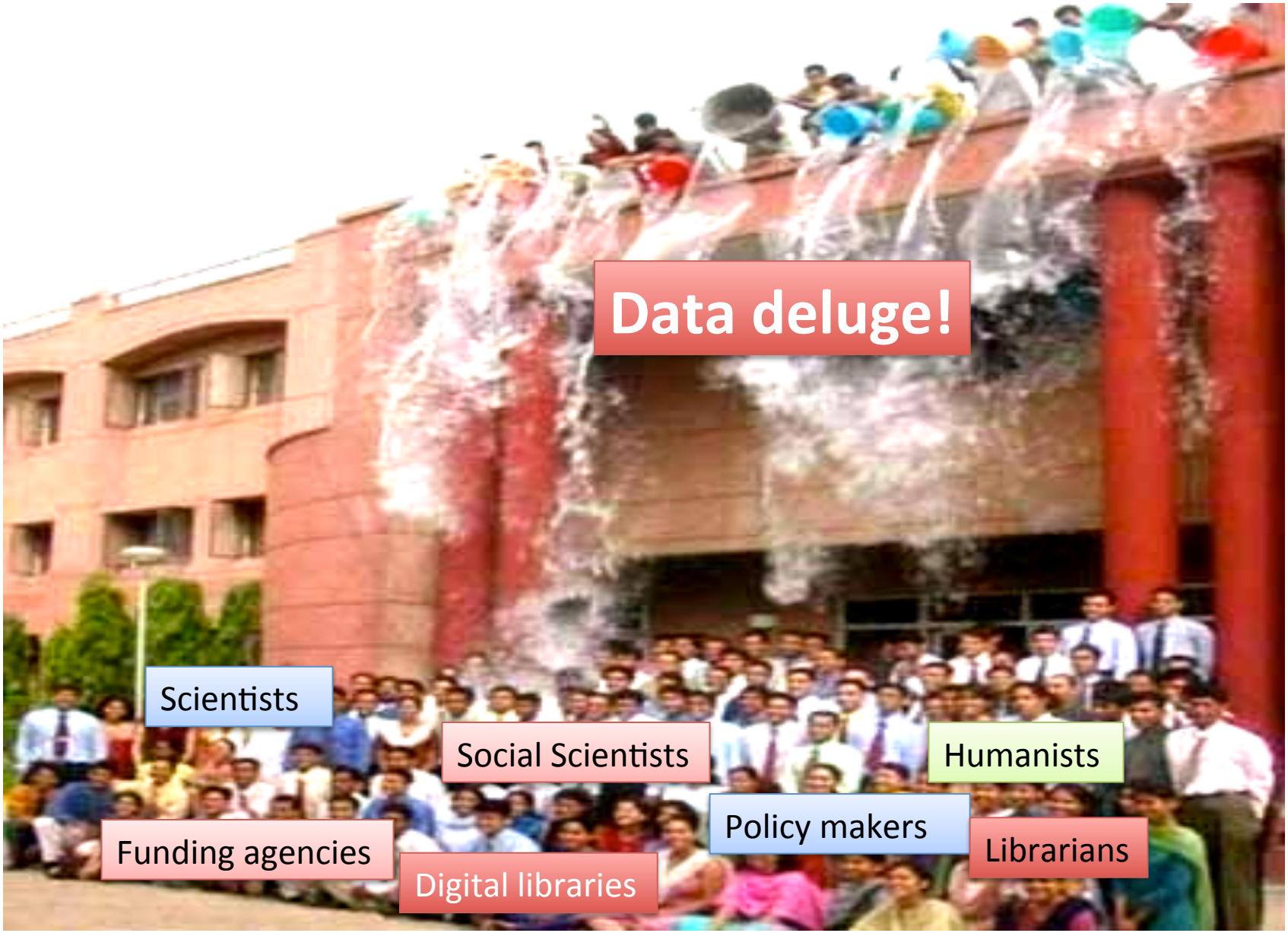
Big data and the long tail: Use and reuse of little data

Christine L. Borgman
Oliver Smithies Visiting Fellow and Lecturer, Balliol College, Oxford
Visiting Fellow, Oxford eResearch Centre
Visiting Fellow, Oxford Internet Institute

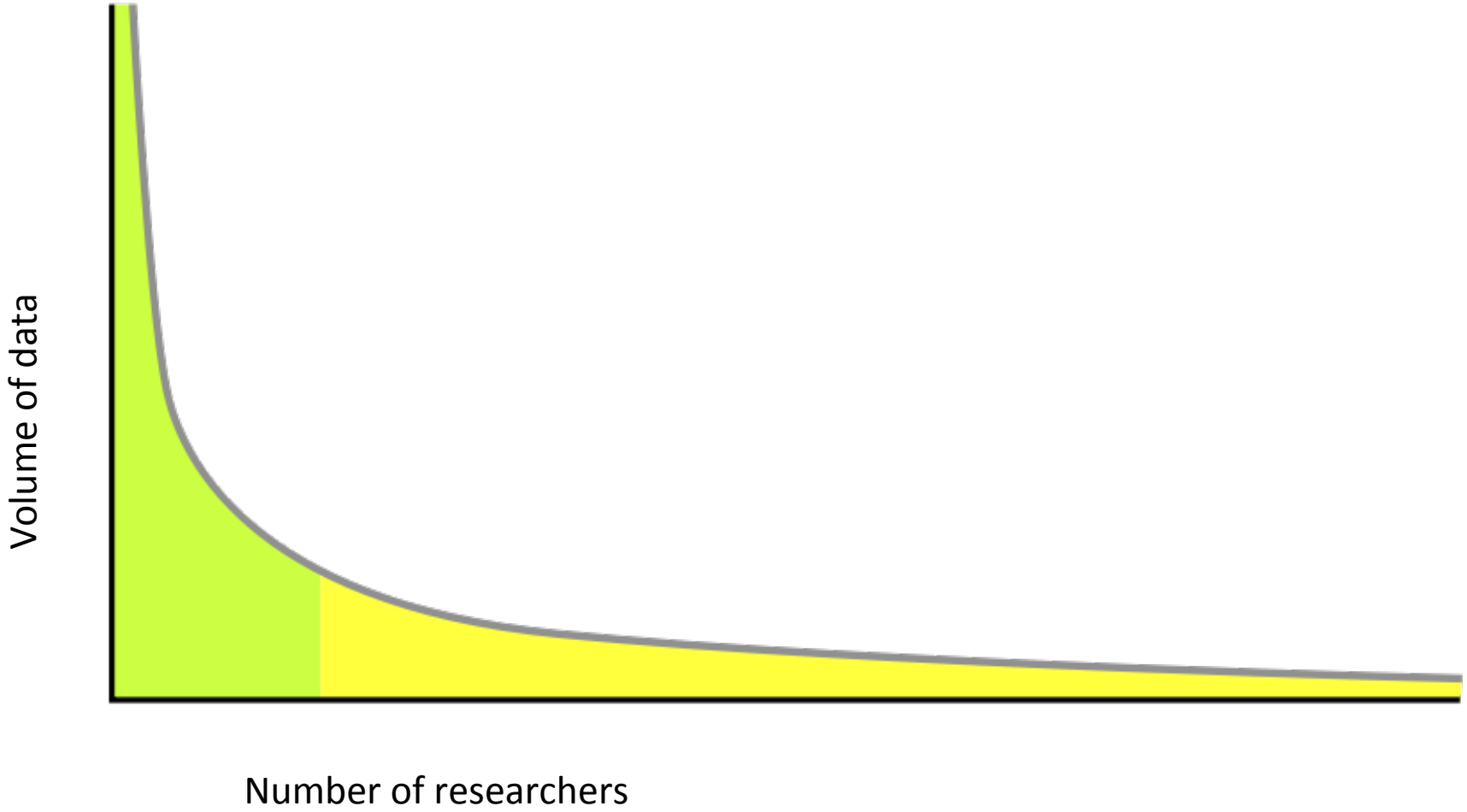
Professor and Presidential Chair in Information Studies
University of California, Los Angeles

Oxford eResearch Centre Talk, 12 March 2013



http://www.guzer.com/pictures/suprise_suprise.jpg

The long tail of data



Empowering Long Tail Research

A study funded by the National Science Foundation

Big science. Small labs.

The challenge of "big data" is felt by everyone, from international teams to "teams of one." Small laboratories need services and tools to help them make full use—*and be good stewards*—of the valuable research data they collect and create.

Why small labs?

Who are we?

Contact us

Receive our monthly newsletter

Privacy by SafeSubscribe℠

Recent announcements

Why we like SaaS Why do we think SaaS is so important to an institute for empowering research in small labs?As our team contemplated how to empower research in small labs, we realized ...
Posted Feb 1, 2013, 3:13 PM by Lee Liming

Principal investigators

Christine Borgman, UCLA Dept. of Information Studies
Ian Foster, U. of Chicago Computation Institute
Bryan Heidorn, U. of Arizona School of Information Resources and Library Science
Bill Howe, U. of Washington Computer Science & Engineering
Carl Kesselman, USC Information Sciences Institute

The Conundrum of Sharing Research Data

*If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.**

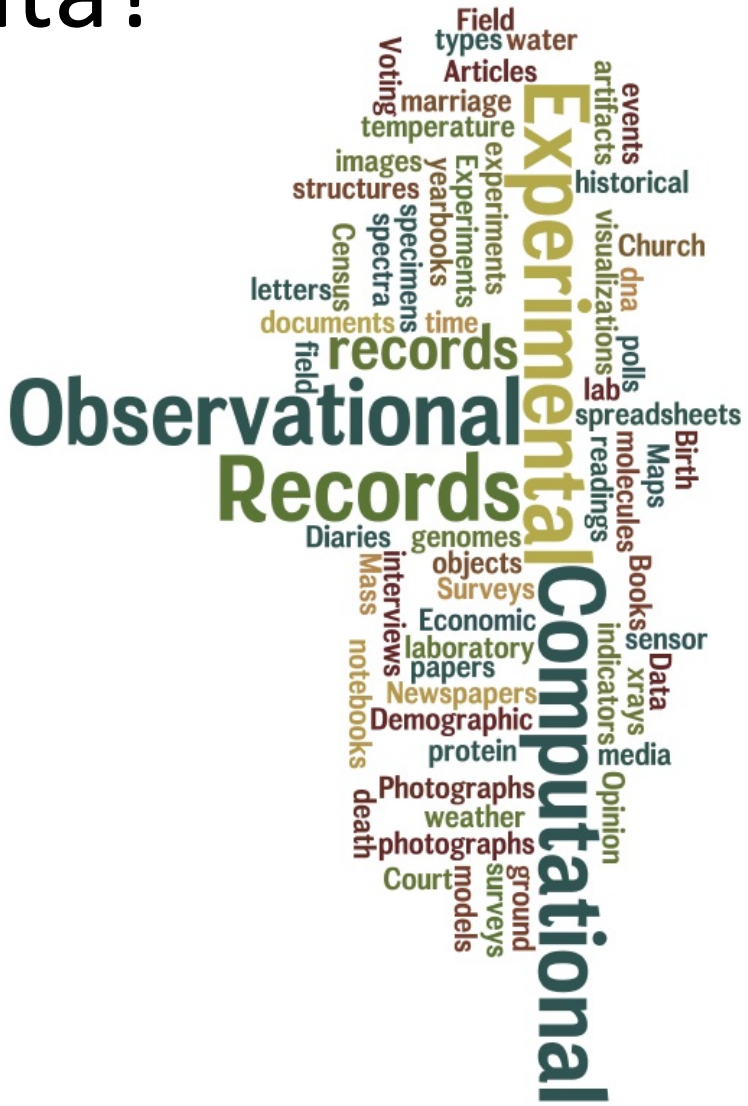
*Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

http://www.tzanis.org/tzanisblog/archives/images/push-pull-thumb.jpg

Long-Lived Digital Data Collections. (2005). National Science Board. http://www.nsf.gov/pubs/2005/nsb0540/

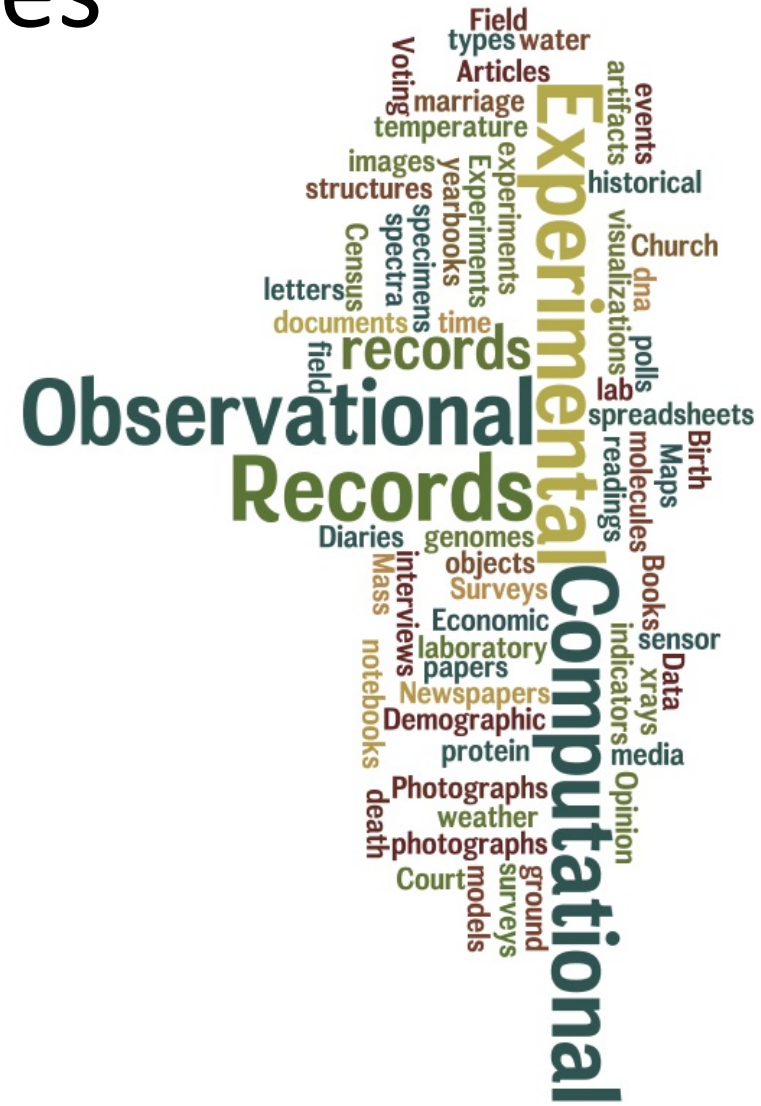
What are data?

- Size matters
  - Big science, little science
  - Big data, long tail
- Origins of data
  - Sources and resources
  - External factors
  - Purposes for collecting data
- Processing of data
  - Metadata
  - Provenance
  - Handling data



Case studies

- Sciences
  - Astronomy
  - Sensor networks
- Social sciences
  - Social networks: Twitter
  - Qualitative interviews
- Humanities
  - Buddhist philology
  - Distributed collections




Copyright Sydney

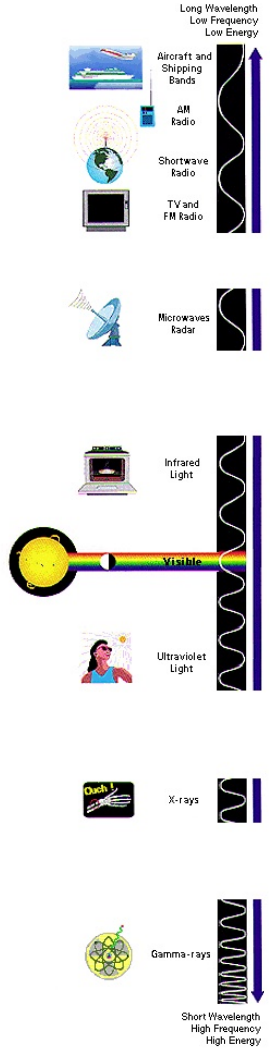
Astronomy: An Info Perspective

- Size matters
  - Big science, little science
  - Big data, long tail
- Origins of data
  - Sources and resources
  - External factors
  - Purposes for collecting data
- Processing of data
  - Metadata
  - Provenance
  - Handling data

NASA Astronomy Picture of the Day

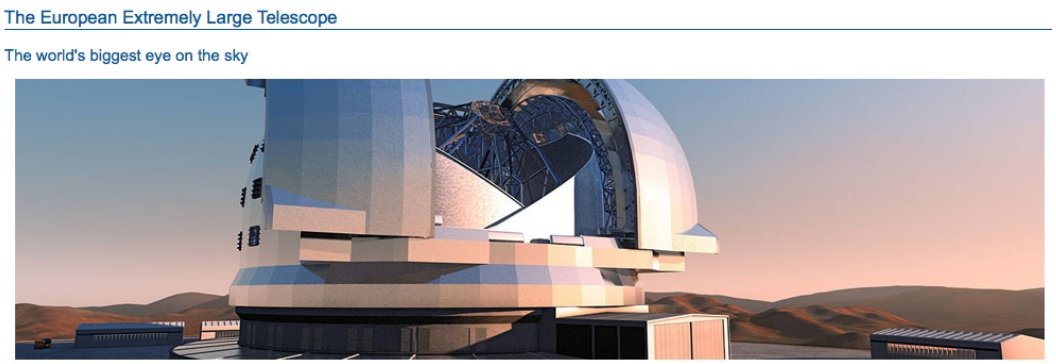
Astronomy Sources and Resources

- Phenomena of interest: Celestial objects
- Organizing principles
  - Coordinates on the sky
  - Electromagnetic spectrum
- Data acquisition
  - Observations of the sky
    - Telescopes
    - Instruments
  - Output of computational models



Astronomy Sources and Resources

- Data disposition:
  - Repositories / archives
    - Scientific mission
    - Spectrum
    - Region
    - Time
  - Desktop / lab servers
    - Observing proposals
    - Computational models



Purposes for collecting astronomy data

COMPLETE

The COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions

Project Description

The COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions (COMPLETE) provides a range of data complementary to the Spitzer Legacy Program "From Molecular Cores to Planet Forming Disks" (c2d) for the Perseus, Ophiuchus and Serpens regions. In combination with the Spitzer observations, COMPLETE will allow for detailed analysis and understanding of the physics of star formation on scales from 500 A.U. to 10 pc.

Phase I, which is now complete, provides fully sampled, arcminute resolution observations of the density and velocity structure of the three regions, comprising: extinction maps derived from the Two Micron All Sky Survey (2MASS) near-infrared data using the NICER algorithm; extinction and temperature maps derived from IRAS 60 and 100um emission; HI maps of atomic gas; 12CO and 13CO maps of molecular gas; and submillimeter continuum images of emission from dust in dense cores.

Click on the "Data" button to the left to access this data.

Phase II (which is still ongoing) uses targeted source lists based on the Phase I data, as it is (still) not feasible to cover every dense star-forming peak at high resolution. Phase II includes high-sensitivity near-IR imaging (for high resolution extinction mapping), mm-continuum imaging with MAMBO on IRAM and high-resolution observations of dense gas tracers such as N2H+. These data are being released as they are validated.

COMPLETE Movies: Check-out our movies page for animations of the COMPLETE data cubes in 3D.

Referencing Data from the COMPLETE Survey

COMPLETE data are non-proprietary. Please reference **Ridge, N.A. et al., *"The COMPLETE Survey of Star Forming Regions: Phase 1 Data"*, 2006, AJ, 131, 2921** as the data source. However, we would like to keep a record of work that is using COMPLETE data, so please send us an email (with a reference if possible) if you make use of any data provided here.

Recent COMPLETE Publications

1. NEW Arce, Hector G.; Borkin, Michelle A.; Goodman, Alyssa A.; Pineda, Jaime E.; Beaumont, Christopher N., 2011, *A Bubbling Nearby Molecular Cloud: COMPLETE Shells in Perseus* ; Submitted for review. (astro-ph)
2. NEW Pineda, Jaime E.; Goodman, Alyssa A.; Arce, Hector; G.; Caselli, Paola; Longmore, Steven; Corder, Stuartt, 2011, ApJ 739,

http://www.cfa.harvard.edu/COMPLETE/

Vol 457|1 January 2009 doi:10.1038/nature07609 nature

LETTERS

A role for self-gravity at multiple length scales in the process of star formation

Alyssa A. Goodman[1,2], Erik W. Rosolowsky[2,3], Michelle A. Borkin[1]†, Jonathan B. Foster[2], Michael Halle[1,4], Jens Kauffmann[1,2] & Jaime E. Pineda[2]

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function[2]. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by ^{13}CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission[3] are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their existence. Turbulent fragmentation simulations without self-gravity—even of unmagnetized isothermal material—can yield mass and velocity power spectra very similar to what is observed in clouds like L1448. But a dendrogram of such a simulation[4] shows that nearly all the gas in it (much more than in the observations) appears to be self-gravitating. A potentially significant role for gravity in 'non-self-gravitating' simulations suggests inconsistency in simulation assumptions and output, and that it is necessary to include self-gravity in any realistic simulation of the star-formation process on subparsec scales.

Spectral-line mapping shows whole molecular clouds (typically tens to hundreds of parsecs across, and surrounded by atomic gas) to be marginally self-gravitating[5]. When attempts are made to further break down clouds into pieces using 'segmentation' routines, some self-gravitating structures are always found on whatever scale is sampled[6,7]. But no observational study to date has successfully used one spectral-line data cube to study how the role of self-gravity varies as a function of scale and conditions, within an individual region.

Most past structure identification in molecular clouds has been explicitly non-hierarchical, which makes difficult the quantification of physical conditions on multiple scales using a single data set. Consider, for example, the often-used algorithm CLUMPFIND[7]. In three-dimensional (3D) spectral-line data cubes, CLUMPFIND operates as a watershed segmentation algorithm, identifying local maxima in the position–position–velocity (p–p–v) cube and assigning nearby emission to each local maximum. Figure 1 gives a two-dimensional (2D) view of L1448, our sample star-forming region, and Fig. 2 includes a CLUMPFIND decomposition of it based on ^{13}CO observations. As with any algorithm that does not offer hierchically nested or overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

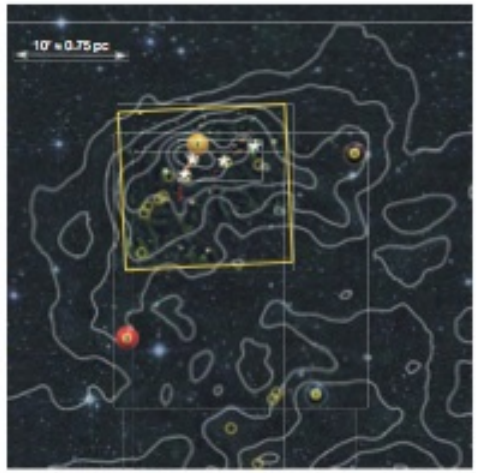


Figure 1 | Near-infrared image of the L1448 star-forming region with contours of molecular emission overlaid. The channels of the colour image correspond to the near-infrared bands J (blue), H (green) and K (red), and the contours of integrated intensity are from ^{13}CO(1–0) emission[8]. Integrated intensity is monotonically, but not quite linearly (see Supplementary Information), related to column density[14], and it gives a view of 'all' of the molecular gas along lines of sight, regardless of distance or velocity. The region within the yellow box immediately surrounding the protostars has been imaged more deeply in the near-infrared (using Calar Alto) than the remainder of the box (2MASS data only), revealing protostars as well as the scattered starlight known as 'Cloudshine'[11] and outflows (which appear orange in this colour scheme). The four billiard-ball labels indicate regions containing self-gravitating dense gas, as identified by the dendrogram analysis, and the leaves they identify are best shown in Fig. 2a. Asterisks show the locations of the four most prominent embedded young stars or compact stellar systems in the region (see Supplementary Table 1), and yellow circles show the millimetre-dust emission peaks identified as star-forming or 'pre-stellar' cores[3].

[1]Initiative in Innovative Computing at Harvard, Cambridge, Massachusetts 02138, USA. [2]Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. [3]Department of Physics, University of British Columbia, Okanagan, Kelowna, British Columbia V1V 1V7, Canada. [4]Surgical Planning Laboratory and Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

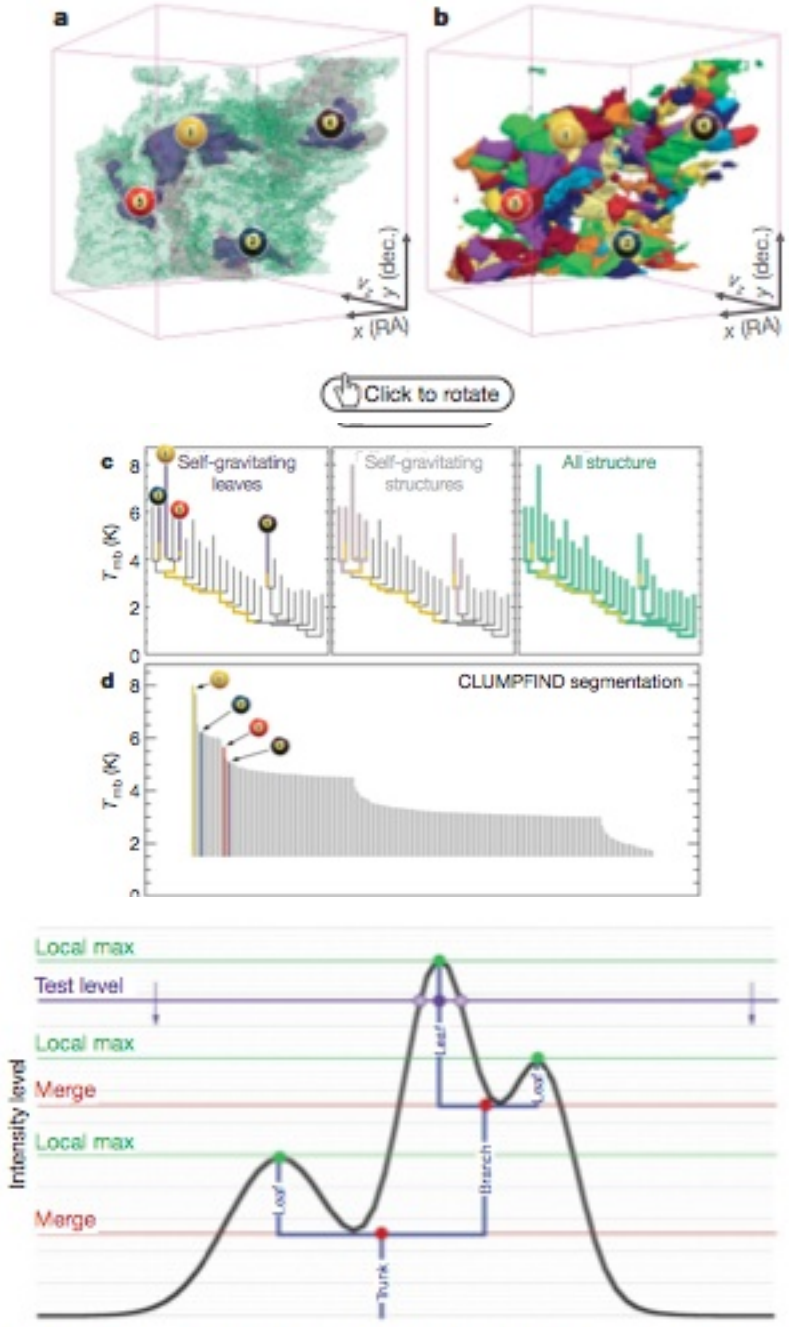


Figure 3 | Schematic illustration of the dendrogram process. Shown is the

Processing data for COMPLETE

- Survey of extant data
  - Three star forming regions
  - Multiple scientific missions
  - Across electromagnet spectrum
- New data from observing proposals
  - Pipeline processing
  - Calibrate, standardize, validate
- Metadata
  - Identify and retrieve data from archives
  - Create metadata for new data
  - Reconcile units and coordinate systems
- Provenance
  - Document data sources
  - Document team workflow

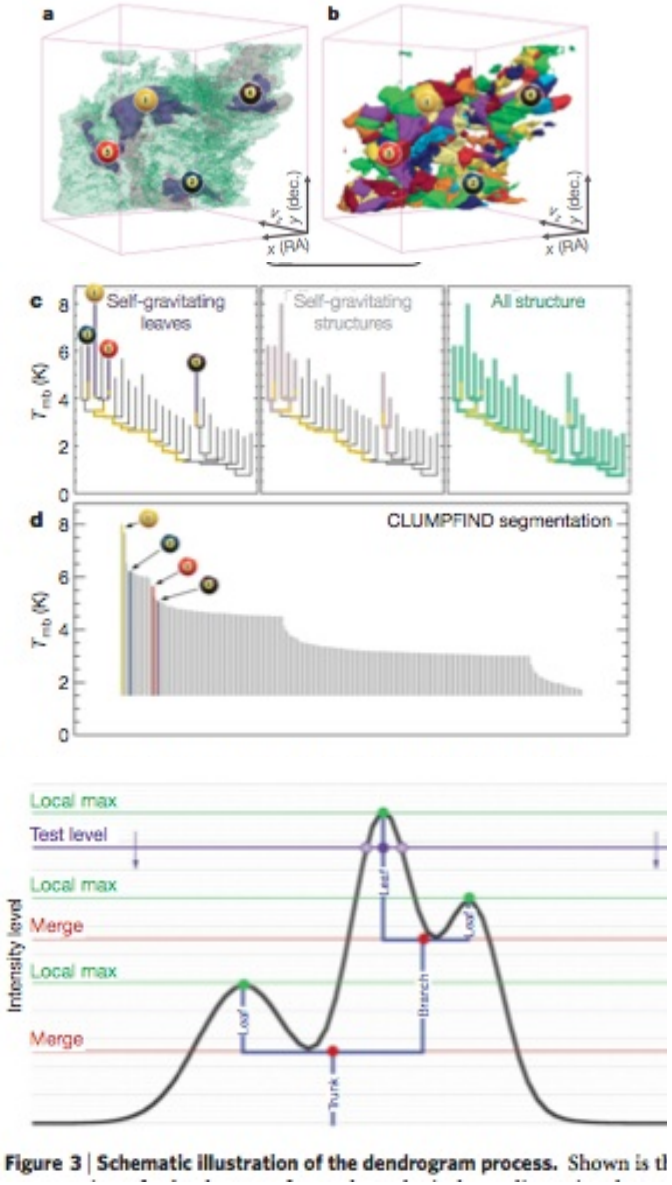
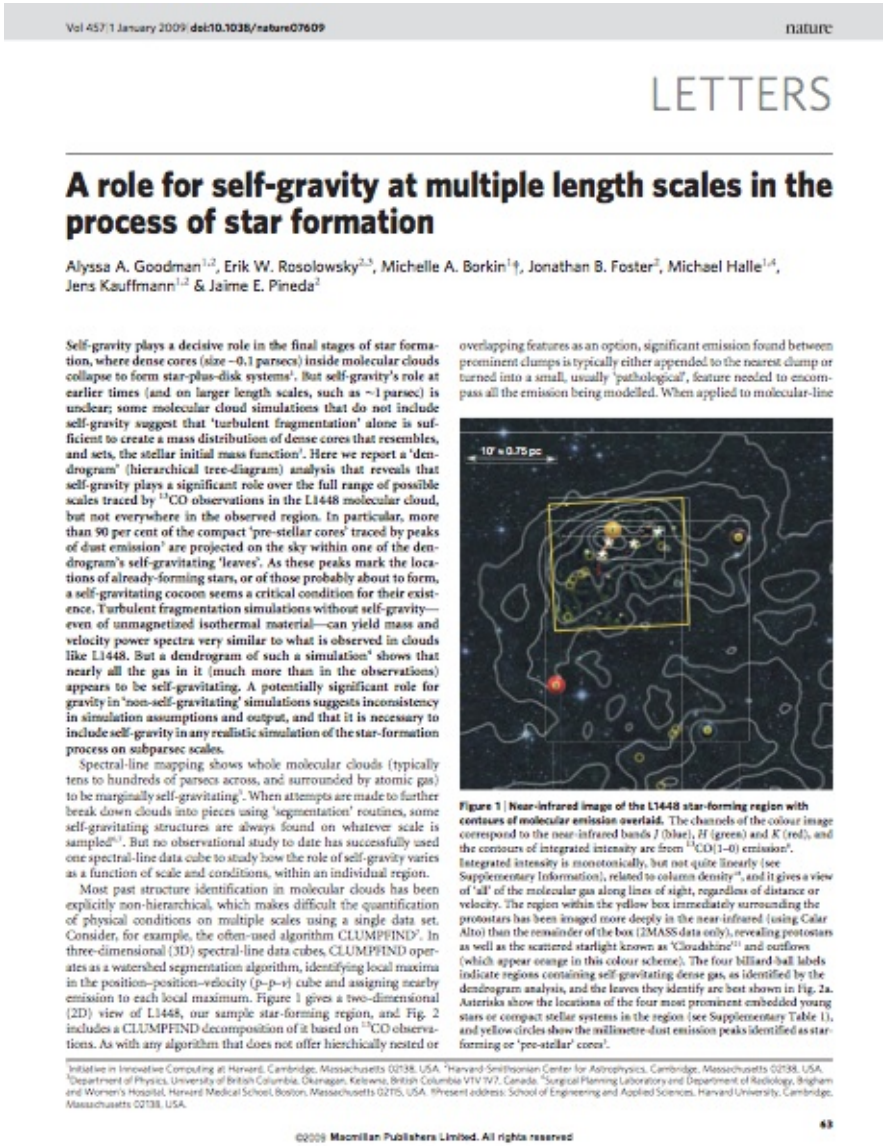


Figure 3 | Schematic illustration of the dendrogram process. Shown is tl

Handling data for COMPLETE

- People involved
  - Multiple scientific missions
  - Pipeline processing
  - CfA team to build COMPLETE
- Collecting the data
  - COMPLETE
  - Other sources as needed
- Processing the data
  - Extant tools
  - New analytical methods
  - New visualization methods
- Releasing the data
  - COMPLETE publicly available
  - Other CfA team data in DataVerse



Vol 457|1 January 2009 doi:10.1038/nature07609 nature

LETTERS

A role for self-gravity at multiple length scales in the process of star formation

Alyssa A. Goodman[1,2], Erik W. Rosolowsky[2,3], Michelle A. Borkin[1]†, Jonathan B. Foster[2], Michael Halle[1,4], Jens Kauffmann[1,2] & Jaime E. Pineda[2]

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function[2]. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by ^{13}CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission[3] are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their existence. Turbulent fragmentation simulations without self-gravity—even of unmagnetized isothermal material—can yield mass and velocity power spectra very similar to what is observed in clouds like L1448. But a dendrogram of such a simulation[4] shows that nearly all the gas in it (much more than in the observations) appears to be self-gravitating. A potentially significant role for gravity in 'non-self-gravitating' simulations suggests inconsistency in simulation assumptions and output, and that it is necessary to include self-gravity in any realistic simulation of the star-formation process on subparsec scales.

Spectral-line mapping shows whole molecular clouds (typically tens to hundreds of parsecs across, and surrounded by atomic gas) to be marginally self-gravitating[5]. When attempts are made to further break down clouds into pieces using 'segmentation' routines, some self-gravitating structures are always found on whatever scale is sampled[6,7]. But no observational study to date has successfully used one spectral-line data cube to study how the role of self-gravity varies as a function of scale and conditions, within an individual region.

Most past structure identification in molecular clouds has been explicitly non-hierarchical, which makes difficult the quantification of physical conditions on multiple scales using a single data set. Consider, for example, the often-used algorithm CLUMPFIND[7]. In three-dimensional (3D) spectral-line data cubes, CLUMPFIND operates as a watershed segmentation algorithm, identifying local maxima in the position–position–velocity (p–p–v) cube and assigning nearby emission to each local maximum. Figure 1 gives a two-dimensional (2D) view of L1448, our sample star-forming region, and Fig. 2 includes a CLUMPFIND decomposition of it based on ^{13}CO observations. As with any algorithm that does not offer hierchically nested or overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

Figure 1 | Near-infrared image of the L1448 star-forming region with contours of molecular emission overlaid. The channels of the colour image correspond to the near-infrared bands J (blue), H (green) and K (red), and the contours of integrated intensity are from ^{13}CO(1–0) emission[8]. Integrated intensity is monotonically, but not quite linearly (see Supplementary Information), related to column density[14], and it gives a view of 'all' of the molecular gas along lines of sight, regardless of distance or velocity. The region within the yellow box immediately surrounding the protostars has been imaged more deeply in the near-infrared (using Calar Alto) than the remainder of the box (2MASS data only), revealing protostars as well as the scattered starlight known as 'Cloudshine'[21] and outflows (which appear orange in this colour scheme). The four billiard-ball labels indicate regions containing self-gravitating dense gas, as identified by the dendrogram analysis, and the leaves they identify are best shown in Fig. 2a. Asterisks show the locations of the four most prominent embedded young stars or compact stellar systems in the region (see Supplementary Table 1), and yellow circles show the millimetre-dust emission peaks identified as star-forming or 'pre-stellar' cores[3].

[1]Initiative in Innovative Computing at Harvard, Cambridge, Massachusetts 02138, USA. [2]Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. [3]Department of Physics, University of British Columbia, Okanagan, Kelowna, British Columbia V1V 1V7, Canada. [4]Surgical Planning Laboratory and Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

What are astronomy data?

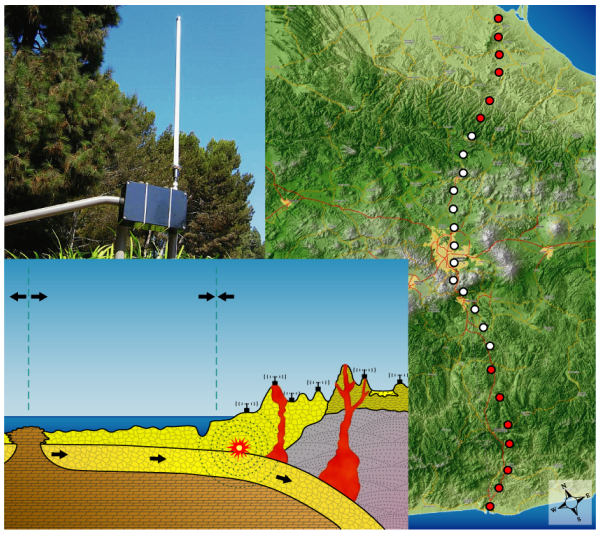
- Scientific knowledge required to
  - Reconcile observations from multiple missions
  - Write new proposals to gather observations
  - Study celestial objects and phenomena
    - Disparate instruments
    - Disparate scientific missions
    - Disparate metadata
    - Distinct user interfaces and data models
- Scientific contributions
  - Creating new data product from extant and new sources
  - Bringing diverse data to bear on extant problem
  - Innovating in methods and visualization

Sensor Networked Science

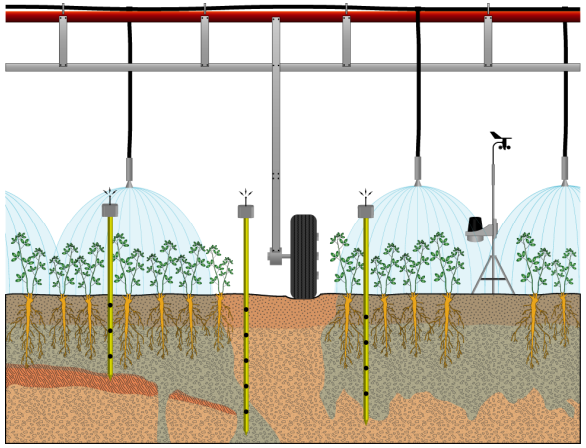
CENTER FOR EMBEDDED NETWORKED SENSING

UCLA USC UCR CALTECH UCM

Seismic



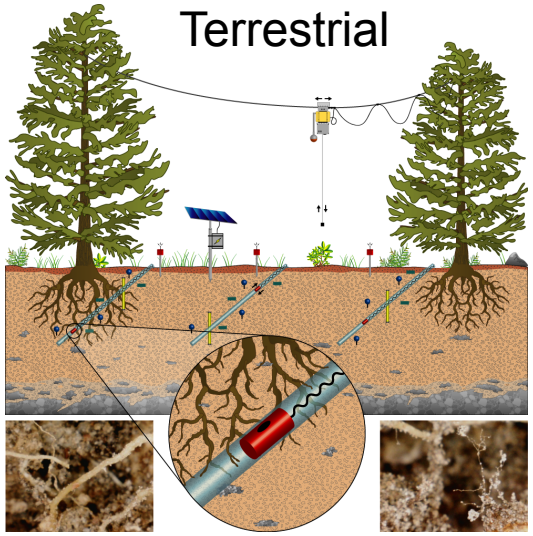
Contaminant transport



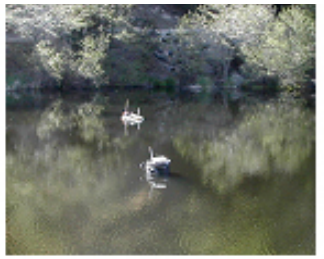
- Create programmable, distributed, multi-modal, multi-scale, multi-use **observatories** to address compelling science and engineering issues
- …and reveal the previously unobservable.
- From the natural to the built environment…
- From ecosystems to human systems…

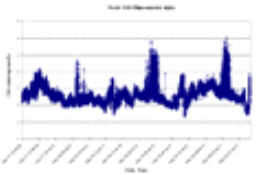
Urban

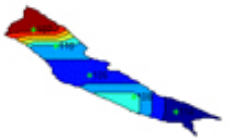
Terrestrial



Aquatic



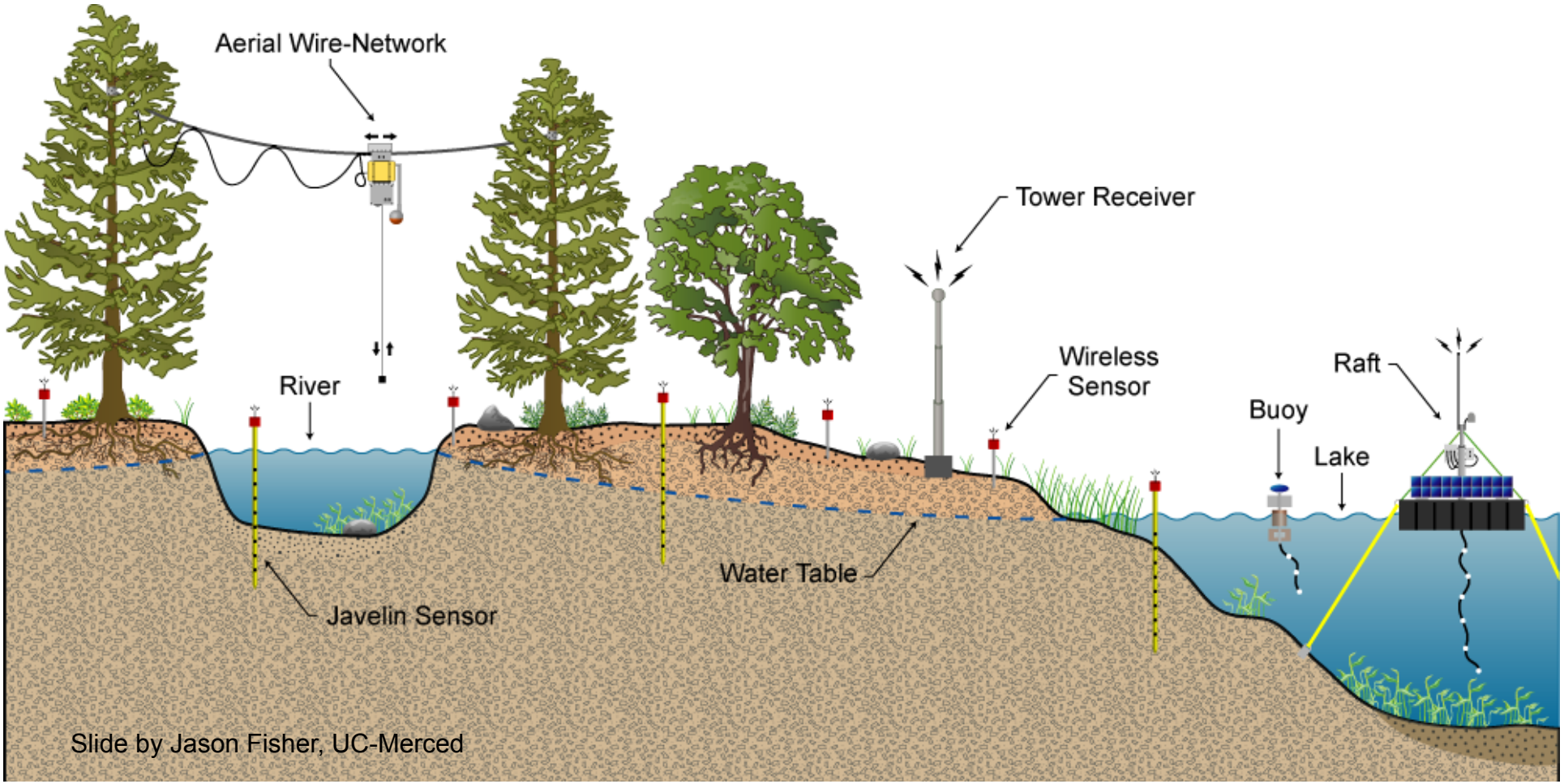




Field Deployment of Embedded Sensor Networks

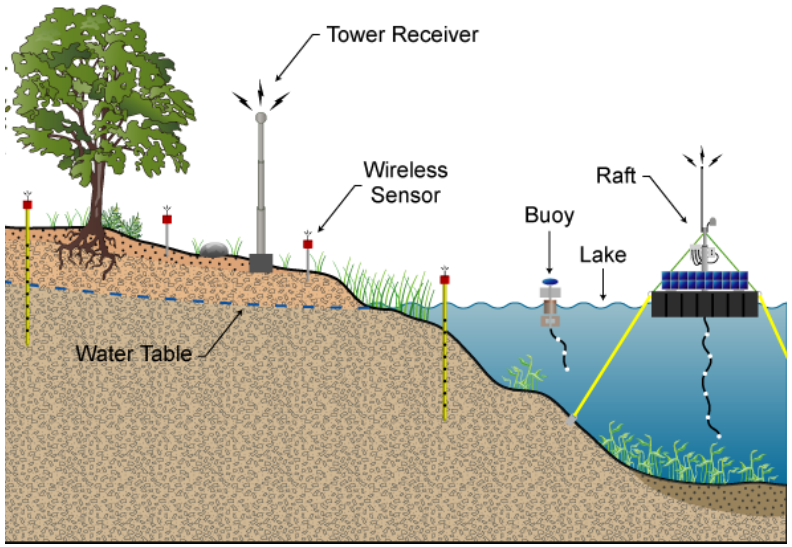
CENTER FOR EMBEDDED NETWORKED SENSING

UCLA USC UCR CALTECH UCM



Sensor networked science: An Information Perspective

- Size matters
  - Big science, little science
  - Big data, long tail
- Origins of data
  - Sources and resources
  - External factors
  - Purposes for collecting data
- Processing of data
  - Metadata
  - Provenance
  - Handling data



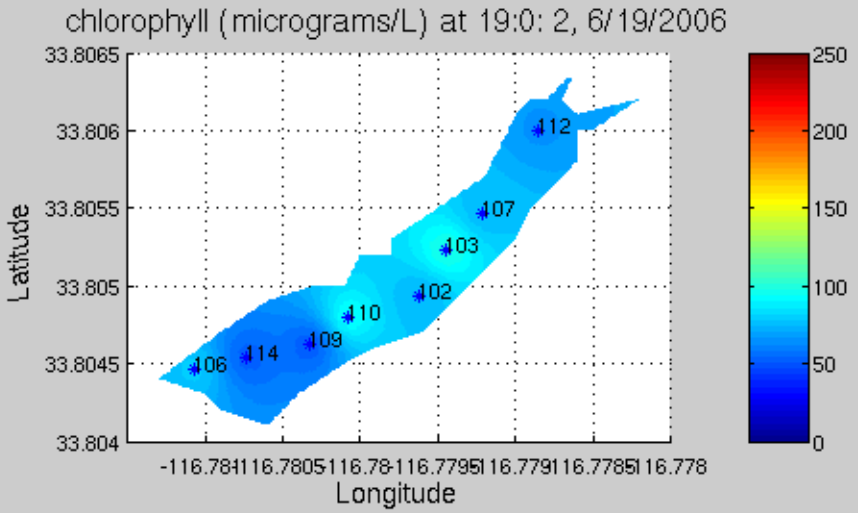
Coupled Human-Observational Systems

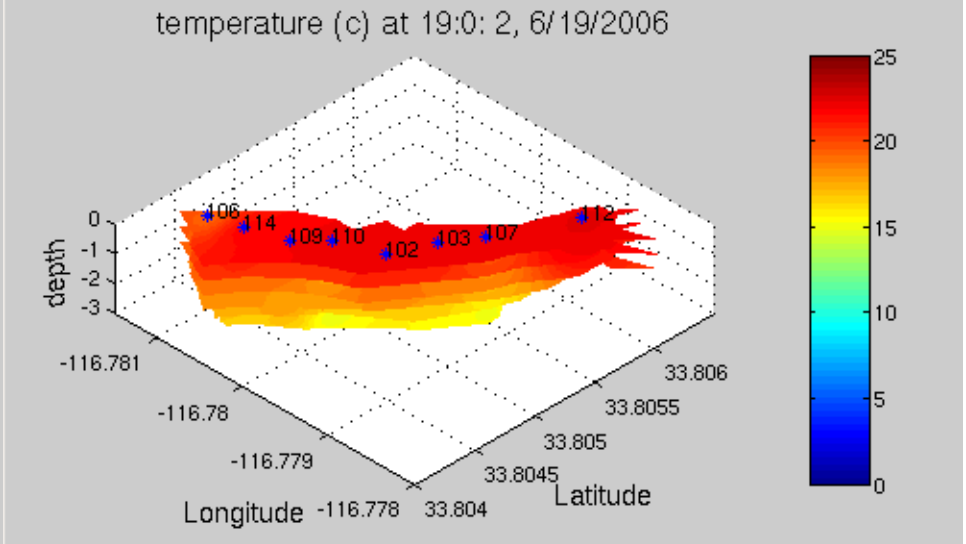
- Physical observations go from batch to interactive process
- Rapid deployments are high value
  - Exploratory research
  - α/β testing sensors
- Take advantage of human observer/ actuator
- Addresses critical issues in the field:
  - Adaptive sampling
  - Topology adjustment
  - Faulty sensor detection
- Require real time data access, model based analysis, and visualization in the field

Heterogeneous Sensing: Harmful Algal Blooms

CENTER FOR EMBEDDED NETWORKED SENSING

UCLA USC UCR CALTECH UCM



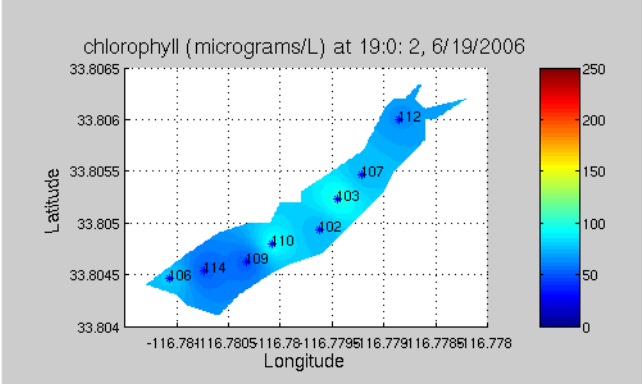


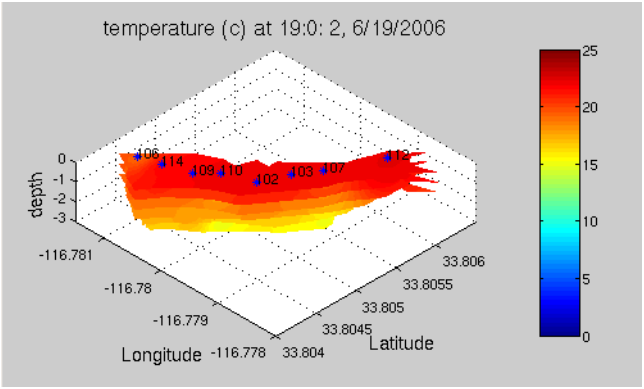
Biological sensor networks
Sources and Resources

- Phenomena of interest: Harmful algal blooms
- Organizing principles
  - Coordinates on land and in water
  - Meteorological conditions
  - Concentrations of algae, plankton, nutrients
- Data acquisition
  - Sensor observations of water: voltage
  - Sensor network data: patterns, flows
  - Samples of water
    - Wet lab, centrifuge, pH concentrations
    - Algae, plankton, nutrients
  - Input to biological models

Harmful algal blooms
Sources and Resources

- Data disposition:
  - Desktop / lab servers
    - Biology team data
    - Engineering team data
  - Refrigerators / freezers
    - Water samples
    - Biological samples





Purposes for collecting HAB data

- Specificity:
  - Biology: Triggers of harmful algal blooms
  - Computer science, engineering:
    - Robotic targeting
    - Network modeling
    - Technology design and testing
- Scope:
  - Biology: Specific lake as exemplar
  - CS&E: Generalizable algorithms and technology
- Goal: Co-innovation of technology for science

Limnol. Oceanogr., 53(5, part 2), 2008, 2333–2349

Macro- to fine-scale spatial and temporal distributions and dynamics of phytoplankton and their environmental driving forces in a small montane lake in southern California, USA

David A. Caron, Beth Stauffer, and Stefanie Moorthi[1]
Department of Biological Sciences, University of Southern California, Los Angeles, California 90089

Amarjeet Singh, Maxim Batalin, and Eric A. Graham
Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095

Mark Hansen
Department of Statistics, University of California Los Angeles, Los Angeles, California 90095

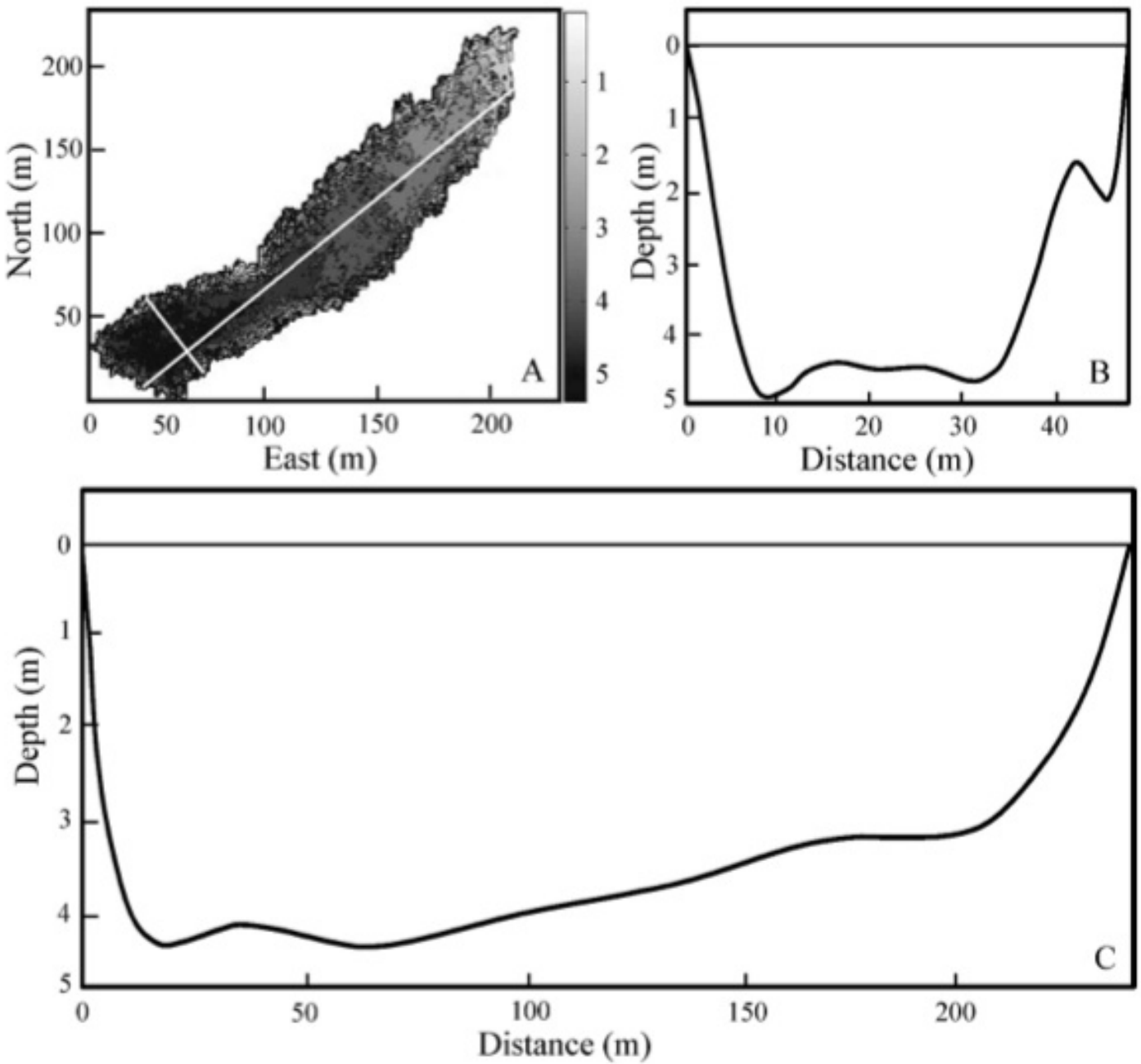
William J. Kaiser
Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095

Jnaneshwar Das, Arvind Pereira, Amit Dhariwal, Bin Zhang, Carl Oberg, and Gaurav S. Sukhatme
Department of Computer Science, University of Southern California, Los Angeles, California 90089

Abstract

A wireless network of buoys, two autonomous robotic boats, and an autonomous tethered vertical profiling system were used to characterize phytoplankton dynamics and spatiotemporal changes in chemical and physical forcing factors in a small montane lake (Lake Fulmor, Idyllwild, California). Three deployments each year were conducted in 2005 and 2006 to examine seasonal changes in the structure of the lake and phytoplankton assemblage, as well as fine-scale temporal and spatial variations. The buoys yielded fine-scale temporal patterns of in situ fluorescence and temperature, while the vertical profiling system yielded two-dimensional, cross-sectional profiles of several parameters. The autonomous vehicles provided information on fluorescence and corresponding temperature patterns across the surface of the lake. Average, lake-wide chlorophyll concentrations increased 10-fold seasonally, and strong anoxia developed in the hypolimnion during the summer. The latter process dramatically affected vertical chemical gradients in the 5 m water column of the lake. Small-scale spatial (<1 m) and temporal (minutes) heterogeneity in fluorescence were surprisingly large. These variations were due predominantly to vertical mixing of the phytoplankton assemblage and to phytoplankton vertical migratory behavior. Large peaks in fluorescence at 0.5-m occurred at very short time intervals (minutes) during all deployments, and appeared to be due to upward mixing of deeper dwelling eukaryotic phytoplankton during early–mid-summer, or downward mixing of surface-associated cyanobacteria during late summer.

Phytoplankton in a montane lake



ornia, and (B) cross-sections of
e solid white lines in panel A
e long axis of the lake), and the
ıe lake). (D) Picture of the lake
e NIMS RD mid-lake, (E) the
a NAMOS buoy.

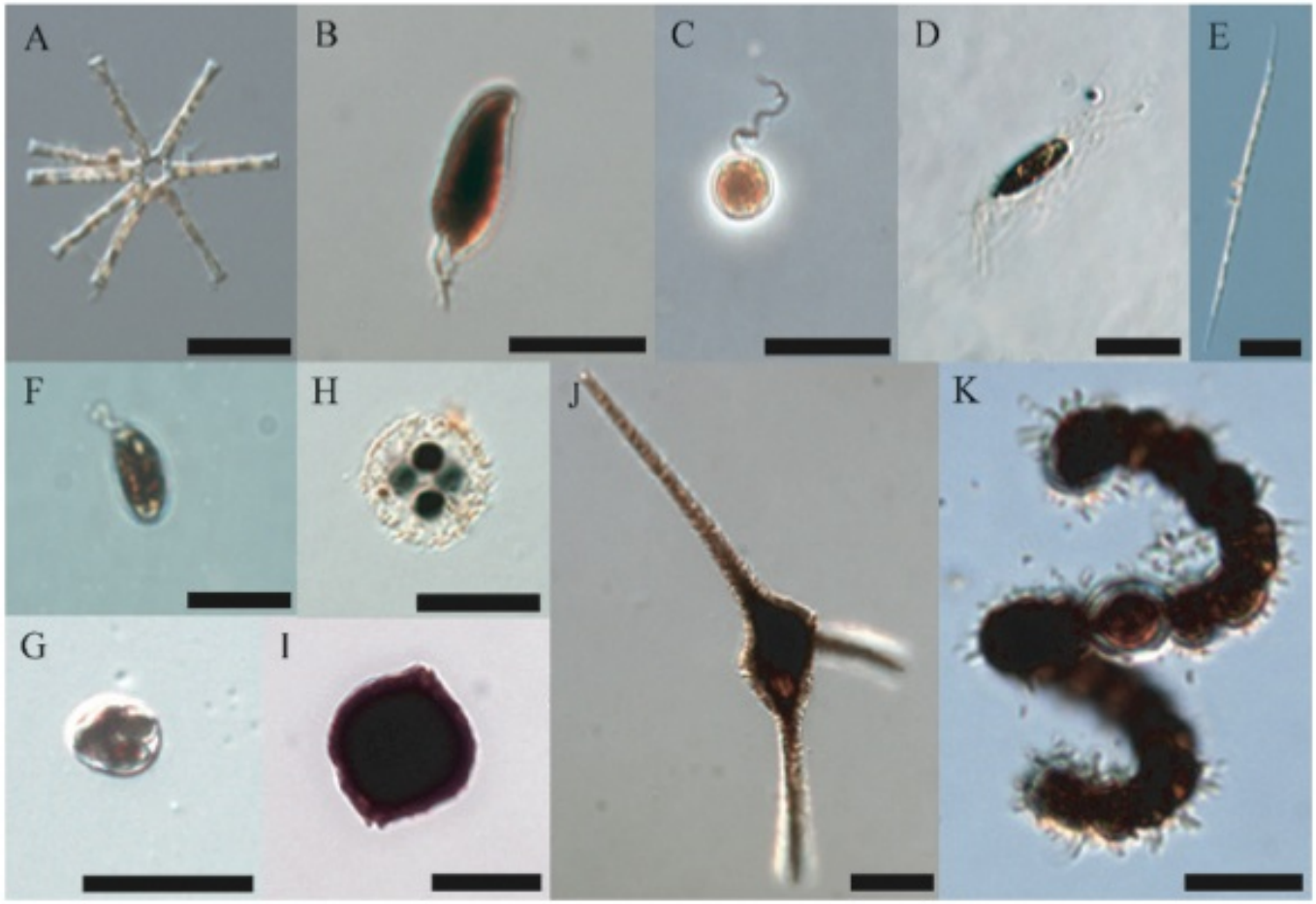
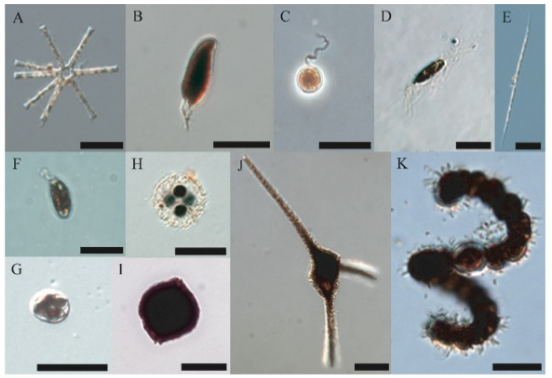
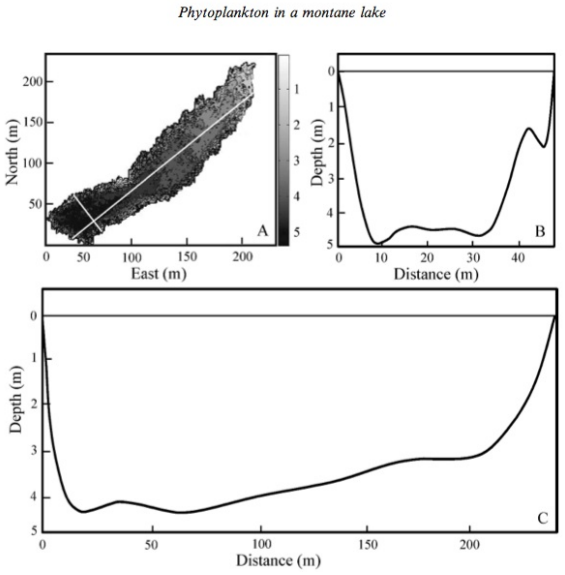


Fig. 2. Phytoplankton communities present in surface waters of Lake Fulmor on 09–10 May, 20–22 June and 29 August–1 September 2006. (A) The assemblage in May was highly dominated by the diatom *Asterionella formosa*. (B) Cryptophytes, (C) a large unidentified flagellate, and (D) small euglenid flagellates were also abundant at the surface and at depth. The plankton community in late June showed strong dominance by (E) diatoms, dinoflagellates, and (G, H) colonial *Gloeocapsa*-like cyanobacteria. (K) A persistent surface scum in late August was composed of the cyanobacterium *Anabaena spiroides* that formed in the late mornings. Dominant phytoplankton in the water column at that time included (I, J) large *Peridinium* and *Ceratium* dinoflagellates, diatoms, and a (F) large aggregation of small flagellates at the 3-m depth. Markers bars in panels D, F, K are 15 μm. Markers bars in all other panels are 35 μm.

Processing data for harmful algal blooms

- Survey of extant data
  - Meteorological data on lake
  - Prior data collected by these teams
  - Available code, algorithms
- Metadata
  - Spreadsheets, Matlab, R scripts
  - File naming conventions
- Provenance
  - Biology team documents biological data
  - Comp sci & eng team documents sensor network data





Handling data for Harmful Algal Blooms

- People involved
  - Biology team
  - Computer science/eng team
- Collecting the data
  - Investigators
  - Graduate students
- Processing the data
  - Biology team
  - Computer science/eng teams
  - Statistics partners
- Releasing the data
  - Genome data deposited
  - Some software code deposited
  - Other data shared on request, after publication

Limnol. Oceanogr., 53(5, part 2), 2008, 2333–2349

Macro- to fine-scale spatial and temporal distributions and dynamics of phytoplankton and their environmental driving forces in a small montane lake in southern California, USA

David A. Caron, Beth Stauffer, and Stefanie Moorthi[1]
Department of Biological Sciences, University of Southern California, Los Angeles, California 90089

Amarjeet Singh, Maxim Batalin, and Eric A. Graham
Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095

Mark Hansen
Department of Statistics, University of California Los Angeles, Los Angeles, California 90095

William J. Kaiser
Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095

Jnaneshwar Das, Arvind Pereira, Amit Dhariwal, Bin Zhang, Carl Oberg, and Gaurav S. Sukhatme
Department of Computer Science, University of Southern California, Los Angeles, California 90089

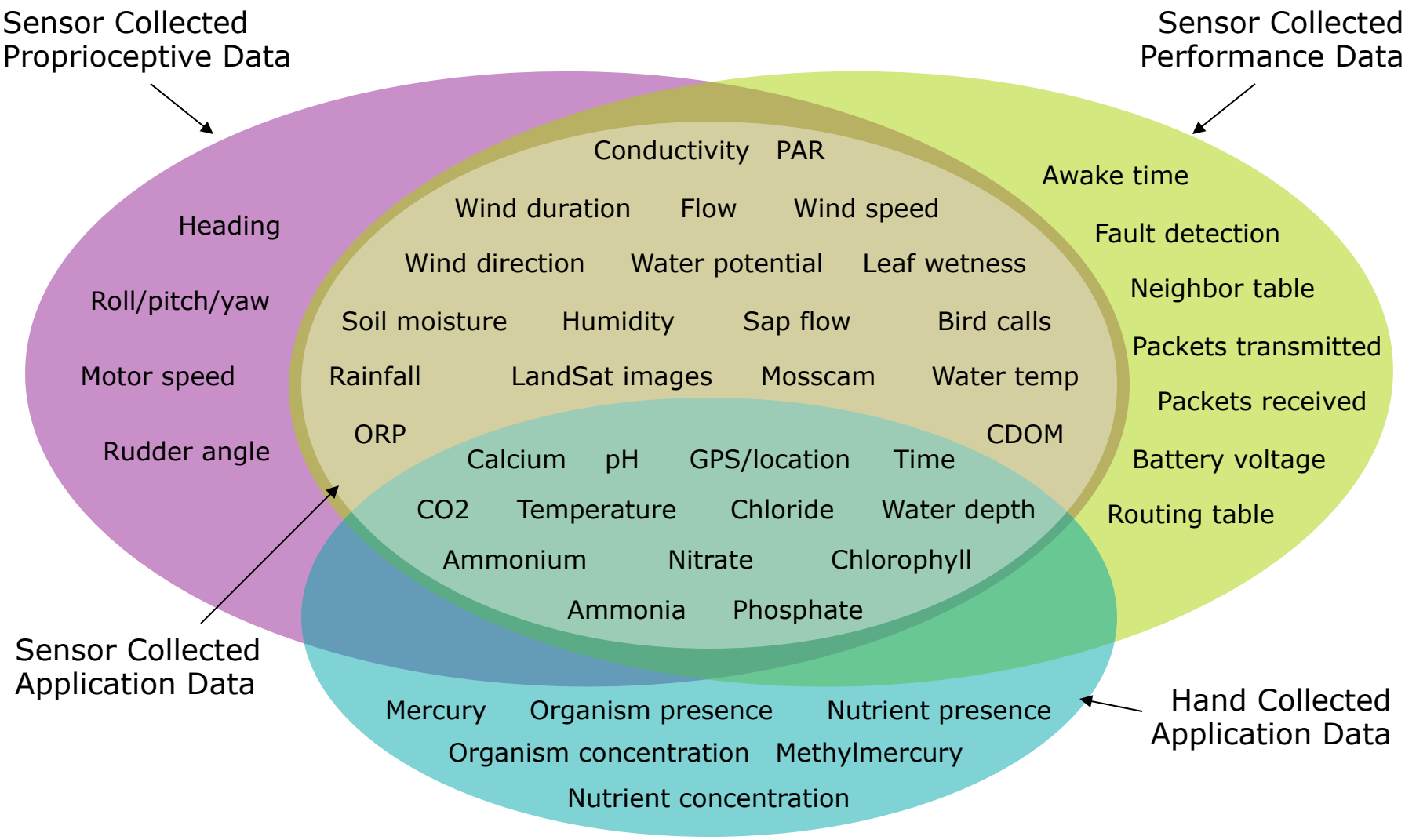
Abstract

A wireless network of buoys, two autonomous robotic boats, and an autonomous tethered vertical profiling system were used to characterize phytoplankton dynamics and spatiotemporal changes in chemical and physical forcing factors in a small montane lake (Lake Fulmor, Idyllwild, California). Three deployments each year were conducted in 2005 and 2006 to examine seasonal changes in the structure of the lake and phytoplankton assemblage, as well as fine-scale temporal and spatial variations. The buoys yielded fine-scale temporal patterns of in situ fluorescence and temperature, while the vertical profiling system yielded two-dimensional, cross-sectional profiles of several parameters. The autonomous vehicles provided information on fluorescence and corresponding temperature patterns across the surface of the lake. Average, lake-wide chlorophyll concentrations increased 10-fold seasonally, and strong anoxia developed in the hypolimnion during the summer. The latter process dramatically affected vertical chemical gradients in the 5 m water column of the lake. Small-scale spatial (<1 m) and temporal (minutes) heterogeneity in fluorescence were surprisingly large. These variations were due predominantly to vertical mixing of the phytoplankton assemblage and to phytoplankton vertical migratory behavior. Large peaks in fluorescence at 0.5-m occurred at very short time intervals (minutes) during all deployments, and appeared to be due to upward mixing of deeper dwelling eukaryotic phytoplankton during early–mid-summer, or downward mixing of surface-associated cyanobacteria during late summer.

What are CENS Data?

CENTER FOR EMBEDDED NETWORKED SENSING

UCLA USC UCR CALTECH UCM

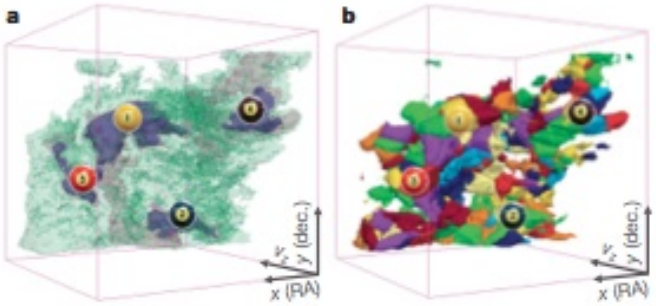


Graphic by Jillian Wallis

Big data and the long tail

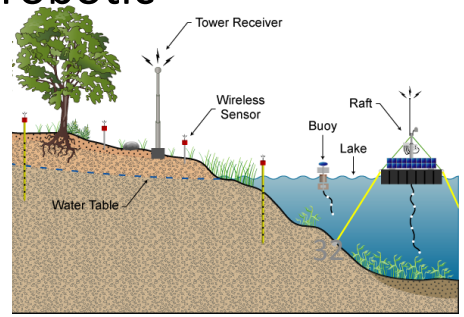
Astronomy: COMPLETE

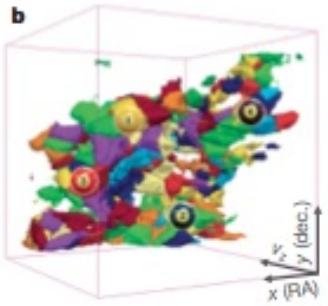
- Big science, big data
- Sky coordinate system
- Models to interpret observations
- Disposition: public
- Purpose: How do stars form?



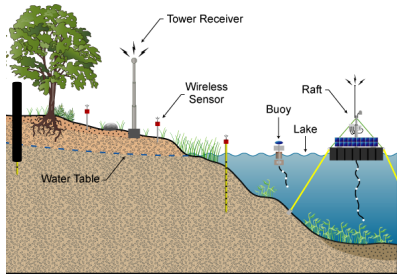
Sensor networks: HAB

- Little science, long tail
- Earth coordinate system
- Models to interpret observations
- Disposition: Private
- Purpose:
  - What triggers HAB?
  - How to target robotic sensors?





Big data and the long tail



Astronomy: COMPLETE

- Resource: extant data
- Source: new observations
- Metadata: community standards
- Provenance
  - Community standards
  - Local practice
- Handling
  - Team
  - Many people prior to project
- Public release of data

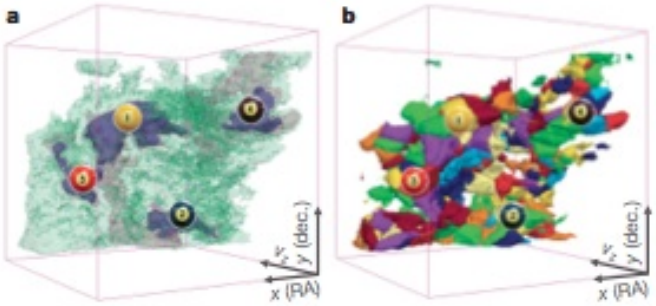
Sensor networks: HAB

- Source: new observations
- Metadata: local practice
- Provenance: local practice
- Handling
  - Science team
  - Biology team
- Data release on request

Big data and the long tail

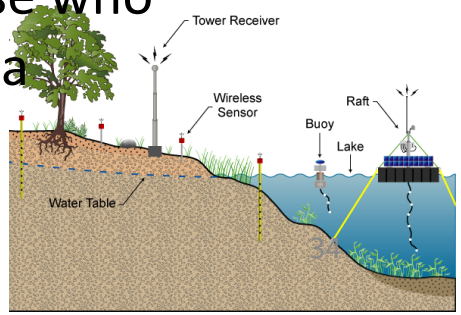
Astronomy: COMPLETE

- Use of observations
  - Foreground
  - Background
- Astronomy data:
  - Observations
  - Data products
- Reusable by domain experts



Sensor networks: HAB

- Use of observations
  - Foreground
  - Background
- Science data
  - Observations
  - Data products
- Computer science / eng data
  - Software code
- Reusable by those who collected the data



Acknowledgements

- National Science Foundation
  - *CENS*: Cooperative Agreement #CCR-0120778, D.L. Estrin, UCLA, PI.
  - *CENS Education Infrastructure*: #ESI- 0352572, W.A. Sandoval, PI; C.L. Borgman, co-PI.
  - *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; T. Finholt, University of Michigan, Co-PI.
  - *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures*: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU)
  - *Data Conservancy*: OCI0830976, Sayeed Choudhury, PI, Johns Hopkins University.
  - Knowledge and Data Transfer: the Formation of a New Workforce. # 1145888. C.L. Borgman, PI; S. Traweek, Co-PI.
- Microsoft External Research: Tony Hey, Lee Dirks, Catherine van Ingen, Catherine Marshall
- Sloan Foundation: The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI. Joshua Greenberg, program director
- Project website: http://knowledgeinfrastructures.gseis.ucla.edu/index.html

Microsoft

Recent papers related to this talk

- Borgman, C. L., Wallis, J. C., & Rolando, E. (2012). If We Share Data, Will Anyone Use Them? Data sharing and reuse in the long tail of science and technology. *Submitted to PLoS One*.
- Borgman, Christine L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634
- Borgman, Christine L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data?: Interdependencies in Science and Technology Collaborations. *Journal of Computer Supported Cooperative Work*. doi:DOI: 10.1007/s10606-012-9169-z
- Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*, 667–690. doi:10.1177/0306312711413314
- Mayernik, M. S. (2011, June). *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators* (PhD Dissertation). UCLA, Los Angeles. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042653
- Mayernik, M. S., Wallis, J. C., & Borgman, C. L. (2012). Unearthing the infrastructure: Humans and sensors in field-based research. *Computer Supported Cooperative Work*. doi:10.1007/s10606-012-9178-y