Wrangling ZTF Light Curves

Mario Juric, Zach Golkhou, Lynne Jones,
Petar Zecevic, Colin Slater, Andy Connolly

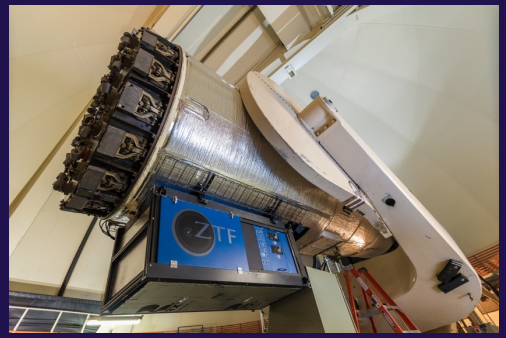
University of Washington DIRAC Institute
& the ZTF Partnership

Before I begin…

Who are we?

Data Intensive Research in Astrophysics and Cosmology

- DiRAC is a new institute housed in the Astronomy Department. It focuses on discoveries about the origins of our universe enabled by new computational and statistical approaches
- The Institute comprises 24 faculty, research scientists, postdoctoral fellows, and students
- The research focus is on Time Domain astronomy, Solar System research, Data Engineering, Astronomical Software related to the Large Synoptic Survey Telescope and Zwicky Transient Facility surveys



Zwicky Transient Facility

Large Synoptic Survey Telescope

ZTF Alerts vs. ~~light curves~~ time series

- ZTF alerts are there to enable rapid response
  - Really designed for machine-to-machine communication, not to be directly used by humans
  - Include 30 days of prior observations, but not more than that
  - Do not include objects that don't change more than +/- 5\sigma
- For work at large-scales that is not tome sensitive, we'd like something closer to a database of time series

MariaDB Time Series Database @ UW

… i.e., Version 1

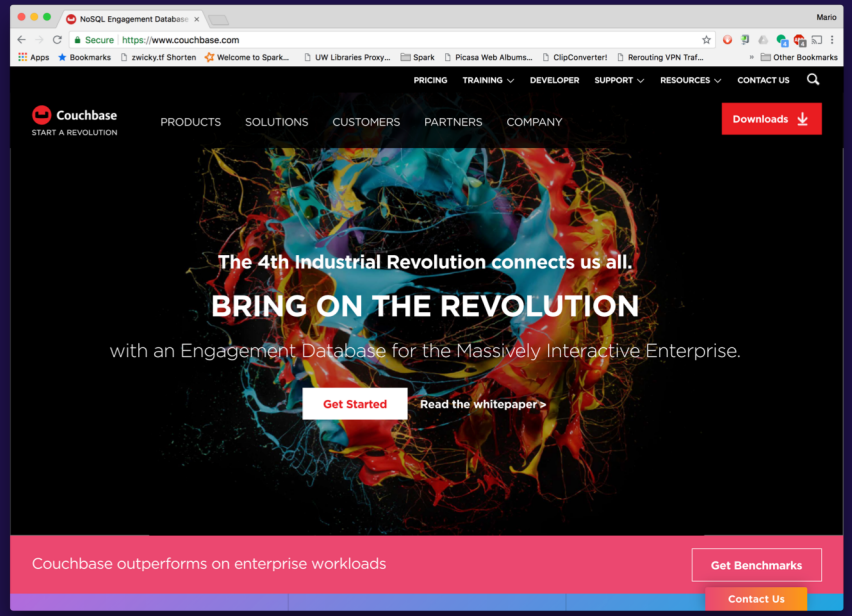
- Since ~June, we've started loading all received alerts (partnership ***and MSIP***) into a MariaDB database.
  - We parse the received alerts, dump CSVs, ingest into MariaDB
  - We also parse and ingest the upper limits
  - We build an "Object" (summary) table: object properties computed from the timeseries
- We keep the database on fast NVMe drive array
  - MariaDB is not the fastest of databases
  - SSDs alleviate that (10GB/s throughput and 2.5M IOPS, observed)

Tables

- `alerts (20,927,972 rows for 9,573,176 distinct objects)`
  - Most of the data from the alerts (99 columns)
- `alerts_limmag (280,731,486 rows)`
  - Magnitude limit for the given JD
- `summary (1,685,092 rows)`
  - Keyed on objectId, contains objects with >= 2 observations
  - Number of observations, mean magnitudes, mean RB scores, mean s/g classification, … (43 columns)
  - More coming

Couchbase thumbnail database

- Image thumbnails stored in Couchbase
- Couchbase:
  - *“Couchbase Server, originally known as Membase, is an open-source, distributed multi-model NoSQL document-oriented database software package that is optimized for interactive applications”*
  - … in other words, it can quickly store/retrieve millions & billions of small “files”.
- It allows us to quickly grab an image, given the candidate ID.



MariaDB + Couchbase + JupyterHub = Easy To Use

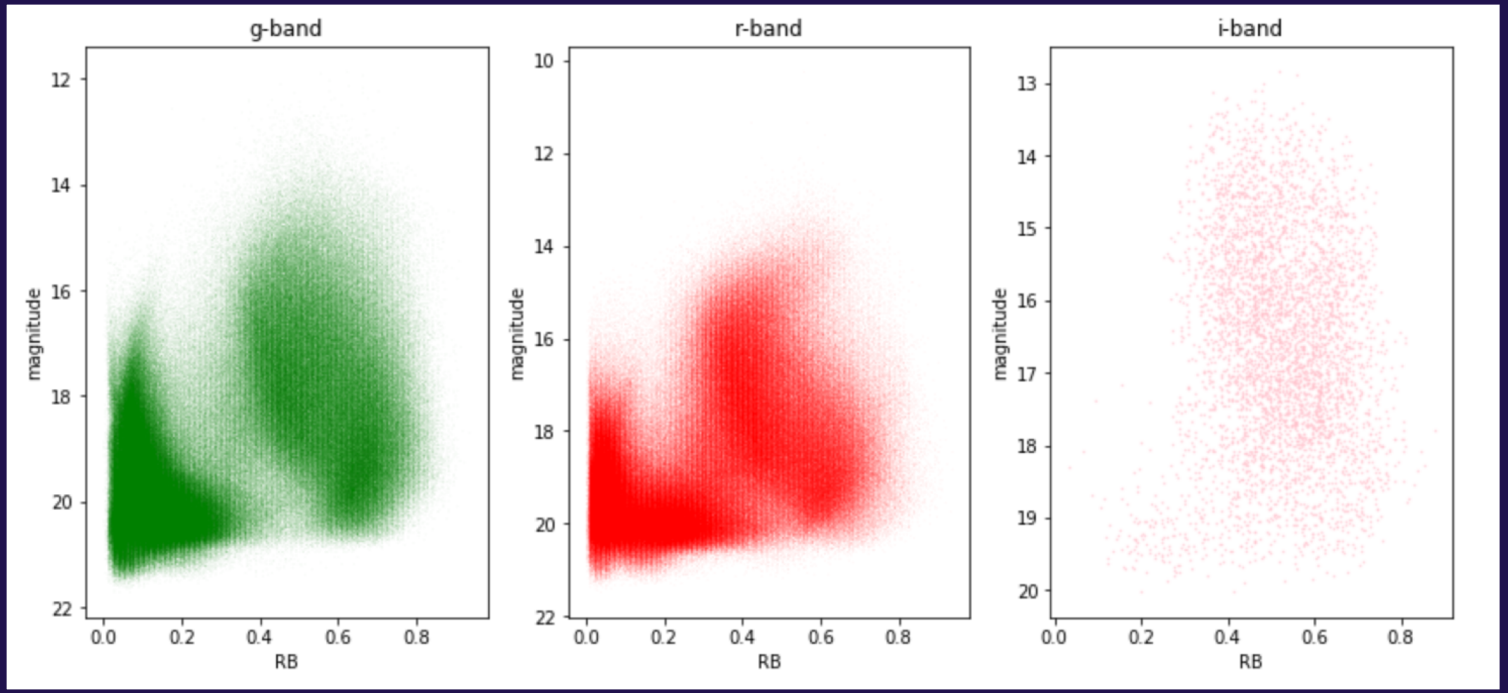
- Both MariaDB and Couchbase have strong Python APIs
- We can do the usual analysis in notebooks

Connecting to the `ztf` database.

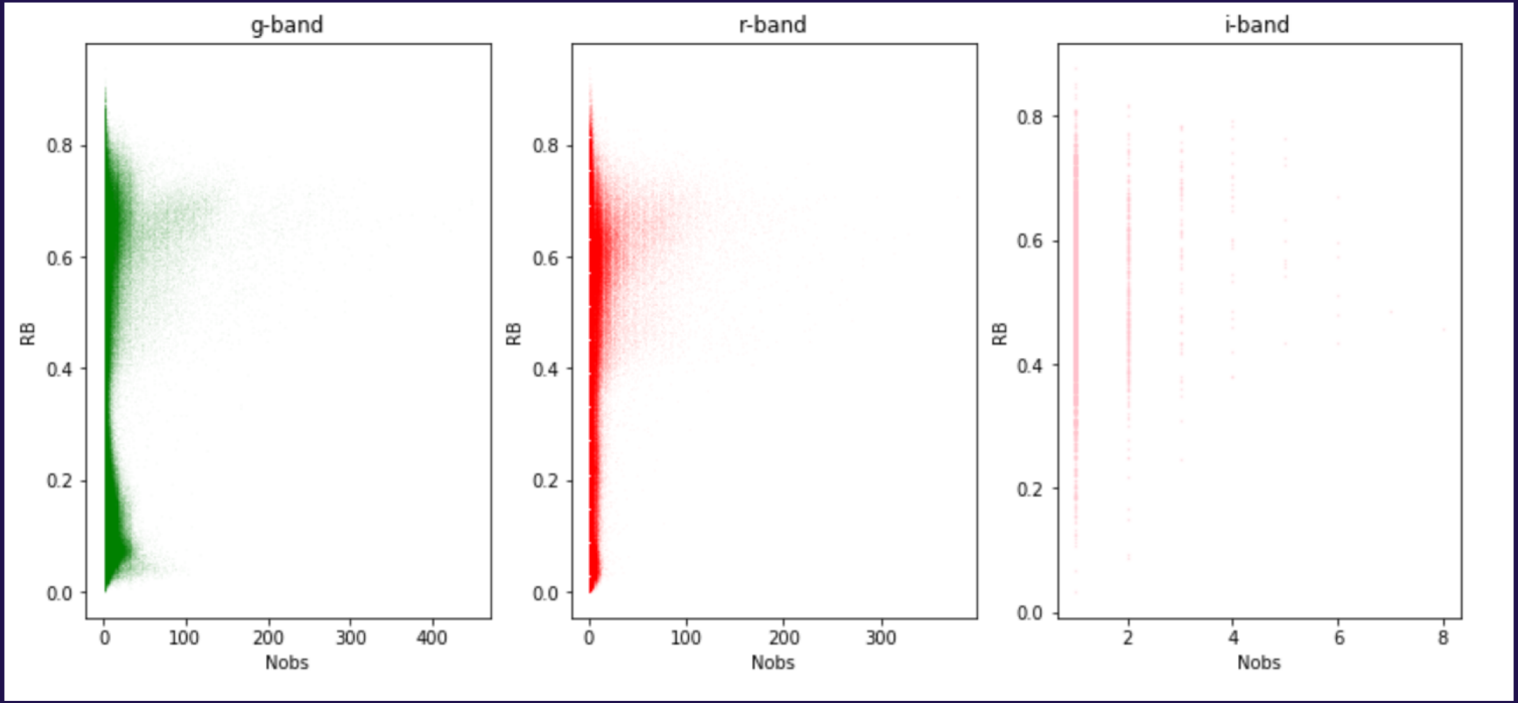
In [3]:
```
con = mariadb.connect(user='ztf', database='ztf', unix_socket='/var/run/mysqld/mysqld.sock')
cur = con.cursor()
```

In [19]:
```
# open connection and bucket for ZTF images
try:
    cluster = Cluster('couchbase://localhost')
    init = cluster.authenticate(PasswordAuthenticator('genesis', '32gigapix!'))
    bucket = cluster.open_bucket("ZTF-images")
except:
    print ("Database connection failed")
```

Examples – QA: mag vs. RB score (averages)

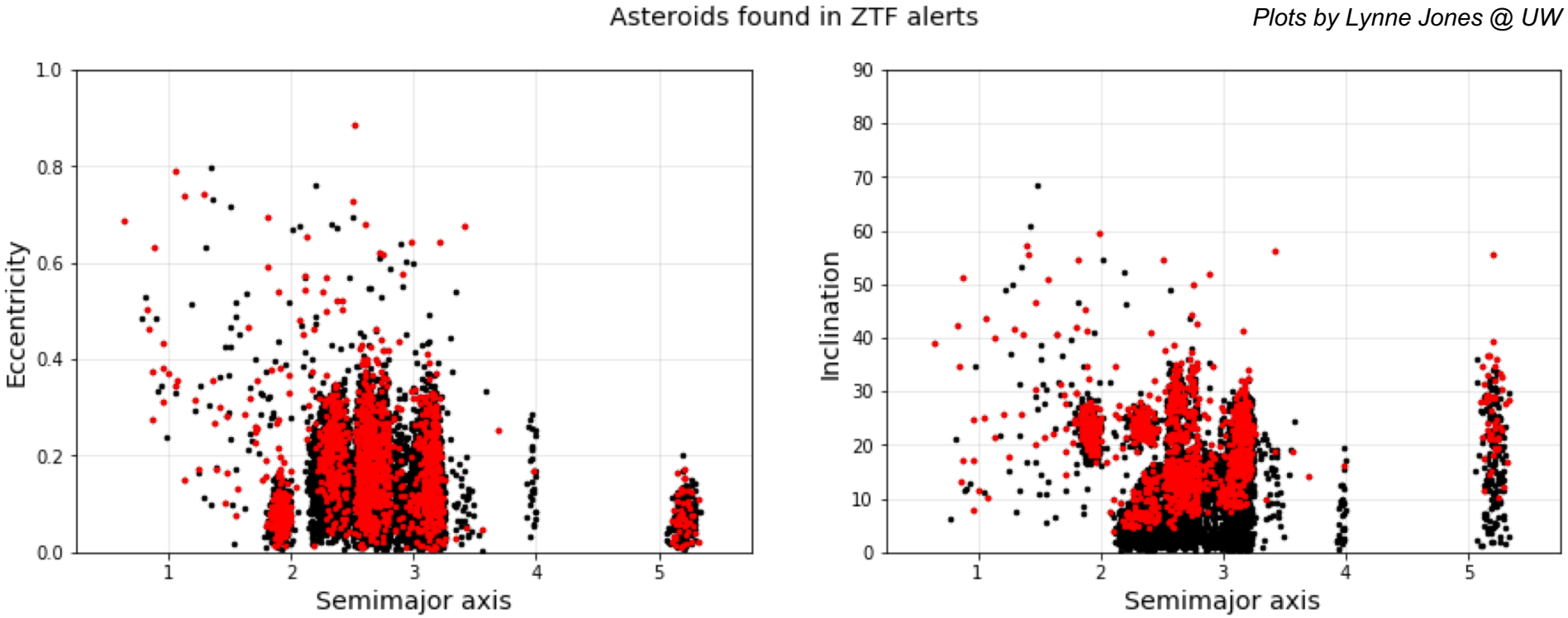


Examples – QA: Avg. RB vs nobs



Example – Science: Asteroids in ZTF Data

Plots by Lynne Jones @ UW

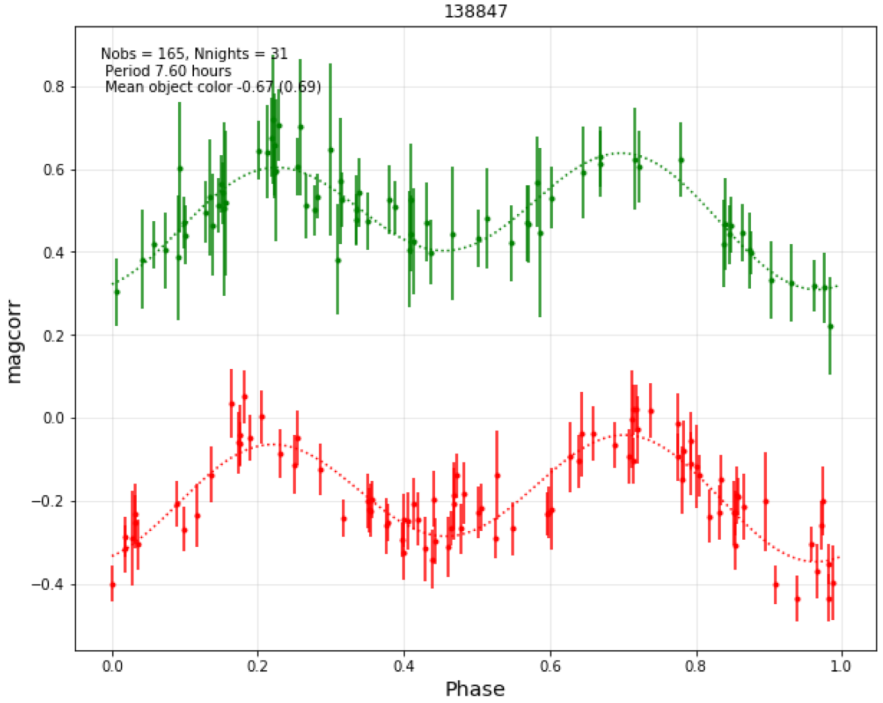


Example – Science: Solar System Ligth Curves

Alert photometry (post photometry update) for a well-sampled asteroid – 138847

This is an NEO with a diameter of ~1km, orbit: a = 1.62 AU, e = 0.29, inc = 22.18 deg

Fit period is 7.6 hours, matching previously reported value.



Plots by Lynne Jones @ UW

Experimental: Partnership Access

- http://ztf.uw.edu/jupyter
- You have to be a member of the `ZwickyTransientFacility` github organization
- Gives access to:
  - The MySQL time-series database
  - The Couchbase thumbnails database
  - The public/partnership alert tarballs

Technology

- Hardware: runs on DIRAC's "home planet" machine, epyc.astro.washington.edu (allocated 16 cores and 128 GB of RAM)
- JupyterHub with OAuth github connector to enable remote notebook access
- Anaconda Python Data Science Distribution
- (Mostly) containerized deployment
  - https://github.com/mjuric/ztf-jupyterhub
  - The aim is to have a fully containerized solution everyone in the partnership can deploy for their own groups

Caveats

- All these services accessible only through JupyterHub, not yet remotely
  - Need to sort out the security issues first
- Planning to load all alerts from the beginning of the survey
  - Also, rebuild the `summary` table with more useful quantities
- MariaDB is not the optimal database solution in terms of performance or scalability
  - Column stores will be better (see next set of slides)
  - We've noticed the Python bindings (serialization/deserialization) are now the bottleneck!
- Think of this service as a "demo verson" – if you find it useful, *talk to us* about getting an account on the full epyc machine.
  - Note: we'll have a ~two week period when we'll have to turn it off due to machine (GPU) being serviced ☹.

Apache Spark™ - Unified Anal

Mario

Secure https://spark.apache.org

APACHE Spark™ *Lightning-fast unified analytics engine*

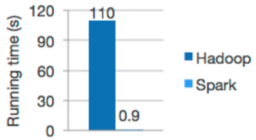
Download | Libraries | Documentation | Examples | Community | Developers | Apache Software Foundation

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Latest News

Spark+AI Summit (October 2-4th, 2018, London) agenda posted (Jul 24, 2018)

Spark 2.2.2 released (Jul 02, 2018)

Spark 2.1.3 released (Jun 29, 2018)

Spark 2.3.1 released (Jun 08, 2018)

Archive

APACHECON North America
September 24-27, 2018
Montréal, Canada

Download Spark

Built-in Libraries:

SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Third-Party Projects

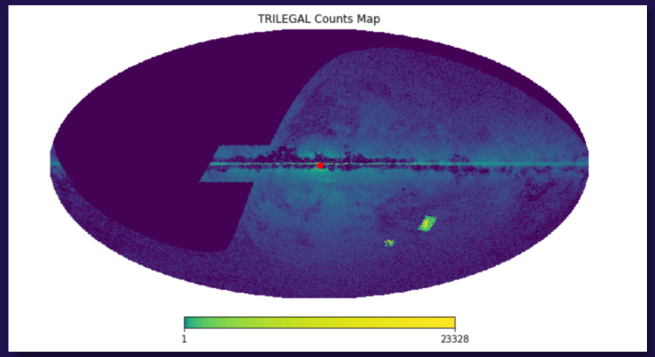
Spark: Time Series Database @ UW

...i.e., Version 2

Petar Zecevic (University of Zagreb and DIRAC Visiting Fellow) and Colin Slater (UW)

Coming up: Apache eXtensions for Spark (AXS)

- For those with iPTF/PS1 background: this is “LSD 2.0”
- Scales to multi-TB datasets (tens of billions of rows)
- (Will) support:
  - Parallel / distributed operation
  - Cross-matching of an arbitrary number of catalogs
  - Spatial selection
  - Easy deployment (`conda install spark-axs`)
  - Python APIs
- Planning to load the match files later this summer/fall.



Above: counts in a bin of a simulated LSST dataset (20 billion objects).

Building NSIDE=1024 CMDs over a 20bn object dataset took 20 minutes on a single machine (IO bound).

Discussion

Contact: mjuric@astro.washington.edu
@mjuric on ZTF Slack