Semi-Supervised Novelty Detection with Adaptive Eigenbases, and Application to Radio Transients

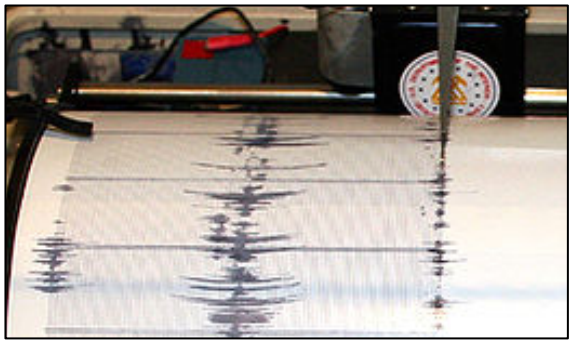
David R. Thompson, Walid A. Majid, Kiri L. Wagstaff, and Colorado J. Reed

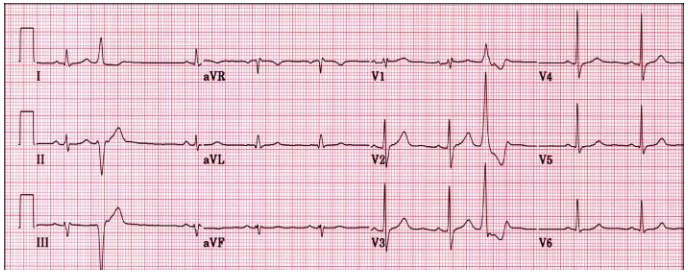
Jet Propulsion Laboratory, California Institute of Technology

October 21, 2011

Novelty/Anomaly Detection

- Anomalies we care about
  - Heart attacks
  - Earthquakes
  - Stock market crashes



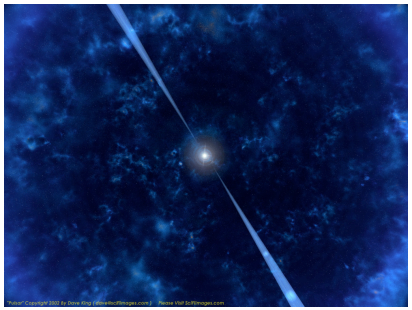


- Subsequent anomalies aren't interesting:
  - Hiccups
  - The boy who cried wolf
  - Allergic reactions

"Statistically anomalous" ≠ "interesting"

Novelty Detection for Radio Astronomy

- Fast transients: brief, energetic pulses
  - X-ray bursts, pulsars, neutron stars, active galactic nuclei, etc.



- RFI: brief, energetic pulses
  - Terrestrial origin: cars, cell phones, satellites
- Low false positive rate is vital
  - Human effort required to review candidates
  - Avoid overflowing data buffer
- State of the art: matched filter

Can we do better?

SSEND Concept

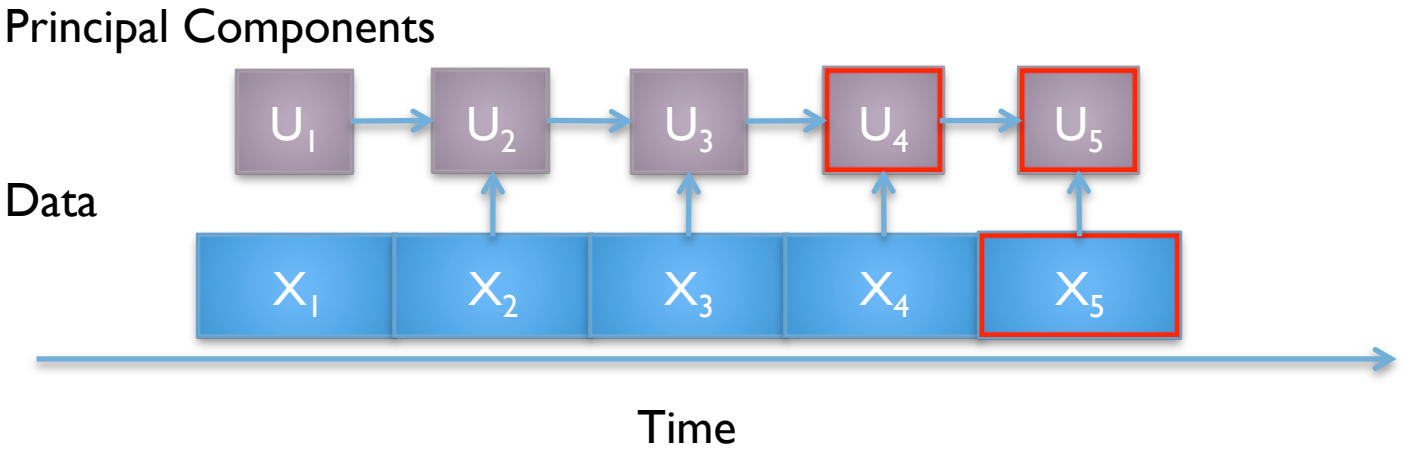
- Construct eigenbasis, then compute novelty score using reconstruction error
- Novel features
  - Online updates based on incoming data
  - Semi-supervised:
    informed by known "ignorable anomalies"

Online Updates

- Compute principal components:

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

- Online PCA [Lim et al., 2004]
  - Given $U_p\Sigma_p V_p^T$, new data X_q, get $U_r\Sigma_r V_r^T$
  - No need to explicitly store X_p



Semi-supervision

- Compute principal components from training data (ignorable anomalies):

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

- Combine bases and use QR decomposition to orthogonalize:

$$U_c = [U_r | U_s]$$

- Retain first few bases in A
- Compute reconstruction error:

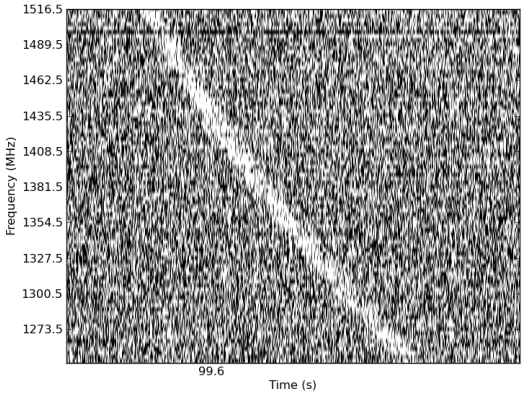
$$f(x_i) = ||x_i - \hat{x}_i|| = ||x_i - AA^T x_i||_2$$

Data

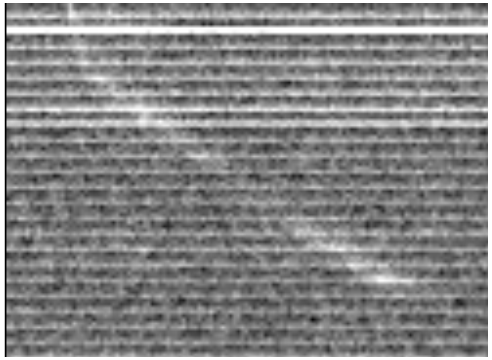
- Parkes Multibeam Survey [Edwards et al., 2001]
  - 1.4 GHz, 125 μs sample time, 96 channels
  - Goal: detect pulsars
    - … but other anomalies also lurk within

Parkes telescope multibeam receiver

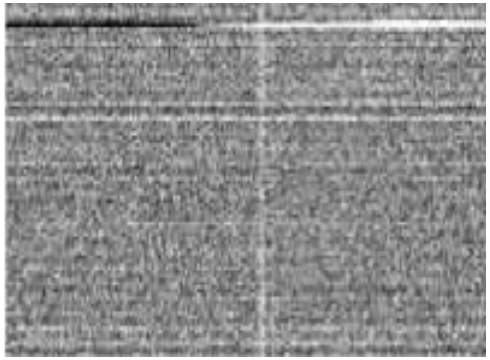
Pulsar pulse



"Peryton"



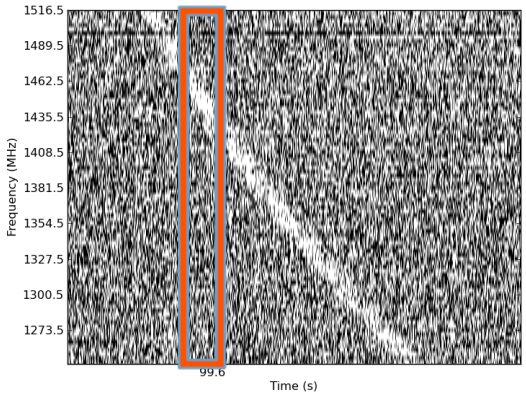
Radio Freq. Interference



Time

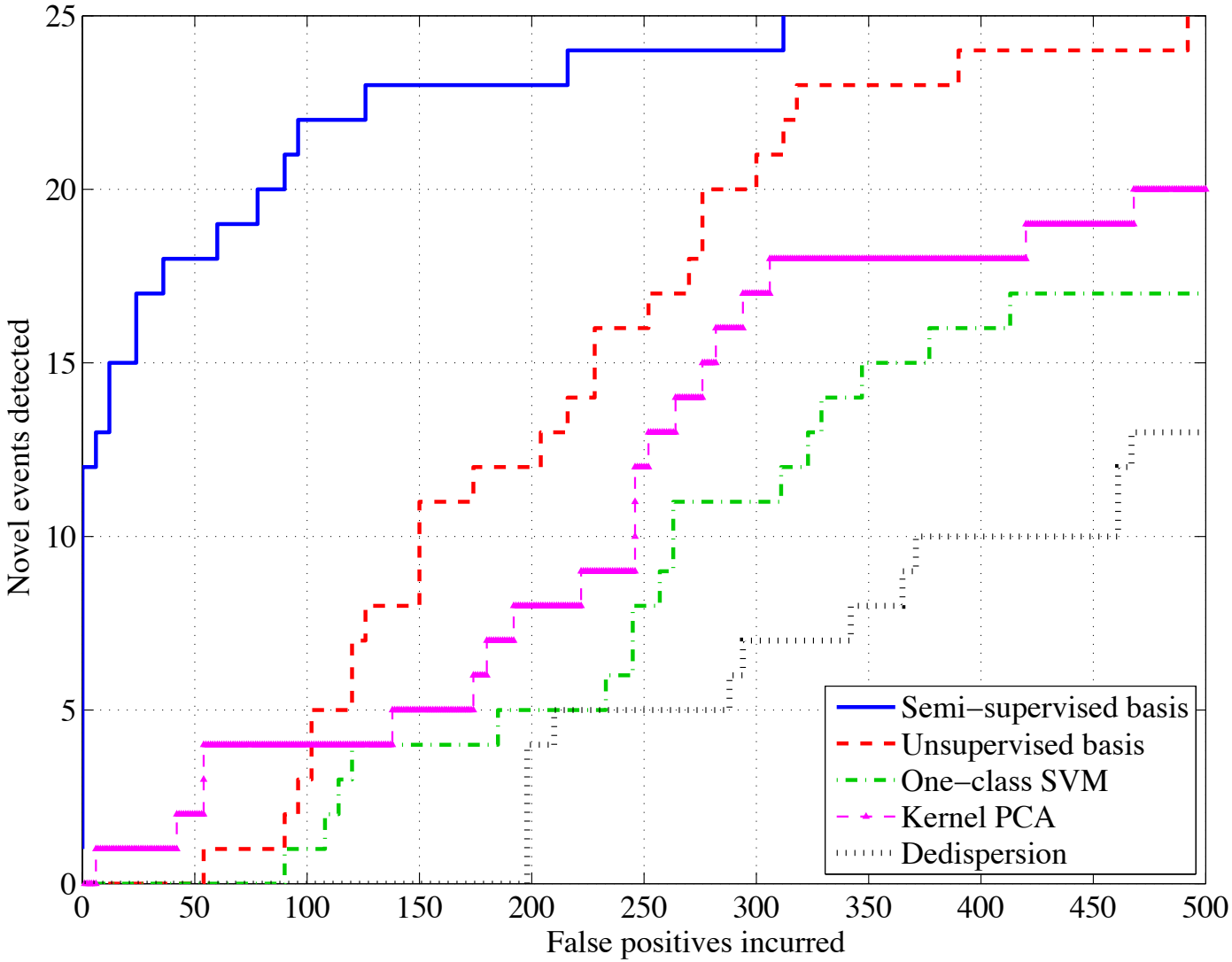
Experiments

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

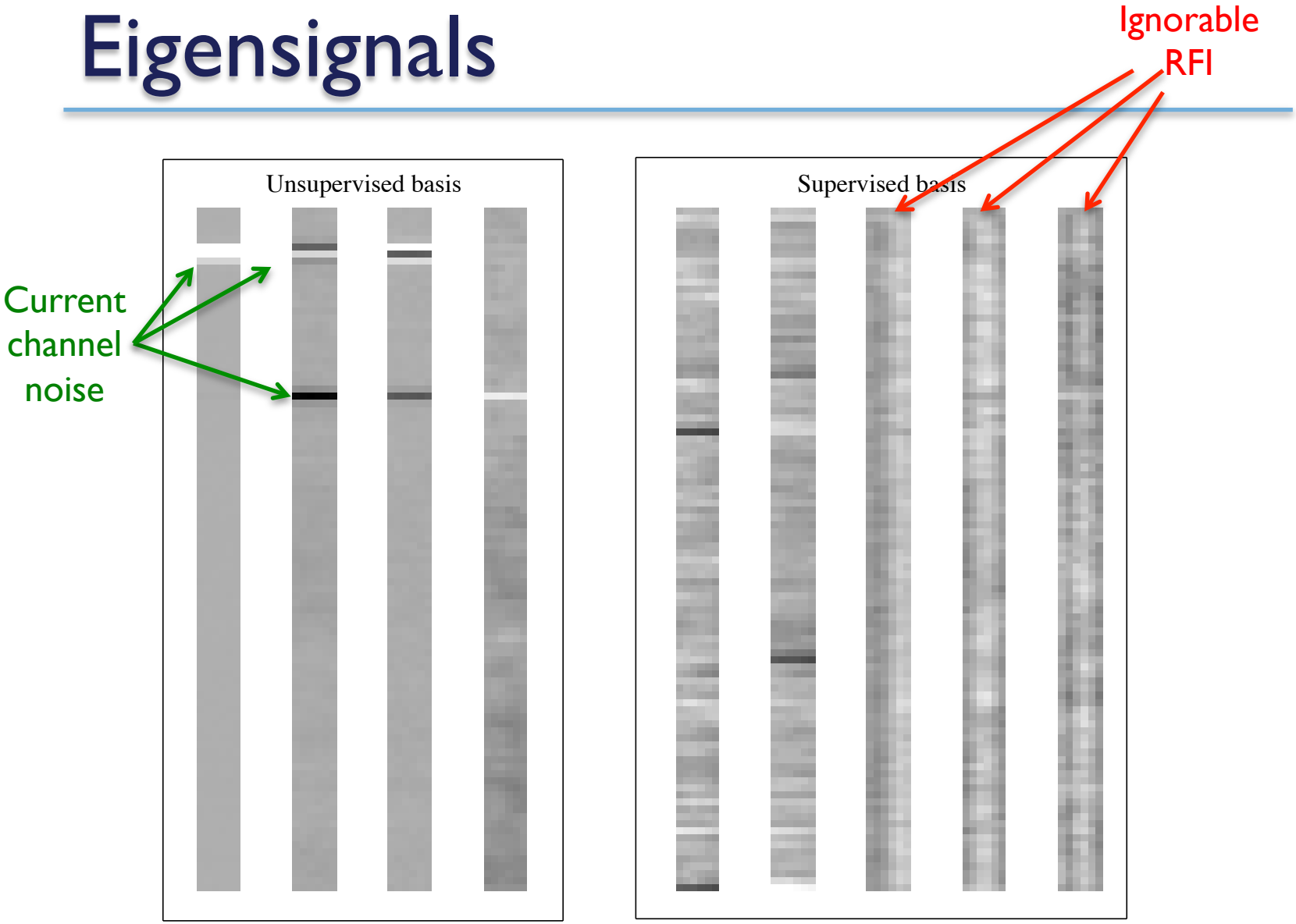


- Construct U_r online; retain 4 bases
- Train U_s using 30 manually selected RFI
  - Collapse to 10 bases

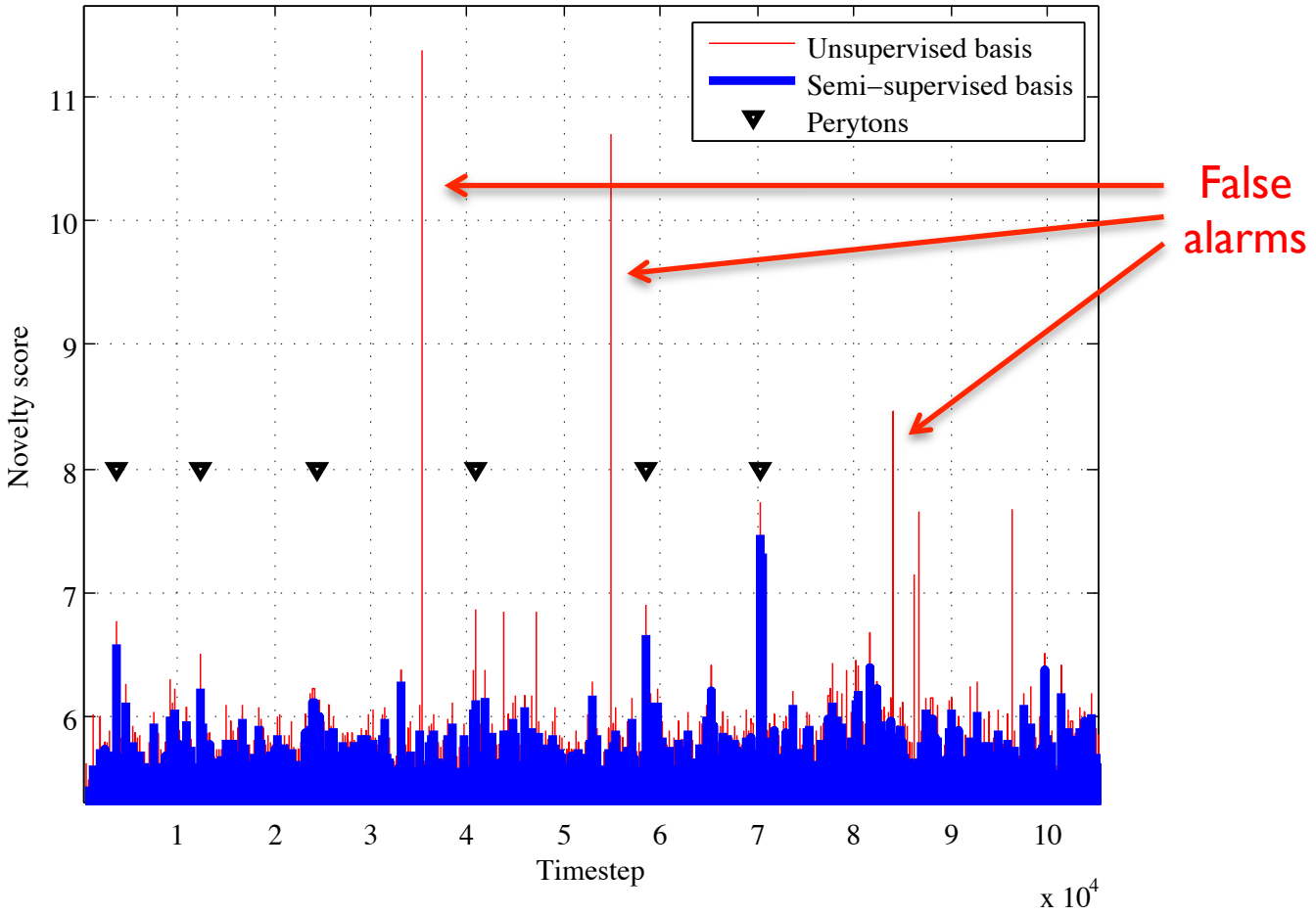
Results


Novel events detected
False positives incurred
0
5
10
15
20
25
0
50
100
150
200
250
300
350
400
450
500
Semi−supervised basis
Unsupervised basis
One−class SVM
Kernel PCA
Dedispersion

Eigensignals



Novelty Scores



Summary

- SSEND: Novelty detection that
  - Adapts to changing data properties
  - Avoids flagging known uninteresting anomalies
- Novelty score:
  - Reconstruction error using combined bases from online PCA + static prior bases
- Application to radio astronomy
  - And anytime false positives are costly

Thank you: Sarah Burke-Spolaor, J-P Macquart, Dayton Jones, Bob Preston, Joseph Lazio, and the SURF program.