PRAGMA Perspective on AI, Data Cyberinfrastructure, and Training

Sept. 13, 2019

Ruth Lee
Korea Institute of Science and Technology Information
South Korea

1. What are the AI applications that your organization/project is working with?

- Several AI applications in KISTI
- Focusing on especially computational science because of our major users of EDISON
- Trying to exploit big data generated by simulation programs, and to build AI models with those simulation data to accelerate large-scale screening to save time
- As a pilot system in materials science, we have built

  an atomic property prediction model for inorganic compounds and

  a pore-geometry similarity quantify model for nanoporous datasets, and so on.
- Developing many AI models for material science is the most important issues for enabling next-generation computational science with AI so far.

2. What are the unique infrastructure, sociotechnical, or legal challenges in AI research involving international partners? Speak from experience if possible.

- Not yet to have any experience…

- Are really domain scientists willing to open their research data?
  - Maybe some domains/scientists are … but many others are not…
  - How to make them to share the data and collaborate with them?
  - Not allow to open especially medical data in Korea even for the research purpose
- Is the data generated from domain scientist exactly the one that data scientist wants or needs?
  - Who has the responsibility to generate proper data for AI?
  - Is it easy to set the rule for the data between domain and data scientists?

- Having enough computing resources and well-defined/-structured dataset, and using the domain customized/automated/web based easy-to-use AI platform for the international AI research could be the first step to move forward….

3. What are the unique training (AI/ML algorithms) challenges for AI research?

- Most datasets for practical AI research in computational science are mutually independent…. Except for some of well-known datasets for AI beginners.

  - VASP and Quantum Espresso have similar goals, but their simulation datasets cannot be shared although target compounds are the same.

  - Because design principals and related algorithms are not the same and also have different units to present data.

- How about a large number of the in-house codes? Since well-structuring all of them is impossible, **the most of AI challenges in computational science is how to handle the unique datasets**.

4. Where do you see AI in 5 years? What are the barriers to getting to that vision?

- Huge attention to AI not only commercial sectors but also science and research sectors in the next 5 years.
- Success of AI definitely depends on DATA… REAL GOOD DATA….
- **Domain scientists** are willing to apply AI technologies to their research and showing a lot of interests.
  - But, not easy to directly apply and not available practical solutions / models for the specific domain yet….
- **Data scientists** need to have well-defined and -structured dataset from the real applications to see the feasibility of AI
  - But, not easy to get the proper and enough dataset….
- AI-related start-ups in South Korea mostly work for consulting to make AI-solutions to SME.
  Most SME do not know how to generate and manage data for AI.
  - Big companies do not share data and know-how
    because they know the power of data and they have own specialized AI group and AI services for their own customers.
- To overcome the biggest huddle for generating good AI data is making a collaborative environment between a domain scientist and a data scientist to well-defined and -structured accurate and enough amount of datasets for AI.

Thank You!!!