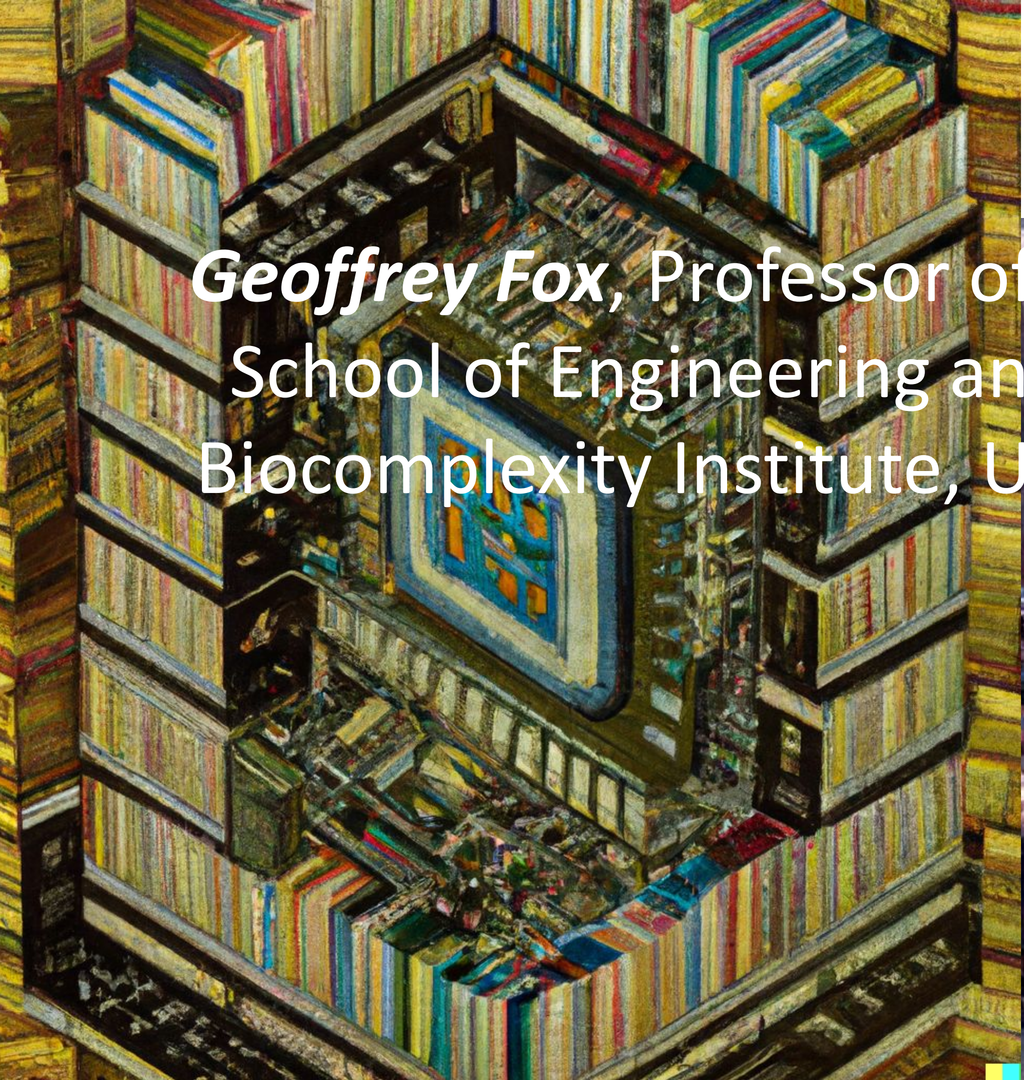
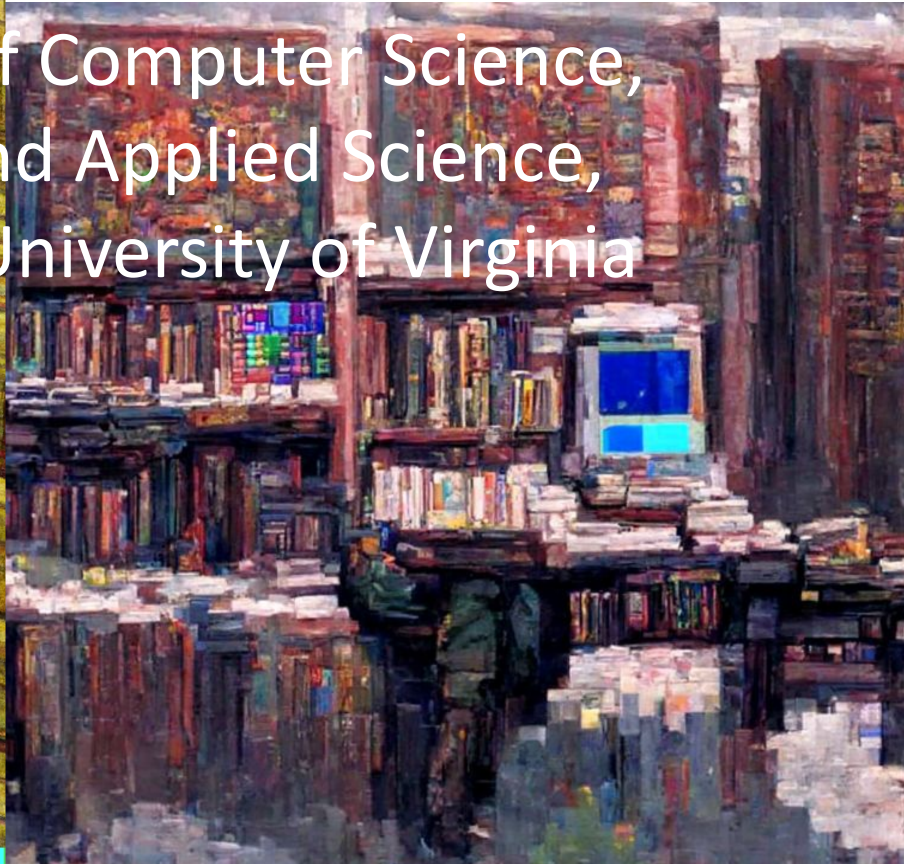
Geoffrey Fox, Professor of Computer Science,
School of Engineering and Applied Science,
Biocomplexity Institute, University of Virginia

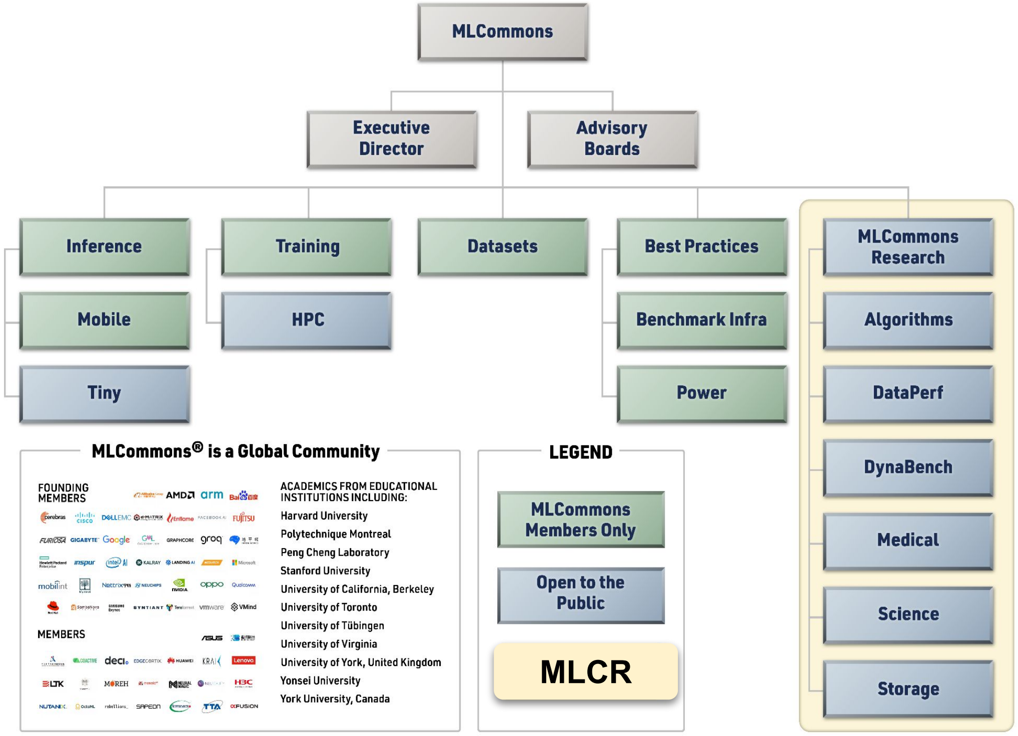




Big Data Computing as envisaged by DALL-E and Disco-Diffusion

MLCommons Research Community

MLCR is **MLCommons Research** led by Janapa Reddi (Harvard) and Pekhimenko (Toronto)
Science WG is part of MLCR and is led by Fox, Hey and Thiyagalingam



Future MLCR WG's

- AR/VR (Metaverse)
- Robustness
- Autonomous Vehicles

- 62 Companies
- 11 Universities
- 6 DOE (mainly Office of Science) Labs
- ~15 WG meetings per week and
- Quarterly community meetings

- >2000 MLCommons members
- >50 FTE's from Industry
- 125 members MLCR mainly MLSys
- 27 Science WG attendees

MLPerf 2018 became **MLCommons** December 2020

"**Accelerating machine learning innovation to benefit everyone**"

- "Grow ML markets and make the world a better place";
- "Get everyone involved";
- "Act through collaborative engineering";
- "Make fast but consensus-supported decisions," and
- "Build a community that people want to be part of."

Data/cyberinfrastructure best practices

- **MLCommons** work is organized around **3 Pillars**:
  - **Best Practice**
  - **Datasets**
  - **“Benchmarks”** -- the model+dataset artifacts that are open realization of mission and principles
- **Benchmarks** measure performance OR Science & CS “Discovery”
- **Models** are critical part of infrastructure
- **DLPerf** rather than **MLPerf**: > 90% activity is **Deep learning**
- MLCommons **Best Practice WG** covers Cyberinfrastructure for managing benchmarks -- Organizing/using GitHub, logging metadata and Container infrastructure **MLCube**
- **SABATH** from UTK supports **FAIR principles** for benchmarks
- Using benchmarks implies Big Data environments such as
  - **Enterprise:** Databricks/Spark on Clouds
  - **Industry:** Kubernetes, Google Pathways (for complex models), NVIDIA RAPIDS (GPU)
  - **HPC:** Parsl, Radical Pilot, Jupyter Notebooks …..
  - **Apache:** Arrow and Parquet optimize vector performance (RAPIDS, Cylon)

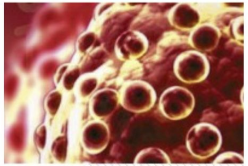
Impact: HDR and larger cyberinfrastructure

- Today **Industry leads** in many aspects of Big Data technology and we should collaborate with them
- Minimize differences between **supercomputers and clouds** as probably best technology aimed at clouds and we certainly want to use both
- In MLPerf, largest systems (4096 TPU-v4-fastest or 4216 A100's on a single training) run on **clouds and not HPC** Systems
- **Models** are "first class" components of cyberinfrastructure
  - Software 2.0 describes next generation programming as training models with Big Data
  - 30 years ago, algorithms (for solving partial differential equation and particle simulations) were at forefront of parallel computing
- **Deep Learning** is dominant

Current data/cyberinfrastructure needs

- Please join and contribute to **MLCommons Research** and **Science WG**
  - https://github.com/mlcommons/science
- Improve and Add to 4 Science Benchmarks to fill out patterns of **Foundation Models** on next slide
- Also from **MLCommons HPC** working group
  - **CosmoFlow** (3D CNN regression on cosmology simulations)
  - **DeepCAM** (2D CNN segmentation, identifying weather phenomena)
  - **OpenCatalyst** (Graph NN predicting energy and forces in atomic catalyst systems)

Fully Connected Network



candle-uno

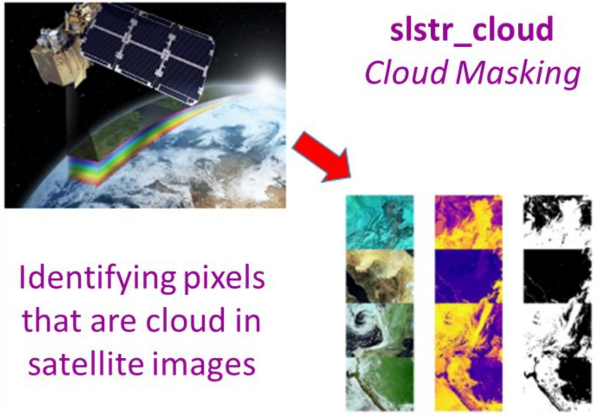
Deep learning-based precision medicine for cancer – drug response measurement

Uses Resnet-50 CNN

Uses U-Net CNN

slstr_cloud

Cloud Masking



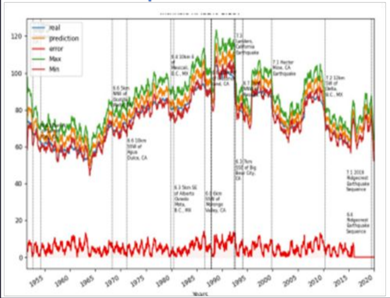
Identifying pixels that are cloud in satellite images

Uses Fusion Transformer + RNN

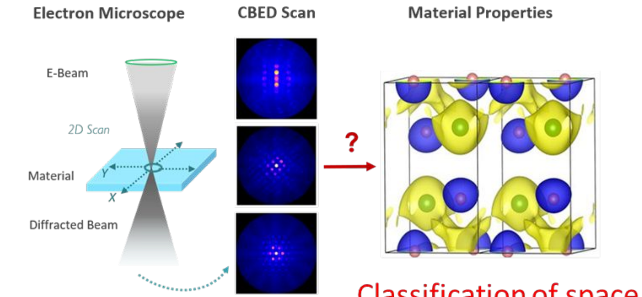
Tevelop

Timeseries Evolution Operator

Physics suggests best data processing



Focusses on extracting how a given timeseries evolves



stemdl: *ImageNet for Solid Materials*

Classification of space groups, 40GB dataset with 230 spacegroups

Current data/cyberinfrastructure needs

- **I just look at models:** Design & Develop **Foundation AI Models** with appropriate high-performance, easy-to-use cyberinfrastructure in end-to-end systems, including data engineering, parallelism, storage and data movement, security, and the user interface
- **Foundation Models** for each of the ~8 different **patterns of Science Data Analytics**
- **Overall Reasoner** based on **reinforcement learning** and large language models to learn the world's knowledge and control experiments, networks, computers;
- **Image-based systems** for astronomy, pathology, microscopy, and light scattering;
- **Graph-based systems** such as in social media and traffic studies; represent molecular and other structure;
- **Dense systems t**o map structure to properties as in drug discovery;
- **Time series and sequence (Recurrent, Transformer)** models as in language, earth, and environmental science;
- **GAN/Diffusion** models to generate scenarios as in datasets to test experimental system.
- **Surrogate** models (deep learning models trained on results of simulations) have distinctive issues where there is no current consensus
- **All Network types** can be mixed together as in text to image system DALL-E

4 out of 7 last slide