Conditional Anomaly Detection with Adaptive Similarity Metric

Michal Valko

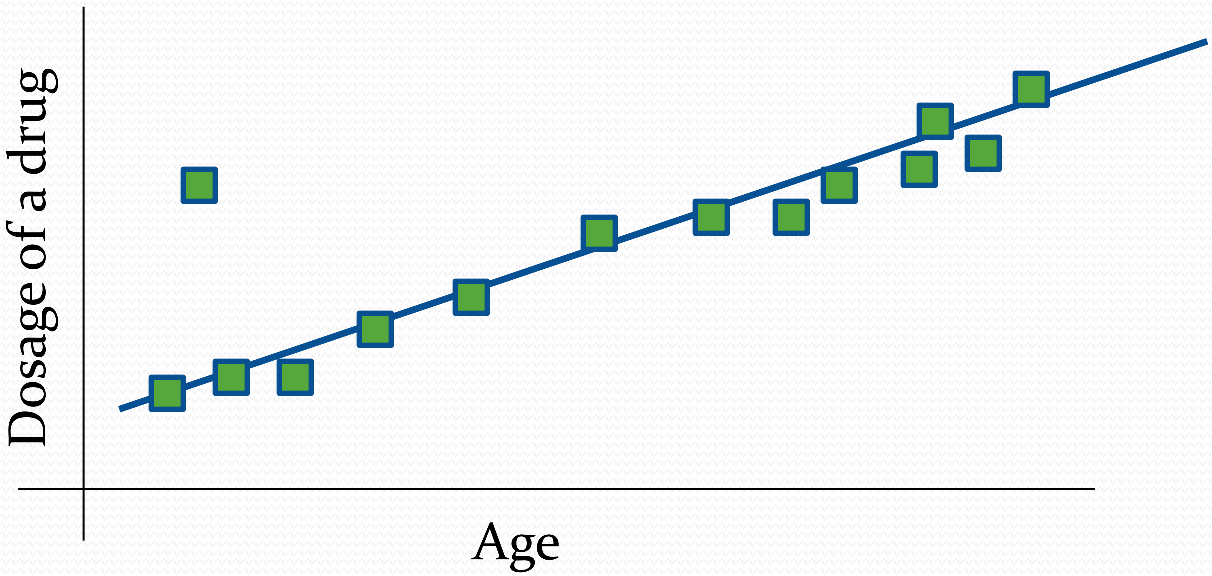
Advisor: Miloš Hauskrecht
Joint work with: Gregory Cooper, Amy Seybert, Shyam Visweswaram, Melissa Saul, James Harrison, Andrew Post

Funded by the National Library of Medicine: Grant R21-LM009102

Anomaly Detection

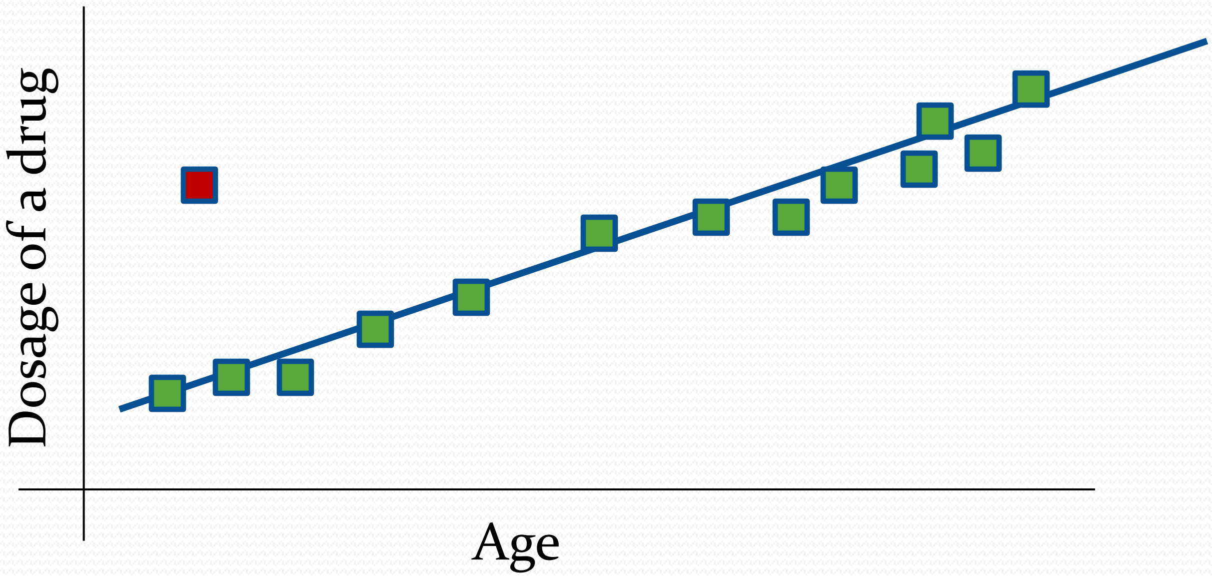
- Goal: Identify unusual patterns in data.
- Methods: from statistics and machine learning
- Contribution: conditional anomaly detection framework
- Application: medical error detection

Conditional Anomaly



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

Conditional Anomaly



Assumption: Anomalies correspond to medical errors

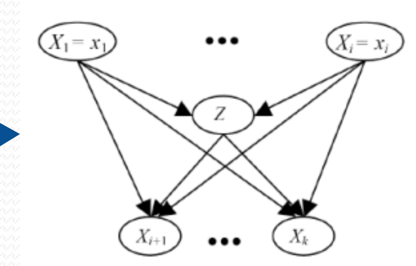
"Medical errors account for 200 000 preventable deaths a year. "

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

Hauskrecht, Valko, Kveton, Visweswaram, Cooper (AMIA 2007)

Medical Database

Group of similar patients

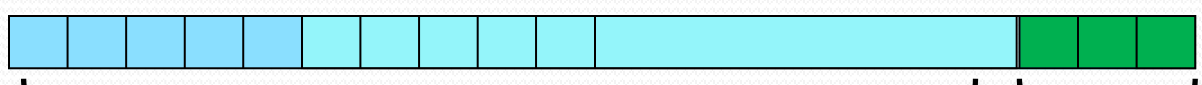


Model

P(Decisions | Description , Model) < α ?

Anomaly Call

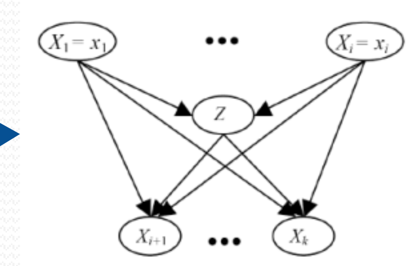
Current patient record



Description (Contex) + Decision(s)

Medical Database

Group of similar patients

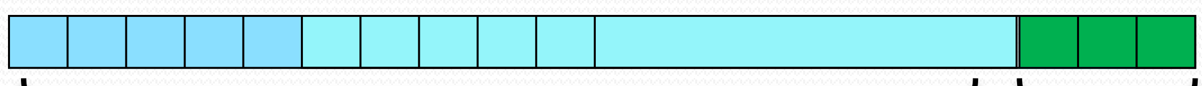


Model

P(Decisions | Description , Model) < α ?

Anomaly Call

Current patient record



Description (Contex) + Decision(s)

Selecting Similar Patients

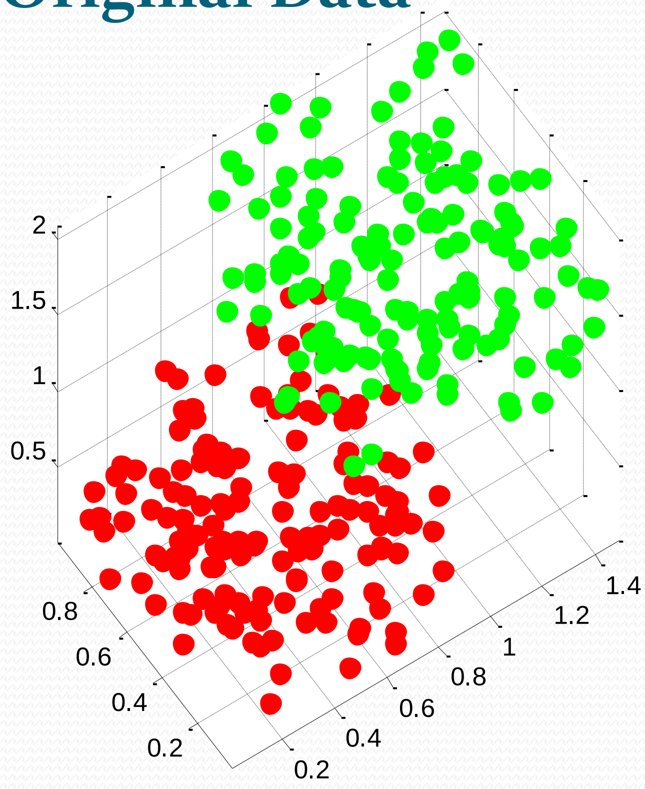
- All other patients in the database
- Select only the closest patients
- What is a good distance metric?
  - Euclidean, Mahalanobis …
    - don’t take into the account the decision variables

- Learn the metric which puts patients with the similar decisions closer together.

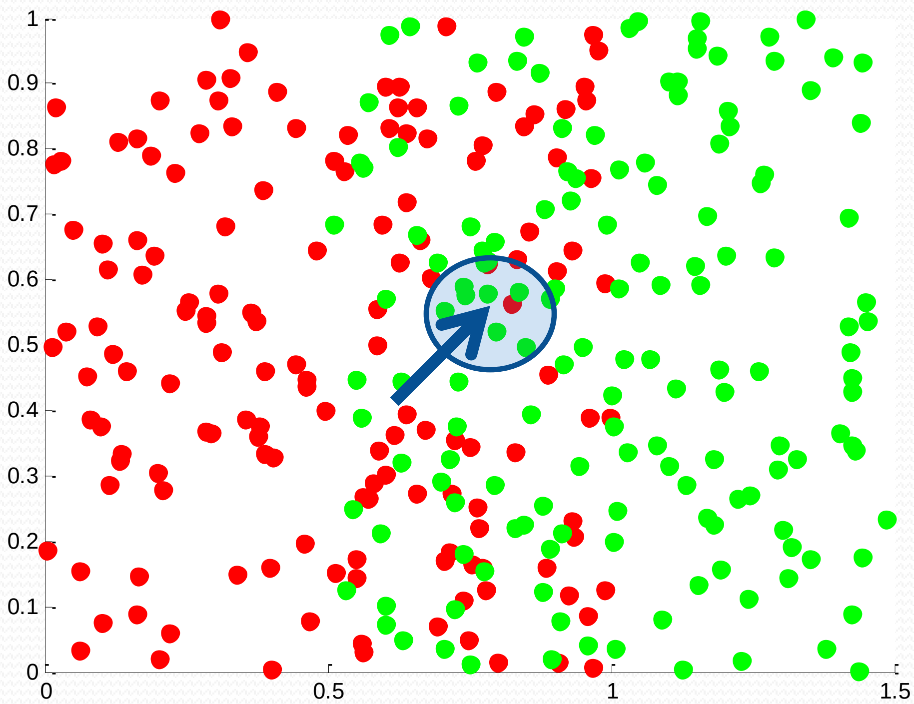
Valko, Hauskrecht (FLAIRS 2008, to appear)

Neighborhood Component Analysis

Original Data



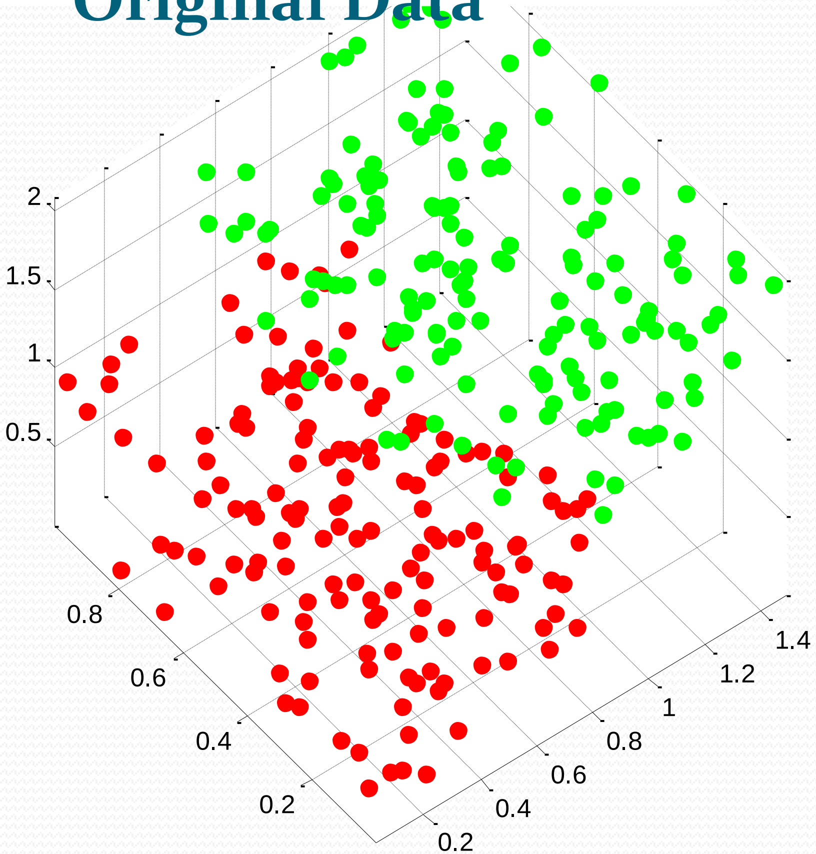
Original Linear Projection



Goldberger et al. NIPS2004

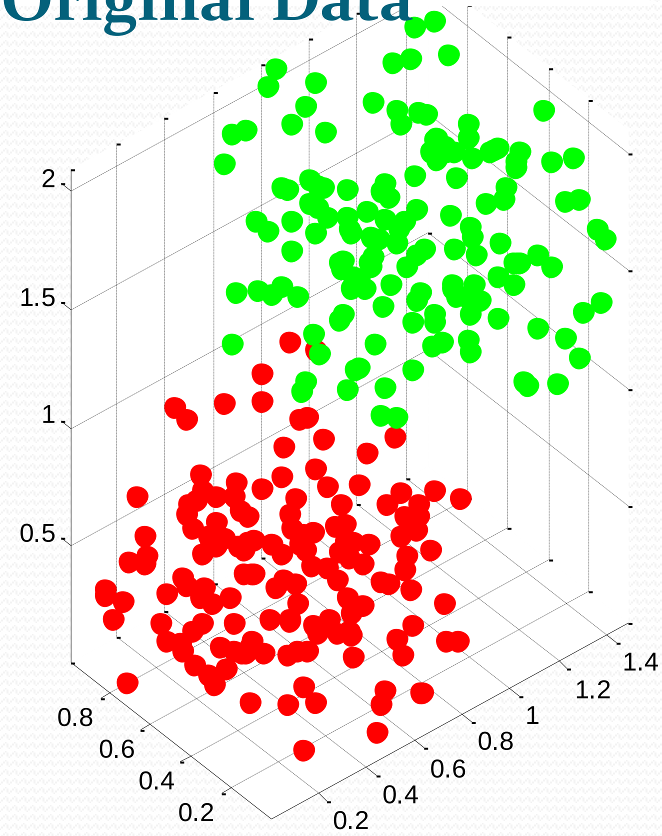
Neighborhood Component Analysis

Original Data



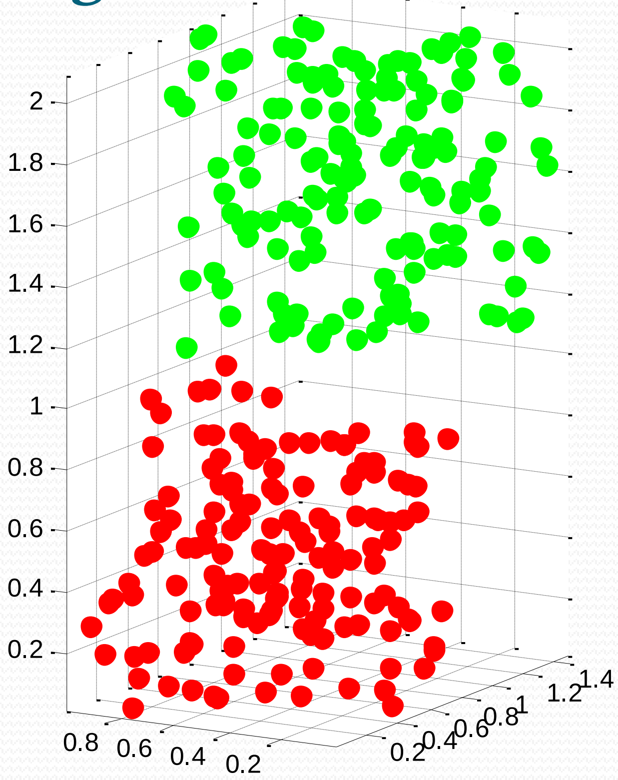
Neighborhood Component Analysis

Original Data

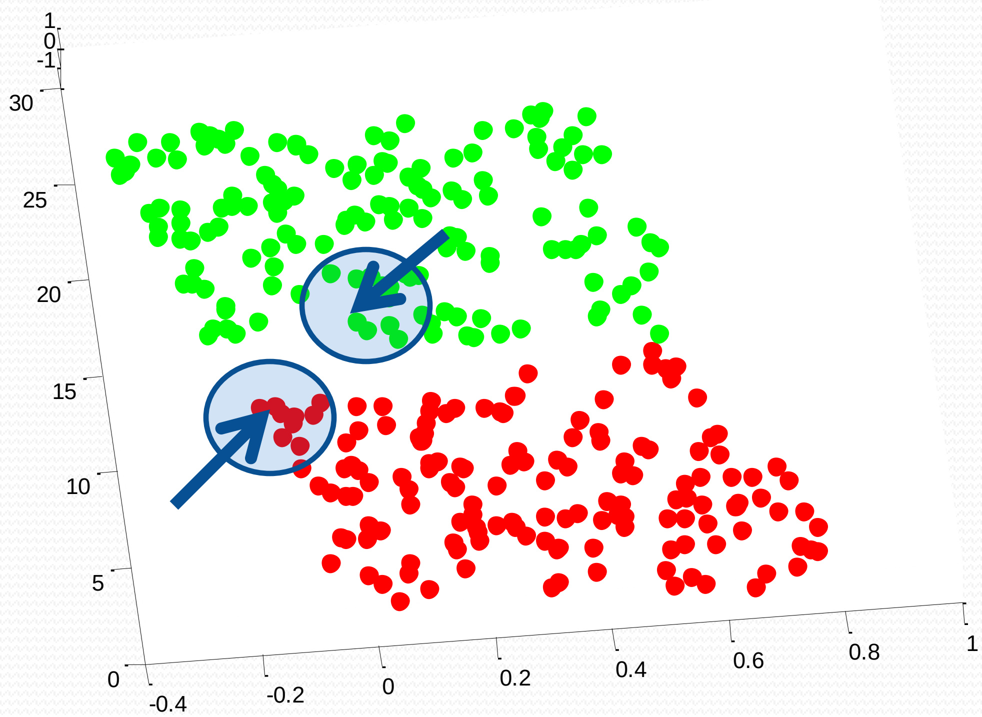


Neighborhood Component Analysis

Original Data

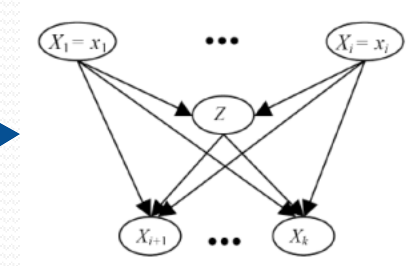


Learned Linear Projection



Medical Database

Group of similar patients



Model

Current patient record

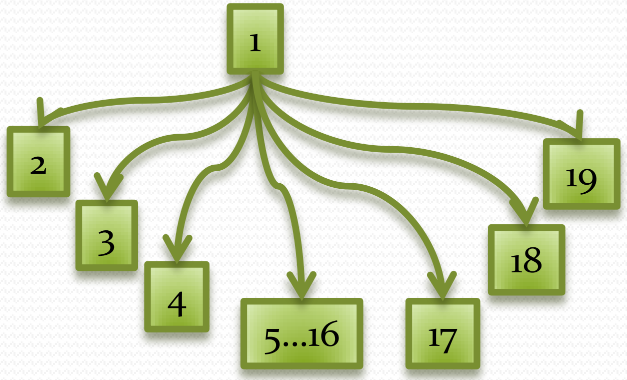
P(Decisions | Description , Model) < α ?

Anomaly Call

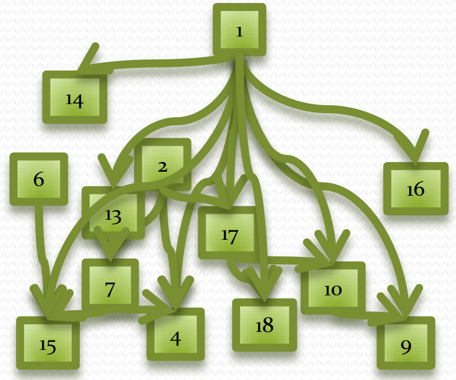
Description (Contex) + Decision(s)

Learn Probabilistic Model

- Bayesian Network with Fixed structure



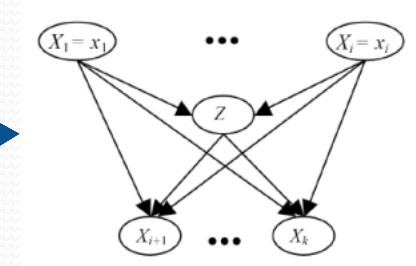
- Learn the Bayesian Network structure and parameters from the data



Eaton & Murphy, UAI 2007

Medical Database

Group of similar patients

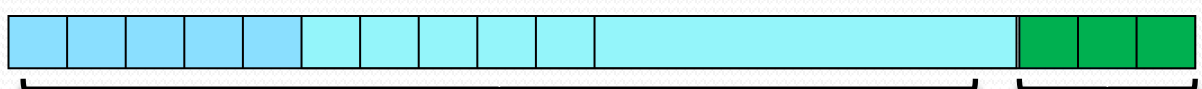


Model

$P(\text{Decisions} \mid \text{Description}, \text{Model}) < \alpha$?

Anomaly Call

Current patient record



Description (Contex) + Decision(s)

Experiments

- PORT dataset (Kapoor 1996)
- Patients diagnosed with the community acquired **pneumonia**

	Target attributes
X_1	Hospitalization
	Prediction attributes
	Demographic factors
X_2	Age > 50
X_3	Gender (male = true, female = false)
	Coexisting illnesses
X_4	Congestive heart failure
X_5	Cerebrovascular disease
X_6	Neoplastic disease
X_7	Renal disease
X_8	Liver disease
	Physical-examination findings
X_9	Pulse ≥ 125 / min
X_{10}	Respiratory rate ≥ 30 / min
X_{11}	Systolic blood pressure < 90 mm Hg
X_{12}	Temperature $< 35\,^\circ$C or $\geq 40\,^\circ$C
	Laboratory and radiographic findings
X_{13}	Blood urea nitrogen ≥ 30 mg / dl
X_{14}	Glucose ≥ 250 mg / dl
X_{15}	Hematocrit $< 30\%$
X_{16}	Sodium < 130 mmol / l
X_{17}	Partial pressure of arterial oxygen < 60 mm Hg
X_{18}	Arterial pH < 7.35
X_{19}	Pleural effusion

Experiments

- 2287 patient cases
- 19 binary attributes
- 100 evaluated by the panel of three physicians
- 23 anomalies

	Target attributes
X_1	Hospitalization
	Prediction attributes
	Demographic factors
X_2	Age > 50
X_3	Gender (male = true, female = false)
	Coexisting illnesses
X_4	Congestive heart failure
X_5	Cerebrovascular disease
X_6	Neoplastic disease
X_7	Renal disease
X_8	Liver disease
	Physical-examination findings
X_9	Pulse ≥ 125 / min
X_{10}	Respiratory rate ≥ 30 / min
X_{11}	Systolic blood pressure < 90 mm Hg
X_{12}	Temperature $< 35\,^{\circ}$C or $\geq 40\,^{\circ}$C
	Laboratory and radiographic findings
X_{13}	Blood urea nitrogen ≥ 30 mg / dl
X_{14}	Glucose ≥ 250 mg / dl
X_{15}	Hematocrit $< 30\%$
X_{16}	Sodium < 130 mmol / l
X_{17}	Partial pressure of arterial oxygen < 60 mm Hg
X_{18}	Arterial pH < 7.35
X_{19}	Pleural effusion

Experiments

- Goal: Detect whether the decision of hospitalization is *anomalous*, conditioning on the description variables

Target attributes	
X_1	Hospitalization
Prediction attributes	
	Demographic factors
X_2	Age > 50
X_3	Gender (male = true, female = false)
	Coexisting illnesses
X_4	Congestive heart failure
X_5	Cerebrovascular disease
X_6	Neoplastic disease
X_7	Renal disease
X_8	Liver disease
	Physical-examination findings
X_9	Pulse ≥ 125 / min
X_{10}	Respiratory rate ≥ 30 / min
X_{11}	Systolic blood pressure < 90 mm Hg
X_{12}	Temperature $< 35\,^\circ$C or $\geq 40\,^\circ$C
	Laboratory and radiographic findings
X_{13}	Blood urea nitrogen ≥ 30 mg / dl
X_{14}	Glucose ≥ 250 mg / dl
X_{15}	Hematocrit $< 30\%$
X_{16}	Sodium < 130 mmol / l
X_{17}	Partial pressure of arterial oxygen < 60 mm Hg
X_{18}	Arterial pH < 7.35
X_{19}	Pleural effusion

Evaluation

- Algorithm catches many anomalies
  - high sensitivity
- Algorithm’s predictions are accurate
  - high specificity

- Combine sensitivity and specificity for various detection thresholds

Results

MODEL	METRIC	SELECTION	RESULT	
Naïve Bayes	any	ALL	11.6%	BASELINE
	Euclidean	CLOSEST 40	16.4%	
	Learned Metric	CLOSEST 40	16.8%	
Learn Bayes Network Structure and Parameters	any	ALL	13.8%	
	Euclidean	CLOSEST 40	17.8%	
	Learned Metric	CLOSEST 40	26.4%	BEST

Conclusion: Two-fold improvement over baseline.

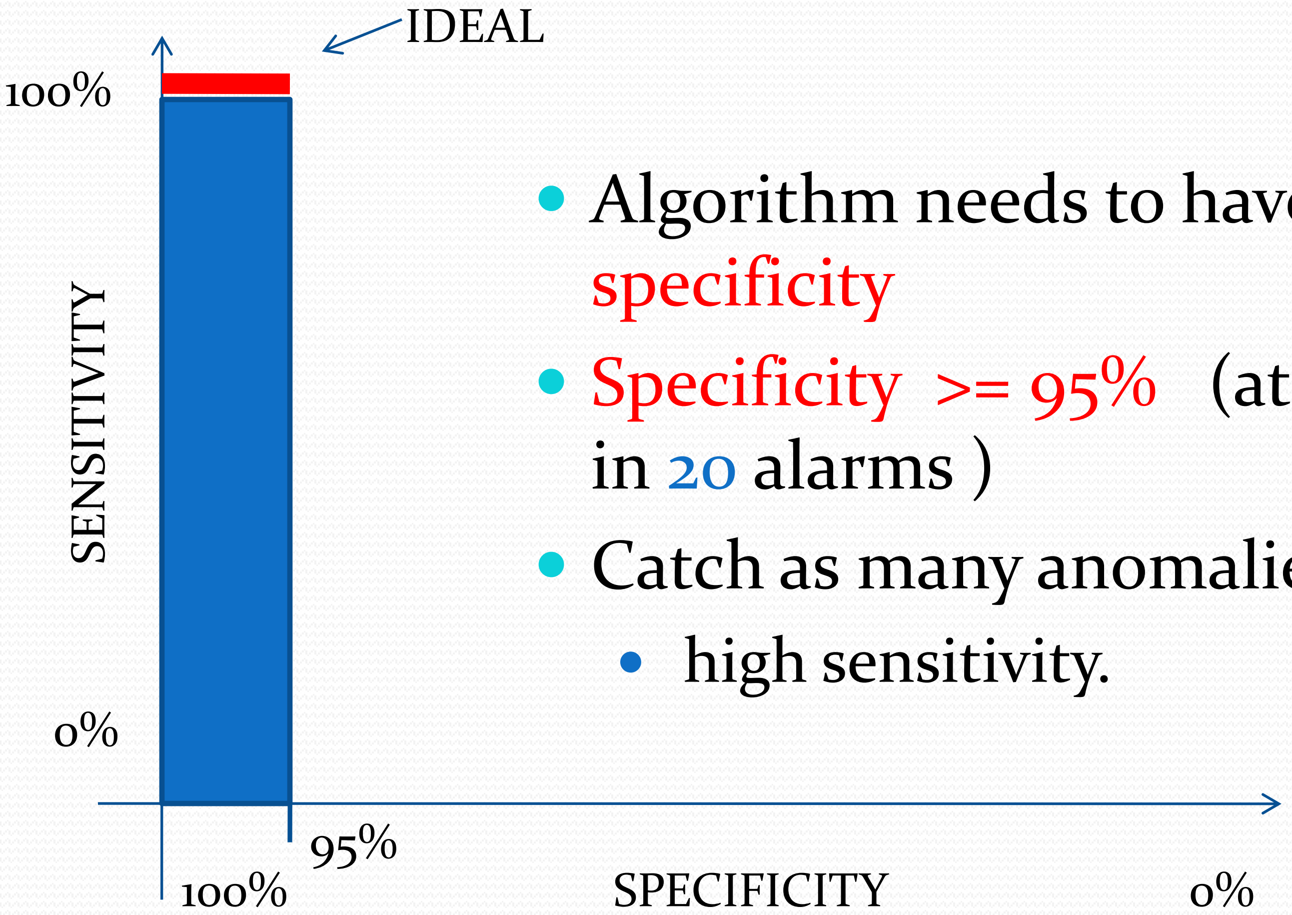
Conclusion

- Selection of closest patients
  - Models tuned to the individual patient
- Metric learning
  - Lowers the influence of irrelevant data
- Structure learning
  - Gives more accurate representation of relation between the variables

Current/Future Work

- Automatic population size selection
- Multiple decisions
- UPMC dataset of patients with cardiac surgery with **thousands** of records per patient
- Anomaly detection in time.

Evaluation



- Algorithm needs to have high specificity
- Specificity >= 95% (at most 1 error in 20 alarms)
- Catch as many anomalies
  - high sensitivity.

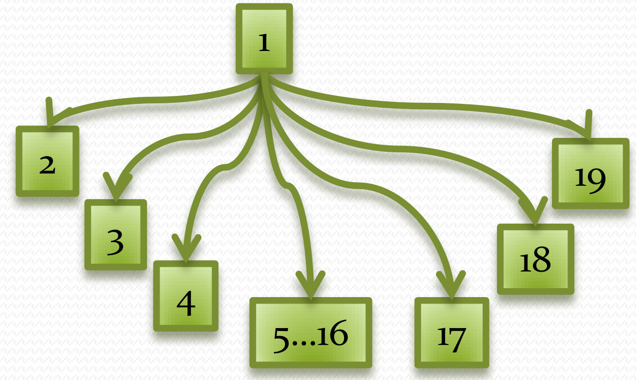
Neighborhood Component Analysis

$$\|Ax_i - Ax_j\|^2$$

$$\sum_{j \in C_i} p_{ij}$$

Learn Probabilistic Model

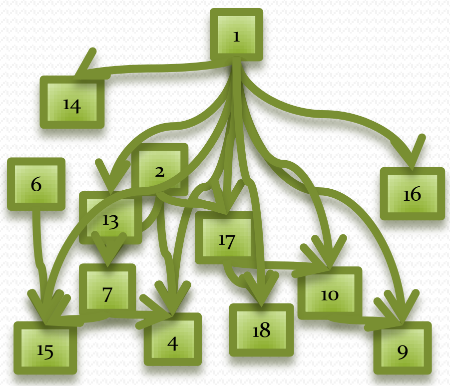
- Bayesian Network with Fixed structure



- Probabilities from metric

$$p_{ij} = \frac{\exp(-\|Ax_i - Ax_j\|^2)}{\sum_{k \neq i} \exp(-\|Ax_i - Ax_k\|^2)} \quad , \quad p_{ii} = 0$$

- Learn the Bayesian Network structure and parameters from the data



Eaton & Murphy, UAI 2007

Results

MODEL	METRIC	SELECTION	RESULT	
Naïve Bayes	any	ALL	11.6%	BASELINE
	Euclidean	CLOSEST 40	16.4%	
	Learned Metric	CLOSEST 40	16.8%	
Probability from the Distance Metric	Euclidean	ALL	8.0%	
	Euclidean	CLOSEST 40	8.0%	
	Learned Metric	ALL	18.0%	
	Learned Metric	CLOSEST 40	20.2%	
Learn Bayes Network Structure and Parameters	any	ALL	13.8%	
	Euclidean	CLOSEST 40	17.8%	
	Learned Metric	CLOSEST 40	26.4%	BEST

Conclusion: Two-fold improvement over baseline.