Scaling biodiversity data processing and analysis workflows with Apache Beam

Jeremy Malczyk
Software Engineer
Map of Life, Yale University

Walter Jetz (PI, Yale University)
Robert Guralnick (Co-I, University of Florida, Gainesville)
Adam Wilson (Co-I, State University of New York, Buffalo)
Map of Life software development team (https://mol.org/team)

Software Workflows and Tools for Integrating Remote Sensing and Organismal Occurrence Data Streams to Assess and Monitor Biodiversity Change

2016 ROSES A.41 Solicitation NNH16ZDA001N
AIST Research Opportunities in Space and Earth Sciences
Grant # AIST-16-0092

Walter Jetz (PI, Yale University)
Robert Guralnick (Co-I, University of Florida, Gainesville)
Adam Wilson (Co-I, State University of New York, Buffalo)

Yale

MOL
MAP OF LIFE

WHY?

“improving the ease with which the biology and ecology communities can understand, select and use appropriately NASA remote sensing data.”

Yale

MOL
MAP OF LIFE

- Types of biodiversity data
- Combining biodiversity data with remote sensing products
- Existing tools for data fusion
- Scaling with Beam

Yale

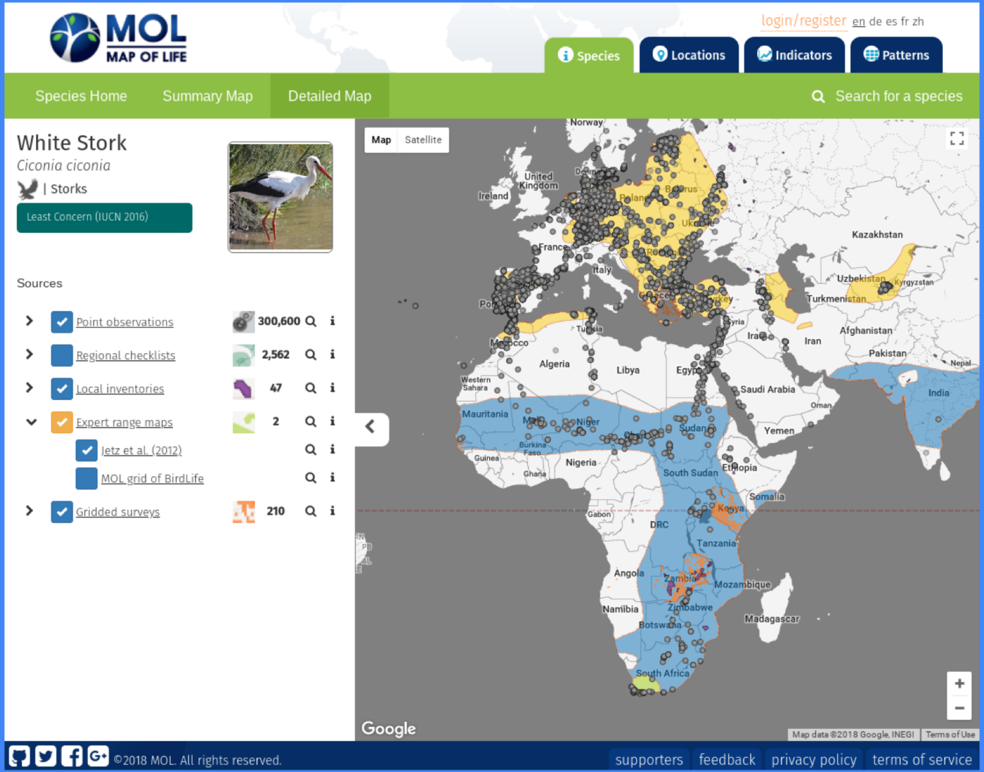
MOL MAP OF LIFE

Types of biodiversity data

Yale

MOL
MAP OF LIFE

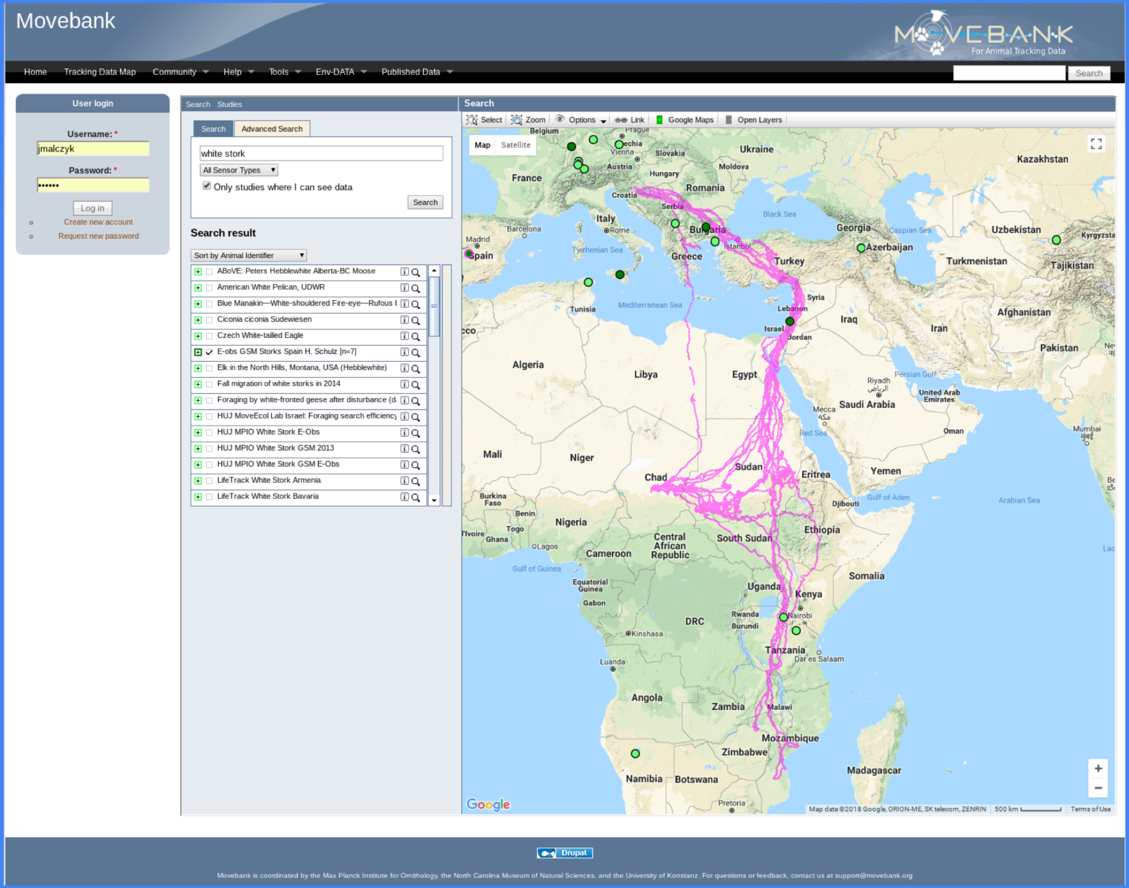
- Observation / occurrence
- Expert range maps
- Local Inventories
- Gridded surveys
- Regional checklists
- Distribution model predictions



Yale

MOL MAP OF LIFE

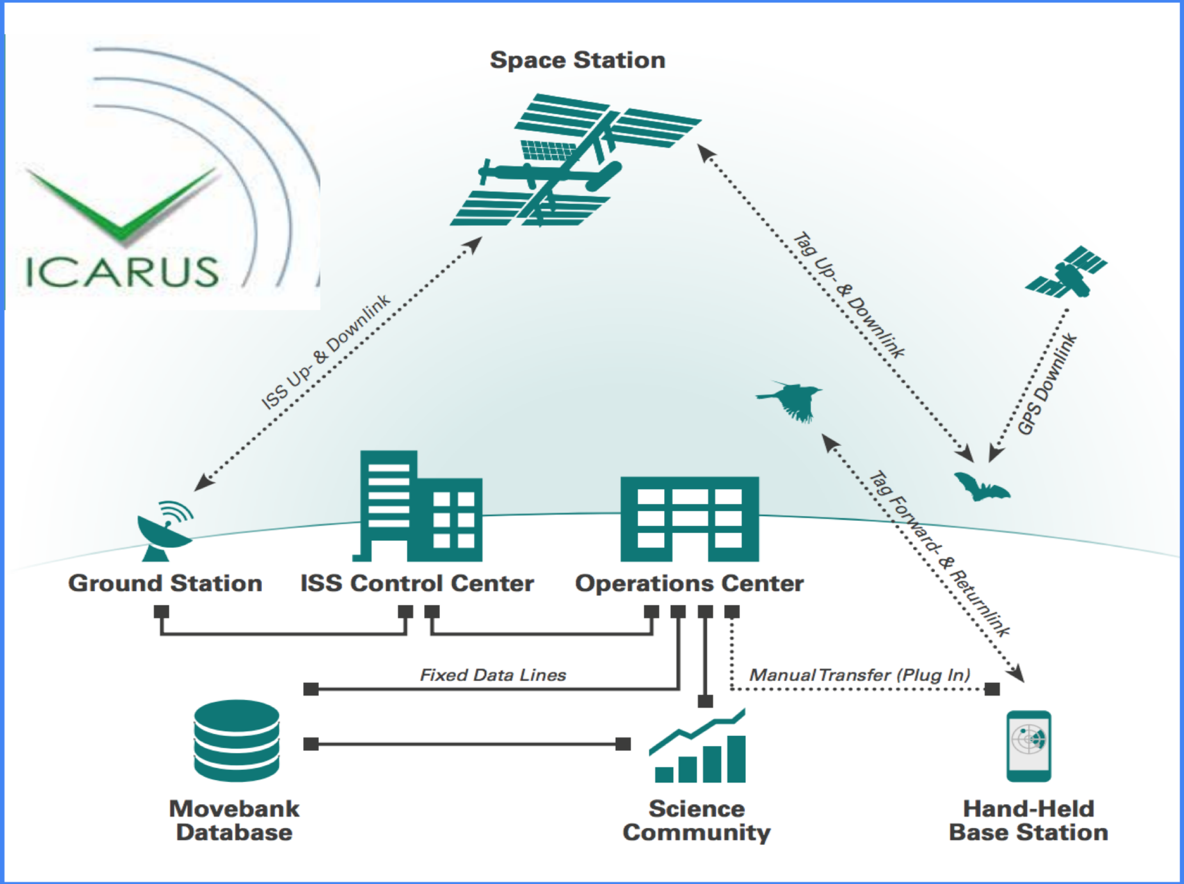
Movement data



Yale

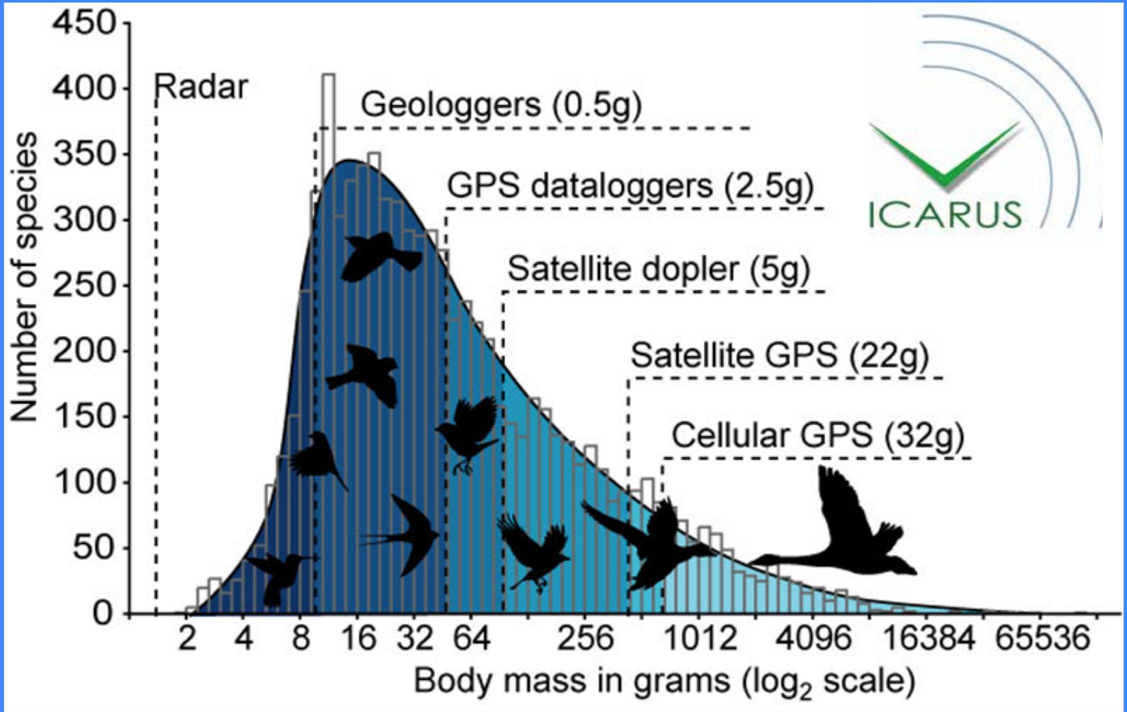
MOL MAP OF LIFE

... and more movement data.



Yale

MOL MAP OF LIFE



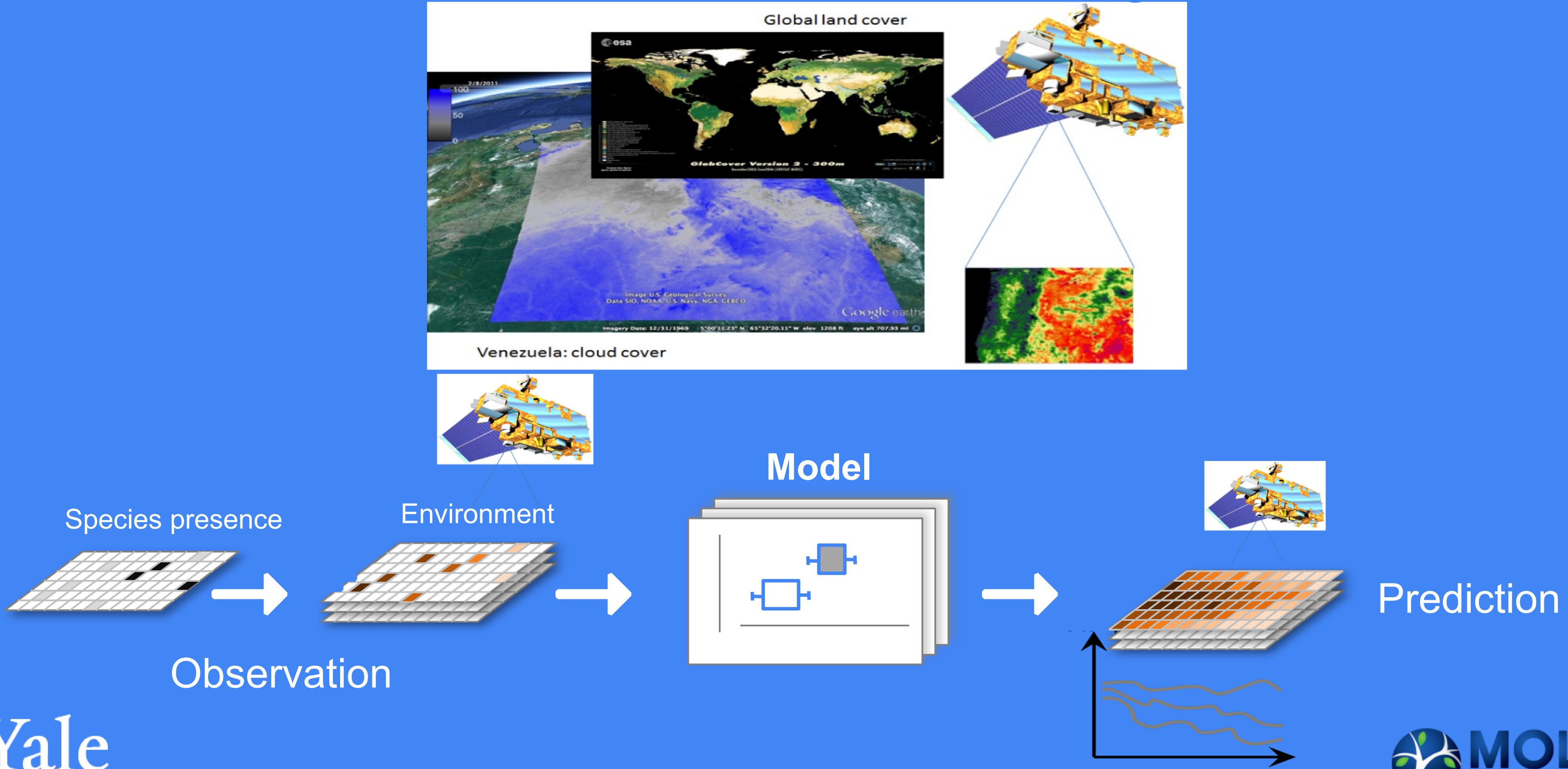
Yale

MOL
MAP OF LIFE

Combining biodiversity data with remote sensing products

Yale

MOL
MAP OF LIFE

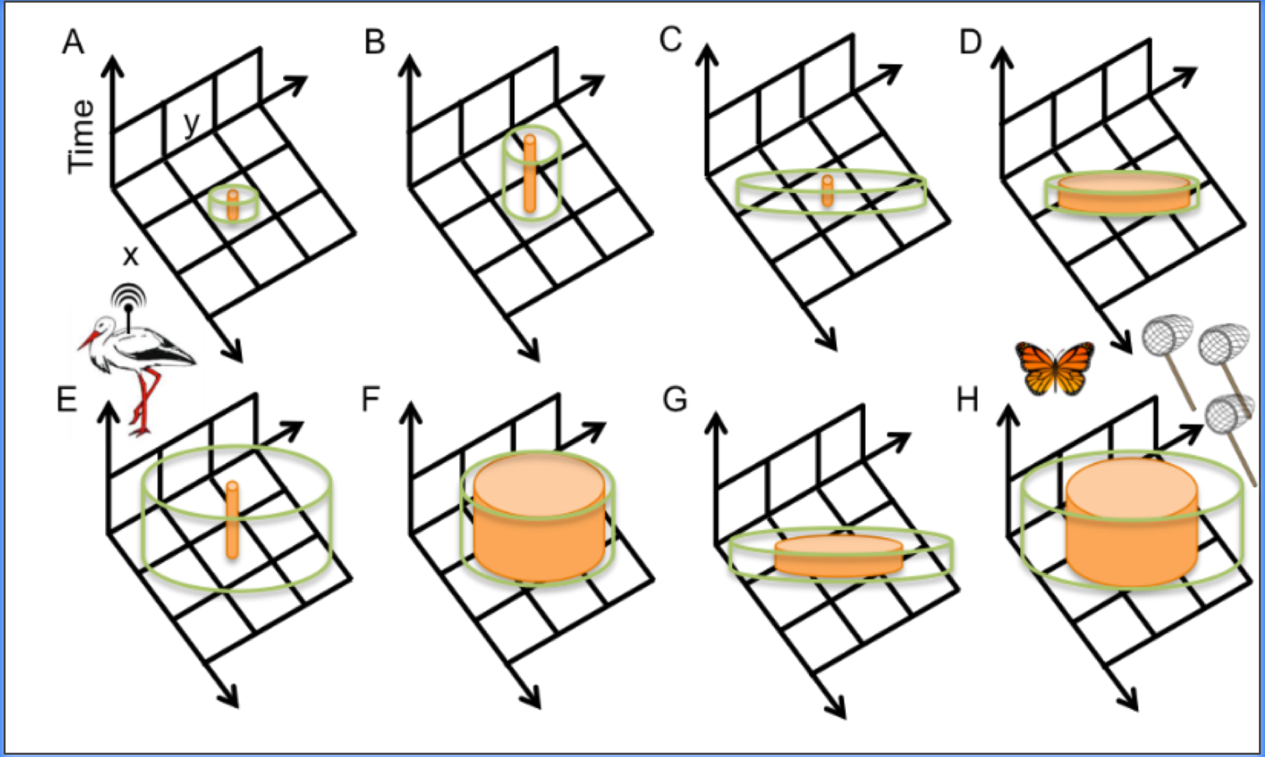


Yale

MOL MAP OF LIFE

Biodiversity data ∩ Environmental data

Not trivial …



Yale

MOL
MAP OF LIFE

Existing tools for data fusion

Yale

MOL
MAP OF LIFE

Desktop applications and libraries

Advantages

- Well documented
- Extensible
- Well integrated with other tools and systems
- Large community of developers and users
- What biology and ecology communities use and understand

Limits

- Requires everything (imagery and data) to be local
- Can’t scale beyond local resources

ENV-Data

Advantages

- Large catalog of public imagery
- Kept up to date
- Specifically designed for annotating movement data

Limits

- Slow (hours - days)
- Not extensible
- No support for raster upload
- No support for spatial or temporal aggregation

Google Earth Engine

Advantages

- Large and growing catalog of imagery
- Kept up to date
- Abstracts complexity away (scales compute, manages tasking)
- Full spatial analysis API - supports aggregation in space-time

Limits

- High vendor lock-in / low portability
- Limited interoperability with traditional tools
- Very limited server-side logging (difficult to debug)
- Quota limited per user
- Fixed node size
- Mediocre performance across large vector datasets.

Yale

MOL
MAP OF LIFE

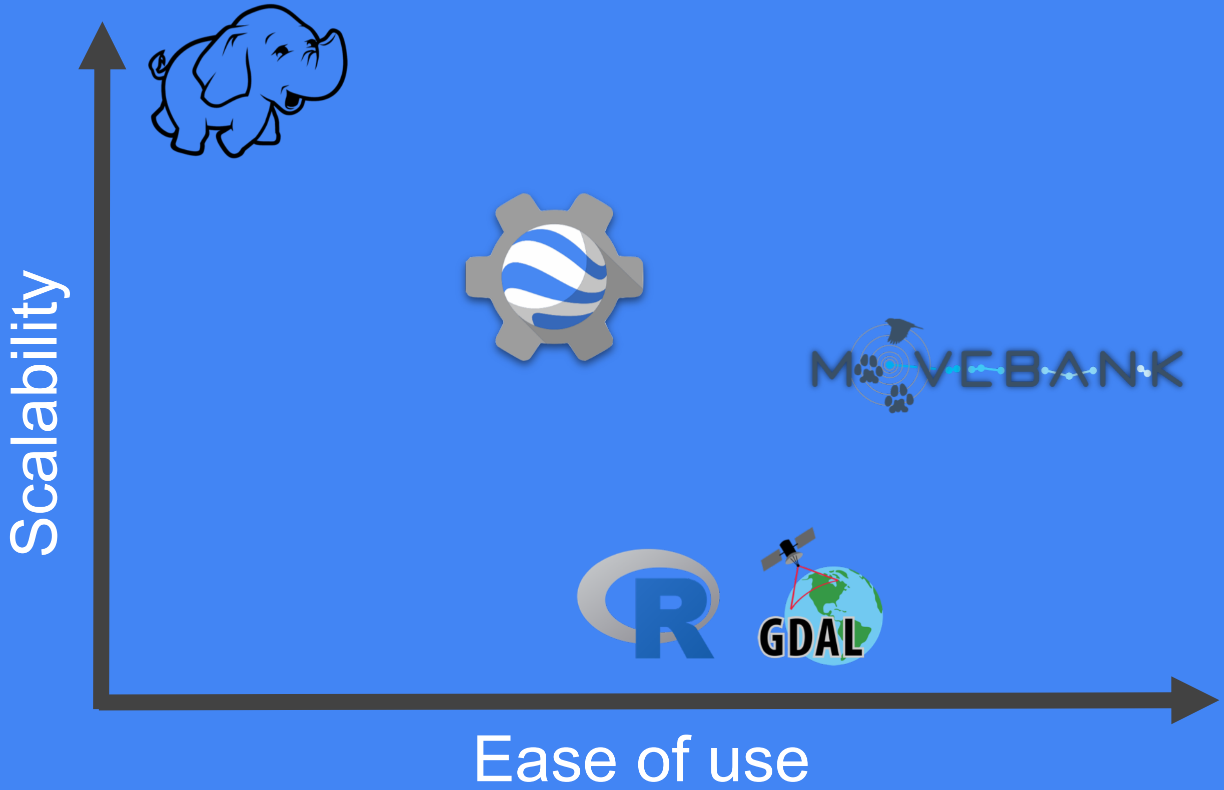
Big-data frameworks

Advantages

- Ultimate flexibility
- Ultimate scalability

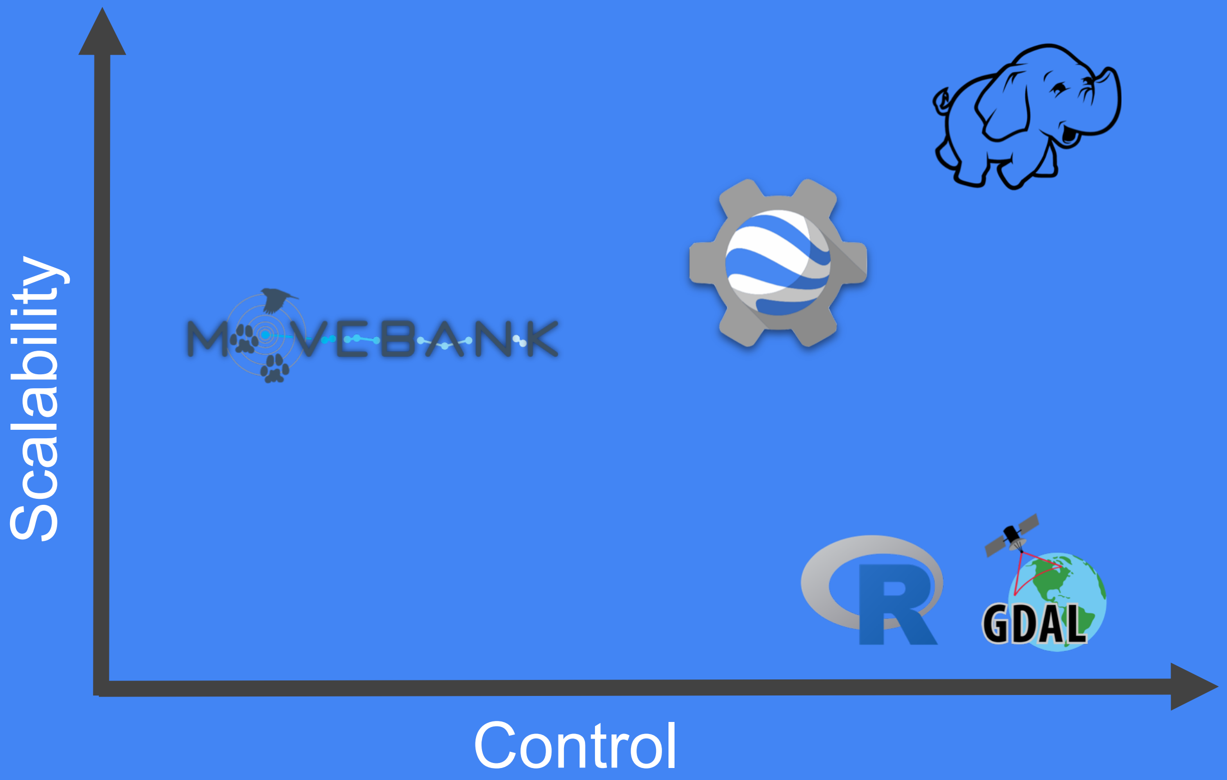
Limits

- Harder to use, less accessible to non-technical users
- Often limited documentation and support for spatial analysis



Yale

MOL
MAP OF LIFE



Yale

MOL
MAP OF LIFE

Bridging the divide with Apache Beam

Yale

MOL MAP OF LIFE

beam

An advanced unified programming model

Implement batch and streaming data processing jobs that run on <u>any execution engine</u>.

APEX

Spark

Yale

The Beam API

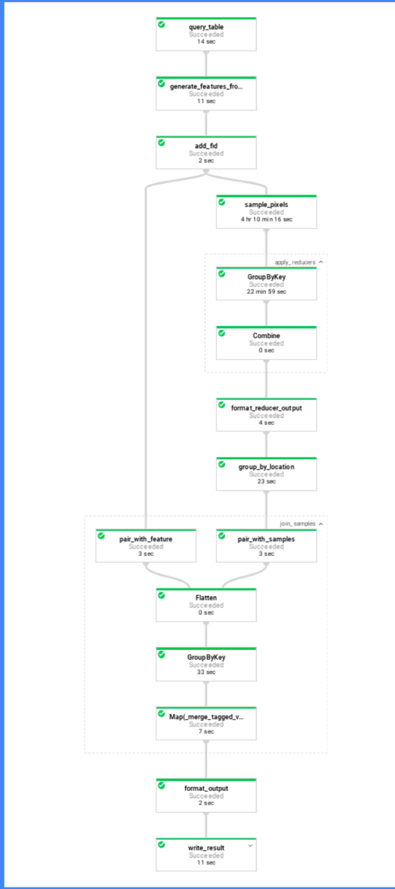
```
# sample images and apply spatiotemporal reducers
samples = (features
      | 'sample_pixels' >> beam.ParDo(sample_region, args, asset)
      | 'apply_reducers' >> beam.CombinePerKey(ReducePixels(args))
      | 'format_reducer_output' >> beam.ParDo(format_reducer_output)
      | 'group_by_location' >> beam.GroupByKey()
)
```

Yale

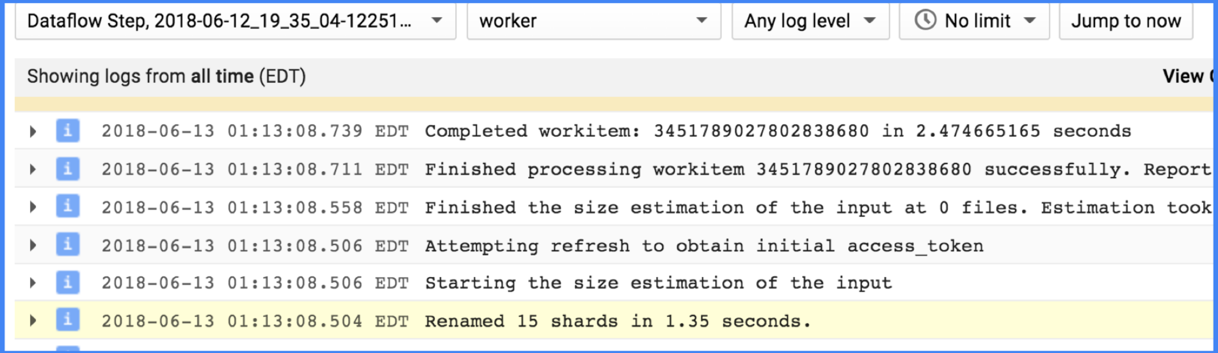
MOL
MAP OF LIFE

Cloud Dataflow features

Live pipeline monitoring



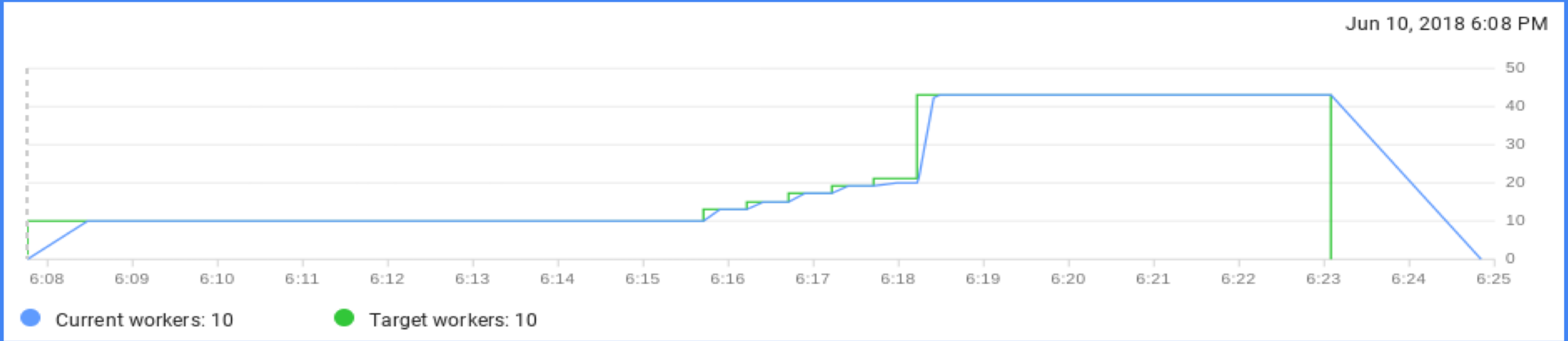
Detailed logging



Yale

MOL MAP OF LIFE

Autoscaling



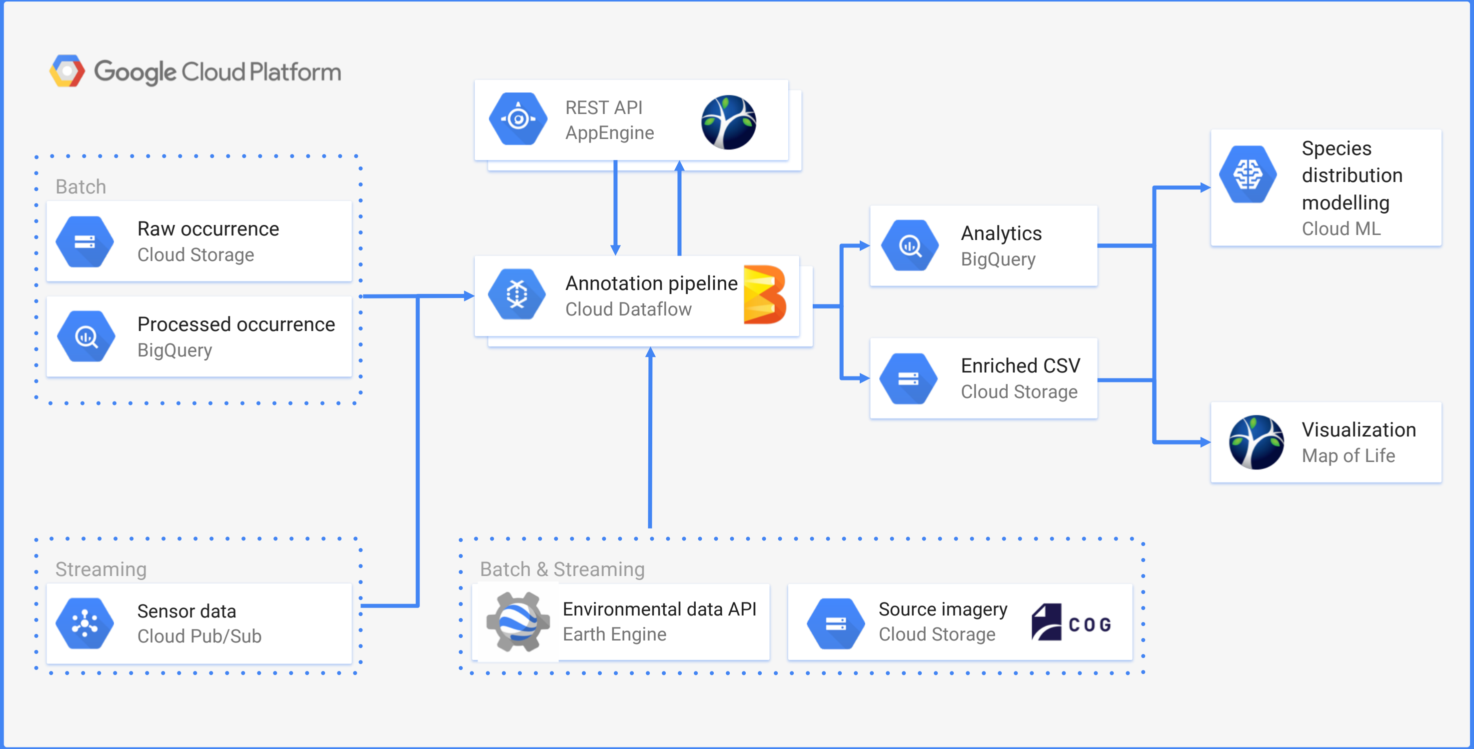
Yale

MOL MAP OF LIFE

Drawbacks

- Very new
- Need high availability access to pixel data (high QPS API or local to compute node)
- Cloud runners (DataFlow) have ~6 minute startup cost
- Less suitable for small requests
- Exotic environments difficult to support and scale
- Google Cloud Dataflow only full-service option

Architecture: Environmental annotation of biodiversity occurrence data



Yale

MOL MAP OF LIFE

Project deliverables

- Open source Apache Beam code to run data fusion requests on local and API accessible datasets
- HTTP API to manage data fusion requests on the Google Cloud Dataflow pipeline runner
- Command-line interface to interact with API
- Web front-end
- Suite of 1km products suitable for conservation science (1km daily temperature and precip)

Yale

MOL MAP OF LIFE