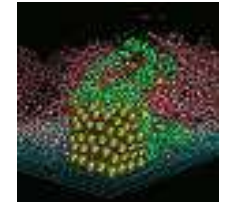
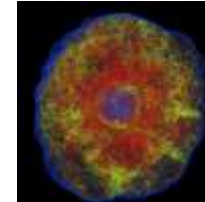
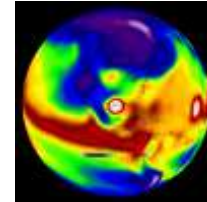
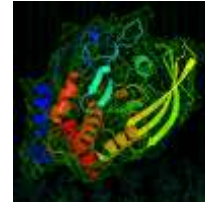
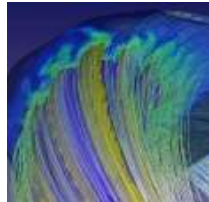
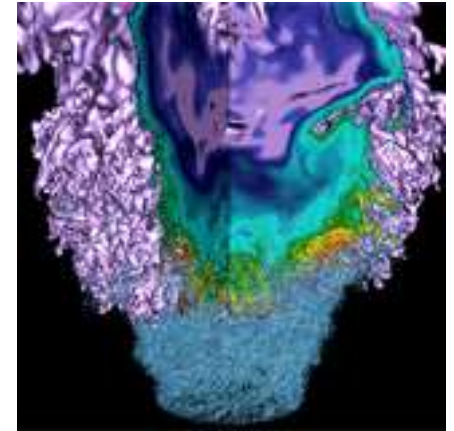


# Machine Learning



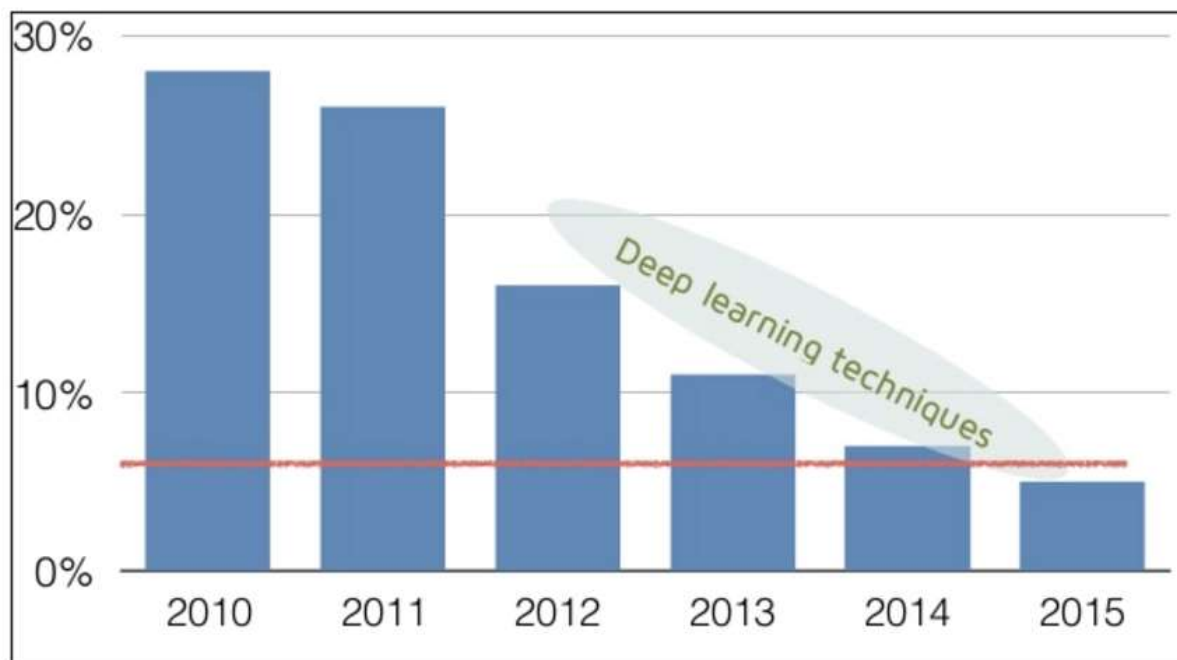
Prabhat

HPC User Forum

April 12, 2016

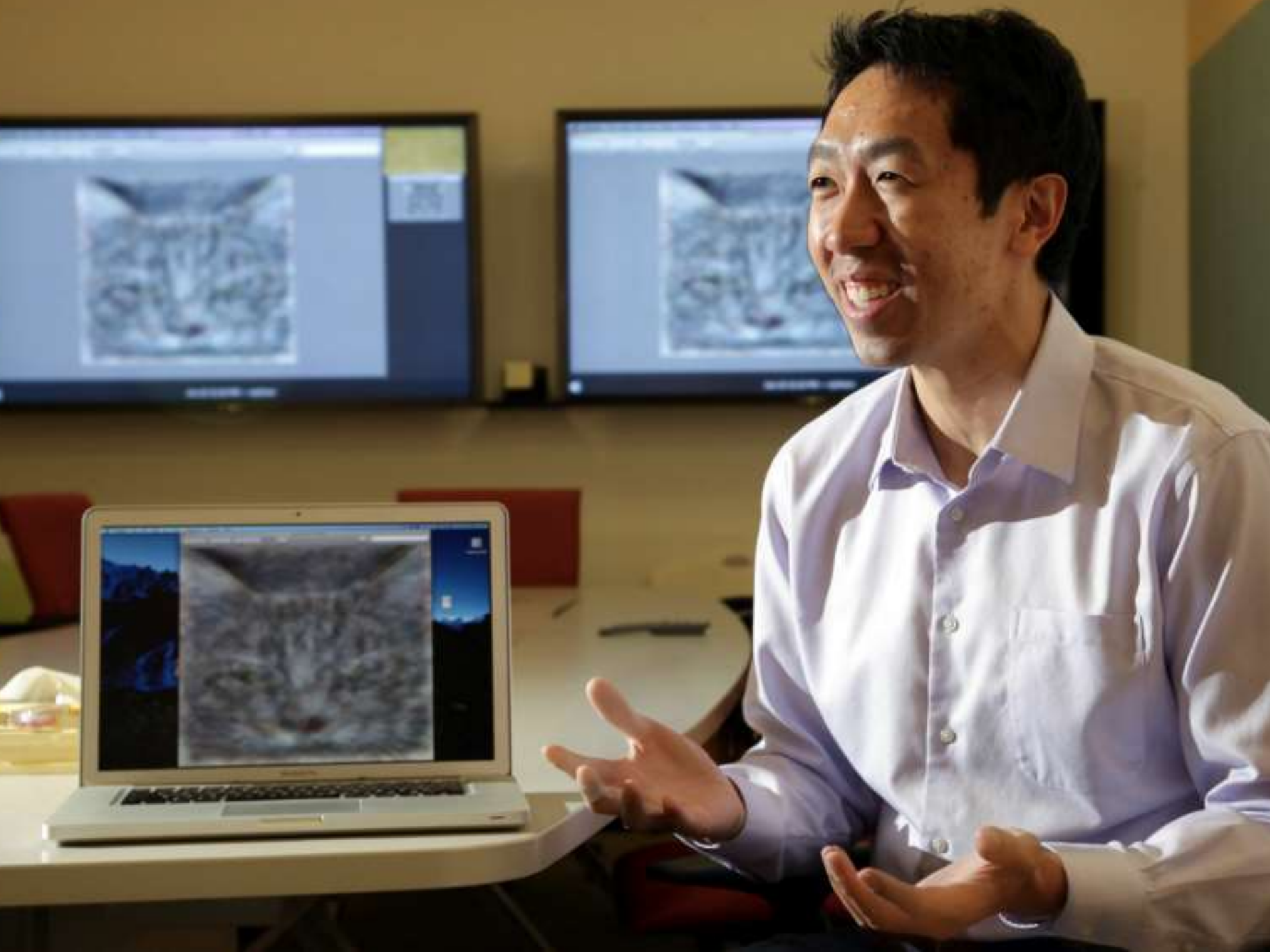
# Imagenet ILSVRC Challenge

Error rate<sup>1</sup>



human  
performance

<sup>1</sup>: Imagenet top 5 error rate  
Source: Imagenet



# (2012) This is all great, but...

---

- **Is Machine Learning relevant to science?**
  
- **Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?**

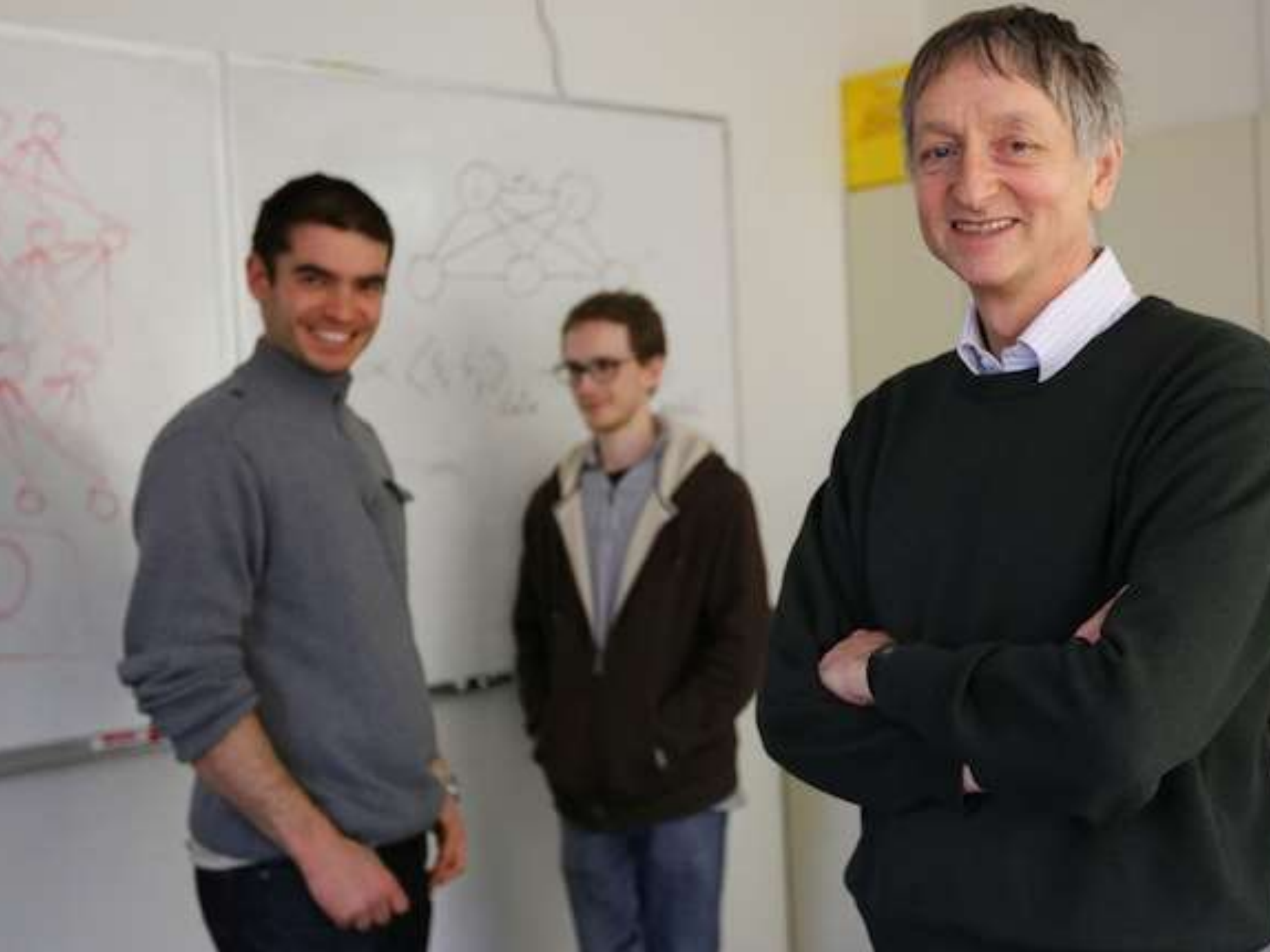
# (2012) This is all great, but...

---

- **Is Machine Learning relevant to science?**
  - Success stories are for images and audio, but how about scientific data?
- **Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?**
  - Our applied mathematicians are content with formulating and solving PDEs
  - The NNSA folks care about Uncertainty Quantification
  - Our data ‘analytics’ folks are happy dealing with meshes, computational geometry, topology





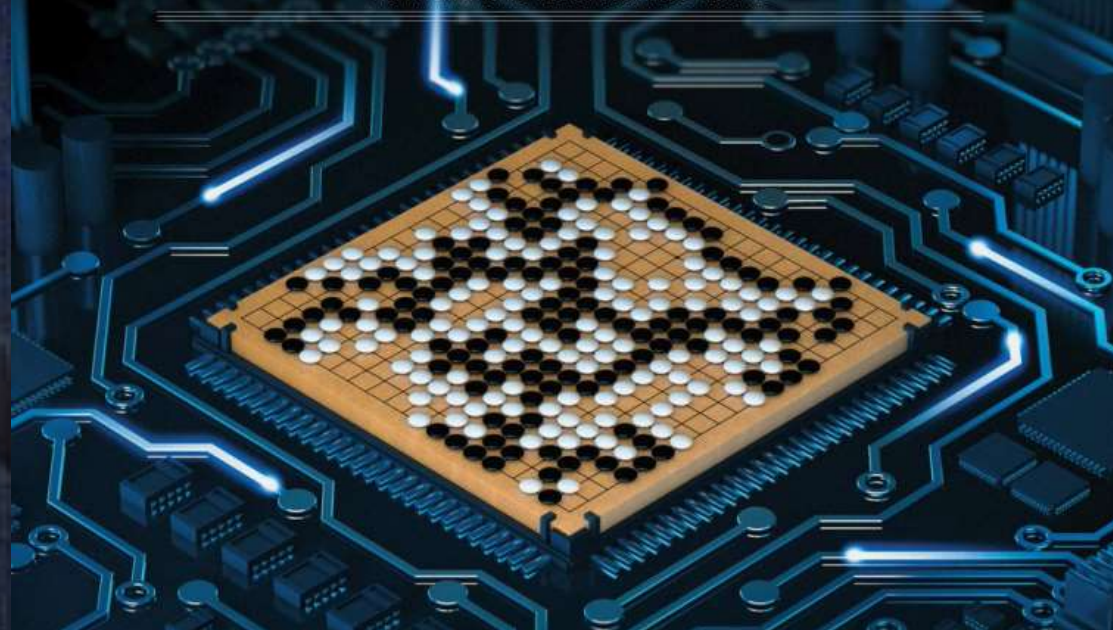






# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



At last — a computer program that can beat a champion Go player **PAGE 484**

## ALL SYSTEMS GO

### CONSERVATION

#### SONGBIRDS A LA CARTE

Illegal harvest of millions  
of Mediterranean birds

PAGE 452

### RESEARCH ETHICS

#### SAFEGUARD TRANSPARENCY

Don't let openness backfire  
on individuals

PAGE 459

### POPULAR SCIENCE

#### WHEN GENES GOT 'SELFISH'

Dawkins's calling  
card 40 years on

PAGE 462

NATUREASIA.COM

28 January 2016

Vol. 529, No. 7587



# (2016) The writing is on the wall

---

- **O(B) \$ worth of investment by industry**
- **Machine Learning and Statistics are established as key disciplines for this decade**
  - Deep Learning has taken off as the most promising ML technique

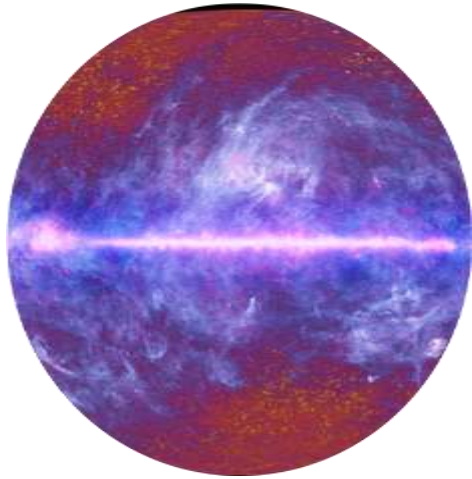
# (2012) Revisited..

---

- **Is Machine Learning relevant to science?**
- **Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?**

# (2010-2016): The Rise of Data-Intensive Science

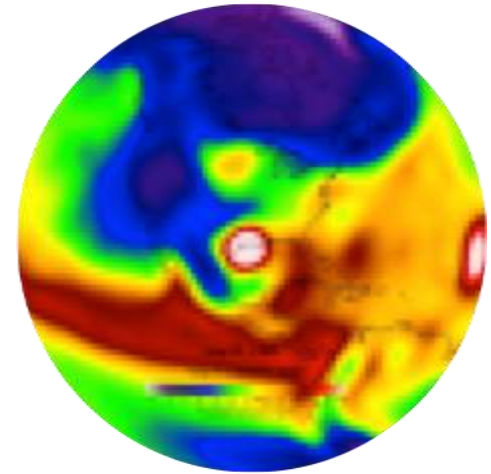
---



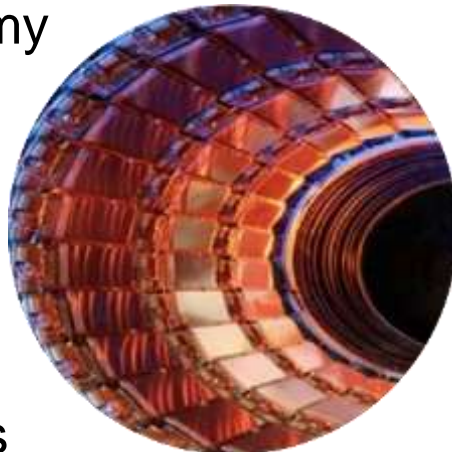
Astronomy



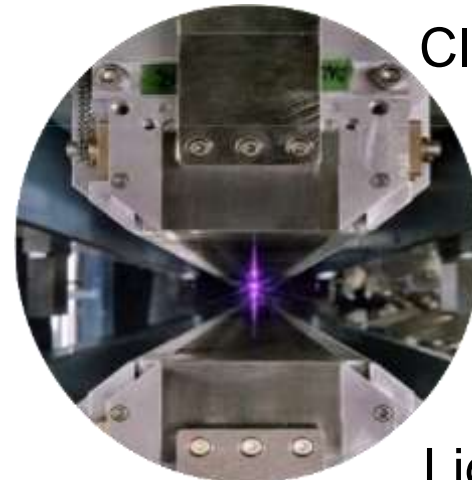
Genomics



Climate



Physics



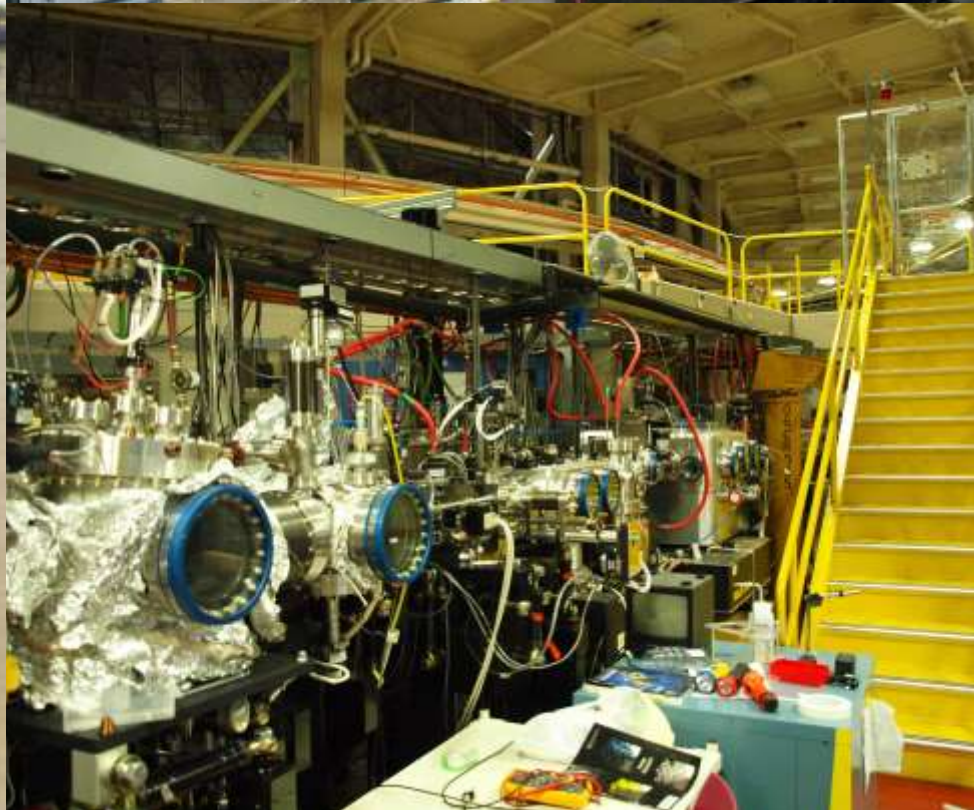
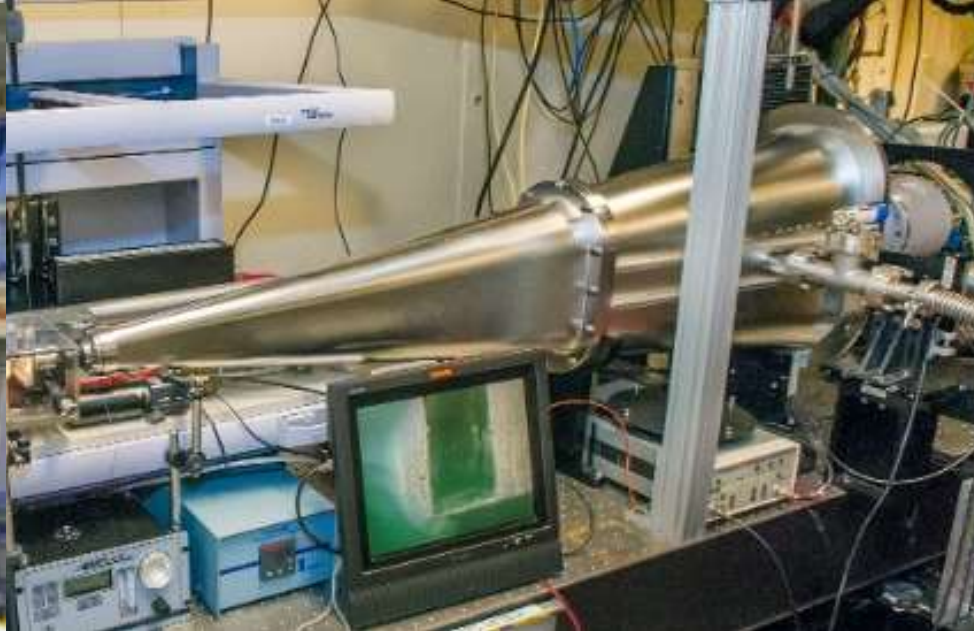
Light Sources

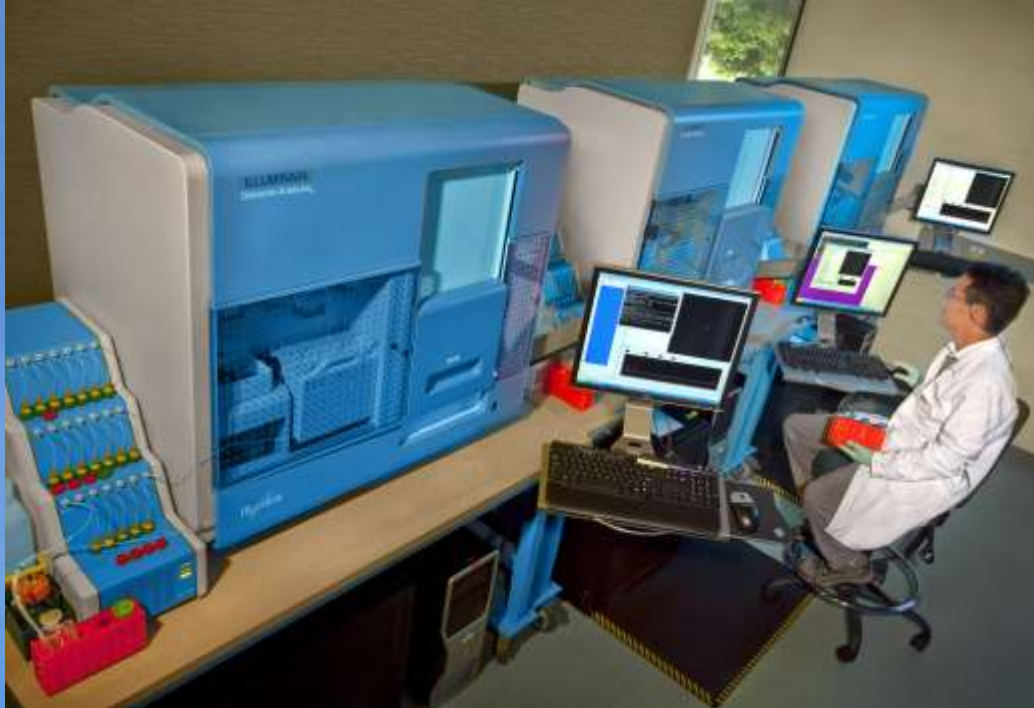














U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science

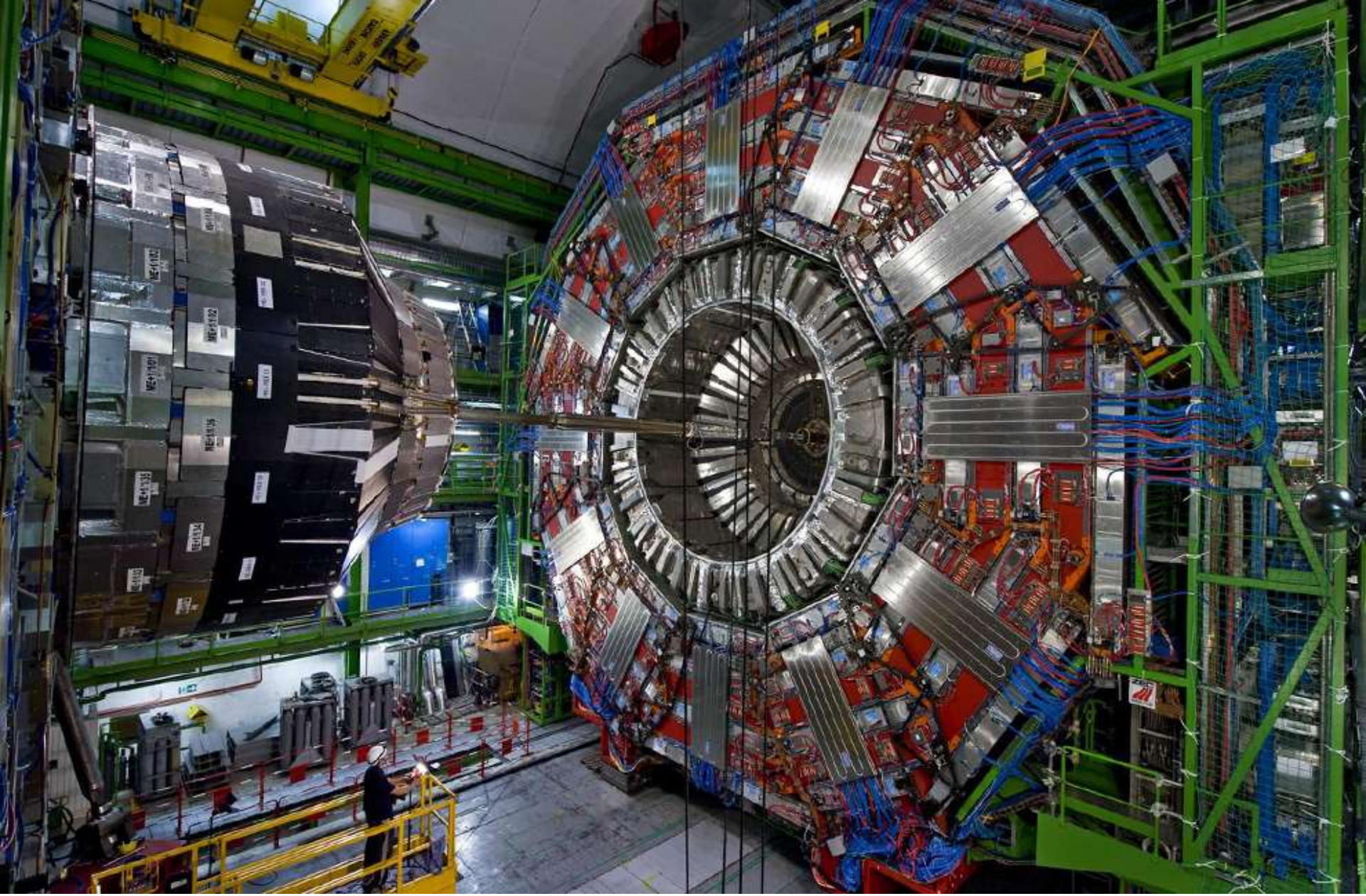




U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science





# 4 V's of Scientific Big Data

Science Domain	Variety	Volume	Velocity	Veracity
Astronomy	Multiple Telescopes, multi-band/spectra	O(100) TB	100 GB/night – 10 TB/night	Noisy, acquisition artefacts
Light Sources	Multiple imaging modalities	O(100) GB	1 Gb/s-1 Tb/s	Noisy, sample preparation/acquisition artefacts
Genomics	Sequencers, Mass-spec, proteomics	O(1-10) TB	TB/week	Missing data, errors
High Energy Physics	Multiple detectors	O(100) TB – O(10) PB	1-10 PB/s reduced to GB/s	Noisy, artefacts, spatio-temporal
Climate	Simulations Multi-variate, spatio-temporal	O(10) TB	100 GB/s	'Clean', need to account for multiple sources of uncertainty

# Does Machine Learning matter?

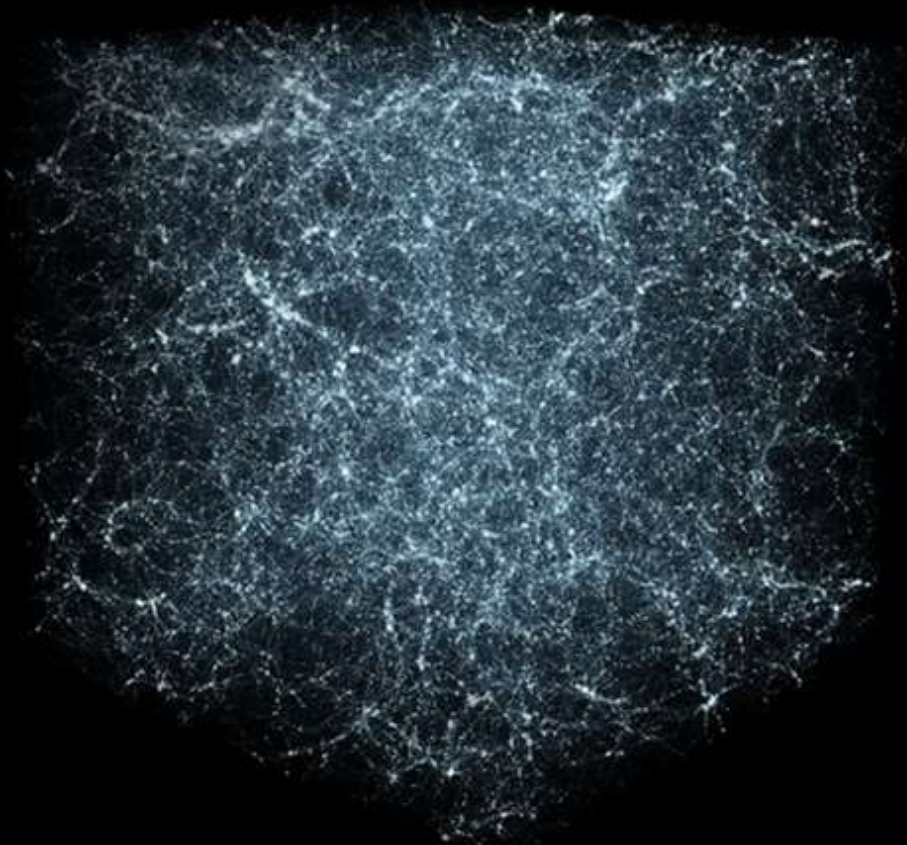
---

- **Is Machine Learning relevant to science?**
  - Yes!
- **Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?**
  - Analytics is *the* key step for gaining scientific insights
  - The nature of questions in data-intensive science are inferential
  - Statistics and Machine Learning deal with inference in presence of noise and errors

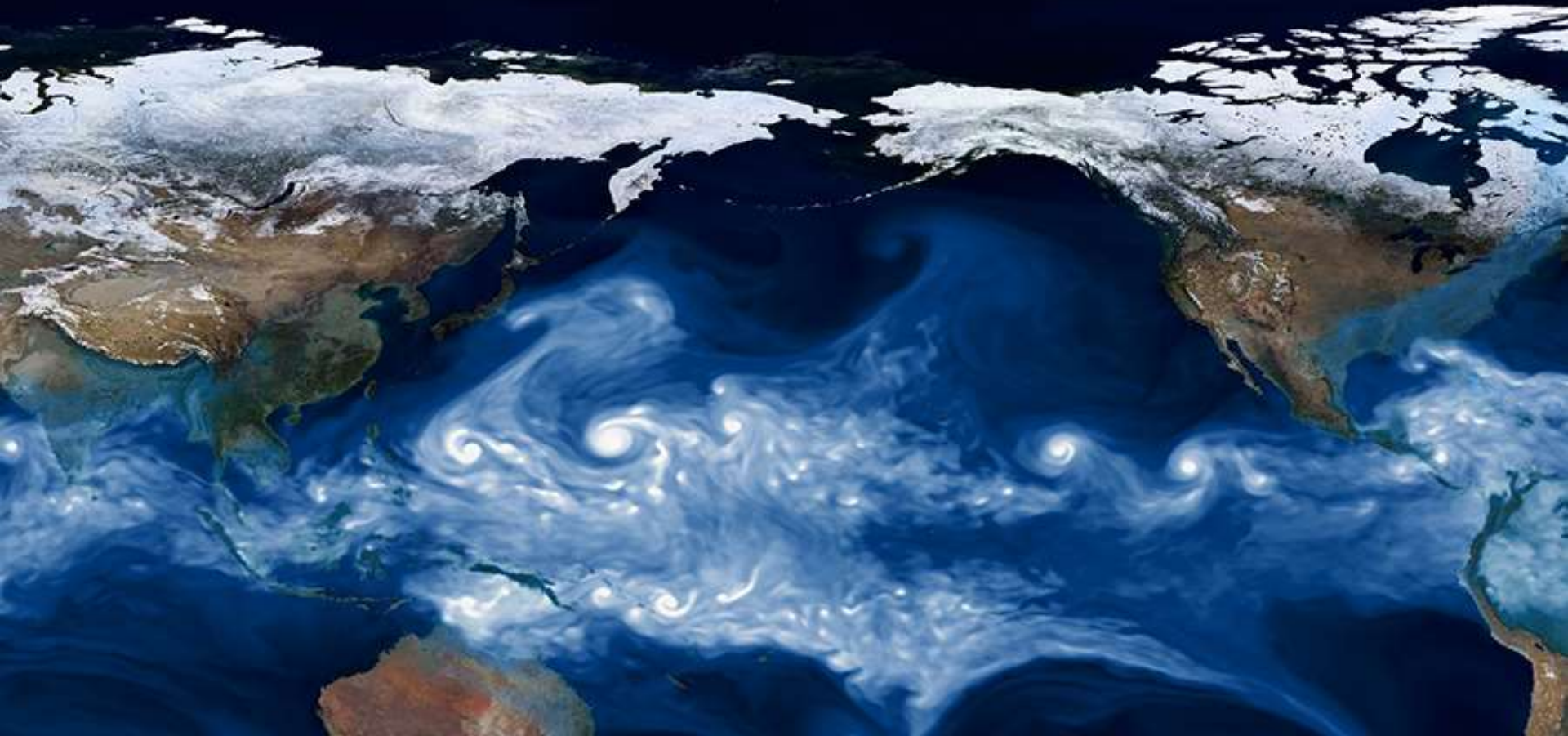
# 1 Creating a catalog of all objects in the Universe



# 2 Fundamental Constants of Cosmology



# 3 Characterizing Extreme Weather in a Changing Climate



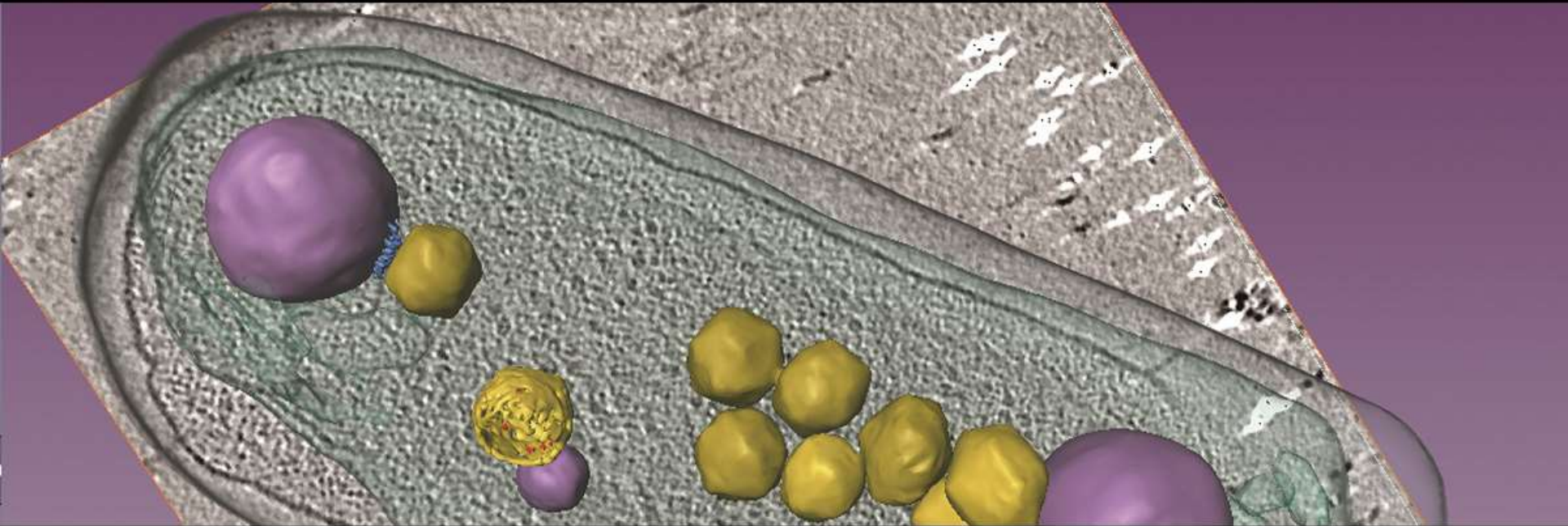
# 4 Knowledge Extraction from Scientific Literature



# 5 Understanding Speech Production



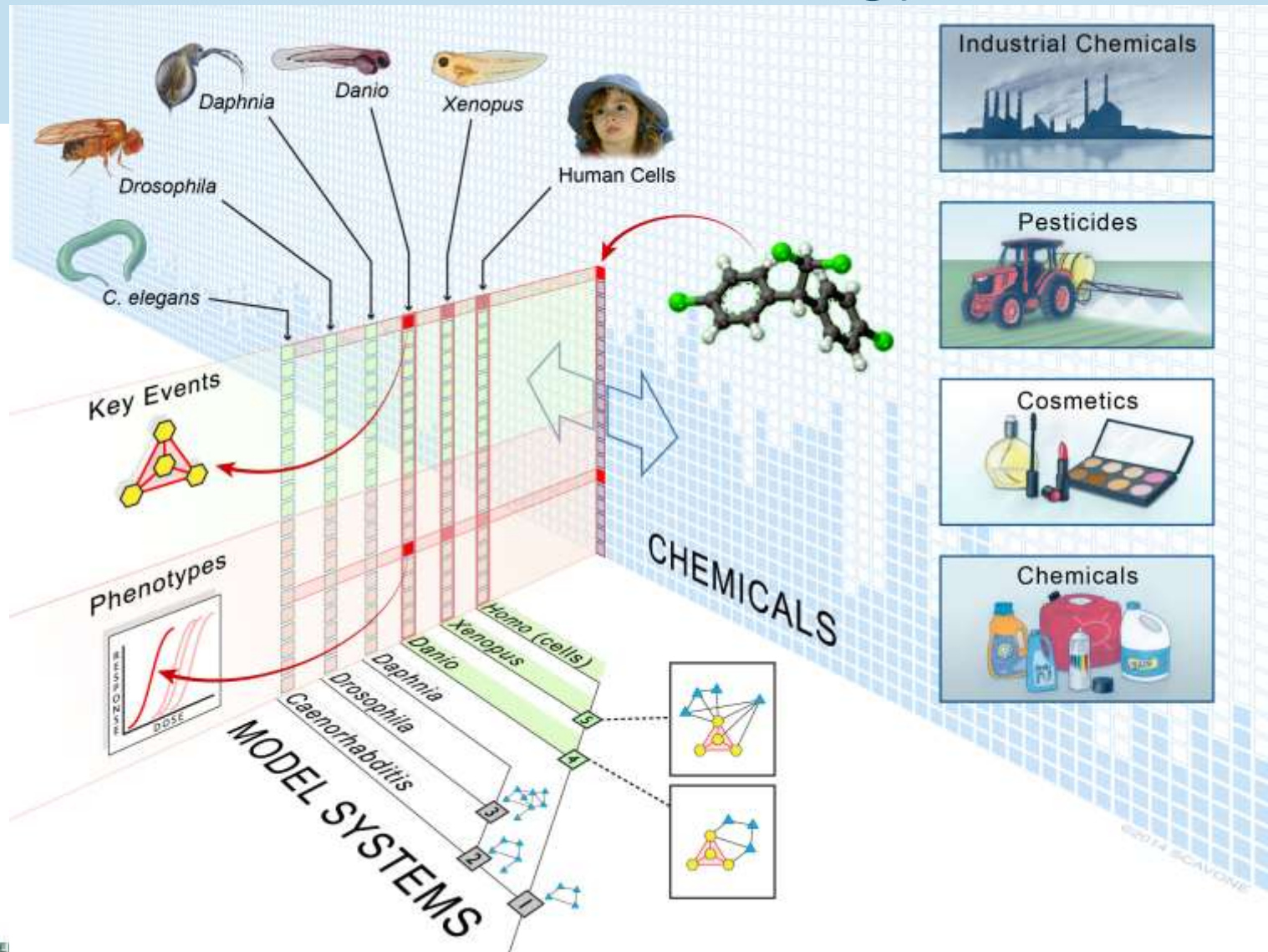
# 6 Quantitative and Predictive Biology



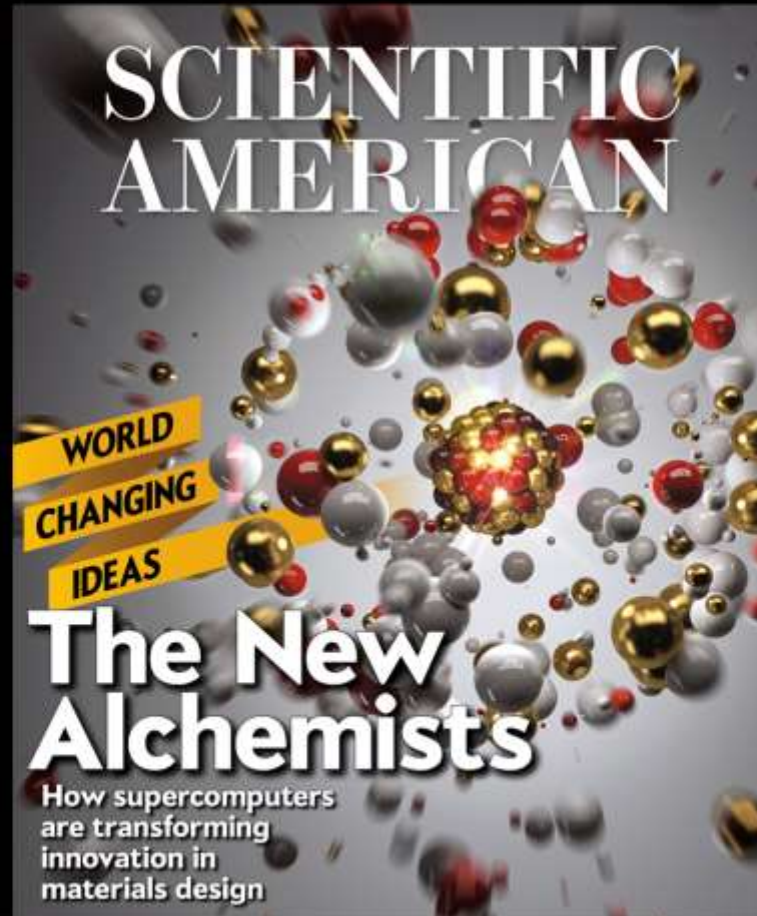
# 7 Understanding the Genetic Code



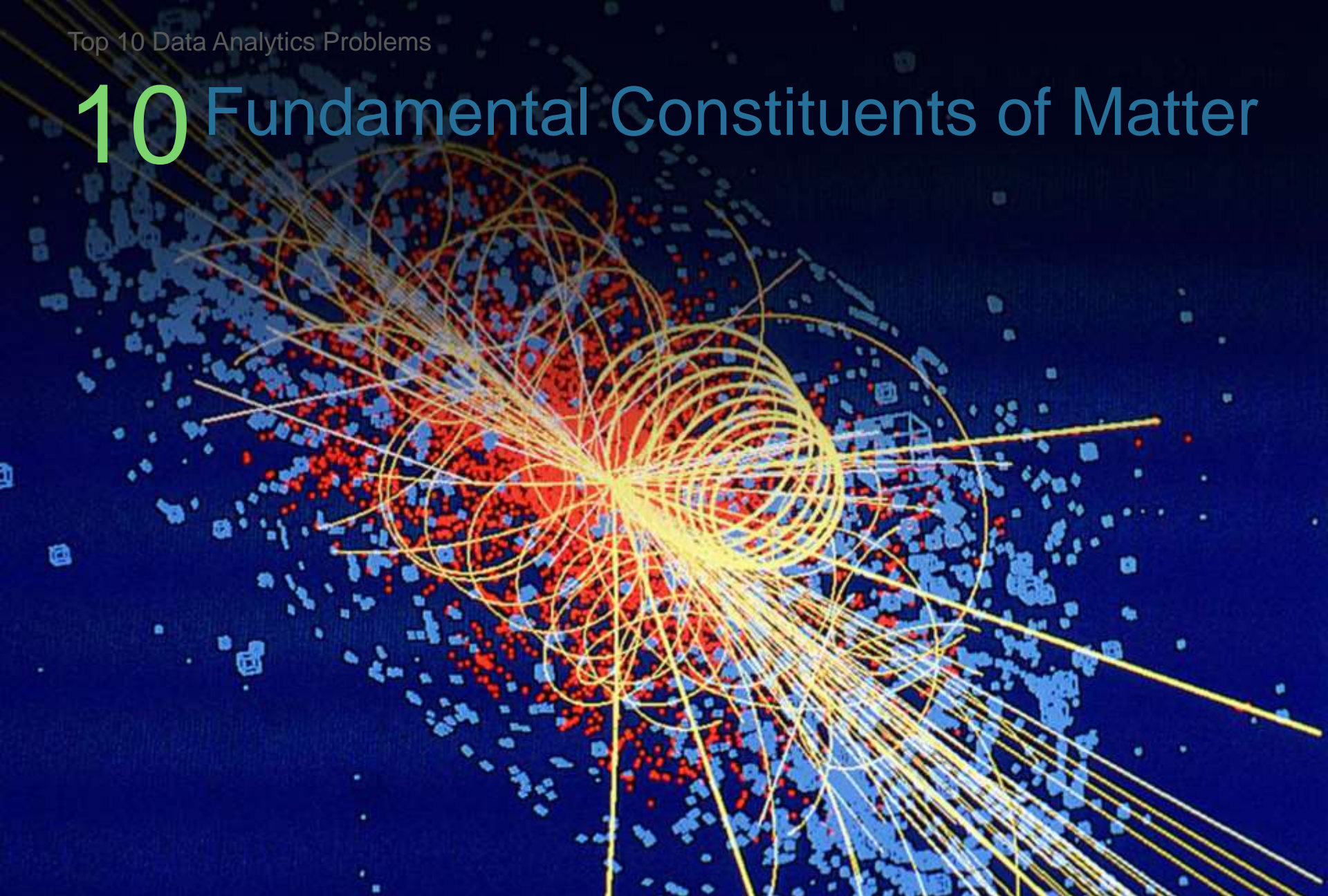
# 8 Personalized Toxicology



# 9 Designer Materials



# 10 Fundamental Constituents of Matter



# Towards Synthesis (and maybe Convergence)

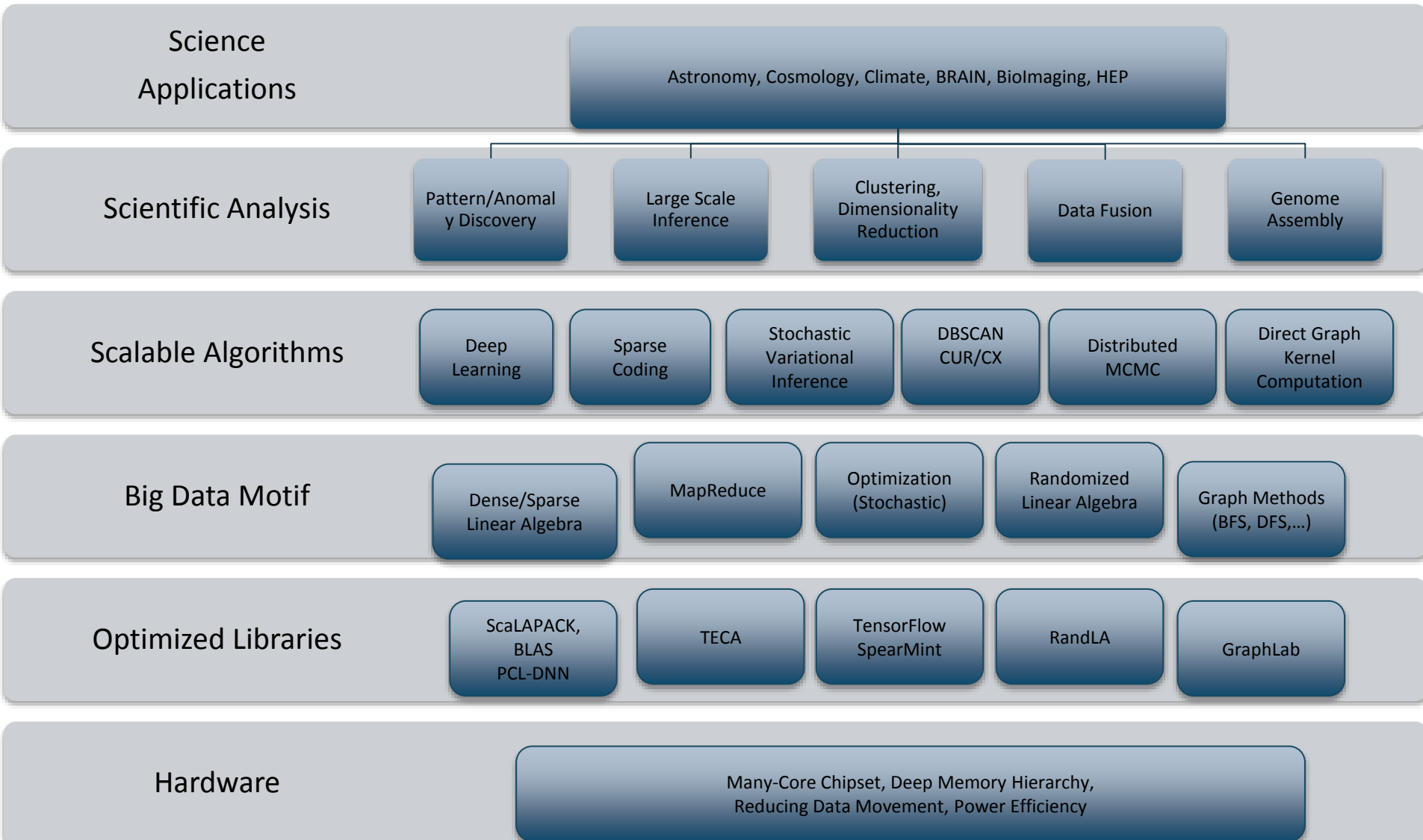
---

- **What is the landscape of Machine Learning problems in science?**
  - Bewildering array of taxonomy and domain-specific terminology
- **What are the key computational motifs?**
  - Need to have a productive conversation with HPC software, hardware vendors

	Astronomy	Cosmology	Climate	Systems Biology	Neuroscience	EM/X-Ray Imaging	Mass-spec Imaging	Personalized Toxicology	Materials	Particle Physics
Classification	X		X		X	X	X			X
Regression					X			X	X	
Clustering		X	X		X	X	X			X
Dimensionality Reduction			X		X		X			
Inference	X						X			X
Model Estimation	X				X			X		
Design of Experiments		X	X						X	
Semantic Analysis			X	X					X	
Feature Learning			X		X	X	X	X	X	X
Anomaly Detection	X		X			X				X

	Astronomy	Cosmology	Climate	Systems Biology	Neuroscience	EM/X-Ray Imaging	Mass-spec Imaging	Personalized Toxicology	Materials	Particle Physics
Classification	X		X		X	X	X			X
Regression					X			X	X	
Clustering		X	X		X	X	X			X
Dimensionality Reduction			X		X		X			
Inference	X						X			X
Model Estimation	X				X			X		
Design of Experiments		X	X						X	
Semantic Analysis			X	X					X	
Feature Learning			X		X	X	X	X	X	X
Anomaly Detection	X		X			X				X

# Machine Learning Research Strategy



# Machine Learning Research Strategy



# Machine Learning: Challenges

---

- **Cultural**

- ML doesn't cleanly 'fit' within Computer Science or Applied Math
- Statistics, CS (Machine Learning, HPC) taxonomy
- Mindshare
  - Attracting the best academic and industry talent is hard

- **Technical**

- Big Data ecosystem has evolved independently of HPC
- Aspirations of Convergence (Software, Hardware)
  - HPC institutions need to do a better job of characterizing their Data Analytics requirements

# Machine Learning: Opportunities

---

- **HPC community is uniquely positioned**
  - Storage and Compute Hardware
  - Meaningful scientific problems
- **Software (Research and Production) is wide open**
- **Most exciting discoveries happen at the intersection of domain sciences and methods**
  - We don't know the limits of Deep Learning methods



# Thanks!

---



Contact: [prabhat@lbl.gov](mailto:prabhat@lbl.gov)