

Title: Data-Centric AI in the era of LLMs (July 29, 2:20PM-3:25PM HST)

Moderator: Praveen Paritosh **Organizer:** Manil Maskey

Key Question: In the era of foundation models, what is the data-centric research agenda?

Confirmed Panelists:

- Ludwig Schmidt, UW
- Megan Ansdell, NASA Planetary science
- Nathan (NATO) Lambert - Hugging Face
- Sang Michael Xie, Stanford (DoReMi paper)

Prep meeting agenda:

- Get to know the panelists
- Go over agenda/format
- Discuss potential topics/questions
- 5-minute intro (Send us slides ahead of time if you want to present)
- Meet us at the [ICML Data Social](https://icml.cc/virtual/2023/social/29264) [Meeting Room 314, Thu Jul 27 5:45 PM - 7:30 PM HST]
 - <https://icml.cc/virtual/2023/social/29264>

Preliminary Panel Agenda:

- 2:20-2:25 **Introduction to the panel (5 min) - Praveen**
- 2:25-2:45 **5 minutes per speaker - background and project (20 min total)** - Slides or no Slides?
- 2:45-3:05 **Panel questions (20 min)**

Broader themes:

- Data questions: *Once you apply LLMs to expert domains, all hell breaks loose!*
 - What are the existing processes in curating domain specific training datasets for LLM?
 - What about structured data?
 - Synthetic vs Natural data?
 - A lot of LLMs are trained on private corpora. What role does public data have?
- Evaluation questions
 - Strategies for consistent evaluation framework
 - How do we evaluate LLMs in new domains requiring expertise?
 - What is missing in having an industry-wide benchmarking standard?
- Call to action, what should we do?
 - Infrastructure: Github for Data
 - Public datasets, benchmarks, research agenda
- 3:05-3:15 **Audience questions (10 min)**