Multi-View Relationships for Analytics and Inference

Eric Lei

Thesis Defense

July 30, 2019

Committee:

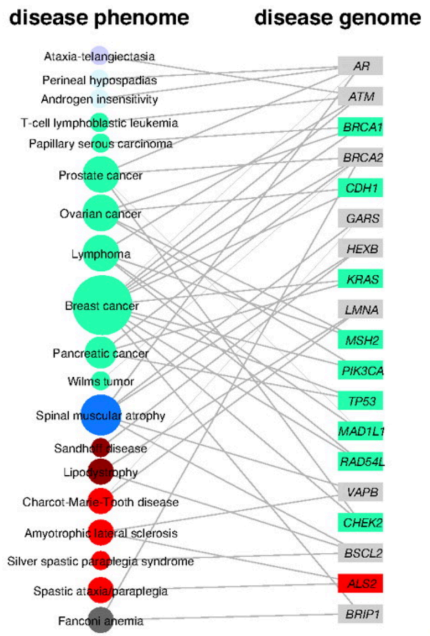
Artur Dubrawski (chair), Barnabas Poczos, Mario Berges, Simon Labov [LLNL]

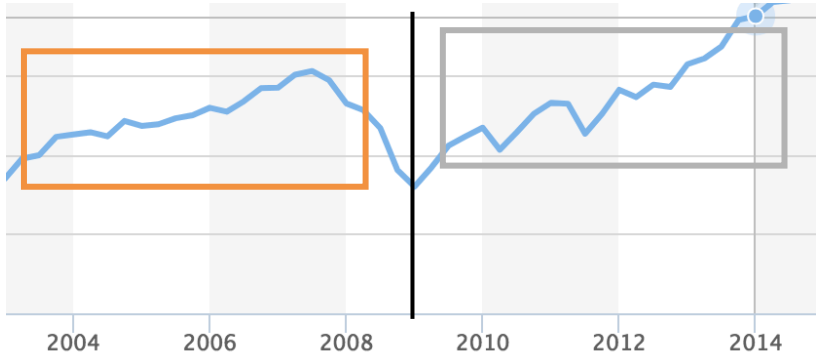
Carnegie Mellon University
School of Computer Science
ML MACHINE LEARNING DEPARTMENT
Auton Lab

Multi-view data

- Features partitioned into multiple sets: *views*
- Latent variables govern the relationship between views
- Multi-view learning finds agreement between views
- Our work <u>explicitly</u> leverages the relationship between views as a unit of analysis

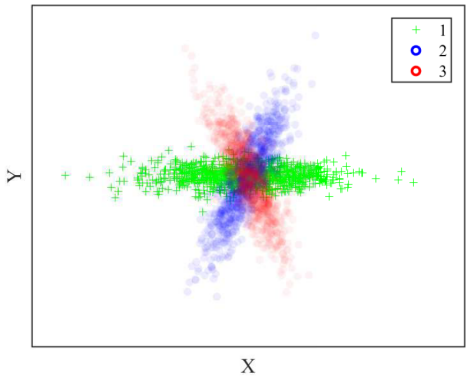
The man at bat readies to swing at the pitch while the umpire looks on.



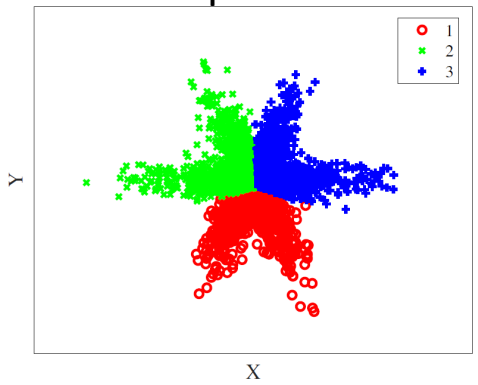


Characterization of multi-view relationships

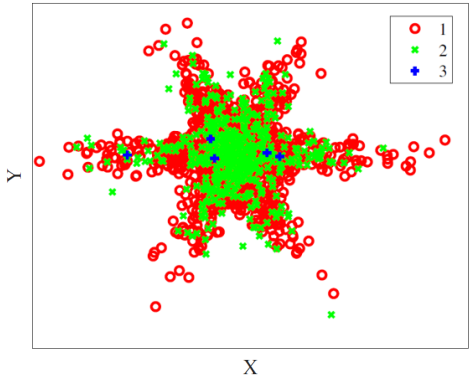
Ground truth



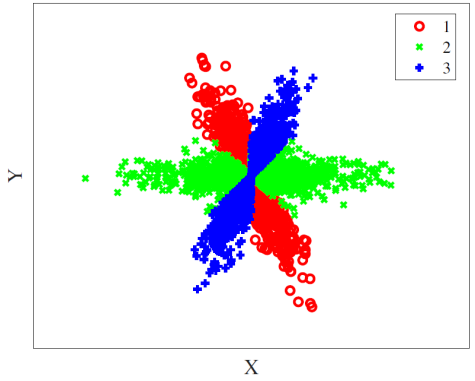
K-means or spectral clustering



Modern multi-view clustering [Zhao 2017]



Clustering aware of multi-view relationships



Zhao et al. (2017). Multi-view clustering via deep matrix factorization. AAAI.

Current multi-view learning

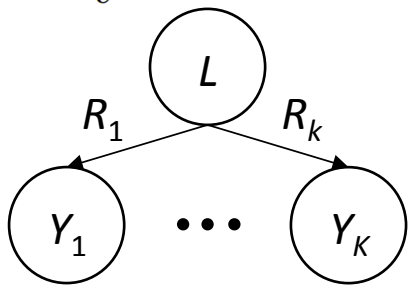
- Regularize single-view solutions $V^{(v)}$ toward each other [Liu 2013]
- Fuse single-view solutions [Greene 2009]
- Apply single-view algorithms in common subspace by learning shared projection Z [Gao 2015]
- **This work:** infer latent variables L that affect all views Y_j
  - L often interpretable
  - Analysis of L and relations R_j

$$\sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$

$$\mathbf{X} \approx \mathbf{PH} \quad \text{such that} \quad \mathbf{P} \geq 0\,, \mathbf{H} \geq 0$$

Stacked solutions from views

$$\lambda \sum_v ||Z - Z_v||$$



Liu et al. (2013). Multi-view clustering via joint nonnegative matrix factorization. ICDM.
Gao et al. (2015). Multi-view subspace clustering. ICCV.
Greene et al. (2009). A matrix factorization approach for integrating multiple data views. ECML PKDD.

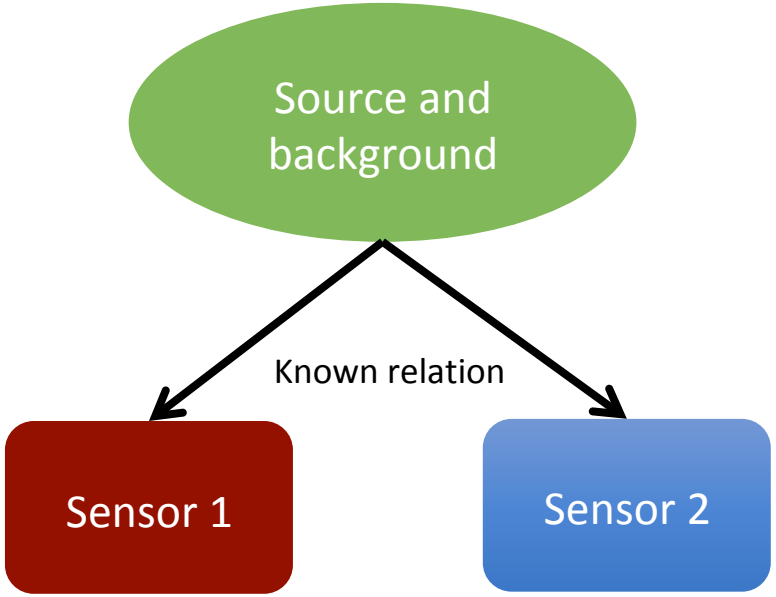
Analysis of multi-view relationships

- **Thesis statement**: It is possible to characterize multi-view relationships and employ them as units of analysis in descriptive analytics and inference

- Present novel methods that characterize multi-view relationships, either using domain knowledge or by learning from data, and employ them as units of analysis
- Reveal structure that alternative methods do not or have competitive empirical performance with the state of the art

Outline

- **Multi-view filtering**
  - Single sensor method for gamma source detection [NSS 2017]
  - Multiple sensor extension
- Learning multi-view relationships
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]



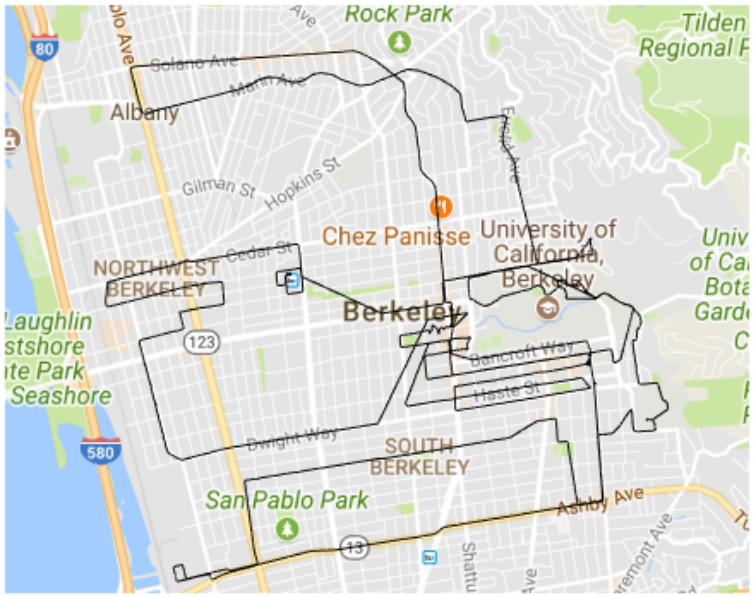
Lei et al. (2016). Radiological threat detection for an unknown energy window by canonical correlation analysis. NSS.
Lei et al. (2017). Robust detection of radiation threat by simultaneous estimation of source intensity and background. NSS.
Lei et al. (2017). Bleeding detection by multi-view correlation clustering of central venous pressure. MLHC.
Lei et al. (2019). Characterization of multi-view hemodynamic data by learning mixtures of multi-output regressors. ISICEM.

Multi-view filtering for gamma source detection

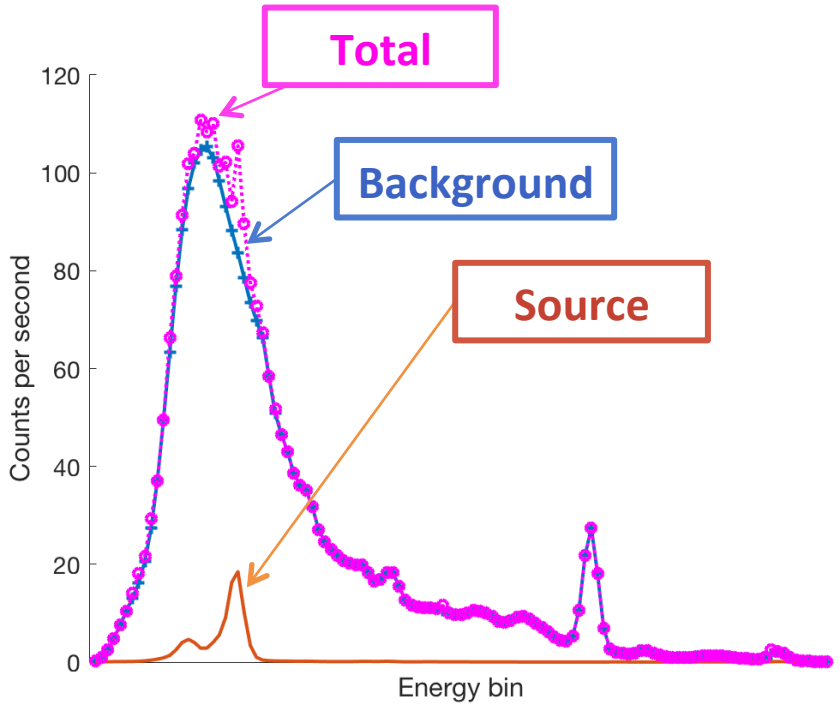
- Multiple sensors = multiple views
- How do we leverage multi-view relationships known through domain knowledge?
- Infer latent variables by collectively filtering views
- Reduce dependence on training data

Truck and Pedestrian carry sensors.
Other Objects are possible source locations.

Challenges of gamma source detection



Mobile sensor collects photon spectra.



Source affects distribution of photon counts.

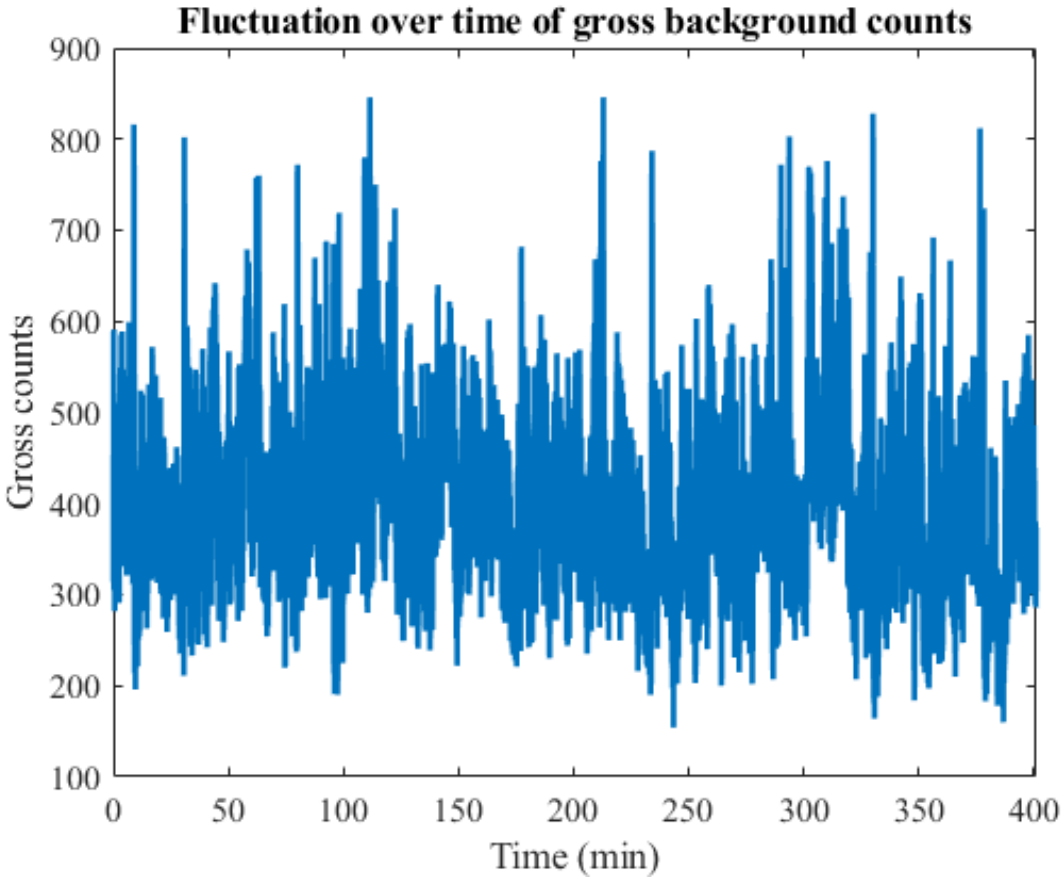
- Unknown background
- Noise
- Hidden sources
- Low signal-to-noise ratio (SNR)

Outline

- Multi-view filtering
  - **Single sensor method for gamma source detection** [NSS 2017]
  - Multiple sensor extension
- Learning multi-view relationships
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]

Single sensor methods

- Stationary characterizations of background from training
  - Matched Filter [Turin 1960]
  - Spectral Anomaly Detection [Nelson 2012]
  - Gaussian-Poisson MAP [Huggins 2014]
- Assume known source type
- Do not want to depend on training data
  - Robustness to unknown background variation
  - Practical utility



Turin (1960). An introduction to matched filters. IRE Transactions on Information theory.
Nelson and Labov (2012). Aggregation of mobile data. LLNL Technical Report.
Huggins et al. (2014). Using Gaussian rate priors with Poisson data likelihoods for improved detection of sources of known types in cluttered background scenes. NSS.

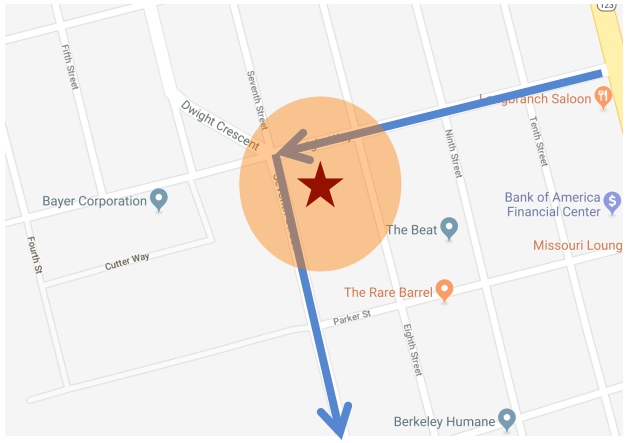
Existing work: State-of-the-art adaptive detector

- Orthonormal Subspace Projection Matched Filter with adaptive background basis (RDAK) [Labov 2019]
  - Residual from spectrum *X* onto background basis *B*
  - Similarity with template *S*
- Warm-up for basis
  - 5 min. minimum
  - 10 min. optimal
- Long warm-up may be infeasible
  - Source contamination
  - No time
  - Warm-up and test mismatch

$$DM(X) = \frac{S^t W(I - B(B^t W B)^{-1} B^t W)X}{k\sqrt{|X|_1}} = \frac{TX}{\sqrt{|X|_1}}$$

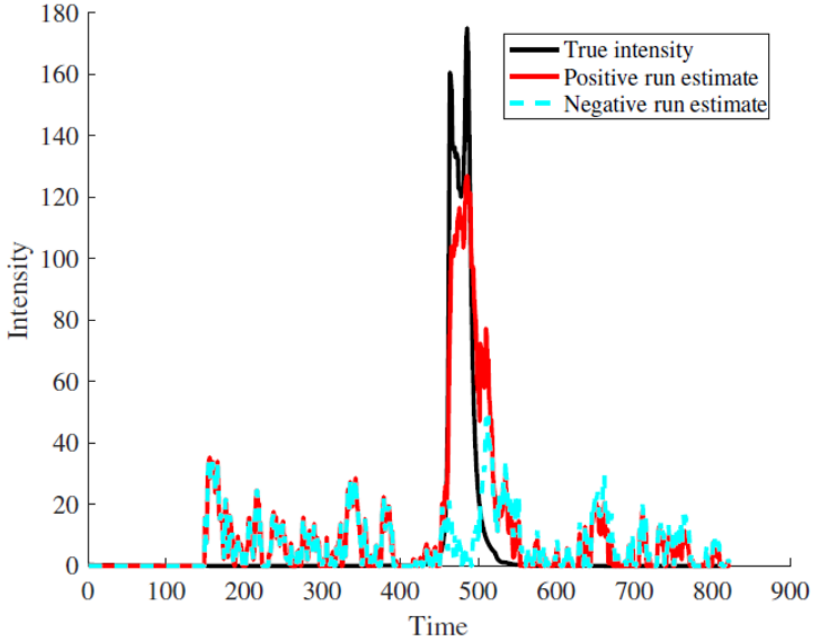
Labov and Nelson (2019). Private communication.

Idea: Simultaneously estimate source intensity and background photon count rates

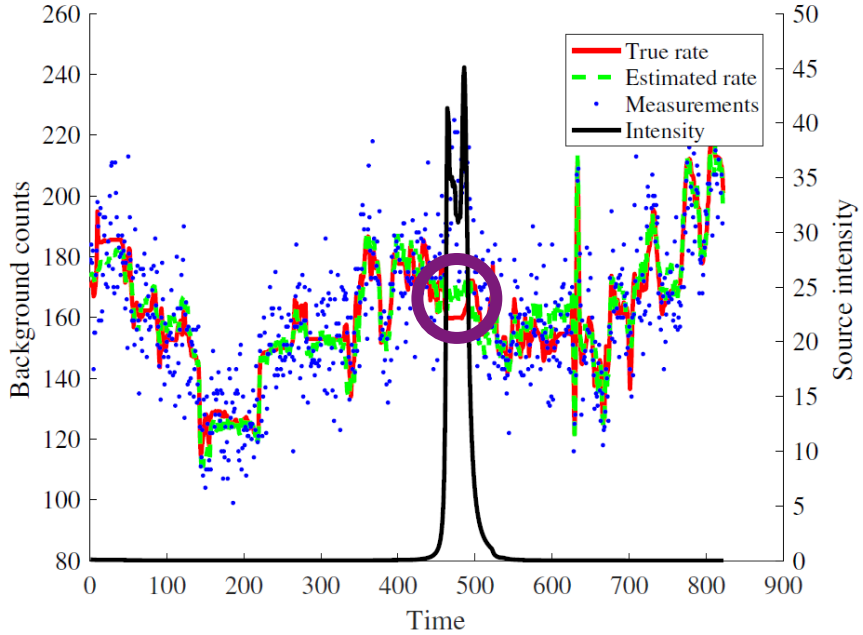


Sensor passes source.

Source intensity



Background rate



Adaptive and less dependent on warm-up or training

Exploits smoothness

Reduced dependence on warm-up or training via Kalman Filter

- Radiation as linear dynamical system
  - State x = <background rates λ, source intensity γ>
  - Observation y = <observed spectrum>

$$x_{t+1} = x_t + w_t$$
$$y_t = \lambda_t + \gamma_t s + v_t$$

- Kalman Filter (KF) estimates mean and covariance of state [Kalman 1960]
- Linear minimum MSE:

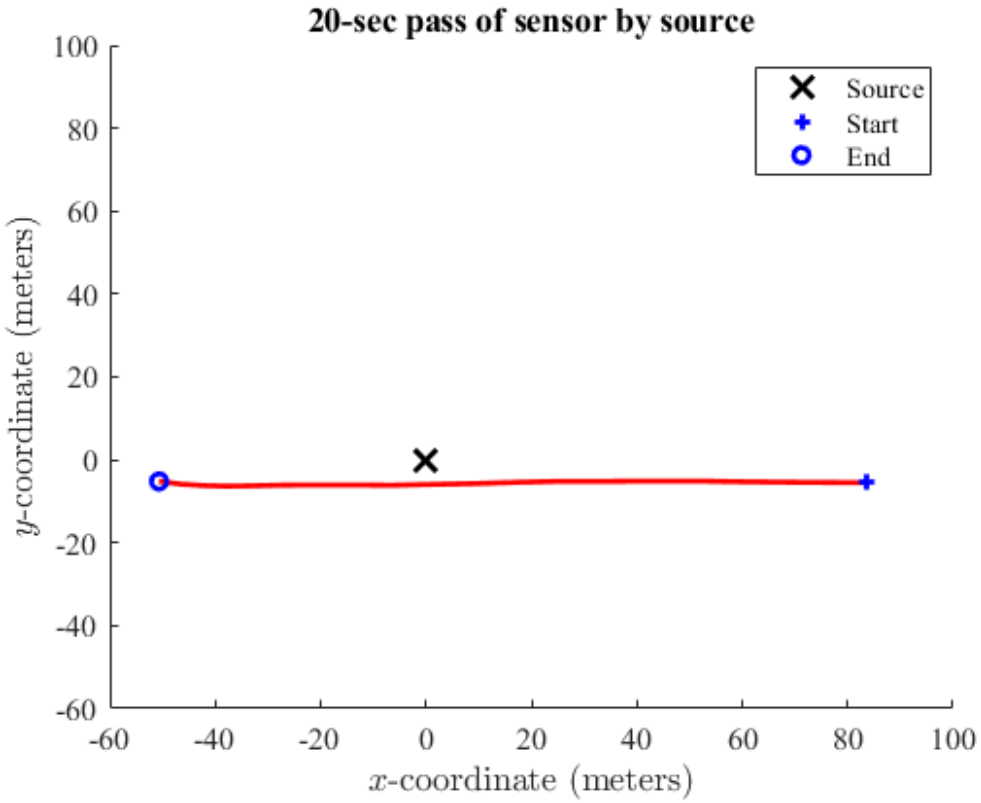
$$w_t \overset{iid}{\sim} N(0, Q_t) \wedge v_t \overset{iid}{\sim} N(0, R_t)$$
$$\Longrightarrow E\left[(\hat{x}_{t,KF} - x_t)^2\right] \leq E\left[(\hat{x}_{t,Linear} - x_t)^2\right]$$

  - Often satisfied because Poisson(λ) → N(λ, λ) as $\lambda \to \infty$
- Warm-up can be about 1 min

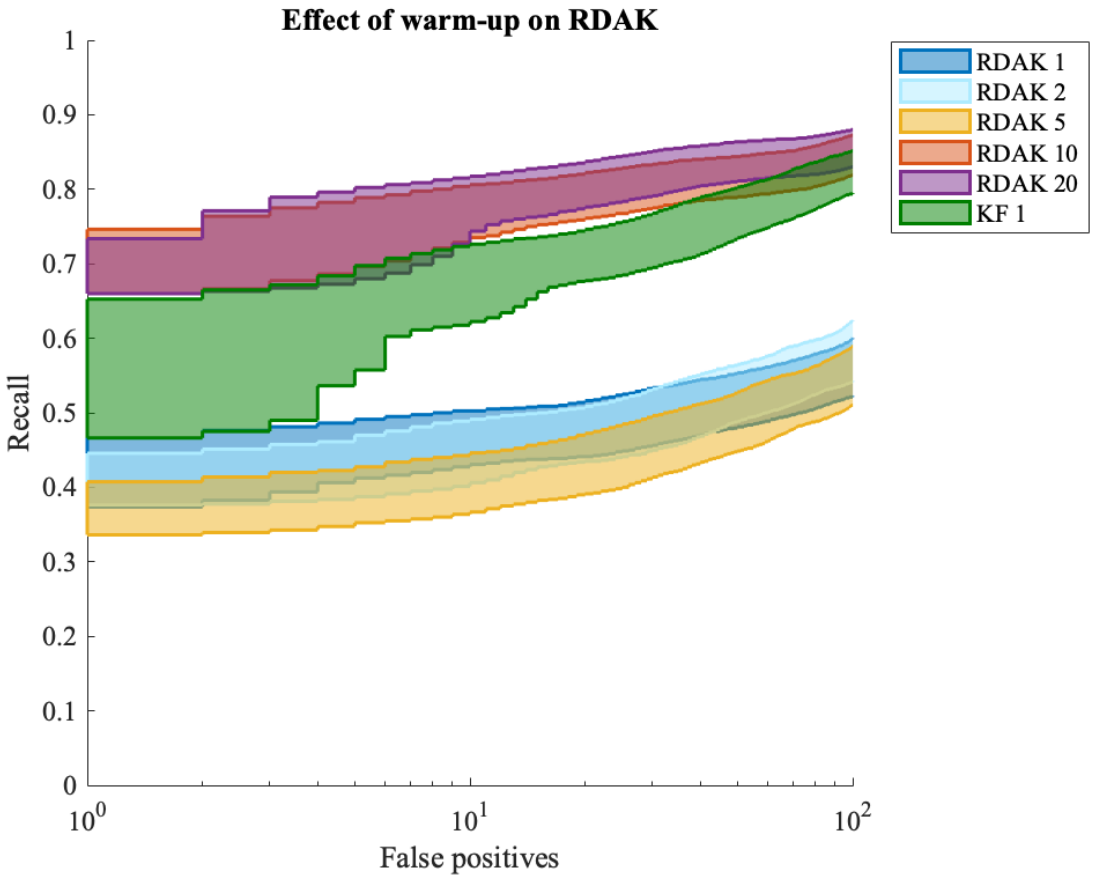
Kalman (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering.

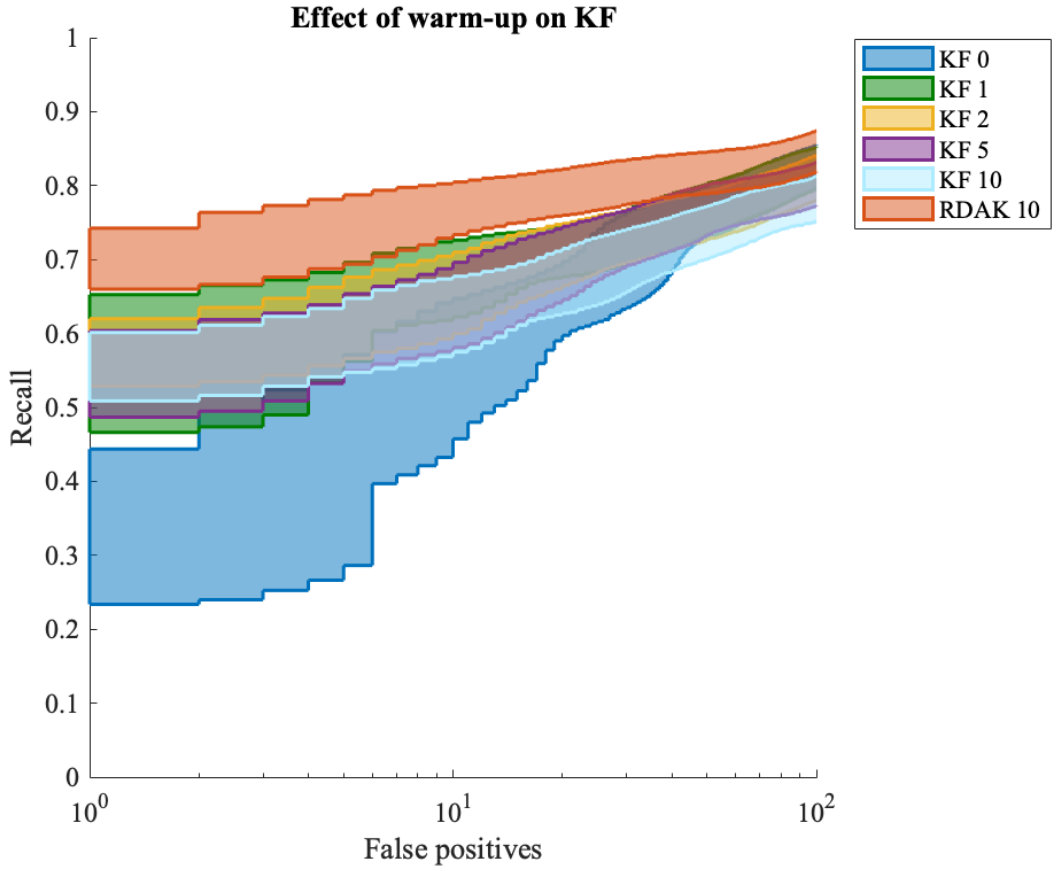
Dataset and experimental design

- Spectroscopic data from a major metropolitan area [not publishable]
- Synthetic injections of industrial isotope at SNR usually in [1, 10]
- Detection of a roadside source in a single pass
- 10 sec before and after
- 20,000 passes
- Disjoint data for warm-up



Our method performs better at short warm-up periods

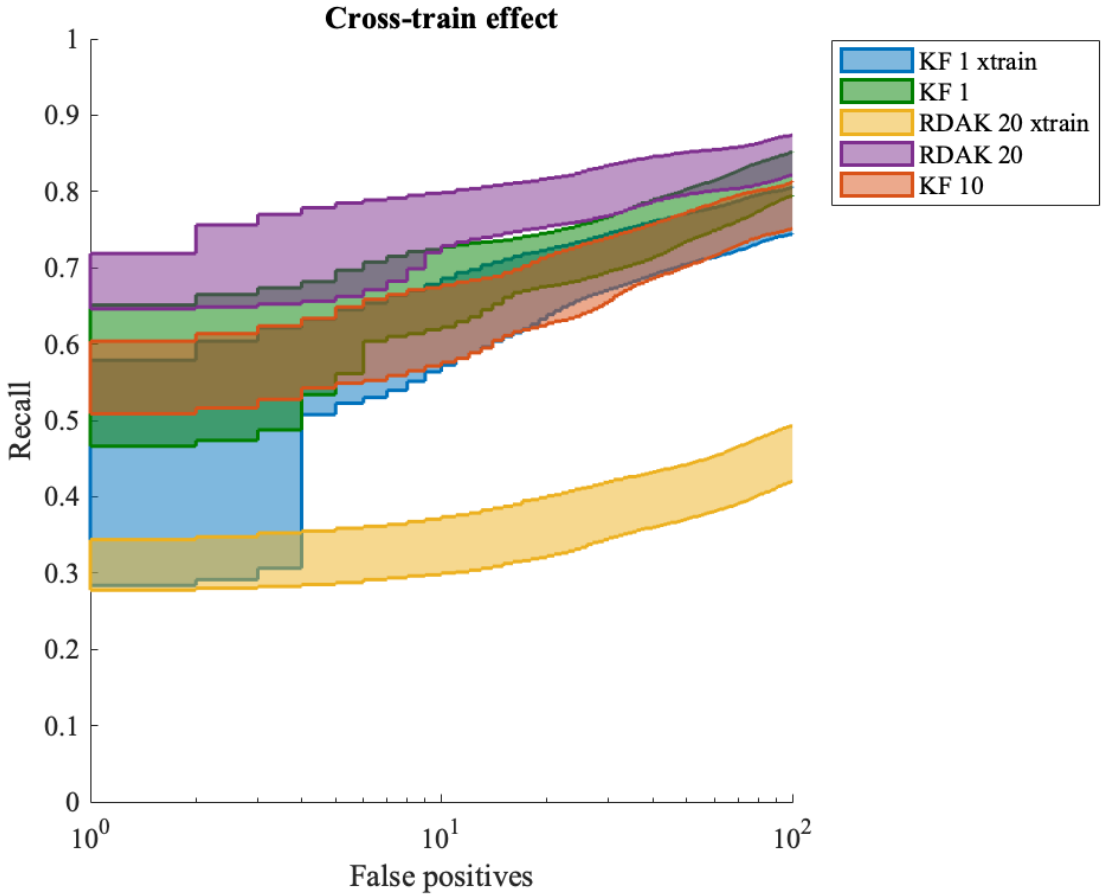




RDAK degrades with 5 min. or less warm-up. KF is indifferent to the amount of warm-up after 1 min.

Our method performs better when warm-up differs from test

- Mismatch between warm-up and test
- Warm-up from different sensor in a different area
- RDAK not designed for this and degrades substantially
- Our method is indifferent and performs much better

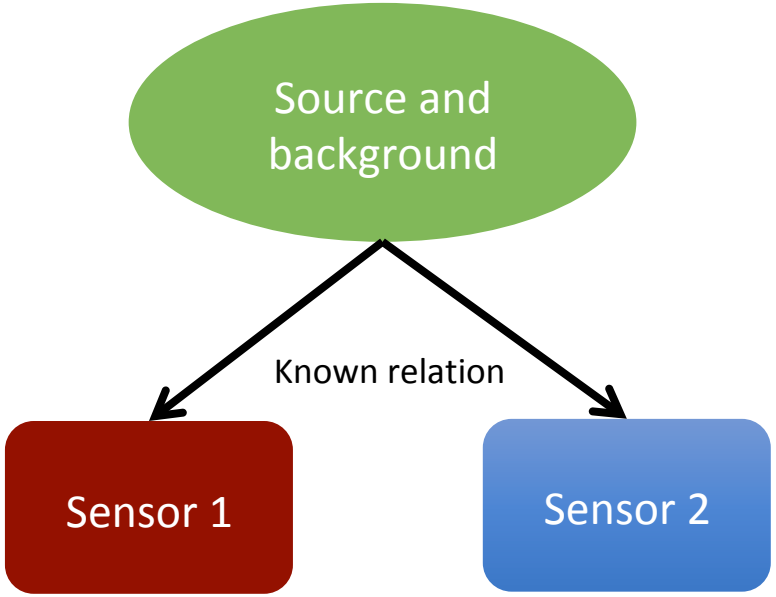


Key takeaway of single sensor problem

Exploit smoothness to adapt to background with little training, which is useful in many practical scenarios

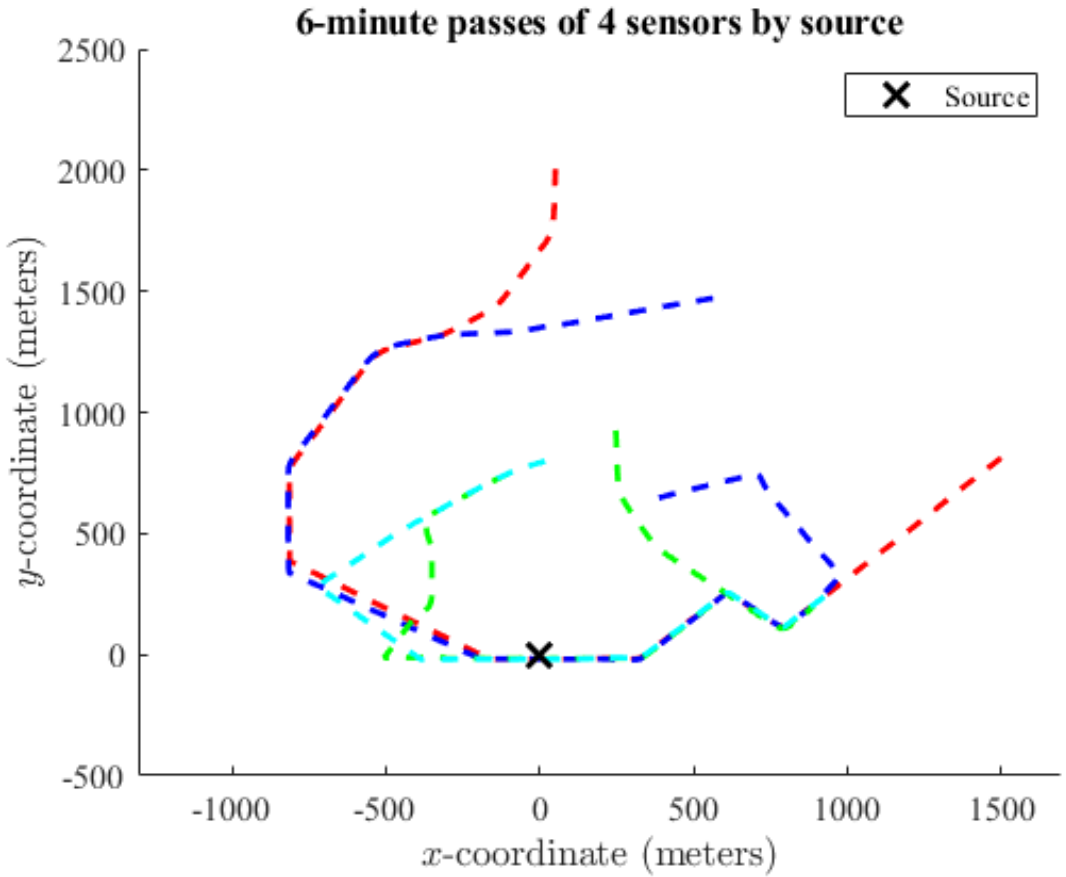
Outline

- Multi-view filtering
  - Single sensor method for gamma source detection [NSS 2017]
  - **Multiple sensor extension**
- Learning multi-view relationships
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]



Multiple sensors

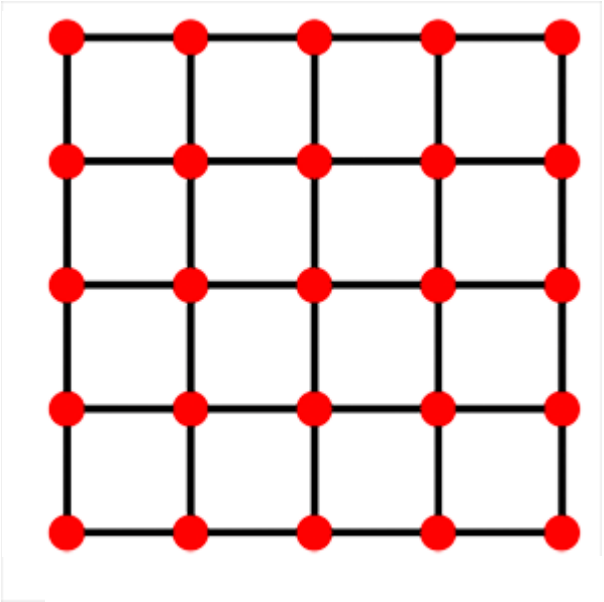
- More than one sensor can be near the source simultaneously
- Simultaneous sensors are related through background and possible source
- How can this contemporaneous multi-view relationship improve inferences?



Related work: Bayesian Aggregation

- Bayesian Aggregation (BA): State-of-the-art method for multiple sensors [Tandon 2016]
- Tracks probability of H_0 and $H_{k,L}$
  - H_0: No source
  - $H_{k,L}$: Source is present with intensity I_k at location L (and possibly other characteristics)
- Requires reference data to train detector to get scores x_i and to fit $\Pr(x_i|H_{k,L})$
- Assumes independence; does not leverage contemporaneous multi-view structure

$$\max_{k,L} \frac{\prod_{i=1}^{T} \Pr(x_i|H_{k,L})}{\prod_{i=1}^{T} \Pr(x_i|H_{0,L})}$$

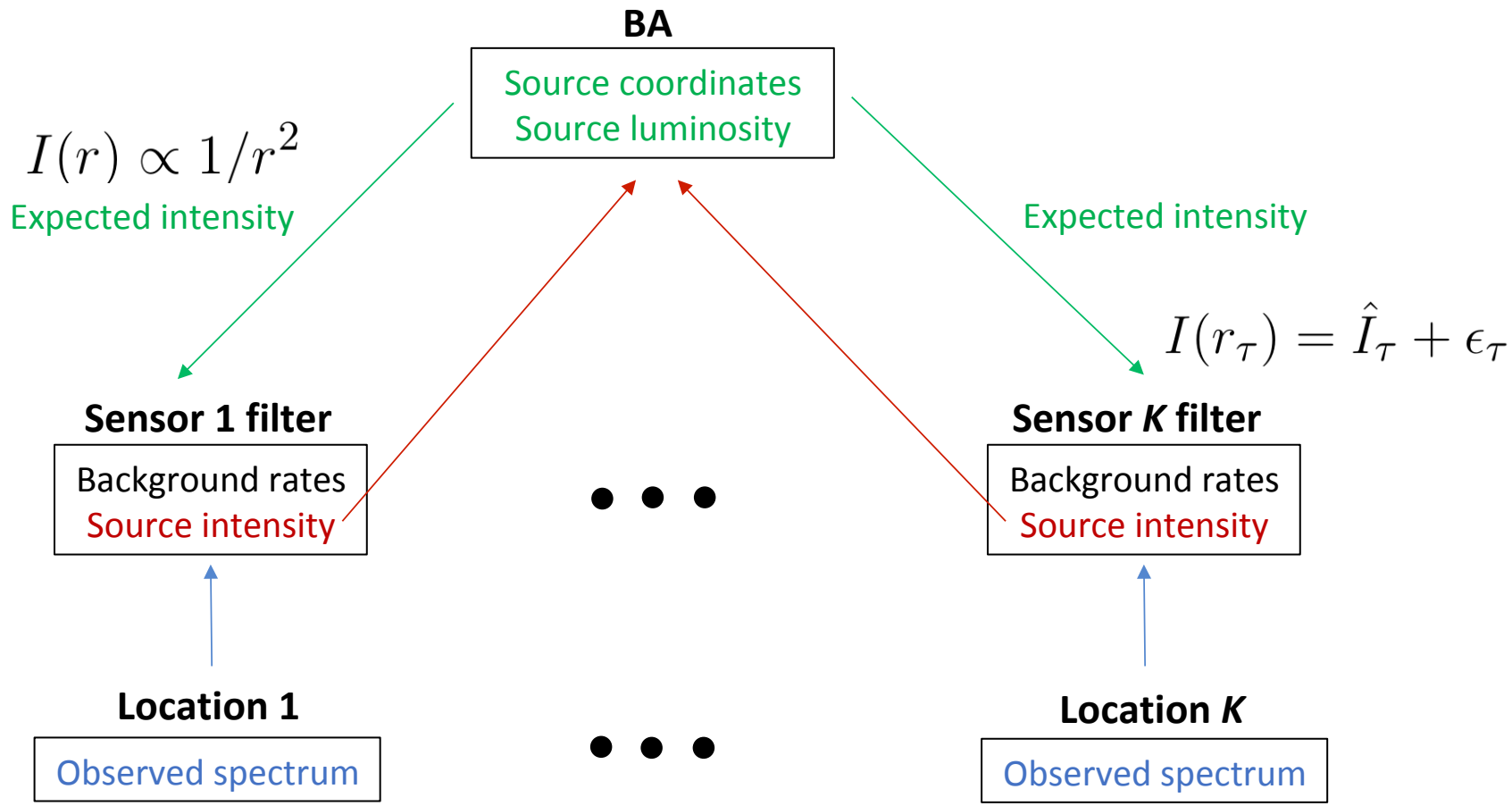


Tandon (2016). Bayesian aggregation of evidence for detection and characterization of patterns in multiple noisy observations. AI Matters.

Our approach: Multi-view filtering

- Collectively filters the inferences from individual sensors in the Bayesian Aggregation Filter (BAF)
- **Reduces dependence on training:** Kalman Filter (KF) at each sensor to bypass detector training
- **Improves detection power over BA:** Share information between sensors using the collective inference of all sensors from the previous time step

Main idea: BA is hub for filters



Dataset and experimental design

Previous experiment

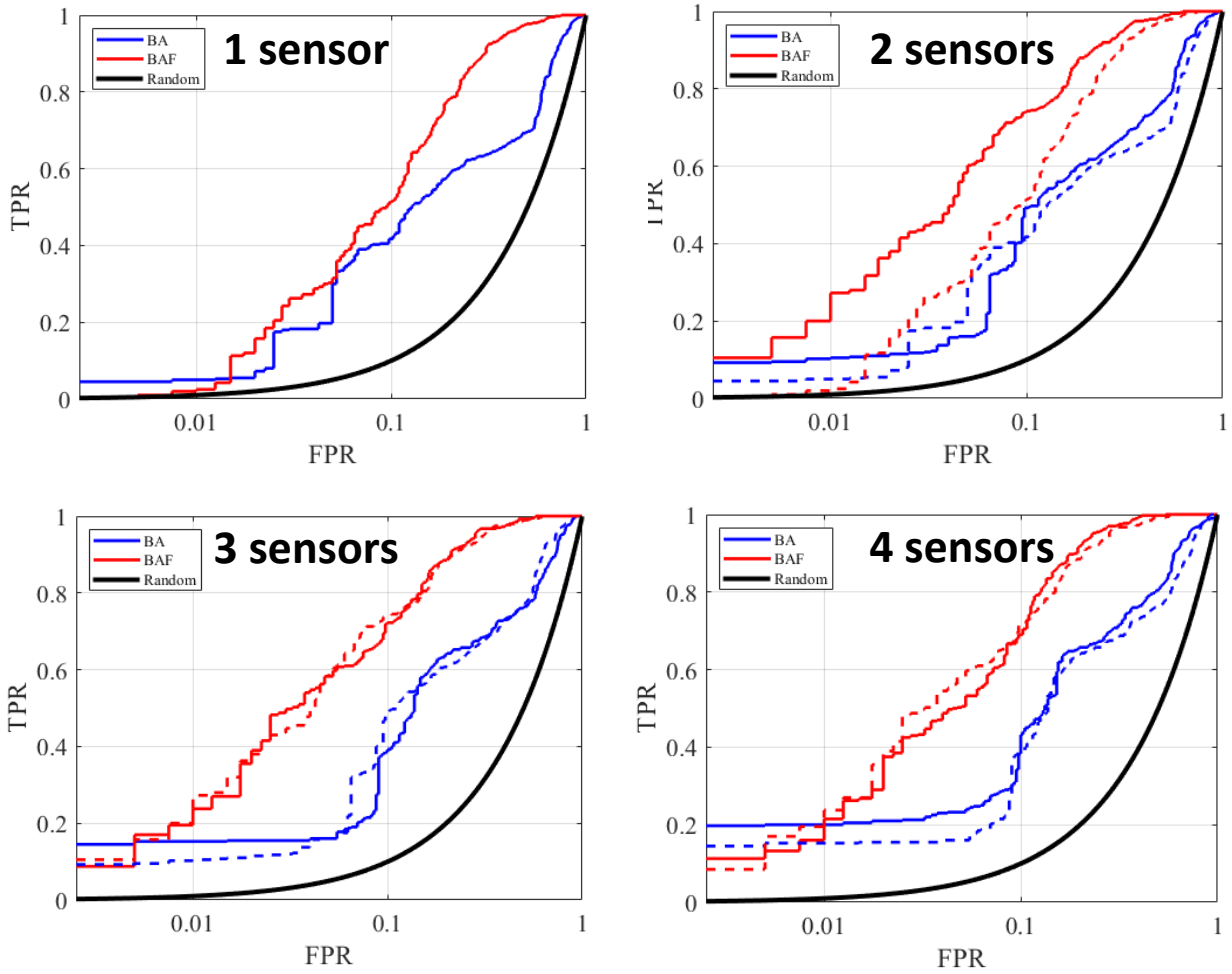
- Spectroscopic data from a major metropolitan area [not publishable]
- Synthetic injections of industrial isotope
- Detection of a roadside source
- Baselines trained on disjoint data

Current experiment

- 1, 2, 3, or 4 passes as sensors
- No warm-up for KF
- Fix SNR at 4
- 3 min. before and after source
- 400 trials

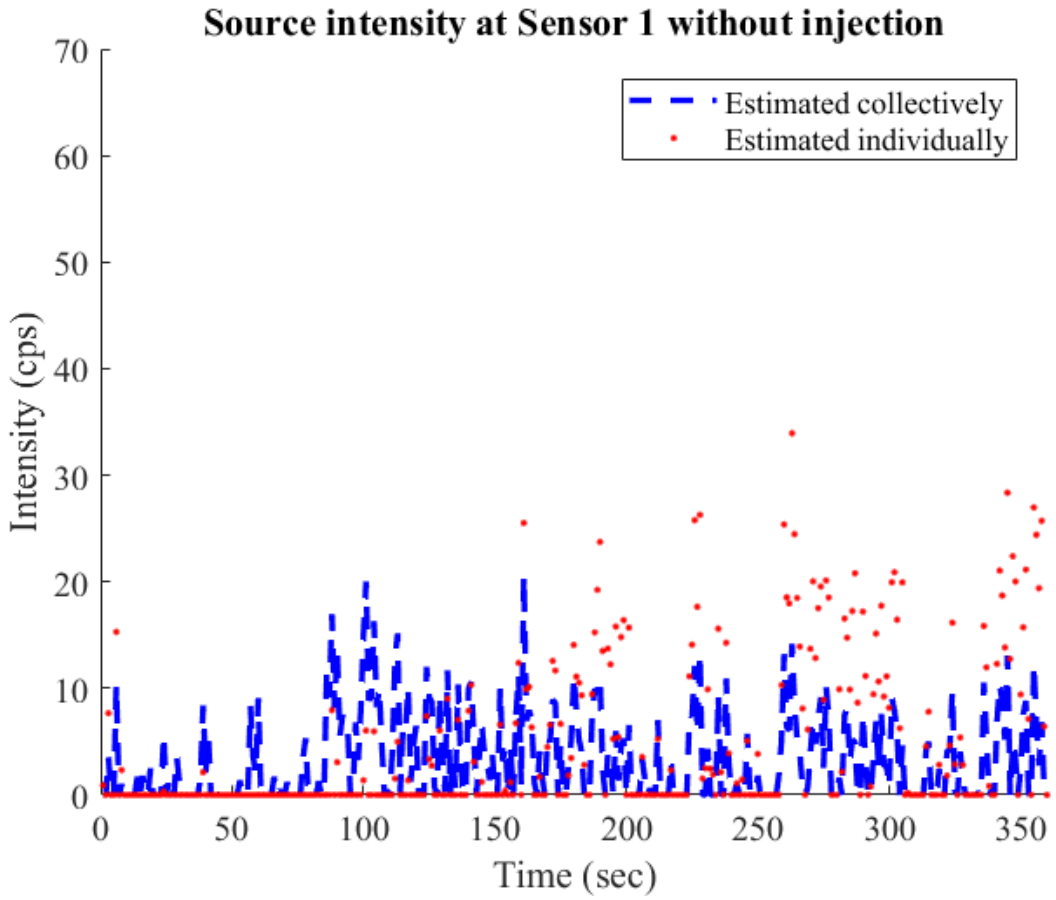
Our method gains more from multiple sensors

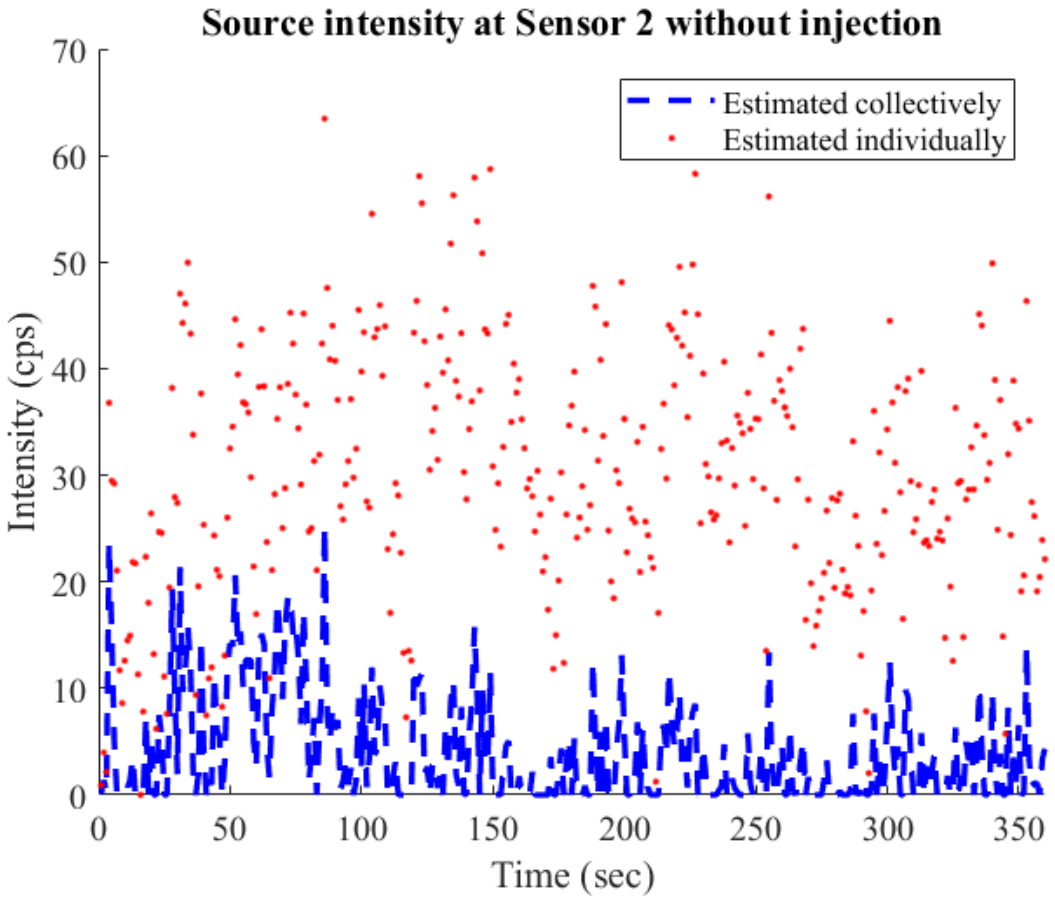
- Our method, BAF, performs much better at 2 sensors than 1
- No gain after 2 sensors
- BA (with classical MF) always weaker than BAF but always benefits from more sensors



Sensors ordered from closest to farthest from source.
Dashed lines are from the previous number of sensors.

Effect of collective filtering





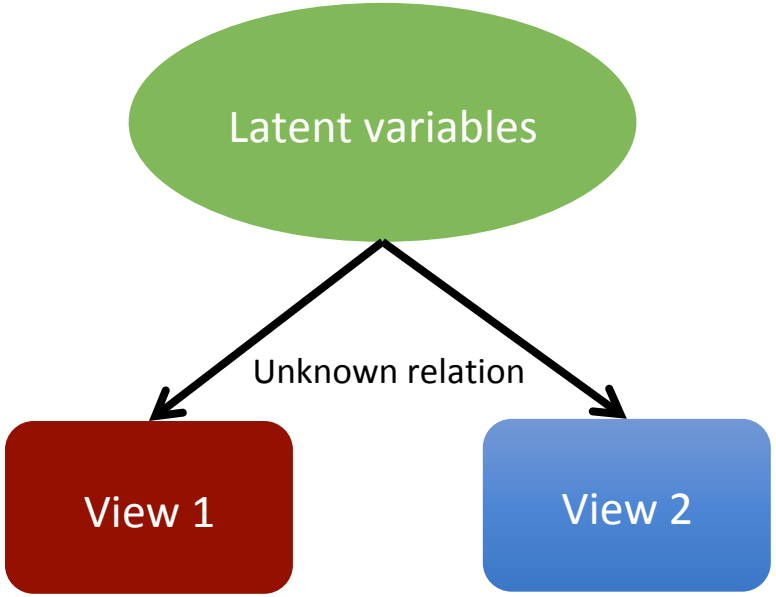
Multi-view filtering greatly decreases false positives from KF

Key takeaway of multiple sensors problem

Exploit smoothness and contemporaneous multiple views to achieve better performance than state of the art in multi-sensor settings with less training data

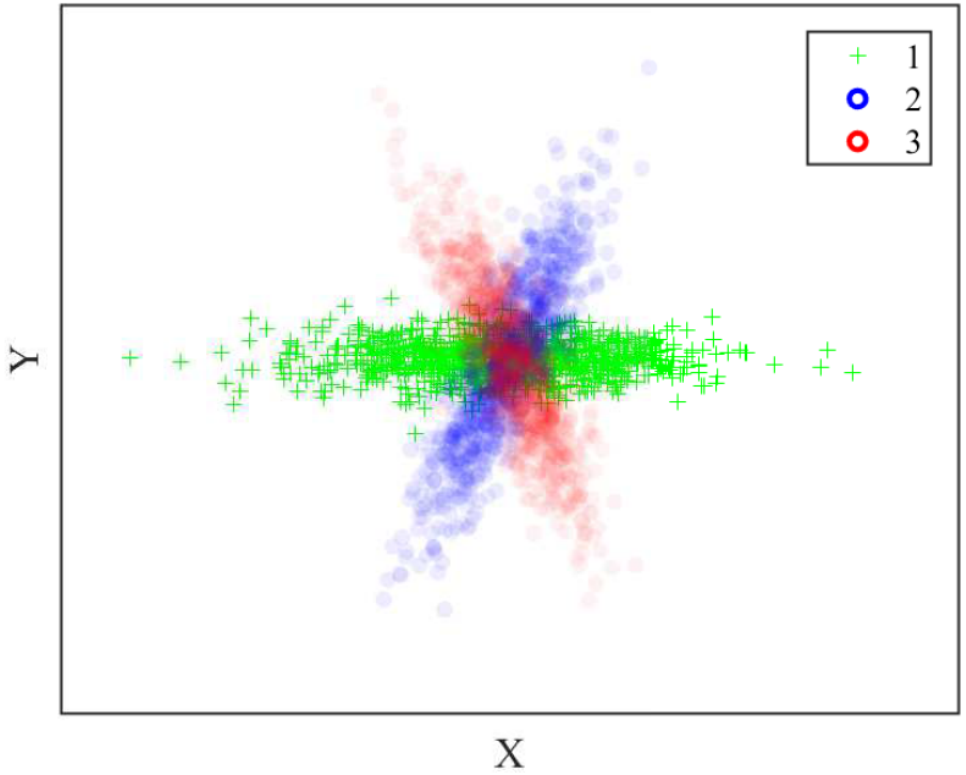
Outline

- Multi-view filtering
  - Single sensor method for gamma source detection [NSS 2017]
  - Multiple sensor extension
- **Learning multi-view relationships**
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]



Learning multi-view relationships

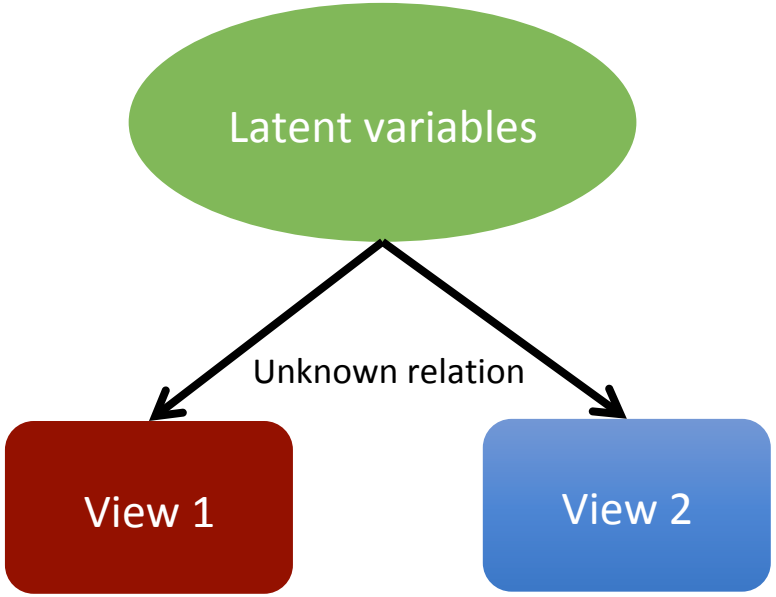
- The filtering work uses domain knowledge about the multi-view relationships
- If we lack this information, can we learn the relationships from data and still utilize them?
  - Linear relationships
  - Nonlinear relationships



Nonlinear multi-view relationship that can be represented by a mixture of 3 linear relationships.

Outline

- Multi-view filtering
  - Single sensor method for gamma source detection [NSS 2017]
  - Multiple sensor extension
- Learning multi-view relationships
  - **Linear multi-view relationships** [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]

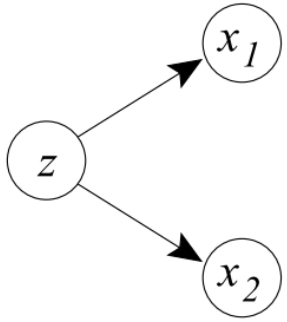


Canonical Correlation Analysis

- Two-view analogue to Principal Components Analysis
- Learns subspace to maximize correlation between two views [Hotelling 1936]

$$\max_{u,v} corr(X_1^T u, X_2^T v)$$

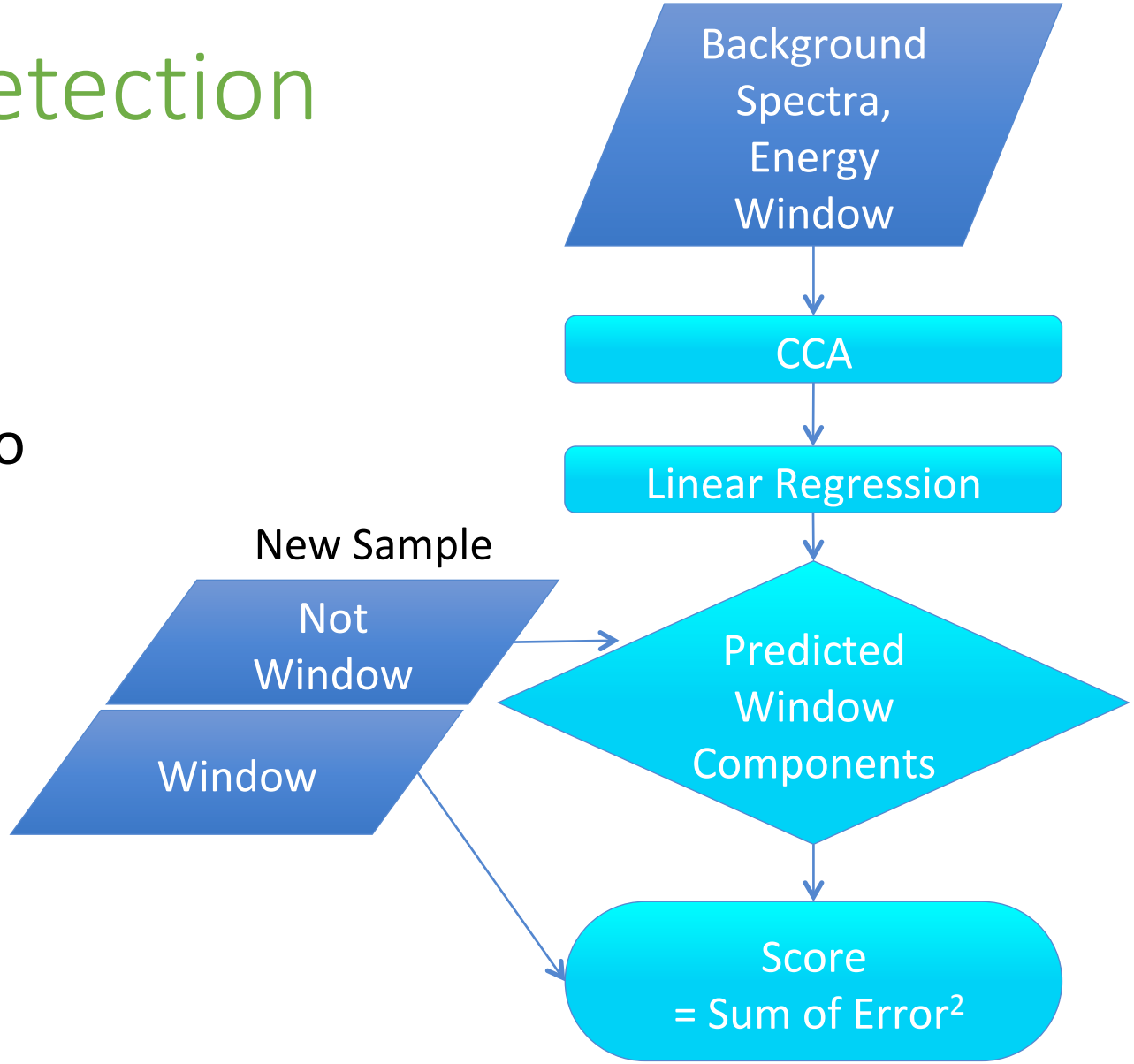
- Non-convex optimization with closed-form solution $A = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^{\mathsf{T}}$
- Gaussian model interpretation [Bach 2006]



Hotelling (1936). Relations between two sets of variates. Biometrika.
Bach (2006). A probabilistic interpretation of canonical correlation analysis. Berkeley Technical Report.

Idea: CCA Anomaly Detection

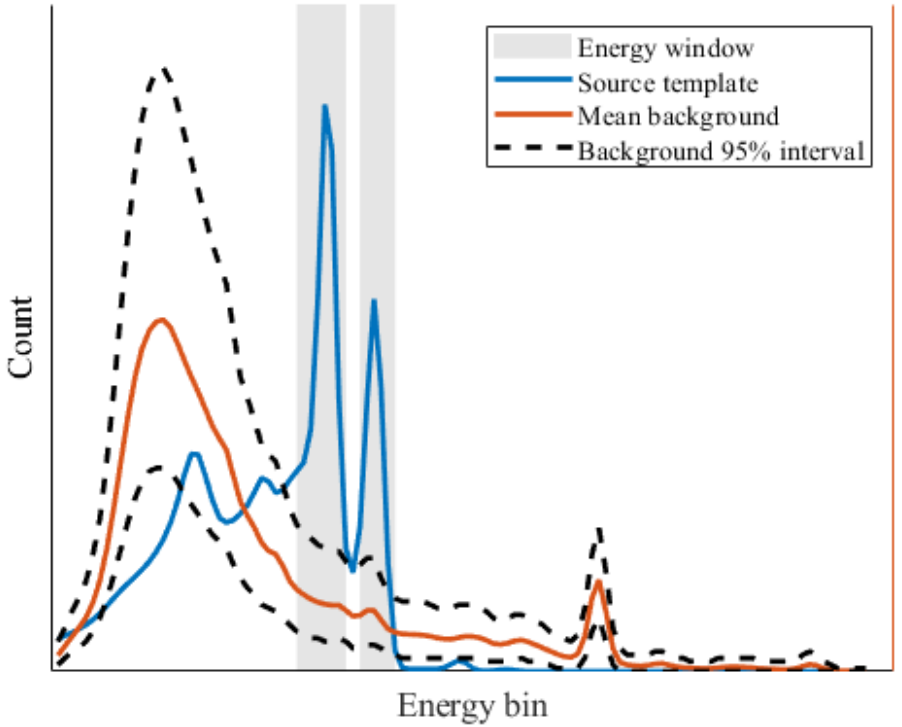
- Characterizes background reference data by CCA
- Checks when new observations do not match this structure
- Apply to imperfect source knowledge
- *Energy windows* as views



Dataset and experimental design

- 24 hours spectroscopic data in one-second intervals
- Synthetic injections of 67 source templates
- Binary classification of each sample
- Censored Energy Window (CEW): multi-view baseline

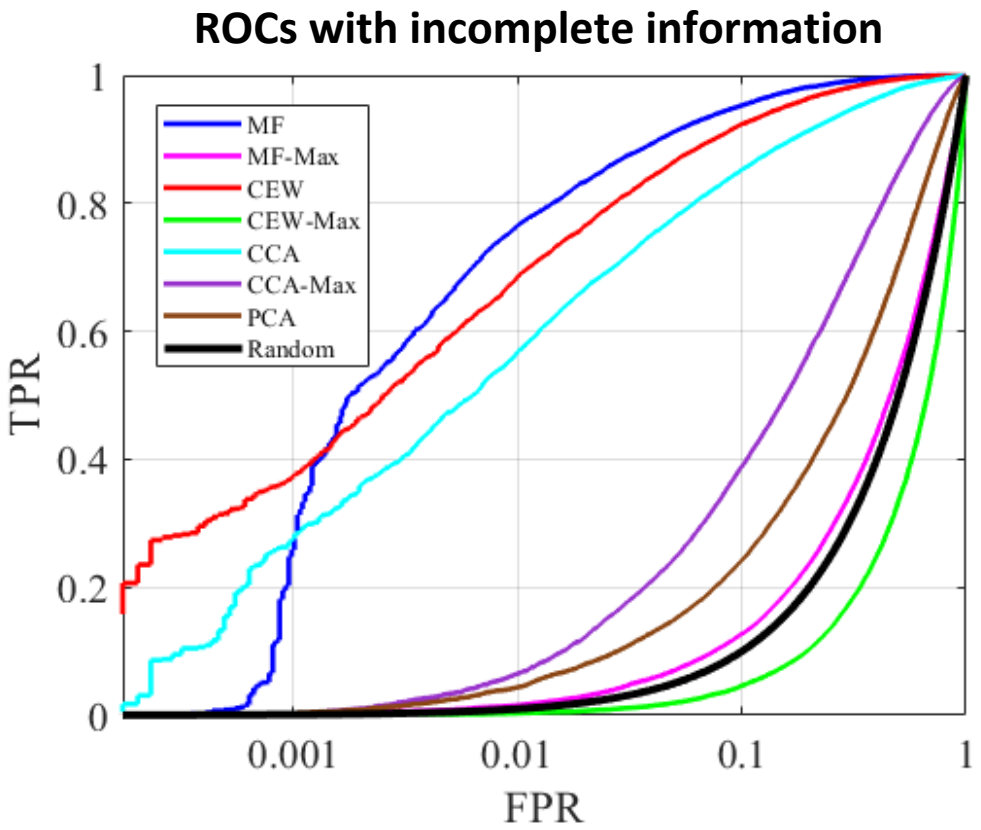
Energy window, source, and background



Expect to see source most clearly in energy window.

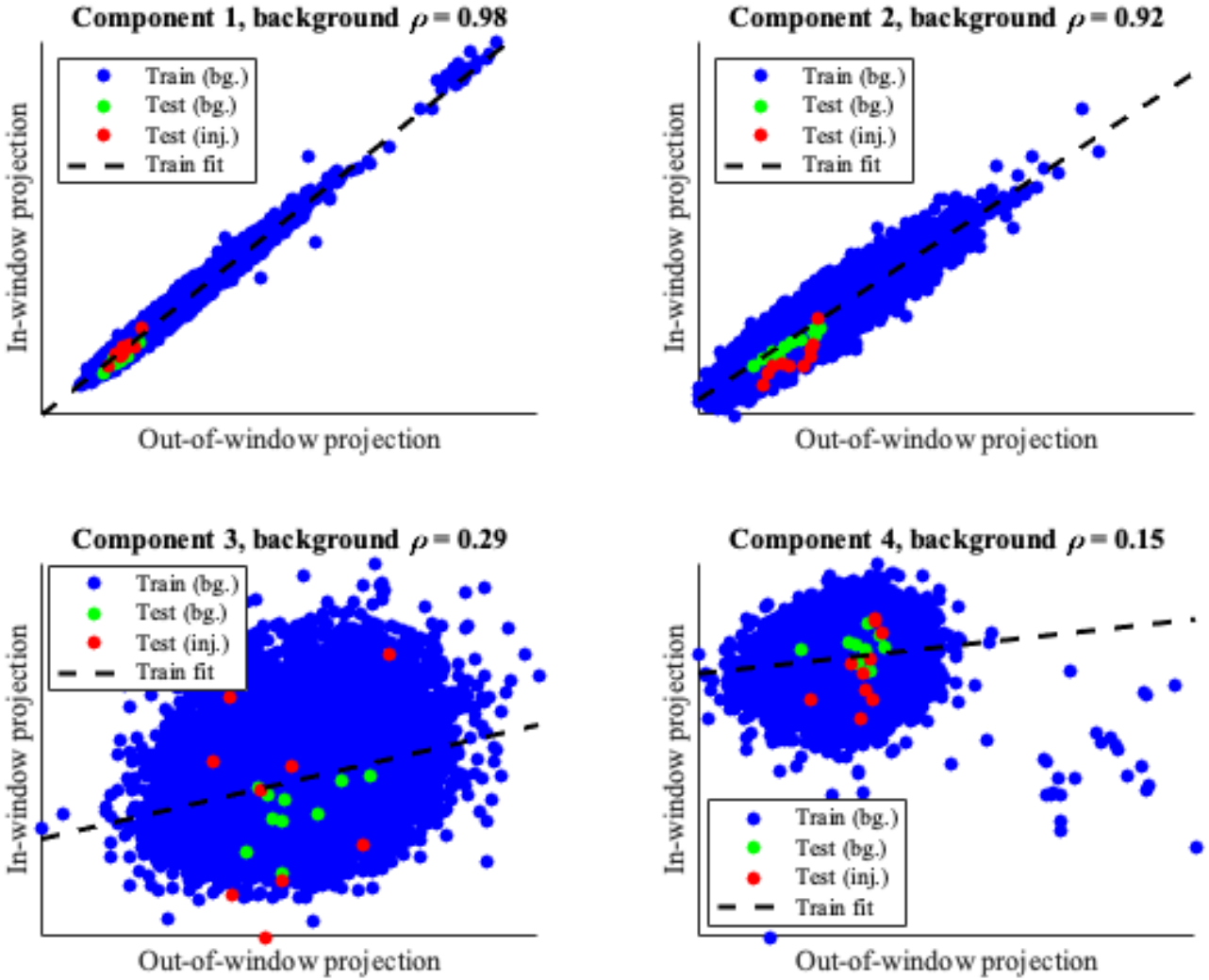
CCA Detection is more robust

- True source template missing from library
- CCA method outperforms MF-Max, CEW-Max, and PCA alternatives
- We expect to do worse than MF and CEW, which have perfect source information



Multiple components aid detection

- CCA Detection score is based on residuals
- With injection, small residuals individually, but noticeable combined
- Weaker correlations can even be more salient

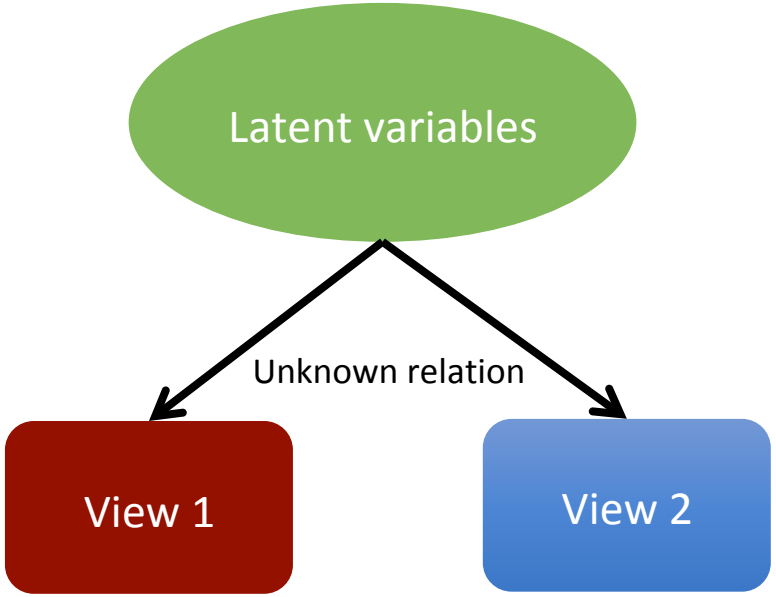


Key takeaway of imperfect source information problem

Leverage multiple linear correlations between views to make detection robust against imperfect information, a practical scenario

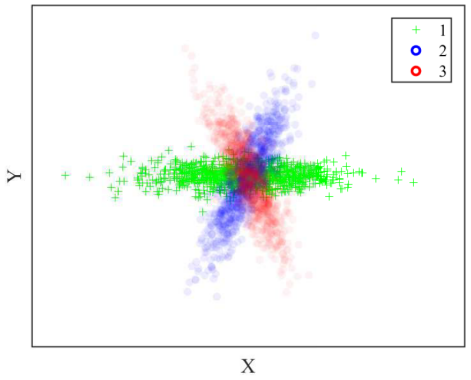
Outline

- Multi-view filtering
  - Single sensor method for gamma source detection [NSS 2017]
  - Multiple sensor extension
- Learning multi-view relationships
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - **Clustering** [MLHC 2017, ISICEM 2019]
    - Classification [MLHC 2017, ISICEM 2019]

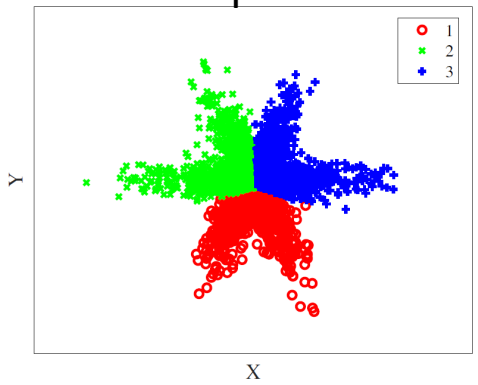


Clustering methods compared

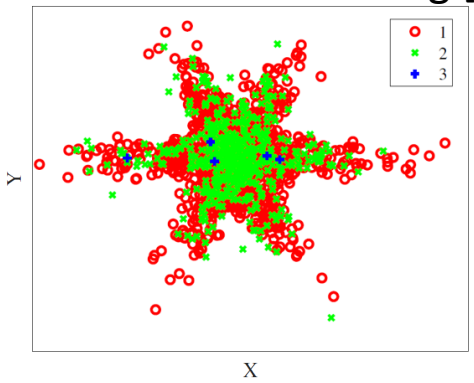
Ground truth



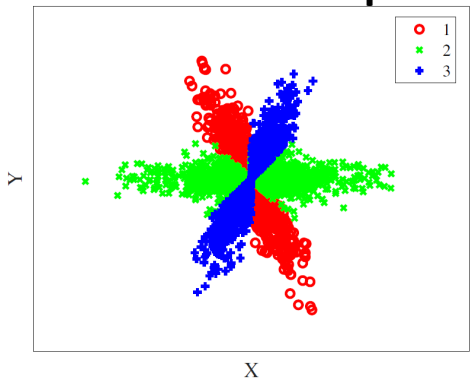
K-means or spectral clustering



Modern multi-view clustering [Zhao 2017]

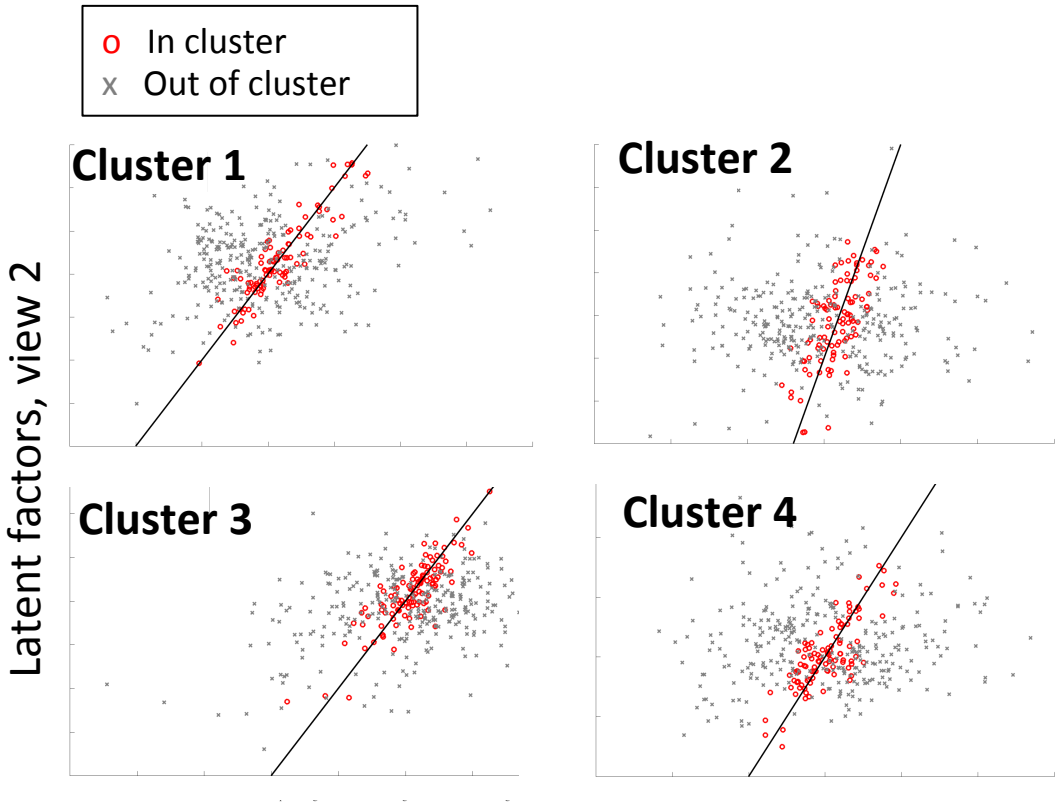


Multi-view relationship clustering



Our approach

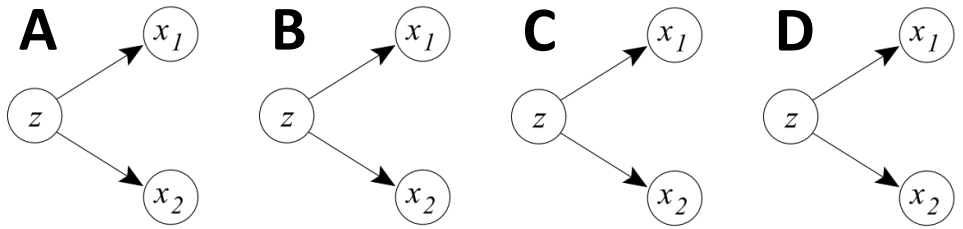
- Clusters observations according to multi-view relationships
- Fits cluster-wise linear relationships
  - Relevant latent factors are discovered in the process
  - Latent factors vary between clusters



In-cluster data in **Red**
Other data in **Grey**

Mixture of Canonical Correlations

- How to model different subsets of observations have different correlations?
- Generative model:



- Each cluster has a conditionally independent CCA structure
- Want to learn both cluster labels and correlations

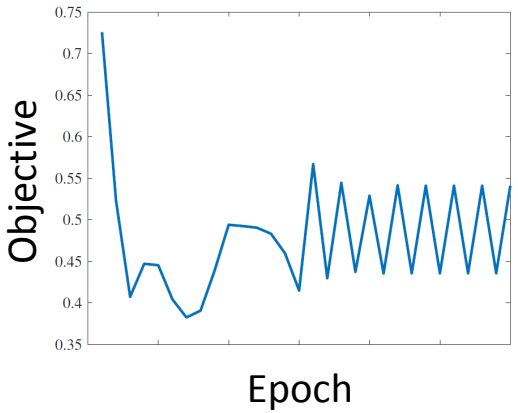
$$\min_{u,v,R} \quad \sum_j \|X^{(j)}u^{(j)} - Y^{(j)}v^{(j)}\|_2^2$$

$$\text{s.t.} \quad u^{(j)\top}X^{(j)\top}X^{(j)}u^{(j)} = v^{(j)\top}Y^{(j)\top}Y^{(j)}v^{(j)} = 1$$

Canonical Least Squares Clustering (CLS)

- Expectation Maximization-like iterative algorithm for CCA clusters [Fern 2005]
- Theoretical and empirical convergence problems
- Replace CCA with novel optimization problem, Canonical Least Squares (CLS)
- Non-convex optimization
  - Closed-form solution in first component
  - Greedy solution performs well empirically

CCA Clustering Loss



Canonical Least Squares

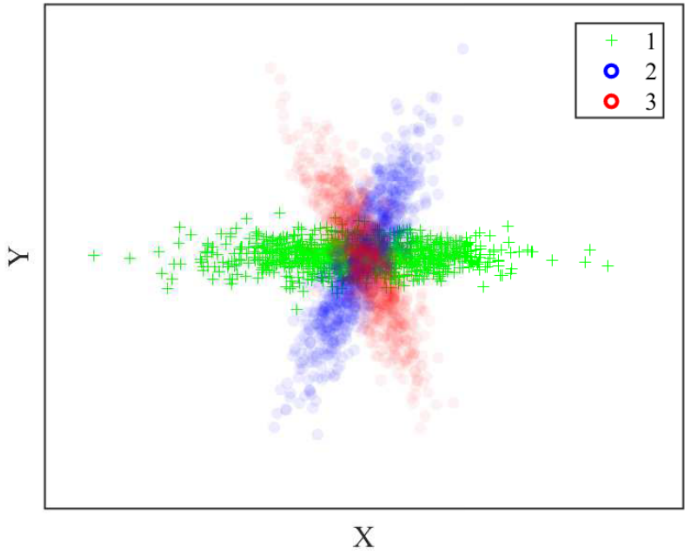
$$\min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad \|Xu - Yv\|_2^2$$

$$\text{subject to} \quad v^\mathsf{T} v = 1.$$

Fern et al. (2005). Correlation clustering for learning mixtures of canonical correlation models. ICDM.

Experiment on synthetic data

- 10 clusters of 1,000 points each in $\mathbf{R}^{100}$
  - Gaussian
  - All overlap at origin
  - Features partitioned in half to make two views



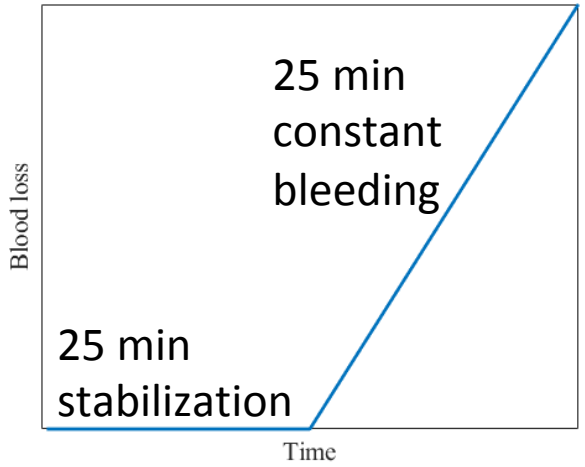
High-dimensional version of this with 10 clusters

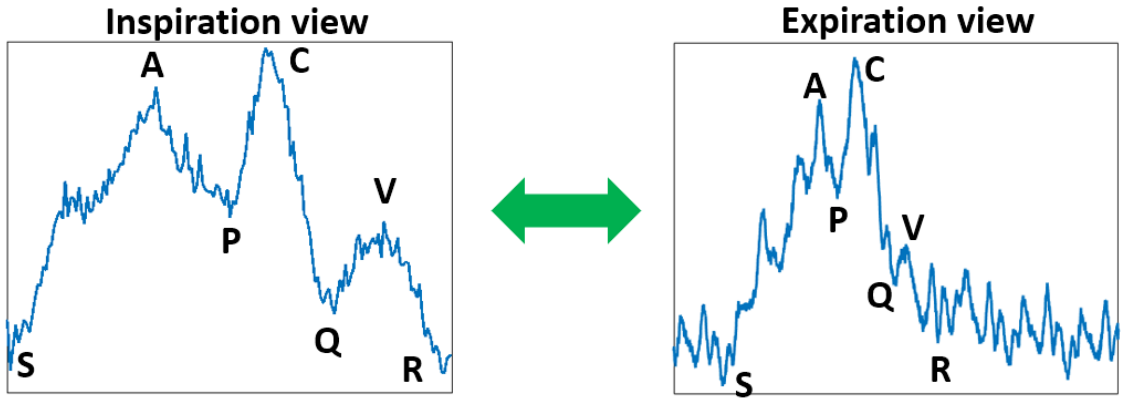
	CLS	CCA	*k*-means	Spectral
Adjusted Rand Index	**.99 ± .01**	.94 ± .02	.005 ± .003	.000 ± .000

CLS Clustering performs best

Experiment on induced bleeding

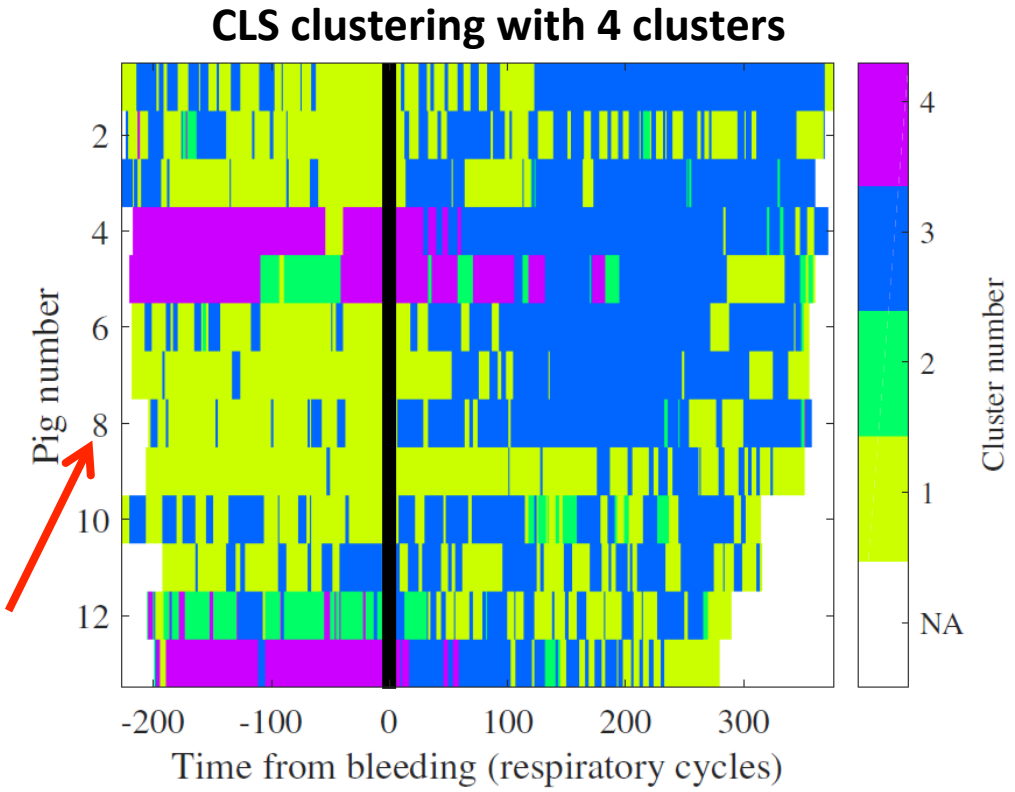
- 38 anesthetized pigs
- Time series of each subject's *central venous pressure* (CVP), sampled at 250Hz and featurized using domain knowledge
- Views correspond to CVP waveforms at the top of inspiration and bottom of expiration of the breathing cycle





Clusters across time and subjects

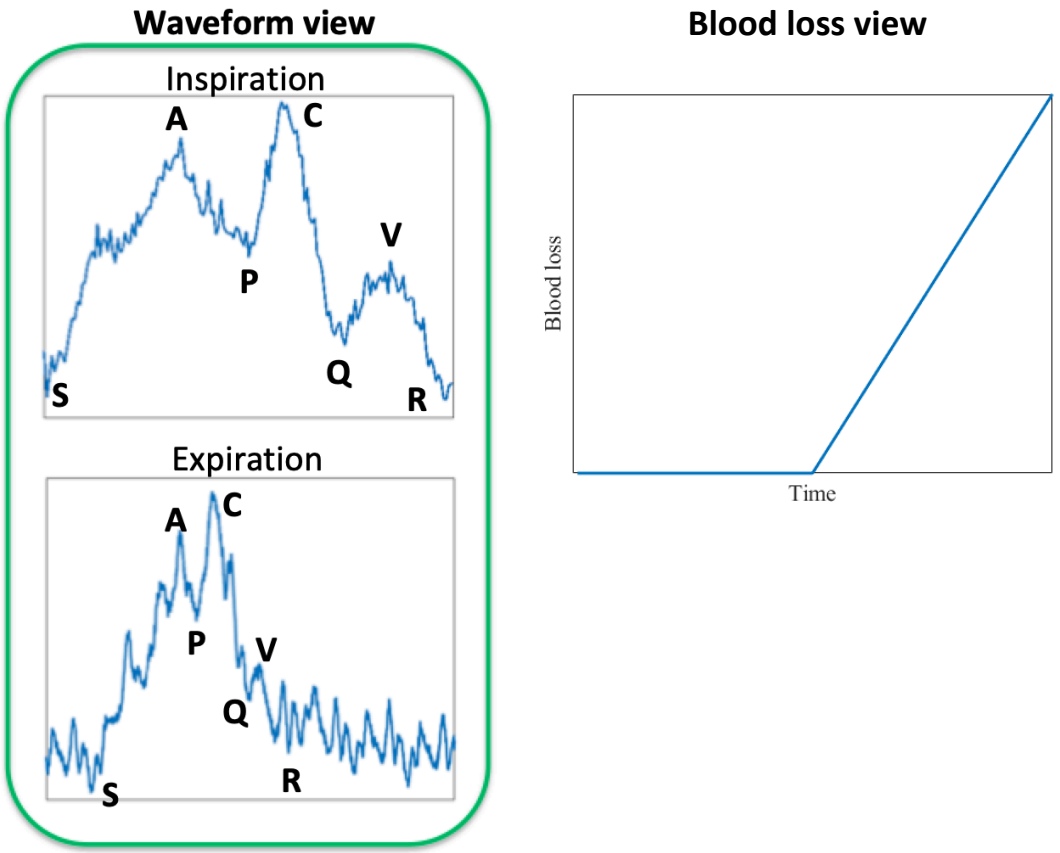
- No explicit knowledge of bleeding
- Variation across time reflects stages of bleeding response
- Variation across subjects reflects different phenotypes
- Our work may be first to characterize structure of response [Pinsky 2005, Boyd 2011, Marik 2013]

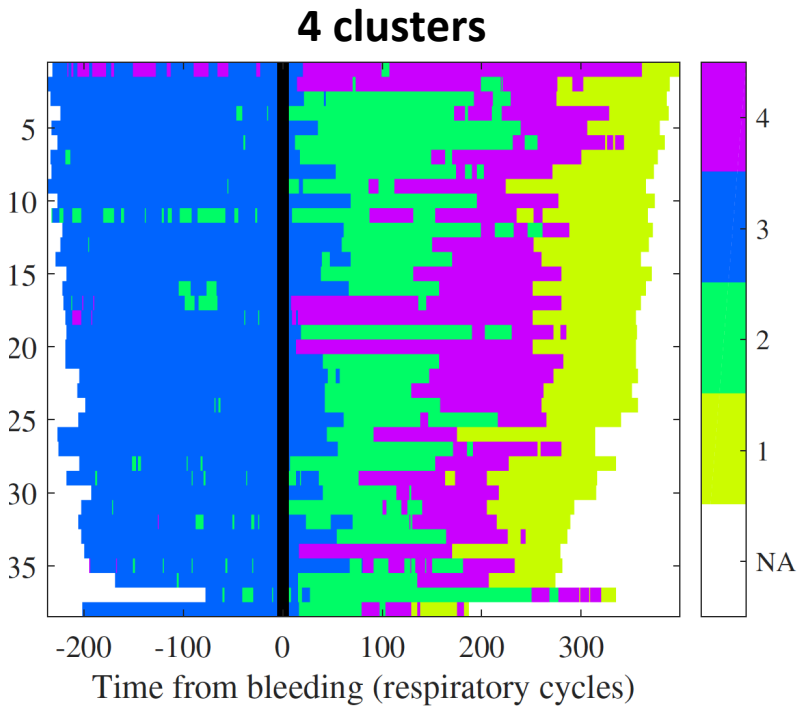


Learned correlation structure corresponds with blood loss

Pinsky and Payen (2005). Functional hemodynamic monitoring. Critical Care.
Boyd et al. (2011). Fluid resuscitation in septic shock. Critical Care Medicine.
Marik and Cavallazzi (2013). Does the central venous pressure predict fluid responsiveness? Critical Care Medicine.

Fully supervised clusters





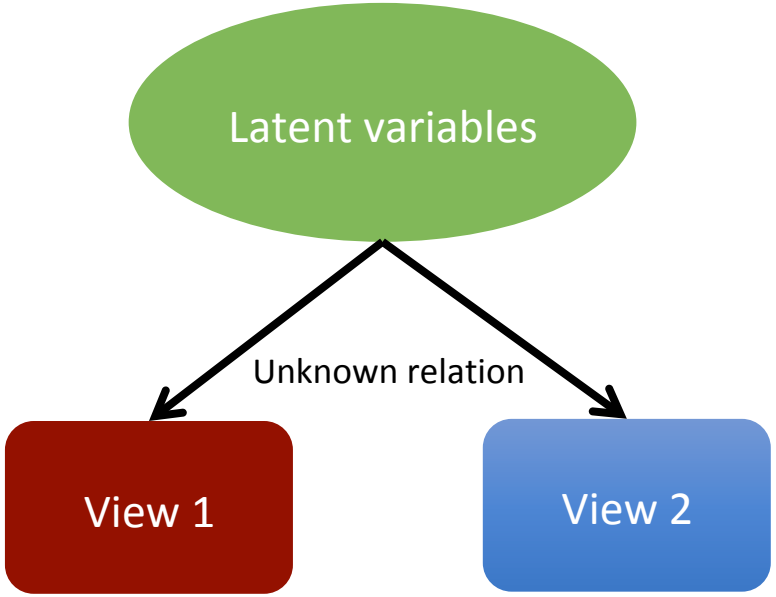
Refined clusters using bleeding information.
Not same clusters as previous slide.

Key takeaway of approach to multi-view clustering

- Novel algorithm identifies clusters based on multi-view relationships
- Performs well quantitatively on synthetic dataset and qualitatively on authentic bleeding dataset

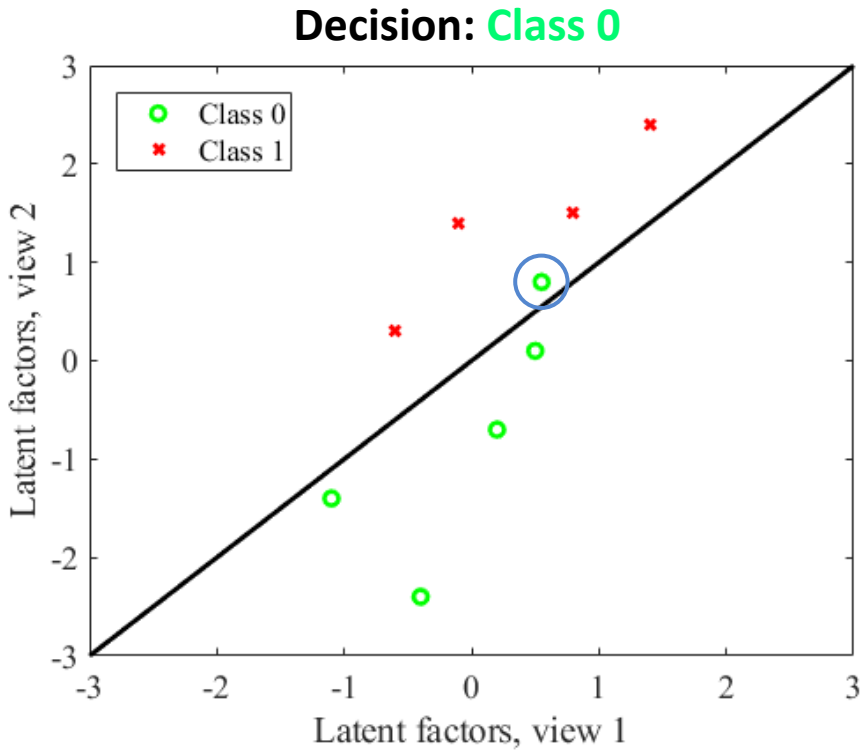
Outline

- Multi-view filtering
  - Single sensor method for gamma source detection [NSS 2017]
  - Multiple sensor extension
- Learning multi-view relationships
  - Linear multi-view relationships [NSS 2016]
  - Nonlinear multi-view relationships
    - Clustering [MLHC 2017, ISICEM 2019]
    - **Classification** [MLHC 2017, ISICEM 2019]



CLS classification

1. Learns CLS clusters independently on each class
2. Checks new point's fit in each cluster
3. Classify according to best fitting cluster



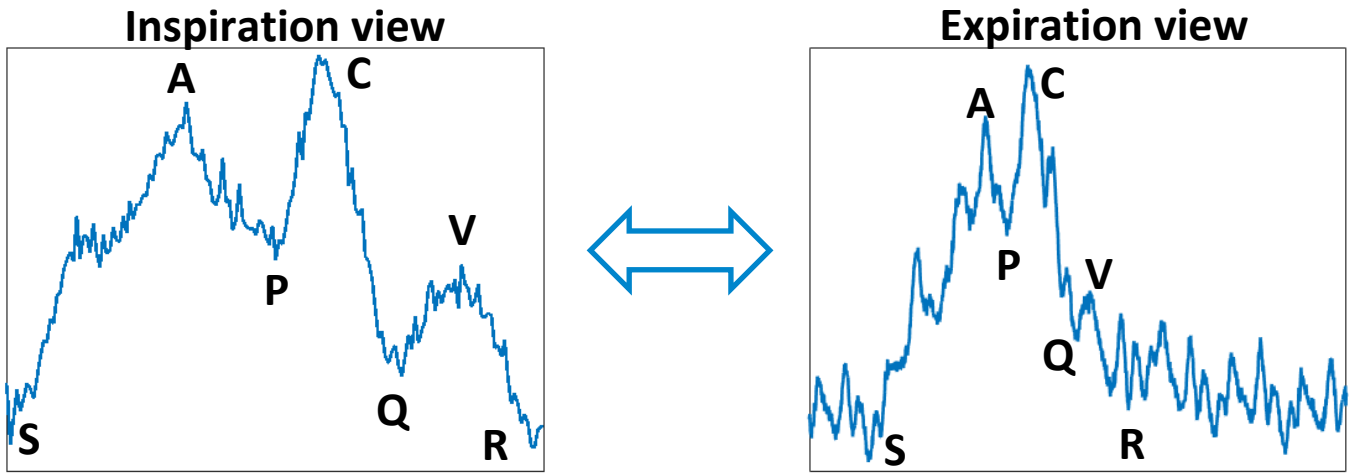
Nonlinear multiclass generalization of CCA anomaly detection

For Class *C* with clusters *j*, example with views *x* and *y* is scored as

$$Score_C((x,y)) = -\min_{j \in C} \|U_j x - V_j y\|_2^2$$

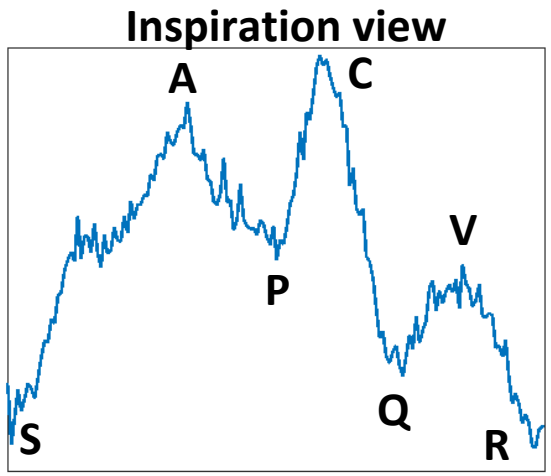
where U_j and V_j are CLS loading matrices in cluster *j*

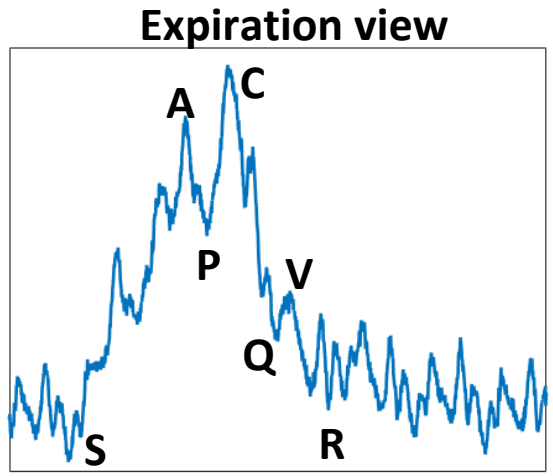
Moving CLS to supervised classification setting



Moving CLS to supervised classification setting

Has bleeding started?
☑ Yes
☐ No

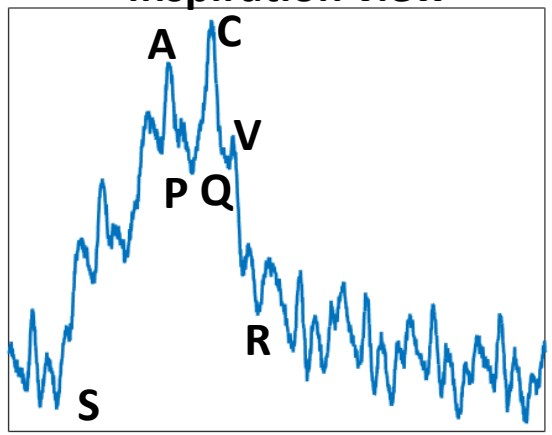


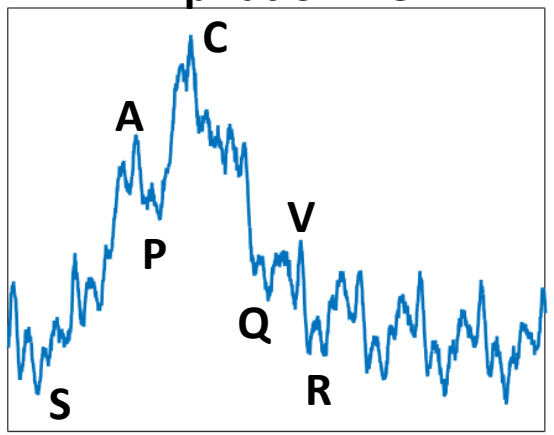


Moving CLS to supervised classification setting

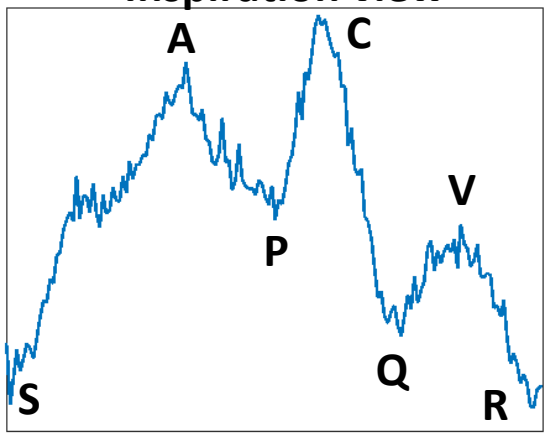
Decide whether a pair of waveforms came from **before** or **during** bleeding

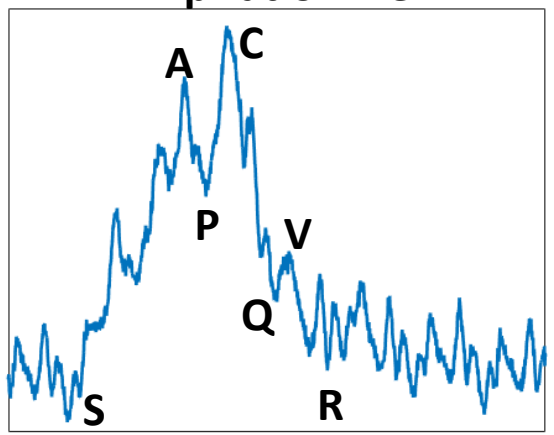
Has bleeding started?
☐ Yes
☑ **No**





Has bleeding started?
☑ **Yes**
☐ No





Classifier performance

	Single cluster CLS	Final CLS	Random forest
AUC	.701 ± .128	.862 ± .064	.891 ± .075
TPR @ .10 FPR	.468 ± .185	.674 ± .145	.762 ± .167
TPR @ .01 FPR	.222 ± .134	.501 ± .185	.610 ± .210
FPR @ .50 TPR	.239 ± .152	.064 ± .055	.073 ± .075

(95% C.I.s)

- Single cluster CLS: CLS classification with one cluster
- Final CLS: CLS classification with multiple clusters
- Random forest: best single-view classifier

Different application: Non-intrusive load monitoring (NILM)

- Pipeline of event-based NILM
- Event detection to identify when appliances are switched on/off; binary classification
- Event classification to identify which appliances are switched; multiclass classification

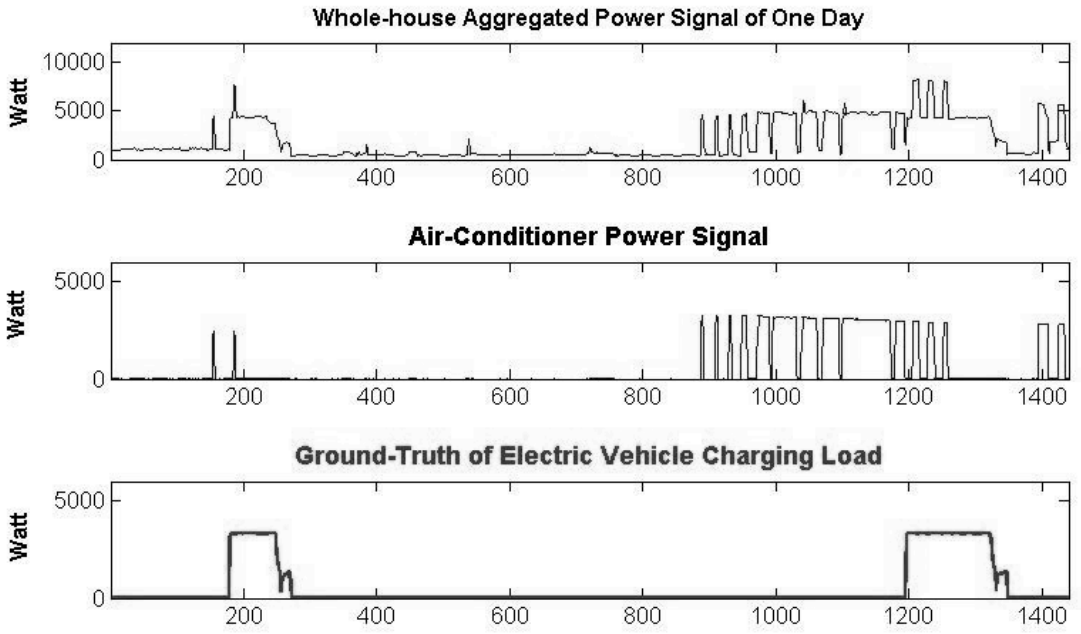


Figure from [Zhang 2019]

Disaggregation of power signals.

Zhang (2019). Non-intrusive load monitoring. Web.

Approach: Multi-view characterization of change over time

- Transient changes in appliance state carry temporal correlation structures
- Fingerprints of different appliances potentially discovered by clusters
- Views are Past and Present

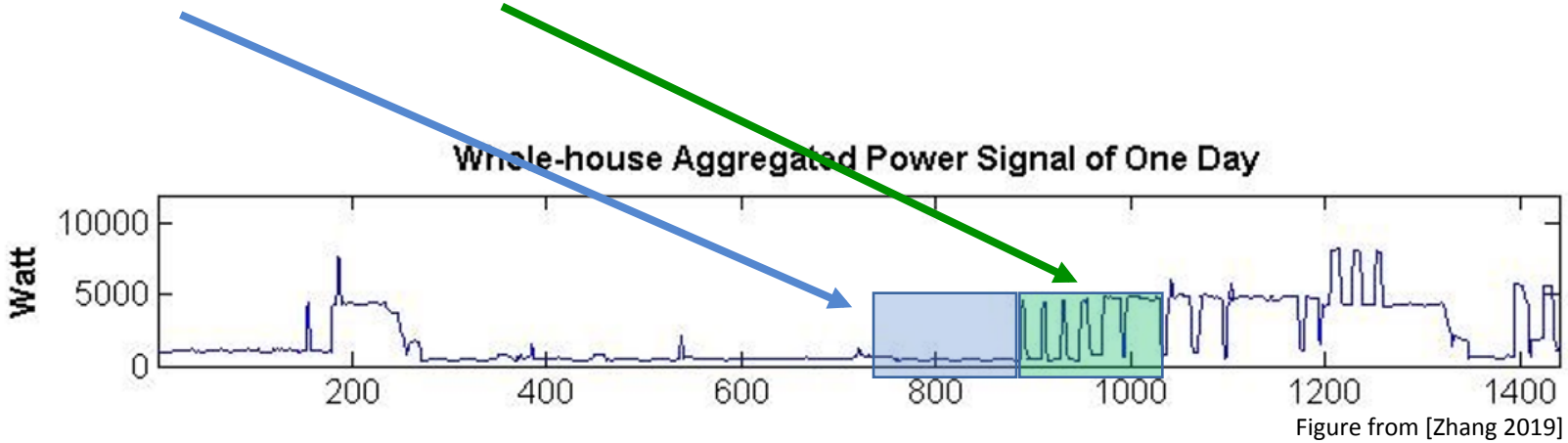


Figure from [Zhang 2019]

Baseline: Goodness-of-fit method

- Chi-squared goodness-of-fit (GOF) test is state-of-the-art [Jin 2011]

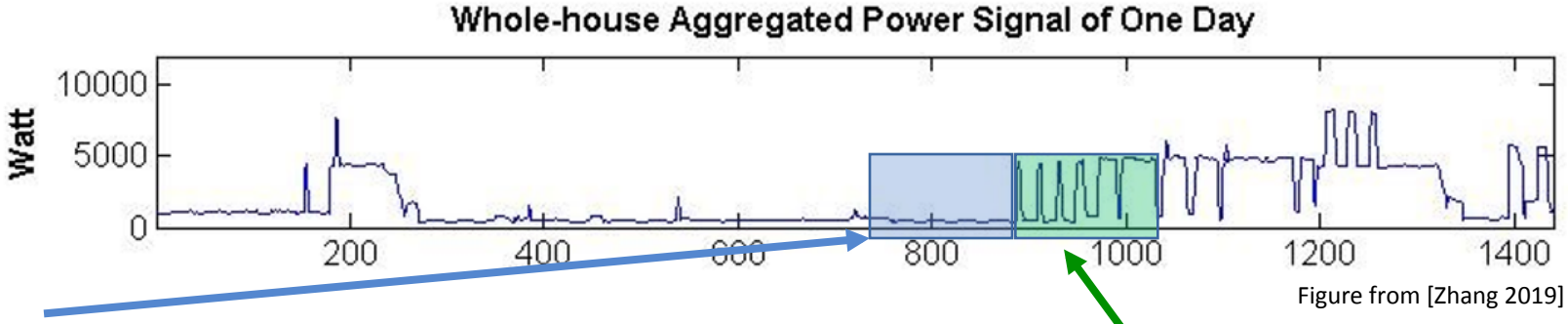


Figure from [Zhang 2019]

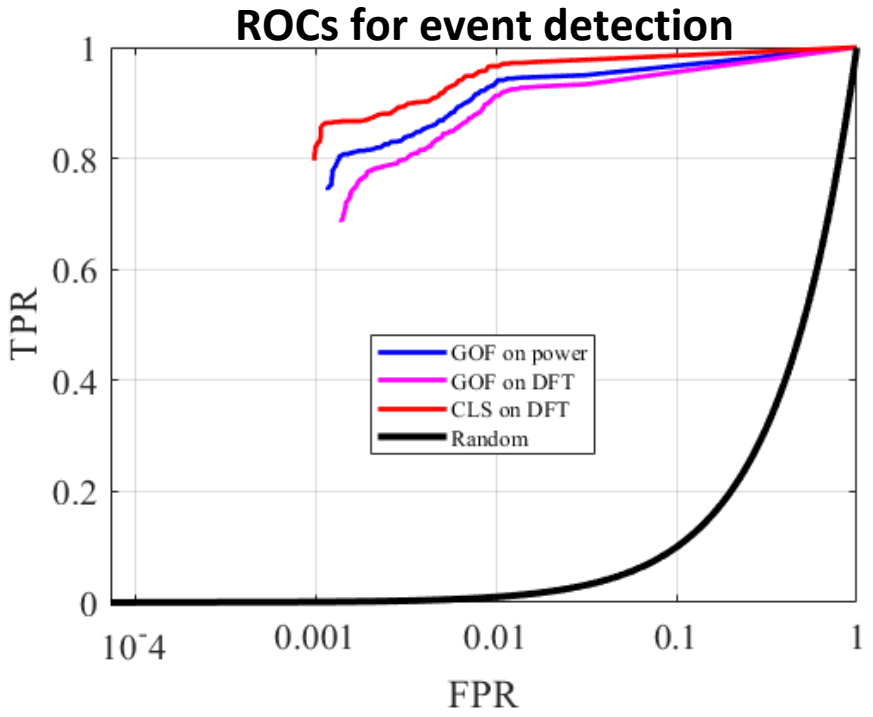
  - H_0: Past was drawn from same distribution as Present
  - H_1: Distributions differ
- Test statistic has chi-squared distribution
  - Paired in order
  - Gaussian assumption

$$\sum_i \frac{(p_i - q_i)^2}{p_i}$$

Jin et al. (2011). A time-frequency approach for event detection in non-intrusive load monitoring. Signal Processing, Sensor Fusion, and Target Recognition.

Event detection experiment

- BLUED [Filip 2011]
  - Power from houses with multiple appliances
  - Frequently used in benchmarking
  - 12kHz power over 7 days with labeled events
- Featurization by discrete Fourier transform of each window; top 5 principal components



FPR at certain levels of TPR

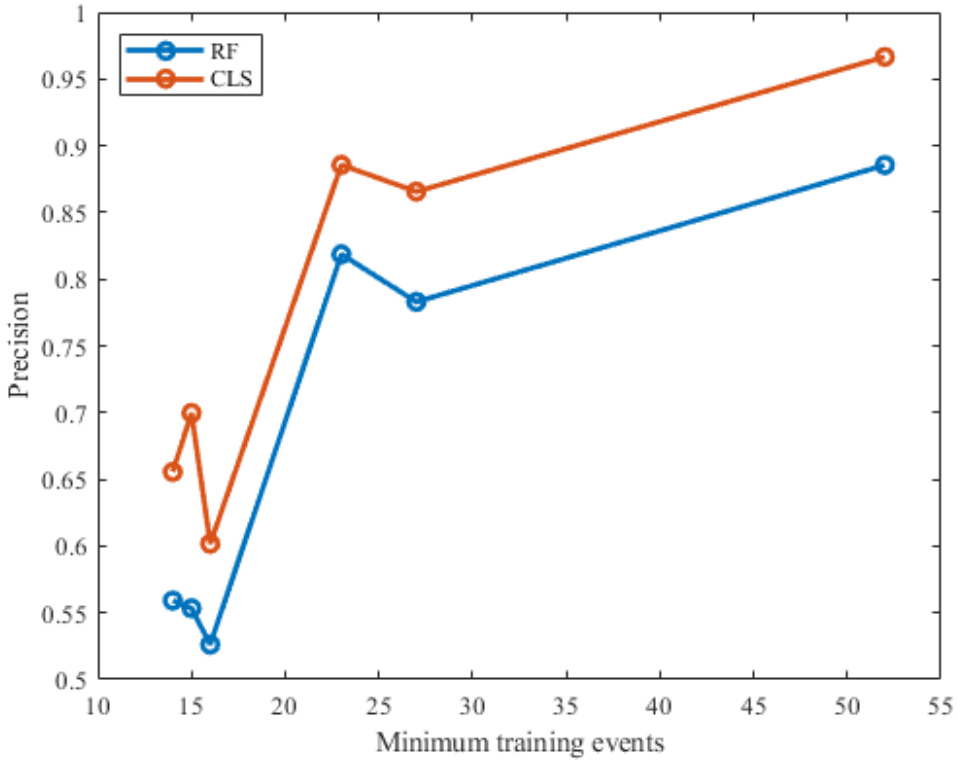
TPR	80%	85%	90%	95%	98%
GOF on power	.10%	.15%	.50%	.81%	5.5%
CLS on Fourier	**.08%**	**.09%**	**.26%**	**.67%**	**3.92%**

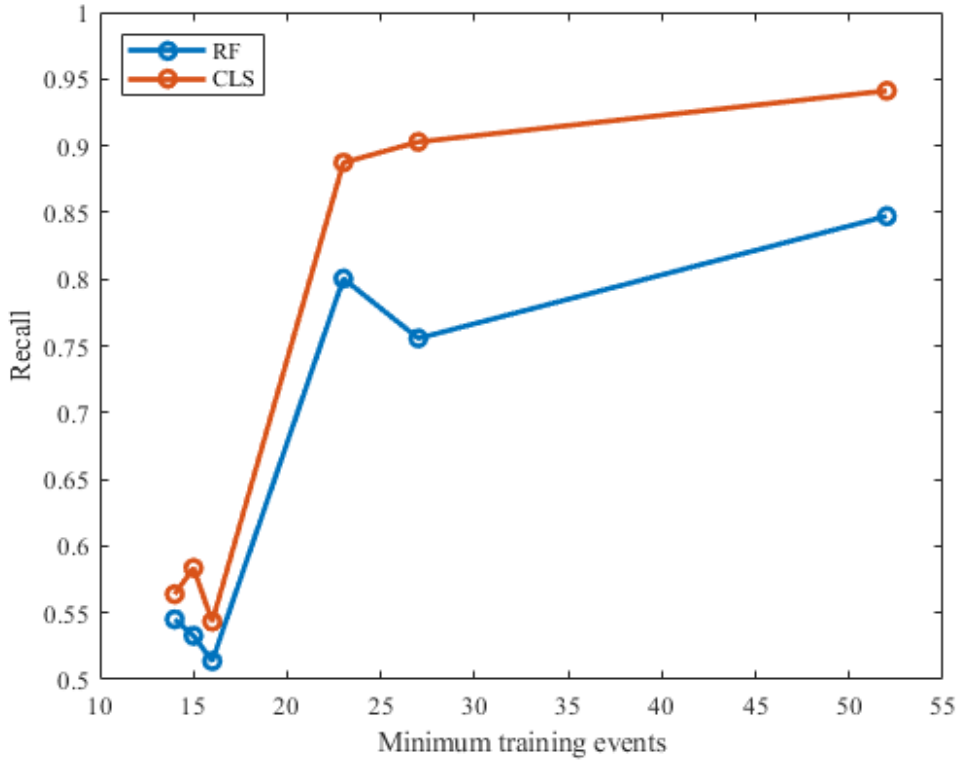
Results statistically significant.

Filip (2011). BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. Workshop on Data Mining Applications in Sustainability.

Event classification experiment

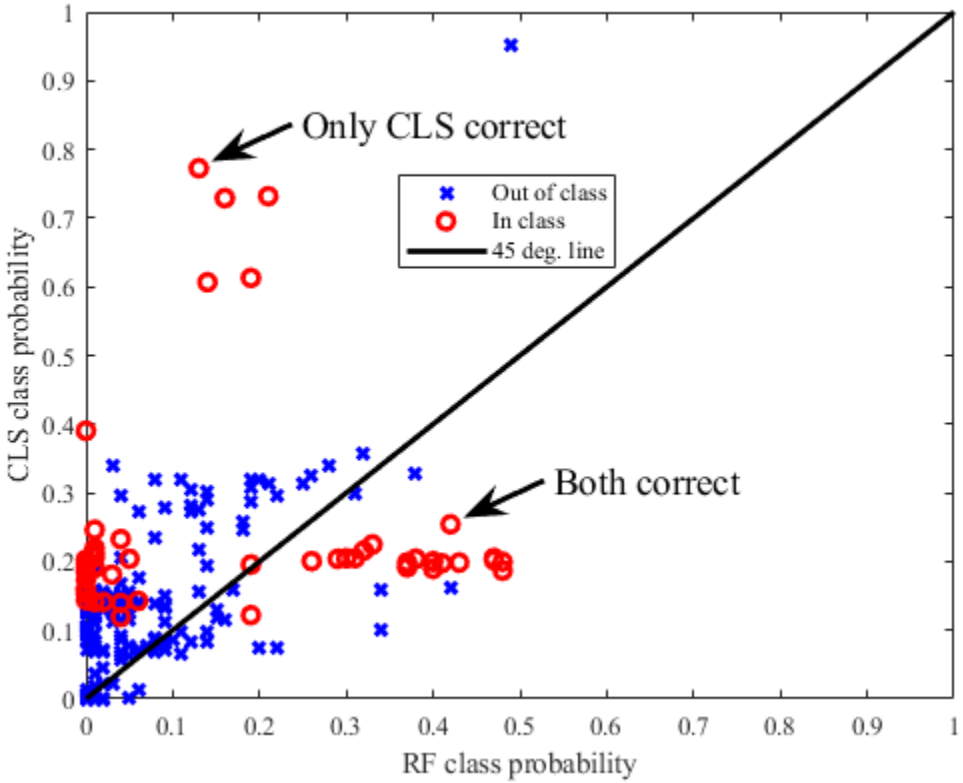
- Varied the threshold of events for a class to be included.
- Compared to Random Forest (RF), the best baseline.



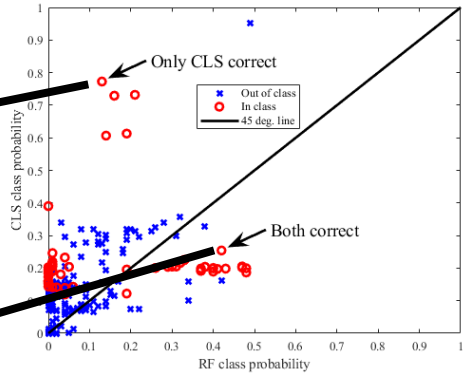


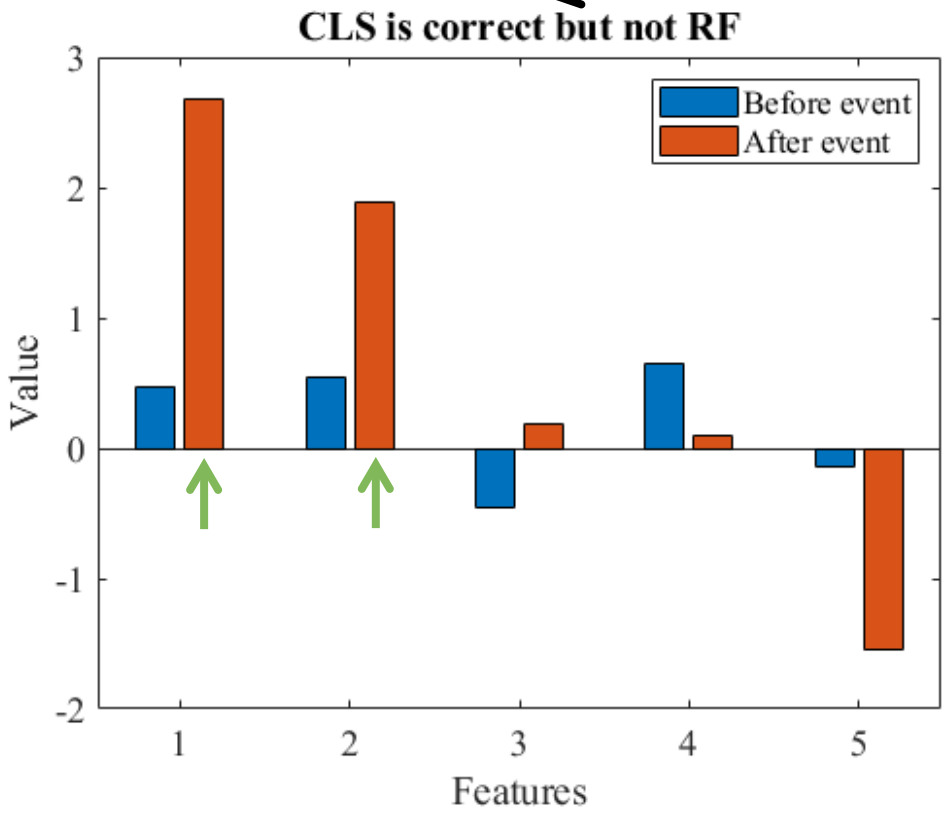
Mostly statistically significant.

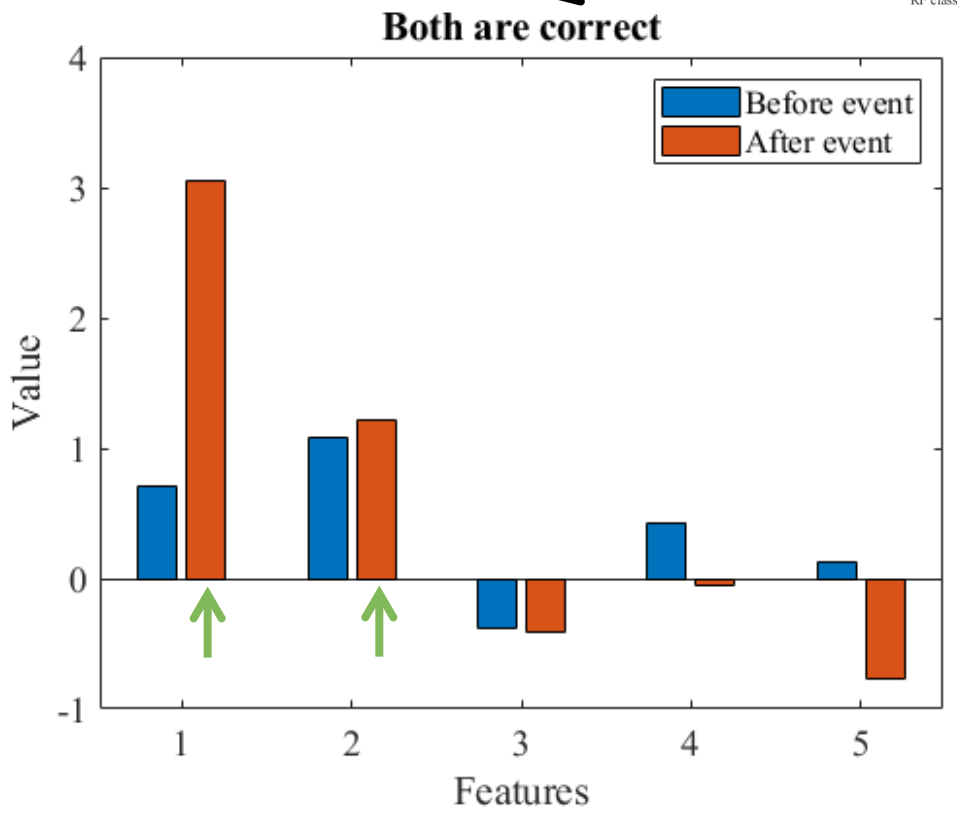
Scores of two examples, one a mistake by random forest



Features of two examples

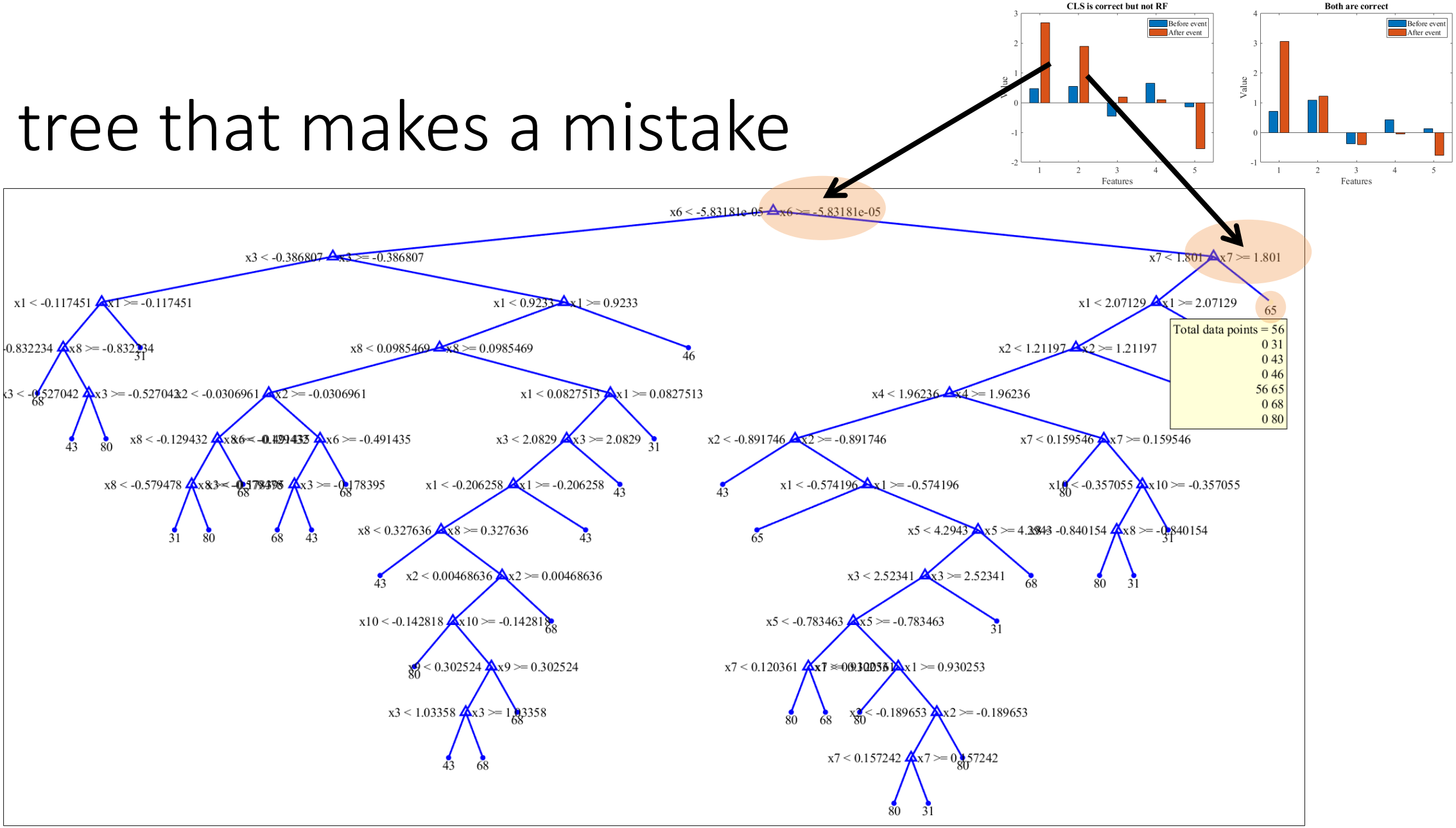






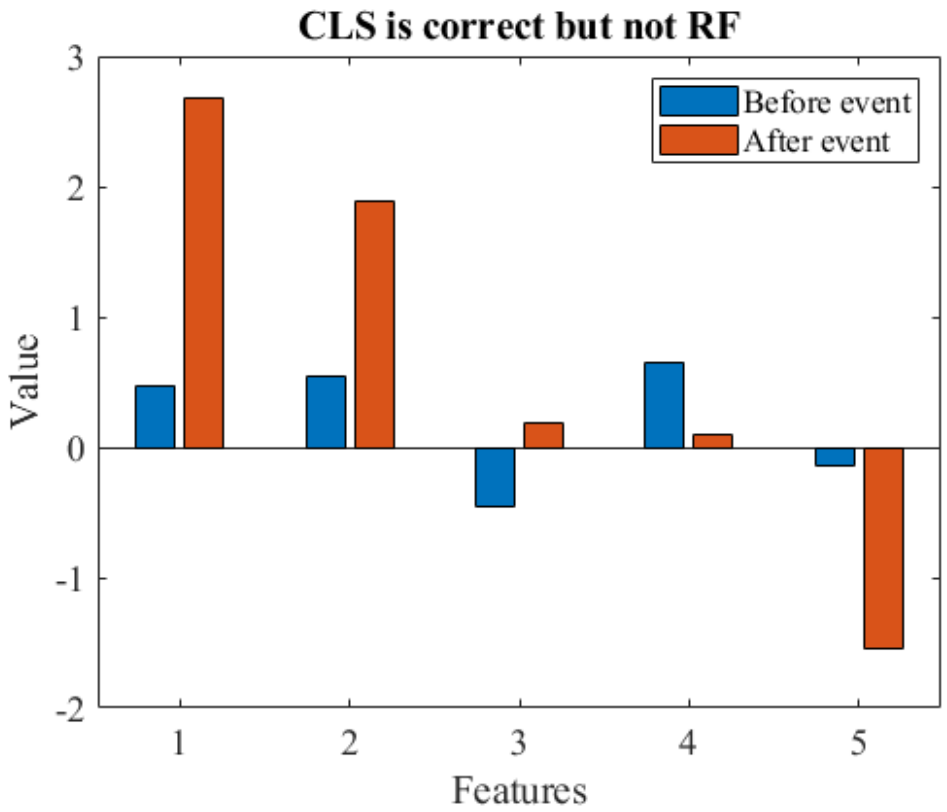
Multi-view relationship is similar between examples.

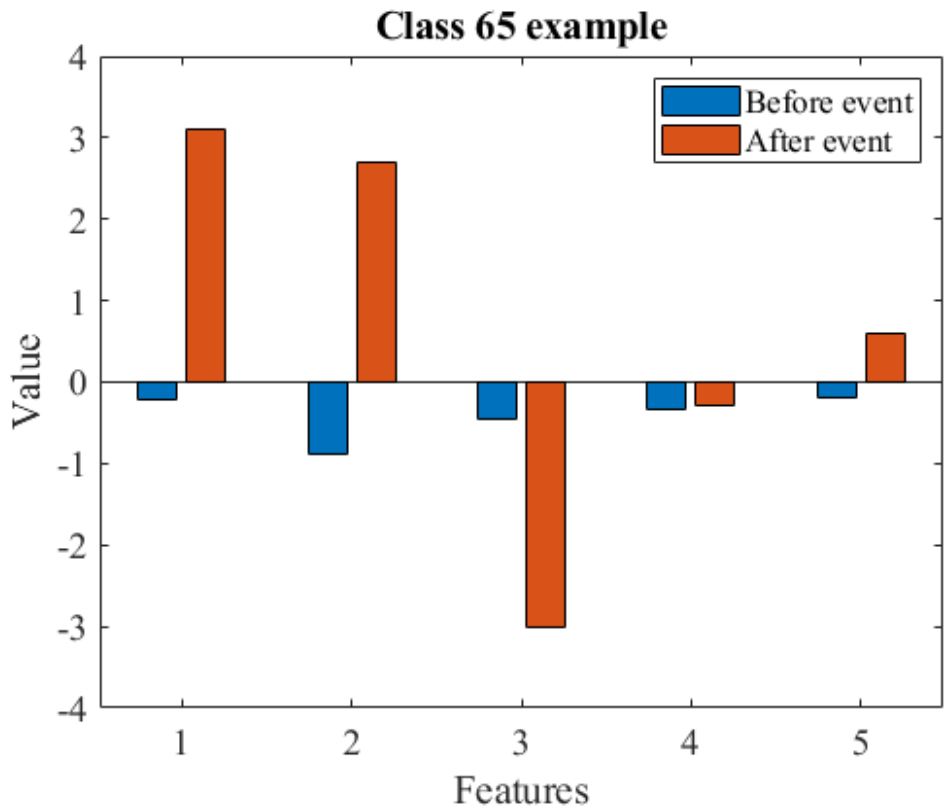
A tree that makes a mistake



Each node only uses one view.

Features of the mistaken class





Tree misclassifies example as Class 65 because it neglects the multi-view relationship.

Key takeaway of approach to multi-view classification

Leverage multi-view relationships as discriminative factors in classification to perform well on bleeding and load monitoring datasets

Multi-view relationships for analytics and inference
Summary

- Operated on multi-view relationships as a unit of analysis, resulting in novel structure and good empirical performance
- Single sensor method for gamma source detection
- Multiple sensor extension with known relationship
- Linear multi-view relationships
- Nonlinear multi-view relationships
  - Clustering
  - Classification
- Multi-view approach to learning on distributions

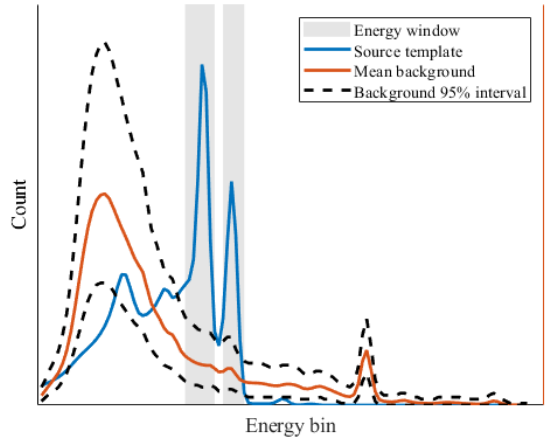
Multi-view relationships for analytics and inference

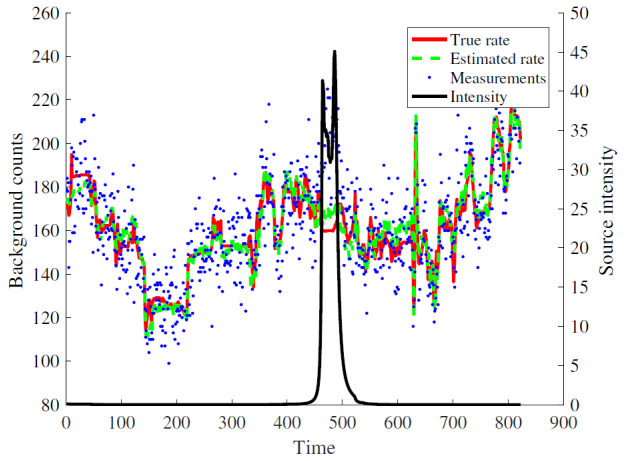
Future work

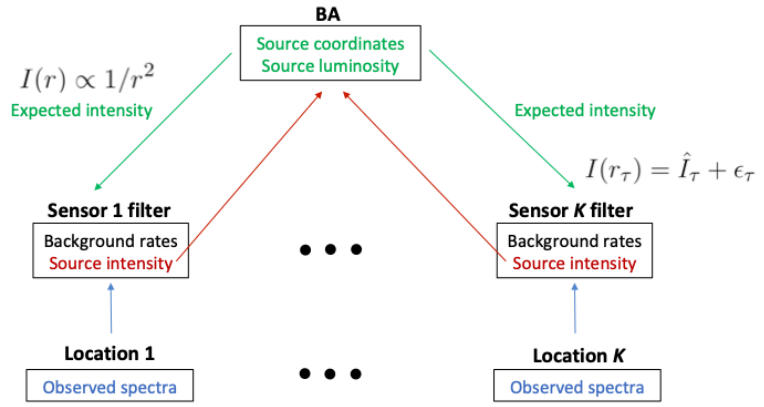
- Regression
- Multi-modal data
- Arbitrary nonlinear relationships
  - Mutual information instead of correlation
  - Kernel CCA [Shotaro 2006] and Deep CCA [Galen 2013]
- Theory to explain performance of CLS clustering/classification

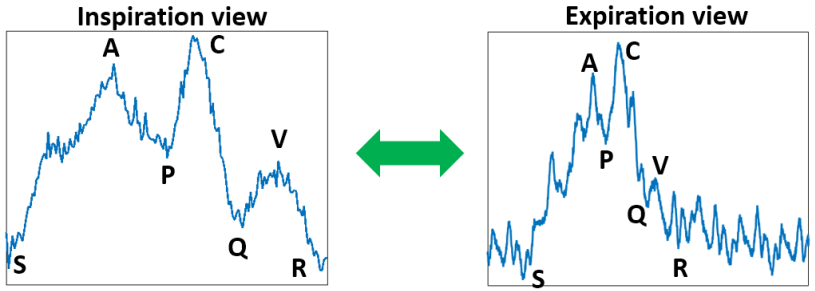
Shotaro (2006). A kernel method for canonical correlation analysis. arXiv preprint.
Galen et al. (2013). Deep canonical correlation analysis. ICML.

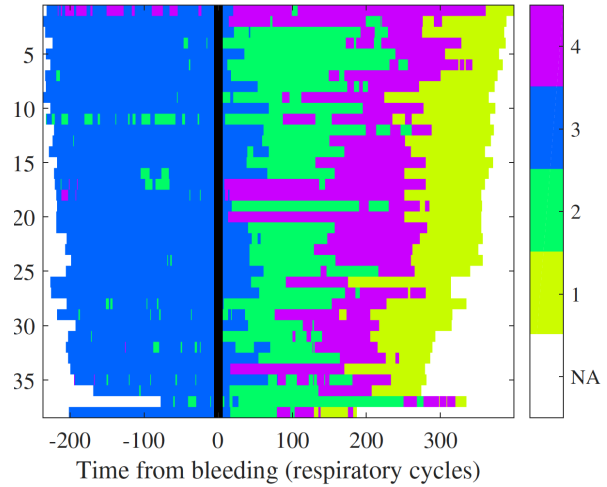
Thank you

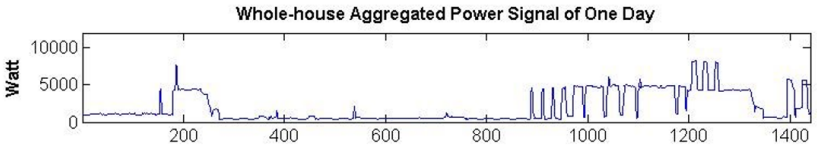












Lei et al. (2016). Radiological threat detection for an unknown energy window by canonical correlation analysis. NSS.
Lei et al. (2017). Robust detection of radiation threat by simultaneous estimation of source intensity and background. NSS.
Lei et al. (2017). Bleeding detection by multi-view correlation clustering of central venous pressure. MLHC.
Lei et al. (2019). Characterization of multi-view hemodynamic data by learning mixtures of multi-output regressors. ISICEM.

Appendix

More than two views

- Generalized CCA [Horst 1961] finds shared representation *G* between all views

$$\underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} \sum_{j=1}^{J} \|G - U_j^\top X_j\|_F^2$$
$$\text{subject to} \quad GG^\top = I_r$$

- Other extensions of CCA loss could be sum or minimax over pairwise loss

Horst (1961). Generalized canonical correlations and their applications to experimental data. University of Washington, Seattle.
Benton (2017). Deep generalized canonical correlation analysis. arXiv preprint.

Adaptive Filtering to Set Up Kalman Filter

- KF is very sensitive to the hyperparameters and
- An adaptive KF assumes they are non-stationary and estimates them in real-time
  - Bayesian, MLE, covariance matching, correlation
  - Often computationally expensive
- We propose a simple method that functions well for this problem

where is a Gaussian filter with length and variance

- Disadvantage 1: Introduces additional uncertainty
- Disadvantage 2: Requires a short burn-in period (measurements)

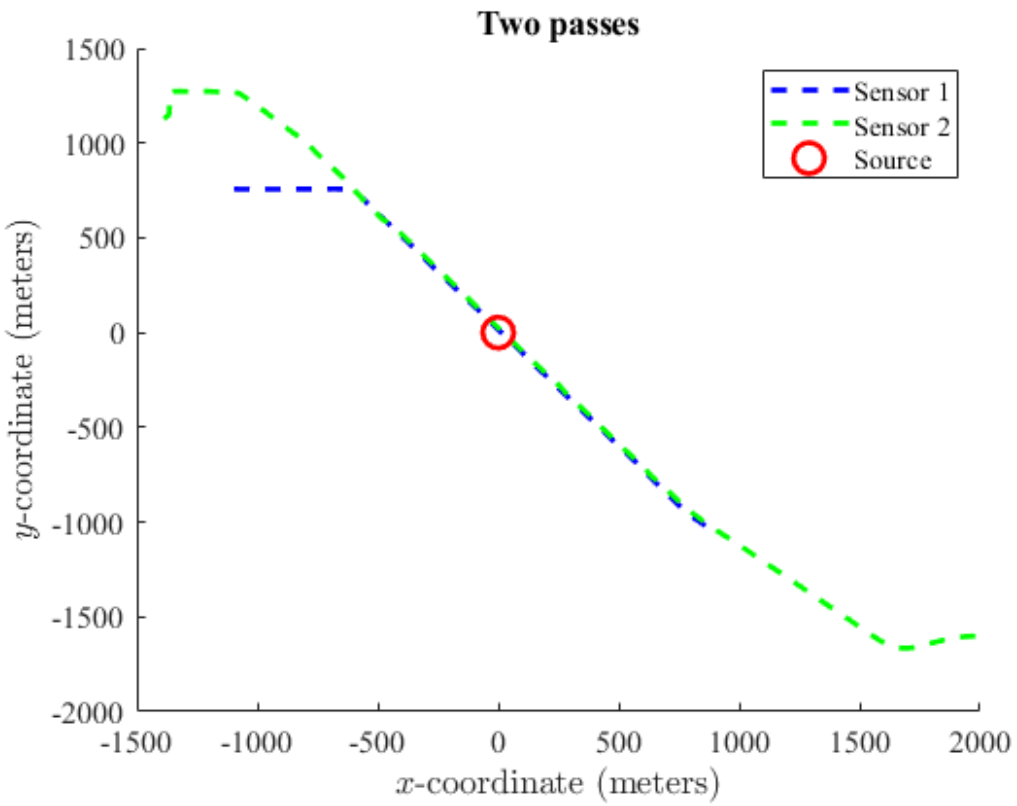
Prediction approaches

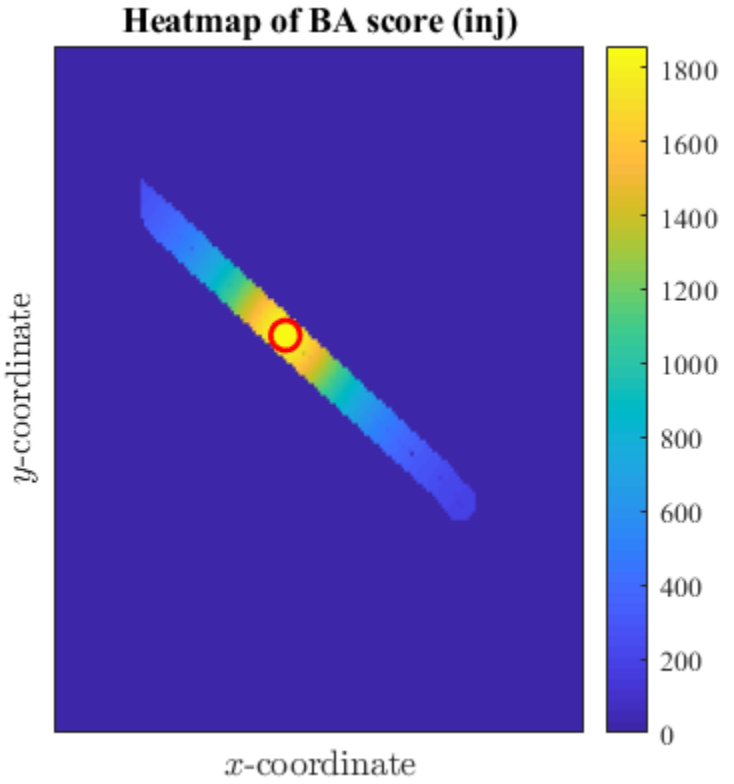
- Estimated background is detection score
- Insert background estimates into other methods
  - Matched Filter maximizes signal-to-noise ratio and is given by:

$$h = \mathrm{cov}(X)^{-1} s$$

    for spectra *X* and threat template *s*
  - Use past *k* estimates of background to get current covariance? No
  - $\mathrm{cov}(X) \rightarrow diag(x_t)$ where x_t is current estimated background

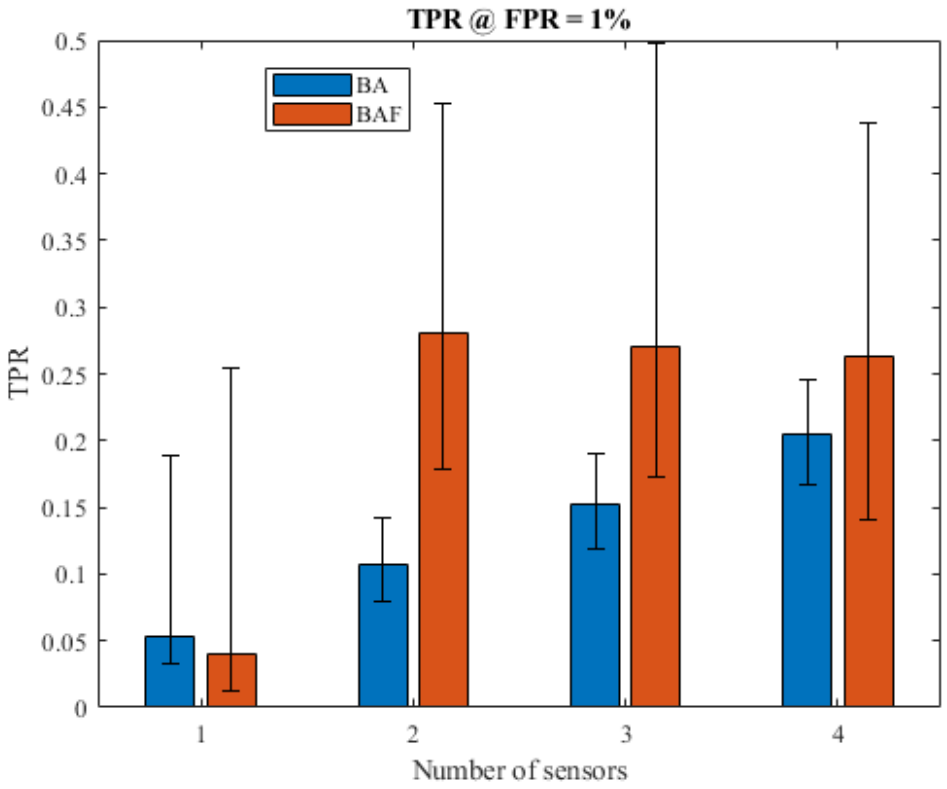
Hypotheses restricted to pass area

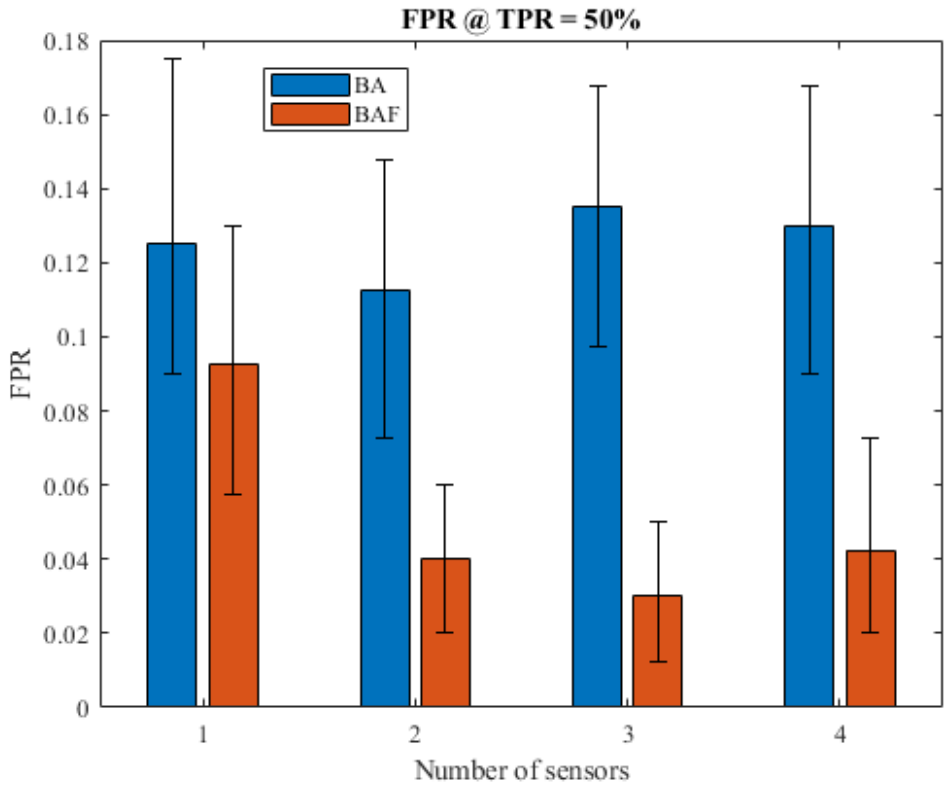




BA and BAF look for source in intersection of sensor paths.

ROC statistics for multi-view filtering

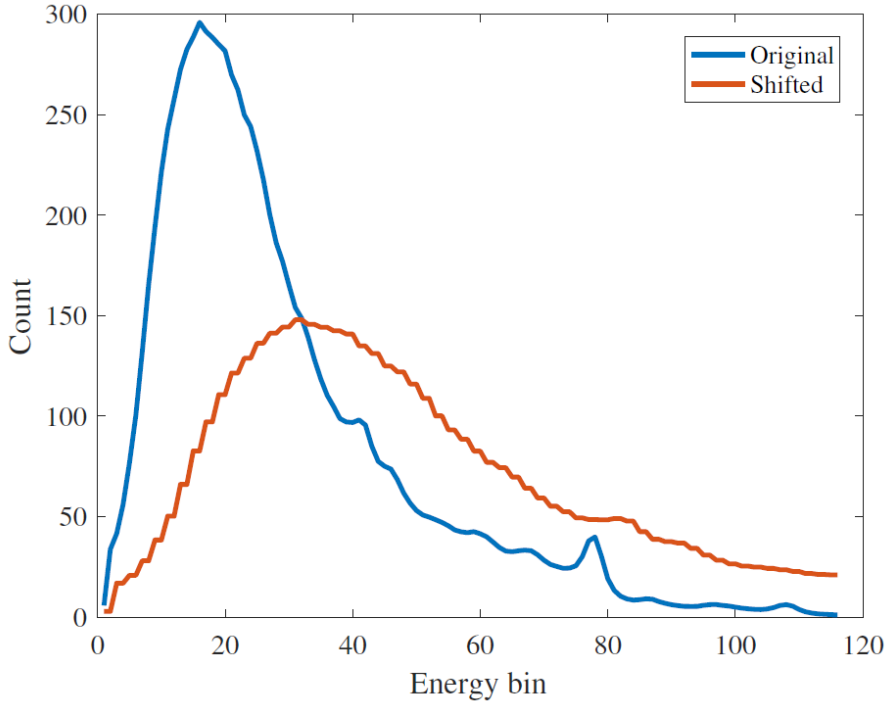




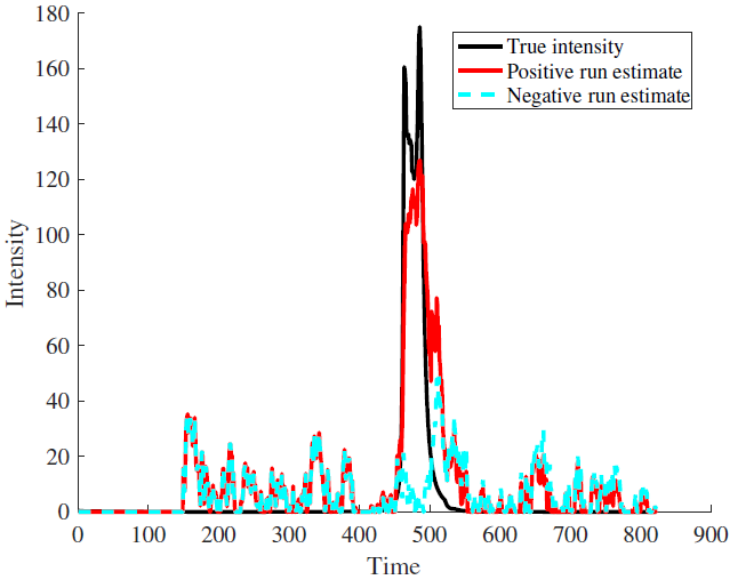
Confidence intervals from bootstrapped passes.

Evaluating the Kalman Filter with Training-Test Mismatch

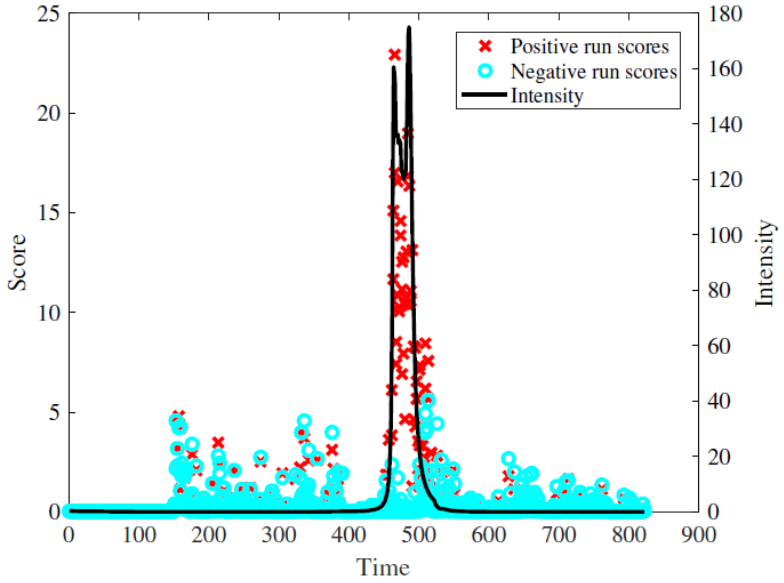
- The Kalman filter does not require training data, but other methods usually do
- It may be naïve to assume that training data match test data
- To induce mismatch between training and test background, test spectra were shifted to higher energy bins, similar to an extreme form of gain drift
- We compared several methods:
  - Oracle: Likelihood ratio using exact background rates and intensity
  - Optimal GP: GP with perfect prior
  - **Kalman GP** (KGP): GP with prior set by KF
  - Moving Average GP (MA GP): Like KGP but with simple moving average
  - GP: GP method with prior set by training data
  - Naïve KGP: KGP with non-adaptive covariance hyperparameters estimated from training data
  - **Intensity**: Intensity estimated by KF



Examining Estimated Intensity and Scores



(a) Kalman filter intensity estimates.

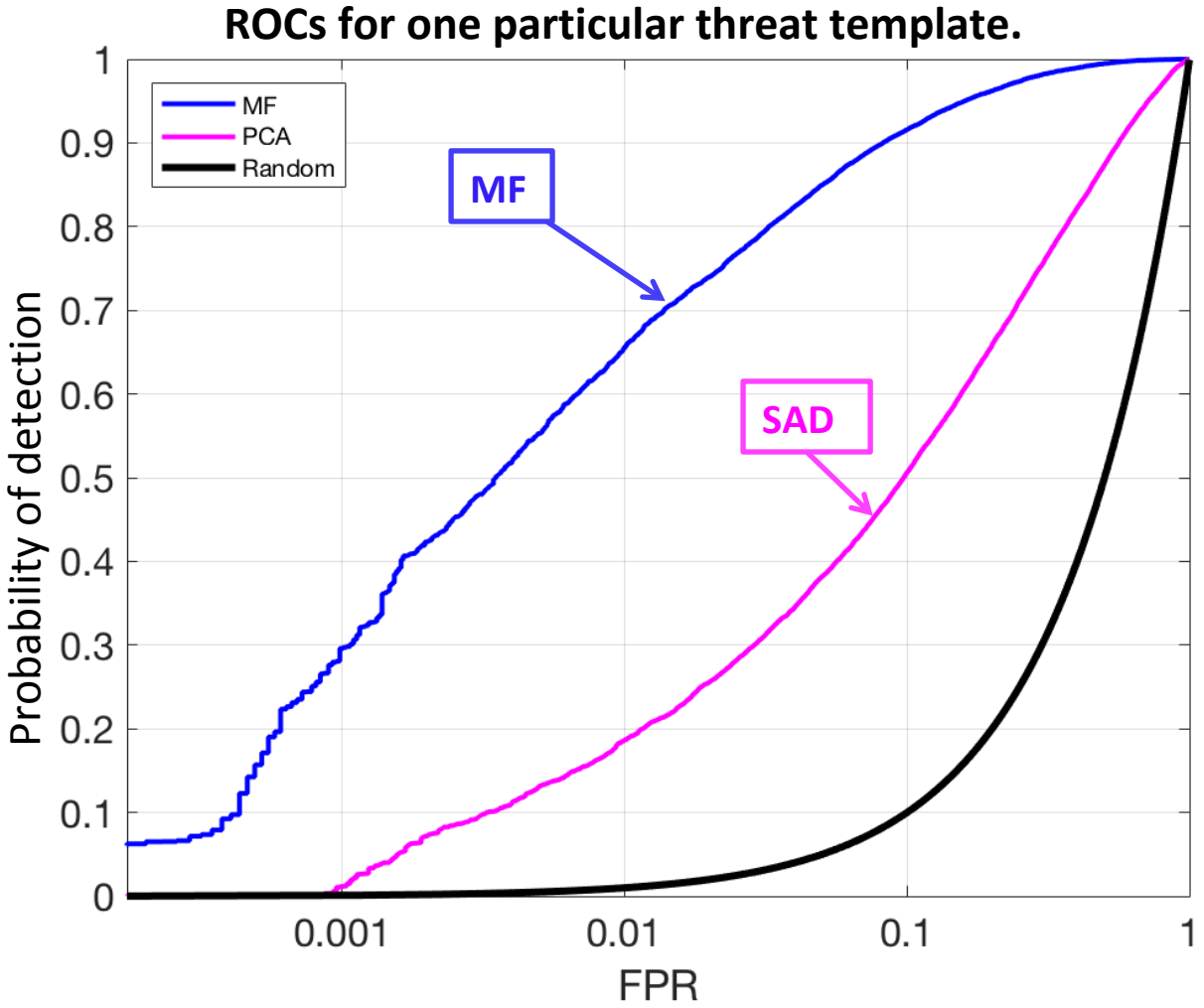


(b) KGP scores.

- Estimates of intensity were compared to the true intensity, which spiked when the detector moved near the source
- Intensity and KGP scores tracked the true intensity well with low lag
- There were not large spikes when there was no source

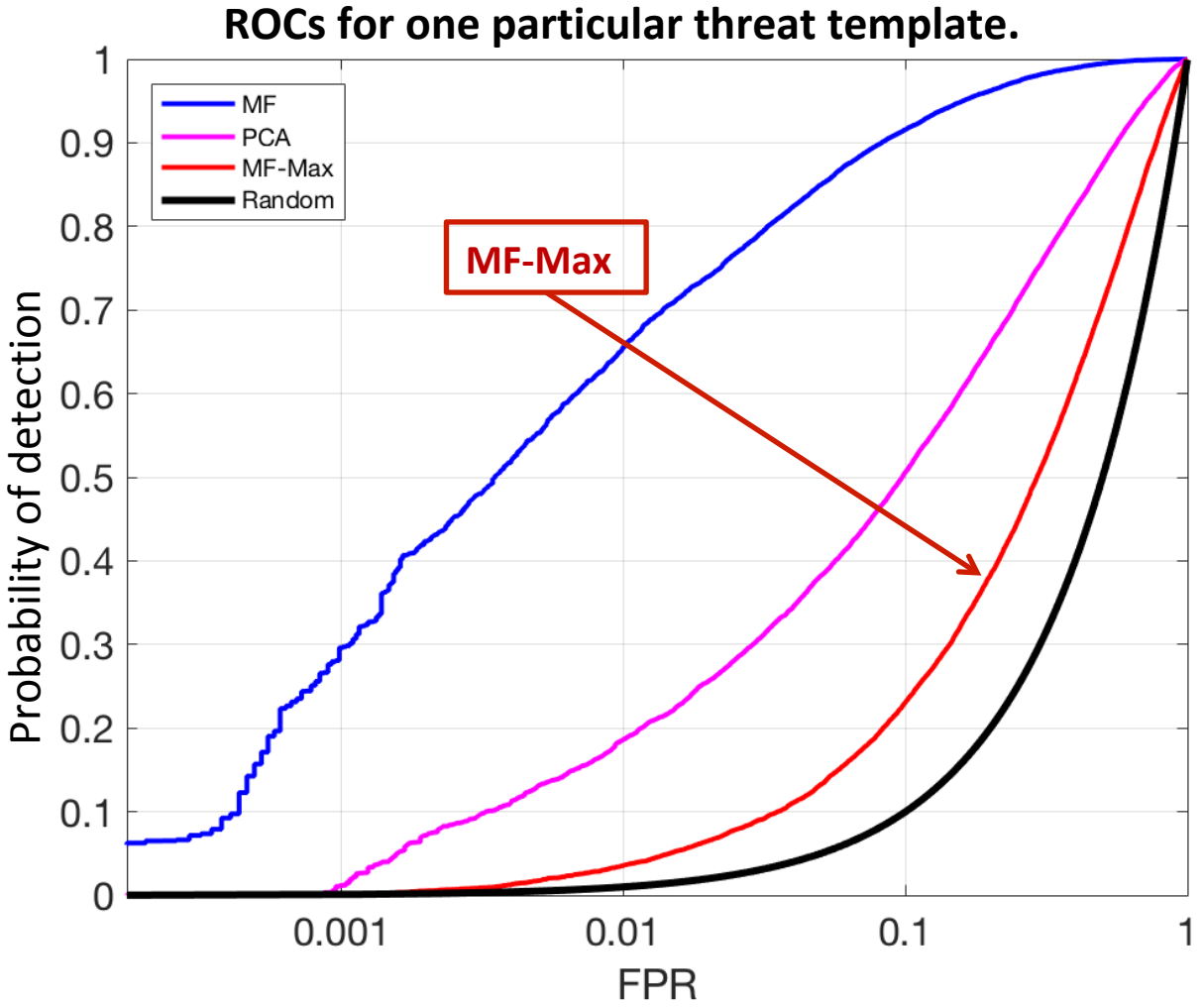
Alternative Circumstances of Source Detection

- If we do not know what threat design to expect, we can use Spectral Anomaly Detection (SAD or PCA) (Tandon 2016).
- If we have perfect knowledge of the shape of threat spectrum, we can use a Matched Filter (MF) (Tandon 2016)
- In practice, we often have an idea of what threat to expect, but our knowledge of it is usually imperfect



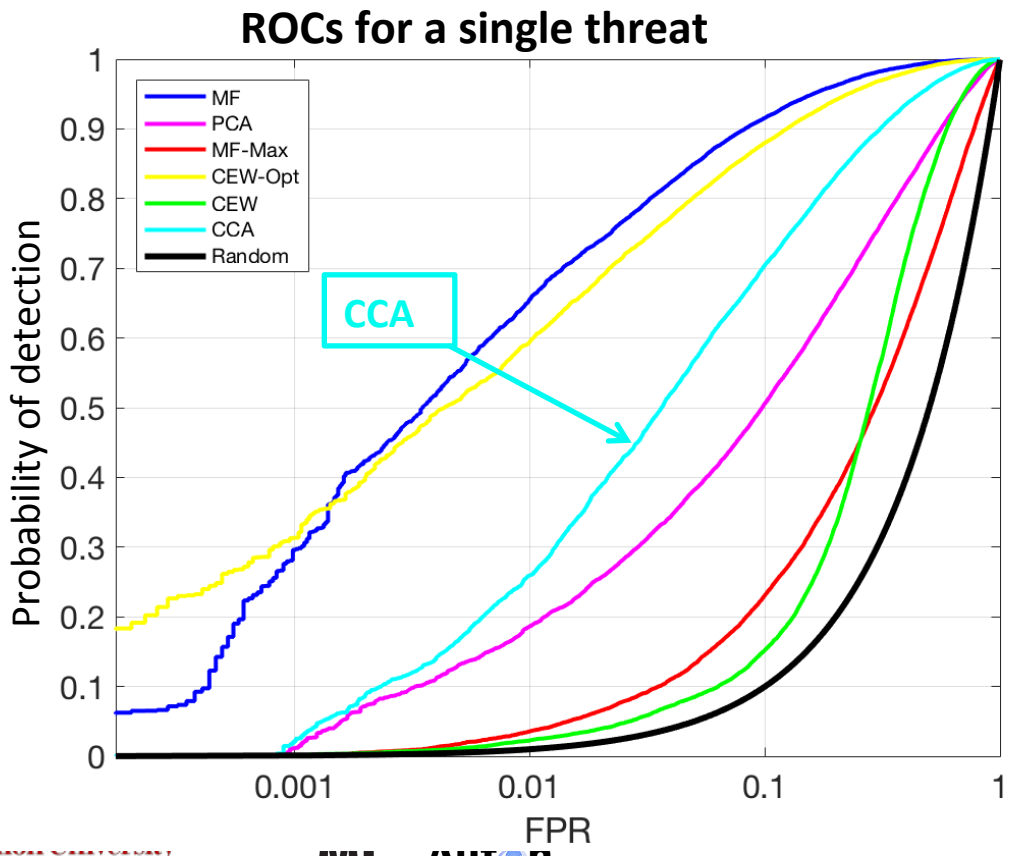
Alternative Circumstances of Source Detection

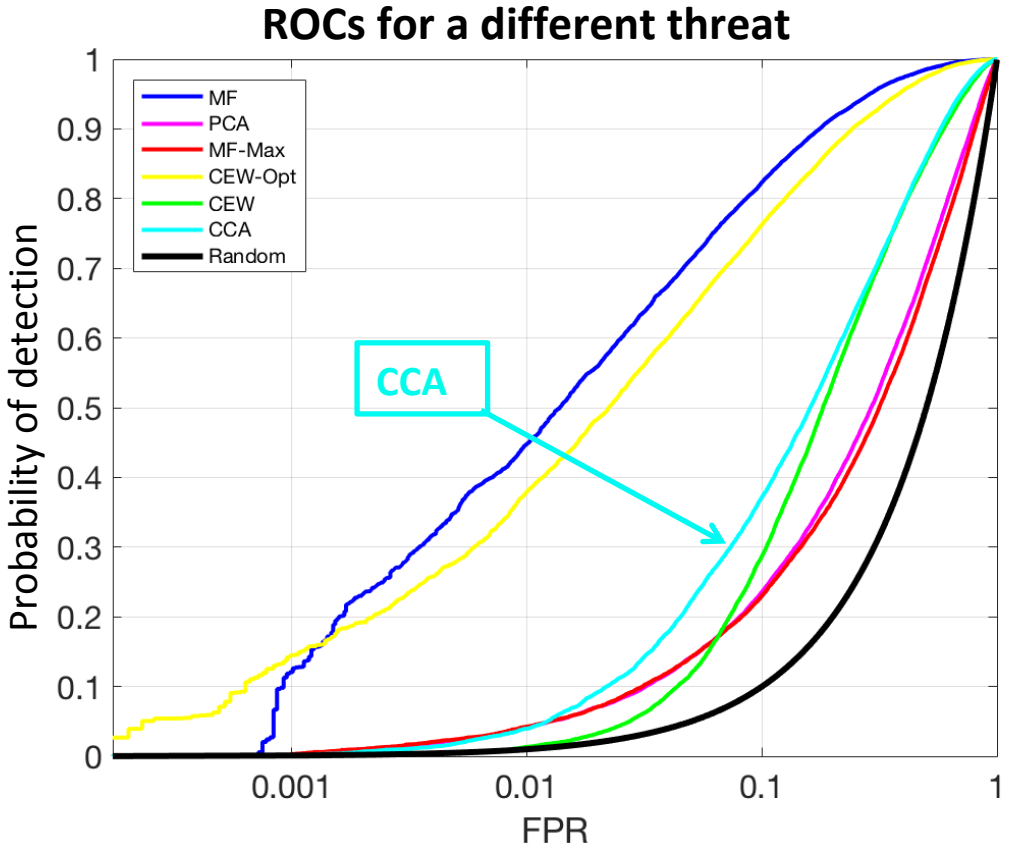
- If we can predict the variety of possible threat templates and form a library of threat templates, we can use marginalized version of Matched Filter, i.e., MF-Max
  - It would work as well as MF if marginalization always correctly picked the right threat template to use



Simulations with Imperfect Information

- We compare MF-Max, CEW, and CCA where we marginalize over a threat library that does not contain the actual threat.
- Our CCA method yields improved performance closer to the optimal information case.



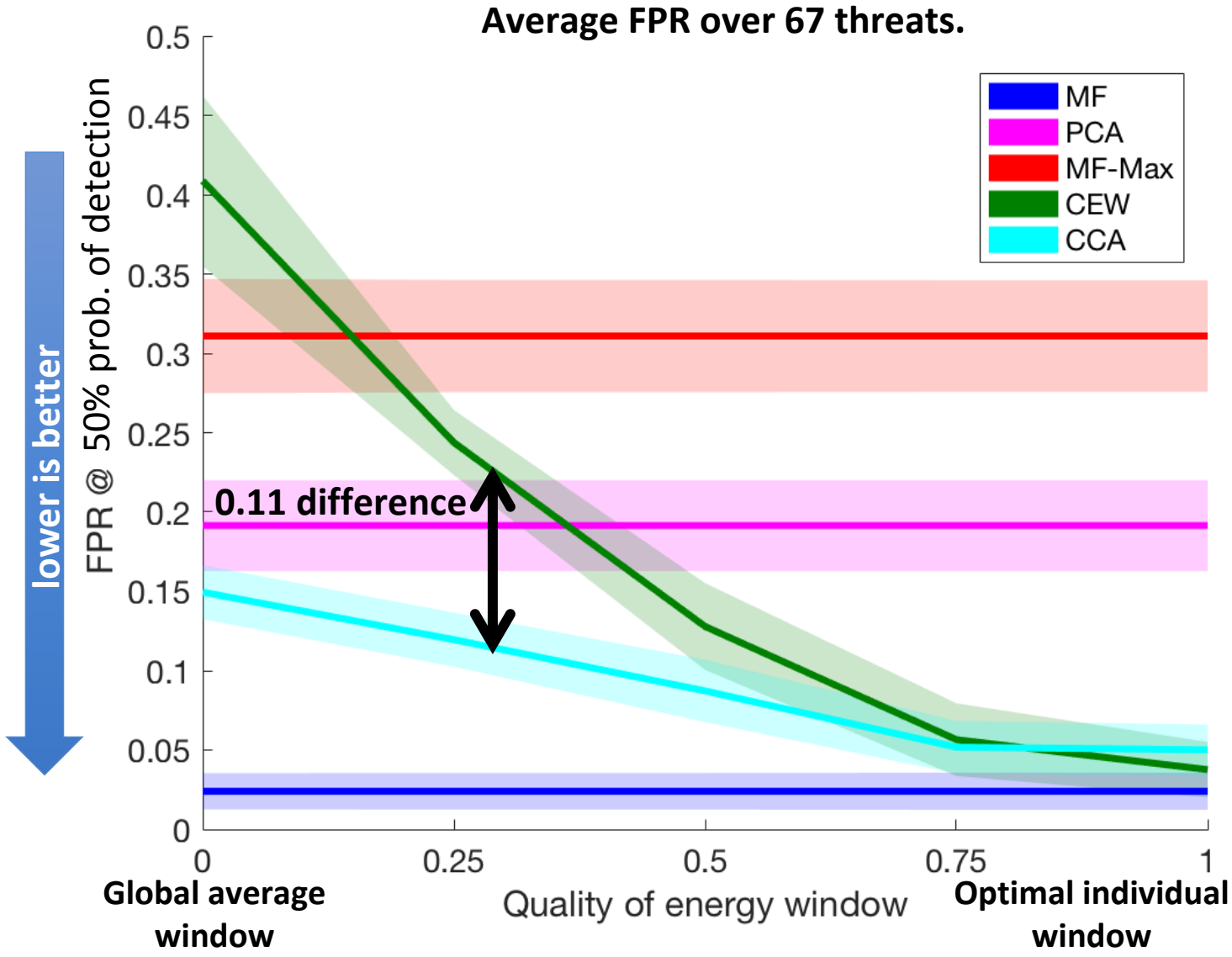


Changing the Energy Window Quality

1. Compute global average threat template
2. Compute convex combinations of average template and actual template
3. Find energy windows of combination templates and pass to CEW and CCA

Changing the Energy Window Quality

- FPR for each method as the window changes from low-quality to optimal
- As information about the threat spectrum decreases, the performance of CEW degrades and becomes much worse than CCA
- (Other methods do not use a window)



Existing approaches include CCA

- A common approach is component analysis, such as Canonical Correlation Analysis (CCA), which fits a linear correlation model between two views

View 1

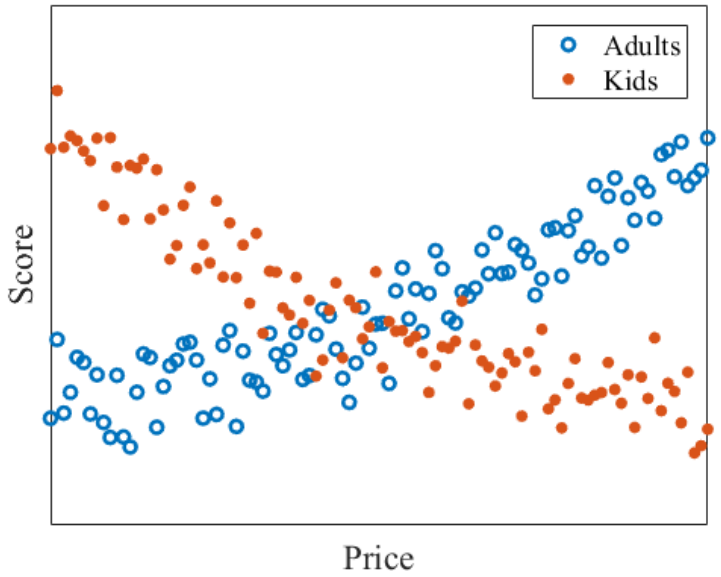
	A	B	C	D	E	F	G	H	I
1		row	range	center_lat	center_lo	plot_id	pi	SF17_VGI	SF17_VGI
2	0	1	1	34.62283	-82.733	SF17-BE-p	Top76	-1	-1
3	1	5	1	34.62281	-82.733	SF17-BE-p	PI_569459	60	-0.01176
4	2	9	1	34.62278	-82.733	SF17-BE-p	PI_569244	66	-0.01835
5	3	13	1	34.62275	-82.733	SF17-BE-p	PI_656035	61	-0.01462
6	4	17	1	34.62273	-82.733	SF17-BE-p	PI_329300	62	0.023038
7	5	21	1	34.6227	-82.733	SF17-BE-p	PI_156178	7	0.012624
8	6	25	1	34.62267	-82.733	SF17-BE-p	PI_329646	75	-0.00574
9	7	29	1	34.62265	-82.733	SF17-BE-p	PI_643016	5	0.077219
10	8	33	1	34.62262	-82.733	SF17-BE-p	PI_255744	-1	-1
11	9	37	1	34.6226	-82.733	SF17-BE-p	PI_152651	-1	-1
12	10	41	1	34.62257	-82.733	SF17-BE-p	Fill	134	-0.0129
13	11	45	1	34.62254	-82.733	SF17-BE-p	PI_569459	-1	-1
14	12	49	1	34.62252	-82.733	SF17-BE-p	PI_569244	30	0.024037
15	13	53	1	34.62249	-82.733	SF17-BE-p	PI_656035	55	-0.01009
16	14	57	1	34.62247	-82.733	SF17-BE-p	PI_329300	-1	-1
17	15	61	1	34.62244	-82.733	SF17-BE-p	PI_156178	-1	-1
18	16	65	1	34.62241	-82.733	SF17-BE-p	PI_329646	77	0.032889
19	17	69	1	34.62239	-82.733	SF17-BE-p	PI_643016	-1	-1
20	18	73	1	34.62236	-82.7329	SF17-BE-p	PI_255744	-1	-1
21	19	77	1	34.62233	-82.733	SF17-BE-p	PI_152651	-1	-1

View 2

	A	B	C	D	E	F	G	H	I
1	phenotyp	date	northing	easting	gridnum	gridletter	row	range	reading
2	SF17_VGI	6/28/2017	341136.2	3832526	17	S	81	1	0.050488
3	SF17_VGI	6/28/2017	341136.1	3832526	17	S	81	1	0.051184
4	SF17_VGI	6/28/2017	341136.1	3832526	17	S	81	1	0.053378
5	SF17_VGI	6/28/2017	341136.1	3832526	17	S	81	1	0.054087
6	SF17_VGI	6/28/2017	341136	3832526	17	S	81	1	0.053768
7	SF17_VGI	6/28/2017	341136	3832526	17	S	81	1	0.055059
8	SF17_VGI	6/28/2017	341136	3832526	17	S	81	1	0.056373
9	SF17_VGI	6/28/2017	341136	3832526	17	S	81	1	0.056966
10	SF17_VGI	6/28/2017	341135.9	3832526	17	S	81	1	0.058119
11	SF17_VGI	6/28/2017	341135.9	3832526	17	S	81	1	0.058122
12	SF17_VGI	6/28/2017	341135.9	3832526	17	S	81	1	0.05825
13	SF17_VGI	6/28/2017	341135.8	3832526	17	S	81	1	0.059209
14	SF17_VGI	6/28/2017	341135.8	3832526	17	S	81	1	0.058273
15	SF17_VGI	6/28/2017	341135.8	3832526	17	S	81	1	0.058624
16	SF17_VGI	6/28/2017	341135.8	3832526	17	S	81	1	0.058834
17	SF17_VGI	6/28/2017	341135.7	3832526	17	S	81	1	0.0584
18	SF17_VGI	6/28/2017	341135.6	3832526	17	S	81	1	0.0519
19	SF17_VGI	6/28/2017	341135.6	3832526	17	S	81	1	0.05027
20	SF17_VGI	6/28/2017	341135.5	3832526	17	S	81	1	0.050918
21	SF17_VGI	6/28/2017	341135.5	3832526	17	S	81	1	0.052112

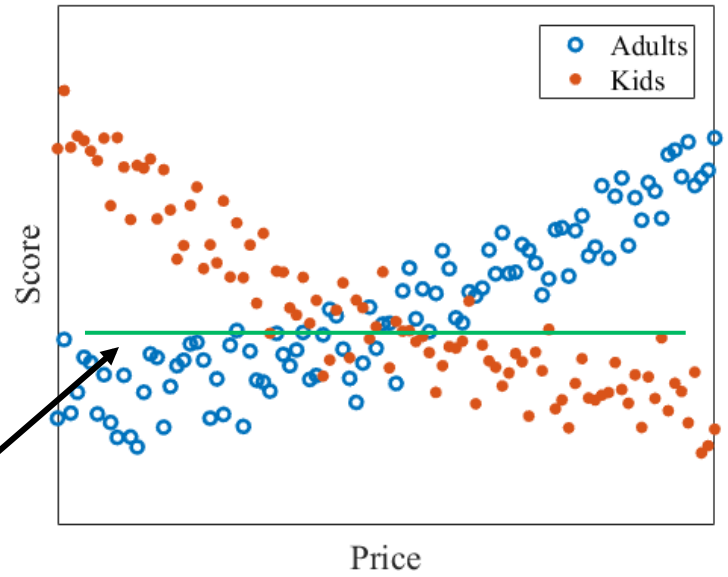
Restaurant example

- Restaurant characterized only by price
- Review characterized only by score
- Two kinds of reviewers: adults and kids
- A reviewer's score is a monotonic function of price
  - Increasing in price for adults
  - Decreasing in price for kids
- We observe restaurant price and review score but not the type of reviewer
- (One variable per view in this example, but usually multivariate)

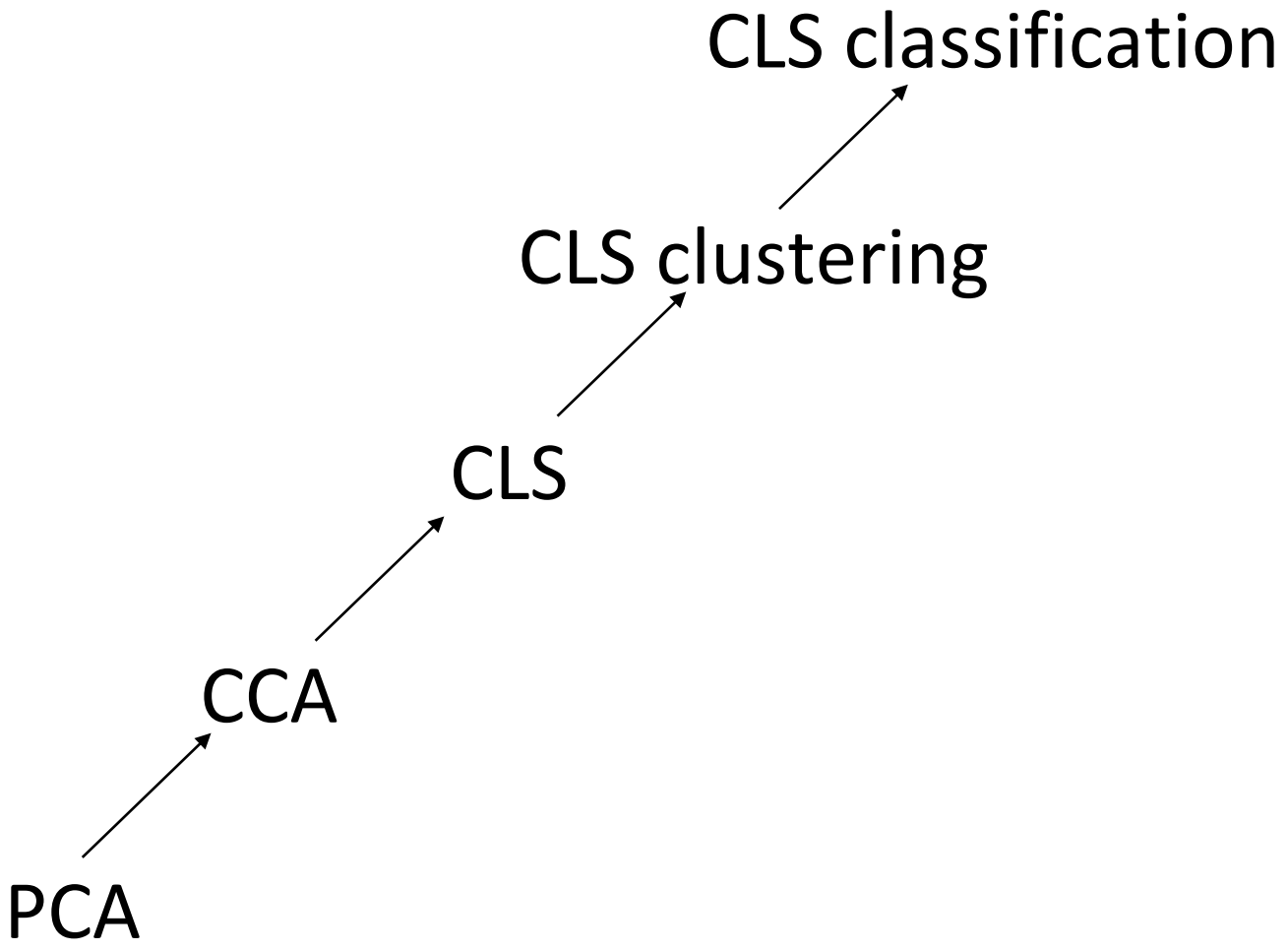


Restaurant example (cont.)

- Given restaurants and reviews, task is to identify the relationship between them
- Since there are two populations of reviewers, this relationship varies
- Existing approaches like CCA may struggle because they only search for global linear structure



Roadmap to our approach



Principal components analysis and canonical correlation analysis

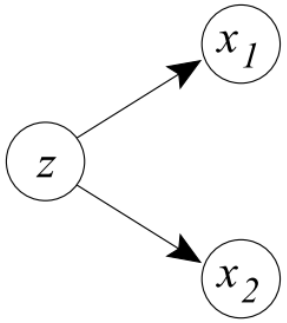
PCA

- Analyzes directions of maximum variance in a single view
- Decomposes view into linear combinations of variables
- Finds multiple orthogonal loadings
- Components are ranked by contribution to variance

CCA

- Analyzes directions of maximum correlation between two views
- Decomposes each view into linear combinations of variables
- Finds multiple orthogonal components
- Components are ranked by contribution to covariance

Formal statement of CCA



- Non-convex optimization with closed-form solution:

$$\max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \operatorname{Corr}(X^\mathsf{T} u, Y^\mathsf{T} v)$$

- $X^\mathsf{T}u$ and $Y^\mathsf{T}v$ are latent factors called canonical variables
- The solution for u is the leading eigenvector of

$$A = \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\mathsf{T}$$

  and similarly for v

- There are multiple components: the mth factor can be found by solving

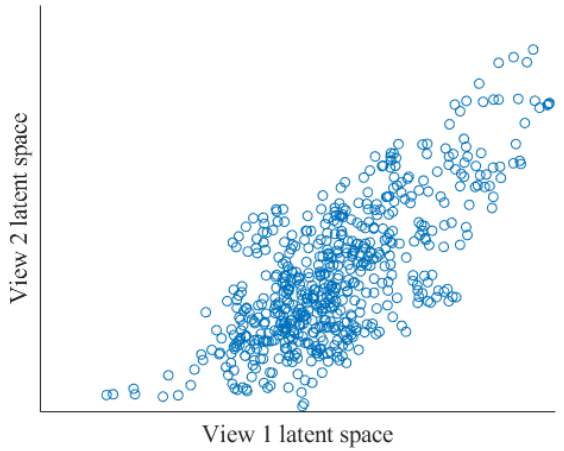
$$\begin{aligned} \max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \operatorname{Corr}(X^\mathsf{T} u, Y^\mathsf{T} v) \\ \text{subject to} \quad & \operatorname{Cov}(Xu, Xu_i) = \operatorname{Cov}(Yv, Yv_i) = 0, \\ & i = 1, \ldots, m-1. \end{aligned}$$

Recap of the linear approach

- A common approach is component analysis, such as Canonical Correlation Analysis (CCA), which fits a linear correlation model between two views

View 1

	A	B	C	D	E	F	G	H	I
1		row	range	center_lat	center_lo	plot_id	pi	SF17_VGI_	SF17_VGI_
2	0	1	1	34.62283	-82.733	SF17-BE-p	Top76	-1	-1
3	1	5	1	34.62281	-82.733	SF17-BE-p	PI_569459	60	-0.01176
4	2	9	1	34.62278	-82.733	SF17-BE-p	PI_569244	66	-0.01835
5	3	13	1	34.62275	-82.733	SF17-BE-p	PI_656035	61	-0.01462
6	4	17	1	34.62273	-82.733	SF17-BE-p	PI_329300	62	0.023038
7	5	21	1	34.6227	-82.733	SF17-BE-p	PI_156178	7	0.012624
8	6	25	1	34.62267	-82.733	SF17-BE-p	PI_329646	75	-0.00574
9	7	29	1	34.62265	-82.733	SF17-BE-p	PI_643016	5	0.077219
10	8	33	1	34.62262	-82.733	SF17-BE-p	PI_255744	-1	-1
11	9	37	1	34.6226	-82.733	SF17-BE-p	PI_152651	-1	-1
12	10	41	1	34.62257	-82.733	SF17-BE-p	Fill	134	-0.0129
13	11	45	1	34.62254	-82.733	SF17-BE-p	PI_569459	-1	-1
14	12	49	1	34.62252	-82.733	SF17-BE-p	PI_569244	30	0.024037
15	13	53	1	34.62249	-82.733	SF17-BE-p	PI_656035	55	-0.01009
16	14	57	1	34.62247	-82.733	SF17-BE-p	PI_329300	-1	-1
17	15	61	1	34.62244	-82.733	SF17-BE-p	PI_156178	-1	-1
18	16	65	1	34.62241	-82.733	SF17-BE-p	PI_329646	77	0.032889
19	17	69	1	34.62239	-82.733	SF17-BE-p	PI_643016	-1	-1
20	18	73	1	34.62236	-82.7329	SF17-BE-p	PI_255744	-1	-1
21	19	77	1	34.62233	-82.733	SF17-BE-p	PI_152651	-1	-1



View 2

	A	B	C	D	E	F	G	H	I
1	phenotyp	date	northing	easting	gridnum	gridletter	row	range	reading
2	SF17_VGI_	6/28/2017	341136.2	3832526	17	S	81	1	0.050488
3	SF17_VGI_	6/28/2017	341136.1	3832526	17	S	81	1	0.051184
4	SF17_VGI_	6/28/2017	341136.1	3832526	17	S	81	1	0.053378
5	SF17_VGI_	6/28/2017	341136.1	3832526	17	S	81	1	0.054087
6	SF17_VGI_	6/28/2017	341136	3832526	17	S	81	1	0.053768
7	SF17_VGI_	6/28/2017	341136	3832526	17	S	81	1	0.055059
8	SF17_VGI_	6/28/2017	341136	3832526	17	S	81	1	0.056373
9	SF17_VGI_	6/28/2017	341136	3832526	17	S	81	1	0.056966
10	SF17_VGI_	6/28/2017	341135.9	3832526	17	S	81	1	0.058119
11	SF17_VGI_	6/28/2017	341135.9	3832526	17	S	81	1	0.058122
12	SF17_VGI_	6/28/2017	341135.9	3832526	17	S	81	1	0.05825
13	SF17_VGI_	6/28/2017	341135.8	3832526	17	S	81	1	0.059209
14	SF17_VGI_	6/28/2017	341135.8	3832526	17	S	81	1	0.058273
15	SF17_VGI_	6/28/2017	341135.8	3832526	17	S	81	1	0.058624
16	SF17_VGI_	6/28/2017	341135.8	3832526	17	S	81	1	0.058834
17	SF17_VGI_	6/28/2017	341135.7	3832526	17	S	81	1	0.0584
18	SF17_VGI_	6/28/2017	341135.6	3832526	17	S	81	1	0.0519
19	SF17_VGI_	6/28/2017	341135.6	3832526	17	S	81	1	0.05027
20	SF17_VGI_	6/28/2017	341135.5	3832526	17	S	81	1	0.050918
21	SF17_VGI_	6/28/2017	341135.5	3832526	17	S	81	1	0.052112

- However, CCA might struggle if the correlations are nonlinear or non-global

Mixture of CCA optimization

- This (simplified) optimization problem looks like

$$\min_{u,v,R} \quad \sum_j \|X^{(j)}u^{(j)} - Y^{(j)}v^{(j)}\|_2^2$$

$$\text{s.t.} \quad u^{(j)\top}X^{(j)\top}X^{(j)}u^{(j)} = v^{(j)\top}Y^{(j)\top}Y^{(j)}v^{(j)} = 1$$

- *R* is cluster labels
- $X^{(j)}$ and $Y^{(j)}$ are subsampled data matrices of cluster *j*
- PSD objective in *u*,*v* but quadratic constraints
- Alternative way of writing CCA that looks like least-squares

Canonical Least Squares

- CLS is our alternative to CCA that is better for clustering
- Minimizes squared error in latent space:

$$\min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \|Xu - Yv\|_2^2$$

$$\text{subject to} \quad v^\mathsf{T} v = 1.$$

- PSD objective and quadratic constraint
  - data-agnostic $v^\mathsf{T} v = 1$ inst $u^\mathsf{T} X^\mathsf{T} X u = v^\mathsf{T} Y^\mathsf{T} Y v = 1$
- Still has closed-form solution: *v* is the lowest eigenvector of $Y^\mathsf{T} H Y$ and $u = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y v$

  where $H = I - X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$
- Lowest eigenvector corresponds to minimizing variance of a regression residual

Multiple CLS components

- Like CCA, CLS can find multiple components
- PSD objective and quadratic constraints:

$$\min_{\substack{U \in \mathbb{R}^{d_X \times m} \\ V \in \mathbb{R}^{d_Y \times m}}} \|XU - YV\|_{\mathcal{F}}^2$$

$$\text{subject to} \quad V^\mathsf{T} V = I.$$

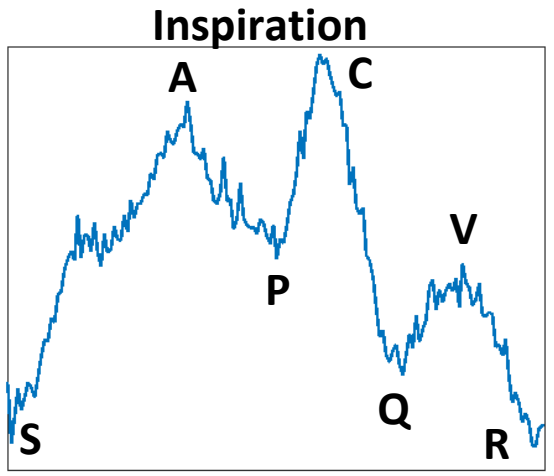
- No known closed-form solution
  - Constraint prevents 0 solution, does not normalize scale
  - Sensitive to scale, so usually standardize data
- We solve for components *V* via a greedy sequential approximation by taking *V* to be the lowest eigenvectors c$Y^\mathsf{T} HY$
- CCA was a multiple correlation problem, while CLS is a multi-output regression problem

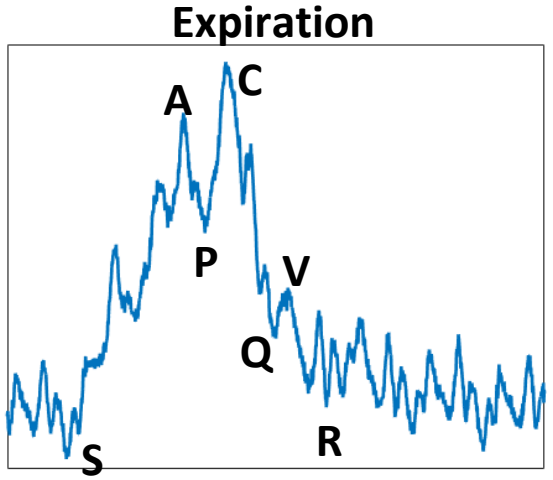
Convergence

- When number of components $m = 1$, algorithm is guaranteed to converge
  - Objective decreases at every step
  - For some problems, first component is most meaningful
- For higher m, the algorithm does not necessarily converge because of greedy approximation
  - Empirically not a problem

Analyzing blood loss with CVP

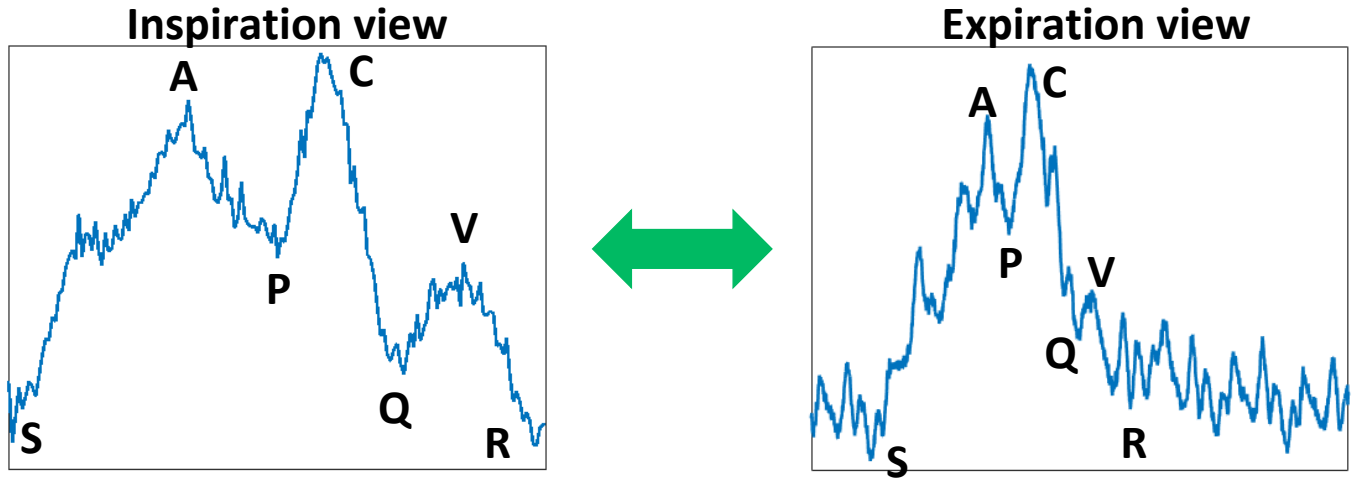
- Dataset contains time series of each pig's **central venous pressure** (CVP), blood pressure near right atrium of heart
- Each waveform is from either inspiration or expiration phase of respiration
- Thirteen features were extracted from each waveform as averages and ratios between different points of the CVP waveform





Unsupervised setting: correlating inspiration and expiration waveforms

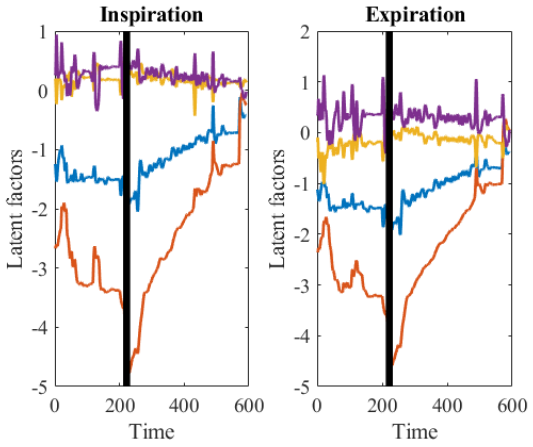
- We consider inspiration and expiration as the views
- We expect the correlations to still depend on bleeding, which is unobserved



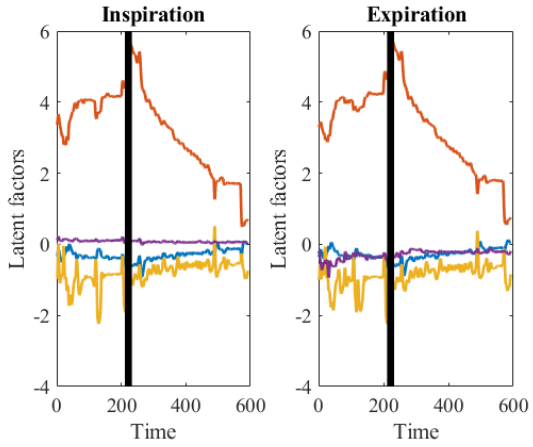
Unsupervised latent variables

— First component
— Second component
— Third component
— Fourth component

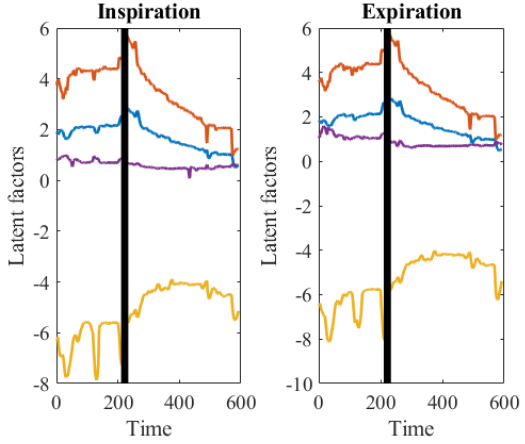
Cluster 1



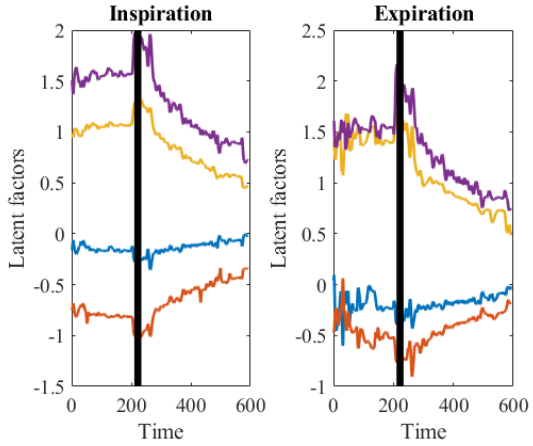
Cluster 2



Cluster 3



Cluster 4



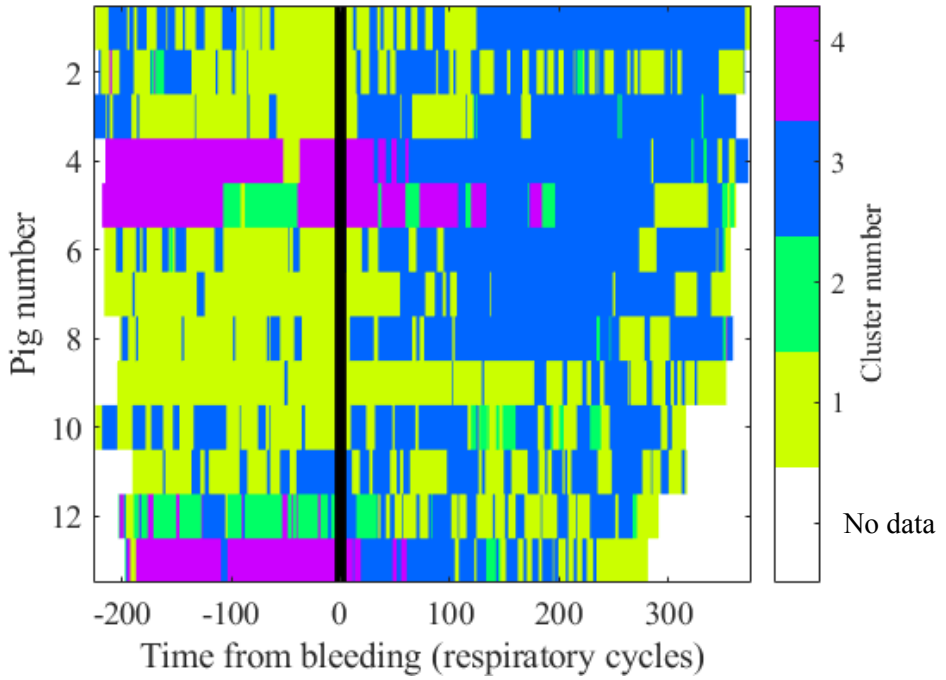
The latent variables resemble blood loss.

Supervised classification setting: correlating inspiration and expiration with knowing whether bleeding had started

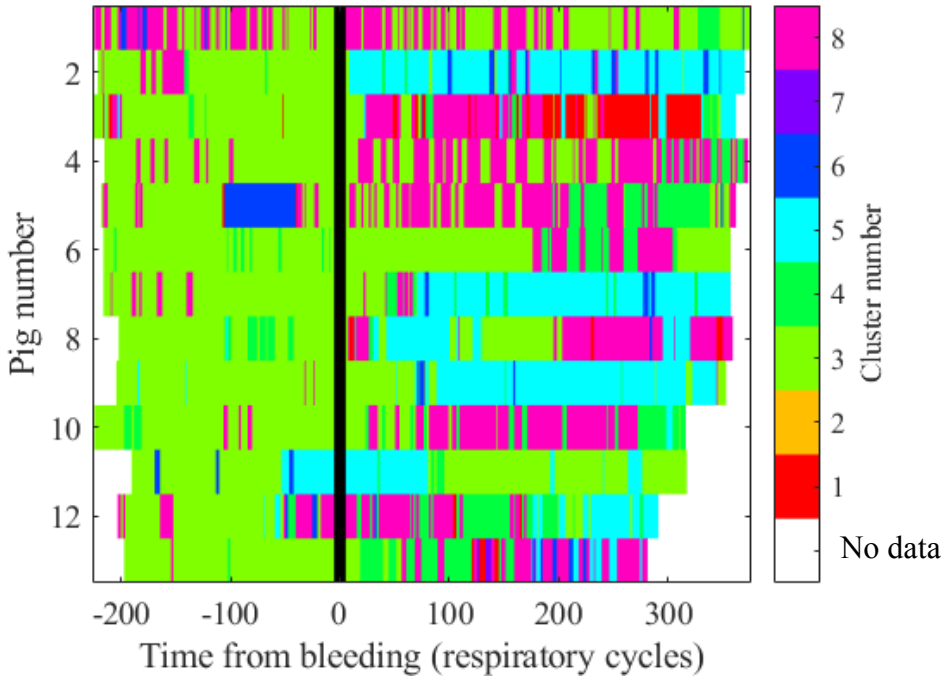
- Task: decide whether a pair of waveforms came from **before** or **after** onset of bleeding

Cluster assignments for test pigs

Unsupervised
(4 clusters)

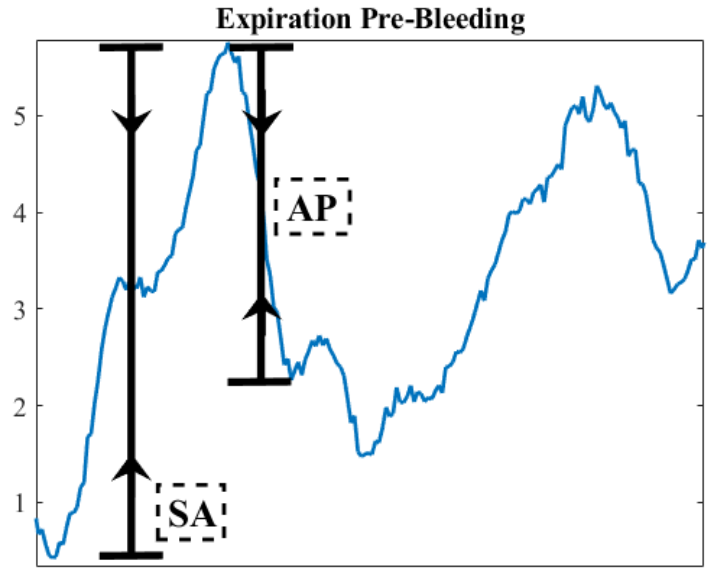


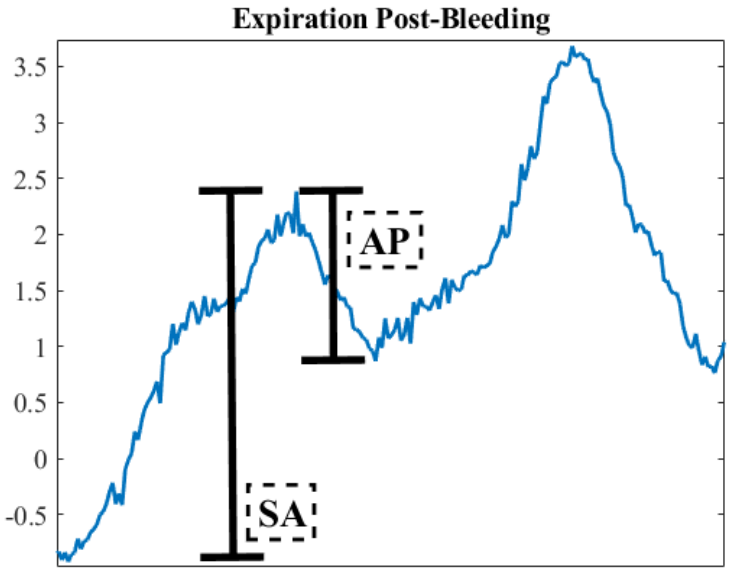
Supervised
(3 clusters non-bleeding, 5 clusters bleeding)



We identified pre-bleeding and post-bleeding clusters, which are usually distinct. The predominant pre-bleeding cluster is light green.

Waveform analysis





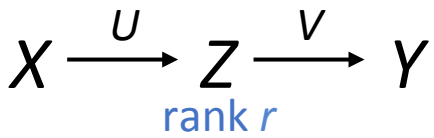
- We analyze the impact of waveform features by analyzing the gradient of classification score
- **Left**: Our model predicts that shrinking the marked lengths is correlated with bleeding in a pre-bleeding observation.
- **Right**: The corresponding lengths have shrunk in a post-bleeding observation.

Global regression model for single-view data

- Reduced-rank regression (RRR) is a multi-output regression method

$$\min_B \mathcal{L} = \|Y - XB\|_{\mathcal{F}}^2 \qquad \text{rank}(B) \leq r$$

- Has closed-form solution using eigendecomposition of XB_{OLS}
- Also can be written $\min_{U,V} \|Y - XUV^T\|_{\mathcal{F}}^2$



Low-rank latent space links inputs and outputs

Global regression model for multi-view data

- Inputs and coefficients are partitioned into *G* views $B = (B_1^\mathsf{T}, \ldots, B_G^\mathsf{T})^\mathsf{T}$
- Want each B_G to be "low rank" – group-wise low rank constraint

$$\min_B \frac{1}{2n}\|Y - XB\|_\mathcal{F}^2 + \lambda \sum_{g=1}^{G} w_g \|B_g\|_*$$

- $||M||_*$ is nuclear norm, the sum of singular values, a convex relaxation of rank
- Exploits multi-view structure because each view connects to a separate low-rank latent space

MoE with iRRR

- Nuclear norm regularization corresponds to a prior on expert parameters
- However, we are unsure if there is a valid probability distribution that leads to the nuclear norm penalty
- We use a pseudo-distribution that suffices mathematically

$$\Pr(B) = \prod_g \Pr(B_g)$$

$$\propto \prod_g \exp(-\lambda w_g \|B_g\|_*) = \exp(-\lambda \sum_g w_g \|B_g\|_*)$$

- From here, EM is straightforward

Multi-view relationships in sets of points

- iRRR does not use relationships between views
- We propose to use these relationships by weighting experts based on correlation structure
- Correlations are only defined on sets of points, so we assume the observations are already partitioned
- All points in a partition are given the same expert weights