The Values Encoded in Machine Learning Research

Abeba Birhane[*]
University College Dublin & Lero
Dublin, Ireland
abeba.birhane@ucdconnect.ie

Pratyusha Kalluri[*]
Stanford University
pkalluri@stanford.edu

Dallas Card[*]
Stanford University
dcard@stanford.edu

William Agnew[*]
University of Washington
wagnew3@cs.washington.edu

Ravit Dotan[*]
University of California Berkeley
ravit.dotan@berkeley.edu

Michelle Bao[*]
Stanford University
baom@stanford.edu

Abstract

Machine learning (ML) currently exerts an outsized influence on the world, increasingly affecting communities and institutional practices. It is therefore critical that we question vague conceptions of the field as value-neutral or universally beneficial, and investigate what specific values the field is advancing. In this paper, we present a rigorous examination of the values of the field by quantitatively and qualitatively analysing 100 highly cited ML papers published at premier ML conferences, ICML and NeurIPS. We annotate key features of papers which reveal their values: how they justify their choice of project, which aspects they uplift, their consideration of potential negative consequences, and their institutional affiliations and funding sources. We find that societal needs are typically very loosely connected to the choice of project, if mentioned at all, and that consideration of negative consequences is extremely rare. We identify 67 values that are uplifted in these papers, and, of these, we find that papers most frequently justify and assess themselves based on performance, generalization, efficiency, researcher understanding, novelty, and building on previous work. We present extensive textual evidence and analysis of how these values are operationalized. Notably, we find that each of these top values is currently being defined and applied with assumptions and implications generally supporting the centralization of power. Finally, we find increasingly close ties between these highly cited papers and tech companies and elite universities.

1 Introduction

Over the past few decades, ML has risen from a relatively obscure research area to an extremely influential discipline, actively being deployed in myriad applications and contexts around the world. The objectives and values of ML research are influenced by many factors, including the personal preferences of researchers and reviewers, other work in science and engineering, the interests of academic institutions, funding agencies, and companies, and larger institutional and systemic pressures, including systems of oppression impacting who is able to do research and on which topics. Together these forces shape patterns in what research gets done and who benefits from this research. As such, it is important to document and understand the emergent values of the field: what the field is prioritizing and working toward. To this end, we perform a comprehensive analysis of 100 highly cited NeurIPS and ICML papers from four recent years spanning more than a decade.

Our key contributions are as follows:

---

[*]equal contribution

(1) We develop and open source a fine-grained annotation scheme for the detection of values in research papers. To our knowledge, our annotation scheme is the first of its kind, and opens the door to further qualitative and quantitative analyses of research.[2]

(2) We apply our annotation scheme to annotate 100 influential ML research papers and extract their value commitments, including identifying a list of 67 values uplifted. These reflect and shape the values of the field more broadly. Like the annotation scheme itself, the resulting repository of annotated papers is available and is valuable not only in the context of this paper's analysis, but also as foundation for further qualitative and quantitative study of ML research.

(3) We perform extensive textual analysis to understand the dominant values: performance, accuracy, state-of-the-art (SOTA), quantitative results, generalization, efficiency, building on previous work, and novelty. Our analysis indicates that while these values may seem on their face to be purely technical, they are socially and politically charged: specifically, we find that these values are currently defined and operationalized in ways that centralize power, i.e., disproportionally benefit and empower the already powerful, such as large corporations, while neglecting society's least advantaged.

(4) We present a quantitative analysis of the affiliations and funding sources of these most influential papers. We find substantive and increasing presence of big tech corporations. For example, in 2008/09, 24% of these top cited papers had corporate affiliated authors, and in 2018/19 this statistic almost tripled, to 71%. Moreover, of these corporations connected to influential papers, the presence of "big-tech" firms, such as Google and Microsoft, increased more than fivefold, from 11% to 58%.

2   Methodology

To understand the values of ML research, we examined the most highly cited papers from NeurIPS and ICML from the years 2008, 2009, 2018, and 2019. We chose to focus on highly cited papers because they both reflect and shape the values of the discipline, drawing from NeurIPS and ICML because they are the most prestigious of the long-running ML conferences.[3] Acceptance to these conferences is a valuable commodity used to evaluate researchers, and submitted papers are explicitly written so as to win the approval of the community, particularly the reviewers who will be drawn from that community. As such, these papers effectively reveal the values that authors believe are most valued by that community. Citations indicate amplification by the community, and help to position these papers as influential exemplars of ML research. To avoid detecting only short-lived trends and enable comparisons over time, we drew papers from two recent years (2018/19) and from ten years earlier (2008/09). We focused on conference papers because they tend to follow a standard format and allow limited space, meaning that researchers must make hard choices about what to emphasize. Collectively, we annotated 100 papers, analyzing over 3,500 sentences drawn from them. In the context of expert qualitative content analysis, this is a significant scale which allows us to meaningfully comment on the values central to ML research.

In more detail, we began by creating an annotation scheme, and then used it to manually annotate each paper, examining the abstract, introduction, discussion, and conclusion: (1) We examined the chain of reasoning by which each paper justified its contributions, which we call the *justificatory chain*, rating the extent to which papers used technical or societal problems to justify or motivate their contributions. (2) We carefully read each sentence of these sections, annotating any and all values from our list that were uplifted or exhibited by the sentence.[4] (3) We documented the extent to which the paper included a discussion of potential negative impacts.

Manual annotation was necessary, both to create the list of emergent values, and to obtain and richly understand the values present in each paper. Automated approaches, such as keyword searches, would suffer significant limitation: these approaches would only capture values that we anticipate; additionally, they would run the risk of systematically skewing the results towards values which are easy to identify, missing or mischaracterizing values which are exhibited in more nuanced ways. The

---

[2]We include our template and all annotations as supplementary material with a CC BY-NC-SA license.

[3]At the time of writing, these two venues, along with the newer ICLR (2013-present), comprised the top 3 conferences according to h5-index (and h5-median) in the AI category on Google Scholar, by a large margin.

[4]We use a conceptualization of "value" that is widespread in philosophy of science in theorizing about values in sciences. In this approach, a value of an entity is a property that is desirable for that kind of entity. For example, speed can be described as valuable in an antelope [28]. Well-know scientific values include accuracy, consistency, scope, simplicity, and fruitfulness [25]. See [27] for a critical discussion of these values.

qualitative approach was key for analyzing the values as well, as it requires a subtle understanding of how the values function in the text and understanding of taken for granted assumptions underlying the values, which methods such as keyword matching would fail to capture.

To assess consistency, 40% of the papers were annotated by two annotators. The intercoder consensus on values in these papers achieved a Cohen kappa coefficient of .61, which indicates substantial agreement [39]. Furthermore, we used several established strategies to increase consistency, including recoding data coded early in the process [23] and conducting frequent discussions and assessments of the coding process, code list, and annotation scheme [24].

To create the list of emergent values specifically (see Figure 1), we followed best practices in manual content analysis. (1) We began with a list of values we expected to be relevant based on prior knowledge, augmenting this list with seven ethical principles of interest from existing literature [6, 17]. (2) We randomly selected a subset of 10 papers for initial annotation, searching for the values on the list sentence by sentence and adding new values as they emerged. (3) Through discussion, we revisited all values and produced a values list. (4) We annotated the full set of papers using this list of values, meeting regularly to discuss difficult examples, and individually nominating and deciding by consensus when sentences justified inductively adding additional, emergent values to the values list. (5) For the final analysis presented here, we combined closely related values into clusters by consensus, such that they could be discussed together (for completeness, all values are treated separately in the Appendix). Formally stated, we establish our codes (short phrases that represent the relevant essence of information, in this case the list of values) using the an inductive-deductive approach. The deductive component involves starting with codes established in existing literature, which ensures we note and can speak to values of interest, including established ethical principles. The inductive component involves the discovery of codes from the data, and impedes inappropriately biased or pre-conceived findings by focusing on emergent codes [8, 24].

The composition of our team confers additional validity to our work. We are a diverse, multi-racial, multi-gender team working closely, including undergraduate, graduate, and post-graduate researchers from machine learning, NLP, robotics, cognitive science, and philosophy. This team captures several advantages that other methods of manual annotation such as crowd sourcing lack: the nature of this team minimizes intra-disciplinary biases, affords the unique combination of expertise required to read the values in ML papers, allows meaningful engagement with relevant work in other fields, and enables best practices including continually clarifying the procedure, ensuring agreement, vetting consistency, reannotating, and discussing themes [24].
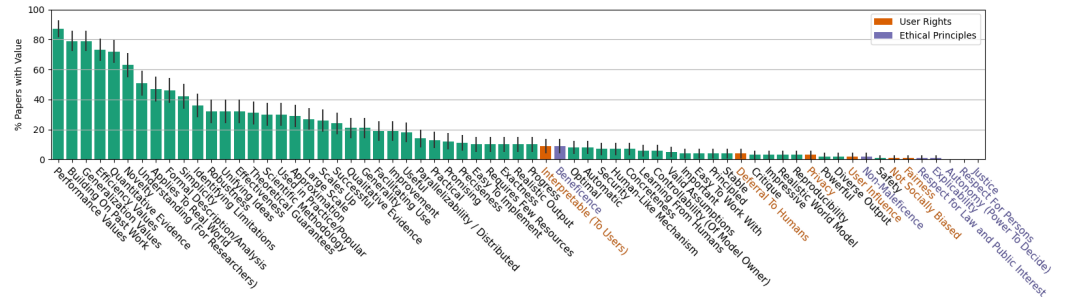

Figure 1: Proportion of annotated papers that uplifted each value.

3 Quantitative Summary

In Figure 1, we plot the prevalence of values in 100 annotated papers. The top values are: performance (87% of papers), building on past work (79%), generalization (79%), efficiency (73%), quantitative evidence (72%), and novelty (63%). Values related to user rights and stated in ethical principles appeared very rarely if at all: none of the papers mentioned autonomy, justice, or respect for persons. In Table 1 (top), we show the distribution of justification scores. Most papers only justify how they achieve their internal, technical goal; 71% don't make any mention of societal need or impact, and only 3% make a rigorous attempt to present links connecting their research to societal needs. In Table 1 (bottom), we show the distribution of negative impact discussion scores. One annotated

Table 1: Annotation scheme and results for justificatory chain (top) and negative impacts (bottom).

Justificatory Chain Condition	% of Papers
Doesn't rigorously justify how it achieves technical goal	1%
Justifies how it achieves technical goal but no mention of societal need	71%
States but does not justify how it connects to a societal need	16%
States and somewhat justifies how it connects to a societal need	9%
States and rigorously justifies how it connects to a a societal need	3%

Negative Impacts Condition	% of Papers
Doesn't mention negative potential	98%
Mentions but does not discuss negative potential	1%
Discusses negative potential	1%
Deepens our understanding of negative potential	0%

paper included a discussion of negative impacts and a second mentioned the possibility; none of the remaining 98 papers contained any reference to potential negative impacts. In Figure 3, we show stated connections (funding and affiliations) of paper authors to institutions. Comparing papers written in 08/09 to those authored 18/19, ties to corporations nearly doubled to 79% of all annotated papers, ties to big tech multiplied over fivefold to 58%, while ties to universities declined to 81%, putting corporations nearly on par with universities in the most cited ML research. In the next sections, we present extensive qualitative examples and analysis of our findings, with additional analyses in the Appendix.

4 Qualitative Analysis of Justifications and Negative Potential

4.1 Justificatory Chain

Papers typically motivate their projects by appealing to the needs of the ML research community and rarely mention potential societal benefits. Research-driven needs of the ML community include researcher understanding (e.g., understanding the effect of pre-training on performance/robustness, theoretically understanding multi-layer networks) as well as more practical research problems (e.g., improving efficiency of models for large datasets, creating a new benchmark for NLP tasks). Some papers do appeal to needs of broader society, such as building models with realistic assumptions, catering to more languages, or understanding the world. However, even when societal needs are mentioned as part of the justification of the project, the connection is often loose. Almost no papers explain how their project is meant to promote a social need they identify by giving the kind of rigorous justification that is typically expected of and given for technical contributions.

4.2 Negative Potential

Two of the 100 papers discussed potential harms, whereas the remaining 98 did not mention them at all. The lack of discussion of potential harms is especially striking for papers which deal with socially contentious application areas, such as surveillance and misinformation. For example, the annotated corpus includes a paper advancing the identification of people in images, a paper advancing face-swapping, and a paper advancing video synthesis. These papers contained no mention of the well-studied negative potential of facial surveillance, DeepFakes, or misleading videos, respectively.

Furthermore, among the two papers that do mention negative potential, the discussions were mostly abstract and hypothetical, rather than grounded in the negative potential of their specific contributions. For example, authors may acknowledge "possible unwanted social biases" when applying the model to a real-world setting, without discussing the social biases encoded in the authors' proposed model.

5 Stated values

The dominant values in ML research, e.g., accuracy and efficiency, may seem purely technical. However, the following analysis of several of these values shows how they can become politically loaded in the process of prioritizing and operationalizing them: sensitivity to the way that they are

Table 2: Random examples of *performance*, the most common emergent value.

"Our model significantly outperforms SVM's, and it also outperforms convolutional neural nets when given additional unlabeled data produced by small translations of the training images."
"We show in simulations on synthetic examples and on the IEDB MHC-I binding dataset, that our approach outperforms well-known convex methods for multi-task learning, as well as related non-convex methods dedicated to the same problem."
"Furthermore, the learning accuracy and performance of our LGP approach will be compared with other important standard methods in Section 4, e.g., LWPR [8], standard GPR [1], sparse online Gaussian process regression (OGP) [5] and v-support vector regression (v-SVR) [11], respectively."

operationalized, and to the fact that they are uplifted at all, reveals value-laden assumptions that are often taken for granted. We thus challenge a conception of prevalent values as politically neural and consider alternatives to their dominant conceptualization that may be equally or more intellectually interesting or socially beneficial.[5] We have encouraged ourselves, and now encourage the reader, to remember that values once held to be intrinsic, obvious, or definitional have been in many cases transformed over time.

To provide a sense of what the values look like in context, we include three randomly selected examples of sentences annotated for each value discussed here (Tables 2-5), with extensive additional examples in the Appendix. Note that most sentences are annotated with multiple values.[6]

5.1 Performance

Performance, accuracy, and achieving SOTA form the most common cluster of related values in annotated papers. While it might seem intrinsic for the field to care about performance, it is important to remember that models are not simply "well-performing" or "accurate" in the abstract but always in relation to and as *quantified* by some metric on some dataset. Examining prevalent choices of operationalization reveals political aspects of performance values. First, we find performance values are consistently and unquestioningly operationalized as correctness averaged across individual predictions, giving equal weight to each instance. However, choosing equal weights when averaging is a value-laden move which might deprioritize those underrepresented in the data or world, as well as societal and evaluee needs and preferences. Extensive research in ML fairness and related fields has considered alternatives, but we found no such discussions among the papers we examined.

Choices of datasets are revealing. They are often driven purely by past work, so as to demonstrate improvement over a previous baseline (see also §5.4). Another common justification for using a certain dataset is applicability to the "real world". Assumptions about how to characterize the real world are also value-laden. One common assumption is the availability of very large datasets. However, presupposing the availability of large datasets is power centralizing because it encodes favoritism to those with resources to obtain and process them [16]. Further overlooked assumptions include that the real world is binary or discrete, and that datasets come with a predefined ground-truth label for each example, presuming that a true label always exists "out there" independent of those carving it out, defining and labelling it. This contrasts against marginalized scholars' calls for ML models that allow for non-binaries, plural truths, contextual truths, and many ways of being [13, 18, 26].

---

[5]By challenging a politically neutral conception of the top values in machine learning research, this paper also contributes to the literature in philosophy. Philosophers of science have been working to understand the roles of values in science for decades. For example, Thomas Kuhn [25] presented a list of five scientific values which he deems as important in scientific research (accuracy, consistency, scope, simplicity, and fruitfulness). Helen Longino [27] and others have argued that prominent values are politically loaded, focusing mostly on how some of these values function in disciplines such as biology and social sciences. However, "technical" values, such as accuracy, are often left out of this type of critical discussion. Our paper shows that even the "technical" values aren't politically neutral, and it does so in the context of machine learning, which is often conceived as a less politically loaded discipline than biology or social sciences.

[6]To avoid the impression that we are drawing attention to anything special about these randomly chosen example sentences, we omit attribution, and include a list of all annotated papers in the Appendix.

Table 3: Random examples of *generalization*, the third most common emergent value.

"The range of applications that come with generative models are vast, where audio synthesis [55] and semi-supervised classification [38, 31, 44] are examples hereof."
"Furthermore, the infinite limit could conceivably make sense in deep learning, since over-parametrization seems to help optimization a lot and doesn't hurt generalization much [Zhang et al., 2017]: deep neural nets with millions of parameters work well even for datasets with 50k training examples."
"Combining the optimization and generalization results, we uncover a broad class of learnable functions, including linear functions, two-layer neural networks with polynomial activation $\phi(z) = z^{2l}$ or cosine activation, etc."

The prioritization of performance values also requires scrutiny. Valuing these properties is so entrenched in the field that generic success terms, such as "success", "progress", or "improvement" are often used as synonyms for performance and accuracy. However, one might alternatively invoke generic success to mean increasingly safe, consensual, or participatory ML that reckons with impacted communities and the environment. In fact, "performance" itself is a general success term that could have been associated with properties other than accuracy and SOTA.

5.2 Generalization

A common way of appraising the merits of one's work in ML is to claim that it generalizes well. Typically, generalization is understood in terms of performance or accuracy: a model generalizes when it achieves good performance on a range of samples, datasets, domains, or applications. Uplifting generalization raises two kinds of questions. First, which datasets, domains, or applications show that the model generalizes well? Typically, a paper shows that a model generalizes by showing that it performs well on multiple tasks or datasets. However, the choice of particular tasks and datasets is rarely justified; the choice of tasks can often seem arbitrary, and authors rarely present evidence that their results will generalize to more realistic settings, or help to directly address societal needs.

Second, uplifting generalization itself reveals substantive assumptions. The prizing of generalization means that there is an incentive to harvest many datasets from a variety of domains, and to treat these as the only datasets that matter for that space of problems. Generalization thus prioritizes distilling every scenario down to a common set of representations or affordances, rather than treating each setting as unique. Critical scholars have advocated for valuing *context*, which stands at the opposite side of striving for generalization [15]. Others have argued that this kind of totalizing lens (in which model developers have unlimited power to determine how the world is represented) leads to *representational* harms, due to applying a single representational framework to everything [14, 1].

Finally, the belief that generalization is even possible implicitly assumes a conservative approach in which new data will be sufficiently similar to previously seen data. When used in the context of ML, the assumption that the future resembles the past is also normative and often problematic as past societal stereotypes and injustice can be encoded in the process [33]. Furthermore, to the extent that predictions are performative [35], especially predictions that are enacted, those ML models which are deployed to the world will contribute to shaping social patterns. Yet, no papers attempt to counteract this quality or acknowledge its presence.

5.3 Efficiency

Efficiency is another common value in ML research. Abstractly, saying that a model is efficient typically means saying that the model uses less of some resource, such as time, memory, energy, or number of labeled examples. In practice however, efficiency is commonly referenced to, ironically, imply the ability to scale up: a more efficient inference method allows you to do inference in much larger models or on larger datasets, using the same amount of resources. This is reflected in our value annotations, where 72% of papers mention valuing efficiency, but only 14% of those value requiring *few* resources. In this way, valuing efficiency facilitates and encourages the most powerful actors to scale up their computation to ever higher orders of magnitude, making their models even less accessible to those without resources to use them and decreasing the ability to compete with them. Alternative usages of efficiency could encode accessibility instead of scalability, aiming to create more equitable conditions for ML research.

6

Table 4: Random examples of *efficiency*, the fourth most common emergent value.

"Our model allows for controllable yet efficient generation of an entire news article – not just the body, but also the title, news source, publication date, and author list."
"We show that Bayesian PMF models can be efficiently trained using Markov chain Monte Carlo methods by applying them to the Netflix dataset, which consists of over 100 million movie ratings."
"In particular, our EfficientNet-B7 surpasses the best existing GPipe accuracy (Huang et al., 2018), but using 8.4x fewer parameters and running 6.1x faster on inference."

Table 5: Random examples of *building on past work* and *novelty*, the second and sixth most common emergent values, respectively.

Building on past work
"Recent work points towards sample complexity as a possible reason for the small gains in robustness: Schmidt et al. [41] show that in a simple model, learning a classifier with non-trivial adversarially robust accuracy requires substantially more samples than achieving good 'standard' accuracy."
"Experiments indicate that our method is much faster than state of the art solvers such as Pegasos, TRON, SVMperf, and a recent primal coordinate descent implementation."
"There is a large literature on GP (response surface) optimization."

Novelty
"In this paper, we propose a video-to-video synthesis approach under the generative adversarial learning framework."
"Third, we propose a novel method for the listwise approach, which we call ListMLE."
"The distinguishing feature of our work is the use of Markov chain Monte Carlo (MCMC) methods for approximate inference in this model."

5.4 Novelty and Building on Past Work

Most authors devote space in the introduction to positioning their paper in relation to past work, and describing what is novel. Mentioning past work serves to signal awareness of related publications, to establish the new work as relevant to the community, and to provide the basis upon which to make claims about what is new. Novelty is sometimes suggested implicitly (e.g., "we develop" or "we propose"), but frequently it is emphasized explicitly (e.g. "a new algorithm" or "a novel approach").

This combined focus on novelty and building on recent work establishes a continuity of ideas, and might be expected to contribute to the self-correcting nature of science [29]. However, this is not always the case [21] and attention to the ways novelty and building on past work are implemented reveals value commitments. In particular, we find a clear emphasis on technical novelty, rather than critique of past work, or demonstration of measurable progress on societal problems, as has previously been observed [40]. Although introductions sometimes point out limitations of past work (so as to further emphasize the contributions of their own paper), they are rarely explicitly critical of other papers in terms of methods or goals. Indeed, papers uncritically reuse the same datasets for years or decades to benchmark their algorithms, even if those datasets fail to represent more realistic contexts in which their algorithms will be used [7]. Novelty is denied to work that rectifies socially harmful aspects of existing datasets in tandem with strong pressure to benchmark on them and thereby perpetuate their use, enforcing a fundamentally conservative bent to ML research.

6 Corporate Affiliations and Funding

Our analysis shows substantive and increasing corporate presence in the most highly-cited papers. In 2008/09, 24% of the top cited papers had corporate affiliated authors, and in 2018/19 this statistic almost tripled, to 71%. Furthermore, we also find a much greater concentration of a few large tech firms, such as Google and Microsoft, with the presence of these "big tech" firms (as identified in [4]) increasing more than fivefold, from 11% to 58% (see Figure 2). The fraction of the annotated papers with corporate ties, by author affiliation or funding, dramatically increased from 43% in
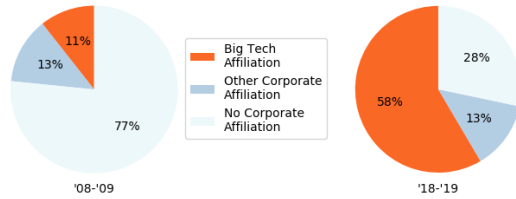
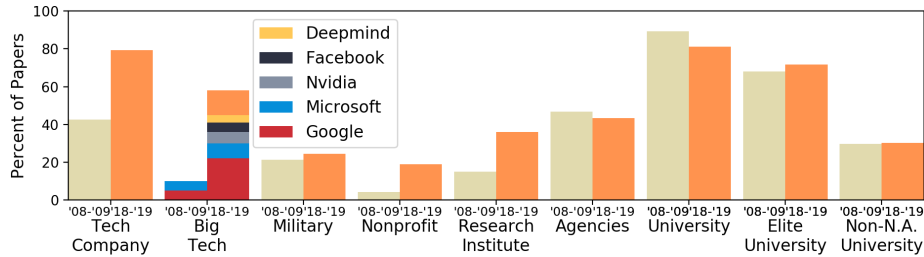Figure 2: Corporate and Big Tech author affiliations.

Figure 3: Corporate affiliations and funding ties. Non-N.A. Universities are those outside the U.S. and Canada.

2008/09 to 79% in 2018/19. In addition, we found paramount domination of elite universities in our analysis as shown in Figure 3. Of the total papers with university affiliations, we found 82% were from elite universities (defined as the top 50 universities by QS World University Rankings, following past work [4]). These findings are consistent with previous work indicating a pronounced corporate presence in ML research. In an automated analysis of peer-reviewed papers from 57 major computer science conferences, Ahmed and Wahed [4] show that the share of papers that have at least one corporate affiliated co-author increased from 10% in 2005 for both ICML and NeurIPS to 30% and 35% respectively in 2019. Our analysis shows that corporate presence is even more pronounced in those papers from ICML and NeurIPS that end up receiving the most citations.

The influence of powerful players in ML research is consistent with field-wide value commitments that centralize power. Others have also argued for causal connections. For example, Abdalla and Abdalla [2] argue that the strategies that big tech uses to sway and influence academic and public discourse, closely resemble that of Big Tobacco. Moreover, examining the prevalent values of big tech, critiques have repeatedly pointed out that objectives such as efficiency, scale, and wealth accumulation [33, 34, 19] drive the industry at large, often at the expense of individuals rights, respect for persons, consideration of negative impacts, beneficence, and justice. The top stated values of ML that we presented in this paper such as performance, generalization, and efficiency may not only enable and facilitate the realization of big tech's objectives, they may also suppress values such as beneficence, justice, and inclusion. A "state-of-the-art" large image dataset, for example, is instrumental for large scale models, further benefiting ML researchers and big tech in possession of huge computing power. In the current climate where values such as accuracy, efficiency, and scale, as currently defined, are a priority, user safety, informed consent, or participation may be perceived as costly and time consuming, evading social needs.

7   Discussion and Related Work

ML research is often perceived as value-neutral, and emphasis is placed on positive applications or potential. This fits into a historical strain of thinking which has tended to frame technology as "neutral", based on the notion that new technologies can be unpredictably applied for both beneficial and harmful purposes [43]. Ironically, this claim of neutrality frequently serves as an insulation from critiques of AI and as a permission to emphasize the benefits of AI [38, 41]. Although it is rare to see anyone explicitly argue in print that ML is neutral, related ideas are part of contemporary conversation, including these canonical claims: long term impacts are too difficult to predict; socio-logical impacts are outside the expertise or purview of ML researchers [20]; critiques of AI are really

misdirected critiques of those deploying AI with bad data ("garbage in, garbage out"), again outside the purview of many AI researchers; and proposals such as broader impact statements represent merely a "bureaucratic constraint" [3]. A recent qualitative analysis of required broader impact statements from NeurIPS 2020 similarly observed that these statements leaned towards positive consequences (often mentioning negative consequences only briefly and in some cases not at all), emphasized uncertainty about how a technology might be used, or simply omit any discussion of societal consequences altogether [31].

Importantly, there is a foundational understanding in Science, Technology, and Society Studies (STSS), Critical Theory, and Philosophy of Science that science and technologies are inherently value-laden, and these values are encoded in technological artifacts, many times in contrast to a field's formal research criteria, espoused consequences, or ethics guidelines [44, 11, 9]. There is a long tradition of exposing and critiquing such values in technology and computer science. Winner [44] introduced several ways technology can encode political values. This work is closely related to Rogaway [37], who notes that cryptography has political and moral dimensions and argues for a cryptography that better addresses societal needs. Weizenbaum [42] argued in 1976 that the computer has from the beginning been a fundamentally conservative force which solidified existing power: in place of fundamental social changes, he argued, the computer renders technical solutions that allow existing power hierarchies to remain intact.

Our paper extends these critiques to the field of ML. It is a part of a rich space of interdisciplinary critiques and alternative lenses used to examine the field. Works such as [30, 10] critique AI, ML, and data using a decolonial lens, noting how these technologies replicate colonial power relationships and values, and propose decolonial values and methods. Others [9, 32, 15] examine technology and data science from an anti-racist and intersectional feminist lens, discussing how our infrastructure has largely been built by and for white men; D'Ignazio and Klein [15] present a set of alternative principles and methodologies for an intersectional feminist data science. Similarly, Kalluri [22] denotes that the core values of ML are closely aligned with the values of the most privileged and outlines a vision where ML models are used to shift power from the most to the least powerful. Dotan and Milli [16] argue that the rise of deep learning is value-laden, promoting the centralization of power among other political values. Many researchers, as well as organizations such as Data for Black Lives, the Algorithmic Justice League, Our Data Bodies, the Radical AI Network, Indigenous AI, Black in AI, and Queer in AI, explicitly work on continuing to uncover particular ways technology in general and ML in particular can encode and amplify racist, sexist, queerphobic, transphobic, and otherwise marginalizing values [12, 36].

We present this paper in part in order to expose the contingency of the present state of the field; it could be otherwise. For individuals, communities, and institutions wading through difficult-to-pin-down values of the field, as well as those striving toward alternative values, it is a useful tool to have a characterization of the way the field is now, for understanding, shaping, dismantling, or transforming what is, and for articulating and bringing about alternative visions.

As with all methods, our chosen approach — coding important sections of highly-cited papers — has limitations. Most notably, this approach requires human expertise and does not automatically scale or generalize to other data, which limits our ability to draw strong conclusions about other conferences or different years. Similarly, this approach is less reproducible than fully automated approaches, and for both our final list of values and specific annotation of individual sentences, different researchers might make somewhat different choices. However, given the overwhelming presence of certain values, the high agreement rate among annotators, and the similarity of observations made by our team, we strongly believe other researchers following a similar approach would reach similar conclusions about what values are most frequently uplifted by the most influential papers in this field. Lastly, we cannot claim to have identified every relevant value in ML. However, by including important ethical values identified by past work, and specifically looking for these, we can confidently assert their relative absence in this set of papers, which captures an important aspect of influential work in ML.

8   Conclusion

We reject the vague conceptualization of the discipline of ML as value-neutral. Instead, we investigate the ways that the discipline of ML is inherently value-laden. Our analysis of highly influential papers in the discipline finds that they not only favor the needs of research communities and large

firms over broader social needs, but also that they take this favoritism for granted. The favoritism manifests in the choice of projects, the lack of consideration of potential negative impacts, and the prioritization and operationalization of values such as performance, generalization, efficiency, and novelty. These values are operationalized in ways that disfavor societal needs, usually without discussion or acknowledgment. Moreover, we uncover an overwhelming and increasing presence of big tech and elite universities in highly cited papers, which is consistent with a system of power-centralizing value-commitments. The upshot is that the discipline of ML is not value-neutral. We find that it is socially and politically loaded, frequently neglecting societal needs and harms, while prioritizing and promoting the concentration of power in the hands of already powerful actors.

Acknowledgements

References

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness in representation: Quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 2019.

[2] Mohamed Abdalla and Moustafa Abdalla. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. *arXiv preprint arXiv:2009.13676*, 2020.

[3] Grace Abuhamad and Claudel Rheault. Like a researcher stating broader impact for the very first time. *arXiv preprint arXiv:2011.13032*, 2020.

[4] Nur Ahmed and Muntasir Wahed. The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.

[5] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of NAACL*, 2018.

[6] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. Technical report, U.S. Department of Homeland Security, Aug 2012.

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*, 2021.

[8] Mariette Bengtsson. How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2:8–14, 2016.

[9] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley, 2019.

[10] Abeba Birhane. Algorithmic colonization of Africa. *SCRIPTed*, 17(2), 2020.

[11] Geoffrey C. Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 2000.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2018.

[13] Sasha Costanza-Chock. Design justice, AI, and escape from the matrix of domination. *Journal of Design and Science*, 2018.

[14] Kate Crawford. The trouble with bias. NeurIPS Keynote, 2017.

[15] Catherine D'Ignazio and Lauren F Klein. *Data Feminism*. MIT Press, 2020.

[16] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. *arXiv preprint arXiv:1912.01172*, 2019.

[17] Luciano Floridi and Josh Cowls. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 2019.

[18] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings CHI*, 2018.

[19] Alex Hanna and Tina M. Park. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850*, 2020.

[20] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of CHI*, 2019.

[21] John P. A. Ioannidis. Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6):645–654, 2012.

[22] Pratyusha Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169–169, 2020.

[23] Laura Krefting. Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy*, 45(3):214–222, 03 1991.

[24] Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, 2018.

[25] Thomas S. Kuhn. Objectivity, value judgment, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pages 320–39. University of Chicago Press, 1977.

[26] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, et al. Indigenous protocol and artificial intelligence position paper. 2020.

[27] Helen E. Longino. Cognitive and non-cognitive values in science: Rethinking the dichotomy. In Lynn Hankinson Nelson and Jack Nelson, editors, *Feminism, Science, and the Philosophy of Science*, pages 39–58. Springer Netherlands, 1996.

[28] Ernan McMullin. Values in science. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982.

[29] Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago press, 1973.

[30] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial rheory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659—684, 2020.

[31] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. Unpacking the expressed consequences of AI research in broader impact statements. *CoRR*, abs/2105.04760, 2021.

[32] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

[33] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

[34] Frank Pasquale. *The black box society*. Harvard University Press, 2015.

[35] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of ICML*, 2020.

[36] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A Pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[37] Phillip Rogaway. The moral character of cryptographic work. Cryptology ePrint Archive, Report 2015/1162, 2015. https://eprint.iacr.org/2015/1162.

[38] Daniela Rus. Rise of the robots: Are you ready? *Financial Times Magazine*, March 2018.

[39] Anthony J. Viera and Joanne M. Garrett. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–3, 2005.

[40] Kiri Wagstaff. Machine learning that matters. In *Proceedings of ICML*, 2012.

[41] Joseph Weizenbaum. On the impact of the computer on society. *Science*, 176(4035):609–614, 1972.

[42] Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. WH Freeman & Co, 1976.

[43] Langdon Winner. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. MIT Press, 1977.

[44] Langdon Winner. Do artifacts have politics? *Daedalus*, 109(1):121–136, 1980.

A Appendix

A.1 Relevance to NeurIPS

The ML community has recently introduced a number of efforts to understand the societal impacts of the field, such as NeurIPS's requirement for the inclusion of broader impacts statements for all submitted papers in 2020, the Resistance AI Workshop at NeurIPS 2020 which investigated how AI concentrates power, and the Navigating the Broader Impacts of AI Research Workshop at NeurIPS 2020 which sought to understand the impacts of ML research as a whole on society. Understanding the social impact of a paper, let alone the discipline, is difficult. Merely looking at various benchmarks, abstract assessments, or broader impact statements, for example, is insufficient to get at the many concrete impacts encoded in the research itself. This paper attempts to begin to bridge this gap by seeking to understand the value commitments in papers published at NeurIPS and a closely related conference, ICML. As such, this paper is highly relevant and a timely contribution to the NeurIPS audience. As research into technical ML topics – reinforcement learning, deep learning, optimization, etc. – are core to NeurIPS, it is apparent that it is increasingly important and of interest to understand where these research areas stand with regard to societal impact, both in a positive and negative manner, as well as the benefits they bring and to whom.

A.2 Additional Methodological Details

A.2.1 Data Sources

To determine the most-cited papers from each conference, we rely on the publicly-available Semantic Scholar database [5], which includes bibliographic information for scientific papers, including citation counts.[7] Using this data, we chose the most cited papers from each of 2008, 2009, 2018, 2019 published at NeurIPS and ICML.

Like all bibliographic databases, Semantic Scholar is imperfect. Upon manual review, we wish to document that our selection includes one paper that was actually published in 2010, and one that was retracted from NeurIPS prior to publication (see §A.9 for details). In addition, the citations counts used to determine the most cited papers reflect a static moment in time, and may differ from other sources.

Because our artifacts of study are papers that have been previously published at NeurIPS or ICML, we surmise that the authors normatively expect and consent to their papers and themselves as authors being referenced and analyzed in future papers, e.g. this paper. Accordingly, we chose not to seek explicit permission from the original authors to reference, annotate, and analyze their papers. The annotations we generated do not introduce any new personally identifying information or offensive content. The sentences from the original published papers are necessarily part of our annotations; to the extent that these papers have these issues, these sentences may contain personally identifying information or offensive content. Given the original authors contributed their work to the same venues as our own work, we believe that the potential to cause new harm from this inclusion is minimal.

A.2.2 Defining elite university

To determine the list of elite universities, we follow Ahmed and Wahed [4], and rely on the QS World University Rankings for the discipline of computer science. For 2018/19, we take the top 50 schools from the CS rankings for 2018. For 2008/09, we take the top 50 schools from the CS rankings for 2011, as the closest year for which data is available.

A.2.3 Defining big tech

We used Abdalla and Abdalla's [2] criterion to what is considered "big tech", which is comprised of: Alibaba, Amazon, Apple, Element AI, Facebook, Google, Huawei, IBM, Intel, Microsoft, Nvidia, Open AI, Samsung, and Uber. Furthermore, we added DeepMind to this list, which Google acquired in 2014. We considered all other companies as "non-big tech".

---

[7] http://s2-public-api.prod.s2.allenai.org/corpus/

A.3 Annotations

We include the annotations of all papers in the supplementary zip file. To present a birds-eye view of the value annotations, we present randomly selected examples of annotated sentences in section §A.8. In addition, here we present the frequency of occurrence for all values prior to clustering (see §A.4) in Figure A.1.
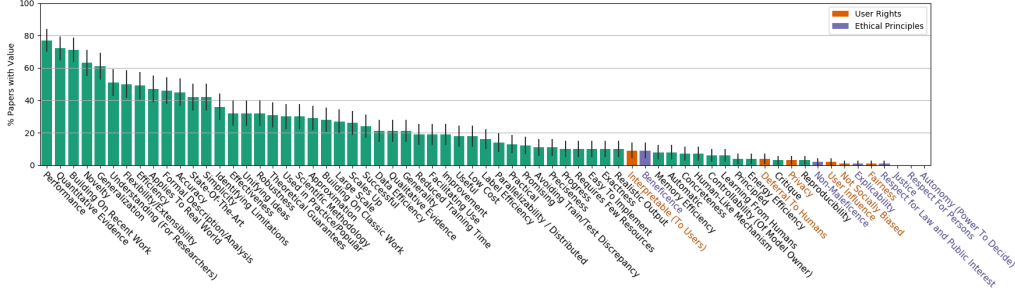

Figure A.1: Value occurrences across papers.

A.4 Value Hierarchy

During ongoing discussions held throughout the annotation process, several values emerged as particularly salient to all annotators, including *Performance*, *Efficiency*, *Generalization*, *Novelty*, and *Building on Past Work*. In some cases, these had strong overlap with related values (e.g., *Performance* is closely related to *Accuracy*). In other cases, we had annotated for several fine-grained values that we felt could be combined (e.g., *Data Efficiency* and *Label Efficiency* are types of *Efficiency*). As such, we decided to group together certain related sets of values into clusters for presentation in the main paper. The values that were combined into each cluster are listed in Table 6 below.

Table 6: Sets of values combined for discussion in the main paper

Cluster	Values
Performance values	Performance, Accuracy, State-of-the-art
Building on past work	Building on classic work, Building on recent work
Generalization values	Generalization, Avoiding train/test discrepancy, Flexibility/extensibility
Efficiency values	Efficiency, Data efficiency, Energy efficiency, Fast, Label efficiency, Low cost, Memory efficiency, Reduced training time

A.5 Experiments with Using Text Classification to Identify Values

Although it was not our primary purpose in annotating highly-cited papers, we include here a brief report on using the annotations we generated as potential training data for classifiers that could in principle be used to estimate the prevalence of these values in a larger set of ML papers. This is something that we should approach with great caution for several reasons: i) we only have a relatively small training set of annotated examples with respect to machine learning best practices; ii) these annotations are taken from a non-random set of papers, and any models trained on these data may not generalize to all papers; iii) an automated approach will fail to detect additional, previously unobserved, emergent values; and iv) based on our experiences annotating these papers, we expect that many would be expressed subtly and in varied ways that would be difficult to detect automatically, at least without considerably more training data.

To present a baseline for testing the potential of this approach, while avoiding any biases that might be introduced by pretrained language models, we make use of simple regularized logistic regression classifiers operating on unigram features. We trained models separately for each value (for all values that had at least 20 relevant sentences, using all relevant sentences for the higher-order grouped values), treating each sentence as an instance with a binary label (present or not), tokenizing each sentence using *spaCy* and converting each to a binary feature representation indicating the presence or absence of each word in the vocabulary (all words occurring at least twice in the corpus). These
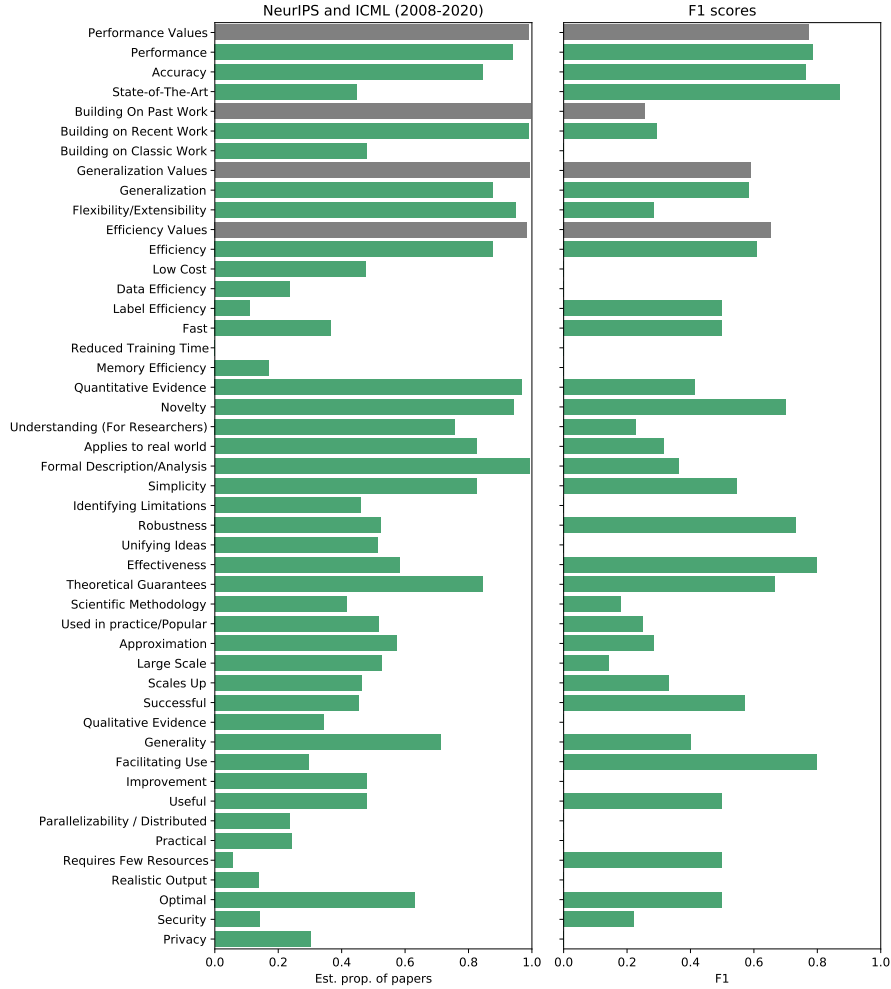
Figure A.2: Proportion of papers in from 2008–2020 (combining NeurIPS and ICML) predicted to have at least one sentence expressing each value (left), and estimated performance (F1) of the corresponding classifiers (right). Note that the overall performance of most classifiers is generally poor, indicating that the estimates on the left should be treated as unreliable in most cases. Grey bars represent the clustered values. Classifiers were not trained for values with less than 20 representative sentences.

choices were not tuned. We randomly selected 300 sentences to use as a held out test set (using the same test set for each value), and trained a model using the remaining data, using 5-fold cross validation to tune the regularization strength.

F1 scores on the test set for the various models are shown in Figure A.2 (right), and can generally be seen to be unimpressive. The F1 score for most values is on the order of 0.5 or less, and some values, even relatively common ones such as *Unifying Ideas*, ended up with an F1 score of 0. The most highly-weighted features for most classifiers were quite reasonable, but this is evidently a relatively difficult task, at least given this amount of data. The exceptions to this poor performance included the Performance-related values (*Performance*, *Accuracy*, and *State-of-the-art*), as well as *Effectiveness*, and *Facilitating Use*, all of which had F1 scores greater than 0.75, and most of which were typically represented by a relatively small set of terms (e.g., "accurate", "accuracy", "accurately", "inaccurate", "accuracies", "errors", etc. for *Accuracy*).

Although the poor performance of these classifiers means we should interpret any use of them with caution, we explore applying them to a broader set of papers for the sake of completeness. To do so, we download pdfs of all papers published at NeurIPS and ICML for the years 2008 through 2020,

15

convert these to text using *pdftotext*, and extract sentences from this text, excluding references, as well as very short sentences (less than 6 tokens) or lines without alphabetic characters. Note that due to the difficulty of automatically parsing papers into sections, these textual representations are not limited to the abstract, introduction, discussion, and conclusion, in contrast to our annotations, thus we would expect most values to occur more frequently, especially those that are likely to occur in sections about experiments and results.

We then apply the classifiers trained above to each sentence in each paper. For each value, we then compute the proportion of papers (combining NeurIPS and ICML for this entire time period) that had at least one sentence predicted to exhibit that value. The overall proportions are shown in Figure A.2 (left). As can be seen, the relative prevalence of values is broadly similar to our anntoated sample, though many are predicted to occur with greater frequency, as expected. However, to reiterate, we should be highly skeptical of these findings, given the poor performance of the classifiers, and we can view this analysis to be useful more so to deepen our understanding of appropriate methodology.

Finally, as an additional exploration, we focus on the Performance-related values (*Performance*, *Accuracy*, and *State-of-the-art*), which represent the overall most prevalent cluster in our annotations and were relatively easy to identify using classification due to their typically simple and explicit expression. We plot the estimated frequency over time for both conferences. For NeurIPS, which has better archival practices, we extend the analysis back to 1987. We should again treat these results with caution, given all the caveats above, as well as the fact that we are now applying these classifiers outside the temporal range from which the annotations were collected. Nevertheless, the results, shown in Figure A.3, suggest that these values have gradually become more common in NeurIPS over time, reinforcing the contingent nature of the dominance of the current set of values. Further investigation is required, however, in order to verify this finding.
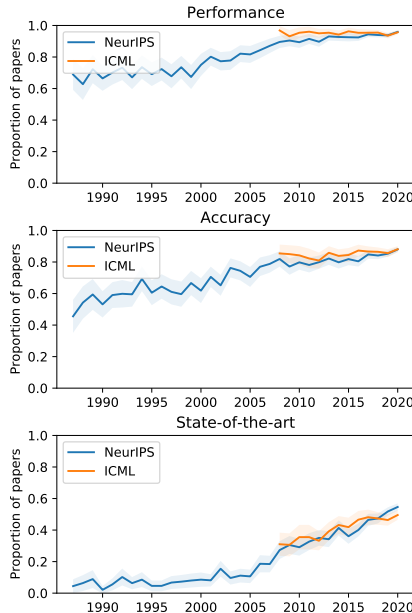

Figure A.3: Proportion of papers per year (of those published in ICML and NeurIPS) that are classified as having at least one sentence expressing *Performance*, *Accuracy*, or *State-of-the-art*, (top, middle, and bottom), based on simple text classifiers trained on our annotations. Bands show ± 2 standard deviations, reflecting the changing overall number of papers per year.

A.6 Code and Reproducibility

We include, in the supplementary zip file, the code used for all data analysis; in particular, we include the code used to run the text classification experiments and generate all figures in the paper. The text classification experiments were run on a 2019 Macbook Air.

A.7 Potential Negative Societal Impacts

Because this paper primarily relies on socially conscientious manual annotation of papers already published at NeurIPS and ICML, we believe that the potential negative societal impacts of carrying out these annotations and sharing them are minimal. However, we still briefly comment on this here.

First, in terms of the specific concerns highlighted in the NeurIPS call for papers, we believe our annotation work poses no risk to living beings, human rights concerns, threats to livelihoods, etc. Similarly, all annotators are co-authors on this paper, thus there was no risk to participants, beyond what we chose to take on for ourselves. We have further discussed these aspects of the data in §A.2.1. Our computational experiments are done locally and have resource usage on par with everyday computer usage.

One area of potential concern to readers, particularly researchers, may be regarding adopting a punitive stance toward individuals, unintentionally casting certain authors in a negative light, or unintentionally contributing to harmful tensions within the ML community. We wish to directly express that throughout this paper we have sought to avoid punitive language toward individuals and adopt language emphasizing systematic patterns. In order to further minimize the former, we have chosen to include randomly selected examples omitting author attributions from quoted sources in the main paper. To complement this and meet the need for completeness, transparency, and reproducibility of our work, we include a full list of cited papers below, so as to acknowledge this work without drawing unnecessary attention to any one particular source.

Although our intention is to broaden and deepen the conversation, we acknowledge that some authors may perceive our work as being not representative of the type of work they would like to see at NeurIPS, and possibly detrimental to the conference. However, because of the prominence and influence of machine learning today, it is especially important to have these conversations at the premier venues, and we hope that our paper will be the basis for useful conversations and future work. As expressed in the main paper, these perceptions and norms may be precisely those that are more contingent than the community realizes; these norms may be shaped, dismantled, transformed, or reenvisioned for the better.

A.8 Random Examples

The list below contains 100 random examples drawn from the annotated data, along with the set of annotated values for each. These sentences were annotated for values within the context of the paper.

- *The problem of minimizing the rank of a matrix variable subject to certain constraints arises in many fields including machine learning, automatic control, and image compression.* ***Used in practice/Popular***

- *Locality-sensitive hashing [6] is an effective technique that performs approximate nearest neighbor searches in time that is sub-linear in the size of the database* ***Approximation***, ***Building on recent work***, ***Effectiveness***, ***Fast***

- *In the finitecase, analysis of optimization and generalization of fully-trained nets is of course an open problem* ***Formal description/analysis***, ***Generalization***

- *So to achieve adversarial robustness, a classifier must generalize in a stronger sense.* ***Generalization***, ***Robustness***

- *Robustness to label corruption is similarly improved by wide margins, such that pre-training alone outperforms certain task-specific methods, sometimes even after combining these methods with pre-training.* ***Performance***, ***Robustness***, ***Understanding (for researchers)***

- *RBMs have been particularly successful in classification problems either as feature extractors for text and image data (Gehler et al., 2006) or as a good initial training phase for deep neural network classifiers (Hinton, 2007).* ***Building on recent work***, ***Flexibility/Extensibility***, ***Successful***

- *Our theoretical analysis naturally leads to a new formulation of adversarial defense which has several appealing properties; in particular, it inherits the benefits of scalability to large datasets exhibited by Tiny ImageNet, and the algorithm achieves state-of-the-art performance on a range of benchmarks while providing theoretical guarantees.* ***Robustness***, ***Scales up***, ***Security***, ***Theoretical guarantees***

- *The current paper focuses on the training loss, but does not address the test loss. **Generalization***

- *This result is significant since stochastic methods are highly preferred for their efficiency over deterministic gradient methods in machine learning applications. **Efficiency***

- *Ranking, which is to sort objects based on certain fac- tors, is the central problem of applications such as in- formation retrieval (IR) and information filtering. **Applies to real world**, **Used in practice/Popular***

- *This subspace is important, because, when projected onto this subspace, the means of the distributions are well-separated, yet the typical distance between points from the same distribution is smaller than in the original space. **Important***

- *Overall, the existence of such adversarial examples raises concerns about the robustness of current classifiers. **Identifying limitations**, **Robustness***

- *We have shown that biased compressors if naively used can lead to bad generalization, and even non-convergence. **Formal description/analysis**, **Generalization***

- *Bartlett and Mendelson [2002] provide a generalization bound for Lipschitz loss functions. **Building on classic work**, **Generalization***

- *The principal advantage of taking this "lateral" approach arises from the fact that compact representation in trajectory space is better motivated physically than compact representation in shape space **Realistic world model***

- *In this paper, we show that gradient descent on deep overparametrized networks can obtain zero training loss **Formal description/analysis**, **Theoretical guarantees***

- *Moreover, web queries often have different meanings for different users (a canonical example is the query jaguar) suggesting that a ranking with diverse documents may be preferable. **Diverse output**, **User influence***

- *We include human performance estimates for all benchmark tasks, which verify that substantial headroom exists between a strong BERT-based baseline and human performance. **Learning from humans**, **Performance***

- *In this paper we propose a simple and fast algorithm SVP(Singular Value Projec-tion) for rank minimization under affine constraints (ARMP) and show that SVP recovers the minimum rank solution for affine constraints that satisfy a restricted isometry property(RIP). **Fast**, **Novelty**, **Simplicity***

- *We use standard formalization of multiclass classification, where data consists of sample x and its label y (an integer from 1 to k). **Building on classic work***

- *A number of recent work has shown that the low rank solution can be recovered exactly via minimizing the trace norm under certain conditions (Recht et al., 2008a; Recht et al., 2008b; Candès and Recht, 2008). **Building on recent work***

- *This difficulty has necessitated the use of a heuristic inference procedure, that nonetheless was accurate enough for successful learning. **Accuracy**, **Successful***

- *We illustrate such potential by measuring search space properties relevant to architecture search. **Quantitative evidence (e.g. experiments)***

- *Deep architectures consist of feature detector units arranged in layers. Lower layers detect simple features and feed into higher layers, which in turn detect more complex features. **Simplicity***

- *This makes the updates hard to massively parallelize at a coarse, dataparallel level (e.g., by computing the updates in parallel and summing them together centrally) without losing the critical stochastic nature of the updates. **Large scale**, **Parallelizability / distributed***

- *This suggests future work on model robustness should evaluate proposed methods with pretraining in order to correctly gauge their utility, and some work could specialize pretraining for these downstream tasks. **Robustness***

- *Adversarial training remains among the most trusted defenses, but it is nearly intractable on largescale problems. **Scales up**, **Security***

- *For complex robots such as humanoids or light-weight arms, it is often hard to model the system sufficiently well and, thus, modern regression methods offer a viable alternative [7,8].* **Realistic world model**

- *In contrast to prior work that operates in this goal-setting model, we use states as goals directly, which allows for simple and fast training of the lower layer.* **Reduced training time**, **Simplicity**

- *Meanwhile, using less resources tends to produce less compelling results (Negrinho and Gordon, 2017; Baker et al., 2017a).* **Requires few resources**

- *This finding represents an exciting opportunity for defense against neural fake news: the best models for generating neural disinformation are also the best models at detecting it.* **Applies to real world**

- *Our strong empirical results suggest that randomized smoothing is a promising direction for future research into adversarially robust classification.* **Quantitative evidence (e.g. experiments)**, **Robustness**, **Security**

- *We then turn our attention to identifying the roots of BatchNorm's success.* **Successful**, **Understanding (for researchers)**

- *We also report the results of large-scale experiments comparing these three methods which demonstrate the benefits of the mixture weight method: this method consumes less resources, while achieving a performance comparable to that of standard approaches.* **Large scale**, **Performance**, **Requires few resources**

- *This paper does not cover the the generalization of over-parameterized neural networks to the test data.* **Avoiding train/test discrepancy**, **Generalization**

- *While there has been success with robust classifiers on simple datasets [31, 36, 44, 48], more complicated datasets still exhibit a large gap between "'standard" and robust accuracy [3, 11].* **Applies to real world**, **Robustness**, **Successful**

- *In this paper, we have shown theoretically how independence between examples can make the actual effect much smaller.* **Novelty**, **Theoretical guarantees**

- *We provide empirical evidence that several recently-used methods for estimating the probability of held-out documents are inaccurate and can change the results of model comparison.* **Accuracy**, **Building on recent work**, **Quantitative evidence (e.g. experiments)**

- *This agreement is robust across differentarchitectures, optimization methods, and loss functions* **Robustness**

- *Unfortunately, due to the slow-changing policy in an actor-critic setting, the current and target value estimates remain too similar to avoid maximization bias.* **Accuracy**

- *As a future work, we are pursuing a better understanding of probabilistic distributions on the Grassmann manifold.* **Understanding (for researchers)**

- *We also view these results as an opportunity to encourage the community to pursue a more systematic investigation of the algorithmic toolkit of deep learning and the underpinnings of its effectiveness.* **Effectiveness**, **Understanding (for researchers)**

- *This challenge is further exacerbated in continuous state and action spaces, where a separate actor network is often used to perform the maximization in Q-learning.* **Performance**

- *The vulnerability of neural networks to adversarial perturbations has recently been a source of much discussion and is still poorly understood.* **Robustness**, **Understanding (for researchers)**

- *Most of the evaluation methods described in this paper extend readily to more complicated topic models— including non-parametric versions based on hierarchical Dirichlet processes (Teh et al., 2006)—since they only require a MCMC algorithm for sampling the latent topic assignments z for each document and a way of evaluating probability $P(w \mid z, \Phi, \alpha m)$.* **Flexibility/Extensibility**, **Understanding (for researchers)**

- *In a formulation closely related to the dual problem, we have: $\hat{w} = argmin$ $w:F(w) \leq c$ $\frac{1}{n}$ $\sum_{i=1}^{n} \ell(hw, xii, yi)$ (2) where, instead of regularizing, a hard restriction over the parameter space is imposed (by the constant c).* **Formal description/analysis**

- *Second, we evaluate a surrogate loss function from four aspects: (a) consistency, (b) soundness, (c) mathemat- ical properties of continuity, differentiability, and con- vexity, and (d) computational efficiency in learning.* **Efficiency**

- *This leads to two natural questions that we try to answer in this paper: (1) Is it feasible to perform optimization in this very large feature space with cost which is polynomial in the size of the input space?* **Performance**

- *Despite its pervasiveness, the exact reasons for BatchNorm's effectiveness are still poorly understood.* **Understanding (for researchers)**

- *We have presented confidenceweighted linear classifiers, a new learning method designed for NLP problems based on the notion of parameter confidence.* **Novelty**

- *In addition, the experiments reported here suggest that (like other strategies recently proposed to train deep deterministic or stochastic neural networks) the curriculum strategies appear on the surface to operate like a regularizer, i.e., their beneficial effect is most pronounced on the test set.* **Beneficence**, **Quantitative evidence (e.g. experiments)**

- *These give further inside into hash-spaces and explain previously made empirical observations.* **Understanding (for researchers)**

- *This means that current algorithms reach their limit at problems of size 1TB whenever the algorithm is I/O bound (this amounts to a training time of 3 hours), or even smaller problems whenever the model parametrization makes the algorithm CPU bound.* **Memory efficiency**, **Reduced training time**

- *Much of the results presented were based on the assumption that the target distribution is some mixture of the source distributions.* **Valid assumptions**

- *Empirical investiga- tion revealed that this agrees well with actual training dynamics and predictive distributions across fully-connected, convolutional, and even wide residual network architectures, as well as with different optimizers (gradient descent, momentum, mini-batching) and loss functions (MSE, cross-entropy).* **Generalization**, **Quantitative evidence (e.g. experiments)**, **Understanding (for researchers)**

- *We design a new spectral norm that encodes this a priori assumption, without the prior knowledge of the partition of tasks into groups, resulting in a new convex optimization formulation for multi-task learning.* **Novelty**

- *Recent progress in natural language generation has raised dual-use concerns.* **Progress**

- *These kernel functions can be used in shallow architectures, such as support vector machines (SVMs), or in deep kernel-based architectures that we call multilayer kernel machines (MKMs).* **Flexibility/Extensibility**

- *Using MCMC instead of variational methods for approximate inference in Bayesian matrix factorization models leads to much larger improvements over the MAP trained models, which suggests that the assumptions made by the variational methods about the structure of the posterior are not entirely reasonable.* **Understanding (for researchers)**

- *In particular, the deep belief network (DBN) (Hinton et al., 2006) is a multilayer generative model where each layer encodes statistical dependencies among the units in the layer below it; it is trained to (approximately) maximize the likelihood of its training data.* **Approximation**, **Data efficiency**

- *Furthermore, the learning accuracy and performance of our LGP approach will be compared with other important standard methods in Section 4, e.g., LWPR [8], standard GPR [1], sparse online Gaussian process regression (OGP) [5] and ν-support vector regression (ν-SVR) [11], respectively* **Accuracy**, **Performance**, **Quantitative evidence (e.g. experiments)**

- *• propose a simple method based on weighted minibatches to stochastically train with arbitrary weights on the terms of our decomposition without any additional hyperparameters.* **Efficiency**, **Simplicity**

- *For example, Ng (2004) examined the task of PAC learning a sparse predictor and analyzed cases in which an $\ell 1$ constraint results in better solutions than an $\ell 2$ constraint.* **Building on recent work**

20

- *Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) are an efficient variant of Convolutional Neural Networks (CNNs) on graphs. GCNs stack layers of learned first-order spectral filters followed by a nonlinear activation function to learn graph representations.* **Efficiency**

- *This is a linear convergence rate.* **Building on recent work**, **Efficiency**, **Quantitative evidence (e.g. experiments)**, **Theoretical guarantees**

- *However, as we observe more interactions, this could emerge as a clear feature.* **Building on recent work**, **Data efficiency**

- *Here we propose the first method that supports arbitrary low accuracy and even biased compression operators, such as in (Alistarh et al., 2018; Lin et al., 2018; Stich et al., 2018).* **Accuracy**, **Novelty**

- *Much recent work has been done on understanding under what conditions we can learn a mixture model.* **Understanding (for researchers)**

- *For this reason, we present an extension of the standard greedy OMP algorithm that can be applied to general struc- tured sparsity problems, and more importantly, meaningful sparse recovery bounds can be obtained for this algorithm.* **Building on recent work**

- *In this paper we show that this assumption is indeed nec-essary: by considering a simple yet prototypical exampleof GAN training we analytically show that (unregularized) GAN training is not always locally convergent* **Formal description/analysis**

- *Overestimation bias is a property of Q-learning in which the maximization of a noisy value estimate induces a consistent overestimation* **Accuracy**

- *This drawback prevents GPR from applications which need large amounts of training data and require fast computation, e.g., online learning of inverse dynamics model for model-based robot control* **Fast**, **Large scale**

- *This is problematic since we find there are techniques which do not comport well with pre-training; thus some evaluations of robustness are less representative of real-world performance than previously thought.* **Applies to real world**, **Performance**, **Robustness**

- *Approximation of this prior structure through simple, efficient hy-perparameter optimization steps is sufficient to achieve these performance gains* **Approximation**, **Efficiency**, **Performance**, **Simplicity**

- *The second mysterious phenomenon in training deep neural networks is "deeper networks are harder to train."* **Performance**

- *However, the definition of our metric is sufficiently general that it could easily be used to test, for example, invariance of auditory features to rate of speech, or invariance of textual features to author identity.* **Generalization**

- *In Sec. 6 we test the proposed algorithm for face recognition and object categorization tasks.* **Applies to real world**, **Quantitative evidence (e.g. experiments)**

- *It is possible to train classification RBMs directly for classification performance; the gradient is fairly simple and certainly tractable.* **Performance**

- *Figure 1 contrasts these two approaches. Defining and evaluating models using ODE solvers has several benefits:* **Beneficence**

- *They claim to achieve 12% robustness against non-targeted attacks that are within an '2 radius of 3 (for images with pixels in [0, 1]).* **Generalization**, **Robustness**

- *Two commonly used penalties are the 1- norm and the square of the 2-norm of w.* **Used in practice/Popular**

- *What should platforms do? Video-sharing platforms like YouTube use deep neural networks to scan videos while they are uploaded, to filter out content like pornography (Hosseini et al., 2017).* **Applies to real world**

- *We mention various properties of this penalty, and provide conditions for the consistency of support estimation in the regression setting. Finally, we report promising results on both simulated and real data* **Applies to real world**

- *There could be a separate feature for "high school student," "male," "athlete," and "musician" and the presence or absence of each of these features is what defines each person and determines their relationships.* **Building on recent work**

- *So, the over-parameterized convergence theory of DNN is much simpler than that of RNN.* **Simplicity**, **Understanding (for researchers)**

- *Other threat models are possible: for instance, an adversary might generate comments or have entire dialogue agents, they might start with a human-written news article and modify a few sentences, and they might fabricate images or video.* **Learning from humans**

- *More generally, we hope that future work will be able to avoid relying on obfuscated gradients (and other methods that only prevent gradient descent-based attacks) for perceived robustness, and use our evaluation approach to detect when this occurs.* **Generality**, **Robustness**

- *For example, the learned linear combination does not consistently outperform either the uniform combination of base kernels or simply the best single base kernel (see, for example, UCI dataset experiments in [9, 12], see also NIPS 2008 workshop).* **Performance**

- *Our main contributions are: • We analyze GP-UCB, an intuitive algorithm for GP optimization, when the function is either sampled from a known GP, or has low RKHS norm.* **Optimal**

- *For the standard linear setting, Dani et al. (2008) provide a near-complete characterization explicitly dependent on the dimensionality. In the GP setting, the challenge is to characterize complexity in a different manner, through properties of the kernel function.* **Building on classic work**

- *This allows us to map each architecture A to its approximate hyperparameteroptimized accuracy* **Accuracy**

- *Unfortunately, they could only apply their method to linear networks.* **Flexibility/Extensibility**

- *The strength of the adversary then allows for a trade-off between the enforced prior, and the data-dependent features.* **Understanding (for researchers)**

- *We observe that the computational bottleneck of NAS is the training of each child model to convergence, only to measure its accuracy whilst throwing away all the trained weights.* **Accuracy**

- *We show that the number of subprob-lems need only be logarithmic in the total number of possible labels, making thisapproach radically more efficient than others.* **Efficiency**

- *We establish a new notion of quadratic approximation of the neural network, and connect it to the SGD theory of escaping saddle points.* **Novelty**, **Unifying ideas or integrating components**

- *In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors.* **Accuracy**, **Robustness**, **Theoretical guarantees**

- *A limit on the number of queries can be a result of limits on other resources, such as a time limit if inference time is a bottleneck or a monetary limit if the attacker incurs a cost for each query.* **Applies to real world**, **Low cost**, **Requires few resources**

- *Preliminary experiments demonstrate that it is significantly faster than batch alternatives on large datasets that may contain millions of training examples, yet it does not require learning rate tuning like regular stochastic gradient descent methods.* **Quantitative evidence (e.g. experiments)**, **Reduced training time**

- *SuperGLUE is available at super.gluebenchmark.com.* **Facilitating use (e.g. sharing code)**

A.9   Full List of Cited Papers

The full list of annotated papers is given below, along with the annotated scores (in square brackets) for *Discussion of Negative Potential* [left] (0 = Doesn't mention negative potential; 1 = Mentions

but does not discuss negative potential; 2 = Discusses negative potential) and *Justification* [right] (0 = Doesn't rigorously justify how it achieves technical goal; 1 = Justifies how it achieves technical goal but no mention of societal need; 2 = States but does not justify how it connects to a societal need; 3 = States and somewhat justifies how it connects to a societal need; 4 = States and rigorously justifies how it connects to a a societal need). Note that due to minor errors in the data sources used, the distribution of papers over venues and years is not perfectly balanced. For the same reason, the list also contains one paper from 2010 (rather than 2009), as well as one paper that was retracted before publication at NeurIPS (marked with a *).

- Mingxing Tan, Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of ICML*, 2019. [0/1]

- Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *Proceedings of ICML*, 2019. [0/1]

- Jeremy Cohen, Elan Rosenfeld, Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of ICML*, 2019. [0/1]

- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, Michael Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of ICML*, 2019. [0/2]

- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of ICML*, 2019. [0/1]

- Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, Kilian Weinberger. Simplifying Graph Convolutional Networks. In *Proceedings of ICML*, 2019. [0/1]

- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of ICML*, 2019. [0/2]

- Justin Gilmer, Nicolas Ford, Nicholas Carlini, Ekin Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In *Proceedings of ICML*, 2019. [0/1]

- Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, Frank Hutter. NAS-Bench-101: Towards Reproducible Neural Architecture Search. In *Proceedings of ICML*, 2019. [0/2]

- Dan Hendrycks, Kimin Lee, Mantas Mazeika. Using Pre-Training Can Improve Model Robustness and Uncertainty. In *Proceedings of ICML*, 2019. [0/1]

- Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, Martin Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *Proceedings of ICML*, 2019. [0/1]

- Anastasia Koloskova, Sebastian Stich, Martin Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *Proceedings of ICML*, 2019. [0/2]

- Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena. Self-Attention Generative Adversarial Networks. In *Proceedings of ICML*, 2019. [0/1]

- Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *Proceedings of ICML*, 2019. [0/1]

- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *Proceedings of ICML*, 2019. [0/1]

- Anish Athalye, Nicholas Carlini, David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proceedings of ICML*, 2018. [0/2]

- Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, Jeff Dean. Efficient Neural Architecture Search via Parameters Sharing. In *Proceedings of ICML*, 2018. [0/1]

- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of ICML*, 2018. [0/2]

- Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, Koray Kavukcuoglu. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of ICML*, 2018. [0/1]

- Scott Fujimoto, Herke Hoof, David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of ICML*, 2018. [0/1]

- Hyunjik Kim, Andriy Mnih. Disentangling by Factorising. In *Proceedings of ICML*, 2018. [0/0]

- Lars Mescheder, Andreas Geiger, Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *Proceedings of ICML*, 2018. [0/1]

- Sanjeev Arora, Rong Ge, Behnam Neyshabur, Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of ICML*, 2018. [0/3]

- Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of ICML*, 2018. [0/2]

- Niranjan Srinivas, Andreas Krause, Sham Kakade, Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of ICML*, 2010. [0/1]

- Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of ICML*, 2009. [0/1]

- Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of ICML*, 2009. [0/1]

- Yoshua Bengio, Jerome Louradour, Ronan Collobert and Jason Weston. Curriculum learning. In *Proceedings of ICML*, 2009. [0/1]

- Laurent Jacob, Guillaume Obozinski and Jean-Philippe Vert. Group Lasso with Overlaps and Graph Lasso. In *Proceedings of ICML*, 2009. [0/3]

- Chun-Nam Yu and Thorsten Joachims. Learning structural SVMs with latent variables. In *Proceedings of ICML*, 2009. [0/2]

- Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford and Alex Smola. Feature hashing for large scale multitask learning. In *Proceedings of ICML*, 2009. [0/2]

- Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of ICML*, 2009. [0/1]

- Kamalika Chaudhuri, Sham Kakade, Karen Livescu and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of ICML*, 2009. [0/2]

- Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of ICML*, 2009. [0/3]

- Junzhou Huang, Tong Zhang and Dimitris Metaxas. Learning with structured sparsity. In *Proceedings of ICML*, 2009. [0/1]

- Rajat Raina, Anand Madhavan and Andrew Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of ICML*, 2009. [0/2]

- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*, 2008. [0/2]

- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML*, 2008. [0/1]

- Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of ICML*, 2008. [0/1]

- John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of ICML*, 2008. [0/1]

- Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of ICML*, 2008. [0/1]

- Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of ICML*, 2008. [0/1]

- Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted Boltzmann machines. In *Proceedings of ICML*, 2008. [0/1]

- Jihun Hamm and Daniel Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of ICML*, 2008. [0/1]

- Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank - Theory and Algorithm. In *Proceedings of ICML*, 2008. [0/1]

- Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of ICML*, 2008. [0/1]

- Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *Proceedings of ICML*, 2008. [0/1]

- Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of ICML*, 2008. [0/1]

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of NeurIPS*, 2019. [0/1]

- Alexis CONNEAU, Guillaume Lample. Cross-lingual Language Model Pretraining. In *Proceedings of NeurIPS*, 2019. [0/4]

- Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In *Proceedings of NeurIPS*, 2019. [0/1]

- Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Proceedings of NeurIPS*, 2019. [0/1]

- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin A. Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proceedings of NeurIPS*, 2019. [0/1]

- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of NeurIPS*, 2019. [0/1]

- Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Proceedings of NeurIPS*, 2019. [0/1]

- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Proceedings of NeurIPS*, 2019. [0/1]

- Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, Tom Goldstein. Adversarial Training for Free! In *Proceedings of NeurIPS*, 2019. [0/3]

- Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proceedings of NeurIPS*, 2019. [0/1]

- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proceedings of NeurIPS*, 2019. [1/1]

- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi. Defending Against Neural Fake News. In *Proceedings of NeurIPS*, 2019. [2/4]

- Yuan Cao, Quanquan Gu. Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks. In *Proceedings of NeurIPS*, 2019. [0/1]

- Florian Tramer, Dan Boneh. Adversarial Training and Robustness for Multiple Perturbations. In *Proceedings of NeurIPS*, 2019. [0/2]

- Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, Percy S. Liang. Unlabeled Data Improves Adversarial Robustness. In *Proceedings of NeurIPS*, 2019. [0/1]

- Lars Maaløe, Marco Fraccaro, Valentin Liévin, Ole Winther. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. In *Proceedings of NeurIPS*, 2019. [0/1]

- Zeyuan Allen-Zhu, Yuanzhi Li, Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. In *Proceedings of NeurIPS*, 2019. [0/1]

- Durk P. Kingma, Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Proceedings of NeurIPS*, 2018. [0/2]

- Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, David K. Duvenaud. Neural Ordinary Differential Equations. In *Proceedings of NeurIPS*, 2018. [0/1]

- Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, Jure Leskovec. Hierarchical Graph Representation Learning with Differentiable Pooling. In *Proceedings of NeurIPS*, 2018. [0/1]

- Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, David K. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Proceedings of NeurIPS*, 2018. [0/1]

- Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, Baoquan Chen. PointCNN: Convolution On X-Transformed Points. In *Proceedings of NeurIPS*, 2018. [0/1]

- Arthur Jacot, Franck Gabriel, Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Proceedings of NeurIPS*, 2018. [0/1]

- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro. Video-to-Video Synthesis. In *Proceedings of NeurIPS*, 2018. [0/1]

- Yuanzhi Li, Yingyu Liang. Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In *Proceedings of NeurIPS*, 2018. [0/1]

- Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, Aleksander Madry. Adversarially Robust Generalization Requires More Data. In *Proceedings of NeurIPS*, 2018. [0/2]

- Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Madry. How Does Batch Normalization Help Optimization? In *Proceedings of NeurIPS*, 2018. [0/1]

- Harini Kannan, Alexey Kurakin, Ian Goodfellow. Adversarial Logit Pairing. In *Proceedings of NeurIPS**, 2018. [0/2]

- Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, Sergey Levine. Data-Efficient Hierarchical Reinforcement Learning. In *Proceedings of NeurIPS*, 2018. [0/3]

- Prateek Jain, Raghu Meka, Inderjit Dhillon. Guaranteed Rank Minimization via Singular Value Projection. In *Proceedings of NeurIPS*, 2010. [0/1]

- Hanna Wallach, David Mimno, Andrew McCallum. Rethinking LDA: Why Priors Matter. In *Proceedings of NeurIPS*, 2009. [0/4]

- Geoffrey E. Hinton, Russ R. Salakhutdinov. Replicated Softmax: an Undirected Topic Model. In *Proceedings of NeurIPS*, 2009. [0/1]

- Daniel J. Hsu, Sham M. Kakade, John Langford, Tong Zhang. Multi-Label Prediction via Compressed Sensing. In *Proceedings of NeurIPS*, 2009. [0/1]

- Youngmin Cho, Lawrence Saul. Kernel Methods for Deep Learning. In *Proceedings of NeurIPS*, 2009. [0/1]

- Kurt Miller, Michael Jordan, Thomas Griffiths. Nonparametric Latent Feature Models for Link Prediction. In *Proceedings of NeurIPS*, 2009. [0/3]

- Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, Andrew Ng. Measuring Invariances in Deep Networks. In *Proceedings of NeurIPS*, 2009. [0/1]

- Vinod Nair, Geoffrey E. Hinton. 3D Object Recognition with Deep Belief Nets. In *Proceedings of NeurIPS*, 2009. [0/1]

- Martin Zinkevich, John Langford, Alex Smola. Slow Learners are Fast. In *Proceedings of NeurIPS*, 2009. [0/1]

- Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, Gideon Mann. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. In *Proceedings of NeurIPS*, 2009. [0/1]

- Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh. Learning Non-Linear Combinations of Kernels. In *Proceedings of NeurIPS*, 2009. [0/1]

- Laurent Jacob, Jean-philippe Vert, Francis Bach. Clustered Multi-Task Learning: A Convex Formulation. In *Proceedings of NeurIPS*, 2008. [0/1]

- Kamalika Chaudhuri, Claire Monteleoni. Privacy-preserving logistic regression. In *Proceedings of NeurIPS*, 2008. [0/3]

- Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy, Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Proceedings of NeurIPS*, 2008. [0/3]

- Ilya Sutskever, Geoffrey E. Hinton, Graham W. Taylor. The Recurrent Temporal Restricted Boltzmann Machine. In *Proceedings of NeurIPS*, 2008. [0/1]

- Wenyuan Dai, Yuqiang Chen, Gui-rong Xue, Qiang Yang, Yong Yu. Translated Learning: Transfer Learning across Different Feature Spaces. In *Proceedings of NeurIPS*, 2008. [0/3]

- Yishay Mansour, Mehryar Mohri, Afshin Rostamizadeh. Domain Adaptation with Multiple Sources. In *Proceedings of NeurIPS*, 2008. [0/1]

- Sham M. Kakade, Karthik Sridharan, Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *Proceedings of NeurIPS*, 2008. [0/1]

- Francis Bach. Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning. In *Proceedings of NeurIPS*, 2008. [0/1]

- Ijaz Akhter, Yaser Sheikh, Sohaib Khan, Takeo Kanade. Nonrigid Structure from Motion in Trajectory Space. In *Proceedings of NeurIPS*, 2008. [0/1]

- Prateek Jain, Brian Kulis, Inderjit Dhillon, Kristen Grauman. Online Metric Learning and Fast Similarity Search. In *Proceedings of NeurIPS*, 2008. [0/1]

- Duy Nguyen-tuong, Jan Peters, Matthias Seeger. Local Gaussian Process Regression for Real Time Online Model Learning. In *Proceedings of NeurIPS*, 2008. [0/1]

- Lester Mackey. Deflation Methods for Sparse PCA. In *Proceedings of NeurIPS*, 2008. [0/1]

A.10 Checklist

Many conferences, including NeurIPS, have begun requiring reproducability checklists. We include a modified version of the NeurIPS checklist here to provide a quick summary of common reproducability questions, and encourage this practice in future papers.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Discussion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Included in the Appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Available on Github

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Included in Appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Included in Appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Full listing of annotated papers is given in the Appendix.

(b) Did you mention the license of the assets? [Yes] See Footnote 1.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Included in supplementary zipfile.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Discussed in Appendix A.2. Additional Methodological Details.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Discussed in Appendix A.2. Additional Methodological Details.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]