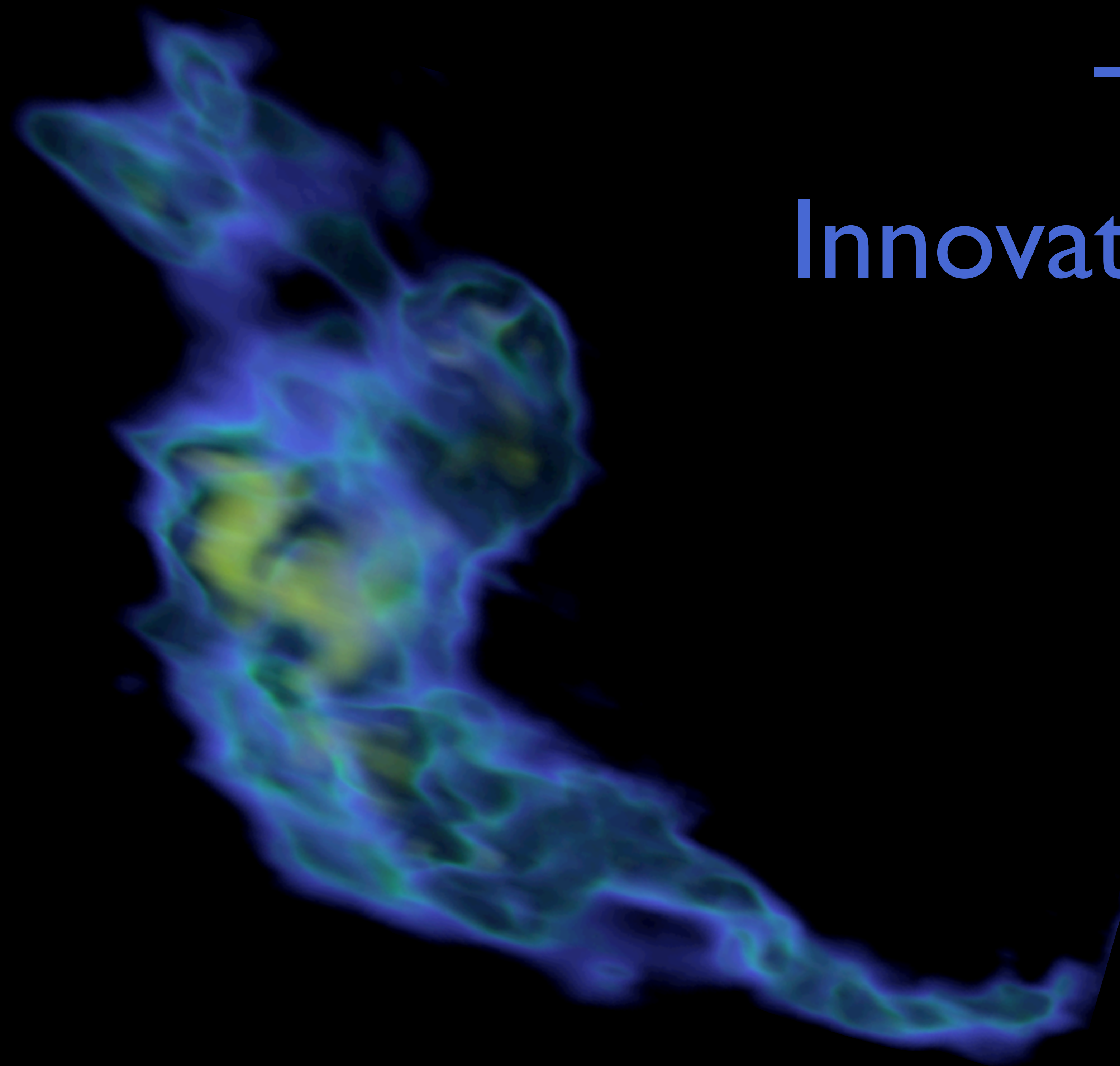


The Initiative in Innovative Computing at Harvard

Alyssa A. Goodman

Professor of Astronomy & Founding Director of the Initiative in Innovative Computing, Harvard University

Scholar-in-Residence, WGBH Boston

TED Fellows The TED Fellows Directory > Michelle Borkin

from the 2009 IIC Report...

TEDGlobal 2009

Bio

Michelle Borkin works on creating new appro
interdisciplinary scientific imaging, data explo
image analysis with a focus on 3D visualizatio
wrote her undergraduate junior and senior the
the application of medical imaging programs t
astronomical data and has continued this rese
part of the "Astronomical Medicine" project at I
Initiative in Innovative Computing. She works
developers of medical visualization tools to im

A Student's Story

My IIC journey began in a Cambridge Starbucks on Broadway on a cold, cloudy day in 2005. I was a Harvard undergraduate concentrating in astrophysics and looking for a junior thesis topic. I had been working on star formation research with Alyssa Goodman and expected to have a conversation about some small project within that research area—perhaps studying the mass distribution in a cloud, or how fast young stars move. But when Prof. Goodman arrived, the conversation—and my career—took a sharp turn. She had just returned from a joint NSF/NIH workshop on scientific visualization, where she had presented the leading challenges in astronomy visualization, including the unsolved problem of visualizing 3D radio data cubes. Another Harvard researcher, Michael Halle, had approached her with a possible solution. They bounced around the crazy idea that maybe 3D medical visualization software could be used with astronomy data. I didn't think twice before saying "I'll do it!"

Next thing I knew, I was riding the M2 shuttle bus from Cambridge to Longwood Medical District with pre-med friends, heading over to Brigham and Women's Hospital. Working with Dr. Halle and the Surgical Planning Lab, I learned about the difference between a CT and MRI image, what software is used, and how the surgeons use it in practice in the operating room. By the end of my junior year, I had managed to demonstrate that astronomy data could be read into medical imaging software, and that viewing these data cubes in 3D (which no astronomer had done previously) reveals features that are invisible or obscured in traditional 2D projections. Thus the IIC's Astronomical Medicine (AM) pilot project was born. I went on in my senior thesis to get important astronomical results using the 3D visualization techniques.

Instead off dashing of to graduate school in astronomy, I was intrigued by this new world of interdisciplinary research and this field called "visualization." So I chose to work at the newly formed IIC to continue my senior thesis research with the AM project. The AM team grew to include a dedicated software engineer, Douglas Alan, and a full-time astronomy postdoc, Jens Kauffmann.

In the two years I worked full-time at the IIC, the AM team accomplished much in terms of creating software tools for scientific visualization and in applying these tools in both astronomy and medicine. I gained the real-life research experience one cannot obtain in a classroom. I learned how to work as part of a team and to write grant proposals and journal articles. I traveled to astronomy, computer science and visualization conferences to present papers, give talks, and at the same time run exhibit booths. I learned how to conduct myself in a radio or magazine interview, make videos for outreach purposes, and make videos for sharing my visualizations with other scientists. I've had the joy of seeing how the physics and math I studied as an undergraduate is applied to real research: One has to know statistical mechanics to understand thermal noise in MRI machines or radio telescopes, so the algorithms to reduce the images can be coded identically, and one needs to understand differential equations and line integrals to create surfaces and render a 3D model. I've learned that a scientist cannot just be an "experimentalist" or a "theorist." To truly understand data, a scientist must compare observations and simulations using the same analysis and visualization techniques. The IIC has been invaluable not only in helping me acquire skills and knowledge, but also in developing my perspective on how to do research and advance the state of the art.

Where did IIC come from, and go?

Short Version

Response to Harvard's "expansion" in Science, and into Allston.

See IIC Whitepaper (2004) & Task Force on Science & Technology report (2005) for more.

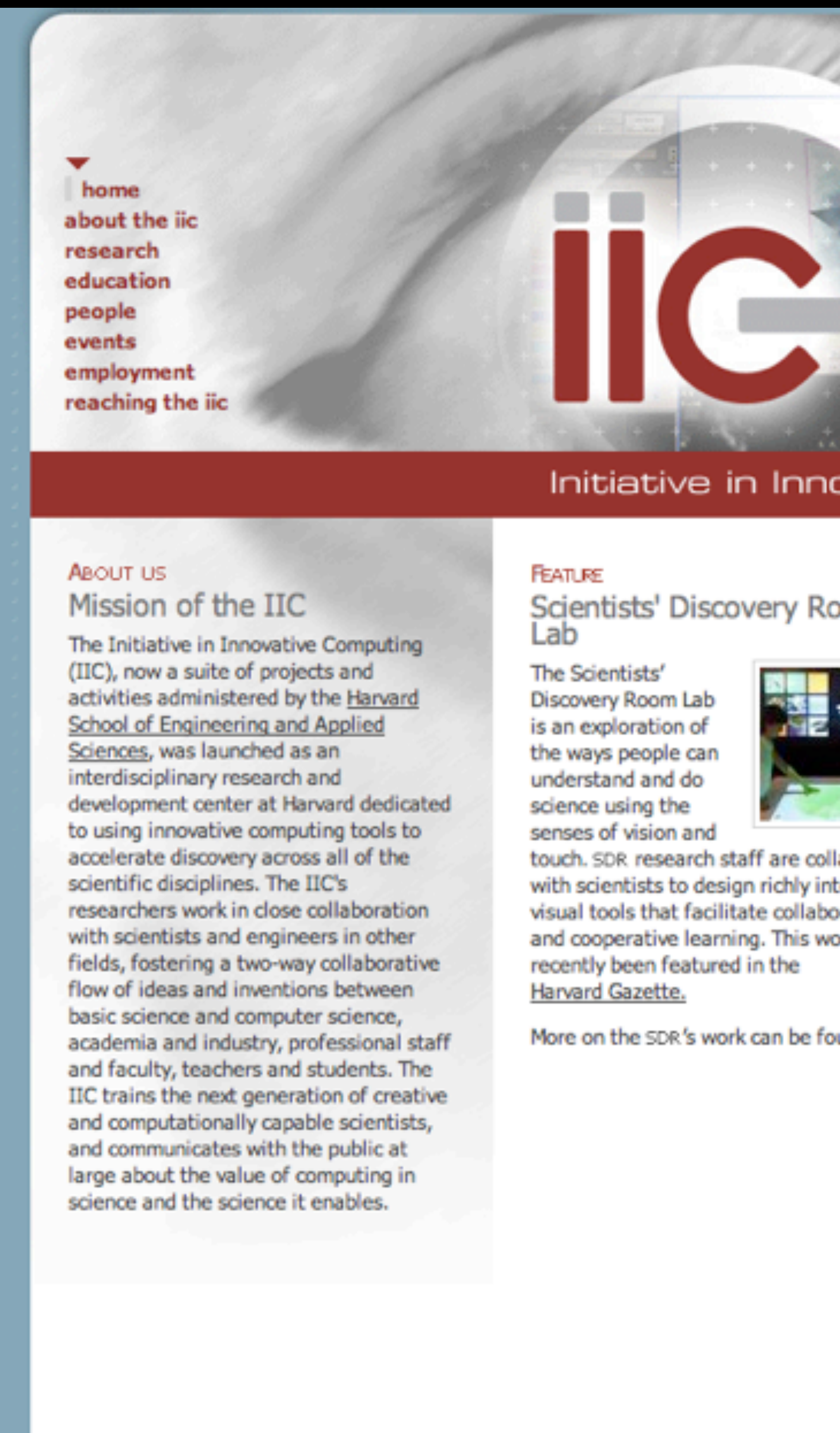
Long Version

(see the News on Allston & Harvard)

Current Status

Part of Harvard's New School of Engineering and Applied Science

IIC 2010: Part of Harvard's New Engineering School



Filling the "Gap" between Science and Computer Science

Increasingly, core problems in science require computational solution

Typically hire/"home grow" computationalists, but often lack the expertise or funding to go beyond the immediate pressing need

Focused on finding elegant solutions to basic computer science challenges

Often see specific, "applied" problems as outside their interests

Continuum

“Computational Science”
Missing at Most Universities

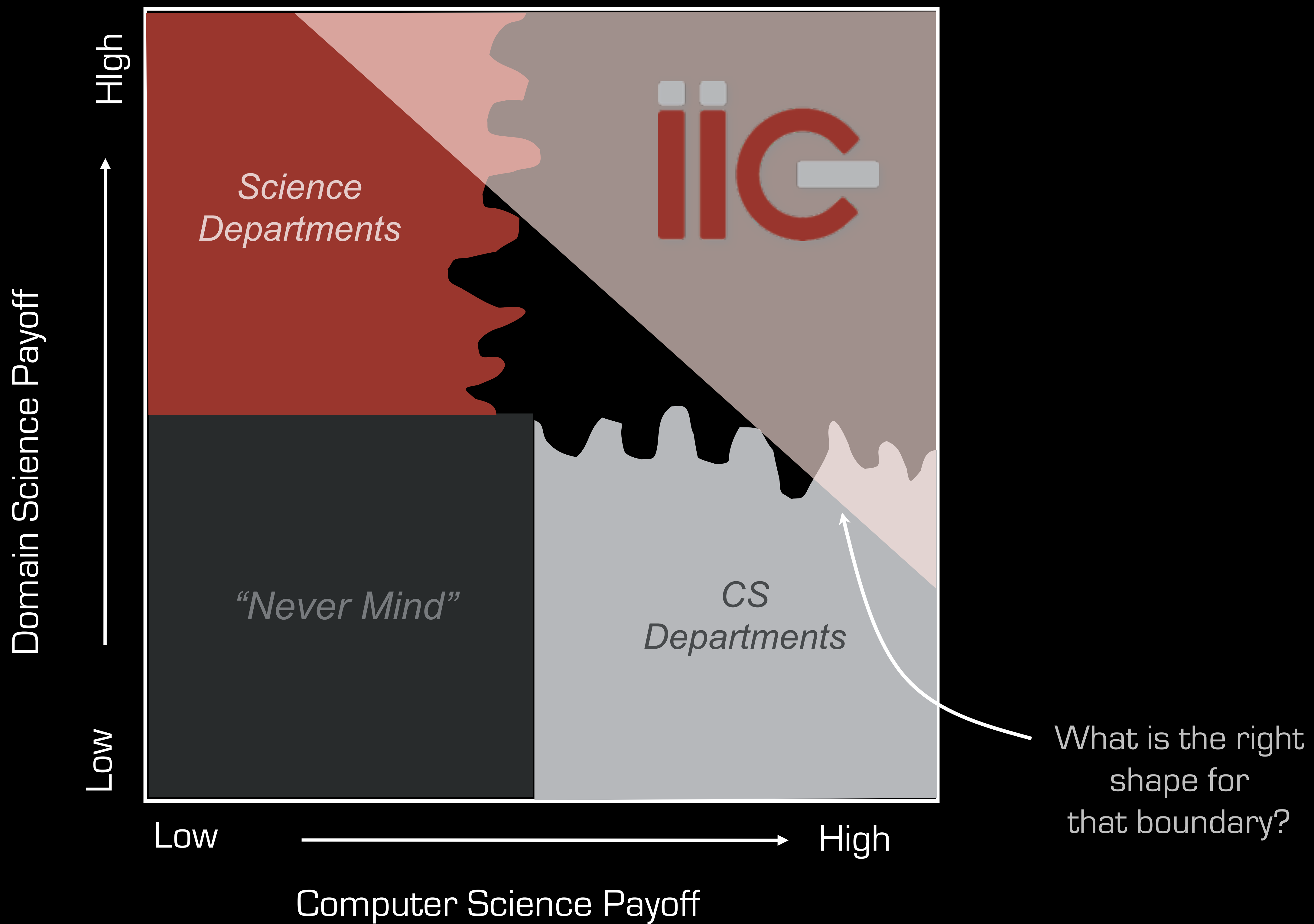
“Pure” Discipline Science

(e.g. Galileo)

“Pure” Computer Science

(e.g. Turing)

Where are the optimal "IIC" problems?



IIC Report
2009

THE IIC REPORT

INITIATIVE IN INNOVATIVE COMPUTING
AT HARVARD, 2006–09

lessons and results
from an interfaculty initiative

THE HARVARD INSTITUTE FOR INNOVATIVE COMPUTING
NOVEMBER 22, 2004

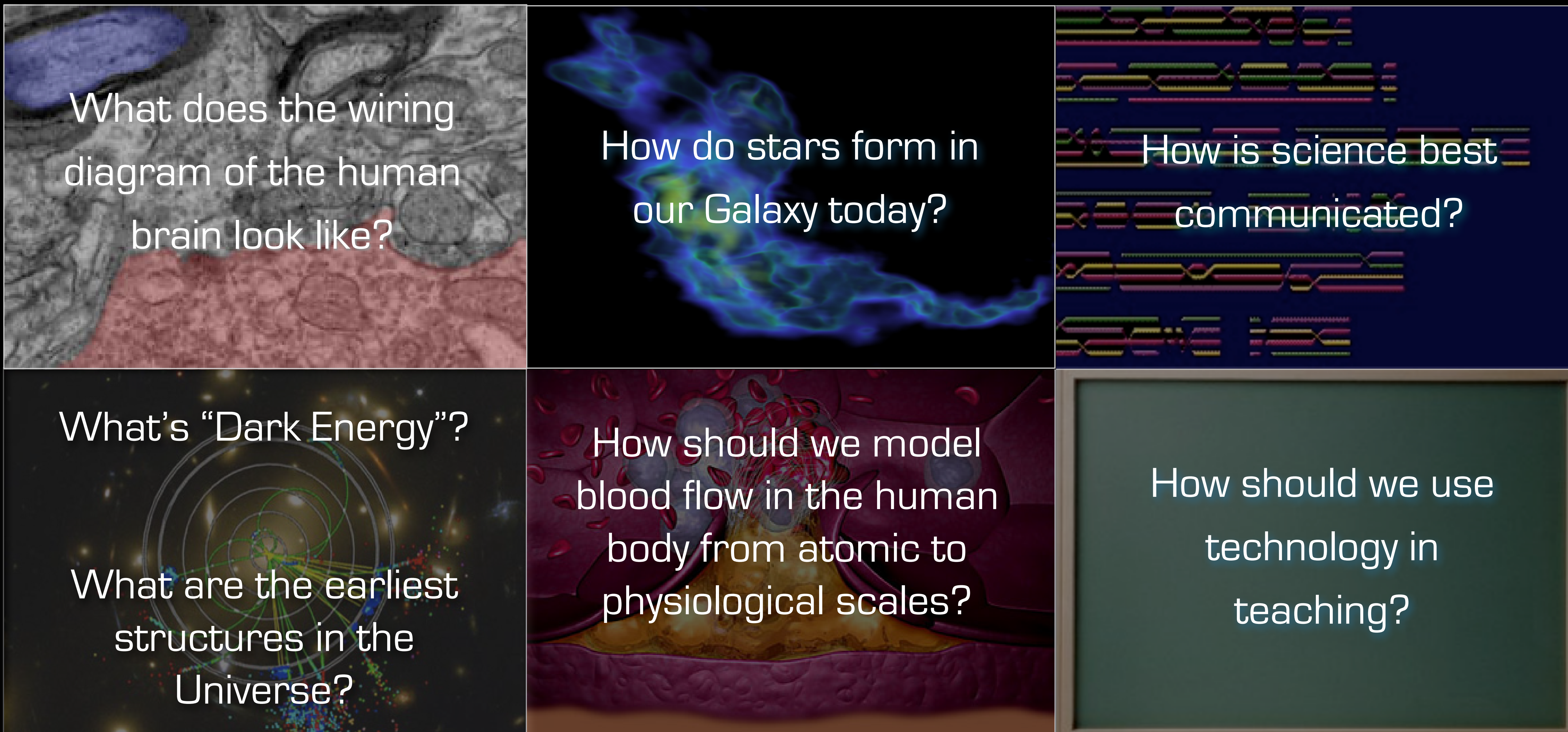
Principal Authors: Timothy Clark (HMS), Alyssa Goodman (FAS) & Christopher Stubbs (FAS)

Principal Editors: Felice Frankel (MIT), John Huth (FAS), Hanspeter Pfister (Mitsubishi Electric Research Laboratory), Joy Sircar (DEAS)[1]

IIC "Whitepaper"
2004

TABLE OF CONTENTS

IIC Symposium 2007


What does the wiring diagram of the human brain look like?
How do stars form in our Galaxy today?
How is science best communicated?
What's "Dark Energy"?
What are the earliest structures in the Universe?
How should we model blood flow in the human body from atomic to physiological scales?
How should we use technology in teaching?

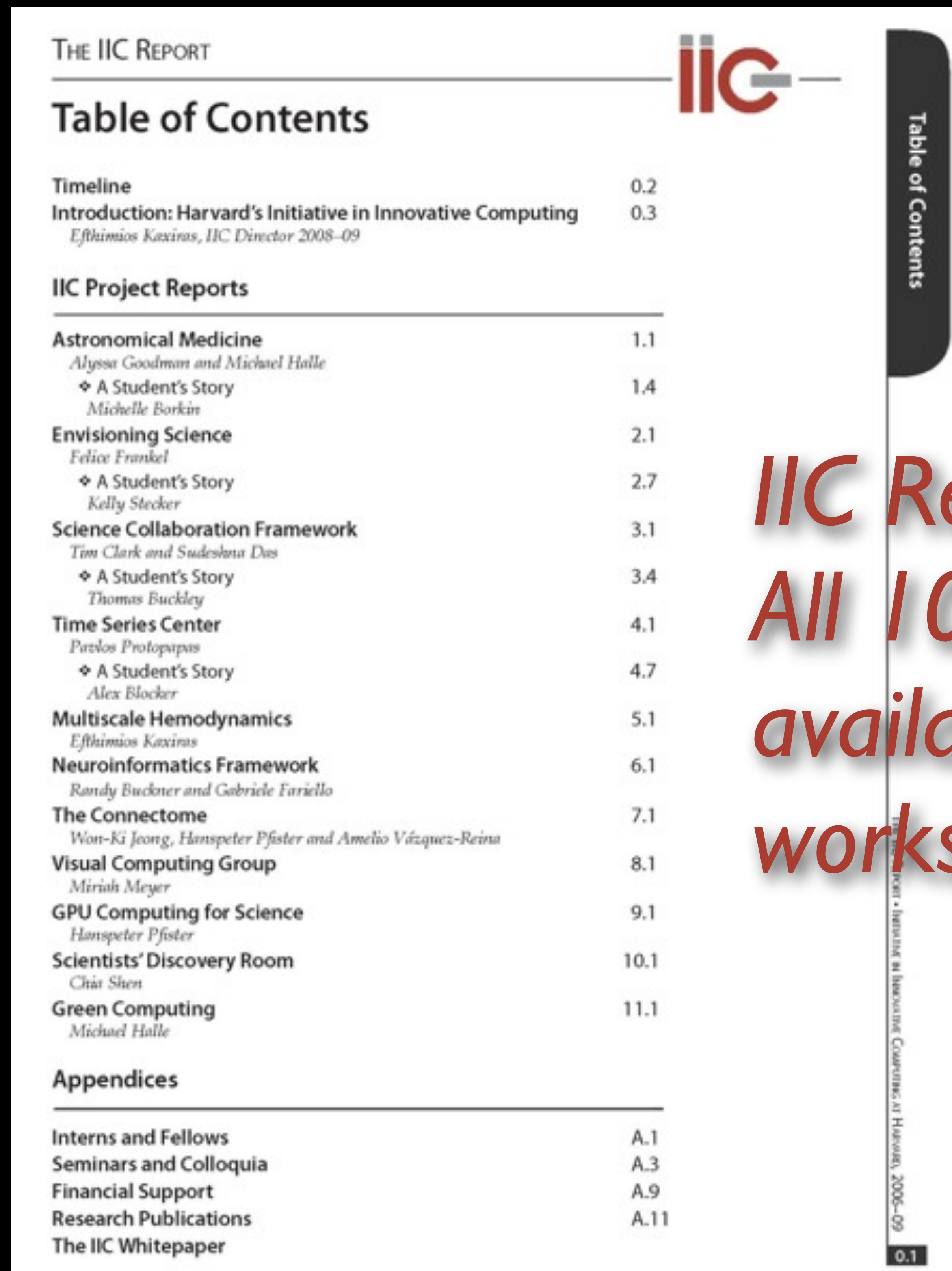


Table of Contents

IIC Report 2009: All 100+ pages available at this workshop's website

One example...

THE HARVARD INSTITUTE FOR INNOVATIVE COMPUTING
NOVEMBER 22, 2004

Principal Authors: Timothy Clark (HMS), Alyssa Goodman (FAS) & Christopher Stubbs (FAS)

Principal Editors: Felice Frankel (MIT), John Huth (FAS), Hanspeter Pfister (Mitsubishi Electric Research Laboratory), Joy Sircar (DEAS)[1]

TA

PARSING, DISPLAYING, AND SERVING THE COMPLETE DATABASE ON THE WEB: A STEPPING STONE TO THE VIRTUAL OBSERVATORY

Alyssa A. Goodman
Department of Astronomy, Faculty of Arts & Sciences, Harvard University

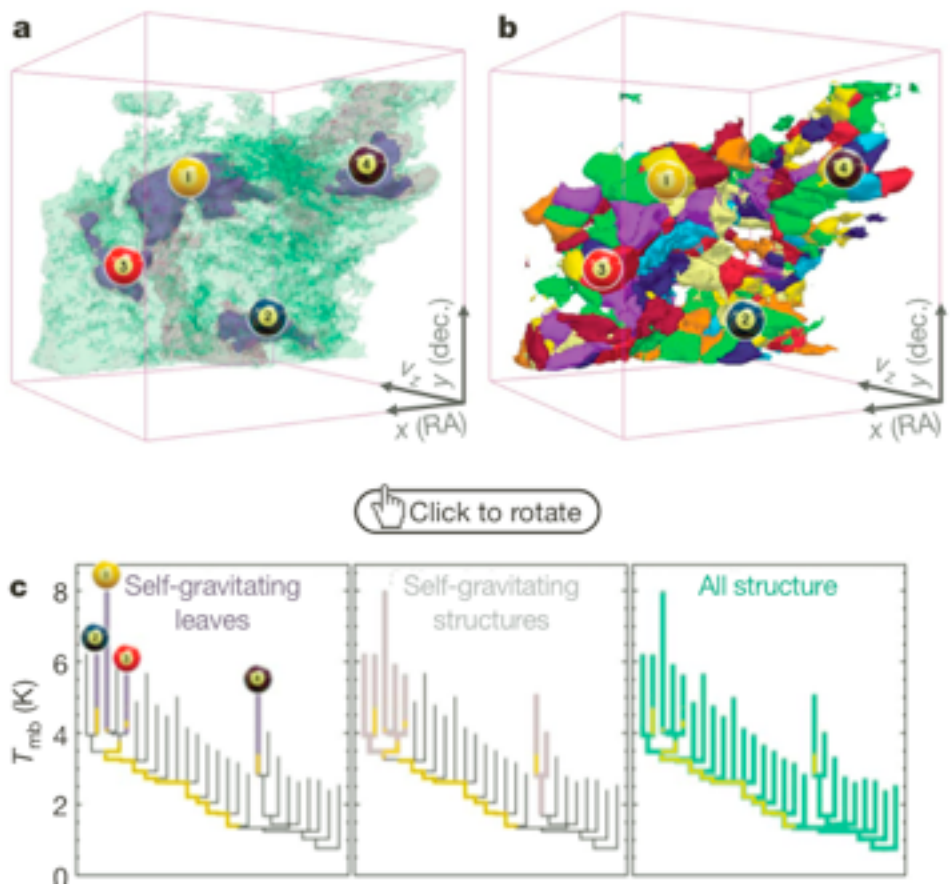
The COMPLETE Survey (cfa-www.harvard.edu/COMPLETE) currently underway is the largest, most diverse, systematic, multi-wavelength study of star-forming regions ever undertaken. Star formation, in spite of its central role in the evolution of the Universe, is currently very poorly understood, and the COMPLETE database will be used to answer a tremendous variety of physical questions, including ones like "How did the Sun form?". Without systematic surveys like COMPLETE, the data available are just too statistically sparse and poorly archived to make headway on these questions.

The COMPLETE data set includes maps made at wavelengths from optical through radio, some of which have an added "velocity" dimension beyond the three dimensions (position-position-intensity) astronomical images usually contain. Each of the mapping techniques employed produces a unique "kind" of data that is ordinarily dealt with almost exclusively by experts on that "type" of data, and very few researchers are expert in all the techniques. The tools for displaying and interpreting these kinds of data are, at this point in time, technique-specific. Trying to retrieve, use, and display COMPLETE and other similarly diverse astronomical data now is akin to trying to assemble a car from all of the one-thousand needed parts, but using directions partially in Chinese, partially in English, and partially in French, with some illustrations that can only be viewed on Windows PCs, others only on Macs, and some others only with the right brand of 3D glasses. Making all of COMPLETE's data readily searchable and accessible to astronomers worldwide, regardless of their expertise or computing platform, is an unprecedented—but definitely achievable—challenge in both database design and visualization.

We envision two potential members of the IIC who would work together in making the COMPLETE dataset accessible to all on the Web. One IIC collaborator would be expert in the design and searching of large multi-dimensional databases, and the other would be more focused on designing tools to use existing and new visual interfaces to access and analyze the database. We expect that a total of four person-years (two years per participant, in parallel) would be needed to create a streamlined working system.

The result of this IIC work would be a new data retrieval/display system that would be the first to handle what is known as "velocity resolved" or "spectral line data cube" data sets that intrinsically have four dimensions. This system would be incorporated into all of the Virtual Observatory interfaces now being developed (see http://www.us-vo.org/) to provide this kind of access to multi-wavelength, multi-dimensional data sets across astronomy.

Astronomical Medicine/3D
2005-2009



Parsing, Displaying, and Serving the COMPLETE Database on the Web: A Stepping Stone to the Virtual Observatory

Alyssa A. Goodman
Department of Astronomy, Faculty of Arts & Sciences, Harvard University

The COMPLETE Survey (cfa-www.harvard.edu/COMPLETE) currently underway is the largest, most diverse, systematic, multi-wavelength study of star-forming regions ever undertaken. Star formation, in spite of its central role in the evolution of the Universe, is currently very poorly understood, and the COMPLETE database will be used to answer a tremendous variety of physical questions, including ones like "How did the Sun form?". Without systematic surveys like COMPLETE, the data available are just too statistically sparse and poorly archived to make headway on these questions.

The COMPLETE data set includes maps made at wavelengths from optical through radio, some of which have an added "velocity" dimension beyond the three dimensions (position-position-intensity) astronomical images usually contain. Each of the mapping techniques employed produces a unique "kind" of data that is ordinarily dealt with almost exclusively by experts on that "type" of data, and very few researchers are expert in all the techniques. The tools for displaying and interpreting these kinds of data are, at this point in time, technique-specific. Trying to retrieve, use, and display COMPLETE and other similarly diverse astronomical data now is akin to trying to assemble a car from all of the one-thousand needed parts, but using directions partially in Chinese, partially in English, and partially in French, with some illustrations that can only be viewed on Windows PCs, others only on Macs, and some others only with the right brand of 3D glasses. Making all of COMPLETE's data readily searchable and accessible to astronomers worldwide, regardless of their expertise or computing platform, is an unprecedented—but definitely achievable—challenge in both database design and visualization.

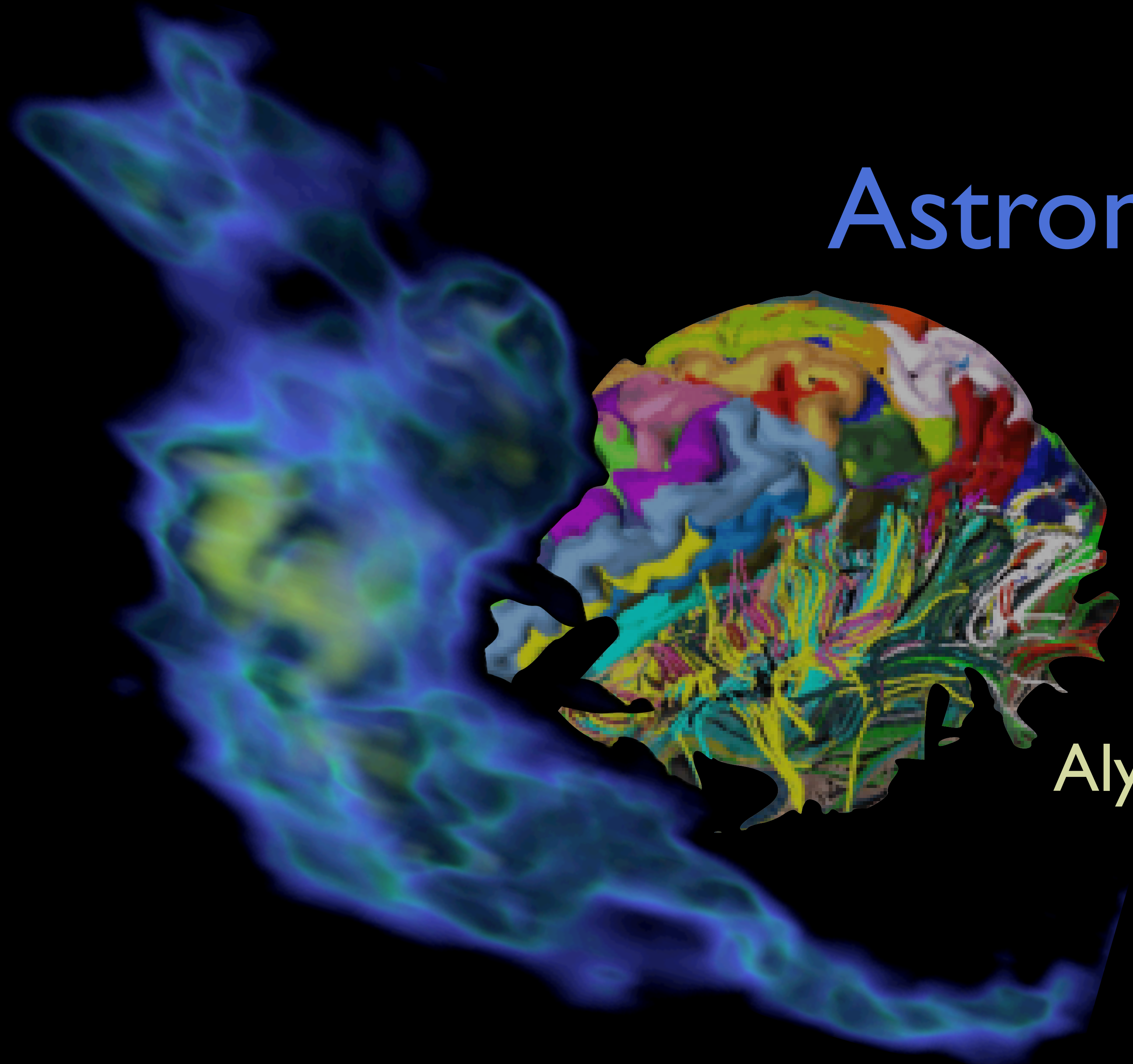
We envision two potential members of the IIC who would work together in making the COMPLETE dataset accessible to all on the Web. One IIC collaborator would be expert in the design and searching of large multi-dimensional databases, and the other would be more focused on designing tools to use existing and new visual interfaces to access and analyze the database. We expect that a total of four person-years (two years per participant, in parallel) would be needed to create a streamlined working system.

The result of this IIC work would be a new data retrieval/display system that would be the first to handle what is known as "velocity resolved" or "spectral line [illegible]" data sets that intrinsically have four dimensions. This system would be incorporated into a [illegible]irtual Observatory interfaces now being developed (see http://www.us-vo.org/) to provide thi[illegible]ccess to multi-wavelength, multi-dimensional data sets across astronomy.

Seamless Astronomy
(2008-...)

Results are fantastic- but not instant!

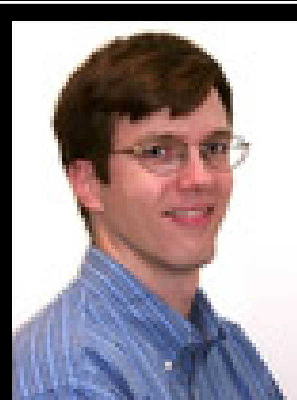
Astronomical Medicine



Alyssa Goodman (IIC/CfA/FAS)
Michael Halle (IIC/SPL/HMS)
Ron Kikinis (SPL/HMS)
Douglas Alan (IIC)
Michelle Borkin (FAS/IIC)
Jens Kauffmann (CfA/IIC)
Erik Rosolowsky (CfA/UBC Okanagan)
Nick Holliman (U. Durham)

The Astronomical Medicine Story

Computer Scientist

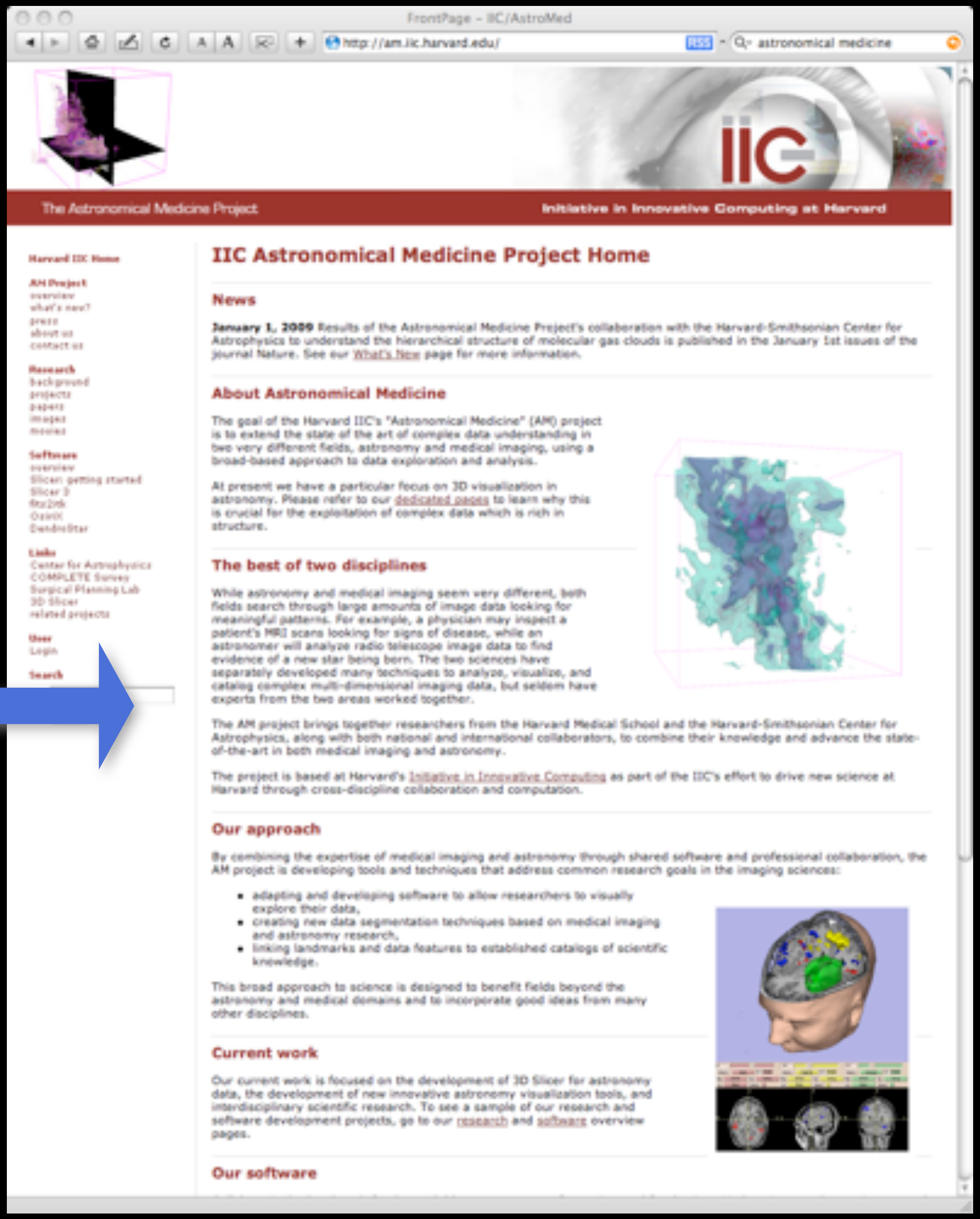


Computer Scientist

"Viz has failed the scientific community…"

Astronomer

Unsuspecting Undergrad



+Nick Holliman (CS, 3D expert)
+Doug Alan (S/W Engineer)
+Jens Kauffmann (postdoc)
+Erik Rosolowsky (postdoc) + …

COMPLETE =COordinated Molecular Probe Line Exinction Thermal Emission

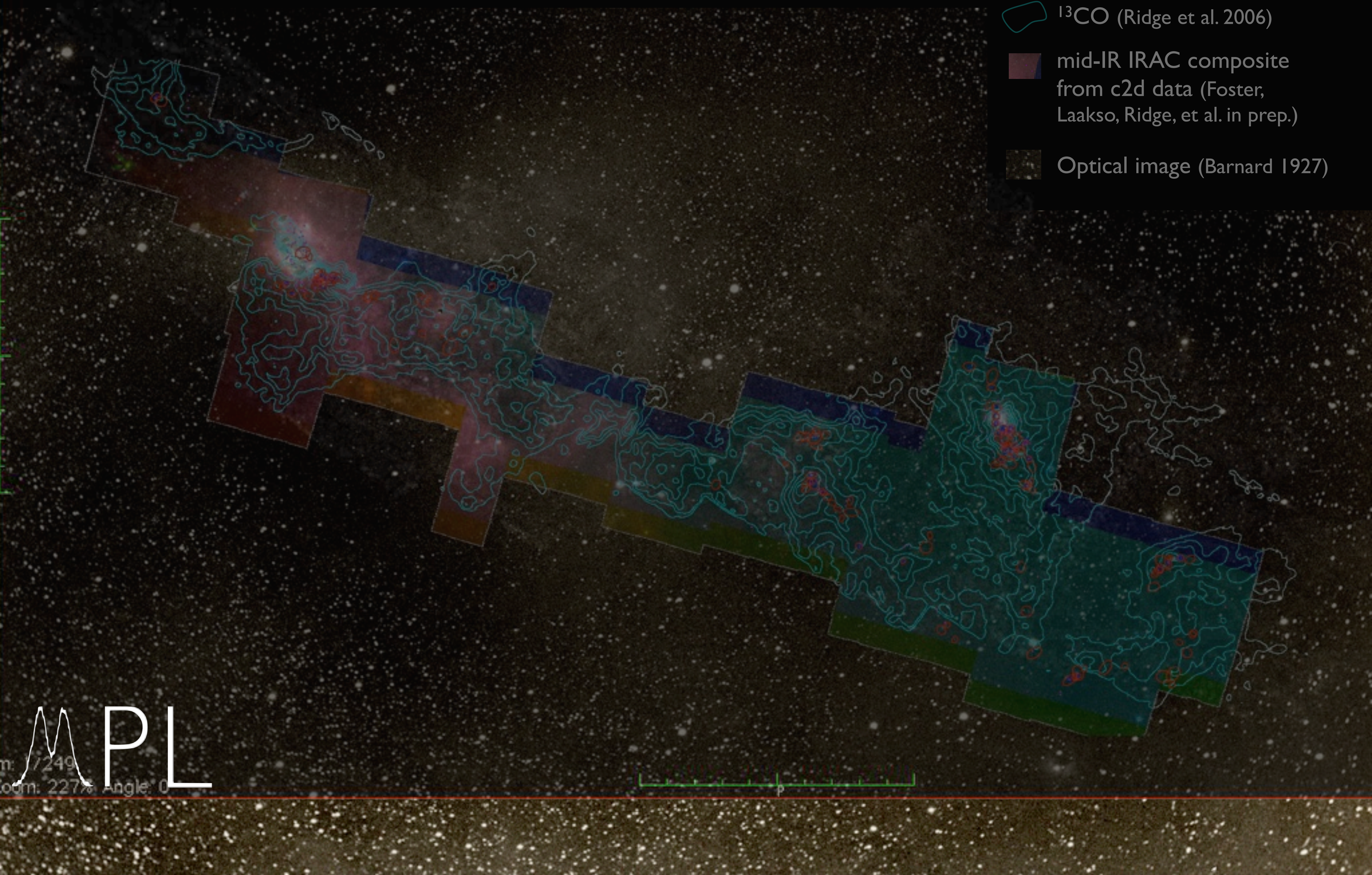
mm peak (Enoch et al. 2006)

sub-mm peak (Hatchell et al. 2005, Kirk et al. 2006)

^{13}CO (Ridge et al. 2006)

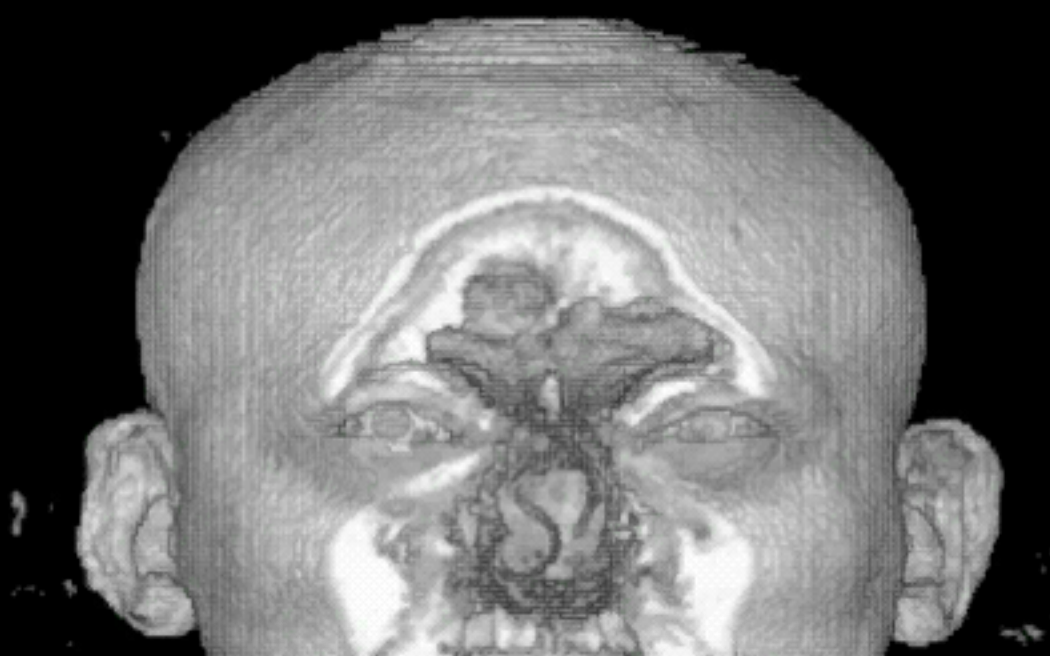
mid-IR IRAC composite from c2d data (Foster, Laakso, Ridge, et al. in prep.)

Optical image (Barnard 1927)



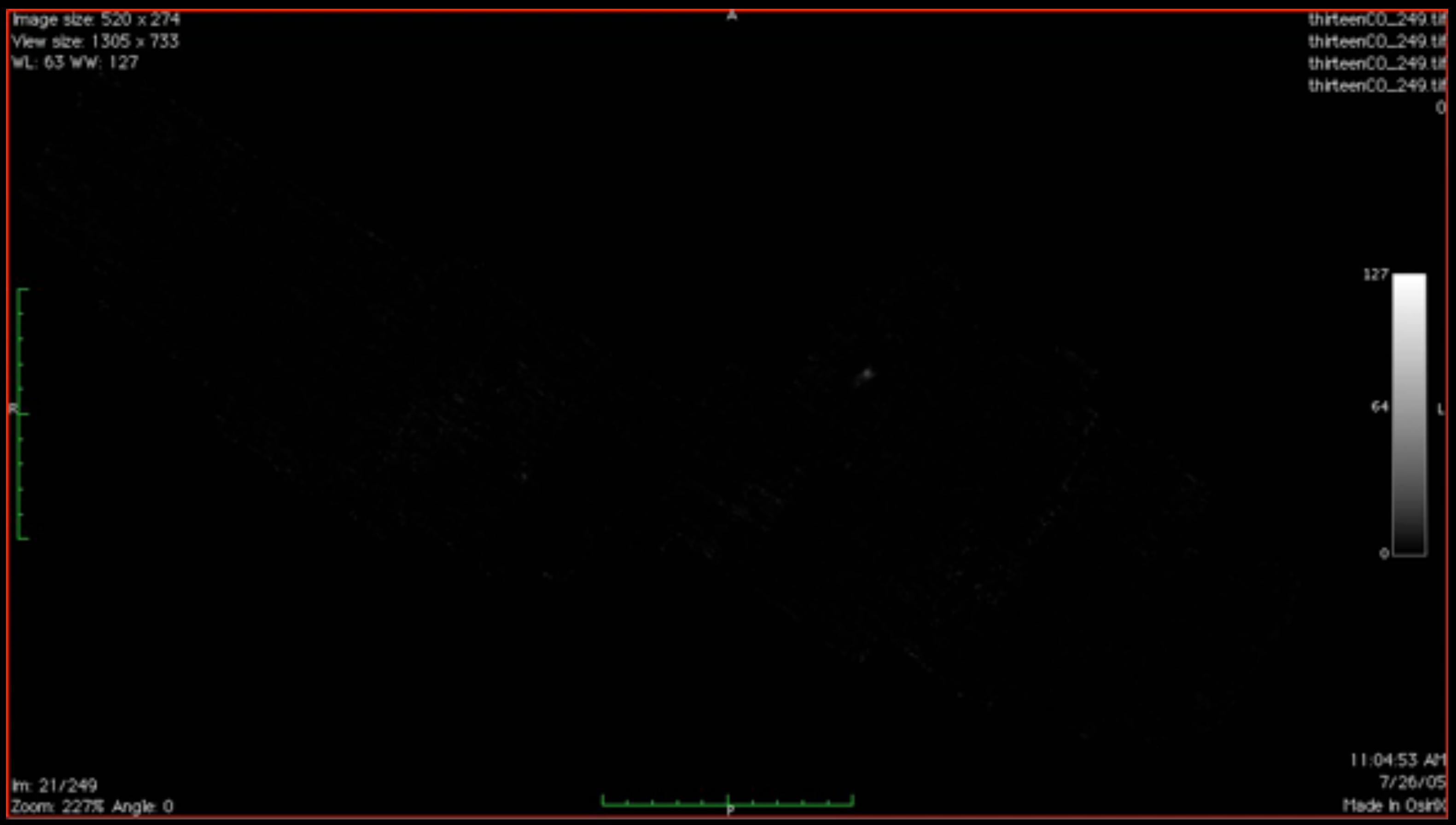
“Astronomical Medicine”

“KEITH”



“z” is depth into head

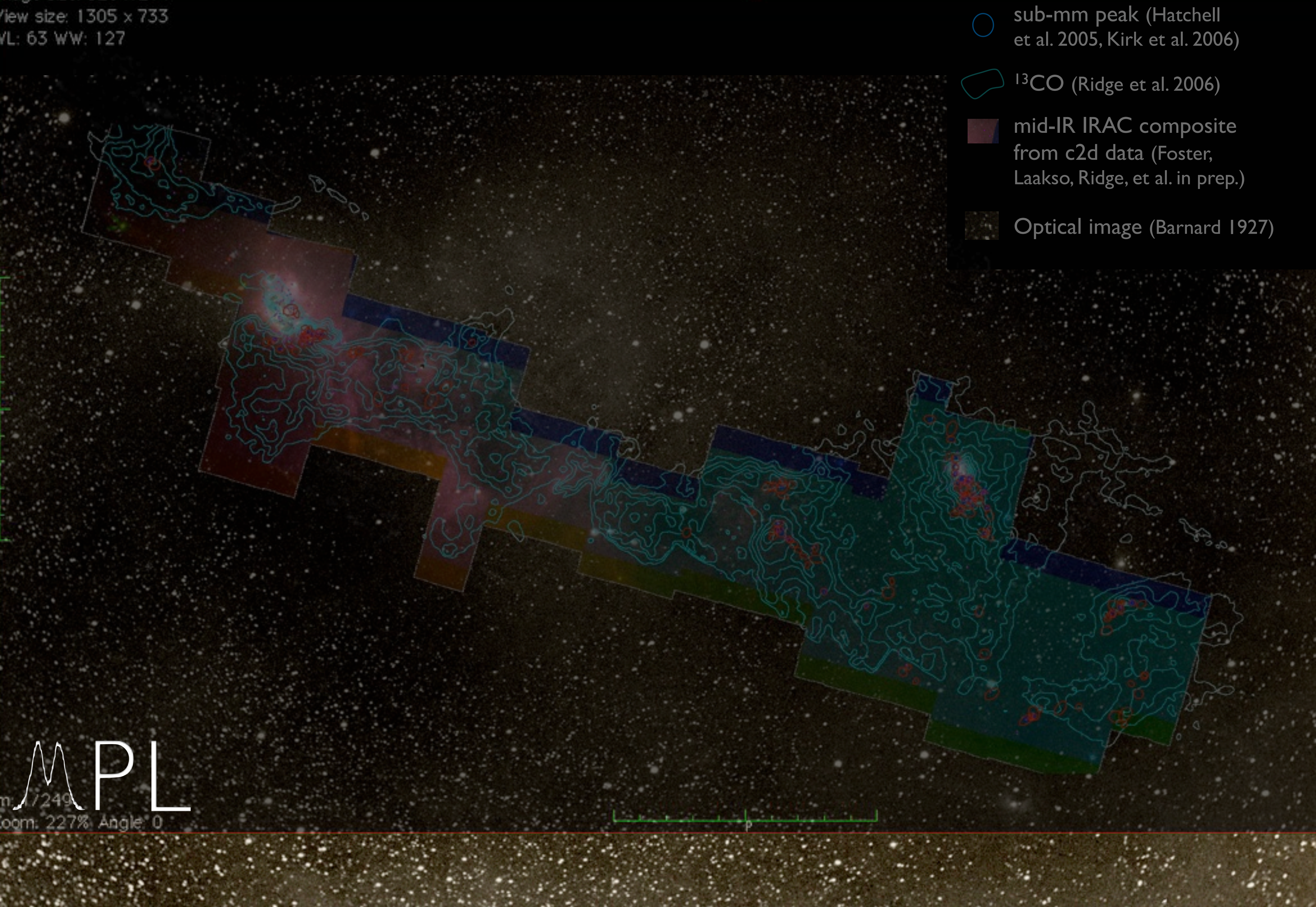
“PERSEUS”

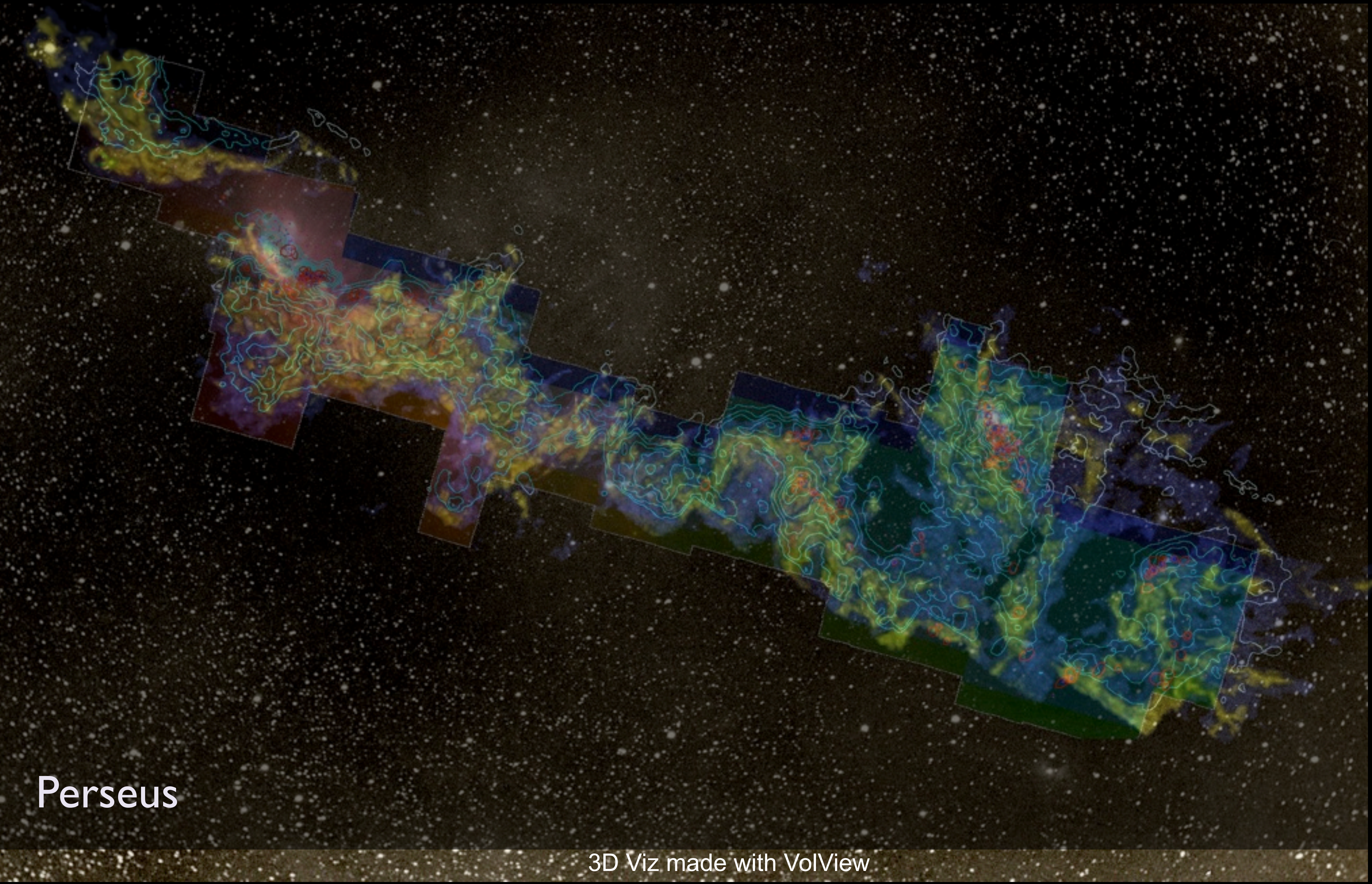


“z” is line-of-sight velocity

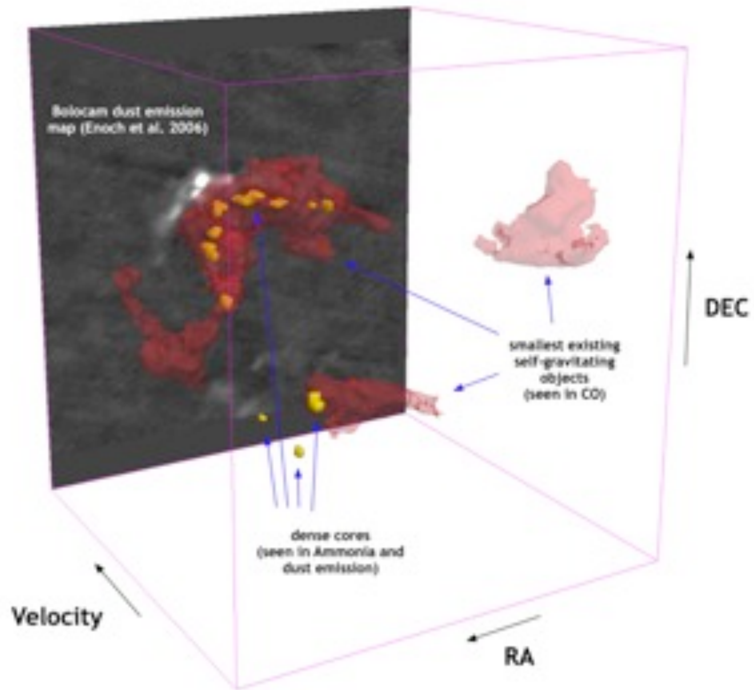
(This kind of “series of 2D slices view” is known in the Viz as “the grand tour”)

COMPLETE Perseus

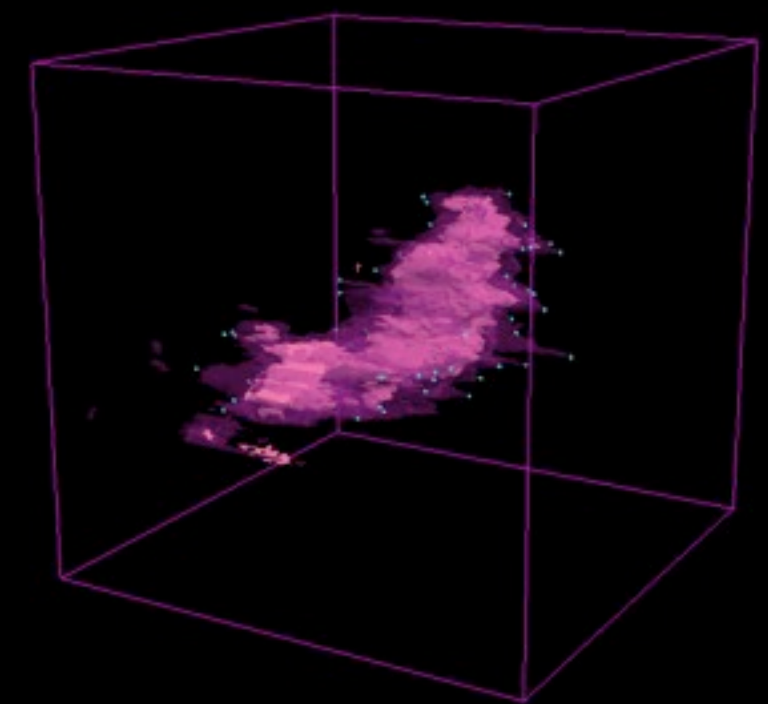



Perseus
3D Viz made with VolView

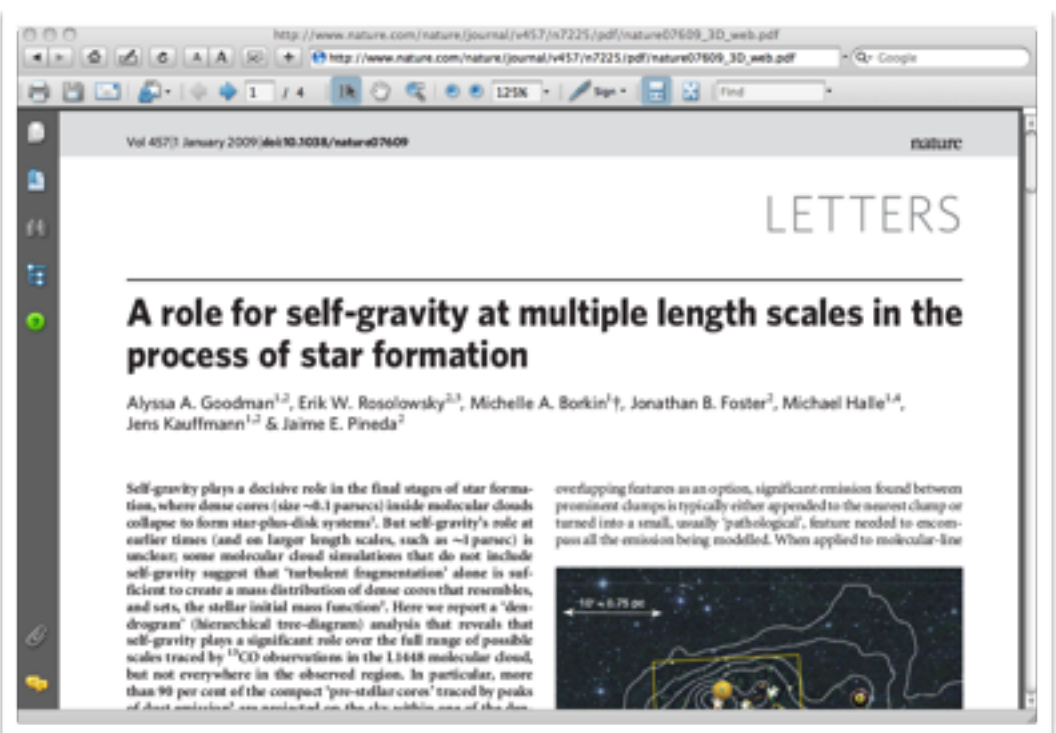
Some of What We've Learned...



Cores nest in coccoons
(Kauffmann et al. 2009)



Tripled Outflows
(Borkin et al. 2008,9)



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

LETTERS

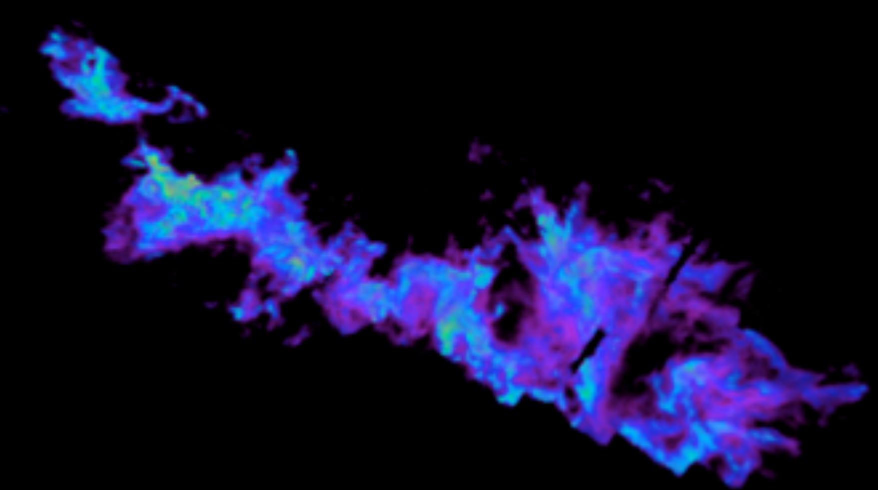
A role for self-gravity at multiple length scales in the process of star formation

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

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

Gravity Matters
(Goodman et al. 2009)



Shells Rule
(Arce et al. 2009)

New ways to share insight

“3D PDF”

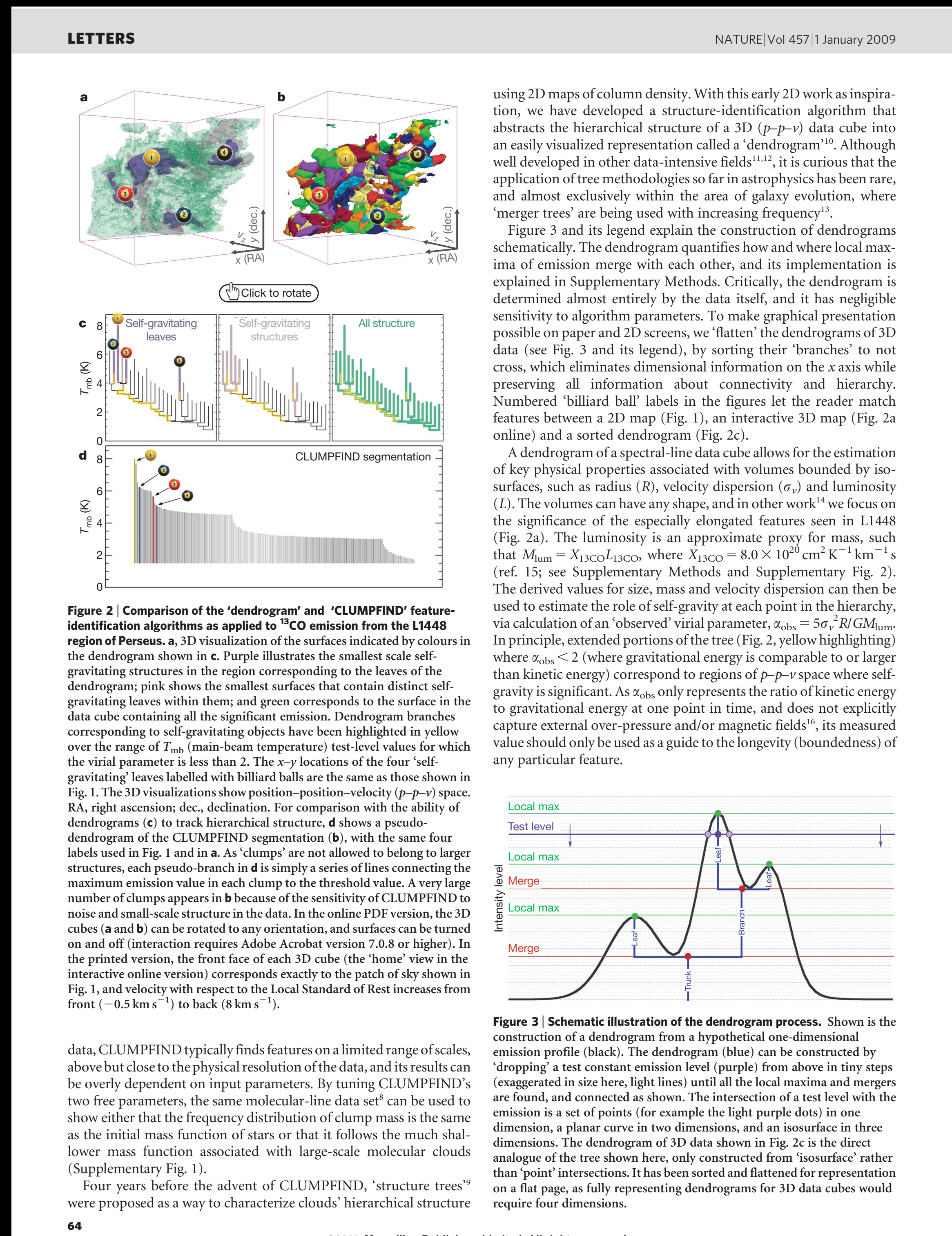


Figure 2 | Comparison of the ‘dendrogram’ and ‘CLUMPFIND’ feature-identification algorithms as applied to ^{13}CO emission from the L1448 region of Perseus. a, 3D visualization of the surfaces indicated by colours in the dendrogram shown in c. Purple illustrates the smallest scale self-gravitating structures in the region corresponding to the leaves of the dendrogram; pink shows the smallest surfaces that contain distinct self-gravitating leaves within them; and green corresponds to the surface in the data cube containing all the significant emission. Dendrogram branches corresponding to self-gravitating objects have been highlighted in yellow over the range of T_{mb} (main-beam temperature) test-level values for which the virial parameter is less than 2. The x–y locations of the four ‘self-gravitating’ leaves labelled with billiard balls are the same as those shown in Fig. 1. The 3D visualizations show position–position–velocity (p–p–v) space. RA, right ascension; dec., declination. For comparison with the ability of dendrograms (c) to track hierarchical structure, d shows a pseudo-dendrogram of the CLUMPFIND segmentation (b), with the same four labels used in Fig. 1 and in a. As ‘clumps’ are not allowed to belong to larger structures, each pseudo-branch in d is simply a series of lines connecting the maximum emission value in each clump to the threshold value. A very large number of clumps appears in b because of the sensitivity of CLUMPFIND to noise and small-scale structure in the data. In the online PDF version, the 3D cubes (a and b) can be rotated to any orientation, and surfaces can be turned on and off (interaction requires Adobe Acrobat version 7.0.8 or higher). In the printed version, the front face of each 3D cube (the ‘home’ view in the interactive online version) corresponds exactly to the patch of sky shown in Fig. 1, and velocity with respect to the Local Standard of Rest increases from front ($-0.5\ km\ s^{-1}$) to back ($8\ km\ s^{-1}$).

data, CLUMPFIND typically finds features on a limited range of scales, above but close to the physical resolution of the data, and its results can be overly dependent on input parameters. By tuning CLUMPFIND’s two free parameters, the same molecular-line data set[8] can be used to show either that the frequency distribution of clump mass is the same as the initial mass function of stars or that it follows the much shallower mass function associated with large-scale molecular clouds (Supplementary Fig. 1).

Four years before the advent of CLUMPFIND, ‘structure trees’[9] were proposed as a way to characterize clouds’ hierarchical structure using 2D maps of column density. With this early 2D work as inspiration, we have developed a structure-identification algorithm that abstracts the hierarchical structure of a 3D (p–p–v) data cube into an easily visualized representation called a ‘dendrogram’[10]. Although well developed in other data-intensive fields[11,12], it is curious that the application of tree methodologies so far in astrophysics has been rare, and almost exclusively within the area of galaxy evolution, where ‘merger trees’ are being used with increasing frequency[13].

Figure 3 and its legend explain the construction of dendrograms schematically. The dendrogram quantifies how and where local maxima of emission merge with each other, and its implementation is explained in Supplementary Methods. Critically, the dendrogram is determined almost entirely by the data itself, and it has negligible sensitivity to algorithm parameters. To make graphical presentation possible on paper and 2D screens, we ‘flatten’ the dendrograms of 3D data (see Fig. 3 and its legend), by sorting their ‘branches’ to not cross, which eliminates dimensional information on the x axis while preserving all information about connectivity and hierarchy. Numbered ‘billiard ball’ labels in the figures let the reader match features between a 2D map (Fig. 1), an interactive 3D map (Fig. 2a online) and a sorted dendrogram (Fig. 2c).

A dendrogram of a spectral-line data cube allows for the estimation of key physical properties associated with volumes bounded by isosurfaces, such as radius (R), velocity dispersion (σ_v) and luminosity (L). The volumes can have any shape, and in other work[14] we focus on the significance of the especially elongated features seen in L1448 (Fig. 2a). The luminosity is an approximate proxy for mass, such that $M_{lum} = X_{13CO} L_{13CO}$, where $X_{13CO} = 8.0 \times 10^{20}\ cm^2\ K^{-1}\ km^{-1}\ s$ (ref. 15; see Supplementary Methods and Supplementary Fig. 2). The derived values for size, mass and velocity dispersion can then be used to estimate the role of self-gravity at each point in the hierarchy, via calculation of an ‘observed’ virial parameter, $\alpha_{obs} = 5\sigma_v^2 R/GM_{lum}$. In principle, extended portions of the tree (Fig. 2, yellow highlighting) where $\alpha_{obs} < 2$ (where gravitational energy is comparable to or larger than kinetic energy) correspond to regions of p–p–v space where self-gravity is significant. As α_{obs} only represents the ratio of kinetic energy to gravitational energy at one point in time, and does not explicitly capture external over-pressure and/or magnetic fields[16], its measured value should only be used as a guide to the longevity (boundedness) of any particular feature.

Figure 3 | Schematic illustration of the dendrogram process. Shown is the construction of a dendrogram from a hypothetical one-dimensional emission profile (black). The dendrogram (blue) can be constructed by ‘dropping’ a test constant emission level (purple) from above in tiny steps (exaggerated in size here, light lines) until all the local maxima and mergers are found, and connected as shown. The intersection of a test level with the emission is a set of points (for example the light purple dots) in one dimension, a planar curve in two dimensions, and an isosurface in three dimensions. The dendrogram of 3D data shown in Fig. 2c is the direct analogue of the tree shown here, only constructed from ‘isosurface’ rather than ‘point’ intersections. It has been sorted and flattened for representation on a flat page, as fully representing dendrograms for 3D data cubes would require four dimensions.

Goodman et al. Nature, 2009

Secret truths I learned from AstroMed at IIC

1. The story about the hammer & the screw is a cliché, but it's true. Don't use "more" software than you must. (e.g. 3DSlicer/Osirix/VolView/AplPy/S2PLOT/custom apps)
2. Smart hackers rule. Software engineering is great when you have the money & time.
3. Distribution is key. There's too much great S/W no one knows about.

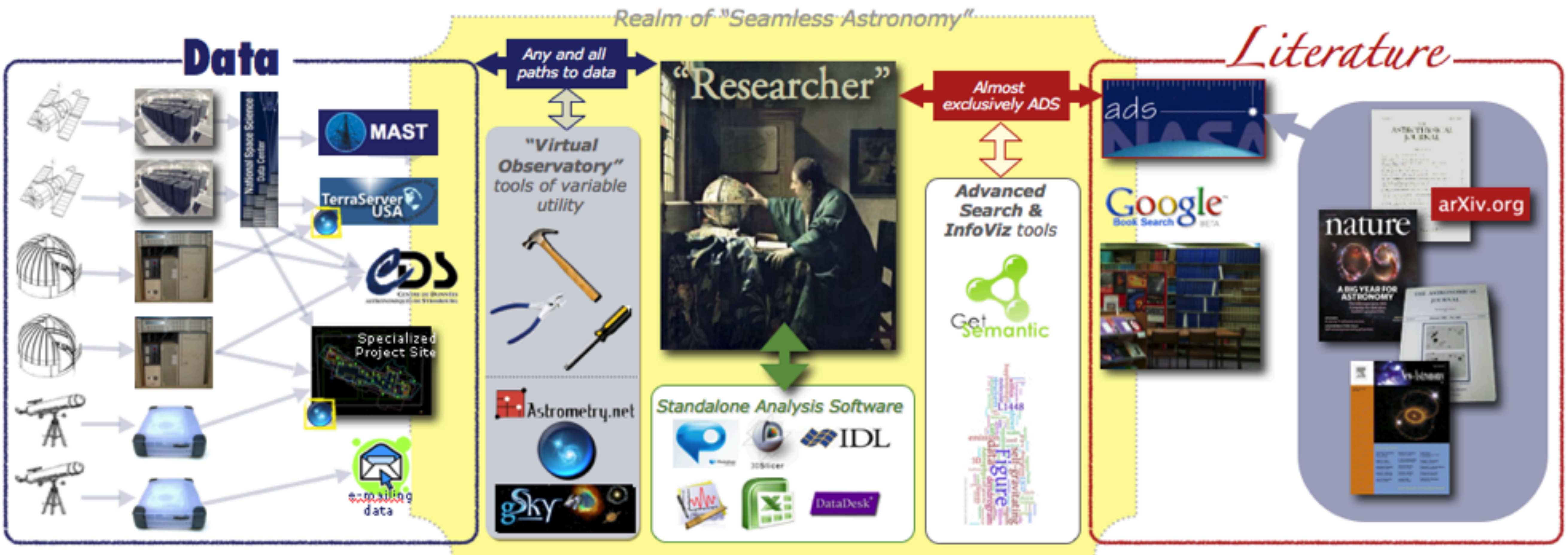
My "advice"
re:IIC-like Enterprises

Alyssa Goodman's Advice to Tim Kaxiras and Ros Reid, about the IIC, May 2008

Key principles to be maintained, or established:

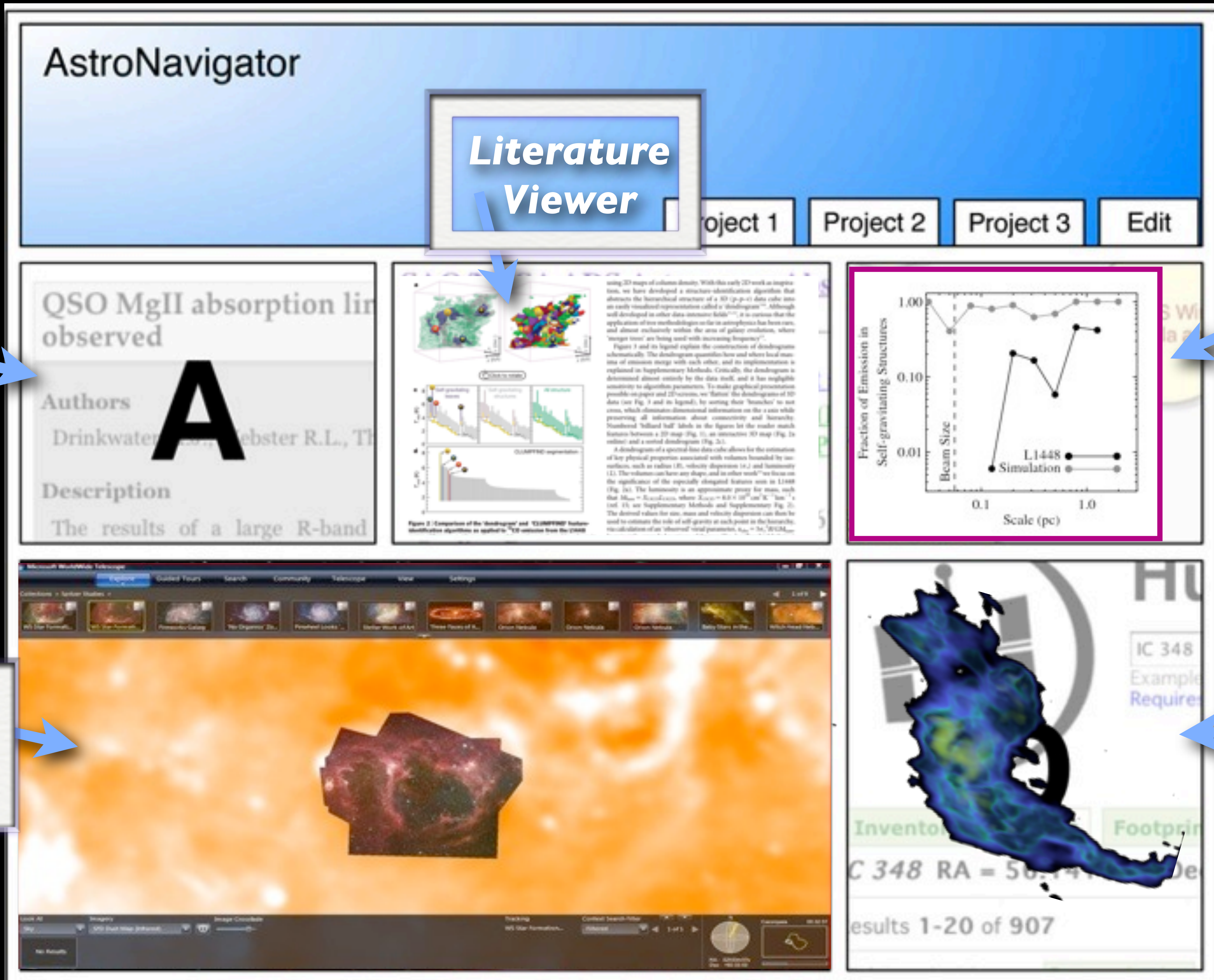
1. Be firm about the expected **life cycle of projects**. Ramp-up and ramp-down periods, and re-evaluation of progress ~bi-annually is essential.
2. Maintain the IIC **"core" technical staff** principles. Do not turn IIC into a "soft-money" institute. A good goal may be 50/50 internal/external funding. Solving the O/H problem could potentially bring some "internal" money into IIC as either "recovered" O/H or services received for O/H.
3. Maintain a strong **Seminar series**. Many "converts" have come to the IIC through the Seminar series.
4. Maintain, and expand, **student-centric programs**. The Summer Internships are good, but there should be more term-time participation by students.
5. Create a set of clear guidelines for what it means to be **faculty** with a partial appointment at IIC (salary from IIC).
6. Create a clear definition of what affiliated faculty (**"Core Members"**) should expect to give/take from IIC. (See Business Plan ideas.)
7. Establish IIC as the "coordinator" for **DISC**. URGENT the IIC convene DISC meetings ASAP. Greenhill is IIC representative, and he and Cuff should host meetings.
8. Establish IIC as a center for experimentation through **internal projects**, e.g. SDR.
9. Establish rigorous proposal **review** process.
10. Form and make good use of **external advisory** committee(s).

What I'm doing now...



"Seamless Astronomy" is collaboration amongst many researchers at CfA, MSR, Princeton, STScI, NYU, RPI, and UCLA, and it is supported by NASA, NSF and Microsoft External Research.

Seamless Astronomy



Semantic Search

Data Viewer (e.g. WWT)

Info-Vis for Analytics

Sift-based Search Results

Archive Viewer

3D Browser

Mockup based on work of Eli Bressert, excerpted from NASA AISRP proposal by Goodman, Muench, Christian, Conti, Kurtz, Burke, Accomazzi, McGuinness, Hendler & Wong, 2008

Touching Insight

The Scientists' Discovery Room

movie courtesy Daniel Wigdor, equipment now in Chia Shen's SDR lab at SEAS