Advancing the Metagenomics Revolution

Invited Talk
Symposium #1816, Managing the Exaflood: Enhancing the Value of Networked Data for Science and Society
San Diego, CA
February 2010

Dr. Larry Smarr
Director, California Institute for Telecommunications and Information Technology
Harry E. Gruber Professor,
Dept. of Computer Science and Engineering
Jacobs School of Engineering, UCSD
lsmarr@twitter.com

camera
Marine Microbial Ecology

it²

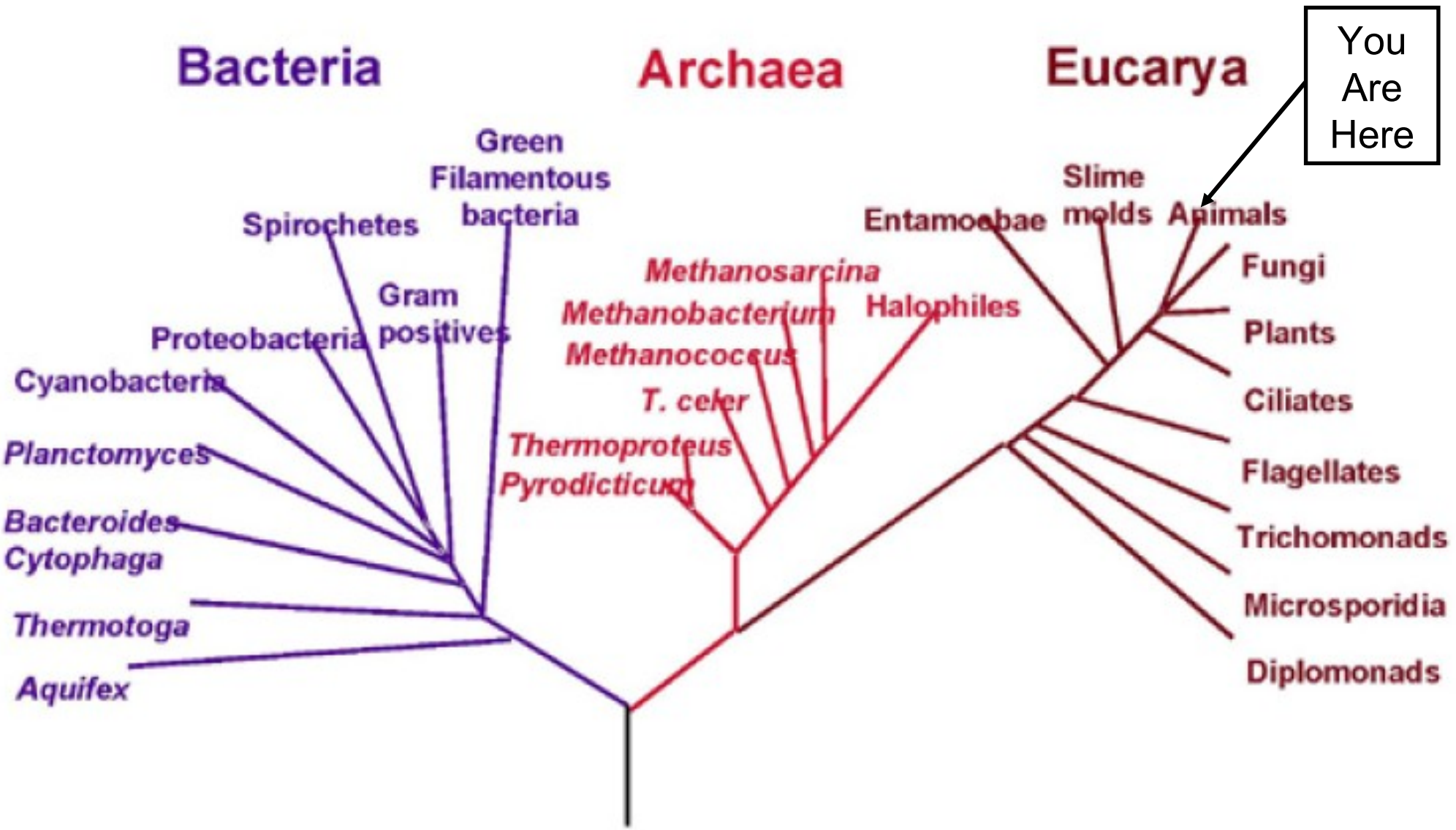
Abstract

The vast majority of life on earth is microbial. Virtually all ecologies rely on the intricate biochemistry of microbial life to sustain themselves. Historically most research on microbes depended on laboratory cultures, but since 99% of microbes cannot be cultured, it is only recently that modern genetic sequencing techniques have allowed determination of the hundreds to thousands of microbial species present at a specific environmental location. The amount of data specifying the “metagenomics” of these microbial ecologies is explosively growing as researchers everywhere are acquiring next generation sequencing devices. Since many genes are related across microbial species, the community needs repositories in which diverse environmental metagenomics samples can be quickly compared, both by comparing genomic data or environmental metadata. I will give a quantitative example of the computing, storage, software, and networking architecture needed to handle this exponentially growing data flood by describing the Gordon and Betty Moore Foundation funded Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) which is hosted by Calit2@UCSD. The CAMERA repository currently contains over 500 microbial metagenomics datasets (including Craig Venter’s Global Ocean Survey), as well as the full genomes of ~166 marine microbes. Registered end users, over 3000 from 70 countries, can access existing and contribute new metagenomics data either via the web or over novel dedicated 10 Gb/s light paths. The user’s BLAST requests transparently activate programs on dedicated and shared parallel computing resources at UCSD. To better support the CAMERA user community, we developed a new component-based cyberinfrastructure, CAMERA Version 2.0. This new cyberinfrastructure will support future needs for data acquisition, data access through diverse modalities, the addition of externally developed tools, and the orchestration of these tools into reproducible analytical pipelines. The management of remote applications and analyses is accomplished via the Kepler workflow engine which supports the natural interaction of automated computational tools that can then be re-utilized and openly shared. Finally, CAMERA 2.0 includes an effective, flexible, and intuitive user interface that facilitates and enhances the process of collaborative scientific discovery for biosciences. I will conclude by examining future trends in metagenomics data generation, data standardization, and the possible use of cloud computing and storage.

camera
Marine Microbial Ecology

it2

Most of Evolutionary Time Was in the Microbial World



Tree of Life Derived from 16S rRNA Sequences

Source: Carl Woese, et al

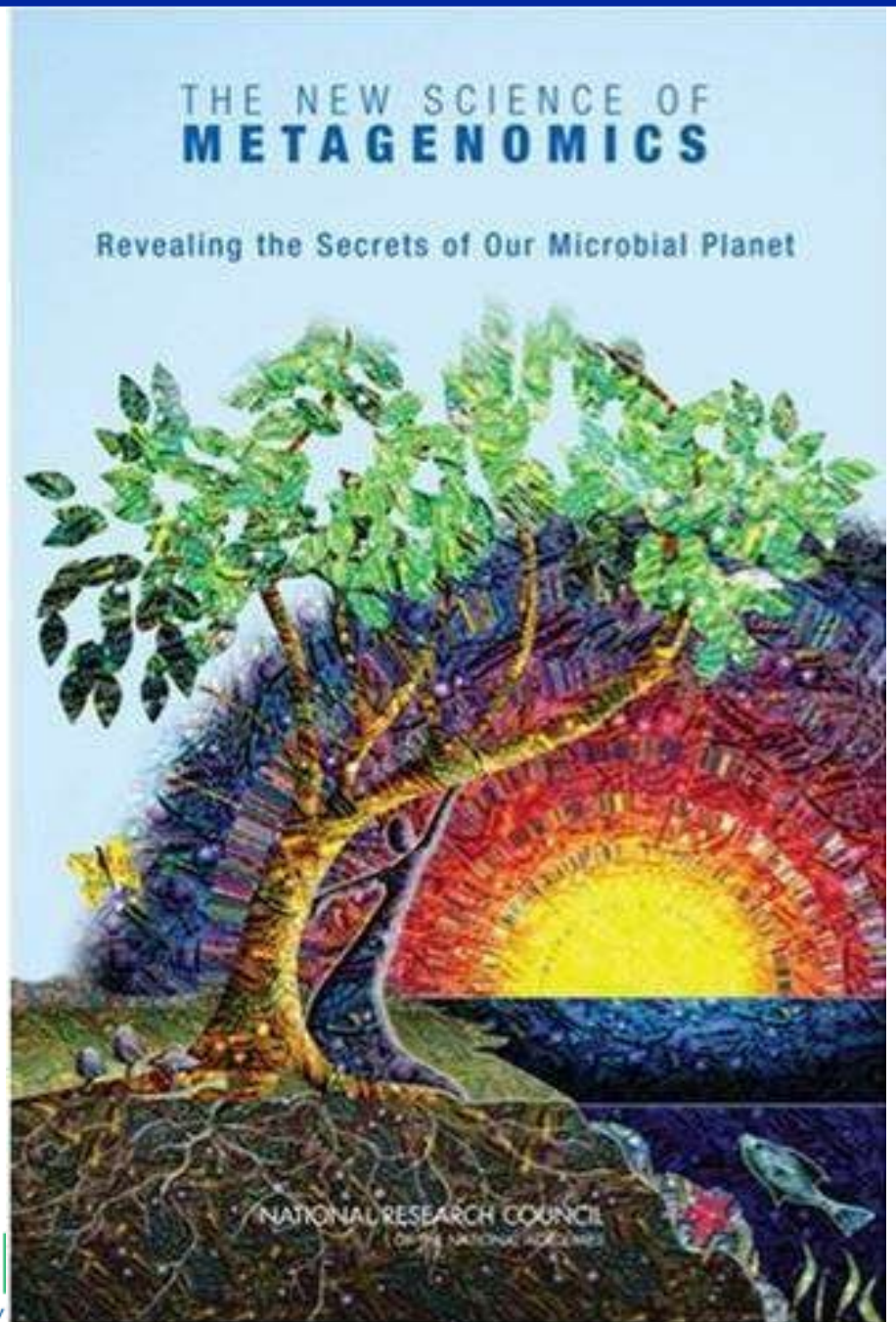
camera Marine Microbial Ecology

it2

The New Science of Metagenomics

NRC Report:

Metagenomic data should be made publicly available in international archives as rapidly as possible.



"The emerging field of metagenomics, where the DNA of entire communities of microbes is studied simultaneously, presents the greatest opportunity -- perhaps since the invention of the microscope – to revolutionize understanding of the microbial world." –

National Research Council
March 27, 2007

came
Marine Microbial Ecology

it2

Enormous Increase in Scale of Known Genes Over Last Decade

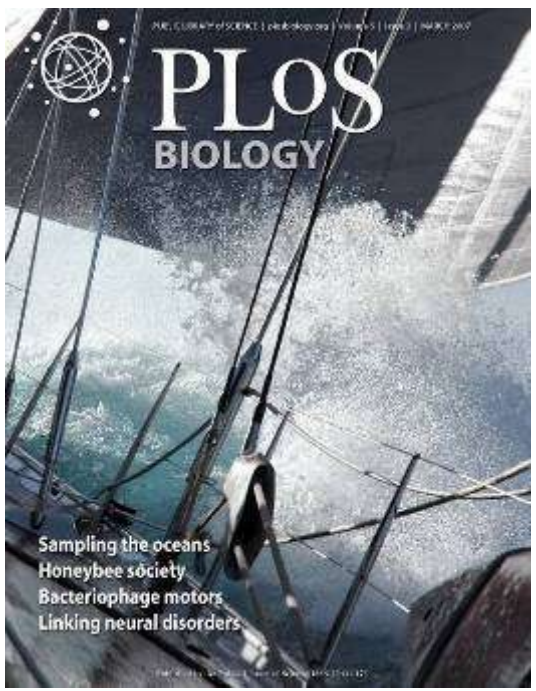
1995
First Microbe Genome



1.8 Million Bases
1749 Genes

J. Craig Venter
INSTITUTE

2007
Ocean Microbial Metagenomics



6.3 Billion Bases
5.6 Million Genes

~3300x

camera
Marine Microbial Ecology

it2

CAMERA: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

National LambdaRail Direct Connect Computation and Storage Complex

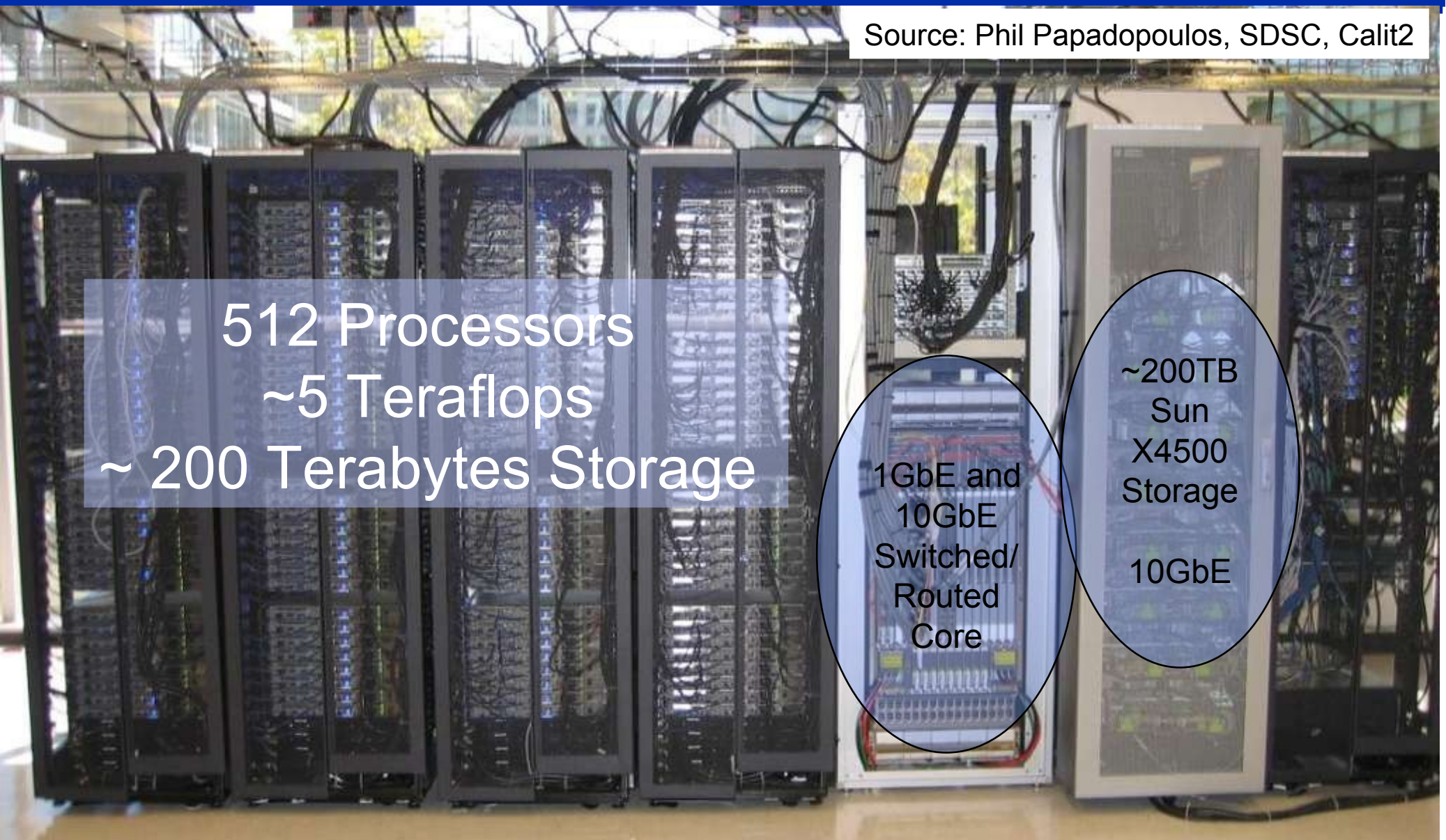
PI Larry Smarr

Funded by: Gordon and Betty Moore Foundation

Joint Partnership of:

Grant Announced January 17, 2006

Calit2 Microbial Metagenomics Cluster- Next Generation Optically Linked Science Data Server



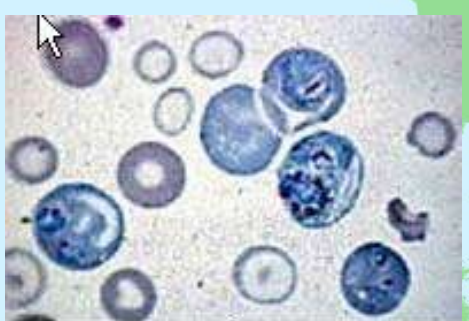
camera Marine Microbial Ecology
SDSC
ROCKS
GORDON AND BETTY MOORE FOUNDATION
J. Craig Venter INSTITUTE
it²

Marine Genome Sequencing Project – CAMERA Anchor Dataset Launched March 13, 2007

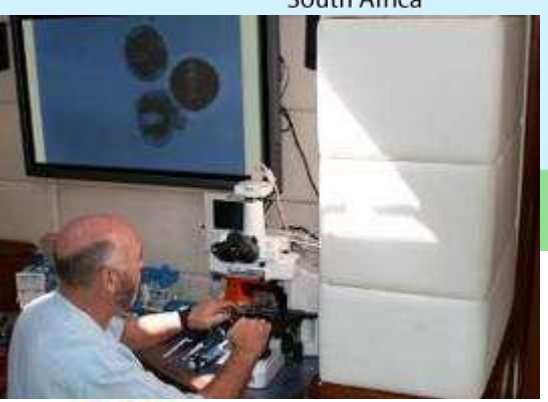


Each Sample ~2000 Microbial Species

Canada
United States
Mexico
Puerto Rico
Honduras
Costa Rica
Panama
Galapagos Islands
x 10 samples
Tanzania/ Zanzibar
x 7 samples
Seychelles
Madagascar
Specify Ocean Data
South Africa
Cocos-Keeling Is.
Vanuatu
x 17 samples
Fiji
Tonga
French Polynesia
Cook Islands
Australia
New Caledonia
x 8 samples



Analyzed
Sampled, but not Analyzed



MARINE MICROBIOLOGY INITIATIVE

camera
Marine Microbial Ecology

Gordon and Betty MOORE FOUNDATION

Measuring the Genetic Diversity of Ocean Microbes

J. Craig Venter INSTITUTE

it²

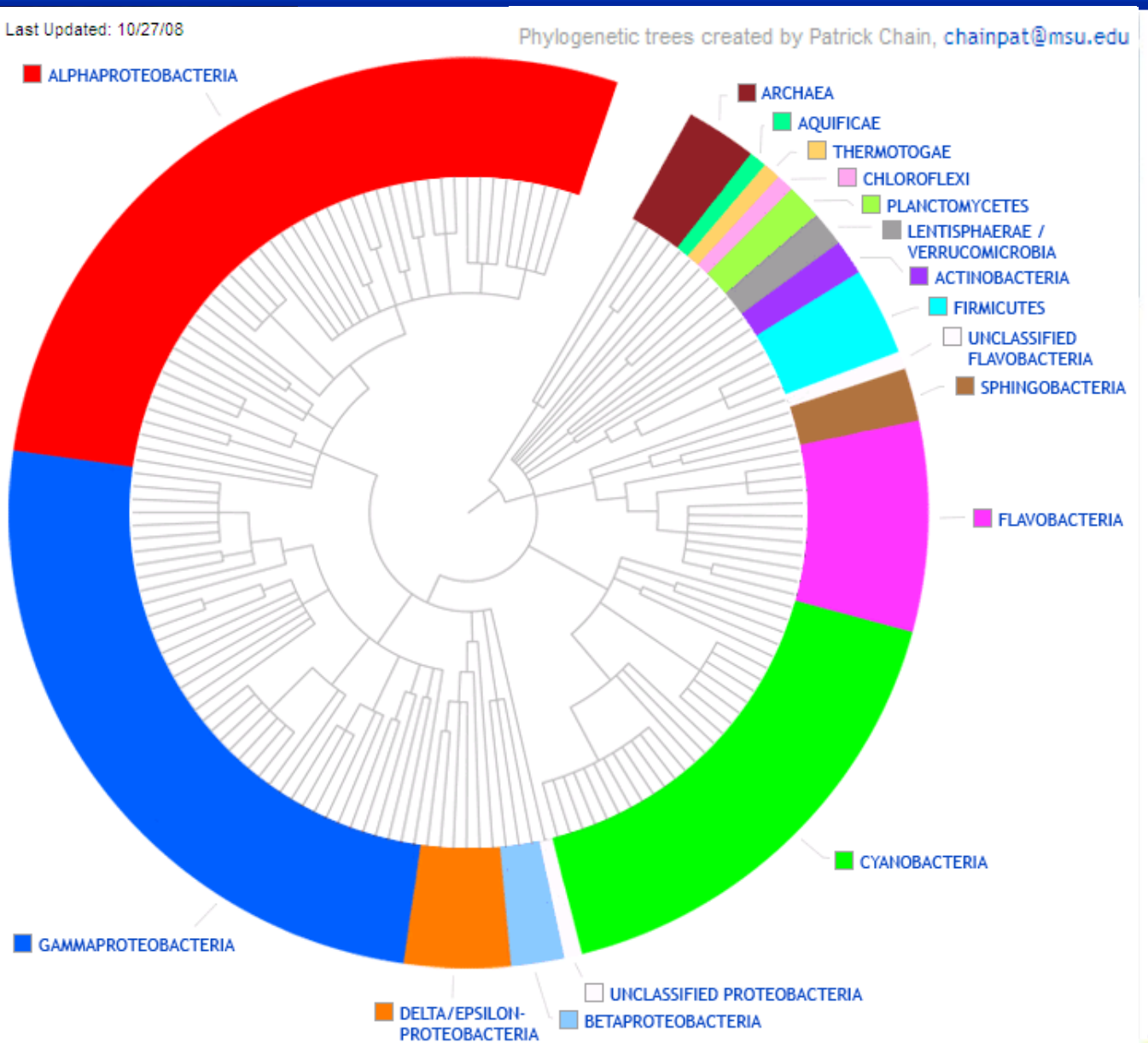
Moore Foundation Enabled the Sequencing of the Full Genome Sequence of 155+ Marine Microbes

GORDON AND BETTY MOORE FOUNDATION

Microb Sequ

J. Craig Venter INSTITUTE





camera Marine Microbial Ecology

www.moore.org/microgenome

it2

CAMERA Houses the Community's Expanding Environmental Metagenomics Datasets

March 16, 2008

Project ▲	Principal Investigator	Organization
Acid Mine Drainage Metagenome	Jillian F. Banfield	DOE Joint Genome Institute
Alvinella pompejana Epibiont Metagenome	Alison Murray	University of Delaware/Desert Research Institute
Chesapeake Bay Virioplankton Metagenome	Eric Wommack	University of Delaware
Dauphin Island Cubitainer Experiment (DICE)	Mary Ann Moran	University of Georgia
Global Ocean Sampling Expedition ★	J. Craig Venter	J. Craig Venter Institute
Marine Bacterioplankton Metagenomes	David Karl	University of Hawaii
Mediterranean Bathypelagic Habitat Metagenome	Francisco Rodriguez-Valera	Evolutionary Genomics Group
Mediterranean Gutless Worm Metagenome	Edward M. Rubin	DOE Joint Genome Institute
Metagenome of Marine NaCl-Saturated Brine	Francisco Rodriguez-Valera	Evolutionary Genomics Group
Metagenomic Analysis of the North Atlantic Spring Bloom	Stephen J. Giovannoni	Oregon State University
Microbial Community Genomics at the HOT/ALOHA	Ed Delong and David Karl	Center for Microbial Oceanography: Research and...
Monterey Bay Microbial Study	Edward F. Delong	Massachusetts Institute of Technology
Moore Marine Microbial Sequencing ★	Various	The Gordon and Betty Moore Foundation
Ocean Viruses	Forest Rohwer	SDSU Center for Universal Microbial Sequencing
Phosphorus Removing (EBPR) Sludge Community	Philip Hugenholtz	Doe Joint Genome Institute
Polyamine Metatranscriptomic Study	Mary Ann Moran	University of Georgia
Sargasso Sea Bacterioplankton Community	Mary Ann Moran	University of Georgia
Surface Water Marine Microbial Community Gene Expression	Jonathan P. Zehr	University of California, Santa Cruz
Viral Metagenomes from Terrestrial Hot Springs	Thomas Schoenfeld	Lucigen Corporation
Waseca County Farm Soil Metagenome	Edward M. Rubin	DOE Joint Genome Institute
Whale Fall Metagenome	Edward M. Rubin	DOE Joint Genome Institute

Rapidly Expanding to Include New Community Datasets
Now Releasing An Additional Dataset Per Week!

camera
Marine Microbial Ecology

GORDON AND BETTY MOORE FOUNDATION

J. Craig Venter INSTITUTE

it²

Current CAMERA Interface
February 19, 2010

camera
Marine Microbial Ecology

Log in Register Contact Help

Search

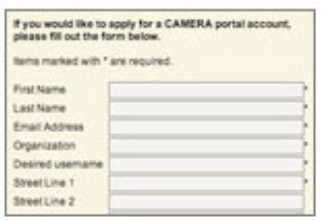
Advanced Search

About CAMERA ▾ Education ▾ Community ▾ Tools ▾ Projects & Data ▾

Analyze data
Run BLAST, Job Results, FRV...



Register
get an account ...



Download datasets
59 datasets available



Learn about ...
CAMERA / metagenomics



Community
get involved



Need Help?
User guides, FAQ



Current Status

- Current application version - 1.3.2.29 (release notes) deployed May 11, 2009
- Available datasets: 59 Last update: Feb 12, 2010
- 3387 registered users as of 02/19/10 (see registration map)

Upcoming events

- RECOMB 2010 - 14th Annual International Conference on Research in Computational Molecular Biology

camera
Marine Microbial Ecology

http://camera.calit2.net/

GORDON AND BETTY MOORE FOUNDATION

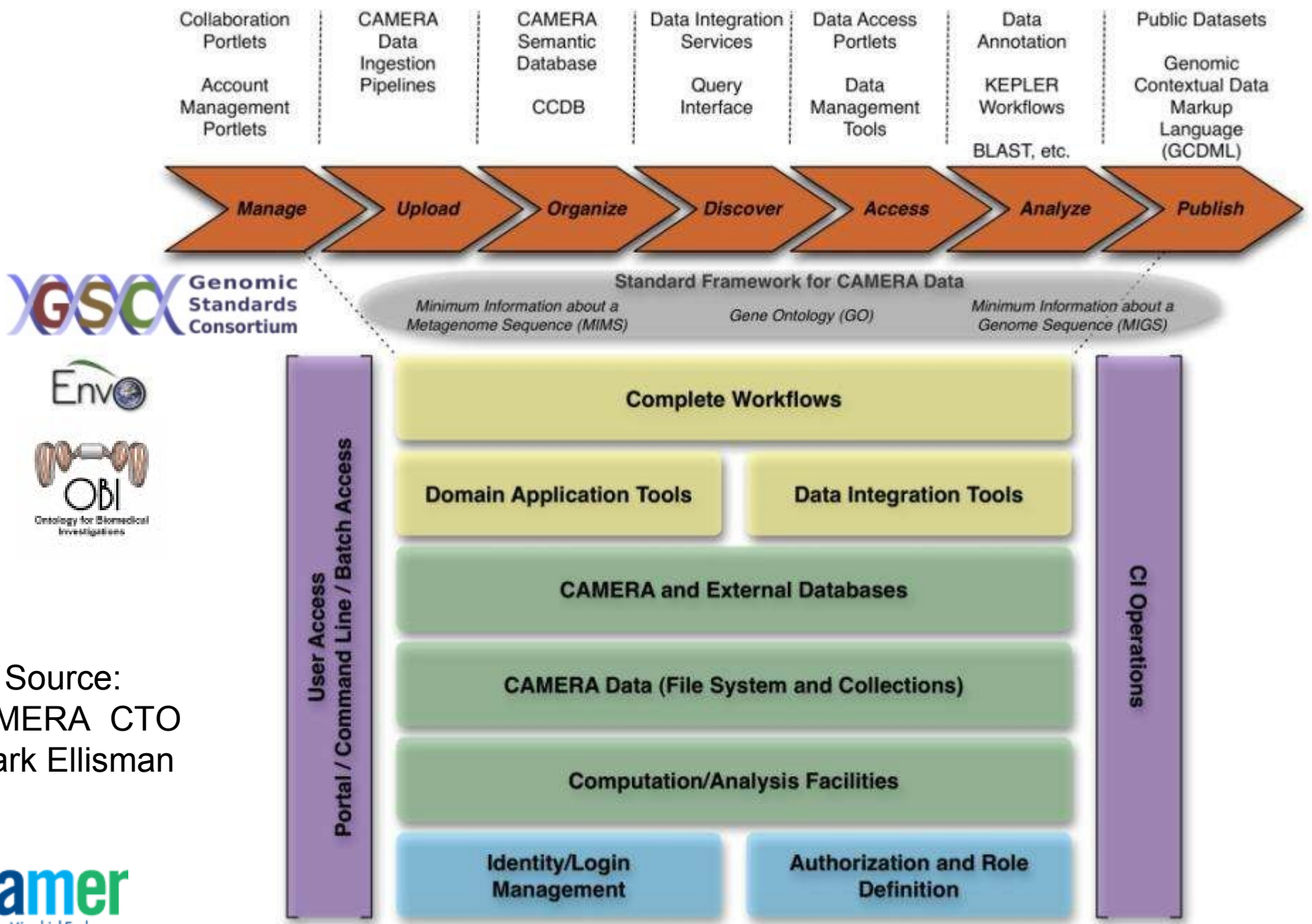
it2

The CAMERA Project Has Established a Global Marine Microbial Metagenomics Cyber-Community



3387 Registered Users
From Over 75 Countries

Creating CAMERA 2.0 -
Advanced Cyberinfrastructure Service Oriented Architecture



Source:
CAMERA CTO
Mark Ellisman

camer
Marine Microbial Ecology

Metagenomic Data Ingestion Growing Rapidly!

	Number of reads	Number of base pairs
CAMERA 1st release (Mar. 2006)	8.23m	8.67b
CAMERA 1.3 (Dec. 2008)	13.42m	12.35b
CAMERA (Jul. 2009)	36.97m	19.27b
CAMERA * (Dec. 2009)	47.87m	22.08b

* All the reference datasets including newly released "All NCBI Environmental Samples (ENV_NT) were not counted

camera
Marine Microbial Ecology

it2

Prototyping a Data Acquisition Pipeline: A New Data Submission Paradigm-Metadata First!

Source: Paul Gilna, Calit2

Solexa and SOLiD Next!

Investigator submits proposal to GBMF

Investigator submits metadata to CAMERA

Metadata now collected *before* sequence data: GSC-compliant

CAMERA sends acknowledgement to Investigator, Seq. Group, GBMF

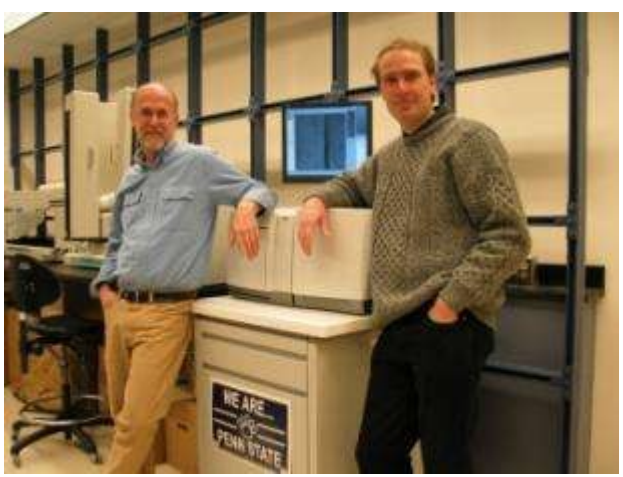
Project-ID serves as acceptance-proof

Seq. Group send barcoded sample "kit" to investigators

Sample is Received and Sequenced

Seq. Group Upload data to CAMERA (& Investigator)

Data & Metadata Released in six months



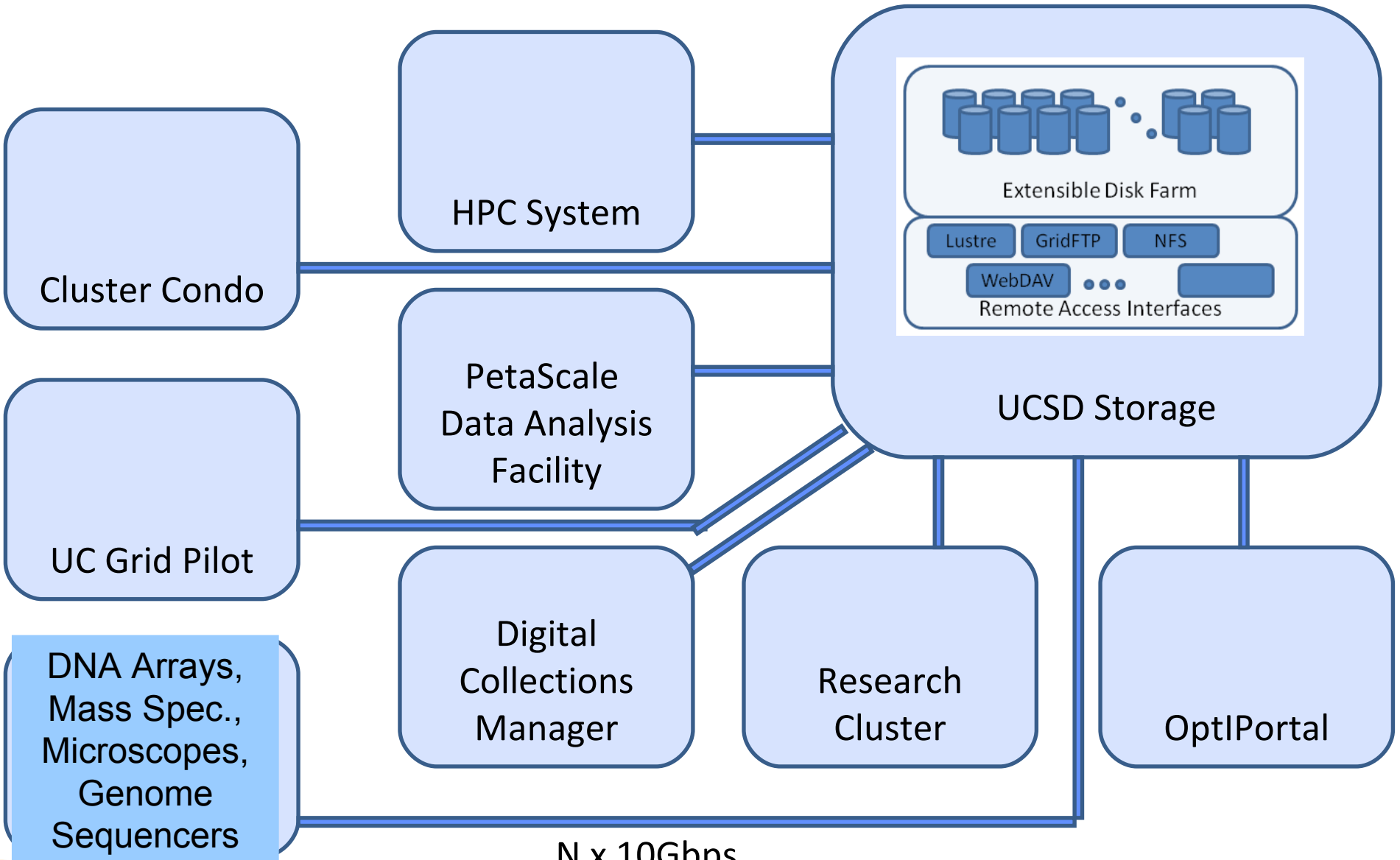
camera
Marine Microbial Ecology

Webb Miller and Stephan C. Schuster, and Roche / 454 Genome Sequencer

GORDON AND BETTY MOORE FOUNDATION

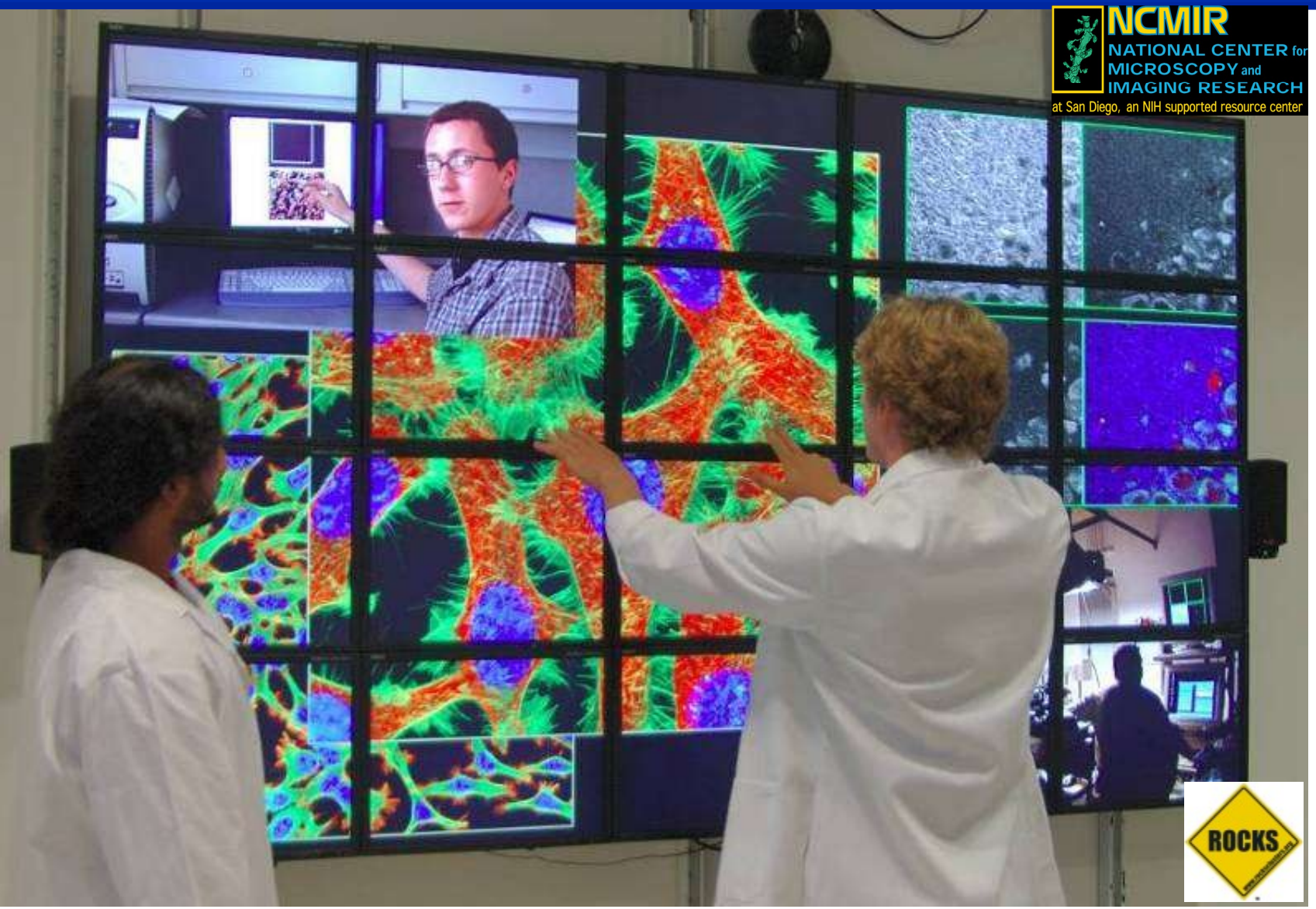
it2

Conceptual Architecture to Physically Connect Campus Resources Using Fiber Optic Networks



Source:Phil Papadopoulos, SDSC/Calit2

The OptIPuter Project: Creating High Resolution Portals Over Dedicated Optical Channels to Global Science Data



Scalable Adaptive Graphics Environment (SAGE)

evl electronic visualization laboratory

NSF

Now in Sixth and Final Year

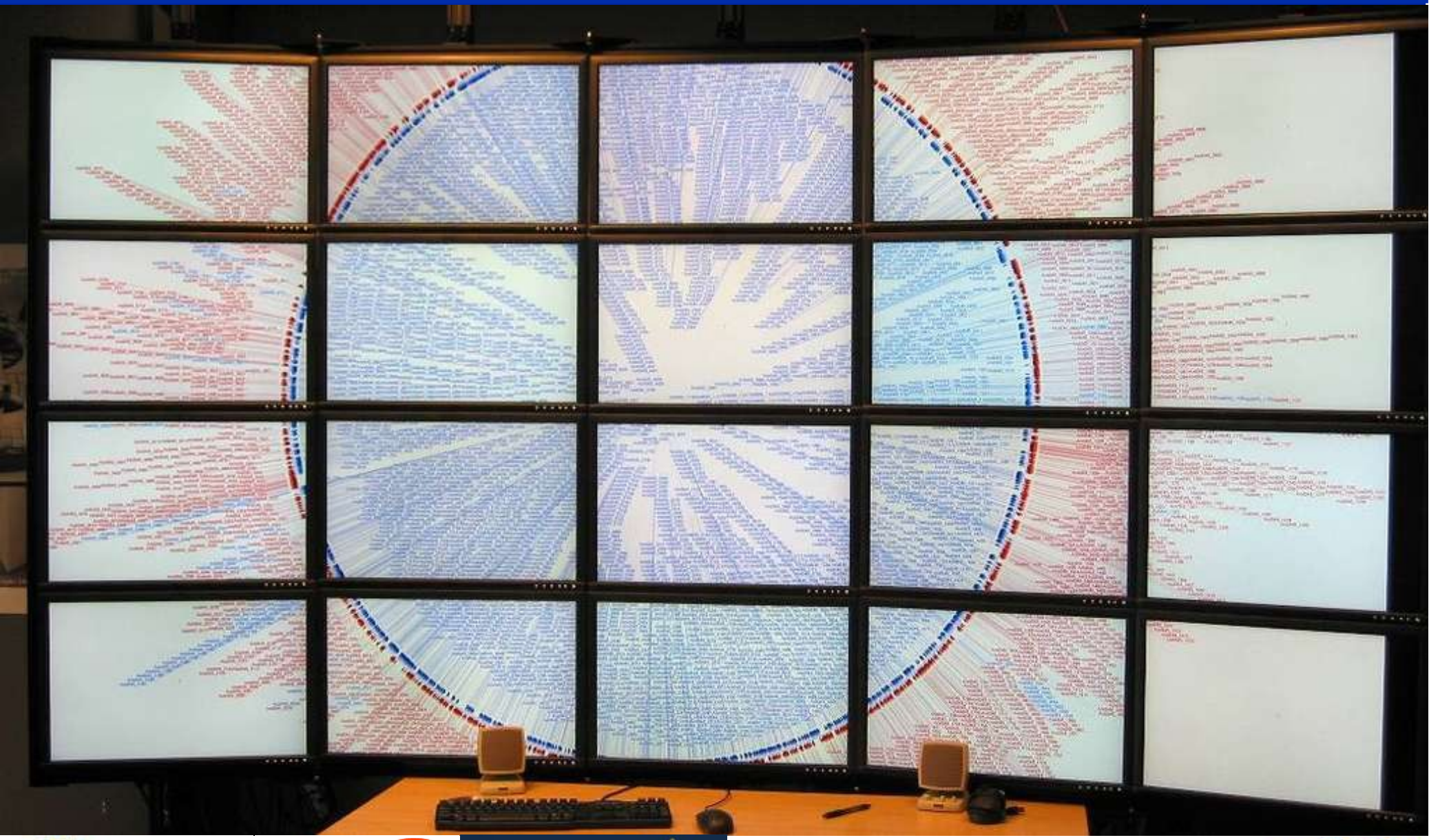
Picture Source: Mark Ellisman, David Lee, Jason Leigh

camera Marine Microbial Ecology

Calit2 (UCSD, UCI), SDSC, and UIC Leads—Larry Smarr PI
Univ. Partners: NCSA, USC, SDSU, NW, TA&M, UvA, SARA, KISTI, AIST
Industry: IBM, Sun, Telcordia, Chiaro, Calient, Glimmerglass, Lucent

it2

Visual Analytics--Use of Tiled Display Wall OptIPortal to Interactively View Microbial Genome (5 Million Bases)



camera Marine Microbial Ecology | OptIPuter | Powered by Rocks ROCKS

Acidobacteria bacterium Ellin345 Soil Bacterium 5.6 Mb; ~5000 Genes

Source: Raj Singh, UCSD

it2

Use of Tiled Display Wall OptIPortal to Interactively View Microbial Genome



camera Marine Microbial Ecology

OptIPuter

Powered by Rocks ROCKS

Source: Raj Singh, UCSD

it²

Use of Tiled Display Wall OptIPortal to Interactively View Microbial Genome



camera Marine Microbial Ecology

OptIPuter

Powered by Rocks ROCKS

Source: Raj Singh, UCSD

it2

MIT's Ed DeLong and Darwin Project Team Using OptIPortal to Analyze 10km Ocean Microbial Simulation



GORDON AND BETTY MOORE FOUNDATION

camera
Marine Microbial Ecology

cross-disciplinary research at MIT, connecting systems biology, microbial ecology, global biogeochemical cycles and climate

it²

Prototyping Next Generation User Access and Analysis- Between Calit2 and U Washington



OptIPuter

Ginger Armbrust's Diatoms: Micrographs, Chromosomes, Genetic Assembly

ROCKS



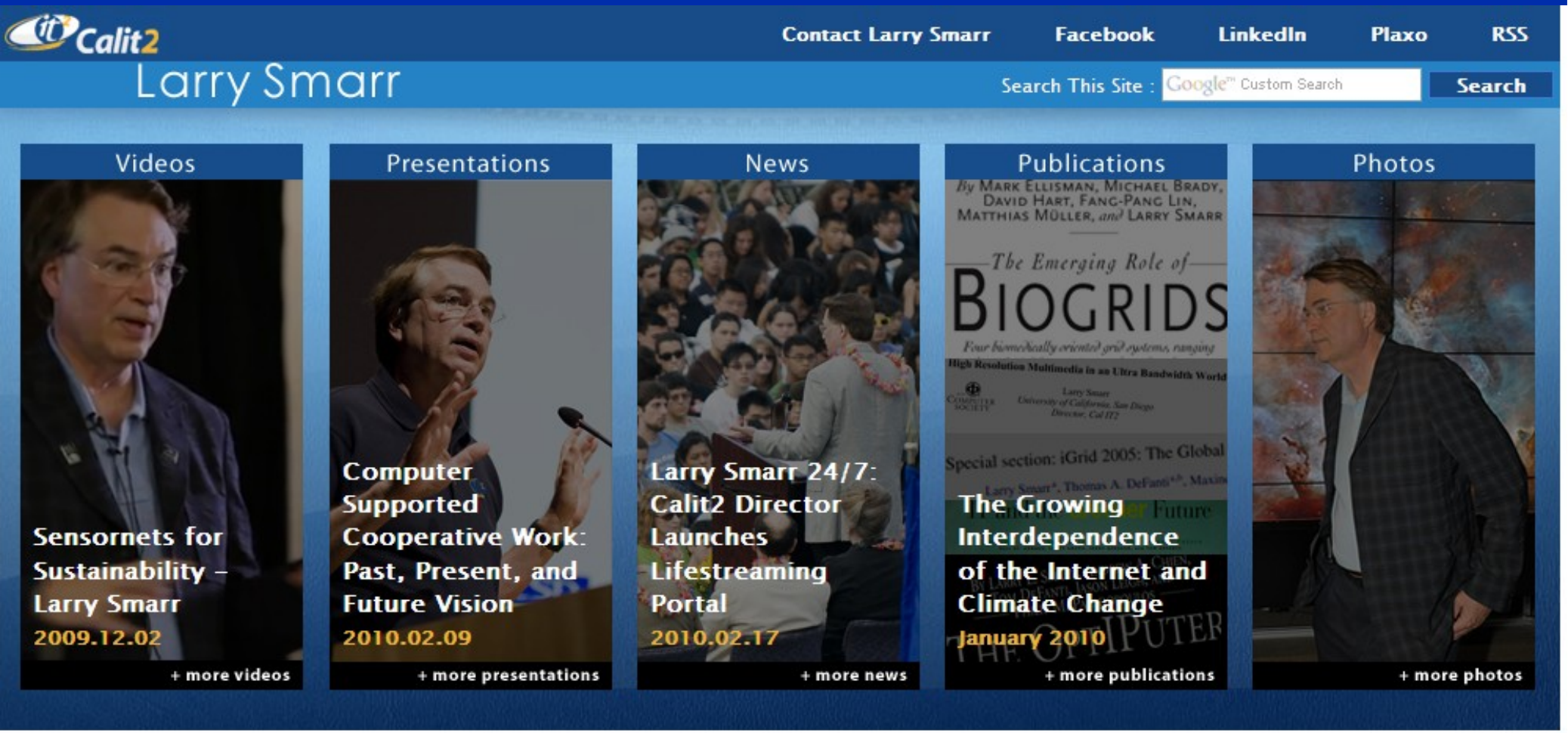


NSF

camera
Marine Microbial Ecology

it2

You Can Download This Presentation at lsmarr.calit2.net



ABOUT LARRY SMARR

Larry Smarr is the founding Director of the California Institute for Telecommunications and Information Technology (Calit2), a UC San Diego/UC Irvine partnership, and holds the Harry E. Gruber professorship in Computer Science and Engineering

WHAT I'M READING

Complete List of Larry's Recommendations

On avoiding dangerous anthropogenic interference with the climate system:

RECENT TWITTER POSTS

Twitter? Click To Follow Me

lsmarr: #AAAS10 Maslowski