The Hitchhiker's Guide to the Galaxy of Artificial Intelligence

EISP Summer Meeting 2021 Edition

Although the synopsis of the H2G2 ("*The Hitchhiker's Guide to the Galaxy*") by Douglas Adams starts with the demolition of our beloved Earth, I assure you this guide is on a much friendlier tone to our home planet. However, there is a striking similarity of the origination of this guide that is the disruptive nature of artificial intelligence (AI) technologies to the Earth and space science (ESS) community. Even if you are passively paying attention to the trend of Earth and space science research, AI and machine learning (particularly deep learning), or AI/ML, are rapidly gaining steam driven by the vast ESS data, powerful computing resources, and revolutionary open science movement.

As the ESS community embraces the potential of AI/ML, ESIP continues to lead the collaboration on advancing the AI/ML innovation with cluster activities and Lab grants. At the upcoming semi-annual gathering of ESIPers, I am particularly excited about the variety of AI/ML relevant events and the dynamic future activities these discussions may lead to. Thus, I want to share my excitement about these activities with the community and hope they can be useful to others (and maybe bring some fellow hitchhikers to the virtual journey with me during the week). This guide is organized by three topics — data, computing, and models — based on my own interpretation of three major components of AI/ML innovation.

Data – the fuel to power the fleet

All AI/ML models are **powered** and also **constrained** by data. In ESS, this data ranges from satellite observations to model simulations to in situ and suborbital sensor measurements. To develop an AI/ML application using ESS (and other) data, the diversity of ESS data is a blessing and a curse. We are blessed to have so much ESS data that can power our AI/ML model, but we often spend most of our development time preprocessing data from different sources so they can work together in the model. There could also be missteps in the data preprocessing stage that lead to misguided AI/ML models. Wouldn't it make sense to define a community standard for ESS data that enables efficient and appropriate AI/ML development?

ESIPers, including both AI/ML researchers and data providers, step up collaborative effort between Data Readiness Cluster and NASA Earth Science Data System Working Group (ESDSWG) on data interoperability with a two-session track on "AI Data Readiness".

- AI Data Readiness: Designing A Community-Driven Road Map for Data Standards and Tools (2021-07-20 13:30 ET)

- AI Data Readiness: What Does ML Training Data Interoperability Mean to You? Examples and Use Cases (2021-07-22 16:00 ET)

These sessions will benefit from the broad expertise and collaboration of the ESIP community on data quality, usability, interoperability, cloud, and semantic technology. I will also closely follow the session on data quality, cloud-optimized data, heterogeneous tabular data, and non-tabular ecological data as they all may affect the development of AI-readiness standards.

Computing – the engine to lead the fleet

Assuming that we spent all our time making ESS data ready for AI/ML development, the next critical topic that we face is the computing infrastructure. Although there are AI/ML models used for small sample applications, the majority of AI/ML models are used for "big data" problems. During model training and development, we need substantial computing resources to find the optimal model structure and parameters that can best fit the large set of data points without overfitting. We are always looking for innovative solutions in both hardware and software spaces to address the computational demand and accelerate model development. Thus, I am very excited to see tutorials and discussions led by the Machine Learning Cluster in this frontier.

- New Frontiers in AI for Earth and Space: Big Data and Parallel Computing (2021-07-21 14:30 ET)

Model – the rapid growing fleet

With appropriate data and sufficient computing power, AI/ML models can show the potential to help the ESS community to be more efficient and innovative in scientific discovery. As a disruptive technology, AI/ML models are also rapidly evolving and continuously integrated with other technologies such as semantic technologies. There have been ongoing discussions within the ESIP community on how we can combine knowledge graphs and AI/ML to take advantage of the structured data and domain knowledge and make data science methods more efficient and explainable. This summer, you can learn more about graph-based data science workflows with hands-on tutorials by Paco Nathan.

- Graph-Based Data Science (2021-07-21 11:00 ET)

As a forward-looking community, ESIP is always concerned about how to ensure research objects (e.g., data, software, and models) are reusable and reproducible. Given the complexity of AI/ML models and their reliance on appropriate data, the ESS community is tackling the reusability challenge for AI/ML models head-on through community-driven guidelines supported by NASA ESDS.

- Best Practices for Reusability of Machine Learning Models: Guideline and Specification (2021-07-22 11:00 ET)

Additionally, you can also find robust discussions on how to ensure reproducibility of AI/ML research through standard code-sharing. During Wednesday's *Research Showcase* (2021-07-21 18:00 ET), there will be a poster on developing community recommendations on reproducible code-sharing for AI/ML research by leveraging existing work from the NeurIPS conference. If reproducibility or AI/ML is your jam, you should try to stop by and share your wisdom

Please consider this guide as an invitation for you to discover your own connections with the "AI Galaxy" no matter what your expertise is. Nonetheless, I will feel remiss in not mentioning "Deep Thought" in this H2G2 inspired guide. If Earth is a supercomputer designed by Deep Thought to determine what is the Ultimate question, can we, the ESIP community, collaborate to decipher the designing principle of Deep Thought via AI/ML using the vast ESS data? Until then, see you all at ESIP meetings!