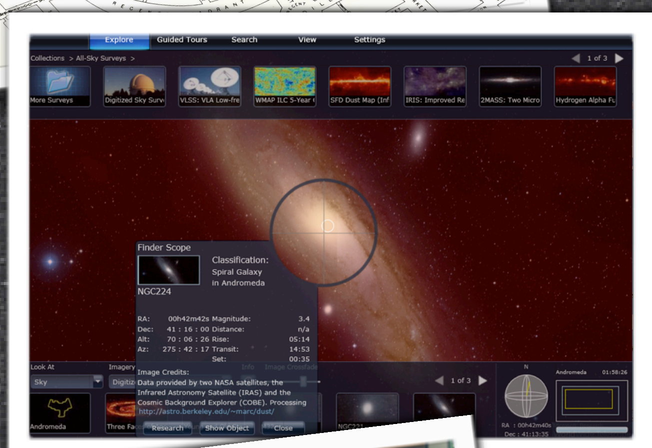
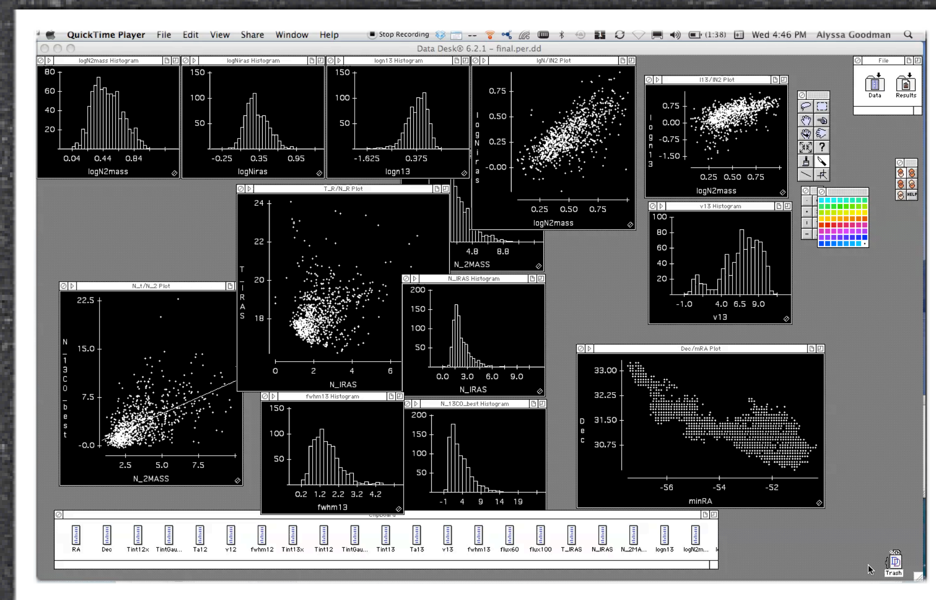
WorldWide Telescope

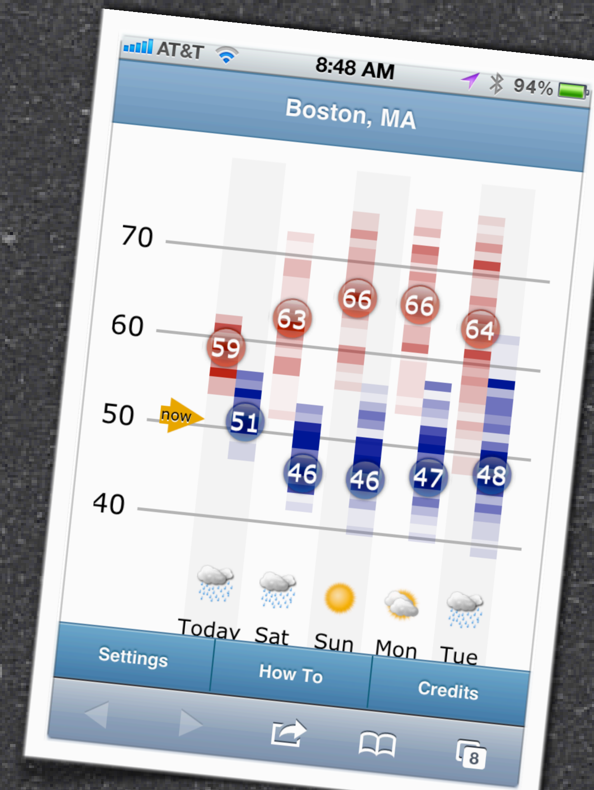
Microsoft Research

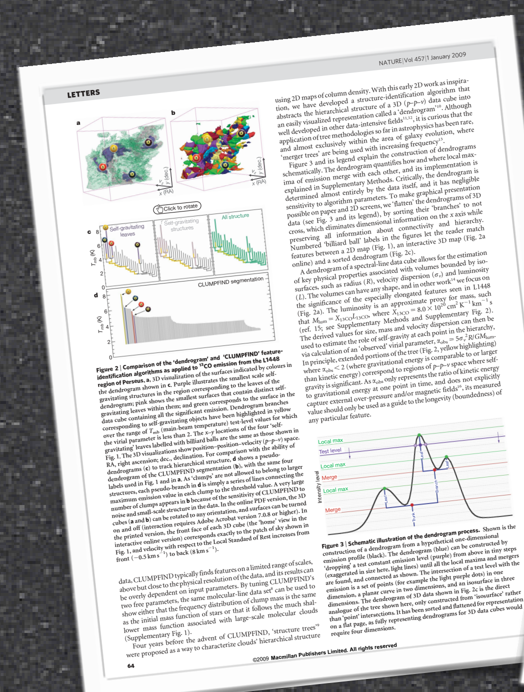
Overcoming Narrow-Mindedness in Modern Science

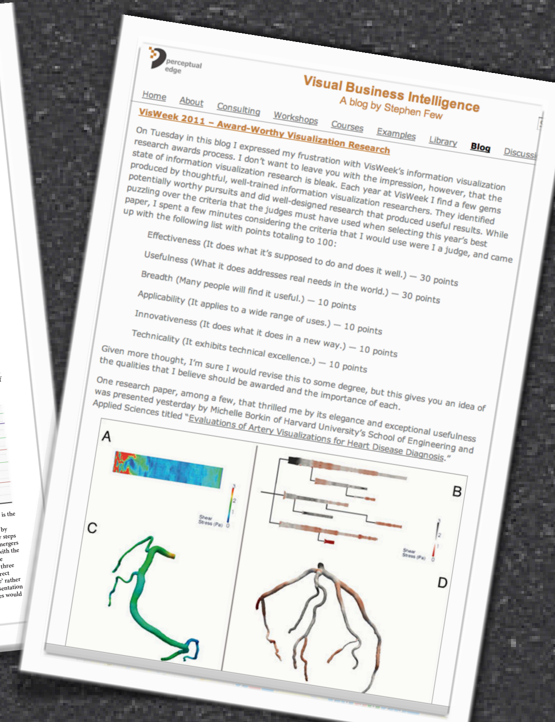
Why Contextual Data Visualization is Becoming so Essential

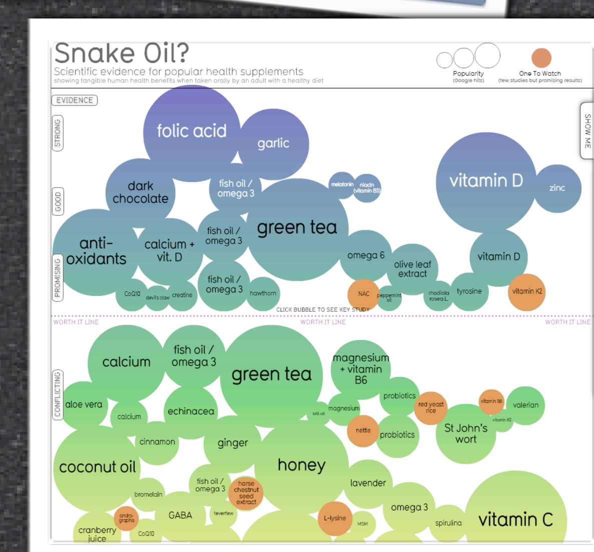












Alyssa A. Goodman, Professor of Astronomy, Harvard-Smithsonian Center for Astrophysics

3500 years of Observing

Stonehenge, 1500 BC

Ptolemy in Alexandria, 100 AD

Observatory Tower, Lincolnshire, UK, c. 1300

Galileo, 1600

The "Scientific Revolution"

Reber's Radio Telescope, 1937

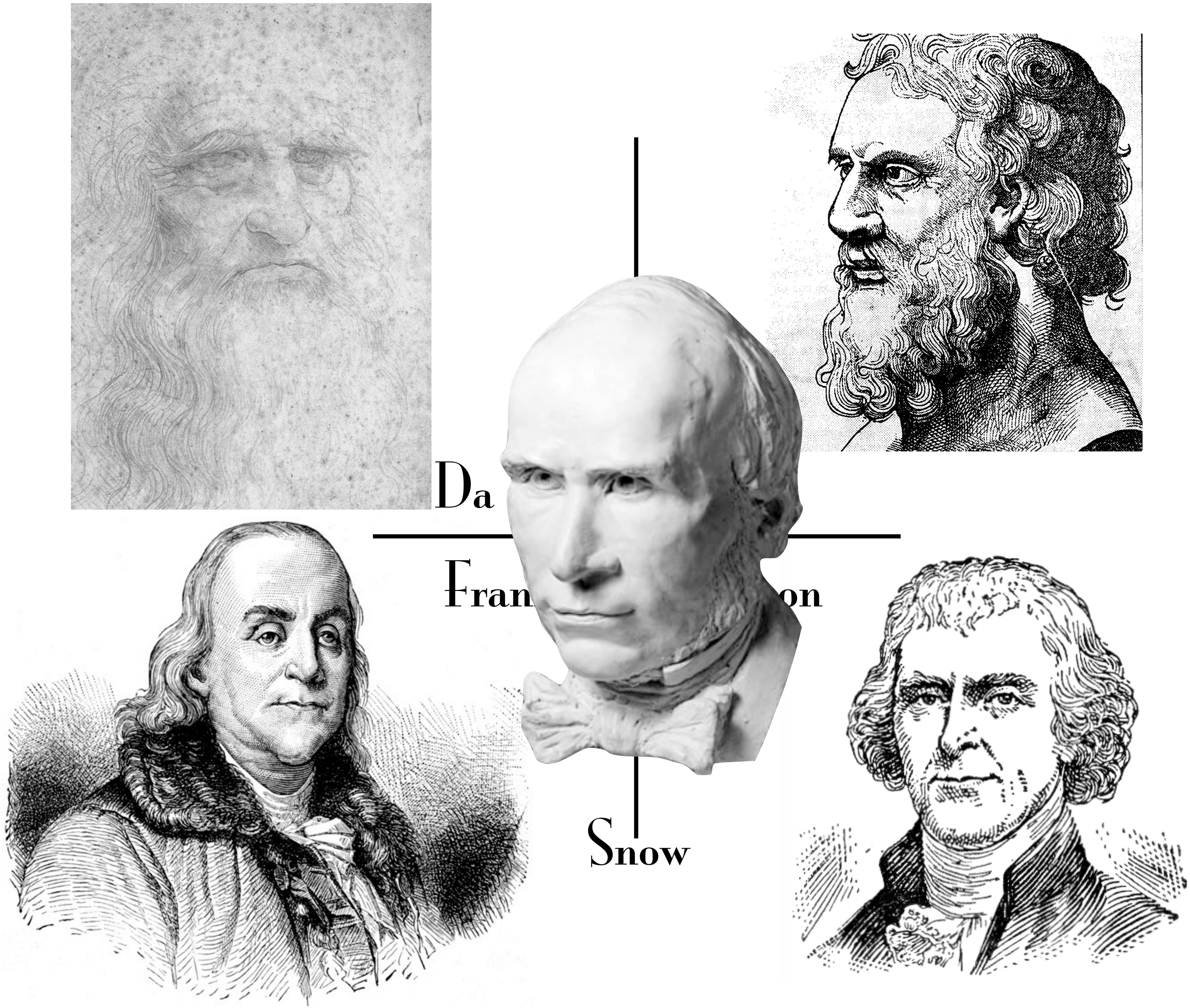
NASA/Explorer 7 (Space-based Observing) 1959

"The Internet"

Long-distance remote-control/ "robotic" telescopes 1990s

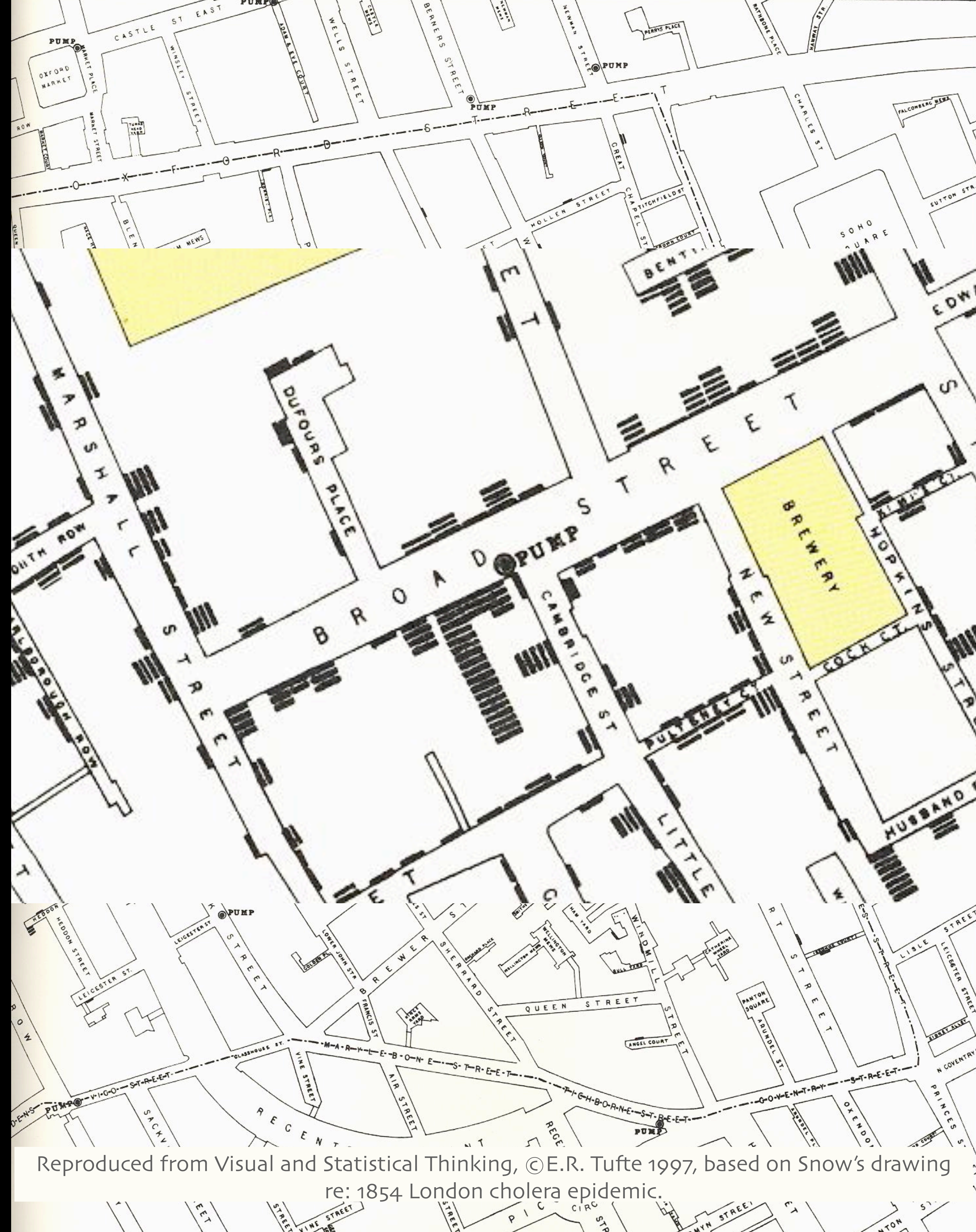
NVO IVOA gSky

"Virtual Observatories" 21st century

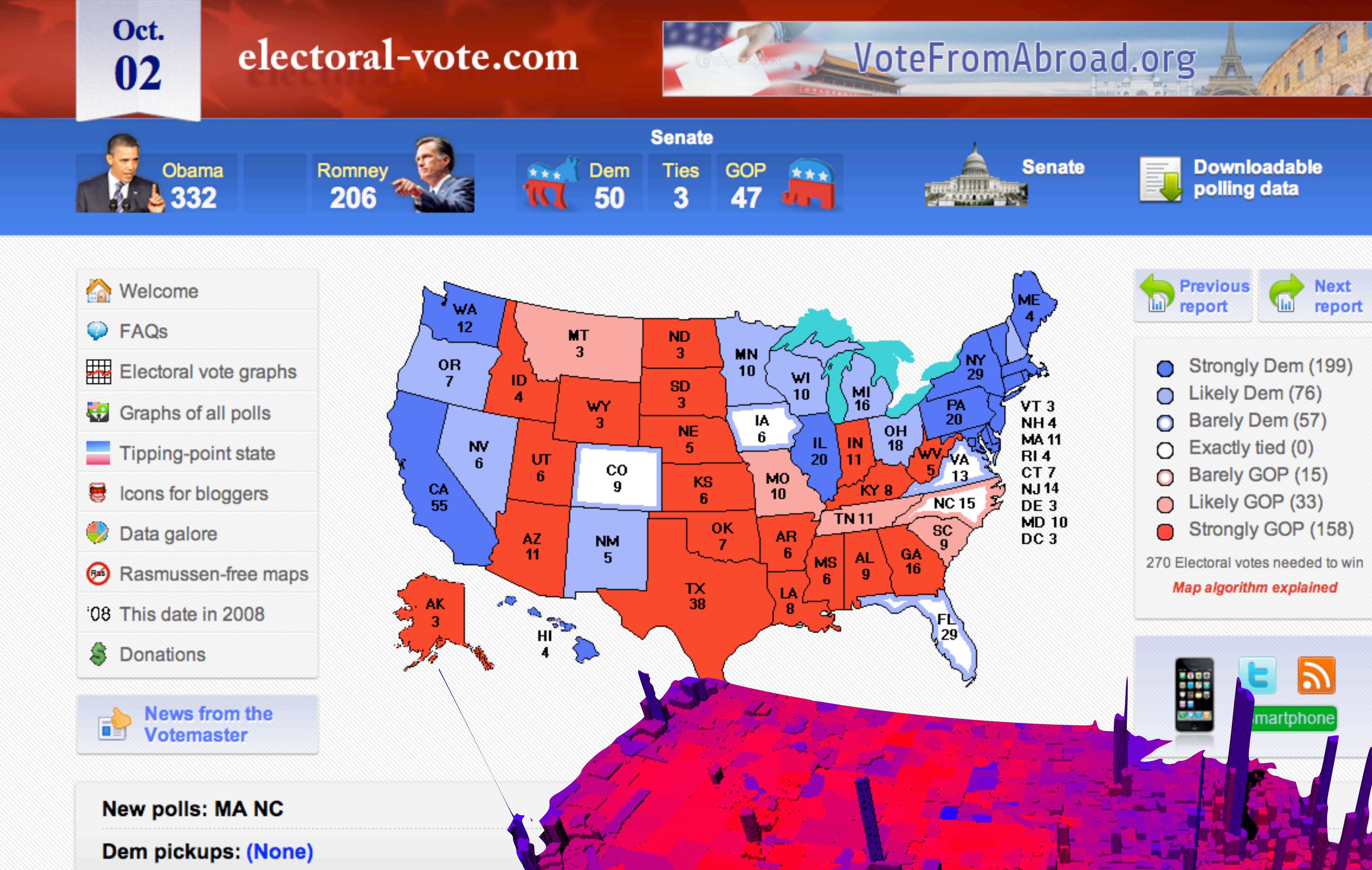

Da
Fran
on
Snow

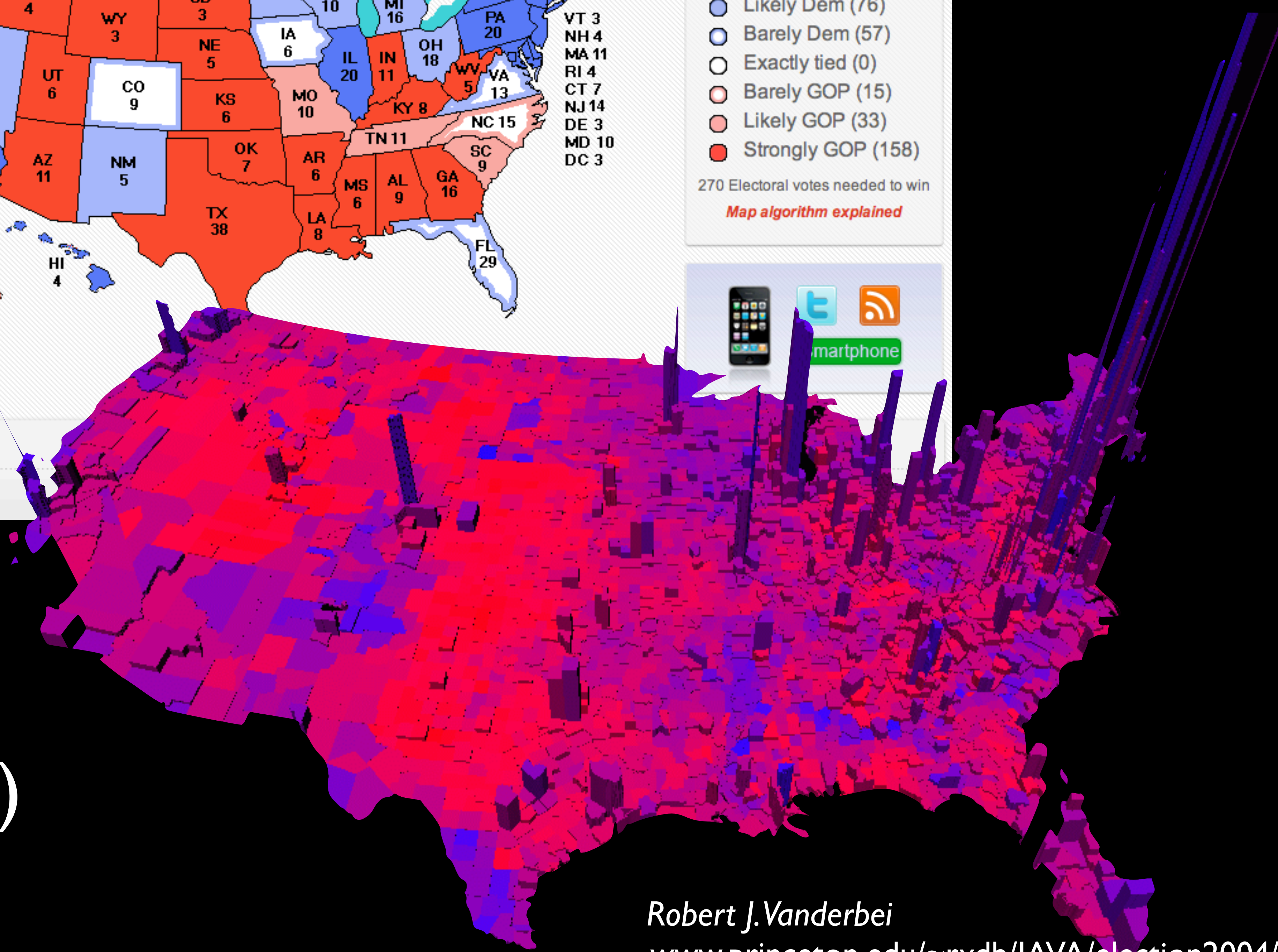
Context

(Snow's 1854 "GIS")



Reproduced from Visual and Statistical Thinking, ©E.R. Tufte 1997, based on Snow's drawing re: 1854 London cholera epidemic.





Context

(“tonight’s GIS”)

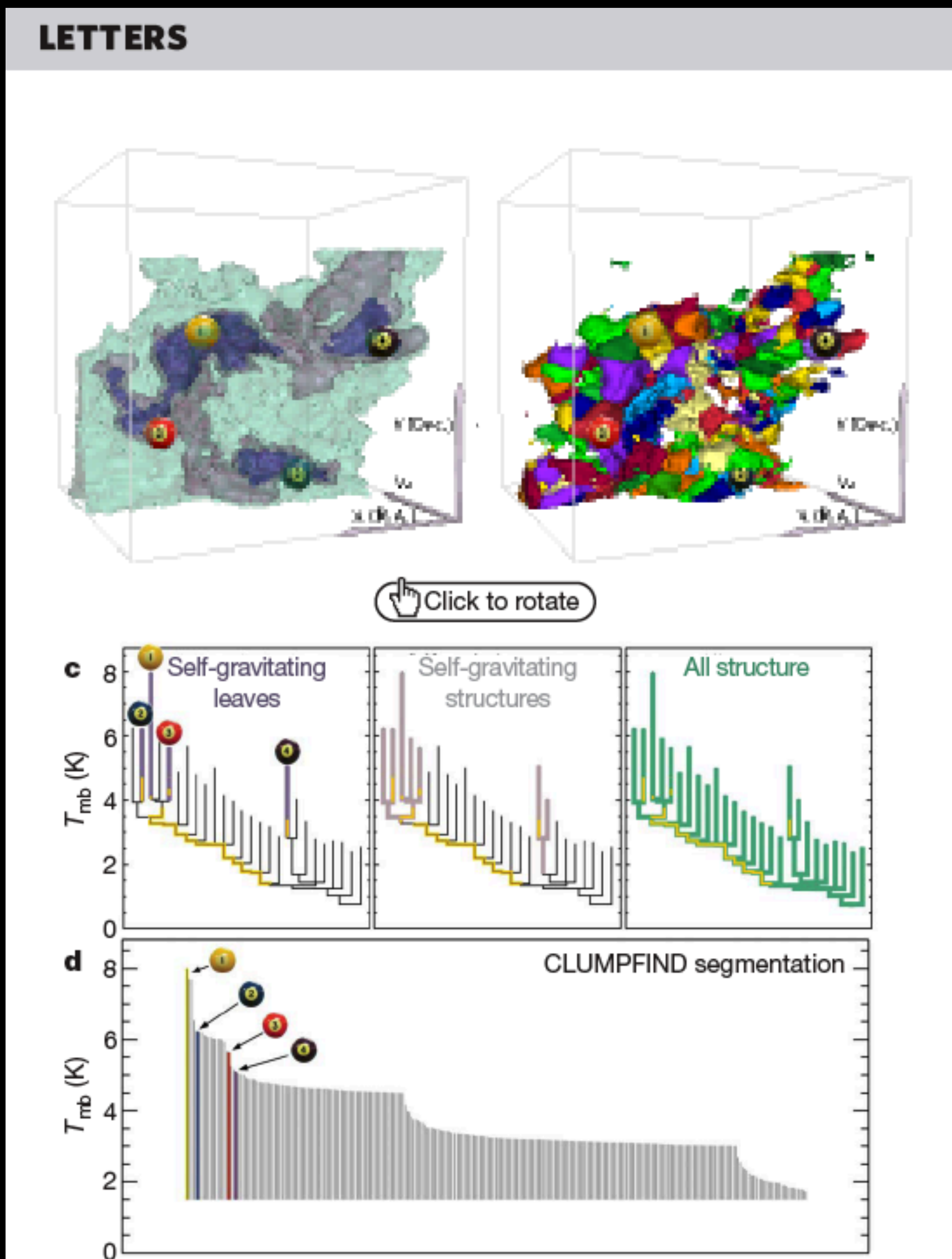
Robert J. Vanderbei
www.princeton.edu/~rvdb/JAVA/election2004/

The context of context—

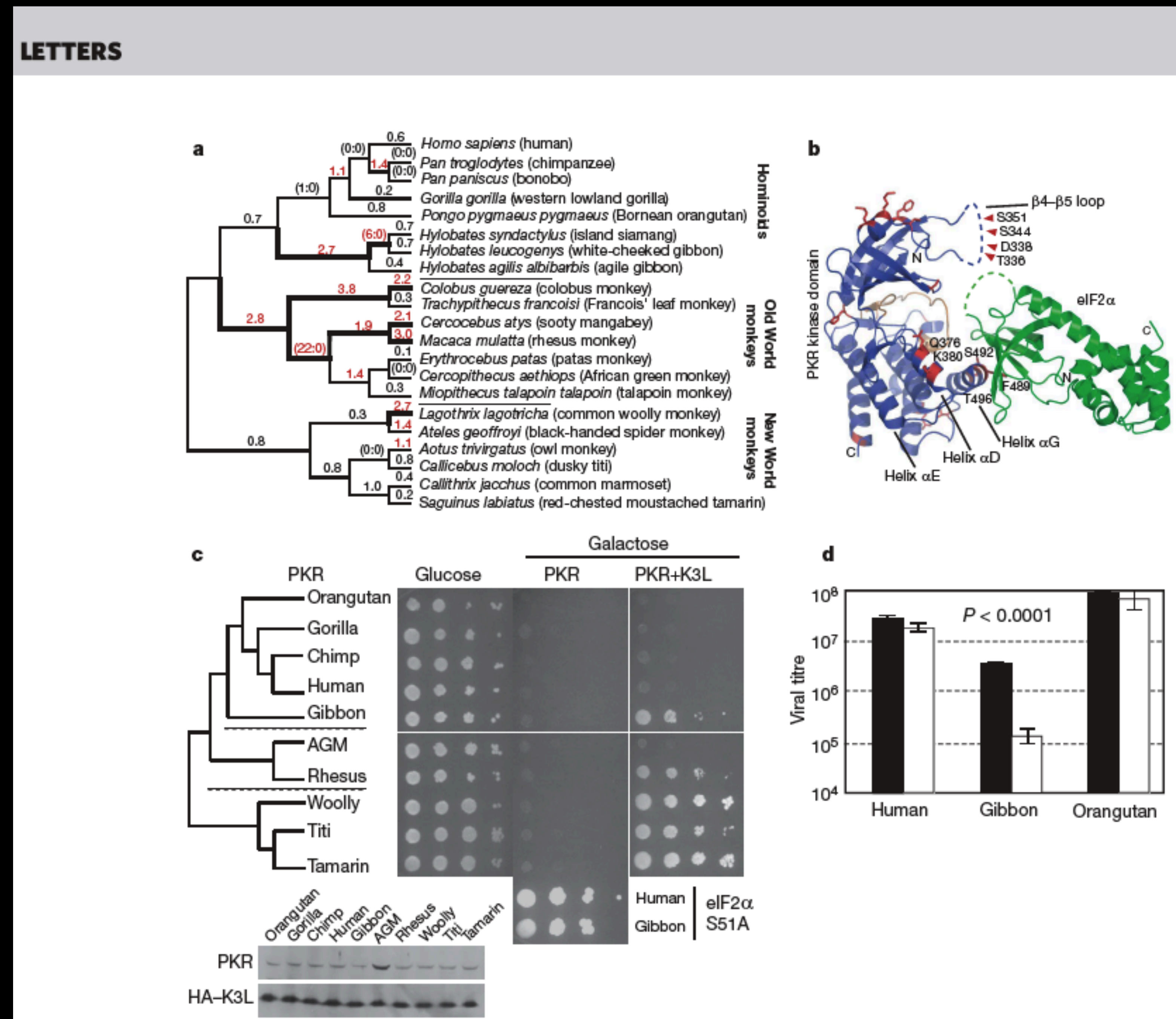
DATA • DIMENSIONS • DISPLAY

“High-dimensional” or “Multivariate” Data

(Astronomy=Biology)



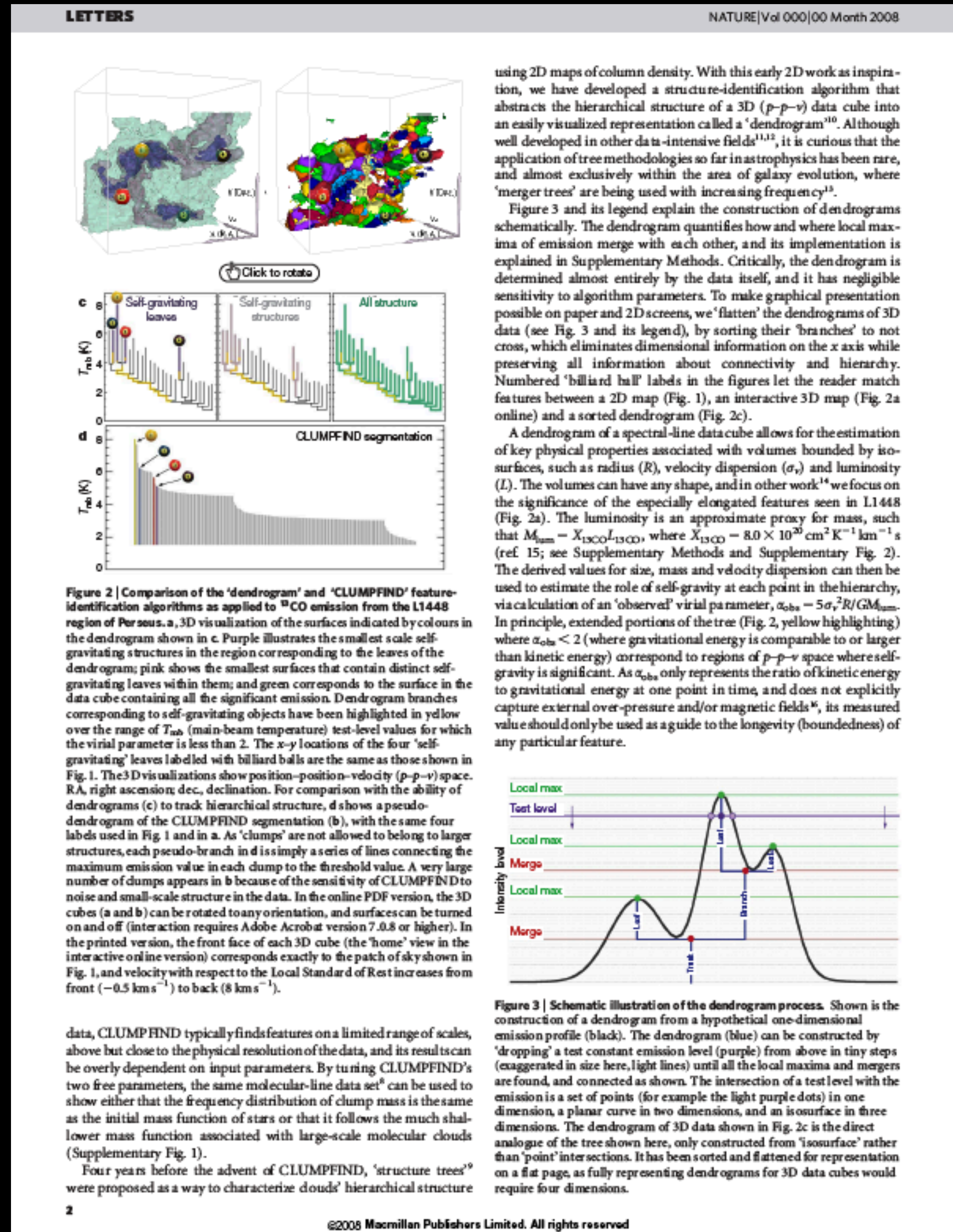
Goodman et al. *Nature*, 2009



Elde et al. *Nature*, 2008

DATA • DIMENSIONS • DISPLAY

3D PDF


LETTERS

NATURE|Vol 000|00 Month 2008

Click to rotate

c Self-gravitating leaves | Self-gravitating structures | All structure

d CLUMPFIND segmentation

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.

Local max
Test level
Local max
Merge
Local max
Merge
Intensity level

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.

2

©2008 Macmillan Publishers Limited. All rights reserved

Relative Strengths

Pattern Recognition

Creativity

Calculations

CIRCOS – an information aesthetic for comparative genomics

M Krzywinski, J Schein, I Birol, S Jones, M Marra

Canada's Michael Smith Genome Sciences Centre // British Columbia Cancer Research Centre // *www.bcgsc.ca*

circos - round is the new square

We have created a visualization tool, called Circos, to facilitate exploring relationships between genomes and in general any kind of position *n*-tuples that relate genomic intervals. Structural variation data such as these, produced by sequence alignment and hybridization arrays, underpin comparative studies but are opaque to conventional visualization methods designed for 2D data sets. Compared to other tools [1,2], Circos is unique in its combination of **circular data domain layout**, support for a large number of **diverse data tracks, global and local length scale control**, extensive **customization and automation**, and maintaining a **high data-to-ink ratio** [3] without sacrificing clarity of presentation. Circos has been used within the genomics community [4-6] and its flexibility and aesthetic has garnered interest from mainstream periodicals and newspapers [7-9] and, recently, illustrate the dynamics of a US presidential debate [10].

DOWNLOAD CIRCOS AT http://mkweb.bcgsc.ca/circos

At present, laboratories are hard-pressed not only to store and analyze, but to visualize the reams of data produced by ultra-high-throughput technologies, such as massively parallel sequencing. Because analytically extracting informative patterns from these large data sets is very difficult, automated visualization tools that generate informative vignettes of the data are valuable in data mining and formulating hypotheses.

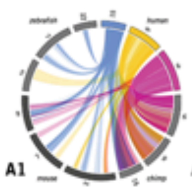
The design of Circos is based on the fact that a circularly composited ideogram layout can encode relationships between genomic regions more informatively than a linear layout. These relationships are visually encoded by links which can be either straight lines or Bezier curves whose control point location can be highly customizable. Other data types that are supported are **scatter plots, line plots, histograms, tile plots, heatmaps, text and glyph labels, highlights, ideograms**, and **labeled ticks**. The radial position of data tracks is controlled by the user and their angular extent is a function of the extent of the data domain. Data tracks such as the tile and text label have their individual elements automatically positioned to avoid overlap.

Circos uses **plain text files** for both **input data and configuration**. The latter controls the placement and format of each data track. The ability to generate both data and configuration files automatically makes Circos highly amenable to incorporation in web-based database mining and visualization.

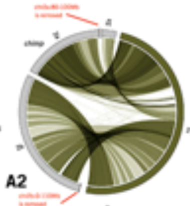
A feature unique to Circos is not only the ability to adjust the length scale for each ideogram (e.g. display chr 17 at 3x normal size), but to **smoothly vary the length scale** locally, effectively zooming (or contracting) regions of interest while still displaying the entire data domain (Figures A1-A3, B1). This global and local scale adjustment is useful when illustrating genomic regions in which data density is highly variable. Furthermore, to help draw attention to important data, the ideograms can be divided into any number of disjoint regions, which in turn can be drawn in any order. The resulting **axis breaks** can be marked up in various styles on the final image to clearly mark the disruption.

Every aspect of the final image is customizable and output can be generated in either **bitmap or SVG** format. For example, the thickness, outline and color of the ideogram track is customizable, as are the corresponding features of the cytogenetic bands. The radial position for each ideogram can be independently set. Each data track, and individual primitives within a track, has an associated z-depth value, which controls how elements stack. Finally, every data type format characteristic, such as color, thickness, data value, label and visibility, can be adjusted by **dynamic formatting rules** based on data position and other format values at run-time (Figures C1-C3). These rule sets are stored in the configuration files and separate the definition and storage of formatting rules from the raw data.

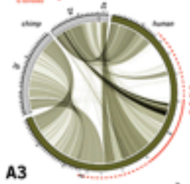
1. Ekdahl, S. & Sonnhammer, E.L. Bioinformatics 20, 576-7 (2004).
2. Kozik, A., Kochetkova, E. & Michelmore, R. Bioinformatics 18, 335-6 (2002).
3. Tufte, E. Visual Display of Quantitative Information (Graphics Press, 1992).
4. Jaillon, O. et al. Nature 449, 463-7 (2007).
5. Aury, J.M. et al. Nature 444, 171-8 (2006).
6. Campbell, P.J. et al. Nat Genet (2008).
7. Duncan, D.E. Conde Nast Portfolio (November 2007) http://www.portfolio.com/news-markets/national-news/portfolio/2007/10/15/23andMe-Web-Site
8. Ostrander, E.A. American Scientist 95, 406-413 (2007).
9. Constantine, New York Times (23 January 2007) http://www.nytimes.com/imagepages/2007/01/22/science/20070123_SCI_ILLO.html
10. Corum, J. & Hossain, F. New York Times (15 December 2007) http://www.nytimes.com/interactive/2007/12/15/us/politics/DEBATE.html



A1



A2



A3

GLOBAL AND LOCAL SCALE ADJUSTMENTS

Ideogram position, order and direction is highly customizable. In Figure A1 a subset of the ideograms shown in the main figure is displayed.

Ideograms can be **partitioned into regions**, which can be shown (or supressed) in any order. This is illustrated in figure A2 where chromosome 2a is shown in two pieces and the leading region of 2b is removed.

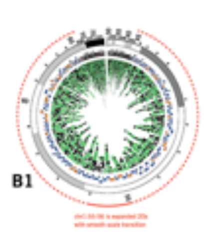
The **scale of any ideogram can be globally adjusted**. In the main figure the ideograms are scaled by the corresponding species' genome size, in order for each genome to occupy the same fraction of the image.

In addition to global scale adjustment, the **scale can be adjusted locally**. Any number of regions of an ideogram can be expanded or contracted to create a **continuously varying zoom factor** along the length of the ideogram. In Figure A3 a region of ideogram 2 is expanded to reveal details in the link position. Length scale directives of overlapping regions combine automatically, with the consensus adjustment taken as the largest scale factor (zoom or constriction) in the region.

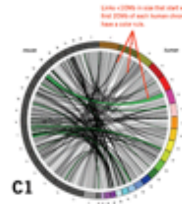
TILE TRACKS

Tiles are spans on ideograms that are drawn in layers to avoid overlap. Tiles are ideal for showing **clones, genes, transcripts**, and any regions which represent **coverage, sampling or structure**. Tile location is optimized automatically, based on the number of layers, padding and orientation of stacking.

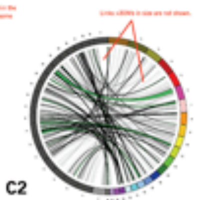
The tiles in this figure represent the **golden path** (outer track) and **end sequenced clones** (inner track). Color is used to indicate source library (e.g. RP11 green, CTD black). Figure B1 also incorporates a local scale adjustment (see Figure A for more details about scale).



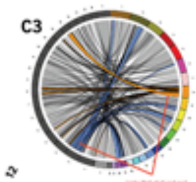
B1



C1



C2



C3

DYNAMIC DATA FORMAT RULES

To highlight patterns in data, **dynamic format rules are used to adjust properties** of data points (e.g. color or visibility). Property adjustments associated with a rule are applied when the condition criteria for the rule pass. The conditions can be based on the position, size, value and even other format characteristics of a data point.

In Figure C1, all links are initially grey. Links that begin within the first 20Mb of each human chromosome are dynamically (i.e., at run-time) colored green. In Figure C2, an additional rule suppresses the display of links whose extent is greater than 30Mb.

In Figure C3, links that begin on human chromosome 8 are orange, and those that end on mouse chromosome 2 are blue.

Rules can be chained and the chain can be **evaluated multiple times**. Thus, one rule can adjust a parameter which is then modified further by subsequent rules. **Rules can adjust any property of a data point**, including format (e.g. color, thickness, transparency) as well as the data point's value, position, or label.

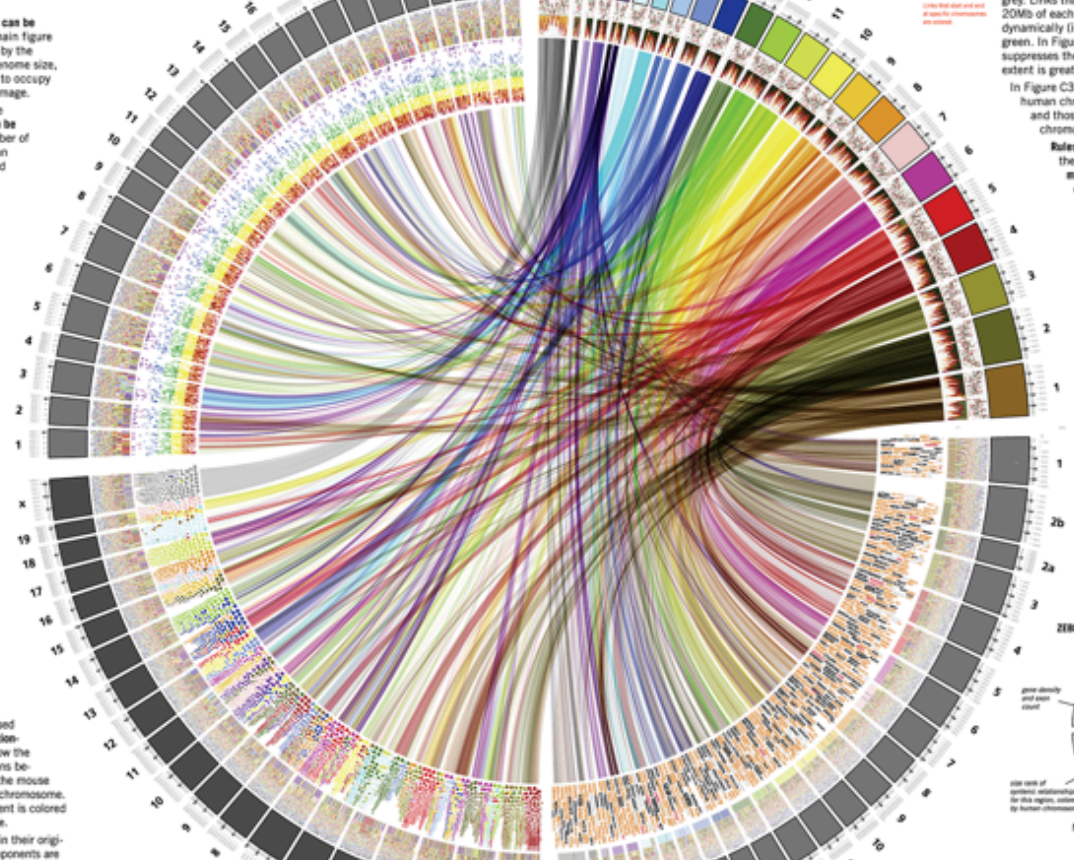
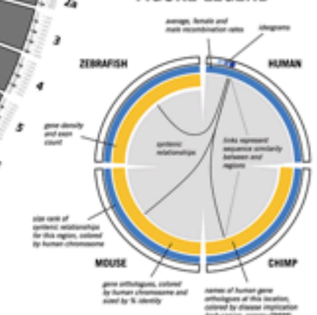


FIGURE LEGEND



2D VIGNETTE: THE STACKED BAR PLOT

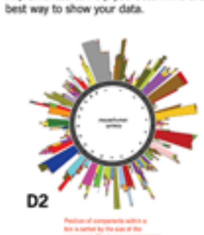
Figures D1-D4 show how **stacked bar plots** can be used to **summarize syntenic relationships**. The figures here show the total size of syntenic regions between a given window on the mouse genome and each human chromosome. Each stacked bar component is colored by the human chromosome.

Figure D1 shows the data in their original order, where bin commponents are ordered by chromosome (1, 2, 3, ...). In Figure D2, components within a bin are reordered by their size. Thus, for any bin, the largest contributing stacked element is shown first.

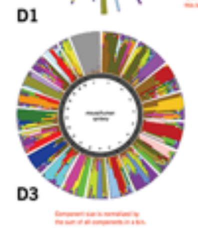
Figures D3 and D4 show the same data set, but normalized by the sum of each bin's stacked elements.

Although Circos does not perform any data analysis, features such as **variable stack order** and **stack normalization** are implemented to help you determine the best way to show your data.

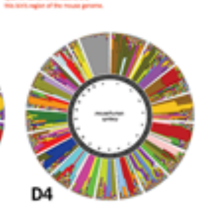
D1



D2

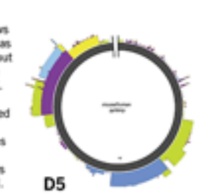


D3

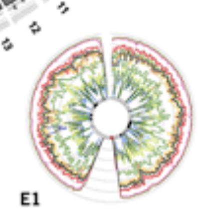


D4

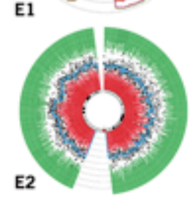
Figure D5 shows the same data as in Figure D2, but only for mouse chromosome 7. Circos can be easily automated and creating multiple images with different data domains is straightforward.



D5



E1



E2

DRAWING 2D DATA

Drawing links is a core feature of Circos. However, it is usually not long before you will need to add other data types to your image, such as 2D sets best represented by **scatter**, **line** or **histogram** plots. Circos supports these, as shown in Figures E1 and E2.

Just like for links, **every data point** in a data drack **can be adjusted using dynamic rules**. For example, glyphs in a scater plot can have their size, color and/or visibility adjusted, among other properties.

Histograms can be drawn with an **underfill** and **bins can be enlarged** to fill the entire data domain.

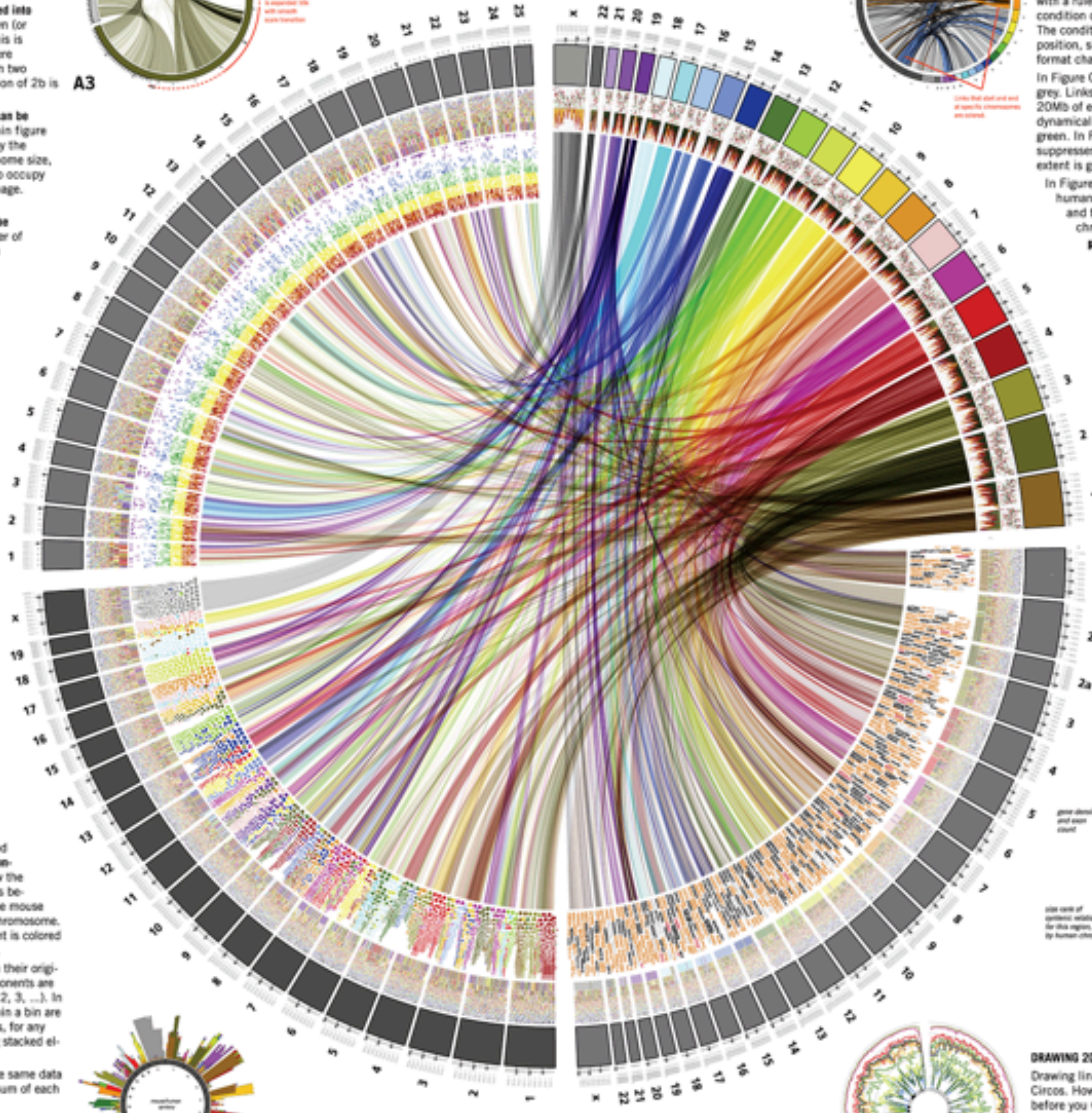
All data plots can be drawn with **axis divisions** and a background color to provide context to the magnitude to the data points and to separate the track from other elements in the image.

Ideograms can be **partitioned into regions**, which can be shown (or supressed) in any order. This is illustrated in figure A2 where chromosome 2a is shown in two pieces and the leading region of 2b is removed.

The **scale of any ideogram can be globally adjusted**. In the main figure the ideograms are scaled by the corresponding species' genome size, in order for each genome to occupy the same fraction of the image.

In addition to global scale adjustment, the **scale can be adjusted locally**. Any number of regions of an ideogram can be expanded or contracted to create a **continuously varying zoom factor** along the length of the ideogram. In Figure A3 a region of ideogram 2 is expanded to reveal details in the link position. Length scale directives of overlapping regions combine automatically, with the consensus adjustment taken as the largest scale factor (zoom or constriction) in the region.

A3



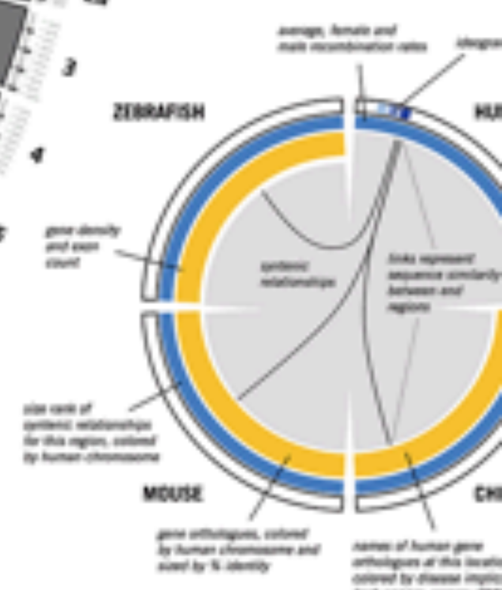
with a rule are applied when the condition criteria for the rule pass. The conditions can be based on the position, size, value and even other format characteristics of a data point.

In Figure C1, all links are initially grey. Links that begin within the first 20Mb of each human chromosome are dynamically (i.e., at run-time) colored green. In Figure C2, an additional rule suppresses the display of links whose extent is greater than 30Mb.

In Figure C3, links that begin on human chromosome 8 are orange, and those that end on mouse chromosome 2 are blue.

Rules can be chained and the chain can be **evaluated multiple times**. Thus, one rule can adjust a parameter which is then modified further by subsequent rules. **Rules can adjust any property of a data point**, including format (e.g. color, thickness, transparency) as well as the data point's value, position, or label.

FIGURE LEGEND



2D VIGNETTE: THE STACKED BAR PLOT

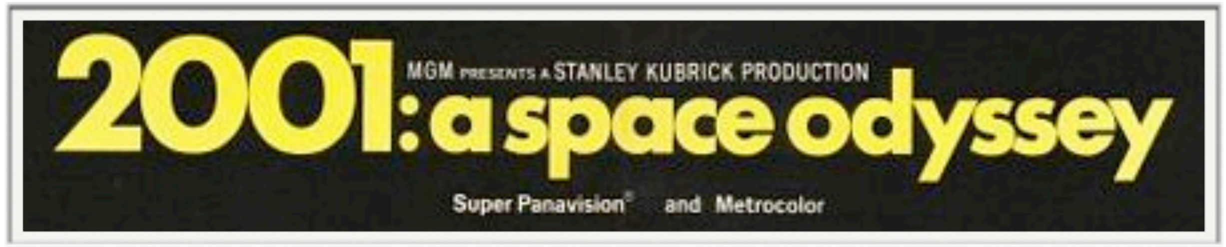
Figures D1-D4 show how **stacked bar plots** can be used to **summarize syntenic relationships**. The figures here show the total size of syntenic regions between a given window on the mouse genome and each human chromosome. Each stacked bar component is colored by the human chromosome.

Figure D1 shows the data in their original order, where bin commponents are ordered by chromosome (1, 2, 3, ...). In Figure D2, components within a bin are reordered by their size. Thus, for any bin, the largest contributing stacked element is shown first.

Figures D3 and D4 show the same data set, but normalized by the sum of each bin's stacked elements.

DRAWING 2D DATA

Drawing links is a core feature of Circos. However, it is usually not long before you will need to add other data



The World-Wide Telescope, an Archetype for Online Science

Jim Gray
Microsoft Research
Gray@Microsoft.com

Alex Szalay
The Johns Hopkins University
Szalay@jhu.edu

Abstract: Most scientific data will never be directly examined by scientists; rather it will be put into online databases where it will be analyzed and summarized by computer programs. Scientists increasingly see their instruments through online scientific archives and analysis tools, rather than examining the raw data. Today this analysis is primarily driven by scientists asking queries, but scientific archives are becoming active databases that self-organize and recognize interesting and anomalous facts as data arrives. In some fields, data from many different archives can be cross-correlated to produce new insights. Astronomy presents an excellent example of these trends; and, federating Astronomy archives presents interesting challenges for computer scientists.

Introduction

Computational Science is a new branch of most disciplines. A thousand years ago, science was primarily *empirical*. Over the last 500 years each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding. Today most disciplines have both empirical and theoretical branches. In the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)

statistics among sets of data points in a metric space. Pair-algorithms on N points scale as N^2. If the data increase a thousand fold, the work and time can grow by a factor of a million. Many clustering algorithms scale even worse. These algorithms are infeasible for terabyte-scale datasets.

The new online science needs new data mining algorithms that use near-linear processing, storage, and bandwidth, and that can be executed in parallel. Unlike current algorithms that give exact answers, these algorithms will likely be heuristic and give approximate answers [Connolly, Szapudi].

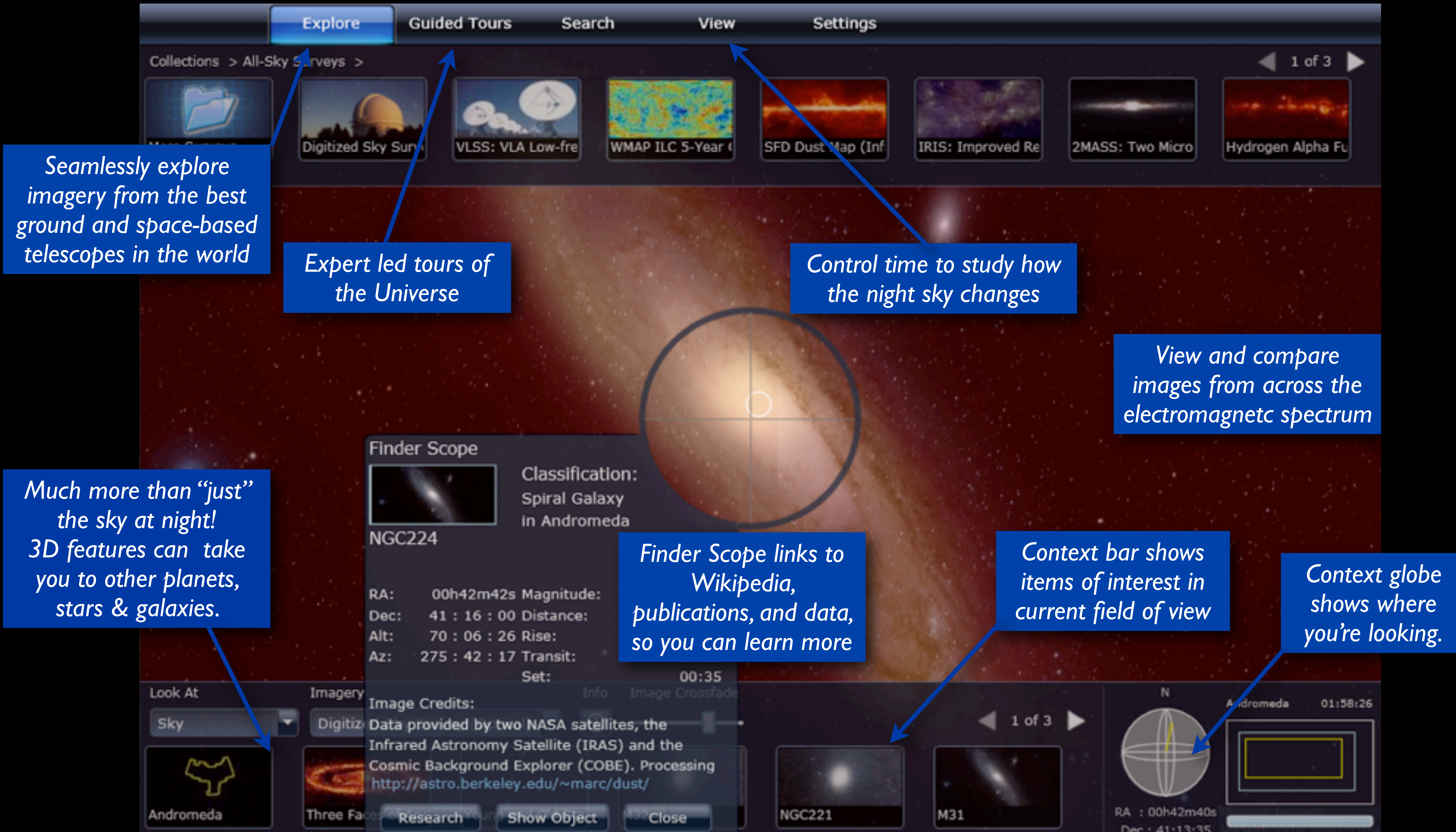
Astronomy as an Archetype for Online Science

Astronomy exemplifies these phenomena. For thousands of years astronomy was primary empirical with few theoretical models. Theoretical astronomy began with Kepler is now co-equal with observation. Astronomy was early to adopt computational techniques to model stellar and galactic formation and celestial mechanics. Today, simulation is an important part of the field – producing new science, and solidifying our grasp of existing theories.

Astronomers are building telescopes that produce terabytes of data each year -- soon terabytes per night. In the old

Microsoft® Research WorldWide Telescope

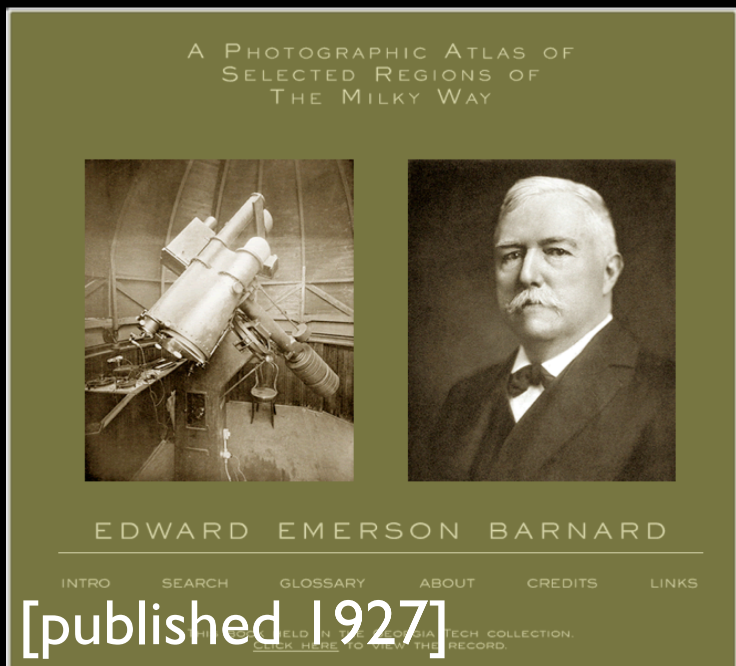
Experience WWT at **worldwidetelescope.org**

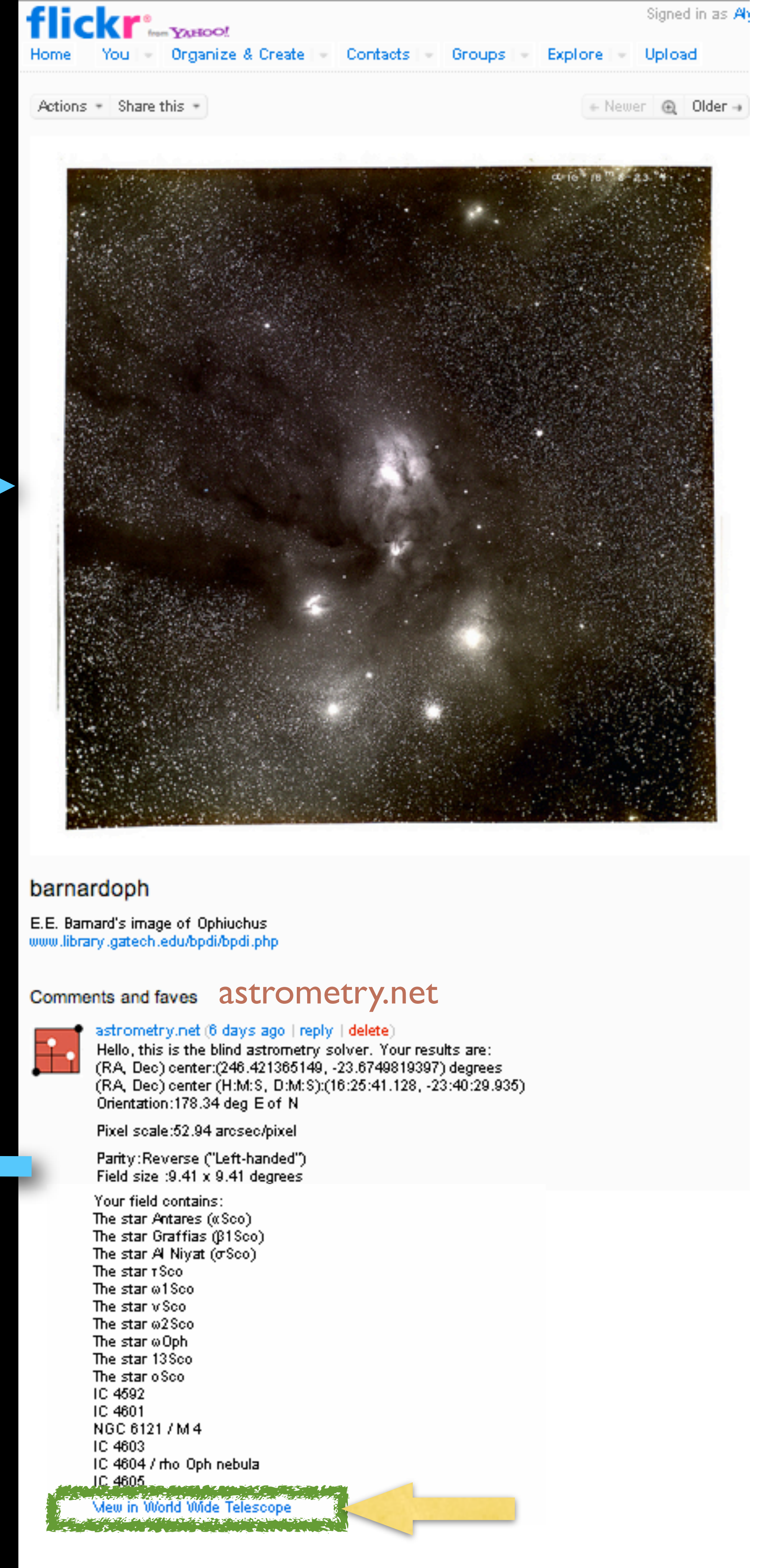


Experience WWT at **worldwidetelescope.org**

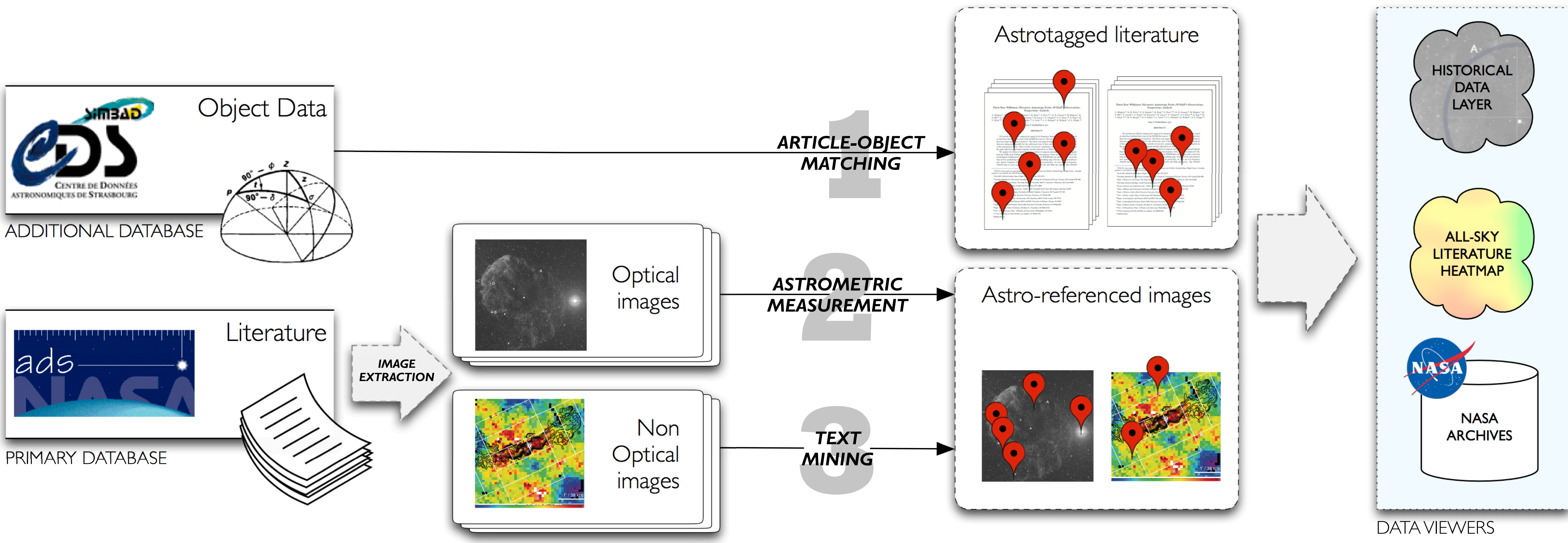
Images in Context

astrometry.net + flickr + WWT

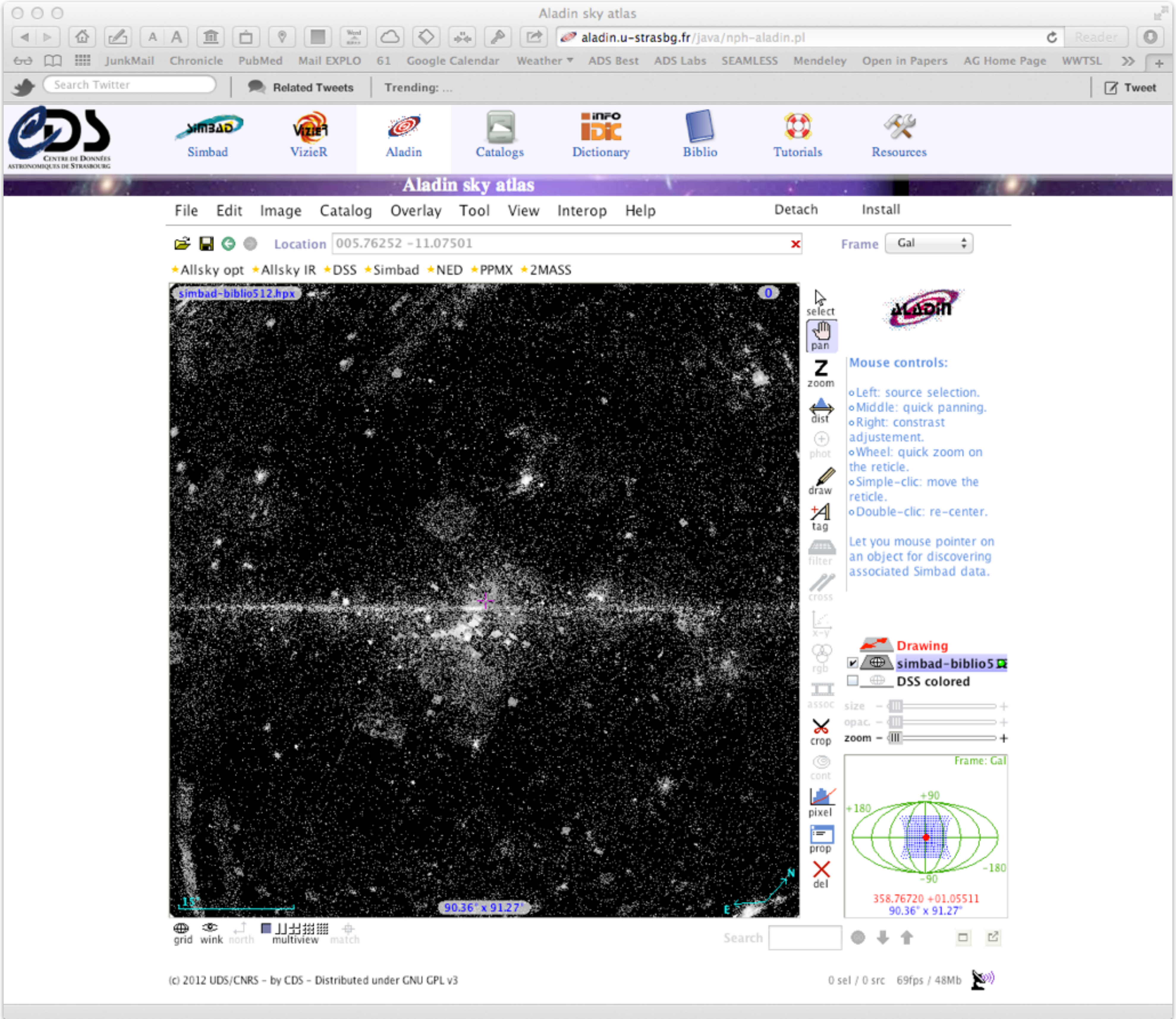




ADS All Sky Survey:
MILLIONS of Articles & Images in Context



ADS All Sky Survey



[prototype: using CDS tools]

How (context and)
John Huchra
fixed the Universe...


WorldWide Telescope Ambassadors

Search this site: Search

HOME | ABOUT | LEARN WWT | FIND TOURS | EDUCATORS | AMBASSADORS | COMMUNITY | GET WWT

- Home
- About
- Learn WWT
- Find Tours
- Educators
- Ambassadors
- Community
- Get WWT

Home

John Huchra's Universe

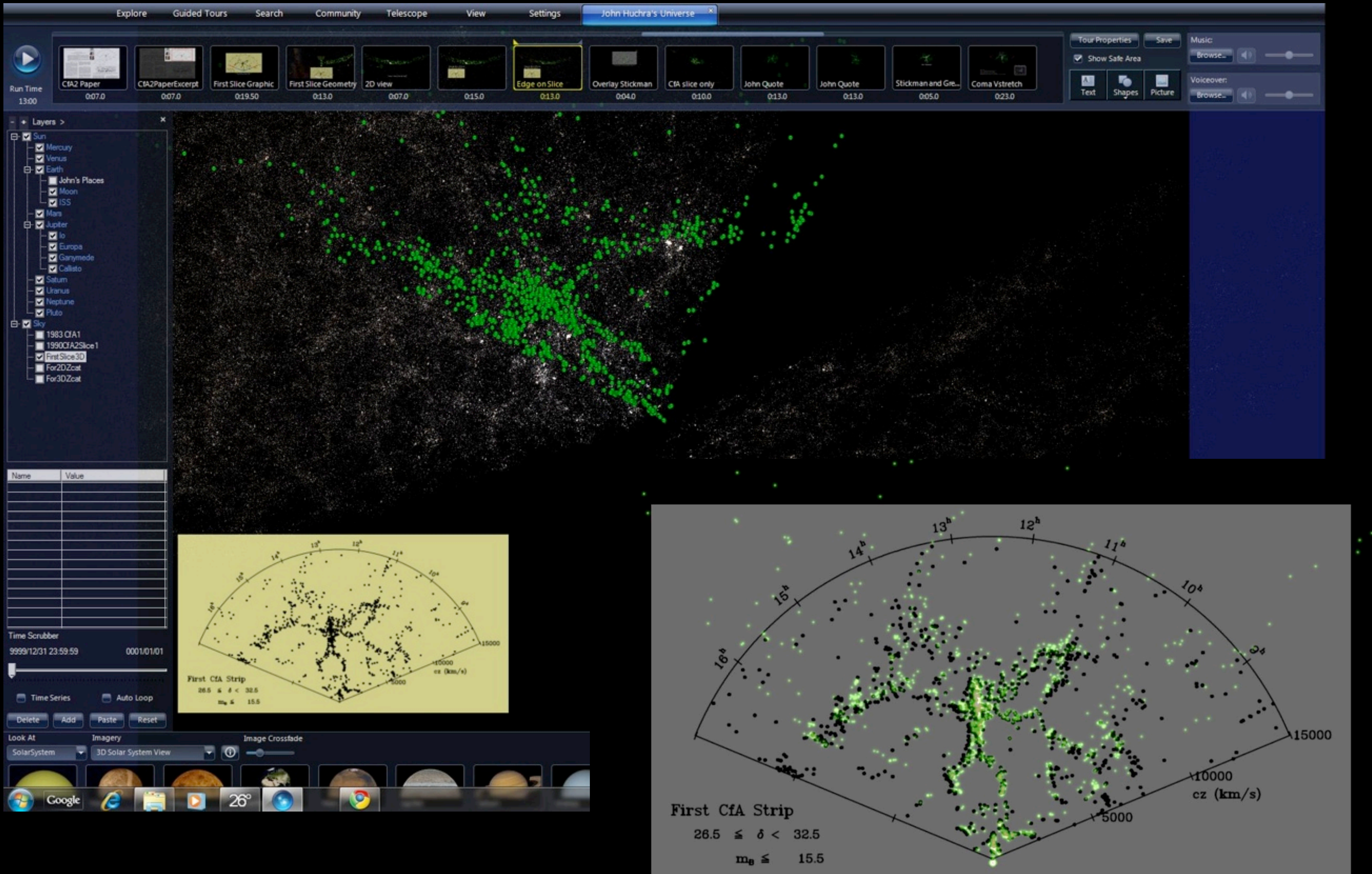
Submitted by patudom on Jan. 11

Download YouTube

John Huchra, former president of the **American Astronomical Society**, passed away on October 8, 2010.

John's colleagues at the Harvard-Smithsonian Center for Astrophysics, in collaboration with the creators of WorldWide Telescope at Microsoft Research, have created a new, interactive, WWT Tour to honor John and his career.

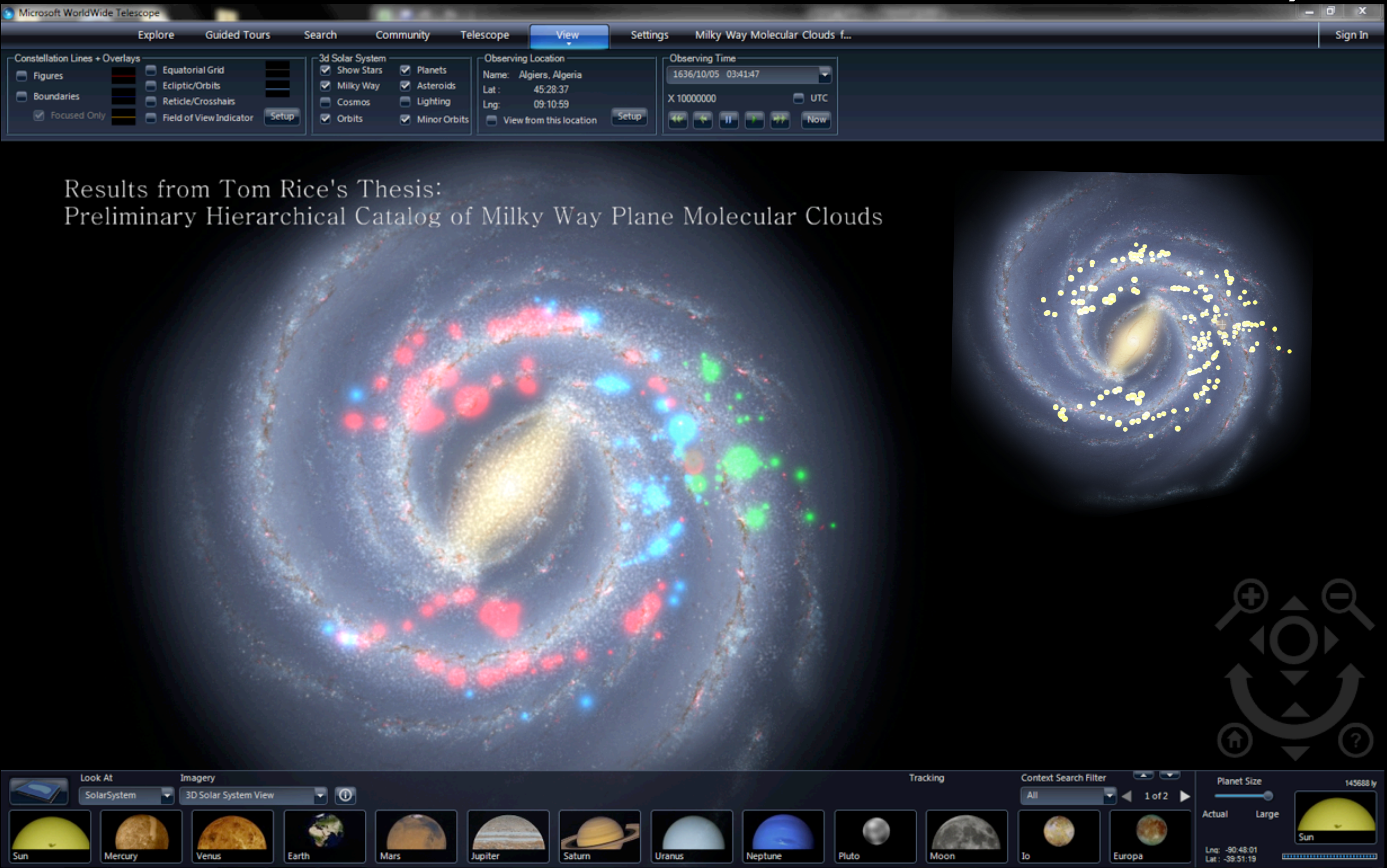
This WorldWide Telescope Tour was created to thank
John Huchra (1948-2010) for the knowledge and cheer he gave us all.



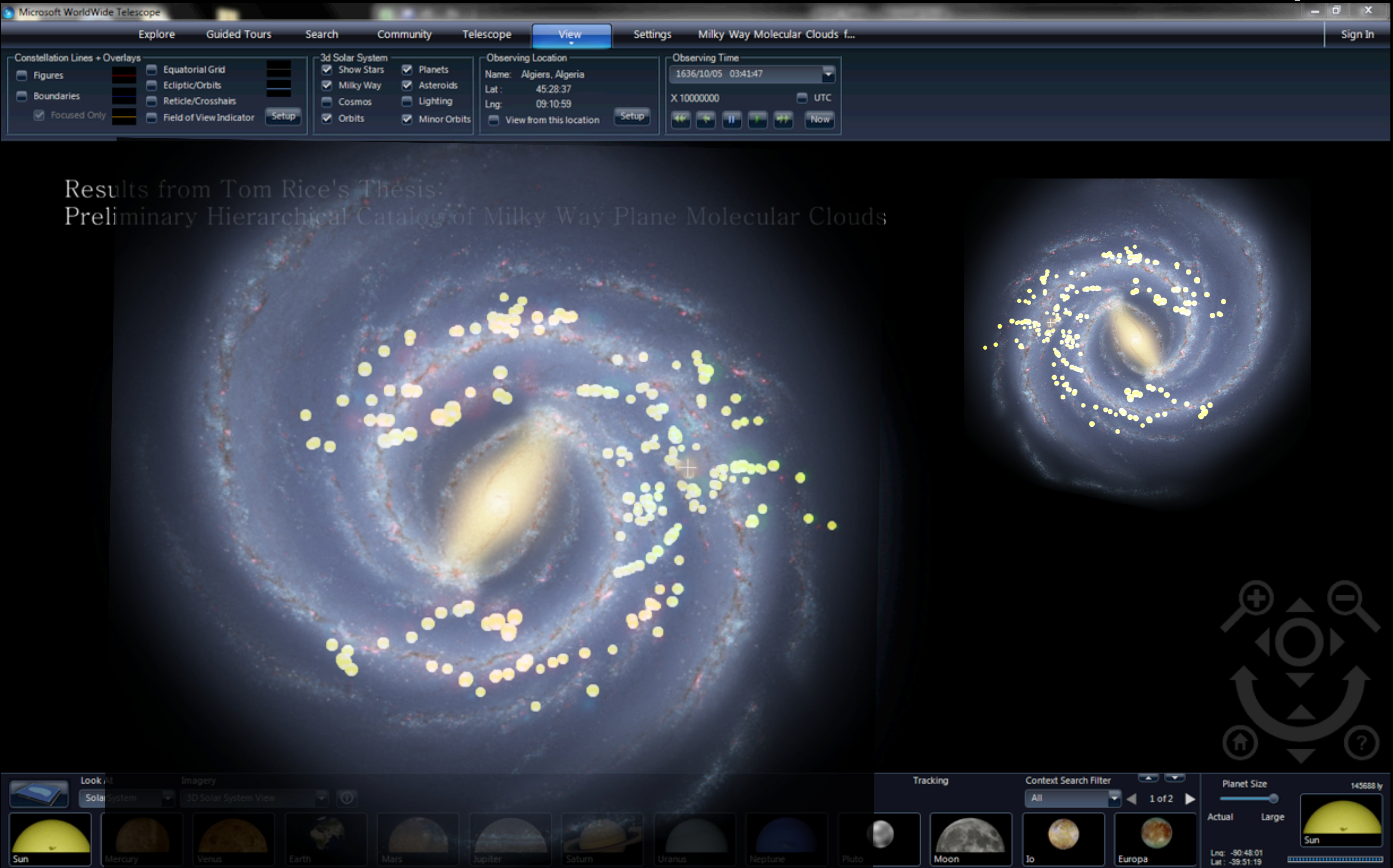
[demo]

This section cannot be "finished" until left-right reversal is addressed by Jonathan. Here, I've spun the Universe around 180 degrees to make everything match.

and the Galaxy...



and the Galaxy...



wwtambassadors.org

Microsoft® Research
WorldWide Telescope

Ambassadors Program

Home | About | My Account | Find Tours | Resources | Log Out | Search

About the WWT Telescope Ambassadors Program

WorldWide Telescope (WWT) is a rich visualization environment that functions as a virtual telescope, allowing anyone to make use of professional astronomical data to explore and understand the universe. As of early 2010, the new WWT Ambassadors Program is recruiting astronomically-literate volunteers, including retired scientists engineers—all of whom will be trained to be experts in using WWT as a teaching tool. Ambassadors will give volunteer presentations at public libraries, community centers, museums, and schools, demonstrating WWT's power to help laypeople visualize and understand our universe.

Read more

John Huchra's Universe

Submitted by patudom on Jan. 11

John Huchra, former president of the **American Astronomical Society**, passed away on October 8, 2010.

John's colleagues at the Harvard-Smithsonian Center for Astrophysics, in collaboration with the creators of WorldWide Telescope at Microsoft Research, have created a new, interactive, WWT Tour to honor John and his career. The Tour primarily focuses on John's quest to map the Universe in three dimensions. It is 12.5 minutes long.

The Tour is best experienced inside the WorldWide Telescope program itself. (**Note: You must have the version of WWT released on 1/13/2011 to view all of this Tour's content. You can download it from here.**) As viewed within the WWT program, the Tour content is interactive, allowing users to pause and explore the parts of the Universe featured in the tour, explore web hyperlinks, and more. For those who do not have the desktop client, the Tour has been posted as a video as well.

Video (Interactive WWT features will be disabled)

John Huchra's Universe

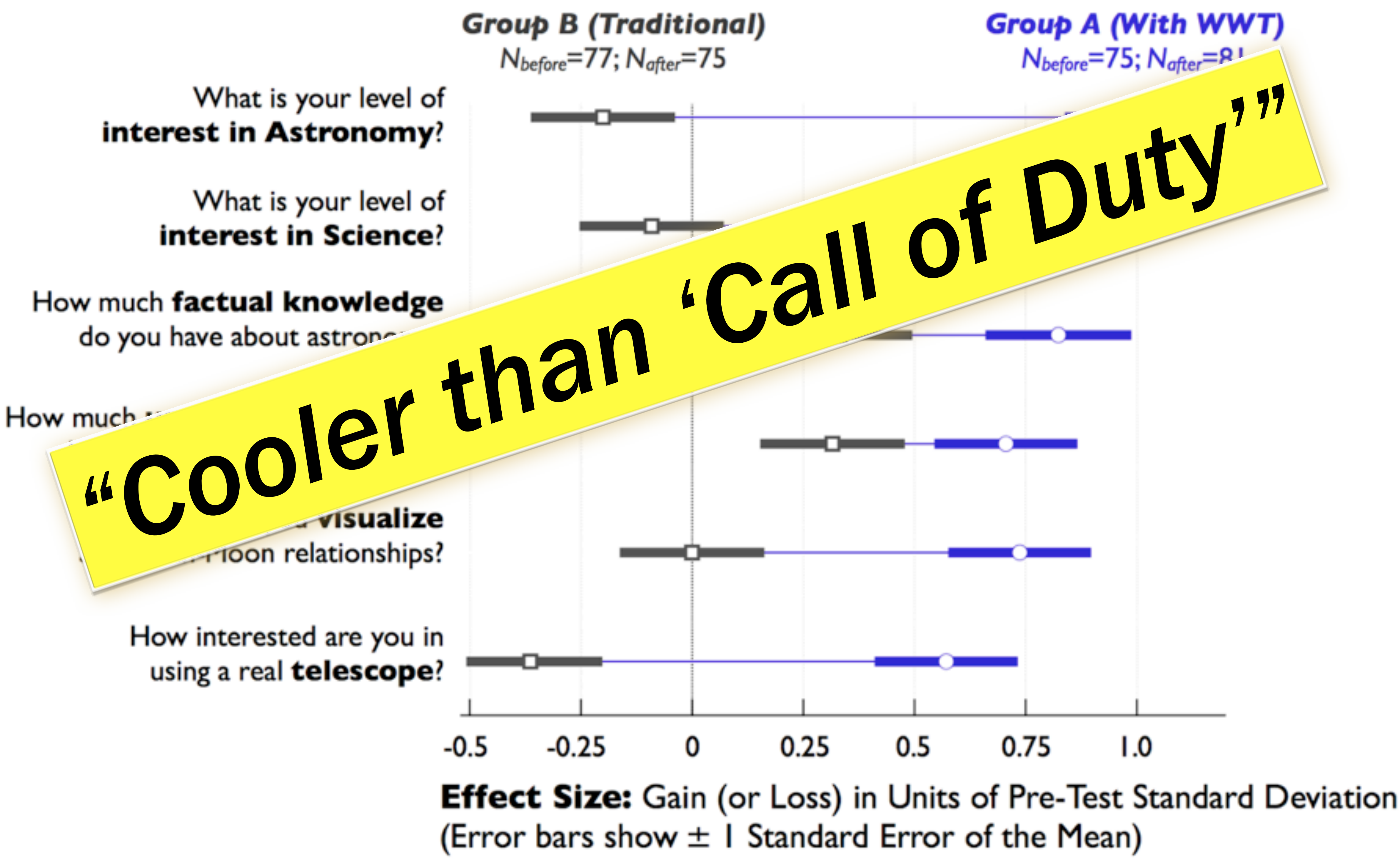
Friends of John Huchra have released a new WWT Tour to honor John and his work. The Tour primarily focuses on John's quest to map the Universe in three dimensions. You can view the Tour here.

Upcoming

- Cyberlearning Tools for STEM Education Conference
  Mar. 8 - Mar. 9
- Cambridge Science Festival
  Apr. 30 - May. 10

Gains in Student Interest and Understanding

("Traditional Way" vs "WWT Way")



EMR19/12

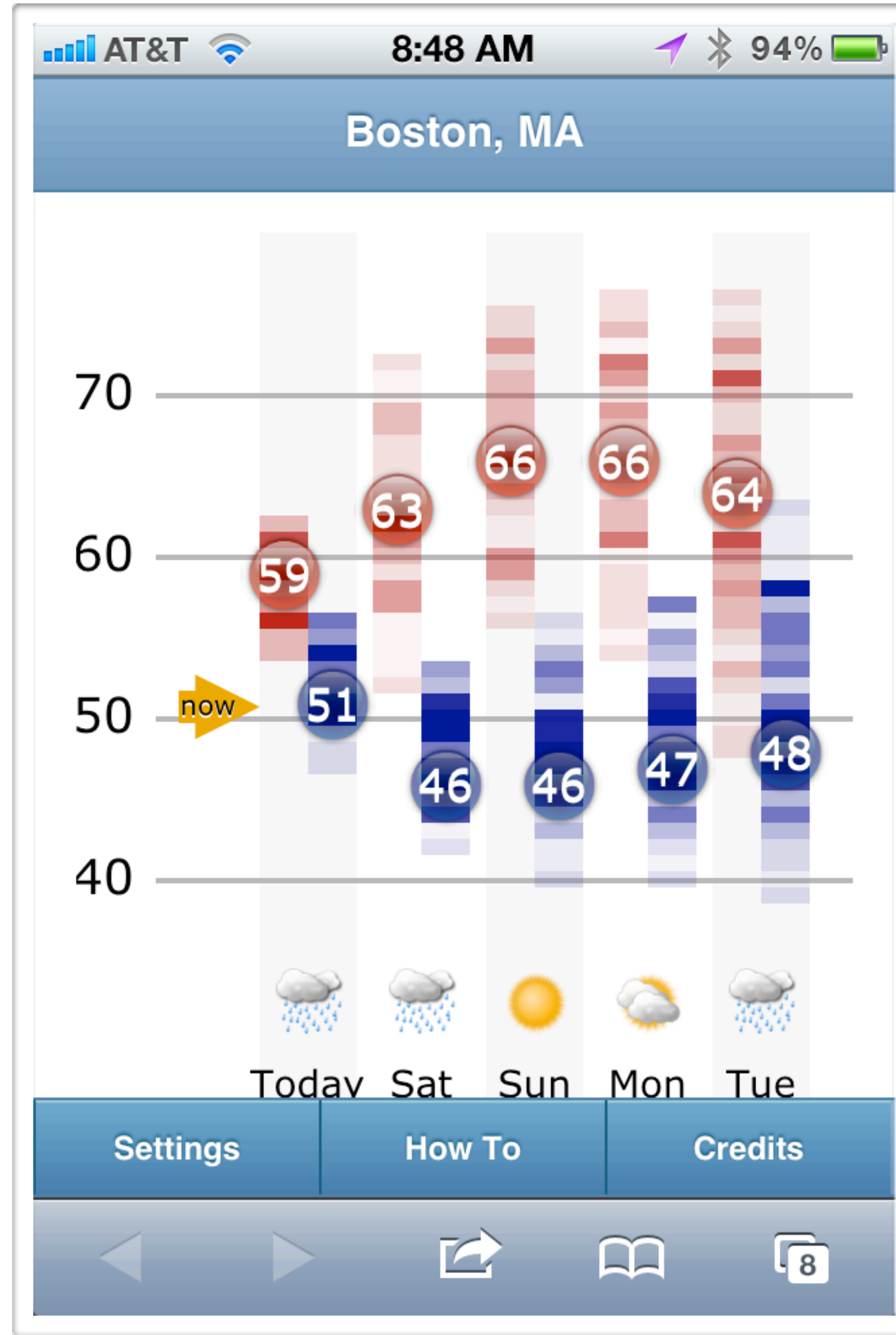
takeasweater?

Alyssa A. Goodman

Software development, 2010-12
Bill Barthelmy, Harvard FAS
Academic Technology Group

WGBH collaborators, 2008-9
Annie Valva, Howard Cutler, et al.

Data provider, 2011-12
Eric Floehr of ForecastWatch



Astron. Nachr. / AN **333**, No. 5/6, 505 – 514 (2012) / **DOI** 10.1002/asna.201211705

Principles of high-dimensional data visualization in astronomy

A.A. Goodman*

Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

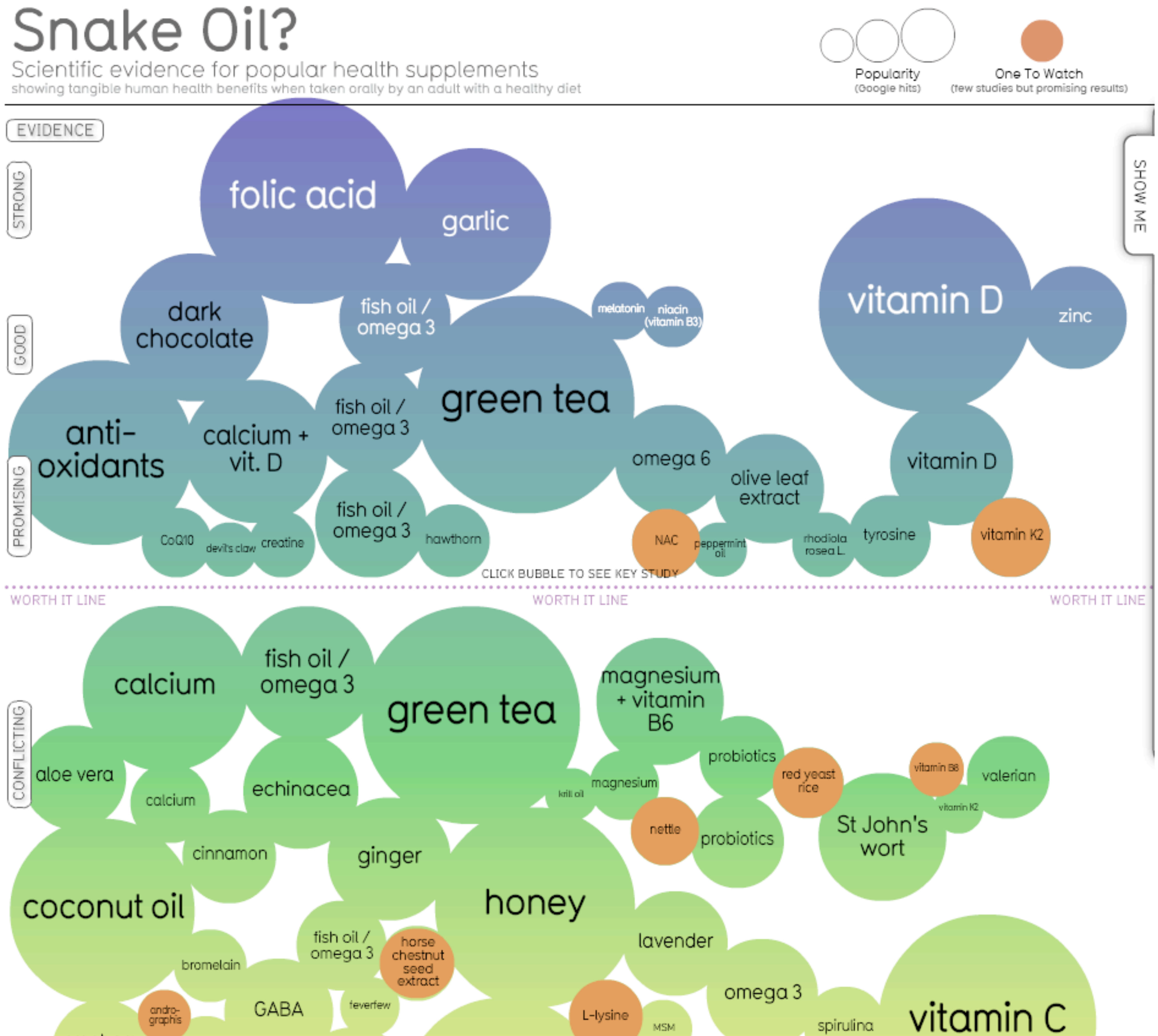
Received 2012 May 3, accepted 2012 May 4
Published online 2012 Jun 15

Key words cosmology: large-scale structure – ISM: clouds – methods: data analysis – techniques: image processing – techniques: radial velocities

Astronomical researchers often think of analysis and visualization as separate tasks. In the case of high-dimensional data sets, though, interactive *exploratory data visualization* can give far more insight than an approach where data processing and statistical analysis are followed, rather than accompanied, by visualization. This paper attempts to charts a course toward "linked view" systems, where multiple views of high-dimensional data sets update live as a researcher selects, highlights, or otherwise manipulates, one of several open views. For example, imagine a researcher looking at a 3D volume visualization of simulated or observed data, and simultaneously viewing statistical displays of the data set's properties (such as an x-y plot of temperature vs. velocity, or a histogram of vorticities). Then, imagine that when the researcher selects an interesting group of points in any one of these displays, that the same points become a highlighted subset in all other open displays. Selections can be graphical or algorithmic, and they can be combined, and saved. For tabular (ASCII) data, this kind of analysis has long been possible, even though it has been under-used in astronomy. The bigger issue for astronomy and other "high-dimensional" fields, though, is that no extant system allows for full integration of images and data cubes within a linked-view environment. The paper concludes its history and analysis of the present situation with suggestions that look toward cooperatively-developed open-source modular software as a way to create an evolving, flexible, high-dimensional, linked-view visualization environment useful in astrophysical research.

DATA • DIMENSIONS • DISPLAY

DAVID MCCANDLESS: WWW.INFORMATIONISBEAUTIFUL.NET/PLAY/SNAKE-OIL-SUPPLEMENTS/

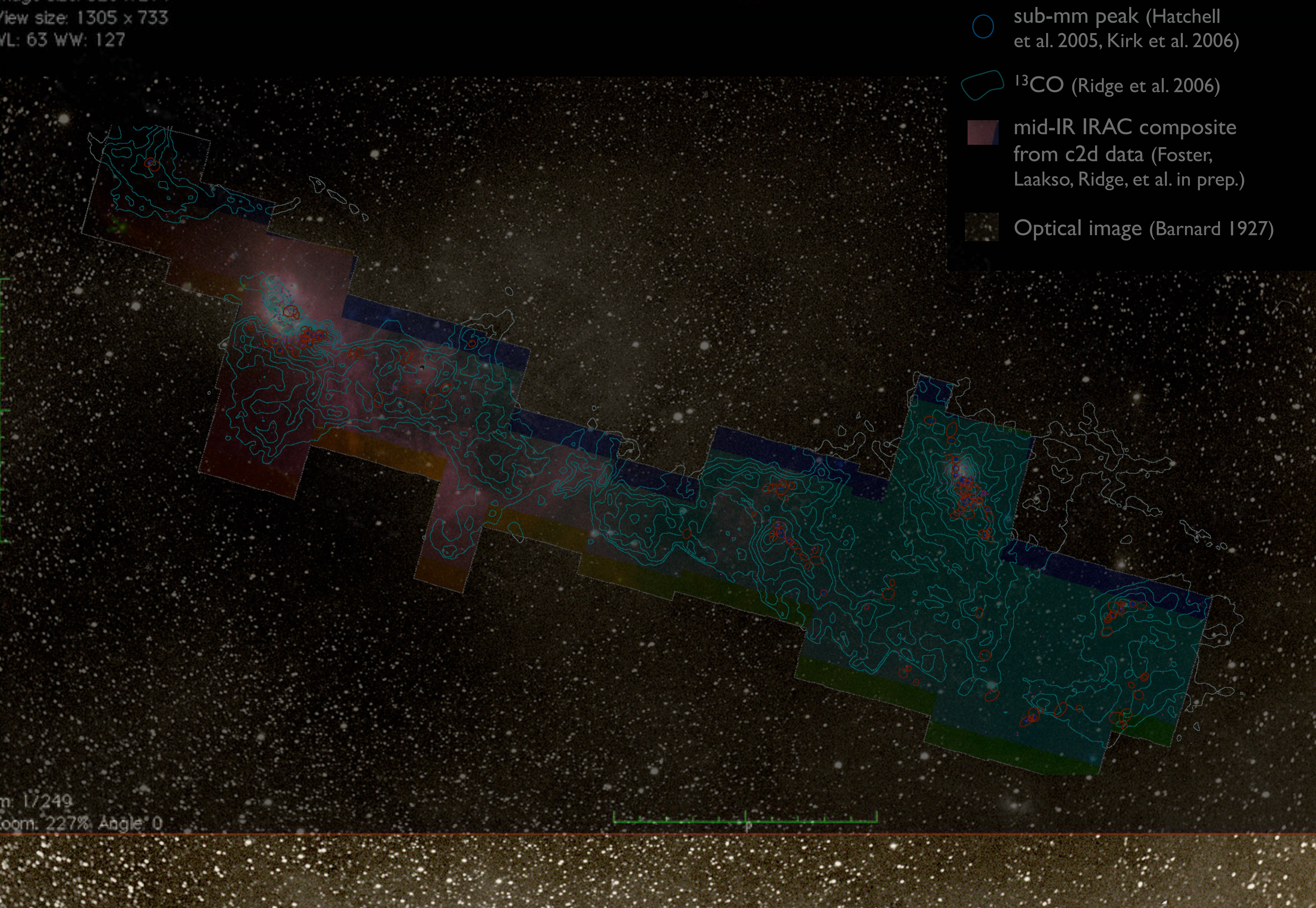


“PERSEUS IN 3D”


© 2007 Kurt Wenner

COMPLETE Perseus

- 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)



The AstroMed Story

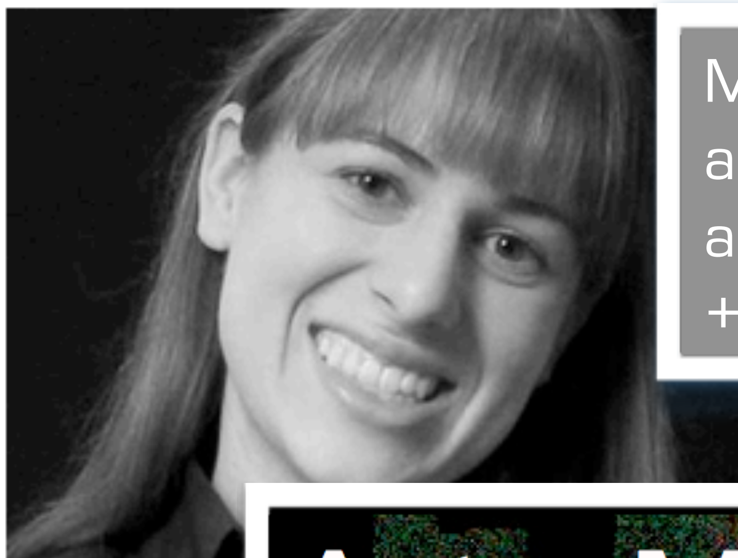
Michelle Borkin is now a SEAS PhD Student, advised by Profs. Alyssa Goodman (Astronomy) and Hanspeter Pfister (SEAS), and IIC +AstroMed became the bases for the Viz-e-Lab

TED Ideas worth spreading

Themes | Speakers | Talks | Translations | TED Conferences | TEDx Events | TED Prize | TED Fellows

TED Fellows The TED Fellows Directory > Michelle Borkin

2009



TEDGlobal 2009

Bio

Michelle Borkin
interdisciplinary
and image analy
She wrote her u
on the applicati
astronomical da
part of the "Astro
Harvard's Initiati
works with the de
tools to improve

erting a stent

perceptual edge

2011

Visual Business Intelligence
A blog by Stephen Few

Home | About | Consulting | Workshops | Courses | Examples | Library | **Blog** | Discussio

VisWeek 2011 – Award-Worthy Visualization Research

On Tuesday in this blog I expressed my frustration with VisWeek's information visualization research awards process. I don't want to leave you with the impression, however, that the state of information visualization research is bleak. Each year at VisWeek I find a few gems produced by thoughtful, well-trained information visualization researchers. They identified potentially worthy pursuits and did well-designed research that produced useful results. While puzzling over the criteria that the judges must have used when selecting this year's best paper, I spent a few minutes considering the criteria that I would use were I a judge, and came up with the following list with points totaling to 100:

Effectiveness (It does what it's supposed to do and does it well.) — 30 points

Usefulness (What it does addresses real needs in the world.) — 30 points

10 points

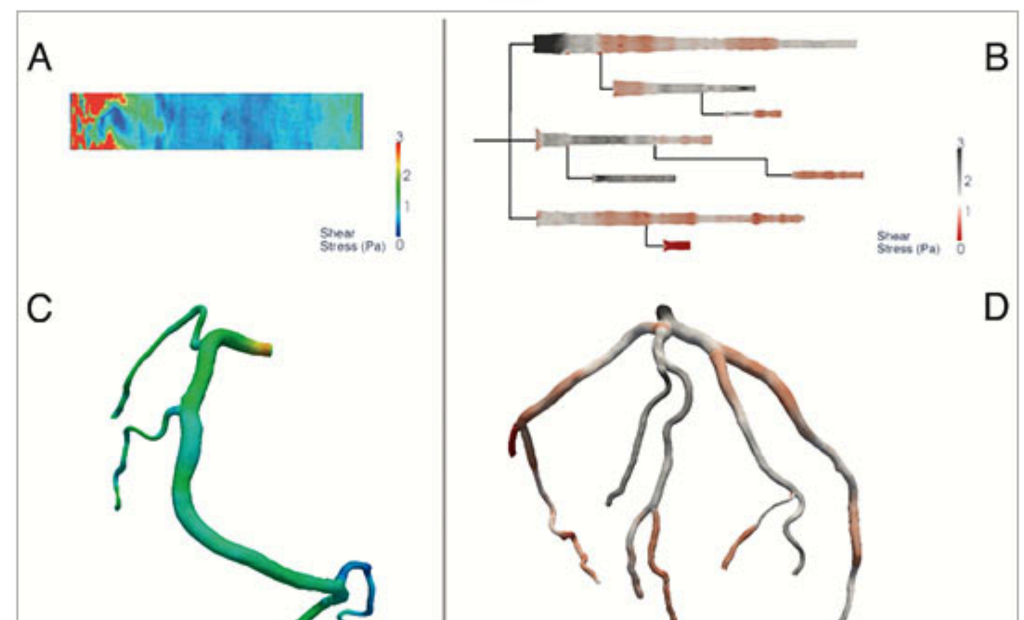
ses.) — 10 points

ew way.) — 10 points

e.) — 10 points

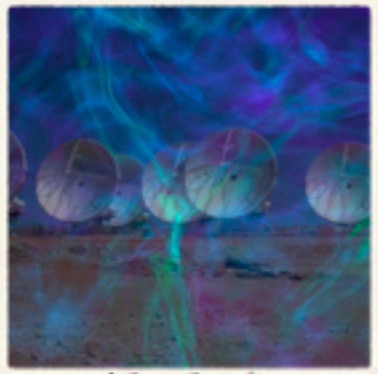
to some degree, but this gives you an idea of
the importance of each.

e by its elegance and exceptional usefulness
rvard University's School of Engineering and
Applied Sciences titled "Evaluations of Artery Visualizations for Heart Disease Diagnosis."

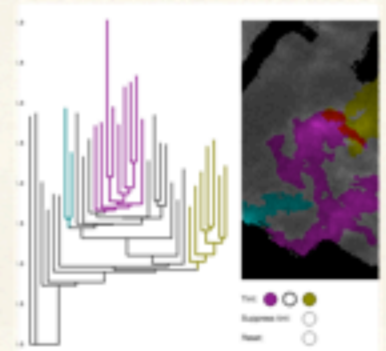


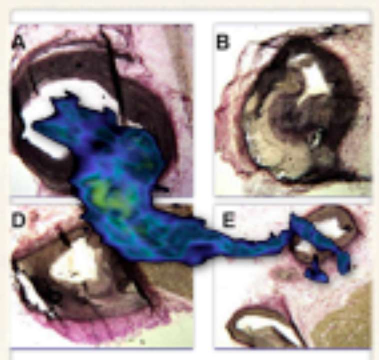
Viz-e-Lab

Projects 2011


"Taste-Testing"


WorldWide Telescope/Ambassadors


Linked Views/Dendrograms


High-D Data Viz


Seamless Astronomy


Wolbach User Experience Lab

ADS Labs

CfA "Astronomy Dataverse"

VAO/Online Astronomy User Group

collaborators/**contacts** at CfA

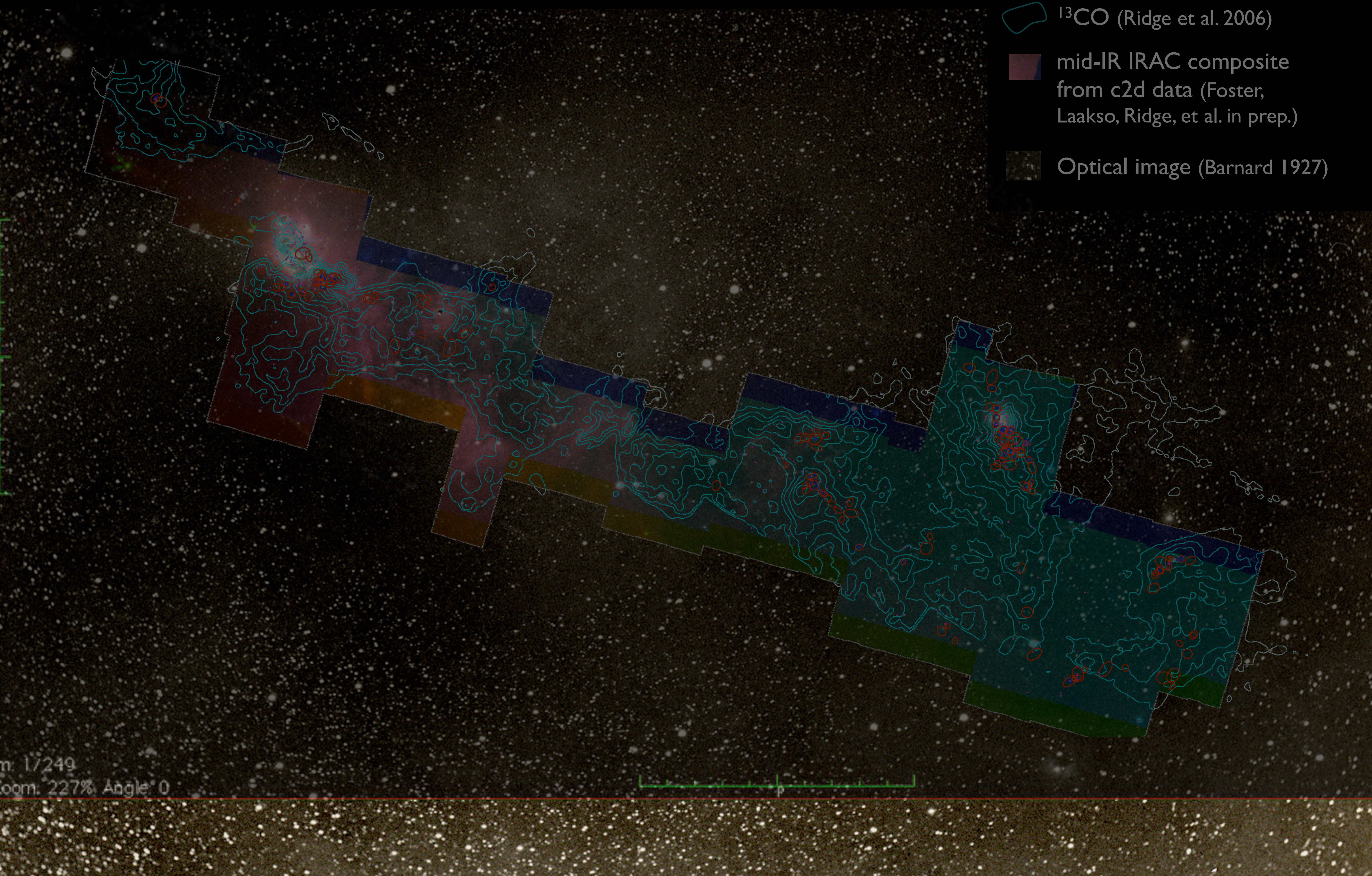
Seamless Astronomy: Alyssa **Goodman** **Online Astronomy** Group, CfA Data **Archives**: Gus **Muench** **ADS** Group: Alberto **Accomazzi**

WorldWide Telescope Ambassadors: Pat **Udomprasert** High-Dimensional **Data Visualization & Interaction**: Michelle **Borkin**

Wolbach Library Lab at CfA : Christopher **Erdmann** **VAO** at CfA: Pepi **Fabbiano** Social **Networks** in Science: Alberto **Pepe**

Questions about using the Viz-e-Lab? Contact Sarah Block, 5-7331, sblock@cfa.harvard.edu

COMPLETE Perseus



mm peak (Enoch et al. 2006)

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

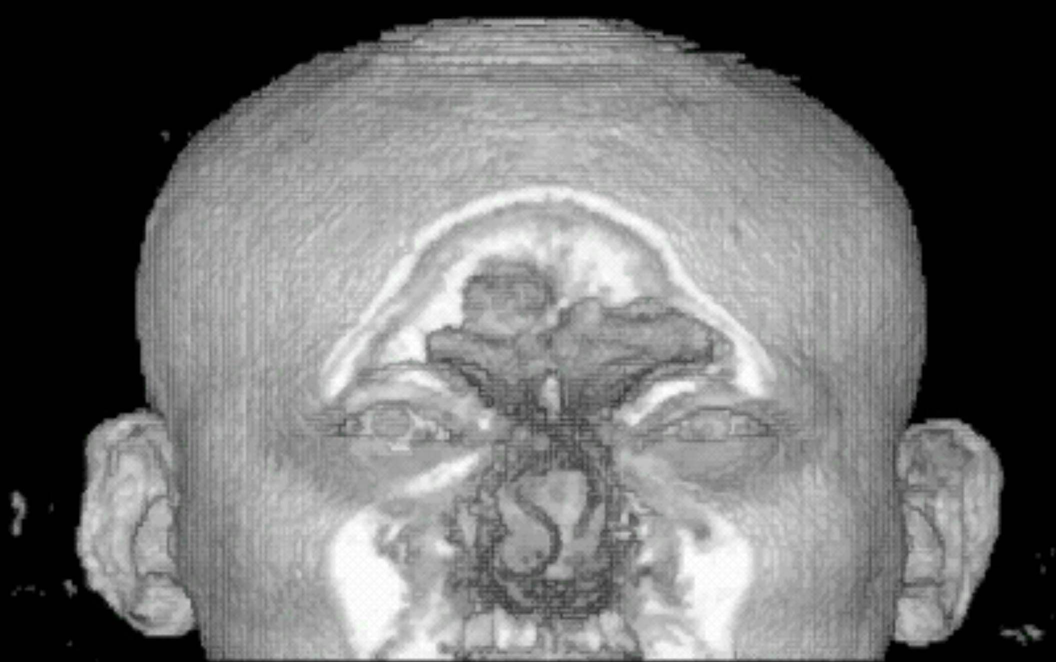
^{13}CO (Ridge et al. 2006)

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

Optical image (Barnard 1927)

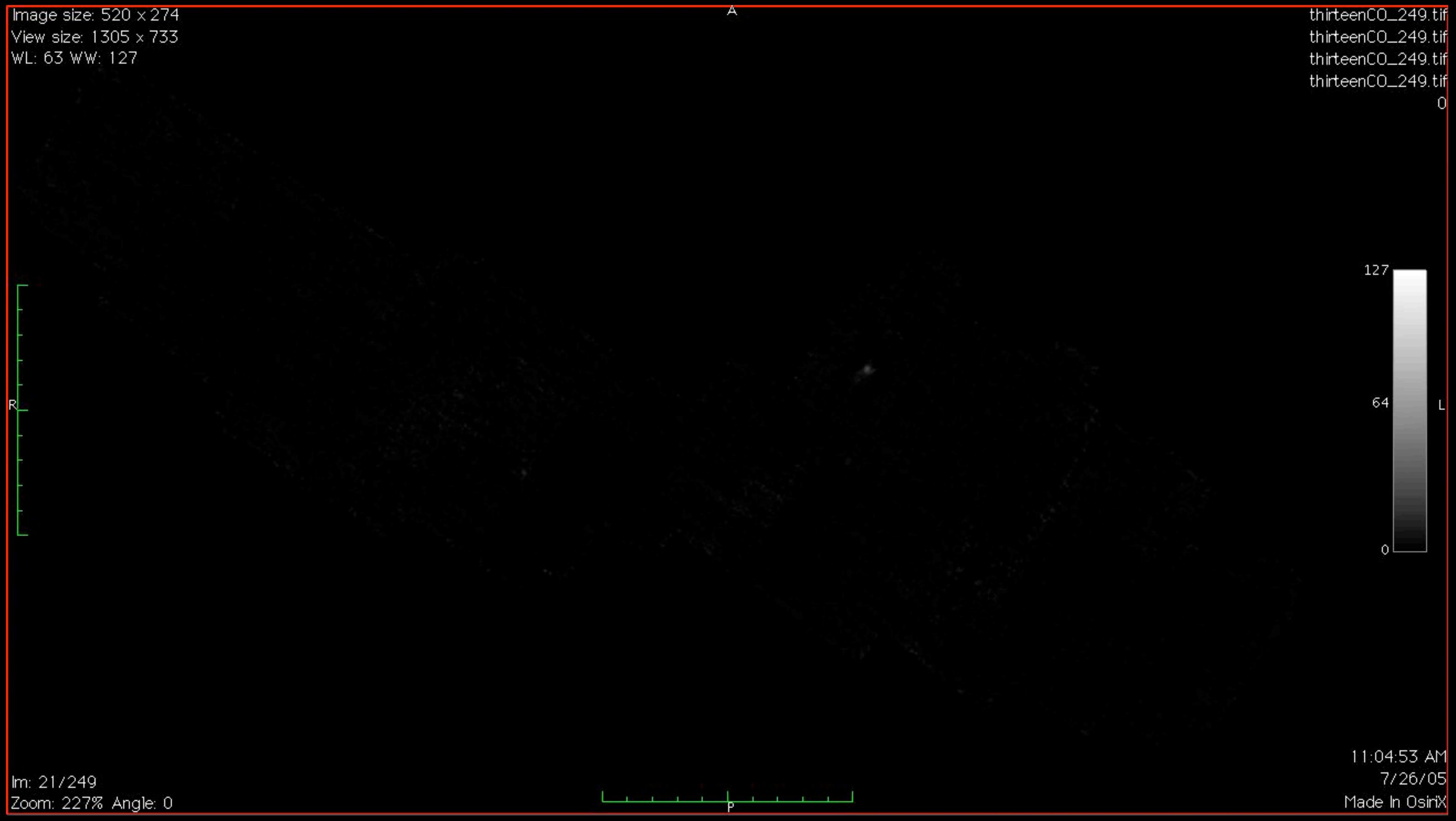
"Astronomical Medicine"

"KEITH"



"z" is depth into head

"PERSEUS"

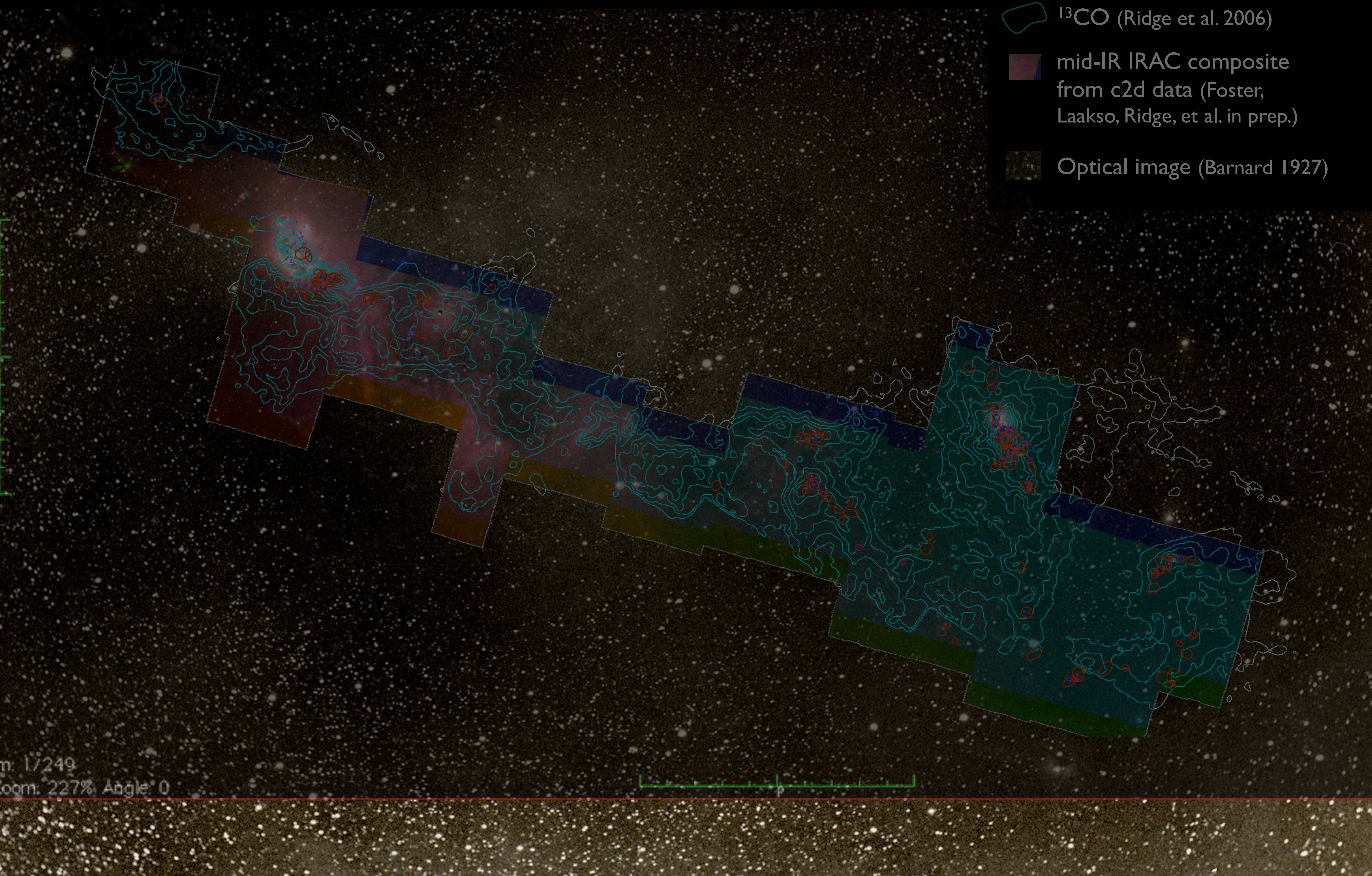


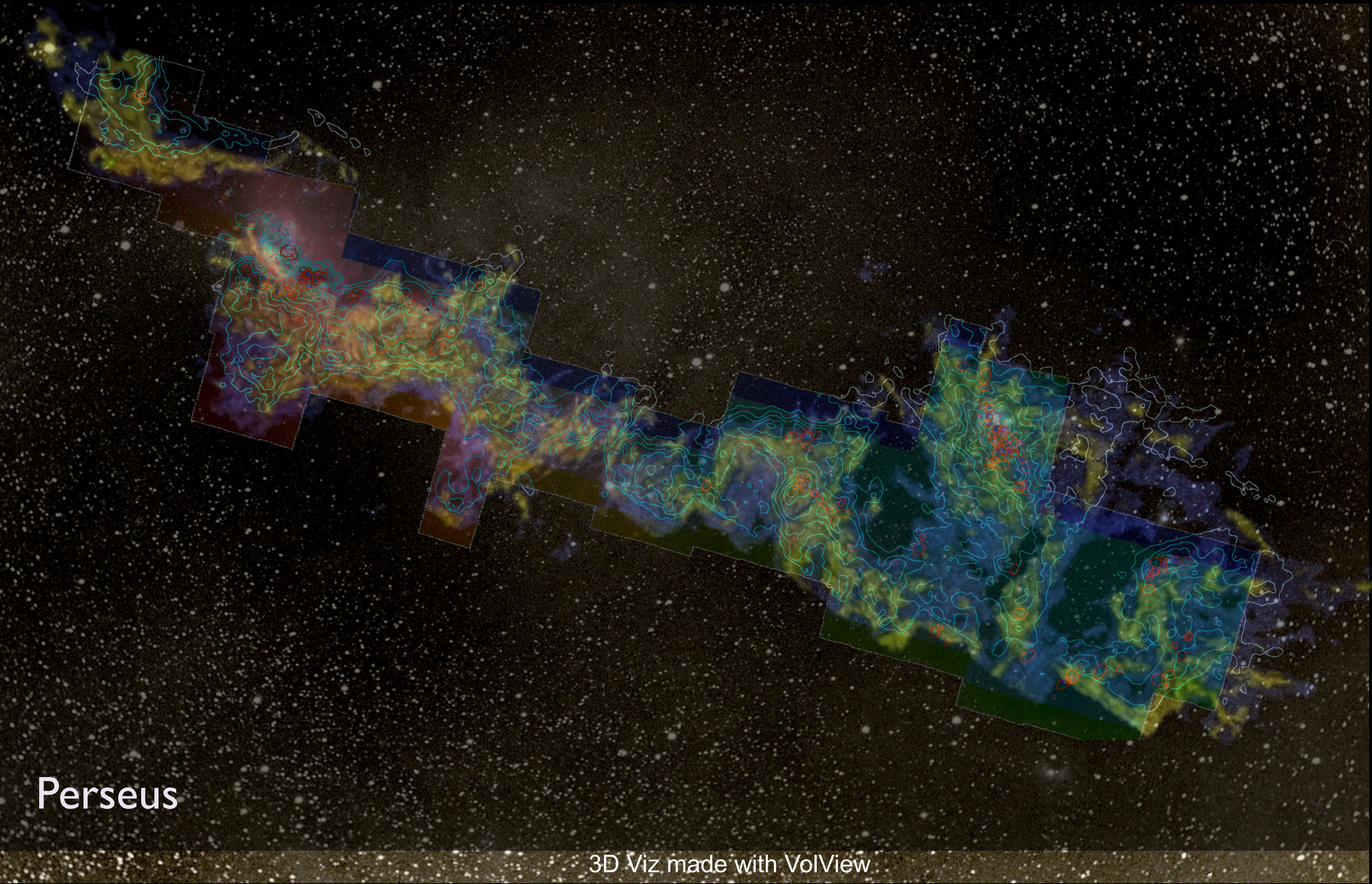
"z" is line-of-sight velocity

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

COMPLETE Perseus

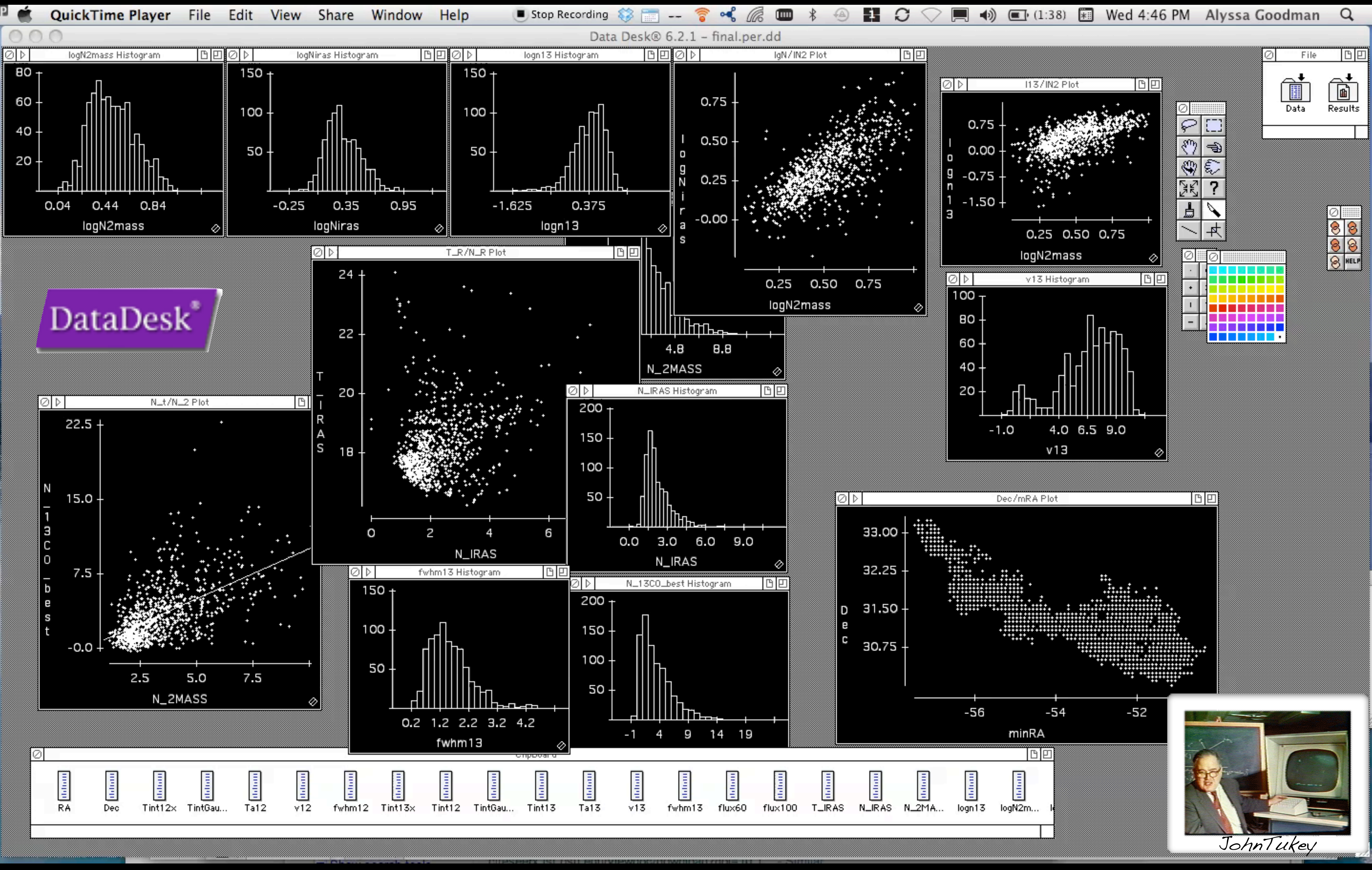
- mm peak (Enoch et al. 2006)
- sub-mm peak (Hatchell et al. 2005, Kirk et al. 2006)
- ^{13}CO (Ridge et al. 2006)
- mid-IR IRAC composite from c2d data (Foster, Laakso, Ridge, et al. in prep.)
- Optical image (Barnard 1927)



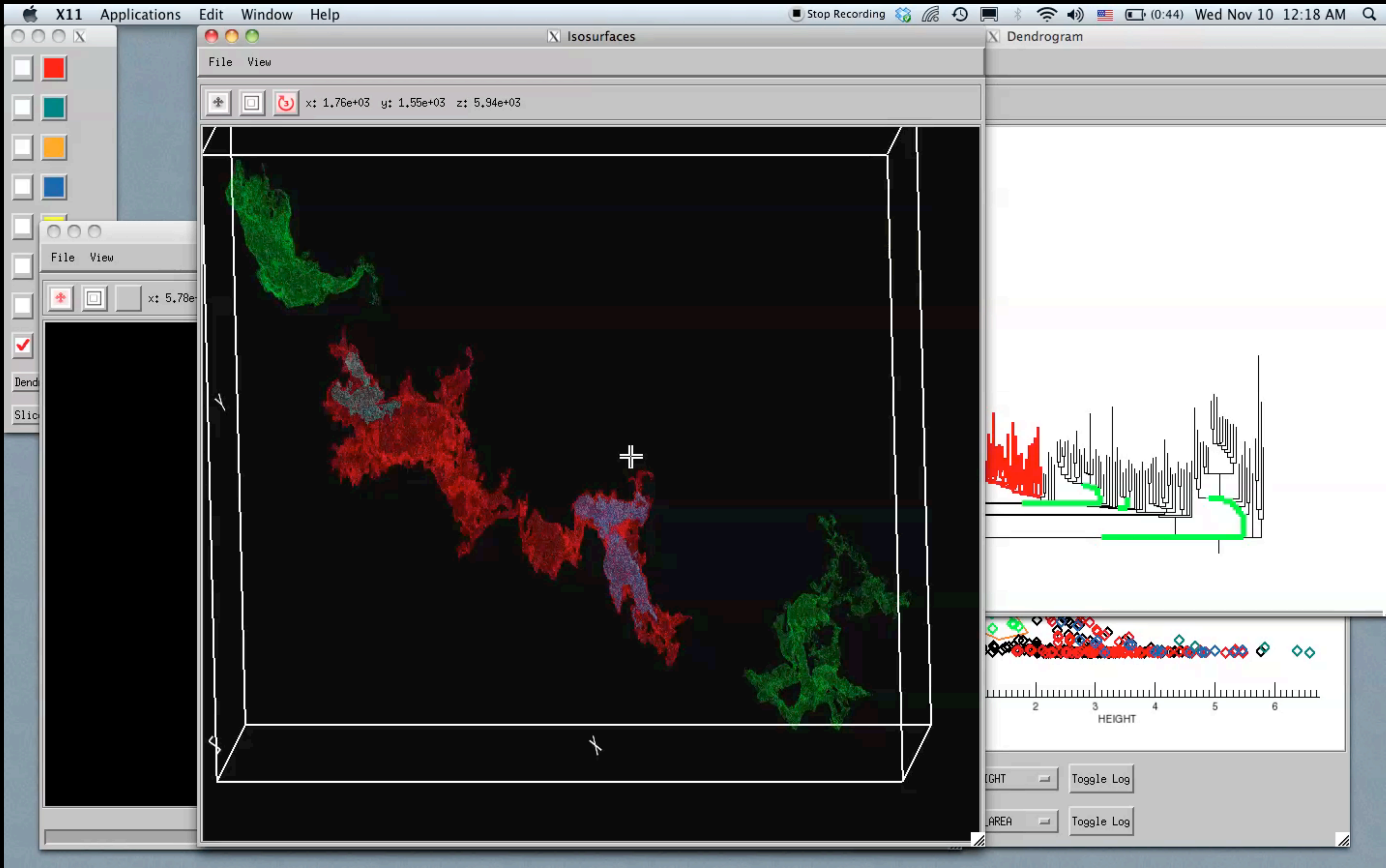

Perseus
3D Viz made with VolView

DataDesk (est. 1986)

"EXPLORATORY DATA ANALYSIS"



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

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

JOHN TUKEY'S LEGACY

PRIM-9

PRIM-H

DataDesk

XGobi

GGobi

RGGobi

glue

Spotfire

Polaris

tableau software

1970 1980 1990 2000 2010

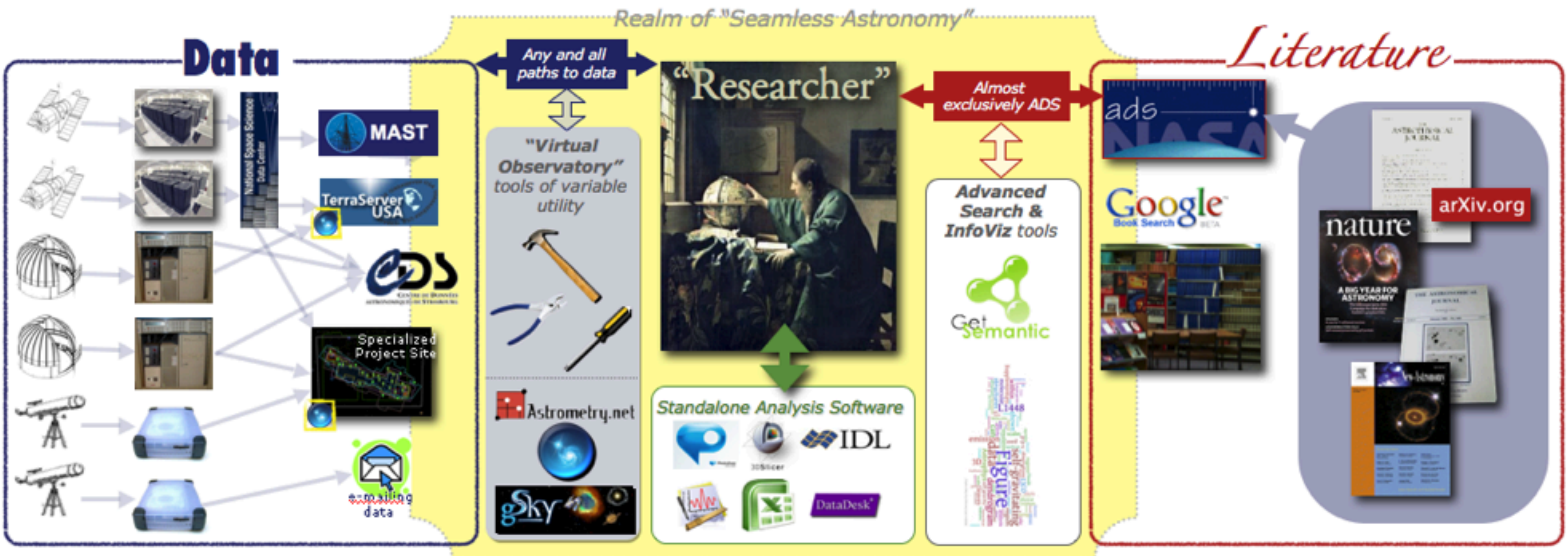
glue

Note: This is a recent screenshot from a context-rich!) story about the Loch Ness monster & the "bones" of the Milky Way--ask me later!

Glue collaboration: Beaumont, Borkin, Goodman, Pfister, Robitaille

SEAMLESS ASTRONOMY

Linking scientific data, publications, and communities



projects.iq.harvard.edu/seamlessastronomy/

What is Universe3D.org?

The intention of Universe3D.org is to host links to web content that enable the enhancement of our three-dimensional view of the Universe.

Recently added Dataset

SLOAN Digital Sky Survey The Sloan Digital Sky Survey or SDSS is a major multi-filter imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. The main galaxy sample has a median redshift of z = 0.1; there are redshifts for luminous red galaxies as far as z = 0.7, and for quasars as far as z = 5; and the imaging survey has been involved in the detection of quasars beyond a redshift z = 6.

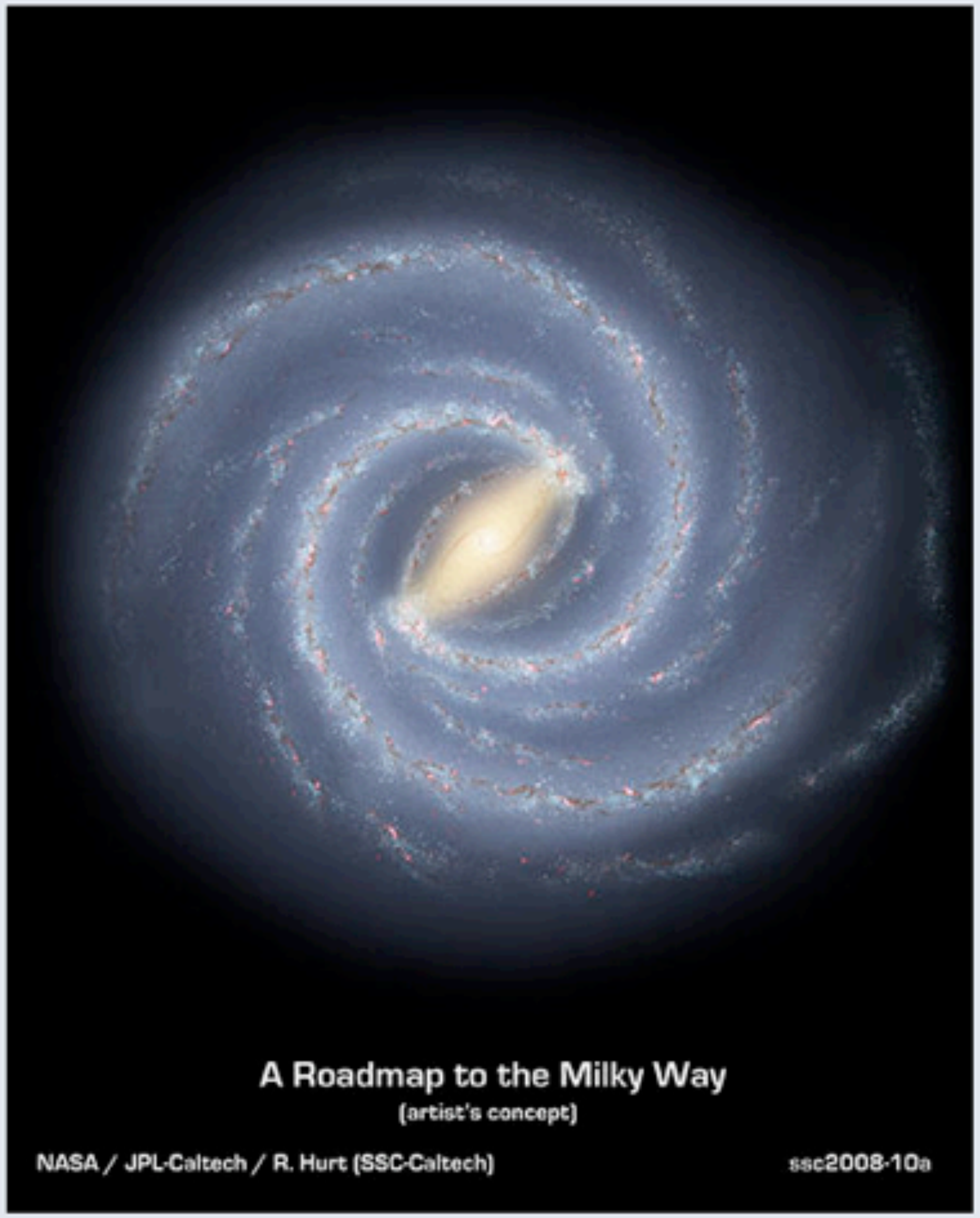
Astronomy News

- *June 26, 2012:* Astronomers use supercomputer to explore role of dark matter in galaxy formation
- *June 25, 2012:* Moon to pass by Mars tonight
- *June 24, 2012:* Astronomers find planets so close they 'see' each other in night sky
- *June 14, 2012:* Huge Asteroid to fly by Earth
- *June 13, 2012:* Astronomers may have discovered the oldest galaxy in the Universe
- *June 5, 2012:* Last Transit of Venus for the 21st century

Announcements

- *July 05, 2012:* Website moved to the URL universe3d.org!
- *June 11, 2012:* Website moved to MediaWiki!
- *December 5, 2011:* Site established!
  To make good on Alyssa Goodman's promise at the "Milky Way 2011" meeing held in Rome this past September, the site "universe3d.org" has been established. By 2012, it will be populated with links to existing data

The Milky Way



Navigation

Home
3D Viewers
Datasets
Images
Publications & Presentations
People

More

Calendar
Announcements
Help

Toolbox

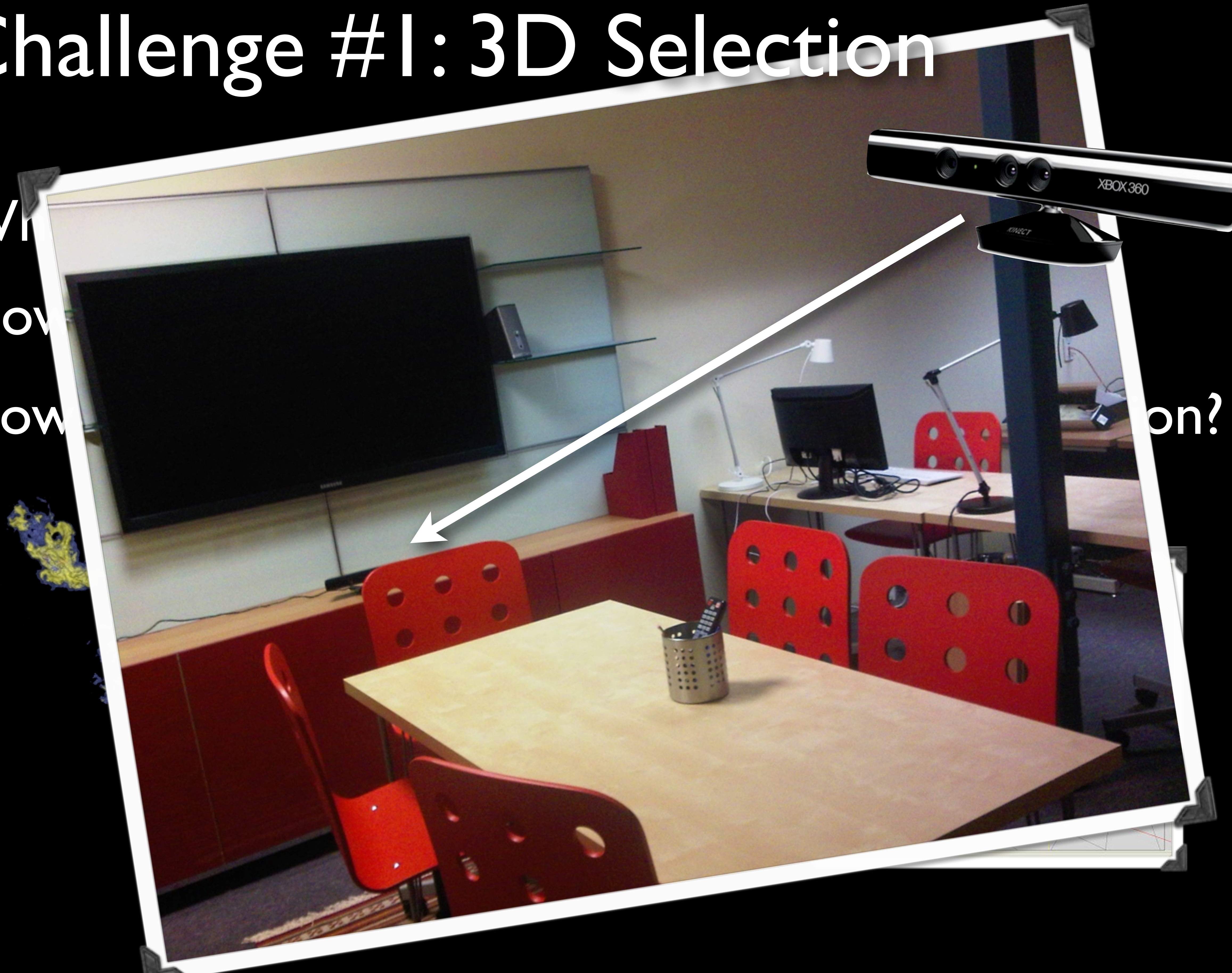
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link

Challenge #1: 3D Selection

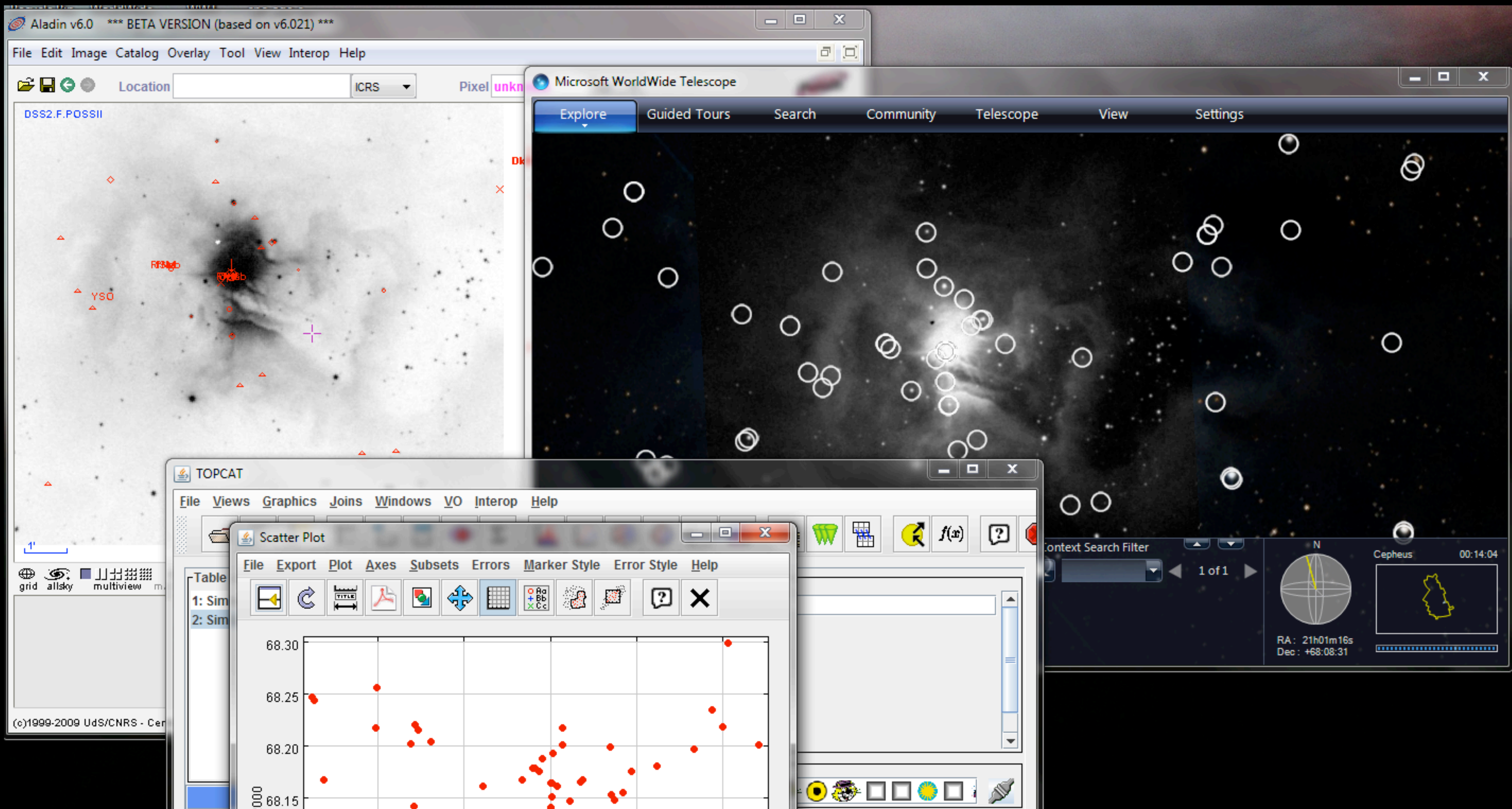
Wh

How

How on?



Challenge #2: Too many windows...



Challenge #3:

What does "Publication-Quality" Graphics Mean in an Interactive 3D World?

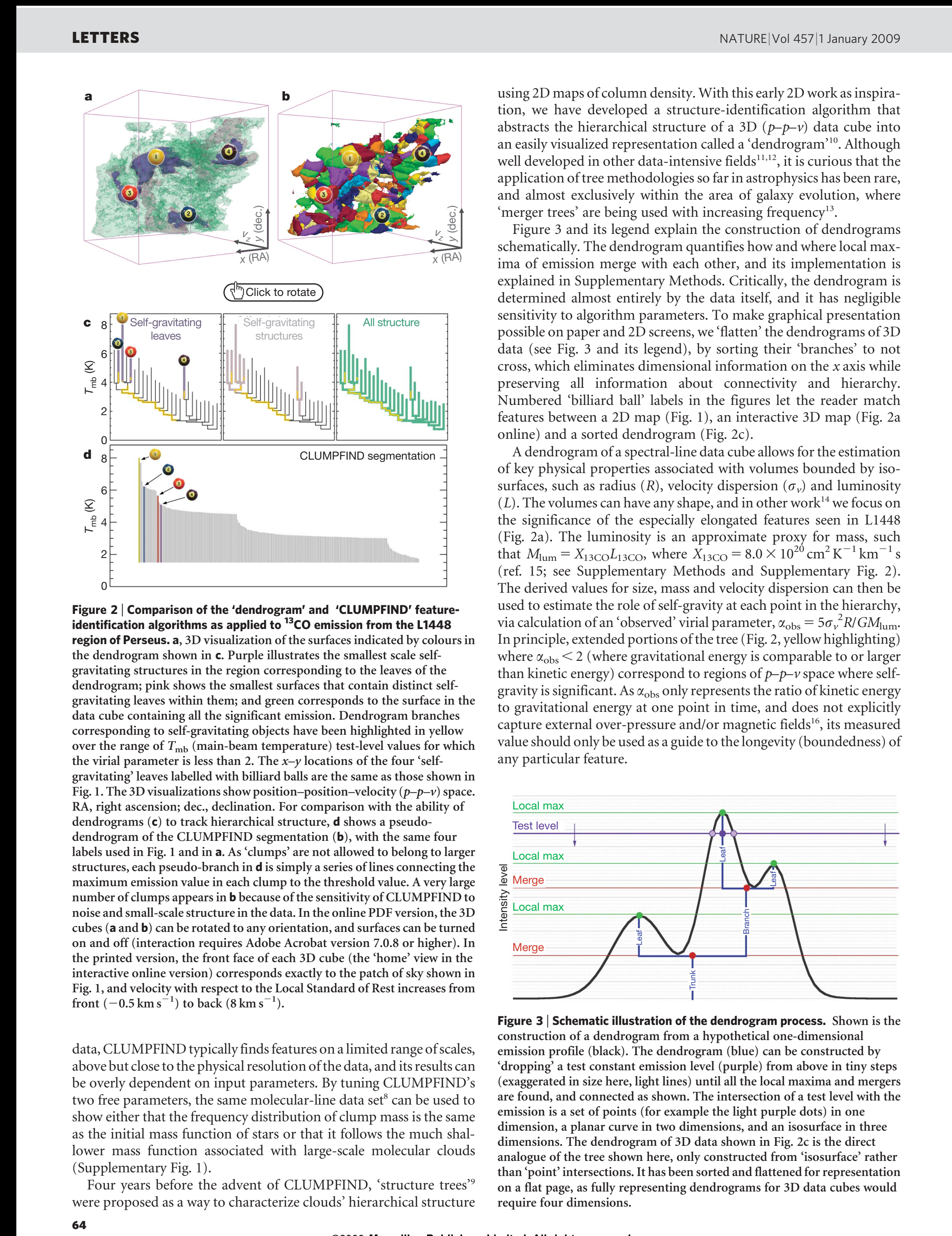


Figure 2 | Comparison of the 'dendrogram' and 'CLUMPFIND' feature-identification algorithms as applied to ^{13}CO emission from the L1448 region of Perseus. a, 3D visualization of the surfaces indicated by colours in the dendrogram shown in **c**. Purple illustrates the smallest scale self-gravitating 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, Rosolowsky, Borkin, Foster, Halle, Kauffmann & Pineda, **Nature**, 2009*