Research Statement

Roxana Geambasu, Columbia University

My research spans broad areas of computer systems, including distributed systems, operating systems, security and privacy, plus the application of machine learning (ML), causal inference, and cryptography to systems problems. Across these areas, my work anticipates and addresses the unseen costs posed by emerging technologies. For example, I have studied the move to cloud computing; the pervasive adoption of small-form, mobile devices; and the increased collection and mining of personal data by myriad ML-driven services and applications. In all of these domains, I focus on exposing problems and inventing solutions to threats to our data security and privacy, and I build scalable distributed infrastructures that secure both.

At Columbia, I have particularly focused on the development of a *new model for privacy* in our data-driven web world. Today's web, a complex ecosystem, is largely driven by the collection, mining, and monetization of personal data. Many web services, mobile applications, and third-party trackers collect and use our personal data for varied purposes, e.g., to target advertisements, personalize recommendations, and fine-tune prices. Some of these uses offer benefits to us (e.g., recommendations on Netflix or Pandora), but others may impose serious personal costs. At present, we have no window into how our data is being managed and used, and there is limited accountability over the services, raising the risk for deceptive, unfair, or insecure practices.

I seek a new model for how we address such personal privacy and equity issues. I envision a web environment where users are more aware of the privacy consequences of their online actions and make more informed decisions about the services they use. In my model, services and applications are held accountable for their actions and are explicitly constructed to protect user privacy and best interests. To forge this new environment, I design, build, and evaluate:

1. *New transparency systems for privacy watchdogs* to increase society's oversight regarding how applications use personal data in order to detect and deter unfair and deceptive practices.

2. *New testing and certification tools for programmers* to assist in development of data-driven applications that are more privacy-preserving, equitable, and robust *by design*.

3. *New protection abstractions* that facilitate and promote a more rigorous and selective approach to data access, sharing, and retention in today's data-driven companies.

My privacy work has garnered the following recognition: two best paper awards; technology transfers to two companies; invitations to address the Federal Trade Commission (FTC), members of the Cybersecurity Caucus of the U.S. House of Representatives, and members of the Intelligence Community; ongoing collaborations with the FTC, the Federal Drug Administration (FDA), and the New York Presbyterian (NYP) hospital; media coverage (e.g., The New York Times, National Public Radio, and The Economist); and prestigious faculty awards, including Alfred P. Sloan, NSF CAREER, Microsoft Faculty Fellowship, Google Research, and Popular Science Brilliant 10.

1 Transparency Systems for Privacy Watchdogs

Are third-party web trackers watching our children's online activities and targeting them with ads for objectionable products? Do shopping, loan, or housing websites tailor their prices based on what they know about us? Is our data being shared with unexpected parties and, if so, what are those parties doing with it? Today, we lack verifiable answers to questions about how our data is being used by the black-box web services and third-party groups that collect it. We must rely on explanations or privacy statements that companies make, which are often vague, incomplete [3], or (as our research found) inaccurate [40]. And *privacy watchdogs* – such as FTC investigators and investigative journalists – lack the tools needed to track online personal data flows to discover deceptive or unfair practices.

Under my leadership, my collaborators and I have been building the first scalable, generic, and interpretable systems that detect data flows within and across web services. We call these *transparency systems* and their purpose is to facilitate measurement studies of web service use of personal data by privacy watchdogs. Thus far, we've published two transparency systems – XRay [37] and Sunlight [40] – both focused on detecting data use for targeting and personalization based on causal inference from randomized experiments with synthetic user accounts.

XRay [37]. The key insight in our first system, XRay, is to infer targeting by correlating user inputs (such as searches, emails, or locations) to service outputs (such as ads, recommendations, or prices) based on observations obtained from synthetic user accounts populated with randomized subsets of the inputs. XRay was the very first system to apply this methodology to detect online targeting in a generic way. After XRay's introduction in 2014, other systems appeared with similar goals (e.g., CMU's AdFisher [13]). A unique property of XRay remains its scalability to a large number of inputs, enabling large-scale studies of targeting on many inputs. Indeed, XRay's algorithms need a number of synthetic accounts that grows only logarithmically with the number of inputs to track, a property that my collaborators and I showed both experimentally and theoretically. However, the lack of tests of statistical validity for XRay's predictions make its conclusions difficult to interpret.

Sunlight [40]. Our second system, *Sunlight* [40], offers statistical justification plus a clean, causal interpretation for its inferences while preserving the same scaling properties as XRay. To date, Sunlight remains the most scalable system for detecting the causal effect of online targeting. Two key insights enable Sunlight to determine targeting causes at scale. First, Sunlight decouples the formulation and the statistical testing of targeting hypotheses. This lets us use scalable ML models (such as Lasso regressions) to formulate hypotheses on a training set, and assess statistical significance on a testing set. Second, statistical methods to control for multiple hypothesis testing heavily penalize incorrect assumptions. Sunlight thus favors high-precision

algorithms to generate hypotheses which, counter-intuitively, result in better recall after p-value correction. We've developed several such algorithms and evaluated their properties with multiple real-world studies of targeting of ads on Gmail, ads on arbitrary websites, recommendations on YouTube and Amazon, and prices in several travel websites. Finally, the correlations Sunlight detects have a clean, causal interpretation because input values are randomly assigned in each synthetic account independent of other factors.

Using Sunlight, we ran the *largest-scale studies of ad targeting* on Gmail and on the web [40]. We found solid evidence that *contradicted two of Google's own statements* related to targeting on sensitive topics. For example, in an FAQ about Gmail ad targeting, Google used to state that "[they] will not target ads based on sensitive information, such as race, religion, sexual orientation, health, or sensitive financial categories." We created Gmail accounts that exchanged emails containing keywords related to each of these topics. We collected 20M unique ads that Gmail delivered to those accounts over a period of 30 days in September-October 2014. We then used Sunlight to reverse the ads' targeting against single or small combinations of our emails, and kept only the targeting results for which we had solid statistical evidence (corrected p-values < 0.05). We found numerous ads that targeted *each and every of the sensitive topics* that Google promised to prohibit (examples are in our paper and the full dataset is available online [40]). Some of the targeting we found was also quite disturbing. For example, there were numerous ads for questionable financial products, such as payday and subprime loans, that targeted users who expressed financial concerns in their private communications (e.g., used "unemployed" or "bankrupt" keywords in their emails). There were ads for shamanic healing and alcohol-related products that targeted users who expressed health concerns (e.g., used "cancer," "Alzheimer's," "depressed," or "sad" keywords in their emails).

Current Work and Impact. These results attracted attention not only from the media but also from the FTC, who are interested in scalable systems to automate their largely manual studies of deceptive targeted advertising. This has opened a collaboration channel with them that has led to one of my students doing an internship at the FTC to better understand their processes and formulate suitable system requirements. They

specifically expressed interest in leveraging observational data supplied by cohorts of volunteer users to study targeting effects at scale. The observational setting is more meaningful for the FTC because synthetic accounts are expensive to create and maintain at scale. However, inferring causality from observational data is notoriously challenging. We are thus developing a new transparency system that combines methods from causal inference in observational settings (such as difference in differences and factor analysis) with a system for automatic experiment generation that (1) formulates targeting hypotheses from observational data, (2) filters them based on watchdog-provided filters, and (3) subsequently generates small-scale, randomized experiments to confirm causation.

Since I started working on web transparency, the topic has gained substantial traction. I believe that my work was the first and remains one of the few to address the opacity of web services in a generic, scalable, and systems-oriented way. I too have been recognized as a core systems contributor to this space: ACM introduced a new conference dedicated to transparency and related topics (FAT*) and in 2018 they invited me to serve as a track chair for this conference. Before that conference was created, I was part of the program committee for the DAT workshop that ultimately led to FAT*. DARPA's advisory group, ISAT – of which I served as a member for three years – commissioned me to organize a workshop around the topic of data transparency to galvanize a community around it.

2 Testing and Certification Tools for ML Developers

By constructing transparency tools and placing them into the hands of privacy watchdogs, I hope to increase society's oversight on the web services' activities and incentivise developers to construct services that are more protective of users' privacy and best interests. This leads to the second thrust of my privacy research: creating development tools that facilitate the construction of applications that are more privacy-preserving, equitable, and robust *by design*. Today, creating such applications is hard particularly in the context of ML-driven applications, which can have unexpected behaviors, such as biased [6, 4] or offen-

sive predictions [18], or predictions that change unnaturally upon very slight changes to the input [34]. Such behaviors can instill mistrust in ML-driven applications by the public [33], prevent their adoption in legally constrained domains (such as hiring, housing, and credit, which have strict laws preventing discriminatory effects), and even open new vectors of attack (such as adversarial examples [34]). Currently, such unexpected or discriminatory behaviors, which I consider to be a new class of bugs characteristic of ML applications, are difficult to weed out by developers because of a lack of systematic tools to discover and debug them.

My collaborators and I are developing tools for systematic and rigorous testing of fairness, robustness, and privacy properties of ML models by the developers that create these models. To date, we have published two tools: *FairTest* [36], a tool that detects violations of desirable fairness properties in ML-driven applications, reports them as bugs to programmers, and assists in debugging them; and *PixelDP* [41], a method to construct an ML model to guarantee a level of robustness against adversarial example attacks. Perhaps surprisingly, the two classes of properties are related, and we are now working on a single framework for formally verifying a variety of fairness, robustness, and privacy properties of ML models.

FairTest [36]. FairTest is a tool that developers can use to systematically check for discriminatory effects of an ML model on slices of their user population. A programmer submits: the trained model (such as a model to predict hospital readmission based on patient medical history); a test set that includes not only patient features that the model takes as input for inference but also protected attributes, such as race, age, or gender, on which the programmer wants to ensure his model does not discriminate. The programmer also supplies the fairness criteria from a variety of group fairness definitions that FairTest supports, including statistical parity, equalized odds, and equal opportunity. FairTest evaluates the model on the test set, and searches for interpretable contexts of the user population (such as users with particular pre-existing conditions) where the model's outputs violate the fairness criteria.

When I started working on FairTest, significant amount of definitional/theoretical work already existed (e.g., [27, 23, 7, 32, 25, 48, 19, 16, 14, 47]), but the space was very fragmented, with techniques for detection/enforcement designed for specific, limited-scope definitions of fairness. There were also no tools for systematic detection of discriminatory effects. The conceptual contributions in FairTest were two-fold. First, it unified and rationalized 13 fairness definitions that existed at the time and supported them with a common, association-based abstraction. Second, it provided an efficient algorithm to detect violations of specified fairness criteria not only at full population level, but also within smaller contexts where the discriminatory effect was stronger. Inspired by decision-tree classifiers, our algorithm recursively split the user space into smaller subsets so as to maximize the strength of the violation of the fairness criteria. Each step yielded subpopulations of decreasing size and increasingly strong violations. FairTest exposed these violations to programmers as "bugs," sorting them in decreasing order of their strength. Using these tools, a programmer could explore how their programs impacted the users in their population.

Using FairTest on a model for hospital readmission prediction that won a Heritage Health Competition, we discovered that while the model had high accuracy overall, its errors were unevenly concentrated on elderly patients, and the discriminatory effect was particularly strong within populations with certain pre-existing conditions, where error could reach levels as high as 45%. An insurance company that used this algorithm to tune insurance premium might involuntary discriminate against these elderly patients. Also using FairTest, we discovered that when conditioning on the model's confidence in the prediction, the bias against the elderly population disappeared. Thus, one way to address the imbalance would have the insurance company only use the algorithm's high-confidence decisions to make automatic decisions about premiums.

PixelDP [41]. PixelDP is a method for constructing an ML model that gives a certified level of robustness against adversarial example attacks. In these attacks, an adversary finds a small perturbation to an otherwise correctly classified input, which results in a misclassification. For example, an adversary wearing makeup may fool a face recognition model into classifying him/her as someone else, enabling access to a phone or

building if the model's predictions are used for authentication. These attacks are debilitating to undefended networks, bringing their accuracy to zero while remaining un-noticeable to human observers. Numerous defenses have been proposed, but most were best-effort approaches (e.g., [26, 20, 22, 10, 17, 11, 29, 21]. that were broken by subsequent attack versions [9, 8, 5]. Recently, a new class of *certified defenses* have emerged that come with a guaranteed level of robustness against arbitrary attack implementations [12, 30, 46]. However, they either do not scale to large models or are not sufficiently flexible to apply to all relevant model structures.

PixelDP is the first certified defense that applies to large models and is agnostic to model structure. It is built on *differential privacy*, a theory from the privacy domain. Briefly, the expected output of a differentially private mechanism can be shown to be bounded under small changes in its input. We use this fact to assess whether any adversarial attack below a given size can change the prediction of a PixelDP model on a given input. If it cannot, the prediction is deemed certifiably robust against attack up to that size. This *robustness certificate* for an individual prediction can be used in two ways. First, a building authentication system could use it to decide whether a prediction is sufficiently robust to rely on the face recognition model to make an automated decision, or whether a human should be consulted. Second, a model designer can use robustness certificates for predictions on a test set to assess a lower bound on the accuracy under attack. This bound holds for arbitrary norm-bounded attacks, so there is no danger of it ever being broken by future instantiations of such attacks.

Using PixelDP, we produced the first version of Google's Inception deep neural network for ImageNet that has non-trivial guaranteed accuracy under arbitrary, norm-bounded adversarial examples. This network is orders of magnitude larger than what previous certified defenses handled. The guaranteed accuracy is reasonable for small attacks that are still invisible to a human eye (e.g., 60% for 2-norm attacks of size 0.1). On state-of-the-art contemporary attacks, PixelDP does better than the guarantee predicts: it maintains an accuracy above 60% for four times as larger attacks. Moreover, if one is willing to only act on predictions

that PixelDP certifies as robust (as in the preceding building authentication scenario), then accuracy jumps significantly for the approximately 70% of the predictions that PixelDP deems robust.

Current Work and Impact. Although PixelDP has yet to be presented at IEEE S&P in May 2019, it has already gathered attention and there are several papers on arxiv that build on our approach to improve the bounds at a given noise level [24], use other noise distributions [28], and adapt the optimization to the noise [31]. Through a recent Google Faculty award, my collaborators and I are also working to make the guarantee practical for larger attacks by reducing the impact of the differentially private noise added during training and prediction.

More broadly, FairTest and PixelDP are examples of my broader vision of developing a framework for systematic testing and certification of important properties for ML systems, such as those related to fairness, robustness against attacks, and privacy. Perhaps surprisingly, these three classes of properties are related. For example, individual fairness [15] requires that whenever two users are "similar" (according to some distance function), an ML model's predictions should also be close. This is very similar to a distance-based definition of robustness against adversarial examples. And as our use of differential privacy to guarantee robustness shows, robustness and privacy are also related. This suggests that we may be able to build a unified and comprehensive framework for certifying many types of desirable properties for ML models, and perhaps leverage techniques from certifying robustness (including ours) to certify fairness. This would constitute a significant advancement of the algorithmic fairness literature, which thus far been characterized by best-effort, empirical assessments on limited testing sets.

3 Protection Abstractions for Data Ecosystems

The third thrust of my privacy research revisits decades' old protection abstractions from operating systems (OS), which are unfit for emerging data-driven workloads. Traditional protection units, such as files, directories, or database tables, fail to support the data access and sharing patterns common in ML ecosys-

tems. This leads some companies to adopt either too loose, wide-access policies on the data (e.g., all engineers and processes within the company get access to all user data), or too restrictive, siloing policies (e.g., the data from service X is beyond reach for any engineer or process outside X's scope). Neither extreme is good: the former can result in wide exposure to hackers or snooping employees; the latter can disable potentially valuable uses of the data.

I believe that the right protection abstraction for ML ecosystems is not files/directories but *ML models*, particularly *feature models*. Most ML predictive models are not built directly on the raw data, but on features that encode, aggregate, and otherwise transform the raw data so learning can be done more efficiently on it. It is these features, such as user embeddings, covariates, and various statistical aggregates, that are often being shared across teams in big companies. I thus propose the notion of a *private feature model*, a new abstraction for protected data access and sharing for ML ecosystems. A private feature module is a feature model that is learned incrementally over historical data, is made differentially private to bound leakage of that data through its parameters, and is made broadly accessible within the company such that any engineer can use it to improve their service, ideally in lieu of sharing the raw data.

Pyramid [38, 39]. My collaborators and I developed, evaluated, and published an initial system, called Pyramid, that implements a simple, special-case of this abstraction based on a particular feature model called count featurization. Count featurization is a frequently used method for scaling ML learning to giant datasets. It works by replacing the features of an example with the probability of its label, conditioned on the feature values. For instance, for a movie rating, the feature $userId$ becomes $P(rating|userId)$. Because the new features are lower dimension, predictive models tend to require much less data to fit. Pyramid keeps a set of count tables to compute these probabilities and adds differentially private noise to safely store and share these count tables. Using a set of new mechanisms, like a count-sketch that interacts well with differential privacy to store large count tables, Pyramid comes within 4% of state-of-the-art models' accuracy while training on, and thus exposing, less than 1% of the raw data.

I am now generalizing this work to support arbitrary feature models [43]. This generality raises substantial technical challenges, because each feature model exposes a different privacy-accuracy tradeoff when made DP. This complicates the design of a uniform abstraction and system. I am addressing these challenges through a combination of novel DP theory and systems techniques for resource allocation to manage the privacy resource judiciously and enforce accuracy targets for all private feature models in the system.

Pebbles [42]. A similar mismatch between traditional protection abstractions and the needs of emerging workloads occurs in mobile devices. While traditional OS standards provide low-level storage abstractions – files and directories – modern OSes embed higher-level abstractions, such as relational databases or object-relational models. Despite the change in abstraction, many crucial protection systems, such as encryption or deniable systems, continue to operate at the old file level, which often renders them ineffective (i.e., files are both too fine grained and too coarse grained for effective protection). In response to this abstraction mismatch, my collaborators and I developed *Pebbles*, an OS-level service that introduces a new and more meaningful abstraction for protecting persistent data in modern OSes: *logical data objects*. These objects correspond to user-relevant objects, such as emails, documents, and pictures. Pebbles reconstructs these objects based on pieces stored in various database and blob files, providing them as abstractions to protection tools. These tools can then enable data protection at a level that is relevant to users, such as encrypting, hiding, or properly deleting individual emails, documents, or bank accounts along with all pieces of data related to them.

CleanOS [35]. XXX Will have a brief description of this paper here.

Current Work and Impact. XXX Will talk about project with NY Presbyterian and the FDA on sharing differentially private models of clinical data to enhance and democratize biomedical research.

4 Beyond Privacy

My primary research focuses on privacy, the fragile state of which drives my relentless efforts. But my curiosity and aptitude span other aspects of computer systems, as well, especially distributed and operating systems. Across these areas, my research aims to explore and resolve other challenges posed by emerging technologies.

As one example, I observed that the growing demand for data-driven features in today's web applications – such as targeting, recommendations, or predictions – has transformed those applications into ad-hoc, overly complex ecosystems of services that operate on distinct databases and attempt to integrate their data in the absence of a coherent systems architecture. We developed *Synapse*, an easy-to-use system that allows the clean integration of heterogeneous-database web services into a uniform, managed architecture [45]. Synapse lets independent services run atop their own databases (such as Oracle, MySQL, Cassandra, MongoDB, and others) and incorporate read-only views of each others' shared data. Synapse synchronizes these views in real-time and at scale and provides abstractions that let programmers handle impedance mismatches between different database engines. We deployed Synapse at a NYC startup, where it has been running successfully for two years.

As another example, I observed that today's applications running on modern operating systems (OSes) – such as Android, iOS, and OSX – require very different abstractions from the abstractions offered by traditional OS standards, such as POSIX. The new abstractions, currently implemented atop POSIX abstractions, are offered as part of user-space libraries. A question that arises is how well are the new abstractions supported by the traditional ones? To find out, my collaborators and I measured POSIX in Android, iOS, and OSX, to better understand POSIX abstraction use in modern workloads [44]. We found that many of the new OS abstractions rely upon POSIX's unstructured extension interfaces (such as the `ioctl` function call) to implement their functionality, suggesting serious mismatches between traditional and modern OS abstrac-

tions. We also found that the file system XXX. This observation guided our design of the Pebbles system for mobile data protection (see previous section).

Finally, I worked on improving virtual machine migration by incorporating past state access histories [1, 2]. With this information, we show that we stream large virtual machines over the wide area and even cellular networks with limited loss in interactivity.

5 Summary

Across all these research areas, I bring a passionate vision of how emerging computing technologies can enrich our lives without imposing undue or unknowable personal costs. Much of my Columbia research focuses on a new model for privacy that is suitable for the emerging "big data" world. My model involves building and deploying a scalable external transparency infrastructure for the web that will increase society's oversight on web services' use of personal data while providing the critical missing tools needed by the services to safeguard this data. Unlike other models for privacy, which rely on protection or prevention of data access by the web services, my model puts forward accountability and personal responsibility as key new components on which to build a more private data-driven world.

References

[1] Yoshihisa Abe, Roxana Geambasu, Kaustubh Joshi, H. Andres Lagar-Cavilla, and Mahadev Satyanarayanan. vTube: Efficient streaming of virtual appliances over last-mile networks. In *Proc. of the Symposium on Cloud Computing (SoCC)*, 2013.

[2] Yoshihisa Abe, Roxana Geambasu, Kaustubh Joshi, and Mahadev Satyanarayanan. Urgent virtual machine migration with enlightened post-copy. In *Proceedings of the Conference of Virtual Execution Environments (VEE)*, 2016.

[3] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and

Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of Facebooks explanationn. In *Proc. of the Annual Network and Distributed System Security Symposium (NDSS)*, 2018.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016.

[5] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. 2018.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, 2016.

[7] Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017.

[9] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[10] Chuan Guo, Mayank Rana, Moustapha Cisse, Laurens van der Maaten. Countering adversarial images using input transformations. *International Conference on Learning Representations*, 2018.

[11] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille. Mitigating adversarial effects through randomization. *International Conference on Learning Representations*, 2018.

[12] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[13] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In *Proc. of Privacy Enhancing Technologies Symposium (PETS)*, 2015.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 2012.

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[17] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. *International Conference on Learning Representations*, 2018.

[18] Jessica Guynn. Google photos labeled black people 'gorillas'. USA Today, July 2015.

[19] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 25(7):1445–1459, 2013.

[20] Dan Hendrycks and Kevin Gimpel. Early methods for detecting adversarial images. In *ICLR (Workshop Track)*, 2017.

[21] Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G. Dimakis. The robust manifold defense: Adversarial training using generative models. *CoRR*, abs/1712.09196, 2017.

[22] Jacob Buckman, Aurko Roy, Colin Raffel, Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. *International Conference on Learning Representations*, 2018.

[23] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE, 2010.

[24] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. *arXiv:1809.03113*, 2018.

[25] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510. ACM, 2011.

[26] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. of IEEE Symposium on Security and Privacy (Oakland)*, 2016.

[27] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568. ACM, 2008.

[28] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization: the case of the exponential family. *arXiv:1902.01148*, 2019.

[29] Pouya Samangouei, Maya Kabkab, Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representations*, 2018.

[30] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[31] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv:1811.09310*, 2018.

[32] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law*, 18(1):1–43, 2010.

[33] Aaron Smith and Monica Anderson. Americans' attitudes toward hiring algorithms. `http://www.pewinternet.org/2017/10/04/americans-attitudes-toward-hiring-algorithms/`, October 2017.

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[35] Yang Tang, Phillip Ames, Sravan Bhamidipati, Ashish Bijlani, Roxana Geambasu, and Nikhil Sarda. CleanOS: Mobile OS abstractions for managing sensitive data. In *Proc. of the USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2012.

[36] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven

applications. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, 2017.

[37] Mathias Lecuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. XRay: Increasing the web's transparency with differential correlation. In *Proc. of USENIX Security*, 2014.

[38] Mathias Lecuyer, Riley Spahn, Roxana Geambasu, Tzu-Kuo Huang, and Siddhartha Sen. Pyramid: Enhancing selectivity in big data protection with count featurization. In *Proceedings of the IEEE Security and Privacy Symposium (IEEE S&P)*, 2017.

[39] Mathias Lecuyer, Riley Spahn, Roxana Geambasu, Tzu-Kuo Huang, and Siddhartha Sen. Enhancing selectivity in big data. In *IEEE Security and Privacy Magazine*, 2018.

[40] Mathias Lecuyer, Riley Spahn, Yannis Spiliopoulos, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proc. of the ACM Conference on Computer and Communications Security (CCS)*, 2015.

[41] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *Proceedings of the IEEE Security and Privacy Symposium (IEEE S&P)*, 2019.

[42] Riley Spahn, Jonathan Bell, Michael Lee, Sravan Bhamidipati, Roxana Geambasu, and Gail Kaiser. Pebbles: Fine-grained data management abstractions for modern operating systems. In *Proc. of the USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2014.

[43] Riley Spahn, Mathias Lecuyer, Kiran Vodrahalli, Roxana Geambasu, and Daniel Hsu. Differentially private management of machine learning models. In preparation, 2019.

[44] Vaggelis Atlidakis, Jeremy Andrus, Roxana Geambasu, Dimitris Mitropoulos, and Jason Nieh. POSIX abstractions in modern operating systems: The old, the new, and the missing. In *Proceedings of the IEEE European Conference on Computer Systems (EuroSys)*, 2016.

[45] Nicolas Viennot, Mathias Lecuyer, Jonathan Bell, Roxana Geambasu, and Jason Nieh. Synapse: New data integration abstractions for agile web application development. In *Proc. of the ACM European Conference on Computer Systems (EuroSys)*, 2015.

[46] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

[47] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

[48] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 992–1001. IEEE, 2011.