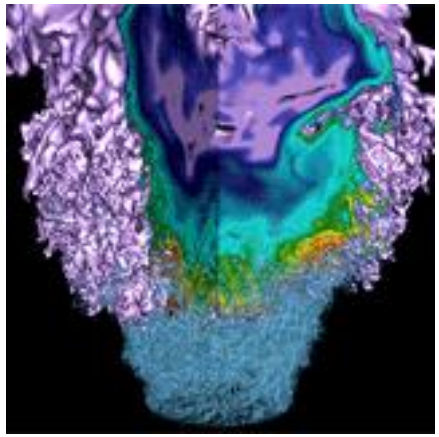
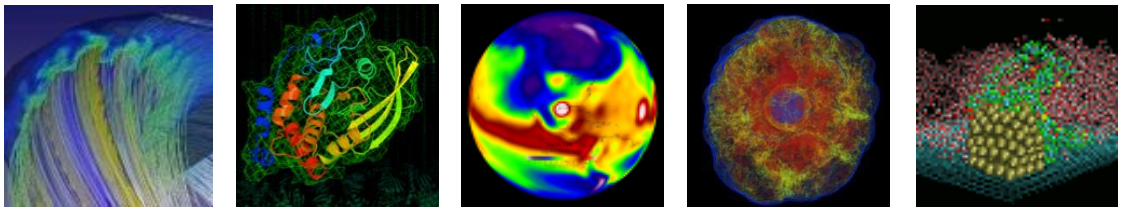
Big Data and machine learning at NERSC





Debbie Bard
Group Leader,
Data Science Engagement

Mission:
"Enable Data-Intensive Science at Scale"

Internal Goals

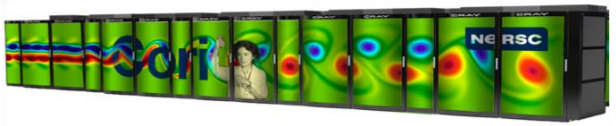
- **Provide world-class, production quality software services for all major Data capabilities**:
  - Analytics, Management, Workflows, Transfer, Access, Visualization
- **Pioneer evaluation, research and deployment of Big Data technologies**
  - Focusing on productivity and performance
- **Engage with stakeholders to enable scientific discovery in a data-driven world**
  - Users, Vendors, CS staff, Researchers (Industry, Academia)

Production Data Stack

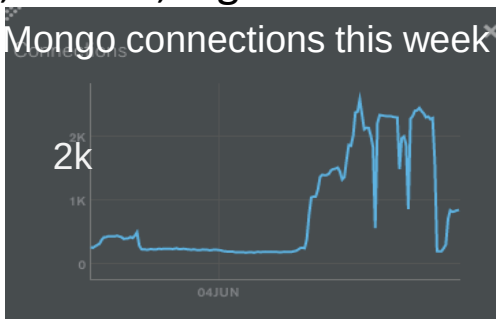
Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, php, jupyter, python django, newt, !m
Workflows	FireWorks, swift, TaskFarmer
Data Management	HDF, netCDF, ROOT Data Analysis Framework, SciDB, mongoDB, MySQL, PostgreSQL
Data Analytics	python, R, IP[y]: IPython, Spark, scikit learn, TensorFlow, theano, Caffe
Data Visualization	VisIt, ParaView

Data-friendly HPC...

Now : Cori (P1) Data Features (with CSG etc.)



- **NVRAM Burst Buffer**
- **High-performance Lustre filesystem: Distributed metadata etc.**
- **External connectivity from compute (SDN)**
- **Workflow/Additional services on logins:** Jupyter; Grid; User-specific ...
- **Flexible queues/qos on SLURM:** realtime; interactive; shared; bigmem…
- **Virtualization capabilities with Shifter**



Other services (with ISG , SSG etc)

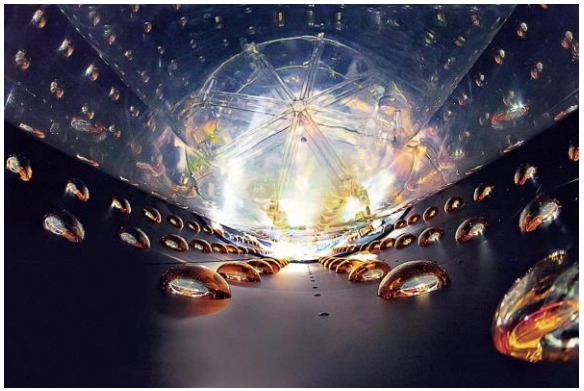
e.g Databases : MongoDB; MySQL and Postgres

The Future: Nersc-9 and beyond: 'Data Users' needs

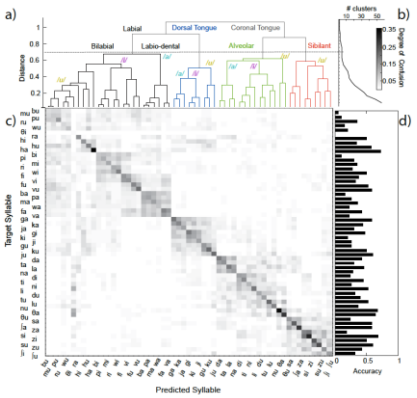
Workflows; Storage (I/O) ; External Network...

Deep Learning for Science

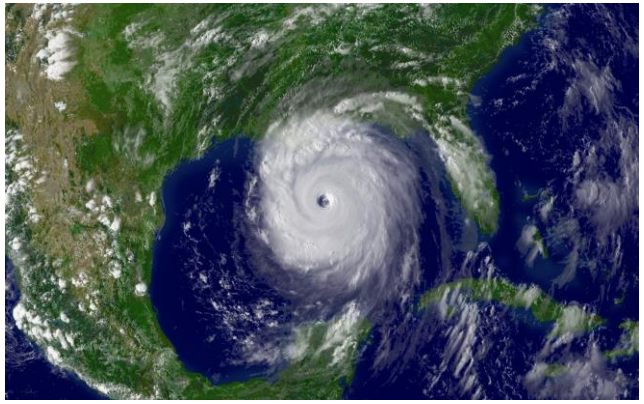
Modeling galaxy shapes



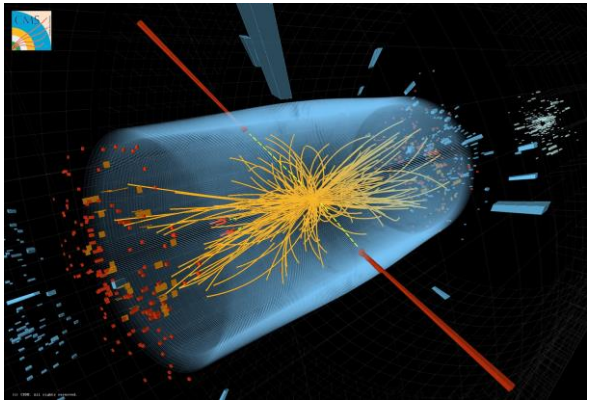
Clustering Daya Bay events



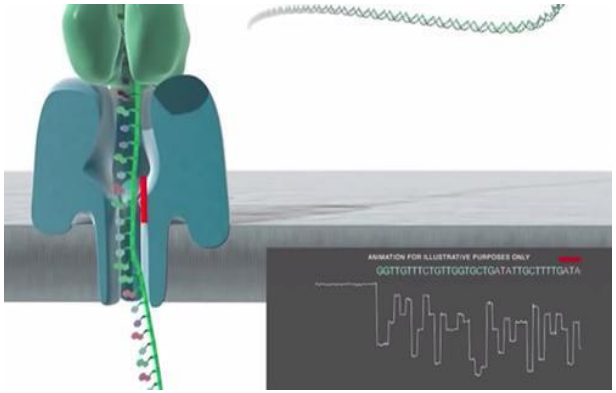
Decoding speech from ECoG



Detecting extreme weather



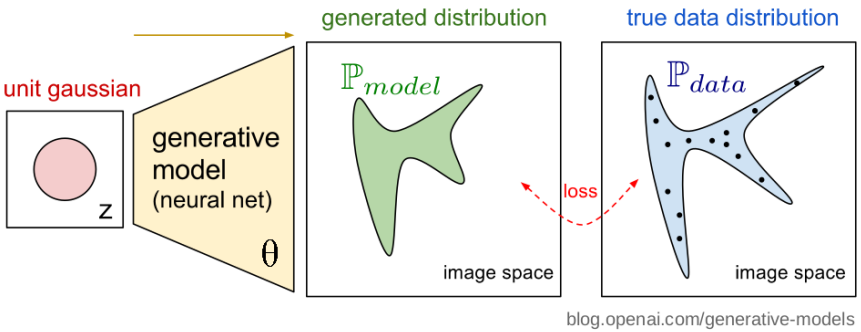
Classifying LHC events

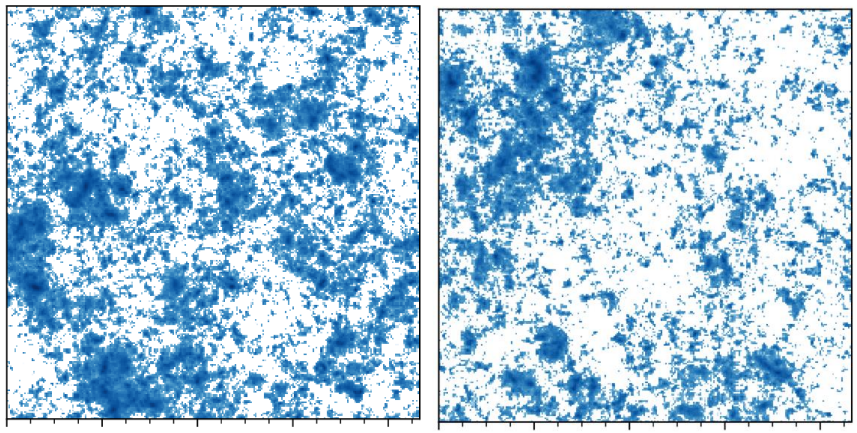


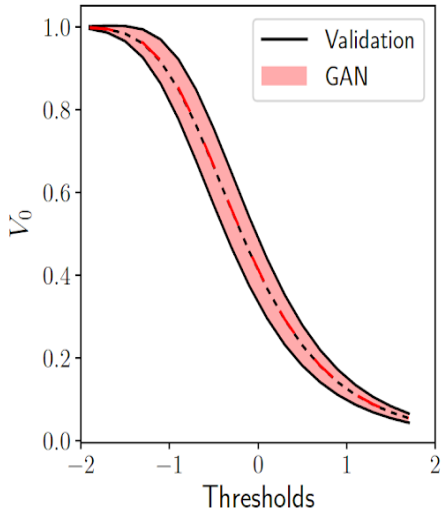
Oxford Nanopore sequencing

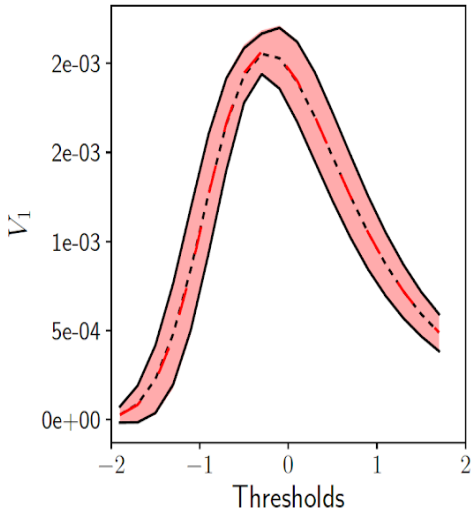
Generative Adversarial Networks

Creating Virtual Universes Using Generative Adversarial Networks. Mustafa, Debbie, Wahid, Rami, Zarija arXiv:1706.02390





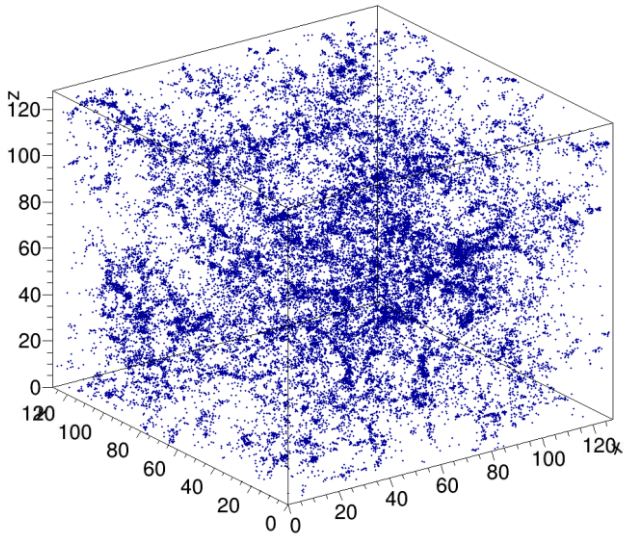


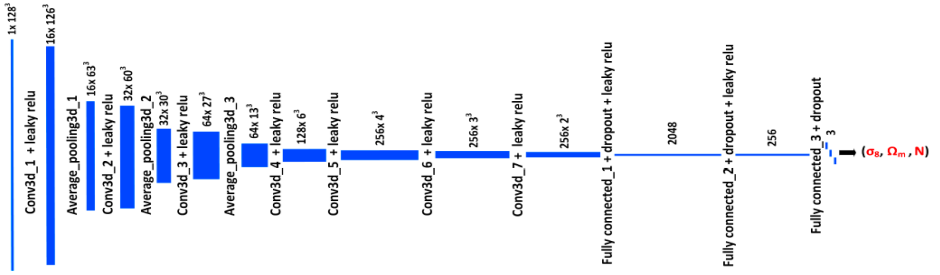


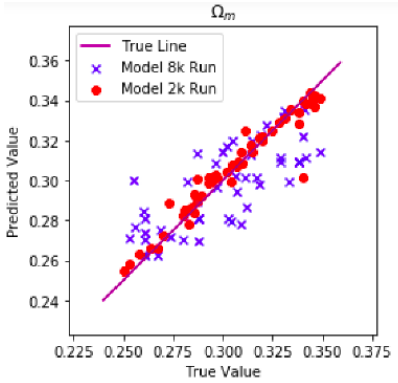
Towards a cheap emulator:

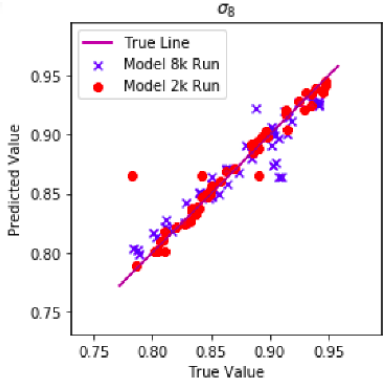
- GAN-generated maps exhibit same gaussian AND non-gaussian structures as those produced by computationally-expensive full simulations.

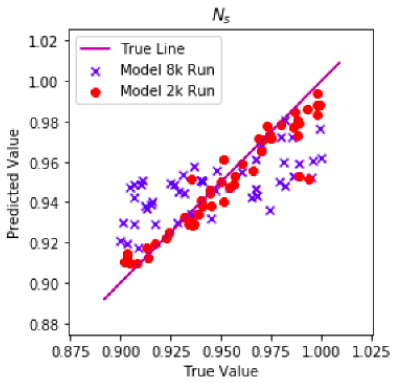
3D volumes: Machine Learning to model the universe











- **3D convolutions are computationally hard**
- **Trained network in <20min on 8192 compute nodes**
- **Estimate cosmological parameters with unprecedented accuracy with TensorFlow**

National Energy Research Scientific Computing Center