Mental Health Markers in Language and Brain Data: Potential Diagnostic Use and Privacy Concerns

Denisa Qori McDonald
Drexel University
Philadelphia, PA 19104, USA
denisa.qori@drexel.edu

Girija Kaimal
Drexel University
Philadelphia, PA 19104, USA
gk27@drexel.edu

Rachel Greenstadt
Drexel University
Philadelphia, PA 19104, USA
rachel.a.greenstadt@drexel.edu

Erin T. Solovey
Drexel University
Philadelphia, PA 19104, USA
erin.solovey@drexel.edu

Abstract
This paper discusses technology's potential role in the inadvertent leaking of information related to mental health conditions, a particularly sensitive and legally protected part of one's identity. Social media as well as emerging technologies such as brain-computer interfaces (BCIs) are changing the way that we interact with each other and the world. They also offer new windows into deeply personal and previously private aspects of our identity, and may leak this information. For example, with the vast amounts of public data on sites such as Twitter, we can now identify individuals even when they are using anonymous user accounts [35]. The internal state of an individual's neural activity is, in many senses, the most private of personal data. Until recently it was impossible or highly inconvenient to gain access to this type of information. With these barriers lowering rapidly, it is critical to look carefully at the privacy implications.

Author Keywords
mental health; design guidelines; privacy; discrimination; diagnosis; mental health detection; brain data; text data; fNIRS; stylometry; BCI.

Introduction
Functional near-infrared spectroscopy (fNIRS) brain imaging is increasingly being used in mental health clinical settings [18, 50, 31, 27]. It has the potential for both contin-

ual lightweight monitoring and in diagnosis. At the same time, we have seen recent progress toward consumer-grade BCIs devices that are capable, affordable, safe, and portable, moving their use out of clinical settings and into new domains. There are proposals to leverage these devices' ability to detect various cognitive states for applications such as authentication [10, 47], games, learning [46, 2, 20], and monitoring cognitive load in high-stress environments [4]. The fact that the same devices can be used for clinical purposes as well as entertainment and productivity purposes leads to a critical question: What personal health information might be leaked as people use BCIs in work or recreation? To facilitate the respectful usage of such technology, we must proactively study the feasibility of attacks on privacy, and explore potential mitigations.

Parallel to these developments, there has been increasing amounts of research in stylometry and sociolinguistics aimed at detecting and diagnosing mental health conditions from text, especially text on social media [11, 14, 12].

Diagnostic & Side Channel Potential of Language

Stylometry is the use of statistical analysis of written language, to uniquely identify a person. Features used may include the lengthe of words, pairs of words, vocabulary usage, sentence structure, etc. Several resources give an overview of stylometry methods [51, 29]. Recent developments have led to robust classifiers using machine learning and other AI techniques [21, 49, 51]. This field has advanced sufficiently that its findings are routinely accepted as evidence in court [8].

The study of how language is reflected in mental health is a relatively new, but active area of study. Initial, pioneering work was done to quantify the psychometric properties of language use [38], and to build a program Linguistic Inquiry and Word Count to help in analysis. Language use has been shown to be a precursor of cognitive decline in studies of popular authors writing over time [52] and in the longitudinal Nuns study [42]. More recently, researchers have shown that it is possible to detect depression [13], PTSD [15], suicidal ideation [14], and many other mental health conditions via text published on social media platforms. In fact, for the past three years, there has been a workshop on Computational Linguistics and Clinical Psychology attached to the North American Association of Computational Linguistics (NAACL) conference devoted to the subject.

Stylometry is a wide field, and besides being recently used to explore the relationship between language and mental health, it has also been largely focused on profiling authorial traits [3, 16, 28, 33] such as gender, age, education, or even authorial personality [33, 40, 41]. For example, the detection of native language and language family from English text has been explored [45]. Some of this work has been used to predict sensitive information such as attrition in organizations [37]. *The ability to detect private information presents a significant privacy concern*, which will have implications as these methods are used for diagnosing mental health conditions such as Alzheimer's, depression, psychopathy, even suicide risk.

Diagnostic & Side Channel Potential of BCIs

Several techniques measure the changing state of the brain: functional magnetic resonance imaging (fMRI), positron emission tomography, electroencephalography (EEG), and magnetoencephalography and functional near-infrared spectroscopy (fNIRS). EEG and fNIRS are the two main methods that have seen adoption outside of clinical settings due to their portability, relative low cost, and safety. EEG detects electrical impulses coming from the neurons firing in the brain, while fNIRS measures blood flow and blood oxygen changes related to the hemodynamic response, and is more similar to fMRI. The methods are complementary and can be measured simultaneously.

Research in mental health settings using fNIRS indicates that it could serve as a diagnostic tool for illness conditions related to affective disorders as well as disorders of self regulation. There is emerging evidence of biological correlates of mental status. For example, PTSD has been found to affect the speech production center of the brain and depression and schizophrenia have been associated with atypical activity [18, 26, 50]. fNIRS has been found to be related to specific brain pathways of reward perception in eating disorders, reduced pre-frontal cortex activation in depression and differential activation in schizophrenia [26, 31, 50]. These markers can be considered a brain signature for disorders.

Brain data can pose privacy threats as well. There has been considerable interest in its use as a biometric authenticator. Both fNIRS and EEG technology have been used as authentication metrics achieving high accuracies [36, 19, 39, 9]. Thorpe et al. [47] also researched the feasibility of using BCI for authentication, bringing up the ethical and privacy concerns of developing such a system. Venkatasubramanian et al. [53] show that certain physiological data is unique enough to be used to create an encryption key for sending patient data. Inherent in this is the fact that each person has a unique brain signature. Researchers have also explored side channel leakage of private information in BCI settings. For example, researchers have demonstrated the future potential for co-opting an EEG setup intended for one purpose, such as gaming, to extract unrelated, potentially private information [30].

Potential Scenarios in Mental Health

Research on reward pathways for eating disorders [32, 34] shows distinct patterns of participant gender and preferences which act as an individual signature that is objectively determinable using fNIRS. These markers can be considered a brain signature for disorders and potentially misused and abused by limited privacy protections. For example, the wireless fNIRS sensors available now could detect variations in reward perception leading to selective marketing to individual with a predisposition to addictive behavior and limitations in self regulation. Similarly, hemodynamic indicators of mood and affective disorders could be misused to discriminate and target patients. It might also be used to challenge diagnoses based on behavior or self report measures alone. For example, a patient's request for disability or treatment might be challenged if a parallel fNIRS brain signature does not indicate the behavioral symptoms. This data might be collected as part of routine care without adequate informed consent. Patients need to be aware of the limits of privacy when participating in clinical procedures and the use of such data. Of particular note is the concern around the digital data generated from fNIRS. If not adequately protected, the digital data on hemodynamic responses could be manipulated and misrepresented by malicious individuals, as well as, for-profit ventures agencies seeking to make illegal profits including insurance companies, legal agencies and healthcare professionals. An additional issue with data generated by fNIRS is the archiving and secondary analysis of large databases as well as individual files for further analysis and research. Patients often might not know if and how their data might be used and what personal information might be shared as part of research or business and marketing efforts.

Users expose their language and brain data whenever they use brain-computer interfaces or enter textual information to online platforms, opening the possibility of the data being used to infer private information about them. This might lead to unequal treatment, including in the workplace, in online forums, and less access to benefits that they would otherwise have access to.

Toward Effective and Responsible Systems

Given the potential for detecting mental health state that brain and text data can offer, when machine learning techniques are applied to them, it is important to design and build systems that leverage this capability that help the target population realize that they need help. At the same time, we should prevent mental health information from being inadvertently leaked and used by third parties putting subjects at a disadvantage.

We are currently investigating side channel leaks emerging in the use of brain-computer interfaces and social media, with a focus on mental illness. We combine our backgrounds in studying mental health including depression, disorders of self regulation and post traumatic stress disorders (PTSD) [55, 23, 25, 24, 54], with our work researching implicit brain-computer interfaces [43, 44, 48] and our experience investigating the privacy implications when using machine learning on complex, personal data [5, 6, 1, 7, 17, 22] to study privacy threats and develop appropriate mitigations. The questions described above will inform public debate on technologies that may leak private mental health information. This will enable the consideration of regulations early, before the technologies are widespread and their exploitation has already occurred. In addition, our findings in identifying mental health issues from everyday tasks could be integrated into future work in which these methods are used to inform the appropriate people (e.g. therapist, psychiatrist, or self) about a condition. When used appropriately, such a tool would be valuable in the assessment of mental health conditions and risks. The investigation of the potential disconnect between self-report and brain-report will have implications for brain-computer interfaces more generally. It will provide some early evidence about how much we can rely on brain data and how much people are able to manipulate it.

References

[1] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Damon McCoy, and Rachel Greenstadt. 2014. Doppelgänger Finder: Taking Stylometry to the Underground. In *IEEE Symposium on Security and Privacy (S&P)*.

[2] John R. Anderson, Shawn Betts, Jennifer L. Ferris, and Jon M. Fincham. 2010. Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences* 107, 15 (2010), 7018–7023. DOI: http://dx.doi.org/10.1073/pnas.1000942107

[3] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.

[4] M. Boyer, M. L. Cummings, L. B. Spence, and E. T. Solovey. 2015. Investigating Mental Workload Changes in a Long Duration Supervisory Control Task. *Interacting with Computers* (2015), iwv012–. DOI: http://dx.doi.org/10.1093/iwc/iwv012

[5] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security (TISSEC)* 15, 3 (2012).

[6] Michael Brennan, Stacey Wrazien, and Rachel Greenstadt. 2010. Using Machine Learning to Augment Collaborative Filtering of Community Discussions. In *Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[7] Aylin Caliskan-Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy Detective: Detecting Private

Information and Collective Privacy Behavior in a Large Social Network.

[8] Carole Chaski. 2007. The keyboard dilemma and authorship identification. In *IFIP International Conference on Digital Forensics*. Springer, 133–146.

[9] John Chuang, Hamilton Nguyen, Charles Wang, and Benjamin Johnson. 2013a. I think, therefore i am: Usability and security of authentication using brainwaves. In *International Conference on Financial Cryptography and Data Security*. Springer, 1–16.

[10] John Chuang, Hamilton Nguyen, Charles Wang, and Benjamin J ohnson. 2013b. I Think, Therefore I Am: Usability and Security of Authenticati on Using Brainwaves. In *Financial Cryptography and Data Security*. Springer.

[11] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

[12] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring Post Traumatic Stress Disorder in Twitter.. In *AAAI International Conference on Weblogs and Social Media (ICWSM)*.

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media.. In *ICWSM*. 2.

[14] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2098–2110.

[15] Glen Coppersmith Craig Harman Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. *Proceedings of ICWSM* (2014).

[16] Lex Fridman. 2014. *Learning of Identity from Behavioral Biometrics for Active Authentication*. Ph.D. Dissertation. Drexel University.

[17] L. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam. 2014. Multi-Modal Decision Fusion for Continuous Authentication. *Computers and Electrical Engineering* 41 (January 2014).

[18] M Fukuda and M Mikuni. 2011. Clinical application of near-infrared spectroscopy (NIRS) in psychiatry: the advanced medical technology for differential diagnosis of depressive state. *Seishin shinkeigaku zasshi= Psychiatria et neurologia Japonica* 114, 7 (2011), 801–806.

[19] D Heger, C Herff, F Putze, and T Schultz. 2013. Towards biometric person identification using fnirs. In *Proceedings of the Fifth International Brain-Computer Interface Meeting: Defining the Future*.

[20] Alicia Heraz and Claude Frasson. 2009. Predicting Learner Answers Correctness Through Brainwaves Assesment and Emotional Dimensions. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 49–56. http://dl.acm.org/citation.cfm?id=1659450.1659462

[21] David I Holmes and Richard S Forsyth. 1995. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic computing* 10, 2 (1995), 111–127.

[22] P. Juola, J. Noecker Jr, A. Stolerman, M. Ryan, P. Brennan, and R. Greenstadt. 2013. Keyboard Behavior Based Authentication for Security. *IEEE IT Professional* 15, 4 (July/August 2013).

[23] Girija Kaimal and William R Beardslee. 2010. Emerging adulthood and the perception of parental depression. *Qualitative Health Research* 20, 9 (2010), 1213–1228.

[24] Girija Kaimal and William R Beardslee. 2015. The Perceived Impact of Parental Depression on the Narrative Construction of Personal Identity: Reflections from Emerging Adults. *Narrative Works* 5, 1 (2015).

[25] Girija Kaimal, Kendra Ray, and Juan Muniz. 2016. Reduction of Cortisol Levels and Participants' Responses Following Art Making. *Art Therapy* 33, 2 (2016), 74–80.

[26] Shinsuke Koike, Yukika Nishimura, Ryu Takizawa, Noriaki Yahata, and Kiyoto Kasai. 2013. Near-infrared spectroscopy in schizophrenia: a possible biomarker for predicting clinical outcome and treatment. (2013).

[27] Shinsuke Koike, Yukika Nishimura, Ryu Takizawa, Noriaki Yahata, and Kiyoto Kasai. 2015. Near-infrared spectroscopy in schizophrenia: a possible biomarker for predicting clinical outcome and treatment. *Neuropsychopharmacology of Psychosis: Relation of Brain Signals, Cognition and Chemistry* (2015), 32.

[28] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 624–628.

[29] Mikhail B Malyutov. 2006. Authorship attribution of texts: A review. In *General Theory of Information Transfer and Combinatorics*. Springer, 362–380.

[30] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros, and Dawn Song. 2012. On the Feasibility of Side-Channel Attacks with Brain-Computer Interfaces.. In *USENIX security symposium*. 143–158.

[31] Daisuke Matsuzawa, Kotaro Takeda, Hiroyuki Ohtsuka, Jun Takasugi, Takashi Watanabe, Junko Maeda, Saeka Nagakubo, Chihiro Sutoh, Ichiro Shimoyama, Ken Nakazawa, and others. 2012. Correlation of prefrontal activity measured by near-infrared spectroscopy (NIRS) with mood, BDNF genotype and serum BDNF level in healthy individuals. (2012).

[32] JA Nasser, H Ayaz, RP Golen, B Makwana, E Albajri, MB Price, S Mogil, G Cucalon, and A DelParigi. 2016. Changes in neural activity of the prefrontal cortex during eating in humans. *Appetite* 107 (2016), 688–689.

[33] John Noecker, Michael Ryan, and Patrick Juola. 2013. Psychological profiling through textual analysis. *Literary and Linguistic Computing* 28, 3 (2013), 382–387.

[34] Yumie Ono. 2012. Prefrontal activity correlating with perception of sweetness during eating. In *Complex Medical Engineering (CME), 2012 ICME International Conference on*. IEEE, 125–130.

[35] Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2016, 3 (2016).

[36] RB Paranjape, J Mahovsky, L Benedicenti, and Z Koles. 2001. The electroencephalogram as a biometric. In *Electrical and Computer Engineering, 2001. Canadian Conference on*, Vol. 2. IEEE, 1363–1366.

[37] Akshay Patil, Juan Liu, Jianqiang Shen, Oliver Brdiczka, Jie Gao, and John Hanley. 2013. Modeling attrition in organizations from email communication. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 331–338.

[38] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[39] Abdul Serwadda, Vir V Phoha, Sujit Poudel, Leanne M Hirshfield, Danushka Bandara, Sarah E Bratt, and Mark R Costa. 2015. fNIRS: A new modality for brain activity-based biometric authentication. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 1–7.

[40] Jianqiang Shen, Oliver Brdiczka, Nicolas Ducheneaut, Nicholas Yee, and Bo Begole. 2012. Inferring personality of online gamers by fusing multiple-view predictions. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 261–273.

[41] Jianqiang Shen, Oliver Brdiczka, and Juan Liu. 2013. Understanding email writers: Personality prediction from email messages. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 318–330.

[42] David A Snowdon, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery. 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Jama* 275, 7 (1996), 528–532.

[43] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: Enhancing Interactive Systems with Streaming Fnirs Brain Input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), 2193–2202. DOI:http://dx.doi.org/10.1145/2207676.2208372

[44] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert J K Jacob. 2009. Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines. In *Proc. UIST '09*. 157–166. DOI:http://dx.doi.org/10.1145/1622176.1622207

[45] Ariel Stolerman, Aylin Caliskan, and Rachel Greenstadt. 2013. From Language to Family and Back: Native Language and Language Family Identification from English Text.. In *HLT-NAACL*. 32–39.

[46] Daniel Szafir and Bilge Mutlu. 2013. ARTFul: Adaptive Review Technology for Flipped Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1001–1010. DOI:http://dx.doi.org/10.1145/2470654.2466128

[47] Julie Thorpe, Paul C van Oorschot, and Anil Somayaji. 2005. Pass-thoughts: authenticating with our minds. In *Proceedings of the 2005 workshop on New security paradigms*. ACM, 45–56.

[48] Erin Treacy Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J. K. Jacob. 2015. Designing Implicit Interfaces for Physiological Computing: Guidelines and Lessons Learned Using fNIRS. *ACM Trans. Comput.-Hum. Interact.* 21, 6, Article 35 (Jan. 2015), 27 pages. DOI:http://dx.doi.org/10.1145/2687926

[49] Fiona J Tweedie, Sameer Singh, and David I Holmes. 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities* 30, 1 (1996), 1–10.

[50] Masahide Usami, Yoshitaka Iwadare, Masaki Kodaira, Kyota Watanabe, and Kazuhiko Saito. 2014. Near infrared spectroscopy study of the frontopolar hemodynamic response and depressive mood in children with major depressive disorder: a pilot study. *PloS one* 9, 1 (2014), e86290.

[51] Özlem Uzuner and Boris Katz. 2005. A comparative study of language models for book and author recognition. In *International Conference on Natural Language Processing*. Springer, 969–980.

[52] Marjolein H van Velzen, Luca Nanetti, and Peter P de Deyn. 2014. Data modelling in corpus linguistics: How low may we go? *Cortex* 55 (2014), 192–201.

[53] Krishna K Venkatasubramanian, Ayan Banerjee, and Sandeep Kumar S Gupta. 2010. PSKA: Usable and secure key agreement scheme for body area networks. *IEEE Transactions on Information Technology in Biomedicine* 14, 1 (2010), 60–68.

[54] Kaimal G. Myers-Coffman K. Gonzaga A.M.L. DeGraba T. J. Walker, M. 2017. Active duty military service members visual representations of PTSD and TBI. *International Journal of Qualitative Studies on Health and Well-being* in press, 1 (2017).

[55] Melissa S Walker, Girija Kaimal, Robert Koffman, and Thomas J DeGraba. 2016. Art therapy for PTSD and TBI: A senior active duty military service memberâĂŹs therapeutic journey. *The Arts in Psychotherapy* 49 (2016), 10–18.