Rensselaer

Fonds de recherche sur la nature et les technologies Québec

Multiple Instance Ranking

http://reccr.chem.rpi.edu/MIRank

Charles Bergeron
Kristin P. Bennett
Mathematical Sciences

Jed Zaretzki
Curt Breneman
Chemistry

Rensselaer Polytechnic Institute
Troy, New York, United States of America

© *Charles Bergeron* et al.*, 2008. All rights reserved.*

Summary

- This presentation introduces a framework that tackles a novel machine learning question arising from an investigation into an important chemistry problem.
- Multiple Instance Ranking (MIRank) is defined and formulated.
- A first working algorithm produces excellent results on several real and synthetic problems.

© *Charles Bergeron* et al.*, 2008. All rights reserved.*

Outline

- Motivation
  - from a Chemistry perspective
  - from a Machine Learning perspective
- Formulation
  - definition of Multiple Instance Ranking
  - algorithm as a bilinear optimization problem
- Experiments
  - datasets and experimental design
  - results and conclusions
- Outlook
  - from a Chemistry perspective
  - from a Machine Learning perspective

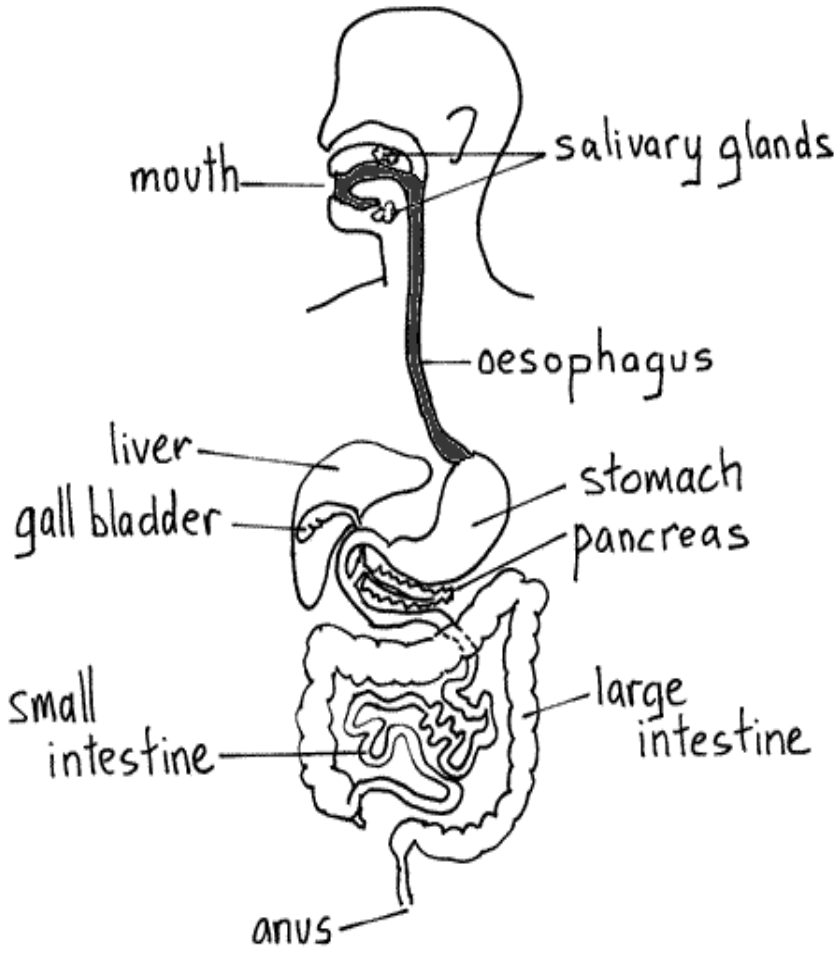
© Charles Bergeron* et al.*, 2008. All rights reserved.

Outline

- Motivation
  - from a Chemistry perspective
  - from a Machine Learning perspective
- Formulation
  - definition of Multiple Instance Ranking
  - algorithm as a bilinear optimization problem
- Experiments
  - datasets and experimental design
  - results and conclusions
- Outlook
  - from a Chemistry perspective
  - from a Machine Learning perspective

© *Charles Bergeron* et al.*, 2008. All rights reserved.*

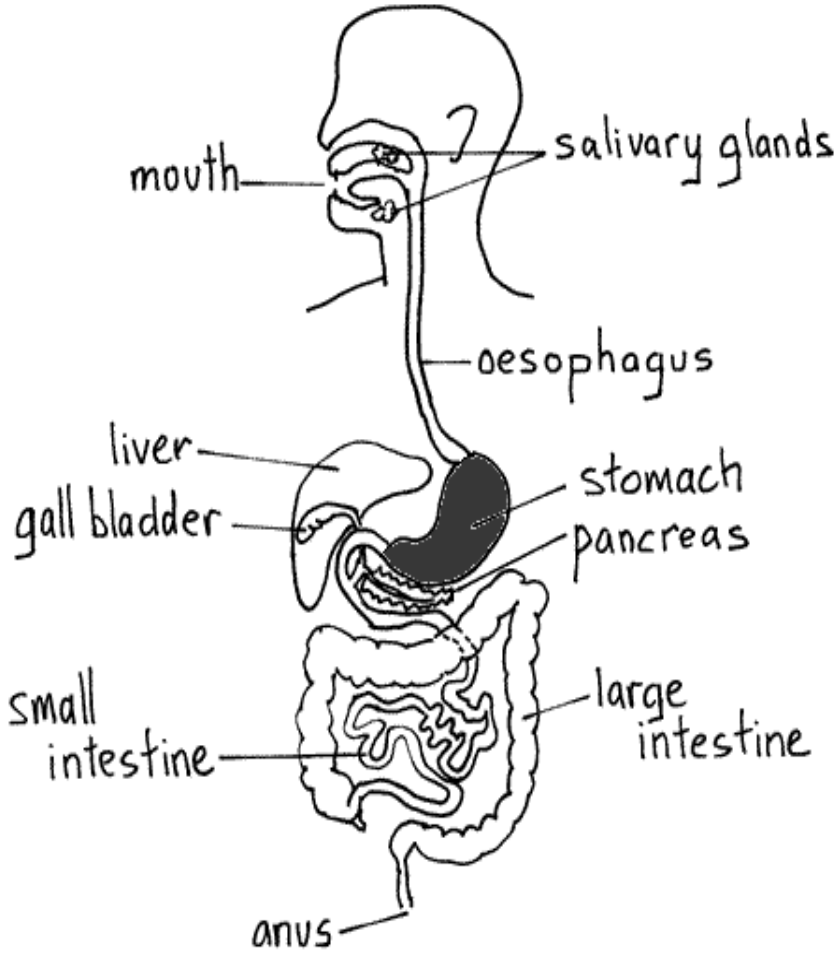
Oesophagus



- Bioavailability (the ability of a drug administered orally to reach the bloodstream) is an important consideration to the pharmaceutical industry.

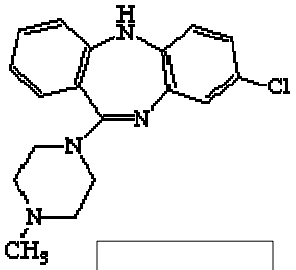
© Charles Bergeron* et al.*, 2008. All rights reserved.

Stomach



- Pill is broken down into basic components.

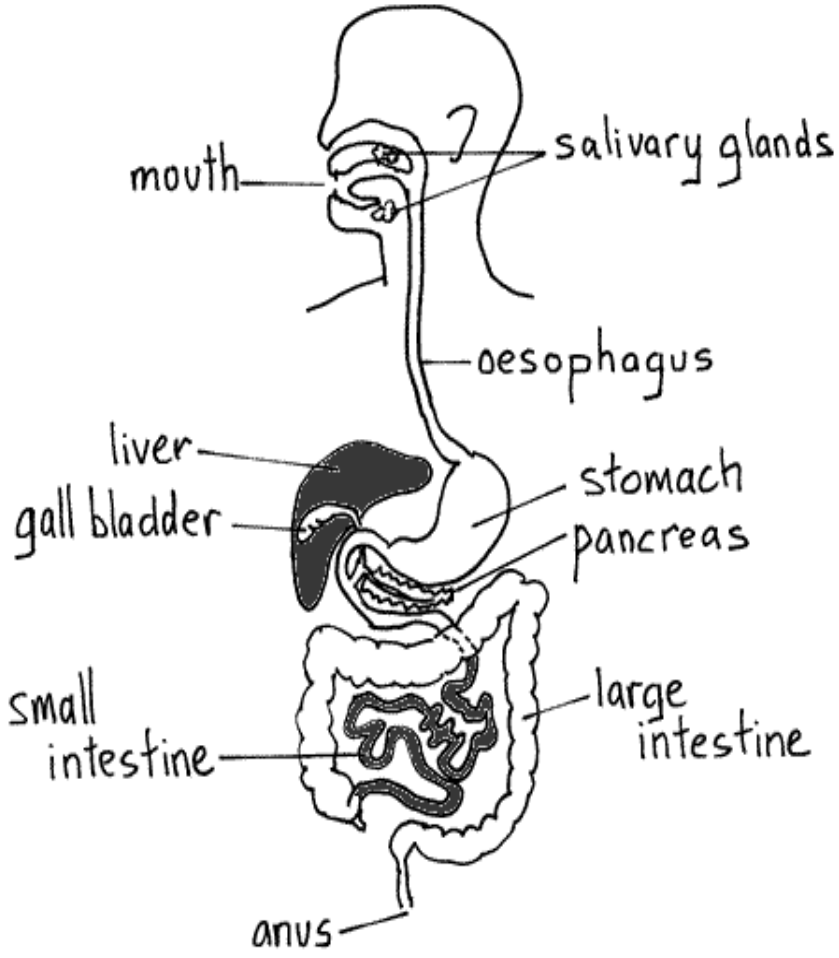
Clozaril pill



Clozapine molecule

© Charles Bergeron* et al.*, 2008. All rights reserved.

Liver & small intestine



- Successful drug compounds cross the hepatic and intestinal lining and make it to the bloodstream without being degraded, so that their medicinal effect may be felt.
- The rate limiting step in the metabolism of drugs by enzyme cytochrome CYP3A4 is hydrogen atom abstraction (removal).

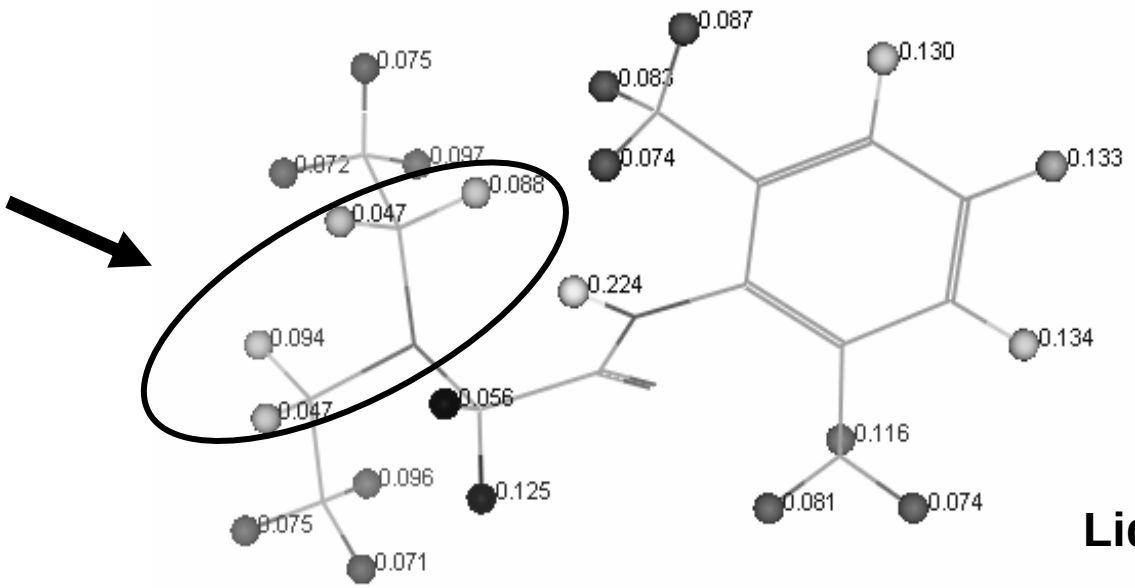
© Charles Bergeron* et al.*, 2008. All rights reserved.

Chemistry research goal

- To better understand the process of drug metabolism.
- To obtain knowledge about the site of metabolism where drug molecules undergo hydrogen abstraction under the effect of cytochrome CYP3A4.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Available data

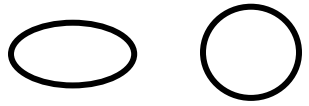


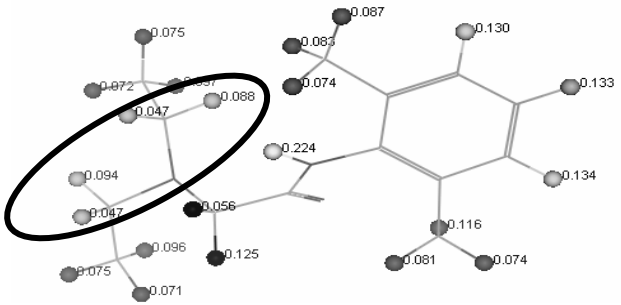
Lidocaine

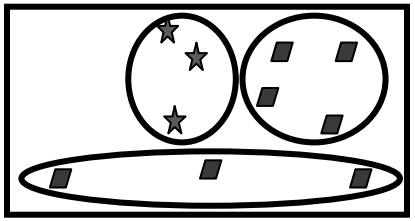
- Database consists of small drug-like molecules (stick diagram).
- Features are computed for each hydrogen atom (small spheres).
- Hydrogens are grouped into sites of metabolism (different colors).
- For each molecule, the preferred site of metabolism is known.
- It is not known which hydrogen actually gets abstracted.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Chemistry to machine learning

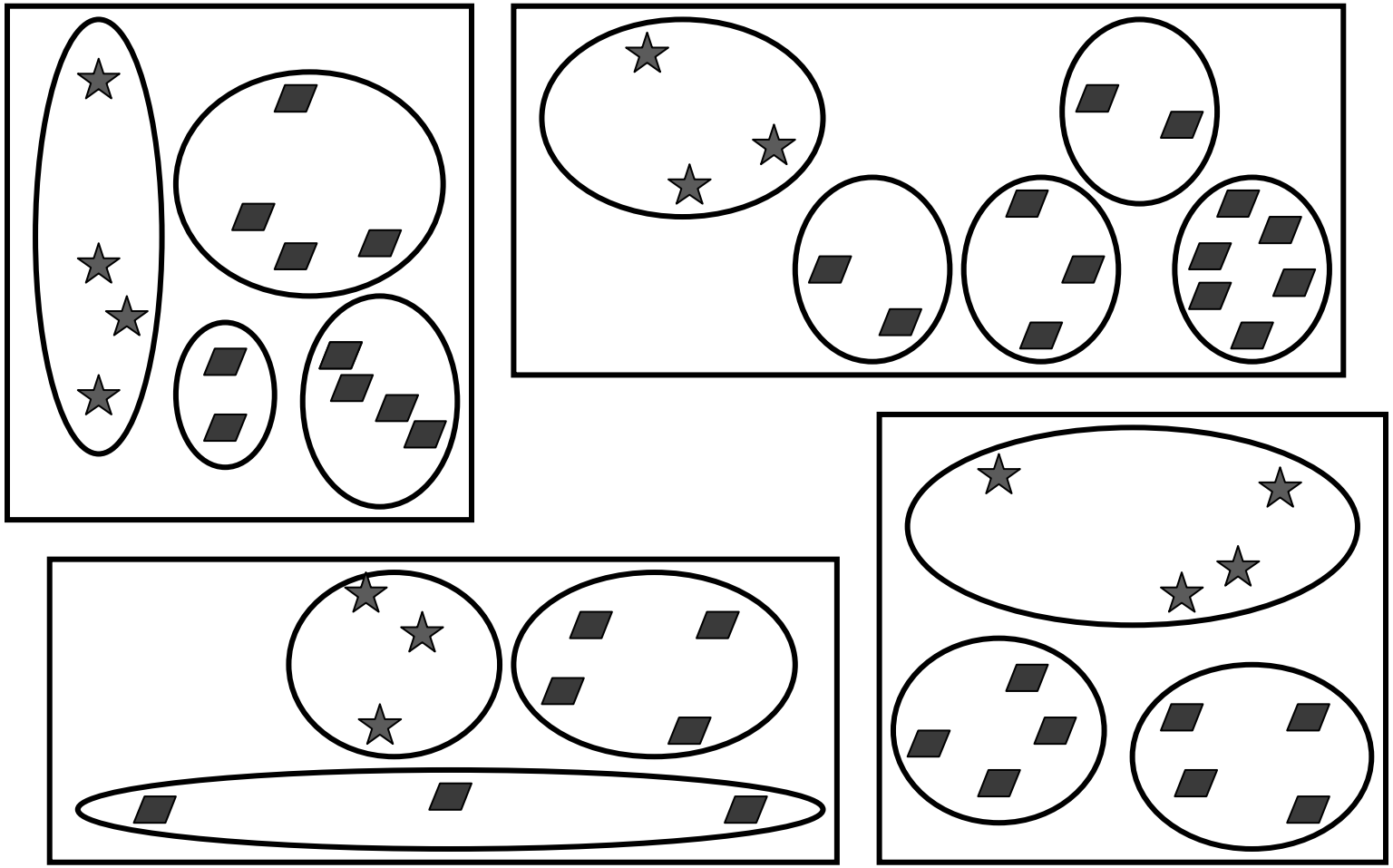
	Computational chemistry	**Machine learning**
Problem statement	For each molecule, find the site of metabolism	For each box, find the preferred bag
Top-level	Molecules	Boxes 
Middle-level	Sites of metabolism	Bags 
Bottom-level	Hydrogen atoms	Items 





© Charles Bergeron* et al.*, 2008. All rights reserved.

Data structure



For each box (red rectangle), predict the preferred bag (green ellipse). Items are represented as parallelograms and stars, and the other bags are blue ellipses.

© Charles Bergeron* et al.*, 2008. All rights reserved.

The need for MIRank

Dataset particularities	Machine learning consequences
• Descriptors are known for each item. • Each box has exactly one preferred bag. • An ambiguity exists as to which item in that bag determines the preference.	This is a multiple instance problem.
• It is not known how other bags within a box rank with respect to each other. • It is not known how bags compare against each other across boxes.	This is a partial ranking problem within each box.
• Boxes may be very different from each other.	This is a hard problem.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Machine learning research goal

- To define the Multiple Instance Ranking (MIRank) setting.
- To develop a first working algorithm to solve MIRank problems.
- To prove the concept of MIRank on the CYP3A4 substrate and other real and synthetic datasets.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Outline

- Motivation
  - from a Chemistry perspective
  - from a Machine Learning perspective
- Formulation
  - definition of Multiple Instance Ranking
  - algorithm as a bilinear optimization problem
- Experiments
  - datasets and experimental design
  - results and conclusions
- Outlook
  - from a Chemistry perspective
  - from a Machine Learning perspective

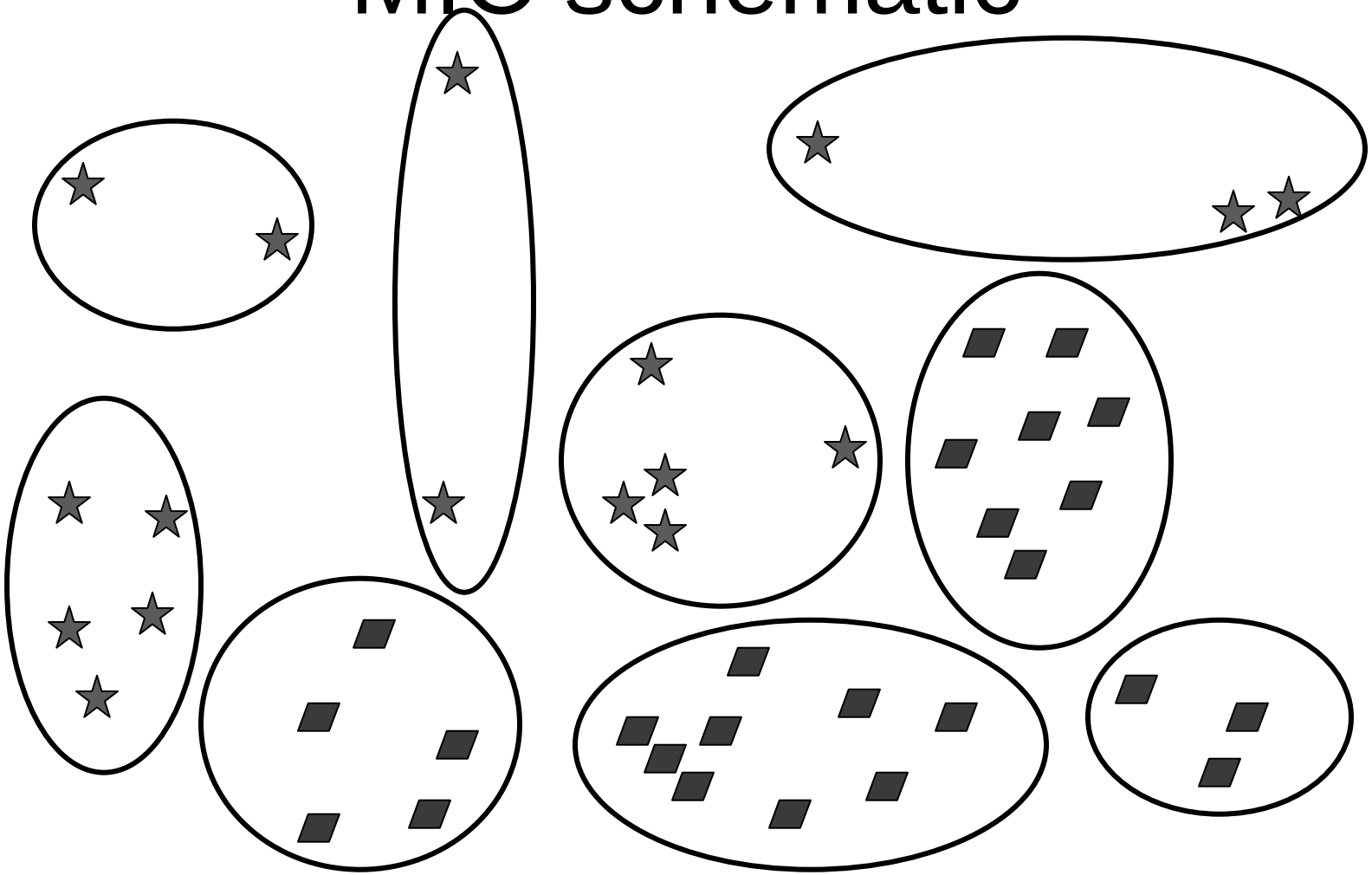
© *Charles Bergeron* et al.*, 2008. All rights reserved.*

Starting point: MIC

- Multiple Instance Classification (MIC) classifies bags based on their item descriptors.
- An active bag contains at least one active item.
- An inactive bag contains exclusively inactive items.
- There exists the ambiguity of what item in the active bag is active.

© Charles Bergeron* et al.*, 2008. All rights reserved.

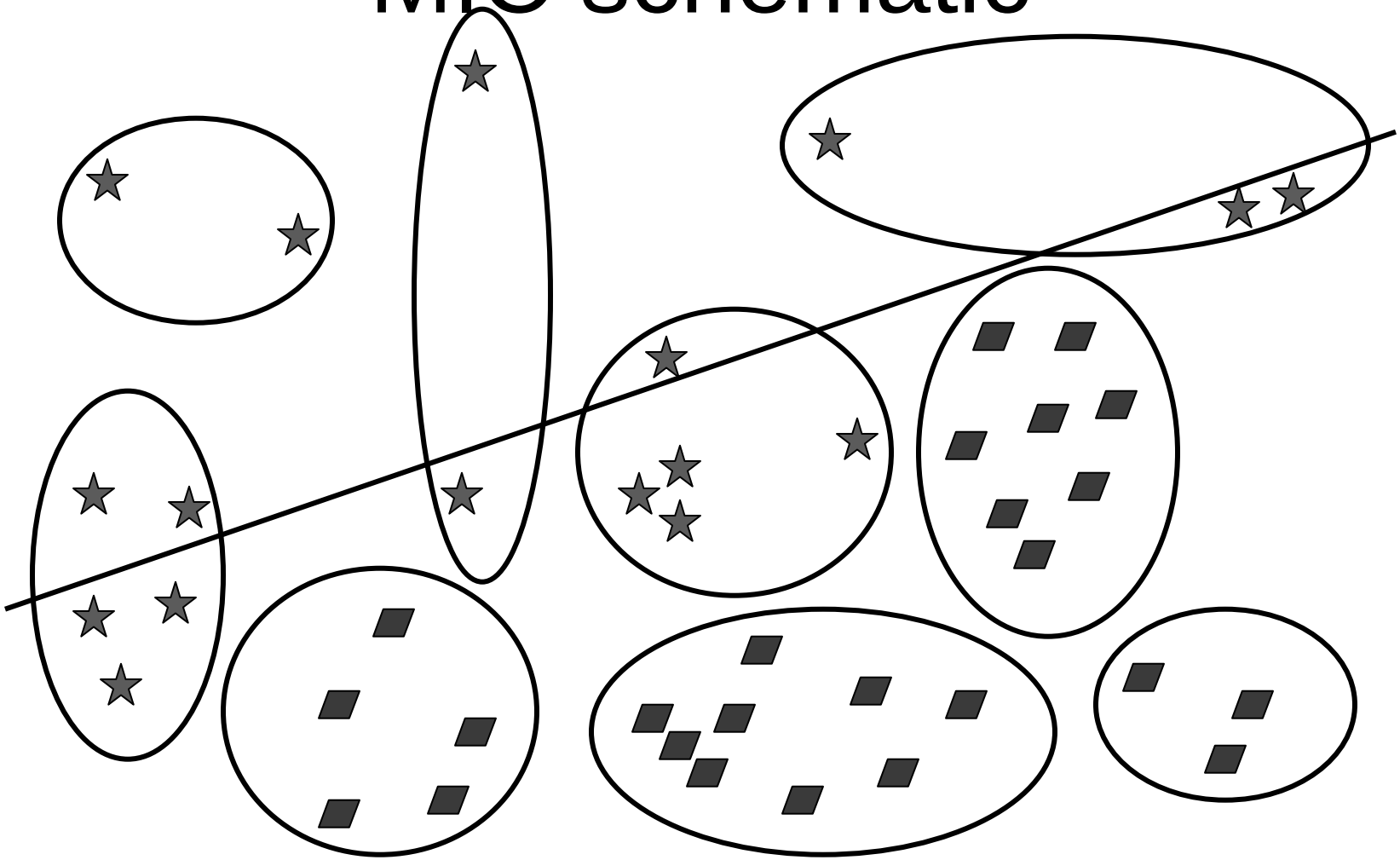
MIC schematic



Items (stars and parallelograms) in active bags (green ellipses) and inactive bags (blue ellipses) are shown.

© Charles Bergeron* et al.*, 2008. All rights reserved.

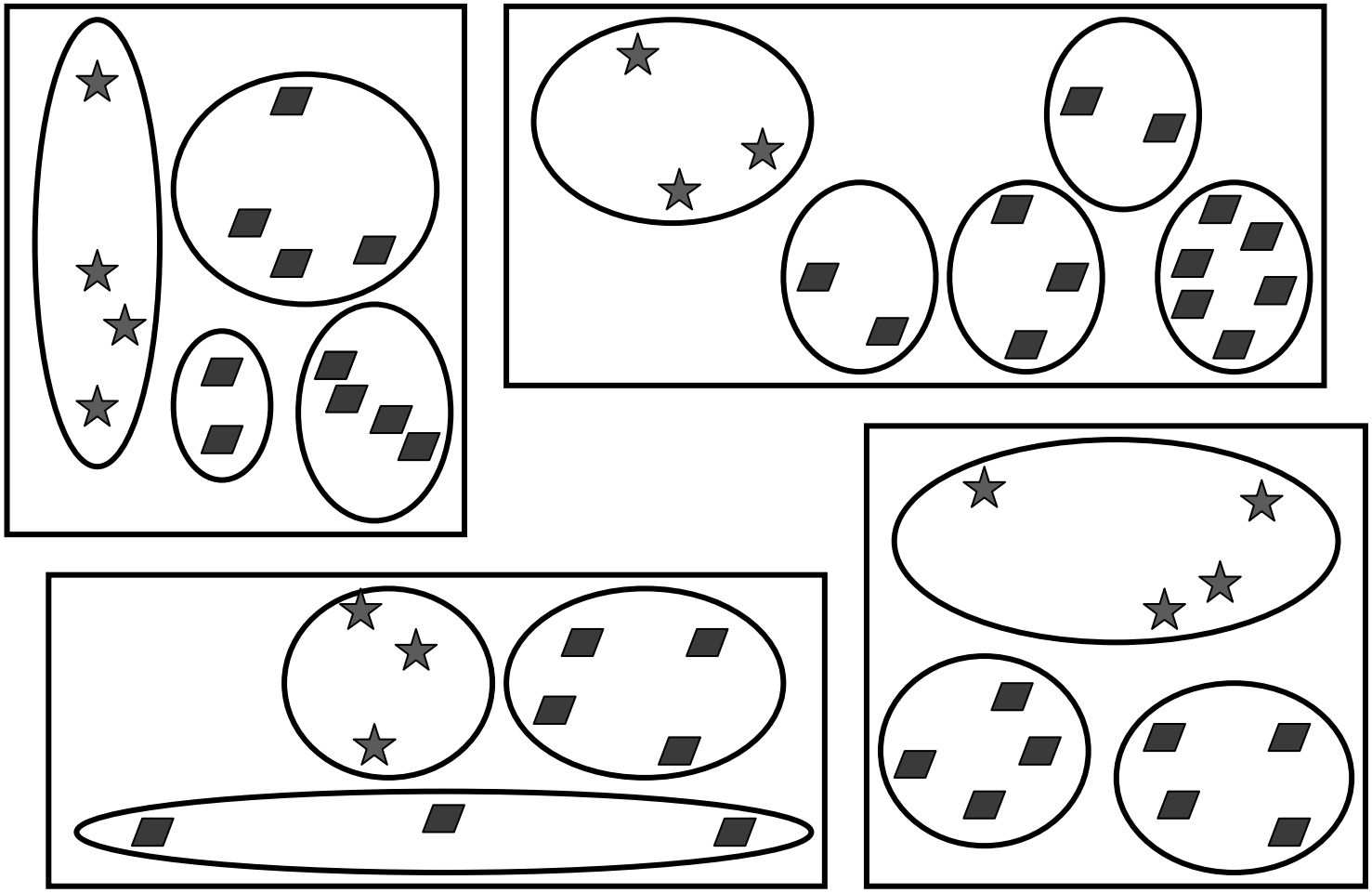
MIC schematic



At least one item (star) in each active bag (green ellipse) is above the decision curve (orange line).

© Charles Bergeron* et al.*, 2008. All rights reserved.

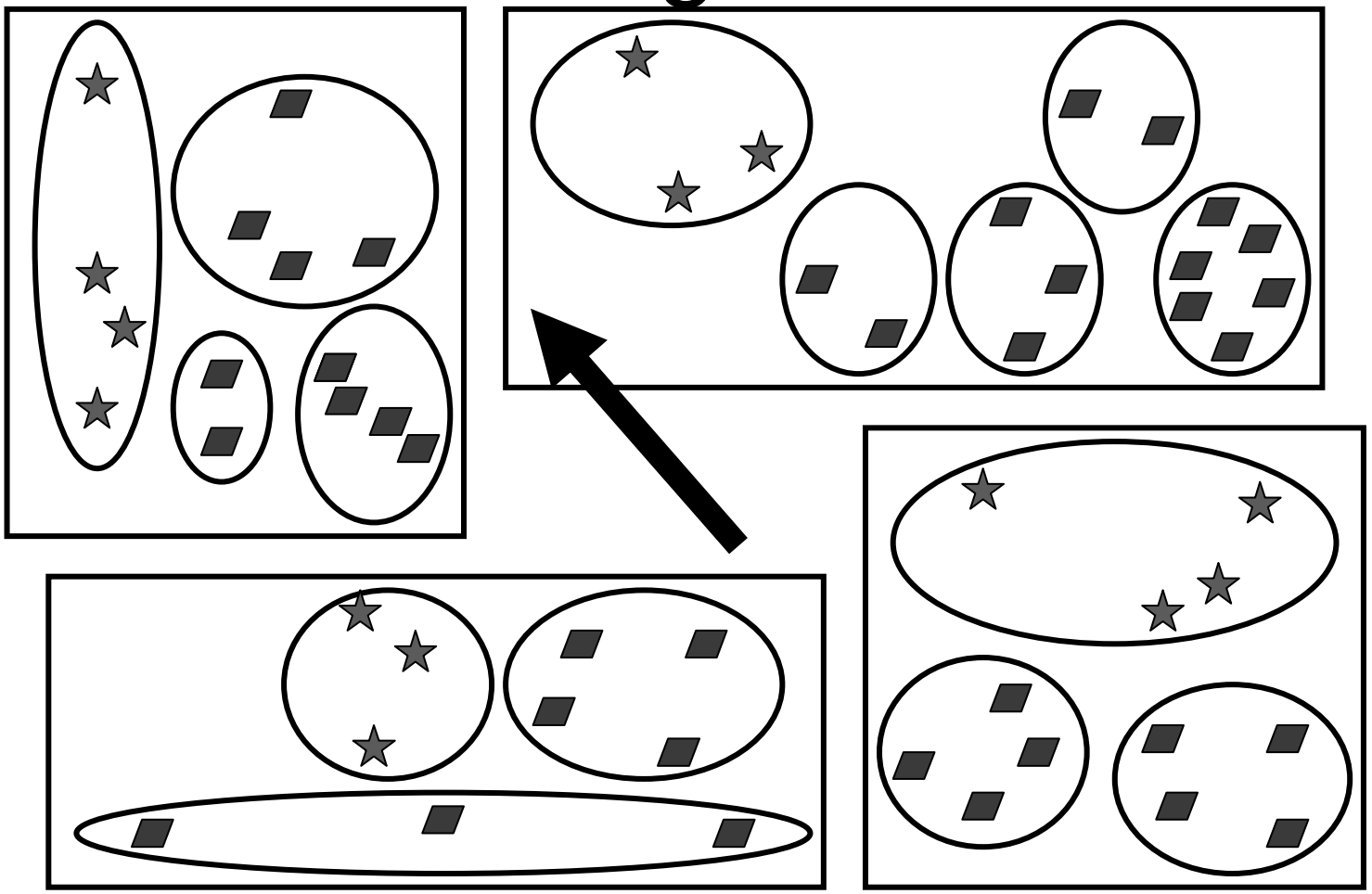
MIRank Schematic



For each box (red rectangle), predict the preferred bag (green ellipse). Items are represented as parallelograms and stars, and the other bags are blue ellipses.

© Charles Bergeron* et al.*, 2008. All rights reserved.

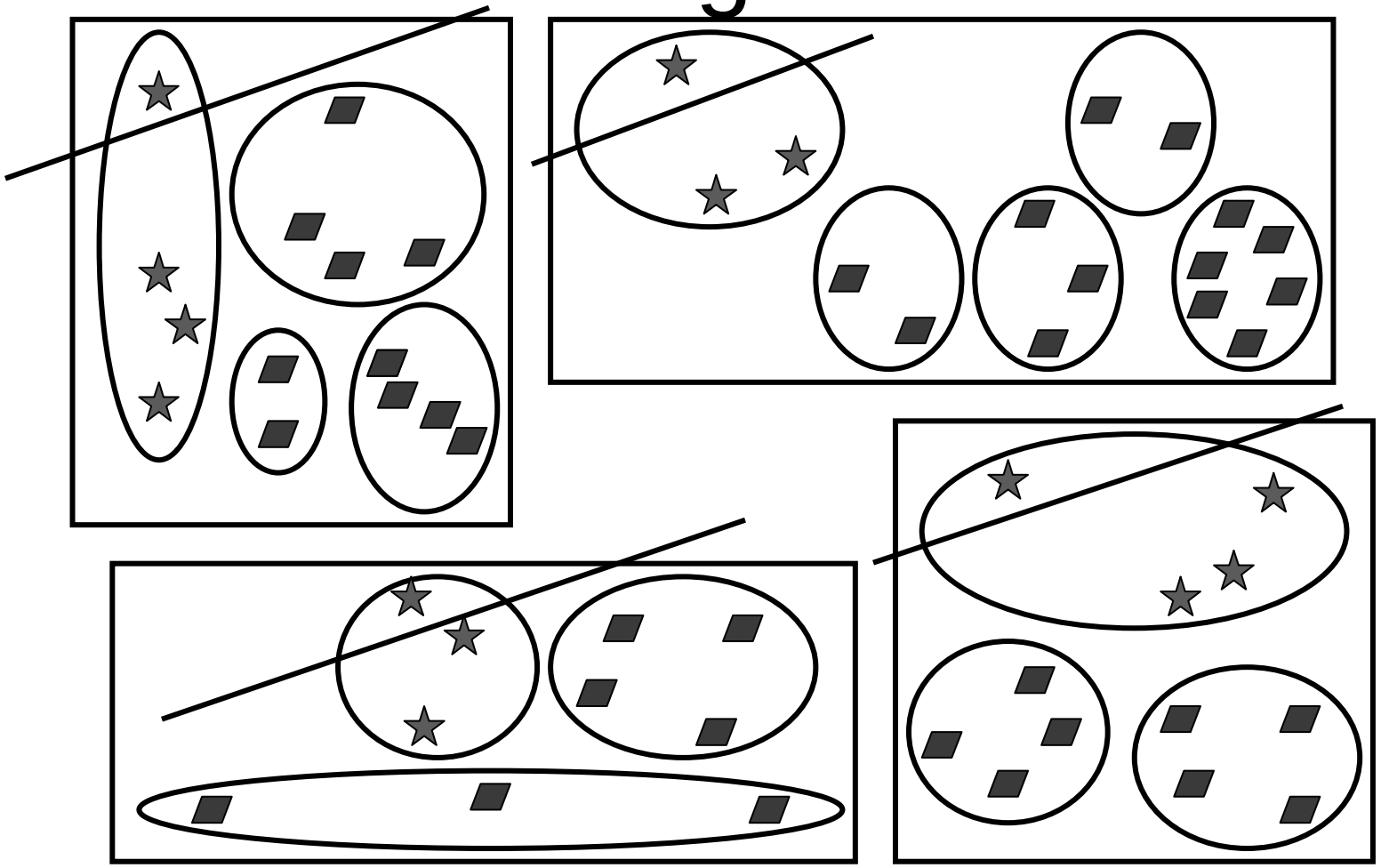
Ranking function



For each box (red rectangle), the ranking function (orange arrow) ranks highest the preferred bag (green ellipse).

© Charles Bergeron* et al.*, 2008. All rights reserved.

Ranking function



For each box (red rectangle), the ranking function (sliding orange line) ranks highest the preferred bag (green ellipse).

© Charles Bergeron* et al.*, 2008. All rights reserved.

Mathematical formulation

$$\max_{i \in I} f(\boldsymbol{x}_i) > \max_{j \in J} f(\boldsymbol{x}_j)$$

- Bag I is greater than bag J

$$\max_{i \in I} f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j) \quad \forall \quad j \in J$$

- RHS max replaced with all items

$$f(X_I^T \boldsymbol{v_{I,J}}) > f(\boldsymbol{x}_j) \quad \forall \quad j \in J$$

- LHS max replaced with convex combination of all items in that bag

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$$

- Assume a linear model

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} > \boldsymbol{x}_j^T \boldsymbol{w}$$

- Apply model

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} - \boldsymbol{x}_j^T \boldsymbol{w} \geq 1 - \xi_{I,j}$$

- Allow for error by introducing empirical risk and margin

© Charles Bergeron* et al.*, 2008. All rights reserved.

Bilinear optimization problem

Tradeoff parameter

Empirical risk

$$\min_{\boldsymbol{\xi},\boldsymbol{w},\boldsymbol{v_{I,J}}} \quad \nu \boldsymbol{e}^T \boldsymbol{\xi} + \|\boldsymbol{w}\|_1$$

Regularization

Model

subject to

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} - \boldsymbol{x}_j^T \boldsymbol{w} \geq 1 - \xi_{I,j} \quad \forall \quad (I, J, j)$$

Bilinear constraint.

$$\boldsymbol{e}^T \boldsymbol{v_{I,J}} = 1 \quad \forall \quad (I, J)$$

Convex combination weights sum to one.

$$\boldsymbol{v_{I,J}} \geq \boldsymbol{0} \quad \forall \quad (I, J)$$

Convex combination weights are nonnegative.

$$\boldsymbol{\xi} \geq \boldsymbol{0}.$$

Empirical risk terms are nonnegative.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Modeling choices

- A linear model is chosen, because chemists are interested in easily interpretable models.
- The 1-norm is used to regularize, resulting in sparse models, again to facilitate chemical interpretation.
- The 1-norm is also used to penalize empirical risk, as is done in SVM. Also, this keeps the objective linear.
- The Mangasarian & Wild (2008) strategy of using the convex combination as a slick way of handling the uncertainty as to which item determines the preference is chosen.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Previous choices

- Several formulations and algorithms: diverse density (Maron & Ratan, 1998), EM-DD (Zhang & Goldman, 2001), neural networks (Ramon & Raedt, 2000)
- Support vector machine generalizations exist using integer variables (Andrews et al., 2003) and convex combinations (Mangasarian and Wild, 2008) to handle the ambiguity

© Charles Bergeron* et al.*, 2008. All rights reserved.

Algorithm outline

- Solve the linear program for $\boldsymbol{\xi}$ and $\boldsymbol{w}$ while keeping the $\boldsymbol{v}$'s fixed.
- Solve the linear program for $\boldsymbol{\xi}$ and the $\boldsymbol{v}$'s while keeping $\boldsymbol{w}$ fixed.
- Repeat.

$$\min_{\boldsymbol{\xi},\boldsymbol{w},\boldsymbol{v_{I,J}}} \quad \nu \boldsymbol{e}^T \boldsymbol{\xi} + \|\boldsymbol{w}\|_1$$

$$\text{subject to}$$

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} - \boldsymbol{x}_j^T \boldsymbol{w} \geq 1 - \xi_{I,j} \ \forall \ (I, J, j)$$

$$\boldsymbol{e}^T \boldsymbol{v_{I,J}} = 1 \ \forall \ (I, J)$$

$$\boldsymbol{v_{I,J}} \geq \boldsymbol{0} \ \forall \ (I, J)$$

$$\boldsymbol{\xi} \geq \boldsymbol{0}.$$

© Charles Bergeron* et al.*, 2008. All rights reserved.

Outline

- Motivation
  - from a Chemistry perspective
  - from a Machine Learning perspective
- Formulation
  - definition of Multiple Instance Ranking
  - algorithm as a bilinear optimization problem
- Experiments
  - datasets and experimental design
  - results and conclusions
- Outlook
  - from a Chemistry perspective
  - from a Machine Learning perspective

© *Charles Bergeron* et al.*, 2008. All rights reserved.*

CYP3A4 substrate dataset

- 227 molecules (boxes)
- 2272 sites of metabolism (bags)
- 4847 hydrogen atoms (items)
- 32 features per item
  - charge
  - hydrogen surface area
  - nonhydrogen surface area
  - hydrophobic moment
  - span
  - topological neighborhood properties
- Output information consists of one preferred bag per box

© Charles Bergeron* et al.*, 2008. All rights reserved.

Experimental design

- Randomly split the dataset into training, validation and testing subsets consisting of 60%, 20% and 20% of boxes, respectively.
- Train MIC and MIRank on the training subset for 19 values of the tradeoff parameter *v*.
- The validation subset is used to select the best *v*.
- Results are recorded over the testing subset.
- This process is repeated 32 times.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Results

Dataset	MIC	MIRank	p-value
CYP3A4 substrate	67.2% ± 6.6	70.8% ± 6.4	$4.10 \cdot 10^{-3}$

- MIC algorithm is that of Mangasarian and Wild (2008).
- Metric is the percentage accuracy in predicting the preferred bag for each box, allowing for 2 guesses per box.
- Results are presented as a mean and std over the 32 runs.
- MIRank statistically outperforms MIC at a 5% significance level.
- MIRank improves the model by over 5%.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Further results

Dataset	MIC	MIRank	p-value
Synthetic-1	90.8 % ± 8.6	99.8 % ± 0.53	$1.62 \cdot 10^{-6}$
Synthetic-2	96.8 % ± 4.6	99.1 % ± 1.8	$1.31 \cdot 10^{-2}$
Synthetic-3	95.5 % ± 8.3	99.9 % ± 0.38	$5.84 \cdot 10^{-3}$
Synthetic-4	95.7 % ± 5.2	99.7 % ± 0.91	$1.46 \cdot 10^{-4}$
Census-16h	52.8 % ± 17.4	60.3 % ± 15.1	$4.51 \cdot 10^{-2}$
Census-16l	46.2 % ± 17.7	57.5 % ± 16.0	$3.92 \cdot 10^{-4}$

- For all datasets, MIRank statistically outperforms MIRank.
- For the census datasets, MIRank improves the model by 14-24%.
- The paper goes into greater depth about these datasets.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Discussion

- Problems fitting into the Multiple Instance Ranking paradigm are better solved using MIRank models than other methods.
- Forcing MIRank problems into a MIC paradigm is not successful.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Software and data

http://reccr.chem.rpi.edu/MIRank

- We are making our CYP3A4 substrate data available.
- We are also making our MIRank algorithm Matlab source codes available, as well as our implementation of the MIC algorithm (Mangasarian & Wild, 2008).
- Look out for it online!
- Contact me for further information.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Outline

- Motivation
  - from a Chemistry perspective
  - from a Machine Learning perspective
- Formulation
  - definition of Multiple Instance Ranking
  - algorithm as a bilinear optimization problem
- Experiments
  - datasets and experimental design
  - results and conclusions
- Outlook
  - from a Chemistry perspective
  - from a Machine Learning perspective

© *Charles Bergeron* et al.*, 2008. All rights reserved.*

Chemistry extensions

- Chemical interpretation of results is a paper of its own.
- Increase the number of molecules in CYP3A4 database.
- Build databases and models for new substrates, such as CYP2D6 and CYP2C9.
- Develop novel descriptors believed to be indicative of hydrogen abstraction.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Machine learning extensions

- Implementation of nonlinear model using kernels (already formulated in the paper) to compare results with linear model.
- Compare model choice of 1-norm with other possibilities for empirical risk and regularization terms.
- Adapt recent large scale SVM algorithms to make MIRank more scalable and efficient at finding local minima.
- Use integer programming or cutting plane algorithms to find global minima (at much greater computational cost).

© Charles Bergeron* et al.*, 2008. All rights reserved.

Future applications

- For each country, predict the city that contains the most profitable coffee shop.
- For a given state/province/*länder*/*département*, predict the electoral district that contains the most *effective* politician (the one that delivers the most subsidies to his constituents).
- For a given molecular class, predict the molecule that contains the conformation with the highest *efficacy* in inhibiting the human immunodeficiency virus (HIV).
- For each Olympic event, predict the nationality of the winning athlete.
- For each document, find the paragraph/passage that contains the most *relevant* phrase/sentence/word.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Summary

- This presentation introduces a framework that tackles a novel machine learning question arising from an important chemistry problem.
- Multiple Instance Ranking (MIRank) is defined and formulated.
- A first working algorithm produces excellent results on several real and synthetic problems.

© Charles Bergeron* et al.*, 2008. All rights reserved.

Rensselaer

Fonds de recherche sur la nature et les technologies Québec

Multiple Instance Ranking

http://reccr.chem.rpi.edu/MIRank

Charles Bergeron
Kristin P. Bennett
Mathematical Sciences

Jed Zaretzki
Curt Breneman
Chemistry

Rensselaer Polytechnic Institute
Troy, New York, United States of America

© *Charles Bergeron* et al., *2008. All rights reserved.*