Graph-Based Anomaly Detection with Soft Harmonic Functions

Michal Valko

Advisor: Milos Hauskrecht

Computer Science Department, University of Pittsburgh, Computer Science Day 2011, March 18th, 2011.

Anomaly (Outlier) Detection

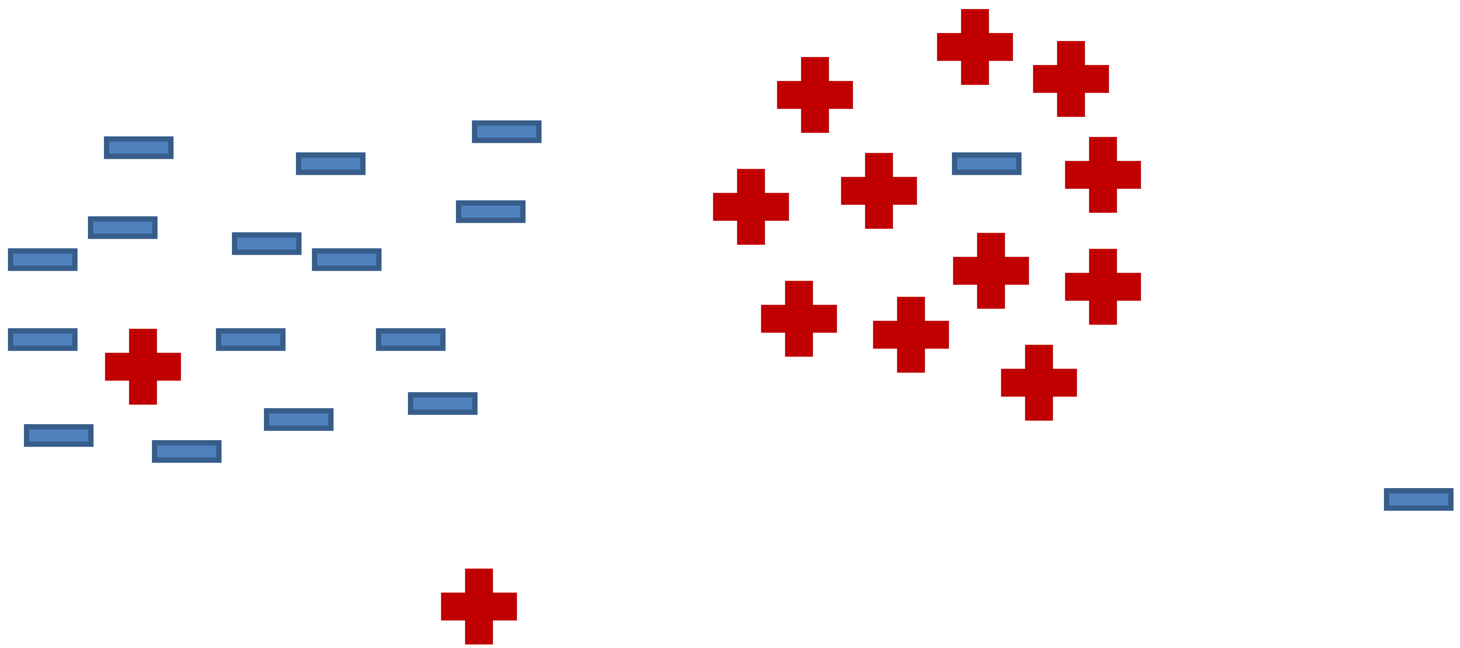
- Goal: identify unusual patterns in data

- Focus: <u>conditional</u> anomalies

- Contribution: graph-based method for conditional anomaly detection

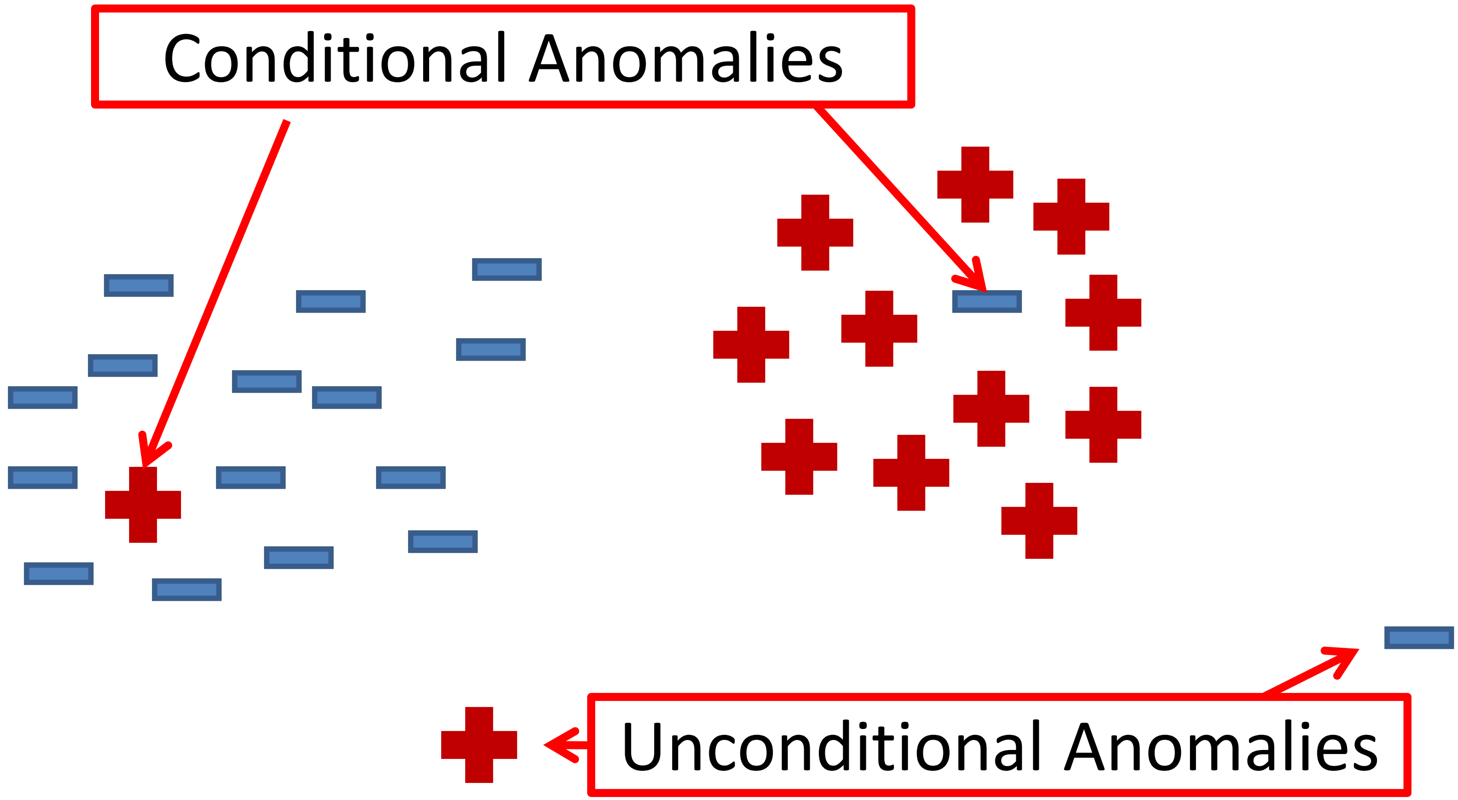
- Application: medical error detection

Conditional Anomaly



- Patient electronic records have: demographics, conditions, labs, medications administered, procedures performed,…

Conditional Anomaly



Assumption: *Conditional* anomalies correspond to medical errors

"Medical errors account for 200 000 preventable deaths a year."

(HealthGrades study, Wall Street Journal, July 27th 2004)

"Controlling overspending becoming necessary" Dr. Reinhardt

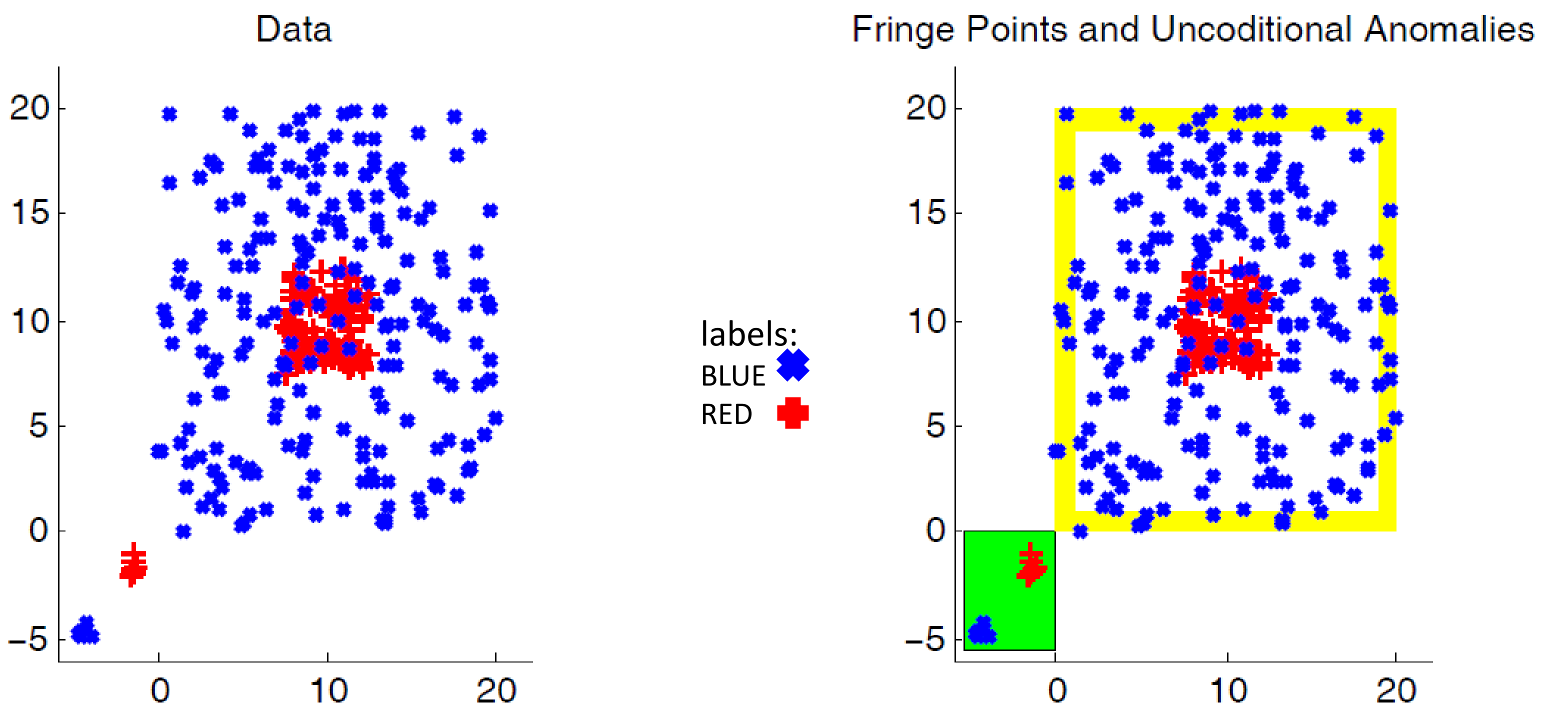
(Health Care Costs, New York Times, December 19th 2010)

Traditional Anomaly Detection

- Nearest Neighbor
  - Distance – anomalies are distant (NN)
  - Density – anomalies in low density regions (LOF, COF, LOCI)
- Classification
  - Model based (separate models for (ab)normal distributions)
  - 1-class (1-class SVM)
  - Classify normal vs. abnormal (when labels available)
- Statistical
  - > 3std

Challenges for CAD

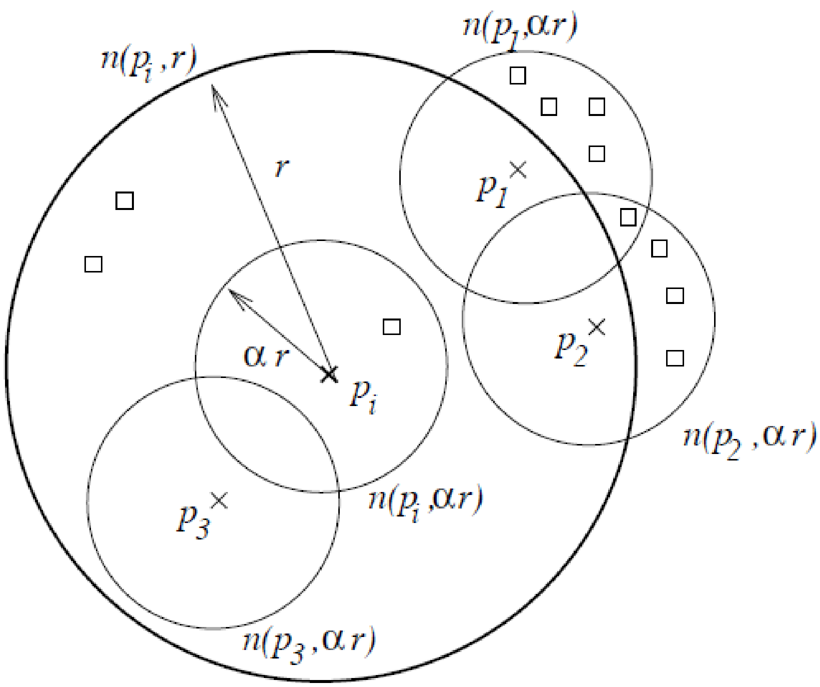
Task: detect anomalies in labels

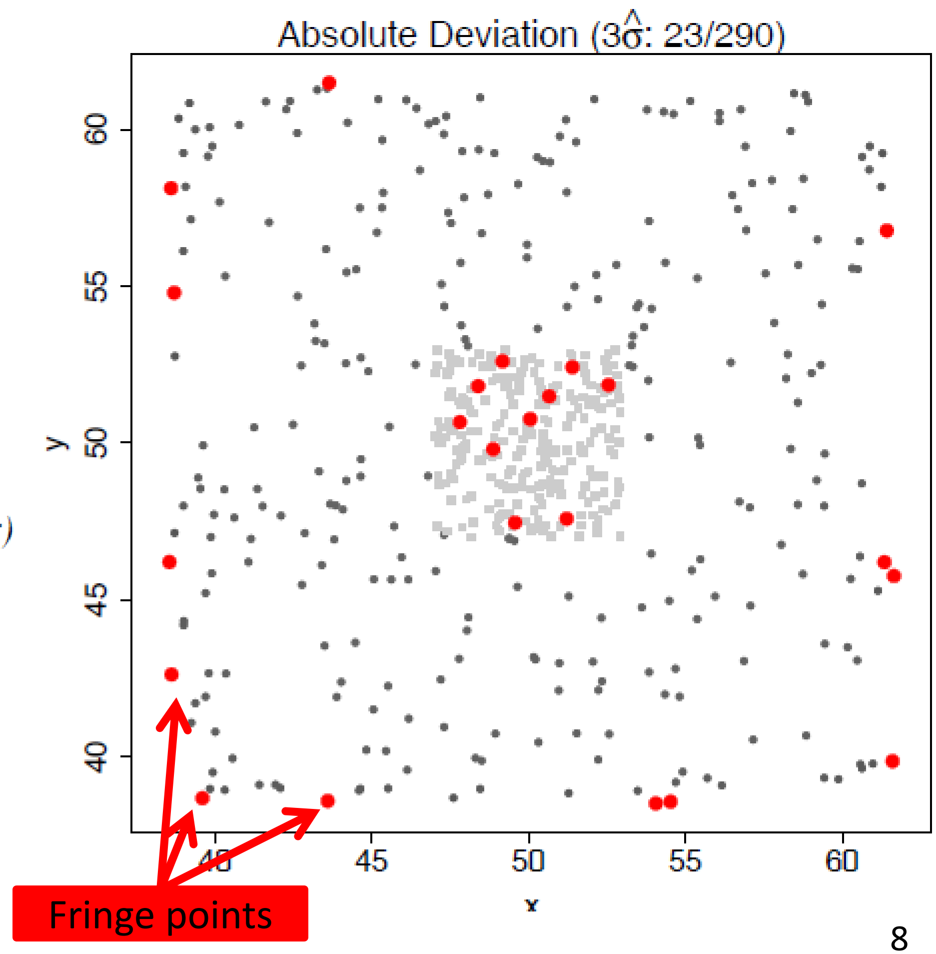


- Dataset adopted from [Papadimitriou and Faloutsos, 2003]

Related Work (CAD)

- Cross Outlier Detection (Papadimitriou, 2003)





CAD approaches

Class Outlier Approach

- OneClass SVM, LOF, …

Discriminative Approach

- SVM-CAD

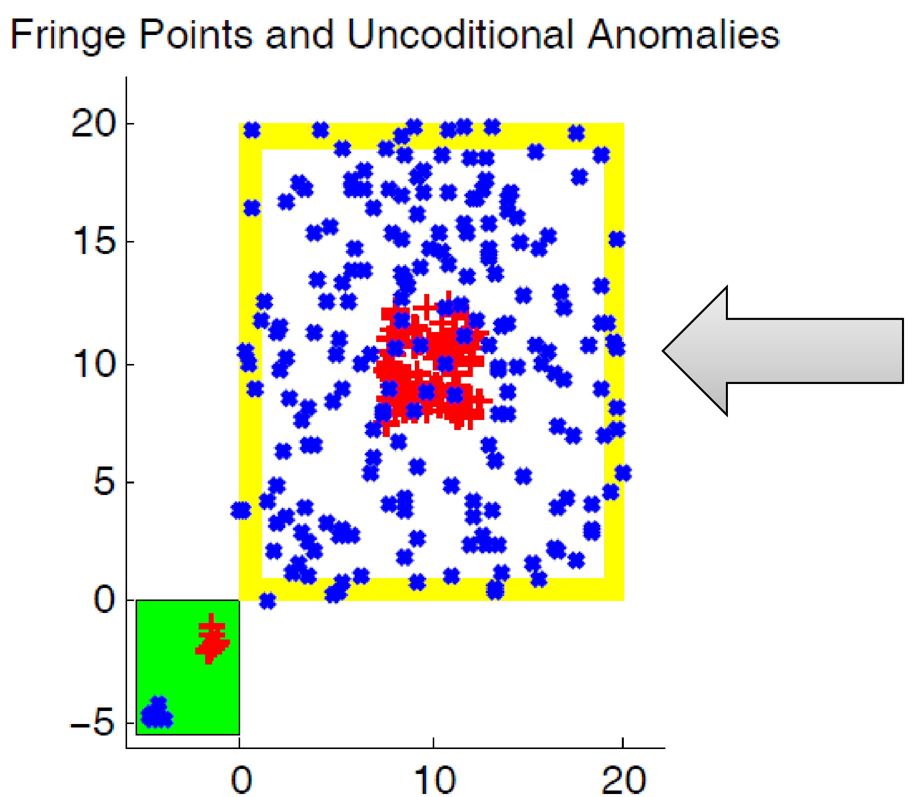
Regularized Discriminative Approach

- Connectivity AD, **Soft Harmonic AD**

regularizing unconditional outliers

Conditional Anomaly Detection Goal

Problem statement (★): For a dataset $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n}$ find pairs of $(\mathbf{x}_i, \mathbf{y}_i)$ such that $P(\mathbf{y} \neq \mathbf{y}_i | \mathbf{x}_i)$ is high.



... and at the same time avoid unwanted anomalies

both train and test data are labeled

Class Outlier Approach

- Take a test case (**x**,y)
- Take any unconditional anomaly method
- Find out if **x** is anomalous wrt { x | x has class y }
- **Problems:**
  - Fringe points
  - Unconditional outliers
  - Anomaly (alert) scores for class 1 and class 2 may not be comparable

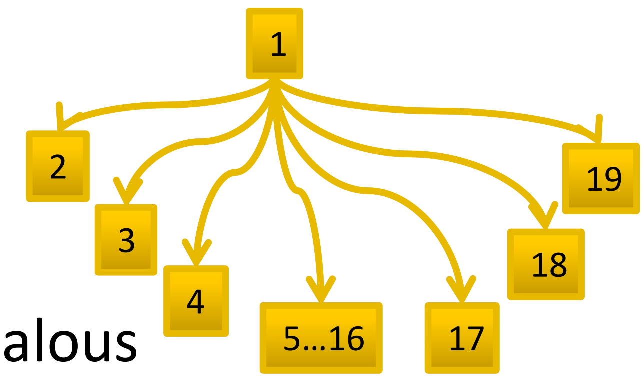
ignores the other class(es)

Discriminative Approach

- $P(y'|\mathbf{x})$ is high → conditional anomaly
- Learn Model/Build Projections
- Bayes Network

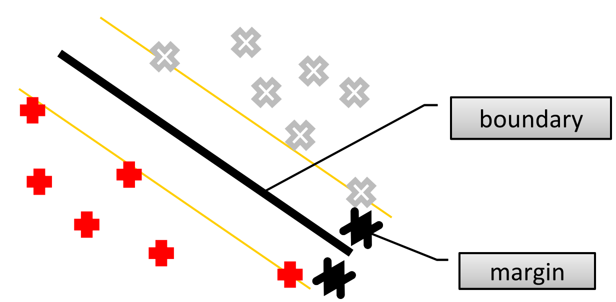
$$d(y|\mathbf{x}) = P(y'|\mathbf{x}) \quad y' \neq y$$

  - bigger the alert score→ more anomalous



- Support Vector Machines projections

$$d(y|\mathbf{x}) = -y(\mathbf{w}^{\top}\mathbf{x} + w_0)$$

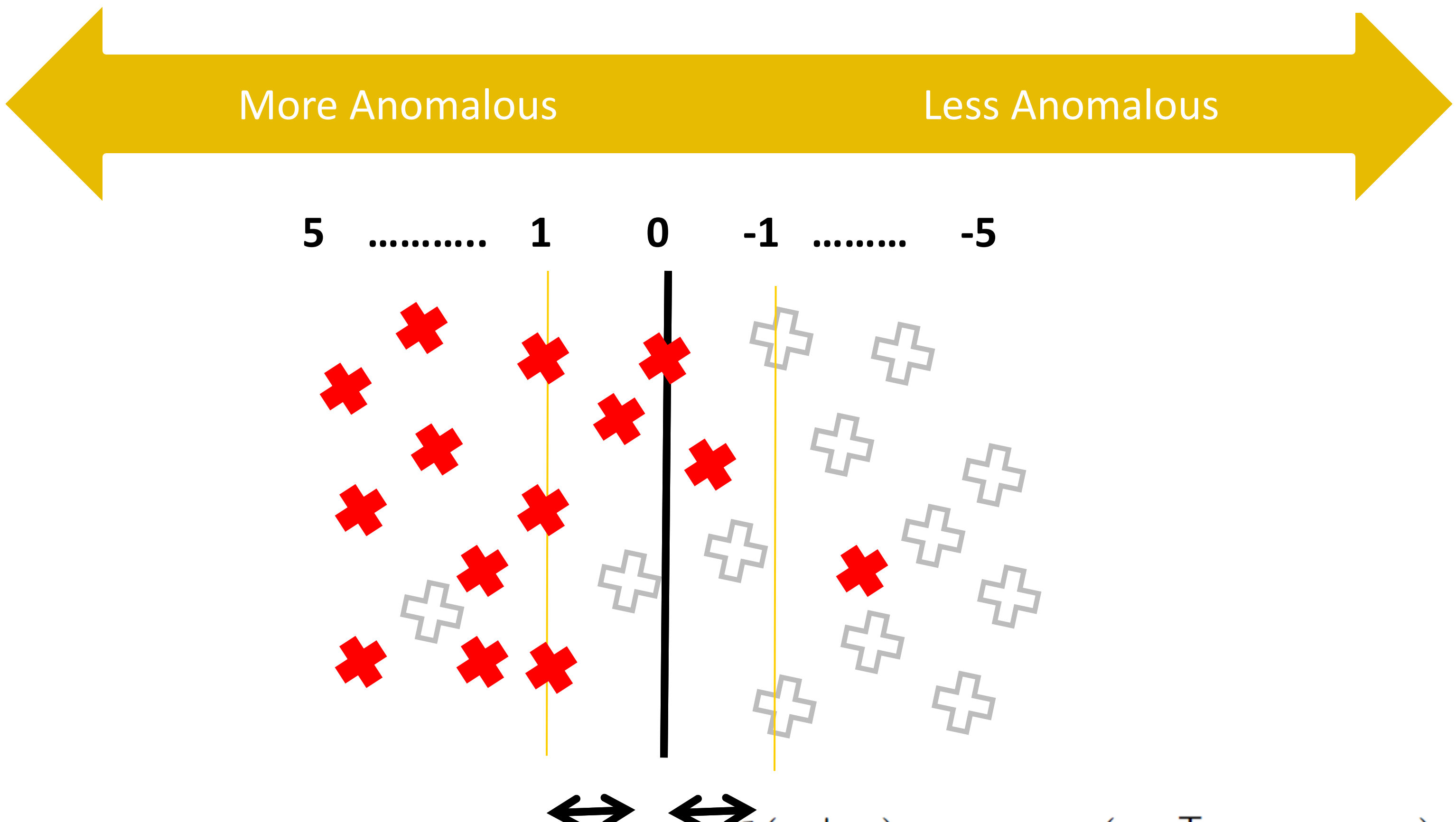


Support Vector Machines projections

[Valko et al., 2008]

[Valko and Hauskrecht, 2008]

[Hauskrecht et al., 2010]



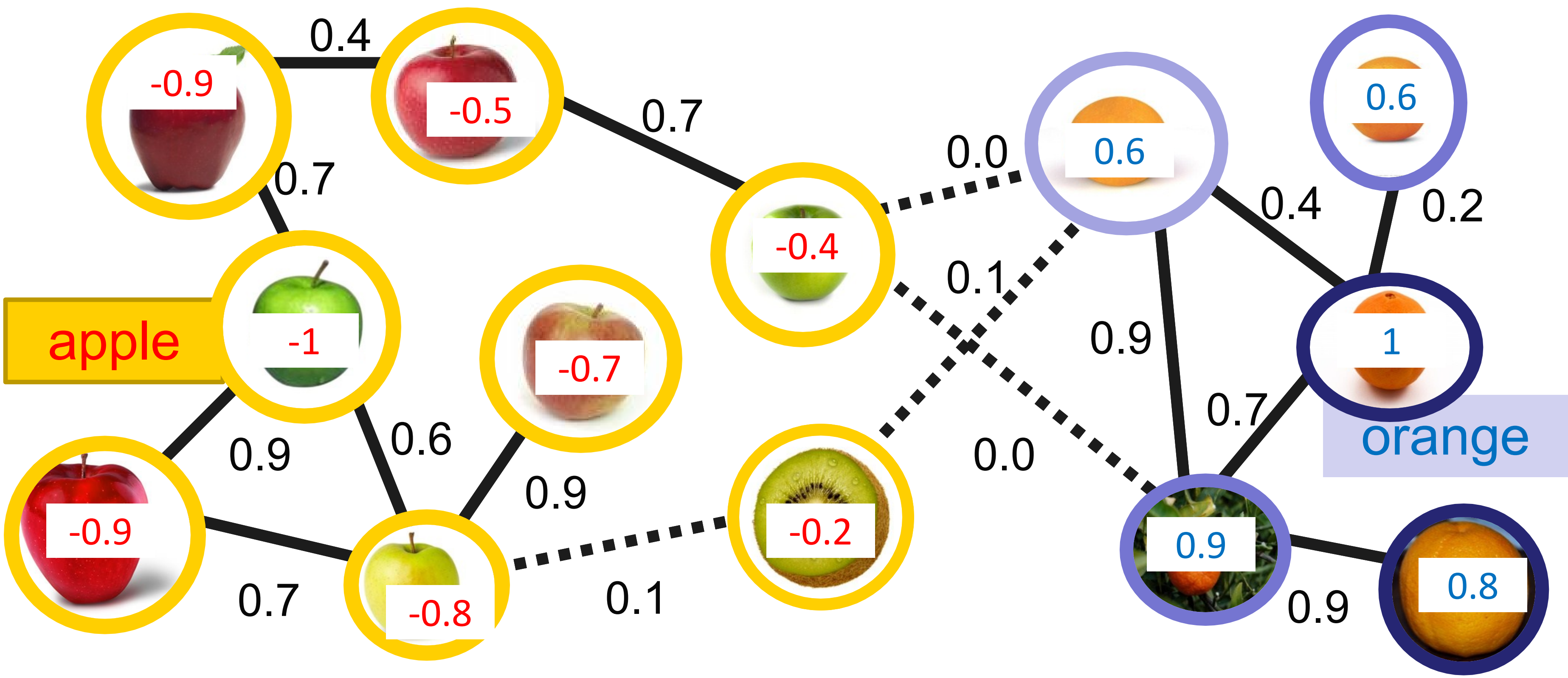
$$d(y|\mathbf{x}) = -y(\mathbf{w}^\top \mathbf{x} + w_0)$$

A new approach

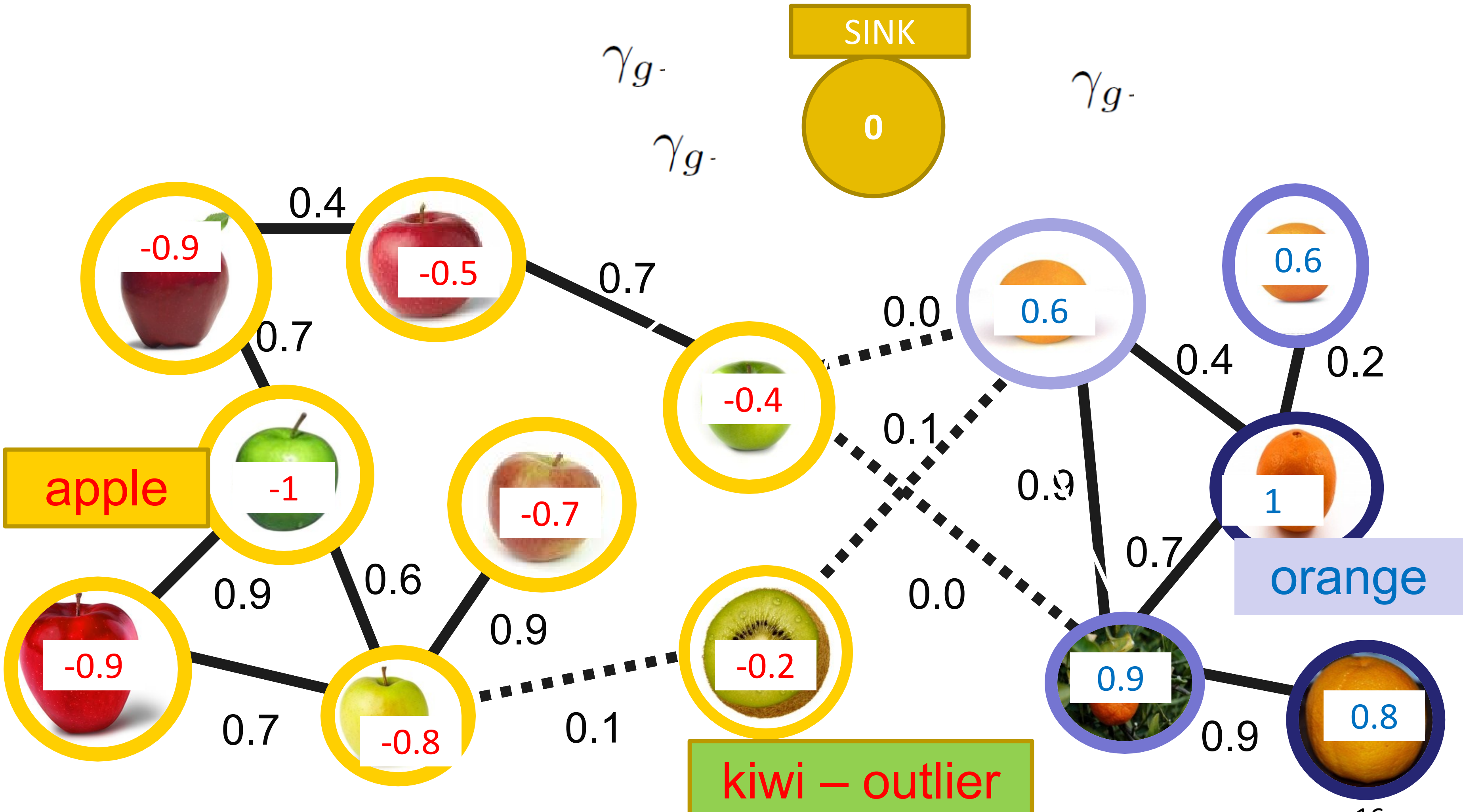
- Disadvantages of the SVM-CAD
  - only linear decision boundary
  - can become overly confident in the areas with little data
    - Isolated points (unconditional outliers)
- Soft Harmonic Anomaly Detection
  - Non-parametric
  - Graph-based
  - Regularization
    - Control the influence of unconditional outliers
  - Can incorporate unlabeled examples
    - Missing medical records
    - Tests not done frequently because of the budget constraints

Harmonic Solution

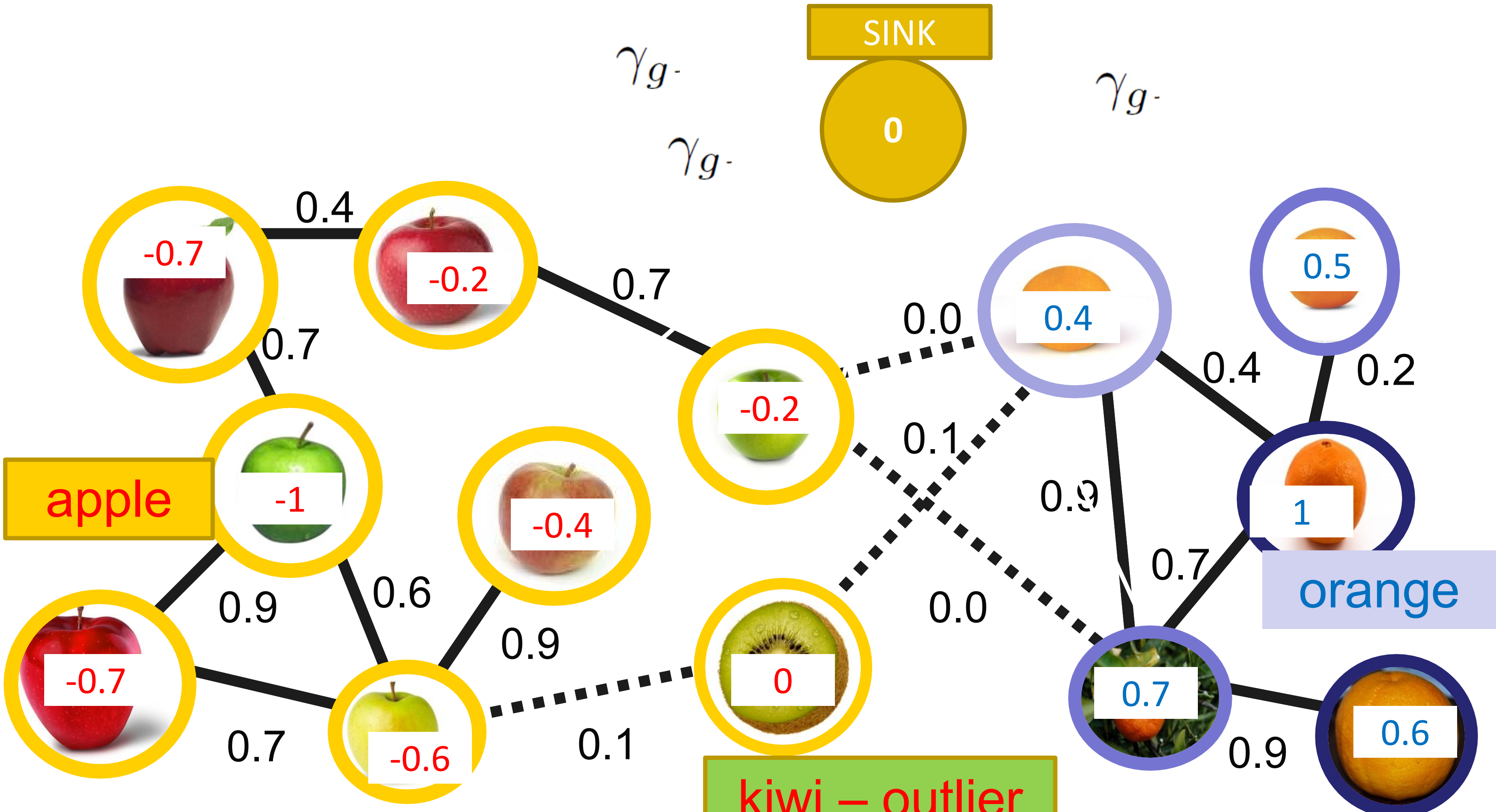
[Zhu et al., 2003]



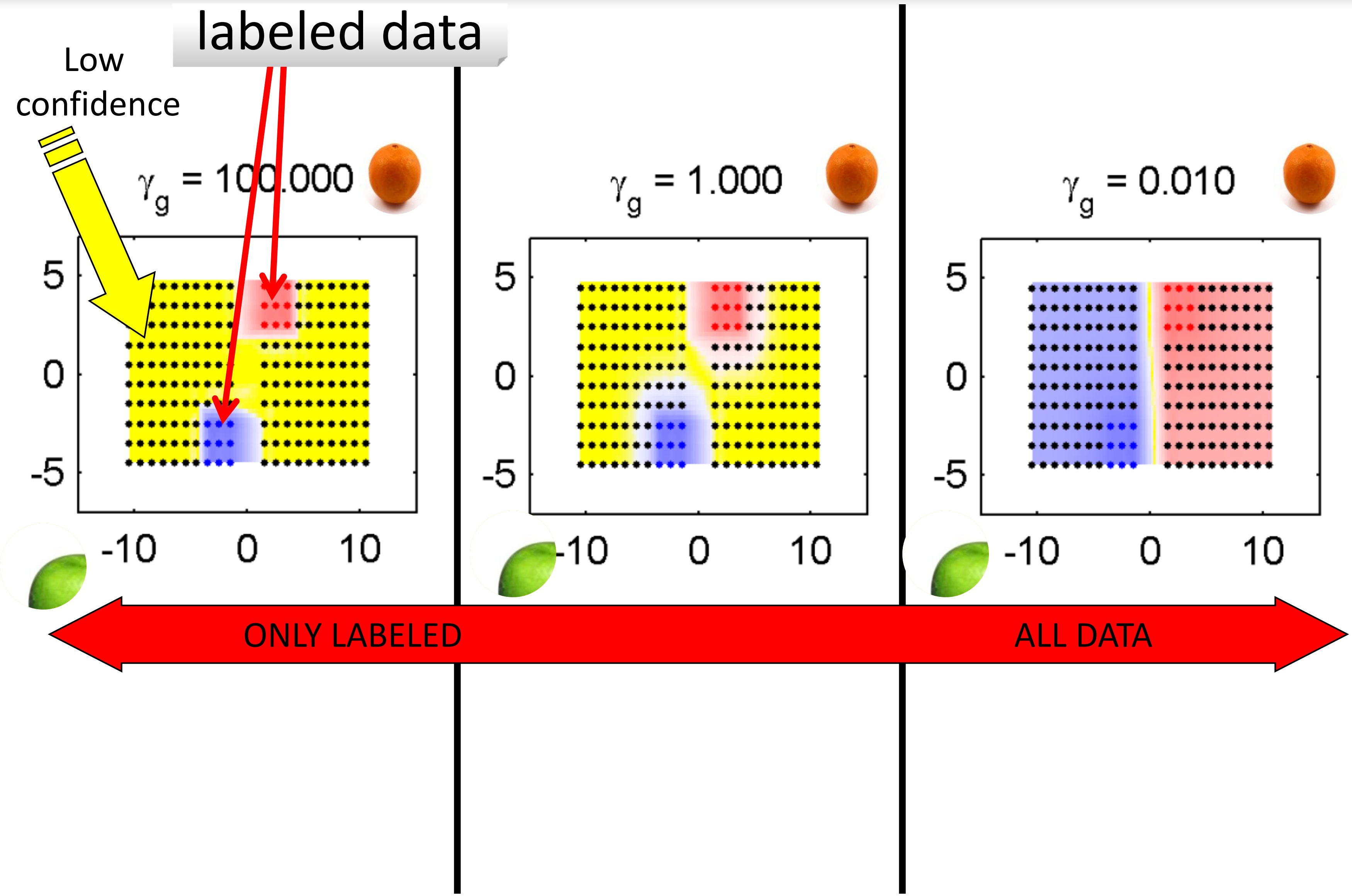
Dealing with Outliers



Dealing with Outliers



Regularization



Soft Harmonic Solution

- Unconstrained Regularization

$$\min_{\ell \in \mathbb{R}^n} (\ell - \mathbf{y})^{\mathsf{T}} C (\ell - \mathbf{y}) + \ell^{\mathsf{T}} K \ell$$

fit to data

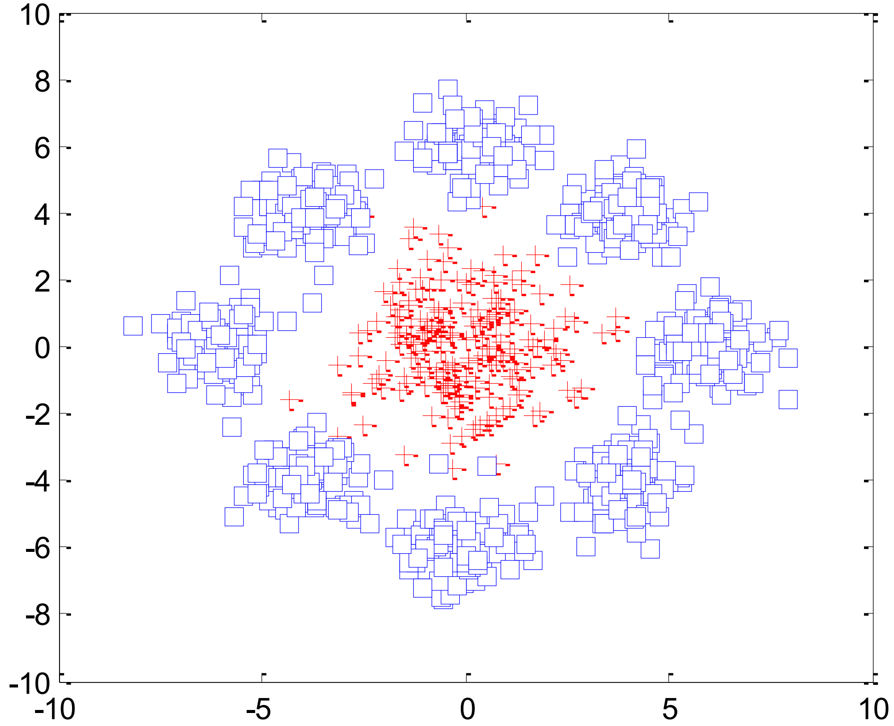
regularizer

- Close form solution $\ell = (C^{-1}K + I)^{-1}\mathbf{y}$

- when ℓ_i is rewritten as $|\ell_i| \operatorname{sgn}(\ell_i)$
- $|\ell_i|$ can be interpreted as a confidence
- $|\ell_i|$ >> 0.5 and $\operatorname{sgn}(\ell_i) \neq y_i$ Conditional Anomaly!

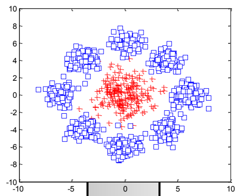
Synthetic Data

- evaluation of conditional anomaly methods is challenging
- synthetic data with known distribution
- flip 3% of the labels
- compare how the anomaly score agrees with true score



Synthetic Data: Results

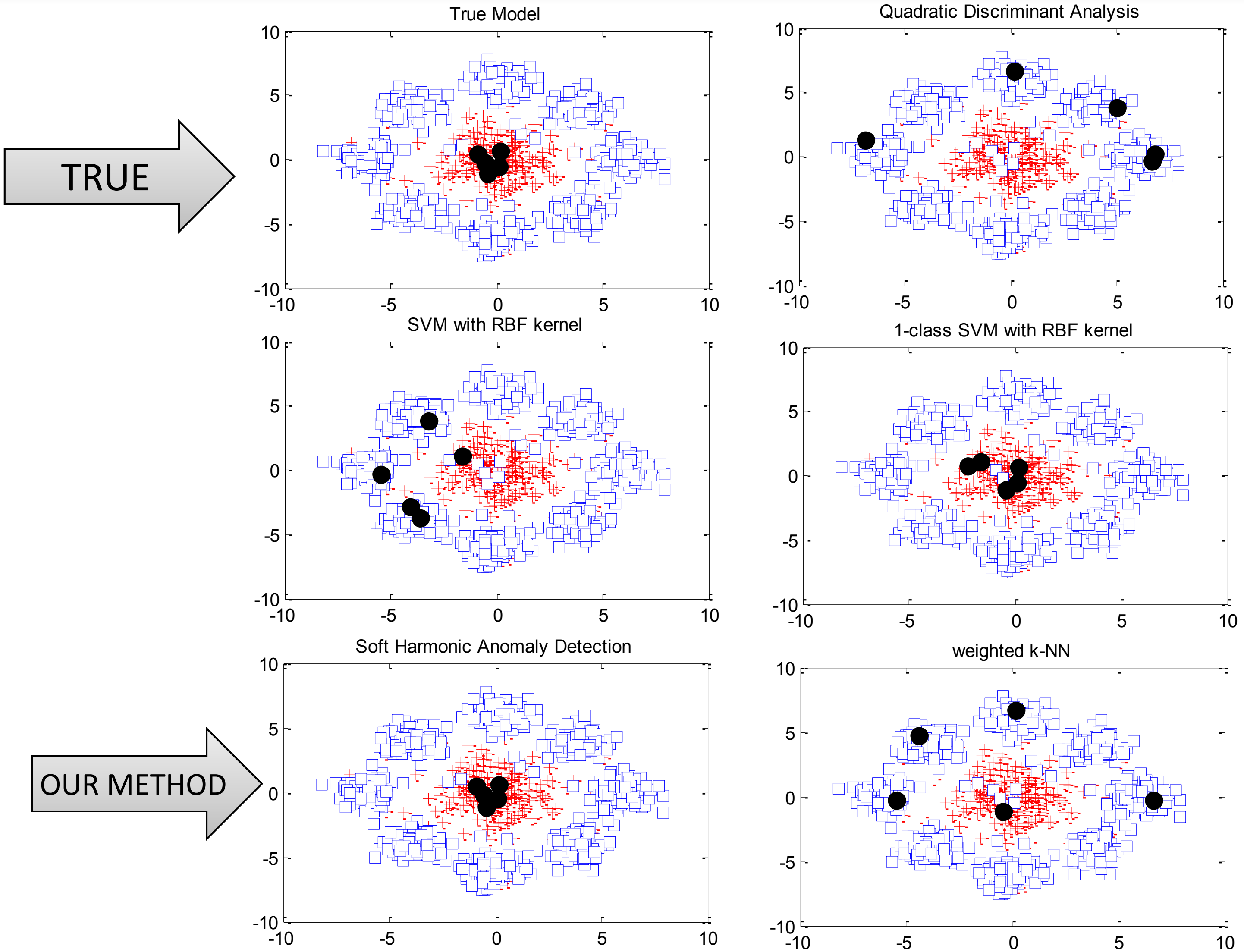
- Evaluation metric:
  - How the anomaly score agrees with the true score



	Dataset **D1**	Dataset **D2**	Dataset **D3**
QDA	73.8% (2.1)	29.4% (5.2)	61.0% (1.2)
SVM	58.8% (7.0)	49.8% (1.7)	46.1% (3.1)
1-class SVM	51.3% (0.9)	47.7% (0.6)	64.7% (0.7)
wk–NN	74.2% (1.9)	56.5% (1.7)	61.4% (2.1)

BETTER

Top 5 best scoring anomalies for different methods on the synthetic dataset D3



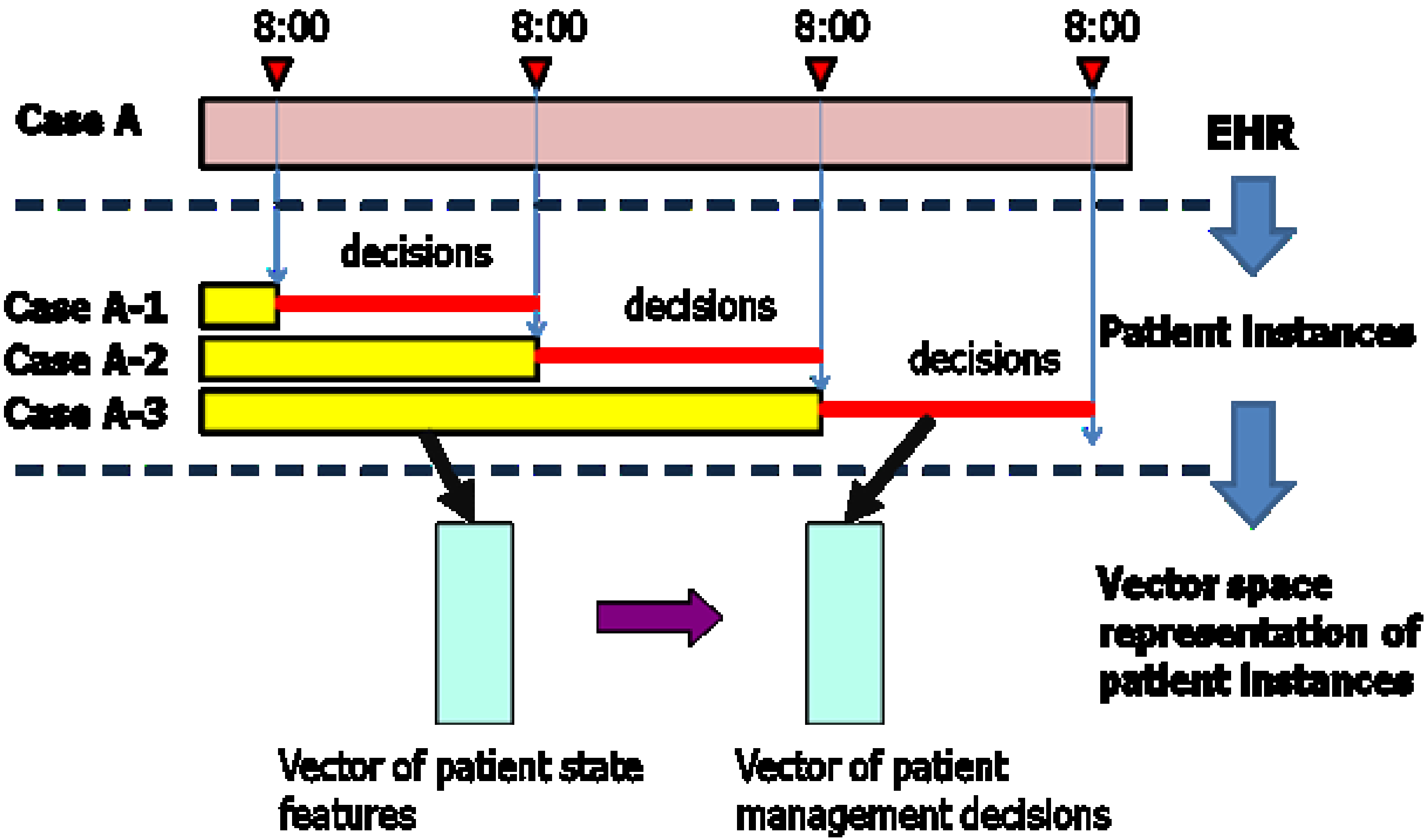
Medical Data

- 4486 patients from UPMC
- Cardiac surgery (2002-2007)
- 45767 patient-day events/states
- 9K attributes
- 222 states evaluated by 15 experts

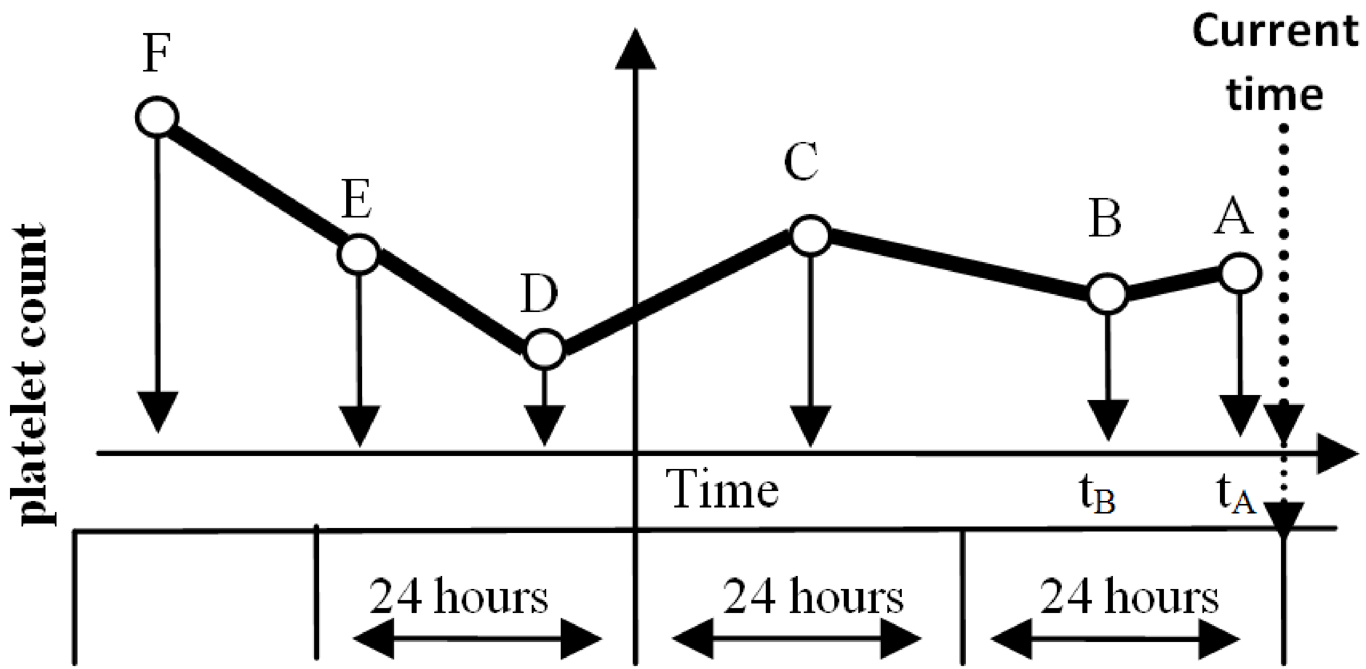
1. Laboratory tests (LABs)
2. Medications (MEDs)
3. Visit features/demographics
4. Procedures
5. Heart support devices

- nearest neighbor graph
- Metric: How much the score agrees with the experts.

PCP data set: Segmentation



PCP Dataset: PLT Lab feature



Last value: A

Last value difference = B-A

Last percentage change = (B-A)/B

Last slope = (B-A) / (tB-tA)

Nadir = D

Nadir difference = A-D

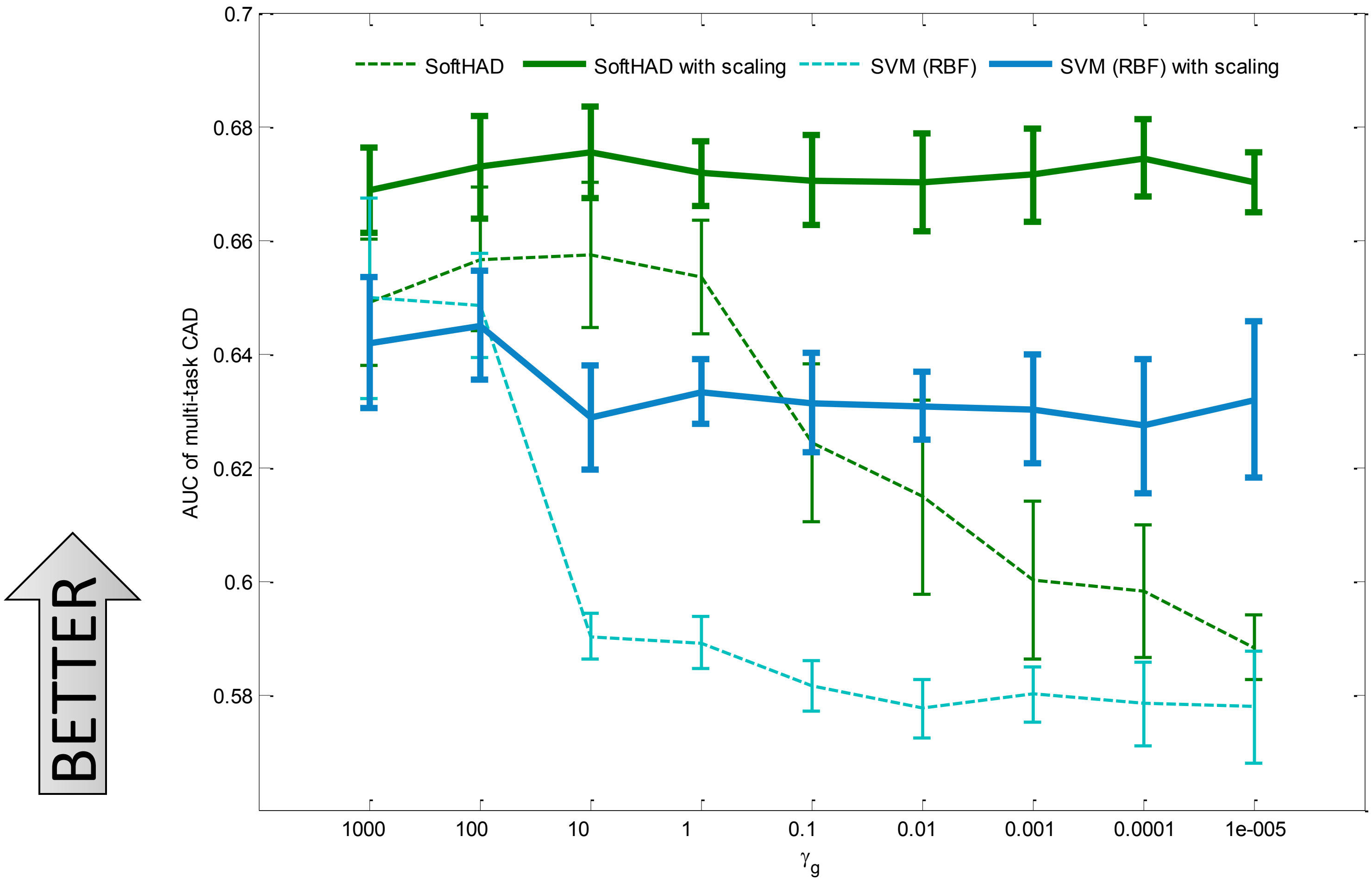
Nadir percentage difference = (A-D)/D

Baseline = F

Drop from baseline = F-A

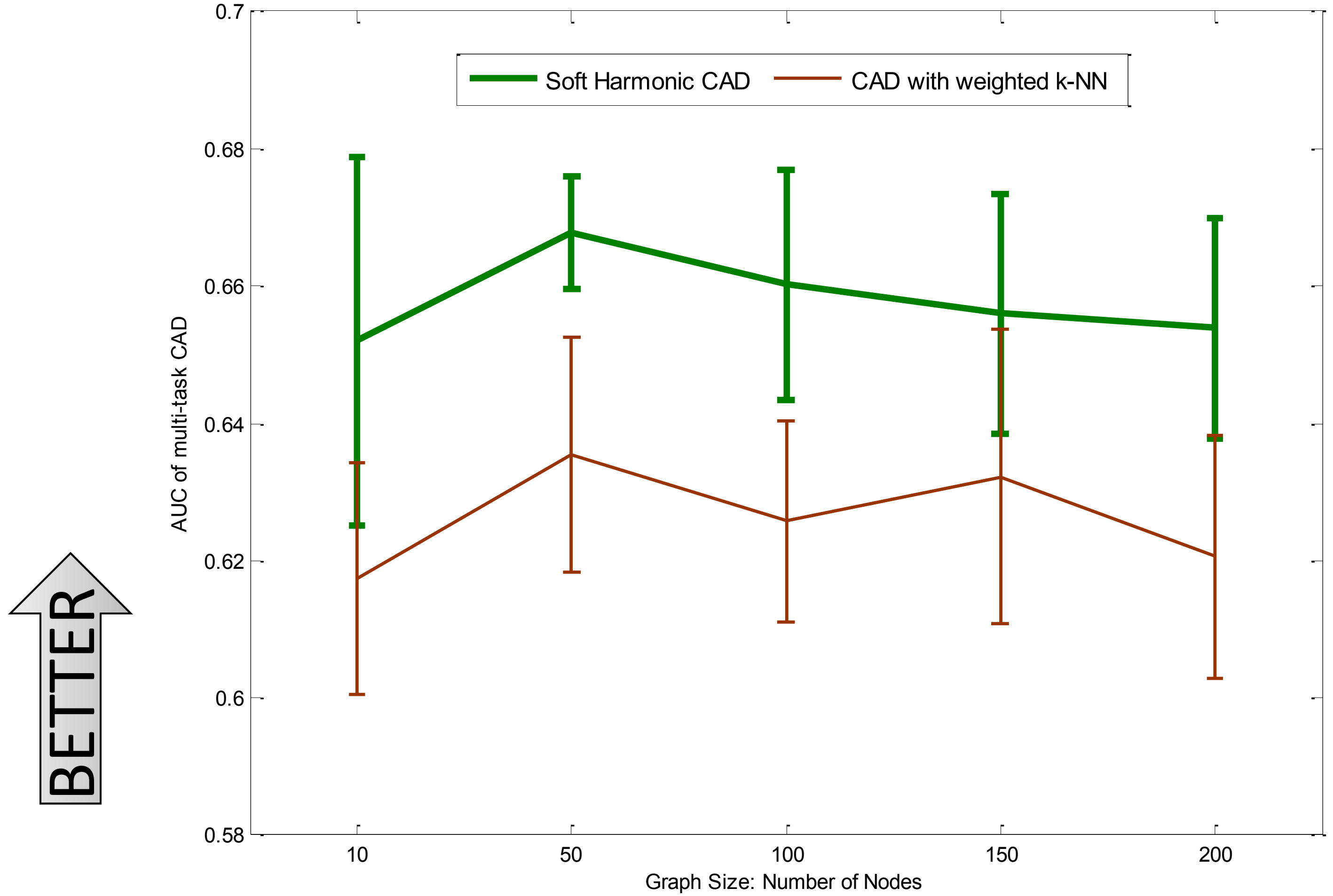
Medical Data Results

- Outperforming SVM method over the range of settings of regularization parameters



Medical Data Results

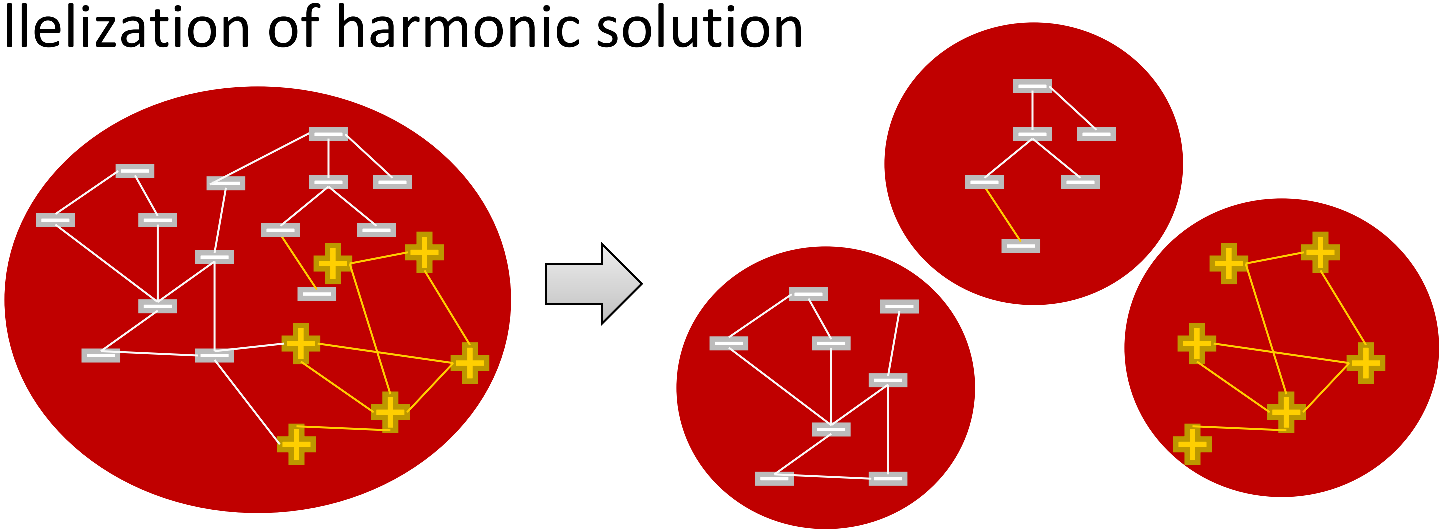
- Outperforming standard weighted nearest neighbors on the same graph



Conclusion & Future Work

- A non-parametric graph-based approach
  - Successfully detect conditional anomalies
- Future work
  - Online Soft Harmonic Anomaly Detection
  - Parallelization of harmonic solution

Adapt to changes in medical practice



Come to my poster

Thanks to:
Branislav Kveton, Greg Cooper, Tomas Singliar, Shyam Visweswaram, Iyad Batal, Amy Seybert, Hamed Valizadegan, Saeed Amizadeh, Quang Nguyen, Dave Krebs