Adaptive Graph-Based Algorithms for Conditional Anomaly Detection

Michal Valko

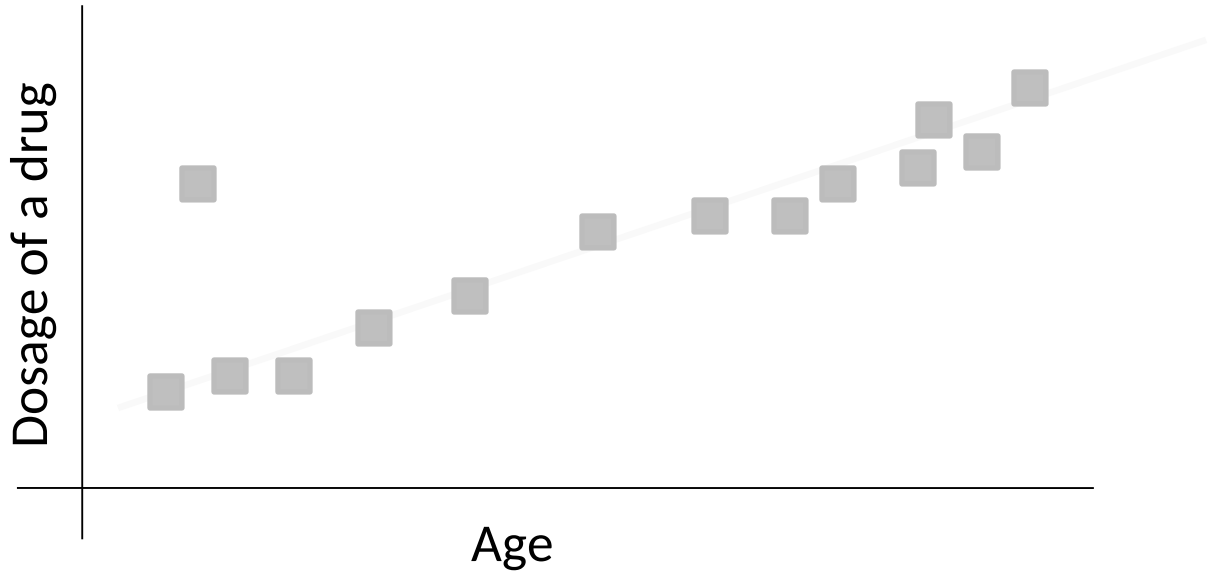
Advisor: Milos Hauskrecht
Committee: Liz Marai, Diane Litman, John Lafferty (CMU)

University of Pittsburgh, December 20th, 2010.

Anomaly (Outlier) Detection

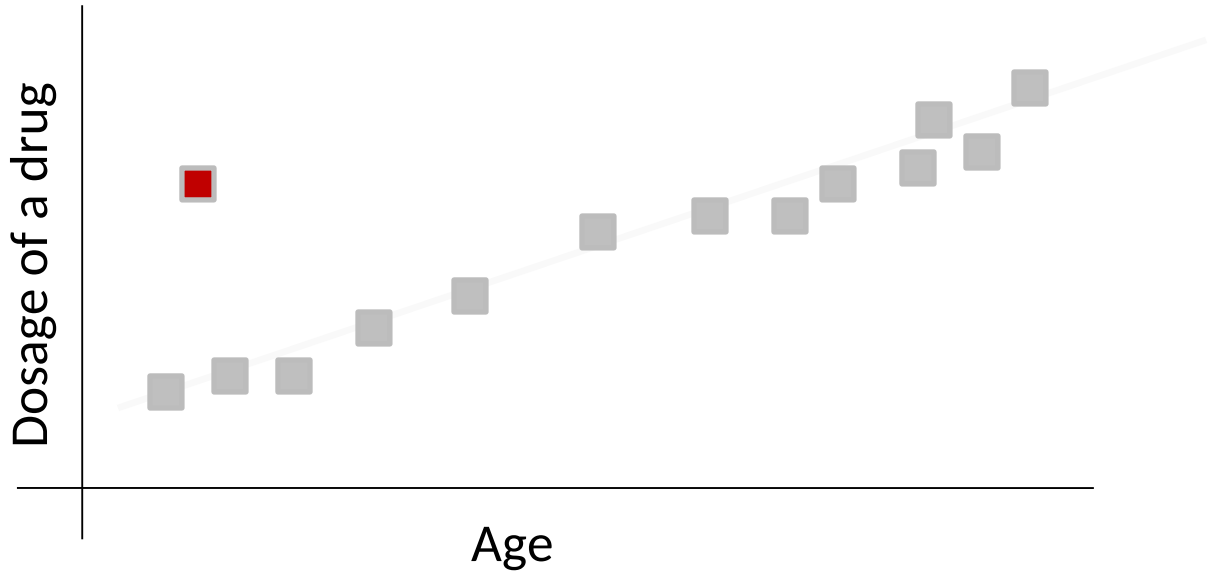
- Goal: Identify unusual patterns in data.
- Methods: from statistics and machine learning
- Contribution: conditional anomaly detection framework
- Application: medical error detection

Conditional Anomaly



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

Conditional Anomaly



Assumption: Anomalies correspond to medical errors

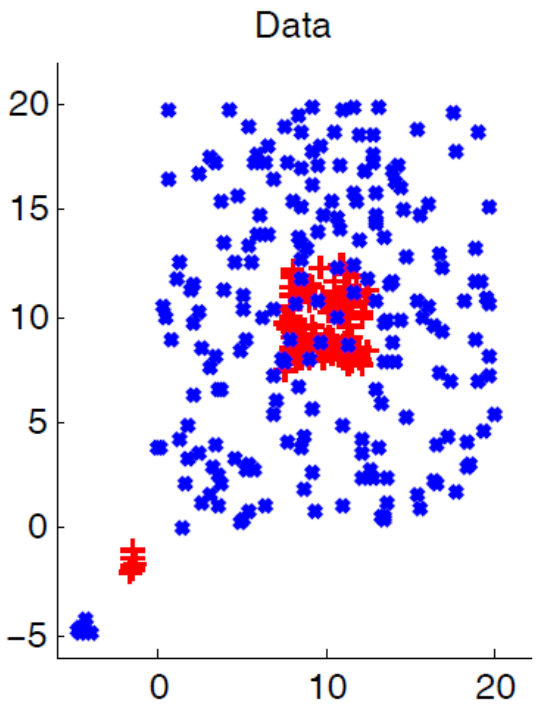
"Medical errors account for 200 000 preventable deaths a year. "

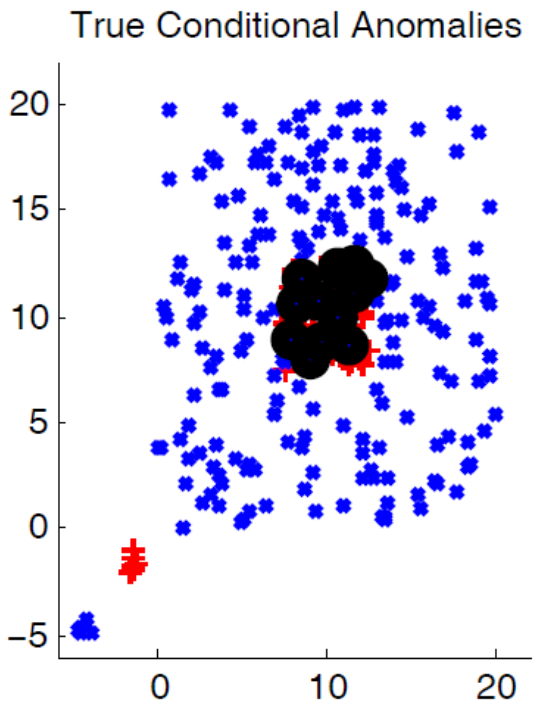
(HealthGrades study, Wall Street Journal, July 27th 2004)

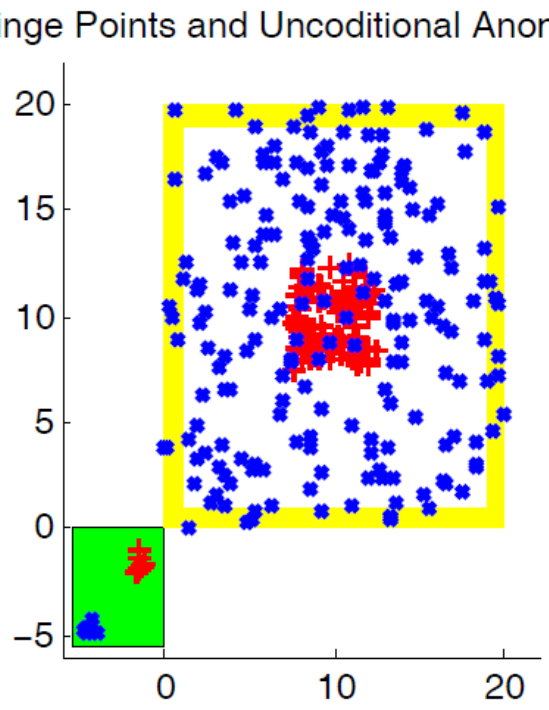
Traditional Anomaly Detection

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

Challenges for CAD



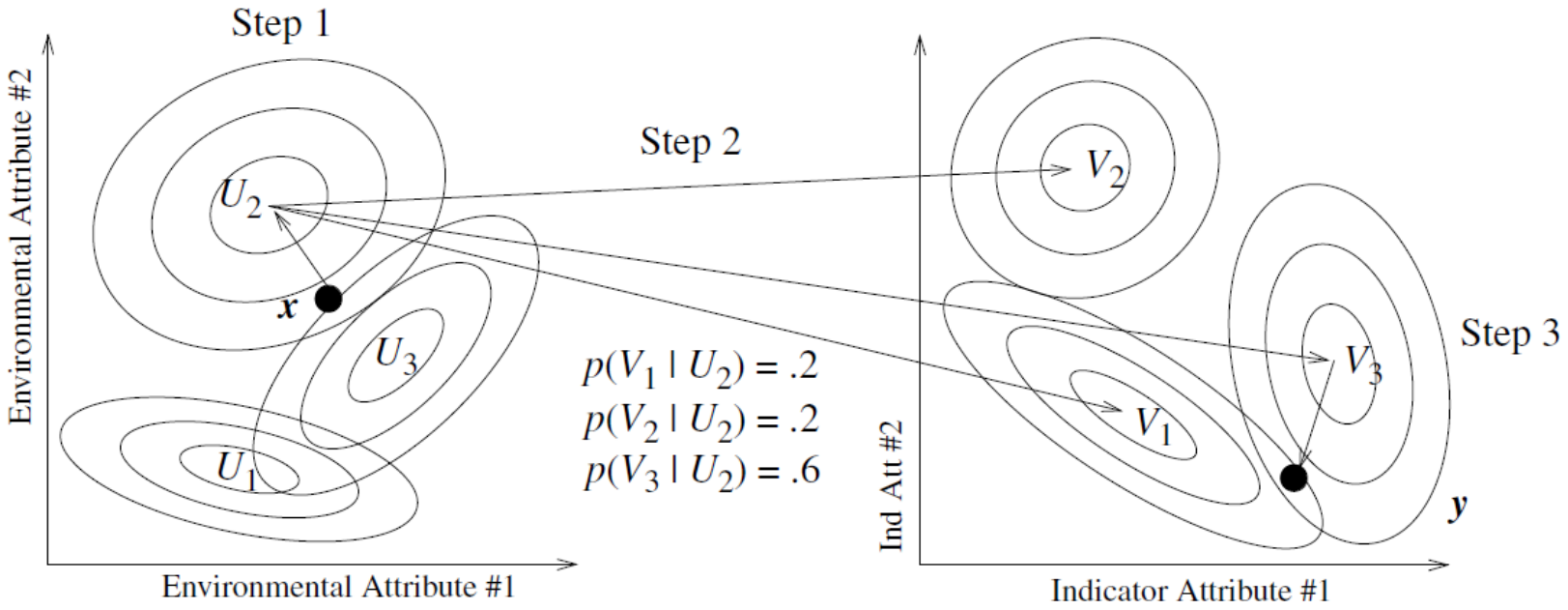




- Dataset adopted from [Papadimitriou and Faloutsos, 2003]

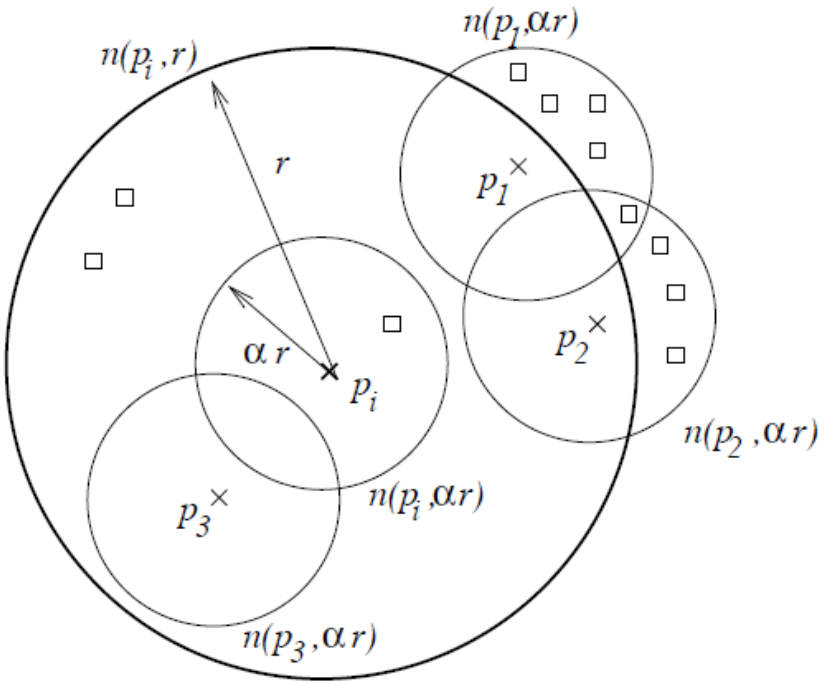
Related Work (CAD)

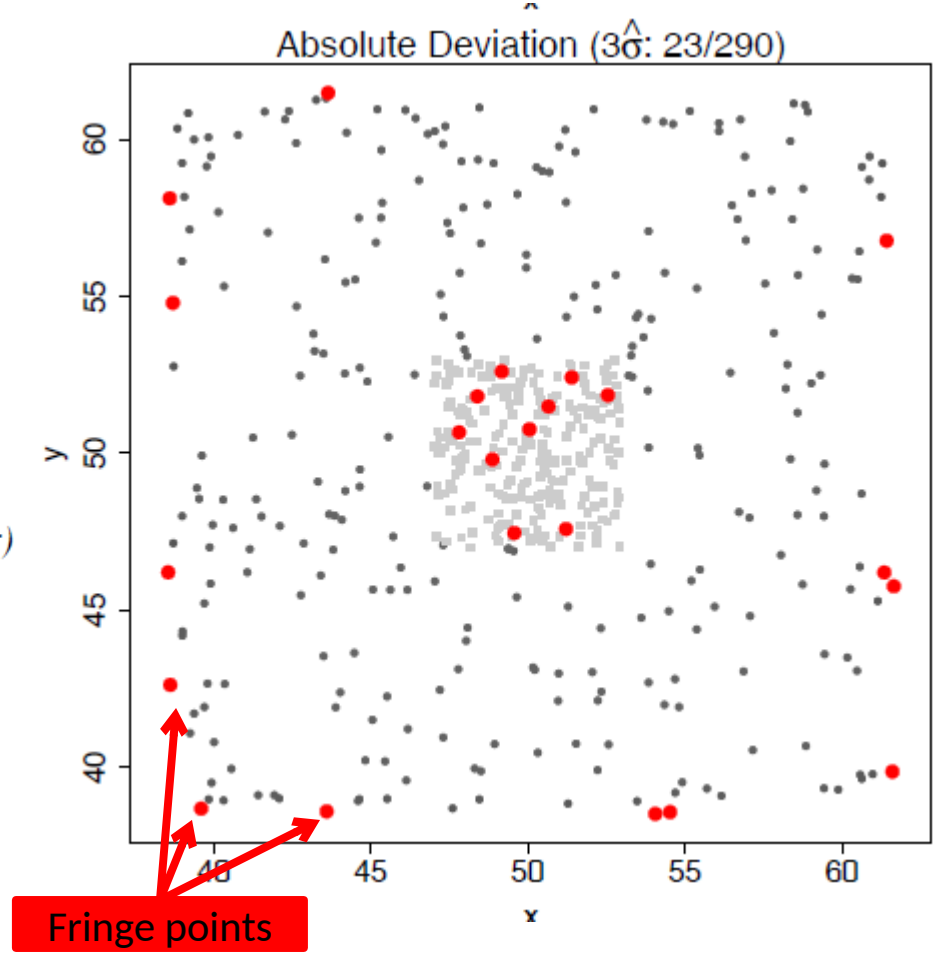
- Conditional Anomaly Detection - Song (2007)
- Use EM to fit mixture of multivariate Gaussians



Related Work (CAD)

- Cross Outlier Detection (Papadimitriou, 2003)





CAD approaches

Class Outlier Approach

- OneClass SVM, LOF, …

Discriminative Approach

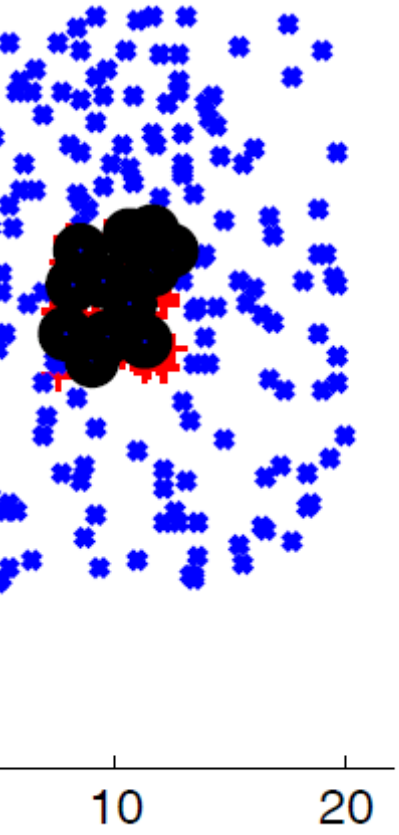
- SVM-CAD

Regularized Discriminative Approach

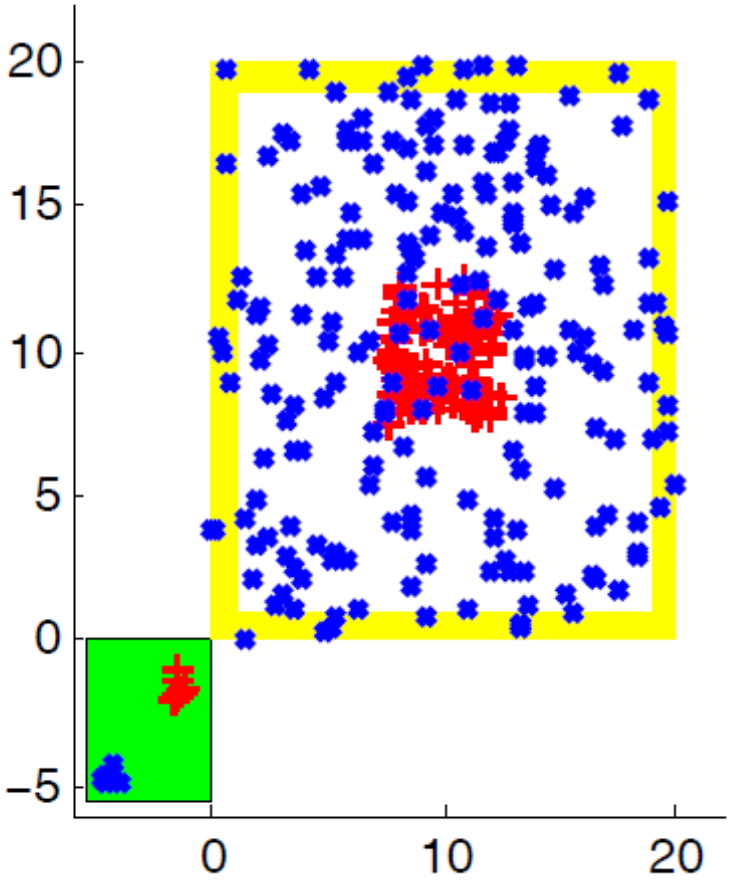
- Connectivity AD, Soft Harmonic AD

Challenges for CAD

itional Anomalies



Fringe Points and Uncoditional Anomalies



Class Outlier Approach

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

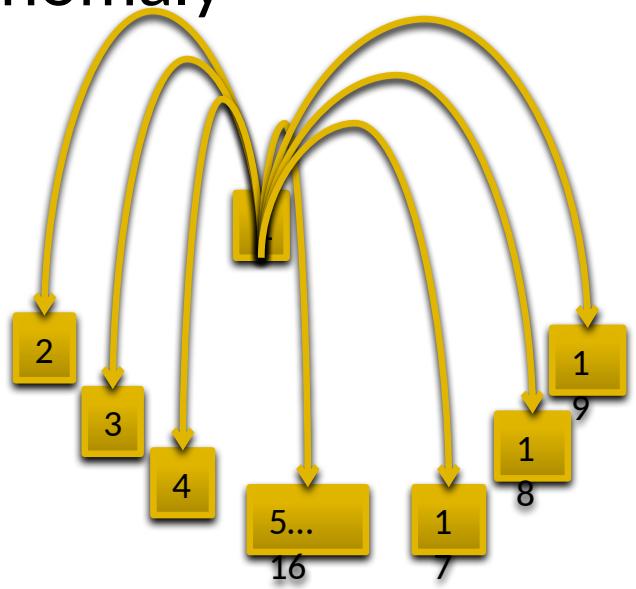
- **Problems:**
  - Anomaly scores for class 1 and class 2 may not be compar
  - Fringe p
  - Unconditional outliers

ignores the other class(es)

Discriminative Approach

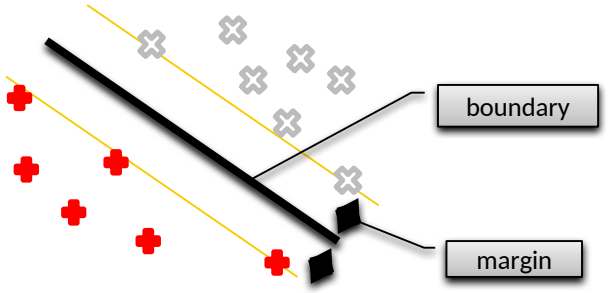
- $P(y|\mathbf{x})$ is small -> conditional anomaly
  - Learn Model/Build Projections
  - Bayes Network

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



- Support Vector Machines projections

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

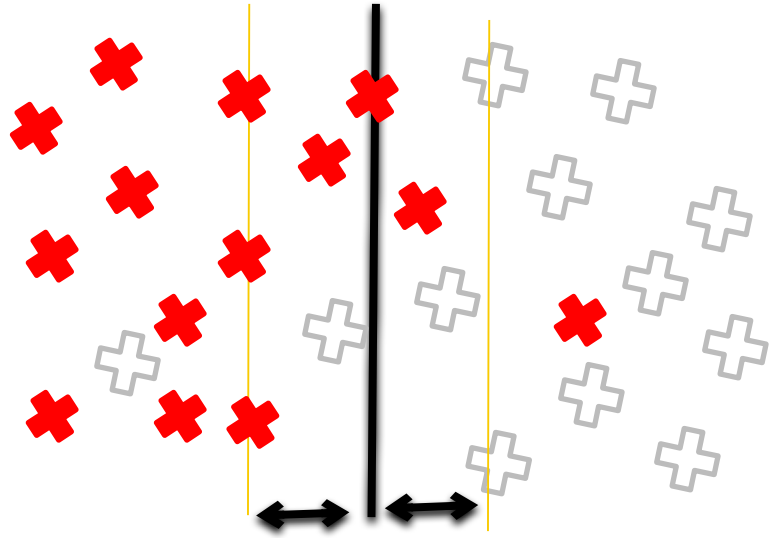


Support Vector Machines projections

[Valko et al., 2008]

[Valko and Hauskrecht, 2008]

[Hauskrecht et al., 2010]



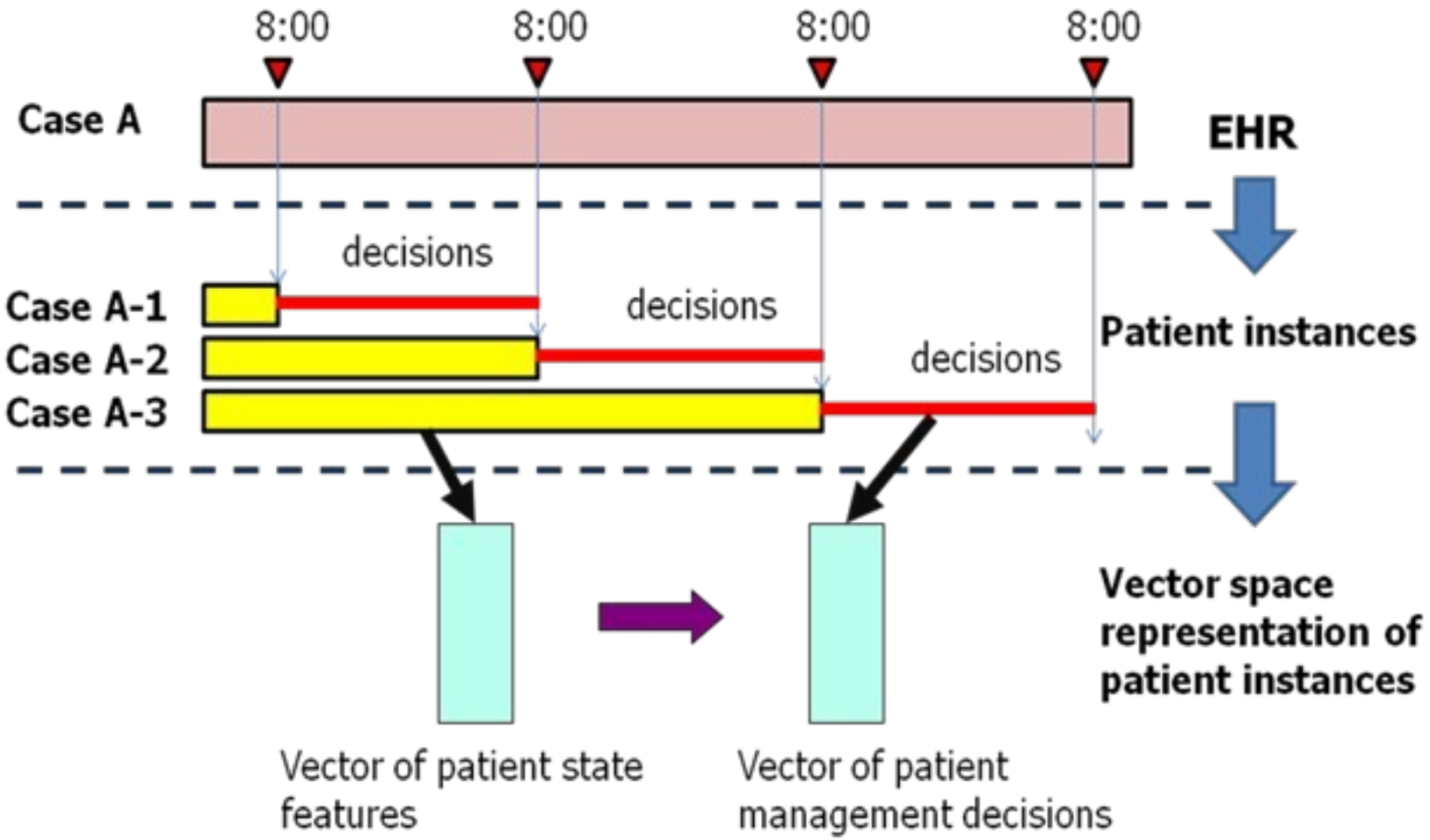
$$d(y|\mathbf{x}) = y(\mathbf{w}^T\mathbf{x} + w_0)$$

Pilot (2009) on Medical PCP Data

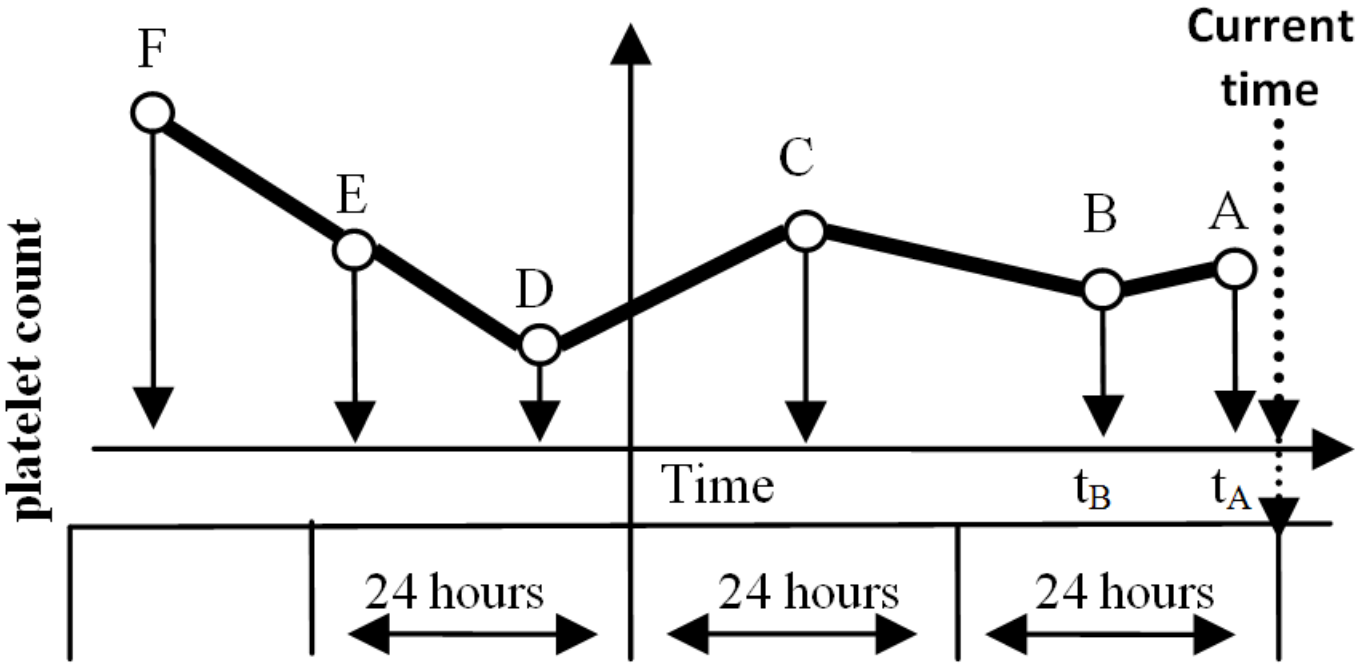
- 4486 patients from UPMC
- Cardiac surgery (2002-2007)
- 45767 patients states
- 9K attributes
- 222 cases evaluated by the 3 pharmacy experts

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

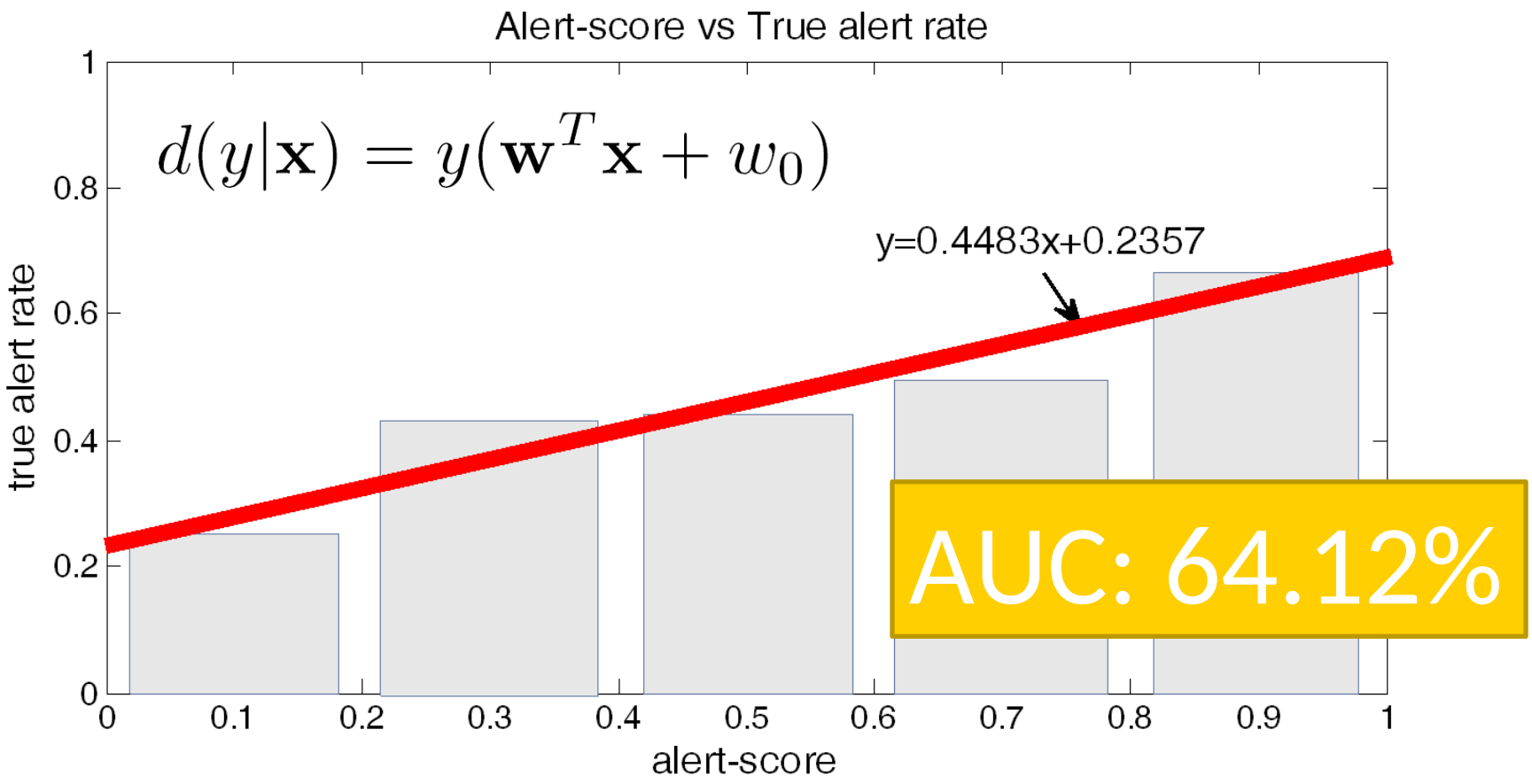
PCP data set: Segmentation



PCP Dataset


F
E
D
C
B
A
Current time
platelet count
Time
t_B
t_A
24 hours
24 hours
24 hours

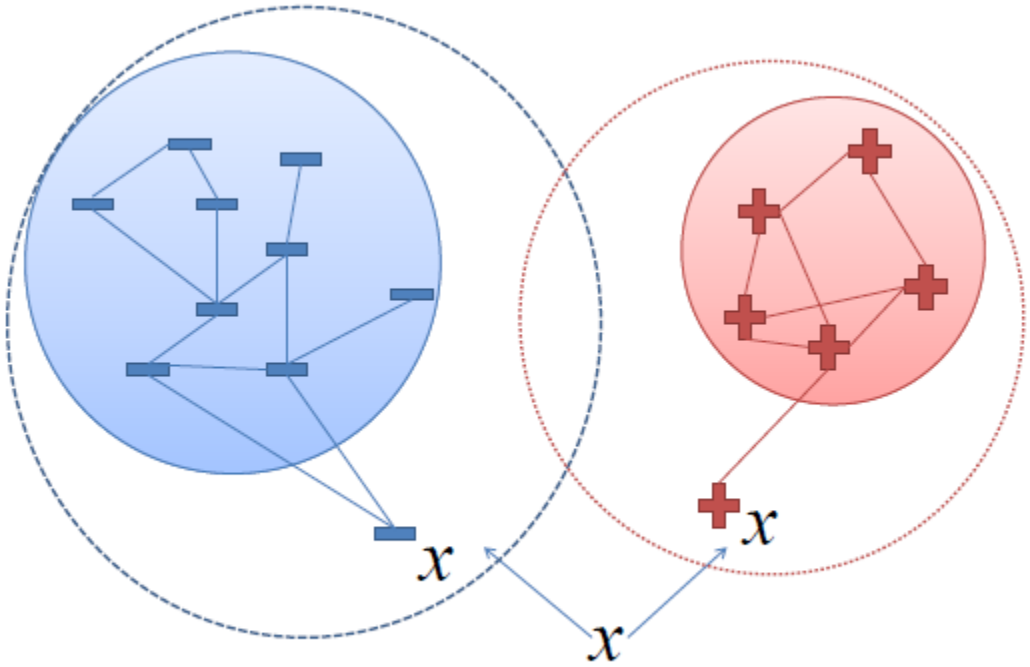
PCP Dataset



- Alert score correlates with true alert rate
- Clinical useful anomalies can be learned

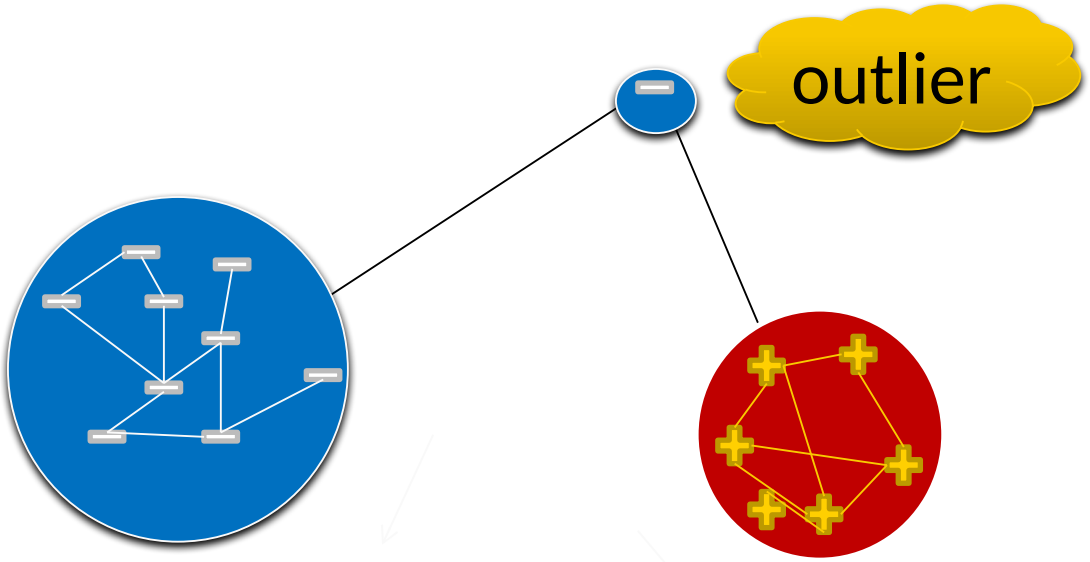
Connectivity AD

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=1)P(y=1) + P(\mathbf{x}|y=0)P(y=0)}$$



Regularization of Connectivity AD

- What happens if we encounter an unconditional outlier?



$$\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=1)P(y=1)+P(\mathbf{x}|y=0)P(y=0)}$$

both small

one magnitudes larger

Algorithm

Algorithm 1 Algorithm Regularized GAD for probability of a positive label.

Inputs:

new example $\mathbf{x}$

similarity metric $K(\cdot,\cdot)$

D^+ for $G_{y=1}$ and D^- for $G_{y=0}$

regularization coefficient λ

Algorithm:

$W_{ix}^+ = K(\mathbf{x},\mathbf{x}_i), \forall i$ positive

$W_{ix}^- = K(\mathbf{x},\mathbf{x}_i), \forall i$ negative

$P(\mathbf{x}|y=1) = \sum_i W_{ix}^+ / (\sum_i D_i^+ i + 2 \times \sum_i W_{ix}^+)$

$P(\mathbf{x}|y=0) = \sum_i W_{ix}^+ / (\sum_i D_i^- i + 2 \times \sum_i W_{ix}^-)$

compute $P(y=1|\mathbf{x})$ according to

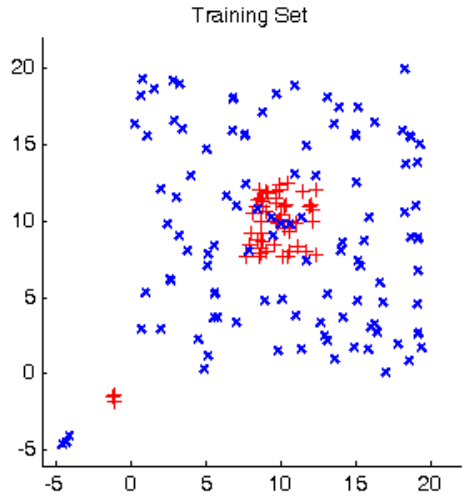
$$\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=1)P(y=1) + P(\mathbf{x}|y=0)P(y=0) + \lambda}$$

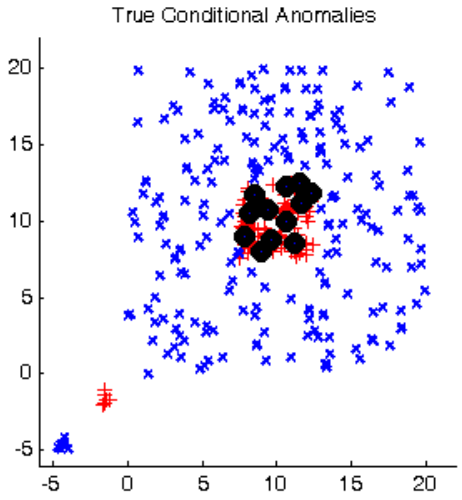
Outputs:

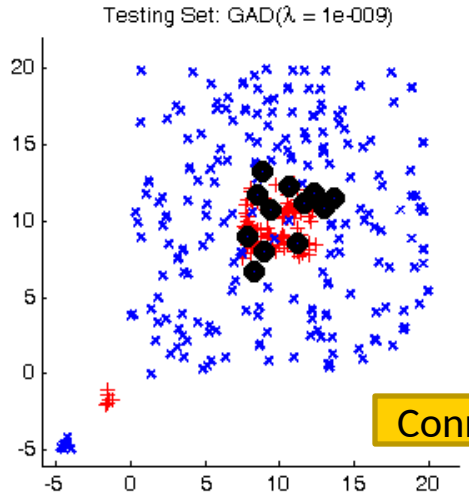
$P(y=1|\mathbf{x})$

Synthetic Core Dataset

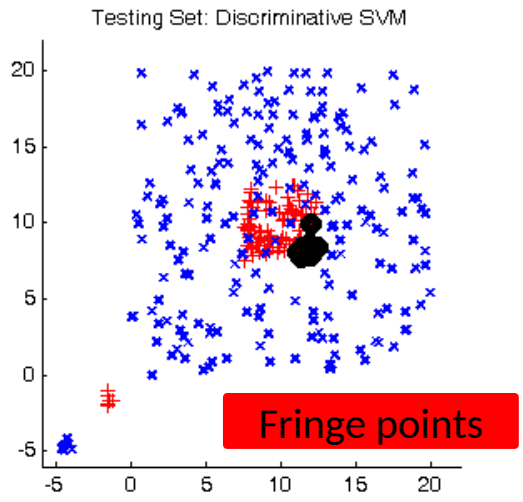
- TOP 10 Scoring anomalies for each method



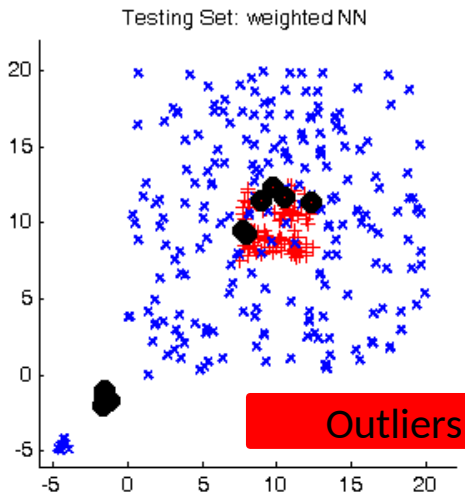




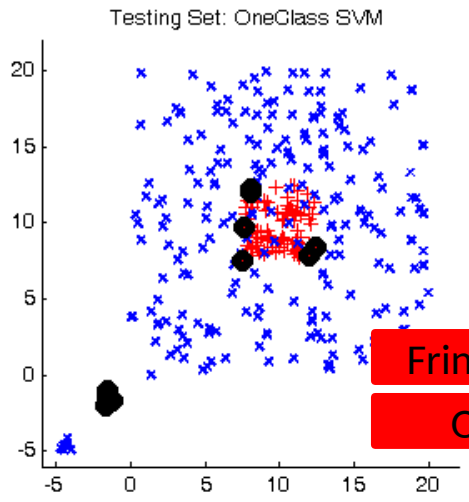
Connectivity AD



Fringe points



Outliers



Fringe points

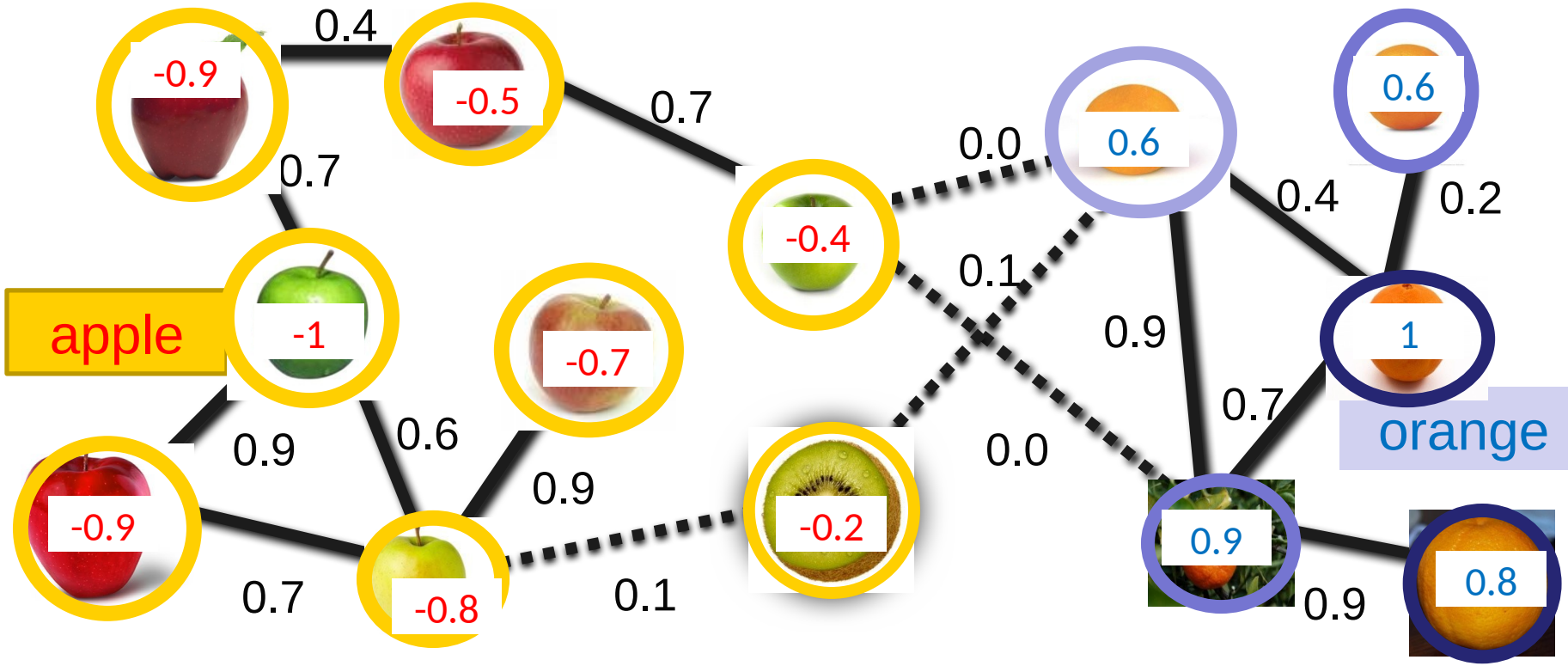
Outliers

ConnectivityAD summary

- Pros
  - Can handle
    - Fringe points
    - Unconditional Anomalies
  - Fast
    - O(n) memory,
    - O(|E|) time
- Cons
  - Calibration
  - Connectivity vs. Density

Harmonic Solution for SSL

[Zhu et al., 2003]



Dealing with Outliers

$$\min_f \sum V(f, \mathbf{x}_i, \mathrm{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$

$$\min_f \sum_{i:|\ell_i^*| \geq \varepsilon} V(f, \mathbf{x}_i, \mathrm{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$

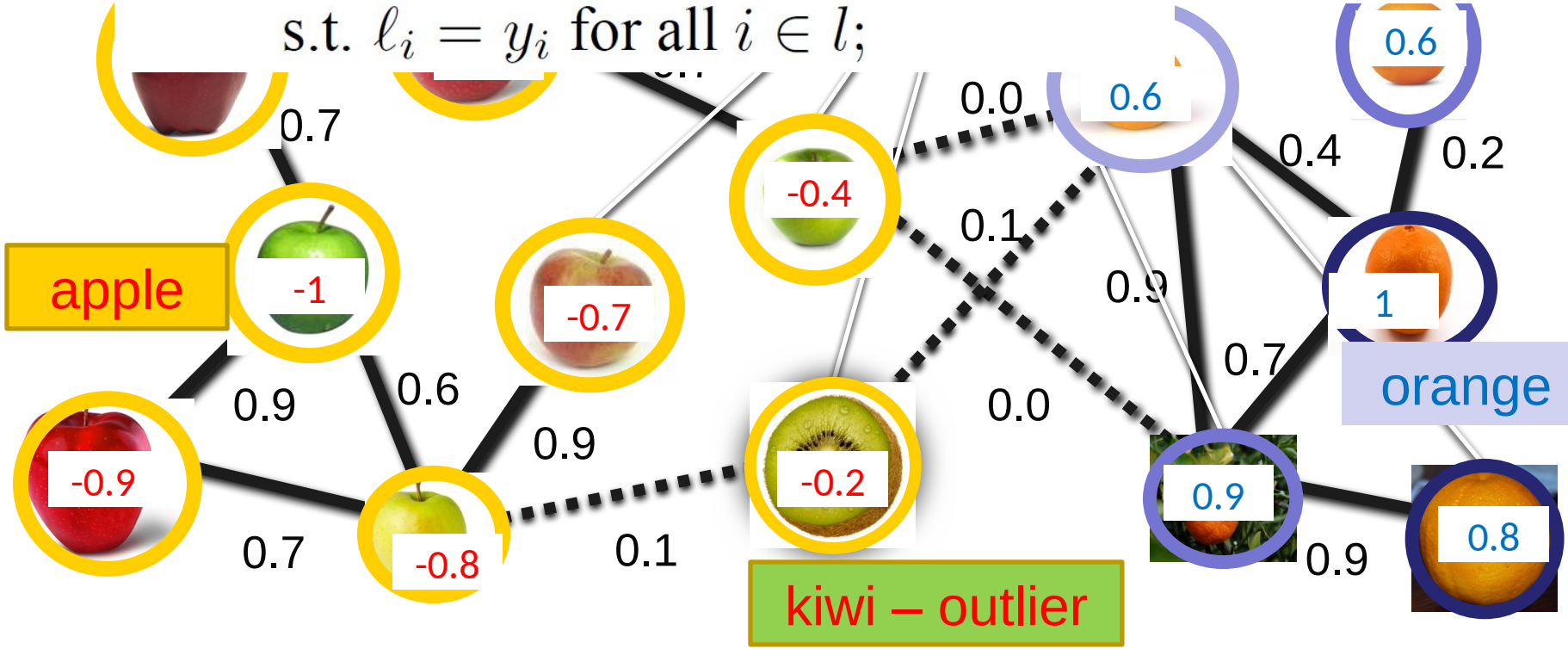
s

$$\gamma_g I + L)\ell$$

$$\text{s.t.} \quad \ell^* = \arg\min_{\ell} \ell^\top (\gamma_g I + L)\ell$$

$$i \in l;$$

$$\text{s.t. } \ell_i = y_i \text{ for all } i \in l;$$



Dealing with Outliers

$$\min_f \sum V(f, \mathbf{x}_i, \mathrm{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$

$$\min_f \sum_{i:|\ell_i^*| \geq \varepsilon} V(f, \mathbf{x}_i, \mathrm{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$

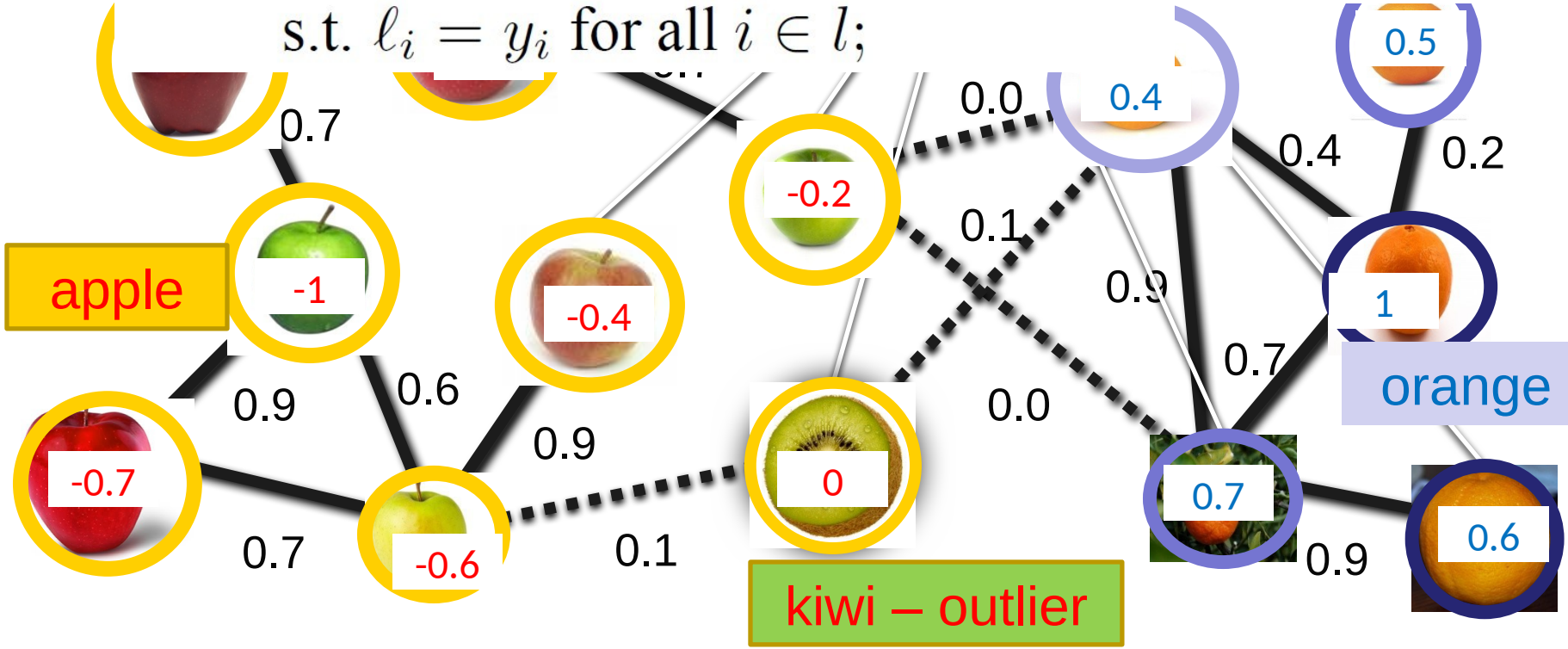
s

$$\gamma_g I + L)\ell$$

$$\text{s.t.} \quad \ell^* = \arg\min_{\ell} \ell^\top (\gamma_g I + L)\ell$$

$$i \in l;$$

$$\text{s.t. } \ell_i = y_i \text{ for all } i \in l;$$



Harmonic Solution

- A standard approach:

$$\min_{f \in \Re^N} \sum_{i,j} w_{ij} (f_i - f_j)^2 \quad \text{s.t.} f_i = y_i \text{ for all } i \in l$$

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

$$\min_{f \in \Re^N} f^{\mathrm{T}} L f \quad \text{s.t.} f_i = y_i \text{ for all } i \in l$$

- Properties of the solution:
  - Smoothness $f_i = \frac{1}{d_i} \sum_{j \sim i} w_{ij} f_j$
  - Computable in a closed form $f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$
  - Interpretable as a random walk on the graph W with the transition matrix $P = D^{-1} W$

Regularized HS

- To control the confidence of labeling unlabeled examples, we compute the regularized HS:

$$\min_{f \in \Re^N} f^{\mathrm{T}}(L + \gamma_g I) f \quad \text{s.t. } f_i = y_i \text{ for all } i \in l$$

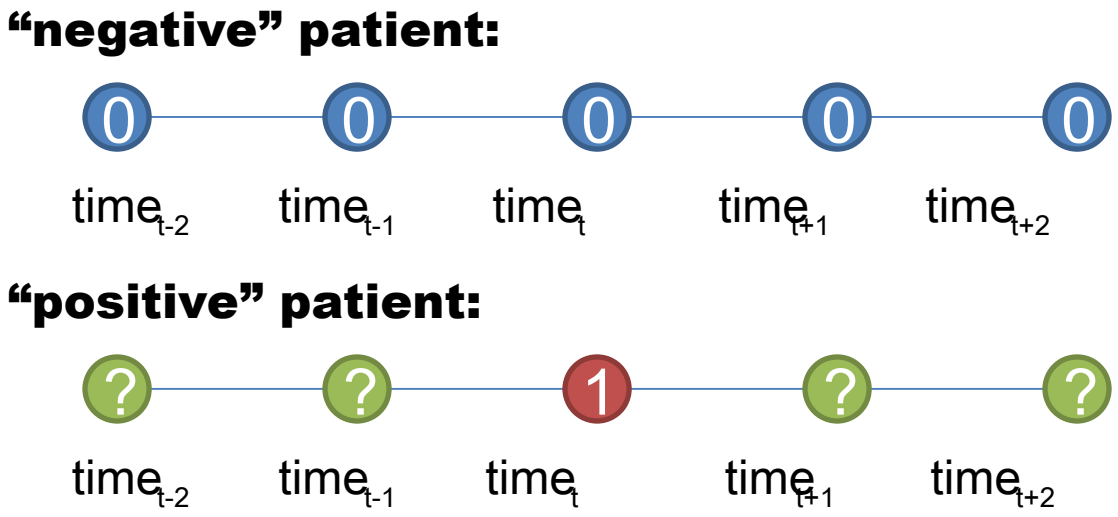
- Properties of the solution:
  - Computable in a closed form $f_u = (L_{uu} + \gamma_g I)^{-1} W_{ul} f_l$
  - Interpretable as a random walk on the graph W with an extra sink. At every step, the walk may terminate at the sink with probability $\gamma_g / (d_i + \gamma_g)$
  - A neat way of filtering outliers and fringe points

Regularized HS for CAD

$$
\begin{aligned}
\ell_i &= (I - P_{uu})^{-1}_{iu} P_{ul} \boldsymbol{\ell}_l \\
&= \underbrace{\sum_{j:y_j=1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^1} - \underbrace{\sum_{j:y_j=-1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^{-1}} \\
&= p_i^1 - p_i^{-1}
\end{aligned}
$$

- 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!

SSL for medical data



- missing data
- unknown labels in nearby times

Soft Harmonic Solution

- Unconstrained Regularization

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

fit to data

regularizer

$$K = L + \gamma_g I$$

- Close form solution

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

- Why Soft Harmonic Function for Conditional AD?

Why Online Learning?

- Data arrive in the streams
- Need to adapt (Train ≠ Test)
  - Eg. Medical practices change
- Online Learning naturally fits the problem

Online HFS

Inputs: an example x_t, a data adjacency graph W

Algorithm:

What is wrong with this algorithm?

Add x_t to the graph W and compute the Laplacian L

O(t)

Infer labels on the graph:

$$\min_{\mathbf{l}\in\Re^N} \mathbf{l}^{\mathrm{T}}(L+\gamma_g I)\mathbf{l} \quad \text{s.t.}\, \mathbf{l}_i = y_i \text{ for all } i \in l$$

O(t^3)

Predict $\hat{y}_t = \mathbf{l}_t$

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

Online HFS

Inputs: an example x_t, a data adjacency graph W

Algorithm:

If the graph W has more than M vertices, quantize it

Add x_t to the graph W and compute the Laplacian L

Infer labels on the graph:

$$\min_{\mathbf{l}\in\Re^N} \mathbf{l}^{\mathrm{T}}(L+\gamma_g I)\mathbf{l} \quad \text{s.t.}\, \mathbf{l}_i = y_i \text{ for all } i \in l$$

O(M)

Predict $\hat{y}_t = \mathbf{l}_t$

O(M^3)

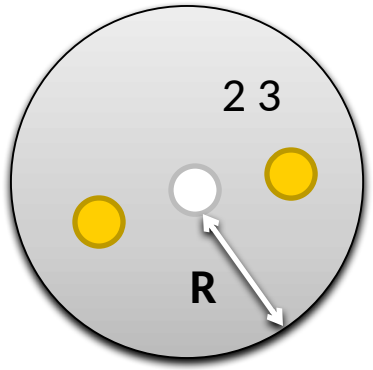
Outputs: a prediction $\hat{y}_t$, an updated data adjacency graph W

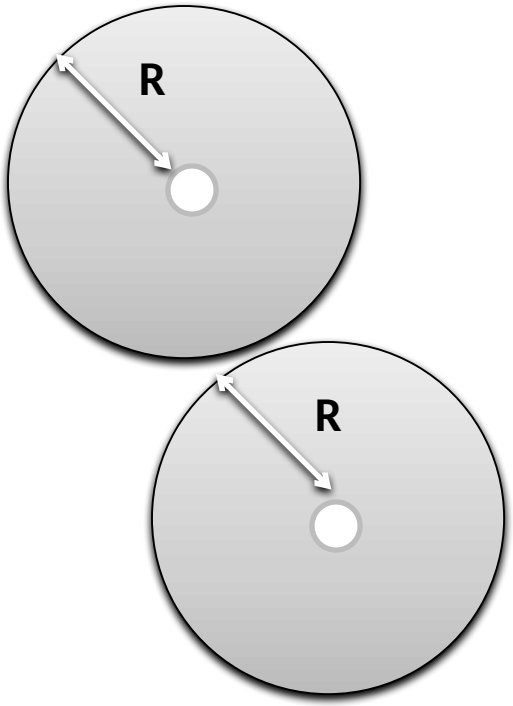
Why Online Learning?

- Data arrive in the streams
- Need to adapt (Train ≠ Test)
  - Eg. Medical practices change
- Online Learning naturally fits the problem
- Complexity issues
  - $\Omega(n^2)$ in general - construction of an $n \times n$ matrix
  - inverse operation on an $n \times n$ matrix - $O(n^{2.4})$
- What about Nyström?

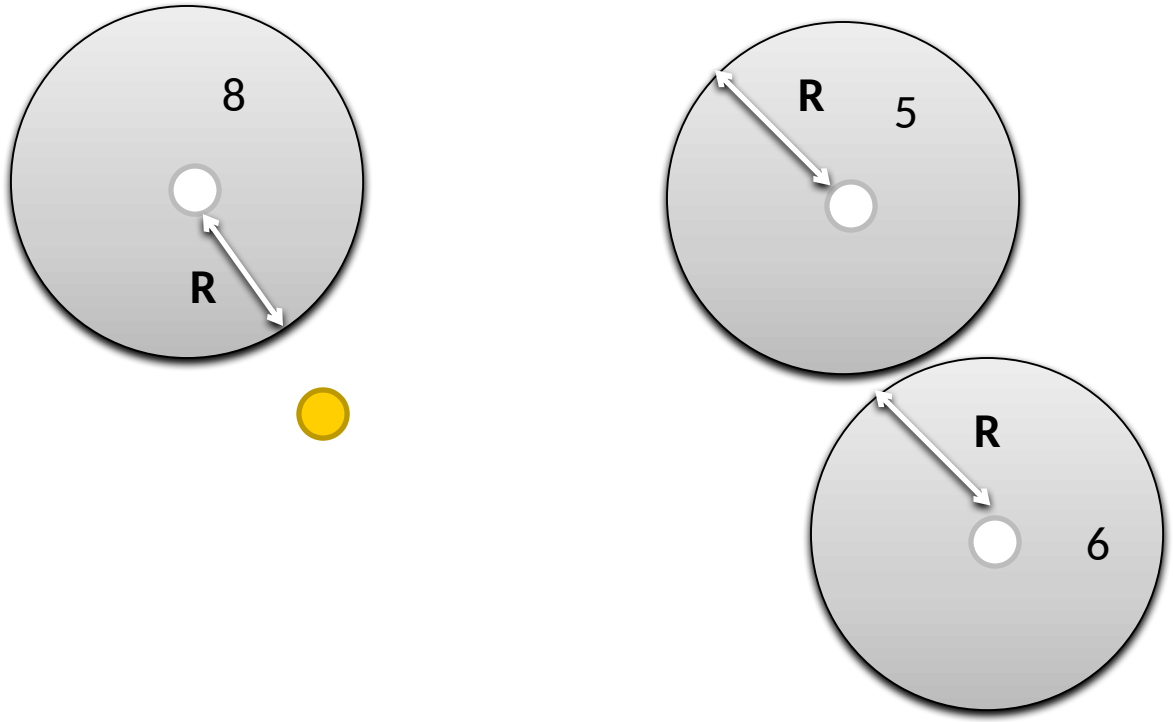
Incremental k-centers

[Charikar et al., 1997]

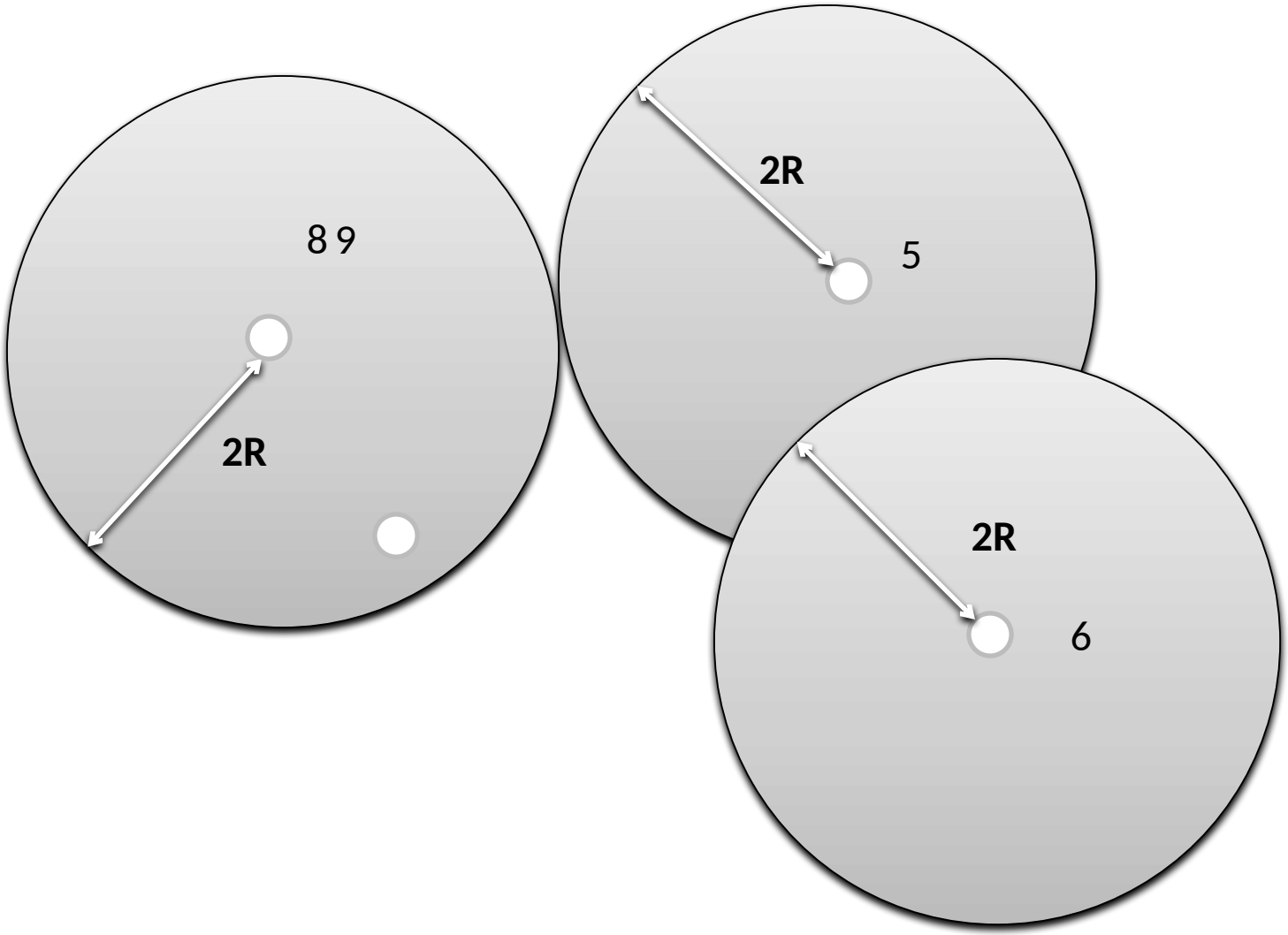




Incremental k-centers



Incremental k-centers



Theoretical Guarantees

- We seek a regret bound of the form:

$$\frac{1}{N}\sum_t (\hat{y}_t - y_t)^2 \quad \leq \quad \frac{1}{N}\sum_t (y_t^* - y_t)^2 + \frac{1}{N}\sum_t (y_t' - y_t^*)^2 + \frac{1}{N}\sum_t (\hat{y}_t - y_t')^2$$

Online learning risk

Offline learning error

Online learning error

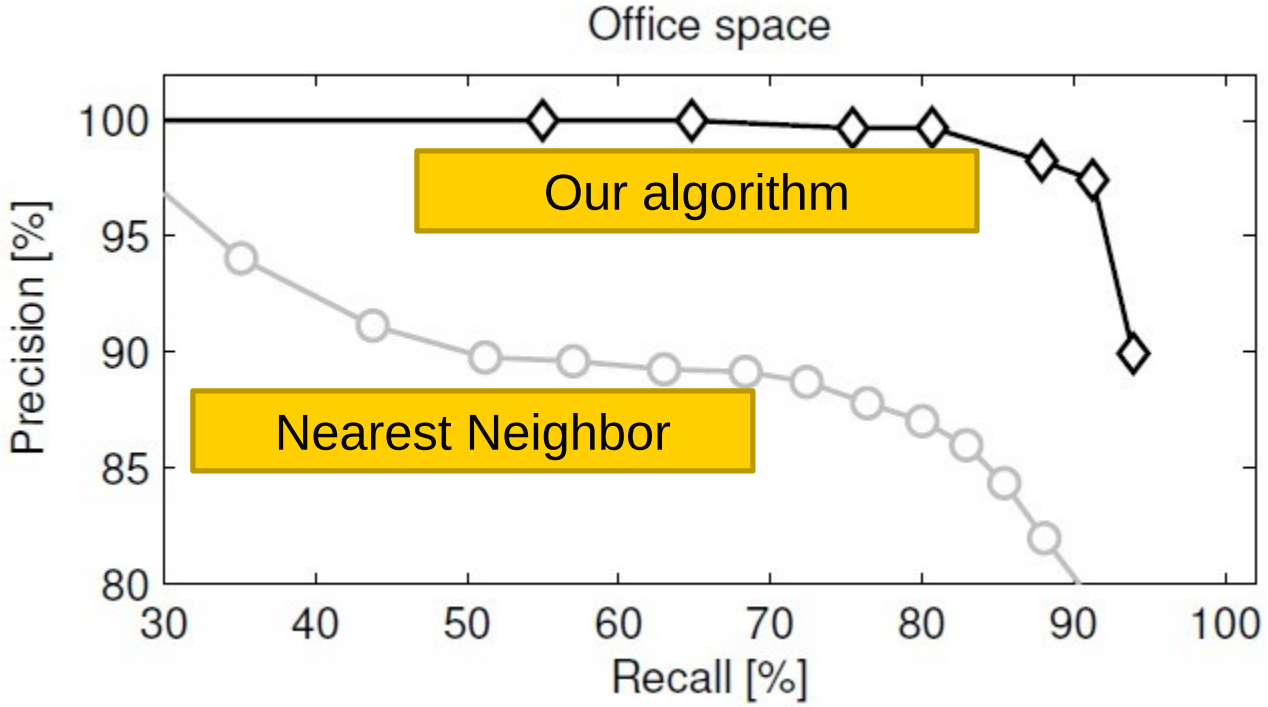
Quantization error

- The errors should be bounded on the order of $\mathrm{O}\left(\sqrt{N}\right)$

SSL experiments

- Office space

- When a person shows up on the camera for the first time, we label four faces of the person.



Medical Dataset

- Biased selection of cases
- Distance metric:
- 3-NN graph

$$w_{ij} = \exp\left[-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2_{2,\psi}}{p\sigma^2}\right]$$

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

	SVM-AD	HAD
ROC - AUC	64.12%	66.68%

Future Work

- Online Soft Harmonic Anomaly Detection
- Quantization with multiple Classes
- Theory for Anomaly Detection
- New human expert evaluation
- Incremental clustering with forgetting the history
- Parallelization of harmonic solution

Thanks to:
Branislav Kveton, Greg Cooper, Tomas Singliar, Shyam Visweswaram, Iyad Batal, Hamed Valizadegan, Quang Nguyen