The Nearby Supernova Factory: Efforts in Optimization and Visualization

Kara Ponder
kponder@berkeley.edu
Computational Data Science Fellow
UC Berkeley / BCCP / LBNL

2018/06/21
Data Visualization and Exploration in the LSST Era
Outline

❖ Brief Overview of Supernova Cosmology
❖ Overview of the Nearby Supernova Factory
❖ Historical efforts
  ❖ Machine Learning: Transient Identification
  ❖ Situation Awareness: Sunfall
  ❖ Cloud Computing: SNfactory Pipeline on AWS
❖ SNfactory Zoo
Type Ia Supernova Cosmology

Perlmutter, Physics Today, 2003
The Nearby Supernova Factory

- Gather ~300 SNe Ia in the Nearby Smooth Hubble Flow: 0.03 < z < 0.08
  - Make up the lever arm to anchor higher redshift SNe Ia
- Spectrophotometric time series every 3-5 days from -15 days to +45 days
  - SuperNova Integral Field Spectrograph (SNIFS) mounted on the University of Hawaii 2.2m telescope on Mauna Kea
- Supernova search utilized imaging by the Near Earth Asteroid Tracking (NEAT) program from 2004-2007
  - 8.5 deg2 per exposure from QUEST-II camera on Palomar Oschin 1.2 m
  - 30,000 images per night = 50 GB of compressed images
  - 600,000 possible candidates per night
- Afterwards, received possible candidates from the Palomar Transient Factory (PTF) from 2009-2013 and La Silla Quest from 2012-2013
Machine Learning: Transient Identification
Problem: Transient Identification

Examples from the Dark Energy Survey (DES) present in:

**Difference images:**

- Actual transients or injected fakes
- Artifacts that may be falsely identified as SN
Methods for Identification

- **Threshold cuts:** Used in SDSS
  - S/N, FWHM, astrometric positions, multi-filter, multi-night, etc...
- **Decision Trees**
  - Generalization of thresholds
  - **Boosted Trees (AdaBoost):** Gives more weights to objects that are harder to classify
- **Random Forests:** Uses random subset of features
- **Support Vector Machines**
  - Maps features to higher dimensional space in a non-linear fashion to optimize distance between classes
  - **C-SVM:** Better for noisy data as it allows some objects to be across the decision border

Results

- **Boosted Trees** perform the best with Random forests performing similarly
- SVM does better than original but much worse than the DTs.
- Inability to deal with events close to decision plane such as dim SNe in bright galaxies confuse for statistical fluctuation or subtraction error

SNfactory chose Boosted Trees which resulted in 10 times fewer false positives

Conclusion: Transient Identification

- Boosted Trees, Random Forests, and SVM are better methods of transient classification than simple threshold cuts
- Efficiency increased and workload decreased
- Goldstein et al (2015) used this same concept for the Dark Energy Survey
  - Found that Random Forests worked the best and reduced human time by a factor of 13.8 with 1% failure
  - Future methods may include active learning
- These models are crucial for surveys like LSST which will have more data than people could visually scan
Situation Awareness: Sunfall
Sunfall: Supernova Factory Assembly Line

- **Search**
  - Image Processing and subtraction w/ML and Fourier contour descriptors

- **Workflow Status Monitor**
  - Web-based monitor to facilitate collaboration and situation awareness

- **Data Forklift**
  - Middleware mechanism to automate data transfers

- **Supernova Warehouse**
  - data management
  - workflow visualization
  - collaborative science analysis tool

---

**Citations:**

SNwarehouse: Overview Page

- Written in Java with underlying relational database
- Improve visualizing observing plan through airmass plot:
  - Hour Angle versus Declination Graph
    - Green lines: Lines of constant airmass
    - Green dots: different supernova at specific time
- List all candidates for the night with summary information

Click!
SNwarehouse: Vetting

- Quickly check subtractions to see if brightness is evolving like a normal SN Ia
- Helps to determine if an object should continue to be observed
SNwarehouse: Post Mortem

- Check quick extractions to see first pass of data reductions and slice of photometry
- Quickly shows the accuracy and Signal to Noise of the observations
Conclusion: Situation Awareness

- Sunfall gathers all necessary information into one place so that you have better situation awareness
- Improves candidate identification
  - Sunfall reduced the time to determine a candidate from 2 minutes to 30 seconds!
- Helps scheduling through linear visualizations of airmass and ease of checking previous observations
- Makes it easier for collaborators in different parts of the world to work together
Utilizing the Pipeline via Amazon Web Services Cloud Computing
Overview of Pipeline

- Raw SNIFS data transferred to IN2P3 Computing Center (CC-IN2P3) in Lyon, France
- Data processing runs at the CC-IN2P3
  - Operating system and software architecture is rolled over every ~18 months
- Designed to work on traditional High Performance Computing clusters
- Go from 2D CCD coordinates to 3D flux calibrated and host galaxy subtracted data cubes \((x, y, \lambda)\)
Amazon Elastic Compute Cloud (EC2)

- Ability to select any flavor of Linux
- Able to choose 32 bit over 64 bits (due to legacy code)
- Install any choice of Linux binary packages and third party packages
- Immunity to forced OS and architecture changes
- Storage resource through
  - Elastic Block Storage (EBS) and/or
  - Simple Storage Service (S3)
- 1 word summary: CONTROL!

Reference for this section: Jackson et al (2011), J. Scientific Programming, 19, 107
Results and Conclusions

❖ Explored different storage uses (EBS versus S3) by evaluating the amount of time used and the cost.

❖ The best results for a SNfactory like pipeline was
  ❖ Store raw inputs and codes in EBS volume
  ❖ Send to virtual cluster with 80 workers using 32-bit high CPU medium instances plus a head node
  ❖ Send the outputs to Amazon S3 once all outputs are created
    ❖ Takes about 3 hours per node with 80 nodes plus a head node
  ❖ Ultimately, we decided this was not good enough for our needs

We are currently exploring containers such as Singularity to be more robust to the OS and software changes being implemented at the CC-IN2P3
SNfactory Zoo
Attempt to live Demo!

http://snf.in2p3.fr/zoo/SNF-0203-NEWYORKa
http://snf.in2p3.fr/zoo/SNF-0203-ALLEG2a
http://snf.in2p3.fr/lc/SNF-0203-ALLEG2a/SN2012cu/

If unsuccessful, continue.
Attempt to live Demo!


If unsuccessful, continue.
Flux Calibrated Supernovae
Including non-ia and Standard stars
Flux calibrated and host galaxy subtracted SNe Ia
Supernova Example: SN2012cu

- Options to toggle on and off different information about the SN or calibration information
  - Light curve, MFR (calibration information), Spectra, Extinction, 2D images
- Options to look at different productions
  - Light curves for SNf UBVRI with SALT2 fits
  - Late time observations past SALT2 fits
- Nights with observations, but did not make it through the pipeline for some reason
  - Clicking on the points with give you the error
- Table with summary statistics

Basic information
SALT2 fit results
nMAD
Selection Criteria
Host Gal Reference Images
Lets you know which stars were used to calibrate this night in both channels.
Multi Filter Ratios
Designed to calibrate non-photometric nights by relating them to photometric nights

Explicitly states issue
Extinction as a function of wavelength for 3 main contributors+ 1 total curve
The Future of the Zoo:
Support Tool for SNfactory Publications & Data Releases

- Originally designed for internal assessment of different productions and per-SN results, but we have decided that it will be useful for others interacting with our data
- Before releasing the data, upgrades need to be made
  - Adding spectral time series plot and possibly movies for well sampled SNe
  - Improving user interface via search bars, easy ways to download specific data the user is interested in, etc…
  - I’m also interested in ways to speed up the interface. Currently slow as it creates some of the plots and downloads the data as you request the information
  - Open to other suggestions!
❖ SNeIa continue to be important in isolating properties of Dark Energy.
❖ Machine learning, especially decision trees, improve efficiency and ease of dealing with large sets of transient data.
❖ Isolating the information needed to make decisions improves observations and scheduling when dealing with large data streams.
❖ Cloud computing is feasible for a SNfactory-sized data set.
❖ We are actively working towards a data release and updating the Zoo to accompany it.