PRIVACY INTERNATIONAL

# Examples of Data Points Used In Profiling

**THE FUTURE OF TRACKING**

## 1. WHAT IS POSSIBLE


*1.1 IDENTITY INFERENCES*
*Input: public data*
*Input: device metadata/data*
*Input: interaction metadata*
*Input: interaction data*
*1.2 PERSONAL INFORMATION INFERENCES*
*Input: public data*
*Input: device metadata/data*
*Input: interaction metadata*
*Input: interaction data*
*Input: unexpected surveillance*
*1.3 ECONOMIC INFERENCE*
*Input: device metadata/data*
*Input: interaction metadata/data*
*Input: unexpected surveillance*
*1.4 EMOTIONAL INFERENCES*
*Input: interaction metadata*
*Input: interaction data*
*Input: unexpected surveillance*
*1.5 BIOMETRIC INFERENCES*
*Input: unexpected surveillance*
*1.6 SPATIAL INFERENCES*
*Input: device metadata/data*
*Input: unexpected surveillance*
*1.7 PREDICTIVE INFERENCES*
*Input: interaction data/metadata*
*1.8 VISUAL INFERENCES*
*Input: unexpected surveillance*
*1.9 AUDIO INFERENCES*
*Input: unexpected surveillance*
*1.10 OTHER*


## 2. ILLUSTRATING HARM


*2.1 POLICING*

*2.2 'SHARING'/'FLEXIBLE' ECONOMY*
*2.3 DEVICES*
*2.4 SOCIAL MEDIA / INTERNET*
*2.5 POLITICS*
*2.6 ECONOMICS*
*2.7 MACHINE LEARNING (GENERAL)*

**3. RESOURCES**

*3.1 SURVEYS/LITERATURE REVIEWS*

**4. DIGITAL RESEARCH METHODS**

*4.1 GUIDES/TUTORIALS*
*4.2 EXAMPLES*
*4.3 ADVANCED METHODS*

When we browse the internet, go to work, drive down the street, go shopping, interact with institutions, or simply move through the city, data is collected about us. For advertisers, companies, and government agencies, what that data can say about us has become a valuable commodity. For data to become meaningful, however, requires it to be aggregated, sorted, and analyzed, often using automated computational processes. Through the use of machine learning and other statistical methods, computers are able to take seemingly mundane or innocuous data—what you like on Facebook or the frequency and length of your phone calls, for example—and make surprisingly accurate predictions of highly personal, sensitive information like your identity, political views, personality traits, or sexual orientation. While they may only be predictive, these methods produce data that is considered actionable by companies and governments as they construct an understanding of who you are and how you might act in the future through the data you leave behind everywhere you go. As more data becomes aggregated and linked to your identity, more accurate predictions or inferences can be made about you, raising important privacy concerns.

This report mostly draws from computer science literature to (1) show types of sensitive information that can be inferred through the analysis of common forms of data, (2) illustrate concrete harms that these inferences can produce, and (3) point researchers to other resources for understanding issues in data exploitation. While many of the studies in this report are at the cutting edge of computational social science and privacy studies, meaning many of these methods

have not yet been implemented at the time of their writing, they point towards possible privacy threats that can arise from the analysis of seemingly innocuous data. Other methods, however, represent refinements of existing methods that have been implemented in various technologies.

Some of these studies traverse complex ethical questions in ways that are not always critical or self-reflexive, but they do give valuable insights into the computational horizons of data exploitation and privacy violations. Also, most of the studies are peer-reviewed and/or cited, but there could certainly be methodological problems that would undermine some of their findings and claims.

† = Inferences that do not reveal personal information (but which might be used in combination with other data to do so)

## 1. WHAT IS POSSIBLE

Many of the studies in this section are validated based on probabilities, often expressed as percentages, that convey the accuracy of the data model. These models are created using training data sets that contain the data to be analyzed along with actual values for attributes that will be inferred. The latter might include a unique identity, gender, race, quantified personality traits, or any other number of data points that can be retrieved through surveys. The validation of the model often relies on binary decisions (although not always) that decide if the inference does or does not match the 'real' information. For example, in reference to a training data set, a study might claim that the model correctly infers the identity of 90% of people in the training data set using a set of defined variables. In some models, the inference being made is expressed as a probability that the inference matches the 'real' information. For example, a model might infer that there is a 75% chance that a subject fits into a given data category.

### *1.1 IDENTITY INFERENCES*

Identity inferences often involve ascertaining an individual's identity from a dataset that appears to lack personally-identifiable information. These inferences might link data to a particular name or they might show how a limited amount of data can be used to re-identify or uniquely identify an individual.

*Input: public data*

[medical info + news articles —> identity/medical records]

Sweeney, L., 2013. Matching known patients to health records in Washington State data. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2289850

> "News information uniquely and exactly matched medical records in the state database for 35 of the 81 cases (or 43 percent) found in 2011, thereby putting names to patient records."

[demographics on website + public records —> identity (names)]
Sweeney, L., Abu, A. and Winn, J., 2013. Identifying participants in the personal genome project by name. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2257732

> "We linked names and contact information to publicly available profiles in the Personal Genome Project. These profiles contain medical and genomic information, including details about medications, procedures and diseases, and demographic information, such as date of birth, gender, and postal code. By linking demographics to public records such as voter lists, and mining for names hidden in attached documents, we correctly identified 84 to 97 percent of the profiles for which we provided names."

[gender + zip + dob —> identity]
Golle, P., 2006, October. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society* (pp. 77-80). ACM.

> "we find that disclosing one's gender, ZIP code and full date of birth allows for unique identification of fewer individuals[1] (63% of the US population)"

*Input: device metadata*

[mobility —> identity]
De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, p.1376.
https://www.nature.com/articles/srep01376

> "In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals."

[installed phone apps —> identity]

1. As compared to earlier study: Sweeney, L., 2000. *Uniqueness of simple demographics in the US population*. Technical report, Carnegie Mellon University.

Achara, J.P., Acs, G. and Castelluccia, C., 2015, October. On the unicity of smartphone applications. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society* (pp. 27-36). ACM.

> "Our study finds that any 4 apps installed by a user are enough (more than 95% times) for the re-identification of the user in our dataset."

[browser configuration —> identity]
Eckersley, P., 2010, July. How unique is your web browser?. In Privacy Enhancing Technologies (Vol. 6205, pp. 1-18).
http://dl.acm.org/citation.cfm?id=1881152&CFID=805742437&CFTOKEN=62333013

> "if we pick a browser at random, at best we expect that only one in 286,777 other browsers will share its fingerprint. Among browsers that support Flash or Java, the situation is worse, with the average browser carrying at least 18.8 bits of identifying information. 94.2% of browsers with Flash or Java were unique in our sample.
>
> By observing returning visitors, we estimate how rapidly browser fingerprints might change over time. In our sample, fingerprints changed quite rapidly, but even a simple heuristic was usually able to guess when a fingerprint was an "upgraded" version of a previously observed browser's fingerprint, with 99.1% of guesses correct and a false positive rate of only 0.86%."

*Input: interaction metadata*

[credit card transactions (spatiotemporal) —> identity]
De Montjoye, Y.A., Radaelli, L. and Singh, V.K., 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221), pp.536-539.
http://science.sciencemag.org/content/347/6221/536

> "We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average."

[movie ratings —> identity & political preferences]
Narayanan, A. and Shmatikov, V., 2008, May. Robust de-anonymization of large sparse datasets. In *Security and Privacy*, 2008. SP 2008. IEEE Symposium on (pp. 111-125). IEEE.
http://ieeexplore.ieee.org/abstract/document/4531148/?reload=true

> "We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix,the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information."

[social network structure —> identity]
Narayanan, A. and Shmatikov, V., 2009, May. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on* (pp. 173-187). IEEE. https://arxiv.org/pdf/0903.3276.pdf

> "We present a framework for analyzing privacy and anonymity in social networks and develop a new
> re-identification algorithm targeting anonymized social-network graphs. To demonstrate its effectiveness on real-
> world networks, we show that a third of the users who can be verified to have accounts on both Twitter, a popular
> microblogging service, and Flickr, an online photo-sharing site, can be re-identified in the anonymous Twitter graph with only a 12% error rate.
>
> Our de-anonymization algorithm is based purely on the network topology"

[social network groups —> identity]
Wondracek, G., Holz, T., Kirda, E. and Kruegel, C., 2010, May. A practical attack to de-anonymize social network users. In *Security and Privacy (SP), 2010 IEEE Symposium on* (pp. 223-238). IEEE. http://ieeexplore.ieee.org/abstract/document/5504716/

> "we show that information about the group memberships of a user (i.e., the groups of a social network to which a user belongs) is sufficient to uniquely identify this person, or, at least, to significantly reduce the set of possible candidates. That is, rather than tracking a user's browser as with cookies, it is possible to track a person. To determine the group membership of a user, we leverage well-known web browser history stealing attacks. Thus, whenever a social network user visits a malicious website, this website can launch our de-anonymization attack and learn the identity of its visitors"

[social media data (location, time, writing style) —> identity]

Goga, O., Lei, H., Parthasarathi, S.H.K., Friedland, G., Sommer, R. and Teixeira, R., 2013, May. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 447-458). ACM.
http://dl.acm.org/citation.cfm?id=2488428

> "We study how potential attackers can identify accounts on different social network sites that all belong to the same user, exploiting only innocuous activity that inherently comes with posted content. We examine three specific features on Yelp, Flickr, and Twitter: the geo-location attached to a user's posts, the timestamp of posts, and the user's writing style as captured by language models. We show that among these three features the location of posts is the most powerful feature to identify accounts that belong to the same user in different sites. When we combine all three features, the accuracy of identifying Twitter accounts that belong to a set of Flickr users is comparable to that of existing attacks that exploit usernames. Our attack can identify 37% more accounts than using usernames when we instead correlate Yelp and Twitter. Our results have significant privacy implications as they present a novel class of attacks that exploit users' tendency to assume that, if they maintain different personas with different names, the accounts cannot be linked together; whereas we show that the posts themselves can provide enough information to correlate the accounts."

*Input: interaction data*

[writing —> identity]
Narayanan, A., Paskov, H., Gong, N.Z., Bethencourt, J., Stefanov, E., Shin, E.C.R. and Song, D., 2012, May. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 300-314). IEEE.
http://ieeexplore.ieee.org/abstract/document/6234420/

> "We study techniques for identifying an anonymous author via linguistic stylometry, i.e., comparing the writing style against a corpus of texts of known authorship. We experimentally demonstrate the effectiveness of our techniques with as many as 100,000 candidate authors. Given the increasing availability of writing samples online, our result has serious implications for anonymity and free speech - an anonymous blogger or whistleblower may be unmasked unless they take steps to obfuscate their writing style."

### *1.2 PERSONAL INFORMATION INFERENCES*

Researchers have been exploring a myriad of ways that machine learning and other statistical data approaches can be used to infer all kinds of personal data using a range of data inputs.

*Input: public data:*

[public data —> SSN]
Acquisti, A. and Gross, R., 2009. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, 106(27), pp.10975-10980.
http://www.pnas.org/content/106/27/10975.full

> "Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites."

*Input: device metadata/data*

[Smartphone usage (ie: calls, texts, app usage) —> personality traits]
Chittaranjan, G., Blom, J. and Gatica-Perez, D., 2011, June. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Wearable Computers (ISWC)*, 2011 15th Annual International Symposium on (pp. 29-36). IEEE.
https://infoscience.epfl.ch/record/192371/files/Chittaranjan_ISWC11_2011.pdf

> "From the analysis, we show that aggregated features obtained from smartphone usage data can be indicators of the Big-Five personality traits [extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience]. Additionally, we develop an automatic method to infer the personality type of a user based on cellphone usage using supervised learning. We show that our method performs significantly above chance and up to 75.9% accuracy."

[Phone data [calls & proximity] —> personality traits]
Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N. and Pentland, A., 2012, September. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 321-330). ACM.
https://static1.squarespace.com/static/55b64ce8e4b030b2d9ed3c6a/t/55c10c1de4b06bb56befb9f9/1438714909246/ubicomp2012.pdf

> "we believe that our results have provided compelling evidence that mobile phones-based behavioral data [calls logs & proximity via bluetooth] can be superior to survey ones for the purposes of personality classification [big five] and that egonet-based features can improve performance over actor-based ones"

[Phone logs —> personality traits]
de Montjoye, Y.A., Quoidbach, J., Robic, F. and Pentland, A., 2013, April. Predicting Personality Using Novel Mobile Phone-Based Metrics. In *SBP* (pp. 48-55).
https://link.springer.com/content/pdf/10.1007/978-3-642-37210-0.pdf#page=63

> "The present study provides the first evidence that personality can be reliably predicted from standard mobile phone logs."

[phone data —> friendship networks]
Eagle, N., Pentland, A.S. and Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36), pp.15274-15278.

> "We also demonstrate that it is possible to accurately infer 95% of friendships based on the observational data alone [call logs, proximity via bluetooth, cell towers, app usage, phone status], where friend dyads demonstrate distinctive temporal and spatial patterns in their physical proximity and calling patterns. These behavioral patterns, in turn, allow the prediction of individual-level outcomes such as job satisfaction."

[location [via phone] —> friendship]
Dong, W., Lepri, B. and Pentland, A.S., 2011, December. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia* (pp. 134-143). ACM.
https://static1.squarespace.com/static/55b64ce8e4b030b2d9ed3c6a/t/55c11290e4b0cae5c6bf24b5/1438716560956MobileUbicompMultiMedia.pdf

> "We demonstrate that by modeling the dynamics in sensor data, we can predict friendship, and can synthesize useful and accurate behavior and interaction projections. "

*Input: interaction metadata*

[facebook likes —> 'ethnic affinity']
Angwin, J. and Parris, T. 2016. Facebook Lets Advertisers Exclude Users by Race. *ProPublica*
https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

"Facebook assigns members an "Ethnic Affinity" based on pages and posts they have liked or engaged with on Facebook."

[Facebook likes —> personal characteristics (ie: views, sexual orientation, personality, age, gender, etc)]
Kosinski, M., D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110 (15):5802–5805. http://www.pnas.org/content/110/15/5802

"We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender…. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases."

[social network structure —> romantic partner]
Backstrom, L. and Kleinberg, J., 2014, February. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 831-841). ACM. https://arxiv.org/pdf/1310.6753v1.pdf

"given all the connections among a person's friends, can you recognize his or her romantic partner from the network structure alone? Using data from a large sample of Facebook users, we find that this task can be accomplished with high accuracy, but doing so requires the development of a new measure of tie strength that we term 'dispersion' — the extent to which two people's mutual friends are not themselves well-connected."

[clickstream —> personal characteristics]
De Bock, K. and Van den Poel, D., 2010. Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, 98(1), pp.49-70. https://www.researchgate.net/profile/Dirk_Van_den_Poel/publication/46443425_Predicting_Website_Audience_Demographics_for_Web_Advertising_Targeting_Using_Multi-Website_Clickstream_Data

"the transformation of website visitors' clickstream patterns to a set of features and the training of Random Forest classifiers that generate predictions for gender, age, level of education and occupation category"

[search queries —> personal characteristics & political views]
Bi, B., Shokouhi, M., Kosinski, M. and Graepel, T., 2013, May. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 131-140). ACM. http://dl.acm.org/citation.cfm?id=2488401

> 'showing how user demographic traits such as age and gender, and even political and religious views can be efficiently and accurately inferred based on their search query histories.'

[social network structure —> political views]
Barberá, P., 2014. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1), pp.76-91.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2108098

> "In this paper I show that the structure of the social networks in which they are embedded has the potential to become a source of information about policy positions. Under the assumption that social networks are homophilic, I develop a Bayesian Spatial Following model that scales Twitter users along a common ideological dimension based on who they follow. I apply this network-based method to estimate ideal points for a large sample of Twitter users in the US, the UK, Spain, Germany, Italy, and the Netherlands. The resulting positions of the party accounts on Twitter are highly correlated with offline measures based on their voting records and their manifestos. Similarly, this method is able to successfully classify individuals who state their political orientation publicly, and a sample of users from the state of Ohio whose Twitter accounts are matched with their voter registration history."

[space/time social media data —> social tie]
Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D. and Kleinberg, J., 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52), pp.22436-22441. http://www.pnas.org/content/107/52/22436.short

> "We investigate the extent to which social ties between people can be inferred from co-occurrence in time and space: Given that two people have been in approximately the same geographic locale at approximately the same time, on multiple occasions, how likely are they to know each other? Furthermore, how does this likelihood depend on the spatial and temporal proximity of the co-occurrences? Such issues arise in data originating in both online and offline domains as well as settings that capture interfaces between online and offline behavior. Here we develop a framework for quantifying the answers to such questions, and we

apply this framework to publicly available data from a social media site, finding that even a very small number of co-occurrences can result in a high empirical likelihood of a social tie."

*Input: interaction data*

[social media language —> personality traits]
Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H. and Seligman, M.E., 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6), p.934. http://psycnet.apa.org/record/2014-45458-001

"We describe a method for assessing personality using an open-vocabulary analysis of language from social media. We compiled the written language from 66,732 Facebook users and their questionnaire-based self-reported Big Five personality traits, and then we built a predictive model of personality based on their language."

*Input: unexpected surveillance*

[audio —> keystrokes]
Zhu, T., Ma, Q., Zhang, S. and Liu, Y., 2014, November. Context-free attacks using keyboard acoustic emanations. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 453-464). ACM. http://dl.acm.org/citation.cfm?id=2660296

"It proves possible for an attacker to recover the keystrokes by acoustic signal emanations…. Using off-the-shelf smartphones to record acoustic emanations from keystrokes, this design estimates keystrokes' physical positions based on the Time Difference of Arrival (TDoA) method. We conduct extensive experiments and the results show that more than 72.2% of keystrokes can be successfully recovered."

[wifi —> keystrokes]
Ali, K., Liu, A.X., Wang, W. and Shahzad, M., 2015, September. Keystroke recognition using wifi signals. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking* (pp. 90-102). ACM. http://dl.acm.org/citation.cfm?id=2790109

"WiFi signals can also be exploited to recognize keystroke"

*1.3 ECONOMIC INFERENCES*

*Input: device metadata/data*

[phone data —> spending habits]
Singh, V.K., Freeman, L., Lepri, B. and Pentland, A.P., 2013. Classifying spending behavior using socio-mobile data.
https://static1.squarespace.com/static/55b64ce8e4b030b2d9ed3c6a/t/55c10bbce4b015abaf6d8cc3/1438714812545/Classifying-Spending-Behavior-using-Socio-Mobile-Data-1.pdf

> "Using a data set involving 52 adults (26 couples) living in a community for over a year, we find that social
> behavior measured via face-to-face interaction [measured via bluetooth], call, and SMS logs, can be used to predict the spending behavior for couples in terms of their propensity to explore diverse businesses, become loyal customers, and overspend."

[mobile phone metadata —> wealth]
Blumenstock, J., Cadamuro, G. and On, R., 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), pp.1073-1076.
http://science.sciencemag.org/content/350/6264/1073.full

> "We show that an individual's past history of mobile phone use ["total volume, intensity, timing, and directionality of communication; the structure of the individual's contact network; patterns of mobility and migration based on geospatial markers in the data"] can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households."

[browser config, history, and account —> prices steering/discrimination]
Hannak, A., Soeller, G., Lazer, D., Mislove, A. and Wilson, C., 2014, November. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference* (pp. 305-318). ACM. http://dl.acm.org/citation.cfm?id=2663744

> "we ran controlled experiments to investigate what features e-commerce personalization algorithms take into account when shaping content. We found cases of sites altering results based on the user's OS/browser, account on the site, and history of clicked/purchased products. We also observe two travel sites conducting A/B tests that steer users towards more expensive hotel reservations."

[OS —> price discrimination]
Mattioli, D., 2012. On Orbitz, Mac users steered to pricier hotels. *Wall Street Journal*, 23, p.2012. https://www.wsj.com/articles/SB10001424052702304458604577488822667325882

> "Orbitz Worldwide Inc. has found that people who use Mac computers spend as much as 30% more a night on hotels, so the online travel agency is starting to show them different, and sometimes costlier, travel options than Windows visitors see."

[mobile phone metadata —> loan decisions]
http://money.cnn.com/2016/08/24/technology/lenddo-smartphone-battery-loan/

> "While certain mobile behavior could impact the outcome of a credit score (like always running out of battery power), Stewart said extremely well-maintained smartphones raise a red flag in the system, too."

[social media data/metadata —> loan decisions]
https://www.economist.com/news/finance-and-economics/21571468-lenders-are-turning-social-media-assess-borrowers-stat-oil

> "Some firms piece together scores by analysing applicants' online social networks."
> "As statistics accumulate, algorithms get better at spotting correlations in the data."
> "Facebook data already inform lending decisions"

*Input: interaction metadata/data*

†[web search —> economic indicators]
Choi, H. and Varian, H., 2012. Predicting the present with Google Trends. Economic Record, 88(s1), pp.2-9. http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/full

> "In this paper we show how to use search engine data to forecast near-term values of economic indicators. Examples include automobile sales, unemployment claims, travel destination planning and consumer confidence."

†[web search —> future housing prices/sales]
Wu, L. and Brynjolfsson, E., 2009. The future of prediction: How Google searches foreshadow housing prices and sales. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293

> "We demonstrate how data from search engines such as Google provide an accurate but simple way to predict future business activities. Applying our methodology to predict housing market trends, we find that a housing search index is strongly predictive of future housing market sales and prices. For state-level predictions in the US, the use of search data produces

out-of-sample predictions with a smaller mean absolute error than the baseline model that uses conventional data but lacks search data."

[clickstream —> personal preferences (for ads)]
Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y. and Chen, Z., 2009, April. How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World wide web* (pp. 261-270). ACM. http://dl.acm.org/citation.cfm?id=1526745

"Behavioral Targeting (BT) [= web browsing data] is a technique used by online advertisers to increase the effectiveness of their campaigns, and is playing an increasingly important role in the online advertising market. However, it is underexplored in academia how much BT can truly help online advertising in search engines. In this paper we provide an empirical study on the click-through log of advertisements collected from a commercial search engine. From the experiment results over a period of seven days, we draw three important conclusions: (1) Users who clicked the same ad will truly have similar behaviors on the Web; (2) Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising in a sponsored search; (3) Using short term user behaviors to represent users is more effective than using long term user behaviors for BT."

*Input: unexpected surveillance*

[video (face) —> economic decision]
Rossi, F., Fasel, I. and Sanfey, A.G., 2011. Inscrutable games? Facial expressions predict economic behavior. *BMC Neuroscience*, 12(1), p.P281.
https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-12-S1-P281

"Neuroscientific and behavioral evidence shows that when subjects are engaged in simple economic games, they pay attention to the face of their opponents. Is this a good idea? Does the face of a decision-maker contain information about his strategy space? We tested this hypothesis by modeling facial expressions of subjects playing the Ultimatum Game. We recorded videos of 60 participants, and automatically extracted time-series of facial actions (12 action units [1], shown in Fig. 1A., as well as pitch, yaw, and roll of the head) using the real-time facial coding system of [2, 3]. We then trained non-linear support vector machines (SVM) to predict the decision of the second player from a segment of video acquired after the offer was received and before the decision was entered (n = 376). To separate the dynamics of facial behavior into different temporal scales, the data was preprocessed with a

bank of Gabor filters. With this method we achieved a between-subjects cross-validation accuracy of 0.66 (chance = 0.50) in predicting decisions."

*1.4 EMOTIONAL INFERENCES*

*Input: interaction metadata*

[location —> emotion]
Frank, M.R., Mitchell, L., Dodds, P.S. and Danforth, C.M., 2013. Happiness and the patterns of life: A study of geolocated tweets. arXiv preprint *arXiv*:1304.1296.
https://arxiv.org/abs/1304.1296

"We use a collection of 37 million geolocated tweets to characterize the movement patterns of 180,000 individuals, taking advantage of several orders of magnitude of increased spatial accuracy relative to previous work. Employing the recently developed sentiment analysis instrument known as the 'hedonometer', we characterize changes in word usage as a function of movement, and find that expressed happiness increases logarithmically with distance from an individual's average location."

*Input: interaction data*

[positive/negative expressions (in news feed) —> emotional state]
Kramer, A.D., Guillory, J.E. and Hancock, J.T., 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), pp.8788-8790. http://www.pnas.org/content/111/24/8788.full[2]

"In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks."

*Input: unexpected surveillance*

2. Privacy concerns have been raised concerning the methods used in this study: http://www.pnas.org/content/111/29/10779.1.full

[RF signals —> emotional state]

Zhao, M., F. Adib, and D. Katabi. 2016. Emotion recognition using wireless signals. 95–108. *ACM Press* http://dl.acm.org/citation.cfm?doid=2973750.2973762 (last accessed 15 August 2017).

> "EQ-Radio transmits an RF signal and analyzes its reflections off a person's body to recognize his emotional state (happy, sad, etc.). The key enabler underlying EQ-Radio is a new algorithm for extracting the individual heartbeats from the wireless signal at an accuracy comparable to on-body ECG monitors."

[typing —> emotional state]

Epp, C., Lippold, M. and Mandryk, R.L., 2011, May. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 715-724). ACM. http://hci.usask.ca/uploads/203-p715-epp.pdf

> "determine user emotion by analyzing the rhythm of their typing patterns on a standard keyboard"

*1.5 BIOMETRIC INFERENCES*

*Input: unexpected surveillance*

[video —> heartbeat/bloodflow]

Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F. and Freeman, W., 2012. Eulerian video magnification for revealing subtle changes in the world. http://dl.acm.org/citation.cfm?id=2185561

> "Our method, which we call Eulerian Video Magnification, takes a standard video sequence as input, and applies spatial decomposition, followed by temporal filtering to the frames. The resulting signal is then amplified to reveal hidden information. Using our method, we are able to visualize the flow of blood as it fills the face and also to amplify and reveal small motions."

[RF signals —> sleep stage]

Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S. and Bianchi, M.T., 2017, July. Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture. In *International Conference*

*on Machine Learning* (pp. 4100-4109).
https://people.csail.mit.edu/tommi/papers/Zhao_etal_ICML2017.pdf

"We focus on predicting sleep stages from radio measurements without any attached sensors on subjects. We introduce a new predictive model that combines convolutional and recurrent neural networks to extract sleep-specific subject-invariant features from RF signals and capture the temporal progression of sleep"

*1.6 SPATIAL INFERENCES*

Long used in criminology, geographic profiling attempts to assign probabilities to space that predict the location of a subject in question, thus narrowing searches for a suspect. This section also contains other attempts to make spatial inferences and predictions using geographically-referenced data.

*Input: device metadata/data*

[correlated mobility —> movement prediction]
De Domenico, M., Lima, A. and Musolesi, M., 2013. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6), pp.798-807.https://arxiv.org/pdf/1210.2376.pdf

"it is possible to increase the forecasting accuracy [of human movement prediction] by considering movements of friends, people, or more in general entities, with correlated mobility patterns"

[locations —> home address (probability)]
Hauge, M. V., M. D. Stevenson, D. K. Rossmo, and S. C. Le Comber. 2016. Tagging Banksy: using geographic profiling to investigate a modern art mystery. *Journal of Spatial Science* :1–6. http://www.tandfonline.com/doi/abs/10.1080/14498596.2016.1138246?journalCode=tjss20

"Here, we use a Dirichlet process mixture (DPM) model of geographic profiling, a mathematical technique developed in criminology and finding increasing application within ecology and epidemiology, to analyse the spatial patterns of Banksy artworks in Bristol and London. The model takes as input the locations of these artworks, and calculates the probability of 'offender' residence across the study area. Our analysis highlights areas

associated with one prominent candidate (e.g., his home), supporting his identification as Banksy."

[locations —> home/workplace (probability)]

Le Comber, S. C., and M. D. Stevenson. 2012. From Jack the Ripper to epidemiology and ecology. *Trends in Ecology & Evolution* 27 (6):307–308. http://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(12)00068-7

"Geographic profiling (GP) is a statistical technique developed in criminology to identify likely candidates from large lists of suspects in cases of serial crime such as murder or rape [3]. With large lists of suspects (268,000 names in the Yorkshire Ripper investigation in the UK in the 1980s), it is difficult or impossible to investigate each name, and a prioritisation strategy is useful. GP uses the spatial locations of crime sites to make inferences about the location of the offender's 'anchor point' (usually a home, but sometimes a workplace). The model depends on two key concepts: distance decay and the buffer zone. Distance decay reflects the fact that criminals are more likely to commit crimes nearer their anchor point rather than further away, since travel requires time and effort. The buffer zone describes an area surrounding the criminal's anchor point in which he/she is less likely to commit crimes, either because of an increased risk of recognition, or for geometric reasons (because area increases with distance squared, the number of potential crime sites will also increase with distance). The model assigns a score to each point in the search area; the higher the score, the greater the probability that the offender's anchor point is located there. The resulting jeopardy surface can be overlaid on a map of the search area to produce a geoprofile (Figure 1). Investigators then search their list of suspects in rank order according to the height of their homes (or other anchor points) on the geoprofile. Note that GP therefore describes a search strategy, rather than giving a point estimate of the offender's home."

[locations —> home (probability)]

Rossmo, D. K. 2014. Geographic Profiling. In *Encyclopedia of Criminology and Criminal Justice*, eds. G. Bruinsma and D. Weisburd, 1934–1942. New York, NY: Springer New York http://link.springer.com/10.1007/978-1-4614-5690-2_678 (last accessed 15 August 2017).

"Geographic profiling has turned out to be a robust and versatile methodology. Originally developed for analyzing serial murder cases, it was subsequently applied to rape, arson, robbery, bombing, kidnapping, burglary, auto theft, credit card fraud, and graffiti investigations. A number of innovative applications outside law enforcement also exist, with geographic profiling being used in military operations, intelligence analysis, biology, zoology, epidemiology, and archaeology."

> "Geographic profiling is based on the theories, concepts, and principles of environmental criminology. Crime locations are not distributed randomly in space but rather are influenced by the road networks and features of the physical environment. This focus on the crime setting – the "where and when" of the criminal act – offers a conceptual framework for determining the most probable area of offender residence."

[location tweets —> home —> demographics]
Liccardi, I., Abdul-Rahman, A. and Chen, M., 2016, May. I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). ACM.
http://dl.acm.org/citation.cfm?id=2858272

> "…with a small number of data points [tweets, in this study], people's locations can be inferred. This kind of information can lead to several privacy disclosures. Using publicly available data, the type of locations can be used to estimate someone's average income based on one's neighborhood, average housing cost, debt, and other demographic information, such as political views etc."

*Input: unexpected surveillance*
[RF signals —> body position]
Adib, F., Hsu, C.Y., Mao, H., Katabi, D. and Durand, F., 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)*, 34(6), p.219.
http://dl.acm.org/citation.cfm?id=2818072

> "RF-Capture tracks the 3D positions of a person's limbs and body parts even when the person is fully occluded from its sensor, and does so without placing any markers on the subject's body. In designing RF-Capture, we built on recent advances in wireless research, which have shown that certain radio frequency (RF) signals can traverse walls and reflect off the human body, allowing for the detection of human motion through walls."

[video —> car location]
Data Brokers Are Now Selling Your Car's Location For $10 Online
https://www.forbes.com/sites/adamtanner/2013/07/10/data-broker-offers-new-service-showing-where-they-have-spotted-your-car/#671fef99470b

> "a prominent data broker announced two weeks ago that it had begun selling locational information on license plates that have been filmed and identified. In recent years, police have also widely embraced license plate recognition to track suspected criminals. Repo men use the technology to recover vehicles; casinos in Las Vegas employ it to monitor cars in their

parking lots. And now data broker TLO has begun selling information about the time and location at which cars have been sighted."

### *1.7 PREDICTIVE INFERENCES*

These examples use analysis to predict future trends at an aggregate level. So while they do not contain inferences about specific individuals, they might be combined with other types of data to target individuals.

*Input: interaction data/metadata*

†[social interaction —> future trends]
Altshuler, Y., Pan, W. and Pentland, A.S., 2012, April. Trends prediction using social diffusion models. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 97-104). Springer, Berlin, Heidelberg.
https://static1.squarespace.com/static/55b64ce8e4b030b2d9ed3c6a/t/55c11631e4b06bb56bf0009b/1438717489174/Behavior-shaping.pdf

> "Whereas many works focus on the detection of anomalies in networks, there exist little theoretical work on the prediction of the likelihood of anomalous network pattern to globally spread and become "trends". In this work we present an analytic model for the social diffusion dynamics of spreading network patterns. Our proposed method is based on information diffusion models, and is capable of predicting future trends based on the analysis of past social interactions between the community's members."

†[transit data —> contagious outbreaks]
Sun, L., Axhausen, K.W., Lee, D.H. and Cebrian, M., 2014. Efficient detection of contagious outbreaks in massive metropolitan encounter networks. *Scientific reports*, 4, p.5099.
https://arxiv.org/pdf/1401.2815.pdf

> "Physical contact remains difficult to trace in large metropolitan networks, though it is a key vehicle for the transmission of contagious outbreaks. Co-presence encounters during daily transit ["Trip records were collected from Singapore's smart-card-based fare collection system"] use provide us with a city-scale time-resolved physical contact network, consisting of 1 billion contacts among 3 million transit users. Here, we study the advantage that knowledge of such co-presence structures may provide for early detection of contagious outbreaks"

*1.8 VISUAL INFERENCES*

*Input: unexpected surveillance*

[sound —> objects in scene]
Aytar, Y., Vondrick, C. and Torralba, A., 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems* (pp. 892-900). https://arxiv.org/abs/1610.09001

> "Given a video, our model recognizes objects and scenes from sound only."

[video (from phone/tablet) —> eye tracking]
Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W. and Torralba, A., 2016. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2176-2184). https://arxiv.org/abs/1606.05814

> "We believe that we can put the power of eye tracking in everyone's palm by building eye tracking software that works on commodity hardware such as mobile phones and tablets, without the need for additional sensors or devices. We tackle this problem by introducing GazeCapture, the first large-scale dataset for eye tracking…"

*1.9 AUDIO INFERENCES*

*Input: unexpected surveillance*

[video —> audio]
Bear, H.L. and Harvey, R., 2016, March. Decoding visemes: Improving machine lip-reading. In *Acoustics, Speech and Signal Processing* (ICASSP), 2016 IEEE International Conference on (pp. 2009-2013). IEEE. https://ueaeprints.uea.ac.uk/57978/1/Template.pdf

> "To undertake machine lip-reading, we try to recognise speech from a visual signal."

[video —> audio]
Davis, A., M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman. 2014. The visual microphone: passive recovery of sound from video. *ACM Transactions on Graphics* 33 (4):1–10. http://dl.acm.org/citation.cfm?id=2601119

> "When sound hits an object, it causes small vibrations of the object's surface. We show how, using only high-speed video of the object, we can extract those minute vibrations and partially recover the sound that produced them, allowing us to turn everyday objects---a glass of water, a potted plant, a box of tissues, or a bag of chips---into visual microphones."

[wifi signal + talking —> audio]
Wang, G., Zou, Y., Zhou, Z., Wu, K. and Ni, L.M., 2016. We can hear you with wi-fi!. *IEEE Transactions on Mobile Computing*, 15(11), pp.2907-2920.
http://ieeexplore.ieee.org/abstract/document/7384744/

> We present WiHear, which enables Wi-Fi signals to "hear" our talks without deploying any devices. To achieve this, WiHear needs to detect and analyze fine-grained radio reflections from mouth movements.

*1.10 OTHER*

[name —> (racist) search result]
Sweeney, L., 2013. Discrimination in online ad delivery. *Queue*, 11(3), p.10.
https://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf

> "This writing investigates the delivery of of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184 racially associated personal names across two websites. First names, previously identified by others as being assigned at birth to more black or white babies, are found predictive of race (88% black, 96% white), and those assigned primarily to black babies, such as DeShawn, Darnell and Jermaine, generated ads suggestive of an arrest in 81 to 86 percent of name searches on one website and 92 to 95 percent on the other, while those assigned at birth primarily to whites, such as Geoffrey, Jill and Emma, generated more neutral copy: the word "arrest" appeared in 23 to 29 percent of name searches on one site and 0 to 60 percent on the other."

[playing video games —> employability]
Morgan, J., 2014. *The future of work: Attract new talent, build better leaders, and create a competitive organization*. John Wiley & Sons.

> "…companies like Knack can help organizations move toward more diverse working environments by helping eliminate biases…. Having them play games and then looking at the data allows organization to not only move beyond biases but to actually understand the

potential candidate in a very deep way, thus truly being able to focus on the best potential person for the job."

[personal information —> employability]
https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/

"Evolv's tests allow companies to capture data about everybody who applies for work, and everybody who gets hired—a complete data set from which sample bias, long a major vexation for industrial-organization psychologists, simply disappears. The sheer number of observations that this approach makes possible allows Evolv to say with precision which attributes matter more to the success of retail-sales workers (decisiveness, spatial orientation, persuasiveness) or customer-service personnel at call centers (rapport-building). And the company can continually tweak its questions, or add new variables to its model, to seek out ever stronger correlates of success in any given job. For instance, the browser that applicants use to take the online test turns out to matter, especially for technical roles: some browsers are more functional than others, but it takes a measure of savvy and initiative to download them."

"Meyerle told me that what most excites him are the possibilities that arise from monitoring the entire life cycle of a worker at any given company. This is a task that Evolv now performs for Transcom, a company that provides outsourced customer-support, sales, and debt-collection services, and that employs some 29,000 workers globally."

"The way Gild arrives at these scores is not simple. The company's algorithms begin by scouring the Web for any and all open-source code, and for the coders who wrote it. They evaluate the code for its simplicity, elegance, documentation, and several other factors, including the frequency with which it's been adopted by other programmers. For code that was written for paid projects, they look at completion times and other measures of productivity. Then they look at questions and answers on social forums such as Stack Overflow, a popular destination for programmers seeking advice on challenging projects. They consider how popular a given coder's advice is, and how widely that advice ranges.

The algorithms go further still. They assess the way coders use language on social networks from LinkedIn to Twitter; the company has determined that certain phrases and words used in association with one another can distinguish expert programmers from less skilled ones. Gild knows these phrases and words are associated with good coding because it can correlate them with its evaluation of open-source code, and with the language and online behavior of programmers in good positions at prestigious companies."

[personal info, browsing habits, etc —> consumer profiles]
http://www.nybooks.com/articles/2014/01/09/how-your-data-are-being-deeply-mined/

> "Acxiom provides "premium proprietary behavioral insights" that "number in the thousands and cover consumer interests ranging from brand and channel affinities to product usage and purchase timing." In other words, Acxiom creates profiles, or digital dossiers, about millions of people, based on the 1,500 points of data about them it claims to have. These data might include your education level; how many children you have; the type of car you drive; your stock portfolio; your recent purchases; and your race, age, and education level. These data are combined across sources—for instance, magazine subscriber lists and public records of home ownership—to determine whether you fit into a number of predefined categories such as "McMansions and Minivans" or "adult with wealthy parent."3 Acxiom is then able to sell these consumer profiles to its customers, who include twelve of the top fifteen credit card issuers, seven of the top ten retail banks, eight of the top ten telecom/media companies, and nine of the top ten property and casualty insurers."

†[network structure —> anomalous event]
Altshuler, Y., Fire, M., Shmueli, E., Elovici, Y., Bruckstein, A.M., Pentland, A. and Lazer, D., 2013, April. Detecting Anomalous Behaviors Using Structural Properties of Social Networks. In *SBP* (pp. 433-440).
https://static1.squarespace.com/static/55b64ce8e4b030b2d9ed3c6a/t/55c10bdee4b07af73978e0e1/1438714846639/SBP13-Amplifier.pdf

> "In this paper we discuss the analysis of mobile networks communication patterns in the presence of some anomalous "real world event". We argue that given limited analysis resources (namely, limited number of network edges we can analyze), it is best to select edges that are located around 'hubs' in the network, resulting in an improved ability to detect such events."

[brain waves —> thoughts]
Pustilnik, A.C., 2012. Neurotechnologies at the intersection of criminal procedure and constitutional law. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2143187

> "The rapid development of neurotechnologies poses novel constitutional issues for criminal law and criminal procedure. These technologies can identify directly from brain waves whether a person is familiar with a stimulus like a face or a weapon, can model blood flow in the brain to indicate whether a person is lying, and can even interfere with brain processes

themselves via high-powered magnets to cause a person to be less likely to lie to an investigator."

[image —> important person]
Solomon Mathialagan, C., Gallagher, A.C. and Batra, D., 2015. Vip: Finding important people in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4858-4866). http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Mathialagan_VIP_Finding_Important_2015_CVPR_paper.pdf

"We introduce a measure of importance of people in images and investigate the correlation between importance and visual saliency. We find that not only can we automatically predict the importance of people from purely visual cues, incorporating this predicted importance results in significant improvement in applications such as im2text (generating sentences that describe images of groups of people)."

## 2. ILLUSTRATING HARM

Harm can be illustrated through a number of methods that include algorithmic audits, reverse-engineering, investigative research, and legal analysis. Many of these examples here deploy statistical analyses that compare algorithmic outcomes with known data to determine if certain classes of people are the subject of discriminatory inferences. Others build test data profiles to run through algorithms to determine how certain variables affect the results, giving insight into how the data model functions. Finally, some of these examples are based on research that shows specific harm to individuals (which is not always explainable) or policies and practices whose intent seems to be to exploit data in ways that will cause harm (the collection of personal data at borders, for examples).

### *2.1 POLICING*

[profiling]
Asher, J., and R. Arthur. 2017. Inside the Algorithm That Tries to Predict Gun Violence in Chicago. *The New York Times* 13 June. https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html

" • Violence in the city is less concentrated at the top — among a group of about 1,400 people with the highest risk scores — than some public comments from the Chicago police have suggested.

• Gangs are often blamed for the devastating increase in gun violence in Chicago, but gang membership had a small predictive effect and is being dropped from the most recent version of the algorithm.

• Being a victim of a shooting or an assault is far more predictive of future gun violence than being arrested on charges of domestic violence or weapons possession.

• The algorithm has been used in Chicago for several years, and its effectiveness is far from clear. Chicago accounted for a large share of the increase in urban murders last year."

"But using the publicly available data that they have released, we reverse-engineered the impact of each characteristic on the final risk scores with a linear regression model. Because the department didn't release all the information that the algorithm uses, our estimates of the significance of each characteristic are only approximate. But using what was available to us, we could predict risk score very accurately, suggesting that we are capturing much of the important information that goes into the algorithm."

[arrest]
Saunders, J., P. Hunt, and J. S. Hollywood. 2016. Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology* 12 (3):347–371. https://link.springer.com/article/10.1007/s11292-016-9272-0

"Results
Individuals on the SSL [Strategic Subjects List] are not more or less likely to become a victim of a homicide or shooting than the comparison group, and this is further supported by city-level analysis. The treated group is more likely to be arrested for a shooting."

[arrest]
How the NYPD is using social media to put Harlem teens behind bars: The untold story of Jelani Henry, who says Facebook likes landed him in Rikers
https://www.theverge.com/2014/12/10/7341077/nypd-harlem-crews-social-media-rikers-prison

""The mix of social media and conspiracy statutes creates a dragnet that can bring almost anybody in," says Andrew Laufer, a New York City attorney who has worked on numerous cases involving teenagers wrongly arrested by police. "It's a complete violation of the Fourth Amendment and the worst kind of big brother law enforcement." To build the case for the

Harlem raid, police had begun social media surveillance of children well before they had built up a serious criminal record.

Affiliation with a crew, even a tangential one, can be a deciding factor in getting locked up. "I find it disturbing and scary," says Christian Bolden, a professor of criminology at Loyola University. "In many states, if police see you together with someone three times — and this can be in real life or in a picture they find online — that is enough to prove conspiracy. That puts the onus on young people to be smart and careful about who they are with and what they post. And if we know one thing about teenagers, it's that they are rabidly social and often quite reckless." It was this exact mix of neighborhood affiliations and social media that entangled the fates of the Henry brothers."

[race discrimination]
Garvie, C., 2016. The Perpetual Line-Up: Unregulated Police Face Recognition in America. *Georgetown Law, Center on Privacy & Technology*. https://www.perpetuallineup.org/

"Yet an FBI co-authored study suggests that face recognition may be less accurate on black people. Also, due to disproportionately high arrest rates, systems that rely on mug shot databases likely include a disproportionate number of African Americans. Despite these findings, there is no independent testing regime for racially biased error rates. In interviews, two major face recognition companies admitted that they did not run these tests internally, either."

[race & class discrimination]
Lum, K. and Isaac, W., 2016. To predict and serve?. *Significance*, 13(5), pp.14-19.

"We find that rather than correcting for the apparent biases in the police data, the model reinforces these biases. The locations that are flagged for targeted policing are those that were, by our estimates, already over-represented in the historical police data. Figure 2(b) shows the percentage of the population experiencing targeted policing for drug crimes broken down by race. Using PredPol in Oakland, black people would be targeted by predictive policing at roughly twice the rate of whites. Individuals classified as a race other than white or black would receive targeted policing at a rate 1.5 times that of whites. This is in contrast to the estimated pattern of drug use by race, shown in Figure 2(c), where drug use is roughly equivalent across racial classifications. We find similar results when analysing the rate of targeted policing by income group, with low-income households experiencing targeted policing at disproportionately high rates. Thus, allowing a predictive policing algorithm to allocate police resources would result in the disproportionate policing of low-income communities and communities of colour."

[data leaks]
https://www.wired.com/story/how-peter-thiels-secretive-data-company-pushed-into-policing

> "When Sergeant Lee DeBrabander marked a case confidential in the Long Beach drug squad's Palantir data analysis system in November 2014, he expected key details to remain hidden from unauthorized users' eyes. In police work, this can be crucial—a matter of life and death, even. It often involves protecting vulnerable witnesses, keeping upcoming operations hush hush, or protecting a fellow police officer who's working undercover.
>
> Yet not long after, someone working in the gang crimes division ran a car license plate mentioned in his case and was able to read the entire file. "Can you please look at this?" DeBrabander wrote to a Palantir engineer in an email, which was obtained by Backchannel in response to public records requests."

[data leaks]
https://www.eff.org/deeplinks/2016/03/eff-pressure-results-increased-disclosure-abuse-californias-law-enforcement

> "Given that ability to access so much private data, local authorities and the CADOJ place restrictions on how and when law enforcement can use CLETS. Violations take many forms. In some cases, police officers have used CLETS data to screen online dates and to stalk former partners. In one incident, an officer allegedly accessed CLETS with the intent of providing a convicted killer's family with information on witnesses in the case."

[misidentification/privacy violations]
Senate, U.S., 2012. Federal support for and involvement in state and local fusion centers. Washington, DC: Permanent Subcommittee on Investigations. Committee on Homeland Security and Governmental Affairs. https://www.hsdl.org/?view&did=723145

> "Sharing terrorism-related information between state, local and federal officials is crucial to protecting the United States from another terrorist attack. Achieving this objective was the motivation for Congress and the White House to invest hundres of millions of taxpayer dollars over the last nine years in support of dozens of state and local fusion centers across the United States…
>
> The Subcommittee investigations found that DHS-assigned detailees to the fusion centers forwarded "intelligence" of uneven quality – oftentimes shoddy, rarely timely, sometime

endangering citizens' civil liberties and Privacy Act protections, occasionally taken from already-published public sources, and more often than not unrelated to terrorism."

*2.2 'SHARING'/FLEXIBLE ECONOMY*

[race & gender discrimination]
Ge, Y., Knittel, C.R., MacKenzie, D. and Zoepf, S., 2016. Racial and gender discrimination in transportation network companies (No. w22776). *National Bureau of Economic Research*. http://www.nber.org/papers/w22776

> "Results indicated a pattern of discrimination, which we observed in Seattle through longer waiting times for African American passengers—as much as a 35 percent increase. In Boston, we observed discrimination by Uber drivers via more frequent cancellations against passengers when they used African American-sounding names. Across all trips, the cancellation rate for African American sounding names was more than twice as frequent compared to white sounding names. Male passengers requesting a ride in low-density areas were more than three times as likely to have their trip canceled when they used a African American-sounding name than when they used a white-sounding name. We also find evidence that drivers took female passengers for longer, more expensive, rides in Boston. We observe that removing names from trip booking may alleviate the immediate problem but could introduce other pathways for unequal treatment of passengers.

[race discrimination]
Edelman, B., Luca, M. and Svirsky, D., 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2), pp.1-22. http://www.hbs.edu/faculty/Publication%20Files/16-069_5c3b2b36-d9f8-4b38-9639-2175aaf9ebc9.pdf

> "In a field experiment on Airbnb, we find that requests from guests with distinctively African-American names are roughly 16% less likely to be accepted than identical guests with distinctively White names."

[tracking]
https://www.theverge.com/2014/11/19/7245447/uber-allegedly-tracked-journalist-with-internal-tool-called-god-view

> "Uber is investigating its top New York executive after he was alleged to have tracked a journalist's location without her permission using an internal company tool called "God View.""

[race & gender discrimination]
Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M. and Wilson, C., 2017, February. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *CSCW* (pp. 1914-1933). http://claudiawagner.info/publications/cscw_bias_olm.pdf

> "In this paper, we study whether two prominent online freelance marketplaces—TaskRabbit and Fiverr—are impacted by racial and gender bias. From these two platforms, we collect 13,500 worker profiles and gather in-formation about workers' gender, race, customer reviews, ratings, and positions in search rankings. In both market-places, we find evidence of bias: we find that gender and race are significantly correlated with worker evaluations, which could harm the employment opportunities afforded to the workers."

*2.3 DEVICES*

[data leaks]
Zang, J., Dummit, K., Graves, J., Lisker, P. and Sweeney, L., 2015. Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps. *Technology Science*, 30. https://techscience.org/a/2015103001/

> "We found that the average Android app sends potentially sensitive data to 3.1 third-party domains, and the average iOS app connects to 2.6 third-party domains. Android apps are more likely than iOS apps to share with a third party personally identifying information such as name (73% of Android apps vs. 16% of iOS apps) and email address (73% vs. 16%). For location data, including geo-coordinates, more iOS apps (47%) than Android apps (33%) share that data with a third party. In terms of potentially sensitive behavioral data, we found that 3 out of the 30 Medical and Health & Fitness category apps in the sample share medically-related search terms and user inputs with a third party. Finally, the third-party domains that receive sensitive data from the most apps are Google.com (36% of apps), Googleapis.com (18%), Apple.com (17%), and Facebook.com (14%). 93% of Android apps tested connected to a mysterious domain, safemovedm.com, likely due to a background process of the Android phone. Our results show that many mobile apps share potentially sensitive user data with third parties, and that they do not need visible permission requests to access the data."

[data leaks]
https://www.eff.org/files/2017/04/13/student-privacy-report.pdf

"Throughout EFF's investigation over the past two years, we have found that educational technology services often collect far more information on kids than is necessary and store this information indefinitely. This privacy-implicating information goes beyond personally identifying information (PII) like name and date of birth, and can include browsing history, search terms, location data, contact lists, and behavioral information. Some programs upload this student data to the cloud automatically and by default. All of this often happens without the awareness or consent of students and their families.

[tracking]
https://www.accessnow.org/cms/assets/uploads/archive/AIBT-Report.pdf

"The researchers found Verizon Wireless and AT&T using so-called supercookies — special tracking headers that the carriers inject beyond the control of the user. These revelations led to an investigation by the U.S. Federal Communications Commission, action by legislators in the U.S. Congress, and several lawsuits. Despite these small victories, tracking headers are still being used around the world, and important questions remain."

"Although tracking headers are popularly called "supercookies," "zombie cookies," or "perma-cookies," these terms are inaccurate. Cookies are injected locally and can be manipulated by end users in a web browser. Tracking headers are in fact not cookies at all because they are injected at the network level, out of the reach of the user."

[deportation]
https://www.theguardian.com/uk-news/2017/aug/19/home-office-secret-emails-data-homeless-eu-nationals?CMP=share_btn_tw

"A chain of emails sent by senior Home Office immigration officials show how they used information that was designed to protect rough sleepers to target vulnerable individuals for deportation. The internal correspondence shows the Home Office repeatedly requesting and finally gaining access to a map created by the Greater London Authority (GLA) that identified and categorised rough sleepers by nationality.

The secret arrangement meant frontline outreach workers tasked with helping the homeless by collating data for the GLA were inadvertently helping the Home Office to remove people who were from the EU or central eastern Europe. In May 2016, the Home Office introduced guidance enabling immigration enforcement teams to deport EU nationals, purely on the grounds that they were sleeping rough."

*2.4 SOCIAL MEDIA / INTERNET*

[immigration]
http://www.bbc.co.uk/news/technology-40132506

> "The Trump administration has approved plans to ask US visa applicants for details of their social media use…. Consular officials can now ask for social media usernames going back five years via a new questionnaire."

[legal/arrest]
https://www.dreamhost.com/blog/we-fight-for-the-users/

> "At the center of the requests is disruptj20.org, a website that organized participants of political protests against the current United States administration. While we have no insight into the affidavit for the search warrant (those records are sealed), the DOJ has recently asked DreamHost to provide all information available to us about this website, its owner, and, more importantly, its visitors." [this request was later dropped by the DOJ, but only after resistance from the hosting company and public outcry]

[gender discrimination]
Kay, M., Matuszek, C. and Munson, S.A., 2015, April. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3819-3828). ACM.
http://dl.acm.org/citation.cfm?id=2702520

> "In this paper, we present the results of studies in which we characterize the gender bias present in [Google] image search results for a variety of occupations. We experimentally evaluate the effects of bias in image search results on the images people choose to represent those careers and on people's perceptions of the prevalence of men and women in each occupation. We find evidence for both stereotype exaggeration and systematic underrepresentation of women in search results. We also find that people rate search results higher when they are consistent with stereotypes for a career, and shifting the representation of gender in image search results can shift people's perceptions about real-world distributions."

[gender discrimination]
Information Flow Experiments: Determining Information Usage from the Outside
https://www.cs.cmu.edu/~mtschant/ife/

"Using our rigorous statistical methodology, we have analyzed ads served by Google. We explored how they are related to the interests Google claims to infer about people at its Ad Settings webpage. We found:

Discrimination: gender-based discrimination in job-related ads

Opacity: browsing substance abuse websites leads to rehab ads despite Google's own Ad Settings showing no evidence of such tracking

Choice: Google's Ad Settings allows some control over the ads you see"

[data leaks]
Krishnamurthy, B. and Wills, C.E., 2009, August. On the leakage of personally identifiable information via online social networks. In Proceedings of the 2nd ACM workshop on Online social networks (pp. 7-12). ACM. http://dl.acm.org/citation.cfm?id=1592668

"The results of our study clearly show that the indirect leakage of PII [Personally identifiable information"] via OSN ["Online Social Networks"] identifiers to third-party aggregation servers is happening. OSNs in our study consistently demonstrate leakage of user identifier information to one or more third-parties via Request-URIs, Referer headers and cookies. In addition, two of the OSNs directly leak pieces of PII to third parties with one of the OSNs leaking zip code and email information about users that may not be even publicly available within the OSN itself."

*2.5 POLITICS*

[political persuasion]
Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E. and Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), pp.295-298. https://www.nature.com/nature/journal/v489/n7415/abs/nature11421.html

"Here we report results from a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during the 2010 US congressional elections. The results show that the messages directly influenced political self-expression, information seeking and real-world voting behaviour of millions of people. Furthermore, the messages not only influenced the users who received them but also the users' friends, and friends of friends."

*2.6 ECONOMICS*

[race & space discrimination]
https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk

> "Our analysis of premiums and payouts in California, Illinois, Texas and Missouri shows that some major insurers charge minority neighborhoods as much as 30 percent more than other areas with similar accident costs."

[at-risk targeting]
https://www.aclu.org/blog/privacy-technology/senate-report-opens-window-hidden-world-data-aggregators (re: Senate Report on data brokers)

> "These data elements can be extraordinarily personal. Profiles include not only basic demographic information (names, addresses, telephone numbers, e-mail addresses, gender, age, marital status, presence of and ages of children in household), but also such things as profession, education level, income level, religious and political affiliations, real estate information, and sensitive health and financial information. Even information about the weight of household member can be included.
>
> Furthermore, the report found that such data aggregation is conducted "behind a veil of secrecy" from both the public and the government."
>
> "While having an idea of the financial status of individuals can be a helpful tool, the industry's categories appear to focus on our most at-risk populations. This information can result in predatory business practices that target the poor, elderly, or other vulnerable populations. It can also result in differential pricing."
>
> "The report also raises questions about whether the industry is skirting key existing legal protections."

*2.7 MACHINE LEARNING (GENERAL)*

[gender discrimination]
Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K.W., 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv preprint arXiv*:1707.09457. https://homes.cs.washington.edu/~my89/publications/bias.pdf

> "In this work, we study data and models associated with multilabel object classification and visual semantic role labeling. We find that (a) datasets for these tasks contain significant

gender bias and (b) models trained on these datasets further amplify existing bias. For example, the activity cooking is over 33% more likely to involve females than males in a training set, and a trained model further amplifies the disparity to 68% at test time."

[gender discrimination]
Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357). https://arxiv.org/abs/1607.06520

"The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases."

[human semantic bias]
Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), pp.183-186. http://science.sciencemag.org/content/356/6334/183.full

"Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names."

## 3. RESOURCES

### *3.1 SURVEYS/LITERATURE REVIEWS*

[tracking by web companies]
Bujlow, T., Carela-Español, V., Solé-Pareta, J. and Barlet-Ros, P., 2015. Web tracking: Mechanisms, implications, and defenses. *arXiv preprint arXiv*:1507.07872. https://arxiv.org/abs/1507.07872

> "This articles surveys the existing literature on the methods currently used by web services to track the user online as well as their purposes, implications, and possible user's defenses."

[data tracking/analysis/use by corporations]
Christl, W., Kopp, K. and Riechert, P.U., 2017. *CORPORATE SURVEILLANCE IN EVERYDAY LIFE.*
http://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf

> "How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions"

[data tracking/analysis/use by corporations]
Christl, W., and S. Spiekerman. "Networks of control: A report on corporate surveillance, digital tracking, big data & privacy." *Facultas Verlags-und Buchhandels AG*, Wien (2016).
http://crackedlabs.org/en/networksofcontrol

> "A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy"

[data tracking/analysis/use by corporations]
Aaron Rieke, Harlan Yu, David Robinson, and Joris von Hoboken. "Data Brokers in an Open Society". *Uptern / Open Society Foundations*.
https://www.opensocietyfoundations.org/sites/default/files/data-brokers-in-an-open-society-20161121.pdf

> "However, on the whole, policymakers, civil society groups, academics, and journalists have struggled to articulate concrete harms related to emerging data broker practices, aside from an abstract erosion of privacy and a lack of awareness of individuals about what is going on. These difficulties are likely to persist" (26).

[transparency and machine learning]
Edwards, L., and M. Veale. 2017. Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You are Looking For.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855.

[deanonymizing data]
Ohm, P., 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. http://www.uclalawreview.org/pdf/57-6-3.pdf

[social network privacy]
Zheleva, E. and Getoor, L., 2011. Privacy in social networks: A survey. In *Social network data analytics* (pp. 277-306). Springer US. https://link.springer.com/chapter/10.1007/978-1-4419-8462-3_10

> "In this chapter, we survey the literature on privacy in social networks. We focus both on online social networks and online affiliation networks. We formally define the possible privacy breaches and describe the privacy attacks that have been studied. We present definitions of privacy in the context of anonymization together with existing anonymization techniques."

[advertising data / data brokers]
Everything We Know About What Data Brokers Know About You
https://www.propublica.org/article/everything-we-know-about-what-data-brokers-know-about-you

[web tracking (fingerprinting, evercookies, and "cookie syncing)]
Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A. and Diaz, C., 2014, November. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 674-689). ACM. https://securehomes.esat.kuleuven.be/~gacar/persistent/the_web_never_forgets.pdf

## 4. DIGITAL RESEARCH METHODS

### *4.1 GUIDES/TUTORIALS*

Diakopoulos, N., 2014. Algorithmic accountability reporting: On the investigation of black boxes. https://academiccommons.columbia.edu/catalog/ac:2ngf1vhhn4

> Outlines possibilities for reverse engineering algorithms (both where inputs are know and unknown) and shows some examples of how people have done it. Ends with legal and ethical implications.
>
> Three examples:
>
> (1) attempt to figure out google autocomplete blocking: " is case illustrates an ideal situation for the use of algorithmic-accountability reporting. Some transparency by the services

through their FAQ's and blogs suggest a hypothesis and tip as to what types of input the algorithm might be sensitive to (i.e., pornography and violence-related words). More- over, the algorithms themselves, both their inputs and outputs, are observable and accessible through APIs, which make it relatively easy to quickly collect a wide range of observations about the input-output relationship." (16)

(2) "The Message Machine,28 as it came to be called, tried to reverse engineer how the campaign was using targeting information to adapt and personalize email messages for different recipients. In addition to collecting the emails, ProPublica solicited the recipients to fill out a survey asking about basic demographic information, where they lived, and if they had donated or volunteered for the campaign before. These survey answers then served as the input to the algorithm they were trying to dissect. In this case, the output was observable—crowdsourced from thousands of people—but the types of inputs used by the targeting algorithm were hidden behind the campaign wall and thus not controllable by journalists." (18)

correlation != causation as reverse engineers claimed age was important variable, when it was actually donation history

(3) price differences online: "To get the story the WSJ had to simulate visiting the various sites from different computers and browsers in different geographies.30 is initially required using various proxy servers that made it appear like the website was being loaded from different geographies. The publication's staff also created different archetype users and built user profiles using cookies to see how those user profiles might impact the prices recorded. is case again mimics Figure 1(A), wherein both inputs and outputs are fully observable. Yet, it was more complex than that of the autocomplete algorithm since a straightforward API wasn't available. Instead, the journalists had to painstakingly reconstruct profiles that simulated inputs to the algorithm, and look to see if any of the variables in those profiles led to significant differences in output (prices)." (19)

Tramér, F., F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. *USENIX Security Symposium* :601–618. https://www.usenix.org/sites/default/files/conference/protected-files/security16_slides_tramer.pdf

Detailed and highly technical tutorial on how to 'duplicate the functionality' of a black-boxed machine learning model using data culled from the model's API.

Ateniese, G., L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici. 2013. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks* 10 (3):137–150. https://arxiv.org/abs/1306.4447

> "We show that it is possible to infer unexpected but useful information from ML classifiers. In particular, we build a novel meta-classifier and train it to hack other classifiers, obtaining meaningful information about their training sets."
>
> example: "we hacked a speech recognition system and were able to determine the accent of speakers employed during its training."

*4.2 EXAMPLES*

Sweeney, L. 2013. Discrimination in online ad delivery. Queue 11 (3). http://dl.acm.org/citation.cfm?id=2460278

> To determine racial bias in google ads, researcher:
>
> 1) searched for predominately black & white first names + degree (ie: MA, PHD, etc) to find full names of real professionals (and check race from pics)
> searched for black/white names on peekyou to get full names of 'netizens'
> 2) searched for names on google, clear cookies and cache and repeat on Reuters
> 3) analyzed content of returned ads, comparing difference between white/black names

Mukherjee, A., V. Venkataraman, B. Liu, and N. S. Glance. 2013. What yelp fake review filter might be doing? *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

> Compares dataset of filtered and non-filtered reviews on Yelp to infer what the website's fake review filter is using as criteria. Through linguistic analysis and behavioral analysis, they show the latter is more likely used in filtering.

*4.3 ADVANCED METHODS*

Tickle, A. B., R. Andrews, M. Golea, and J. Diederich. 1998. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9 (6):1057–1068.