Ensembles for the Discovery of Compact Structures in Data

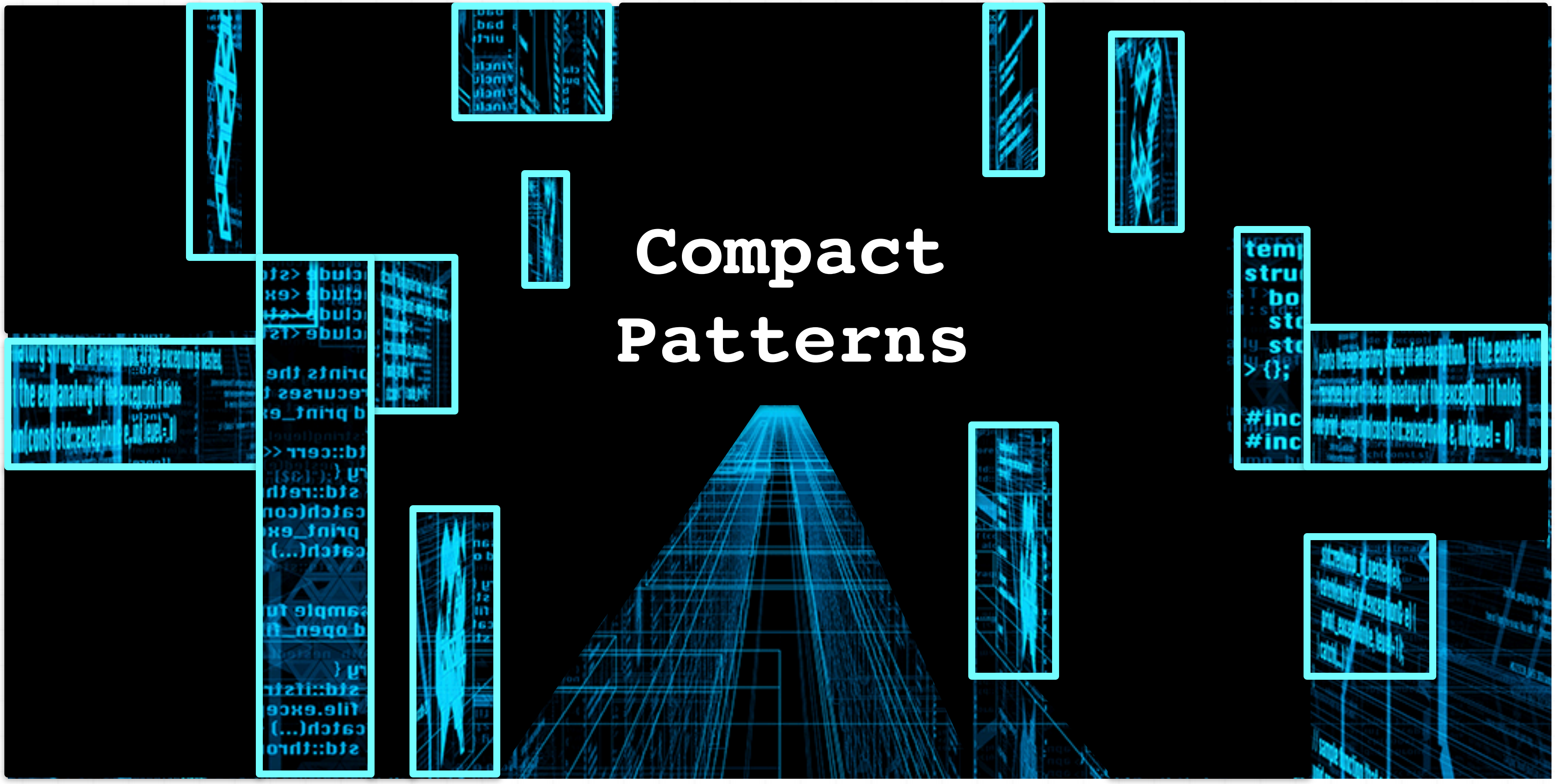
Madalina Fiterau
Carnegie Mellon University
School Of Computer Science
Machine Learning Department

Gatsby Unit, UCL
9th of March 2015

The Big Data Paradox

The Big Data Paradox

The Big Data Paradox



Talk Outline

Informative Projection Recovery (IPR)

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

Applications to Data Diagnostics

- Pattern Discovery in Clinical Data
- Finding Gaps in Training Data for Radiation Threat Detection

Back-propagation Forests

- Learning Fuzzy Decision Trees Using Back-propagation
- Potential Extensions of BP Forests

Informative Projection Recovery

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

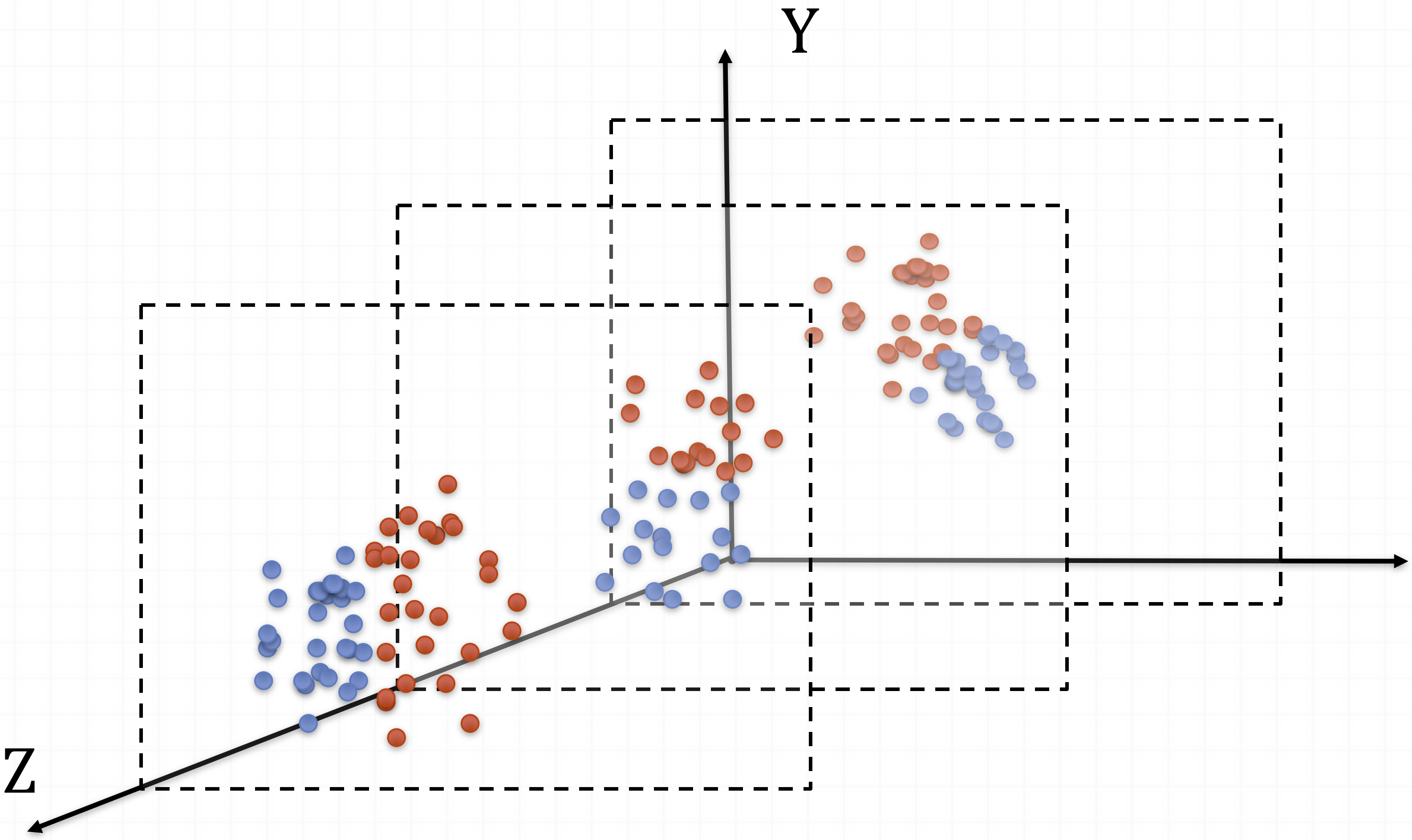
BORDER CONTROL

DIAGNOSTICS

VEHICLE CHECKS

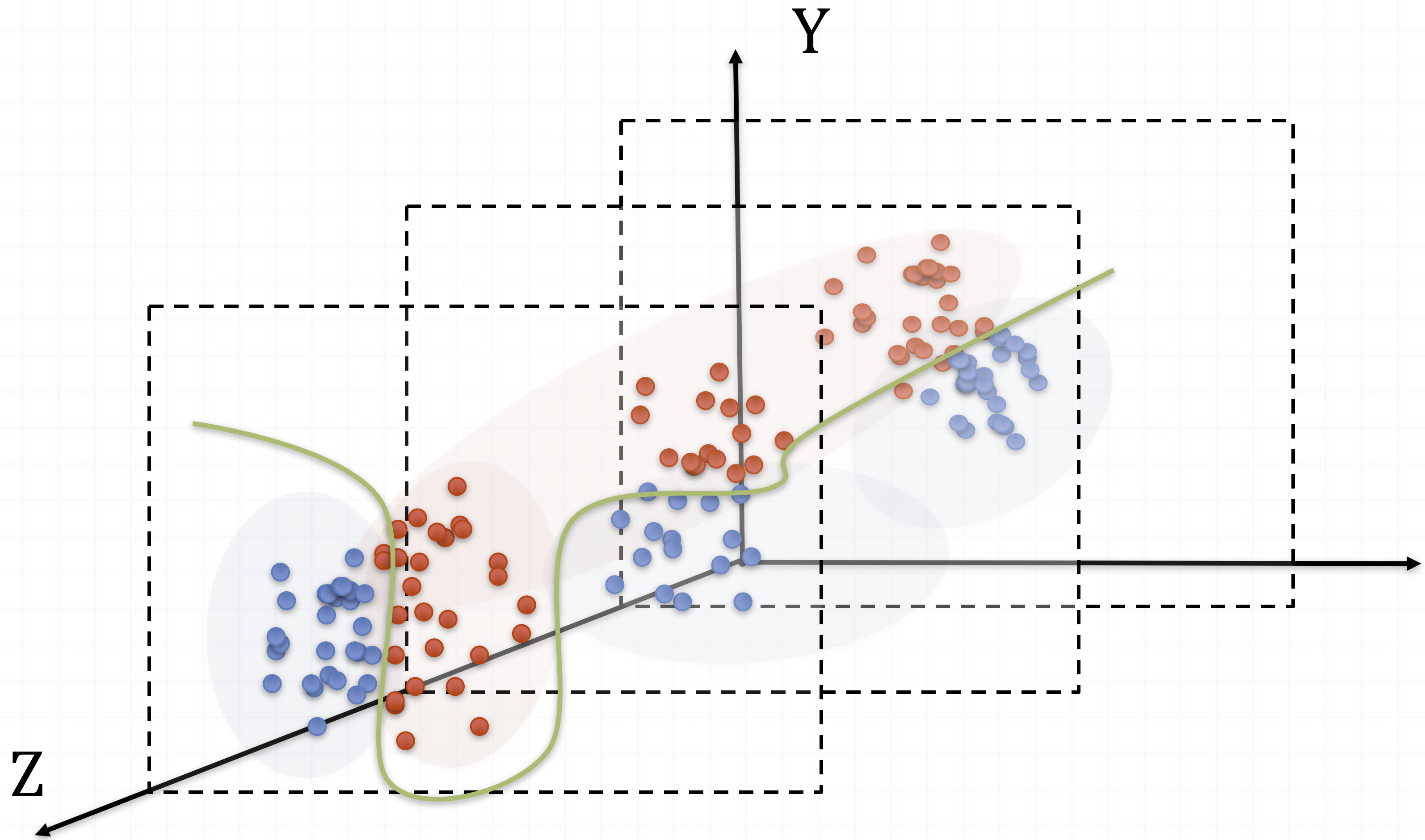
Sparse Predictive Structures

Heterogeneous data



Sparse Predictive Structures

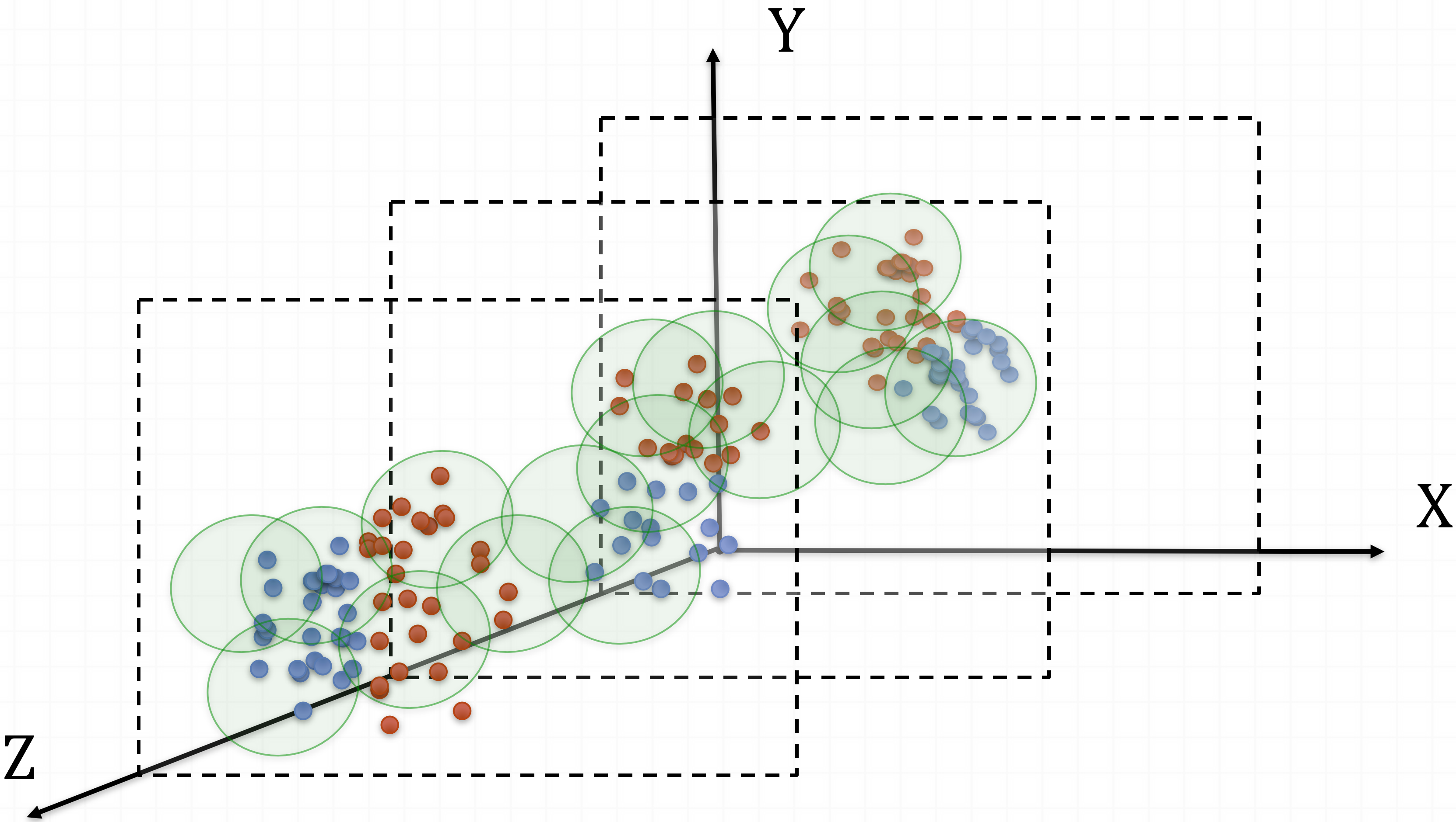
Learning Global Models



Issue: different features are relevant in different parts of the input space.

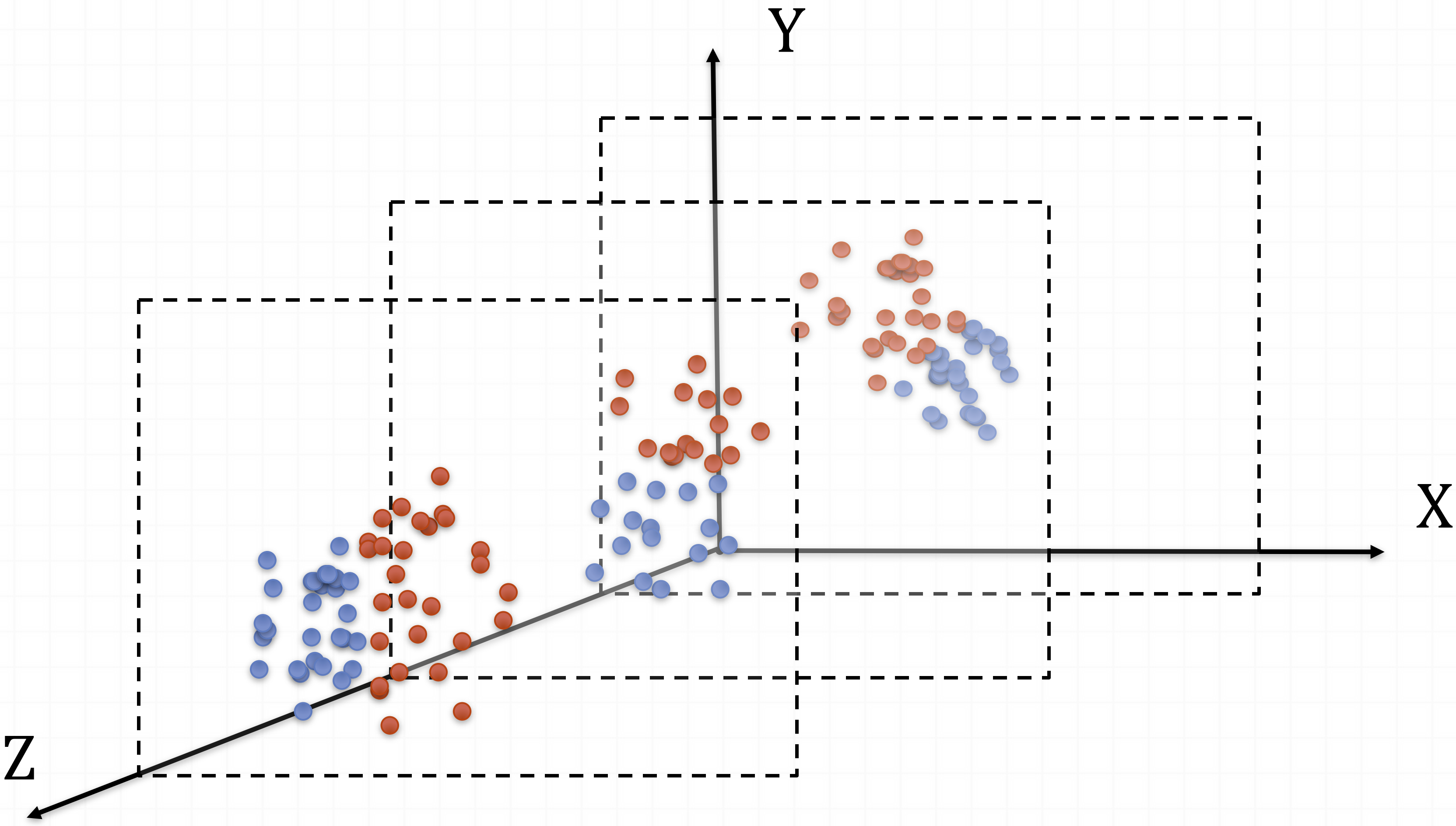
Sparse Predictive Structures

Learning Local Models

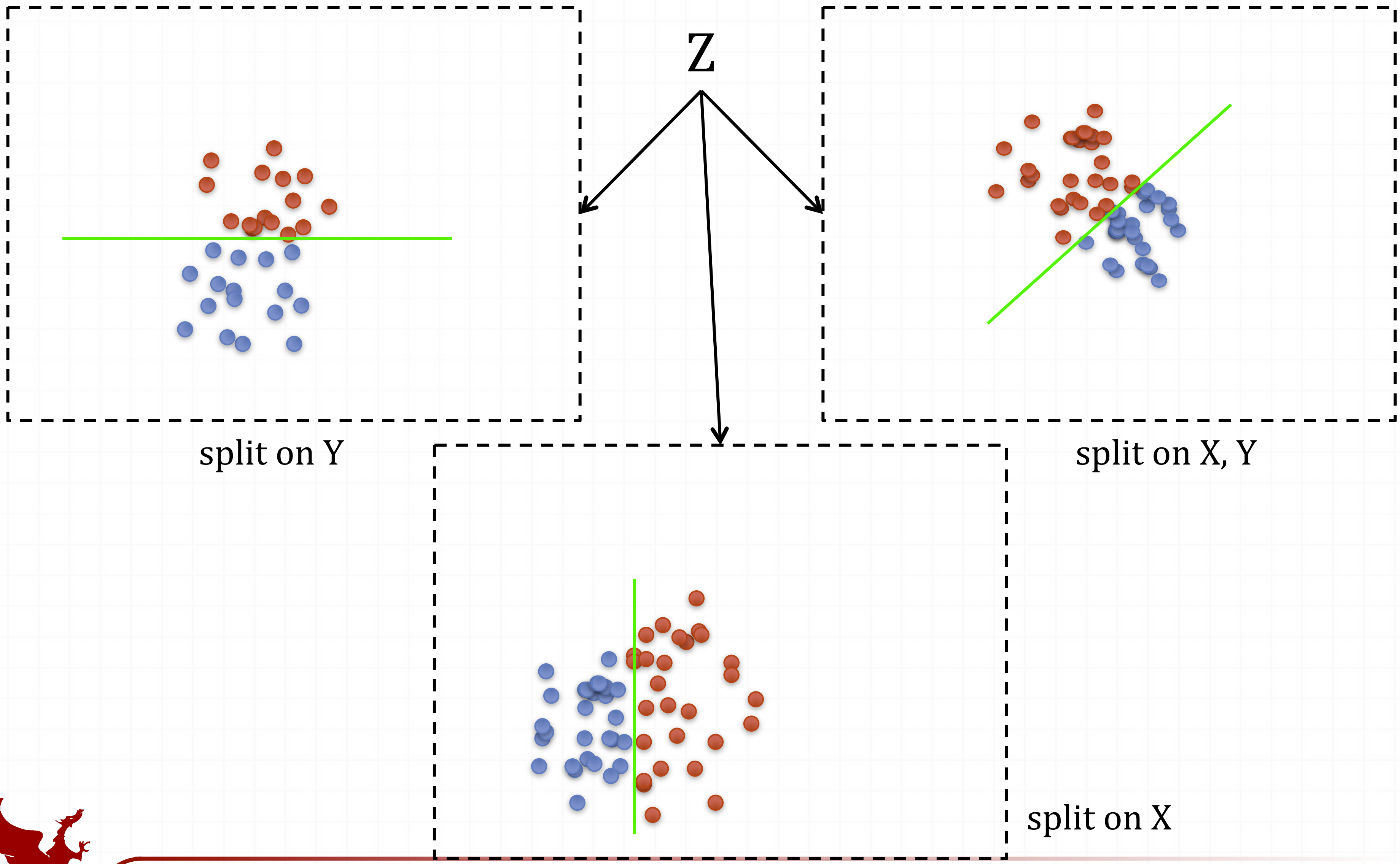


Issue: insufficient training data in the neighborhood or the sample.

Compact Partitioning Models



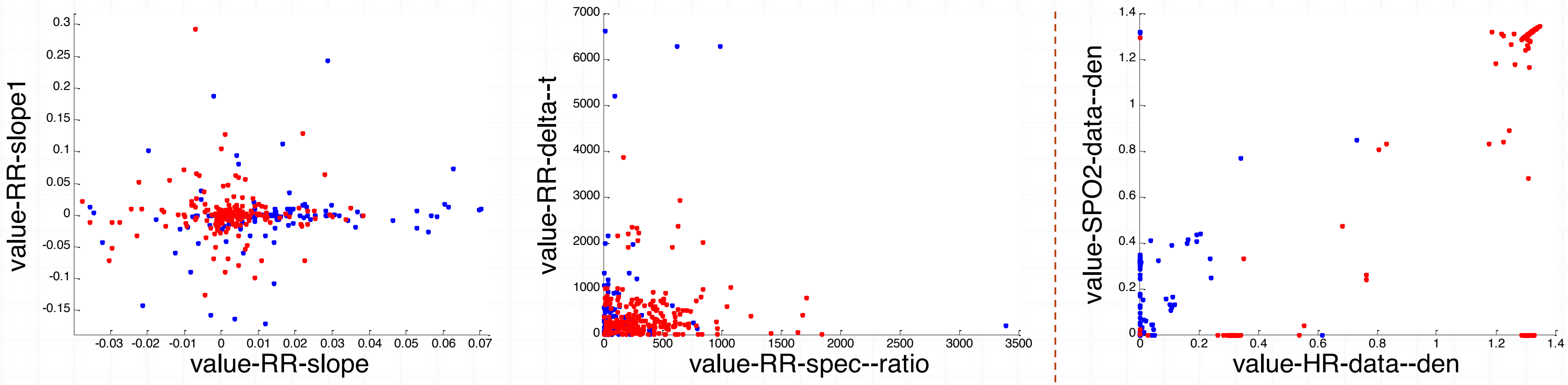
Compact Partitioning Models



Informative Projection Retrieval

- Select low-d subspaces which allow confident classification

- Clinical data example: vital signs and derived features



Noisy data

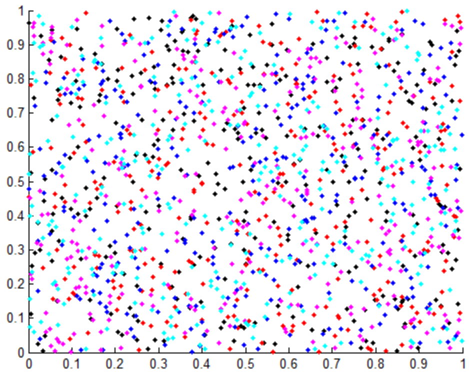
Clear separation

- **RIPR** = Regression-based Informative Projection Retrieval[1]

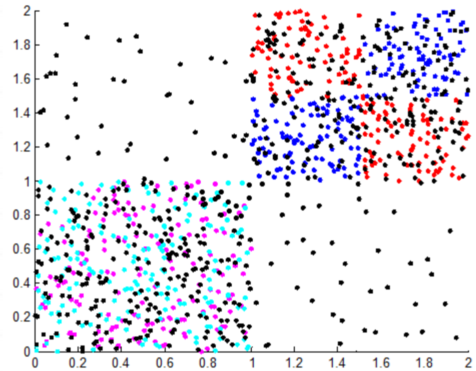
[1] Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In Advances in Neural Information Processing Systems 25 (NIPS), pages 3032–3040, 2012.

Dataset Assumptions

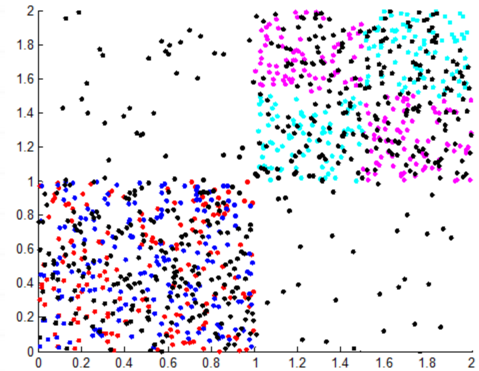
- Only a small subset of the projections have useful structure
- Projections are complementary, dealing with different samples



Aspect of most projections



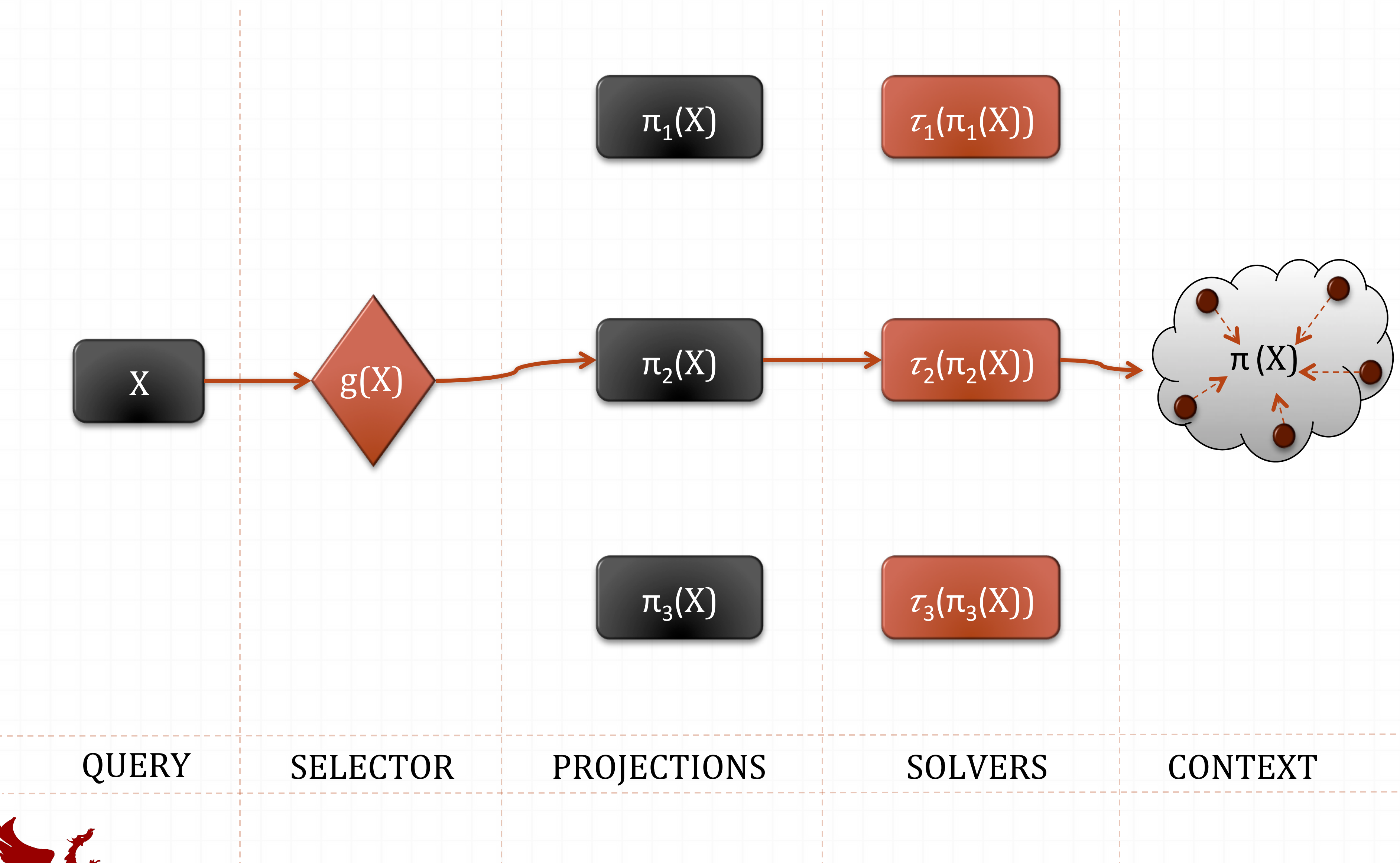
IP for blue/red group



IP for light blue/purple group

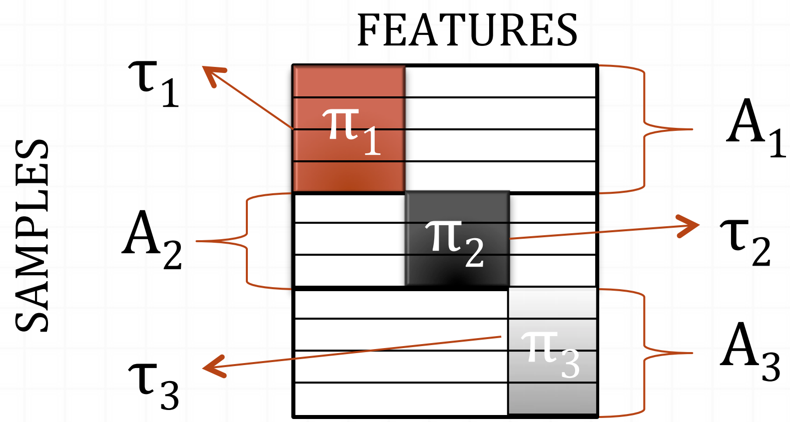
- Engineered data - unintentionally introduced artifacts usually show in low-dimensional patterns
- Clinical data - multiple sub-models reflect specifics of particular conditions and patient characteristics

RIPR Framework



Dual-Objective Training Process

1. Data is split across informative projections
2. Train solvers using data assigned to each projection



RIPR Model

Model Components

- P - set of axis-aligned sub-spaces, max. d features
- T - set of solvers trained on each of the projections in P
- g – determines the projection/solver for a point x, $(\pi_{g(x)}, \tau_{g(x)})$
- $\ell(\tau_{g(x)}(\pi_{g(x)}(x)), y)$ represents the model loss at point x

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

$\mathcal{M}_d$

Small set of projections

Target model

Selection function

Solvers

RIPR Objective Function

Model Components

- P - set of axis-aligned sub-spaces, max. d features
- T - set of solvers trained on each of the projections in P
- g – determines the projection/solver for a point x, $(\pi_{g(x)}, \tau_{g(x)})$
- $\ell(\tau_{g(x)}(\pi_{g(x)}(x)), y)$ represents the model loss at point x

Model Components

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

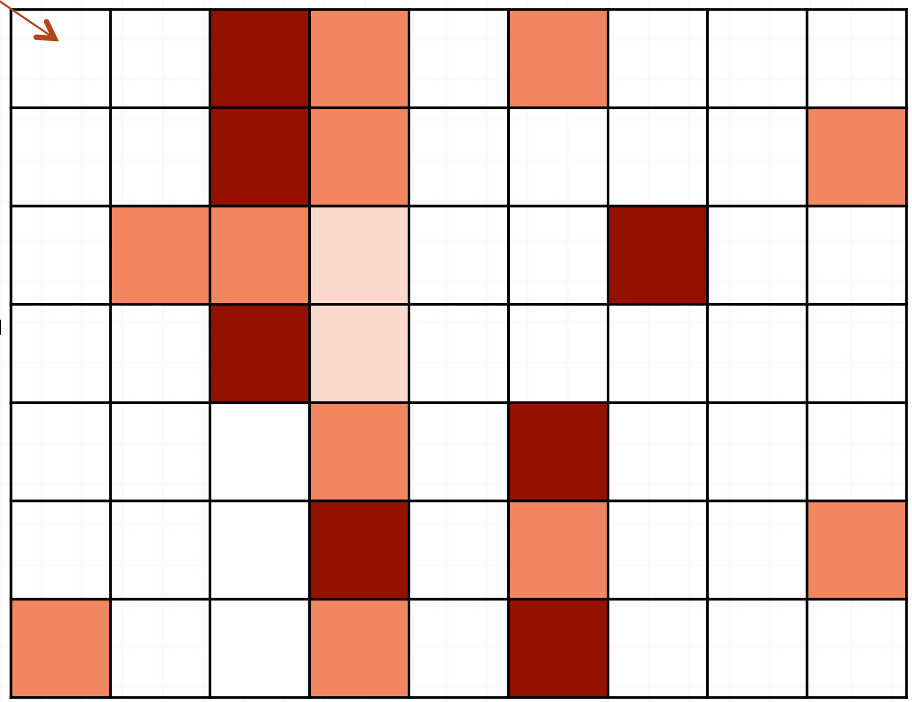
Expected loss for task solver trained on projection assigned to point

Starting point: the loss matrix

Loss estimators

Projections

Samples

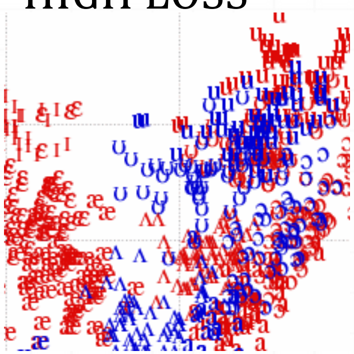


low loss

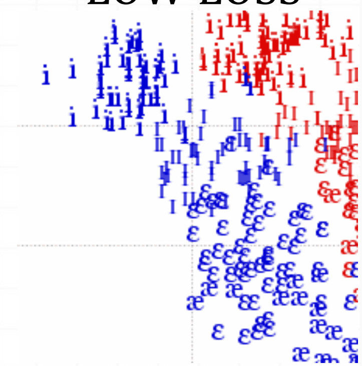
moderate loss

high loss

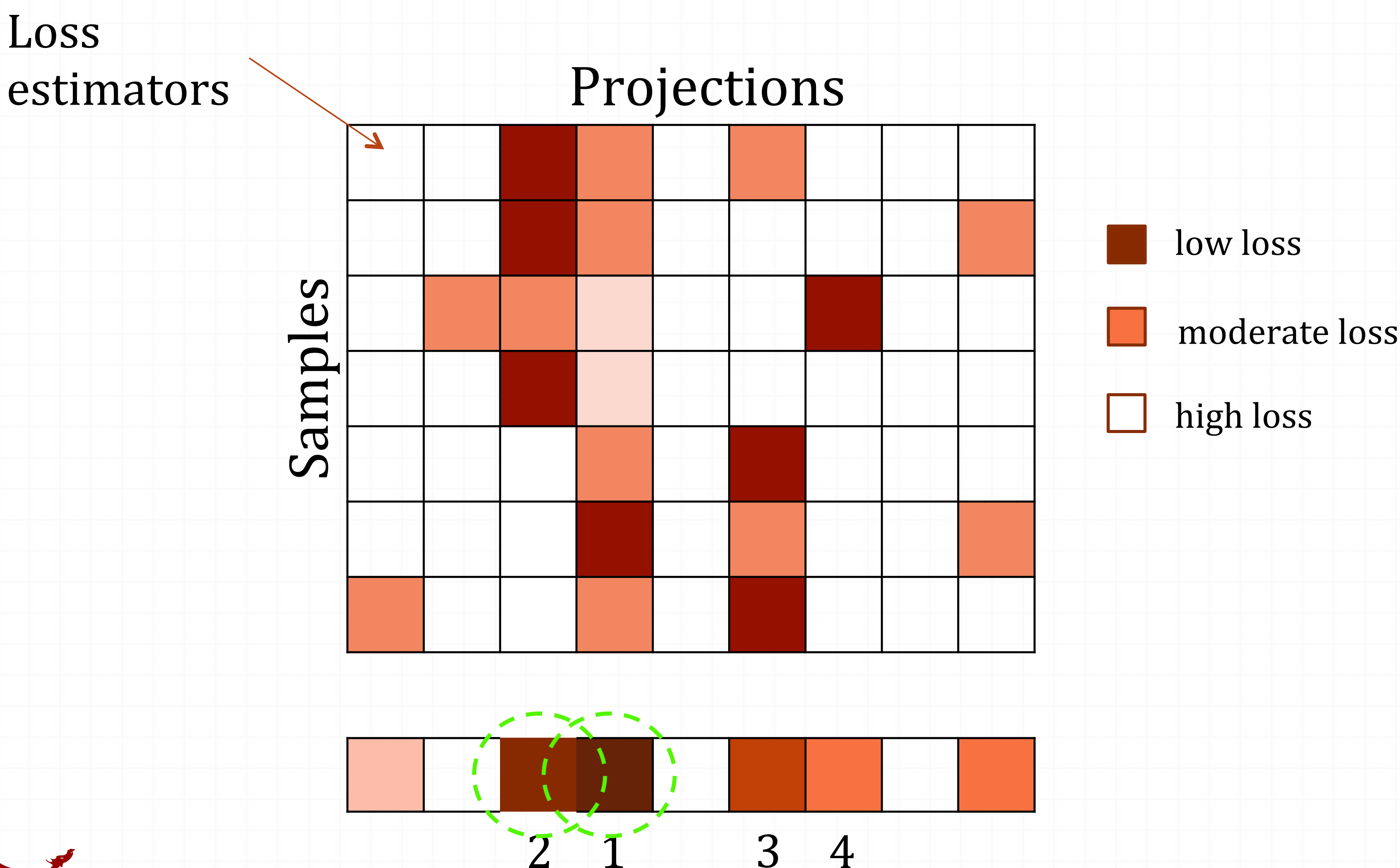
HIGH LOSS



LOW LOSS

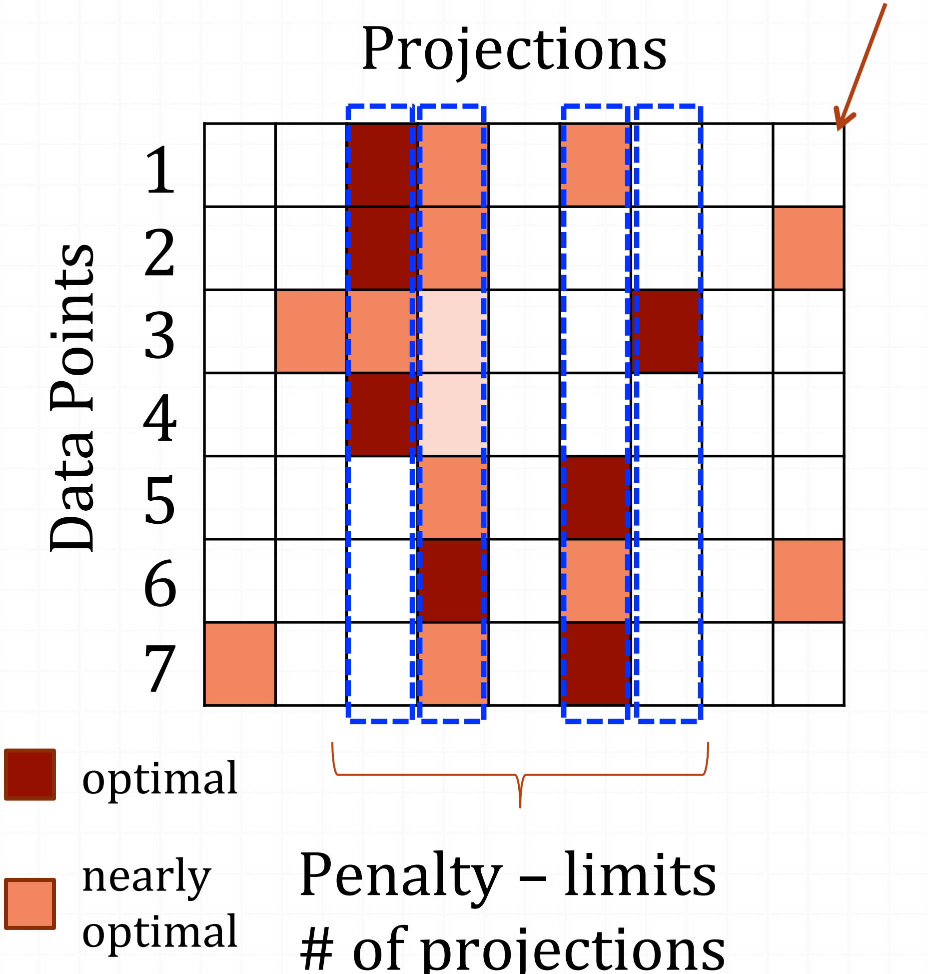


Starting point: the loss matrix



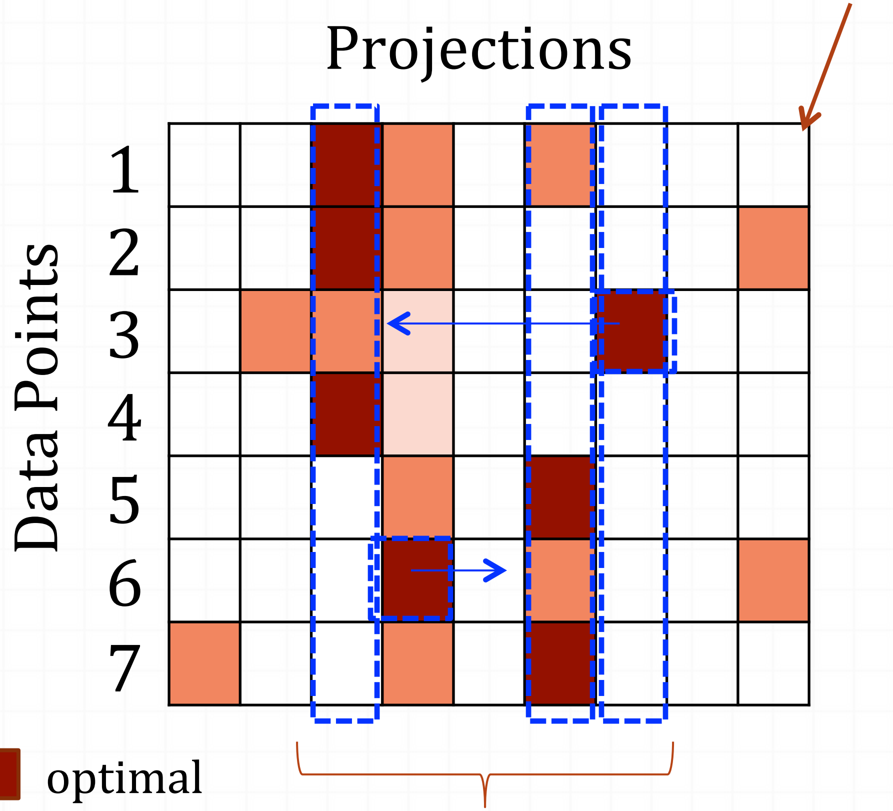
The Optimization Procedure

Matrix of Loss Estimators (L)



The Optimization Procedure

Matrix of Loss Estimators (L)



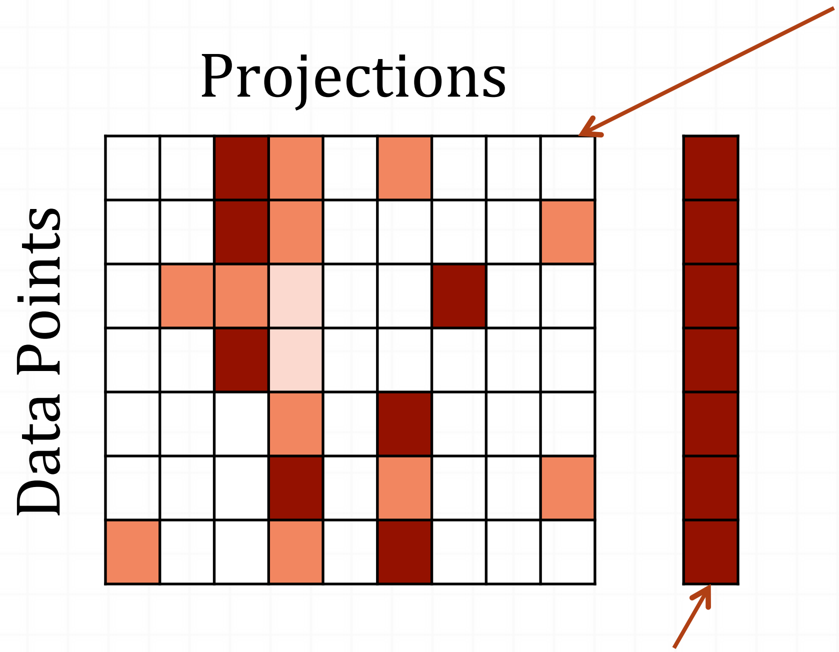
optimal

nearly optimal

some points use suboptimal projections

The Optimization Procedure

Matrix of Loss Estimators (L)



optimal

nearly optimal

- L_{ij} is the loss for 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
- Convex program: $B = \min_B ||T - L \odot B||_1 + \text{reg}(B)$

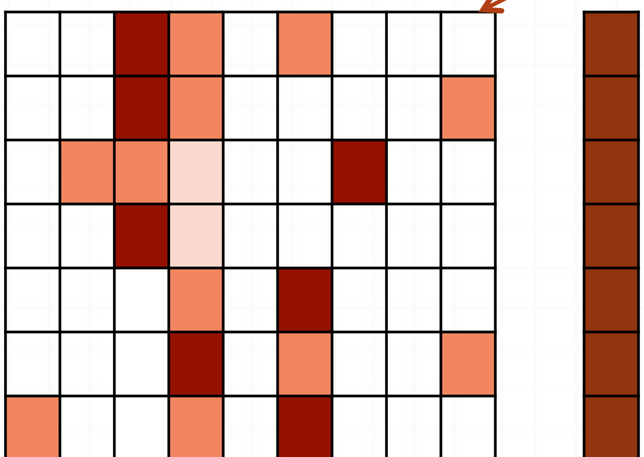
where $L \odot B \stackrel{def}{=} \sum_{j=1}^{m} L_{.,j} B_{.,j}$

The Optimization Procedure

Matrix of Loss Estimators (L)

Projections

Data Points



Target Loss (T)

IPR problem - solved through this regression

optimal

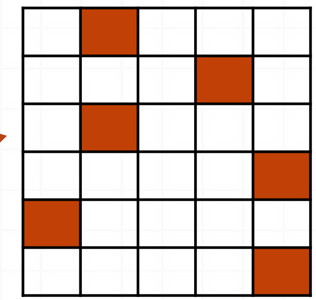
nearly optimal

- Convex program: $B = \min_B ||T - L \odot B||_1 + \text{reg}(B)$

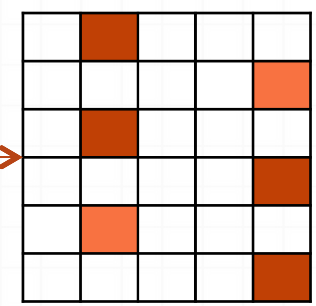
where $L \odot B \overset{def}{=} \sum_{j=1}^{m} L_{.,j} B_{.,j}$

Performing the Regression

- 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]:

- (Semi-supervised) classification
- Clustering
- Regression
- Active learning

Loss matrix computed differently for each task

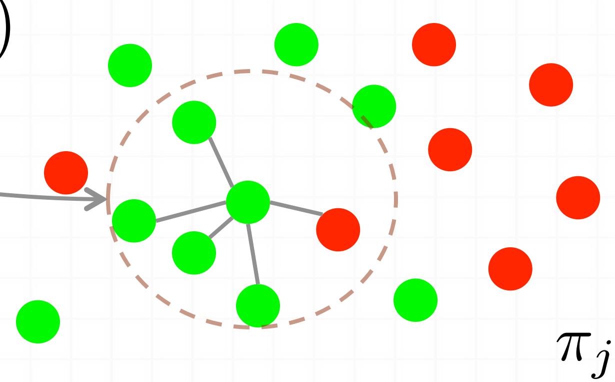
RIPR can solve any learning task for which the risk can be decomposed using consistent loss estimators.

[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.

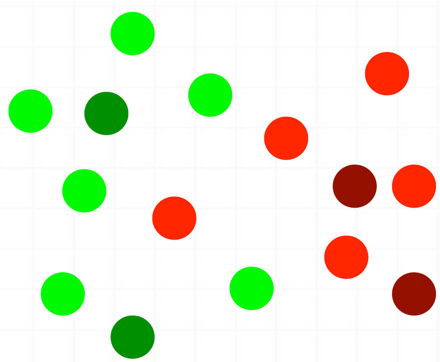
Loss Estimators: Classification

- Neighbor-based estimator for conditional entropy*

$$H(Y|\pi_j(X); g(X) = j)$$



- For unlabeled samples, assume label with lowest loss



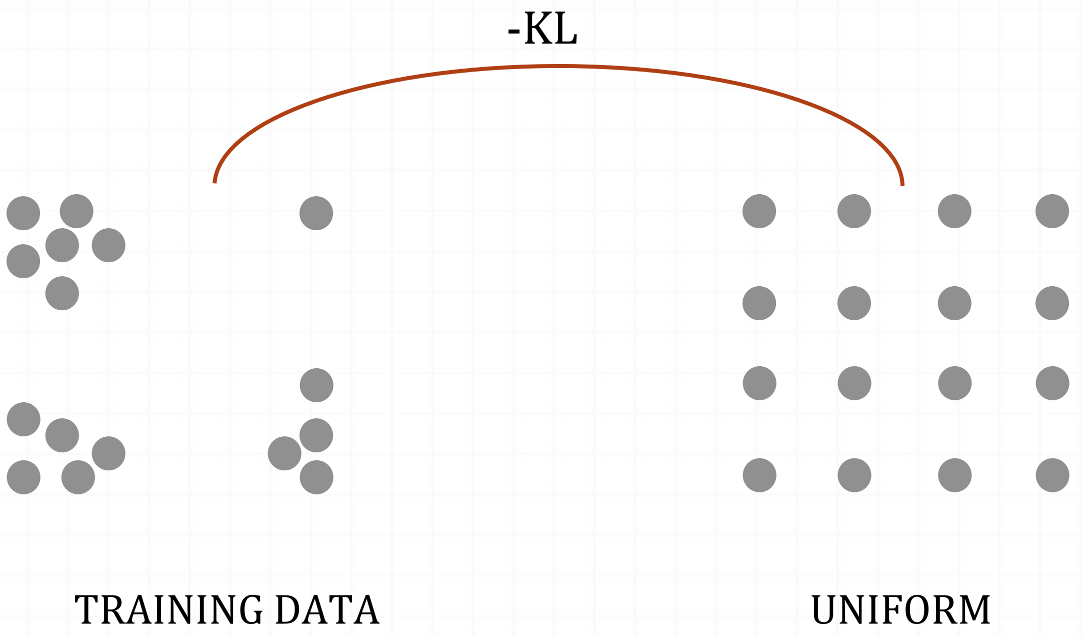
*Based on the divergence estimator by Poczos and Schneider, *"On the estimation of alpha-divergences"* (AISTATS 2011)

Classification Results

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

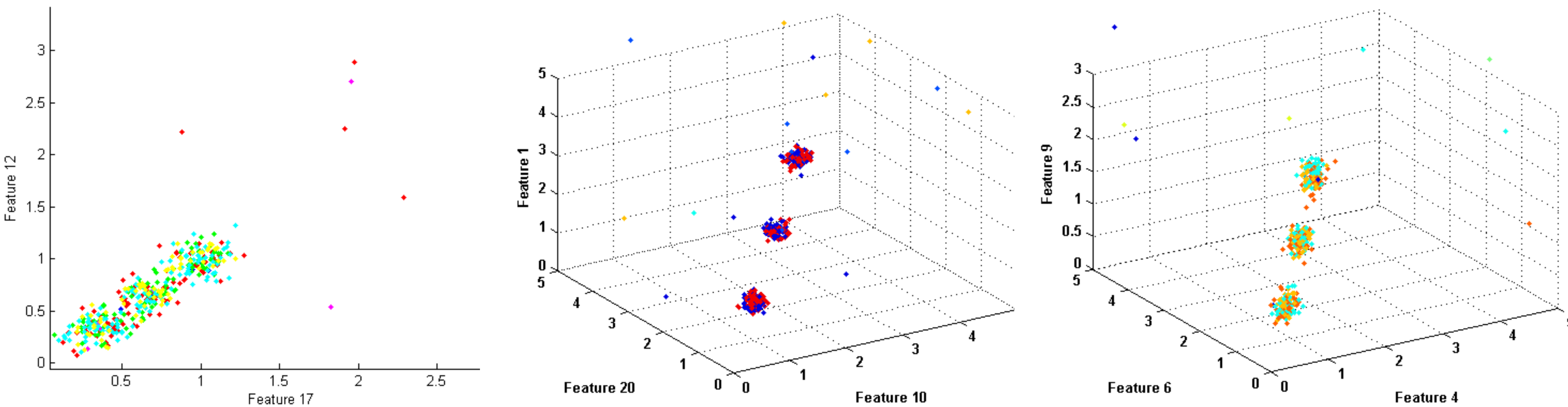
Loss Estimators: Clustering

- Density-based clustering
- Loss is lower for high density areas
- Negative KL divergence to uniform

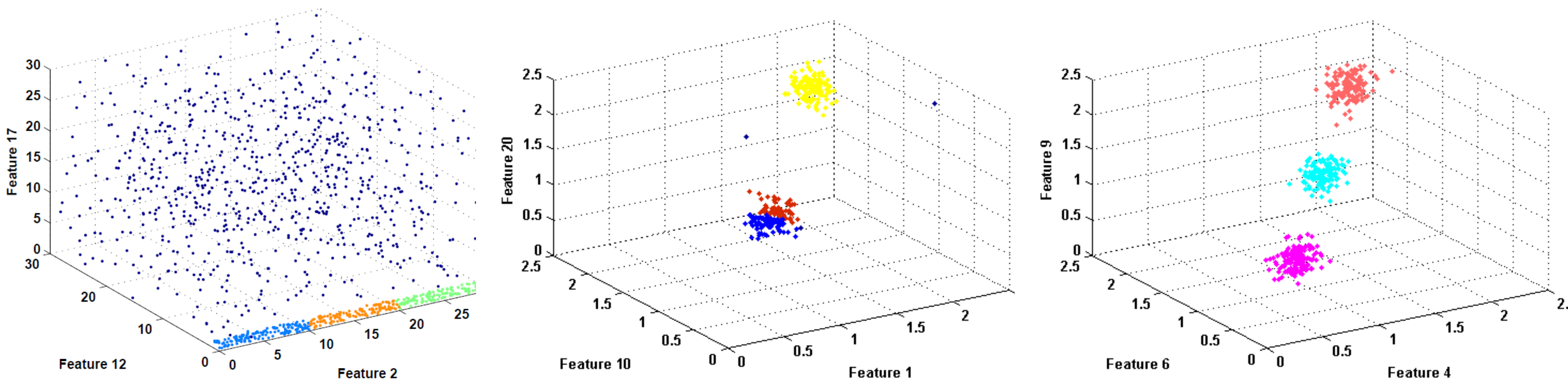


Low-d Clustering: Why it Works

K-Means model projected on (known) informative features



Representation of RIPR model – recovered projections and assigned data

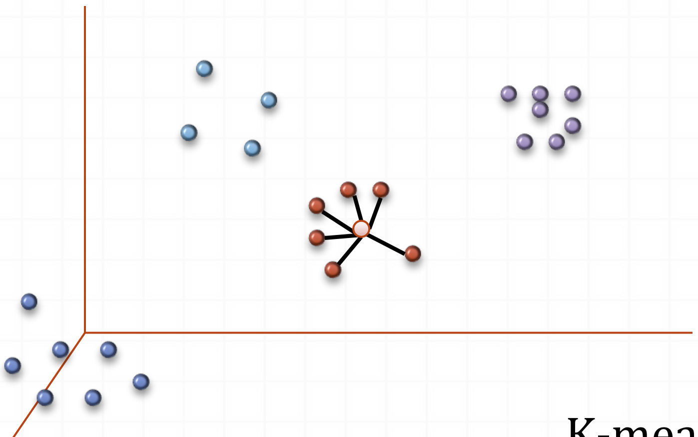


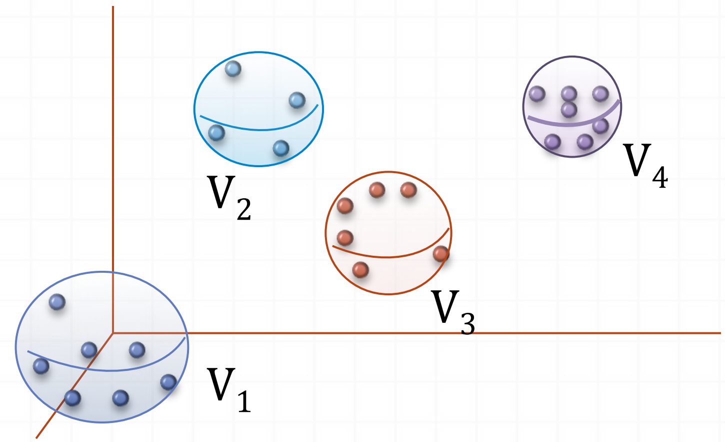
The hidden structure in data is clearly revealed by the RIPR model.

Clustering Evaluation Metrics

DISTORTION (goodness-of-fit)

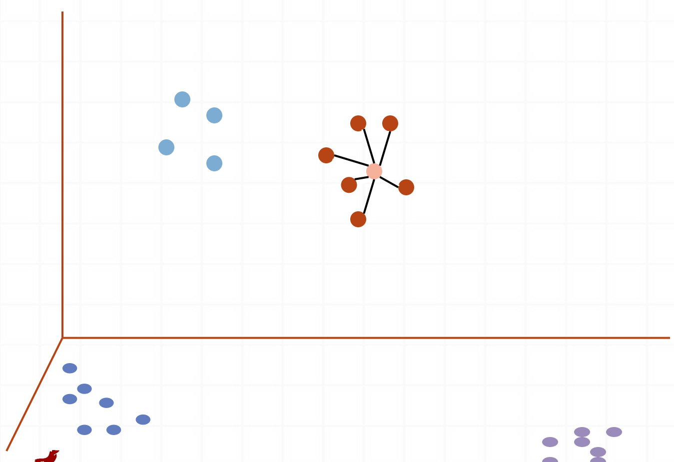
LOG CLUSTER VOLUME (compactness)

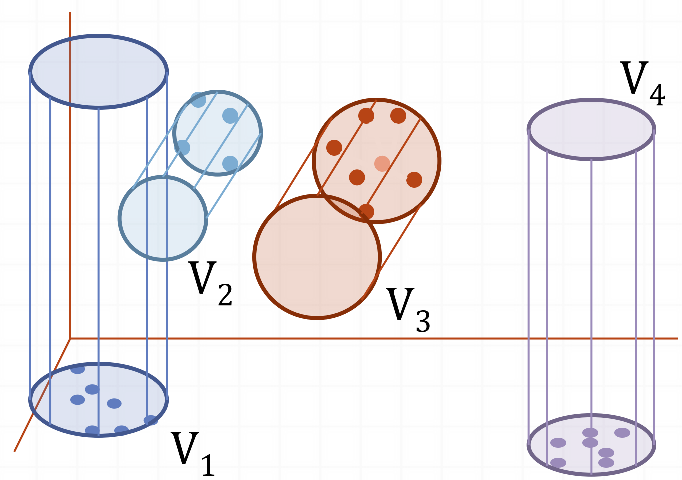




K-means Model

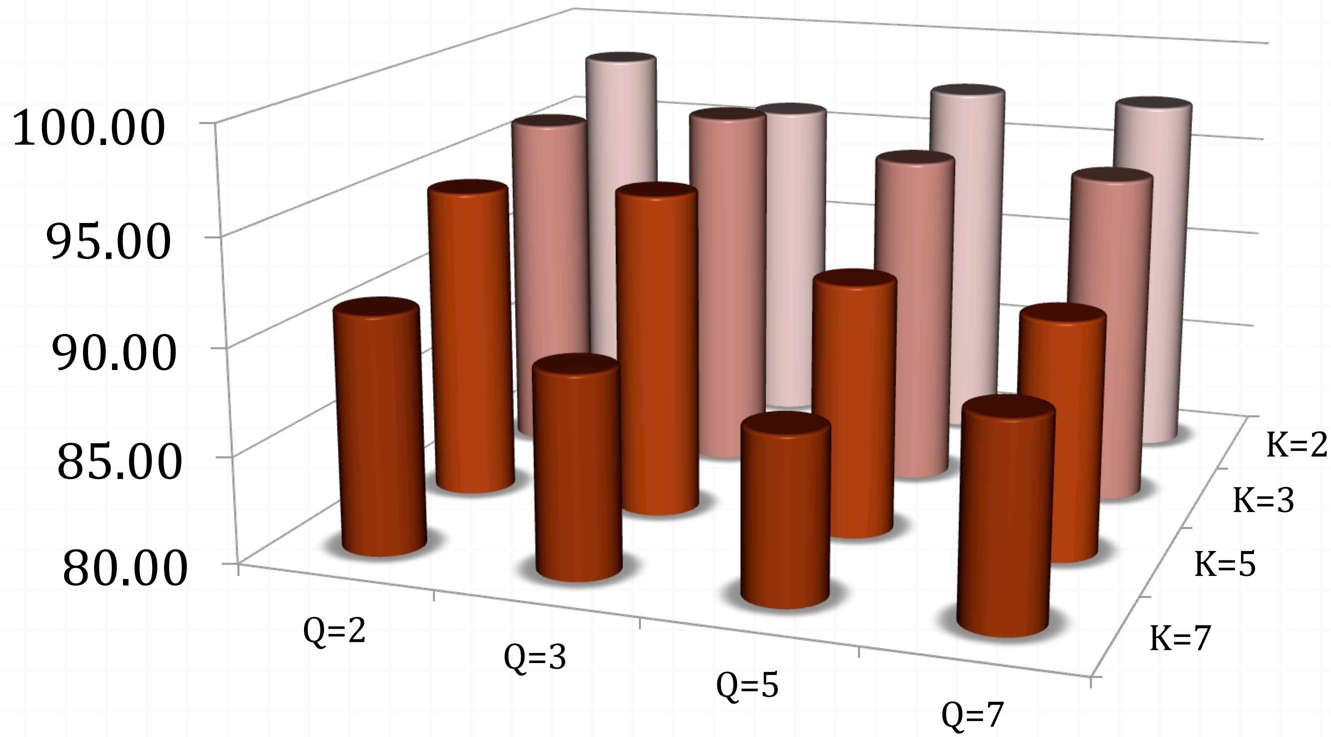
Ripped K-means Model





Clustering on 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 on 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

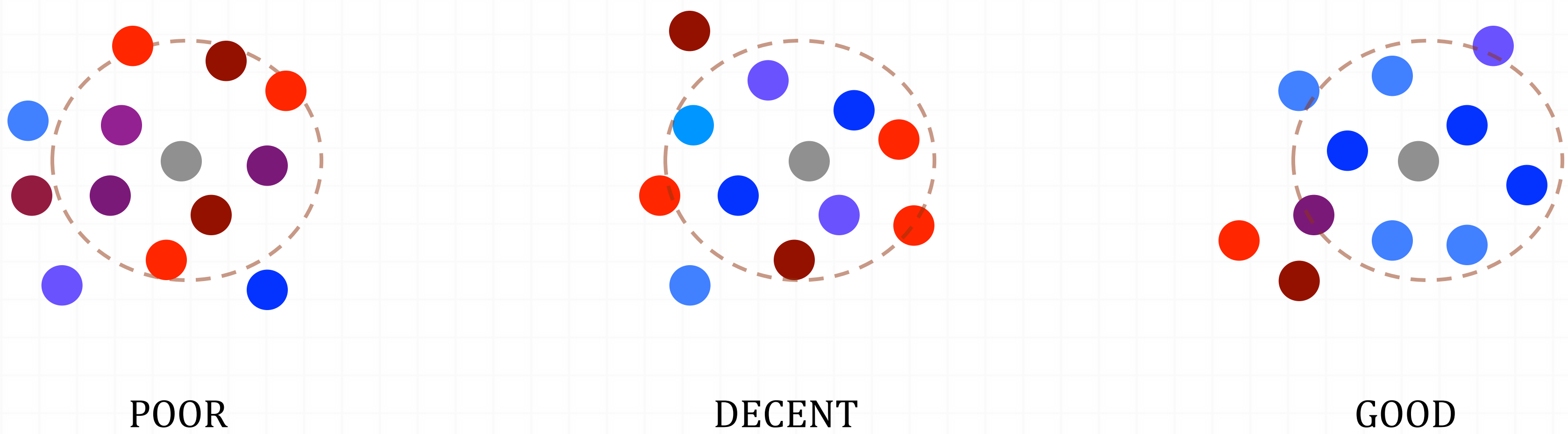
LOWER IS BETTER. RIPR MODELS ALWAYS HAVE A SMALLER TOTAL VOLUME.

Loss Estimators: Regression

- Estimates error in point neighborhood

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

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



Regression on Artificial Data

ACCURACY OF RIPPED SVM COMPARED TO ACCURACY OF STANDARD SVM

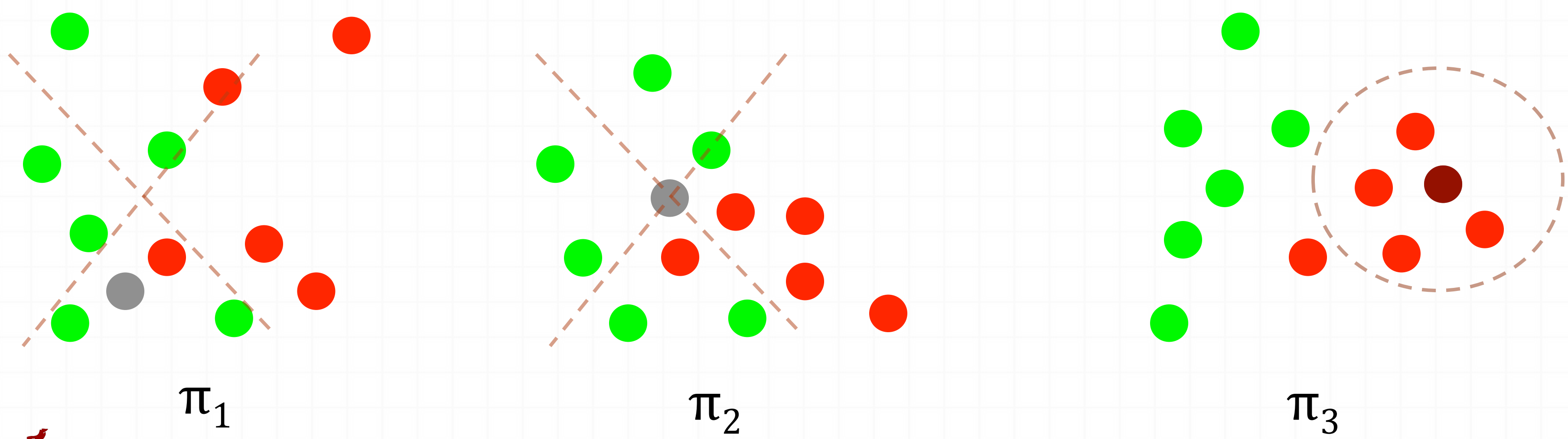
- THE NUMBER OF INFORMATIVE PROJECTIONS : 2-10 (OUT OF 45)
- PERCENTAGE OF NOISY SAMPLES: 0-50% (OUT OF 1600)

NOISY SAMPLES	IP #	2	3	5	7	10
	MSE RIPPED-SVM/MSE SVM					
	0%	0.19	0.23	0.45	0.20	0.53
	6.25%	0.53	1.24	0.57	0.48	0.55
	12.5%	0.52	0.68	2.76	0.49	0.69
	25%	1.58	1.17	0.82	0.94	1.38
	50%	1.33	6.33	1.23	0.76	0.95

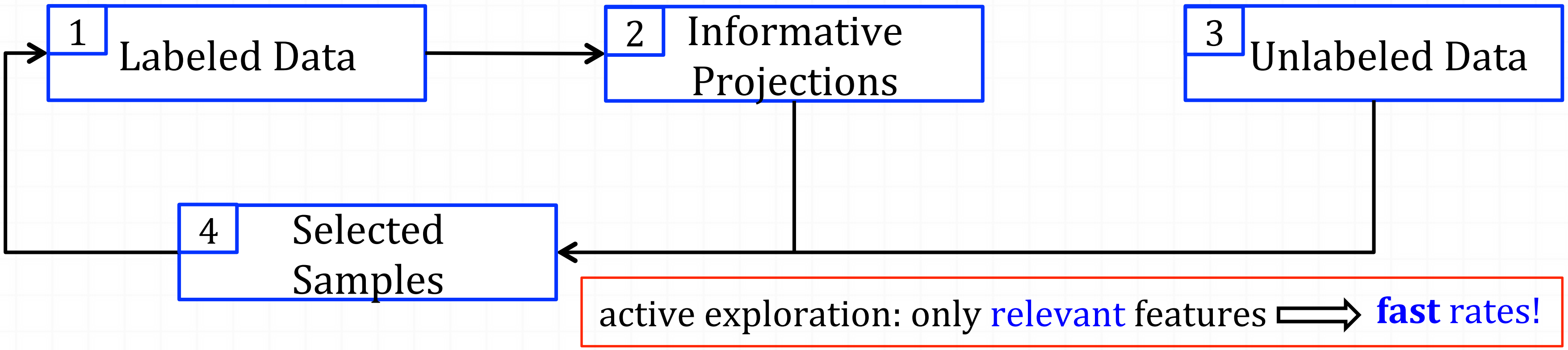
Query Assignment

- Problem: how to select the appropriate projection for a specific query x?
- Solution: select the projection in the **learned subset P** for which the estimated loss is the lowest.

$$(k*, y*) = argmin_{(k \in \{1 \ldots |P|\}, y \in \mathcal{Y})} \ell(\tau_k(\pi_k(x)), y)$$



Active Learning with RIPR



- Scoring functions
  - Uncertainty sampling
  - Query by committee
  - Information gain (best performance)
  - Low conditional entropy
- Clinical application: framework requests ***half of the labels*** requested by random forests with active learning

Talk Outline

Informative Projection Recovery (IPR)

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

Applications to Data Diagnostics

- Pattern Discovery in Clinical Data
- Finding Gaps in Training Data for Radiation Threat Detection

Back-propagation Forests

- Learning Fuzzy Decision Trees Using Back-propagation
- Potential Extensions of BP Forests

Applications to Data Diagnostics

Healthcare Alert Prediction

Nuclear Threat Detection

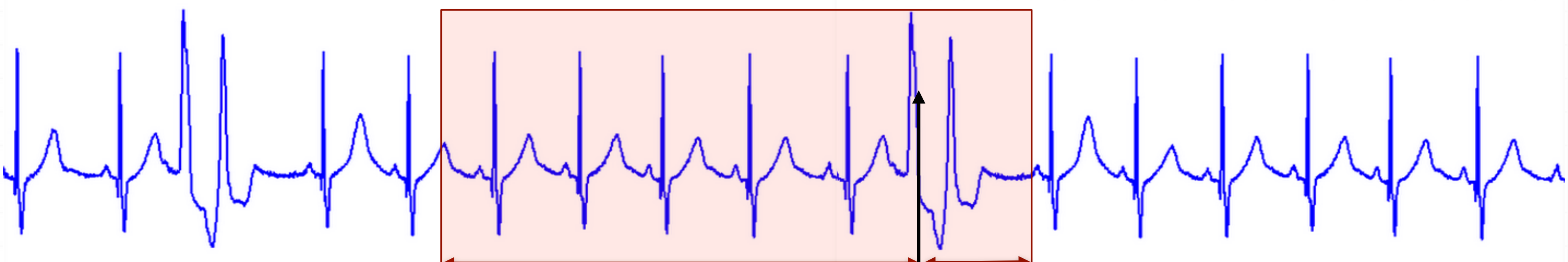
Learning from Multiple Datasets

Artifacts in Clinical Alerts

- Hear Rate<40 or >140
- Respiratory Rate<8 or >36
- Systolic Blood Pressure<80 or >200
- Diastolic Blood Pressure>110
- SPO_2<85%

Health alerts

some are
artifacts, not
true alerts

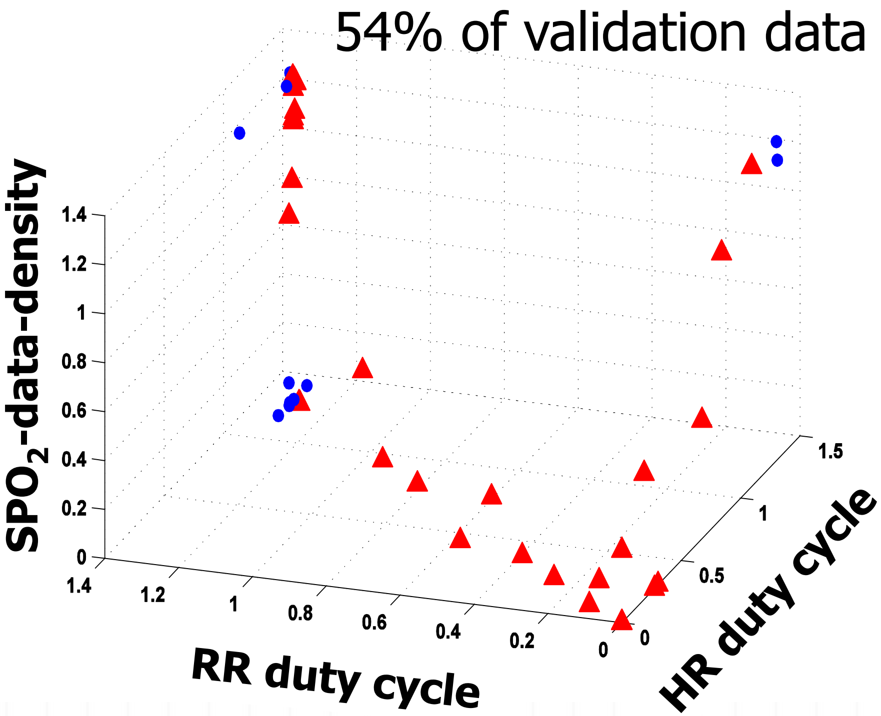


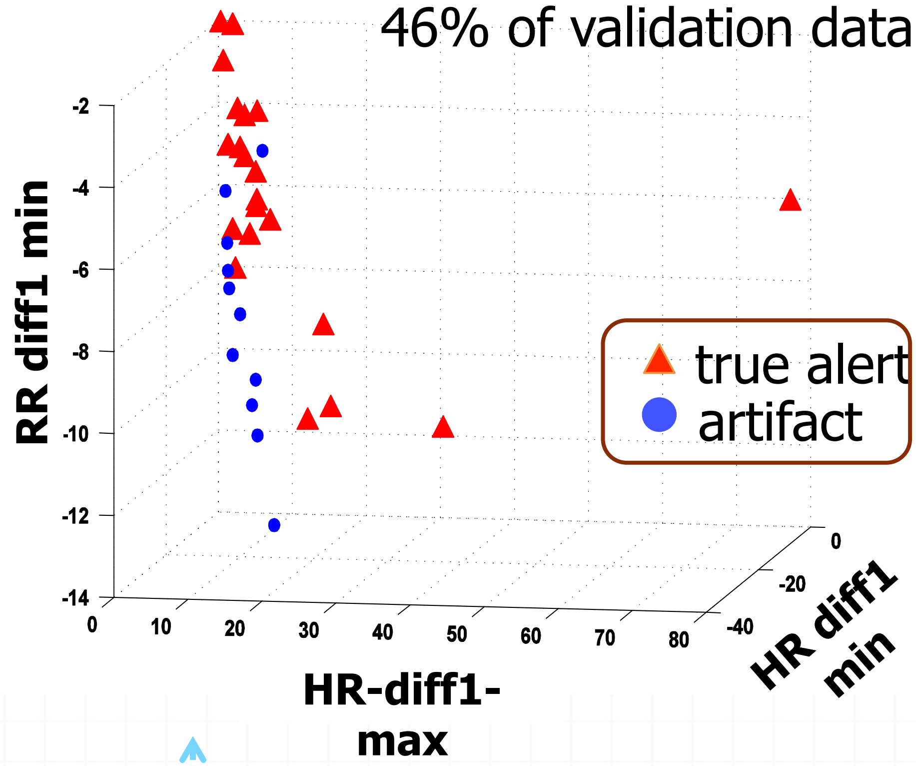
window of 4 minutes
preceding alert onset

alert duration

Features computed from time series include common statistics of each VS: mean, stdev, min, max, range of values, duty cycle ...

Artifacts in Clinical Alerts



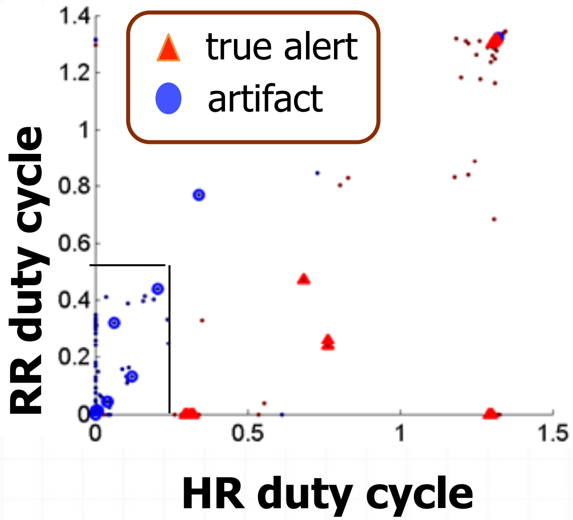


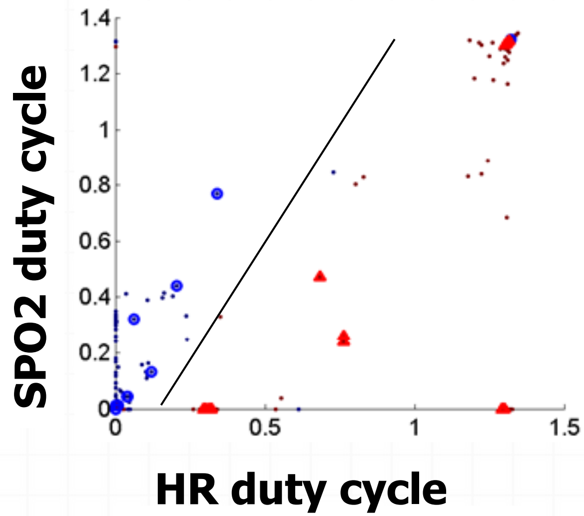
Alarm Type	RR	BP		SPO$_2$	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

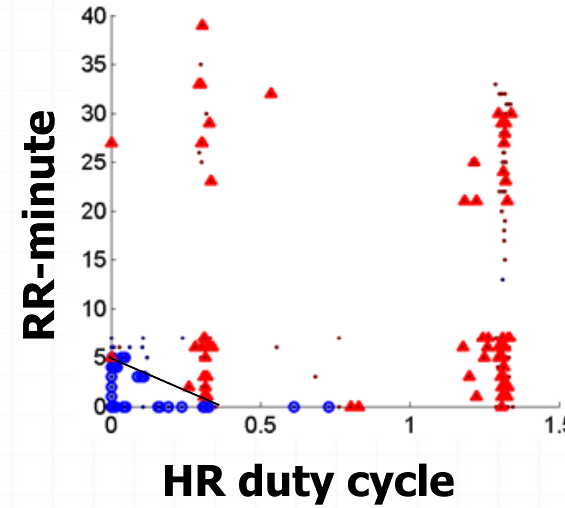
The retrieved projections enable domain experts to quickly validate alert labels.

Artifact Adjudication Rules

Informative Projections for SPO_2 alerts allow derivation of rules.







RR duty cycle <= 0.6
and
HR duty cycle <= 0.3

HR duty cycle –
SPO_2 duty cycle <= 0.2

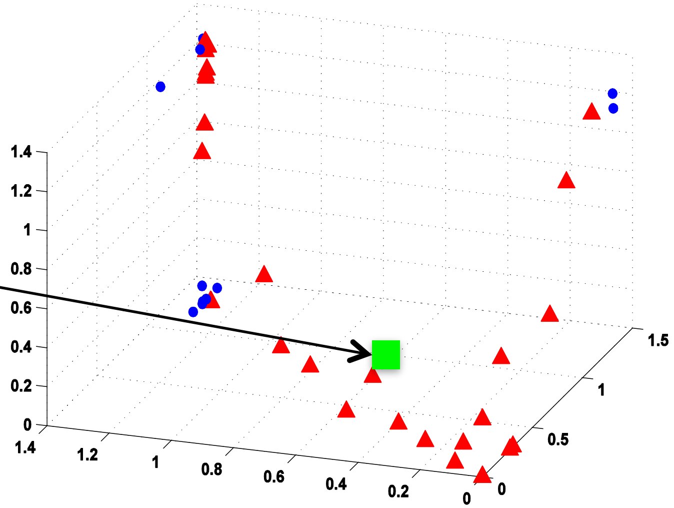
HR duty cycle/0.3
+ RR-min/5 <= 1

Annotation with Active Learning

We applied the active learning procedure to artifact annotation.

Annotation **without** Informative Projections

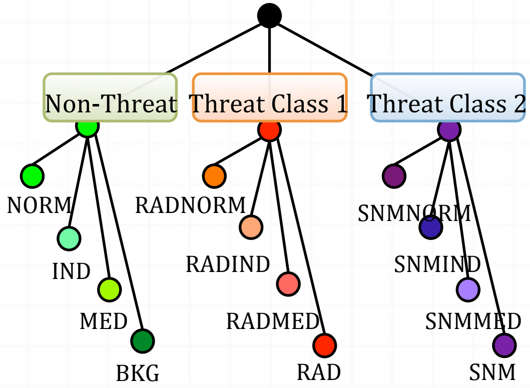
Annotation **with** Informative Projections



SPO$_2$ alerts: RIPR achieves max. accuracy with 25% of the data.
BP alerts: RIPR achieves max. accuracy with 50% of the data.

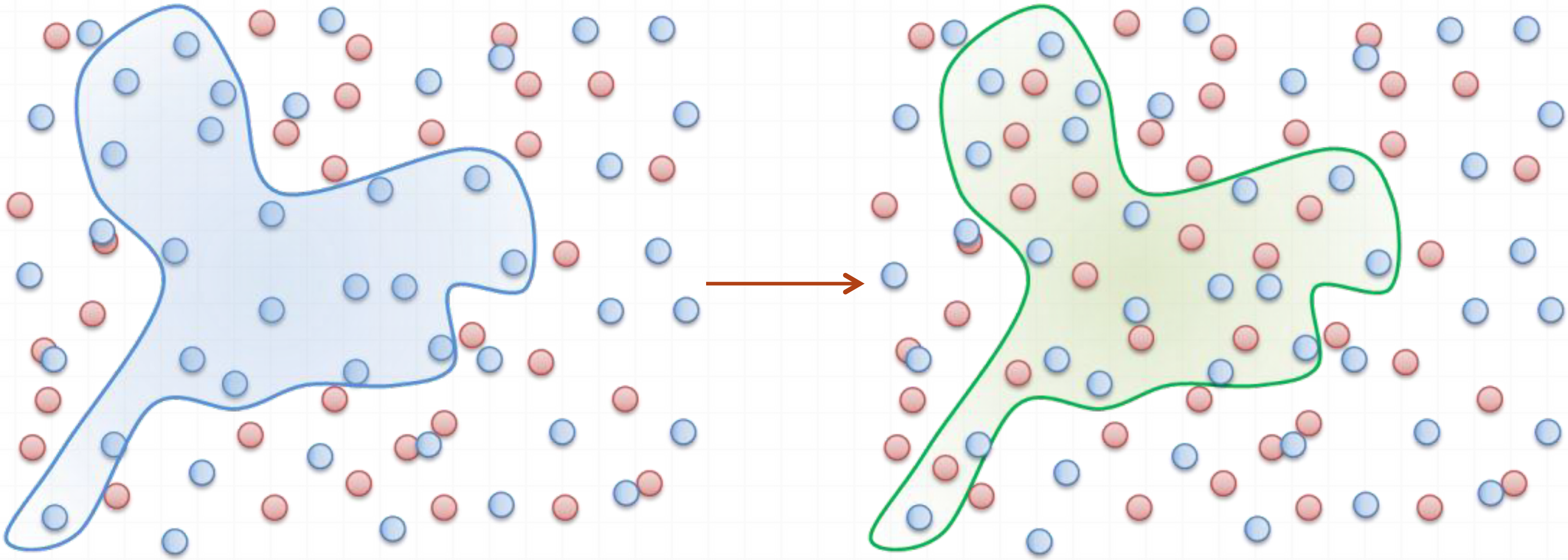
Nuclear Threat Detection

- Vehicles scanned at US border
- Radiation measurements
- Classify threat posed by vehicle
- Threats are rare in practice
- Training ‘threats’ are simulated
- We trained 2-D classification models



Identifying Gaps in Datasets

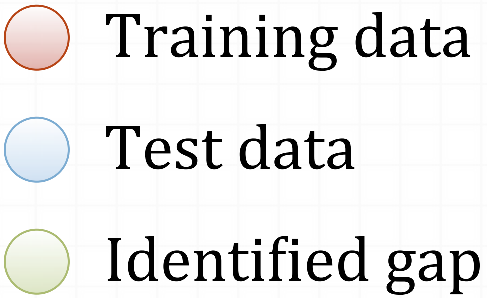
Joint work with Nick Gisolfi (ngisolfi@andrew.cmu.edu)



Training data is incomplete

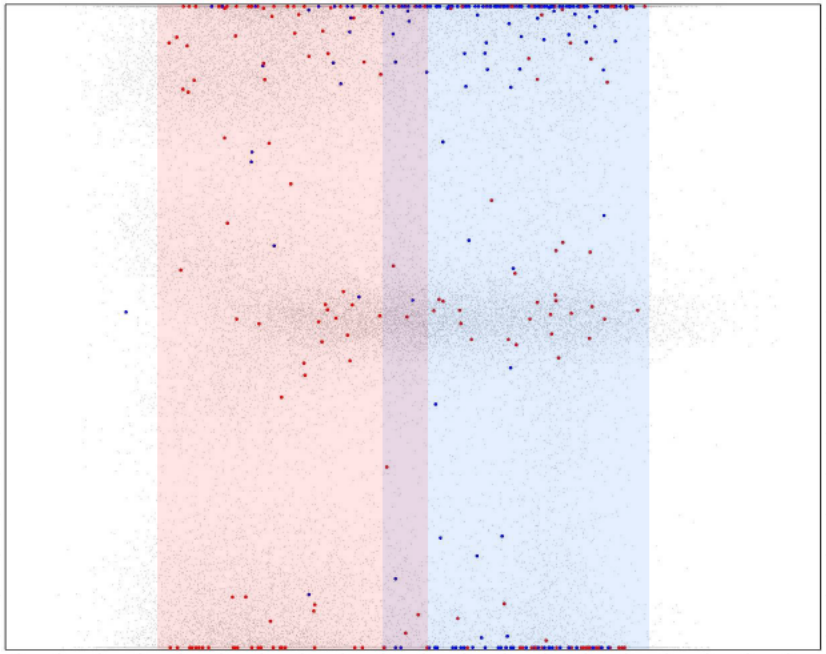
Additional samples requested

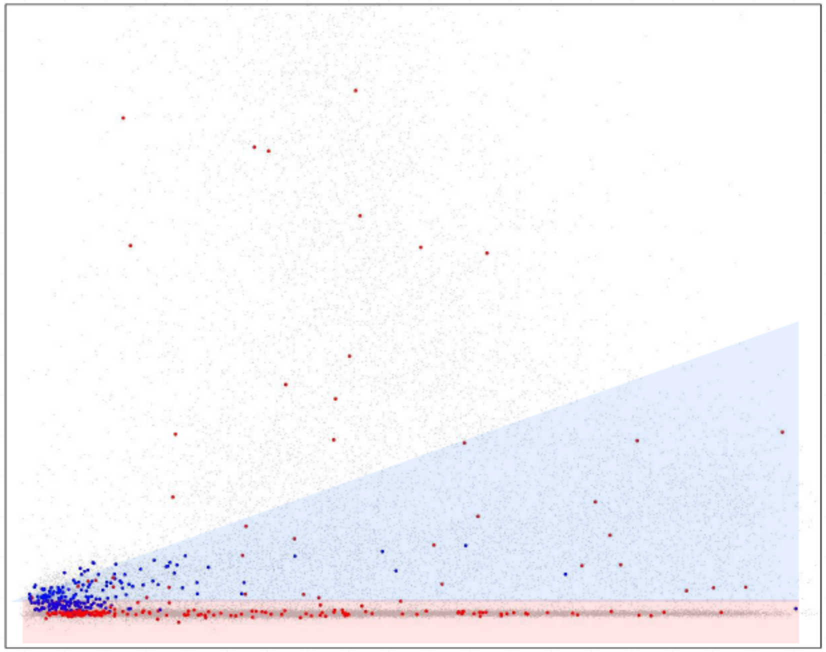
RIPR can express training data gaps in terms of low-dimensional projections.



Gap-Finding Examples

DIRECT GAP-FINDING: finding mismatches between distributions of training and testing data.

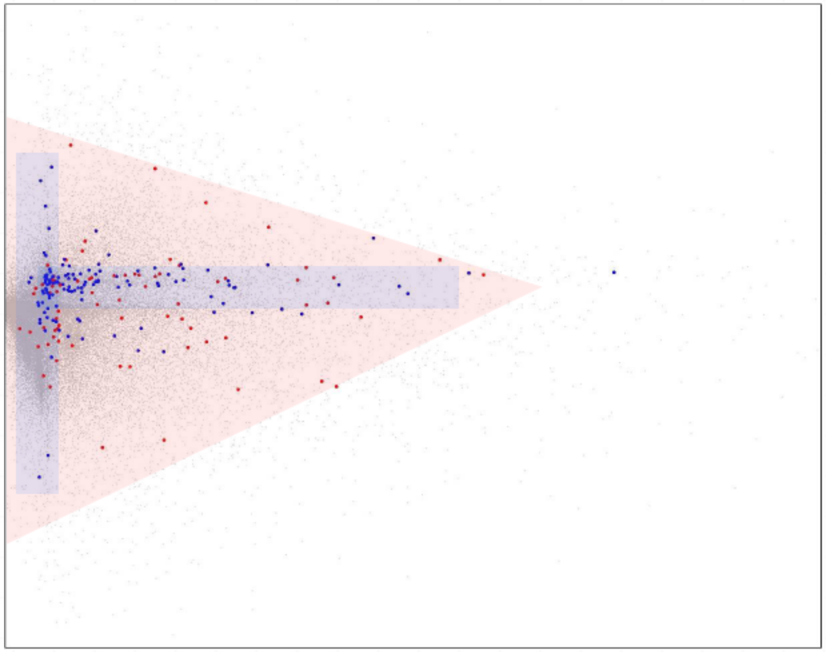


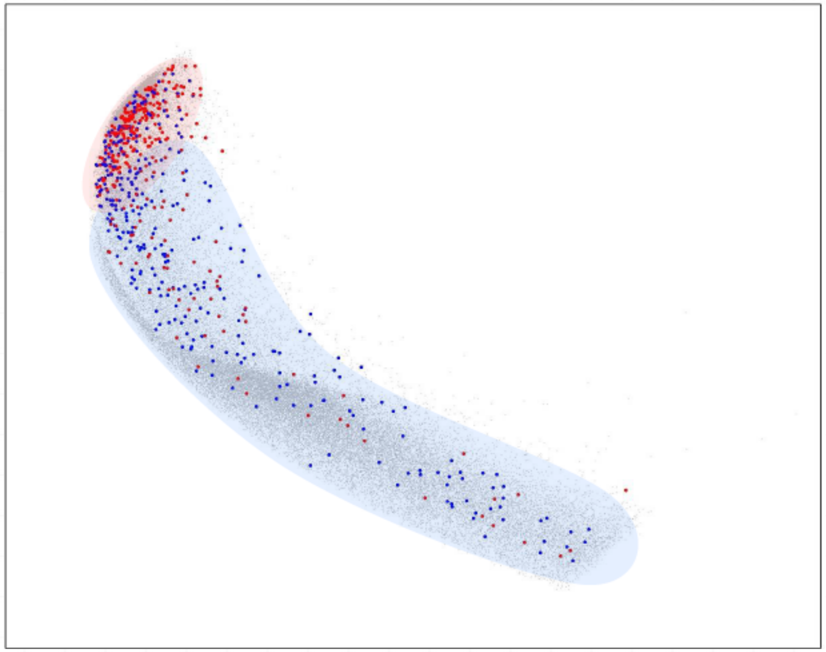


Gaps found in nuclear threat data.

Gap-Finding Examples

DIAGNOSTIC GAP-FINDING: finding areas of the test data where the classifier behaves poorly.





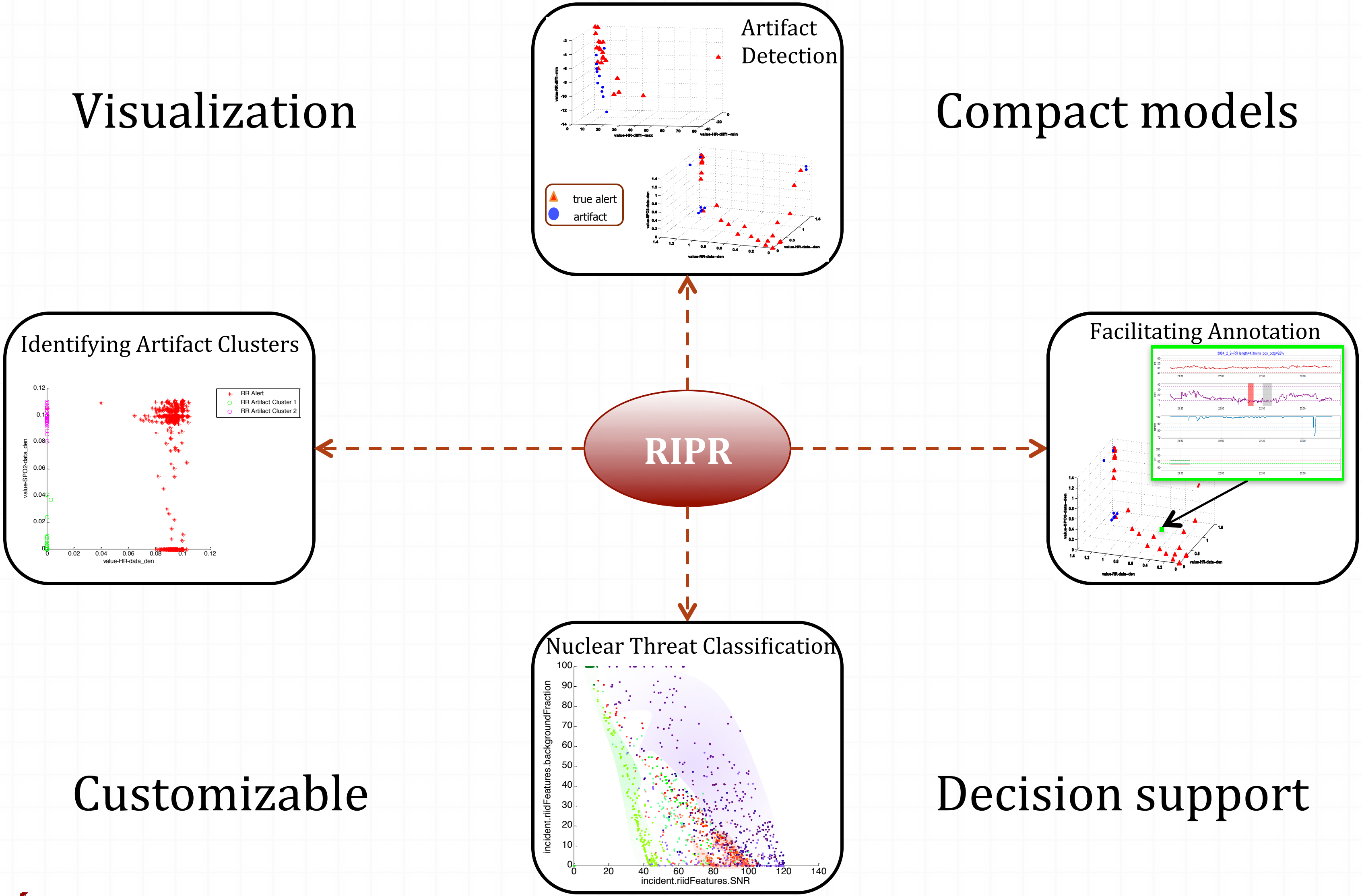
Accuracy improves from 75% to 75.7% by filling the gap compared to 75.2% by randomly adding data.

Conclusions

Projects in Chronological Order

- Using Dynamic Bayes Nets for Online Vital Sign Monitoring
- Using MRFs to obtain Elevation Map of Lunar Surface from LRO LIDAR and LCROSS imagery
- Explanation-Oriented Classification via Subspace Partitioning
- ✓ Regression for Informative Projection Recovery
- ✓ Detecting Artifacts in Clinical Data via Projection Retrieval
- Feature Task Bi-clustering in Multitask Regression
- Sparsistent Additive Modeling in Multi-task Learning
- ✓ Finding Gaps in Training Data to Guide Development of a Radiation Threat Detection System
- ✓ Interpretable Active Learning in Support of Data Annotation
- Improving prediction Across Related Datasets

Informative Projections



Thanks!

Collaborators:

- Artur Dubrawski, CMU, Auton Lab (advisor)
- Matt Barnes, CMU, RI and Auton Lab
- Karen Chen, CMU, Auton Lab
- Gilles Clermont, University of Pittsburgh
- Nick Gisolfi, CMU, Robotics
- Mathieu Guillaume-Bert, CMU, Auton Lab
- Peter Kontschieder, MSR Cambridge
- Marilyn Hravnak, University of Pittsburgh
- Michael R. Pinsky, University of Pittsburgh
- Donghan Wang, CMU, Auton Lab

References

[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