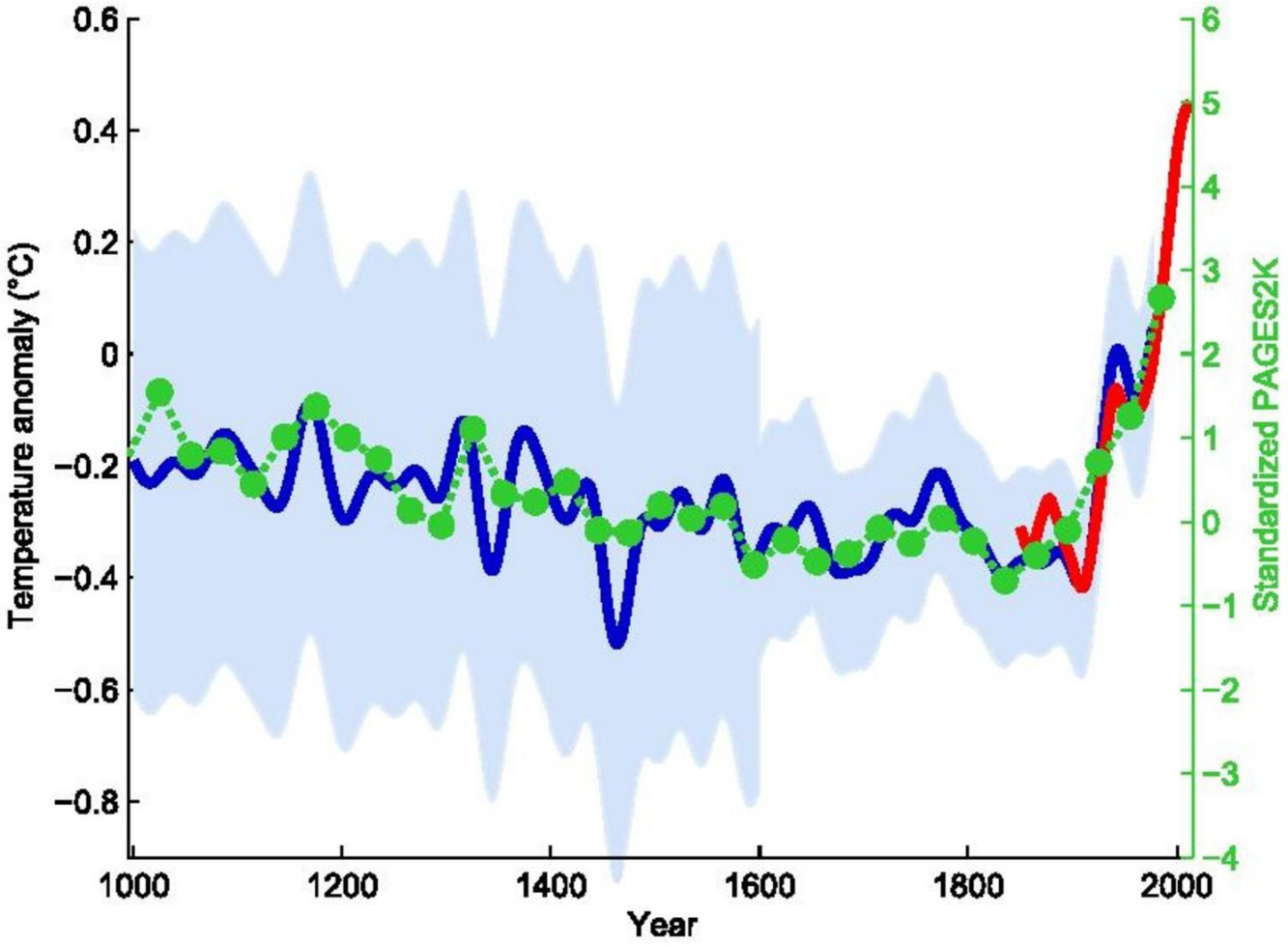
Annotating Paleoclimate Data

Yincheng Lin(MS DS), Shravya Manety (MS CS)

Dr. Deborah Khider (Information Sciences Institute) &
Dr. Julien Emile-Geay (Dept. of Earth Sciences)

Motivation

The hockey stick debate :



data: Mann et al, 1998. PAGES 2k Consortium (2013). HadCRUT4 temperature (Morice et al, 2012)

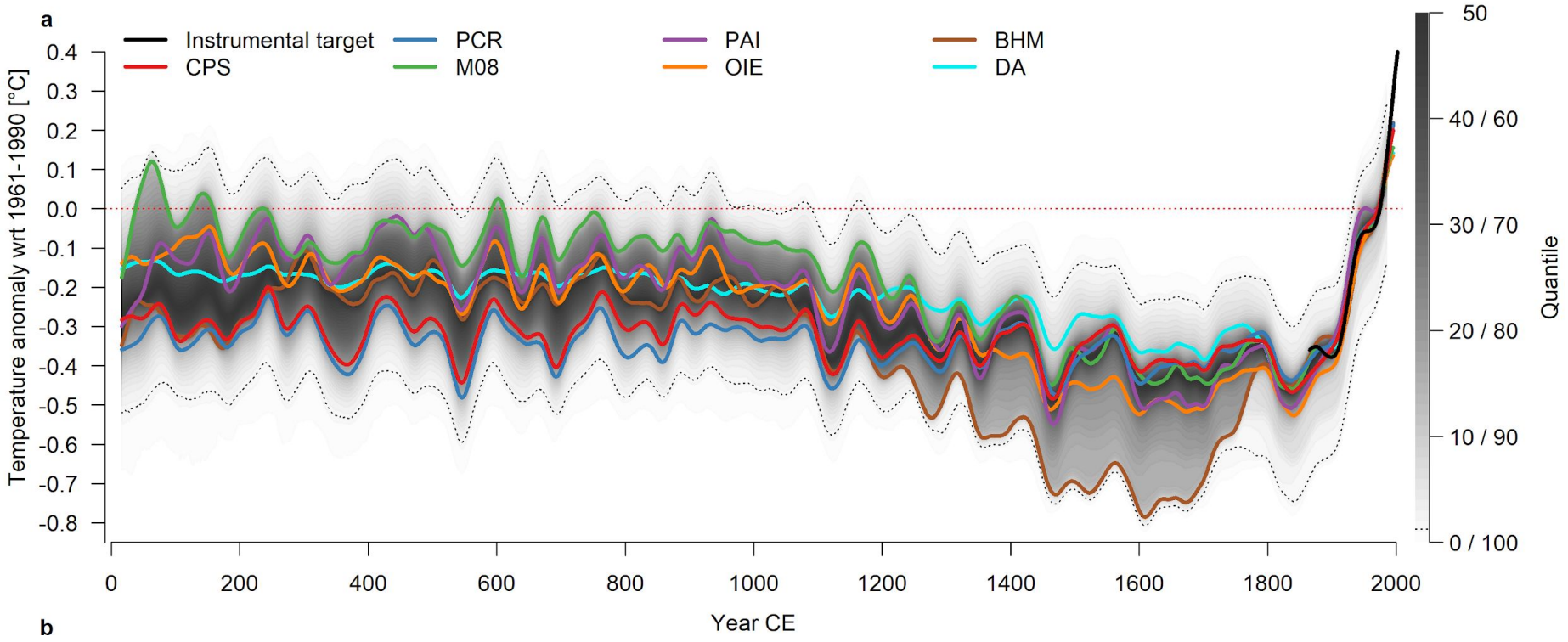
Motivation

From the time the hockey stick graph was published by Mann et al. [1998], climate researchers have probed its validity with new data and methods.

The latest effort, PAGES 2k, represents an ongoing community effort to update and refine the Mann et al study. In the process, scientists have annotated 692 climate datasets covering the past 2000 years. In turn, these datasets form the basis of statistical reconstructions (e.g. PAGES 2k Consortium, 2019; Barboza et al 2019) and climate assessments (e.g. IPCC).

A key bottleneck in the development and use of these datasets is annotation (detailed description of the metadata). Here we wish to develop a simple recommendation system that will help automate this process.

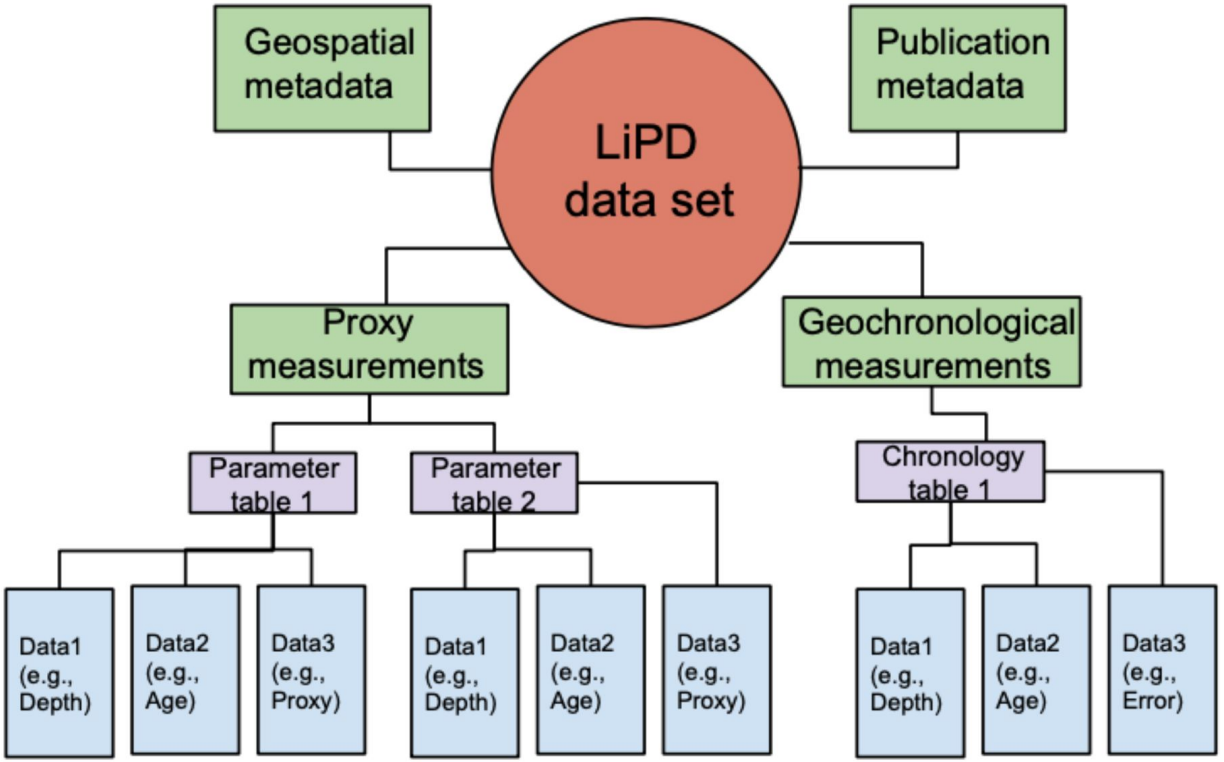
The Hockey Stick 2.0



Neukom, R., L. A. Barboza, M. P. Erb, F. Shi, J. Emile-Geay, M. N. Evans, J. Franke, D. S. Kaufman, L. Lücke, K. Rehfeld, A. Schurer, F. Zhu, S. Brönnimann, G. J. Hakim, B. J. Henley, F. C. Ljungqvist, N. McKay, V. Valler, and L. von Gunten (2019), Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nature Geoscience, 12(8), 643–649, doi:10.1038/s41561-019-0400-0.

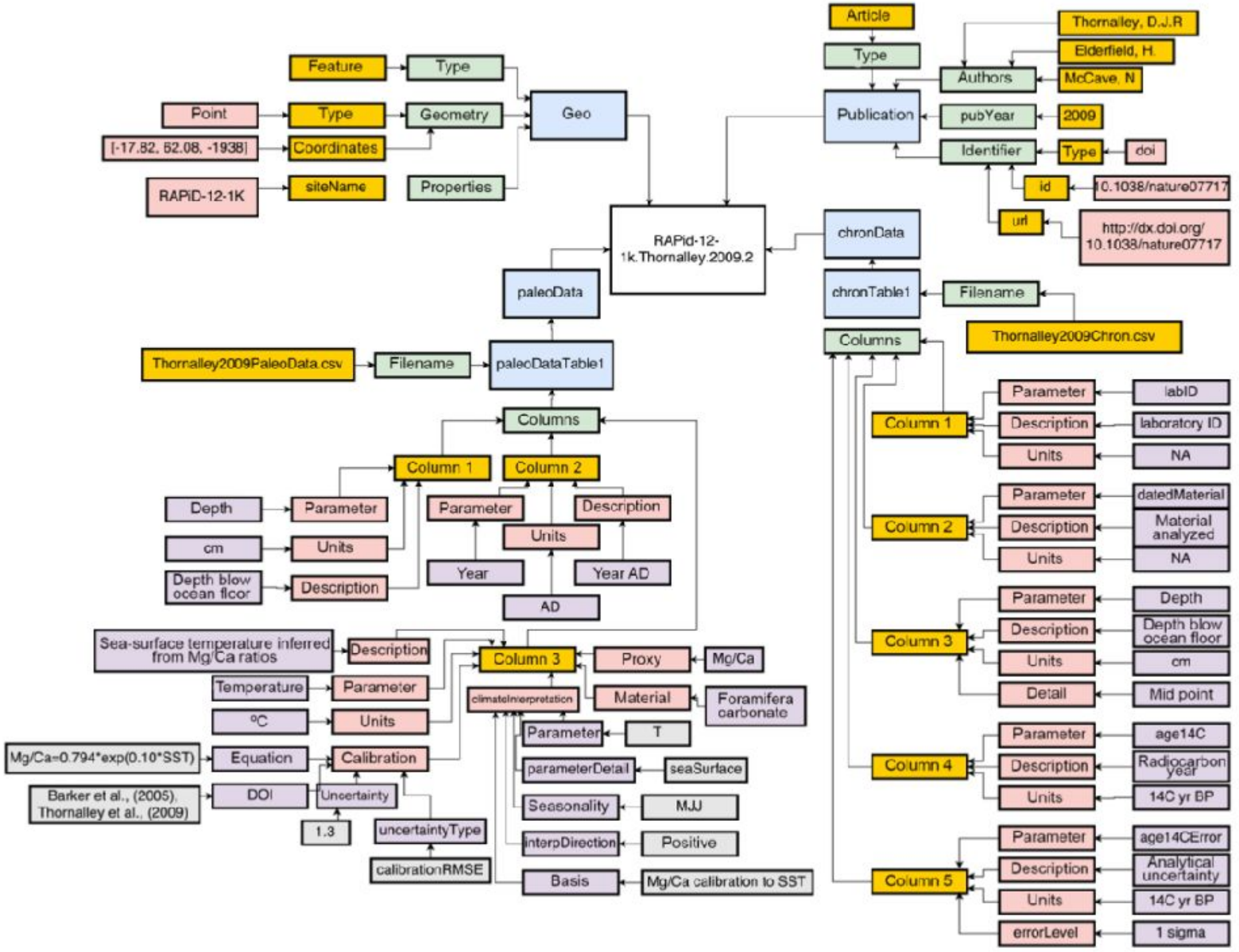
LiPD: Linked Paleo Data Format

Paleoclimate Data is organized in a series of csv files while the metadata is stored in JSON-LD format.



McKay and Emile-Geay, 2016

LiPD Format


Feature
Type
Point
Type
Geometry
[-17.82, 62.08, -1938]
Coordinates
RAPiD-12-1K
siteName
Properties
Geo
Article
Type
Thornalley, D.J.R
Elderfield, H.
McCave, N
Authors
Publication
pubYear
2009
Identifier
Type
doi
id
10.1038/nature07717
url
http://dx.doi.org/10.1038/nature07717
RAPid-12-1k.Thornalley.2009.2
chronData
paleoData
chronTable1
Filename
Thornalley2009Chron.csv
Columns
Thornalley2009PaleoData.csv
Filename
paleoDataTable1
Parameter
labID
Description
laboratory ID
Units
NA
Column 1
Column 2
Depth
Parameter
Parameter
Description
cm
Units
Units
Year
Year AD
Depth blow ocean floor
Description
AD
datedMaterial
Material analyzed
Sea-surface temperature inferred from Mg/Ca ratios
Description
Column 3
Proxy
Mg/Ca
Temperature
Parameter
climateInterpretation
Material
Foraminifera carbonate
°C
Units
Parameter
T
Depth
Depth blow ocean floor
cm
Mid point
Mg/Ca=0.794*exp(0.10*SST)
Equation
Calibration
parameterDetail
seaSurface
Barker et al., (2005), Thornalley et al., (2009)
DOI
Uncertainty
Seasonality
MJJ
1.3
uncertaintyType
interpDirection
Positive
calibrationRMSE
Basis
Mg/Ca calibration to SST
Column 4
Column 5
age14C
Radiocarbon year
14C yr BP
age14CError
Analytical uncertainty
errorLevel
1 sigma

Example:

```
"paleoData": [{
      "paleoDataTableName": "data",
      "filename": "
          Atlantic0220Thornalley2009.csv",
      "columns": [{
            "number": 1,
            "variableName": "depth",
            "variableType": "measured",
            "description": "depth below
                ocean floor",
            "units": "cm",
            "datatype": "csvw:NumericFormat
                ",
            "notes": "depth refers to top of
                 sample"
      },
      {
            "number": 2,
            "variableName": "year",
            "variableType": "inferred",
            "description": "calendar year AD
                ",
            "units": "AD",
            "datatype": "csvw:NumericFormat
                ",
            "method": "linear interpolation"
      },
      {
            "number": 3,
            "variableName": "temperature",
            "variableType": "inferred",
            "description": "sea-surface
                temperature inferred from Mg/
                Ca ratios",
            "datatype": "csvw:NumericFormat
                ",
            "material": "foraminifera
                carbonate",
            "calibration": {
                  "equation": "BAR2005: Mg/
                      Ca=0.794*exp(0.10*SST)
                      ",
                  "reference": "Barker et al
                      ., (2005), Thornalley
                      et al., (2009)",
                  "uncertainty": 1.3
            },
            "units": "deg C",
            "proxy": "Mg/Ca",
            "climateInterpretation": {
                  "variable": "T",
                  "variableDetail": "
                      seaSurface",
                  "seasonality": "MJJ",
                  "interpDirection": "
                      positive",
                  "basis": "Mg/Ca
                      calibration to SST"
            }
      }]
}]
```

Thornalley et al (2009)

Approach

Given the Archive Type, we need to recommend the inferredVariableType and proxyObservationType with the units. Given that, we proceed to recommend the Variable and Variable Details under the Interpretation Data.

Approach:

1. Develop a probability table based on the sample dataset assigning probabilities for each combination of the required parameters.
2. Using Conditional Probabilities to develop a basic interface for a Recommendation System.

Collecting Data

The data has been queried from LinkedEarth wiki using SPARQL. We had a total of 705 lipd files for various archive types. This is the sample data for the system.

To read each of the lipd files we utilised the lipdutilities python library to generate a csv with the required parameters .

	filename	archiveType	inferredVariableType	units	proxyObservationType	interpretation/variable	interpretation/variableDetail
0	A7.Oppo.2005	marine sediment	Age	yr BP	NaN	Age	calendar
1	A7.Oppo.2005	marine sediment	Age	yr BP	NaN	Age	calendar
2	A7.Oppo.2005	marine sediment	NaN	per mil	D18O	d18Osw	sea surface
3	A7.Oppo.2005	marine sediment	NaN	mmol/mol	Mg/Ca	Temperature	sea surface
4	A7.Oppo.2005	marine sediment	Sea Surface Temperature	deg C	NaN	Temperature	sea surface

Probabilistic Measures to Recommend Data

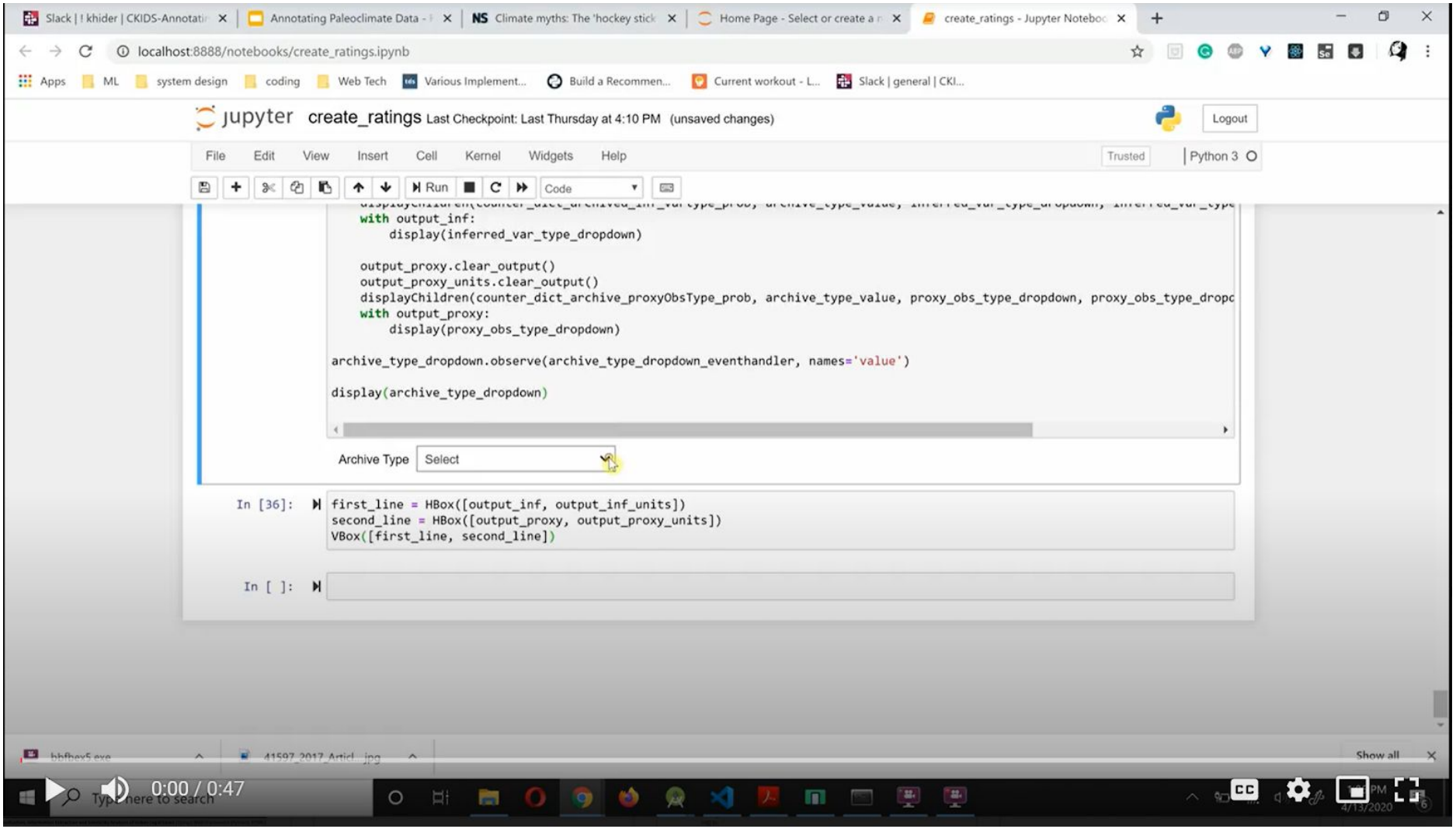
We used Bayes' rule to compute the probability for the Inferred Variable Type given the Archive Type.

P(SeaSurfaceTemperature | MarineSediment)


```
{'marine sediment': {'Age': 0.20357142857142857,
  'summation': 280,
  'Sea Surface Temperature': 0.24642857142857144,
  'Temperature': 0.18928571428571428,
  'Year': 0.19285714285714287,
  'D18O': 0.1,
  'Relative Sea Level': 0.007142857142857143,
  'Sedimentation Rate': 0.010714285714285714,
  'Thermocline Temperature': 0.010714285714285714,
  'Sea Surface Salinity': 0.007142857142857143,
  'Carbonate Ion Concentration': 0.014285714285714285,
  'Radiocarbon Age': 0.0035714285714285713,
  'Subsurface Temperature': 0.0035714285714285713,
  'Salinity': 0.0035714285714285713,
  'Accumulation rate': 0.0035714285714285713,
  'D18o': 0.0035714285714285713},
```

```
{'Age': {'yr BP': 0.1794871794871795,
  'summation': 78,
  'BP': 0.20512820512820512,
  'kyr BP': 0.5,
  'ka BP': 0.01282051282051282,
  'cal. BP': 0.01282051282051282,
  'year': 0.038461538461538464,
  'kaBP': 0.01282051282051282,
  'yr. BP': 0.01282051282051282,
  'ky': 0.01282051282051282,
  'yr B.P.': 0.01282051282051282},
```

Working Prototype



Future Work

1. Data Cleaning

   a. Temperature (deg C, degC)

   b. Year (yr BP, yr B.P.)

2. Separate Geochronological and Proxy Metadata.
3. Currently we are considering that scientists will progress in a sequential manner by filling the Archive Type, then the Inferred Variable Type and Proxy Observation Type. Further we also need to calculate the the back probabilities for the Inferred Variable Type and Proxy Observation Type to fill in the missing values from the LiPD file.