Position: Machine learning data curator
Job Type: Permanent, Full-time
Location: Canada, United States
Reporting to: CEO/CTO

Overview:
Theia is seeking a dynamic and experienced **Machine learning data curator** to extend our cloud-based data generation pipeline. In this role, you will serve as the data guardian, developing and maintaining data generation processes that seamlessly integrate into machine learning workflows.

High level, this position is responsible for full data life cycle management, including intake, analysis, consolidation, and organization, primarily for next generation R&D. The ideal candidate has experience in this type of role and thrives in a small team environment that is data and results driven, with a strong emphasis on scalable process and cloud environments.

Why Theia?
At Theia, we are redefining motion capture through our industry-leading markerless motion capture software. Our team of scientists, engineers, and developers have extensive experience in biomechanics and ML research and are leaders in the field of movement tracking and analysis. We are committed to providing accurate, repeatable, and reliable solutions for academic and commercial applications, bringing a high standard of excellence to everything we do, whether it's for our internal teams or for our thousands of customers. Our approach to any problem is results and objective based, making the work environment flexible to cater to the individual requirements of the team member.

Responsibilities:
- **Data Intake** - refine existing data intake process from remote data collection sites, extending on existing systems that assess data quality and completeness
- **Pipelines and Analysis** - refine existing cloud-based analysis procedures, to scalably analyze data set used downstream for biomechanical analysis and machine learning ingestion
- **Quality Assurance and Integration** - establish systems to allow for manual QA on analyzed data sets, as well as broad automated processes to refine and improve data quality
- **Data Curation** - combine data sets from multiple data sources, establishing scalable pipelines for final data generation used in machine learning experiments

Qualifications and Skills:

- Master's or PhD in Computer Science or related degree
- Mastery of the English language with demonstrated writing expertise of technical concepts
- Mastery of scalable cloud systems (google cloud compute, linux environments)
- Very strong skill in python, C++
- Experience in bazel build systems QT, CUDA, ML data formats (webdataset, arrayrecords), preprocessing pipelines for ML workflows
- Proven track record of developing high-quality cloud-based pipelines for data generation
- Strong understanding of pytorch and jax
- Experience with preprocessing pipelines in machine learning frameworks

Benefits:

- Possibility of remote work
- Flexible working hours
- Comprehensive health benefits
- Collaborative and inclusive work environment
- Extended holiday weekends and flexible time off

How to Apply:

Please submit your resume and cover letter outlining your qualifications and why you would be a great fit for this role to jobs@theiamarkerless.com. Application deadline is December 15th, 2025.