Predicting code quality: how?

```
219     public static float[] arithmetic_average(int[][] mat) {
220         float [] vec = new float [mat.length];
221         float temp=0;
222
223         for (int i=0;i<=mat.length-1;i++){
224             for (int j=0;j<=mat[0].length-1;j++){
225                 temp+=mat[i][j];
226             }
227             temp=temp/mat[0].length;
228             vec[i]=temp;
229             temp=0;
230         }
231         return vec;
232     }
```

```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '    %s [label="%s' % (nodename, label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s"];' % ast[1]
        else:
            print '"]'
    else:
        print '"];'
        children = []
        for  in n, childenumerate(ast[1:]):
            children.append(dotwrite(child))
        print ,'    %s -> {' % nodename
        for  in :namechildren
            print '%s' % name,
```

```
Sub Macro1()
Dim Number
Worksheets("Visual Basic").Activate
Number = Cells(2, "A").Value

If Number >= 50 Then
ActiveSheet.Shapes("Freeform 1").Fill.ForeColor.RGB = RGB(204, 0,

End If

If Number >= 100 Then
ActiveSheet.Shapes("Freeform 1").Fill.ForeColor.RGB = RGB(0, 255,

End If

End Sub
```

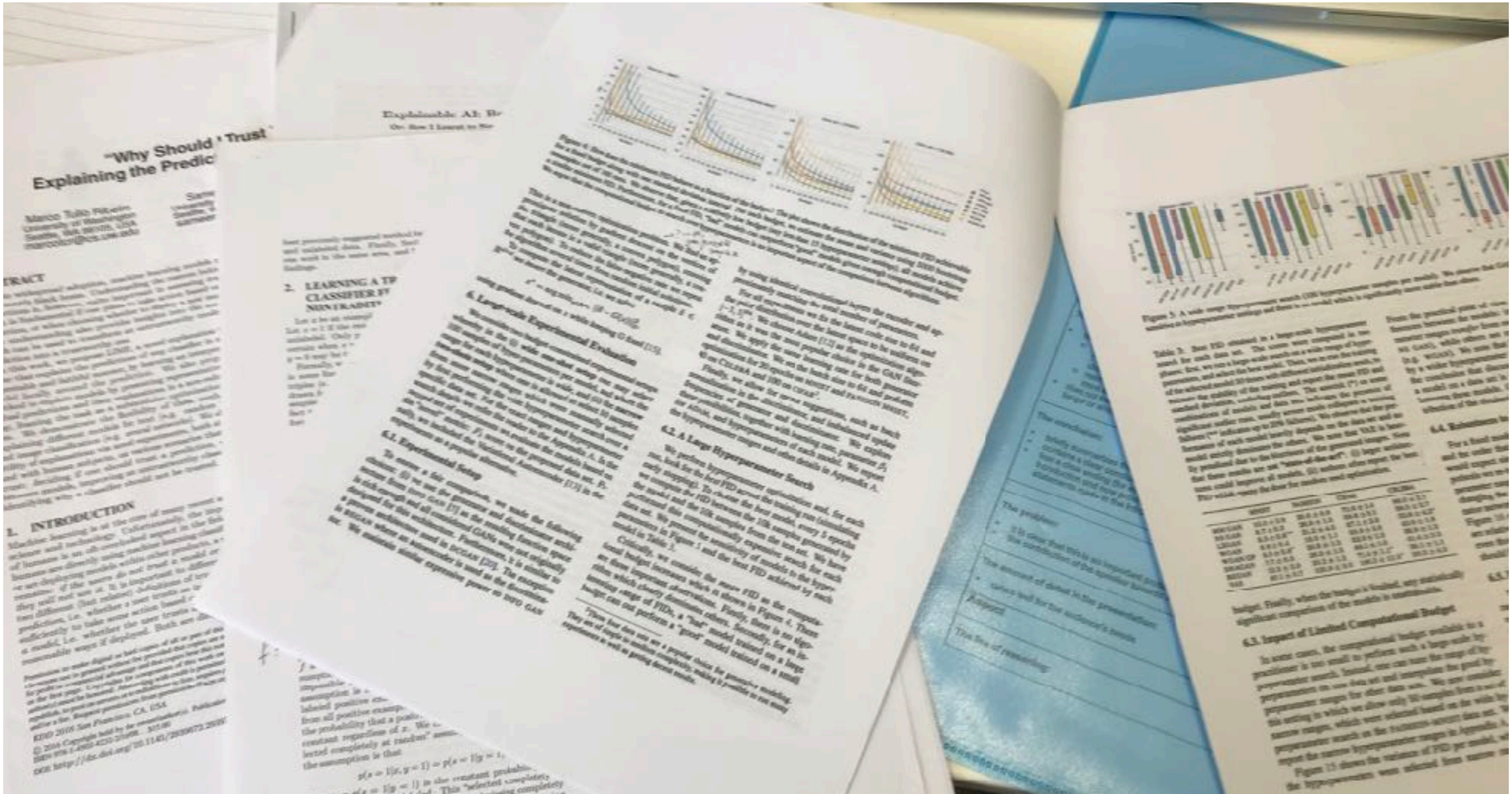
```
for i in people.data.users:
    response = client.api.statuses.user_timeline.get(screen_name=i.scre
    print 'Got', len(response.data), 'tweets from', i.screen_name
    if len(response.data) != 0:
        ltdate = response.data[0]['created_at']
        ltdate2 = datetime.strptime(ltdate,'%a %b %d %H:%M:%S +0000 %Y'
        today = datetime.now()
        howlong = (today-ltdate2).days
        if howlong < daywindow:
            print i.screen_name, 'has tweeted in the past' , daywindow,
            totaltweets += len(response.data)
            for j in response.data:
                if j.entities.urls:
                    for k in j.entities.urls:
                        newurl = k['expanded_url']
                        urlset.add((newurl, j.user.screen_name))
        else:
            print i.screen_name, 'has not tweeted in the past', daywind
```

```
$_='while(read+STDIN,$_,2048)
{$a=29;$b=73;$c=142;$t=255;@t=map{$_%16or$t^=
$c^=( $m=(11,10,116,100,11,122,20,100)[$_/
16%8])&110;$t^=(72,@z=(64,72,$a^=12*($_%16 -2?0:$m&17)),
$b^=$_%64?12:0,@z)[$_%8]}(16..271);if((@a=unx"C*",$_)
[20]&48){$h =5;$_=unxb24,join"",@b=map{xB8,unxb8,chr($_^
$a[--$h+84])}@ARGV;s/...$/1$&/;$ d=unxV,xb25,$_;$e=256|
(ord$b[4])<<9|ord$b[3];$d=$d8^($f=$t&($d12^$d4^ $d^$d/
8))<<17,$e=$e8^($t&($g=($q=$e14&7^$e)^$q*8^$q<<6))<<9,$_=
$t[$_]^ (($h=8)+=$f+(~$g&$t))for@a[128..$#a]}print
+x"C*",@a}';s/x/pack+/g;eval
```

Good representation for this image?

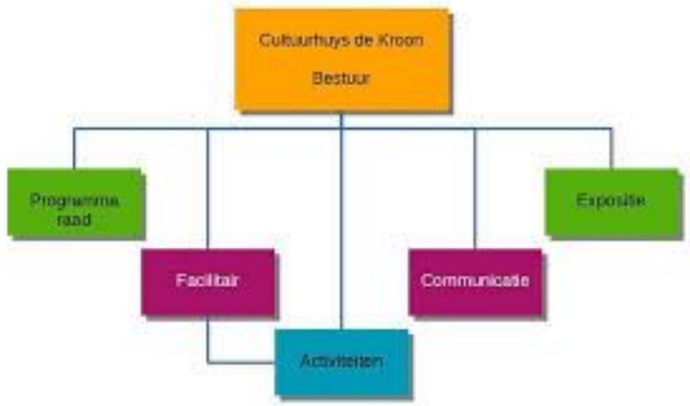
Classify project reports?

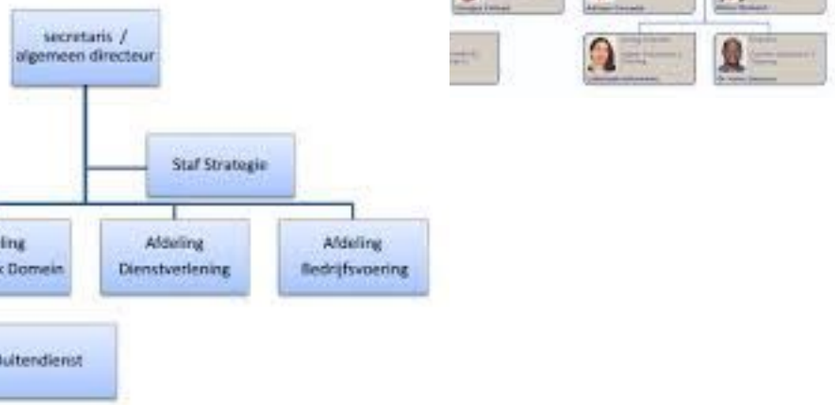
- Given that the structure of the texts are, more or less, comparable, how to (automatically) evaluate/classify?

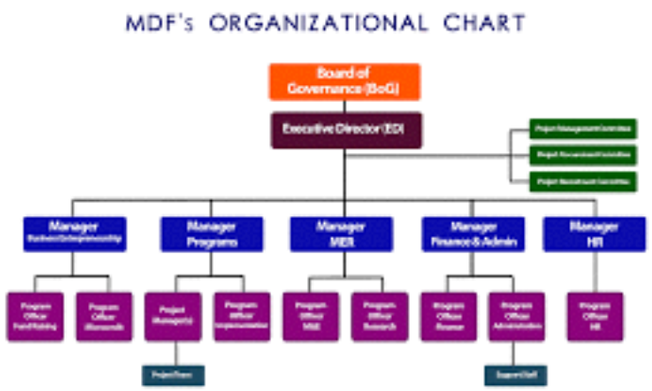


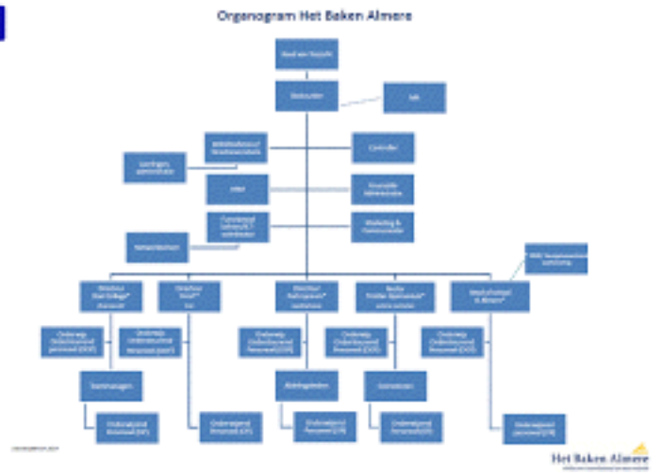
Representation for organisations?

- Which organisation will develop a problem?



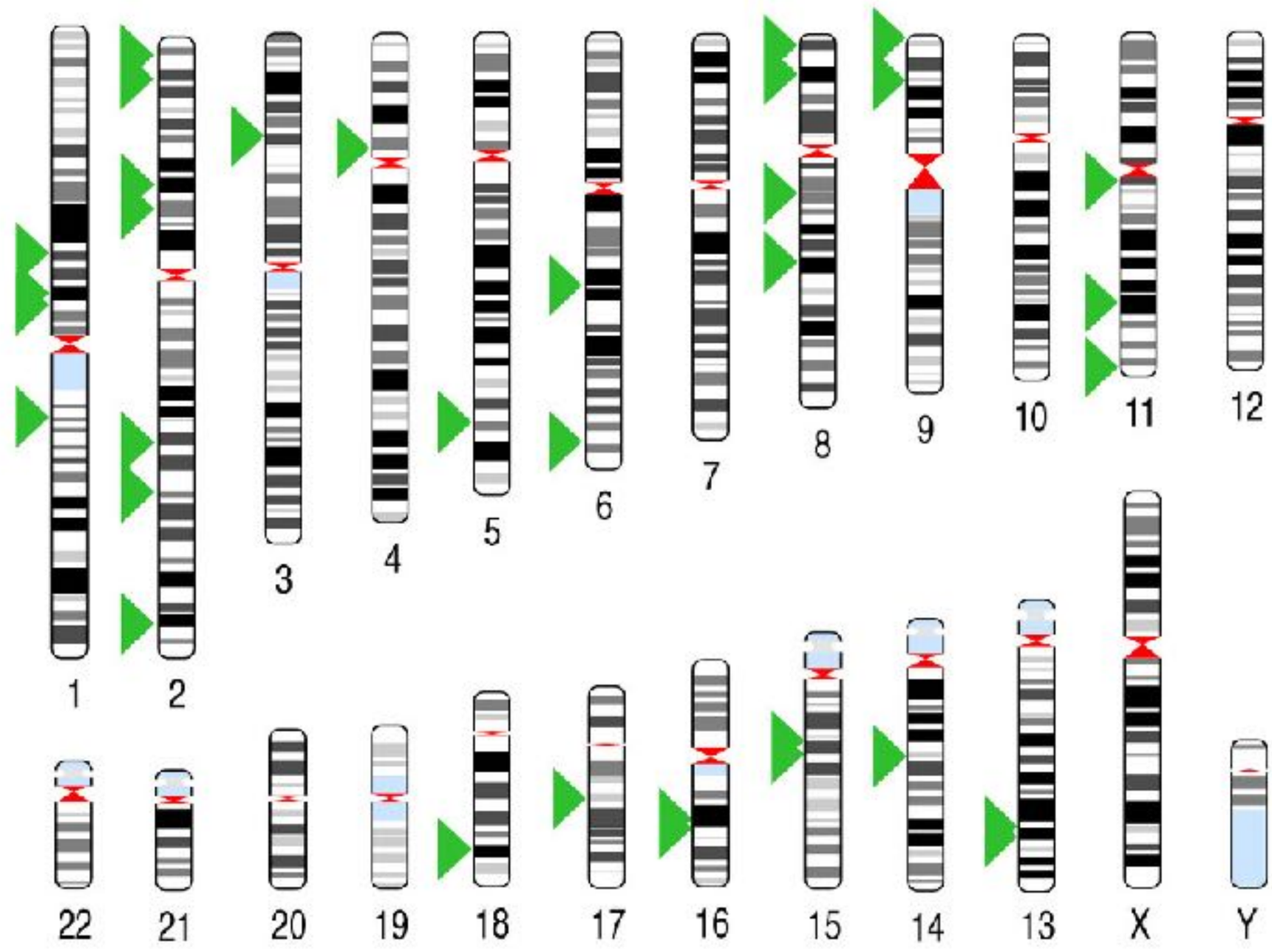






For tumor classification?

- Classify a piece of tumor tissue, given mutations in the genome



Concept to learn?

Multiple-Instance Learning

or:

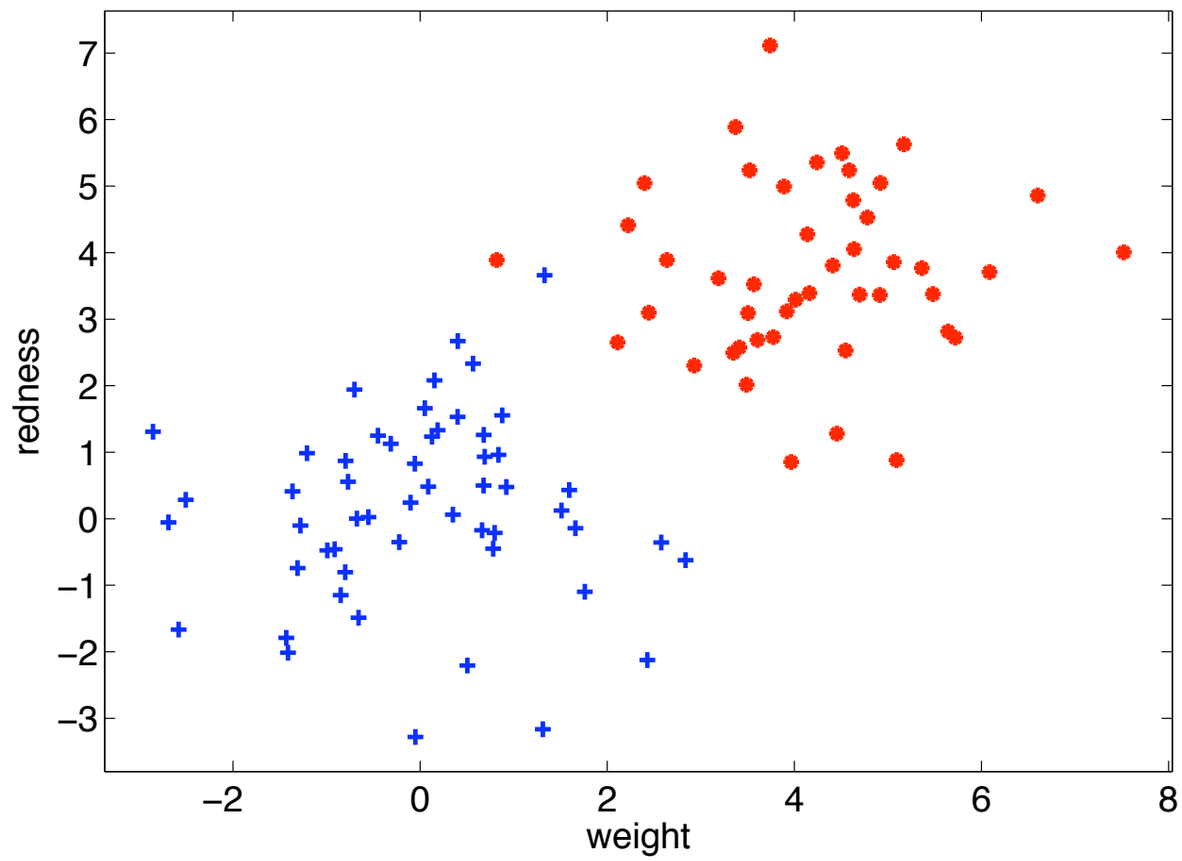
Set Classification

D.M.J. Tax
Pattern Recognition Laboratory
Delft University of Technology

Multiple instance learning (MIL)

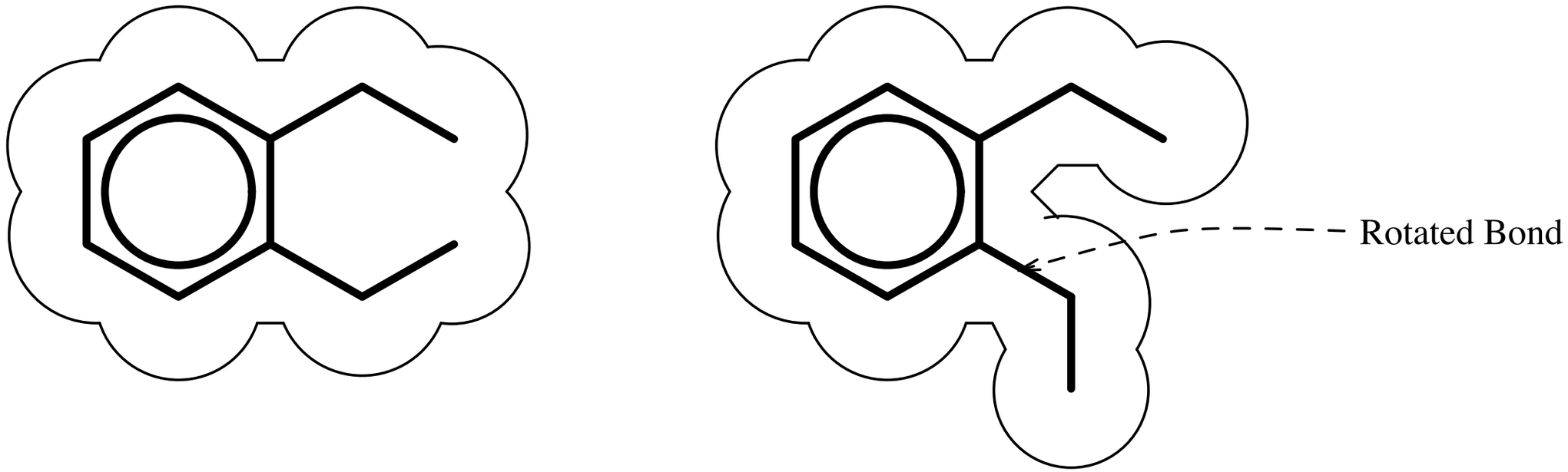
- Multiple-instance learning, background history
- 'Pure' MIL Algorithms:
  - Diverse Density
  - naive approach, MI-SVM, MIL-Boost
- MIL bag-representations
  - Bag vectors
  - Bag similarities
  - 'Bag of Words'
  - MILES
- Some applications/future research questions

Traditional classification



- Objects are represented by a (fixed) set of features
- Objects have (one) label
- Confusion between classes caused by noisy measurements and poor features

Origin: Drug discovery, MUSK

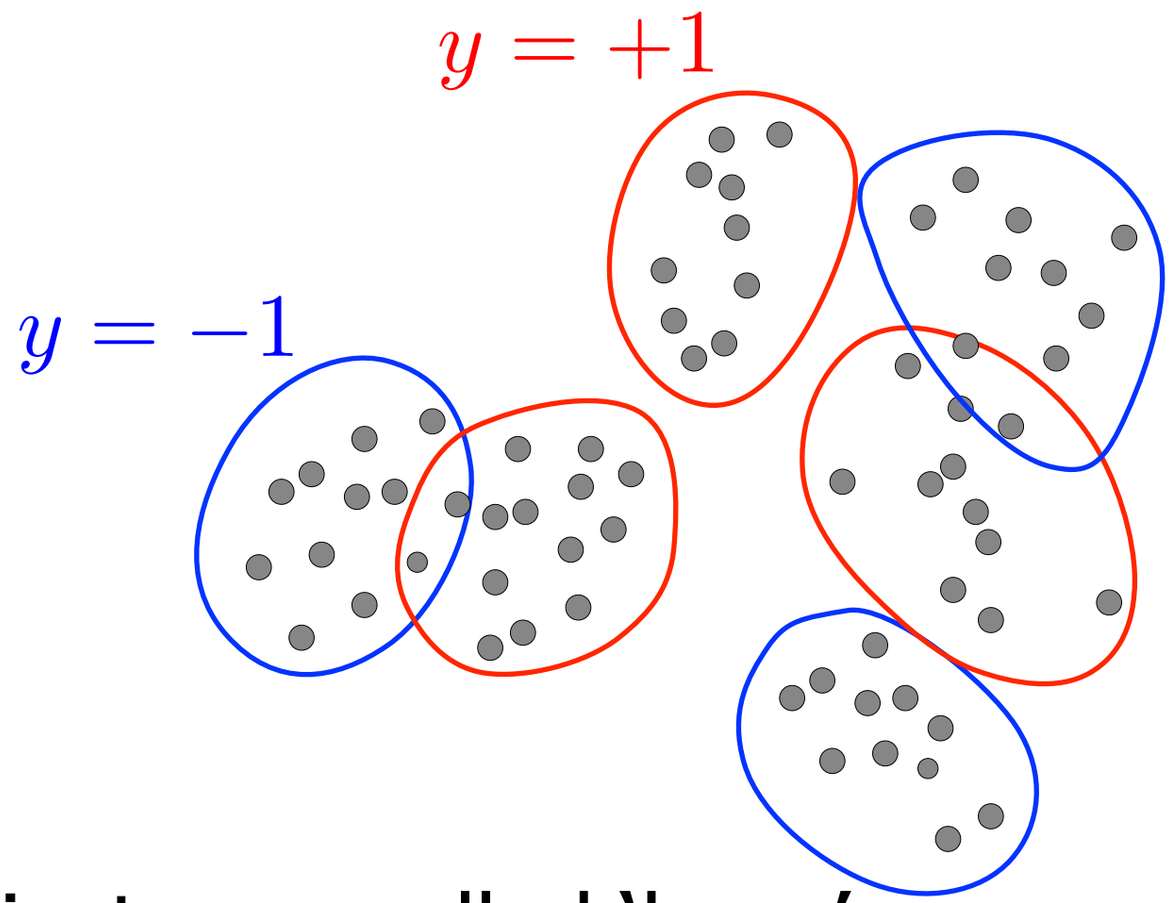


- Predict if a molecule smells ‘musky’
- Chemical formula is fixed, physical shape can change: different **conformations** are possible.
- When one conformation smell musky, the molecule smells musky
- Each conformation is an ‘instance’

Multi(ple)-instance learning


- In multi-instance learning, it is assumed that a set (or **bag**) of objects (**instances**) is labeled
- All instances are represented by feature vectors in the same feature space
- It is a two-class problem:
  - In a negative bag, all instances are assumed to be negative
  - In a positive bag, some (or, at least one) instance is assumed to be positive

Multiple-instance learning



- Objects are called ‘bags’
- Feature vectors are called ‘instances’
- The bags are labeled, **not** the instances

Notation

- Assume we have N bags of instances
- Each bag B_i has n_i instances

$$B_i = \{\mathbf{x}_{i1}, ..., \mathbf{x}_{ij}, ..., \mathbf{x}_{in_i}\}$$

- In training, each bag is labeled

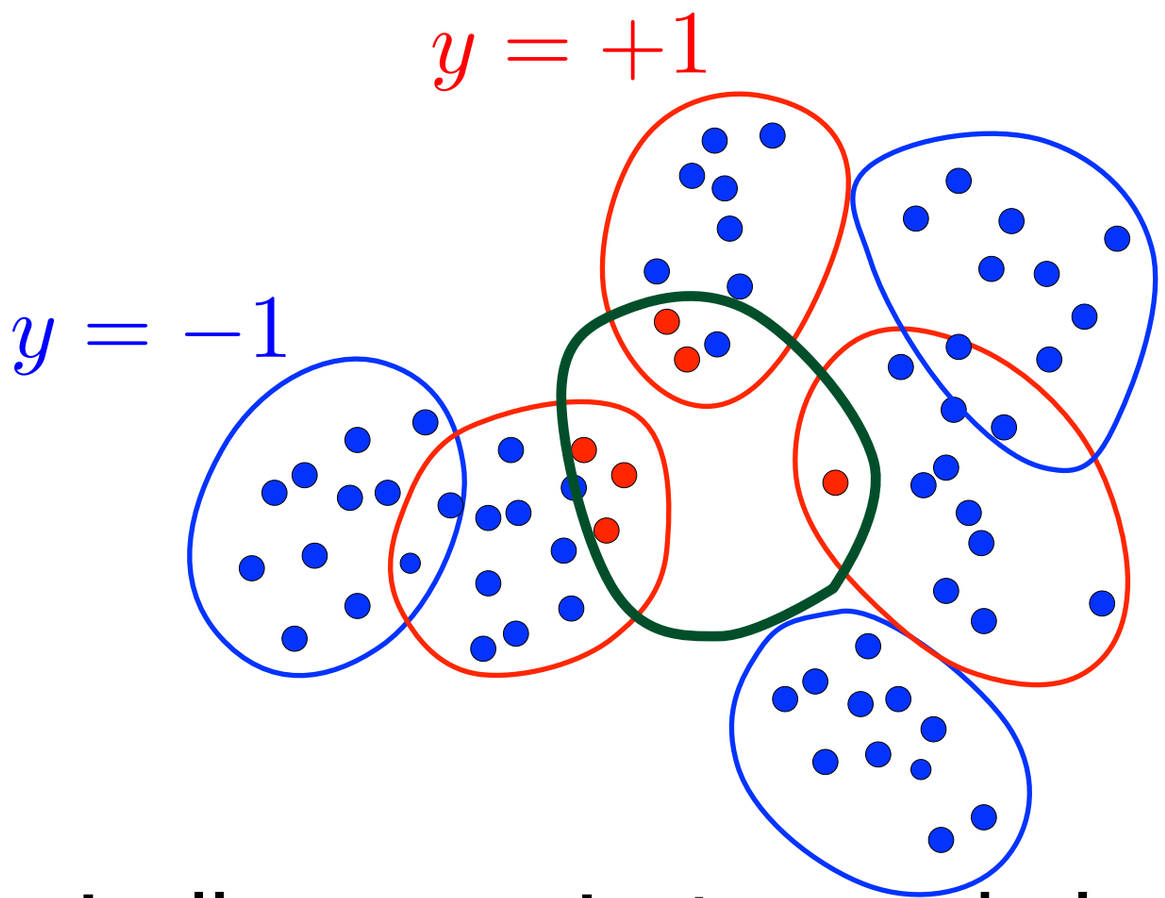
where $\{(B_i, y_i), \quad i = 1, .., N\}$

$$y \in \{-1, +1\}$$

- We are asked to find a bag-classifier:

$$\hat{y}_i = f(B_i)$$

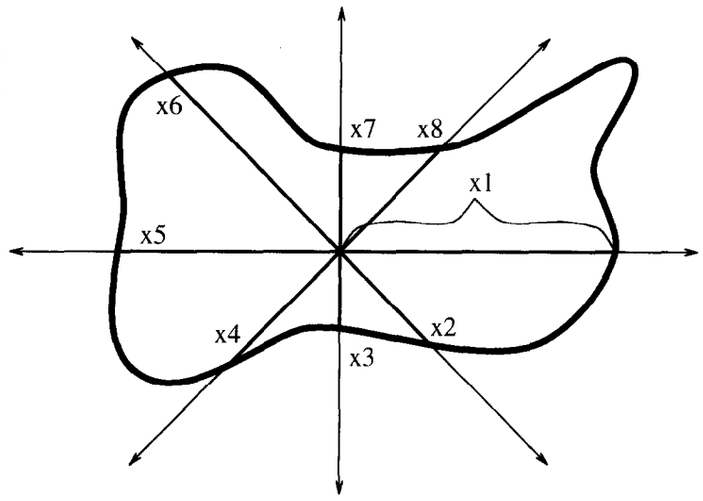
Classical: find the concept



- Classically: some instances belong to a ‘concept’, or not
- Learning means finding the ‘concept’

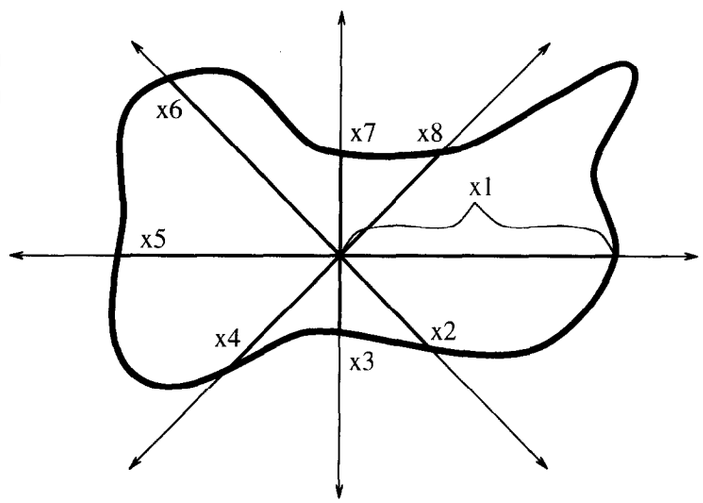
Original solution drug discovery

- Make a 3D model of the molecule: each conformation is another model
- Describe the shape by ‘rays’

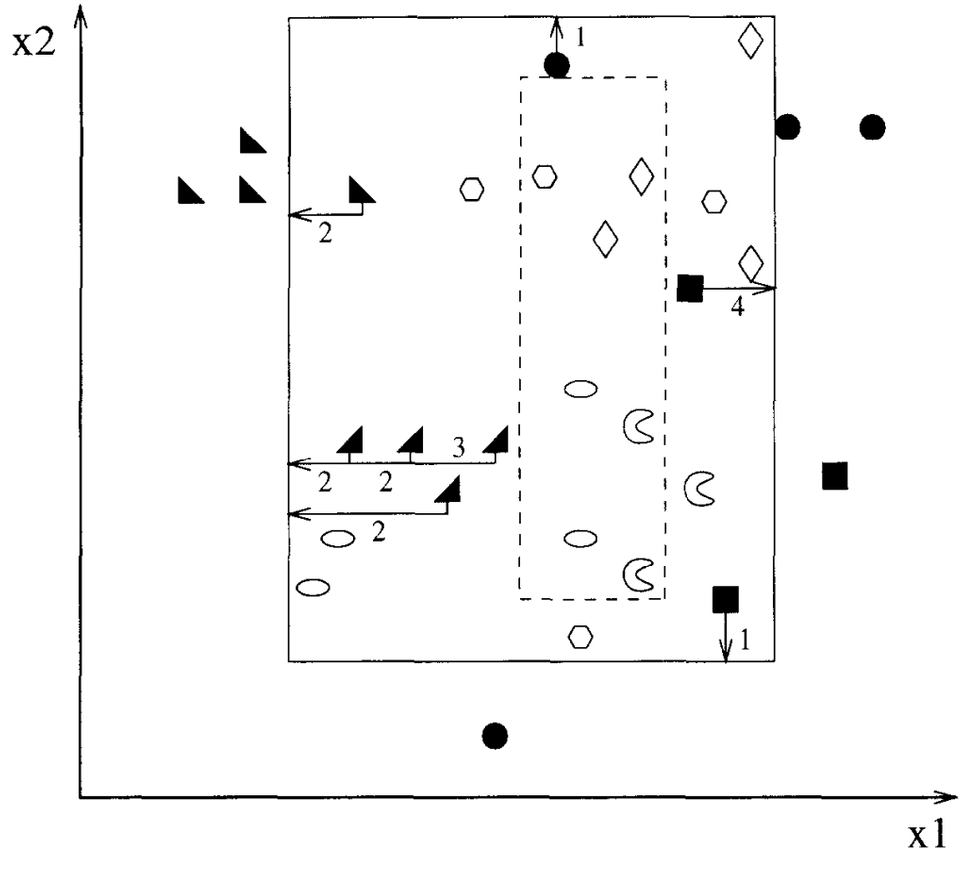


Original solution drug discovery

- Make a 3D model of the molecule: each conformation is another model
- Describe the shape by ‘rays’



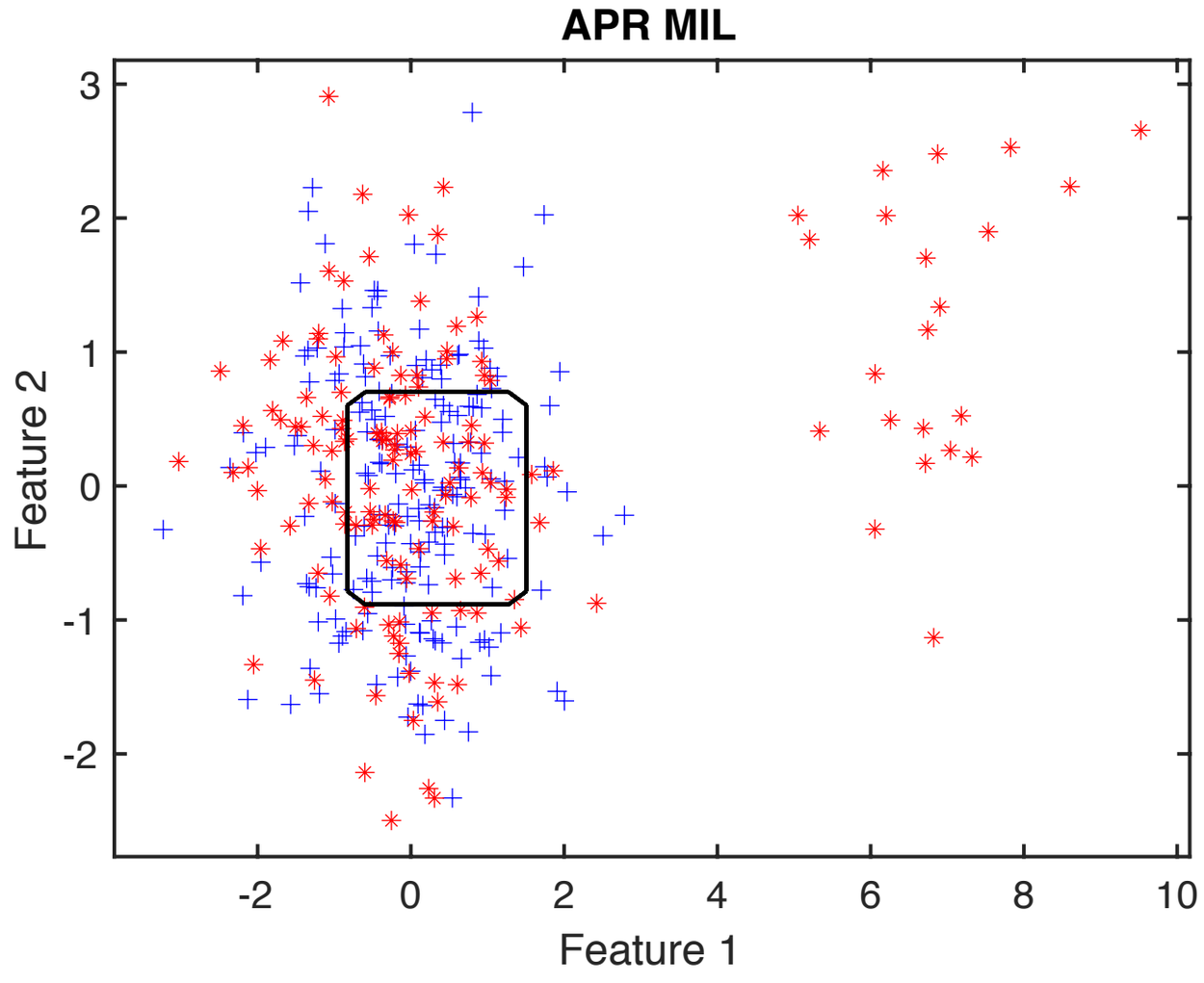
- Now fit the axis-parallel rectangle (see lecture on computational learning theory)
- Take the MIL assumption into account



Solving the multiple instance problem with axis-parallel rectangles
Thomas G. Dietterich, Richard H. Lathrop, Tomás Lozano-Pérez Artificial Intelligence, 1997

On artificial data

- 10 positive, 10 negative bags
- Instances are draw from two Gaussian distributions
- Positive bags have instance in the right Gaussian (the concept)
- To plot the decision boundary: use bags of size 1



MIL methods

- Dietterich's approach did not work so well
- Many different approaches:
  - Often quite heuristic (but it works)
  - Sometimes nicely founded (but hard to optimise)

- Naive approach
- Diverse Density
- MI-SVM
- Boosting approach
- MILES and friends

The 'naive' approach

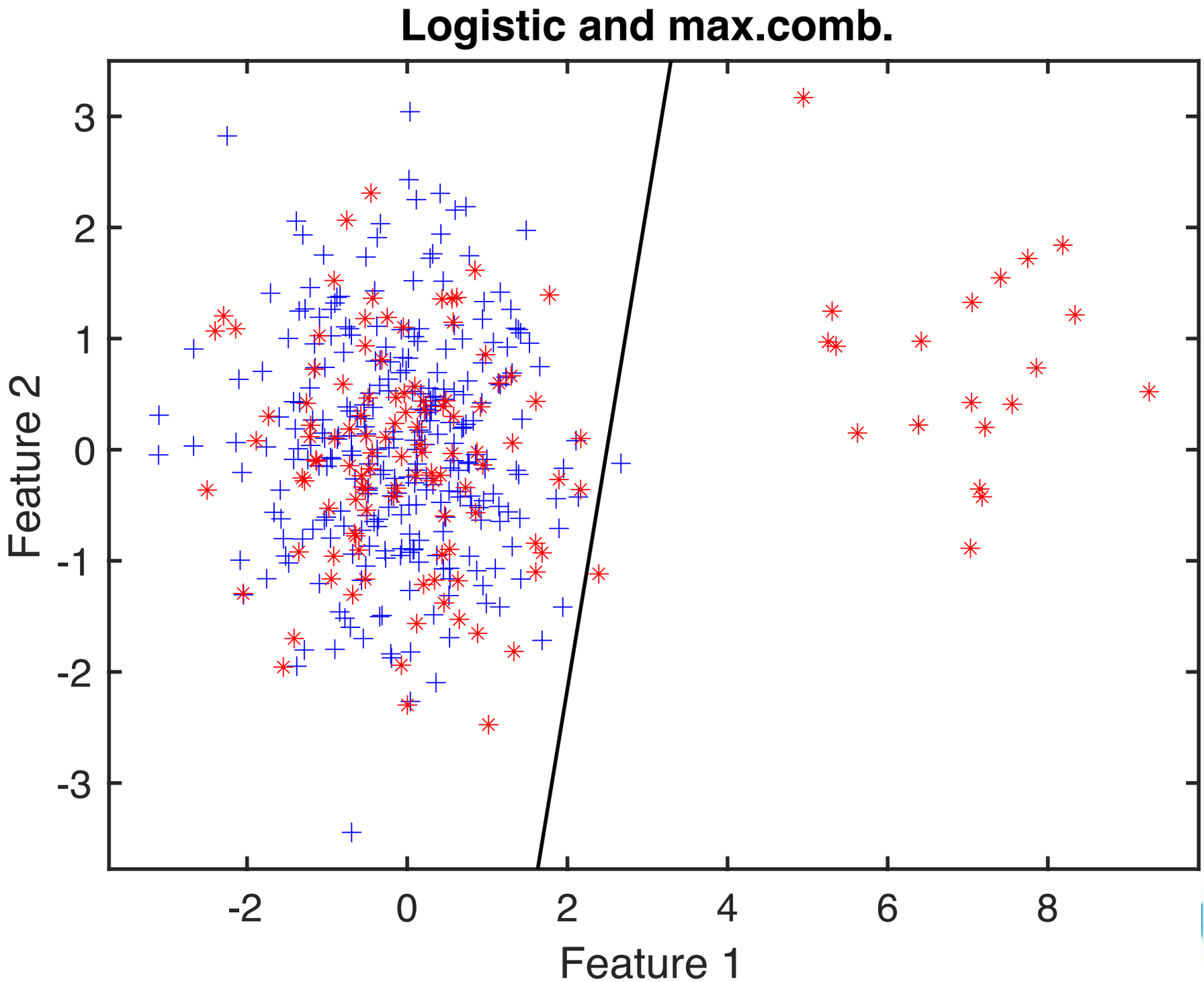
- Many people do multiple-instance learning, but do not realize it
- Often the straightforward approach is followed:

1. Copy the bag-labels to the instances $y_{ij} = y_i, \quad \forall j$
2. Train a standard classifier $\hat{y}_{ij} = h(\mathbf{x}_{ij})$
3. Combine the classifier outcomes on the instance using a combining rule $\hat{y}_i = C(h(\mathbf{x}_{i1}), ..., h(\mathbf{x}_{in_i}))$

- Often the max-combining rule is used
- Suffers from heavily overlapping classes

Simple combination of logistic cl.

- Max-combining of logistic:



Diverse density

- A probabilistic approach to MIL

- Idea:
  if none of the instances are from the concept: negative bag

  if one, **or more**, instances are from concept: positive bag

Diverse density

- For positive bags, we want to optimize

$$Pr(+|B) = 1 - Pr(-|B)$$

  and for negative bags

$$Pr(-|B)$$

- For negative bags, **all** instances should have high prob. of negative class:

$$Pr(-|B_i) = \prod_{j \in i} Pr(-|\mathbf{x}_{ij})$$

Noisy-OR combination

- The negative class posterior is modeled:

$$Pr(-|\mathbf{x}_{ij}) = 1 - Pr(+|\mathbf{x}_{ij}) = 1 - \exp\left(-\frac{(\mathbf{x}_{ij} - \mathbf{t})^2}{s^2}\right)$$

- For positive bags:

$$Pr(+|B_i) = 1 - \prod_{j \in i}\left(1 - \exp\left(-\frac{(\mathbf{x}_{ij} - \mathbf{t})^2}{s^2}\right)\right)$$

A framework for multiple instance learning

Oded Maron, Tomas Lozano-Pérez NIPS, 1998

Diverse Density

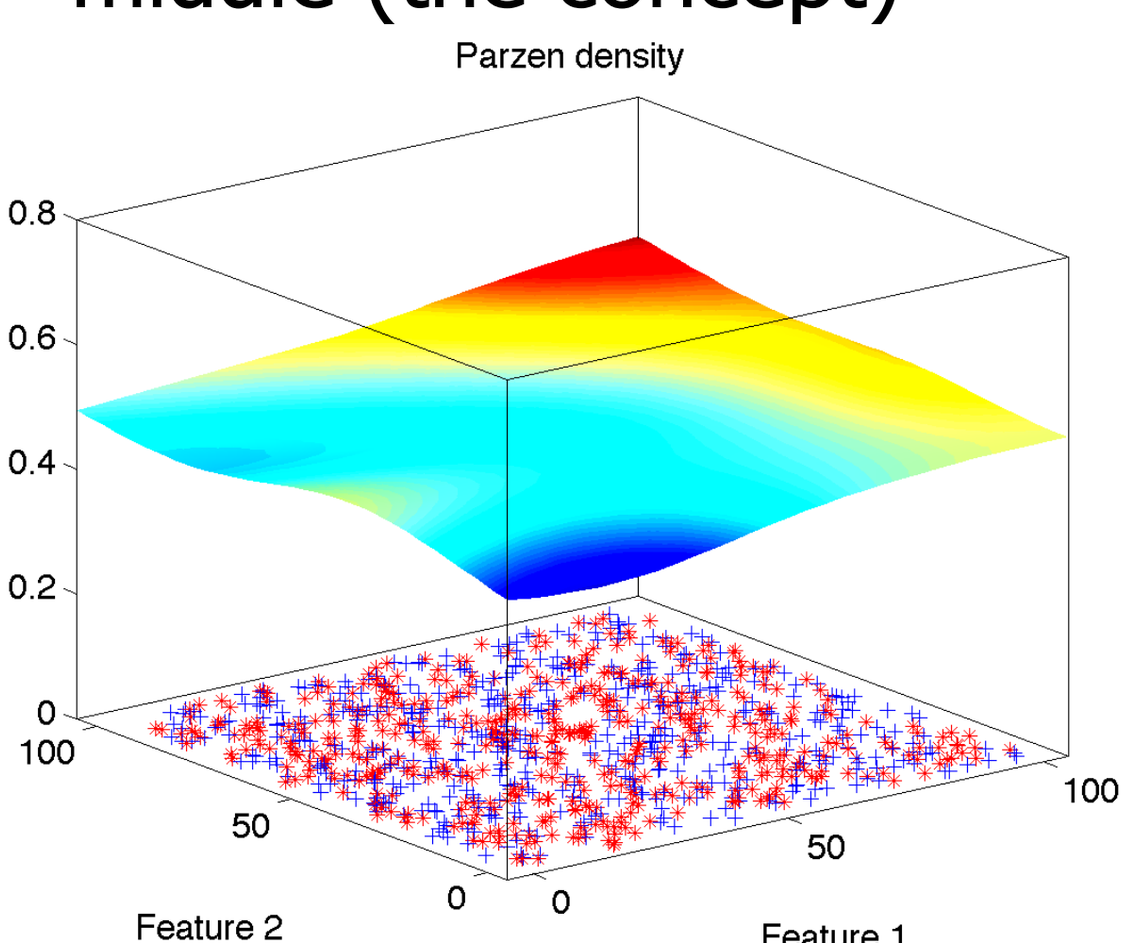
- How to optimize this? Maximize the likelihood:

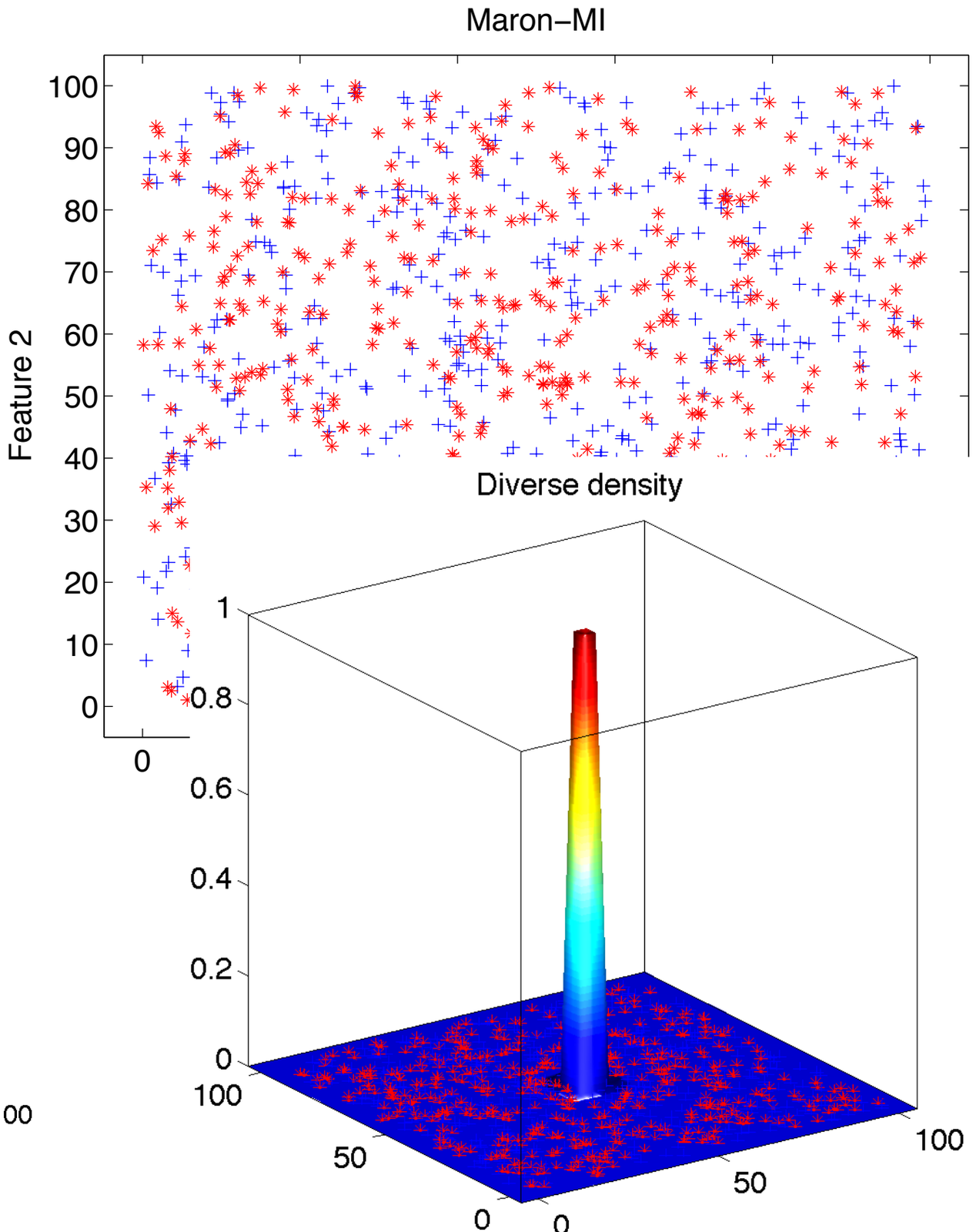
$$L = \prod_{i \in +} Pr(+|B_i) \prod_{i \in -} Pr(-|B_i)$$

- We need to optimize the location of the concept $\mathbf{t}$, and the scale per feature $\mathbf{s}$
- It is hard to optimize. A gradient ascent procedure is proposed. L has many local optima, so multiple initializations are necessary to find a good one
- In practice this is **very** slow, but it works

Artificial example

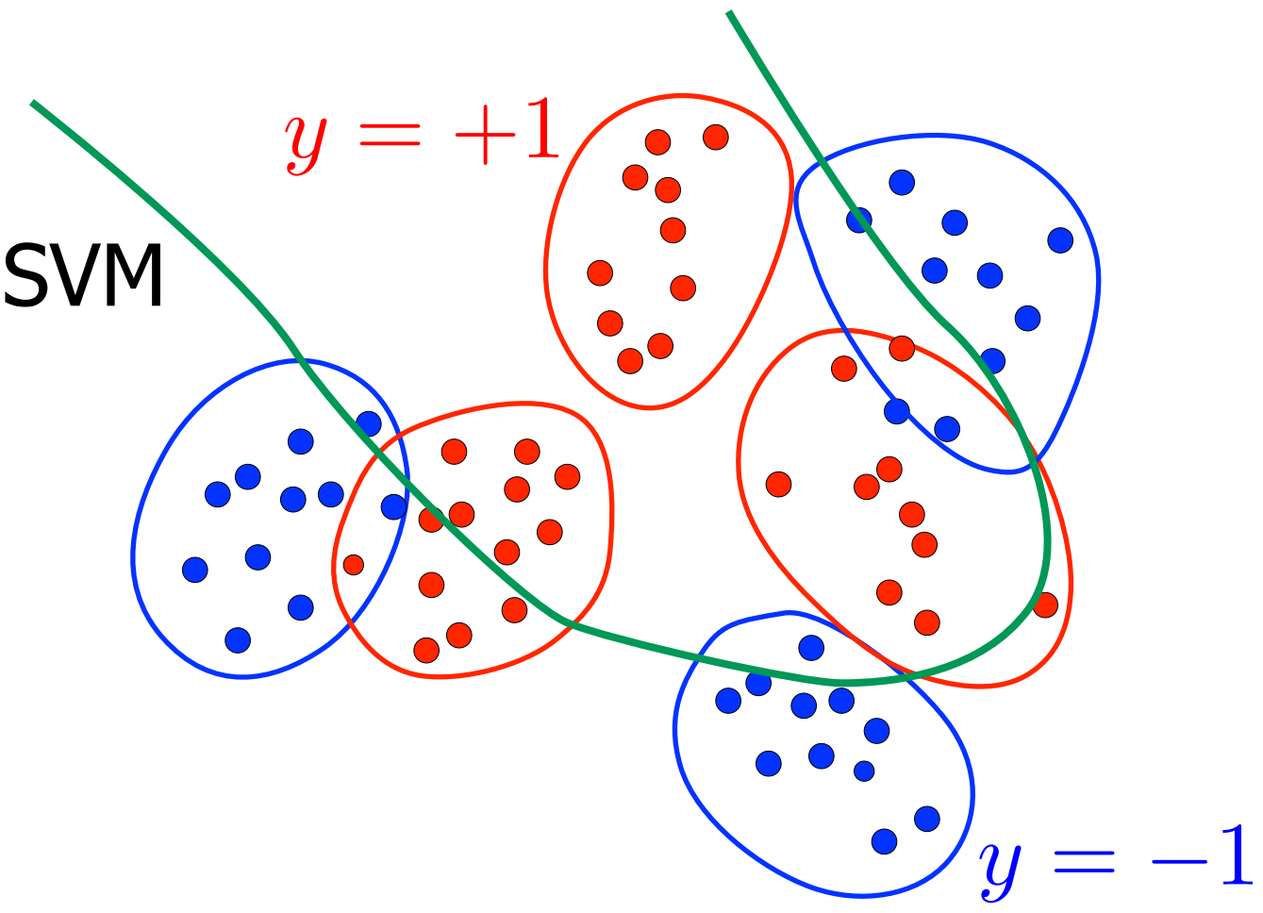
- Generate 10 positive and 10 negative bags
- The positive bags have at least one instance in a small square in the middle (the concept)





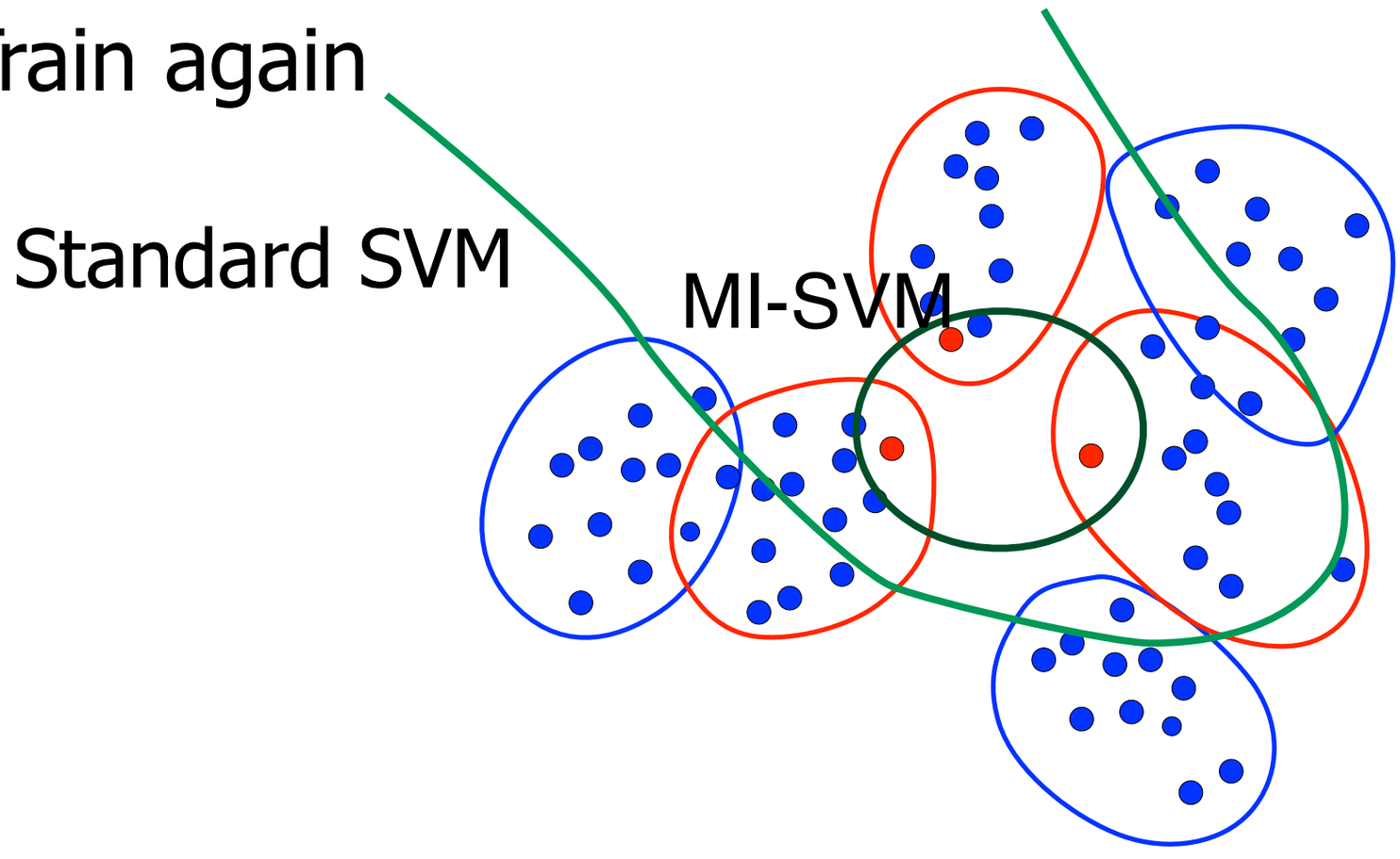
MI-SVM

- Extension of a standard Support Vector machine
- In the first iteration, copy the bag label to the instance label
- Train a standard SVM



MI-SVM

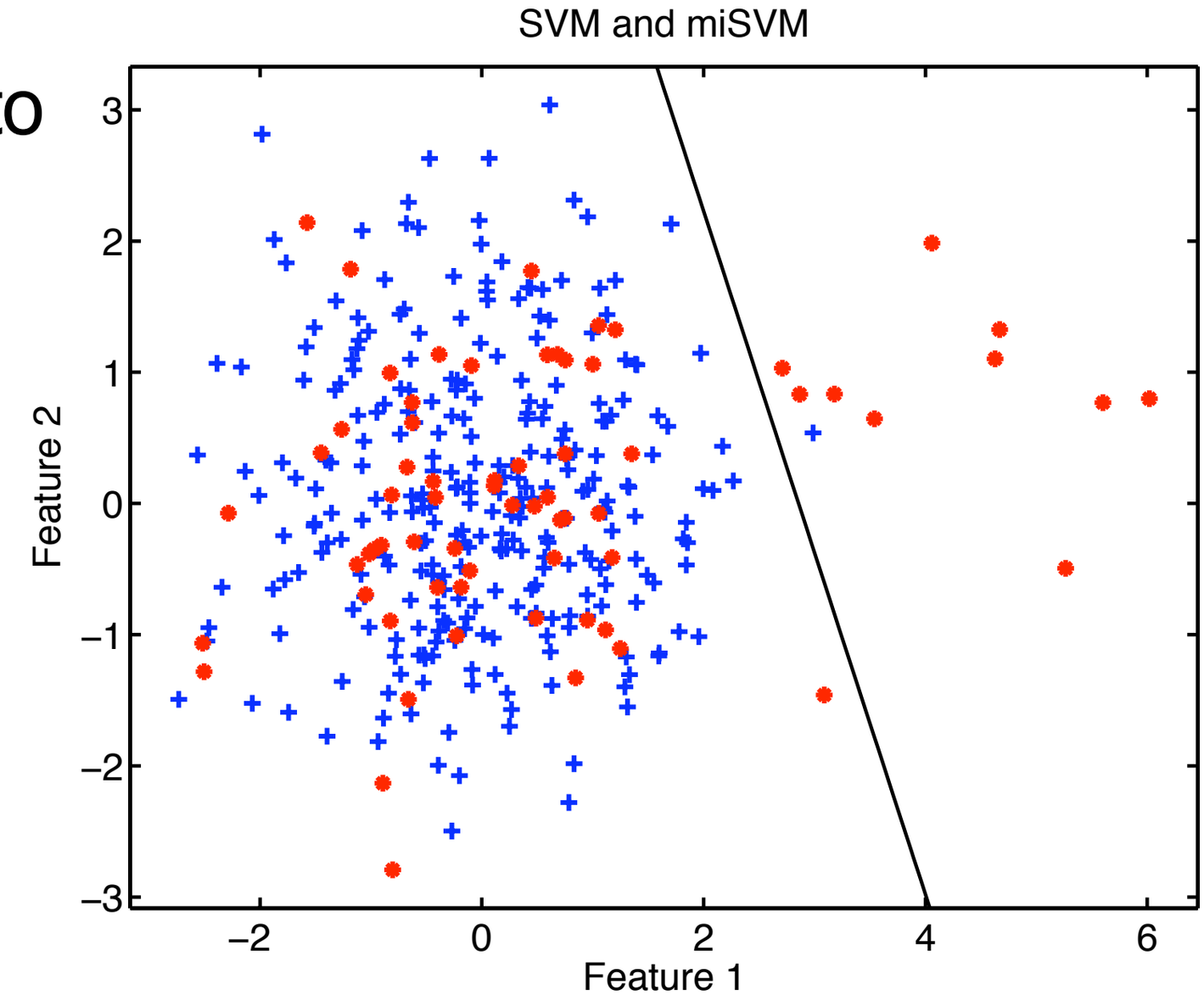
- Extension of a standard Support Vector machine
- After the first iteration, allow that just a **single** instance in a positive bag is on the correct side of the decision boundary
- Train again



Support vector machines for multiple instance learning
S.Andrews, I.Tsochantaridis, T.Hofmann NIPS, 2002

On artificial data

- The miSVM manages to detect the ‘concept’



- By the overwhelming presence of the negative class, the standard SVM classifies everything as negative

Generalising MI-SVM?

- Why would you only use an SVM here?
- Why not generalise?
- How…?

MIL-Boost

- Assume we have a model $\mathbf{h}(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
- Assume we have a performance

  where $\mathcal{L} = \prod_{i \in +} p(+|B_i) \prod_{i \in -} p(-|B_i)$

$$p(+|B_i) = \mathrm{softmax}_i(\mathrm{sigmoid}(\mathbf{h}(\mathbf{x}_{ij})))$$

- Softmax can be define in different ways, for instance

$$\left(\frac{1}{m}\sum_{\ell} v_{\ell}^{r}\right)^{\frac{1}{r}}$$

- Then optimising the loss iteratively

Simultaneous Learning and Alignment: multi-instance and Multi-Pose Learning

Boris Babenko, Piotr Dollar, Houwen Tu, Serge Belongie Workshop on faces in 'real-life' images, 2008

MIL Boost

Input: Dataset $\{X_1, \ldots, X_n\}$, $\{y_1, \ldots, y_n\}$,$y_i \in \{-1, 1\}$

1: **for** $t = 1$ to T **do**

2: Compute weights $w_{ij} = -\frac{\partial \mathcal{L}}{\partial \mathbf{h}_{ij}}$

3: Train weak classifier h_t using weights $|w_{ij}|$

$h_t = \text{argmin}_h \sum_{ij} \mathbf{1}(h(x_{ij}) \neq y_i)|w_{ij}|$

4: Find α_t via line search to minimize $\mathcal{L}(\mathbf{h})$

$\alpha_t = \text{argmin}_\alpha \mathcal{L}(\mathbf{h} + \alpha h_t)$

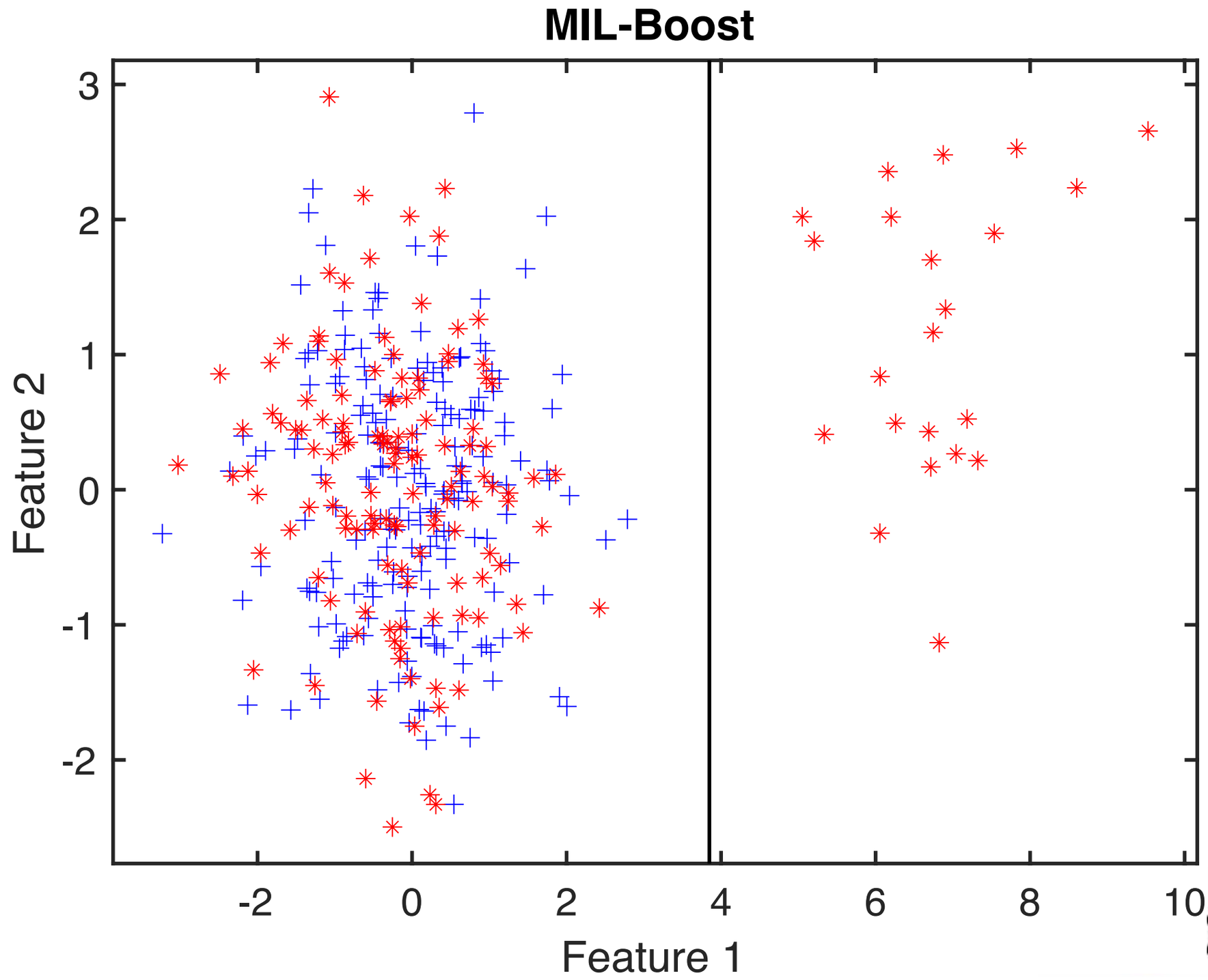
5: Update strong classifier $\mathbf{h} \leftarrow \mathbf{h} + \alpha_t h_t$.

6: **end for**

- Here, the tough part is to derive $-\frac{\partial \mathcal{L}}{\partial h(x_{ij})}$
- Chain-rule after chain rule…

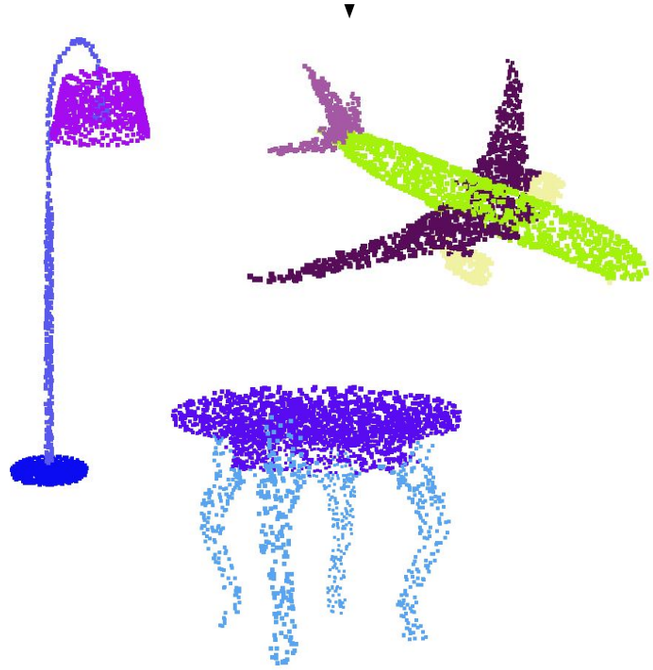
On artificial data

- Weak classifier is a decision stump:

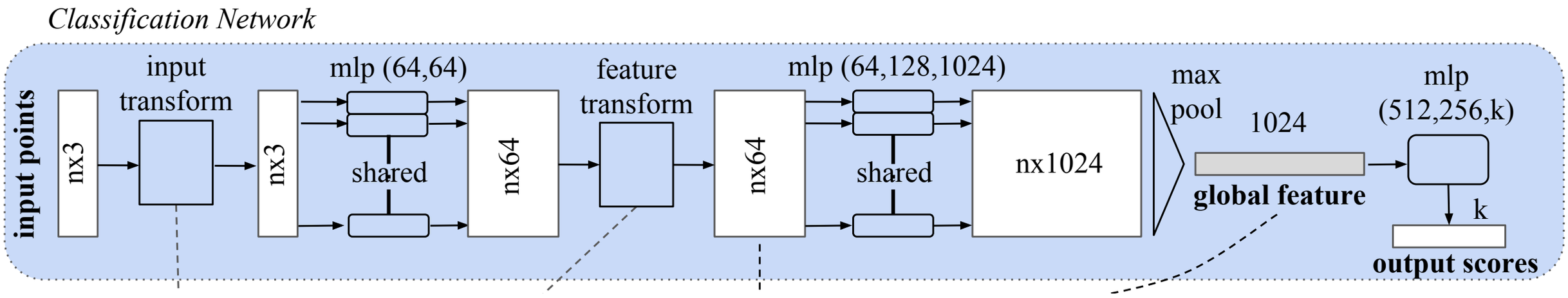


PointNet

- Deep learning approach to classify/segment point clouds



- Each 3D point is an instance
- The n instances are combined using a max-pooling/combining
- Does **one** instance determine the bag label?



Conclusion first approaches MIL

- General idea: detect concept
- Diverse Density: probabilistic implementation, works, but (very) slow
- MI-SVM, naive approach: heuristic approach, very fast, often works reasonably
- MIL-Boost: more flexible, also using the noisy-OR idea of Diverse Densiy
- Point-net: a deep learning approach for point clouds
- Concept detection allows for interpretation of the classification
- Competing idea: general bag classification, avoid concept detection

Bag representation

- Bag approaches:
  make new representations that capture all bag info

- Extract features from a bag of instances:

$$\mathbf{x}(B) = (\text{mean}(B), \min(B), \max(B), \text{cov}(B), ...)$$

- Define a (dis-)similarity measure between bags:

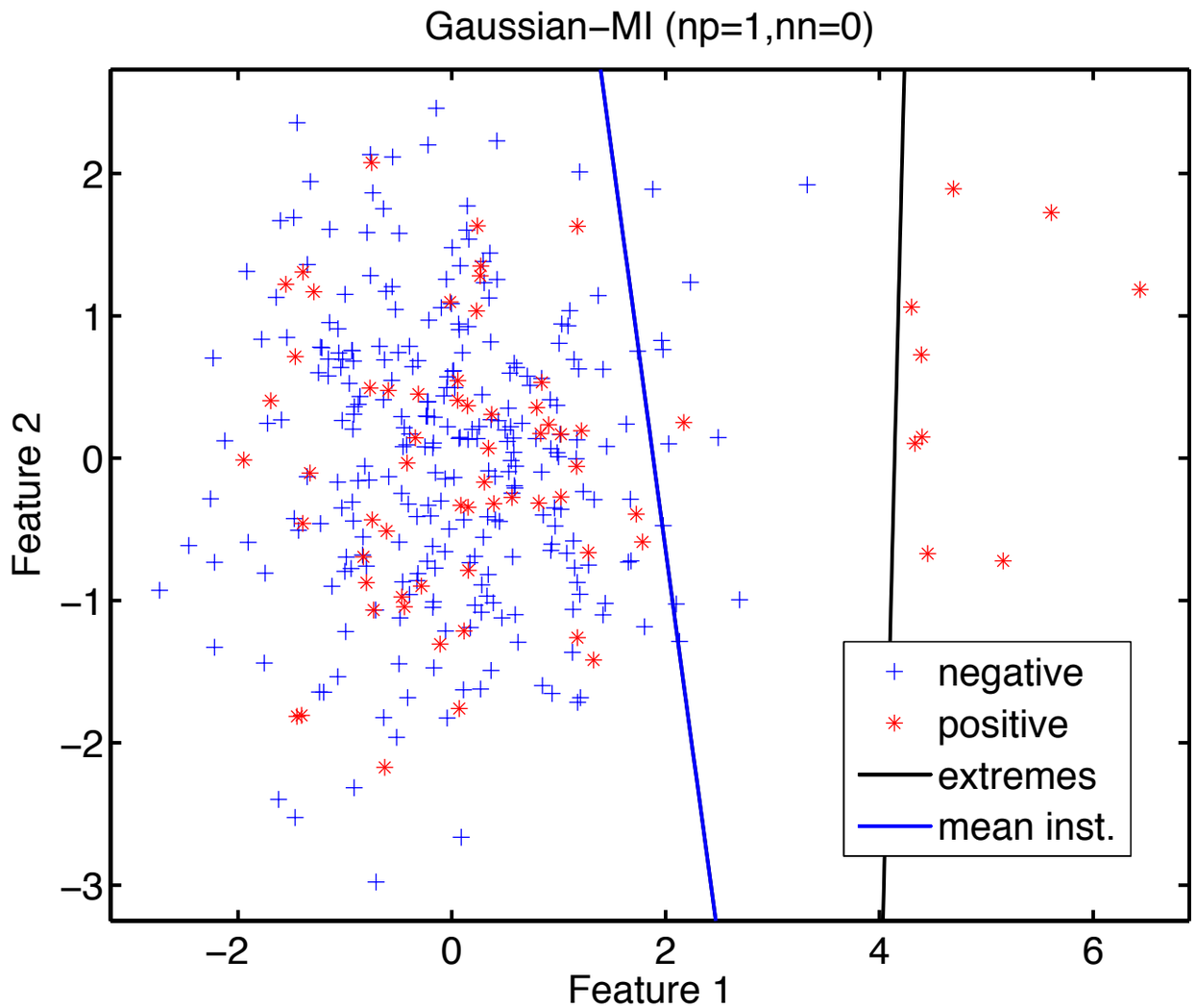
$$K(B_i, B_j) = (m(B_i) - m(B_j))^T \Sigma^{-1} (m(B_i) - m(B_j))$$

$$m(B_i) = \text{mean}(B_i)$$

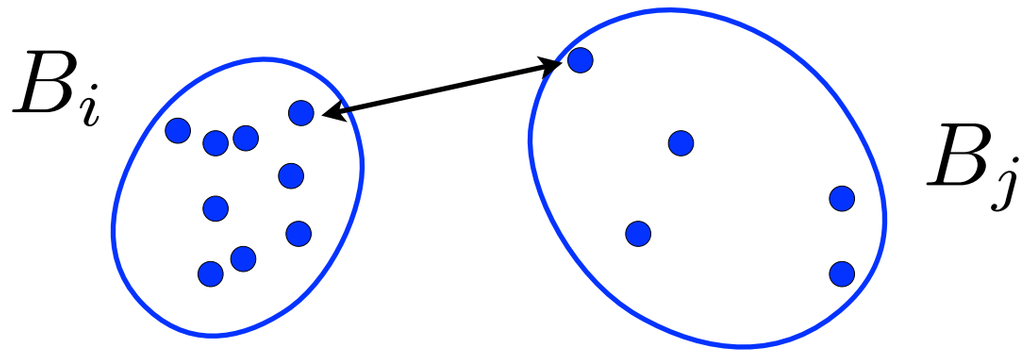
- A Bag-of-Words approach: ...

Simple feature representation

- Use the mean-instance, or the min.and max. values per bag, and train an SVM
- Not perfect, but very fast and simple:



Bag dissimilarities



- Define a distance between bags:

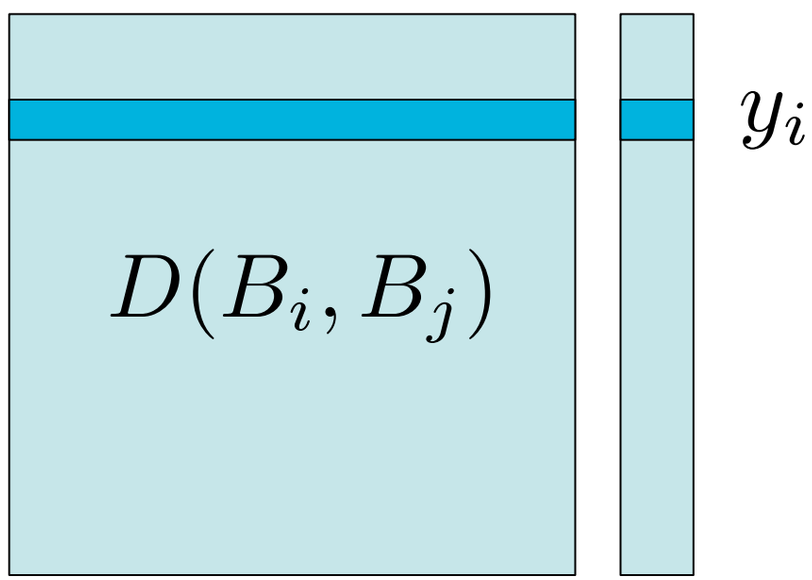
$$D(B_i, B_j) = \min_{k,l} \|\mathbf{x}_{ik} - \mathbf{x}_{jl}\|^2$$

- or

$$D(B_i, B_j) = \max\left(\min_{l} \|\mathbf{x}_{i} - \mathbf{x}_{jl}\|^2, \min_{l} \|\mathbf{x}_{j} - \mathbf{x}_{il}\|^2\right)$$

- or ...

Bag dissimilarities

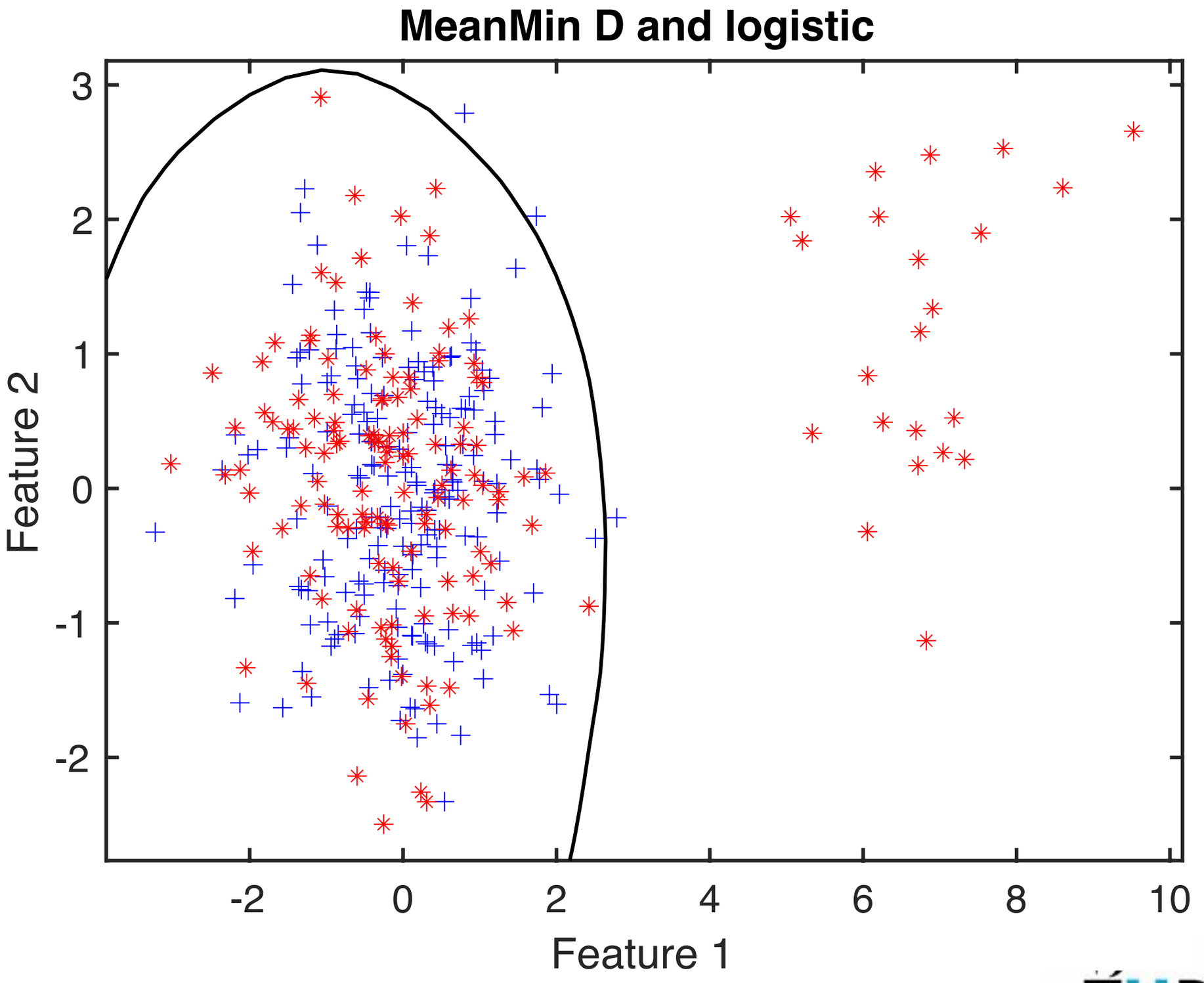


- Given a bag-distance matrix, train any classifier you want!

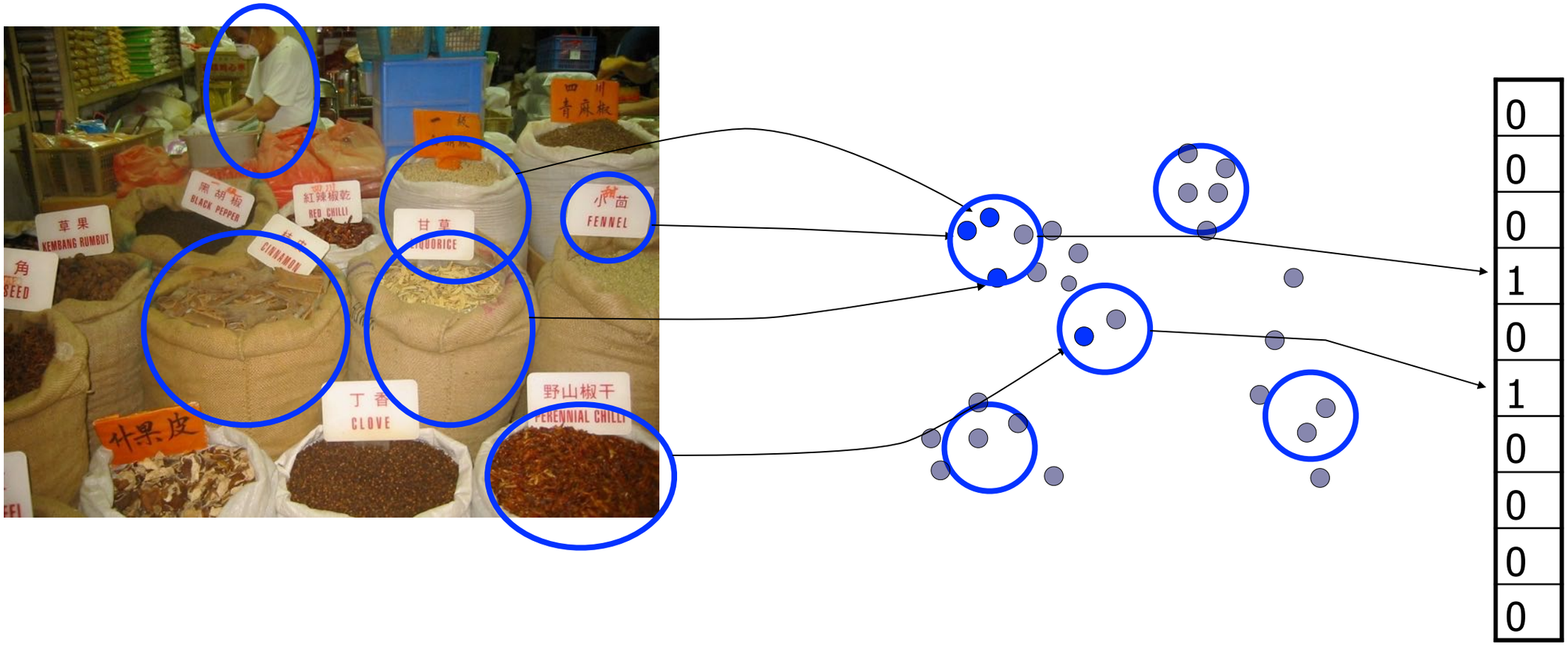
$$\hat{y}_i = f([D(B_i, B_1), ..., D(B_i, B_N)])$$

- For evaluating a new bag, you need to compute the distances to the original prototype

On artificial data

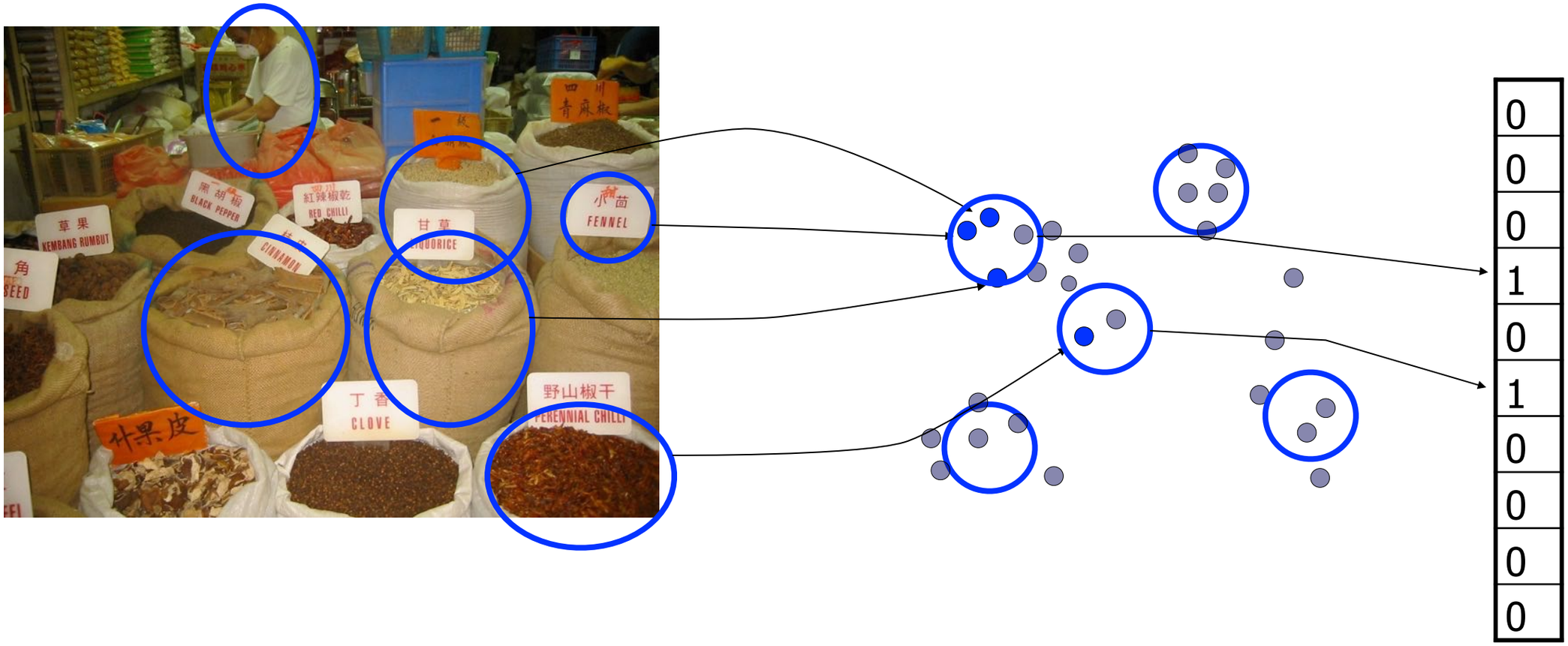


Bag-of-... representation



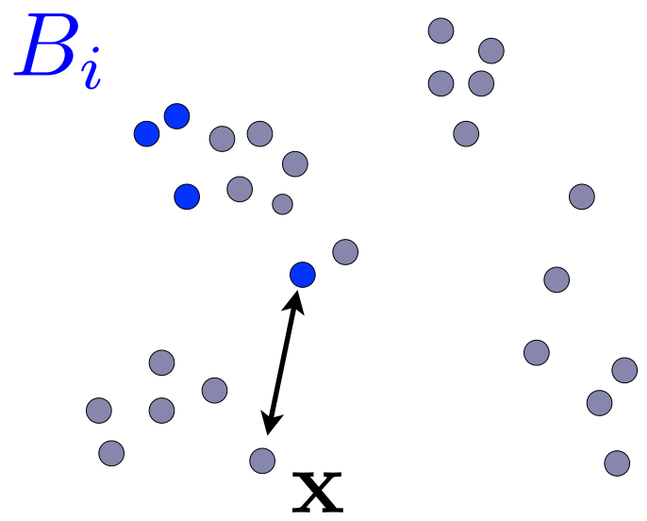
- A dictionary is constructed (often by clustering of all instances)
- The presence of a (visual) word is counted
- A bag is represented by a histogram

Bag-of-… representation



- The histogram often has many bins: high-dimensional feature vector!
- Typically, a support vector classifier is trained
- The clustering typically does not involve labels

MILES



- Instead of distance to cluster centers
  - use ALL instances as clusters
  - use a similarity to the instance

$$s(B_i, \mathbf{x}) = \max_j \exp\left(-\frac{\|\mathbf{x}_{ij} - \mathbf{x}\|^2}{\sigma^2}\right)$$

$$\mathbf{m}(B_i) = [s(B_i, \mathbf{x}_1), ..., s(B_i, \mathbf{x}_N)]$$

Classifiers on the bag representation

- Note that bag representations typically are large (high dimensional)
- Need for robust, sparse, classifiers

- Typically used is Liknon, or L_1-support vector classifier:

such that

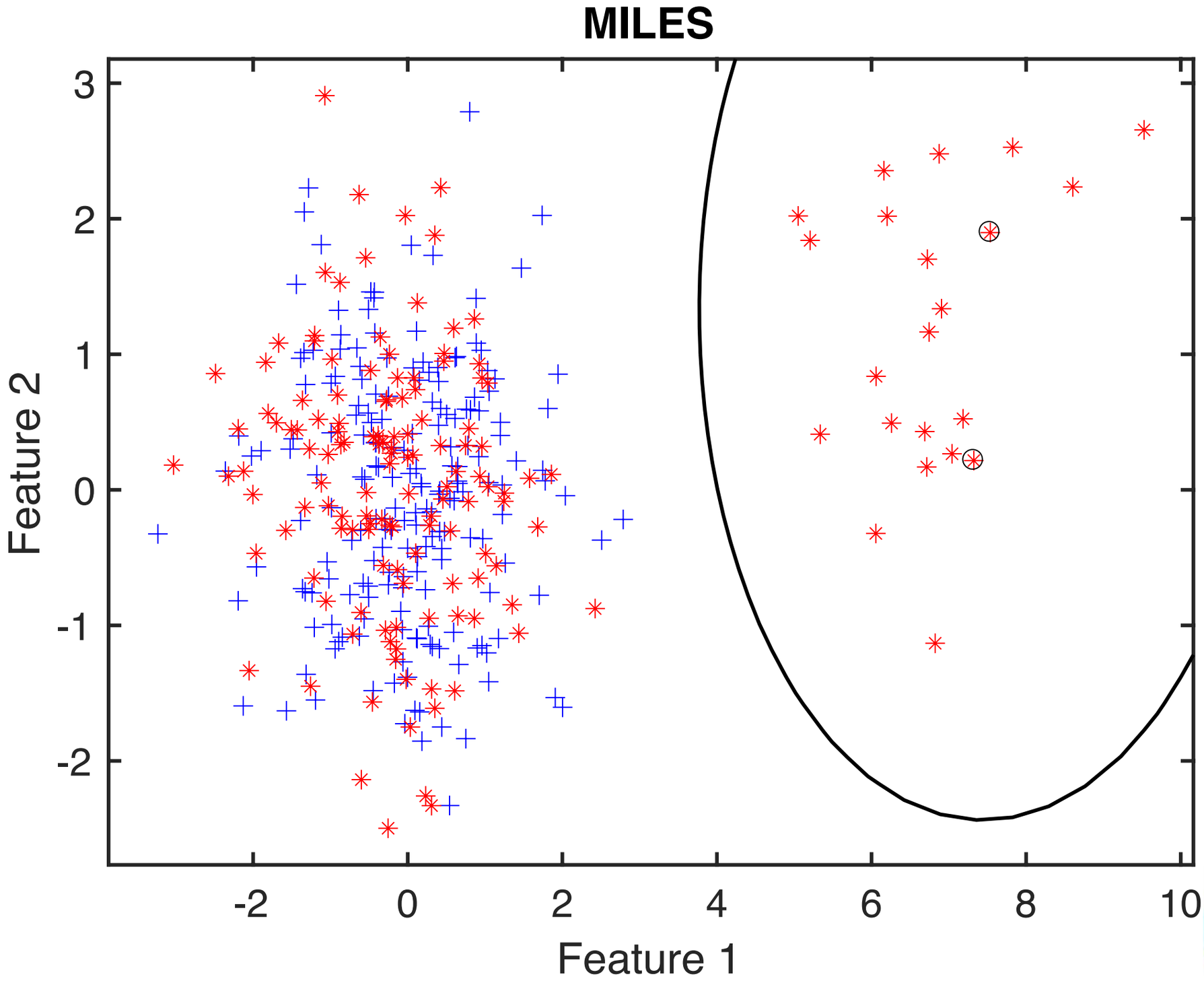
$$\min |\mathbf{w}|_1 + C \sum_{i=1}^{N} \xi_i$$

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i$$

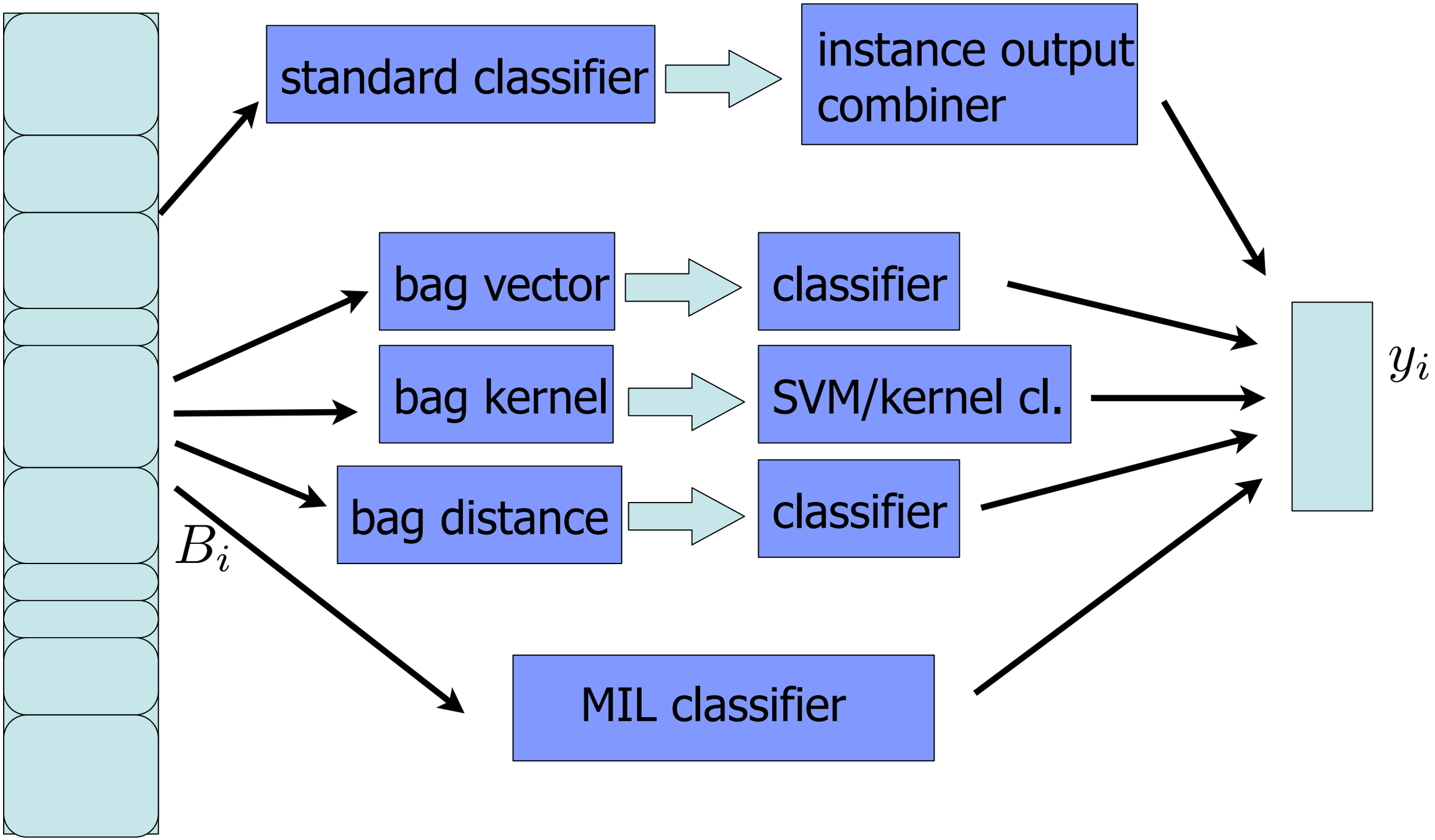
MILES: Multiple-instance Learning via Embedded Instance Selection
Y. Chen, J.Bi, J.Z.Wang Transactions on PAMI, 2006

On artificial data

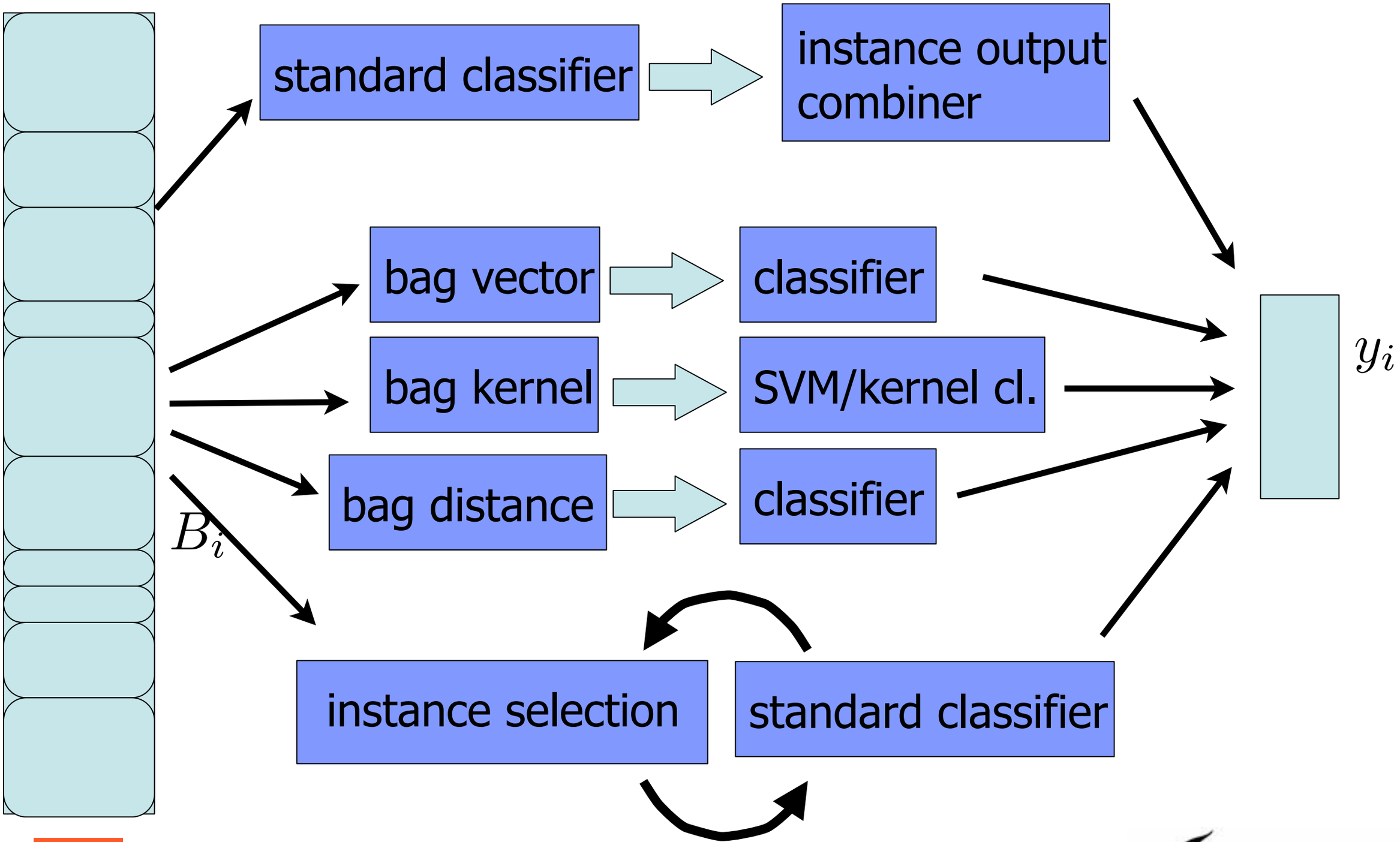
- Only two prototype instances selected:



Overview approaches



Overview approaches



Many other approaches

- Rule-based classifiers (RELIC, FOIL, ...)
- Expectation-Maximization Diverse Density (extension of the Diverse Density)
- Citation-kNN, a variant of the k-nearest neighbor classifier
- (probably someone already proposed Deep MIL classifiers, or convolutional MIL neural networks…)
- …

Some results on MUSK

- See http://homepage.tudelft.nl/n9d04/milweb/ :-)

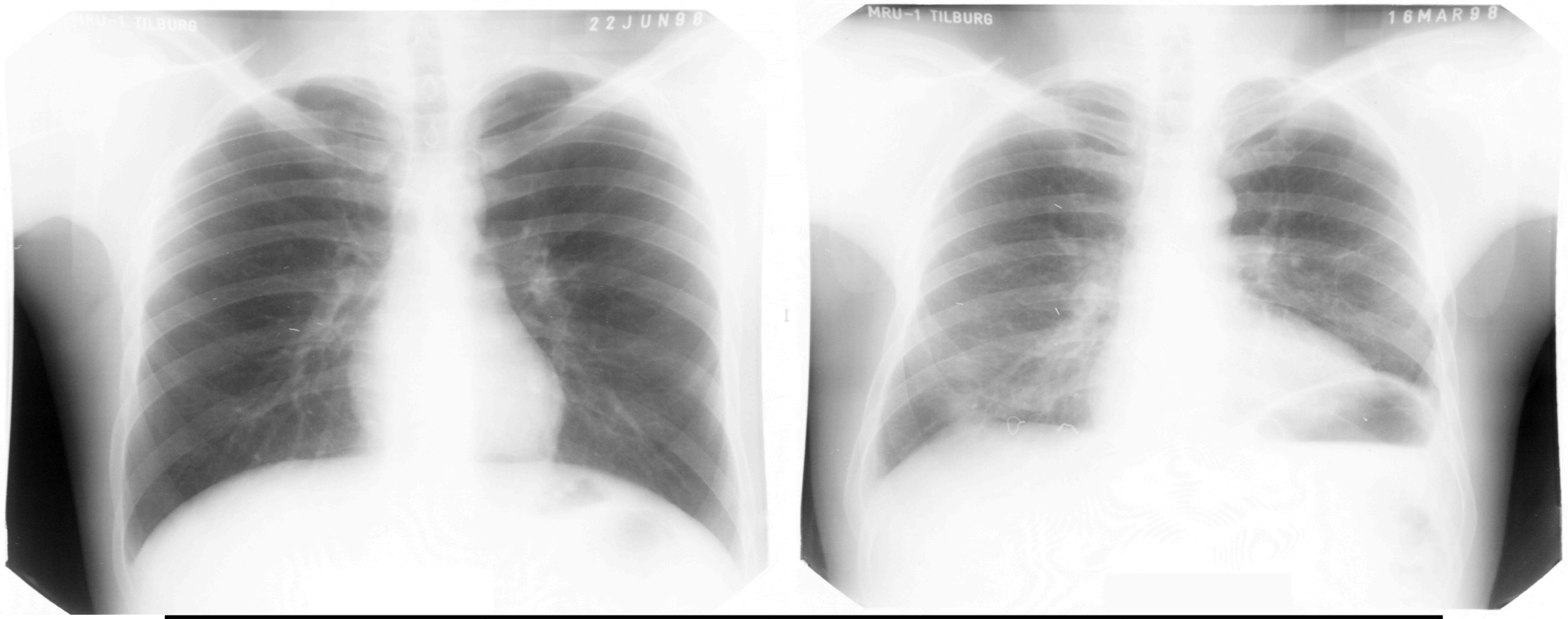
Experiments

- Interstitial Lung Disease
- Overlapping speech classification
- Action-Unit classification
- (Tumor prediction)
- (Company consultancy)

- Future…?

ILD/Tuberculosis detection

- Try to predict if a patient has a lung disease



healthy	ill	instances	features
241	147	$\{1.7, 1.0\}x10^6$	158

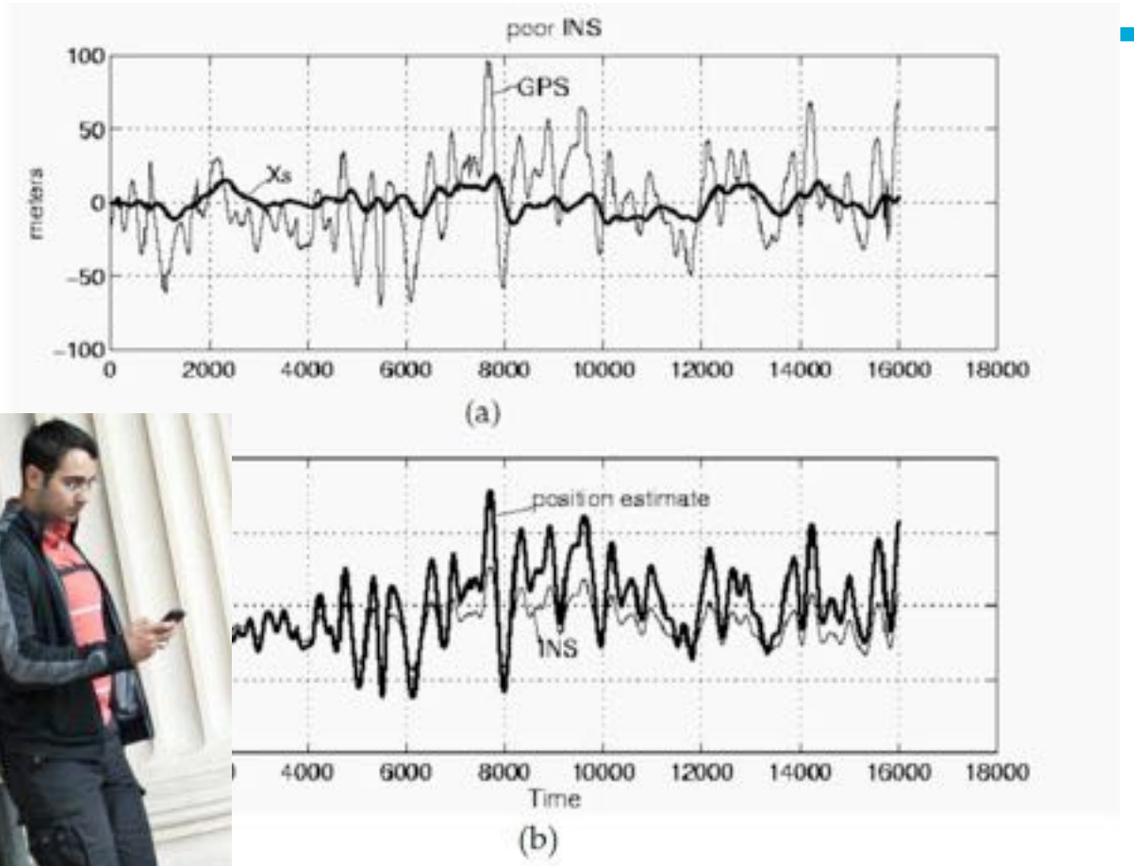
Clean-speech detection

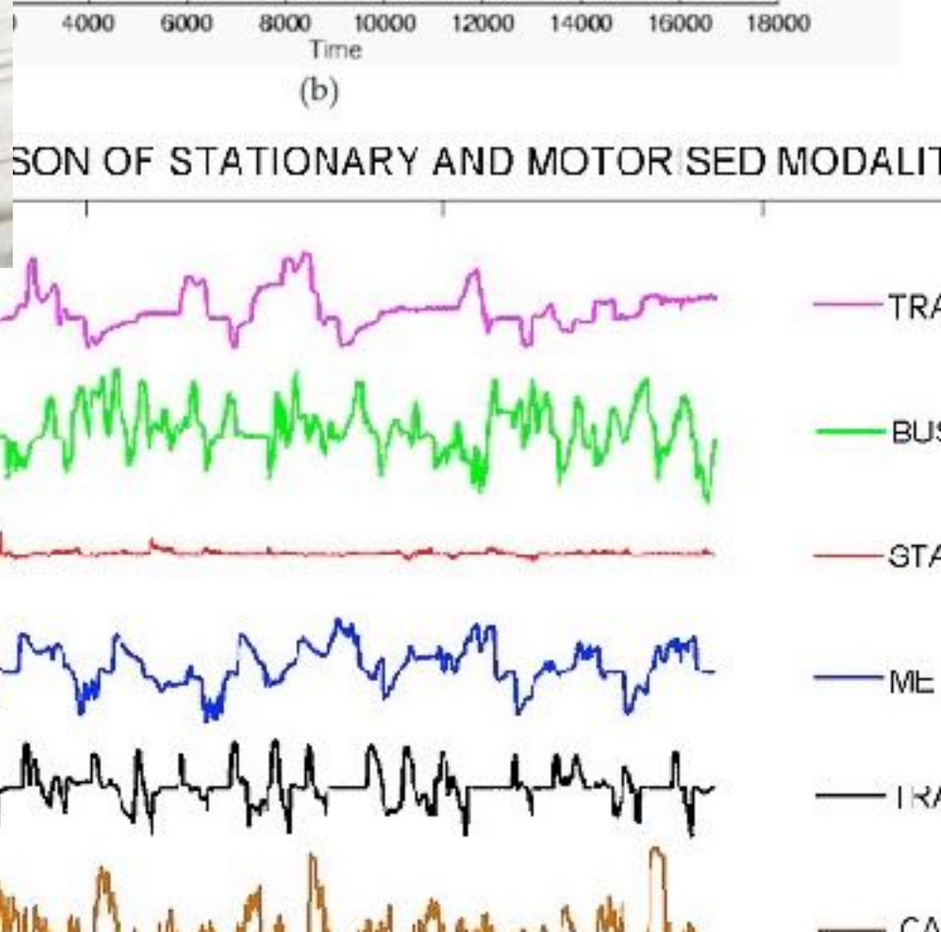
- Find the ‘clean’ speech (where a single person is speaking)
- The detection of multiple speakers should not depend on the speakers themselves (it should generalize over different movies)

Future: situation classification

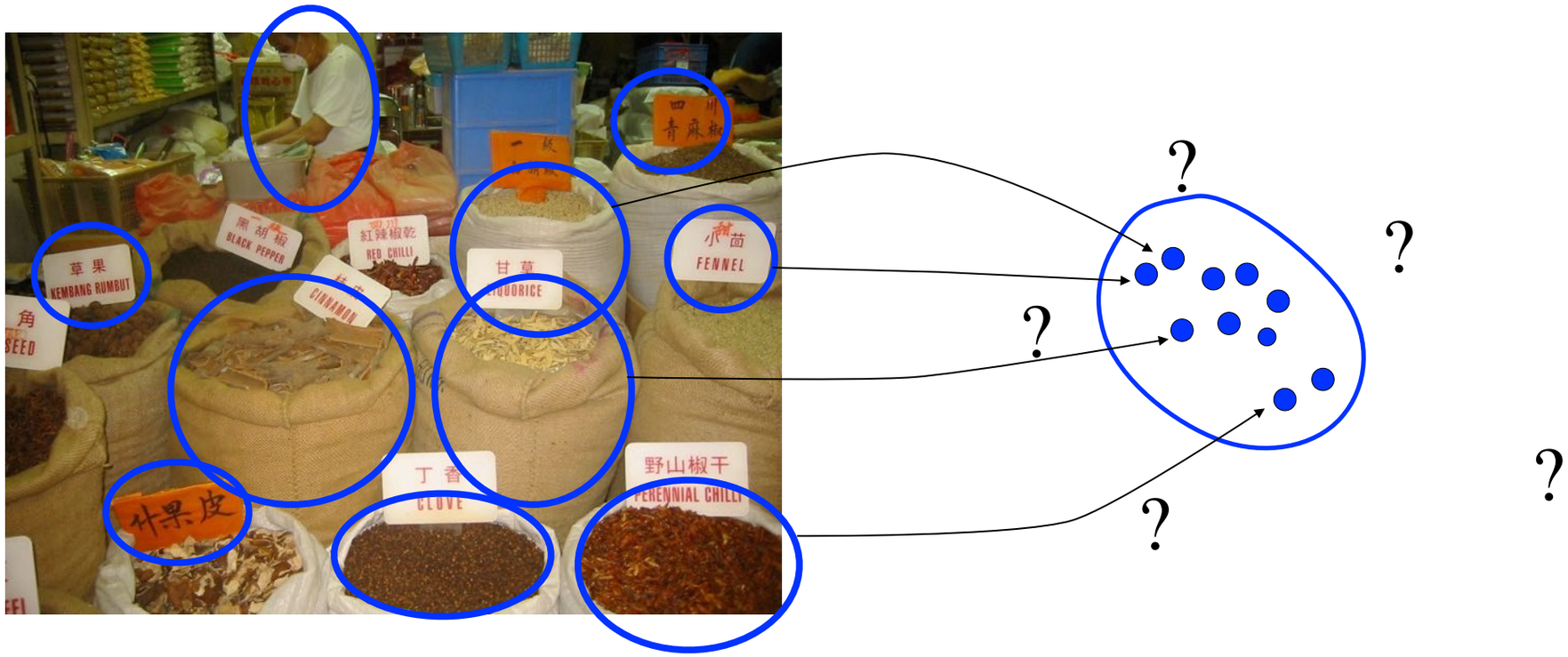
- What are you doing?
- Can we suggest something?

- How to represent ‘situations’?





Multiple-instance **learning??**



- In many situations, the ‘instances’ are not realisations of one concept
- Noisy, background instances appear
- Combinations of and relations between instances are not considered

Set classification

- In practice the multiple-instance learning has to be **generalised**
- Allow background instances to be ignored
- Allow several concepts (combined with AND/OR)
- Allow the modelling of relations between instances

Open (fundamental) questions…

- What is the sample size of a MIL problem?
  (what is the influence of the #instances in a bag, the fraction of 'truly' positive instances?)
- How to characterise MI problem (bags/dim)?
- What is the complexity of a MI problem? How to vary the complexity of a MI classifier?
- How to efficiently train a MI classifier? Is the idea of a 'concept' fruitful?
- How to include more structure in a MI classifier?
  (use more than one concept, use temporal structure, spatial structure, …)
- How to incorporate (partial) instance labels?
- How to optimise the features used in MIL? Deep learning!?
- Multi-class MIL?

Conclusions

- For many real-world classification problems complex/ compound objects have to be represented
- One possibility: represent the object by a **set** of feature vectors
- Many procedures have been proposed, not very clear when one will work well:
  - Diverse-Density is very good, but very slow
  - Straightforward methods sometimes surprisingly good
  - Represent bags with feature vectors is good
  - Current recommendation: MILES

Questions?

- …?