Carnegie Mellon University
Auton Lab

Discovering Compact and Informative Structures through Data Partitioning

Madalina Fiterau

Thesis Proposal

15th October 2014

Thesis Committee
Artur Dubrawski
Geoff Gordon
Andreas Krause
Alex Smola

Sparse Predictive Structures

Considerable effort expended on building *complex models* from *vast* amounts of data, not enough to make models *comprehensible.*

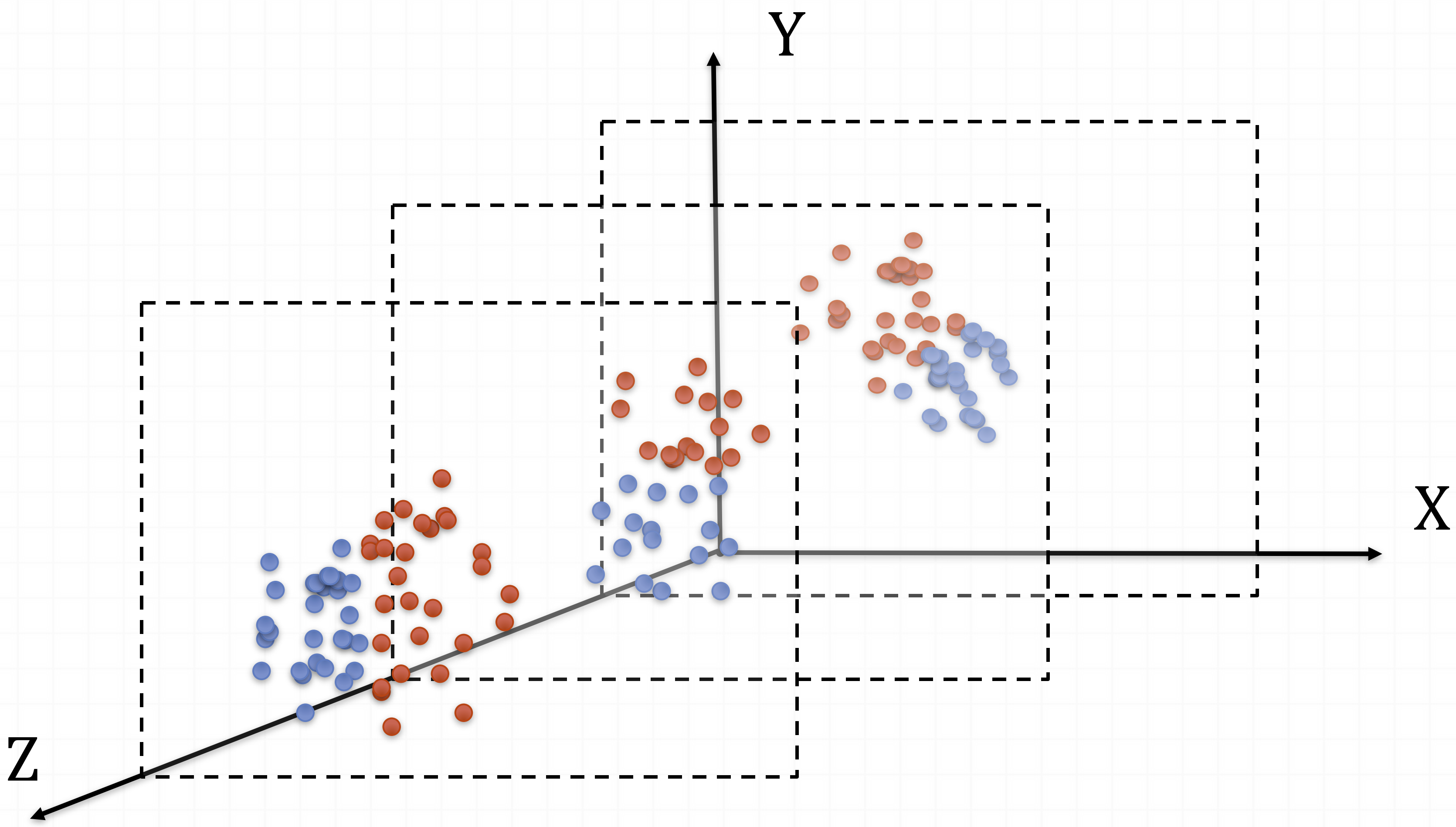
1. NEED COMPACT MODELS TO ENABLE ANALYSIS AND VISUALIZATION
2. LEVERAGING EXISTING STRUCTURE IN DATA → HIGH PERFORMANCE
3. COMPACT ENSEMBLES OF COMPLEMENTARY LOW-D SOLVERS

BORDER CONTROL

DIAGNOSTICS

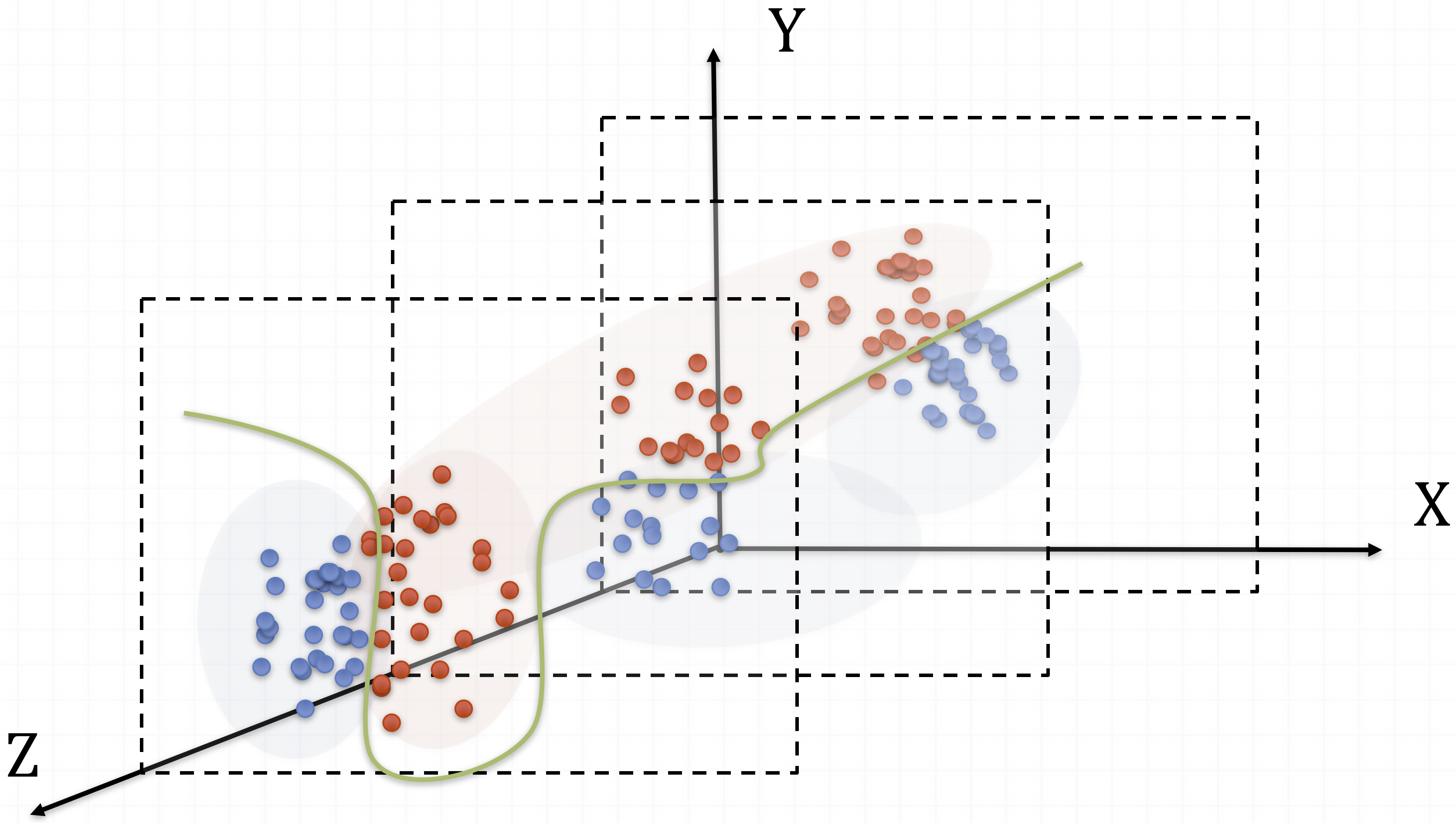
VEHICLE CHECKS

Sparse Predictive Structures



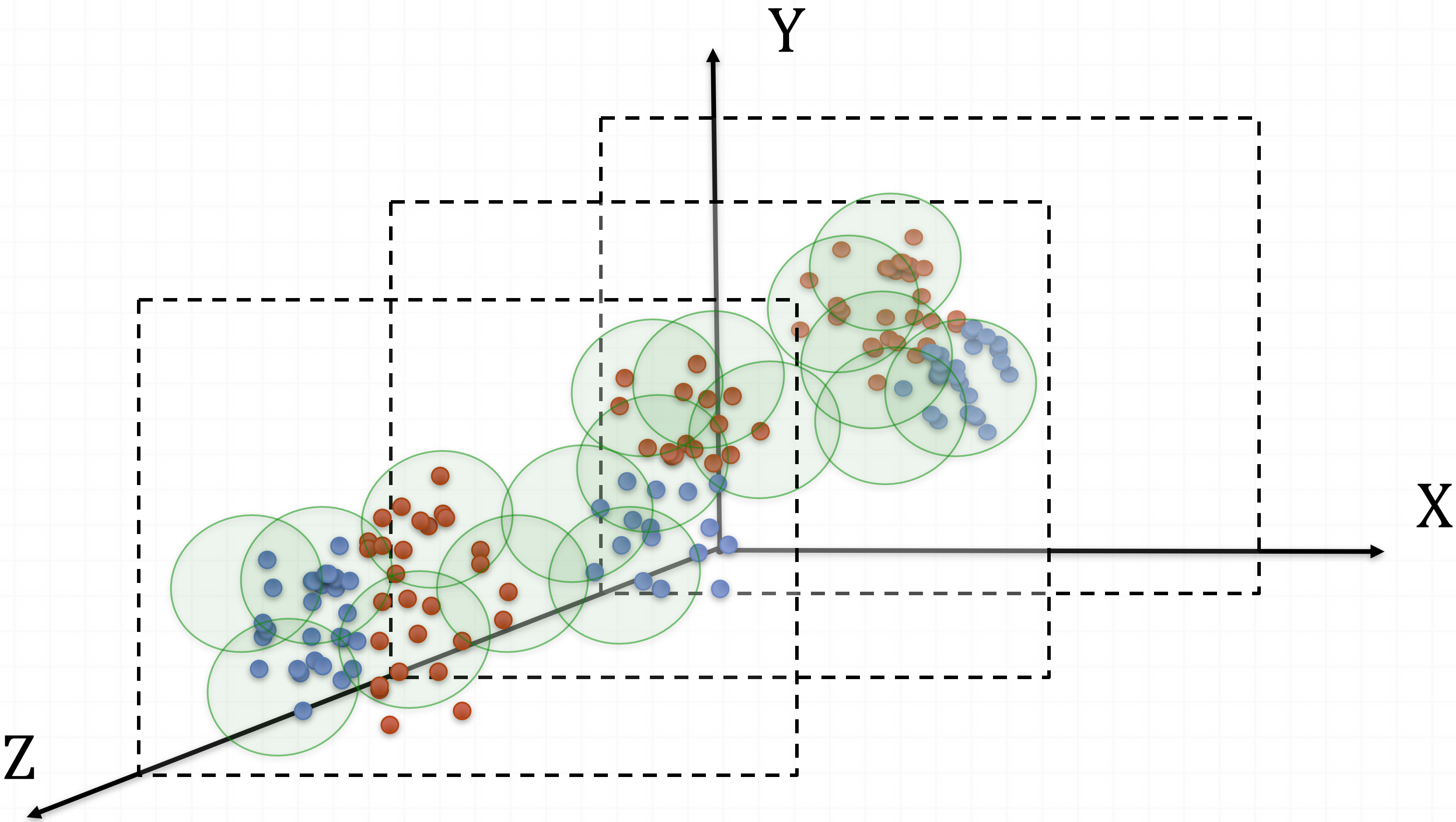
High dimensional data is often heterogeneous

Learning Sparse Predictive Structures



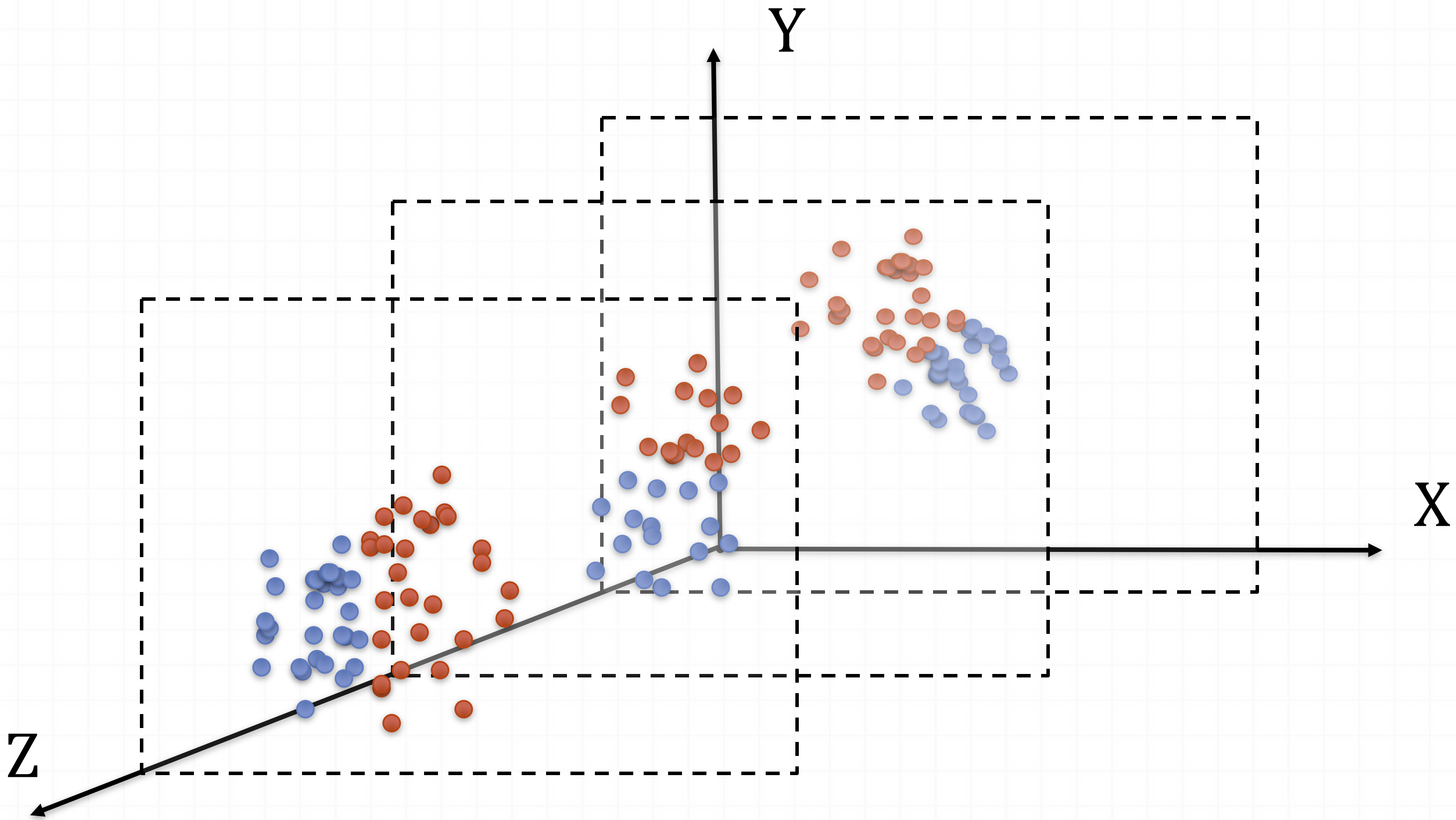
Global Models

Learning Sparse Predictive Structures



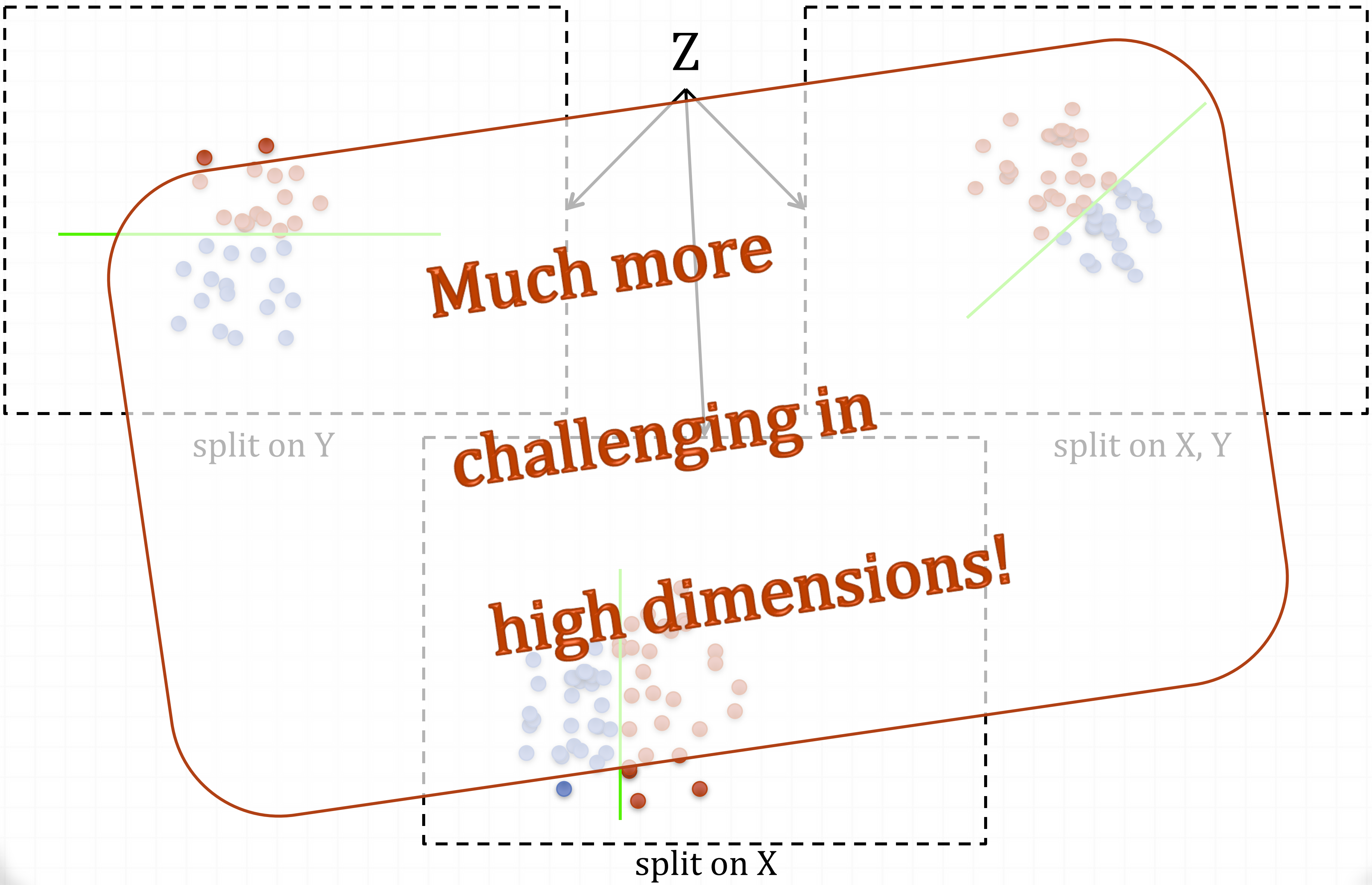
Local Models

Learning Sparse Predictive Structures



Trade-off: compact data partitioning models

Learning Sparse Predictive Structures



Thesis

It is possible to identify low dimensional structures in complex high-dimensional data, if such structures exist, and leverage them to construct compact interpretable models for various machine learning tasks.

Thesis Outline

Informative Projection Retrieval

- Projection Retrieval as a combinatorial problem
- Optimization procedure for IPR
- RIPR for classification, clustering, regression, active learning

Applying RIPR to Clinical Alert Classification

- Building interpretable classification models for clinical alerts
- Annotation Framework using Active RIPR

Proposed research

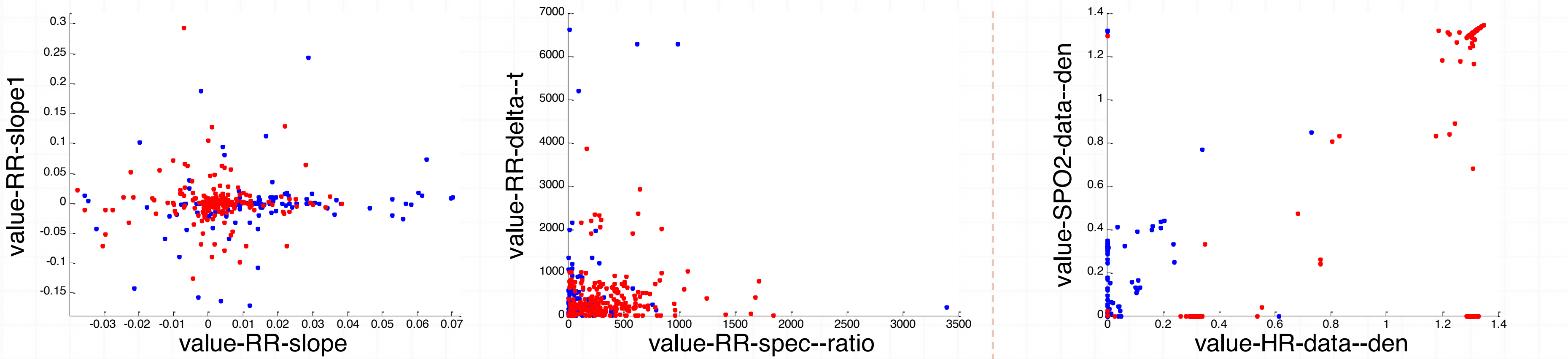
- IPR for multi-task learning and time series
- Low-dimensional model learning for feature hierarchies
- Online cost-constrained subset selection policies

Informative Projection Retrieval (IPR)

Projection Retrieval for a Learning Task

- problem of **selecting low-d (2D, 3D) subspaces**
- s.t. queries are resolved with **high-confidence**
- models perform the task with **low expected risk**

example: features represent vital signs and derived features; considering only the duty cycles of the signals might be sufficient



A multitude of projections where data is 'noisy'

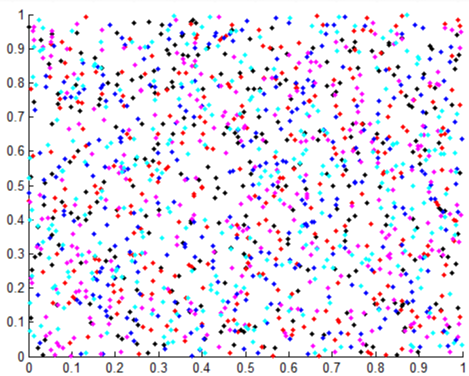
A small set where there is a clear separation

RIPR = Regression-based Informative Projection Retrieval*

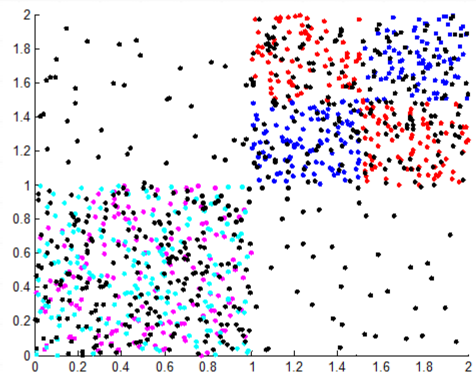
[1] Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In Advances in Neural Information Processing Systems 25 (NIPS), pages 3032–3040, 2012.

RIPR Target Datasets

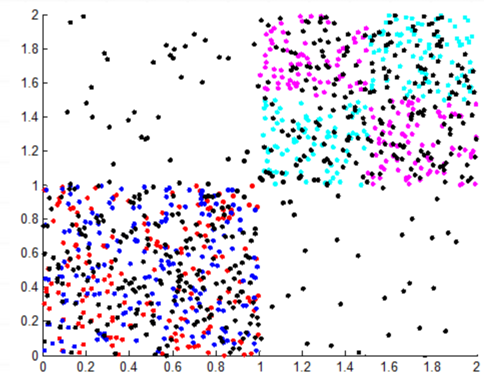
- Most of the low-dimensional projections are non-informative
- But there are at least a few with useful structure
- Each such structure could only involve a subset of data
- But jointly, these subsets cover all data



Aspect of most projections



IP for blue/red group

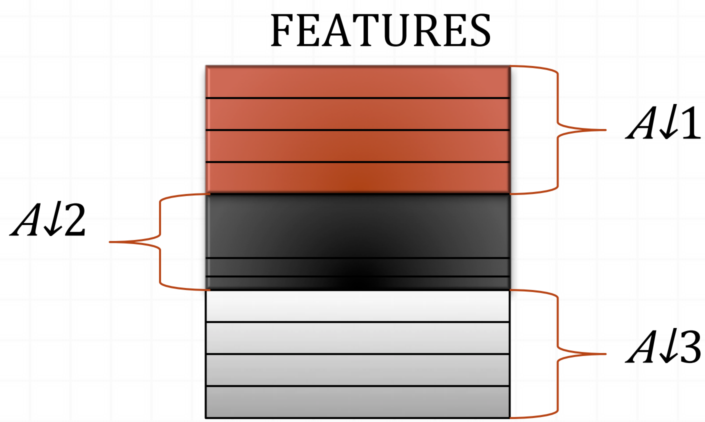


IP for light blue/purple group

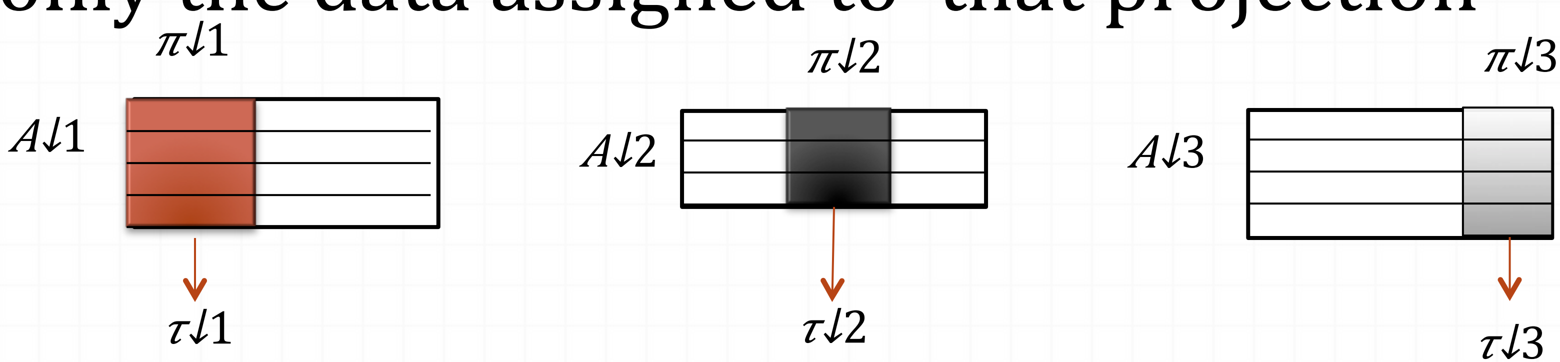
- Engineered data - unintentionally introduced artifacts usually show in low-dimensional patterns
- Clinical data - multiple sub-models reflect specifics of particular conditions and patient characteristics

A Dual-Objective Training Process

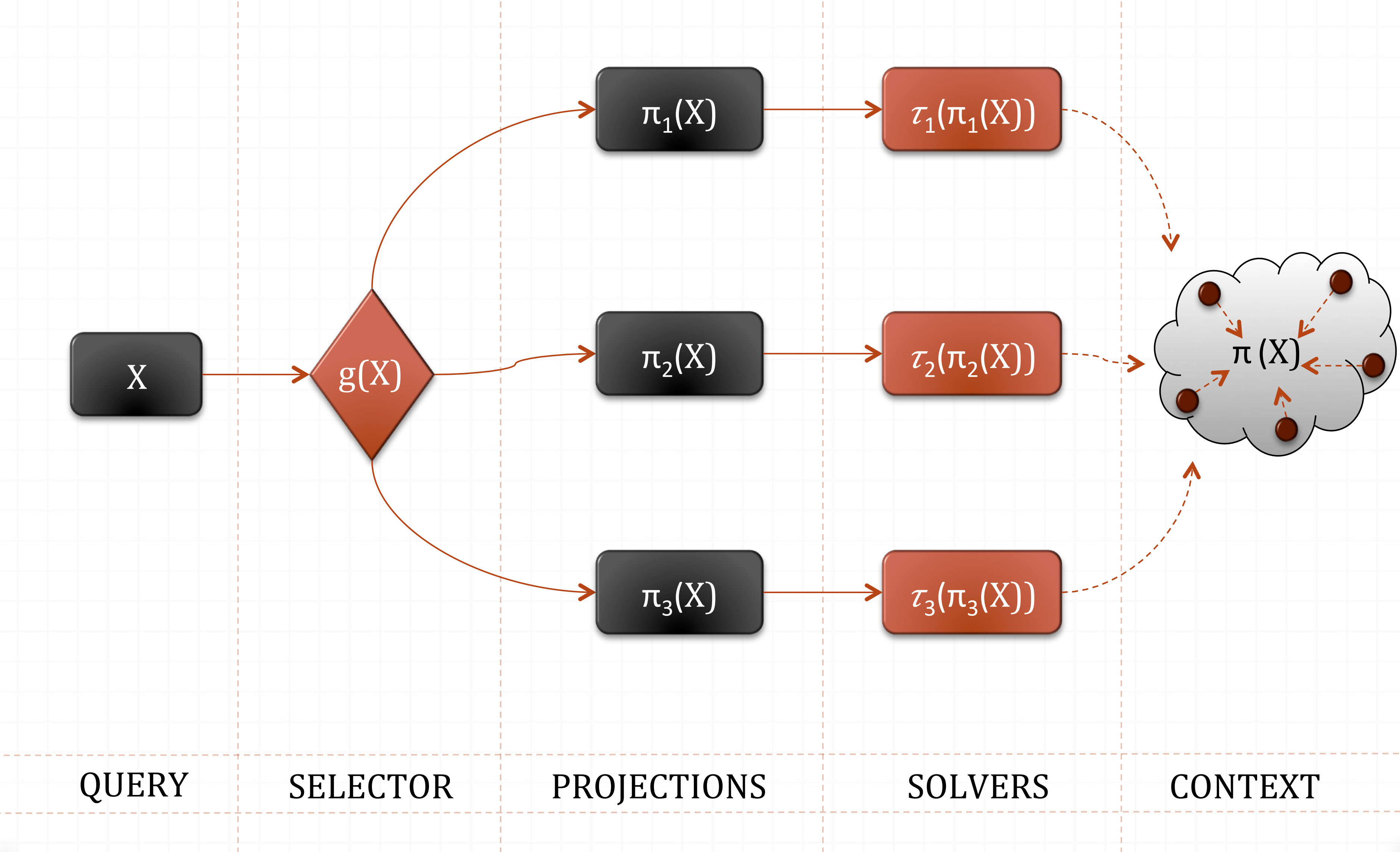
1. Data is split across informative projections



2. Each projection has a solver trained using only the data assigned to that projection



RIPR Framework



RIPR Model

Model components:

- Set of d-dimensional, axis-aligned sub-spaces of the original feature space P ϵ Π
- Each projection has an assigned solver of the task T; the solvers are selected from some solver class $\mathcal{T}$
- A selection function g, which yields, for a query point x, the projection/solver pair $(\pi_{g(x)}, \tau_{g(x)})$ for the point;
- $\ell(\tau_{g(x)}(\pi_{g(x)}(x)), y)$ represents the model loss at point x

Dataset ➤ $X = \{x_1 \ldots x_n\} \in \mathcal{X}^n$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^m$

$$\mathcal{M}_d = \{\Pi = \{\pi;\ \pi \in \mathbf{\Pi}, |\pi| \leq d\},$$
$$T = \{\tau;\ \tau_i \in \mathcal{T}, \tau_i : \pi_i(\mathcal{X}) \rightarrow \mathcal{Y} \quad \forall i = 1 \ldots |\Pi|\}$$
$$g \in \{f : \mathcal{X} \rightarrow \{1 \ldots |\Pi|\}\}\ \}\ .$$

Small set of projections

Solvers

Target model

Selection function

RIPR Objective Function

Model components:

- Set of d-dimensional, axis-aligned sub-spaces of the original feature space $P \in \Pi$
- Each projection has an assigned solver of the task T; the solvers are selected from some solver class $\mathcal{T}$
- A selection function g, which yields, for a query point x, the projection/solver pair $(\pi_{g(x)}, \tau_{g(x)})$ for the point;
- $\ell(\tau_{g(x)}(\pi_{g(x)}(x)), y)$ represents the model loss at point x

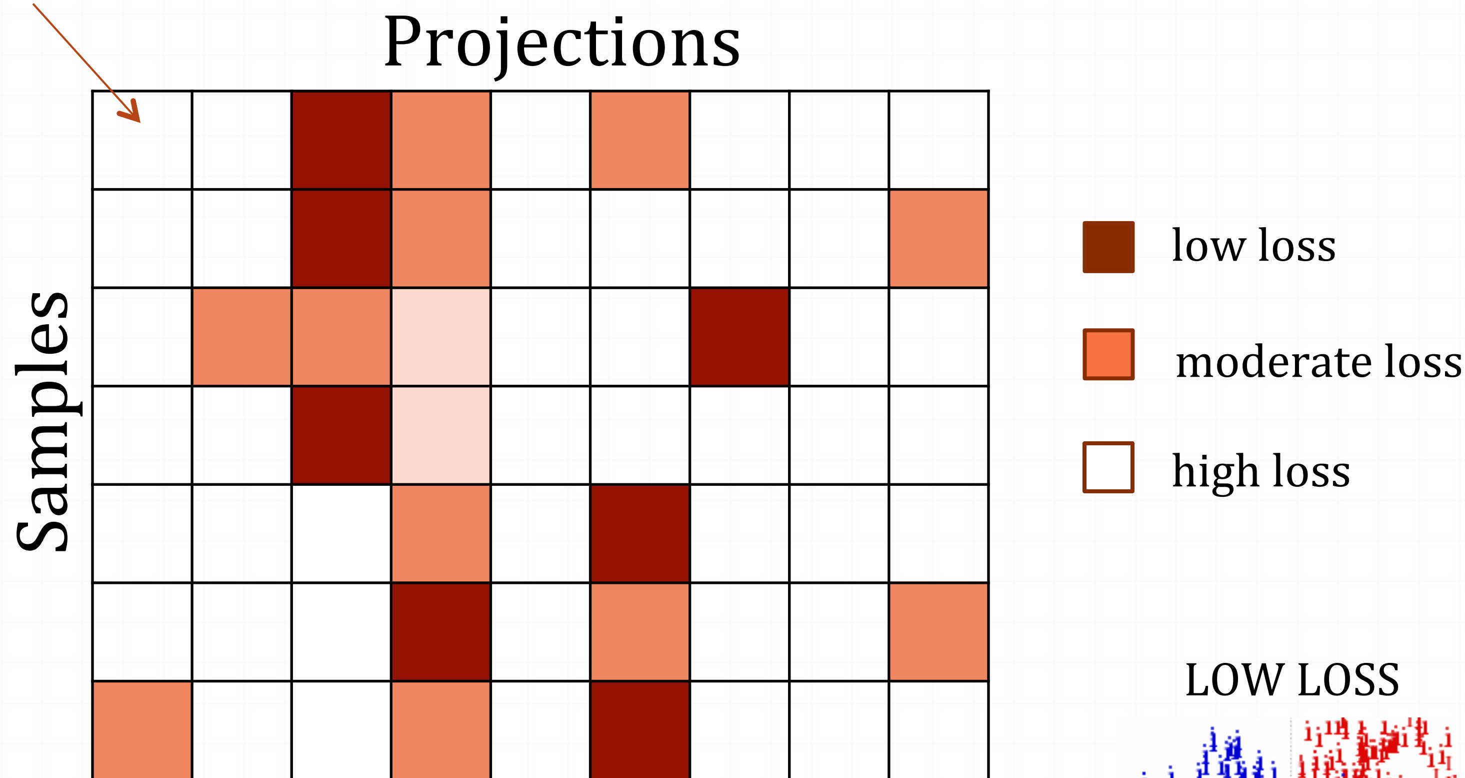
Minimization:

$$M^* = argmin_{M \in \mathcal{M}_d} \mathbb{E}_{\mathcal{X},\mathcal{Y}}\left[y \neq h_{g(x)}(\pi_{g(x)}(x)) \right]$$

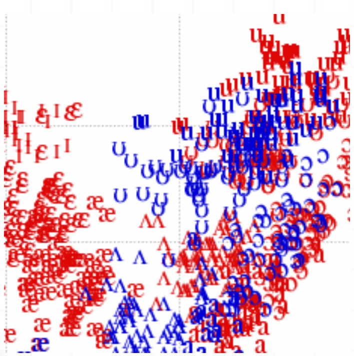
Expected loss for task solver trained on projection assigned to point

Starting point: the loss matrix

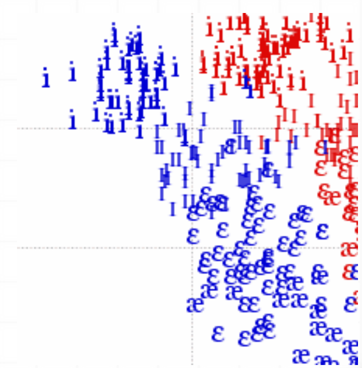
Loss estimators



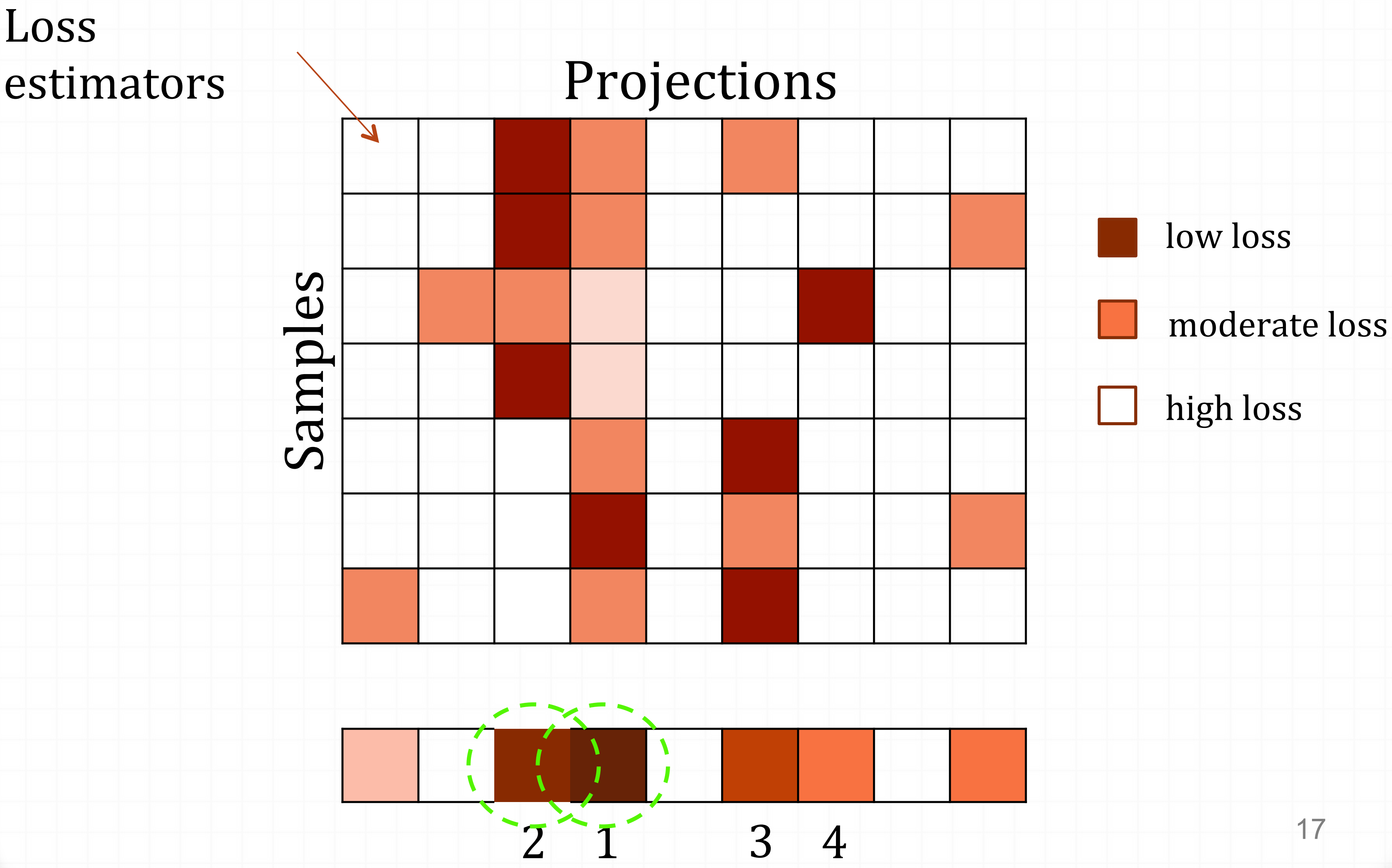
HIGH LOSS



LOW LOSS

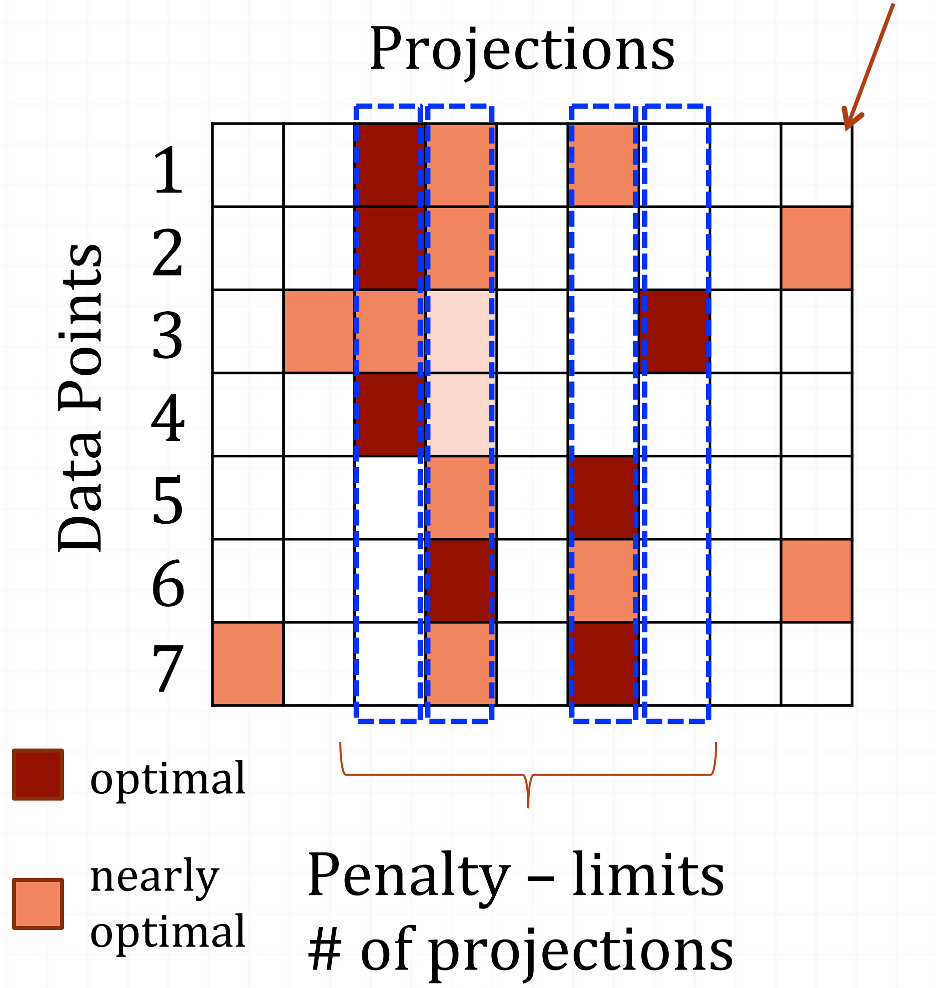


Starting point: the loss matrix



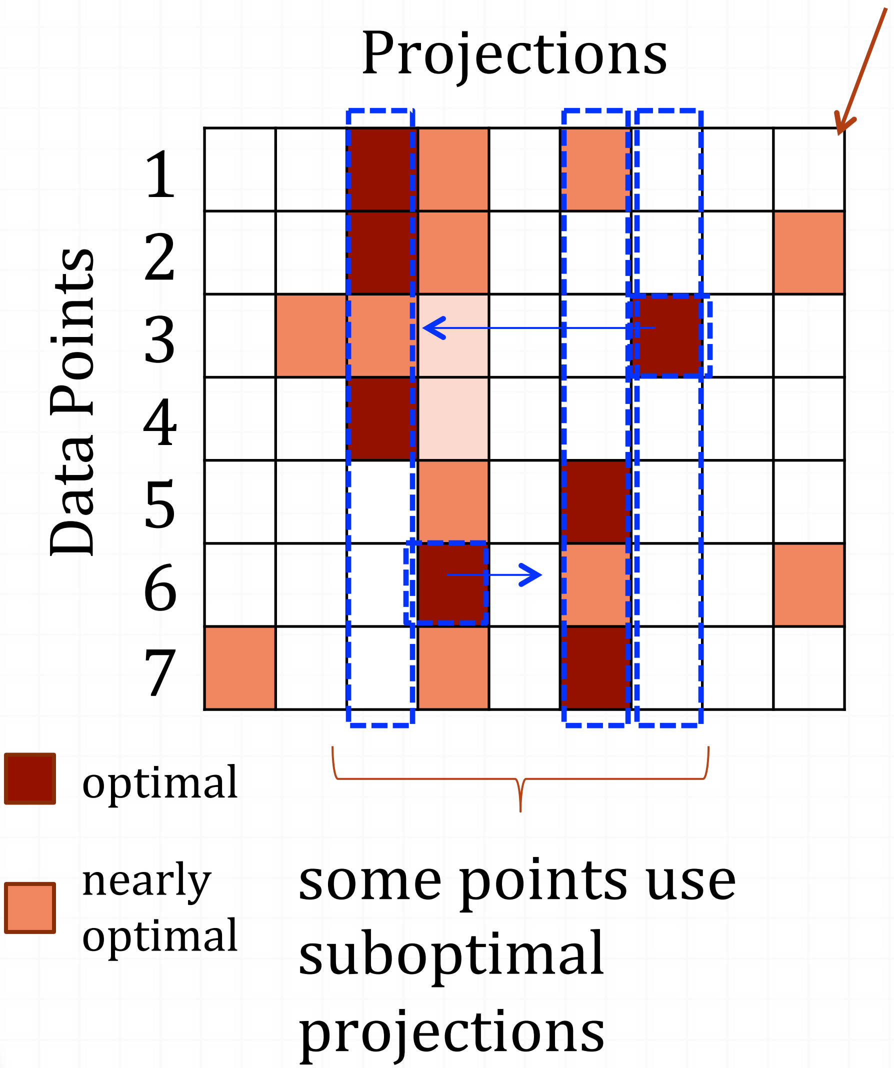
The Optimization Procedure

Matrix of Loss Estimators (L)

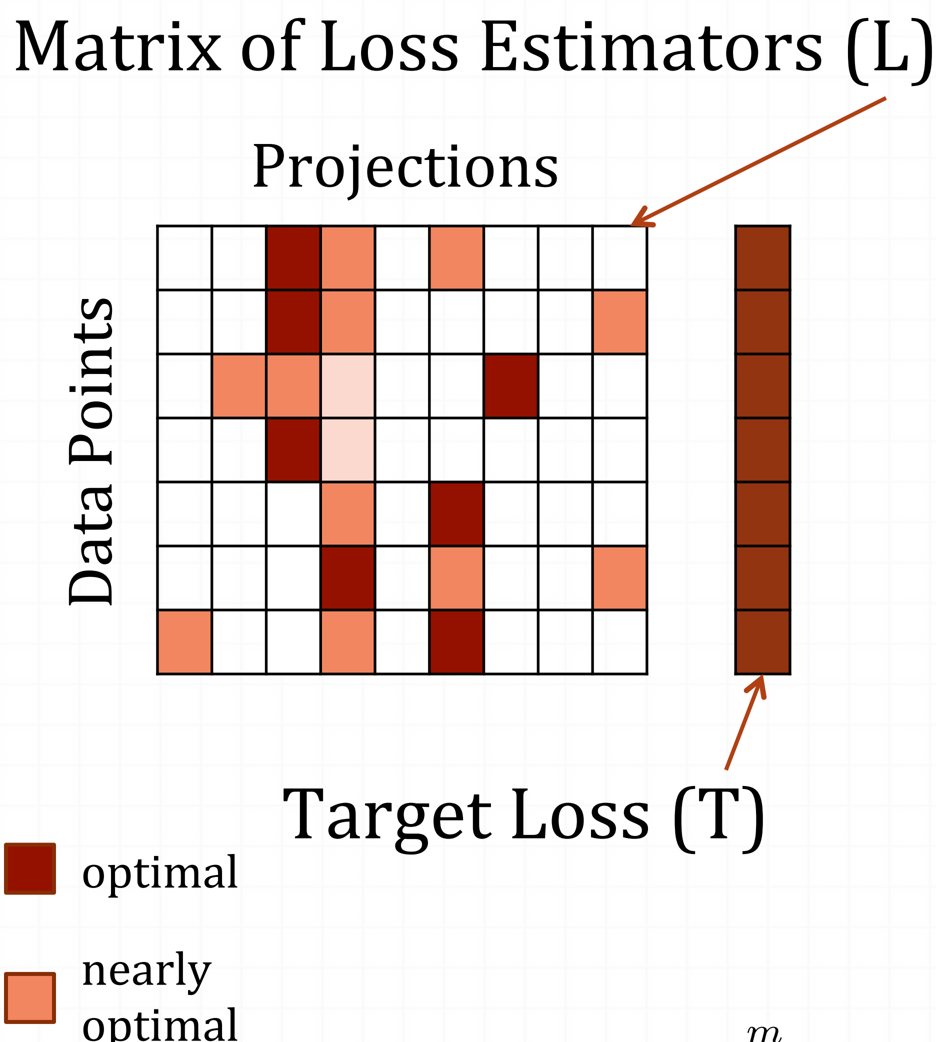


The Optimization Procedure

Matrix of Loss Estimators (L)



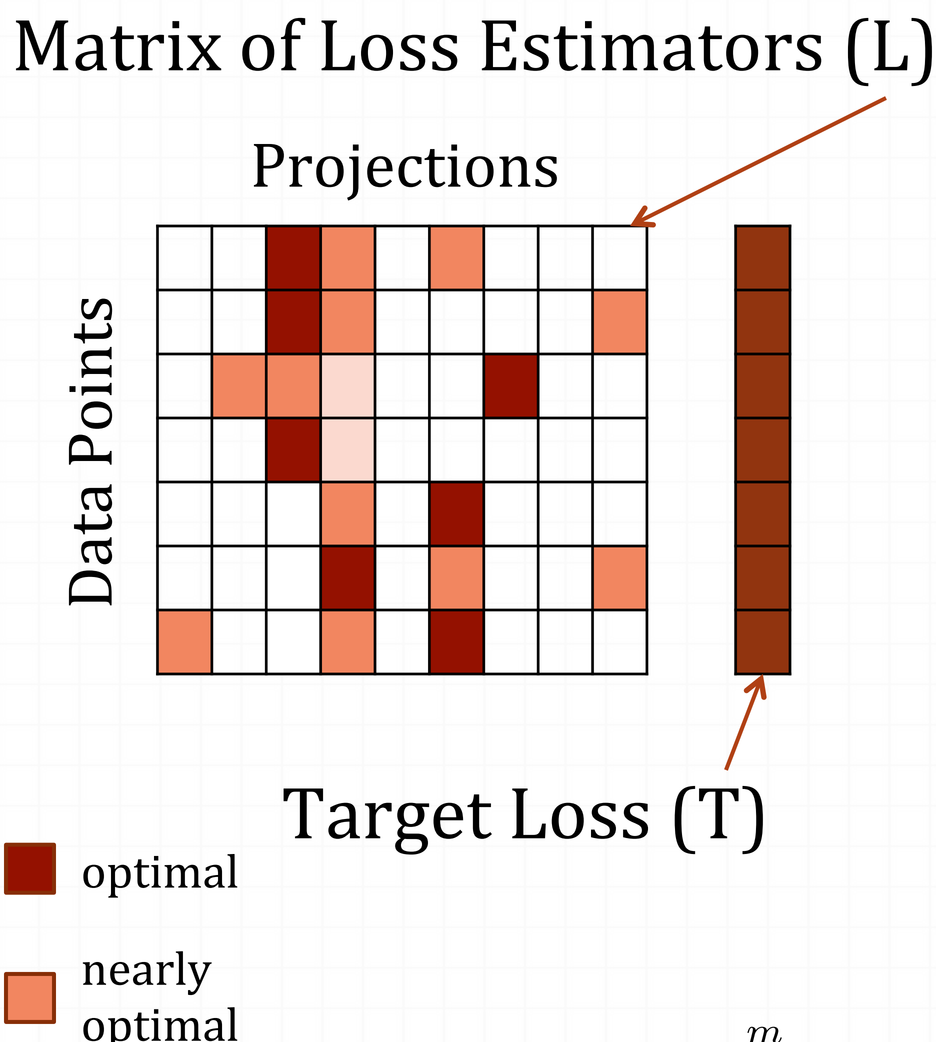
The Optimization Procedure



where $$L \odot B \stackrel{def}{=} \sum_{j=1}^{m} L_{.,j} B_{.,j}$$

- L_{ij} is the loss of sample i at projection j
- For each point i, let T_i be the lowest loss over the projections $T_i = \min L_{ij}$
- B binary selection matrix
- B_{ij} is 1 if projection j is to be used to solve point i and 0 otherwise
- $B = \min_B ||T - L \odot B||_1 +$ regularization (B)

The Optimization Procedure



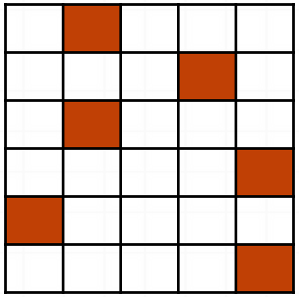
IPR problem - solved through this regression

- $B = \min_B ||T - L \odot B||_1$ + regularization (B)

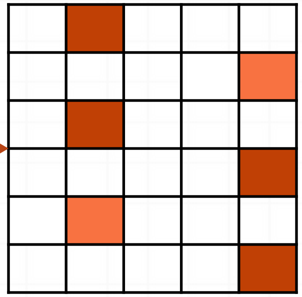
where $L \odot B \stackrel{def}{=} \sum_{j=1}^{m} L_{.,j} B_{.,j}$

Regression for Informative Projection Recovery (RIPR)

- RIPR learns the binary selection matrix B in a manner resembling the adaptive lasso



- Iterative procedure
  - Initialize selection matrix B
  - Compute multiplier δ inv. prop. with projection popularity
  - Use penalty $|B\delta|_1 \rightarrow$ new B



Applicability to Learning Tasks

RIPR can solve the following tasks[2]:

- Classification
- Semi-supervised classification
- Clustering
- Regression

Loss matrix computed differently for each task

Generality:

RIPR can solve any learning task for which the risk can be decomposed using consistent loss estimators.

[2] Madalina Fiterau and Artur Dubrawski. Informative projection recovery for classification, clustering and regression. In International Conference on Machine Learning and Applications, volume 12, 2013.

Loss Estimators: Classification

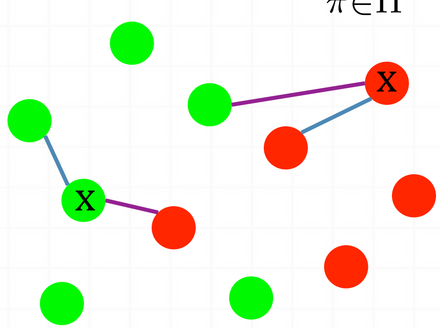
Neighbor-based estimator for conditional entropy*:

$$\hat{H}(Y|\pi(X); X \in \mathcal{A}(\pi)) \propto \frac{1}{n}\sum_{i=1}^{n} I[x_i \in \mathcal{A}(\pi)]\left(\frac{(n-1)dist_k(\pi(x_i), \pi(X_{y(x_i)}) \setminus \pi(x_i))^d}{n dist_k(\pi(x_i), \pi(X_{\neg y(x_i)} \setminus x_i))^d}\right)^{1-\alpha}$$

For a projection π, the estimator is $H(Y|\pi(X); g(X) \to \pi)$

The optimal model can be computed through the minimization:

$$M = \min_{M \in \mathcal{M}_d} \sum_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} I[g(x_i) \to \pi]\left(\frac{(n-1)\nu_k(\pi(x_i), \pi(X_{y(x_i)}) \setminus \pi(x_i))^d}{n\nu_k(\pi(x_i), \pi(X_{\neg y(x_i)} \setminus x_i))^d}\right)^{1-\alpha}$$



Selection matrix B_{ij}

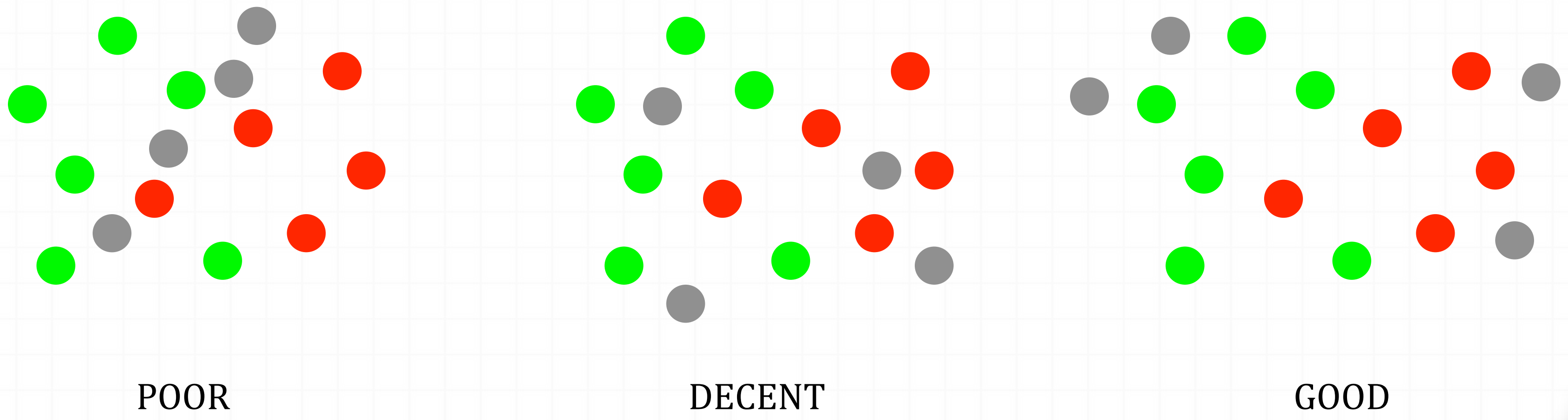
Loss matrix L_{ij}

Based on the divergence estimator by Poczos and Schneider, *"On the estimation of alpha-divergences"* (AISTATS 2011)

Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification
- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most ‘confident’ label (with smallest loss)

  Equivalent to
  - Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned



Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification
- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most 'confident' label (with smallest loss)

  Equivalent to

  - Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned

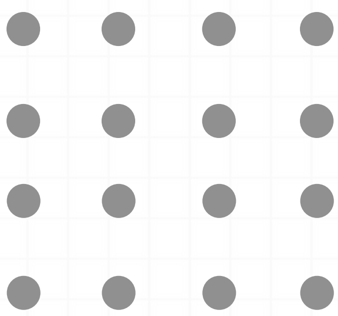
$$R_{ssc}(X, \tau_\pi^k) = \sum_{x \in X_+} \left(\frac{\nu_{k+1}(\pi(x), \pi(X_+))}{\nu_k(\pi(x), \pi(X_-))} \right)^{(1-\alpha)|\pi|} + \sum_{x \in X_-} \left(\frac{\nu_{k+1}(\pi(x), \pi(X_-))}{\nu_k(\pi(x), \pi(X_+))} \right)^{(1-\alpha)|\pi|} + \sum_{x \in X_u} \min \left(\frac{\nu_k(\pi(x), \pi(X_-))}{\nu_k(\pi(x), \pi(X_+))}, \frac{\nu_k(\pi(x), \pi(X_+))}{\nu_k(\pi(x), \pi(X_-))} \right)^{(1-\alpha)|\pi|}$$

Loss Estimators: Clustering

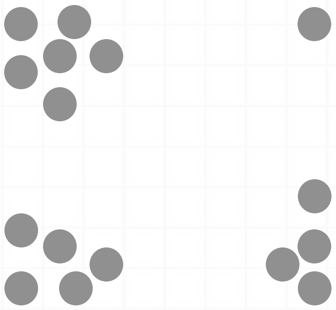
- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering

Loss Estimators: Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
- We focus on density-based clustering
- The loss is lower for densely packed regions
- We eliminate dimensionality issues by considering negative KL divergence from uniform on the same space



POOR



GOOD

Loss Estimators: Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
- We focus on density-based clustering
- The loss is lower for densely packed regions
- We eliminate dimensionality issues by considering negative KL divergence to uniform on the same space*

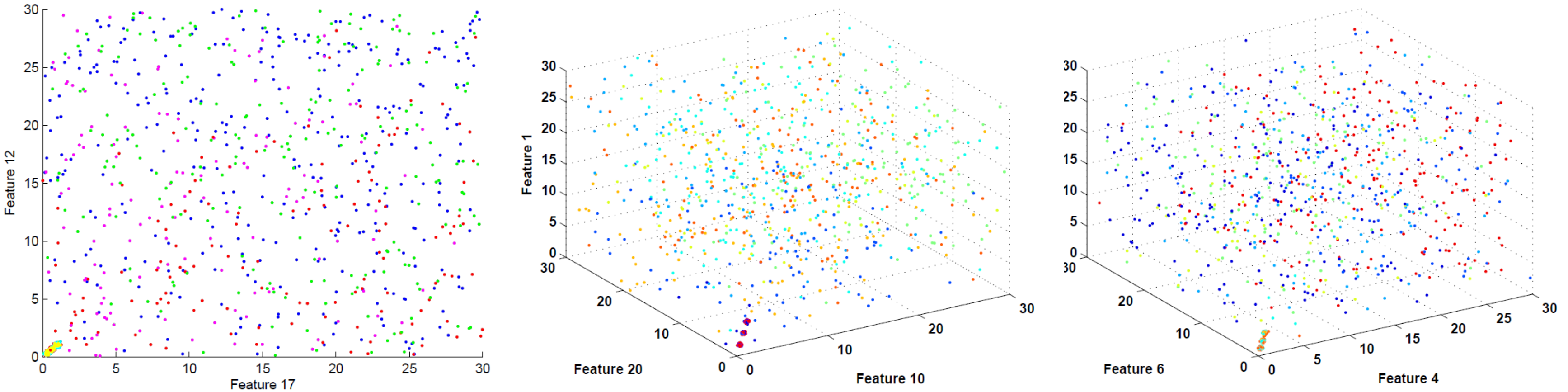
$$\hat{\mathcal{R}}_{clu}(\pi_i(x), \tau_i^{clu}) \rightarrow -KL(\pi_i(X) \| \pi_i(U))$$

$$\hat{\ell}_{clu}(\pi_i(x), \tau_i^{clu}) \approx \left(\frac{d(\pi_i(x), \pi_i(X))}{d(\pi_i(x), U)} \right)^{|\pi_i|(1-\alpha)}$$

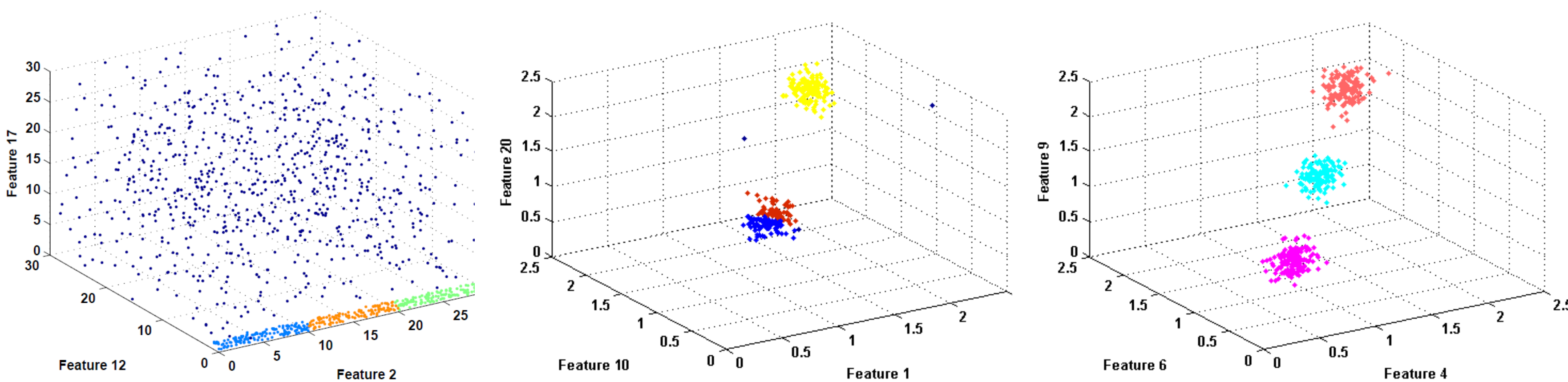
* some scaling issues remain

Low-d Clustering: Why it Works

K-Means model projected on (known) informative features



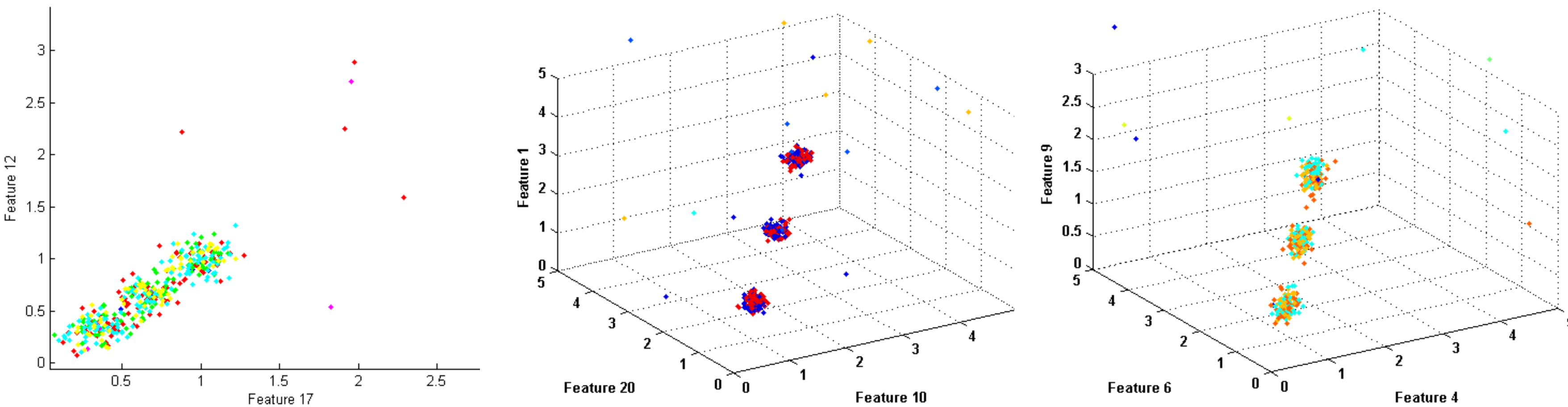
Representation of RIPR model – recovered projections and assigned data



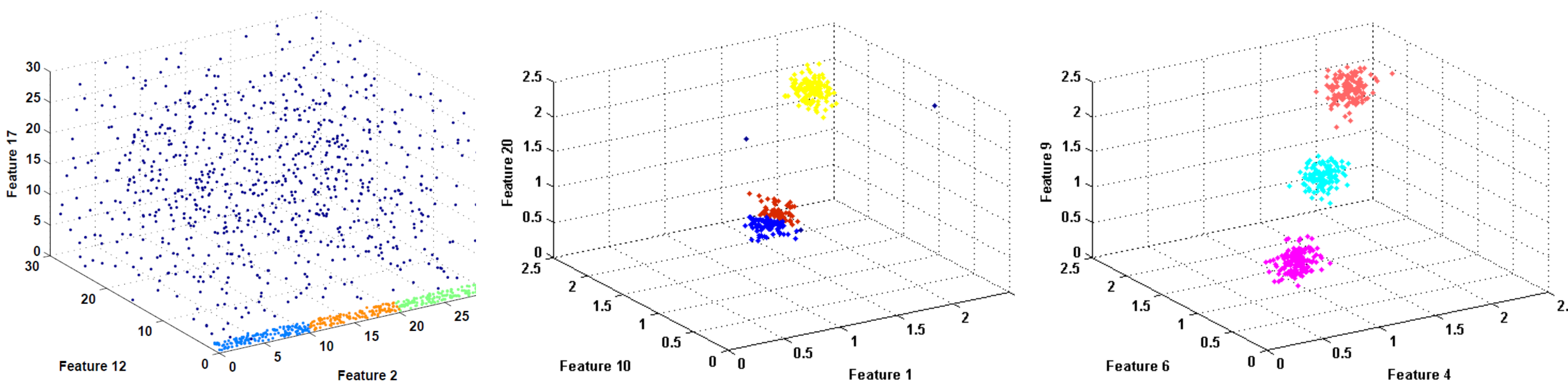
The hidden structure in data is clearly revealed by the RIPR model.

Low-d Clustering: Why it Works

K-Means model projected on (known) informative features



Representation of RIPR model – recovered projections and assigned data



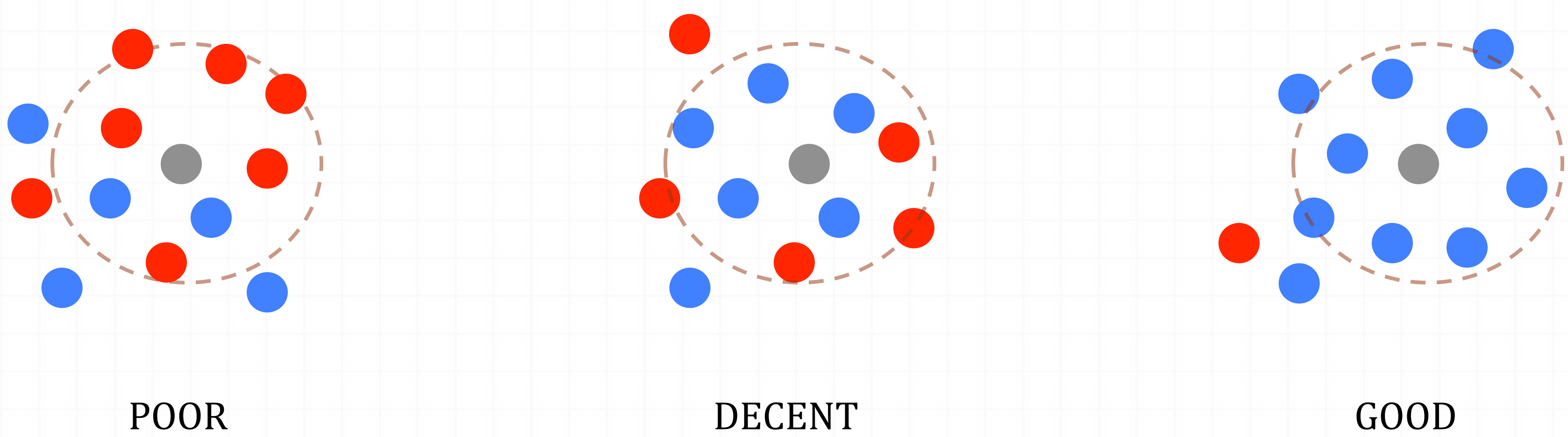
The hidden structure in data is clearly revealed by the RIPR model.

Loss Estimators: Regression

- Estimates error in point neighborhood

$$\hat{\ell}_{reg}(\pi_i(x), \tau_i(\pi_i(x))) = (\hat{\tau}(\pi_i(x)) - y)^2 \qquad \hat{\ell}_{reg} \to 0$$

$$\hat{\tau}_i(\pi_i(x)) = \frac{\sum_{i=1}^{k} w_{(i)} y_{(i)}}{\sum_{i=1}^{k} w_{(i)}}, \qquad \text{where } w_{(i)} = \frac{1}{||x - x_{(i)}||_2}$$



Loss/Risk for common Learning Tasks

Learning Task	Loss/Risk
Classification[1]	Classification error approximated by conditional entropy
Semi-supervised classification[2]	Conditional entropy for labeled samples plus best case entropy over label assignments for unlabeled samples
Clustering[2]	Negative divergence between distribution of data and a uniform distribution on the same sample space
Regression	Mean squared error

[1] Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In Advances in Neural Information Processing Systems 25 (NIPS), pages 3032–3040, 2012.

[2] Madalina Fiterau and Artur Dubrawski. Informative projection recovery for classification, clustering and regression. In International Conference on Machine Learning and Applications, volume 12, 2013.

Assigning a projection to a query

- Problem: how to select the appropriate projection for a specific query x?
- Solution: select the projection in P for which the estimated loss is the lowest
  $$(k*, y*) = argmin_{(k \in \{1 \ldots |P|\}, y \in \mathcal{Y})} \ell(\tau_k(\pi_k(x), y)$$
- For classification, the selection function and label are
  $$g^k(x) := argmin_{(\pi,\tau) \in (\Pi^k, T^k)} \hat{h}(\tau(\pi(x))|\pi(x))$$
  $$\hat{y}(x) := \tau^k_{g^k(x)}(x)$$
- For clustering, the loss estimator is computed considering the cluster assignments determined during learning

Active Sampling Approach[4]

- At iteration k, samples X_ℓ^k are labeled as Y_ℓ^k
- Samples X_u^k are unlabeled
- The RIPR model built so far is $M^k = \{\Pi^k, T^k, g^k\}$
- The expected error of the model is

$$Err(M^k) = \mathbb{E}_{x \in \mathcal{X}}[I(\tau^k_{g^k(x)}(\pi^k_{g^k(x)}(x)) \neq y)]$$

- Key issue: find the appropriate scoring function $s : \mathcal{M} \times \mathcal{X} \rightarrow \mathbb{R}$
- Next sample to be labeled $x^{k+1} = argmax_{x \in X_n^k} s(M^k, x)$
- We use the notation M_s^k to refer to a model obtained after k iterations using scoring function s
- Given maximum acceptable error ϵ and a set $\mathcal{S}$ of scoring functions, the optimal selection strategy can be expressed as

$$s^* = argmin_{s \in \mathcal{S}} \min_k \{k \text{ s.t. } Err(M_s^k) \leq \epsilon\}$$

- The algorithm starts with r_0 randomly selected samples
- The stopping criterion is based on error on a hold-out set

[4]Fiterau M, Dubrawski A, Chen L, Hravnak M, Clermont G, Bose E, Guillame-Bert M, Pinsky MR. Artifact adjudication for vital sign step-down unit data can be improved using Active Learning with low-dimensional models. Intensive Care Medicine. 2014.

Active Sampling Strategies

Let $\hat{h}$ be the conditional entropy estimator for a label given a subset of the features and $\hat{y}(x)$ the prediction made for a sample x.

Sample selection: $x^{k+1} = argmax_{x \in X_n^k} s(M^k, x)$

Sampling Type	Formula for RIPR model
Uncertainty	$s_{uncrt}(x) = \min_{\pi \in \Pi_{uncrt}^k, \tau \in T_{uncrt}^k} \hat{h}(\tau(\pi(x))\|\pi(x))$
Query by Committee	$s_{qbc}(x) = \max_{\tau_i, \tau_j \in T_{qbc}^k} I(\tau_i(\pi_i(x)) \neq \tau_j(\pi_j(x)))$
Information Gain	$s_{ig}(x) = \hat{H}_{X_\ell, Y_\ell}^k(X_{u,ig}^k)$ $- p(y=0)\hat{H}_{X_\ell \cup \{x\}, Y_\ell \cup \{0\}}^k(X_{u,ig}^k)$ $- p(y=1)\hat{H}_{X_\ell \cup \{x\}, Y_\ell \cup \{1\}}^k(X_{u,ig}^k), \quad \forall x \in X_{u,ig}^k$
Low Conditional Entropy	$s_{mc}(x) = 1 - \min_{\pi \in \Pi_{mc}^k, \tau \in T_{mc}^k} \hat{h}(\tau(\pi(x))\|\pi(x))$

RIPR Results

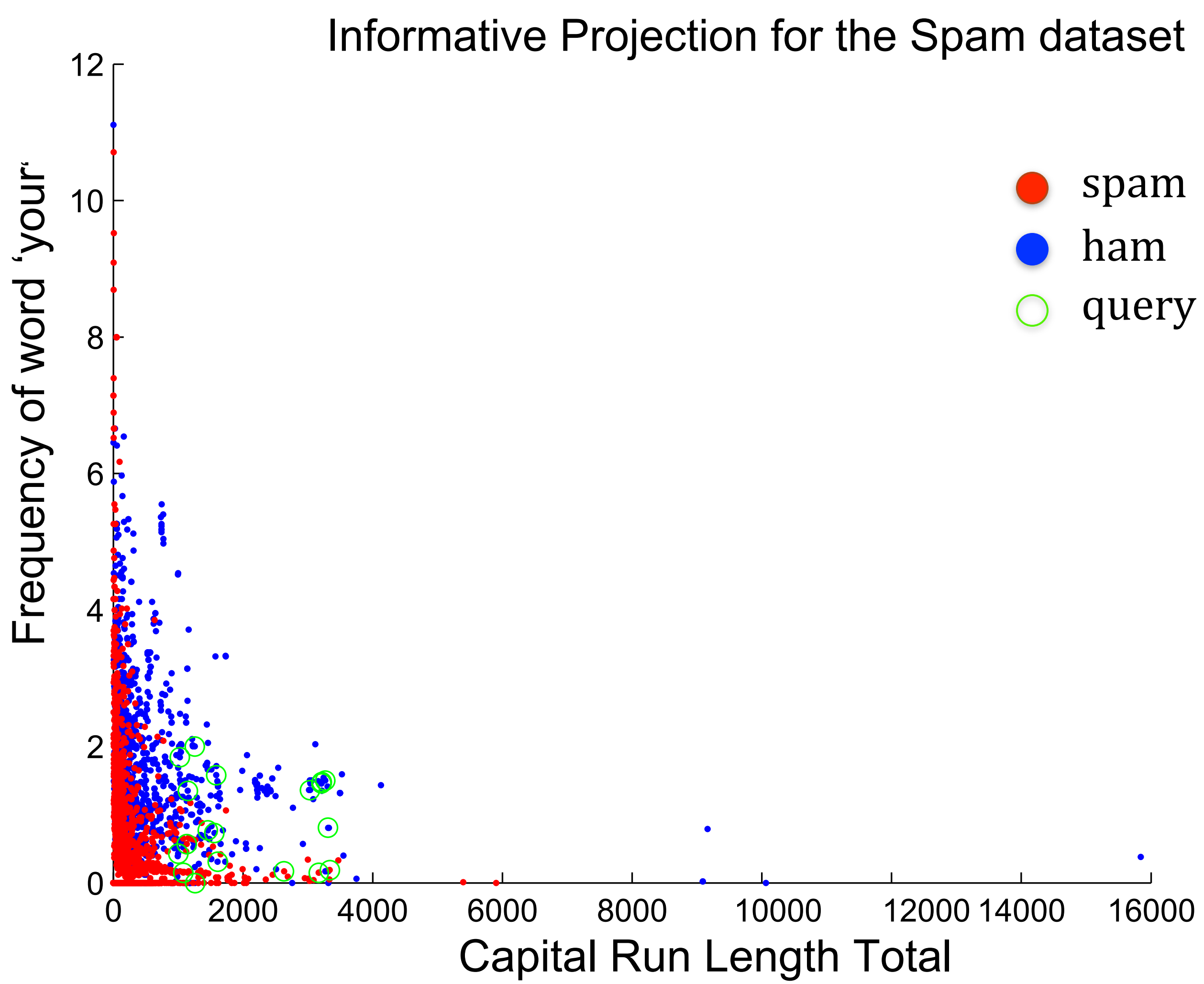
Classification

Classification
- UCI data -

Comparison of Classification Accuracy						
Dataset	# Features	# Instances	K-NN	RIPPED K-NN	# RIPR projections	#features in projection
Breast Tissue	10	106	1.000	1.000	1	2
Cell	6	200	0.707	**0.7640**	4	{1,2,2,2}
Mini BOONE	50	130065	**0.790**	0.740	1	1
Nuclear Threat	50	200	0.7788	0.7807	3	2
SPAM	57	4601	0.7680	0.7680	5	{1,2,3,3,3}
Vowel	10	528	0.984	0.984	1	10

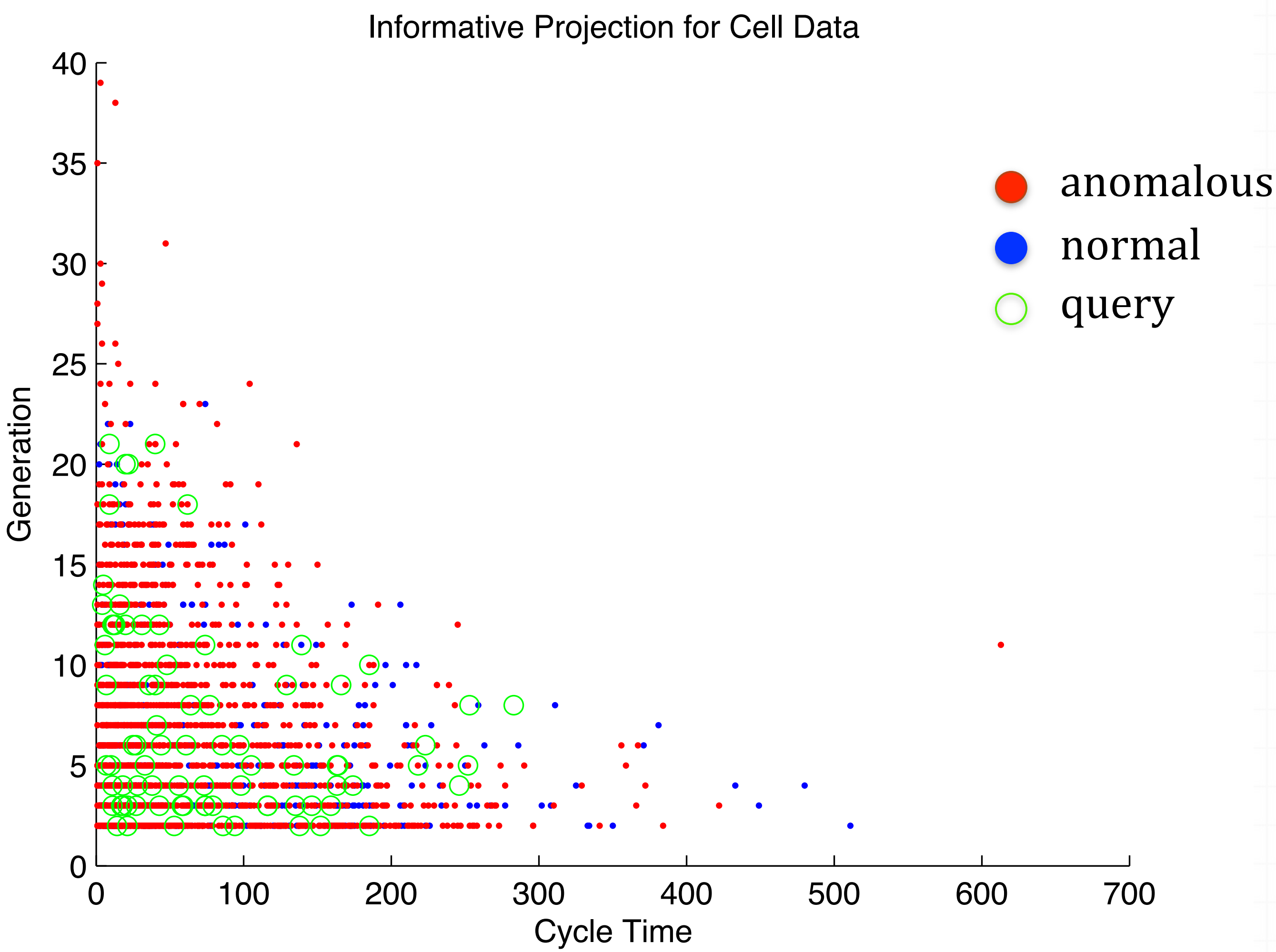
Classification
- Informative Projections -

The main advantage is the low-dimensional representation that RIPR provides.

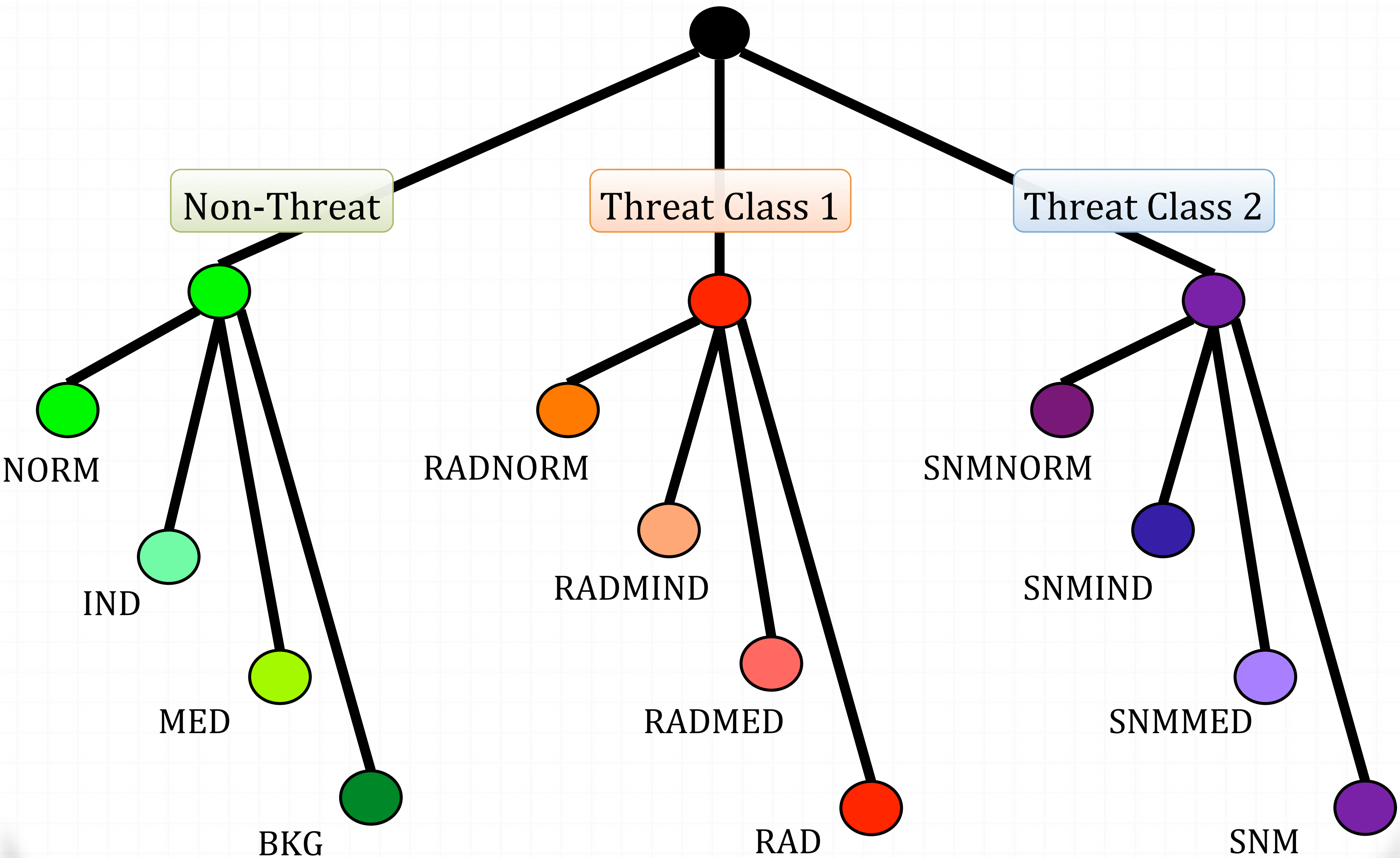


Classification
- Informative Projections -

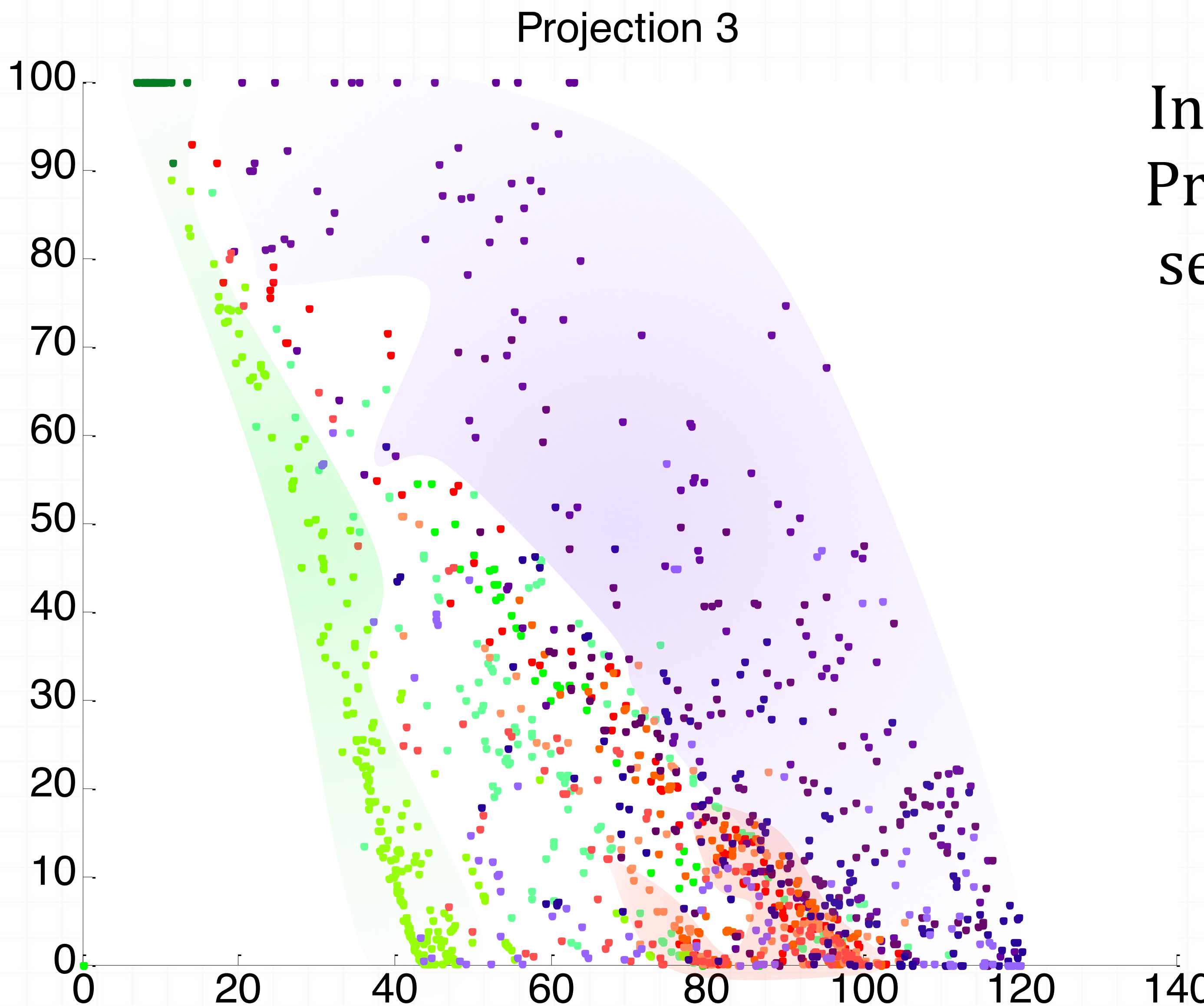
The main advantage is the low-dimensional representation that RIPR provides.



Nuclear Threat Detection

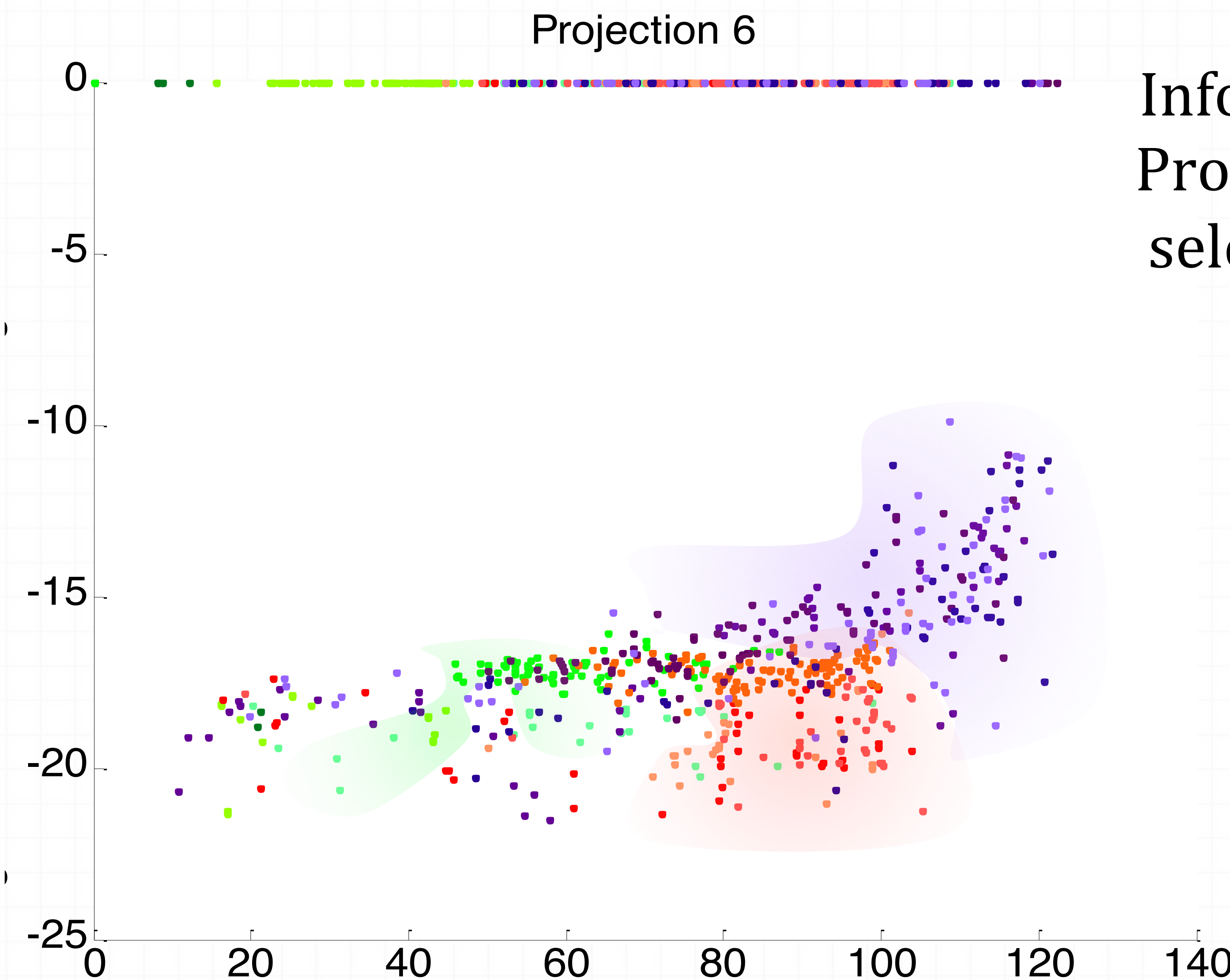


Nuclear Threat Detection



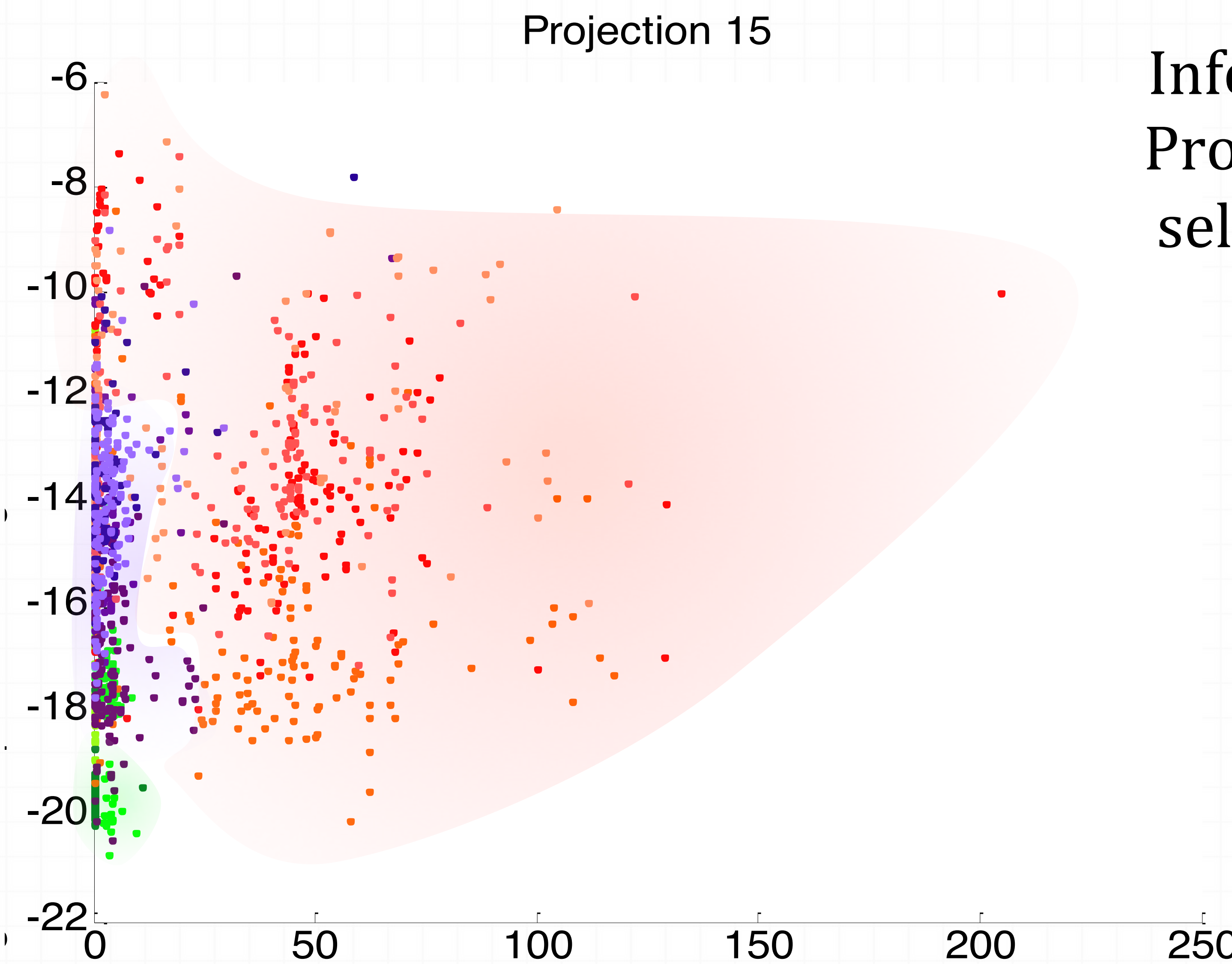
Informative Projection 1 selected by RIPR

Nuclear Threat Detection



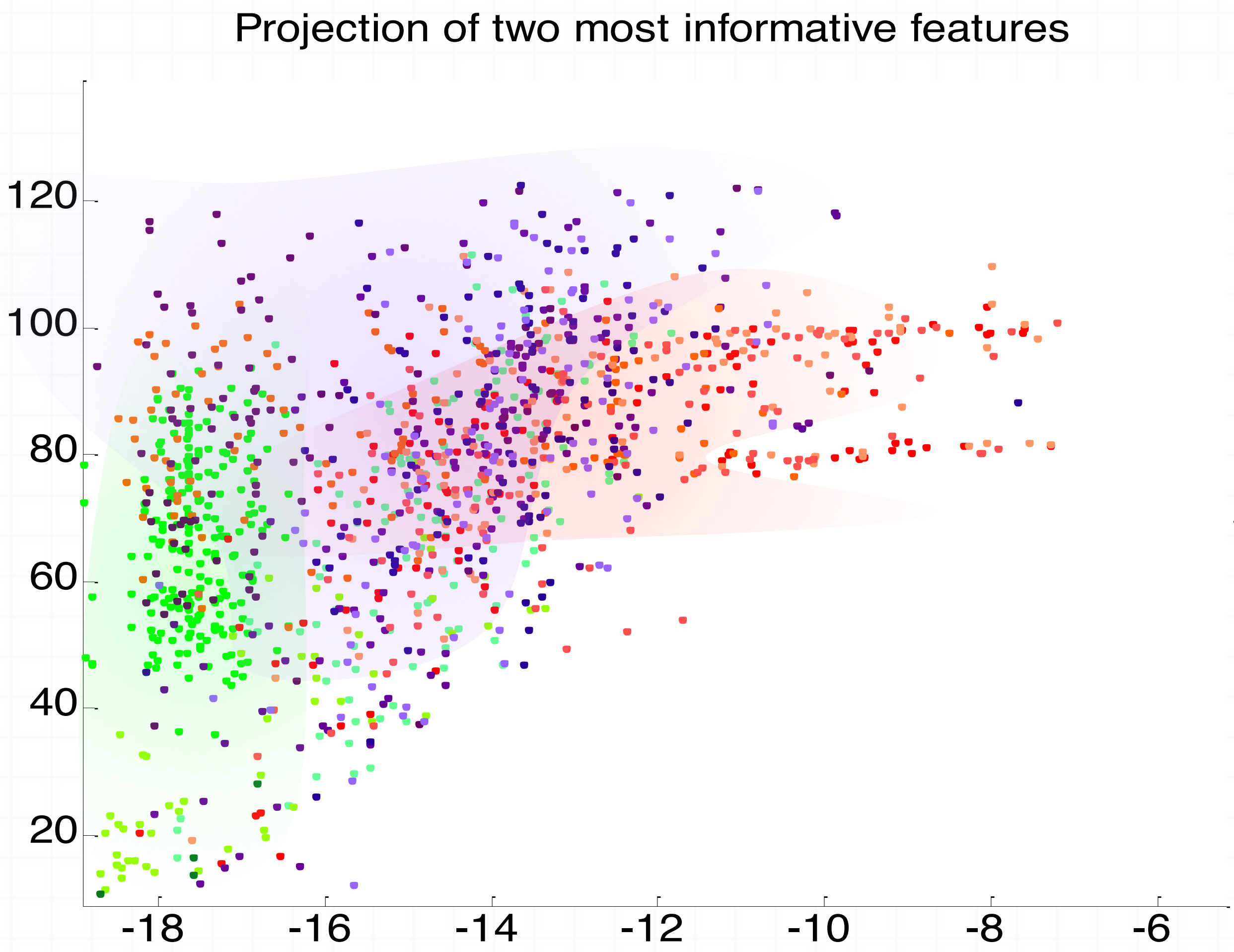
Informative Projection 2 selected by RIPR

Nuclear Threat Detection



Informative Projection 3 selected by RIPR

Nuclear Threat Detection



An informative projection that domain experts would use.

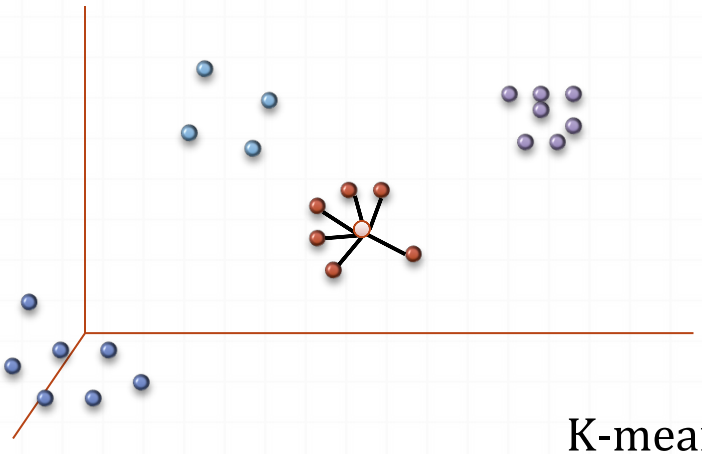
RIPR Results

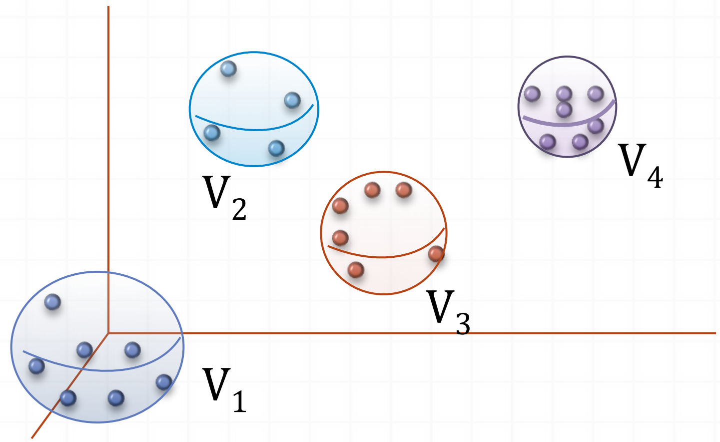
Clustering

Clustering
- evaluation metrics -

DISTORTION – mean distance to cluster centers

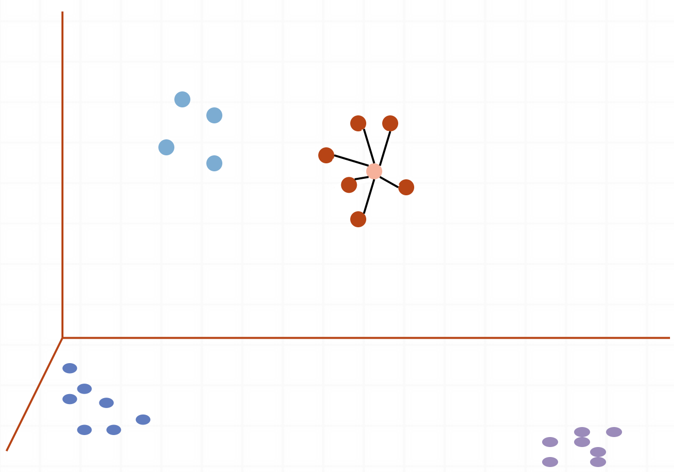
LOG CLUSTER VOLUME

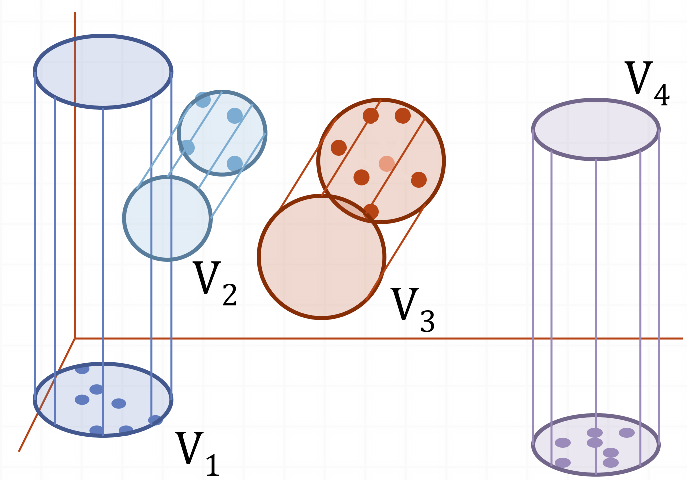




K-means Model

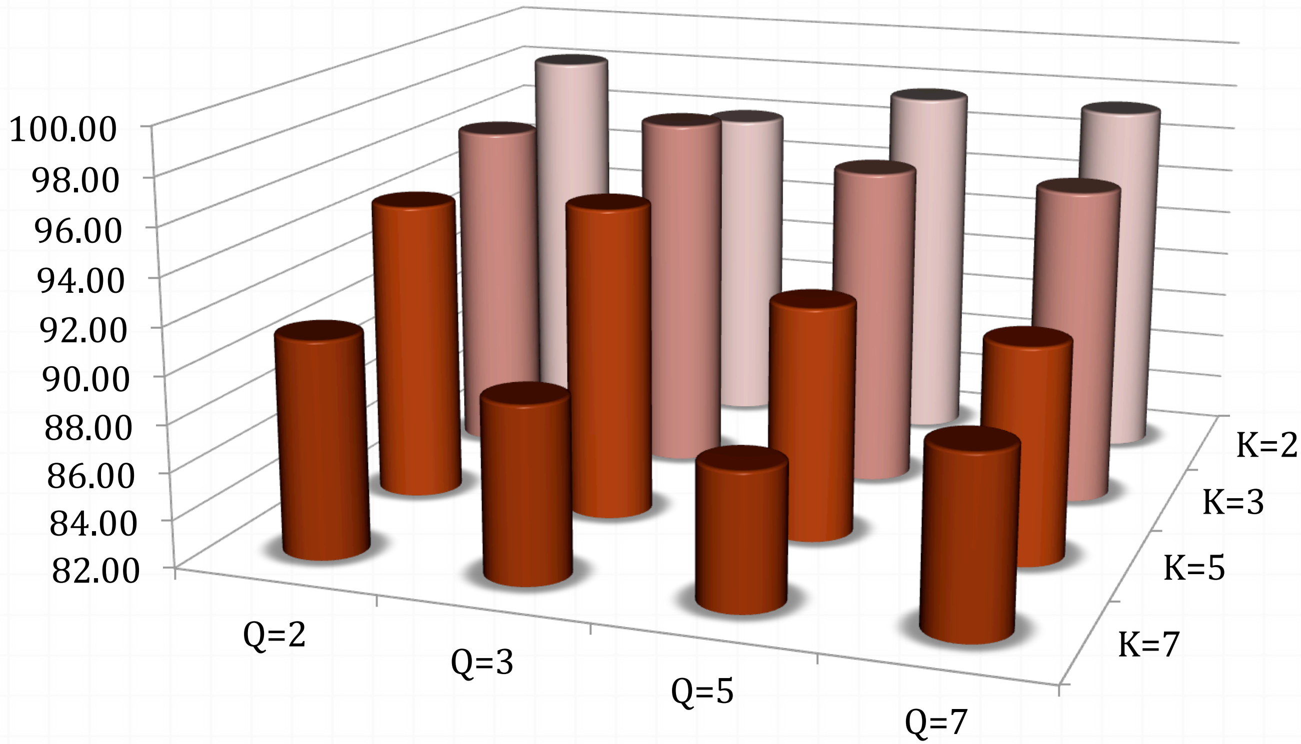
Ripped K-means Model





Clustering - artificial data -

PERCENTAGE REDUCTION IN SUM OF CLUSTER LOG VOLUMES



Q = NUMBER OF INFORMATIVE PROJECTIONS
K = NUMBER OF CLUSTERS ON EACH PROJECTION

COMPRESSION IS REDUCED AS MORE CLUSTERS/PROJECTIONS ARE ADDED

NOTE: THE K-MEANS AND RIPR MODELS HAVE THE NUMBER OF CLUSTERS.

Clustering
- UCI data -

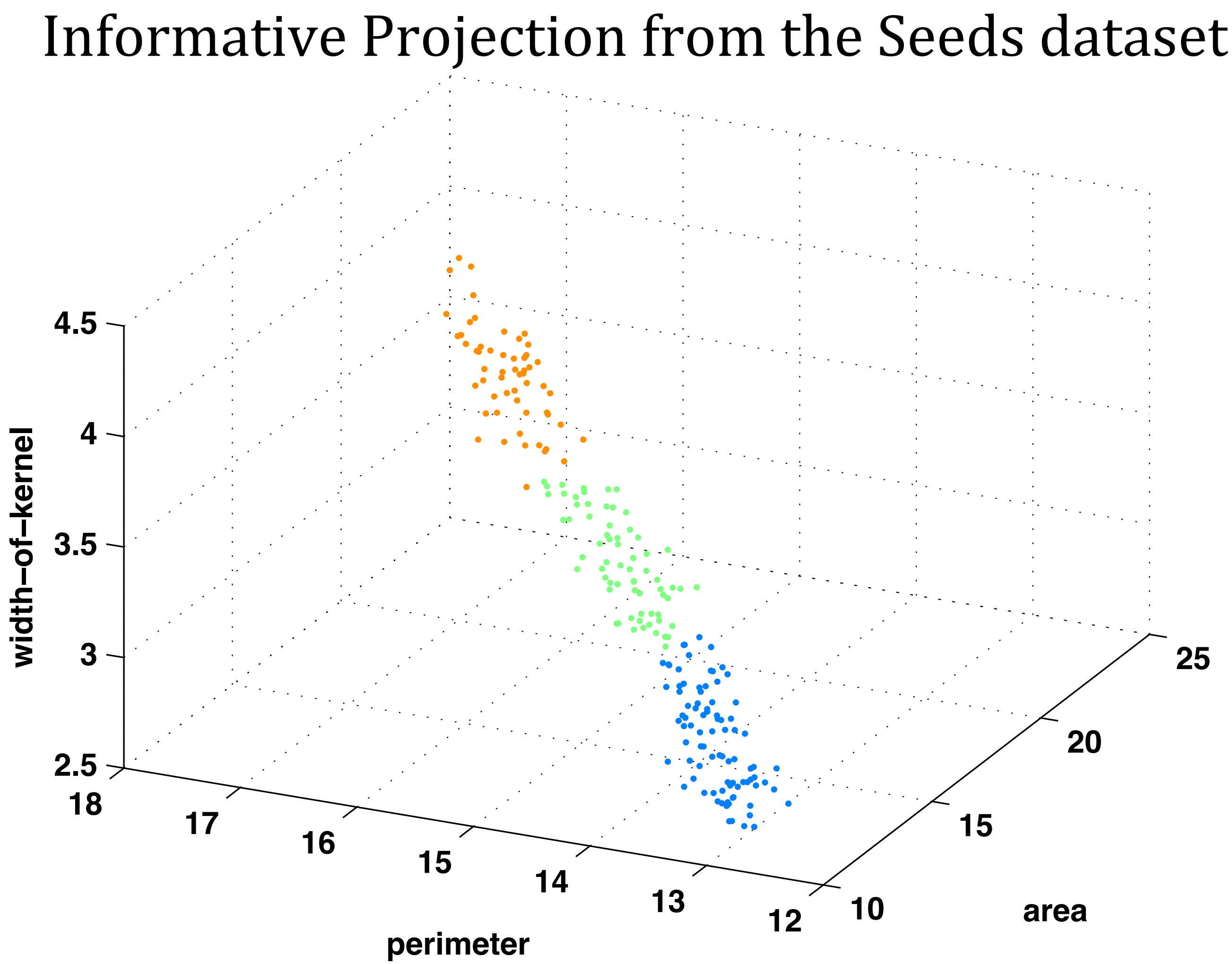
SUM OF MEAN DISTANCES TO CLUSTER CENTERS AND LOG CLUSTER VOLUME

UCI Dataset	Mean Distortion		% Distortion Reduction	Log Volume of Clusters on All Dimensions		% Volume Reduction
	RIPR	Kmeans		RIPR	Kmeans	
Seeds	**16**	107	90.73	**3.33**	4.21	86.83
Libras	**9**	265	98.54	**-2.52**	3.15	99.00
MiniBOONE	**125**	1,154,704	99.99	**104.23**	107.77	99.97
Cell	**40,877**	8,181,327	99.78	**23.75**	29.39	99.00
Concrete	**1,370**	55,594	98.01	**21.39**	22.91	97.01

LOWER IS BETTER. RIPR MODELS ALWAYS HAVE A SMALLER TOTAL VOLUME.

Clustering
- UCI data -

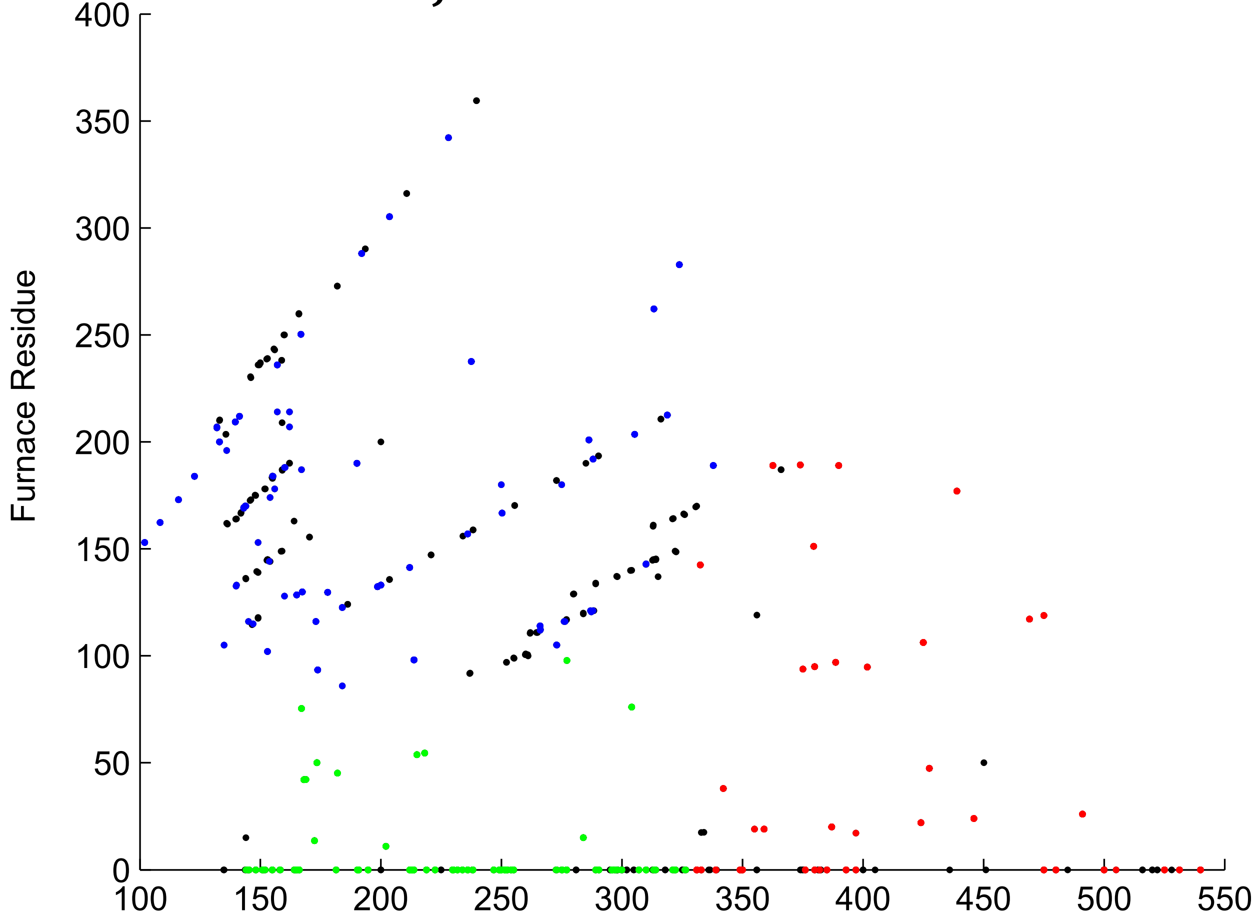
The main advantage is the low-dimensional representation that RIPR provides.



Clustering
- UCI data -

The main advantage is the low-dimensional representation that RIPR provides.

Informative Projection from the Concrete dataset



RIPR Results

Regression

Regression
- artificial data -

ACCURACY OF RIPPED SVM COMPARED TO ACCURACY OF STANDARD SVM
- THE NUMBER OF INFORMATIVE PROJECTIONS : 2-10 (OUT OF 45)
- PERCENTAGE OF NOISY SAMPLES: 0-50% (OUT OF 1600)

	IP #	2	3	5	7	10		2	3	5	7	10
		MSE RIPPED-SVM					MSE SVM					
NOISY SAMPLES	0%	**0.05**	**0.27**	**0.05**	**0.02**	**0.23**		0.27	1.16	0.11	0.1	0.43
	6.25%	**0.42**	1.26	**0.34**	**1.45**	**0.52**		0.8	**1.02**	0.6	2.99	0.94
	12.5%	**0.5**	**0.86**	0.8	**0.33**	**0.99**		0.97	1.27	**0.29**	0.68	1.44
	25%	0.63	1.47	1.34	**1.61**	0.11		**0.4**	**1.26**	**1.64**	1.71	**0.08**
	50%	0.69	0.38	1.12	0.68	**1.1**		**0.52**	**0.06**	**0.91**	**0.9**	1.16

Thesis Outline

Informative Projection Retrieval

- Projection Retrieval as a combinatorial problem
- Optimization procedure for IPR
- Customizing RIPR for classification, clustering, regression
- Projection Discovery in an Active Learning setting

Applying RIPR to Clinical Alert Classification

- Building interpretable classification models for clinical alerts
- Annotation Framework using Active RIPR

Proposed research

- IPR for multi-task learning and time series
- Low-dimensional model learning for feature hierarchies
- Online cost-constrained subset selection policies

Case Study – Alert Classification[3] - importance of artifact adjudication -

- Step-down Unit vital sign monitoring system
- Alerts are raised when patient health status deteriorates
- One alert is issued every 90s

- A significant amount of alerts are artifacts
- Frequent alerts cause alarm fatigue in medical staff
- 812 labeled samples, each associated with a vital sign
- Extracted temporal features and derived metrics

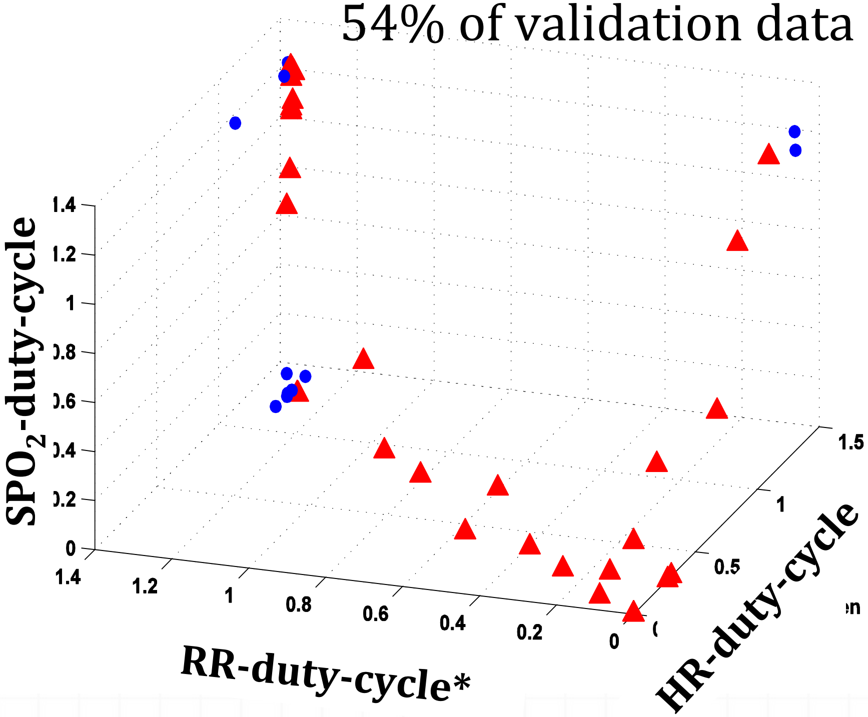
- RIPR provides interpretable artifact adjudication models

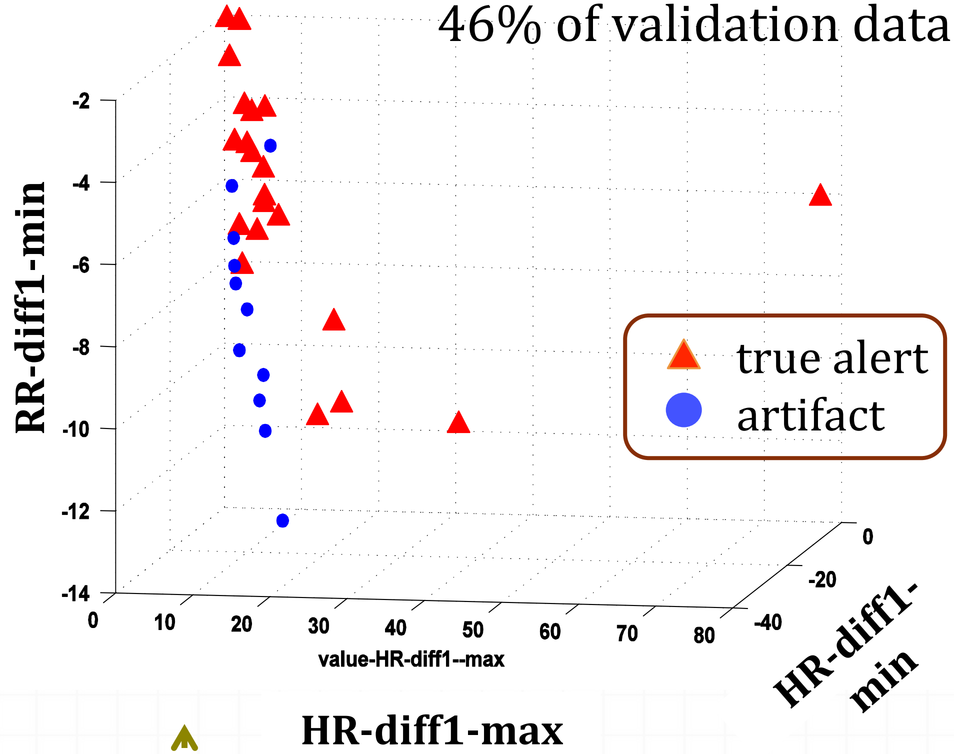
[3] Fiterau M, Dubrawski A, Chen L, Hravnak M, Clermont G, Pinsky MR. Automatic identification of artifacts in monitoring critically ill patients. Intensive Care Medicine. 2013; 39 (Suppl 2]: S470.

Case Study – Alert Classification - performance -

Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

Case Study – Alert Classification
- RIPR model for blood pressure -



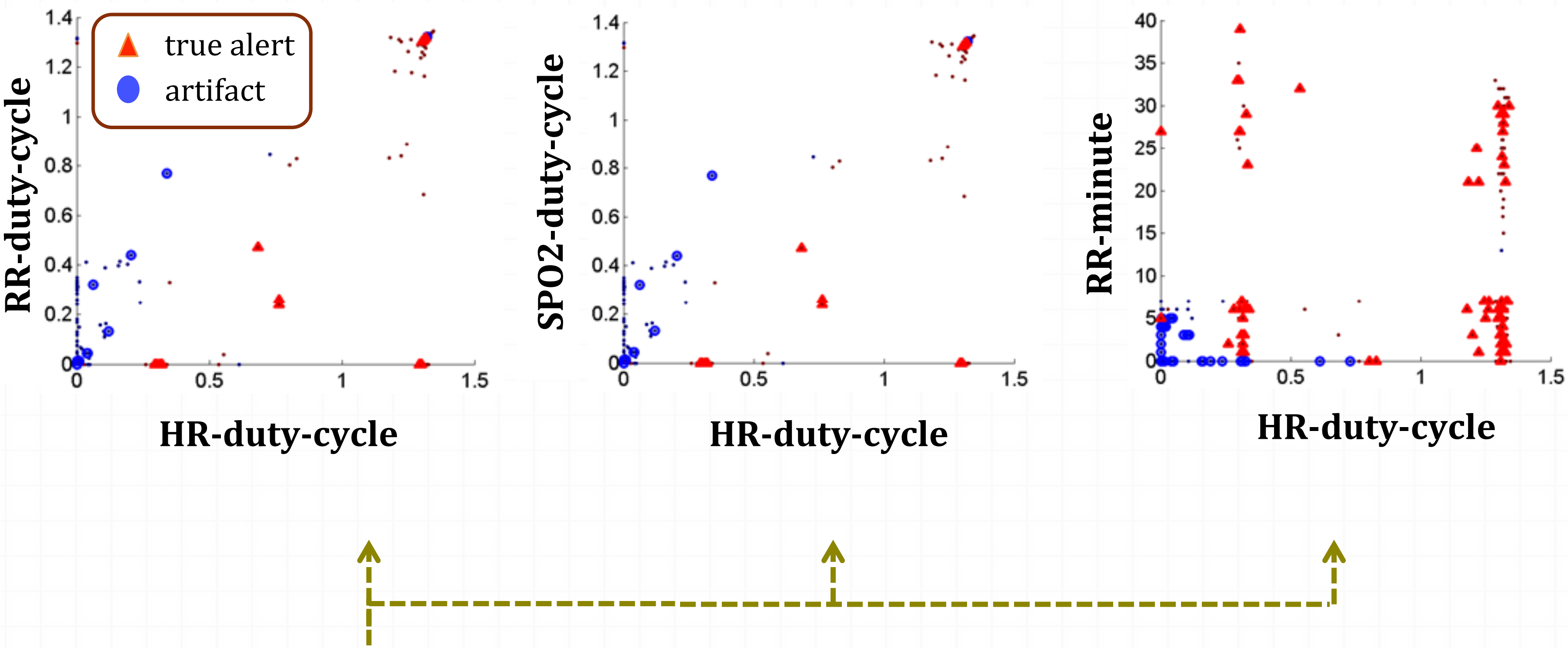


Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

RIPR identifies interpretable projections which adjudicate alerts.

*duty cycle = number of readings over time units: a low value indicates high sparseness

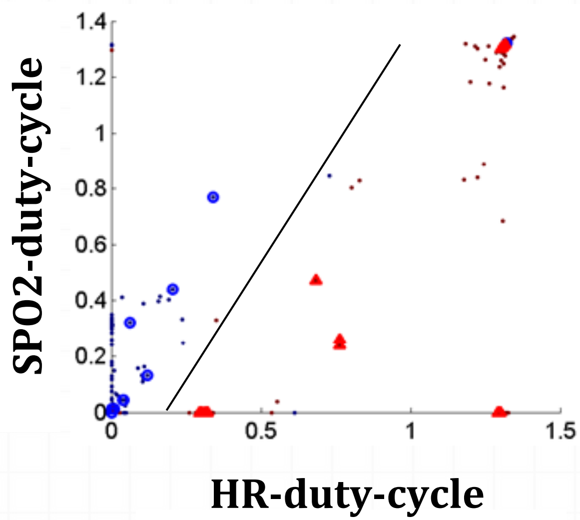
Case Study – Alert Classification - deriving rules -

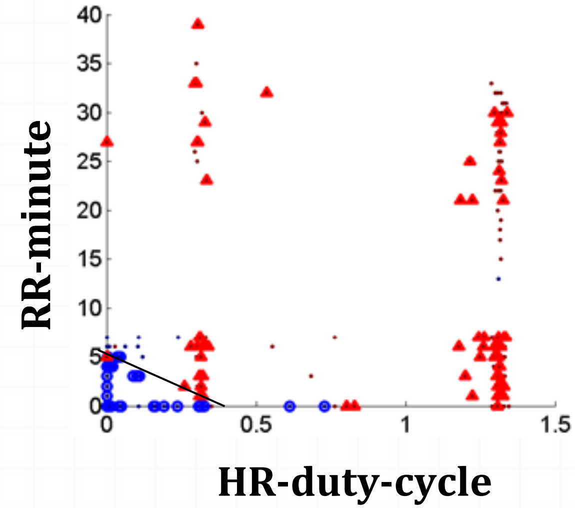


Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

*duty cycle = number of readings over time units: a low value indicates high sparseness

Case Study – Alert Classification - deriving rules -





RR-duty-cycle* <= 0.6
and
HR-duty-cycle <= 0.25

HR-duty-cycle –
SPO_2-duty-cycle <= 0.2

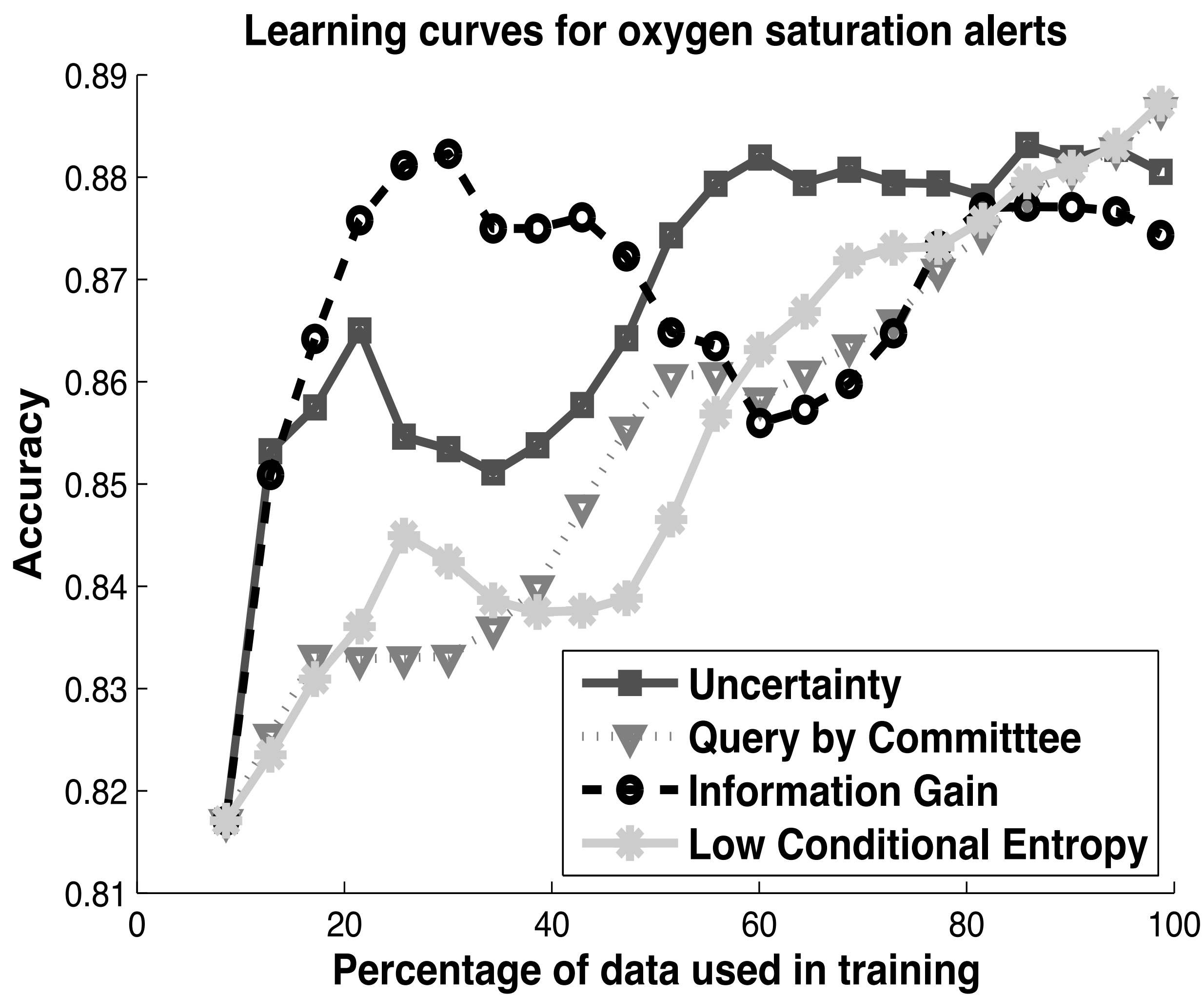
HR-duty-cycle/0.3
+ RR-min/5 <= 1

Decreasing expert annotation effort[6]

- Only ~10% of the data is currently labeled
- Initial set could be different from the rest
- Clinicians will need to annotate some of the remaining samples
- Annotation objectives:
  - Provide informative projections
  - Minimize expert effort
  - Maintain high classification accuracy
- We use *ActiveRIPR:*
  - Projections available during annotations
  - Samples selected based on current RIPR models

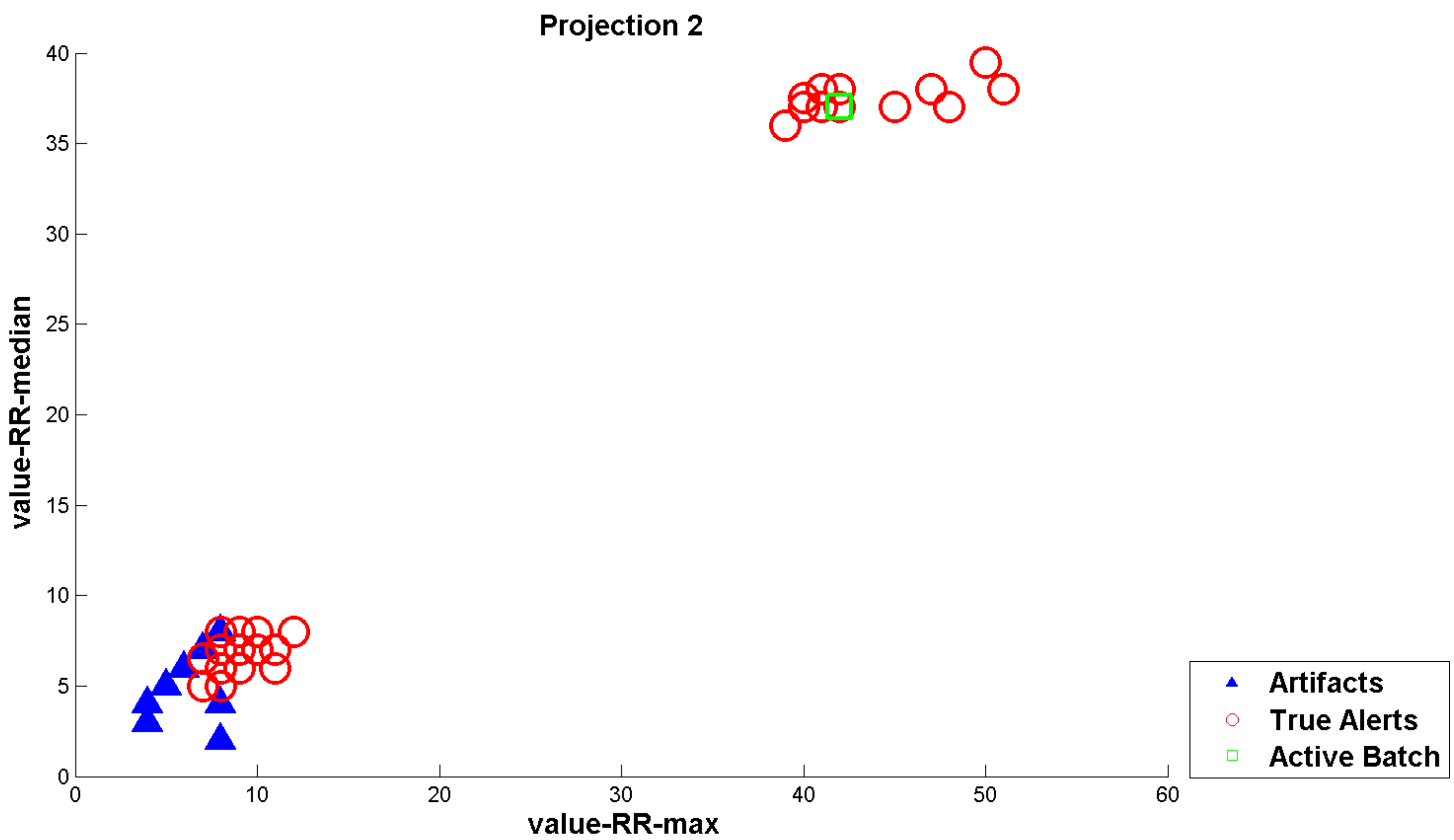
[6] Wang D, Fiterau M, Dubrawski A, Hravnak M, Clermont G, Pinsky MR. Interpretable active learning in support of clinical data annotation. SSCM 2015

Adjudication of oxygen saturation alerts



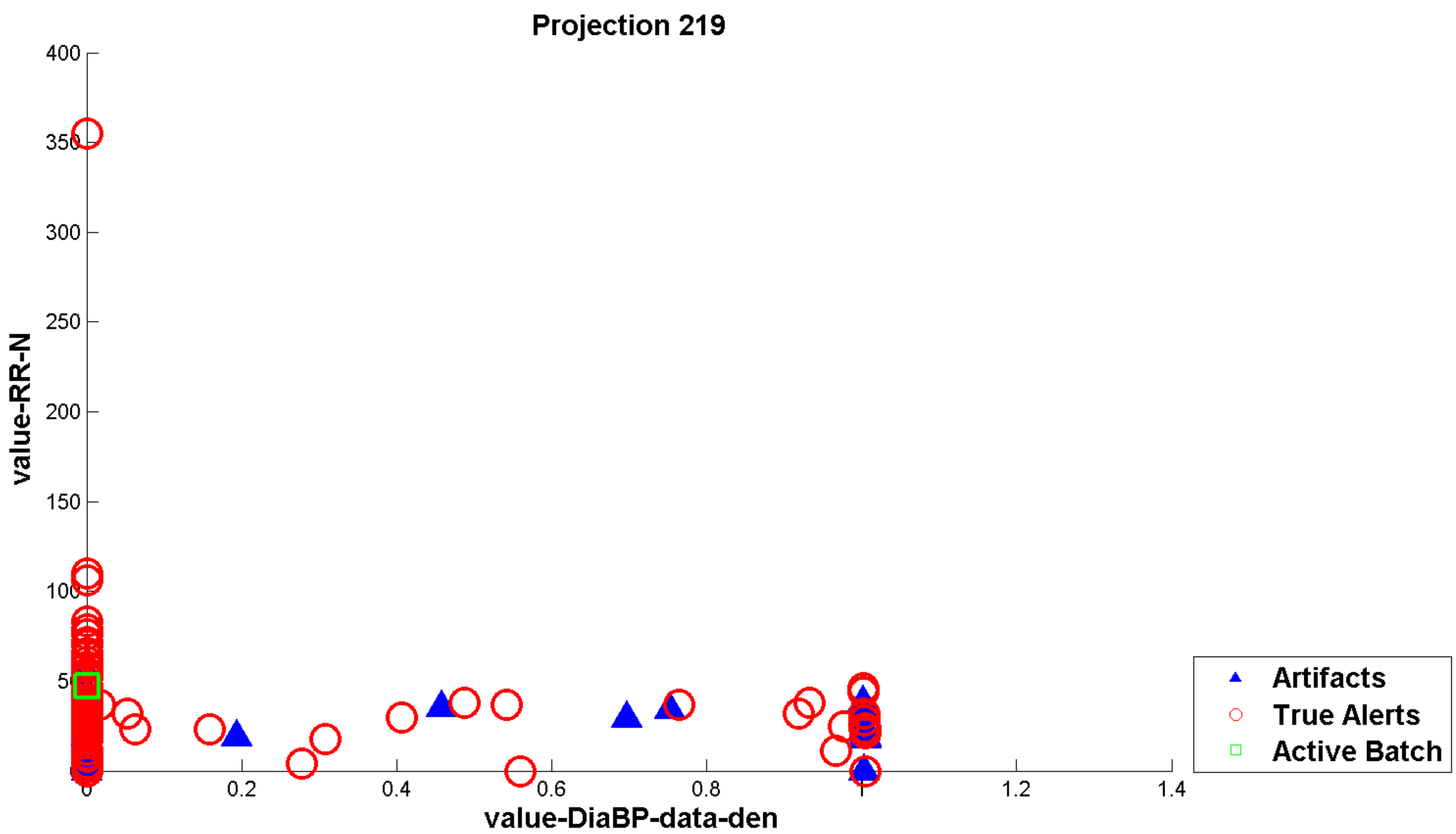
We performed 10-fold cross-validation, training the ActiveRIPR model on 90% of the samples and using the remainder to calculate the learning curve.

Projections assisting annotation (RR)



The retrieved few low-dimensional projections make it possible for domain experts to quickly adjudicate alert labels.

Projections assisting annotation (SPO_2)



The retrieved few low-dimensional projections make it possible for domain experts to quickly adjudicate alert labels.

Contribution Summary

- Informative Projection Retrieval is relevant to many applications requiring interaction with human users
- We generalized RIPR, our solution to the IPR problem, to a wide range of learning tasks (classification, regression, clustering)
- RIPR expresses loss though divergence estimators
  - Semi-supervised models: penalize unlabeled data that cannot be confidently assigned to a class
  - Clustering models: favor high data density
- RIPR models are compact and well-performing in practice
- Overall, RIPR provides an intuitive solution problem of classifying alerts issues by clinical monitoring systems

Alert data issues worth considering

- Feature cost (invasiveness, computational cost)
- Means of deriving the features
  (feature hierarchies)
- Determining alert subcategories
- Timestamp information
- Online execution

Thesis Outline

Informative Projection Retrieval

- Projection Retrieval as a combinatorial problem
- Optimization procedure for IPR
- Customizing RIPR for classification, clustering, regression
- Projection Discovery in an Active Learning setting

Applying RIPR to Clinical Alert Classification

- Building interpretable classification models for clinical alerts
- Annotation Framework using Active RIPR

Proposed research

- IPR for multi-task learning and time series
- Low-dimensional model learning for feature hierarchies
- Online cost-constrained subset selection policies

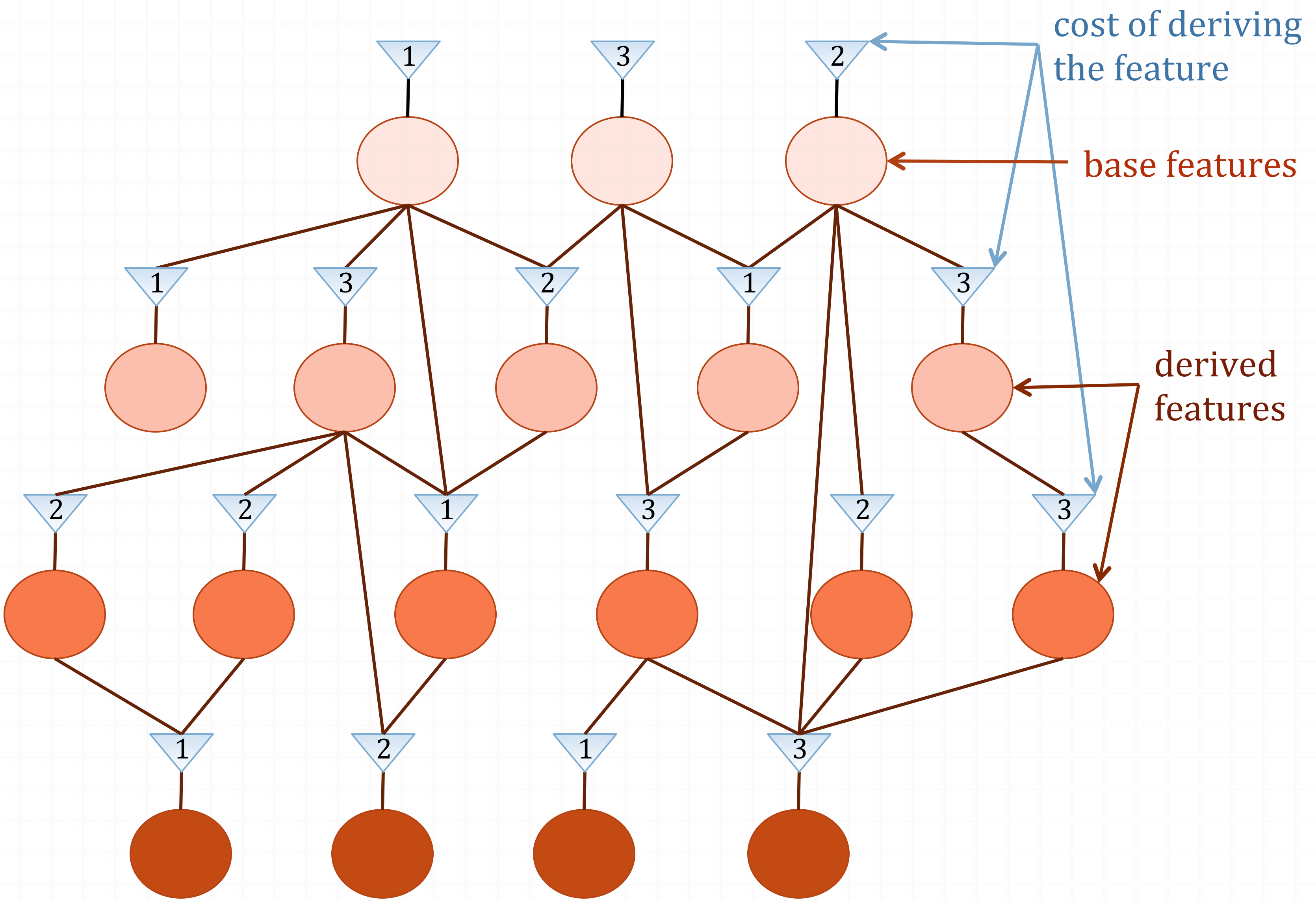
IPR for Multitask Learning

- Generalize of RIPR to multitask learning
  - Multiple types of nuclear threats
  - Sub-categories of clinical alerts
- Not only are we grouping features/samples, but also features/samples/tasks
- The loss matrix becomes a loss tensor
- Assignment procedure is an optimization, with the appropriate constraints, over the loss tensor.
- Modify RIPR to perform multi-model low-d CCA
- Outcome: set of canonical parameter pairs.

IPR for Time Series

- Extend the concept of projections to time series data
- Learn time-varying models
- Impose smoothness constraints over parameters at consecutive timestamps (fused lasso)
- Ensemble coherence constraints needed across samples, to ensure use of a small number of projections
- Transition constraints which will prevent the model switching to become too sample-specific
- Trends in the data, as well as the actual feature values, will have to be considered.
- A usage example is instability prediction due to blood loss under the assumption that the mode of response to a health crisis is patient-dependent

Feature hierarchies



Feature hierarchy example

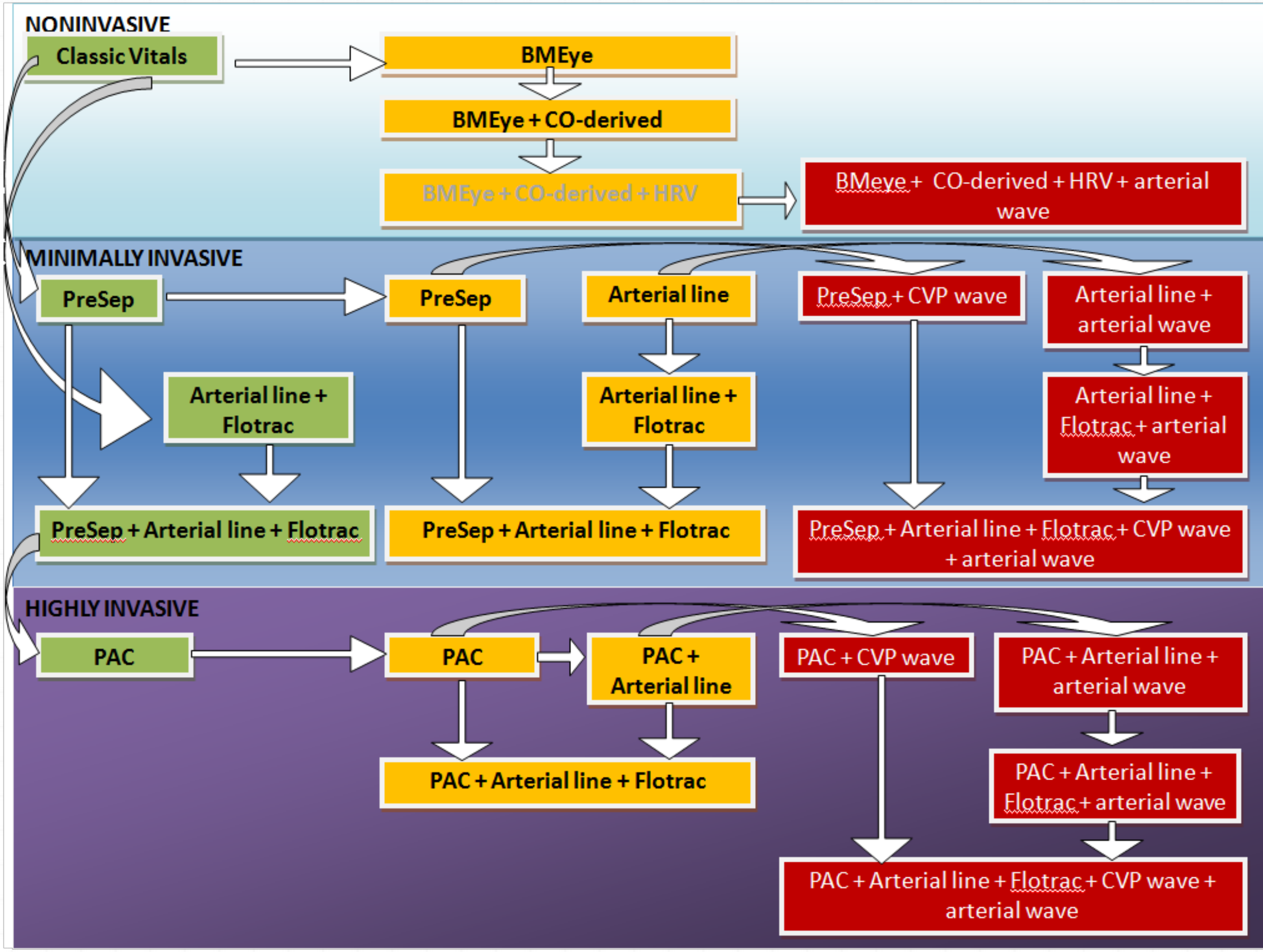


Image and corresponding data courtesy of Andre Holder and Mathieu Guillaume-Bert

Penalty for feature dependency

- Feature set $A = \{a_1 \ldots a_m\}$
- Cost function $c : 2^A \to \mathbb{R}$
- Feature dependencies: directed graph (A, D)
- $(a_i, a_j) \in D$ ⇔ feature j depends on feature i
- Weight learning involves the minimization

$$w^* = argmin_w \sum_{i=1}^{n} f(w, x_i, y_i) + g(w)$$

penalty function according to cost

- Weighted lasso typically used

$$g_{\ell_1}(w) = \sum_{i=1}^{m} c(a_i)|w_i|$$

- Does not account for cost already expended for parent features in the hierarchy

Penalty for feature dependency

- Feature set $A = \{a_1 \ldots a_m\}$
- Cost function $c : 2^A \rightarrow \mathbb{R}$
- Feature dependencies: directed graph (A, D)
- $(a_i, a_j) \in D$ $\Leftrightarrow$ feature j depends on feature i
- Weight learning involves the minimization

$$w^* = argmin_w \sum_{i=1}^{n} f(w, x_i, y_i) + g(w)$$

penalty function according to cost

- We link each feature to its children through ℓ_2 norms
- Index set of children of a_i is $\phi(a_i) = \{1 \leq j \leq m | (a_i, a_j) \in D\}$
- Penalty $g_{c,D}(w) = \sum_{i=1}^{m} c(a_i) ||w_{i,\phi(i)}||_2$
- encourages parent weight to be 0 only when all weights of children are 0
- Equal to ℓ_1 norm for features without children

Penalty for feature redundancy

- Feature redundancy is present in some cases
- Examples: vital signal readings obtained through procedures with different levels of invasiveness
- Only one feature in such a group is needed at a time
- 'OR' constraint distributes weight across the features
- Assume a_i can be obtained from either of $a_i^1 \ldots a_i^r$

$$g_{OR}(w_i) = c(a_i)||w_{i,\phi(i)}||_2 + \sum_{j=1}^{r} \sum_{k \neq j}^{r} c(a_i^j)||\bar{w}_i^j, w_i^k||_2$$

- where w_i decomposes as $\sum_{j=1}^{r} w_i^j = w_i$

and

$$\bar{w}_i^j = \max\left(\frac{1}{w_i^j + 0.5} - 0.5, 0\right)$$

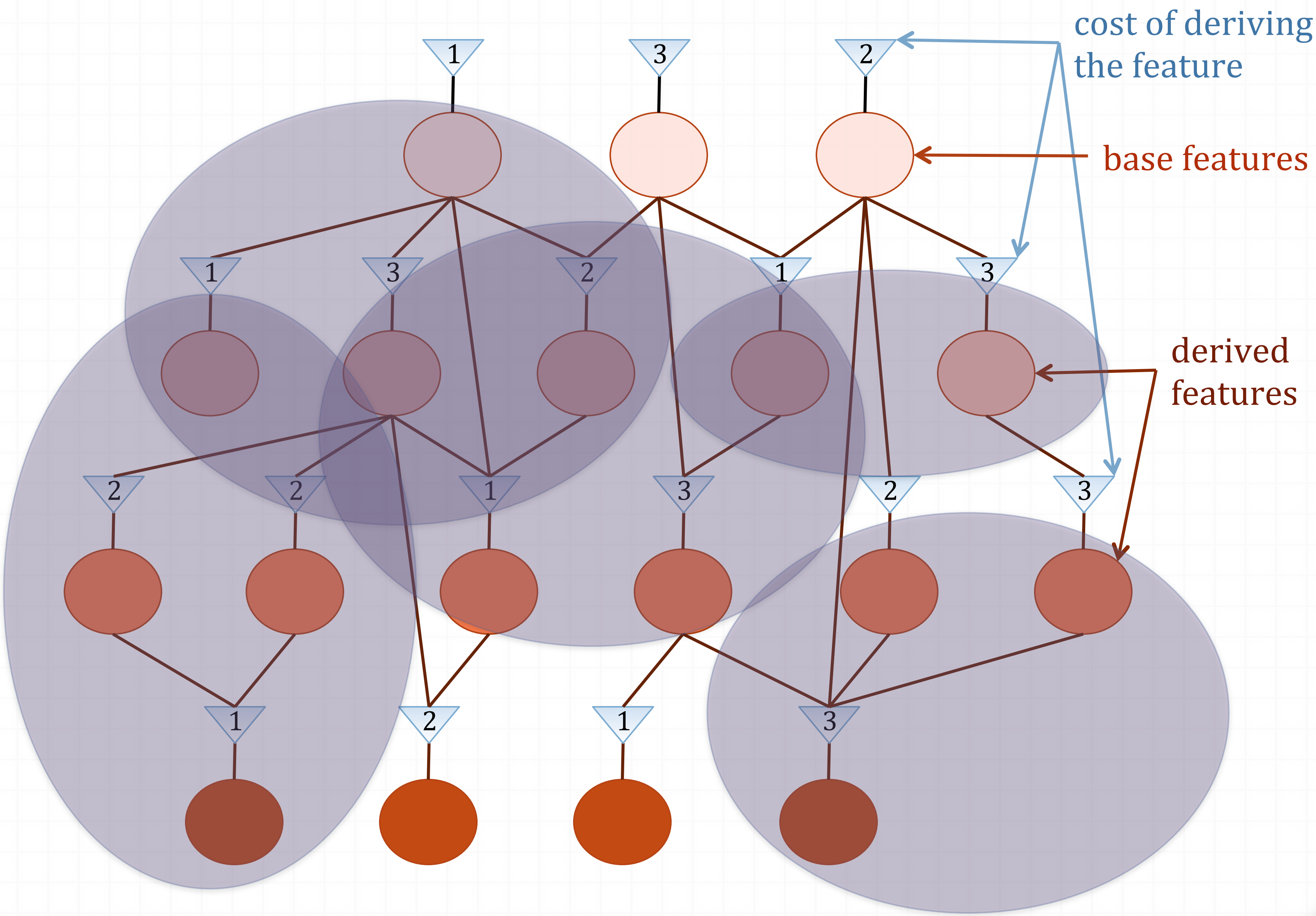
Preliminary Results

We applied the procedure to the vital sign monitoring data. There are a total of 150 interdependent features.

Cost	MSE (CFS)	MSE (lasso)
0	0.777	0.777
1	**0.344**	0.435
2	**0.246**	0.250
4	**0.244**	0.250
6	**0.244**	0.250
12	0.244	0.244

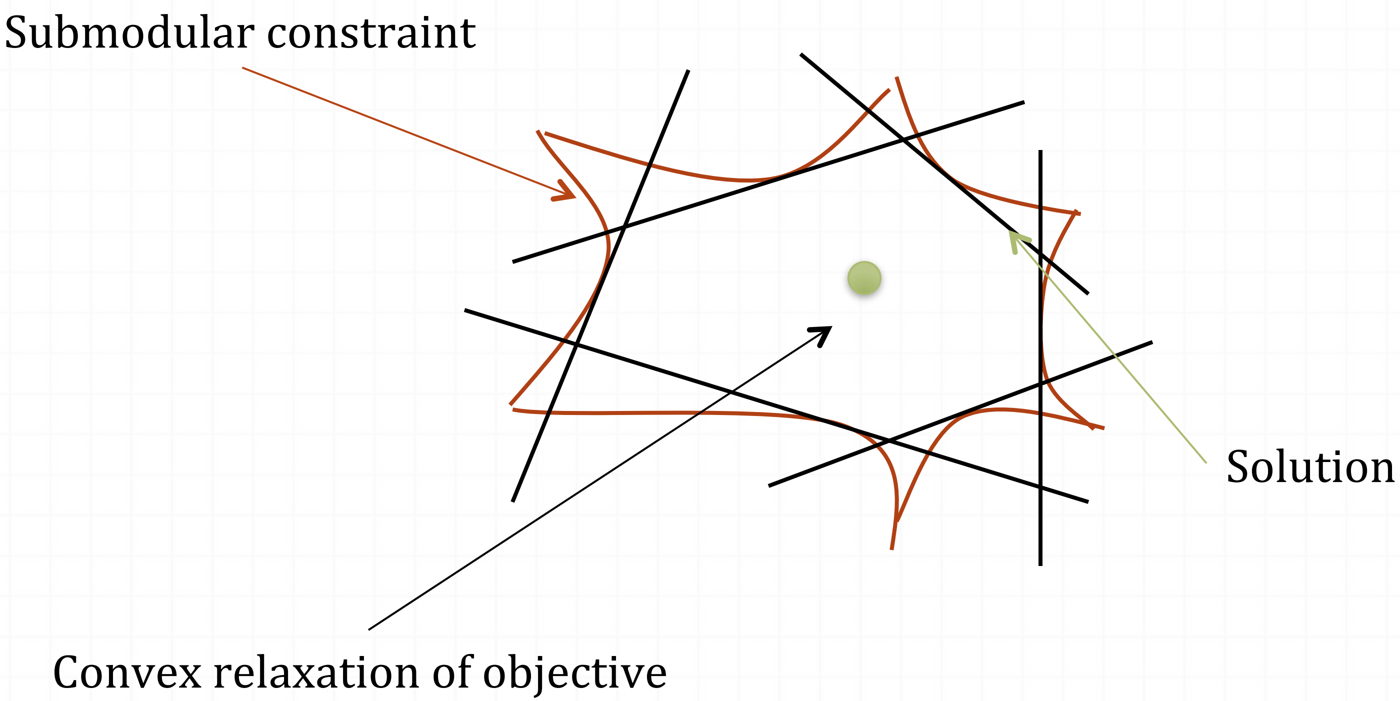
Here, the cost of all base features is a unit, and one cost unit is added for each additional operation which needs to be performed to obtained derived features.

Adding submodular cost constraints



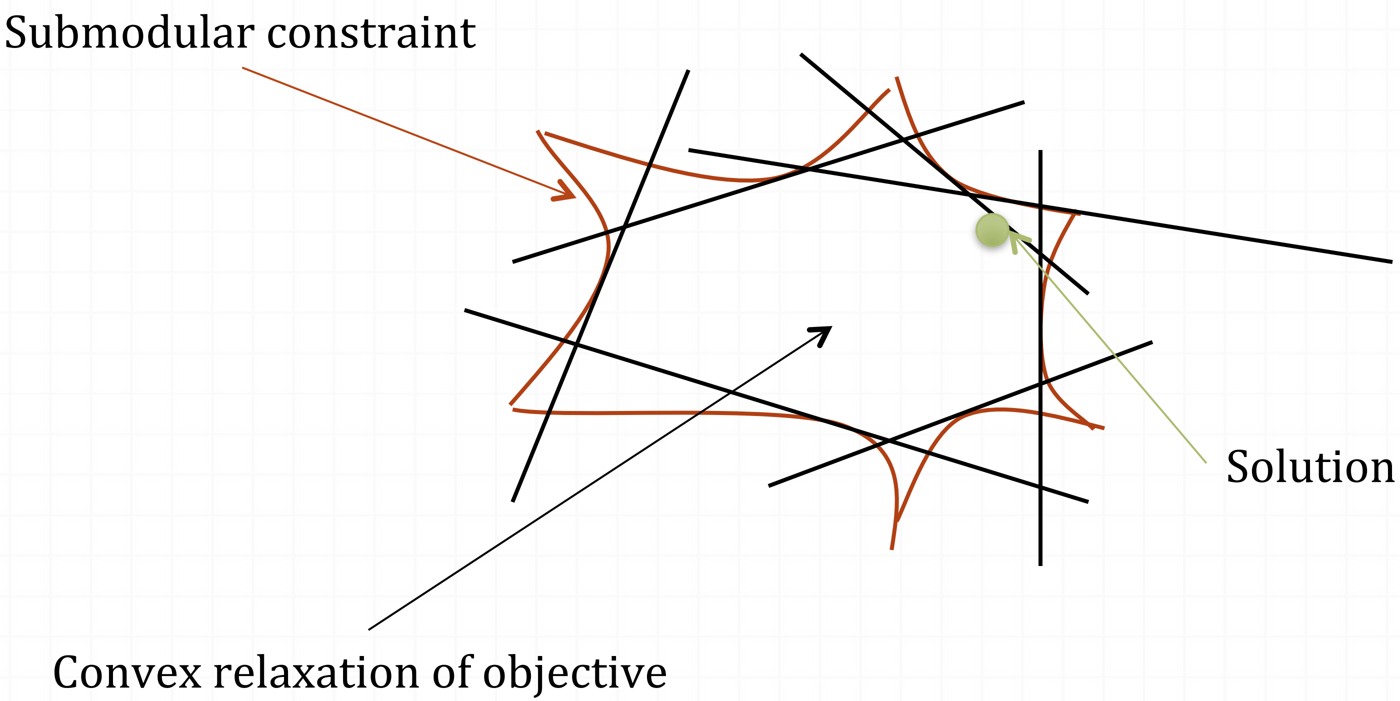
Adding submodular cost constraints

- We express this as an optimization with an approximately submodular objective with submodular cost constraints
- Idea: linearize, solve, re-linearize, improve solution …



Adding submodular cost constraints

- We express this as an optimization with an approximately submodular objective with submodular cost constraints
- Idea: linearize, solve, re-linearize, improve solution …



Timeline

Contribution	Status	Estimated completion	References
Informative Projection Recovery	completed	Spring 2013	[1],[2],[3],[5]
Active IPR Framework	completed	Spring 2014	[4]
Low-dimensional Model Learning for Feature Hierarchies	in progress	Winter 2015	
Online Cost Constrained Subset Selection Policies	future work	Spring 2015	
Efficient IPR and extensions	in progress	Summer 2015	

[1] Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In Advances in Neural Information Processing Systems 25 (NIPS), pages 3032–3040, 2012.

[2] Madalina Fiterau and Artur Dubrawski. Informative projection recovery for classification, clustering and regression. In International Conference on Machine Learning and Applications, volume 12, 2013.

[3] Fiterau M, Dubrawski A, Chen L, Hravnak M, Clermont G, Pinsky MR. Automatic identification of artifacts in monitoring critically ill patients. Intensive Care Medicine. 2013; 39 (Suppl 2]: S470.

[4]Fiterau M, Dubrawski A, Chen L, Hravnak M, Clermont G, Bose E, Guillame-Bert M, Pinsky MR. Artifact adjudication for vital sign step-down unit data can be improved using Active Learning with low-dimensional models. Intensive Care Medicine. 2014.

[5] Fiterau M, Dubrawski A, Chen L, Hravnak M, Bose E, Gilles, Michael. Archetyping artifacts in monitored noninvasive vital signs data. SSCM 2015.

[6] Wang D, Fiterau M, Dubrawski A, Hravnak M, Clermont G, Pinsky MR. Interpretable active learning in support of clinical data annotation. SSCM 2015