Jet Propulsion Laboratory
California Institute of Technology

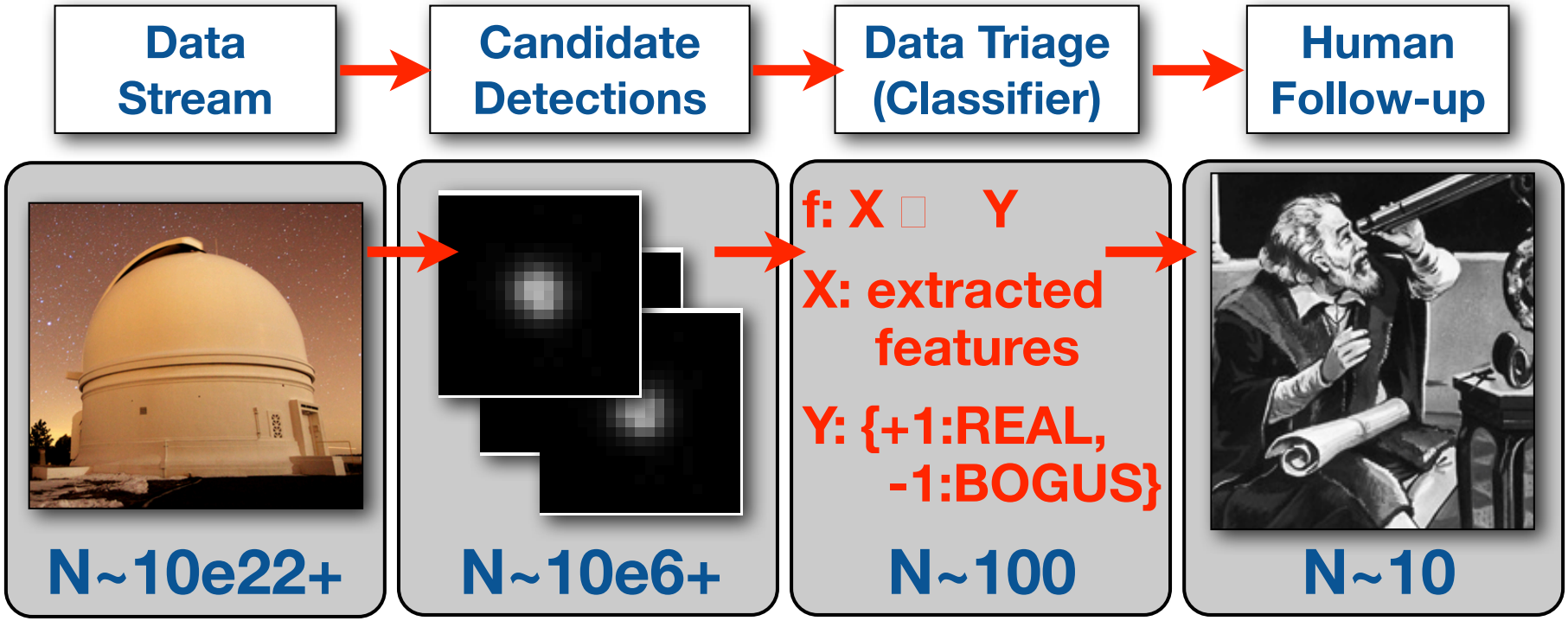
Leveraging Annotated Archival Data with Domain Adaptation to Improve Data Triage in Optical Astronomy

Brian D. Bue, Umaa D. Rebbapragada

Machine Learning and Instrument Autonomy Group
Jet Propulsion Laboratory, California Institute of Technology

Tools for Astronomical Big Data
Tucson, AZ. March 9, 2015

Automated Triage of Astronomical Data



Supervised Learning for Data Triage

Training Data (labeled)

$$(X,Y)^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$
$$\mathbf{x}_i \in \mathbb{R}^n \text{ (features)}$$
$$y_i \in \{\text{real}, \text{bogus}\}$$

Test Data (unlabeled)

$$(X,Y)^{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$$
$$\mathbf{x}_i \in \mathbb{R}^n$$
$$y_i \text{ unknown}$$

Trained Classifier

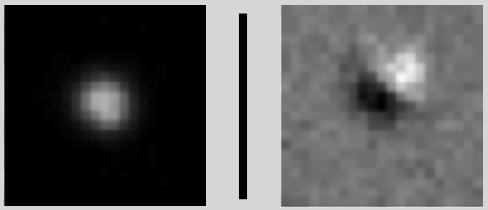
$$f(\mathbf{x}) \rightarrow y$$

Test Predictions

$$\{y_i\}_{i=1}^{M}$$

Follow-up

Real



Bogus

Assumption: $(X, Y)^{\text{train}}$, $(X, Y)^{\text{test}}$ drawn i.i.d. *from the same distribution* $\mathcal{P}(X, Y)$

Violated when measurement conditions change

- Due to: sensor modifications / changes to imaging pipeline

How can we use existing labeled data when such changes occur?

Applications

- Bootstrap detection systems for new surveys using old data
- Provide data continuity for ongoing surveys

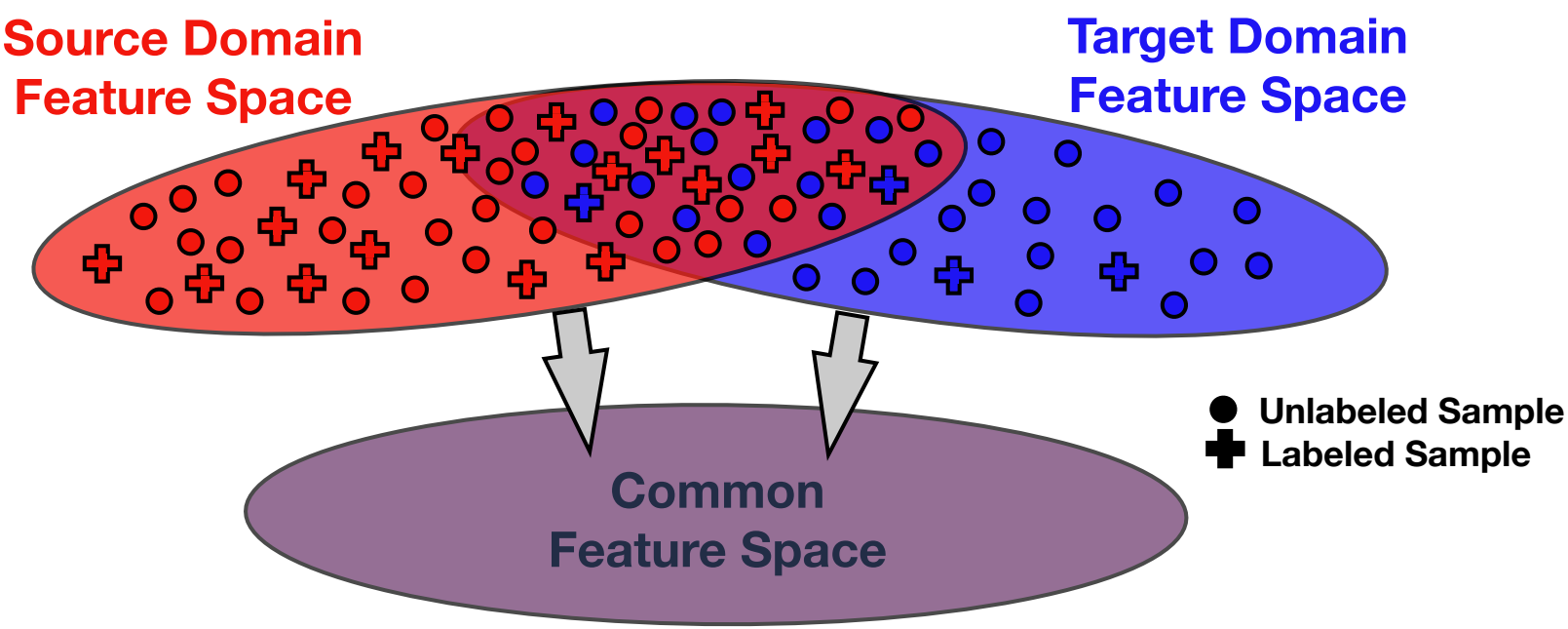
Domain Adaptation

Proposal: use **domain adaptation** to adapt data/classifiers from earlier surveys with similar science goals

Given: labeled **source** (training) samples $\mathcal{D}^S \sim \mathcal{P}^S(X, Y)$

(mostly) unlabeled **target** (test) samples $\mathcal{D}^T \sim \mathcal{P}^T(X, Y)$

Compute: mapping between source and target feature spaces

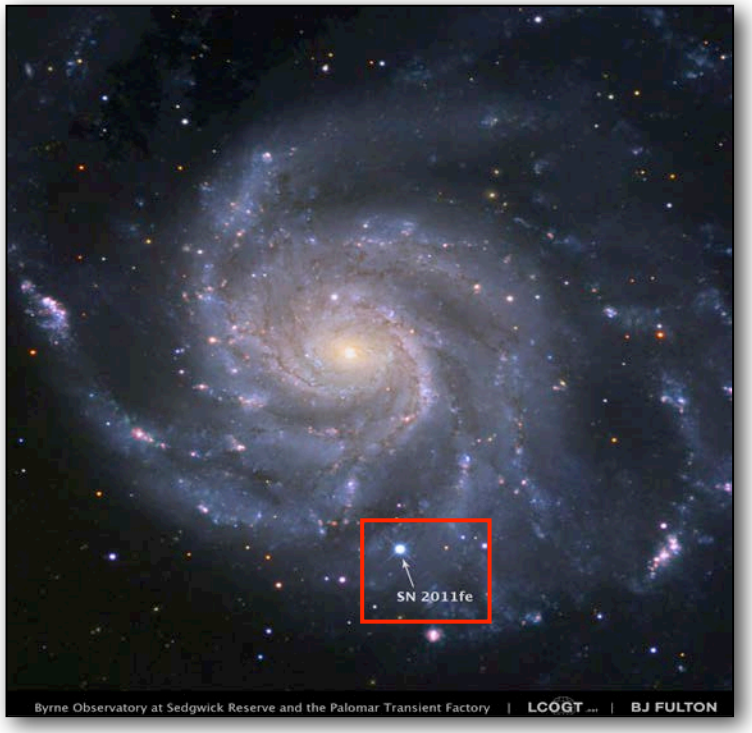


Case study: Transient Detection with the Intermediate Palomar Transient Factory (iPTF)

Fully-automated, wide-field optical transient survey

Focus on supernovae (short duration), variable stars (persistent)

iPTF succeeded Palomar Transient Factory (PTF) in January 2013



Palomar P48 + Large field camera captures 7.8 deg^2 on the sky

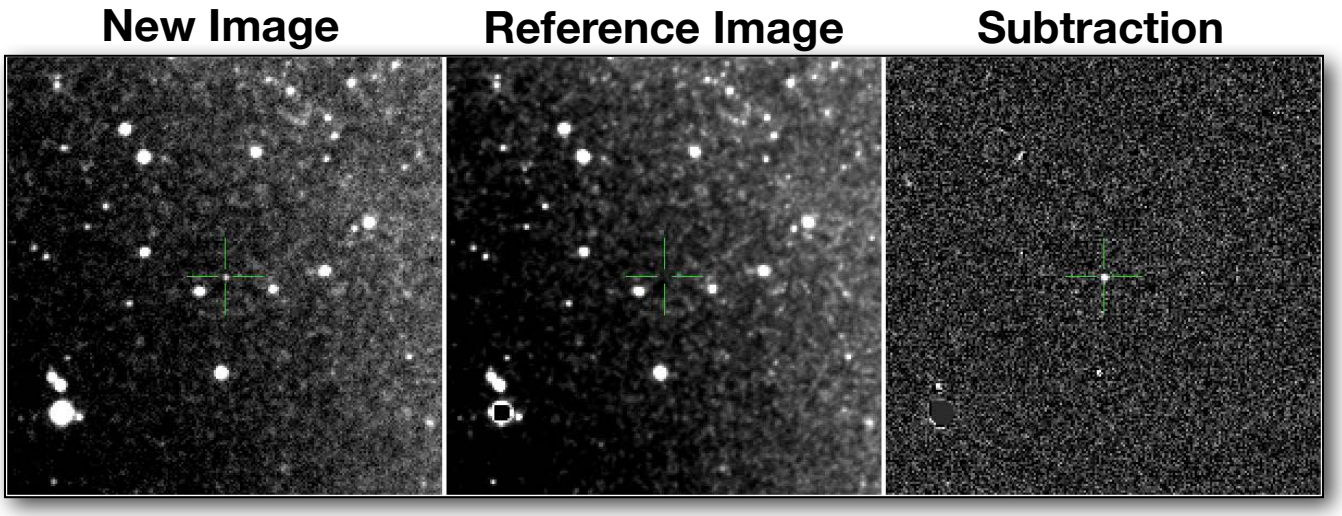
http://ptf.caltech.edu/iptf/

Detects 500K – 1M candidate sources per night

≈0.1% of all point source detections real transients

iPTF Candidate Detection + Classification

Point source detection: via image subtraction pipeline

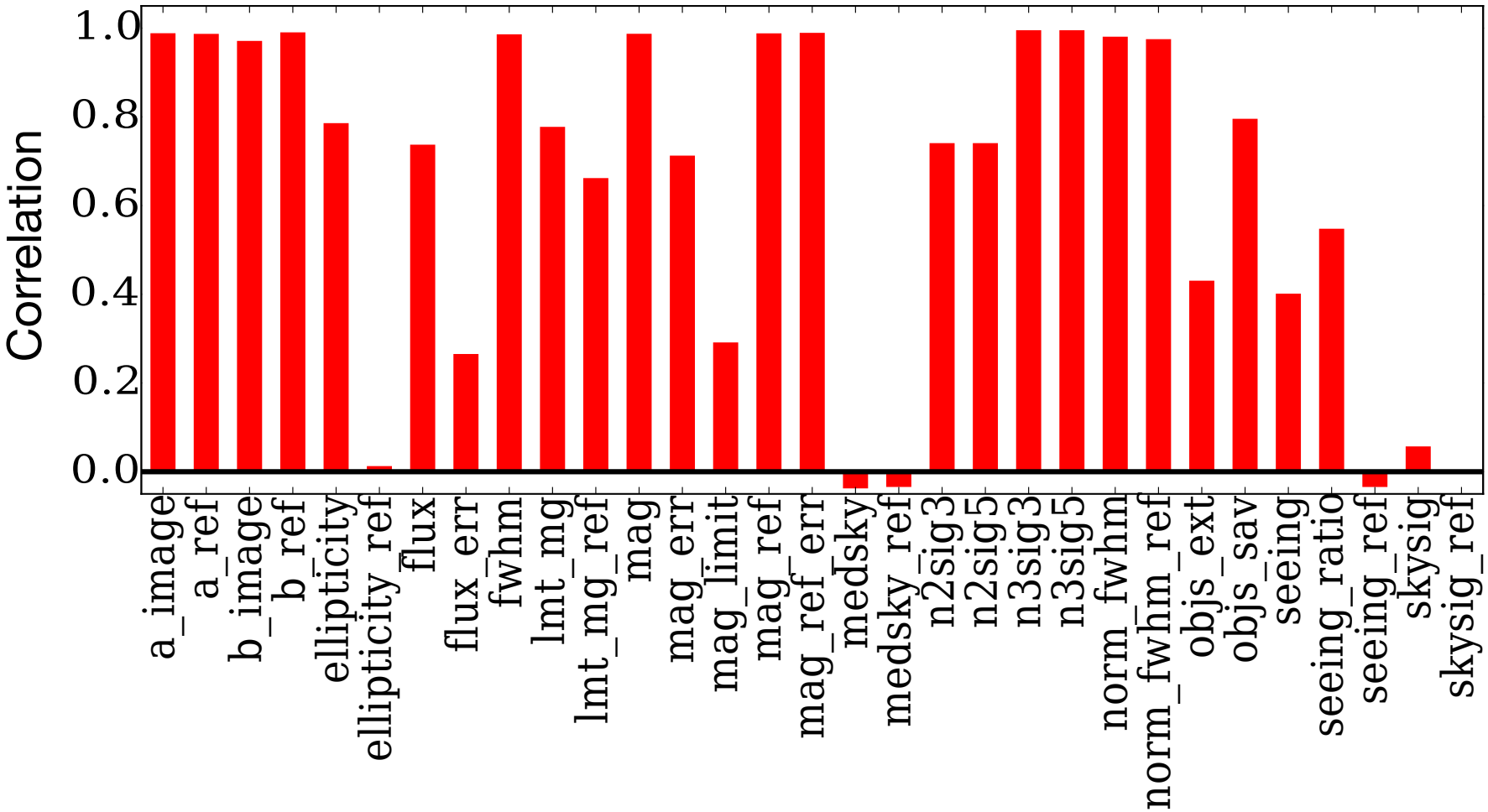


Real-Bogus classifier: (Bloom et al., 2011; Brink et al., 2012)

- Classifies point source detections; prioritizes high-confidence detections for human review/confirmation
- **Trained using PTF data**
  ⇒ iPTF upgrade changed image processing pipeline

  ⇒ **produced suboptimal predictions on new iPTF data**

Feature drift due to iPTF pipeline upgrade



Correlation between PTF and iPTF features for real (spectroscopically-verified) transient sources

Experiment: PTF → iPTF Transient Detection

Goal: adapt PTF-trained classifier to iPTF domain

Proof of concept for future surveys

- PTF/iPTF → Zwicky Transient Facility (ZTF, 2017)

Validation Data

- **PTF:** 37,028 verified reals, 39,613 bogus
- **iPTF:** 18,433 verified reals, 19,000 bogus

Features (31 total)

- candidate: mag, mag_err, flux, flux_err, a_image, b_image, fwhm, mag_ref, mag_ref_err, a_ref, b_ref, n2sig3, n3sig3, n2sig5, n3sig5
- subtraction: objs_extracted, objs_saved, lmt_mg_new, lmt_mg_ref, medsky_new, medsky_ref, seeing_new, seeing_ref, skysig_new, skysig_ref
- misc: mag_from_limit, ellipticity, ellipticity_ref, normalized_fwhm, normalized_fwhm_ref, seeing_ratio

PTF → iPTF Algorithm Comparison

Geodesic Flow Kernel (GFK, Gong et al., CVPR12)

- Computes domain-invariant features from intermediate subspaces "between" source & target domains
- Unsupervised: requires no labeled target data

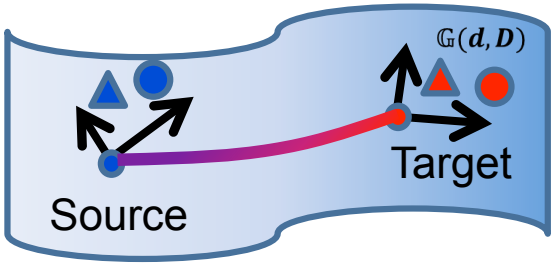


Figure credit: Gong et al., 2012

Co-Training for Domain Adaptation (CODA, Chen et al., NIPS11)

- Partitions feature space, trains a classifier for each partition
- Adds high-confidence predictions on target data to training set
- Learns informative features across domains

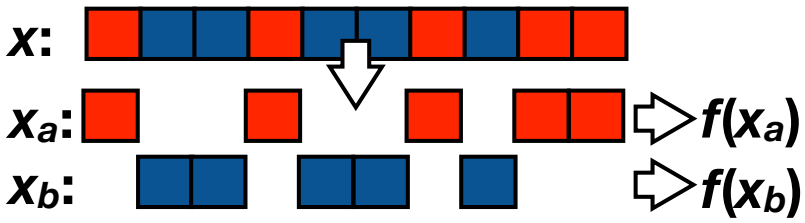
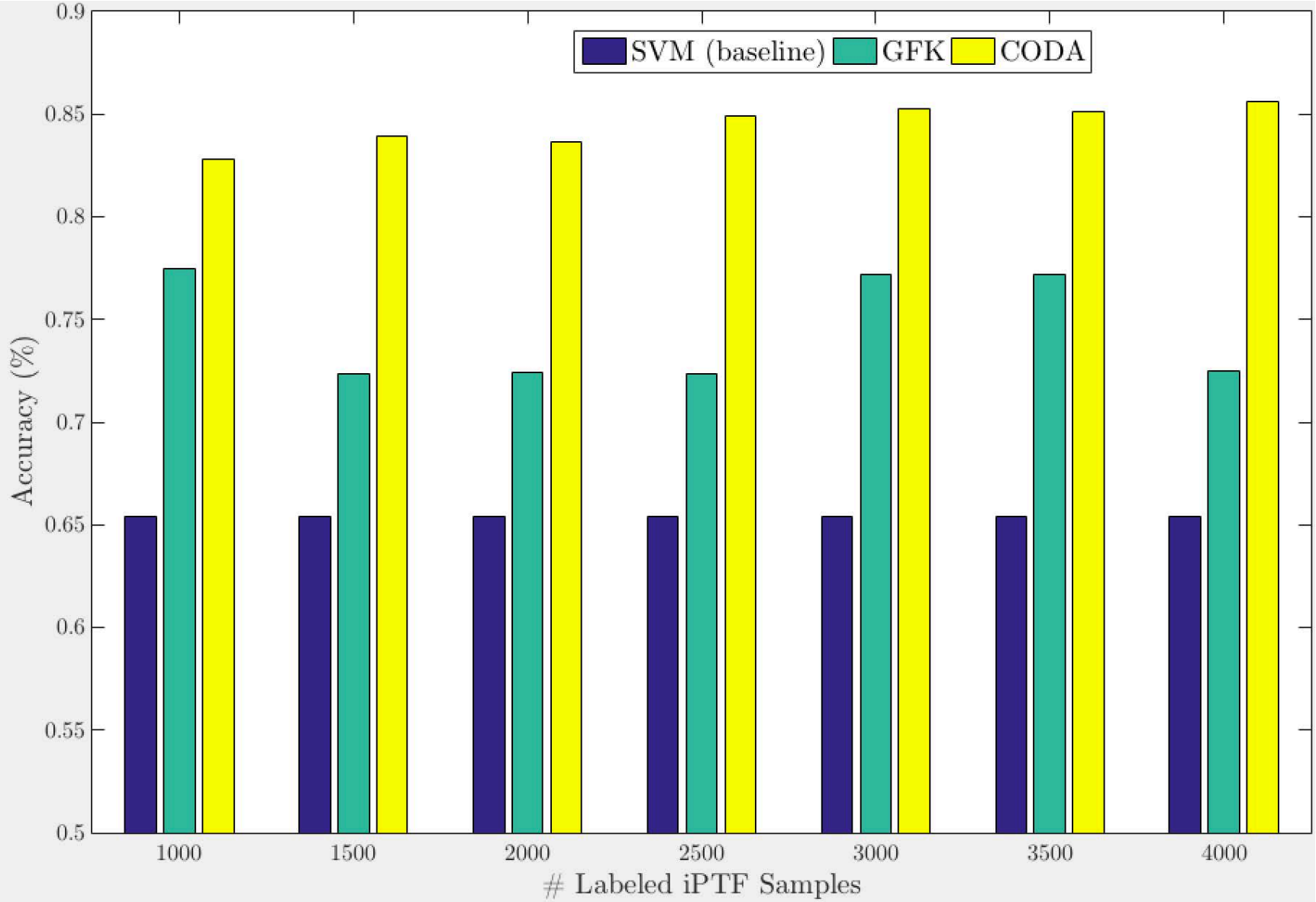


Figure credit: Chen et al., 2011

PTF → iPTF Prediction Accuracy



Bue et al., Domain Adaptation / Data Triage / Optical Astronomy

Summary

- Automated transient detection accuracy suffers when training/test distributions differ
- Domain adaptation can help
  - GFK and CODA show 5-20% improvements in accuracy over baseline

Ongoing/future work

- Diagnose iPTF image processing pipeline artifacts
- Use PTF/iPTF data to bootstrap Real-Bogus classifier for Zwicky Transient Facility

---

bbue@jpl.nasa.gov
http://imbue.jpl.nasa.gov