# Facilitating scientific investigations from long-tail data with Python


**Deborah Khider**
khider@usc.edu


## Background

The standard NumPy datetime64 objects restricted analyses with Pandas to relatively short timespans (~584 years), limiting its usefulness in fields such as paleoclimatology, the study of how climate has changed in the past.

Introducing a non-nanosecond dtype for datetime64 objects and thus allowing resolutions as coarse as 1s (and therefore a timespan of a few billion years) has broadened the range of analyses that can be performed on these datasets. Here we show how Pandas has become a linchpin for paleoclimate studies, from data wrangling to analysis.

## The Linked Paleo Data Format and PyLiPD

LiPD is a flexible data container using CSV files to store the timeseries data and a JSON-LD file for metadata. The format has given rise to an ontology that explicitly defines the relationships in paleoclimate datasets.

PyLiPD is a Python toolbox that can read LiPD files and convert them internally to RDF graphs for complex querying. The toolbox returns Pandas DataFrames for further data analyses.

```
from pylipd.lipd import LiPD

path = '../data/Pages2k/'

D = LiPD()
D.load_from_dir(path)
```

```
D_geo = D.filter_by_geo_bbox(lonMin=0, latMin=40, lonMax=90, latMax=70)
```

```
ts = D_filt.to_lipd_series()
```

```
df = ts.get_timeseries_essentials()

df.head()
```

| | dataSetName | archiveType | name | TSID | values | units | proxy |
|---|---|---|---|---|---|---|---|
| 0 | Eur-SpanishPyrenees.Dorado-Linan.2012 | tree | trsgi | Eur_020 | [-1.612, -0.703, -0.36, -0.767, -0.601, -0.733... | None | TRW |
| 1 | Eur-SpanishPyrenees.Dorado-Linan.2012 | tree | year | PYT2K8MIA3N | [1260.0, 1261.0, 1262.0, 1263.0, 1264.0, 1265.... | AD | None |
| 2 | Eur-FinnishLakelands.Helama.2014 | tree | temperature | Eur_005 | [14.603, 14.643, 12.074, 13.898, 13.671, 13.41... | degC | MXD |
| 3 | Eur-FinnishLakelands.Helama.2014 | tree | year | PYTUSB62S0A | [2000.0, 1999.0, 1998.0, 1997.0, 1996.0, 1995.... | AD | None |
| 4 | Eur-NorthernScandinavia.Esper.2012 | tree | year | PYTECO66XAD | [-138.0, -137.0, -136.0, -135.0, -134.0, -133.... | AD | None |

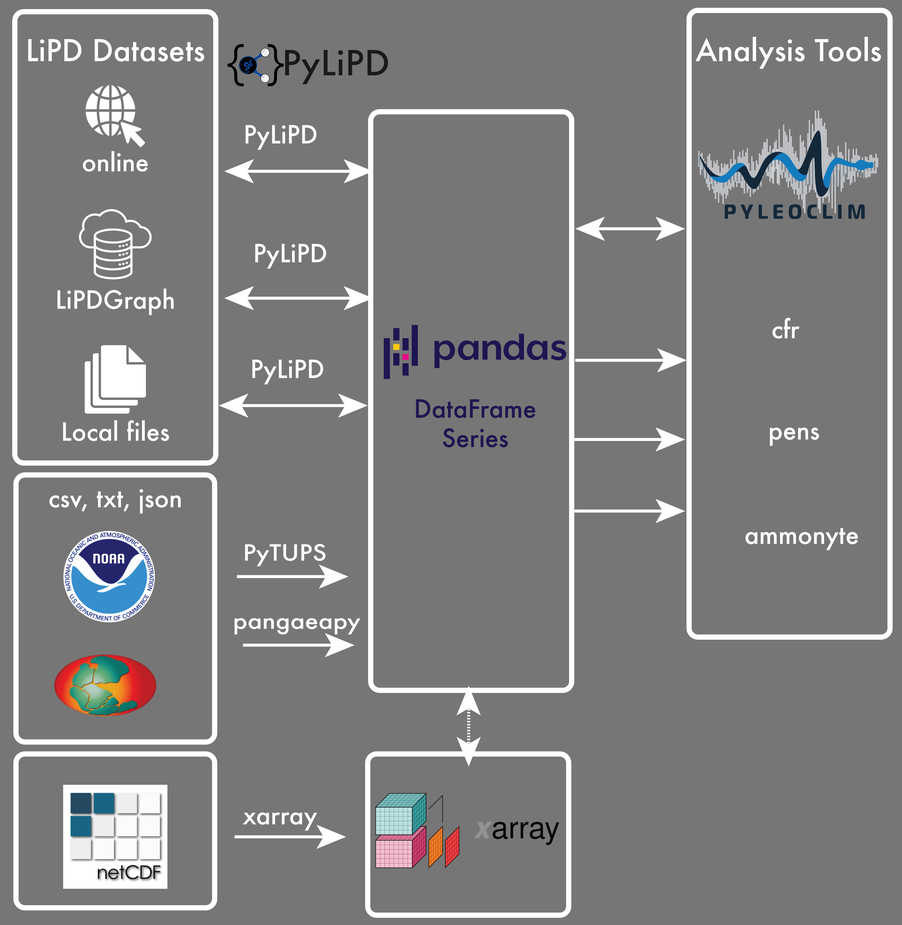
```
df_loc = D_filt.get_all_locations()
```

| | dataSetName | geo_meanLat | geo_meanLon | geo_meanElev |
|---|---|---|---|---|
| 0 | Eur-SpanishPyrenees.Dorado-Linan.2012 | 42.5 | 1.000 | 1200.0 |
| 1 | Eur-NorthernScandinavia.Esper.2012 | 68.0 | 25.000 | 300.0 |
| 2 | Eur-FinnishLakelands.Helama.2014 | 62.0 | 28.325 | 130.0 |

```
df_merged = df.merge(df_loc,how='inner', on='dataSetName')
```

https://pylipd.readthedocs.io/en/latest/
http://linked.earth/pylipdTutorials/intro.html

# The new non-nanosecond dtype in Pandas and its integration with paleoclimate-focused Python libraries have greatly eased data analysis.




Take a picture to download the full poster

## Pandas `datetime64[s]`

```
import pandas as pd
import numpy as np
```

```
idx = pd.Index(np.arange(-1000000000000,2,100000000), dtype='datetime64[s]')
signal = np.cos(2*np.pi*1/40*np.arange(len(idx)))
ser = pd.Series(data=signal, index=idx, name = 'Periodic Signal')

ser.head()
```

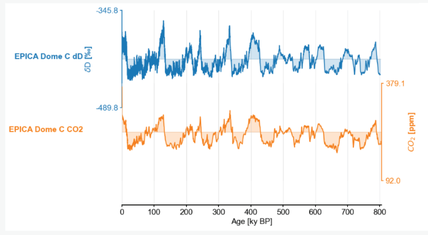
```
-29719-04-05 22:13:20    1.000000
-29716-06-06 08:00:00    0.987688
-29713-08-07 17:46:40    0.951057
-29710-10-08 03:33:20    0.891007
-29707-12-08 13:20:00    0.809017
Name: Periodic Signal, dtype: float64
```

## Pyleoclim

Pyleoclim is a Python package for the analysis of paleoclimate time series data.

Pyleoclim uses the new Pandas non-nanosecond datetime to resample timeseries and provide an interoperable format to other scientific libraries such as `Seaborn`.

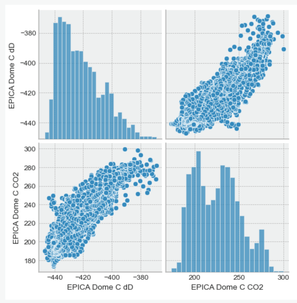
```
edc = pyleo.utils.load_dataset('EDC-dD')

ms =  edc.convert_time_unit("ky BP") & co2ts

ms.stackplot()
```




```
df = ms.to_pandas(use_common_time=True)
df.head()
```

| datetime | EPICA Dome C dD | EPICA Dome C CO2 |
|---|---|---|
| 1911-08-18 06:32:30 | -390.900000 | 299.817617 |
| 1630-04-19 08:12:23 | -400.382865 | 274.220786 |
| 1348-12-20 09:52:15 | -401.167982 | 278.367365 |
| 1067-08-23 11:32:07 | -398.662864 | 282.612398 |
| 786-04-25 13:11:59 | -396.520050 | 279.625252 |

```
sns.set(font_scale=0.8)
with plt.style.context('bmh'):
    sns.pairplot(df)
```




https://pyleoclim-util.readthedocs.io/en/latest/
http://linked.earth/PyleoTutorials/intro.html


Deborah Khider[1], Varun Ratnakar[1], Julien Emile-Geay[2], Kim Pevey[3], Marco Gorelli[3], Nicholas McKay[4], Alexander James[2], Jordan Landers[2]
[1]USC Information Science Institute, [2]USC Department of Earth Sciences, [3]Quansight, [4]Northern Arizona University