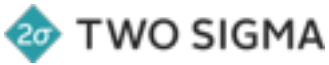
WHATS NEW IN PANDAS

Jeff Reback

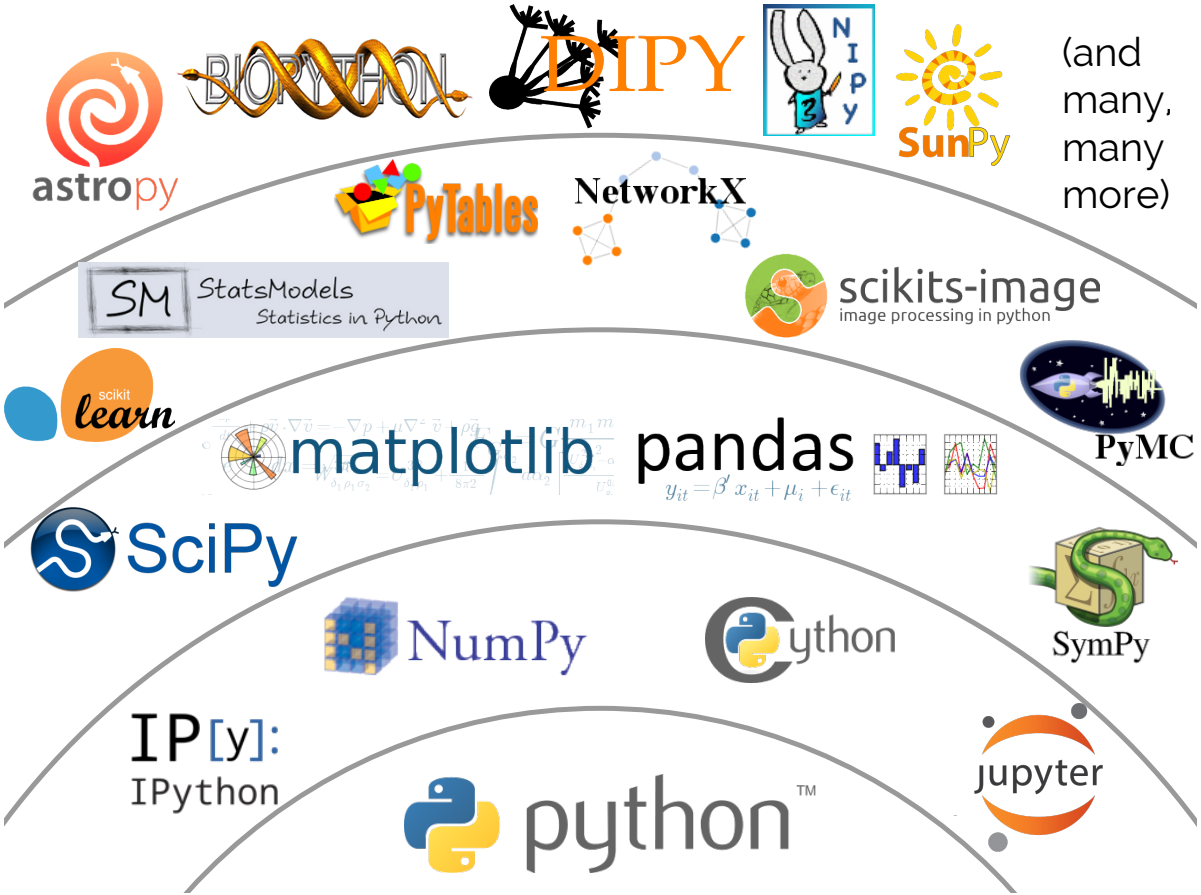
About me

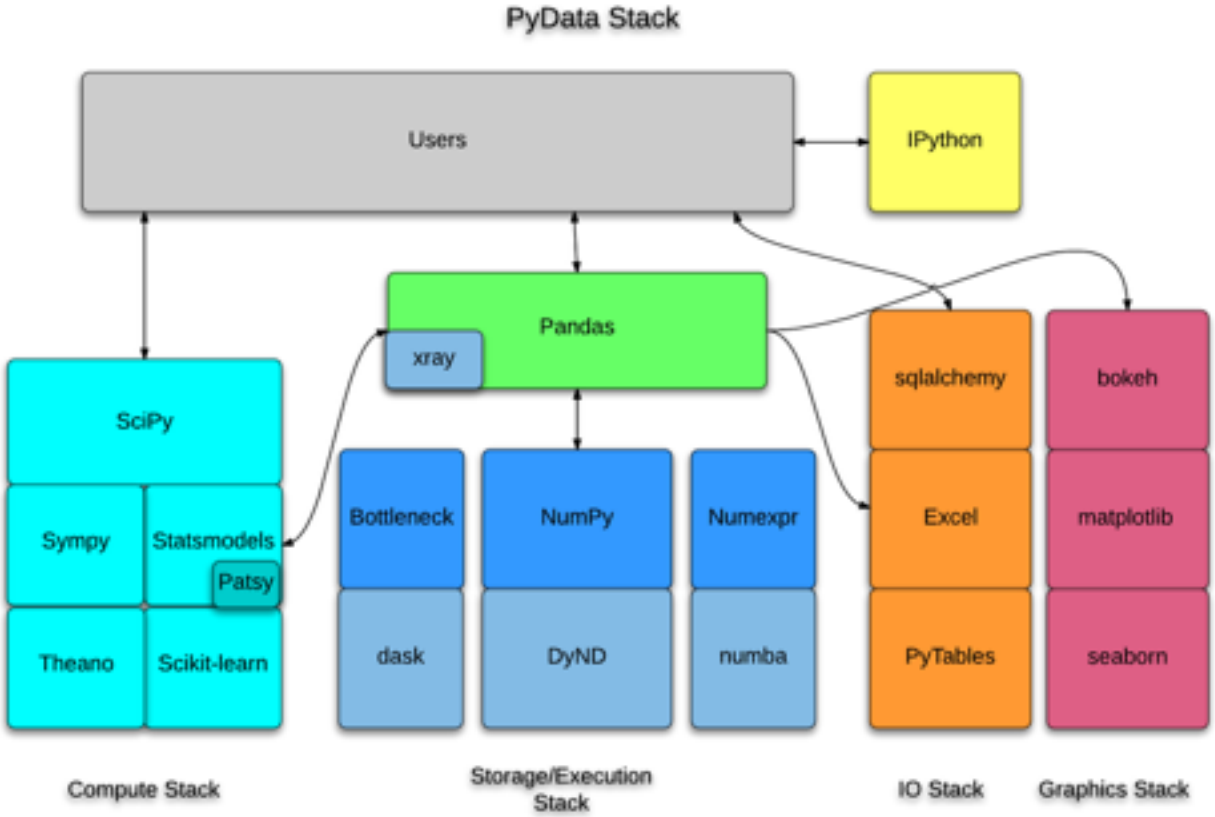
Jeff is a senior software developer for financial companies. As a former quant he has much experience in building financial trading systems, using python and working with very large data. He has been a core committer to the pandas project for the past few years, and currently manages the project.

Jeff holds a B.S. in Computer Science from the Massachusetts Institute of Technology.



PYDATA STACK

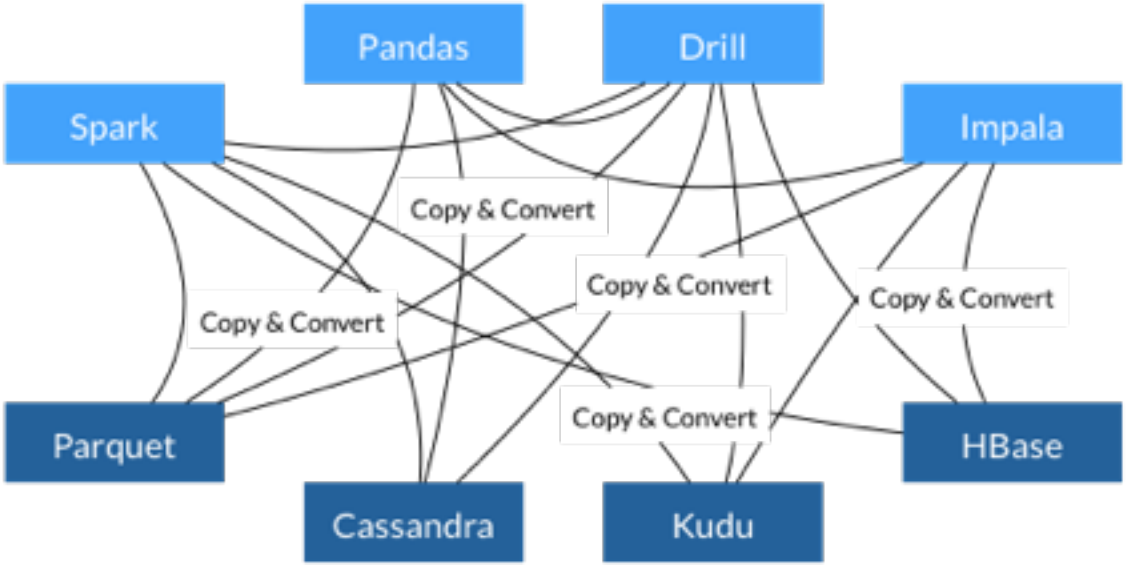

astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM
StatsModels
Statistics in Python
scikits-image
image processing in python
scikit
learn
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
IP[y]:
IPython
python
jupyter


PyData Stack
Users
IPython
Pandas
xray
SciPy
Sympy
Statsmodels
Patsy
Theano
Scikit-learn
Bottleneck
dask
NumPy
DyND
Numexpr
numba
sqlalchemy
Excel
PyTables
bokeh
matplotlib
seaborn
Compute Stack
Storage/Execution Stack
IO Stack
Graphics Stack

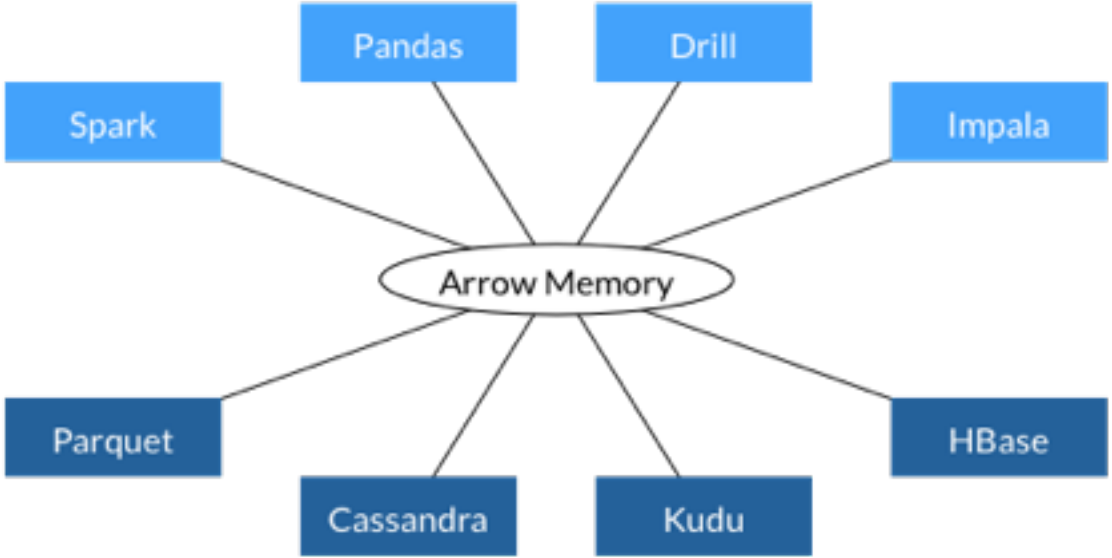
Why PyData?

- Amazing variety out there
  - scikit-learn, statsmodels, keras
  - matplotlib, seaborn, bokeh
- Tools to Scale up
  - numpy, numexpr, cython, numba
- and Scale Out
  - IPython, dask, PySpark

PANDAS 2



https://arrow.apache.org/



https://arrow.apache.org/

Its Happening

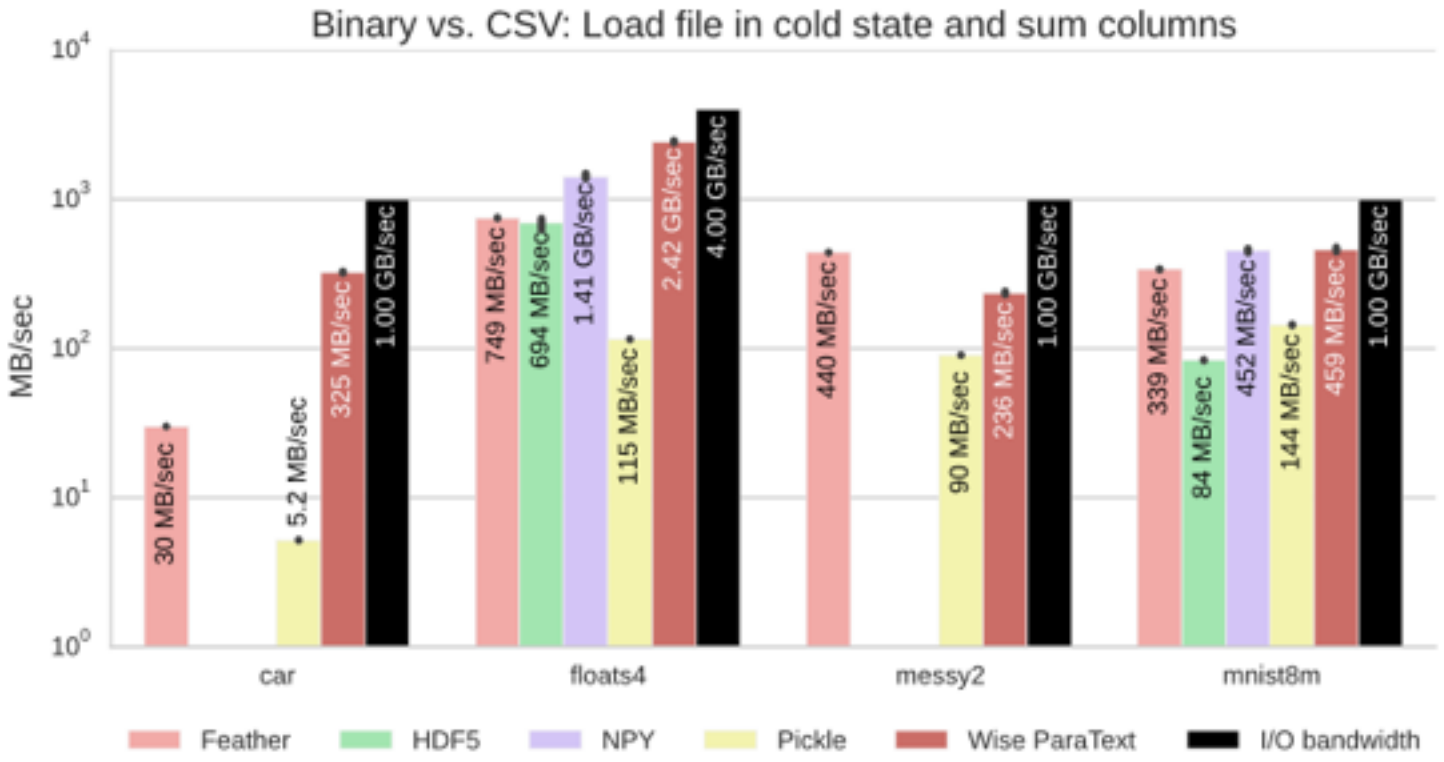
WHATS COMING FOR PANDAS

What has been recently added

- .ix Deprecation
- Panel Deprecation
- Asof Time Series Merging
- Time-aware Rolling
- `DataFrame.agg()`
- Interval Dtype & Index
- `to/from_feather()`
- `to/from_parquet()`

Whats coming in the longer term

- `libpandas`
  - setup clear back-end API / c-API
  - pandas unified data types
    - gain real types with missing values
    - boolean, integer
    - categorical Strings
  - column oriented DataFrame
  - lazy evaluation
  - PyData & Apache friendly
  - High performance IO / CSV & JSON

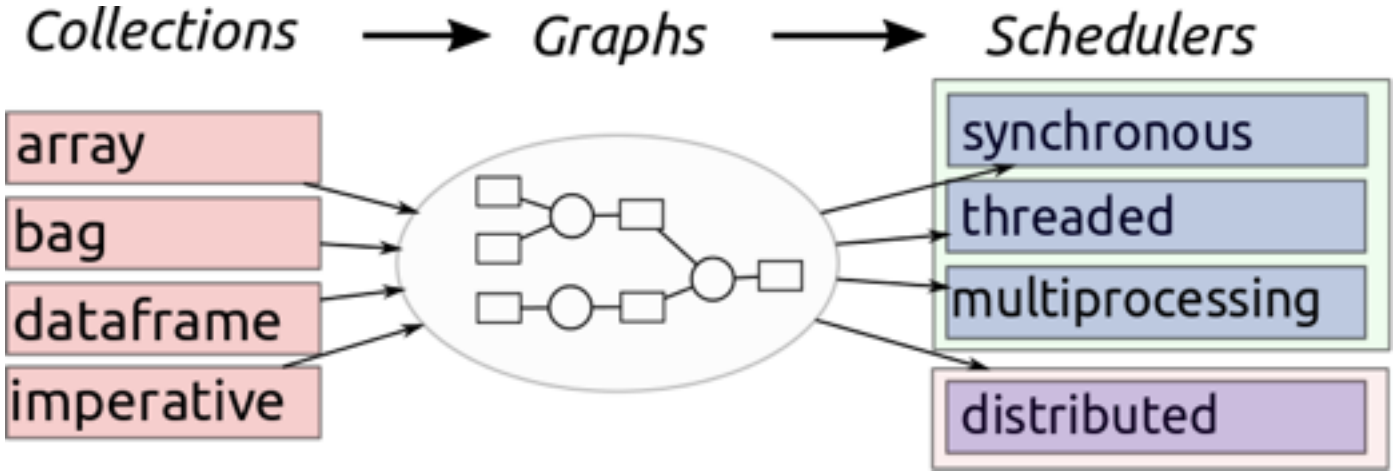


- http://www.wise.io/tech/paratext

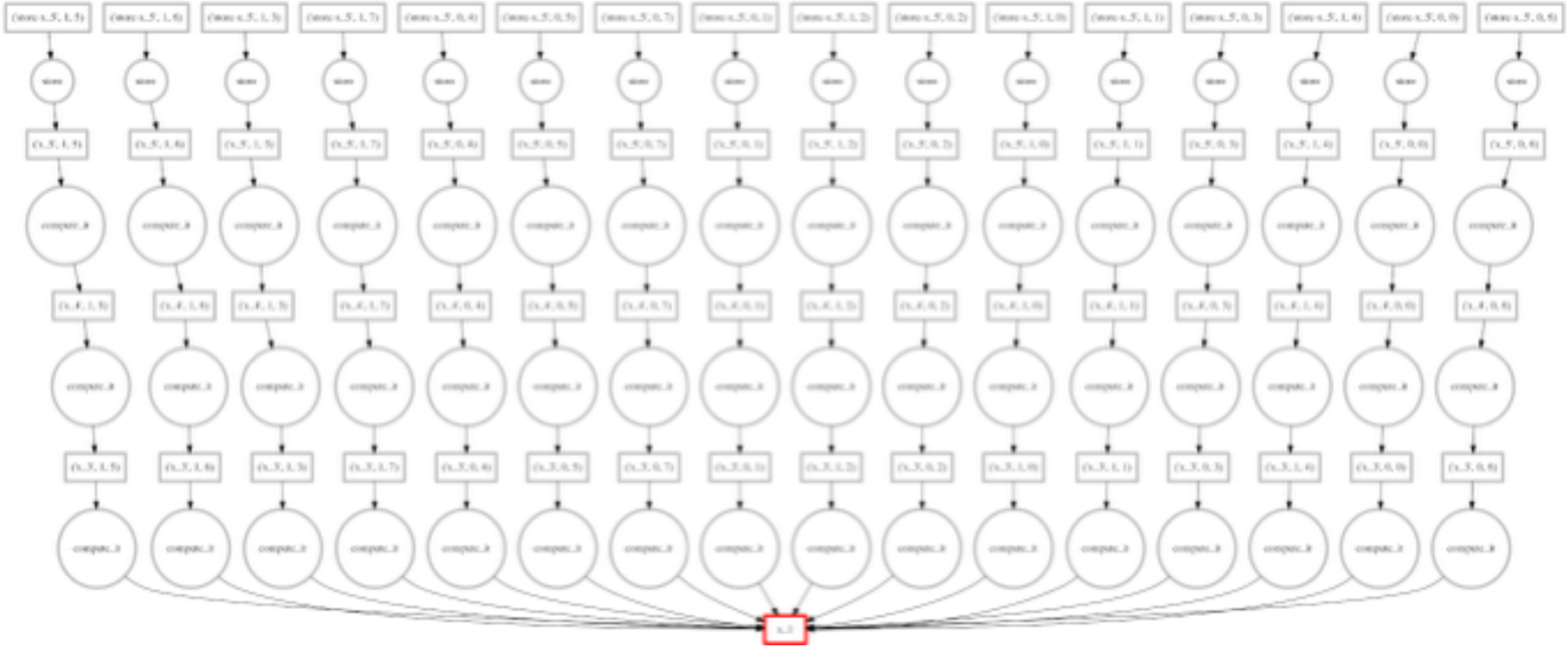
Whats coming in the longer term

- `.plot(engine=bokeh | seaborn | mpl)`
- `.apply(engine=dask | numba)`
- `.groupby(engine=dask)`
- `.to_dask()`

DASK

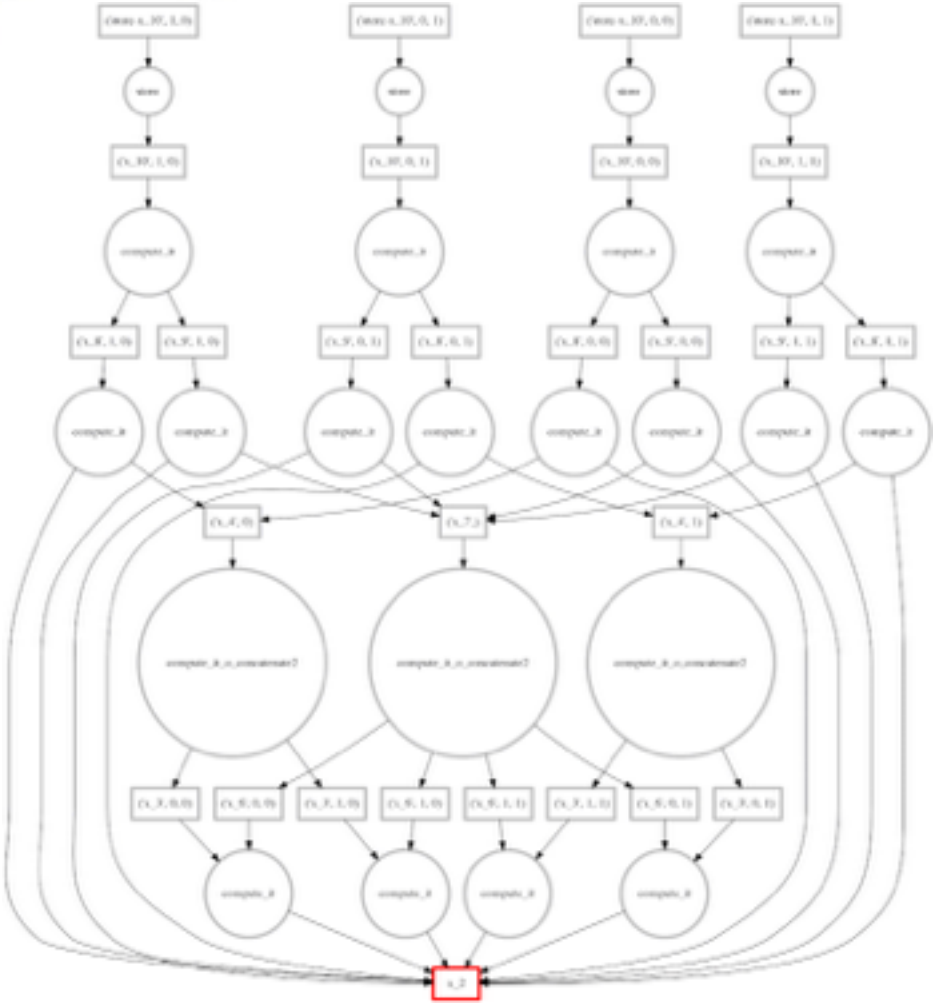


- Collections build task graphs
- Schedulers execute task graphs
- Graph specification = uniting interface
- A generalization of RDDs



```
(((A + 1) * 2) ** 3)
```

```
(B - B.mean(axis=0))
+ (B.T / B.std())
```



dask dataframe

pandas

```
>>> import pandas as pd

>>> df = pd.read_csv('iris.csv')

>>> df.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
>>> max_sepal_length_setosa = df[df.species ==
'setosa'].sepal_length.max()

5.7999999999999998
```

dask

```
>>> import dask.dataframe as dd

>>> ddf = dd.read_csv('*.csv')

>>> ddf.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

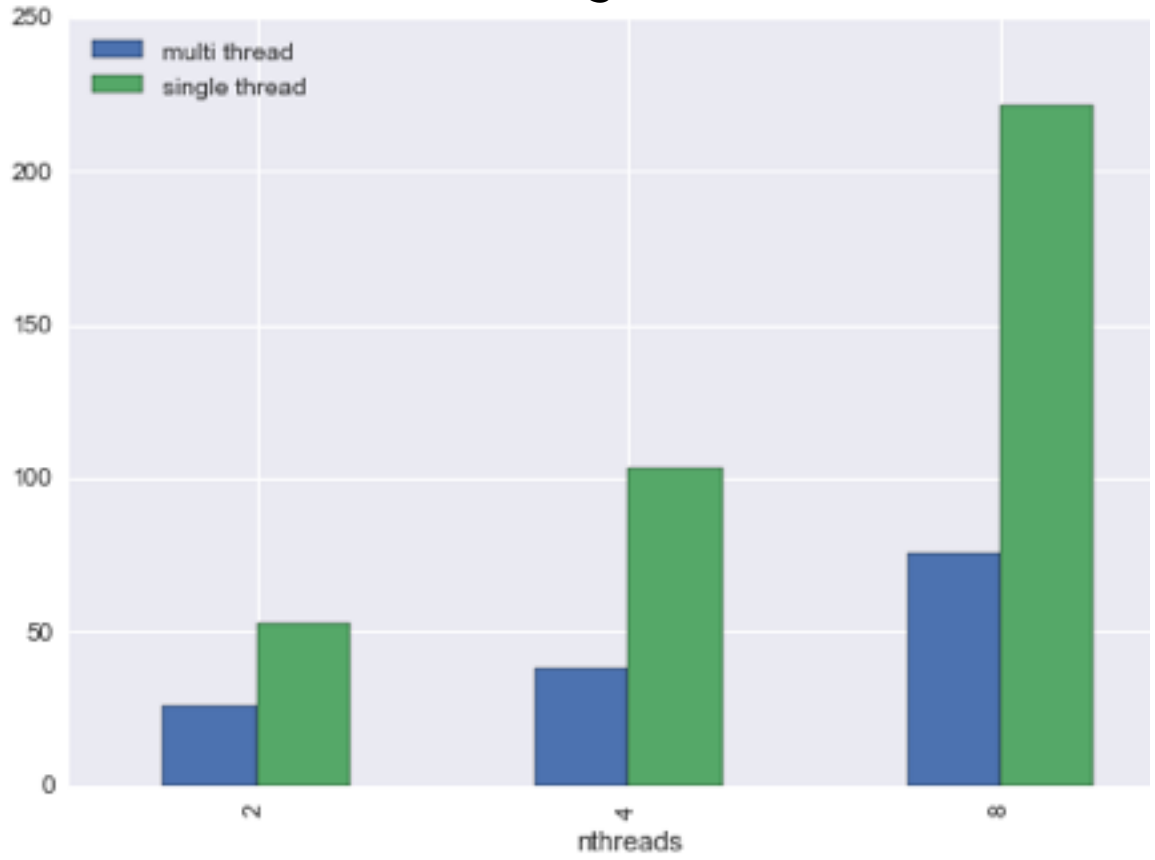
…

```
>>> d_max_sepal_length_setosa = ddf[ddf.species ==
'setosa'].sepal_length.max()

>>> d_max_sepal_length_setosa.compute()

5.7999999999999998
```

Releasing the GIL



Thanks!

https://github.com/jreback/PandasTalks/tree/master/whatsnew/april_2017