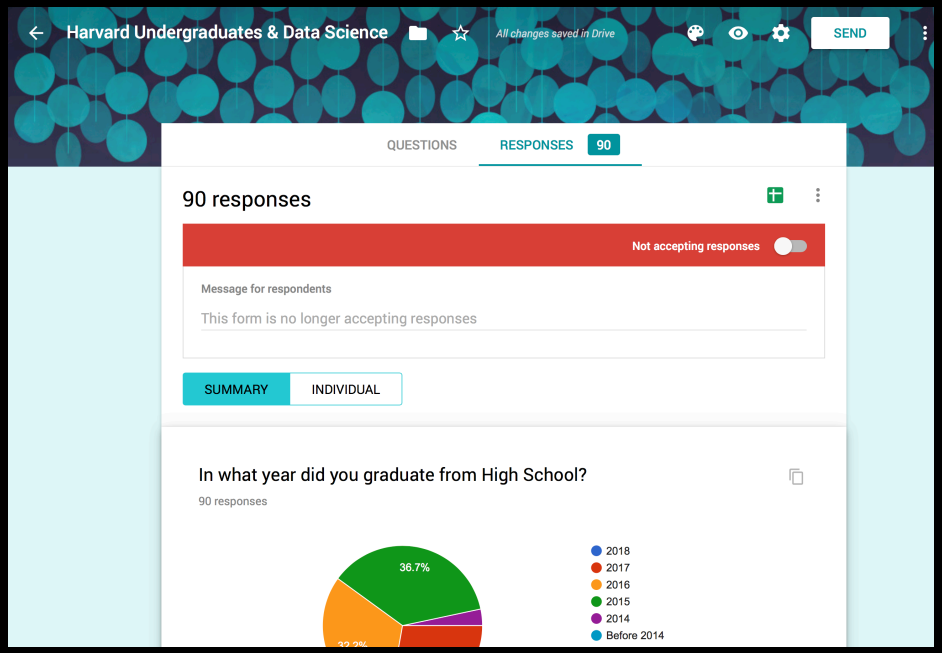
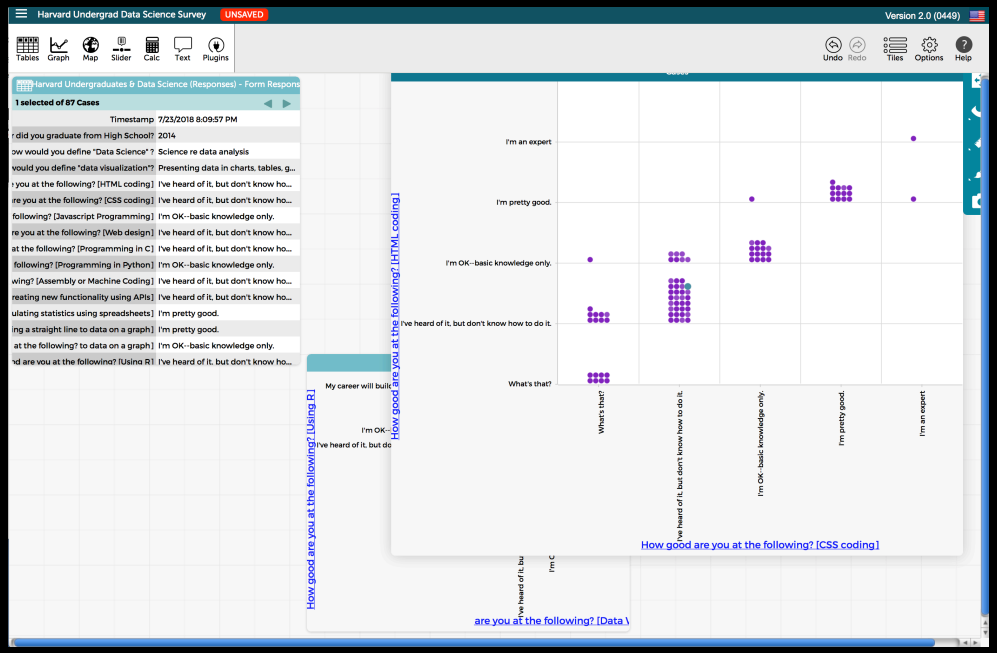
What does "Data Science" mean to me, and to you?

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

Data science & you



https://docs.google.com/forms/d/1kBPtjFrkyKzQL33ivYuEmWKvPFJxh0wbAuHQy2uOlaE/edit#



https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=46183

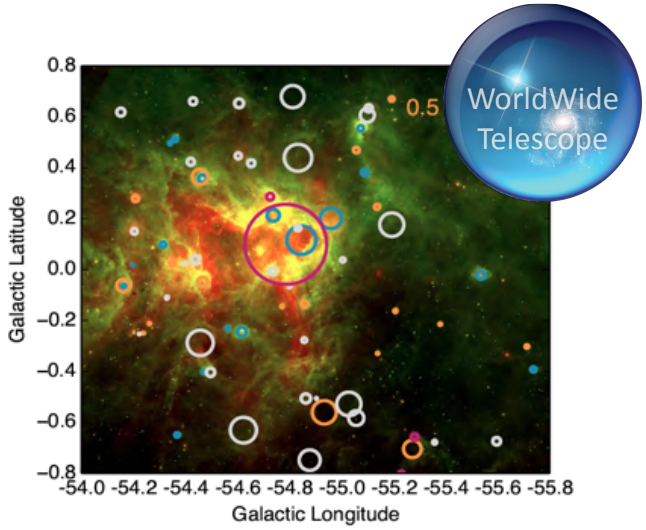
Data science & me

glue
multidimensional data exploration

glueviz.org

PREDICTIONX

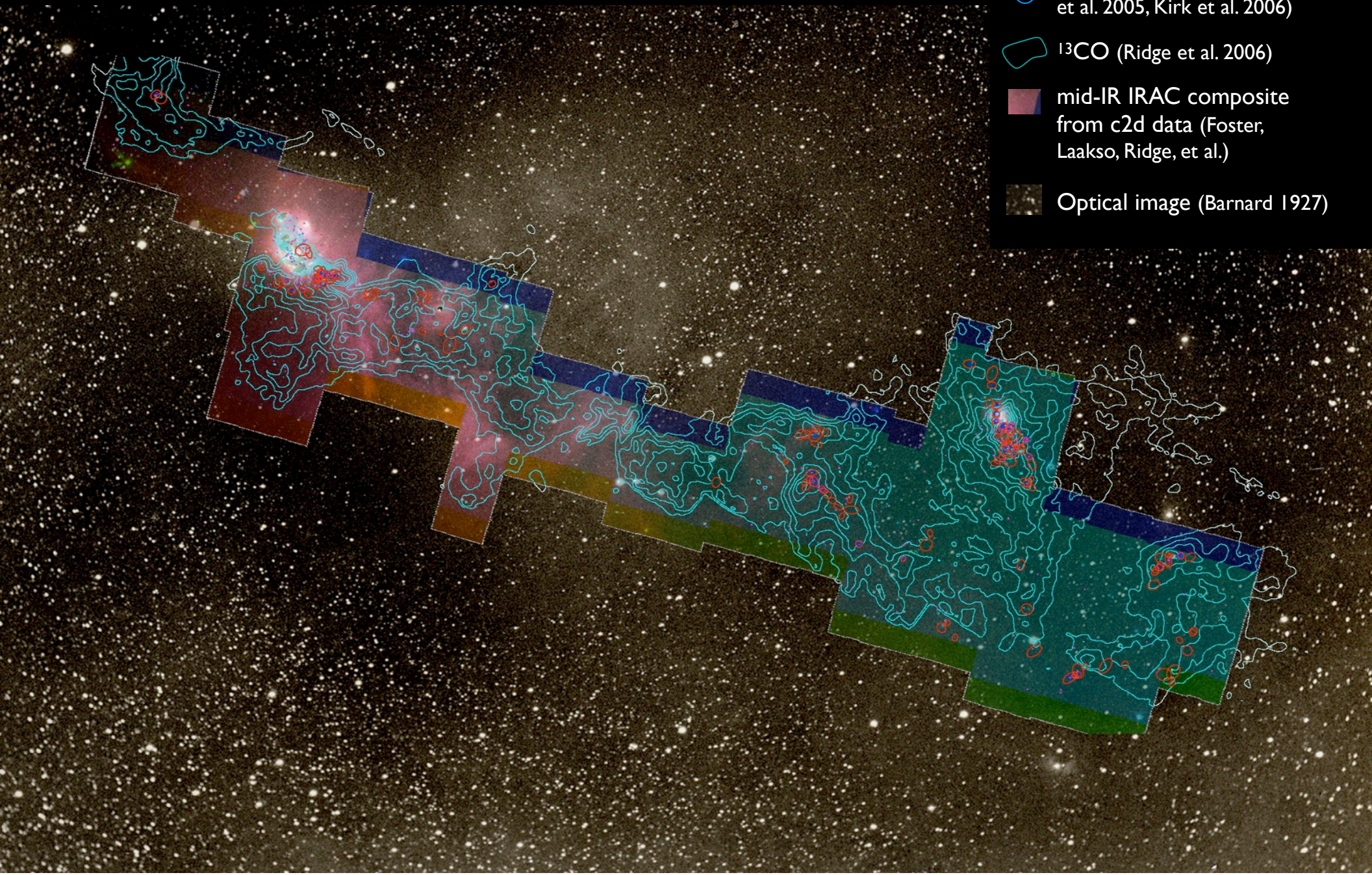
predictionx.org



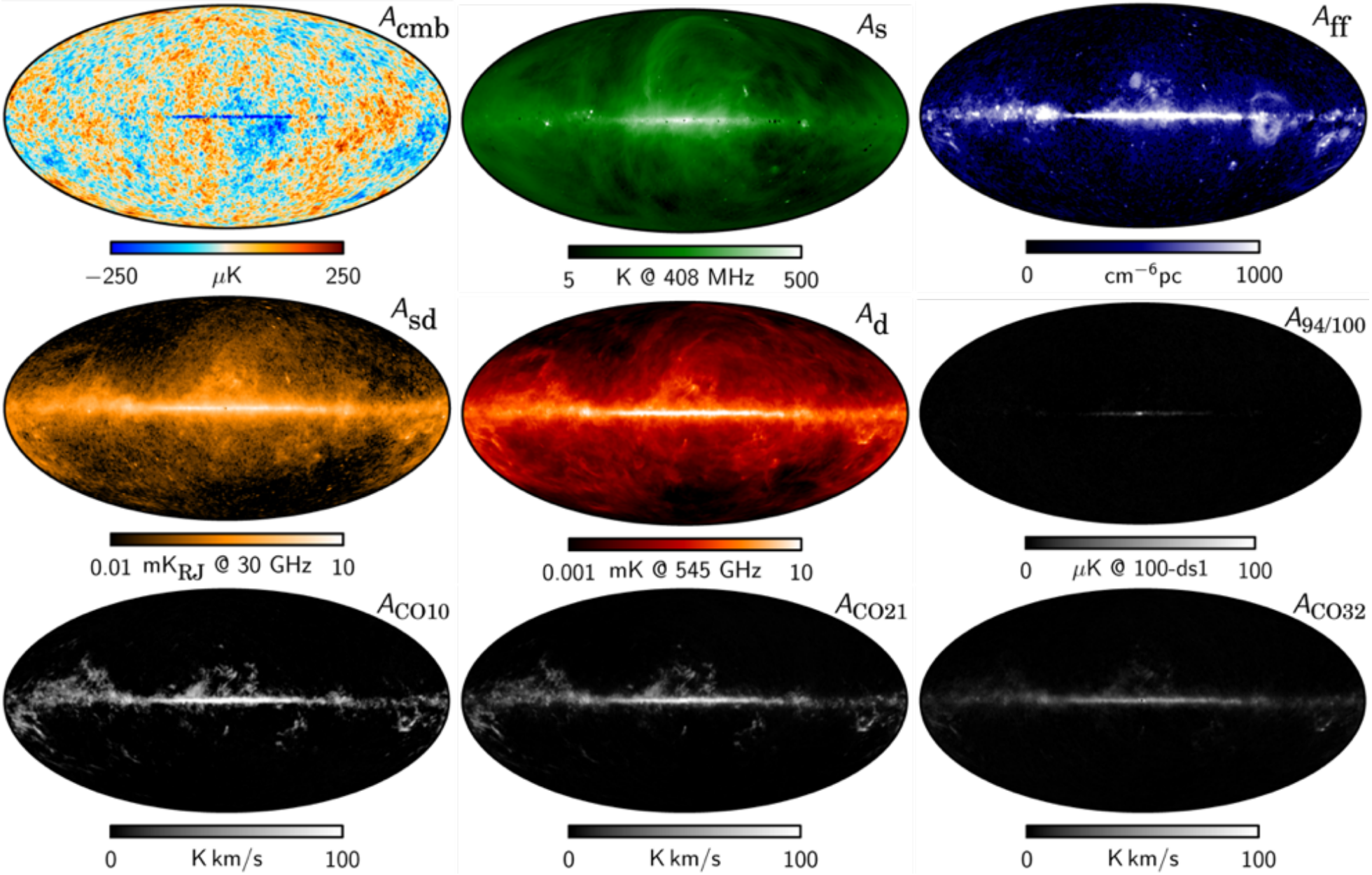
Machine Learning, Data Sharing, Citizen Science

tinyurl.com/paperofthefuture

WIDE DATA


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

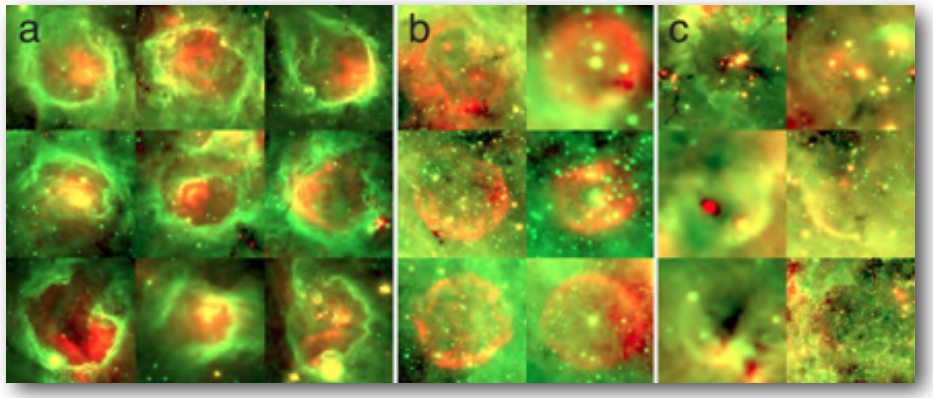
WIDE DATA



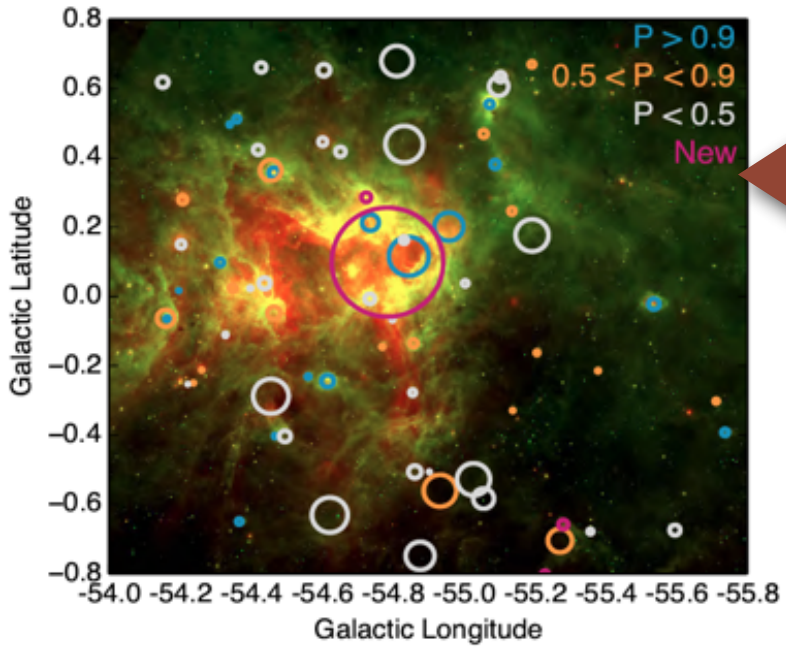
Temperature Foreground amplitudes from Commander, Planck Data [Feb 2015]

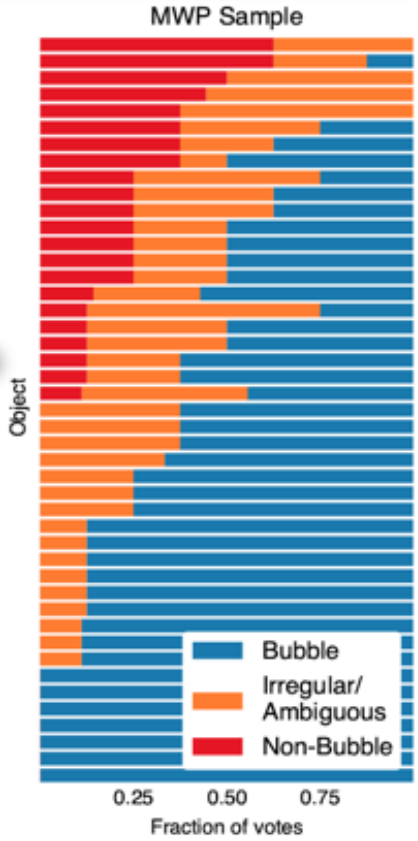
Explore
Guided Tours
Search
View
Settings
Install Windows Client
Use Layer Manager to Control User Settings
Name My Location
Lat 37:47:15
Alt 0 m
Lng -123:35:23
View From This Location
2015/02/11 04:40:33
Real Time
Now
Galactic Plane Mode
WorldWide Telescope
ZOONIVERSE
BIG DATA
Look At
Sky
Imagery
Digitized Sky Survey (Color)
Image Crossfade
Tracking
GLIMPSE/MIPSGAL
1 of 3
N
Scorpius
RA: 17h28m14s
Pismis 24 and
NGC6334
NGC6357
NGC6374
NGC6383
NGC6396
NGC6404
Lesath
Shaula
HR6397
HR6405

BIG DATA AND "HUMAN-AIDED COMPUTING"



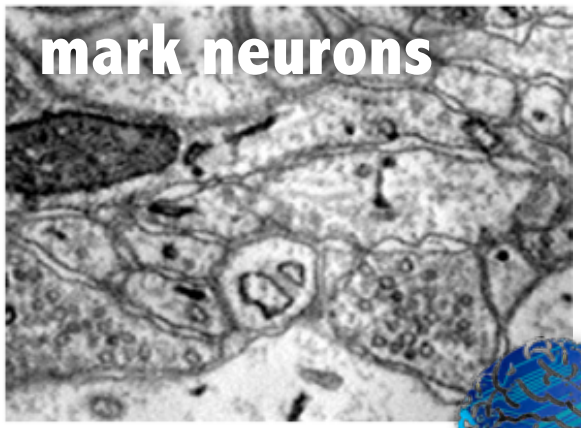
machine-learning algorithm (Brut)

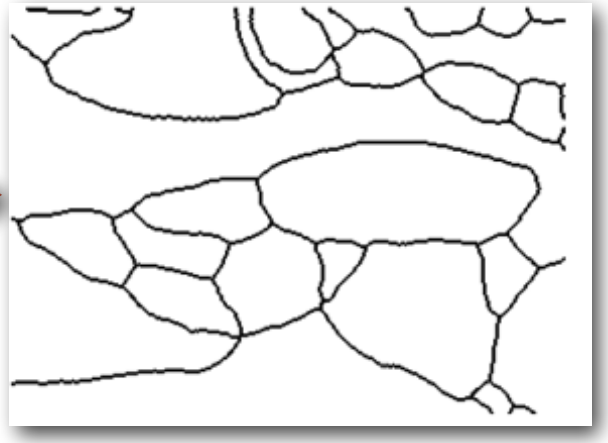


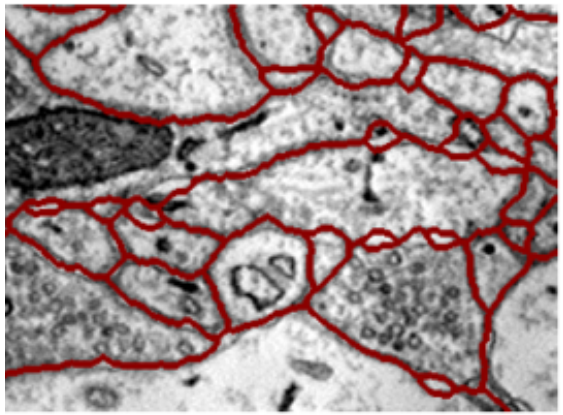


example here from: ***Beaumont****, Goodman, Kendrew, Williams & Simpson 2014; based on* ***Milky Way Project*** *catalog (Simpson et al. 2013), which came from* ***Spitzer/GLIMPSE*** *(Churchwell et al. 2009, Benjamin et al. 2003), cf. Shenoy & Tan 2008 for discussion of HAC;* ***astroml.org*** *for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

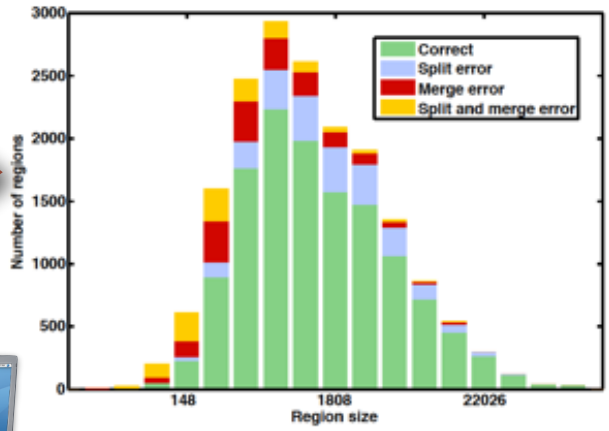
BIG DATA AND "HUMAN-AIDED COMPUTING"



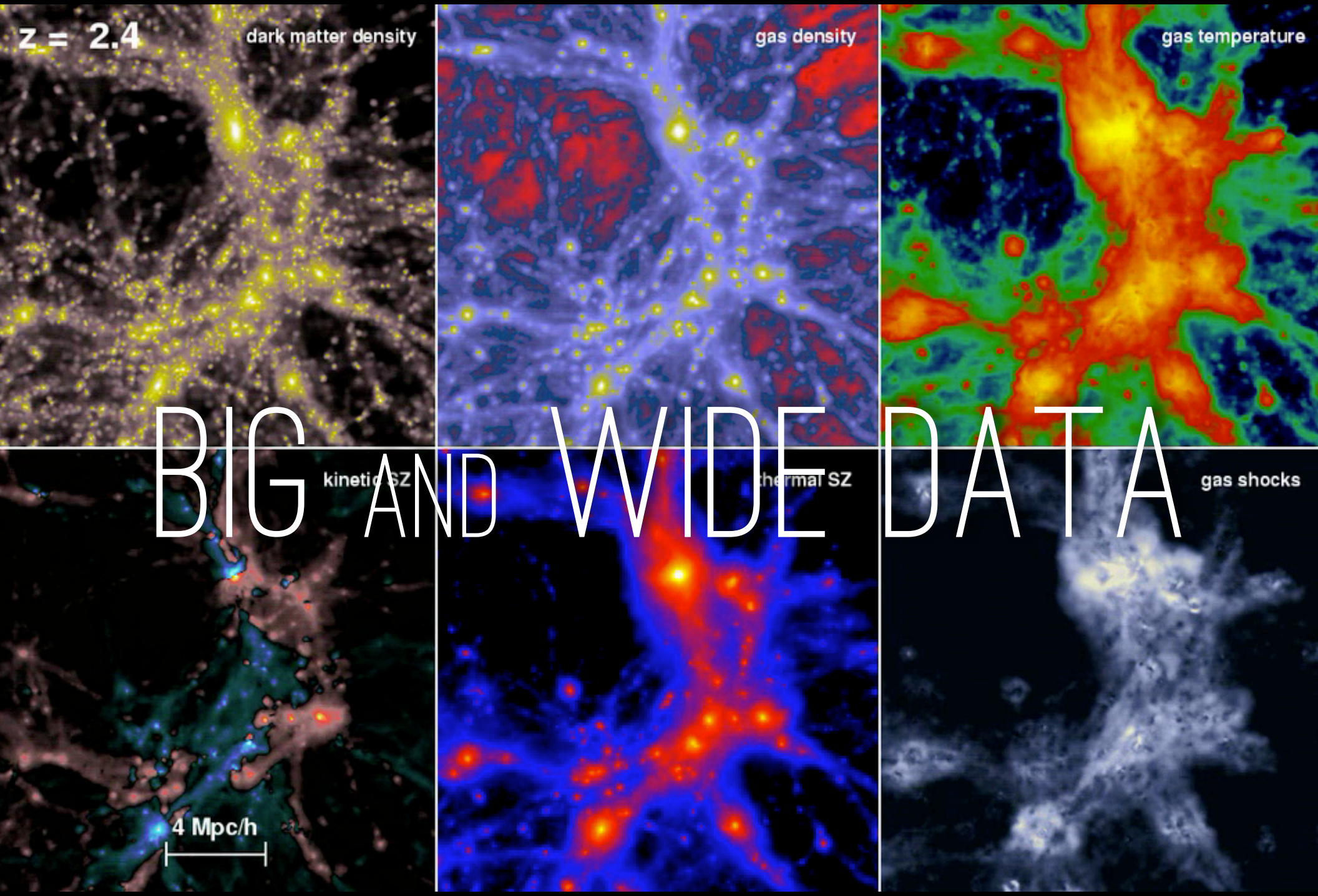




machine-learning algorithm (RF+CRF)

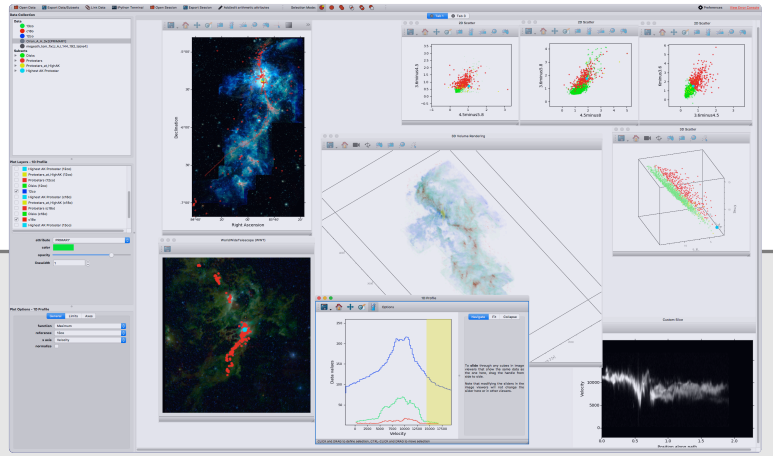


example here from: Kaynig...Lichtman...Pfister et al. 2013, "Large-Scale Automatic Reconstruction of Neuronal Processes from Electron Microscopy Images"; cf. Shenoy & Tan 2008 for discussion of HAC; ***astroml.org*** *for machine learning advice/tools (Note: RF=Random Forest; CRF=Conditional Random Fields.)*



BIG AND WIDE DATA

Movie: Volker Springel, formation of a cluster of galaxies. Millenium Simulation requires 25TB for output.



Preview

New Thinking on, and with, Data Visualization

Alyssa A. Goodman, Michelle A. Borkin, Thomas P. Robitaille

As the complexity and volume of datasets have increased along with the capabilities of modular, open-source, easy-to-implement, visualization tools, scientists' need for, and appreciation of, data visualization has risen too. Until recently, scientists thought of the "explanatory" graphics created at a research project's conclusion as "pretty pictures" needed only for journal publication or public outreach. The plots and displays produced during a research project – often intended only for experts – were thought of as a separate category, what we here call "exploratory" visualization. In this view, discovery comes from exploratory visualization, and explanatory visualization is just for communication. Our aim in this paper is to spark conversation amongst scientists, computer scientists, outreach professionals, educators, and graphics and perception experts about how to foster flexible data visualization practices that can facilitate discovery and communication at the same time. We present an example of a new finding made using the glue visualization environment to demonstrate how the border between explanatory and exploratory visualization is easily traversed. The linked-view principles as well as the actual code in glue are easily adapted to astronomy, medicine, and geographical information science – all fields where combining, visualizing, and analyzing several high-dimensional datasets yields insight. Whether or not scientists can use such a flexible "undisciplined" environment to its fullest potential without special training remains to be seen. We conclude with suggestions for improving the training of scientists in visualization practices, and of computer scientists in the iterative, non-workflow-like, ways in which modern science is carried out.

Comments: Submitted as an invited "Perspectives" Paper for PNAS, in conjunction with the 2018 Sackler Colloquium
License: http://arxiv.org/licenses/nonexclusive-distrib/1.0/

Categories

Primary: Instrumentation and Methods for Astrophysics (astro-ph.IM)

Cross lists: Astrophysics | Instrumentation and Methods for Astrophysics | Add

This article is currently submitted.

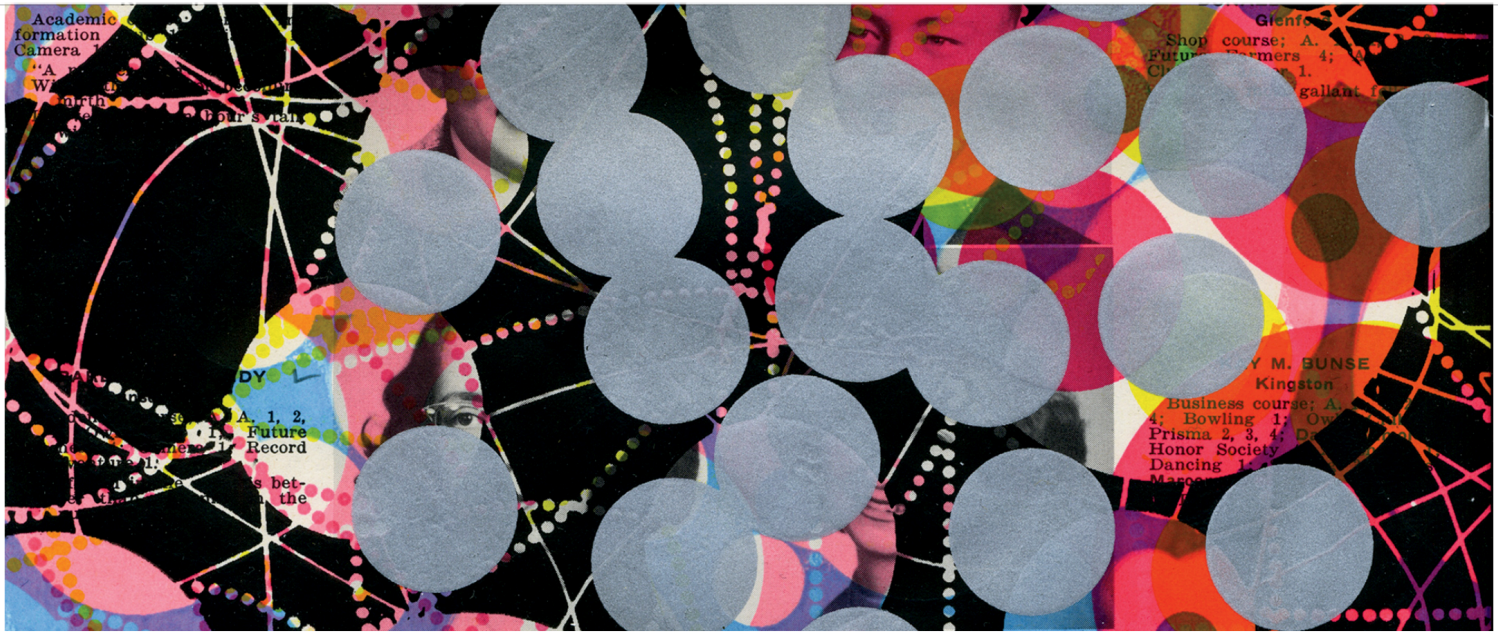
View Article

tinyurl.com/newVizthinking

(Hyperlinked) List of Web Sites visited during the discussion on 24 July 2018

The "Paper" of the Future
Harvard Undergrad Data Science Survey - CODAP
The Prediction Project
The Framework | The Prediction Project
The 10 Questions - 10QViz
WorldWide Telescope Web Client
The Dataverse Project - Dataverse.org
Harvard Undergraduates & Data Science - Google Forms
Glue: multi-dimensional linked-data exploration

Harvard Business Review



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 13 COMMENT HH TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8

WHAT TO READ NEXT

Big Data: The Management Revolution

VIEW MORE FROM THE

October 2012 Issue