Machine learning for exploring data streams, with lessons from the Very Long Baseline Array

David R. Thompson
Jet Propulsion Laboratory,
California Institute of Technology
david.r.thompson@jpl.nasa.gov

With W. Brisken (NRAO), A. Deller (ASTRON, NL), C. Reed (Cambridge), S. Tingay (ICRAR), K. L. Wagstaff (JPL), R. Wayth (ICRAR)

Image: NRAO

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

AGENDA

1. Why online data mining matters
2. Commensal transient detection on the VLBA
3. Other noteworthy applications

Image: NRAO

Traditional data mining

Random-access storage

Source detection and Classification

Statistical inference

Clustering, anomaly detection

Manifold learning, visualization

Automated acquisition "to spec"

Asychronous manually-supervised

Image credit: Keck Observatory

Traditional data mining

- A managed, asynchronous process
- Algorithms see the entire dataset
- Algorithms often scale with n^k (naïve)

  … or *n log n* (clever)

Images: SKATelescope.org, LSST.org, Keck Observatory

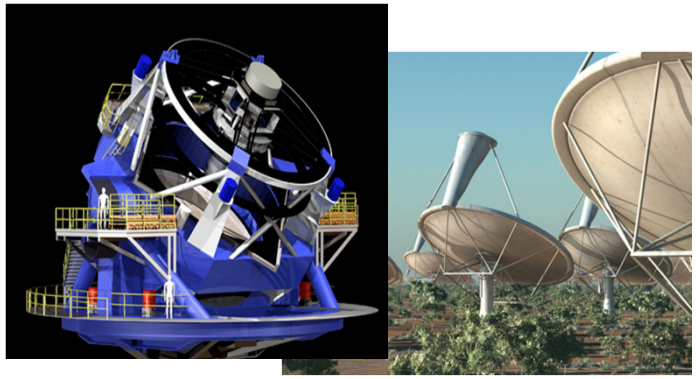
Traditional data mining

- A managed, asynchronous process
- Algorithms see the entire dataset
- Algorithms often scale with n^k (naïve)

  … or *n log n* (clever)

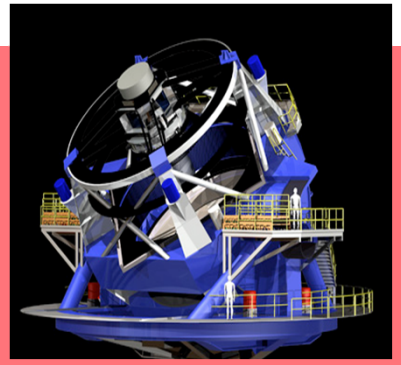
Next-generation instruments

- Too much data to store or analyze
- Desire immediate triage, followup
- Continuous online stream mining
- Constant-time algorithms



Images: SKATelescope.org, LSST.org, Keck Observatory

Next-generation instruments



Source detection and Classification

Statistical inference

Clustering, anomaly detection

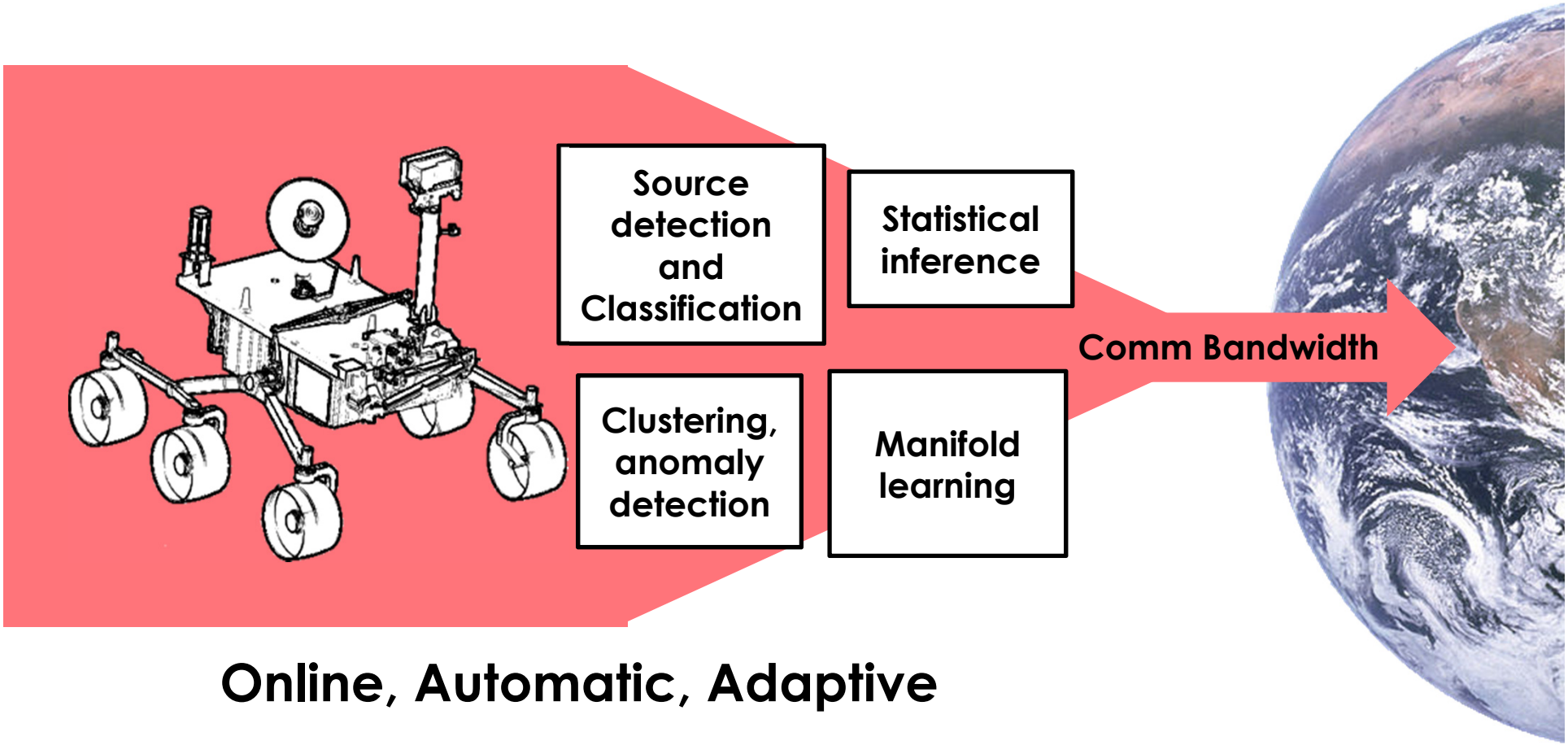
Manifold learning

Random-access storage

Online, Automatic, Adaptive

Images: SKATelescope.org, LSST.org

Remote explorers



Online, Automatic, Adaptive

Images: NASA / JPL / Caltech

This presentation: Two archetypal stream mining scenarios

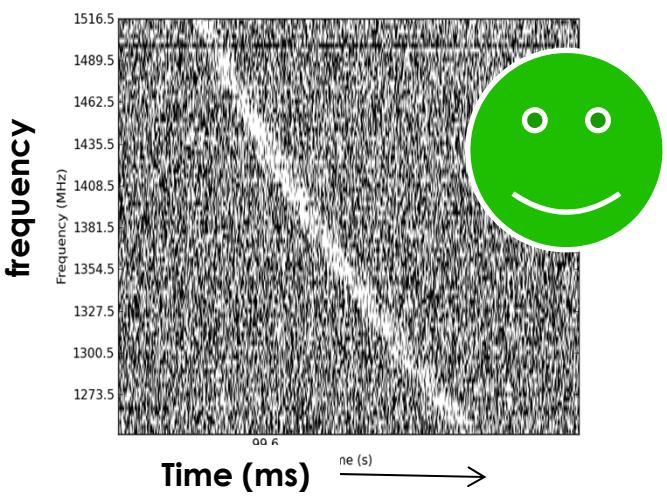
We know exactly what we're looking for
…but don't know the background noise

We can characterize background noise
…but don't know what we're looking for

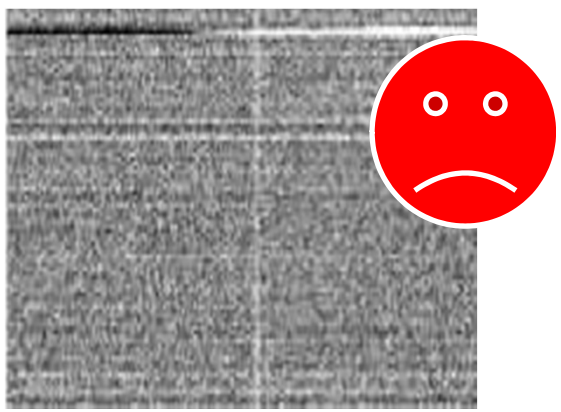
Case I: We know what we're looking for, but can't anticipate the background noise

Radio transients are millisecond pulses from sources such as:

- Collapsing black holes
- Gamma Ray Bursts
- RRATs (intermittent pulsars)



Dispersed pulse – save it!



Radio Frequency Interference – Ignore it!

V-FASTR: Commensal transient detection on the Very Long Baseline Array [S. Tingay et al.]

- Passive analysis piggybacks on all daily correlator operations
- Detects pulses in real time, while removing interference
- Saves interesting events to the archive for asynchronous manual followup

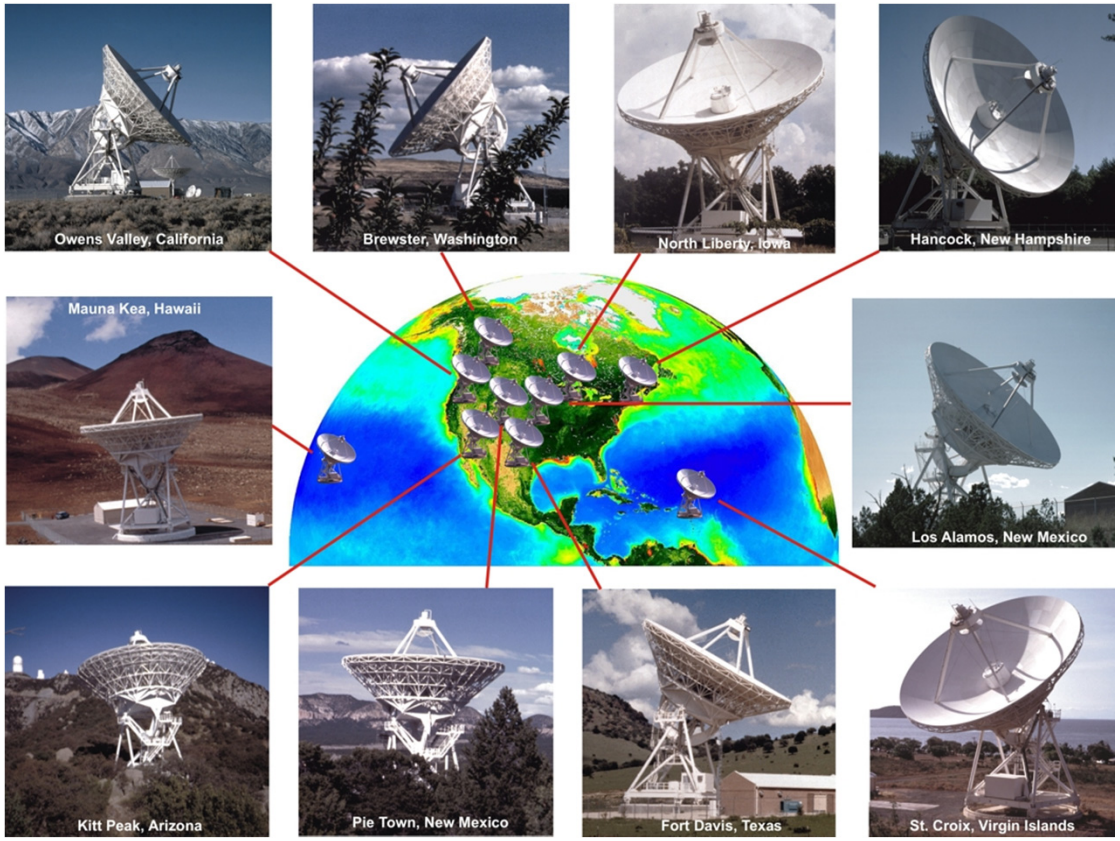
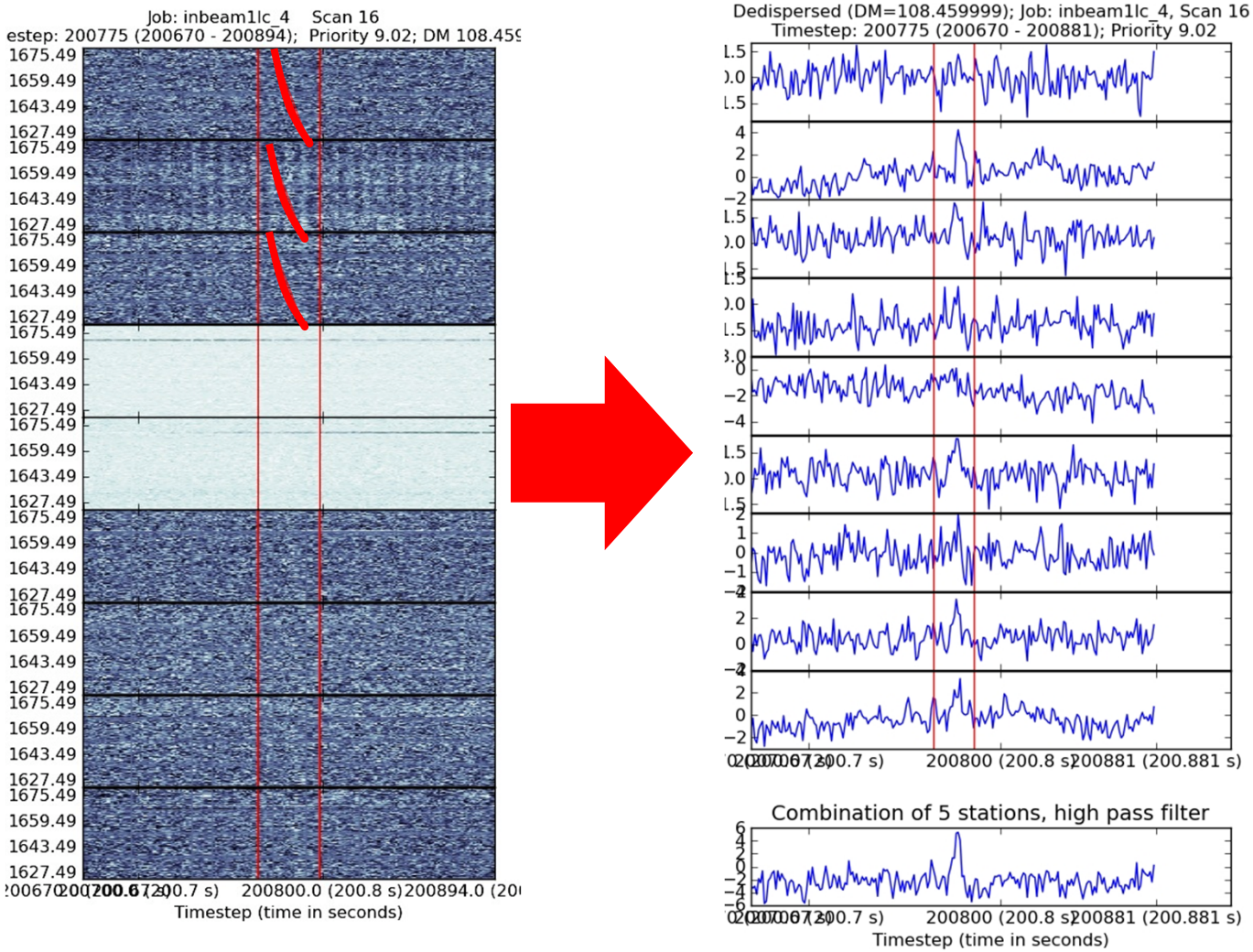


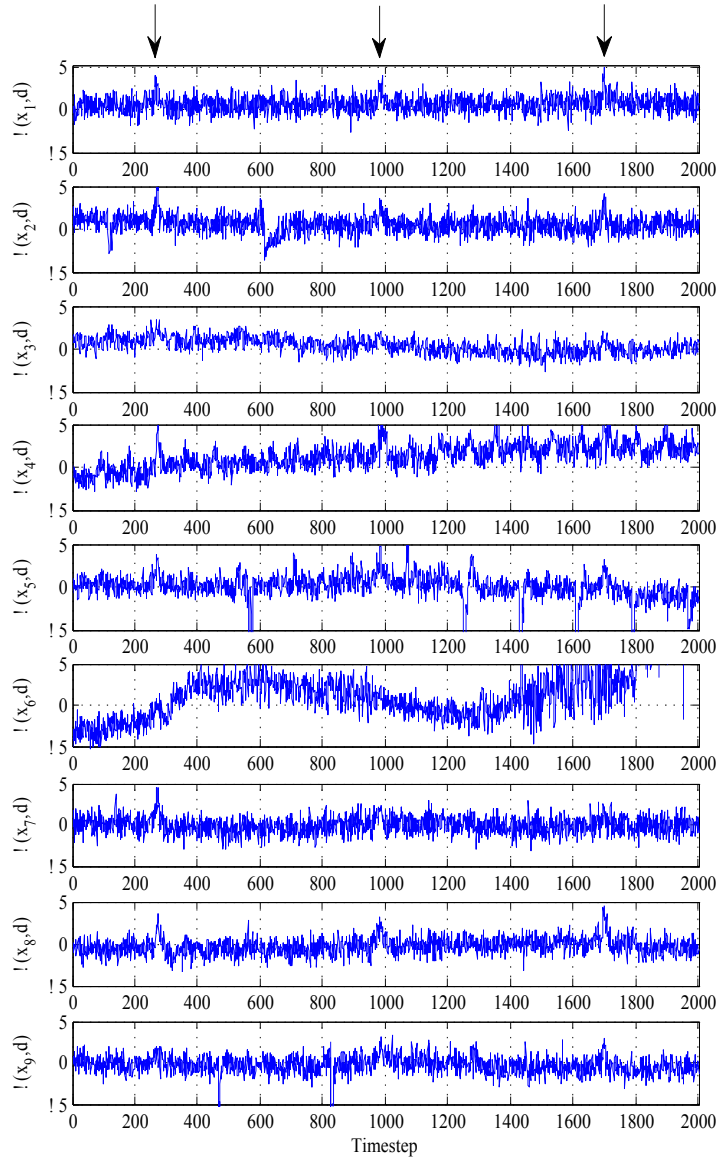
Image: NRAO

The Dedispersion search



Problem: severe interference

- RFI properties change constantly
  - Differ across observations
  - Differ across frequencies
  - Differ across stations
  - Differ across correlation parameters
- Both impulsive and slowly-varying interference



Multi-station detection algorithms

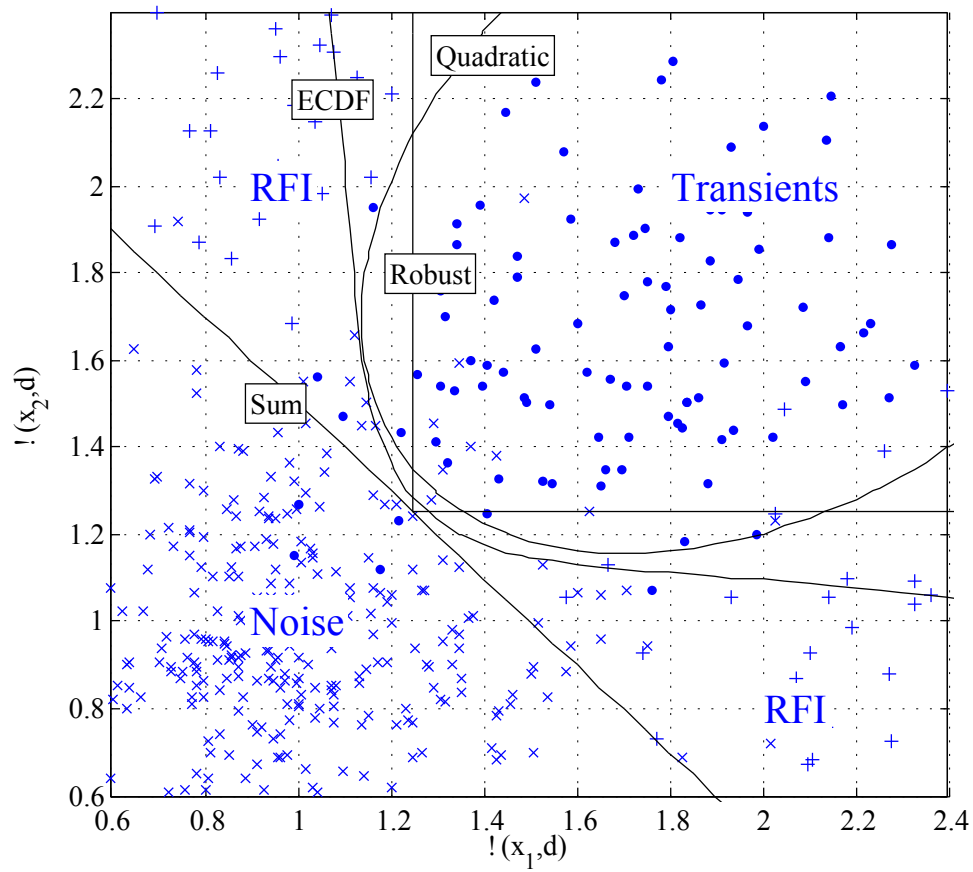
Signal strength is shown in each of two stations (arbitrary units).

Curves show the decision boundary of different 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 $$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} P(X \leq S_a(DM))$$



Multi-station detection algorithms

Signal strength is shown in each of two stations (arbitrary units).

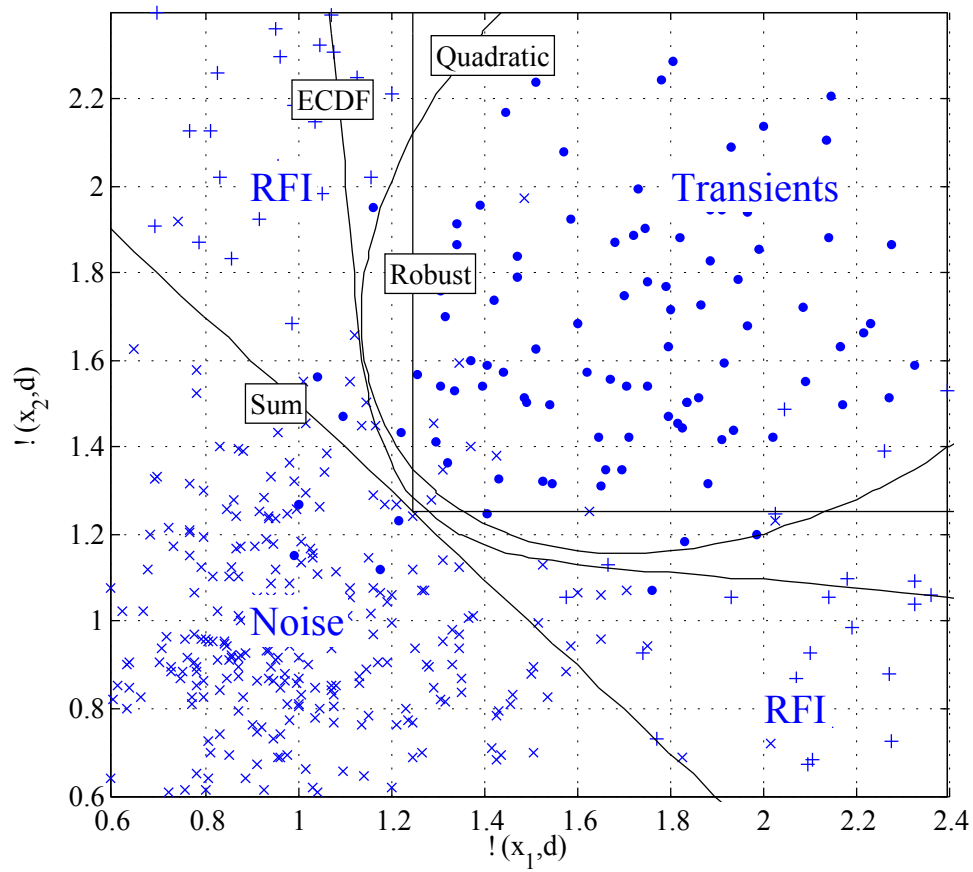
Curves show the decision boundary of different detection rules

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

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

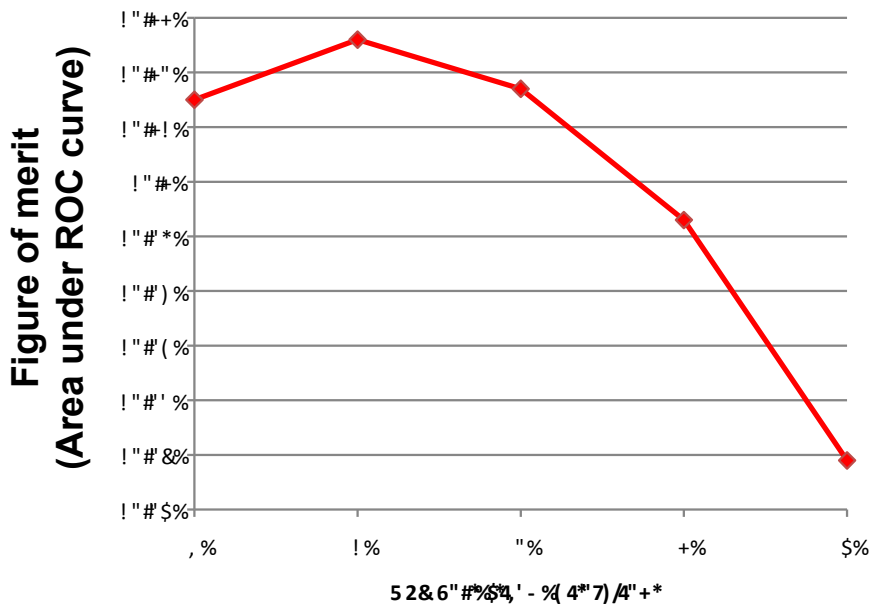
ECDF $$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} P(X \leq S_a(DM))$$

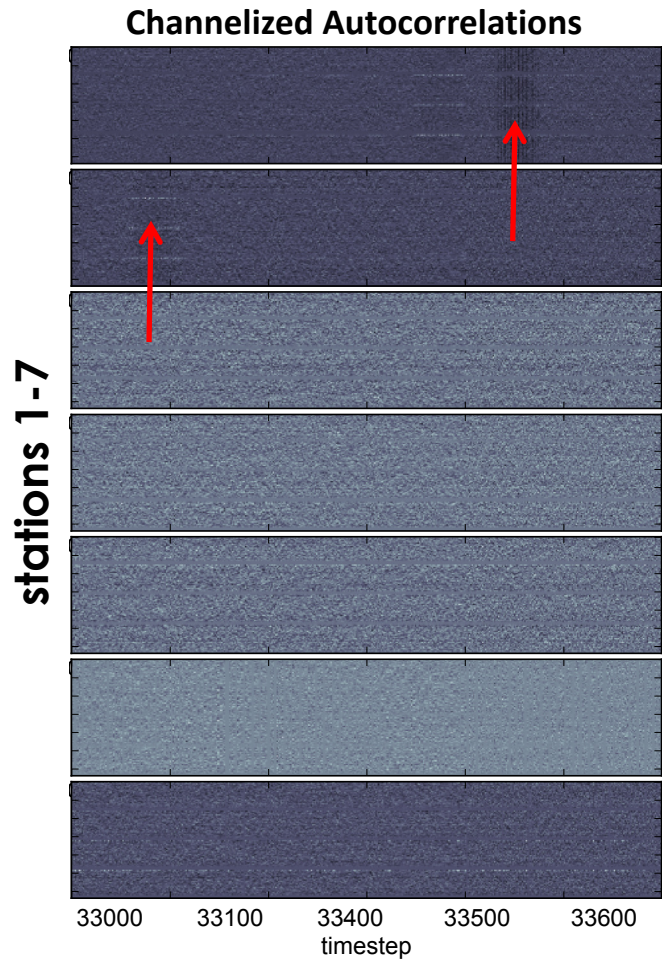
BUT… noise and RFI distributions change constantly



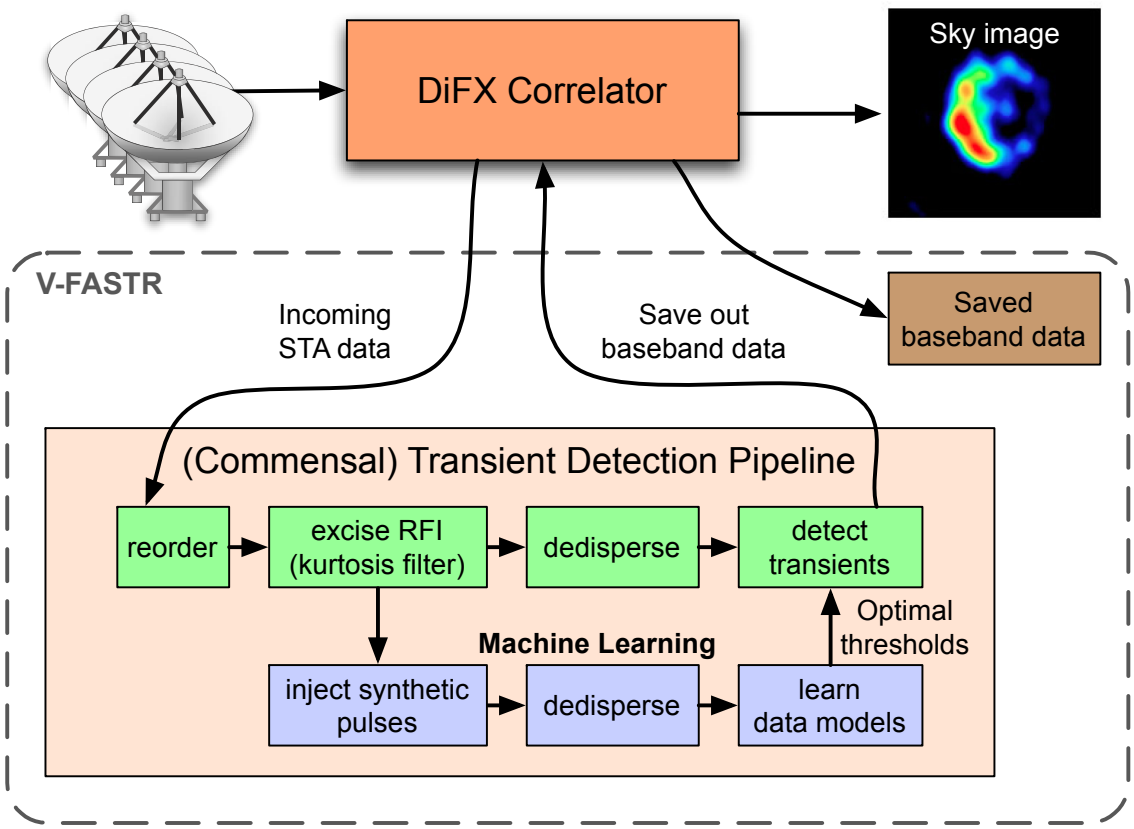
Robust detection example

- **Robust estimator excludes stations to excise RFI**
- **How many antennas to exclude? The answer varies depending on noise conditions!**



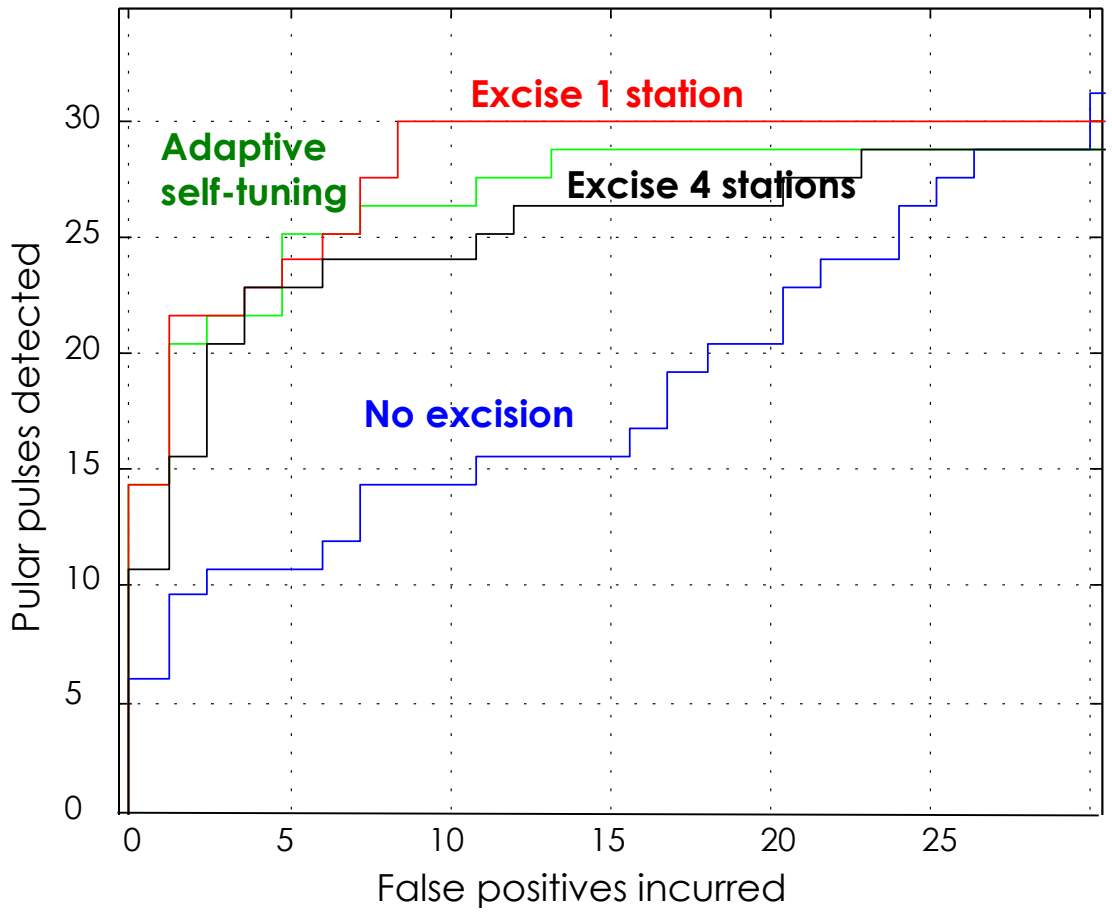


Adaptive self-tuning: designs new classifiers in real time

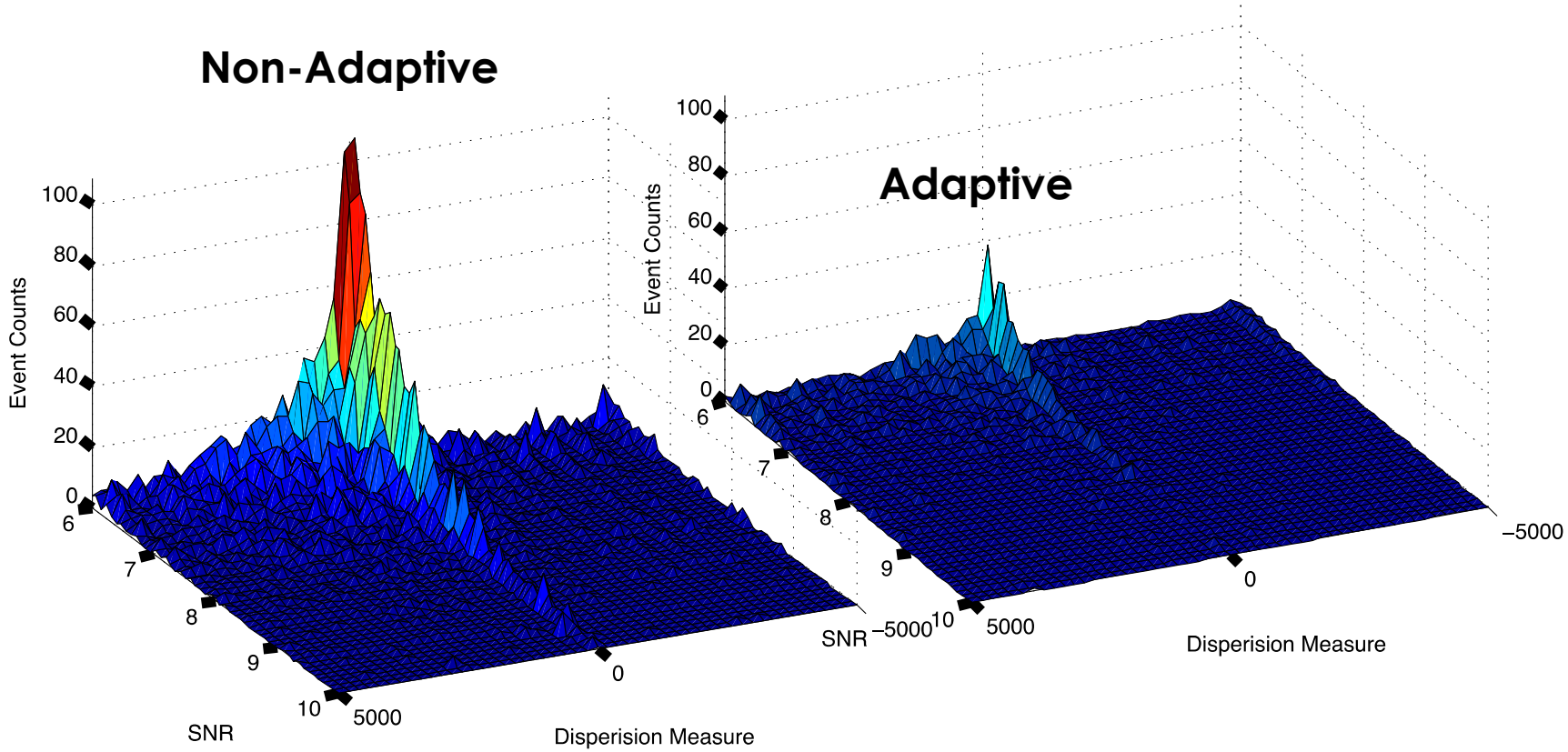


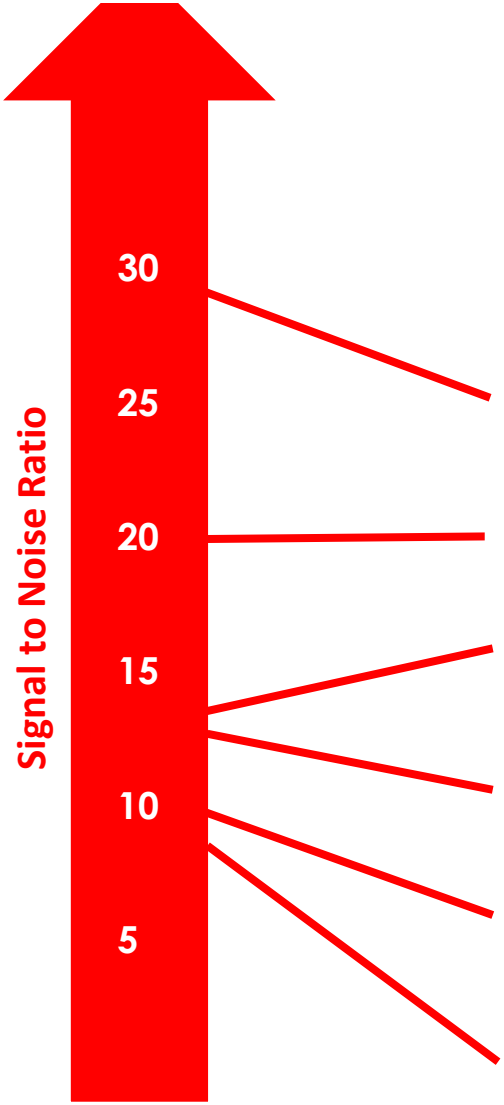
See: Wayth et al., ApJ 2011, Thompson et al., ApJ 2011,

Empirical performance (pulsar)



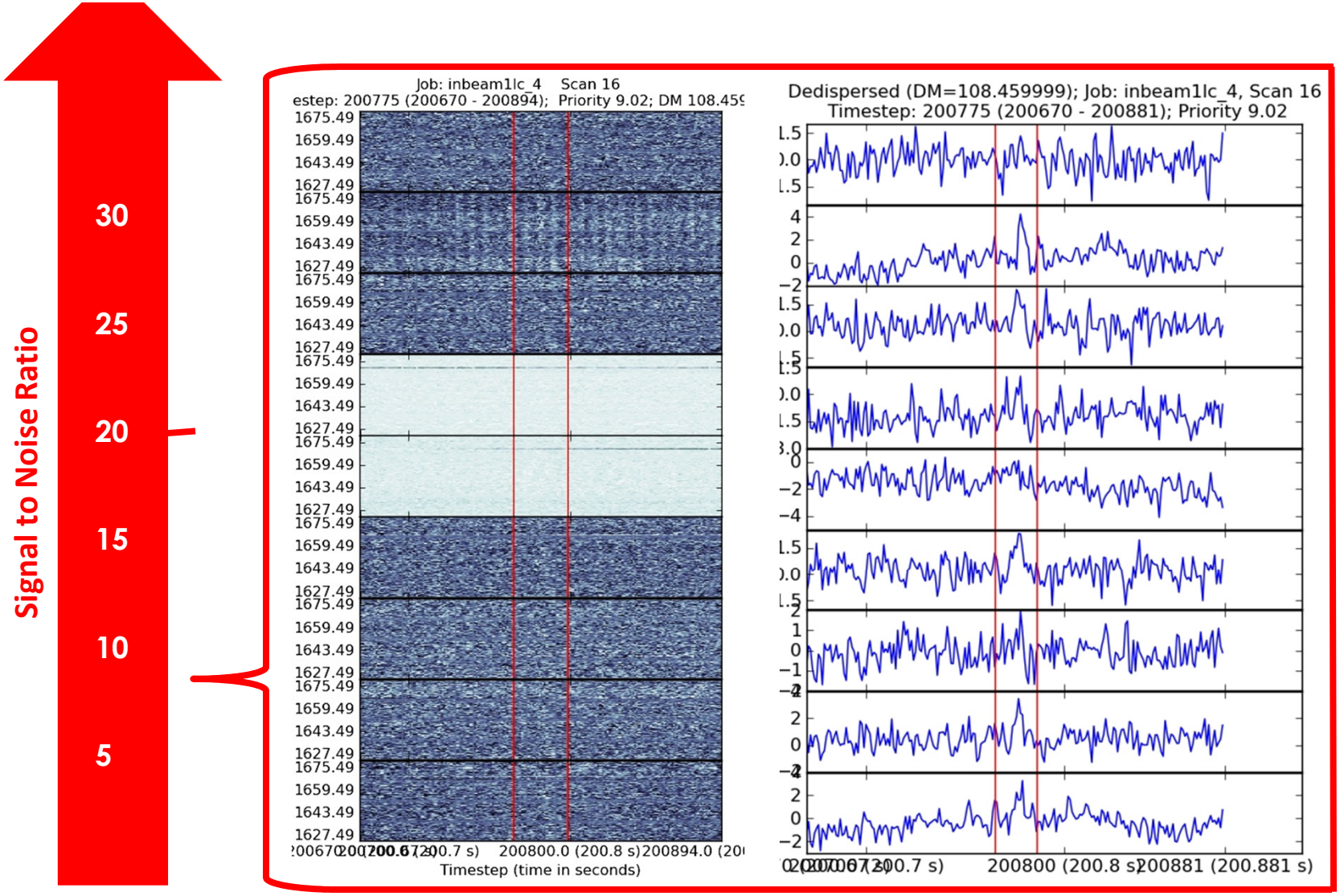
Alerts for April 2012



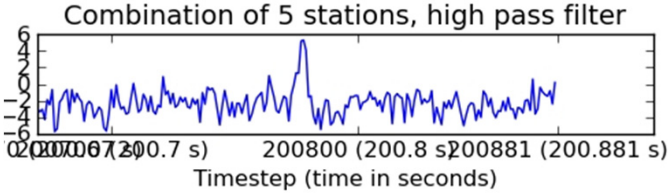


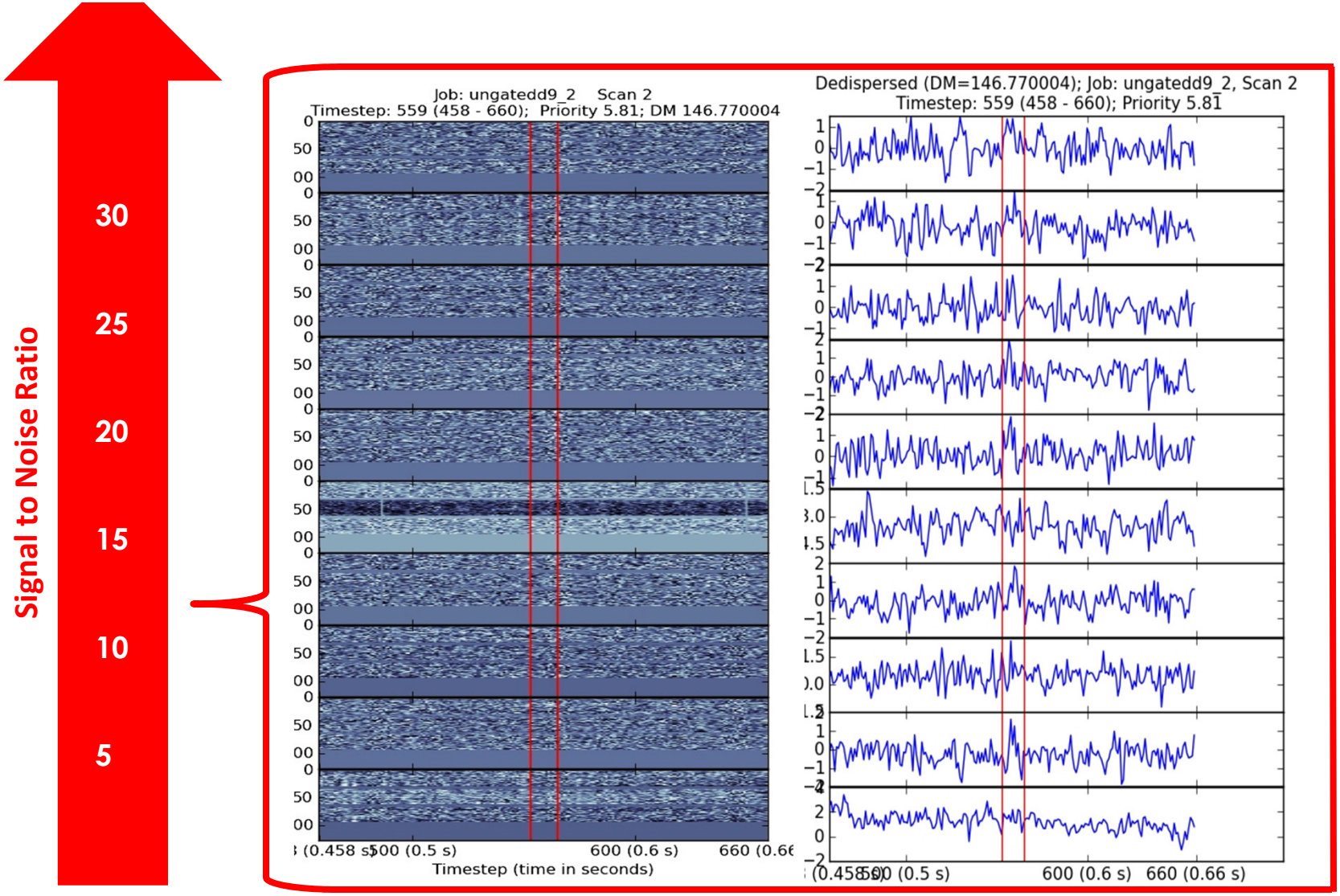
Pulsar name	Max SNR	DM	On-pulse flux density (Jy)
J0332+5434	>50	26	14.0
J1136+1551	20	5	1.0
J0826+2637	13	20	0.5
J1935+1616	12	158	1.0
J0157+6212	10	30	0.14
J1919+0021	9	90	0.05

Blind pulsar detections

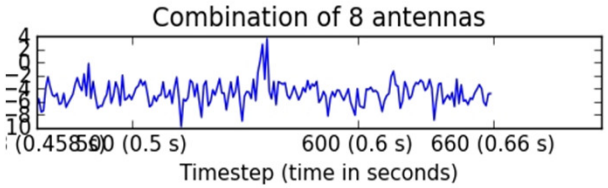


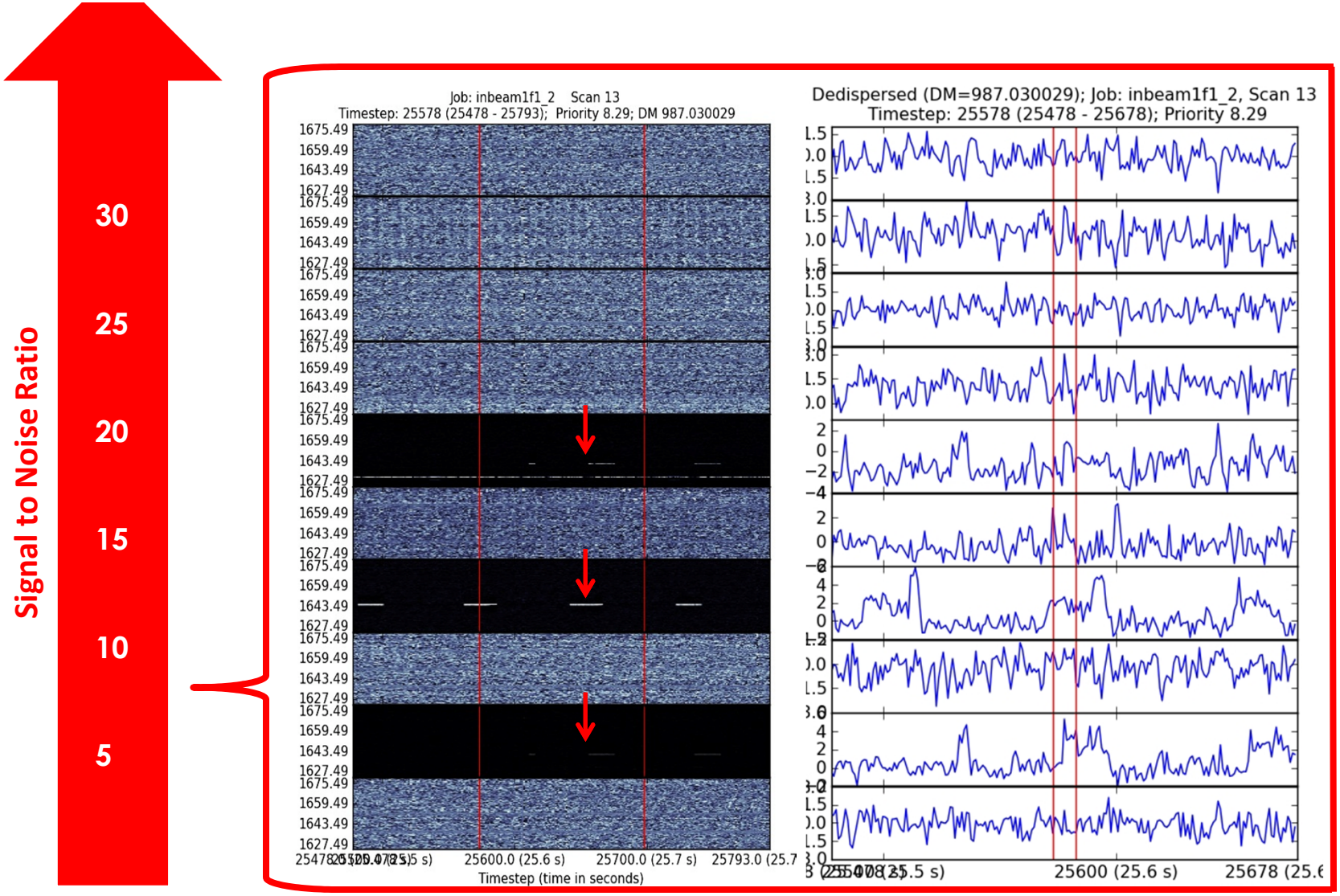
PSR J1919+0021



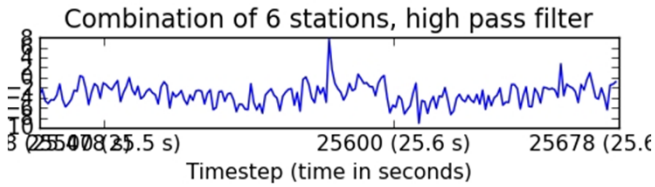


PSR J1935+1616

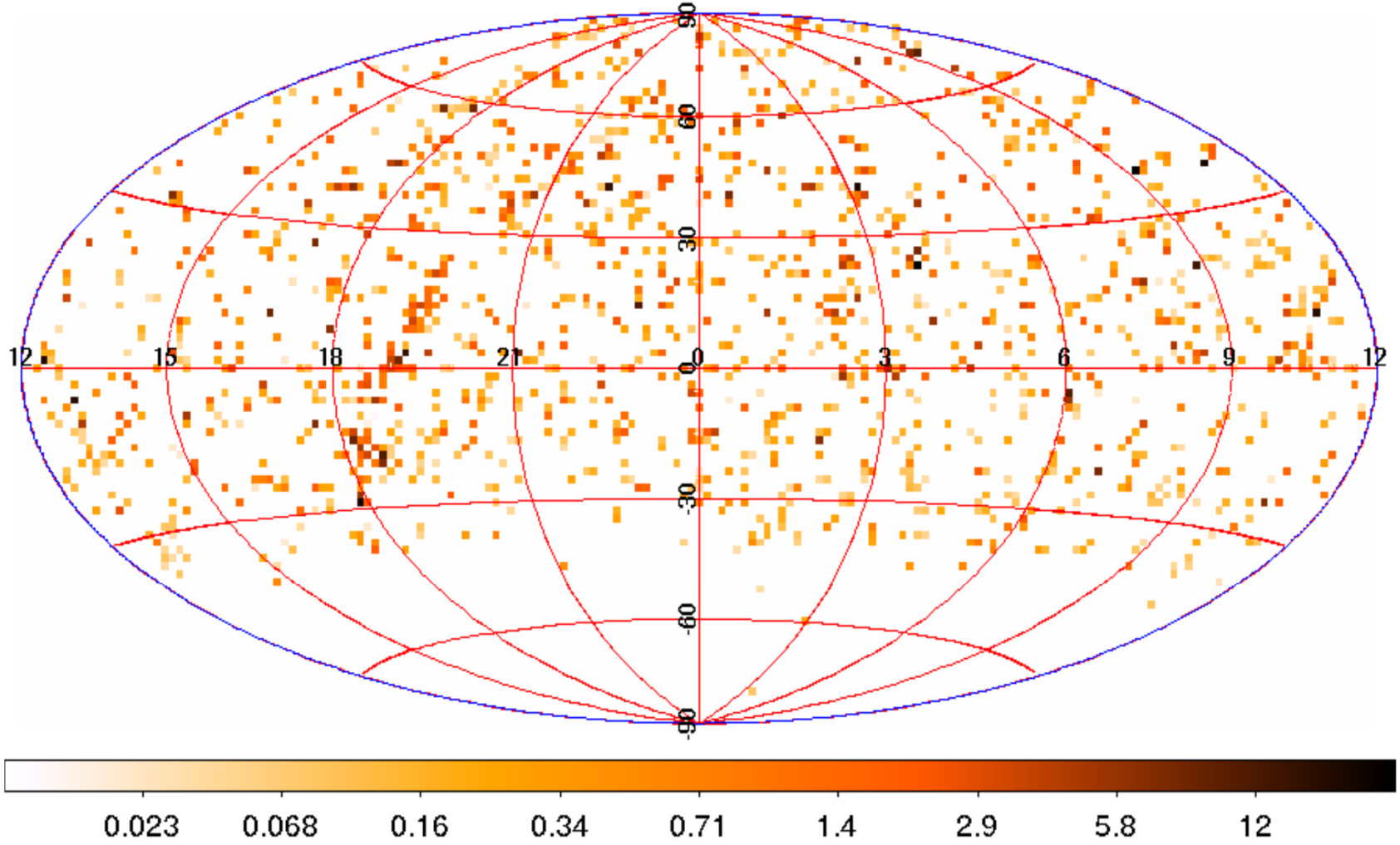


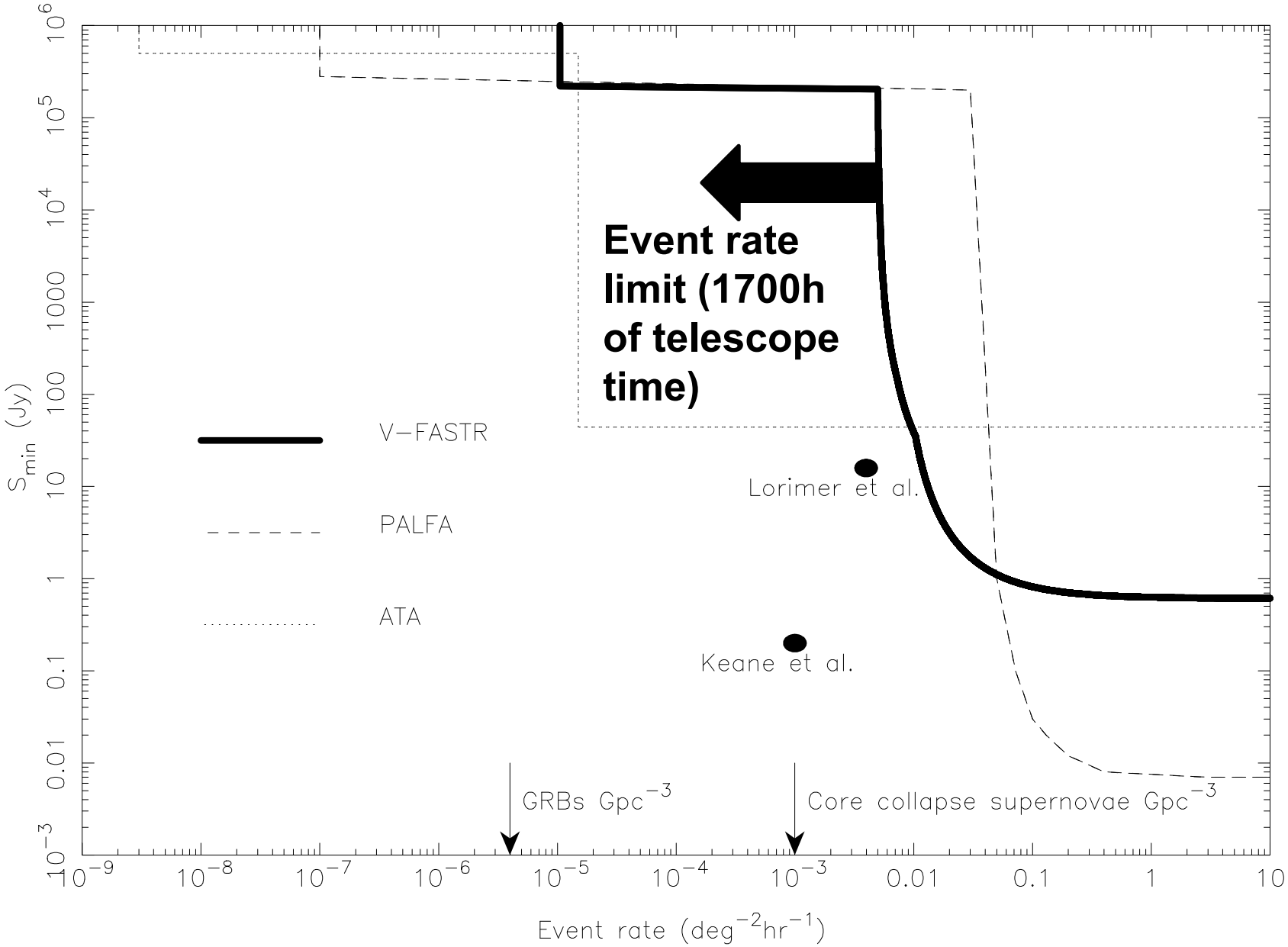

Signal to Noise Ratio
30
25
20
15
10
5
Job: inbeam1f1_2 Scan 13
Timestep: 25578 (25478 - 25793); Priority 8.29; DM 987.030029
Timestep (time in seconds)
Dedispersed (DM=987.030029); Job: inbeam1f1_2, Scan 13
Timestep: 25578 (25478 - 25678); Priority 8.29
25600 (25.6 s)
25700.0 (25.7 s)
25600 (25.6 s)

False alarm!


Combination of 6 stations, high pass filter
25600 (25.6 s)
Timestep (time in seconds)

V-FASTR observations (hours)




Event rate limit (1700h of telescope time)
V−FASTR
PALFA
ATA
Lorimer et al.
Keane et al.
GRBs Gpc^{-3}
Core collapse supernovae Gpc^{-3}
S_{min} (Jy)
Event rate (deg^{-2}hr^{-1})
10^{-3}
0.01
0.1
1
10
100
1000
10^4
10^5
10^6
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
0.01
0.1
1
10

Case II: Have examples of background, but don't know all patterns of interest

Novelty detection looks for outliers

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

BUT… not all statistical outliers are *interesting*

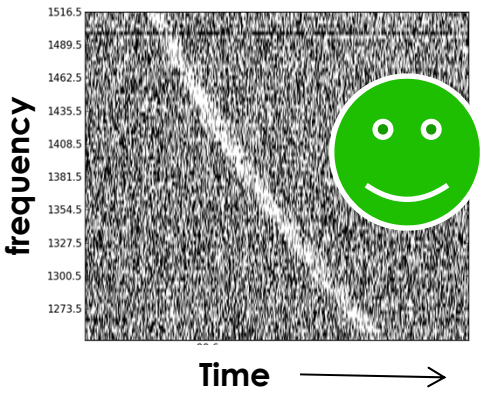
Case II: Have examples of background, but don't know all patterns of interest

Novelty detection looks for outliers

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

BUT… not all statistical outliers are *interesting:*

Pulsar pulse



"Peryton"

Radio Freq. Interference

Parkes Multibeam Survey [Edwards et al., 2001, S. Burke Spolaor 2011] 1.4 GHz, 96 channels, 125 μs samples

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

Semi-Supervised Eigenbasis Novelty Detection (SSEND)

Novelty detection that combines constant-time adaption with robustness to known false alarms

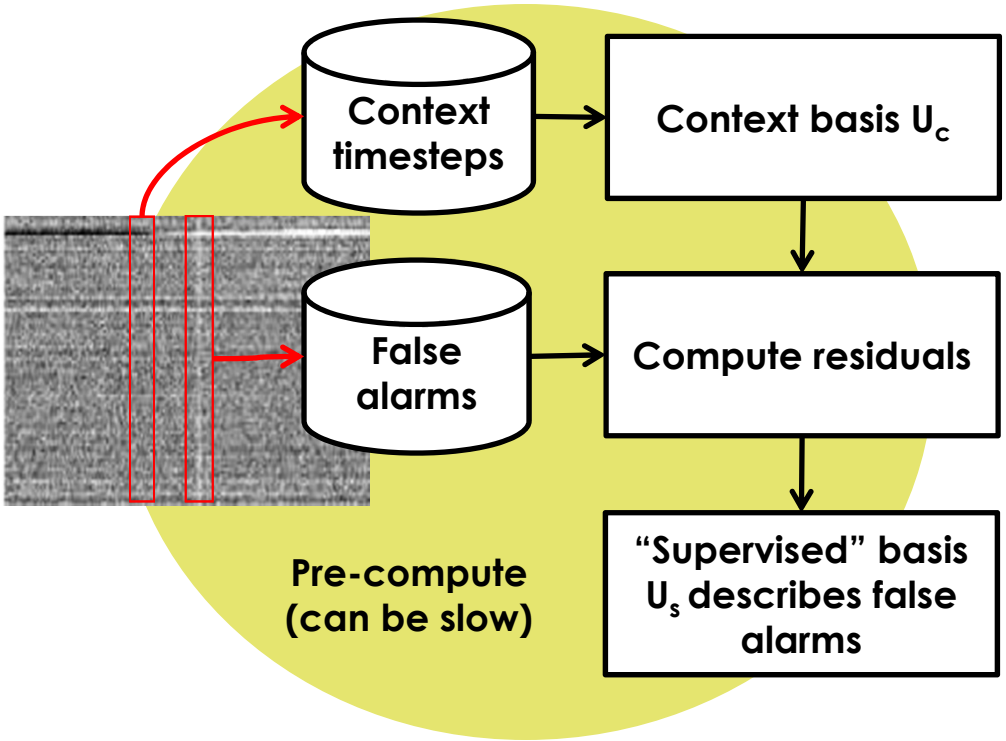
1. **Offline: train a false alarm model (linear subspace)**
2. **Online: learn a model of streaming data [Ross et al. 2004]**
3. **Combine the two, and use *reconstruction error as a novelty score***

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

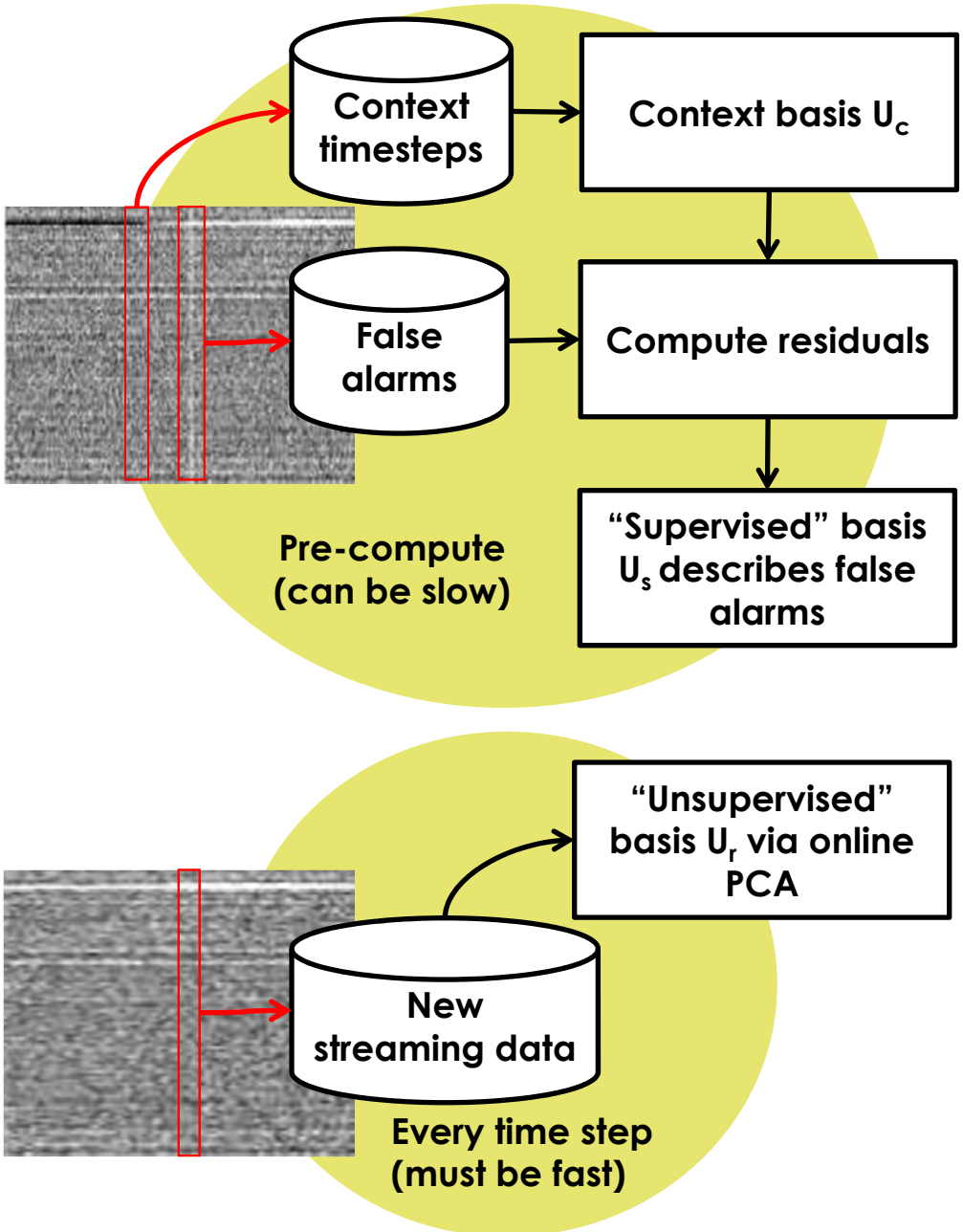
Novelty score — New datapoint — Combined basis

[Thompson et al., 2012]

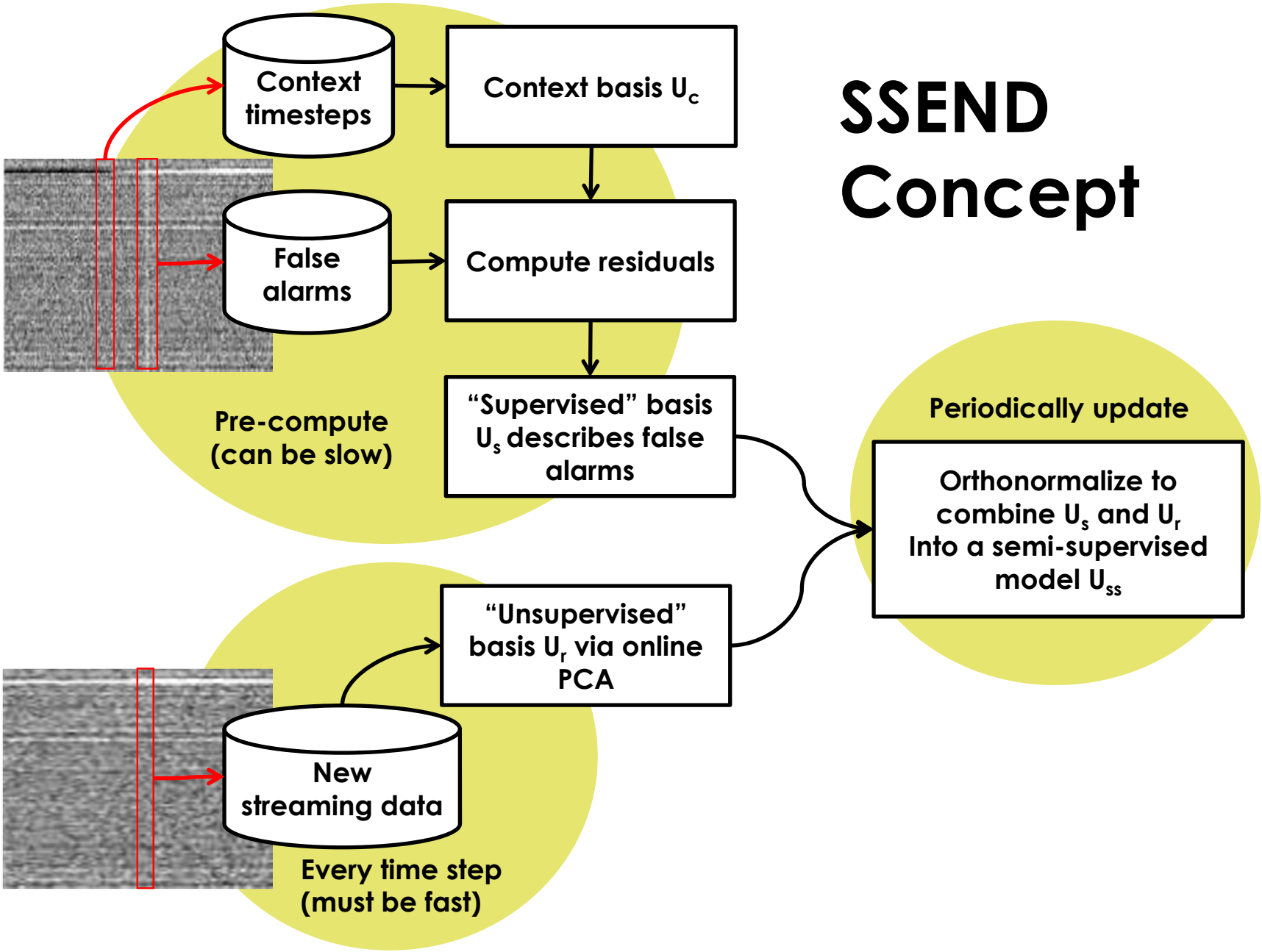
SSEND Concept



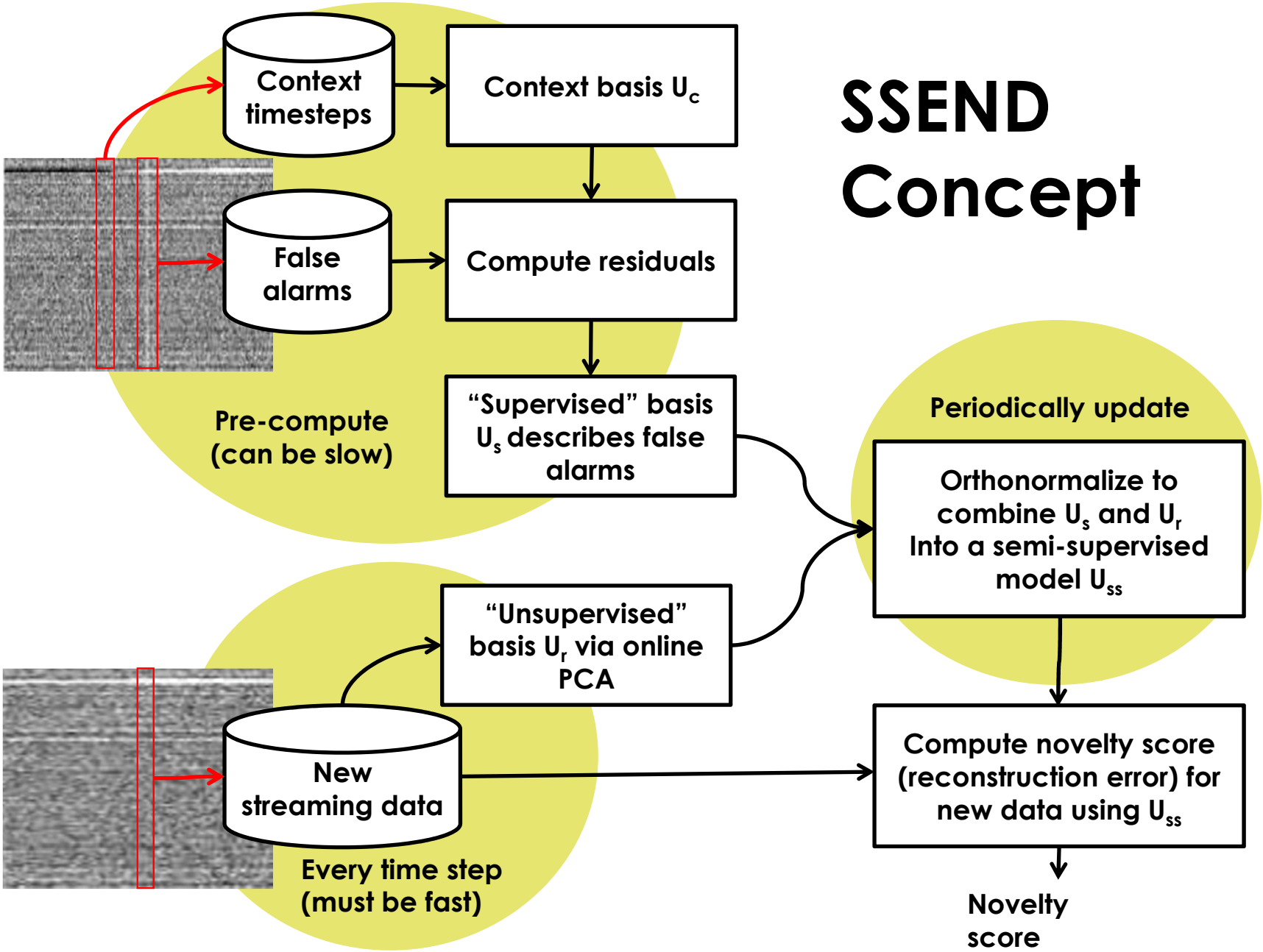
SSEND Concept



SSEND Concept



SSEND Concept



Eigensignals

channel noise

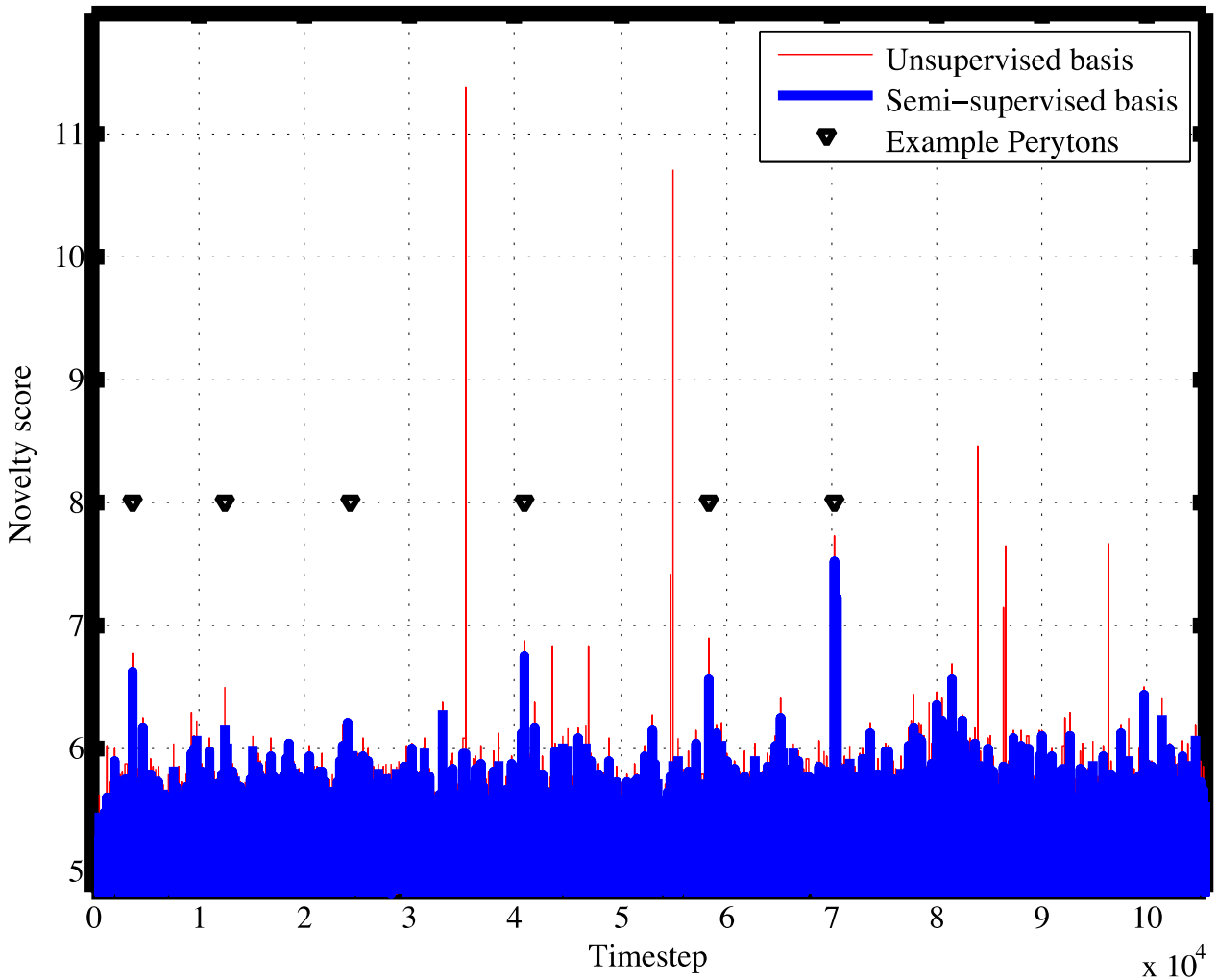
momentary RFI

Adaptive (online)basis vectors

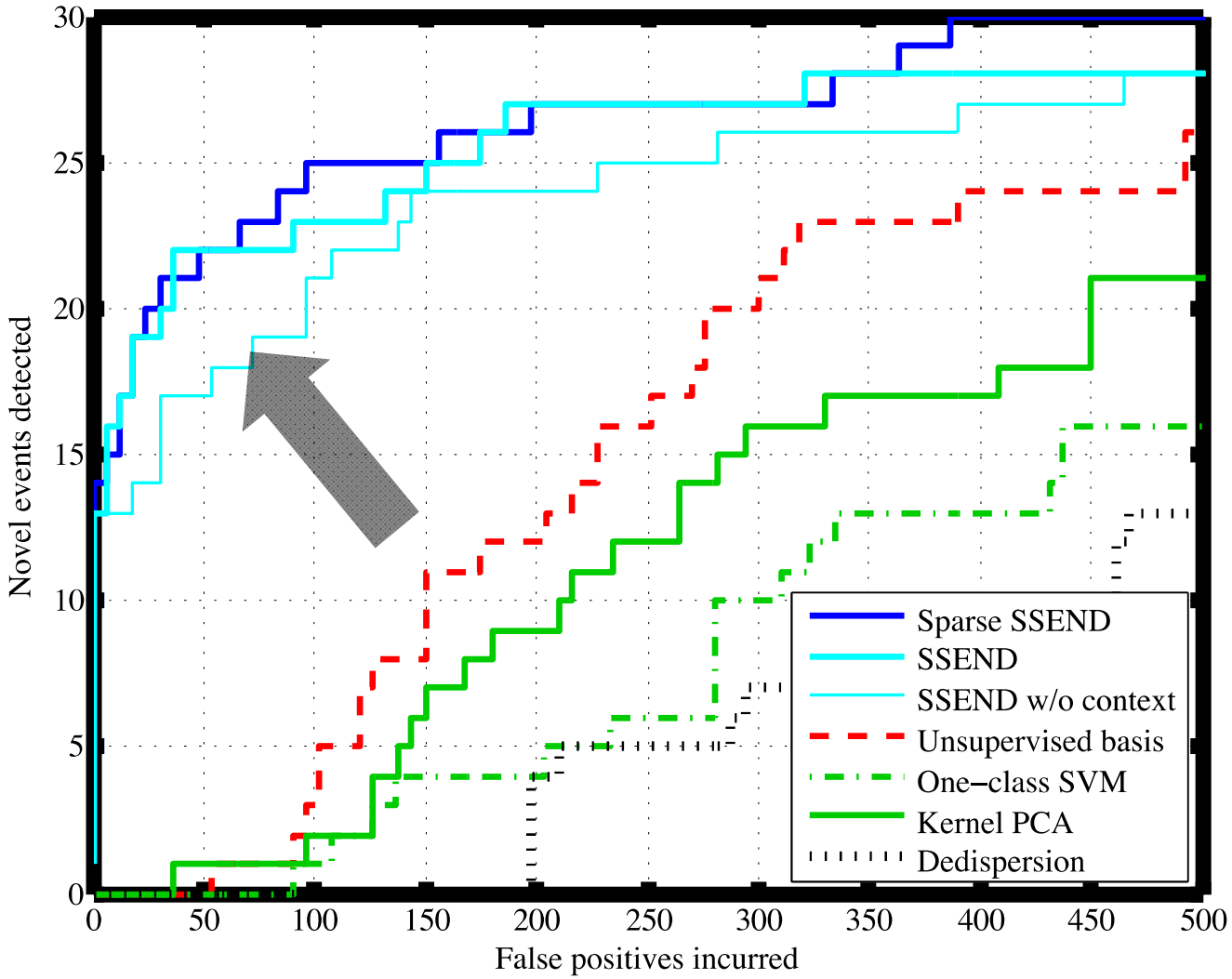
False alarm basis vectors

[Thompson et al., CIDU 2011]

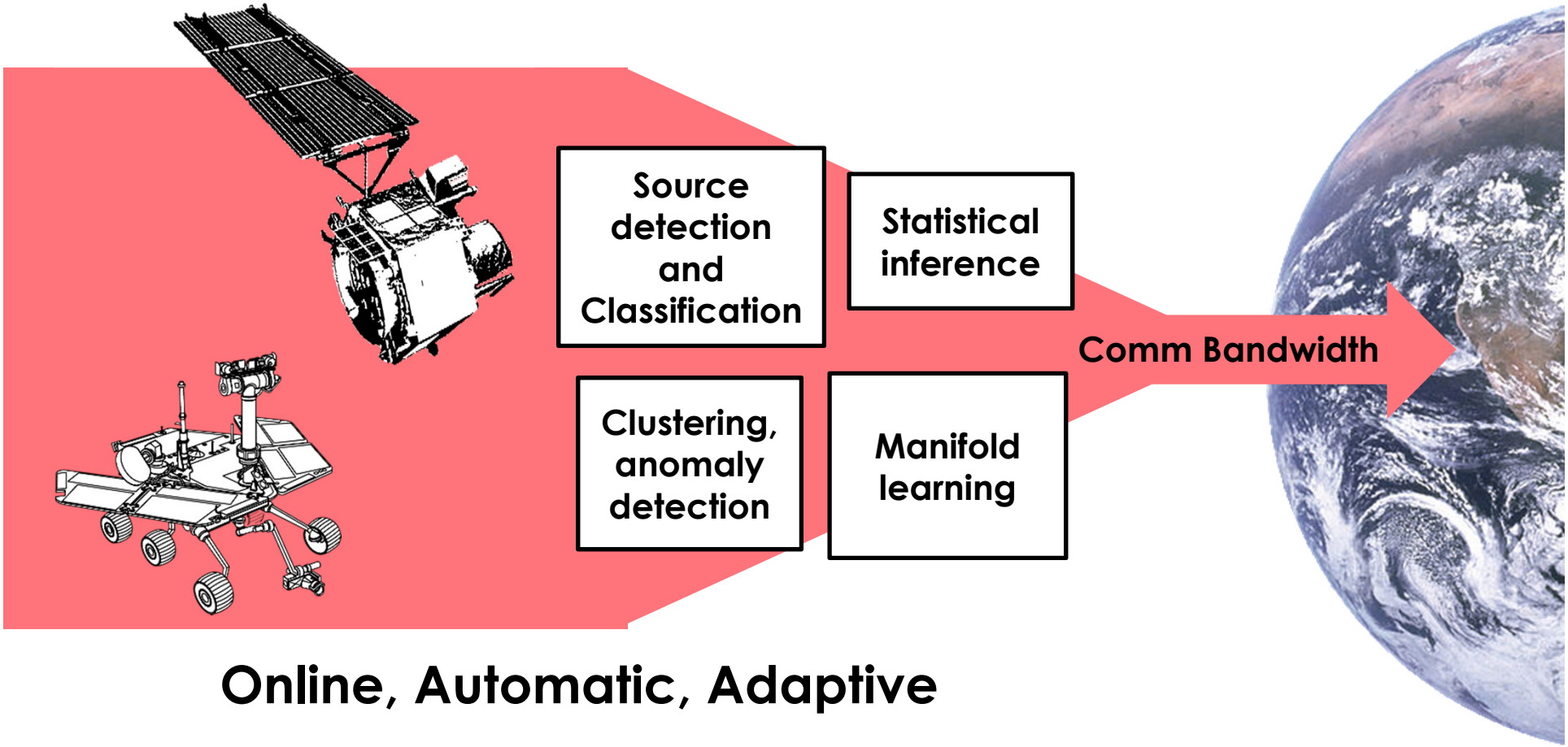
Performance



Performance



A few more applications…



Online, Automatic, Adaptive

Images: NASA / JPL / Caltech

Automatic hyperspectral mapping on EO-1

EO-1 satellite automatically maps features of interest in spectral images

Reduces downlink volumes by >100x

Statistical endmember detection identifies pure minerals and spectral anomalies

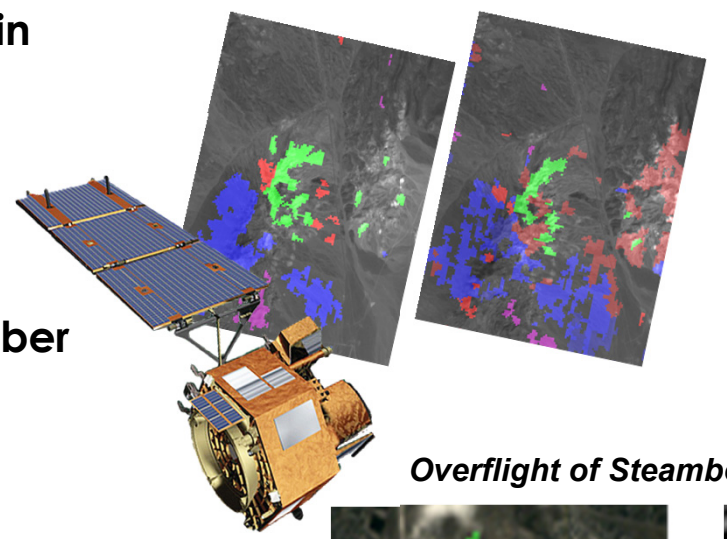
[Thompson et al., in preparation]

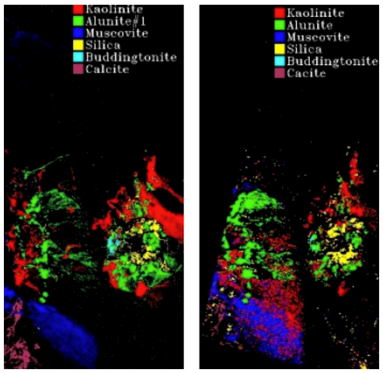
Overflights of Cuprite, NV

Hii-HAT Onboard Sept. 21, 2011

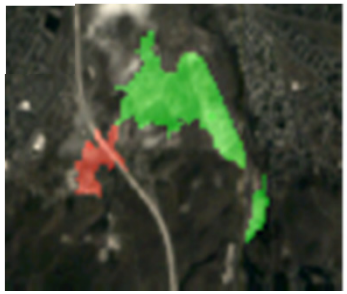
Hii-HAT Onboard Sept. 27, 2011

Maps constructed by Kruse et al. 2003 (via AVIRIS, Hyperion)

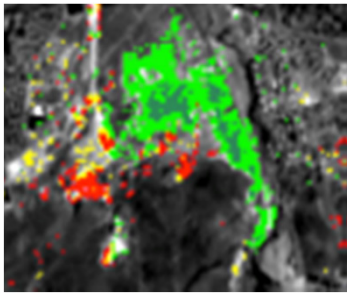




Overflight of Steamboat Springs, NV



Hii-HAT Onboard



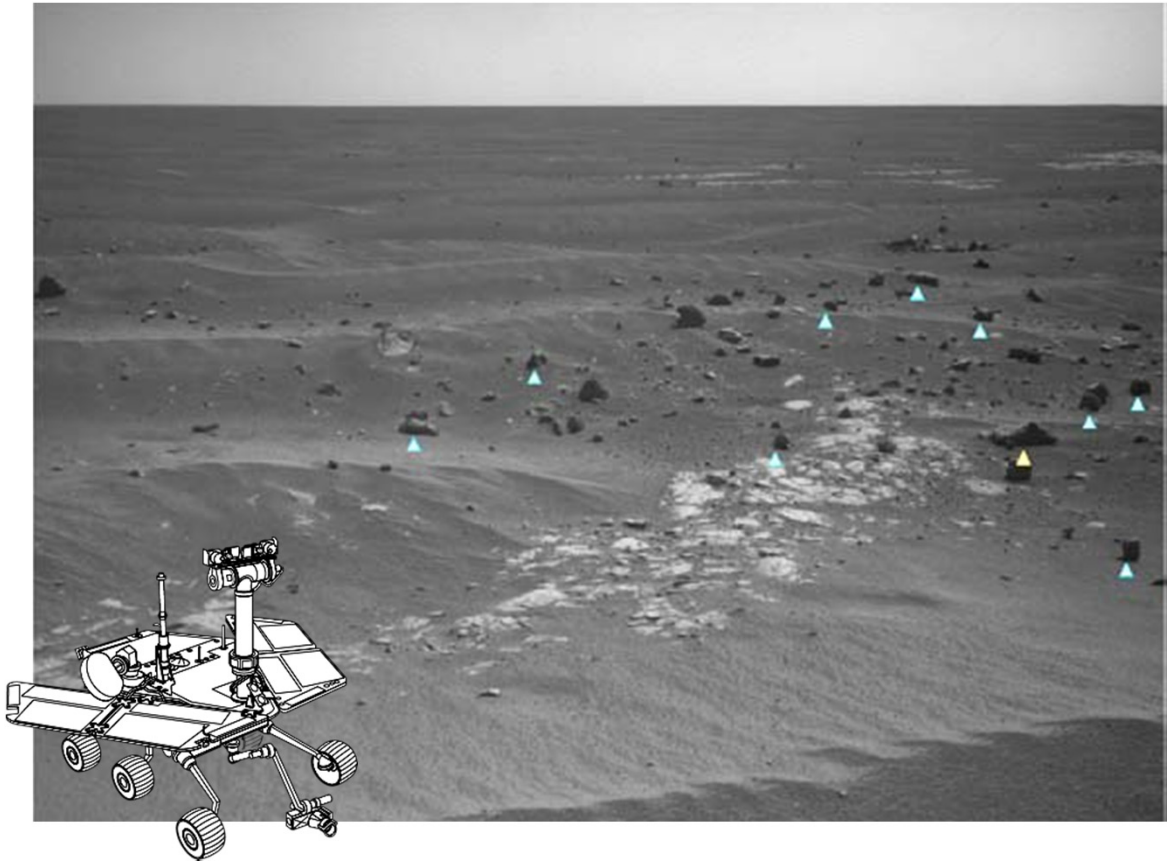
Map by Kruse et al. 2003

Images: IEEE / USGS / GSFC / NASA/ JPL / Caltech

AEGIS target detection

Onboard the Mars Rover *Opportunity*

Meridiani planum, Mars



Automatic data collection from top target matching "dark" profile

Images: Cornell / NASA/ JPL / Caltech

Summary

- Next-generation instruments could benefit from stream mining
  - Machine learning without random access
- Online self-tuning
  - When you can't anticipate observing conditions
  - Automatically train a classifier on the fly
- Semi-supervised anomaly detection
  - Training data distinguishes "statistically anomalous" from "scientifically interesting"

Lessons learned

Further reading

J. S. Bloom , J. W. Richards, Data Mining and Machine Learning in Time Domain Discovery and Classification. In *Advances in Machine Learning and Data Mining for Astronomy 2012*. arXiv:1104.3142v1.

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)

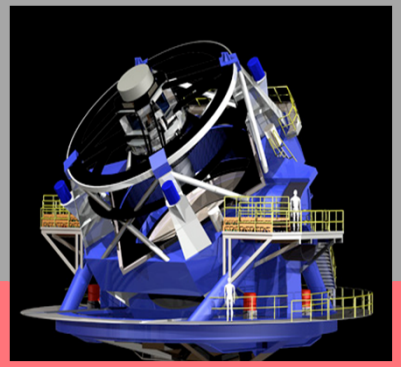
D. R. Thompson, W. A. Majid, C. J. Reed, K. L. Wagstaff, "Semi-Supervised Eigenbasis Novelty Detection," *Statistical Analysis and Data Mining* 2012.

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, Joeseph Lazio, J-P Macquart

Extra slides

Summary



Source detection and Classification

Statistical inference

Clustering, anomaly detection

Manifold learning

Random-access storage

Online, Automatic

Data representation

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

X(t) = channel_noise(t) + false_alarms(t) + anomaly(t)

slowly varying — intermittent

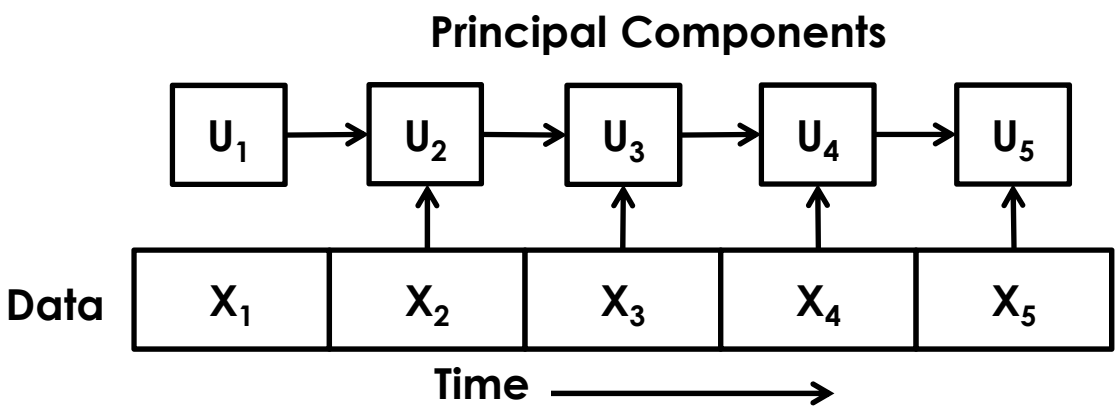
We can model the channel noise, but how do we avoid known false alarms?

Online Updates

Principal components via 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:

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