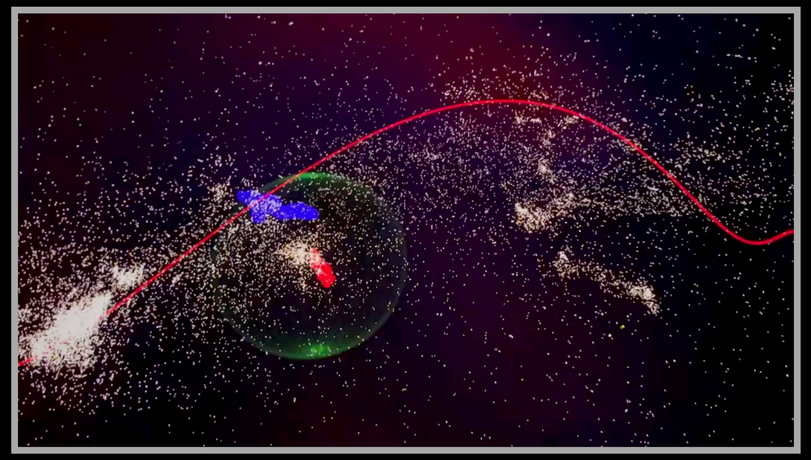
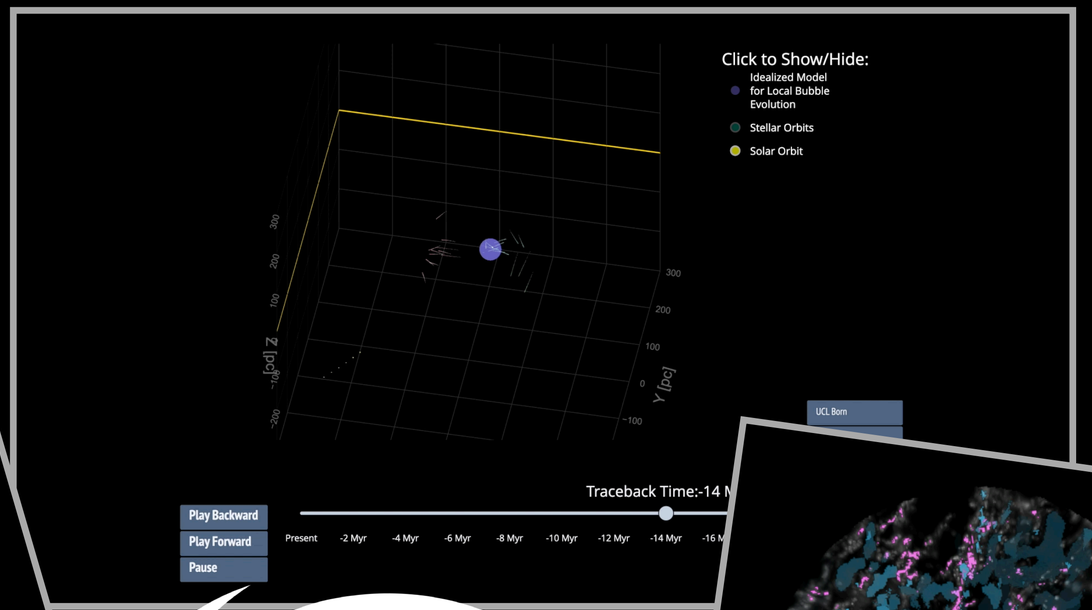
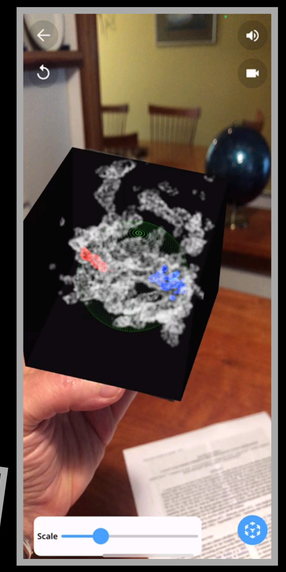
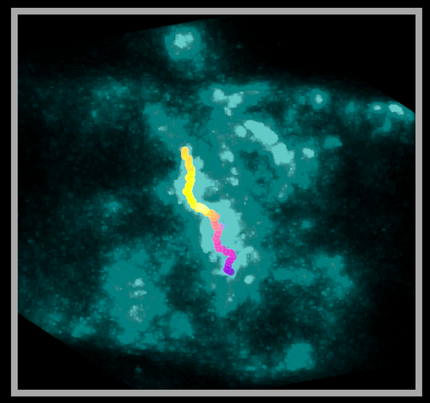
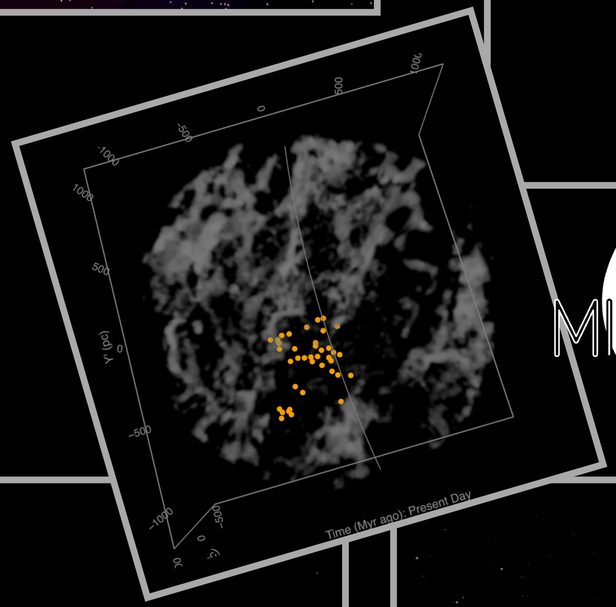
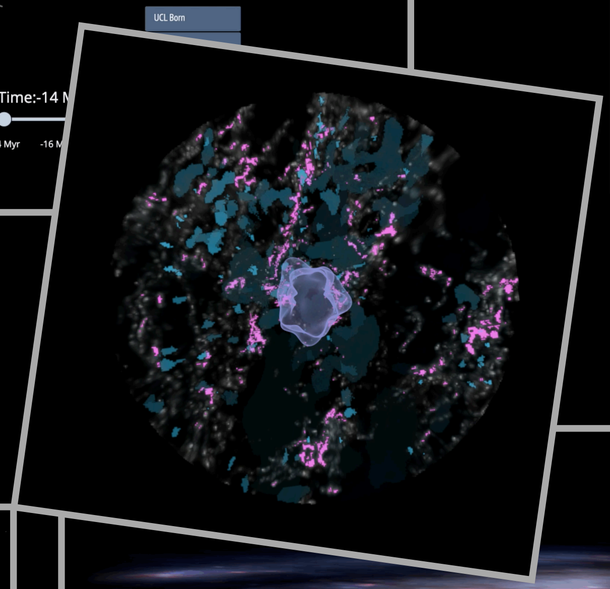




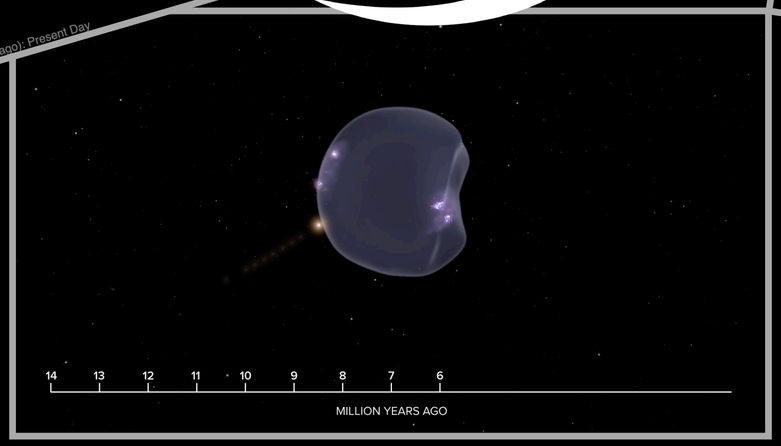


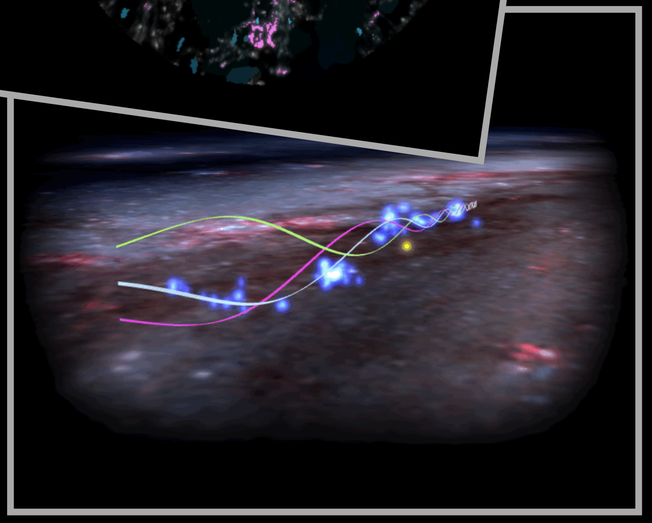




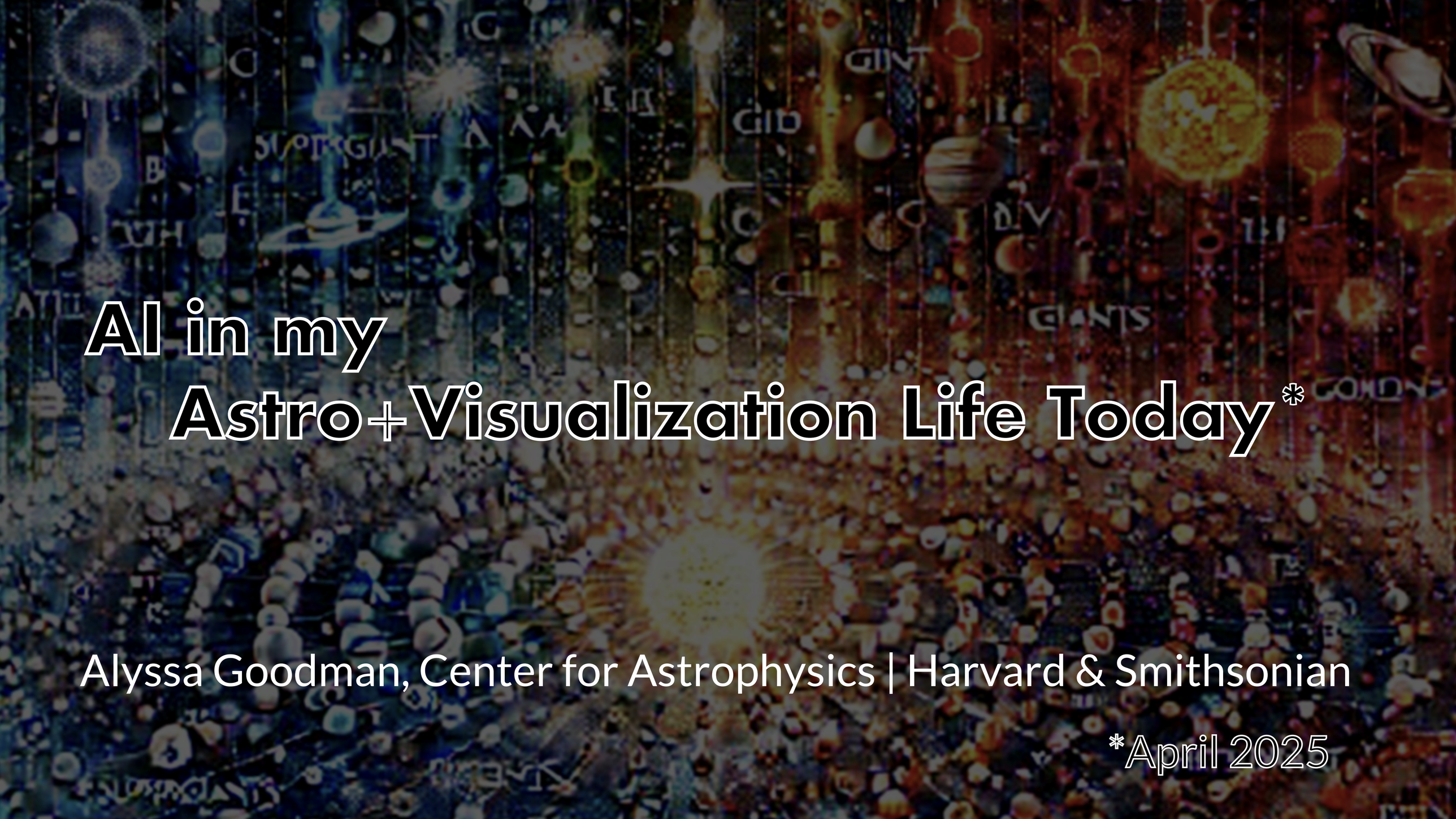


MILKYWAY3D.org





Zucker et al. 2021, Bialy et al. 2021; Zucker et al. 2022, Konietizka et al. 2024, O'Neill et al. 2024, Swiggum et al. 2024



AI in my Astro+Visualization Life Today*

Alyssa Goodman, Center for Astrophysics | Harvard & Smithsonian

*April 2025

AI in my Astro+Visualization Life Today

I have been lucky enough in my career so far to watch, and I hope help, computational technology change what we can learn about our Universe. Today, in 2025, I somewhat unexpectedly find myself involved in a broad range of AI-based and AI-enhanced efforts designed to speed learning and discovery in astrophysics and in science. In this talk, I will offer glimpses into a handful of ongoing AI-enhanced efforts, each of which is very different from the others, yet which work together in a researcher/educator's life to speed progress. Work to be highlighted includes: 1) automated data-set linking in the "glue" and LIVE-Environments visualization environments; 2) The "Reading Time Machine" which uses AI to "read" graphics and images and ingest their content into the ADS Literature archive, as "data," 3) approaches to 3D selection in volumetric data, using both AI and augmented reality (AR); 4) a quest to understand why LLMs are so good at describing infographics, but so terrible at creating them; 5) capabilities of AI for writing code for visualization, in both research and education. The plan of the talk will be to present an overview of each of these efforts, in order to inspire broader discussion of whichever topics evolve as most interesting to the assembled audience.

Thematic questions for discussion

1. visualization and infographics — how good is AI at abstraction, and how good will it become?
2. If we give “all” the MilkyWay3D.org data to AI, what (kinds of) new physical insights might we expect—should we have expectations *a priori*—or is that approach too restrictive? (Note, 3D dust also AI!)

My Astro+Visualization Life *a Decade Ago*: The "Paper" of the *Future* (2015)

connecting language, data, and pictures, for communication & cognition

Fig. 1

*The Paper of the Future should include seamless linkages amongst **data**, **pictures**, and **language**, where "language" includes both words and math. When an individual attempts to understand each of these kinds of information, different cognitive functions are utilized: communication is inefficient if the channel is restricted primarily to language, without easy interconnection to data and pictures.*

Goodman, Peek, et al. 2015

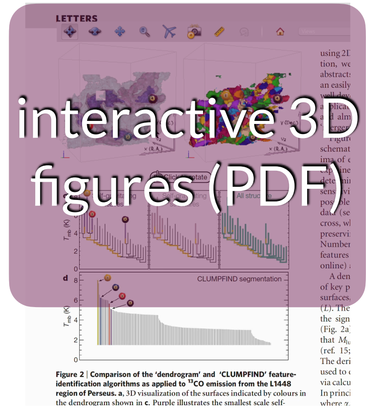
open data, data published with papers

Homepage of the Astronomy [illegible] automatically extracts metadata from, a wide variety of file formats, including the Astronomical standard known as "FITS".

executable code (notebooks) IN papers

Fig. 4

inch (and run) the code that created this image in your brower. The Dataverse e data file used to make this plot is linked through hdl:10904/10081.



interactive 3D figures (PDF)

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-

contextual, live, views of data

Fig. 7

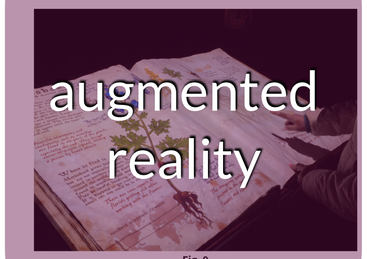
Click here to see this image on the Sky in your browser (using HTML5 WorldWide Telescope). Original image source.

exploratory data analysis INSIDE papers

After Galileo discovered the first four moons of Jupiter, it took nearly three hundred years to discover the next one.

	ShareLatex	Authorea	Overleaf	Google Docs	Microsoft Word	
Version Control	full (3 users)	full	full	full	full	
Simultaneous Editing	full	by section	full	full	Sharepoint	
Offline Editing	Dropbox	browser/git	browser/git	Google Drive	native	
Ease of Use	***	***	***	****	****	
Math Support	Latex	Latex / others	Latex	Limited GUI	GUI	
Output Formats	PDF	PDF	HTML	PDF	PDF	Word PDF
Commenting	Latex	By section or text	By section	Track Changes	Track changes	
Citations	bibtex	integrated	bibtex	in add-ons	integrated	
Interactivity	none	full javascript	none	in add-ons	None?	

online collaborative authoring

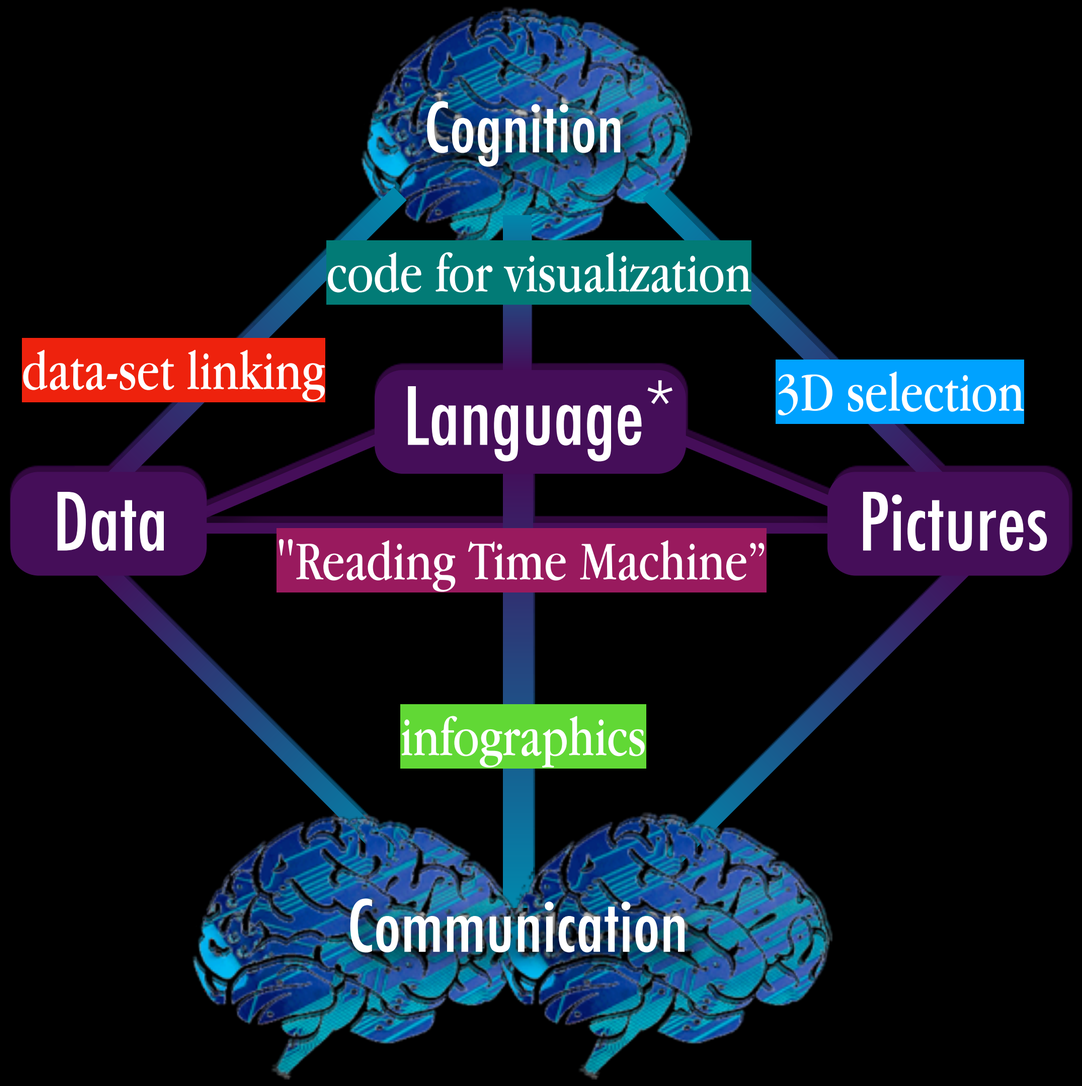


augmented reality

Fig. 9

A "Enchanted" book, with moving, changeable images, and interactive content, on each physical page. On view at "The Power of Poison" display at the American Museum of Natural History, in New York City from November 2013 -- August 2014. An online version of the material is at http://www.amnh.org/the-power-of-poison#page/, but it is key to point out the the experience online, on a screen, is a very poor substitute for the physical object--where each physical page updates as it is turned.

AI in My Astro+Visualization Life TODAY



*"Language" includes words & math

"Reading Time Machine”

code for visualization

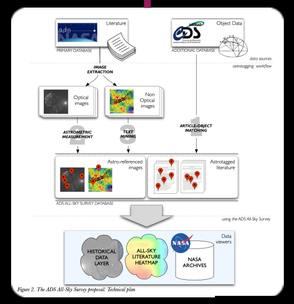
data-set linking

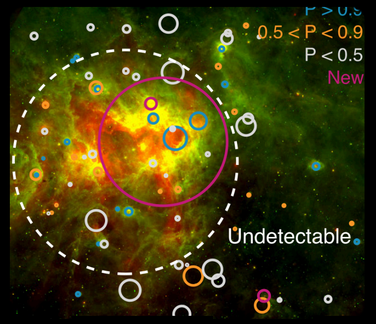
3D selection

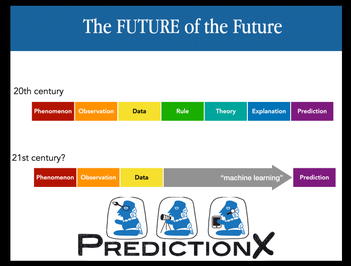
infographics

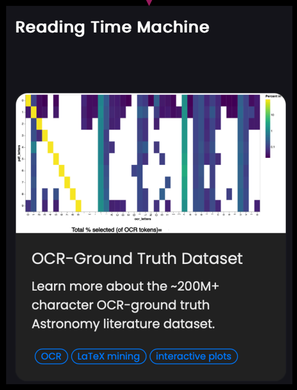
"Reading Time Machine"

code for visualization

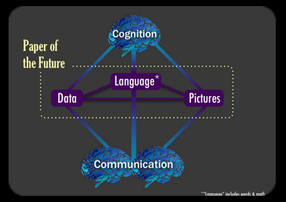


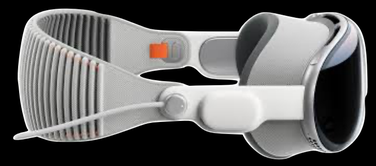






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024





data-set linking

3D selection

infographics

“State-of-the-Art” ML in 2014

The Astrophysical Journal Supplement Series, 214:3 (18pp), 2014 September doi:10.1088/0067-0049/214/1/3
© 2014. The American Astronomical Society. All rights reserved. Printed in the U.S.A.

THE MILKY WAY PROJECT: LEVERAGING CITIZEN SCIENCE AND MACHINE LEARNING TO DETECT INTERSTELLAR BUBBLES

Christopher N. Beaumont[1,2], Alyssa A. Goodman[2], Sarah Kendrew[3], Jonathan P. Williams[1], and Robert Simpson[3]
[1] Institute for Astronomy, University of Hawai’i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA; beaumont@ifa.hawaii.edu
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[3] Department of Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
Received 2013 October 3; accepted 2014 June 7; published 2014 August 20

ABSTRACT

We present Brut, an algorithm to identify bubbles in infrared images of the Galactic midplane. Brut is based on the Random Forest algorithm, and uses bubbles identified by >35,000 citizen scientists from the Milky Way Project to discover the identifying characteristics of bubbles in images from the *Spitzer Space Telescope*. We demonstrate that Brut’s ability to identify bubbles is comparable to expert astronomers. We use Brut to re-assess the bubbles in the Milky Way Project catalog, and find that 10%–30% of the objects in this catalog are non-bubble interlopers. Relative to these interlopers, high-reliability bubbles are more confined to the mid-plane, and display a stronger excess of young stellar objects along and within bubble rims. Furthermore, Brut is able to discover bubbles missed by previous searches—particularly bubbles near bright sources which have low contrast relative to their surroundings. Brut demonstrates the synergies that exist between citizen scientists, professional scientists, and machine learning techniques. In cases where “untrained” citizens can identify patterns that machines cannot detect without training, machine learning algorithms like Brut can use the *output* of citizen science projects as *input* training sets, offering tremendous opportunities to speed the pace of scientific discovery. A hybrid model of machine learning combined with crowdsourced training data from citizen scientists can not only classify large quantities of data, but also address the weakness of each approach if deployed alone.

Key words: H II regions – ISM: bubbles – methods: data analysis – stars: formation

Online-only material: color figures

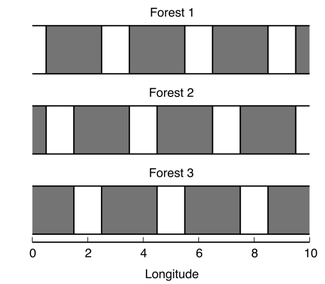


Figure 7. Illustration of the zones used to train each Random Forest. Each forest is trained using the data from the shaded regions in its zone, and used to classify the light regions.

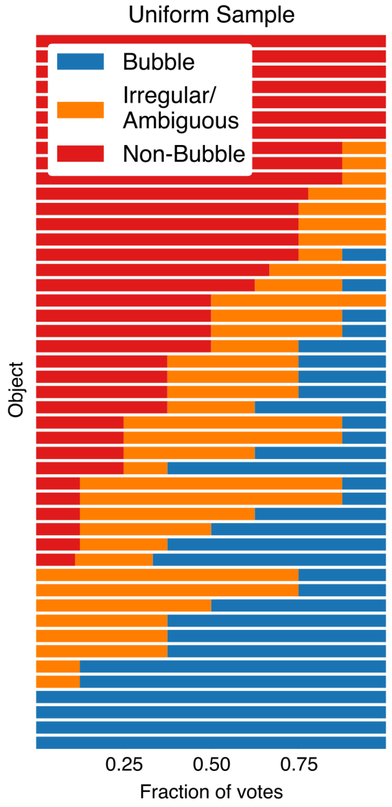


Figure 12. Same as Figure 10, for the uniform sample.

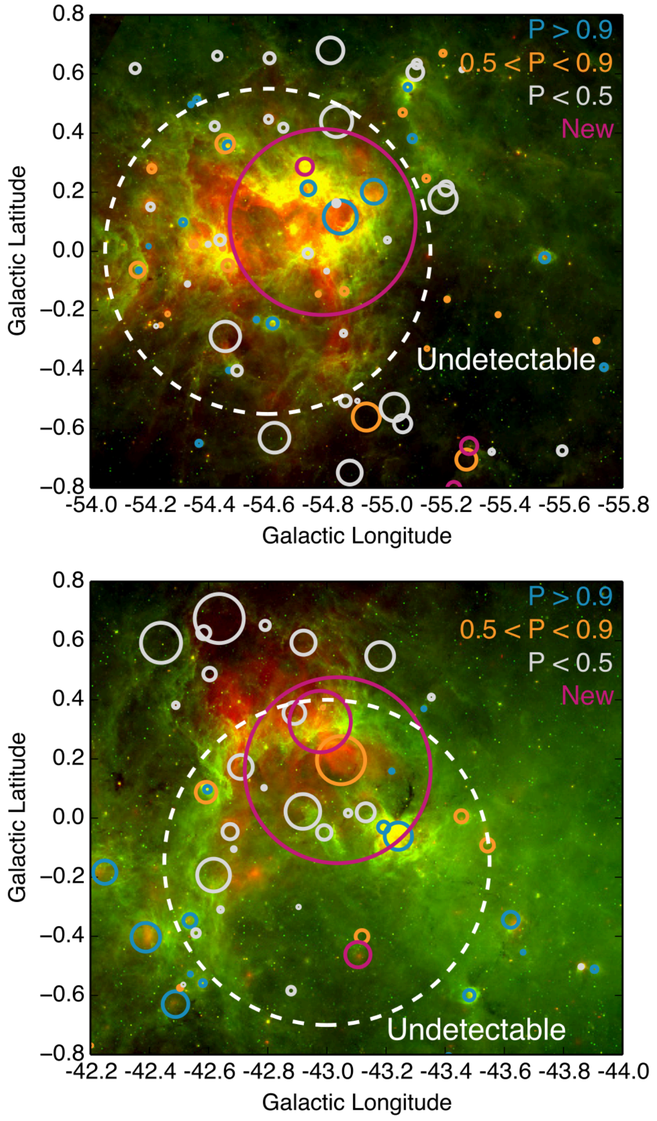
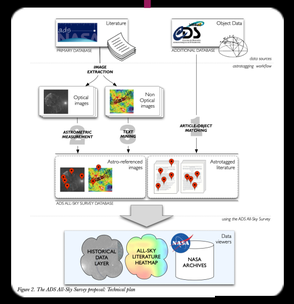
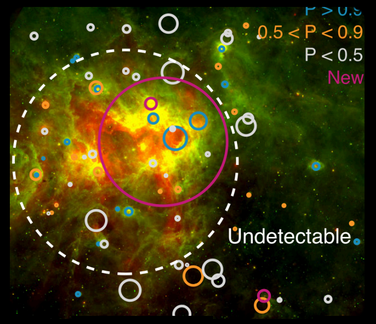


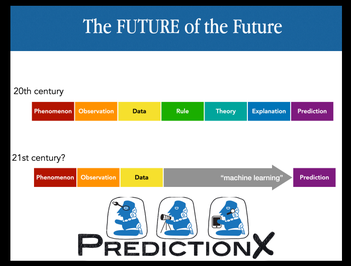
Figure 16. Two fields with overabundances of low-probability bubbles in the MWP catalog.

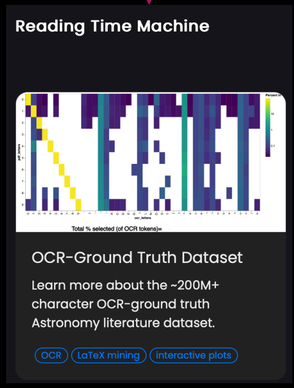
"Reading Time Machine"

code for visualization

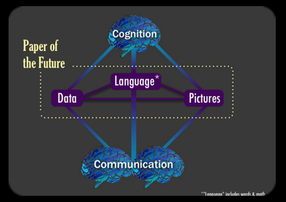


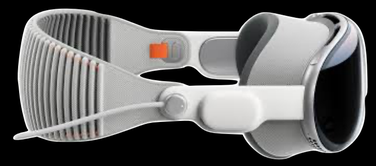






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024





data-set linking

3D selection

infographics

ADS All-Sky Survey & Astronomy Rewind

"putting articles and images (back) on the Sky"

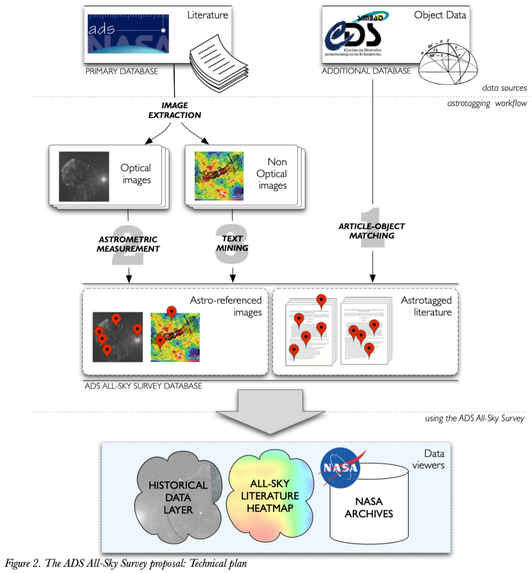


Figure 2. The ADS All-Sky Survey proposal: Technical plan

1. Images Extracted from Journal Articles

ON A GREAT NEBULOUS REGION AND ON THE QUESTION OF ABSORBING MATTER IN SPACE AND THE TRANSPARENCY OF THE NEBULAE

By E. E. BARNARD

While photographing the region of the great nebula of *ρ Ophiuchi* (which I had found with the Willard lens) at the Lick Observatory in 1893, the plates with the small lantern lens (1¼ inches diameter, also attached to the Willard mounting) showed a remarkable nebula involving the 4.5 magnitude star *ν Scorpii* (Plate I). It had not been noticed on the Willard lens photograph, where it was very faint and near the edge of the plate. The discovery of this object therefore is due to the small lantern lens.

astrophysics data system

"historical" images

© 2006. The American Astronomical Society. All rights reserved. Printed in U.S.A.

THE COMPLETE SURVEY OF STAR-FORMING REGIONS: PHASE I DATA

Naomi A. Ridge,[1] James Di Francesco,[2] Helen Kirk,[2,3] Di Li,[1,4] Alyssa A. Goodman,[5] João F. Alves,[1] Héctor G. Arce,[6] Michelle A. Borkin,[5] Paola Caselli,[7] Jonathan B. Foster,[5] Mark H. Heyer,[8] Doug Johnstone,[2,3] David A. Kosslyn,[5] Marco Lombardi,[9] Jaime E. Pineda,[1] Scott L. Schnee,[1] and Mario Tafalla[10]

Received 2005 November 8; accepted 2006 February 22

ABSTRACT

We present an overview of data available for the Ophiuchus and Perseus molecular clouds from Phase I of the COMPLETE Survey of Star-Forming Regions. This survey provides a range of data complementary to the *Spitzer*

astrophysics data system + astronomy image explorer

Key words: ISM: clouds — stars: formation — surveys

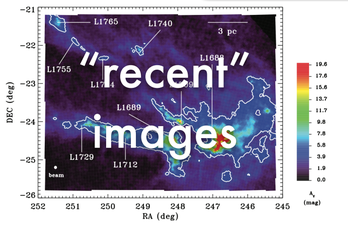


FIG. 3.—Map of extinction in Ophiuchus derived using 2MASS NICER. The contour indicates an A_V of 3 mag and is repeated in subsequent figures for orientation. Note that the small "hole" at the center of the L1688 cluster is an artifact due to the high extinction at that position.

2. Missing coordinate metadata added back to images, either...

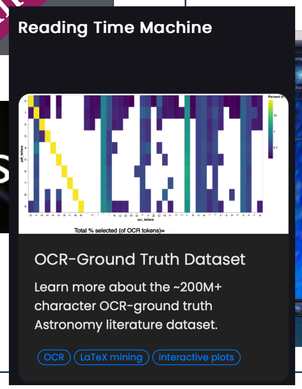
...automatically, applying astronometry.net to wide-field optical images, or

Astrometry.net

via "Astronomy Rewind" Zooniverse Science Project

"Reading Time Machine"

ZOONIVERSE

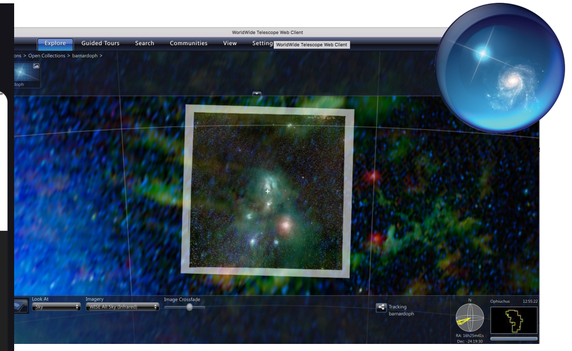


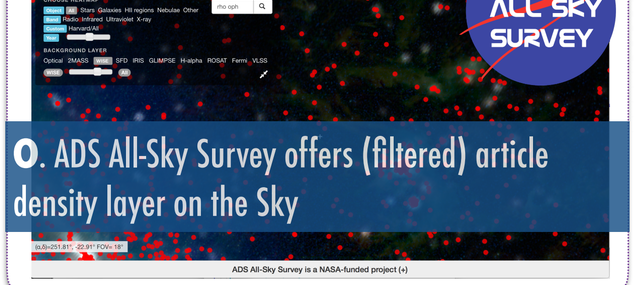
3. "Solved" images returned to ADS & Astronomy Image Explorer

astrophysics data system

astronomy image explorer

4. New button in Astronomy Image Explorer offers image-in-context, using AAS' WorldWide Telescope in the browser





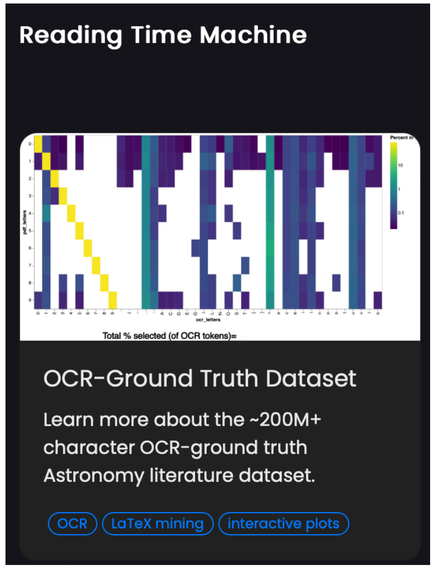
0. ADS All-Sky Survey offers (filtered) article density layer on the Sky

click entries on the timeline to try out services

ADS	WorldWide Telescope	Zooniverse	Astrometry.net	ADS All Sky Survey	Astronomy Image Explorer	Astronomy Rewind
1992	2008	2009	2011	2014	2014	2017

"Reading Time Machine"

International Journal on Digital Libraries (2024) 25:471–491
https://doi.org/10.1007/s00799-023-00350-9



The digitization of historical astrophysical literature with highly localized figures and figure captions

Jill P. Naiman[1,2] · Peter K. G. Williams[3] · Alyssa Goodman[3]

Received: 3 November 2022 / Revised: 14 February 2023 / Accepted: 20 February 2023 / Published online: 22 March 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Scientific articles published prior to the "age of digitization" in the late 1990s contain figures which are "trapped" within their scanned pages. While progress to extract figures and their captions has been made, there is currently no robust method for this process. We present a YOLO-based method for use on scanned pages, after they have been processed with Optical character recognition (OCR), which uses both grayscale and OCR features. We focus our efforts on translating the intersection-over-union (IOU) metric from the field of object detection to document layout analysis and quantify "high localization" levels as an IOU of 0.9. When applied to the astrophysics literature holdings of the NASA astrophysics data system, we find F1 scores of 90.9% (92.2%) for figures (figure captions) with the IOU cut-off of 0.9 which is a significant improvement over other state-of-the-art methods.

Keywords Scholarly document processing · Document layout analysis · Astronomy

"Reading Time Machine"

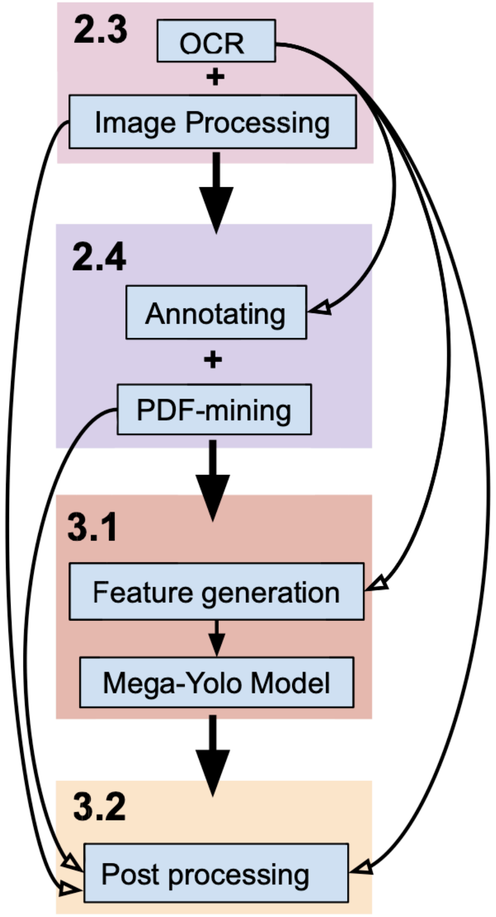
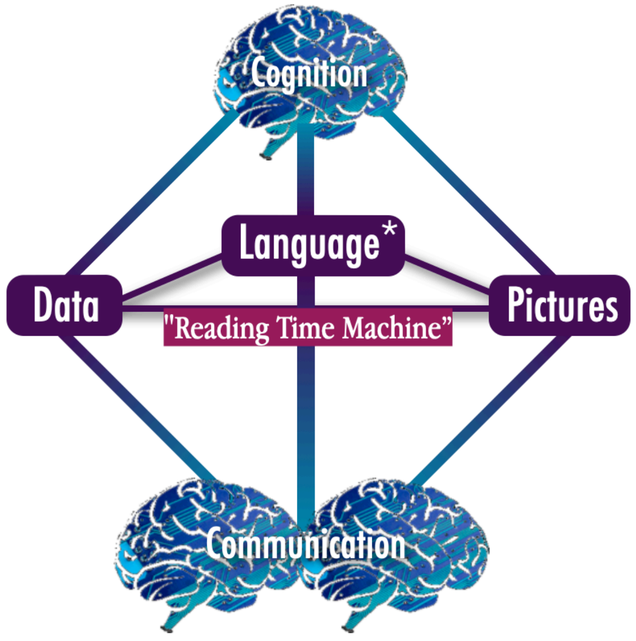


Fig. 3 Our overall pipeline is shown as four main steps. OCR and image processing is discussed in Sect. 2.3. Class definitions and "codebook", along with annotations and PDF mining are discussed in Sect. 2.4. Feature selection and deep learning model description is housed in Sect. 3.1, and post-processing techniques are discussed in Sect. 3.2. (See Appendix 6 and Fig. 10 for a larger breakdown of these steps.)



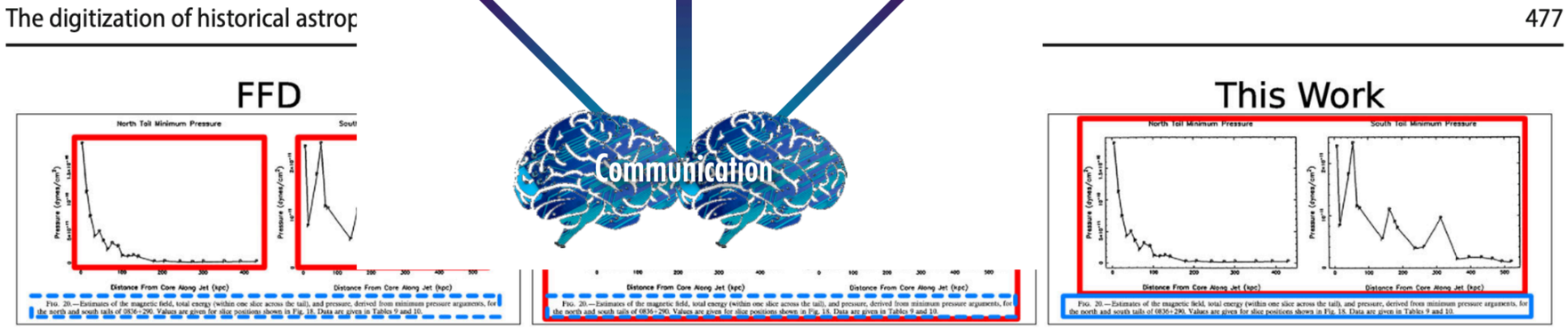


Fig. 4 Illustration of the "codebook" for defining figures and figure captions (right) in comparison with other works (left, center). For example, some works [e.g., [19]] (left red boxes) typically split subfigures into separate figures while other works [14,18,32] (center red box) combine figure captions and figures into a single "figure" box. Caption boxes (shown in blue) in other methods are delineated with dashed squares as applications often focus on detection of figures alone [e.g., [32]]. Note as there is no overlap in datasets for direct comparison, the boxes from other works are shown for illustrative purposes only. See Sect. 4.2 for a comparison of our category definitions and methodology with a subset of several of these datasets (color figure online)

RTM Project led by Jill Naiman, UIUC, funded by NASA ADAP

ARTICLE

Large Synthetic Data from the arχiv for OCR Post Correction of Historic Scientific Articles

"Reading Time Machine"

Authors: J. P. Naiman, Morgan G. Cosillo, Peter K. G. Williams, Alyssa Goodman | Authors Info & Claims

Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26–29, 2023, Proceedings
Pages 265 - 274 • https://doi.org/10.1007/978-3-031-43849-3_23

Published: 26 September 2023 Publication History

ChatGPT 4o ›

The paper **"Large Synthetic Data from the arχiv for OCR Post Correction of Historic Scientific Articles"** by Naiman et al. presents a new approach to improving the accuracy of **optical character recognition (OCR)** for **historic astrophysical literature**, particularly from the **NASA Astrophysics Data System (ADS)**. Here's a concise summary:

🔍 Goal

To improve OCR accuracy for historical scientific texts by generating a **large synthetic dataset** from arχiv LaTeX source files, and using it to train deep learning models for **post-OCR correction**.

🛠️ Methodology

- **Source**: Mined LaTeX files from ~65k arχiv astronomy papers (1991–2011); ~7,850 successfully aligned with OCR outputs.
- **Synthetic Ground Truth (SGT)**: Created by marking words, equations, citations, and references in LaTeX source using the `tikzmark` package.
- **OCR Process**: Rendered PDFs were OCR'd using **Tesseract**.
- **Alignment**: An algorithm matched OCR output with ground truth via bounding box overlap and edit distance.
- **Dataset Size**: 1.5M sentence pairs with **203 million character-level alignments**.

🤖 Model Training

- Trained a transformer model (**ByT5**) on the dataset.
- Baseline results showed:
  - **67% improvement in Character Error Rate (CER)** and
  - **60% improvement in Word Error Rate (WER)** on synthetic data.
- When tested on **real historic ADS data**, improvements were smaller but significant:
  - **CER improved by 7.71%**
  - **WER improved by 18.82%**

📊 Key Contributions

- **Largest known scientific OCR post-correction dataset**.
- **Open-source tools** and **interactive dashboards** for exploration.
- **Generalizable pipeline** for other scientific domains with LaTeX sources.

⚠️ Limitations

- Not all arχiv articles compile successfully due to outdated LaTeX.
- Public sharing of full dataset is restricted (working with arχiv to resolve).
- More robust post-correction for ambiguous LaTeX constructs (like multiple representations of ≤90%) remains future work.

🔗 Resources

- Project site & dashboard: https://readingtimemachine.github.io/projects/1-ocr-groundtruth-may2023
- Code & subset of data: GitHub repository

Let me know if you'd like a shorter version, a diagram of the pipeline, or a breakdown of how this might apply to another field!

🔁 Diagram: OCR Post-Correction Pipeline

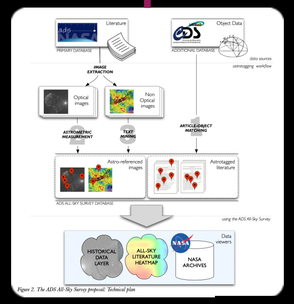
plaintext

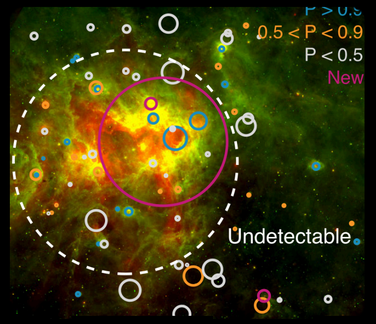
```
arχiv LaTeX Sources (1991–2011)
          │
          ▼
 Parse & Mark Text (Tikzmark in LaTeX)
          │
          ▼
 Compile to PDF → OCR with Tesseract
          │                │
          └──────→ Extract Words & Locations
                           │
                           ▼
        Align OCR with Ground Truth
                           │
                           ▼
      Train ByT5 Correction Model
                           │
                           ▼
   Apply to Historic ADS OCR Text
```

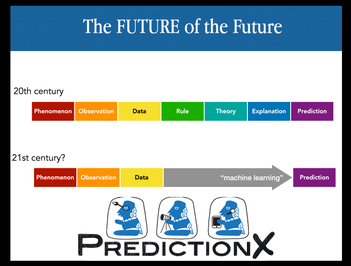
RTM Project led by Jill Naiman, UIUC, funded by NASA ADAP

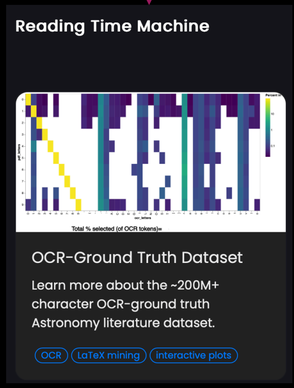
"Reading Time Machine"

code for visualization

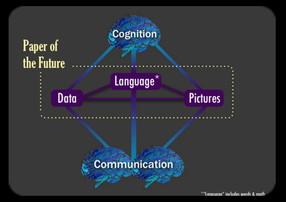


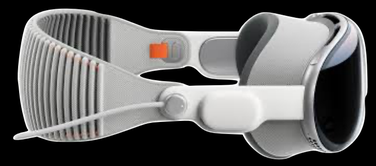






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024





data-set linking

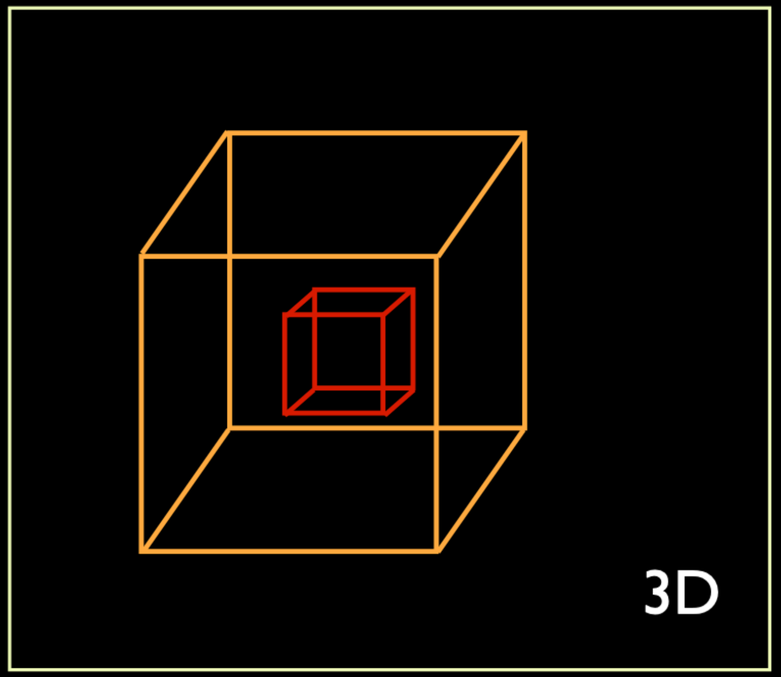
3D selection

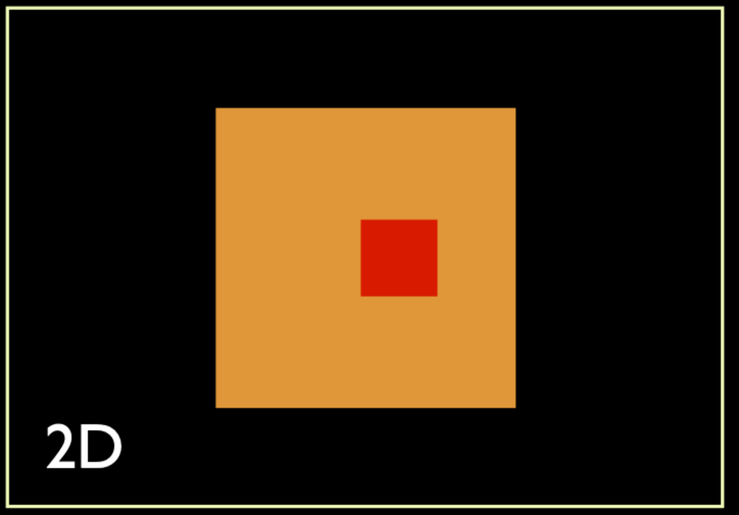
infographics

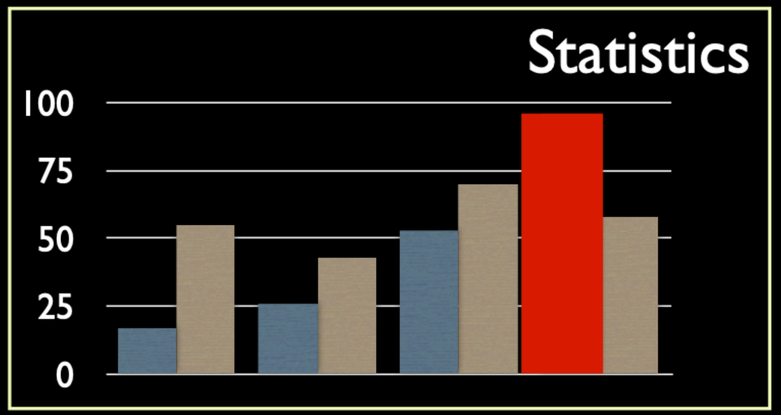
glue
multidimensional data exploration
glueviz.org

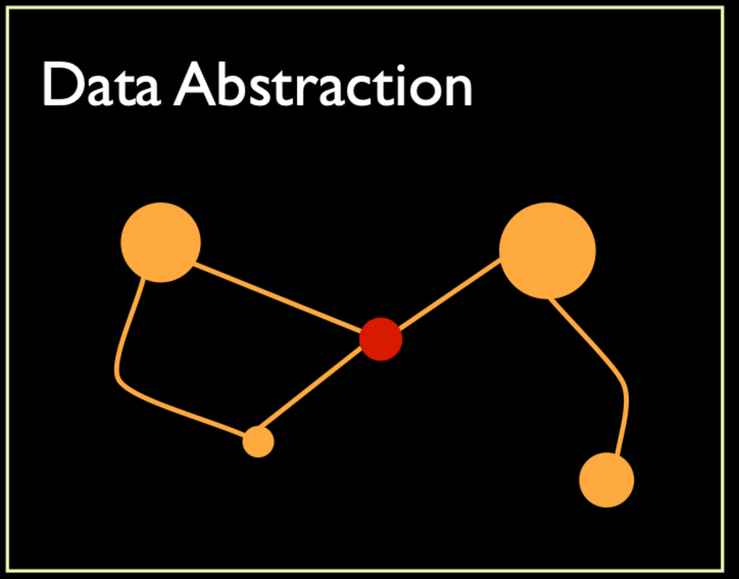
glupyter
multidimensional data exploration

"Linked Views of High-dimensional Data"

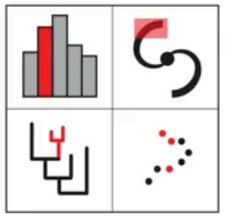








data-set linking

What is  glue?

multidimensional data exploration

data-set linking

data files' common attributes are glued

glue

multidimensional data exploration

data files' common attributes are **glued**

LIVE Environments

About LIVE | Capabilities | Tech | Who ar

LIVE

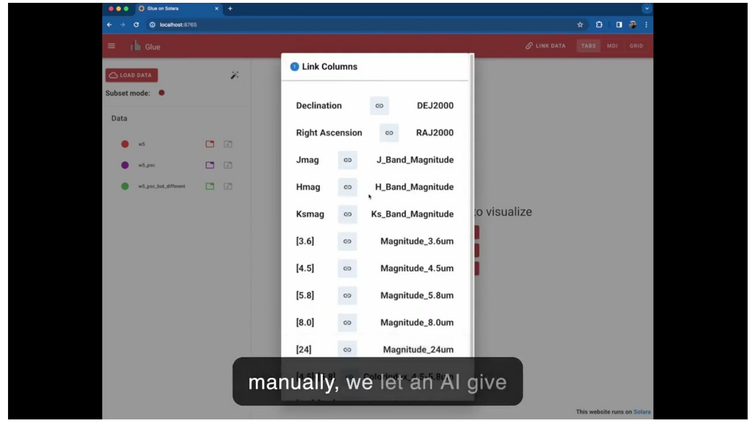
Artificial Intelligence

The ever-expanding capabilities of large-language models and other machine learning techniques can be leveraged to improve LIVE environments. A few areas of focus are identified below, and more are yet to come.

AI-Enhanced Linking Options

Large Language Models (LLMs) can be effectively used to help users automatically create links between diverse datasets with column names that don't conform to specific metadata standards (Zhang et al. 2023, Lobo et al. 2023) . By encoding the variety of columns naming conventions present in a large amount of real world data, LLMs can effectively predict when columns in different datasets refer to the same entity (a form of Entity Resolution), and provide these proposals to the user for acceptance/verification.

We have a short demo showing an **AI-enhanced data linking tool** working in glue-solara. We will post additional updates here so *"watch this space!"*



AI-Enhanced Selection Functions

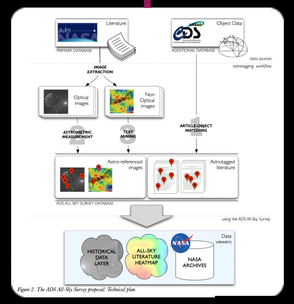
Currently, many of the LIVE tools allow for brushing & linking, as well as algorithmic selection of data subsets, to enable exploratory data analysis. Many recent and ongoing research projects offer clever, often AI-enahnced, tools for selecting data subsets, and it's a near-term goal of the LIVE project to begin incorporating machine intelligence into selection options. While many display types and tables can and will benefit from smart selection, the LIVE team is especially interested in the ~unsolved problem of intelligent 3D selection withing volumetric data, so stay-tuned--and do be in touch if you have ideas!

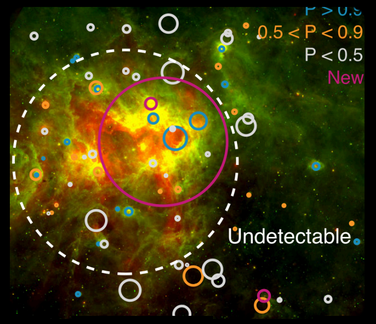
Video courtesy of Maarten Breddels

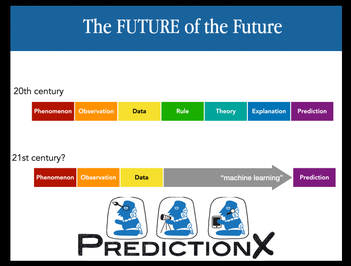
data-set linking

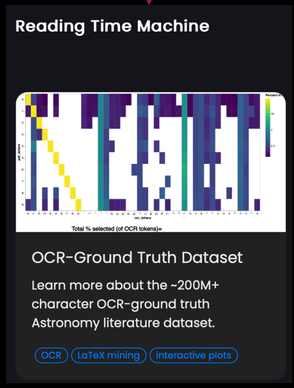
"Reading Time Machine"

code for visualization

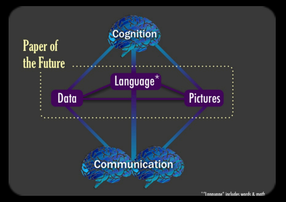


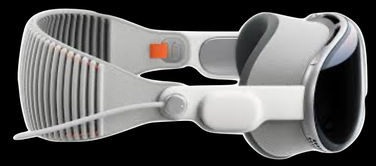






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024



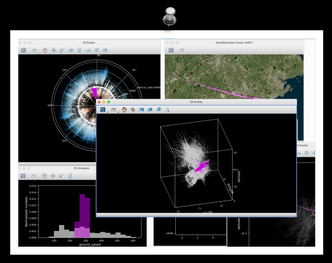


data-set linking

3D selection

infographics

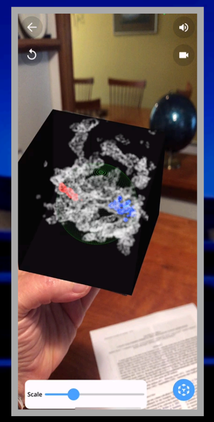
3D selection

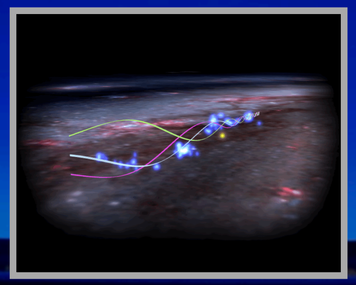


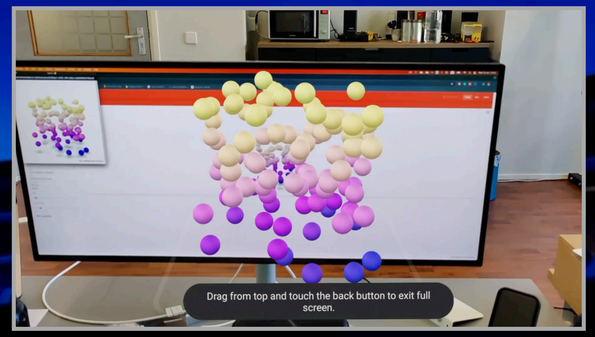
ask (later) for LIVE selection demo…

(and “smart” selection more generally!)

3D selection



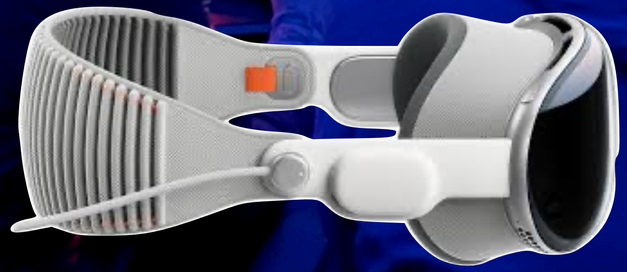




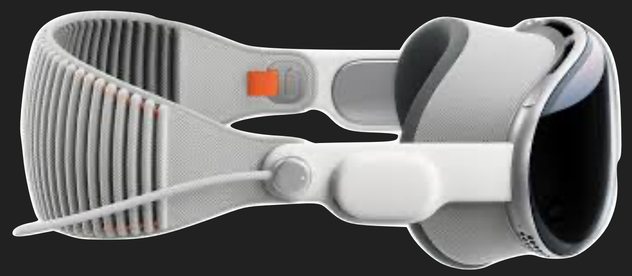
AR for Science and Outreach

WHAT'S NEXT IN AR?

glueviz.org/glue-ar



(and "smart" selection more generally!)

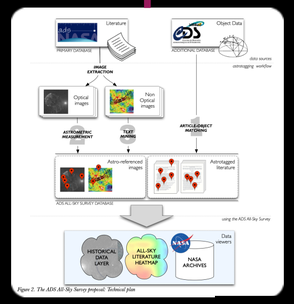


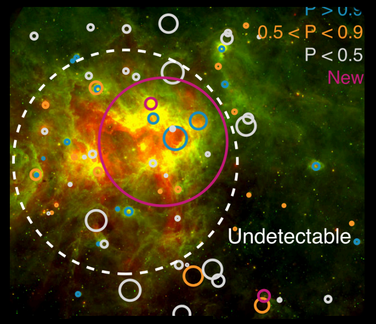
SOLARA

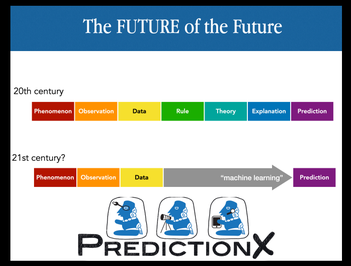
3D selection

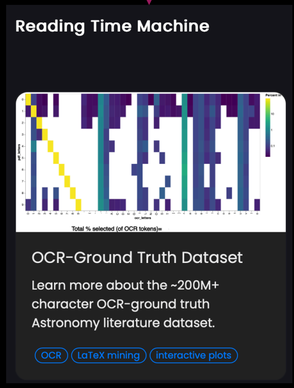
"Reading Time Machine"

code for visualization

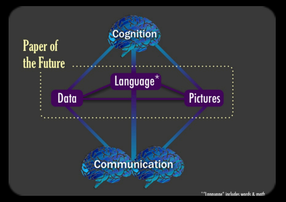


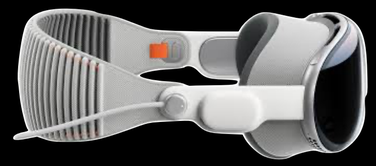






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024



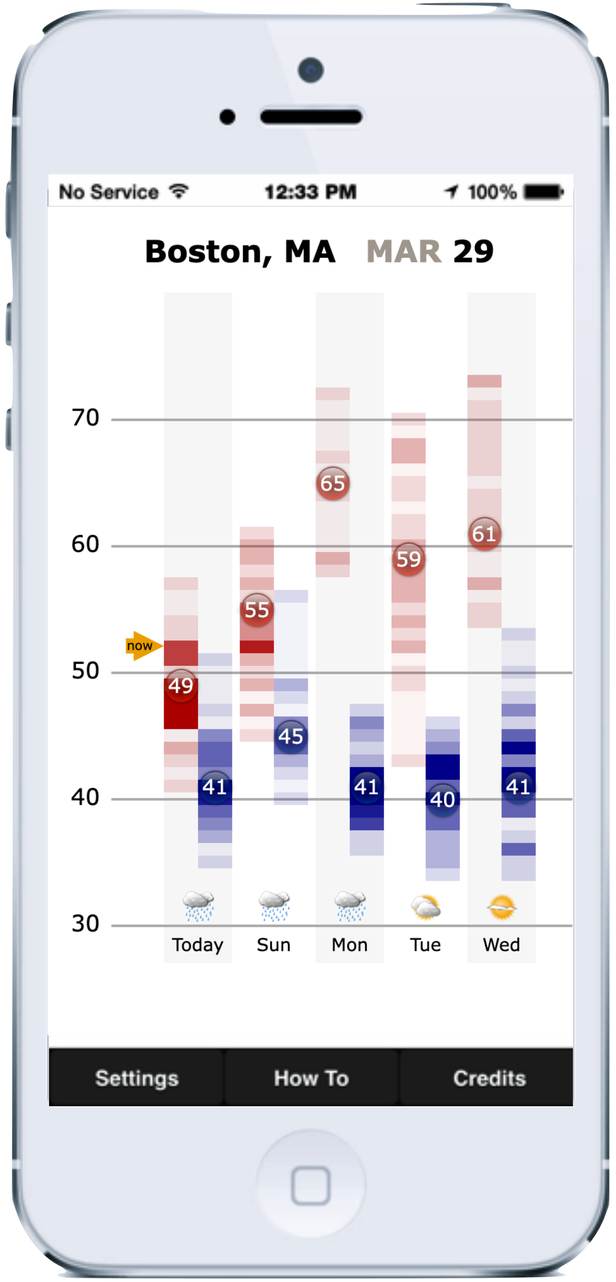


data-set linking

3D selection

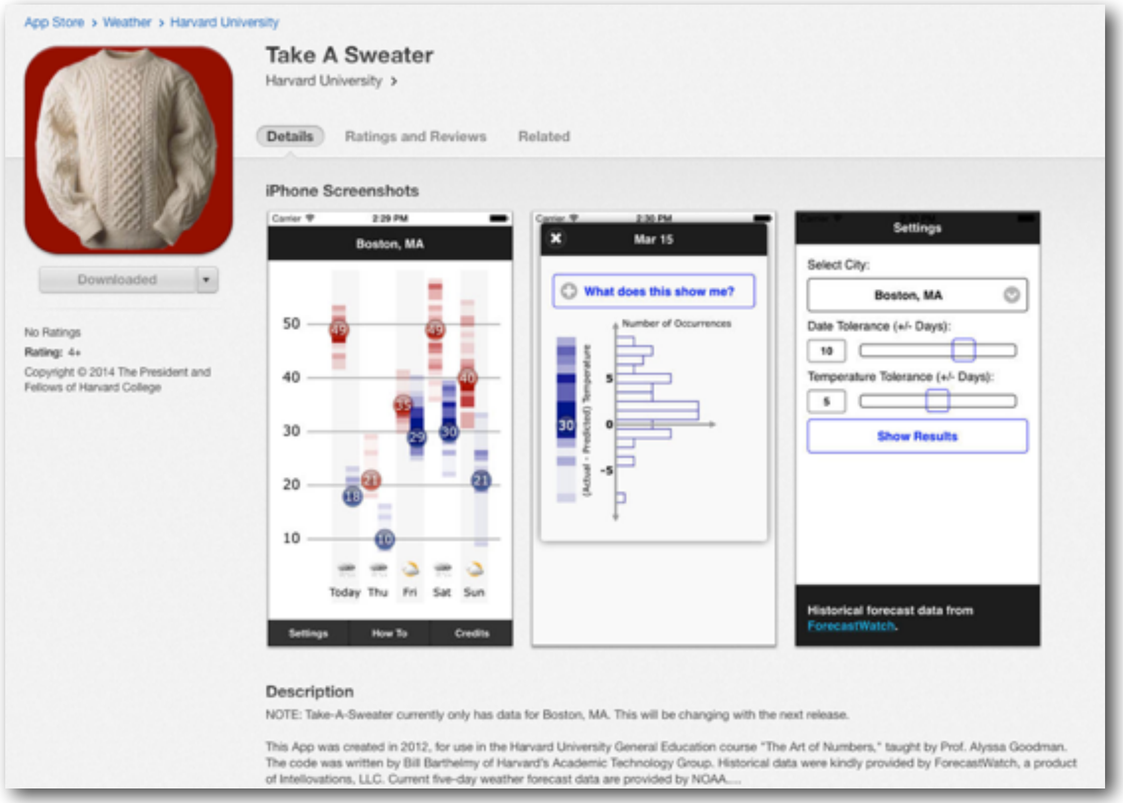
infographics

code for visualization



"Take A Sweater"

2011



takeasweater.com

with thanks to Eric **Floehr** of Forecast Watch and Bill **Barthelmy** of HUIT Academic Technology at FAS

PREDICTIONX

code for visualization

Prediction Journal Visualization

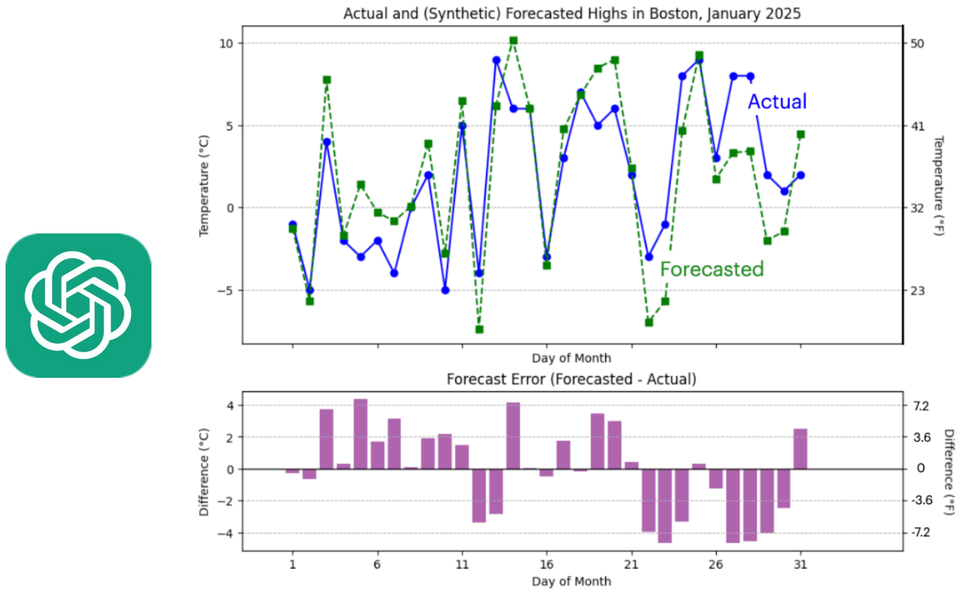
Start Assignment

Due Mar 25 by 11:59pm **Points** 80 **Submitting** a website url

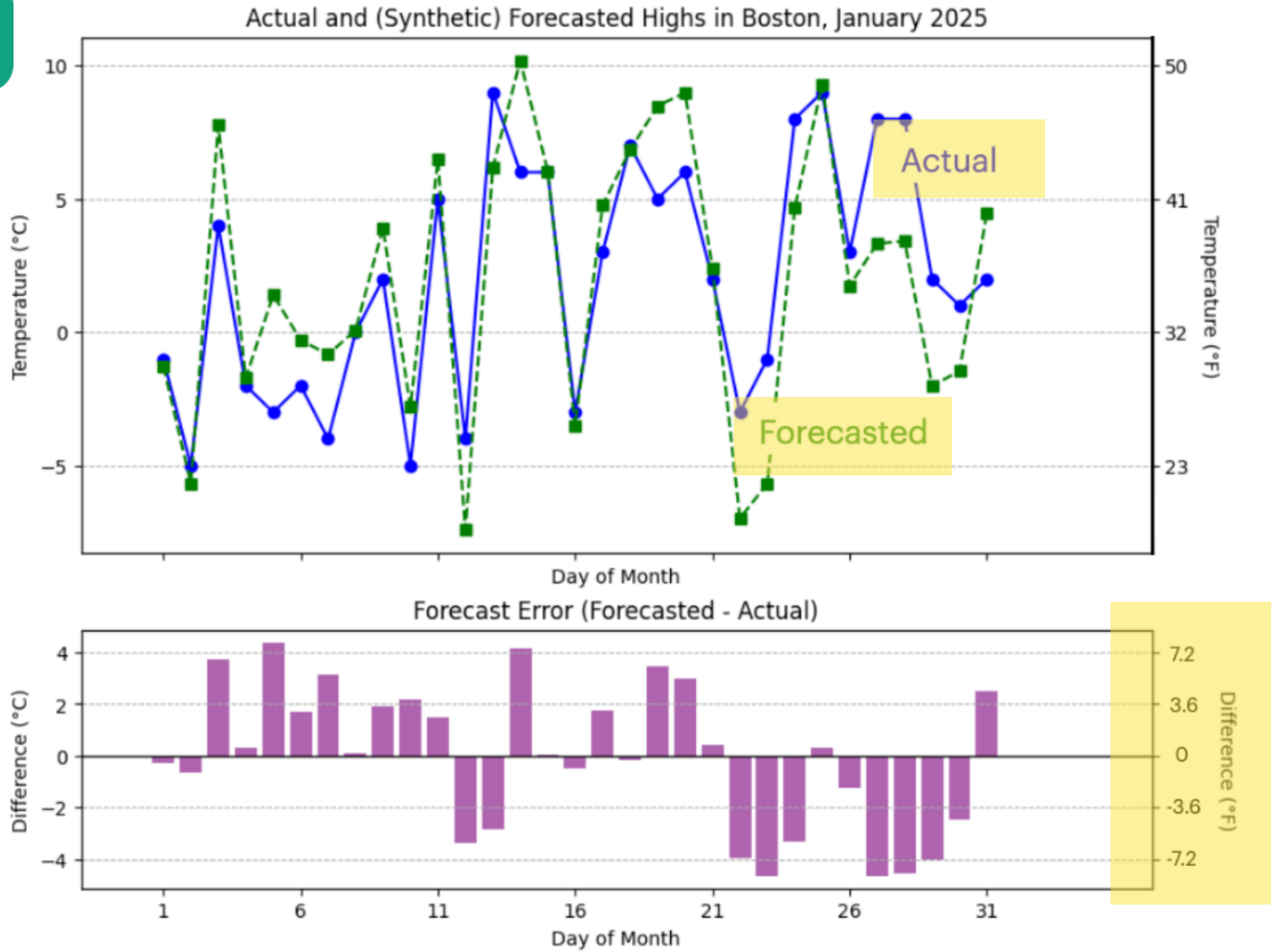
At this point, you should have 4 weeks worth of predictions and actual outcomes. For this assignment, please:

1) **Create a visualization** of the data you've collected (these will likely include, at the very least, your predictions, recorded outcomes, and dates/times). Visualizations can take many forms and **will depend on your predictions and the type of data you collected**. Talk to your TF for ideas and/or help! Feel free to use computer programs or complete visualizations by hand (see example below or explore more options here--and remember, yours need NOT be super-fancy, just informative--hand-plotting is fine).

2) **Reflect on your prediction process over the past 4 weeks** using your visualization and the Padua Rainbow. How have your predictions changed in the past weeks? What parts of the Padua Rainbow have you engaged, what parts have you not engaged, and why? What story does your visualization tell about your own predictive process? Your reflection should be ~500 words.



code for visualization



manual adjustments!

PREDICTIONX

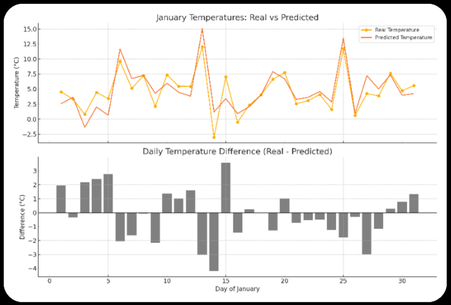
6 days later. . .

Create an image that contains two graphs - one a line graph that shows the month of January in days on the X axis, with the y axis as the temperature. There should be two lines - one showing the real temperature, and the other showing a predicted forecast, so that the difference can be compared. The second graph should quantify that difference for each day - e.g. if one day was 2 degrees warmer than expected, the lower graph would show +2 for that day.

Here are the two graphs:

- The top graph shows the real vs. predicted temperatures for each day in January.
- The bottom graph visualizes the daily difference between real and predicted temperatures (real minus predicted), making it easy to spot where predictions were too warm or too cold.

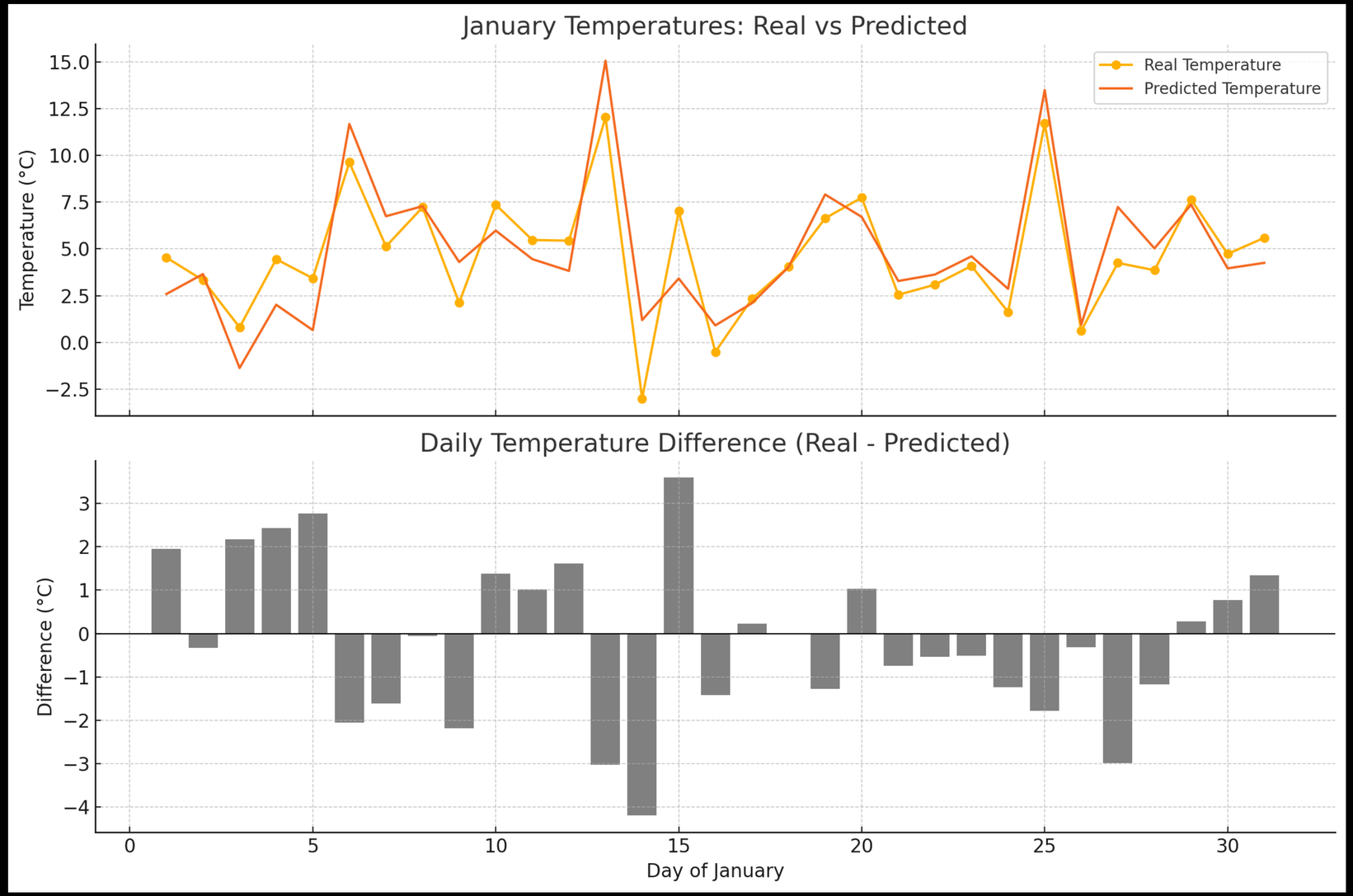
Let me know if you want to adjust the style, range, or add real data. ❶



March 31, 2025
No python, single prompt generation. Holy smokes that's a huge leap of improvement

11:18
Sim Sayer
Remarkably, the graphs appear to correspond

code for visualization

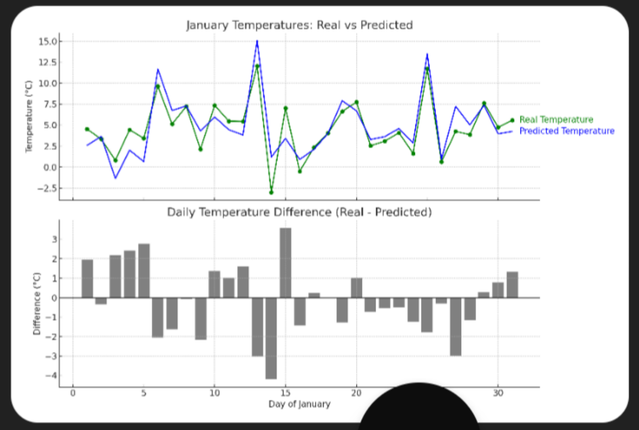


Now instead of a legend, have the text appear near the line instead. And make the real line green and the predicted line blue. Use the same data to build the graph

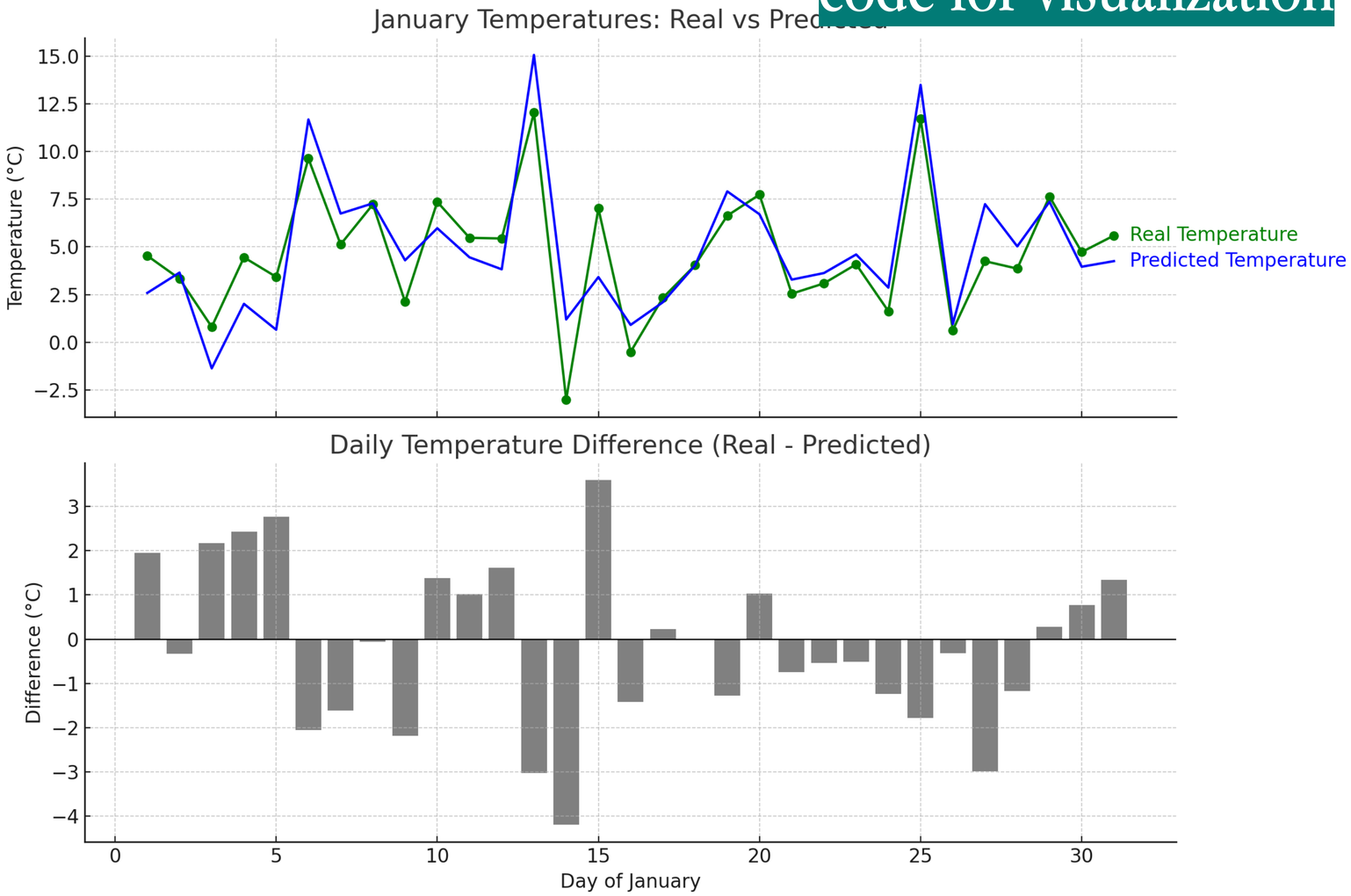
Here's the updated version of the graphs:

- The top graph now uses **green** for the real temperature line and **blue** for the predicted temperature.
- Instead of a legend, the labels appear directly next to the respective lines.

Let me know if you'd like to adjust the font size, move the labels, or apply any other styling.



code for visualization



April 4, five days later: let's add a slider, and get *fancy*… [DEMO, ChatGPT, Jupyter, Colab]

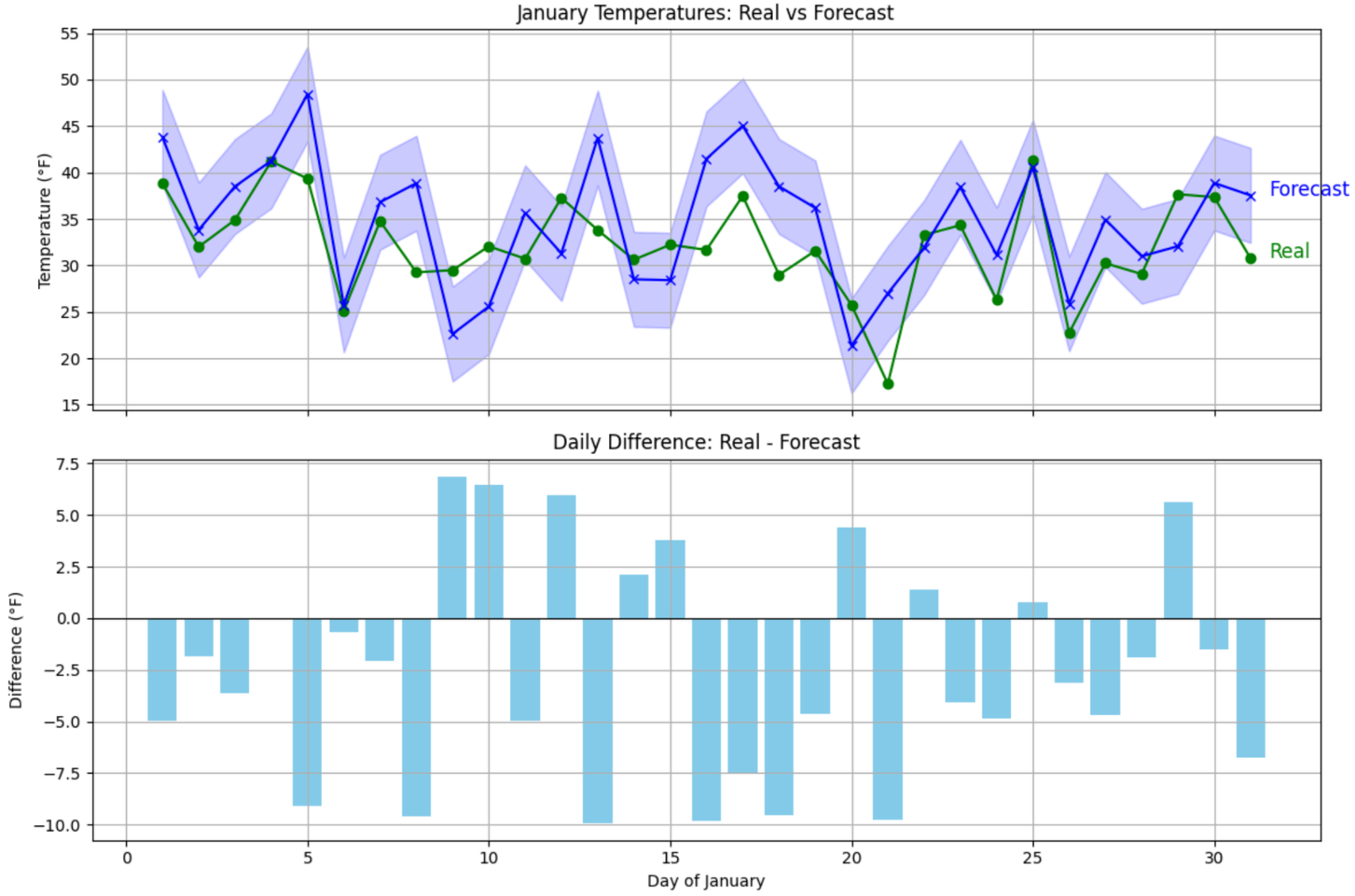
code for visualization

mperature_Forecast.ipynb

ew Insert Runtime Tools Help

Code + Text

uncertainty 5.10




```
plot_forecast
def plot_forecast(uncertainty)
```

<no docstring>

ChatGPT 4o

OK, that's great--can we try a fancier visualization now?

Absolutely, let's level it up! 💡

What kind of *fancy* are you thinking?

Here are a few ideas to get us started:

✨ **Fancy Visualization Upgrades**

1. Interactive Plotly Dashboard

- Use Plotly for beautiful, zoomable, responsive charts
- Sliders, dropdowns, hover tooltips, and uncertainty bands
- Works great in Colab too!

2. 3D Surface or Heatmap

- Show temperature trends across multiple months or locations
- Useful for visualizing patterns a

3D Forecast Temperature Explorer

Forecast Surface | Difference Surface

3. Animated Forecast Evolution

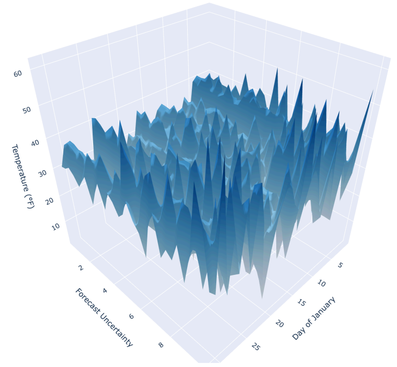
- Animate how the forecast chan
- Could be a `matplotlib.anim`

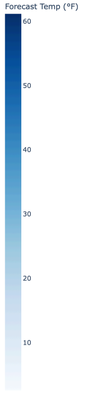
4. AR-Ready or Globe Visualizatic

- For geographic data: World or r

5. Time Series Decomposition (Fa

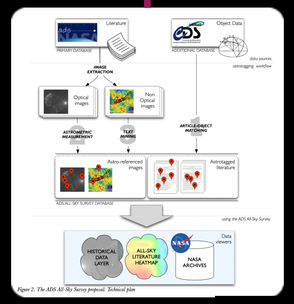
- Show forecast vs. trend, seasor
- Use `statsmodels` or Prophet

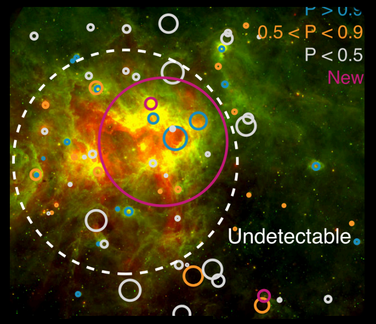


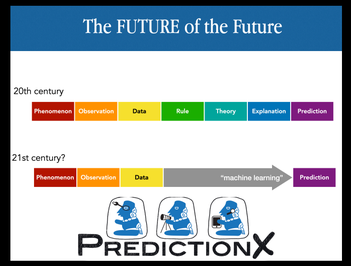


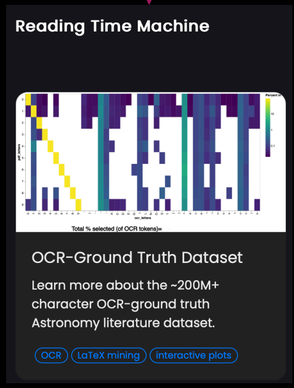
"Reading Time Machine"

code for visualization

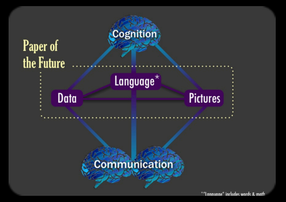


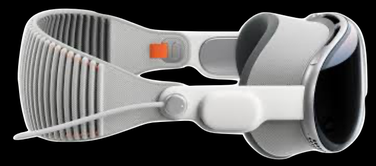






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024

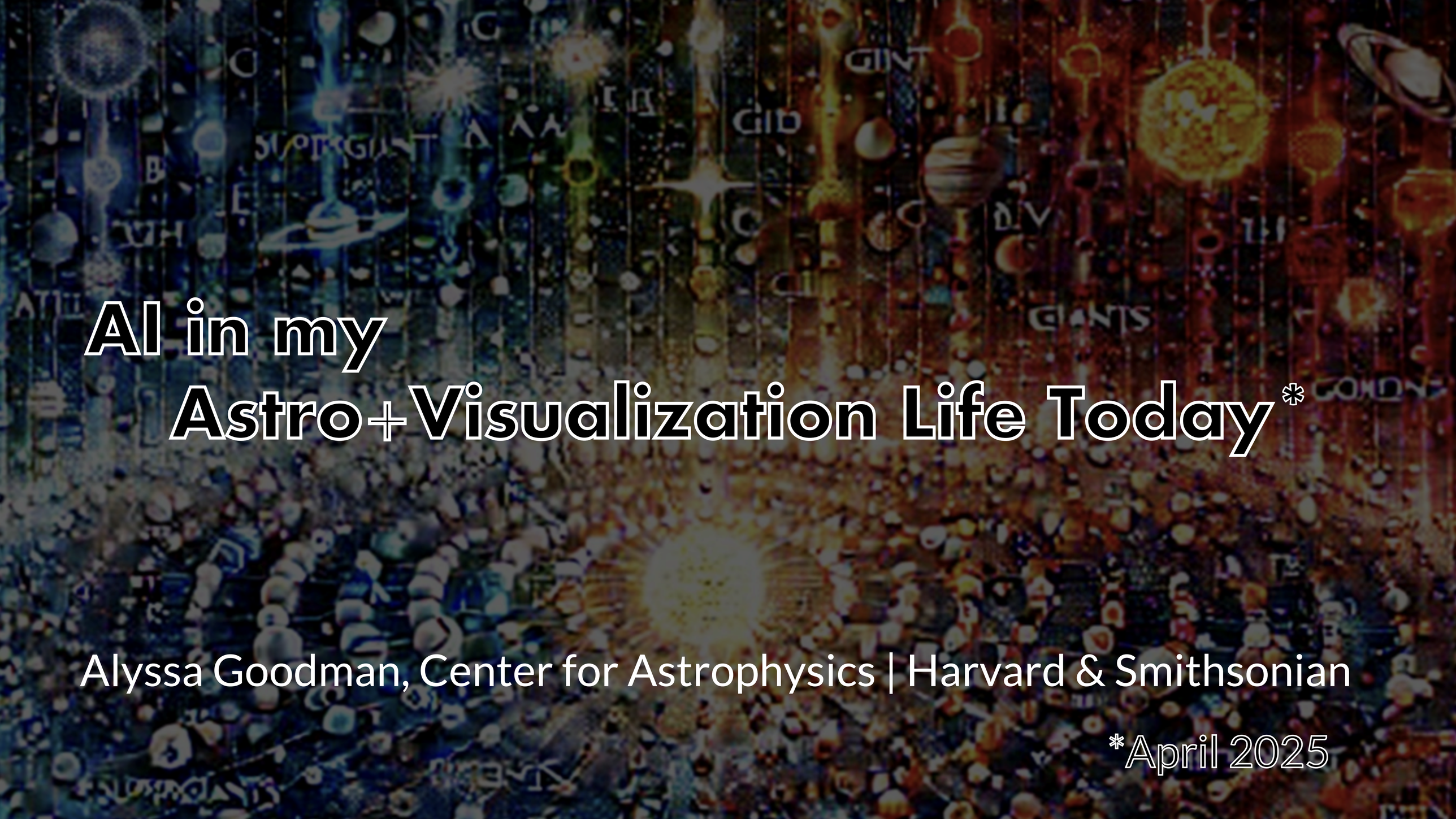




data-set linking

3D selection

infographics



AI in my Astro+Visualization Life Today*

Alyssa Goodman, Center for Astrophysics | Harvard & Smithsonian

*April 2025

AI in my Astro+Visualization Life Today

I have been lucky enough in my career so far to watch, and I hope help, computational technology change what we can learn about our Universe. Today, in 2025, I somewhat unexpectedly find myself involved in a broad range of AI-based and AI-enhanced efforts designed to speed learning and discovery in astrophysics and in science. In this talk, I will offer glimpses into a handful of ongoing AI-enhanced efforts, each of which is very different from the others, yet which work together in a researcher/educator's life to speed progress. Work to be highlighted includes: 1) automated data-set linking in the "glue" and LIVE-Environments visualization environments; 2) The "Reading Time Machine" which uses AI to "read" graphics and images and ingest their content into the ADS Literature archive, as "data," 3) approaches to 3D selection in volumetric data, using both AI and augmented reality (AR); 4) a quest to understand why LLMs are so good at describing infographics, but so terrible at creating them; 5) capabilities of AI for writing code for visualization, in both research and education. The plan of the talk will be to present an overview of each of these efforts, in order to inspire broader discussion of whichever topics evolve as most interesting to the assembled audience.

Thematic questions for discussion

1. visualization and infographics — how good is AI at abstraction, and how good will it become?
2. If we give “all” the MilkyWay3D.org data to AI, what (kinds of) new physical insights might we expect—should we have expectations *a priori*—or is that approach too restrictive? (Note, 3D dust also AI!)

AI in my Astro+Visualization Life Today

I have been lucky enough in my career so far to watch, and I hope help, computational technology change what we can learn about our Universe. Today, in 2025, I somewhat unexpectedly find myself involved in a broad range of AI-based and AI-enhanced efforts designed to speed learning and discovery in astrophysics and in science. In this talk, I will offer glimpses into a handful of ongoing AI-enhanced efforts, each of which is very different from the others, yet which work together in a researcher/educator's life to speed progress. Work to be highlighted includes: 1) automated data-set linking in the "glue" and LIVE-Environments visualization environments; 2) The "Reading Time Machine" which uses AI to "read" graphics and images and ingest their content into the ADS Literature archive, as "data," 3) approaches to 3D selection in volumetric data, using both AI and augmented reality (AR); 4) a quest to understand why LLMs are so good at describing infographics, but so terrible at creating them; 5) capabilities of AI for writing code for visualization, in both research and education. The plan of the talk will be to present an overview of each of these efforts, in order to inspire broader discussion of whichever topics evolve as most interesting to the assembled audience.

Thematic questions for discussion

1. visualization and infographics — how good is AI at abstraction, and how good will it become?
2. If we give “all” the MilkyWay3D.org data to AI, what (kinds of) new physical insights might we expect—should we have expectations *a priori*—or is that approach too restrictive? (Note, 3D dust also AI!)

Abstraction?

Why can't LLMs make Informative Infographics in Astronomy?

Bruna Biz

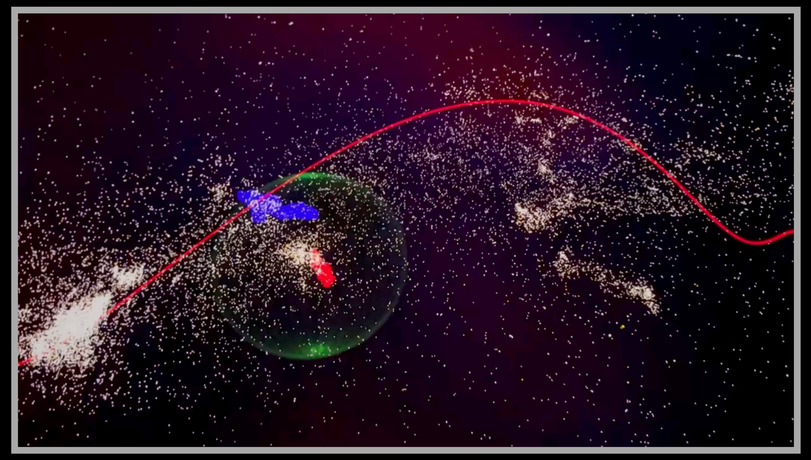
Harvard University

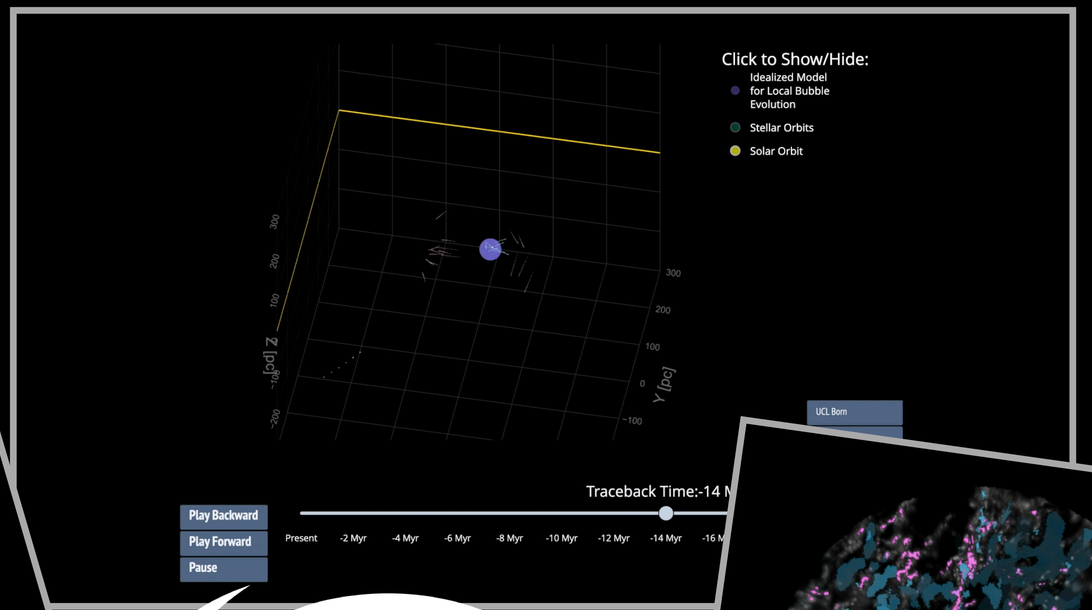
ABSTRACT

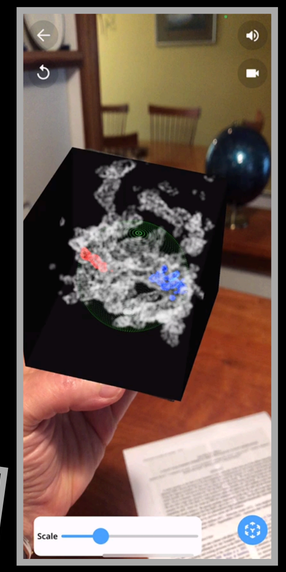
This work is designed to gauge the limitations of generative AI models such as ChatGPT and DALL-E in their ability to accurately produce astrophysical information graphics. We use the Hertzsprung-Russell (H-R) diagram – a fundamental tool in astronomy that maps the relationship between stellar luminosity and temperature – as a test case because of its simple 2D structure and its importance in stellar astronomy. Initial attempts to produce accurate H-R diagrams with ChatgPT led to interesting results which demonstrate the models' tendency to perceive and illustrate cycles (in this case, the stellar life cycle) as circles, resulting in images that deviate from the traditional linear structure of H-R diagrams. Despite high quality H-R diagrams being fed and described to the model, these patterns persist, demonstrating DALL-E's inability to diverge from circular patterns in relation to cyclic phenomena. An investigation into DALL-E's diffusion process also suggests that a big challenge with astrophysical infographic generation is rooted in how tokens are embedded and processed within latent space – an abstract, higher dimensional space where models abstract their vector embeddings. This points to potential improvements that could be made in astrophysical infographic generation to produce more accurate visuals.

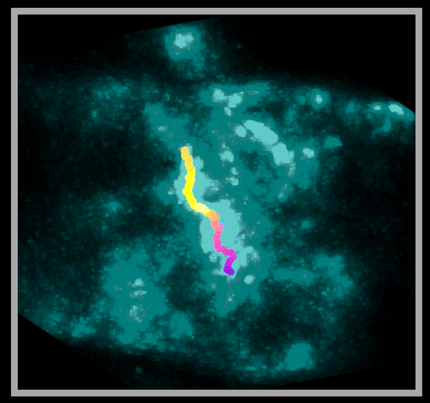
Bruna Biz's Harvard Junior Paper, 2024

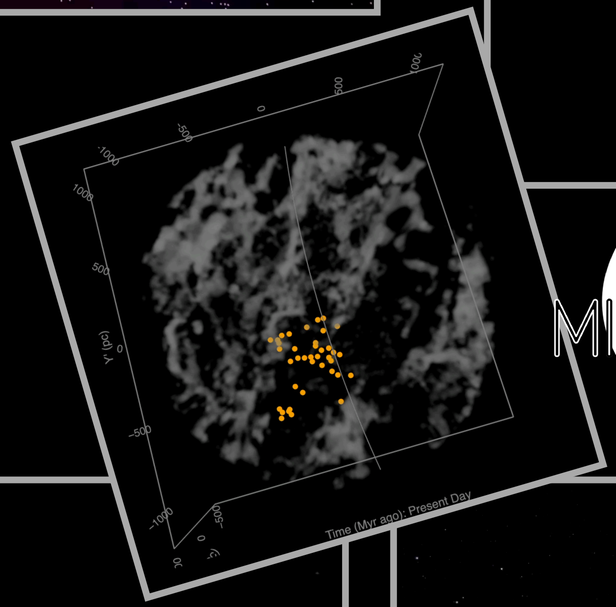
infographics

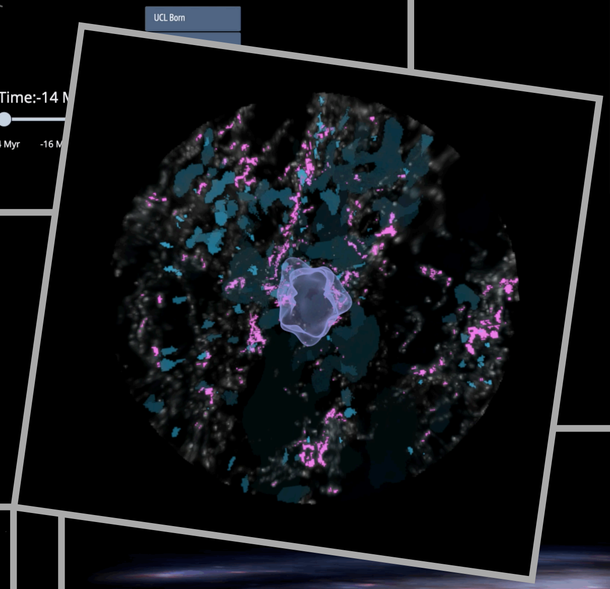




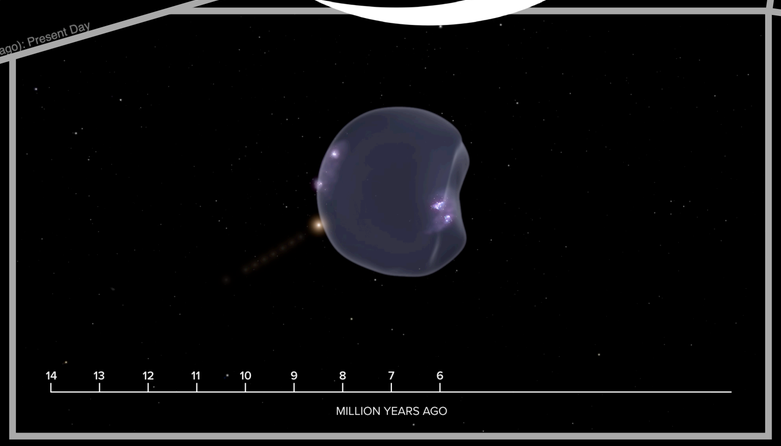


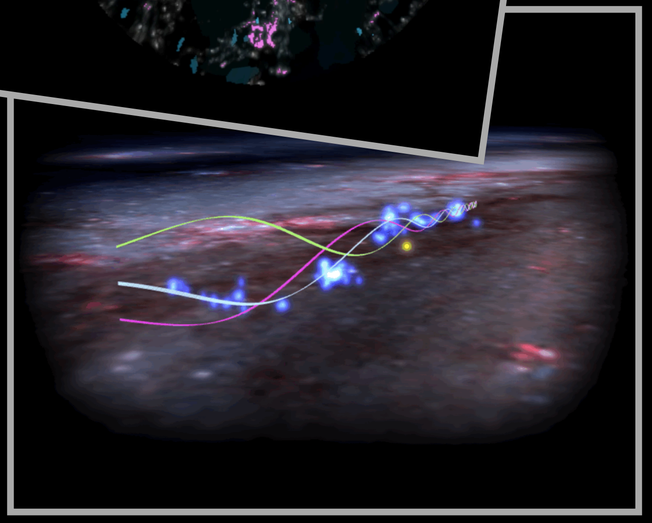






MILKYWAY3D.org



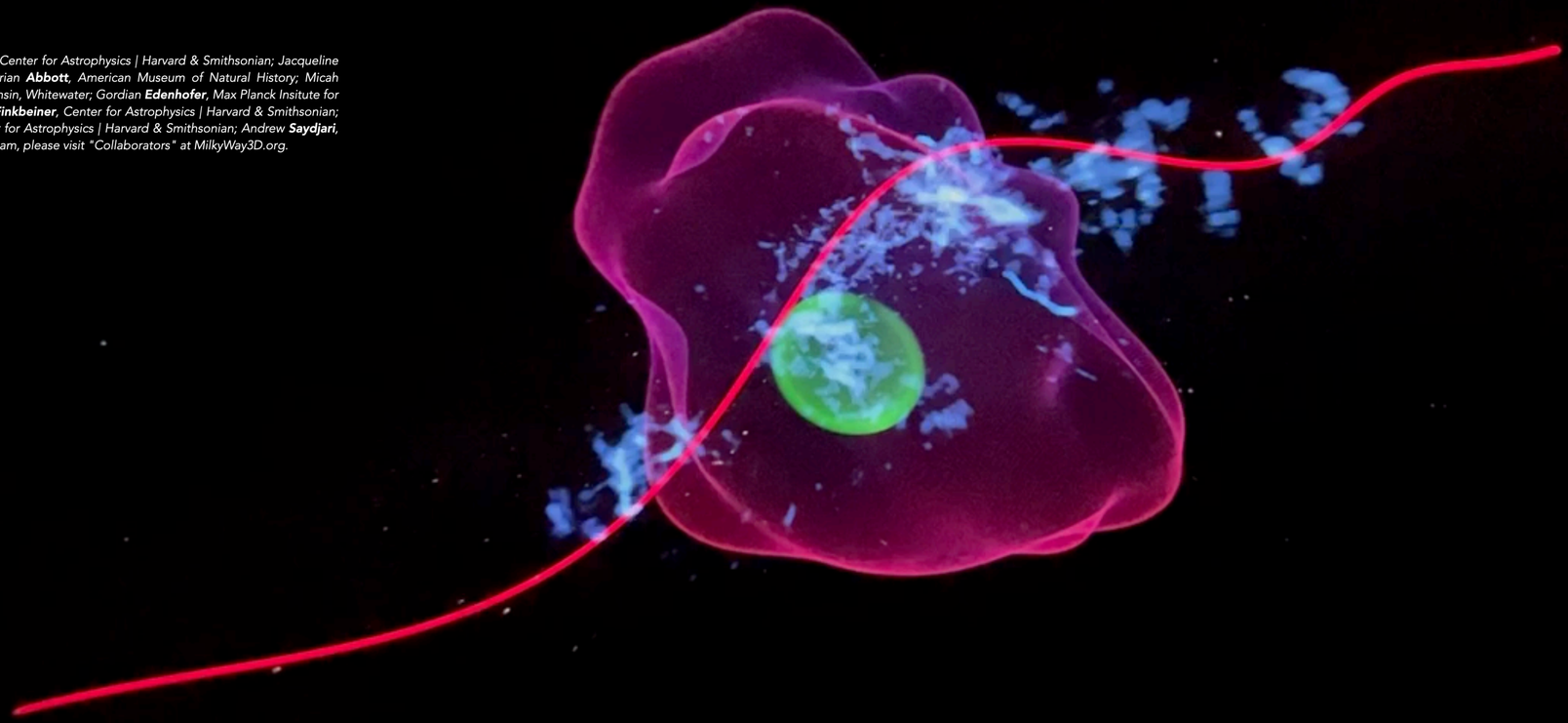


Zucker et al. 2021, Bialy et al. 2021; Zucker et al. 2022, Konietizka et al. 2024, O'Neill et al. 2024, Swiggum et al. 2024

"6D" Movie

A 3D Map of the Star Forming Milky Way, for Everyone

*Alyssa **Goodman**, Center for Astrophysics | Harvard & Smithsonian; Catherine **Zucker**, Center for Astrophysics | Harvard & Smithsonian; Jacqueline **Faherty**, American Museum of Natural History; Joao **Alves**, University of Vienna; Brian **Abbott**, American Museum of Natural History; Micah **Acinapura**, American Museum of Natural History; Robert **Benjamin**, University of Wisconsin, Whitewater; Gordian **Edenhofer**, Max Planck Insitute for Astrophysics—> CfA; Carter **Emmart**, American Museum of Natural History; Douglas **Finkbeiner**, Center for Astrophysics | Harvard & Smithsonian; Ralf **Konietzka**, Center for Astrophysics | Harvard & Smithsonian; Theo **O'Neill**, Center for Astrophysics | Harvard & Smithsonian; Andrew **Saydjari**, Center for Astrophysics | Harvard & Smithsonian—>IAS; and the Full MilkyWay3D.org Team, please visit "Collaborators" at MilkyWay3D.org.*



June 2024 AAS Workshop on Milky Way, plus updated version Monday at CIERA!

Glue (/Users/aagoodman/Library/CloudStorage/GoogleDrive-goodman.alyssa@gmail.com/.shortcut-targets-by-id/1e4RysNClnsaerZT11AaNJ9GKPS9q7vZg/MilkyWay3D.ORG/Sessions/munich_d

Open Session | Export Session | Import Data | Export Data/Subsets | Link Data | x² Arithmetic attributes | Active Subset: None/Create New (the next selection will create a subset)

Data Collection

- Foley2022_arXiv_2212.01405_OrionShell_MW3D[HDU1]
- Reid2019_ApJ_885_131_LocalArmFit_MW3D[HDU1]
- Reid2019_ApJ_885_131_SGNArmFit_MW3D[HDU1]
- Reid2019_ApJ_885_131_MW3D[HDU1]
- Hunt2023_arXiv_2303.13424_MW3D[HDU1]
- Lallement2019_A+A_625_A135_Split_MW3D[HDU1]
- Edenhofer_2023_3D_Dust_XYZ_Revised-2

Subsets

Plot Layers - 3D Volume Rendering

- [x] Edenhofer_2023_3D_Dust_XYZ_Revised-2
- [x] Zucker2021_ApJ_919_35_spines_MW3D[HDU1]
- [x] Bialy2021_ApJL_919_L5_MW3D[HDU1]
- [x] Pelgrims2020_A+A_636_A17_lmax10_MW3D[HDU1]
- [x] Alves2020_Nat_578_237_MW3D[HDU1]
- [x] Leike2020_A+A_639_A138_xyz_cube_MW3D

Attribute: PRIMARY

Limits: 5.89626e-08 | 0.00643762

Color:

Plot Options - 3D Volume Rendering

x axis Pixel Axis 2 [x]

min/max: -73.9532 ⇄ 812.953

stretch: 1.00

y axis Pixel Axis 1 [y]

min/max: -73.9532 ⇄ 812.953

stretch: 1.00

z axis Pixel Axis 0 [z]

min/max: -54.101 ⇄ 593.101

stretch: 1.00

reference: Leike2020_A+A_639_A138_xyz_cube_MW3D

resolution: 256

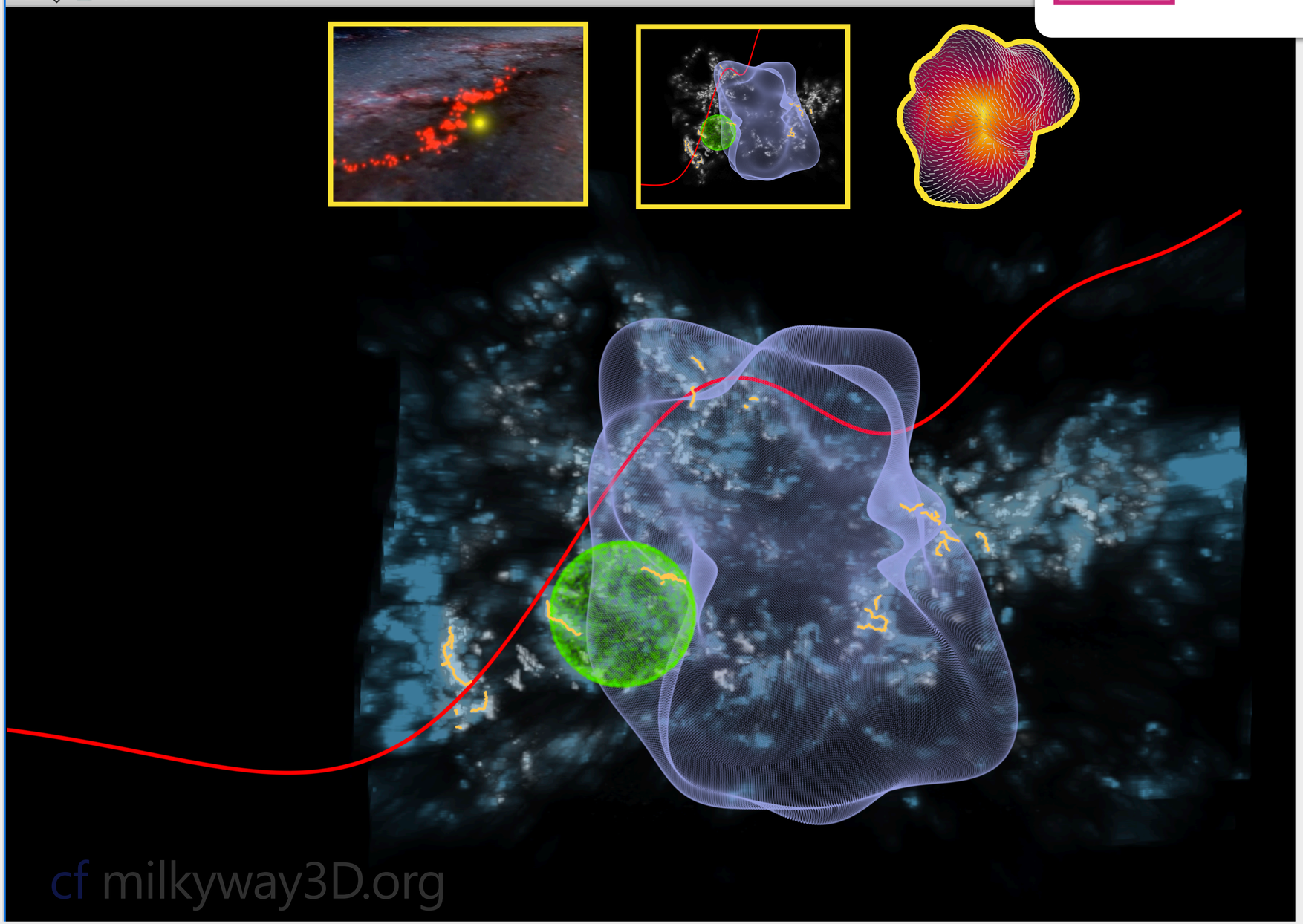
- [x] Native aspect ratio
- [ ] Perspective
- [ ] Downsample when panning

Line Width 1

- [ ] Show axes

Tab 1 | Tab 2 | Tab 3

3D Volume Rendering



glue
multidimensional data exploration

LIVE

MILKYWAY3D.org

LIVE Environments | About LIVE | Capabilities | Tech | Who are we? | Learn More

Linkable Interactive Visualization Exploration (LIVE) Environments

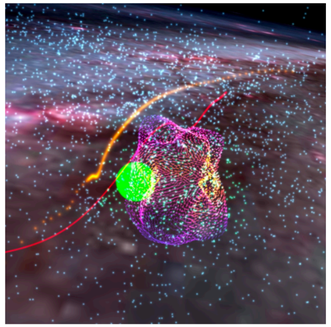
glue

What is LIVE?

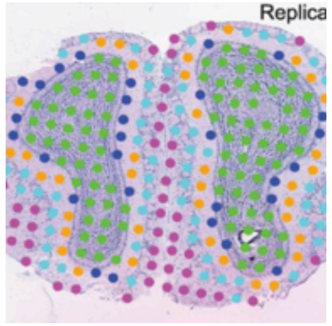
LIVE lets anyone build
"Linkable Interactive Visualization and Exploration" Environments.

LIVE is free, open-source, and helps with shared data and visualization challenges across astronomy (LIVE Astro), biology (LIVE Bio) and GIS (LIVE GIS).

As LIVE's infrastructure is being built, collaborators are ensuring its utility across Astronomy, Biology, and GIS by pursuing LIVE's science demonstration projects.



LIVE Astro



LIVE Bio

LIVE GIS

MilkyWay3D.org | Home | Data | Software | Science | Gallery | Events | Collaborators | Vienna Site (Private)

Welcome to a new view of the Milky Way... in 3D!

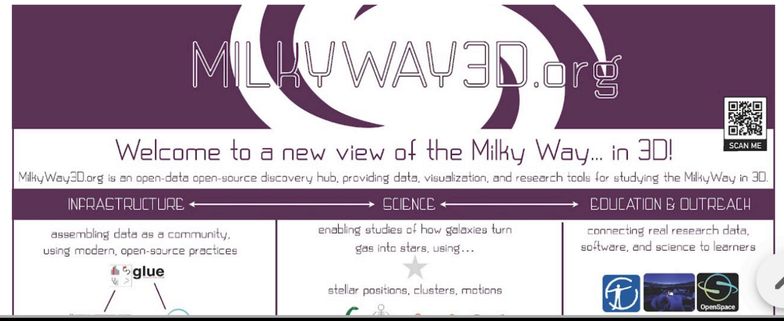
With community input, milkyway3d.org will serve as a hub for the interconnected set of outreach, education, and research resources that allow astronomers to map out and understand the three-dimensional structure of the Milky Way (especially near the Sun).

The project includes new software development, approaches to data sharing, and scientific research questions propelling our collaboration forward. In fact, MilkyWay3D's tech is so innovative, that it's the "demonstration project" for LIVE-Astro, the astronomy portion of the "Linkable, Interactive, Visualization, & Exploration" or "LIVE" Environments project.

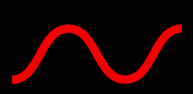
All data sets and software collected, connected, and created will be available to all--including educators and learners!

Debut in Pasadena

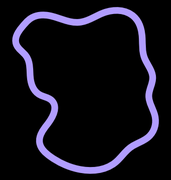
The first time the MilkyWay3D.org was presented to the astronomy research community was a the Surveying the Milky Way: The Universe in Our Own Backyard, in the poster shown here, and in talks by Science PI Catherine Zucker and Harvard graduate student Theo O'Neill.



5 years ago, no one knew that…

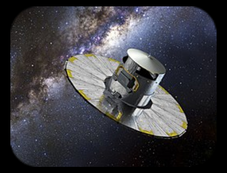


There is a giant (2.7 kpc-long) "wave" of star-forming regions (Alves et al. 2020, named "The Radcliffe Wave") waving up & down (P=90 Myr; Konietzka et al. 2024) along the "Local Arm" of the Milky Way, the Sun passed through it ~13 Myr ago (Maconi et al. 2024*); and a magnetic field runs along it (Panopolou 2024*).



Nearly all star forming regions within 250 pc of the Sun lie on the "surface" of the Local Bubble (Zucker et al. 2022), which is actually a Local Chimney (O'Neill et al. 2024*), and we can estimate what its B-field looks like, in 3D (O'Neill et al. 2024*).

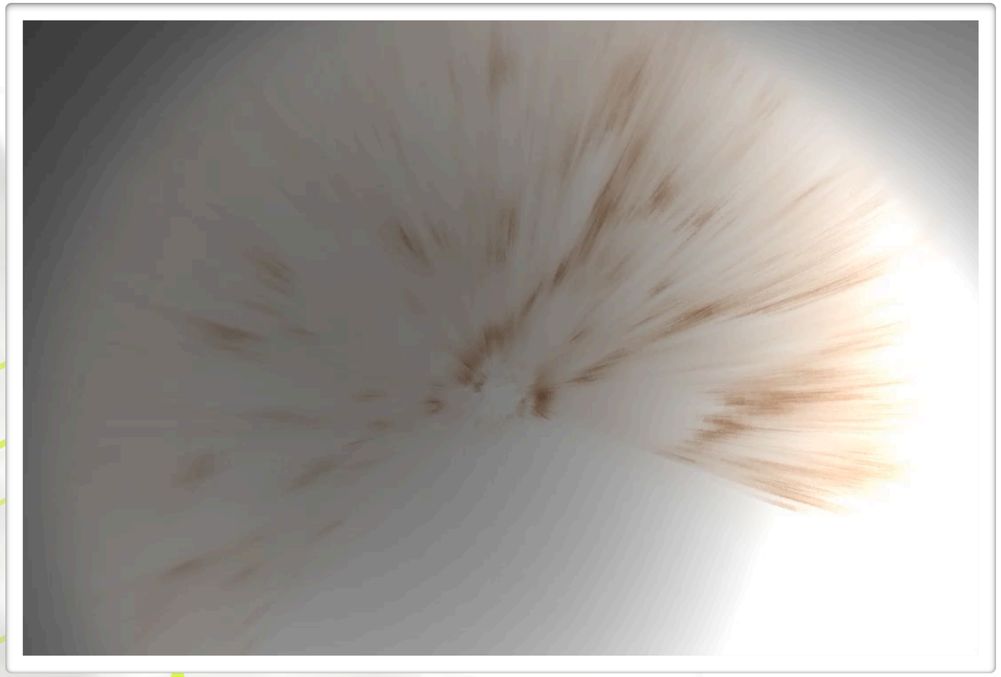
The Perseus and Taurus star-forming regions lie on opposite sides of a "shell," with Perseus in the Radcliffe Wave, and Taurus smooshed between the Local Bubble & the "PerTau Shell" (Bialy et al. 2021).



"Most nearby young star clusters formed in three massive complexes" (Swiggum et al. 2024*).

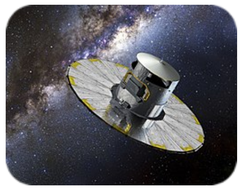
**submitted, embargoed, or in prep, ask for details*

Extinction & Reddening, from Color Imaging



Green et al. 2019

Can infer matter's distance from *dust's* effects on stars.



WARNING: schematic diagram, NOT to scale (credit A. Goodman, 2019)

Abstraction: Why SO bad? →

Why can't LLMs make Informative Infographics in Astronomy?

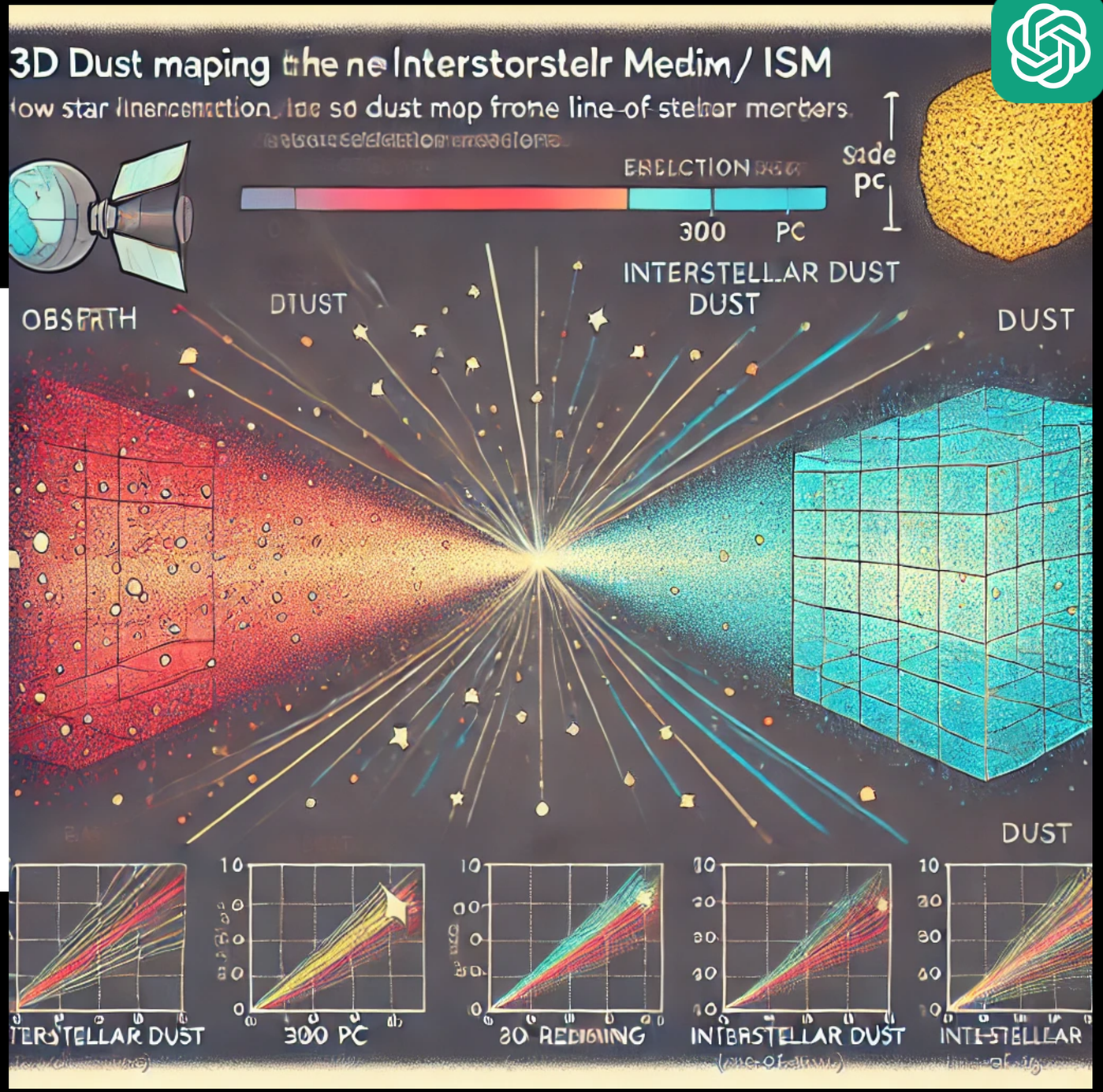
Bruna Biz

Harvard University

ABSTRACT

This work is designed to gauge the limitations of generative AI models such as ChatGPT and DALL-E in their ability to accurately produce astrophysical information graphics. We use the Hertzsprung-Russell (H-R) diagram – a fundamental tool in astronomy that maps the relationship between stellar luminosity and temperature – as a test case because of its simple 2D structure and its importance in stellar astronomy. Initial attempts to produce accurate H-R diagrams with ChatgPT led to interesting results which demonstrate the models' tendency to perceive and illustrate cycles (in this case, the stellar life cycle) as circles, resulting in images that deviate from the traditional linear structure of H-R diagrams. Despite high quality H-R diagrams being fed and described to the model, these patterns persist, demonstrating DALL-E's inability to diverge from circular patterns in relation to cyclic phenomena. An investigation into DALL-E's diffusion process also suggests that a big challenge with astrophysical infographic generation is rooted in how tokens are embedded and processed within latent space – an abstract, higher dimensional space where models abstract their vector embeddings. This points to potential improvements that could be made in astrophysical infographic generation to produce more accurate visuals.

infographics



nature

Science in culture

Books & arts

Will AI jeopardize science photography?

Generative artificial-intelligence illustrations can be helpful, but fall short as scientific records. **By Felice Frankel**

THREE VIEWS

Decisions made during the imaging process result in distinct representations of vials containing fluorescing nanocrystals. The photographs of the real vials contrast sharply with a visual created by prompting an artificial intelligence (AI) program.

FELICE FRANKEL

One of the privileges of being on the campus of the Massachusetts Institute of Technology (MIT) in Cambridge is seeing glimpses of the future, from advances in quantum computing and energy sustainability and production, to designing new antibiotics. Do I understand it all deeply? No, but I am able to wrap my head around much of it when I am asked to create an image to document the research.

The joy of being a science photographer is that I must learn about the things I am documenting to produce communicative and trustworthy images, intended as a form of data, for the researchers who welcome me into their laboratories.

But now, with the wide availability of generative artificial intelligence (genAI) tools, lots of questions must be asked. Will there be a point at which, with just a few keystrokes and prompts, a scientist can create a 'visual' of their research, as I do with my camera, and consider that image a record of the work? Will research-

FELICE FRANKEL

Iron particles in an oil drop on a glass slide respond to magnets placed underneath.

graphics

Essay by MIT's Felice Frankel, Nature, February 2025

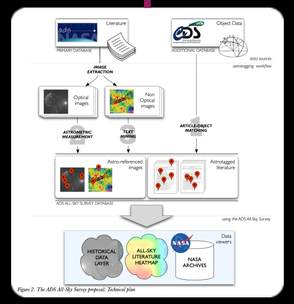
Thematic questions for discussion

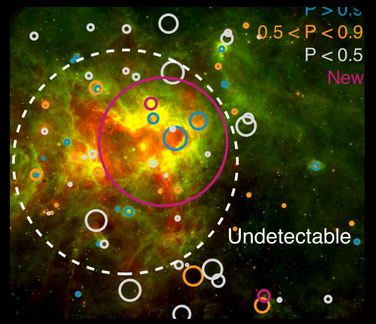
1. visualization and infographics — how good is AI at abstraction, and how good will it become?
2. If we give “all” the MilkyWay3D.org data to AI, what (kinds of) new physical insights might we expect—should we have expectations *a priori*—or is that approach too restrictive? (Note, 3D dust also AI!)

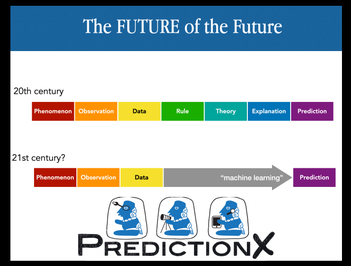
Please feel free to ask me about these. . .

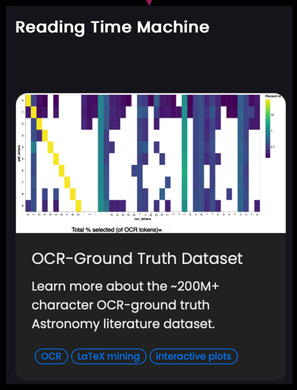
"Reading Time Machine"

code for visualization

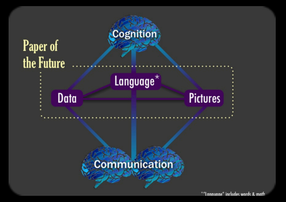


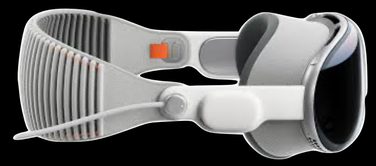






2012 — 2014 — 2016 — 2018 — 2020 — 2022 — 2023 — 2024





data-set linking

3D selection

infographics

extra slides

How 3D dust mapping does/does not use AI

Method	Technique	AI Involvement?	Key strength
Bayestar	Bayesian inference (classic)	❌ No	Statistical rigor; simplicity
Brutus/Augustus	Advanced Bayesian inference	❌ No	Enhanced stellar modeling; precise handling of uncertainties
IFT (Enßlin)	Bayesian inference + AI	✅ Yes	Learned priors; better efficiency
Gaussian splatting	Neural rendering (AI-enhanced)	✅ Yes	Real-time visualization; realism

The Prediction Project

PredictionX: The Past & Present of the Future

Essentials

Predictive Systems Framework

The "Padua Rainbow"

Understanding Uncertainty

Study Design

Timelines

▶ **Why predict?**

Omens, Oracles & Prophecies

Mesopotamian Haruspicy
Egyptian Priests
Yoruba Ifa
Tarot
Roman Augury
Casting Lots
The Diviner's Guide
Chinese Oracle Bones
Greek Astronomy
Oracle of Delphi
Turkish Tasseography
Astrology
Aztec Rituals
Maya Spacetime
Comets of Doom

cross-cultural conversations

The Rise of Theory

Ancient Mesopotamia, Egypt, Greece & Rome
Islamic Science

The Path to Newton

▶ *Indian Mathematics*
European Renaissance

Lost without Longitude (Navigation)

Help, I'm Lost!
Tools of the Navigator

The Royal Society

Modern Prediction

John Snow & Cholera
Cholera Map

Health
▶ *Epidemiology*
▶ *Personal Genomics*
▶ *Population Genetics*

Wealth
▶ *Climate & Wealth*
▶ *Behavioral Economics*

The Future of the Future
▶ *AI, Derek's Day*
▶ *Philosophy*
▶ *Uncertainty*

Earth
▶ *Climate & Energy*
▶ *Climate Policy*
▶ *Tent Tarot*
▶ *Earthquakes*

Space
▶ *Futures of our Universe*
▶ *SETI*

Interactive Resource ▶ video(s) Coming Soon

visit predictionx.org for more information on the Prediction Project

edX HarvardX predictionx.org

Harvard College Program in General Education
Explore. Expand. Engage.

LabXchange™ Science and learning-connected.

NVIDIA
“Imagine a world where we can predict the future...”

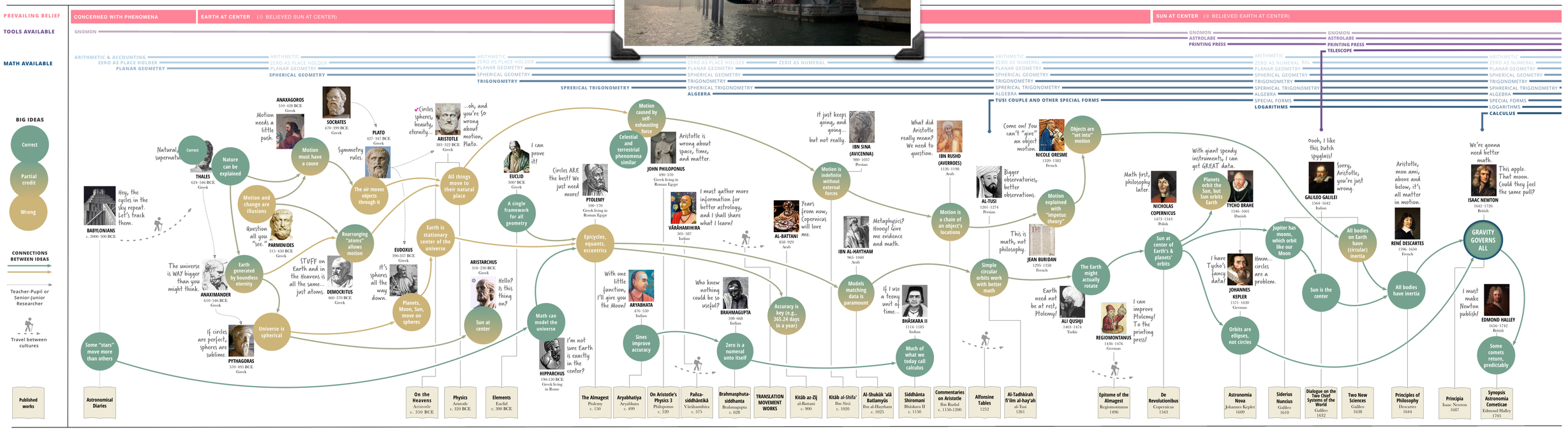
The “Padua Rainbow” & The Path to Newton

PREDICTIVE SYSTEMS

Phenomenon | Observation* | Data | Theory | Explanation | Prediction

The Path to Newton

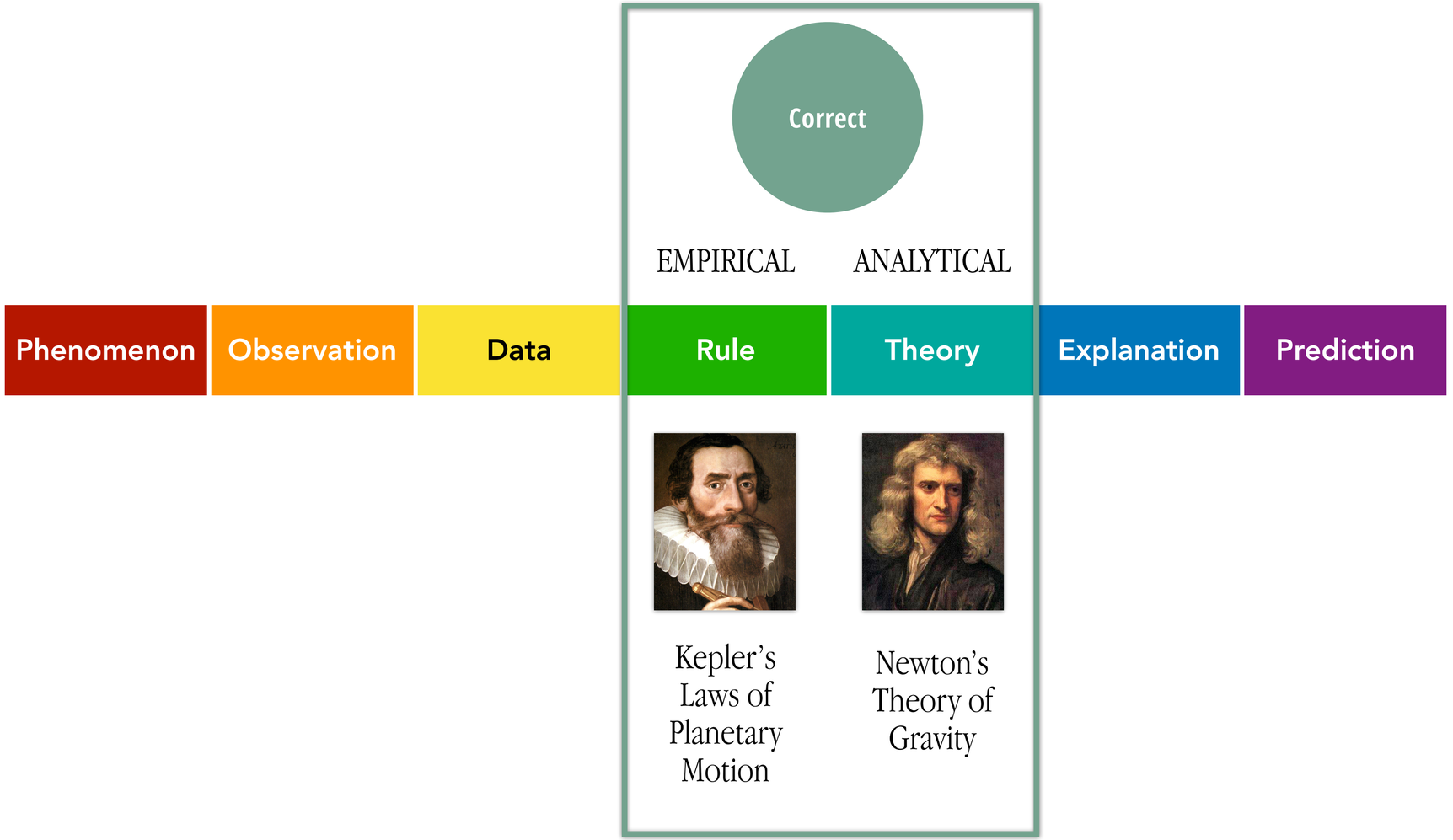
PREDICTIONX



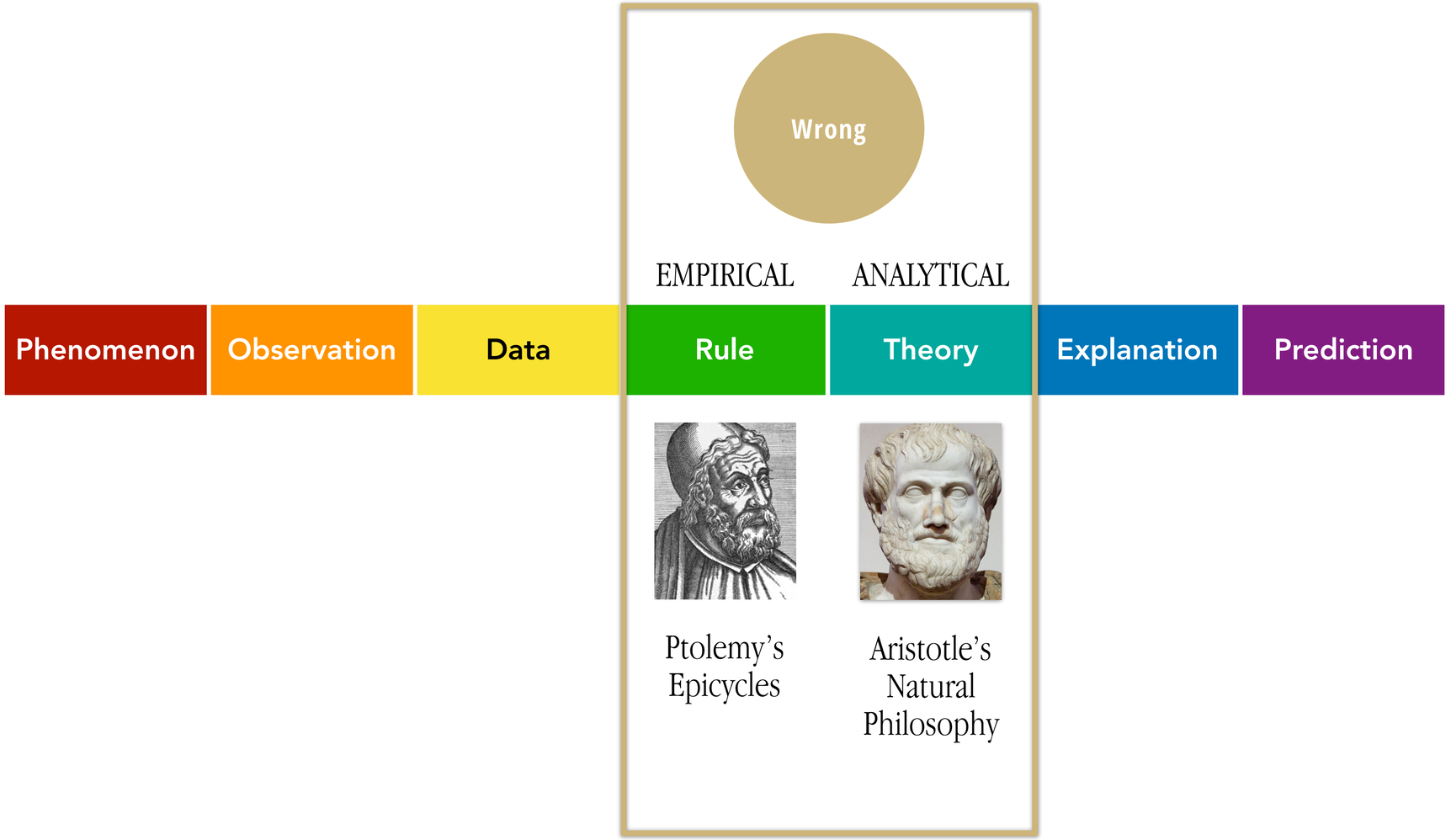
© Harvard University, created by Alyssa Goodman, Jais Brohinsky, Drew Lichtenstein & Katie Peek, re-use is allowed, with attribution, version 1, 2019

or, Experiment

The “Padua Rainbow” & The Path to Newton



The “Padua Rainbow” & The Path to Newton



Mendel

Phenomenon | Observation | Data | Rule | Prediction

Darwin

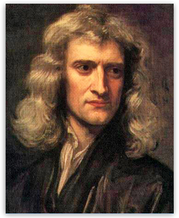
Phenomenon | Observation | Data | Theory | Explanation | Prediction

BIOLOGY

PHYSICS

Kepler

Phenomenon | Observation | Data | Rule | Prediction



Newton

Phenomenon | Observation | Data | Rule | Theory | Explanation | Prediction

Mendel *1865*

Phenomenon | Observation | Data | Rule | Prediction

Darwin *1859*

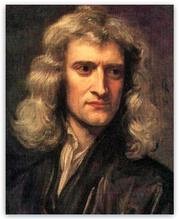
Phenomenon | Observation | Data | Theory | Explanation | Prediction

NO FULLY PREDICTIVE GENERAL THEORY

FULLY PREDICTIVE GENERAL THOERY

Kepler *1609*

Phenomenon | Observation | Data | Rule | Prediction



Newton *1687*

Phenomenon | Observation | Data | Rule | Theory | Explanation | Prediction

The Paths to Science

Path-to-Foundation

path-to-foundation.org

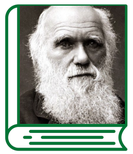
Science is so much more than equations and lab coats. It requires human creativity, ingenuity, argument, and persistence.

The goal of the Paths to Science is to deepen students,' scientists,' and the broader public's appreciation of science by telling stories of its history.

5 small books published by MIT Press, each including a narrative linked to a poster

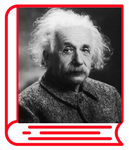
The MIT Press

The Path to Newton



The Path to Darwin

The Path to AI



The Path to Einstein

The Path to Modern Genetics

A growing consortium. . .

The Path-to Science Consortium

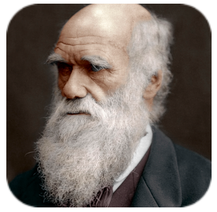
Università degli Studi di Padova

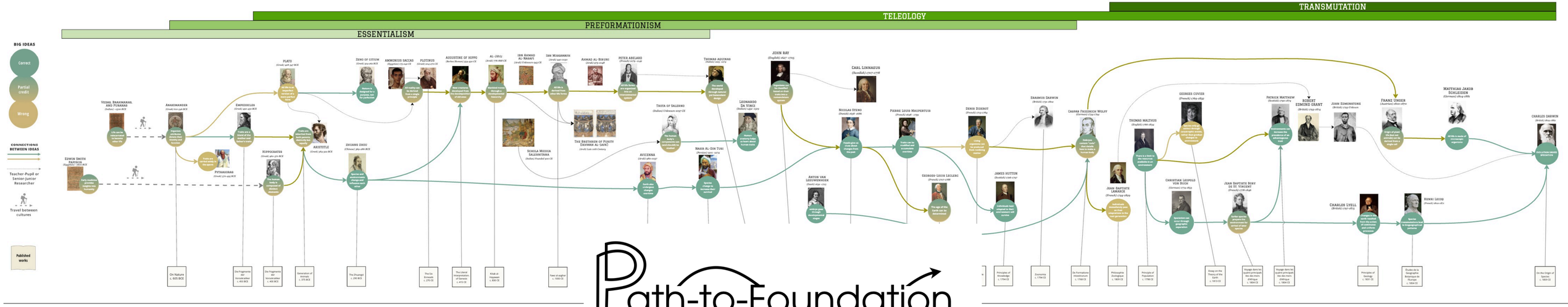
Harvard University

University of Oxford

The Path to Darwin, v. 0.1

Prof. Immaculata De Vivo & Mally Shan (Class of 24)





The Path to AI, v. 0.01

Simeon Sayer, Class of 22

Path to AI Threads - Data

Papyrus Scrolls (3000 BCE)
Clay Tablets (2000 BCE)
Parchment and Vellum (200 BCE)
Invention of Paper (105 CE)
Printing Press (1440)
Card Catalogs (late 19th century)
Magnetic Tape (1928)
Hard Disk Drives (1956)
Floppy Disks (1971) Compact Disc (CD) (1982) Solid State Drives (SSDs) (2009) Cloud Storage (2006 - present)

Path to AI Threads – Human History

Artificial beings created by Gods (1600 BCE)
Designing mechanical humanoid figures (8th - 14th century)
Considering the design for automotive machines (1400)
"beginning of science-fiction (1800s)"
AI is featured in films (1950s - present)
AI is shown to be dangerous in films (1950s - present)

Path to AI Threads - Algorithms

Birth of graph theory (1736)
Boolean Algebra (1847)
Evolution of graph theory (1845 - 1990s)
Game Theory (1928-1944)
Linear Threshold Unit Model (1943)
Prediction Calculus (1956)
Logic Theorist (1956) Decision trees (1970s)
Transformer Paper (2018)

The FUTURE of the Future

20th century