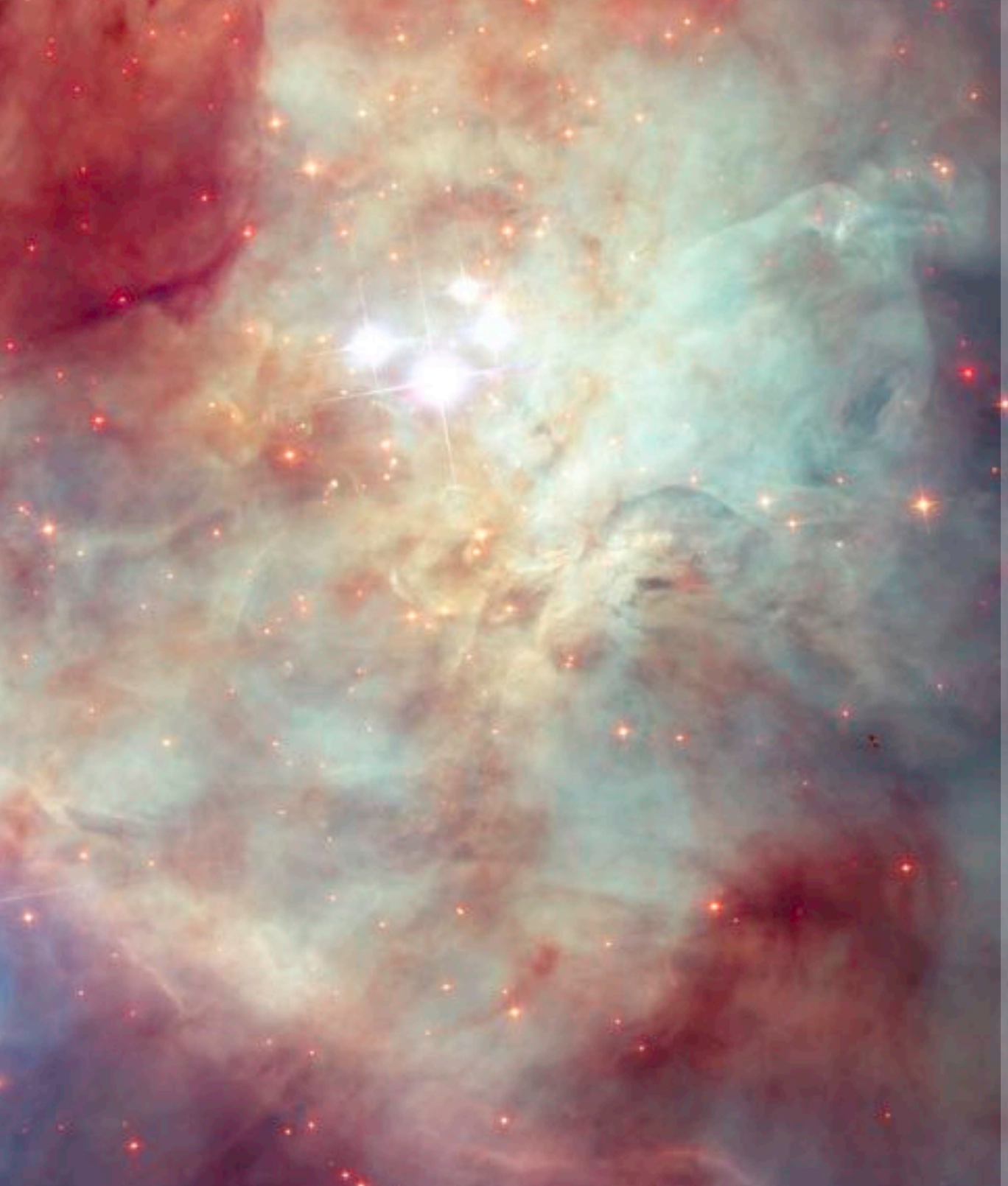


A Cross-Disciplinary Scheme for Scientific Data Descriptors

Demitri Muna

Cold Spring Harbor Laboratory

@demitrimuna

I Annotate • San Francisco • 6 June 2018

Scientific Data User Interfaces

Sloan Digital Sky Survey

http://data.sdss.org

Cold Spring Harbor Laboratory

http://cshl.edu

THE STARCHIVE

http://starchive.org

Nightlight

http://nightlightapp.io

Trillian

http://trillianverse.org

Big Data

Astronomy

- Gaia DR2: 1.3B stars
- Pan-STARRS PS1: 1.3PB data, 3B objects
- Astronomical tables can have hundreds of columns for each object.

Genomics

- Genetic sequences are comprised of hundreds of millions to billions of base pairs, each.
- Sequences may differ only slightly across species (mutations, evolution), meaninful for the way they are expressed (e.g. physical features, disease resistance).

… & most other sciences have large data sets …

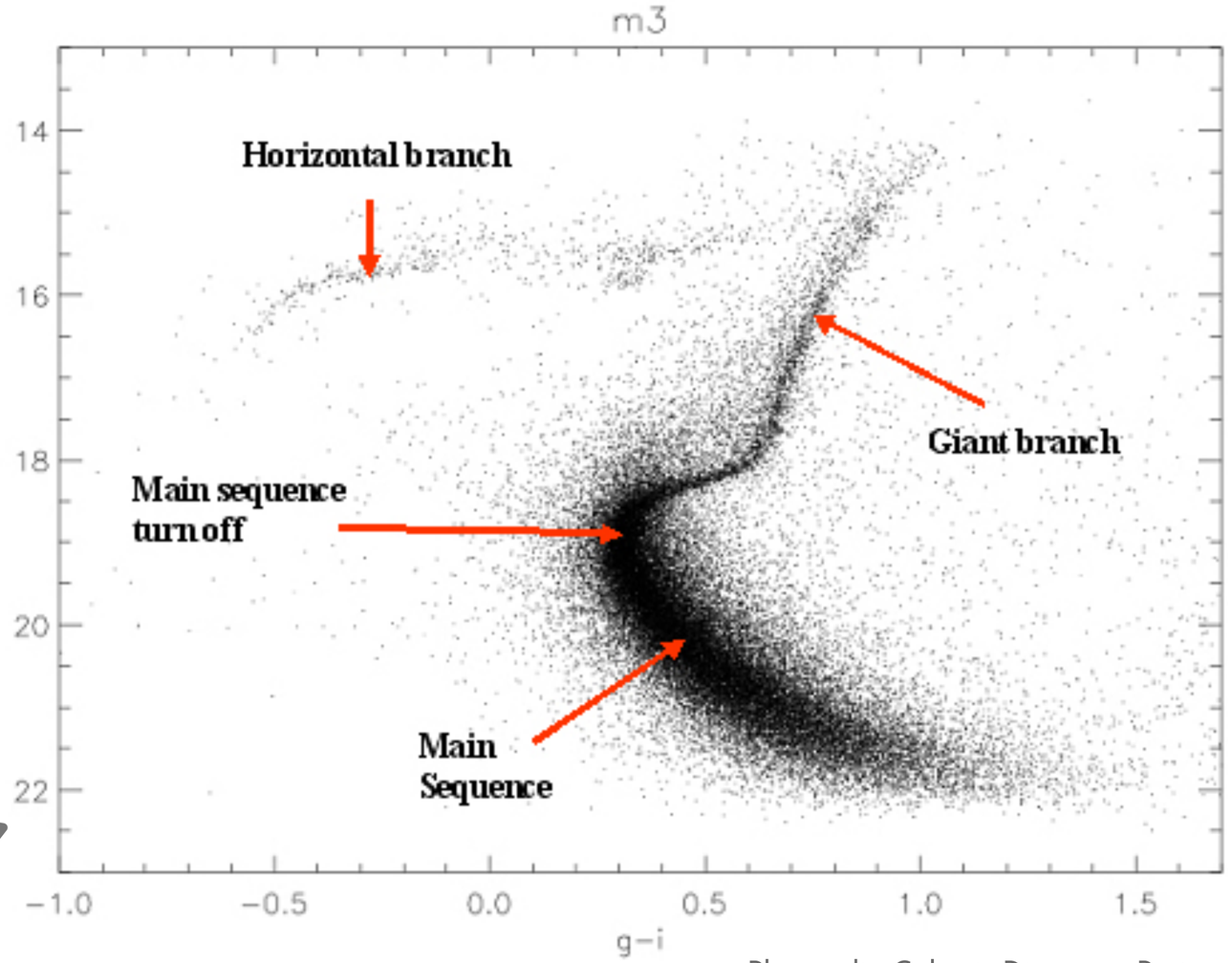
Data is Aggregated

Data is aggregated to handle the large volume: plots, statistics, machine learning – there is too much to view individually.

We look for trends to discover and understand the underlying models.

M3 Globular Cluster

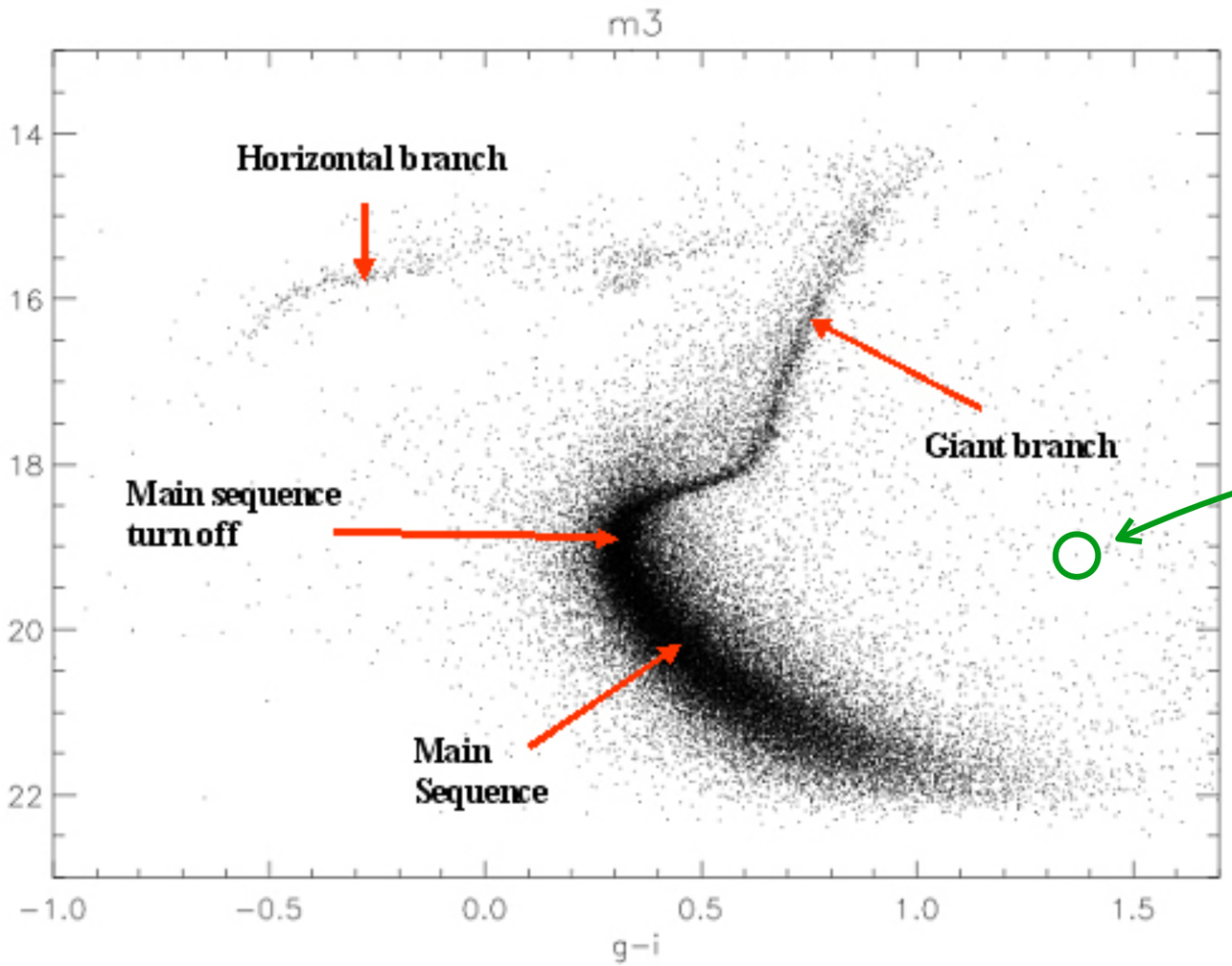
~ 500,000 stars



Plot of all objects identified as "star" from a public data set (SDSS).

Plot credit: Galactic Discovery Project
Photo credit: Sloan Digital Sky Survey

M3 Globular Cluster



Clear trends that match physical models visible… but what is this outlier?

- Two unresolved stars?
- Not a star?
- Star not part of M3?
- Instrument effect?
- Processing error?
- Something new?

Outliers

To study outliers, we look at the individual data point. This requires human intervention, domain knowledge, and expertise. Once the outlier is identified… nothing happens:

- Data set is not updated (releases are static).
- Classification is not updated.
- Bad data is not flagged.

The outlier will need to be identified again by the next scientist.

Scientific Data Descriptors

What if individual elements of data can be annotated?

We would need descriptors for scientific data. Goals:

- Descriptor easily generated.
- Descriptor reasonably human readable.
- Descriptor works with Hypothesis infrastructure.
- Does not require data creator to define descriptor.
- Refer only to publicly released data sets.
- File format agnostic.
- Descriptors are easily citable, searchable.

Resource Identification Initiative

A project called the Resource Identification Initiative (Maryann Martone, UCSD) has pioneered this kind of data descriptor. Research Resource Identifiers (RRID) aim to be: machine readable, free to generate and access, consistent across journal publications. Since they are a string, they can be searched just as text and easily embedded into journal papers.

RRID:MGI:4440559

RRIDs are used in biology. This proposal is to design a (reasonably!) uniform scheme for all disciplines.

Why Not DOIs?

- DOIs are appropriate for the data set itself, not an individual row in a table or gene from a larger sequence.
- Can't expect or wait for data creator to "mint" a DOI. Some won't for just the data, referring one to the methods papers to cite. Other data sets are decades old and the creators have long disbanded.
- Don't want to mint 1.3 billion DOIs for individual data points from one data set. Identifiers should only be created/stored when annotations have been created.

URL vs URN

URL = uniform resource locator

URL vs URN

URN = uniform resource name
(which is a kind of URI → uniform resource identifier)

- identify a resource over a long period of time
- *not* bound to a location
- not resolvable
- remains unique

Hypothesis uses URLs as identifiers to annotations, but URNs can be used.

Example:

`urn:isbn:0345391802`

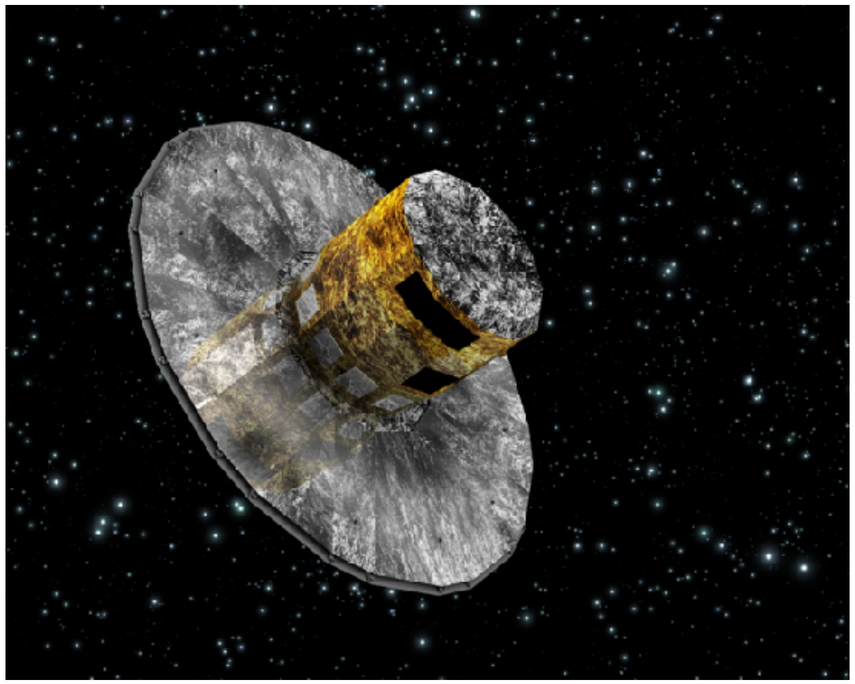
namespace

ISBN of Hitchhiker's Guide to the Galaxy

There is no one canonical copy or location of the book.

Example Data Set: Gaia Data Release

As an example, let's use a recent data release from the Gaia spacecraft. This is a mission to precisely measure the distance to 1% of the stars in the galaxy.



- 1.3 billion records (stars)
- tabular format
- each row has unique integer identifier
- several columns on each row

Scientific Data Descriptor

scientific field (astronomy)

data release/ product/ version

urn:sdd:astro:gaia:dr2:17351859

sdd = scientific data descriptor

data source

unique identifier defined in data set

This descriptor would point to the entire record of one star.

Scientific Data Descriptor

Use URL-style fragments to refer to individual columns.

```
urn:sdd:astro:gaia:dr2:17351859#ra,dec
```

This would refer to the columns "ra" and "dec" (coordinates) of the record. References to subsets of the data can be created using the URL-style query language:

```
urn:sdd:astro:gaia:dr2?ra=10,20&dec=-1,1
```

Aggregate References

We can take advantage of the fact that the data sets (releases) are static and make references to aggregate values. Statistics can be represented by referring to columns. This could refer to a histogram of RA values between 0 and 60:

aggregate

```
urn:sdd:astro:gaia:dr2:agg:hist:ra?ra=0,60
```

Maximum values of two columns:

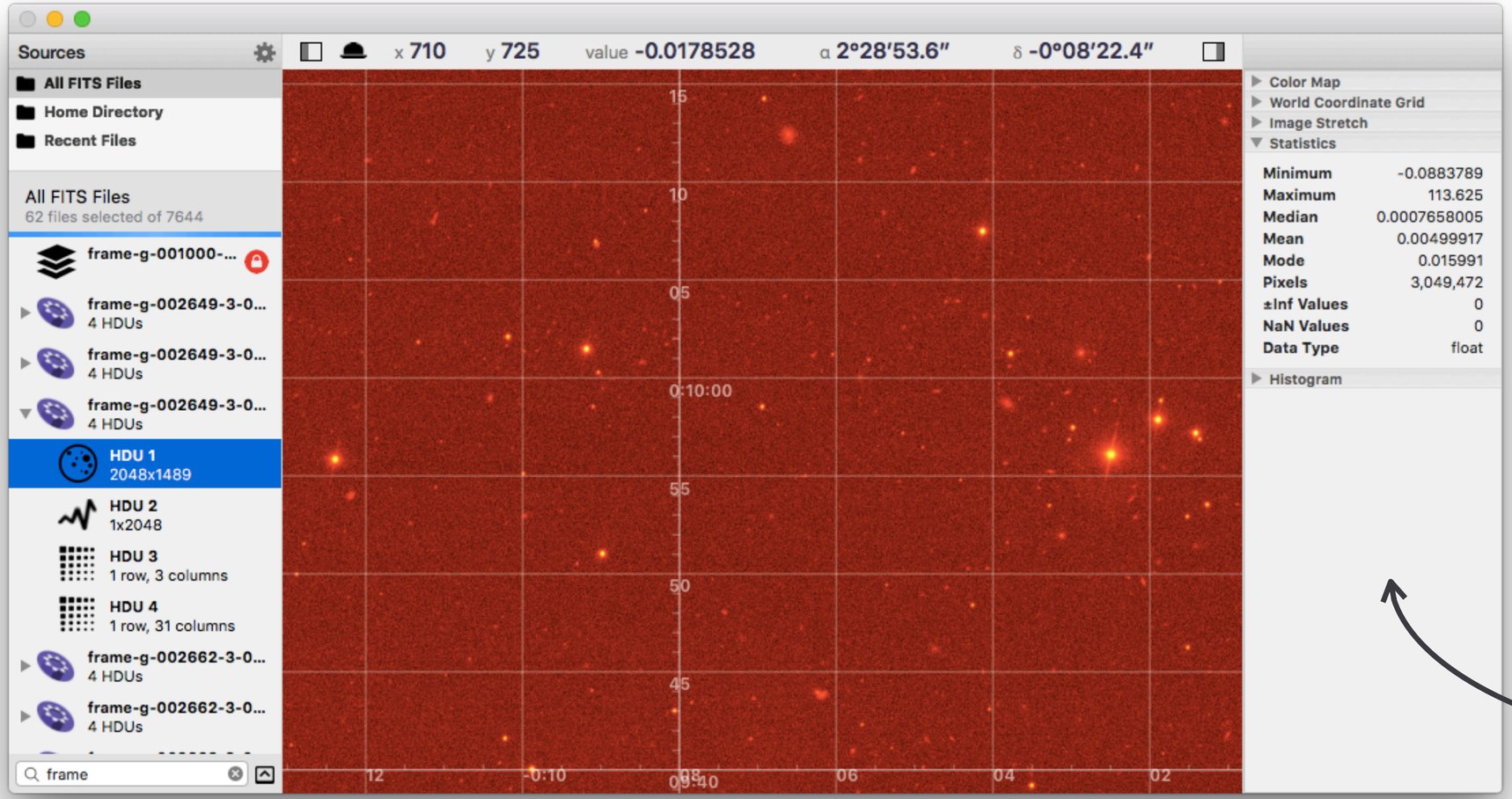
```
urn:sdd:astro:gaia:dr2:agg:max:ra,dec
```

Here, we offload the computation of extremely large data sets to the cloud (where we can take advantage of caching).

Use Cases

- User interfaces can attach annotations to any data visualization (e.g. web, desktop).
- Researchers can mark data ("textbook example of x", "instrument error", etc.) that others can learn from.
- Automated classification, e.g. bots that combine several data sets and assign likelihood values of model fits
- Educational – "looking over shoulder" of domain expert.
- Dramatically lowers bar for shared research – comments can be made and viewed long before the publication process.
- Serendipity – open a file for one purpose, draw connections from comments from another researcher.

Data Visualization

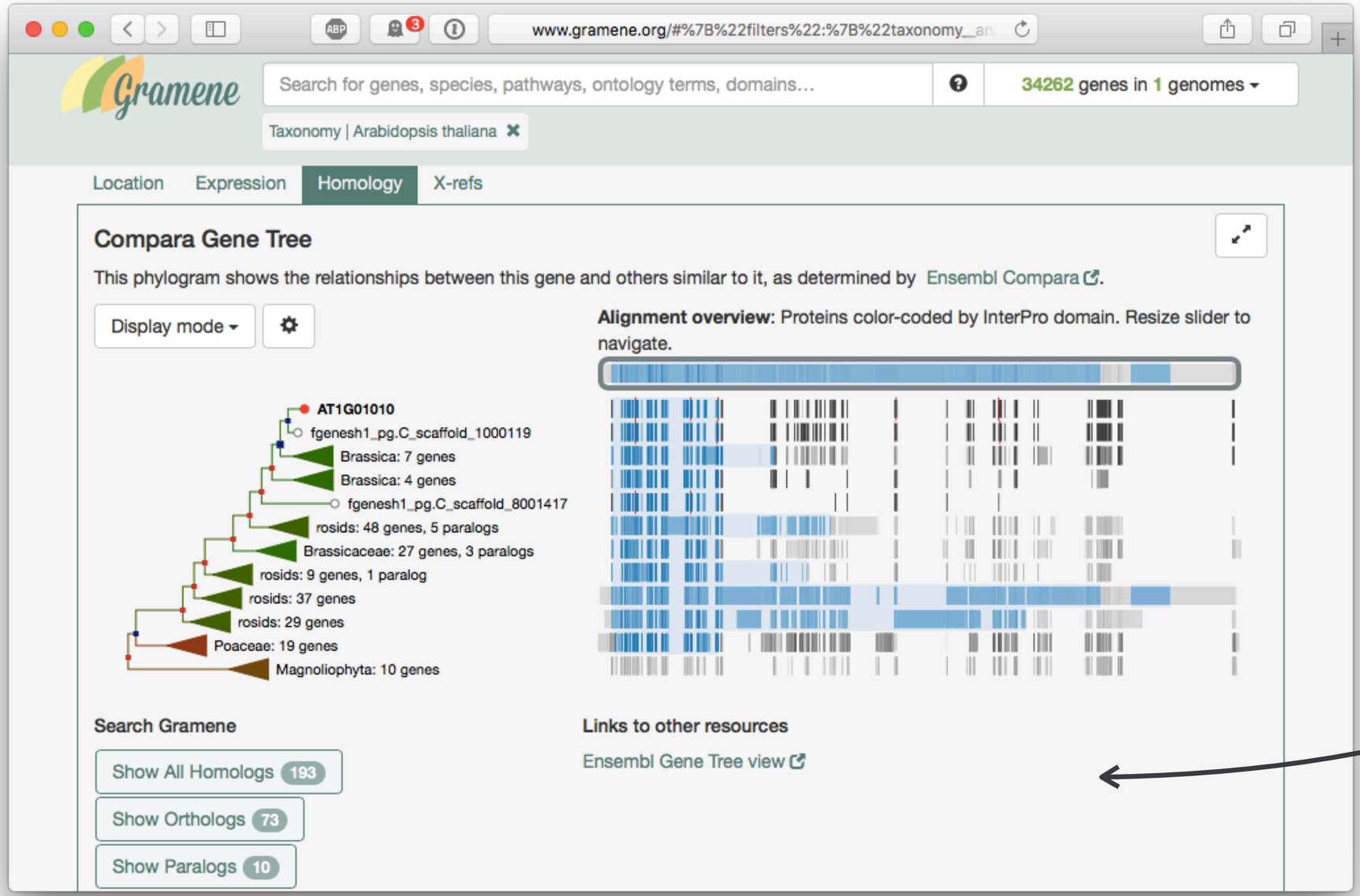


Integrate community annotation to common data visualization tools.

When a file is opened for viewing, automatically identify any objects in the data that have annotations.

Demitri Muna • @demitrimuna

Data Visualization



Integrate community annotation to common data visualization tools.

Attach annotations to web reference tools.

Feedback?

- Use cases not covered?
- Scheme sufficiently applicable to different disciplines?
- Extend scheme to data not in a tabular format, e.g. regions or objects in images? (see Web Annotation Data Model: https://www.w3.org/TR/annotation-model/)
- URIs are not locatable (by design), but this would be useful. Perhaps a registration scheme?