

# intake-esm Tutorial

Aaron Spring (MPI-M)

# What will happen today?

- 10' intro: me  you 
- 10' tutorial: tutorial.ipynb 
- 20-100' hands-on: tasks.ipynb 

# What is Pangeo?

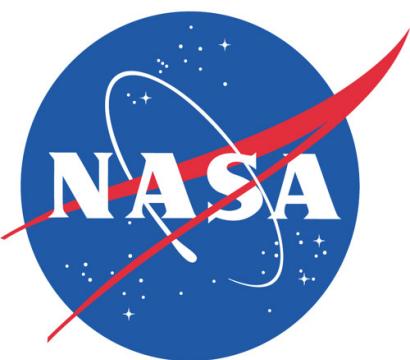
*“A community platform for Big Data geoscience”*

- Open Community
- Open Source Software
- Open Source Infrastructure

# Pangeo Community

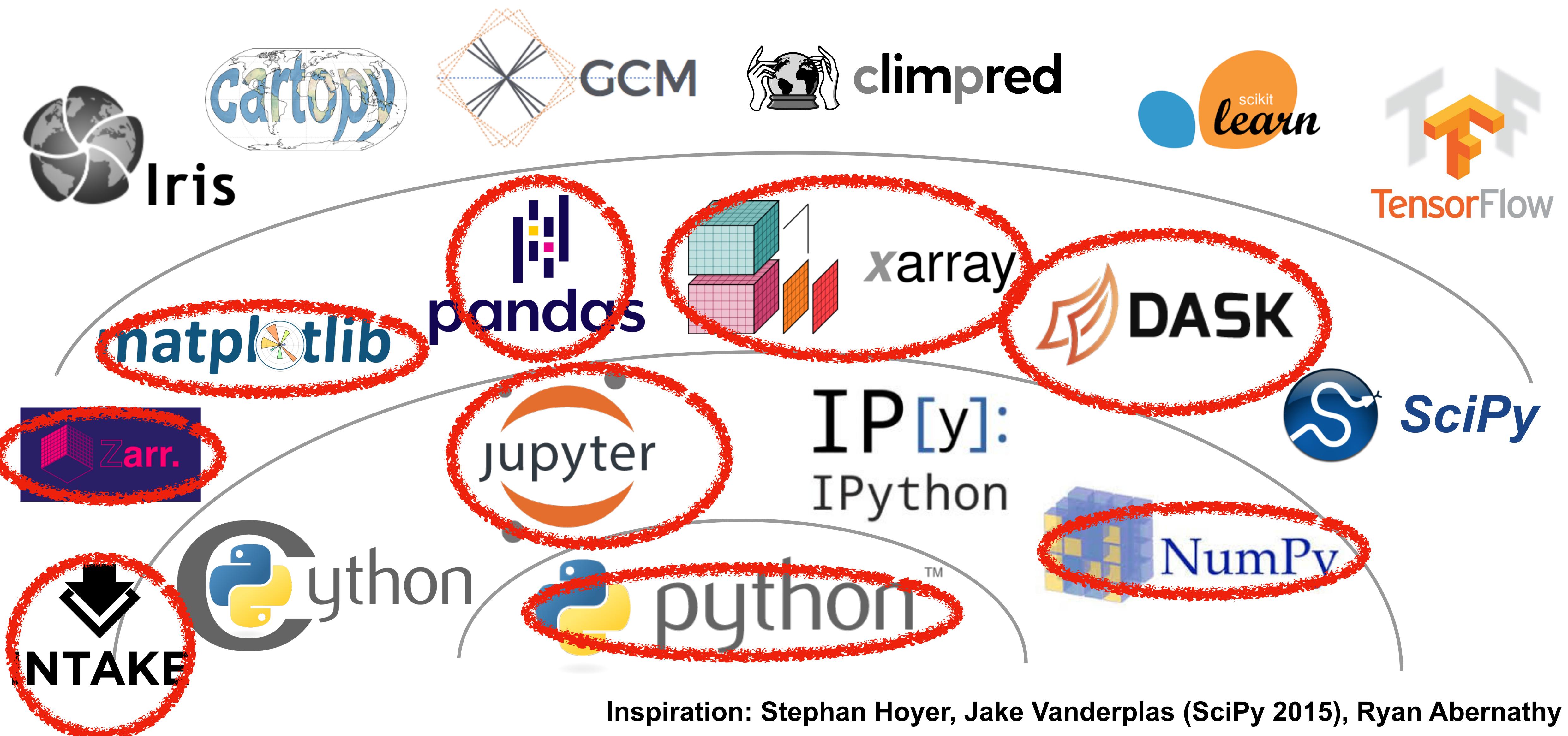


Lamont-Doherty Earth Observatory  
COLUMBIA UNIVERSITY | EARTH INSTITUTE



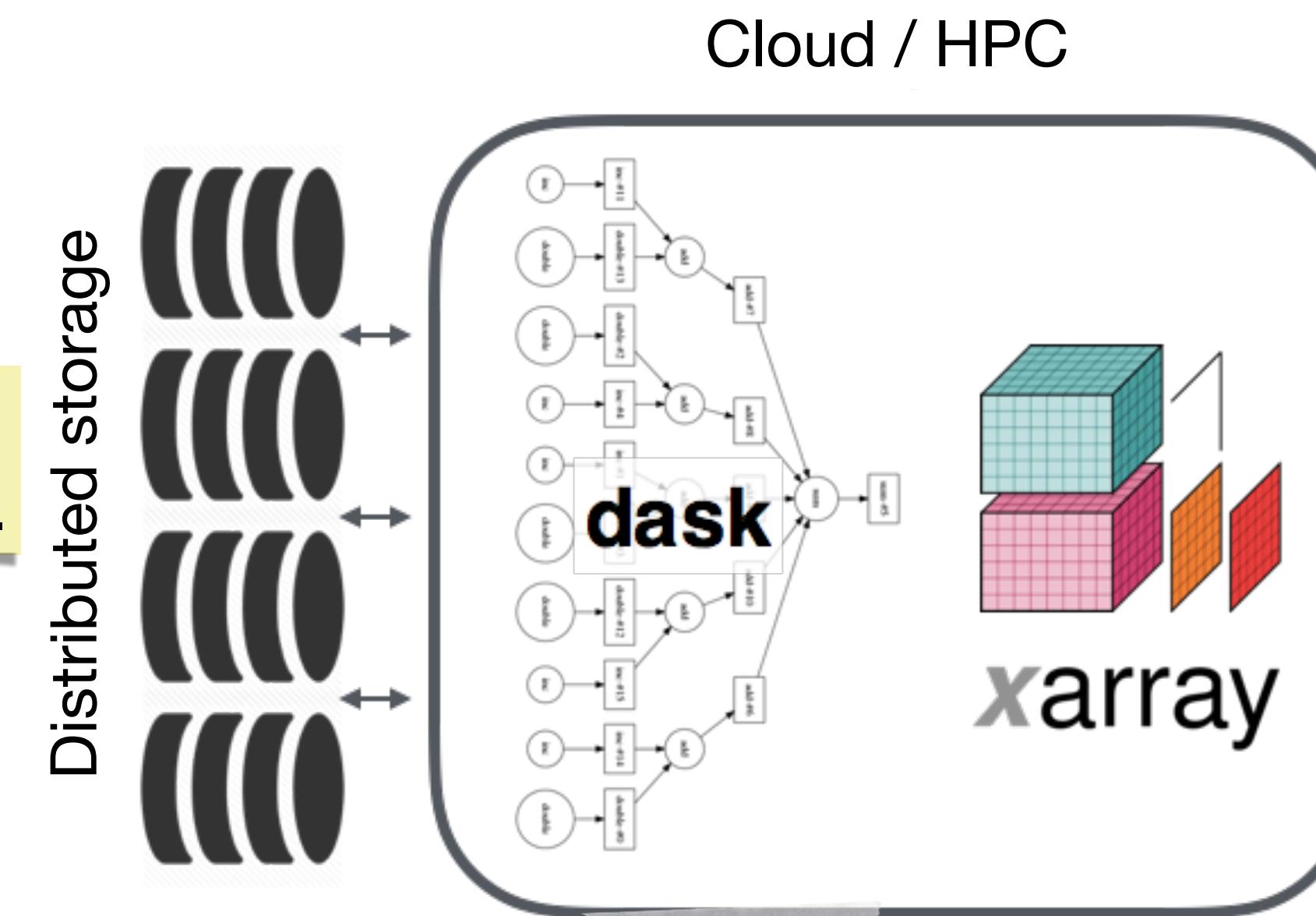
<http://pangeo.io>

# Pangeo Software Ecosystem



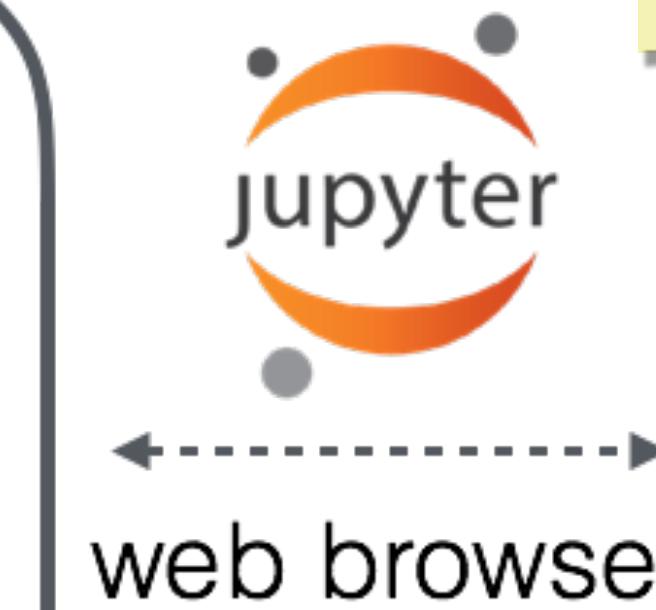
# HPC Architecture

**“Analysis Ready Data”**  
stored on distributed storage.



Parallel computing system allows users deploy clusters of compute nodes for data processing.

Dask tells the nodes what to do.



Jupyter for interactive access remote systems

end user



Xarray provides data structures and intuitive interface for interacting with datasets

# Pangeo Cloud Data Catalog

[catalog.pangeo.io](https://catalog.pangeo.io)

 PANGEO CATALOG    [Blog](#)    [Forum](#)

## PANGEO CATALOG

master

### MASTER

Pangeo Master Data Catalog

<https://raw.githubusercontent.com/pangeo-data/pangeo-datastore/master/intake-catalogs/master.yaml>

### Child Catalogs

ocean  
Pangeo Oceanography Dataset Catalog

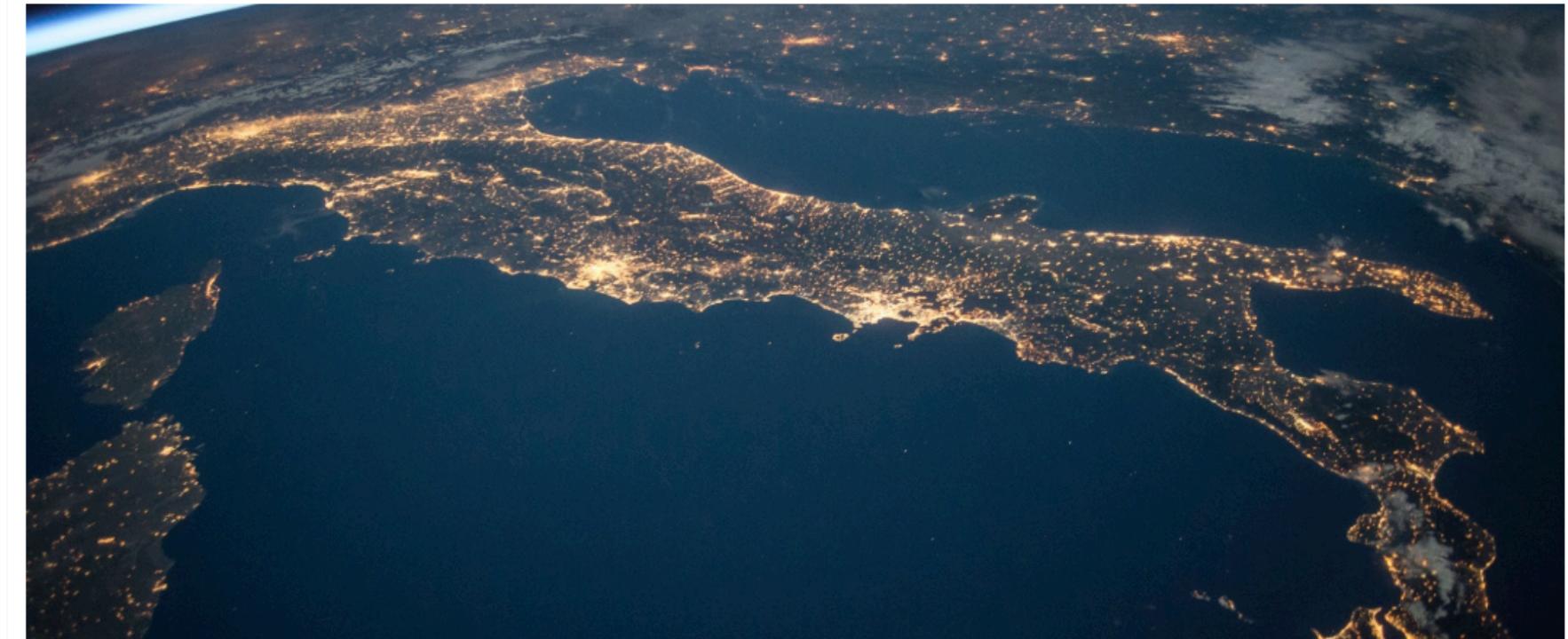
atmosphere  
Pangeo Atmospheric Science Dataset Catalog

climate  
Pangeo Climate Dataset Catalog. Include model ensembles such as CMIP6 and LENS.

hydro  
Pangeo Hydrology Dataset Catalog

DATA ANALYTICS

### New climate model data now in Google Public Datasets



Shane Glass  
Program Manager, Google Cloud Public Dataset Program  
December 9, 2019

Exploring [public datasets](#) is an important aspect of modern data analytics, and all this gathered data can help us understand our world. At Google Cloud, we maintain a collection of public datasets, and we're pleased to collaborate with the [Lamont-Doherty Earth Observatory](#) (LDEO) of Columbia University and the Pangeo Project to host the latest climate simulation data in the cloud.



# Culture of sharing

- Together we achieve more.
- Currently we make reproducibility unnecessary hard.
  - ▶ I try my best: [https://github.com/aaronspring/Spring\\_etal\\_2020\\_Code](https://github.com/aaronspring/Spring_etal_2020_Code)
- Let's put code examples out there:
  - ▶ [gitlab.dkrz.de](https://gitlab.dkrz.de)
  - ▶ [github.com](https://github.com)
  - ▶ [gist.github.com](https://gist.github.com)

# intake-esm

- Developed at NCAR, mostly by Anderson Banihirwe (**andersy005**)
- <https://intake-esm.readthedocs.io/en/latest/>
- Taking the pain out of data access: [link to wiki.mpimet](#)
- Search and load ESM output:
  - ▶ Query in `pandas.DataFrame`
  - ▶ Load data with `dask` into `xarray`

# Live demo



tutorial.ipynb

 launch binder

# Try intake-esm yourself!



OR



```
git clone https://gitlab.dkrz.de/  
m300524/lunchbytes_intake-esm
```