Federal Register Notice 86 FR 46278, https://www.federalregister.gov/documents/2021/08/18/2021-17737/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence, October 1, 2021.

Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource: Responses

The 3Ds of AI: Data, Developer and Democratization

Ben Freed and Howie Choset
School of Computer Science, Carnegie Mellon

Abstract

As our AI tools become more advanced, they are increasingly created and controlled by a select few organizations. By limiting the breadth of institutions, groups, and people who can create, use, and inspect AI tools, the AI oligarchy has negative impacts for individuals, society, as well as the progress of the field. We firmly believe that AI has the potential to bring a tremendous amount of good to the world, but only if developed and used responsibly, which is a conversation in which everyone should have a voice. We identify three key ingredients to advance AI, which we call *the three D's of AI:* data availability, developer accessibility, and democratization of AI tools. It is our view that to unlock the true potential of deep learning, and retain American competitiveness, we must conquer the three Ds of deep learning: data, developer, and democratization.

The Three-D's	Data availability	Developer	Democratization
Benefit	Benefits everyone Removes biases	More talent Uncork talent	Society engagement Equitable ownership
Problem	Cost to create Privacy Biases / silos	Cost to obtain Cost to maintain Lack of tools	Unchecked firms Monopoly / barriers No easy-to-use tools
Solution	Surrogate data Public funds reqs. Data efficient apps	Tool develop Shared resources K12 education Retain international	Easy-to-use tools User-owned data K12 competition

Introduction

In recent years, AI has seen a boom of development, spurred by deep learning. Deep learning has revolutionized the way in which artificial intelligence is applied to domains such as manufacturing, finance, medicine, energy, agriculture, security, retail, just to name a few. Deep learning can be viewed as a type of data science that can model and predict future outcomes from (an enormous amount of) data that is provided to it, during a training process, say of a multi-layer neural network. Deep learning technologies are typically classified as *data driven*, because they primarily focus on extracting patterns from data, rather than relying on the knowledge of AI engineers or domain experts. This

shift in perspective from the *good old-fashioned AI* (GOFAI) techniques of past decades has the benefit that it removes human bias from the system, allowing deep learning algorithms to discover their own data representations and decision-making procedures, yielding higher levels of performance compared to hand-engineered approaches.

The salient feature that delivers deep learning's greatest strength- its ability to process an enormous amount of data - is also a drawback: it requires an overwhelmingly large amount of data to be effective. Such data may not be available to the "common" developer. In fact, lack of access to data is just one barrier of entry to enjoy the benefits of deep learning: an extensive and often time-consuming and expensive education is another requirement and therefore limits deep learning to highly educated and specialized PhDs with years of important education and training. These PhDs are great, but only represent the tip of the iceberg of potential developers that can contribute to and benefit from deep learning - not everyone can get into Carnegie Mellon. Finally, the computational resources to develop and use deep learning tend to be limited to the Google's, Facebook's, and perhaps some Universities of the world and yet many can contribute, if resources or low-overhead deep learning approaches were available.

An increasing portion of AI breakthroughs are being made using resources far outside the budget of the typical academic lab. Freelance and small companies also offer us opportunities that large companies and universities cannot, such as niche applications of AI to problems that might not be appealing to large companies; we do not want to lose them. Finally, improving access to data and AI tools also has the potential to reduce harmful bias in our AI technologies. If datasets are free and open, they can be inspected and are open to criticism by experts in fairness and ethics in AI. America may be at the lead of AI research, development and use, but frankly we are still doing it with one arm tied behind our back.

D1: Data availability

Deep learning-based approaches require a large amount of data to be effective. As can be expected, data availability plays a crucial role in the performance of our ML-based technology. High-quality datasets are necessary for the advances made in research settings to percolate into applied technologies, because availability of quality data plays a large role in determining the efficacy of a learned model in the real world. While it is a strength to process and ``learn from'' a large amount of data, often high quantities of data is required for development and training of AI approaches. Typical datasets used to train deep neural networks used for supervised vision models contain hundreds of thousands to millions of labeled datapoints (e.g., ImageNet, CIFAR10, CIFAR100). For example, state of the art natural language processing (NLP) models, such as GPT3,

have been trained on hundreds of billions of words. State of the art reinforcement learning systems trained for two-player game-play, such as AlphaZero and AlphaStar, were trained on games. For typical academic labs, these large data requirements limit the applicability of deep learning to situations in which large datasets are freely available.

Challenges limiting data availability include 1) high cost or high required investment, 2) privacy concerns, 3) data is siloed and 4) difficulty in obtaining high-quality labels.

D1P1: High cost / investment to acquire data: Gathering large datasets for custom applications is often a high-cost endeavor. For example, one recent publication that used reinforcement learning to learn robotic grasping required 14 expensive robotic arms and over 800,000 grasp attempts to achieve 80-90% grasp success rates using a 2-finger gripper [CITE Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection]. For many academic labs, this would be a prohibitively expensive undertaking. Finally, in order to get the most out of our tax-payer dollars, we must share and document our publicly supported datasets.

D1P1.1: Data efficiency: To ameliorate the high costs associated with large dataset collection, we suggest that more funding be allocated to improving data efficiency in ML, through techniques such as transfer learning, semi-supervised learning, domain adaptation, and data augmentation. These approaches improve data efficiency on some target task by enabling data from another task, or unlabeled data, to contribute to the learning process.

D1P1.2 Leverage public funds: We can also leverage our existing investments by requiring that datasets generated by public funds should be available to the public: it was paid for by the public. Also, we believe that such a practice will inevitably promote the scientific process of validating results. In fact, we already see this practice taking place, as it is in the best interest of the scientist to promote their work and supporting data facilitates such promotion.

D1P1.3: Synthetic Data: Finally, we advocate for increased funding for research on synthetic data creation. Fake data is free but making it meaningful is hard; however, the ability to generate realistic synthetic data could at least allow AI researchers and practitioners to validate approaches and identify weaknesses before making costly investments in dataset collection. An additional potential benefit of synthetic data generation is that it avoids the privacy concerns typically associated with sensitive (e.g. medical) data.

D1P2: Privacy. Privacy permeates all issues that involve datasets. Obviously, privacy at the extreme inhibits the proliferation of datasets in order to protect the owners and subjects of the data. Failure to recognize this importance could be catastrophic. For one thing, we can compromise personal, organizational, and national security. Next, we could potentially lose the trust of those who contribute to the dataset. By no means do we, the authors of this document, claim to be privacy experts, nor understand the bounds of the implications of privacy. Therefore, we suggest that experts in privacy be included in the ideation of data availability and defer to them for specific suggestions. With such experts, we advise that approaches be developed to either develop surrogate data sets and other methods be created to strip private information from datasets and yet retain their salient properties.

D1P3: Data Silos. Improving data availability has the potential to both improve AI both as a research field and a technology. Data is the raw ore from which useful models are smelt, and machine learning is a fundamentally empirical science; hypotheses must be tested on *real-world data* that are truly reflective of the situations in which they are intended to be used. For this reason, in many fields such as computer vision and natural language processing, large, representative, and high-quality datasets are absolutely crucial for fundamental advancement. We strongly advocate an increase in funding for transfer learning and imitation learning, but require disparate problem domains for which this research would be funded.

Additionally, we suggest that measures be taken to encourage inter-agency sharing of data, when possible. It stands to reason that different agencies may have some common denominators in the datasets they collect. It would be meaningful to understand the commonalities, to see what shared problems they're all solving, as well as the differences to see how we can round off each others' limited datasets. Moreover, agency x can stress-test its approaches using agency y's datasets. The problem is that, from the authors' distant perspective, agencies often have a hard time cooperating and sharing at deep levels. We suggest that the White House look at examples of where inter-agency cooperation has been successful, and one such example is the National Robotics Initiative, based out of the NSF.

D1P4: Labels

Labels typically refer to some form of identification or annotation placed on data by people. Obtaining high-quality labels can in many cases be the most expensive aspect of data collection. Often, gathering unlabeled data is cheap because it requires little human oversight (e.g., downloading text from wikipedia or images from Google images). In some applications, such as labelling of medical data or data from particle accelerator experiments, data must be labeled by domain experts, who’s time is very valuable.

D1P4.1: To ameliorate the difficulties associated with labeling large datasets, we advocate for increased funding in dataset generation. Generation of high-quality quality datasets with high-quality labels is not a flashy job, but it often spurs advances in the field (e.g., the ImageNet dataset, which was an expensive undertaking, but since its inception has served as an invaluable tool for the computer vision community). Of course, we must acknowledge that incorrectly labeled datasets can have a detrimental effect, but

D1P4.2: We additionally suggest increased funding for machine learning approaches that make more efficient use of human experts, for example *active learning*. Active learning is a form of machine learning that allows the ML system to query an expert or other knowledge source (e.g., a person, or a simulator) for labels during the learning process. Typically, active learning algorithms are designed so as to query the expert for the highly useful information, thereby reducing the number of labels that must be provided by the expert.

D2: Developer Access

Developer access relates to the resources and capabilities that people who develop AI technologies must possess in order to develop, and in many cases advance the state of the art, in AI and deep learning. One resource, as described above, is data. However, other resources are needed: computers, software tools, developer communities, etc. Just like data, a tremendous problem faced by deep learning developers is the quantity of computational resources and other developer access tools required. Most academic labs cannot compete with the massive GPU (and now TPU) clusters used by the likes of OpenAI and Google. As a result, high-powered private industrial companies such as Facebook and Google, would be the only ones who could enjoy the benefits of developer tools to advance the state of the art. This means that the most powerful AI algorithms are controlled by a few large companies. We should seek a policy of supporting research and education in empowering people outside these centers of machine learning excellence to create novel AI tools.

D2P1: Computing resources. The primary obstacle to developer accessibility is cost. The hardware cost for a single AlphaGo Zero system in 2017, including the four TPUs, has been quoted as around $25 million (according to wikipedia). Certainly, a tier 1 University lab, let alone a small company or citizen-scientist, cannot afford such computational resources. The trend towards ever larger models that yield better performance on popular benchmarks while requiring more computational power to train

makes it increasingly difficult for labs with modest resources to compete with state-of-the-art (SOTA) performance on ML benchmarks.

D2P1.1: Shared Resources. Therefore, we suggest, just as the physicists can band together to raise funds for a common platform, such as a telescope, so should the academic AI researchers form a similar consortium for a shared resource. This could follow the already existing model of the Super Computer Centers, but some careful consideration must be given to the special needs of AI researchers and perhaps the more broad user community of such a resource.

D2P1.2: Efficient tool development. To better enable labs with smaller budgets and modest compute resources to compete with well-funded companies, we recommend that the NSF fund research in *computationally efficient* machine learning approaches, and *low-cost computing hardware (perhaps including robotics).* Investing more in computationally efficient ML approaches would have several benefits: firstly, it would provide more avenues of possible funding for labs that are capable of contributing, but cannot match the SOTA performance on benchmarks simply due to computational limitations. Secondly, the development of computationally efficient ML approaches would allow academic labs with modest budgets *to be competitive* with SOTA performance. Finally, efficient ML algorithms have the potential to lower the carbon footprint of ML research.

D2P2: Another challenge limiting developer accessibility is K-12 education. Opportunities for K-12 students to engage with computer science are not evenly distributed, putting segments of the population at a disadvantage when entering college. We therefore advocate for the expansion of computer science education in K-12. Computer science is unique in that compute resources, and even IT support for students can be easily shared by multiple schools.

The barrier to entry for becoming an AI developer for the community at large is unnecessarily high. Even state-of-the-art advances in machine learning can be broken down into a few basic steps: define the model, train the model, validate the model. While programming libraries exist (e.g., Keras) that massively streamline the machine learning development process, even just installing and using these tools requires a high degree of programming expertise and understanding of computer infrastructure, for example, proficiency in python and linux.

To lower the barrier to entry for potential AI developers, both in K-12 and in the community at large, we advocate for the creation and development of web-based tools, accessible to anyone with an internet connection, that allow machine learning workflows

such as data set handling, model definition, and training, to be represented through a simple and easy-to-use graphical interface. Any program created in this interface could then be converted to (e.g. python) code for the purposes of further development or sharing with the AI community.

D3: Democratization of AI Tools and Data.

Beyond data and developers, AI tools are often out of reach of most people who want to use AI tools to solve problems for their own businesses, or just personal research and education. For the United States to reach its full potential in using advanced computing to compete and collaborate with our peers in Europe and Asia, we must get the AI solutions into the hands of everyone. We believe that in doing so, everyone has an opportunity to voice how AI tools are used - in other words, we must democratize the use of AI tools.

As stated already, having solutions in the hands of a few large companies will limit our ability to solve complex problems. We are quite fortunate to have Tenosor Flow and Pytorch, owned by Google and Facebook, but as AI tools increasingly shape our lives, it is increasingly important that the power of tech giants does not go unchecked. Having the citizen-AI-scientist using AI tools to solve similar problems may actually serve as a check and balance to large companies, whose initial goals were to generate profit, who may misuse or abuse their capabilities.

One major challenge toward democratizing AI tools is the fact that large tech companies (such as Apple, Google, and Facebook) control much of the data generating pipelines, because much of the data these companies run on is generated by users using their products. While these tech giants offer a tremendous benefit to our economy and society, we cannot allow them to monopolize the AI market in perpetuity. We are inhibiting our growth if we sustain long-term difficulties for small companies or non-profit open-source ventures to break into the market. This challenge overlaps heavily with the issues discussed in data availability; however, here we are mainly focused on the assumption that companies own the data generated by their platforms, and how this limits democratization of AI tools.

To encourage competition in the AI market, as well as decouple data from data-generation platforms, we advocate for measures to be taken that allow users of AI technology to *own their own data.* Users could then choose to sell their data at free market price on a data market, thus lowering the barrier to entry of smaller companies and research groups. Such a data ownership model would also give users the ability to

vote with their data: if users do not like the way a particular company uses their data, they can choose to withhold their data from that company. Changing the data ownership model gives users of AI technology a seat at the table, instead of simply allowing big tech companies to be the sole arbitrator of how to use data and AI tools in whatever way makes them the most profit. Finally, it is our belief that having a free data market will have the positive side-effect of encouraging citizens to use and develop AI tools.