DAY ONE PROJECT

A National AI for Good Initiative

Anna Mitchell
Jasmine Sun
Jonathan Mak
Nick Rose

January 2021

Summary

Artificial intelligence (AI) and machine learning (ML) models can solve well-specified problems, like automatically diagnosing disease or grading student essays, at scale. But applications of AI and ML for major social and scientific problems are often constrained by a lack of high-quality, publicly available data—the foundation on which AI and ML algorithms are built.

The new administration should launch a multi-agency initiative to coordinate the academic, industry, and government research community to support the identification and development of datasets for applications of AI and ML in domain-specific, societally valuable contexts. The initiative would include activities like generating ideas for high-impact datasets, linking siloed data into larger and more useful datasets, making existing datasets easier to access, funding the creation of real-world testbeds for societally valuable AI and ML applications, and supporting public-private partnerships related to all of the above.

Challenge and Opportunity

Open-source data challenges are a proven way to attract top researchers to develop ML models. For example, the 14-million-picture ImageNet dataset was released in 2007 as a computer-vision challenge in which researchers competed to produce the best image-processing algorithm. But because assembling big datasets is a lengthy and expensive process, private-sector companies often have little incentive to share the big datasets they do create.

Funding the creation of big datasets and other open, shared AI resources is a powerful way for the federal government to drive AI and ML talent toward socially impactful and nationally strategic ends. Datasets can significantly accelerate progress in research related to (a) core AI technologies and techniques (e.g., computer vision, natural-language processing, and meta-learning); (b) applications of AI in science and engineering; and (c) applications of AI to societal problems. For instance, U.S. traffic and transportation data are currently dispersed across thousands of jurisdictions and companies. Integrating these data into a single accessible and responsibly managed dataset would help experts optimize freight routes, reduce transportation emissions, and anticipate supply-chain disruptions.[1] Other high-leverage domains for AI and ML include energy demand forecasting, medical diagnostics, and automated legal assistance.

There is little time to waste, as our nation's technological competition with China intensifies. The Chinese government is investing significantly in applied AI via its Made in China 2025 plan, data-

[1] David Rolnick et al., "Tackling climate change with machine learning," arXiv preprint (2019), arXiv:1906.05433, https://arxiv.org/pdf/1906.05433.pdf.

sharing partnerships with tech companies, and policy support.[2] As a result, many AI products achieved widespread Chinese public adoption before reaching the market in the U.S. (e.g. automated loan underwriting and facial recognition[3]), despite the fact that average citations for Chinese AI papers lag behind those for American AI research.[4] The new administration must close the data and deployment gap in order to shape global AI governance around our nation's values.

The Trump administration's American AI Initiative increased the budget for non-defense AI R&D from $1.118 billion in FY2020 to $1.503 billion in FY 2021, including $868 million to the NSF, $125 million to the DOE, $100 million to the USDA, and $50 million to the NIH.[5] Furthermore, OSTP reports have cited the importance of shared datasets, compute, and testbeds.[6] However, work remains to be done in identifying the datasets and other shared resources most needed by researchers, so that funding can be directed towards its most impactful uses. We propose several concrete ideas for identifying and funding these shared AI resources.

Plan of Action

The federal government should launch a multi-agency AI for Good initiative with a budget of at least $100 million per year, sourced from the National Science Foundation's FY2021 budget for AI R&D. This initiative funds opportunities for external research talent to work within and outside of the government to identify, create, and maintain shared datasets in domains of crucial public importance.

This initiative would be headed by the newly formed National Artificial Intelligence Initiative Office in the White House Office of Science and Technology Policy (OSTP) and operate in coordination with the NSF and the multi-agency Networking and Information Technology Research and Development Program (NITRD). It would support the following types of activities:

- **Reward New Ideas**: Encourage external researchers and practitioners in academia and the private sector to generate ideas for high-impact datasets across a range of domains. To catalyze public interest, the OSTP can partner with companies and nonprofits to run a widely marketed prize competition for ideas. Ideas for these datasets should be sourced

[2] Arjun Kharpal, "Power is 'up for grabs': Behind China's plan to shape the future of next-generation tech,"*CNBC*, April 26 2020, https://www.cnbc.com/2020/04/27/china-standards-2035-explained.html.
[3] Lulu Yilun Chen, "U.S. Blacklist Hurt Megvii's Sales Before IPO Attempt," *Bloomberg*, April 6, 2020, https://www.bloomberg.com/news/articles/2020-04-06/u-s-blacklist-hurt-china-ai-giant-s-sales-ahead-of-ipo-attempt.
[4] Sarah O'Meara, "Will China lead the world in AI by 2030?" *Nature*, August 21, 2019.
[5] John F. Sargent Jr., *Federal Research and Development (R&D) Funding: FY2021*, Congressional Research Service R46341, December 17, 2020, https://fas.org/sgp/crs/misc/R46341.pdf.
[6] The Select Committee on AI and the NSTC. *The National Artificial Intelligence Research & Development Strategic Plan: 2019 Update*.

through close collaboration with domain and technical experts in academia and the private sector.

- **Embed Research Talent**: Provide funding to support "embedding" postdocs and other researchers at federal agencies, prioritizing institutions with large amounts of data (e.g. the Veterans Health Administration) or funding institutions that generate valuable data (e.g. through the National Institutes of Health). Researchers will identify use cases for existing data, identify opportunities to fund critical datasets that the private sector is not incentivized to create, and lead efforts to share them with the academic community.
- **Release Existing Data**: Create an "ombudsman" function to prioritize public release of data that is owned or has been paid for by the federal government. An example is the collection of tens of millions of slides and tissue samples owned by the Department of Defense's Joint Pathology Center. If digitized, annotated, and shared publicly, this collection would be an invaluable resource for researchers, clinicians, and pathologists seeking to improve our ability to understand and diagnose diseases.
- **Encourage Private Sector Collaboration**: Provide funding incentives to encourage pre-competitive collaboration by companies willing to share the costs of developing shared resources.
- **Pilot Privacy-Preserving Practices**: Shared resources should be protected with cutting-edge privacy and security practices, such as differential privacy, synthetic data, and secure multi-party computation.
- **Promote FAT AI**: Use support for training data and other AI resources to address issues related to safety, bias, transparency, robustness, etc. Federally funded AI initiatives would present an opportunity to consider, test, and enforce standards around important issues in AI. For example, funding for a dataset could be conditioned on representativeness.

Past Precedents

The federal government has previously funded multiple applied AI initiatives to achieve broad policy goals. Examples include:

- **The American AI Initiative** included several investments in shared datasets for researchers.[7] Between 2016 and 2019, agencies sponsored the Naturalistic Driving Study, a dataset of over 5 million vehicle trips and the VA Data Commons, the largest linked medical-genomics dataset in the world.[8] Additionally, the administration has allocated $850 million for non-defense AI R&D at the NSF and issued a public RFI for applied AI datasets; outcomes have not yet been published.[9]

[7] The White House Office of Science and Technology Policy, *American Artificial Intelligence Initiative: Year One Annual Report. The White House*, 2020.
[8] The Select Committee on AI and the NSTC, *The National Artificial Intelligence Research & Development Strategic Plan: 2019 Update*, 2019.
[9] Office of Management and Budget, Executive Office of the President, *Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing; Request for Information*, July

- **SpaceNet** is a partnership between the public sector (through the National Geospatial-Intelligence Agency (NGA) and the Central Intelligence Agency (CIA)'s affiliated nonprofit, In-Q-Tel) and the private sector (through the companies Planet, Maxar, and Amazon Web Services) to release satellite imagery for ML challenge competitions in areas like building identification.[10] Winning models are open-sourced.[11] The SpaceNet competitions attracted top AI and ML researchers to produce high-performing models. This competition design could be replicated in more fields.

Conclusion

Artificial intelligence is beginning to unlock a range of applications, from enabling social workers to identify at-risk youth[12] to helping cities anticipate their exposure to extreme climate events.[13] But scaling up these efforts and enabling many more applications requires greater access to data. A prerequisite for many more transformative applications of AI and ML – to pressing problems in fields like healthcare, energy, and education – will be shared datasets and infrastructure widely available to top researchers.

American innovation has flourished through a decentralized and complex ecosystem of companies and universities. The new administration should therefore closely collaborate with academia and the private sector to find the best ideas for datasets and other shared resources, construct or release these datasets, and create competitions and other mechanisms to encourage the development of applied solutions at scale. With this toolkit, the new administration could have high leverage against specific hard problems. For example, funding a dataset in a field of strategic national importance like energy would allow the U.S. government to define the problem, set the agenda for an entire field, then attract the most talented researchers and engineers to develop the best solutions at scale. A national AI for Good initiative could attract top talent and spur the development of solutions to some of our greatest national challenges.

10, 2019, https://www.federalregister.gov/documents/2019/07/10/2019-14618/identifying-priority-access-or-quality-improvements-for-federal-data-and-models-for-artificial.

[10] "About Us," SpaceNet, accessed 2020, https://spacenet.ai/about-us/.

[11] Ryan Lewis, "SpaceNet Turns Four" *The DownLinQ*, August 10, 2020, https://medium.com/the-downlinq/spacenet-turns-four-fb646e32ba5a.

[12] Aida Ramattalabi et al., "Exploring Algorithmic Fairness in Robust Graph Covering Problems, *33rd Conference on Neural Information Processing Systems* (Vancouver: NeurIPS 2019).

[13] "ClimateNet," National Energy Research Scientific Computing Center (NERSC), last modified October 17, 2019, https://www.nersc.gov/research-and-development/data-analytics/big-data-center/climatenet/

Frequently Asked Questions

What differentiates the AI for Good proposal from past federal data initiatives?
Past federal data initiatives have largely not leveraged advances in AI and ML to build data products, partly due to a lack of domain knowledge, the right datasets, and financial resources.

By contrast, the AI for Good initiative aims to produce AI and ML solutions that can be implemented at scale and adopted by the communities who need them by:

- Coordinating experts across domains to identify datasets and AI/ML models that have the greatest potential for practical impact.
- Ensuring that federally assembled datasets are relevant, labeled, and suited to develop predictive models.
- Identifying and collating datasets not owned by the federal government, such as those at the state and local levels or housed in the private sector.
- Establishing relationships with top AI researchers to develop models using these datasets.
- Using the National Research Cloud proposal signed by 22 top universities[14] to provide researchers access to computational resources.

What are some examples of high-impact datasets to collect?
The Climate Change AI research organization has suggested a range of high-leverage datasets.[15] Some that seem particularly appropriate for government funding include:

- Satellite data: Much is public but scattered, and requires collation. This can support solar energy forecasting, climate modeling, and global intelligence operations.
- Traffic and transport data: Data on arrival times, traffic counts, and more are scattered across localities with different data standards. These datasets can support infrastructure and urban planning and improve supply chain resilience.

This initiative could catalyze similar intra-institutional efforts in other domains such as economics, healthcare, and education.

What shared AI infrastructure should the federal government invest in besides datasets?
Additional resources that could accelerate progress in AI include (1) real-world testbeds for reinforcement learning, (2) open-source libraries, and (3) secure data-labeling tools.

Reinforcement learning is the study of a computer agent as it learns through interactions with the environment. Many researchers have used games to train agents. The federal government

[14] John Etchemendy and Fei-Fei Li, "National Research Cloud: Ensuring the Continuation of American Innovation," Stanford University Human-Centered Artificial Intelligence, March 28, 2020, https://hai.stanford.edu/blog/national-research-cloud-ensuring-continuation-american-innovation.
[15] Climate Change AI, "Tackling Climate Change with ML: Dataset Wishlist," 2020, https://docs.google.com/document/d/1E1vhuGNiUWUbij8WxqNbTfytxfyaS9LnaBEdx58KQQc/edit

can fund more diverse and sophisticated testing environments that more directly relate to envisioned real-world applications. For example, researchers participating in the Autonomous Greenhouse Challenge co-sponsored by Wageningen University in the Netherlands developed algorithms that increased the productivity and sustainability of indoor agriculture.[16]

Open-source libraries expedite development of AI models by allowing researchers to import code for common tasks. For example, libraries might provide code for common tasks like preloading datasets, parallelizing machine learning, and logging progress.

Data-labeling tools are necessary to efficiently create big datasets that can be used to train AI and ML algorithms. The federal government could create tools tailored to specific domains. For instance, secure tools that respect HIPAA privacy will be needed to create big training datasets for medical applications.

How can the federal government build a pipeline from research to applied AI?

The first step in building a pipeline from research to applied AI is to understand AI researchers' existing interests and needs. Armed with this knowledge, the federal government can strategically determine what domains are most in need of support, what valuable data needs to be unlocked, and where grant opportunities will be most impactful. The federal government can help meet these needs by forging partnerships with varied members of the AI community (e.g., universities, companies, and policymakers).

Competitions and challenges—along the lines of Kaggle competitions,[17] SpaceNet,[18] or DARPA Prize Challenges[19]—hold great promise for driving researchers toward high-impact work in applied AI. To be most effective, competitions and challenges should focus on a practical AI problem chosen by domain experts. Participation incentives include monetary awards, computing resources, and/or recognition at top conferences. For federally sponsored competitions and challenges, winning models should be open-sourced, and winning teams connected with resources or partners to implement their solution.

How can we ensure that datasets are secure and privacy-preserving? How can we ensure that datasets and models are unbiased and equitable?

The AI for Good initiative is an opportunity to establish clear standards for ethical, secure, and equitable data science across multiple domains. The following general recommendations should underlie the initiative, with adaptations made as necessary by experts depending on the specific domain in question. In all cases, trust relies on the relevant bodies being transparent about how data is collected and governed.

[16] "Autonomous Greenhouses 2nd Edition," Wageningen University & Research, https://www.wur.nl/en/project/autonomous-greenhouses-2nd-edition.htm.
[17] "Competitions," Kaggle, https://www.kaggle.com/competitions.
[18] "Spacenet™: Accelerating Geospatial Machine Learning," Spacenet, 2020, https://spacenet.ai/
[19] "Prize Challenges," Defense Advanced Research Projects Agency, https://www.darpa.mil/work-with-us/public/prizes.

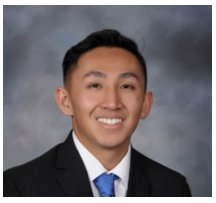
- Collecting data. Data collectors should use an ethics framework such as the "5 Cs: consent, clarity, consistency, control, and consequences."[20] The framework should inform what data is collected and how they are protected before being shared: for instance, by using de-identification and differential privacy techniques.
- Governing datasets. Datasets should be governed by a clear set of security principles, such as those published by the Precision Medicine Initiative to "Identify, Protect, Detect, Respond, and Recover" risks.[21]
- Building ethical models. AI researchers using the data should be encouraged to evaluate their computational models for bias, impact, and fairness. Evaluations may involve technical auditing techniques as well as self-assessments to anticipate potential harms.

[20] DJ Patil, Hilary Mason, Mike Loukides, "The five Cs," O'Reilly, July 24, 2018, https://www.oreilly.com/radar/the-five-cs/.
[21] The White House, "Precision Medicine Initiative: Data Security Policy Principles and Framework," May 25, 2016, https://obamawhitehouse.archives.gov/sites/obamawhitehouse.archives.gov/files/documents/PMI_Security_Principles_Framework_v2.pdf

About the Authors

Anna Mitchell is a Senior Associate Product Manager at Schmidt Futures, where she works on projects related to advancing and protecting free societies using technology. Her writing has been published by *The Atlantic* and Stanford Law School as well as the *Stanford Review*, where she was Editor-in-Chief. Anna holds a B.S. in computer science with a concentration in artificial intelligence from Stanford University.

Jasmine Sun is a Research Intern at Schmidt Futures and a fourth-year undergraduate at Stanford University. She has led curricula and instruction for multiple Stanford computer-science courses in civic technology and has previously worked at Bridgewater Associates, One Concern, and in technology policy research.



Jonathan Mak is an Impact Fellow and works as a Product Manager at COVID Act Now. He previously worked at Apple on CoreMotion Health, the United Nations as a Product Manager, Amazon on the Photos Team, and DoNotPay as one of the first product managers/engineers developing chatbots. Jonathan graduated from Stanford University with an M.S. and B.S. in electrical engineering with a focus in signal processing and machine learning.

Nick Rose is a Product Manager at Google working on applied AI. He graduated from UC Berkeley with a degree in computer science and helped teach a class on artificial intelligence.

About the Day One Project

The Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans. For more about the Day One Project, visit dayoneproject.org