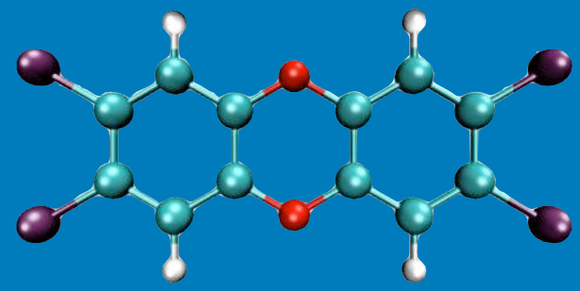
fast effective molecular feature mining by local optimization

Albrecht Zimmermann, Björn Bringmann, Ulrich Rückert

fast effective molecular feature mining by local optimization

Albrecht Zimmermann, Björn Bringmann, Ulrich Rückert

qsar



You are here

(1 1 0 ... 1 0)

?

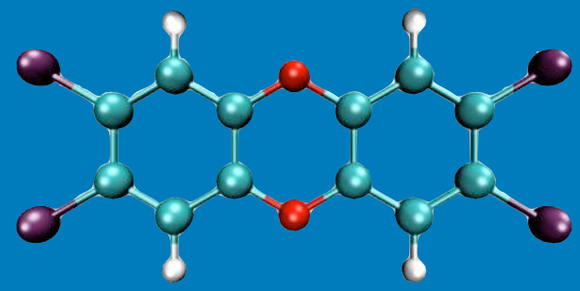
cancer

svm

model

simple approach: **qsar**

dump a load of patterns into svm



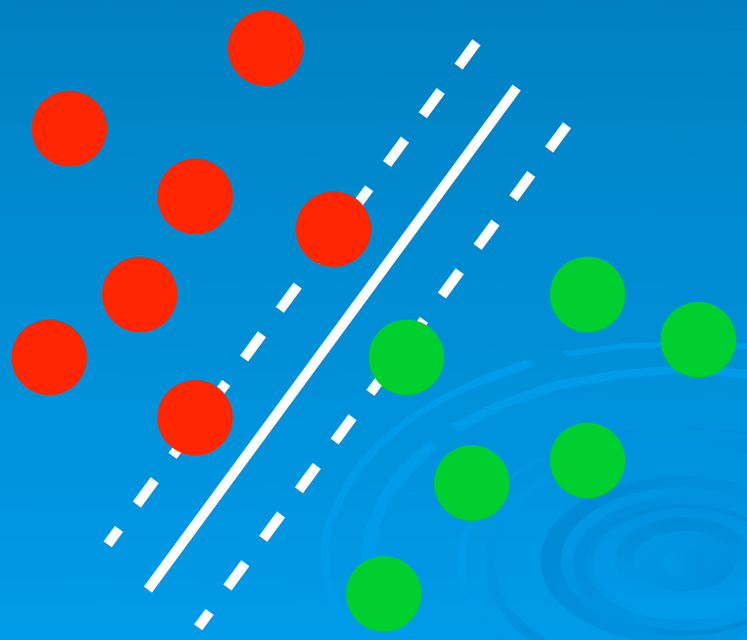
You are here

(1 1 0 ... 1 0)

?

cancer

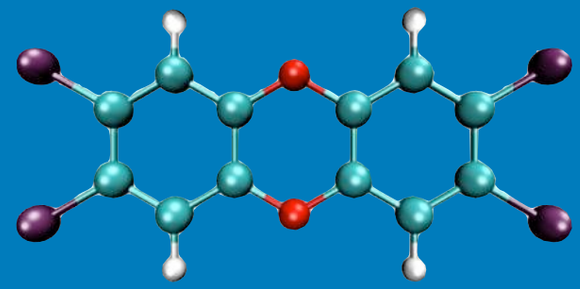
svm



model

simple approach: **qsar**

dump a load of patterns into svm



You are here

(1 1 0 ... 1 0)

?

cancer

svm

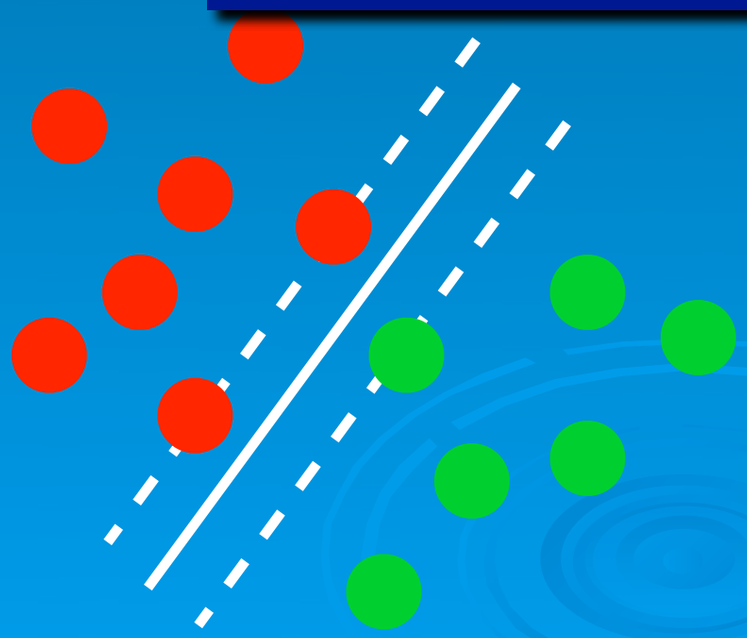
model

better

pattern set mining with

a) post-processing

b) iterative mining



a tale of two paradigms

Supervised

pattern set to
distinguishes classes

correspondences, ig, ...

Bringmann & Zimmermann ’05
Cheng et al. ’08 Thoma et al. ‘09

Unsupervised

pattern set to induce
fine, balanced partition

joint entropy

Knobbe & Ho ‘06
Bringmann & Zimmermann ‘07

a tale of two paradigms

Supervised

pattern set to
distinguishes classes

correspondences, ig, ...

Bringmann & Zimmermann ’05
Cheng et al. ’08 Thoma et al. ‘09

Unsupervised

pattern set to induce
fine, balanced partition

joint entropy

Knobbe & Ho ‘06
Bringmann & Zimmermann ‘07

Trade-off

Combined Scores

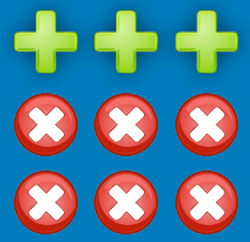
Trade-off

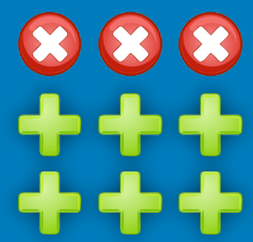
Cheng et al. ‘07
Rückert & Kramer ‘07

class-correlated dispersion score

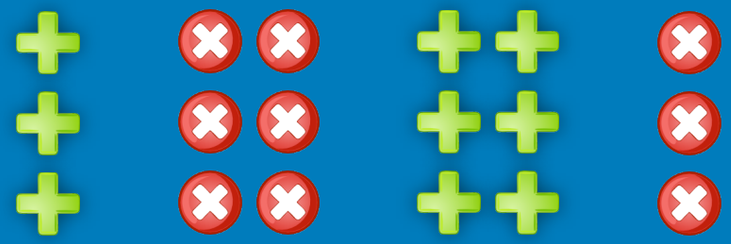
how does this translate?

how does this translate?



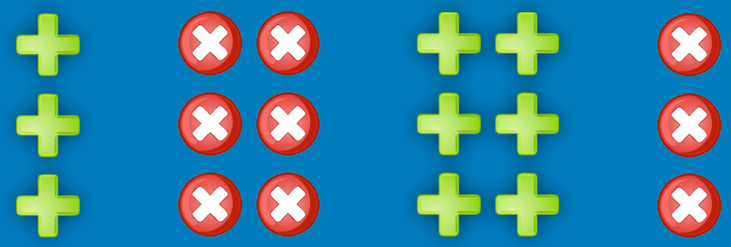


how does this translate?



how does this translate?

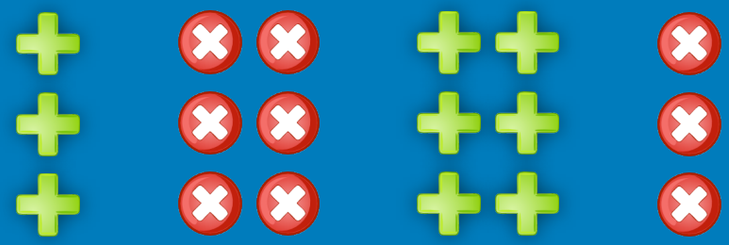
Supervised



how does this translate?

Supervised

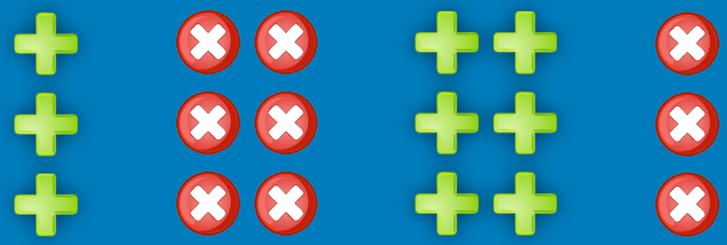
Unsupervised

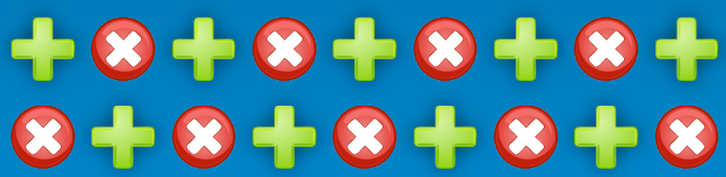


how does this translate?

Supervised

Unsupervised

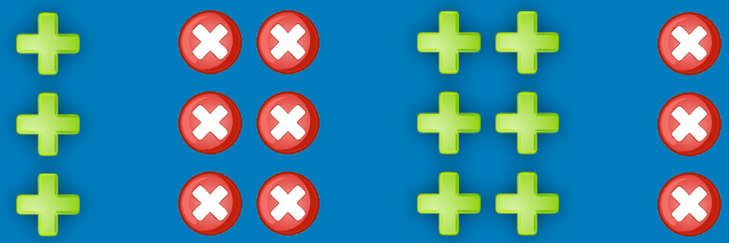


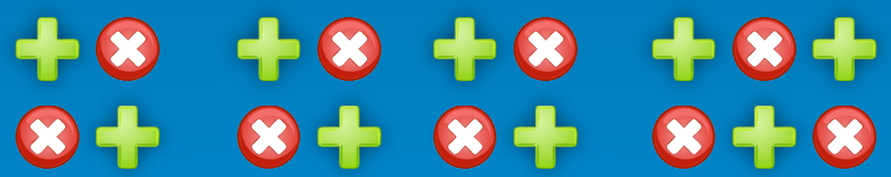


how does this translate?

Supervised

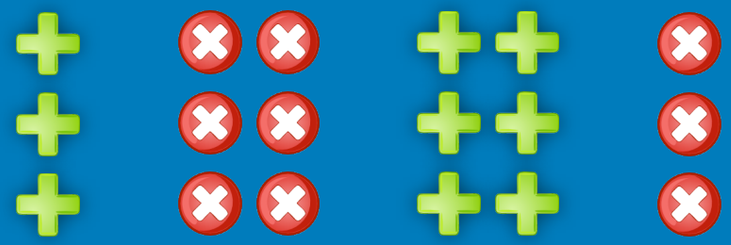
Unsupervised



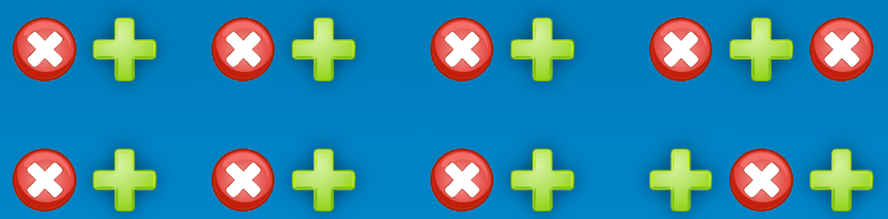


how does this translate?

Supervised

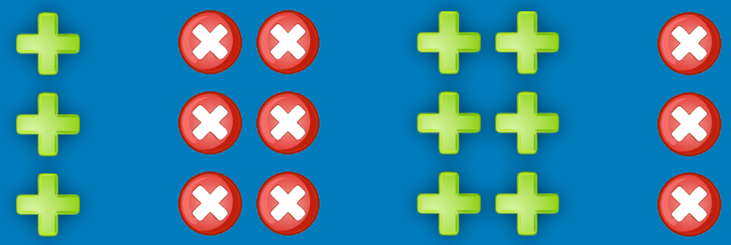


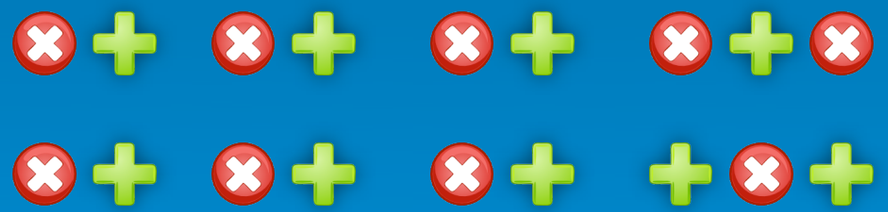
Unsupervised



how does this translate?

Supervised



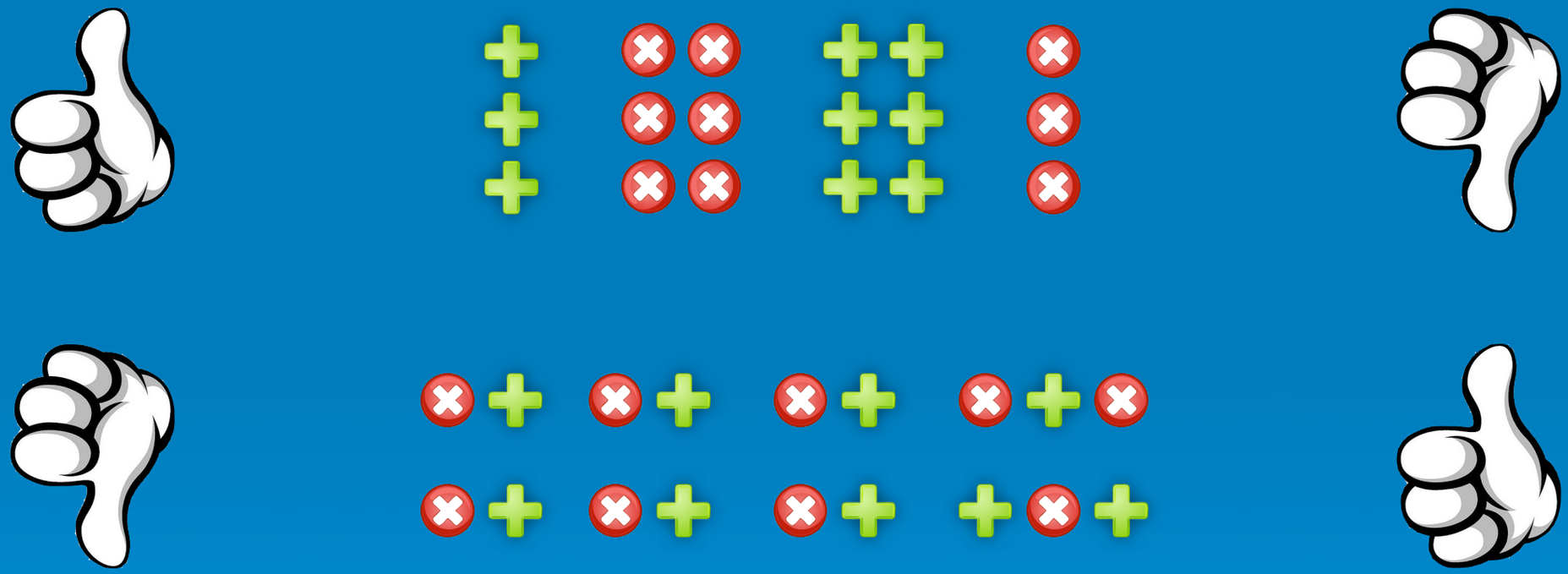


Unsupervised

how does this translate?

Supervised

Unsupervised



which paradigm is justified?

some empirical exploration

algorithm	properties	
baseline	500 most general frequent graphs	
*picker**	cond. pattern prob.	post processing
rückert '07	class-correlated disp-score	iterative
fcork	correspondence-based	iterative
dtm	information gain	iterative

relation to AUC?

some empirical exploration

algorithm	properties	
baseline	500 most general frequent graphs	
*picker**	cond. pattern prob.	**Unsupervised**
rückert '07	class-correlated disp-score	**Combined**
fcork	correspondence-based	**Supervised**
dtm	information gain	**Supervised**

relation to AUC?

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	⚡
# equivalence classes	+ +	✓ / ⚡
high correspondence	− − −	✓
low correspondence	✗	⚠
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ / ⚡
high correspondence	− − −	✓
low correspondence	✗	⚠
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ /
high correspondence	− − −	✓
low correspondence	✗	⚠
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ / ⚡
high correspondence	− − −	Supervised
low correspondence	✗	⚠
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ /
high correspondence	− − −	Supervised
low correspondence	✗	⚠
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ / ⚡
high correspondence	− − −	Supervised
low correspondence	✗	⚠
⊗ ⊗ ⊗ / ⊗ ⊗ ⊗ / ⊗ ⊗ ⊗	+ + + / + + + / + + +	overfitting?!
dispersion score	+	
disp. score w/class	+ +	✓

what we found - correlations

characteristic	correlates w/auc	expected?
# features	✗	✓
joint entropy	+ + +	Unsupervised
# equivalence classes	+ +	✓ /
high correspondence	− − −	Supervised
low correspondence	✗	⚠
⊗ ⊗ ⊗ / ⊗ ⊗ ⊗ / ⊗ ⊗ ⊗	+ + + / + + + / + + +	overfitting?!
dispersion score	+	
disp. score w/class	+ +	✓

what we found - algorithms

characteristic	result
auc	dtm >> fcork
# correspondences	dtm ~ fcork
# equivalence classes	dtm > fcork
# features	dtm >> fcork
running times	dtm >> fcork

what we found - algorithms

characteristic	result
auc	dtm >> fcork
# correspondences	dtm ~ fcork
# equivalence classes	dtm > fcork
# features	dtm >> fcork
running times	dtm >> fcork

💡 it's in the way they work! 🙂

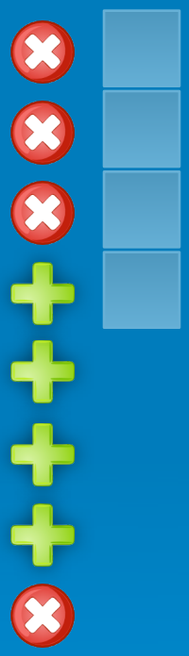
another tale of two paradigms

mine global,
use global

another tale of two paradigms

mine global,
use global

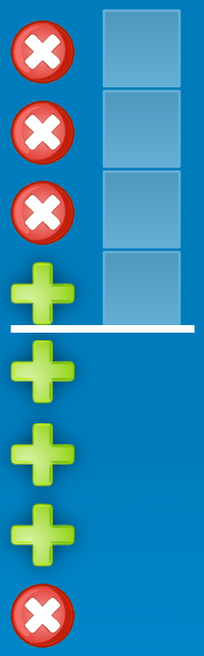
p_1



another tale of two paradigms

mine global,
use global

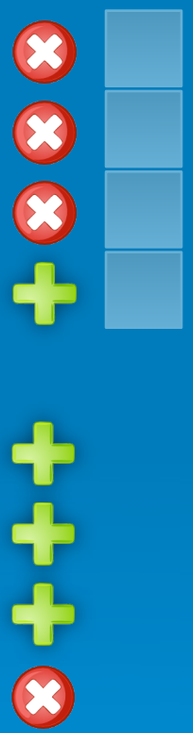
p_1



another tale of two paradigms

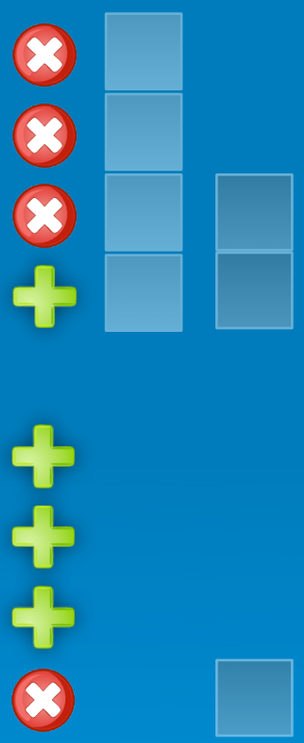
mine global,
use global

p_1



another tale of two paradigms

mine global, use global

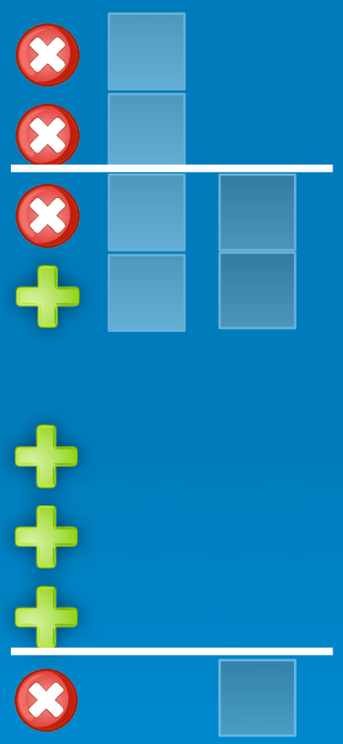


- patterns mined on **all** data

another tale of two paradigms

mine global, use global

p_1 p_2

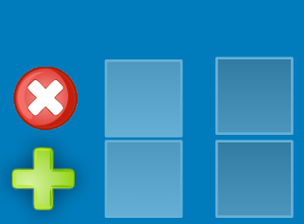


- patterns mined on **all** data
- patterns applied to **all** data

another tale of two paradigms

mine global, use global

p_1 p_2

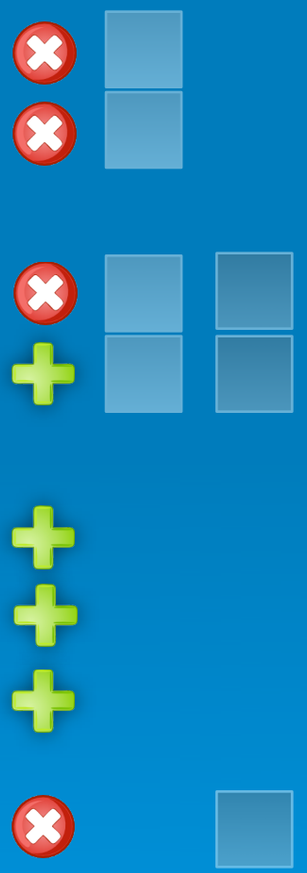


- patterns mined on **all** data
- patterns applied to **all** data

another tale of two paradigms

mine global, use global

p_1 p_2



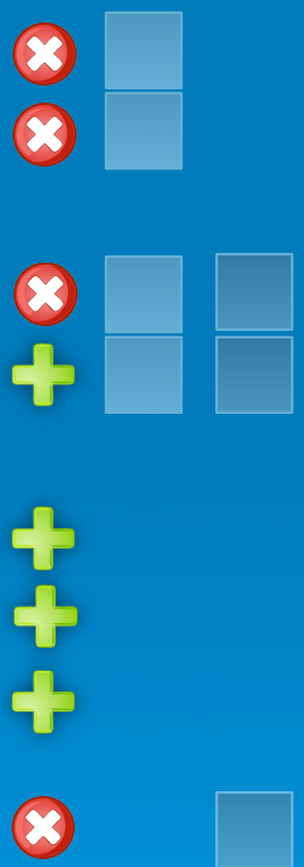
mine local, use local

- patterns mined on **all** data
- patterns applied to **all** data

another tale of two paradigms

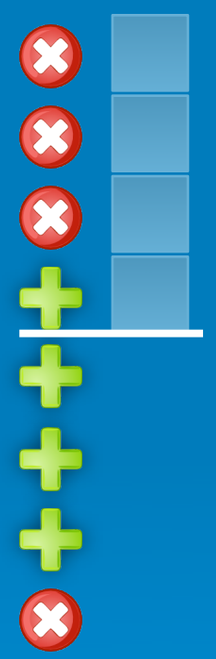
mine global, use global

p_1 p_2



mine local, use local

p_1

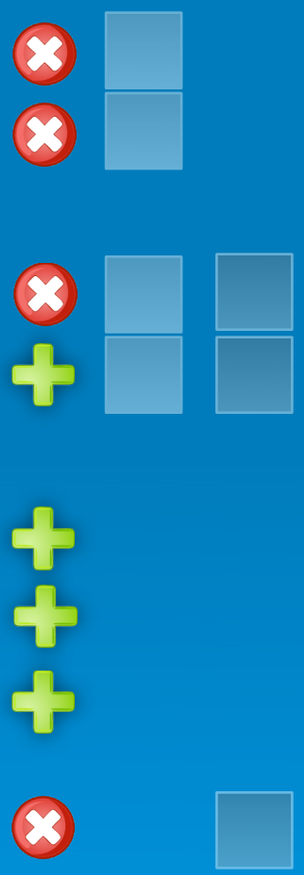


- patterns mined on **all** data
- patterns applied to **all** data

another tale of two paradigms

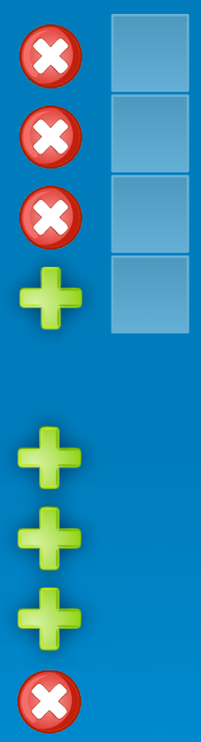
mine global, use global

p_1 p_2



mine local, use local

p_1

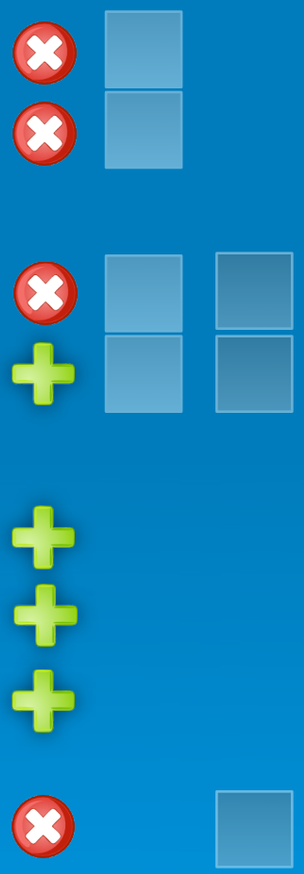


- patterns mined on **all** data
- patterns applied to **all** data

another tale of two paradigms

mine global, use global

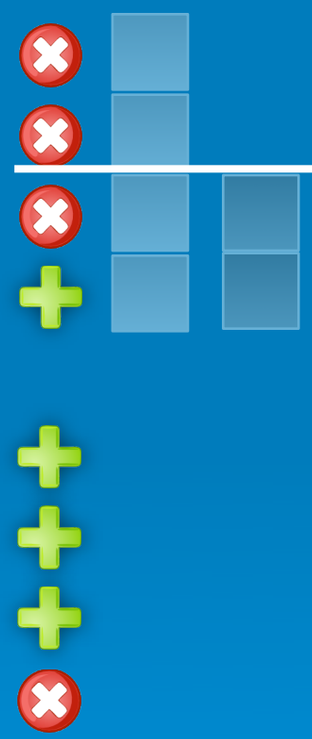
p_1 p_2



- patterns mined on **all** data
- patterns applied to **all** data

mine local, use local

p_1 p_2

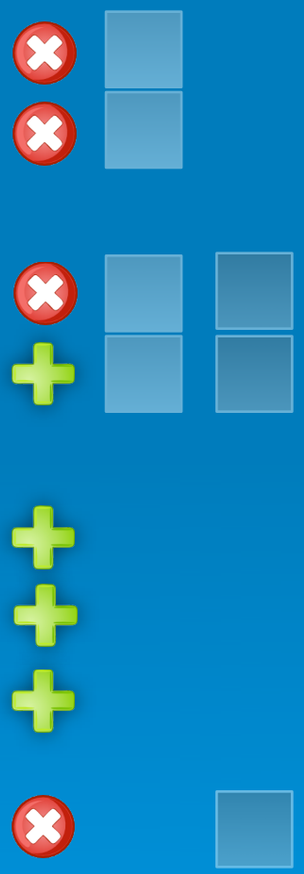


- patterns mined on **subset**

another tale of two paradigms

mine global, use global

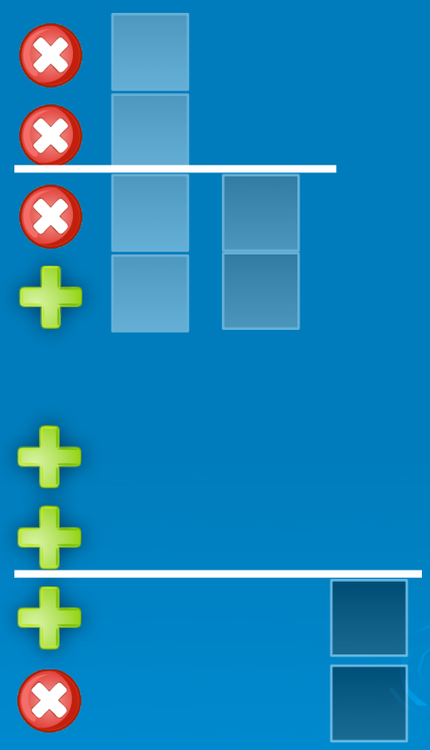
p_1 p_2



- patterns mined on **all** data
- patterns applied to **all** data

mine local, use local

p_1 p_2 p_3

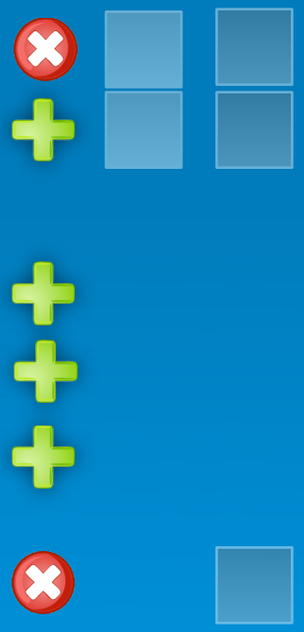


- patterns mined on **subset**
- patterns applied its **subset**

another tale of two paradigms

mine global, use global

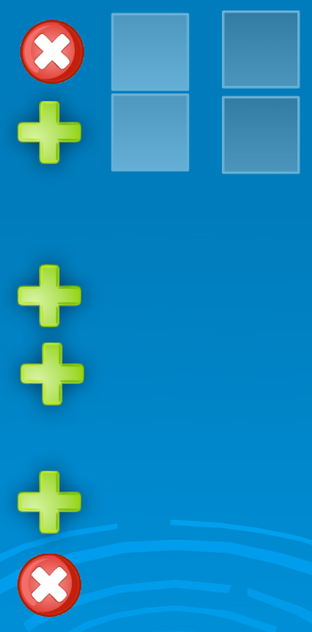
p_1 p_2



- patterns mined on **all** data
- patterns applied to **all** data

mine local, use local

p_1 p_2 p_3

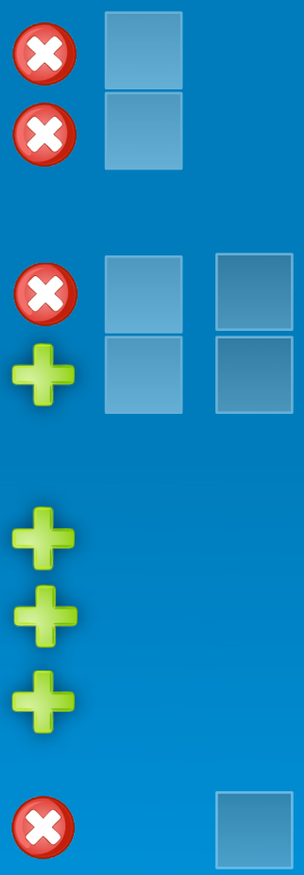


- patterns mined on **subset**
- patterns applied its **subset**

another tale of two paradigms

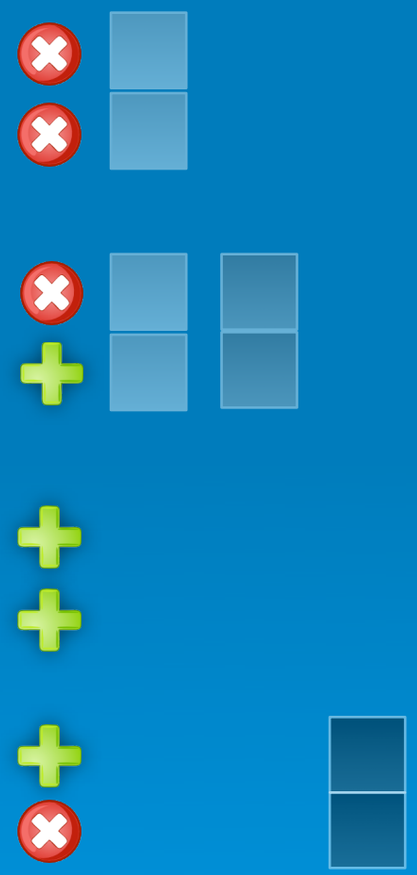
mine global, use global

p_1 p_2



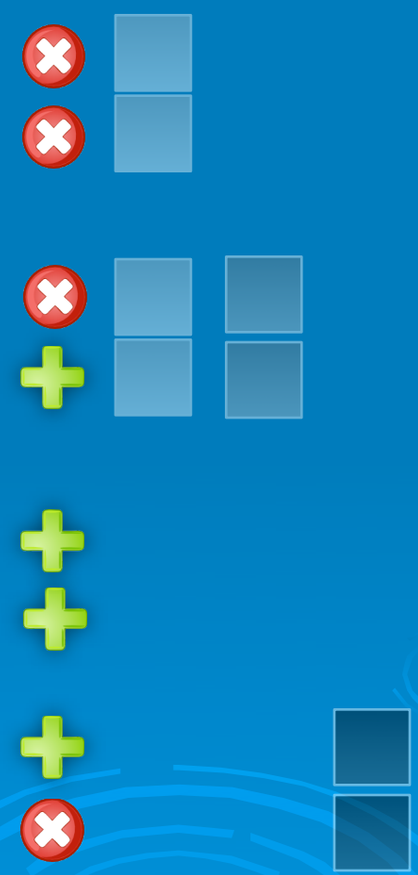
- patterns mined on **all** data
- patterns applied to **all** data

p_1 p_2 p_3



mine local, use local

p_1 p_2 p_3

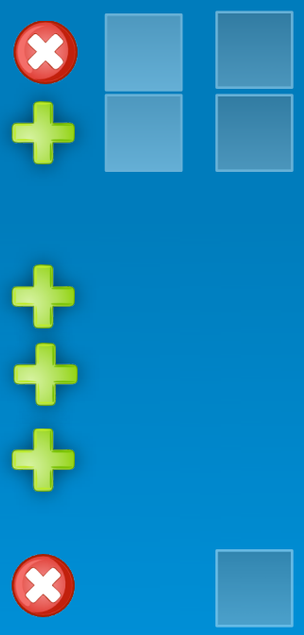


- patterns mined on **subset**

another tale of two paradigms

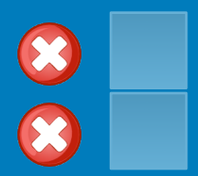
mine global, use global

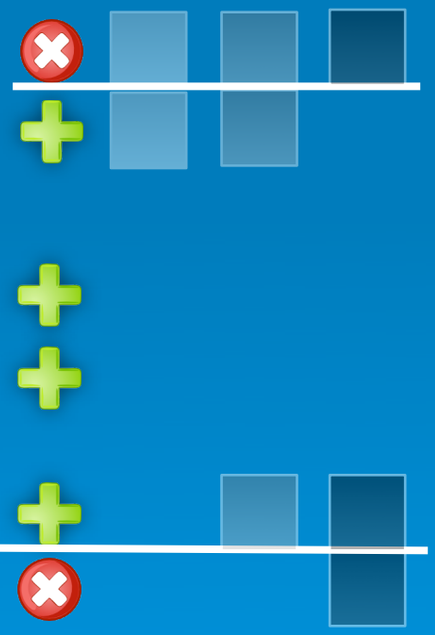
p_1 p_2



- patterns applied to **all** data

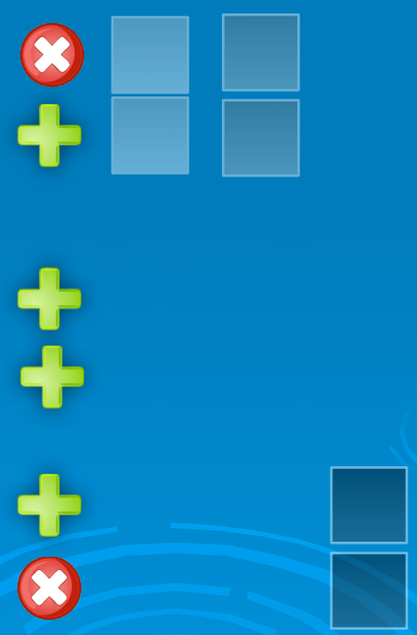
p_1 p_2 p_3





mine local, use local

p_1 p_2 p_3

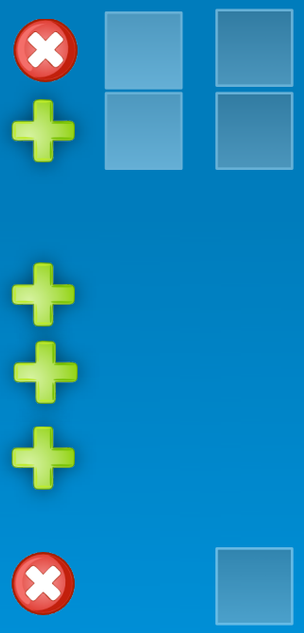


- patterns mined on **subset**

another tale of two paradigms

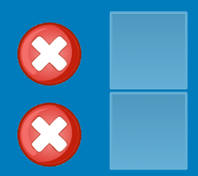
mine global, use global

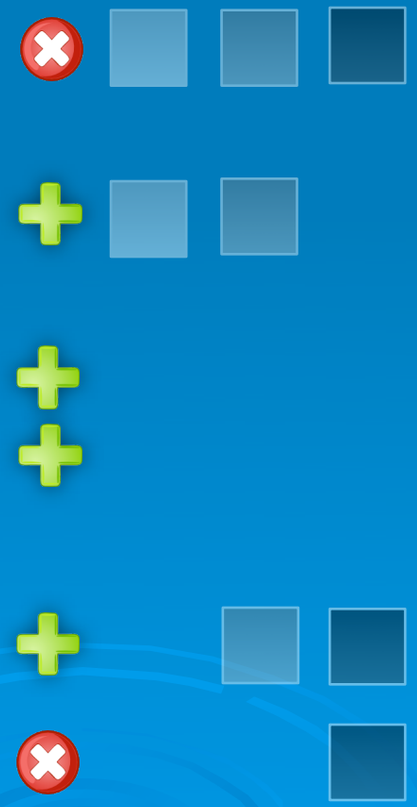
p_1 p_2



- patterns applied to **all** data

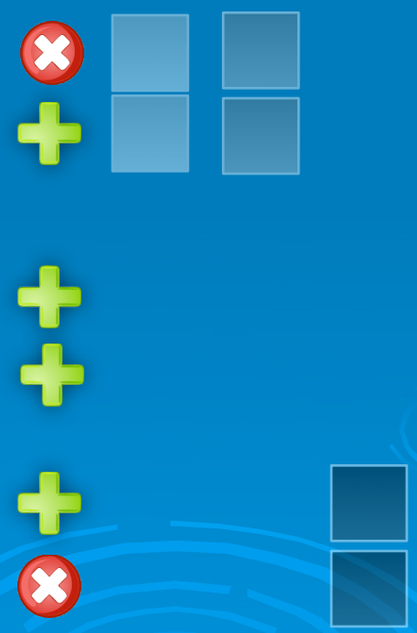
p_1 p_2 p_3





mine local, use local

p_1 p_2 p_3

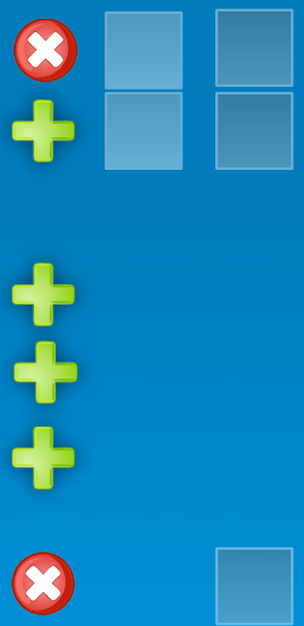


- patterns mined on **subset**

another tale of two paradigms

mine global, use global

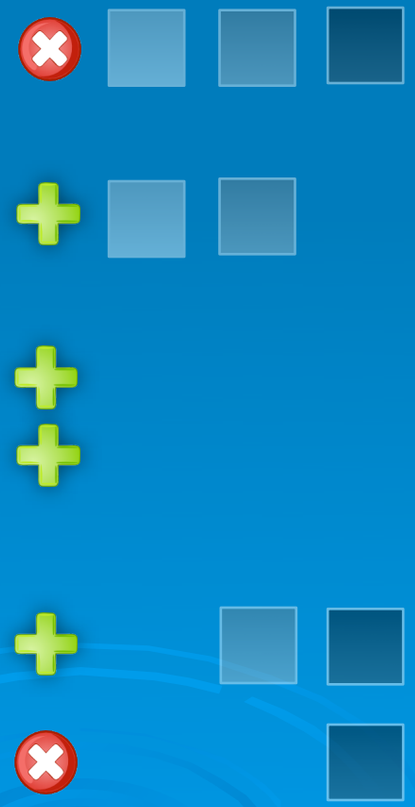
p_1 p_2



- patterns applied to **all** data

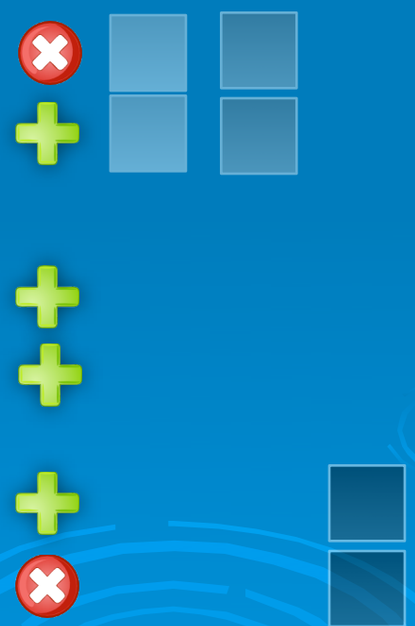
mine local, use global

p_1 p_2 p_3



mine local, use local

p_1 p_2 p_3



- patterns mined on **subset**

another tale of two paradigms

mine global, use global

p_1 p_2

mine local, use global

p_1 p_2 p_3

mine local, use local

p_1 p_2 p_3

introducing **remine**

- patterns applied to **all** data
- patterns mined on **subset**

experimental evaluation

measure	expected	actual	
# equivalence classes	remine ~ dtm	remine ~ dtm	✓
# features (prec slide)	remine < dtm	remine < dtm	✓
	remine ? fcork	remine > fcork	
# correspondences	remine ~ dtm	remine < dtm	⚠
auc	remine ~ dtm	remine ~ dtm	✓
running times	remine < dtm	remine < dtm	✓
	remine ? fcork	remine < fcork	☺

getting picky

measure	evaluation
# correspondences	remine < dtm
# equivalence classes	remine < dtm
auc	remine < dtm

getting picky

measure	evaluation	exploration
# correspondences	remine < dtm	dtm ~ fcork
# equivalence classes	remine < dtm	dtm > fcork
auc	remine < dtm	dtm >> fcork

getting picky

measure	evaluation	exploration
# correspondences	remine < dtm	dtm ~ fcork
# equivalence classes	remine < dtm	dtm > fcork
auc	remine < dtm	dtm >> fcork

⇒ **Unsupervised**

getting picky

measure	evaluation	exploration
# correspondences	remine < dtm	dtm ~ fcork
# equivalence classes	remine < dtm	dtm > fcork
auc	remine < dtm	dtm >> fcork

⇒ **Unsupervised**

# features	fcork < remine < dtm

getting picky

measure	evaluation	exploration
# correspondences	remine < dtm	dtm ~ fcork
# equivalence classes	remine < dtm	dtm > fcork
auc	remine < dtm	dtm >> fcork

⇒ **Unsupervised**

# features/auc	fcork < remine < dtm

getting picky

measure	evaluation	exploration
# correspondences	remine < dtm	dtm ~ fcork
# equivalence classes	remine < dtm	dtm > fcork
auc	remine < dtm	dtm >> fcork

⇒ **Unsupervised**

# features/auc	fcork < remine < dtm

⇒ **overfitting?!**

conclusions

conclusions

preliminary evidence that

Unsupervised > **Supervised**

conclusions

preliminary evidence that

Unsupervised > **Supervised**

& class information can improve partition-scores

conclusions

preliminary evidence that

Unsupervised > **Supervised**

& class information can improve partition-scores

mine global, use global — Trade-off → **remine** ← Trade-off — **mine local, use local**

conclusions

preliminary evidence that

Unsupervised > **Supervised**

& class information can improve partition-scores

mine global, use global — Trade-off → **remine** ← Trade-off — **mine local, use local**

further evidence of importance partition

but also of **overfitting?!**

Thank You for Your Attention