HIGH-DIMENSIONAL DATA VISUALIZATION & THE "PAPER" OF THE FUTURE

Alyssa A. Goodman • Harvard-Smithsonian Center for Astrophysics

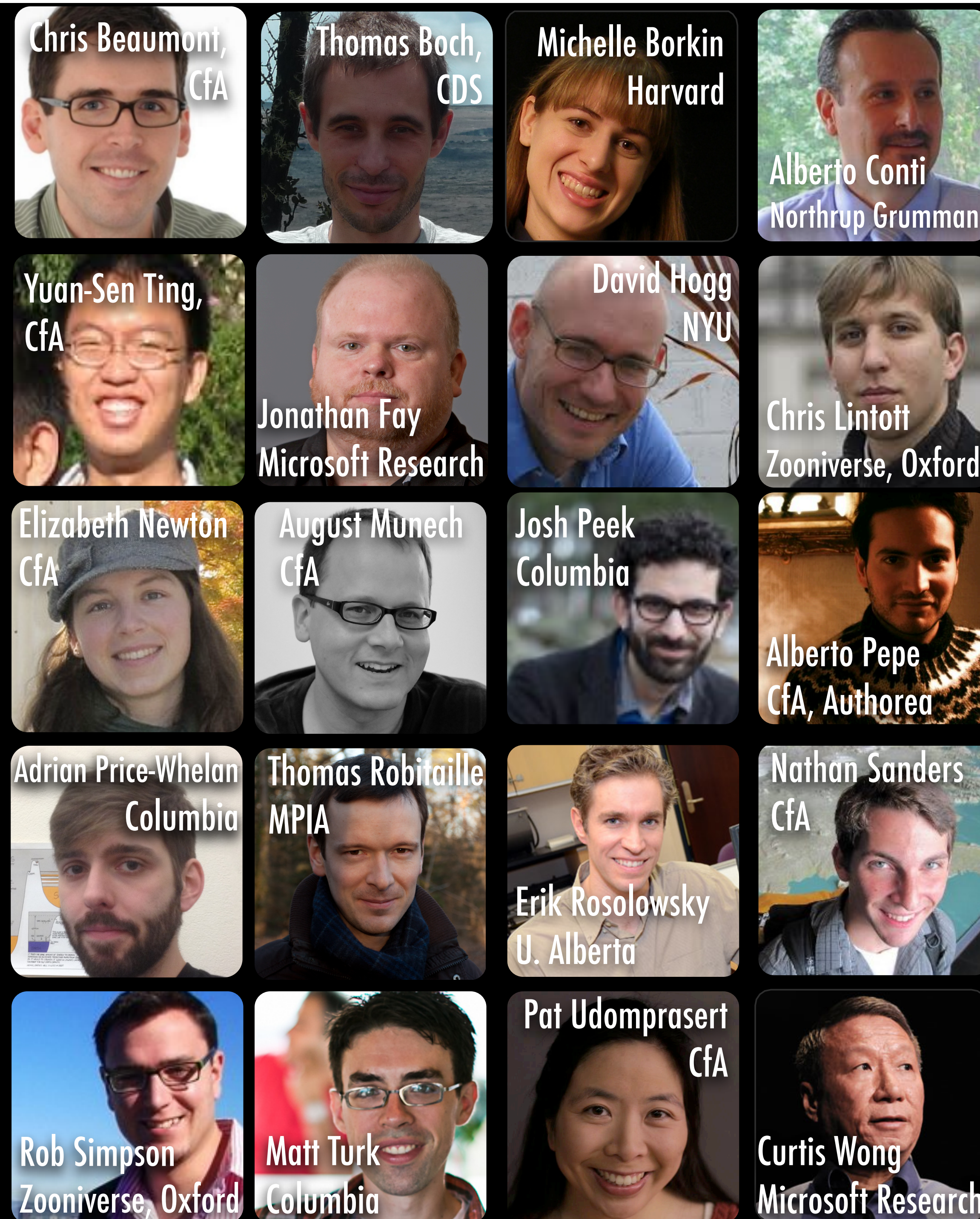
COLLABORATORS

SEAMLESS ASTRONOMY

Linking scientific data, publications, and communities

projects.iq.harvard.edu/seamlessastronomy

...including ADS team (Alberto Accomazzi, Michael Kurtz, Edwin Henneken, et al.) and Wolbach Library staff (Christopher Erdmann et al.)



HIGH-DIMENSIONAL DATA VISUALIZATION & THE "PAPER" OF THE FUTURE

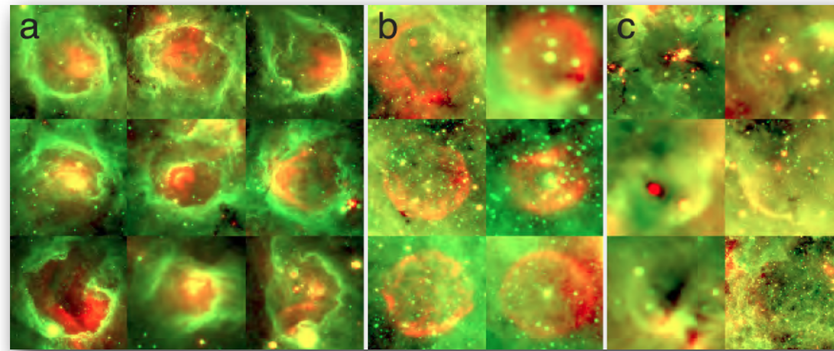
Alyssa A. Goodman • Harvard-Smithsonian Center for Astrophysics

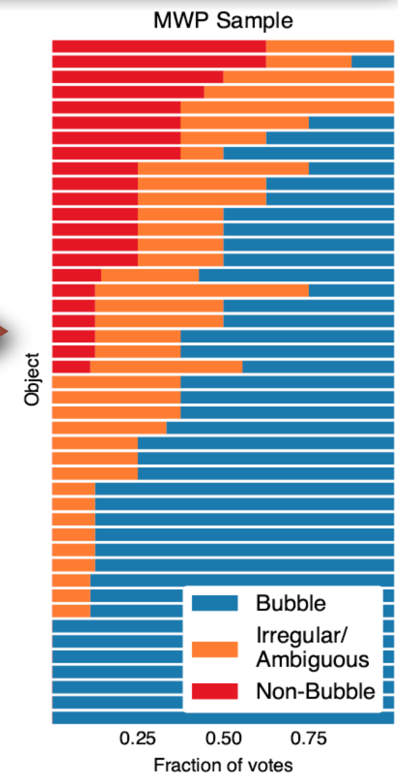
Relative Strengths

Pattern Recognition
Creativity

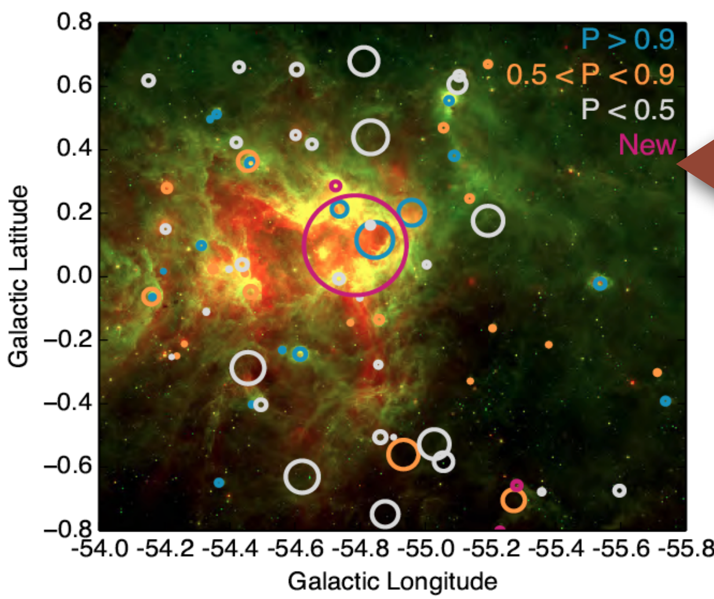
Calculations

HUMAN-AIDED COMPUTING





machine-learning algorithm (Brut)



example here from: ***Beaumont****, Goodman, Kendrew, Williams & Simpson 2014; based on* ***Milky Way Project*** *catalog (Simpson et al. 2013), which came from* ***Spitzer/GLIMPSE*** *(Churchwell et al. 2009, Benjamin et al. 2003), cf. Shenoy & Tan 2008 for discussion of HAC;* ***astroml.org*** *for machine learning advice/tools*

VISUALIZATION RESEARCH TODAY

posted by user: mark_s || 471 views || tracked by 6 users: [display] Add to My List | I'm Organizer | Modify

VIS 2015 : IEEE Visualization: Visual Analytics Science and Technology (VAST), Information Visualization (InfoVis) & Scientific Visualization (SciVis)

Link: http://www.ieeevis.org/

When	Oct 25, 2015 - Oct 30, 2015
Where	Chicago, USA
Submission Deadline	Mar 31, 2015

Call For Papers

TBA

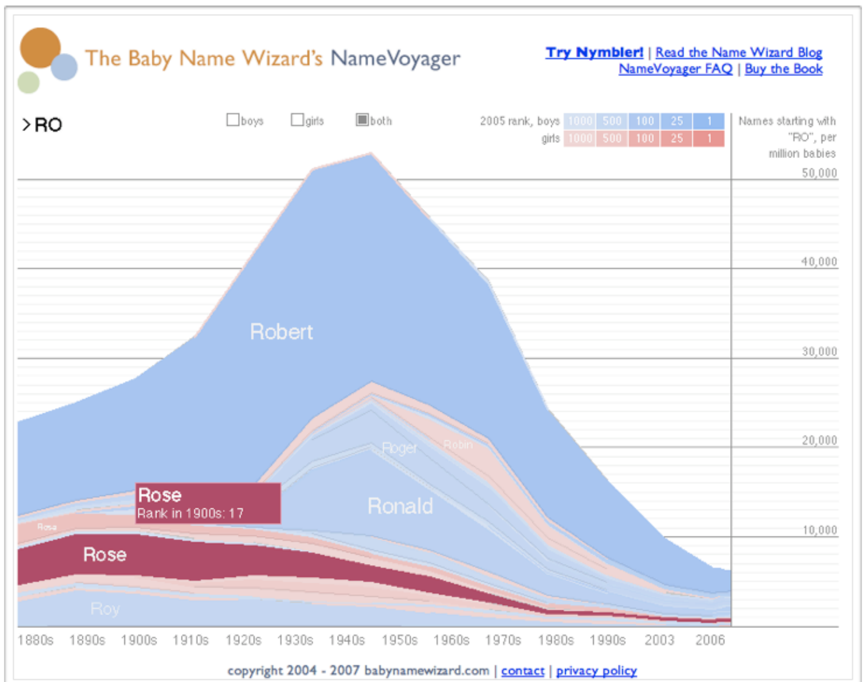
(information from http://infovis-wiki.net/index.php?title=VIS_2015)

"My first reading of your proposal suggests that it is more directed at scientific visualization, computer vision, signal processing and the generation of effective images. My own interests are more on information visualization and visual analytics processes that lead to insights, not images. The description does NOT include mention of multi-variate, time series, tree structures, or networks, which form the bulk of my work. Also, I assume you do not wish to cover infographics designed for the general public."

-excerpt from an email about attending a workshop on "the intersection between human and algorithmic approaches to finding patterns in visual data, within the context of real-world problems" from one prominent visualization researcher to another, 2012

INFOVIS

Information visualization *is the study of (interactive) visual representations of abstract data to reinforce human cognition. The abstract data include both numerical and non-numerical data, such as text and geographic information. However, information visualization differs from scientific visualization: "it's infovis [information visualization] when the spatial representation is chosen, and it's scivis [scientific visualization] when the spatial representation is given".* (Wikipedia, 2014)

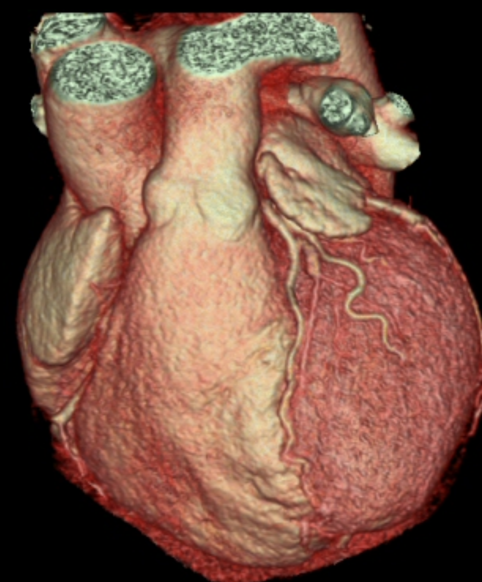


http://www.babynamewizard.com/namevoyager/lnv0105.html

created by Martin Wattenberg

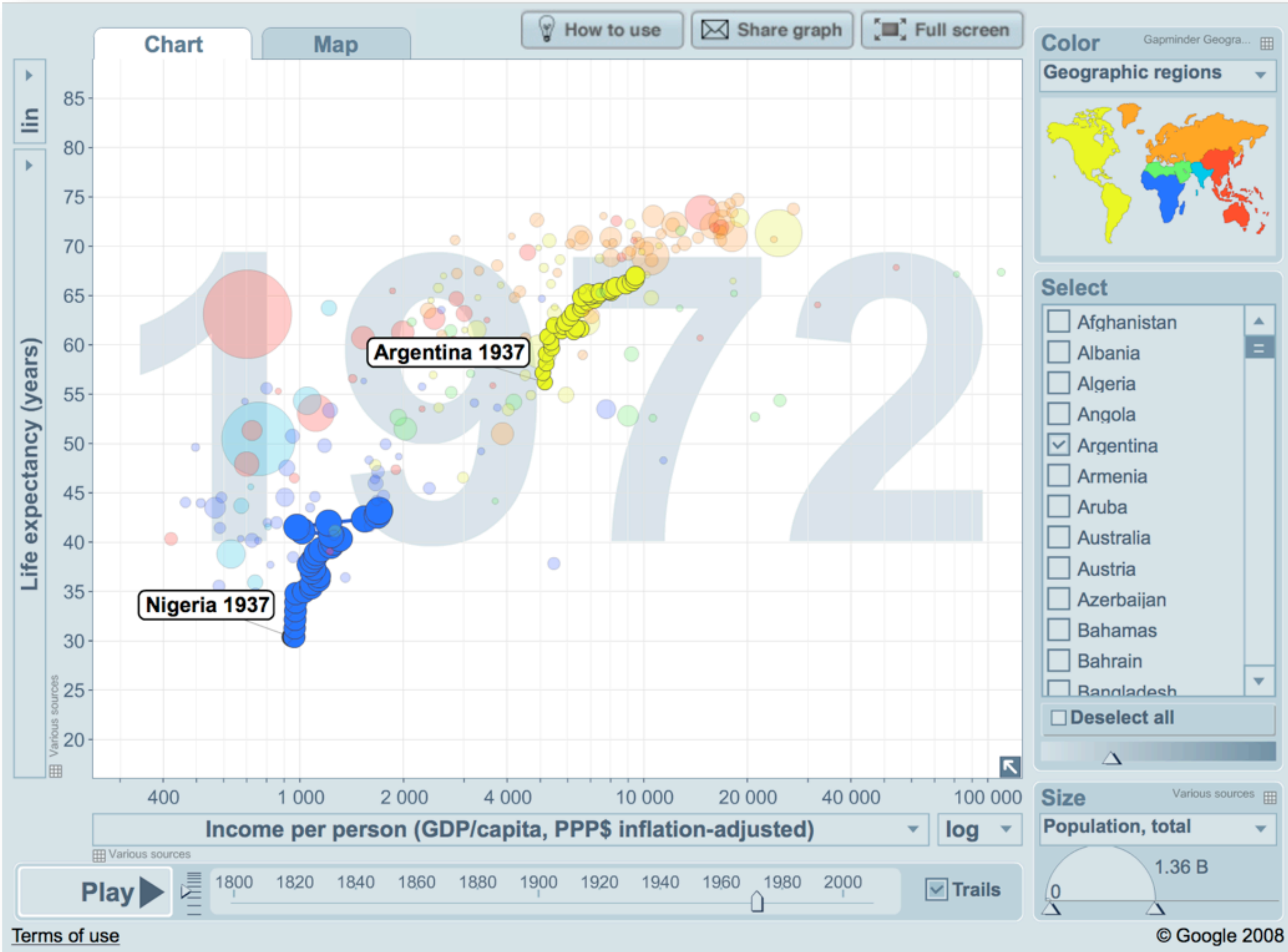
SCIVIS

Scientific visualization is an interdisciplinary branch of science. *According to Friendly (2008), it is "primarily concerned with the visualization of three-dimensional phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic (time) component". It is also considered a branch of computer science that is a subset of computer graphics. The purpose of scientific visualization is to graphically illustrate scientific data to enable scientists to understand, illustrate, and glean insight from their data.* (Wikipedia 2014)



http://www.uphs.upenn.edu/news/News_Releases/sep05/64sliceCT.htm

INFOVIS? SCIVIS? VISUAL ANALYTICS?



GAPMINDER for a fact-based world view

from Hans Rosling, et al.

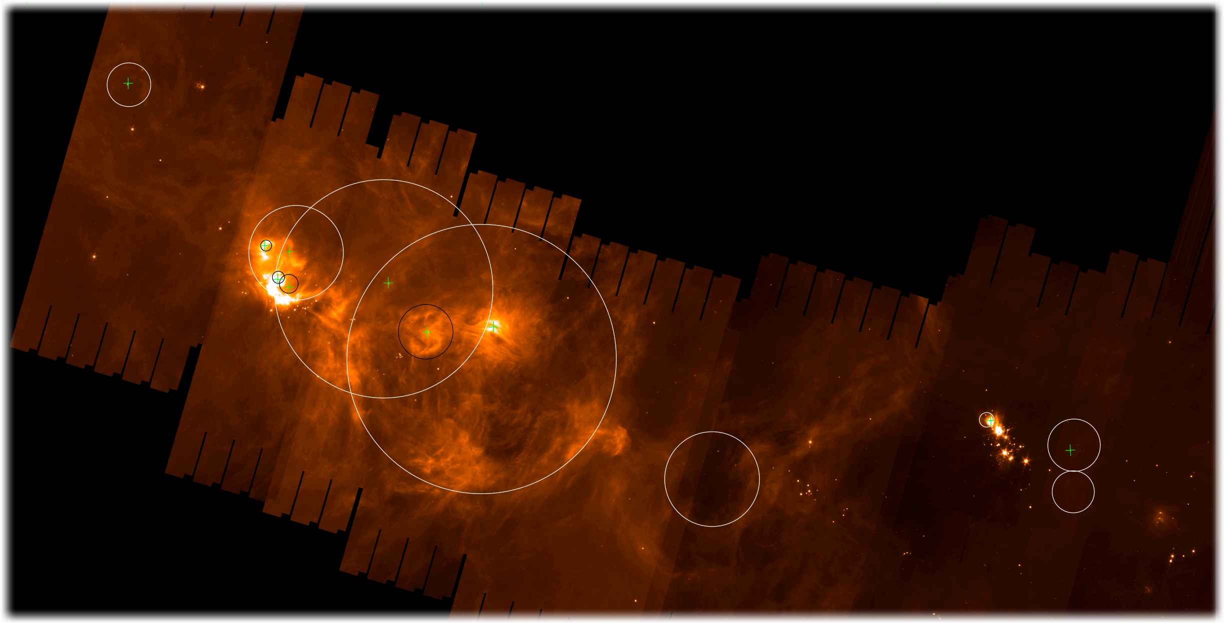
~~INFOVIS? SCIVIS? VISUAL ANALYTICS?~~

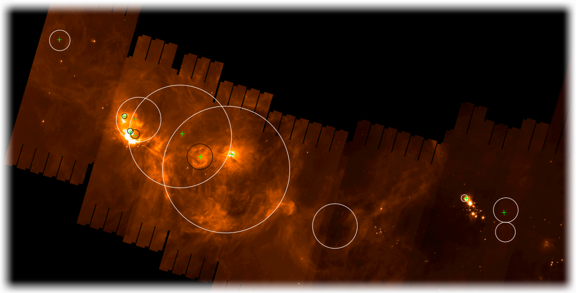
These distinctions need not be made, and they harm progress in research.

DISTINCTIONS AMONGST VIZ TYPES IS UNHELPFUL

e.g. "Shells in Perseus"

Arce et al. 2011, 10.1088/0004-637X/742/2/105





DISTINCTIONS AMONGST VIZ TYPES IS UNHELPFUL

e.g. “Shells in Perseus”

Arce et al. 2011, 10.1088/0004-637X/742/2/105

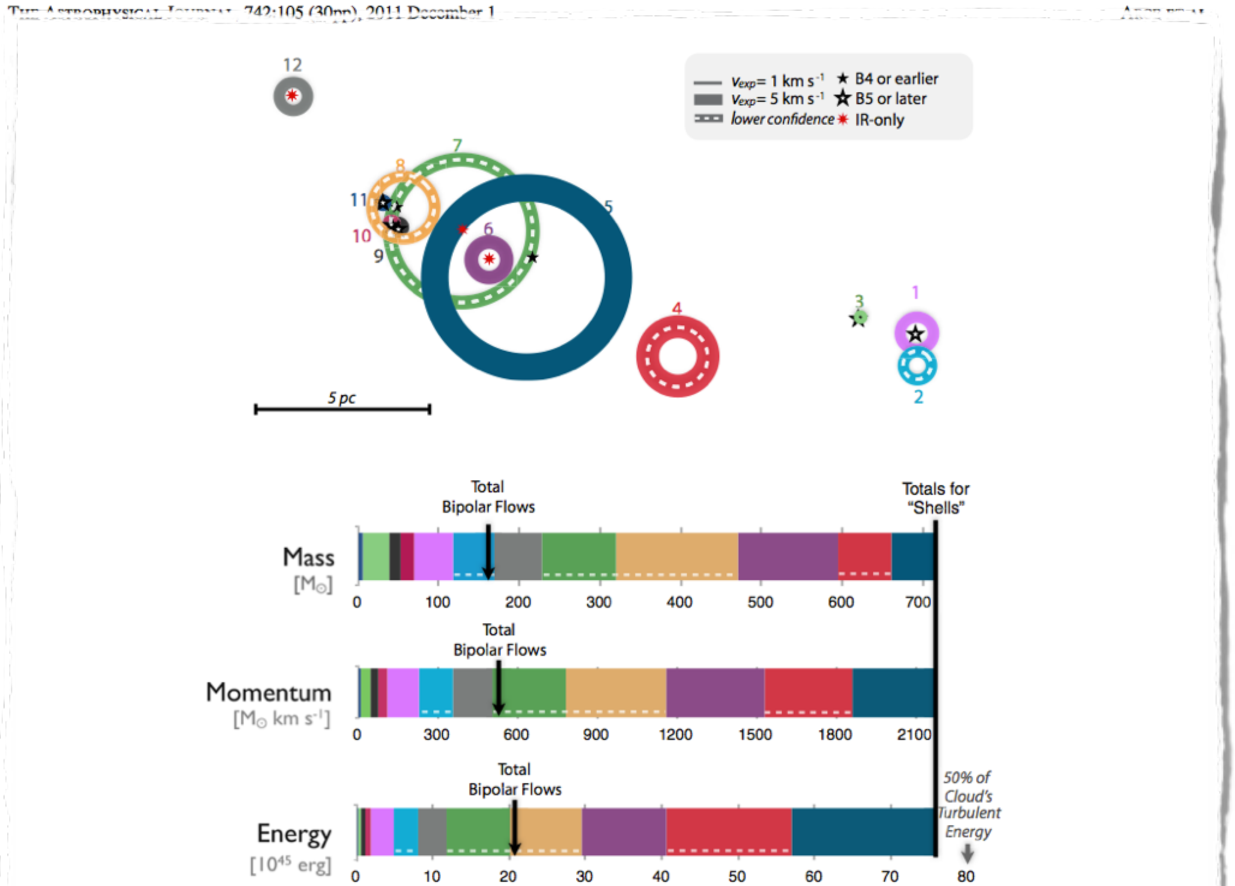


Figure 30. Schematic picture of size, mass, momentum, and energy distribution of shells in Perseus. The radius of each circle (and position) are proportional to the radius (and location) of the shell in the cloud, while the ring thickness is proportional to the expansion velocity (see the legend on the upper right corner). Shells with a confidence level of 3 or less (from Table 4) are indicated by a dashed white line. Candidate powering sources with a B5 spectral type or later are shown as white star symbols, while those with earlier spectral type (i.e., high-mass stars) are shown as black (filled) star symbols. Candidate sources with no known spectral type (but known α) are shown as red stars. The relative mass, momentum, and kinetic energy of the shells are shown in the three horizontal bars (where the colors indicate the value for each shell). The total outflow mass, momentum, and kinetic energy of the molecular outflows in Perseus (from Arce et al. 2010) are shown for comparison.

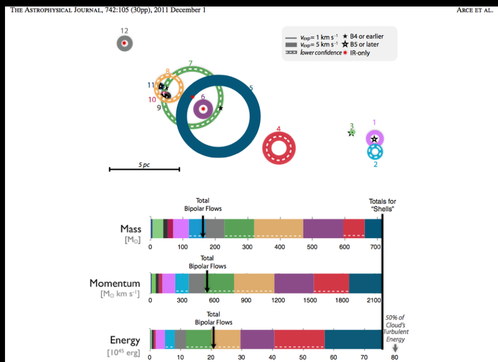
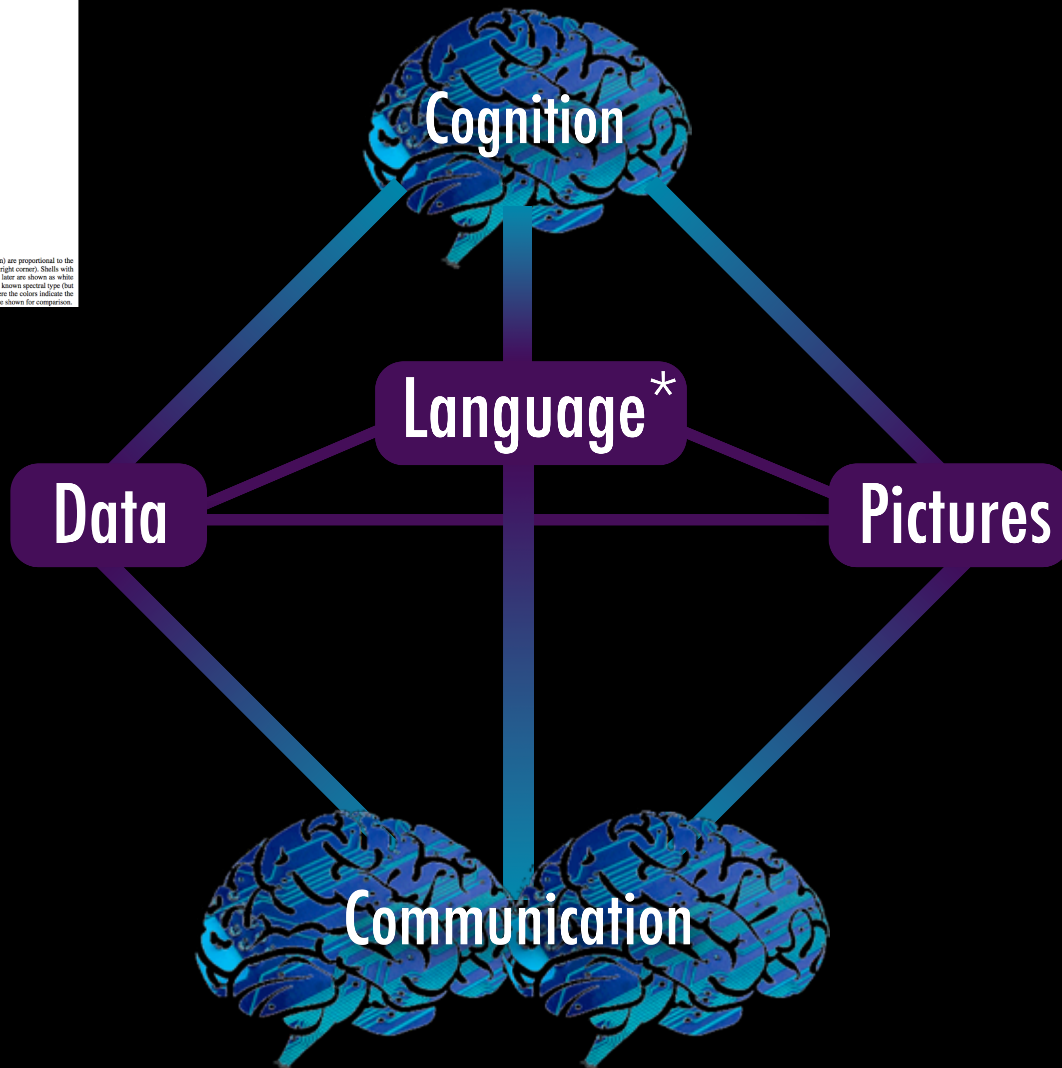


Figure 30. Schematic picture of size, mass, momentum, and energy distribution of shells in Perseus. The radius of each circle (and position) are proportional to the radius (and location) of the shell in the cloud, while the ring thickness is proportional to the expansion velocity (see the legend on the upper right corner). Shells with a confidence level of 3 or less (from Table 4) are indicated by a dashed white line. Candidate powering sources with a B5 spectral type or later are shown as white star symbols, while those with earlier spectral type (i.e., high-mass stars) are shown as black (filled) star symbols. Candidate sources with no known spectral type (but known α) are shown as red stars. The relative mass, momentum, and kinetic energy of the shells are shown in the three horizontal bars (where the colors indicate the value for each shell). The total outflow mass, momentum, and kinetic energy of the molecular outflows in Perseus (from Arce et al. 2010) are shown for comparison.



*"Language" includes words & math

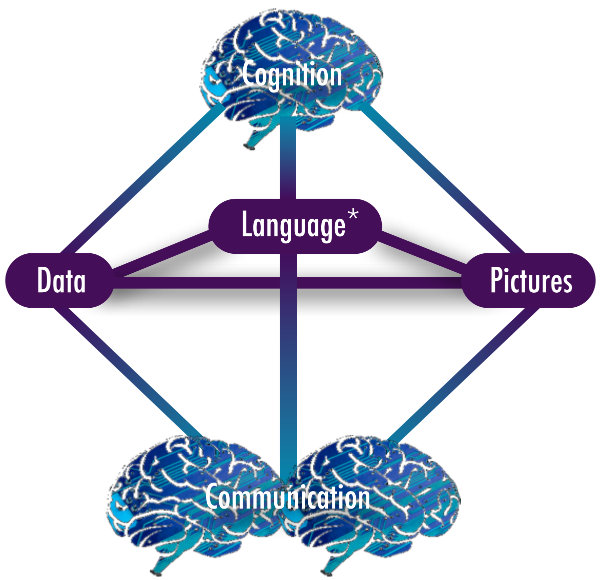
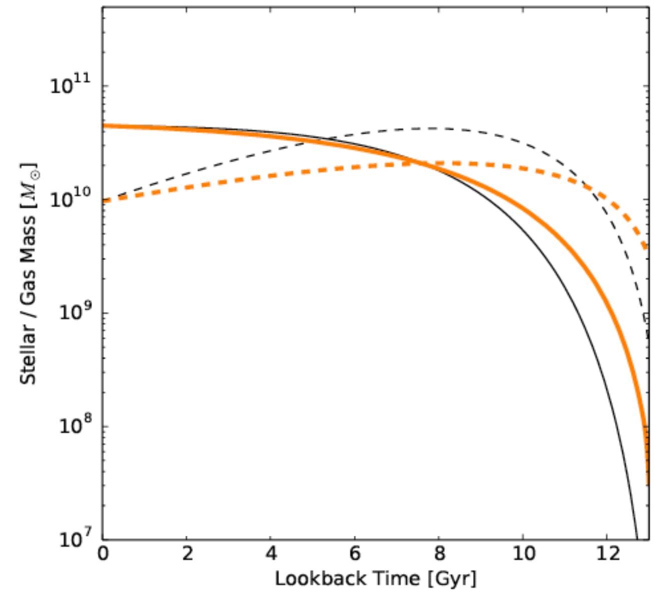
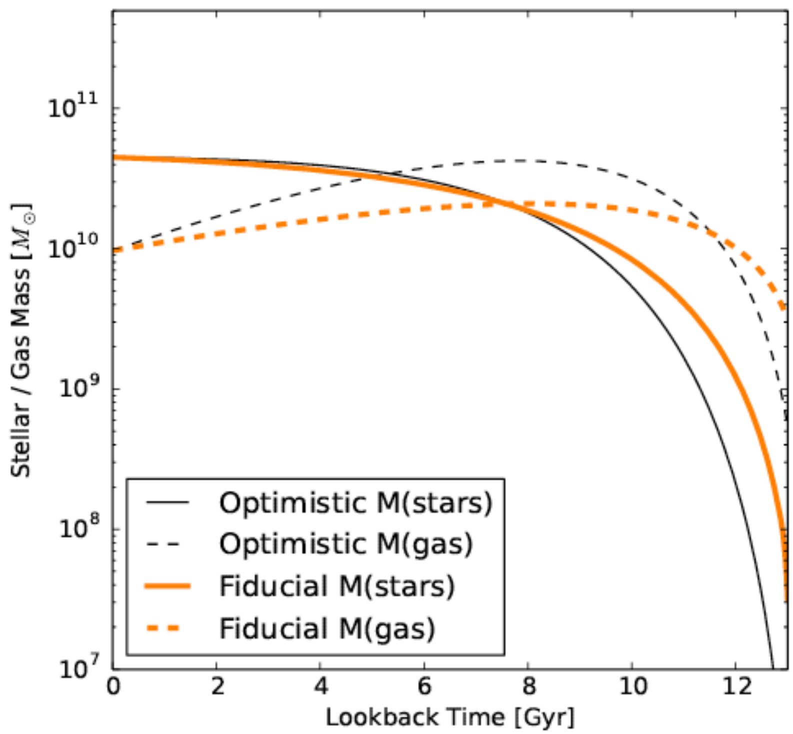
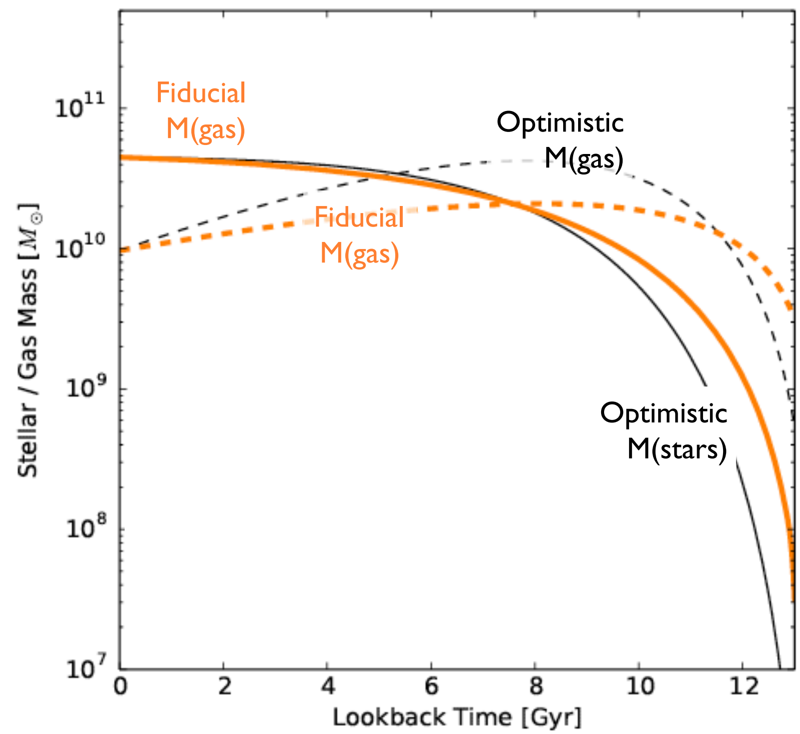


Figure Caption: The solid/solid black line shows the optimistic case for M(stars)/M(gas). The orange lines show the same quantities, for the fiducial case.







HIGH-DIMENSIONAL DATA VISUALIZATION & THE "PAPER" OF THE FUTURE

glue

multidimensional data exploration

Au

Authorea

The Visualization of Astrophysical Data:

Bringing together Science, Art and Education

The visualization of real astrophysical data sets is a powerful tool for communicating science to the public and for teaching. The confluence of high quality data sets such as SDSS & WMAP, advances in computational techniques, and the continued march of Moore's law has enabled the use of stunningly beautiful and scientifically accurate images, animations, and interactives in a variety of settings (e.g. TV programs, museums, websites, digital planetaria, magazines, & undergraduate classrooms). Two dimensional, three dimensional, and hyper-dimensional (e.g. color coded 3D data) representations convey large amounts of information in a visceral fashion that can inform both experts and the public. As the data and techniques have progressed the boundaries between art and science have begun to blur and move towards research. This workshop will bring together astrophysicists, visualizers, and educators to discuss the current status and to debate the future direction of astronomical visualization as a tool for research, education, and public outreach.

Organizers: Randy Landsberg, Josh Frieman, Andrew Hamilton (CU Boulder), Andrey Kravtsov, Mark SubbaRao, & Alex Szalay (JHU).

2005 IN CHICAGO!

thanks Randy, Mark & Andrey

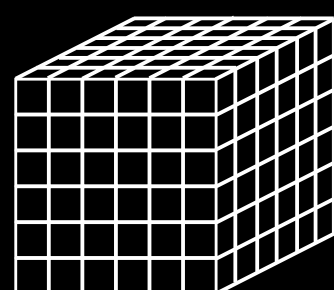
2014 IN CHICAGO!

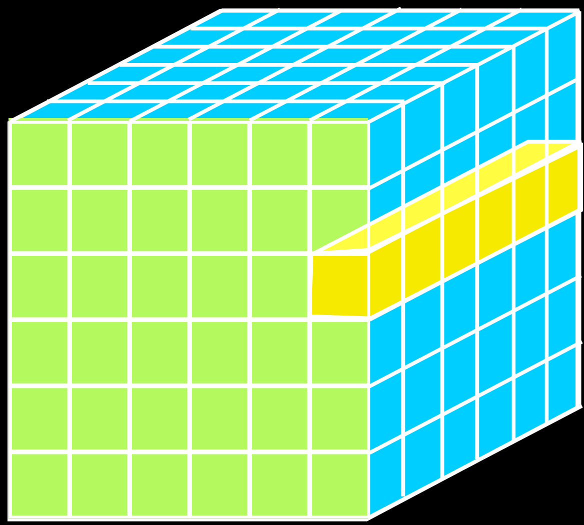
.astronomy

glue

INFOVIS? SCIVIS? VISUAL ANALYTICS?

These distinctions need not be made,
and they harm progress in research,
by ignoring the high-dimensional nature of many data sets.

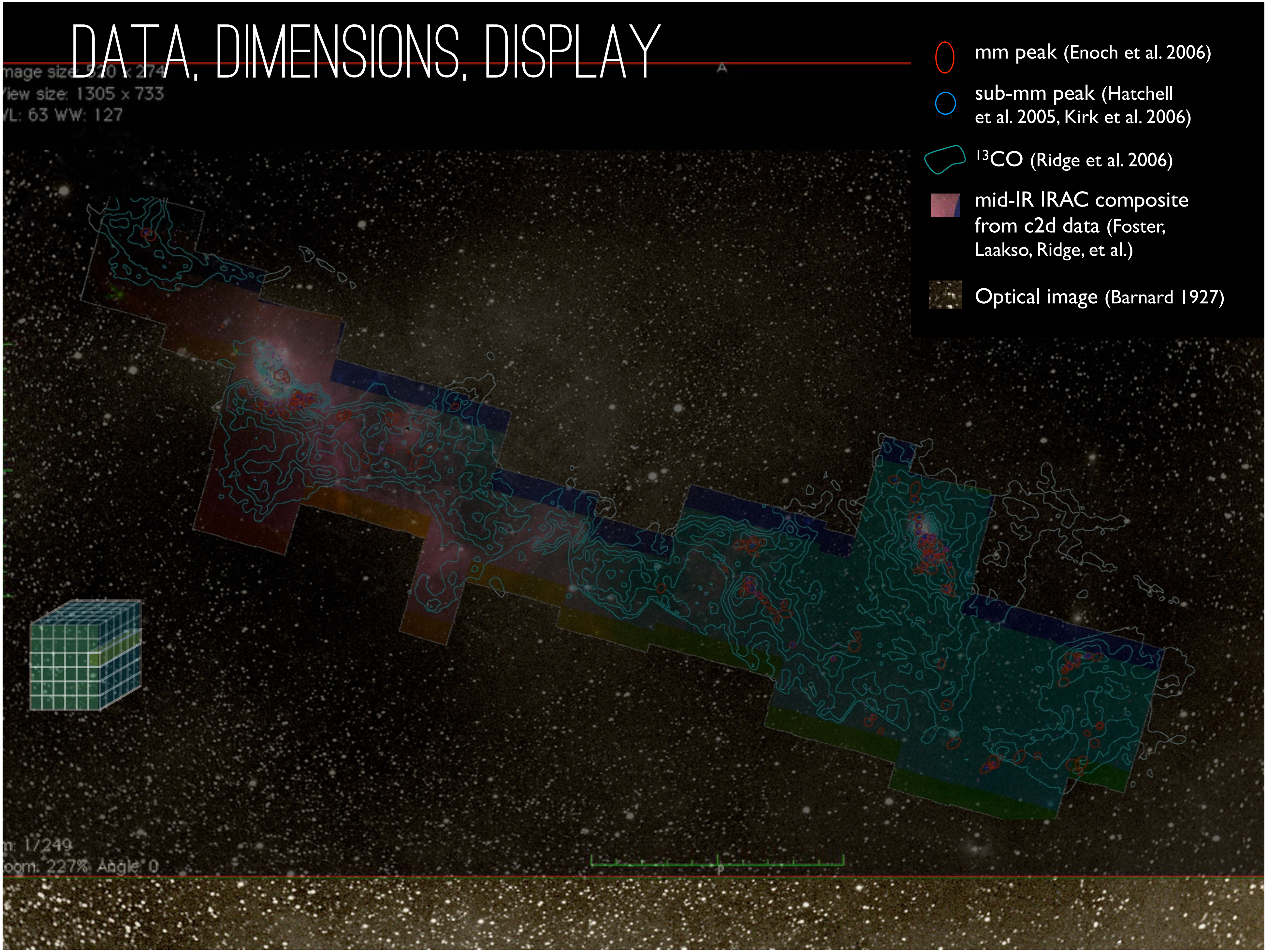




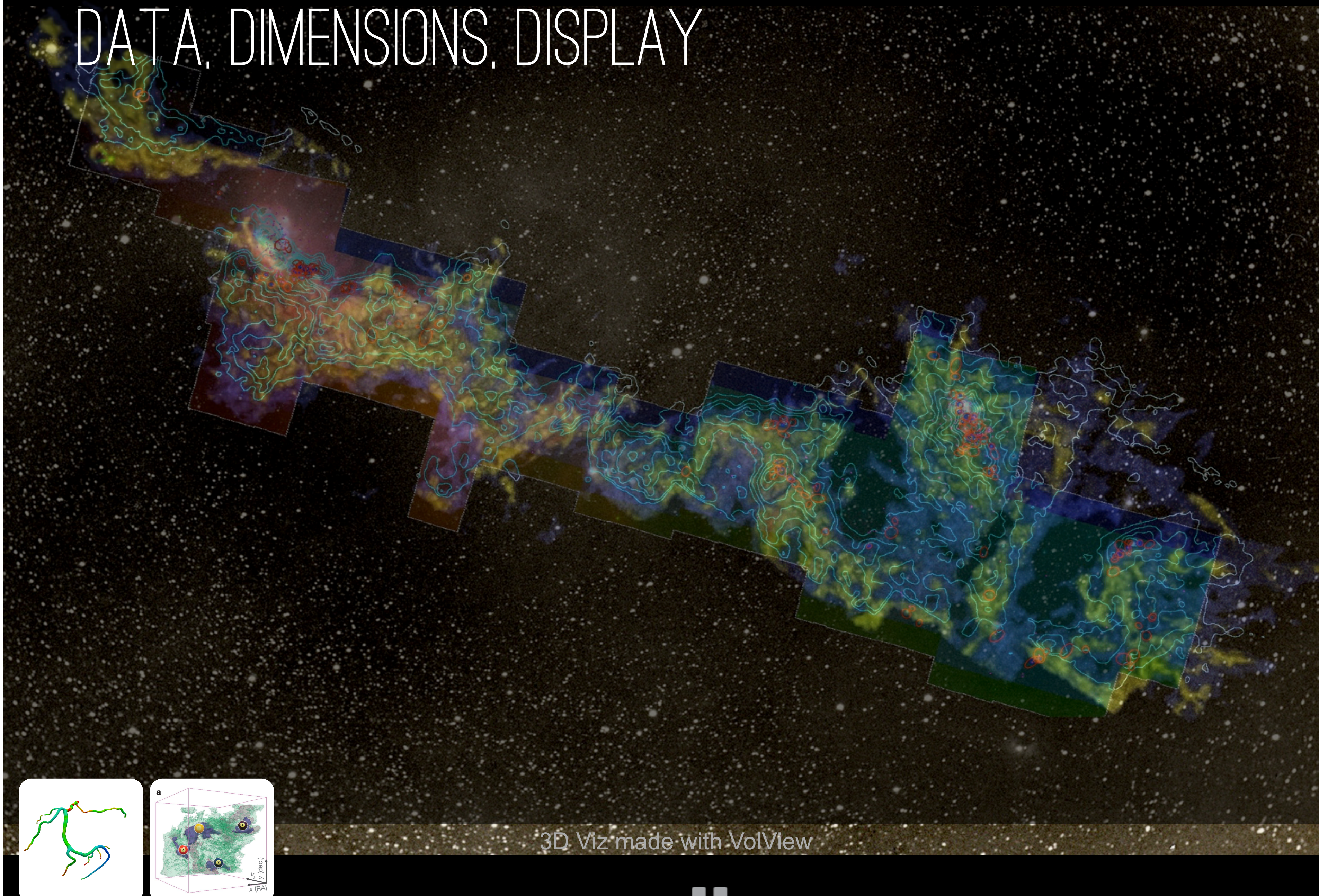
"DATA, DIMENSIONS, DISPLAY"

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"

DATA, DIMENSIONS, DISPLAY



DATA, DIMENSIONS, DISPLAY

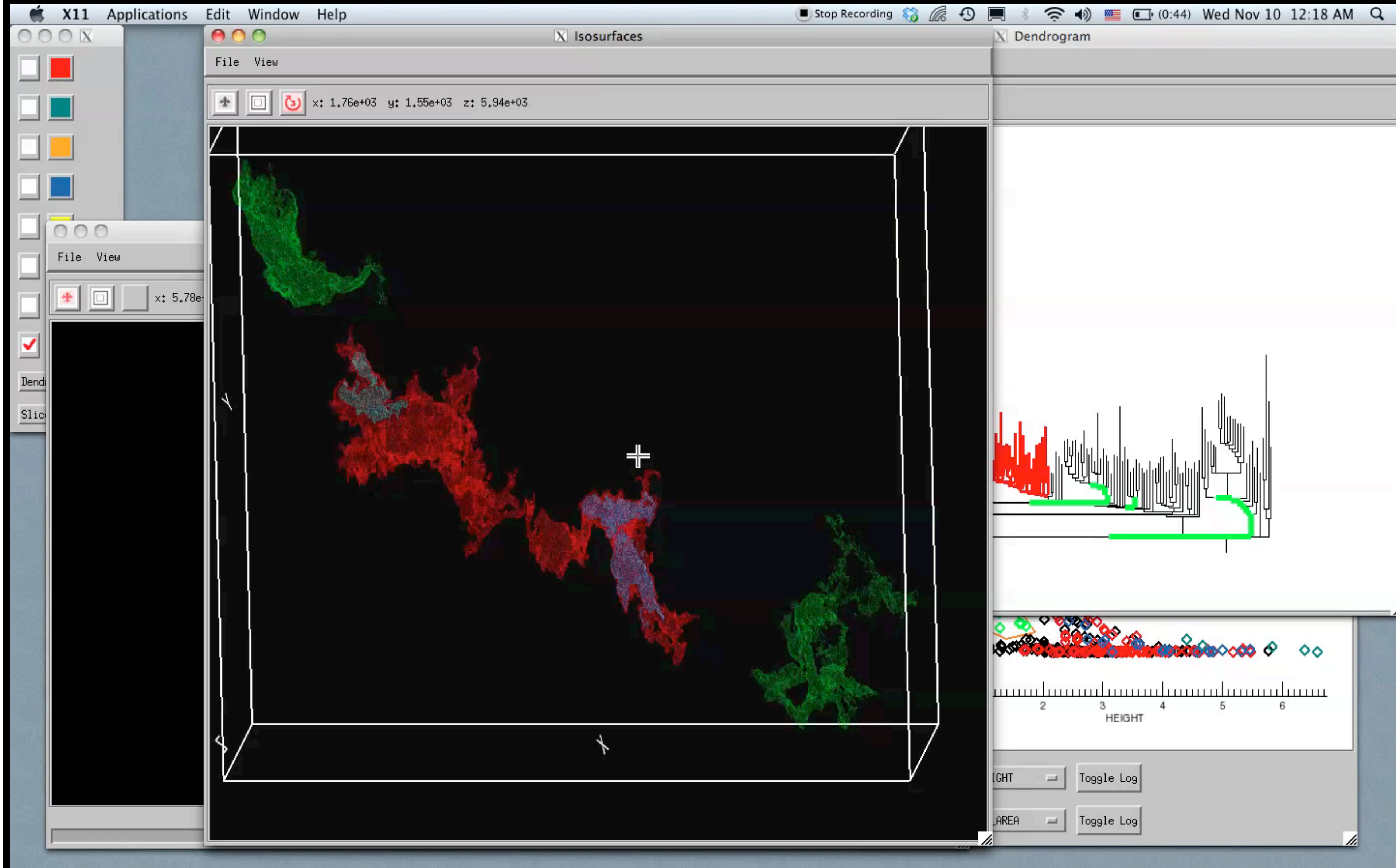


3D Viz made with VolView

AstronomicalMedicine@IIC

COMPLETE

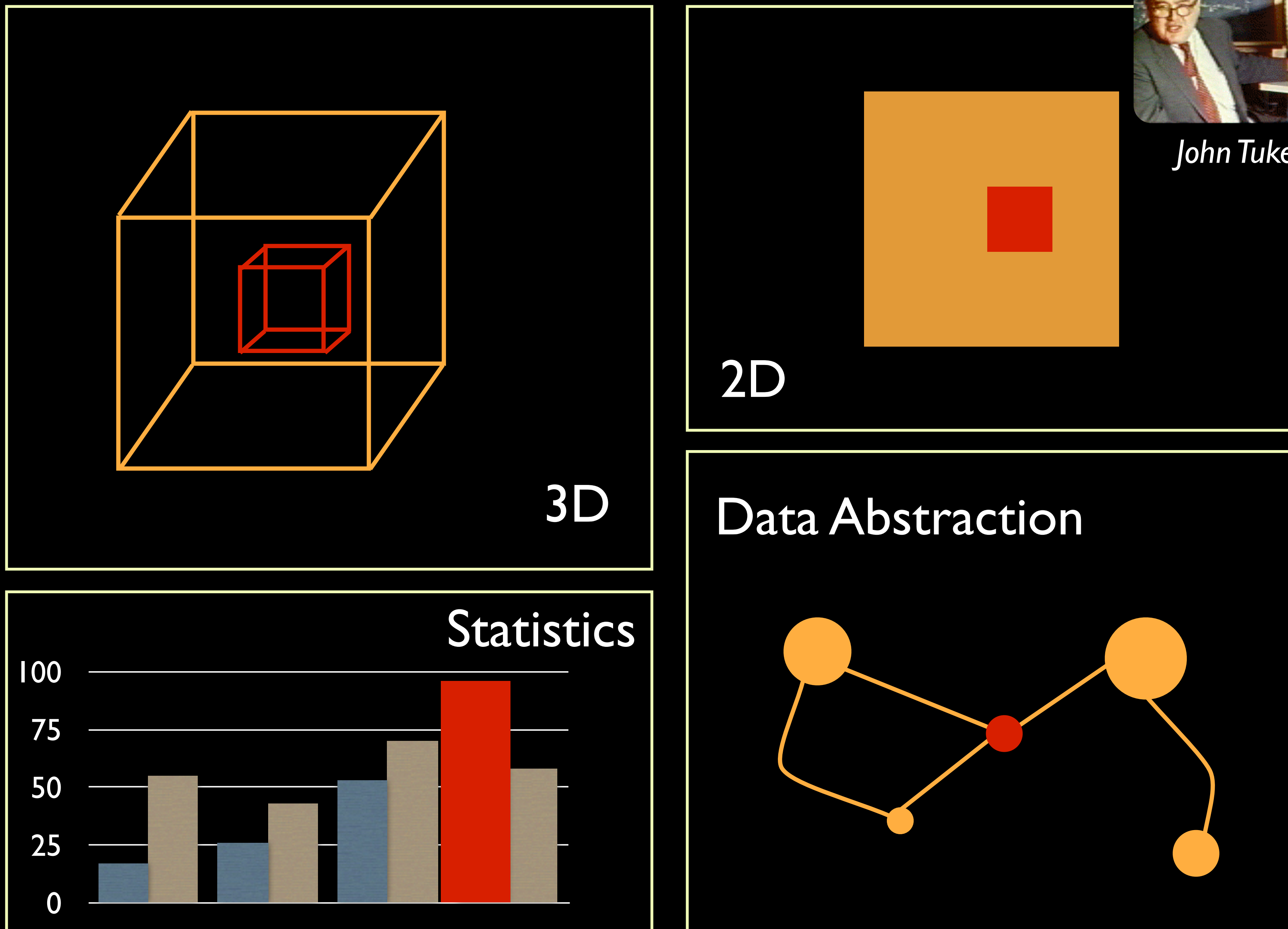
INFOVIS? SCIVIS? VISUAL ANALYTICS?



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

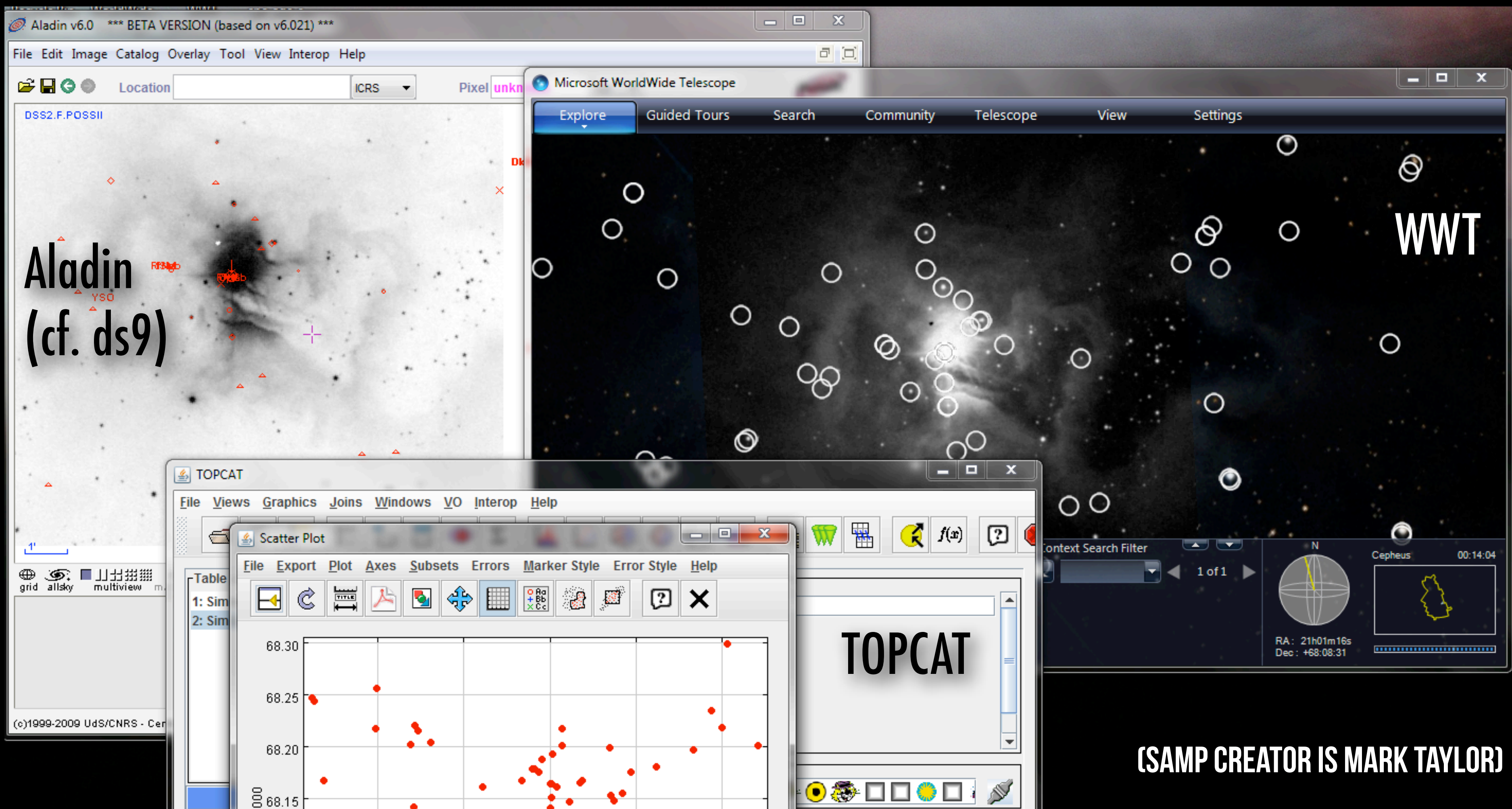
LINKED VIEWS OF HIGH-DIMENSIONAL DATA

John Tukey



figure, by M. Borkin, reproduced from Goodman 2012, "Principles of High-Dimensional Data Visualization in Astronomy"

LINKED VIEWS OF HIGH-DIMENSIONAL DATA "SAMP"

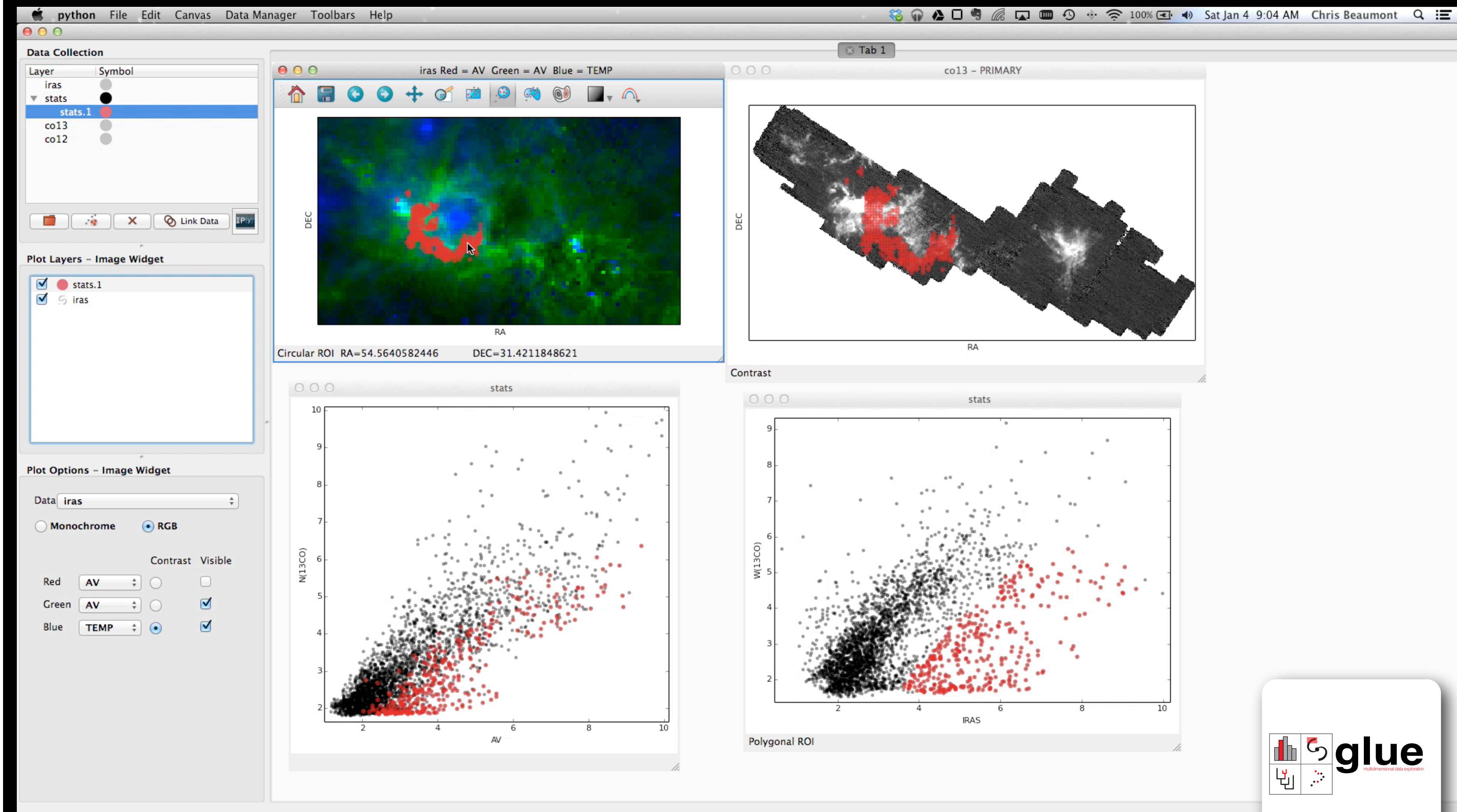


(SAMP CREATOR IS MARK TAYLOR)

figure, showing SAMP screenshot, reproduced from Goodman 2012, "Principles of High-Dimensional Data Visualization in Astronomy"

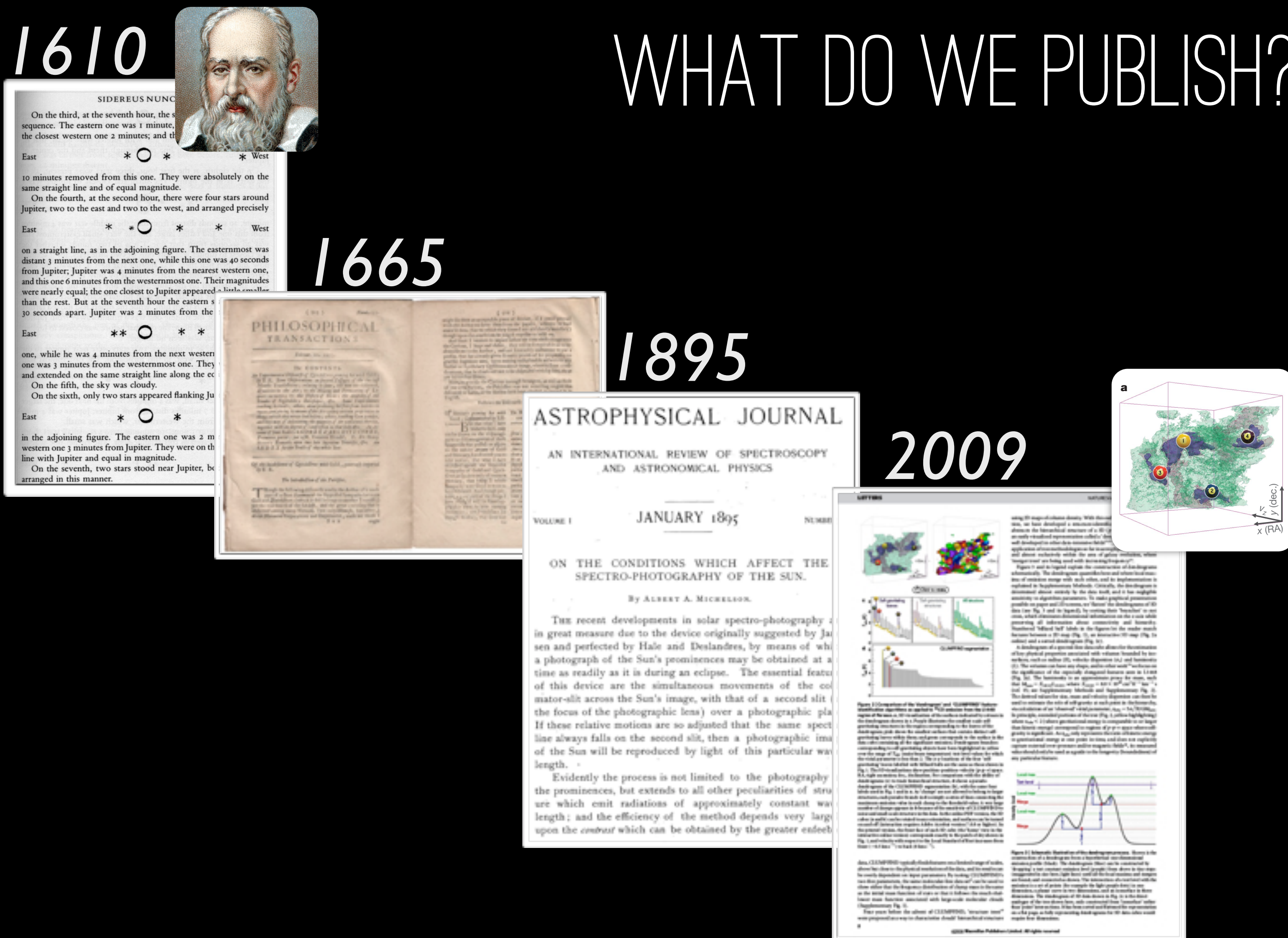
LINKED VIEWS OF HIGH-DIMENSIONAL DATA

GLUE



Beaumont, w/Goodman, Robitaille & Borkin

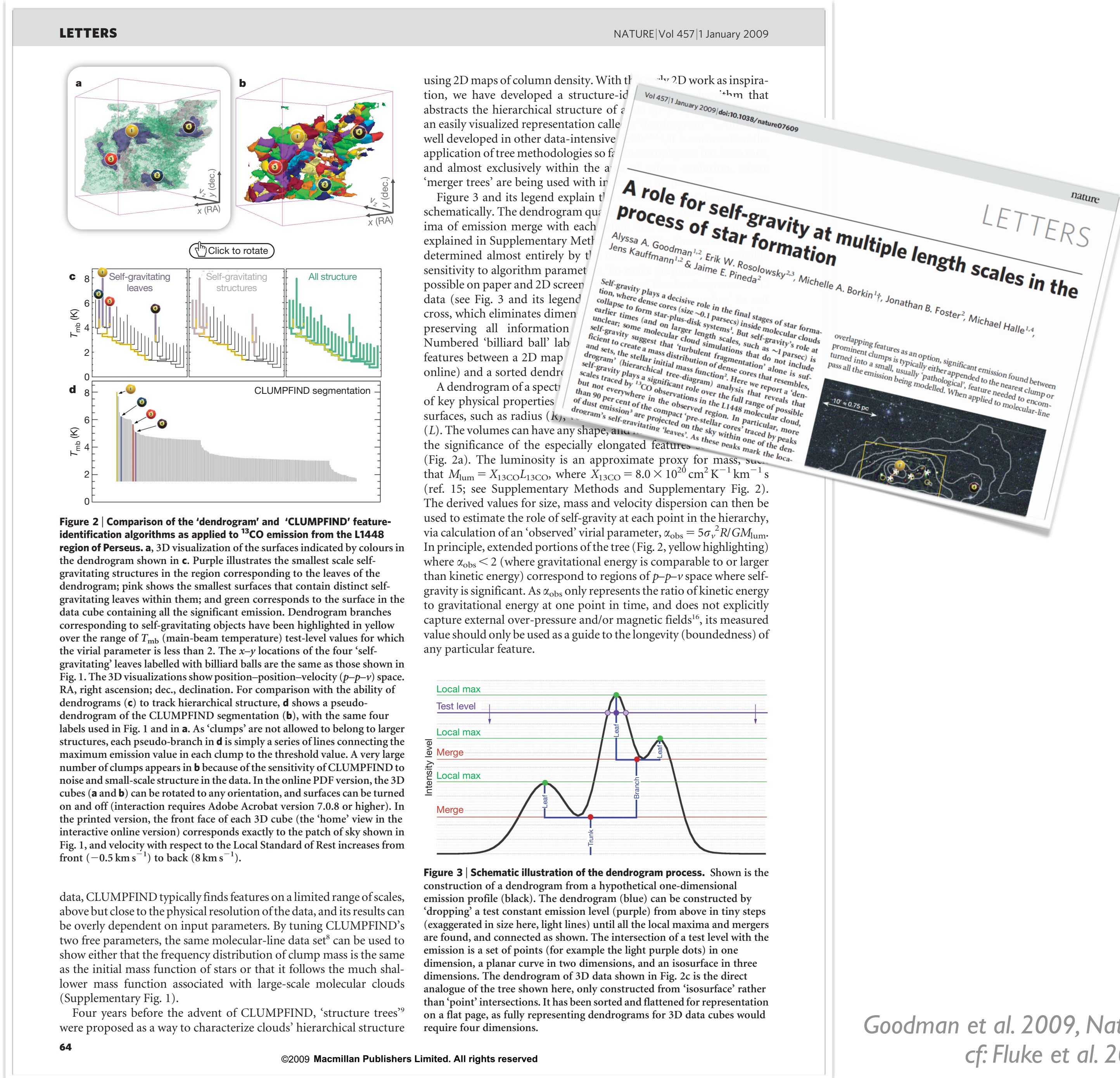
WHAT DO WE PUBLISH?



2009

3D PDF

INTERACTIVITY IN A "PAPER"



LETTERS

NATURE|Vol 457|1 January 2009

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 th... 2D work as inspiration, we have developed a structure-id... ...thm that abstracts the hierarchical structure of a... an easily visualized representation calle... well developed in other data-intensive... application of tree methodologies so fa... and almost exclusively within the a... 'merger trees' are being used with in...

Figure 3 and its legend explain t... schematically. The dendrogram qua... ima of emission merge with each... explained in Supplementary Meth... determined almost entirely by t... sensitivity to algorithm paramet... possible on paper and 2D screen... data (see Fig. 3 and its legend... cross, which eliminates dimen... preserving all information... Numbered 'billiard ball' lab... features between a 2D map... online) and a sorted dendr...

A dendrogram of a spect... of key physical properties... surfaces, such as radius (R)... (L). The volumes can have any shape, and... the significance of the especially elongated features... (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.

64

©2009 Macmillan Publishers Limited. All rights reserved

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

nature

LETTERS

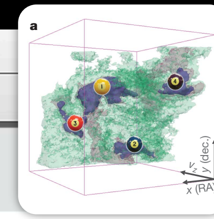
A role for self-gravity at multiple length scales in the process of star formation

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

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function[2]. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by ^{13}CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission[3] are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the loca-

overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

Goodman et al. 2009, Nature, cf: Fluke et al. 2009



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

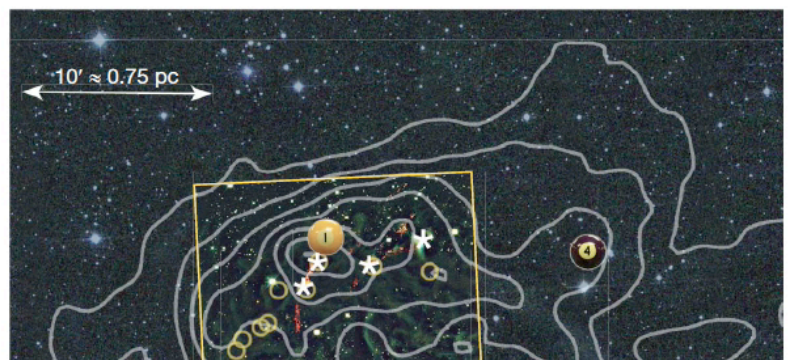
LETTERS

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

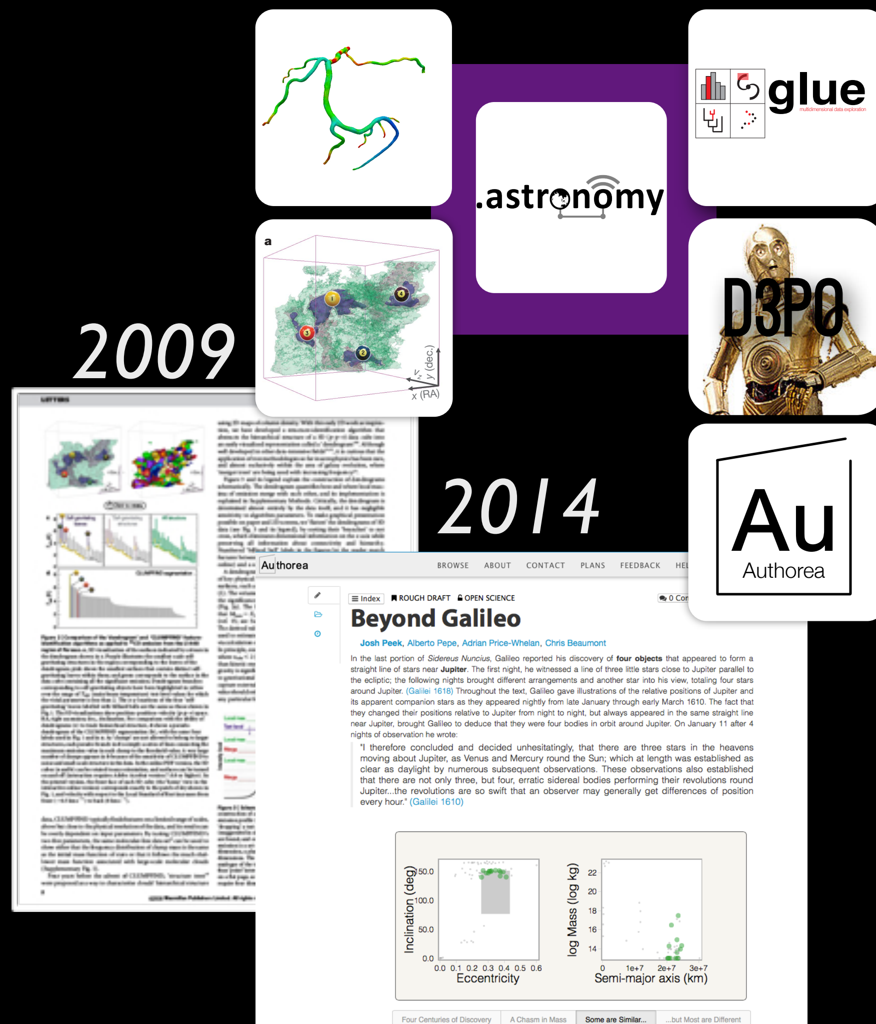
Alyssa A. Goodman[1,2], Erik W. Rosolowsky[2,3], Michelle A. Borkin[1]†, Jonathan B. Foster[2], Michael Halle[1,4], Jens Kauffmann[1,2] & Jaime E. Pineda[2]

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function[2]. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by ^{13}CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission[3] are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their exist-

overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

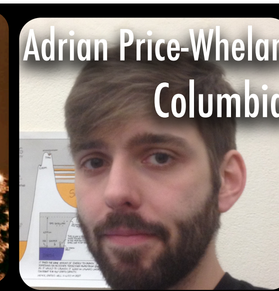


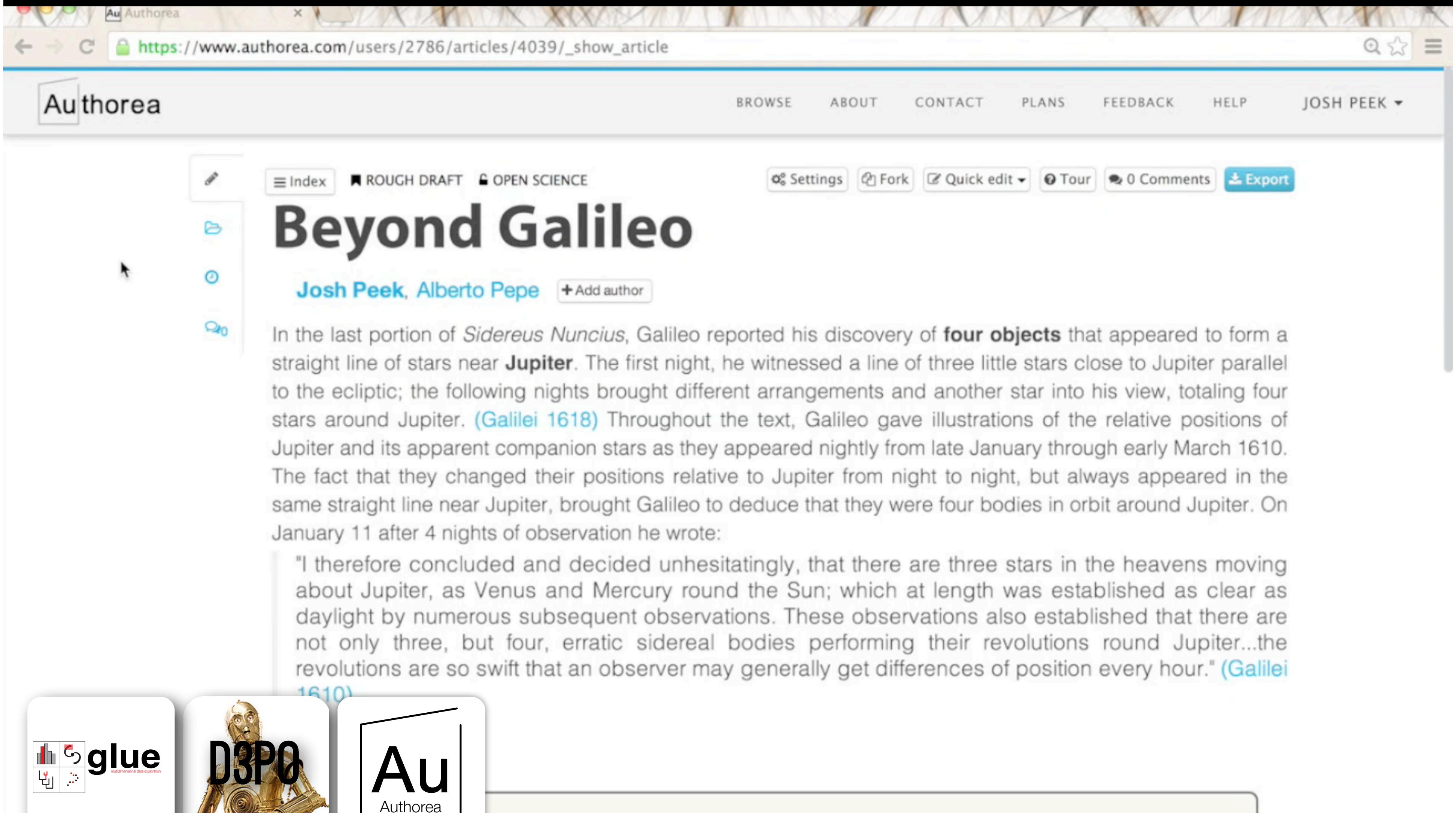
TOWARD
THE
PAPER OF THE FUTURE



"THE STORY & THE SANDBOX"

(GLUE:D3PO:AUTHOREA)

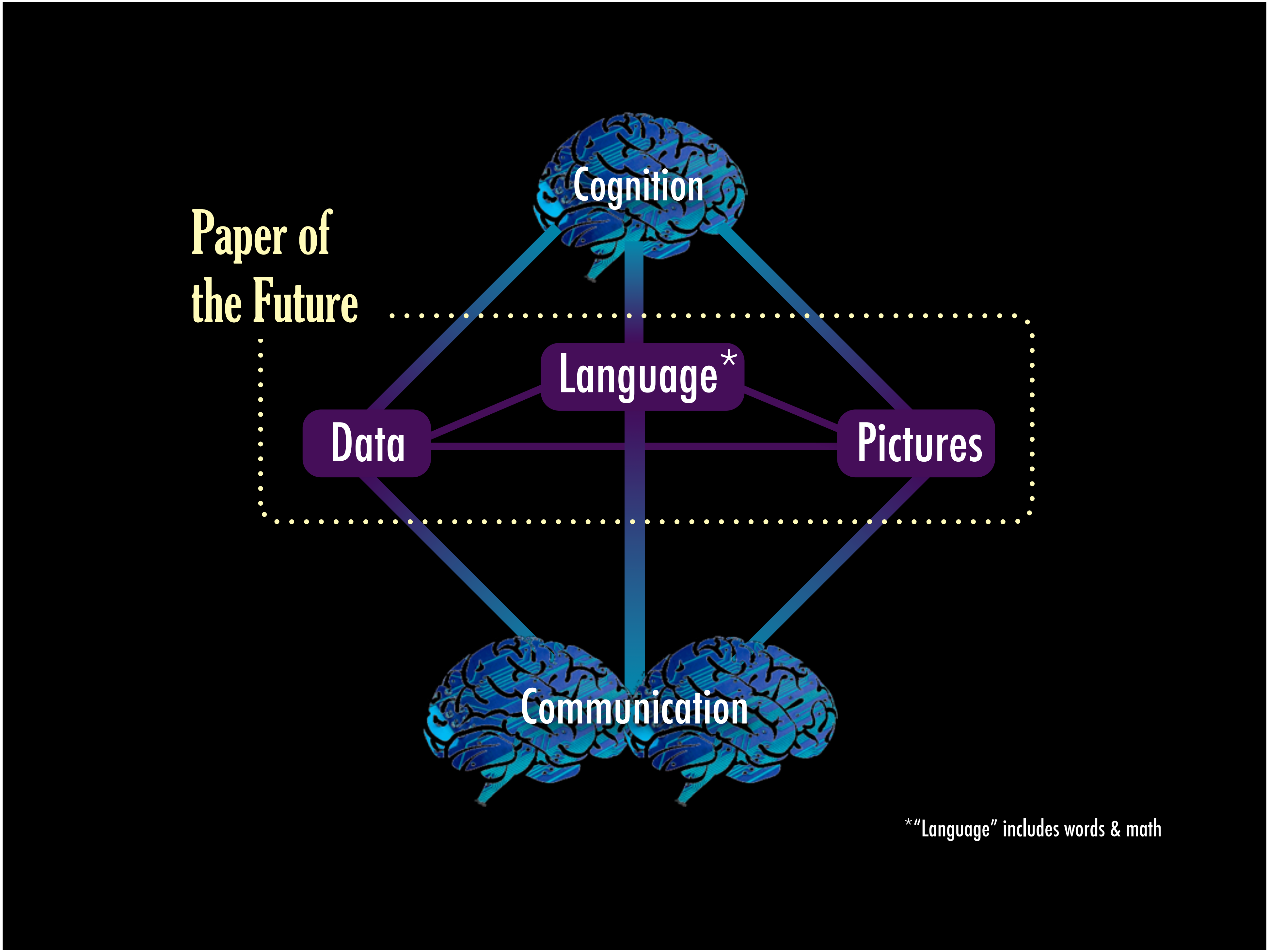




glue

D3PO

Au
Authorea


Paper of the Future
Cognition
Language*
Data
Pictures
Communication
*"Language" includes words & math

Au
Authorea

FEATURED ARTICLES ABOUT PLANS BLOG FEEDBACK HELP ALYSSA GOODMAN

PUBLIC ROUGH DRAFT Index Settings Fork Quickedit Word Count 42 Comments Export Unfollow

The "Paper" of the Future

3

Alyssa Goodman, Josh Peek, Alberto Accomazzi, Chris Beaumont, Christine L. Borgman, How-Huan Hope Chen, Merce Crosas, Christopher Erdmann, August Muench, Alberto Pepe, Curtis Wong + Add author Re-arrange authors

A 5-minute video demonsration of this paper is available at this YouTube link.

2

1 Preamble

A variety of research on human cognition demonstrates that humans learn and communicate best when more than one processing system (e.g. visual, auditory, touch) is used. And, related research also shows that, no matter how technical the material, most humans also retain and process information best when they can put a narrative "story" to it. So, when considering the future of scholarly communication, we should be careful not to do blithely away with the linear narrative format that articles and books have followed for centuries: instead, we should enrich it.

Much more than text is used to commuicate in Science. Figures, which include images, diagrams, graphs, charts, and more, have enriched scholarly articles since the time of Galileo, and ever-growing volumes of data underpin most scientific papers. When scientists communicate face-to-face, as in talks or small discussions, these figures are often the focus of the conversation. In the best discussions, scientists have the ability to manipulate the figures, and to access underlying data, in real-time, so as to test out various what-if scenarios, and to explain findings more clearly. **This short article explains—and shows with demonstrations—how scholarly "papers" can morph into long-lasting rich records of scientific discourse**, enriched with deep data and code linkages, interactive figures, audio, video, and commenting.

0

Konrad Hinsen 3 days ago · Public
Many good suggestions, but if the goal is "long-lasting rich records of scientific discourse", a more careful and critical attitude towards electronic artifacts is appropriate. I do see it concerning videos, but not a word on the much more critical situation in software. Archiving source code is not sufficient: all the dependencies, plus the complete build environment, would have to be conserved as well to make things work a few years from now. An "executable figure" in the form of an IPython notebook wil...
more

Merce Crosas 3 days ago · Public
Konrad, good points; this has been a concern for the community working on reproducibility. Regarding data repositories, Dataverse handles long-term preservation and access of data files in the following way: 1) for some data files that the repository recognizes (such as R Data, SPSS, STATA), which depend on a statistical package, the system converts them into a preservation format (such as a tab/CSV format). Even though the original format is also saved and can be accessed, the new preservation format gua...
more

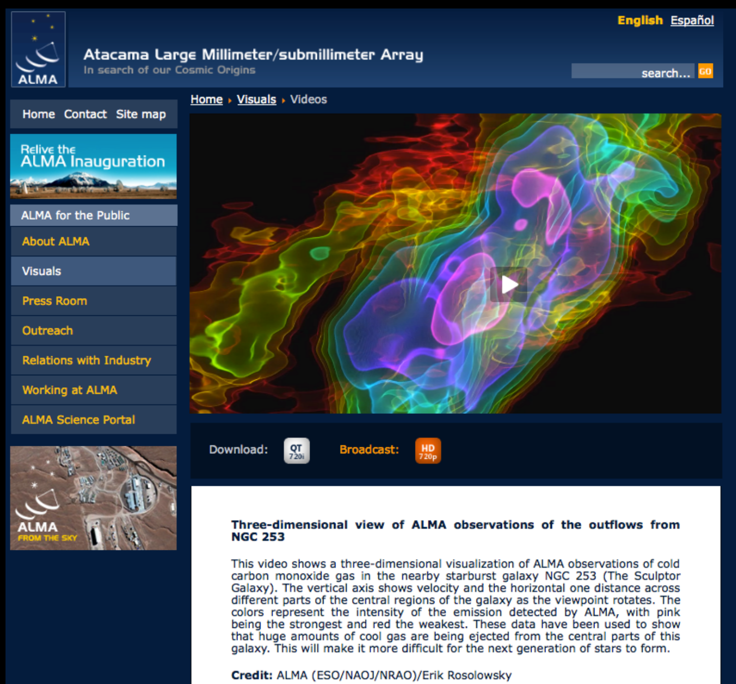
Konrad Hinsen 1 day ago · Public
That sounds good. I hope more repositories will follow the example of Dataverse. Figshare in particular has a very different attitude, encouraging researchers to deposit as much as possible. That's perhaps a good strategy to change habits, but in the long run it could well backfire when people find out in a few years that 90% of those deposits have become useless.

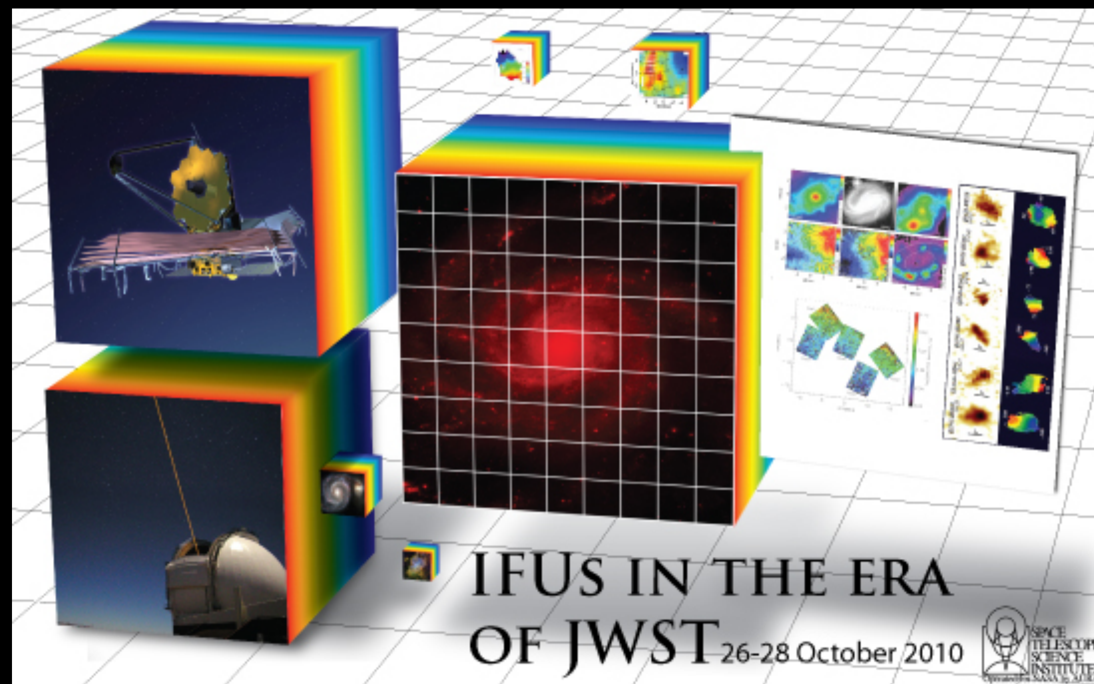
Christine L. Borgman 4 months ago · Private
"publications"

Astrometry.net glue

[video, demos]

THE FUTURE IS IN 3D

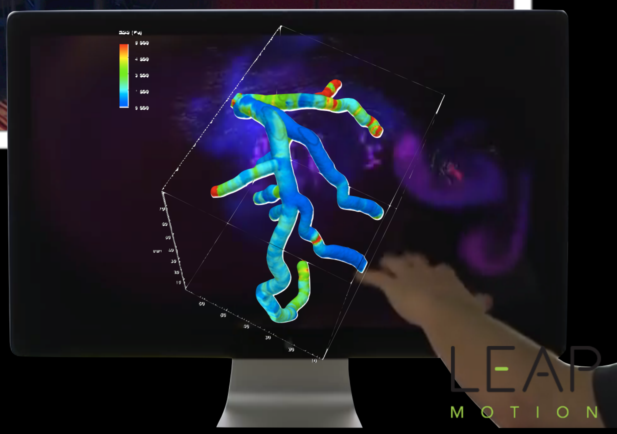




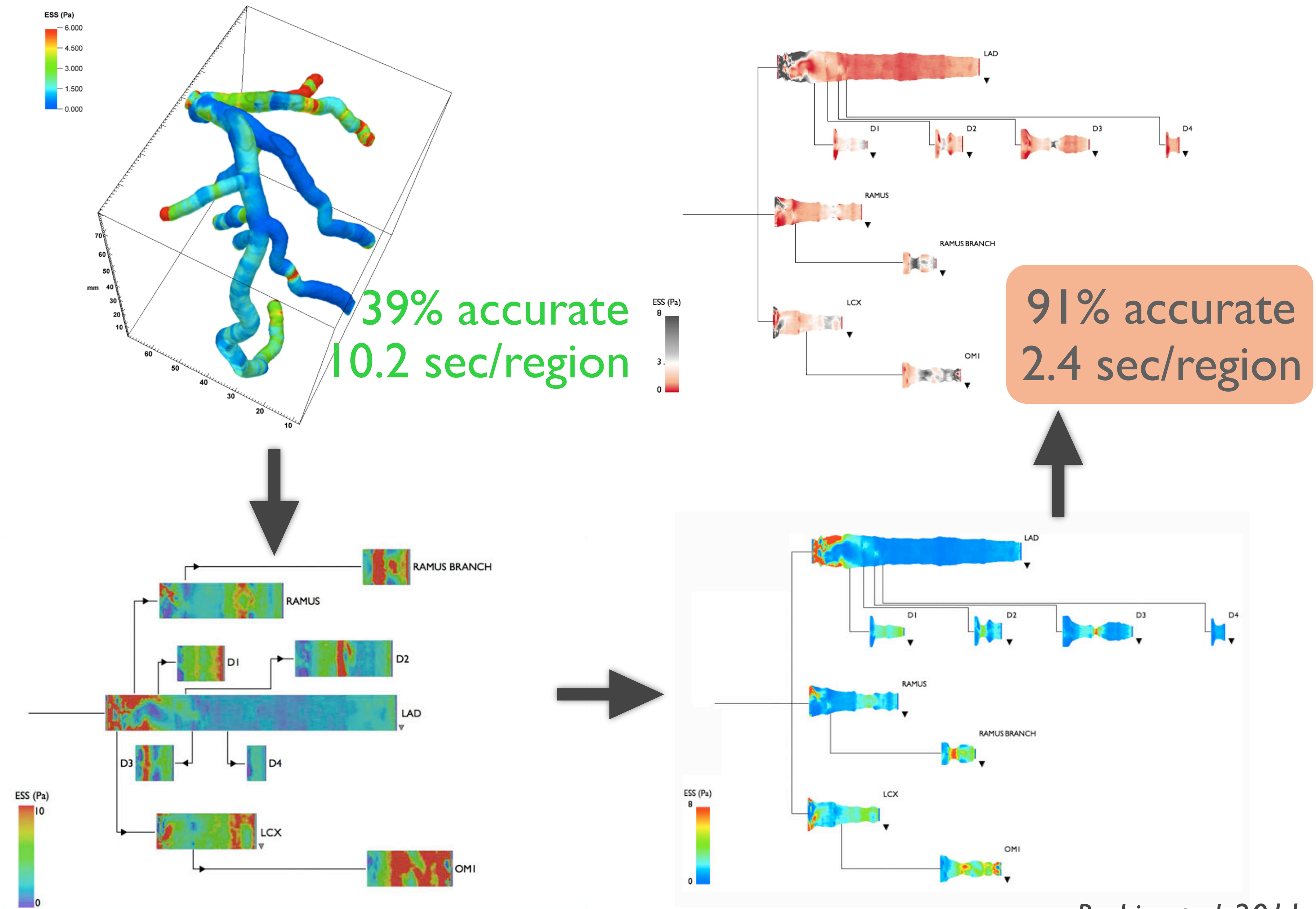
yt viz from ALMA data
(Turk, Rosolowsky)

IFUs on JWST...with Glue!
(Glue funded by NASA)

THE FUTURE IS MODULAR, OPEN-SOURCE, AND NOT (JUST) ON THE DESKTOP



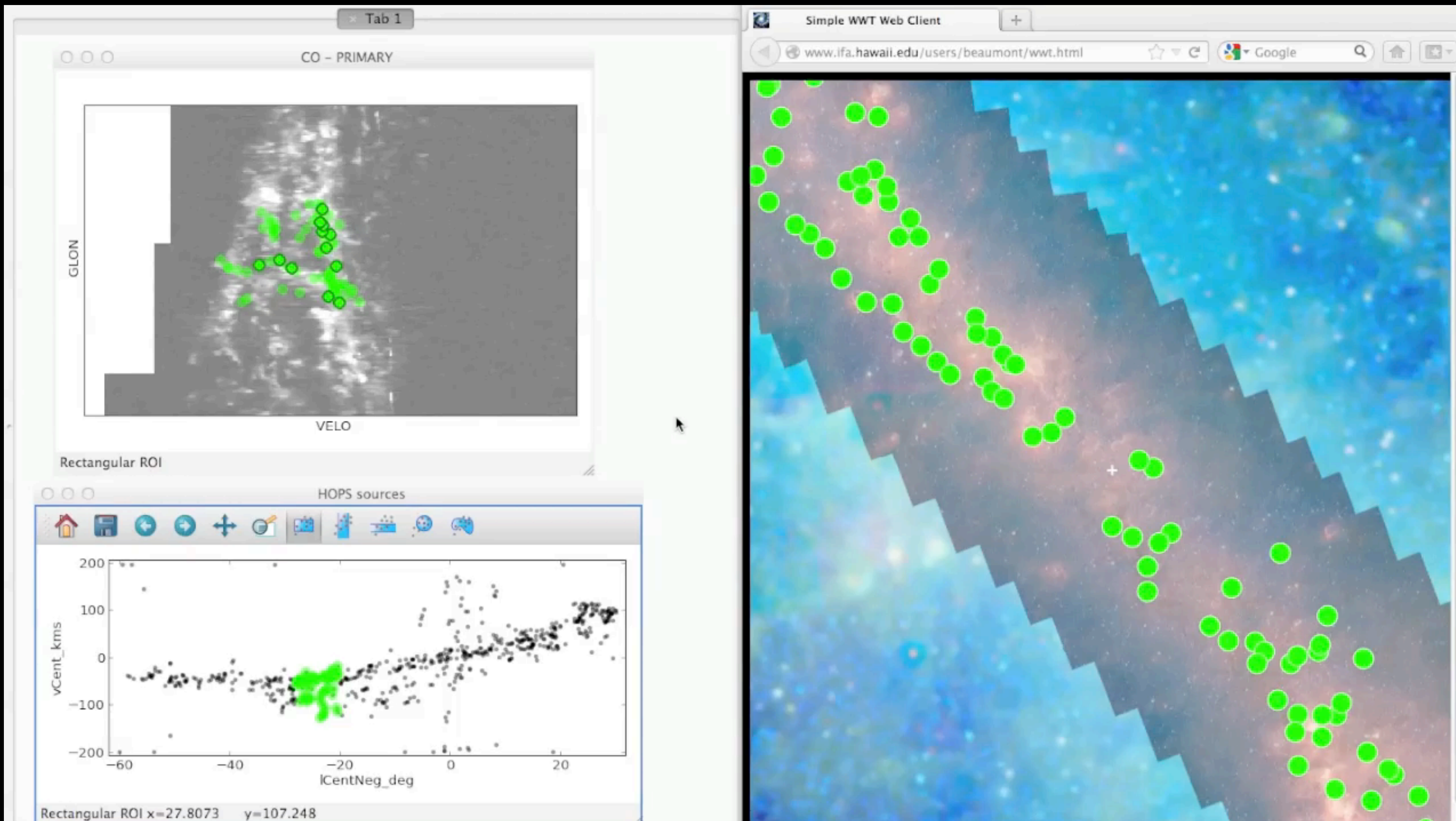
DIMENSIONALITY AND COLOR



Borkin et al. 2011
cf. colorbrewer2.org

glue

WorldWide Telescope



Video courtesy of Chris Beaumont, Lead Glue Architect

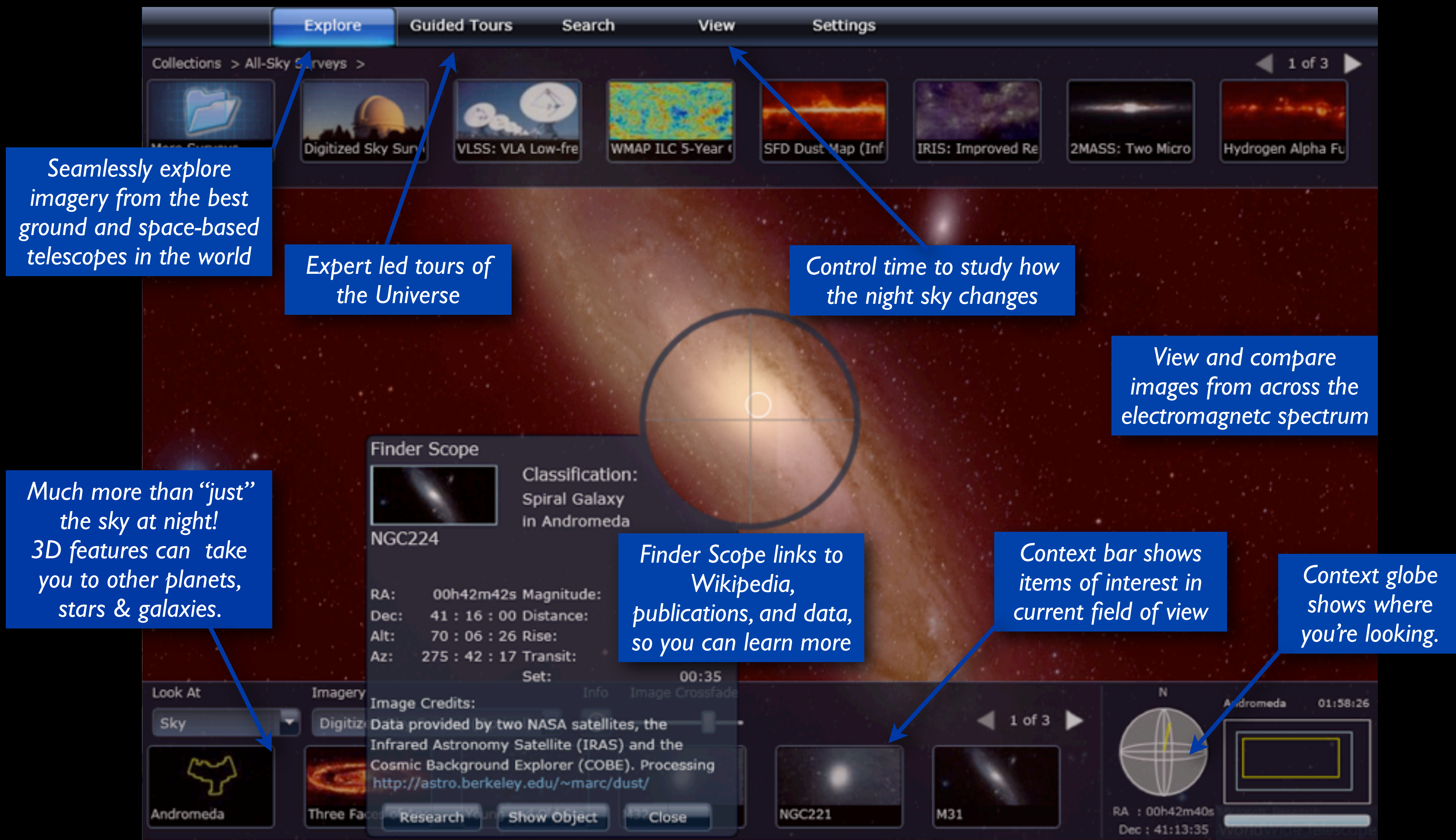
THE FUTURE OFFERS NEW WAYS TO LEARN

WorldWide Telescope Ambassadors

Higher Ed

Microsoft® Research WorldWide Telescope

worldwidetelescope.org



WHAT DOES "Shells in Perseus" LOOK LIKE IN THE FUTURE?

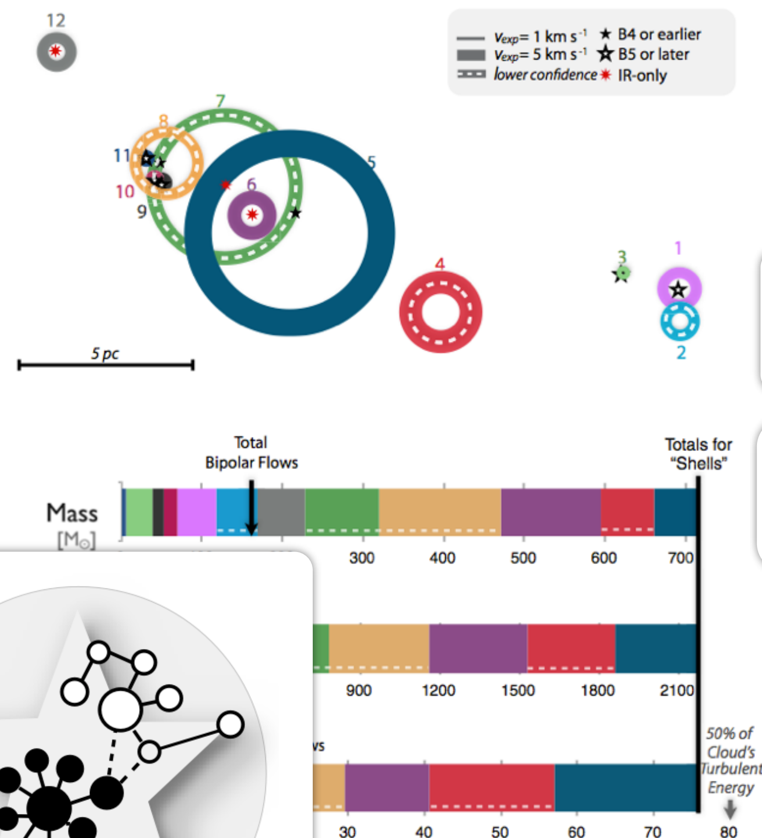


Figure 30. ... bution of shells in Perseus. The radius of each circle (and position) are proportional to the radius (and ... roportional to the expansion velocity (see the legend on the upper right corner). Shells with a confiden... hite line. Candidate powering sources with a B5 spectral type or later are shown as white star symbo... re shown as black (filled) star symbols. Candidate sources with no known spectral type (but known α) ... ic energy of the shells are shown in the three horizontal bars (where the colors indicate the value for e... gy of the molecular outflows in Perseus (from Arce et al. 2010) are shown for comparison.

The Astrophysical Journal, 742:105 (30pp), 2011 December 1

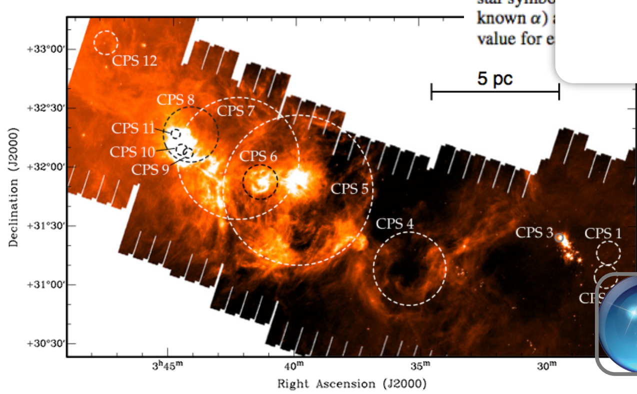
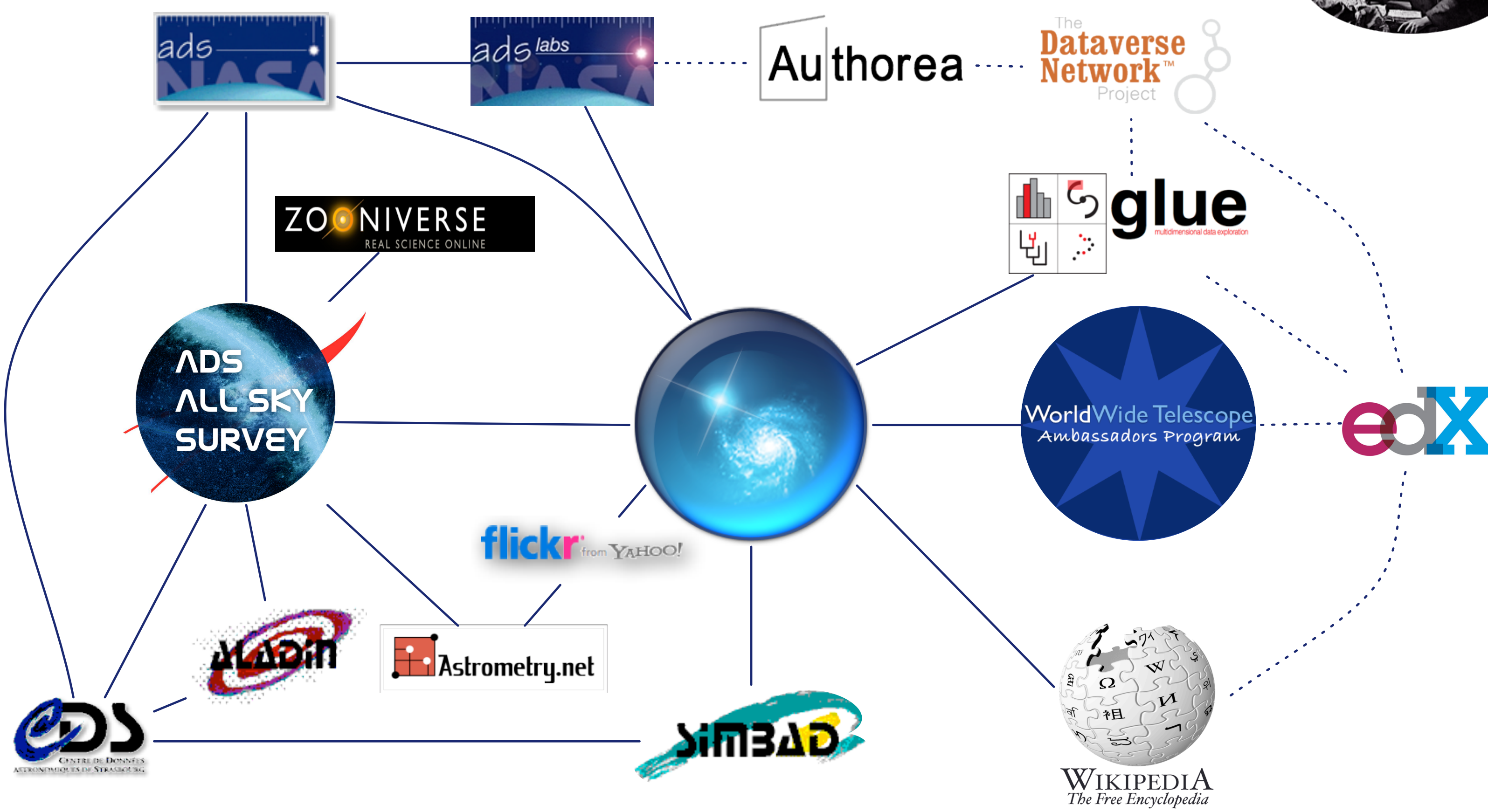


Figure 1. Shells in Perseus. The position of each CPS is shown as a dashed circle overlaid on the c2d *Spitzer* MIPS 1 (24 μm) image (from Rebull et al. 2007). (A color version of this figure is available in the online journal.)

Arce et al. 2011, 10.1088/0004-637X/742/2/105

SEAMLESS ASTRONOMY

Linking scientific data, publications, and communities



https://www.cfa.harvard.edu/~agoodman/seamless/

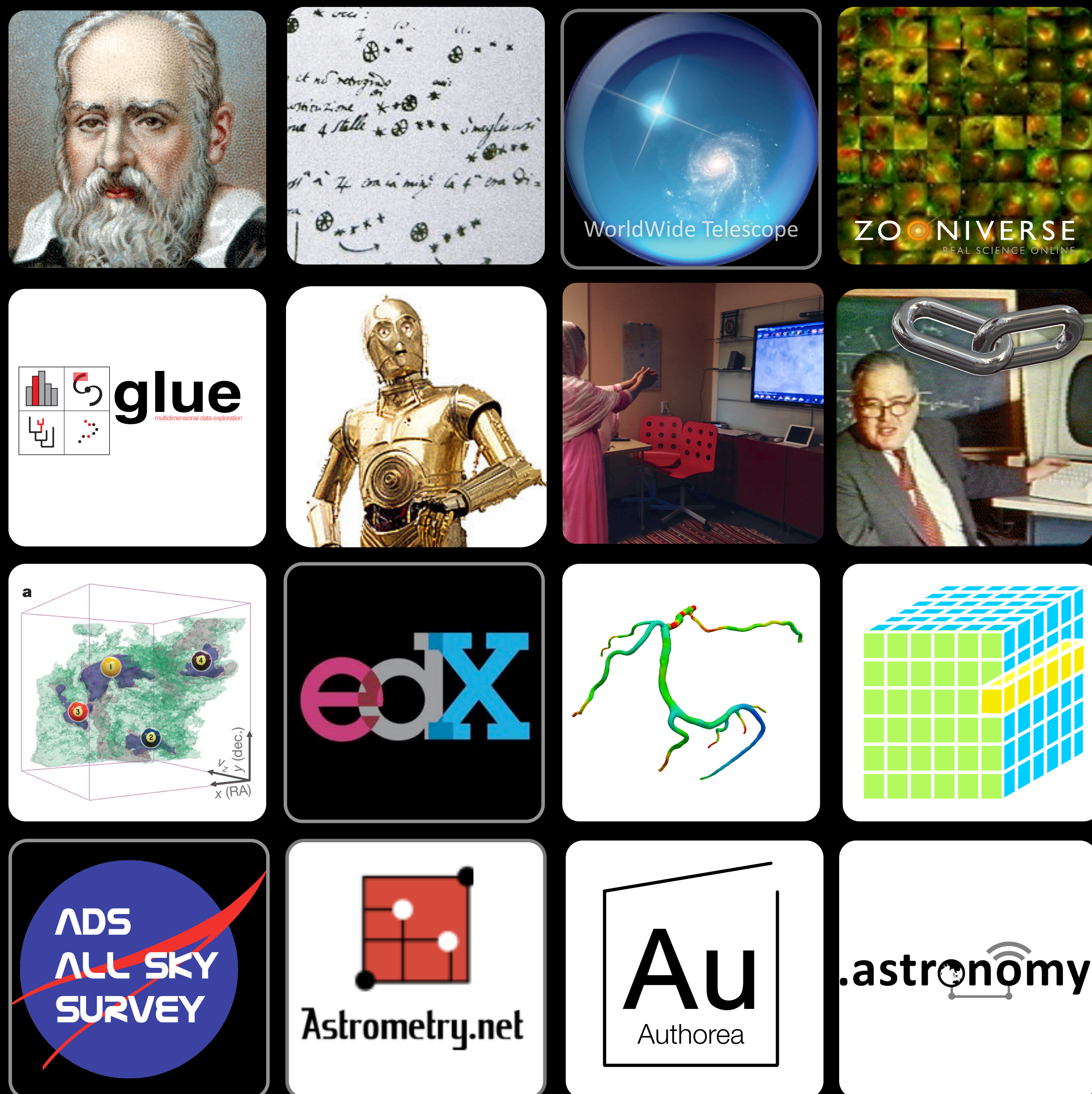
Supported by Microsoft Research NSF NASA Smithsonian

Made possible by MANY collaborators, listed at projects.iq.harvard.edu/seamlessastronomy

HIGH-DIMENSIONAL DATA VISUALIZATION & THE "PAPER" OF THE FUTURE

Alyssa A. Goodman • Harvard-Smithsonian Center for Astrophysics

[redacted]



WorldWide Telescope
ZOONIVERSE
REAL SCIENCE ONLINE
glue
a
y (dec.)
x (RA)
edX
ADS
ALL SKY
SURVEY
Astrometry.net
Au
Authorea
.astronomy

JOIN US?


SEAMLESS
ASTRONOMY
Linking scientific data, publications, and communities
VE RI TAS
HARVARD
NASA
NSF
projects.iq.harvard.edu/seamlessastron