions, measured in characters per second	0.90
iki200_AppendBurst_len_mean	Mean length in characters of bursts of append keystrokes where no pause is longer than 200 milliseconds	0.90
initial_backspace_logIKI_median*	The median log interkey interval for backspace actions that appear first in a series of backspace actions	0.89

Features that are highly correlated between the repeated tests

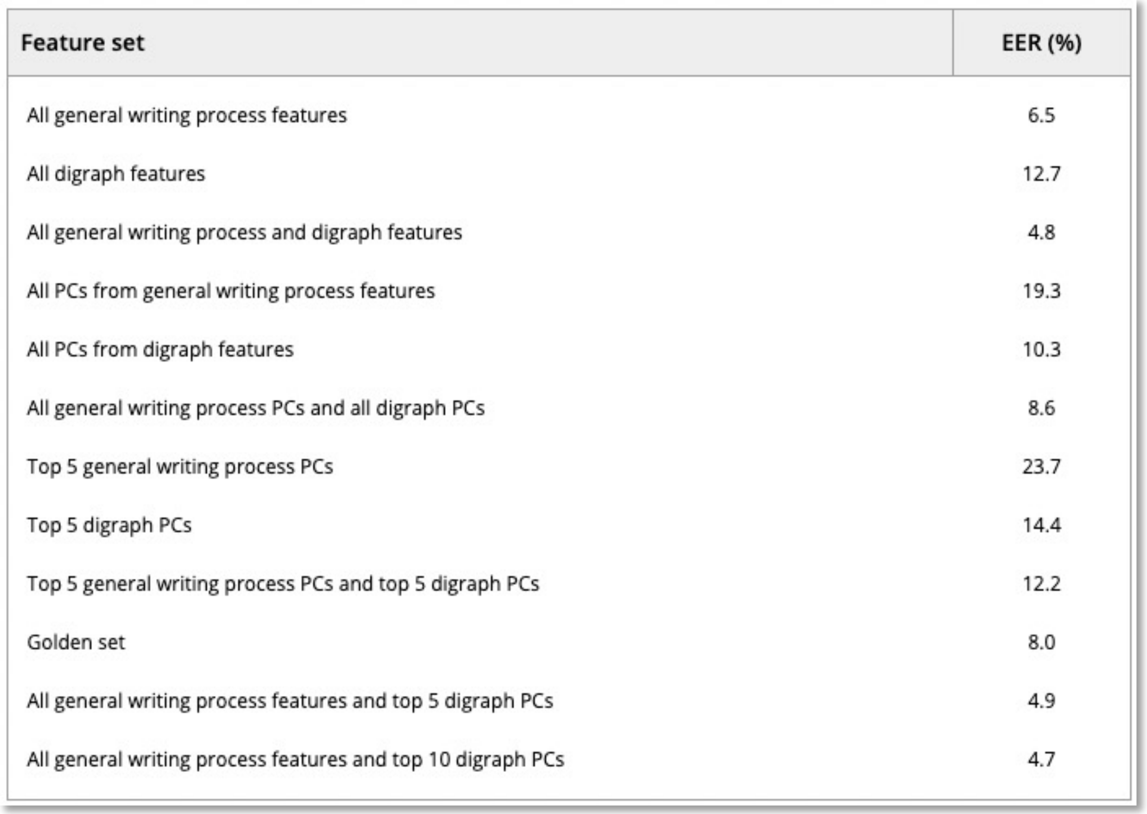
Choi, Hao, Deane, & Zhang, 2019

Biometrics Terminologies

- Terminology remapping
  - Positive -> Acceptance -> Match
  - Negative -> Rejection -> non-match
- False Rejection Rate (FRR) - type II error – False Negative Rate
- False Acceptance Rate – FAR - Type I error – False Positive Rate
- Equal Error Rate: FAR = FRR
  - Voice dynamics: 2%
  - Signature: 2%
  - Fingerprint: 0.2%
  - Iris: 0.01%
  - TouchID: 1/50,000 = 0.002%
  - FaceID: 1/1,000,000 = 0.001%



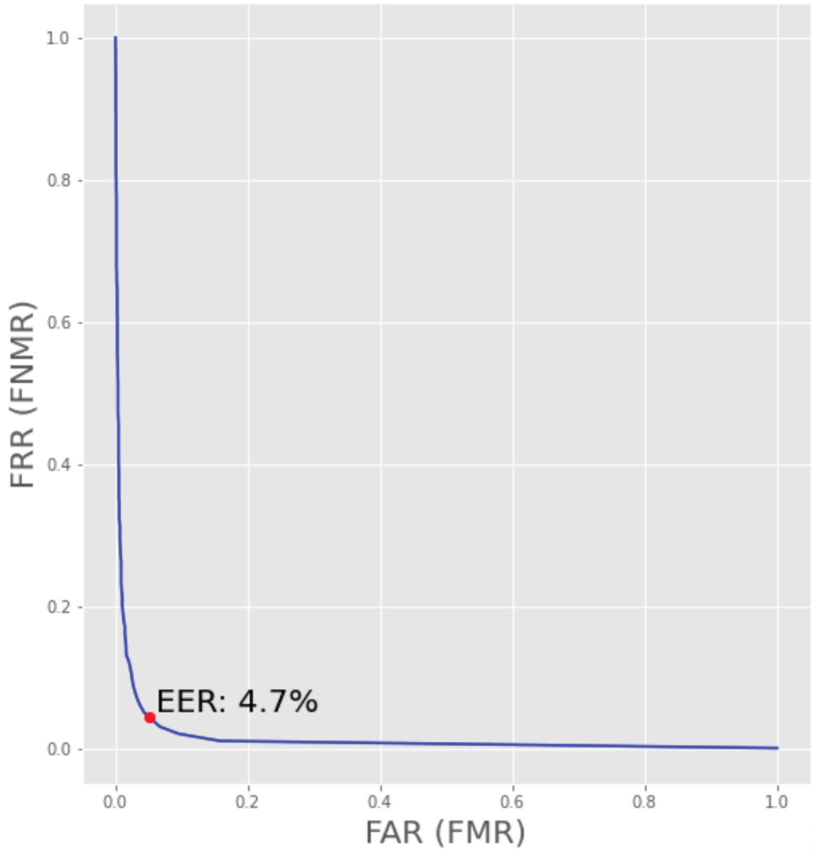
Findings



Feature set	EER (%)
All general writing process features	6.5
All digraph features	12.7
All general writing process and digraph features	4.8
All PCs from general writing process features	19.3
All PCs from digraph features	10.3
All general writing process PCs and all digraph PCs	8.6
Top 5 general writing process PCs	23.7
Top 5 digraph PCs	14.4
Top 5 general writing process PCs and top 5 digraph PCs	12.2
Golden set	8.0
All general writing process features and top 5 digraph PCs	4.9
All general writing process features and top 10 digraph PCs	4.7

- The EER depends on the feature set used in the ML model
- Gradient Boosting Machine gives the best result

Equal Error Rate: 4.7%



Choi, Hao, Deane, & Zhang, 2019

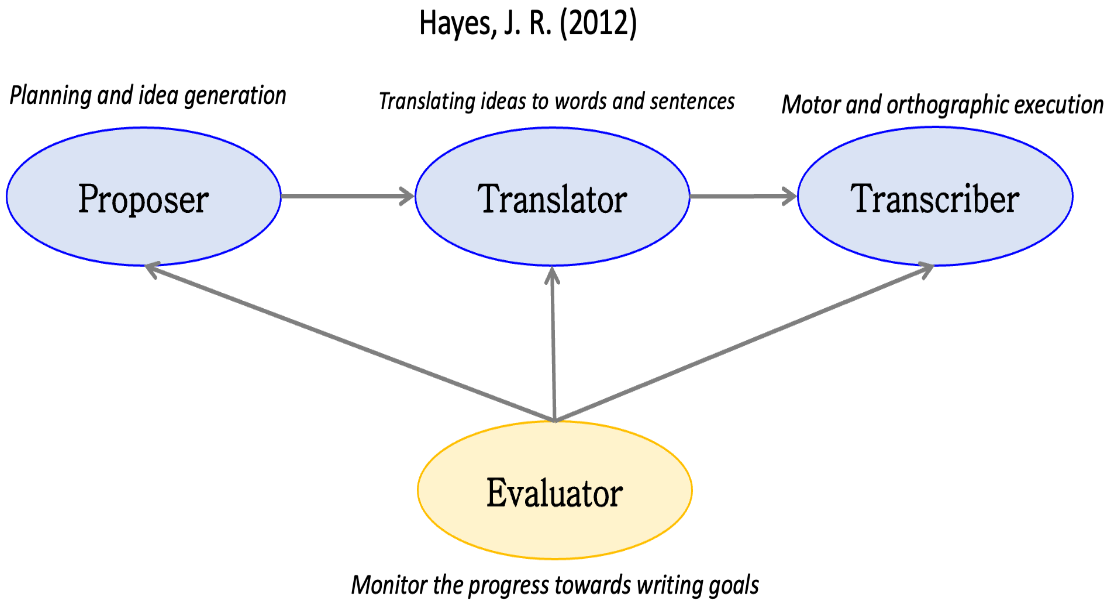
Study 3

Detection of Copy-typing and Draft Writing

Draft Writing vs. Copy Writing

- Draft writing and copy writing have different cognitive processes, which leads to different writing processes
- Keystroke captures fine-grained writing process information
- Can we build an AI classifier to classify draft writing and copy writing based on the keystroke features? What classification accuracy we can achieve?

Cognitive process of writing


Hayes, J. R. (2012)
Planning and idea generation
Translating ideas to words and sentences
Motor and orthographic execution
Proposer
Translator
Transcriber
Evaluator
Monitor the progress towards writing goals

An Empirical Study

Data

- Data were collected from 8th grade students in an urban middle school in a state in the American West.
- This school has a population of students who identify primarily as being from minority groups and from households with low socioeconomic status:
  - 72.1% free and reduced-price lunch;
  - 20% limited English proficiency;
  - 60.3% Hispanic, 27.7% Black, 4.8% White, 2.6% Asian.
- N = 201, each takes both instruments

Instruments

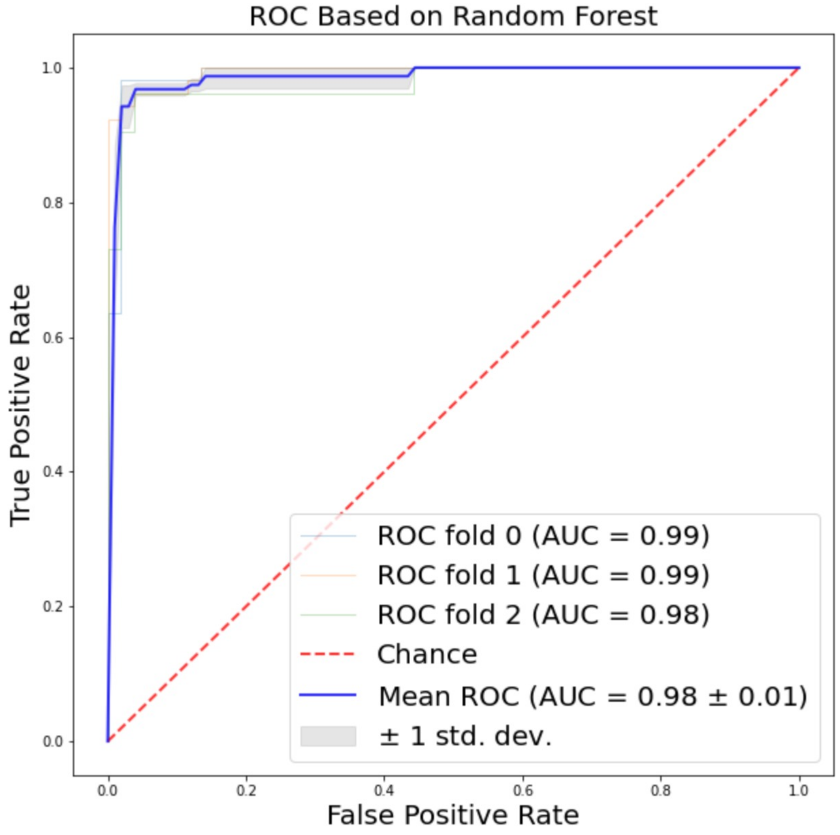
- Re-typing task:
  - Students were asked to retype the article "Mark Twain's Huckleberry Finn"
  - Allotted time: 30 minutes
  - Administered online, keystrokes collected
- Essay drafting task:
  - ETS CBAL® formative assessment module: Junk food
  - Extended essay task is allotted for 30 minutes
  - Administered online, keystrokes collected

Zhang, Hao, & Deane, In prep.

Findings

Best classification result is from Random Forest

Model	Precision	Recall	F1	Accuracy
Support Vector Machine	0.987	0.872	0.924	0.929
Random Forest	0.987	0.962	0.974	0.974
Gradient Boosting Machine	0.949	0.949	0.949	0.949
XGBoost	0.962	0.987	0.974	0.974



Zhang, Hao & Deane, In prep.

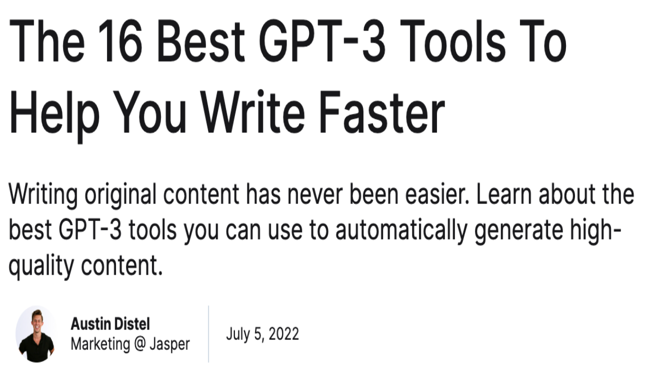
Important Note

- Behavioral biometrics is not as stable as other biometrics
- Many factors may affect the behaviors
  - Keyboards, computers, etc
  - Tasks and environment, etc.
- Should be used in combination with other measures
- Need to constantly monitor the drift of the model in practice
- Valuable for monitoring trends

Study 4

Detection of AI-generated Essays

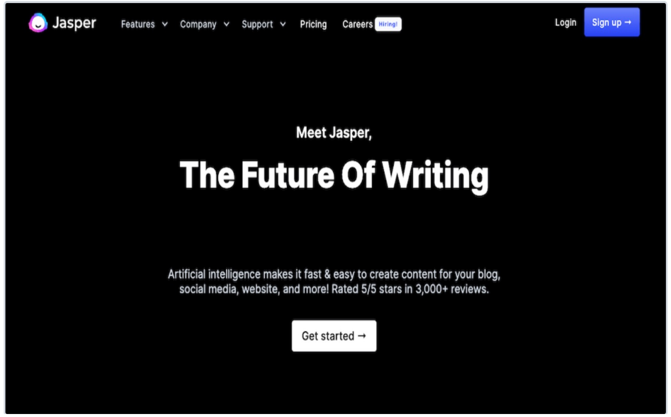
Writing in the AI Age



The 16 Best GPT-3 Tools To Help You Write Faster

Writing original content has never been easier. Learn about the best GPT-3 tools you can use to automatically generate high-quality content.

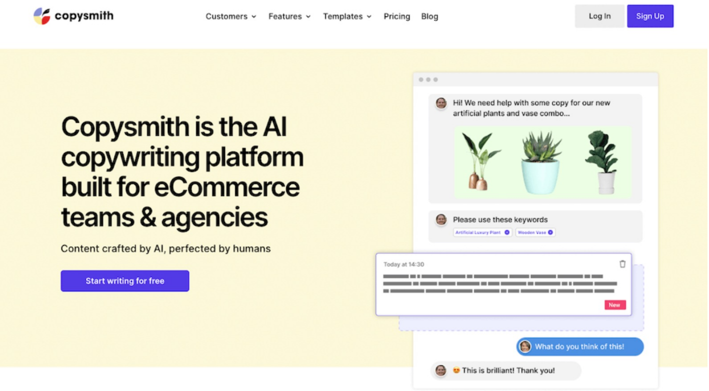
Austin Distel
Marketing @ Jasper | July 5, 2022



Meet Jasper,

The Future Of Writing

Artificial intelligence makes it fast & easy to create content for your blog, social media, website, and more! Rated 5/5 stars in 3,000+ reviews.

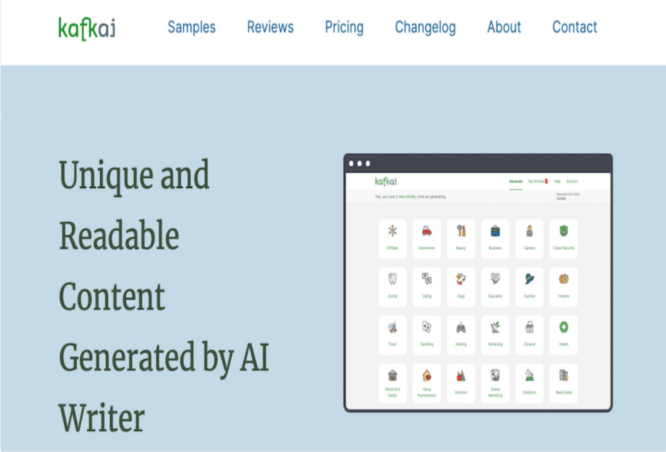


copysmith
Customers
Features
Templates
Pricing
Blog
Log In
Sign Up

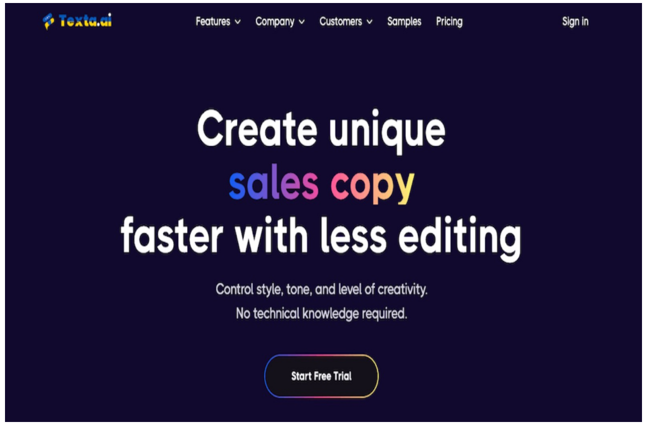
Copysmith is the AI copywriting platform built for eCommerce teams & agencies

Content crafted by AI, perfected by humans

Start writing for free
Hi! We need help with some copy for our new artificial plants and vase combo...
Please use these keywords
What do you think of this!
This is brilliant! Thank you!

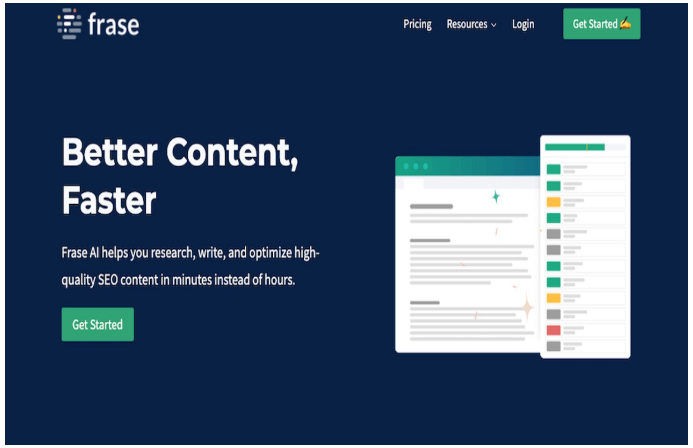


Unique and Readable Content Generated by AI Writer



Create unique sales copy faster with less editing

Control style, tone, and level of creativity.
No technical knowledge required.



Better Content, Faster

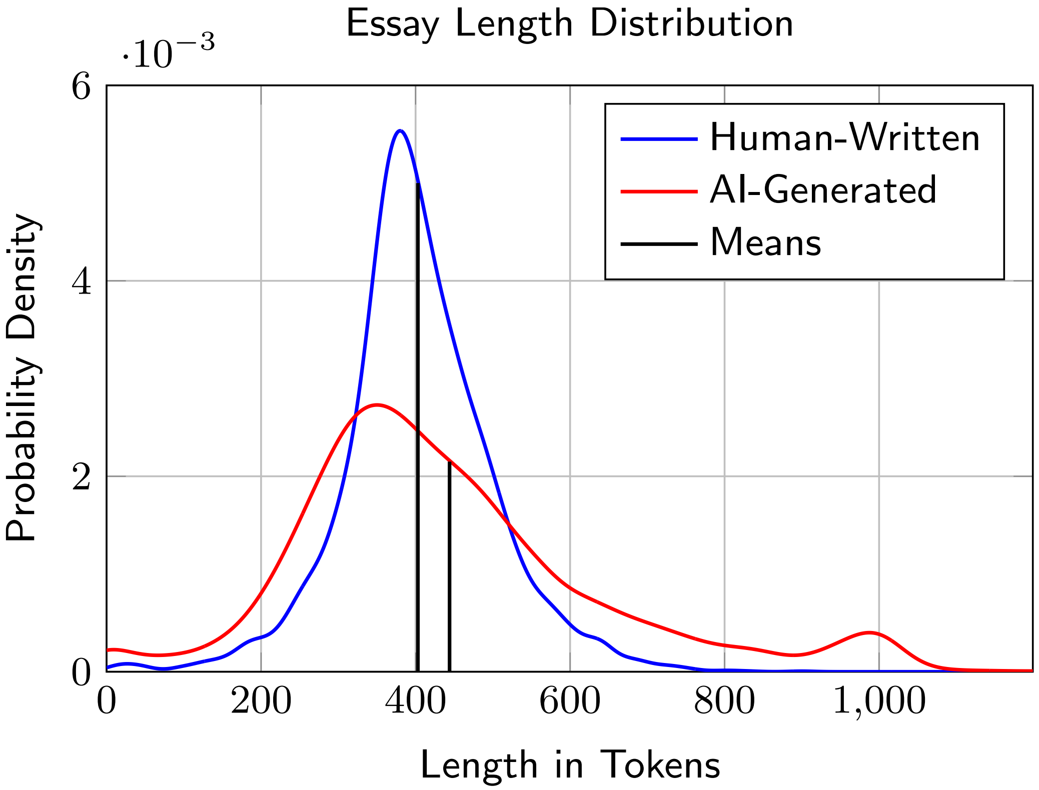
Frase AI helps you research, write, and optimize high-quality SEO content in minutes instead of hours.

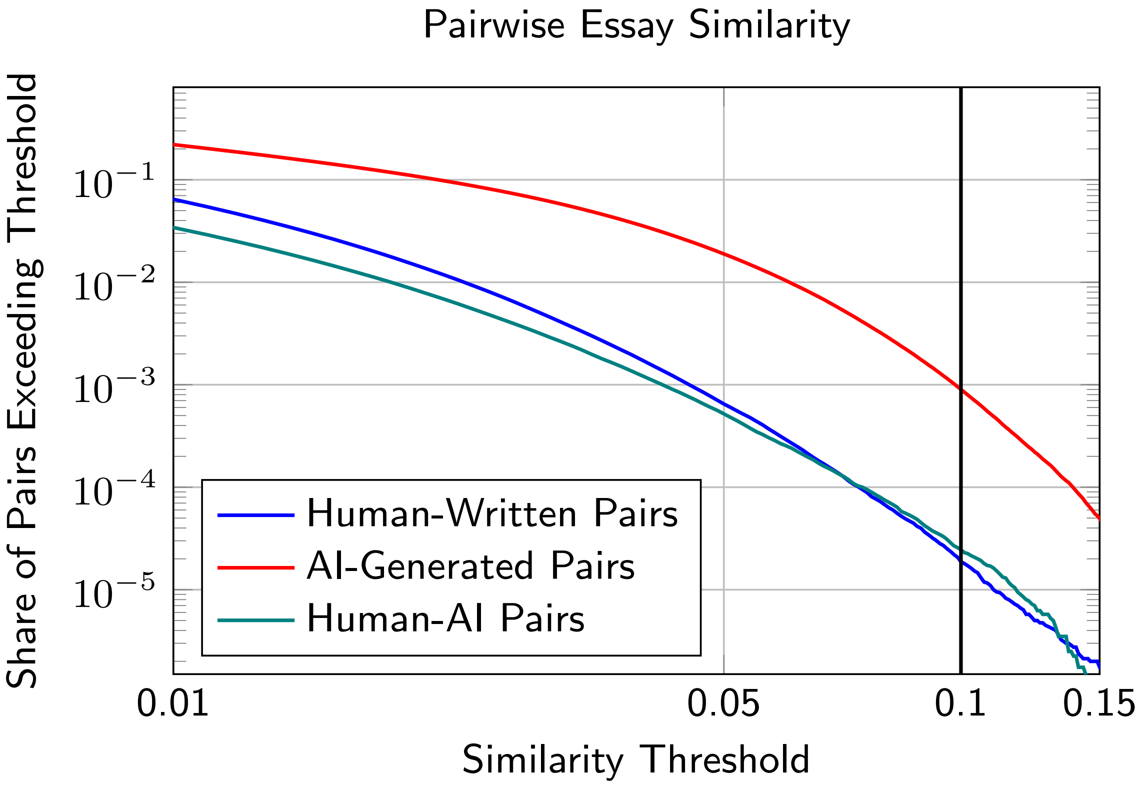
https://www.jasper.ai/blog/gpt3-tools

AI Generated Essays

- It is plausible to think that test takers may use AI generated essays in writing assessment
- Can we detect them?
- An empirical study
  - Use GPT-3 generate 4000 essays based on 4 open prompts of a high-stake writing tests
  - Add artificial typos to each of the generate essays
  - Sample 4000 human written essays from the same 4 prompts at different score levels
  - Build AI classifier to detect AI-generated essays

AI-generated Essay vs. Human-written Essay





Plots from Michael Fauss

AI Classifier

- Method 1: transfer learning – based classifier
  - Pre-trained LLM (RoBERTa)
  - Fine-tuned with our training data (60%, 20%, 20% split for training, validation and test set)
  - Accuracy of classification – 99.5%
  - Black box
- Method 2: hand engineered features + Linear SVM
  - Features are from ETS e-rater engine (~190 features)
  - Accuracy of classification – 95%
  - Semi-black box

Some Insights

- AI-generated essays have fewer
  - Grammar errors/typos
  - Pronouns (he, she, him, her, …)
  - Conjunctive adverb (therefore, thus, etc.)
  - Passive voice
- These observations may be changed for different LLMs and we are developing benchmarks against mainstream LLMs.
- The findings will inform our assessment development to ensure new items are not easily hacked by AI

Summary

- The clickstream process data contain important information about the test-taking process and play an important role for ensuring the quality of our high-stake tests.
- The findings can provide feedback to improve the proctoring protocols and assessment design to reduce the unintended behaviors
- Synergy with psychometric and statistical analyses and other efforts is crucial.
- **Important: AI/analytics only set some indicators and human experts will review each case carefully to make decision!**

Email: jhao@ets.org

Acknowledgement

We thank ETS Research Allocation and Enterprise Test Security Initiative for funding the presented research