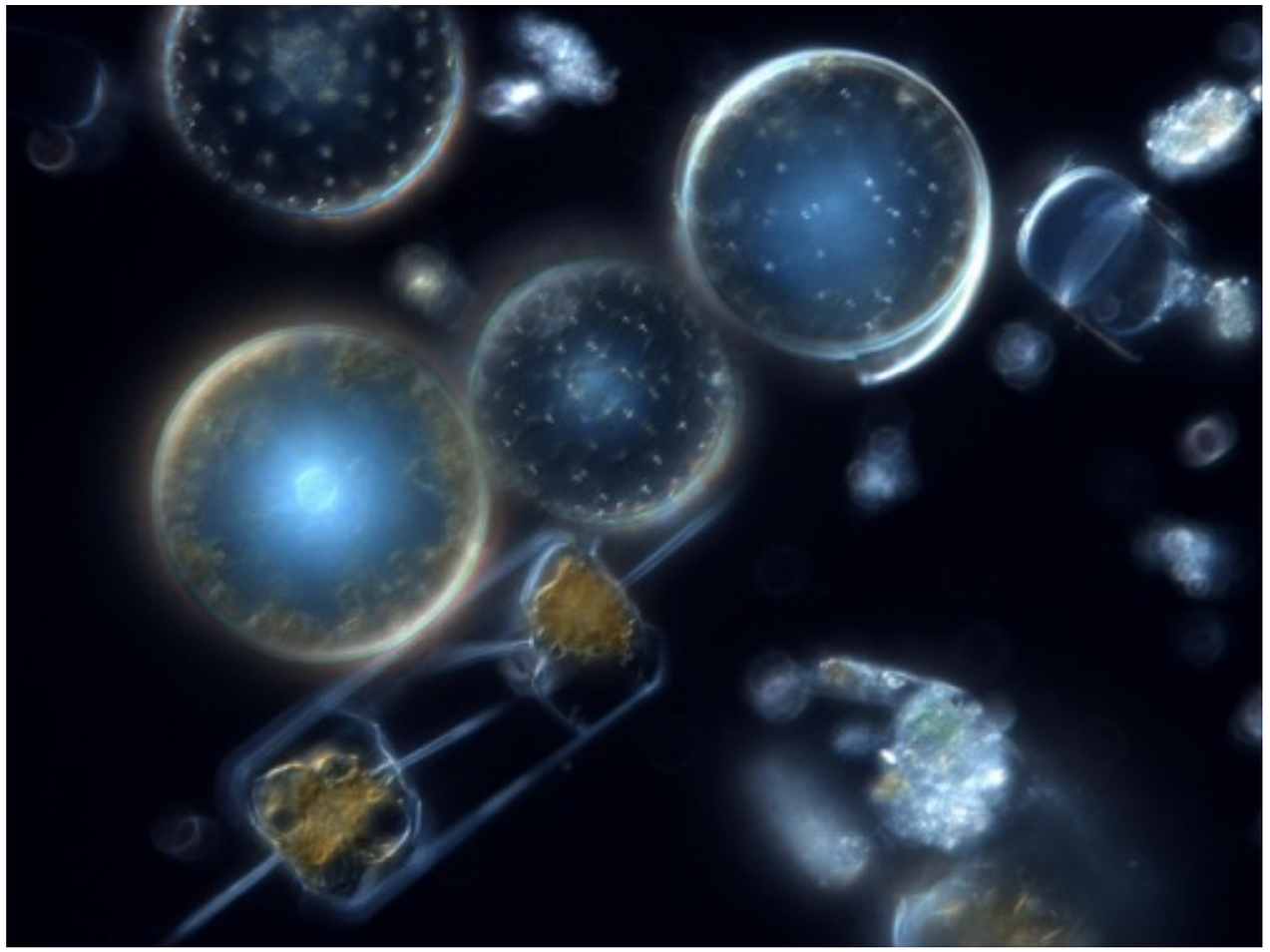
Analysis of large-scale patterns in phytoplankton diversity

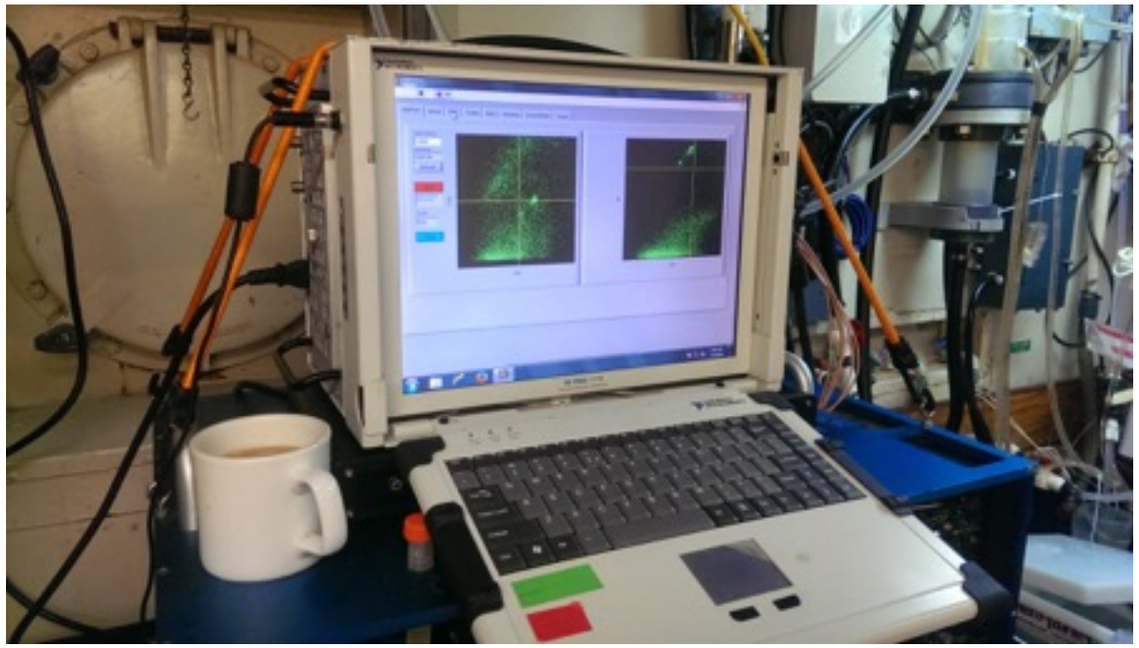
Sophie Clayton
sclayton@uw.edu

December, 2014



School of Oceanography
University of Washington

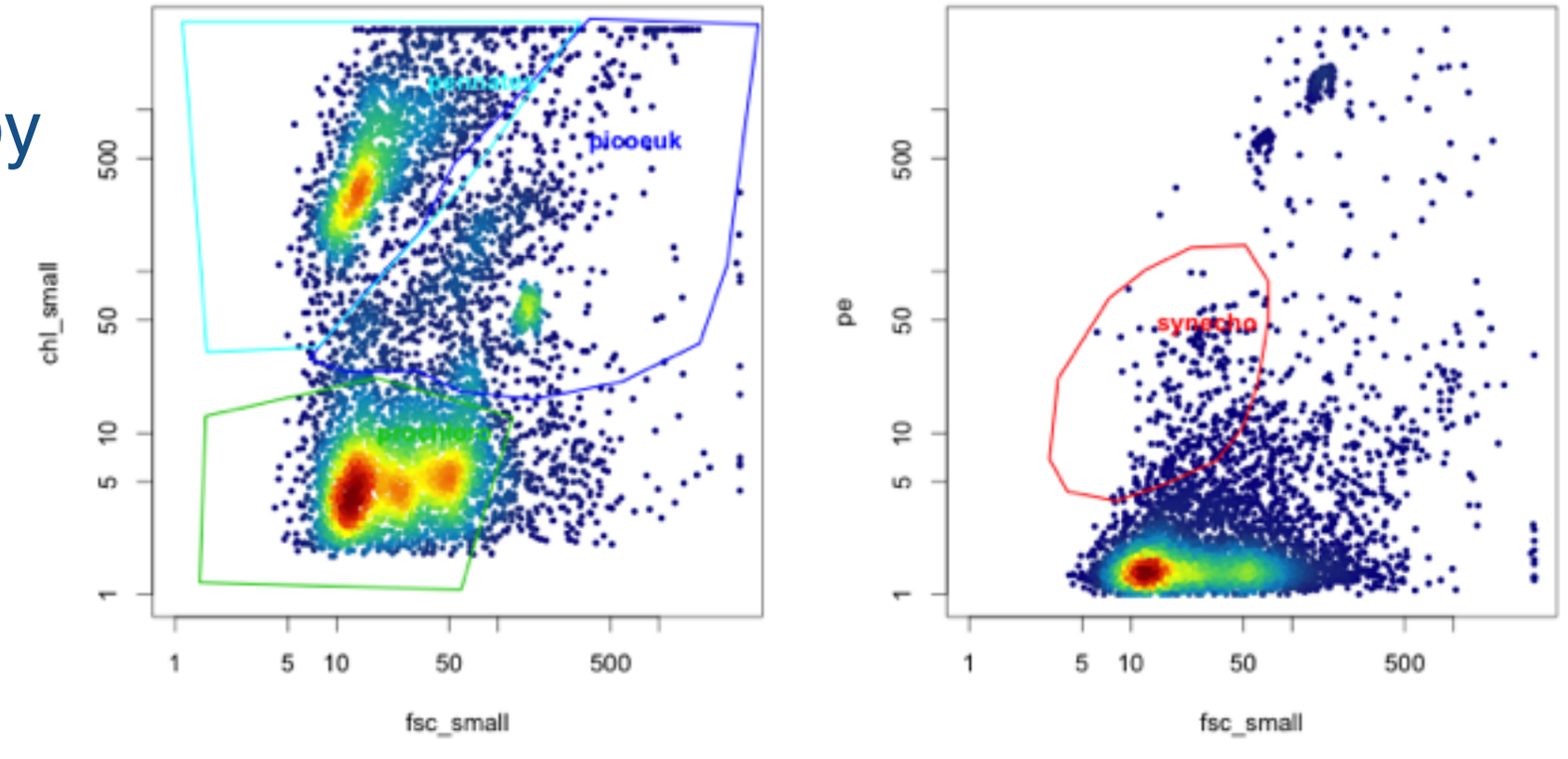
SeaFlow flow cytometer



SeaFlow in action aboard the R/V Melville

- Measures properties of particles in seawater.
- Data collected continuously and stored every 3 minutes.

- Phytoplankton identified by their size and pigments.
- Beads used as internal standard.
- Estimate diversity from cytogram.

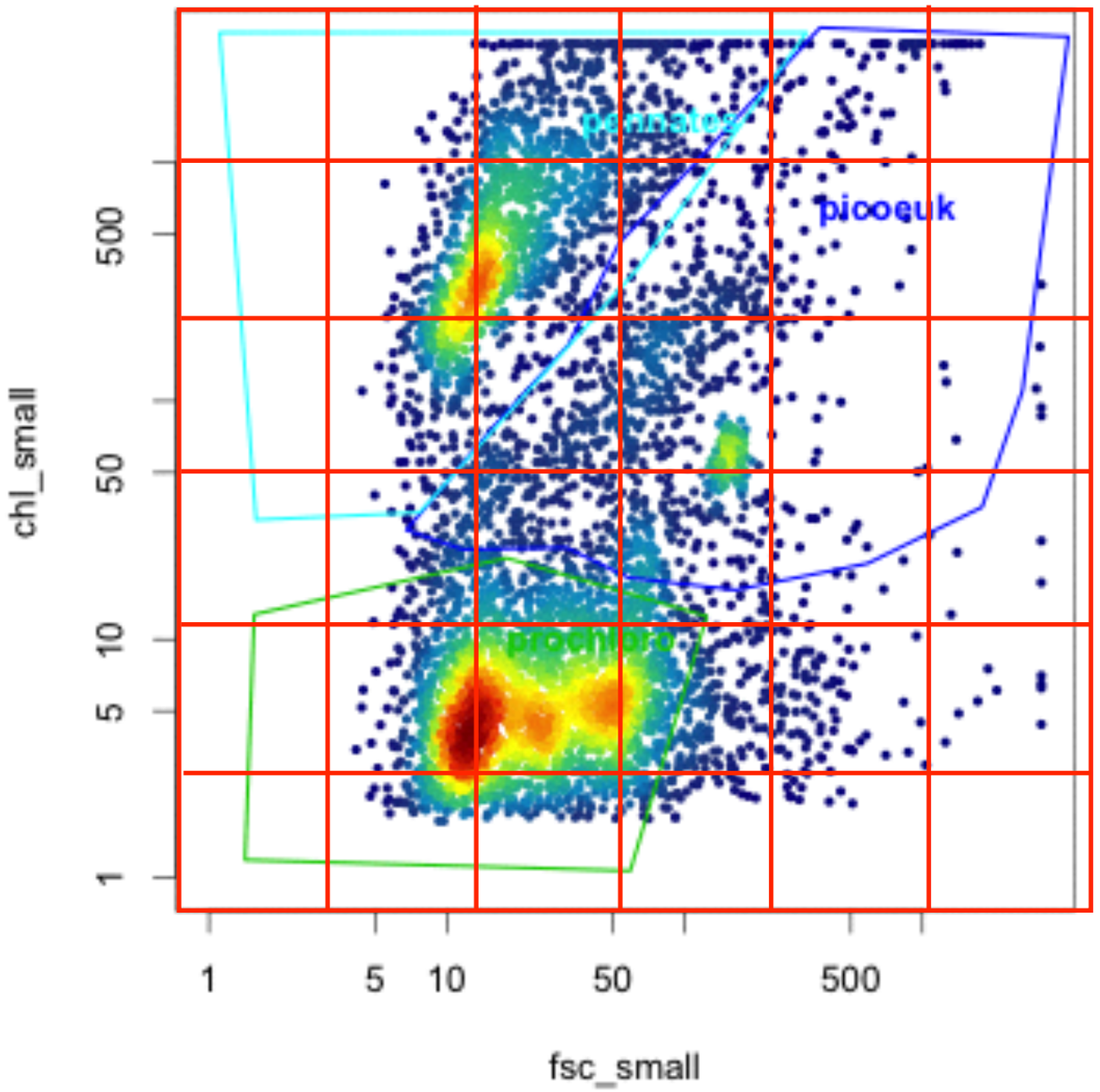


Example cytogram with population assignment

Diversity index from SeaFlow data

Taxonomic determinations of diversity are very labour-intensive.

Diversity indices are related to the area of the cytogram “occupied” by particles - doesn’t require formal identification.



Particles in each cytogram are binned to estimate diversity of observed population

The SeaFlow dataset: overview

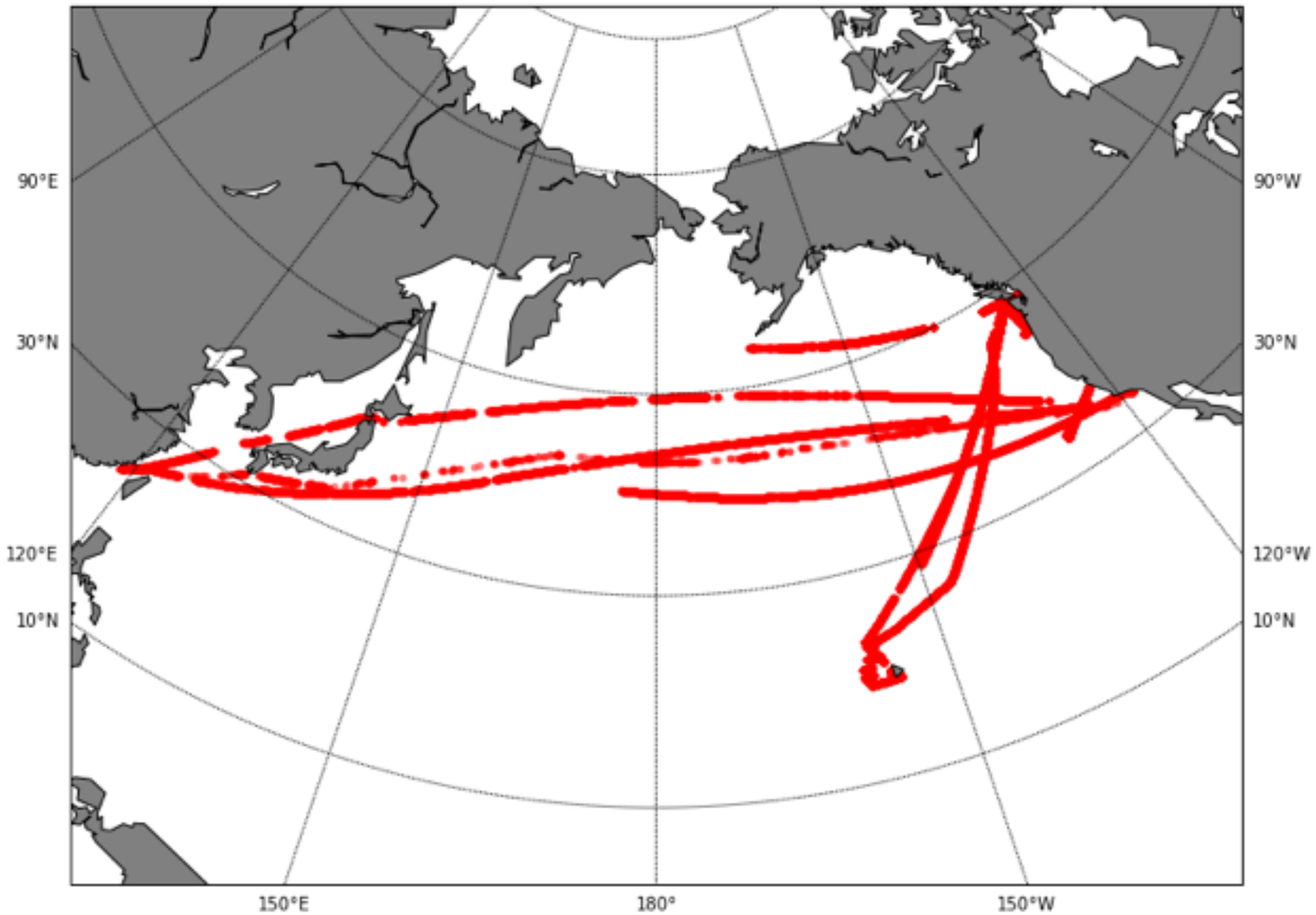
Goal: to explore patterns in phytoplankton diversity and their relationship to the environment in the North Pacific.

Data from 18 cruises in the North Pacific

- > 500GB data
- 1.7B particles
- > 50,000 data points

Data types:

- OPP - particle properties
- VCT - population assignment
- SDS - environmental data

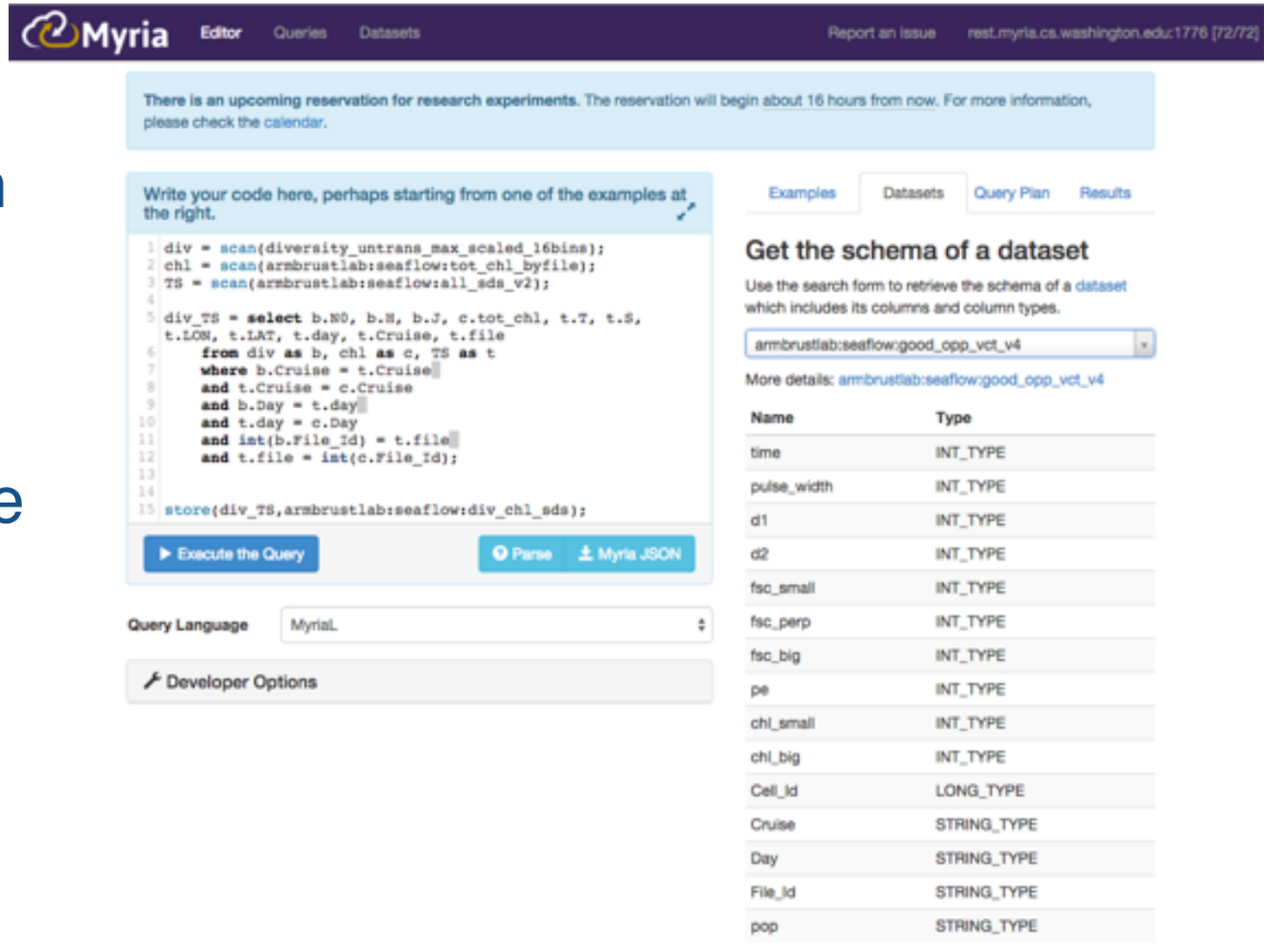


Map of all SeaFlow data points

The SeaFlow dataset: Myria and SQLshare

- OPP and VCT data already uploaded to Myria
- SQLShare for interpolating SDS data onto OPP timestamp

- OPP, VCT and SDS data can be joined using Cruise, Day and File_ID.
- Myria used for filtering data and computationally-intensive queries (e.g. re-scaling, binning)



MyriaDB vs. R: calculating diversity

Run time: several hours

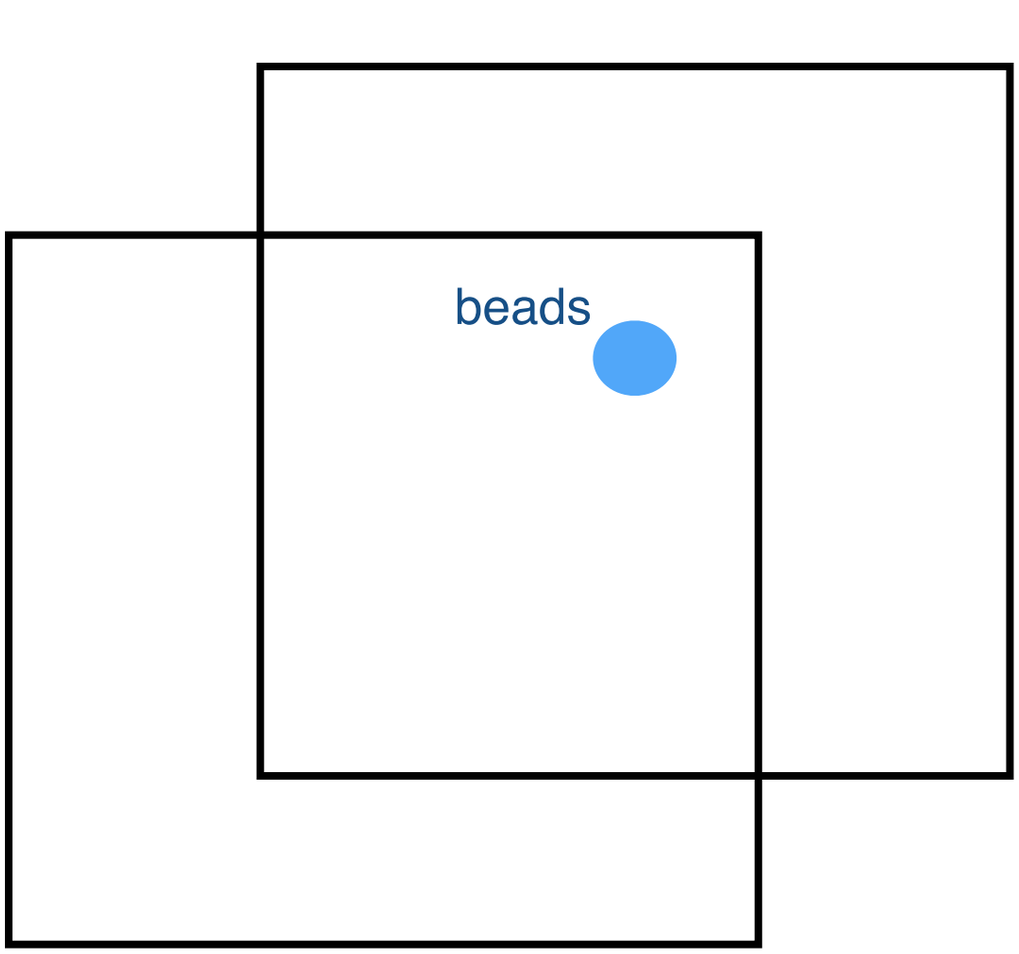
```
-- Load the existing dataset
AllData = scan(armbrustlab:seaflow:good_opp_vct_v4);

-- Assign a linear value into one of 16 bins 0..15
-- N.B.: // is integer division
def makebins(x): x//(pow(2, 16)/16);

-- For each cruise & sample (day + file _id)
-- break the 3-D cytogram given by forward scatter,
-- chlorophyll, and phycoerythrin into a 16x16x16
-- bin space and count the number of cells in each bin.
AllDataBinned = select Cruise, Day, File_Id,
                       makebins(fsc_small) as fsc_bin,
                       makebins(chl_small) as chl_bin,
                       makebins(pe) as pe_bin,
                       count(*) as num_particles
                  from AllData;

-- Compute the Richness N0 as the number of full bins
Richness = select Cruise, Day, File_Id,
                  count(*) as richness
             from AllDataBinned;

store(Richness, richness_untrans);
```

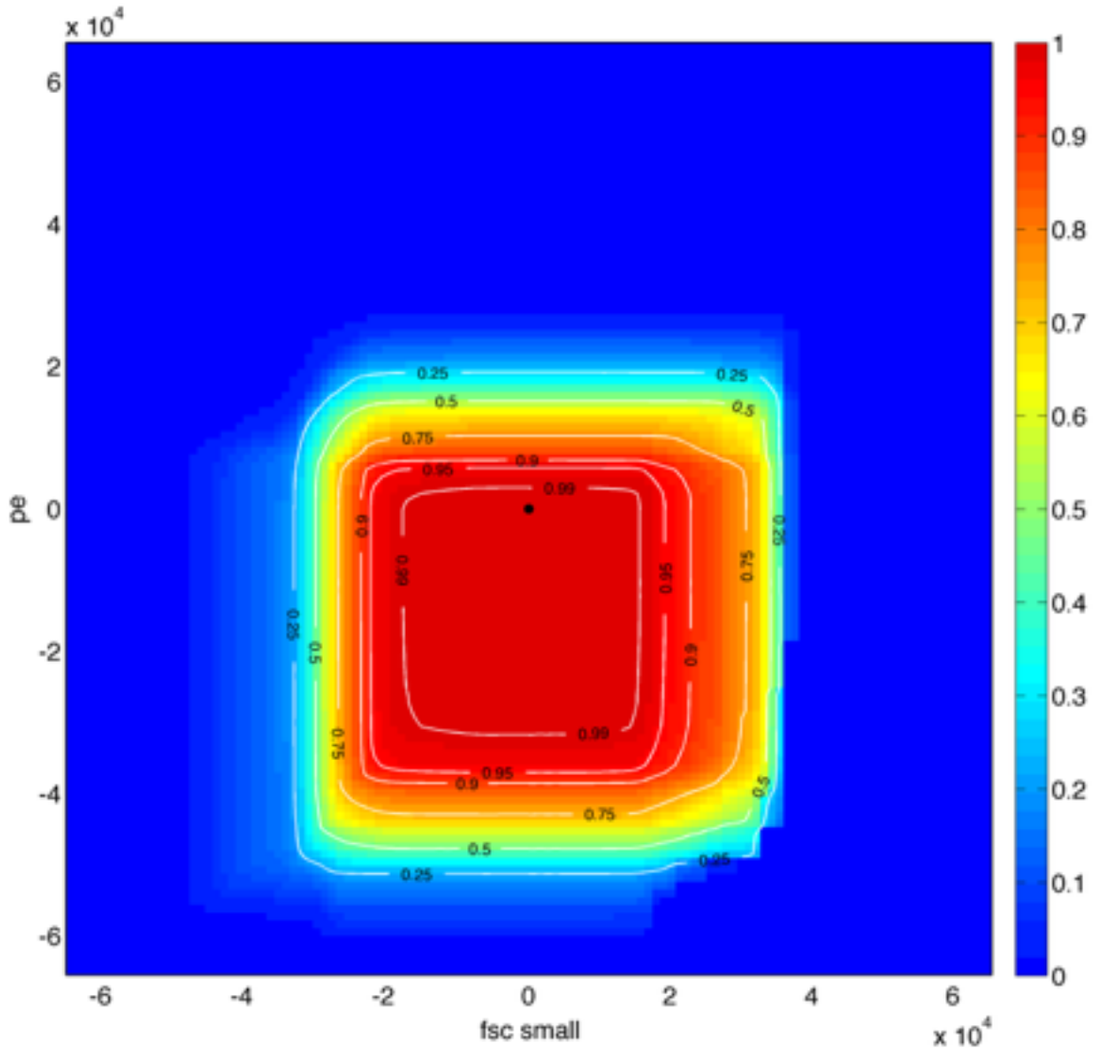
Run time: ~10 minutes

Normalizing SeaFlow data

Need to account for variable SeaFlow settings, and different instruments.

Use beads to standardize.

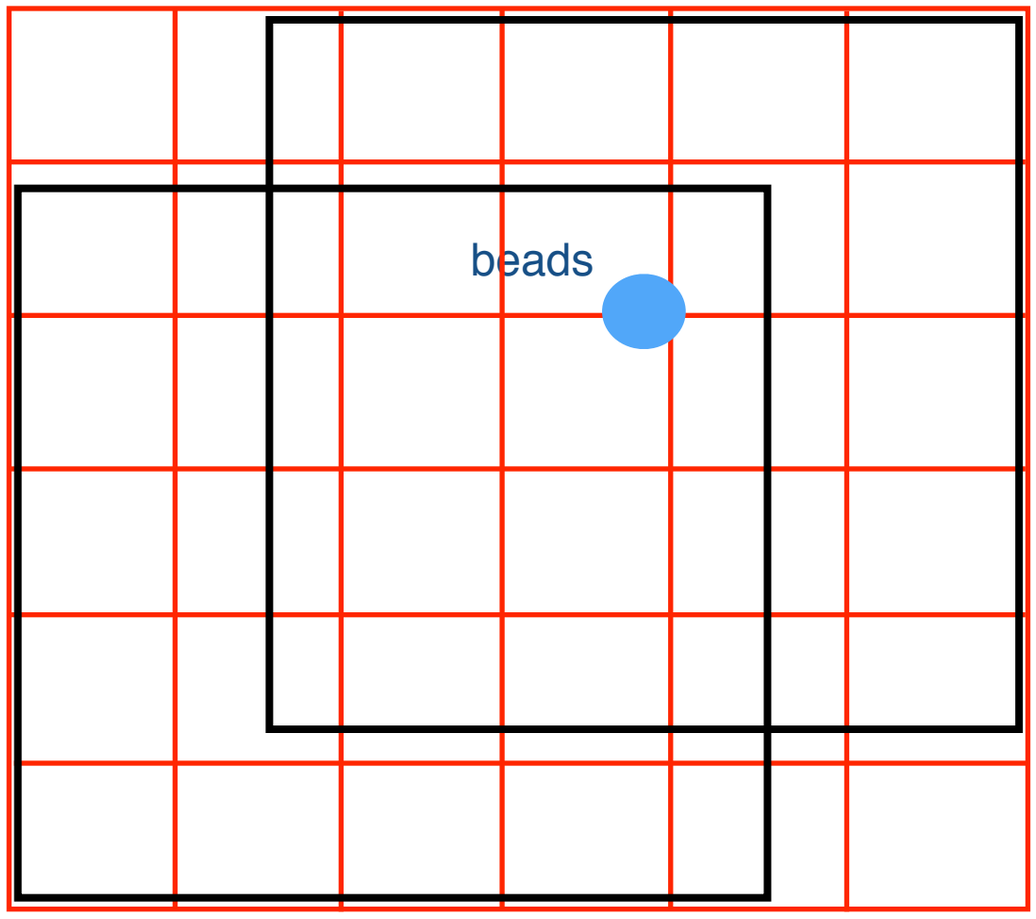




Proportion of file coverage around bead position

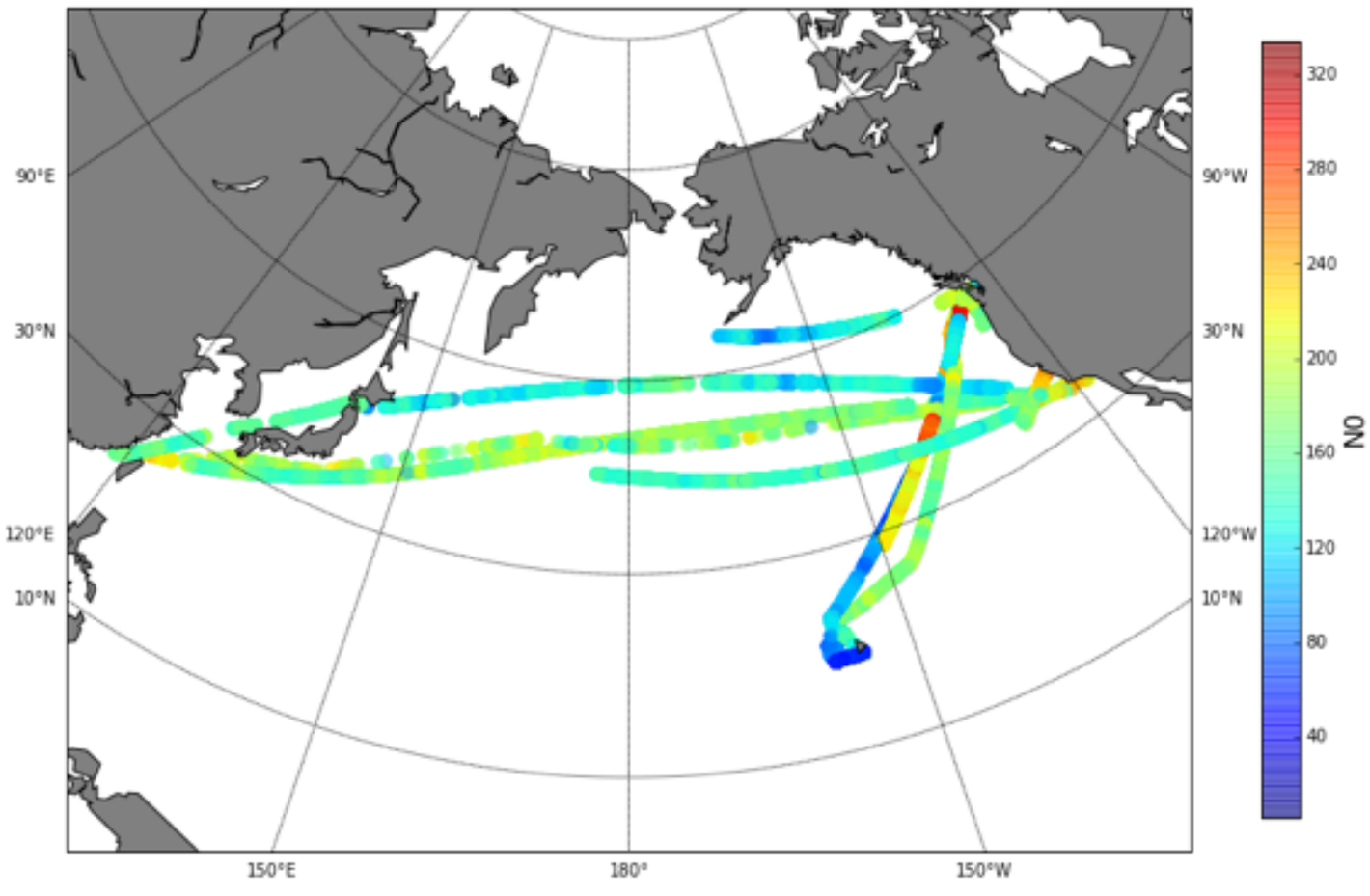
Particle properties are re-scaled to be relative to the value of the bead properties in each file.

Normalized estimates of diversity



Example of how the normalized data is binned

Diversity indices calculated over re-gridded normalized OPP data.



Map of bead normalized cytometric diversity

Further analysis and visualization

Download data directly from Myria REST server into Python.

Document analysis and visualization in iPython notebooks stored in GitHub repo.

IP[y]: Notebook myria_N0_latitude_bins Last Checkpoint: Nov 21 11:37 (autosaved)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

```
from mpl_toolkits.basemap import Basemap, cm
from netCDF4 import Dataset as NetCDFFile
```

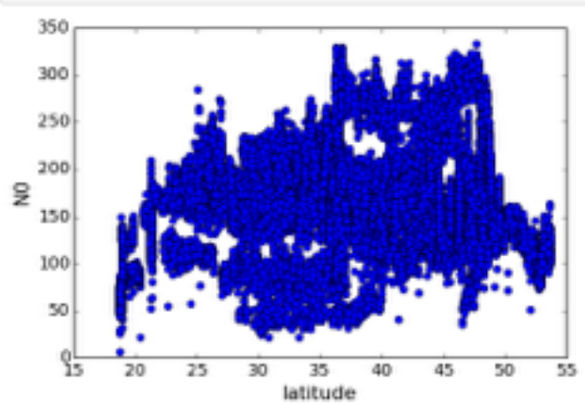
In [2]:
```
# import data directly from myria

# myria query:
# beads_TS = SELECT b.fsc_avg, t.LAT, t.LON
# FROM SCAN(armbrustlab:seaflow:bead_stats_v4_byfile_untrans) AS b,
# SCAN(armbrustlab:seaflow:all_ads_v2) AS t
# WHERE b.Cruise = t.Cruise AND b.Day = t.day AND INT(b.File_Id) = t.file;
# STORE(beads_TS,armbrustlab:seaflow:beads_LATLON);

fileURL = 'https://rest.myria.cs.washington.edu:1776/dataset/user-armbrustlab/program-seaflow/relatic
result = pd.read_csv(fileURL)
print(result.columns)
```

```
Index([N0, H, J, T, S, LON, LAT], dtype=object)
```

In [63]:
```
# plot the mean bead fsc_small value in space
# the data

# pick the variable you want to plot
value = 'N0'

x = (result['LON']).values # longitude
y = (result['LAT']).values # latitude
z = (result[value]).values # bead attribute

# convert all longitude to longitude east
x[x<0]= 360+x[x<0];

print(max(x))
```

```
241.5523
```

In [64]:
```
fig1 = plt.figure(1, figsize=(6,4))

plt.axes([0.025, 0.025, 0.95, 0.95])
plt.plot(y,z, 'o')

plt.xlabel('latitude',fontsize=16)
plt.ylabel(value,fontsize=16)

plt.tick_params(axis='both', which='major', labelsize=14)

plt.show()
```

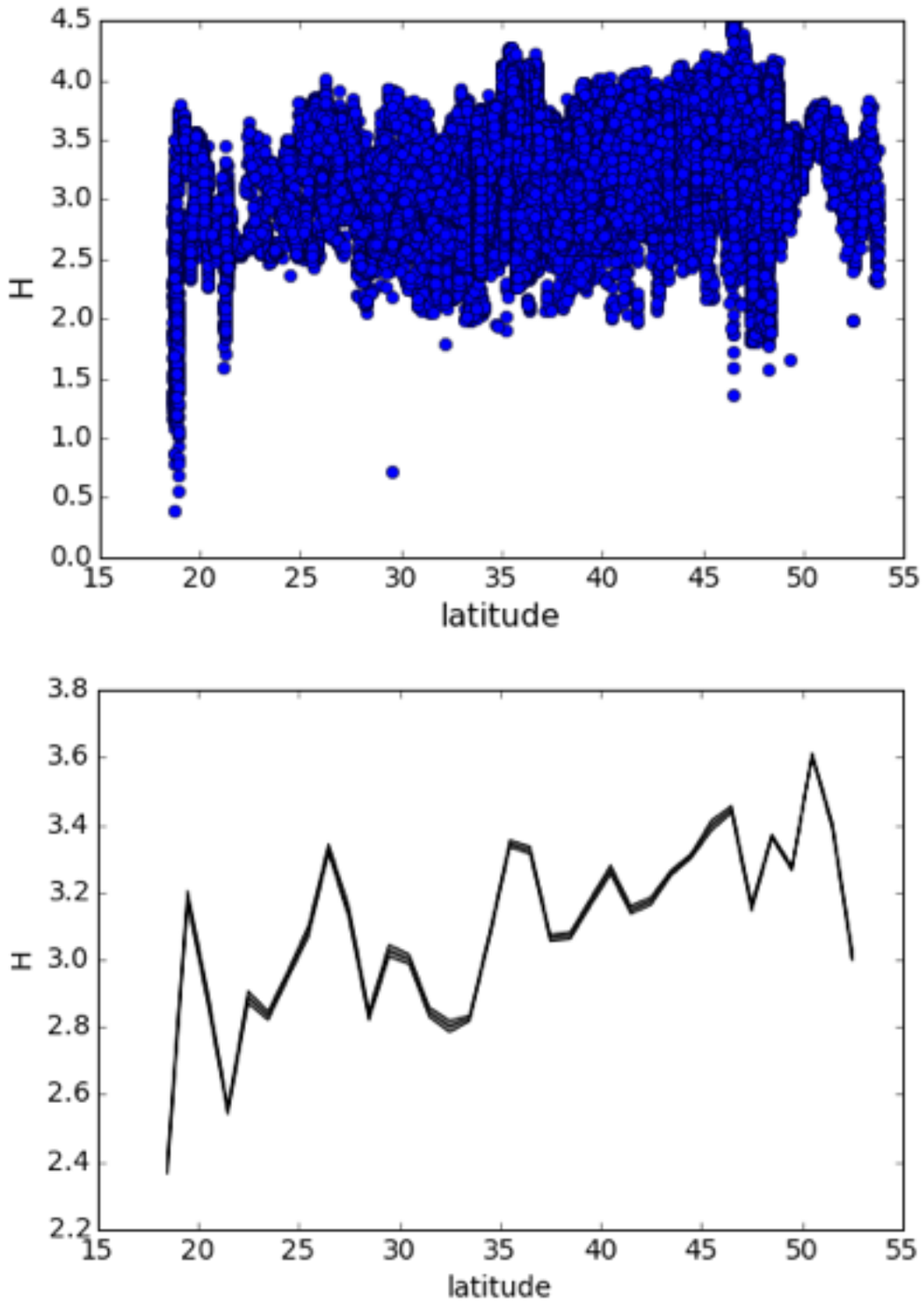


Conclusions & Outlook

We have developed tools to work with the SeaFlow dataset in Myria and Python.

Next steps:

- explore seasonal patterns in data
- split data into coastal vs. open ocean sets
- sensitivity analysis of the binning parameters for the diversity indices



Zonally averaged cytometric diversity

Thanks!

- Dan Halperin
- Bill Howe
- eScience Fall 2014 Incubator staff and participants

sophieclayton commented 10 days ago

just tried this query:

```
good_opp_vct = scan(armbrustlab:seaflow:good_opp_vct_v4);

day = select Cruise, File_Id, substr(Day, 5, len(Day)) as yearday
        from good_opp_vct;

store(day, armbrustlab:seaflow:yearday_opp_vct);
```

and got this error:
Error 400 (Bad Request): Error 400 (Bad Request): The second argument of substr has to be INT.

Since when is 5 not an integer?!...

dhalperi commented 10 days ago — Owner

wtf? I will look into this.