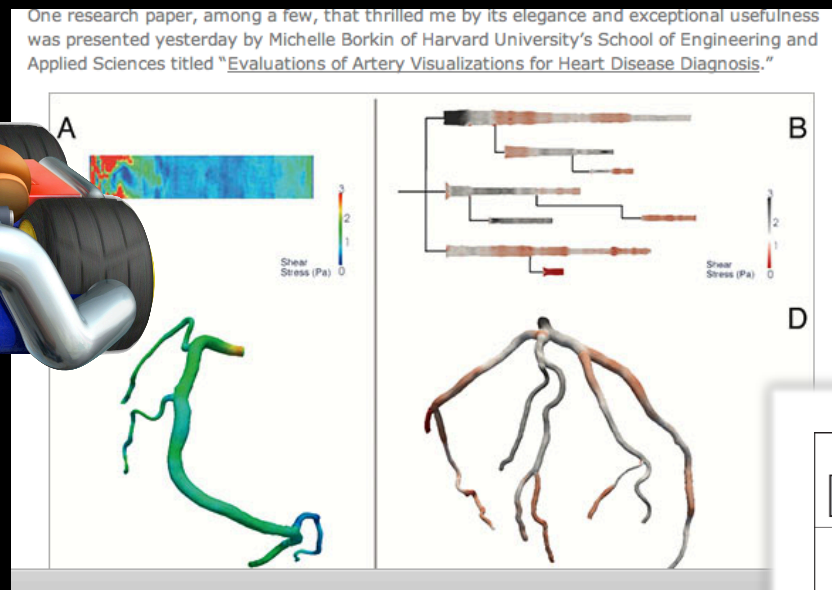
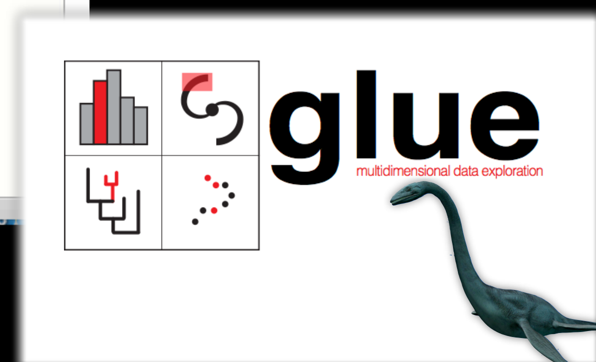
Seeing More in Data



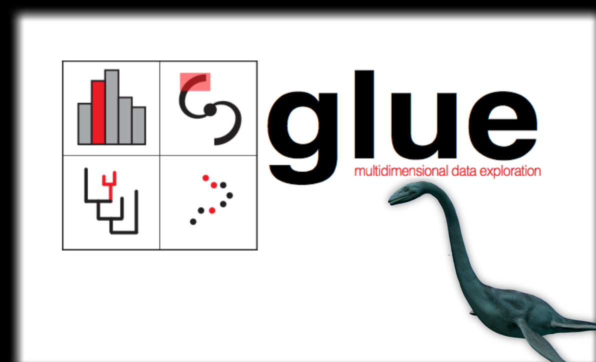


Viz-e-Lab

Alyssa A. Goodman

Harvard-Smithsonian Center for Astrophysics

Seeing More in Data



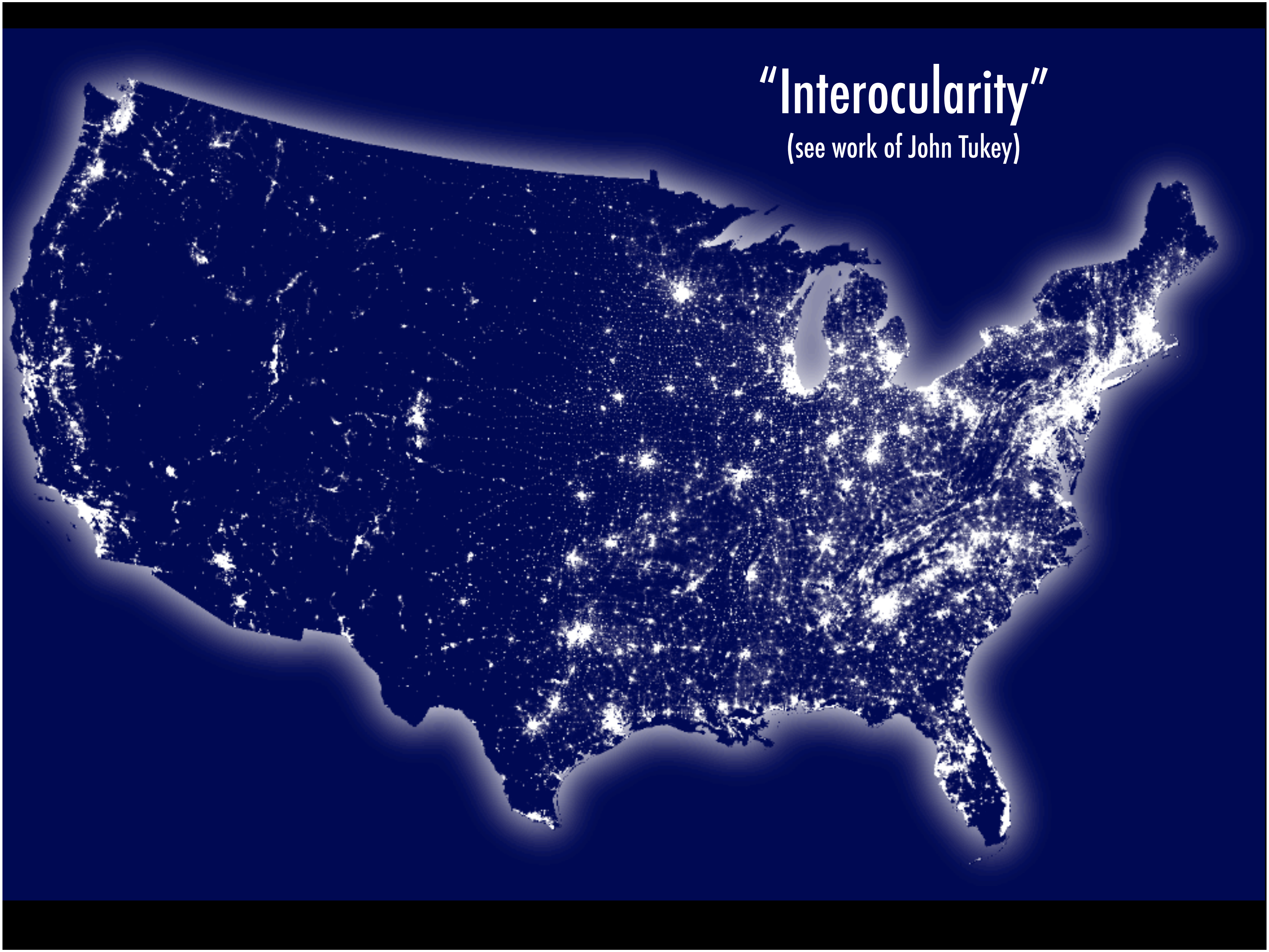
Viz-e-Lab

Alyssa A. Goodman
Harvard-Smithsonian Center for Astrophysics

Relative Strengths

Pattern Recognition
Creativity

Calculations


"Interocularity"
(see work of John Tukey)

What...

...is easier now than before?

fast computation, animation, 3D

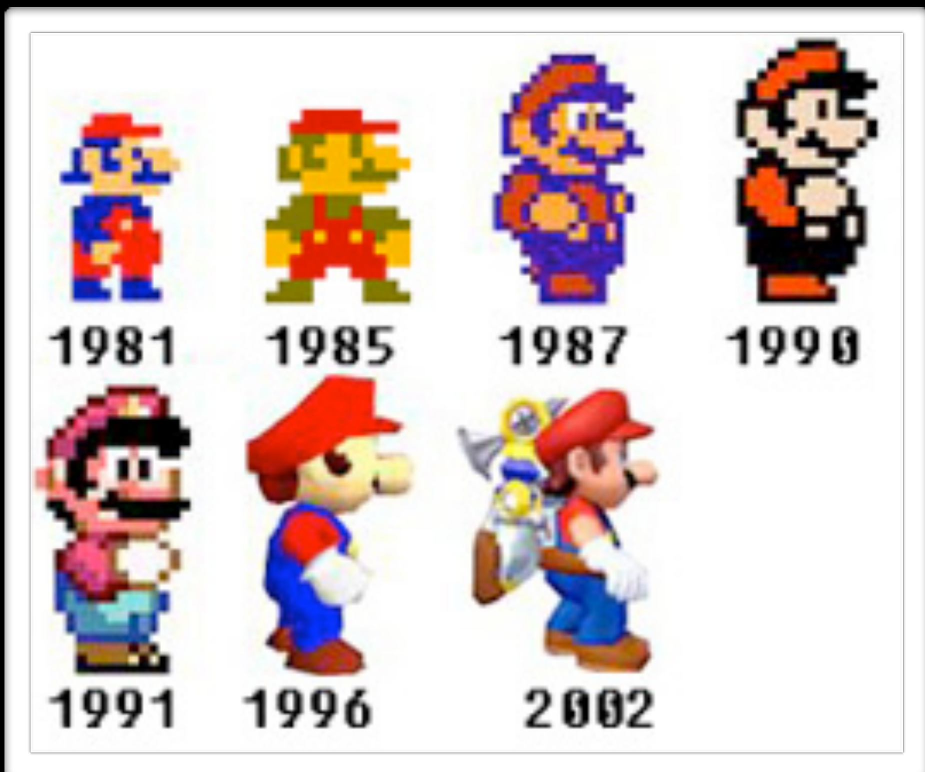
...was easier before than now?

craftsmanship

...should be easier in the future?

modular craftsmanship, linked views

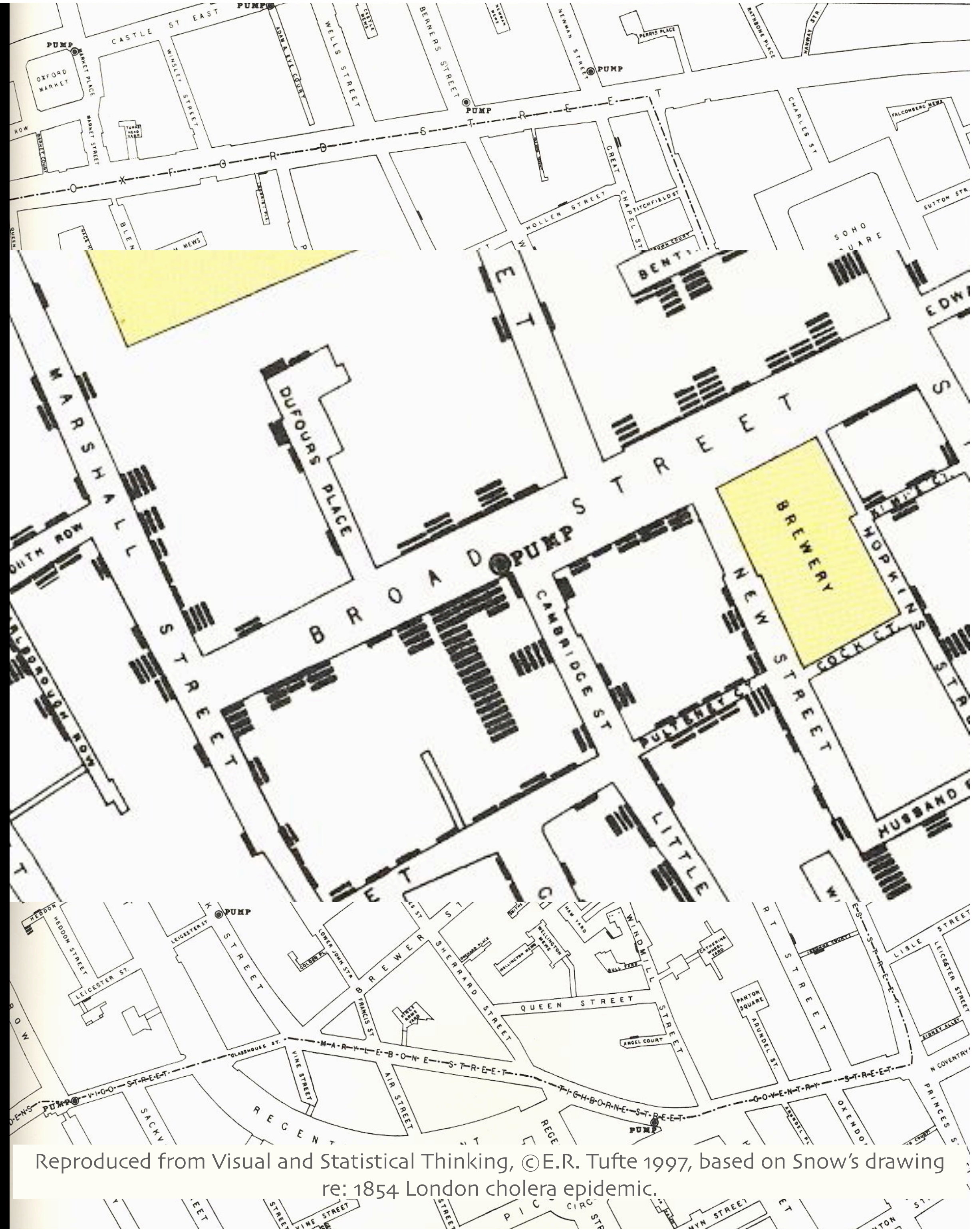
“Easier”



2011

Craftsmanship

(in 1854)



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

Craftsmanship

(in 1854)

Displaying
"high-dimensional" data

with

"multi-functioning
graphical elements"

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

Data • Dimensions • Display

Craftsmanship
(in 1854)

Displaying
"high-dimensional" data

with

"multi-functioning
graphical elements"

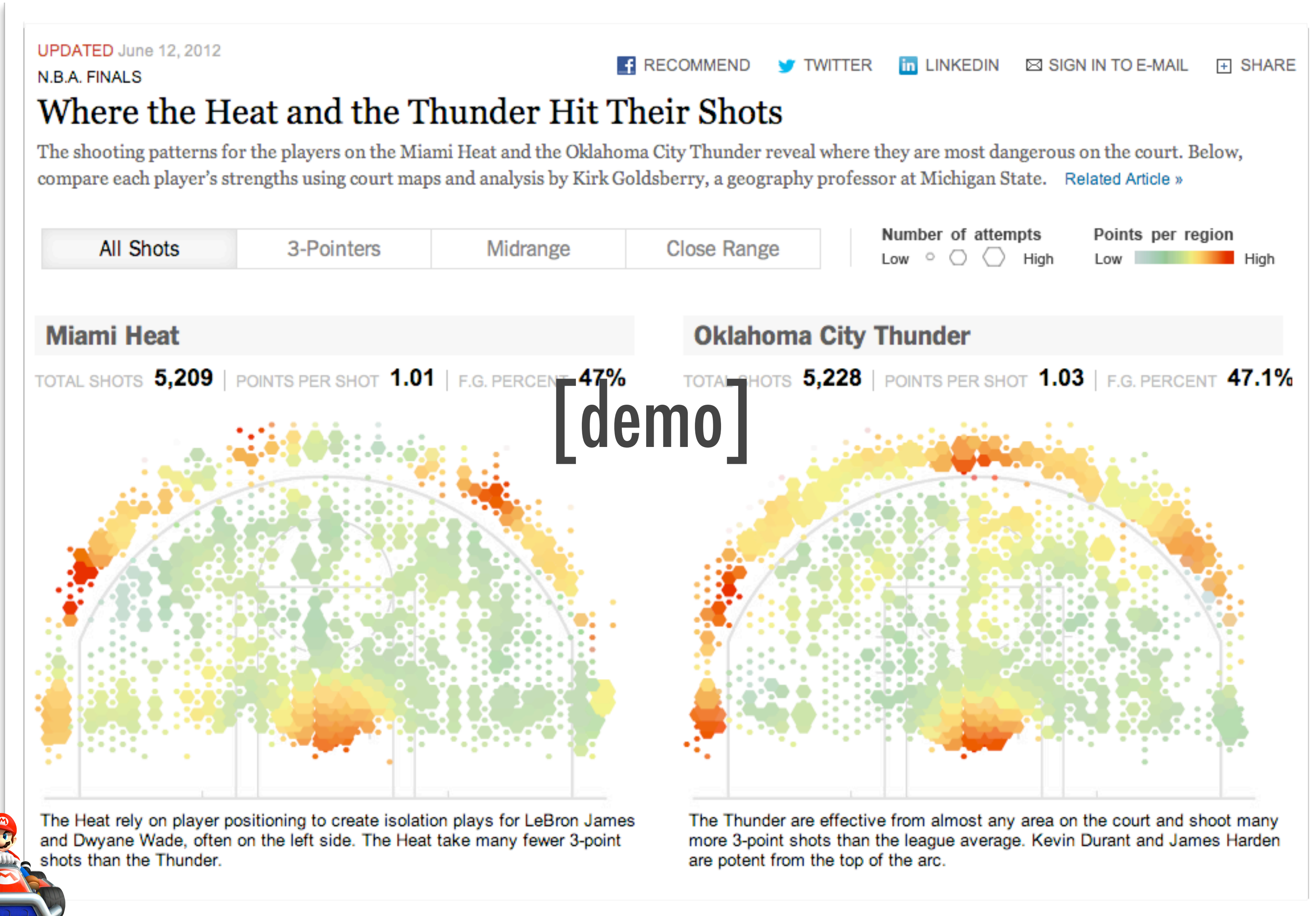
Reproduced from Visual and Statistical Thinking, ©E.R. Tufte 1997, based on Snow's drawing re: 1854 London cholera epidemic.

What Computers Can Let us Craft (2008)

Elements...

- ✓Maps
- ✓Tables
- ✖Graphs
- ✓Charts
- ✓Illustrations
- ✓Combinations

What Computers (D3) Can Let us Craft (2012)



[demo]

http://www.nytimes.com/interactive/2012/06/11/sports/basketball/nba-shot-analysis.html?_r=0

What...

...is easier now than before?

fast computation, animation, 3D

...was easier before than now?

craftsmanship

...should be easier in the future?

modular craftsmanship, linked views

What...

...is easier now than before?

fast computation, animation, 3D

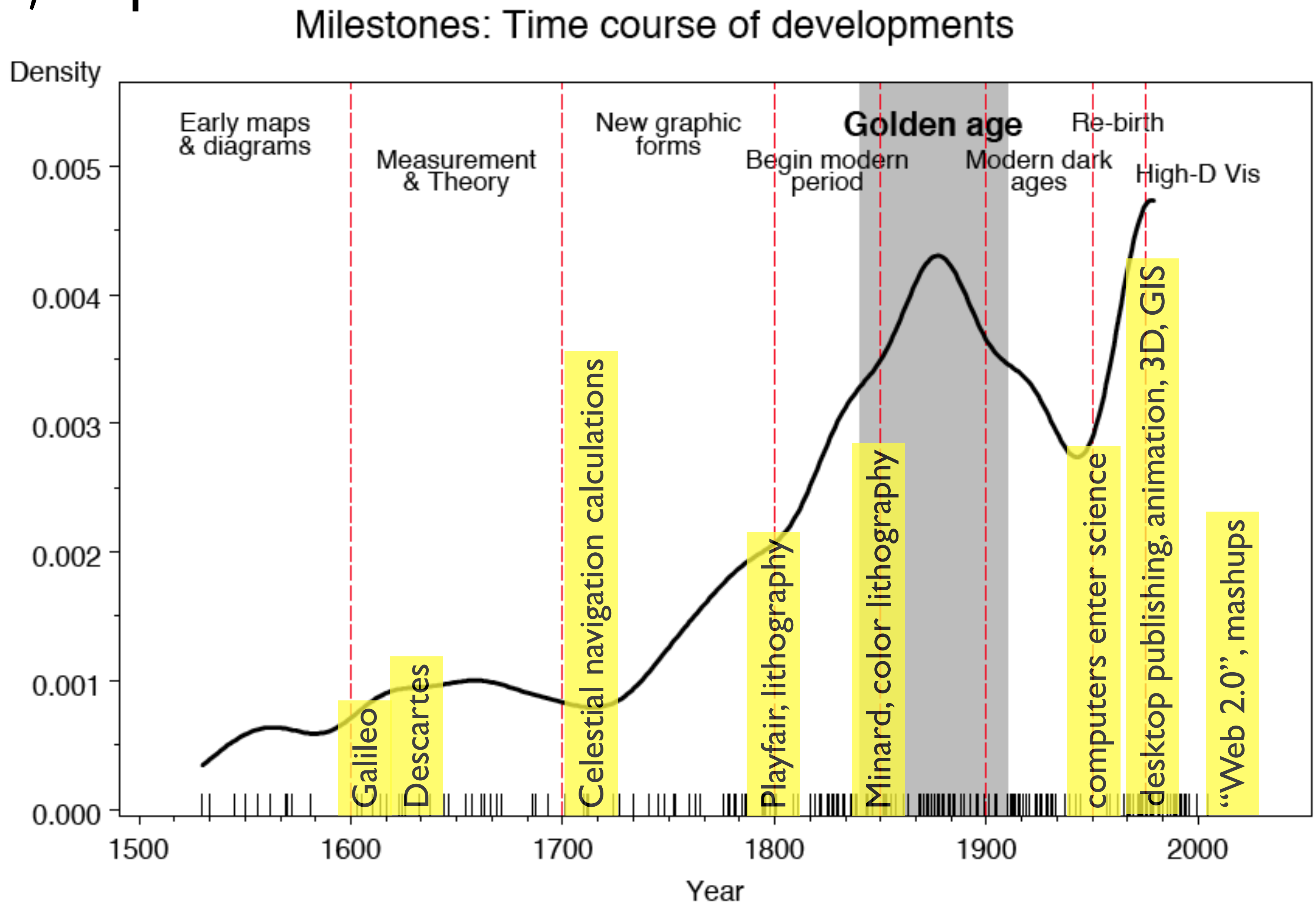
...was easier before than now?

craftsmanship

...should be easier in the future?

modular craftsmanship, linked views

First, the past...

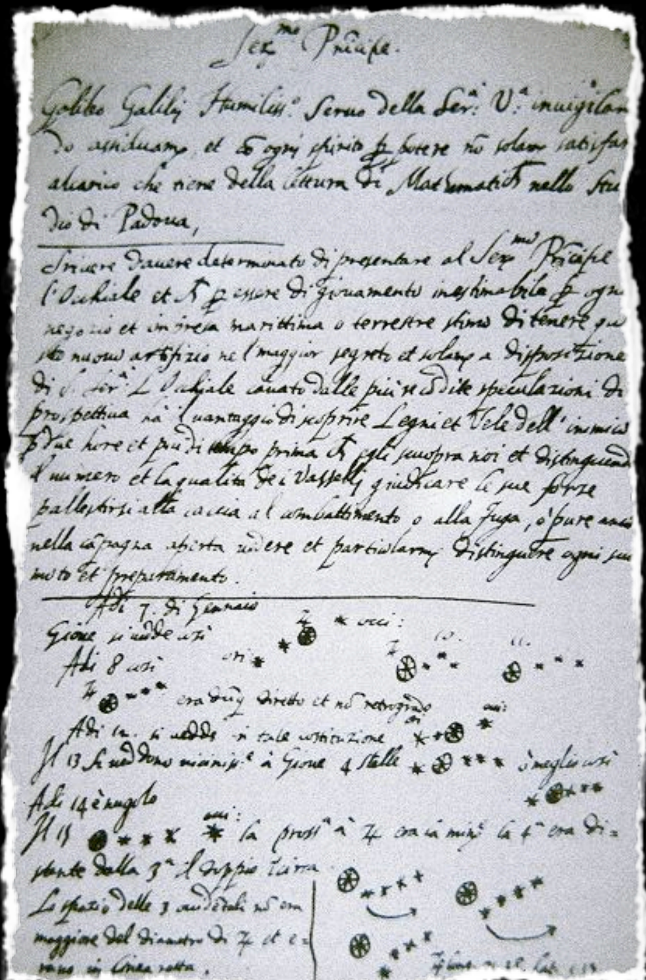


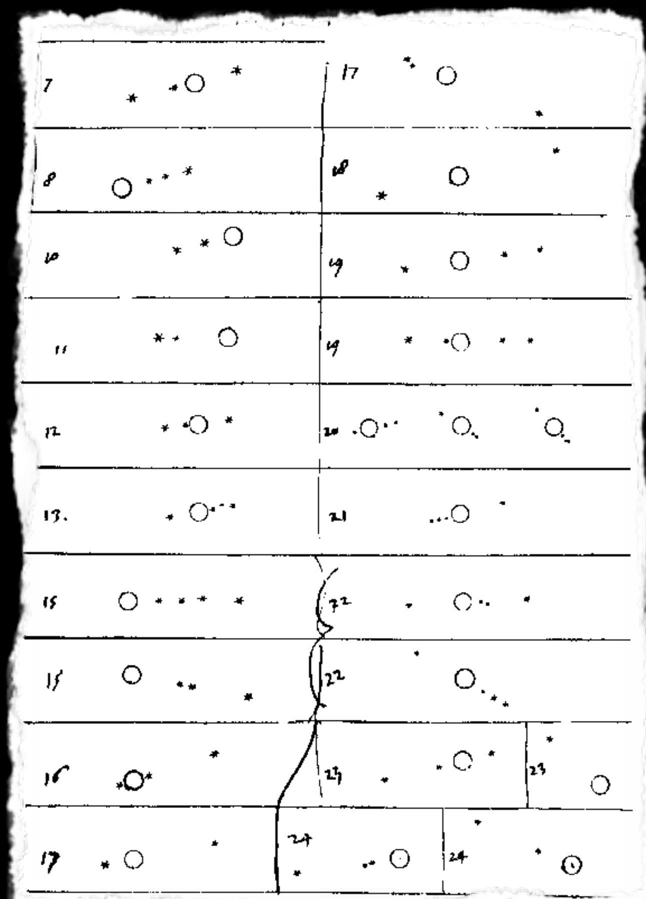
adapted from Friendly, "The Golden Age of Statistical Graphics," *Statistical Science*, 2009

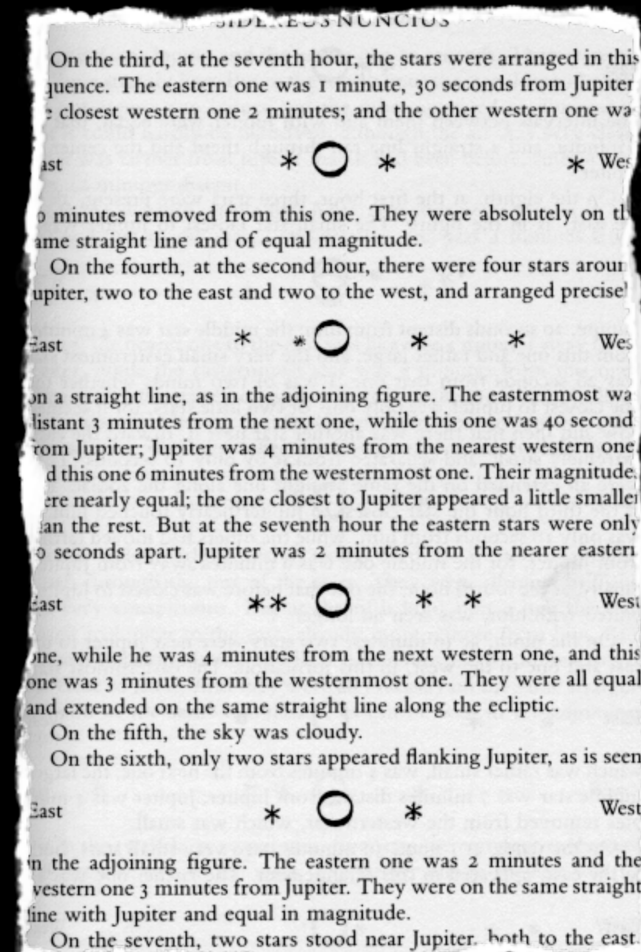
Galileo

Galileo Galilei

(1564-1642)






SIDEREUS NUNCIUS

On the third, at the seventh hour, the stars were arranged in this
quence. The eastern one was 1 minute, 30 seconds from Jupiter
closest western one 2 minutes; and the other western one wa

ast * O * * Wes

o minutes removed from this one. They were absolutely on th
ame straight line and of equal magnitude.
On the fourth, at the second hour, there were four stars arour
upiter, two to the east and two to the west, and arranged precise

East * * O * * Wes

on a straight line, as in the adjoining figure. The easternmost wa
listant 3 minutes from the next one, while this one was 40 second
rom Jupiter; Jupiter was 4 minutes from the nearest western one
d this one 6 minutes from the westernmost one. Their magnitude
ere nearly equal; the one closest to Jupiter appeared a little smalle
an the rest. But at the seventh hour the eastern stars were only
o seconds apart. Jupiter was 2 minutes from the nearer easter

East ** O * * West

one, while he was 4 minutes from the next western one, and this
one was 3 minutes from the westernmost one. They were all equal
and extended on the same straight line along the ecliptic.
On the fifth, the sky was cloudy.
On the sixth, only two stars appeared flanking Jupiter, as is seen

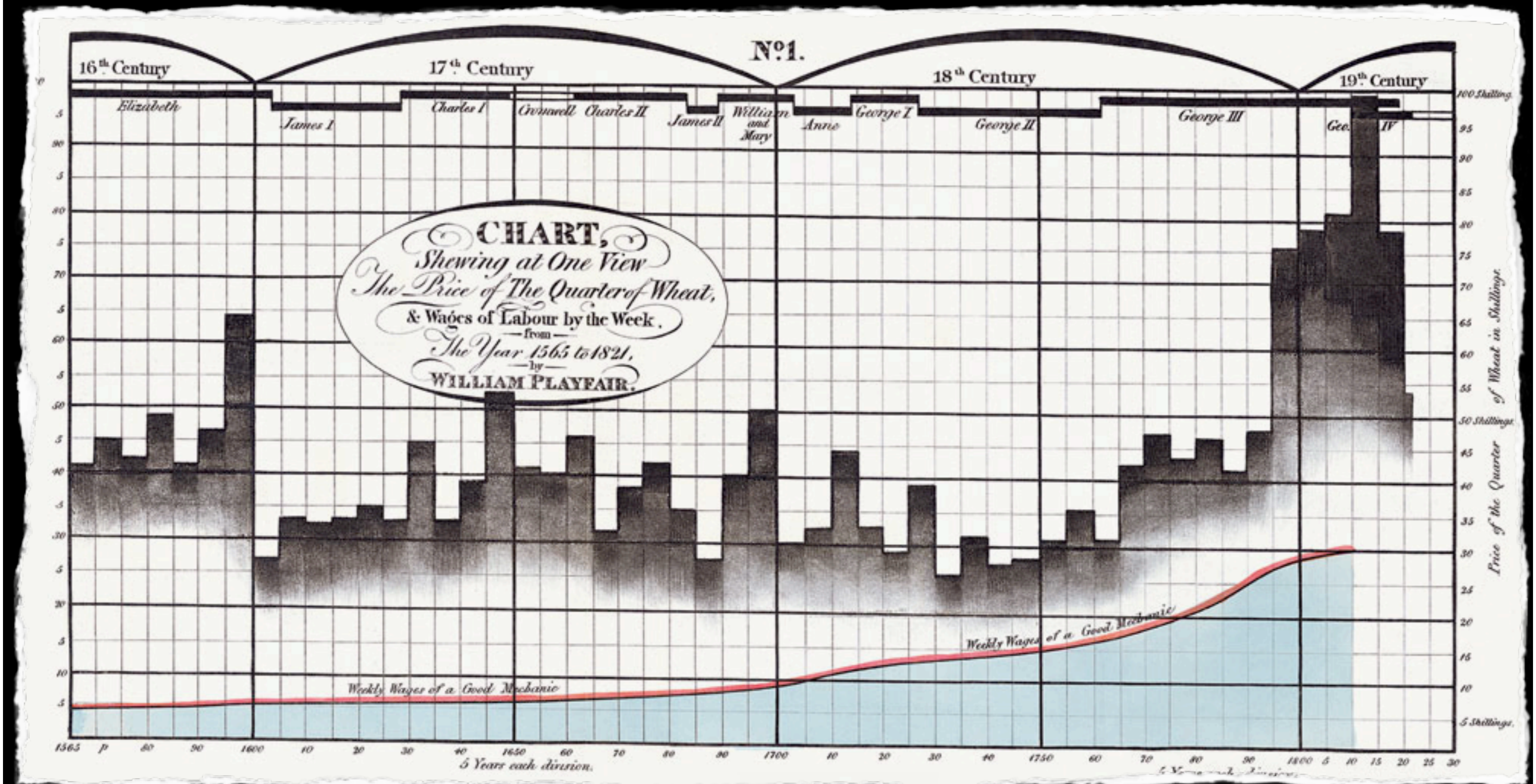
East * O * West

in the adjoining figure. The eastern one was 2 minutes and the
western one 3 minutes from Jupiter. They were on the same straight
line with Jupiter and equal in magnitude.
On the seventh, two stars stood near Jupiter, both to the east,

Notes for & re-productions of Siderius Nuncius

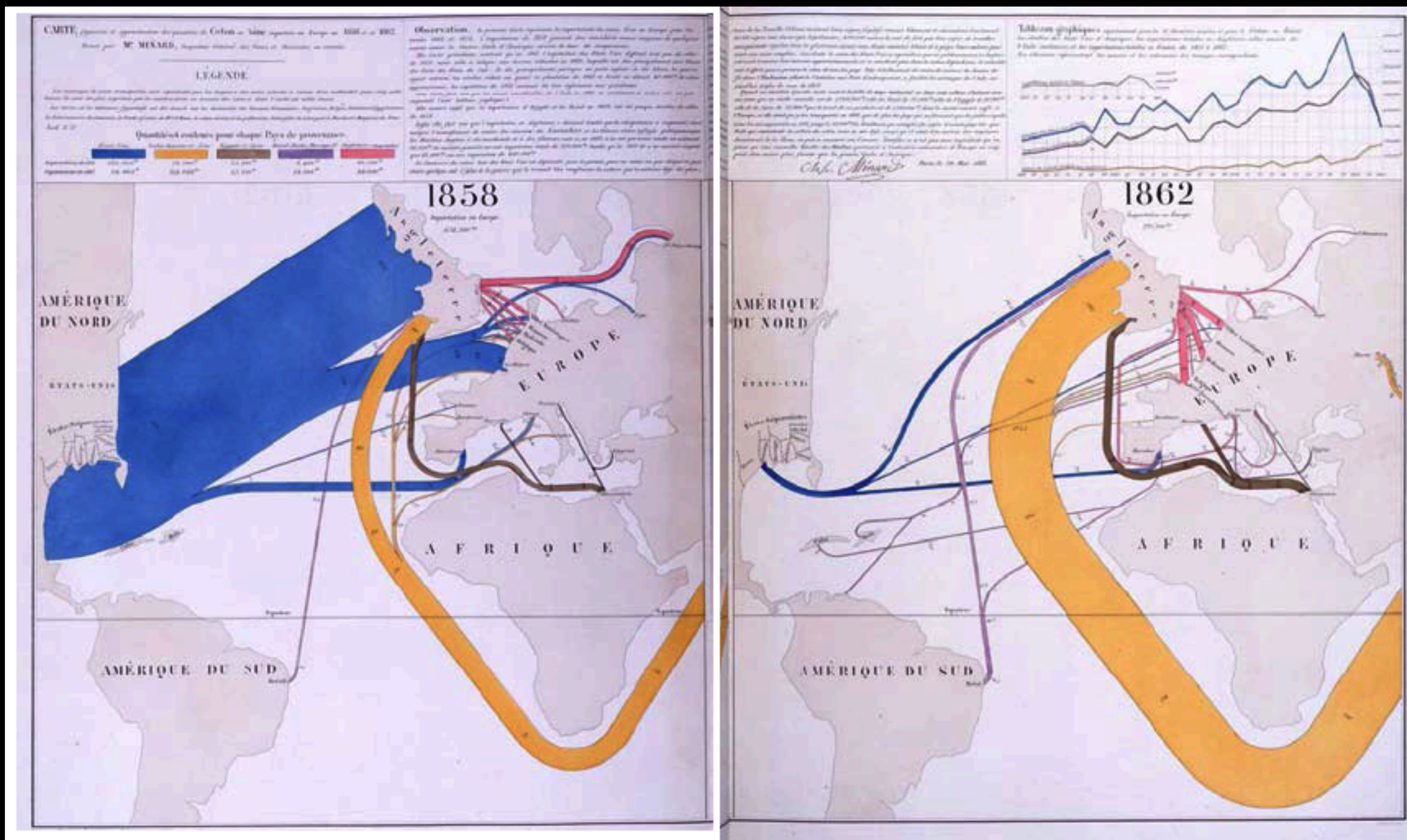
Playfair, lithography

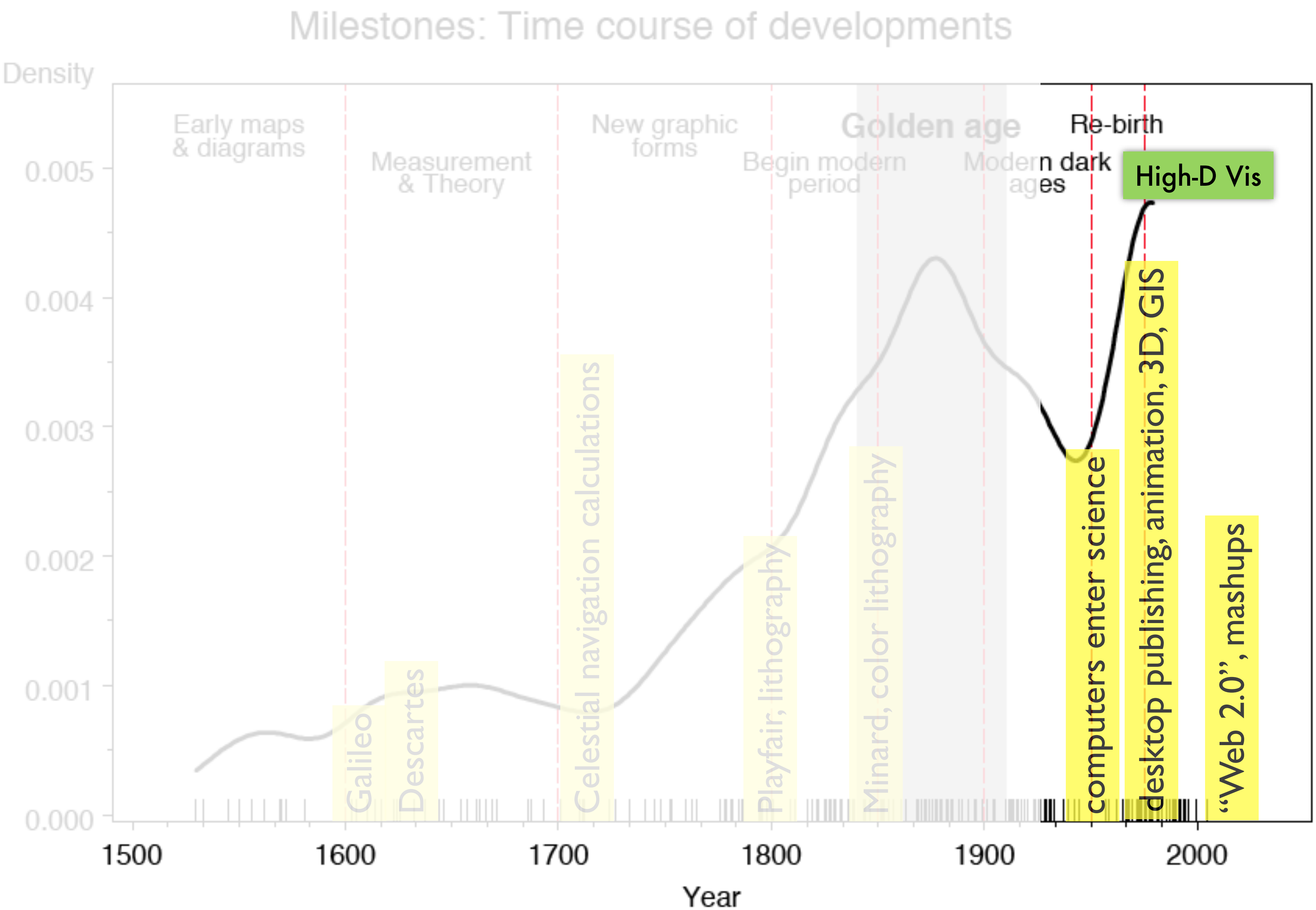
William Playfair
(1759-1823)



Minard, color lithography

Charles Joseph Minard, in color
(1781-1870)

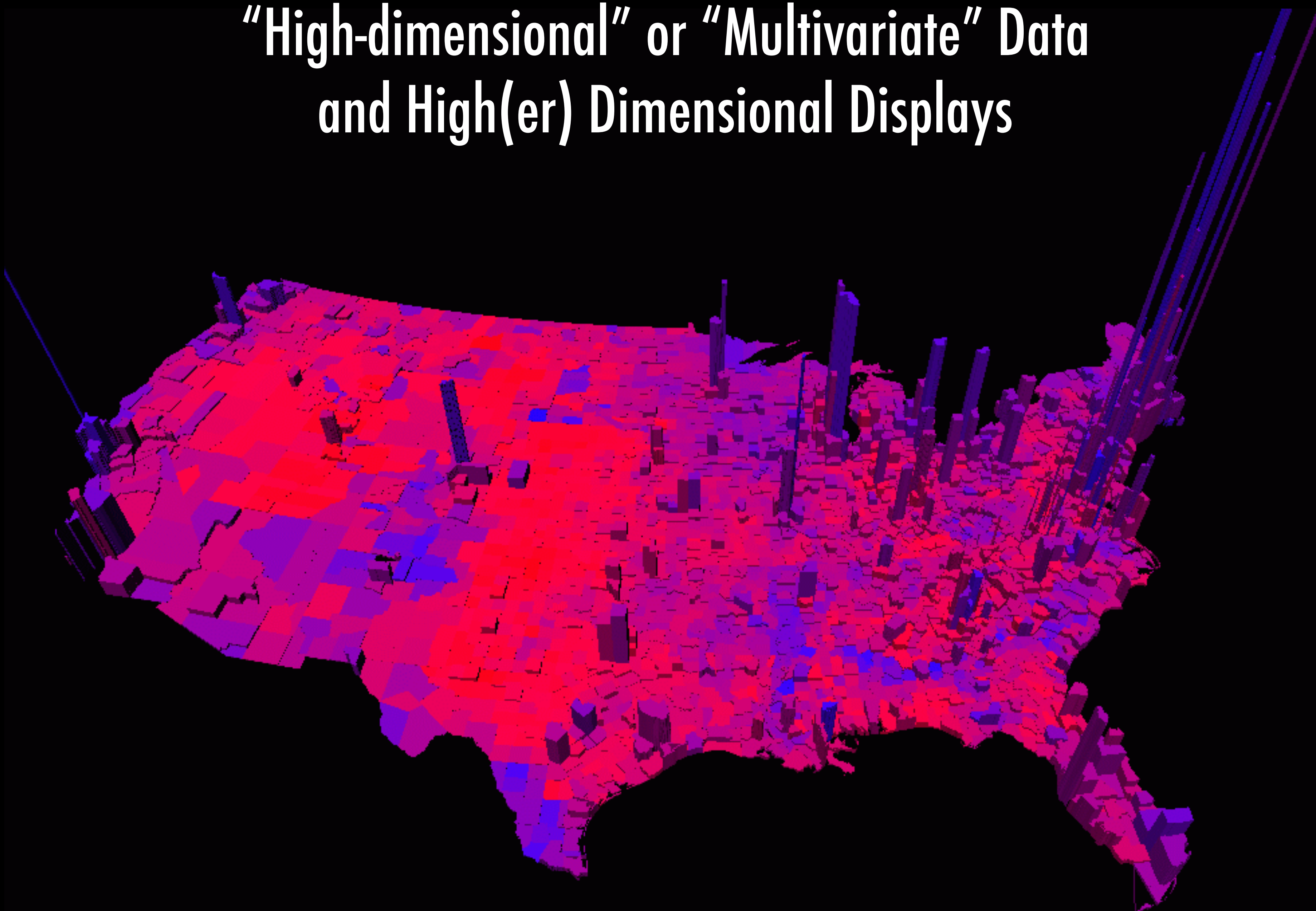




adapted from Friendly, "The Golden Age of Statistical Graphics," *Statistical Science*, in press (2008)

Data • Dimensions • Display

"High-dimensional" or "Multivariate" Data and High(er) Dimensional Displays

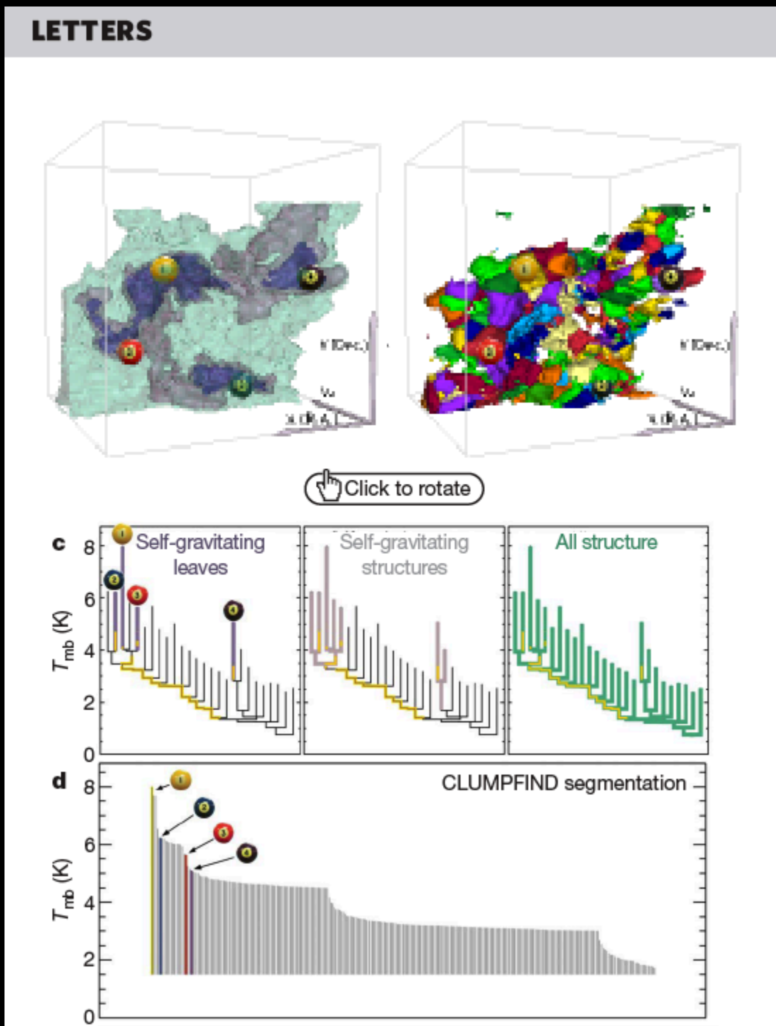


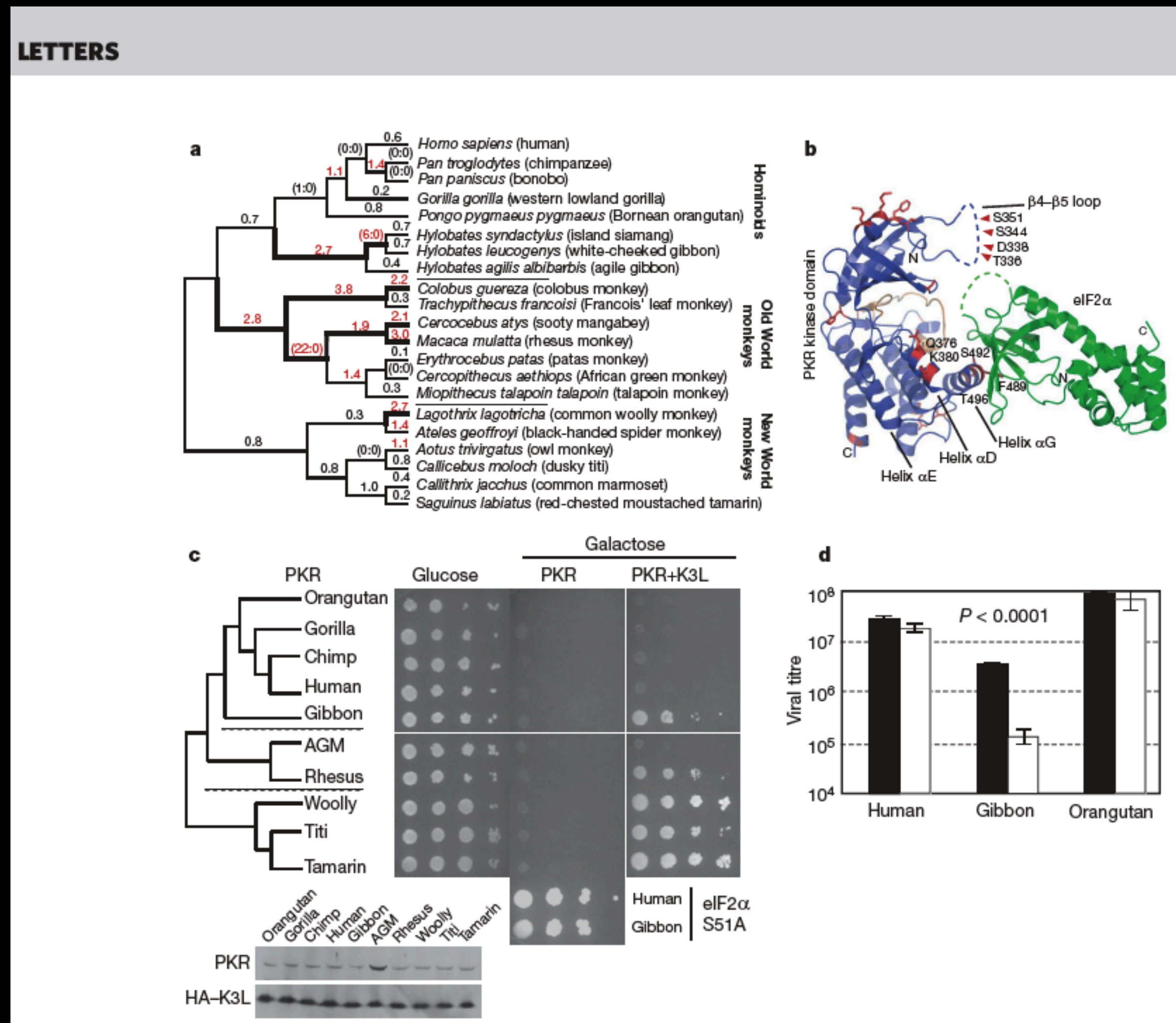
Robert J. Vanderbei

www.princeton.edu/~rvdb/JAVA/election2004/

"High-dimensional" or "Multivariate" Data

(Astronomy=Biology)

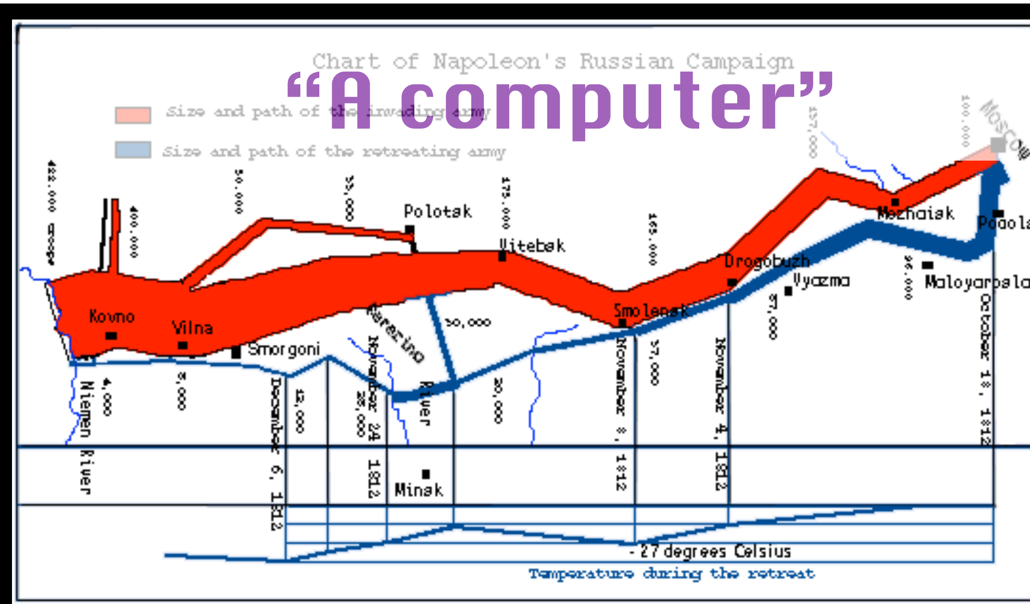


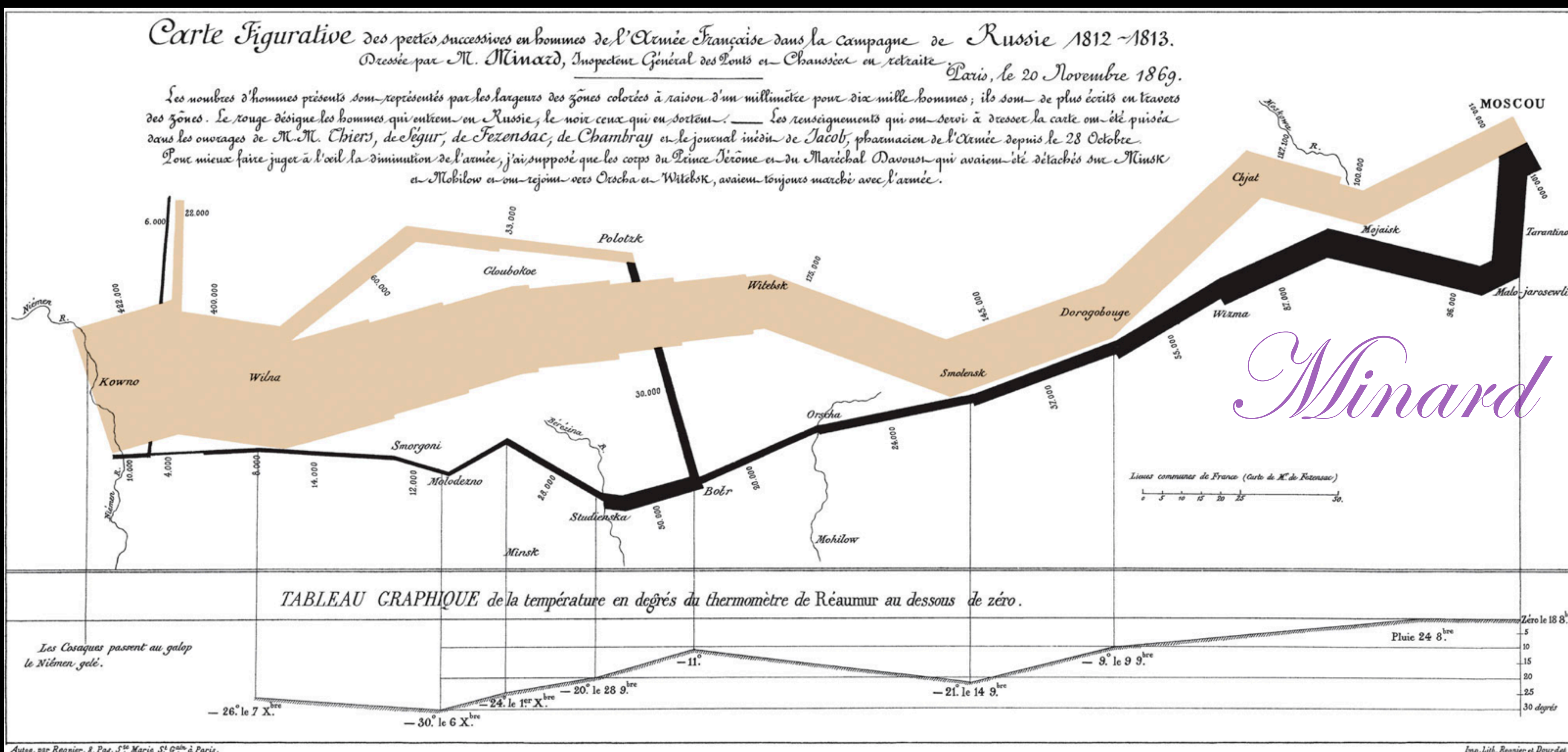


Goodman et al. *Nature*, 2009

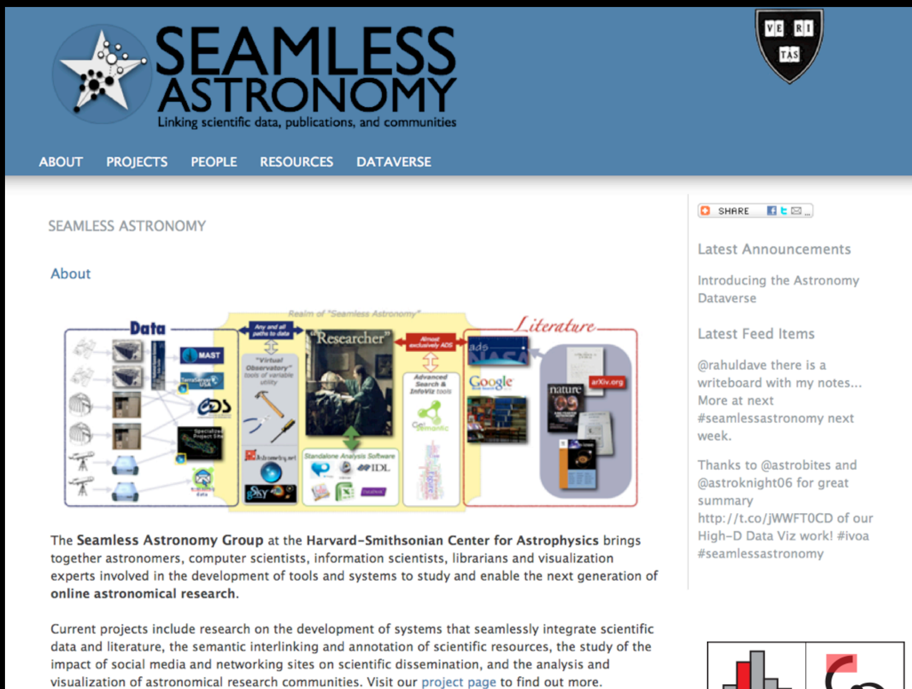
Elde et al. *Nature*, 2008

How much are we held back today by **digital** *tools?*

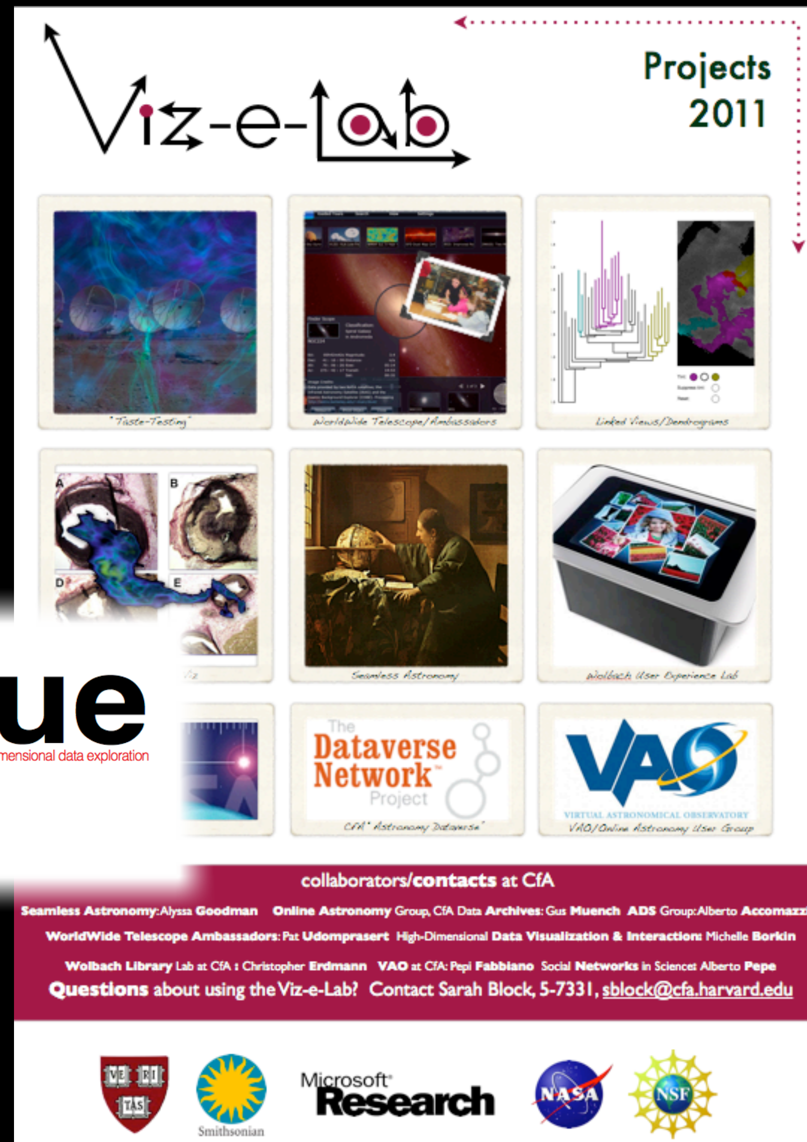




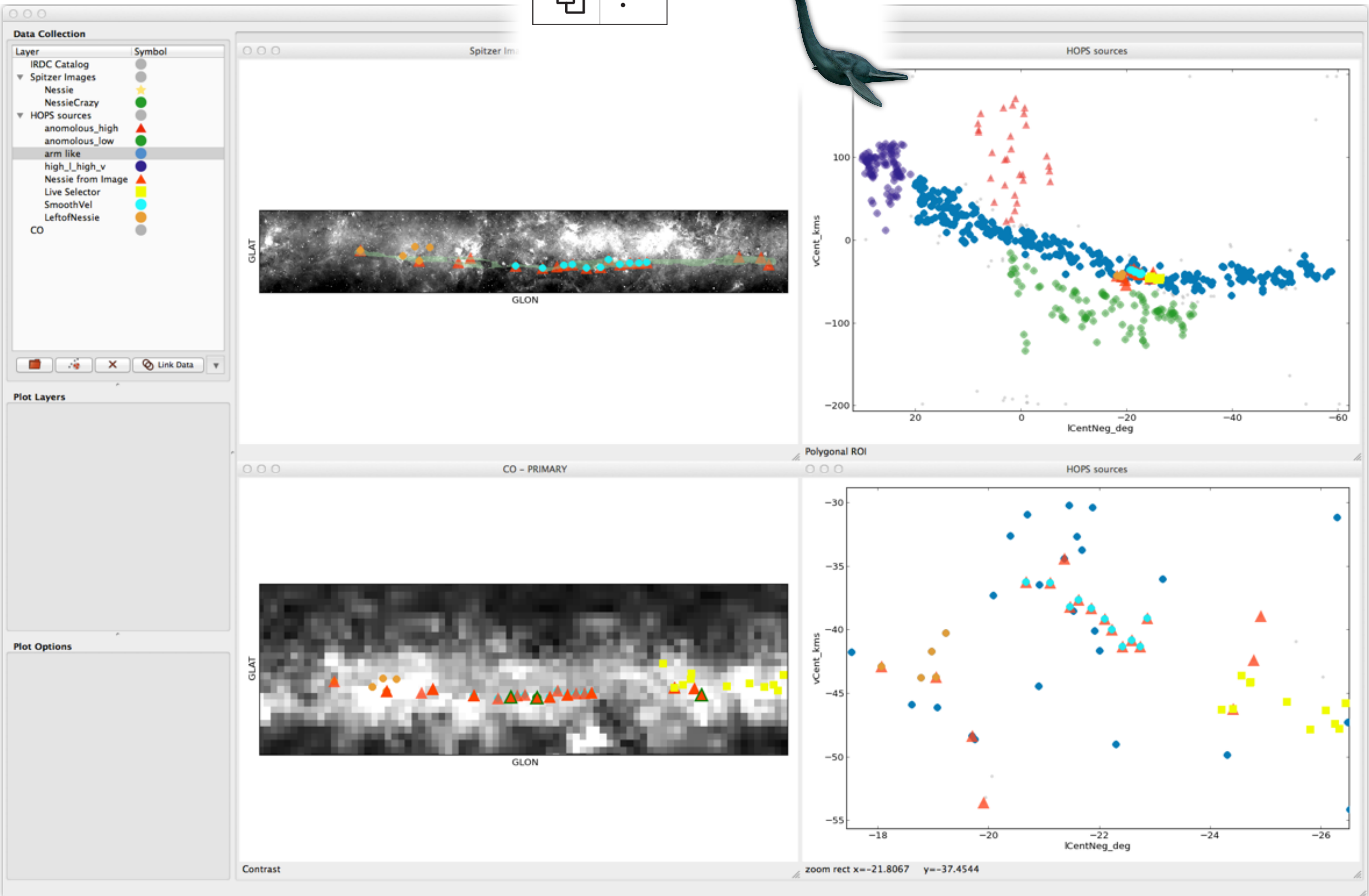
*How to advance the **digital** (visualization) tools for quantitative research?*



glue
multidimensional data exploration

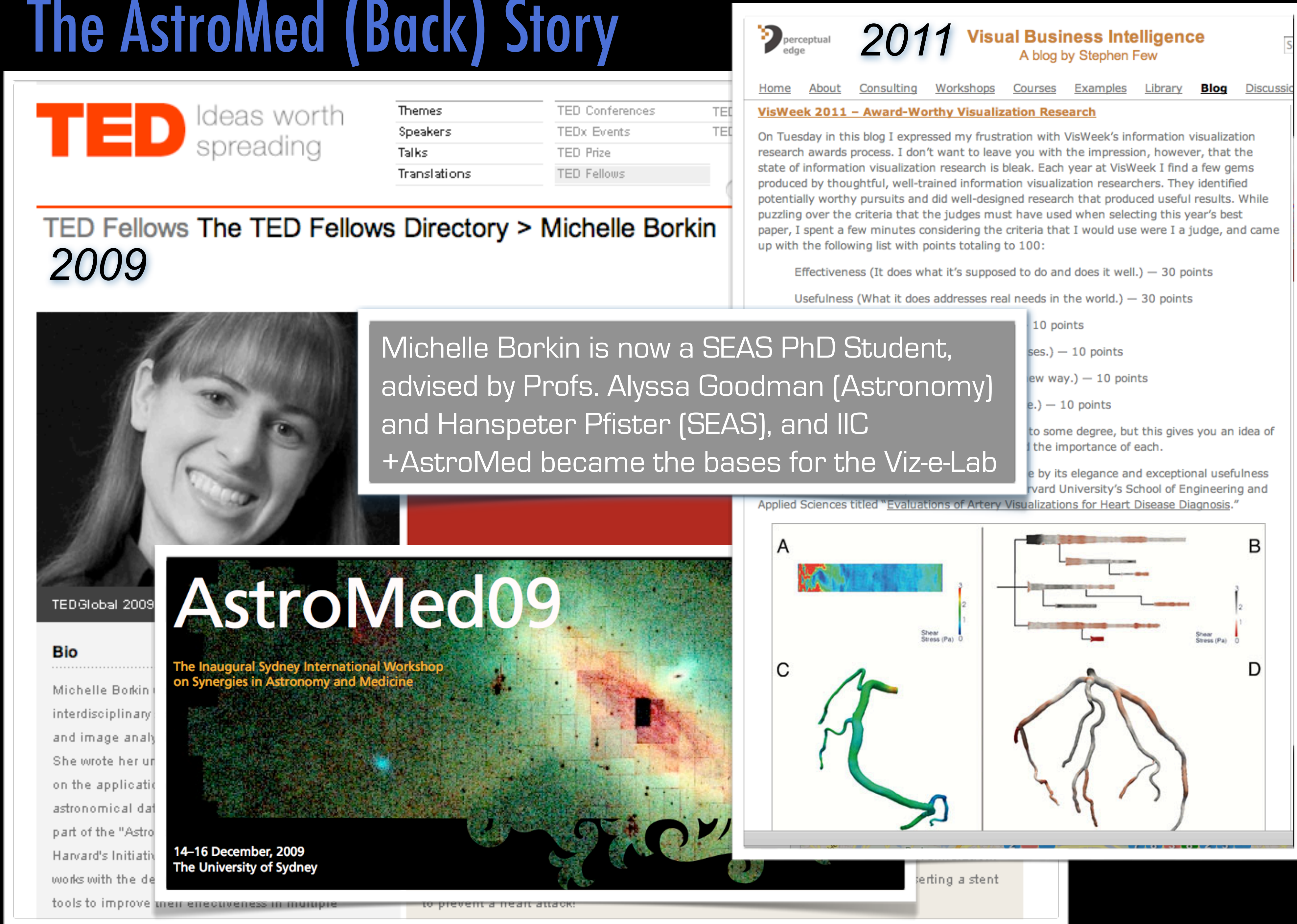


glue
multidimensional data exploration



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

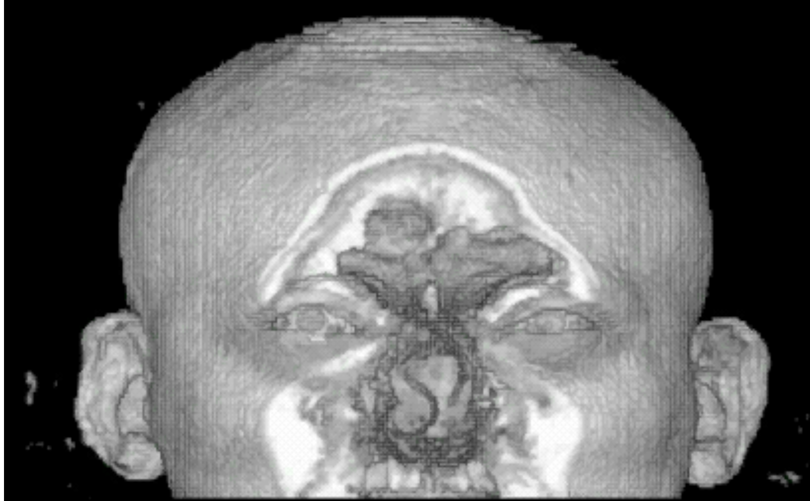
The AstroMed (Back) Story



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

“Astronomical Medicine”

“KEITH”

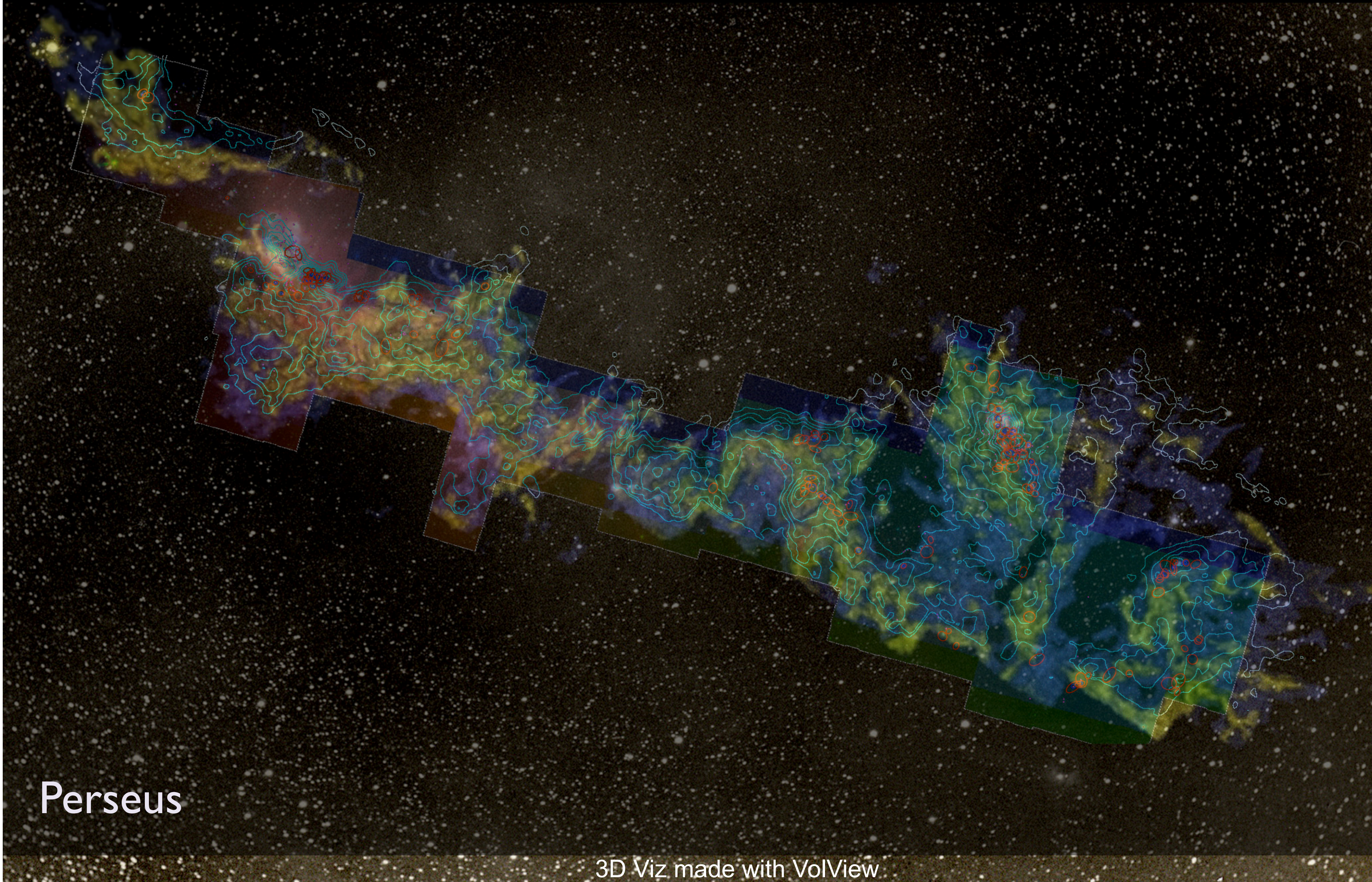


“PERSEUS”

“z” is depth into head

“z” is line-of-sight velocity

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



AstronomicalMedicine@IIC

COMPLETE

How interactive?
How “linked”?

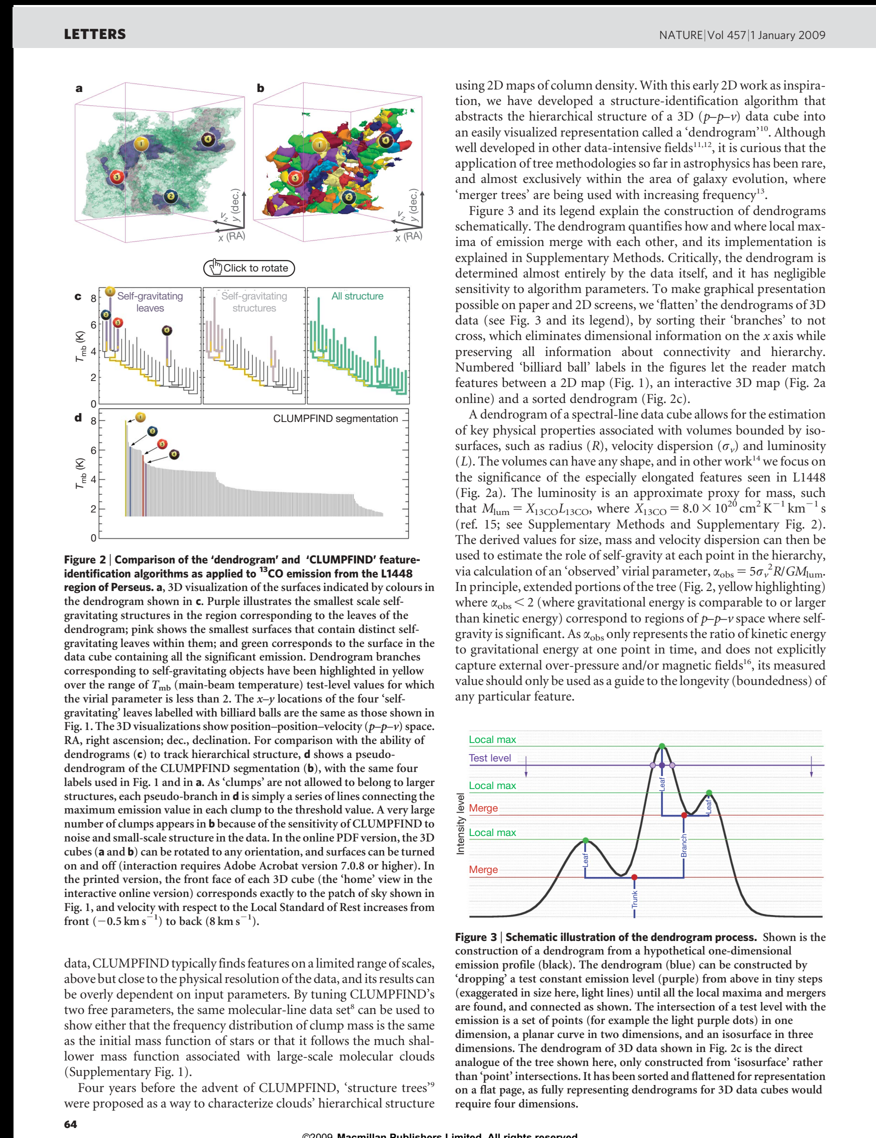


Figure 2 | Comparison of the ‘dendrogram’ and ‘CLUMPFIND’ feature-identification algorithms as applied to ^{13}CO emission from the L1448 region of Perseus. a, 3D visualization of the surfaces indicated by colours in the dendrogram shown in **c**. Purple illustrates the smallest scale self-gravitating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

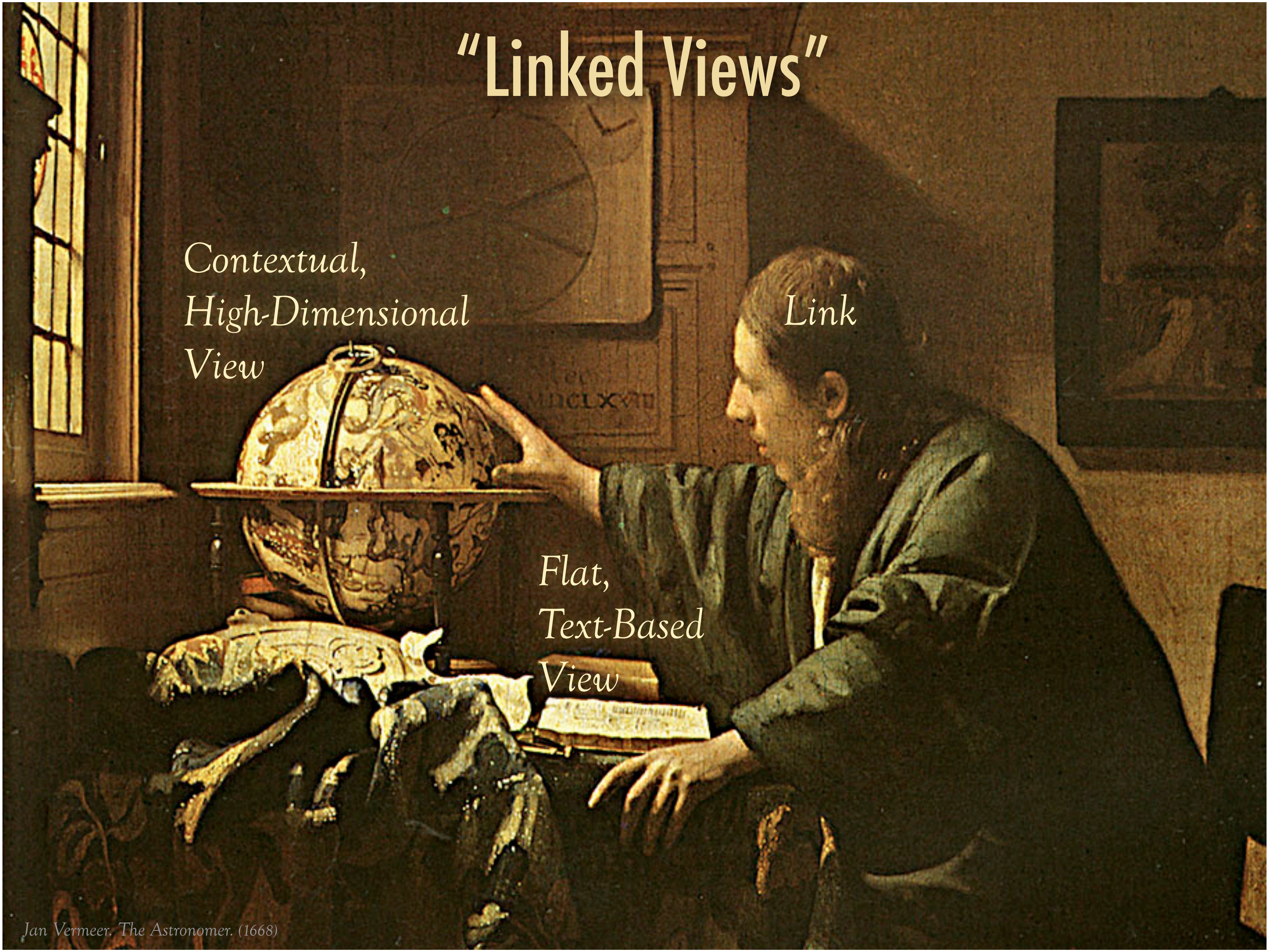
Four years before the advent of CLUMPFIND, ‘structure trees’[9] were proposed as a way to characterize clouds’ hierarchical structure using 2D maps of column density. With this early 2D work as inspiration, we have developed a structure-identification algorithm that abstracts the hierarchical structure 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).

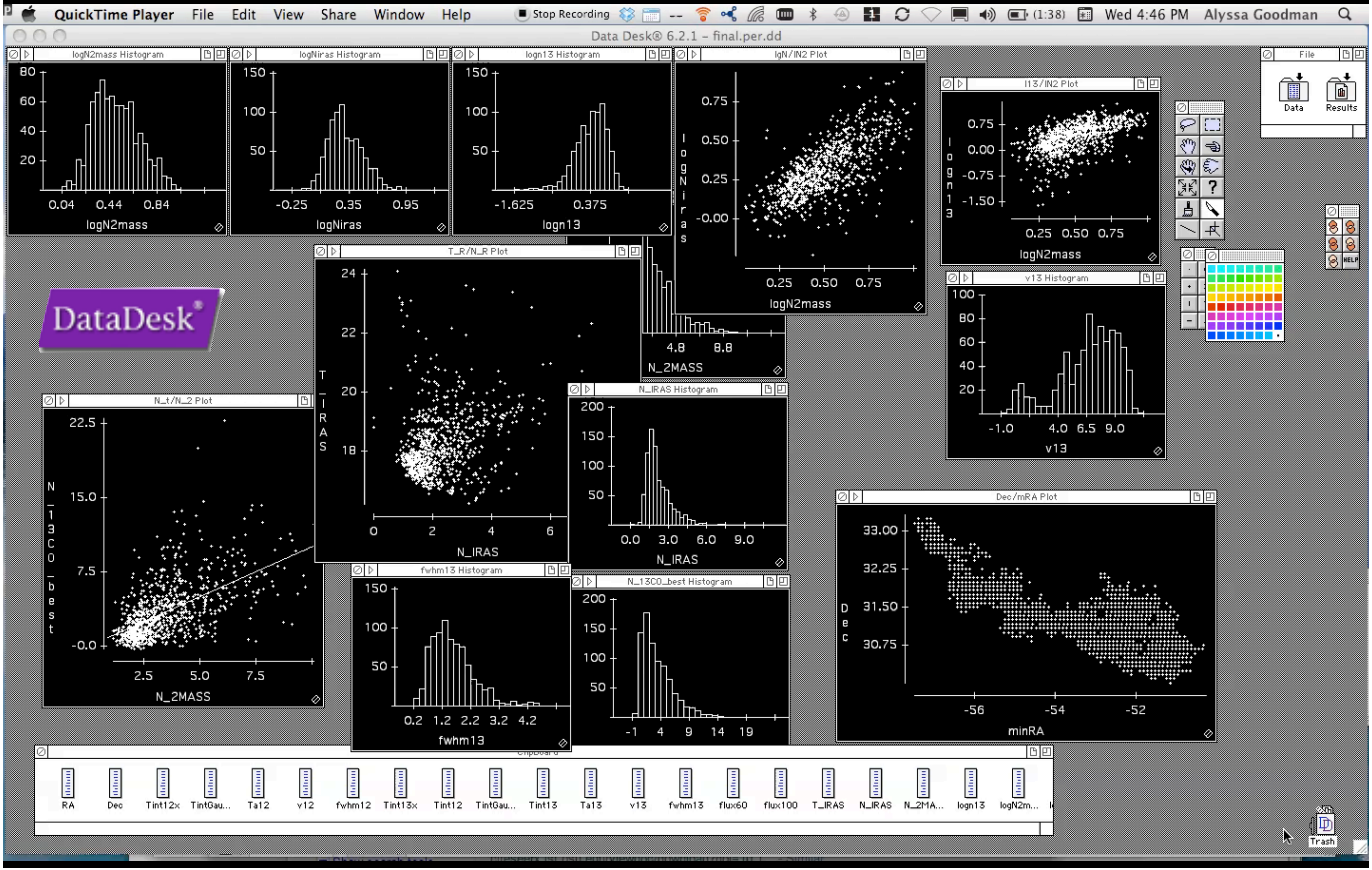
A dendrogram of a spectral-line data cube allows for the estimation of key physical properties associated with volumes bounded by isosurfaces, such as radius (R), velocity dispersion (σ_v) and luminosity (L). The volumes can have any shape, and in other work[14] we focus on the significance of the especially 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*


"Linked Views"
Contextual, High-Dimensional View
Link
Flat, Text-Based View
Jan Vermeer. The Astronomer. (1668)

DataDesk (est. 1986)



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

Picturing | Rotation | Isolation | Masking

Selection

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

Brushing | Linking

Results...

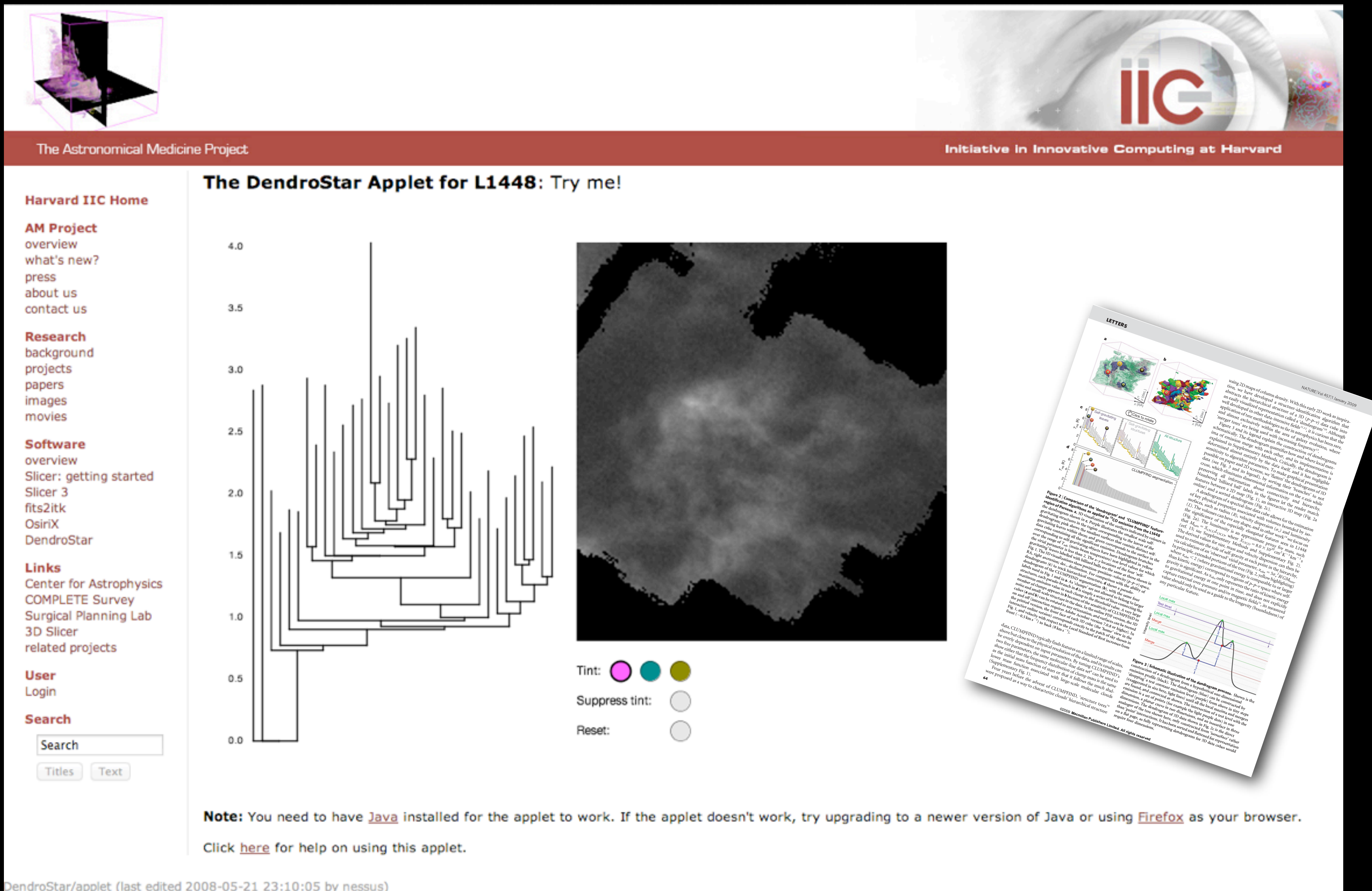
1. for immediate insight
2. as visual source of ideas for statistical algorithms (...relation to SVM)

Warning

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

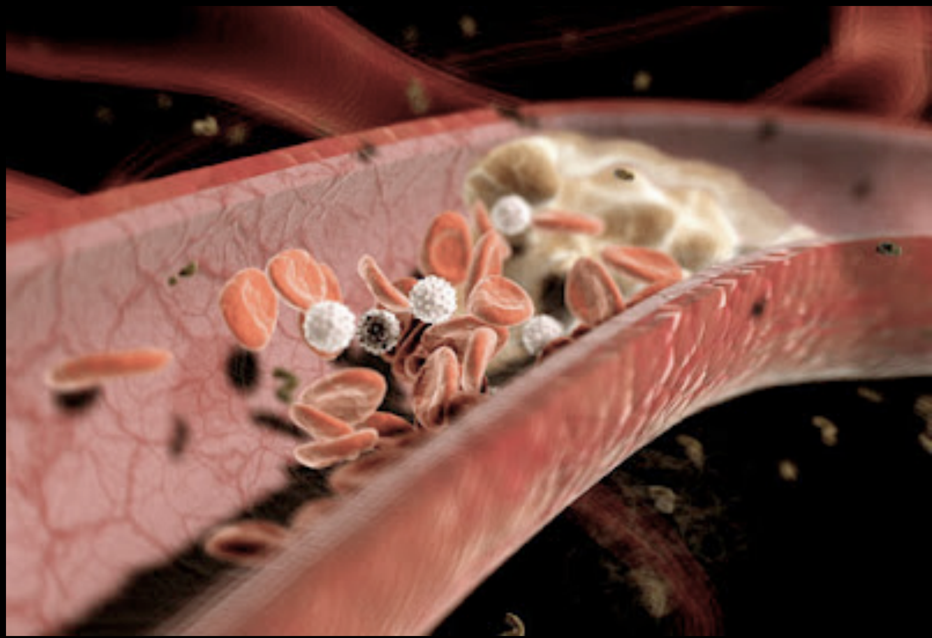
Watch the PRIM-9 video at: http://stat-graphics.org/movies/prim9.html

2008: Dendrostar by Douglas Alan



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

2011: The (Medical) Value of Linked Views...

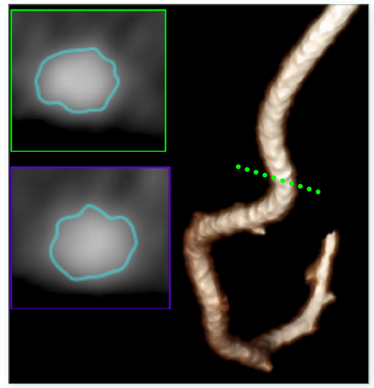


Michelle Borkin

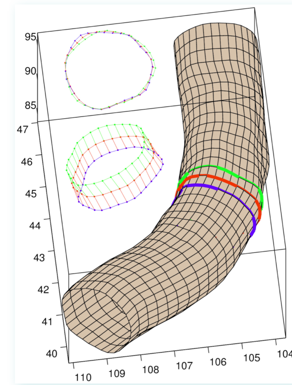
Harvard School of Engineering & Applied Science Ph.D. student,
supervised by Alyssa Goodman (Astronomer) & Hanspeter Pfister (Computer Scientist)

Patients Troubled Hearts, in 3D

Obtain patient CT data

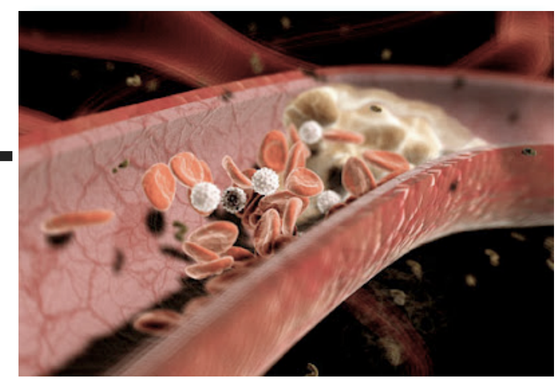


Segment arteries

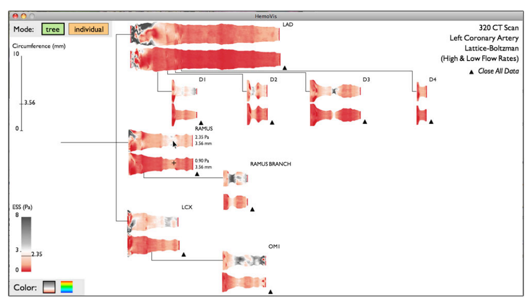


Generate patient geometries

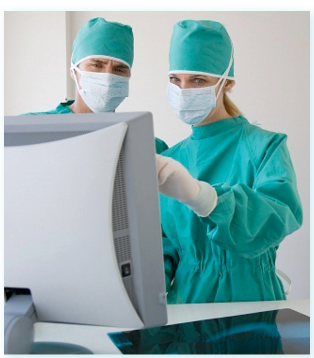
Patient specific flow simulation



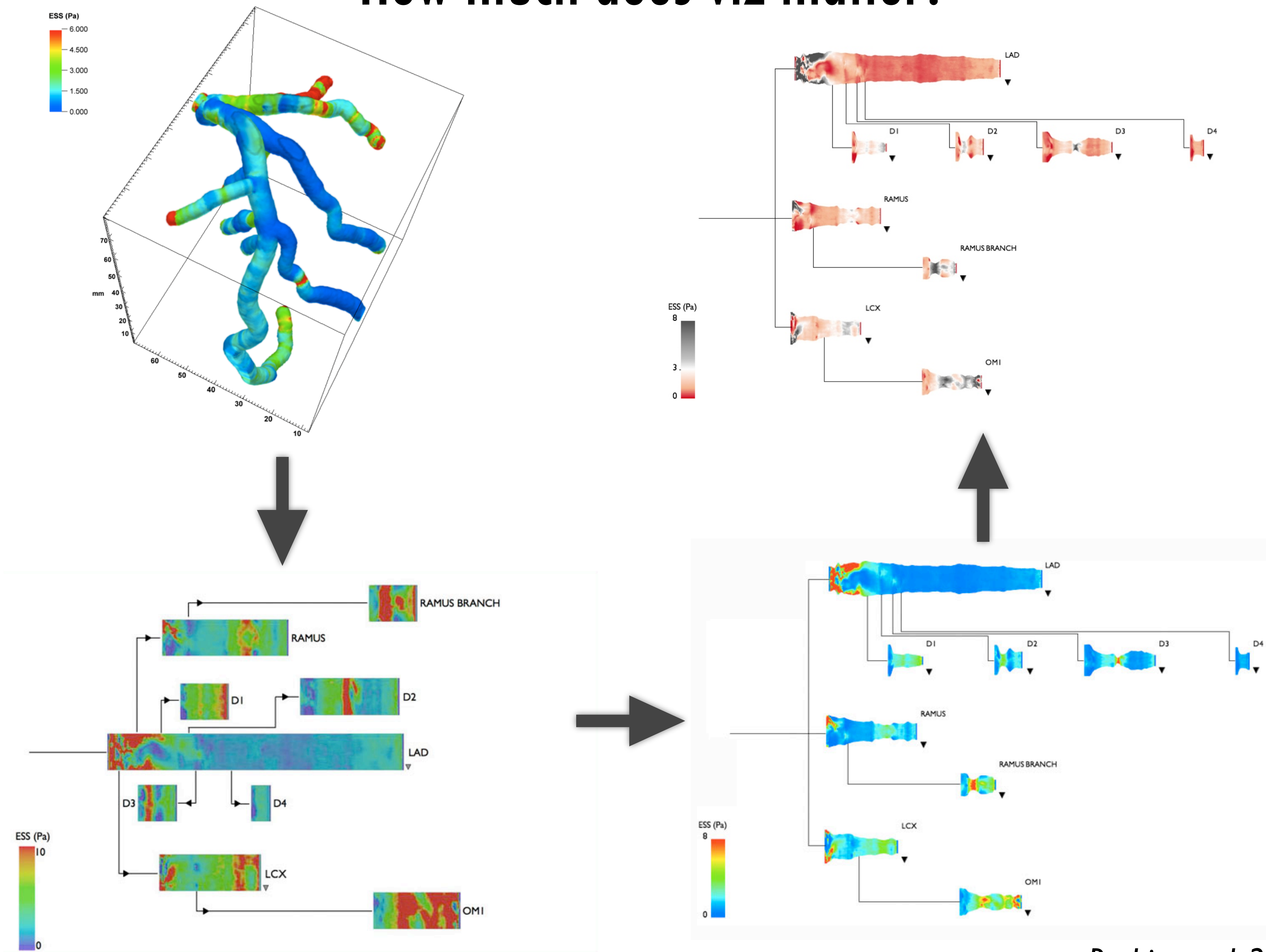
Visualize/analyze data



Clinical decision



How much does viz matter?



Borkin et al. 2011

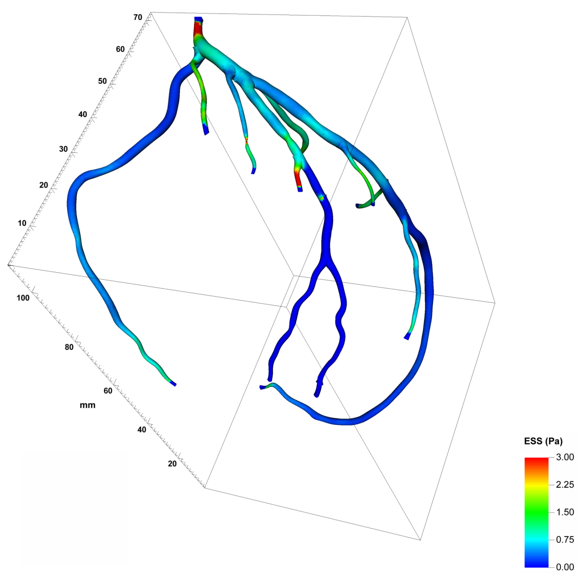
ACCURACY

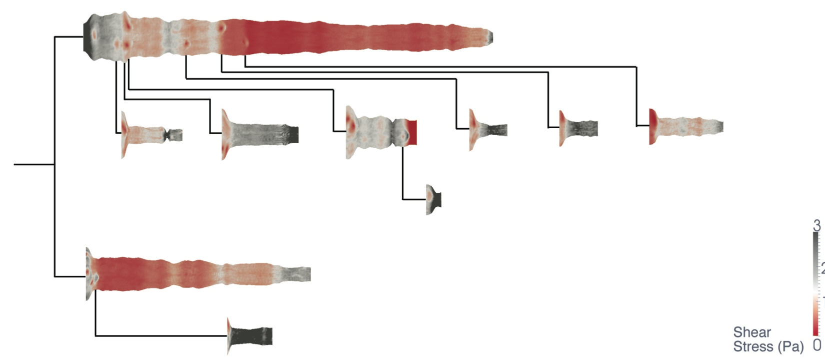
Strong effect of **dimensionality** on accuracy

And strong effect of **color**...

39% *percent low ESS regions found* 62%

91%





Borkin et al. 2011

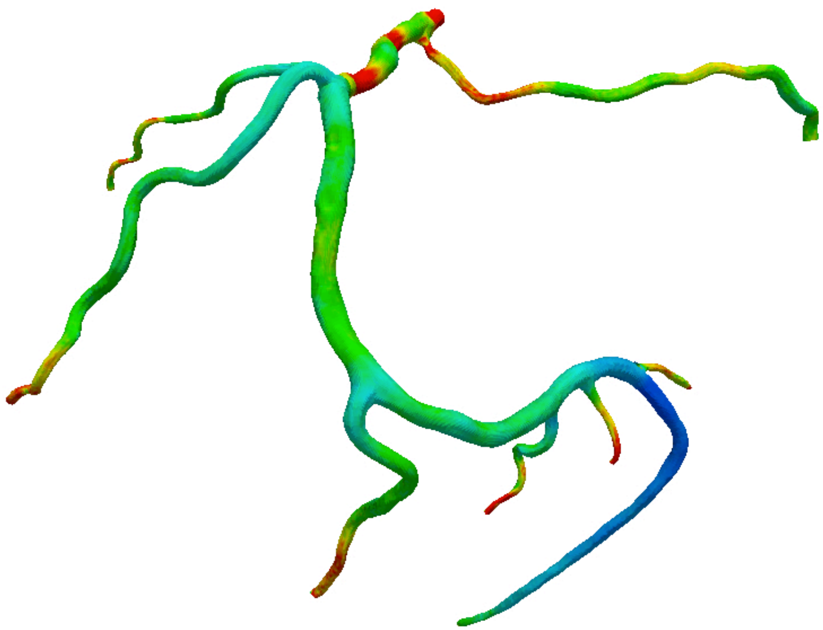
Efficiency

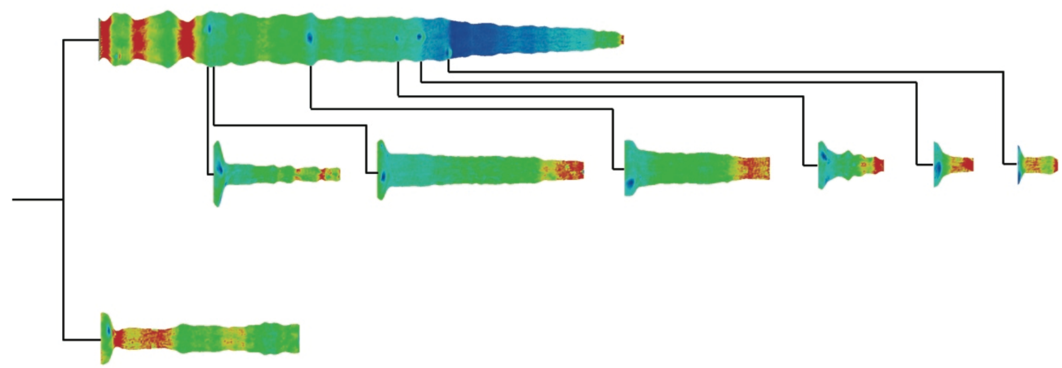
Participants more **efficient** in **2D**.

Rainbow color map has greater detriment in 3D.

10.2 sec/region
5.6 sec/region

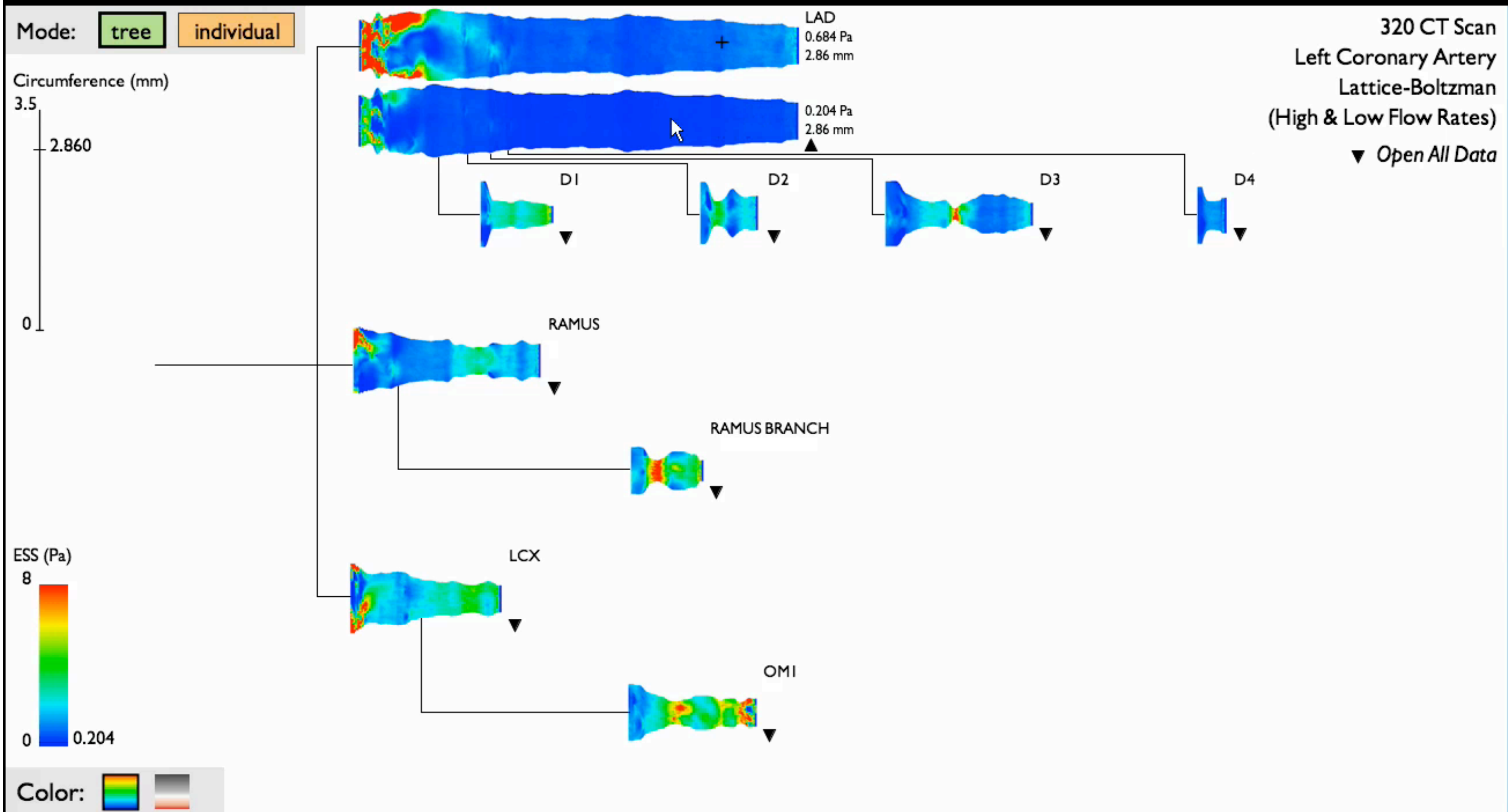
2.6 sec/region
2.4 sec/region





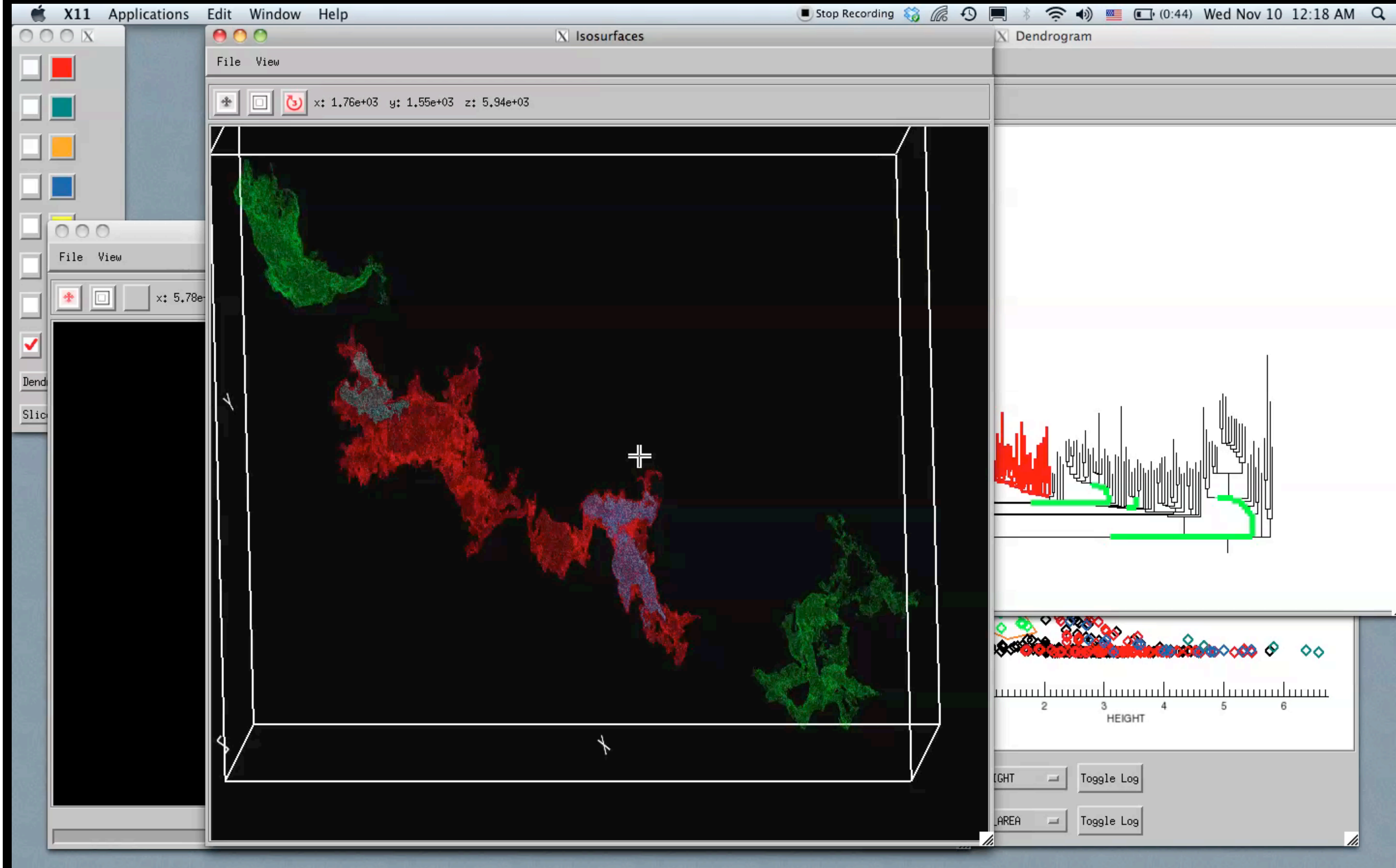
BUT–3D still essential for sugical planning.

Borkin et al. 2011



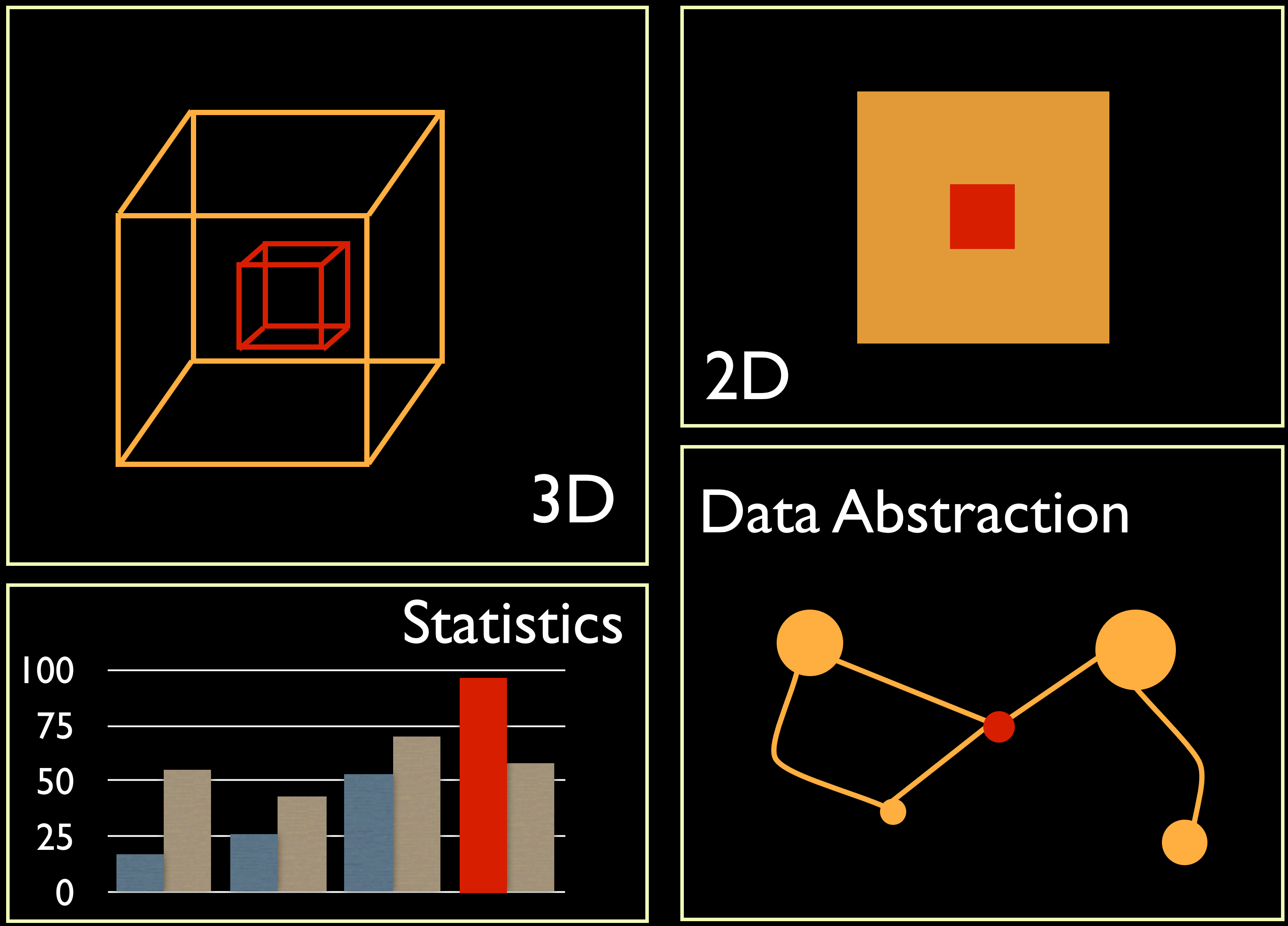
Borkin et al. 2011

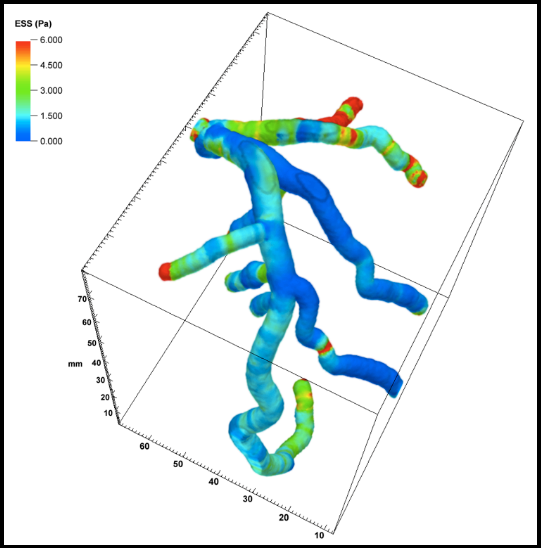
Also in 2011: Linked (Astronomical) Dendrogram Views in IDL



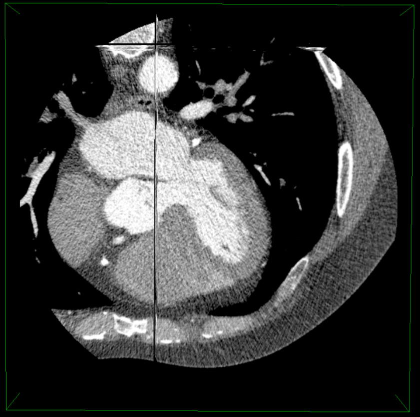
*Video & implementation: Christopher **Beaumont**, CfA/UHawaii;*
inspired by AstroMed work of Douglas Alan, Michelle Borkin, AG, Michael Halle, Erik Rosolowsky

“Linked Views”



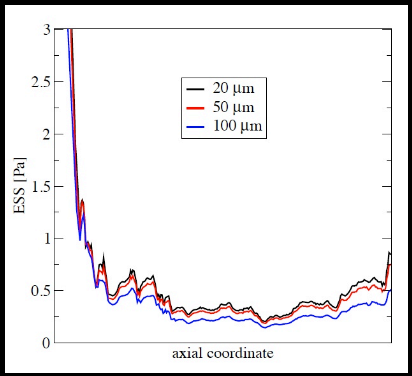


3D

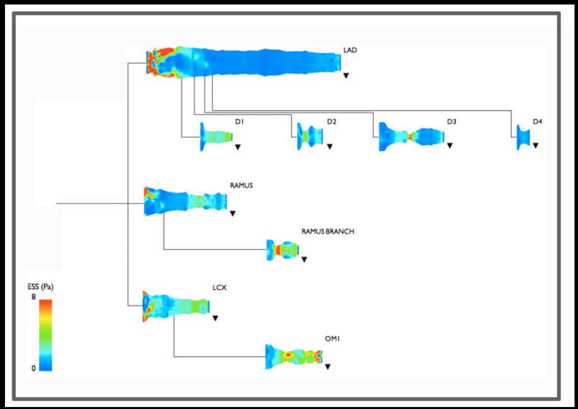


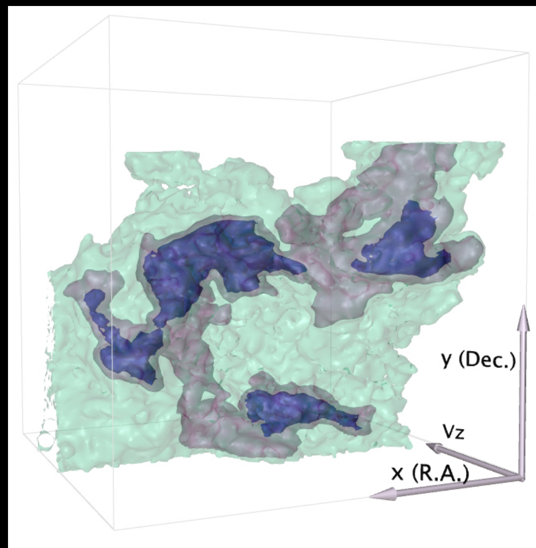
2D

Statistics

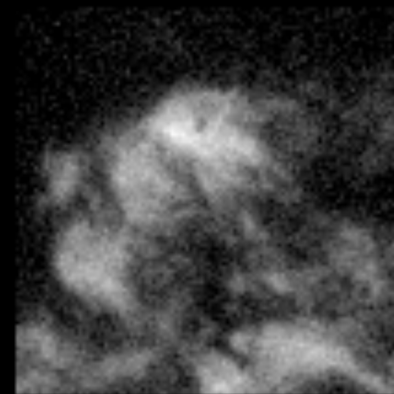


Data Abstraction



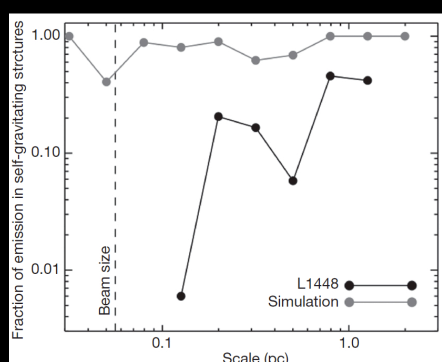


3D

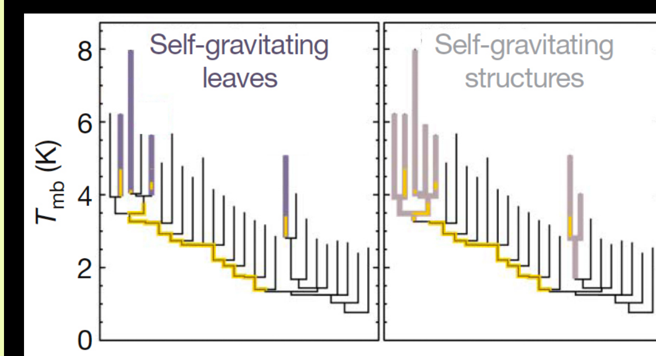


2D

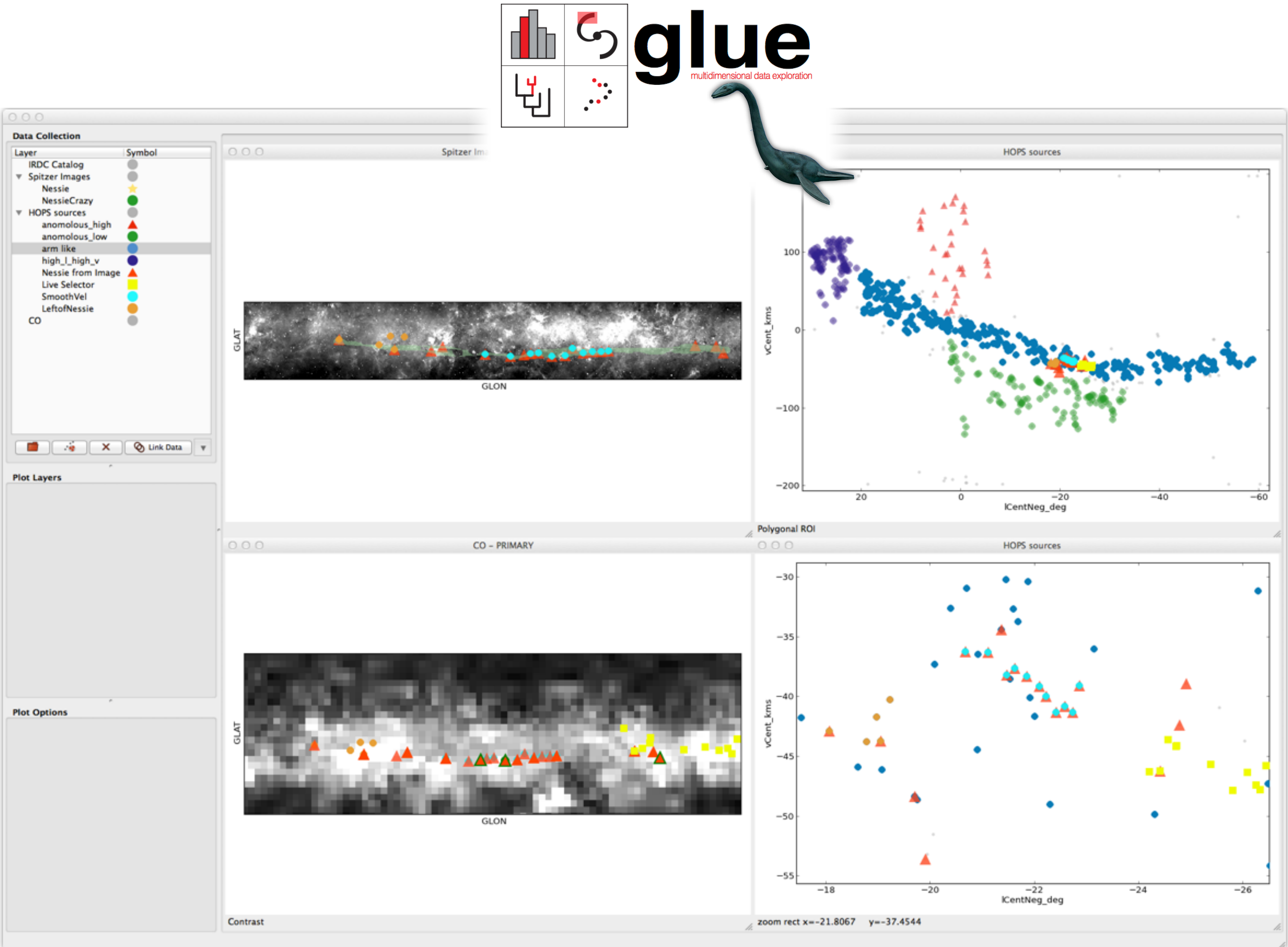
Statistics



Data Abstraction

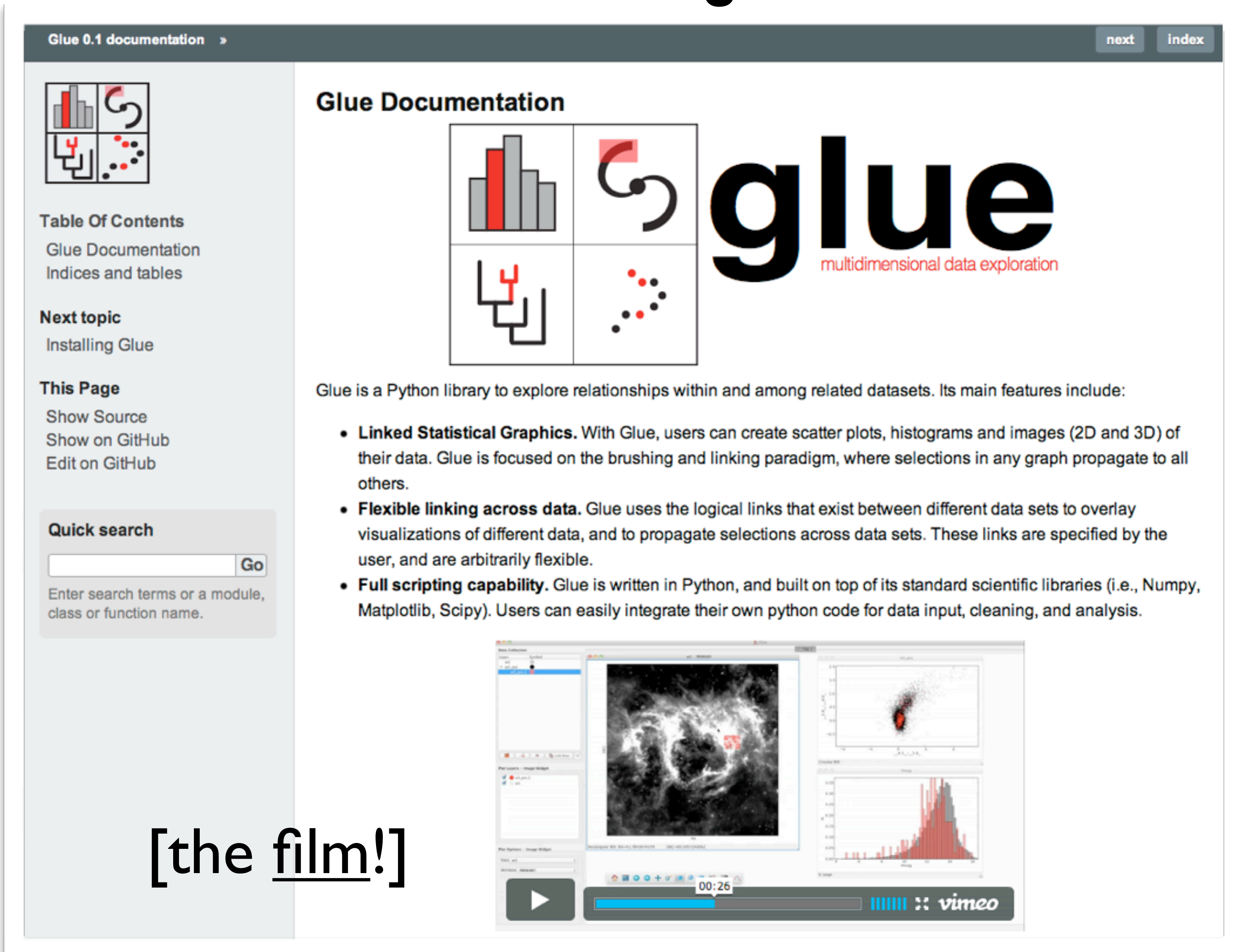


How?



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

What is glue?

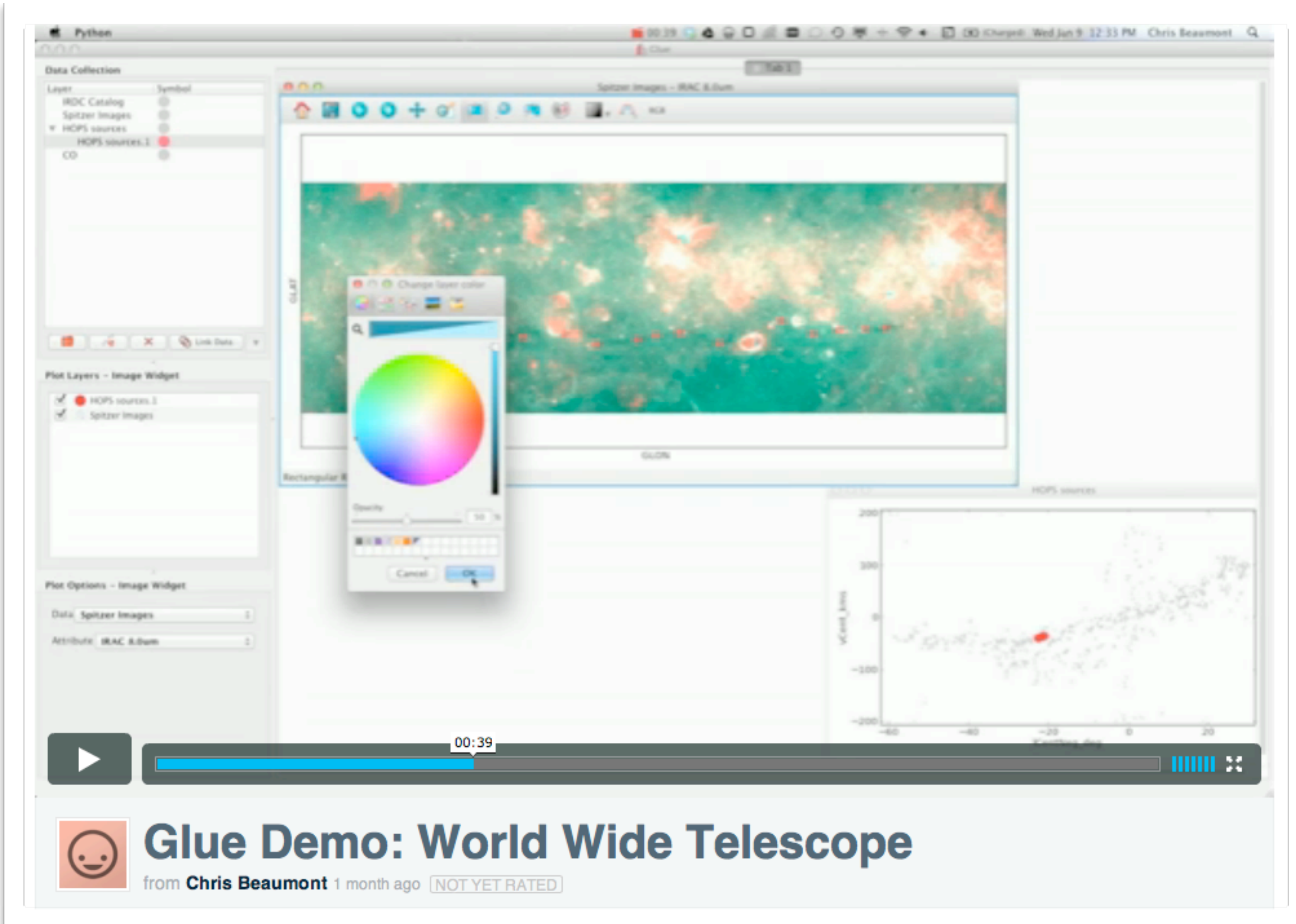


[the film!]

python™

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

Gluing glue to external APIs



http://vimeo.com/57078802

The Bones of the Milky Way: Credits

Seamless Astronomy-style tools used in this project

SEAMLESS ASTRONOMY
Linking scientific data, publications, and communities

- authorea.com (open publishing)
- theastrodata.org (open data)
- glueviz.org (open source tools)
- universe3d.org (collaborative data)
- worldwidetelescope.org (universe information system)
- virtual observatory standards (international online information sharing systems)

The Dataverse Network™ Project

OpenScholar®

Supported by

Microsoft® Research

NSF

NASA

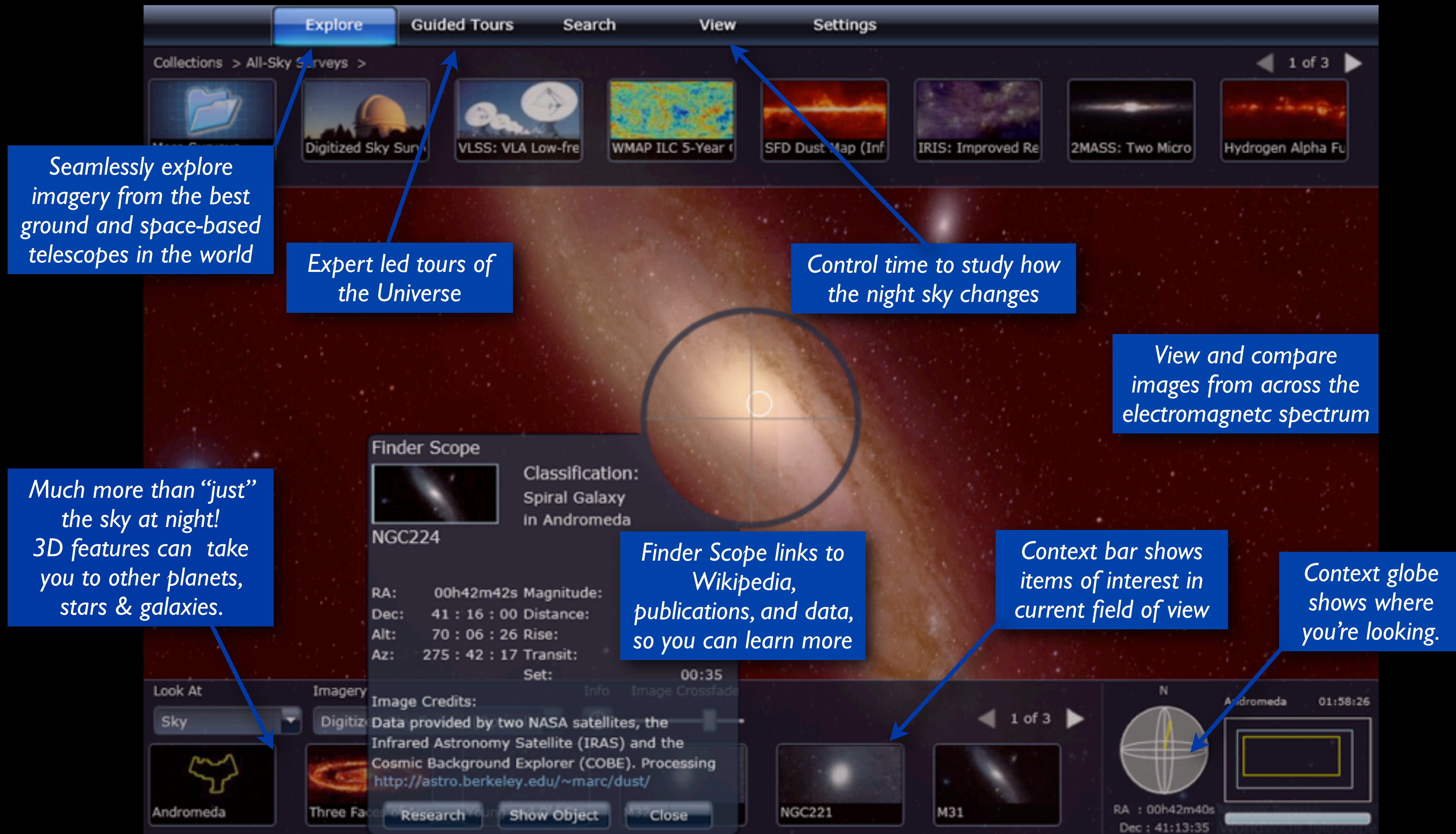
VE RI TAS

Smithsonian

Alyssa Goodman, m:617-230-7080; milkywaybones.org

Microsoft® Research WorldWide Telescope

Experience WWT at **worldwidetelescope.org**



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


JohnTukey

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.