Adaptive Graph-Based Algorithms

for Conditional Anomaly Detection

and Semi-Supervised Learning

Michal Valko

Computer Science Department
University of Pittsburgh

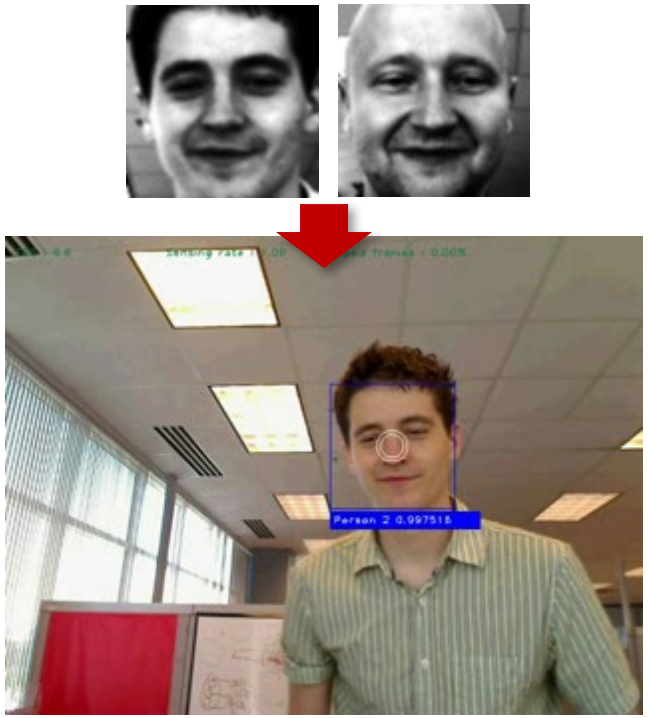
Advisor: **Milos Hauskrecht**

Committee: Liz Marai, Diane Litman, John Lafferty (CMU)

University of Pittsburgh, August 1st, 2011

My Research

- **My research**
  - Learning with minimal feedback
  - Online adaptation
- Example: **Online Face Recognition**
  - Changing light conditions
  - Different backgrounds
  - Faces change
  - Outliers
- **Approach**
  - Similarity graph as a representation
  - Data dependent, non-parametric



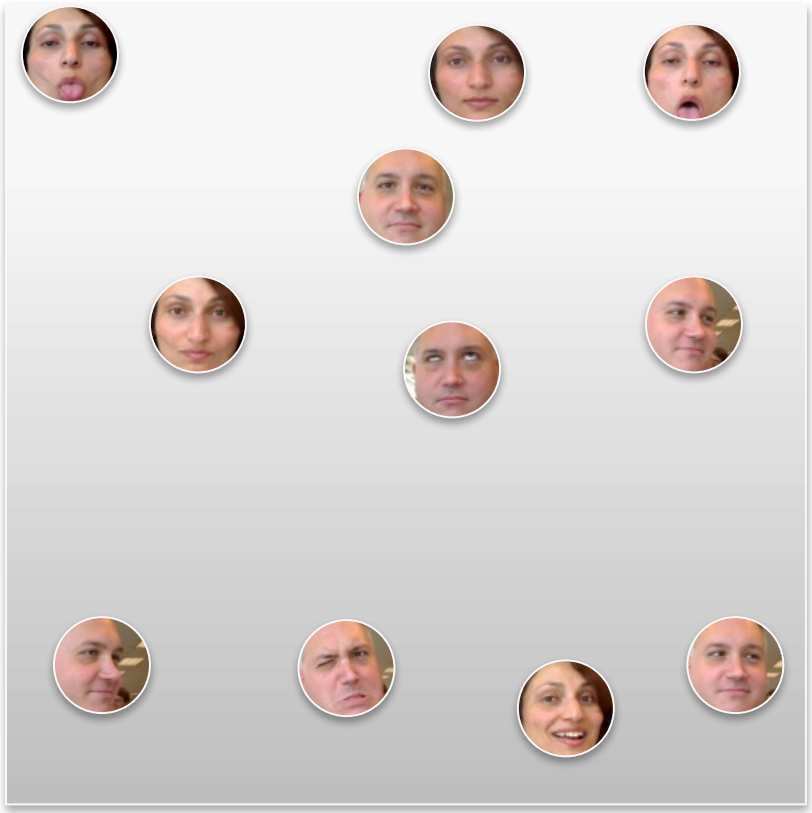
Outline

- Graph Based Learning
- Semi-Supervised Learning (SSL) on Graphs
- Online SSL with Quantized on Graphs
- Application: *Online Face Recognition*
- Conditional Anomaly Detection
- Application: *Medical Error Detection*

Graph Based Learning

- Data adjacency graph W is a graph that represents the structure of data
- The vertices of the graph are data points x_i
- The edges of the graph are weighted by the similarity of the vertices:

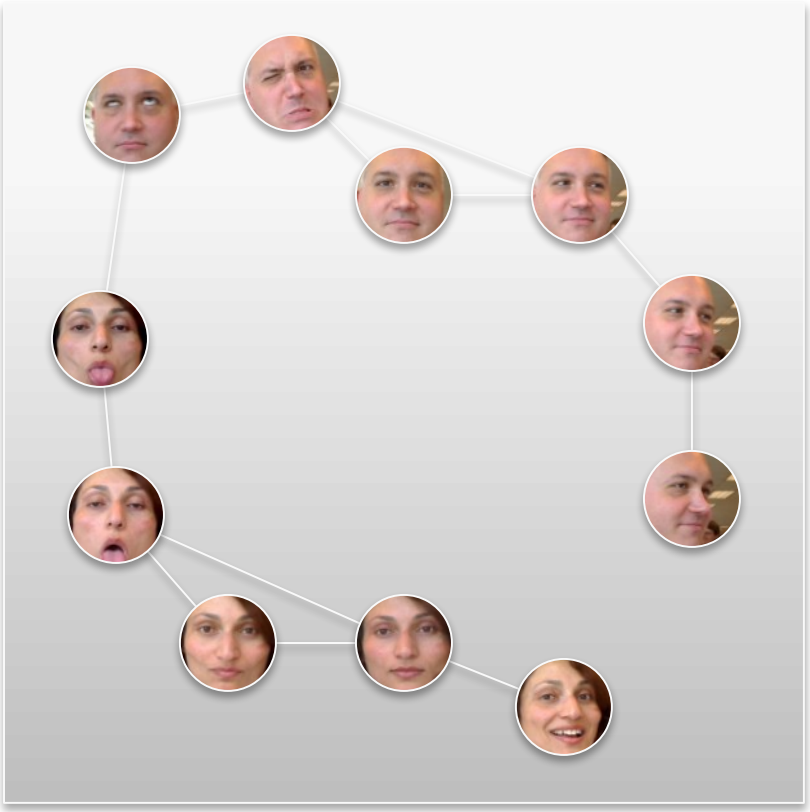
$$w_{ij} = \exp[-d^2(x_i, x_j)/(2\sigma^2)]$$



Graph Based Learning

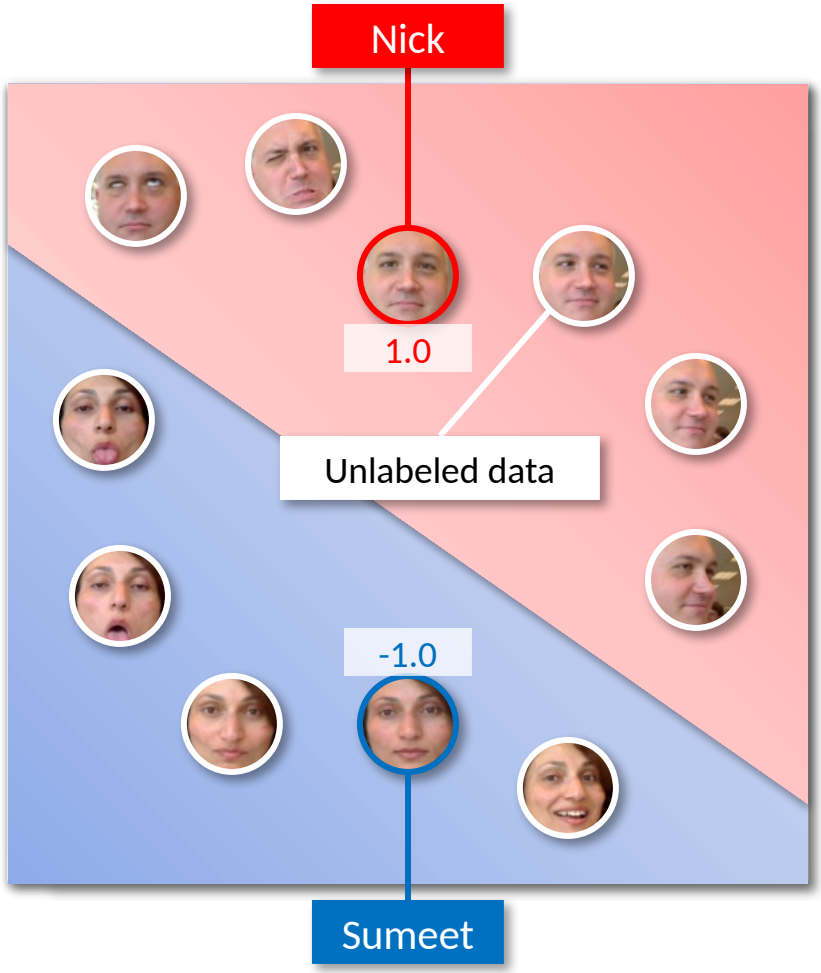
- Data adjacency graph W is a graph that represents the structure of data
- The vertices of the graph are data points x_i
- The edges of the graph are weighted by the similarity of the vertices:

$$w_{ij} = \exp[-d^2(x_i, x_j)/(2\sigma^2)]$$



Semi-Supervised Learning

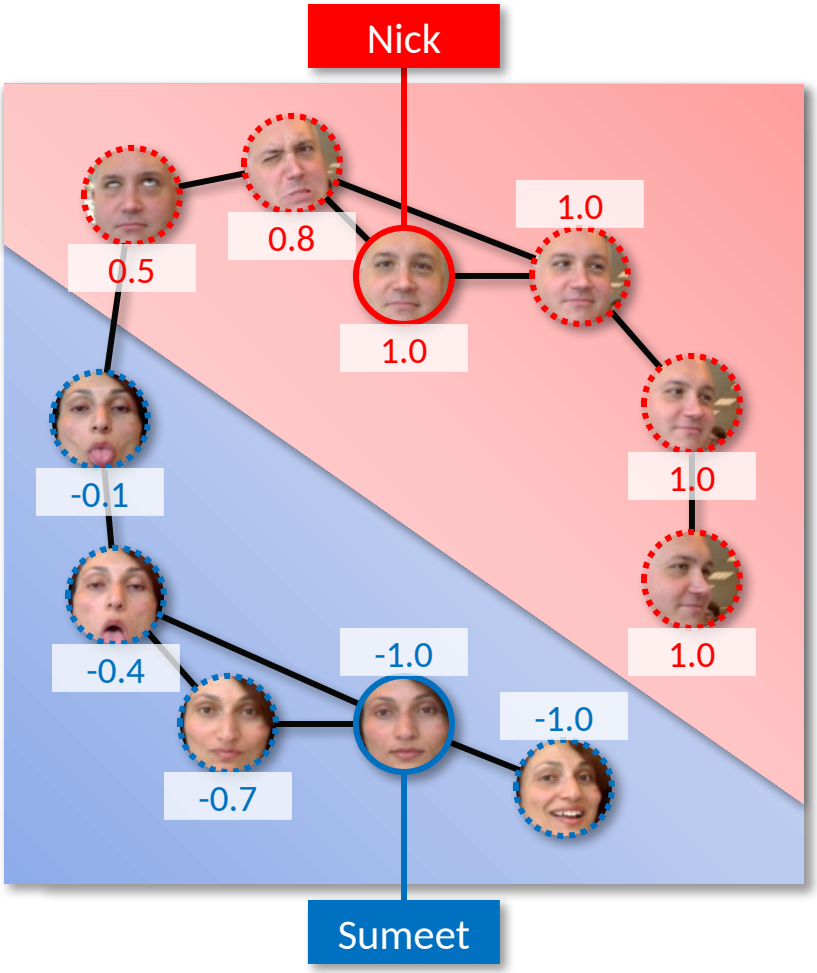
- Semi-supervised learning (SSL) is machine learning paradigm for learning from both labeled and unlabeled data
- If a vertex x_i is labeled, its label is $y_i \in \{1,-1\}$



Semi-Supervised Learning on Graphs

- The structure of the graph W is used to infer labels of unlabeled data $\ell_i \in [1, -1]$
- Unlabeled examples can be labeled based on a random walk on the graph:

$$\ell_i = P(\text{rw from i ends at 1}) - P(\text{rw from i ends at -1})$$



Harmonic Function Solution

- A standard approach to semi-supervised learning:

$$\min_{\ell \in \Re^n} \sum_{i,j} w_{ij} (\ell_i - \ell_j)^2 \quad \text{s.t. } \ell_i = y_i \text{ for all } i \in l$$

- Rewritten in terms of the graph Laplacian L = D – W:

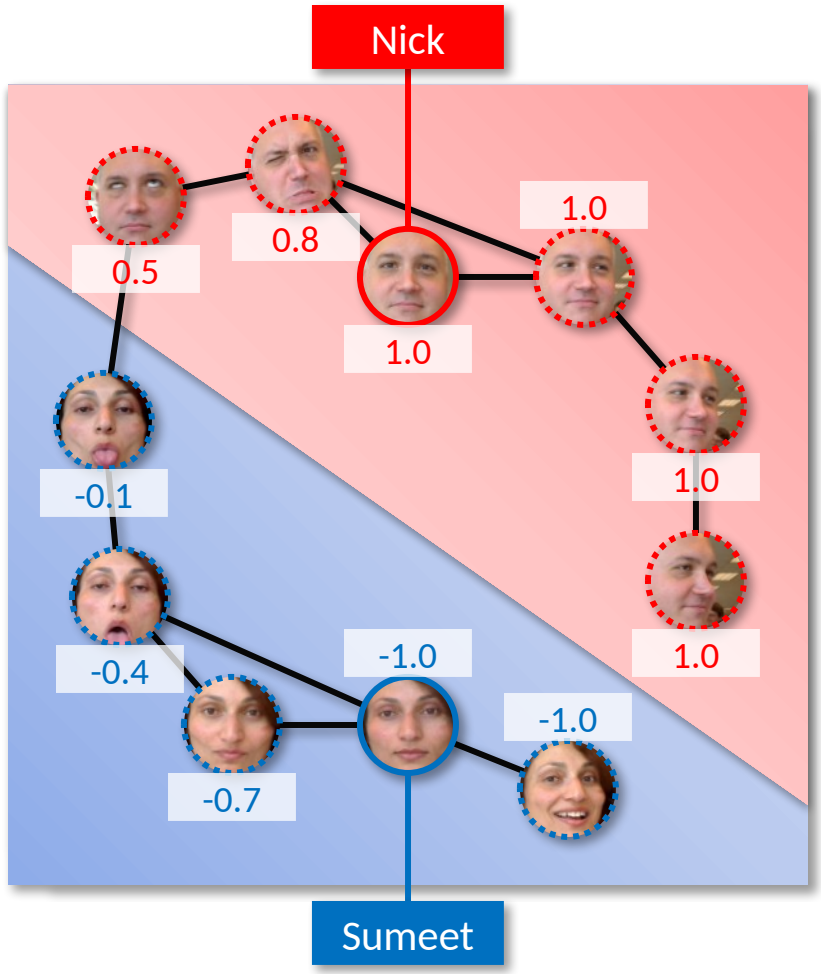
$$\min_{\ell \in \Re^n} \ell^{\mathrm{T}} L \ell \quad \text{s.t. } \ell_i = y_i \text{ for all } i \in l$$

- Properties of the harmonic function solution (HFS):
  - Smoothness $\ell_i = \frac{1}{d_i} \sum_{j \sim i} w_{ij} \ell_j$
  - Closed-form solution $\ell_u = (D_{uu} - W_{uu})^{-1} W_{ul} \ell_l$
  - Can be interpreted as a product of a random walk on the graph W with the transition matrix $P = D^{-1} W$

[Zhu et al., 2003]

Harmonic Function Solution

- Harmonic function solution is a result of a random walk on the graph W
- Advantages:
  - Tracks non-linear patterns
  - Convexity
  - Globally optimal
- Disadvantages:
  - Sensitive to the structure of the graph (problem-specific calibration is usually needed)



Online HFS

Inputs: an example x_t, a similarity graph W

What is wrong with this algorithm?

Algorithm:

Time complexity of $\theta(t)$

Add the example x_t to the graph W

Compute the Laplacian L of W and infer labels:

$$\min \ell^{\mathrm{T}} L \ell \quad \text{s.t.}\ \ell_i = y_i \ \text{for all}\ i \in l$$

Predict $\hat{y}_t = \mathrm{sgn}(\ell_t)$

Time complexity of $\Omega(t^2)$

Outputs: a prediction $\hat{y}_t$, an updated graph W

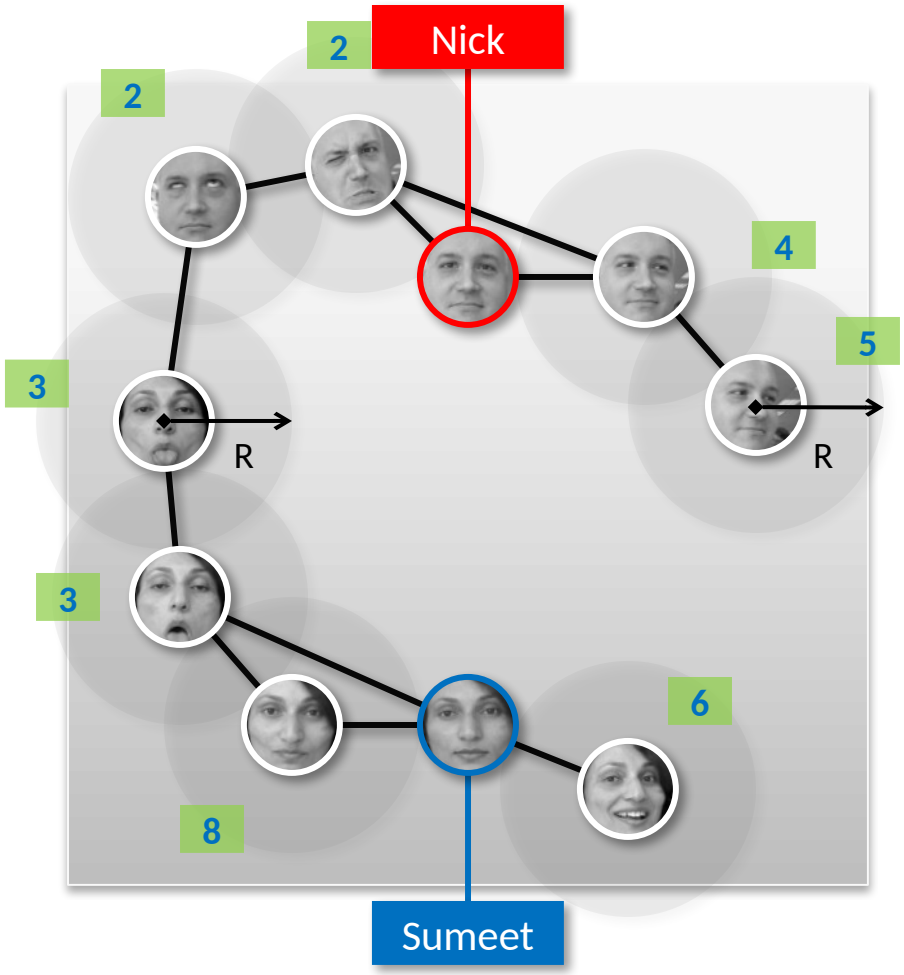
Reduce the number of nodes

Semi-Supervised Perception: Real-Time Learning without Explicit Feedback. In Proceedings of the 4th IEEE Online Learning for Computer Vision Workshop, San Francisco, California, June 2010. Best paper award.

Graph Quantization

- Our solution combines
  - Online Clustering
  - Semi-Supervised Inference

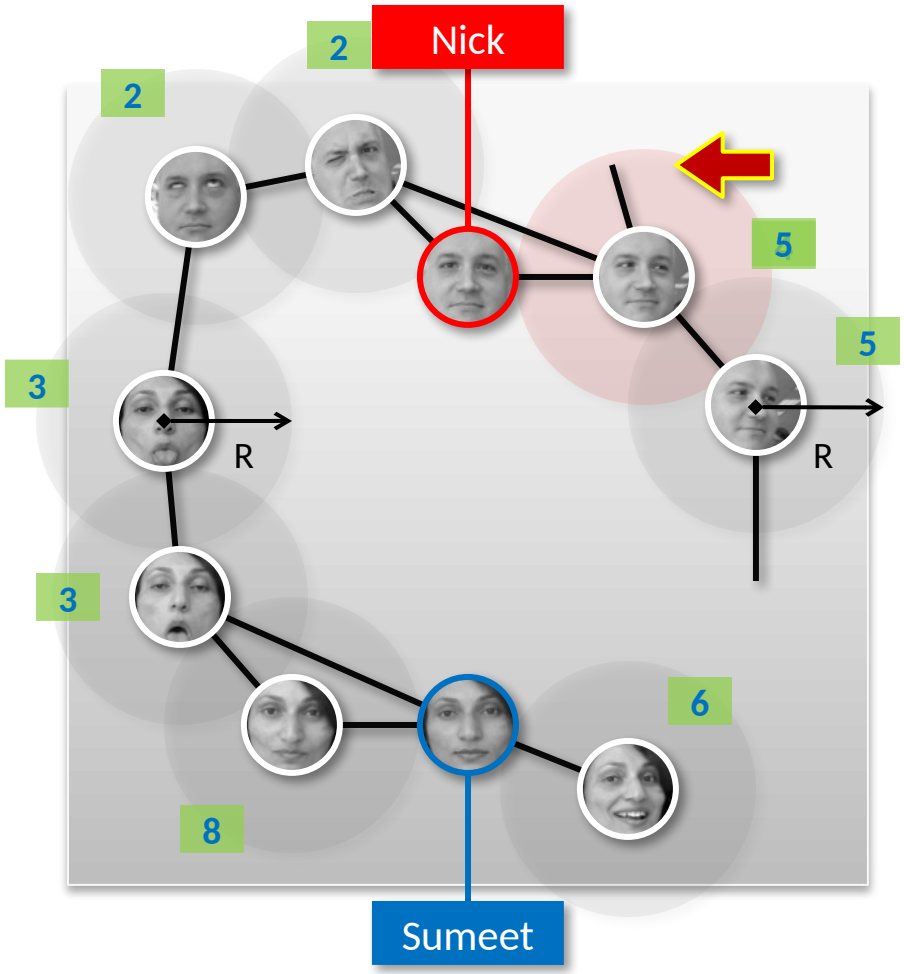
- Online clustering based on k-center clustering algorithm of Charikar *et al.* (1997) incrementally covers data by a set of R-balls



Online Semi-Supervised Learning on Quantized Graphs. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, July 2010.

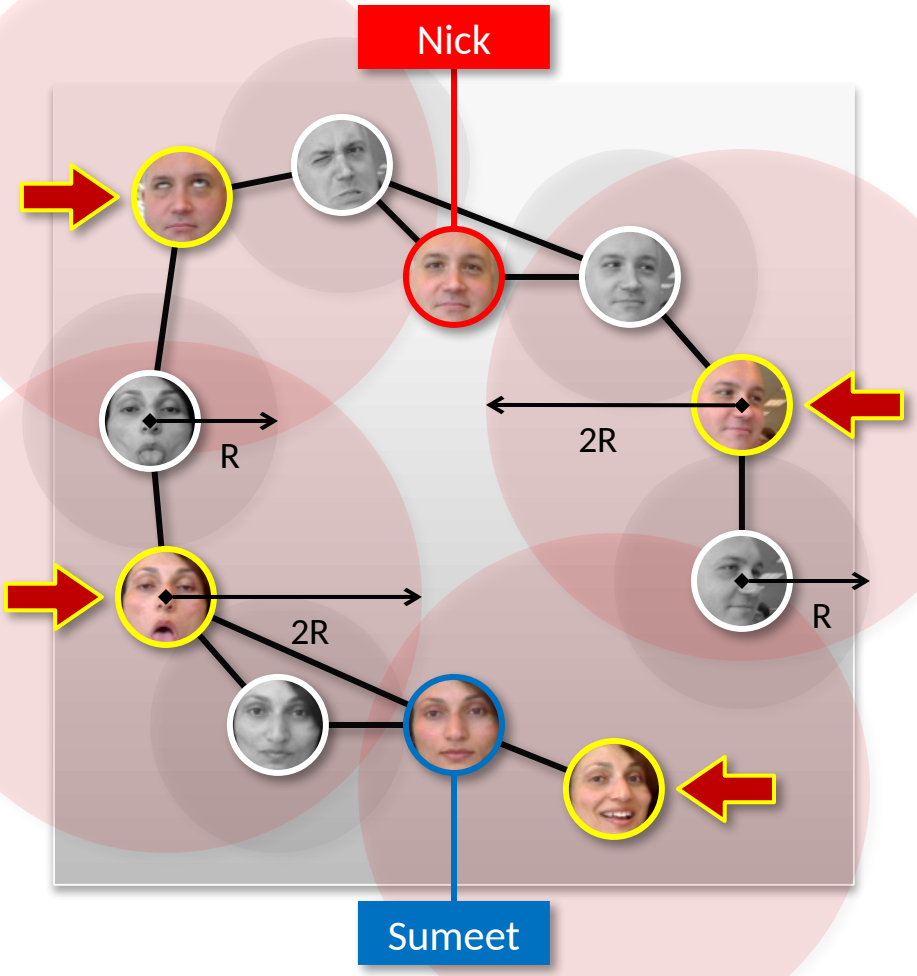
Graph Quantization

- When the example x_t is less than R away a representative vertex, the example is discarded
- When the graph gets large
  - R is doubled
  - vertices are repartitioned
    - no two vertices are closer than R
    - old vertices are covered by at least one new vertex



Graph Quantization

- When the example x_t is less than R away a representative vertex, the example is discarded
- When the graph gets large
  - R is doubled
  - vertices are repartitioned
    - no two vertices are closer than R
    - old vertices are covered by at least one new vertex



Online Quantized HFS

Inputs: an example x_t, a similarity graph W of up to k representative vertices, a diagonal matrix of vertex multiplicities V

Algorithm:

Time complexity of $\theta(k)$

Add the example x_t to the graph W and quantize it

Update the matrix of vertex multiplicities V

Compute the Laplacian L of (V W V) and infer labels:

$$\min \ell^{\mathrm{T}} L \ell \quad \text{s.t.}\ \ell_i = y_i \ \text{for all}\ i \in l$$

Time complexity of $O(k^3)$

Predict $\hat{y}_t = \mathrm{sgn}(\ell_t)$

Outputs: a prediction $\hat{y}_t$, an updated graph W and multiplicities V

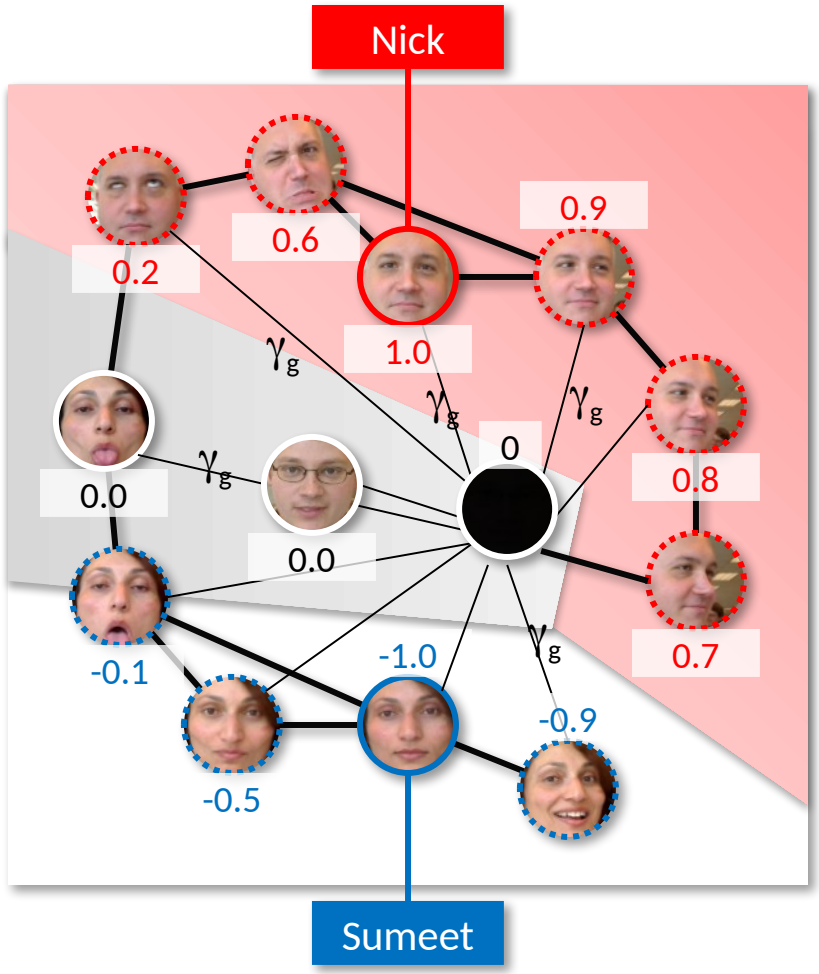
Real-Time Learning without Explicit Feedback. In Proceedings of the 4th IEEE Online Learning for Computer Vision Workshop, San Francisco, California, June 2010. Best paper award.

Outliers

- Real-world problems often involve outliers
- Extrapolation to unlabeled data needs to be controlled
  - The graph W is turned into an ε-neighborhood graph
  - The HFS is regularized as:

$$(L_{rw} + \gamma_g I)\ell_r = W_{rr}\ell_r$$

  - If the label of an example x_t cannot be inferred with a sufficiently high confidence, the example is discarded



Theoretical Analysis

- Prove a bound of the form:

$$\frac{1}{n}\sum_{t}(\ell_t^{\mathrm{q}}[t]-y_t)^2 \le \frac{1}{n_l}\sum_{i\in l}(l_i^*-y_i)^2+\mathrm{O}(n^{-\frac{1}{2}})$$

Error of our solution

Empirical risk on labeled vertices

- Idea of the proof: The regularization parameter γ_g is set such that the error term vanishes as n increases

Online Semi-Supervised Learning on Quantized Graphs. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, July 2010.

Theoretical Analysis

- The error of our solution can be decomposed as:

$$\frac{1}{n}\sum_t (\ell_t^{\mathrm{q}}[t]-y_t)^2 \le \frac{3}{n}\sum_t (\ell_t^{*}-y_t)^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{o}}[t]-\ell_t^{*})^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{q}}[t]-\ell_t^{\mathrm{o}}[t])^2$$

Error of our solution	Offline learning error	Online learning error	Quantization error

Claim: When the regularization parameter is set as $\gamma_g = \Omega(n_l^{3/2})$, the difference between the risks on labeled and all vertices decreases at the rate of $O(n_l^{-1/2})$ (with a high probability)

$$\frac{1}{n}\sum_t (\ell_t^{*}-y_t)^2 \le \frac{1}{n_l}\sum_{i\in l} (\ell_i^{*}-y_i)^2 + \beta + \sqrt{\frac{2\ln(2/\delta)}{n_l}}\left(n_l\beta+4\right)$$

$$\beta \le \left[\frac{\sqrt{2}}{\gamma_g+1} + \sqrt{2n_l}\,\frac{1-\sqrt{c_u}}{\sqrt{c_u}}\,\frac{\lambda_M(L)+\gamma_g}{\gamma_g^2+1}\right]$$

Semi-Supervised Learning with Max-Margin Graph Cuts. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, May 2010.

Theoretical Analysis

- The error of our solution can be decomposed as:

$$\frac{1}{n}\sum_t (\ell_t^{\mathrm{q}}[t]-y_t)^2 \le \frac{3}{n}\sum_t (\ell_t^{*}-y_t)^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{o}}[t]-\ell_t^{*})^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{q}}[t]-\ell_t^{\mathrm{o}}[t])^2$$

Error of our solution	Offline learning error	Online learning error	Quantization error

Claim: When the regularization parameter is set as $\gamma_g = \Omega(n^{1/4})$, the average error between the offline and online HFS predictions decreases at the rate of $O(n^{-1/2})$

$$\frac{1}{n}\sum_t (\ell_t^{\mathrm{o}}[t]-\ell_t^{*})^2 \le \frac{1}{n}\sum_t \left\|\ell^{\mathrm{o}}[t]-\ell^{*}\right\|_2^2 \le \frac{4n_l}{(\gamma_g+1)^2}$$

$$\left\|\ell\right\|_2 \le \frac{\left\|y\right\|_2}{\lambda_m(C^{-1}K+I)} = \frac{\left\|y\right\|_2}{\lambda_m(K)\lambda_M^{-1}(C)+1} \le \frac{\sqrt{n_l}}{\gamma_g+1}$$

Online Semi-Supervised Learning on Quantized Graphs. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, July 2010.

Theoretical Analysis

- The error of our solution can be decomposed as:

$$\frac{1}{n}\sum_t (\ell_t^{\mathrm{q}}[t] - y_t)^2 \le \frac{3}{n}\sum_t (\ell_t^{*} - y_t)^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{o}}[t] - \ell_t^{*})^2 + \frac{3}{n}\sum_t (\ell_t^{\mathrm{q}}[t] - \ell_t^{\mathrm{o}}[t])^2$$

Error of our solution	Offline learning error	Online learning error	Quantization error

Claim: When the regularization parameter is set as $\gamma_g = \Omega(n^{1/8})$, and the Laplacians L^{q} and L^{o} and normalized, the average error between the online and online quantized HFS predictions decreases at the rate of $O(n^{-1/2})$

$$\frac{1}{n}\sum_t (\ell_t^{\mathrm{q}}[t] - \ell_t^{\mathrm{o}}[t])^2 \le \frac{1}{n}\sum_t \left\|\ell^{\mathrm{q}}[t] - \ell^{\mathrm{o}}[t]\right\|_2^2 \le \frac{n_l}{c_u^2 \gamma_g^4}\left\|L^{\mathrm{q}} - L^{\mathrm{o}}\right\|_F^2$$

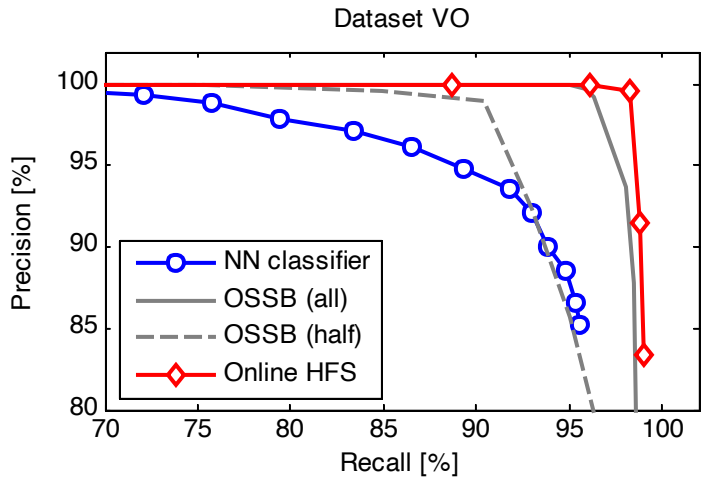
$$\left\|L^{\mathrm{q}} - L^{\mathrm{o}}\right\|_F^2 \propto O(k^{-2/d})$$

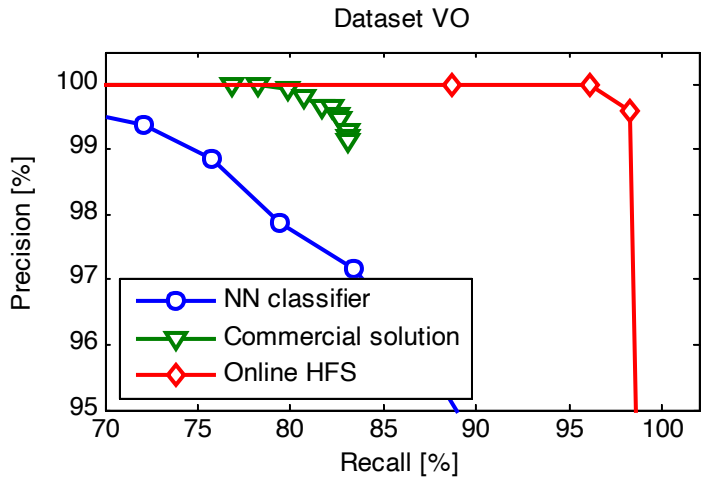
The distortion rate of online k-center clustering is $O(k^{-1/d})$, where d is dimension of the manifold and k is the number of representative vertices

Online Semi-Supervised Learning on Quantized Graphs. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, July 2010.

Experiment 1

- One person moves among various indoor locations
- 4 labeled examples of a person in the cubicle

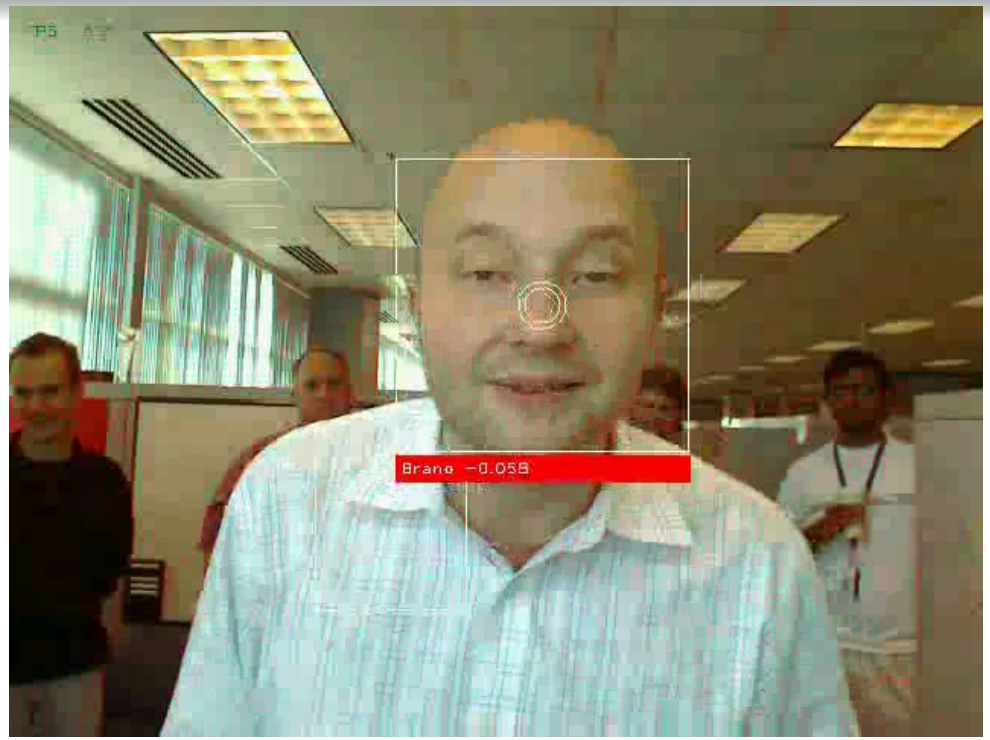




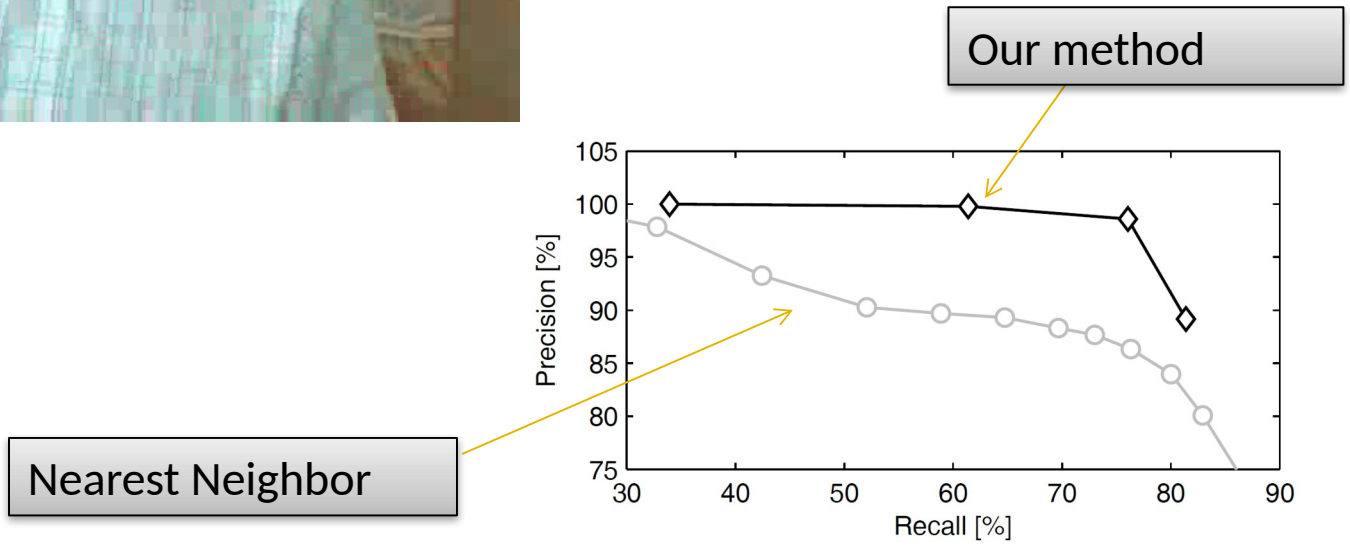
Online HFS outperforms OSSB (even when the weak learners are chosen using future data)

Online HFS yields better results than a commercial solution at 20% of the computational cost

Mutli-class Experiment

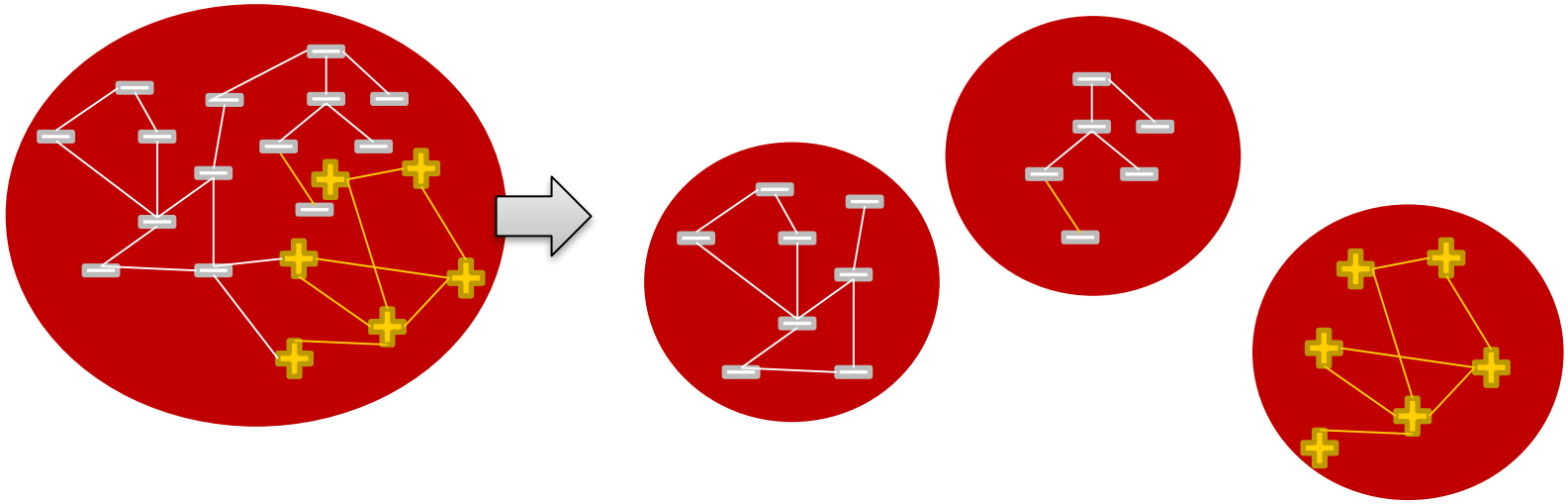


- 8 people classification
- Making funny faces
- 4 faces/person are labeled



Parallel Harmonic Solution

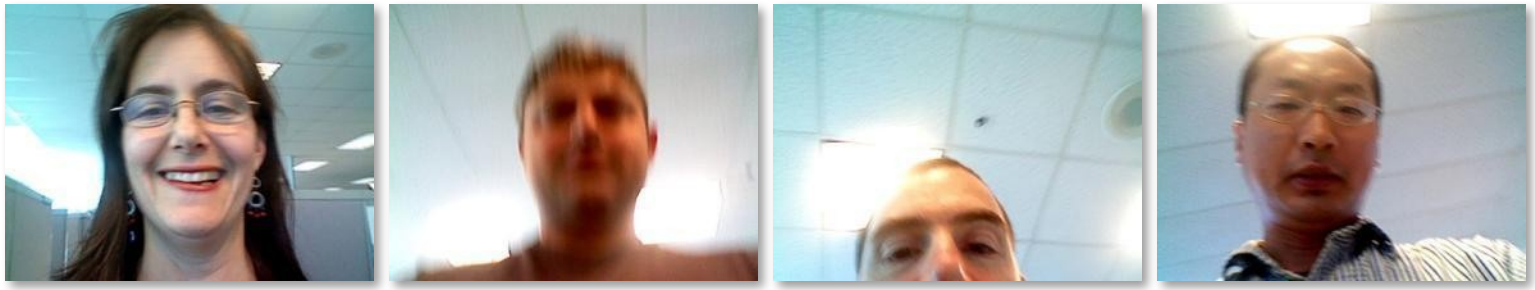
- Addressing scalability



- Learning several smaller manifolds
- Inference in parallel
- Good approximation when the similarity matrix has a block-diagonal structure

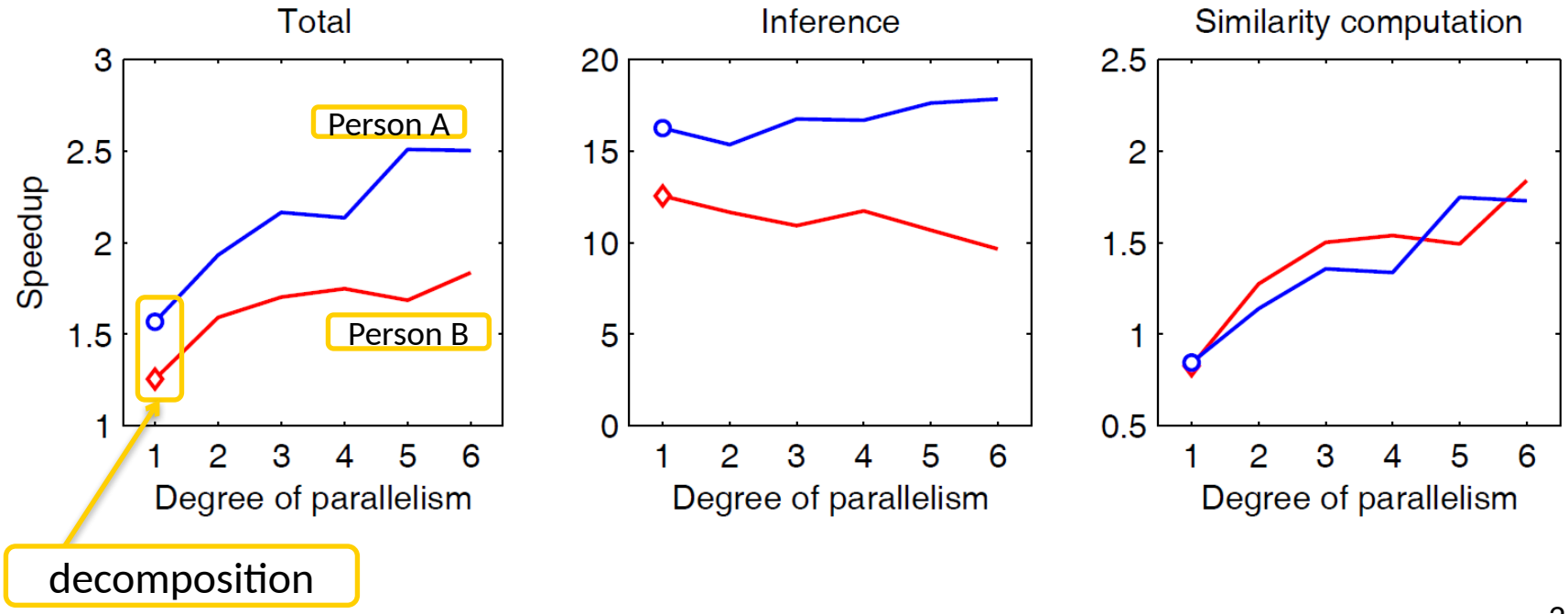
Parallel SSL Experiment

- authentication with a face
- logging into a tablet PC with their face
- 10 different locations
- when the owner is not recognized
  - enters a password
  - we get a new labeled example



Parallel SSL Results

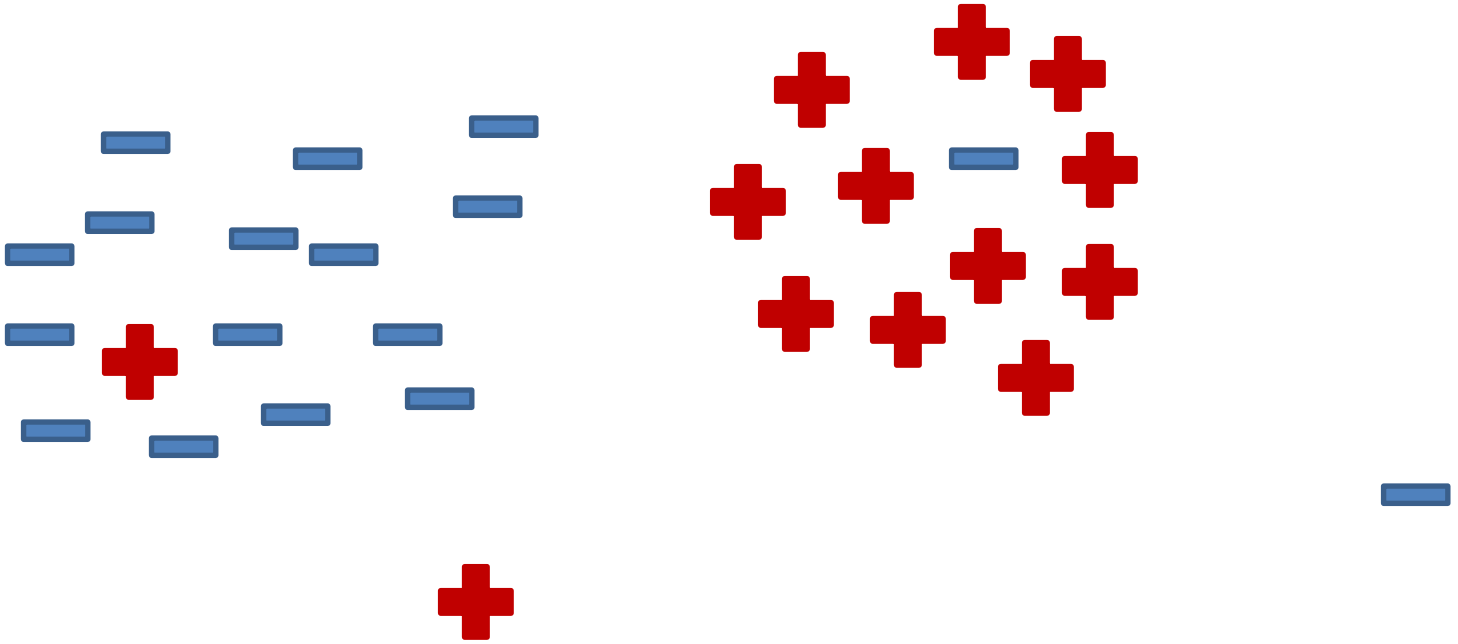
- Nodes: 300 vs. 6x50
- Speedup
  - 35% due to decomposition
  - 2x due to parallelism
- Accuracy
  - Loss less than 3%



Outline

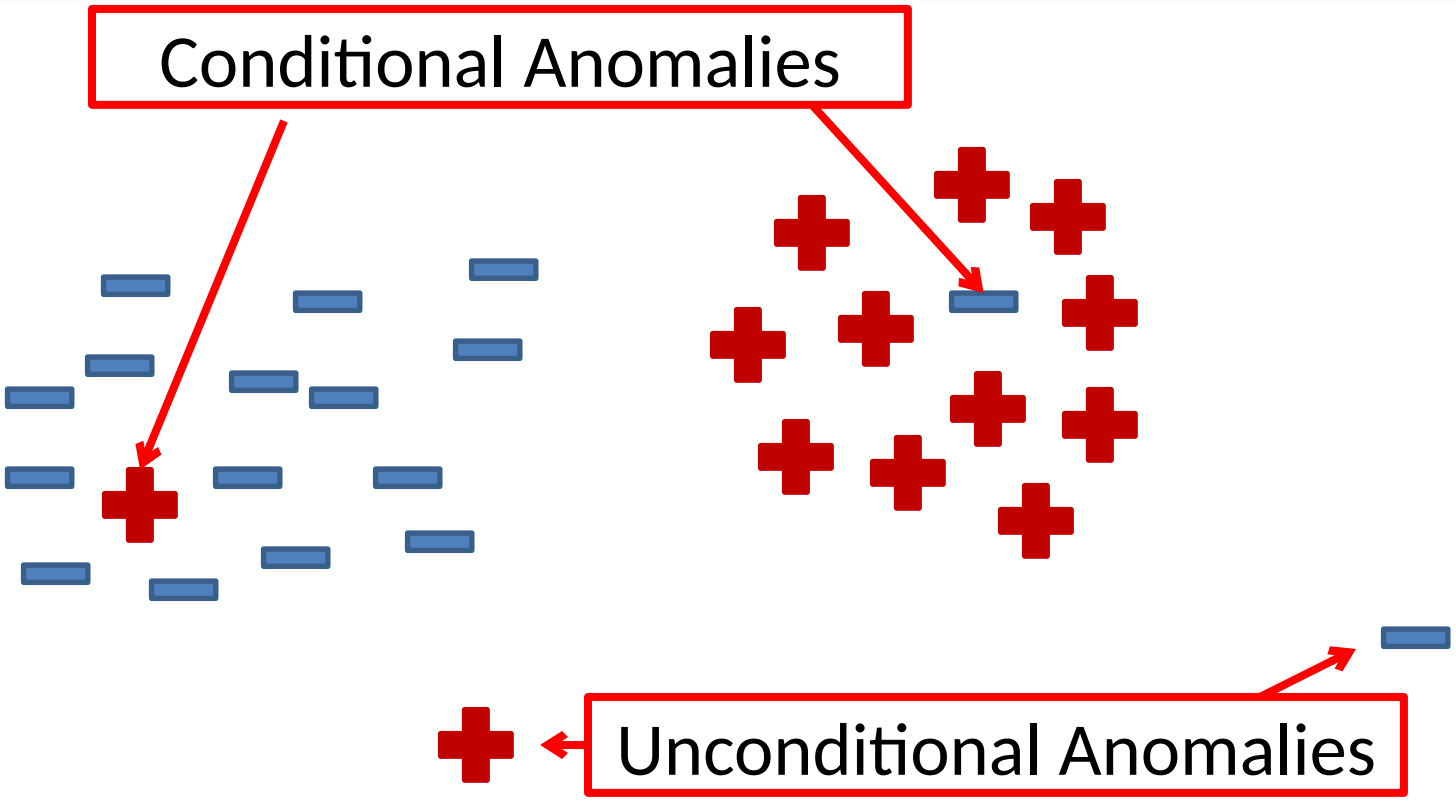
- Graph Based Learning
- Semi-Supervised Learning (SSL) on Graphs
- Online SSL with Quantized on Graphs
- Application: *Face Recognition*
- Conditional Anomaly Detection
- Application: *Medical Error Detection*

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)

Current systems use expert-created rules

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

CAD approaches

Class Outlier Approach

- OneClass SVM, LOF, …

Discriminative Approach

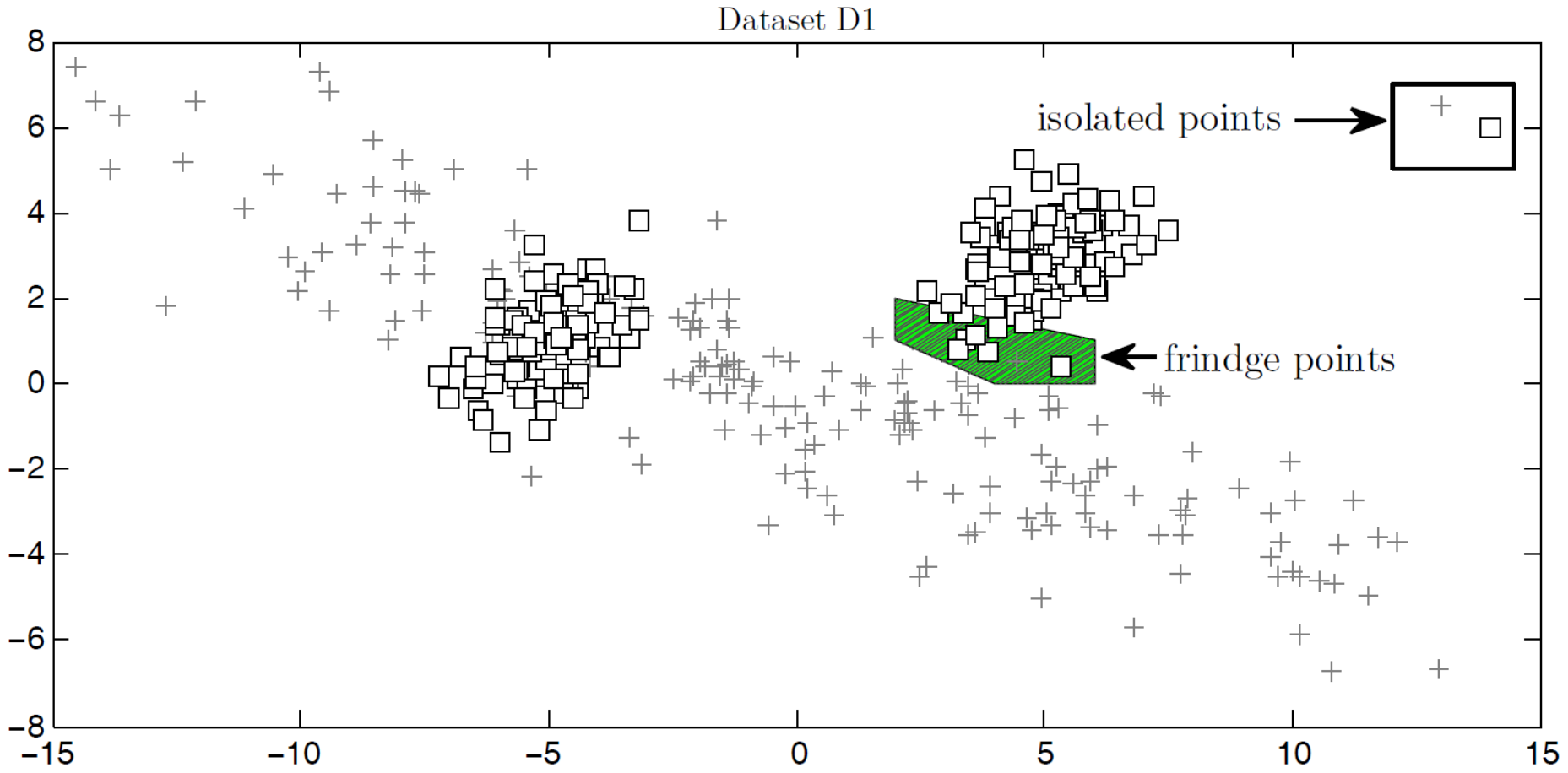
- SVM-CAD

Regularized Discriminative Approach

- Soft Harmonic AD

Challenges for CAD

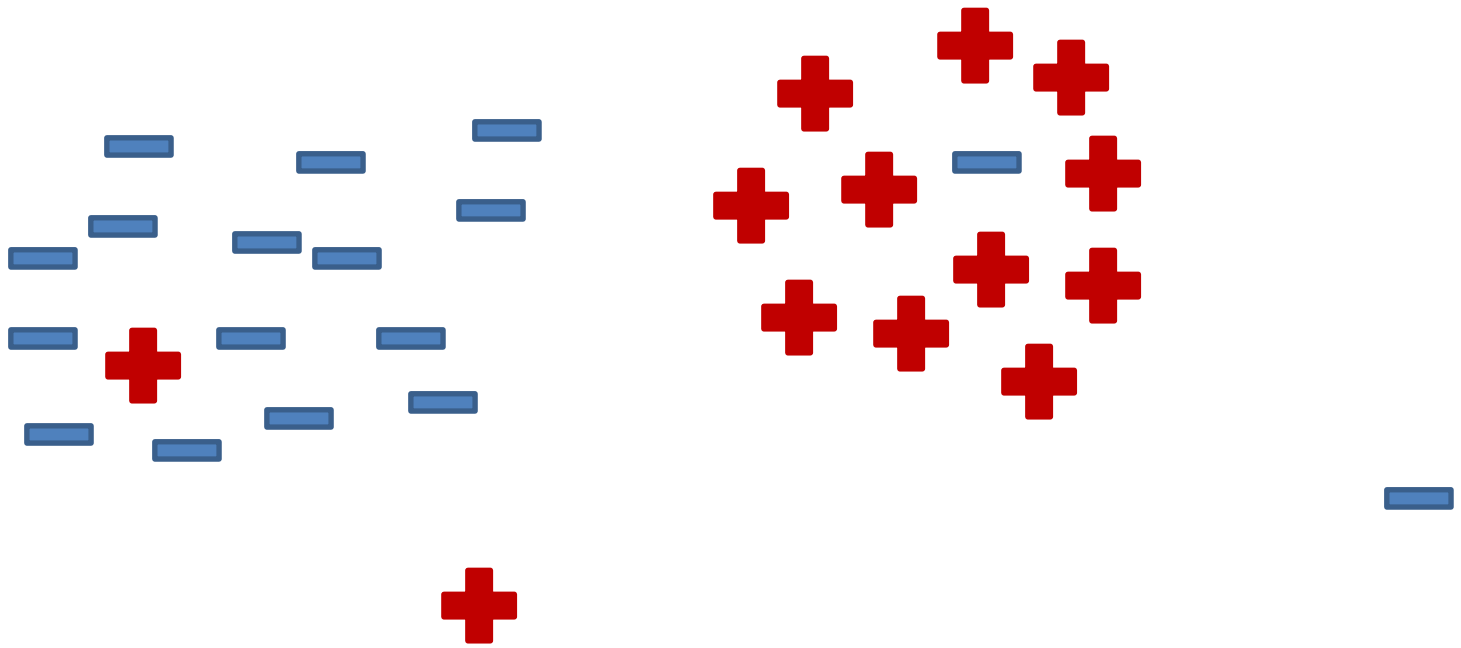
Task: detect anomalies in labels/decisions



Class Outlier Approach

ignores the other class(es)

- Take a test case (**x**,y)
- Take any unconditional anomaly method
- Find out if **x** anomalous wrt { x | x has class y }

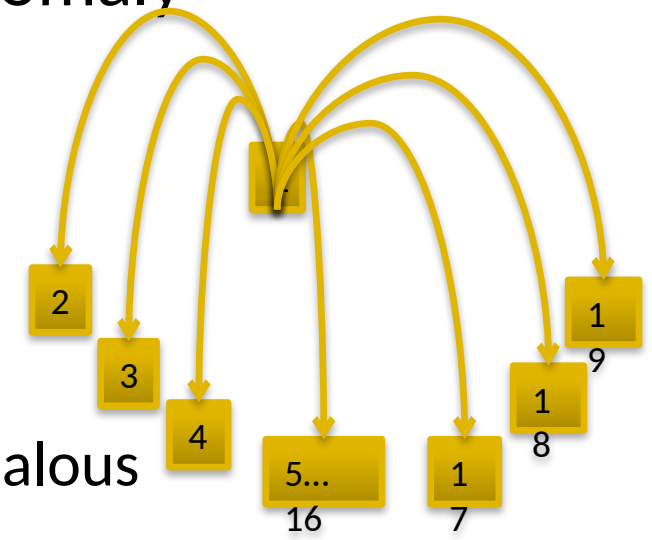


Discriminative Approach

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

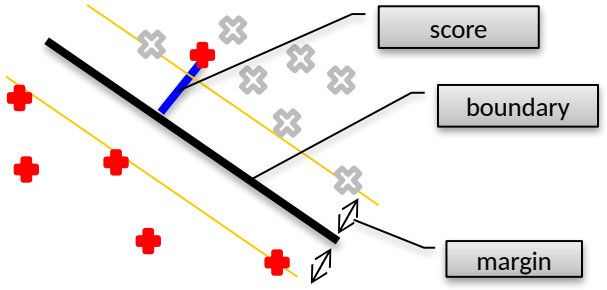
$$d(y|\mathbf{x}) = P(y'|\mathbf{x}) \quad y' \neq y$$

  - bigger the alert score→ more anomalous



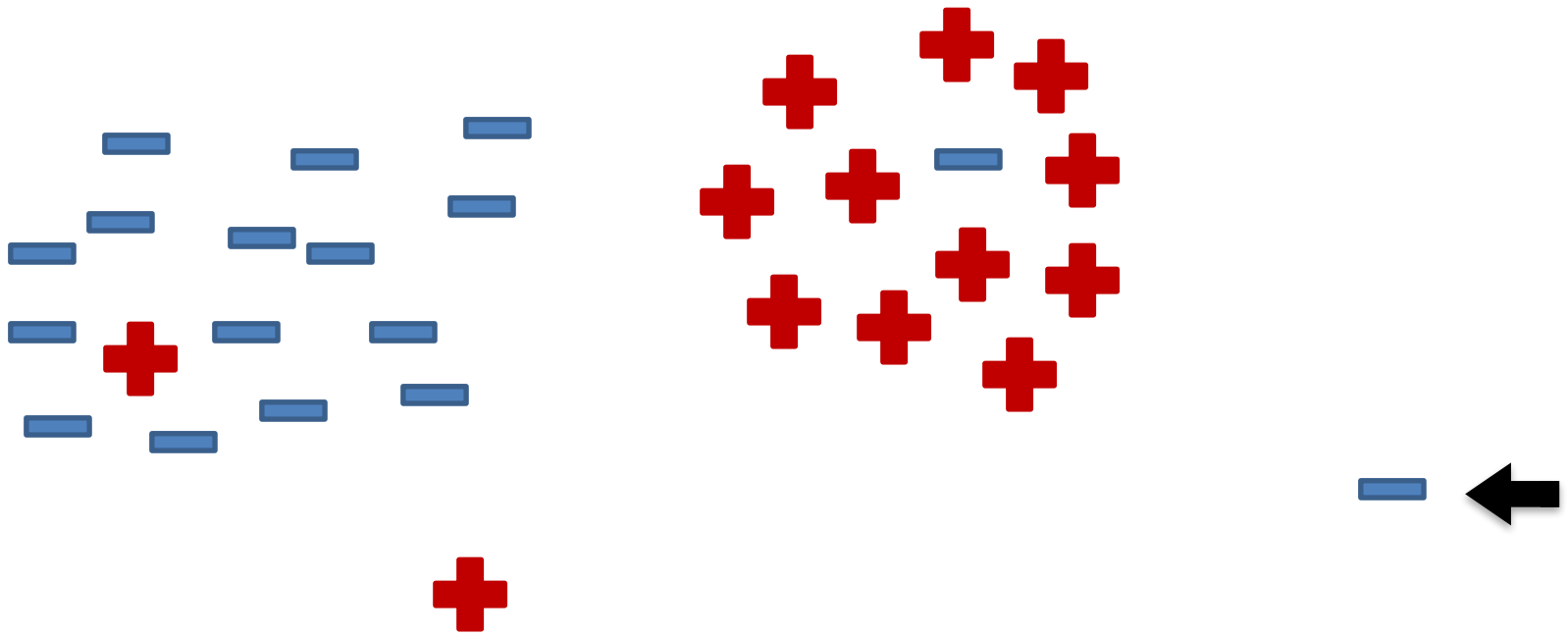
- Support Vector Machines projections

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



Discriminative Approach

- Problem: (unconditional) anomalies
- Can become overly confident in the low density areas



Regularized HS for CAD

$$\ell_i = (I - P_{uu})^{-1}_{iu} P_{ul} \ell_l$$

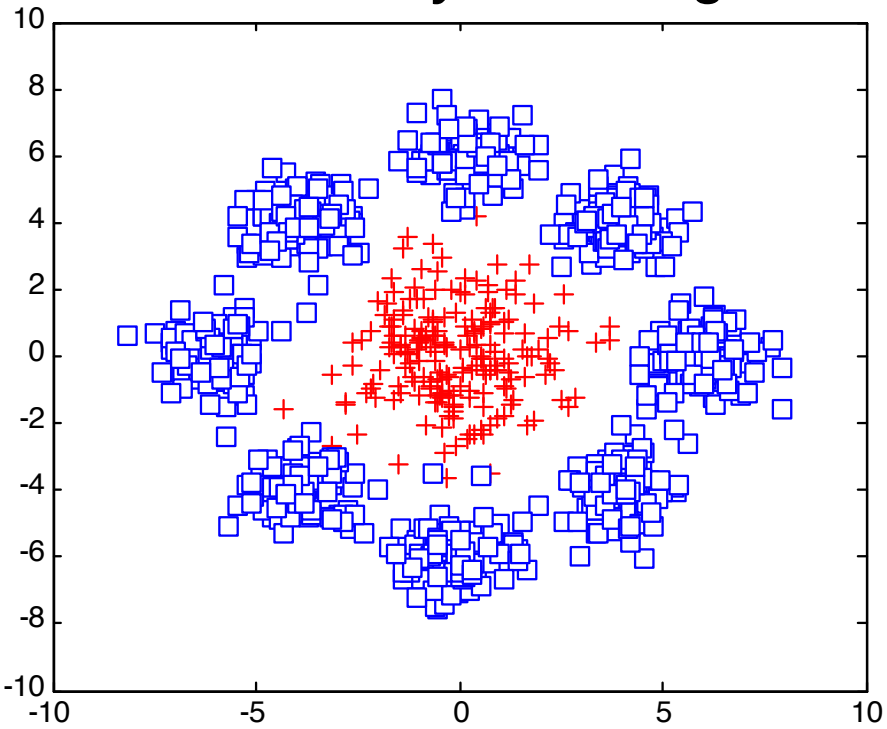
$$= \underbrace{\sum_{j:y_j=1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^{(1)}} - \underbrace{\sum_{j:y_j=-1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^{(-1)}}$$

$$= p_i^{(1)} - p_i^{(-1)}$$

- 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!**
- regularization (with sink) diminishes the effect of outliers

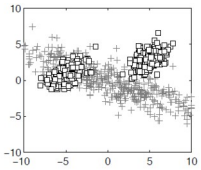
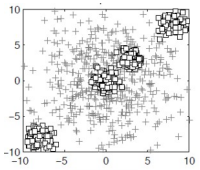
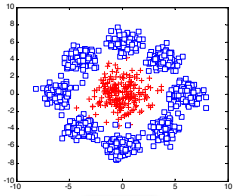
Synthetic Data

- 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

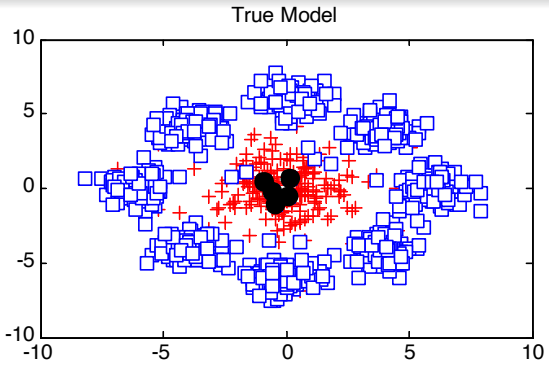




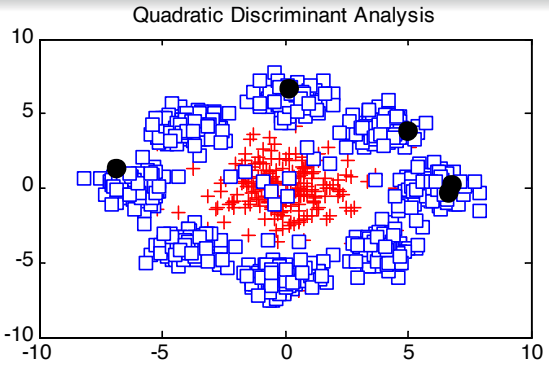
	Dataset **D1**	Dataset **D2**	Dataset **D3**
QDA	73.8% (2.1)	29.4% (5.2)	61.0% (1.2)
SVM	58.8% (7.0)	49.8% (1.7)	46.1% (3.1)
1-class SVM	51.3% (0.9)	47.7% (0.6)	64.7% (0.7)
SoftHAD	86.7% (1.6)	62.8% (1.4)	77.0% (2.7)
wk–NN	74.2% (1.9)	56.5% (1.7)	61.4% (2.1)

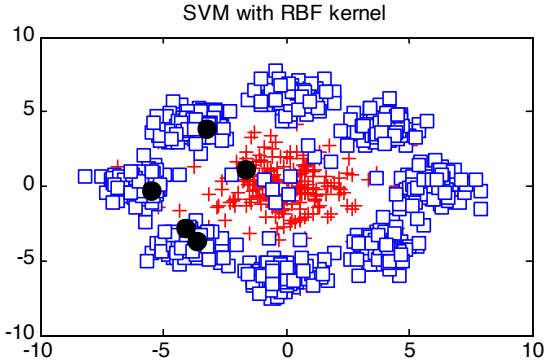
BETTER

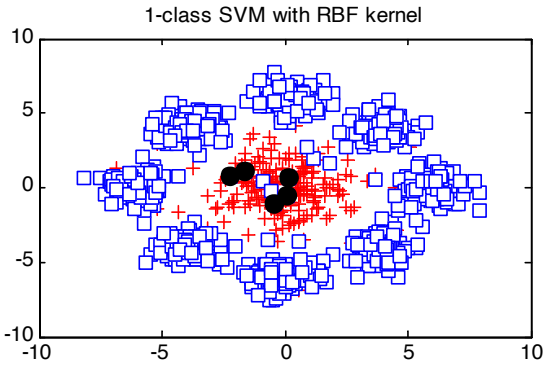
Top 5 best scoring anomalies for different methods on the synthetic dataset D3

TRUE

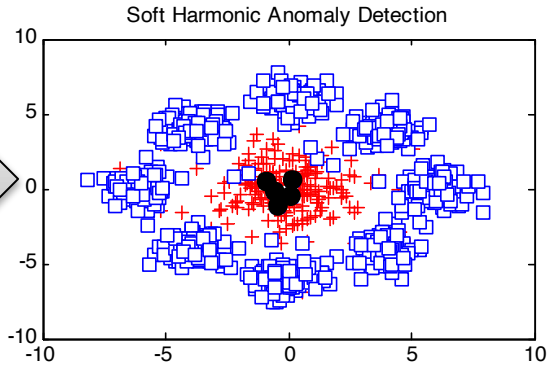


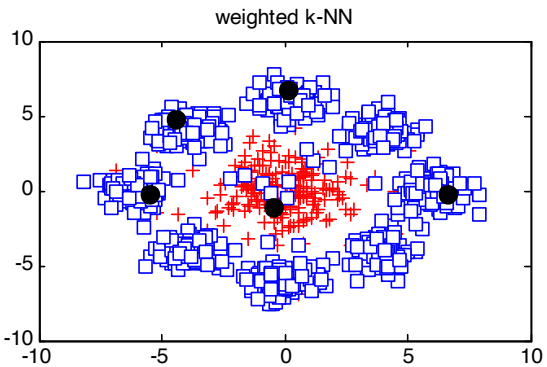






OUR METHOD





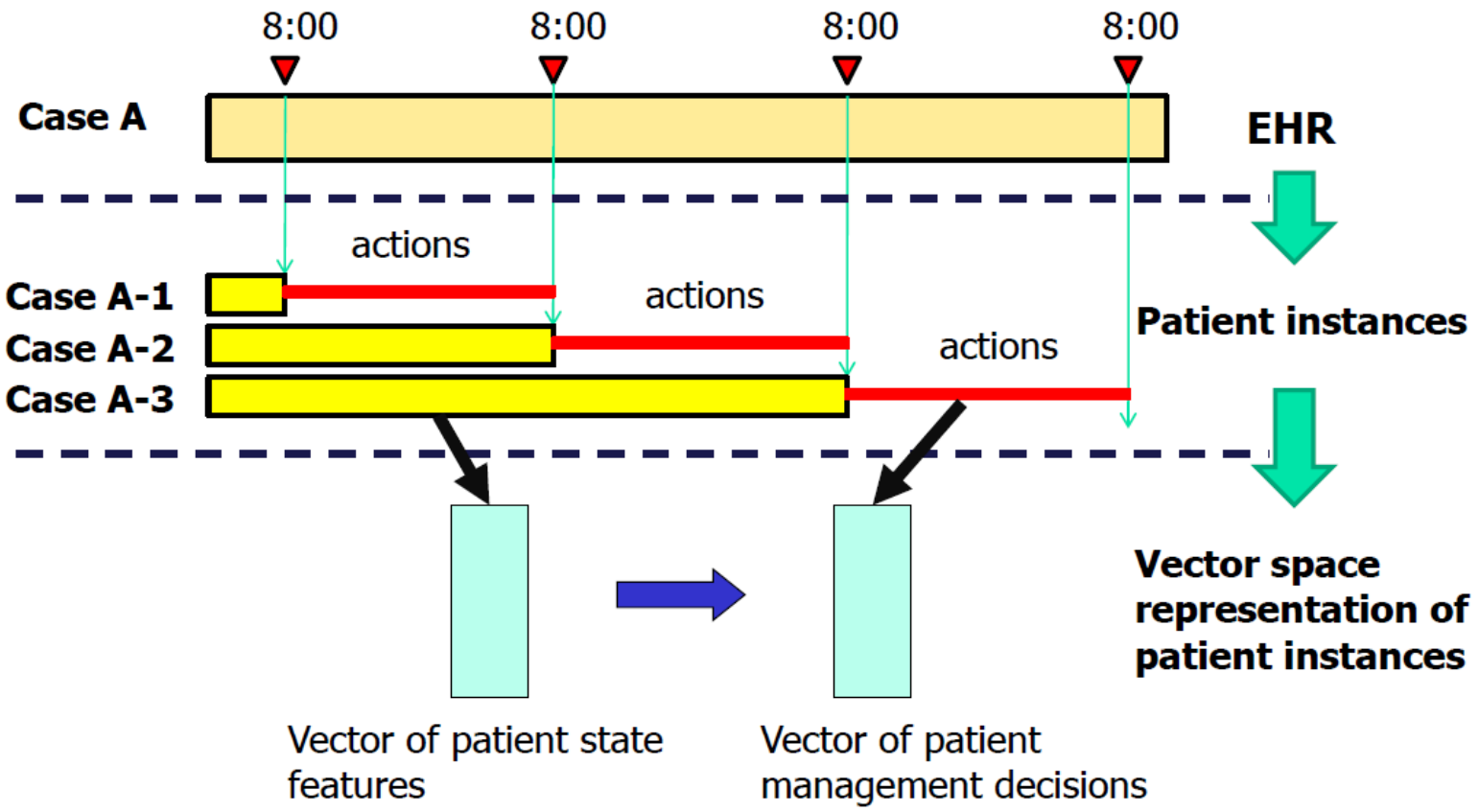
Medical Data

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

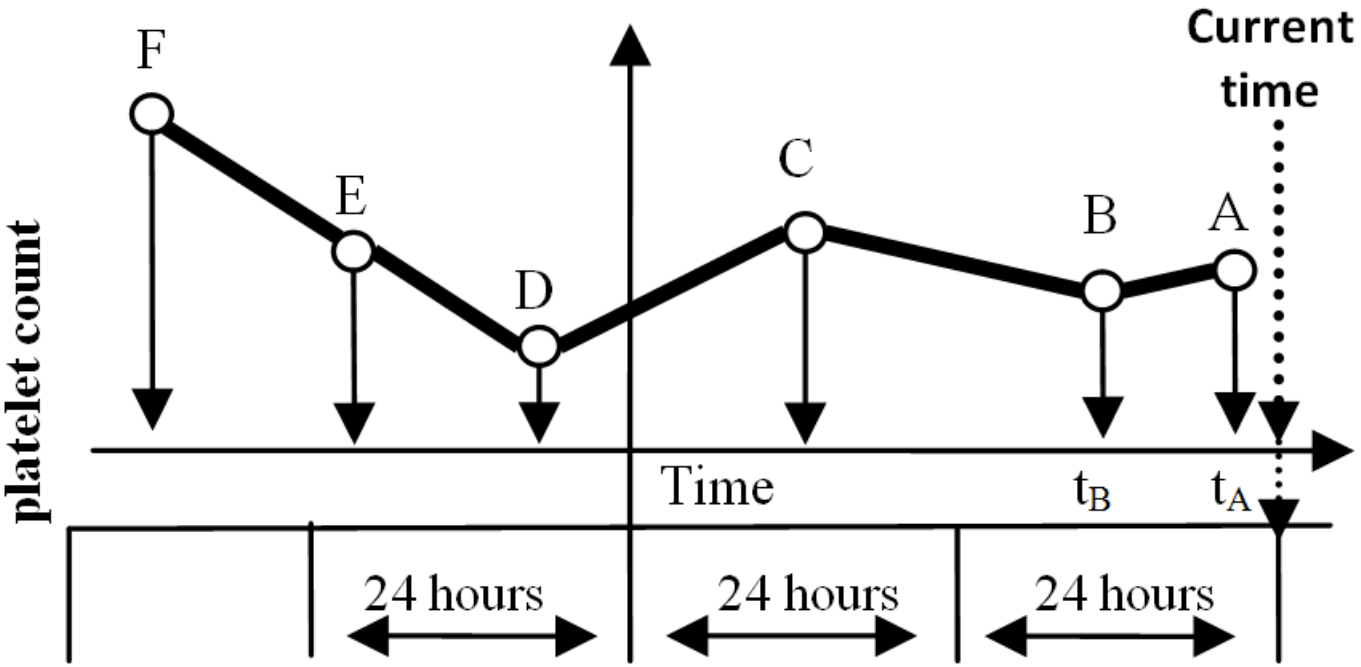
1. Laboratory tests (LABs)
2. Medications (MEDs)
3. Visit features/demographics
4. Procedures
5. Heart support devices

- 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

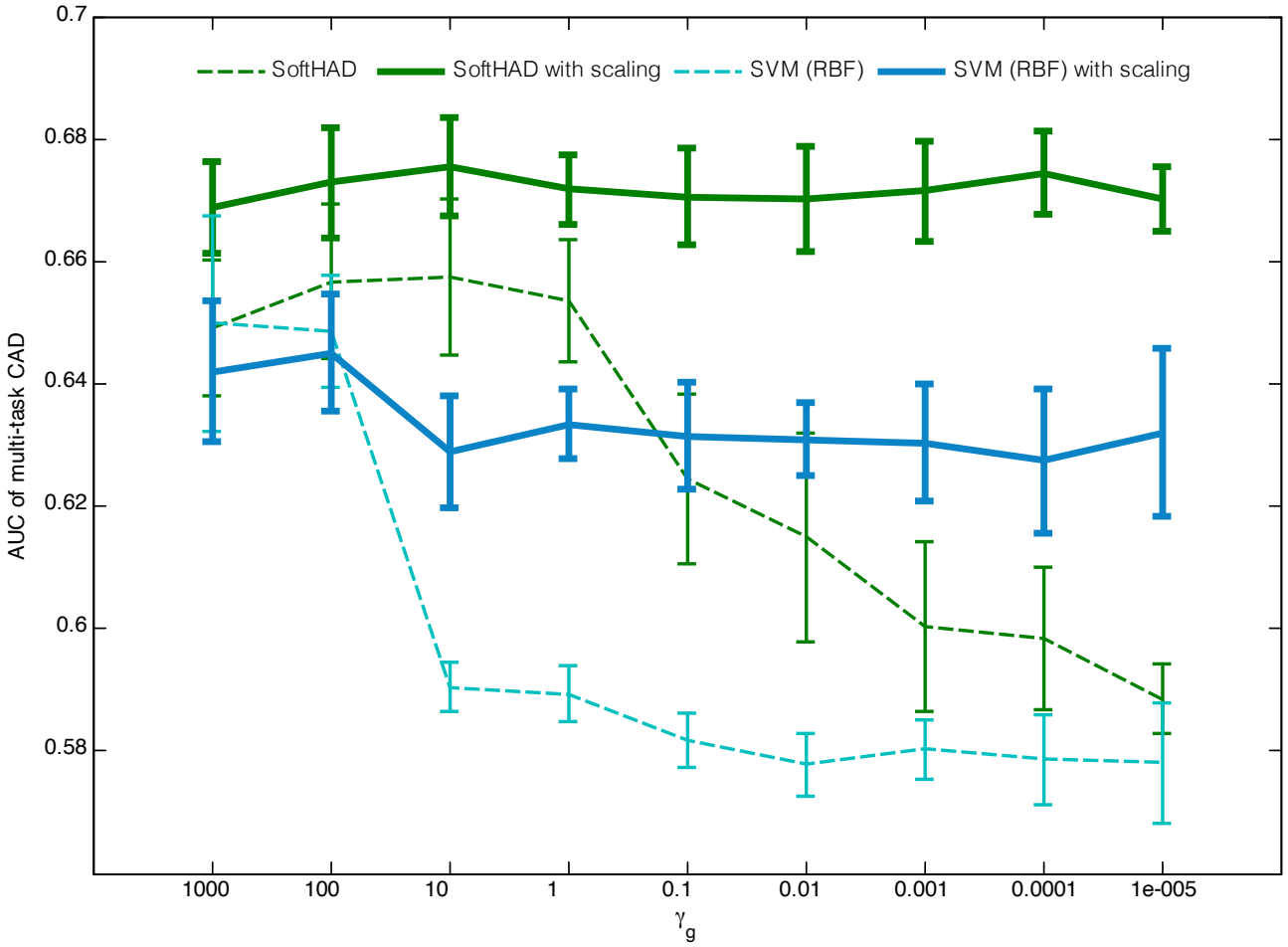
Nadir percentage difference = (A-D)/D

Baseline = F

Drop from baseline = F-A

Medical Data Results

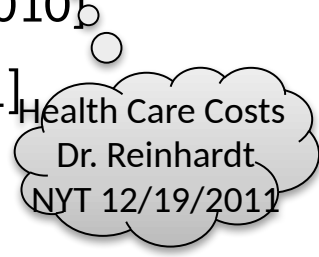
- Outperforming SVM-based method over the range of settings for regularization parameters



BETTER

Summary

- Graph-based methods
- Learning with minimal feedback
- Semi-supervised learning
  - Online Learning with Quantization
    - Online Semi-Supervised Perception [CVPR – OLCV 2010]
    - Bounds on quality of the solution [UAI 2010]
  - Max-Margin Graph Cuts [AISTATS 2010] (in the document)
  - Parallel Semi-Supervised Learning
  - Joint Quantization and Label Propagation (in the document)
- Conditional Anomaly Detection
  - Discriminative Approach [ICML - HEALTH 2008, AMIA 2010]
  - Soft harmonic Anomaly Detection [ICML - GLOBAL 2011]
  - Detecting Unusual Medical Decisions
- Evaluation on challenging real-world problems



Future Work

- Scalability
  - Lifetime online learning
  - “Information exchange” between the manifolds
- Concept Drift
  - Incremental clustering with forgetting the history
  - Recent data more important
- Structured Conditional Anomaly Detection
  - drugs with the same effect do not tend to be given at the same time
  - drugs with the opposite effect do not tend to be given at the same time
  - drugs with negative interactions do not tend to be given at the same time
- Distance Metric Learning for Temporal Data
  - Compare patient states at current time
  - Different kinds, number and time of measurements
  - Missing values
  - Free text

Acknowledgements:

Milos Hauskrecht (PhD advisor), Branislav Kveton (Technicolor Stanford), Greg Cooper, Hamed Valizadegan, Ling Huang (Intel Labs Berkeley), Shyam Visweswaran, Daniel Ting (UC Berkeley), Lily Mummert (Google), Matthai Philipose (Intel Labs Seattle), Ali Rahimi (Intel Labs Berkeley), Saeed Amizadeh, Tomas Singliar (Boeing Research), Amy Seybert, Gilles Clermont, Rich Pelikan, Iyad Batal, Shuguang Wang, Quang Nguyen, Dave Krebs

NIH R21 LM009102–01A1, NIH 1R01LM010019-01A1, Mellon Foundation

Thank you!