Need for classification

Astro datasets getting larger (TB -> PB -> ...)
SDSS/CRTS/PTF/.../LSST/SKA/LIGO
Transient science (multi-epoch surveys)
Spectroscopy is a bottleneck
Early characterization and classification is a must
Separating ordinary and known from unknown and interesting
Given the data volumes, it should be automated
To understand transients, the variables need to be understood too.
Mathematics and Statistics
abstractions and summaries

Computer Science
Efficient algorithms and optimization

Machine Learning

Data Science
galaxy proximity
Galactic latitude etc.

Domain Specific Knowledge
Automated Classification Techniques

• Implementation of clustering algorithms in a machine-learning (ML) or AI setting
  – Examples: star-galaxy separation, automated galaxy morphology classification, stellar or galaxy spectral types, etc.

• *Supervised classifiers*: a set of learning examples is provided; the number of possible classes is known
  – Examples: SVMs, Decision Trees, ...

• *Unsupervised classifiers*: the program decides how many classes are needed to account for the diversity of the data, and classifies on the basis of the data
Variety of available tools

- **Python**
  - PyML
  - scikit-learn
  - astropy/astroML

- **R**
  - [http://cran.r-project.org/web/views/MachineLearning.html](http://cran.r-project.org/web/views/MachineLearning.html)

- **Matlab**
From Python’s scikit-learn
Transient classification

• Characteristic properties
  – proximity to a galaxy
  – Galactic latitude
  – proximity to a radio source

• Lightcurve based quantities
  – amplitude
  – skew
  – Stetson J

Quantify these
make “priors” out of them
Simple(r) classification problem

Stars and galaxies
Enter clustering

• Determine the number of classes
  – Stars
  – Galaxies
Possible complications

- Star - galaxy
- Galaxy - galaxy (E, S0, S, Ir)
- Quasar - star
- Dwarfs - main sequence
Enter clustering

• Determine the number of classes
• Understand their properties
  – Extendedness
  – Light concentration
Measure parameters that are handles for these properties

– Pixels occupied

– Ratio of flux in two apertures
Enter clustering

• Determine the number of classes
• Understand their properties
• Measure parameters that are handles for these properties
• Plot the parameters
• “Separate” the clusters
• Classification is an integral part of A’nomy
• Clustering is the means to separate the classes (in an unsupervised manner)
Simple classification problem
Complications: just stars and galaxies?

- Stars
- Galaxies
- CCD defects
- Cosmic rays
- Bleed trails
- Satellite trails
- Asteroids!

e.g. real-bogus or CRTS’ NN for artifact removal
Complications

• How many classes are there?

• Are they cleanly separated?
  – Brighter stars
  – Distant galaxies
  – Grazing cosmic rays
Complications

How many classes are there?
Are they cleanly separated?
Do all objects belong to these classes?

A Generic Machine-Assisted Discovery Problem: Data Mapping and a Search for Outliers
Complications

• How many classes are there?
• Are they cleanly separated?
• Do all objects belong to these classes?
• Could we add observables to classify better and find rarer objects?
  – Another waveband?
  – A third one?
  – More epochs?
Typical Parameter Space for S/G Classif.

(From DPOSS)
Automated Star-Galaxy Classification: Decision Trees (DTs)

**Fig. 1.** In this sample decision tree, one starts at the top node (root), following the appropriate path to a final leaf (class) based upon the truth of the assertion at each node.

**Fig. 2.** A portion of a much larger actual decision tree generated by the O-Btree algorithm for performing star/galaxy classification. The interval appearing above each node indicates the range in value of the attribute specified in the node above that an object must meet for it to pass along that branch. The dark branches lead to actual classifications. The number in parentheses within each leaf indicates the number of training examples classified correctly at that node.

(Weir et al. 1995)

Ashish Mahabal
An Example: Classification of DPOSS Sources with AutoClass (an unsupervised Bayesian classifier)

Class 1: stellar (PSF)

Class 2: star with a fuzz

Class 3: early-type galaxy

Class 4: late-type galaxy
• Classification is an integral part of A’nymy
• Clustering is the means to separate the classes
• Outliers are the interesting rarer objects which do not belong to the main classes
To understand transients, the variables need to be understood too.
Broad, incomplete hierarchy

- All transients
  - SN
    - SN I
    - SN II
  - CV, blazars, periodic
    - CV, blazars
    - periodic
    - To other classifiers
Light-curve features

• Measure features (metrics) for all light curves
amplitude and std-dev for six classes of variables from CRTS
Separation is better understood when shown as density.
Many features
- not all are independent

Adam Miller
A Variety of parameters that can be used

- Discovery: magnitudes, delta-magnitudes
- Contextual:
  - distance to nearest star
  - Magnitude of the star
  - color of that star
  - normalized distance to nearest galaxy
  - Distance to nearest radio source
  - Flux of nearest radio source
  - Galactic latitude
- Follow-up
  - Colors (g-r, r-I, i-z etc.)
- Prior classifications (event type)
- Characteristics from light-curve
  - amplitude
  - Median buffer range percentage
  - Standard deviation
  - Stetson k
  - Flux percentile ratio mid80
  - Prior outburst statistic

Bayesian Networks best to deal with such datasets as they can deal with missing data and the structure can be learnt from the data – at least in principle

Ashish Mahabal
Relative significance of parameters/parameter selection

Parameters from Richards et al.

Linear trend:
\[
\text{sign(linear trend)} \times \log(\text{linear trend} \mid + 1e^{-06})
\]
\[
\text{sign(linear trend)} \times \sqrt{|\text{linear trend}|}
\]

med_buf_range_per:
\[-\log(1 - \text{med_buf_range_per})\]

Kurtosis:
\[\log(3 + \text{kurtosis})\]

Thomas Fuchs
Ciro Donalek
Julian Faraway
Bits we will leave out

Periodicity
  – Kepler – dense light-curves
  – irregular and sparse light-curves (most surveys)
  – best phasing, characteristic time-scales etc.

GPR
  – interpolation
  – regular grid

Mahabal, Moghaddam, Chen
Using 900 non-SNe and 600 SNe

80-90% completeness using just these parameters

\[
\text{prior outburst} = \frac{1}{t_{span}} \cdot \left( \frac{\sum_i w_i (p_i - p_m)^2}{N} \right)^{1/2}
\]

Ashish Mahabal, Ball
Bigger Bayesian Network picture
Various methods

- Support Vector machines (SVM)
- Random Forests
- Decision Trees

(Deep learning for images)
Citizen science classifications as a path to machine learning

Most data never seen by scientists
Pattern matching techniques not mature enough
(and may never be as mature as humans – but large data makes a difference)

- Hanny’s Voorwerp is an excellent example

citizen sky is a path to better understanding (not an end)
Collaborators and credits

CRTS: Djorgovski, Drake, Donalek, Graham

ML/Stats: Mike Turmon, Thomas Fuchs, Julian Faraway

Students/other: Arun Kumar, Alex ball, Yutong Chen, Tejas Kale

Plus many others ...