AI Data and Model Economy (AIDAME)

Dr. Vijay Gadepally, Dr. William Streilein
{vijayg, wws} @ ll.mit.edu

Lincoln Laboratory
Massachusetts Institute of Technology

Motivation

Artificial Intelligence (AI) has the ability to transform American competitiveness by improving the efficiency of operations, improving the quality of work, simplifying complex transactions and supporting technological innovation [1]. Recent developments in computing and algorithms coupled with "Big Data" have ushered in a new era of tools and technologies [2].

While the United States continues to remain the leader for AI innovation, much of this innovation has been driven by the commercial sector, where applications with commercial potential are realized and data from internal, proprietary sources are generated and maintained. While the commercial sector is making great advances, the academic community focuses on algorithms and models, applied to smaller, publicly available datasets. Across an AI ecosystem comprised of small, medium and large businesses, government, and academia there is a major divide in access to resources, datasets, models and expertise to effectively learn, adopt and deploy AI at a societal level.

To keep up with near-peer adversaries such as China, the United States will need to make strategic investments in technology, workforce development, and in addressing policy issues that are currently positioned to hamper the widespread development, deployment and adoption of critical AI technologies in the coming decade. These investments will shift the emphasis of government investments towards the creation of a meaningful economy based upon data and models that support and maintain AI innovation already within the United States. Such an economy will be a free market ecosystem that rewards contributors that move the state of AI in society, its data and models, and/or advanced research forward.

As an example, adversary states such as China have an inherent advantage in developing AI technologies due to their ability to gather data from their own citizens (e.g. social credit system, mass video surveillance) [3]. While such a surveillance system is antithetical to American values of privacy and personal freedom, this gives China an advantage in developing AI capabilities that learn from the massive quantities of labeled data they are collecting. The United States will need to invest in technologies that are capable of competing with such competitors without compromising personal freedom and privacy. We believe that with appropriate Federal support, an environment can be designed that allows for the protection of core American principles while maintaining the quality and effectiveness of American Artificial Intelligence innovation.

The central thesis of our proposal is to create an economy of AI data and models that incentivizes sharing through legal mechanisms and copyright, in a way similar to the way intellectual property is currently protected. Such a system would intensify the interest in relevant data sets and build models that would support long term community focus on particular "AI hard" problems, leading to the eventual discovery of leap-ahead solutions.

Further, such incentives would spur innovation from small businesses and startups that often lack the resources to collect massive amounts of data and train computationally intensive models that often take days to months on specialty computing platforms [4]. Not only would individual entities be interested in participating in such an economy, but such organizations could leverage mechanisms to ensure a return on their research investment, if the AI technology does not succeed. More time could be spent up-front thinking about the data that is necessary, curating and maintaining it over time to support long-term research.

In addition to a data sharing economy, we emphasize the need for more research into privacy, security and ethics of AI systems and models. For example, can we devise a facial recognition system that does not require access to data on law-abiding citizens but is still able to recognize dangerous actors? Below, we discuss a few proposed solutions:

Proposed Solutions

There are a number of related initiatives that the White House could pursue in support of the AI Data and Model Economy (AIDAME). These initiatives include support for: novel, but necessary scientific research; infrastructure to enable sharing and further active research pursuits; and rewards that would incentivize others to participate at a broader level in the development of a robust data/model ecosystem:

Model.Gov

Building upon the success of the data.gov initiative, a repository should be created that captures and maintains AI models that have been built upon publicly-available data, similar to that which is place on data.gov. Such models could be built from state-of-the-art AI architectures that are freely available on the Internet, such as ResNet [5], AlexNet [6] and VGGNet [7], though others can be leveraged as well. An incentive or reward structure should be created to give back to participants who contribute ground-breaking research, datasets or models.

The Model.Gov portal could also help connect small and medium businesses with partners from other organizations that could license models, subject matter expertise and/or datasets.

Data.Gov

While data.gov is an excellent example of the types of infrastructure needed for AI development, it is currently largely targeted towards Federal and government entities[1]. Expanding participation to academic and commercial organizations could provide a "one-stop-shop" for AI related

[1] https://catalog.data.gov/organization

datasets. Further, policies/procedures are needed to incentivize contributions and updates to data as well as source code needed to extract, transform and load data.gov datasets. Further, new technology is required to protect the confidentiality, integrity and availability of datasets that are shared via the data.gov portal.

Proactive Data and Model Policy

For data and models that are collected and stored on central repositories, such as data.gov and model.gov, policy will be developed that guides how models and data are collected and shared in support of research. Such policy will include stipulations for access to artifacts that result from model-learning and the use of potentially private data sets. Specifications for data and models will enable standardization of practice which will further support open research advances. Finally, proactive policy will provide guidelines for ownership of models that are shared, permitting researchers, developers and model authors to benefit from sharing and further use by others.

Federal support for Data and Model Sharing.

The US Government should provide financial support for organizations, both within the government and from the commercial sector that motivates sharing of data and models on an open register, such as, data.gov. In some cases, the federal government can require and incentivize publishing of AI related data and models to this ecosystem. For example, data sharing on data.gov can be required by recipients or federal funding for which a majority of the funding is going towards data collection or curation.

In support of US economic and strategic goals, a Federated AI Center should be established to oversee and leverage the proposed data and model economy in the development of advanced AI capabilities that further department and agency missions.

Research Funding

In order to keep up with near peer and adversary nations, we encourage strategic investments into a number of under-resourced disciplines within Artificial Intelligence. In particular:

- **Applications to National Security and Government**. An important benefit of sharing openly data and models will be accelerated development of advanced AI capabilities. As AI continues to become an important component in the nation's economy and defense, it will be necessary to further explore aspects of data and model applications which relate to national security and government.
- **Technical issues related to data and model storage and sharing, particularly of multi-modal data**. Current capabilities are good at storing text-based data, such as CSV files. More work needs to be done to improve capabilities for storing non-text data (e.g. video, speech, etc.) and binary data, such as from scientific instruments (e.g. biological sensors). Additionally, research could be pursued into composable models (e.g., mixing and matching subcomponents from different AI models).

- **Research into Data and Model Privacy**. As AIDAME begins to take shape, it will be necessary to consider and develop capabilities that permit, provide and preserve privacy when necessary. Promising capabilities from cryptographic research, such as differential privacy can be further explored and applied specifically to data and models at rest, such as will be on a repository. Data and Model Privacy research can also ensure that it is difficult to reconstruct data from models and provide new tools to enable model training on encrypted or protected data (such as in health care and other privacy-centric applications).
- **Adversarial Considerations**. The open nature of AIDAME will undoubtedly invite attention of adversaries wishing to benefit from available data, but also hinder our ability to use it in developing novel AI capabilities. Research is needed into enhanced mechanisms for protecting the data and models and enabling availability of the repositories, but also the function of the data and model economy.
- **AI Data/Model Ownership**. The open nature of data/model sharing will require incentives from the federal government. These incentives could come in the form of financial support or a robust protection for entities that develop and share data/models such as intellectual property. While there are a number of options, careful study is required to promote desired outcomes.

Conclusion

The current age of AI innovation is due in large part to availability of both increased computational capability and large, truth-marked data sets. The United States is in a position to benefit greatly from these advances in AI both from an economic, but also from a national security perspective.

However, US success with AI is not guaranteed and requires focused initiatives in policy, research and in the creation of supporting infrastructure. As an important complement to the successful and open data.gov, we argue for the creation of model.gov for sharing and leveraging AI models which have been built from both publicly and privately available data sets. Together with a well-designed incentive structure, data.gov and model.gov can be the basis for AIDAME: an AI Data and Model Economy, which will continue to grow and support rapid advancement across the diverse American AI ecosystem.

White House initiatives will be crucial to the successful establishment, nurturing and maintenance of an economy which relies on both data and models and supports a new era of AI innovation.

References

[1] Gadepally, V., Goodwin, J., Kepner, J., Reuther, A., Reynolds, H., Samsi, S., ... & Martinez, D. (2019). AI Enabling Technologies: A Survey. *arXiv preprint arXiv:1905.03592*.
[2] Gadepally, V. N., Hancock, B. J., Greenfield, K. B., Campbell, J. P., Campbell, W. M., & Reuther, A. I. (2016). Recommender systems for the department of defense and intelligence community. *Lincoln Laboratory Journal*, *22*(1).
[3] https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4
[4] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." *arXiv preprint arXiv:1906.02243* (2019).
[5] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
[6] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[7] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.