AI in the Paleogeociences: Progress, Challenges, and Opportunities

Deborah Khider
Information Sciences Institute

USC Viterbi
School of Engineering
Information Sciences Institute

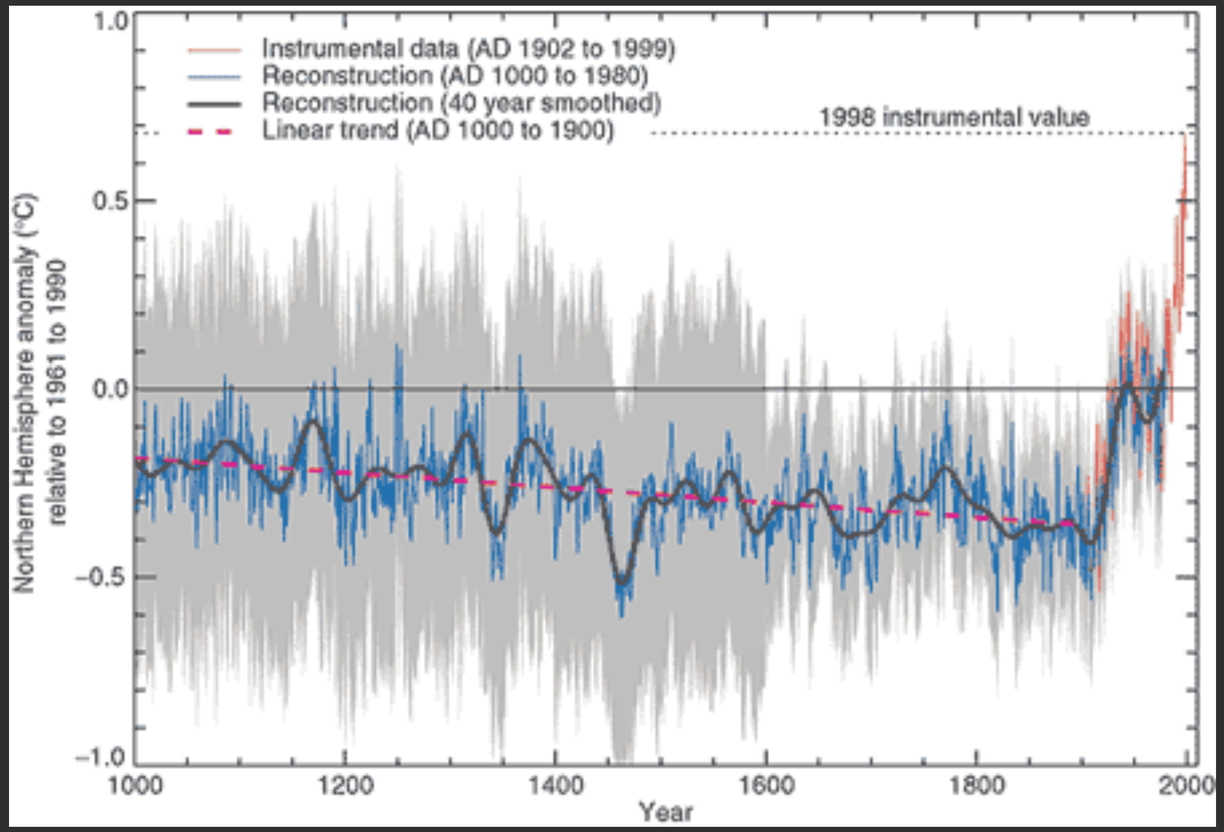
Why study paleoclimatology?

To address questions such as:

- Has climate changed in the past?
- How much?
- How often?
- How fast?
- Why?

To provide test datasets for the verification of climate models

The Hockey Stick



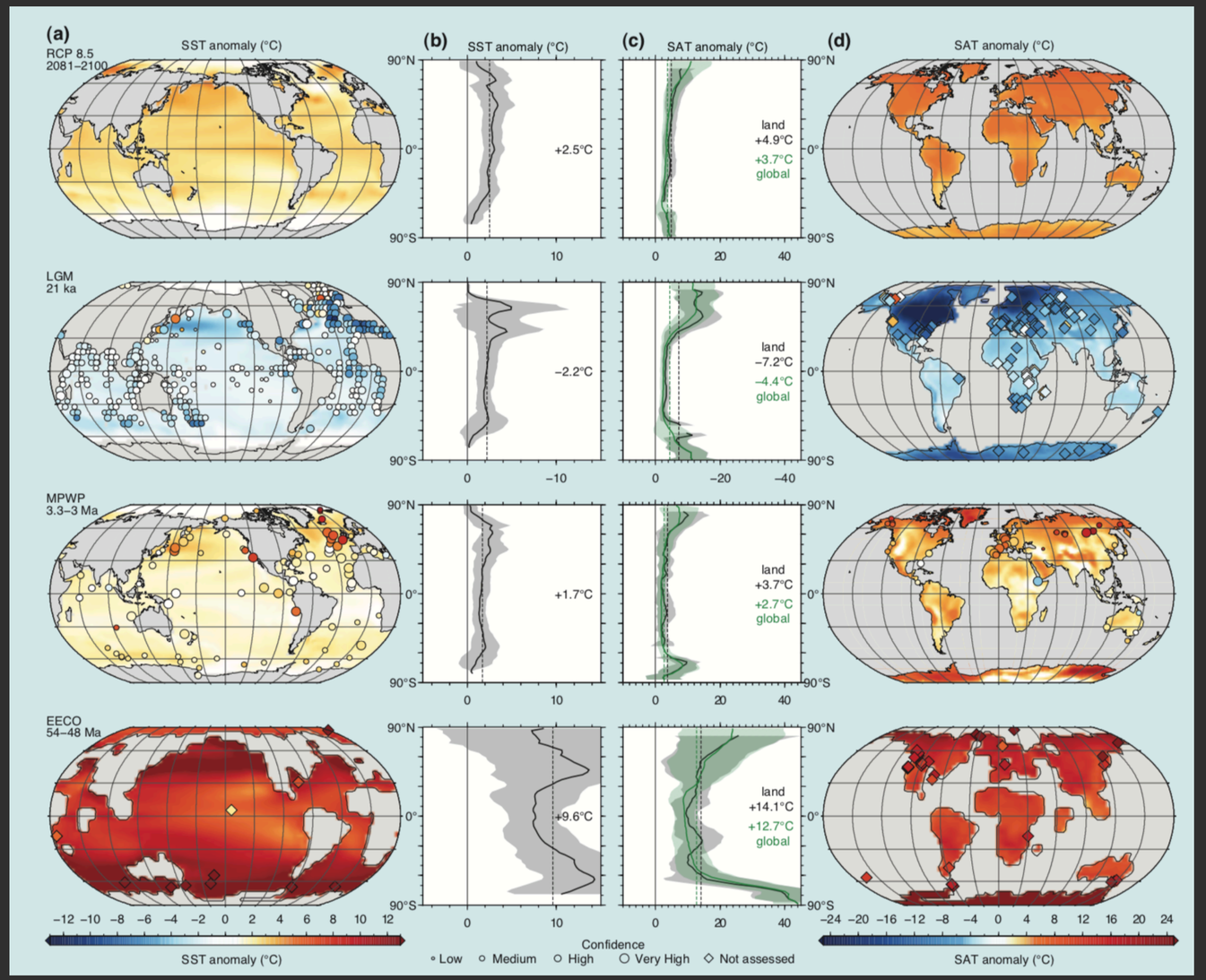
Mann et al. , 1999, Geophysical Research Letters

Why study paleoclimatology?

To address questions such as:

- Has climate changed in the past?
- How much?
- How often?
- How fast?
- Why?

To provide test datasets for the verification of climate models



IPCC AR5, Paleoclimate, 2015

How do we answer these questions?

Past climate quantities are not measured directly

Evans et al., 2013, Quaternary Science Reviews

How do we answer these questions?

Past climate quantities are not measured directly

Proxy Archive

Marine Sediment

Evans et al., 2013, Quaternary Science Reviews

How do we answer these questions?

Past climate quantities are not measured directly

Proxy Archive **Proxy Observation**

Marine Sediment Mg/Ca

Evans et al., 2013, Quaternary Science Reviews

How do we answer these questions?

Past climate quantities are not measured directly

Proxy Archive **Proxy Observation** **Environment**

Marine Sediment Mg/Ca Surface Ocean

Evans et al., 2013, Quaternary Science Reviews

How do we answer these questions?

Past climate quantities are not measured directly

Proxy Archive | **Proxy Observation** | **Proxy Sensor** | **Environment**

Marine Sediment

Mg/Ca

Foraminifera
Globigerinoides ruber

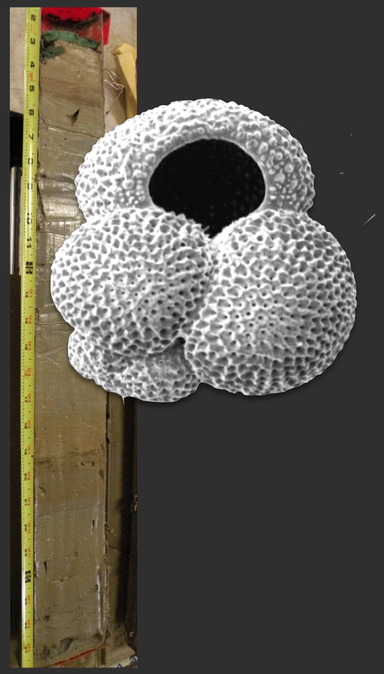
Surface Ocean

Evans et al., 2013, Quaternary Science Reviews

We have multitude of proxy systems!

Leng et al., 2006, Isotopes in Paleoenvironmental Research

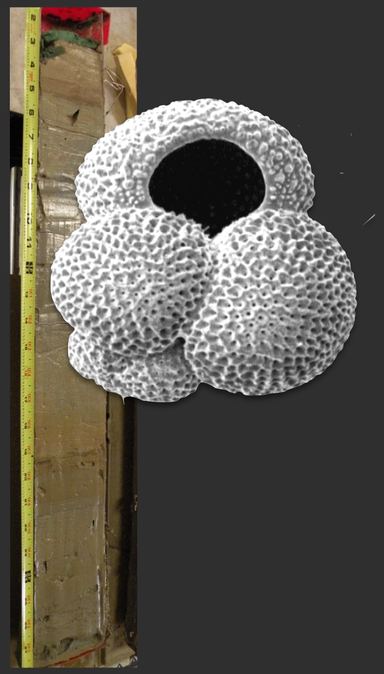
We have multitude of proxy systems!

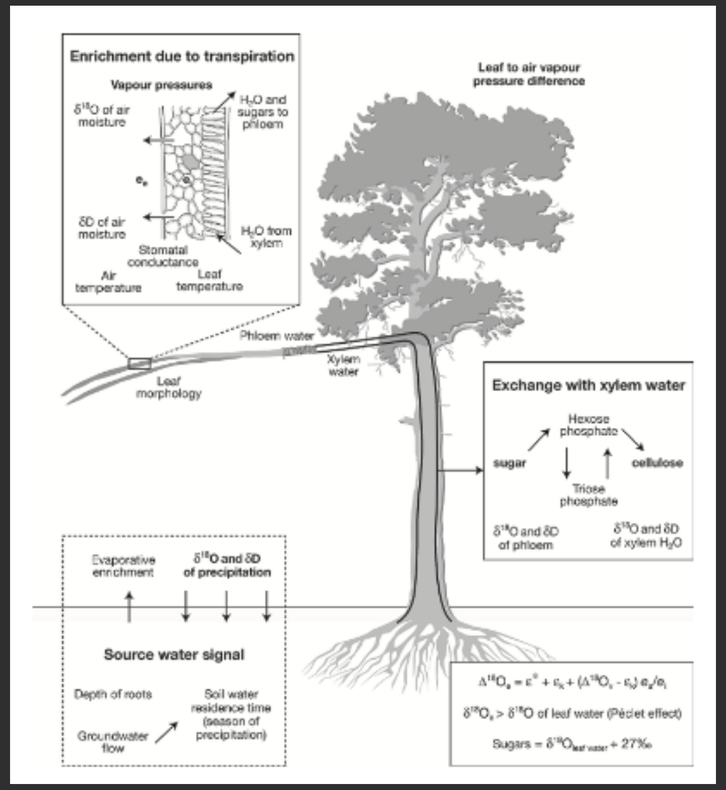


Archive: Marine Sediment
Sensor: Foraminifera
Observation: Mg/Ca

Leng et al., 2006, Isotopes in Paleoenvironmental Research

We have multitude of proxy systems!



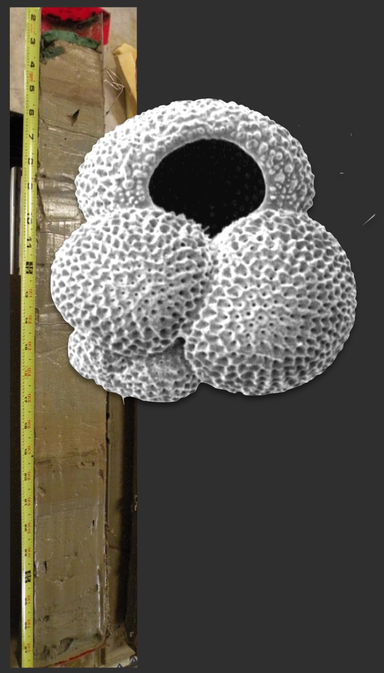


Archive: Marine Sediment
Sensor: Foraminifera
Observation: Mg/Ca

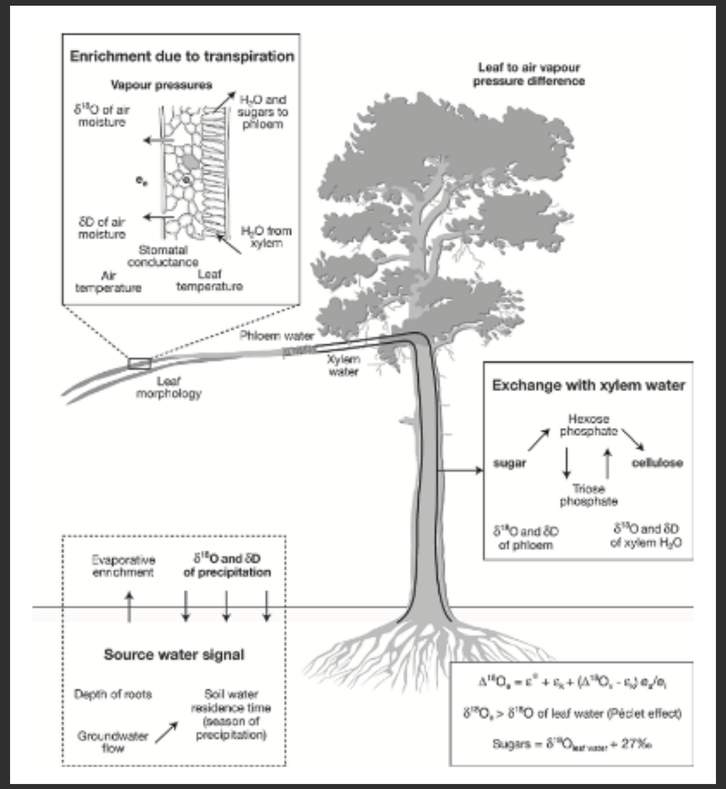
Archive: Wood
Sensor: Tree
Observation: $\delta^{18}O$

Leng et al., 2006, Isotopes in Paleoenvironmental Research

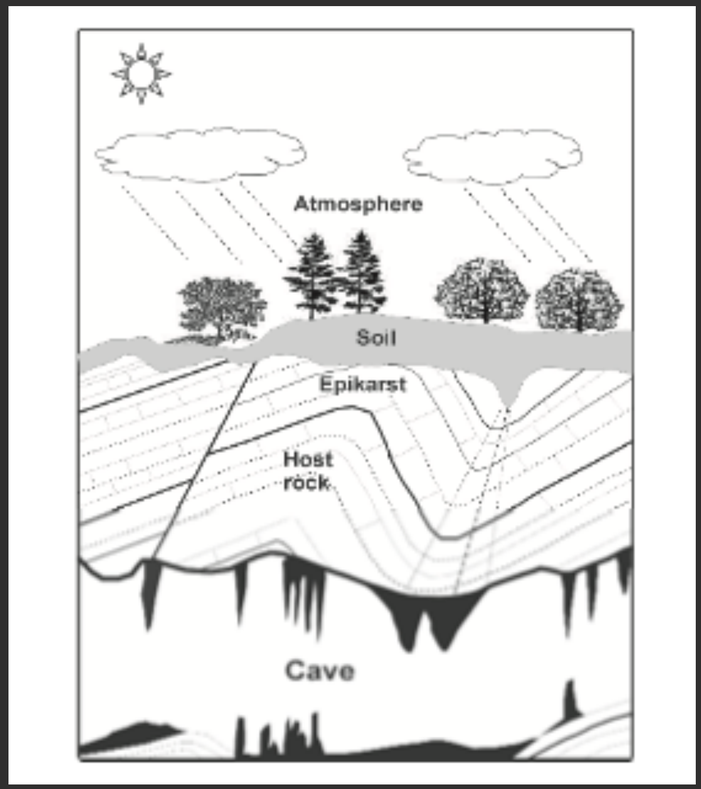
We have multitude of proxy systems!



Archive: Marine Sediment
Sensor: Foraminifera
Observation: Mg/Ca



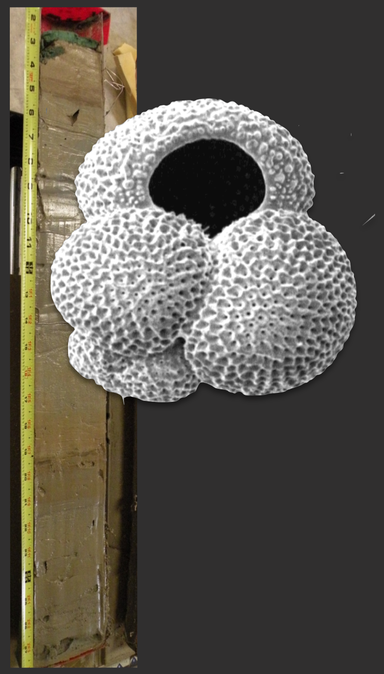
Archive: Wood
Sensor: Tree
Observation: $\delta^{18}O$



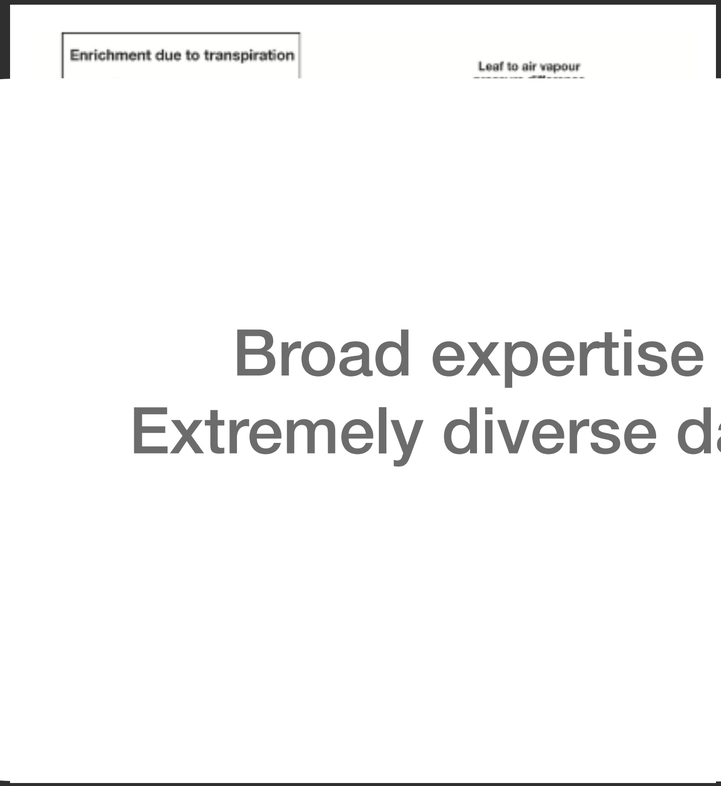
Archive: Speleothem
Sensor: Karst
Observation: $\delta^{18}O$

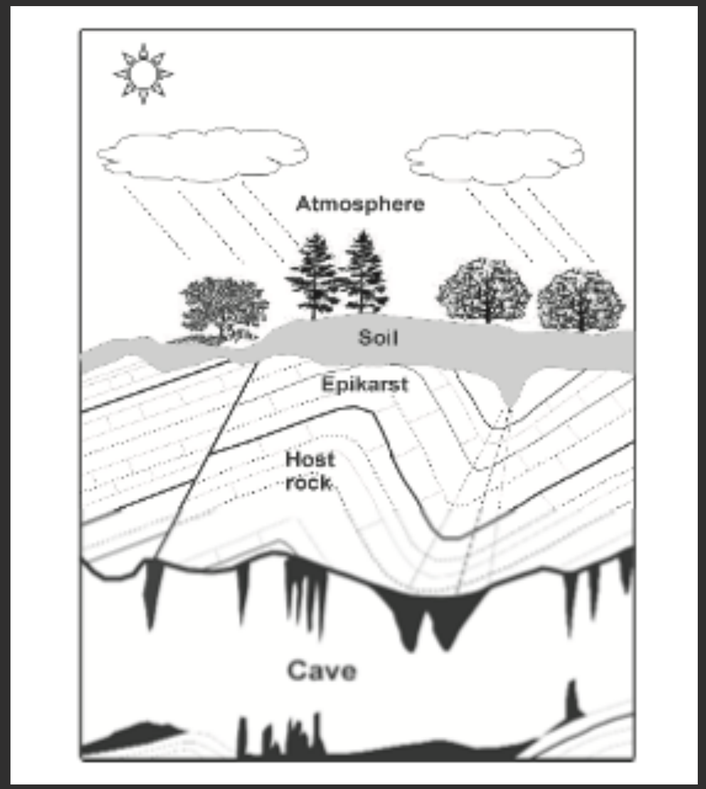
Leng et al., 2006, Isotopes in Paleoenvironmental Research

We have multitude of proxy systems!



1. Broad expertise
2. Extremely diverse data





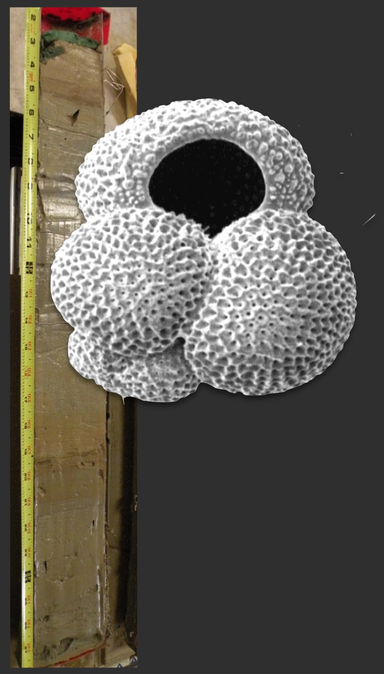
Archive: Marine Sediment
Sensor: Foraminifera
Observation: Mg/Ca

Archive: Wood
Sensor: Tree
Observation: $\delta^{18}O$

Archive: Speleothem
Sensor: Karst
Observation: $\delta^{18}O$

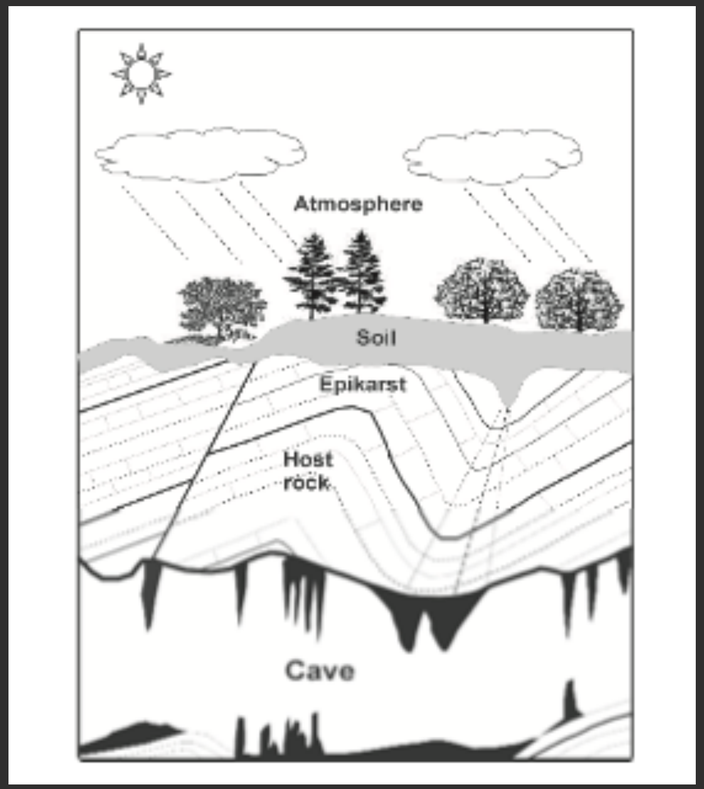
Leng et al., 2006, Isotopes in Paleoenvironmental Research

We have multitude of proxy systems!



1. Broad expertise
2. Extremely diverse data

Long-tail approach to data collection: each record is provided by an individual investigator.



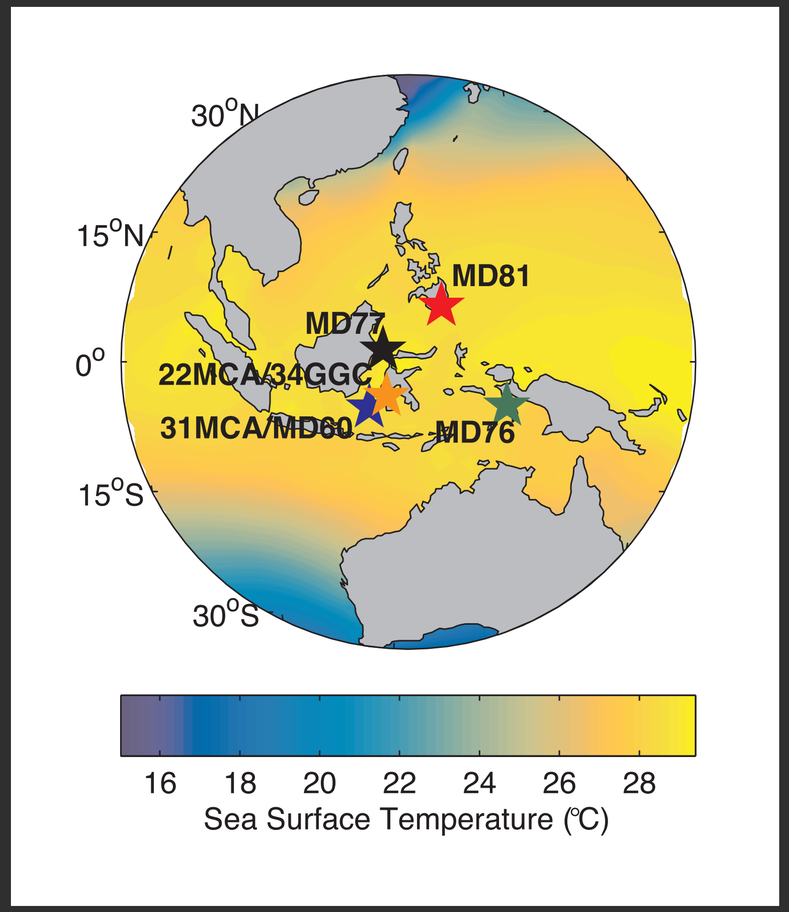
Archive: Marine Sediment
Sensor: Foraminifera
Observation: Mg/Ca

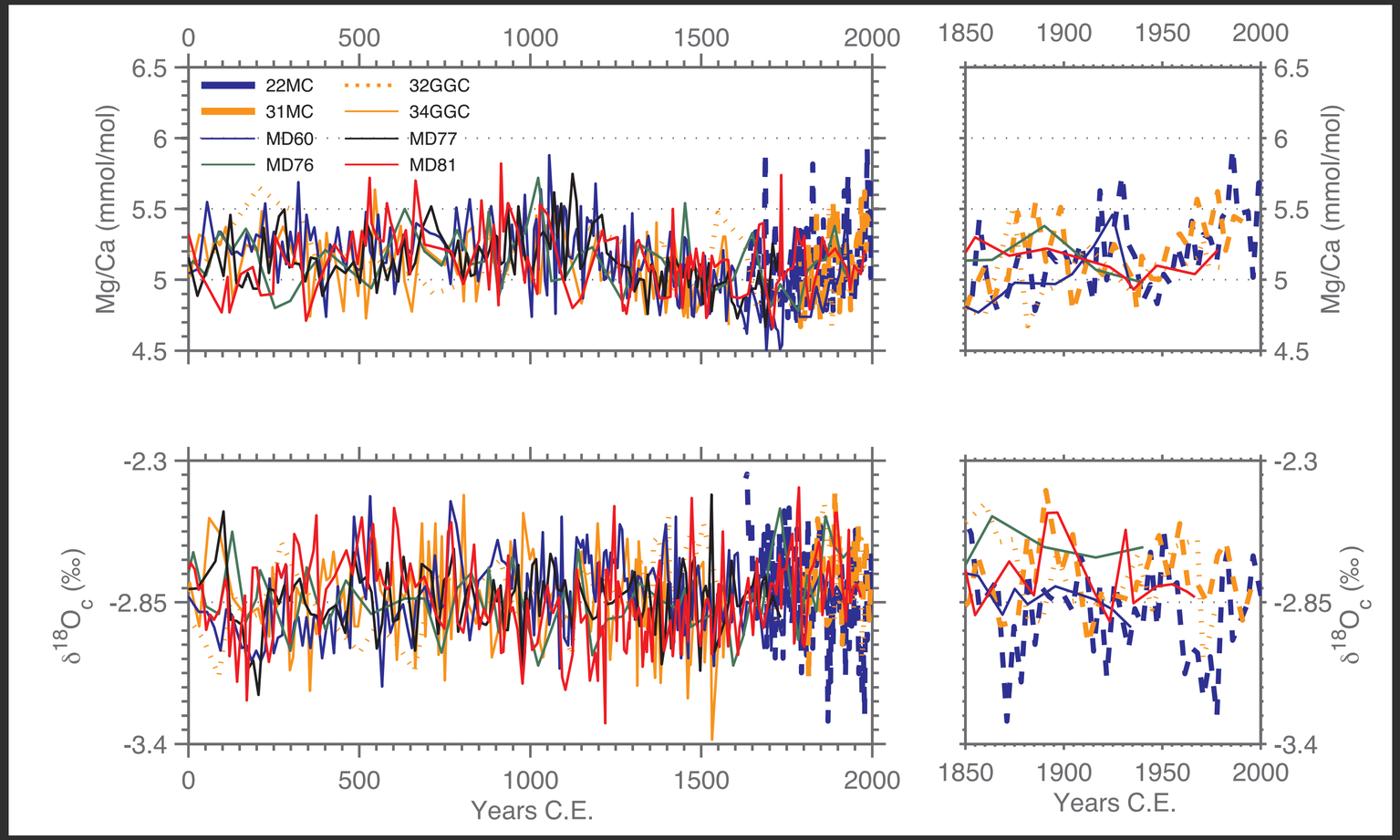
Archive: Wood
Sensor: Tree
Observation: $\delta^{18}O$

Archive: Speleothem
Sensor: Karst
Observation: $\delta^{18}O$

Leng et al., 2006, Isotopes in Paleoenvironmental Research

They offer a picture of the climate at one location through time





Khider et al., 2013, Workshop on Climate Informatics

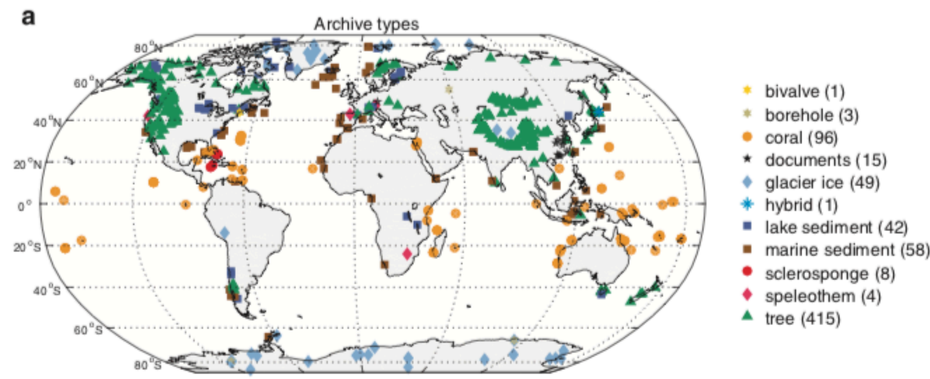


Figure 1. Spatiotemporal data availability in the PAGES2k database. (**a**) Geographical distribution, by archive type, coded by color and shape. (**b**) Temporal resolution in the PAGES2k database, defined here as the median of the spacing between consecutive observations. Shapes as in (**a**), colors encode the resolution in years (see colorbar). (**c**) Temporal availability, coded by color as in (**a**).

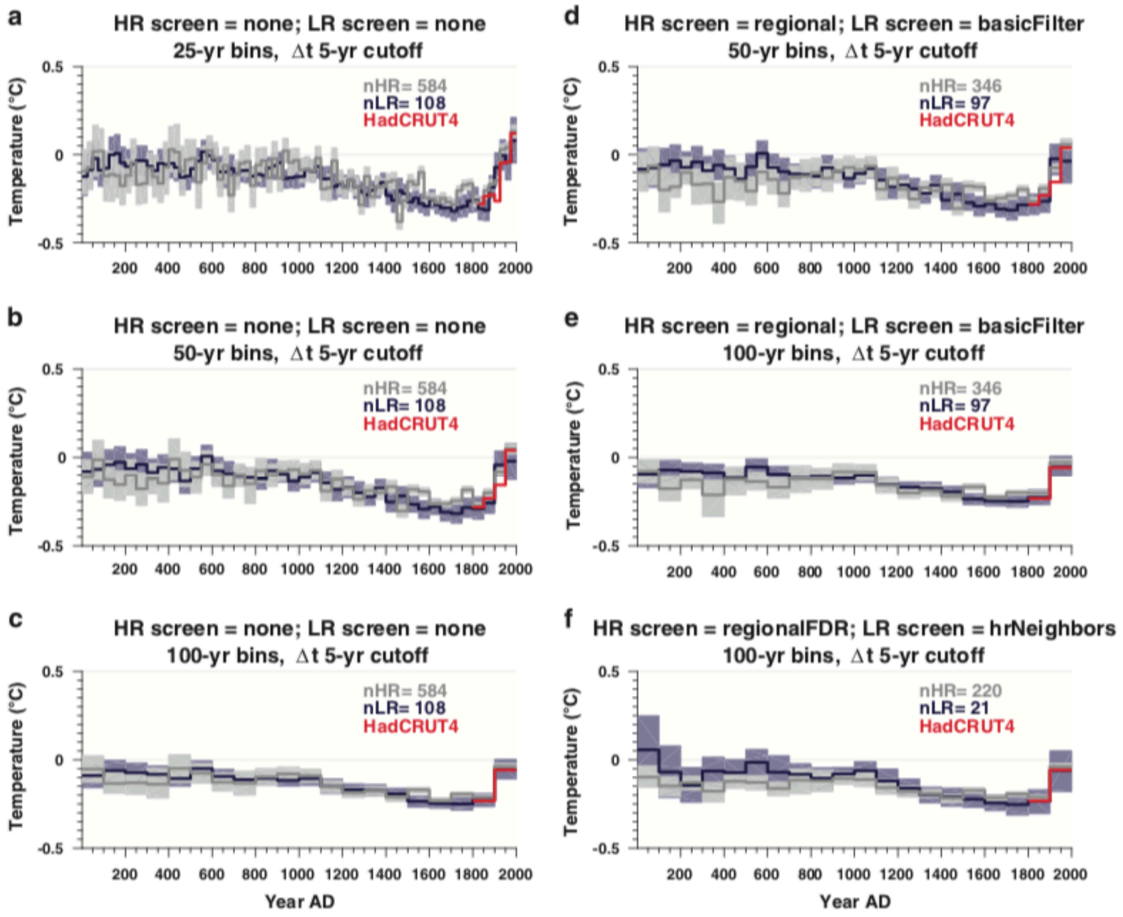


Figure 7. Global composites for various binning intervals and screening criteria, as indicated in subplot titles. The composites are scaled to temperature for comparison, and the shading denotes 95% bootstrap confidence intervals with 500 replicates, to constrain uncertainties. The cutoff between high-resolution (HR) and low-resolution (LR) records is defined as a median resolution of 5 years. Screening options comprise: no screening (*none*), regional temperature screening (*regional*), or regional screening adjusted for the false discovery rate (*regionalFDR*). For low-resolution records, *basicFilter* denotes records that comprise at least 20 values over the Common Era (Supplementary Fig. 2), while *hrNeighbors* denotes records with at least one significantly correlated HR neighbor.

PAGES2k Consortium, 2017, Scientific Data

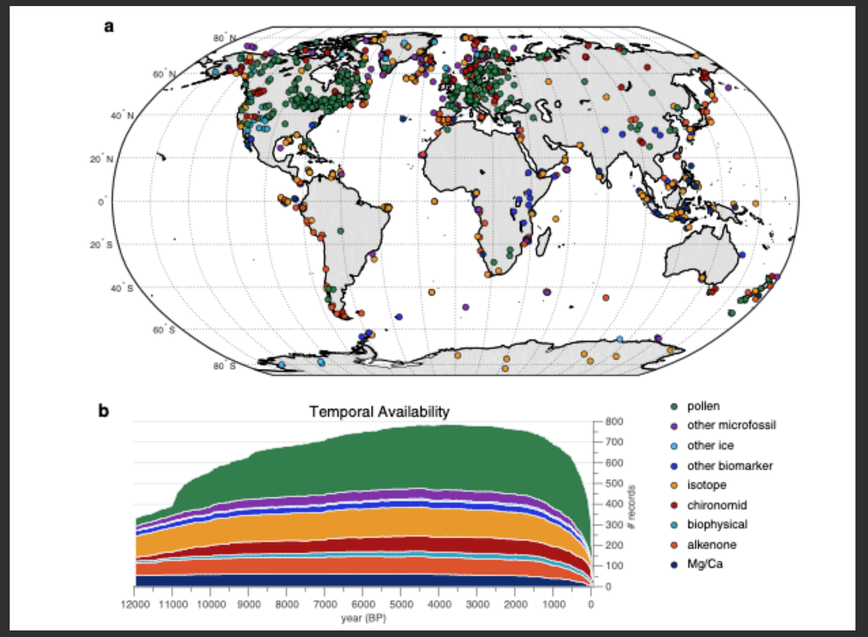
Data Descriptor | Open Access | Published: 14 April 2020

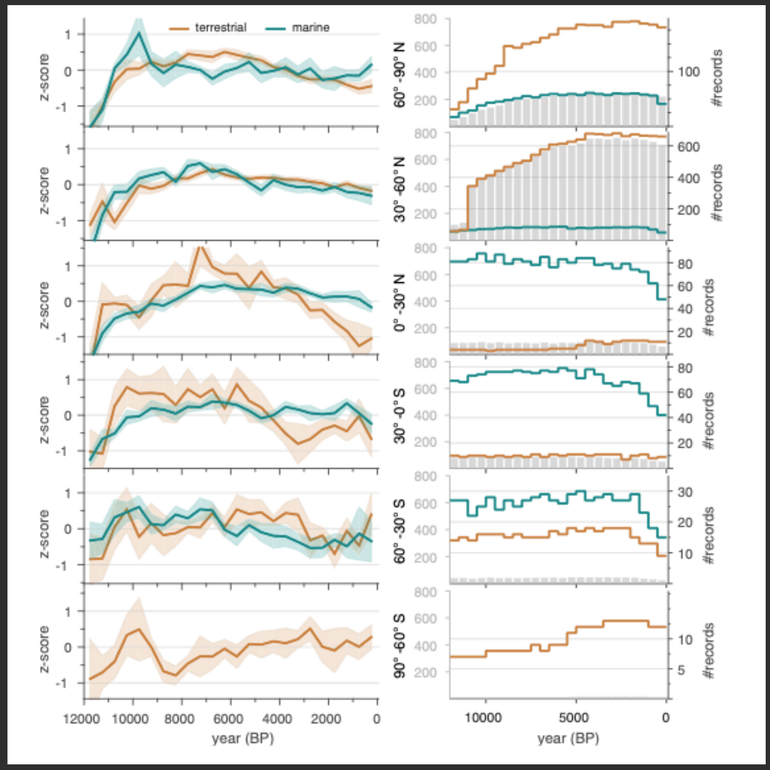
A global database of Holocene paleotemperature records

Darrell Kaufman ✉, Nicholas McKay, Cody Routson, Michael Erb, Basil Davis, Olive Heiri, Samuel Jaccard, Jessica Tierney, Christoph Dätwyler, Yarrow Axford, Thomas Brussel, Olivier Cartapanis, Brian Chase, Andria Dawson, Anne de Vernal, Stefan Engels, Lukas Jonkers, Jeremiah Marsicek, Paola Moffa-Sánchez, Carrie Morrill, Anais Orsi, Kira Rehfeld, Krystyna Saunders, Philipp S. Sommer, Elizabeth Thomas, Marcela Tonello, Mónika Tóth, Richard Vachula, Andrei Andreev, Sebastien Bertrand, Boris Biskaborn, Manuel Bringué, Stephen Brooks, Magaly Caniupán, Manuel Chevalier, Les Cwynar, Julien Emile-Geay, John Fegyveresi, Angelica Feurdean, Walter Finsinger, Marie-Claude Fortin, Louise Foster, Mathew Fox, Konrad Gajewski, Martin Grosjean, Sonja Hausmann, Markus Heinrichs, Naomi Holmes, Boris Ilyashuk, Elena Ilyashuk, Steve Juggins, Deborah Khider, Karin Koinig, Peter Langdon, Isabelle Larocque-Tobler, Jianyong Li, André Lotter, Tomi Luoto, Anson Mackay, Eniko Magyari, Steven Malevich, Bryan Mark, Julieta Massaferro, Vincent Montade, Larisa Nazarova, Elena Novenko, Petr Pařil, Emma Pearson, Matthew Peros, Reinhard Pienitz, Mateusz Płóciennik, David Porinchu, Aaron Potito, Andrew Rees, Scott Reinemann, Stephen Roberts, Nicolas Rolland, Sakari Salonen, Angela Self, Heikki Seppä, Shyhrete Shala, Jeannine-Marie St-Jacques, Barbara Stenni, Liudmila Syrykh, Pol Tarrats, Karen Taylor, Valerie van den Bos, Gaute Velle, Eugene Wahl, Ian Walker, Janet Wilmshurst, Enlou Zhang & Snezhana Zhilich -Show fewer authors

Scientific Data **7**, Article number: 115 (2020) | Cite this article

9846 Accesses | **3** Citations | **353** Altmetric | Metrics





Has paleoclimate fulfilled its promises?

Has paleoclimate fulfilled its promises?

- Mostly yes, especially in the past 10-20 years

Has paleoclimate fulfilled its promises?

- Mostly yes, especially in the past 10-20 years
- Advancements in (geo)informatics have made this possible

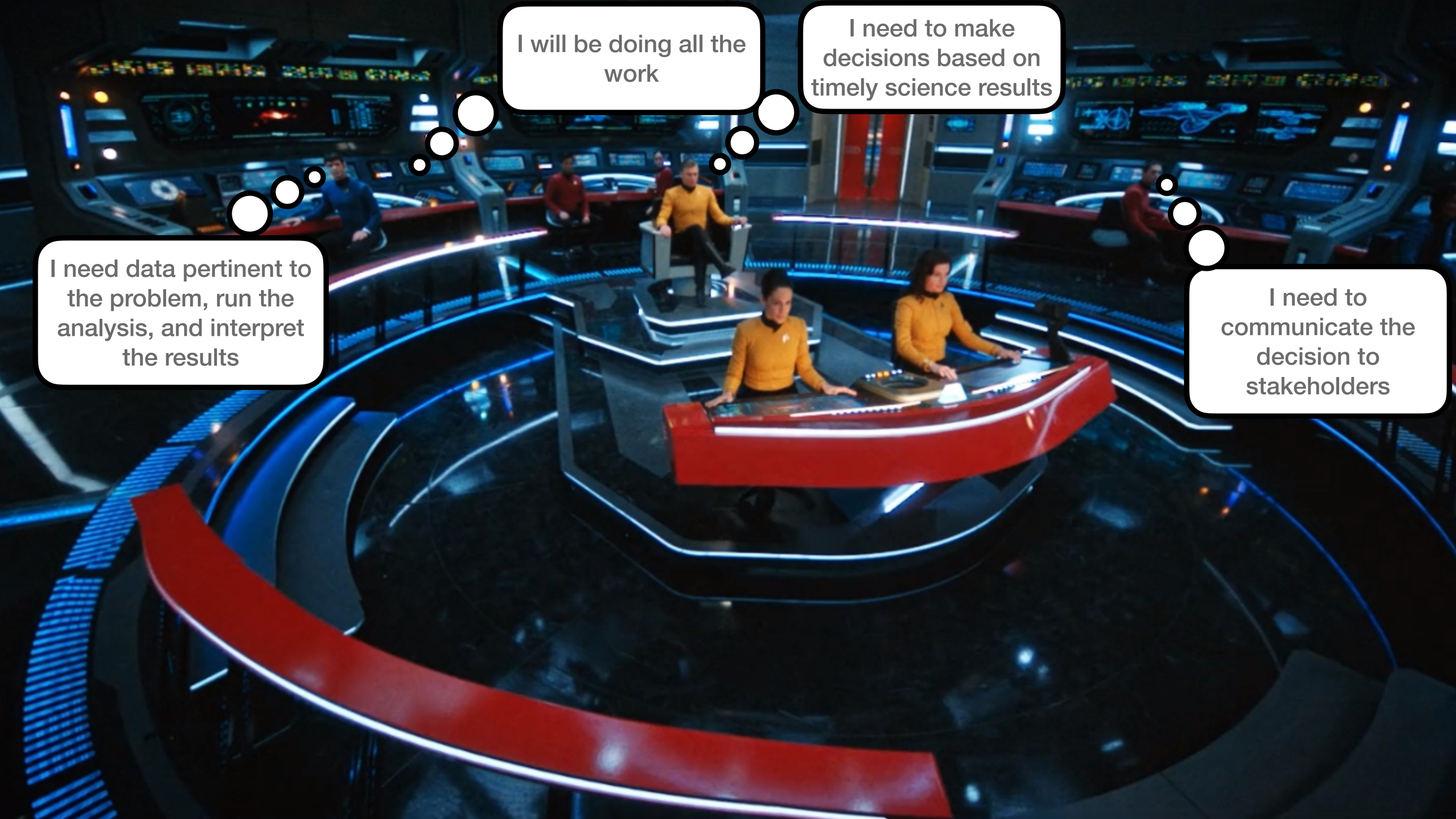
Geosciences in the future

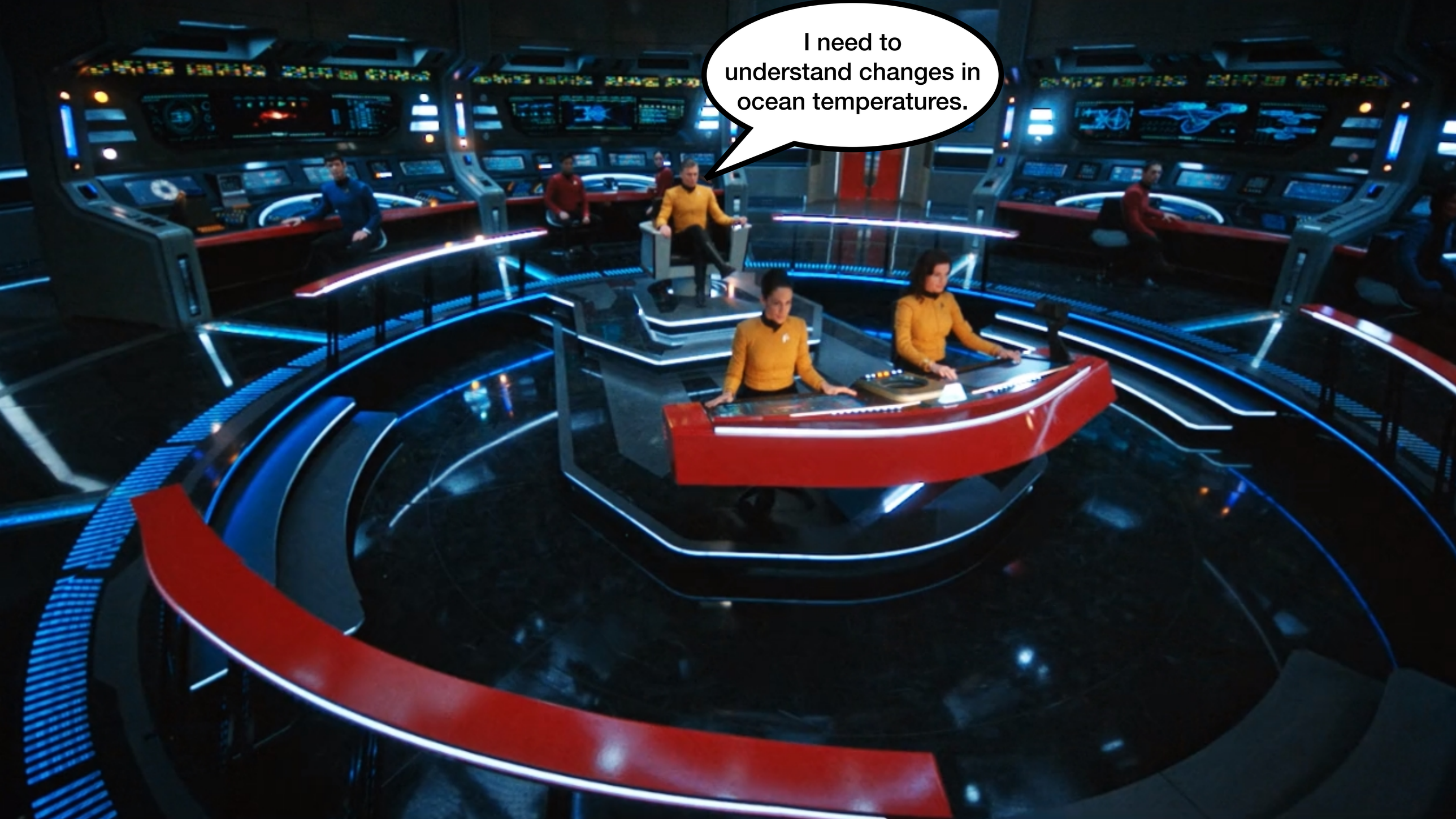
Geosciences in the future

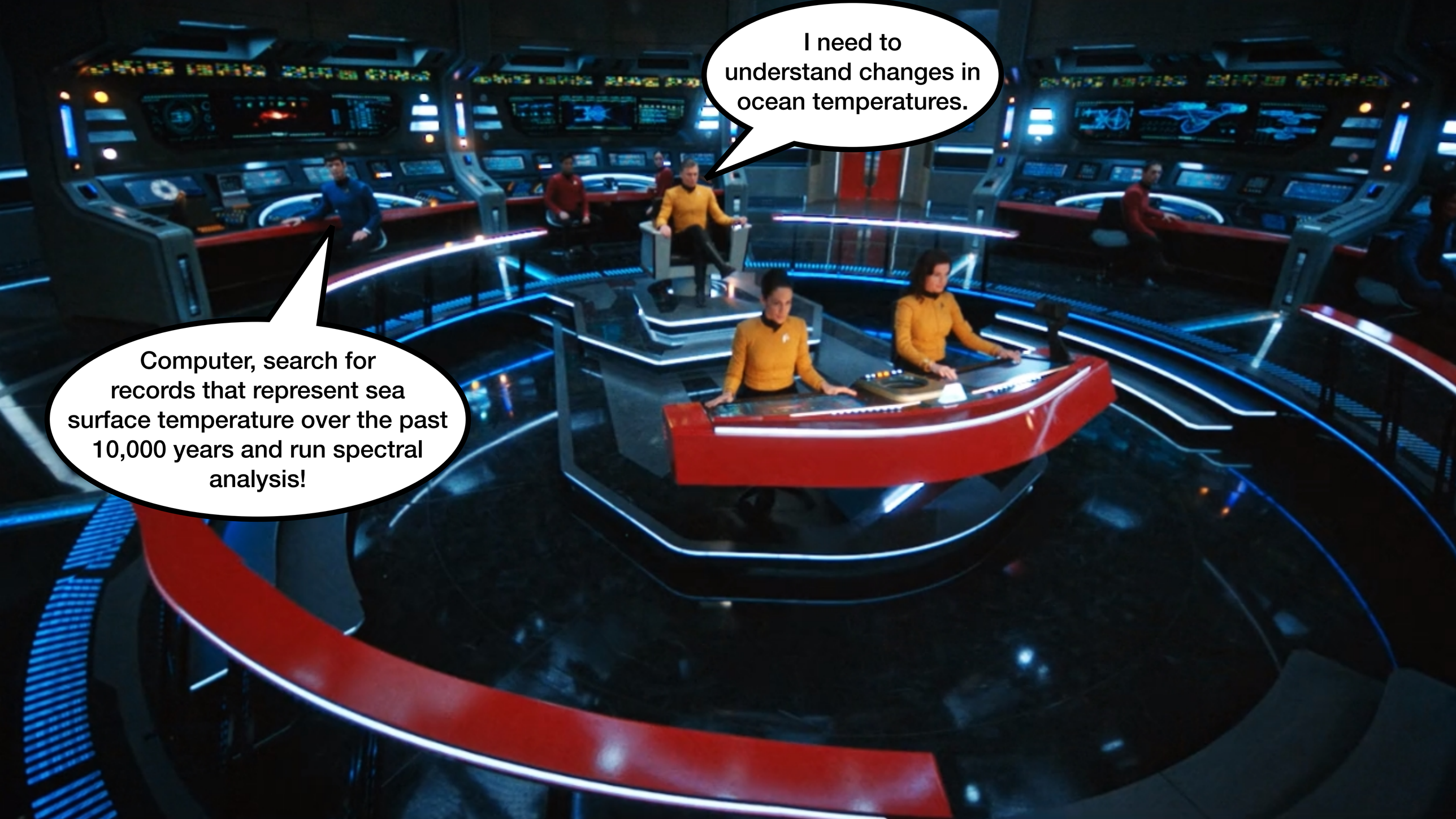
I need to make decisions based on timely science results

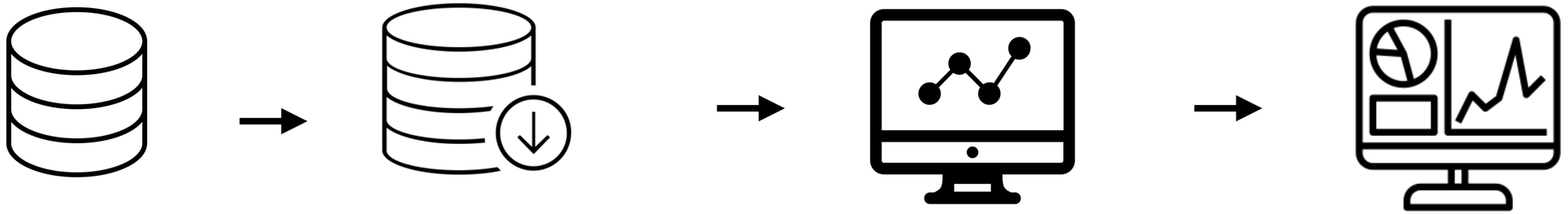
I need to make decisions based on timely science results
I need data pertinent to the problem, run the analysis, and interpret the results

I need to make decisions based on timely science results
I need data pertinent to the problem, run the analysis, and interpret the results
I need to communicate the decision to stakeholders


I will be doing all the work
I need to make decisions based on timely science results
I need data pertinent to the problem, run the analysis, and interpret the results
I need to communicate the decision to stakeholders


I need to understand changes in ocean temperatures.


I need to understand changes in ocean temperatures.
Computer, search for records that represent sea surface temperature over the past 10,000 years and run spectral analysis!



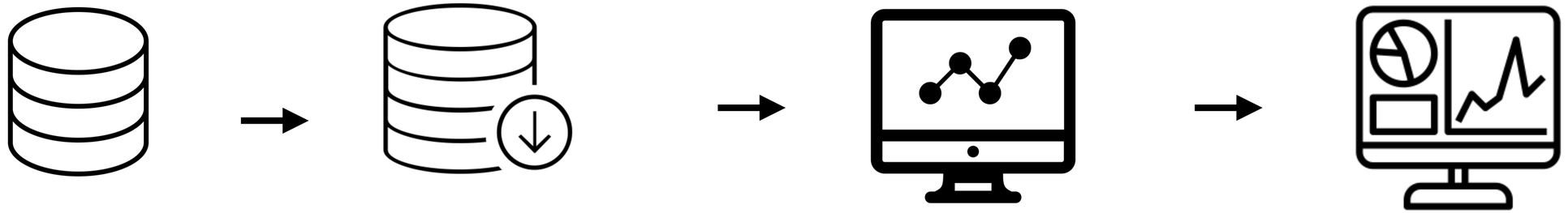
Query Data Catalog

Download datasets

Spectral Analysis

Visualize and Interpret Results

Icons: The Noun Project



Query Data Catalog

Download datasets

Spectral Analysis

Visualize and Interpret Results

1. Standardization is key to enable automation
2. Create analysis code that makes use of this standard

Icons: The Noun Project

LinkedEarth: Bringing paleoclimatology into the 21st century

Yolanda Gil
USC

Deborah Khider
USC

Julien Emile-Geay
USC

Nick McKay
NAU

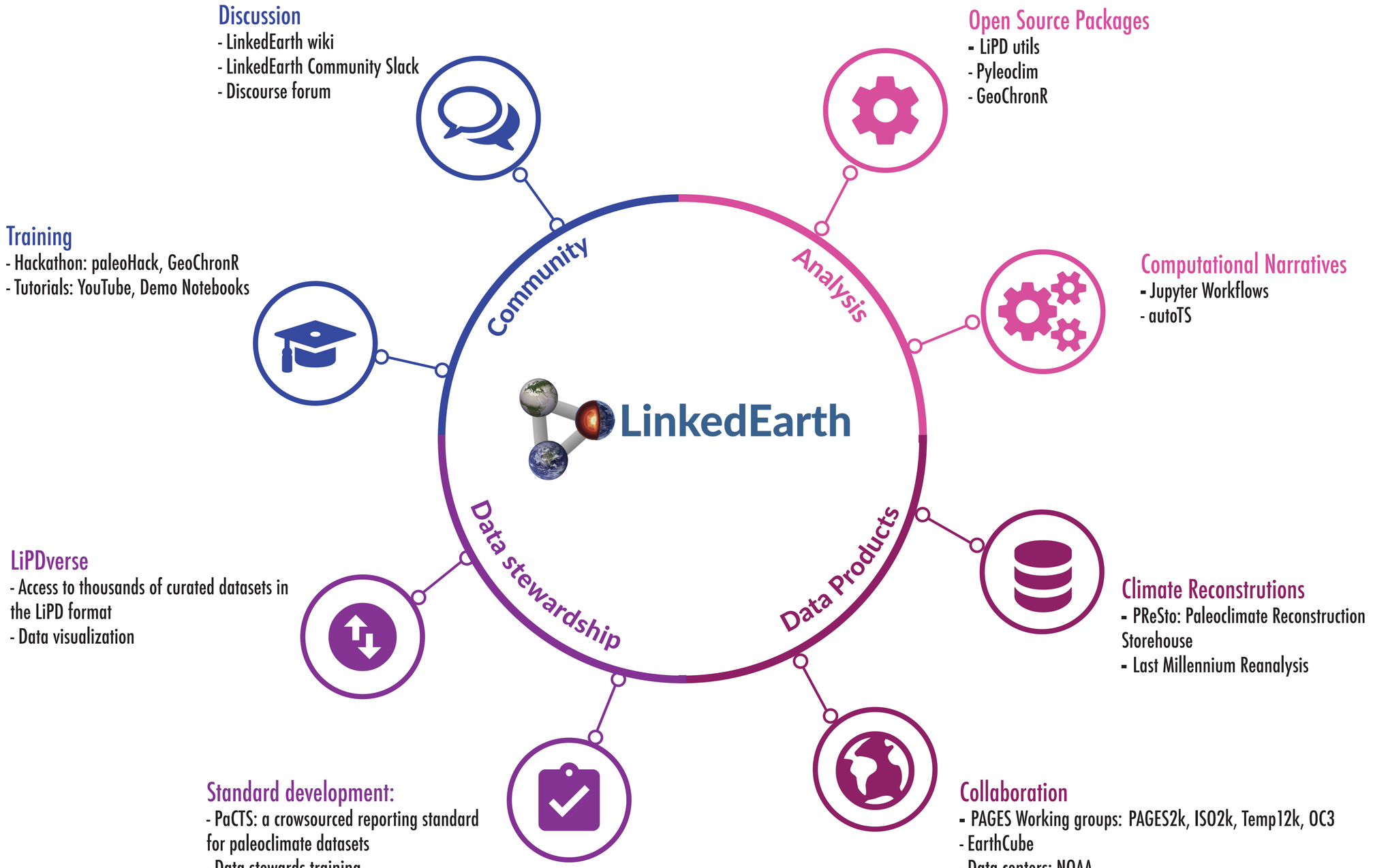
Daniel Garijo
USC

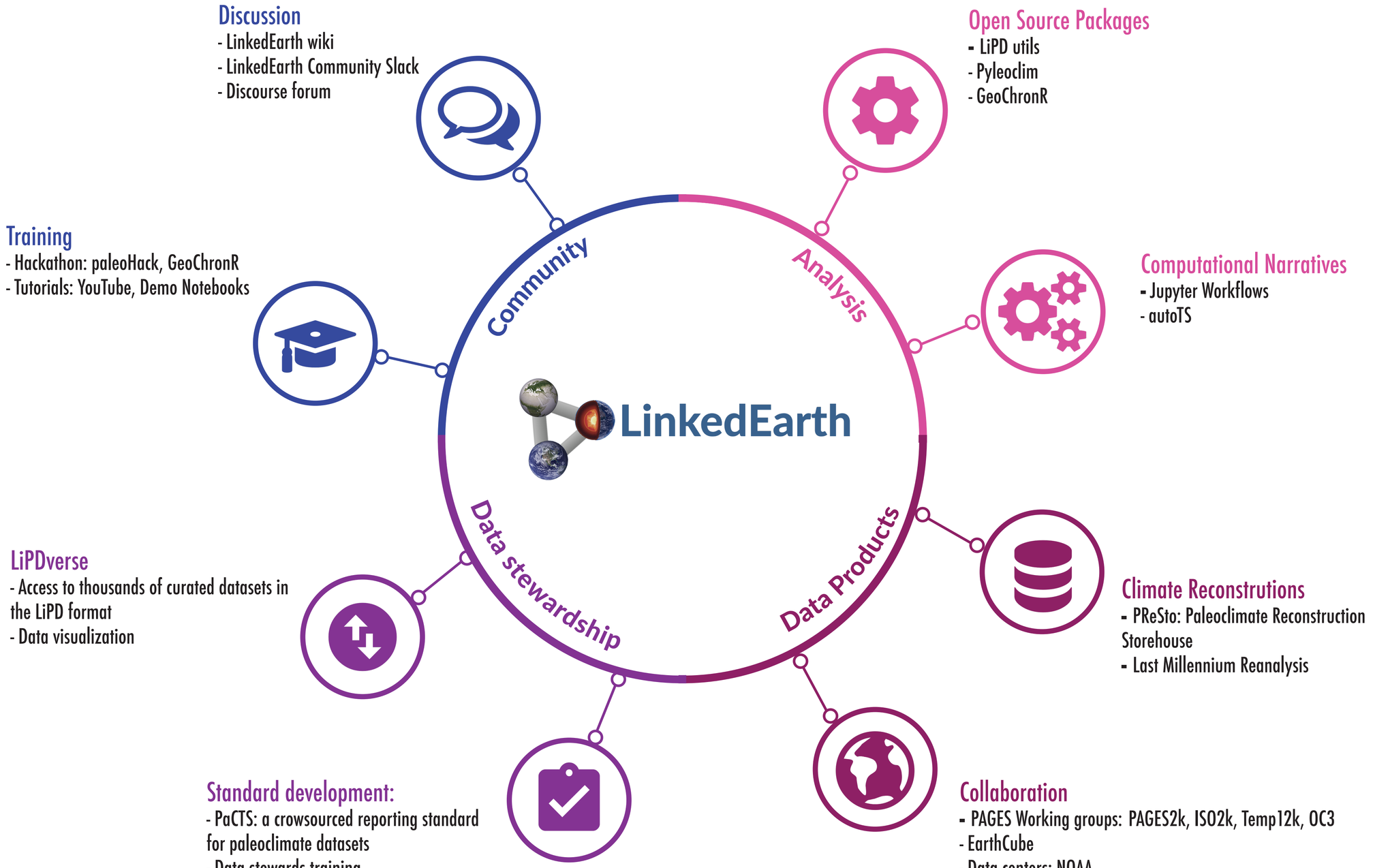
Feng Zhu
USC

Jun Hu
USC

Varun Ratnakar
USC

Chris Heiser
NAU

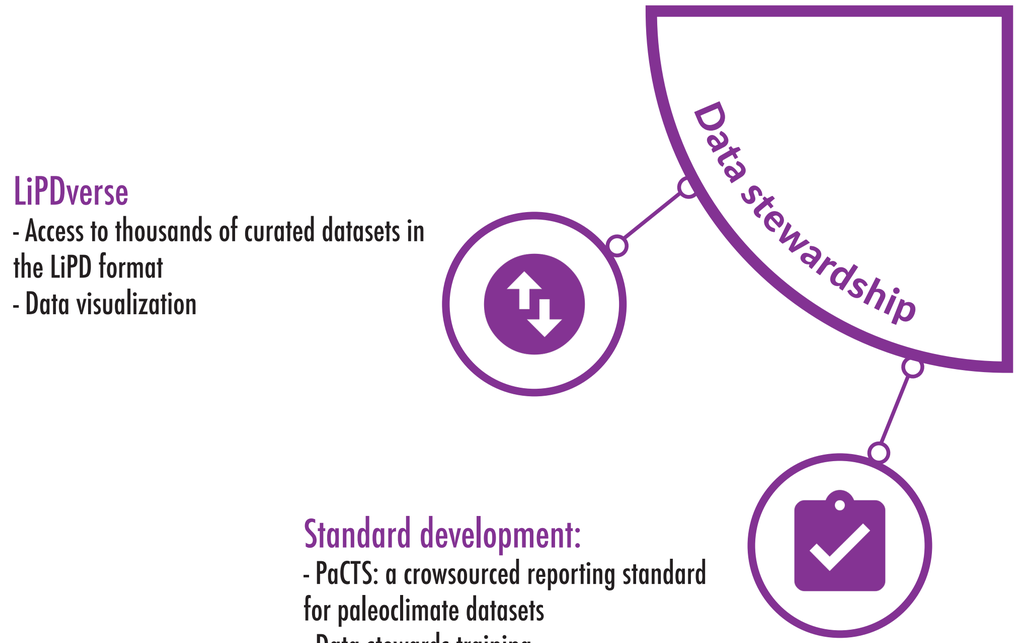

Discussion
- LinkedEarth wiki
- LinkedEarth Community Slack
- Discourse forum
Open Source Packages
- LiPD utils
- Pyleoclim
- GeoChronR
Training
- Hackathon: paleoHack, GeoChronR
- Tutorials: YouTube, Demo Notebooks
Computational Narratives
- Jupyter Workflows
- autoTS
Community
Analysis
LinkedEarth
Data stewardship
Data Products
LiPDverse
- Access to thousands of curated datasets in the LiPD format
- Data visualization
Climate Reconstrutions
- PReSto: Paleoclimate Reconstruction Storehouse
- Last Millennium Reanalysis
Standard development:
- PaCTS: a crowsourced reporting standard for paleoclimate datasets
- Data stewards training
Collaboration
- PAGES Working groups: PAGES2k, ISO2k, Temp12k, OC3
- EarthCube
- Data centers: NOAA


Discussion
- LinkedEarth wiki
- LinkedEarth Community Slack
- Discourse forum
Open Source Packages
- LiPD utils
- Pyleoclim
- GeoChronR
Training
- Hackathon: paleoHack, GeoChronR
- Tutorials: YouTube, Demo Notebooks
Computational Narratives
- Jupyter Workflows
- autoTS
Community
Analysis
LinkedEarth
Data stewardship
Data Products
LiPDverse
- Access to thousands of curated datasets in the LiPD format
- Data visualization
Climate Reconstrutions
- PReSto: Paleoclimate Reconstruction Storehouse
- Last Millennium Reanalysis
Standard development:
- PaCTS: a crowsourced reporting standard for paleoclimate datasets
- Data stewards training
Collaboration
- PAGES Working groups: PAGES2k, ISO2k, Temp12k, OC3
- EarthCube
- Data centers: NOAA

What is a standard?

“A public specification documenting some practice or technology that is adopted and used by a community.”

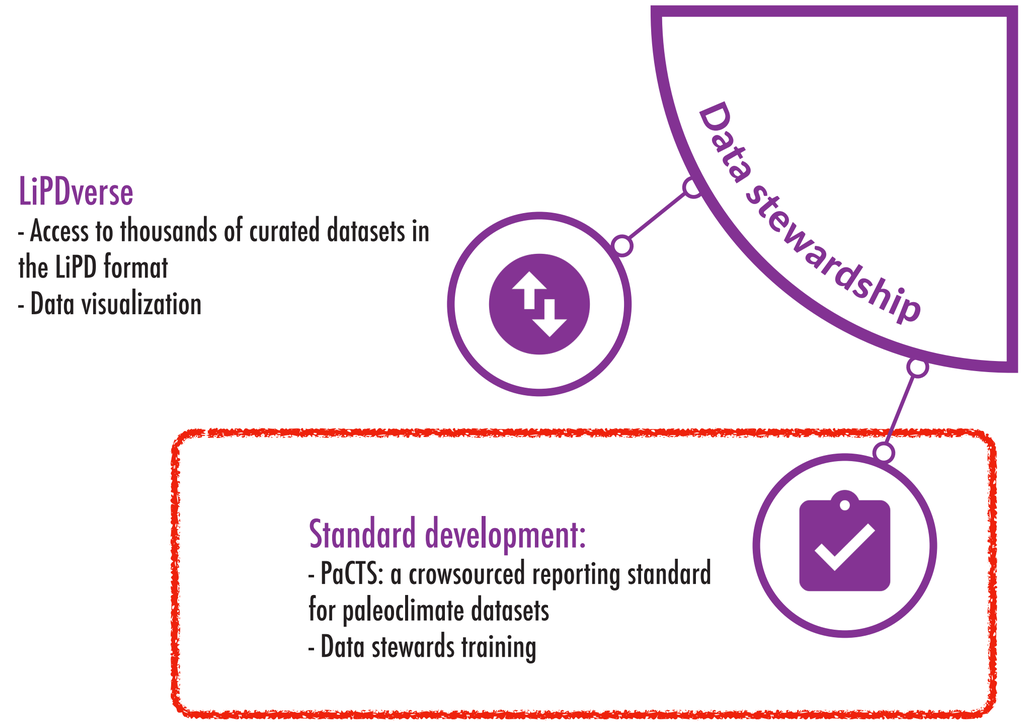
EarthCube 2015



What is a standard?

“A public specification documenting some practice or technology that is adopted and used by a community.”

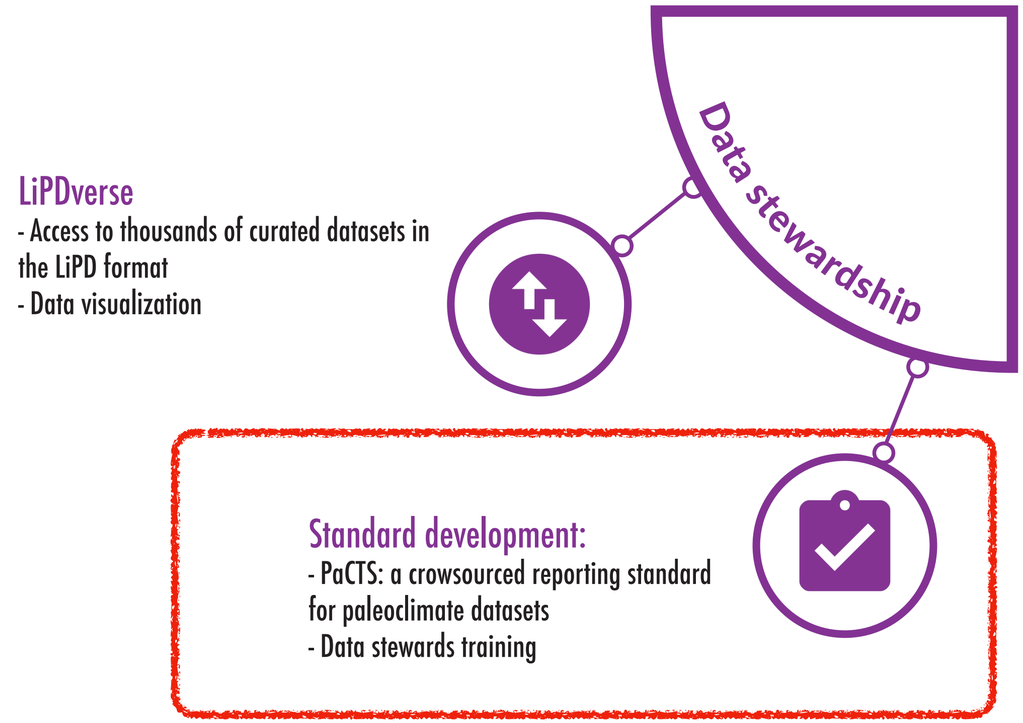
EarthCube 2015



What is a standard?

“A public specification documenting some practice or technology that is adopted and used by a community.”

EarthCube 2015



What is a standard?

A standard representation for the data

A standard terminology for the data

Standard guidelines for reporting data

What is a standard?

A standard representation for the data

A standard terminology for the data

Standard guidelines for reporting data

LiPD/LinkedEarth Ontology

What is a standard?

A standard representation for the data

A standard terminology for the data

Standard guidelines for reporting data

LiPD/LinkedEarth Ontology

PaST Thesaurus (NOAA)

What is a standard?

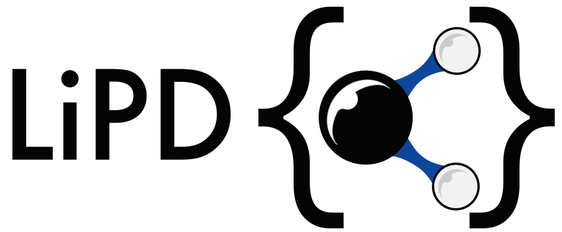
A standard representation for the data	LiPD/LinkedEarth Ontology
A standard terminology for the data	PaST Thesaurus (NOAA)
Standard guidelines for reporting data	PaCTSv1.0

What is a standard?

A standard representation for the data	LiPD/LinkedEarth Ontology
A standard terminology for the data	PaST Thesaurus (NOAA)
Standard guidelines for reporting data	PaCTSv1.0

The Linked Paleo Data Format (LiPD)

The Linked Paleo Data Format (LiPD)



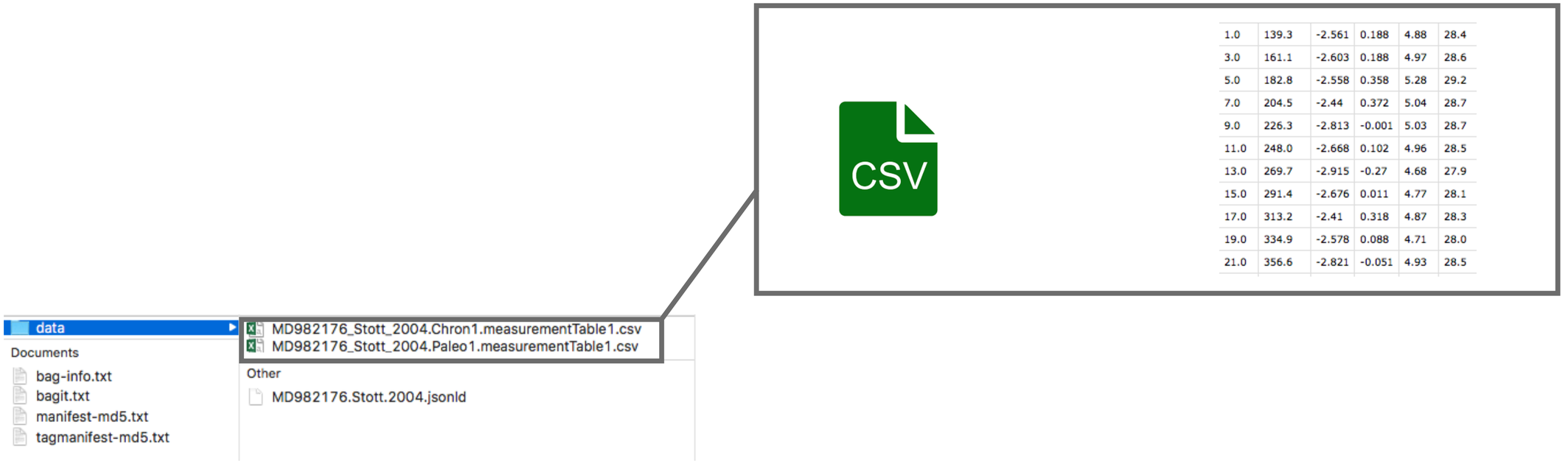
McKay & Emile-Geay, 2016, Climate of the Past

The Linked Paleo Data Format (LiPD)

LiPD

McKay & Emile-Geay, 2016, Climate of the Past

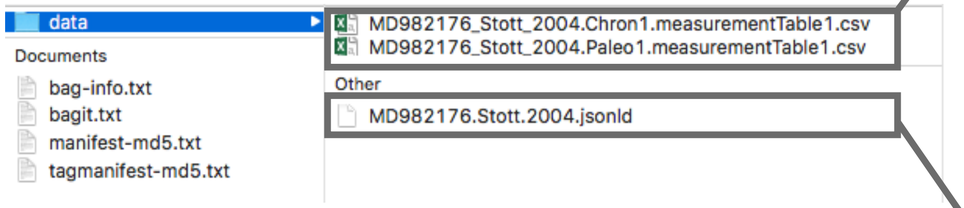
The Linked Paleo Data Format (LiPD)



LiPD

McKay & Emile-Geay, 2016, Climate of the Past

The Linked Paleo Data Format (LiPD)



LiPD

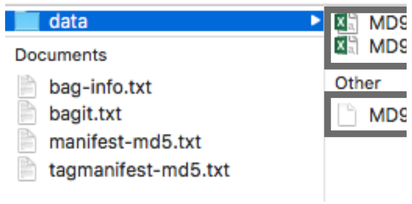
CSV

1.0	139.3	-2.561	0.188	4.88	28.4
3.0	161.1	-2.603	0.188	4.97	28.6
5.0	182.8	-2.558	0.358	5.28	29.2
7.0	204.5	-2.44	0.372	5.04	28.7
9.0	226.3	-2.813	-0.001	5.03	28.7
11.0	248.0	-2.668	0.102	4.96	28.5
13.0	269.7	-2.915	-0.27	4.68	27.9
15.0	291.4	-2.676	0.011	4.77	28.1
17.0	313.2	-2.41	0.318	4.87	28.3
19.0	334.9	-2.578	0.088	4.71	28.0
21.0	356.6	-2.821	-0.051	4.93	28.5

JSON-LD


```
"paleoData": [
    {
        "paleoMeasurementTable": [
            {
                "filename": "MD982176_Stott_2004.Paleo1.measurementTable1.csv",
                "columns": [
                    {
                        "number": "1",
                        "hasMaxValue": 887,
                        "hasMeanValue": 423.07228915663,
                        "hasMedianValue": 419,
                        "hasMinValue": 1,
                        "resolution": null,
                        "units": "cm",
                        "TSid": "PYT28ZWSOMB",
                        "variableName": "depth",
                        "variableType": "measured"
                    },
                    {
                        "number": "2",
                        "hasMaxValue": 14831,
                        "hasMeanValue": 5836.9088353414,
                        "hasMedianValue": 5449.5,
                        "hasMinValue": 139.3,
                        "resolution": null,
                        "units": "yr BP",
                        "TSid": "PYTLE0KGBBD",
                        "inferredVariableType": "Age",
                        "variableName": "yrbp",
                        "takenAtDepth": {
                            "foundInTable": "MD982176 Stott 2004.Paleo1.measurementTable1",
                            "hasColumnNumber": "1",
                            "hasMaxValue": 887,
                            "hasMeanValue": 423.07228915663,
                            "hasMedianValue": 419,
                            "hasMinValue": 1,
                            "hasProxySystem": {
                                "proxyArchiveType": "MarineSediment",
                                "proxySensorType": {
                                    "sensorGenus": "Globigerinoides",
                                    "sensorSpecies": "ruber"
                                }
                            },
                            "hasResolution": {
                                "hasMaxValue": 182,
                                "hasMeanValue": 59.240725806452,
                                "hasMedianValue": 55.8,
                                "hasMinValue": 5.7000000000007
```


McKay & Emile-Geay, 2016, Climate of the Past

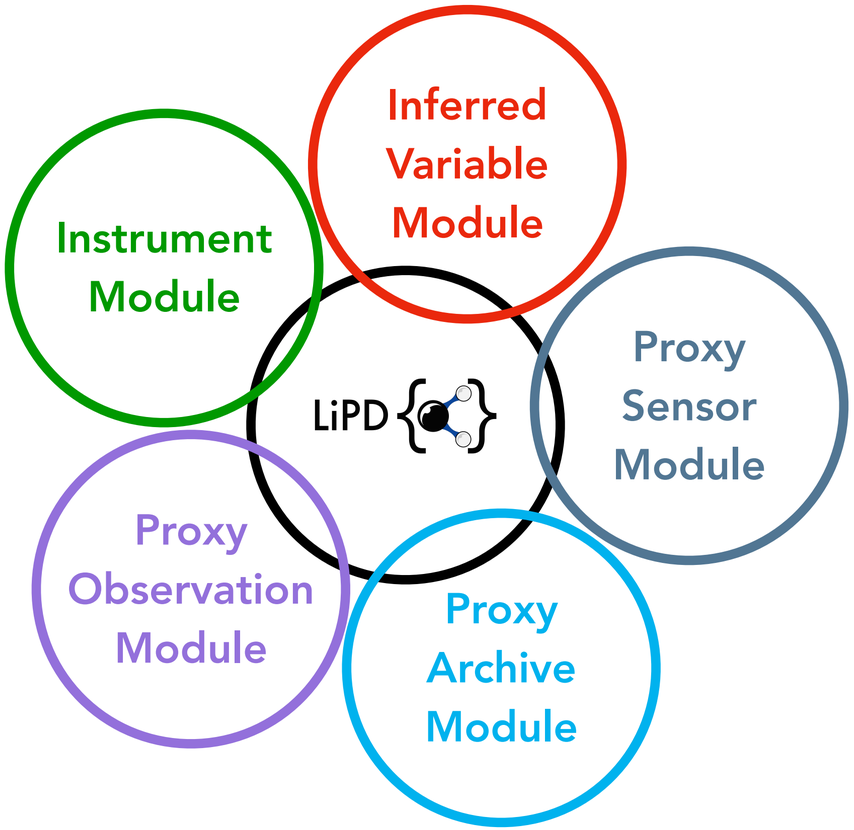


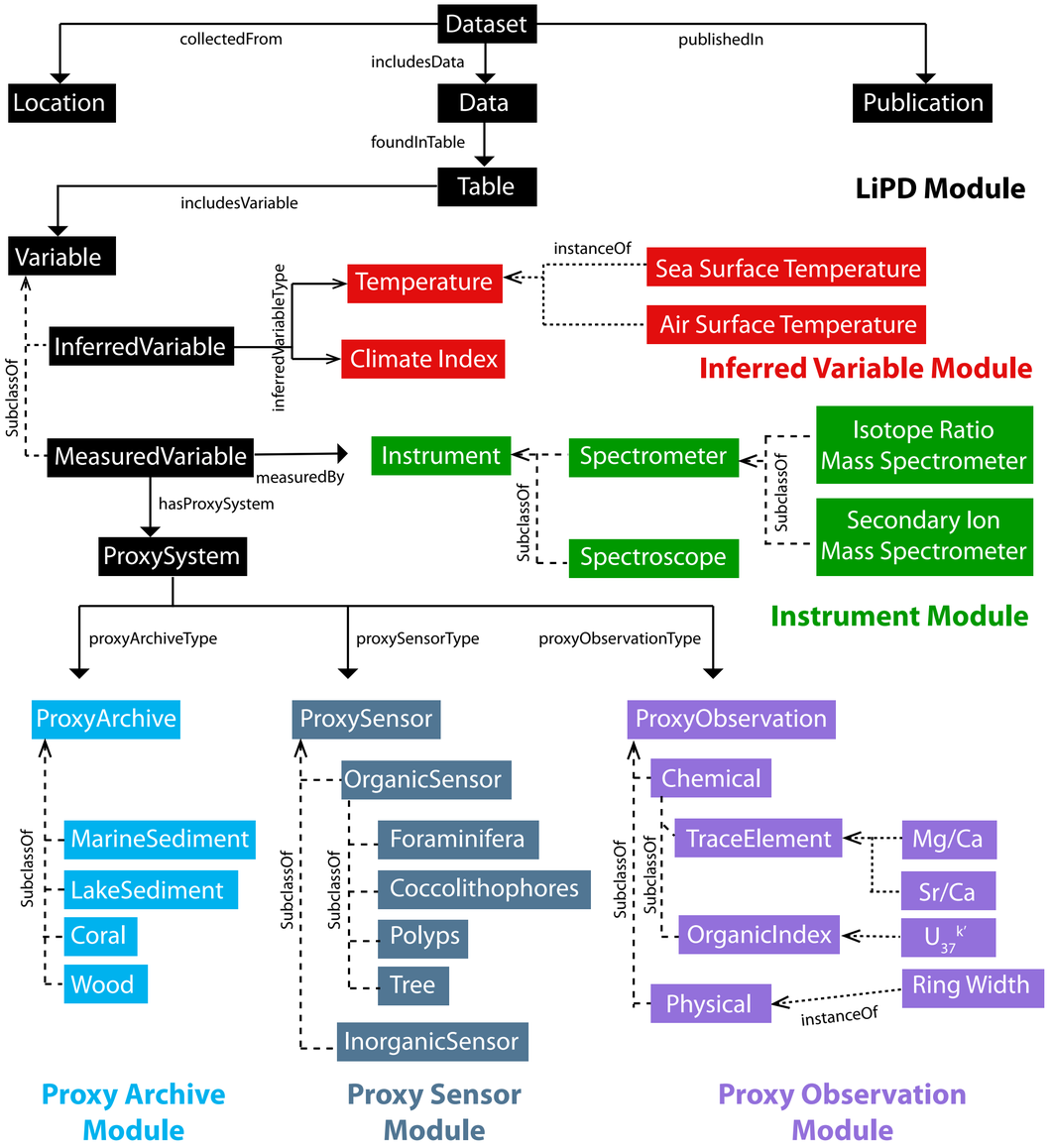
LiPD


```
"paleoData": [
    {
        "paleoMeasurementTable": [
            {
                "filename": "MD982176_Stott_2004.Paleo1.measurementTable1.csv",
                "columns": [
                    {
                        "number": "1",
                        "hasMaxValue": 887,
                        "hasMeanValue": 423.07228915663,
                        "hasMedianValue": 419,
                        "hasMinValue": 1,
                        "resolution": null,
                        "units": "cm",
                        "TSid": "PYT28ZWSOMB",
                        "variableName": "depth",
                        "variableType": "measured"
                    },
                    {
                        "number": "2",
                        "hasMaxValue": 14831,
                        "hasMeanValue": 5836.9088353414,
                        "hasMedianValue": 5449.5,
                        "hasMinValue": 139.3,
                        "resolution": null,
                        "units": "yr BP",
                        "TSid": "PYTLE0KGBBD",
                        "inferredVariableType": "Age",
                        "variableName": "yrbp",
                        "takenAtDepth": {
                            "foundInTable": "MD982176 Stott 2004.Paleo1.measurementTable1",
                            "hasColumnNumber": "1",
                            "hasMaxValue": 887,
                            "hasMeanValue": 423.07228915663,
                            "hasMedianValue": 419,
                            "hasMinValue": 1,
                            "hasProxySystem": {
                                "proxyArchiveType": "MarineSediment",
                                "proxySensorType": {
                                    "sensorGenus": "Globigerinoides",
                                    "sensorSpecies": "ruber"
                                }
                            },
                            "hasResolution": {
                                "hasMaxValue": 182,
                                "hasMeanValue": 59.240725806452,
                                "hasMedianValue": 55.8,
```

McKay & Emile-Geay, 2016,

The LinkedEarth Ontology



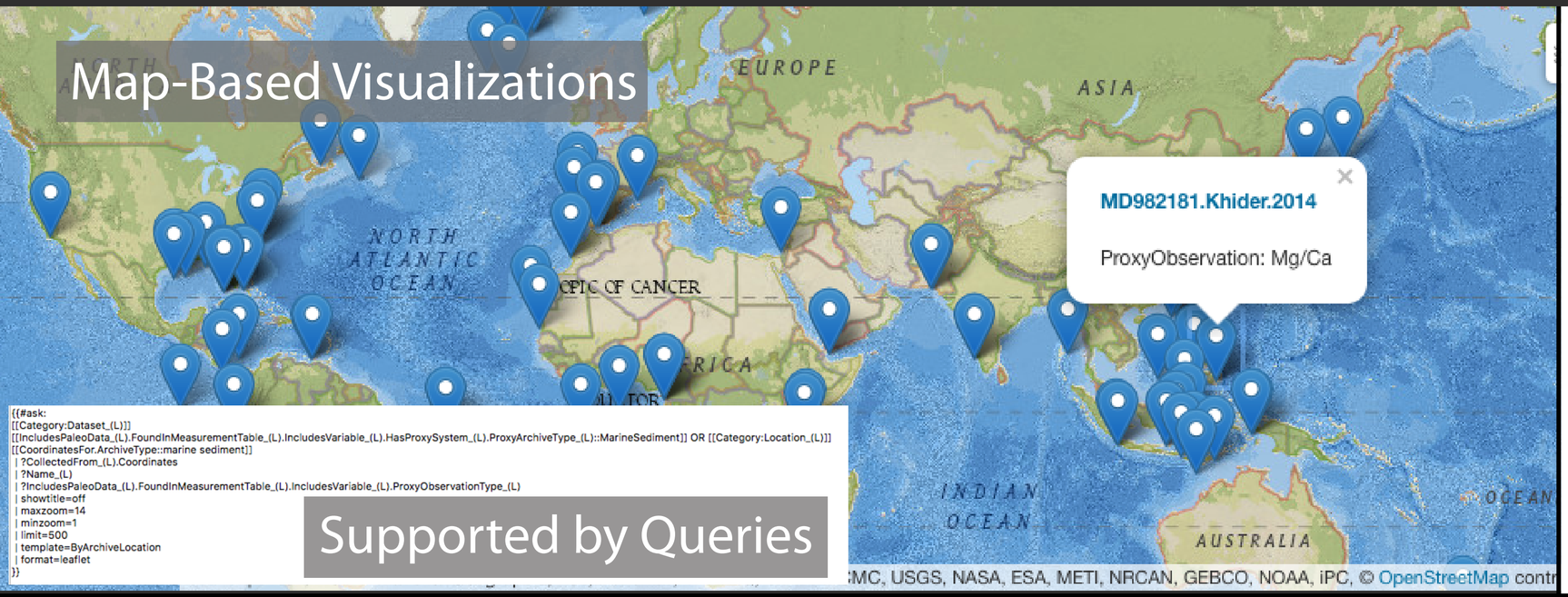


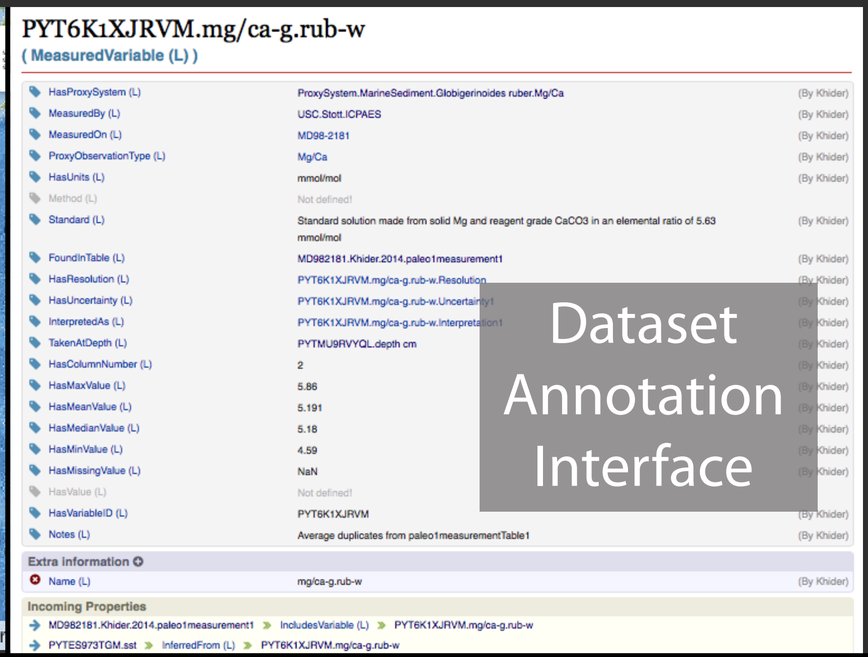
Emile-Geay et al., in revisions, Earth Science Informatics

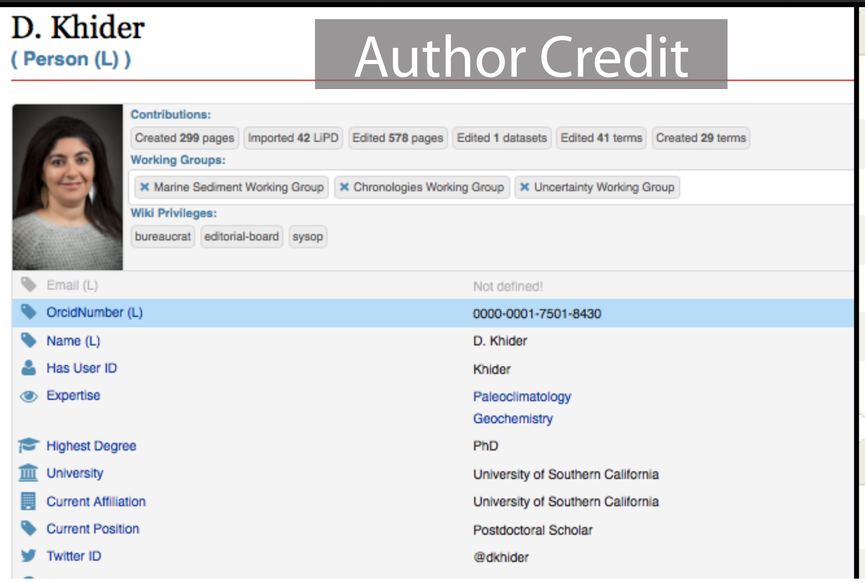
Gil et al.. 2017, ISWC

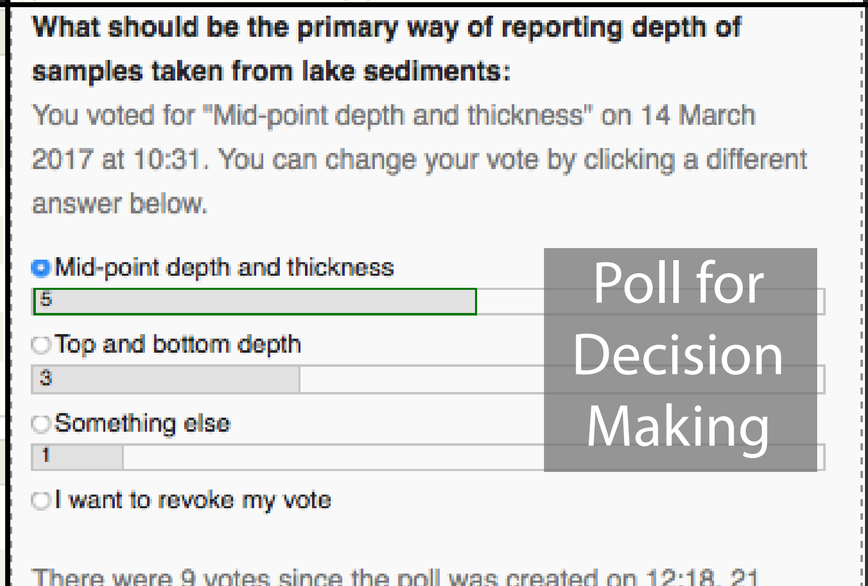
http://linked.earth/ontology

The LinkedEarth Platform



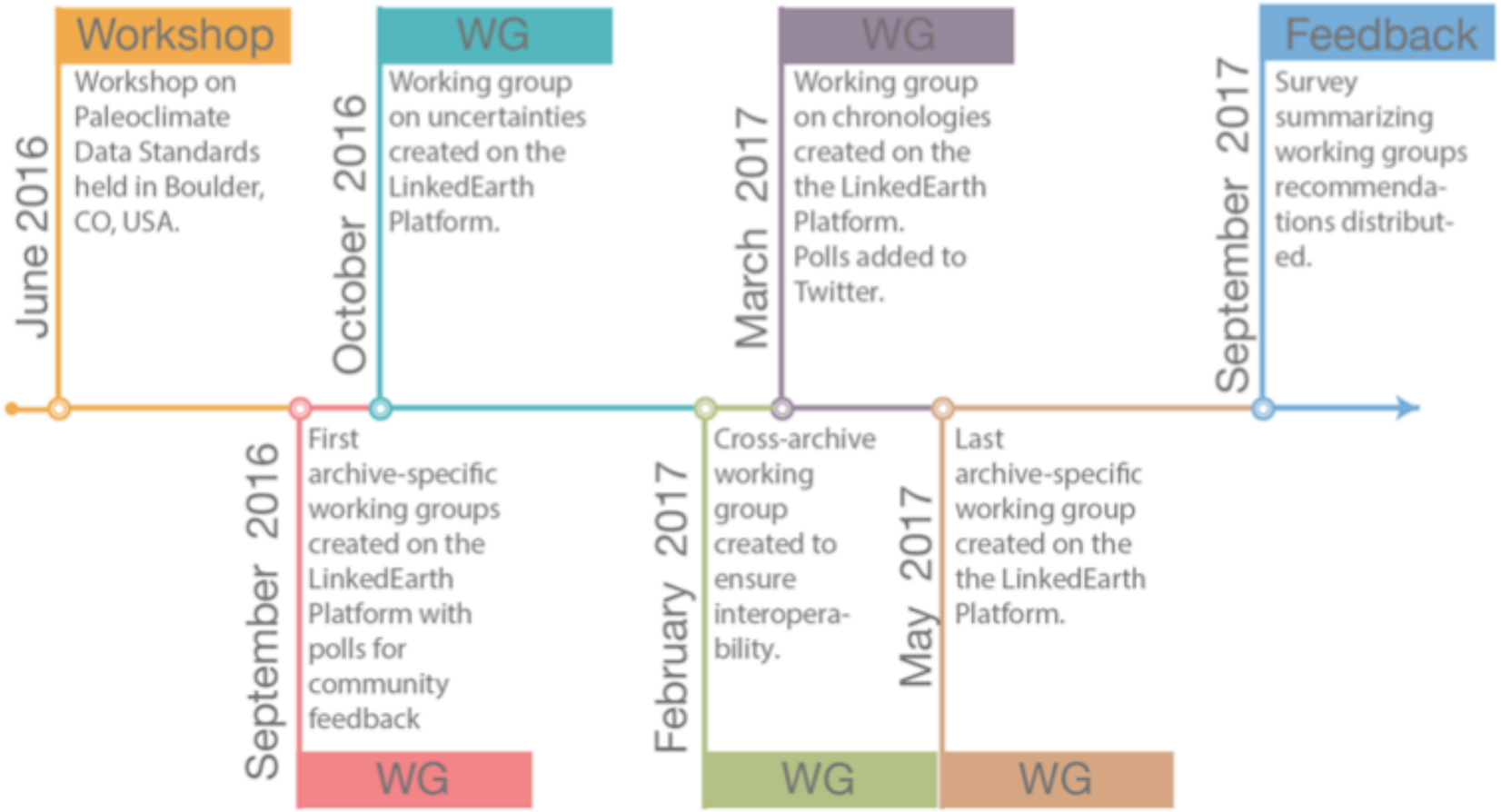






Gil et al., 2017, International Semantic Web Conference

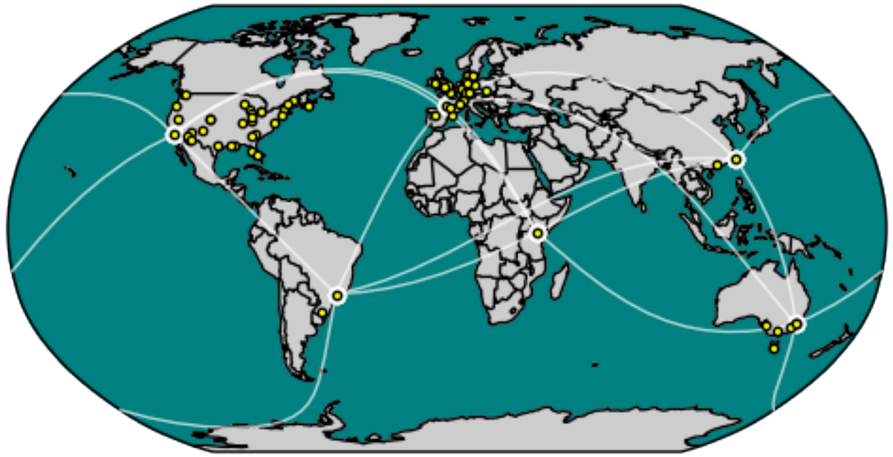
PaCTS: A crowdsourced reporting standard for paleoclimate data



Khider et al., 2019, Paleoceanography and Paleoclimatology

PaCTS: A crowdsourced reporting standard for paleoclimate data

135 researchers…

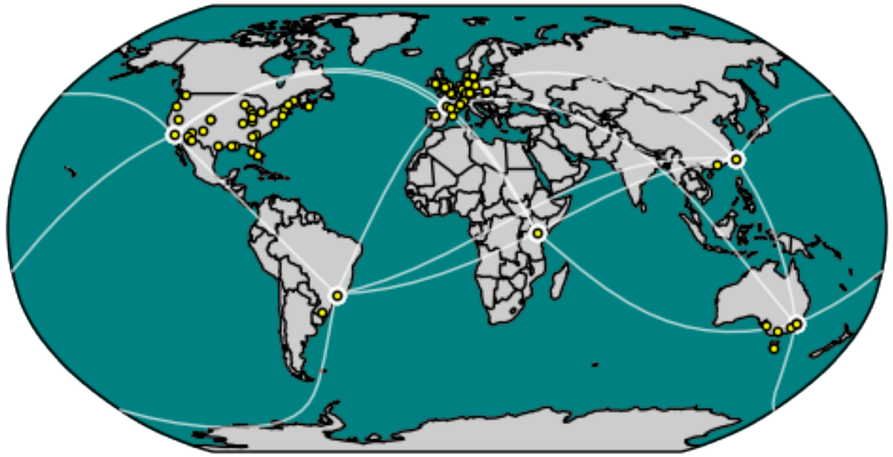


… voted on 603 properties

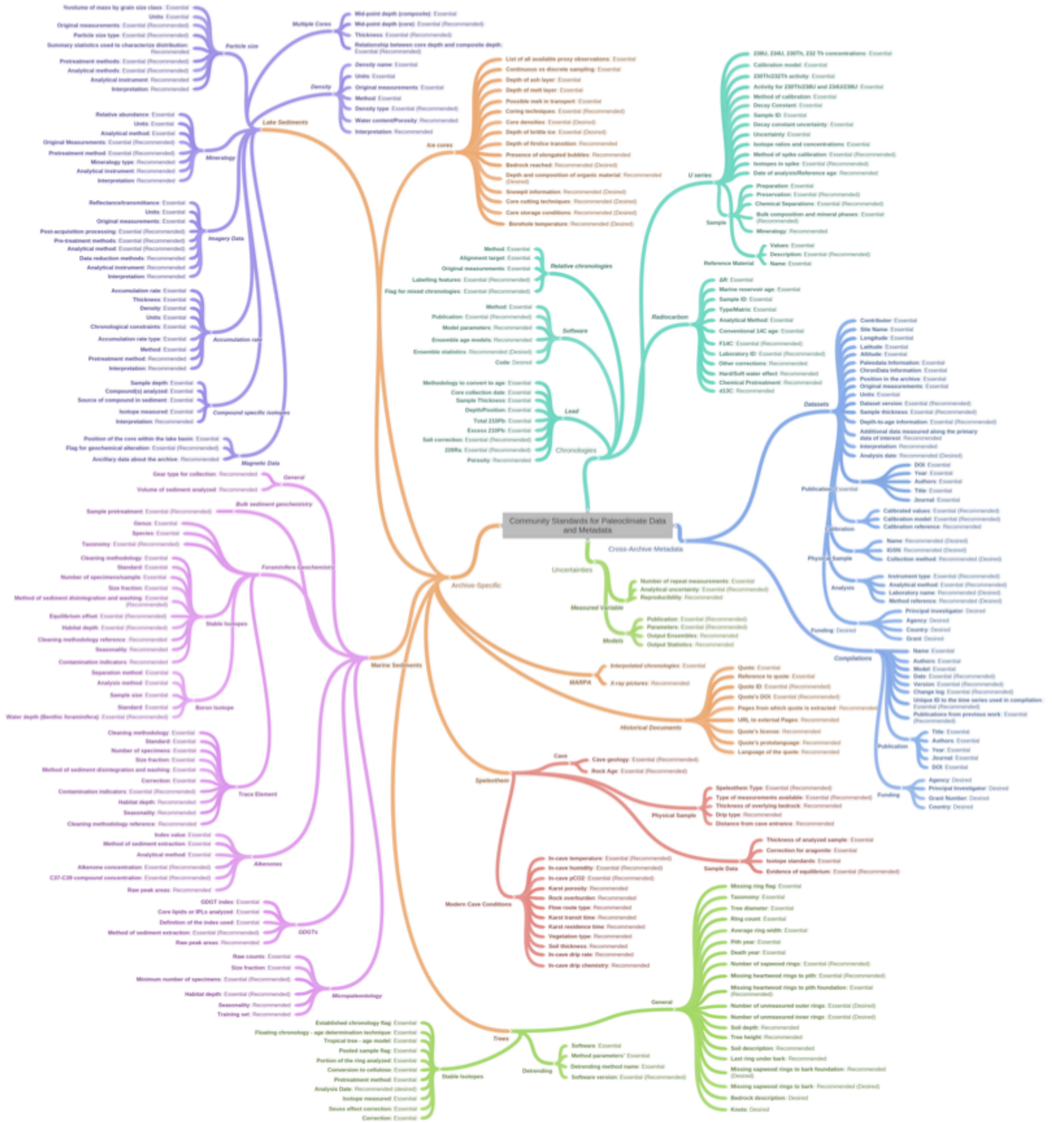
Khider et al., 2019, Paleoceanography and Paleoclimatology

PaCTS: A crowdsourced reporting standard for paleoclimate data

135 researchers…



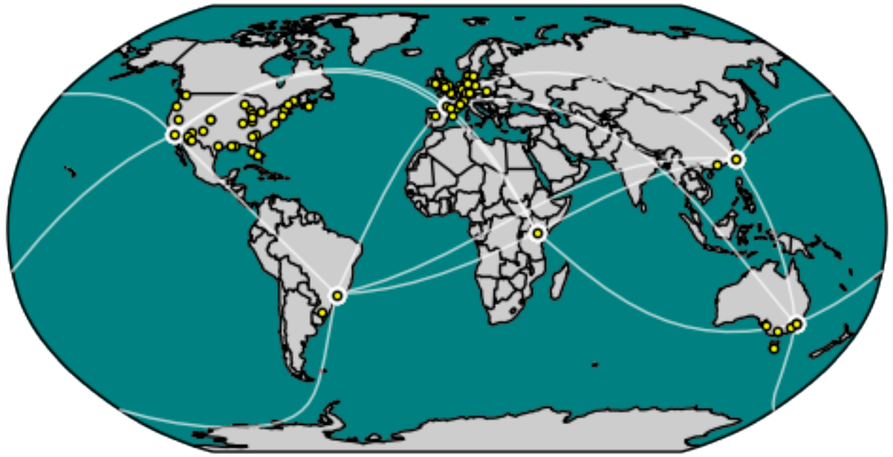
… voted on 603 properties



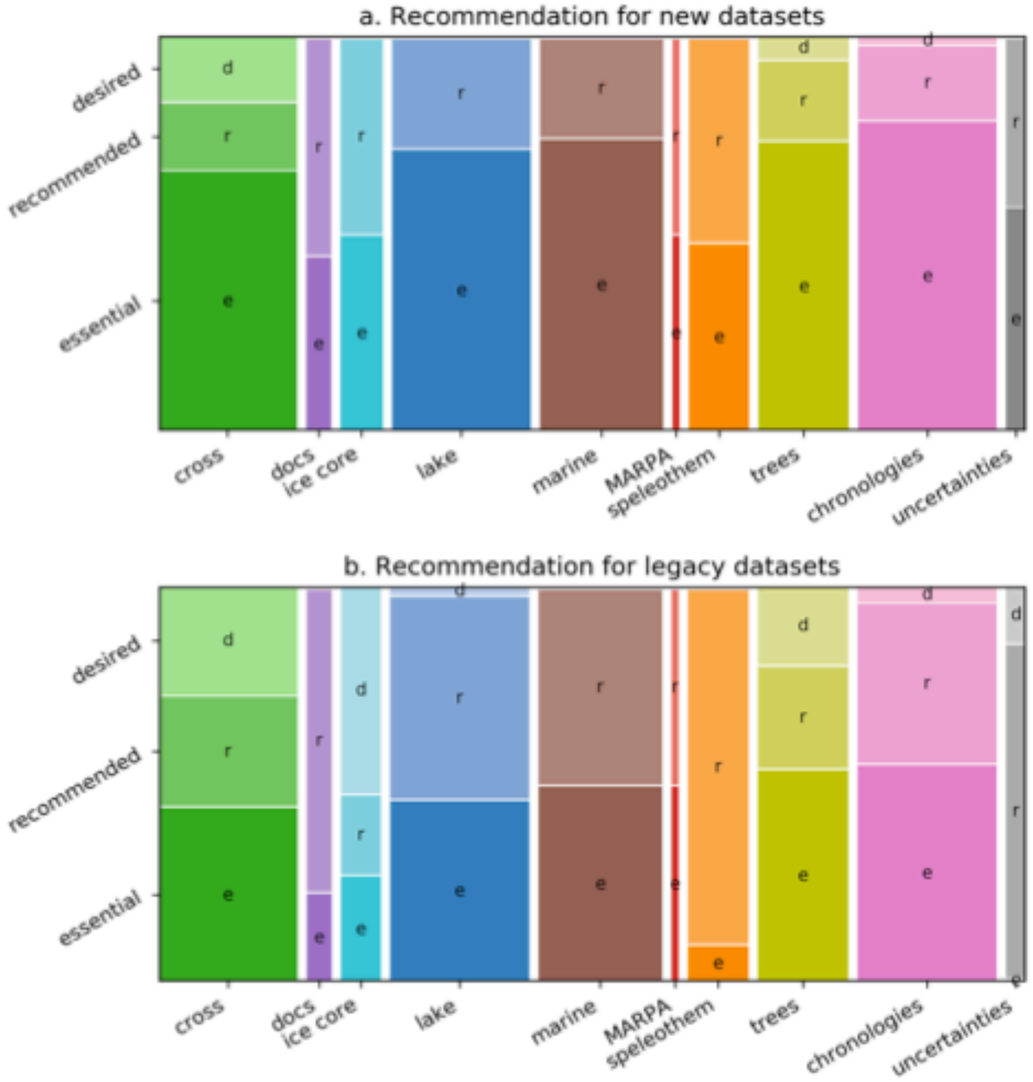
Khider et al., 2019, Paleoceanography and Paleoclimatology

PaCTS: A crowdsourced reporting standard for paleoclimate data

135 researchers…



… voted on 603 properties



Khider et al., 2019, Paleoceanography and Paleoclimatology

Implementing PaCTS

1. Lower the barrier to metadata archiving
2. Change the incentive structure

Funding agencies

Editors and publishers

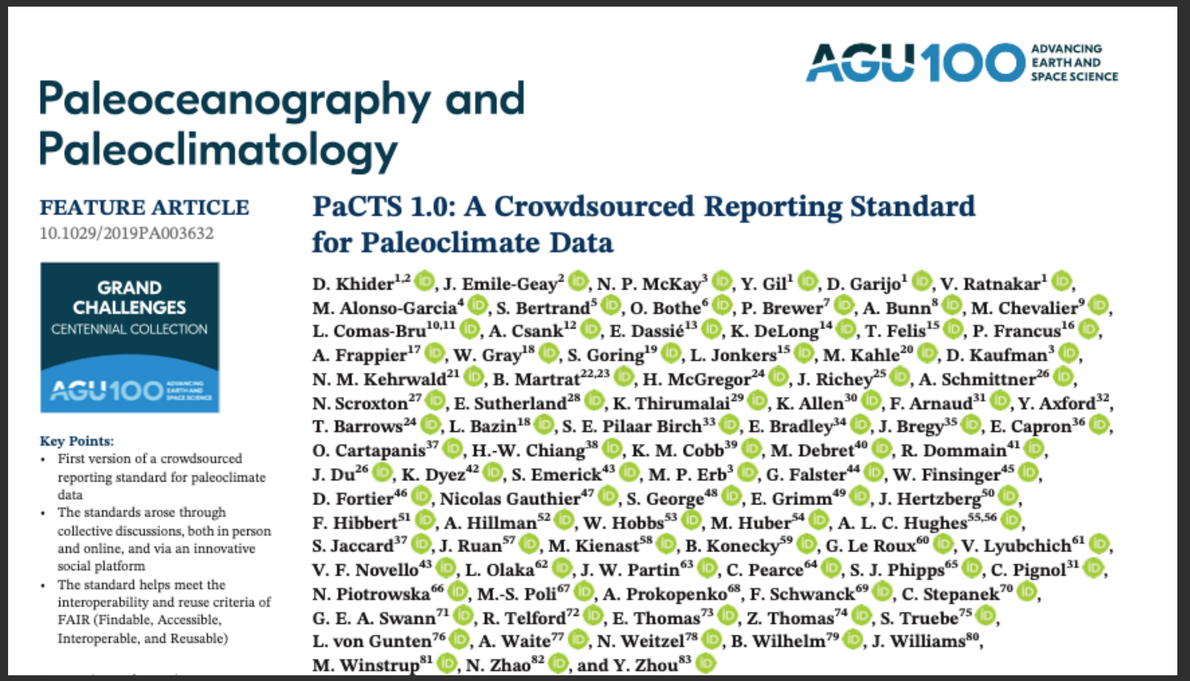
Professional societies

Science-enabling code

Implementing PaCTS

1. Lower the barrier to metadata archiving
2. Change the incentive structure

- $ Funding agencies
- Editors and publishers
- Professional societies
- Science-enabling code



AGU100 ADVANCING EARTH AND SPACE SCIENCE

Paleoceanography and Paleoclimatology

FEATURE ARTICLE
10.1029/2019PA003632

GRAND CHALLENGES CENTENNIAL COLLECTION
AGU100 ADVANCING EARTH AND SPACE SCIENCE

PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data

D. Khider[1,2], J. Emile-Geay[2], N. P. McKay[3], Y. Gil[1], D. Garijo[1], V. Ratnakar[1], M. Alonso-Garcia[4], S. Bertrand[5], O. Bothe[6], P. Brewer[7], A. Bunn[8], M. Chevalier[9], L. Comas-Bru[10,11], A. Csank[12], E. Dassié[13], K. DeLong[14], T. Felis[15], P. Francus[16], A. Frappier[17], W. Gray[18], S. Goring[19], L. Jonkers[15], M. Kahle[20], D. Kaufman[3], N. M. Kehrwald[21], B. Martrat[22,23], H. McGregor[24], J. Richey[25], A. Schmittner[26], N. Scroxton[27], E. Sutherland[28], K. Thirumalai[29], K. Allen[30], F. Arnaud[31], Y. Axford[32], T. Barrows[24], L. Bazin[18], S. E. Pilaar Birch[33], E. Bradley[34], J. Bregy[35], E. Capron[36], O. Cartapanis[37], H.-W. Chiang[38], K. M. Cobb[39], M. Debret[40], R. Dommain[41], J. Du[26], K. Dyez[42], S. Emerick[43], M. P. Erb[3], G. Falster[44], W. Finsinger[45], D. Fortier[46], Nicolas Gauthier[47], S. George[48], E. Grimm[49], J. Hertzberg[50], F. Hibbert[51], A. Hillman[52], W. Hobbs[53], M. Huber[54], A. L. C. Hughes[55,56], S. Jaccard[37], J. Ruan[57], M. Kienast[58], B. Konecky[59], G. Le Roux[60], V. Lyubchich[61], V. F. Novello[43], L. Olaka[62], J. W. Partin[63], C. Pearce[64], S. J. Phipps[65], C. Pignol[31], N. Piotrowska[66], M.-S. Poli[67], A. Prokopenko[68], F. Schwanck[69], C. Stepanek[70], G. E. A. Swann[71], R. Telford[72], E. Thomas[73], Z. Thomas[74], S. Truebe[75], L. von Gunten[76], A. Waite[77], N. Weitzel[78], B. Wilhelm[79], J. Williams[80], M. Winstrup[81], N. Zhao[82], and Y. Zhou[83]

Key Points:

- First version of a crowdsourced reporting standard for paleoclimate data
- The standards arose through collective discussions, both in person and online, and via an innovative social platform
- The standard helps meet the interoperability and reuse criteria of FAIR (Findable, Accessible, Interoperable, and Reusable)

Implementing PaCTS

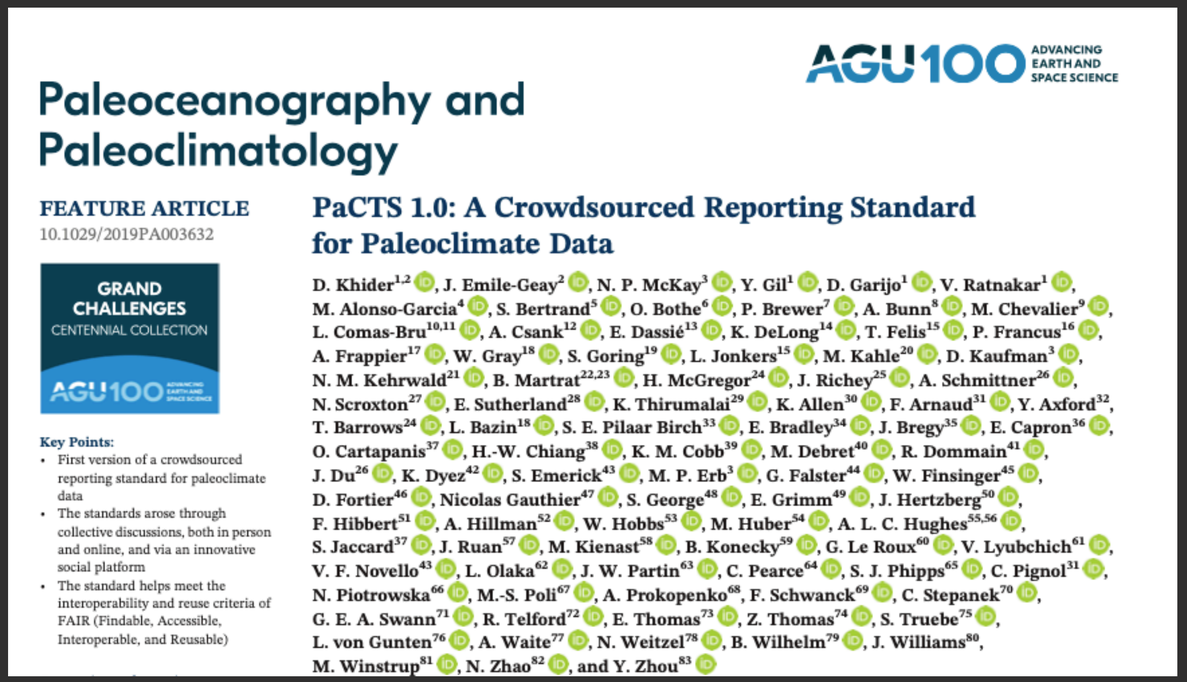
1. Lower the barrier to metadata archiving
2. Change the incentive structure

$ Funding agencies

Editors and publishers

Professional societies

Science-enabling code



AGU100 ADVANCING EARTH AND SPACE SCIENCE

Paleoceanography and Paleoclimatology

FEATURE ARTICLE

10.1029/2019PA003632

GRAND CHALLENGES CENTENNIAL COLLECTION

AGU100

Key Points:

- First version of a crowdsourced reporting standard for paleoclimate data
- The standards arose through collective discussions, both in person and online, and via an innovative social platform
- The standard helps meet the interoperability and reuse criteria of FAIR (Findable, Accessible, Interoperable, and Reusable)

PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data

D. Khider[1,2], J. Emile-Geay[2], N. P. McKay[3], Y. Gil[1], D. Garijo[1], V. Ratnakar[1], M. Alonso-Garcia[4], S. Bertrand[5], O. Bothe[6], P. Brewer[7], A. Bunn[8], M. Chevalier[9], L. Comas-Bru[10,11], A. Csank[12], E. Dassié[13], K. DeLong[14], T. Felis[15], P. Francus[16], A. Frappier[17], W. Gray[18], S. Goring[19], L. Jonkers[15], M. Kahle[20], D. Kaufman[3], N. M. Kehrwald[21], B. Martrat[22,23], H. McGregor[24], J. Richey[25], A. Schmittner[26], N. Scroxton[27], E. Sutherland[28], K. Thirumalai[29], K. Allen[30], F. Arnaud[31], Y. Axford[32], T. Barrows[24], L. Bazin[18], S. E. Pilaar Birch[33], E. Bradley[34], J. Bregy[35], E. Capron[36], O. Cartapanis[37], H.-W. Chiang[38], K. M. Cobb[39], M. Debret[40], R. Dommain[41], J. Du[26], K. Dyez[42], S. Emerick[43], M. P. Erb[3], G. Falster[44], W. Finsinger[45], D. Fortier[46], Nicolas Gauthier[47], S. George[48], E. Grimm[49], J. Hertzberg[50], F. Hibbert[51], A. Hillman[52], W. Hobbs[53], M. Huber[54], A. L. C. Hughes[55,56], S. Jaccard[37], J. Ruan[57], M. Kienast[58], B. Konecky[59], G. Le Roux[60], V. Lyubchich[61], V. F. Novello[43], L. Olaka[62], J. W. Partin[63], C. Pearce[64], S. J. Phipps[65], C. Pignol[31], N. Piotrowska[66], M.-S. Poli[67], A. Prokopenko[68], F. Schwanck[69], C. Stepanek[70], G. E. A. Swann[71], R. Telford[72], E. Thomas[73], Z. Thomas[74], S. Truebe[75], L. von Gunten[76], A. Waite[77], N. Weitzel[78], B. Wilhelm[79], J. Williams[80], M. Winstrup[81], N. Zhao[82], and Y. Zhou[83]

Implementing PaCTS

1. Lower the barrier to metadata archiving
2. Change the incentive structure

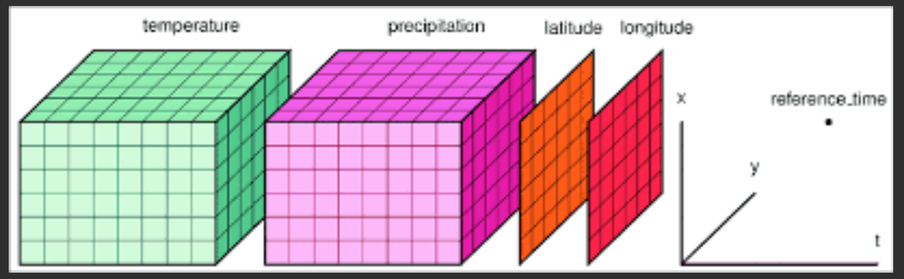
$ Funding agencies

Editors and publishers

Professional societies

Science-enabling code

NetCDF-CF



Implementing PaCTS

1. Lower the barrier to metadata archiving
2. Change the incentive structure

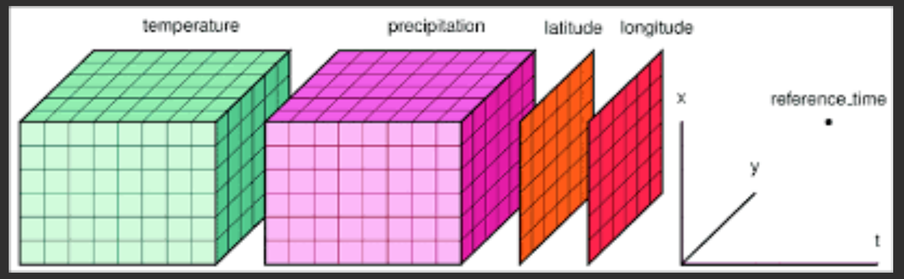
$ Funding agencies

Editors and publishers

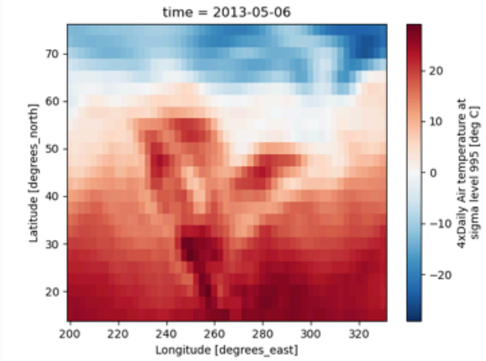
Professional societies

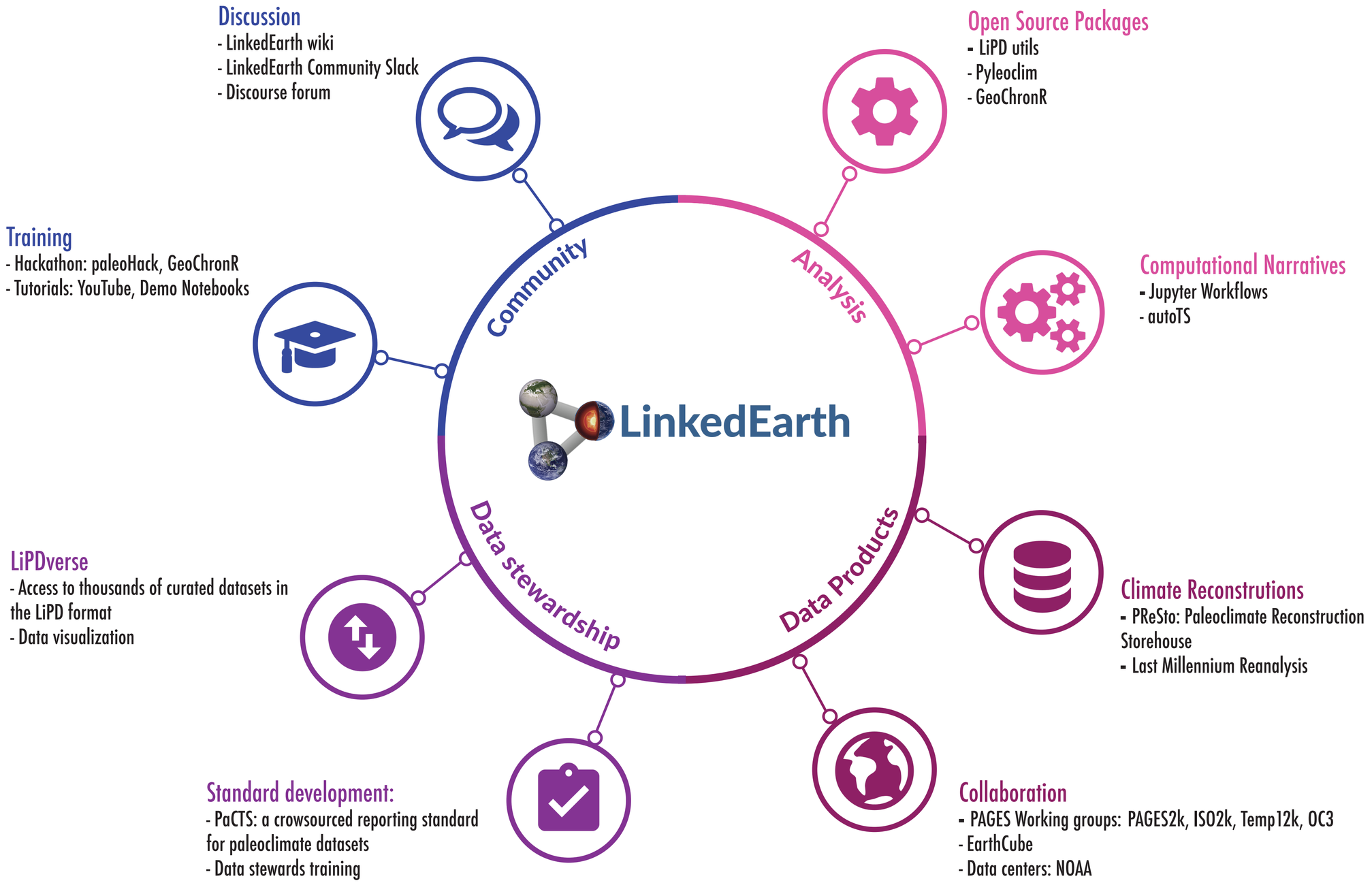
Science-enabling code

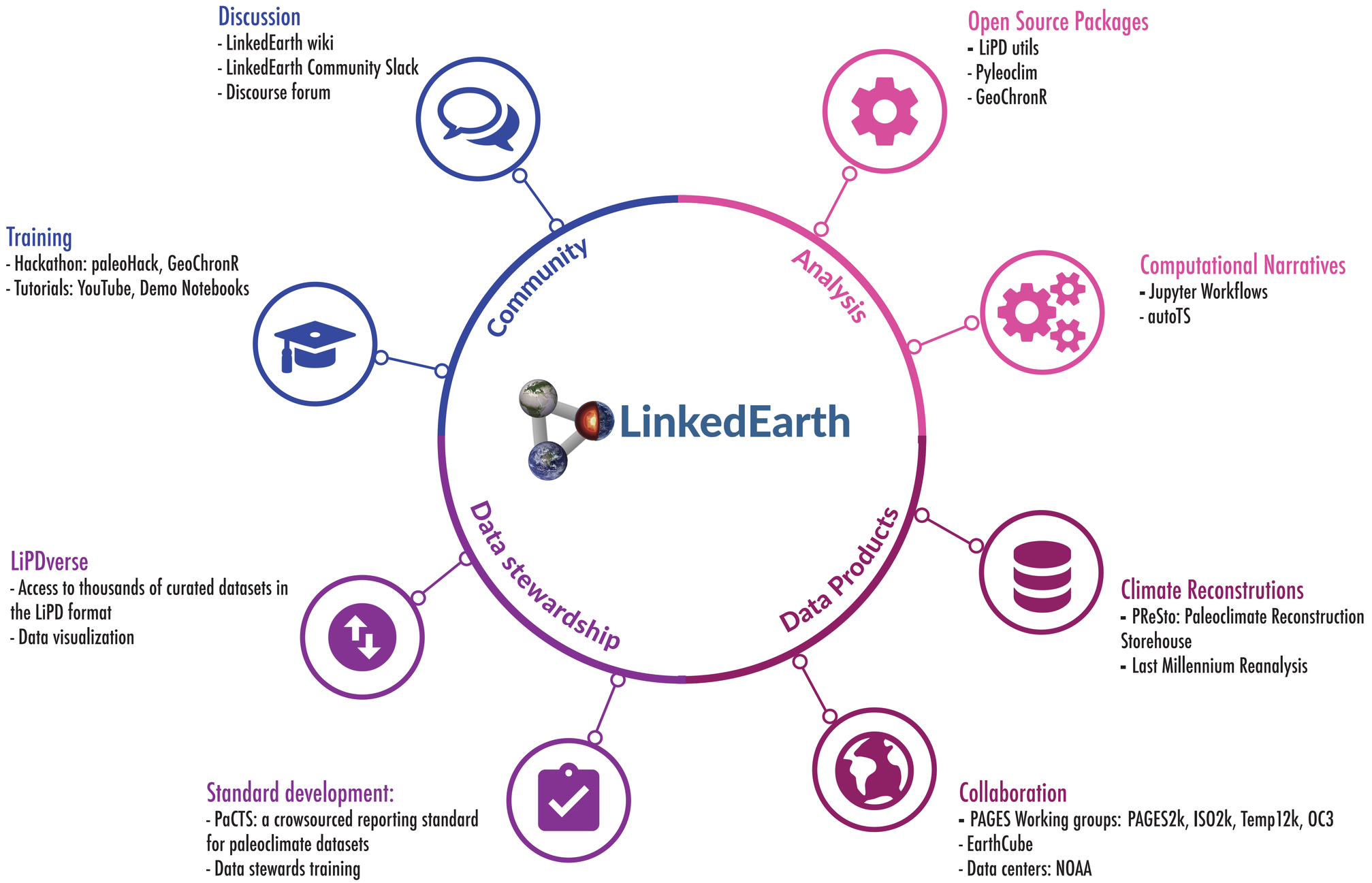
NetCDF-CF

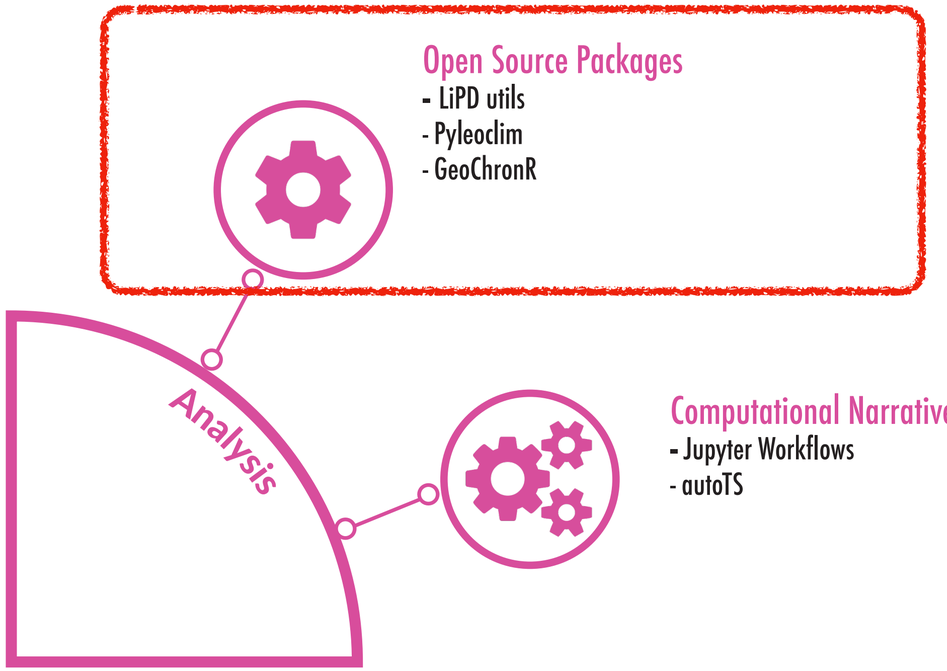



```
In [43]: air2d = air.isel(time=500)

In [44]: air2d.plot()
Out[44]: <matplotlib.collections.QuadMesh at 0x7f4ad45c8400>
```




Discussion
- LinkedEarth wiki
- LinkedEarth Community Slack
- Discourse forum
Open Source Packages
- LiPD utils
- Pyleoclim
- GeoChronR
Training
- Hackathon: paleoHack, GeoChronR
- Tutorials: YouTube, Demo Notebooks
Computational Narratives
- Jupyter Workflows
- autoTS
Community
Analysis
LinkedEarth
Data stewardship
Data Products
LiPDverse
- Access to thousands of curated datasets in the LiPD format
- Data visualization
Climate Reconstrutions
- PReSto: Paleoclimate Reconstruction Storehouse
- Last Millennium Reanalysis
Standard development:
- PaCTS: a crowsourced reporting standard for paleoclimate datasets
- Data stewards training
Collaboration
- PAGES Working groups: PAGES2k, ISO2k, Temp12k, OC3
- EarthCube
- Data centers: NOAA


Discussion
- LinkedEarth wiki
- LinkedEarth Community Slack
- Discourse forum
Open Source Packages
- LiPD utils
- Pyleoclim
- GeoChronR
Training
- Hackathon: paleoHack, GeoChronR
- Tutorials: YouTube, Demo Notebooks
Computational Narratives
- Jupyter Workflows
- autoTS
Community
Analysis
LinkedEarth
Data stewardship
Data Products
LiPDverse
- Access to thousands of curated datasets in the LiPD format
- Data visualization
Climate Reconstrutions
- PReSto: Paleoclimate Reconstruction Storehouse
- Last Millennium Reanalysis
Standard development:
- PaCTS: a crowsourced reporting standard for paleoclimate datasets
- Data stewards training
Collaboration
- PAGES Working groups: PAGES2k, ISO2k, Temp12k, OC3
- EarthCube
- Data centers: NOAA



Open Source Packages

- LiPD utils
- Pyleoclim
- GeoChronR

Computational Narratives

- Jupyter Workflows
- autoTS

Pyleoclim

Pyleoclim

What is it?

- ✦ Python package for the analysis and visualization of paleoclimate data
- ✦ The package works with data stored in the LiPD format.
- ✦ Data transformation is made automatically within the package

Pyleoclim

What is it?

- ✦ Python package for the analysis and visualization of paleoclimate data
- ✦ The package works with data stored in the LiPD format.
- ✦ Data transformation is made automatically within the package

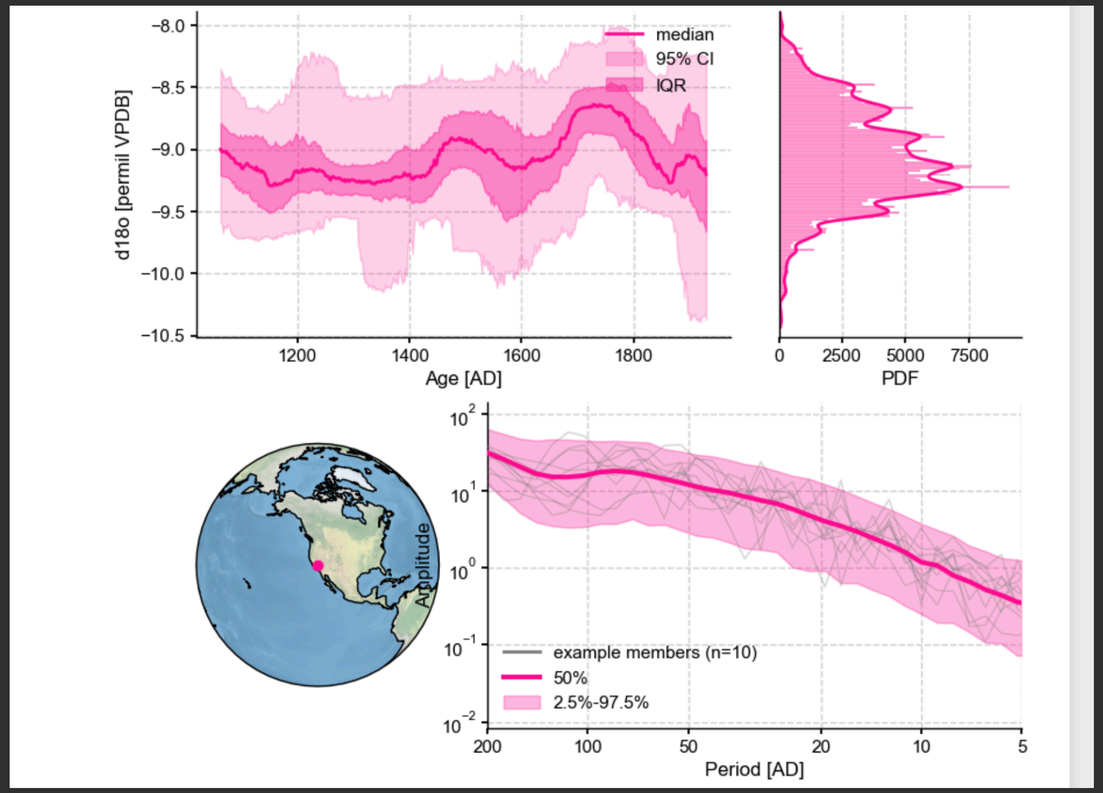
```
In [2]: d_cave = pyleo.Lipd('../data/Crystal.McCabe-Glynn.2013.lpd')

        Disclaimer: LiPD files may be updated and modified to adhere to standards

        reading: Crystal.McCabe-Glynn.2013.lpd
        1.42 MB :That's a big file! This may take a while to load...
        Finished read: 1 record


        Let's load the d18O record.

In [3]: ts = d_cave.to_LipdSeries()

        extracting paleoData...
        extracting: Crystal.McCabe-Glynn.2013
        Created time series: 3 entries
        0 :  Crystal.McCabe-Glynn.2013 :  speleothem :  depth
        1 :  Crystal.McCabe-Glynn.2013 :  speleothem :  age
        2 :  Crystal.McCabe-Glynn.2013 :  speleothem :  d18o
        Enter the number of the variable you wish to use: 2
```

Pyleoclim

What is it?

- ✦ Python package for the analysis and visualization of paleoclimate data
- ✦ The package works with data stored in the LiPD format.
- ✦ Data transformation is made automatically within the package

```
ts.dashboard(metadata=False, ensemble=True, D=d_cave)
```



Pyleoclim

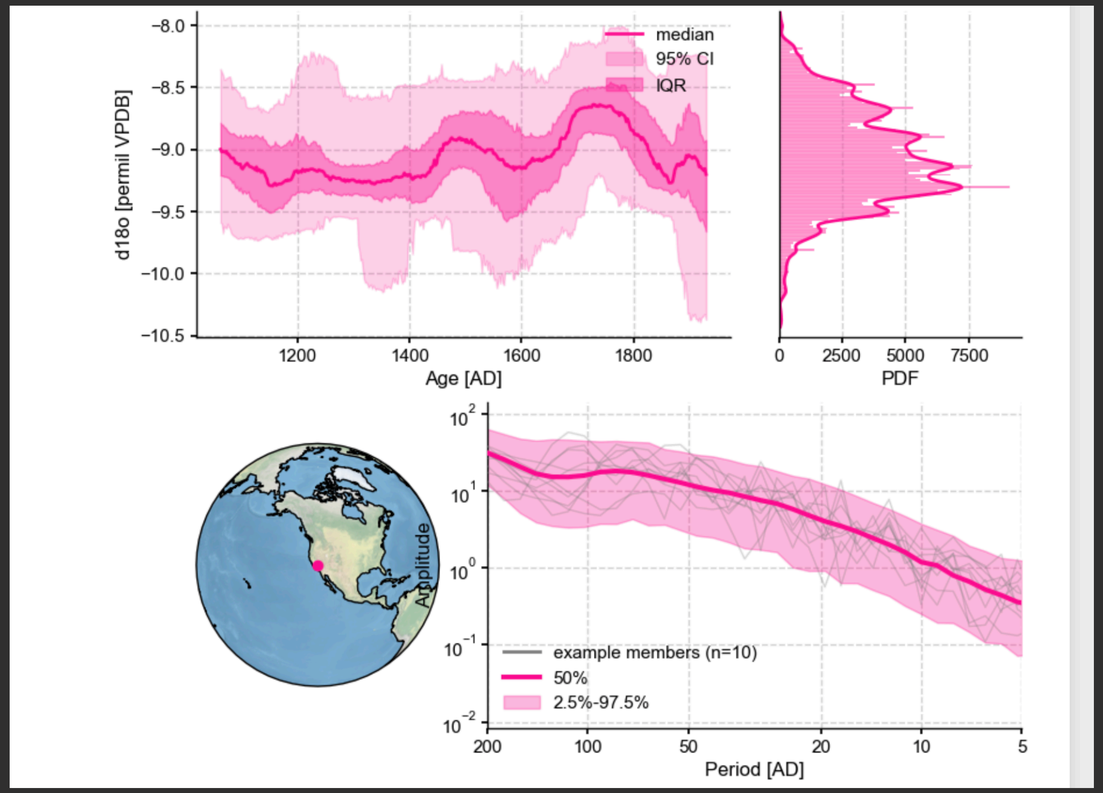
What is it?

- ✦ Python package for the analysis and visualization of paleoclimate data
- ✦ The package works with data stored in the LiPD format.
- ✦ Data transformation is made automatically within the package

What does it do?

- ✦ Mapping
- ✦ Plotting
- ✦ Timeseries analysis
  - ❖ Spectral
  - ❖ Wavelet
  - ❖ Causality
  - ❖ Pre-processing

```
ts.dashboard(metadata=False, ensemble=True, D=d_cave)
```



Pyleoclim

What is it?

- ✦ Python package for the analysis and visualization of paleoclimate data
- ✦ The package works with data stored in the LiPD format.
- ✦ Data transformation is made automatically within the package

What does it do?

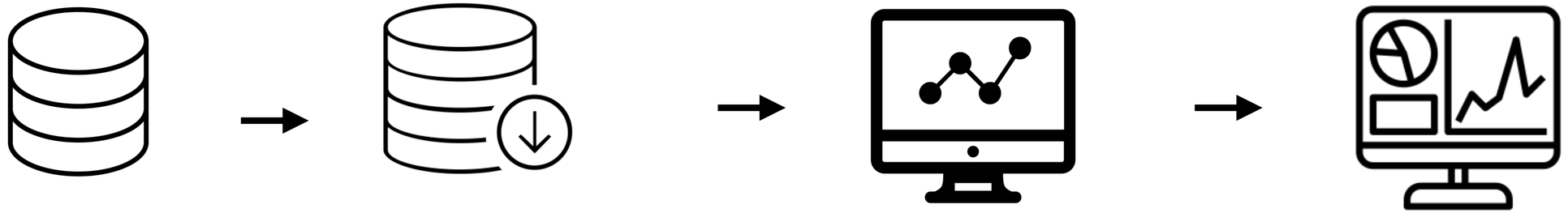
- ✦ Mapping
- ✦ Plotting
- ✦ Timeseries analysis
  - ❖ Spectral
  - ❖ Wavelet
  - ❖ Causality
  - ❖ Pre-processing

Where can I get it?

Github: LinkedEarth

Pypi

https://pyleoclim-util.readthedocs.io/en/master/



Query Data Catalog

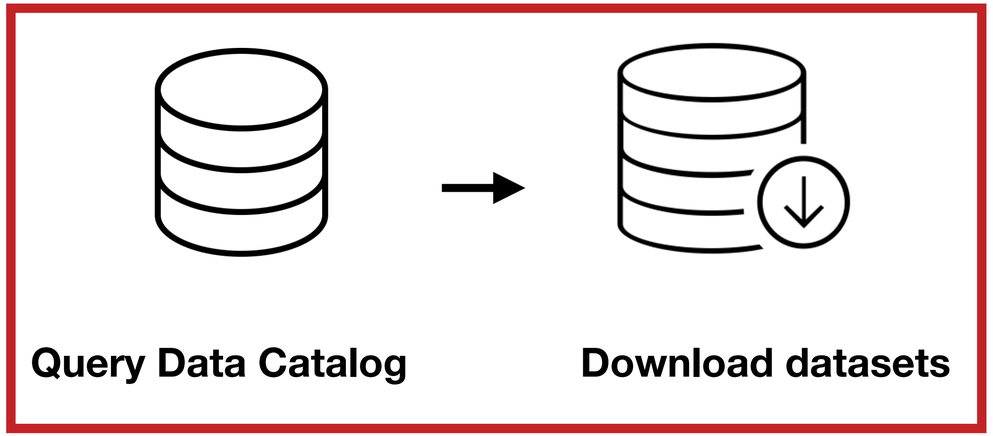
Download datasets

Spectral Analysis

Visualize and Interpret Results

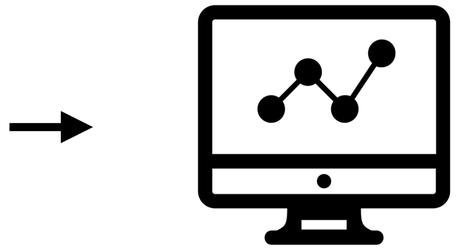
https://github.com/LinkedEarth/paleoHackathon

Icons: The Noun Project



Query Data Catalog

Download datasets



Spectral Analysis

Visualize and Interpret Results

Query

First, we will query the LinkedEarth database for marine sedimentary records recording sea surface temperature spanning the Holocene epoch. The database has a SPARQL endpoint, which is directly accessible through the Pyleoclim package.

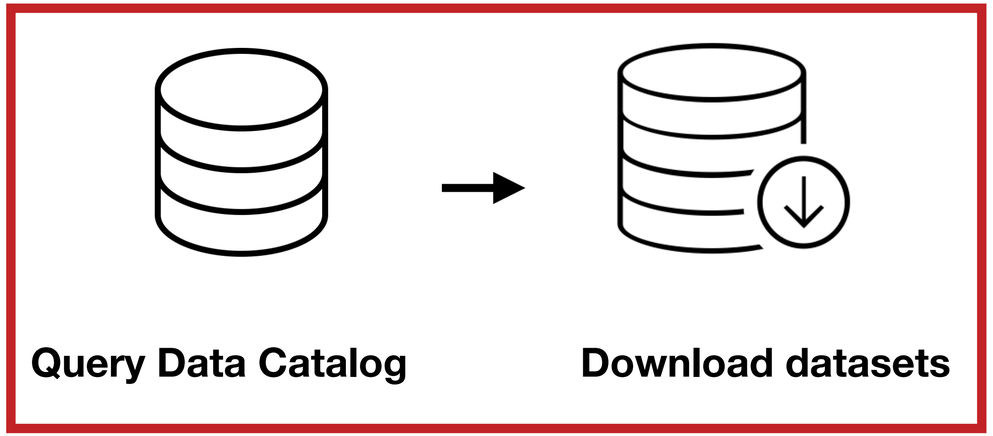
In [2]:

```
# variables and parameters
archiveType=["marine sediment","Marine Sediment"]
proxyObsType=[ ]
infVarType=["Sea Surface Temperature"]
sensorGenus=[ ]
sensorSpecies=[ ]
interpName=["temperature","Temperature"]
interpDetail=["sea surface"]
ageUnits=["kyr BP"]
ageBound=[0,10]
ageBoundType=["any"]
recordLength=[4]
resolution=[ ]
lat=[-14,1.5]
lon=[110,135]
alt=[-10000,0]
#functions
res = pyleo.utils.queryLinkedEarth(archiveType=archiveType, proxyObsType=proxyObsTy
pe,
                                   infVarType = infVarType, sensorGenus=sensorGenus,
                                   sensorSpecies=sensorSpecies, interpName=interpName,
                                   interpDetail =interpDetail, ageUnits = ageUnits,
                                   ageBound = ageBound, ageBoundType = ageBoundType,
                                   recordLength = recordLength, resolution = resolution,
                                   lat = lat, lon = lon, alt = alt,
                                   print_response = True, download_lipd = True,
                                   download_folder = './lipd')
#outputs
```

```
http://wiki.linked.earth/Special:URIResolver/MD01-2D2378.Xu.2008
http://wiki.linked.earth/Special:URIResolver/GeoB10069-2D3.Gibbons.2014
```

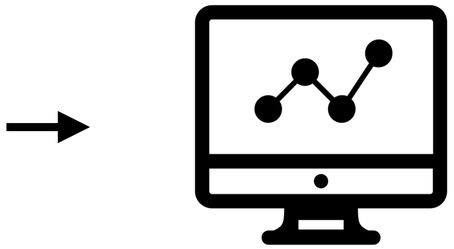
https://github.com/LinkedEarth/paleoHackathon

Icons: The Noun Project



Query Data Catalog

Download datasets



Spectral Analysis

Visualize and Interpret Results

Query

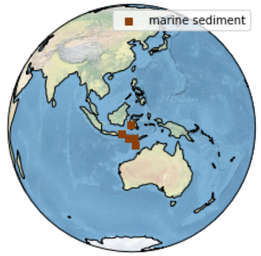
First, we will query the LinkedEarth database for marine sedimentary records recording sea surface temperature spanning the Holocene epoch. The database has a SPARQL endpoint, which is directly accessible through the Pyleoclim package.

In [2]:
```
# variables and parameters
archiveType=["marine sediment","Marine Sediment"]
proxyObsType=[ ]
infVarType=["Sea Surface Temperature"]
sensorGenus=[ ]
sensorSpecies=[ ]
interpName=["temperature","Temperature"]
interpDetail=["sea surface"]
ageUnits=["kyr BP"]
ageBound=[0,10]
ageBoundType=["any"]
recordLength=[4]
resolution=[ ]
lat=[-14,1.5]
lon=[110,135]
alt=[-10000,0]
#functions
res = pyleo.utils.queryLinkedEarth(archiveType=archiveType, proxyObsType=proxyObsTy
pe,
                                   infVarType = infVarType, sensorGenus=sensorGenus,
                                   sensorSpecies=sensorSpecies, interpName=interpName,
                                   interpDetail =interpDetail, ageUnits = ageUnits,
                                   ageBound = ageBound, ageBoundType = ageBoundType,
                                   recordLength = recordLength, resolution = resolution,
                                   lat = lat, lon = lon, alt = alt,
                                   print_response = True, download_lipd = True,
                                   download_folder = './lipd')
#outputs
```

```
http://wiki.linked.earth/Special:URIResolver/MD01-2D2378.Xu.2008
http://wiki.linked.earth/Special:URIResolver/GeoB10069-2D3.Gibbons.2014
```

In [9]:
```
# variables and parameters
#functions
D = pyleo.Lipd(usr_path=os.getcwd())
D.mapAllArchive(projection='Orthographic', proj_default={'central_longitude':120, '
central_latitude':-5})
#outputs
```

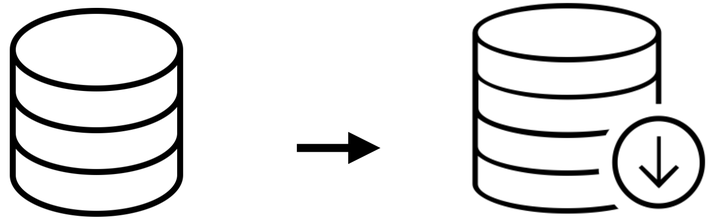
```
Disclaimer: LiPD files may be updated and modified to adhere to standards

Found: 5 LiPD file(s)
reading: BJ8-03-13GGC.Linsley.2010.lpd
reading: GeoB10069-3.Gibbons.2014.lpd
reading: MD98-2165.Levi.2007.lpd
reading: BJ8-03-70GGC.Linsley.2010.lpd
reading: MD01-2378.Xu.2008.lpd
Finished read: 5 records
```



Out[9]:
```
(<Figure size 432x288 with 1 Axes>,
 <cartopy.mpl.geoaxes.GeoAxesSubplot at 0x1a29230e48>)
```

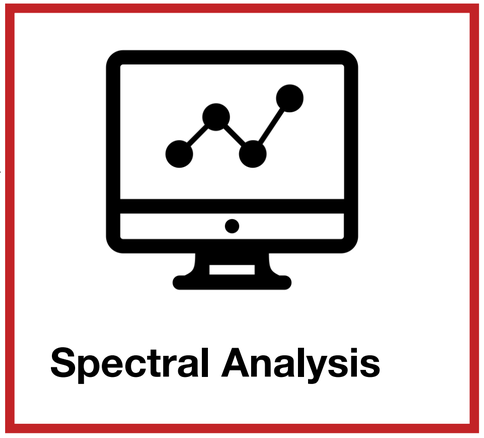
https://github.com/LinkedEarth/paleoHackathon

Icons: The Noun Project



Query Data Catalog

Download datasets



Spectral Analysis

Visualize and Interpret Results

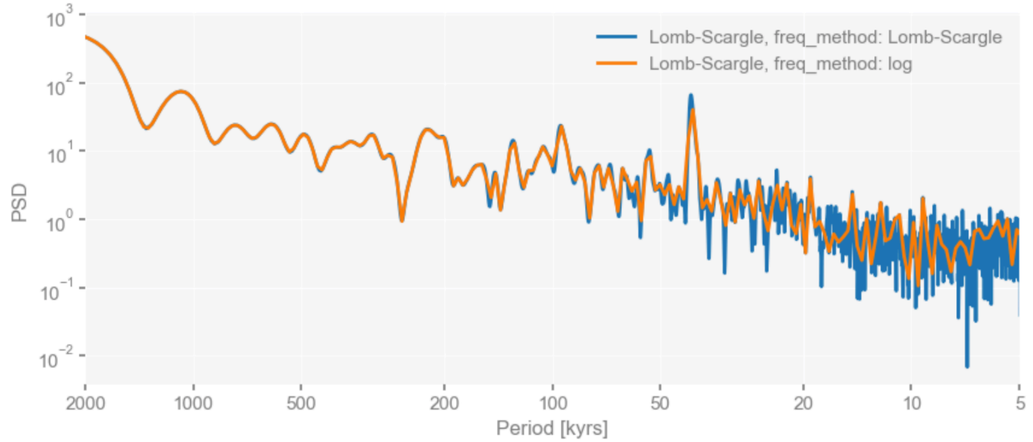
Query

First, we will query the LinkedEarth database for marine sedimentary records recording sea surface temperature spanning the Holocene epoch. The database has a SPARQL endpoint, which is directly accessible through the Pyleoclim package.

In [2]:
```
# variables and parameters
archiveType=["marine sediment","Marine Sediment"]
proxyObsType=[ ]
infVarType=["Sea Surface Temperature"]
sensorGenus=[ ]
sensorSpecies=[ ]
interpName=["temperature","Temperature"]
interpDetail=["sea surface"]
ageUnits=["kyr BP"]
ageBound=[0,10]
ageBoundType=["any"]
recordLength=[4]
resolution=[ ]
lat=[-14,1.5]
lon=[110,135]
alt=[-10000,0]
#functions
res = pyleo.utils.queryLinkedEarth(archiveType=archiveType, proxyObsType=proxyObsTy
pe,
                                   infVarType = infVarType, sensorGenus=sensorGenus,
                                   sensorSpecies=sensorSpecies, interpName=interpName,
                                   interpDetail =interpDetail, ageUnits = ageUnits,
                                   ageBound = ageBound, ageBoundType = ageBoundType,
                                   recordLength = recordLength, resolution = resolution,
                                   lat = lat, lon = lon, alt = alt,
                                   print_response = True, download_lipd = True,
                                   download_folder = './lipd')
#outputs
```

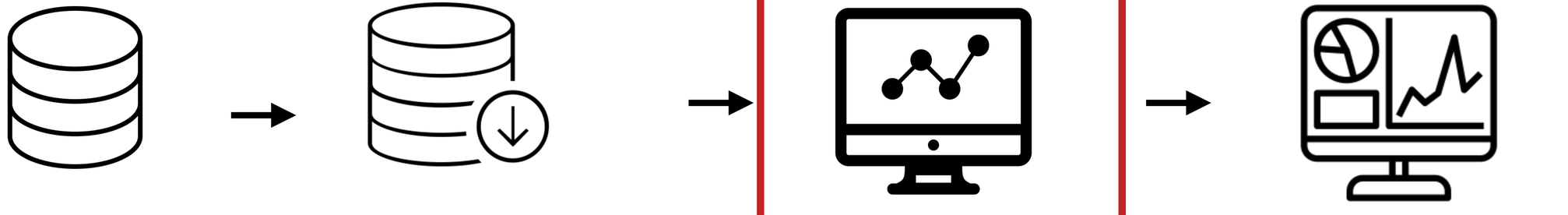
```
http://wiki.linked.earth/Special:URIResolver/MD01-2D2378.Xu.2008
http://wiki.linked.earth/Special:URIResolver/GeoB10069-2D3.Gibbons.2014
```

In [8]:
```
psd_ls_ls = ts.standardize().spectral(method='lomb_scargle',freq_method = 'lomb_scargle')

fig, ax = psd_ls_ls.plot(label='Lomb-Scargle, freq_method: Lomb-Scargle', mute = True)
psd_ls.plot(label='Lomb-Scargle, freq_method: log', mute = True, ax=ax)
pyleo.showfig(fig)
```



https://github.com/LinkedEarth/paleoHackathon

Icons: The Noun Project



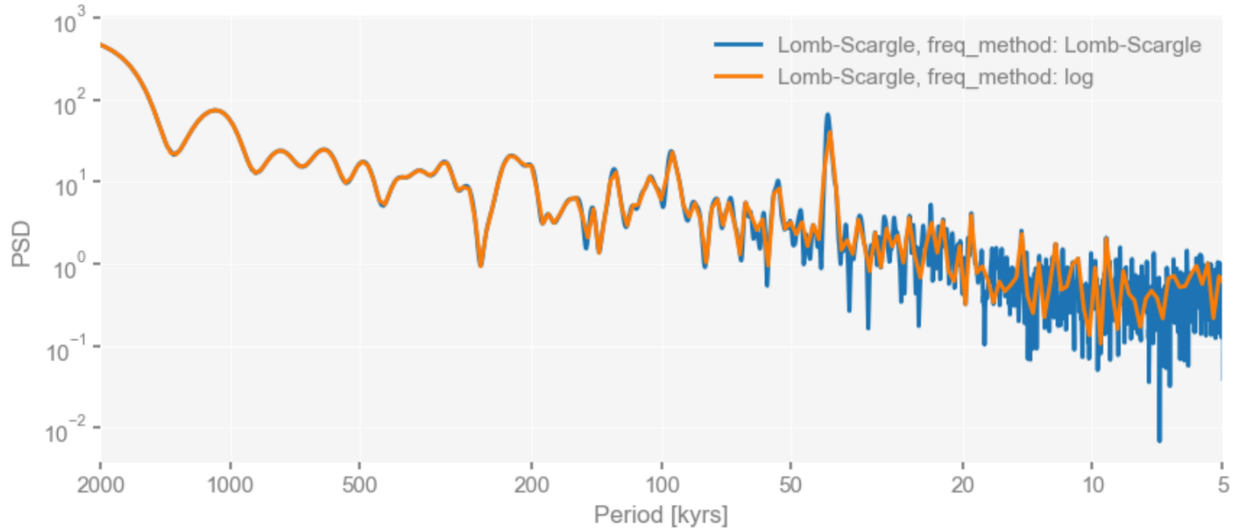
Query Data Catalog

Download datasets

Spectral Analysis

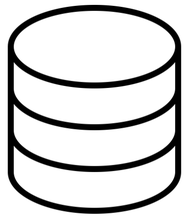
Visualize and Interpret Results

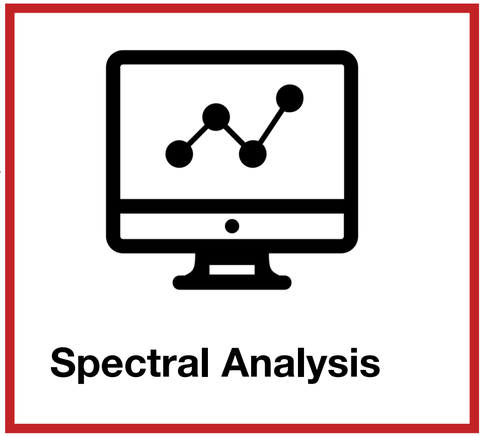
```
In [8]: psd_ls_ls = ts.standardize().spectral(method='lomb_scargle',freq_method = 'lomb_scargle')

        fig, ax = psd_ls_ls.plot(label='Lomb-Scargle, freq_method: Lomb-Scargle', mute = True)
        psd_ls.plot(label='Lomb-Scargle, freq_method: log', mute = True, ax=ax)
        pyleo.showfig(fig)
```



https://github.com/LinkedEarth/paleoHackathon

Icons: The Noun Project

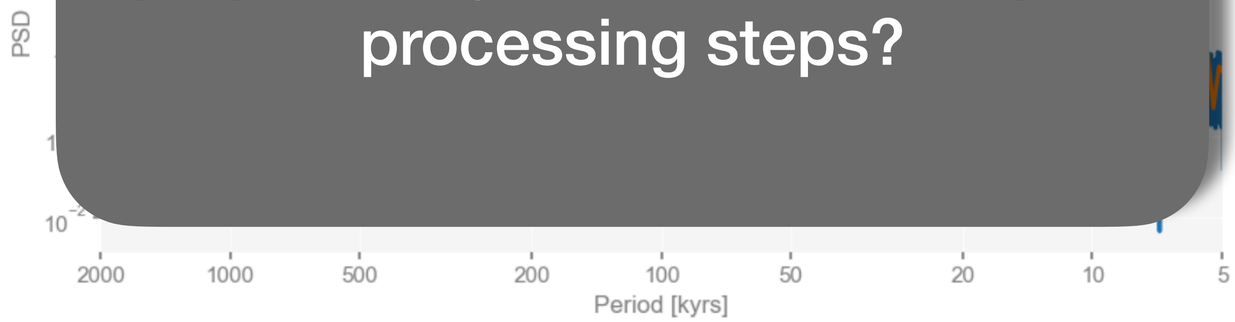




Query Data Catalog → **Download datasets** → **Spectral Analysis** → **Visualize and Interpret Results**

```
In [8]: psd_ls_ls = ts.standardize().spectral(method='lomb_scargle',freq_method = 'lomb_scargle')

        fig, ax = psd_ls_ls.plot(label='Lomb-Scargle, freq method: Lomb-Scargle', mute = True)
        psd_ls
        pyleo
```

Can we train an AI to choose the proper analysis methods and pre-processing steps?



https://github.com/LinkedEarth/paleoHackathon

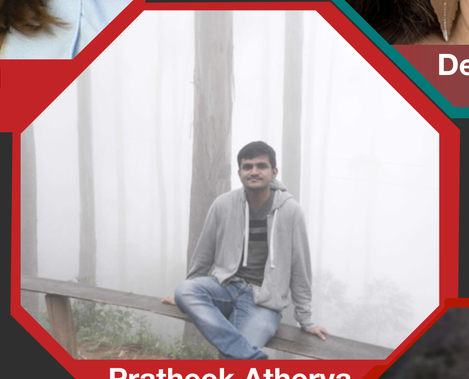
Icons: The Noun Project

autoTS: Automated Machine Learning for Time Series Analysis

Yolanda Gil
USC

Deborah Khider
USC

Feng Zhu
USC

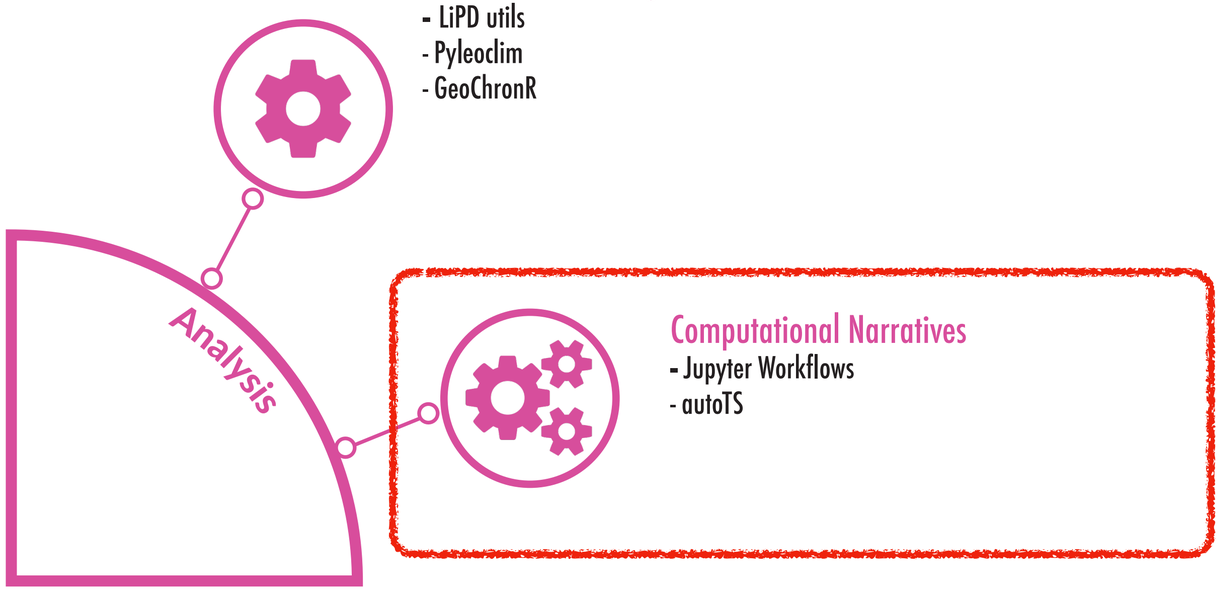


Pratheek Atherya
USC

Myron Kwan
USC

Varun Ratnakar
USC

J.P.Morgan



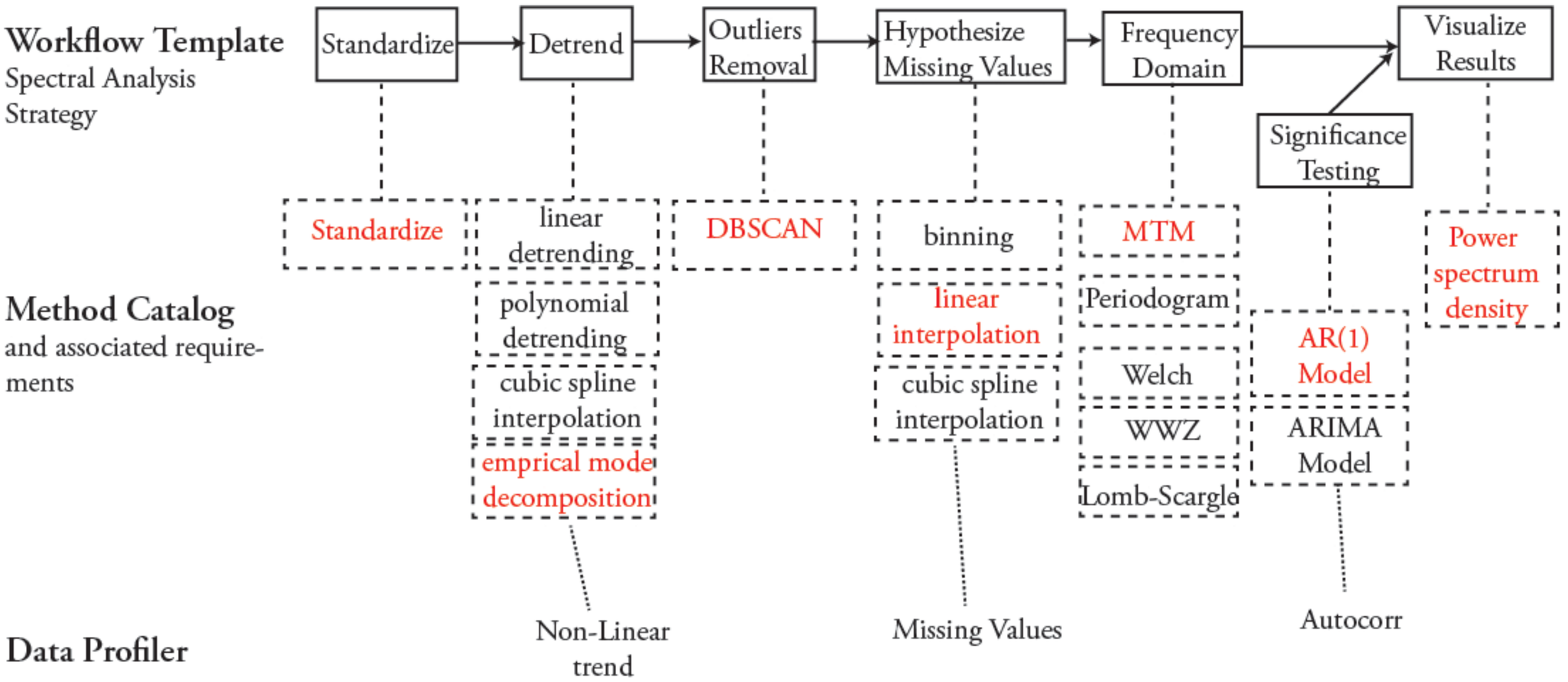
Open Source Packages

- LiPD utils
- Pyleoclim
- GeoChronR

Computational Narratives

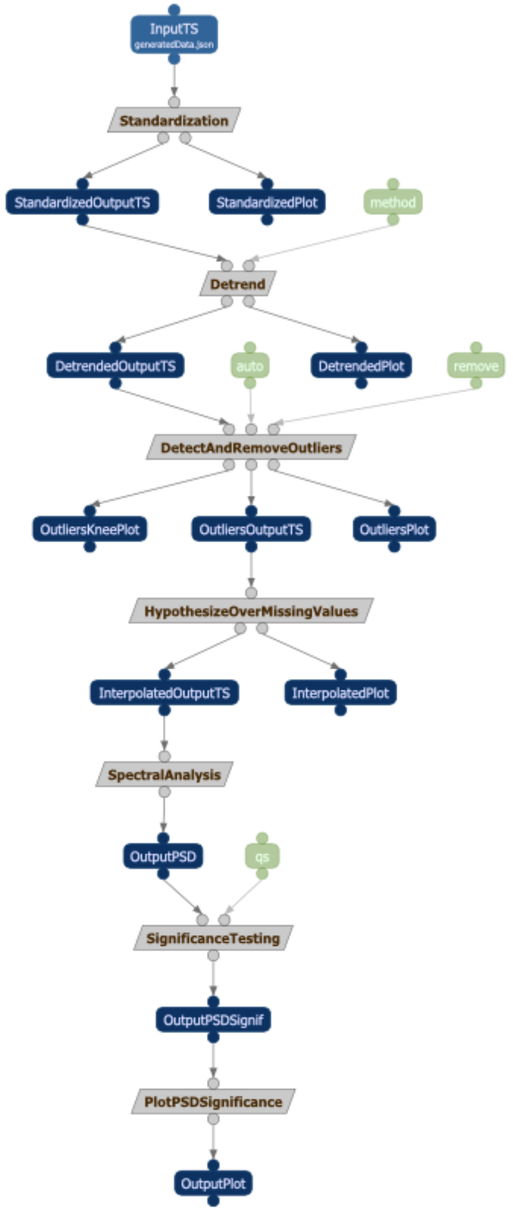
- Jupyter Workflows
- autoTS

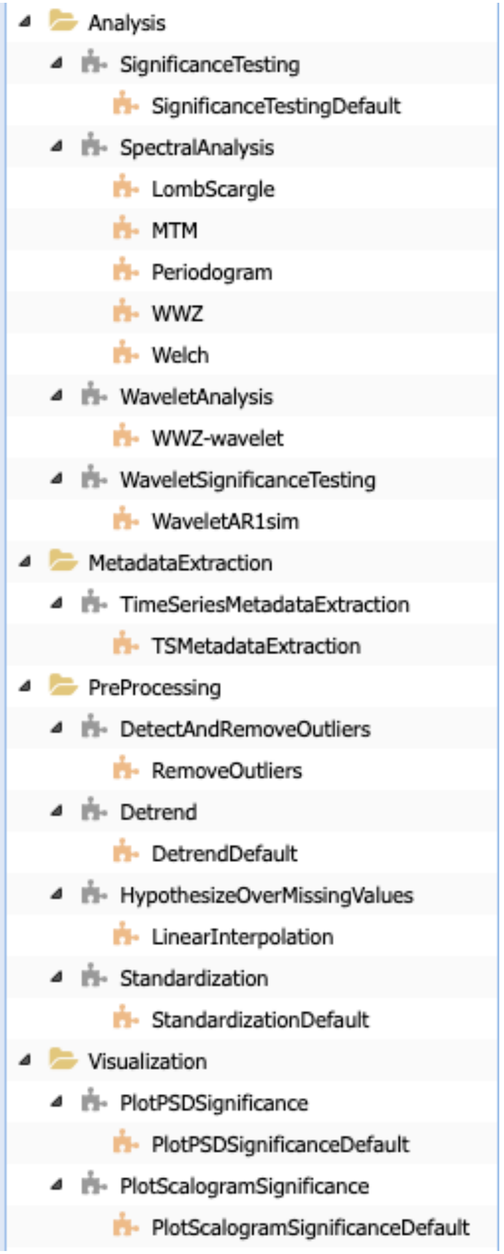
Strategy



Khider et al., 2020, MiLeTS

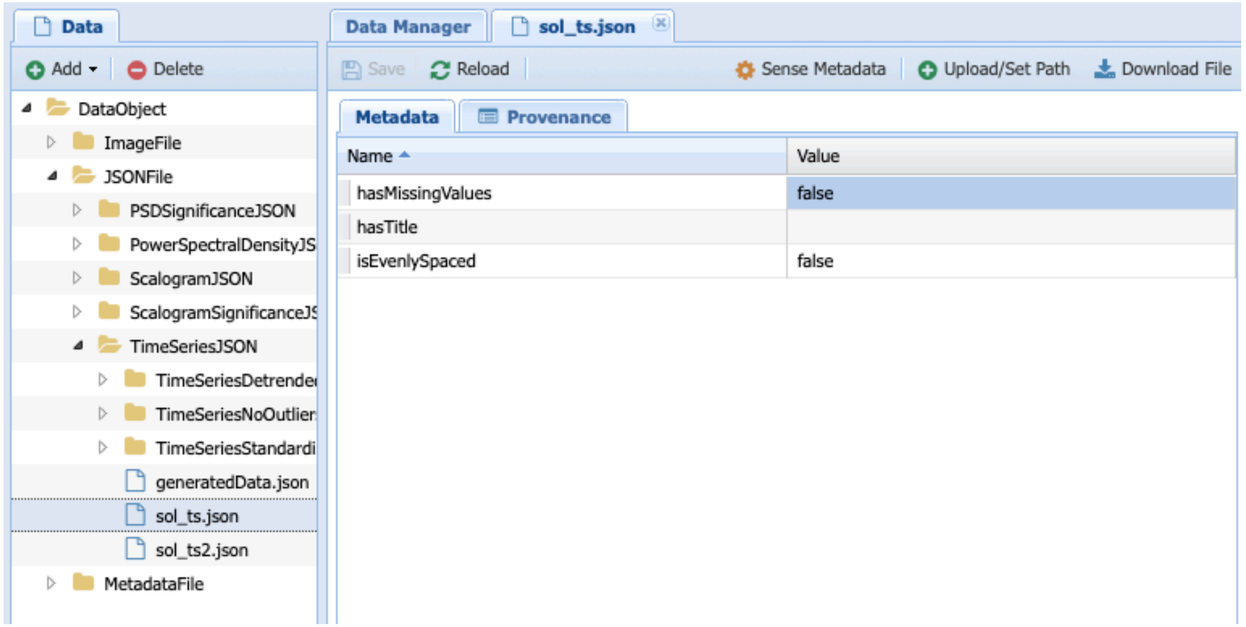
WINGS can capture all the steps involved in spectral analysis in a abstract workflow template





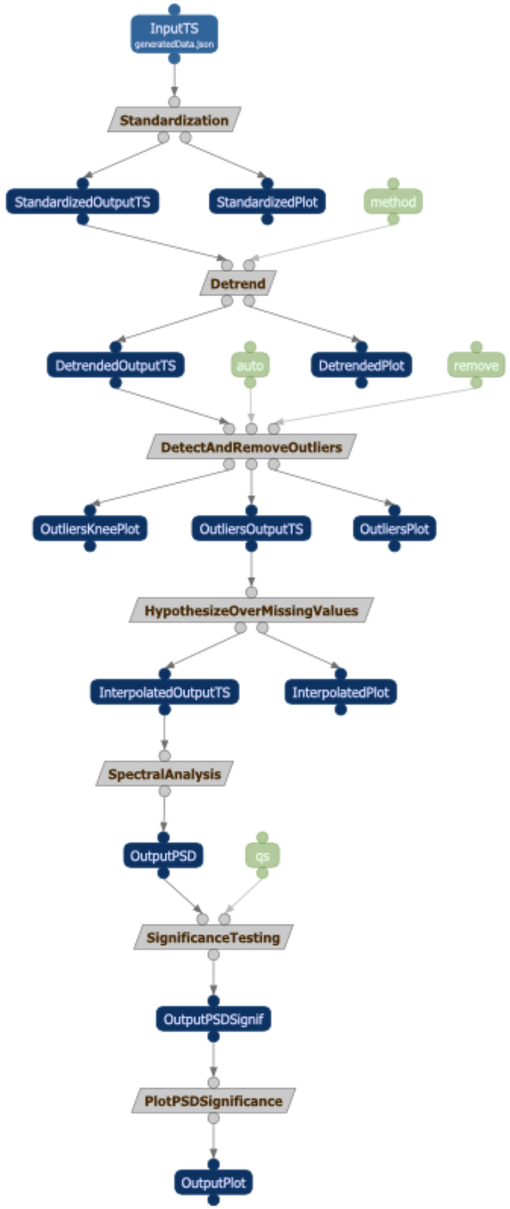
Khider et al., 2020, MiLeTS

Each dataset has metadata properties that reflect its characteristics, and are used to express constraints that result in the generation of the most effective workflows



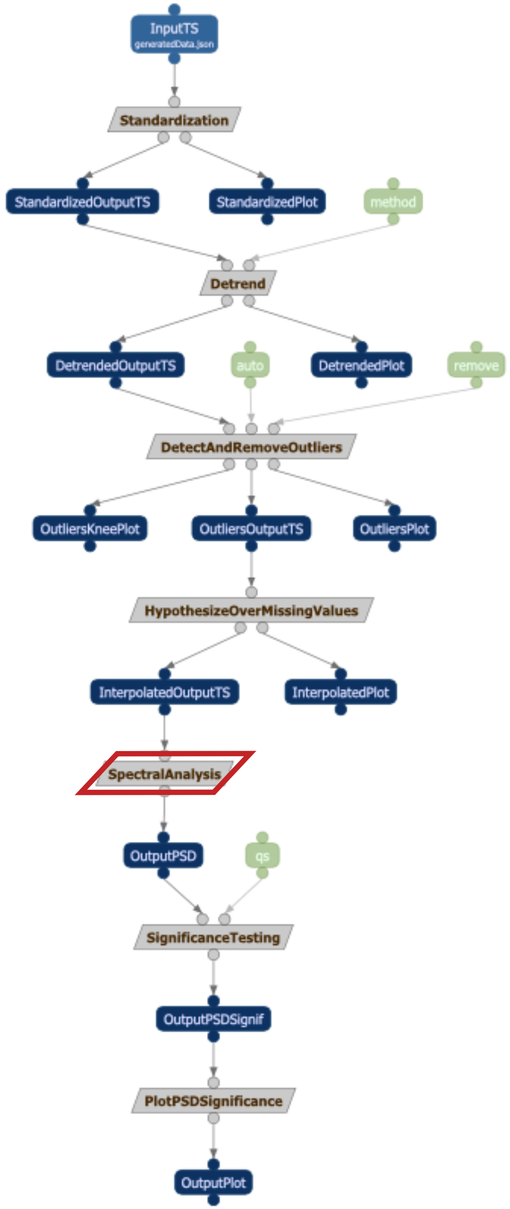
Khider et al., 2020, MiLeTS

WINGS was extended to allow for a type of constraints that skips an abstract workflow step



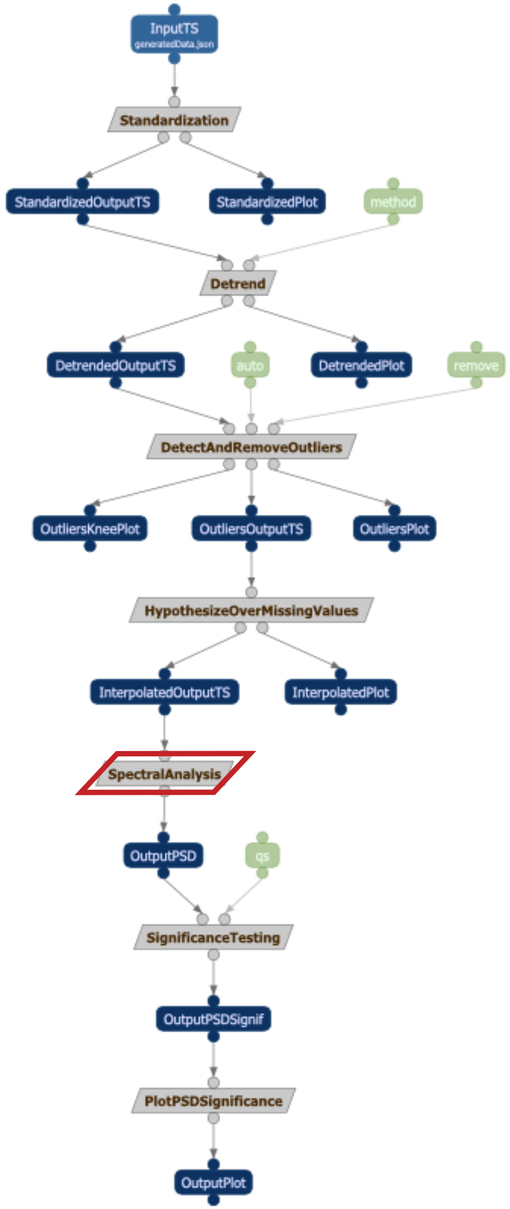
Khider et al., 2020, MiLeTS

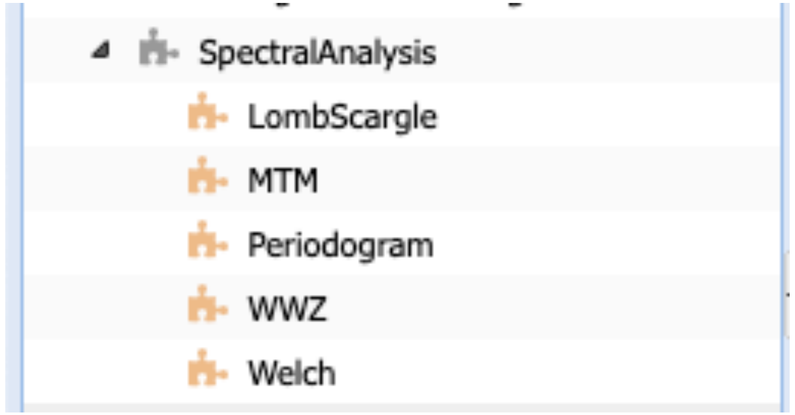
WINGS was extended to allow for a type of constraints that skips an abstract workflow step



Khider et al., 2020, MiLeTS

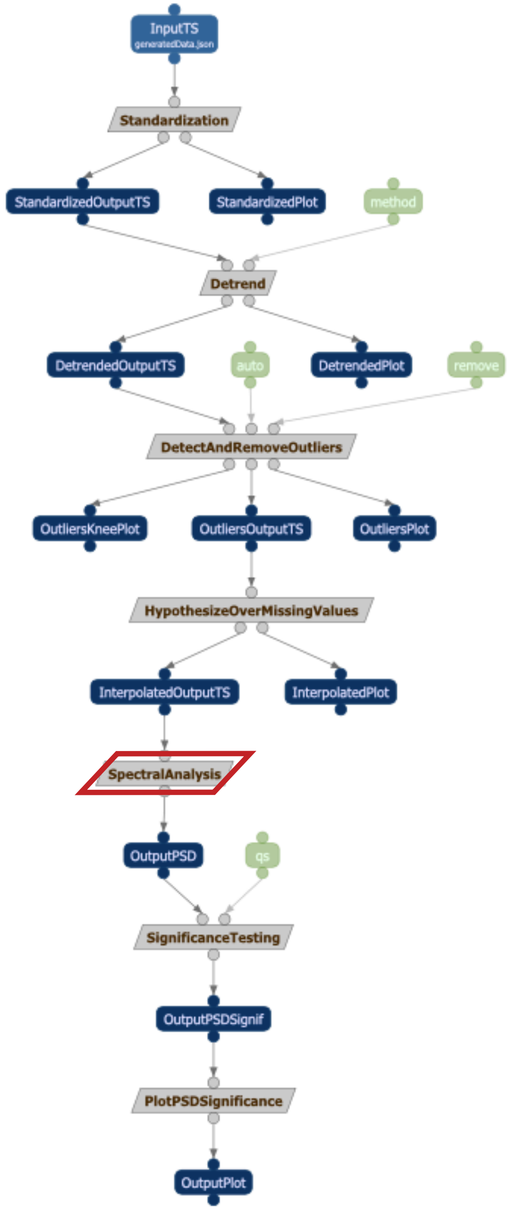
WINGS was extended to allow for a type of constraints that skips an abstract workflow step

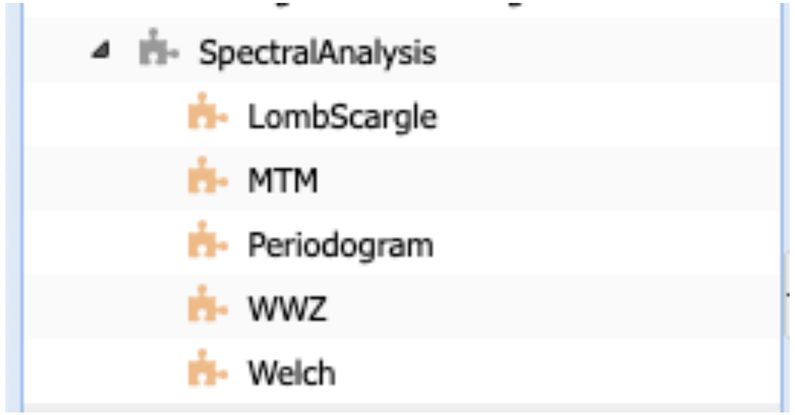




Khider et al., 2020, MiLeTS

WINGS was extended to allow for a type of constraints that skips an abstract workflow step



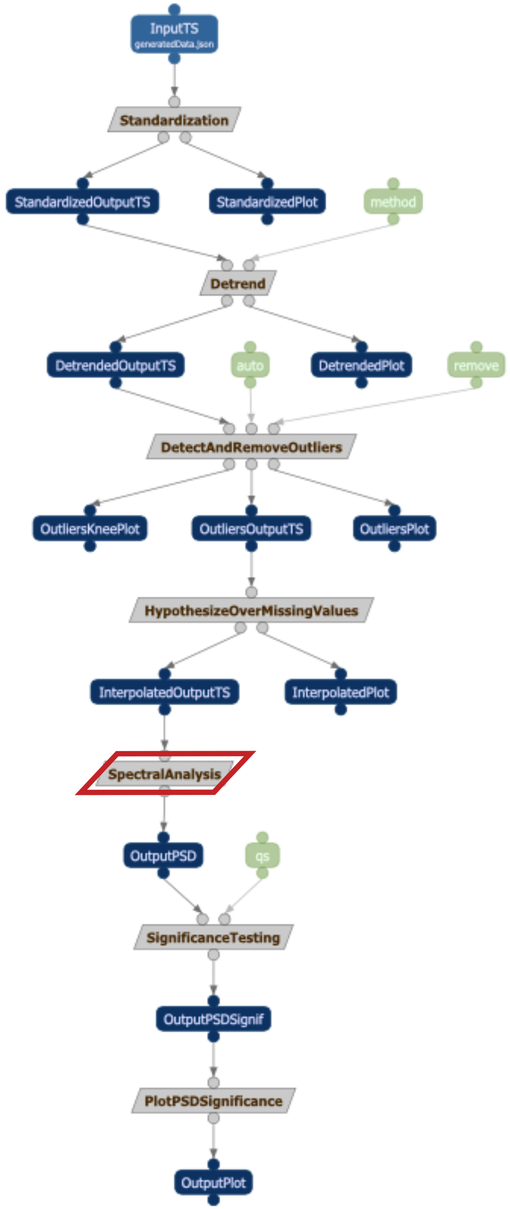


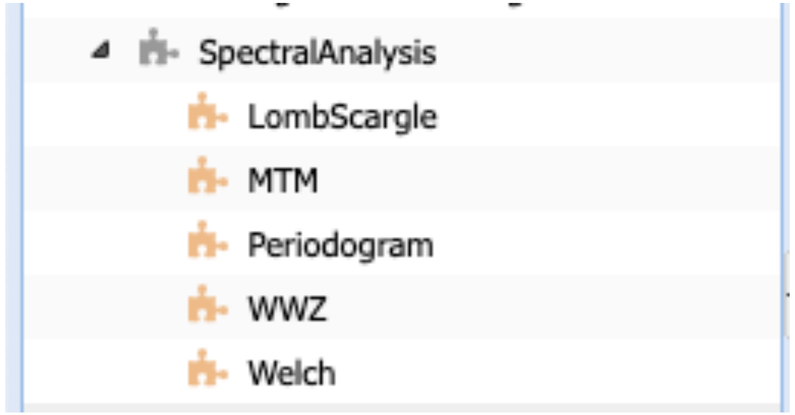
Add Rule

```
[ RequirementOfInputToBeEvenlySpaced:
  (?c rdf:type acdom:MTMClass)
  (?c ac:hasInput ?InputTS)
  (?InputTS ac:hasArgumentID 'InputTS')
  -> (?InputTS dcdom:isEvenlySpaced 'true'^^xsd:boolean)
  print(?c 'MTM requires evenly spaced data')
]
```

Khider et al., 2020, MiLeTS

WINGS was extended to allow for a type of constraints that skips an abstract workflow step





Add Rule

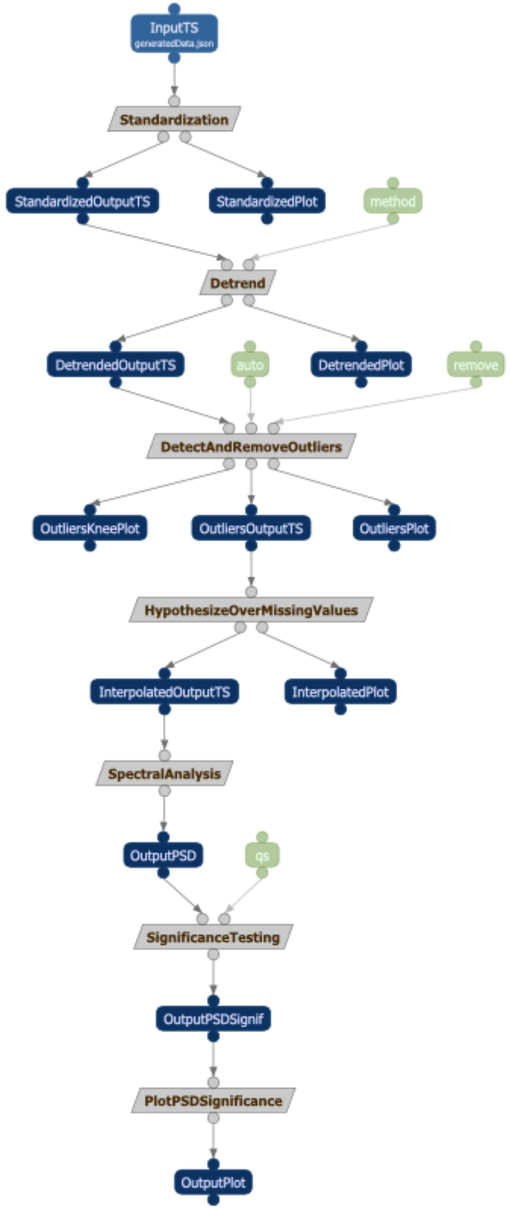
```
[ RequirementOfInputToBeEvenlySpaced:
  (?c rdf:type acdom:MTMClass)
  (?c ac:hasInput ?InputTS)
  (?InputTS ac:hasArgumentID 'InputTS')
  -> (?InputTS dcdom:isEvenlySpaced 'true'^^xsd:boolean)
  print(?c 'MTM requires evenly spaced data')
]
```

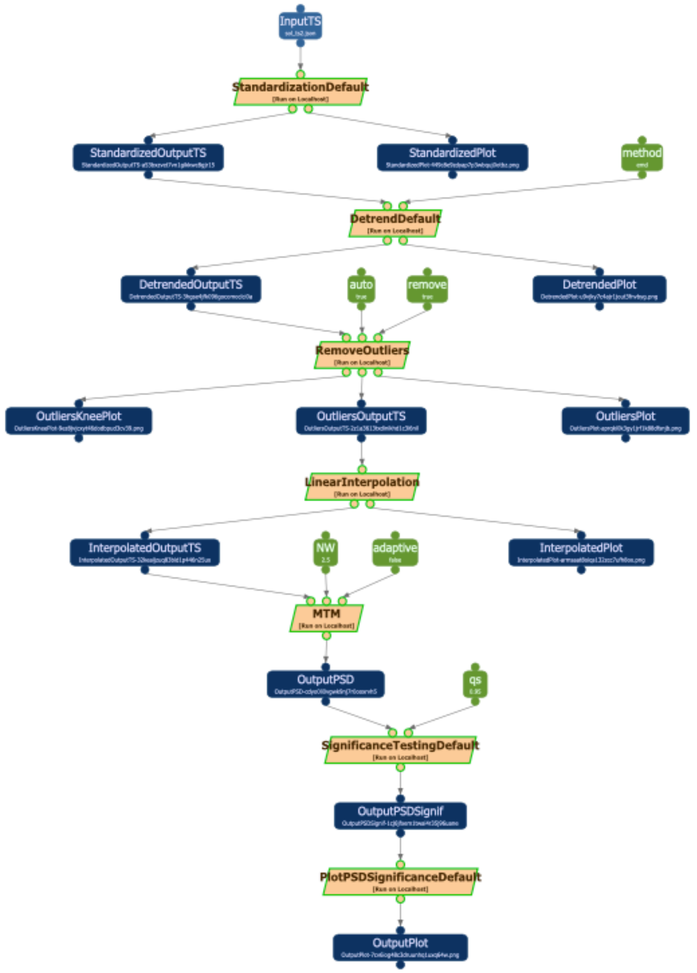
Add Rule

```
[ NoRequirementOfInputToBeEvenlySpaced:
  (?c rdf:type acdom:WWZClass)
  (?c ac:hasInput ?InputTS)
  (?InputTS ac:hasArgumentID 'InputTS')
  -> (?InputTS dcdom:isEvenlySpaced 'false'^^xsd:boolean)
  print(?c 'WWZ does not require evenly spaced data')
]
```

Khider et al., 2020, MiLeTS

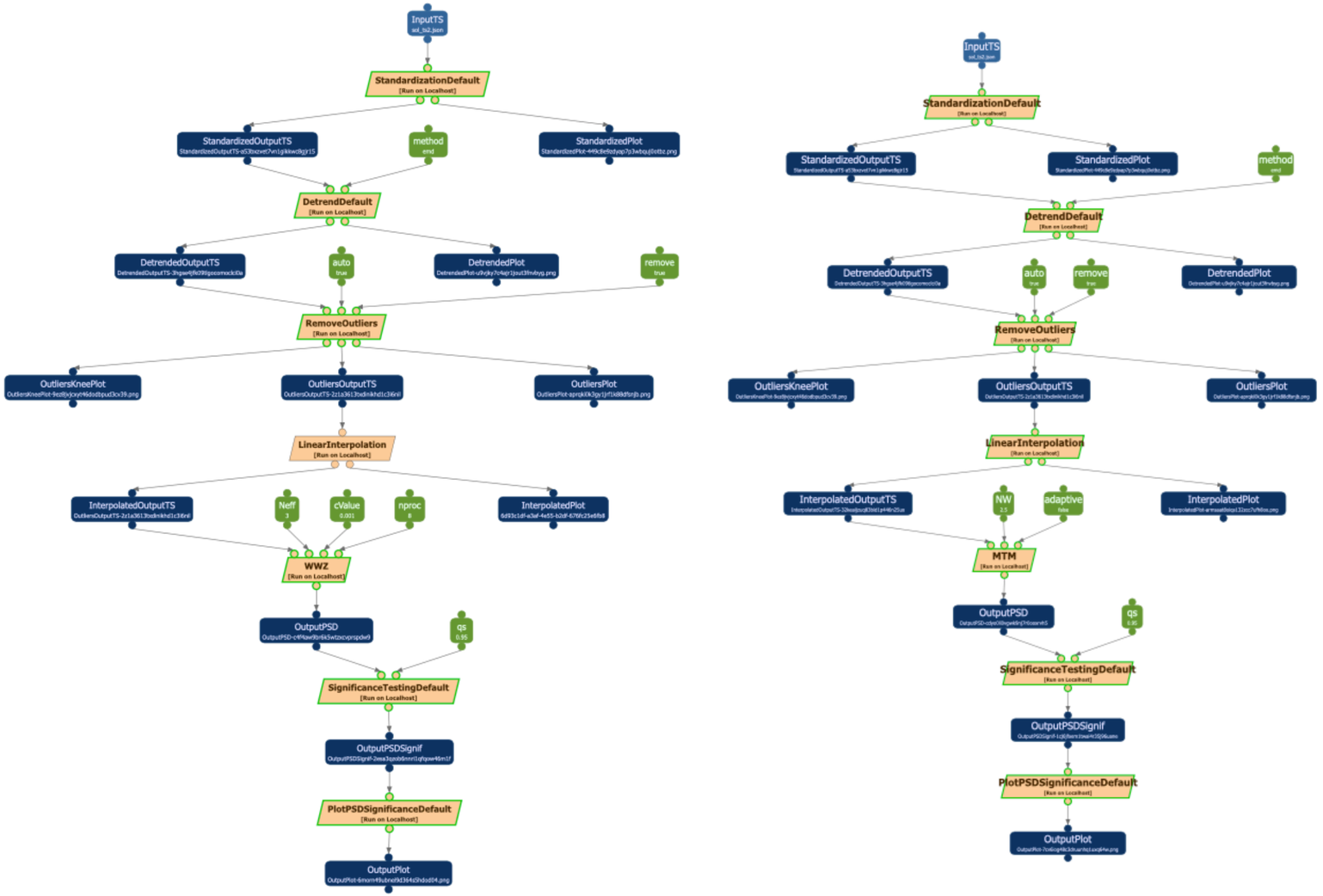
WINGS was extended to allow for a type of constraints that skips an abstract workflow step





Khider et al., 2020, MiLeTS

WINGS was extended to allow for a type of constraints that skips an abstract workflow step



Khider et al., 2020, MiLeTS

Welcome to the Wings Portal

This portal provides access to the **Wings semantic workflow system** by allowing you to browse, set up, and run workflows. Some unique features of Wings that can assist a user include:

- Choosing datasets that are valid for a workflow you have selected, based on the metadata properties of the datasets and the semantic constraints of that workflow
- Setting parameter values that are appropriate for the datasets you have selected, based on the semantic constraints of the workflow steps
- Handling collections of data in a compact way, based on compact collection reasoning in the workflow templates
- Specializing abstract workflow steps, based on the constraints imposed by both datasets and other workflow steps

You can walk through some Wings Tutorials and Videos that guide you through the basics of creating and using workflows.

About Wings

Workflow systems can automate many aspects of workflow creation and execution. The WINGS workflow system developed at USC/ISI illustrates these capabilities, with WINGS providing user assistance and automatic workflow validation and generation. The WINGS portal is setup by default to run locally, but it can use different execution engines such as OODT and Pegasus

WINGS (Workflow INstance Generation and Specialization) uses AI planning and semantic reasoners to assist users in creating workflows while validating that the workflows comply with the requirements of the software components and datasets. WINGS can reason about the constraints of the components and the characteristics of the data and propagate them through the workflow structure. WINGS reasons over *semantic workflow representations* that consist of both a traditional dataflow graph as well as a network of constraints on the data and components of the workflow.

Wings workflow repositories are available through other portal sites installed for use of specific research groups. These collections include workflows for population genomics, for educational student assessment, and for social network analysis. Each of these sites contains datasets and the executable codes for all workflow steps in their collection, so the workflows can submitted for execution at the location where that particular Wings portal installation is set up.

The WINGS workflow system has an open modular design and can be easily integrated with other existing workflow systems and execution frameworks to extend them with semantic reasoning capabilities. We have integrated the Wings semantic workflow system with other user interfaces, and submitted workflows with a variety of execution engines. WINGS is built on open web standards from the World Wide Web Consortium (W3C) such as the Web Ontology Language (OWL), the Resource Description Framework (RDF), and the SPARQL query language for RDF.

If you are interested in using the Wings semantic workflow reasoners with your own workflow framework, or in setting up your own installation of a Wings workflow portal, please contact us.

Publications providing overviews and technical descriptions of how Wings works are available in the Wings web site.

Welcome to the Wings Portal

This portal provides access to the **Wings semantic workflow system** by allowing you to browse, set up, and run workflows. Some unique features of Wings that can assist a user include:

- Choosing datasets that are valid for a workflow you have selected, based on the metadata properties of the datasets and the semantic constraints of that workflow
- Setting parameter values that are appropriate for the datasets you have selected, based on the semantic constraints of the workflow steps
- Handling collections of data in a compact way, based on compact collection reasoning in the workflow templates
- Specializing abstract workflow steps, based on the constraints imposed by both datasets and other workflow steps

You can walk through some Wings Tutorials and Videos that guide you through the basics of creating and using workflows.

About Wings

Workflow systems can automate many aspects of workflow creation and execution. The WINGS workflow system developed at USC/ISI illustrates these capabilities, with WINGS providing user assistance and automatic workflow validation and generation. The WINGS portal is setup by default to run locally, but it can use different execution engines such as OODT and Pegasus

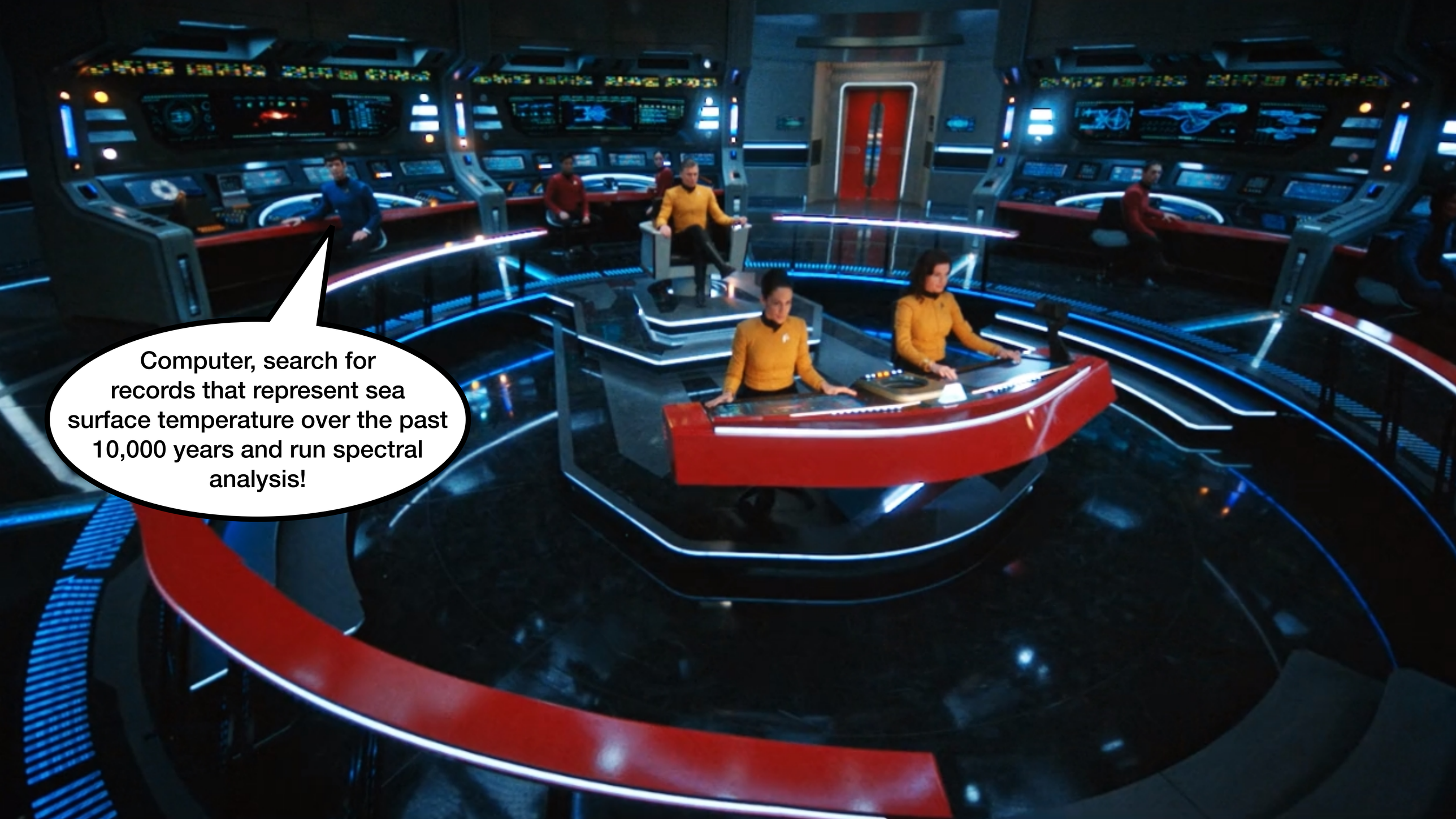
WINGS (Workflow INstance Generation and Specialization) uses AI planning and semantic reasoners to assist users in creating workflows while validating that the workflows comply with the requirements of the software components and datasets. WINGS can reason about the constraints of the components and the characteristics of the data and propagate them through the workflow structure. WINGS reasons over *semantic workflow representations* that consist of both a traditional dataflow graph as well as a network of constraints on the data and components of the workflow.

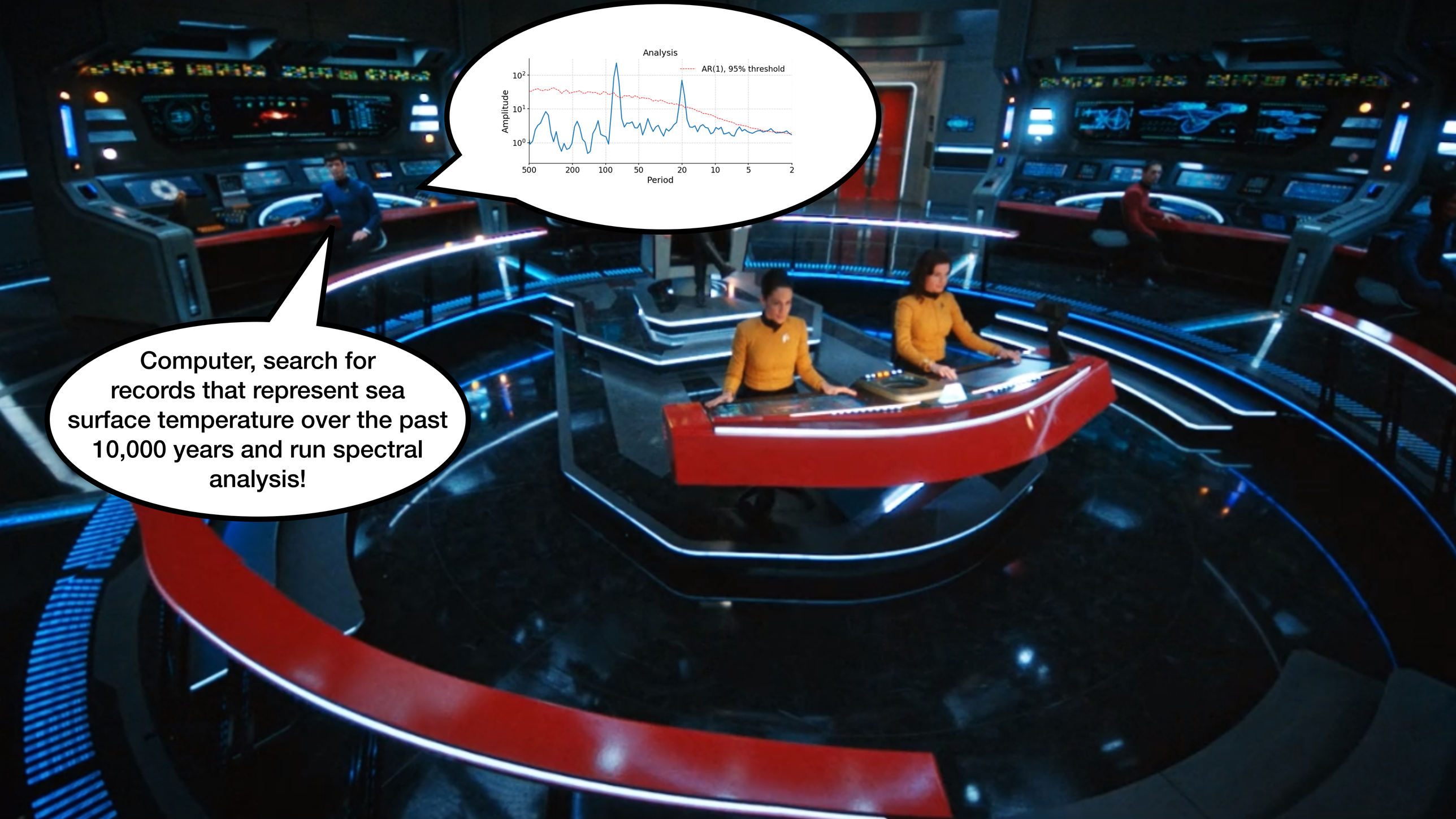
Wings workflow repositories are available through other portal sites installed for use of specific research groups. These collections include workflows for population genomics, for educational student assessment, and for social network analysis. Each of these sites contains datasets and the executable codes for all workflow steps in their collection, so the workflows can submitted for execution at the location where that particular Wings portal installation is set up.

The WINGS workflow system has an open modular design and can be easily integrated with other existing workflow systems and execution frameworks to extend them with semantic reasoning capabilities. We have integrated the Wings semantic workflow system with other user interfaces, and submitted workflows with a variety of execution engines. WINGS is built on open web standards from the World Wide Web Consortium (W3C) such as the Web Ontology Language (OWL), the Resource Description Framework (RDF), and the SPARQL query language for RDF.

If you are interested in using the Wings semantic workflow reasoners with your own workflow framework, or in setting up your own installation of a Wings workflow portal, please contact us.

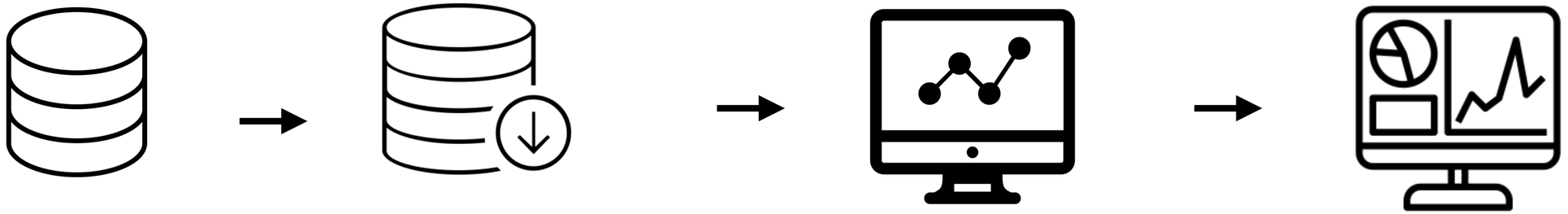
Publications providing overviews and technical descriptions of how Wings works are available in the Wings web site.


Computer, search for records that represent sea surface temperature over the past 10,000 years and run spectral analysis!


Analysis
AR(1), 95% threshold
Amplitude
10^2
10^1
10^0
500
200
100
50
20
10
5
2
Period
Computer, search for records that represent sea surface temperature over the past 10,000 years and run spectral analysis!

I need to make decisions based on timely science results

I need to make decisions based on timely science results



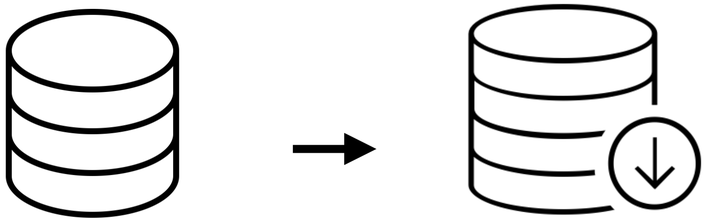
Query Data Catalog

Download datasets

Spectral Analysis

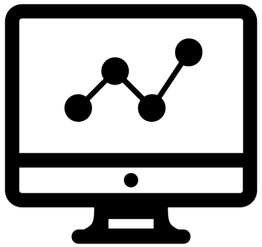
Visualize and Interpret Results

Icons: The Noun Project



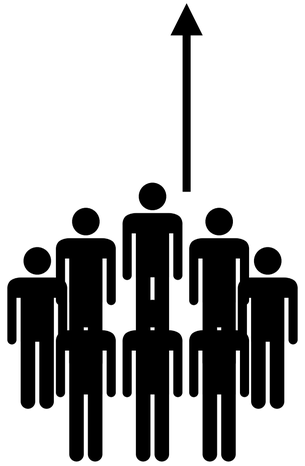
Query Data Catalog

Download datasets

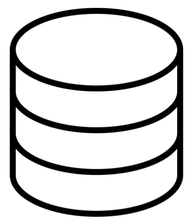


Spectral Analysis

Visualize and Interpret Results

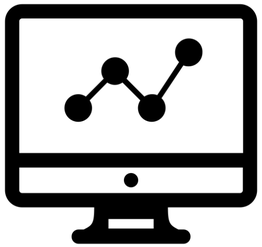


Icons: The Noun Project



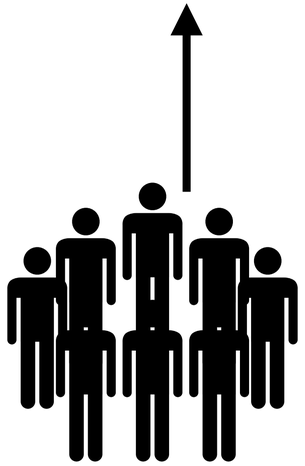
Query Data Catalog

Download datasets



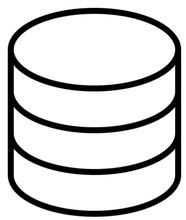
Spectral Analysis

Visualize and Interpret Results



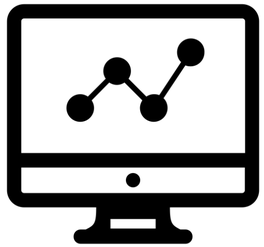
Can we leverage machine learning to help with metadata annotation?

Icons: The Noun Project



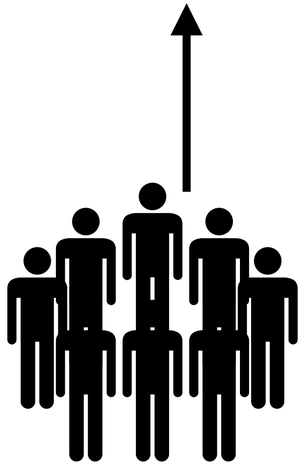
Query Data Catalog

Download datasets



Spectral Analysis

Visualize and Interpret Results

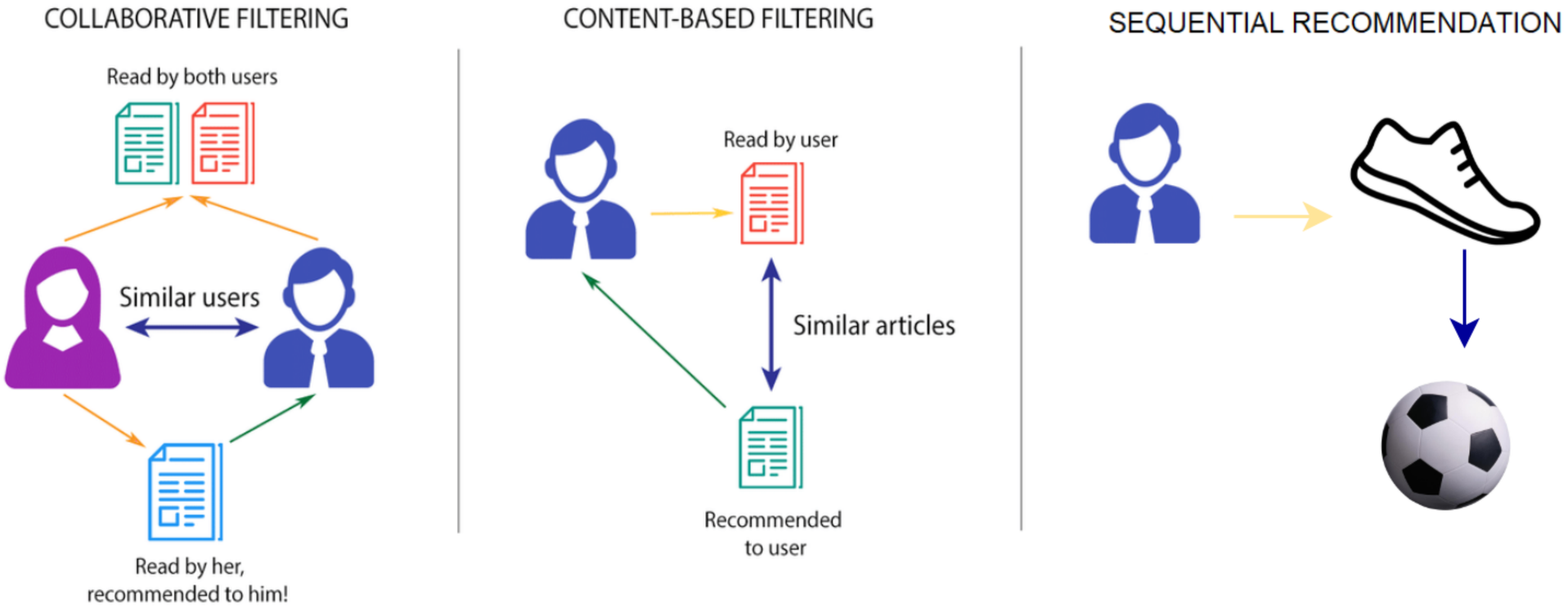


Can we leverage machine learning to help with metadata annotation?

Shravya Manety

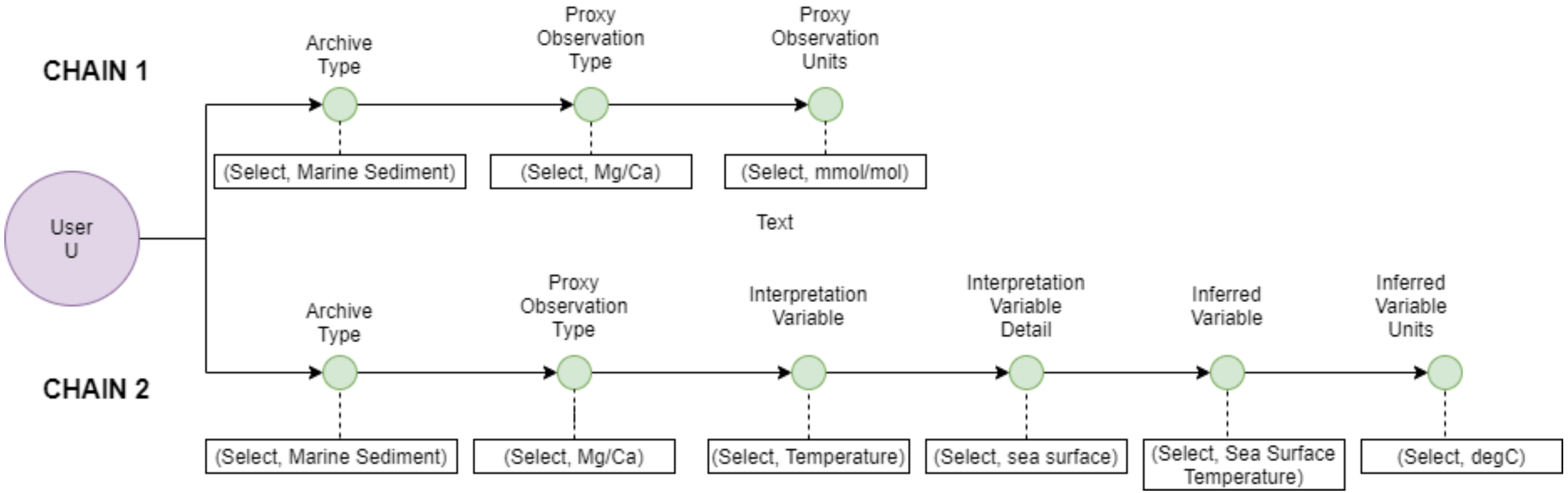
Icons: The Noun Project

Types of recommender systems



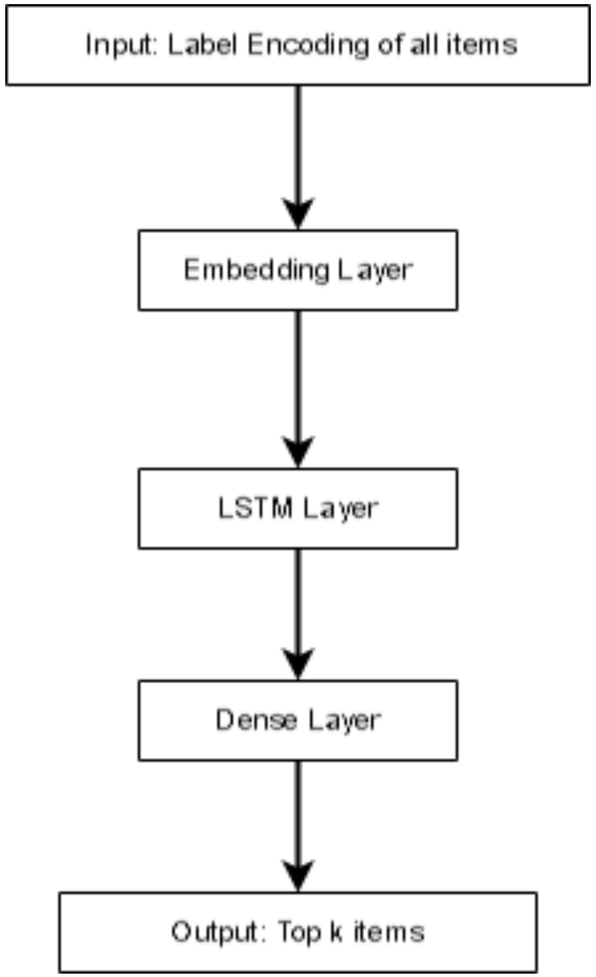
Ngoupeyou Tondj, 2018

PaleoRec: A transaction-based sequential recommender system



Manety et al., in review, Environmental Data Science

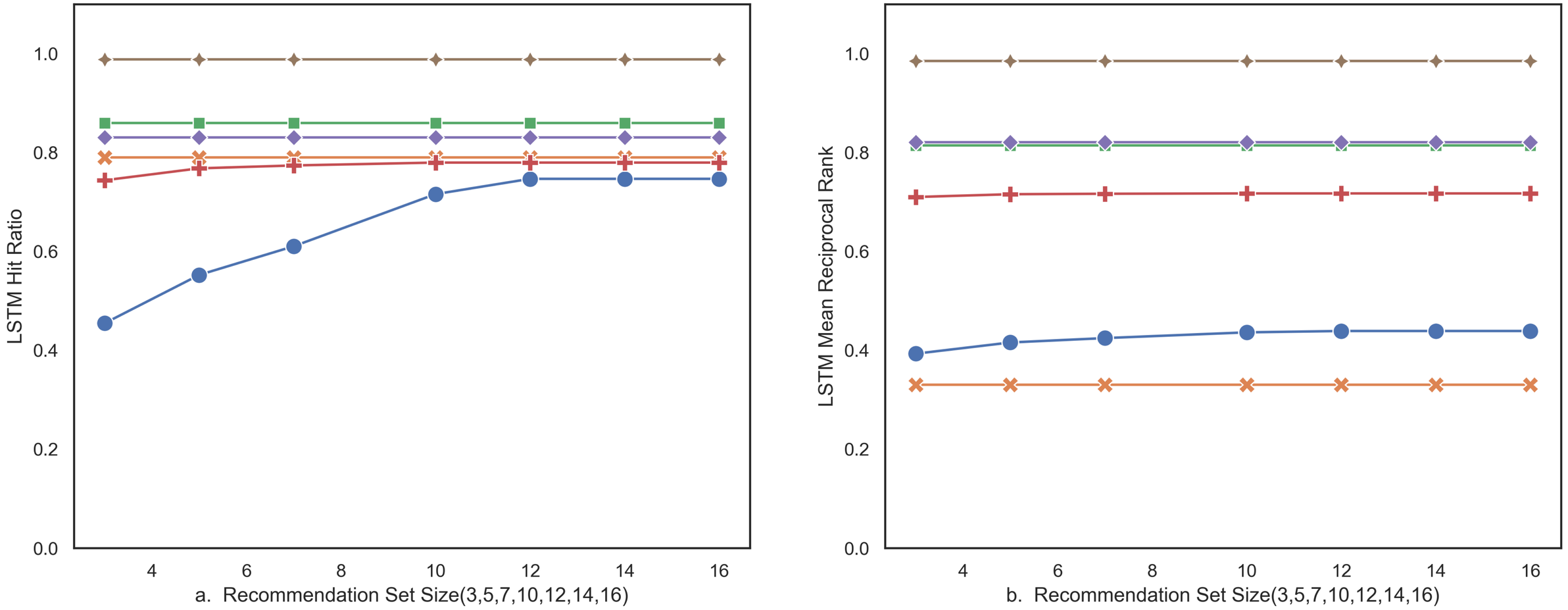
PaleoRec: A transaction-based sequential recommender system



	archiveType	proxyObservationType	units	interpretation/variable	interpretation/variableDetail	inferredVariable	inferredVarUnits
0	Wood	Rbar	NA	NA	NA	NA	NA
1	Wood	Trsgi	NA	Temperature	Air Surface	Air Surface Temperature	deg C
2	MarineSediment	TEX86	NotApplicable	Temperature	Sea Surface	Sea Surface Temperature	deg C
3	MarineSediment	D18O	permil	Temperature	Sea Surface	Sea Surface Temperature	deg C
4	LakeSediment	Sampleid	NotApplicable	NA	NA	NA	NA
5	Wood	Trsgi	NA	Temperature	Air Surface	Air Surface Temperature	deg C
6	MarineSediment	D18O	permil	D18Osw	Bottom Water	Temperature	deg C
7	Wood	Corrs	NA	NA	NA	NA	NA
8	Wood	Ringwidth	cm	NA	NA	NA	NA
9	MarineSediment	Depth	cm	NA	NA	NA	NA

Manety et al., in review, Environmental Data Science

PaleoRec: Metrics



Manety et al., in review, Environmental Data Science

Playground

MAP GRAPH

Are you new here? Take the tour to learn more about how to use this page.

ake the tour! Questions, comments, concerns

https://lipd.net

Playground

MAP GRAPH

Are you new here? Take the tour to learn more about how to use this page.

ake the tour! Questions, comments, concerns

https://lipd.net

I need to make decisions based on timely science results

I need to make decisions based on timely science results

Forecasting with paleoclimate data

- From the data
  - Sparse (extremely!)
  - At different resolution
  - Not the variables we are interested in

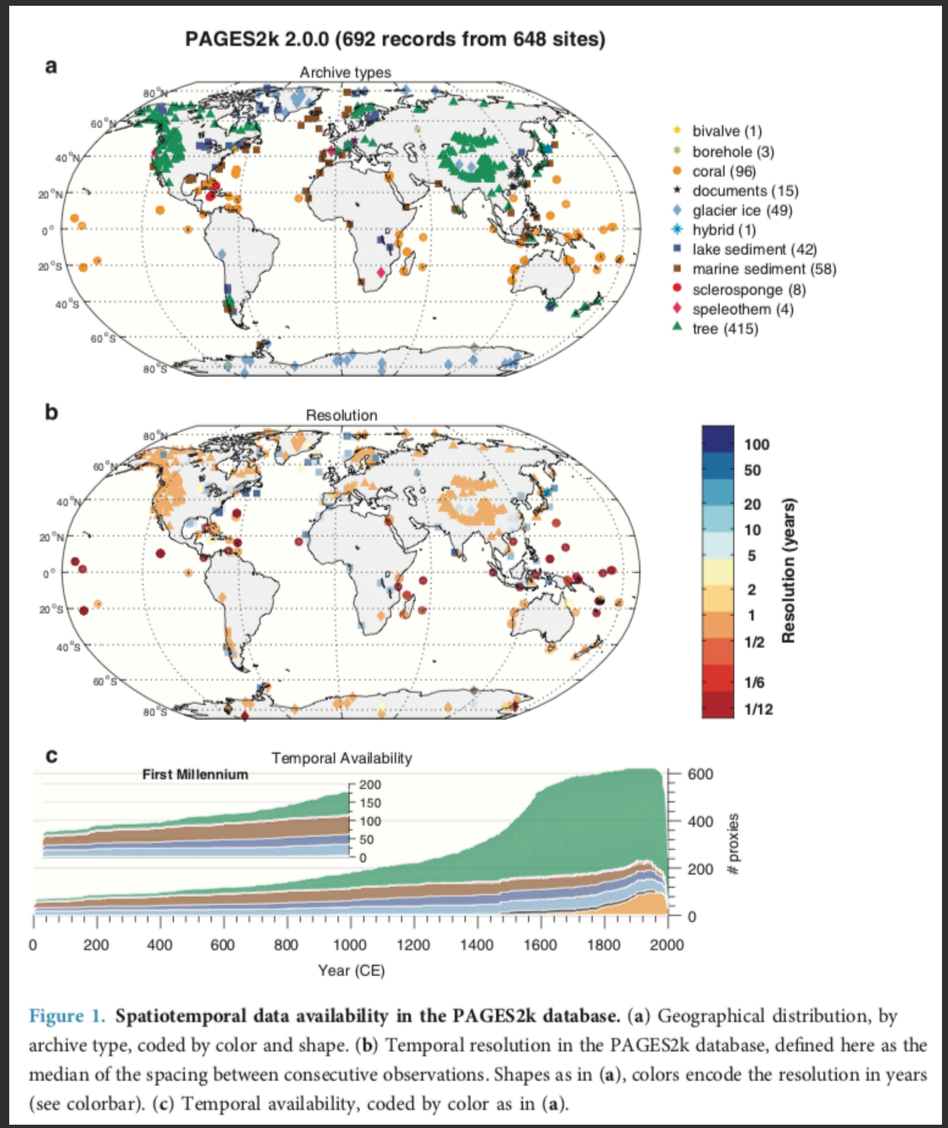
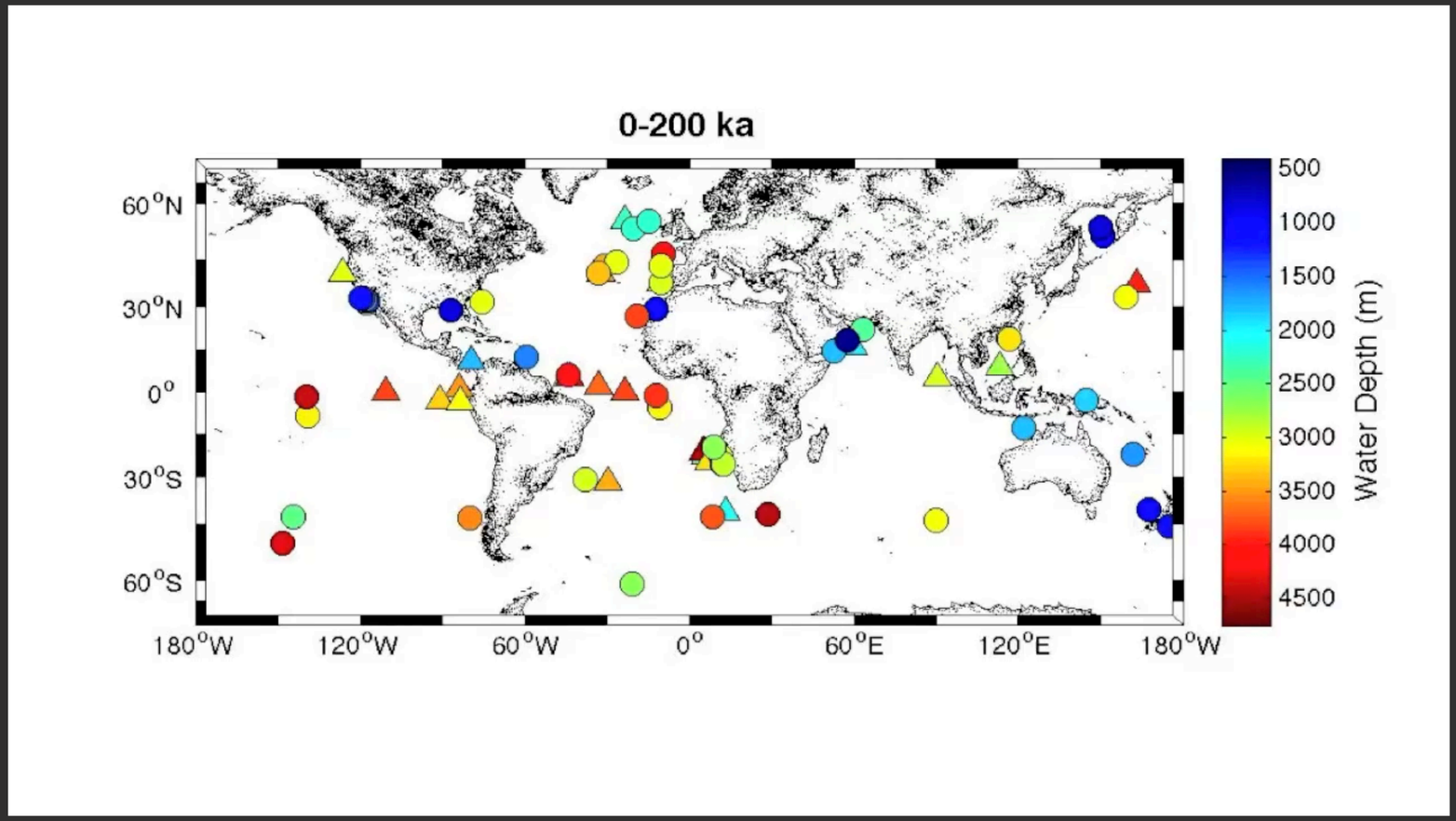


Figure 1. **Spatiotemporal data availability in the PAGES2k database.** (**a**) Geographical distribution, by archive type, coded by color and shape. (**b**) Temporal resolution in the PAGES2k database, defined here as the median of the spacing between consecutive observations. Shapes as in (**a**), colors encode the resolution in years (see colorbar). (**c**) Temporal availability, coded by color as in (**a**).

Forecasting with paleoclimate data

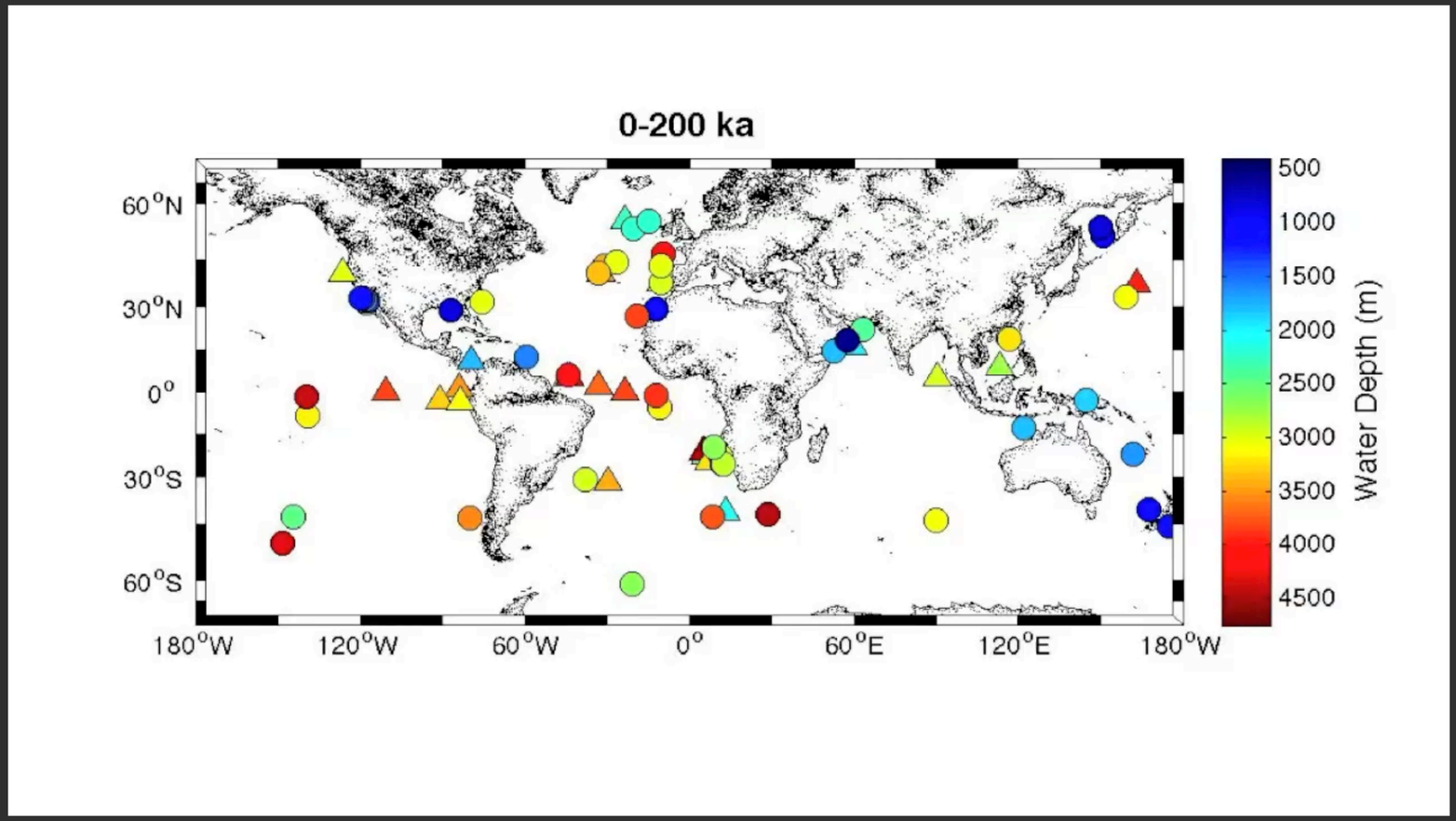
- From the data
  - Sparse (extremely!)
  - At different resolution
  - Not the variables we are interested in



Ahn et al., 2017

Forecasting with paleoclimate data

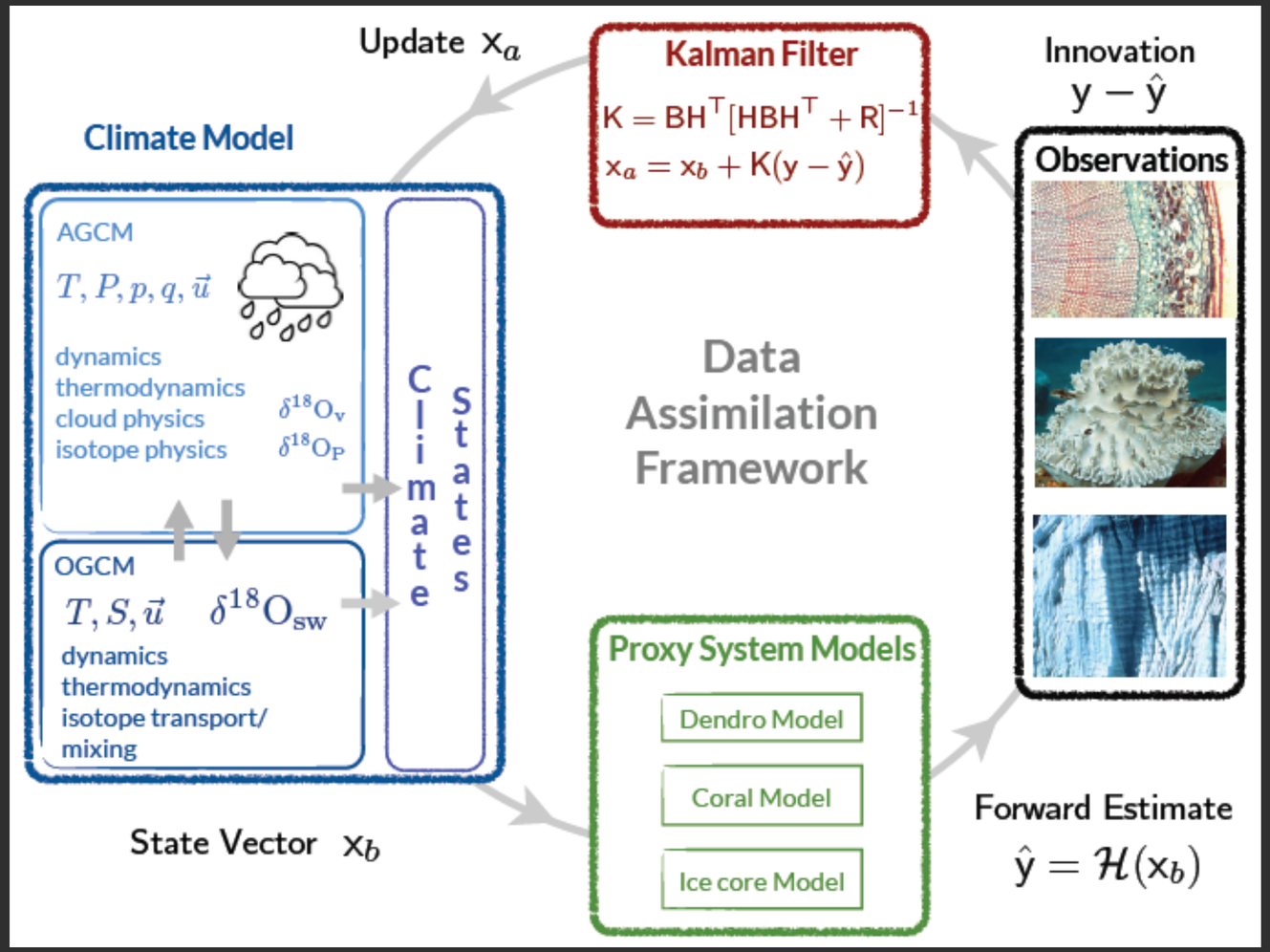
- From the data
  - Sparse (extremely!)
  - At different resolution
  - Not the variables we are interested in



Ahn et al., 2017

Forecasting with paleoclimate data

- From reanalysis:
  - Model Bias
  - PSM representation is simplistic



Tardif et al., 2019; image: NOAA -LMR project

I need to communicate the decision to stakeholders

I need to communicate the decision to stakeholders

Characterizing the counter-narratives of climate change

Abhilash Pandurangan
USC

Vanshika Sridharan
USC

Aditya Jajodia
USC

Fred Morstatter
USC

Sushmitha Ravikumar
USC

Deborah Khider
USC

USC CKIDS Center for Knowledge-Powered Data Science

Climate scientists post their findings and views regularly on social media

Climate scientists post their findings and views regularly on social media

These views are often met with harassment…

Climate scientists post their findings and views regularly on social media

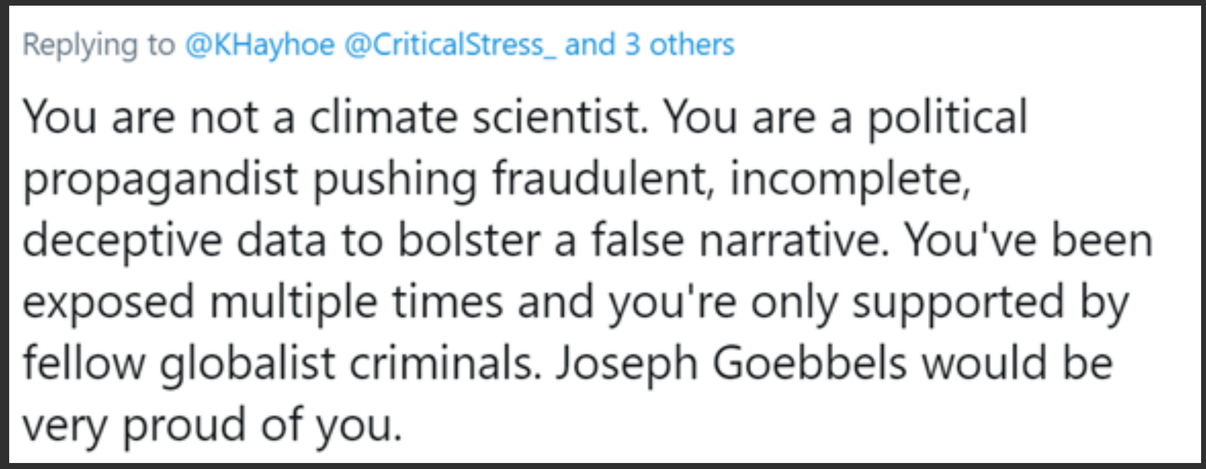
These views are often met with harassment…

… which is not picked up by already trained classifiers

Harassment of scientist on the Twitter platform

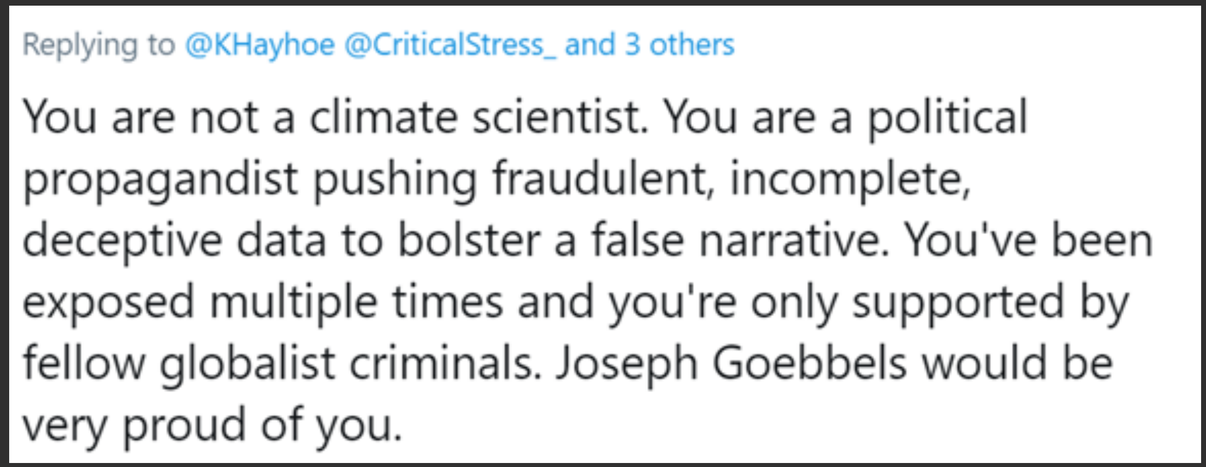
Harassment of scientist on the Twitter platform

- Harassment: attacking someone based on their race, gender, character…



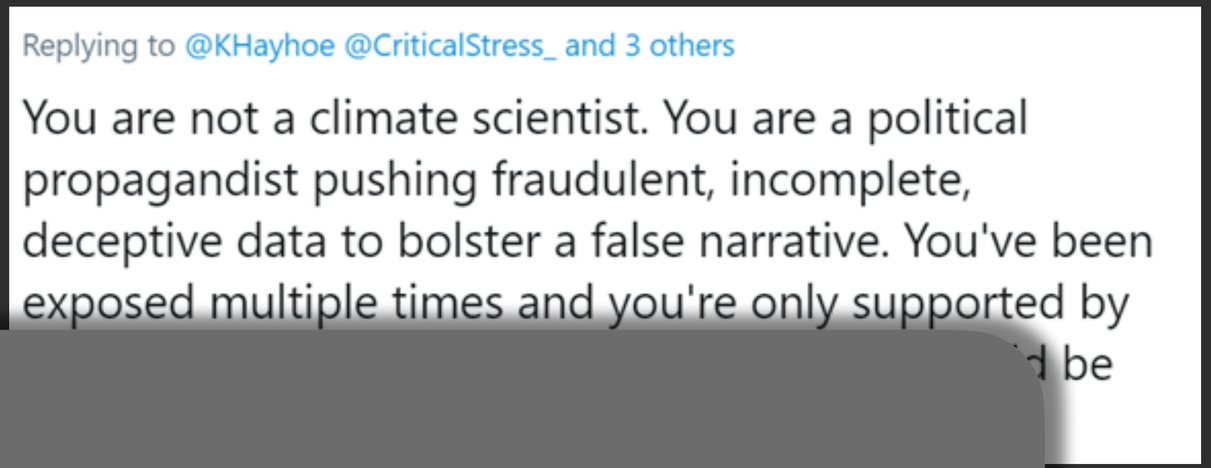
Harassment of scientist on the Twitter platform

- Harassment: attacking someone based on their race, gender, character…
- Laymansplaining: prevalent form of harassment where user condescendingly explains to a scientist their research



Harassment of scientist on the Twitter platform

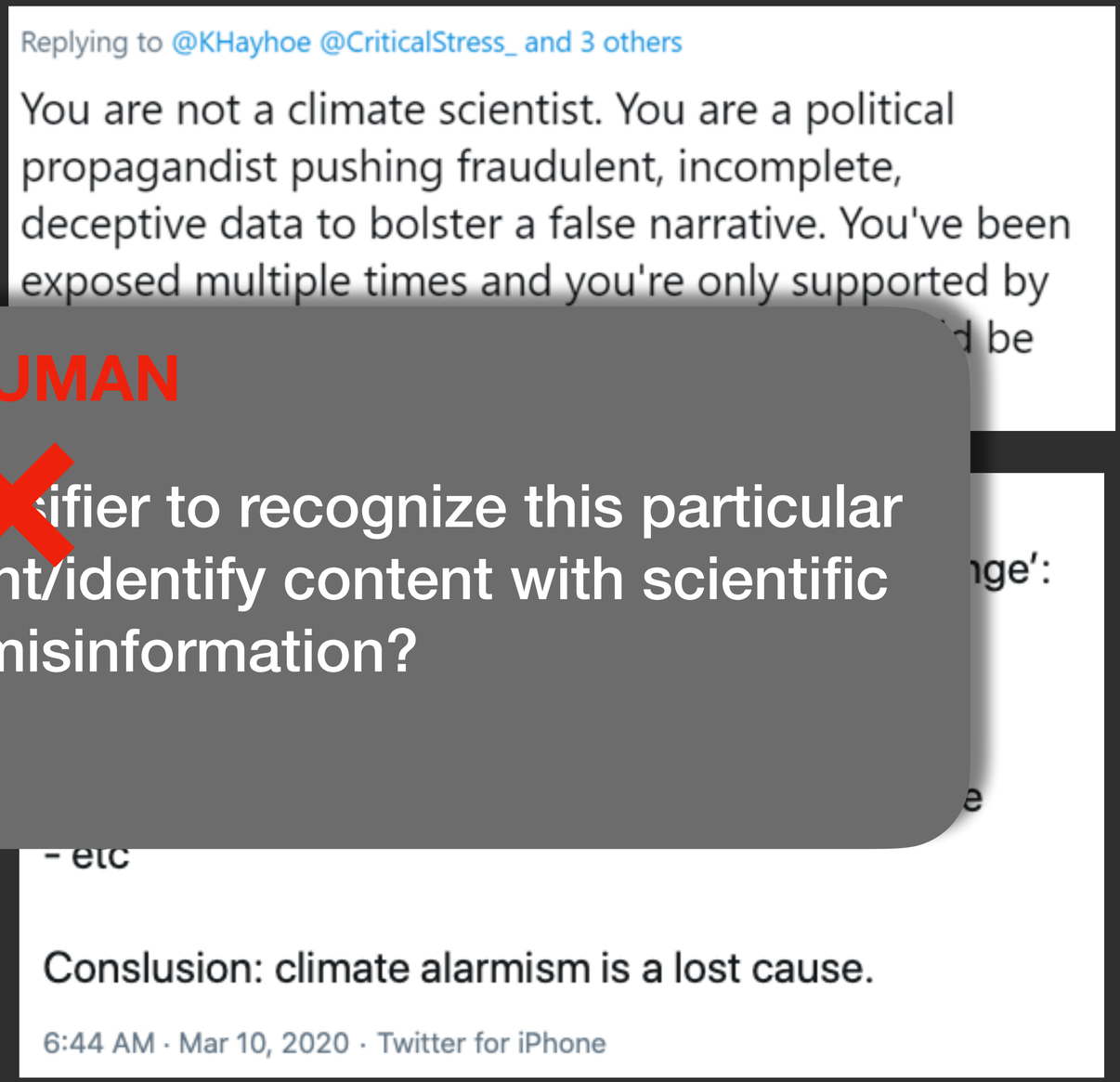
- Harassment: attacking someone ba… race, gend…
- Laymansp… prevalent f… harassmen… condescen… to a scienti…



Can we train a classifier to recognize this particular form of harassment/identify content with scientific misinformation?

Harassment of scientist on the Twitter platform

- Harassment: attacking someone ba… race, gend…
- Laymansp… prevalent f… harassmen… condescen… to a scienti…

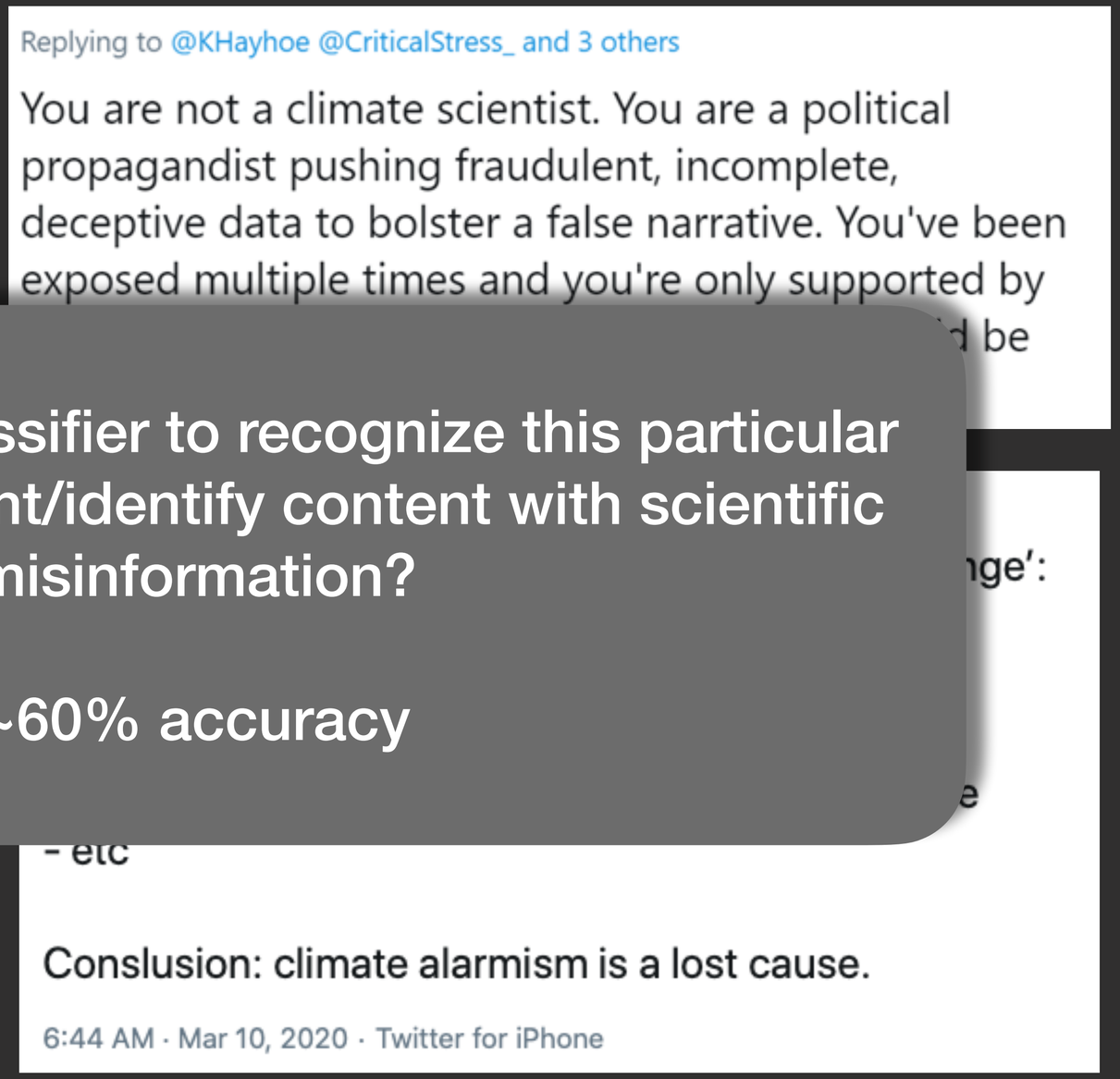


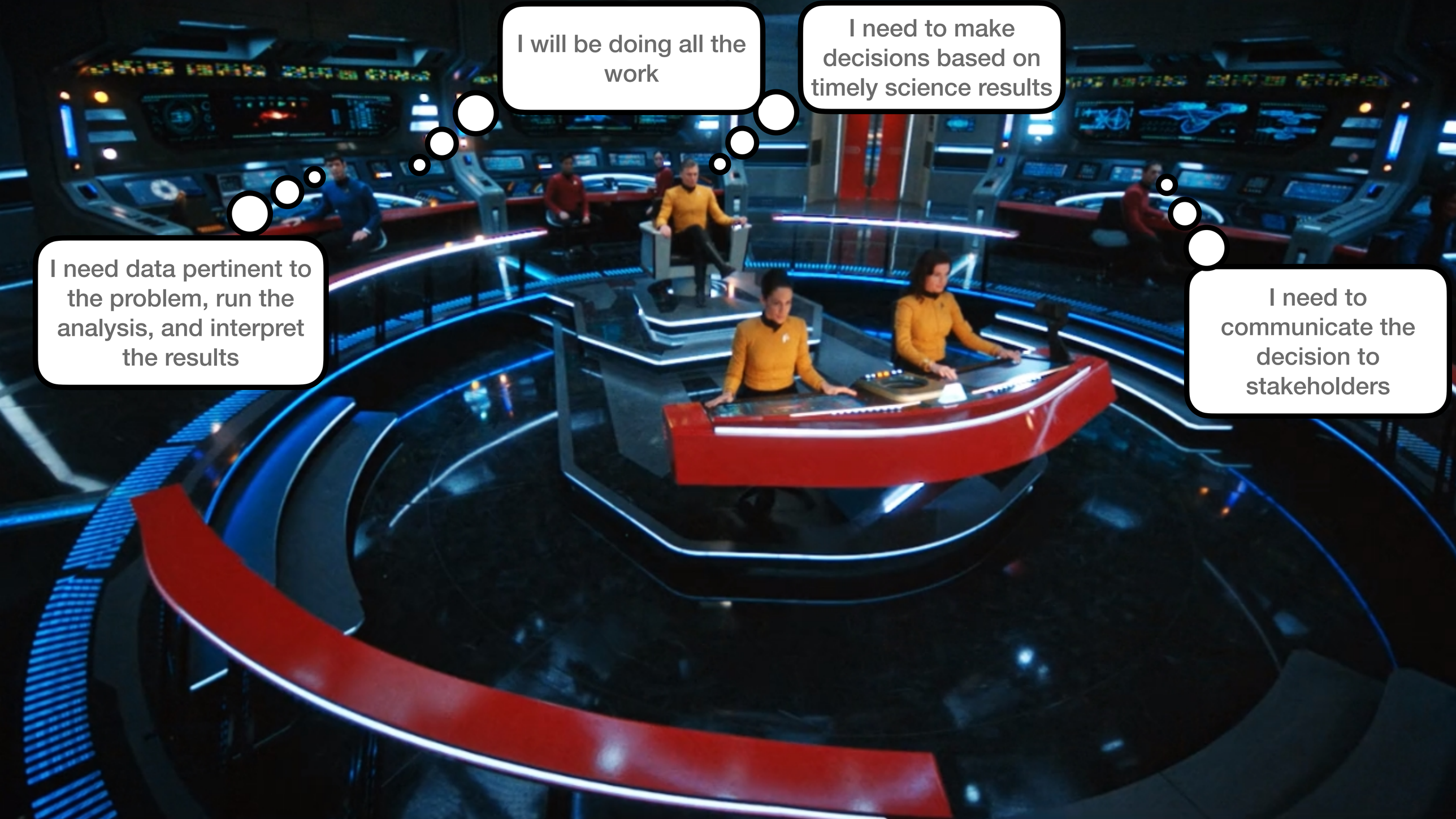
Harassment of scientist on the Twitter platform

- Harassment: attacking someone ba… race, gend…
- Laymansp… prevalent f… harassmen… condescen… to a scienti…

Can we train a classifier to recognize this particular form of harassment/identify content with scientific misinformation?

~60% accuracy




I will be doing all the work
I need to make decisions based on timely science results
I need data pertinent to the problem, run the analysis, and interpret the results
I need to communicate the decision to stakeholders

khider@usc.edu
@dkhider