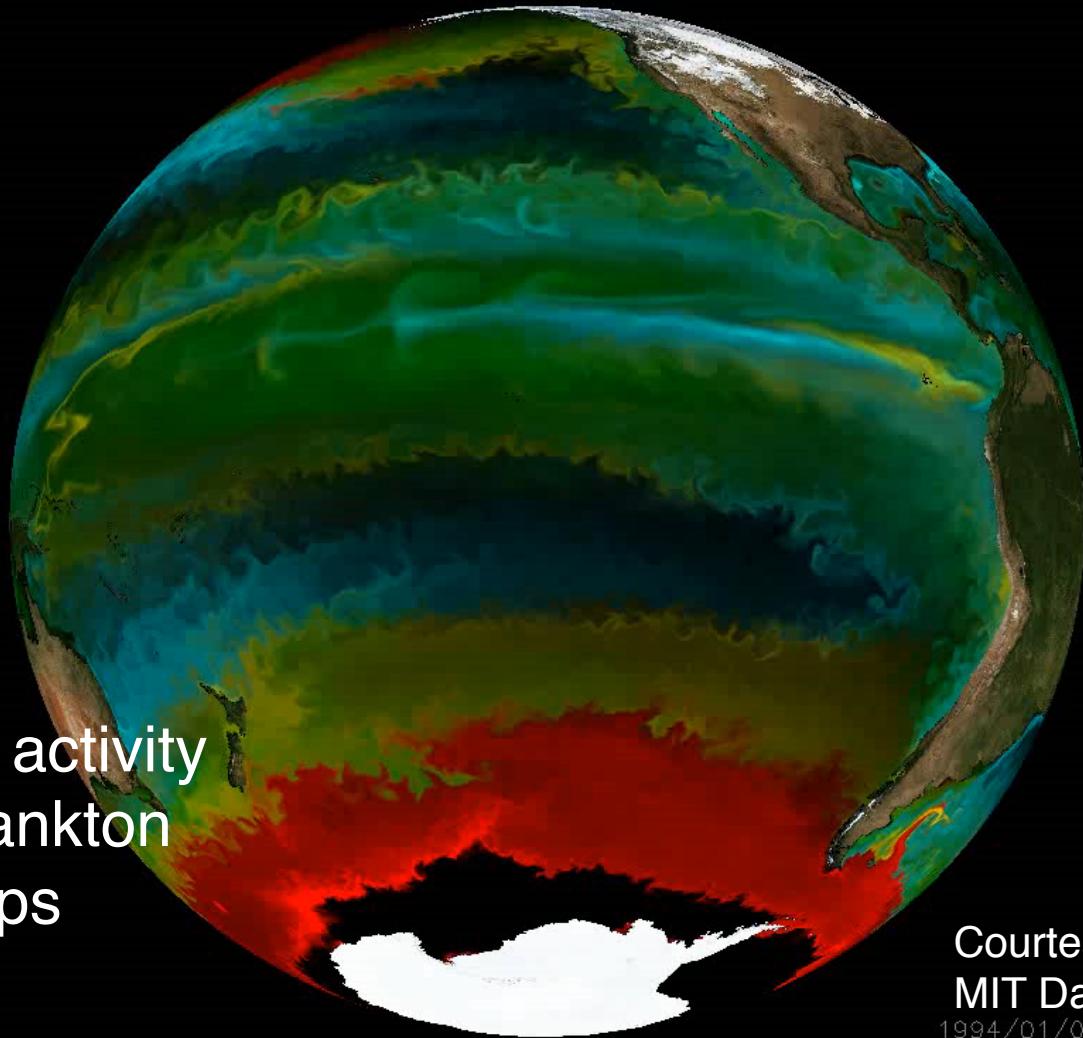


# Harnessing the power of scientific Python to investigate metaproteomes and biogeochemistry of the tropical Pacific

Noelle Held, Jaclyn Saunders, Joe Futrelle, Mak Saito  
and the Ocean Protein Portal Team  
Woods Hole Oceanographic Institution



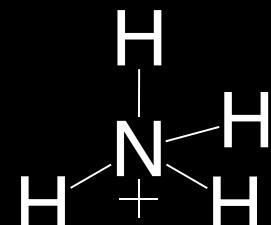
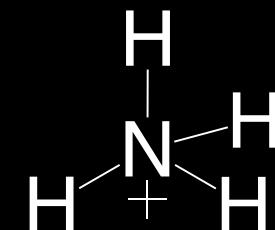
Photosynthetic activity  
of four phytoplankton  
functional groups  
over time



Courtesy of Mick Follows/  
MIT Darwin Project  
1994/01/01



Dinitrogen

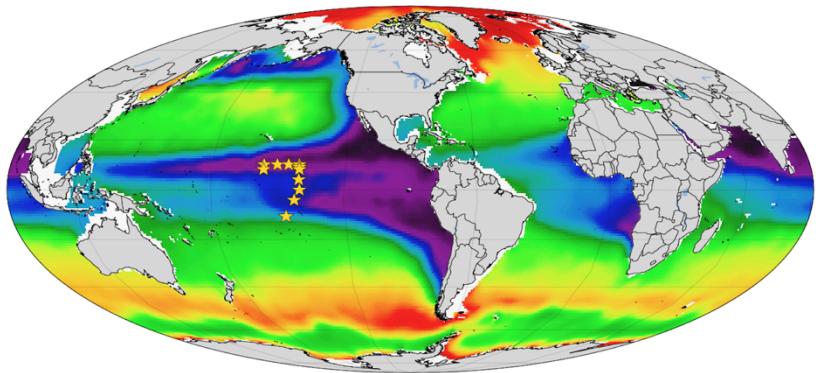
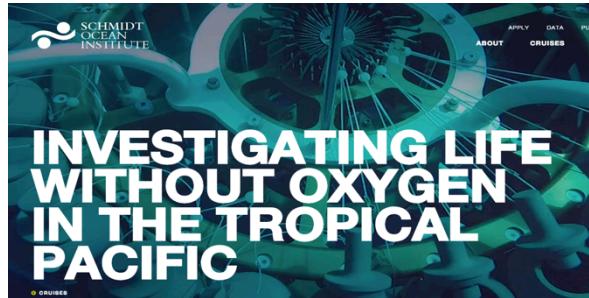


Ammonia (2)

Crystal structure of nitrogenase enzyme  
Tezcan et al 2005 (Science)

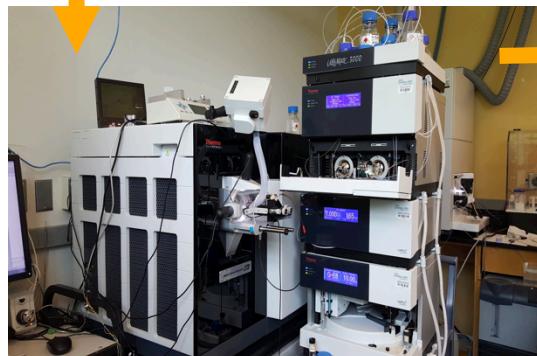
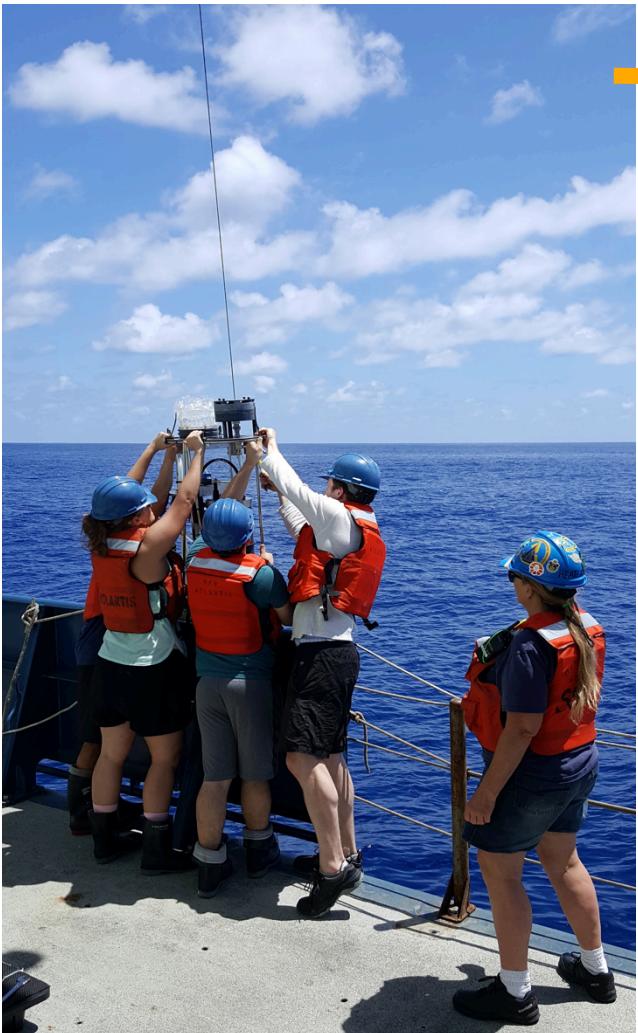
# The ProteOMZ research expedition

OMZ = “Oxygen Minimum Zone”  
aka “Oxygen Deficient Zone”



Oxygen Data: World Ocean Atlas

For more background and videos of sampling efforts in action, see:  
<https://schmidtocean.org/cruise/investigating-life-without-oxygen-in-the-tropical-pacific/>



Analysis by tandem mass spectrometry

**Abundance data  
of 60,000 unique  
microbial  
proteins**

↑  
**Bioinformatics  
analyses/protein  
and peptide  
identifications**

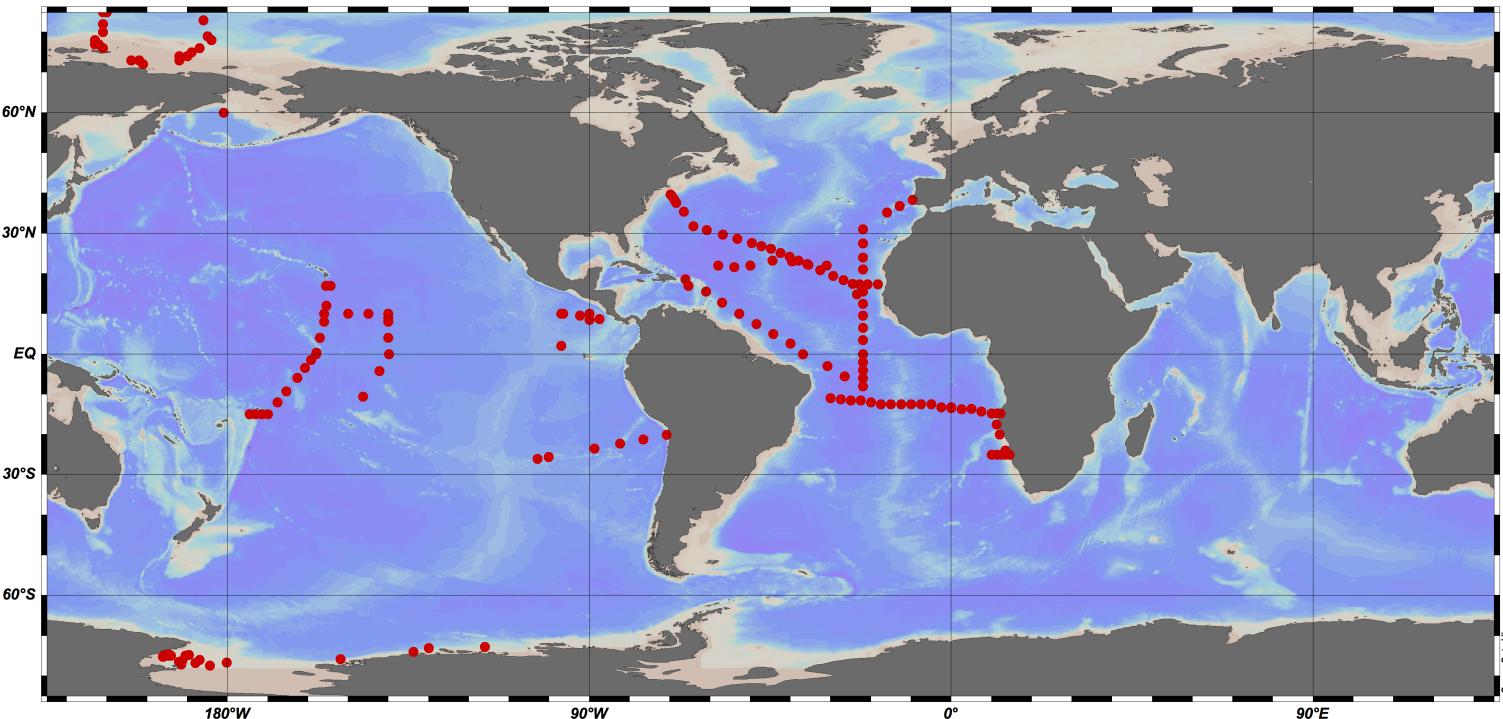
11 Stations

8-16 Depths per  
Station

60,000 proteins  
per depth

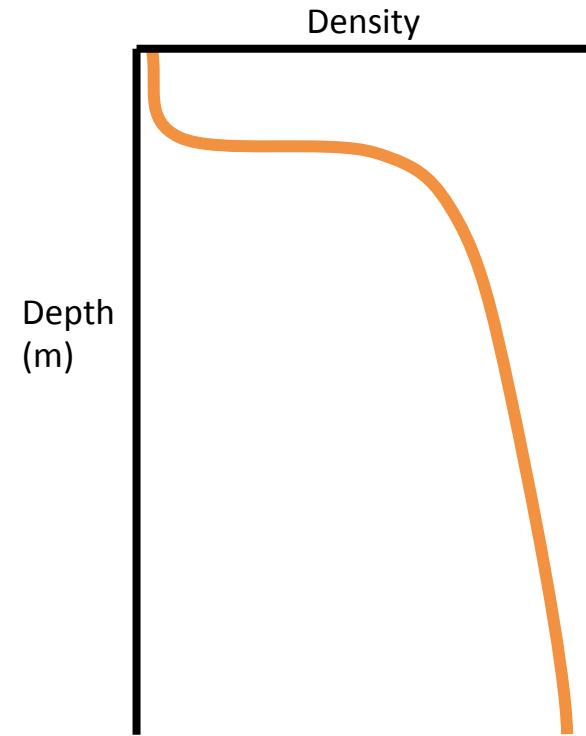
**= 6 million data points  
+ 6,600 contextual data  
points (chemical and  
physical features)**

# The data analysis bottleneck is just beginning



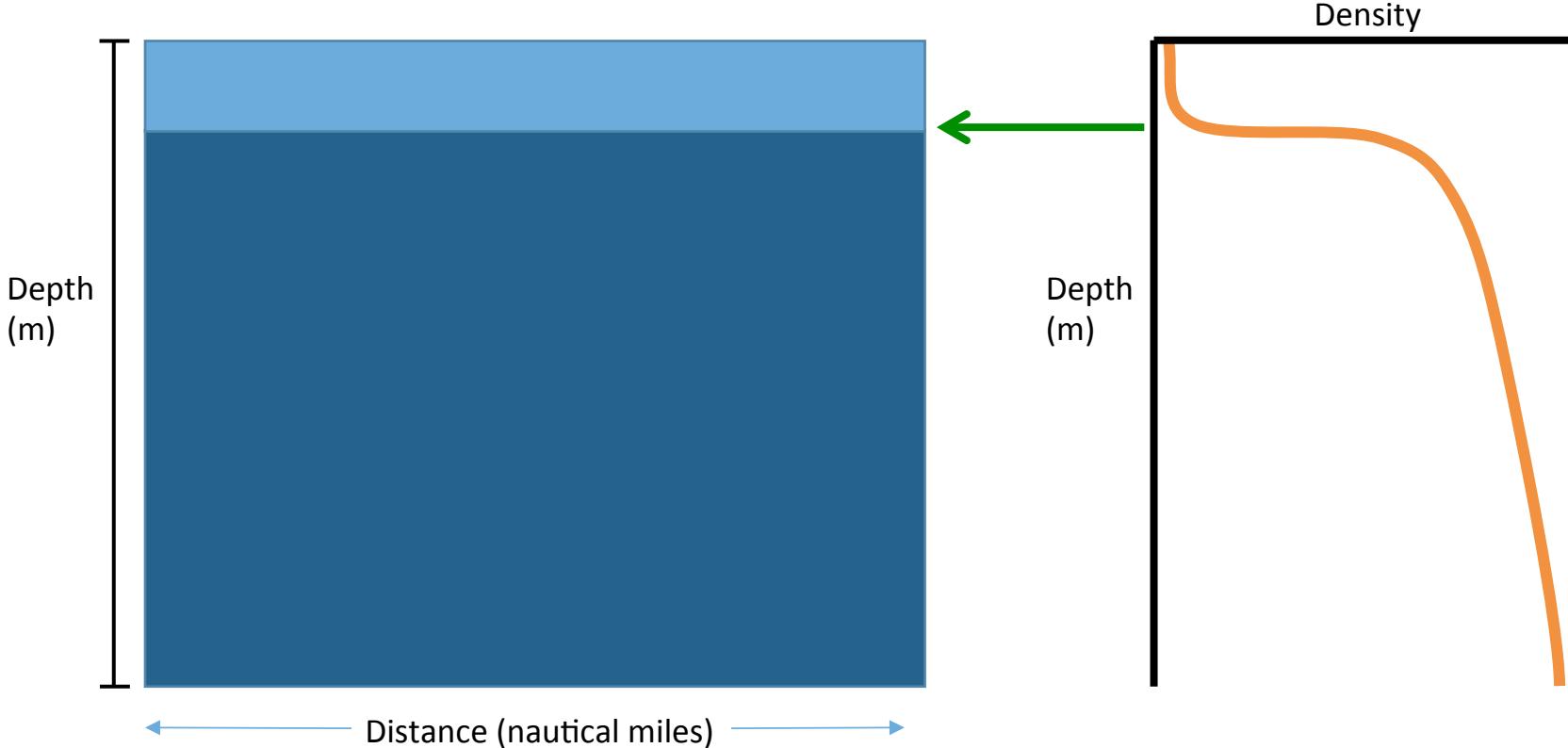
Locations of metaproteome samples in Saito lab collections

# Oxygen Deficient Zones





# Oxygen Deficient Zones



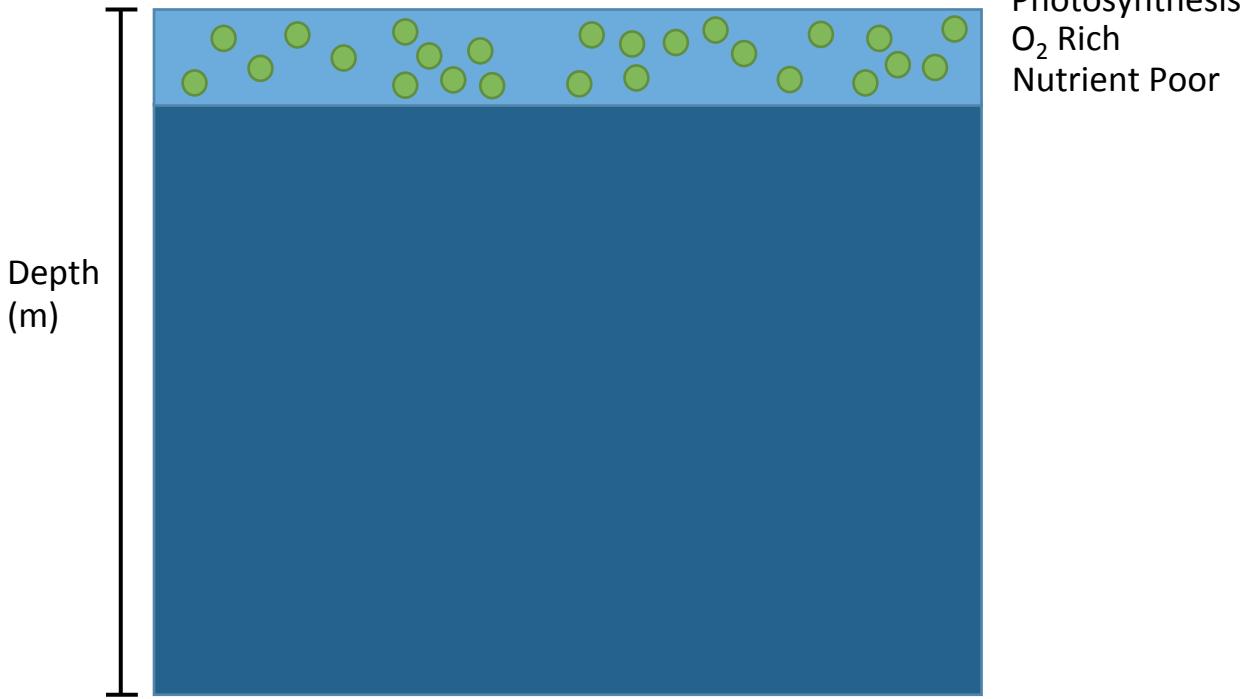


# Oxygen Deficient Zones



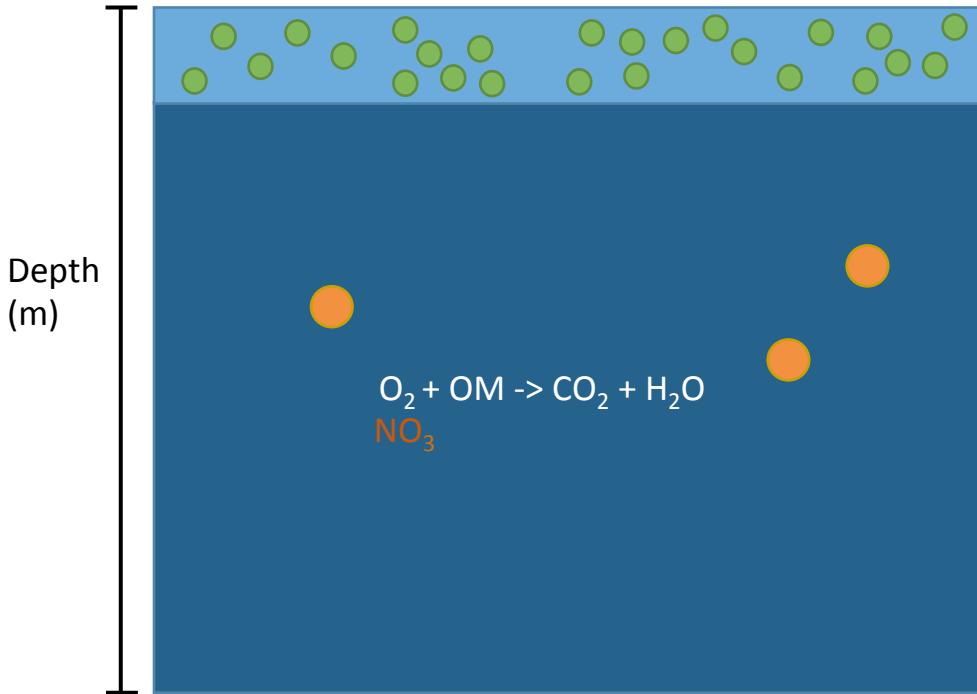


# Oxygen Deficient Zones

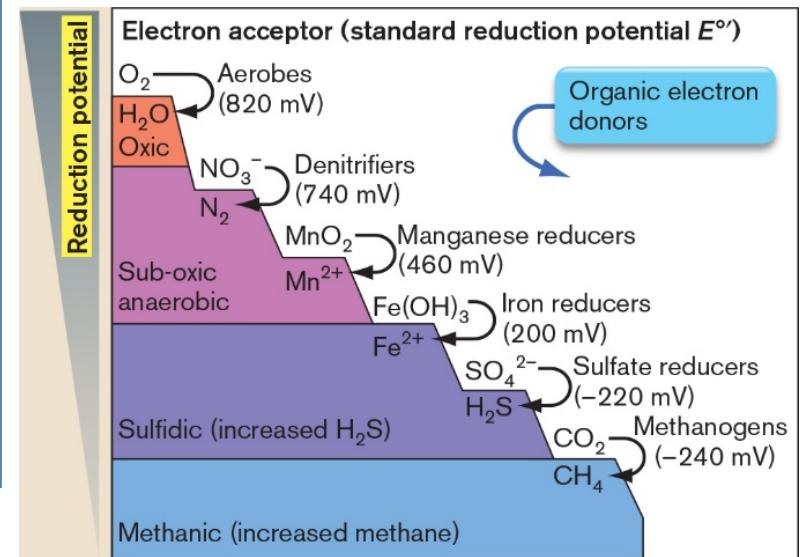




# Oxygen Deficient Zones



Photosynthesis  
O<sub>2</sub> Rich  
Nutrient Poor  
Nutrient Rich  
O<sub>2</sub> Poor



# Visualizing Hydrographic Features

[https://github.com/maksaito/  
proteOMZ hydrography visualization](https://github.com/maksaito/proteOMZ_hydrography_visualization)

# Visualizing Metaproteomics Data

[https://github.com/maksaito/  
proteOMZ\\_visualization\\_app\\_public](https://github.com/maksaito/proteOMZ_visualization_app_public)

See also:

[https://github.com/naheld/15000lines\\_datashader](https://github.com/naheld/15000lines_datashader)

## Benefits of Bokeh:

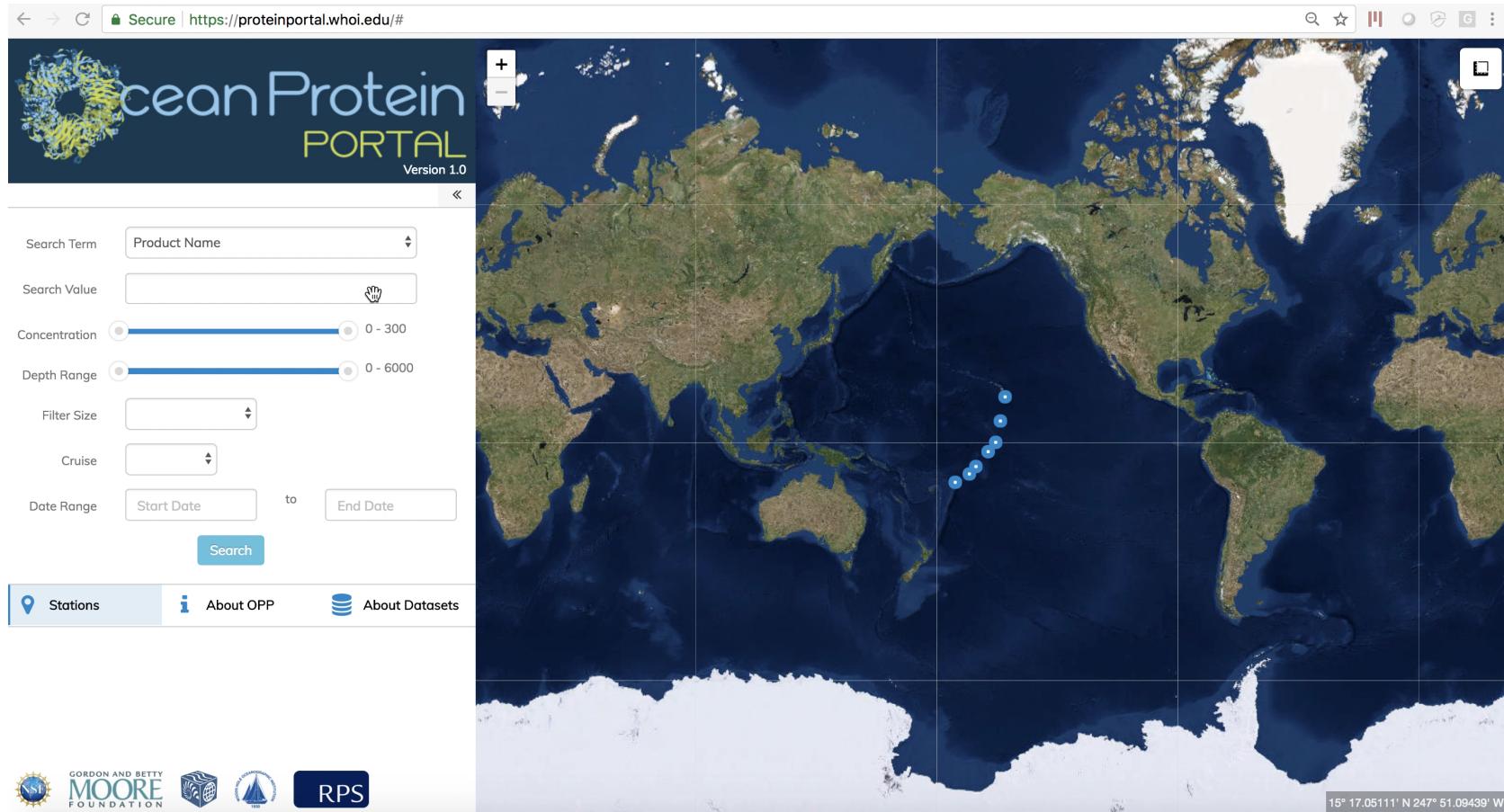
- Beautiful plots “out of the box”
- Interactive
- HTML outputs for colleagues
- Python based and open source

## Challenges to our approach:

- Sustainability/Stability
- ColumnDataSource is non-intuitive
  - Interactivity requires all data to be in a single CDS
- Performance
  - Platform can only display 5% of the data at a time
  - Server document currently runs on individual laptops; slow

# Toward the future...

[www.oceanproteinportal.org](http://www.oceanproteinportal.org) (Beta)



# Acknowledgements

**Joe Futrelle\*\***  
**Mak Saito\*\***

**ProteOMZ Expedition:**  
Tristan Horner  
Alyson Santoro  
Caleb Hsu  
Science party and crew  
of RV Falkor

**Ocean Protein Portal:**

Nick Symmonds\*\*  
David Gaylord\*\*  
Danie Kinkade  
Adam Shepherd  
Michael Chagnon  
Paul Duffy



**Data Generation:**

Matt McIlvin  
Dawn Moran  
Chris Dupont (Metagenomics)

**Contributors to our favorite tools!**

Bokeh	Jupyter
Holoviews	Conda
Datashader	Cartopy
Matplotlib	Biopython
Pandas	Numpy
SciPy	Binder

