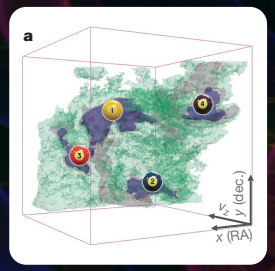
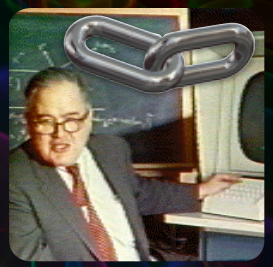
THE VALUE OF HIGH-DIMENSIONAL DATA VISUALIZATION IN SCIENCE





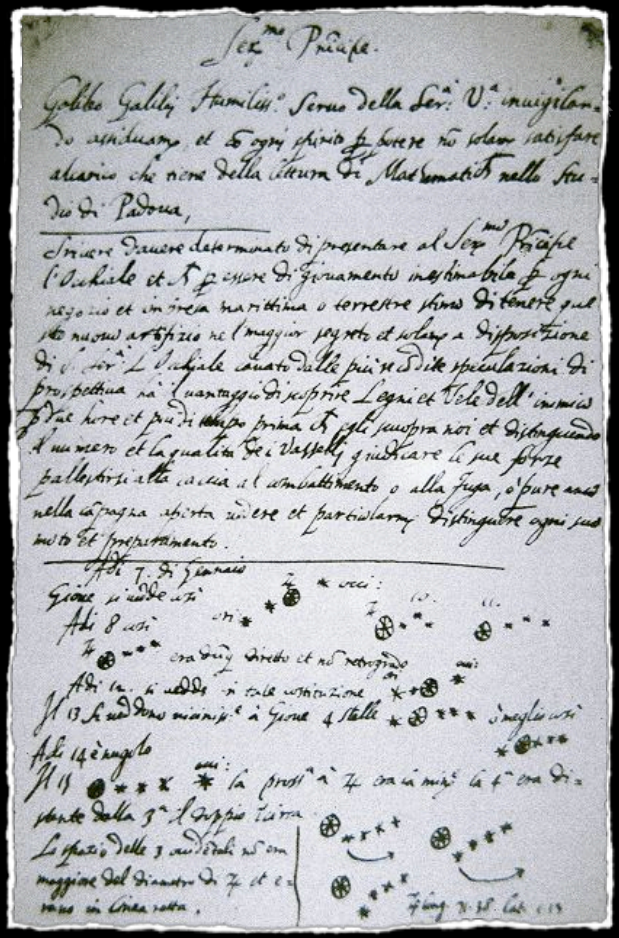



Alyssa A. Goodman
Harvard-Smithsonian Center for Astrophysics
& Radcliffe Institute
@aagie

Thomas Robitaille
lead Developer for glue
@astrofrog

background image credit: ALMA (ESO/NAOJ/NRAO)/Erik Rosolowsky

IN GALILEO'S MIND'S EYE



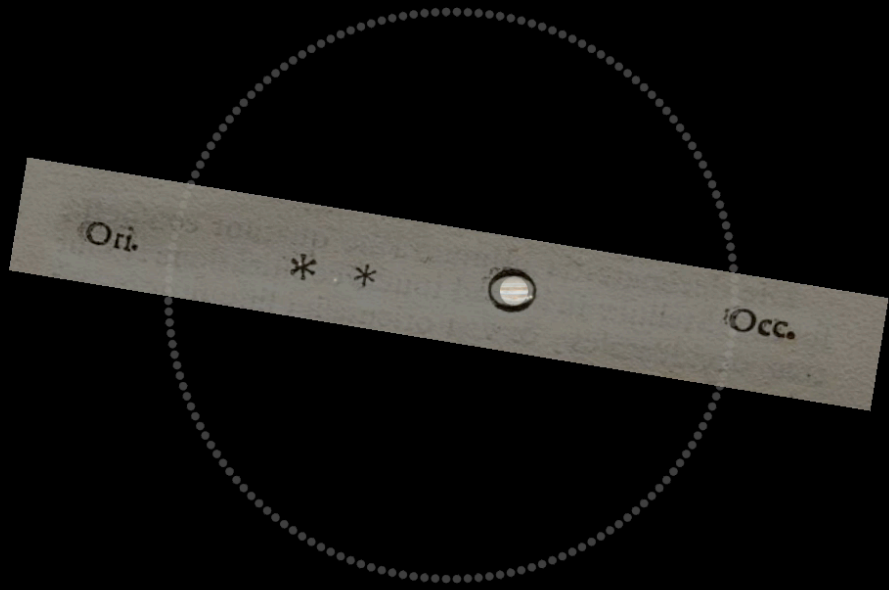
Sex.mo Prencipe.

Galileo Galilei Humilis.o Servo della Ser.a V.a inuigilan-
do assiduam.te et di ogni spirito di potere no solam.te sadisfare
alluffitio che tiene della Lettura di Mathematich nello stu-
dio di Padoua,

stimare douere determinato di presentare al Ser.mo Prencipe
l'Vnicale et N.ro essere di giouamento inestimabile p ogni
negotio et impresa marittima o terrestre stimo di tenere qual
sia nuouo artificio nel maggior segreto et silenzo a disposi.one
di V. Ser.a L'Vnicale causato dalle più sode speculazioni di
prospettiua nè i uantaggi di scoprire Leghe et Vele dell'inimico
2 hore et più di lontano prima di quelli scuoprano noi et distinguendo
il numero et la qualità de i Vasselli giudicare le sue forze
p allestirsi alla caccia al combattimento o alla fuga, o pure essere
nella campagna aperta uedere et particolarm.te distinguere ogni suo
moto et preparamento.

Adi 7. di Genaio
Giouue si uedde così
Adi 8 così
era duq. diretto et no retrogrado
Adi 10. si uedde in tale costituzione
Il 13 si uedeua mia.ta à Gioue 4 Stelle
Adi 14 è rotto
Il 15 la prossa à 4 era in mis.ta la 4.a era di-
stante della 3.a il doppio Tairra
Lo spatio delle 3 uedute no era
maggiore del diametro di 4 et e-
rano in linea retta.

January 11, 1610



IN GELLER, HUCHRA & DELAPPARENT'S MINDS' EYES

THE ASTROPHYSICAL JOURNAL, 302:L1–L5, 1986 March 1
© 1986. The American Astronomical Society. All rights reserved. Printed in U.S.A.

A SLICE OF THE UNIVERSE[1]

VALÉRIE DE LAPPARENT,[2,3] MARGARET J. GELLER,[2] AND JOHN P. HUCHRA[2]
Received 1985 November 12; accepted 1985 December 5

ABSTRACT

We describe recent results obtained as part of the extension of the Center for Astrophysics redshift survey to $m_B = 15.5$. The new sample contains 1100 galaxies (we measured 584 new redshifts) in a $6° \times 117°$ strip going through the Coma cluster. Several features of the data are striking. The galaxies appear to be on the surfaces of bubble-like structures. The bubbles have a typical diameter of $\sim 25h^{-1}$ Mpc. The largest bubble in the survey has a diameter of $\sim 50h^{-1}$ Mpc, comparable with the most recent estimates of the diameter of the void in Bootes. The galaxy density in the region of the largest void contained in the survey is only 0.20 of the mean. The edge of the largest void in the survey is remarkably sharp.

All of these features pose serious challenges for current models for the formation of large-scale structure. The best available model for generating these structures is the explosive galaxy formation theory of Ostriker and Cowie, published in 1981. These new data might be the basis for a new picture of the galaxy and cluster distributions.

Subject headings: cosmology — galaxies: clustering — galaxies: formation

I. INTRODUCTION

The behavior of the distribution of galaxies on scales $\geq 10h^{-1}$ Mpc ($H_0 = 100h$ km s^{-1} Mpc^{-1}) is a critical constraint on models for the formation of large-scale structure. Even the currently most popular cold dark matter model with biased galaxy formation (Davis *et al.* 1985) cannot produce voids as large as the one in Bootes discovered by Kirshner *et al.* (1981). In other words, if large voids are common, standard gravitational clustering models are unlikely to match the observations.

On the observational side there have been a number of discoveries of structures which extend for tens of Mpc. Zeldovich, Einasto, and Shandarin (1982) emphasize a general cell-like structure in the galaxy distribution on these and larger scales. Two striking examples of individual large structures are the Perseus-Pisces chain (Gregory, Thompson, and Tifft 1981), an apparently filamentary structure extending for $40h^{-1}$ Mpc, and the void in Bootes which has a volume of order 10^3h^{-3} Mpc^3. Taken at face value, the break in the correlation function at a scale of $15h^{-1}$ Mpc (Groth and Peebles 1977; Davis and Peebles 1983) indicates that structures as large as these must not be common. However, the determination of the behavior of the correlation function on large scales may be suspect because of biases in the galaxy catalogs (Geller, de Lapparent, and Kurtz 1984; de Lapparent, Kurtz, and Geller 1985).

This *Letter* is a preliminary discussion of recent results obtained as part of the extension of the Center for Astrophysics redshift survey. Several features of the results are striking. The distribution of galaxies in the redshift survey slice looks like a slice through the suds in the kitchen sink: it appears that the galaxies are on the surfaces of bubble-like structures with diameter $25–50h^{-1}$ Mpc. This topology poses serious challenges for current models for the formation of large-scale structure.

II. THE DATA

The sample is restricted to objects listed by Zwicky *et al.* (1961) with $m_B \leq 15.5$, $8^h \leq \alpha \leq 17^h$ and $26°.5 \leq \delta \leq 32°.5$. On the sky, this survey covers a strip of $\sim 117° \times 6°$, centered near the north Galactic pole. We assigned magnitudes to the individual objects in multiple systems: galaxies are included in the sample only if they satisfy $m_B \leq 15.5$. The resulting catalog contains 1099 galaxies. Redshifts for the 186 galaxies brighter than $m_B = 14.5$ have been compiled by Huchra *et al.* in the CfA redshift survey (1982). Among the remaining redshifts, 327 have been published by other groups. The other 584 redshifts were measured with the Mount Hopkins 60 inch (1.5 m) telescope and the MMT, yielding a complete redshift survey. The data will be published elsewhere (Huchra *et al.* 1987).

Figure 1*a* shows a plot of the observed velocity versus right ascension for the galaxies brighter than $m_B = 15.5$ in the full 6° thick wedge; we only plot the 1061 objects with velocities less than 15,000 km s^{-1}. The effective depth of the sample is $\sim 100h^{-1}$ Mpc ($M^* = -19.4$; Davis and Huchra 1982). For contrast, we plot in Figure 1*b* the 182 objects with $m_B \leq 14.5$ and $V \leq 10{,}000$ km s^{-1}.

Figures 1*a* and 1*b* differ significantly in appearance. In Figure 1*a*, nearly every galaxy with $V \leq 10{,}000$ km s^{-1} appears to be in a large structure. We argue below that the size of the largest of these bubble-like structures is comparable with the depth of the earlier survey. Thus undersampling explains the difference in the appearance of the surveys.

In order to compare the redshift-space distribution with the distribution projected on the sky, Figure 1*c* shows the positions of all the galaxies from the Zwicky *et al.* catalog which satisfy $m_B \leq 15.5$, $8^h \leq \alpha \leq 17^h$ and $8°.5 \leq \delta \leq 50°.5$. The grid is Cartesian in R.A. and decl. The deficiency of galaxies west of 9^h and east of 16^h is caused by Galactic obscuration. The tick marks show the 6° declination region of the redshift survey in Figure 1*a*. The Coma cluster is the dense region at 13^h in the 6° strip.

III. ANALYSIS

The cellular pattern of Figure 1*a* and the smoothness of Figure 1*c* can be simply understood if the galaxies are distributed on the surfaces of shells tightly packed next to each other. If shell-like structures are common in the universe, any sufficiently deep wedge-shaped redshift survey will show a pattern of voids surrounded by connected filaments of galaxies similar to that in Figure 1*a*.

One impressive feature of the new data in Figure 1*a* is the presence of several large regions almost devoid of galaxies. The galaxies appear to be distributed in elongated structures

[1] Based partly on results obtained with the Multiple Mirror Telescope, a facility operated jointly by the Smithsonian Institution and the University of Arizona.
[2] Harvard-Smithsonian Center for Astrophysics.
[3] Ecole Normale Supérieure de Jeunes Filles and Université Paris VII, Paris.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

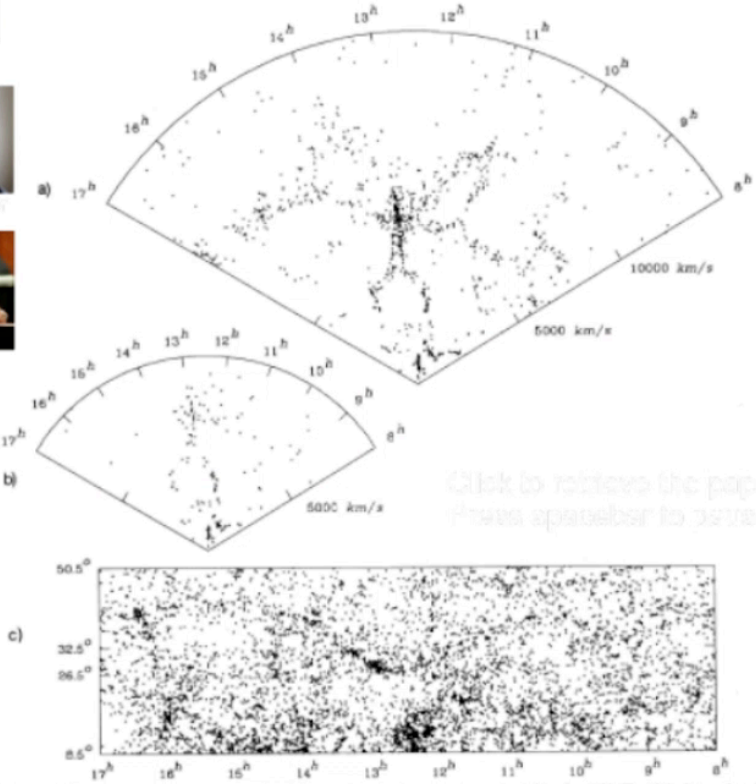


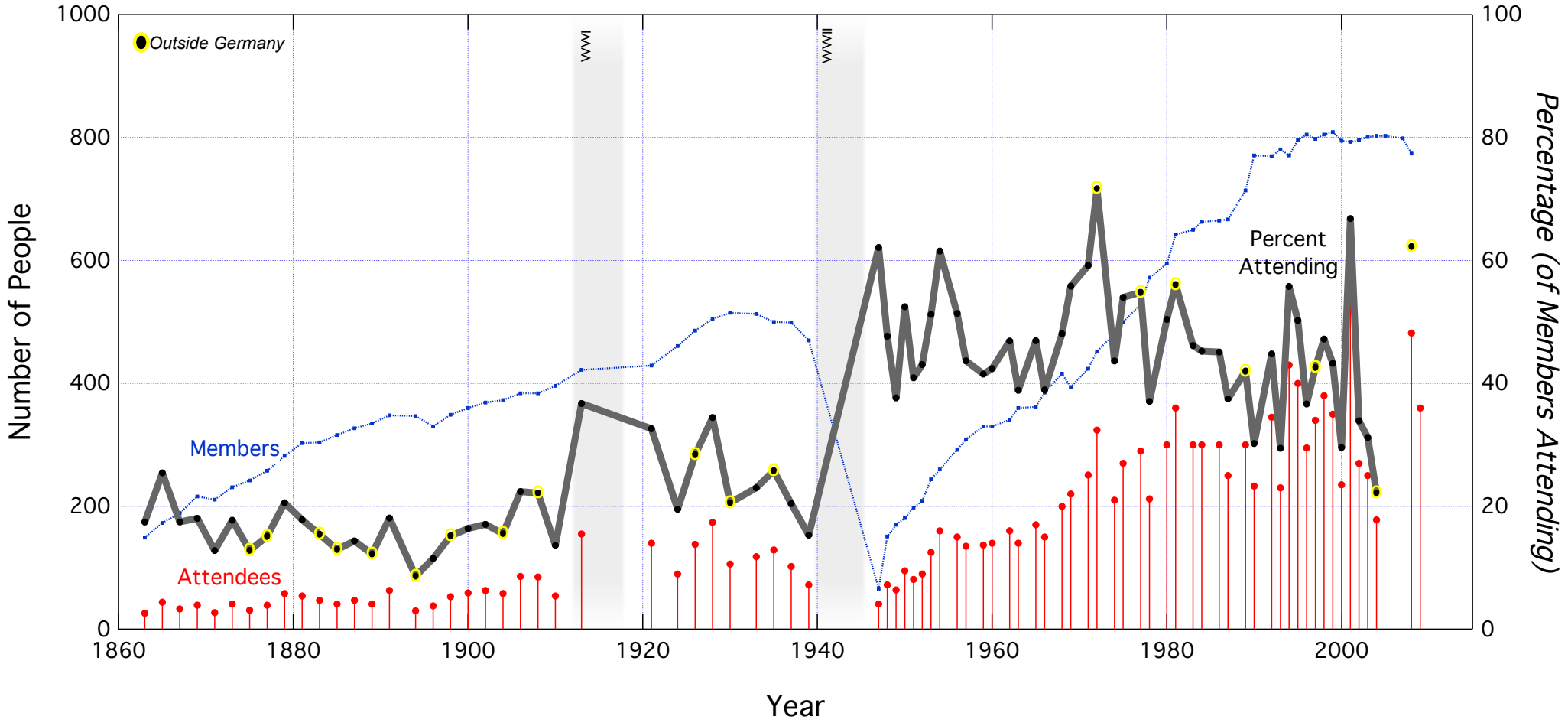
FIG. 1.—(*a*) Map of the observed velocity plotted vs. right ascension in the declination wedge $26°.5 \leq \delta \leq 32°.5$. The 1061 objects plotted have $m_B \leq 15.5$ and $V \leq 15{,}000$ km s^{-1}. (*b*) Same as Fig. 1*a* for $m_B \leq 14.5$ and $V \leq 10{,}000$ km s^{-1}. The plot contains 182 galaxies. (*c*) Projected map of the 7031 objects with $m_B \leq 15.5$, listed by Zwicky *et al.* in the region bounded by $8^h \leq \alpha \leq 17^h$ and $8°.5 \leq \delta \leq 50°.5$.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

"HIGH-DIMENSIONAL"

need not mean spatial

need not be displayed in >2D



example from Goodman 2012: "Principles of High-Dimensional Data Visualization in Astronomy"

"HIGH-DIMENSIONAL"

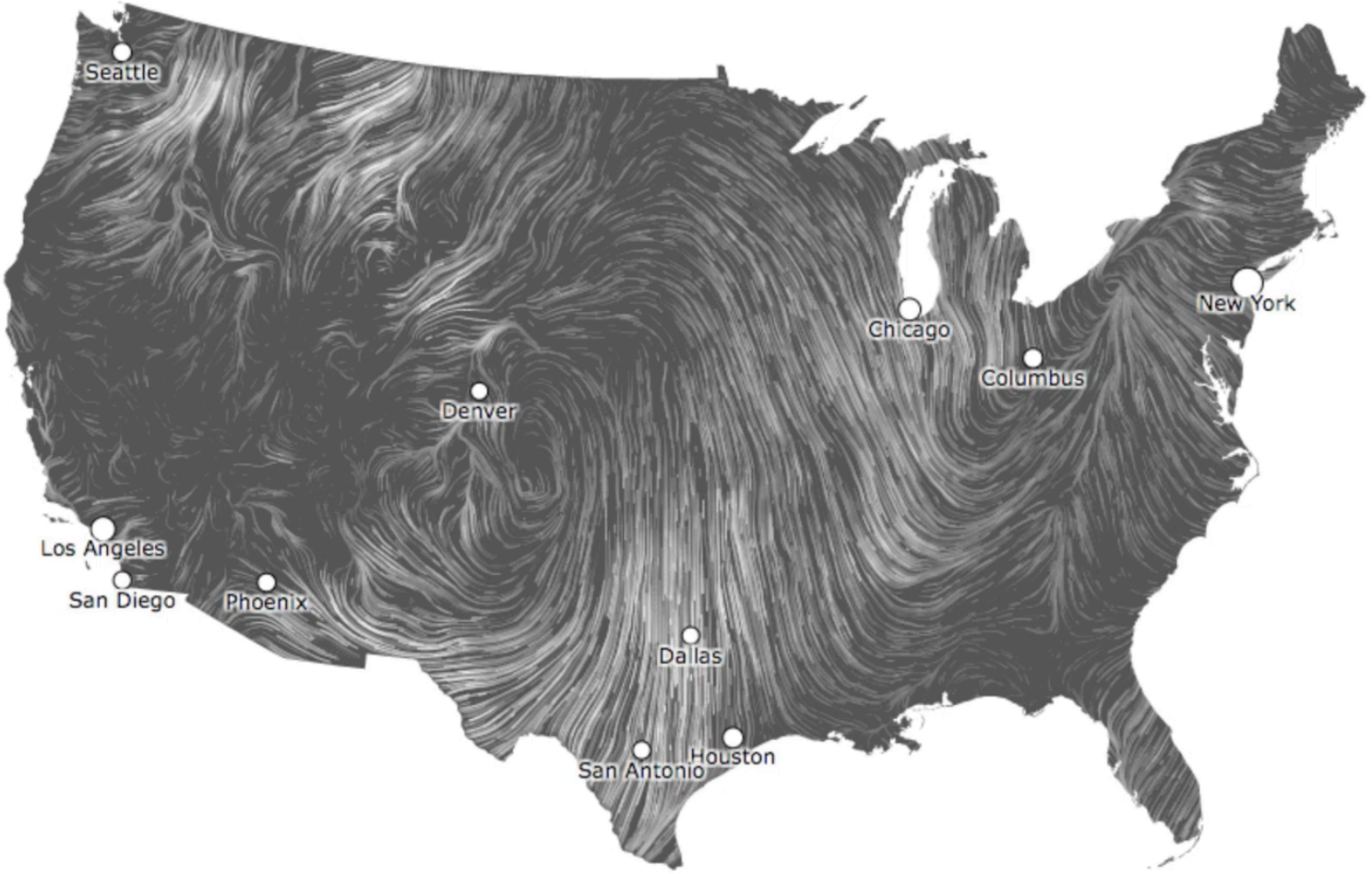
need not mean spatial

need not be displayed in >2D

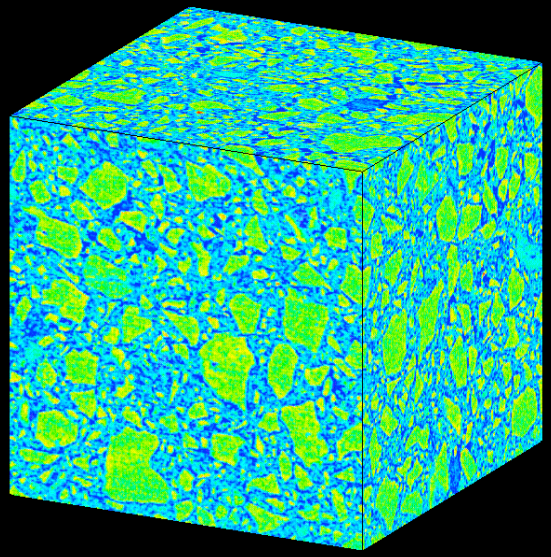
wind map

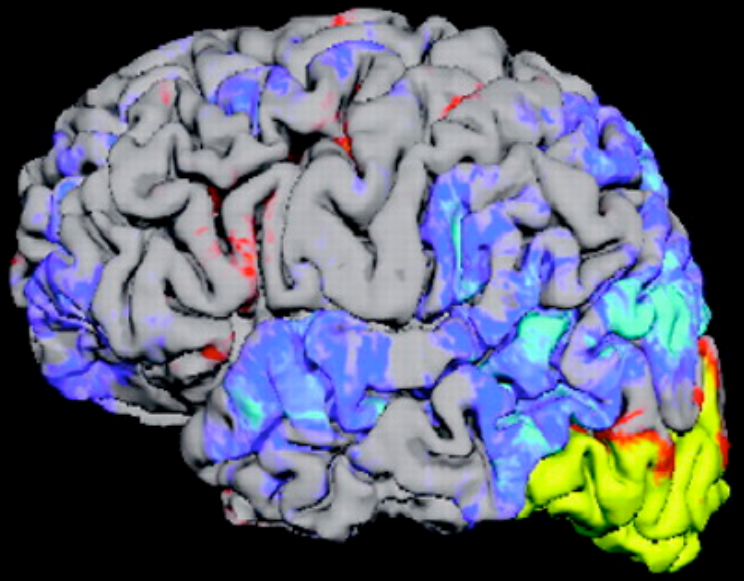
October 20, 2016
4:35 pm EST
(time of forecast download)

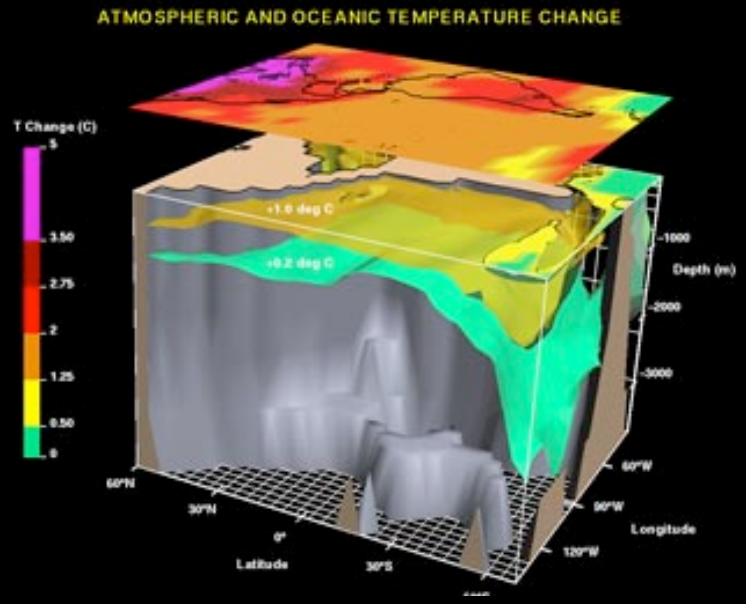
top speed: **33.2 mph**
average: **8.5 mph**

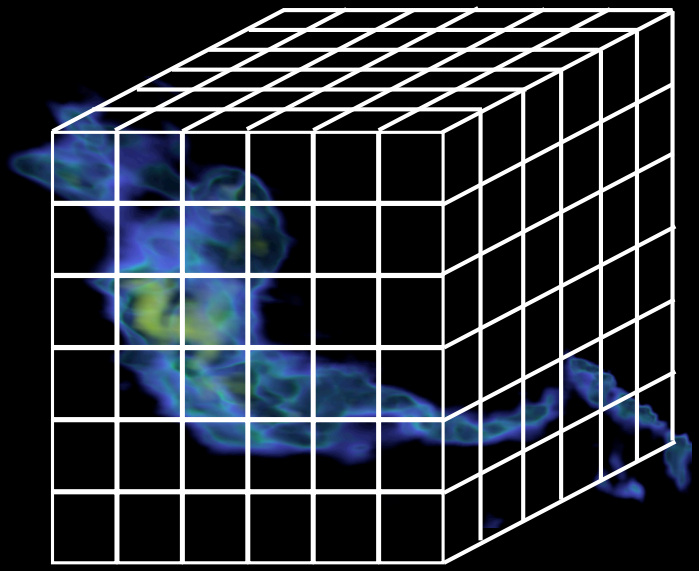


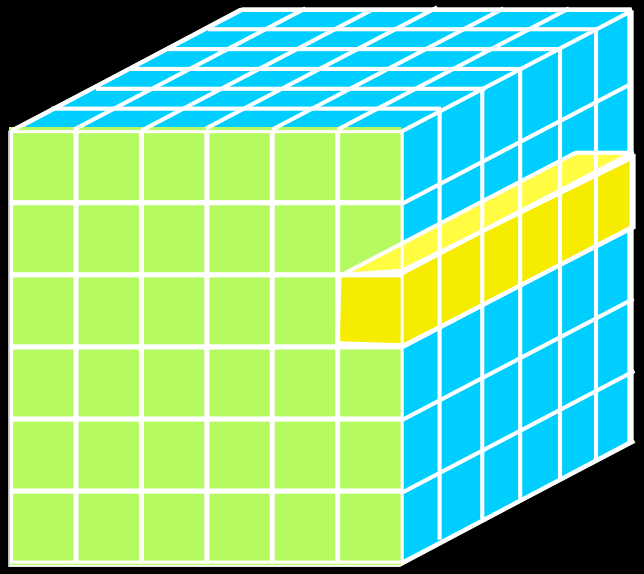
"HIGH-DIMENSIONAL" DATA







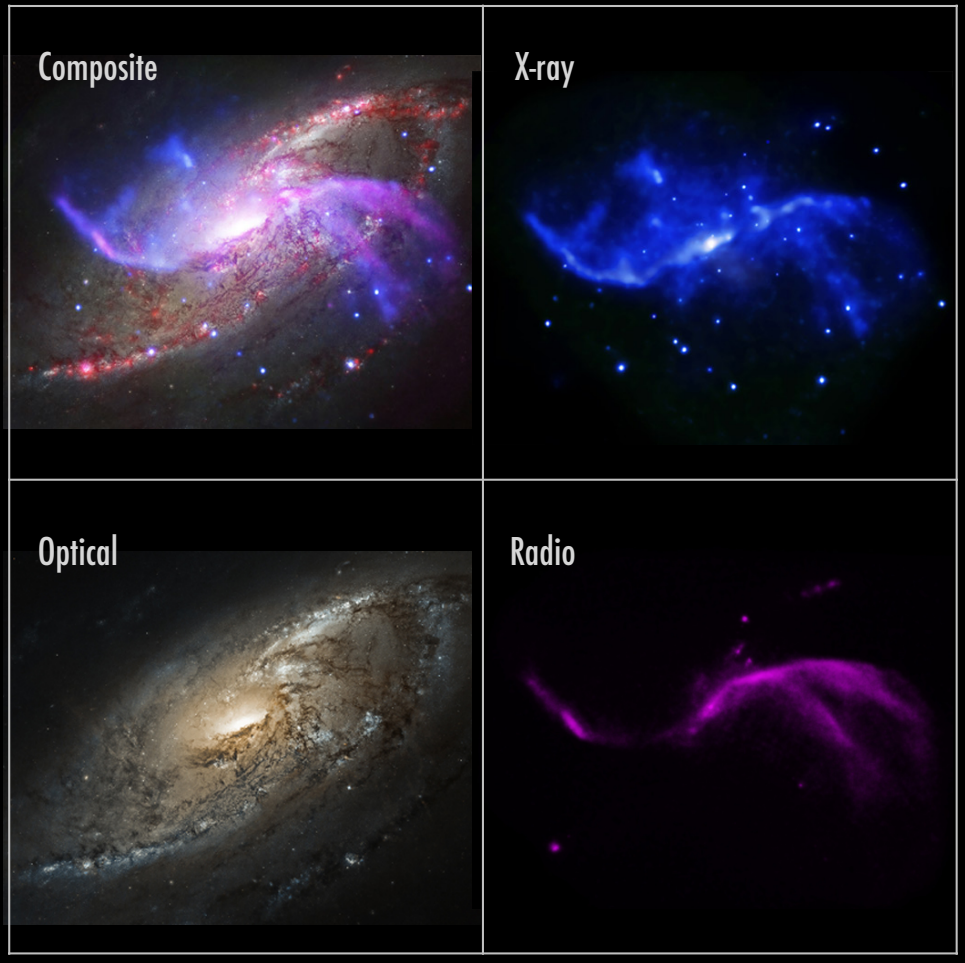




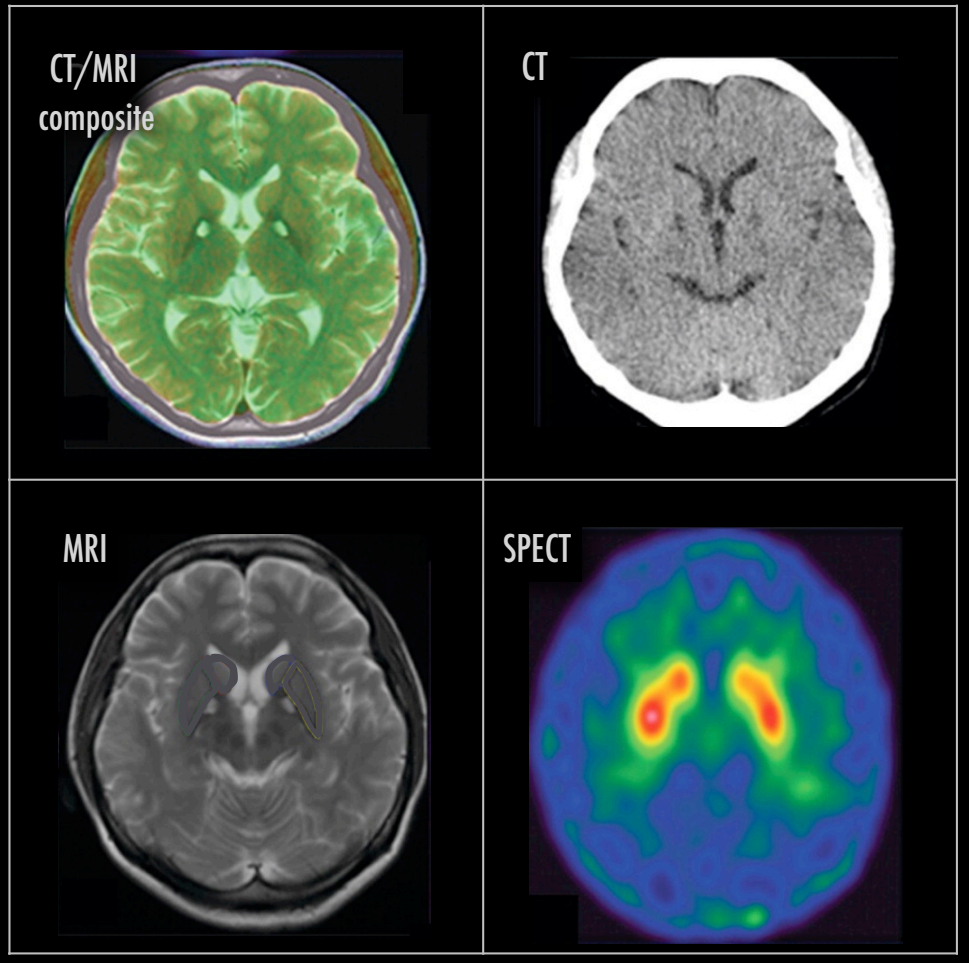
DATA-DIMENSIONS-DISPLAY

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

ASTRONOMICAL MEDICINE



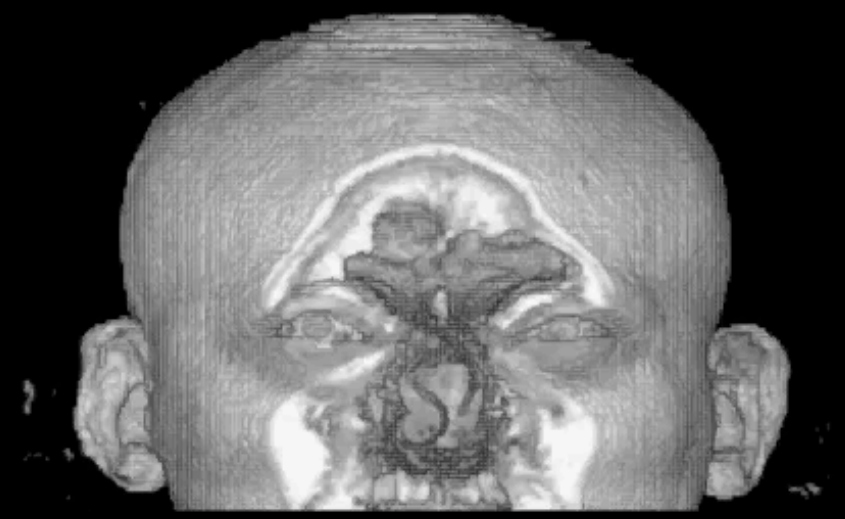
chandra.harvard.edu/photo/2014/m106/



Chang, et al. 2011, brain.oxfordjournals.org/content/134/12/3632

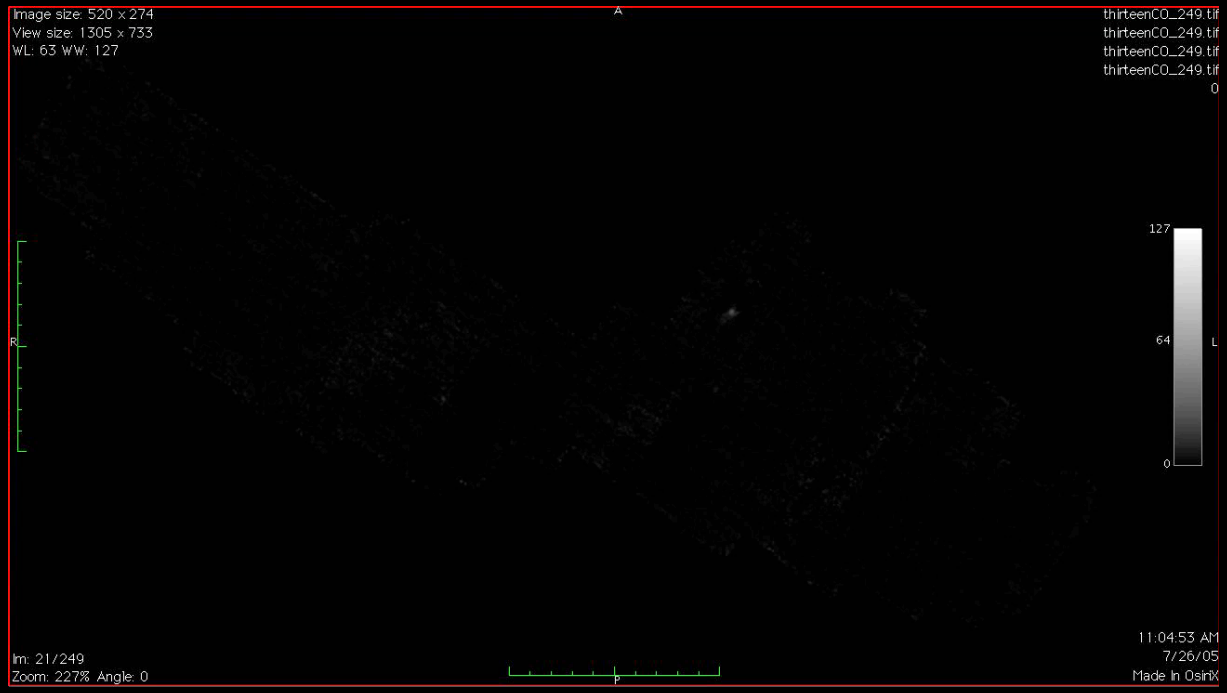
ASTRONOMICAL MEDICINE

"KEITH"



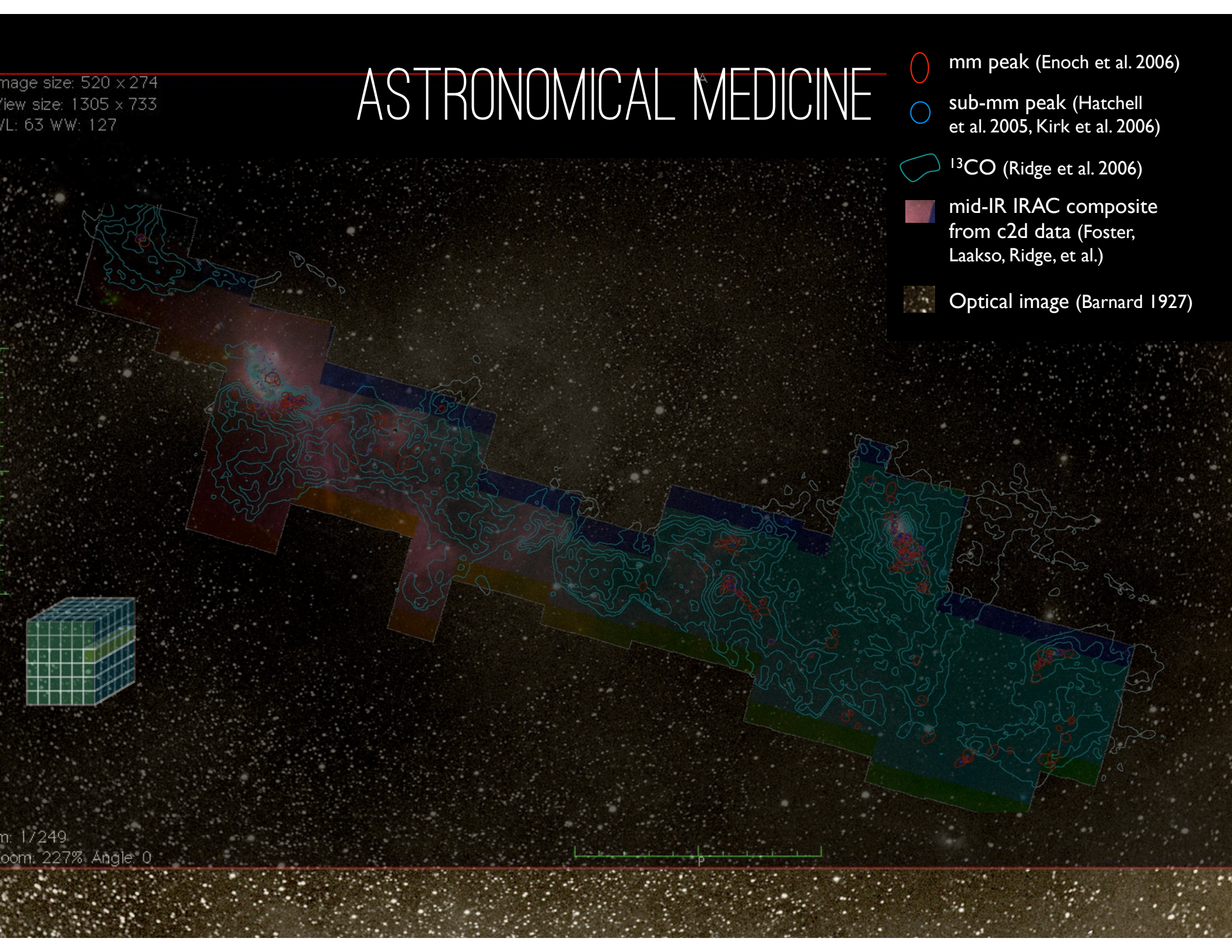
"z" is depth into head

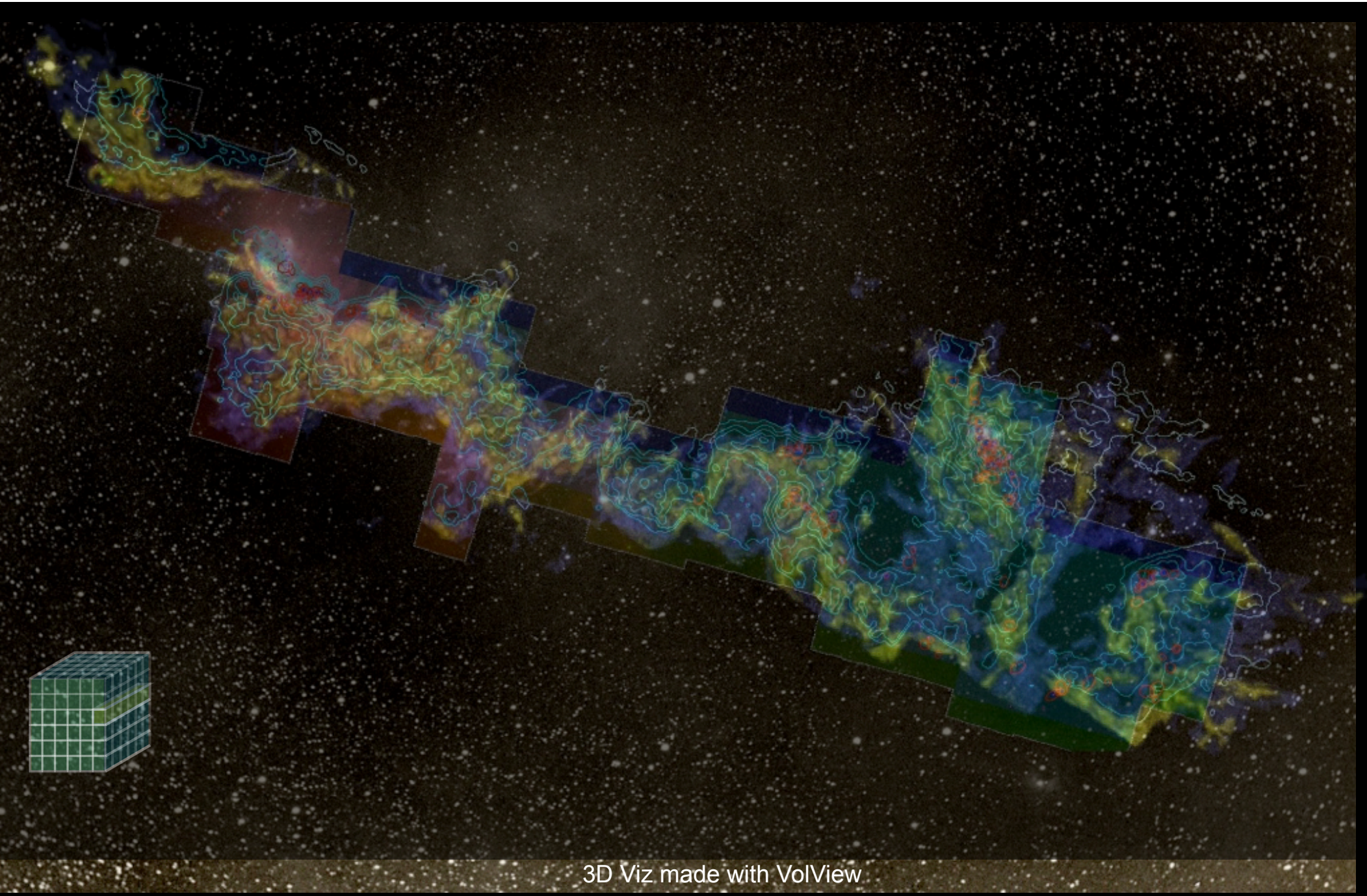
"PERSEUS"



"z" is line-of-sight velocity

ASTRONOMICAL MEDICINE





COMPLETE

2009

3D PDF

HIGH-DIMENSIONAL DATA IN A "PAPER"

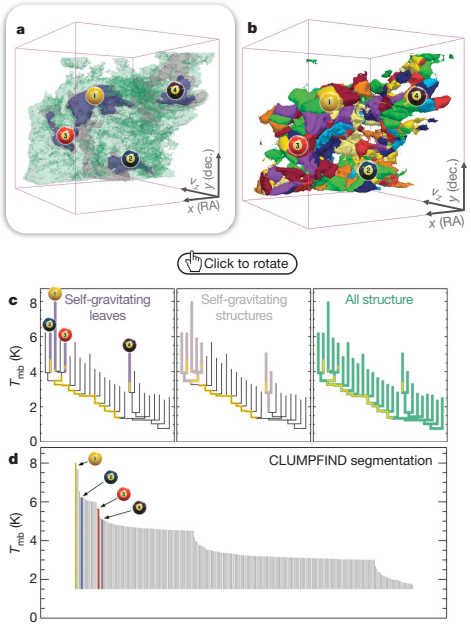


Figure 2 | Comparison of the 'dendrogram' and 'CLUMPFIND' feature-identification algorithms as applied to ^{13}CO emission from the L1448 region of Perseus. a, 3D visualization of the surfaces indicated by colours in the dendrogram shown in **c**. Purple illustrates the smallest scale self-gravitating 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…rly 2D work as inspiration, we have developed a structure-id…rithm 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 spectr… 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, su… 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.

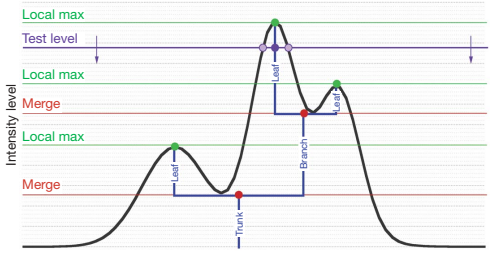
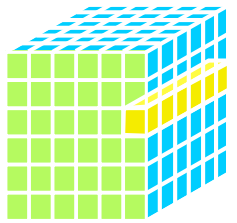


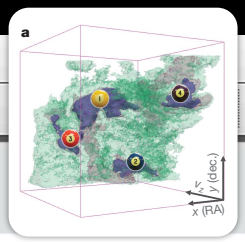
Figure 3 | Schematic illustration of the dendrogram process. Shown is the construction of a dendrogram from a hypothetical one-dimensional emission profile (black). The dendrogram (blue) can be constructed by 'dropping' a test constant emission level (purple) from above in tiny steps (exaggerated in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

©2009 Macmillan Publishers Limited. All rights reserved



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

Create
1 / 4
131%
Share



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

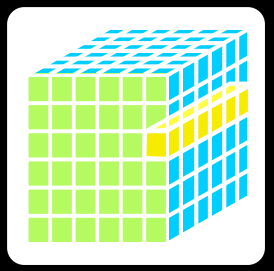
LETTERS

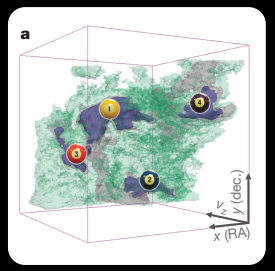
A role for self-gravity at multiple length scales in the process of star formation

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-

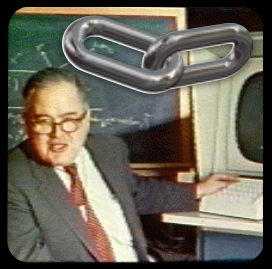
overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line





So, by 2009 we could “publish” what Galileo had in mind.

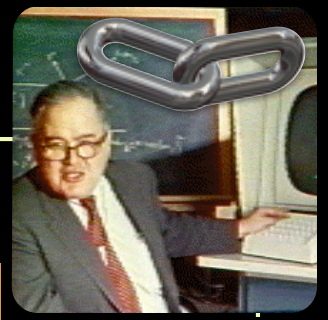
Today, though, we want to explore big, wide, data sets, live.



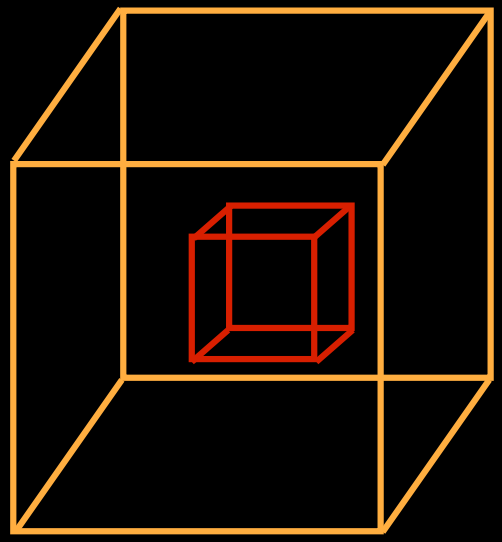
glue
multidimensional data exploration

Au
Authorea

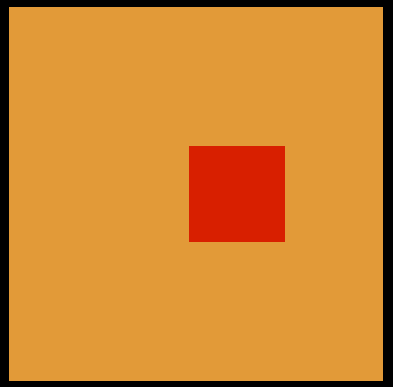
LINKED VIEWS OF HIGH-DIMENSIONAL DATA



John Tukey

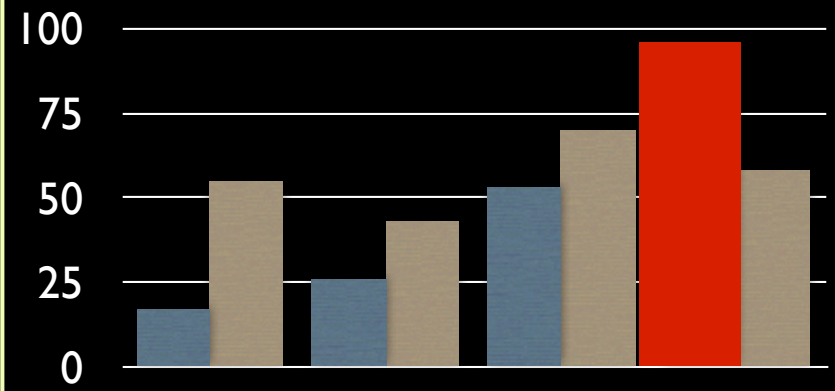


3D

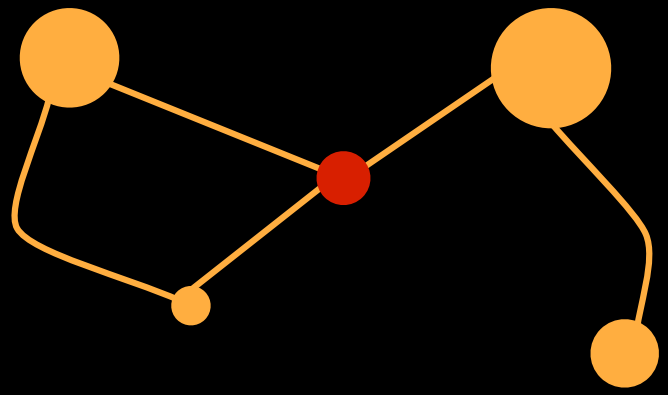


2D

Statistics



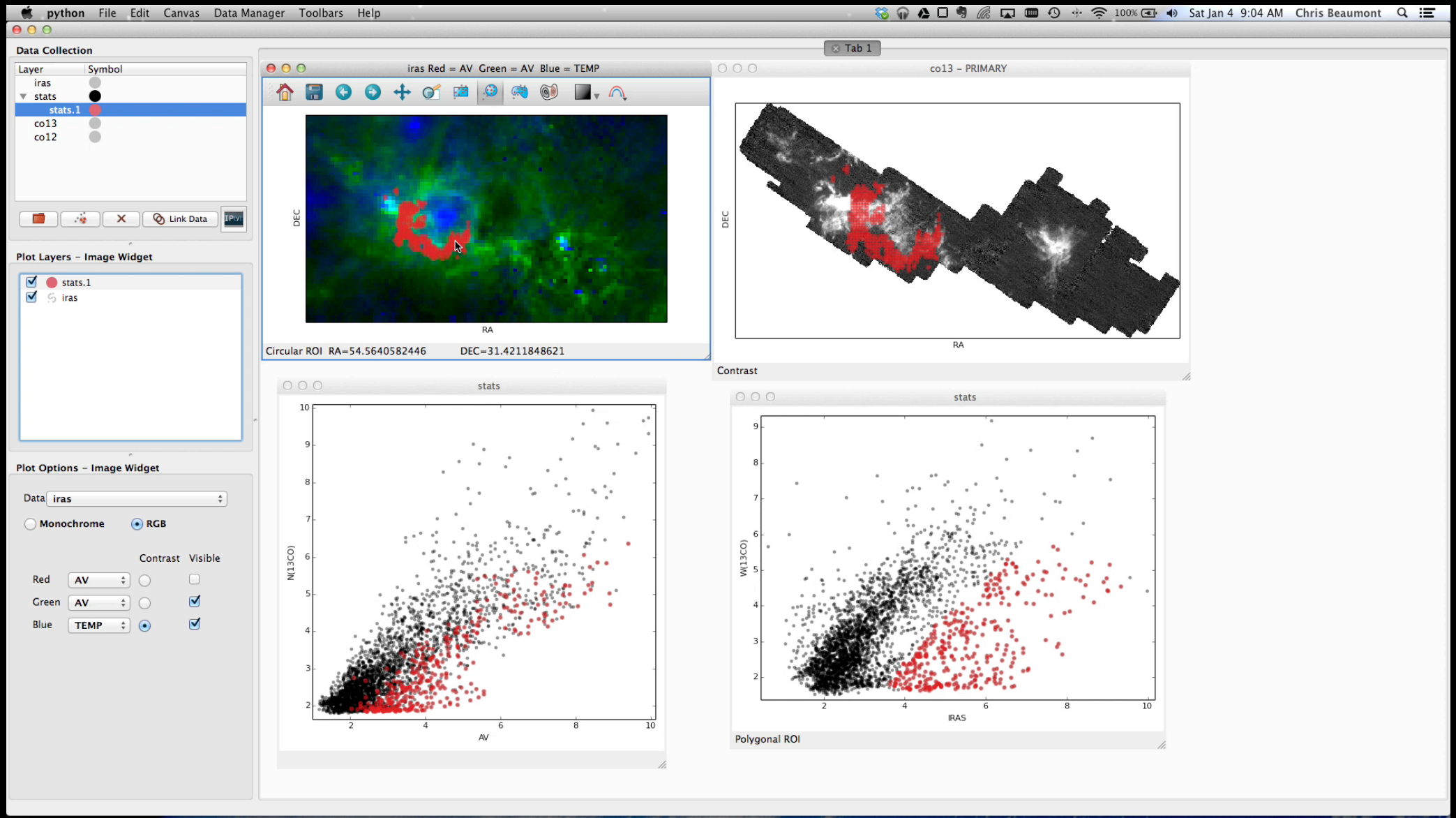
Data Abstraction



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

LINKED VIEWS OF HIGH-DIMENSIONAL DATA (IN PYTHON)

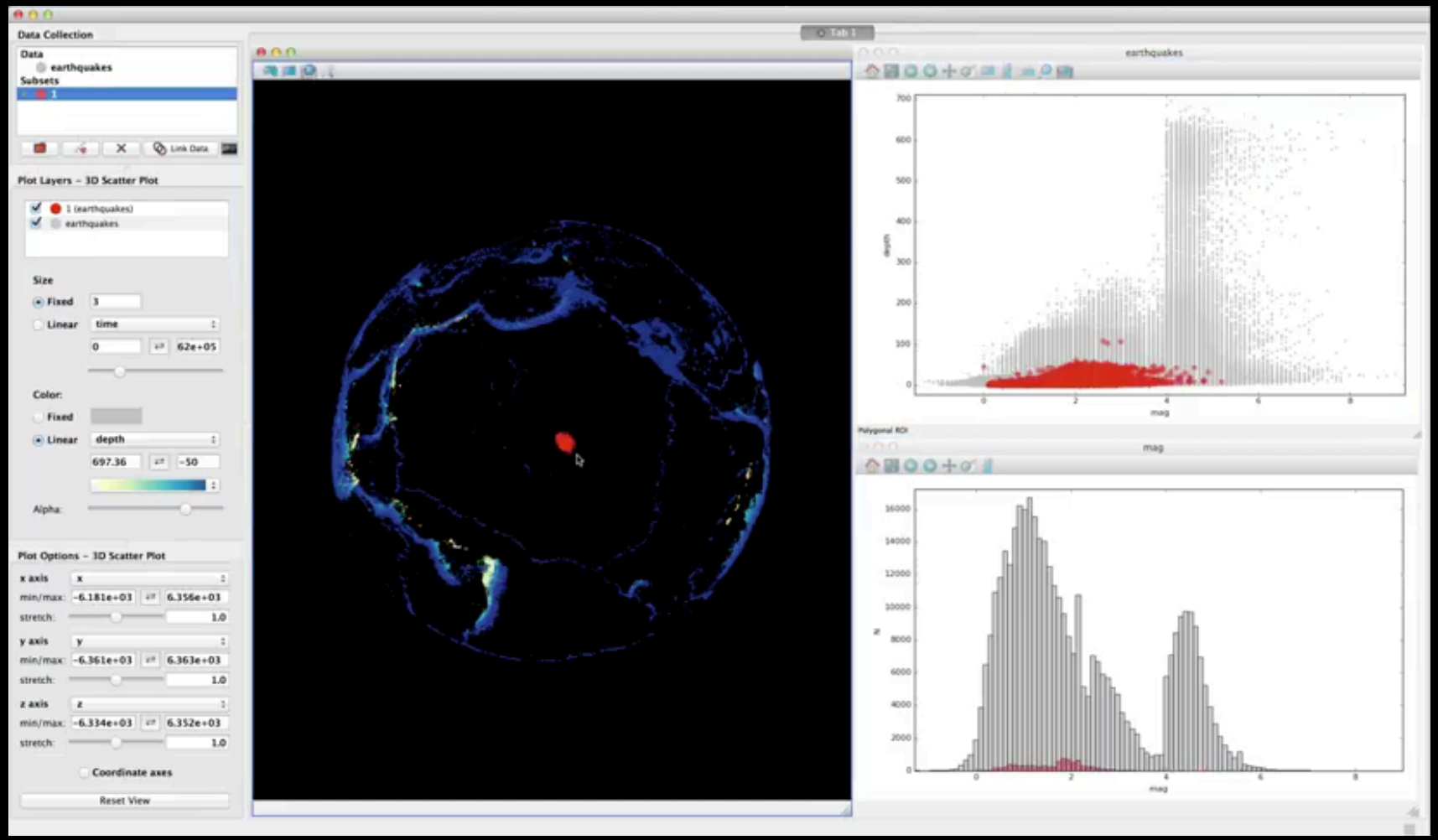
GLUE



Christopher Beaumont, w/A. Goodman, T. Robitaille & M. Borkin

LINKED VIEWS OF HIGH-DIMENSIONAL DATA (IN PYTHON)

GLUE

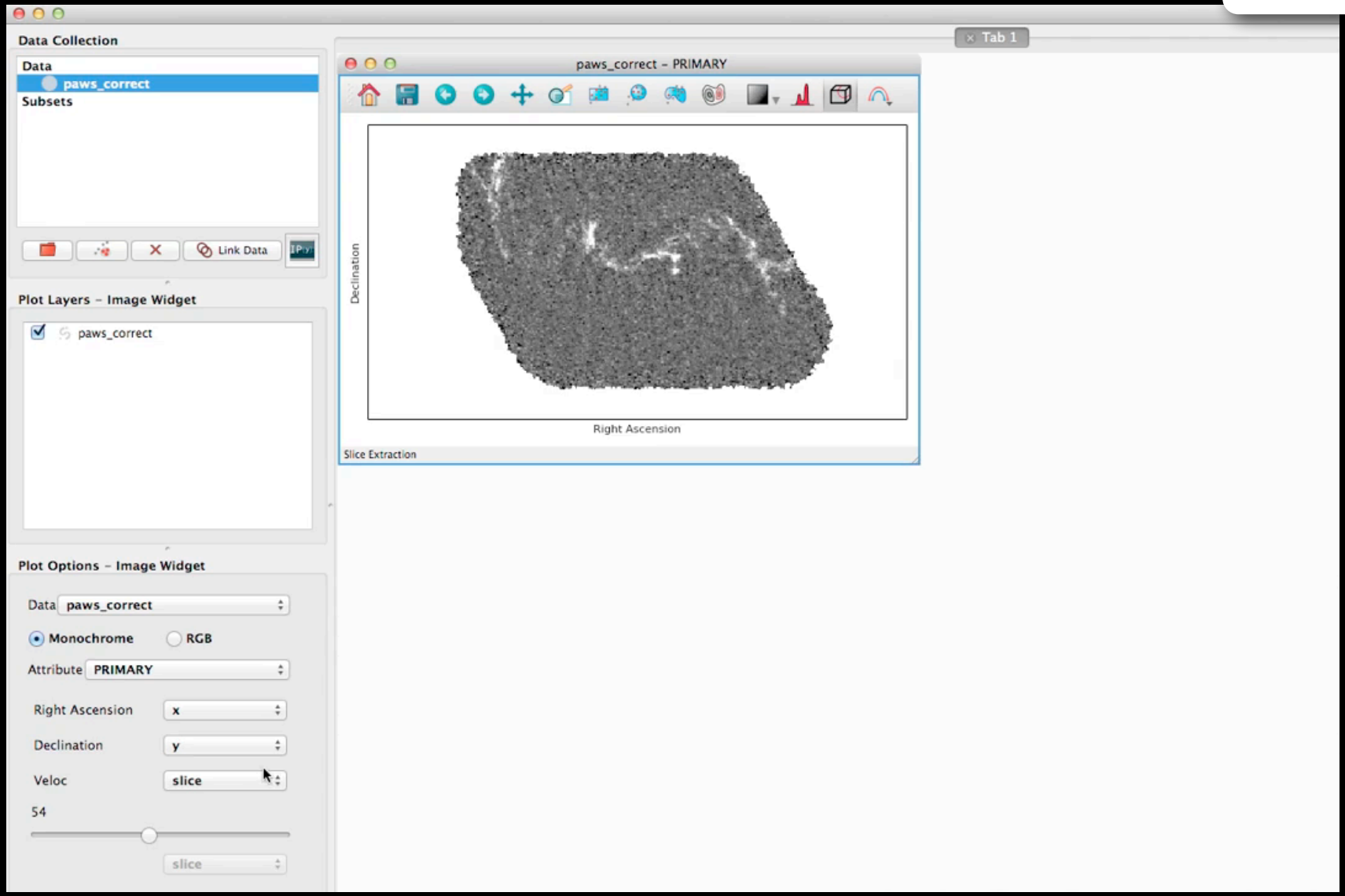


video by Tom Robitaille, lead glue developer

glue created by: C. Beaumont, M. Borkin, P. Qian, T. Robitaille, and A. Goodman, PI

LINKED VIEWS OF HIGH-DIMENSIONAL DATA (IN PYTHON)

GLUE

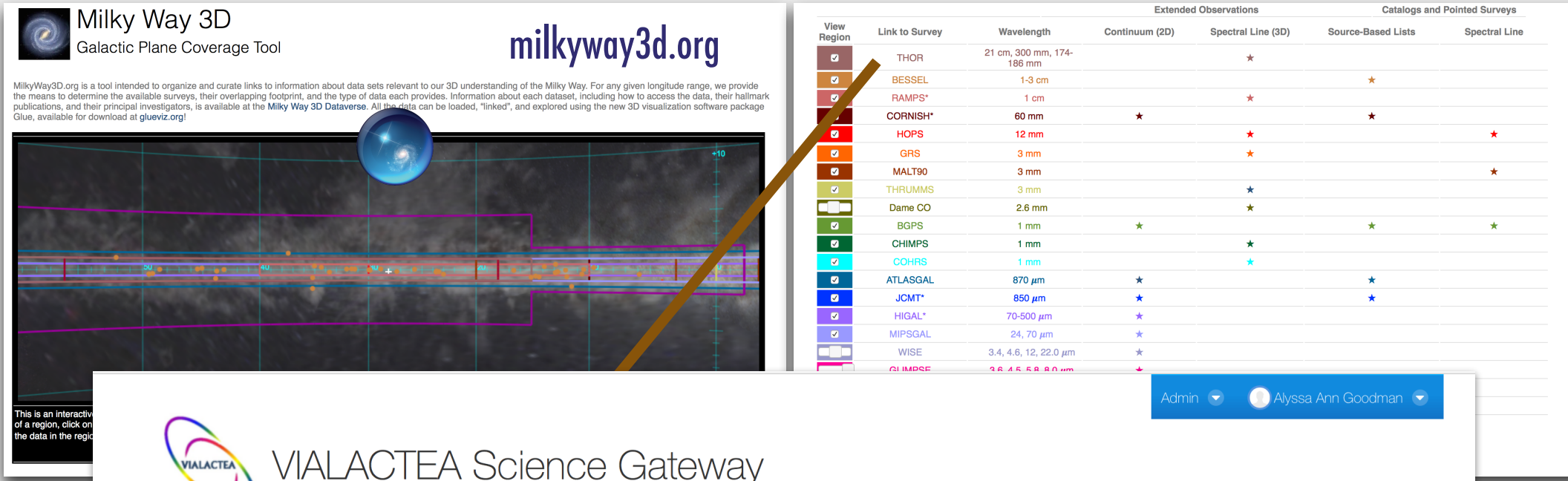


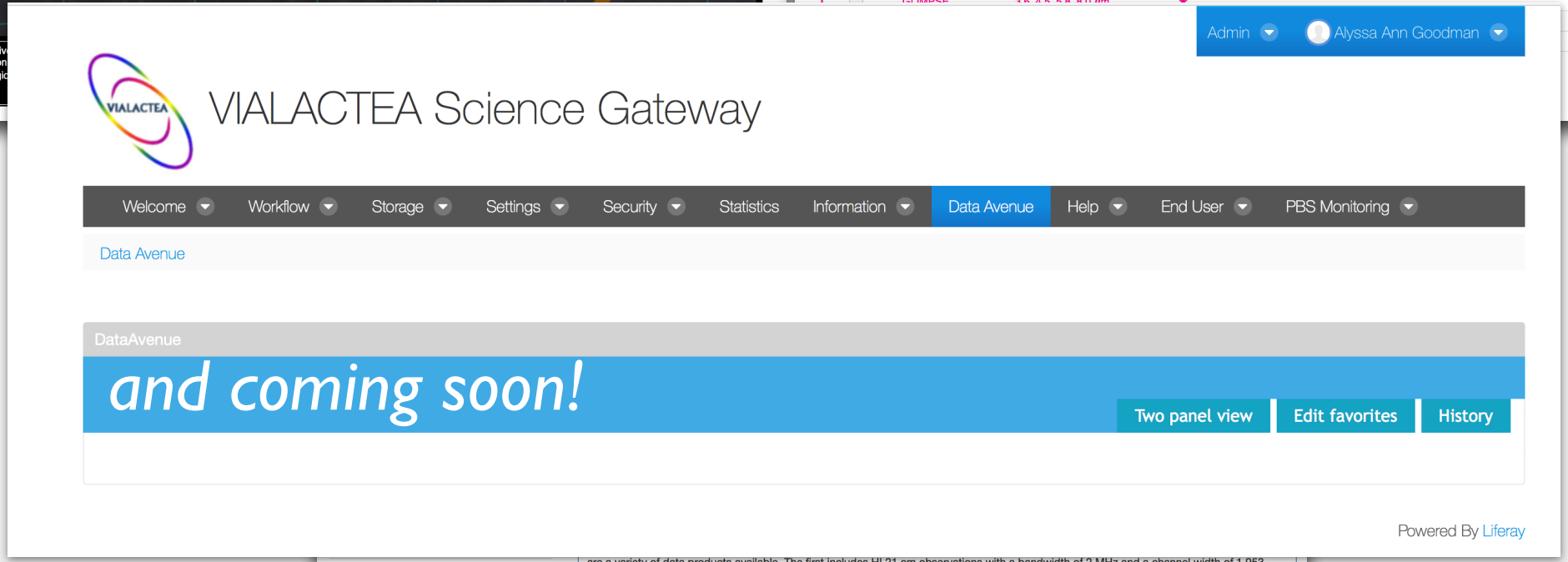
video by Chris Beaumont, glue developer

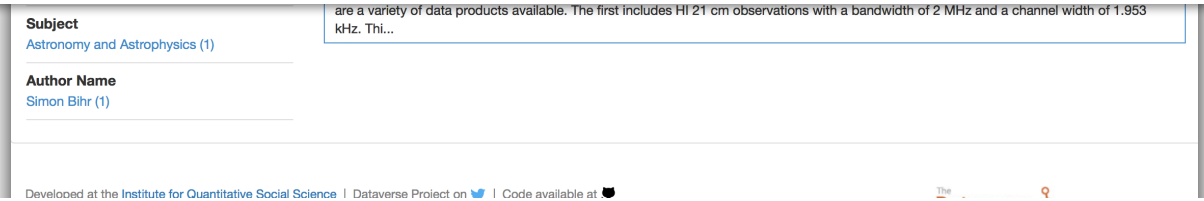
glue created by: C. Beaumont, M. Borkin, P. Qian, T. Robitaille, and A. Goodman, PI

[+demo]

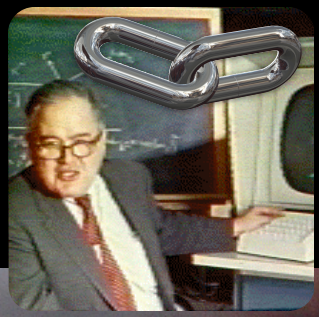
WIDE, BIG, OPEN DATA







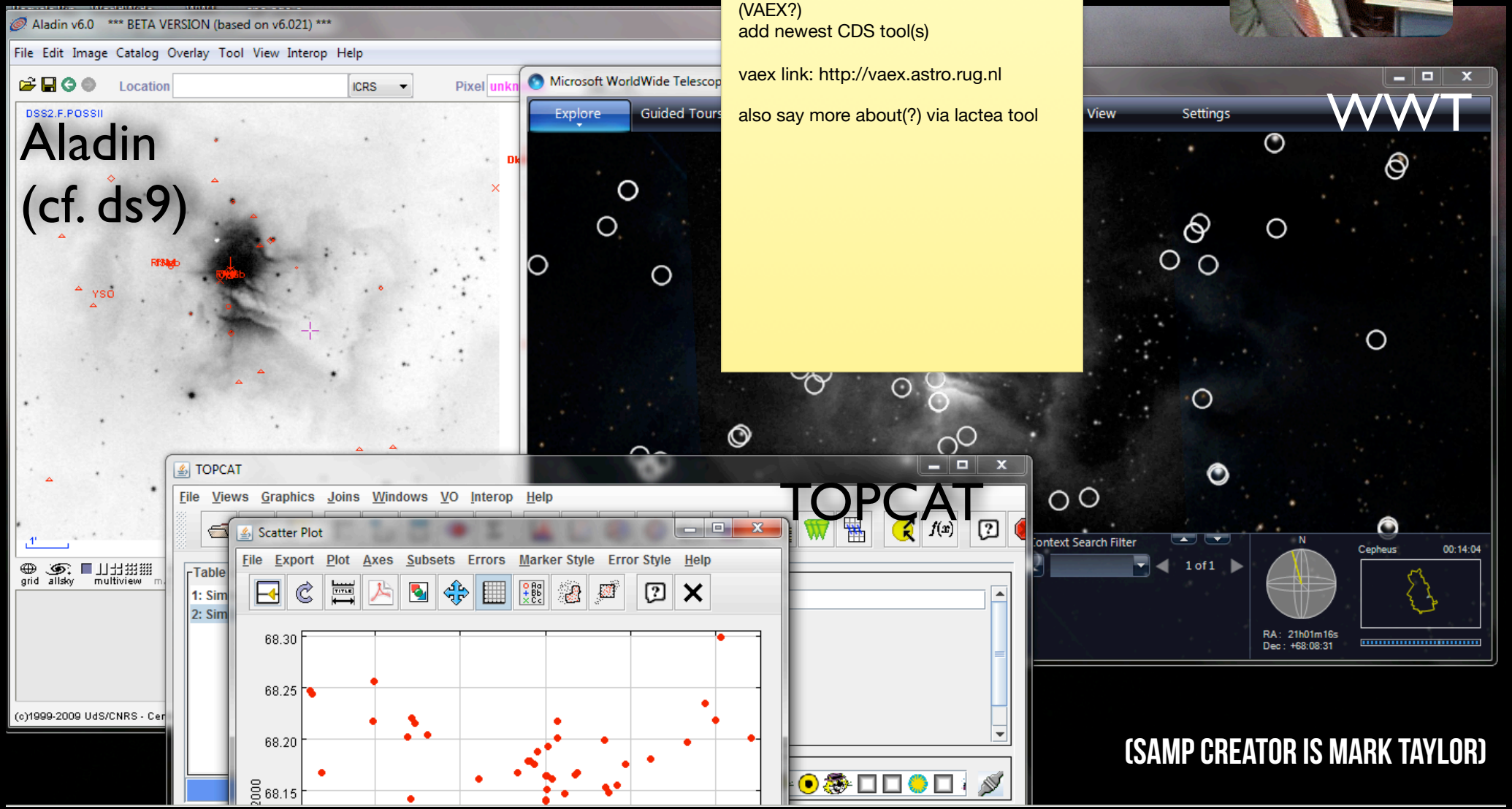
LINKED VIEWS OF HIGH-DIMENSIONAL DATA "SAMP"



update this slide re:TOPCAT, SAMP, (VAEX?)
add newest CDS tool(s)

vaex link: http://vaex.astro.rug.nl

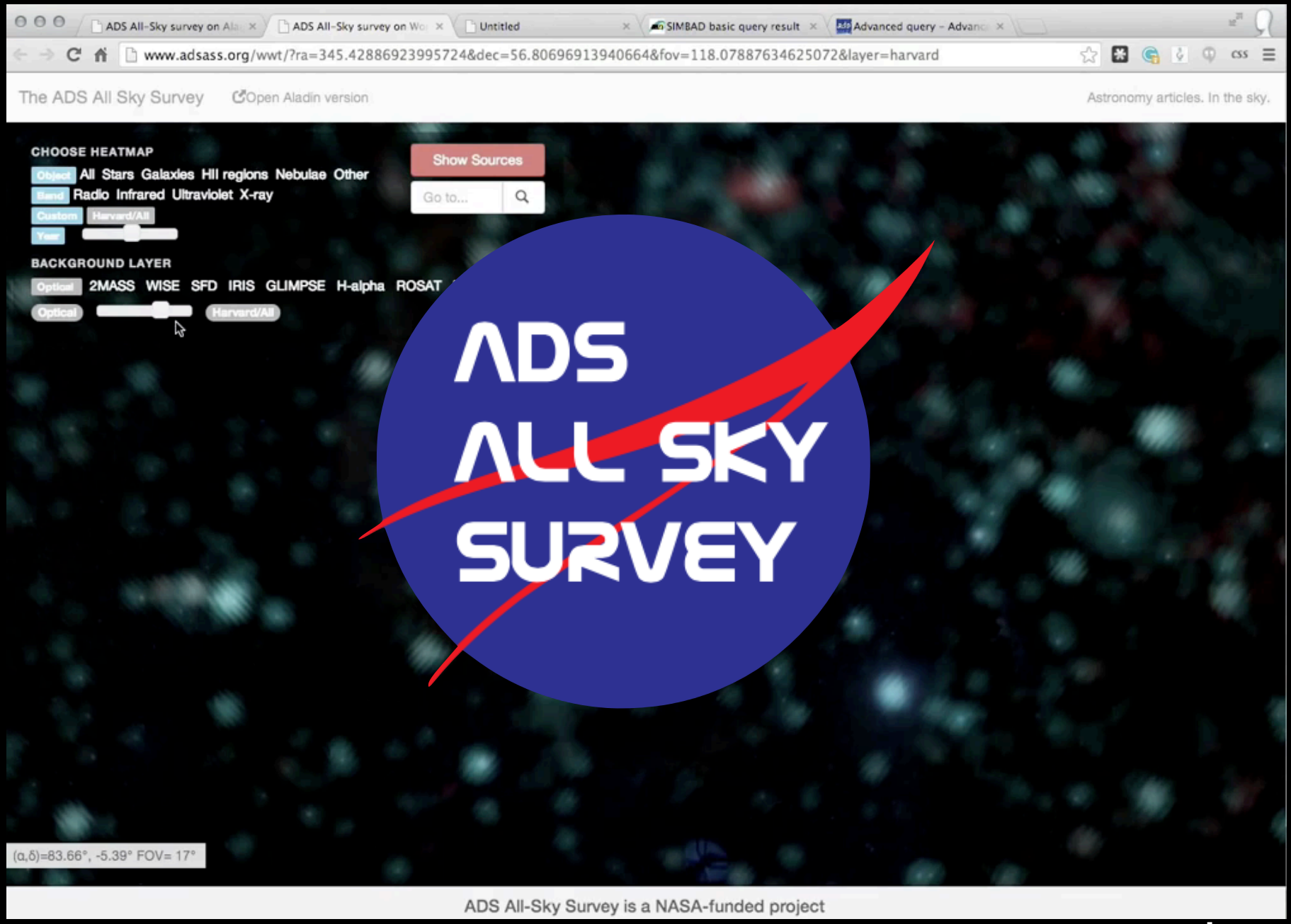
also say more about(?) via lactea tool



(SAMP CREATOR IS MARK TAYLOR)

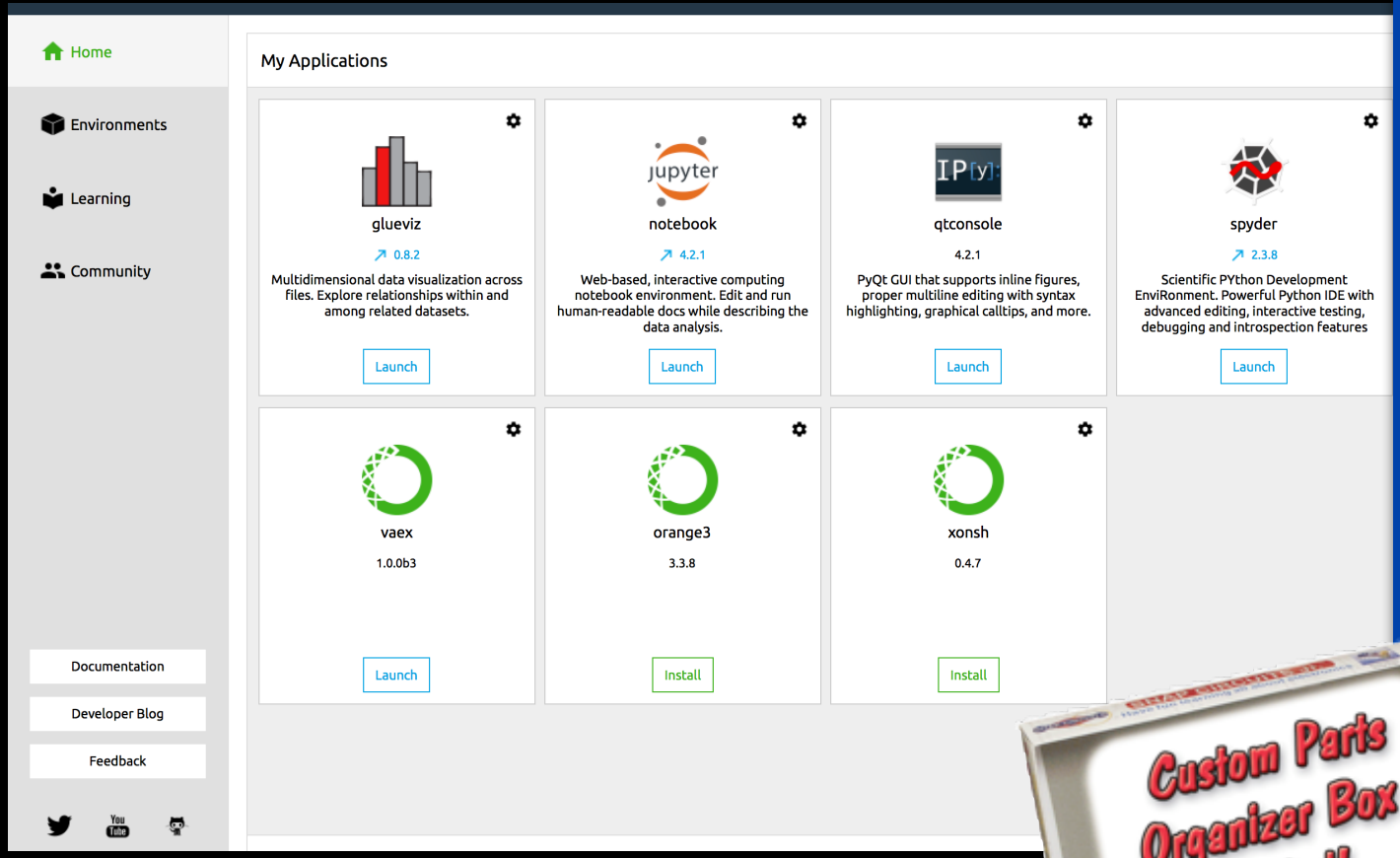
figure, showing SAMP screenshot, reproduced from Goodman 2012, "Principles of High-Dimensional Data Visualization in Astronomy"

ADDING METADATA "DIMENSIONS"



adsass.org

THE FUTURE IS ABOUT **INTEGRATION**



python™

SciPy

jupyter

astropy
A Community Python Library for Astronomy

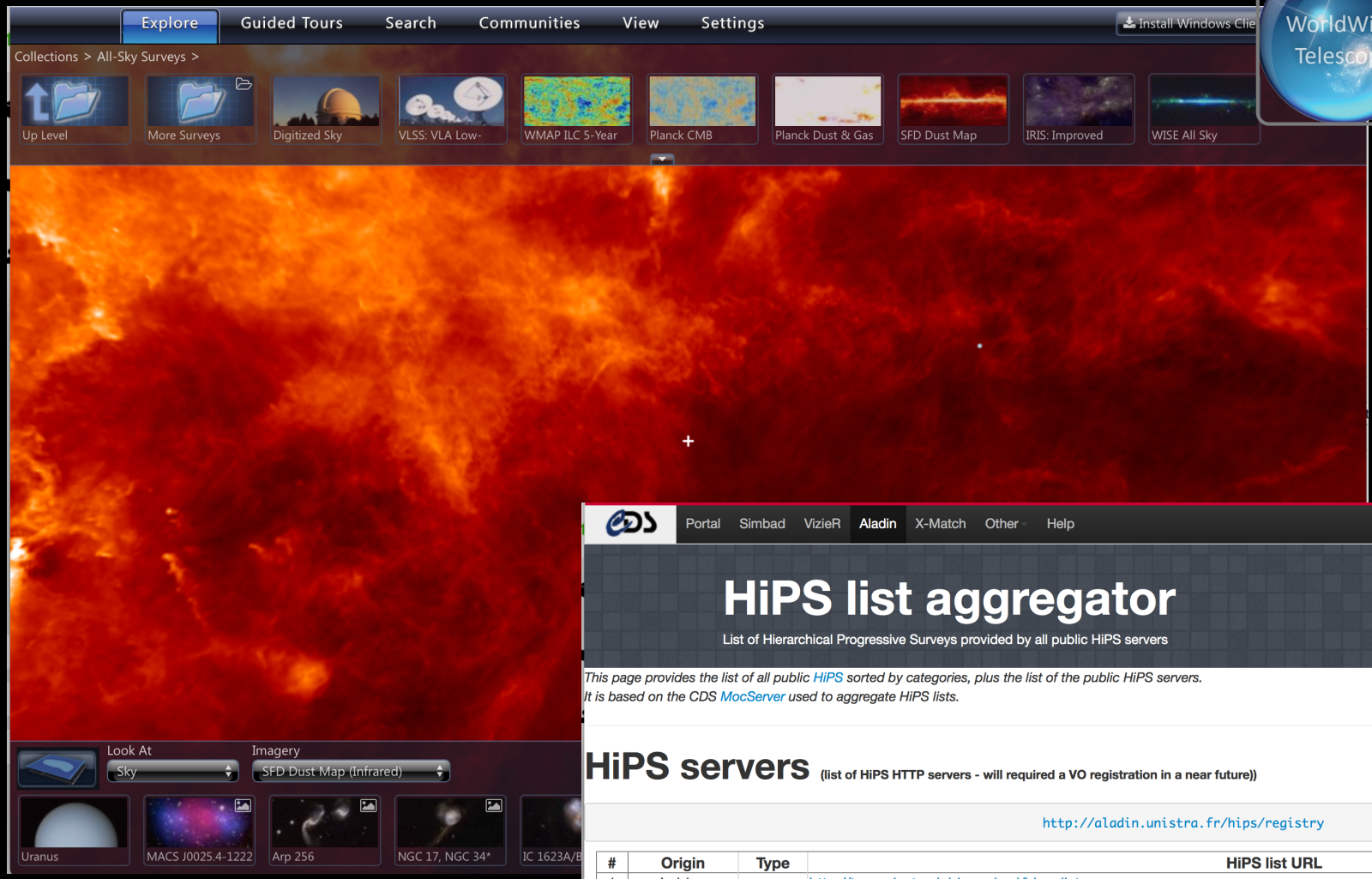
JavaScript

d3.js

plotly

HTML

THE FUTURE IS ABOUT **INTEGRATION**



WorldWide Telescope

Portal Simbad VizieR Aladin X-Match Other Help

HiPS list aggregator

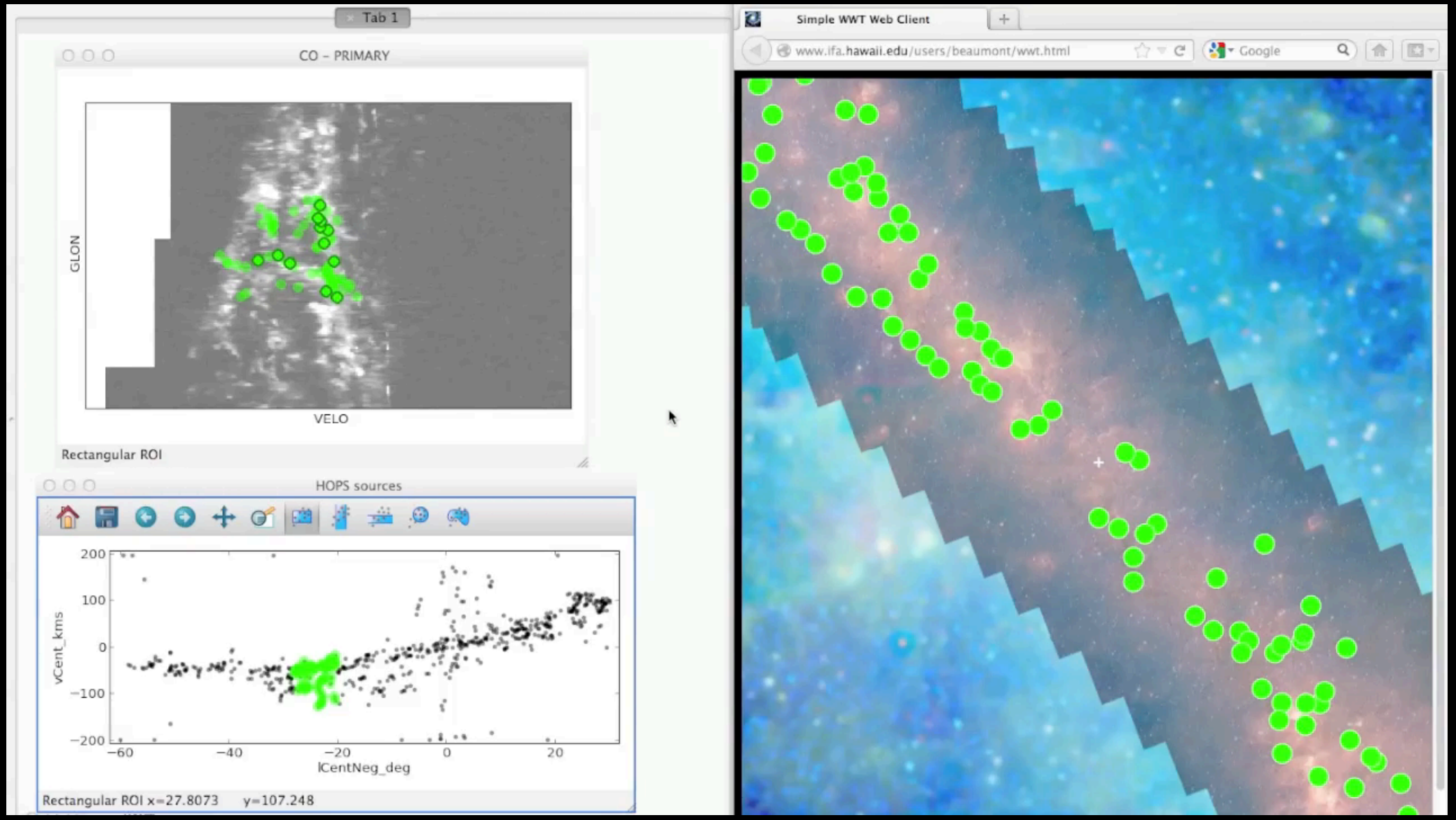
List of Hierarchical Progressive Surveys provided by all public HiPS servers

This page provides the list of all public HiPS sorted by categories, plus the list of the public HiPS servers.
It is based on the CDS MocServer used to aggregate HiPS lists.

HiPS servers (list of HiPS HTTP servers - will required a VO registration in a near future))

http://aladin.unistra.fr/hips/registry

#	Origin	Type	HiPS list URL
1	Leiden		http://tgssadr.strw.leidenuniv.nl/hips_list
2	IRAP		http://cade.irap.omp.eu/documents/Ancillary/4Aladin/hipslist-IRAP.txt
3	SSC		http://saada.unistra.fr/cgi-bin/hipslist
4	CDS		http://alasky.u-strasbg.fr/hipslist
5	CDS		http://alaskybis.u-strasbg.fr/hipslist
6	CDS	*catalog*	http://axel.u-strasbg.fr/HiPSCatService/hiplist
7	AMIGA		http://amiga.iaa.es/hipslist
8	svo.cab		http://gtc.sdc.cab.inta-csic.es/hips/hipslist
9	IAS		http://healpix.ias.u-psud.fr/hipslist
10	ESAC		http://skies.esac.esa.int/hipslist
11	JAXA		http://darts.isas.jaxa.jp/astro/judo2/common/hipslist.cgi



Video courtesy of Chris Beaumont

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.

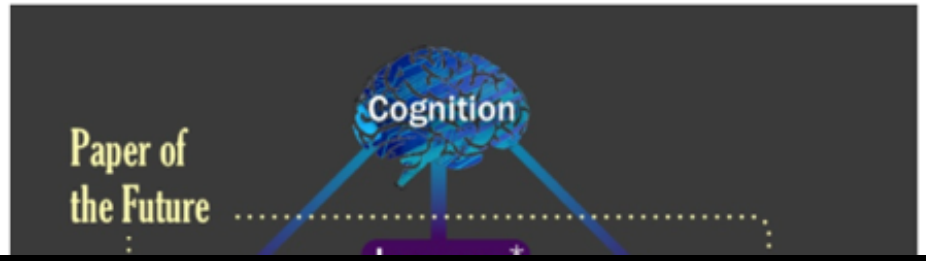
2

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

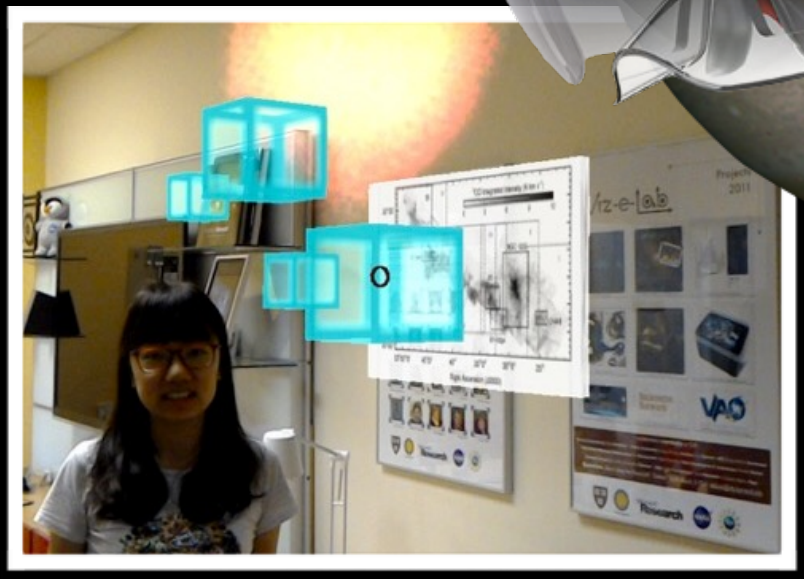
Konrad Hinsen 1 day ago · Public
That sounds good. I hope more repositories will follow the example of Dataverse. Figshare in particular has a very different attitude, encouraging researchers to deposit as much as possible. That's perhaps a good strategy to change habits, but in the long run it could 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"

d3po/Authorea: Peek, Price-Whelan, Pepe, Beaumont, Borkin, Newton; PoF: Goodman, Peek; WWT: Wong, Fay et al.; Astrometry.net: Hogg, Lang, Roweis et al.

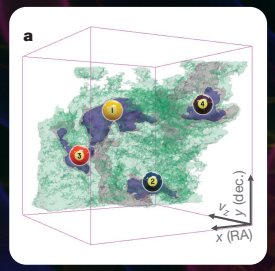
[video, demos]

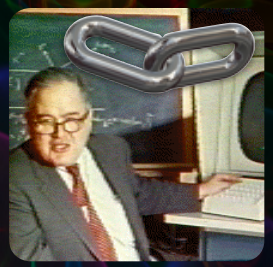
WHAT'S NEXT?



THE VALUE OF HIGH-DIMENSIONAL DATA VISUALIZATION IN SCIENCE








Alyssa A. Goodman
Harvard-Smithsonian Center for Astrophysics
& Radcliffe Institute
@aagie

Thomas Robitaille
lead Developer for glue
@astrofrog

background image credit: ALMA (ESO/NAOJ/NRAO)/Erik Rosolowsky

extra slides

MORE?

projects.iq.harvard.edu/
seamlessastronomy/
presentations

JOHN TUKEY'S LEGACY

PRIM-9

PRIM-H

DataDesk®

glue

XGobi

GGobi

RGGobi

Spotfire®

Polaris

tableau SOFTWARE

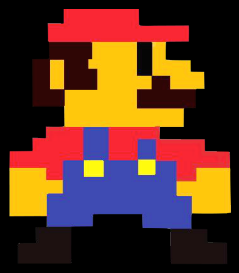
1970 1980 1990 2000 2010

1992

Super Mario Kart: Rainbow Road (1992)

"Today"

Mario Kart 8: Rainbow Road (2014)



Mario franchise, copyright Nintendo