Relational Differential Prediction

Houssam Nassif[1] Vítor Santos Costa[2]
Elizabeth S. Burnside[1] David Page[1]

[1]University of Wisconsin, Madison, USA

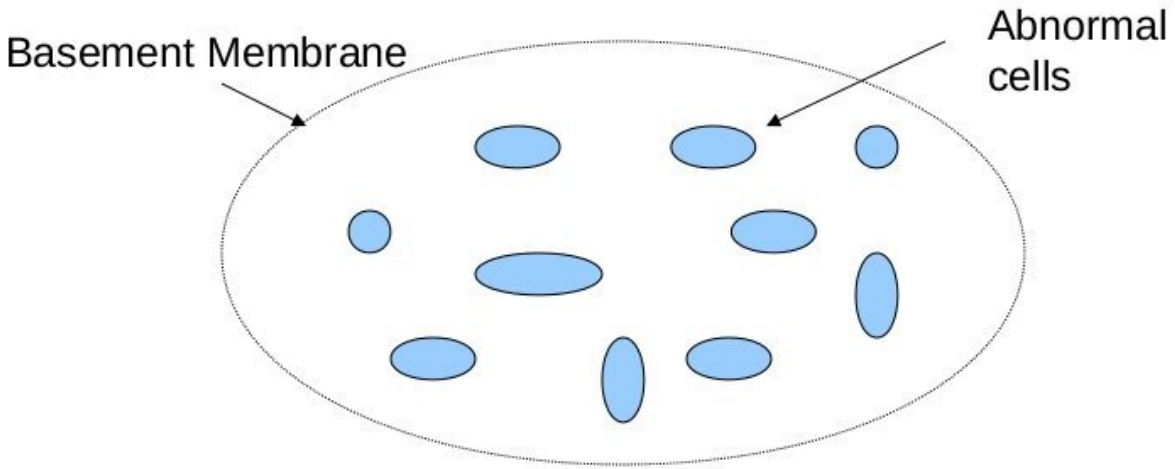
[2]University of Porto, Portugal

ECML'12

Outline

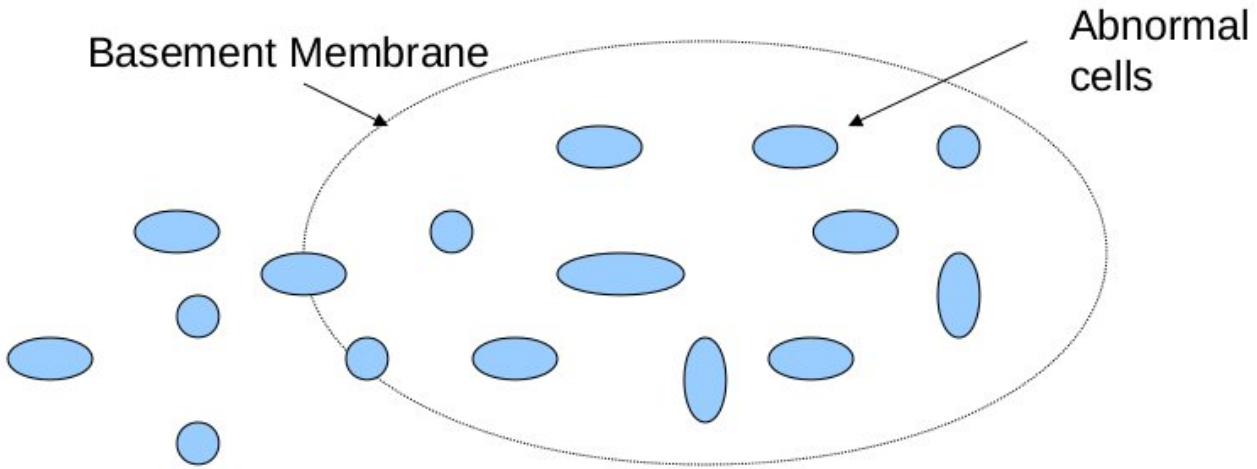
Outline

Breast-Cancer Stages



In-Situ Cancer Stage

Breast-Cancer Stages



Invasive Cancer Stage

Cancer Stage Features

- In Situ can develop into Invasive
  - Current practice: Always treat In Situ
- Time to spread may be very long
  - Patient may die of other causes
  - Over-diagnosis (and over-treatment)

- What features characterize In Situ in older patients?
- What features change between older and younger?

Cancer Stage Features

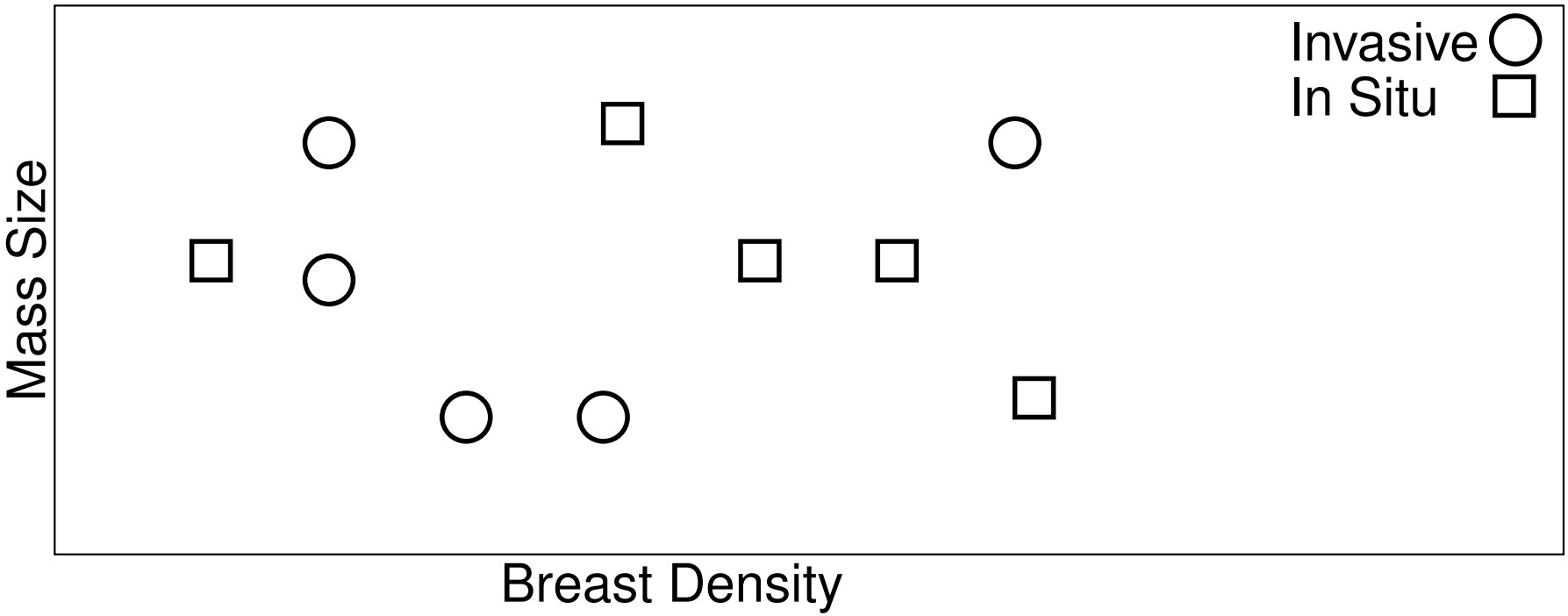
- In Situ can develop into Invasive
  - Current practice: Always treat In Situ
- Time to spread may be very long
  - Patient may die of other causes
  - Over-diagnosis (and over-treatment)

- What features characterize In Situ in older patients?
- What features change between older and younger?

Differential Prediction

Definition

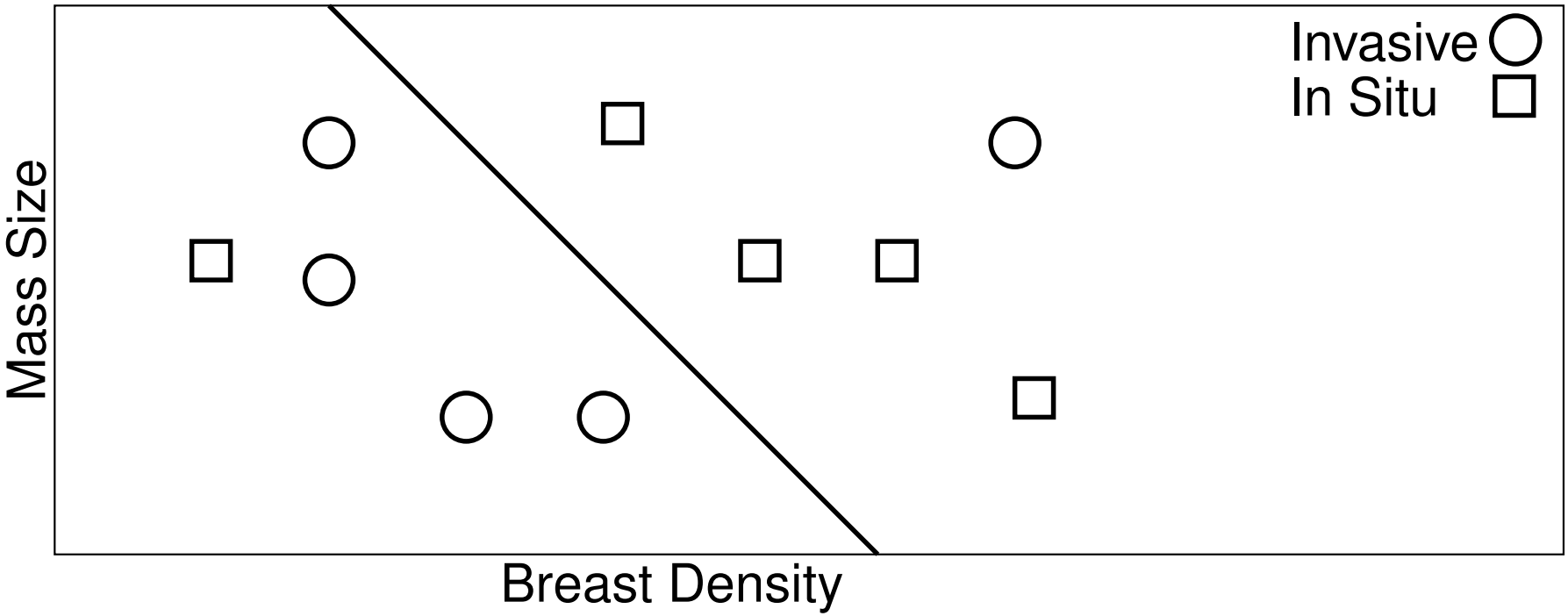
Differential Prediction (DP): Classifier exhibits significant performance differences over particular instance subgroups



Differential Prediction

Definition

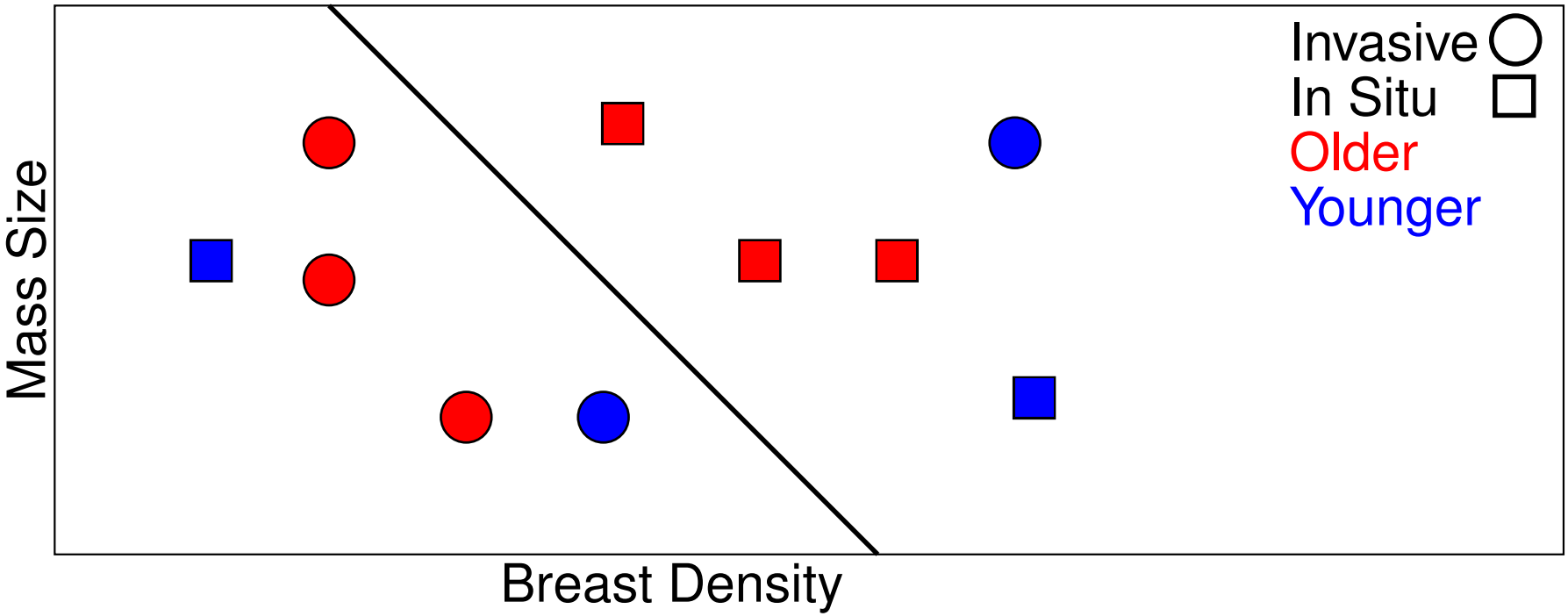
Differential Prediction (DP): Classifier exhibits significant performance differences over particular instance subgroups



Differential Prediction

Definition

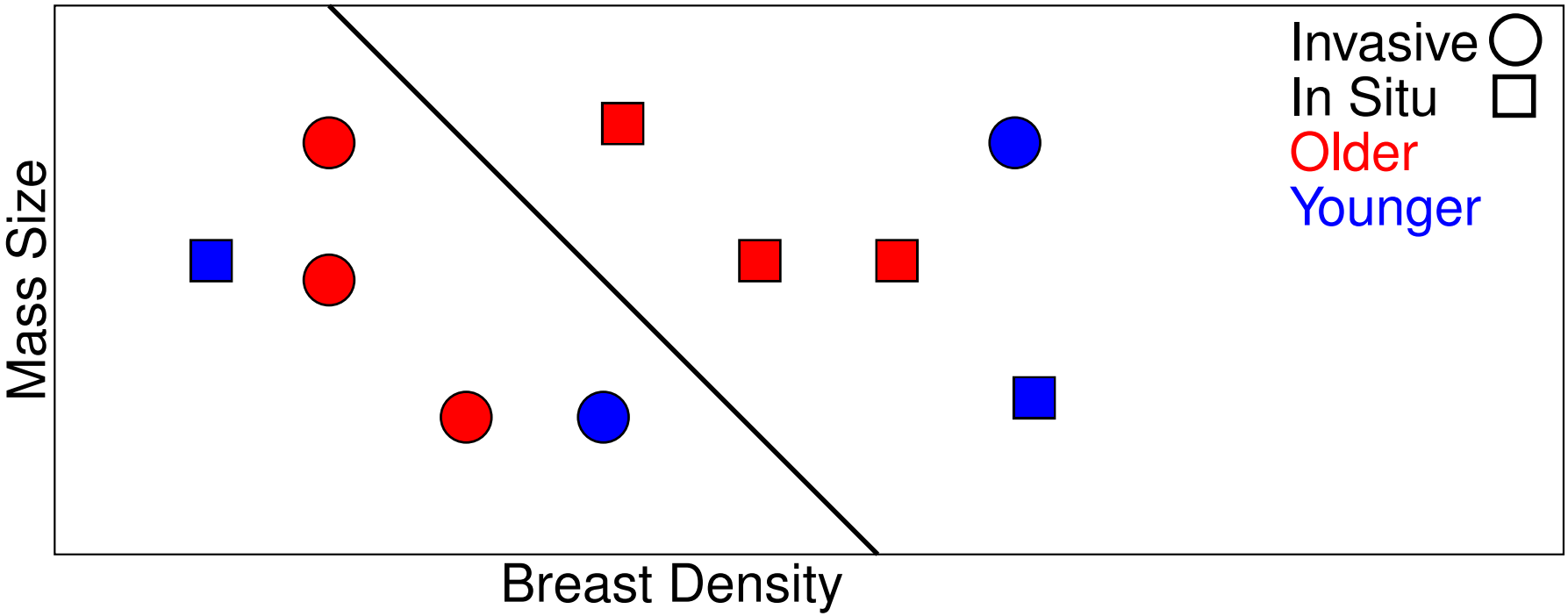
Differential Prediction (DP): Classifier exhibits significant performance differences over particular instance subgroups



Differential Prediction

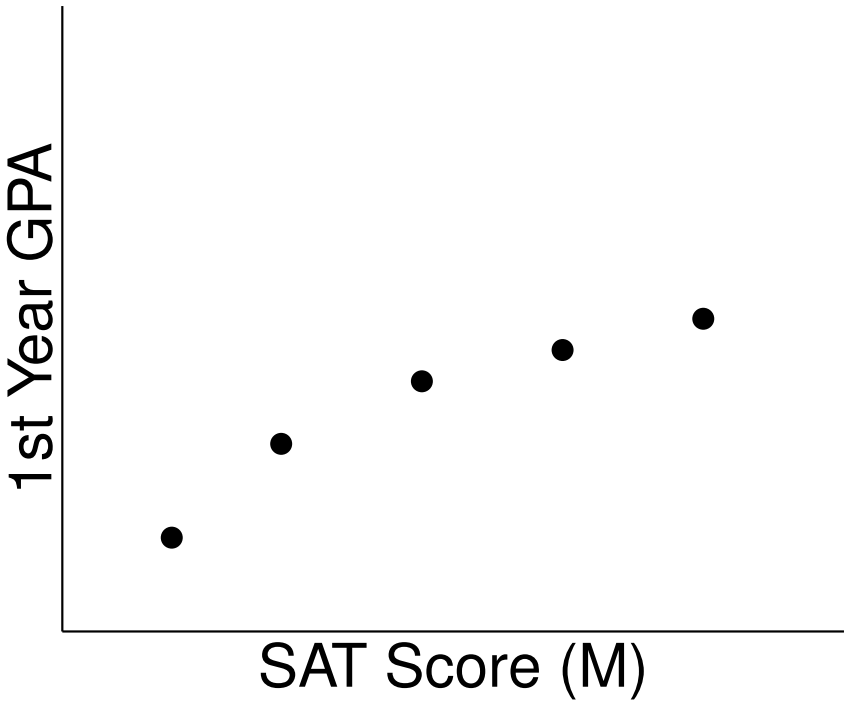
Definition

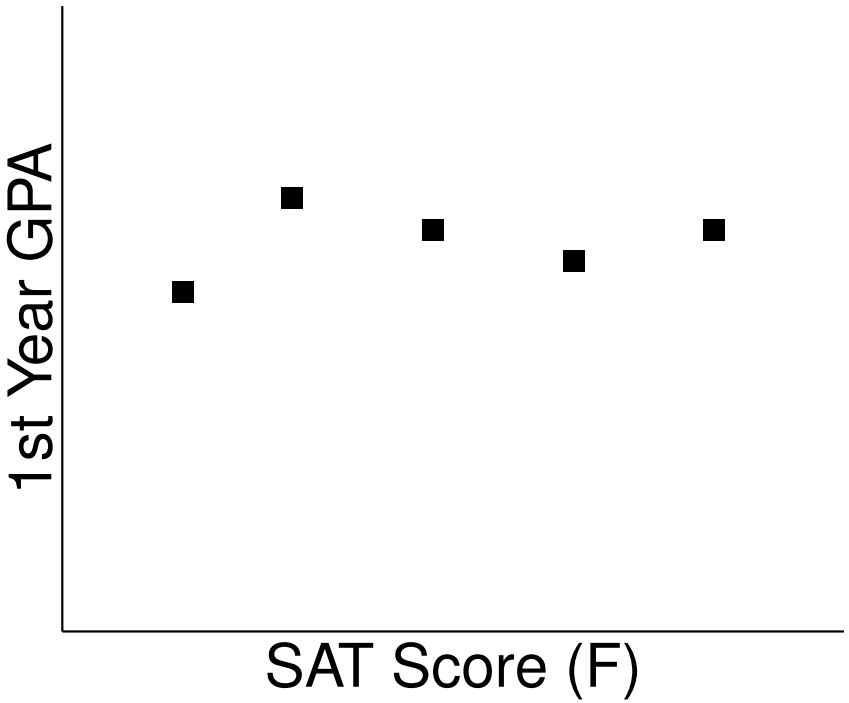
Differential Prediction (DP): Classifier exhibits significant performance differences over particular instance subgroups



Using Regression to Detect DP

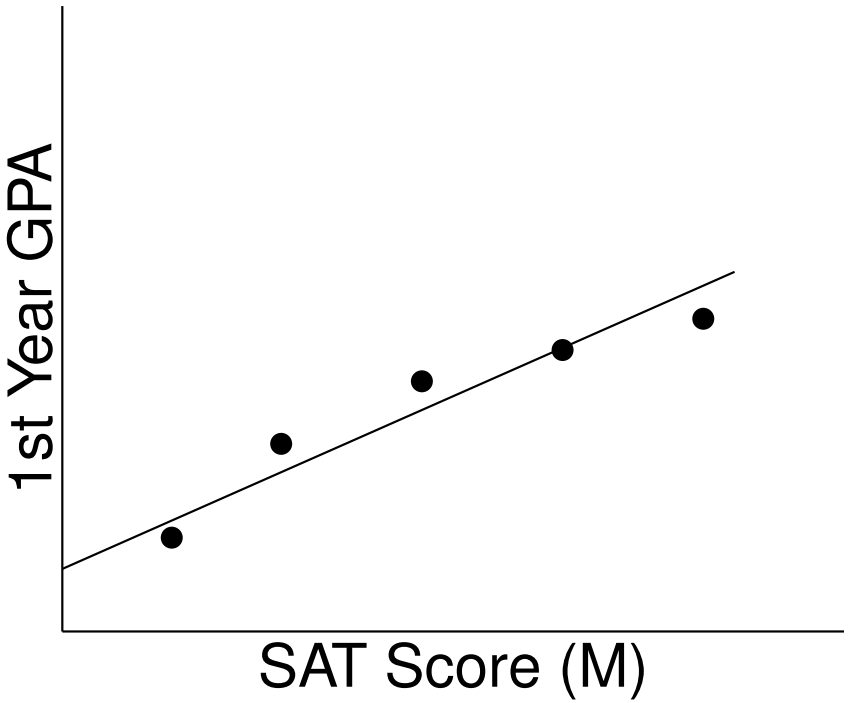
- Validate educational and psychological tests
- Detect discrepancies related to race or gender

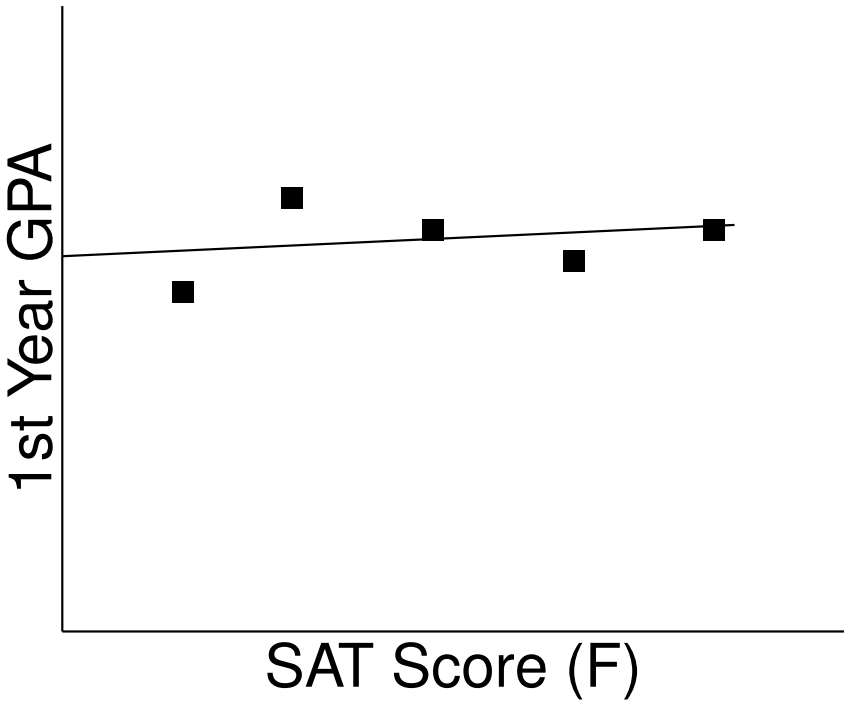




Using Regression to Detect DP

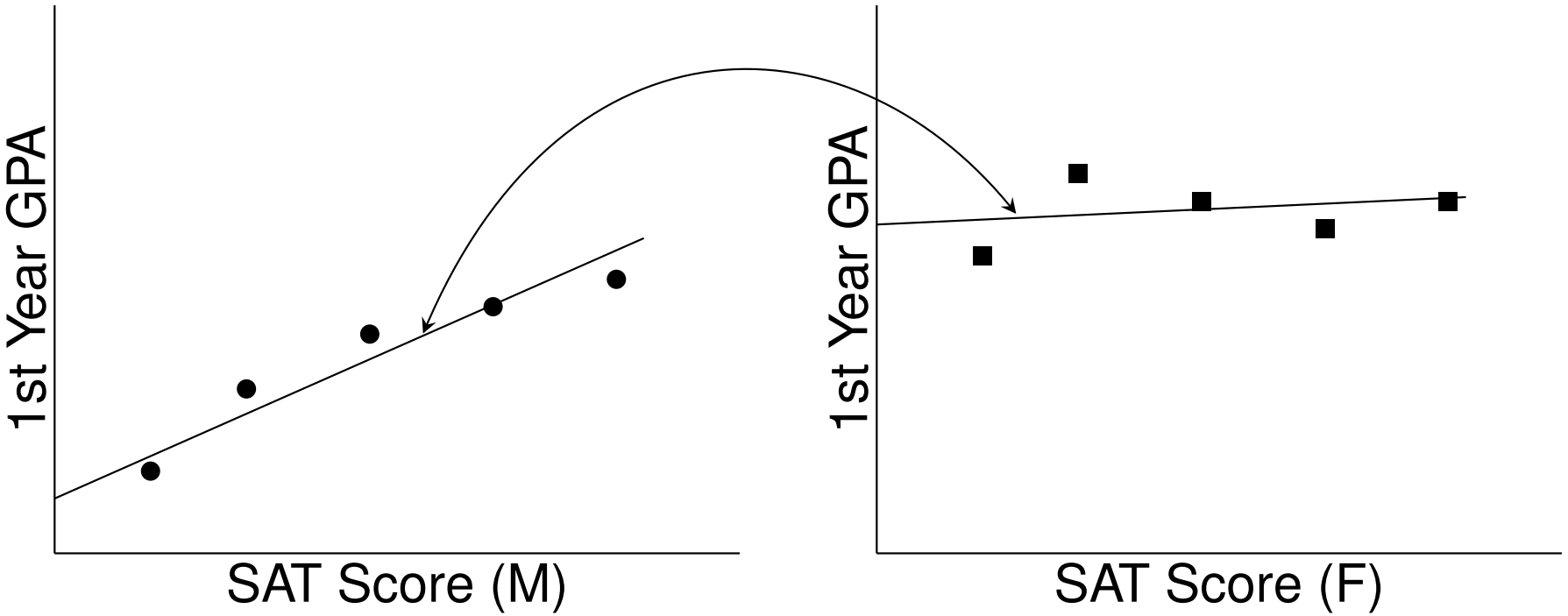
- Validate educational and psychological tests
- Detect discrepancies related to race or gender





Using Regression to Detect DP

- Validate educational and psychological tests
- Detect discrepancies related to race or gender



DP in Machine Learning

- Detected by:
  - Comparing classifiers built on distinct data subgroups
  - Checking classifier performance on multiple subgroups
- Related to:
  - Uplift Modeling in marketing
  - Subgroup Discovery
  - Differential misclassification cost

Aim

- Classifier to maximize DP over given data subsets
- Extend DP to relational sets
- Insight into DP features

DP in Machine Learning

- Detected by:
  - Comparing classifiers built on distinct data subgroups
  - Checking classifier performance on multiple subgroups
- Related to:
  - Uplift Modeling in marketing
  - Subgroup Discovery
  - Differential misclassification cost

Aim

- Classifier to maximize DP over given data subsets
- Extend DP to relational sets
- Insight into DP features

Stratified Dataset

Stratified Dataset

Strata are disjoint, each strata should contain at least one example of each target class

Definition (**Stratified Dataset**)

Let tc be a target class defined over the set of instances X, and let $D = \{\langle x, tc(x)\rangle\}$ be a set of examples labeled according to tc. Let $\{D_1, \ldots, D_n\}$ be n disjoint subsets of D, and let D_i^l be the set of examples of D_i with class label l, such that:

$$(\forall (i,j) \in [1,n],\ i \neq j)\ D_i \subset D,\ D_j \subset D,\ D_i \cap D_j = \emptyset, \quad (1)$$

$$\forall (i,l)\ D_i^l \neq \emptyset. \quad (2)$$

Differential Predictive Rule

Differential Predictive Rule

Given a stratified data, a rule whose performance is significantly better over one stratum as compared to the others

Definition (**Differential Predictive Rule**)

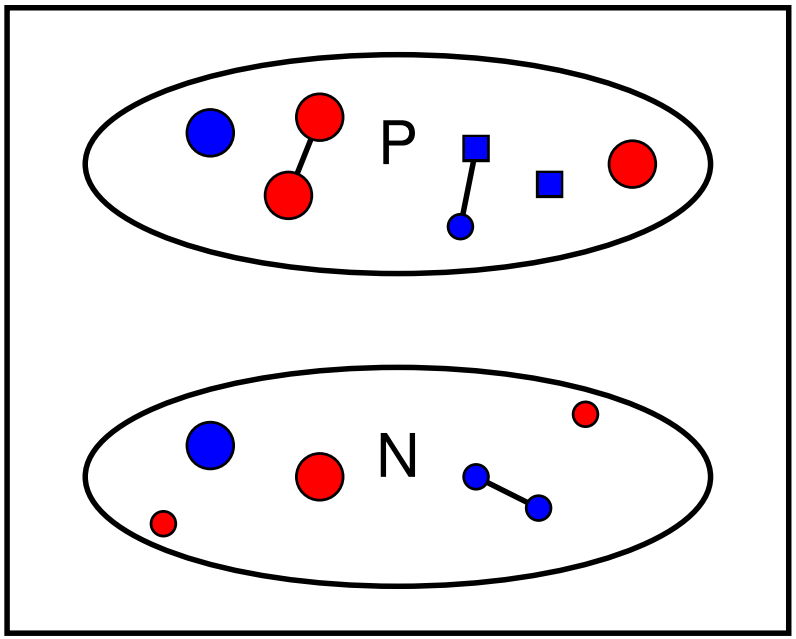
Let c be a rule over the set of instances X, and let $\mathcal{D}$ be a stratified dataset. Let $S(c|D_i)$ be the classification performance score for c over the subset D_i. A **stratum-j specific differential predictive rule** is a rule c_j such that:

$$\forall i \neq j,\ S(c_j|D_j) \gg S(c_j|D_i). \tag{3}$$

- Often want $S(c_j|D_j)$ to achieve a good performance

Outline

Inductive Logic Programming Example

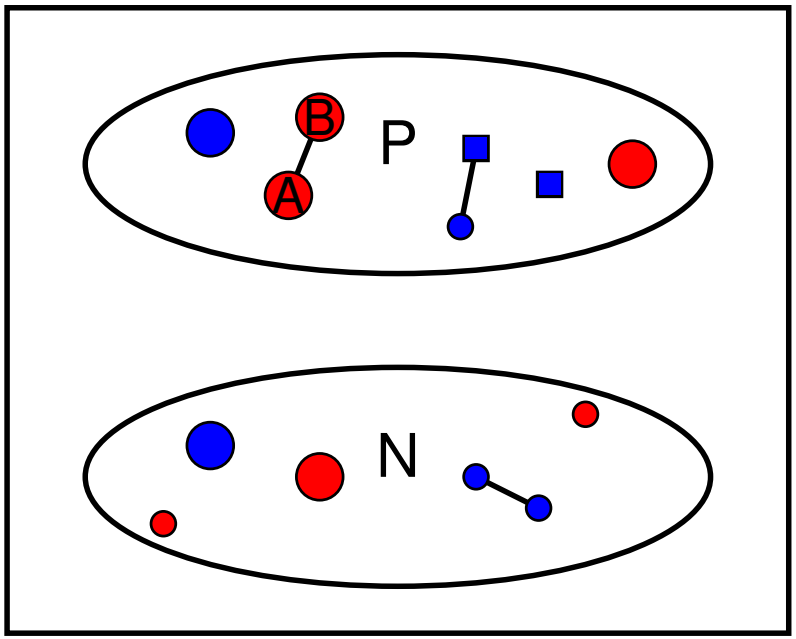


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

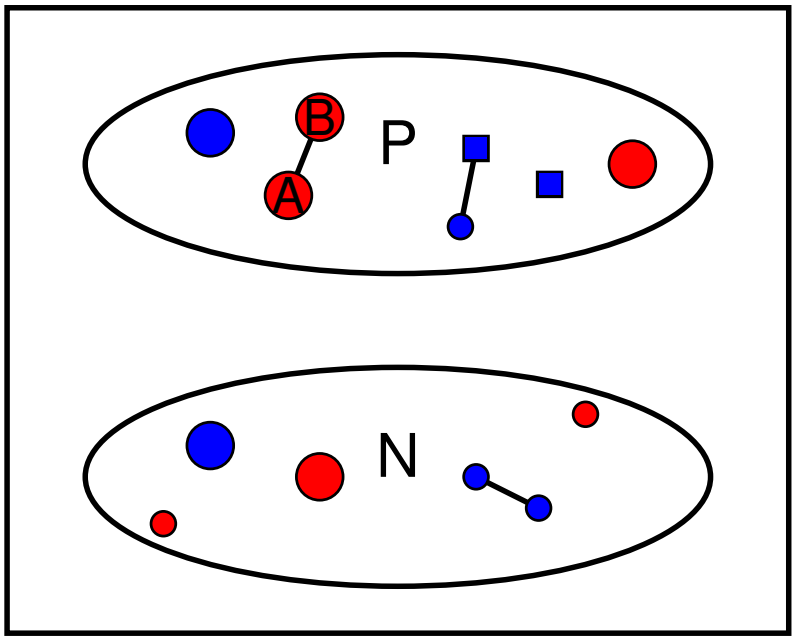


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

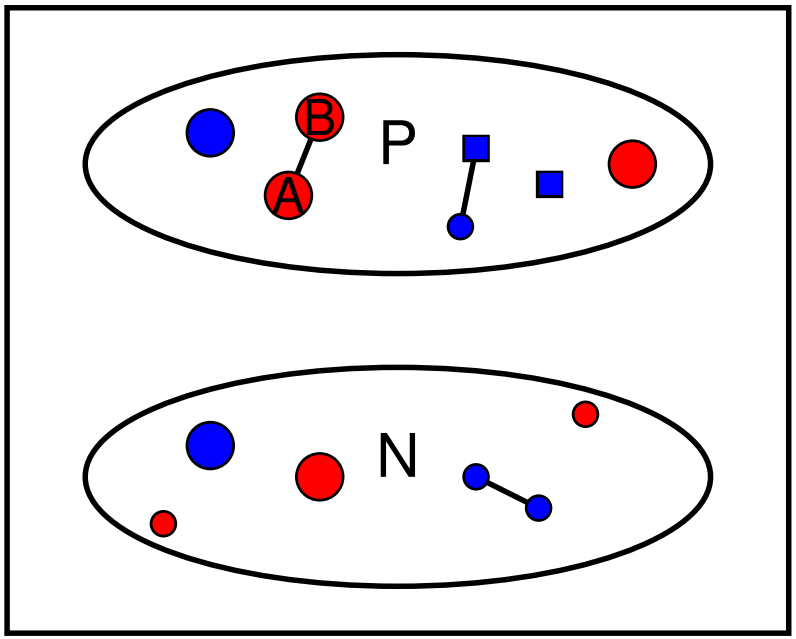


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

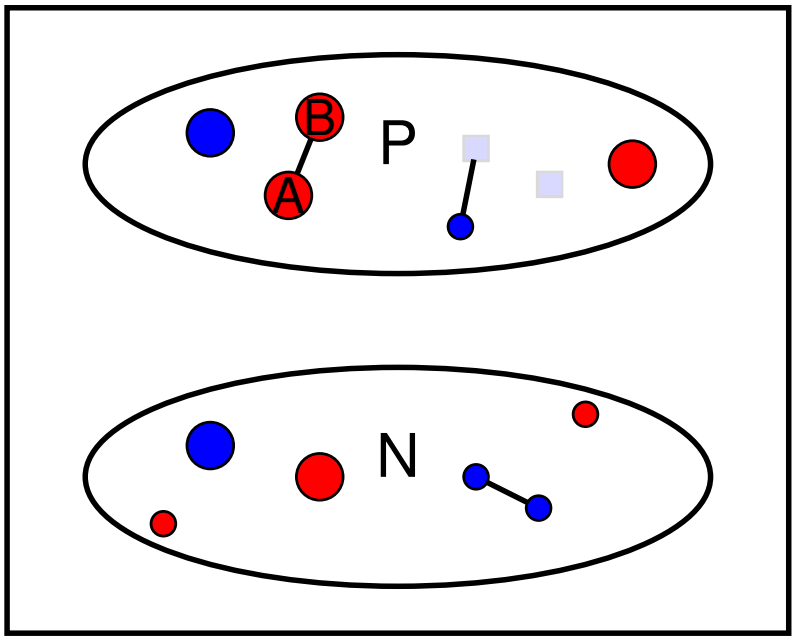


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

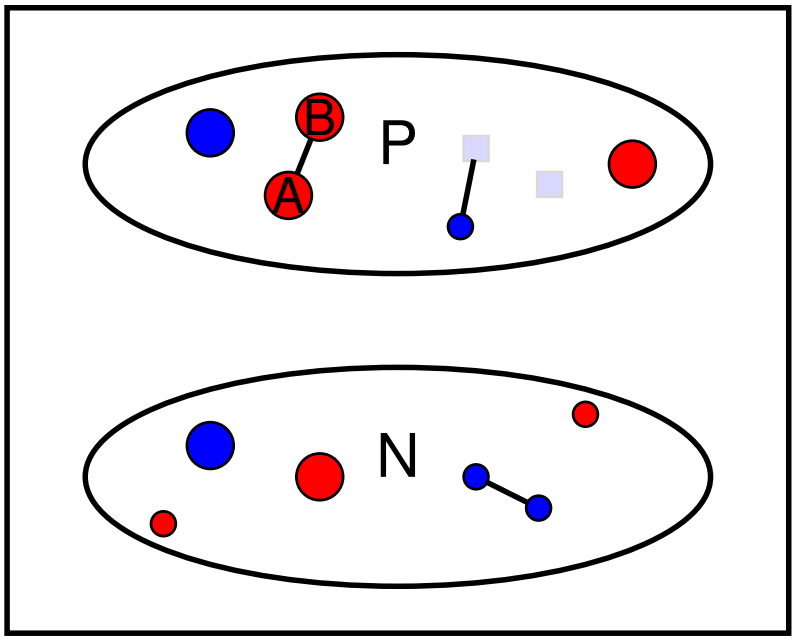


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

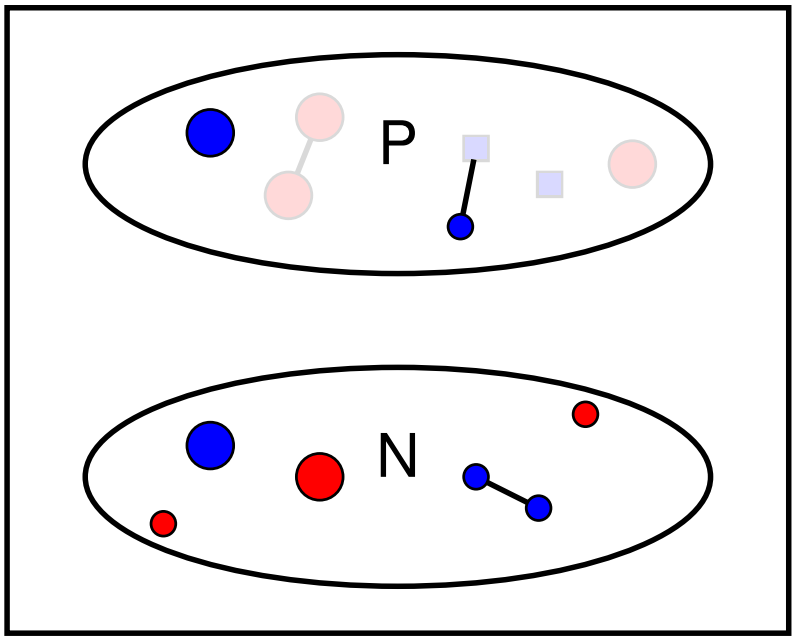


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

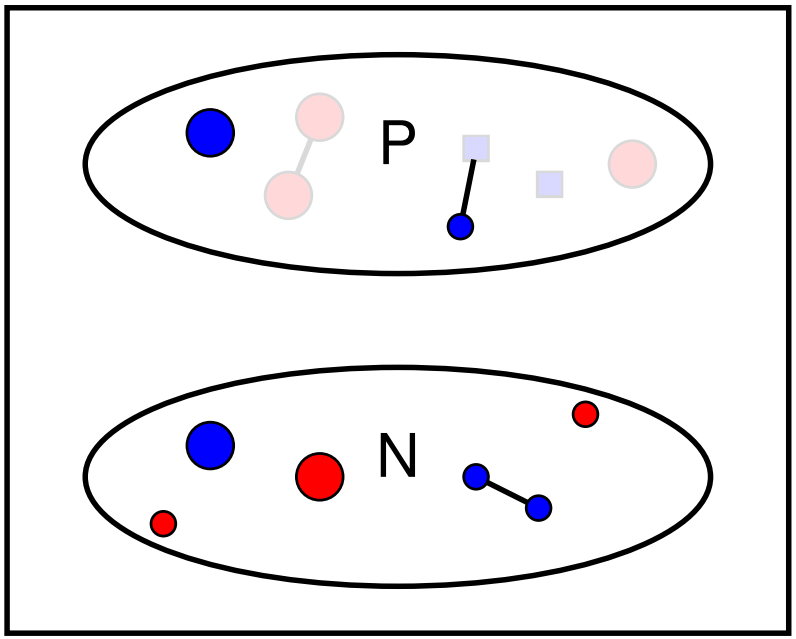


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

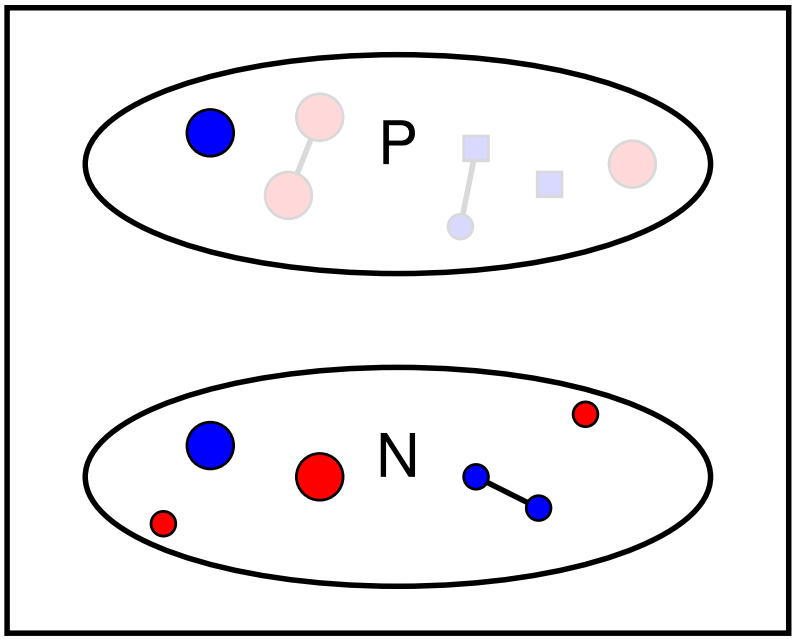


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

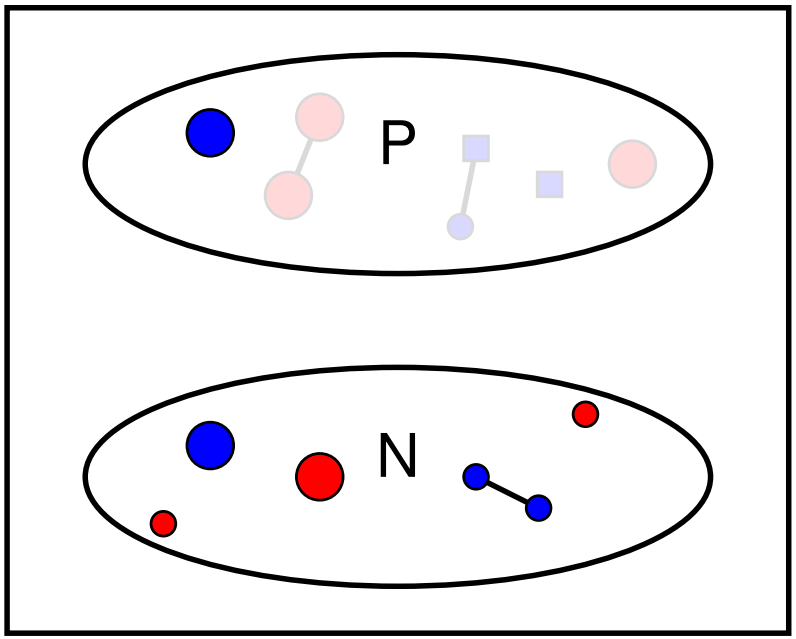


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

Inductive Logic Programming Example

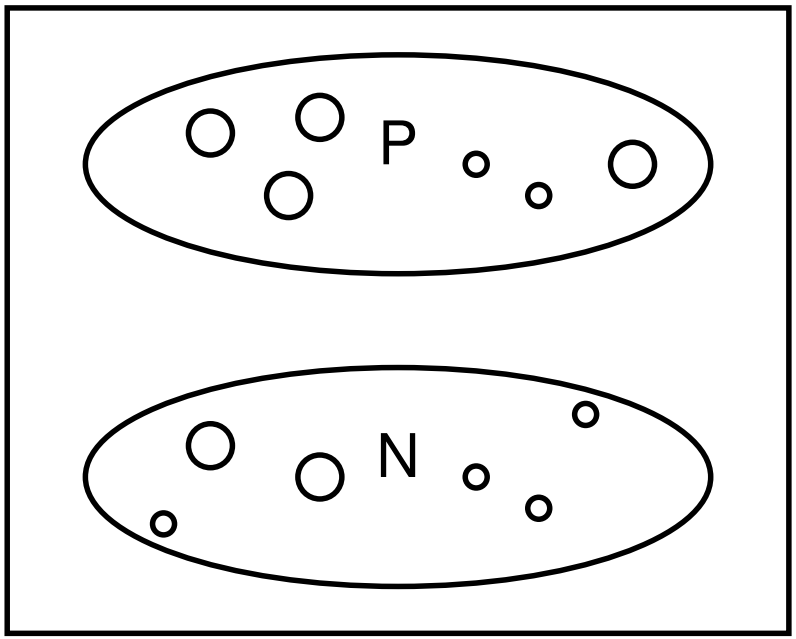


Example

$P(A), red(A), big(A), round(A)$
$sibling(A, B)$

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Baseline Method

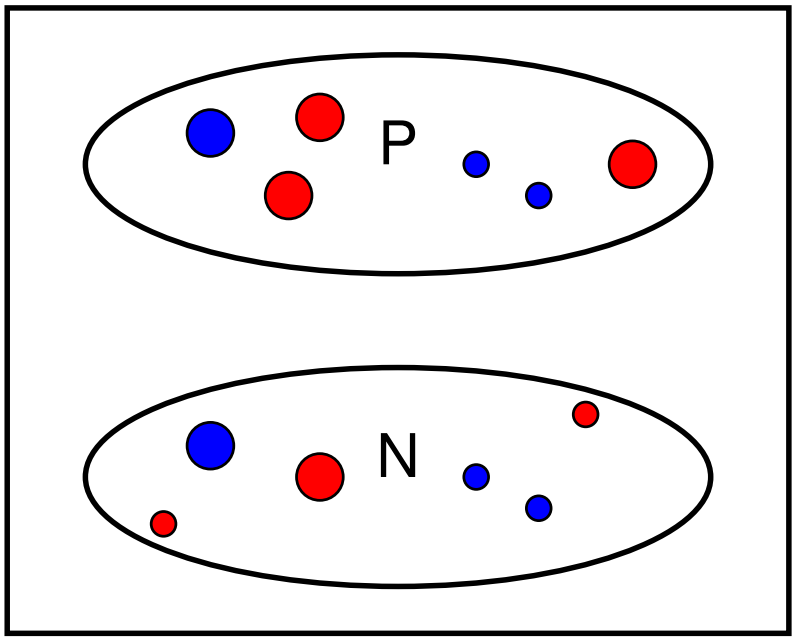


- Include stratifying attribute as a predicate p
- Run ILP over whole dataset
- Select rules containing the predicate p
- Rules specific to the stratum the predicate p refers to

Example

$P(X)$ *if* $red(X) \wedge big(X)$

Baseline Method

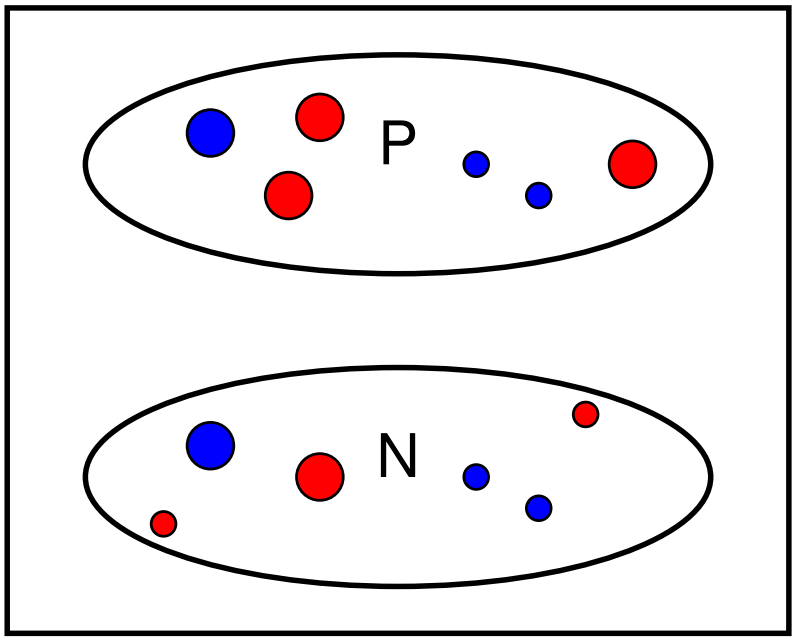


- Include stratifying attribute as a predicate p
- Run ILP over whole dataset
- Select rules containing the predicate p
- Rules specific to the stratum the predicate p refers to

Example

$P(X)$ *if* $red(X) \wedge big(X)$

Baseline Method

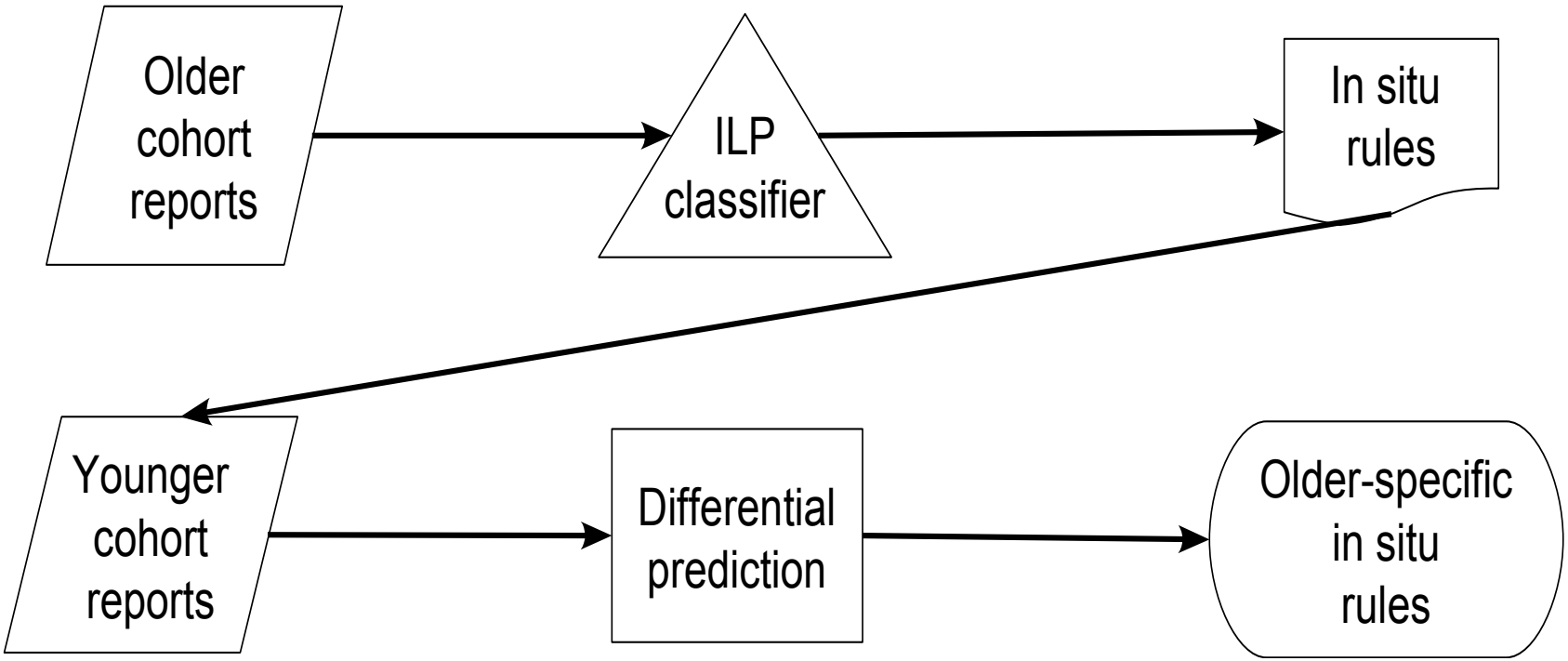


- Include stratifying attribute as a predicate p
- Run ILP over whole dataset
- Select rules containing the predicate p
- Rules specific to the stratum the predicate p refers to

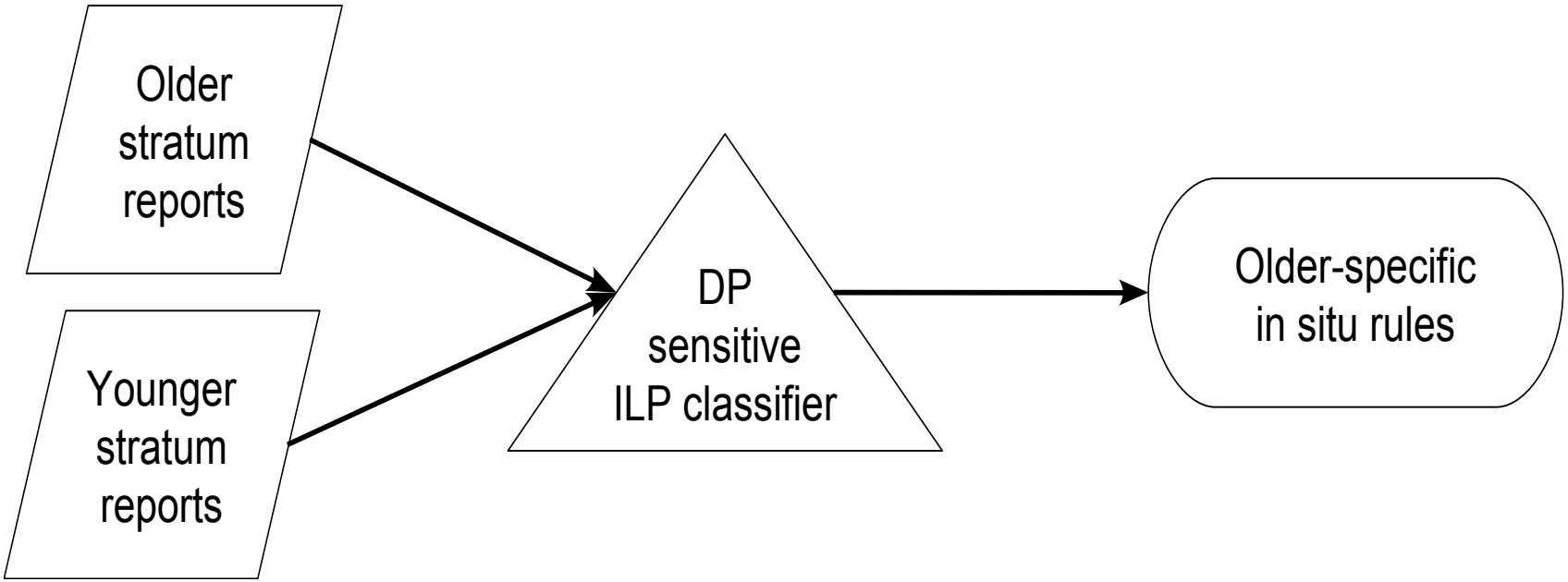
Example

$P(X)$ *if* $red(X) \wedge big(X)$

Model Filtering Method



Differential Prediction Search Method



DP-Sensitive Scoring Function

Definition (DP-Sensitive Scoring Function)

Let R be a rule over the set of instances X, and let $\mathcal{D}$ be a 2-strata dataset over X. We define a **differential-prediction-sensitive scoring function** Q as a function of R, D_t and D_o, such that Q is positively correlated to the performance of R over D_t, and negatively correlated to the performance of R over D_o.

Example

$$Q(R|D_t, D_o) = S(R|D_t) - S(R|D_o)$$

DP-Sensitive Scoring Function

Definition (DP-Sensitive Scoring Function)

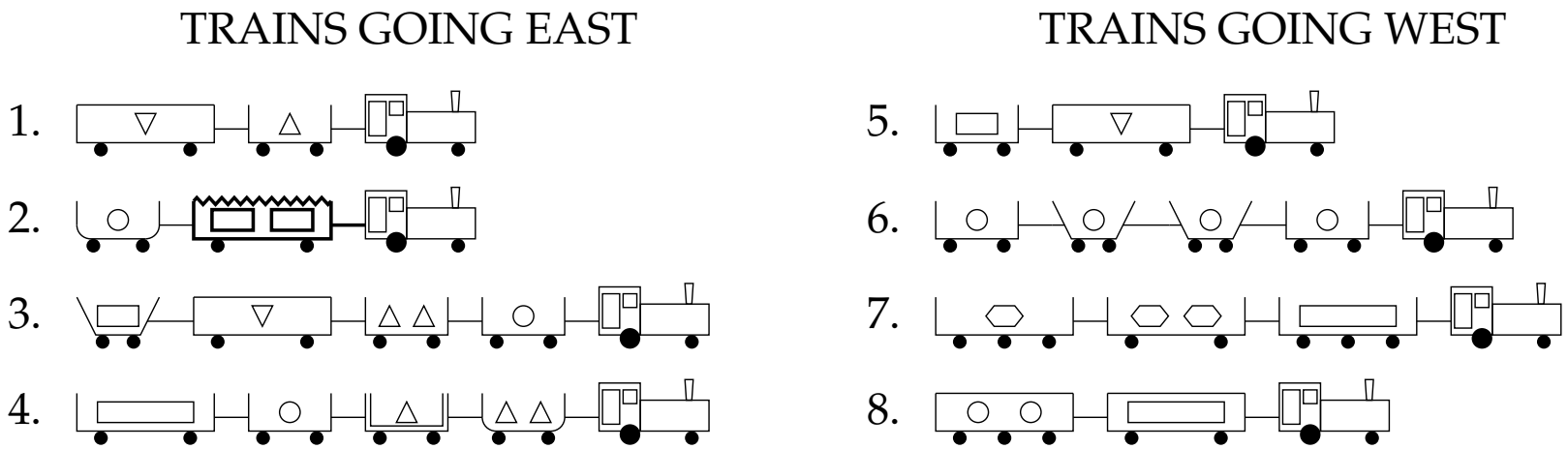
Let R be a rule over the set of instances X, and let $\mathcal{D}$ be a 2-strata dataset over X. We define a **differential-prediction-sensitive scoring function** Q as a function of R, D_t and D_o, such that Q is positively correlated to the performance of R over D_t, and negatively correlated to the performance of R over D_o.

Example

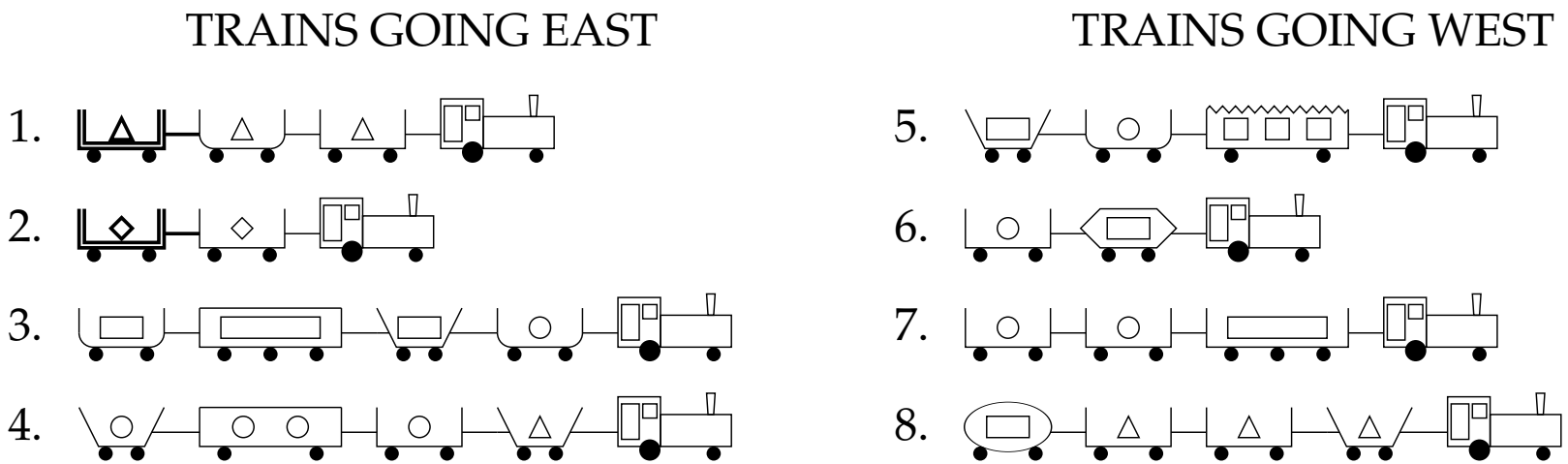
$$Q(R|D_t, D_o) = S(R|D_t) - S(R|D_o)$$

Outline

Michalski Trains (*Larson'77, Muggleton'98*)



(a) East *A* trains: short infront of long; jagged-roof



(b) East *B* trains: short infront of long; double-hulled

Michalski Trains Experiment

- Size: 100, 1000 (per class and stratum)
- Data: clean, noisy (5% swap)
- Scenarios: one, up to 5 target rules
- Common rules: 1-5

- Rank theory rules by score
- Match rules to target ground truth rules
- PR curve on recovered rules

Michalski Trains Experiment

- Size: 100, 1000 (per class and stratum)
- Data: clean, noisy (5% swap)
- Scenarios: one, up to 5 target rules
- Common rules: 1-5

- Rank theory rules by score
- Match rules to target ground truth rules
- PR curve on recovered rules

Michalski Trains Results

Mean AUC PR for 30 experiments in each block

Size	clean BASE	clean MF	clean DPS	noisy BASE	noisy MF	noisy DPS
One target rule scenario						
100	0.73	**0.83**	0.62	0.57	**0.62**	0.54
1000	0.87	**0.90**	0.88	0.63	0.80	**0.87**
Multiple target rules scenario						
100	0.61	**0.70**	0.42	0.38	**0.52**	0.31
1000	0.75	**0.86**	0.77	0.52	0.55	**0.65**

- DPS more appropriate for real-world (large + noisy) data

Michalski Trains Results

Mean AUC PR for 30 experiments in each block

Size	clean BASE	clean MF	clean DPS	noisy BASE	noisy MF	noisy DPS
One target rule scenario						
100	0.73	**0.83**	0.62	0.57	**0.62**	0.54
1000	0.87	**0.90**	0.88	0.63	0.80	**0.87**
Multiple target rules scenario						
100	0.61	**0.70**	0.42	0.38	**0.52**	0.31
1000	0.75	**0.86**	0.77	0.52	0.55	**0.65**

- DPS more appropriate for real-world (large + noisy) data

Breast Cancer Diagnosis

- Baseline: No returned DP rule
- Model Filtering: Rule 1
- Differential Prediction Search: Rules 1-5

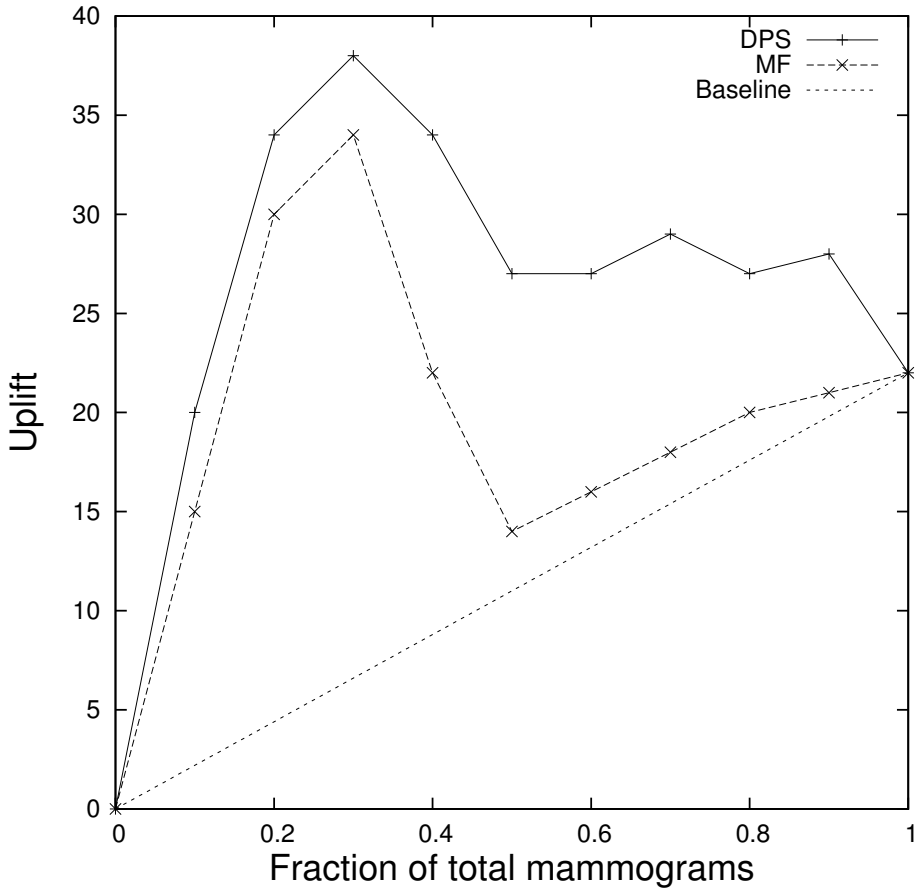
- A tumor is older-specific if:
  1. It has a calcification
  2. It is a class 2 breast density
  3. It has a prior in situ biopsy
  4. It's BI-RADS score increased
  5. It is a screening visit

Uplift Curve

Lift : Nb of positives in top ranking fraction p

Uplift : $p \in [0, 1]$, plot $\{p, Lift_t - Lift_o\}$

Use theory to form TAN classifier to assign example probability



Contribution Summary

- Extended differential prediction to relational datasets
- Proposed three methods for learning DP rules
- Recommend Differential Prediction Search for large and noisy data
- Recommend Model Filtering for small or clean data

- This work is supported by US National Institute of Health (NIH) grants R01-CA127379-01, R01-LM010921-01A1, and R01-CA165229-01A1.

Outline

Age Matters

- Stratify our data (*Nichols'06*):

  Younger: $<$ 50 years, pre-menopausal
  Middle: $[50, 65)$ years, peri-menopausal
  Older: $>=$ 65 years, post-menopausal

- Apply logistic regression on older and younger
- Uncover a differential ability in predicting invasive and in-situ cancer in older vs. younger women

Age Strata (*Nassif'10*)

Stratum	Invasive	In-Situ	Total
Younger	264	110	374
Middle	398	170	568
Older	401	132	533
Total	1063	412	1475

Mammography Features

Structured	NLP Extracted (*Nassif'09*)
Family breast cancer history	Mass margin
Personal breast cancer history	Mass shape
Prior surgery	Calcification distribution
Palpable lump	Calcification morphology
Screening v/s diagnostic	Architectural distortion
Indication for exam	Associated findings
Breast Density	Mammary lymph node
BI-RADS code left	Asymmetric breast tissue
BI-RADS code right	Focal asymmetric density
BI-RADS code combined	Tubular density
Principal finding	Mass size

Older-specific In Situ Rules

1. Calcification
   - Tumor indolent in older women
   - Asymptomatic in situ detected due to micro-calcifications
   - Novel finding
2. Class 2 breast density
   - Lower breast density increases mammogram sensitivity, easier micro-calcification detection
3. Prior in situ biopsy
   - Tumor indolent in older women
4. BI-RADS score increase
5. Screening visit
   - Regular screening age > 40