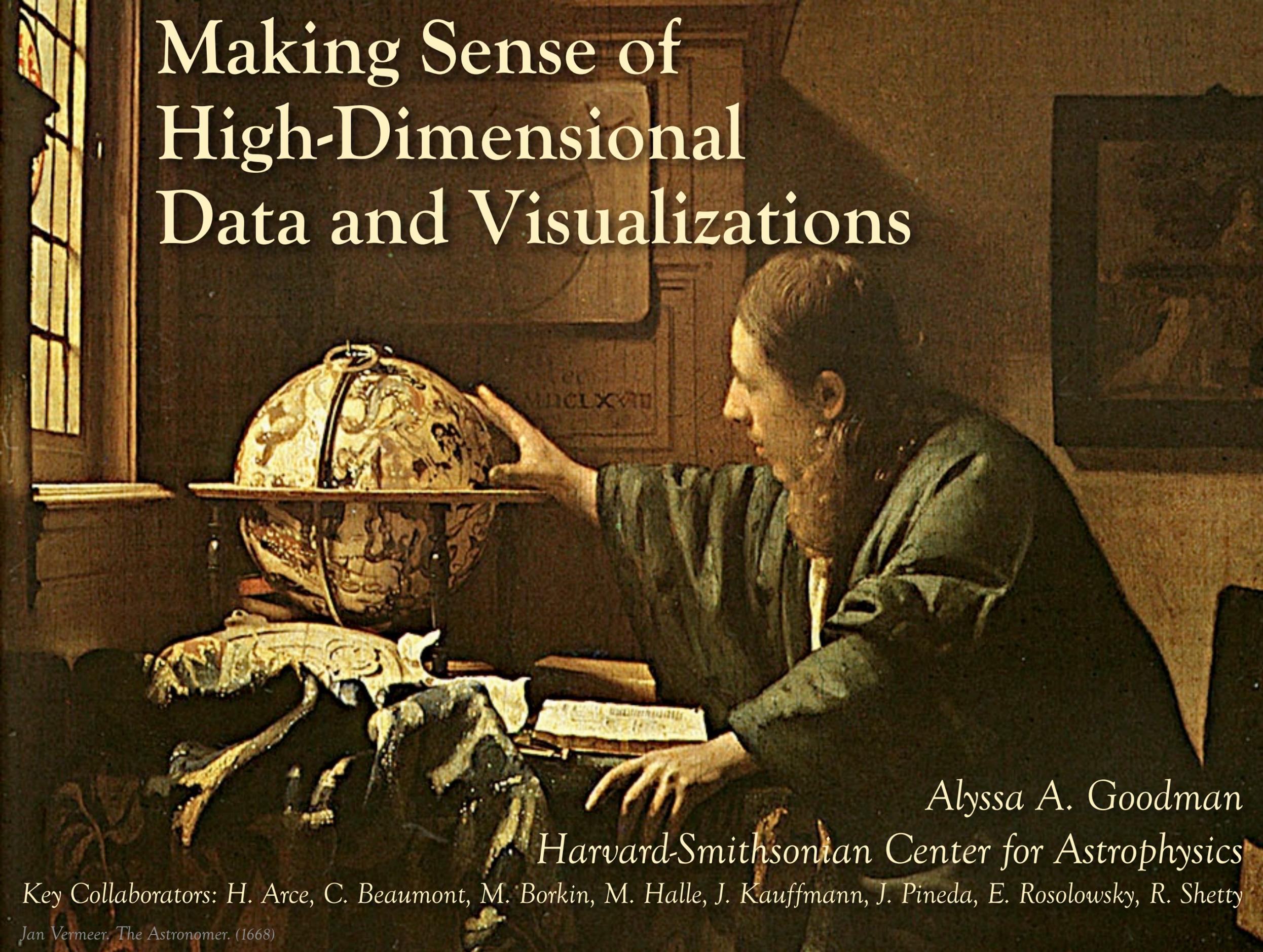


# Making Sense of High-Dimensional Data and Visualizations



*Alyssa A. Goodman*

*Harvard-Smithsonian Center for Astrophysics*

*Key Collaborators: H. Arce, C. Beaumont, M. Borkin, M. Halle, J. Kauffmann, J. Pineda, E. Rosolowsky, R. Shetty*

*Jan Vermeer. The Astronomer. (1668)*

Tuesday, March 22, 2011

The "data deluge" in science is old news. Now, it's pouring, and we need working tools to collect, sort out, understand, and keep what is falling down on us. In astronomy, the greatest insights very often come from studies where more than one "band" of data (e.g. optical, infrared, radio, X-ray) is combined. And, data sets aren't just large--they are often also high-dimensional, in that they contain information about flux as functions not just of position on the sky, but also as functions of a third dimension (e.g. frequency, velocity), and/or of time. Life science, geophysical, and geospatial data all present similar challenges.

In this talk, I will focus on examples drawn from my group's research on star formation in molecular clouds. In particular, I will show how new visualization and statistical analysis techniques relying on interactive high-dimensional views of data and on automated algorithms for "segmenting" data give new insight. "Segmentation" in imaging terms refers to extracting the meaningful structures from data, and I will show results from both dendrogram (tree-hierarchy) and machine-learning approaches. I will emphasize how the visualization of segmentation results is critical for understanding. The highlighted science results will show how we can now--for the first time--quantitatively but intuitively understand the connections between the "real" (position-position-position) space where simulations (e.g. of star formation) can be made and the "observational" (e.g. position-position-velocity) space available to earthbound astronomers. As a result of this newfound understanding, we can place important limits on the validity of virial-theorem-based calculations of the properties of gas--allowing, for example, for better estimates of which gas in star-forming regions is most likely to stay bound long enough to form stars.

Even though this abstract may sound technical to non-star-formation or non-computational researchers, my goal will be to keep the talk accessible to non-experts, so people from other fields faced with high-dimensional data and visualization challenges should feel free to join in--and to ask questions

# Introduction

High-Dimensional  
Data

AstroMed

Simulations

3D PDF

Linked Views,  
e.g. Dendro...

Machine Learning

WWT—"NUIs"—Seamless  
Astronomy

Star Formation

"p-p-v" cubes

True 3D  
Structure

What's bound?/  
Virial Theorem

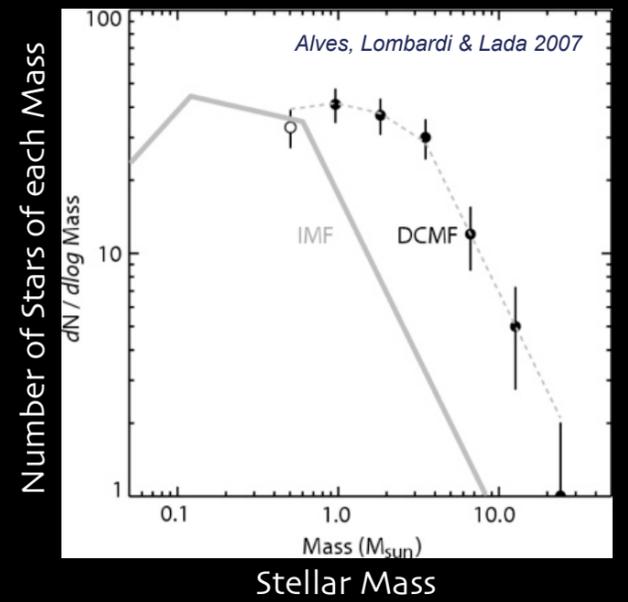
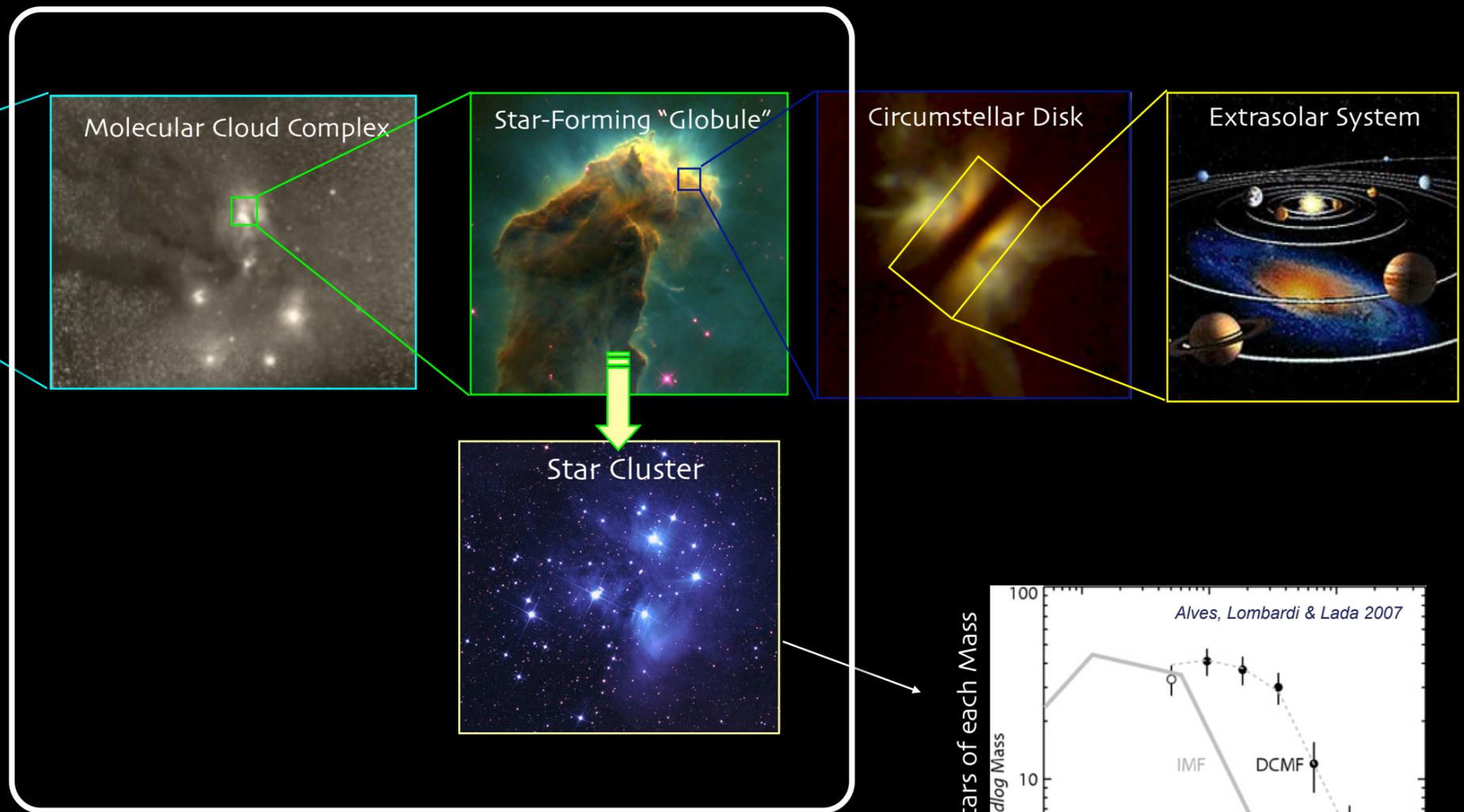
What riles up  
the ISM?

3D Milky Way,  
Predictive KS?

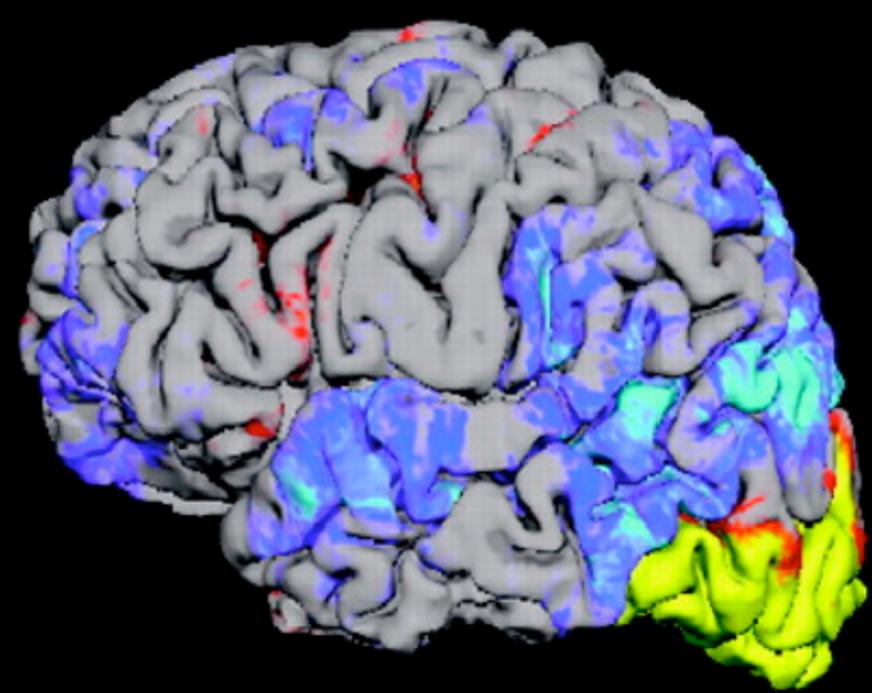
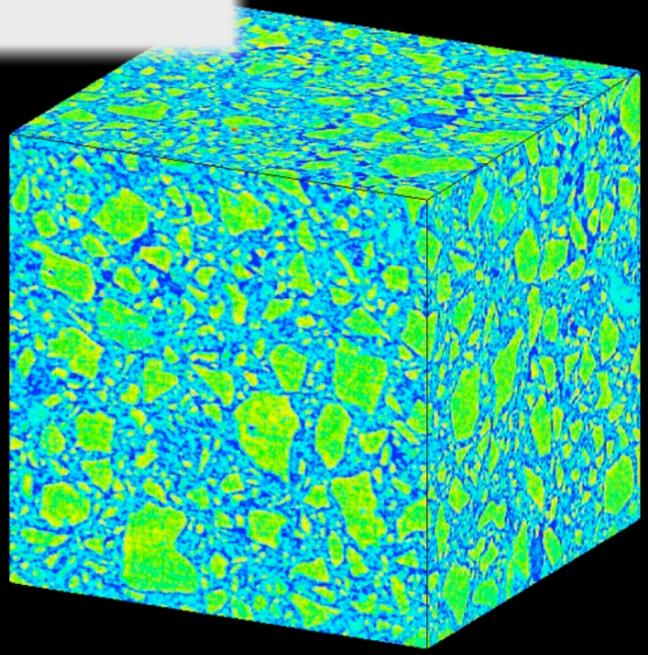
# The Future



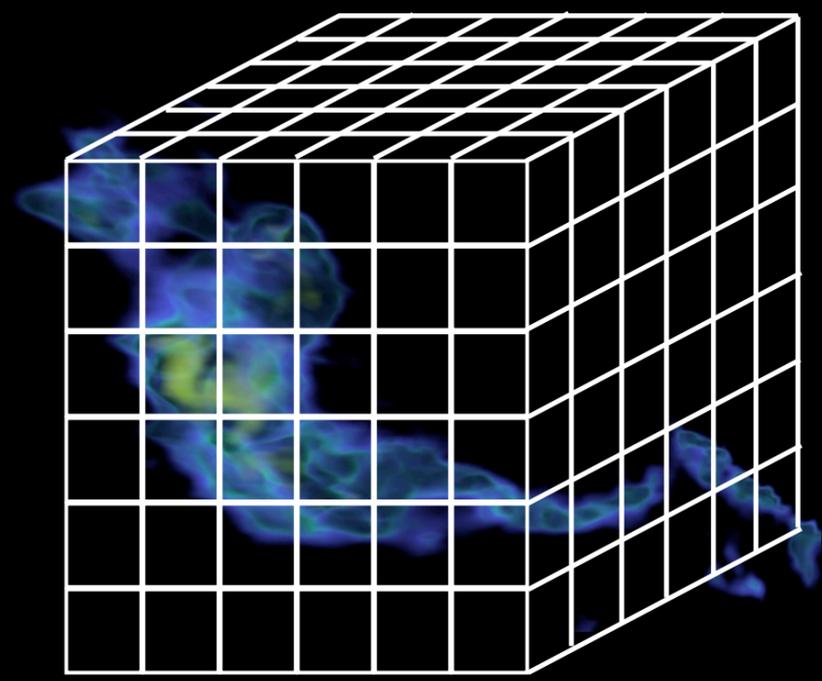
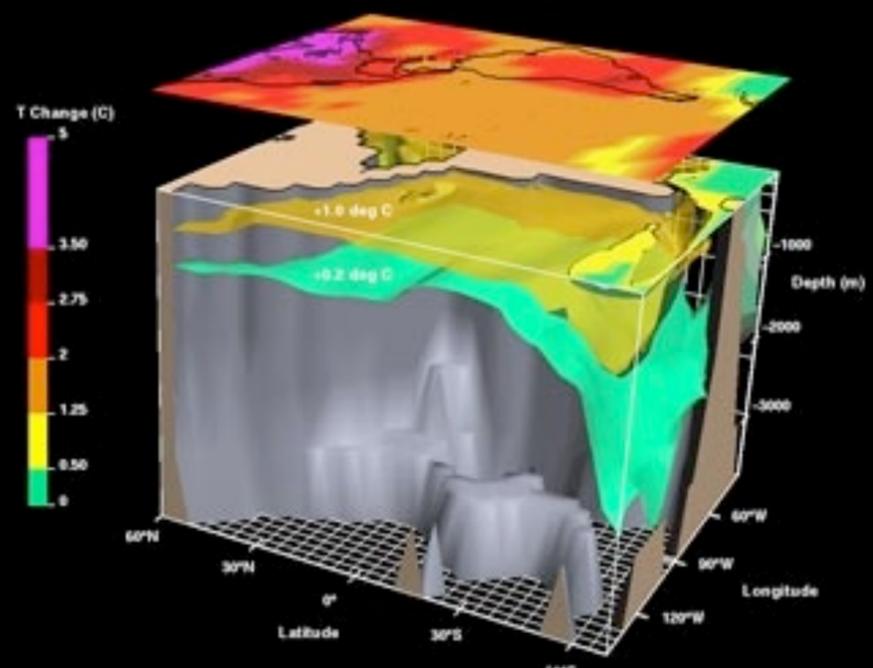
# Star (and Planet, and Moon) Form Star Formation

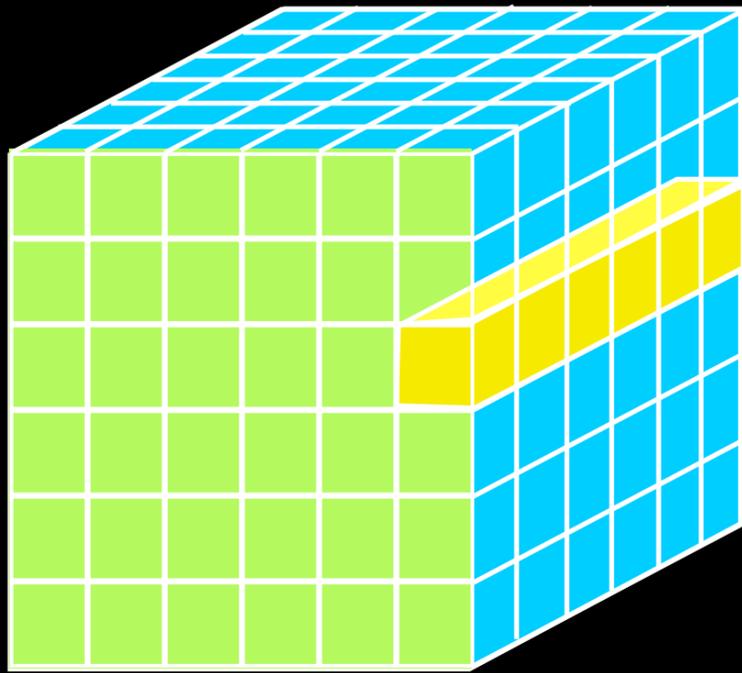


# High-Dimensional Data



ATMOSPHERIC AND OCEANIC TEMPERATURE CHANGE





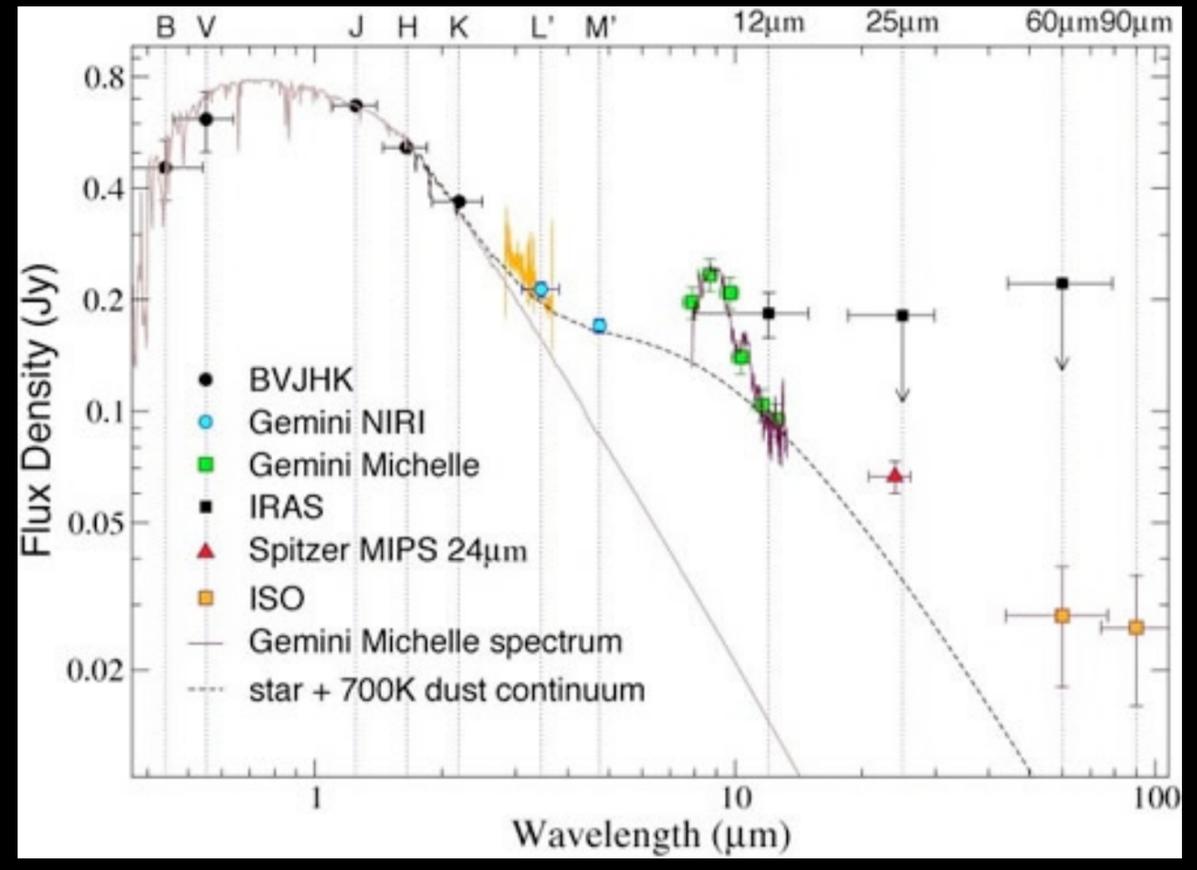
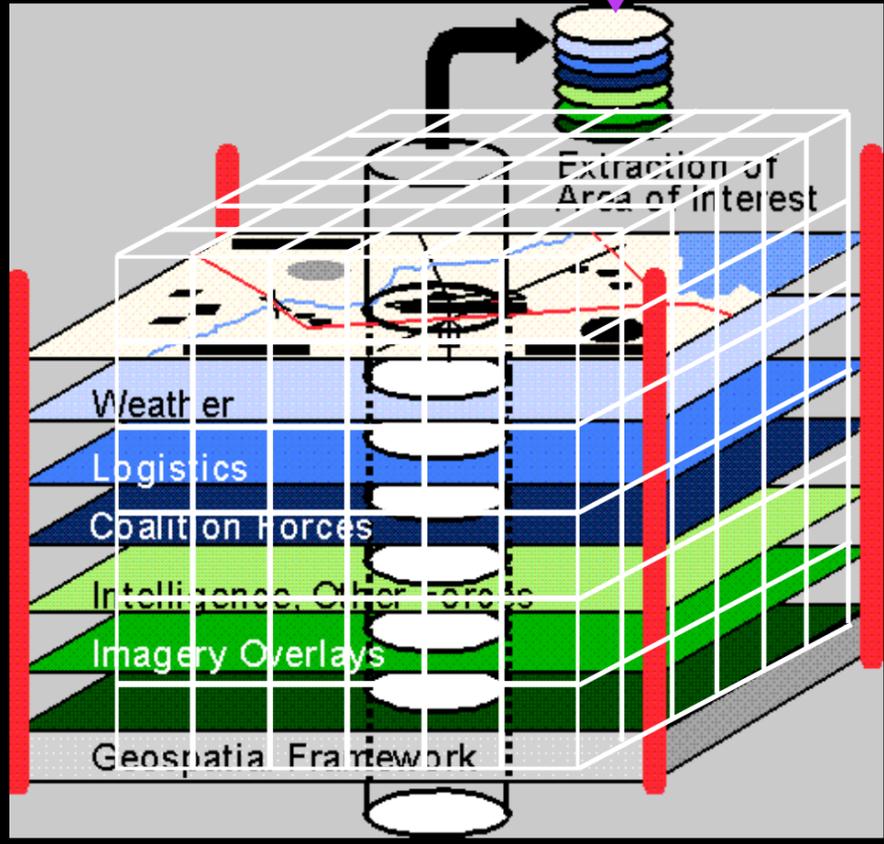
## GENERALLY

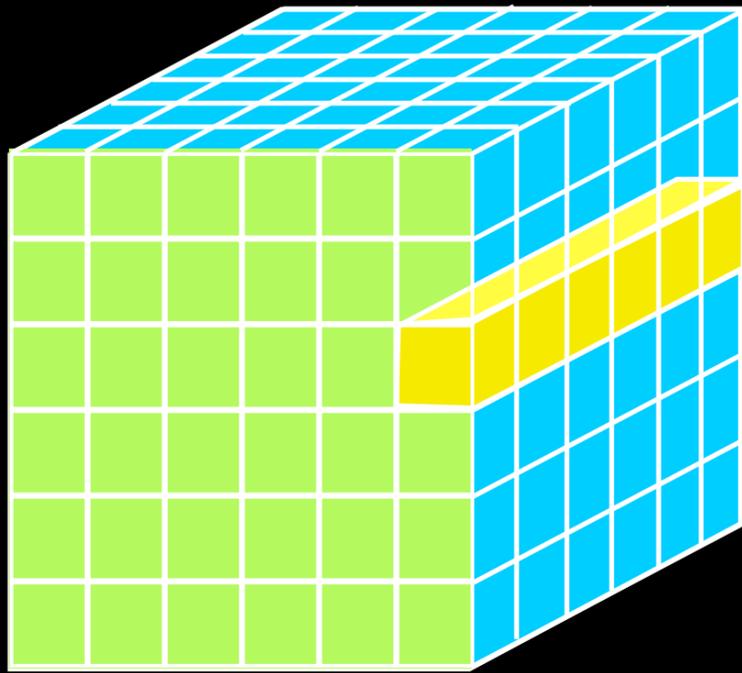
- 1D:** Columns = “Spectra”, “SEDs” or “Time Series”
- 2D:** Faces or Slices = “Images”
- 3D:** Volumes = “3D Renderings”
- 4D:** Time Series of Volumes = “3D Movies”

High-Dimensional  
Data

This  
↓

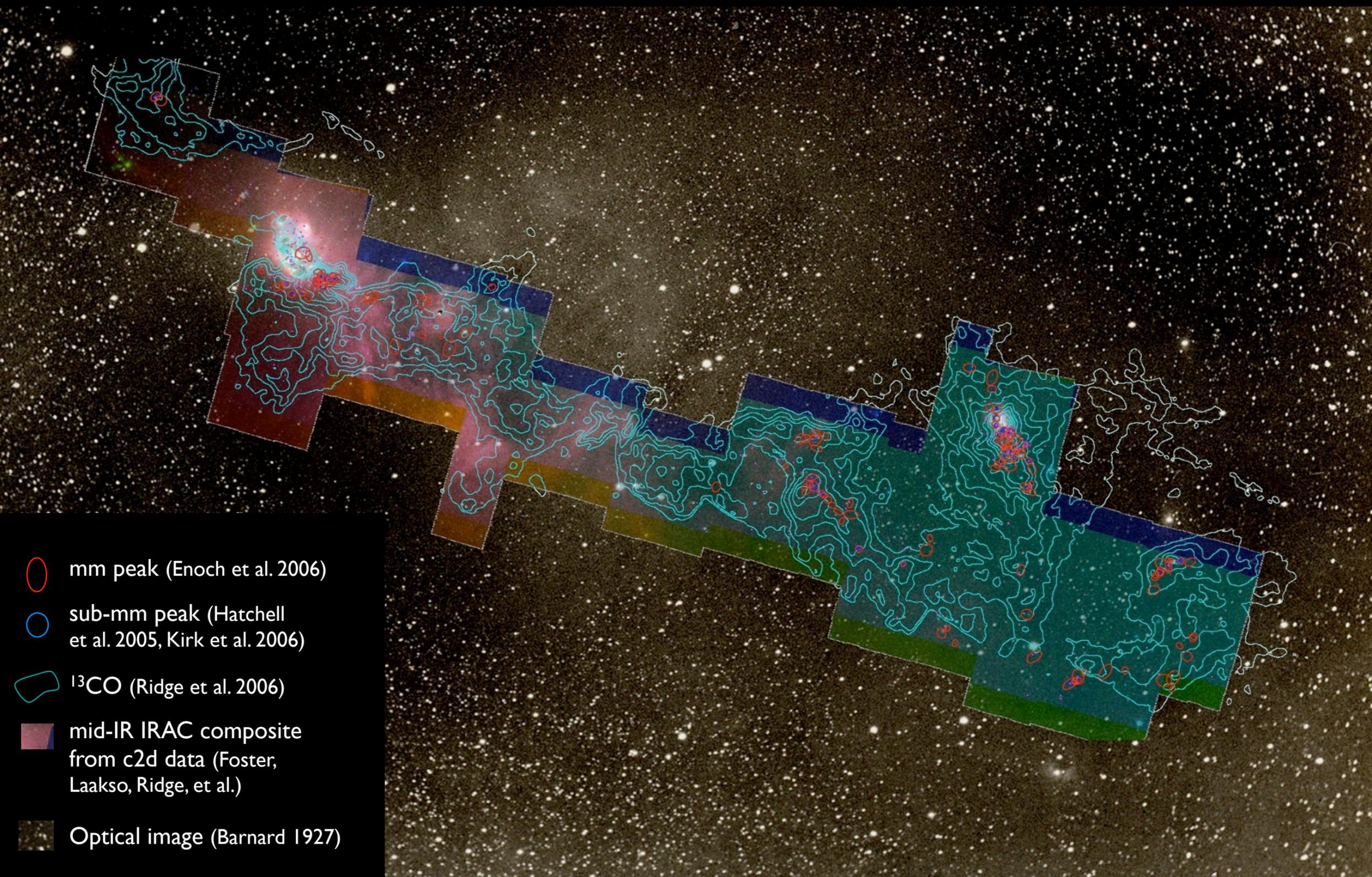
is a “spectral energy distribution”





## GENERALLY

- 1D:** Columns = “Spectra”, “SEDs” or “Time Series”
- 2D:** Faces or Slices = “Images”
- 3D:** Volumes = “3D Renderings”
- 4D:** Time Series of Volumes = “3D Movies”

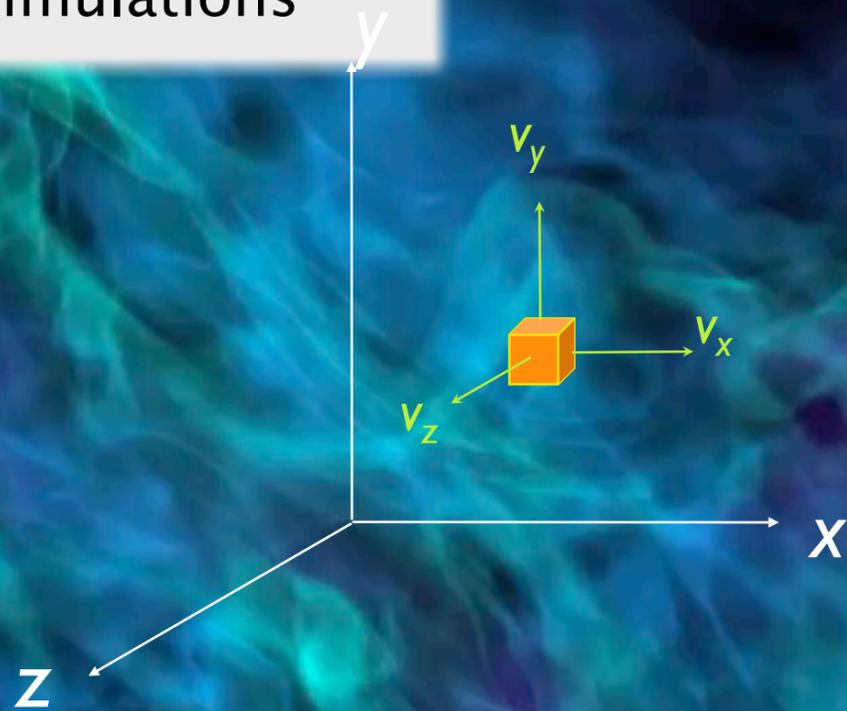


# “Three” Dimensions: Spectral-Line Mapping

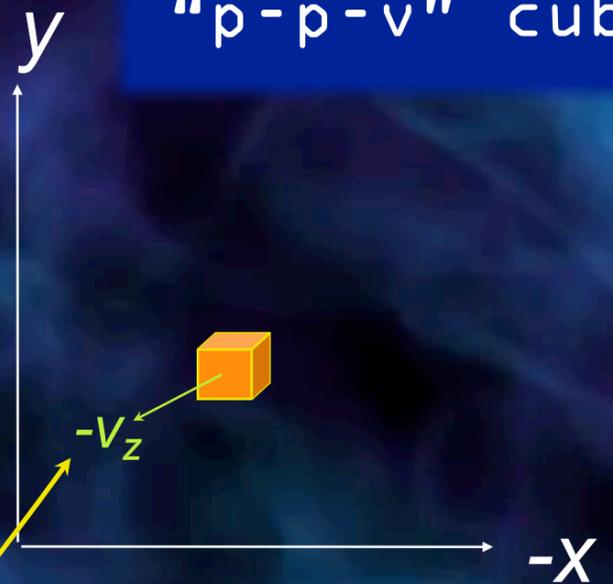
We wish we could measure...

*But* we can measure...

Simulations



“p-p-v” cubes

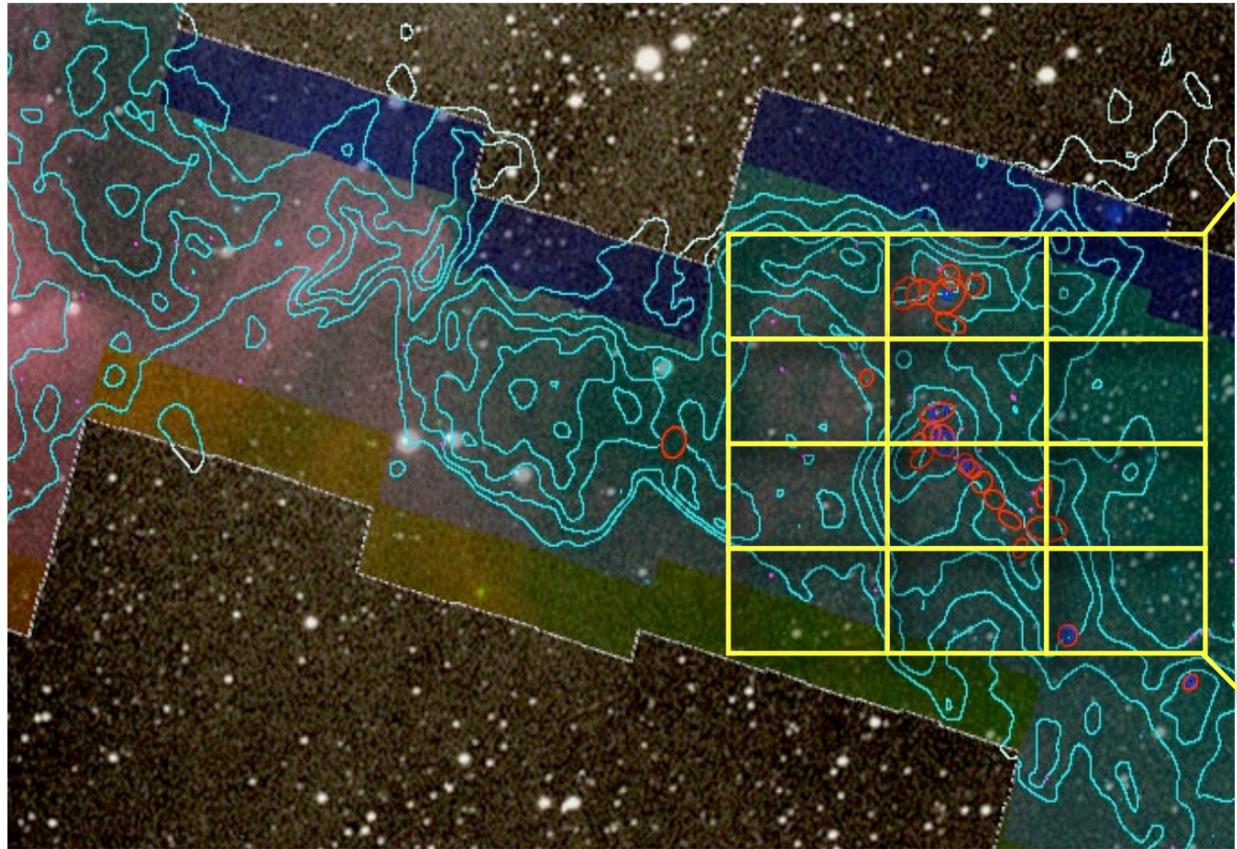


$v_z$  *only* from  
“spectral-line  
maps”

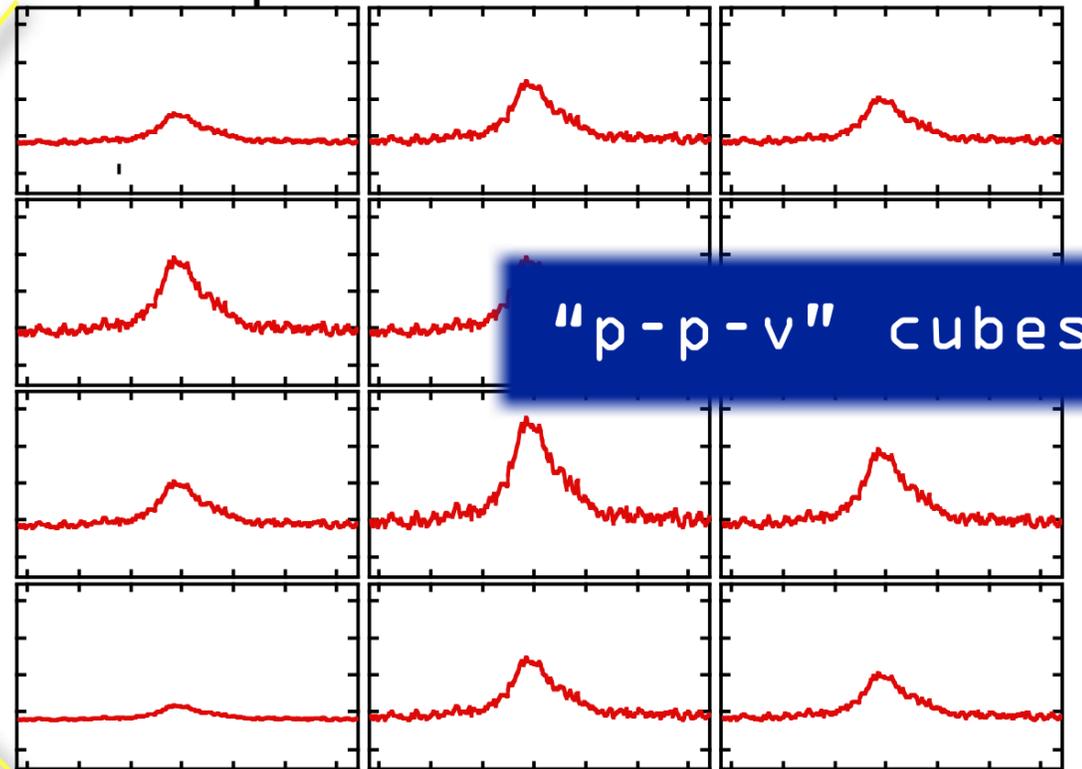
*Hydrodynamic AMR Simulation, courtesy Stella Offner*



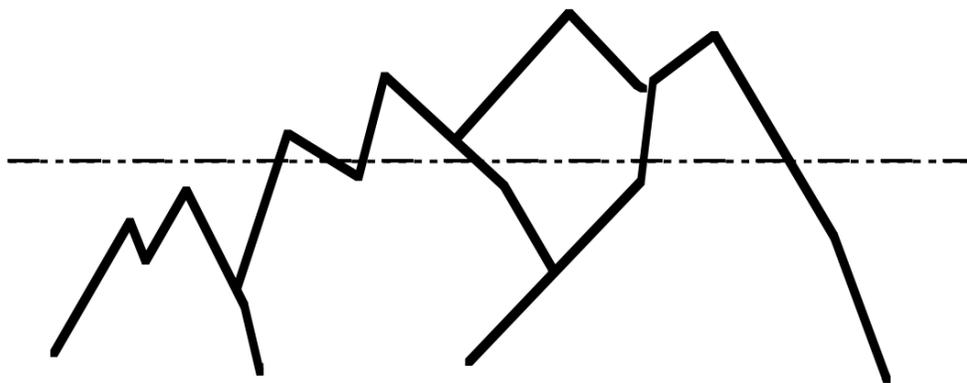
# Spectral-Line Mapping



Spectral Line Observations



"p-p-v" cubes



Mountain Range

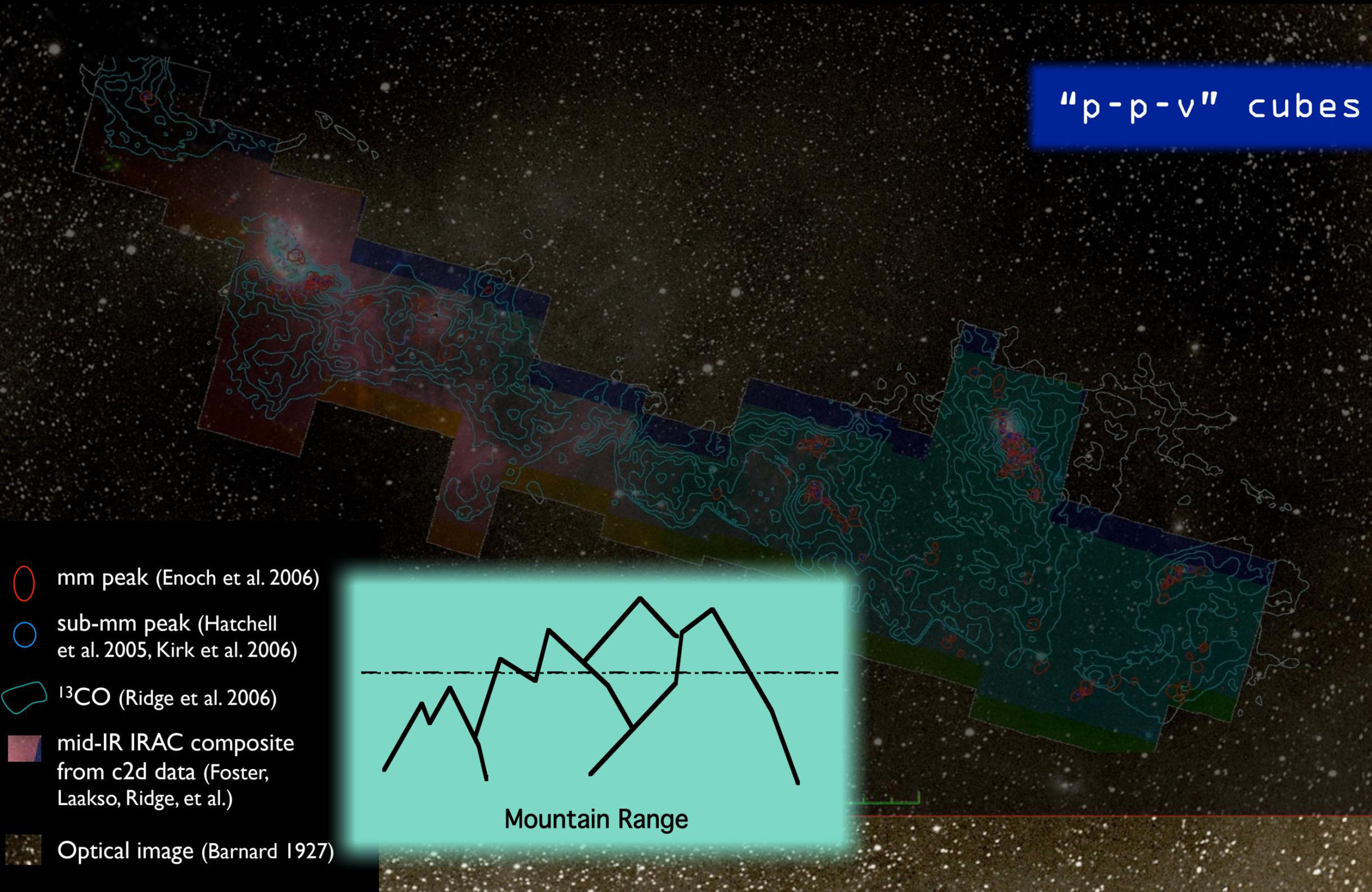


No loss of information



Loss of  
1 dimension

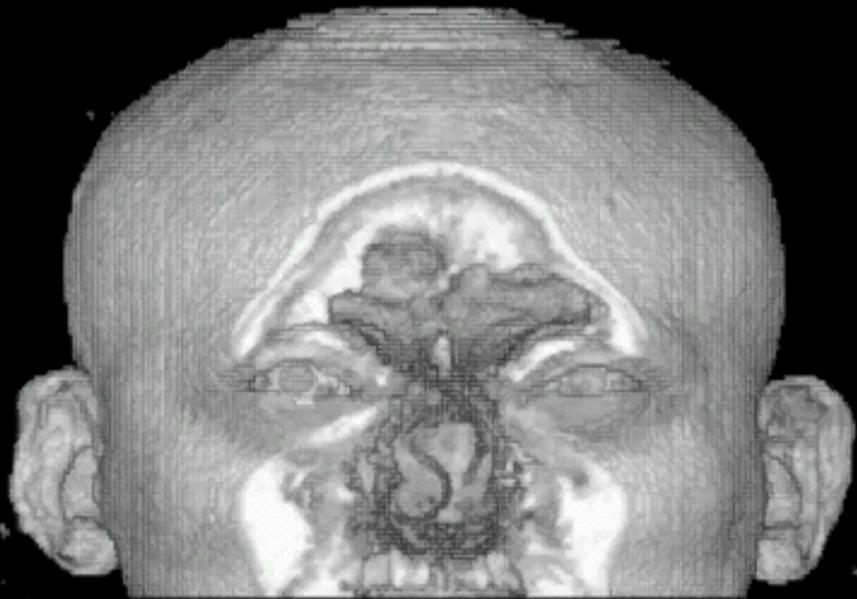
"p-p-v" cubes



# “Astronomical Medicine”

AstroMed

“KEITH”



“z” is depth into head

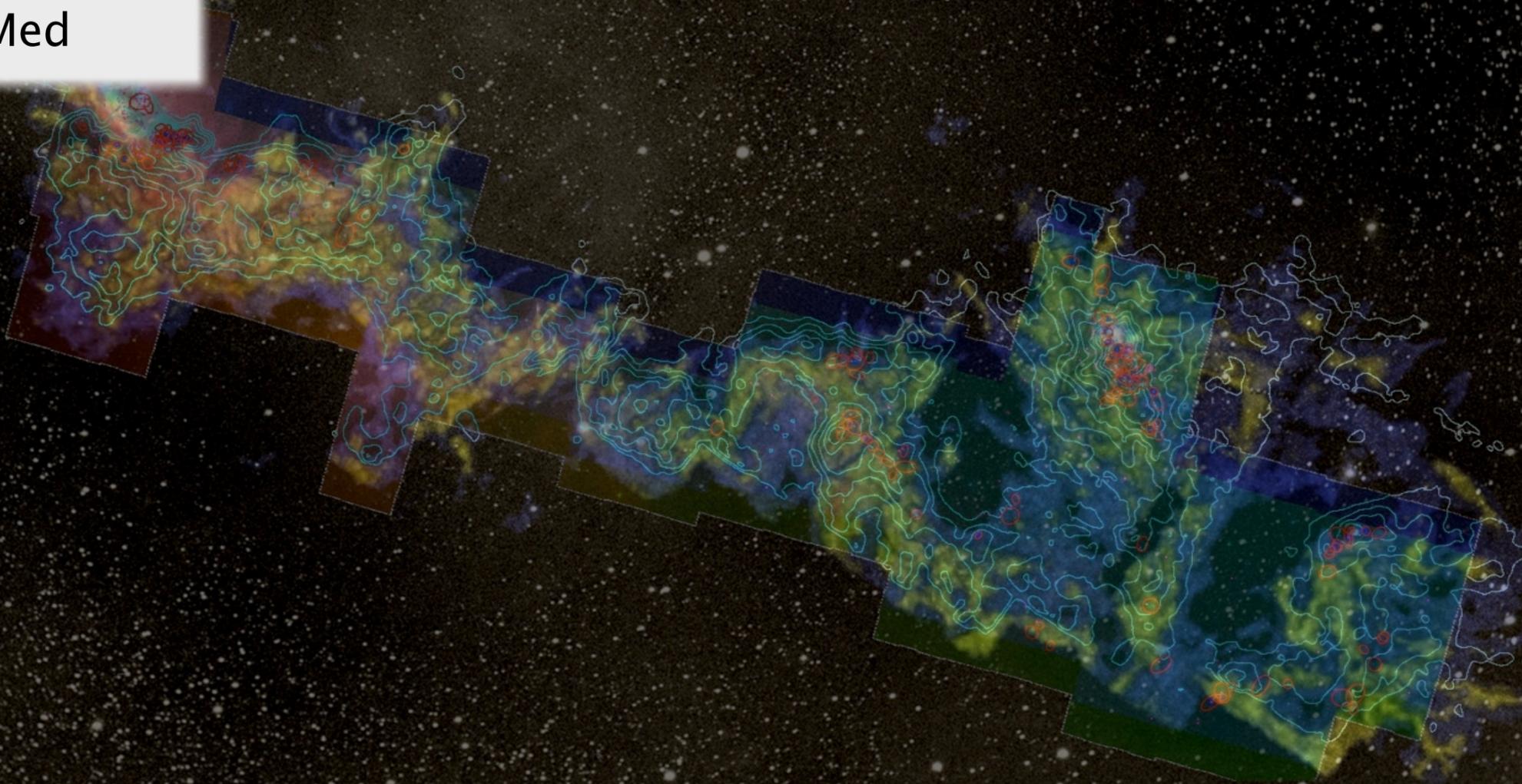
“PERSEUS”



“z” is line-of-sight velocity

<http://am.iic.harvard.edu/>

AstroMed



Perseus

3D Viz made with VolView

AstronomicalMedicine@iig

COMPLETE

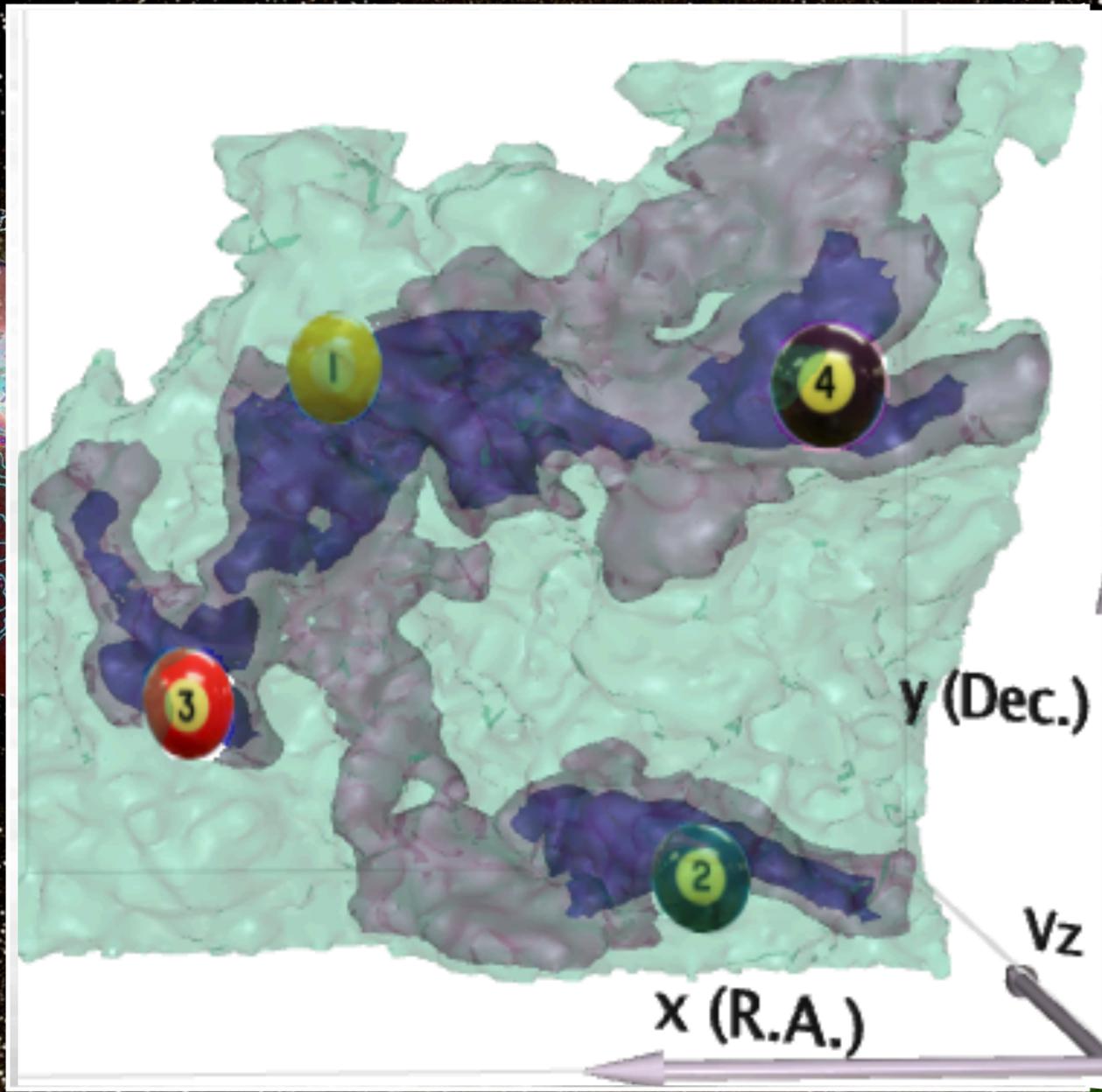
Tuesday, March 22, 2011

Star Formation

True 3D Structure

What's bound? / Virial Theorem

"L1448+"

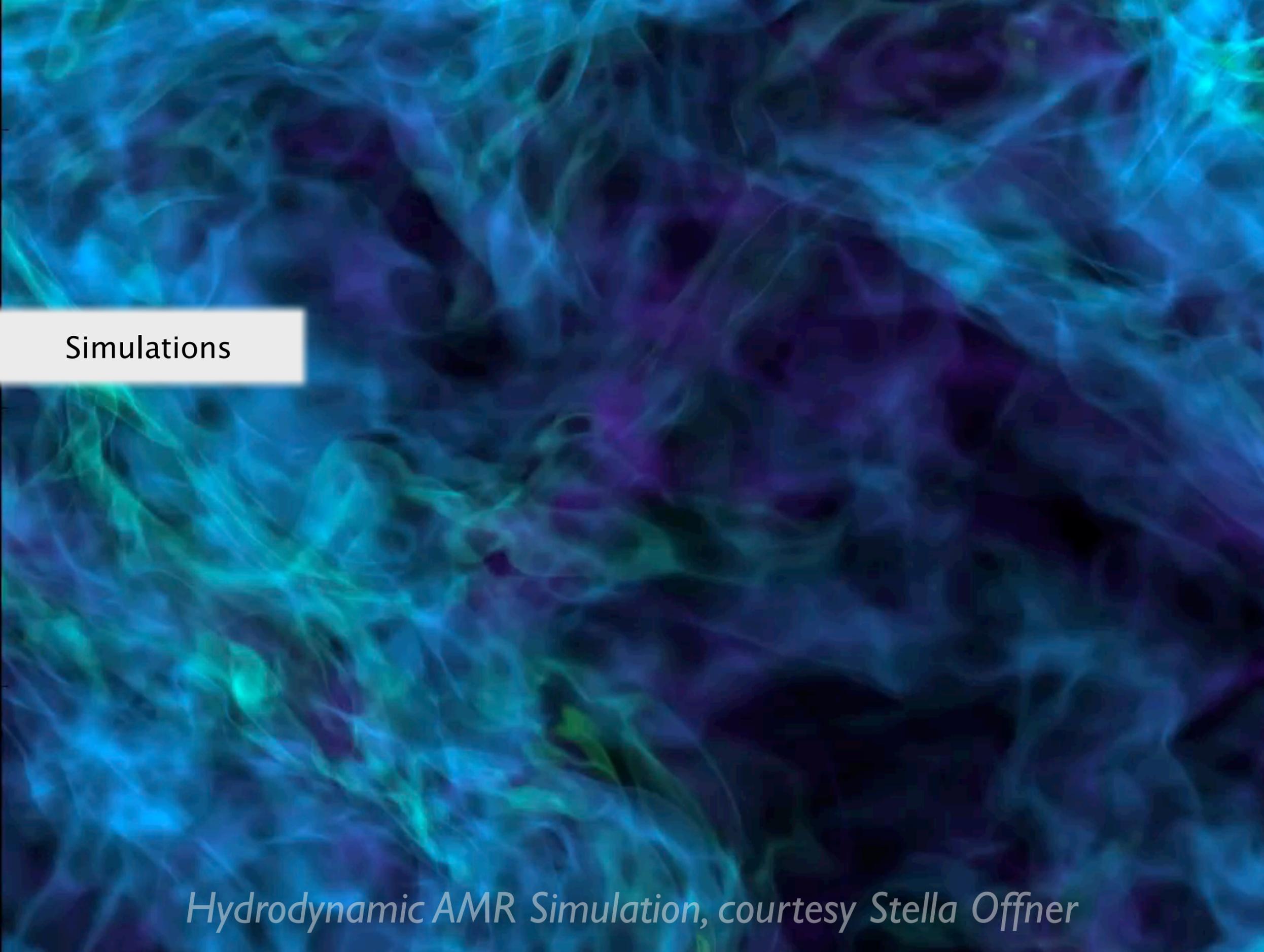


Perseus

WWT-"NUIs"-Seamless Astronomy

COMPLETE

Tuesday, March 22, 2011

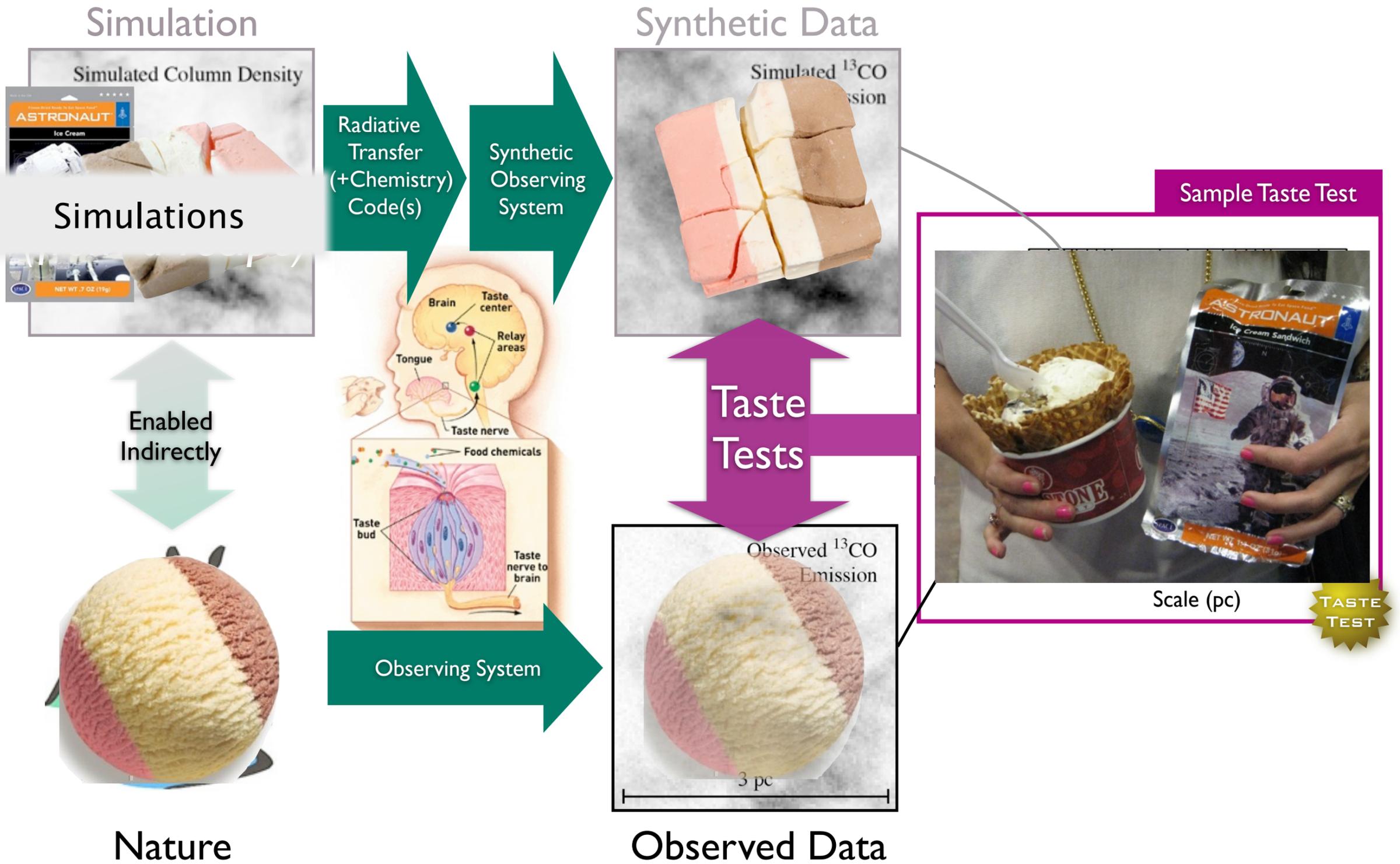


Simulations

*Hydrodynamic AMR Simulation, courtesy Stella Offner*

Tuesday, March 22, 2011

# “Taste-Testing” Simulations



Magnetic  
Fields

Gravity

Chemical & Phase  
Transformations

~1 pc

Radiation

Thermal  
Pressure

“Turbulence”  
(Random Kinetic Energy)

Outflows  
& Winds

Image Credit: Jonathan Foster & Jaime Pineda CfA/COMPLETE Deep Megacam Mosaic of West End of Perseus

High-Dimensional Data

# Taste-Testing "Gravity"

Star Formation

AstroMed

"p-p-v" cubes

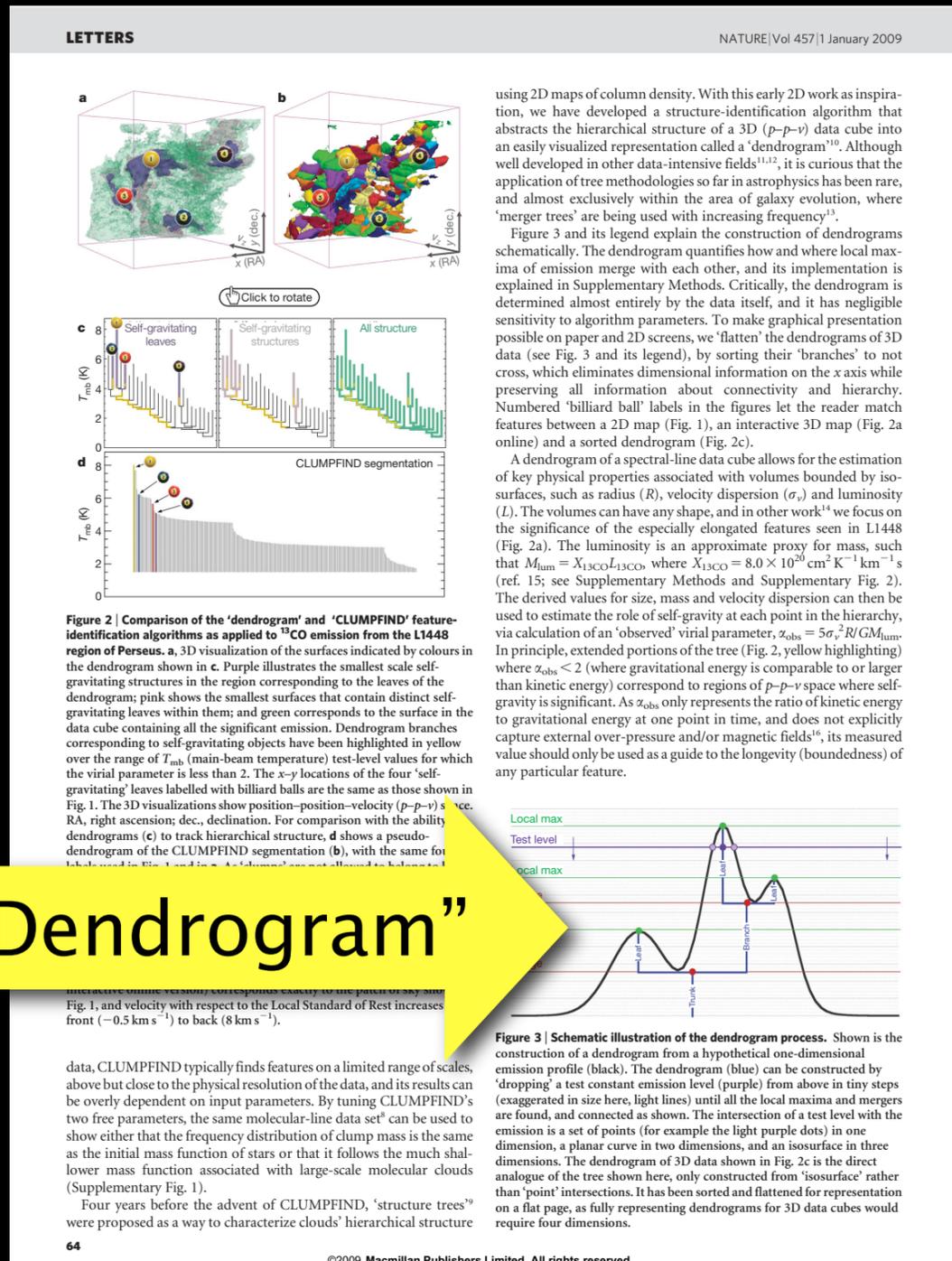
Simulations

True 3D Structure

3D PDF

What's bound? / Virial Theorem

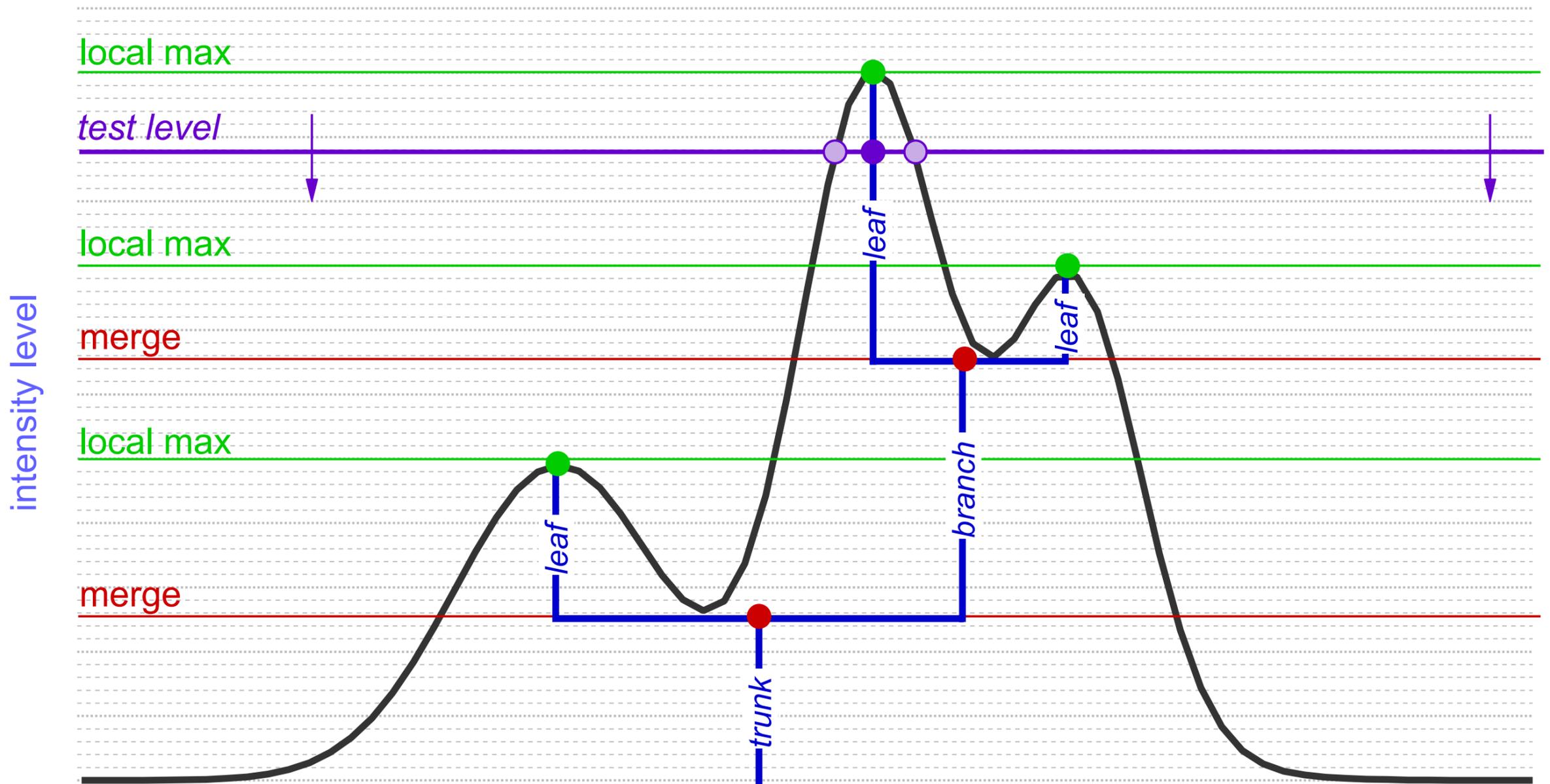
"Dendrogram"



Goodman et al. Nature, 2009



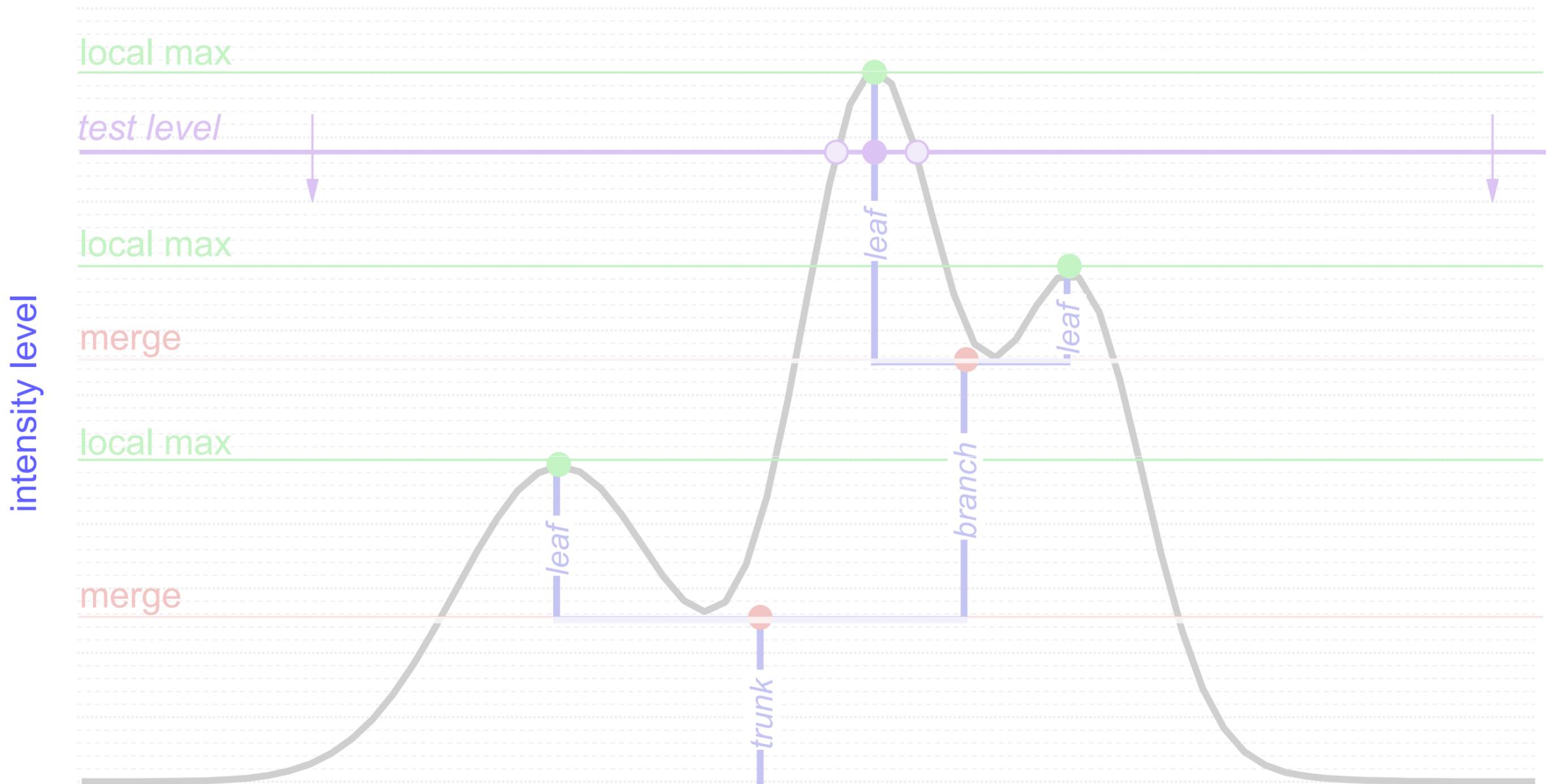
# Dendrograms



## Hierarchical "Segmentation"

*Rosolowsky, Pineda, Kauffmann & Goodman 2008*

# Dendrograms



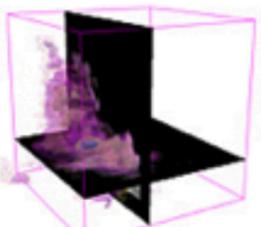
I-D: points; 2-D closed curves (contours); 3-D surfaces enclosing volumes

see 2D demo at <http://am.iic.harvard.edu/index.cgi/DendroStar/applet>

DendroStar/applet - IIC/AstroMed

http://am.iic.harvard.edu/index.cgi/DendroStar/applet

astronomical medicine



The Astronomical Medicine Project Initiative In Innovative Computing at Harvard

### Harvard IIC Home

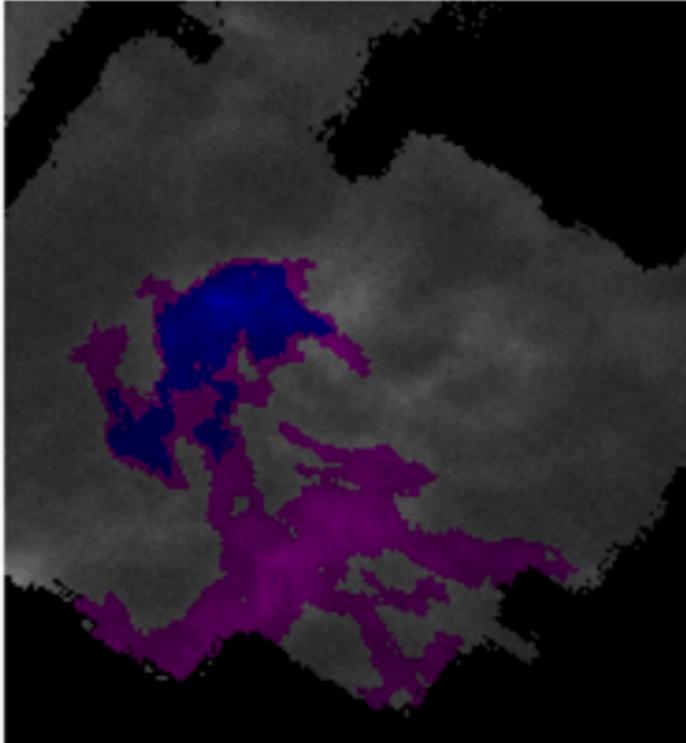
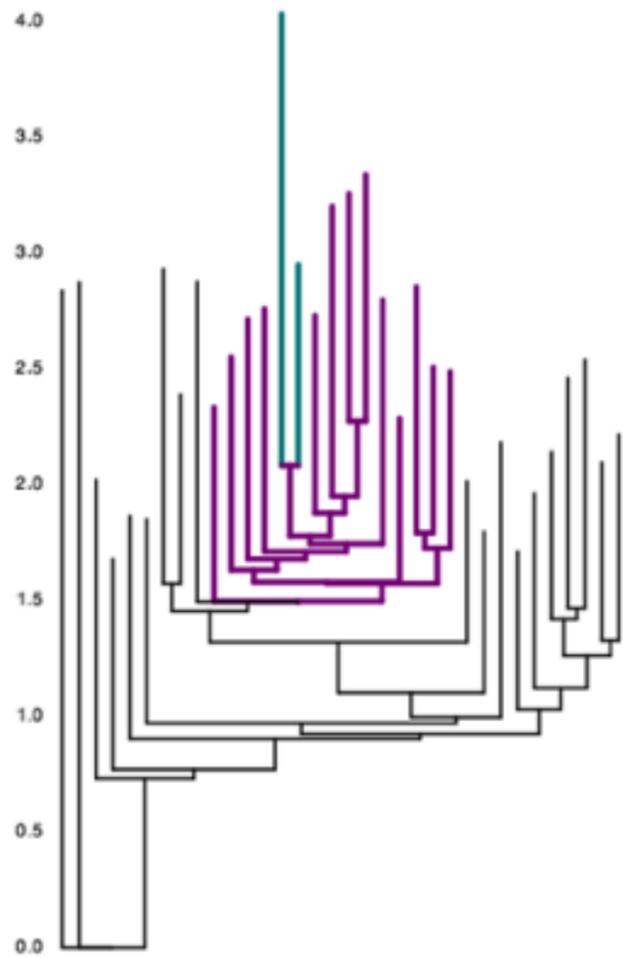
- AM Project**
  - overview
  - what's new?
  - press
  - about us
  - contact us
- Research**
  - background
  - projects
  - papers
  - images
  - movies
- Software**
  - overview
  - Slicer: getting started
  - Slicer 3
  - fits2itk
  - OsiriX
  - DendroStar

related projects

**User**  
Login

**Search**  
Search  
Titles Text

### The DendroStar Applet for L1448: Try me!



Tint:

Suppress tint:

Reset:

Applet DendroStar started

Linked Views,  
e.g. Dendro...

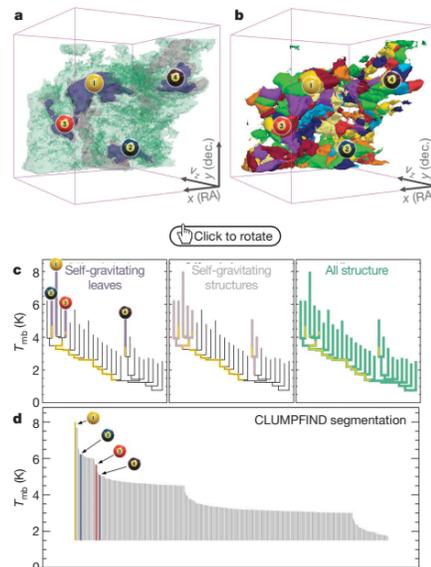
<http://am.iic.harvard.edu/index.cgi/DendroStar/applet>  
Dendrogram Algorithm by Erik Rosolwosky; Applet by Douglas Alan

# Taste-Testing "Gravity"

3D PDF

LETTERS

NATURE | Vol 457 | January 2009



**Figure 2** | Comparison of the 'dendrogram' and 'CLUMPFIND' feature-identification algorithms as applied to  $^{13}\text{CO}$  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  $T_{\text{mb}}$  (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 ( $-0.5 \text{ km s}^{-1}$ ) to back ( $8 \text{ km s}^{-1}$ ).

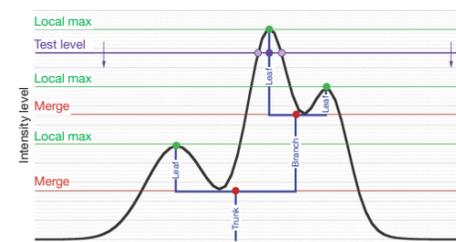
data, CLUMPFIND typically finds features on a limited range of scales, above but close to the physical resolution of the data, and its results can be overly dependent on input parameters. By tuning CLUMPFIND's two free parameters, the same molecular-line data set<sup>6</sup> can be used to show either 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'<sup>9</sup> 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'<sup>10</sup>. Although well developed in other data-intensive fields<sup>11,12</sup>, 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<sup>13</sup>.

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 ( $\sigma_v$ ) and luminosity ( $L$ ). The volumes can have any shape, and in other work<sup>14</sup> 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  $M_{\text{lum}} = X_{13\text{CO}} L_{13\text{CO}}$ , where  $X_{13\text{CO}} = 8.0 \times 10^{20} \text{ cm}^2 \text{ K}^{-1} \text{ s}$  (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,  $\alpha_{\text{obs}} = 5\sigma_v^2 R/GM_{\text{lum}}$ . In principle, extended portions of the tree (Fig. 2, yellow highlighting) where  $\alpha_{\text{obs}} < 2$  (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  $\alpha_{\text{obs}}$  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<sup>16</sup>, 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.

True 3D  
Structure

What's bound?/  
Virial Theorem

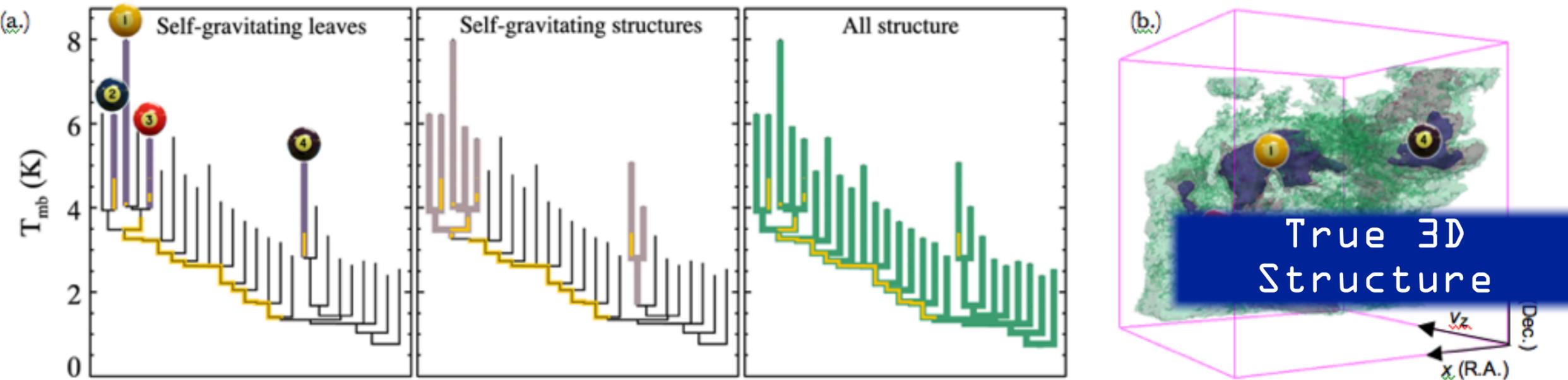


Goodman et al. Nature, 2009



Tuesday, March 22, 2011

# What's Bound? Can we Know?



What's bound? /  
Virial Theorem

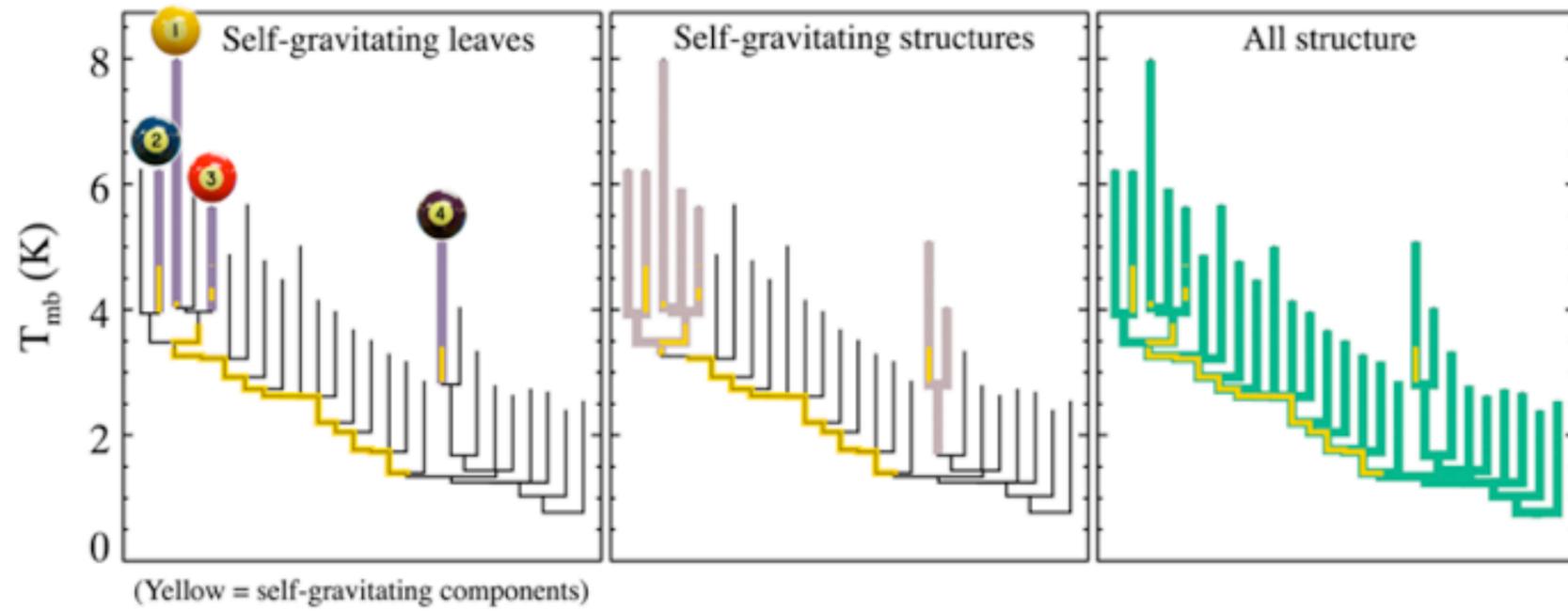
**Yellow** highlighting= “self-gravitating”

“Self-gravitating” here just means  $\alpha_{vir} (=5s_v^2R/GM_{lum}) < 2$   
(à la Bertoldi & McKee 1992—*BUT*—see Shetty et al. 2010)

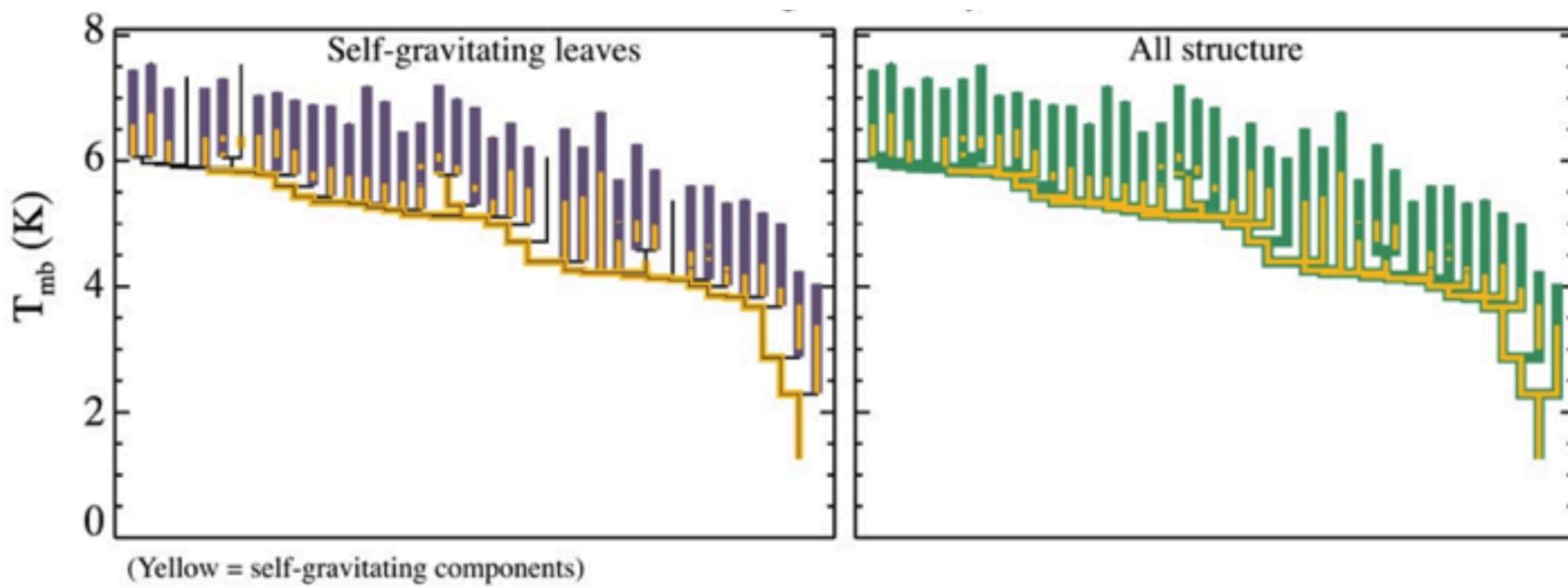
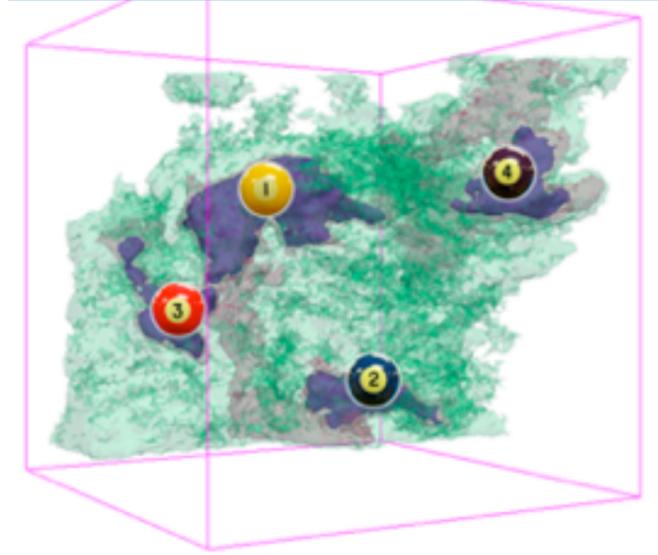
*Rosolowsky et al. 2008 (ApJ) &  
Goodman et al. 2009 (Nature)*

see PDF...

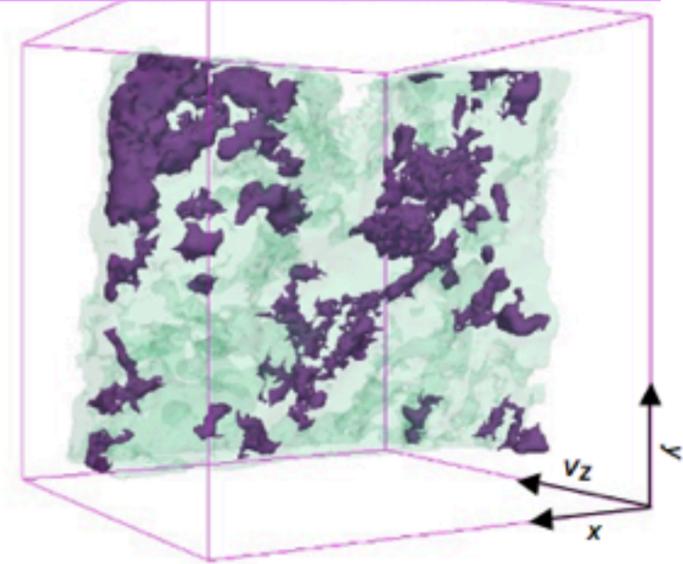
# Real and Simulated $^{13}\text{CO}$



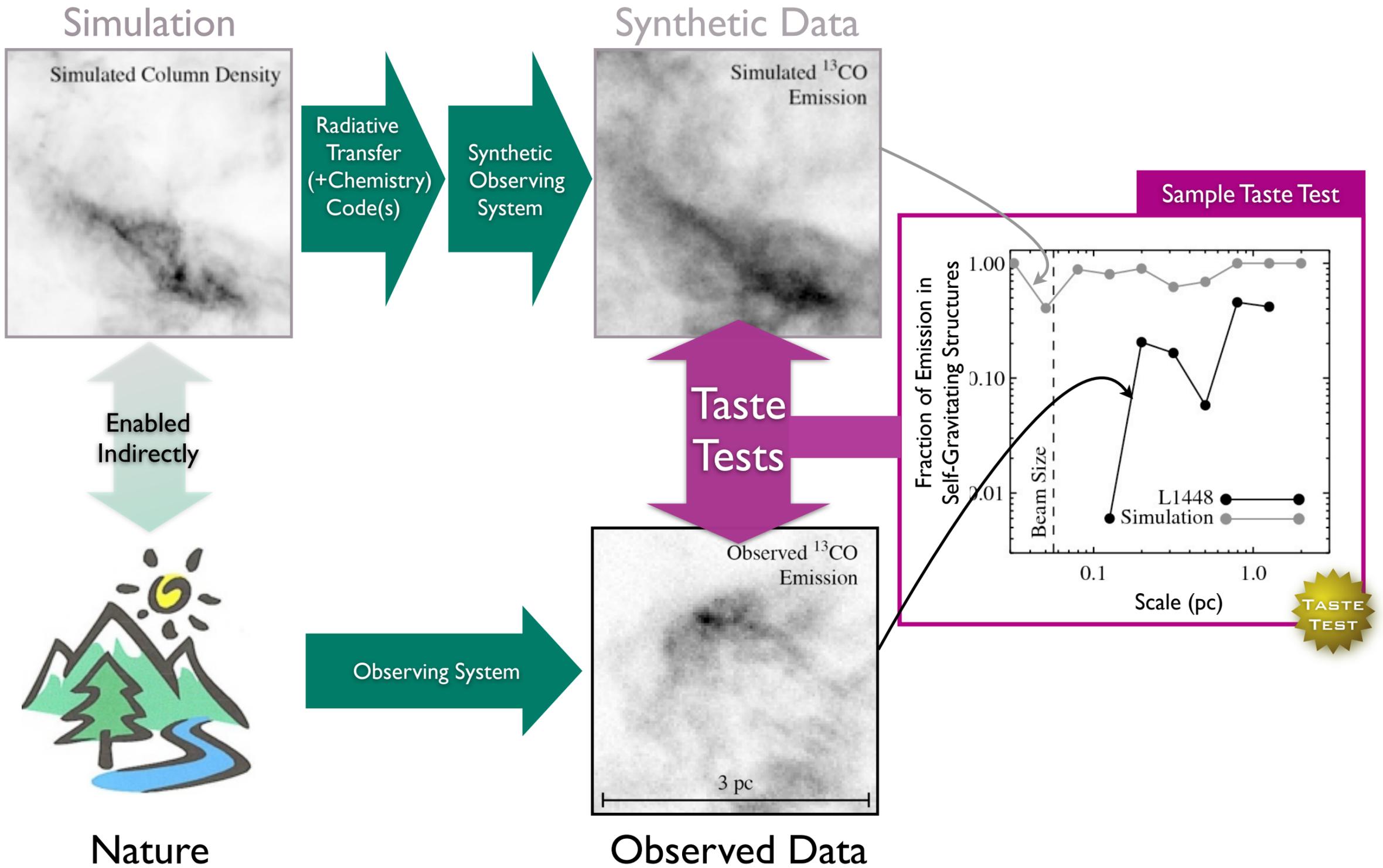
Real



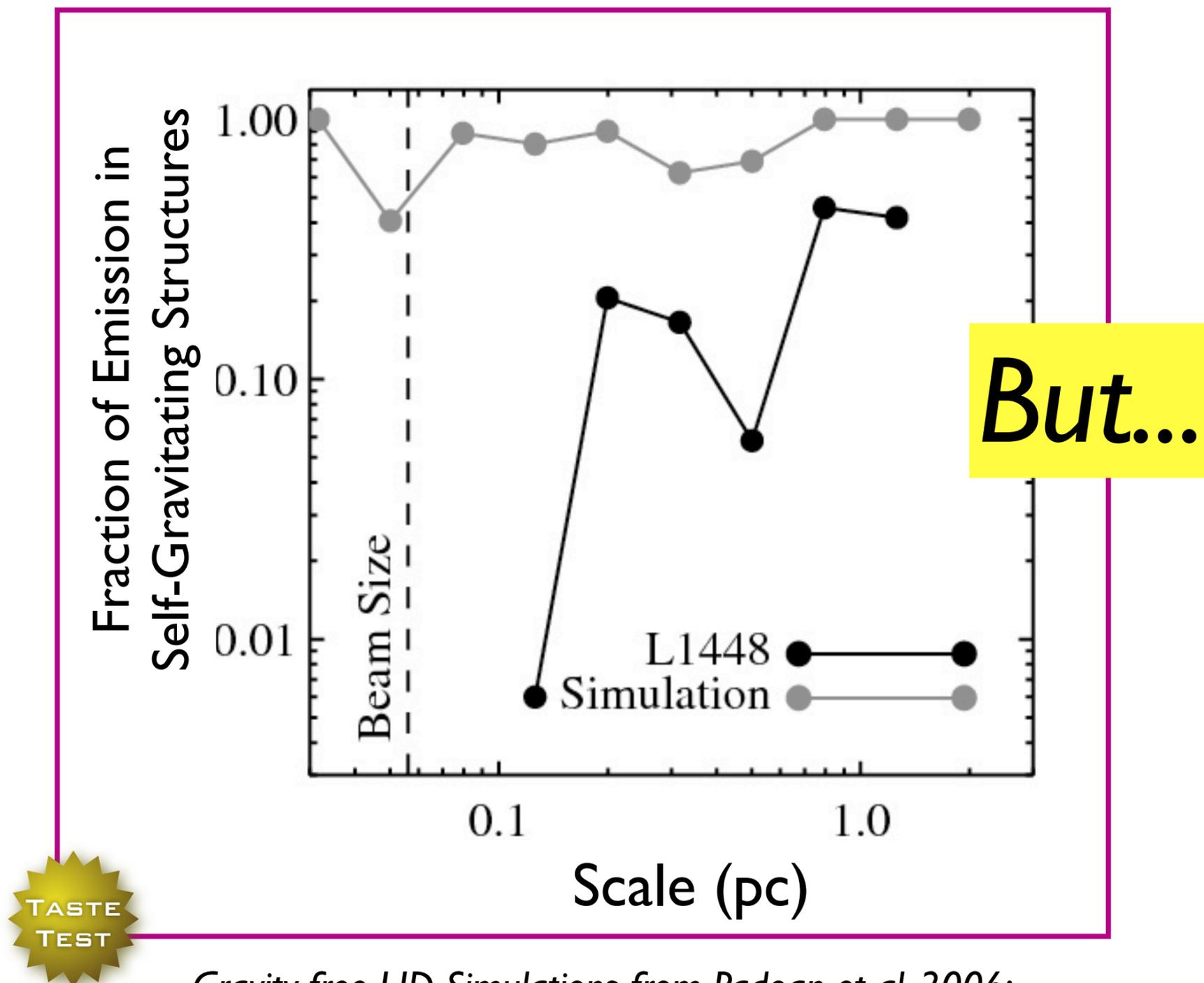
Simulated



# The Taste-Testing Process



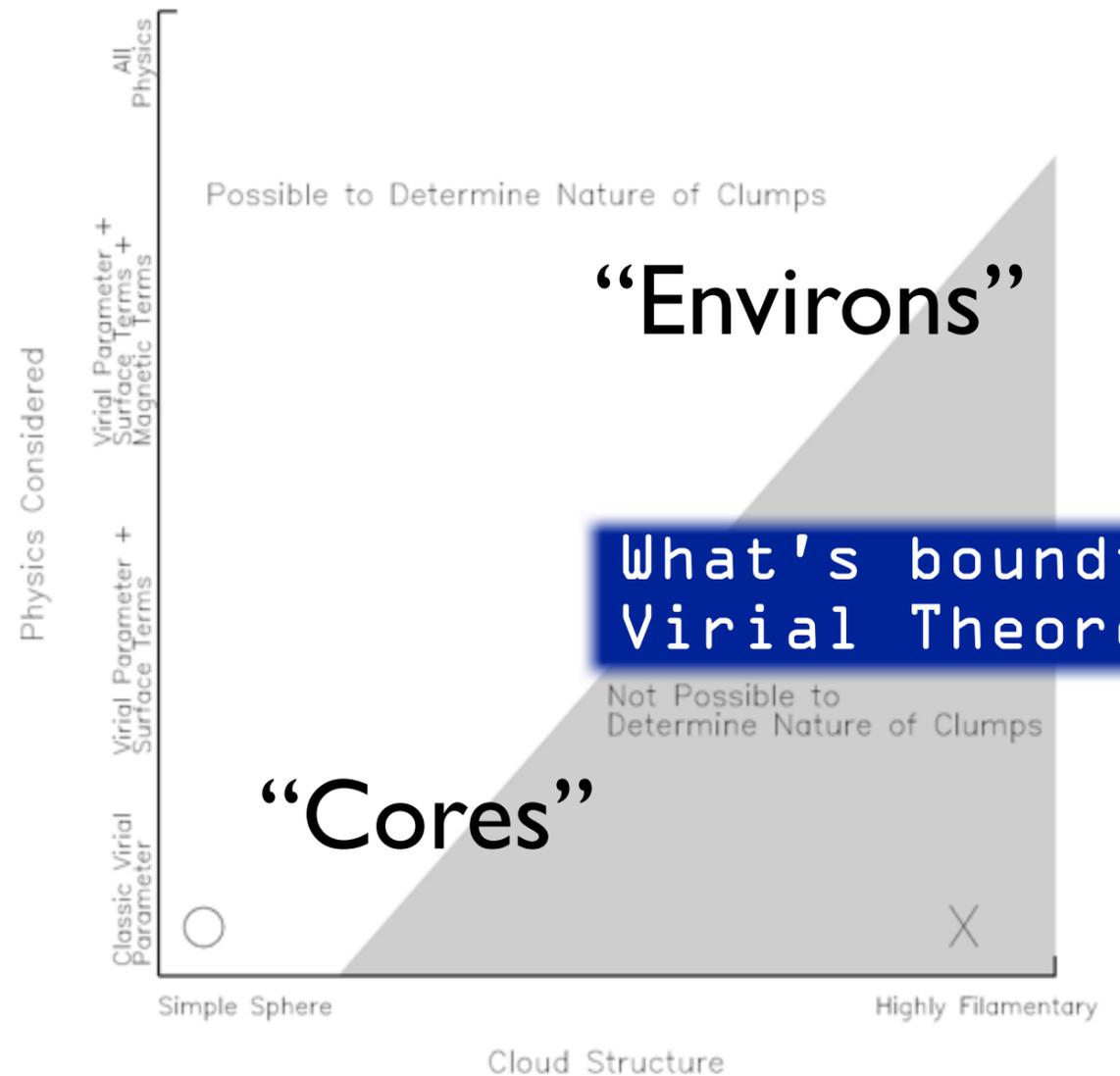
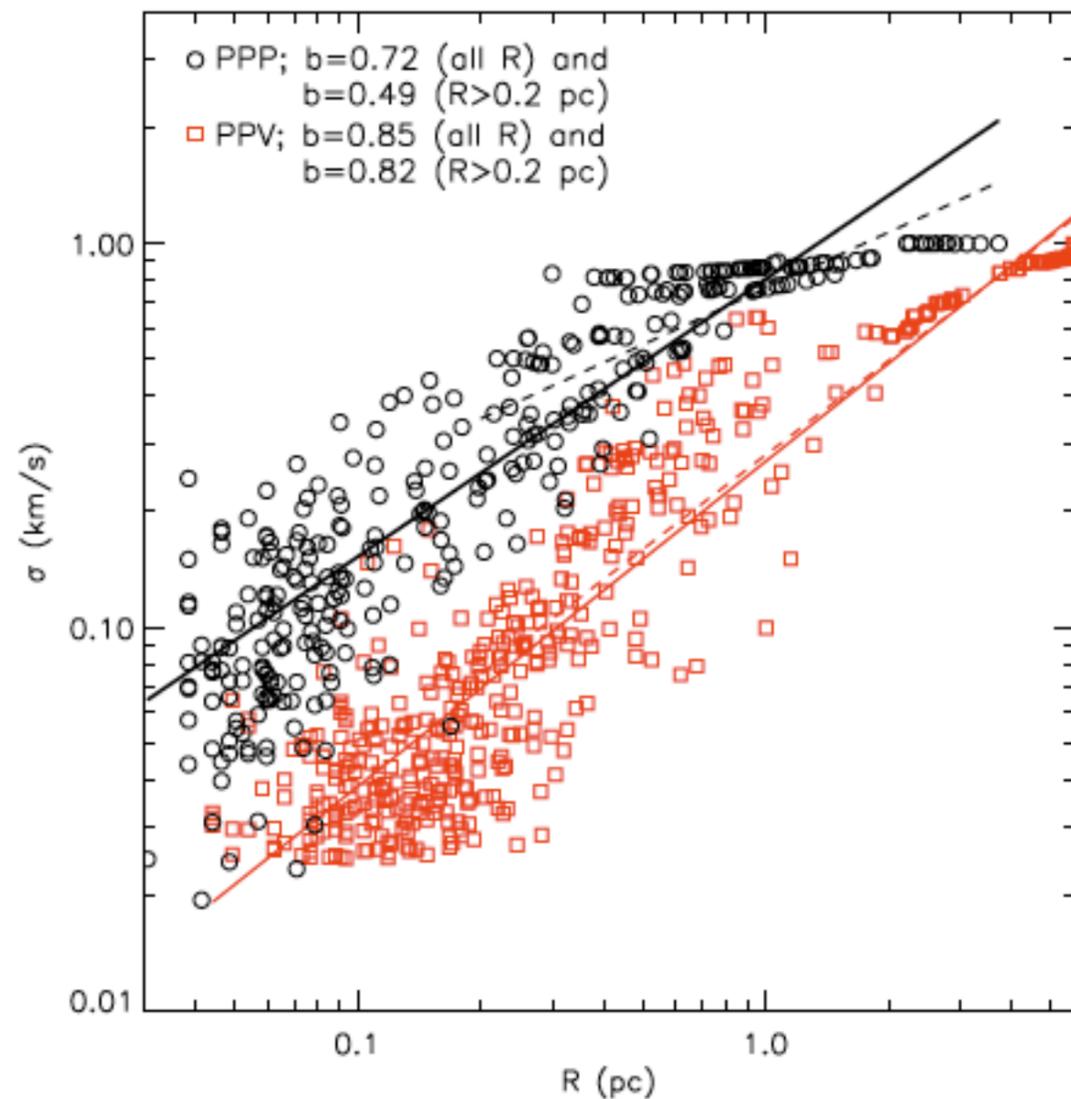
# Taste-Testing “Gravity”



*Gravity-free HD Simulations from Padoan et al. 2006;  
L1448 analysis from Rosolowsky et al. 2008  
both lines derived from  $^{13}\text{CO}$  “observations”*

TASTE  
TEST

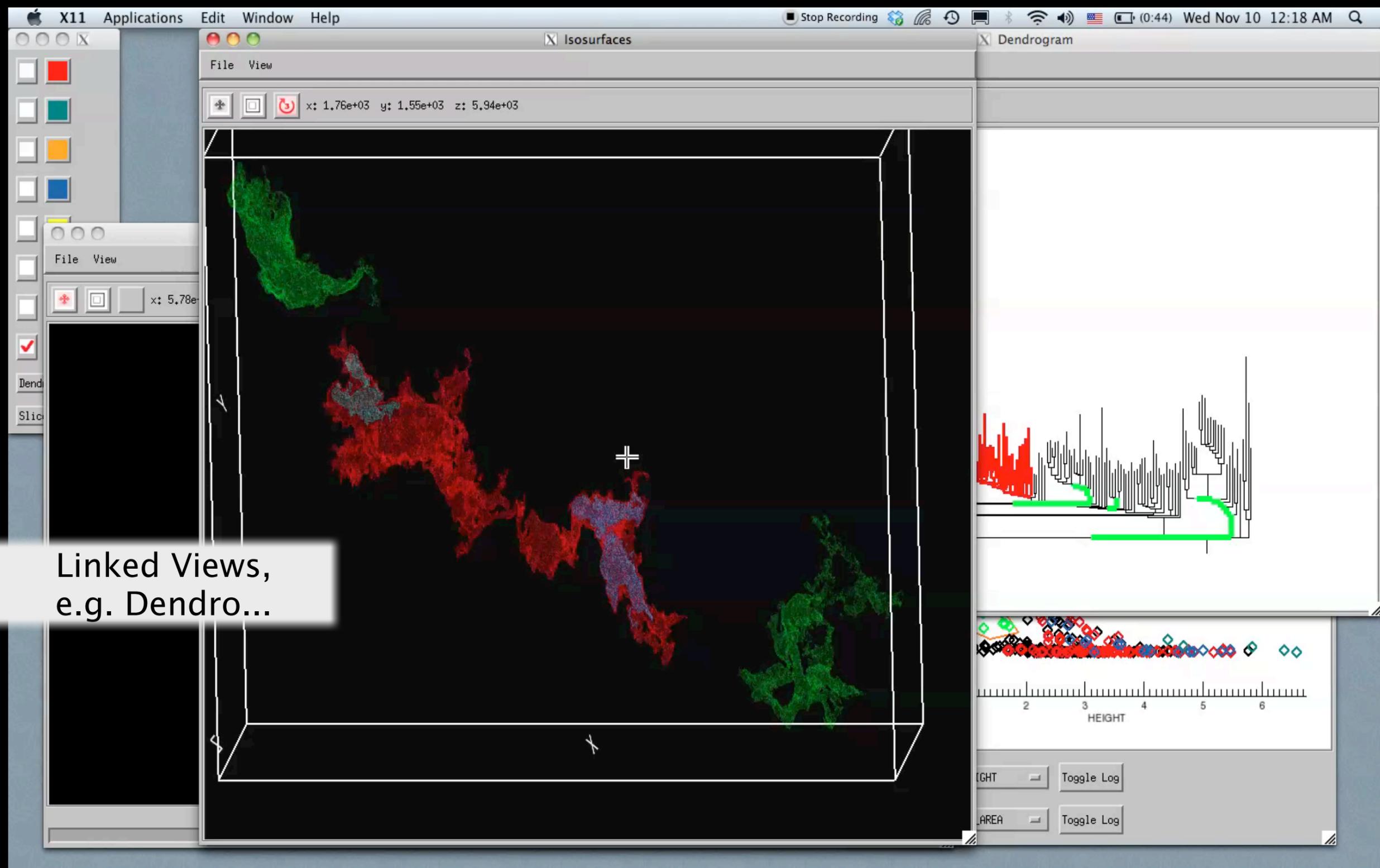
# **But...** Caveats/Worries about $p$ - $p$ - $v$ (bijection) ... and the virial parameter



from **Shetty**, Collins, Kauffmann, Goodman, Rosolowsky & M. Norman 2010;

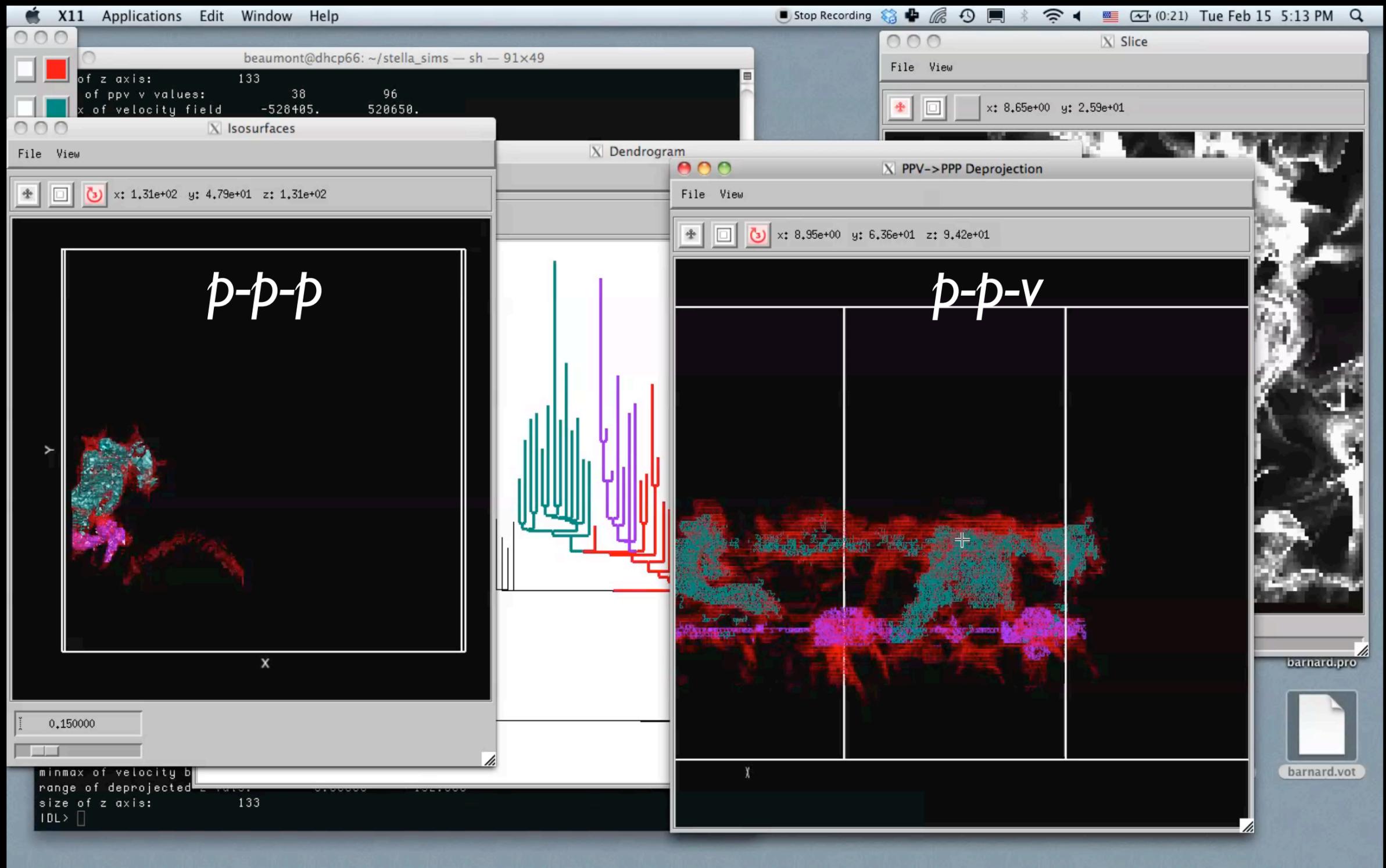
see also recent work of Dib et al., Ostriker et al., Ballesteros-Paredes et al., Myers, and Smith, Clark & Bonnell

# Linked Dendrogram Views in IDL (I)



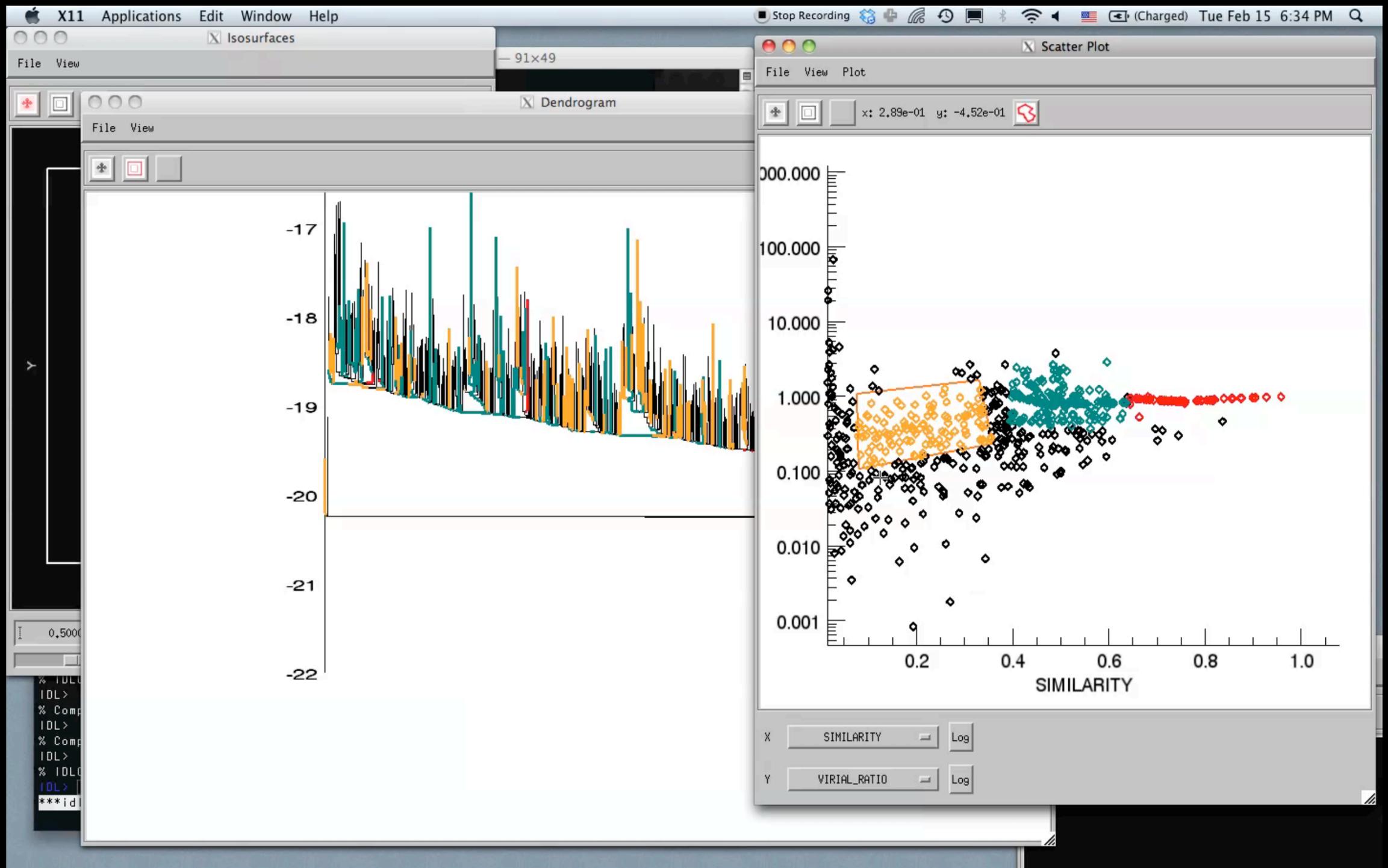
*Video & implementation: Christopher Beaumont, CfA/UHawaii;  
inspired by AstroMed work of Douglas Alan, Michelle Borkin, AG, Michael Halle, Erik Rosolowsky*

# Linked Dendrogram Views in IDL (2)



Tuesday, March 22, 2011

# Linked Dendrogram Views in IDL (3)



Tuesday, March 22, 2011

## SHOCK-GENERATED VORTICITY IN THE INTERSTELLAR MEDIUM AND THE ORIGIN OF THE STELLAR INITIAL MASS FUNCTION

N. KEVLAHAN<sup>1,2</sup> AND RALPH E. PUDRITZ<sup>2</sup>

<sup>1</sup> Department of Mathematics & Statistics, McMaster University, Hamilton, ON L8S 4K1, Canada; kevlahan@mcmaster.ca

<sup>2</sup> Origins Institute, McMaster University, Hamilton, ON L8S 4M1, Canada

Received 2009 February 19; accepted 2009 June 29; published 2009 August 7

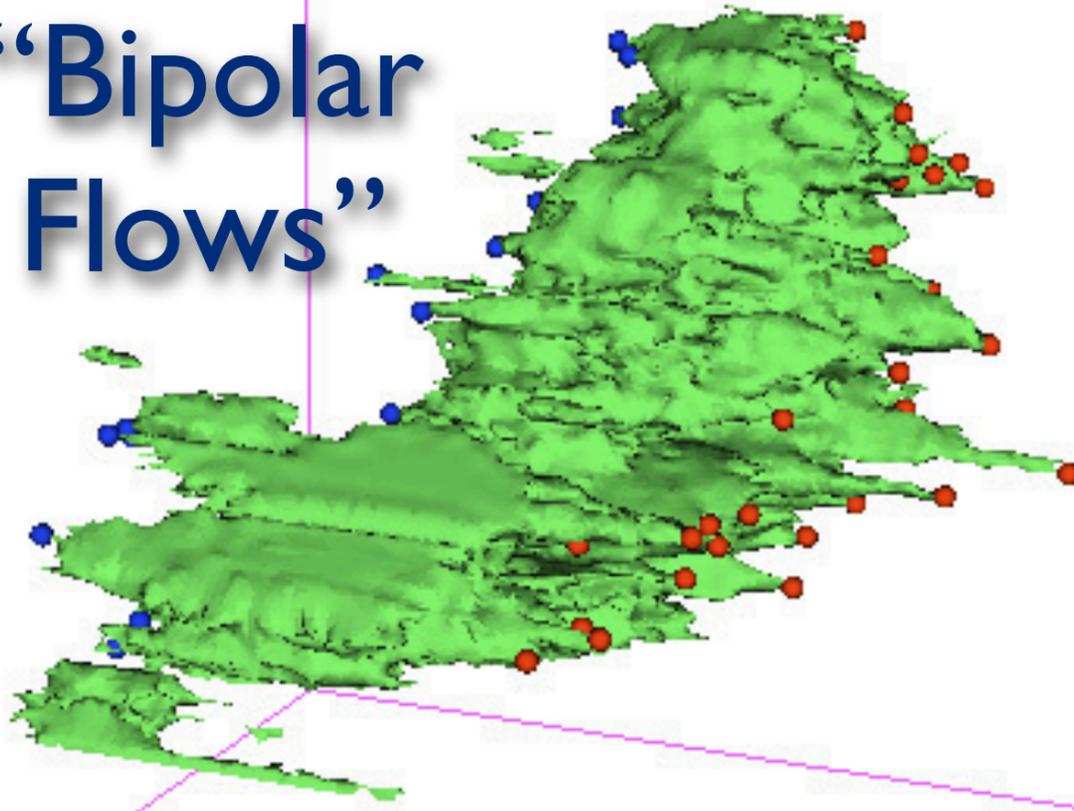
### ABSTRACT

Observations of the interstellar medium (ISM) and molecular clouds suggest these astrophysical flows are strongly turbulent. The main observational evidence for turbulence is the power-law energy spectrum for velocity fluctuations,  $E(k) \propto k^\alpha$ , with  $\alpha \in [-1.5, -2.6]$ . The Kolmogorov scaling exponent,  $\alpha = -5/3$ , is typical. At the same time, the observed probability distribution function (PDF) of gas densities in both the ISM as well as in molecular clouds is a log-normal distribution, which is similar to the initial mass function (IMF) that describes the distribution of stellar masses. In this paper we examine the density and velocity structure of interstellar gas traversed by curved shock waves in the kinematic limit. We demonstrate mathematically that just a few passages of curved shock waves generically produces a log-normal density PDF. This explains the ubiquity of the log-normal PDF in many different numerical simulations. We also show that subsequent interaction with a spherical blast wave generates a power-law density distribution at high densities, qualitatively similar to the Salpeter power law for the IMF. Finally, we show that a focused shock produces a *downstream* flow with energy spectrum exponent  $\alpha = -2$ . Subsequent shock passages reduce this slope, achieving  $\alpha \approx -5/3$  after a few passages. We argue that subsequent dissipation of energy piled up at the small scales will act to maintain the spectrum very near to the Kolmogorov value despite the action of further shocks that would tend to reduce it. These results suggest that fully developed turbulence may *not* be required to explain the observed energy spectrum and density PDF. On the basis of these mathematical results, we argue that the self-similar spherical blast wave arising from expanding H II regions or stellar winds from massive stars may ultimately be responsible for creating a mass, power-law, Salpeter-like tail on an otherwise a log-normal density PDF for gas in star-forming regions. The IMF arises from the gravitational collapse of sufficiently overdense regions within this PDF. Thus the nature of the IMF—a log-normal plus power-law distribution—is shown to be a natural consequence of shock interaction and feedback from the most massive stars that form in most regions of star formation in the galaxy.

*Key words:* ISM: kinematics and dynamics – ISM: structure – shock waves – stars: formation – stars: luminosity function, mass function – turbulence

What piles up  
the ISM?

# “Bipolar Flows”



# “Shells”



THE ASTROPHYSICAL JOURNAL, 715:1170–1190, 2010 June 1  
 © 2010. The American Astronomical Society. All rights reserved. Printed in the U.S.A. doi:10.1088/0004-637X/715/2/1170

THE COMPLETE SURVEY OF OUTFLOWS IN PERSEUS

HÉCTOR G. ARCE<sup>1</sup>, MICHELLE A. BORKIN<sup>2</sup>, ALYSSA A. GOODMAN<sup>3</sup>, JAIME E. PINEDA<sup>3</sup>, AND MICHAEL W. HALLE<sup>4,5</sup>

<sup>1</sup> Department of Astronomy, Yale University, P.O. Box 208101, New Haven, CT 06520, USA; [hector.arce@yale.edu](mailto:hector.arce@yale.edu)  
<sup>2</sup> School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA  
<sup>3</sup> Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA  
<sup>4</sup> Surgical Planning Laboratory, Department of Radiology, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115, USA  
<sup>5</sup> Initiative in Innovative Computing, Harvard University, 60 Oxford Street, Cambridge, MA 02138, USA

Received 2009 October 7; accepted 2010 April 9; published 2010 May 7

ABSTRACT

We present a study on the impact of molecular outflows in the Perseus molecular cloud complex using the COMPLETE Survey large-scale  $^{12}\text{CO}(1-0)$  and  $^{13}\text{CO}(1-0)$  maps. We used three-dimensional isosurface models generated in right ascension–declination–velocity space to visualize the maps. This rendering of the molecular line data allowed for a rapid and efficient way to search for molecular outflows over a large ( $\sim 16 \text{ deg}^2$ ) area. Our outflow-searching technique detected previously known molecular outflows as well as new candidate outflows. Most of these new outflow-related high-velocity features lie in regions that have been poorly studied before. These new outflow candidates more than double the amount of outflow mass, momentum, and kinetic energy in the Perseus cloud complex. Our results indicate that outflows have significant impact on the environment immediately surrounding localized regions of active star formation, but lack the energy needed to feed the observed turbulence in the *entire* Perseus complex. This implies that other energy sources, in addition to protostellar outflows, are responsible for turbulence on a global cloud scale in Perseus. We studied the impact of outflows in six regions with active star formation within Perseus of sizes in the range of 1–4 pc. We find that outflows have enough power to maintain the turbulence in these regions and enough momentum to disperse and unbind some mass from them. We found no correlation between outflow strength and star formation efficiency (SFE) for the six different regions we studied, contrary to results of recent numerical simulations. The low fraction of gas that potentially could be ejected due to outflows suggests that additional mechanisms other than cloud dispersal by outflows are needed to explain low SFEs in clusters.

**Key words:** ISM: clouds – ISM: individual objects (Perseus) – ISM: jets and outflows – ISM: kinematics and dynamics – stars: formation – turbulence

*Online-only material:* color figures

Arce  
 et al.  
 2010,  
 2011

COMPLETE Shells in Perseus

ABSTRACT

We present a study on the shells in the Perseus molecular cloud using the COMPLETE survey large-scale  $^{12}\text{CO}(1-0)$  and  $^{13}\text{CO}(1-0)$  maps. The shells are spread throughout most of the Perseus cloud and have circular or arc-like morphologies with a range in radius of about 0.2 to 3 pc. Most of the CO shells are coincident with near-infrared nebulosity of similar shape and have a candidate powering source near the center. We suggest they are formed by the interaction of spherical or very wide-angle winds powered by young stars inside or near the Perseus molecular cloud complex—a cloud that is commonly considered a low-mass star forming region. It is clear that two of the twelve shells are powered by high-mass stars near the cloud, while the others appear to be powered by low or intermediate-mass stars. We estimate the mass loss rate of the observed shells, which are clearly impacted by the presence of the shells. The age of the pre-main sequence stars in the Perseus cloud is estimated to be about 1 Myr. Our mass loss rate. Our estimates indicate that the energy input from both collimated protostellar outflows and powerful spherical winds from young stars is sufficient to maintain the turbulence in the molecular cloud. Most of the shells had not been detected before, most likely as maps of the region lacked the coverage and resolution needed to distinguish the shells. Large scale molecular line and IR continuum maps of a sample of other clouds will help investigate the frequency of powerful shells from low-mass stars and the impact from stellar winds from nearby massive stars on low-mass star forming regions.

**Subject headings:** star: formation — ISM: jets and outflows — ISM: clouds — ISM: individual (Perseus) — ISM: kinematics and dynamics — turbulence

What riles up  
 the ISM?

Arce, Beaumont, Borkin, Pineda, Goodman



# THE MILKY WAY PROJECT

FOLLOW US ON TWITTER   
VISIT THE BLOG   
MILKY WAY TALK 

HOME TAKE PART ABOUT TUTORIAL LOG IN GALACTOMETER™



## WELCOME

The Milky Way Project aims to sort and measure our galaxy, the Milky Way. Initially we're asking you to help us find and draw bubbles in beautiful infrared data from the Spitzer Space Telescope.

Understanding the cold, dusty material that we see in these images, helps scientists to learn how stars form and how our galaxy changes and evolves with time.

[Click here](#) to see the full tutorial or browse the site to find out more about the science behind the Milky Way Project.

“Shells”

Machine Learning

What riles up the ISM?



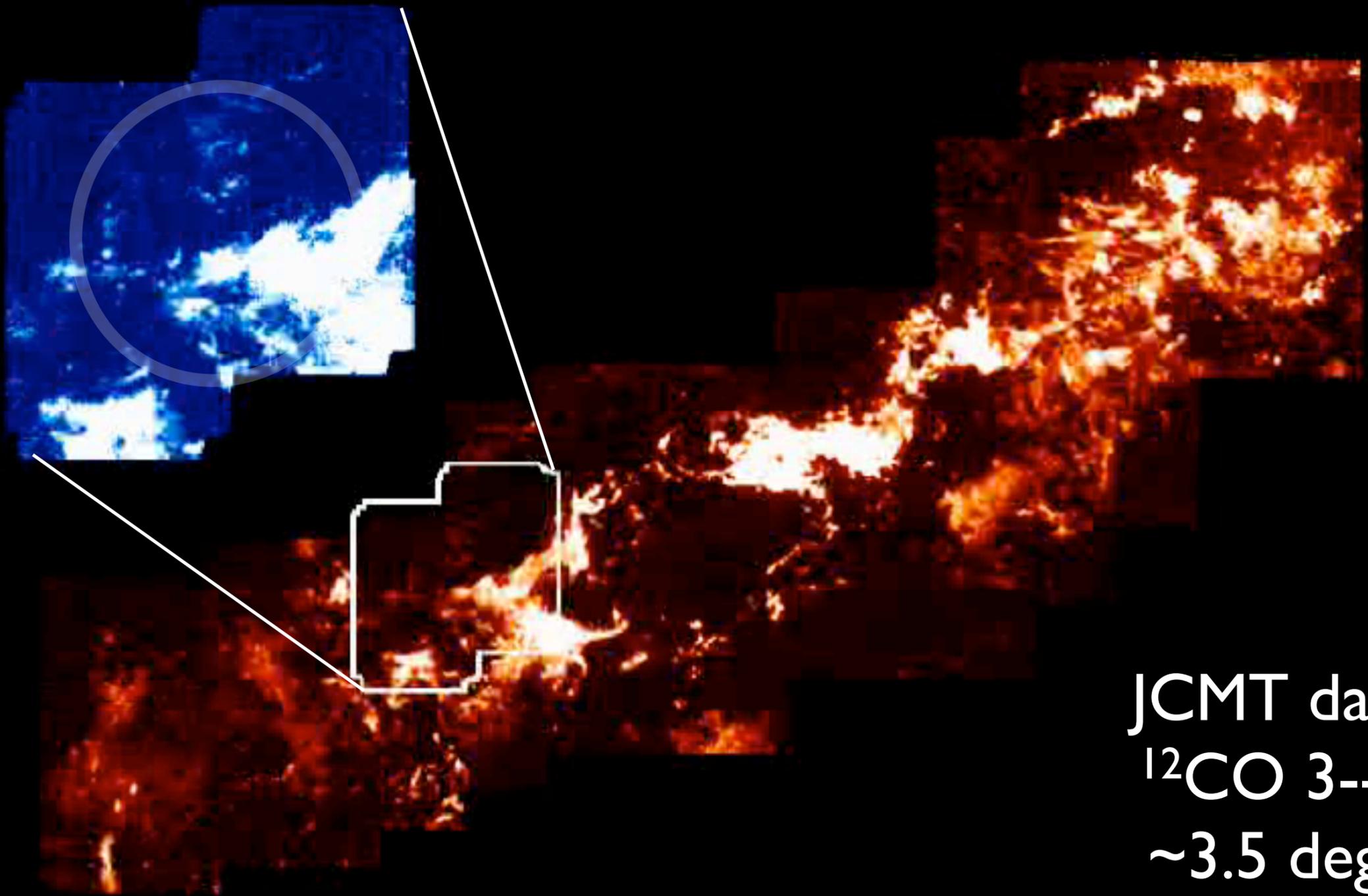
YOU CAN NOW SEE HOW CLOSE WE ARE TO 1,000,000  
194,943 IMAGES SERVED · 252,562 BUBBLES DRAWN ·

PROJECT.ORG/G...  12 DAYS AGO  
DIATE GALAXIES · 597,054 OTHER OBJECTS

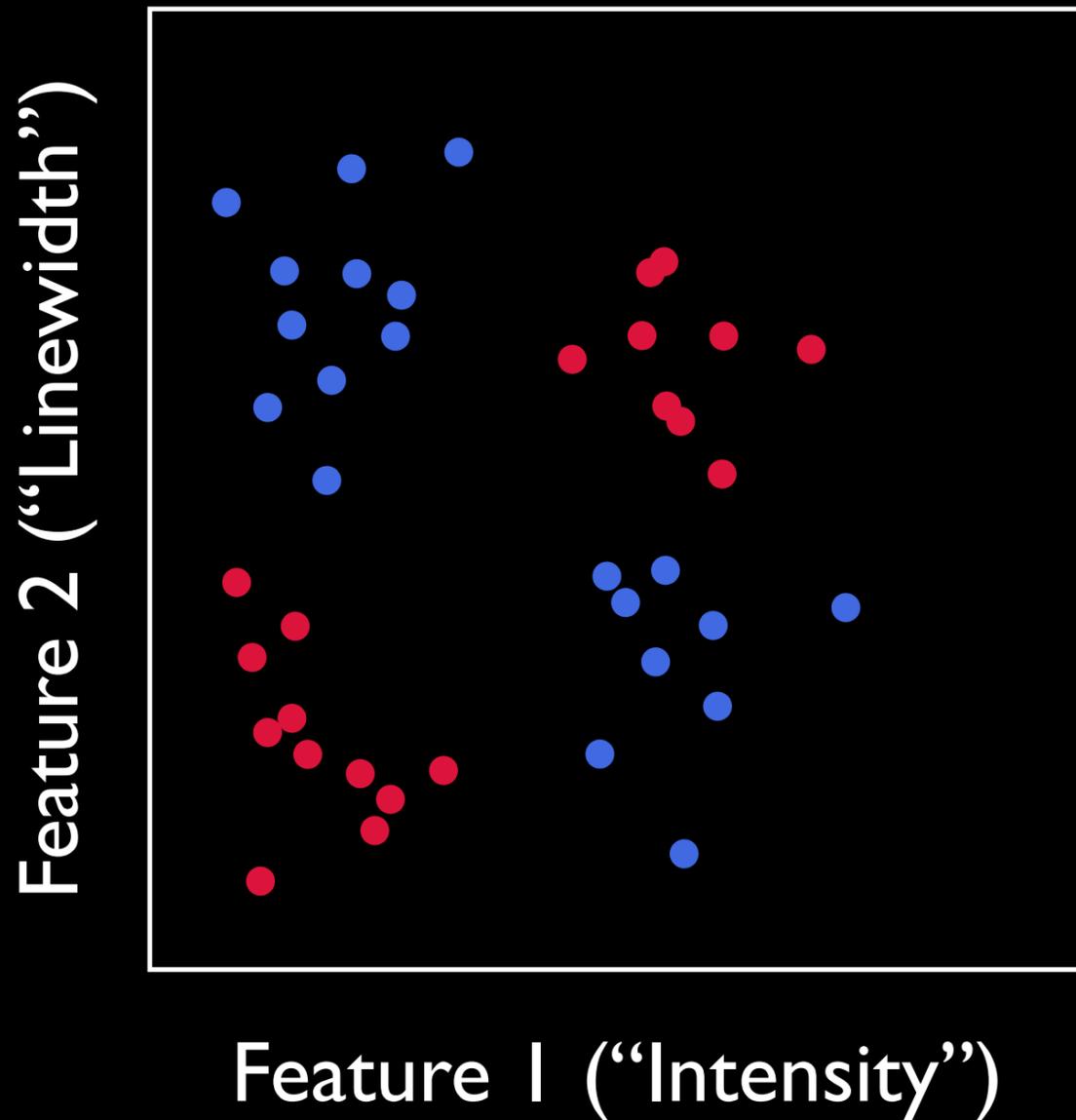
# $(p-p-v)$ Case Study (Beaumont)

“Buried” SNR GI 6.05-0.57

All of **MI7**

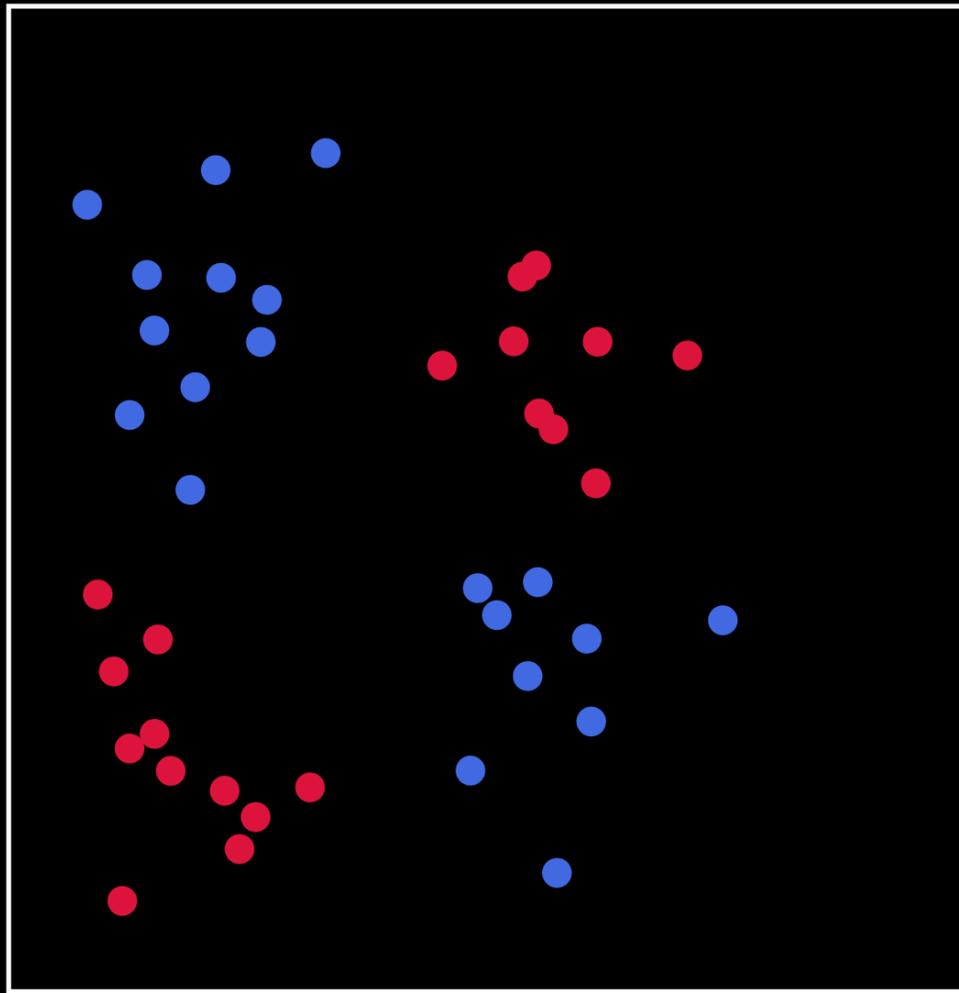


# Support Vector Machines in One Minute (SVM is a kind of “Machine Learning”)

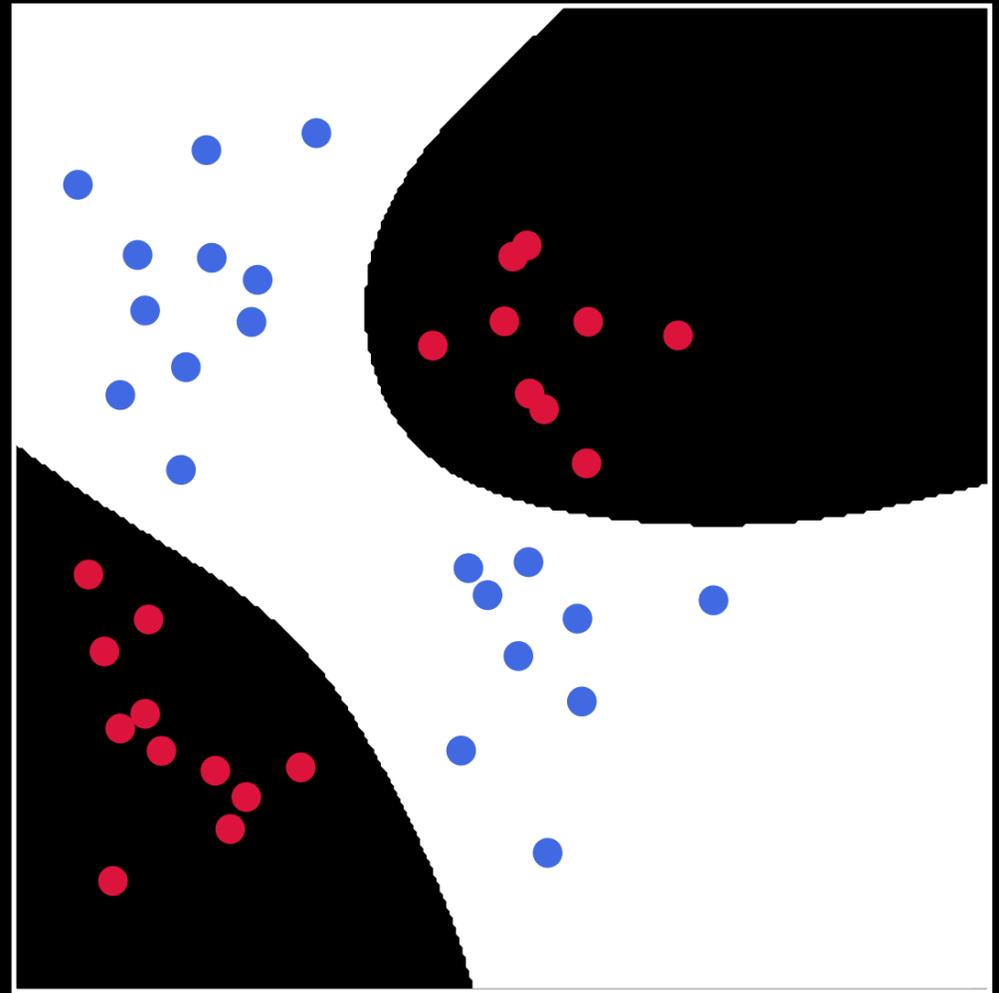


# Support Vector Machines in One Minute

Feature 2 (“Linewidth”)



Feature 1 (“Intensity”)



# Training Set

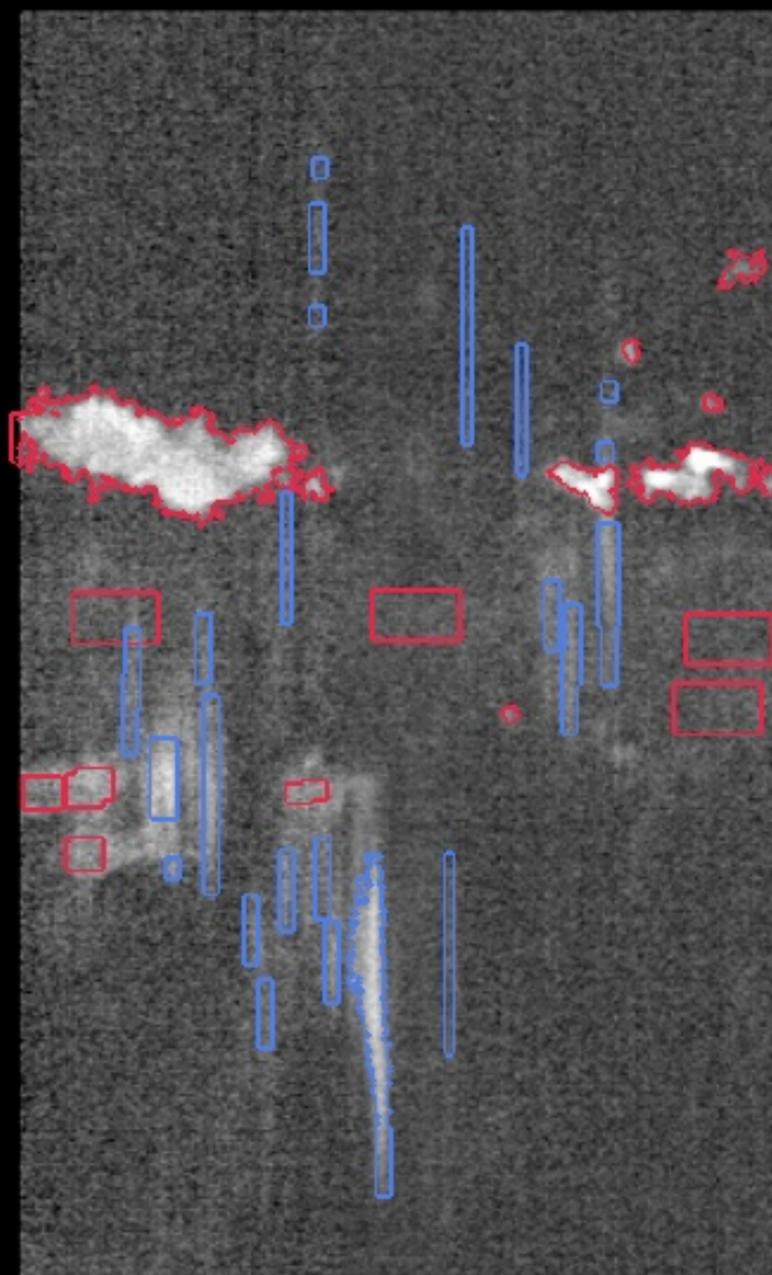
Slice

Classification

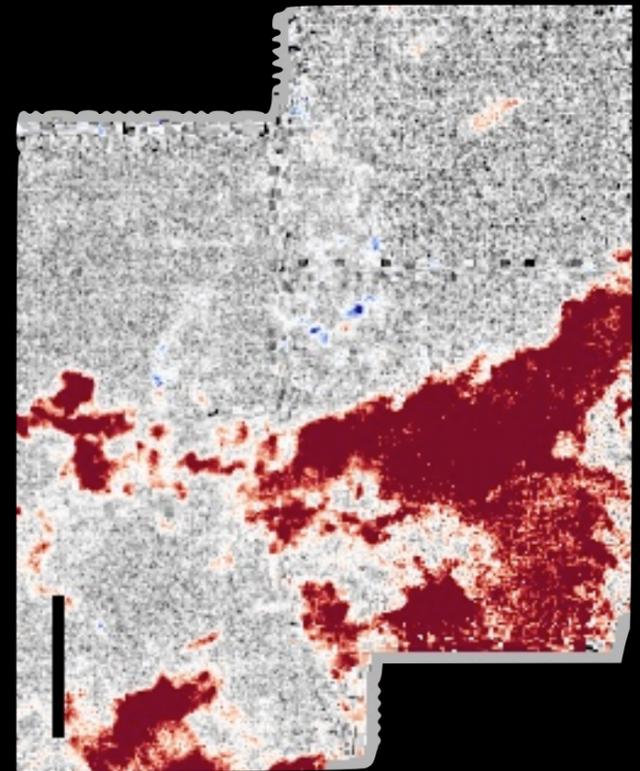
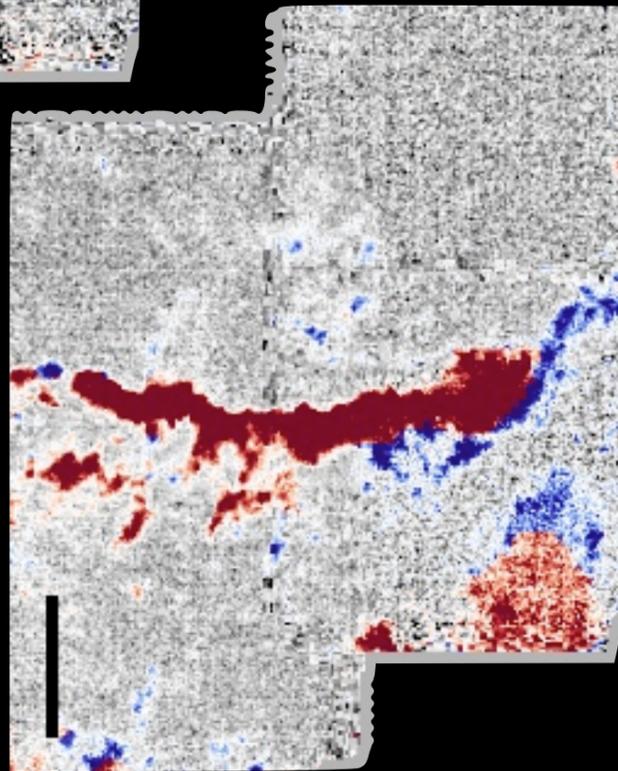
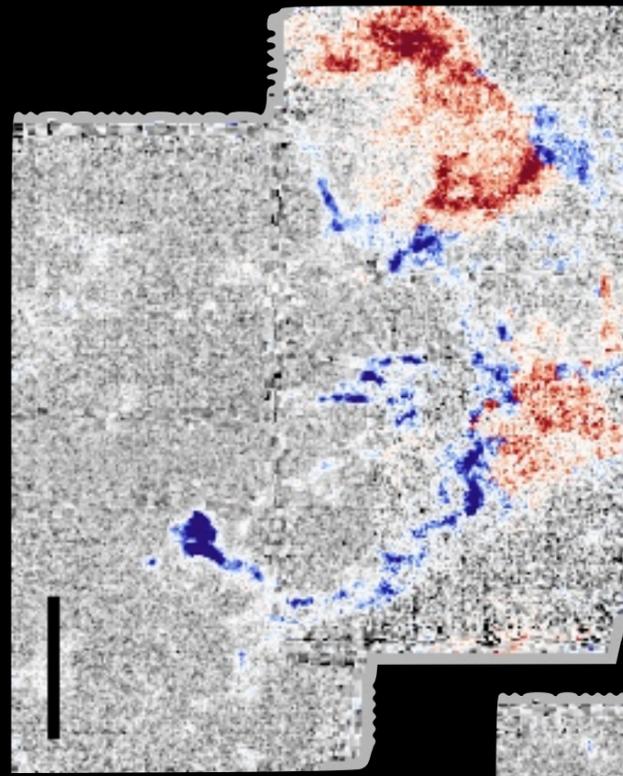
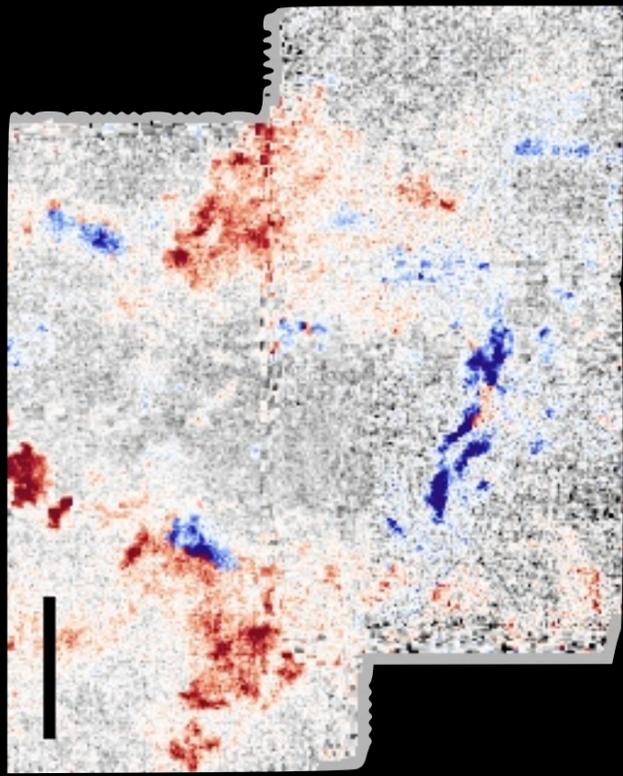


Cloud  
Supernova

(regions defined via  
rectangles and  
contours)



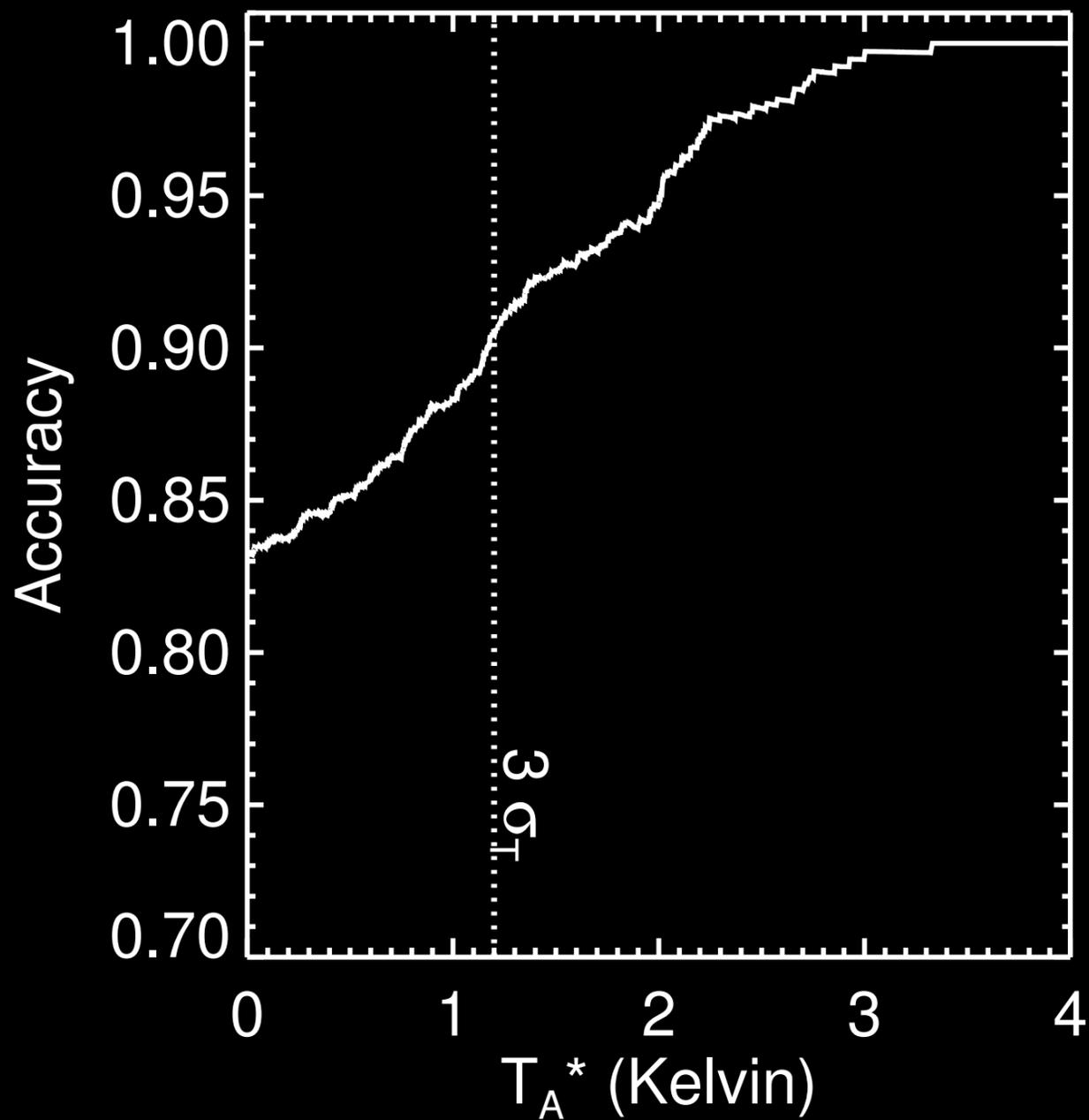
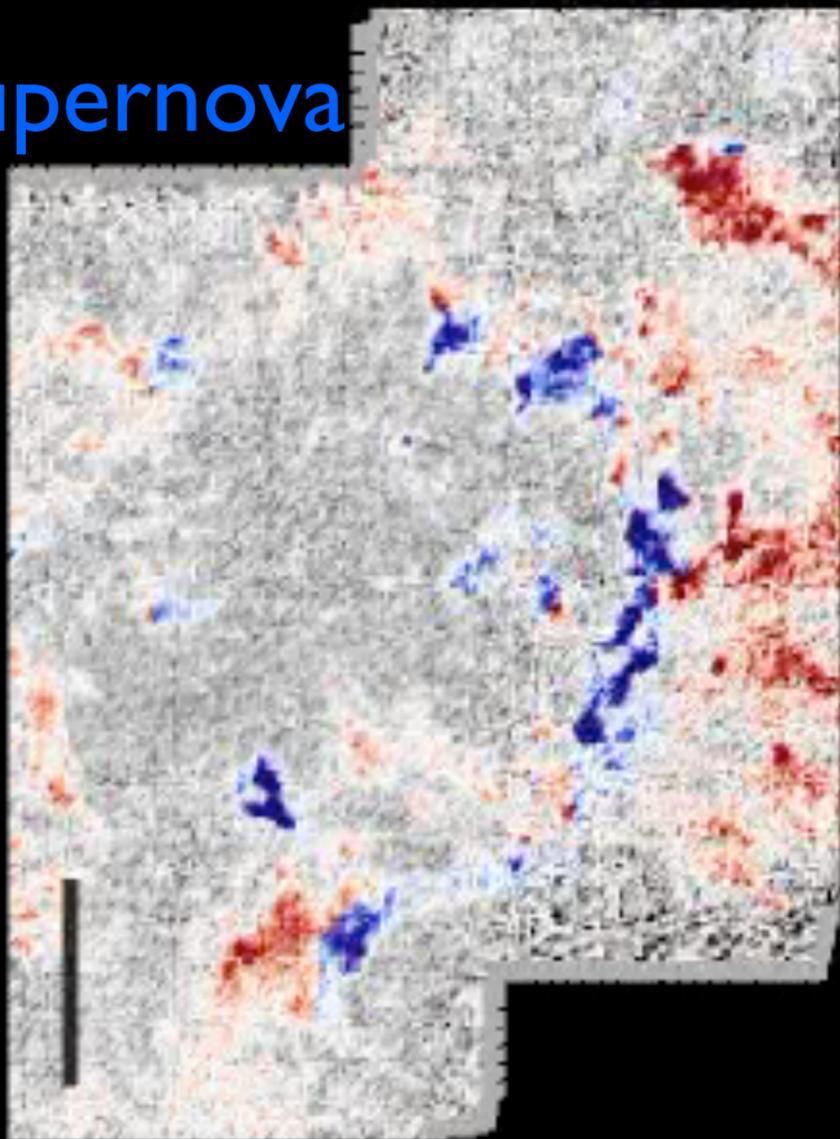
# Results



Cloud  
Supernova

# Results

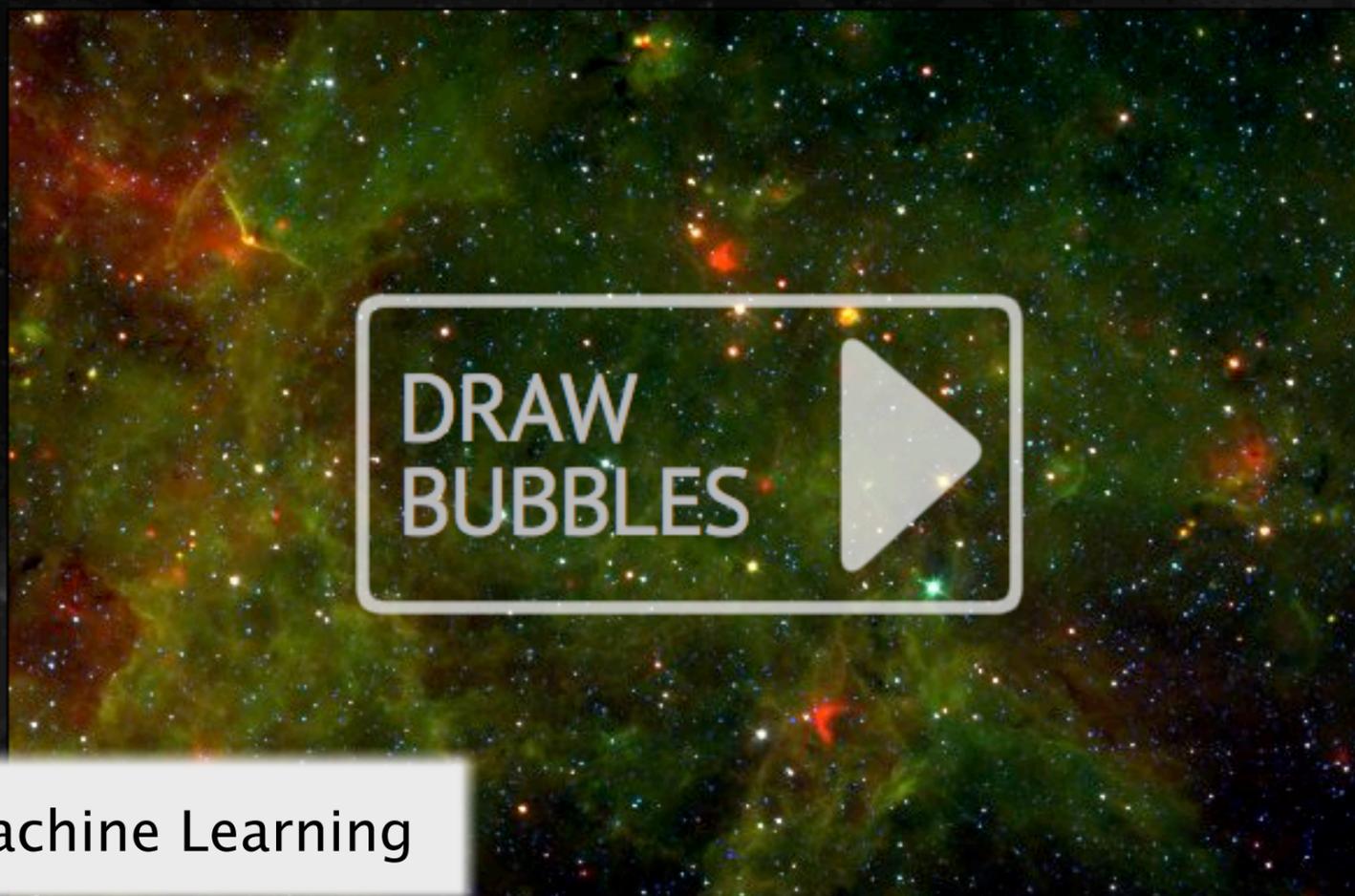
Cloud  
Supernova



# THE MILKY WAY PROJECT

FOLLOW US ON TWITTER   
VISIT THE BLOG   
MILKY WAY TALK 

HOME TAKE PART ABOUT TUTORIAL LOG IN GALACTOMETER™



## WELCOME

The Milky Way Project aims to sort and measure our galaxy, the Milky Way. Initially we're asking you to help us find and draw bubbles in beautiful infrared data from the Spitzer Space Telescope.

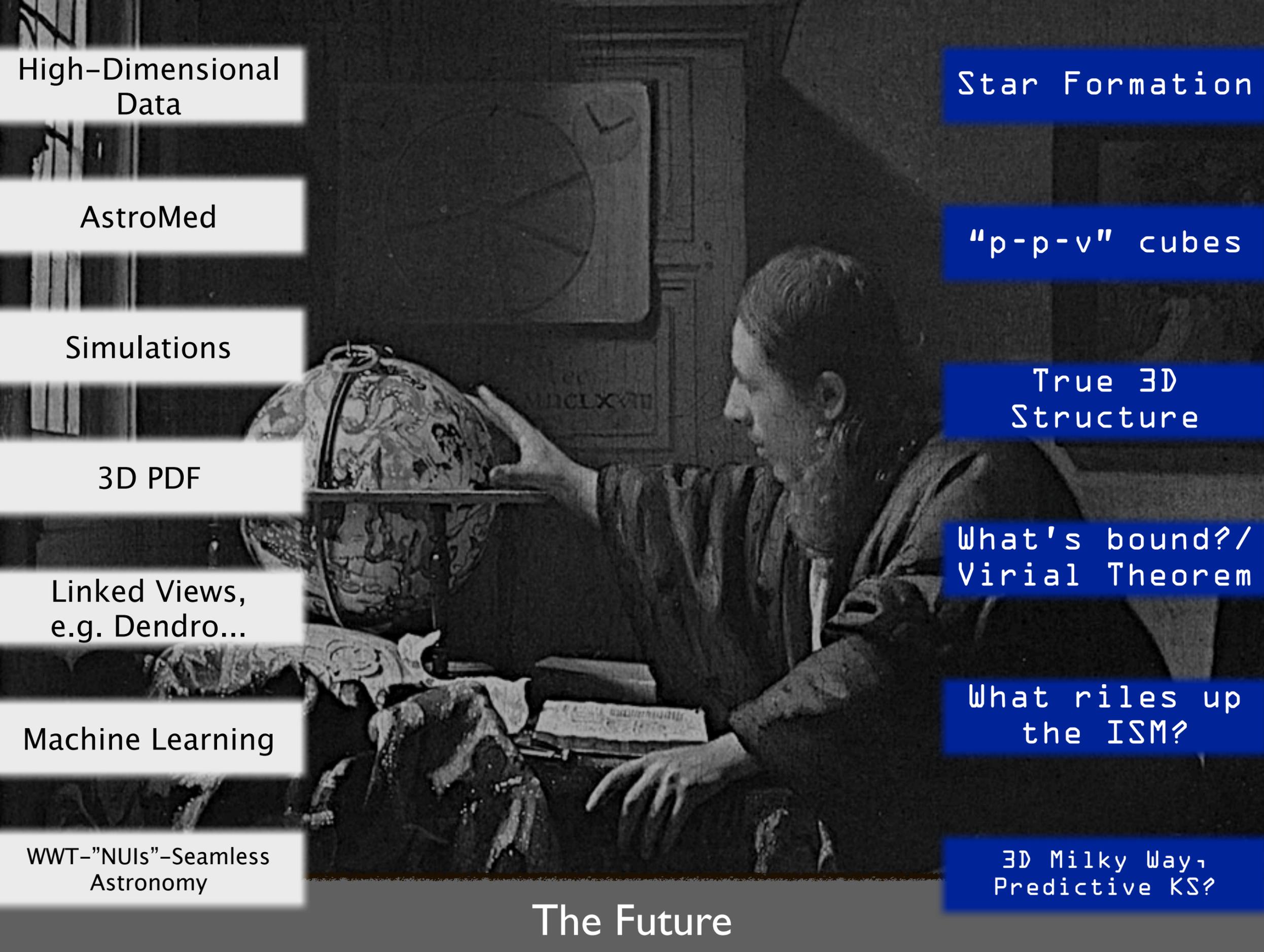
Understanding the cold, dusty material that we see in these images, helps scientists to learn how stars form and how our galaxy changes and evolves with time.

[Click here](#) to see the full tutorial or browse the site to find out more about the science behind the Milky Way Project.

Machine Learning

What rules up  
the ISM?

YOU CAN NOW SEE HOW CLOSE WE ARE TO 1,000,000 DRAWINGS AT [HTTP://WWW.MILKYWAYPROJECT.ORG/G...](http://www.milkywayproject.org/g...)  12 DAYS AGO  
194,943 IMAGES SERVED · 252,562 BUBBLES DRAWN · 24,234 POSSIBLE STAR CLUSTERS · 8,978 CANDIATE GALAXIES · 597,054 OTHER OBJECTS  
© COPYRIGHT 2010 ZOO NIVERSE



High-Dimensional  
Data

AstroMed

Simulations

3D PDF

Linked Views,  
e.g. Dendro...

Machine Learning

WWT—"NUIs"—Seamless  
Astronomy

Star Formation

"p-p-v" cubes

True 3D  
Structure

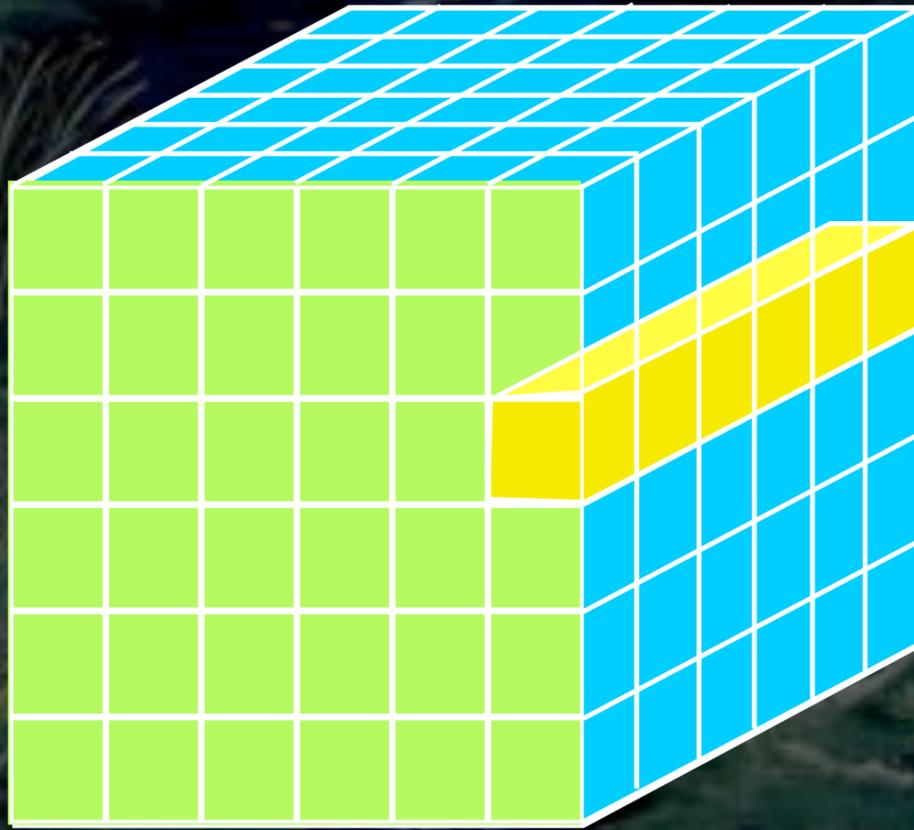
What's bound?/  
Virial Theorem

What riles up  
the ISM?

3D Milky Way,  
Predictive KS?

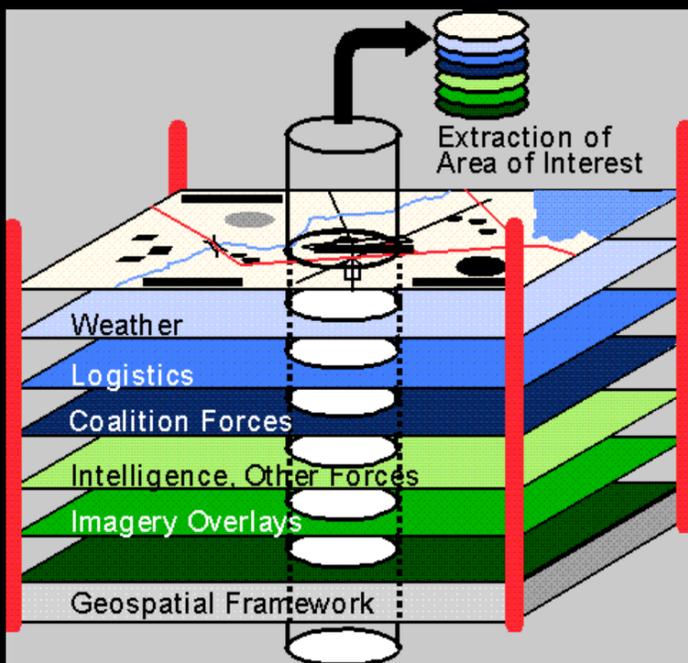
# The Future

# The dream scenario...



WWT—"NUIs"—Seamless  
Astronomy

Tuesday, March 22, 2011



# COMPLETE

## COMPLETE Data Available

Center on Perseus Center on Ophiuchus Center on Serpens

### Full-Cloud Data (Phase I, All Data Available)

Dataset	Show	Perseus	Ophiuchus	Serpens	Link
GBT: HI Data Cube	<input checked="" type="checkbox"/>	✓	✓	∅	<a href="#">Data</a>
IRAS: Av/Temp Maps	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>
FCRAO: 12CO	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>
FCRAO: 13CO	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>
JCMT: 850 microns	<input checked="" type="checkbox"/>	✓	✓	∅	<a href="#">Data</a>
Spitzer c2d: IRAC 1,3 (3.6,5.8 μm)	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>
Spitzer c2d: IRAC 2,4 (4.5,8 μm)	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>
CSO/Bolocam: 1.2-mm	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>
Spitzer MIPS: Derived Dust Map	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>

### Targeted Regions (Phase II, Some Data Not Yet Available)

CTIO/Calar Alto: NIR (J,H,Ks)	<input checked="" type="checkbox"/>	✓	✓	∅	<a href="#">Data</a>
IRAM 30-m: N2H+ and C18O	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>
IRAM 30-m: 1.1-mm continuum	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>
Megacam/MMT: r,i,z images	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>

### Catalogs & Pointed Surveys

NH3 Pointed Survey	<input checked="" type="checkbox"/>	✓	∅	∅	<a href="#">Data</a>
YSO Candidate list (c2d)	<input checked="" type="checkbox"/>	✓	✓	✓	<a href="#">Data</a>

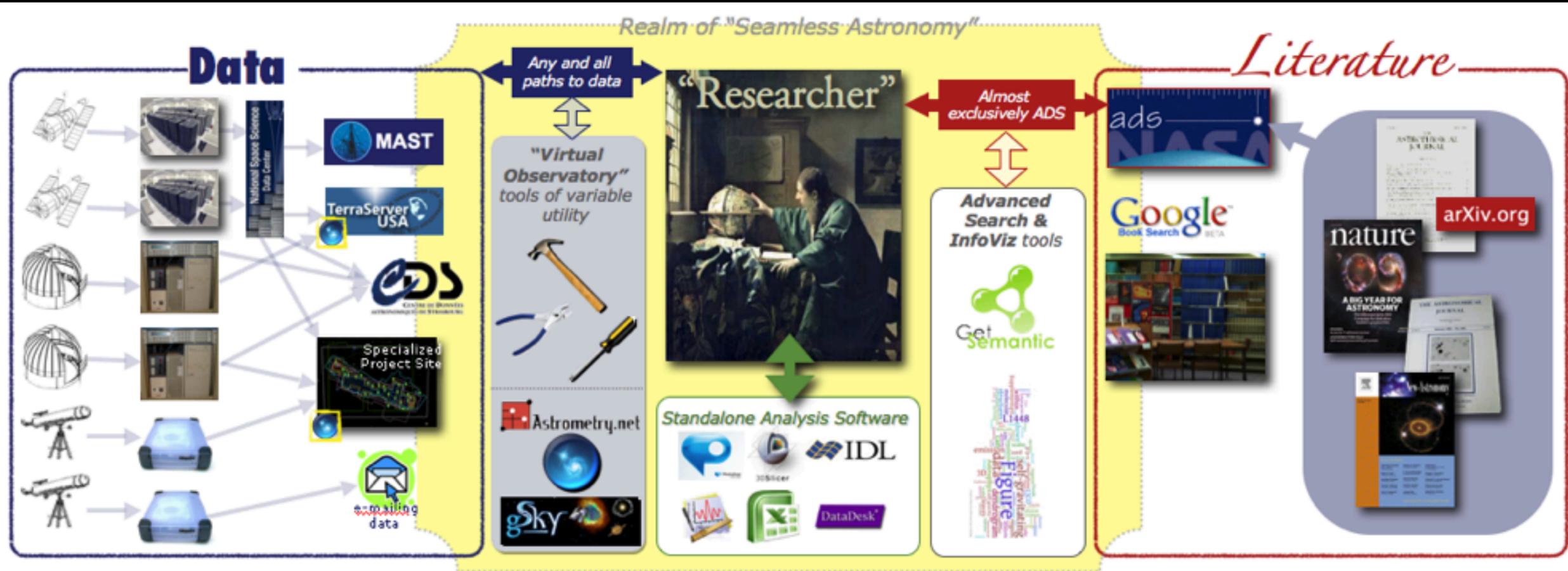


WWT—"NUIs"—Seamless Astronomy

Microsoft Research  
WorldWide Telescope

# Seamless Astronomy

Alberto Accomazzi, Doug Burke, Alberto Conti, Carol Christian, Mercé Crosas, Raffaele D'Abrusco, Rahul Davé, Christopher Erdmann, Jonathan Fay, Jay Luker, Alyssa Goodman, Michael Kurtz, Gus Muench, Alberto Pepe, Curtis Wong



WWT—"NUIs"—Seamless Astronomy





# Microsoft® Research WorldWide Telescope

Experience WWT at [worldwidetelescope.org](http://worldwidetelescope.org)

The screenshot shows the WWT interface with a top navigation bar containing 'Explore', 'Guided Tours', 'Search', 'View', and 'Settings'. Below this is a 'Collections > All-Sky Surveys >' section with a row of eight image thumbnails: Digitized Sky Survey, VLSS: VLA Low-fre, WMAP ILC 5-Year, SFD Dust Map (Inf), IRIS: Improved Re, 2MASS: Two Micro, and Hydrogen Alpha Fu. The main view is a 3D sky with a central circular field of view showing a galaxy. A 'Finder Scope' window is open, displaying details for NGC224, including its classification as a 'Spiral Galaxy in Andromeda' and various astronomical coordinates. At the bottom, there is a 'Look At' dropdown set to 'Sky', an 'Imagery' section with 'Digitized Sky Survey' selected, and a 'Context bar' showing 'NGC221' and 'M31'. A 'Context globe' on the right shows the current field of view on a celestial sphere.

Seamlessly explore imagery from the best ground and space-based telescopes in the world

Expert led tours of the Universe

Control time to study how the night sky changes

View and compare images from across the electromagnetic spectrum

Much more than "just" the sky at night! 3D features can take you to other planets, stars & galaxies.

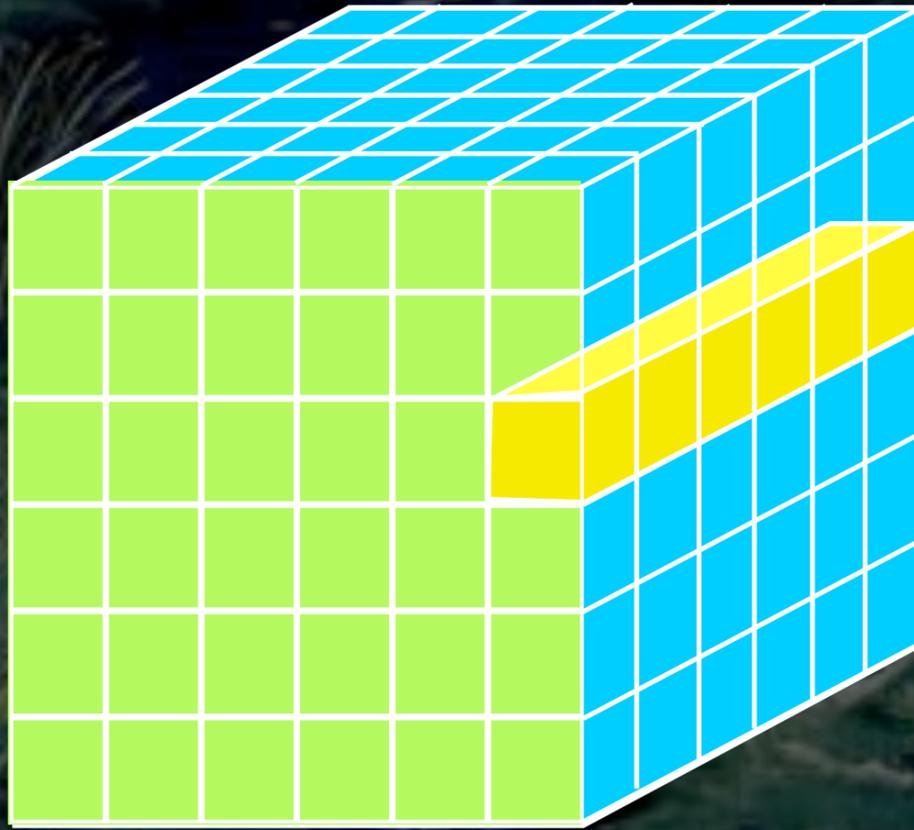
Finder Scope links to Wikipedia, publications, and data, so you can learn more

Context bar shows items of interest in current field of view

Context globe shows where you're looking.

WWT—"NUIs"—Seamless Astronomy

# The dream scenario...



WWT—"NUIs"—Seamless  
Astronomy

Tuesday, March 22, 2011

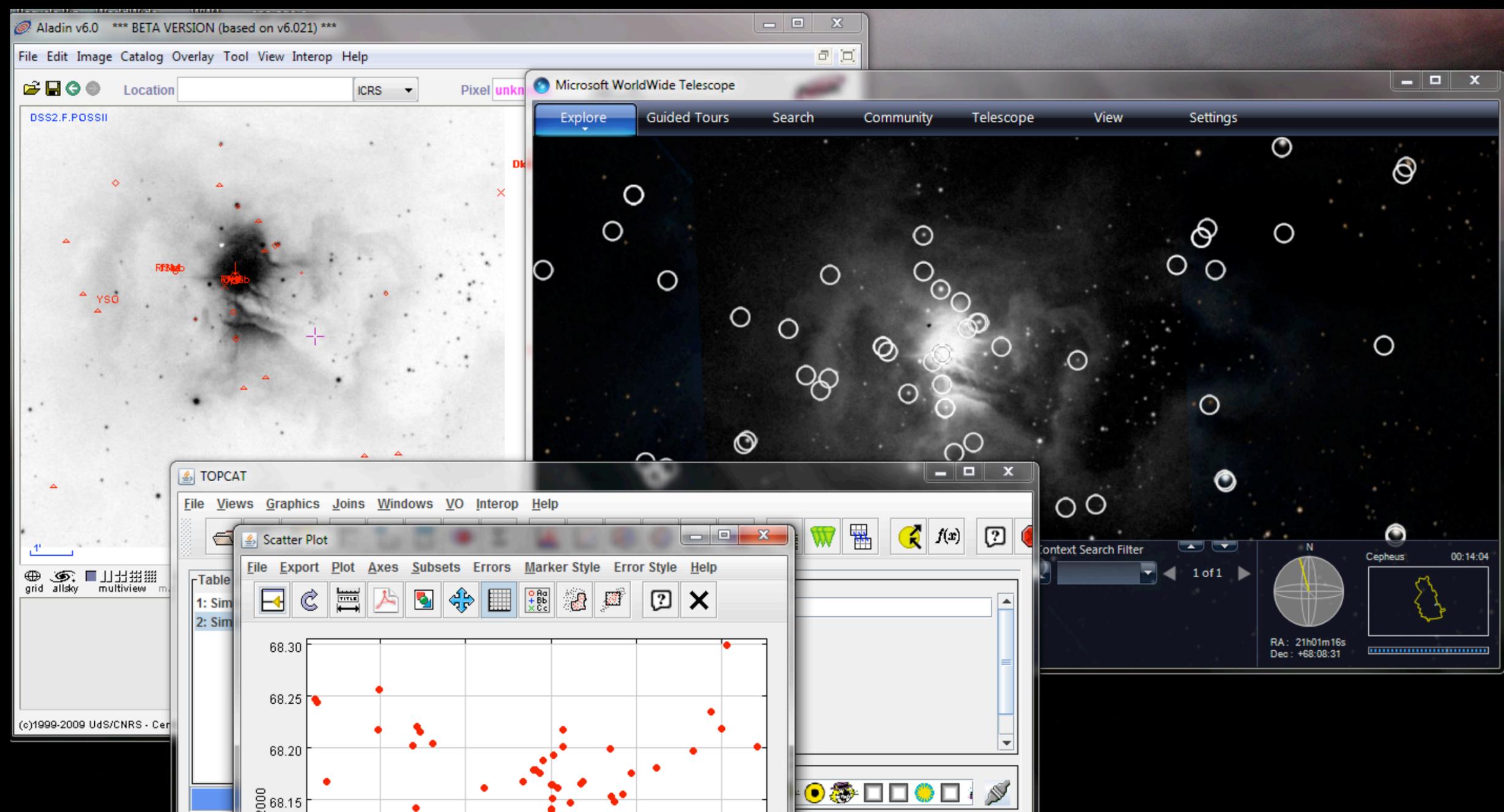
# Challenge #1: 3D Selection

Why?  
How?  
How?



visit our new "Viz-e-Lab"!

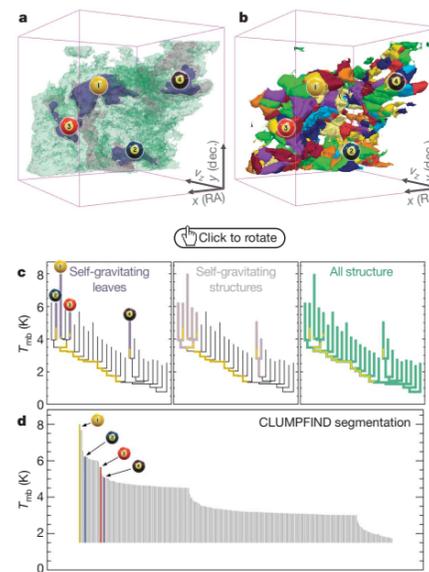
# Challenge #2: Too many windows...



Tuesday, March 22, 2011

# Challenge #3:

## What does “Publication-Quality” Graphics Mean in an Interactive 3D World?



**Figure 2 | Comparison of the 'dendrogram' and 'CLUMPFIND' feature-identification algorithms as applied to  $^{13}\text{CO}$  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  $T_{\text{mb}}$  (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 ( $-0.5 \text{ km s}^{-1}$ ) to back ( $8 \text{ km s}^{-1}$ ).

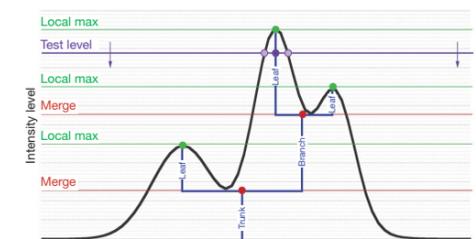
data, CLUMPFIND typically finds features on a limited range of scales, above but close to the physical resolution of the data, and its results can be overly dependent on input parameters. By tuning CLUMPFIND's two free parameters, the same molecular-line data set can be used to show either that the frequency 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' 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'<sup>10</sup>. Although well developed in other data-intensive fields<sup>11,12</sup>, 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<sup>13</sup>.

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 ( $\sigma_v$ ) and luminosity ( $L$ ). The volumes can have any shape, and in other work<sup>14</sup> 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  $M_{\text{lum}} = X_{13\text{CO}} L_{13\text{CO}}$ , where  $X_{13\text{CO}} = 8.0 \times 10^{20} \text{ cm}^{-2} \text{ K}^{-1} \text{ km}^{-1} \text{ s}$  (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,  $\alpha_{\text{obs}} = 5\sigma_v^2 R / GM_{\text{lum}}$ . In principle, extended portions of the tree (Fig. 2, yellow highlighting) where  $\alpha_{\text{obs}} < 2$  (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  $\alpha_{\text{obs}}$  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<sup>16</sup>, 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, Rosolowsky, Borkin, Foster, Halle,  
Kauffmann & Pineda, **Nature**, 2009

# Making Sense of High-Dimensional Data and Visualizations



3D Milky Way  
• Predictive KS?

*Alyssa A. Goodman*

*Harvard-Smithsonian Center for Astrophysics*

*Key Collaborators: H. Arce, C. Beaumont, M. Borkin, M. Halle, J. Kauffmann, J. Pineda, E. Rosolowsky, R. Shetty*

*Jan Vermeer. The Astronomer. (1668)*

Tuesday, March 22, 2011

The "data deluge" in science is old news. Now, it's pouring, and we need working tools to collect, sort out, understand, and keep what is falling down on us. In astronomy, the greatest insights very often come from studies where more than one "band" of data (e.g. optical, infrared, radio, X-ray) is combined. And, data sets aren't just large--they are often also high-dimensional, in that they contain information about flux as functions not just of position on the sky, but also as functions of a third dimension (e.g. frequency, velocity), and/or of time. Life science, geophysical, and geospatial data all present similar challenges.

In this talk, I will focus on examples drawn from my group's research on star formation in molecular clouds. In particular, I will show how new visualization and statistical analysis techniques relying on interactive high-dimensional views of data and on automated algorithms for "segmenting" data give new insight. "Segmentation" in imaging terms refers to extracting the meaningful structures from data, and I will show results from both dendrogram (tree-hierarchy) and machine-learning approaches. I will emphasize how the visualization of segmentation results is critical for understanding. The highlighted science results will show how we can now--for the first time--quantitatively but intuitively understand the connections between the "real" (position-position-position) space where simulations (e.g. of star formation) can be made and the "observational" (e.g. position-position-velocity) space available to earthbound astronomers. As a result of this newfound understanding, we can place important limits on the validity of virial-theorem-based calculations of the properties of gas--allowing, for example, for better estimates of which gas in star-forming regions is most likely to stay bound long enough to form stars.

Even though this abstract may sound technical to non-star-formation or non-computational researchers, my goal will be to keep the talk accessible to non-experts, so people from other fields faced with high-dimensional data and visualization challenges should feel free to join in--and to ask questions