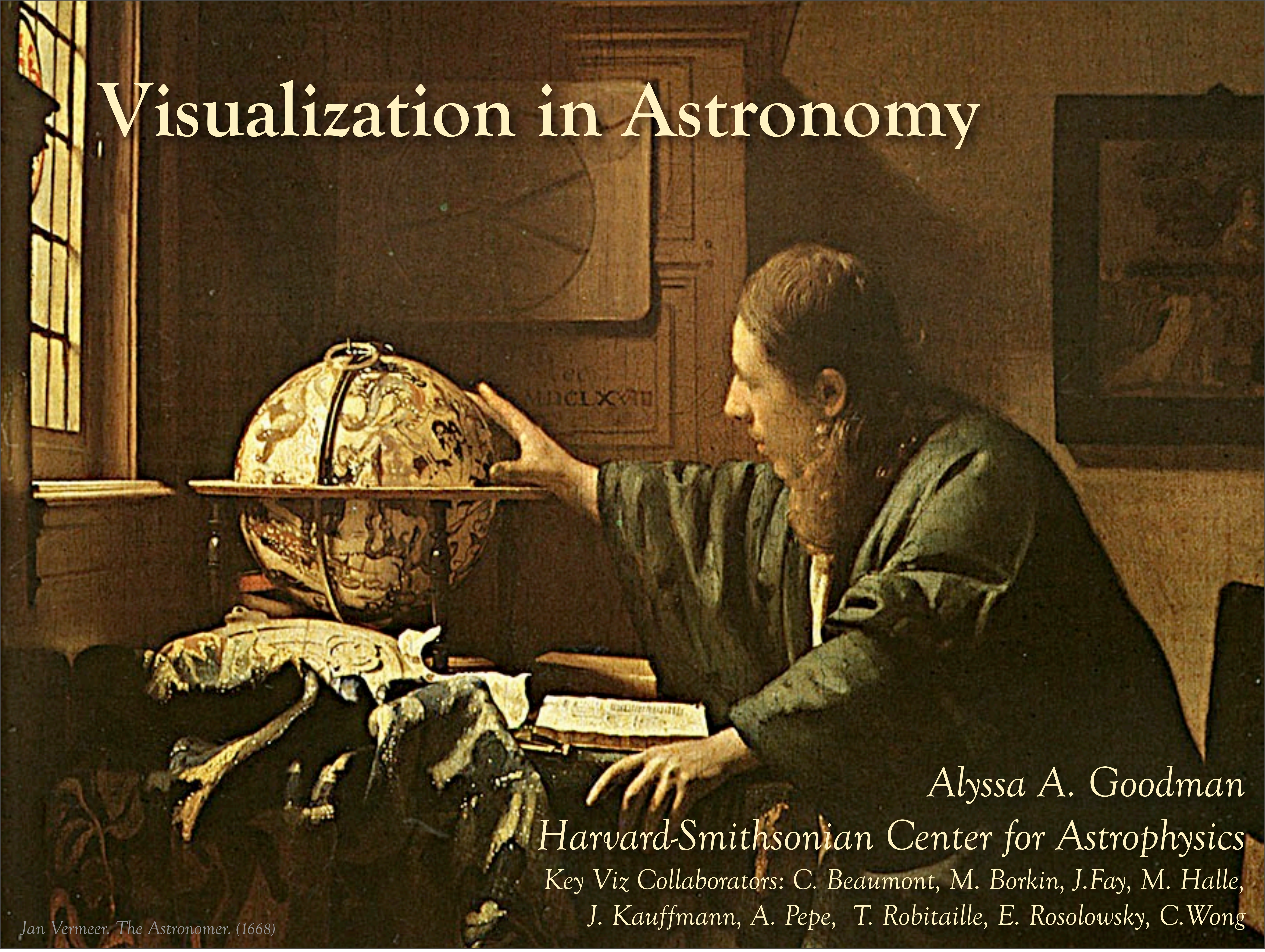
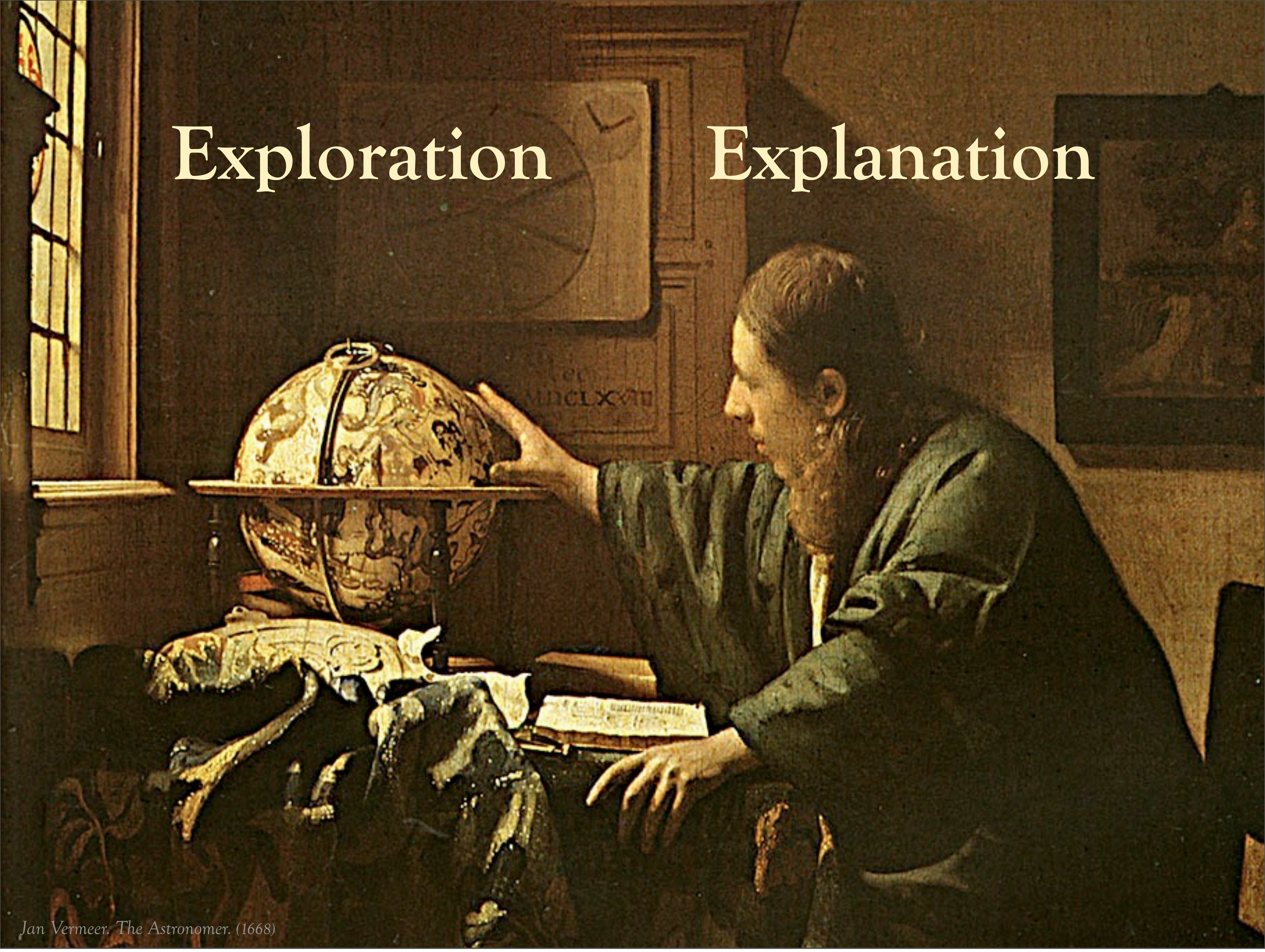



Tuesday, September 27, 2011

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




Tuesday, September 27, 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

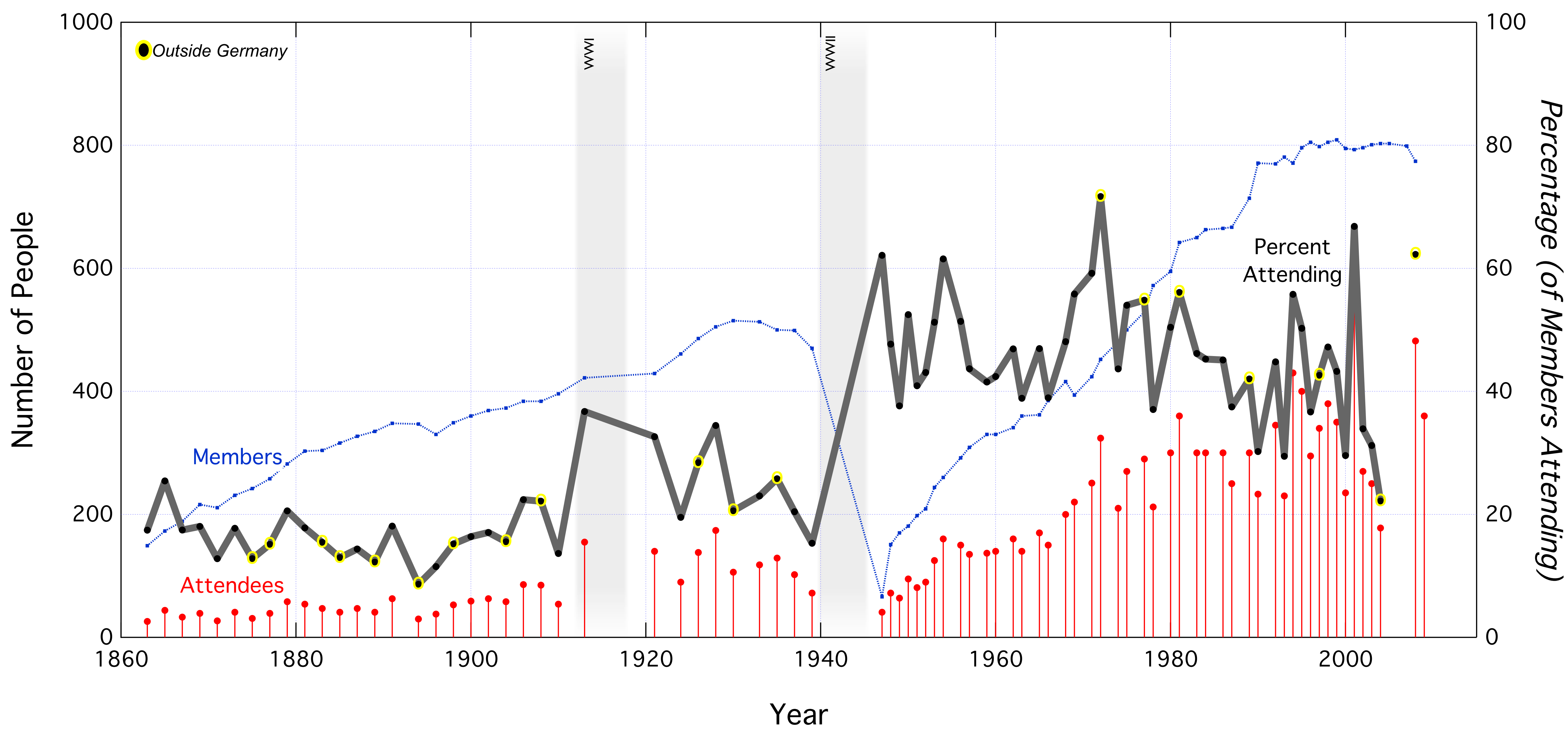
# History of the AG, http://www.astronomische-gesellschaft.org/Tagungen_eng.html

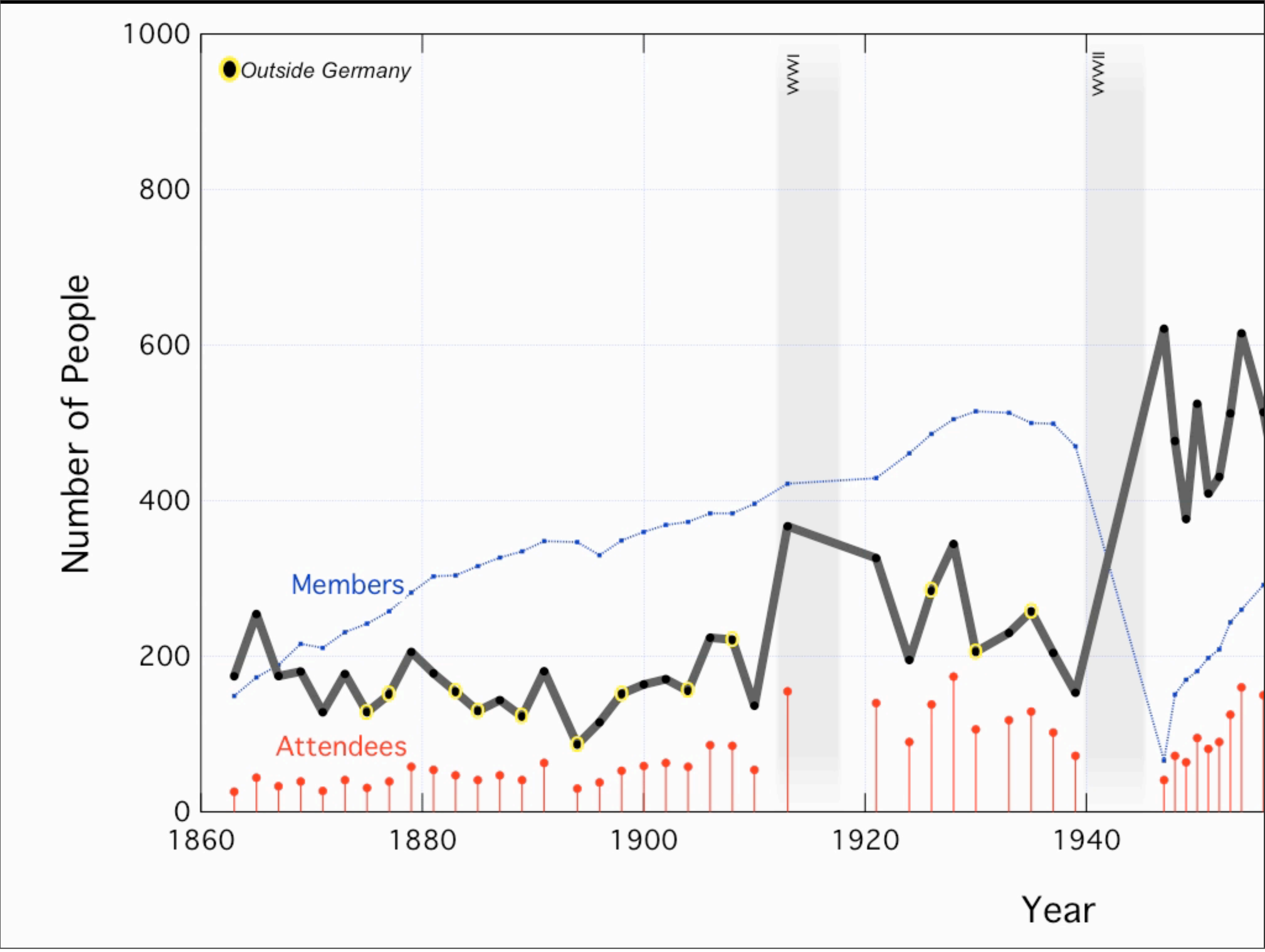
**All meetings since 1863:**

(RGA = regular meeting of the General Assembly)

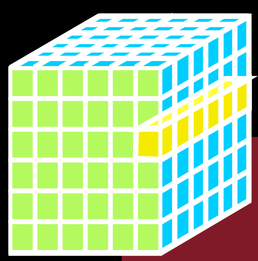
| Year | RGA | City | Date | Members | Atendees | Lectures | Poster |
|---|---|---|---|---|---|---|---|
| 2009 | 82 | Potsdam | Sep. 21-25 | | 360 (138 members, 92 students) | 185 (23 plenary talks, 162 splinter talks) | 102 |
| 2008 | 81 | Vienna (A) | Sep. 07-12 | 774 | 482 | | |
| 2008 | | *Spring:* Strasbourg (F) | Mar. 16-20 | | | | |
| 2007 | 80 | Würzburg | Sep. 24-28 | 799 | | | |
| 2005 | 79 | Cologne | Sep.26-Oct.01 | 803 | | | |
| 2004 | 78 | Prague (CZ) | Sep. 20-25 | 803 | 178 | 107 | 67 |
| 2004 | | Hamburg | July 05-09 | | | | |
| 2003 | 77 | Freiburg | Sep. 15-20 | 801 | 250 | 160 | 110 |
| 2002 | 76 | Berlin | Sep. 23-28 | 796 | 270 | 143 | 120 |
| 2001 | 75 | Munich | Sep. 10-14 | 793 | 530 | 187 | 180 |
| 2000 | 74 | Bremen | Sep. 18-23 | 795 | 235 | 87 | 110 |
| 2000 | | *Spring:* Nördlingen | May 16-20 | | | | |
| 2000 | | Heidelberg | Mar. 20-24 | | | | |
| 1999 | 73 | Göttingen | Sep. 20-25 | 809 | 350 | 163 | 125 |
| 1998 | 72 | Heidelberg | Sep. 14-19 | 805 | 380 | 184 | 155 |
| 1998 | | *Spring:* Gotha | May 11-15 | | | | |
| 1997 | 71 | Innsbruck (A) | Sep. 22-27 | 798 | 340 | 155 | 118 |
| 1996 | 70 | Tübingen | Sep. 16-21 | 805 | 295 | 165 | 98 |
| 1995 | 69 | Bonn | Sep. 18-23 | 796 | 400 | 198 | 116 |
| 1994 | 68 | Potsdam | Sep. 26-30 | 771 | 430 | 113 | 145 |

# History of the AG,



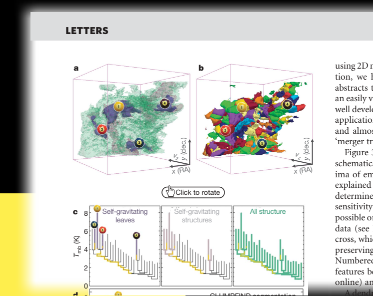



Tuesday, September 27, 2011



Data Dimensions Display

Linked Views

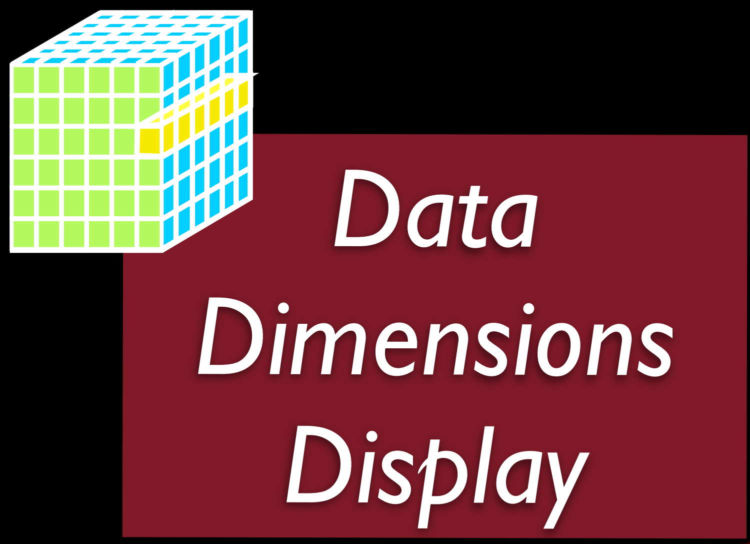


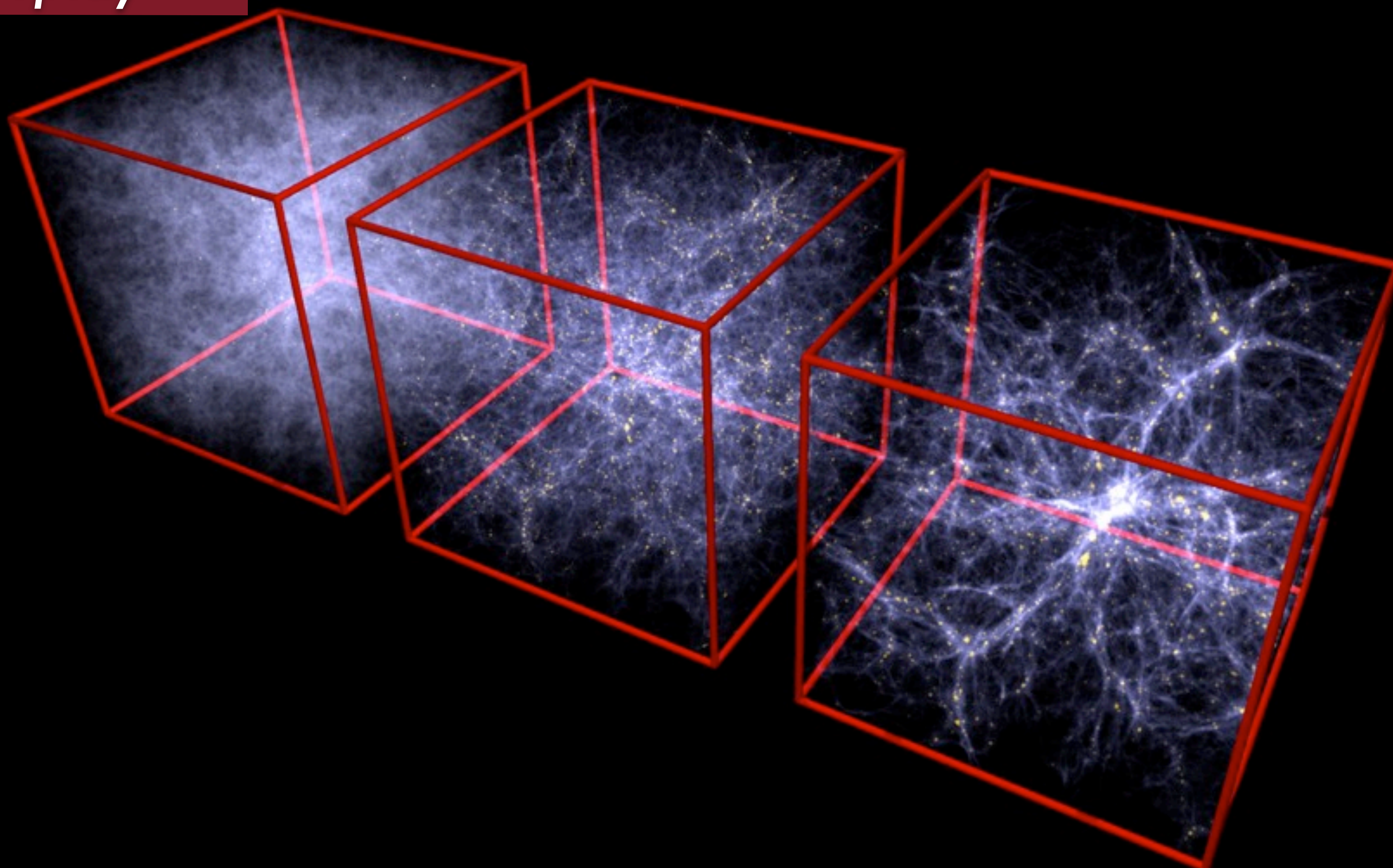
"Taste-Testing"
[Sim:Data]

Seamless Astronomy
[the future]

ADS Labs
[info viz]

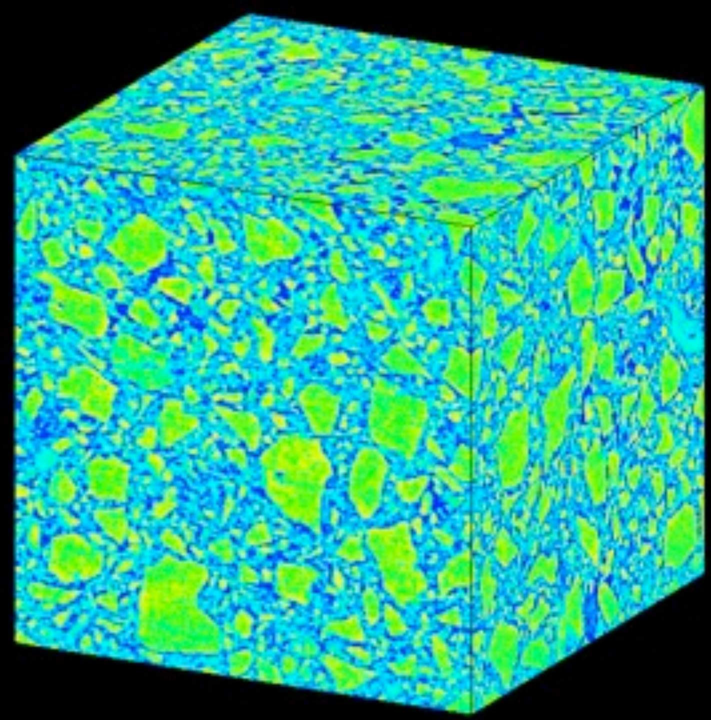
Open Data
Open Tools

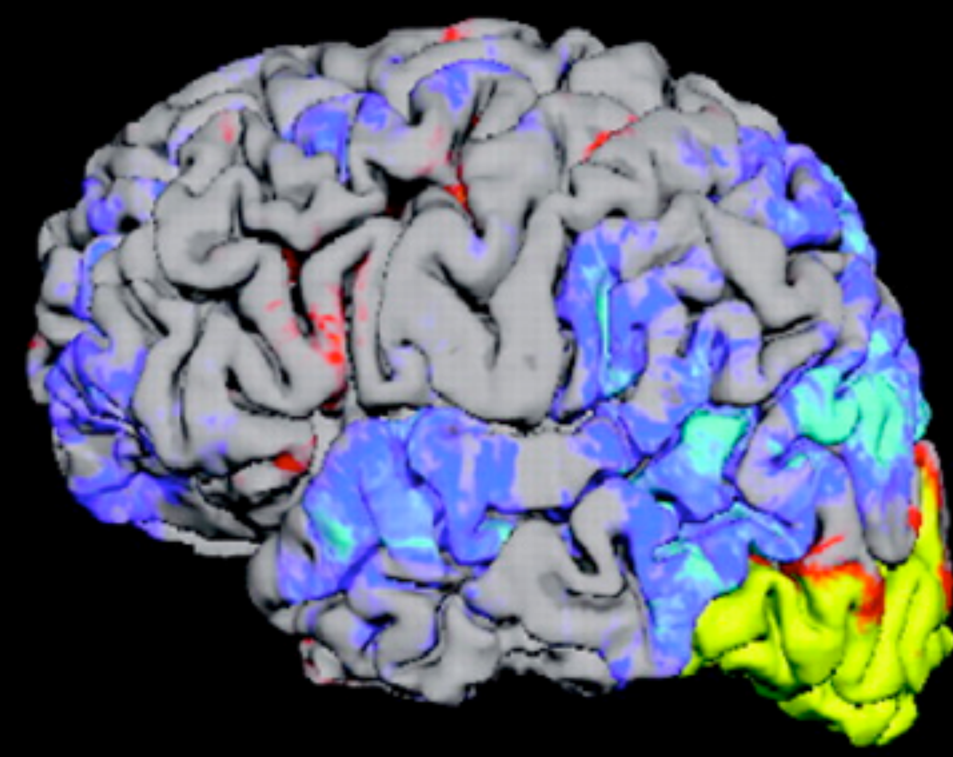


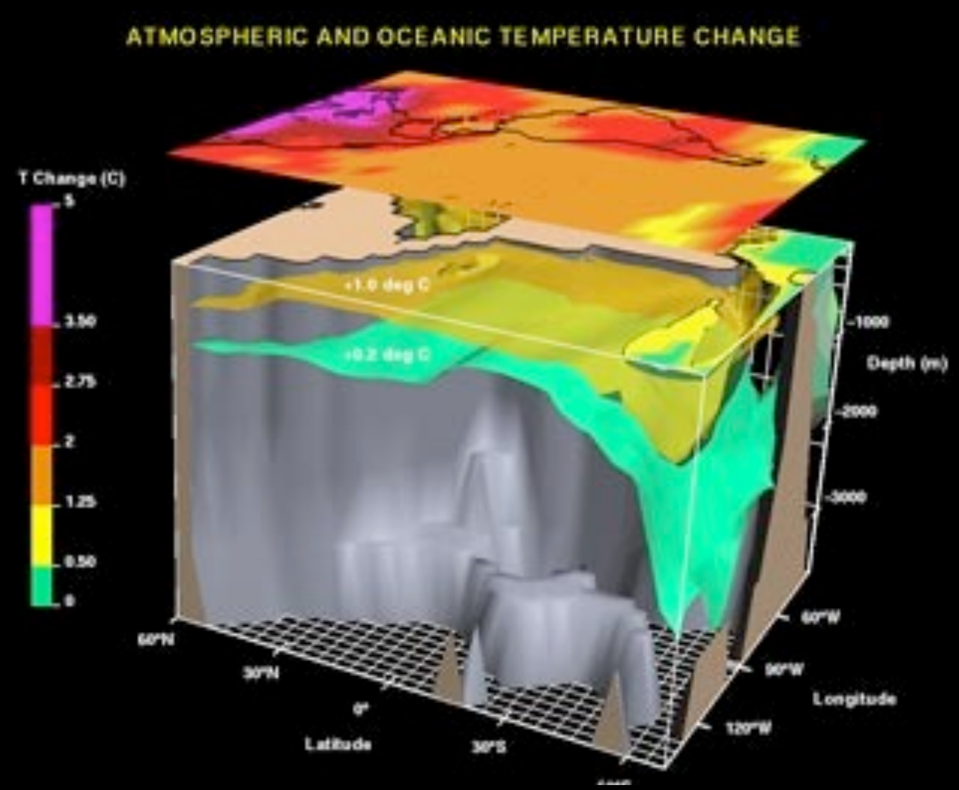



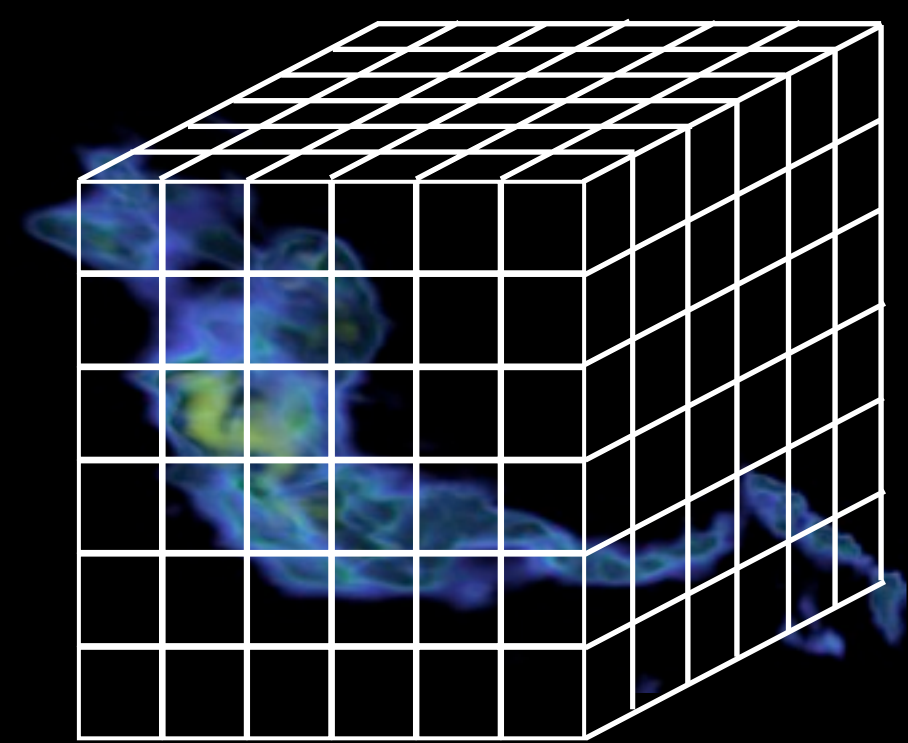
*Structure formation in the Universe (Volker Springel)*

# “High-Dimensional” Data



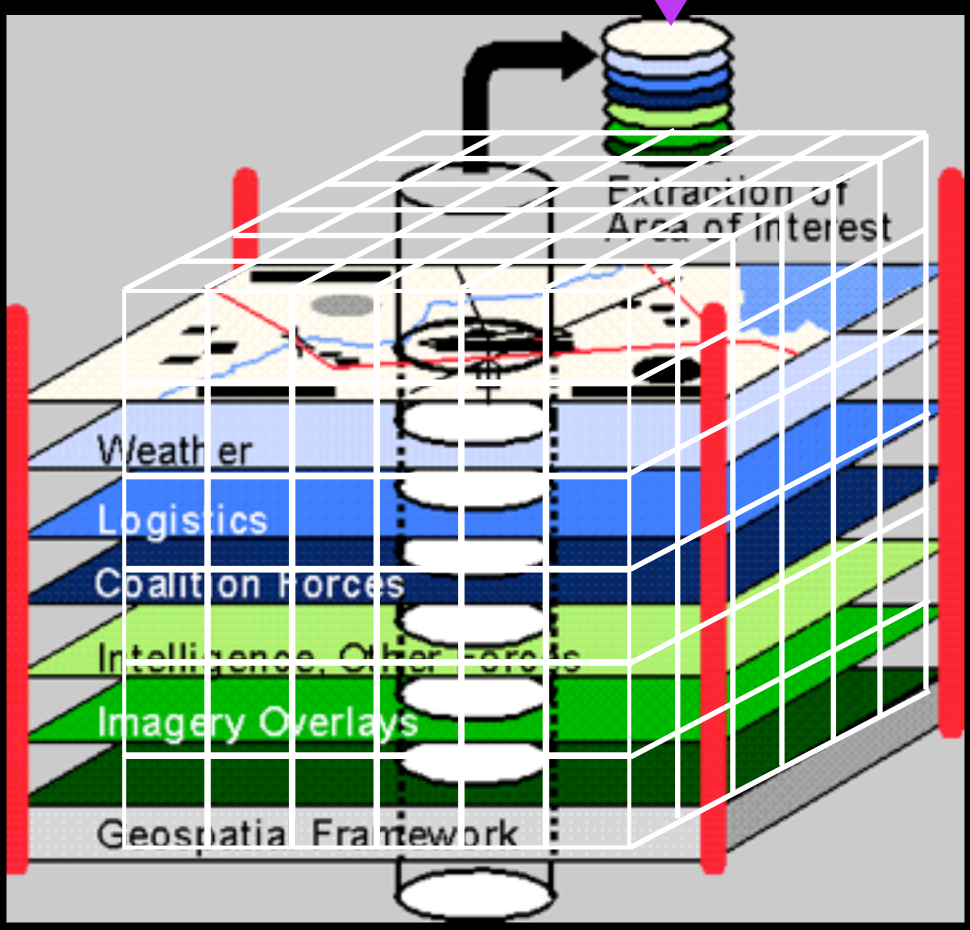




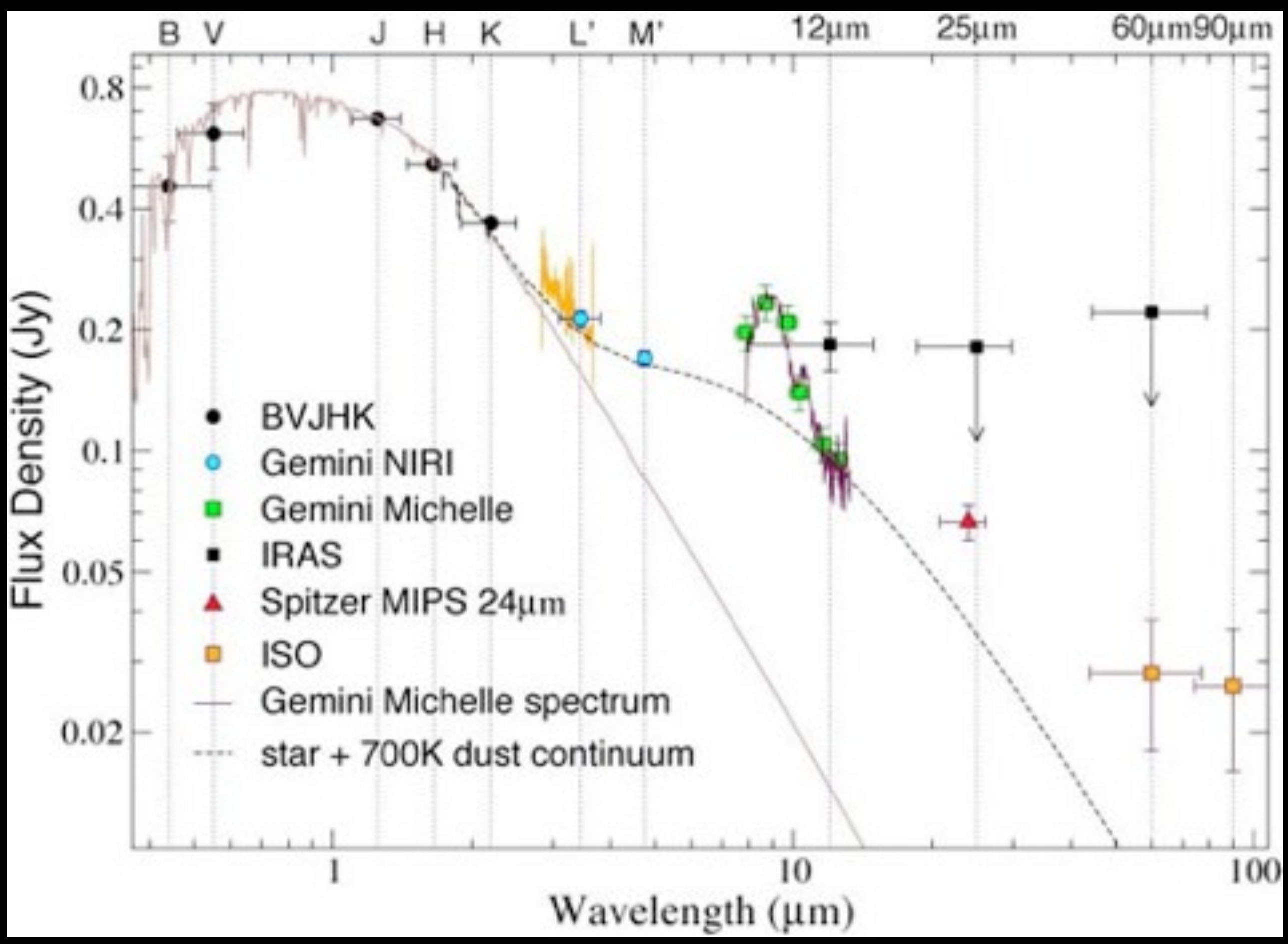



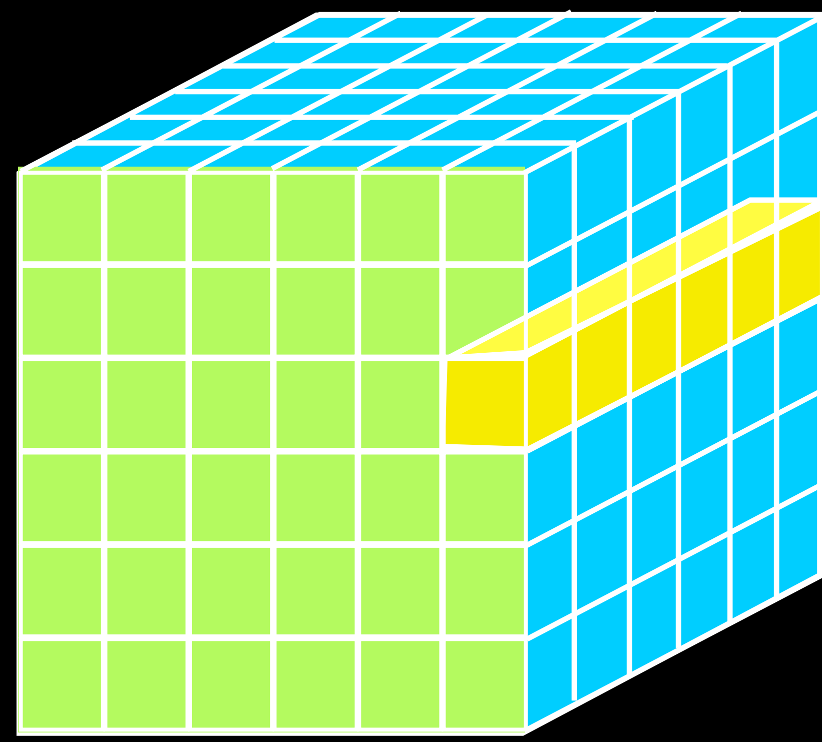
Tuesday, September 27, 2011

This




is a "spectral energy distribution"





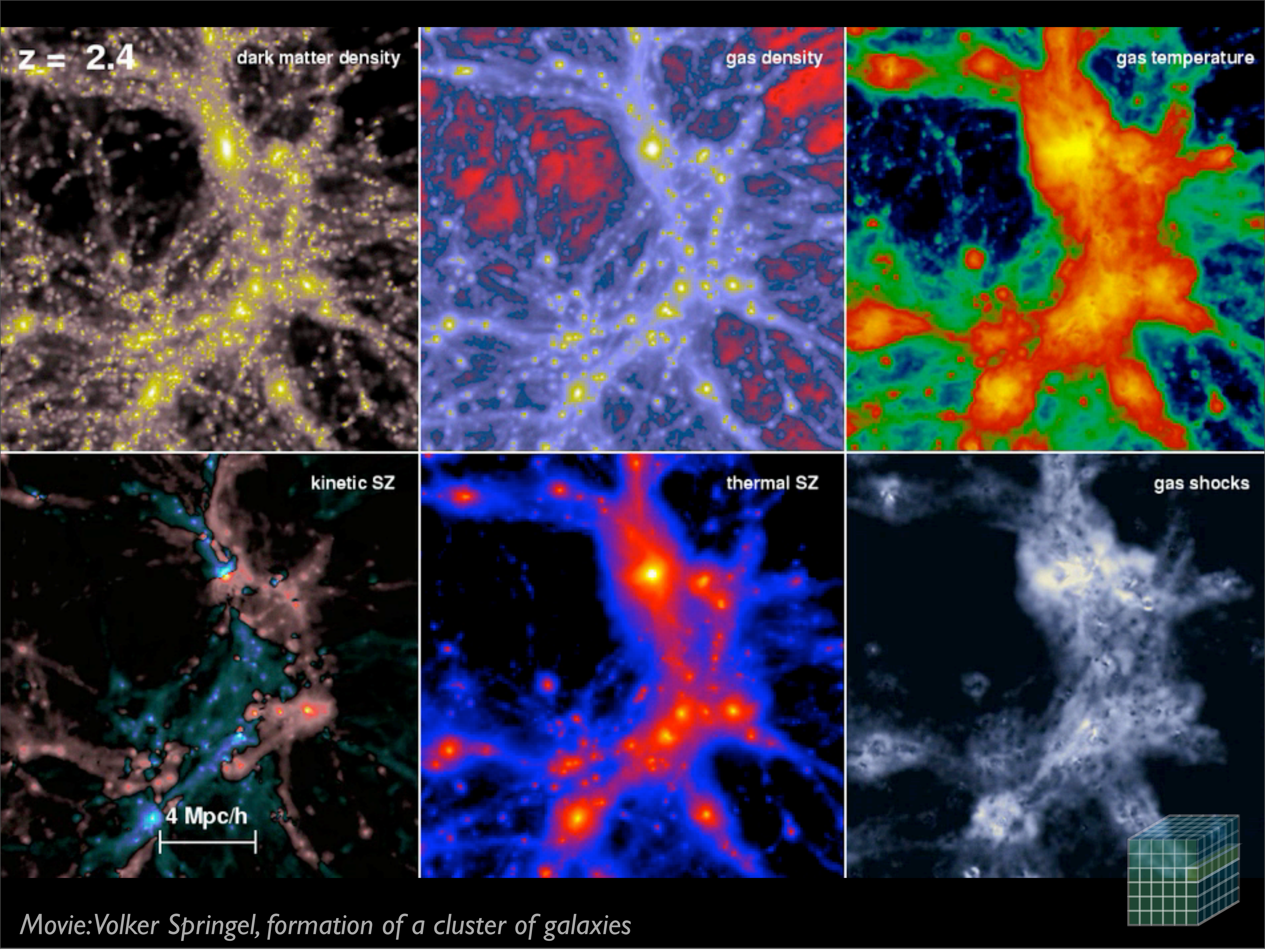
**GENERALLY**

**1D**: Columns = “Spectra”, “SEDs” or “Time Series”

**2D**: Faces or Slices = “Images”

**3D**: Volumes = “3D Renderings”, “2D Movies”

**4D**: Time Series of Volumes = “3D Movies”



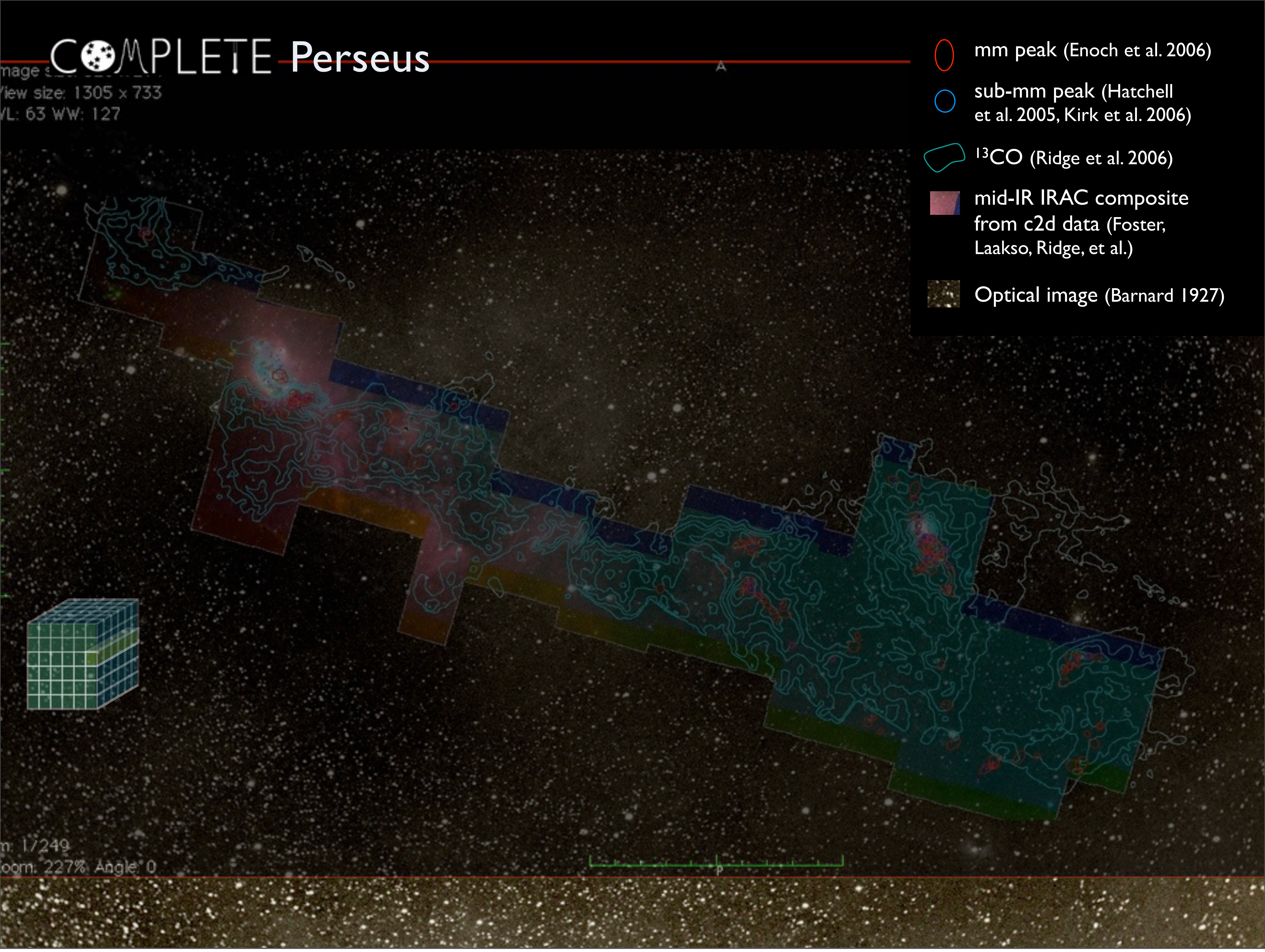

Tuesday, September 27, 2011

This high-resolution animation shows the formation of a rich cluster of galaxies from cosmological initial conditions. The simulation follows dark matter and baryonic gas.

Cluster Movie [divx5, 112 MB, 960x640]
Just dark matter [divx5, 34 MB, 768x768]

Credit: Volker Springel
Simulation code: Gadget-2

# COMPLETE Perseus

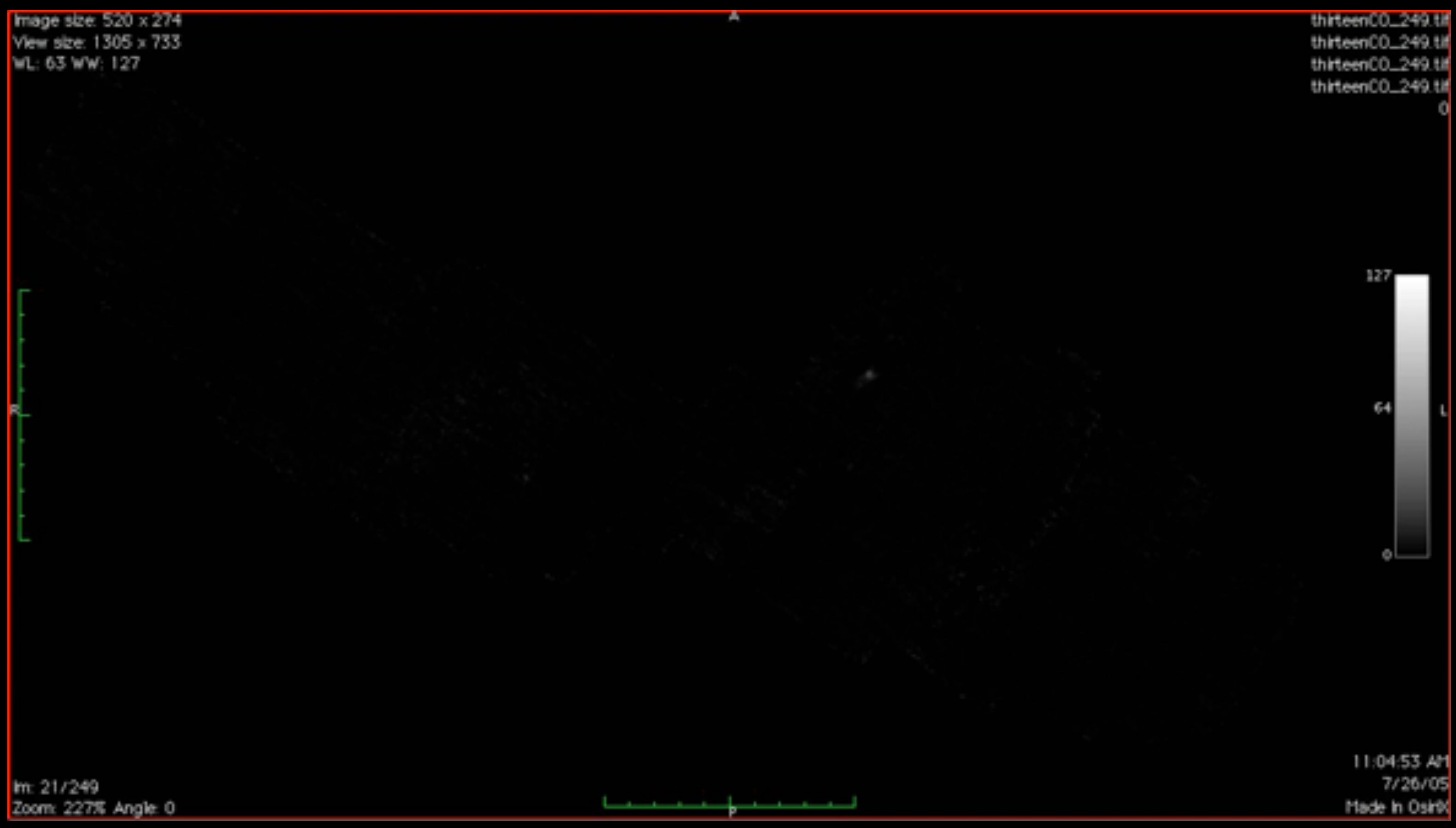


# "Astronomical Medicine"

"Keith"

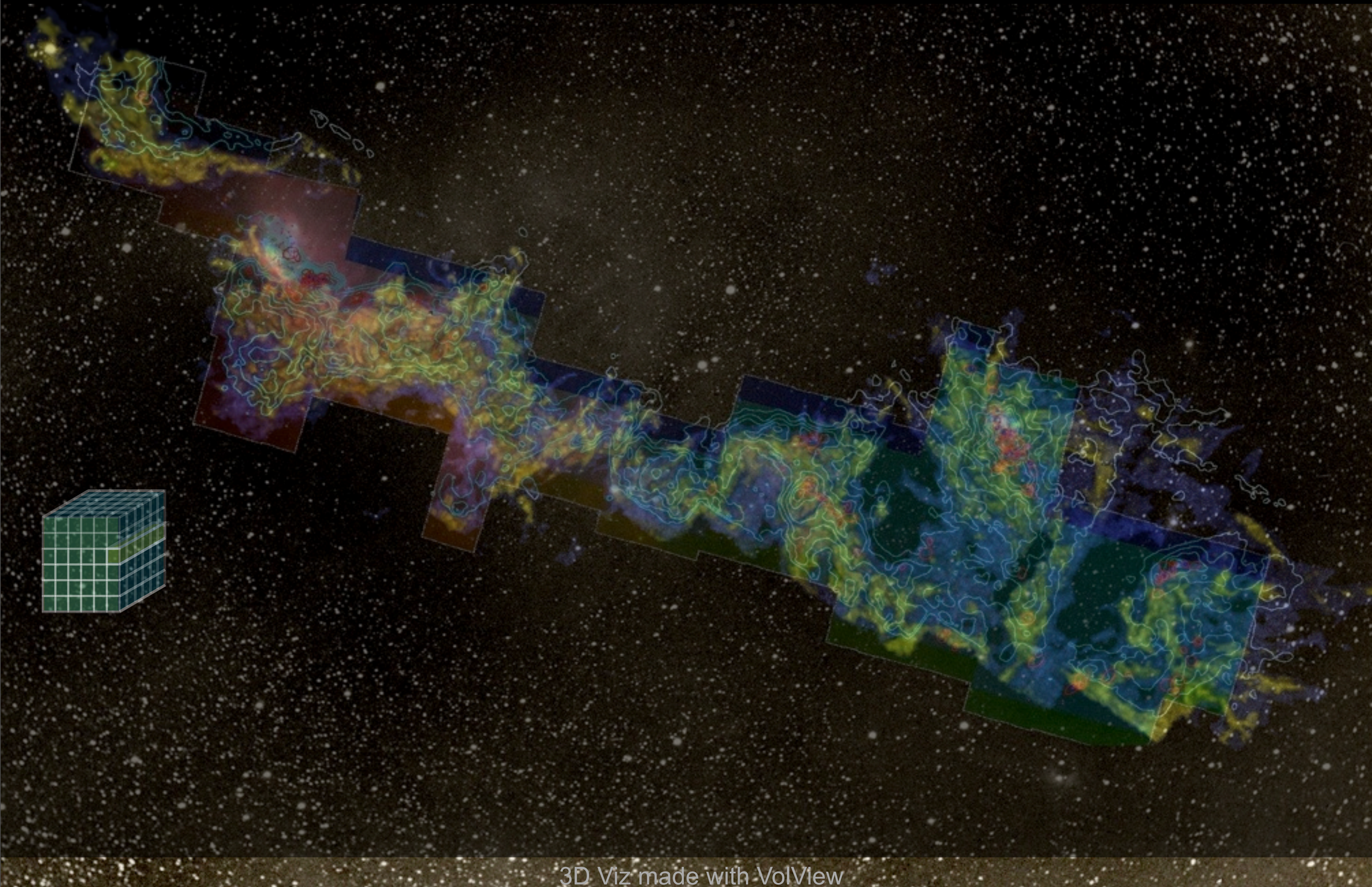
"z" is depth into head

"Perseus"



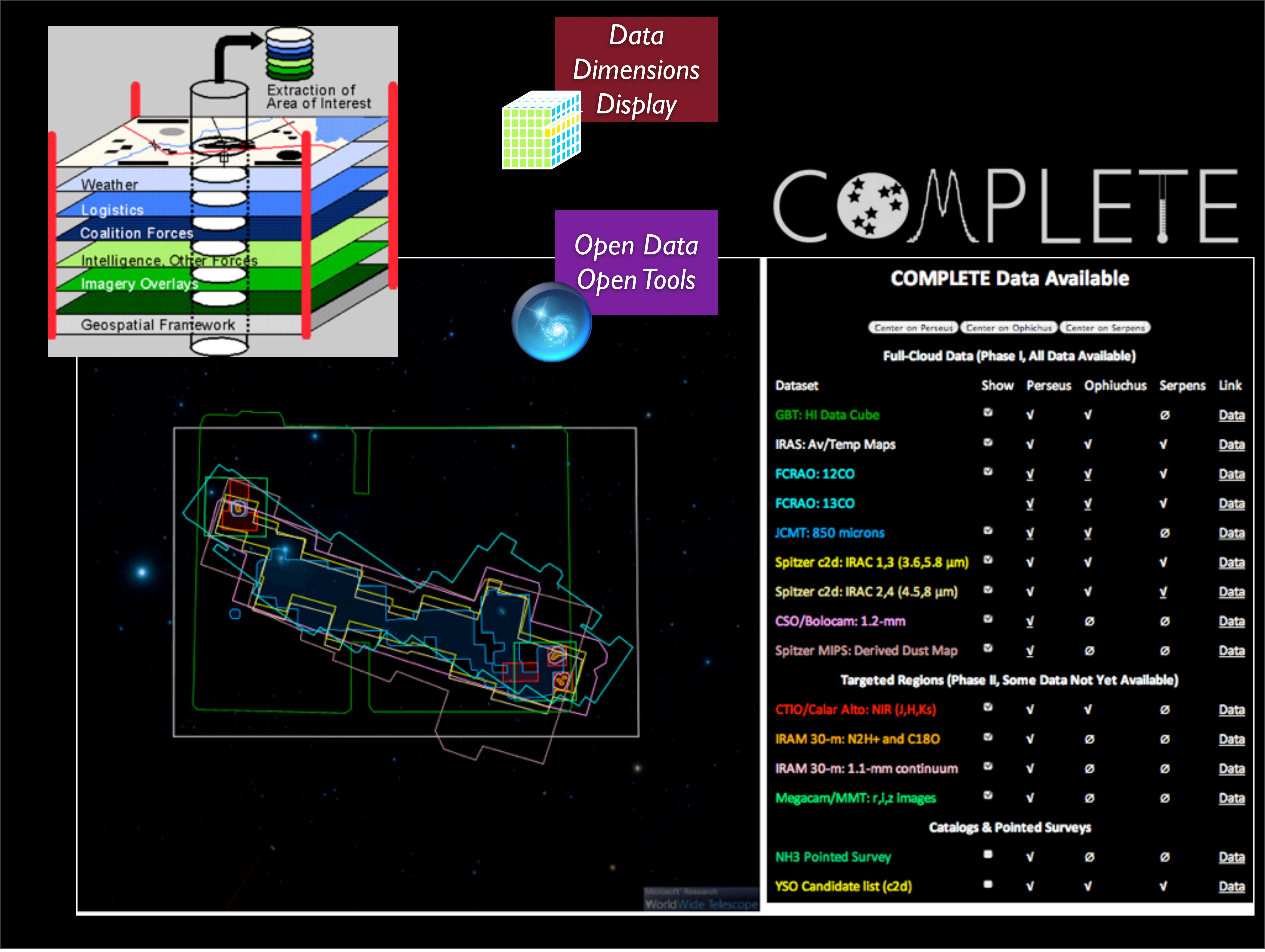

"z" is line-of-sight velocity

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

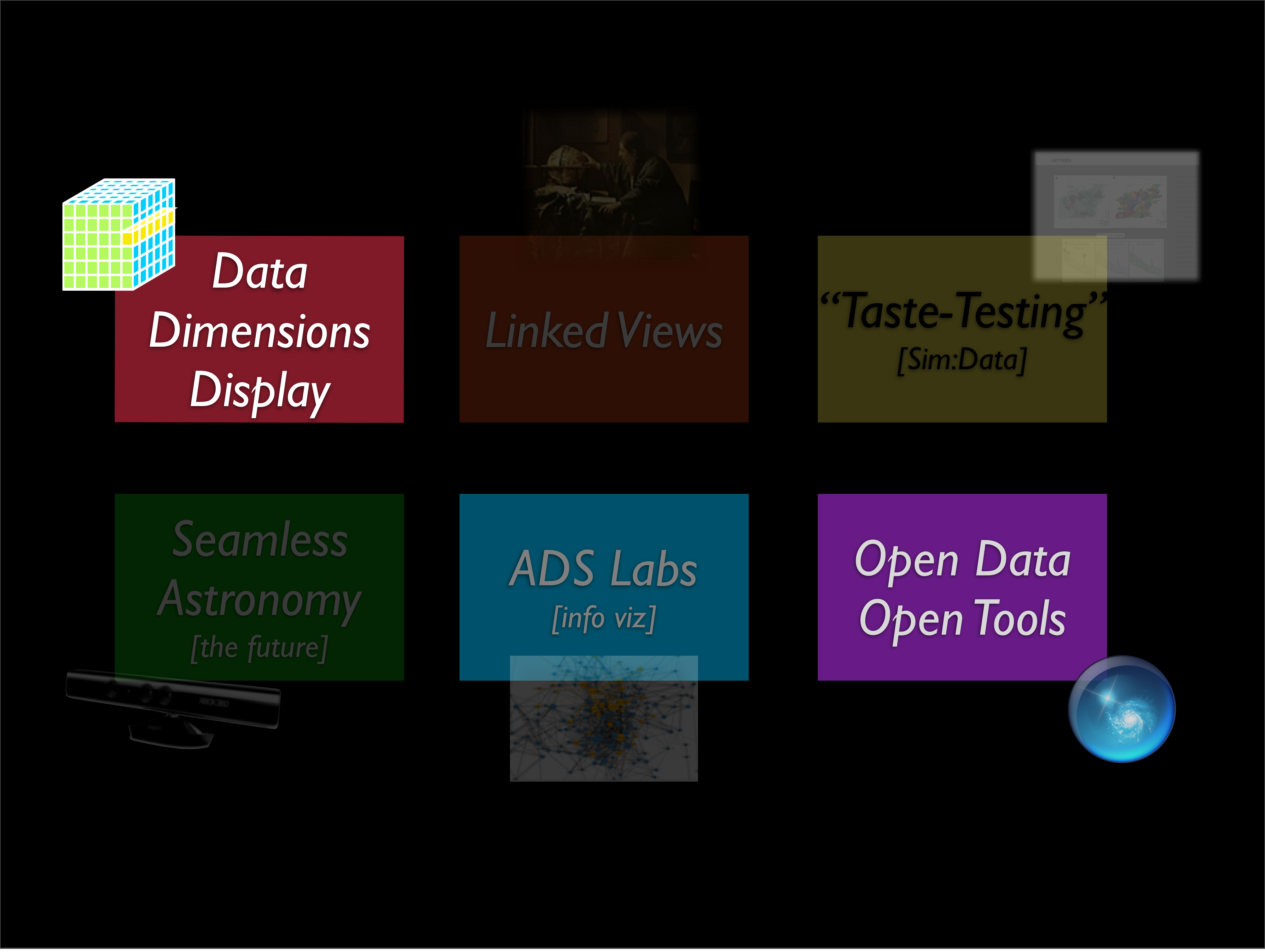



AstronomicalMedicine@IIC

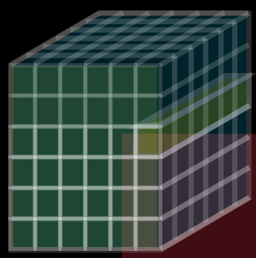
COMPLETE




Tuesday, September 27, 2011




Tuesday, September 27, 2011



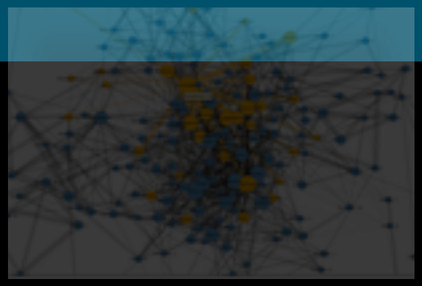
Data Dimensions Display

Linked Views

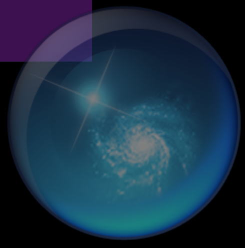
"Taste-Testing"
[Sim:Data]

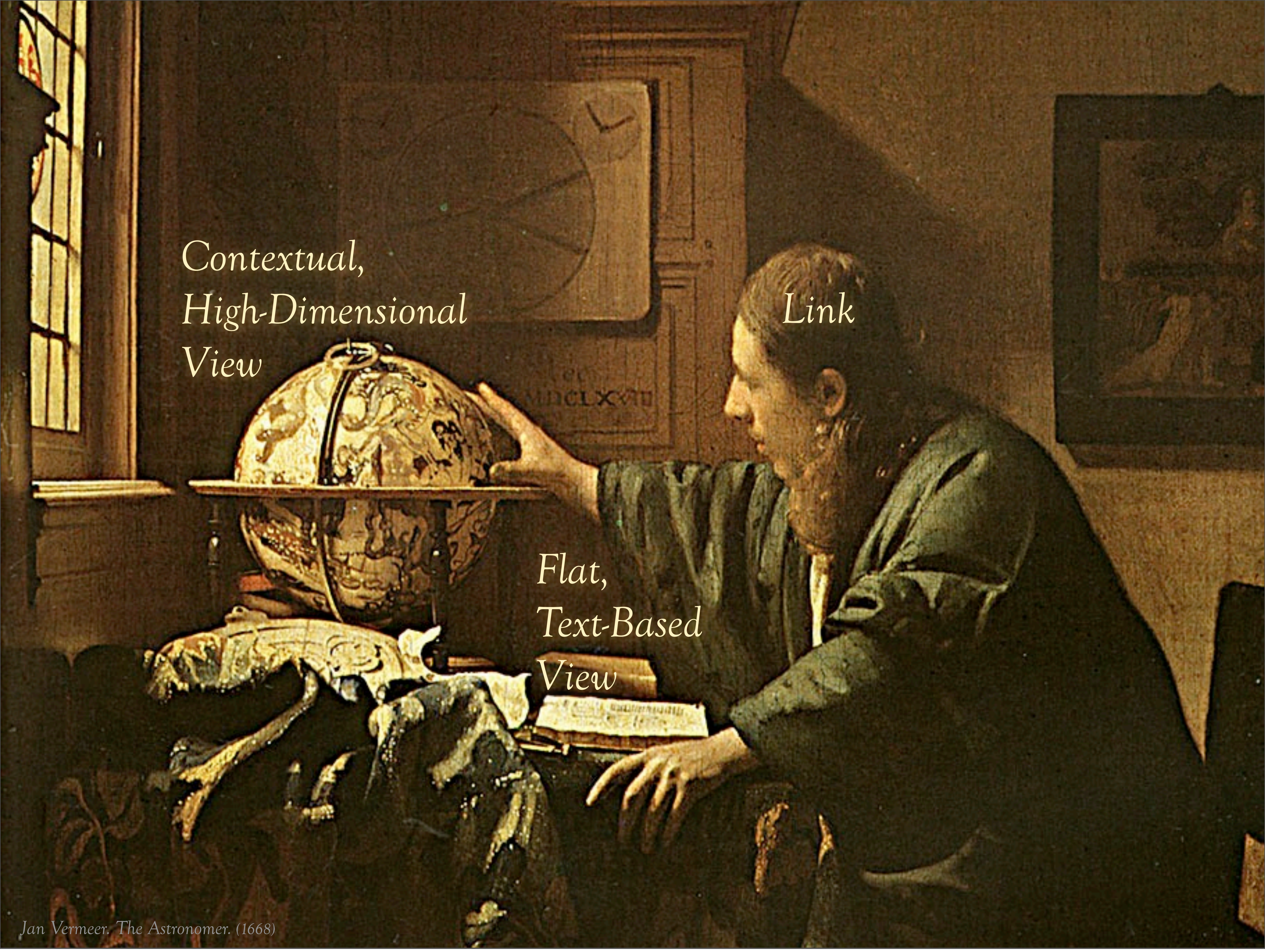
Seamless Astronomy
[the future]

ADS Labs
[info viz]



Open Data Open Tools






Tuesday, September 27, 2011

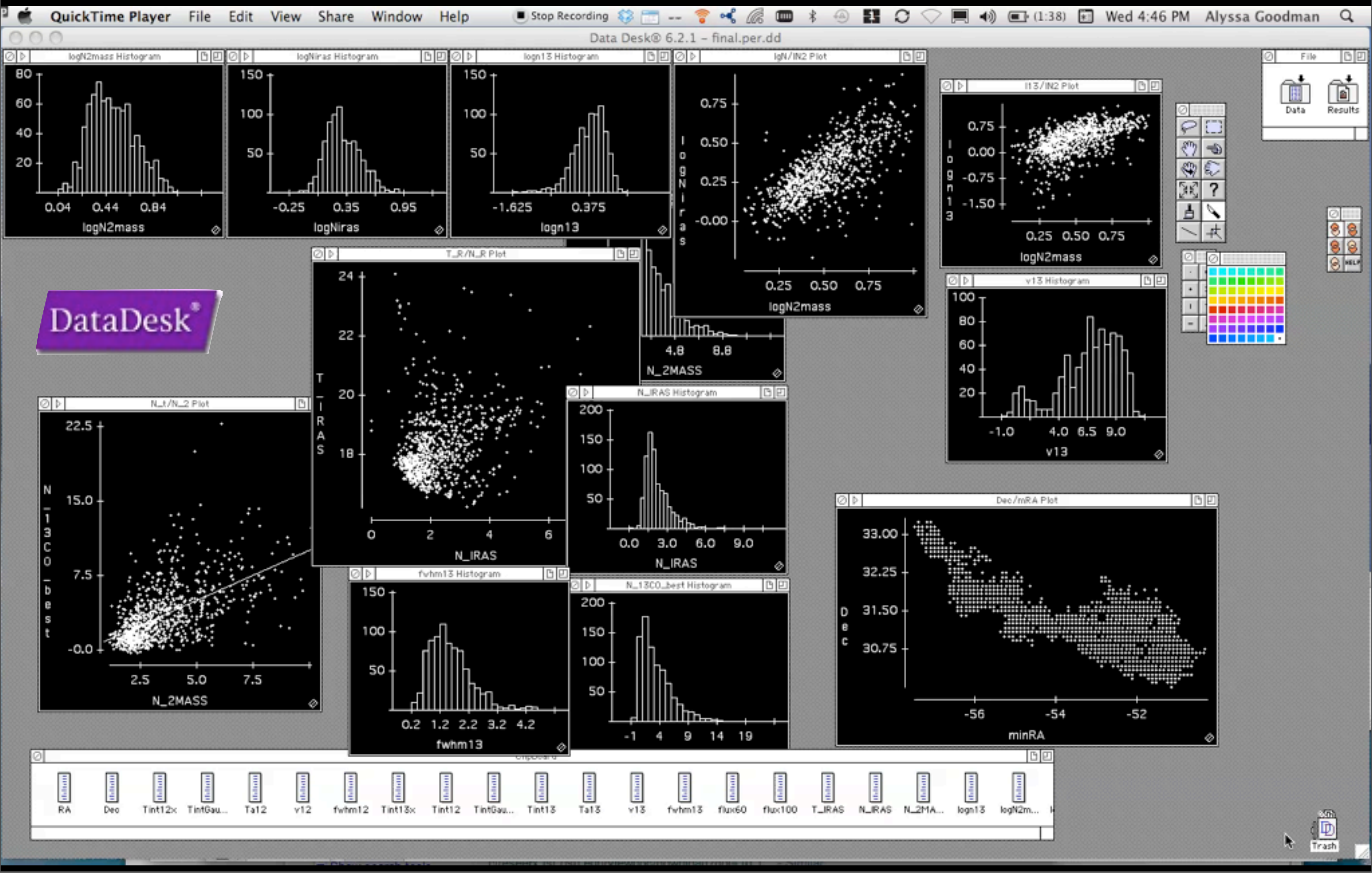
In the late 1960's and early 1970's the statistician John Tukey and his colleagues carried out seminal work on "Exploratory Data Analysis," which relied heavily on what I shall call "linked views" of data. In essence, linked views are those where selection of a sub-set of data in one view (e.g. lassoing a region on a map of points observed) effects the display of different dimensions of the same data in a different view (e.g. highlight all the observations in the map-lassoed region on a graph of temperature versus density). The best uses of linked views make the links happen in real time, and also offer algorithmic as well as graphical options for subset selection.

While a growing need for systems implementing linked views--thanks to expanding data set volumes--were clear twenty years ago (see Buja et al. 1991 and references therein) such systems were not widely implemented. In the commercial world, a notable implementation of linked views came with the advent of DataDesk which began (in 1986) as a Mac-only program, but DataDesk missed the opportunity for widespread adoption thanks to the dominance of Windows. In Astronomy, the VO-compatible program TOPCAT is one of the only current tools that implements linked views effectively. Not only can TOPCAT interlink points within its own display windows, it can also "live link" to other programs via use of the SAMP messaging protocol which can connect catalog statistical (ASCII-based) tools like TOPCAT to image-based applications such as ds9, GAIA, Aladin, and WorldWide Telescope and spectral tools like SPLAT.

In this talk, I will explain and demonstrate live why the future of linked views must address two currently unsolved issues. First, in many fields, including notably astronomy and medical imaging, "maps" have become three dimensional, so a two-dimensional "lasso-like" approach fails. We need three-dimensional selection tools. And second, while protocols like SAMP make linked views possible (at least within astronomy), they do not address the user-interface (too many windows, confusion, lack of screen real-estate) issues that arise when large, diverse, multi-dimensional data sets are to explored.

I shall conclude by urging the ADASS community take the lead within science on effectively interlinking views offered by modern tools, so that Astronomers' near-term options for exploratory data analysis can once again be a future-leaning example for other scientists.

# DataDesk (est. 1986)



# John Tukey's "Four Essentials" (c.1972)

Picturing | Rotation | Isolation | Masking

*Selection*

and these *"need to work together"*
in a *"dynamic display"*

Brushing | Linking

**Results...**

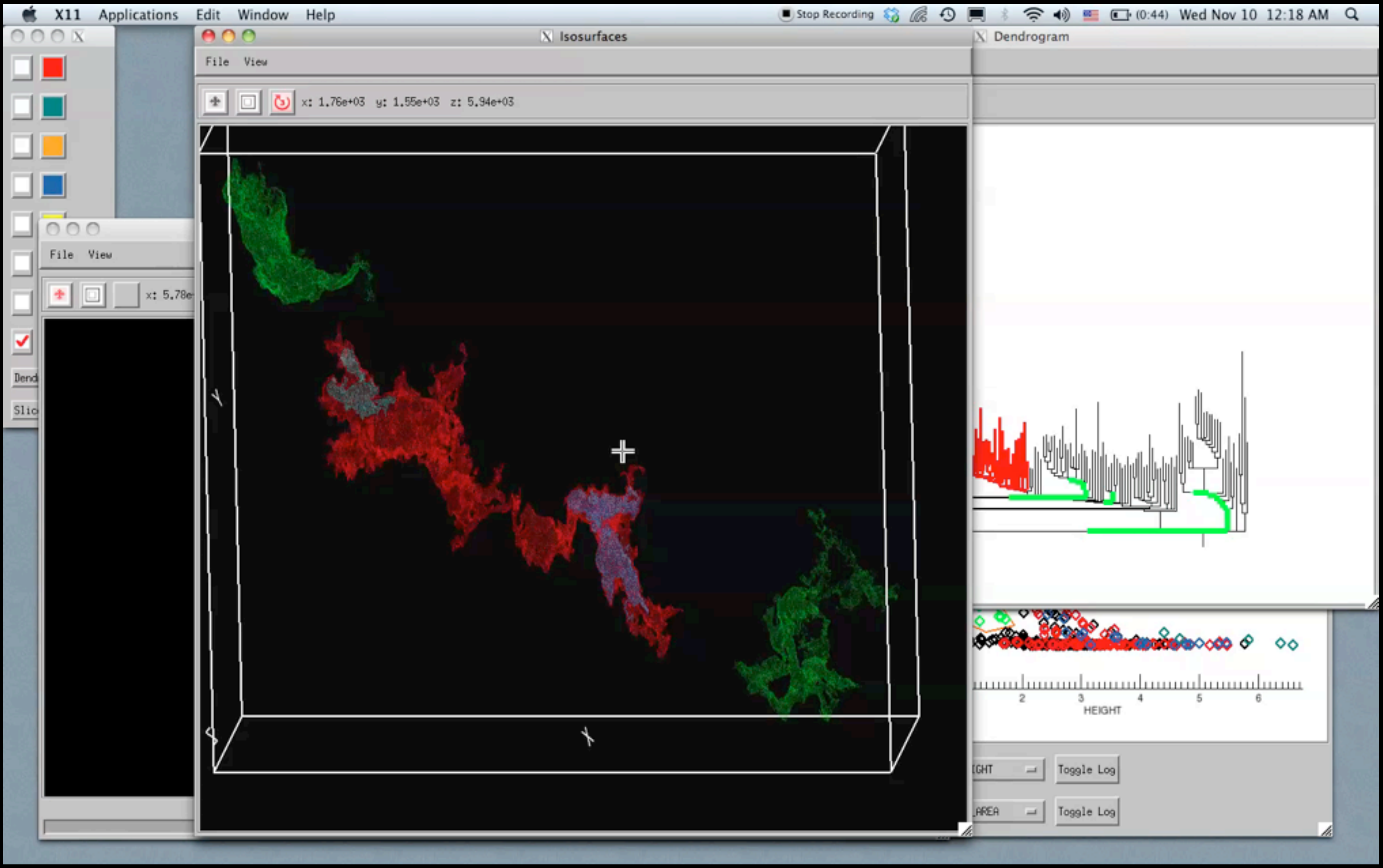
1. for immediate insight
2. as visual source of ideas for statistical algorithms (...relation to SVM)

**Warning**

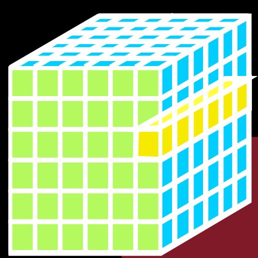
"details of control can make or break such a system"

*Watch the PRIM-9 video at: http://stat-graphics.org/movies/prim9.html*

# Exemplar: Linked Dendrogram Views in IDL

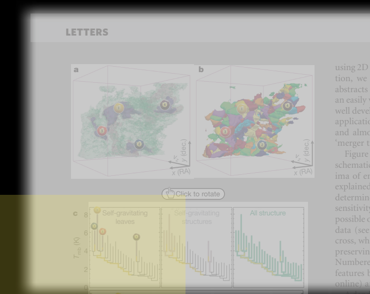



*Video & implementation: Christopher Beaumont, CfA/UHawaii;*
*inspired by AstroMed work of Douglas Alan, Michelle Borkin, AG, Michael Halle, Erik Rosolowsky*



Data Dimensions Display

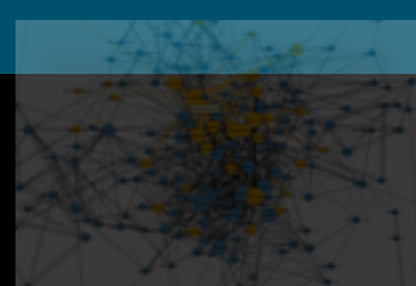
Linked Views



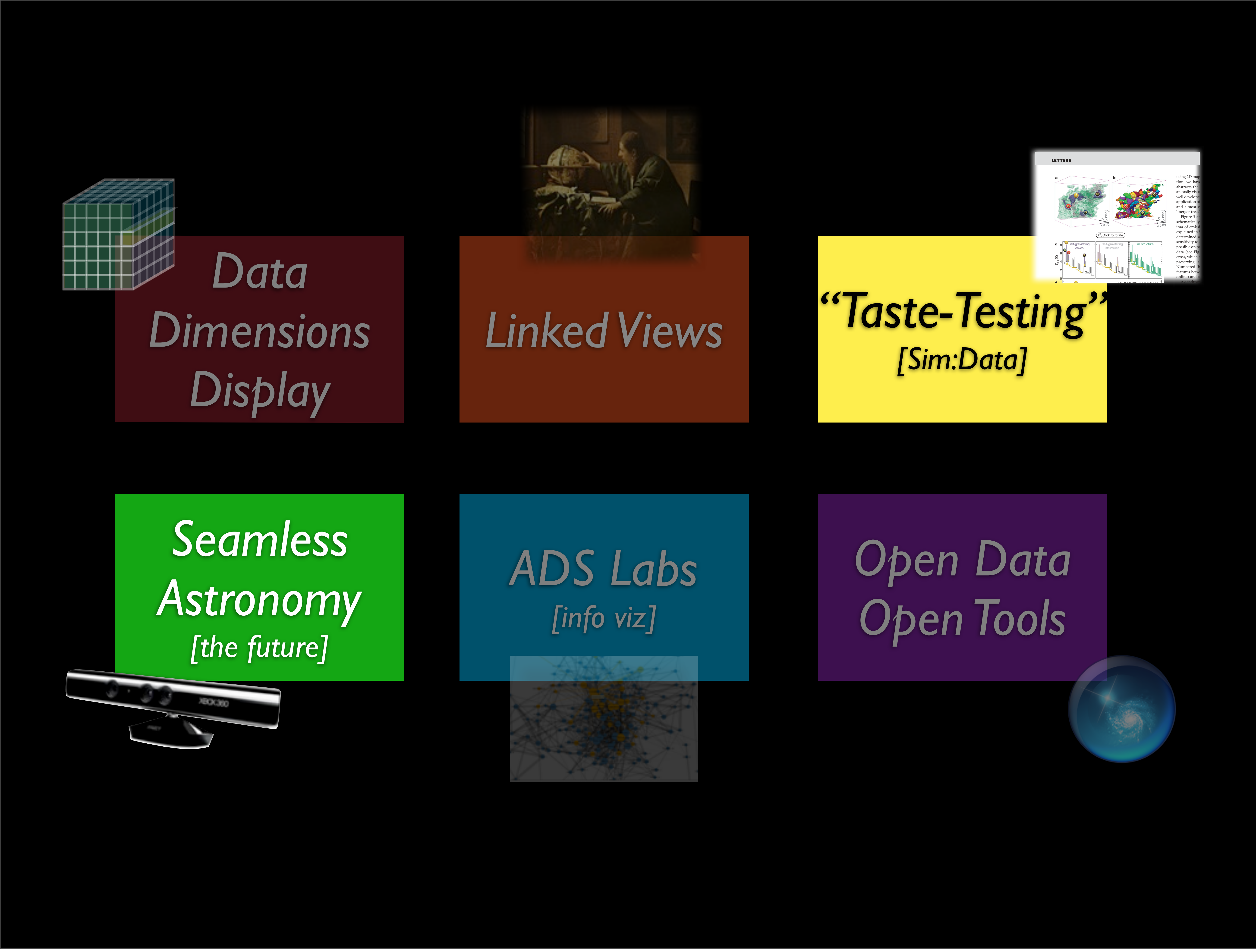
"Taste-Testing"
[Sim:Data]

Seamless Astronomy
[the future]

ADS Labs
[info viz]



Open Data Open Tools


Data Dimensions Display
Linked Views
"Taste-Testing"
[Sim:Data]
Seamless Astronomy
[the future]
ADS Labs
[info viz]
Open Data
Open Tools

# Seamless Astronomy
## [the future]

# “3D PDF”

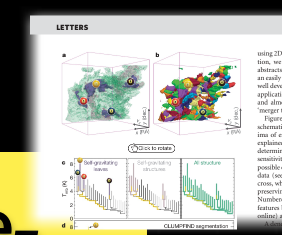
[Demo]

Seamless Astronomy

[the future]

“Taste-Testing”

[Sim:Data]





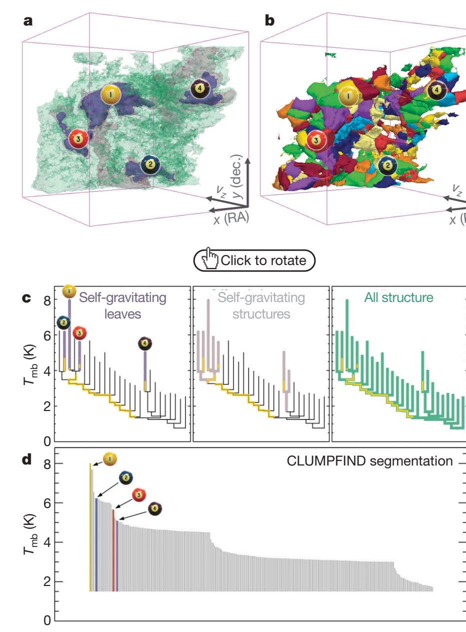



**Figure 2 | Comparison of the ‘dendrogram’ and ‘CLUMPFIND’ feature-identification algorithms as applied to $^{13}CO$ emission from the L1448 region of Perseus. a**, 3D visualization of the surfaces indicated by colours in the dendrogram shown in **c**. Purple illustrates the smallest scale self-gravitating 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 ($-0.5\ \mathrm{km\,s^{-1}}$) to back ($8\ \mathrm{km\,s^{-1}}$).

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 ($\sigma_v$) 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 $X_{13CO} = 8.0 \times 10^{20}\ \mathrm{cm^2\,K^{-1}\,km^{-1}\,s}$ (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 $\alpha_{obs}$ 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.

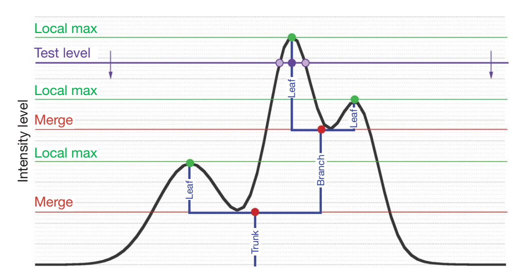



**Figure 3 | Schematic illustration of the dendrogram process.** Shown is the construction of a dendrogram from a hypothetical one-dimensional emission profile (black). The dendrogram (blue) can be constructed by ‘dropping’ a test constant emission level (purple) from above in tiny steps (exaggerated 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.

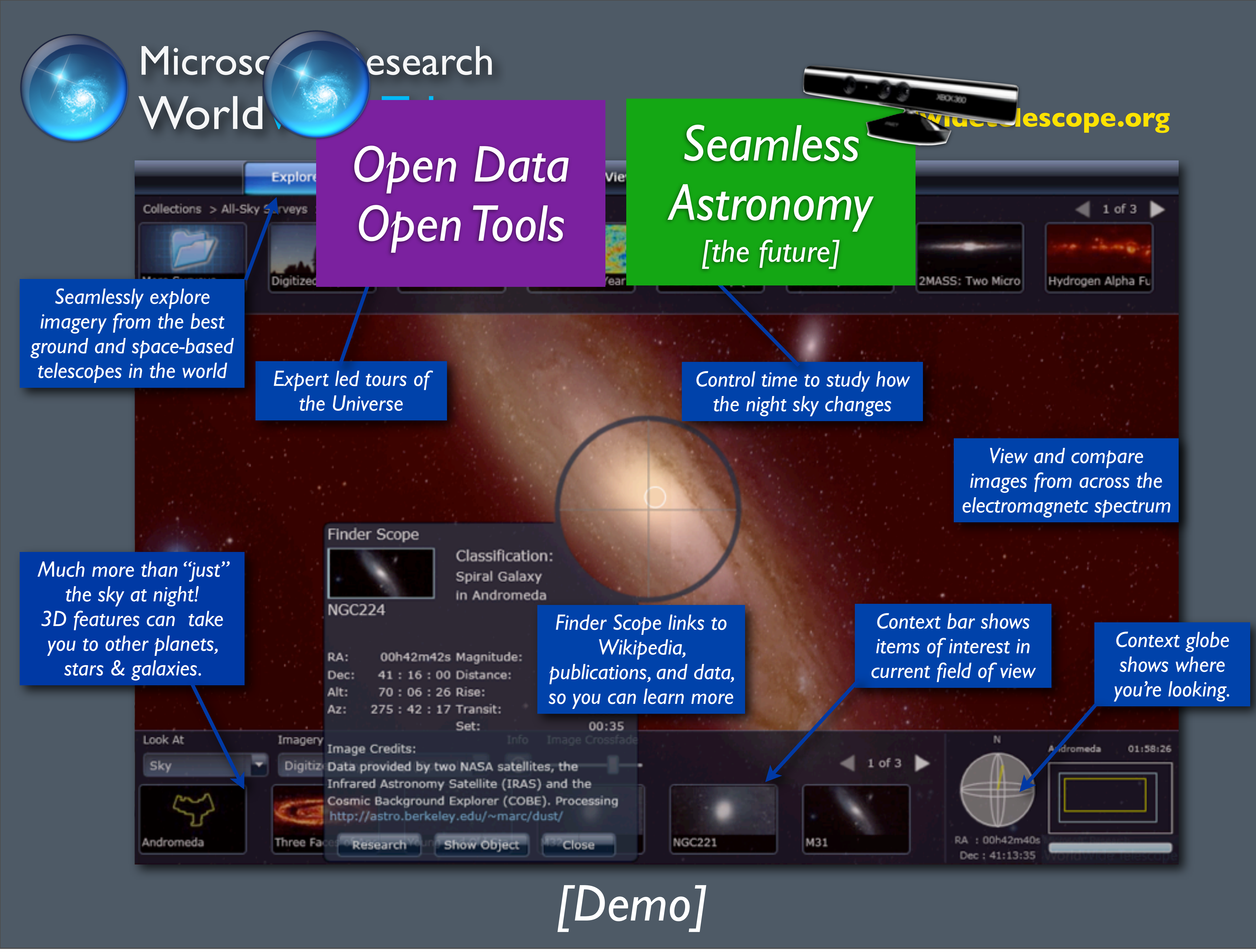


©2009 Macmillan Publishers Limited. All rights reserved

Goodman et al. *Nature*, 2009

Open Data
Open Tools

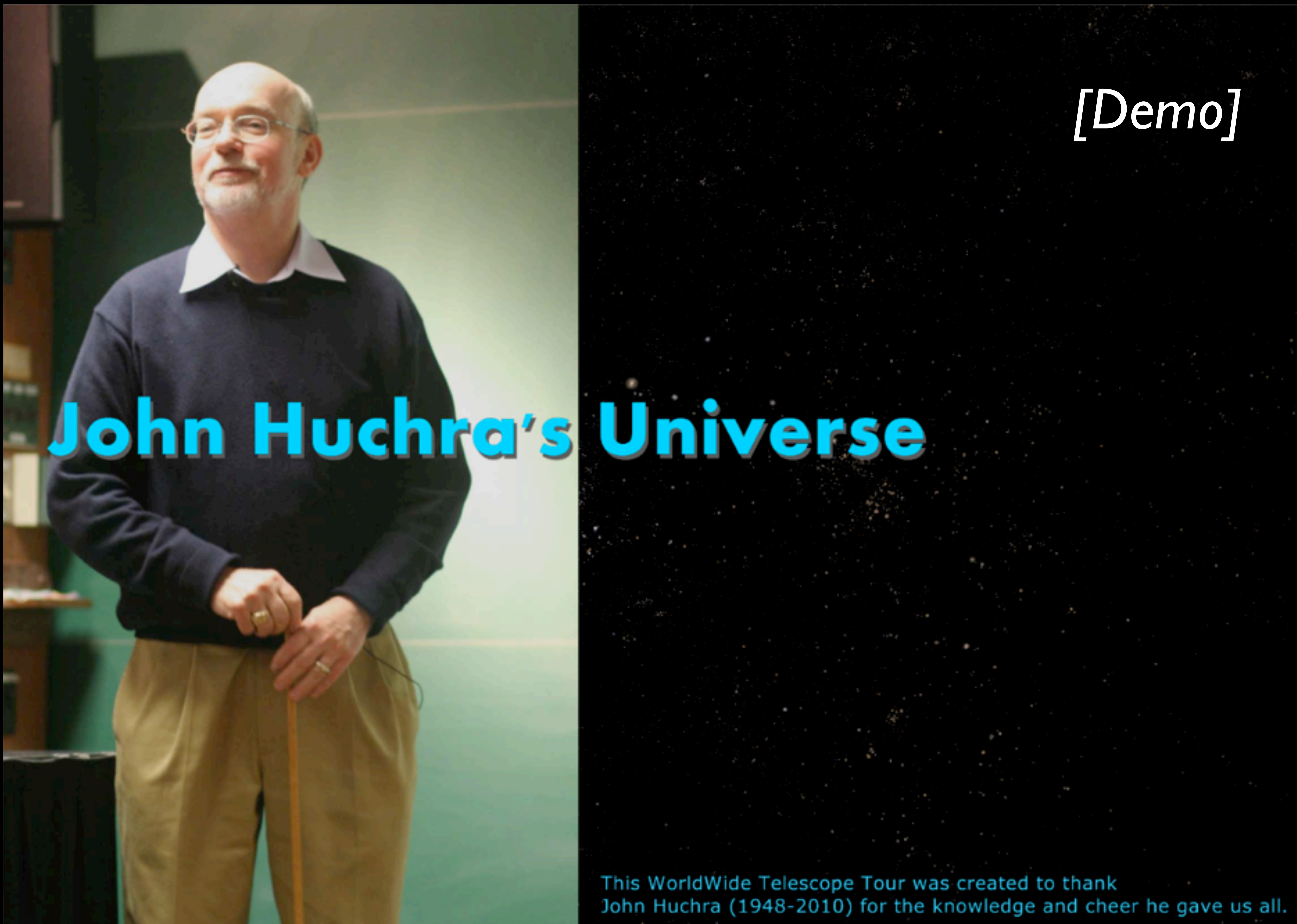
Seamless
Astronomy
*[the future]*

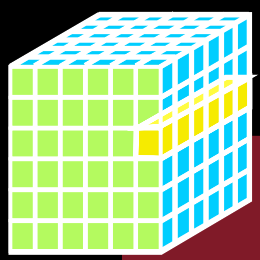

Tuesday, September 27, 2011


John Huchra's Universe
This WorldWide Telescope Tour was created to thank
John Huchra (1948-2010) for the knowledge and cheer he gave us all.

[Demo]

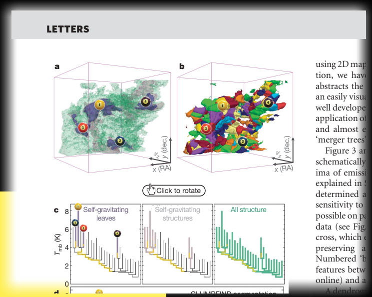



*also available on YouTube (search "John Huchra's Universe")*



Data Dimensions Display

Linked Views

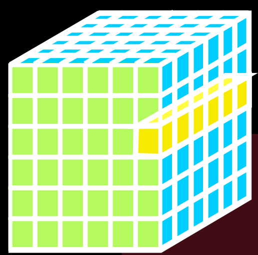



"Taste-Testing"
[Sim:Data]

Seamless Astronomy
[the future]

ADS Labs
[info viz]

Open Data Open Tools



Data Dimensions Display

Linked Views

"Taste-Testing"
[Sim:Data]

Seamless Astronomy
[the future]

ADS Labs
[info viz]

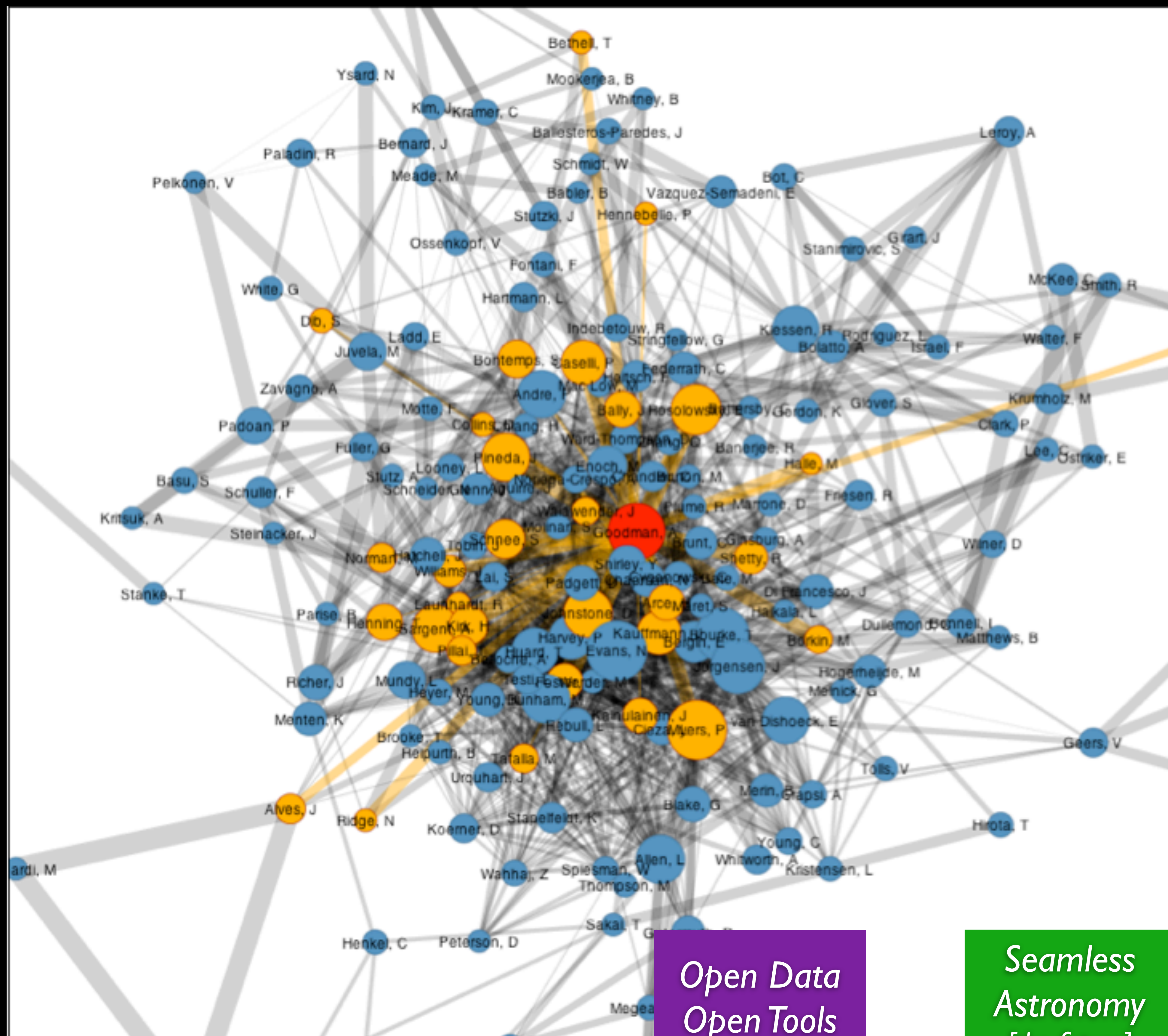
Open Data
Open Tools

COMPLETE

Citation Network

*(from ADS Labs)*

*[Demo]*

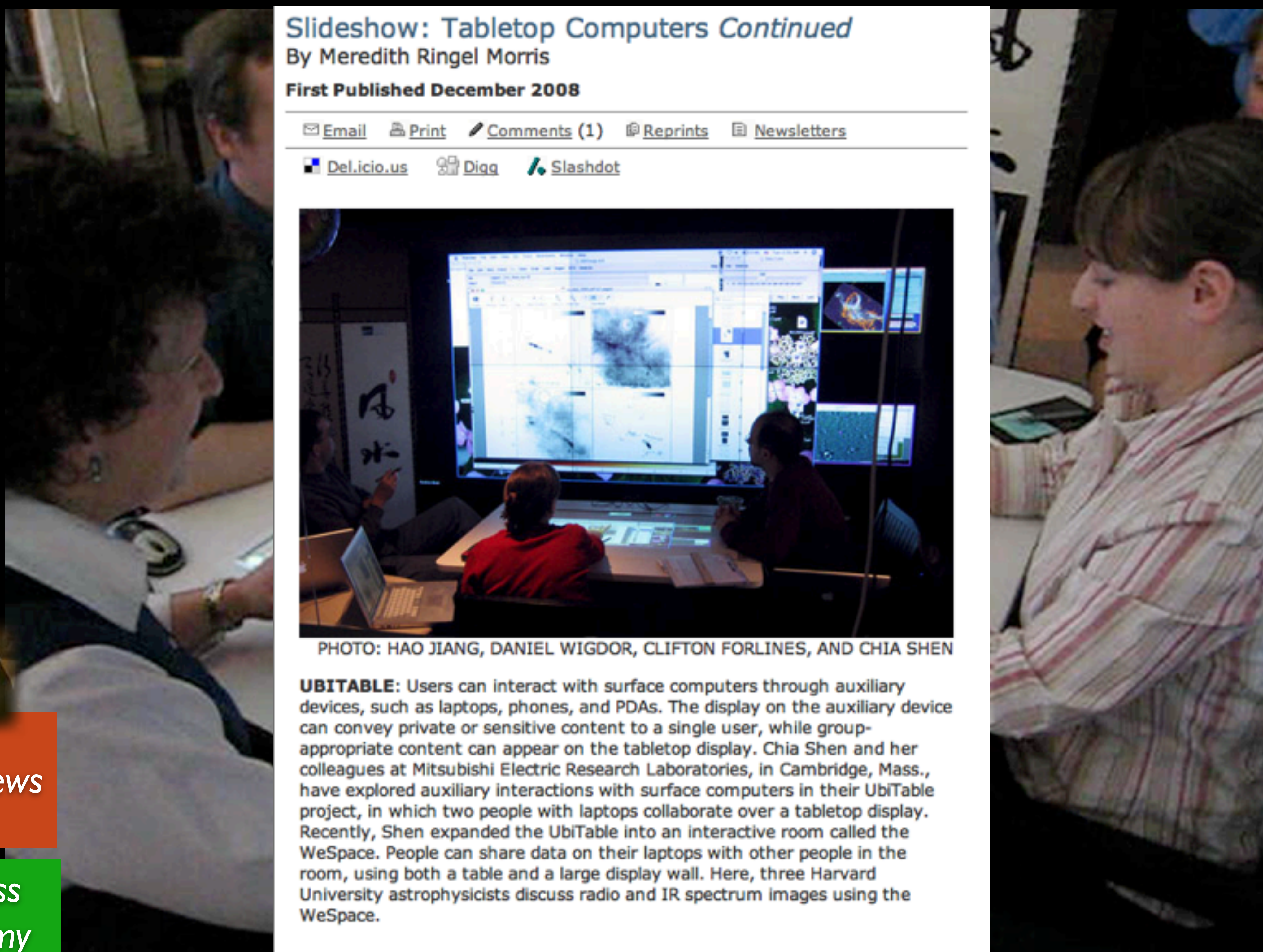



Open Data
Open Tools

*Seamless Astronomy*
*[the future]*

Tuesday, September 27, 2011

# “Off the Desktop”



Slideshow: Tabletop Computers *Continued*
By Meredith Ringel Morris
**First Published December 2008**

Email | Print | Comments (1) | Reprints | Newsletters

Del.icio.us | Digg | Slashdot

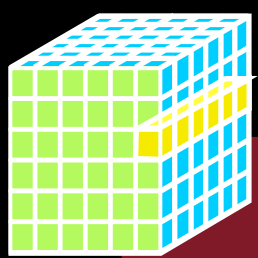
PHOTO: HAO JIANG, DANIEL WIGDOR, CLIFTON FORLINES, AND CHIA SHEN

**UBITABLE:** Users can interact with surface computers through auxiliary devices, such as laptops, phones, and PDAs. The display on the auxiliary device can convey private or sensitive content to a single user, while group-appropriate content can appear on the tabletop display. Chia Shen and her colleagues at Mitsubishi Electric Research Laboratories, in Cambridge, Mass., have explored auxiliary interactions with surface computers in their UbiTable project, in which two people with laptops collaborate over a tabletop display. Recently, Shen expanded the UbiTable into an interactive room called the WeSpace. People can share data on their laptops with other people in the room, using both a table and a large display wall. Here, three Harvard University astrophysicists discuss radio and IR spectrum images using the WeSpace.

*Linked Views*

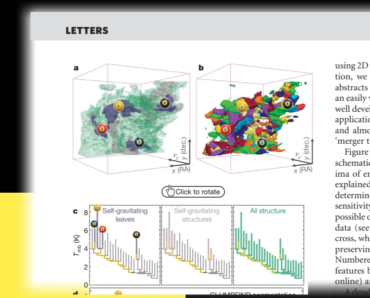
*Seamless Astronomy*
*[the future]*

*movie courtesy Daniel Wigdor, taken at MERL, Kendall Square, Cambridge, 2007*



Data Dimensions Display

Linked Views




"Taste-Testing"
[Sim:Data]

Seamless Astronomy
[the future]

ADS Labs
[info viz]

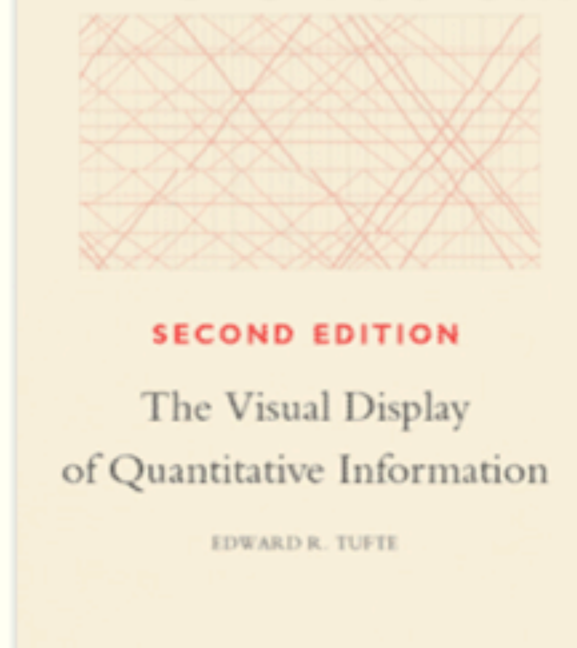
Open Data
Open Tools

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION

The classic book on statistical graphics, charts, tables. Theory and practice in the design of data graphics, 250 illustrations of the best (and a few of the worst) statistical graphics, with detailed analysis of how to display data for precise, effective, quick analysis. Design of the high-resolution displays, small multiples. Editing and improving graphics. The data-ink ratio. Time-series, relational graphics, data maps, multivariate designs. Detection of graphical deception: design variation vs. data variation. Sources of deception. Aesthetics and data graphical displays.

This is the second edition of *The Visual Display of Quantitative Information*. Recently published, this new edition provides excellent color reproductions of the many graphics of William Playfair, adds color to other images, and includes all the changes and corrections accumulated during 17 printings of the first edition.

The classic...




*Edward Tufte 1983, The Visual Display of Quantitative Information*

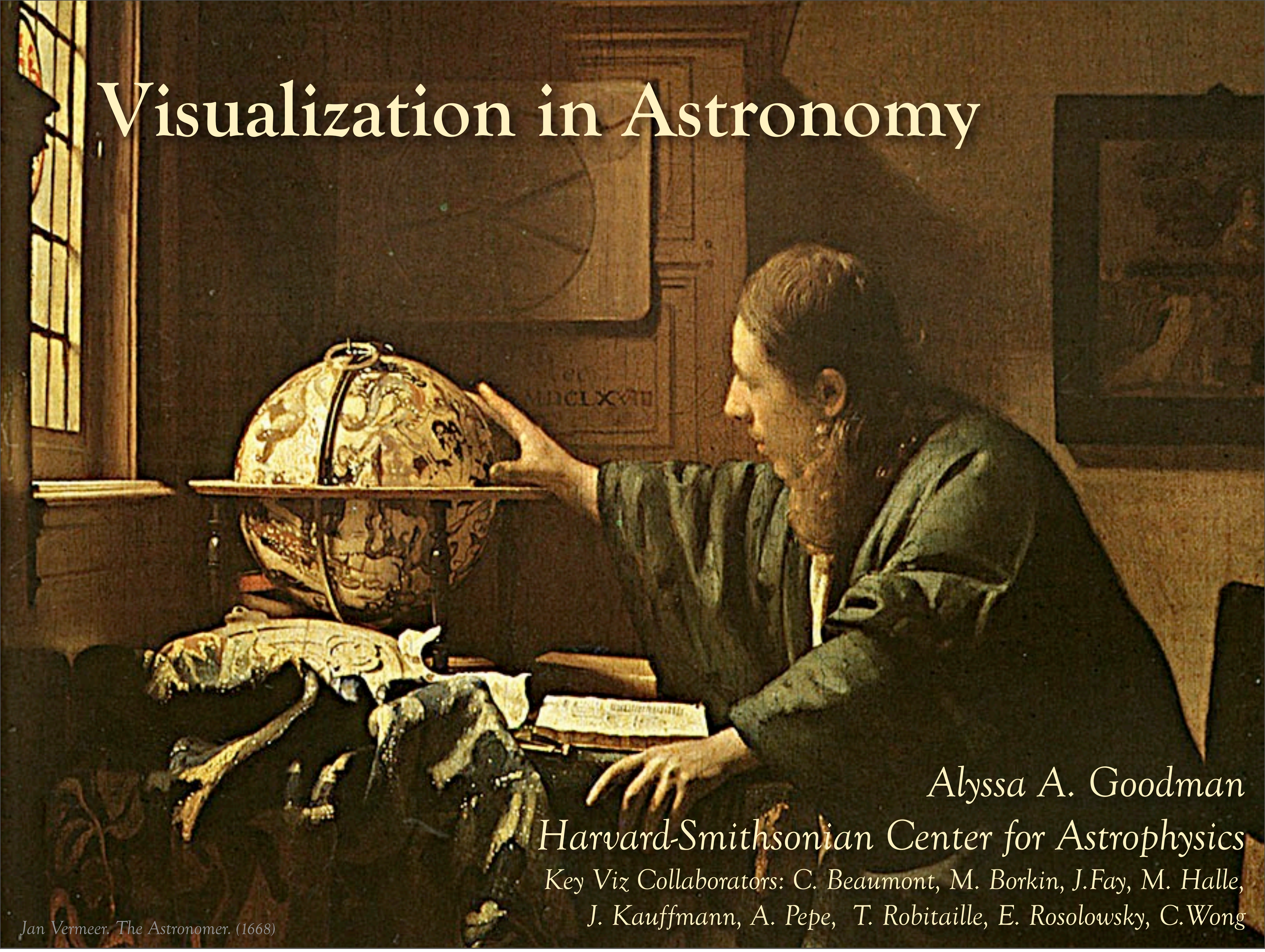
Coming soon...


*Frankel & DePace 2011, Visual Strategies*

*Online **slides** & more...*
*[www.cfa.harvard.edu/~agoodman](www.cfa.harvard.edu/~agoodman)*

***Software**: projects.iq.harvard.edu/seamlessastronomy/software*

*"Everyday" software **blog**...*
*alyssagoodman.tumblr.com*




Tuesday, September 27, 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.

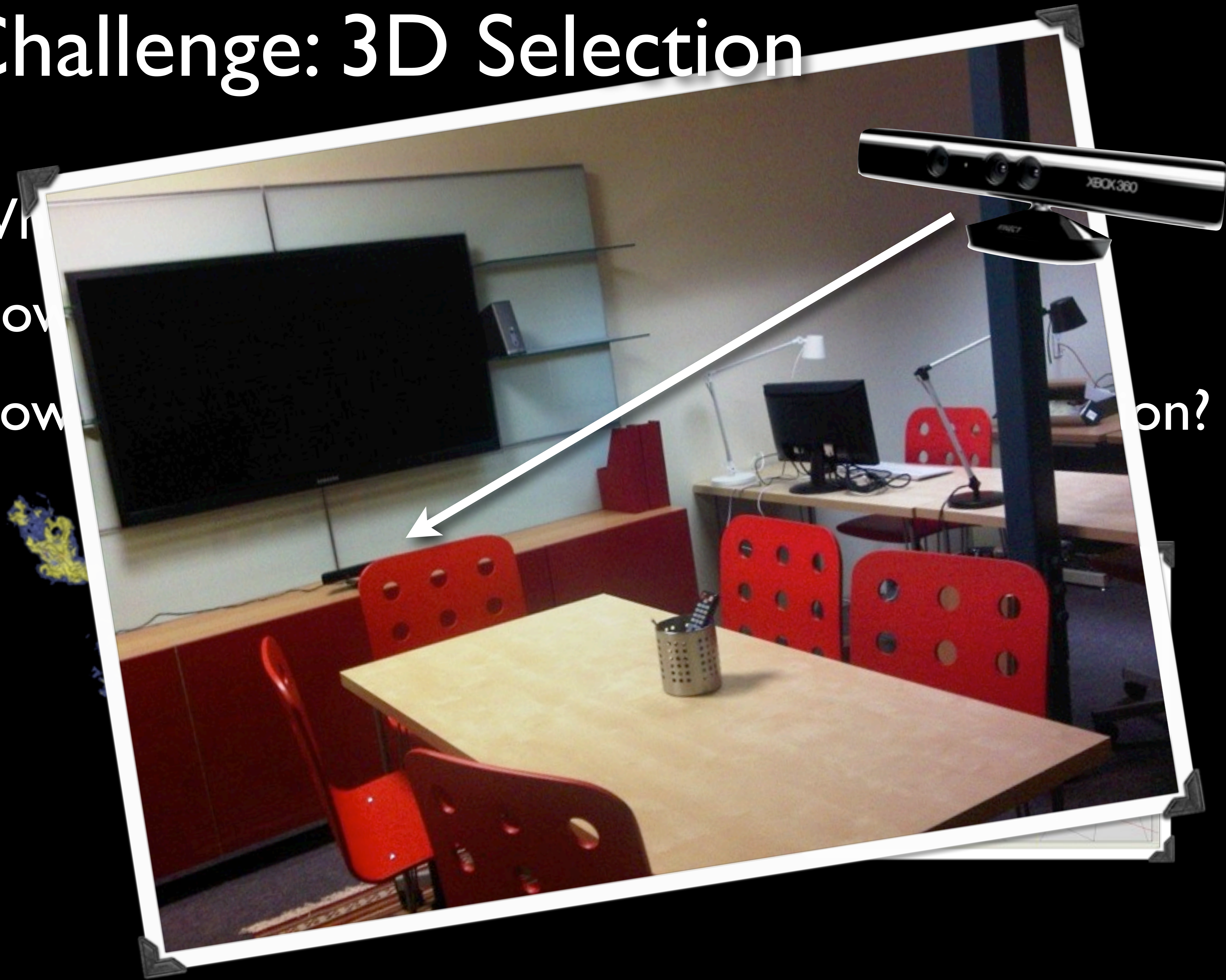
Even though this abstract may sound technical to non-star-formation or non-computational researchers, my goal will be to keep the talk accessible to non-experts, so people from other fields faced with high-dimensional data and visualization challenges should feel free to join in--and to ask questions

# Challenge: 3D Selection

Wh

How

How ... on?



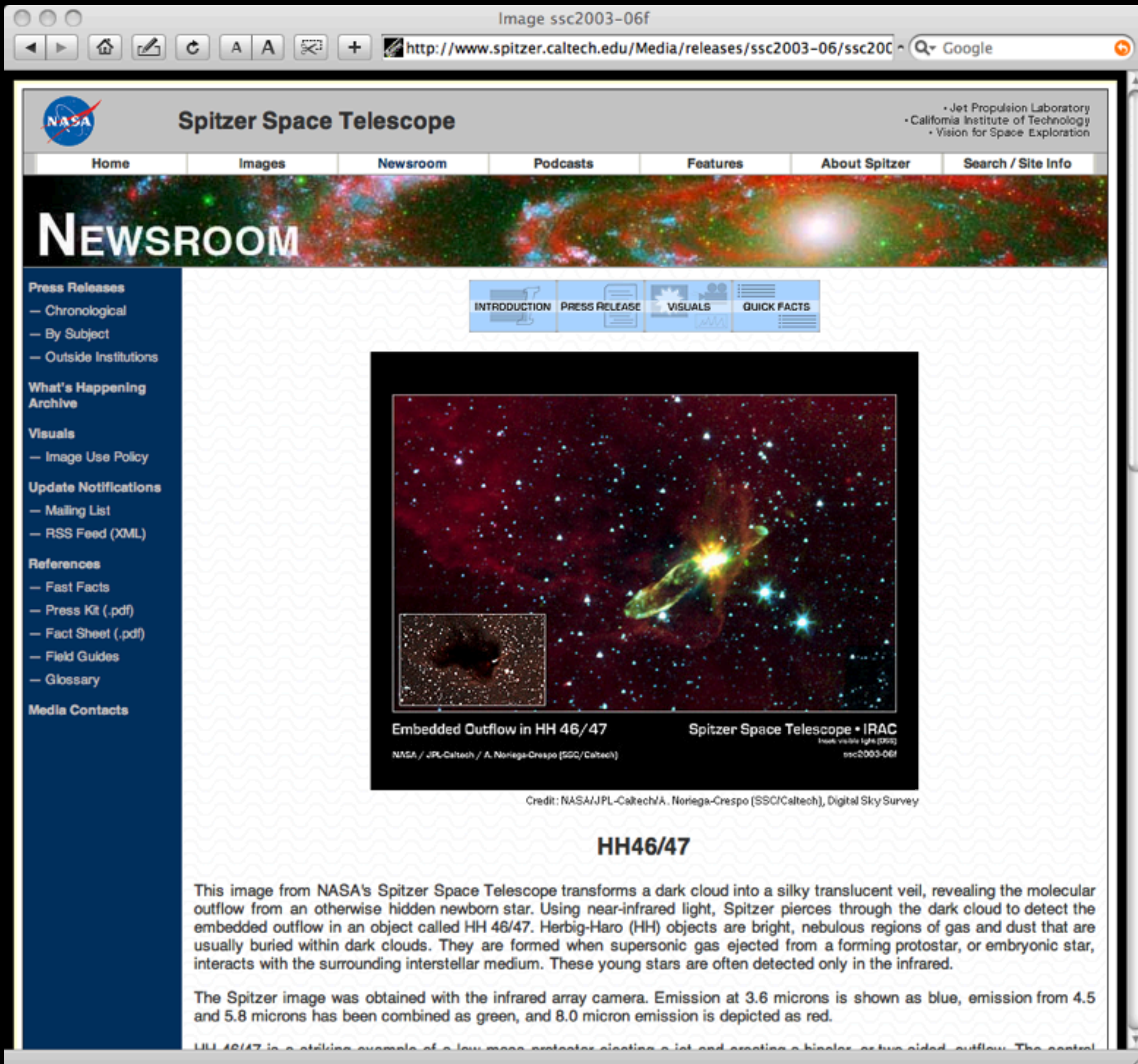


Image ssc2003-06f

http://www.spitzer.caltech.edu/Media/releases/ssc2003-06/ssc200

Spitzer Space Telescope

• Jet Propulsion Laboratory
• California Institute of Technology
• Vision for Space Exploration

Home | Images | Newsroom | Podcasts | Features | About Spitzer | Search / Site Info

NEWSROOM

Press Releases
— Chronological
— By Subject
— Outside Institutions

What's Happening Archive

Visuals
— Image Use Policy

Update Notifications
— Mailing List
— RSS Feed (XML)

References
— Fast Facts
— Press Kit (.pdf)
— Fact Sheet (.pdf)
— Field Guides
— Glossary

Media Contacts

INTRODUCTION | PRESS RELEASE | VISUALS | QUICK FACTS

Embedded Outflow in HH 46/47 — Spitzer Space Telescope • IRAC

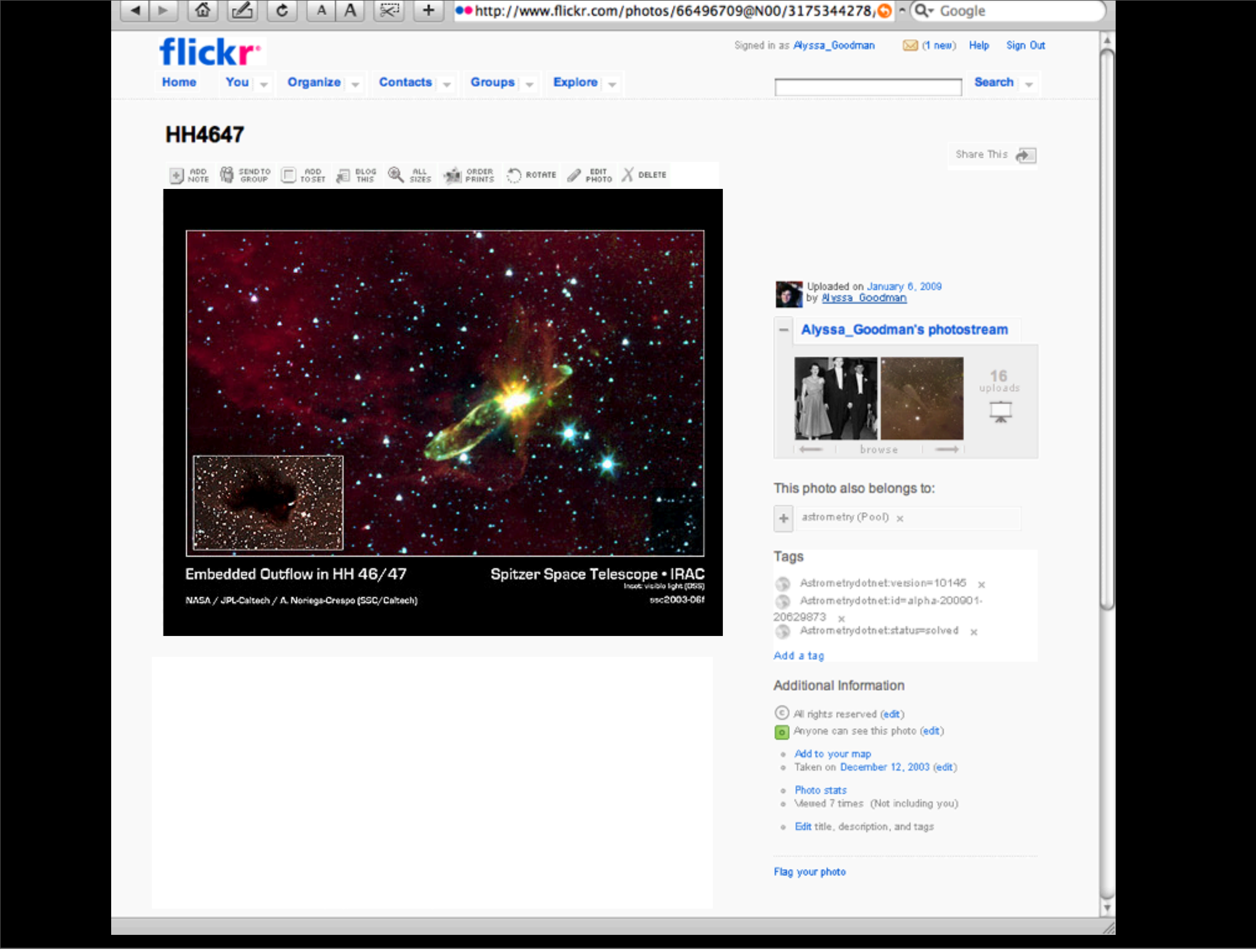
NASA / JPL-Caltech / A. Noriega-Crespo (SSC/Caltech) — ssc2003-06f

Credit: NASA/JPL-Caltech/A. Noriega-Crespo (SSC/Caltech), Digital Sky Survey

## HH46/47

This image from NASA's Spitzer Space Telescope transforms a dark cloud into a silky translucent veil, revealing the molecular outflow from an otherwise hidden newborn star. Using near-infrared light, Spitzer pierces through the dark cloud to detect the embedded outflow in an object called HH 46/47. Herbig-Haro (HH) objects are bright, nebulous regions of gas and dust that are usually buried within dark clouds. They are formed when supersonic gas ejected from a forming protostar, or embryonic star, interacts with the surrounding interstellar medium. These young stars are often detected only in the infrared.

The Spitzer image was obtained with the infrared array camera. Emission at 3.6 microns is shown as blue, emission from 4.5 and 5.8 microns has been combined as green, and 8.0 micron emission is depicted as red.

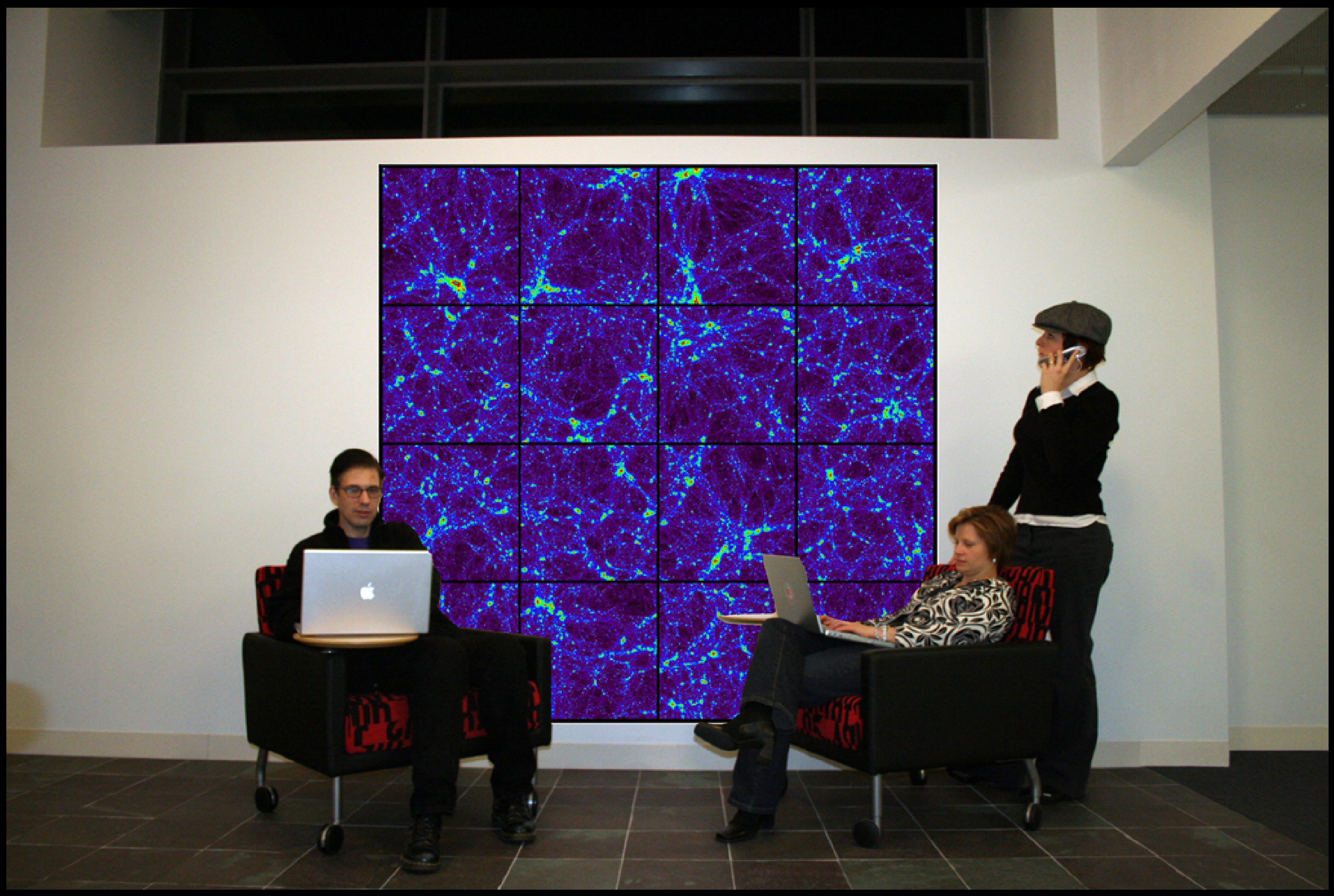


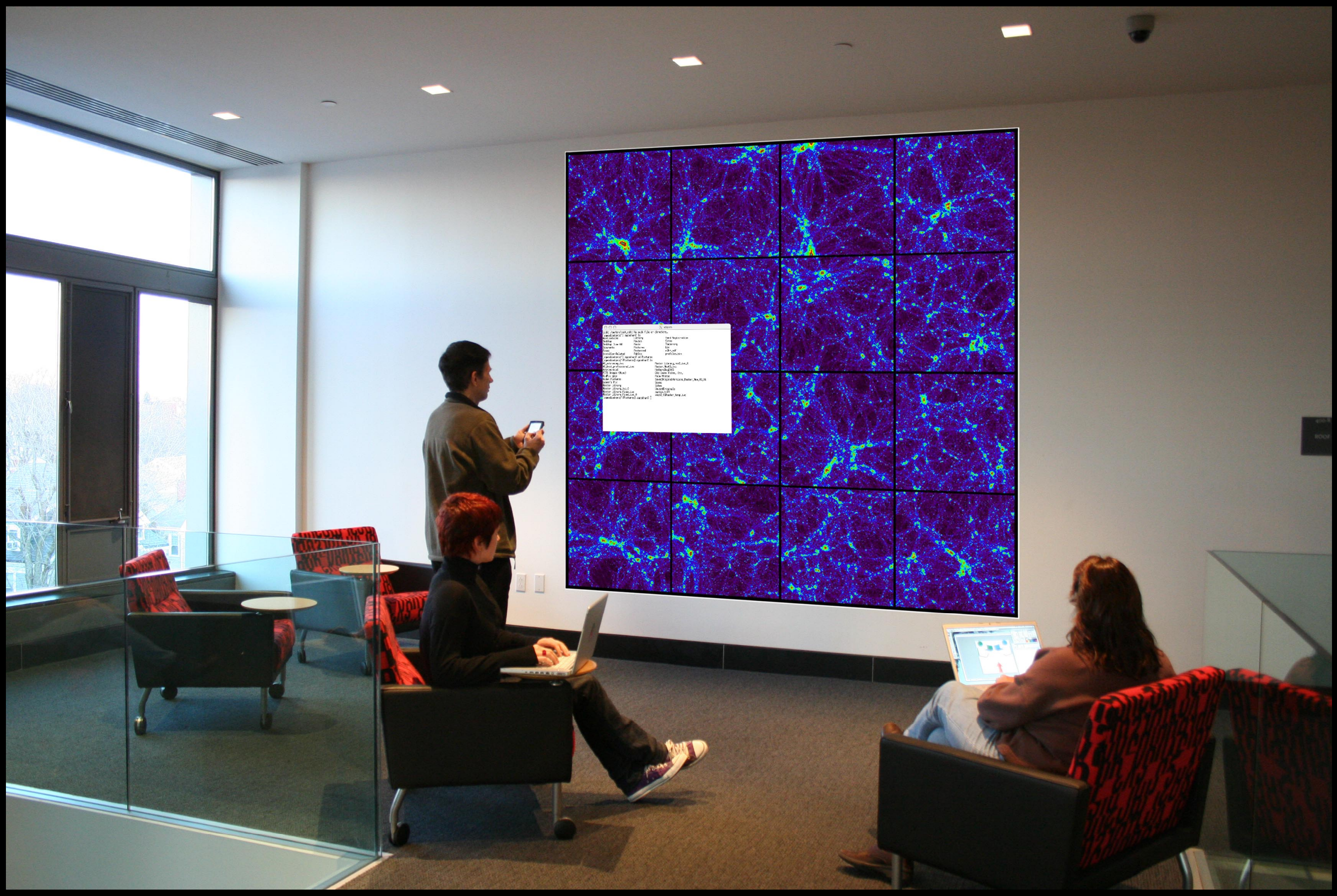

Tuesday, September 27, 2011

display modes

...why we must explain that...

*"This is not art."*





Tuesday, September 27, 2011