Toward on-line data mining and event detection in petascale data streams

David R. Thompson
Kiri L. Wagstaff

Jet Propulsion Laboratory
California Institute of Technology
firstname.lastname@jpl.nasa.gov

Research described in this presentation was carried out at the Jet Propulsion Laboratory under a Research and Technology Development Grant. Images courtesy NASA / Caltech JPL / Swinburne University / Curtin University.
Copyright 2012 California Institute of Technology. All Rights Reserved. US Government Support Acknowledged.

Traditional data mining

Online
Automatic
"To spec"

Random-access storage

Source detection and Classification

Statistical inference

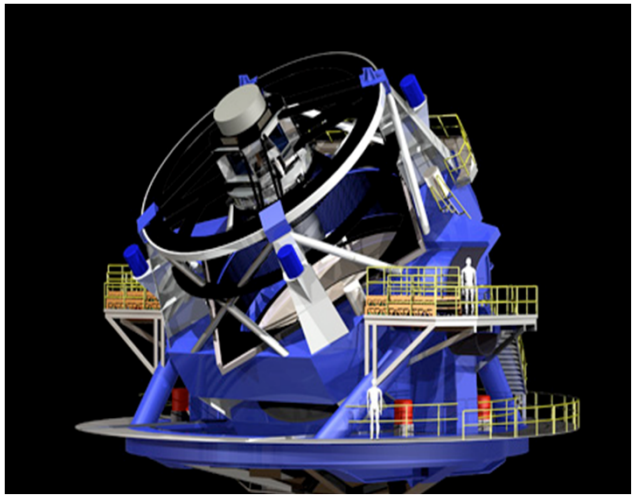
Clustering, anomaly detection

Manifold learning, visualization

Asychronous
Manually-supervised

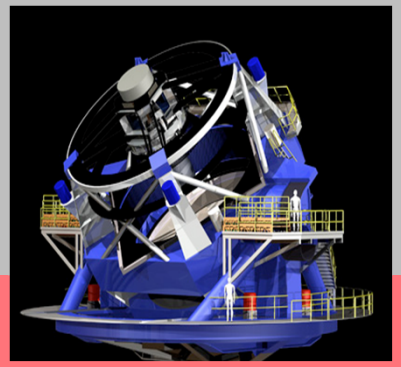
Next-generation instruments

- Cannot store or analyze petascale data volumes.
- Need immediate, real-time followup
- Solution: adaptive processing chains to handle *streaming* data
  - Data mining is continuous
  - Pattern recognition permits immediate triage and followup



Images: SKATelescope.org, LSST.org

Next-generation instruments



Source detection and Classification

Statistical inference

Clustering, anomaly detection

Manifold learning

Random-access storage

Online, Automatic

Examples

Classical Neyman Pearson detection

$$f(x) = \frac{L(x_t\,;\,H_1)}{L(x_t\,;\,H_0)} > \tau$$

Signal model

Background model

Anomaly detection

$$f(x) = \frac{1}{L(x_t\,;\,H_0)} > \tau$$

Background model only

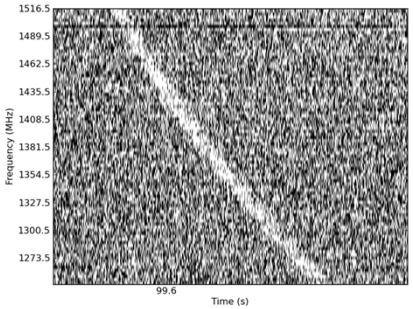
This presentation: two archetypal stream mining scenarios

We have examples of the background
…But don't know what we're looking for

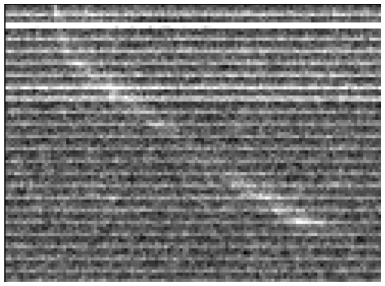
We have examples of desired targets
…but don't know the background noise conditions

Case I: Have examples of background, but don't know targets of interest

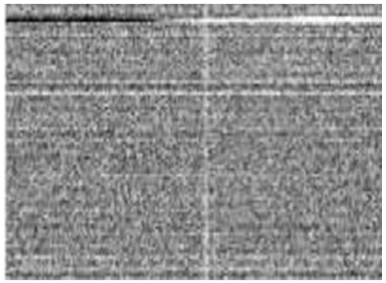
Pulsar pulse



"Peryton"



Radio Freq. Interference



Parkes Multibeam Survey [Edwards et al., 2001]

- 1.4 GHz, 96 channels
- 125 μs sample time

Related methods:

- One-class SVM [Schölkopf et al., 2001]
- Semi-supervised anomaly detection [Blanchard et al., 2010]

Shown: Parkes telescope multibeam receiver, courtesy CSIRO / Swinburne. Other Images courtesy S. Burke-Spolaor

Semi-Supervised Eigenbasis Novelty Detection (SSEND)

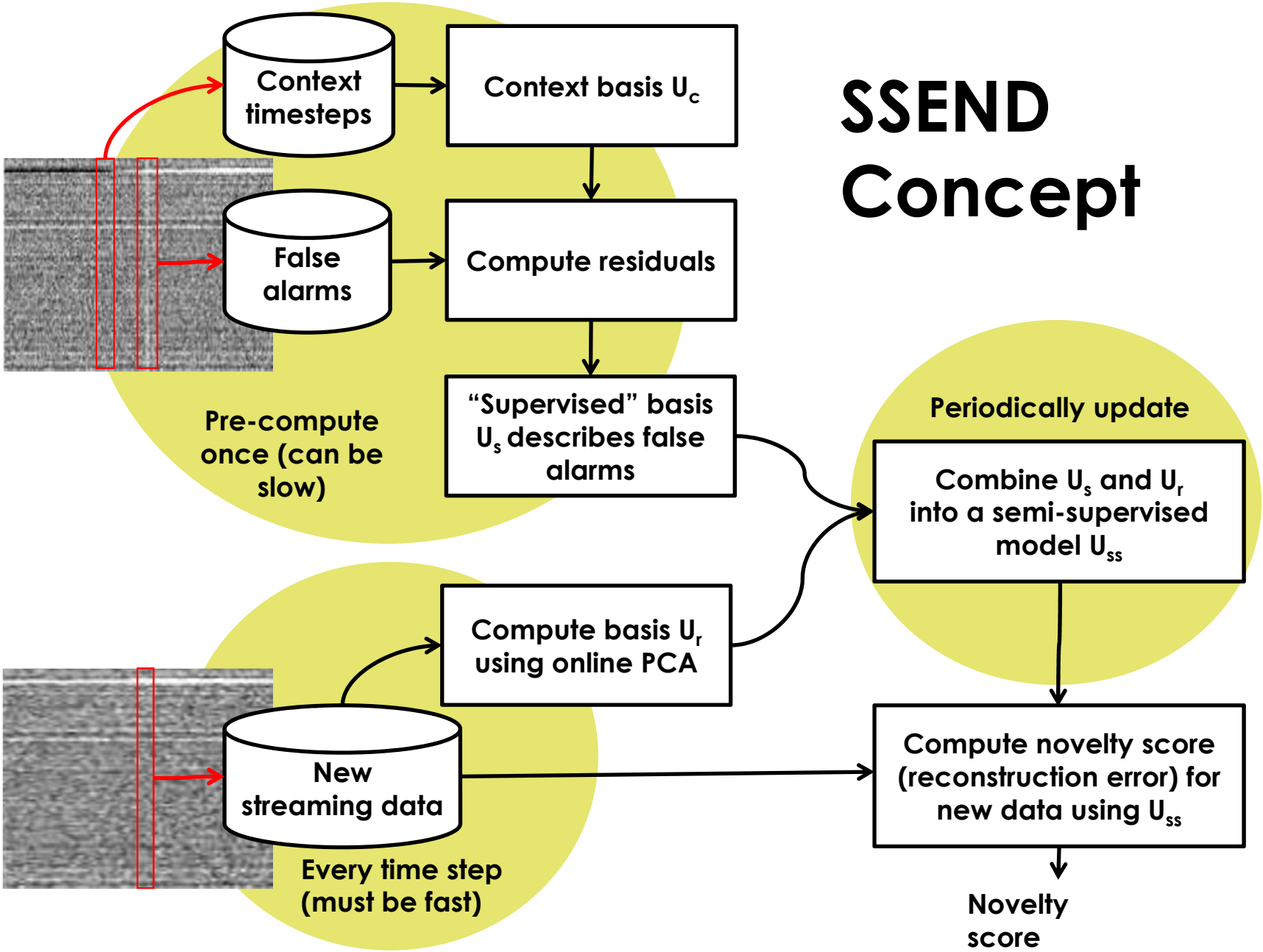
- **Basic assumption: each timestep combines a slowly-varying background, and intermittant "anomaly" or "false alarm" processes**

  X = background(t) + false(t) … + anomaly(t)

- **Online background estimation from incoming data**
- **Train a false alarm model in advance**
- **Use Reconstruction Error from linear basis as novelty score**

$$f(x_i) = ||x_i - \hat{x}_i|| = ||x_i - AA^T x_i||_2$$

SSEND Concept

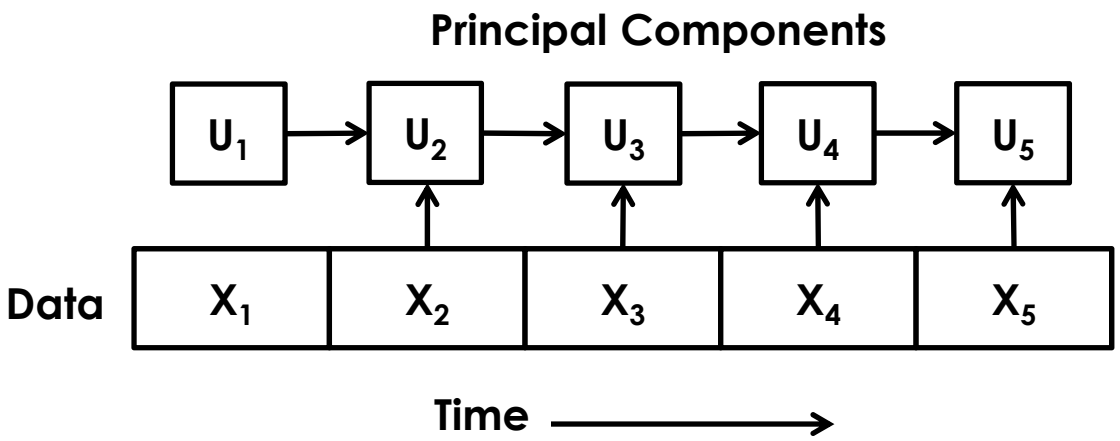


Online Updates

Compute principal components using Online PCA [Ross et al., 2004]

$$X = U\Sigma V^T$$

1. Have old decomposition $U_r \Sigma_r V_r^T$
2. We get a new data point X_q
3. Compute combined $U_p \Sigma_p V_p^T$ without X_p



Semi-supervision

Compute principal components from training data (false alarms):

$$X_s = U_s \Sigma_s V_s^T$$

Concatenate bases and orthogonalize with QR decomposition:

$$U_c = [U_r | U_s]$$

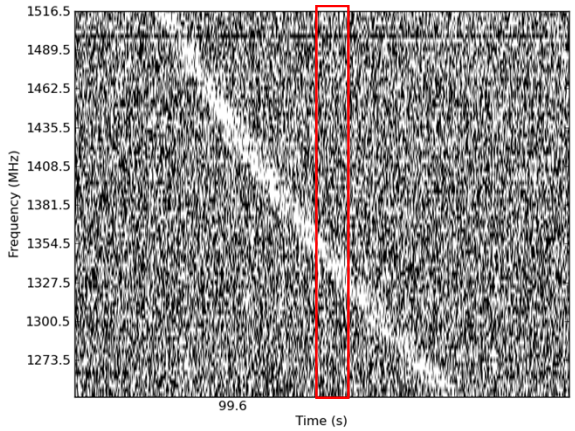
Retain first few bases in A

Compute reconstruction error:

$$f(x_i) = ||x_i - \hat{x}_i|| = ||x_i - AA^T x_i||_2$$

Experiments

- Subsample and segment data every 15 ms
  - 576 dimensions (6 time steps x 96 channels)



- Construct U_r
- Train U_s using 30 manually selected RFI
  - Collapse to 10 bases

Eigensignals

channel noise

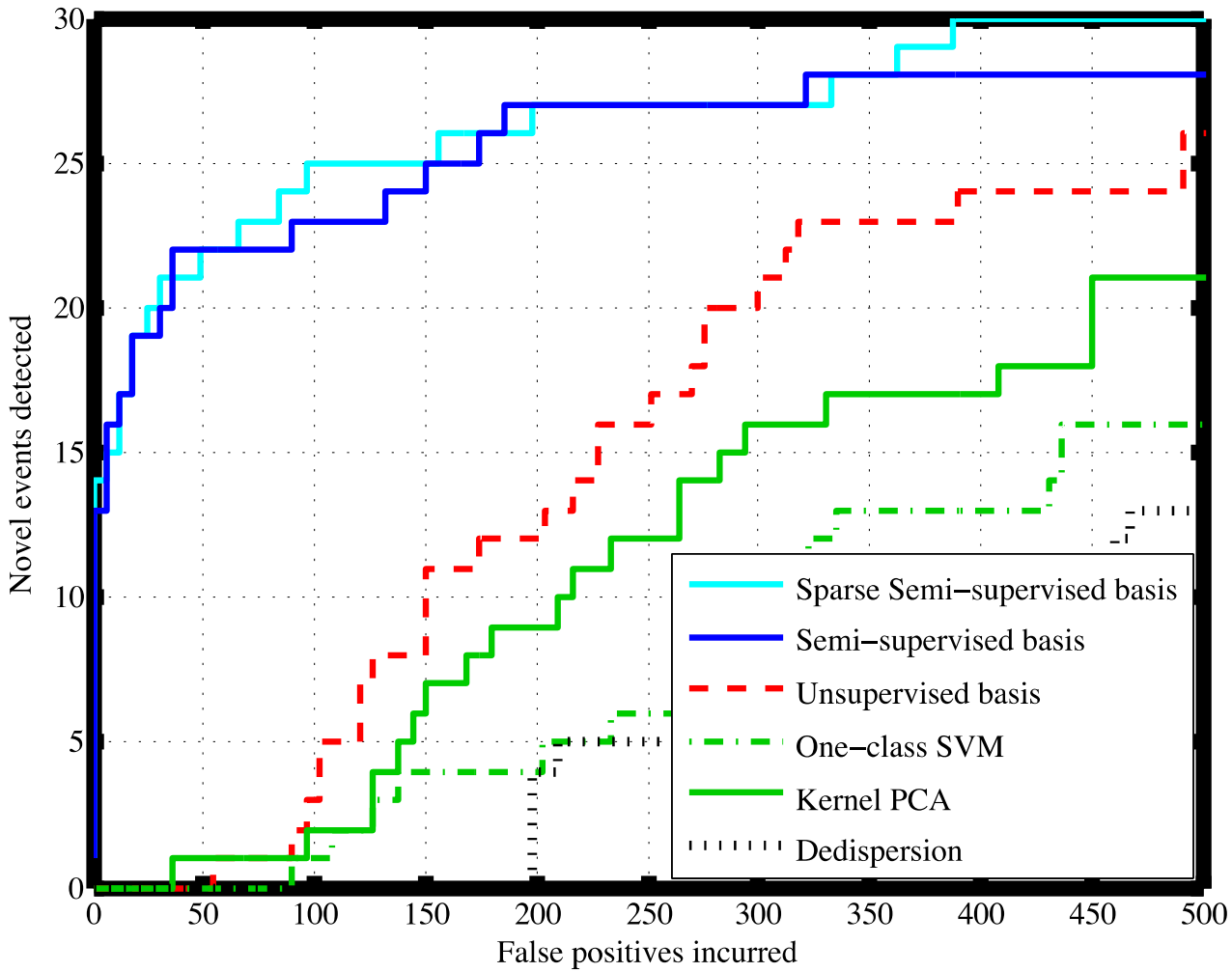
momentary RFI

Adaptive (Online PCA)

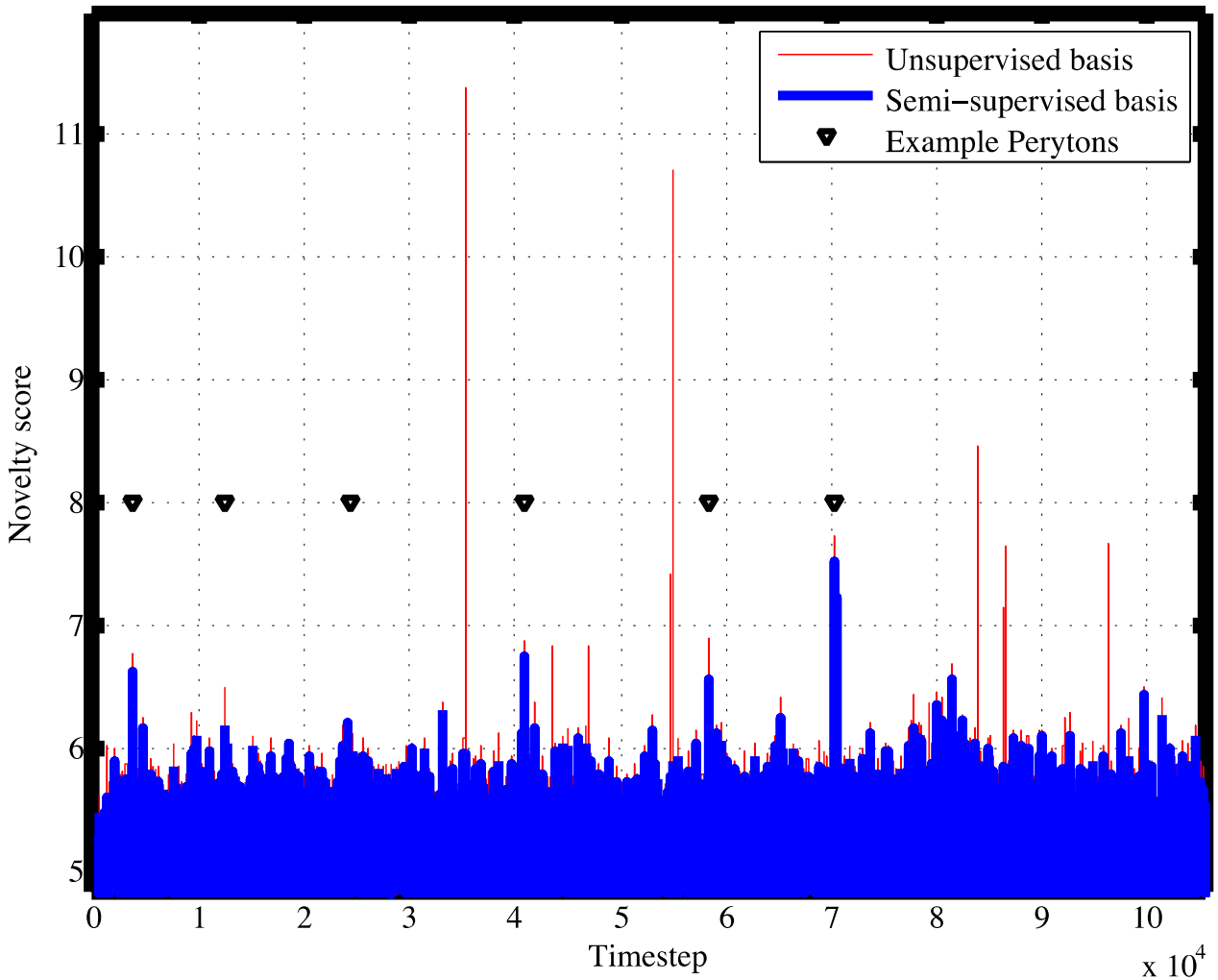
Supervised (PCA)

Supervised (Sparse PCA)

Performance

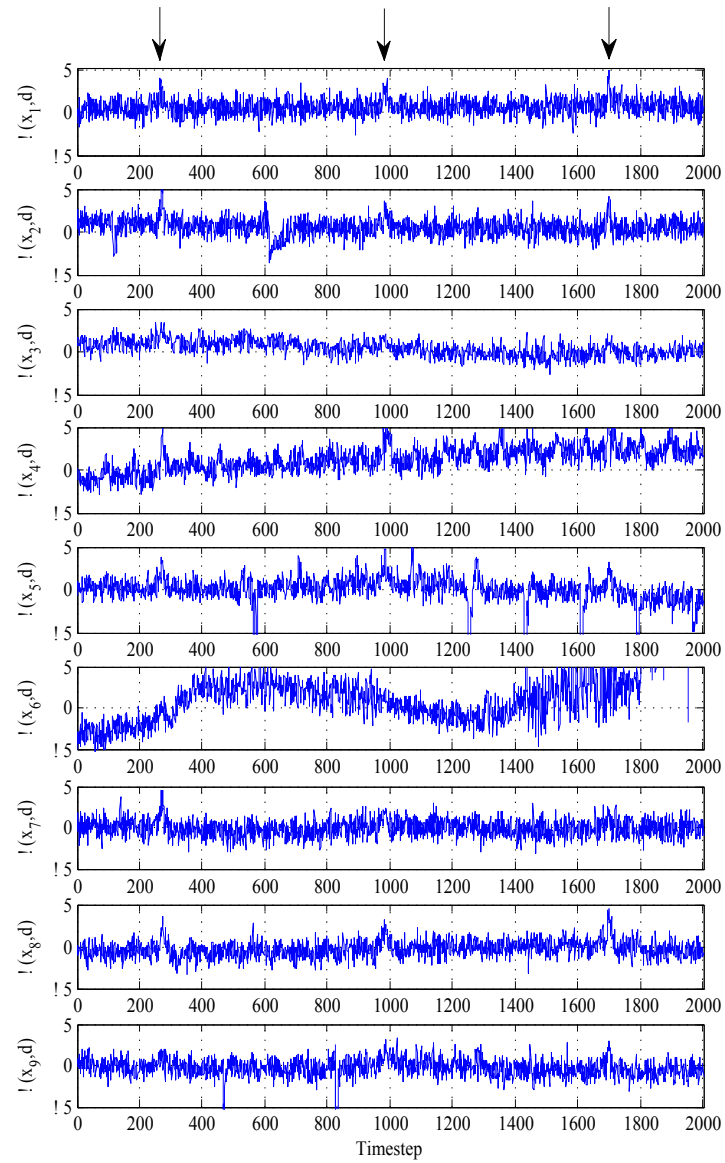


Performance



Case 2: Don't know background, but do know targets

- V-FASTR real-time software fast transient detector
- Operates commensally on the Very Long Baseline Array (VLBA) [Tingay et al.]



Pulses from Pulsar B0329+54 observed by 9 VLBA antennas

Multi-station detection algorithms

- Signal strength is shown in each of two stations (arbitrary units).
- Curves show the decision boundary of various detection rules

Sum

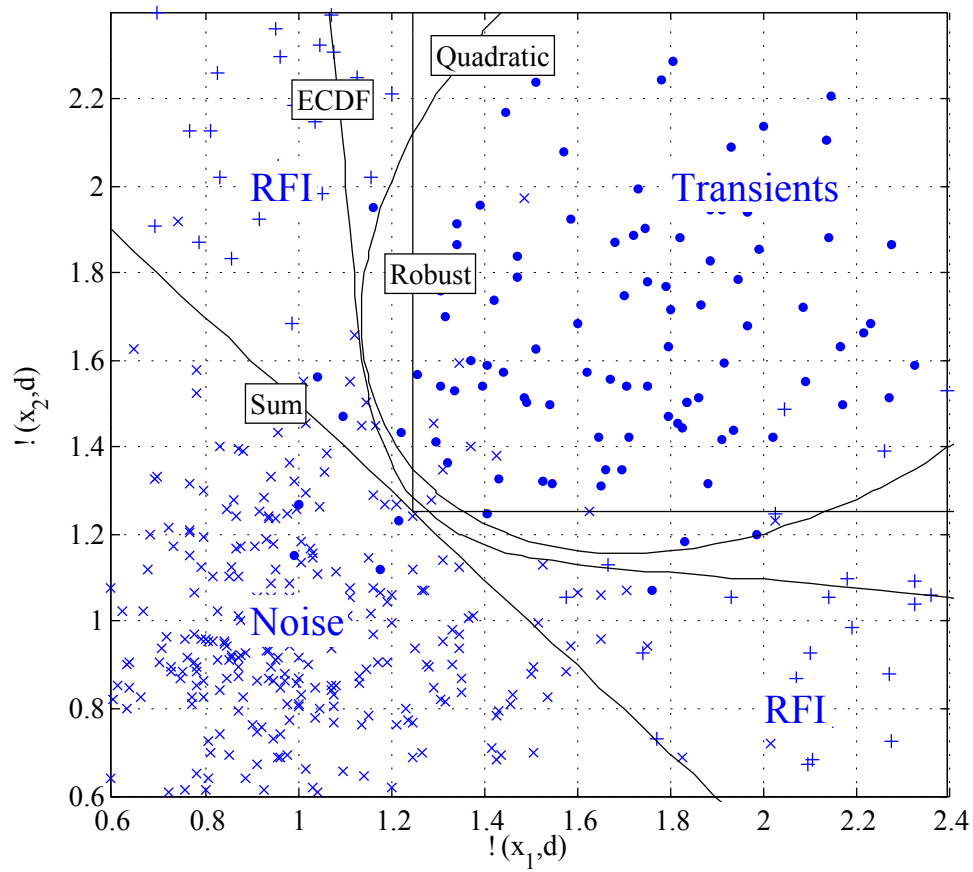
$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} S_a(DM)$$

Robust

$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A-1} S_a(DM)$$

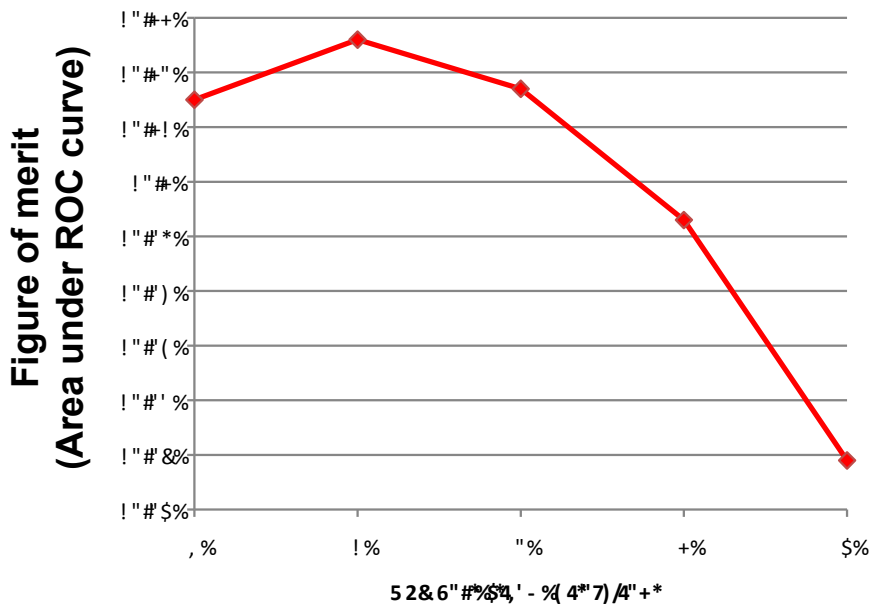
ECDF

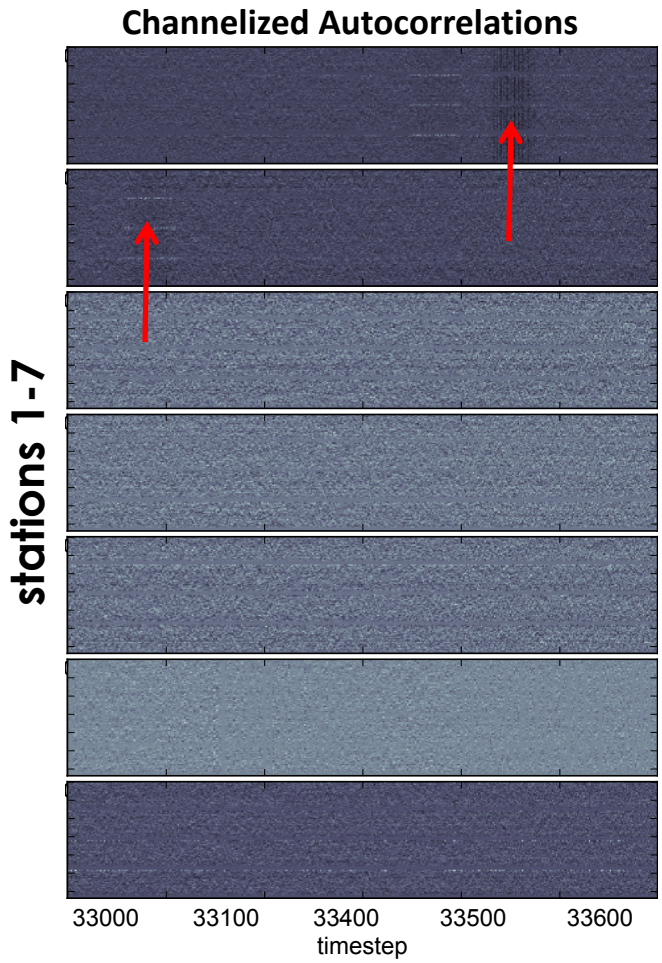
$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} P(X \leq S_a(DM))$$



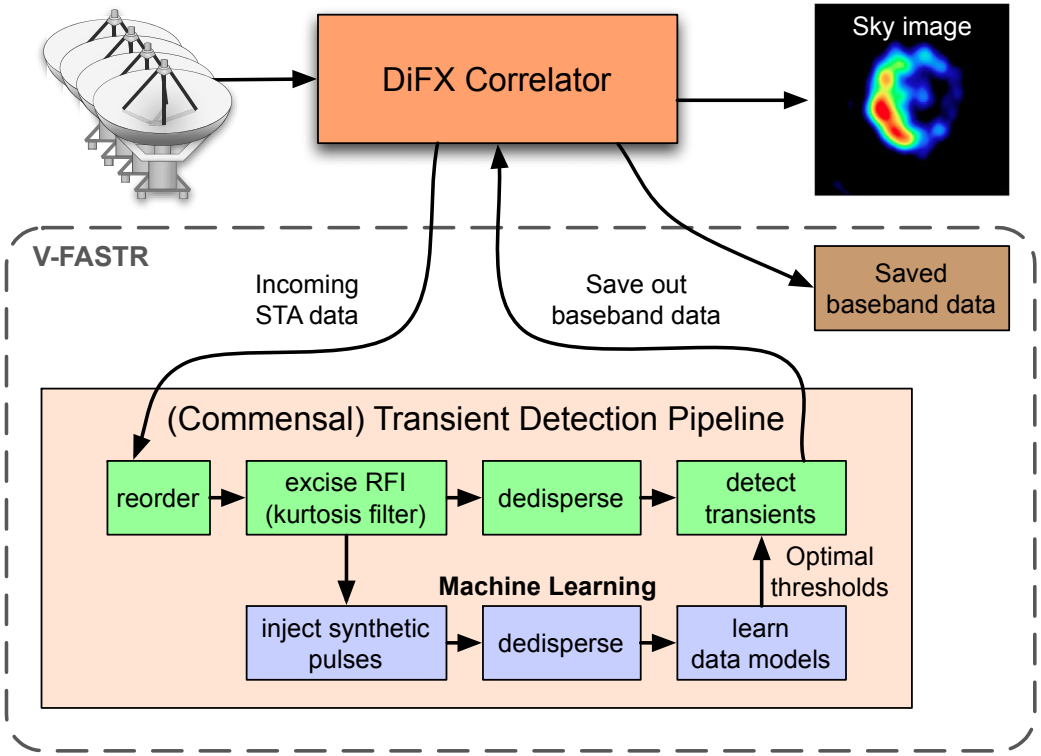
Robust detection example

- **Robust estimator considers all antennas in detection decision**
- **Must set system parameters without knowing background noise in advance**



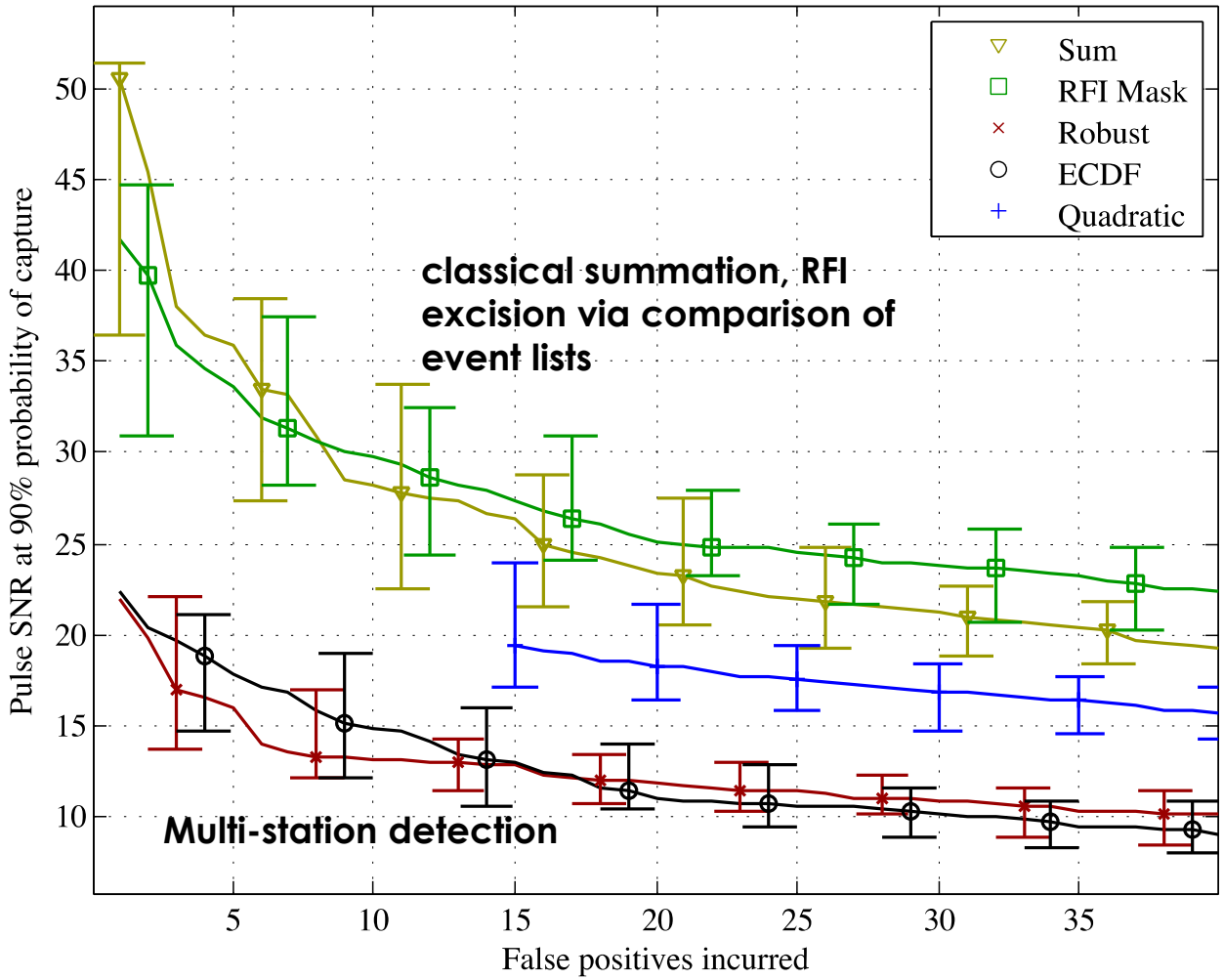


Adaptive self-tuning: train a classifier in real time!



See: Wayth et al., ApJ 2011, Thompson et al., ApJ 2011,

Pulsar B0329+54 observation

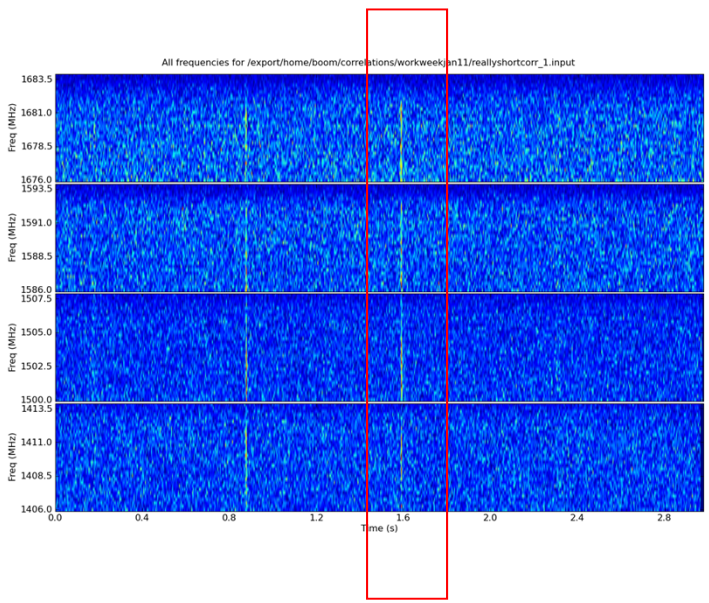


Better performance

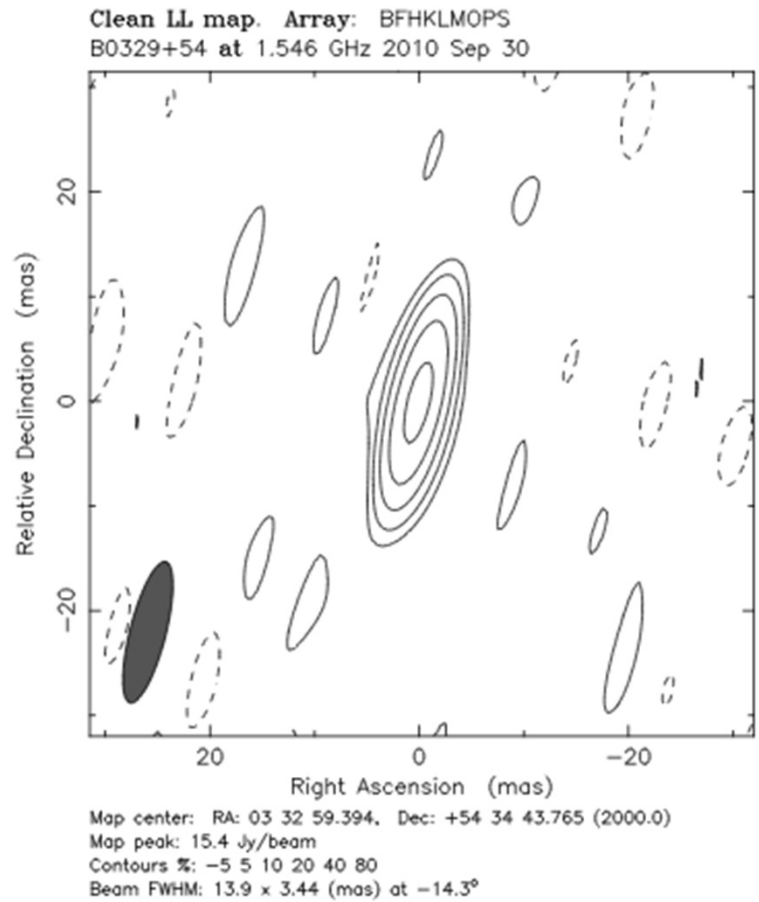
Thompson et al., 2011

Single-pulse imaging

- **Imaging of B0329+54 from a single pulse**
- **No external calibration**

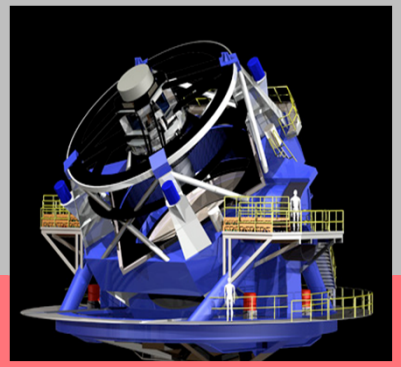


Post-correlation dedispersed



Cleaned image from ~5ms data (-5% contour is dashed)

Summary



Source detection and Classification

Statistical inference

Clustering, anomaly detection

Manifold learning

Random-access storage

Online, Automatic

Further reading

R. B. Wayth, W. F. Brisken, A. T. Deller, W. A. Majid, D. R. Thompson, S. J. Tingay and K. L. Wagstaff, V-FASTR: The VLBA Fast Radio Transients Experiment. *ApJ* 735, 98 (2011)

D. R. Thompson, K. L. Wagstaff, W. F. Brisken, A. T. Deller, W. A. Majid, S. J. Tingay and R. B Wayth, Detection of Fast Radio Transients with Multiple Stations: A Case Study with the Very Long Baseline Array. *ApJ* 735, 98 (2011)

D. R. Thompson, W. A. Majid, C. J. Reed, K. L. Wagstaff, "Semi-Supervised Anomaly Detection with Adaptive Eigenbases, and Application to Radio Transients", Conference on Intelligent Data Understanding 2011.

D. R. Thompson, W. A. Majid, C. J. Reed, K. L. Wagstaff, "Semi-Supervised Eigenbasis Novelty Detection," *Statistical Analysis and Data Mining* (in preparation).

Thanks!

Sarah Burke-Spolaor, Dayton Jones, Peter Hall, the ASKAP/CRAFT team, The National Radio Astronomy Observatory and the Staff of the Array Operations Center, Robert Preston, Joe Lazio, J-P Macquart