Getting things done with jupyter Notebooks

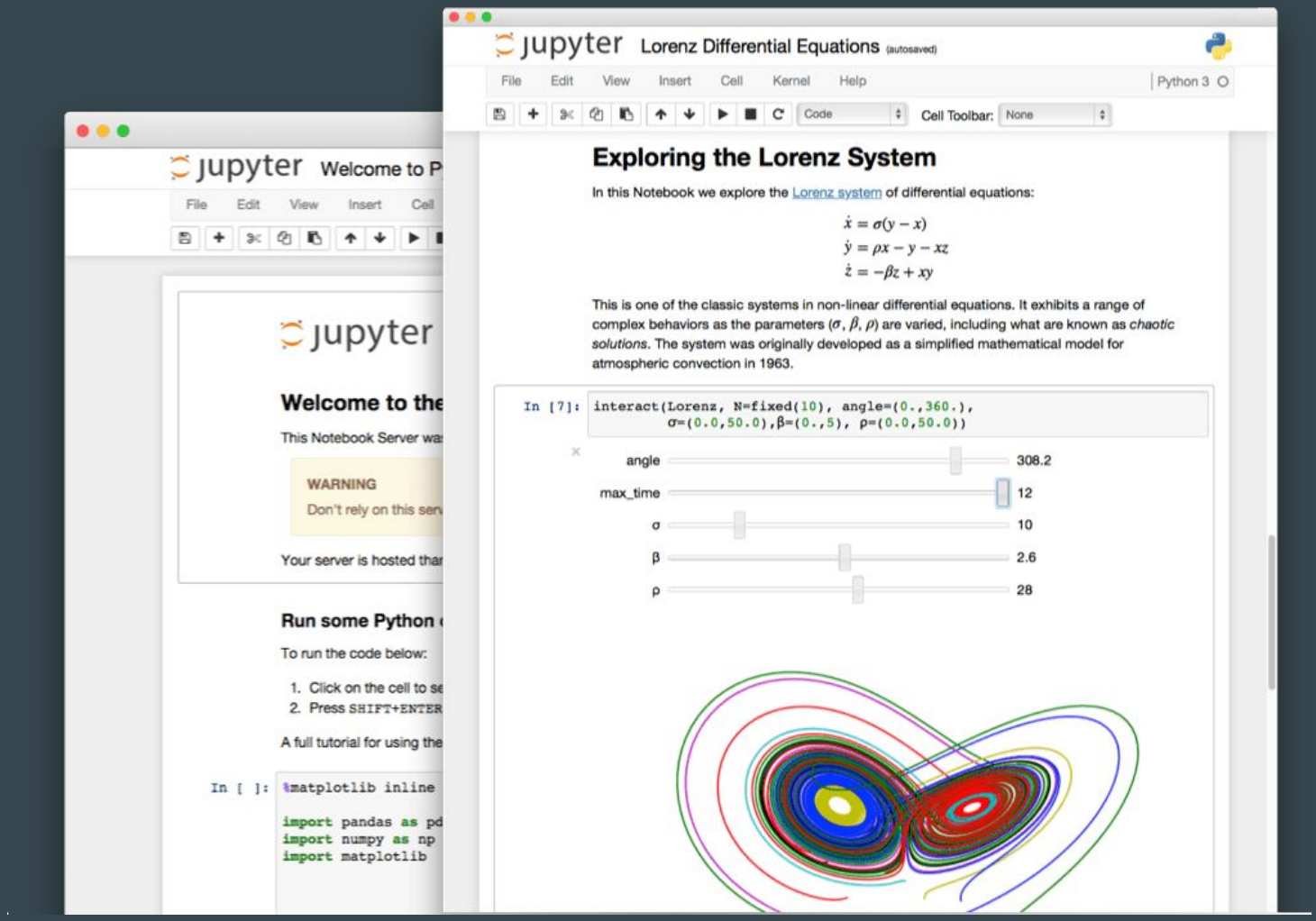
Chris Turner

Data Librarian at Axiom Data Science

Information Manager for NGA LTER

chris@axiomdatascience.com

Getting things done, with what?



A Jupyter Notebook is an IDE, kind of. But really it's just a file format that allows a mix of runnable code blocks, and documentation (markdown).

Getting things done, with what?

Project Jupyter

- Open source
- Non-profit
- Evolved from IPython Project
- Notebooks, and more!

How to get started for free*

Jupyter

- jupyter.org
- Binder

Big Names

- Google Colab
- Azure Notebooks
- Watson Studio Cloud

Research / Publishing

- CodeOcean
- Kogence
- Research Workspace

Data Science / ML

- Kyso
- ModeAnalytics
- Quantopian

Training / Education

- CoCalc
- Notebooks_
- Gryd
- DataQuest
- Kaggle

Reviews and Discussion

- on blogs
- medium
- dataschool.io

… or, host your own server

How do we use jupyter Notebooks?

Data management and processing
- ingest, processing scripts

Demonstration of capabilities and processes
- novel analyses and approaches

Compute near the data
- don’t need to transfer large files or collections

How do we use jupyter Notebooks?

Data Management and Processing Example:
Ingest, style, and map Audubon data

Project:

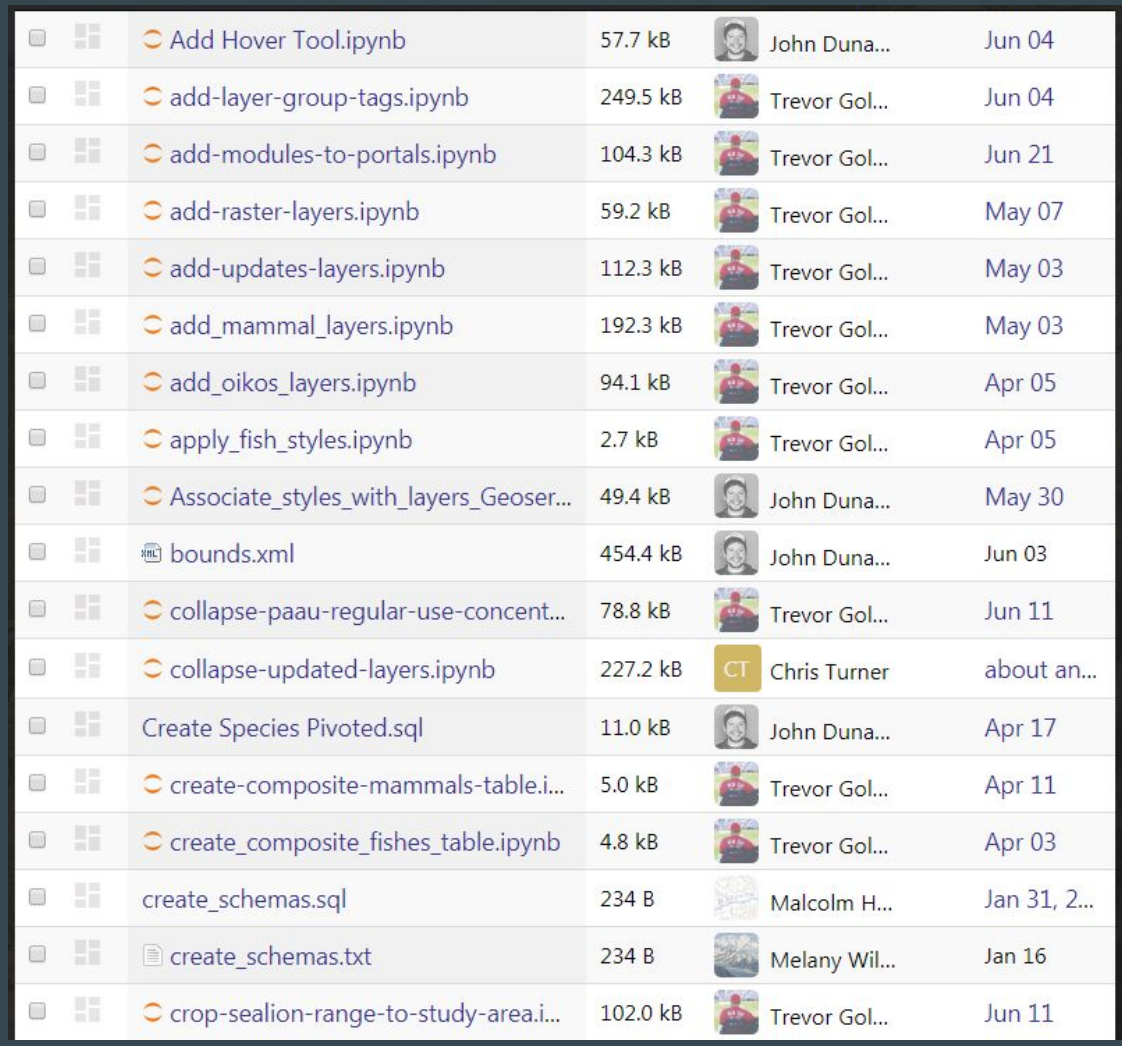
- 1000+ spatial datasets
- Re-style layer to match print version of the EABCBS
- Create interactive data portal

Why Notebooks:

- Unique data
- Transparent to client
- Shareable
- One-off work

The Ecological Atlas of the Bering, Chukchi, and Beaufort Seas

- Print (pdf)
- Web



Add Hover Tool.ipynb	57.7 kB	John Duna...	Jun 04
add-layer-group-tags.ipynb	249.5 kB	Trevor Gol...	Jun 04
add-modules-to-portals.ipynb	104.3 kB	Trevor Gol...	Jun 21
add-raster-layers.ipynb	59.2 kB	Trevor Gol...	May 07
add-updates-layers.ipynb	112.3 kB	Trevor Gol...	May 03
add_mammal_layers.ipynb	192.3 kB	Trevor Gol...	May 03
add_oikos_layers.ipynb	94.1 kB	Trevor Gol...	Apr 05
apply_fish_styles.ipynb	2.7 kB	Trevor Gol...	Apr 05
Associate_styles_with_layers_Geoser...	49.4 kB	John Duna...	May 30
bounds.xml	454.4 kB	John Duna...	Jun 03
collapse-paau-regular-use-concent...	78.8 kB	Trevor Gol...	Jun 11
collapse-updated-layers.ipynb	227.2 kB	Chris Turner	about an...
Create Species Pivoted.sql	11.0 kB	John Duna...	Apr 17
create-composite-mammals-table.i...	5.0 kB	Trevor Gol...	Apr 11
create_composite_fishes_table.ipynb	4.8 kB	Trevor Gol...	Apr 03
create_schemas.sql	234 B	Malcolm H...	Jan 31, 2...
create_schemas.txt	234 B	Melany Wil...	Jan 16
crop-sealion-range-to-study-area.i...	102.0 kB	Trevor Gol...	Jun 11

How do we use jupyter Notebooks?

Example Demonstrations:

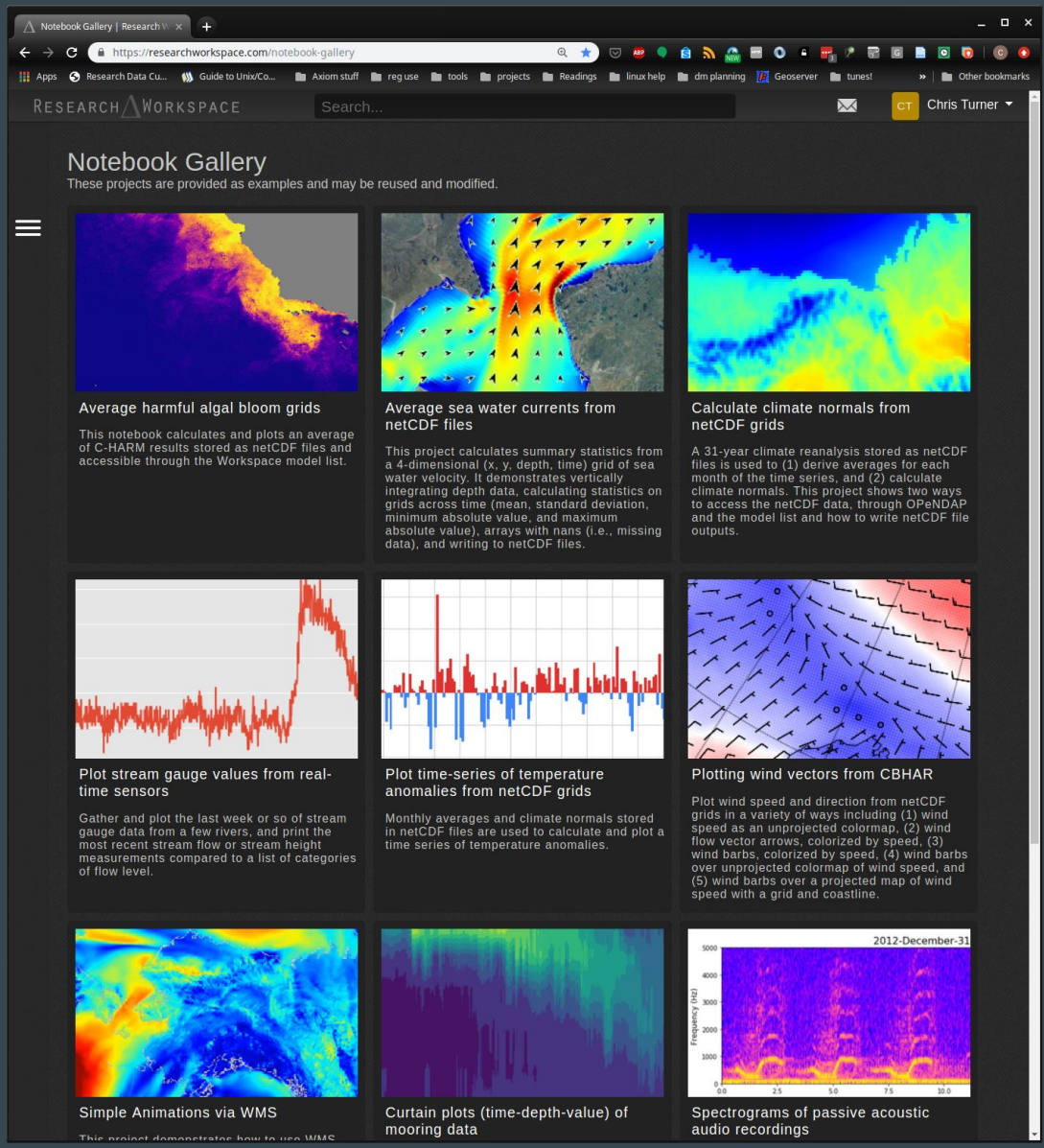
Research Workspace
Notebook Gallery

Project(s):

- Use Notebooks to process or analyze assets in the Axiom data system

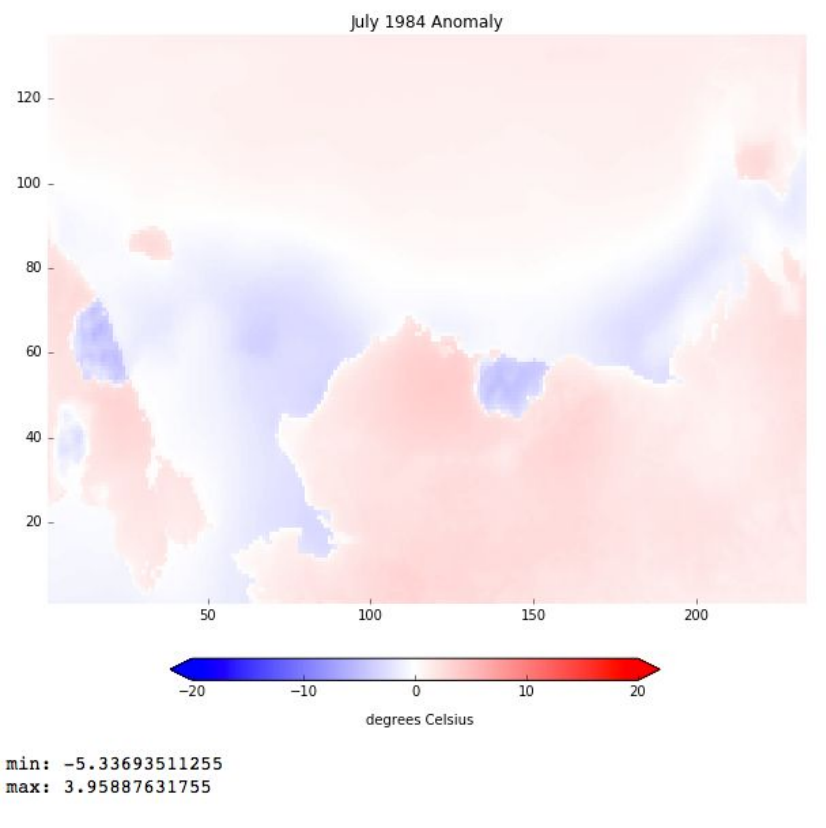
Why Notebooks:

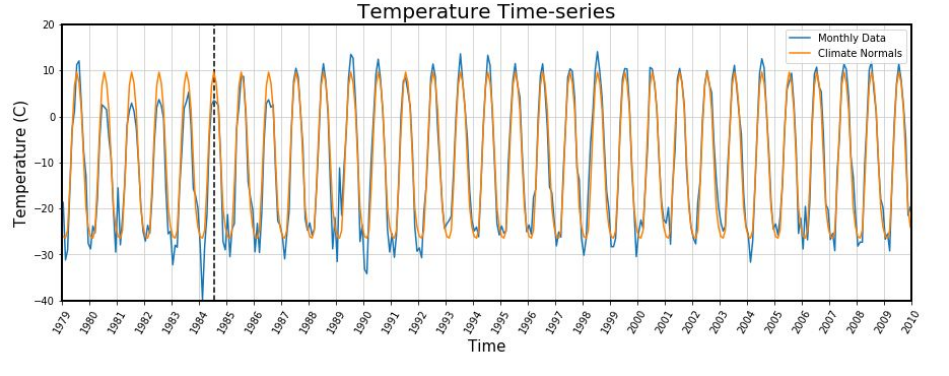
- Shareable
- Documentable
- Interactive

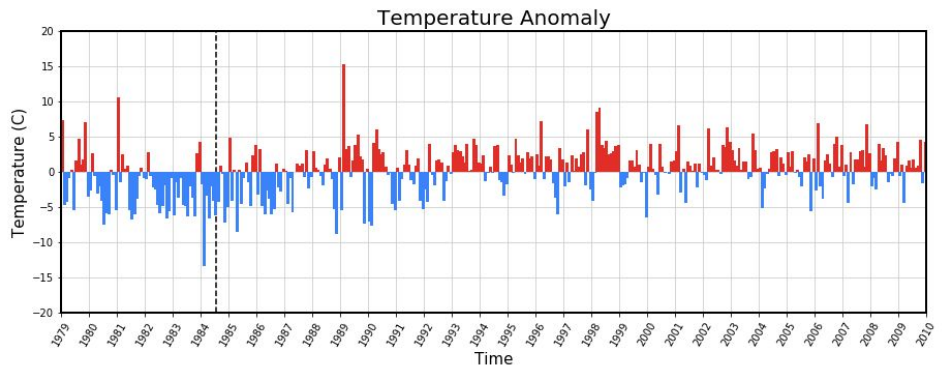


https://researchworkspace.com/notebook-gallery

Time-series Anomalies: CBHAR model







- Calculate climate normals on a 31-year long, multi-terabyte dataset
- Then plot temperature anomalies over a region

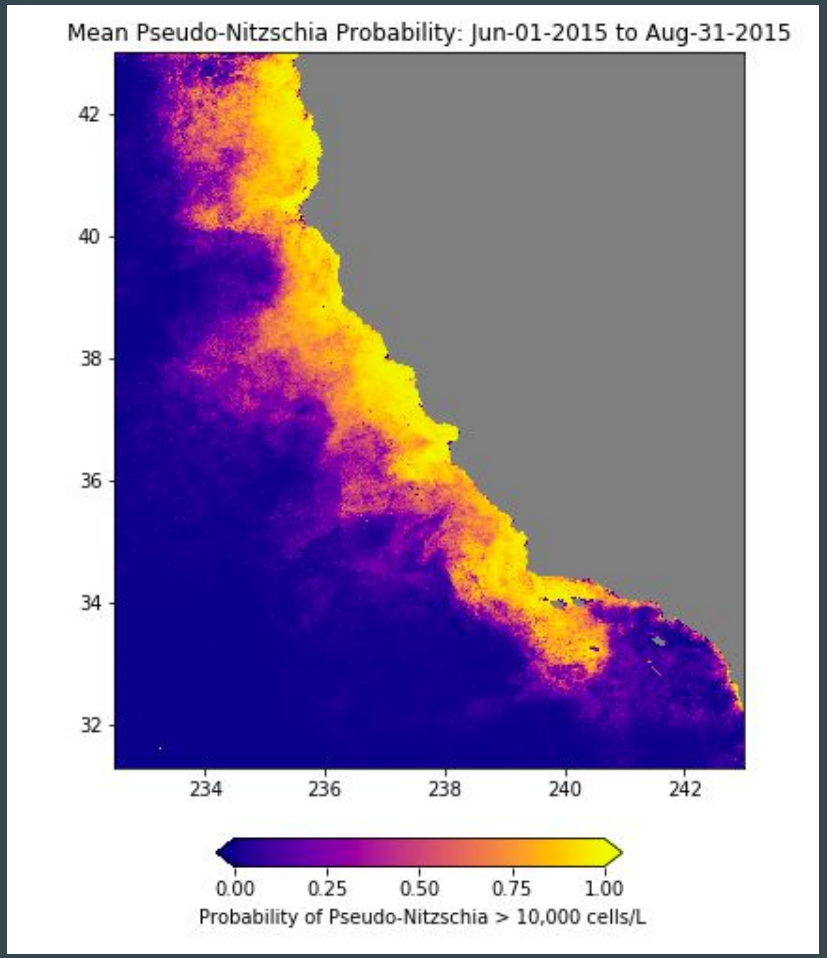
Averaging Many Gridded Files: C-HARM model

```
# We have many files in the directory, but we want to subset those by time. In this example,
# we'll just subset by a start and end date.
start_day = datetime.datetime(2015,6,1,0,0)
end_day = datetime.datetime(2015,8,31,0,0)
```

```
# These lines build a datacube of time slices between our start and end dates,
# expanding the dimensions as it goes.

first = True

for i in trange(nfiles):

    netcdf = netCDF4.Dataset(filenames[i])
    # extract the time, turn it into a date
    t = np.array(netcdf.variables['time'])[0]
    t = netCDF4.num2date(t, time_var.units, time_var.calendar)

    # compare the date of the time slice to our set start and end dates
    if start_day <= t <= end_day:
        # get the data from the netcdf file, remove the first axis (time)
        thisdata = np.array(netcdf.variables[variable_name])[0,:,:]

        if first:
            # If this is the first filename, create an array
            datacube = np.expand_dims(thisdata, axis=0)
            first = False
        else:
            thisdata=np.expand_dims(thisdata, axis=0)
            # If this is not the first filename, add to the existing array
            datacube= np.append(datacube, thisdata, axis=0)
    # close each file that we open
    netcdf.close()
```

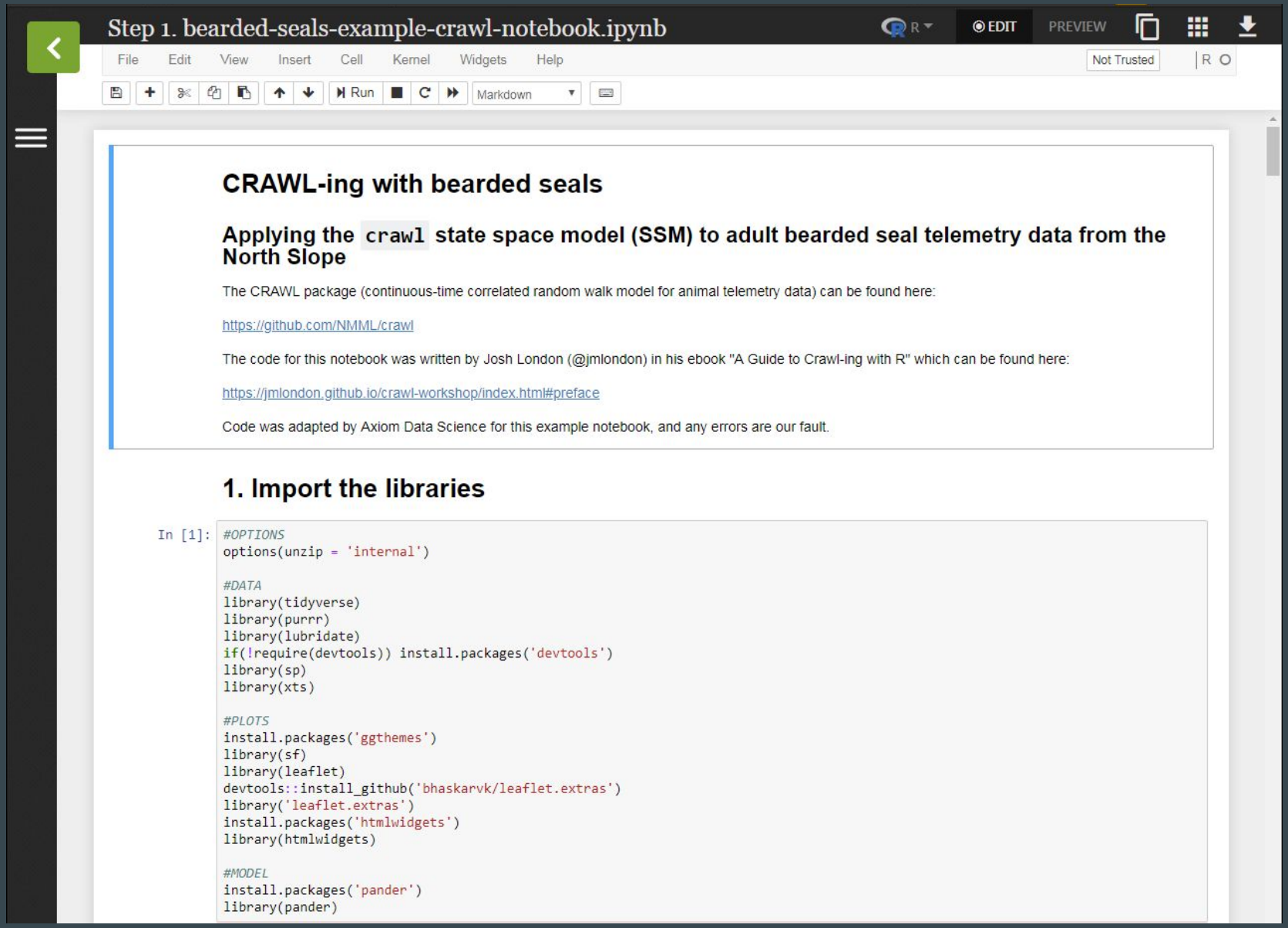
```
100%|██████████| 1222/1222 [00:24<00:00, 50.09it/s]
```

```
# Apply the mask to the datacube
datacube = ma.masked_values(datacube, -999.)
```



- Averaging gridded files between a start and end date

Test Implementation: CRAWL State Space Model



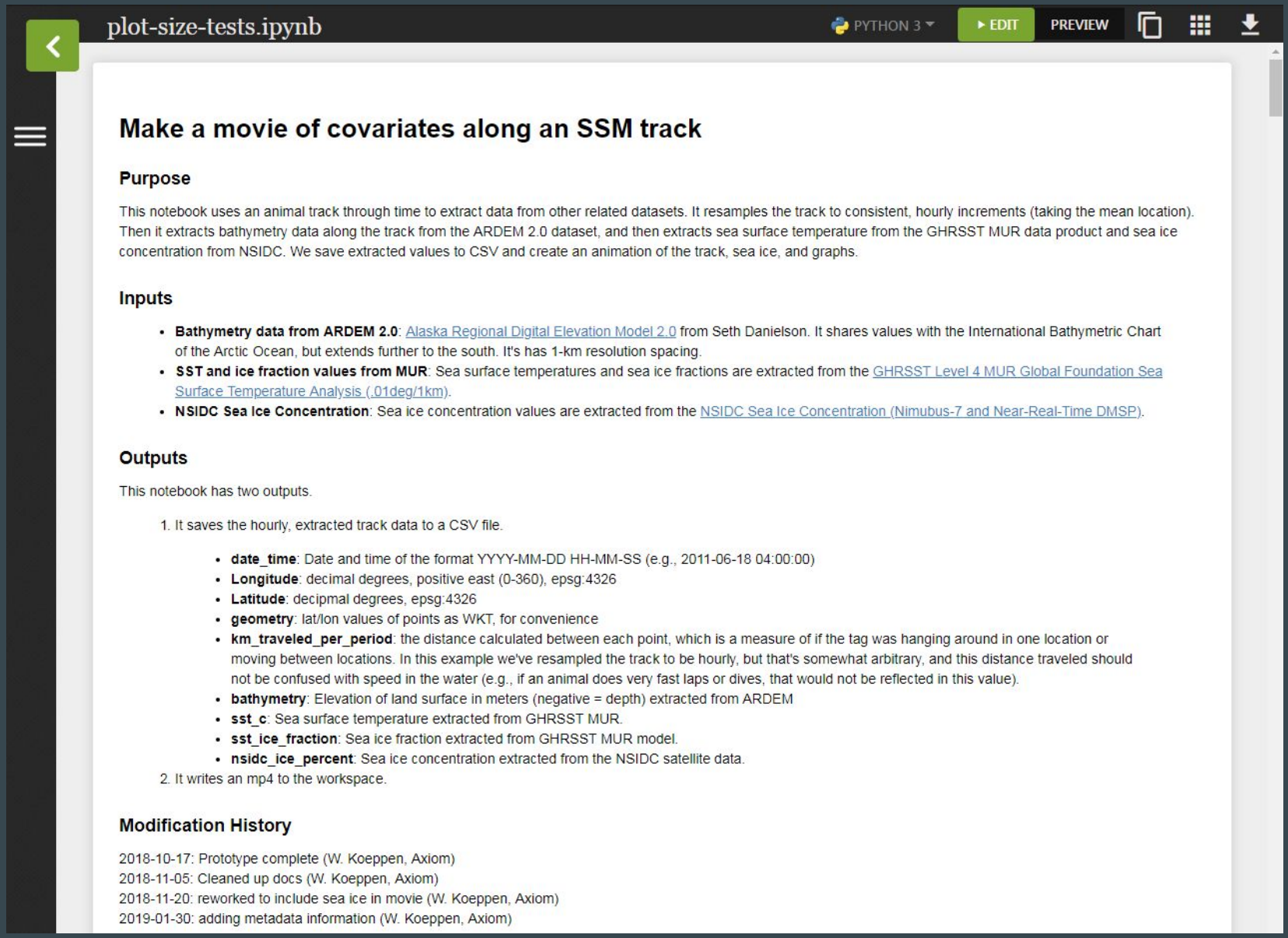
https://jmlondon.github.io/crawl-workshop/crawl-practical.html

Test Implementation: CRAWL State Space Model

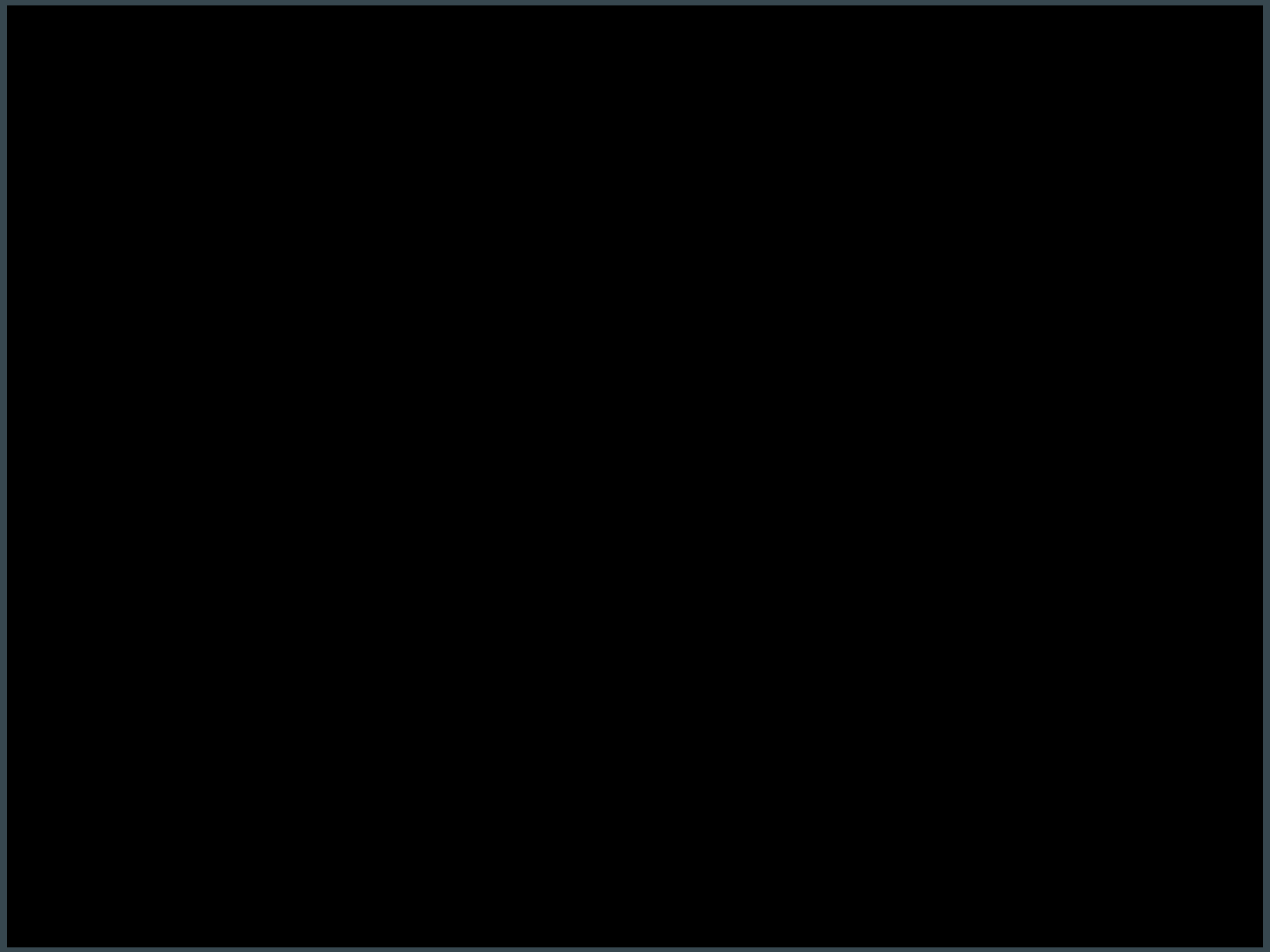
In [33]:

```
n <- length(unique(sf_pred_lines$deployid))
pal <- colorFactor(topo.colors(5),
                   domain = sf_pred_lines$deployid)

pal2 <- colorFactor(ggthemes::hc_pal(palette = "default")(n),
                   domain = sf_pred_lines$deployid)

 #sf::st_transform(4326) %>%
m <- leaflet() %>%
  addProviderTiles("Esri.OceanBasemap") %>%
#  addCircleMarkers(data = sf_locs, radius = 2,
#                   weight = 2, opacity = 1,
#                   color = ~pal(deployid)) %>%
  addPolylines(data = sf_lines, weight = 2, color = ~pal(deployid)) %>%
  addPolylines(weight = 2, data = sf::st_transform(sf_pred_lines,4326), color = ~pal2(deployid)) %>%
#  addLegend(pal = pal,values = ~deployid,labels = ~deployid) %>%
  suspendScroll()

m
```

Animation: CRAWL + models



Animation: CRAWL + models



Strengths of jupyter Notebooks

- They're easy, interactive, and (can be) very readable

- Shareable
  - https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks
  - https://www.researchworkspace.com/notebook-gallery
  - https://plot.ly/ipython-notebooks/
  - https://unidata.github.io/python-gallery/examples/index.html
  - https://proba-v-mep.esa.int/documentation/manuals/notebook-sample-gallery

- Language and kernel options:

  Python, R, Julia, and 130+ other kernels

Downsides of jupyter Notebooks?

- Easy to make mistakes
  - All cells must be run in order
  - No linting, syntax highlighting, etc.

- Encourage bad habits
  - Difficult to version, merge, and test code
  - Environment can be mysterious

Downsides of jupyter Notebooks?

- Easy to make mistakes
  - All cells must be run in order
  - No linting, syntax highlighting, etc.

- Encourage bad habits
  - Difficult to version, merge, and test code
  - Environment can be mysterious

Criticisms:

- Why Jupyter Is Not My Ideal Notebook
- I Don't Like Notebooks
- 5 Reasons Why Jupyter Notebooks Suck

Best Practices:

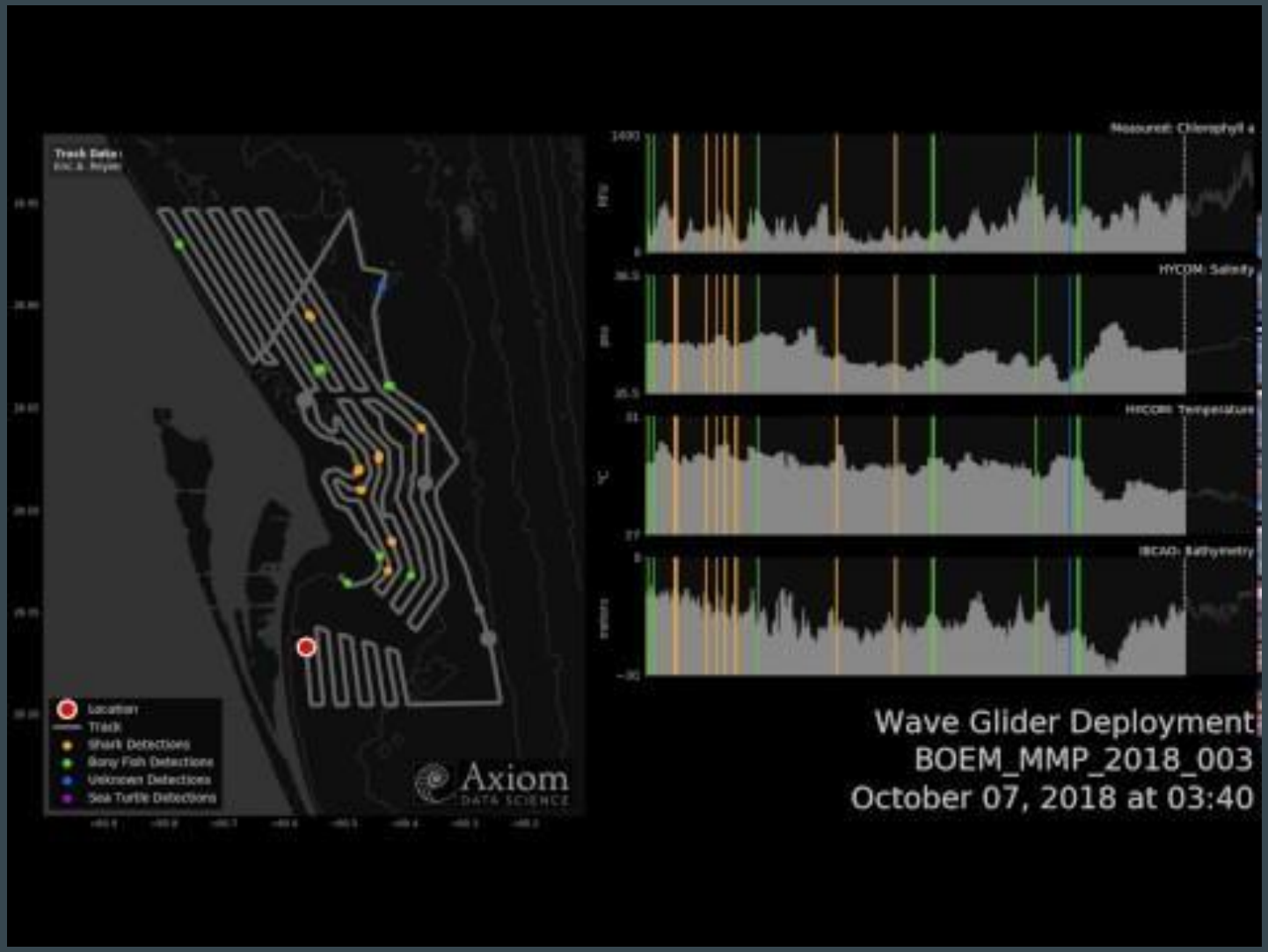
- Jupyter Notebook Manifesto
- How to Version Control Jupyter Notebooks
- Making Publication Ready Jupyter Notebooks

Done.

Questions?

chris@axiomdatascience.com

More Animations

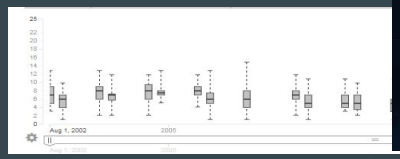


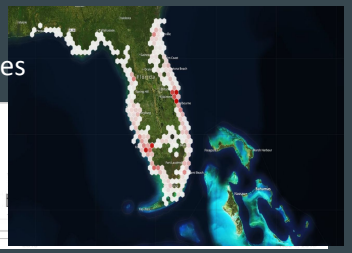
Another example: https://twitter.com/secoora/status/1040379622486147078

Data Types in the Axiom Data System

Biodiversity

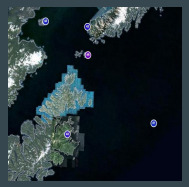
count, richness, diversity indices

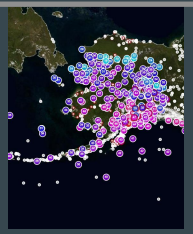




Platforms

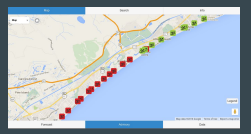
moorings, shore stations

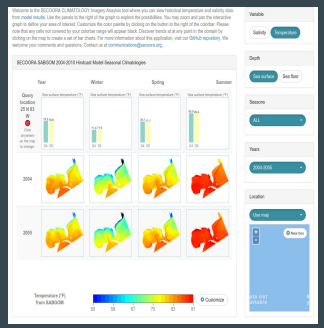


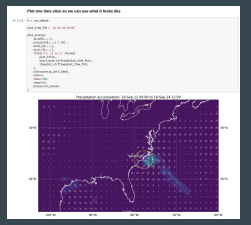


Products

skill assessment, shoreline change, etc.

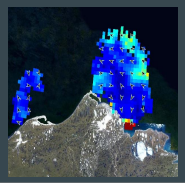


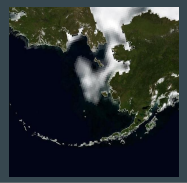


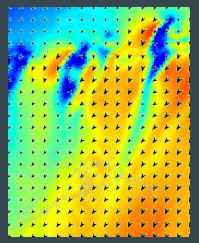


Grids

models, satellite, radar

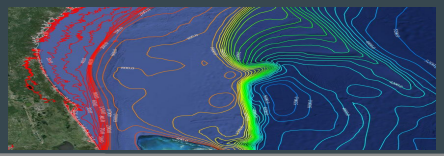






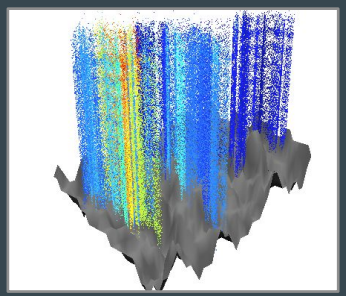
GIS

Habitat types, bathymetry, fishing zones, etc.



Moving Platforms

Gliders, Cruises



Unstructured Data

The Research Workspace

- Organize into projects, research campaigns and organizations
- Coordinate data exchange across networks, groups, programs
- ISO 19115-2/19110 metadata editor
- Execute server side Jupyter Notebooks on uploaded data AND any data in Axiom Data System
- Mint DOIs
- Archive pathway to DataONE (NCEI coming soon)