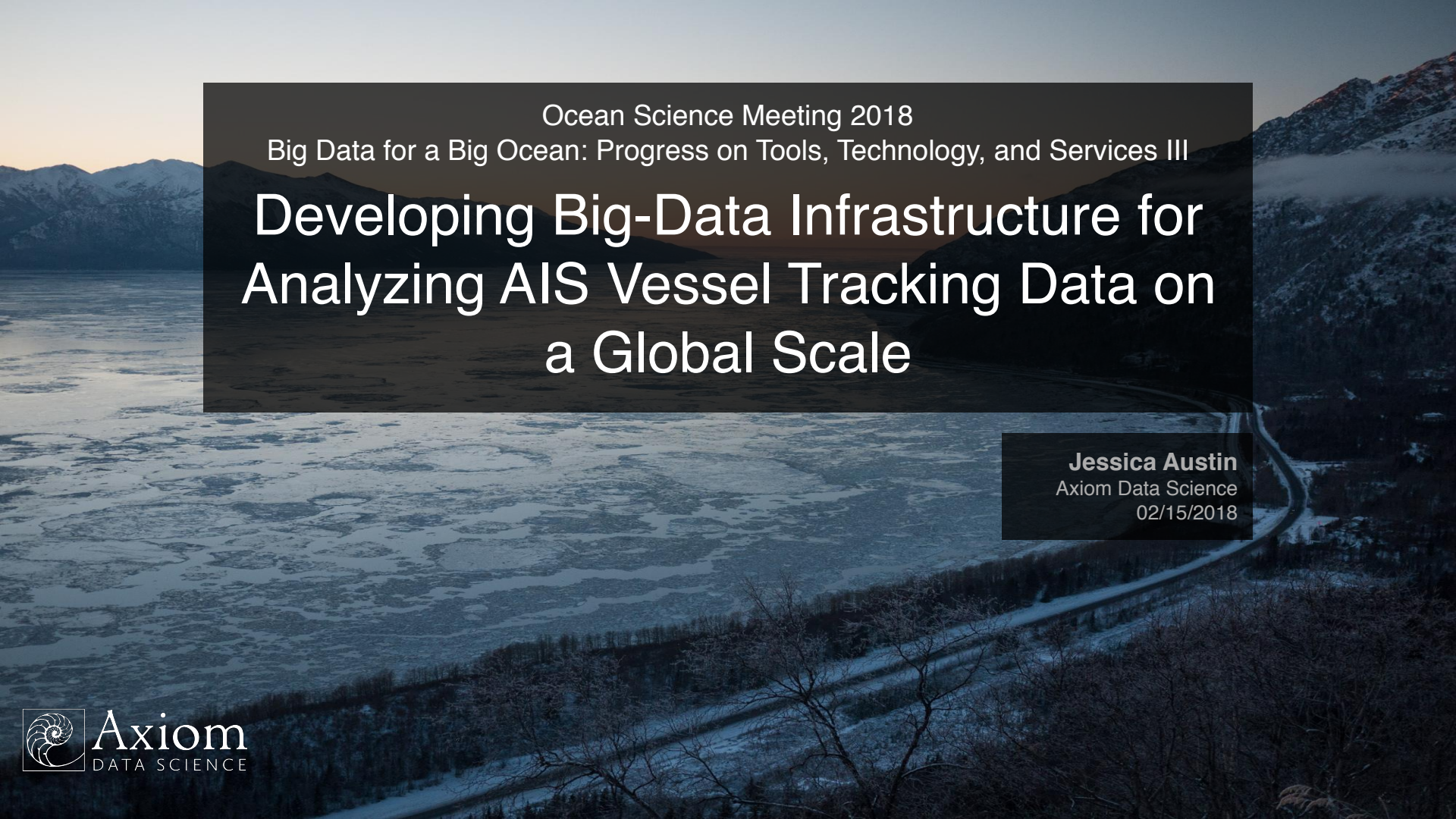


Ocean Science Meeting 2018
Big Data for a Big Ocean: Progress on Tools, Technology, and Services III

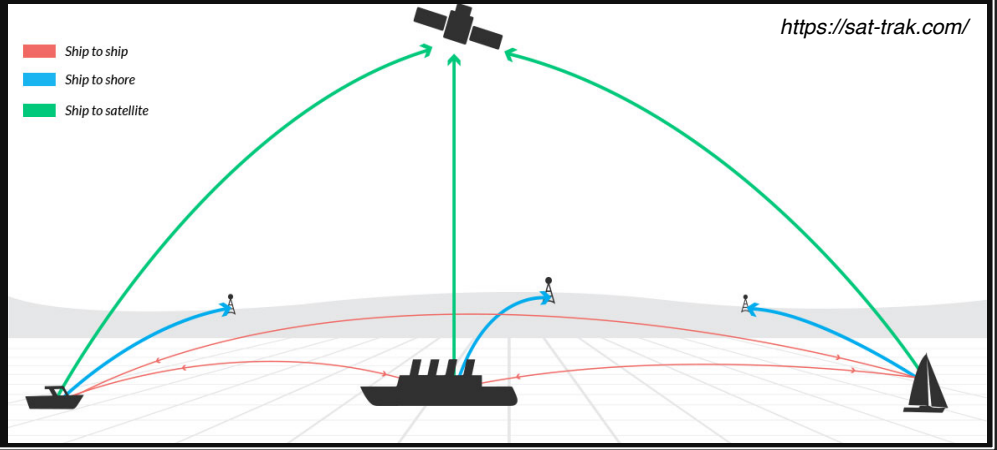
Developing Big-Data Infrastructure for Analyzing AIS Vessel Tracking Data on a Global Scale

Jessica Austin
Axiom Data Science
02/15/2018

Axiom DATA SCIENCE

The Automatic Identification System (AIS)

- On-board tracking system, transmitting over VHF
  - AIS equipment integrates with vessel navigation systems, such as GPS
  - Other AIS-equipped ships can see traffic in the area
  - Land-based (and satellite) receiver networks can collect data for an entire region
- Increasing interest in analyzing historic datasets
  - Prioritizing hydrographic surveys
  - Predicting probability and impact of oil spills
  - Quantifying interaction with marine wildlife
  - etc



Project Overview

Problem: Due to immense size of these datasets—typically 10s of billions of raw messages per year—and limitations on infrastructure and computing power, AIS data must currently be processed in small temporal or spatial subsets.

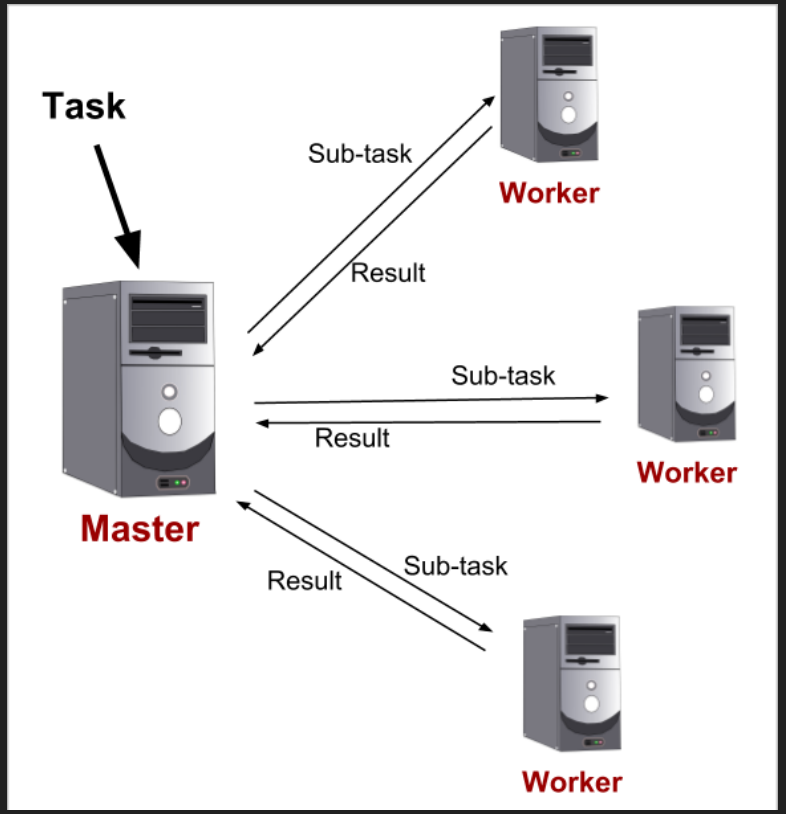
Solution: Develop cluster computing infrastructure and applications to analyze the data across many machines in parallel

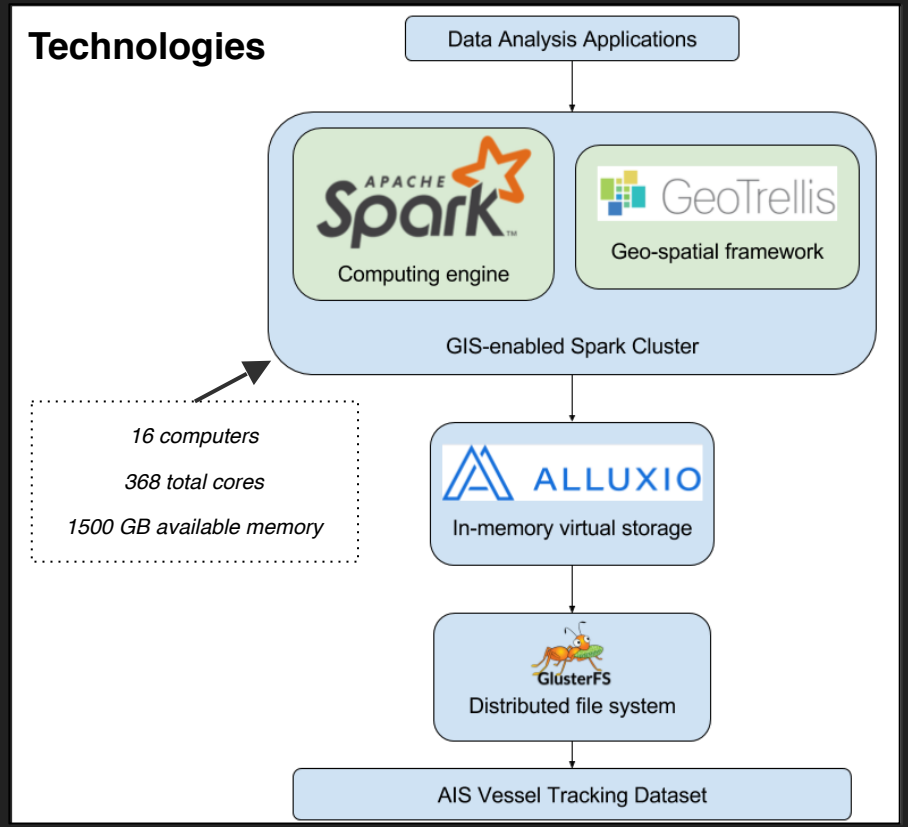
Results:

Previous workflow: single machine, small subset ⟶ days to weeks weeks to process

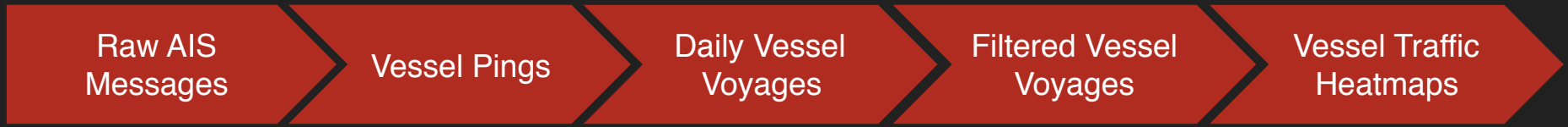
Our workflow: compute cluster, all US waters for one year ⟶ 48 hours

What is Cluster Computing?



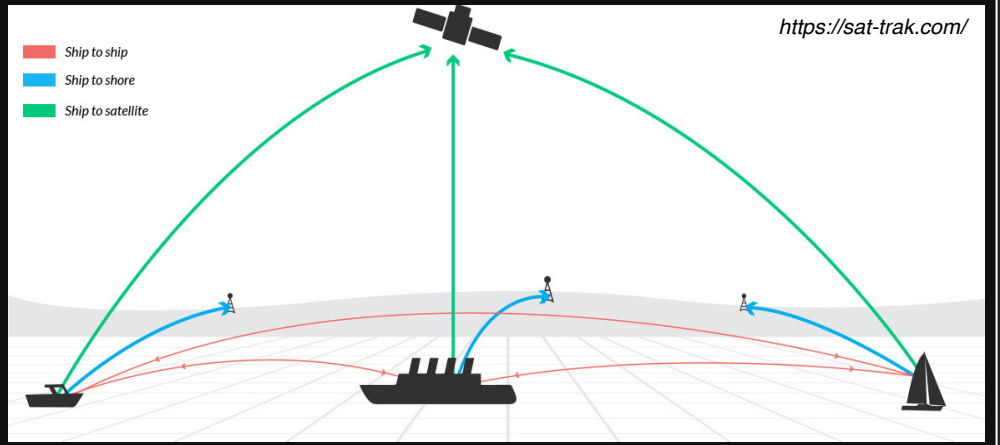


The Data Processing Pipeline



- Hydrographic Health Model establishes survey priorities across all US waters
  - Managed by NOAA Office of Coast Survey (OCS)
  - Includes metrics like: When was this area last surveyed? What is the seafloor complexity? **What is the typical vessel traffic in this area**? etc
- Vessel Traffic Heatmap requirements
  - We only care about ships that are moving and that are in US waters
  - Need to be split out by ship type, metric, and region (30 total files)
    - Tanker, Cargo, Passenger, and Other, and All ships
    - Total Vessel Count and Unique Vessel Count
    - Continental US, Alaska, Hawaii
  - 500m resolution, Albers Equal Area projection
- Current state-of-the-art (an ArcMap plugin) takes days to weeks to process small spatio-temporal subset of data

- AIS Messages are collected by land-based and satellite receivers
  - Agencies aggregate messages, and provide them as a database
  - Each message includes: timestamp, ship MMSI, latitude, longitude, speed over ground, etc
- Vessel Catalog
  - Maps MMSI to vessel attributes
  - Ship type, max draft, length, etc
- Possible issues
  - Duplicate data
  - Limited range for land-based
  - Self-reported data



Raw AIS Messages → Vessel Pings → Daily Vessel Voyages → Filtered Vessel Voyages → Vessel Traffic Heatmaps

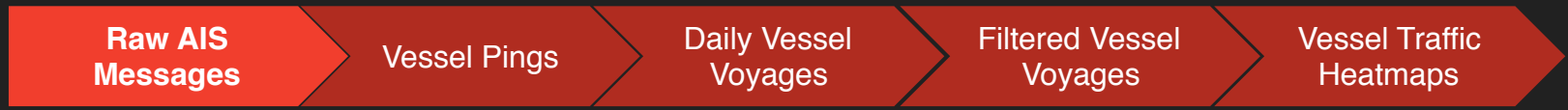
- AIS Messages are collected by land-based and satellite receivers
  - Agencies aggregate messages, and provide them as a database
  - Each message includes: timestamp, ship MMSI, latitude, longitude, speed over ground, etc
- Vessel Catalog
  - Maps MMSI to vessel attributes
  - Ship type, max draft, length, etc
- Possible issues
  - Duplicate data
  - Limited range for land-based
  - Self-reported data

Real-world Example: Hydrographic Health Model

2015 Terrestrial data from Coast Guard network
(via NOAA OCS)

74,212,891,806 messages
(7.5TB uncompressed)

Vessel Catalog: From US Coast Guard, 80,855 vessels

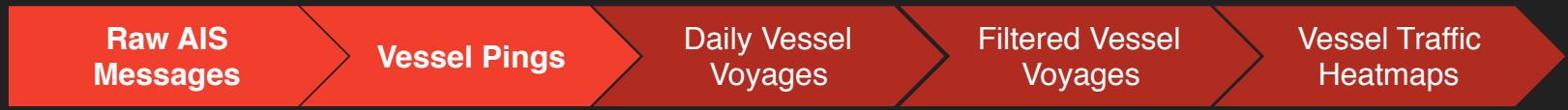


- Parse and clean the messages
  - Discard invalid messages
  - Discard any messages that are not Position Reports
  - Remove duplicate messages
- Result is called a *Vessel Ping*

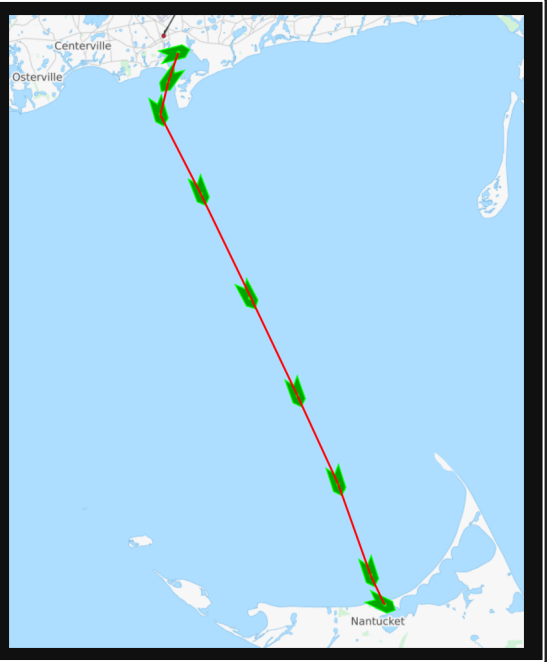
Real-world Example: Hydrographic Health Model

Input: 74,212,891,806 raw messages
Output: 2,017,535,550 pings
(2.72% of raw messages)

Total time: 40.7 hrs
(~6.7 min/day)



- *Vessel Voyages* describe the probable path of a ship
- Voyages are
  - At least two *pings* long
  - Broken up if the points are too sparse
  - Broken up if the ship stopped in one place for a long time



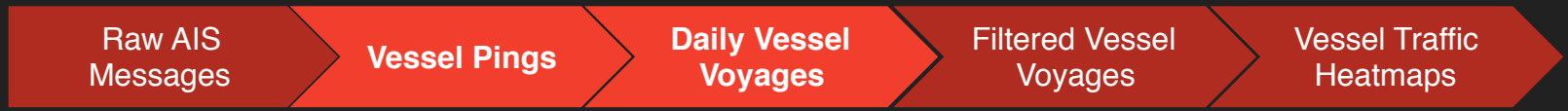
Raw AIS Messages → **Vessel Pings** → **Daily Vessel Voyages** → Filtered Vessel Voyages → Vessel Traffic Heatmaps

- *Vessel Voyages* describe the probable path of a ship
- Voyages are
  - At least two *pings* long
  - Broken up if the points are too sparse
  - Broken up if the ship stopped in one place

Real-world Example: Hydrographic Health Model

Input: 2,017,535,550 pings
Output: 12,297,024 voyages
Avg length: 51 pings
(~31% of pings used)

Total time: 3.2 hrs (~31 sec/day)

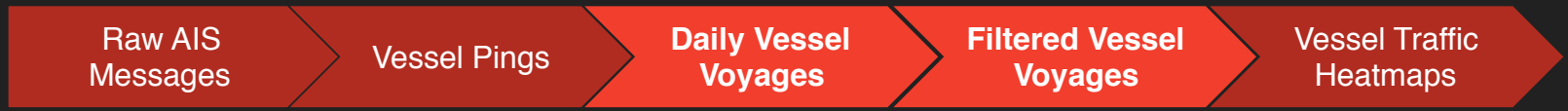


Pipeline: Filtered Vessel Voyages

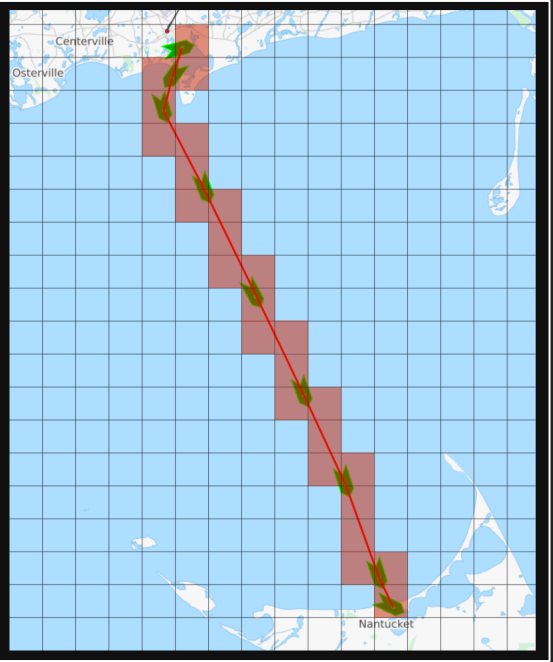
- Join to *Vessel Catalog* by ship identifier (MMSI)
- Segment data by
  - Region
  - Ship Type
  - Time Frame
  - etc

Real-world Example: Hydrographic Health Model

30 total files:
Regions: Continental US, Alaska, Hawaii
Ship Types: Tanker, Cargo, Passenger, and Other, and All ships
Metrics: Total Vessel Count and Unique Vessel Count



- Divide region into a grid, then count crossings
- Heatmap format is configurable
  - GeoTIFF or NetCDF
  - Arbitrary grid size; default is 500 meters
  - Choice of Albers Equal Area projection
- Possible metrics
  - Total traffic volume
  - Unique vessel count
  - Maximum vessel draft
  - etc...



Raw AIS Messages → Vessel Pings → Daily Vessel Voyages → **Filtered Vessel Voyages** → **Vessel Traffic Heatmaps**

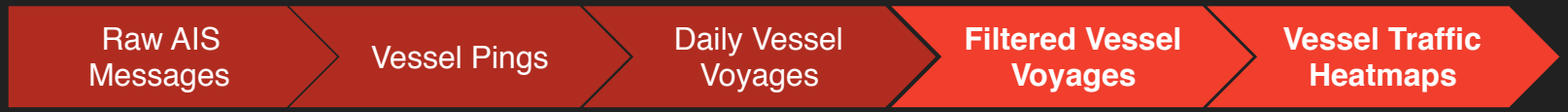
- Divide region into a grid, then count crossings
- Heatmap format is configurable
  - GeoTIFF or NetCDF
  - Arbitrary grid size; default is 500 meters
  - Choice of Albers Equal Area projection
- Possible metrics
  - Total traffic volume
  - Unique vessel count
  - Maximum vessel draft
  - etc...

Real-world Example: Hydrographic Health Model

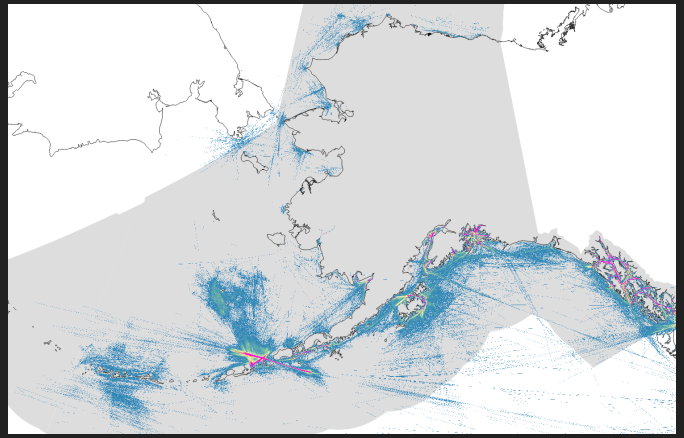
Input: 12,297,024 voyages
Output: 30 heatmap files

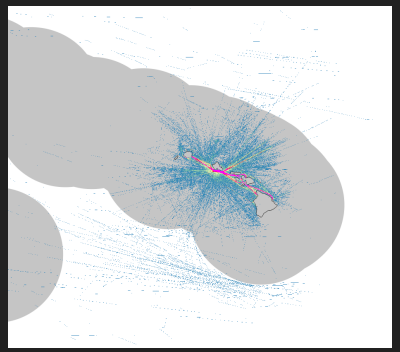
Total time: 90 minutes
(Includes filtering and creating heatmaps)

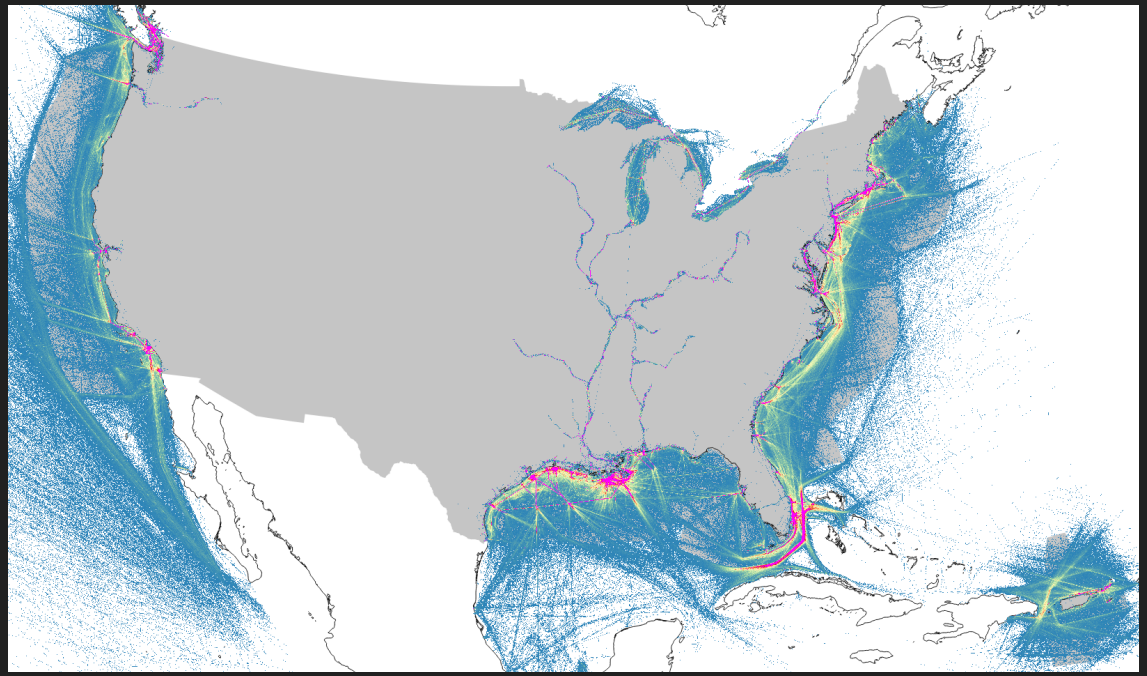
OVERALL TIME: 45.4 hrs



Real-world example: Hydrographic Health Model

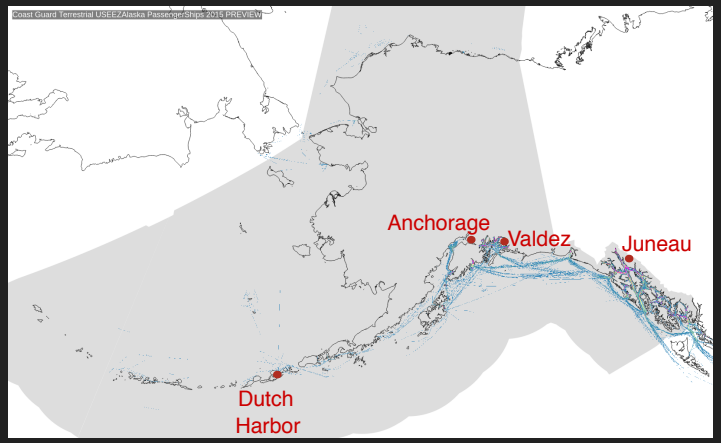




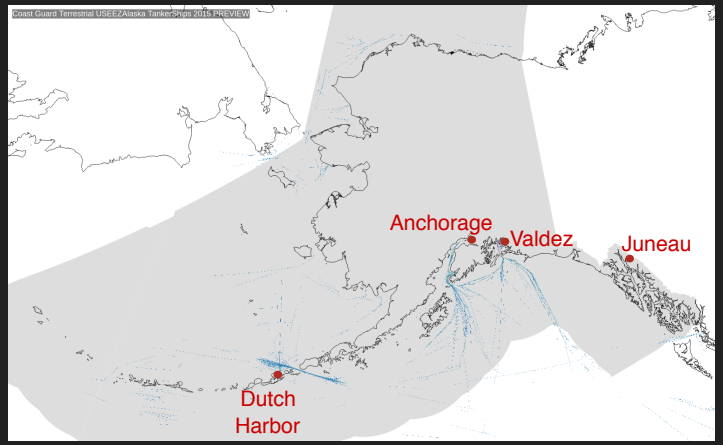


Real-world example: Hydrographic Health Model

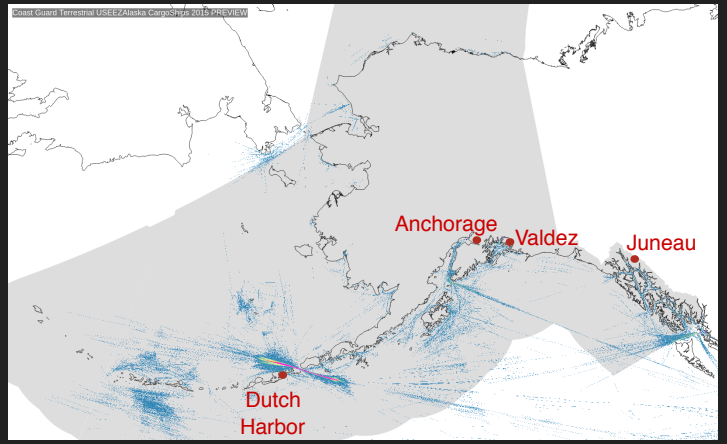
Passenger



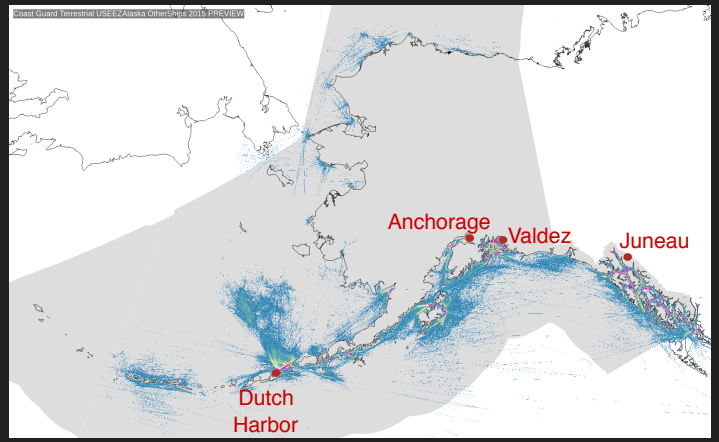
Tanker



Cargo



Other



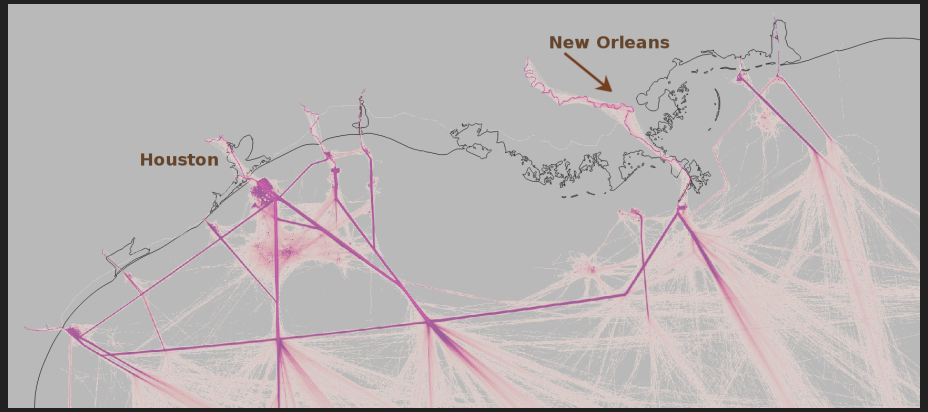
Axiom DATA SCIENCE

What's Next?

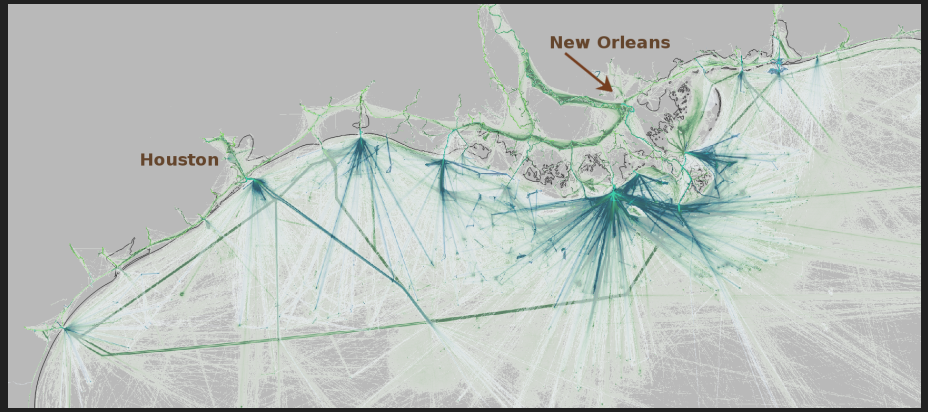
- Publicly available website with static downloads: http://ais.axds.co
- More datasets (currently in analysis):
  - Marine Exchange of Alaska, 2008-2017
  - Satellite AIS Data (global)
  - Danish Maritime Authority (Europe)
- Integration into IOOS and AOOS Data Portals
  - View alongside sea ice models, environmental sensors, wildlife habitat datasets, etc
  - Dataset added to IOOS Data Catalog, archived with NCEI
- Enable ad-hoc querying with GeoTrellis, Jupyter Notebooks

Real-world example: Hydrographic Health Model

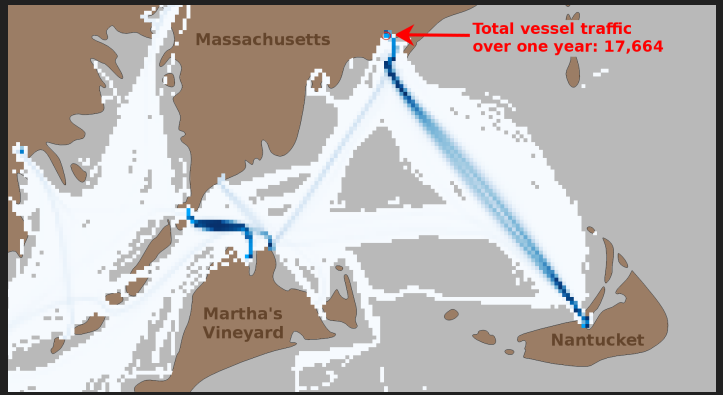
Tanker Ships

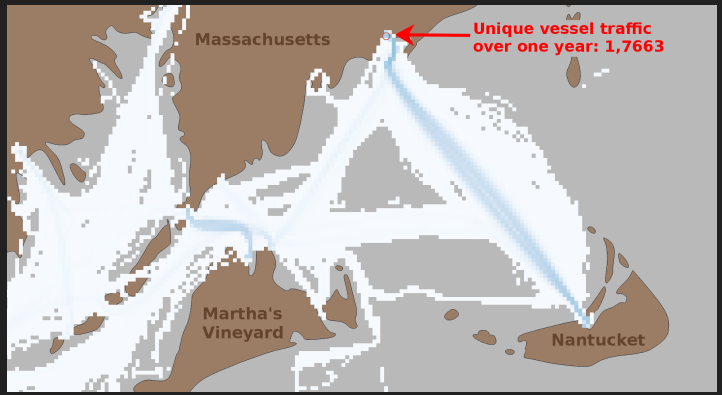


Passenger (blue) and Other (green)



Total vs Unique Traffic:





Axiom DATA SCIENCE