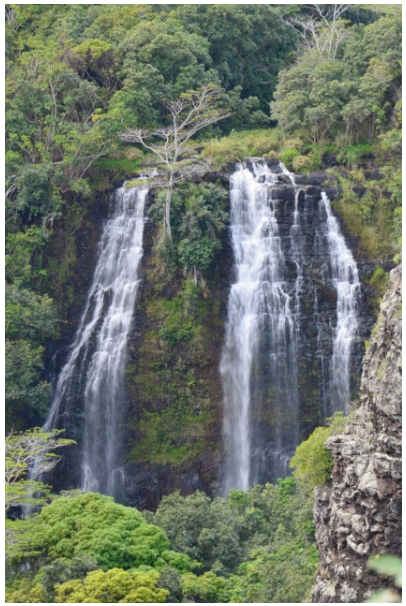
Systems Session
Database Day 2015

Dep. of Computer Science & Engineering
University of Washington

The Exciting Times of "Big Data"



Everyone today has a big data problem

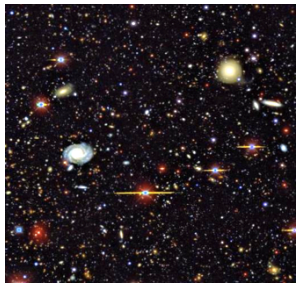
- Whether it is a data lake, data swamp, or data stream
- Whether they call it big data, data science, data wrangling, ..

Photo by Gary Bridgman / CC BY

Challenging Application Requirements

Exciting and challenging requirements of campus applications

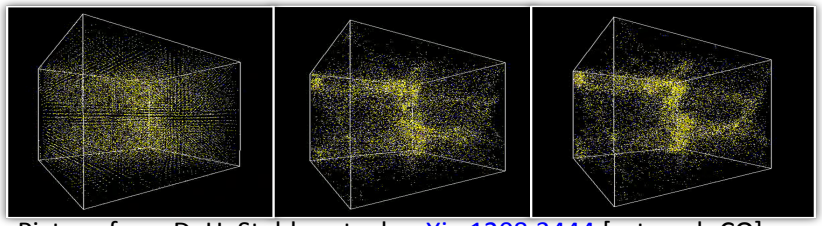
Use to motivate & test -- Often generalize beyond campus



Telescope image:

1. Iterative data cleaning
2. Objects extraction
3. Classification

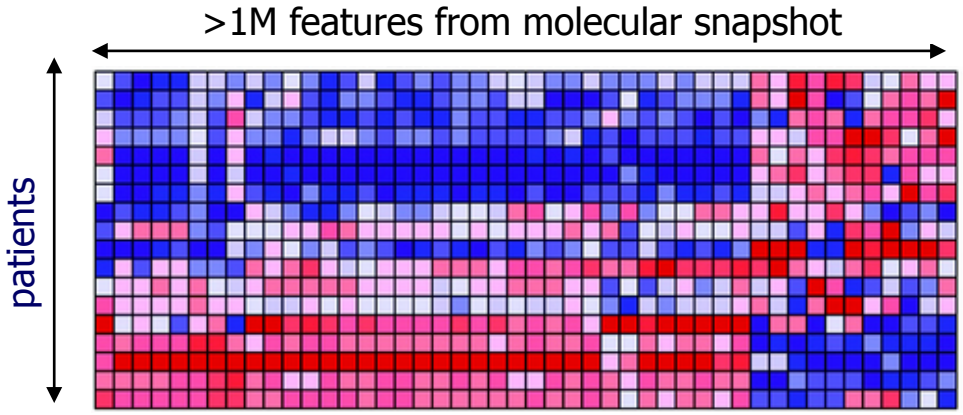
Picture from Deep Lens Survey (DLS: Tyson)



Picture from D. H. Stalder et. al. arXiv:1208.3444 [astro-ph.CO]

N-body simulation data:

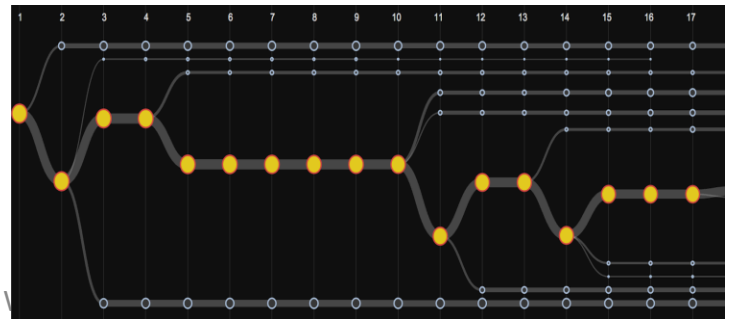
1. Manage hundreds of TB of data
2. Data clustering to extract galaxies
3. Graph analytics to study galaxy evolution



Picture from Su-In Lee

Genome data processing:

1. Linear algebra on large matrices
2. Novel machine learning algorithms



The Challenge

- *Everyone* needs to work with "big data"
  - Big data = data with large volume, velocity, or variety

- Need tools that
  - Can manage big data efficiently
  - Can analyze big data efficiently (complex analytics)
  - Are geared toward being used by data scientists

- Core focus
  - **How to make data scientists maximally productive?**

We Build on Open-Source Tools

Developed **ParaTimer** [SIGMOD10]

- Shows progress of DAGs of Hadoop jobs

Developed **PerfXPlain** [VLDB12]

- Explains the performance of Hadoop jobs

Developed **SkewReduce** [SOCC10] and **SkewTune** [SIGMOD12]

- Based on Hadoop and available as open source

Developed **HaLoop** [VLDB10]

- Faster iterative processing in Hadoop also open source

Developed **Array Proc. Methods** [SSDBM15, ICDE13, SIGMOD11]

- Array storage and query processing in SciDB

Goals of the Myria stack

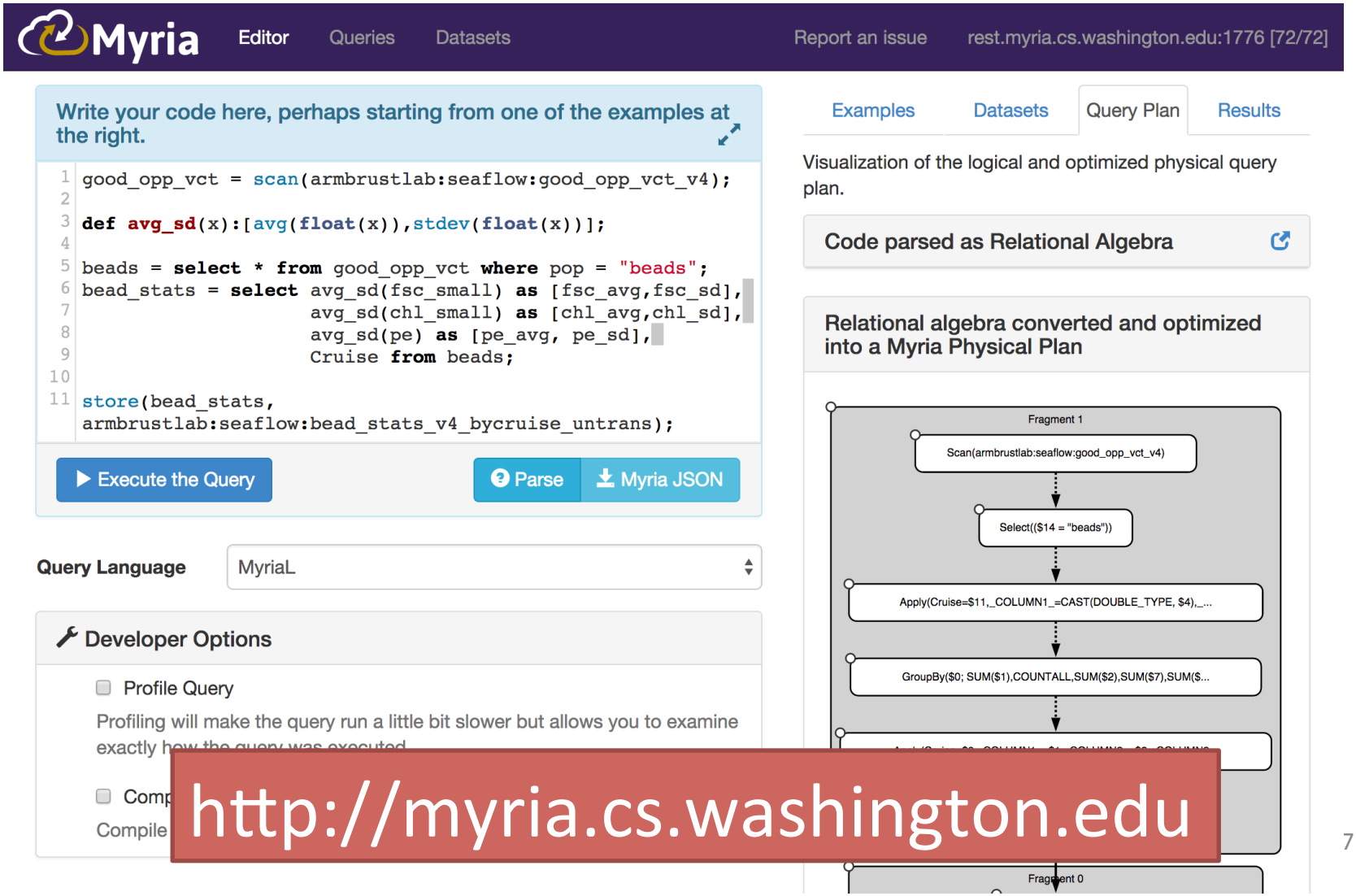
- Advance state-of-the-art in big data systems
- Focus on efficiency and productivity
- Test on real applications and support real users

Deliverables:

- Built a new **big data mgmt & analytics system**
- Deployed and operate Myria as a **service**

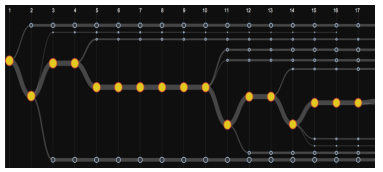
Myria Big Data Management Service

Myria is a cloud service: Just open browser and go!



http://myria.cs.washington.edu

Myria Is a Cloud Service



MyMergerTree

Browser | Python | Specialized Services

RACO

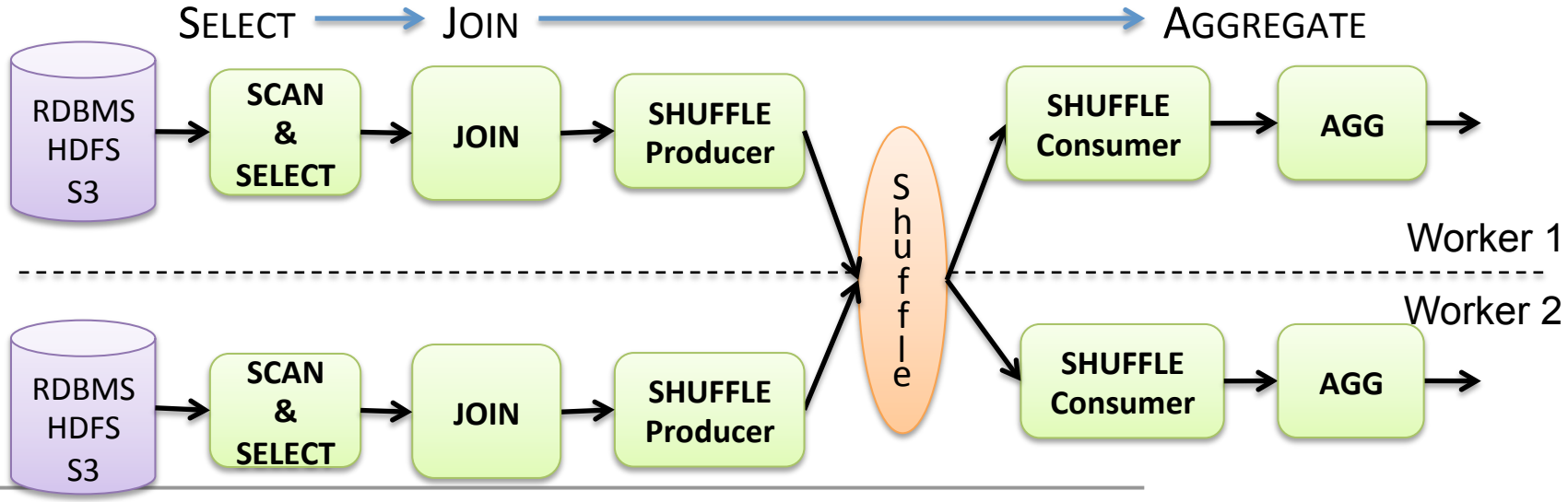
Query Translation, Optimization, and Orchestration

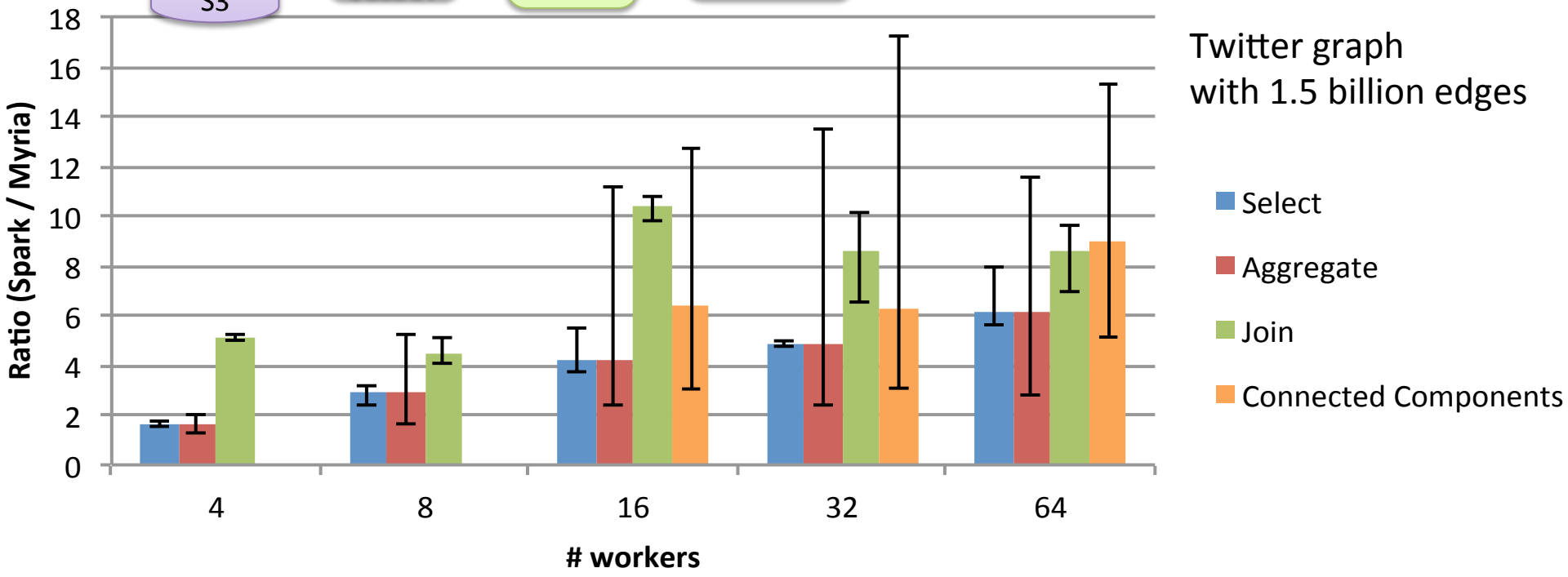
MyriaX

Parallel, Iterative, and Elastic Query Execution

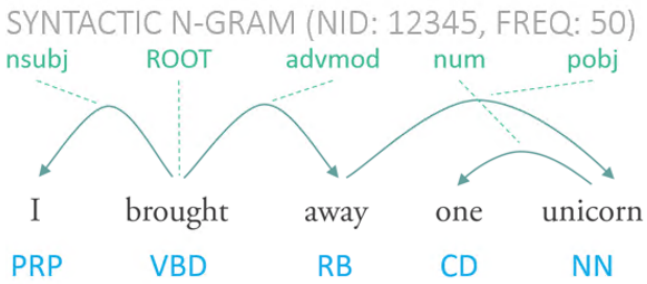
MPI | SciDB | Graphs | NoSQL

MyriaX Query Execution Engine

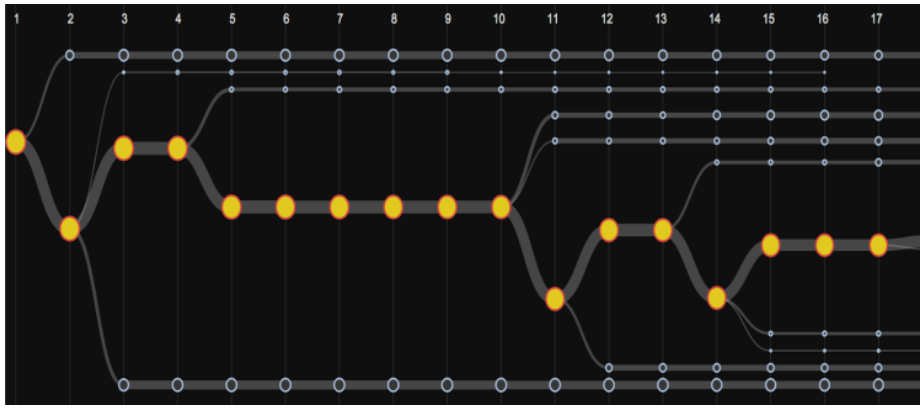




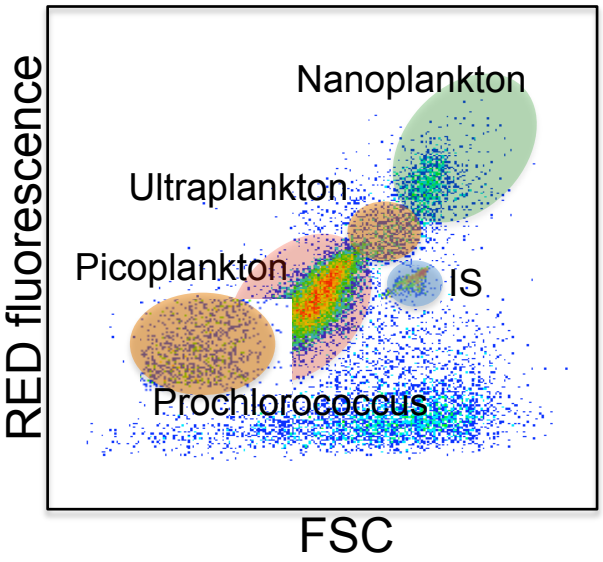
Example Myria Applications



Natural Language Processing

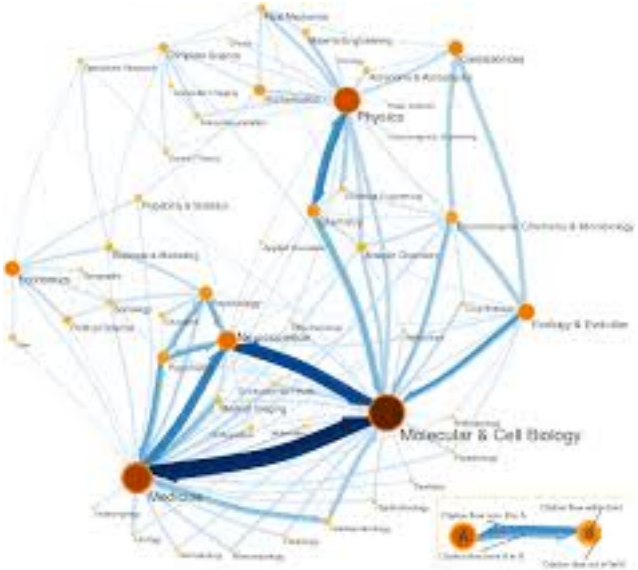


Galaxy Simulations



Environmental Flow Cytometry

Telescope Images



Bibliometrics

Some of Key Research Themes

Efficient big data management and analytics

- Efficient multi-join query processing (**Shumo**)
- Iterative & in-memory query processing (**Jingjing**)
- Data summarization (**Laurel**)

Effective operation as a cloud service

- Personalized Service-Level Agreements (**Jennifer**)
- Query time guarantees (Brendon & **Jennifer**)
- Predictable and explainable performance (**Parmita & Helga**)

Easy to use even for complex tasks

- Cross-system analytics – Auto connectors (**BrandonH**.)
- Cross-system analytics – Algebra (**Dylan**)
- Linear algebra support (**Ryan**)

Beyond Big Data

Transaction Processing

- Making optimistic concurrency control faster (**Bailu**)
- With Johannes (Microsoft) and Lucja (Cornell)