The Weirdest Sources in the CSC
A Machine Learning Approach

Rafael Martínez-Galarza
with Megan Freeman, Andreas Zezas, Pepi Fabbiano,
Raffaele D’Abrusco, Doug Burke, Francesca Civano

Chandra 20 Symposium - December 6, 2019
CSC 2.0 includes measured properties for:

* 317,167 unique compact and extended X-ray sources in the sky.
* 928,280 individual observation detections
* 7,287 stacks made from *
* 10,382 Chandra ACIS and HRC-I imaging observations released publicly through the end of 2014.
CSC2: Detecting the faintest high-energy sources

Chandra Deep Field South in CSC 2.0 — stack of 81 observations, 5.8 Ms total exposure

Source detection in CSC 2.0 is performed on stacked (co-added) observations to enhance source detectability.

Visit: http://cxc.harvard.edu/csc2/
The question-driven approach

Well defined, pre-established research questions, e.g.:

• What is the nature of gravitational wave sources progenitors?
• What is the value of $\Omega_0$ to a precision below 1% percent?
• How common are terrestrial planets around main-sequence stars?

Known unknowns. Based on present knowledge

Design targeted observations/surveys. Model specific objects. TAC panels are tuned for this.
The exploration approach

Gain the capability to pose new questions, to formulate new hypotheses.

- This is a way to deal with the unknown unknowns
- Particularly pressing in astronomy, a science which is mostly exploratory
- We need to expand our discovery space.

The golden age of machine learning in astronomy combined with exquisite datasets like CSC 2.0 provide the perfect opportunity to do so.

Quasi-periodic oscillations in GSN 069 likely due to a tidal disruption event “in pieces.” (Miniutti et al. 2019, Nature). This was a serendipitous discovery.
Making serendipitous discovery systematic - Finding anomalies

"An anomaly is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism"
-- Hawkins (1980)

- We need a metric to measure similarity (or dissimilarity) between them. Not so easy in multidimensional parameter spaces.
- We also need a set of features to characterize each example in the dataset. These features can be defined by the scientist, or found through, e.g., deep learning.
- We then rank all objects in our dataset according to their dissimilarity from other objects.
- **Anomalies are scientifically interesting objects: they are often not explained by current models, and could lead to new discoveries.**
Unsupervised Random Forest

1. Generate synthetic dataset from Baron & Poznanski, 2017

2. Train RF classifier with real+synthetic data

3. Propagate the real data

4. Compute weirdness

Baron & Poznanski, 2017
What CSC 2.0 features matter?

• We performed a feature importance analysis using scikit-learn.

• This is based on what features are used the most to split the trees during the training of the model.

• Features that come at the top are:
  • Hardness ratios
  • Spectral fit parameters
  • Spectral variability
Outliers are sampled from a different distribution
Are we just picking artifacts?

Important: weirdness does not depend on significance

But weirdos are more likely to be spectrally variable

Spectral variability conveniently characterized in CSC2!
Look for `var_inter_hard_prob`, `var_inter_hard_sigma`
Example of a weird object

M 31 in X-rays with Swift - a new transient and an old source gone

Never characterized using Chandra data
Variability and Multi-Wavelength Aspects

The transient seems to have undergone a second outburst a couple of months after the main one.

Figure 5. The light curve of the transient between May 2010 and May 2012, using Chandra and SWIFT observations.
Figure 1. Three color stacked images (red: soft band, green:medium band, blue:hard band) of the transient. The red cross indicates the position of the source as determined in CSC2. The left panel corresponds to the observation made on 2011-05-27, whereas the right panel corresponds to the observation made on 2011-09-01.

Figure 2. Power-law spectral fits to the two Chandra observations in which the transient is observed. The left panel corresponds to the observation made on 2011-05-27, whereas the right panel corresponds to the observation made on 2011-09-01. Note the significant increase in the total flux of the source.
• Chandra peak luminosity: 2E37 erg/s

• Not a lot of star formation expected in the bulge of M31 - High mass binary unlikely.

• This transient is most likely an outburst from a low mass X-ray binary. If so, at the distance to M31 the IR emission might come from the disk.

• If not the disk, then the IR photometry of the source is consistent with the donor being and evolved (AGB?) red star. This is relatively rare.

<table>
<thead>
<tr>
<th>obsid</th>
<th>statistic</th>
<th>( n_H )</th>
<th>( n_H^{err} )</th>
<th>( \Gamma )</th>
<th>( \Gamma^{err} )</th>
<th>ampl</th>
<th>ampl^{err}</th>
</tr>
</thead>
<tbody>
<tr>
<td>12970</td>
<td>5.27584</td>
<td>0.429982</td>
<td>0.280321</td>
<td>2.54135</td>
<td>0.547819</td>
<td>0.000133904</td>
<td>8.13E-05</td>
</tr>
<tr>
<td>14197</td>
<td>2.96245</td>
<td>0.0959689</td>
<td>0.0847944</td>
<td>2.16471</td>
<td>0.3681</td>
<td>9.96E-06</td>
<td>3.04E-06</td>
</tr>
</tbody>
</table>

**Table 1.** Spectral fit parameters for the transient using a single component power-law model.
Example: Flaring cool stars

Flaring events tend to be stronger in non-accreting YSOs. They provide insight about physical processes driving very dynamical events. How do we find them?

Both plasma temperature and hardness ratios indicate a moderately energetic and peculiar flare.

Other similar flaring young stars have been identified in Orion and other SF fields. Several of them are not characterized.
Got weirdos?

• Candidates for containing a stellar-mass black hole in early-type galaxies.

• Pulsating ULXs. - Georgios Vasilopoulos’s told us about these earlier in this meeting.

• Highly obscured quasars and AGNs. Several not reported.

• Young stars with strong Fe line emission at 6.4keV.

• Unknown, rare new types of X-ray sources? Stay tuned!
Take home messages

• Anomaly detection is an excellent way to do “exploration approach” science.

• The Chandra Source Catalog is a fertile ground for discovery. But we are by no means constrained to limit the search to CSC.

• Astronomical outliers (anomalies) represent extreme stages in the evolutionary history of astrophysical objects - they used to be found by chance.

• In the golden age of data mining and machine learning, time allocation committees and funding agencies should be ready to be more welcoming to “exploration approach” proposals.
Thanks for a great conference!
Making serendipitous discoveries systematic

Anomaly detection algorithms offer a natural way to find the weirdest objects in large datasets.

- Unsupervised Random Forest (URF) applied to SDSS spectra: complex velocity structures, extremely strong or rare absorption lines (Baron & Poznanski, 2017).

- Proximity clustering + dimensionality reduction applied to Kepler light curves: cataclysmic variables. (Giles & Walkowicz, 2019).
Unreported X-ray source in the outskirts of Abell 2052
Quasar

Nothing peculiar published about this object. High excess variance (Vagnetti et al. 2016), funny spectral fits, hard spectrum.
High-mass X-ray binary candidate in the SMC
95% confidence position error ellipse
0.71" by 0.71" at 141°

Galactic $n_H$ column density
8.45 \times 10^{20} \text{ cm}^2

Aperture-corrected flux (broad band)
Lower confidence limit
5.995e-14 \text{ erg cm}^2 \text{ s}^{-1}
Upper confidence limit
5.745e-14

Source significance (S/N)
22.51

Hard/Medium band hardness ratio
0.2473
Lower confidence limit
0.2017
Upper confidence limit
0.2929

Medium/Soft band hardness ratio
-0.1611
Lower confidence limit
-0.208
Upper confidence limit
-0.1131

Number of ACIS observations
22

Number of HRC observations
2
Globular cluster in Andromeda

Source is variable (within or between observations).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% confidence position error ellipse</td>
<td>0.71&quot; by 0.71&quot; at 1°</td>
</tr>
<tr>
<td>Galactic $n_H$ column density</td>
<td>$6.77 \times 10^{20}$ cm$^2$</td>
</tr>
<tr>
<td>Aperture-corrected flux (broad band)</td>
<td>$4.159 \times 10^{-13}$ erg cm$^2$ s$^{-1}$</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>4.122e-13</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>4.196e-13</td>
</tr>
<tr>
<td>Source significance (S/N)</td>
<td>129.66</td>
</tr>
<tr>
<td>Hard/Medium band hardness ratio</td>
<td>0.0862</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>0.0768</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.0956</td>
</tr>
<tr>
<td>Medium/Soft band hardness ratio</td>
<td>-0.1012</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>-0.1118</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>-0.0906</td>
</tr>
<tr>
<td>Number of ACIS observations</td>
<td>103</td>
</tr>
<tr>
<td>Number of HRC observations</td>
<td>64</td>
</tr>
</tbody>
</table>
Bremsstrahlung temperature
Blackbody temperature
BH candidate in Andromeda

Source is variable (within or between observations).

95% confidence position error ellipse 0.71" by 0.71" at 21.4°
Galactic n_H column density 6.62 x 10^{20} cm^{-2}
Aperture-corrected flux (broad band) 4.178e-13 erg cm^{-2} s^{-1}
  Lower confidence limit 4.147e-13
  Upper confidence limit 4.207e-13
Source significance (S/N) 127.21
Hard/Medium band hardness ratio 0.0587
  Lower confidence limit 0.0506
  Upper confidence limit 0.0668
Medium/Soft band hardness ratio -0.1137
  Lower confidence limit -0.1218
  Upper confidence limit -0.1056
Number of ACIS observations 105
Number of HRC observations 65
**Young star in Orion**

Source is variable (within or between observations).

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% confidence position error ellipse</td>
<td>0.71&quot; by 0.71&quot; at 69.3°</td>
</tr>
<tr>
<td>Galactic $n_H$ column density</td>
<td>$19.31 \times 10^{20}$ cm²</td>
</tr>
<tr>
<td>Aperture-corrected flux (broad band)</td>
<td>$3.792 \times 10^{-15}$ erg cm⁻² s⁻¹</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>$3.506 \times 10^{-15}$</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>$4.078 \times 10^{-15}$</td>
</tr>
<tr>
<td>Source significance (S/N)</td>
<td>127.21</td>
</tr>
<tr>
<td>Hard/Medium band hardness ratio</td>
<td>0.0287</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>-0.0606</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.1181</td>
</tr>
<tr>
<td>Medium/Soft band hardness ratio</td>
<td>0.3673</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>0.2255</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.5378</td>
</tr>
<tr>
<td>Number of ACIS observations</td>
<td>13</td>
</tr>
<tr>
<td>Number of HRC observations</td>
<td>1</td>
</tr>
</tbody>
</table>
Transient ULX in the vicinity of Centaurus A (similar to M86 tULX-1)

Source is variable (within or between observations).

- **95% confidence position error ellipse**: 0.71" by 0.71"
- **Galactic \( n_H \) column density**: \( 8.4 \times 10^{20} \) cm\(^{-2} \)
- **Aperture-corrected flux (broad band)**: 6.427e-13 erg cm\(^{-2} \) s\(^{-1} \)
  - Lower confidence limit: 6.394e-13
  - Upper confidence limit: 6.458e-13
- **Source significance (S/N)**: 110.61
- **Hard/Medium band hardness ratio**: 0.0175
  - Lower confidence limit: 0.0119
  - Upper confidence limit: 0.0231
- **Medium/Soft band hardness ratio**: -0.0837
  - Lower confidence limit: -0.0893
  - Upper confidence limit: -0.0781
- **Number of ACIS observations**: 22
- **Number of HRC observations**: 2
Seyfert 2 galaxy with hard spectrum and high spectral temperature

- Chandra HIPS

- 95% confidence position error ellipse: 0.71" by 0.71" at 178.3°
- Galactic $n_H$ column density: $3.98 \times 10^{20} \text{ cm}^2$
- Aperture-corrected flux (broad band):
  - Lower confidence limit: 6.894e-13 erg cm$^{-2}$ s$^{-1}$
  - Upper confidence limit: 6.997e-13
- Source significance ($S/N$): 68.24
- Hard/Medium band hardness ratio: 0.9494
  - Lower confidence limit: 0.9463
  - Upper confidence limit: 0.9538
- Medium/Soft band hardness ratio: -0.3548
  - Lower confidence limit: -0.4004
  - Upper confidence limit: -0.3092
- Number of ACIS observations: 5
- Number of HRC observations: 0
Variable young star in Orion with peculiar HRs and high spectral T

Chandra HIPS

Source is variable (within or between observations).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% confidence position error ellipse</td>
<td>0.72&quot; by 0.72&quot; at 28.8°</td>
</tr>
<tr>
<td>Galactic $n_H$ column density</td>
<td>$19.12 \times 10^{20}$ cm$^2$</td>
</tr>
<tr>
<td>Aperture-corrected flux (broad band)</td>
<td>$7.224\text{e-15}$ erg cm$^2$ s$^{-1}$</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>$6.676\text{e-15}$</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>$7.737\text{e-15}$</td>
</tr>
<tr>
<td>Source significance (S/N)</td>
<td>20.93</td>
</tr>
<tr>
<td>Hard/Medium band hardness ratio</td>
<td>0.7071</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>0.6502</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.7626</td>
</tr>
<tr>
<td>Medium/Soft band hardness ratio</td>
<td>-0.6508</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>-0.7177</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>-0.5815</td>
</tr>
<tr>
<td>Number of ACIS observations</td>
<td>12</td>
</tr>
<tr>
<td>Number of HRC observations</td>
<td>1</td>
</tr>
</tbody>
</table>
X-ray source near the nucleus of M51. Weird spectrum

Source is variable (within or between observations).

- **95% confidence position error ellipse**: 0.72" by 0.72" at 28.8°
- **Galactic n_H column density**: $19.12 \times 10^{20}$ cm$^{-2}$
- **Aperture-corrected flux (broad band)**
  - Lower confidence limit: 6.676e-15 erg cm$^{-2}$ s$^{-1}$
  - Upper confidence limit: 7.737e-15 erg cm$^{-2}$ s$^{-1}$
- **Source significance (S/N)**: 20.93
- **Hard/Medium band hardness ratio**: 0.7071
  - Lower confidence limit: 0.6502
  - Upper confidence limit: 0.7626
- **Medium/Soft band hardness ratio**: -0.6508
  - Lower confidence limit: -0.7177
  - Upper confidence limit: -0.5815
- **Number of ACIS observations**: 12
- **Number of HRC observations**: 1
Galaxy in CDFS. Low spectral T, high photon index

Source is variable (within or between observations).

95% confidence position error ellipse: 0.72" by 0.72" at 5.4°
Galactic n_H column density: 0.9 \times 10^{20} \text{ cm}^{-2}
Aperture-corrected flux (broad band):
  Lower confidence limit: 7.057e-16
  Upper confidence limit: 8.502e-16
Source significance (S/N): 16.1
Hard/Medium band hardness ratio: 0.0787
  Lower confidence limit: 0.0006
  Upper confidence limit: 0.1555
Medium/Soft band hardness ratio: -0.1786
  Lower confidence limit: -0.253
  Upper confidence limit: -0.1006
Number of ACIS observations: 83
Number of HRC observations: 0

Please review the current caveats for source properties in CSC 2.0.
Aperture Photometry in CSC2

- Bayesian Model

Sources with overlapping apertures, nearby sources, and background simultaneously.

Joint posterior for source fluxes and background flux (for single observation):

\[ P(s_1 \ldots s_n, b | C_1 \ldots C_n, B) = K \times P(b)P_{\text{Pois}}(B | \phi) \prod_i P(s_i)P_{\text{Pois}}(C_i | \theta_i) \]

\[ \theta_i = E_i \times \left[ \sum_{j=1}^{n} f_{ij} s_j + \Omega_i b \right] \]

\[ \phi = E_b \times \left[ \sum_{i=1}^{n} g_i s_i + \Omega_b b \right] \]

Counts in overlapping regions assigned to brightest source

Master source flux for source \( s_k \) in an \( n \)-source bundle is determined from the Bayesian block for that source with the largest exposure:

\[ P(s_k | \{C^j_i\}, \{B^j\}) \cong P(s_k) \prod_{j=1}^{m} P_{\text{Pois}}(B^j | \hat{\phi}^j) \times P_{\text{Pois}}(C^j_k | \hat{\theta}^j_k) \prod_{i=1, i \neq k}^{n} P_{\text{Pois}}(C^j_i | \hat{\theta}^j_i) \]

Bottom plot to the right: marginalized posteriors

Posterior optimized and sampled using MCMC in Sherpa.

Relevant catalog properties:
- Photflux_aper_x, flux_aper_x, flux_bb_aper_x
- flux_aper90_x, flux_aper90_x

Center for Astrophysics | Harvard & Smithsonian
Aperture Photometry in CSC2

- Spectral Model Fits

For sources with more than 150 counts in the energy band 0.5-7 keV, two spectral model fits are done:

* Absorbed black body (thermal)  \[ f(E) = \exp^{-N_h \sigma E} A \left(\frac{E^2}{\exp^{E/kT}} - 1\right) \]

* Absorbed power law (non-thermal)  \[ f(E) = \exp^{-N_h \sigma E} A E^{-\Gamma} \]

Background-subtracted counts forward fitted using Sherpa ($\chi^2$ stat.), and 1s confidence intervals estimated using Sherpa’s projection method.

Model flux for the energy interval 0.5-7 keV found by integration of the best fit.

Relevant catalog properties: flux_powlaw, flux_bb, flux_brems

Center for Astrophysics | Harvard & Smithsonian
**Hardness Ratios**

- **Single Observation**

For each band pair $x,y$, the hardness ratio ($H_{xy}$) is the flux value for the softer band, subtracted from the harder band, relative to their sum $(x-y)/(x+y)$. Just like $F_x$ and $F_y$ are random variables with associated probabilities, so is $H_{xy}$:

\[
PH_{xy}(H_{xy}) \, dH_{xy} = \int_{F_{xy}=0}^{\infty} P_x \left( \frac{1 + H_{xy}F_{xy}}{2} \right) P_y \left( \frac{1 - H_{xy}F_{xy}}{2} \right) \frac{F_{xy}}{2} \, dH_{xy} \, dF_{xy}
\]

Hardness ratios listed in the catalog (hard_ms, hard_hm, hard_hs) are the values of $H_{xy}$ that maximize the above distribution for the corresponding bands.

- **Multiple observations (Combine all PDFs together)**

Stack: all observations in the stack

Master: all observations in Bayesian block

Relevant catalog properties:

hard_ms, hard_hm, hard_hs
**Variability**

- **Single observation:**
  
  Gregory-Loredo Test: Hypothesis rejection test (i.e., odds ratio of assuming variability vs not assuming it). The resulting `var_prob` gives the probability that events detected are not arriving at a uniform rate. Used to estimate intra-obs variability (pick max `prob` among stack `obsids`).

- **Multiple observations:**
  
  Inter-observation variability. Variability test is based on a likelihood ratio between the null hypothesis of no variability, and the assumption of variability, when several observations are considered. `Var_inter_prob` is a p, value.

Relevant catalog properties:

- `var_prob_x`, `ks_prob_x`, `kp_prob_x`
- `var_inter_prob_x`, `ks_inter_prob_x`, `kp_inter_prob_x`, `var_intra_prob_x`