



Needs for Diverse Translational Biomedical Interoperable Computational Tools

Dr. Nathan Hotaling and Team!

Senior Data Scientist - National Center for Advancing Translational Science -
National Institutes of Health

Senior VP of Data Science - Axle Informatics

2023-12-05

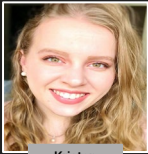
The Team!



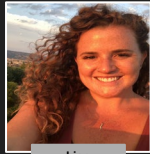
Andrew



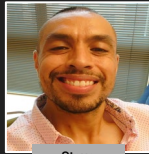
Arjun



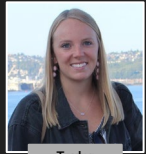
Kristen



Lisa



Steve



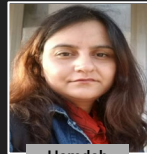
Taylor



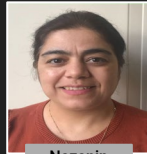
Nick



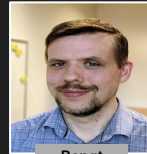
Brandon



Hamdah



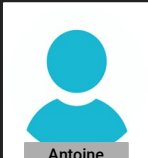
Nazanin



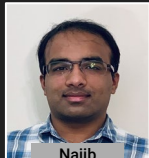
Bengt



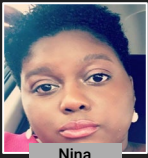
Camilo



Antoine



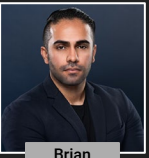
Najib



Nina



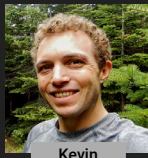
Kevin



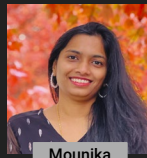
Brian



Chris



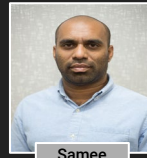
Kevin



Mounika



Tim



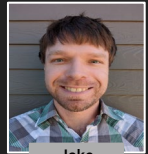
Sameer



Andre



Jesse



Jake



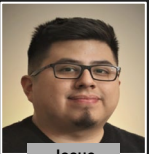
Phil



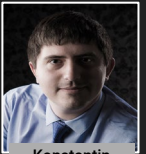
Sunny



Aaron



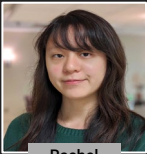
Josue



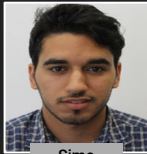
Konstantin



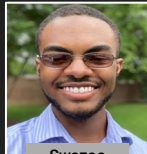
Nikita



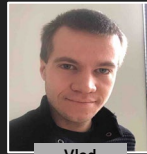
Rachel



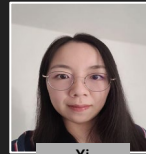
Simo



Swazoo



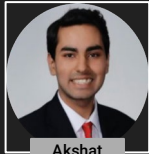
Vlad



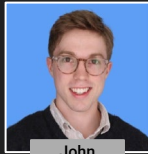
Xi



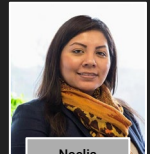
Aditi



Akshat



John



Noelia

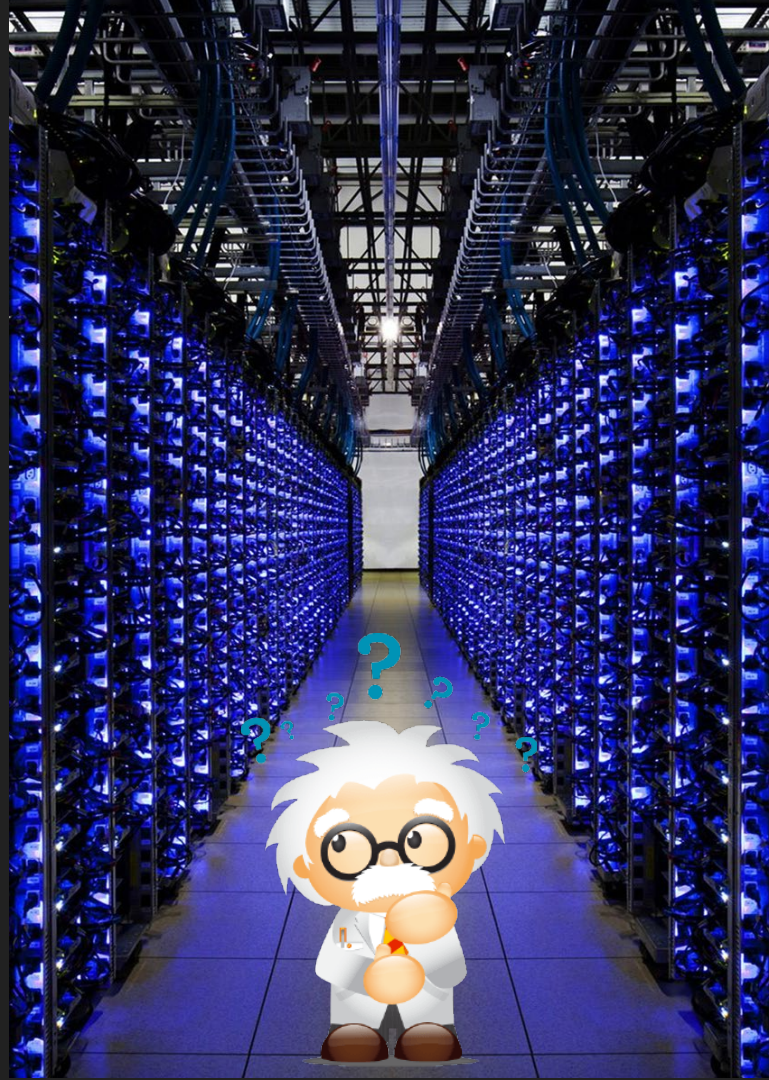


Nathan



Foundational Problem

How can a scientist harness petabytes of data to derive quantitative measurements and gain meaningful insights?



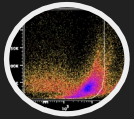
National Center
for Advancing
Translational Sciences

Polus Platform - Open Source Adobe Creative Cloud... for Science



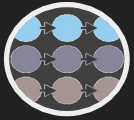
Constant Processing Time For Increasing Data Sizes

- Web-based solution to access cloud/server/ HPC computational and storage resources



Collaborative, Visual, and Interactive

- Web based UI; Simple point-click interface



Traceable and Reproducible Science

- Environment, packages, functions, etc. all recorded automatically; automated hyperlinked provenance to raw data



Analytical

- Validated measurements/algorithms; Supports latest in machine learning, deep learning, and statistics



Flexible and Modular

- Open source Plugins/Applications/Codebase; Interoperable container-based. Shared Metadata, Storage, and AuthN/AuthZ



Secure

- Private research, patient data/PII, and IP is securely and safely stored



National Center
for Advancing
Translational Sciences

Diverse Biomedical Data & Tooling

Began with Imaging

- Microscopy - light, fluorescent, electron, pathology, high content screening
- Clinical - Fundus images (RGB, fluorescent), X-Rays, CTs, MRIs, OCTs, etc.

Expanded Out!

- Cheminformatics, molecular modeling, real world evidence (electronic health records, insurance/payer claims, mortality, etc.)

On the Horizon

- Omics - genomics, transcriptomics, proteomics, metabolomics, etc.



Diverse Tool Needs

Language Support

- Python, Java, R, C/C++, and Rust



Diverse Libraries

- Traditional, Tensorflow, Pytorch, Jax, etc.
- Diverse driver requirements i.e. CUDA



Diverse Hardware

- CPU - x86 and ARM
 - AVX/AVX2
- GPU - Nvidia
 - Historical code written across multiple Nvidia compute capabilities
- RAM - 8GB \rightarrow ∞



Diverse Users Needs

Programmers

- SDKs and libraries

Technical

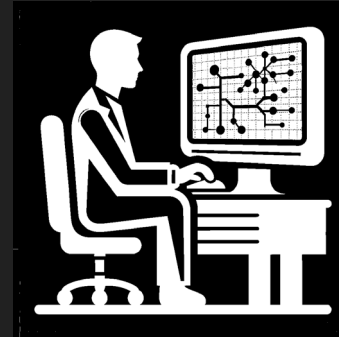
- Command line tools & terminal
- Direct edit YAML/JSON

Advanced Domain Experts

- Complex GUIs
- Configure/optimize hyperparameters
- Some "Macro" scripting

Domain Experts

- Simple GUIs
- Single button if possible!



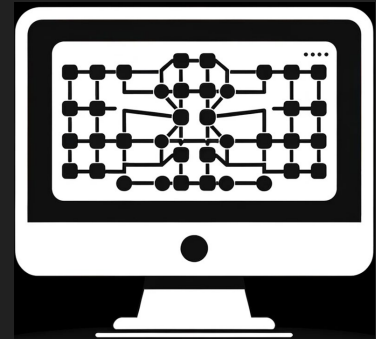
Diverse Data Sizes and Scaling Needs

Dataset Scales Processed

- < 1GB - Text and image data
- > 100s of TBs - Imaging

Scaling Paradigms

- Split on data
 - Wells of plates in HCS
 - Molecules binding to proteins
- Scale on computation
 - Molecular modeling
 - Feature extraction
 - EHR/Claims extraction, transformation, loading, filtering, modeling



Interoperable Analysis Tools

100+ Scalable Interoperable Tools!

- Background correction, Table Merge, etc.

Deep Learning Training/Inference Models

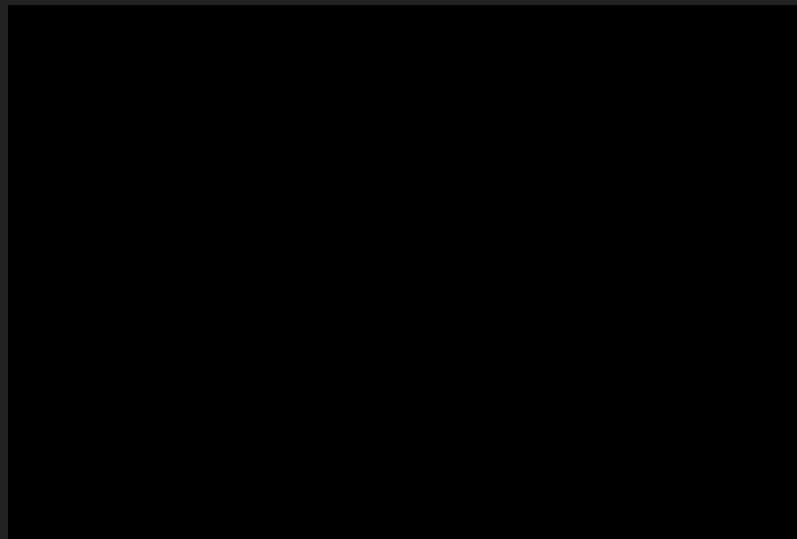
- Point/click interface for training & inference
- Over 400+ classification networks
- Over 4200+ segmentation networks (Video→)

Custom Deep Learning Models

- Image segmentation models - MESMER, CellPose, Spline/Star-Dist, etc.
- LLMs - Llama 2s, Mistral, Falcon, etc.
- Chemi-informatic - MolGAN, diffDock, etc.
- MM - OnionNet-pose, AlphaFold-2, BioExcel/BioBB, etc.

Trained/Implemented

- 4000+ image segmentation models on diverse images



Questions?