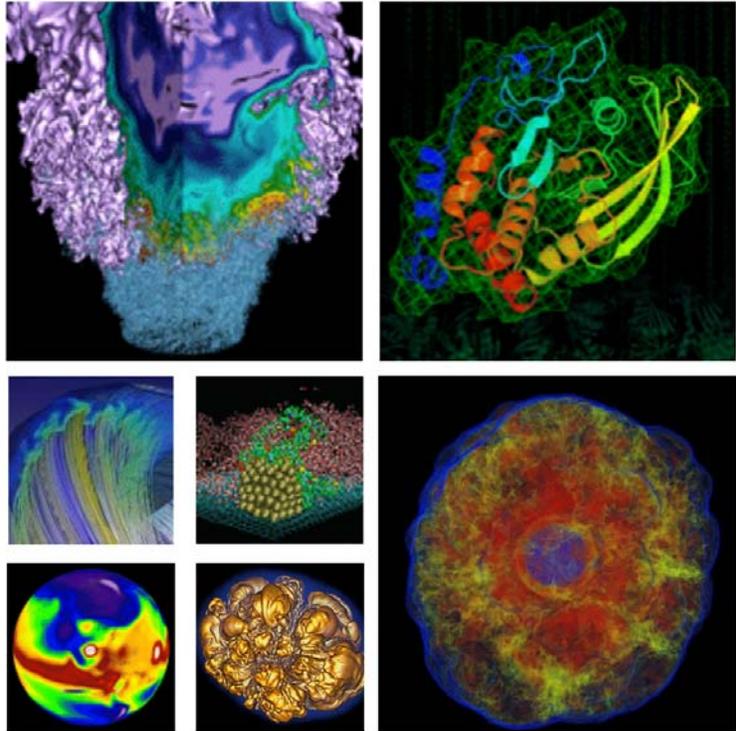


Deep Learning for Science



Prabhat
9/26/2017

ASCAC



- **Emerging User Requirements**
 - NERSC hardware and software strategy
- **Deep Learning in Industry**
- **Deep Learning in Science**
 - Success Stories
 - Challenges
- **The Road Ahead**

- **Emerging User Requirements**
 - NERSC hardware and software strategy
- **Deep Learning in Industry**
- **Deep Learning in Science**
 - Success Stories
 - Challenges
- **The Road Ahead**

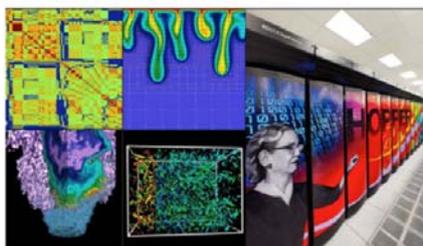
NERSC: the Mission HPC Facility for DOE Office of Science Research



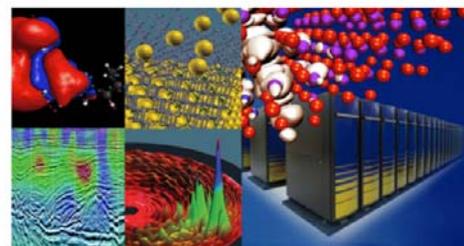
Largest funder of physical science research in the U.S.



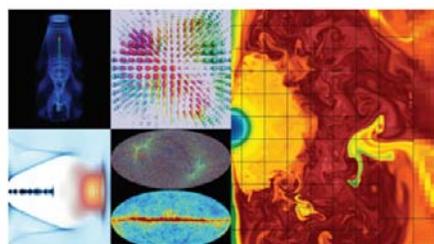
Bio Energy, Environment



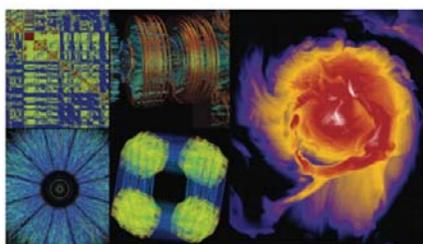
Computing



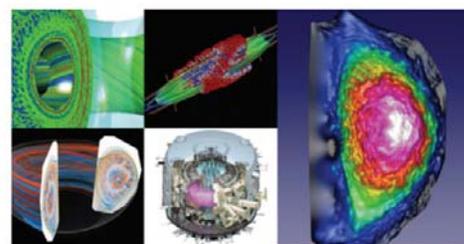
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

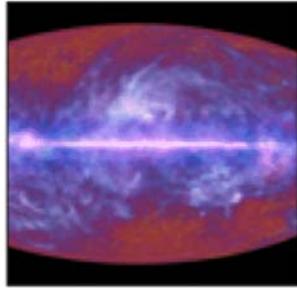
6,000 users, 700 projects, 700 codes, 48 states, 40 countries, universities & national labs



NERSC has a long history of working with experimental and observational science projects



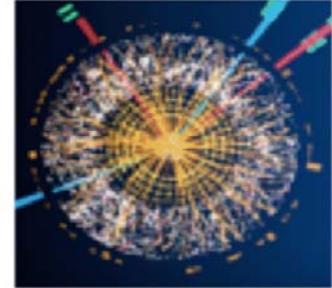
Palomar Transient
Factory
Supernova



Planck Satellite
Cosmic Microwave
Background
Radiation



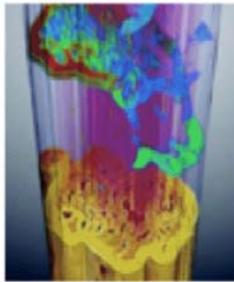
Alice
Large Hadron Collider



Atlas
Large Hadron Collider



Dayabay
Neutrinos



ALS
Light Source



LCLS
Light Source



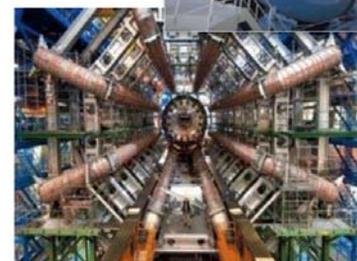
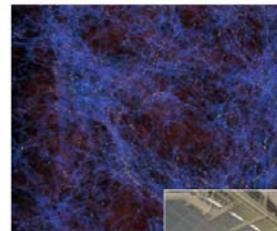
Joint Genome Institute
Bioinformatics



What's different?



- **Proliferation of data from DOE user facilities**
- **Scientific workflows have become more complex**
 - Streaming data to HPC facilities
 - Real-time/Interactive access
 - Rich 'Data' stack
- **Important scientific problems are requiring both simulation and data analytics**
 - Advanced Machine Learning and Statistics methods + tools required



DOE Exascale Requirements Reviews

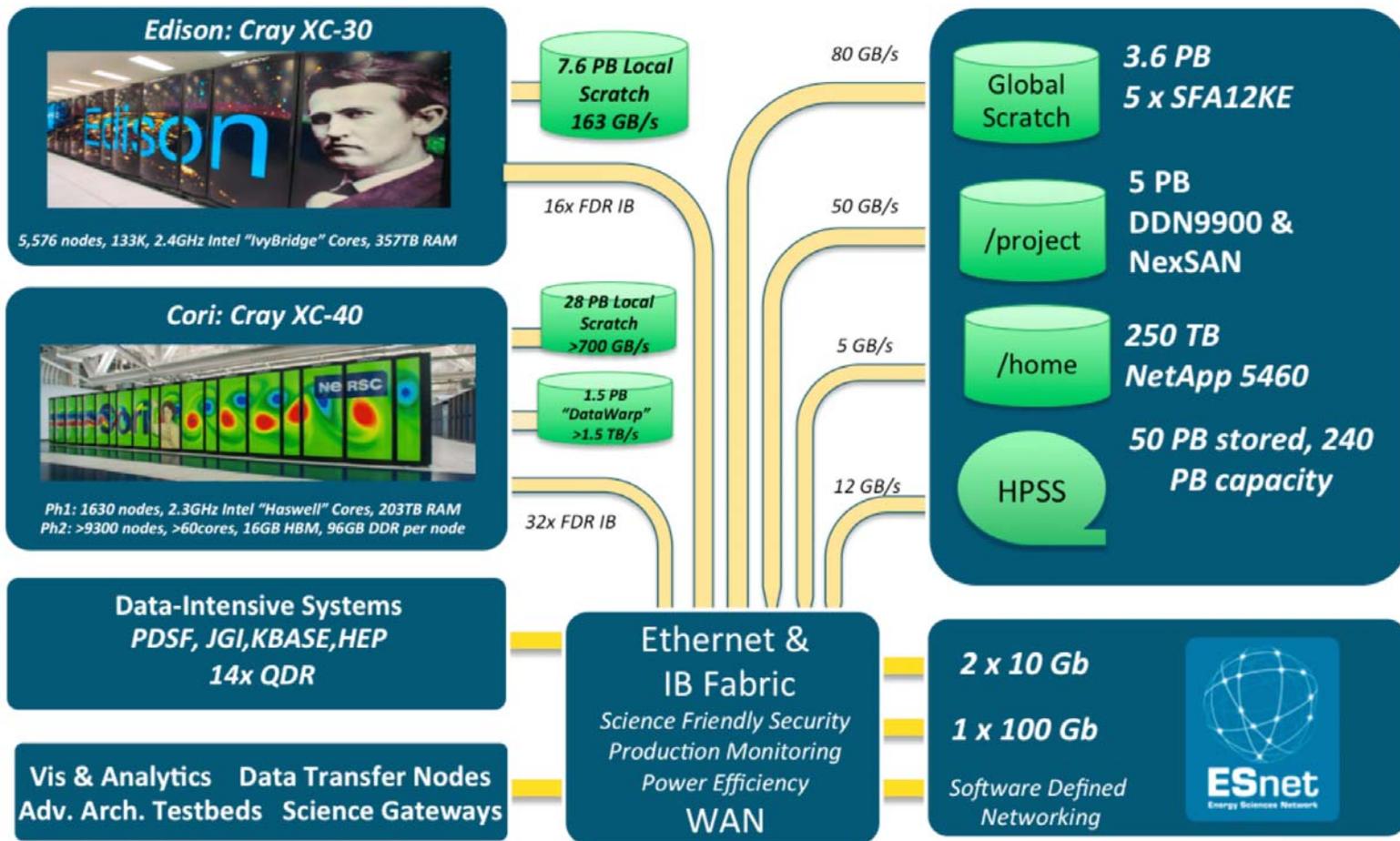


- **Broad input from DOE experimental facilities**
- **Focused on the exascale 'ecosystem', beyond compute**
- **Machine Learning called out as an important cross-cut theme**



	HEP			BER		BES		NP	FES
	Astronomy	Cosmology	Particle Physics	Climate	Genomics	Light Sources	Materials	Heavy Ion Colliders	Plasma Physics
Classification	X		X	X	X	X	X	X	X
Regression		X			X	X	X	X	X
Clustering		X	X	X	X	X	X	X	X
Dimensionality Reduction				X				X	
Surrogate Models	X	X	X				X	X	X
Design of Experiments		X		X			X		X
Feature Learning	X	X	X	X	X	X	X	X	X
Anomaly Detection	X		X	X		X		X	

NERSC Platforms

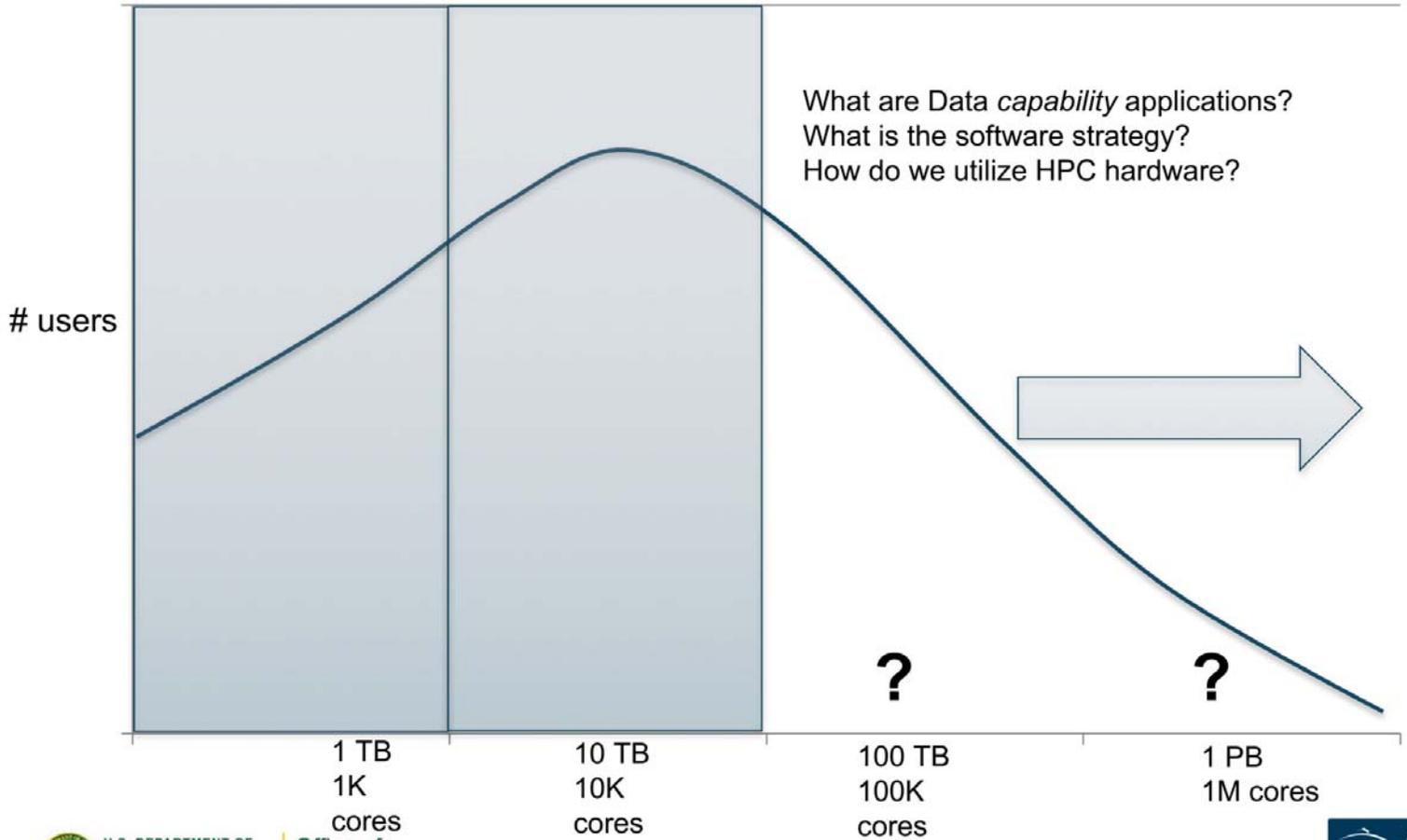


NERSC Big Data Stack

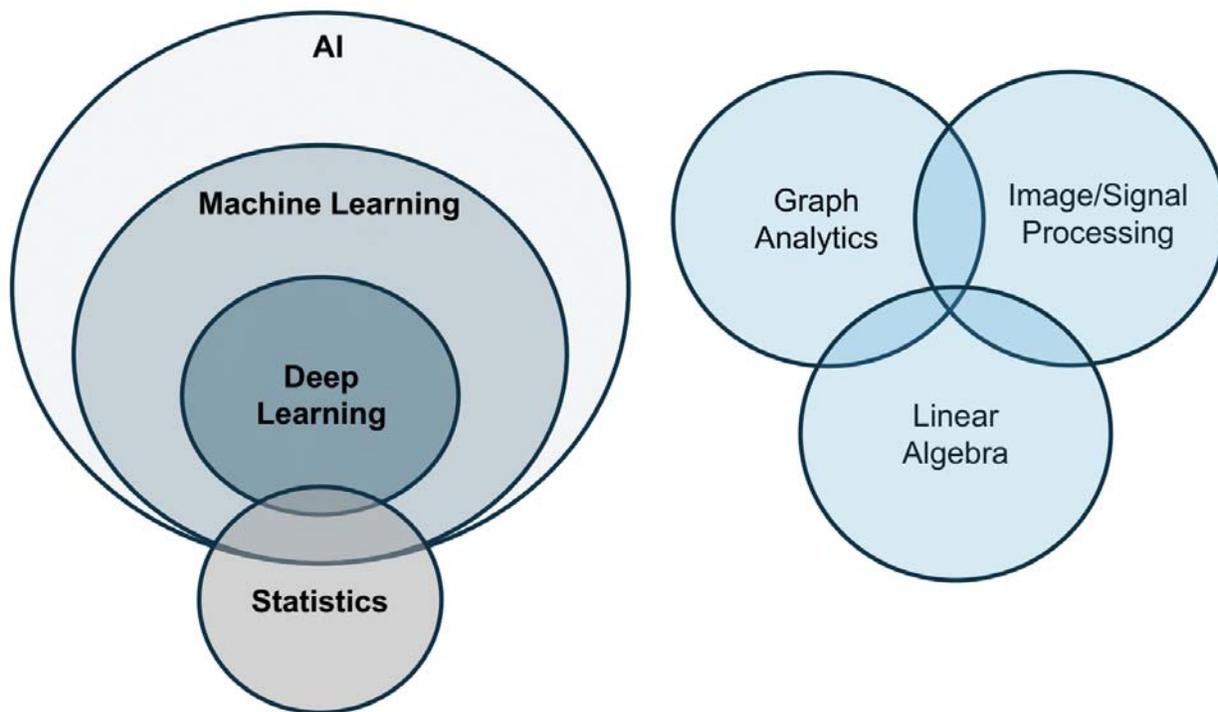


Capabilities	Technologies
Data Transfer + Access	     
Workflows	 
Data Management	      
Data Analytics	      
Data Visualization	 

Big Data Center



Data Analytics Methods



Outline



- **Emerging User Requirements**
 - NERSC hardware and software strategy
- **Deep Learning in Industry**
- **Deep Learning in Science**
 - Success Stories
 - Challenges
- **The Road Ahead**

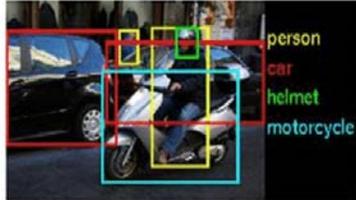
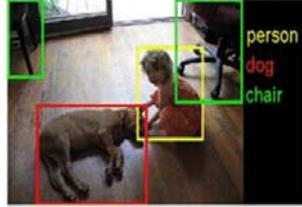
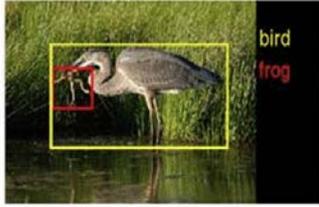




IBM and MIT to pursue joint research in artificial intelligence, establish new MIT-IBM Watson AI Lab

IBM plans to make a 10-year, \$240 million investment in new lab with MIT to advance AI hardware, software, and algorithms.

Breakthrough Results



Can Deep Learning work for Science?

Similarities

Tasks

- Pattern Classification
- Regression
- Clustering
- Feature Learning

Differences

Unique attributes of Scientific Data

- Multi-channel / Multi-variate
- Double precision floating point
- Noise and Artefacts
- Statistics are likely different

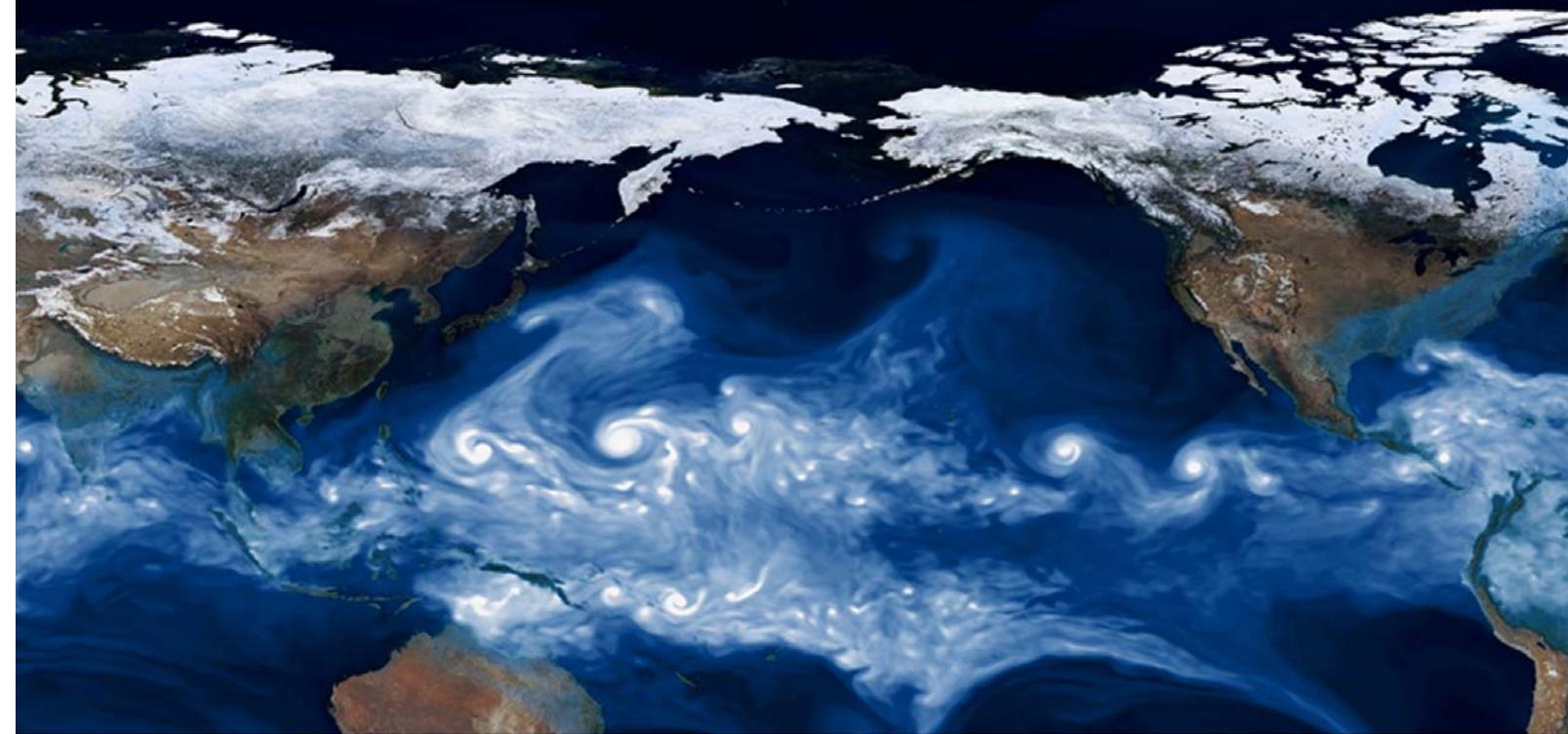
Outline



- **Emerging User Requirements**
 - NERSC hardware and software strategy
- **Deep Learning in Industry**
- **Deep Learning in Science**
 - Success Stories
 - Challenges
- **The Road Ahead**



1 Characterizing Extreme Weather in a Changing Climate



Climate Science Tasks



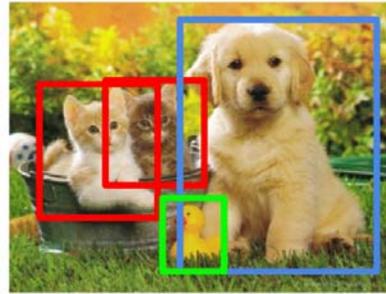
Classification



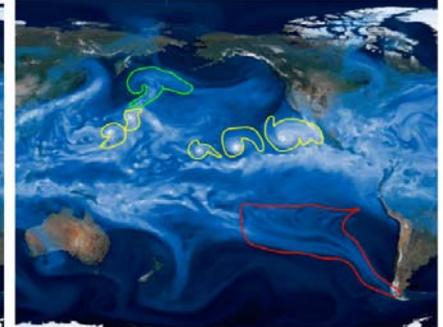
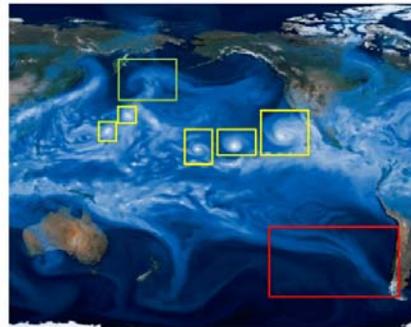
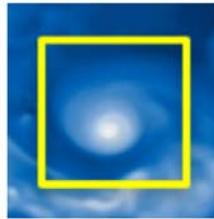
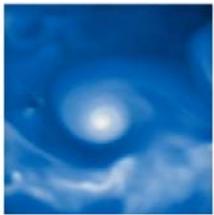
Classification + Localization



Object Detection



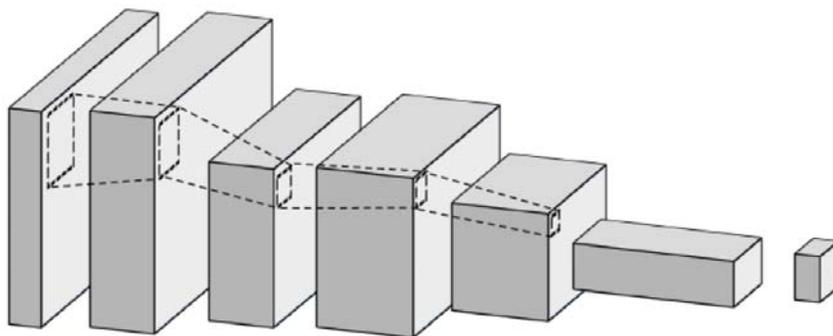
Instance Segmentation



Supervised Classification Accuracy



	Logistic Regression	K-Nearest Neighbor	Support Vector Machine	Random Forest	ConvNet
	Test	Test	Test	Test	Test
Tropical Cyclone	95.85	97.85	95.85	99.4	99.1
Atmospheric Rivers	82.65	81.7	83.0	88.4	90.0
Weather Fronts	89.8	76.45	90.2	87.5	89.4



Input

Convolution

Pooling

Convolution

Pooling

Fully connect

Class score

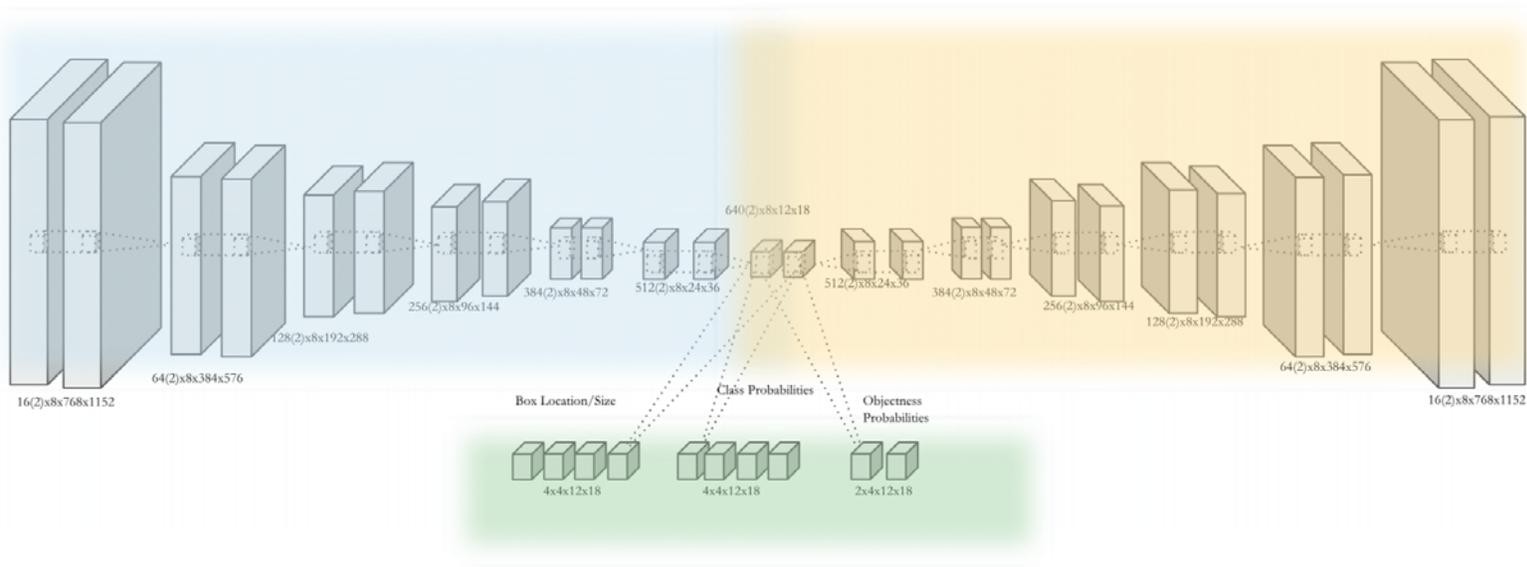


Semi-Supervised Convolutional Architecture



Encoder

Decoder



Classification + Bounding Box Regression



Contributors: Evan Racah, Chris Pal, Chris Beckham, Tegan Maharaj

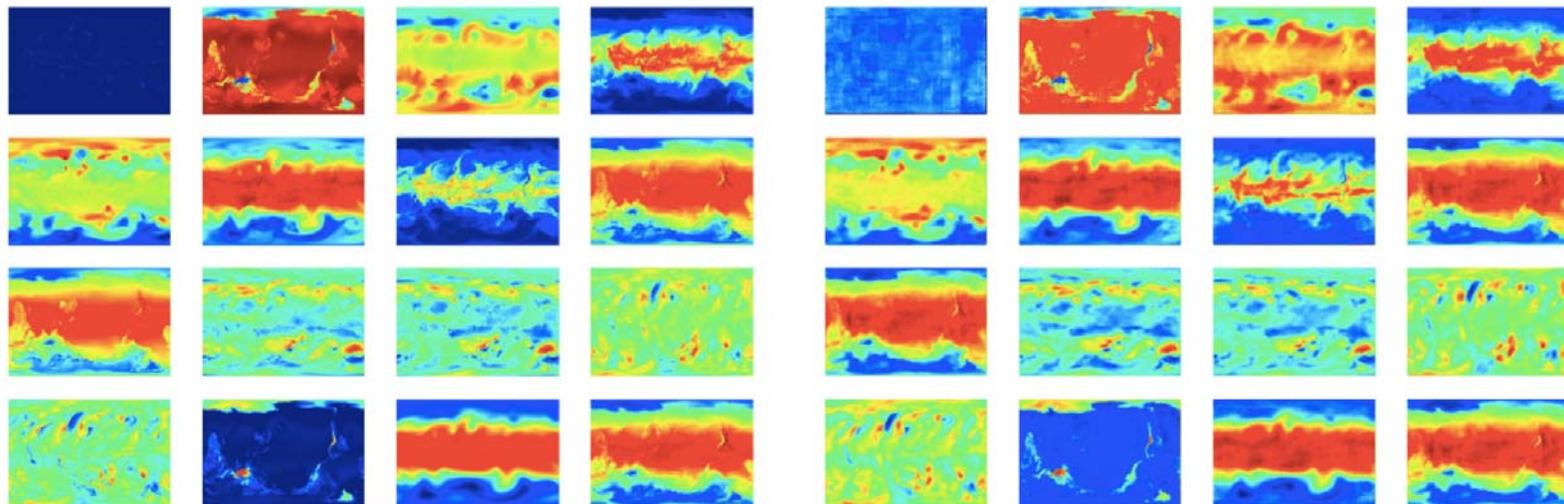


Reconstruction Results

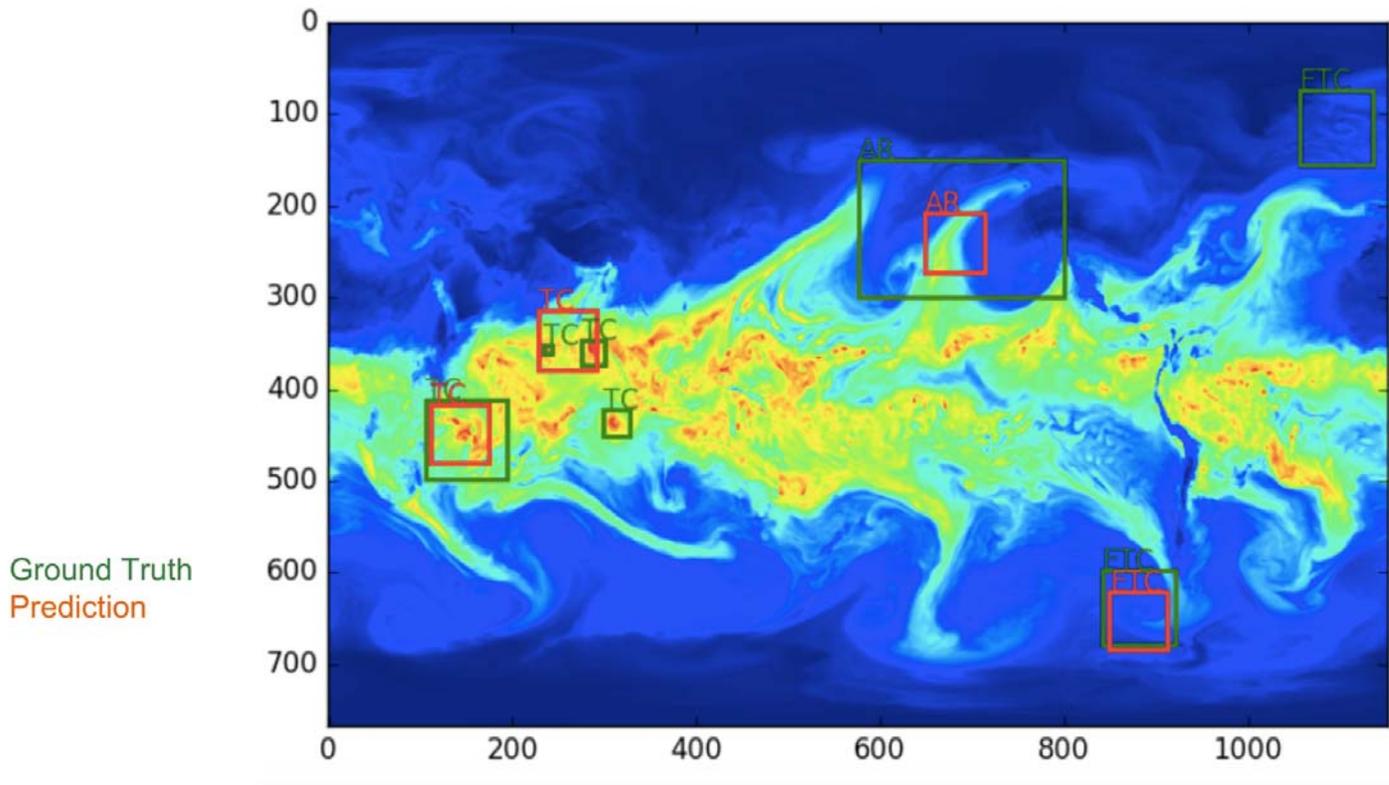


original

reconstruction



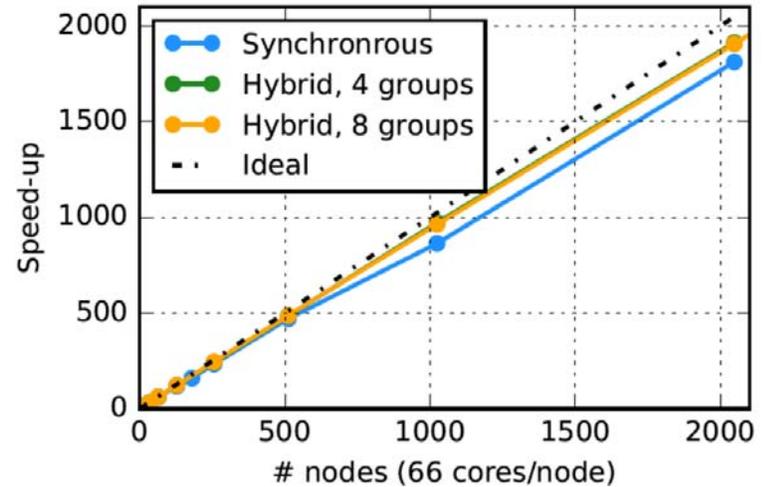
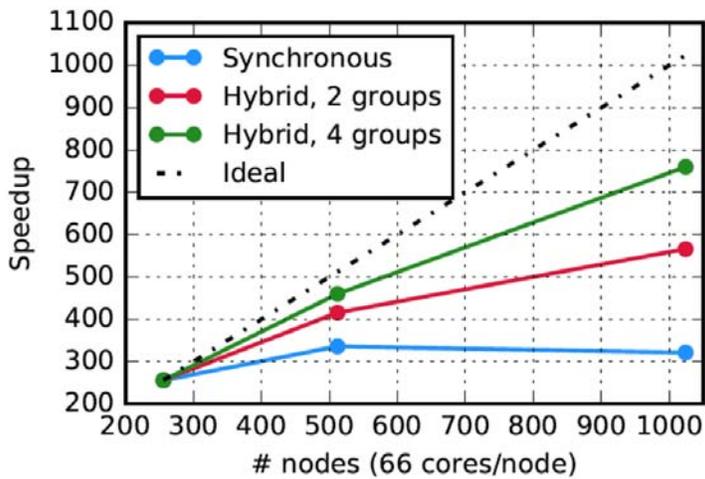
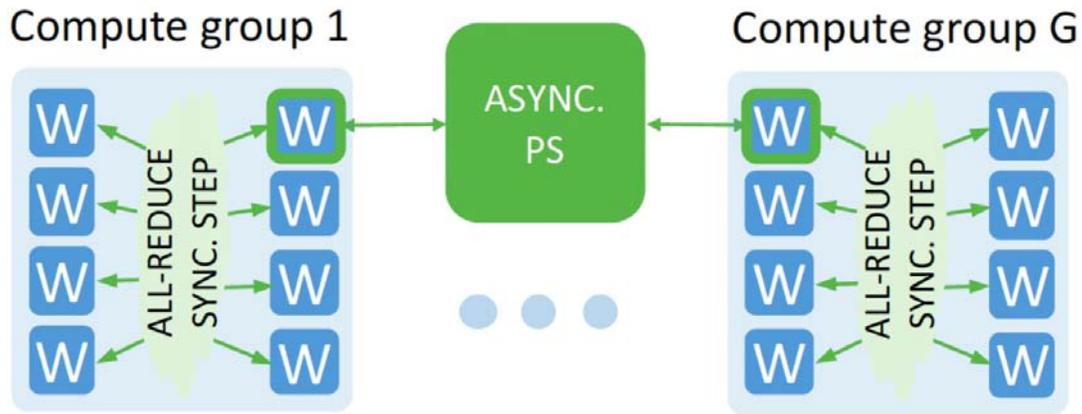
Classification + Regression Results



Contributors: Thorsten Kurth, Jian Yang, Ioannis Mitliagkas, Chris Pal, Nadathur Satish, Narayanan Sundaram, Amir Khosrowshahi, Michael Wehner, Bill Collins.

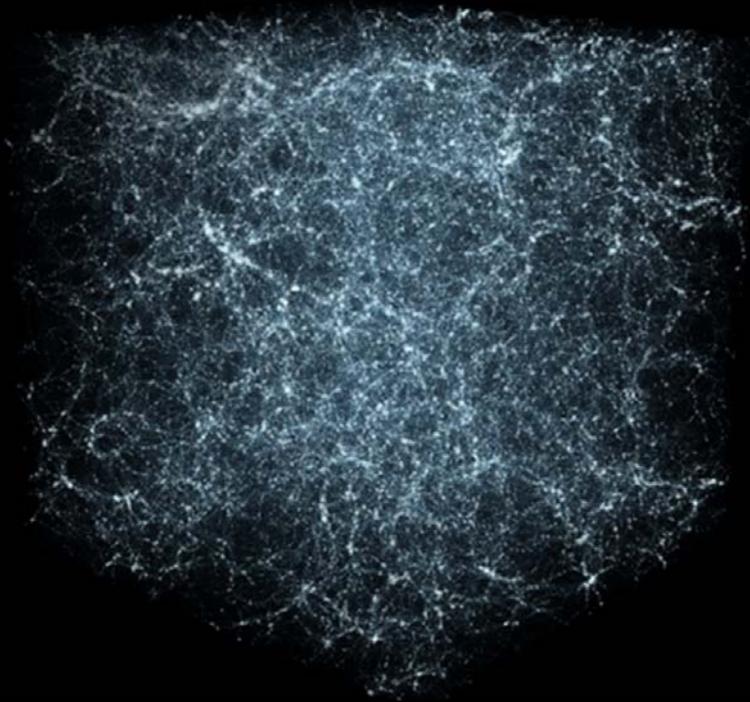


Deep Learning at 15PF (SC'17)



2

Determining the Fundamental Constants of Cosmology



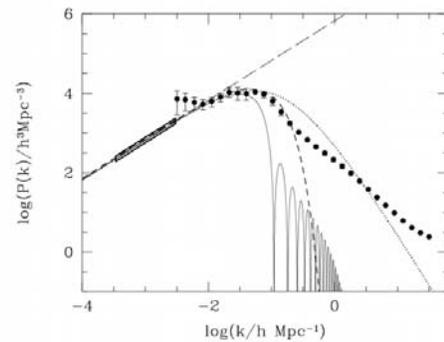
2 Determining the Fundamental Constants of Cosmology

Science challenge

- Comparison of simulation results with observations to correct for observational systematics

Analysis Results

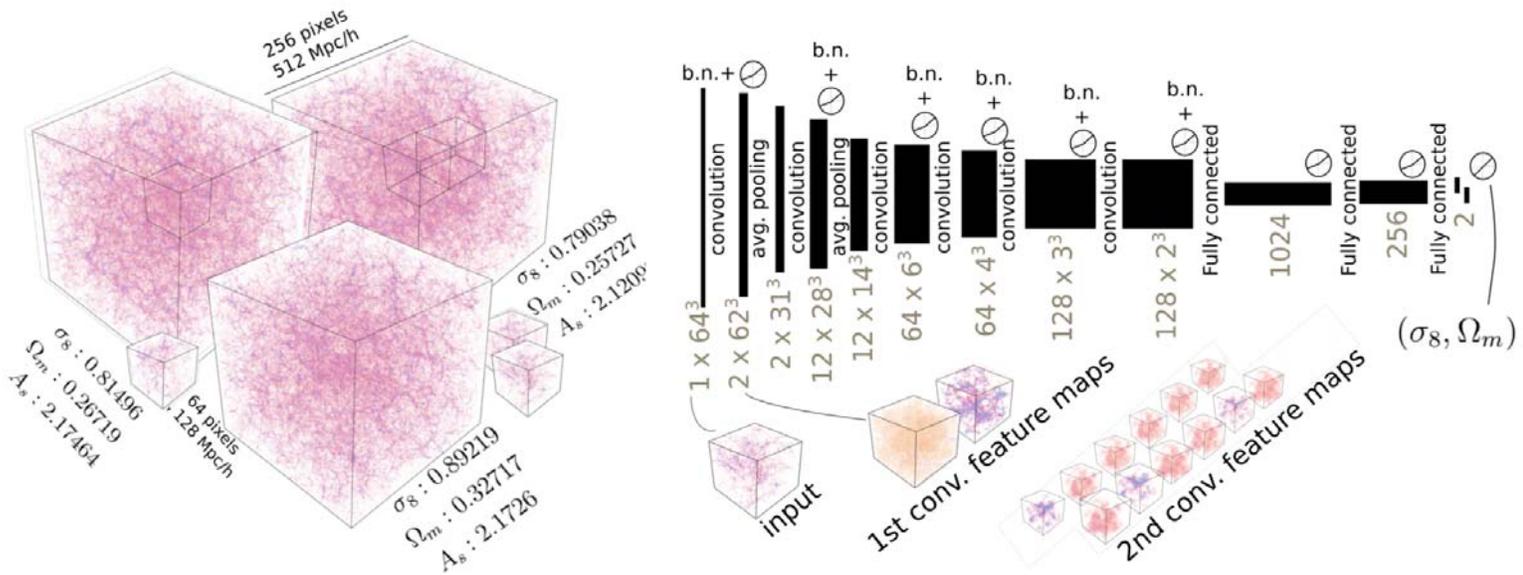
- DB-Scan: applied to 1T HACC simulation dataset; clustering computed in 20 minutes on 100K Edison cores.
- Galactos: $O(N^2)$, 3-pt correlation code processed 2B Outer Rim galaxies in 15 minutes on 650,000 Cori cores. 9.8PF performance. (SC'17)



Contributors: Debbie Bard, Brian Friesen, Mostofa Patwary, Nadathur Satish, Pradeep Dubey



3D Convolutional Network



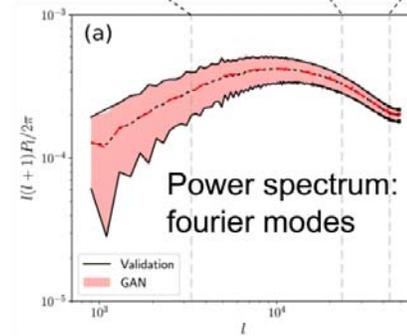
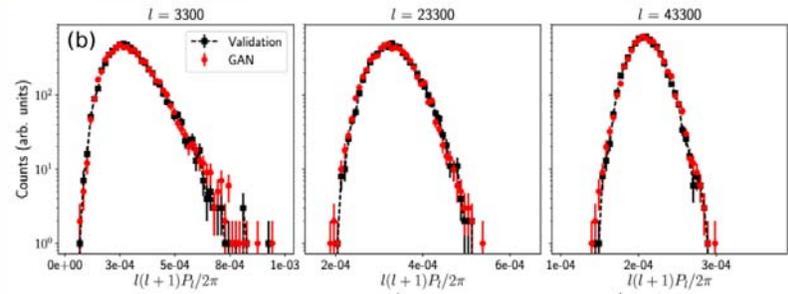
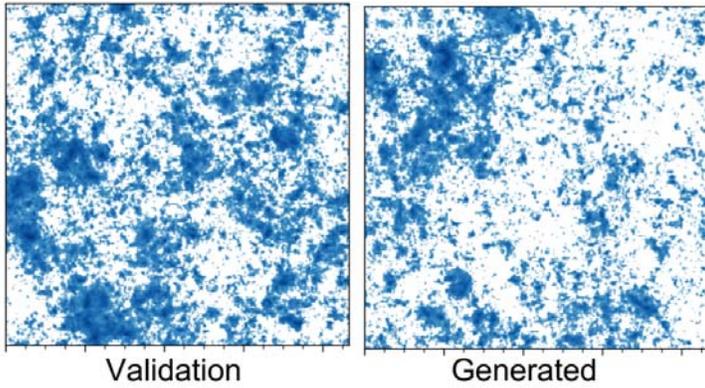
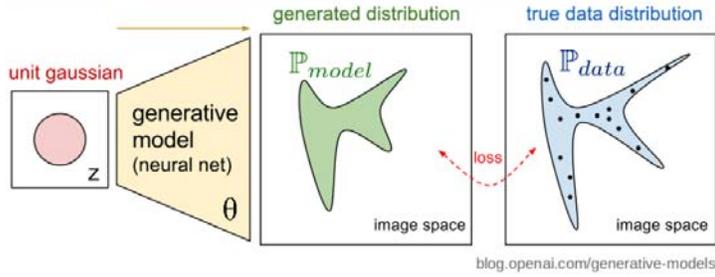
- Regress cosmological constants directly from simulation data
- Reasonable accuracy for 2 constants; currently extending framework to run on Cori



Contributors: Simak Ravanbaksh, Junier Oliva, Sebastian Froenteau, Layne Price, Shirley Ho, Jeff Schneider, Barnabas Poczos



Generative Adversarial Networks



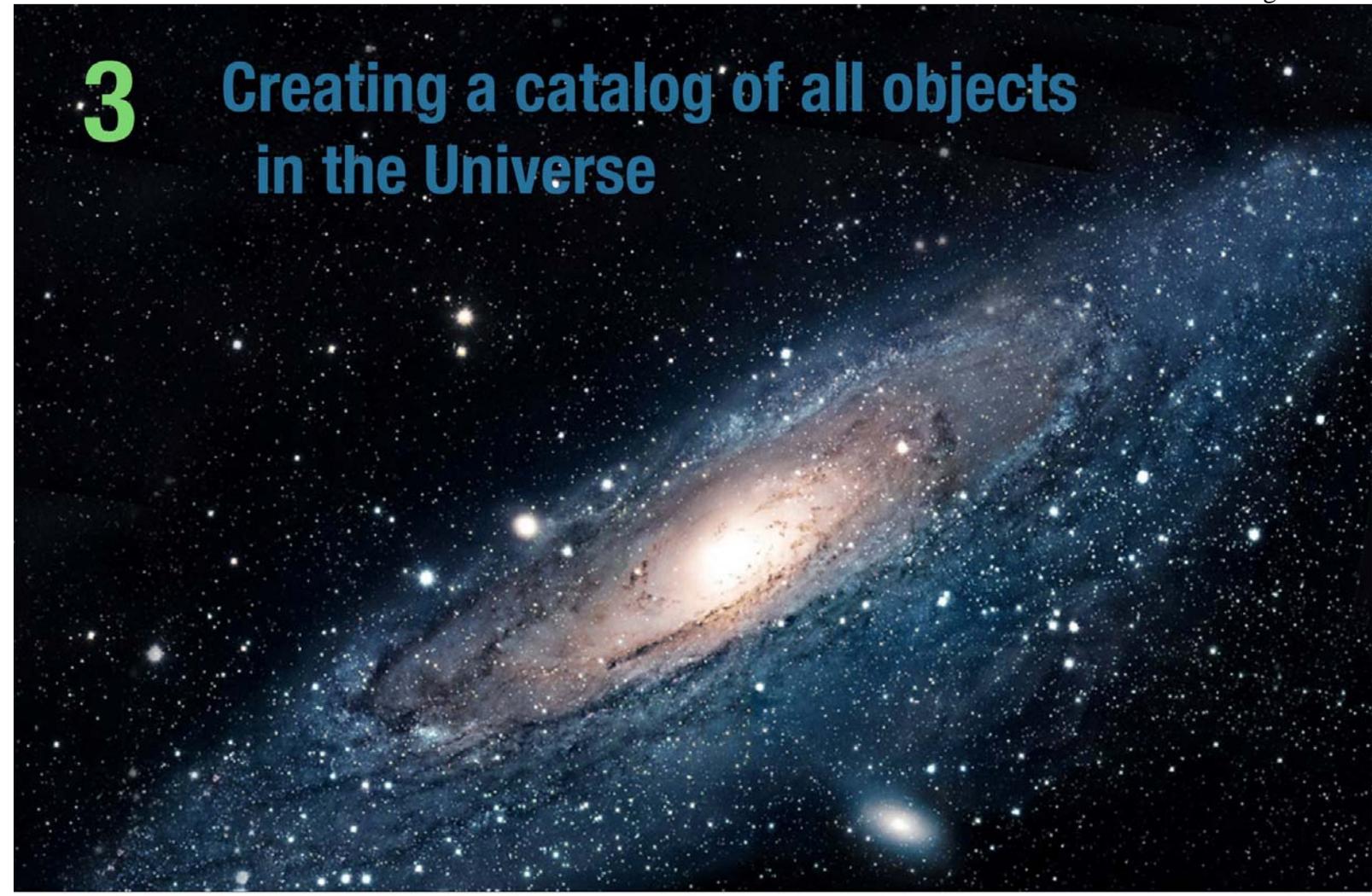
GANs generated maps exhibit the same gaussian and non-gaussian structures as full simulations.



Contributors: Mustafa Mustafa, Debbie Bard, Wahid Bhimji, Rami Al-Rfou, Zarija Lukic



3 Creating a catalog of all objects in the Universe



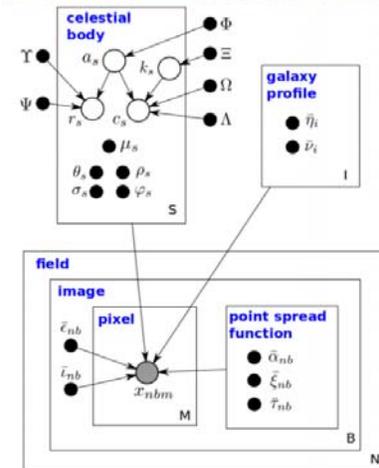
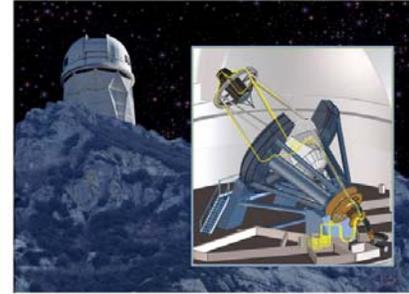
3 Celeste: A Generative Model of Astronomical Images

Astronomy challenge

- Inferring stars and galaxies from all available telescope data

Analysis Results

- Developed Graphical Model and variational inference techniques
 - Demonstrated on 8B parameters, 188M stars and galaxies
- Processed all SDSS data in 15 minutes
- First Julia application to exceed 1PF performance
 - 1.3 M threads on 650,000 KNL cores



Contributors: Jon McAuliffe, Ryan Adams, Jeff Regier, Andy Miller, Keno Fischer, Kiran Pamnany, Rollin Thomas



Celeste Galaxy Model



NGC 4753, an elliptical galaxy with interesting dust filaments.



NGC 60, a spiral galaxy with unusually distorted arms.



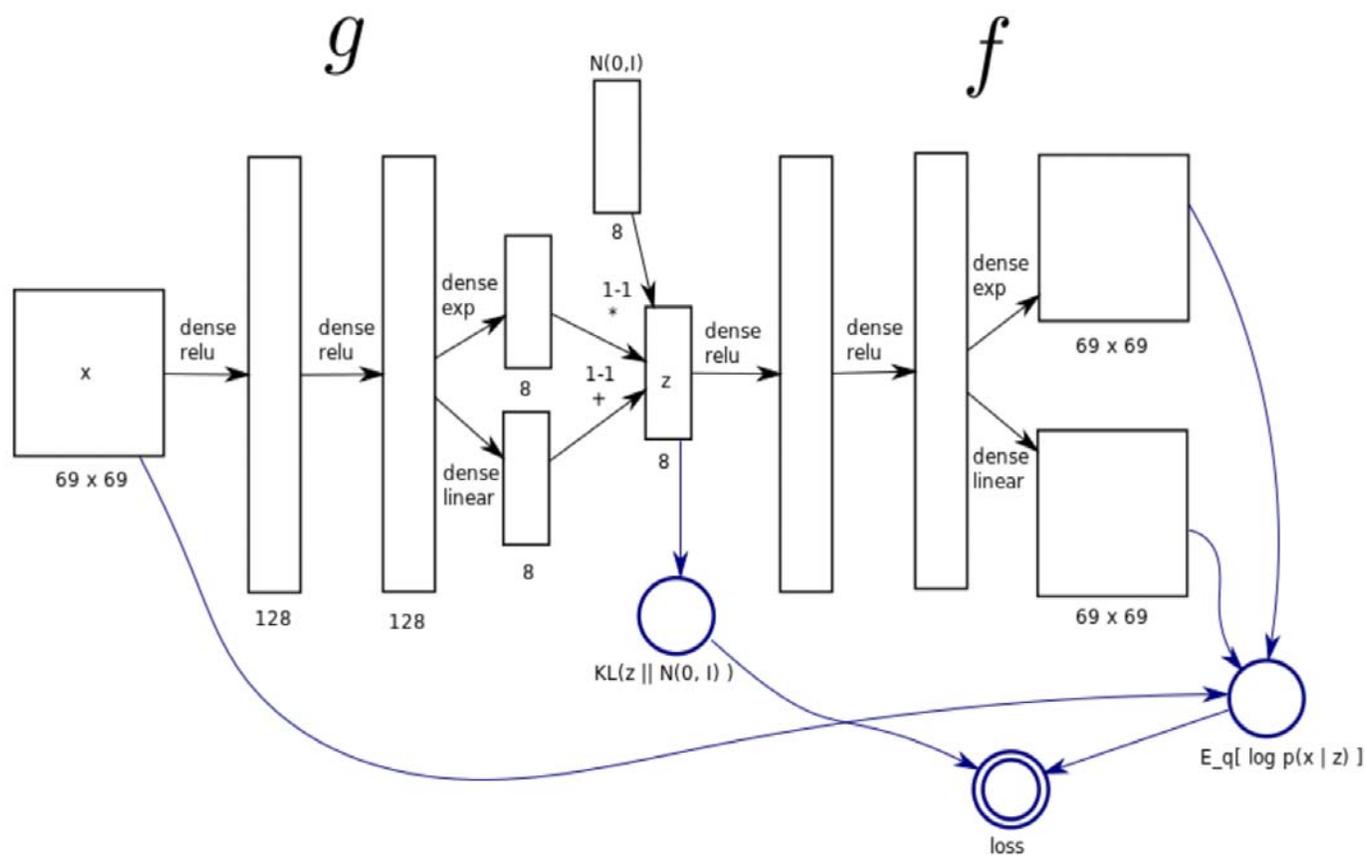
An irregular galaxy.



Contributors: Jeff Regier, Jon McAuliffe



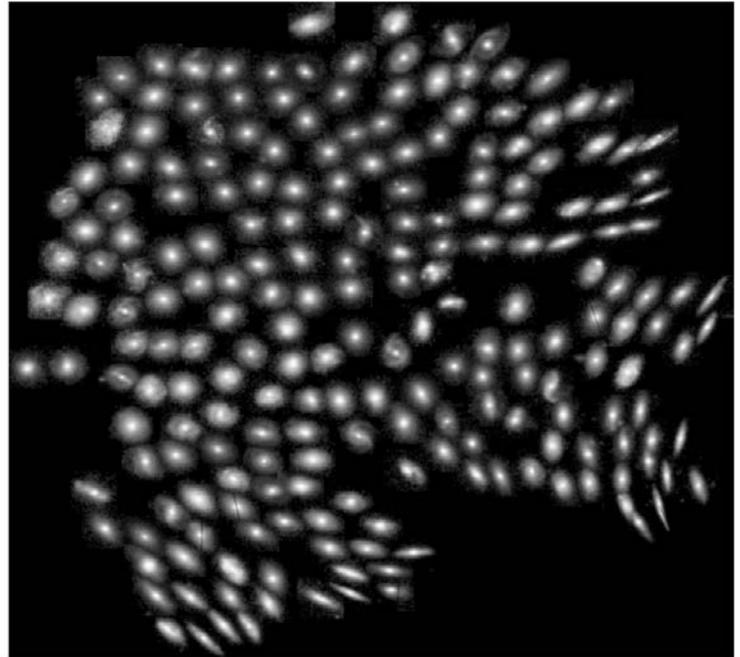
Variational Auto-Encoder



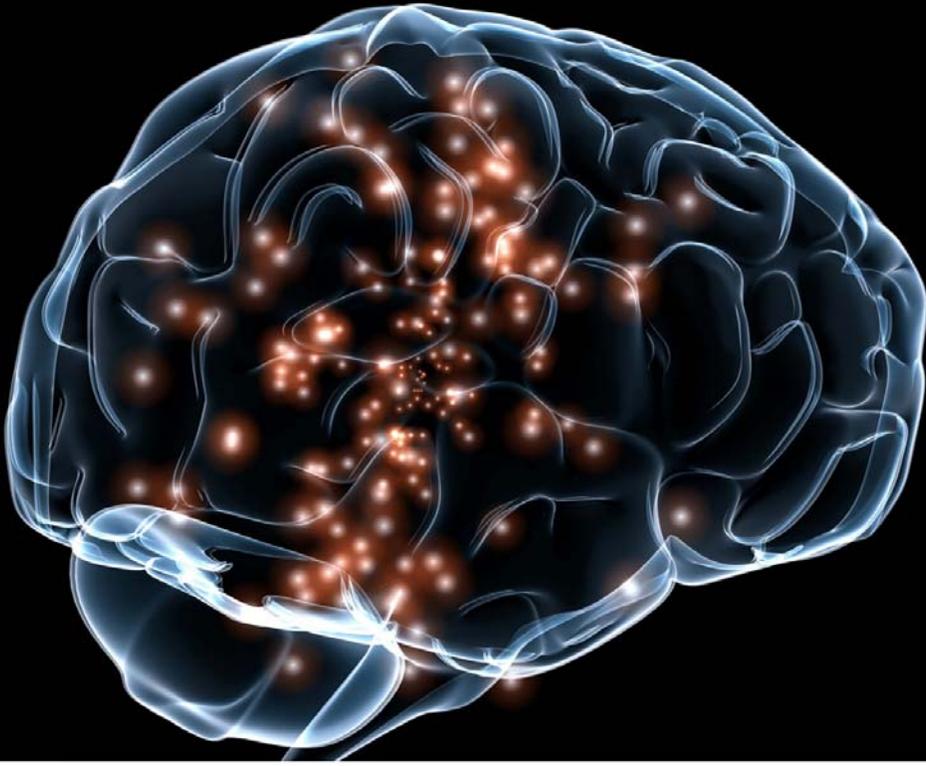
Celeste Galaxy Model Results



- The Celeste galaxy model outperformed bivariate Gaussian densities for 99.3% of galaxy images.
- Qualitative results from t-SNE indicate that the neural network learns a compact representation of galaxy shapes and orientation.



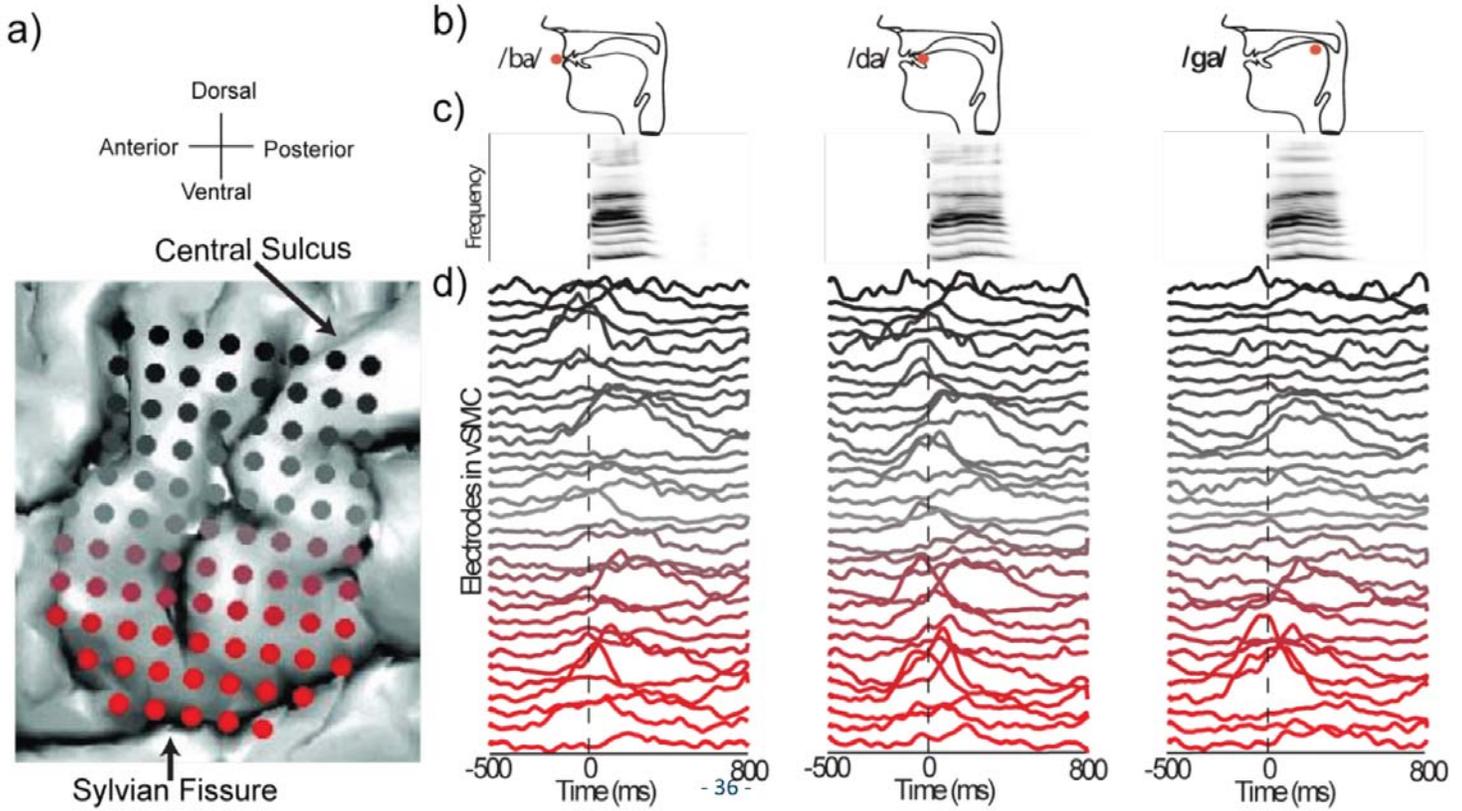
4 Understanding the Brain



Speech Prosthesis



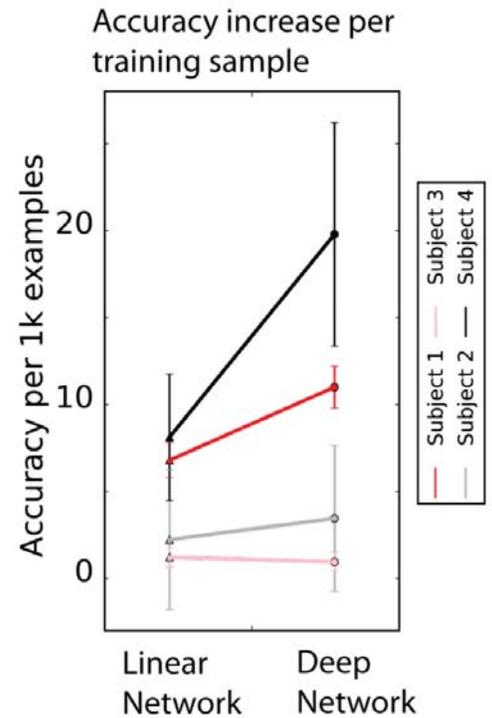
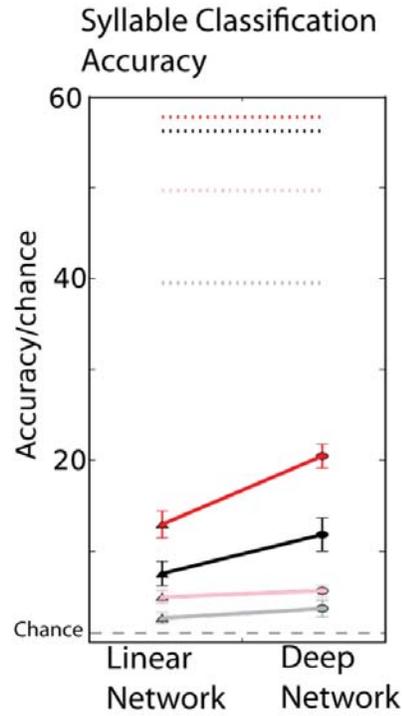
Decoding Speech



DNNs achieve best decoding performance

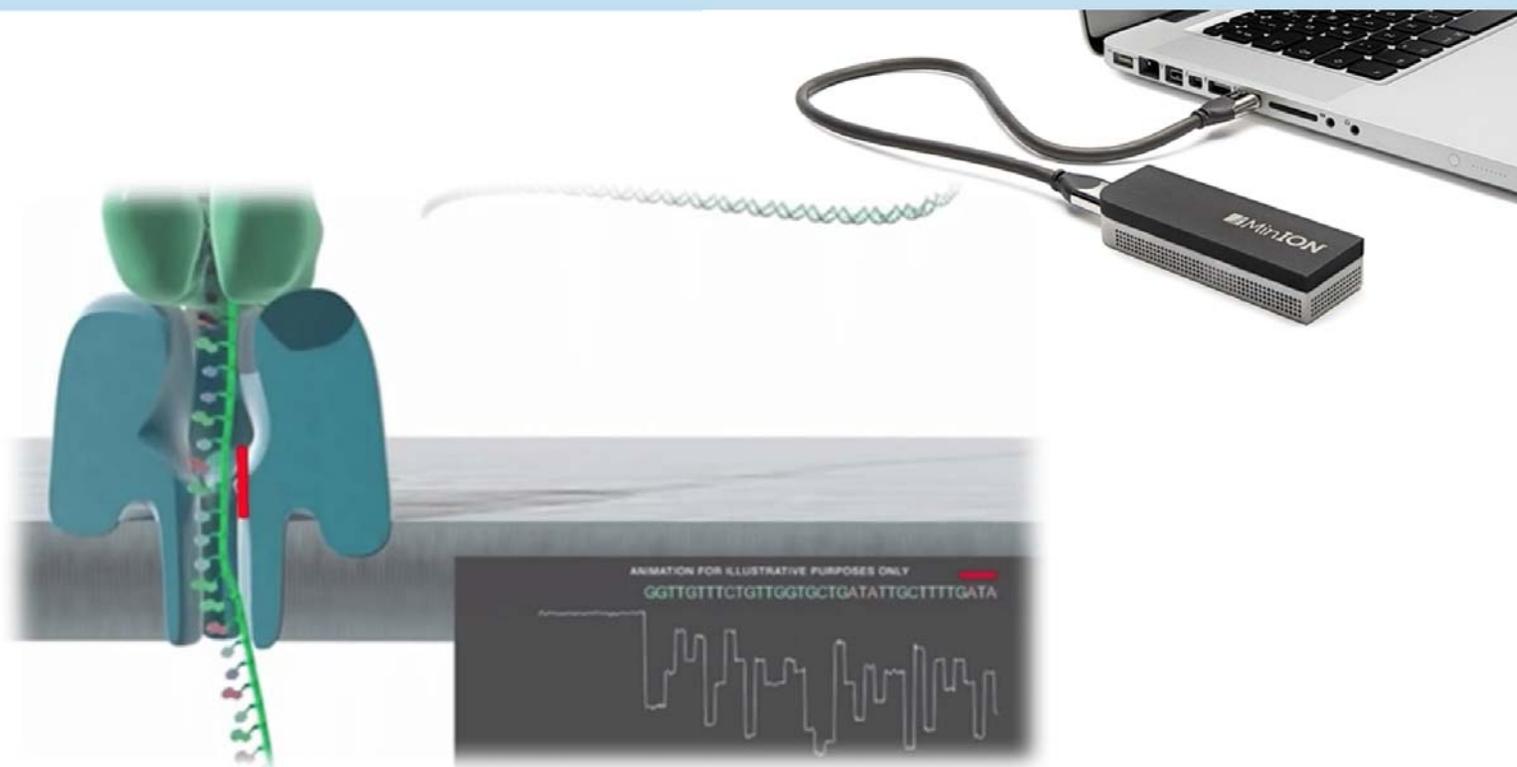


- Classify spoken syllable from spatiotemporal patterns of human neural recordings
- Fully Connected, Feed-forward Network
- All hyper-parameters optimized with Spearmin
- L_2 regularization and dropout

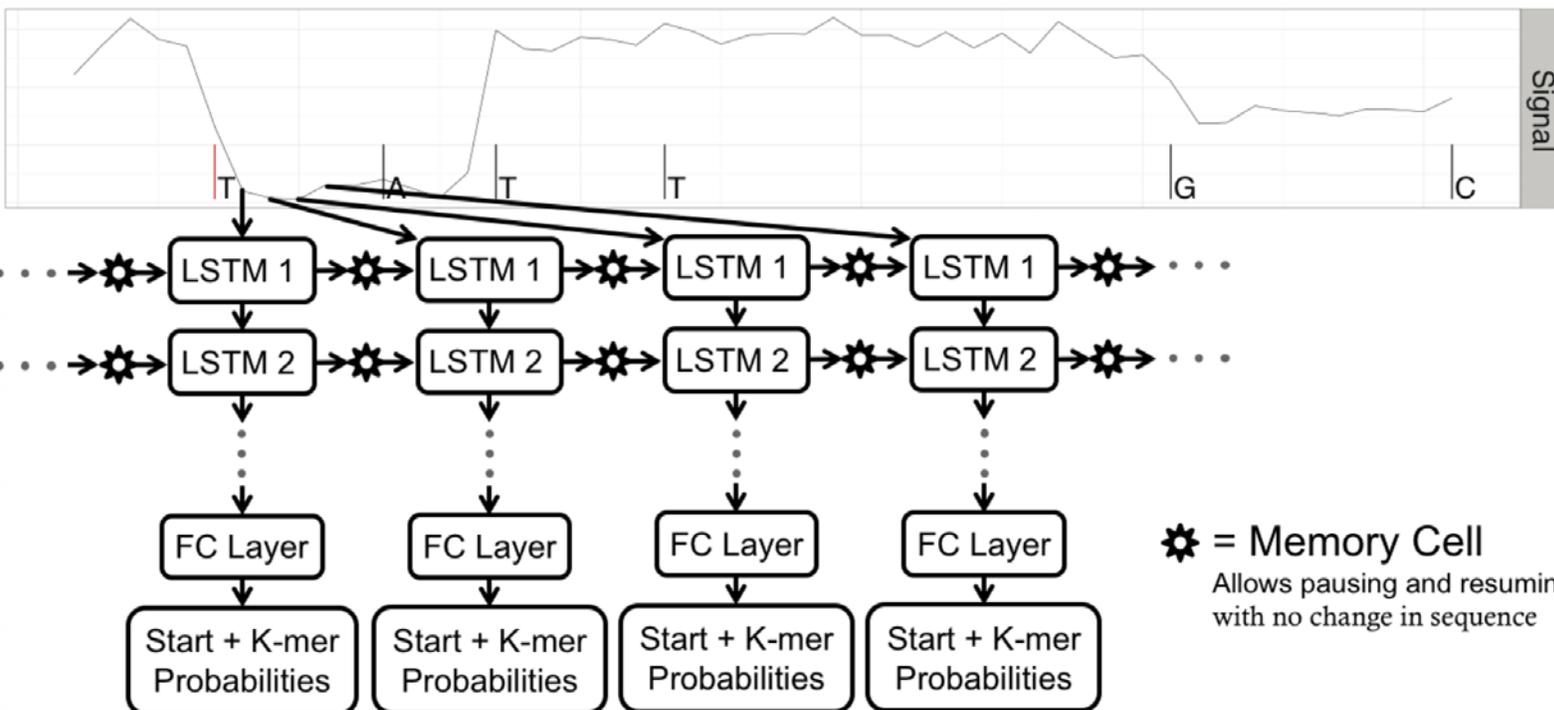


- 37 -

5 Oxford Nanopore Sequencing



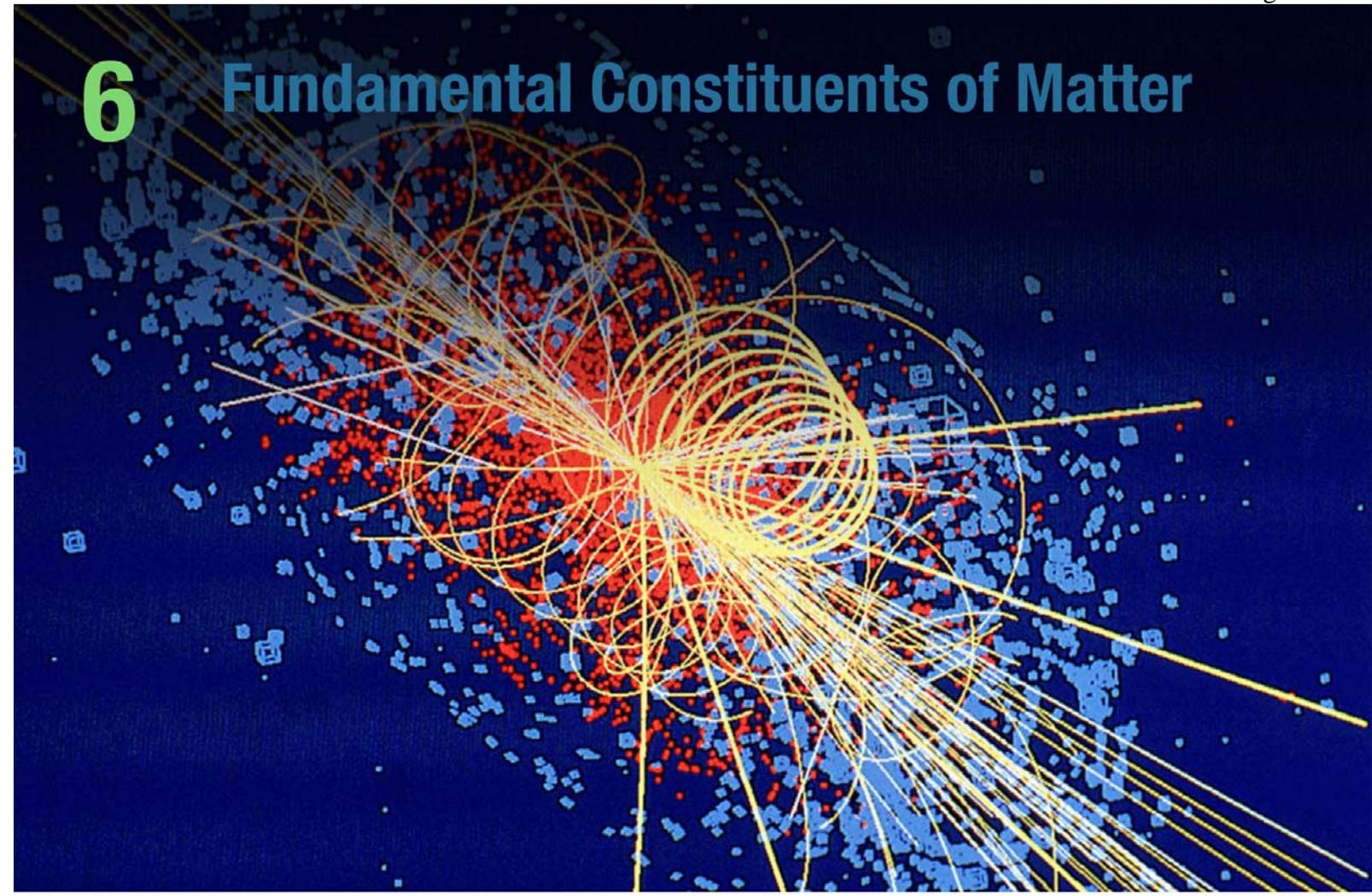
LSTMs provide state-of-the-art performance



Team: Ben Brown, Marcus Stoiber



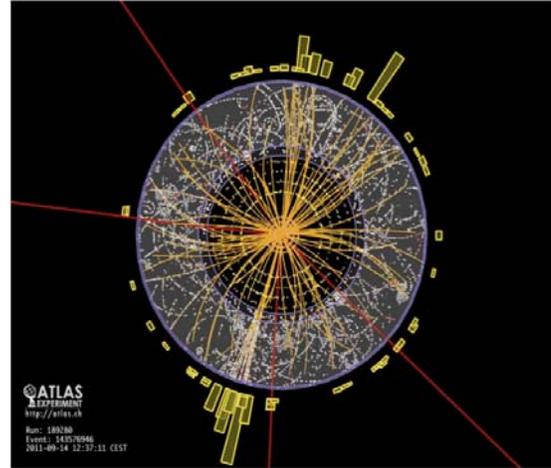
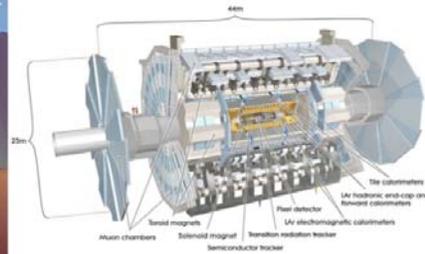
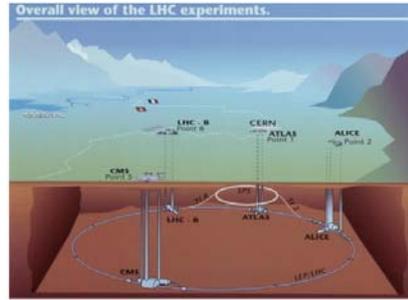
6 Fundamental Constituents of Matter



LHC Experiment



- Colliding protons with high energy
- Particles produced in collision (“event”) hit detector
- Physicist need to decide which events are interesting and which can be described by physics we know
- Large amount of data recorded
 - 1PB/s reduced to 100GB/s
 - 10PB of raw data/year



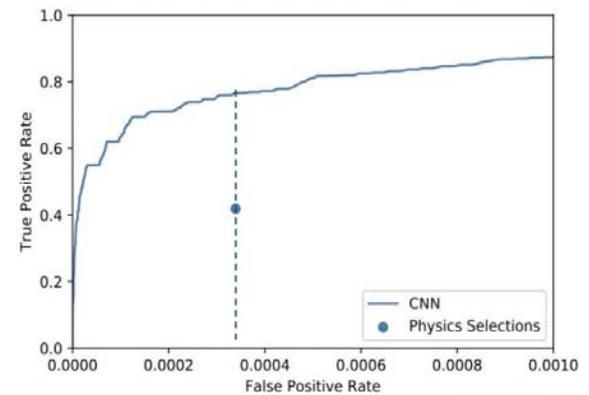
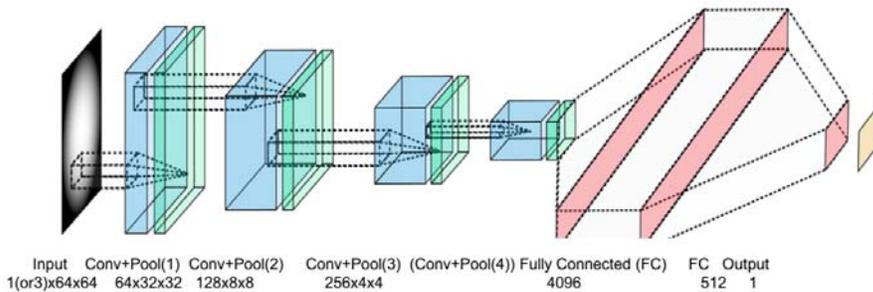
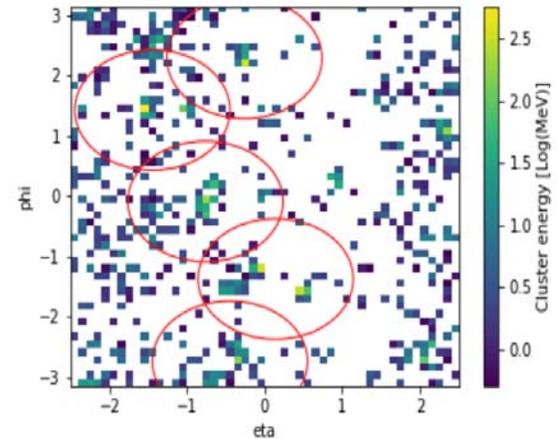
Contributors: Wahid Bhimji, Thorsten Kurth, Steve Farrell, Evan Racah



LHC Classification Approach



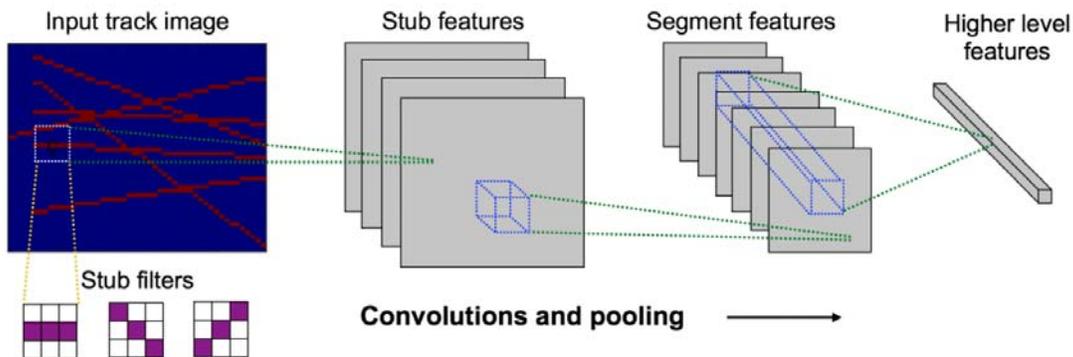
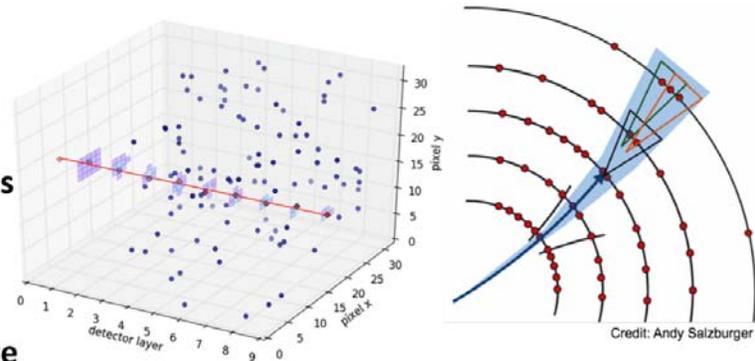
- Bin energy from sub-detector ('calorimeter') and unroll cylinder to form 64x64 or 224x224 image
- Train CNN on labelled data from full detector simulations to directly classify signal ('Supersymmetry') from background
- Benchmark from existing analysis on high-level physics variables
- *Increased signal efficiency at same background rejection without using high-level physics variables*



LHC Particle Tracking

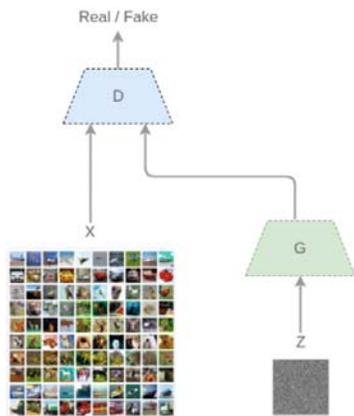


- Reconstruct thousands of particle tracks from tens of thousands of spacepoint “hits”
- Traditional algorithms have limitations
 - Hand-engineered, quadratic (or worse) scaling, linear dynamics
- HEP.TrkX project is exploring ML solutions
- Using recurrent architectures for track dynamics
 - Kalman-filter-like state estimation
 - Smarter combinatorial tree-search
- Using CNNs to classify hits
- Using CNN + LSTM to “caption” a detector image



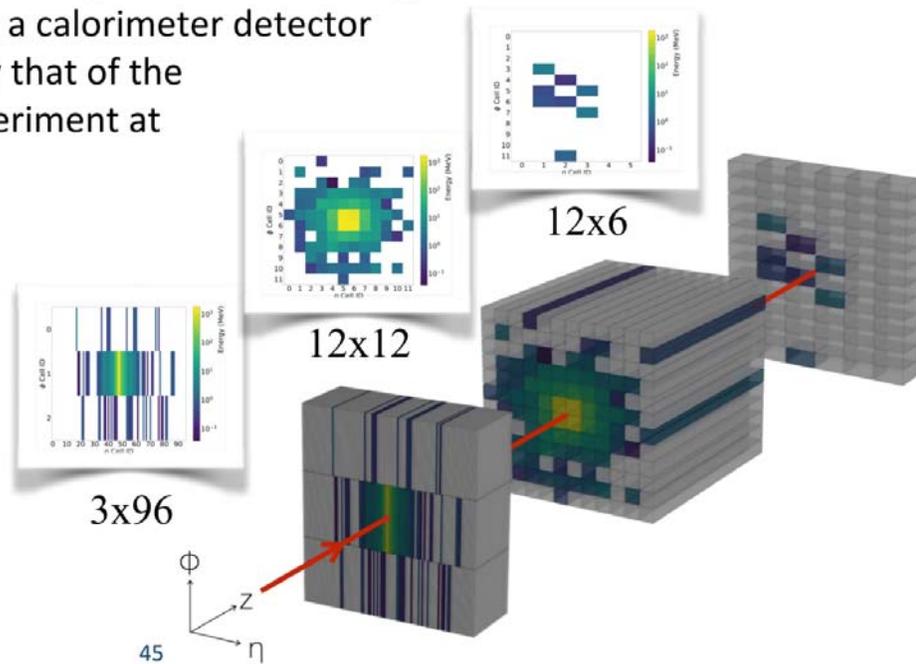
Contributors: Caltech, FNAL, LBL collaborators. ASCR/HEP Pilot project.

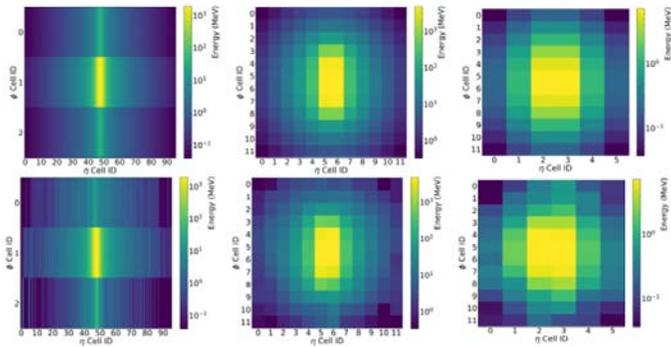




- Goal: accelerating particle physics simulation
- Fast & accurate generation of energy deposits in a calorimeter detector inspired by that of the ATLAS experiment at the LHC

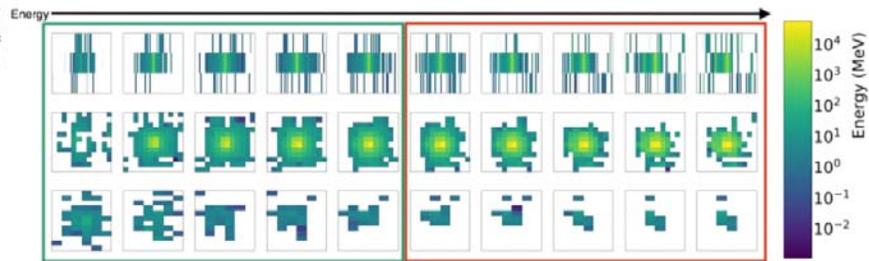
- Ad-hoc design to fit Physics data:
 - sparsity
 - high dynamic range
 - highly location-dependent features





Average energy deposition per calorimeter layer in the GEANT4 training dataset (top) and in the GAN generated dataset (bottom)

- Conditional generation based on physical attributes
- Parameter interpolation and extrapolation



Requested (GeV)	1.0	23.1	45.2	67.3	89.4	111.6	133.7	155.8	177.9	200.0
Generated (GeV)	1.4	23.1	46.8	69.3	91.6	113.6	135.2	159.3	184.4	211.2

Ten positron showers generated by varying shower energy in equal intervals while holding all other latent codes fixed. The three rows are the shower representations in the three calorimeter layers. The energies of showers in the green box were within the range of the training dataset, while the ones in the red box are in the extrapolation regime.

- Realistic average and individual images
- Diverse samples

	HEP			BER		BES		NP	FES
	Astronomy	Cosmology	Particle Physics	Climate	Genomics	Light Sources	Materials	Particle Colliders	Plasma Physics
Classification	X		X	X	X	X	X	X	X
Regression		X			X	X	X	X	X
Clustering		X	X	X	X	X	X	X	X
Dimensionality Reduction				X				X	
Surrogate Models	X	X	X				X	X	X
Design of Experiments		X		X			X		X
Feature Learning	X	X	X	X	X	X	X	X	X
Anomaly Detection	X		X	X		X		X	

	HEP			BER		BES		NP	FES
	Astronomy	Cosmology	Particle Physics	Climate	Genomics	Light Sources	Materials	Particle Colliders	Plasma Physics
Classification	X		X	X	X	X	X	X	X
Regression		X			X	X	X	X	X
Clustering		X	X	X	X	X	X	X	X
Dimensionality Reduction				X				X	
Surrogate Models	X	X	X				X	X	X
Design of Experiments		X		X			X		X
Feature Learning	X	X	X	X	X	X	X	X	X
Anomaly Detection	X		X	X		X		X	

Short-Term Challenges



- **Complex Data**
 - 2D/3D/4D, #channels, dense/sparse, graph structure
- **Hyper-Parameter Optimization**
 - Tuning #layers, #filters, learning rates, schedule is a black art
- **Performance and Scaling**
 - Current networks take days to train on O(10) GB datasets, we have O(100) TB datasets on hand
- **Scarcity of Labeled Data**
 - Communities need to self-organize and run labeling campaigns

Long-Term Challenges



- **Lack of Theory**
 - Limits of supervised, unsupervised, semi-supervised learning
- **Formal protocol for applying Deep Learning**
 - Applied Math has developed methodology over 30 years, no analog in DL
- **Interpretability: ‘Introspect It’ vs. ‘Build It’**
 - Black Box classifier; need to visualize representations
 - Incorporate domain science principles (physical consistency, etc)
- **Uncertainty Quantification**



2018-2020



- **Broad deployment of tools at HPC centers and Cloud**
- **Domain science communities will start self-organizing and conducting labeling campaigns**
- **Researchers will exploit low-hanging fruit**
 - Classification, Regression, Clustering problems will be (nearly) completely solved



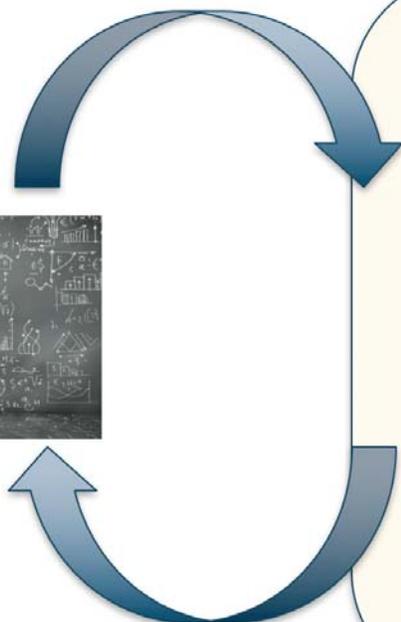
- **Entire data archives are segmented and classified**
 - Anomaly detection; Correlation; Causal Analysis
- **Long-term challenges are formulated and addressed**
 - Generalization limits, UQ
 - Interpretability, incorporating domain science principles
- **What is the ‘value add’ of the scientist?**

2020+ Workflow

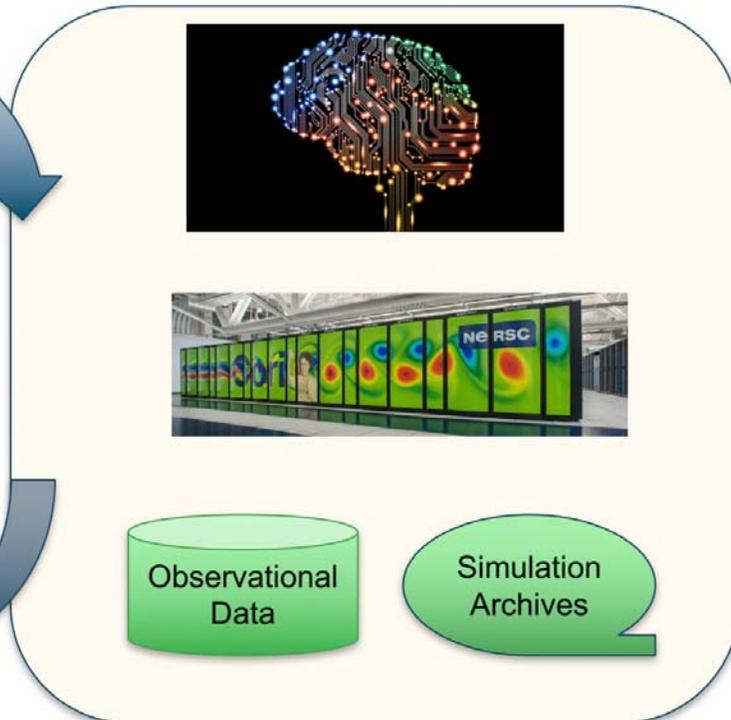


- Interactive Exploration
- Semantic Labels

- Mechanisms
- Hypothesis



- Patterns
- Clusters
- Anomalies



Conclusions



- **Machine Learning is an emerging requirement in the DOE community**
 - NERSC has invested in staff, hardware and software
 - Big Data Center is enabling capability applications
- **Deep Learning has enabled breakthroughs in industry; direct analogs in DOE applications**
 - Current success stories from BER, HEP, NP; broader class of applications poised to benefit
- **Low-hanging fruit can be exploited in the next 2-3 years, but long-term challenges exist**
- **Exciting times!**





Questions?
prabhat@lbl.gov