MyMergerTree:
A Cloud Service For Creating And Analyzing Galactic Merger Trees

Sarah Loebman, Jennifer Ortiz, Lee Lee Choo, Laurel Orr, Lauren Anderson, Daniel Halperin, Magdalena Balazinska, Thomas Quinn, and Fabio Governato

UM ASTRONOMY, UW CSE, UW ASTRONOMY

A Paradigm Shift in Science

Standard:

What data do I have to collect to (dis)prove a theory?

Data-driven:

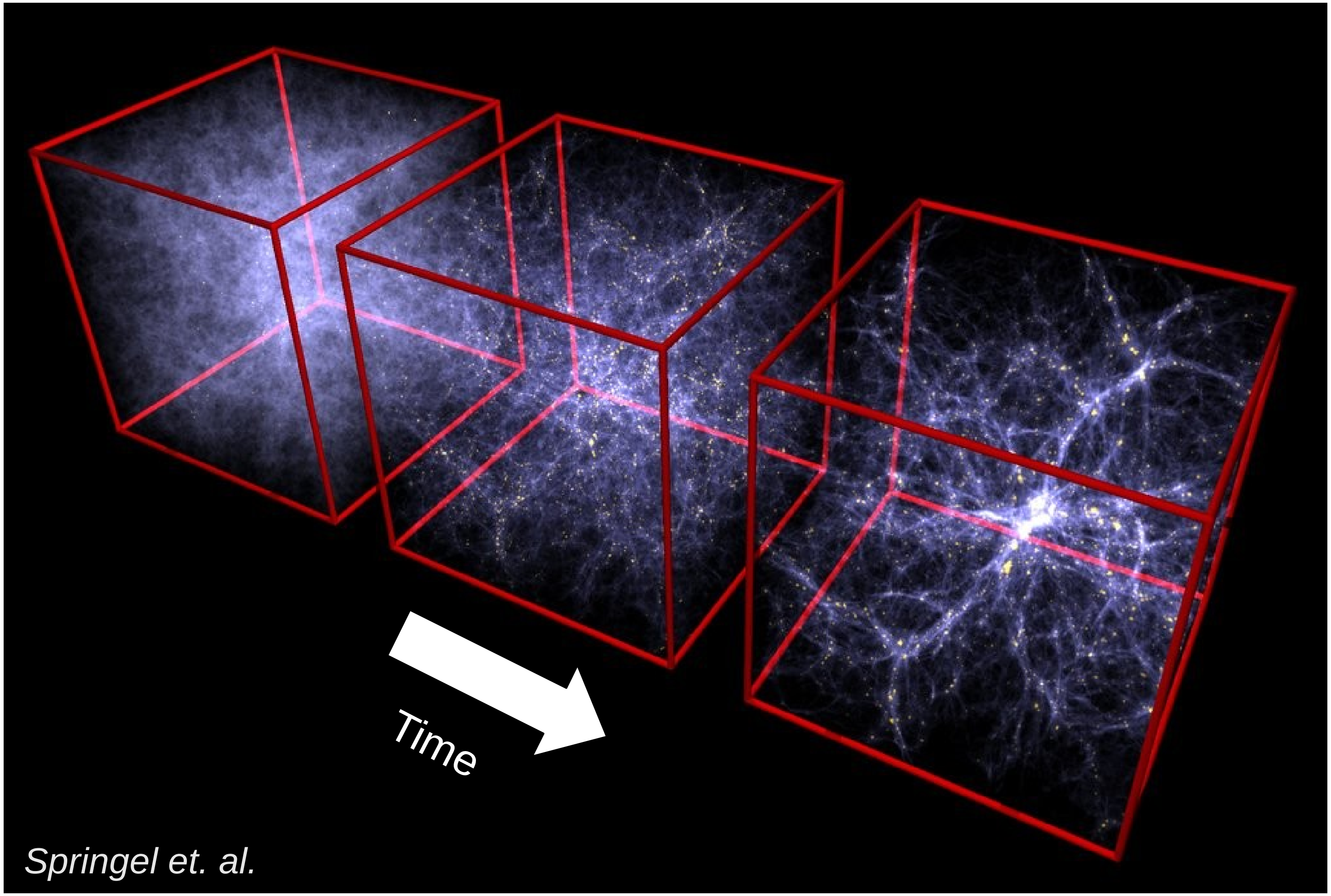
What theories can I test given the data that I already have?

Ex: The Sloan Digital Sky Survey (SDSS)

The Millennium Simulation

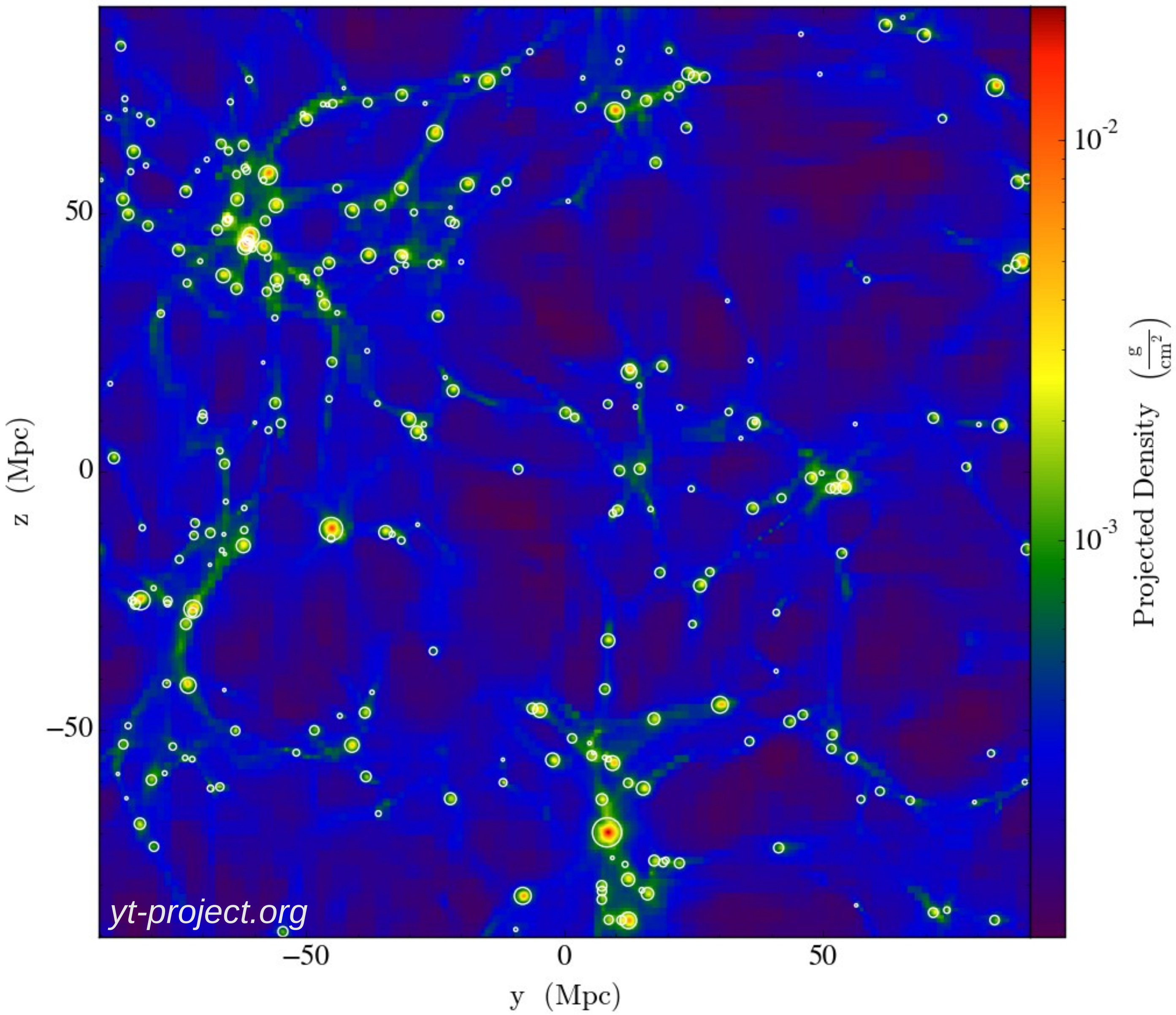
Vast datasets stored at discrete moments in time

Simulated Cosmological Volume



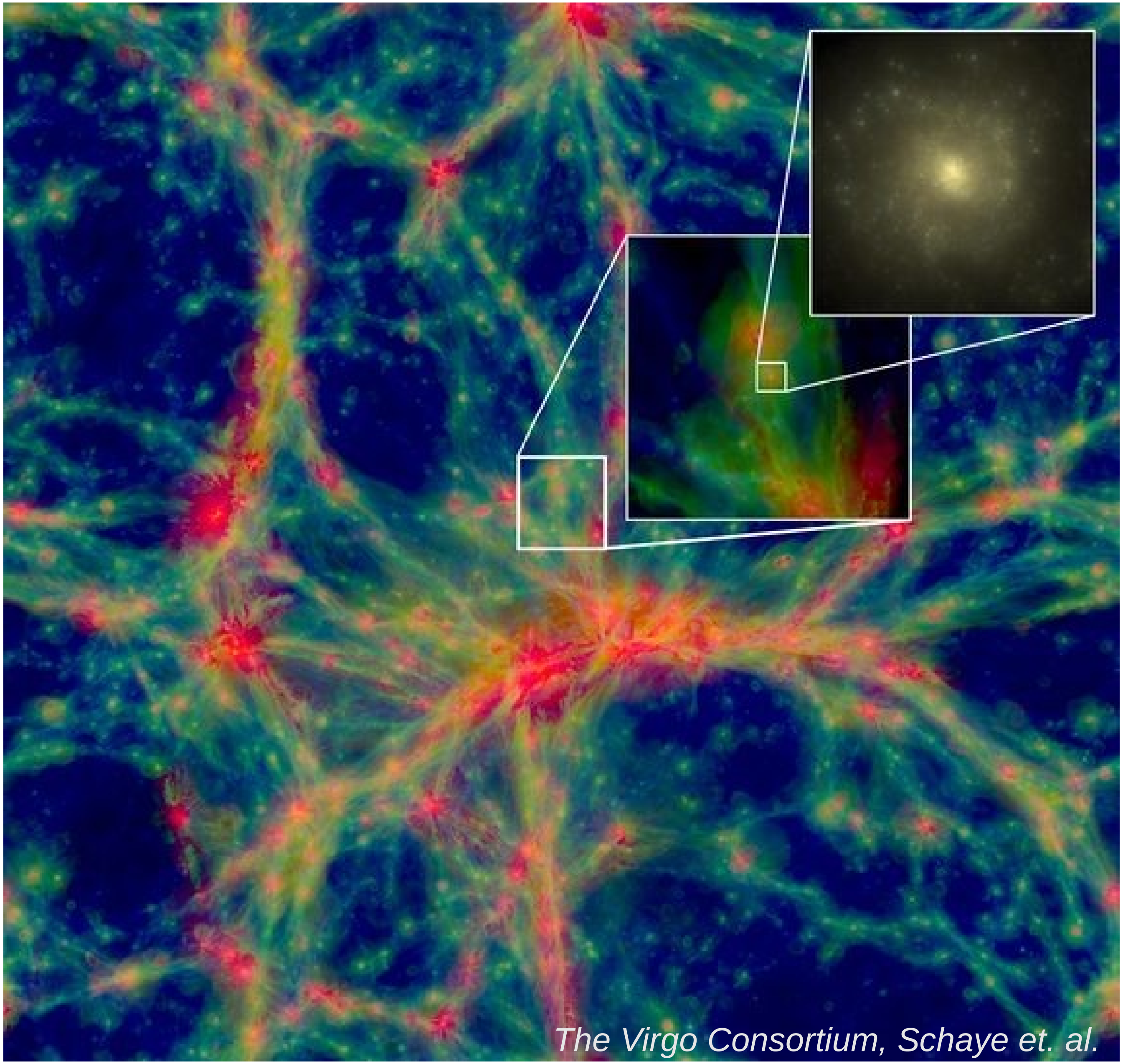
Springel et. al.

At each independent timestep galaxies identified via halo finder



yt-project.org

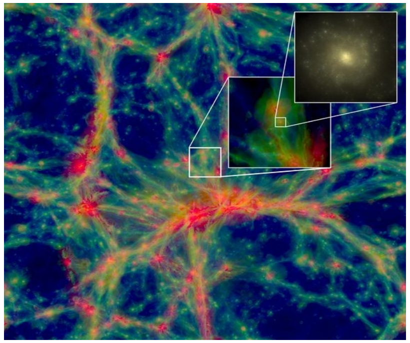
There are many galaxies within the cosmological volume

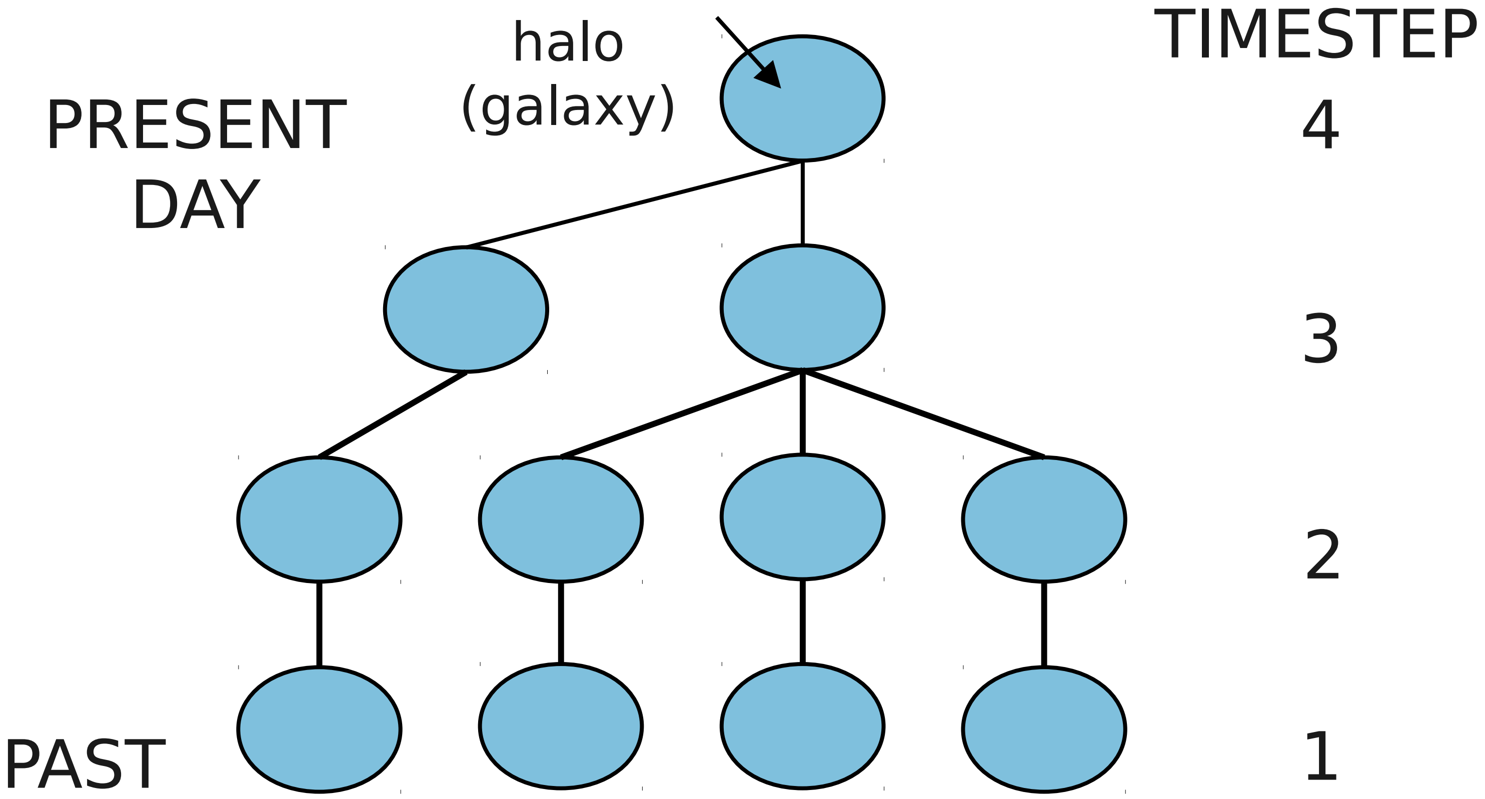


The Virgo Consortium, Schaye et. al.

Use merger trees to visualize how galaxies build up over time

Merger Trees





What tool can generate this structure from the data?

Look to a Big Data system (Myria) to produce merger tree.

Need a data service can match billions of particles across time

Visualization fast to load, platform independent, easy to use/share

Many Data Services

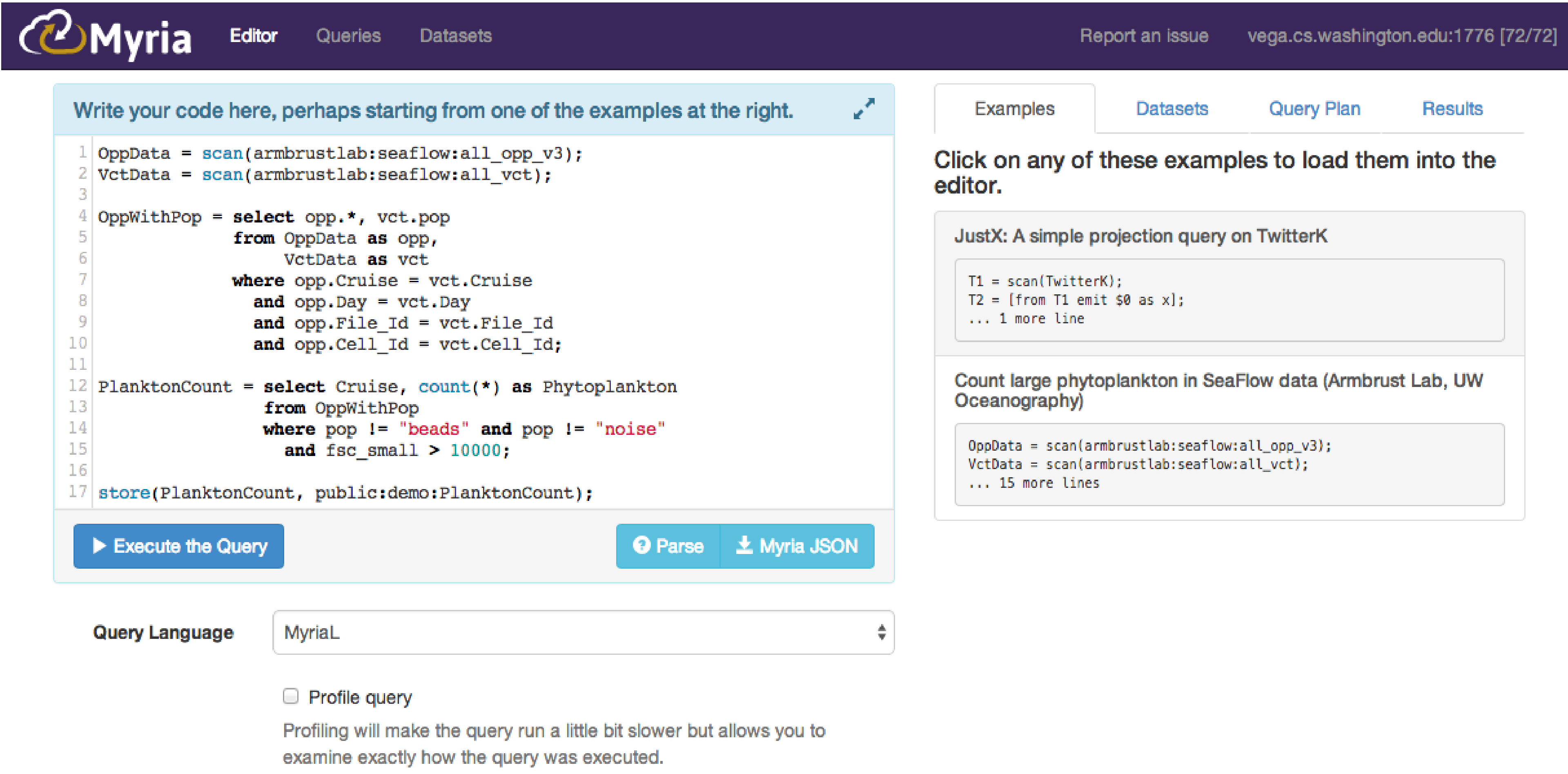
Many Cloud Services

Myria is a distributed database system accessible via browser

Myria

- ***In the Cloud***: A RESTful Query-as-a-Service platform
- ***Expressive***: A compiler framework for multiple iterative RA-based languages
- ***Efficient***: A parallel, shared-nothing, iterative execution engine

Myria in the browser



Write your code here, perhaps starting from one of the examples at the right.

```
OppData = scan(armbrustlab:seaflow:all_opp_v3);
VctData = scan(armbrustlab:seaflow:all_vct);

OppWithPop = select opp.*, vct.pop
             from OppData as opp,
                  VctData as vct
             where opp.Cruise = vct.Cruise
               and opp.Day = vct.Day
               and opp.File_Id = vct.File_Id
               and opp.Cell_Id = vct.Cell_Id;

PlanktonCount = select Cruise, count(*) as Phytoplankton
                from OppWithPop
                where pop != "beads" and pop != "noise"
                  and fsc_small > 10000;

store(PlanktonCount, public:demo:PlanktonCount);
```

Click on any of these examples to load them into the editor.

JustX: A simple projection query on TwitterK

```
T1 = scan(TwitterK);
T2 = [from T1 emit $0 as x];
... 1 more line
```

Count large phytoplankton in SeaFlow data (Armbrust Lab, UW Oceanography)

```
OppData = scan(armbrustlab:seaflow:all_opp_v3);
VctData = scan(armbrustlab:seaflow:all_vct);
... 15 more lines
```

Query Language: MyriaL

□ Profile query

Profiling will make the query run a little bit slower but allows you to examine exactly how the query was executed.

Generating a Merger Tree in Myria

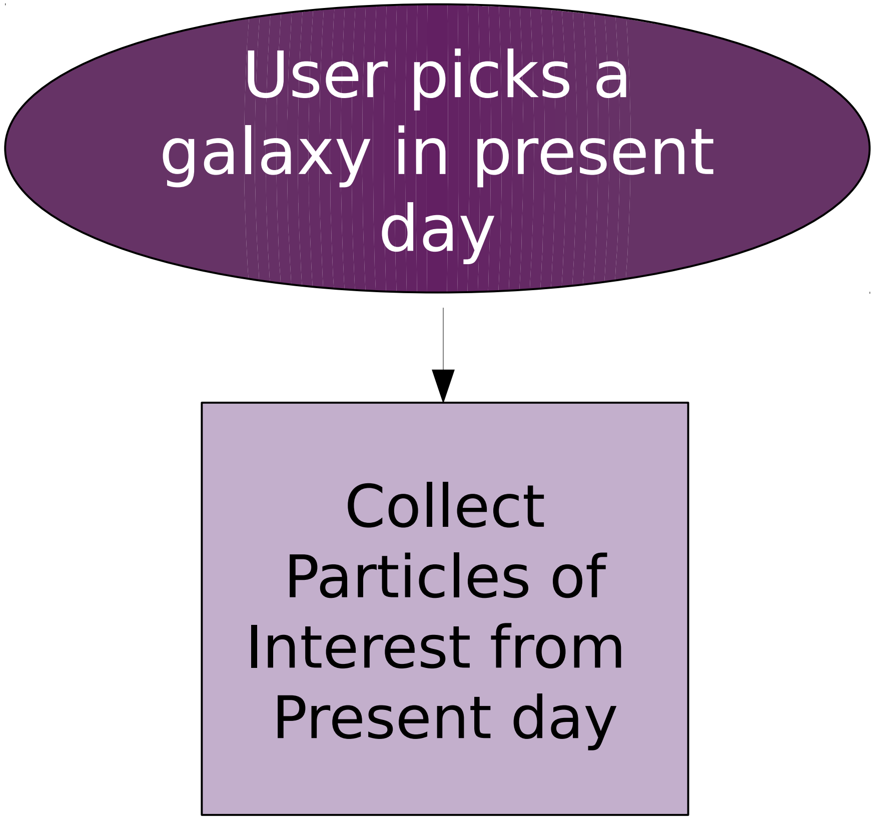
Demo

We saw 3 types of challenges in building this tool.

Challenges:

- Expressing scientific problems declaratively
- Physical Tuning for high performance
- Visualizing naturally and easily

How to generate trees?



User picks a galaxy in present day

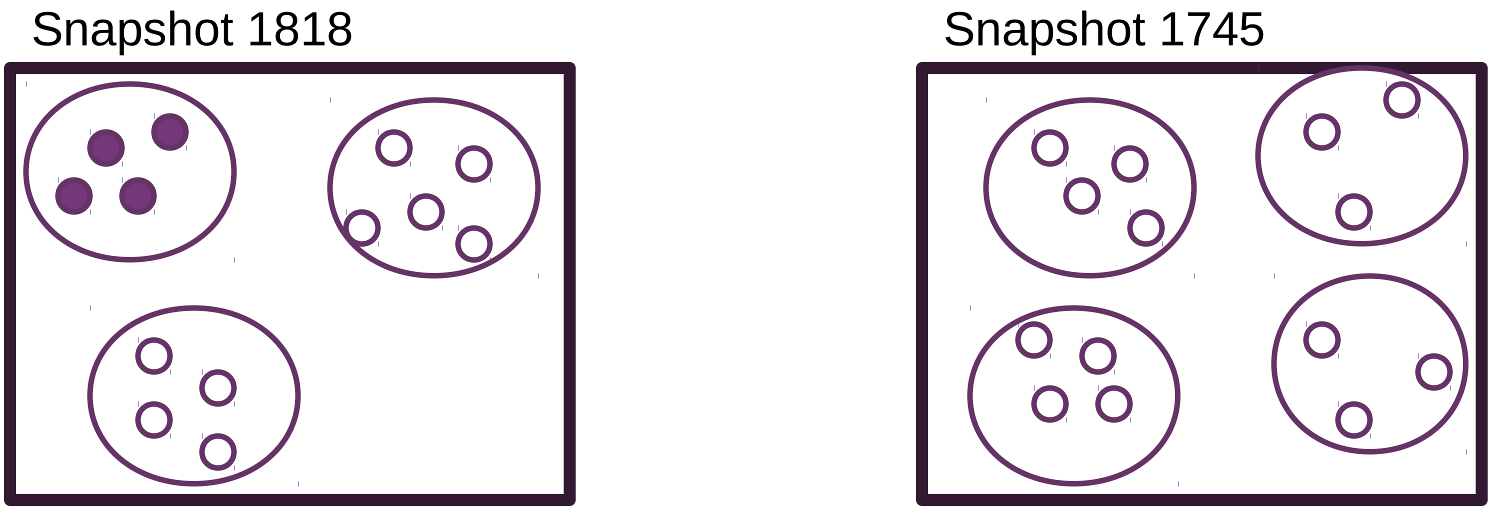
Collect Particles of Interest from Present day

Look earlier in the simulation….
Where are the particles at previous timesteps?

The science question actually easy to express in SQL!

Select *ParticlesOfInterest*

```
SELECT s.iOrder, s.mass, s.type, s.grp
FROM Snapshot1818 s  -- present day snapshot
WHERE s.grp = ‘user selection’
```



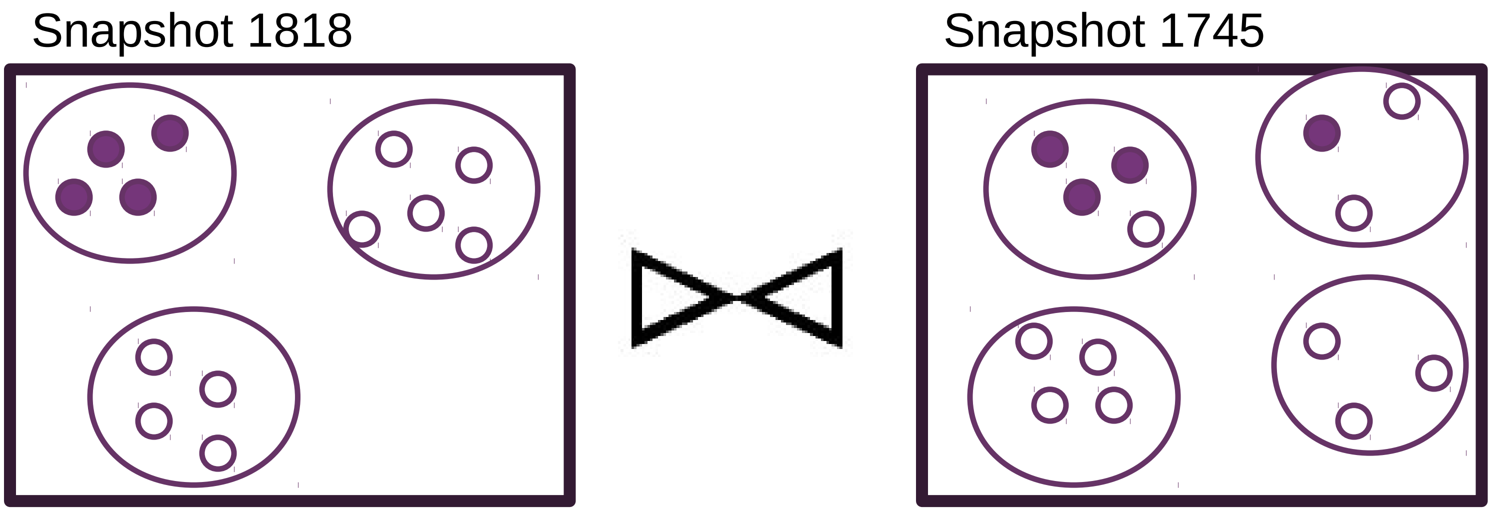
Problem reduces to GROUP BY on a join table

Join across time *AllParticlesTable*

SELECT i.iOrder, i.mass, i.type, i.time, i.grp

FROM *ParticlesOfInterest* i, Snapshot1745 s

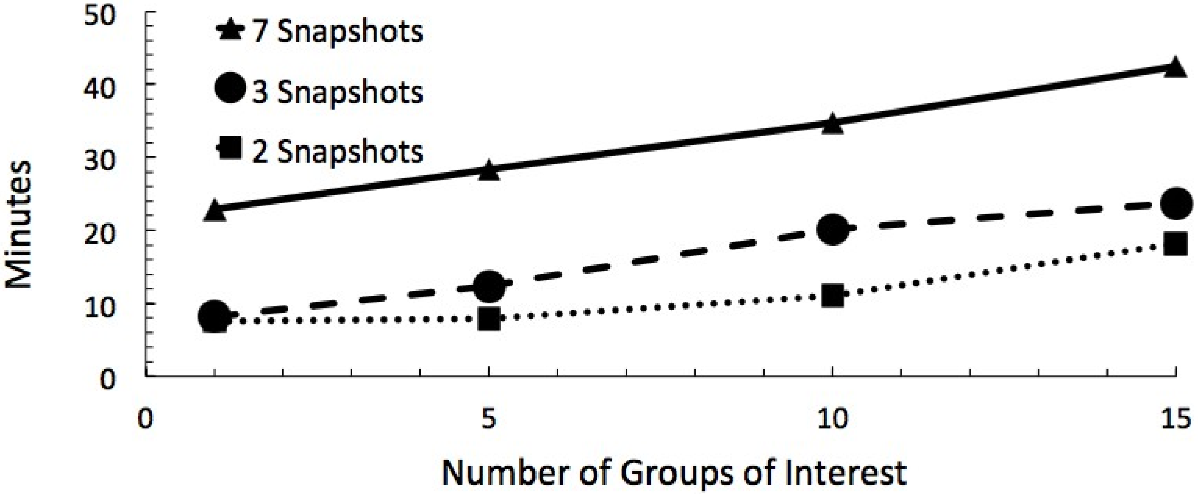
WHERE i.iOrder = s.iOrder



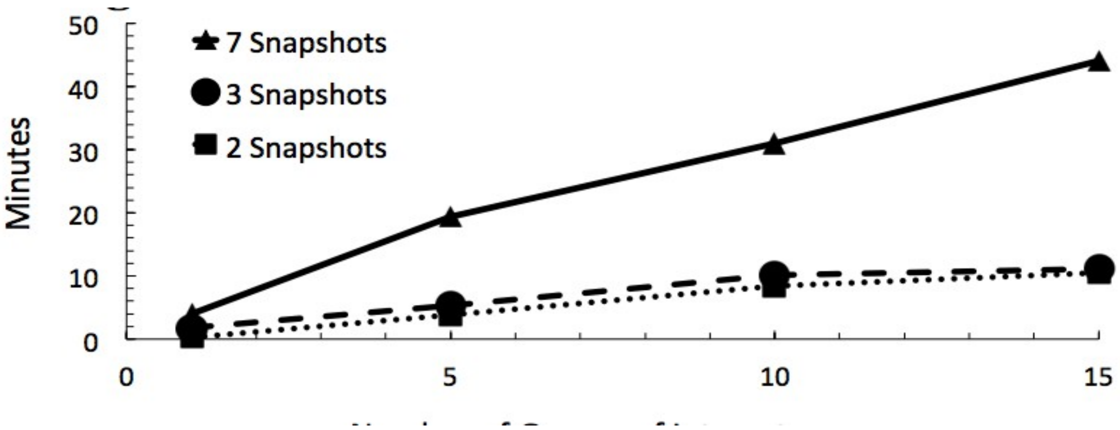
Physical tuning was the greatest challenge

How data was arranged had a huge impact on query speed!

Not Hashed on Ingest

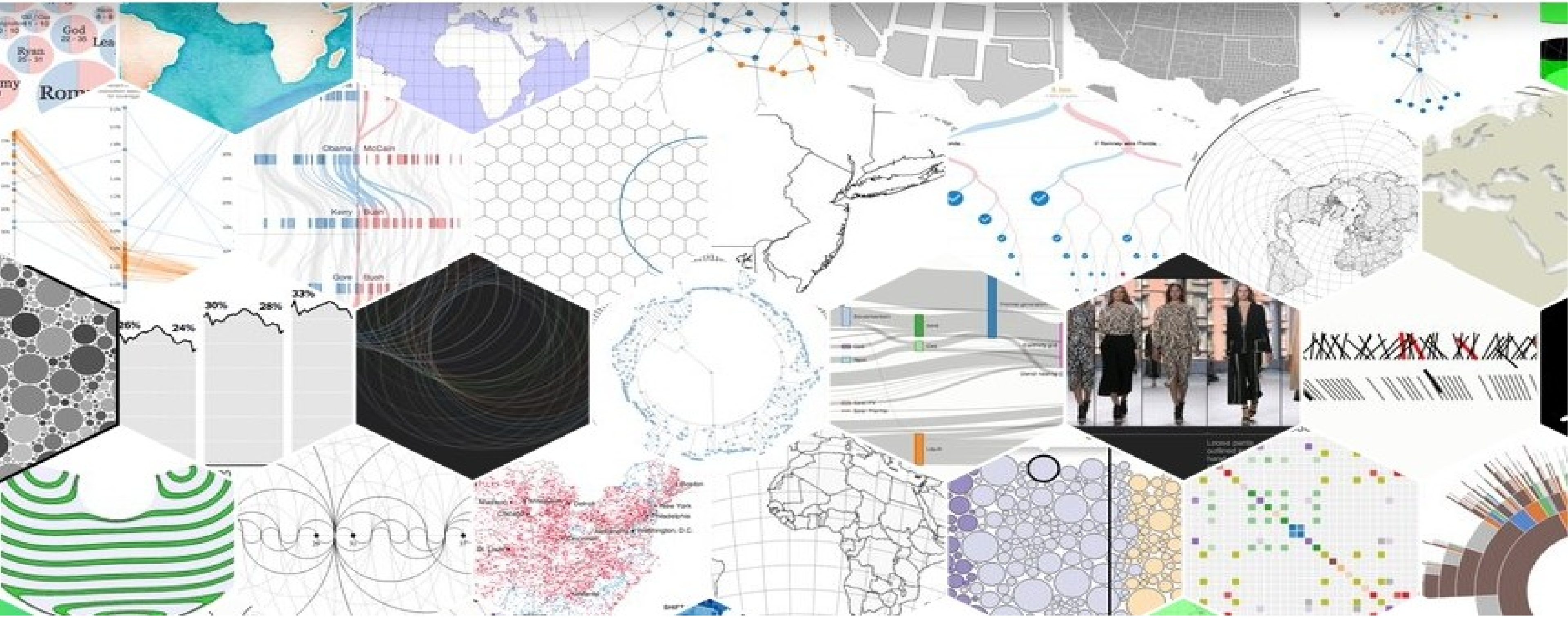


Hashed on Ingest



Visualization straight-forward using D3 javascript

Data-Driven Documents



d3js.org

Conclusions

- Translation of problem easier than predicted
- Physical tuning required most effort
- Specialized visualization tools existed
- Ingesting and validation still needed

How efficiently can we solve other problems?

Exploring dynamic arrangement of data.

Questions?

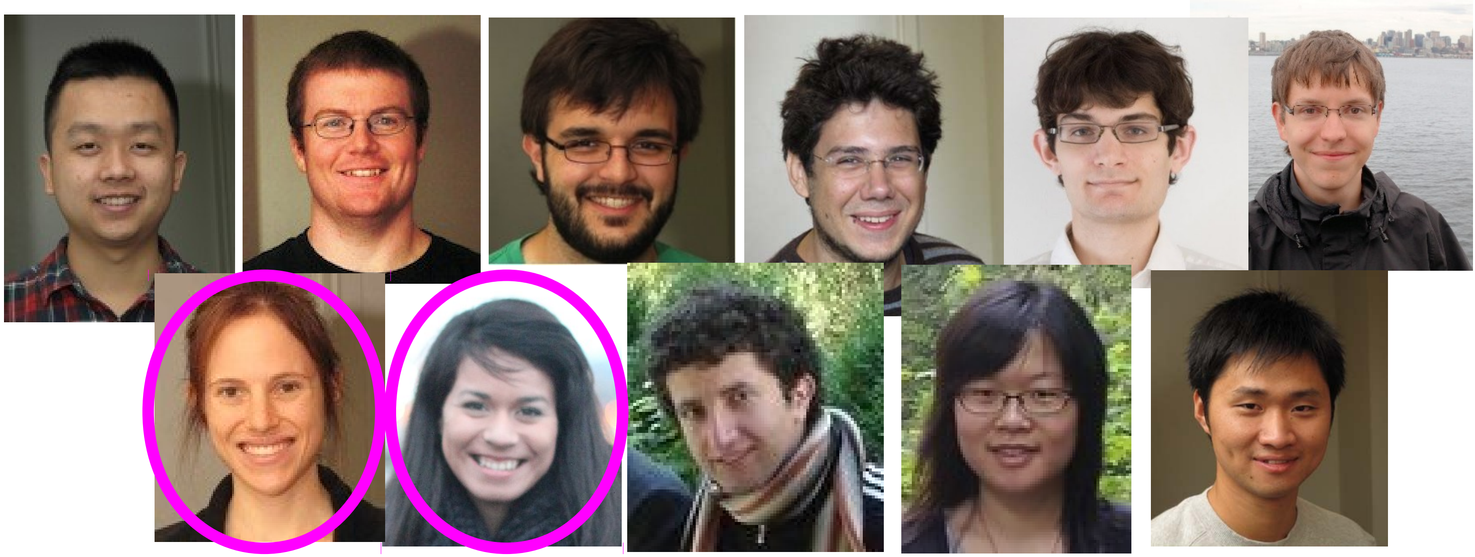
Myria Team

Magda Balazinska, Bill Howe, Dan Suciu (faculty)

Dan Halperin (postdoc, technical lead)
Victor Almeida (postdoc)
Andrew Whitaker (research scientist)

PhD Students

- Shumo Chu
- Eric Gribkoff
- Jeremy Hyrkas
- Paris Koutris
- Ryan Maas
- Dominik Moritz
- **Laurel Orr**
- **Jennifer Ortiz**
- Emad Soroush
- Jingjing Wang
- ShengLiang Xu



Undergraduate Students

- Lee Lee Choo
- Vaspol Ruamviboonsuk