

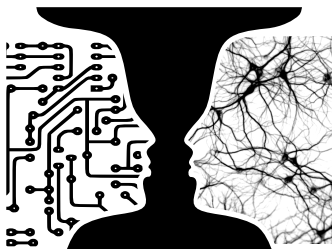
# Apprentissage de modèles parcimonieux à partir de génomes complets avec application à la résistance aux antibiotiques

Alexandre Drouin

Groupe de recherche en apprentissage automatique

Département d'informatique et de génie logiciel

Université Laval



Groupe de  
Recherche en  
Apprentissage  
Automatique de  
Laval



UNIVERSITÉ  
LAVAL

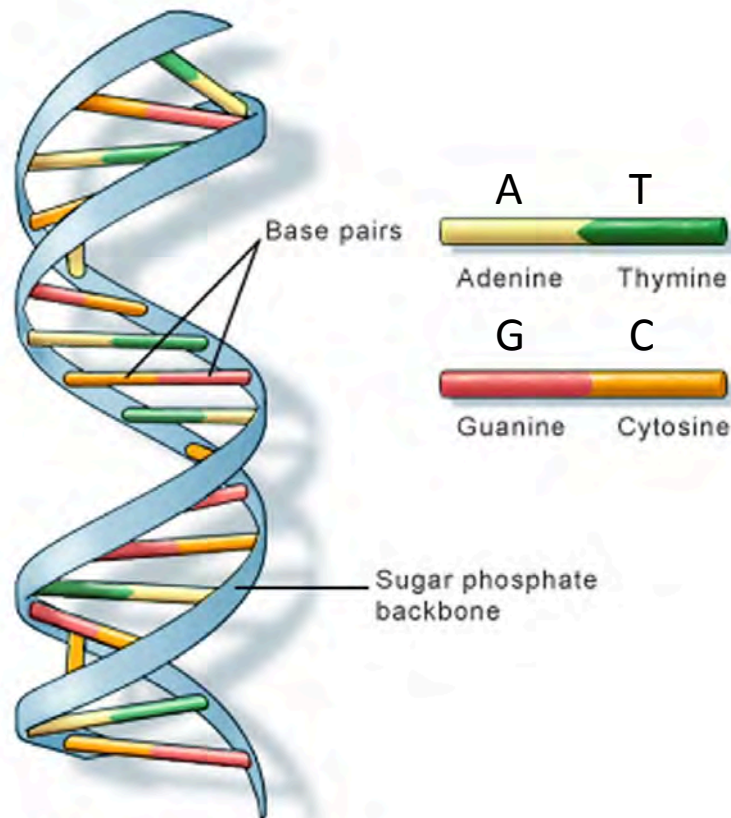
# Plan

- Introduction
  - Génomique
  - Apprentissage automatique
- Méthode
  - Représentation des données
  - Set Covering Machine
- Implémentation
- Résultats
  - Résistance aux antibiotiques
- Conclusion

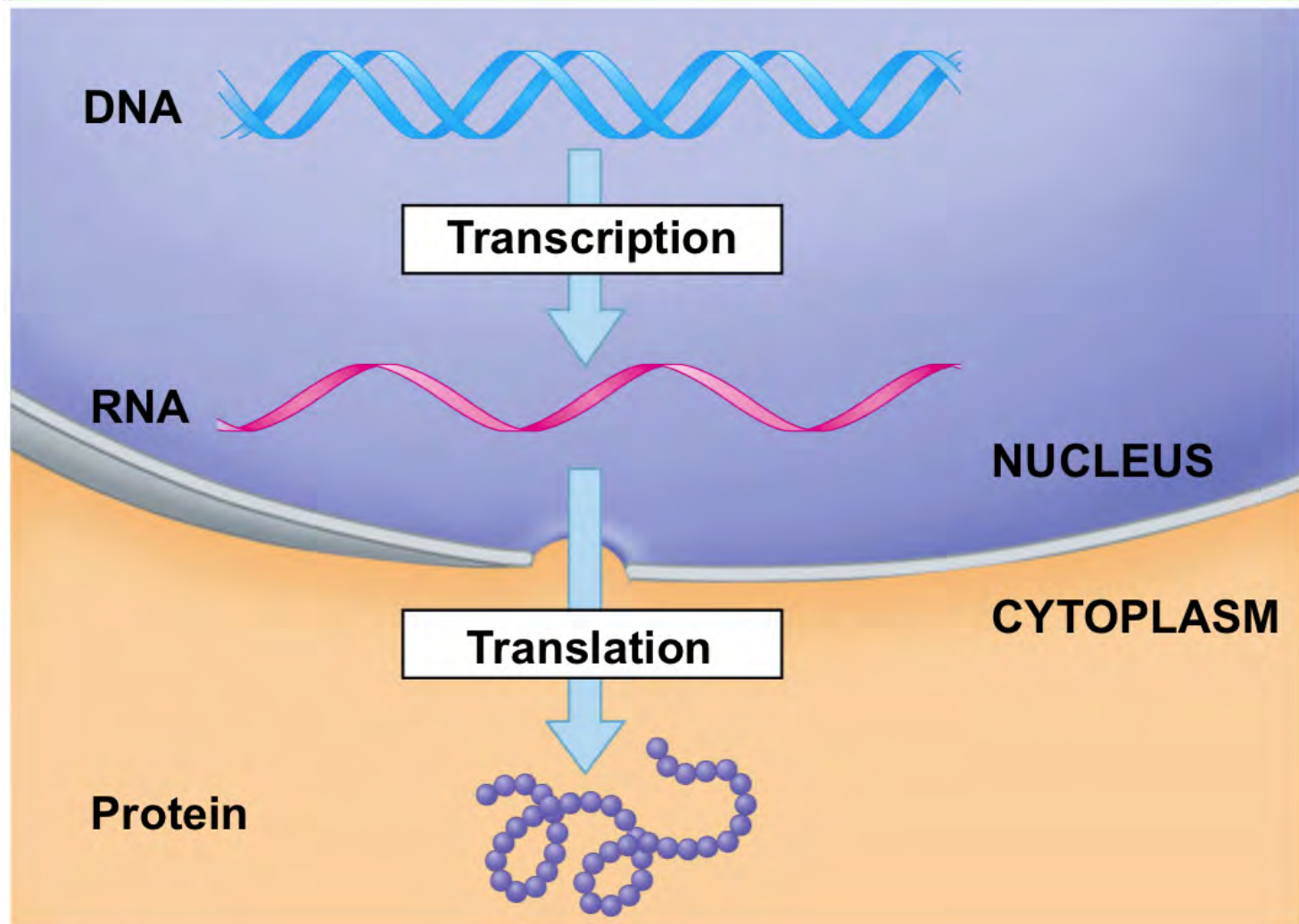
# Introduction

# Génomique

La génomique est un champ d'étude de la biologie portant sur l'étude de l'ensemble de l'ADN (génome).



# Biologie moléculaire

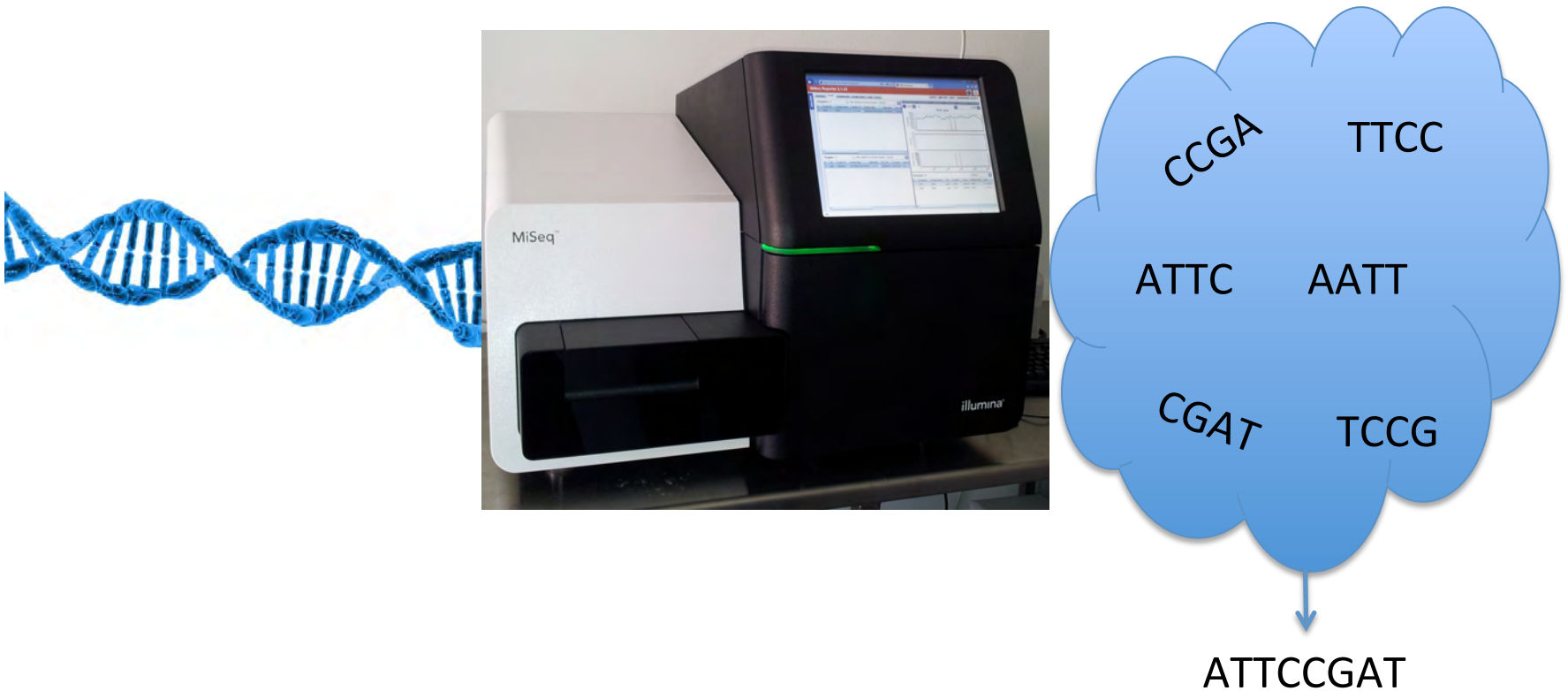


# Séquençage de l'ADN



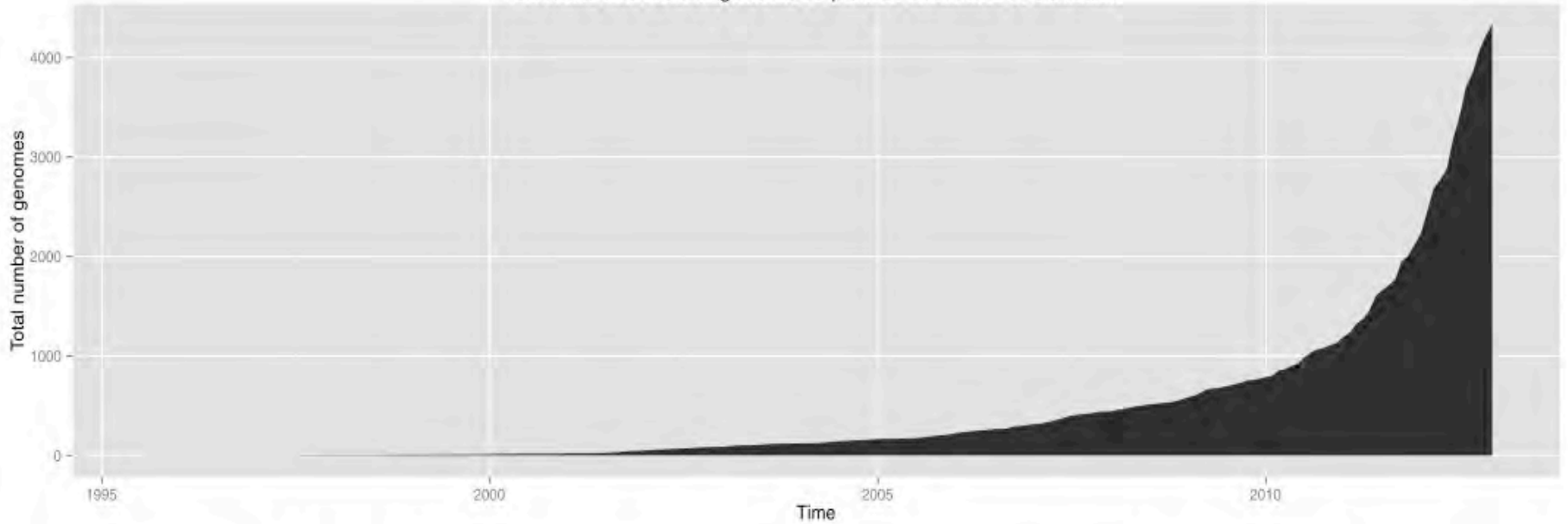
Le séquenceur produit un ensemble de courtes séquences représentant des fragments de la molécule d'ADN.

# Assemblage de l'ADN

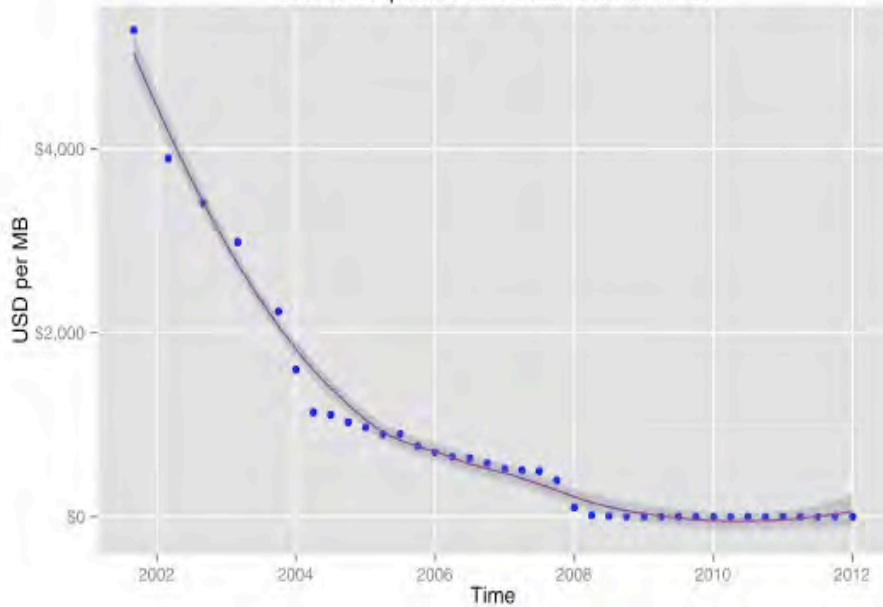


L'assembleur assemble les fragments pour former de longues séquences contigües.

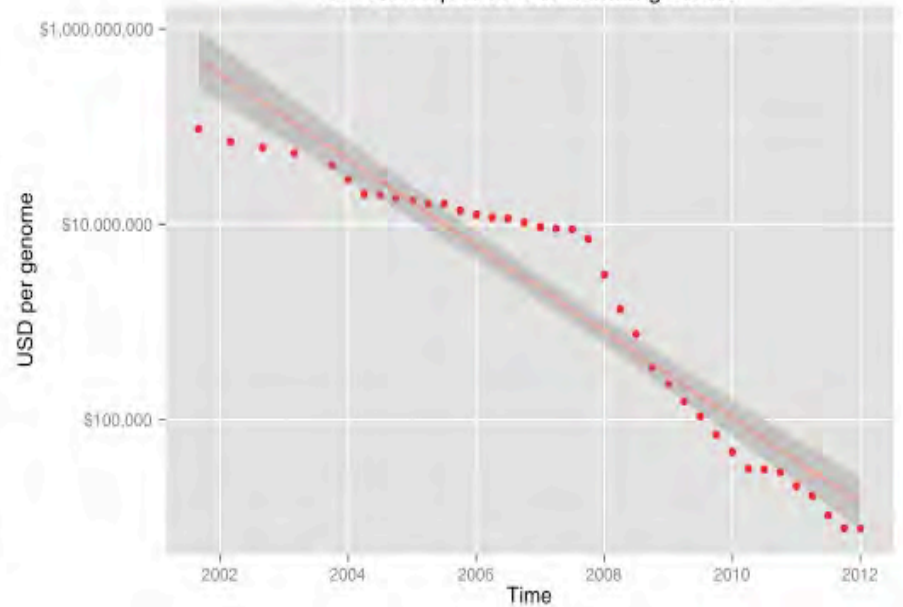
Bacterial and archeal genome sequences submitted to Genbank



Cost to sequence one million nucleotides

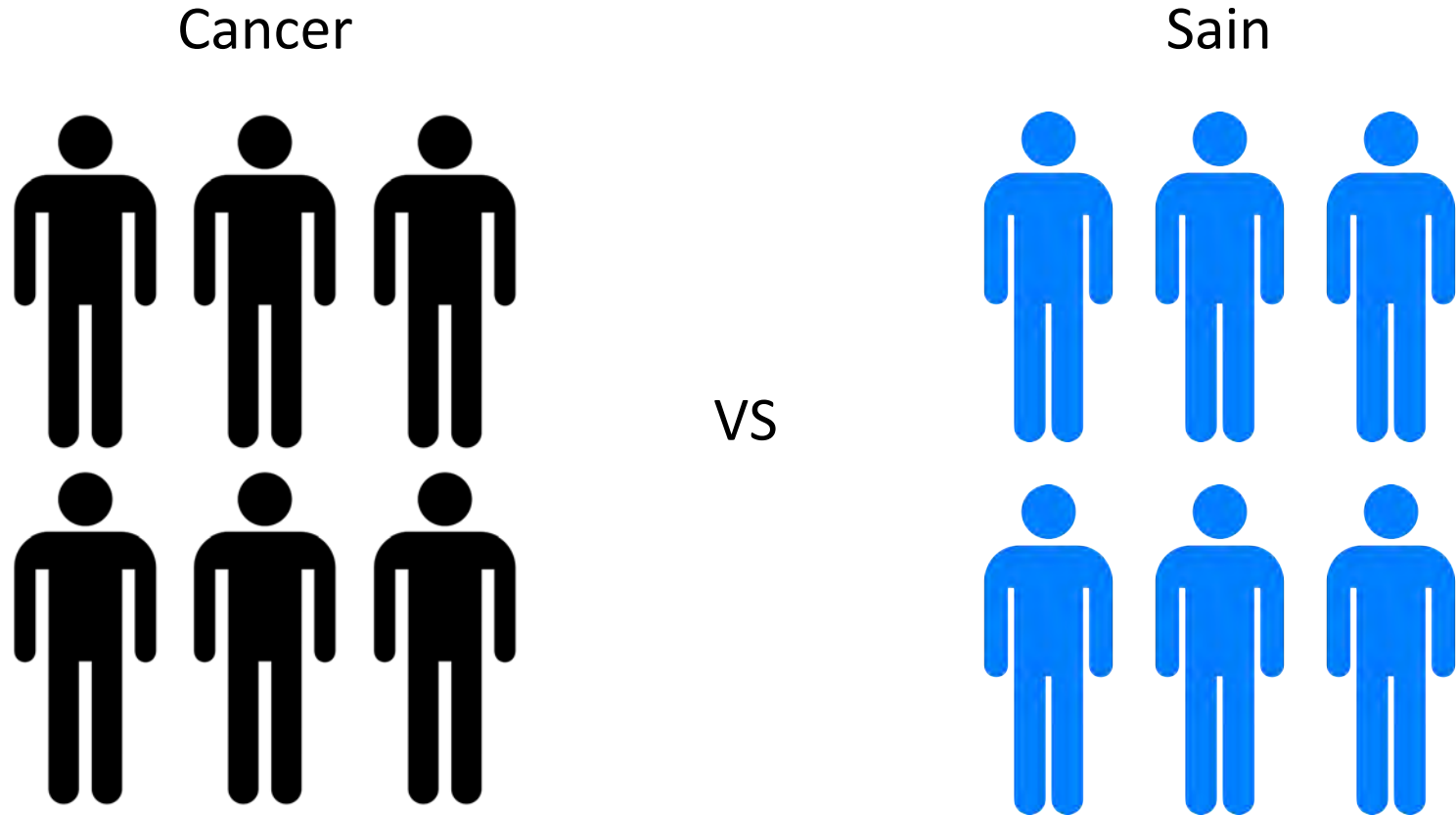


Cost to sequence one human genome





# Études cas-témoin



Les études cas-témoin comparent l'ADN de plusieurs individus, en vue de déterminer ce qui les distingue.

# Apprentissage automatique

*"Field of study that gives computers the ability to learn without being explicitly programmed"*

- Arthur Samuel, 1959

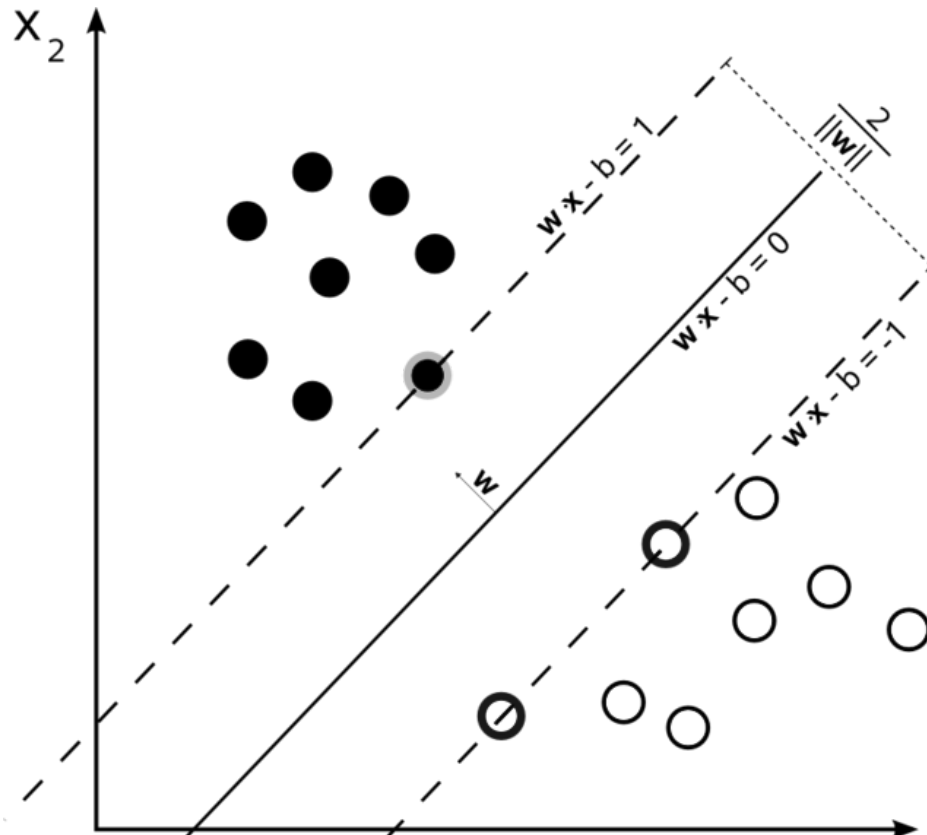
# Apprentissage Supervisé

Ensemble de données	$\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim D$ avec $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
Espace d'entrée $\mathcal{X}$	L'ensemble de tous les génomes
Espace de sortie $\mathcal{Y}$	$\mathcal{Y} = \{0, 1\}$
Hypothèse (modèle)	$h = \mathcal{A}(\mathcal{S}), h : \mathcal{X} \rightarrow \mathcal{Y}$
Risque	$\mathbf{E}_{(\mathbf{x}, y) \sim D} I[h(\mathbf{x}) \neq y]$

# Interprétable

- Les modèles doivent pouvoir être interprétés par des experts du domaine
- Parcimonieux
- Structure du modèle

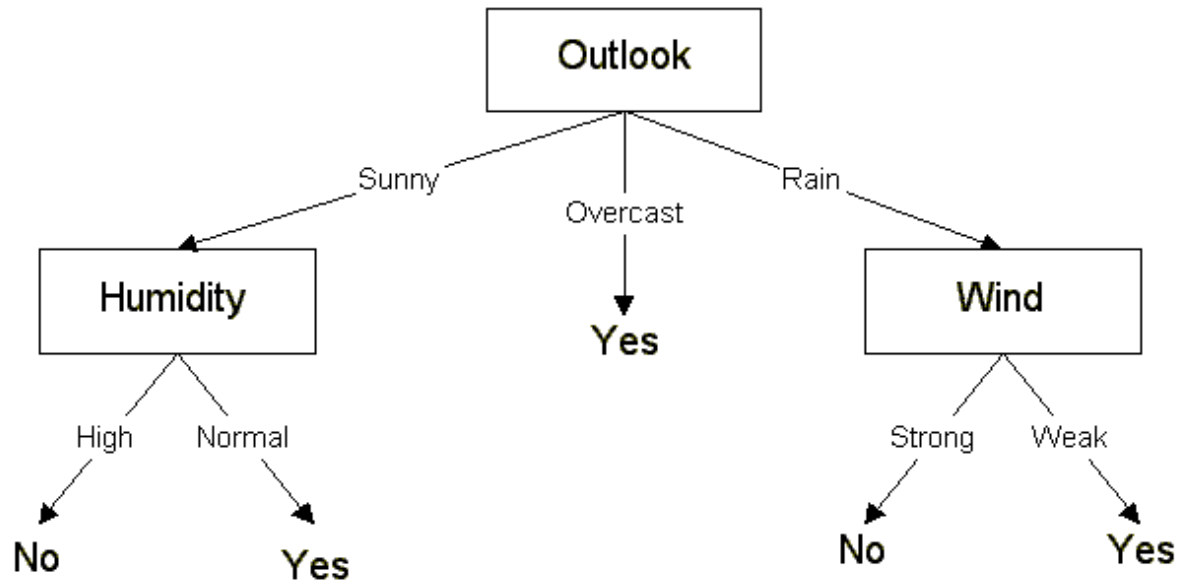
# Support Vector Machine



Supposons que les  $\mathbf{x}$  sont des vecteurs. Le modèle appris par une SVM est dense et a la forme d'une combinaison linéaire.

# Arbre de décision

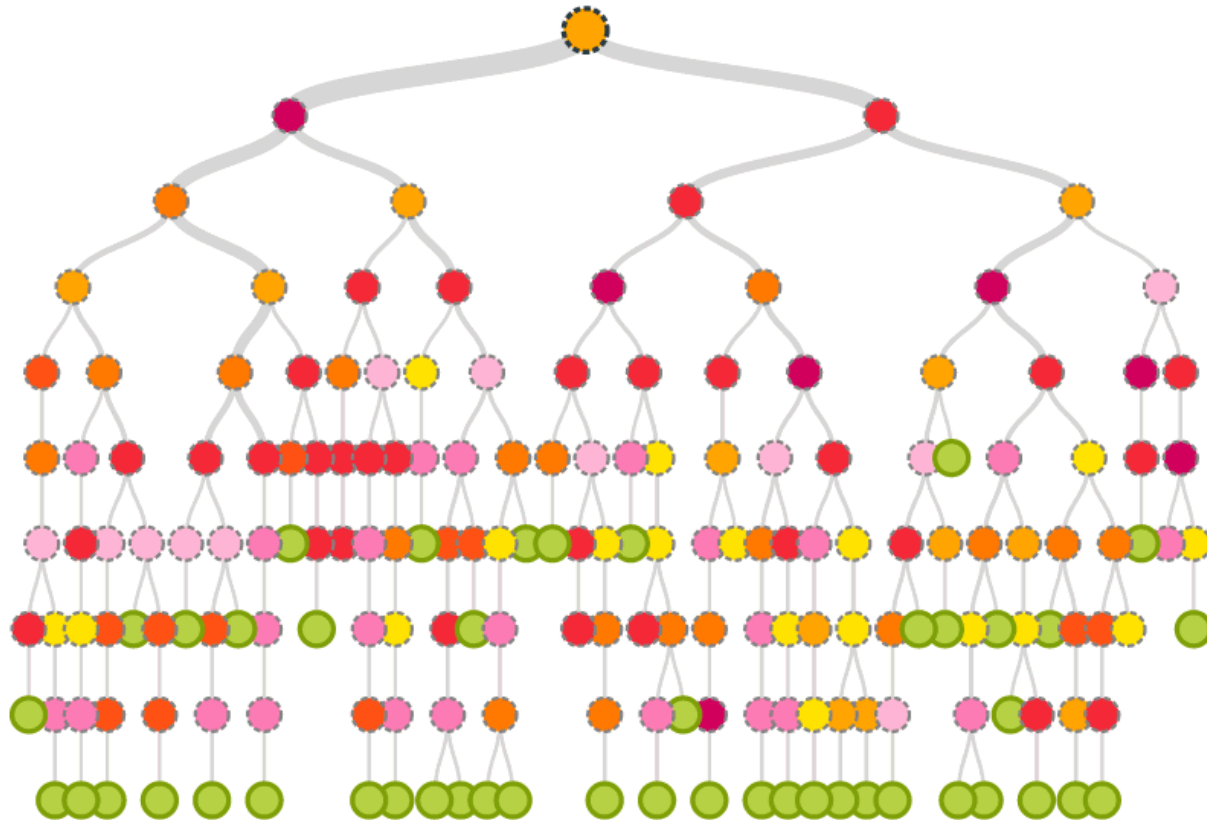
Is it a good time to play baseball?



<http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/Image3.gif>

La structure d'un arbre de décision est simple à interpréter.

# Arbre de décision



Source: [http://littleml.files.wordpress.com/2013/04/decision\\_tree.png](http://littleml.files.wordpress.com/2013/04/decision_tree.png)

Quand il est parcimonieux...

Méthode



# Représentation “bag-of-words”

- k-mer: séquence de k nucléotides
- Soit  $\mathcal{K}$ , l'ensemble de tous les k-mers présents dans les génomes de l'ensemble  $\mathcal{S}$ .
- On représente chaque génome  $\mathbf{x}$  par un vecteur  $\phi(\mathbf{x}) \in \mathbb{R}^{|\mathcal{K}|}$  tel que  $\phi_j(\mathbf{x}) = 1$  si  $k_j \in \mathcal{K}$  et 0 sinon.
- $|\mathcal{K}|$  peut être très grand! (Humain > 30000000000 / individu)

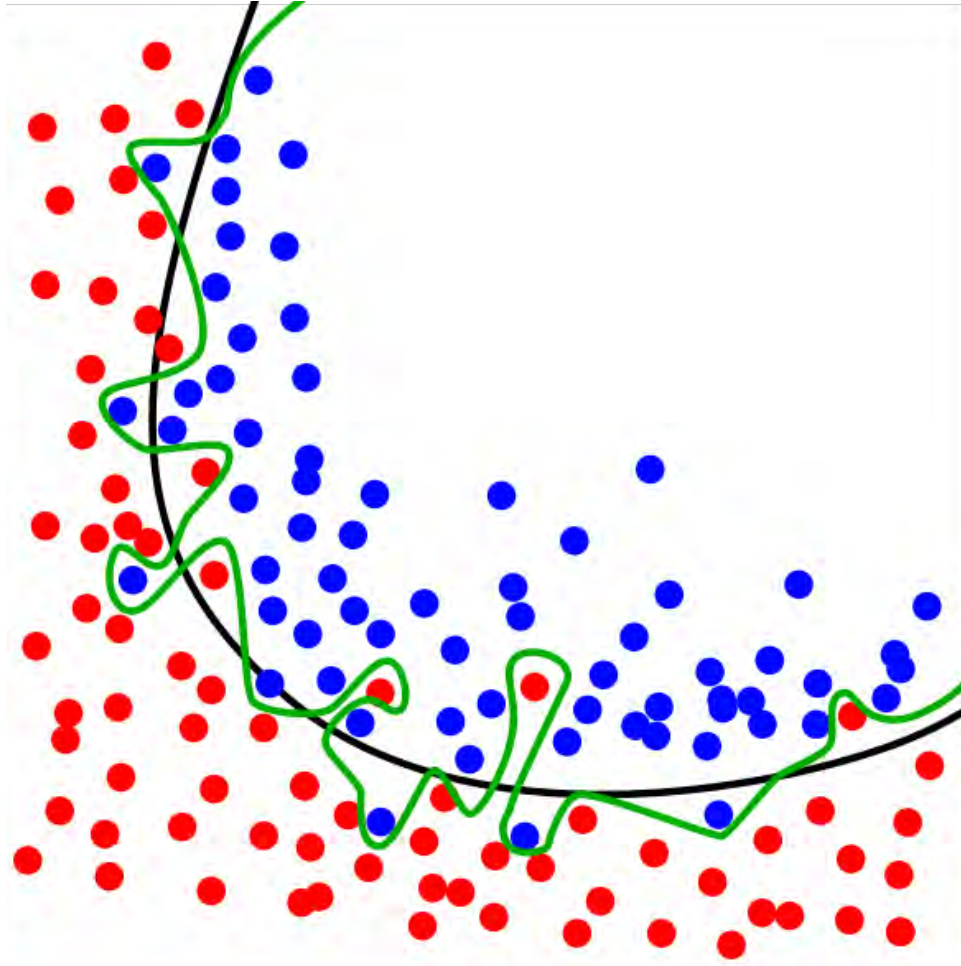
# Représentation “bag-of-words”

$$\mathcal{K} = \left\{ \begin{array}{cccc} \text{CAGATA} & \text{GATAGA} & \text{GAACAG} & \text{CGATGA} \\ \text{AGATAG} & \text{AGAACA} & \text{ATAGAA} & \text{CCGGCT} \\ \text{AACAGC} & \text{TAGAAC} & \text{TTTCGG} & \text{AAATAC} \end{array} \right\}$$

$$\mathbf{x} = \text{CAGATAGAACAGC}$$

$$\phi(\mathbf{x}) = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ \hline \text{CAGATA} & \text{TTTCGG} & \text{AGATAG} & \text{GATAGA} & \text{CGATGA} & \text{AACAGC} & \text{ATAGAA} & \text{CCGGCT} & \text{TAGAAC} & \text{GAACAG} & \text{AGAACA} & \text{AAATAC} \\ \hline \end{array}$$

# Attention au surapprentissage!



# Set Covering Machine

- Proposé par Marchand et Shawe-Taylor en 2003
- Modèles:
  - ★ Parcimonieux
  - ★ Conjonctions/disjonctions de règles à valeur booléenne:
$$r : \mathbb{R}^{|\mathcal{K}|} \rightarrow \{0, 1\}$$
- Complexité algorithmique optimale  $O(m|\mathcal{K}|)$

# Règles à valeur booléenne

- Pour tout k-mer  $k_j \in \mathcal{K}$ :

★ Règle de présence:

$$p_{k_j}(\phi(\mathbf{x})) = I[\phi_j(\mathbf{x}) = 1]$$

★ Règle d'absence:

$$a_{k_j}(\phi(\mathbf{x})) = I[\phi_j(\mathbf{x}) = 0]$$

# Set Covering Machine

---

**Algorithm 1:** TrainConjunctionSCM( $S, \mathcal{R}, p, s$ )

---

**input:**  $S$ : Set of training examples,  $\mathcal{R}$ : Set of boolean-valued rules,  
 $p$ : Class tradeoff parameter,  $s$ : Early stopping parameter.

$\mathcal{R}^* \leftarrow \emptyset$

$\mathcal{P} \leftarrow$  the set of examples in  $S$  with label 1

$\mathcal{N} \leftarrow$  the set of examples in  $S$  with label 0

**while**  $\mathcal{N} \neq \emptyset$  **and**  $|\mathcal{R}^*| < s$  **do**

$\forall r_i \in \mathcal{R}, \mathcal{A}_i \leftarrow$  the subset of  $\mathcal{N}$  correctly classified by  $r_i$

$\forall r_i \in \mathcal{R}, \mathcal{B}_i \leftarrow$  the subset of  $\mathcal{P}$  misclassified by  $r_i$

$\forall r_i \in \mathcal{R}, U_i \leftarrow |\mathcal{A}_i| - p \cdot |\mathcal{B}_i|$

$i^* \leftarrow \operatorname{argmax}_i U_i$

$\mathcal{R}^* \leftarrow \mathcal{R}^* \cup \{r_{i^*}\}$

$\mathcal{N} \leftarrow \mathcal{N} - \mathcal{A}_{i^*}$

$\mathcal{P} \leftarrow \mathcal{P} - \mathcal{B}_{i^*}$

**return**  $h$ , where  $h(\mathbf{x}) = \bigwedge_{r_i^* \in \mathcal{R}^*} r_i^*(\phi(\mathbf{x}))$

---

# Apprentissage de disjonctions

1. Créer un nouvel ensemble de données:

$$\mathcal{S}' = (\mathbf{x}_i, \neg y_i) : (\mathbf{x}_i, y_i) \in \mathcal{S}$$

2. Utiliser SCM pour apprendre à partir de  $\mathcal{S}'$

$$h = \bigwedge_{r_i^* \in \mathcal{R}^*} r_i^*(\phi(\mathbf{x}))$$

3. Appliquer la loi de De Morgan:

$$\neg \bigwedge_{r_i^* \in \mathcal{R}^*} r_i^*(\phi(\mathbf{x})) = \bigvee_{r_i^* \in \mathcal{R}^*} \neg r_i^*(\phi(\mathbf{x}))$$

# Implémentation



# Matrice de Kmers

Nous utilisons Ray Surveyor pour produire une matrice de “bag-of-words” à partir des génomes.

$|\mathcal{K}|$

---

$m$

1	1	0	0	1	0	1	0	0	1	0
1	1	0	1	1	1	0	0	0	1	0
0	1	0	0	1	1	1	1	1	1	1
0	1	0	1	0	1	0	0	1	1	1
1	0	1	1	0	1	0	1	1	1	1
1	1	1	1	0	0	0	0	1	1	1



<https://github.com/zorino/ray>

# Calcul de la fonction d'utilité

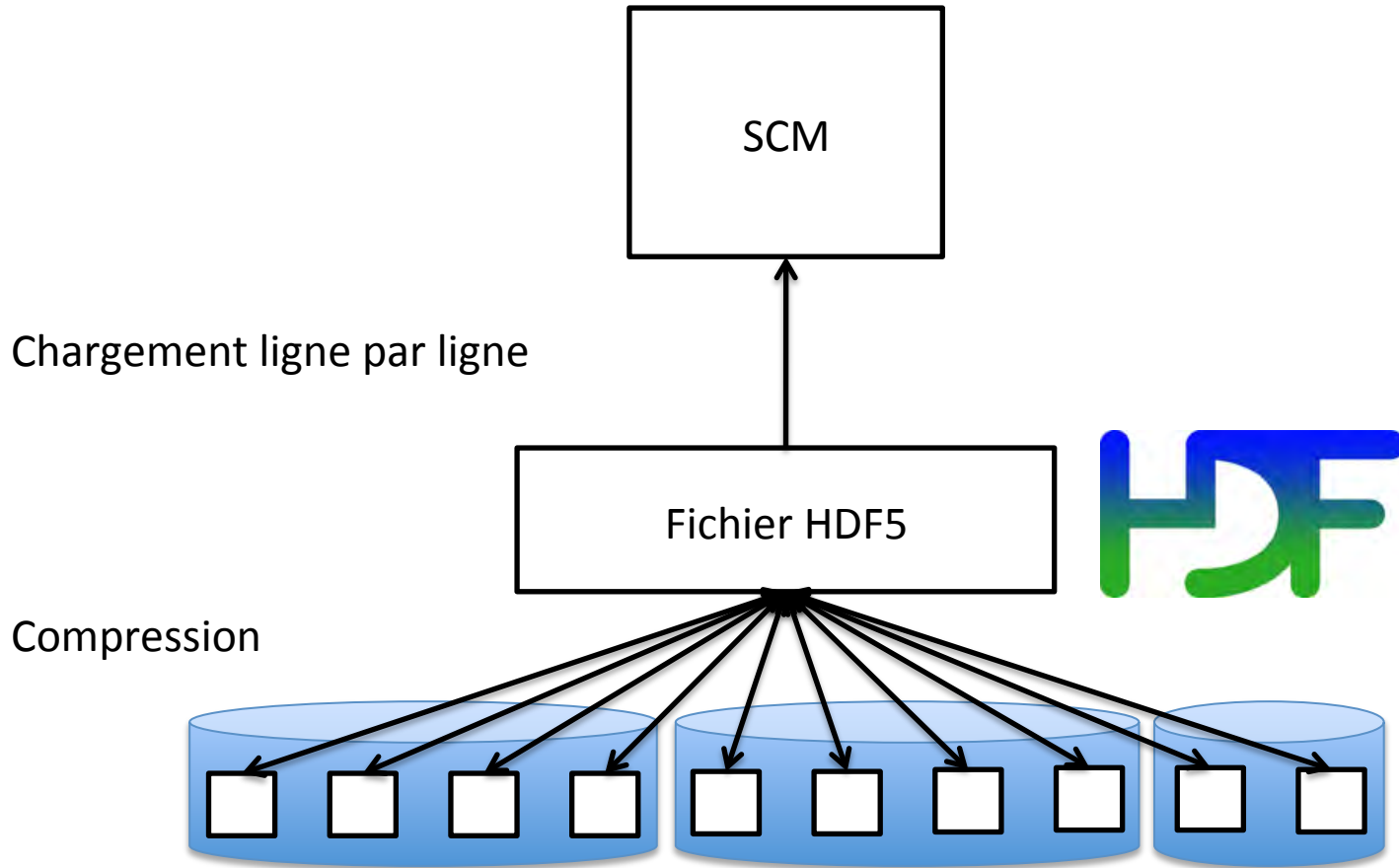
		$ \mathcal{K} $										$ \mathcal{K} $											
$m$		1	1	0	0	1	0	1	0	0	1	0	0	0	1	1	0	1	0	1			
		1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1		
		0	1	0	0	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	
		0	1	0	1	0	1	0	0	1	1	1	1	1	1	0	1	0	1	1	0	0	0
		1	0	1	1	0	1	0	1	1	1	1	1	0	1	0	0	1	0	1	0	0	0
		1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
$ \mathcal{A}_i $	1	1	1	0	3	1	3	2	0	0	0	2	2	2	3	0	2	0	1	3	3	3	
$ \mathcal{B}_i $	1	0	3	2	0	1	1	2	2	0	2	2	3	0	1	3	2	2	1	1	3	1	

# Calcul de la fonction d'utilité

		$ \mathcal{K} $										$ \mathcal{K} $									
$m$		1	1	0	0	1	0	1	0	0	1	0	0	0	1	1	0	1	0	1	
		1	1	0	1	1	1	0	0	0	1	0	0	0	0	1	1	1	0	1	
		0	1	0	0	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	
		0	1	0	1	0	1	0	0	1	1	1	1	1	1	0	1	1	0	0	
		1	0	1	1	0	1	0	1	1	1	1	0	1	0	0	1	0	0	0	
		1	1	1	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	0	

- Astuce: regrouper les bits de chaque colonne dans des int64
  - 8 x moins de mémoire
  - Instruction popcount => 64 x moins d'opérations!

# Stockage et accès aux données



# Résultats

# Résistance aux antibiotiques

“Le rapport 2014 de l’OMS sur la surveillance mondiale de la résistance aux antimicrobiens révèle que **la résistance aux antibiotiques n’est plus un souci pour l’avenir**; c’est une réalité partout dans le monde aujourd’hui, qui risque de **compromettre notre capacité à traiter des infections courantes** dans la communauté comme dans les hôpitaux.”

– Organisation mondiale de la santé, 2014

# Ensembles de données

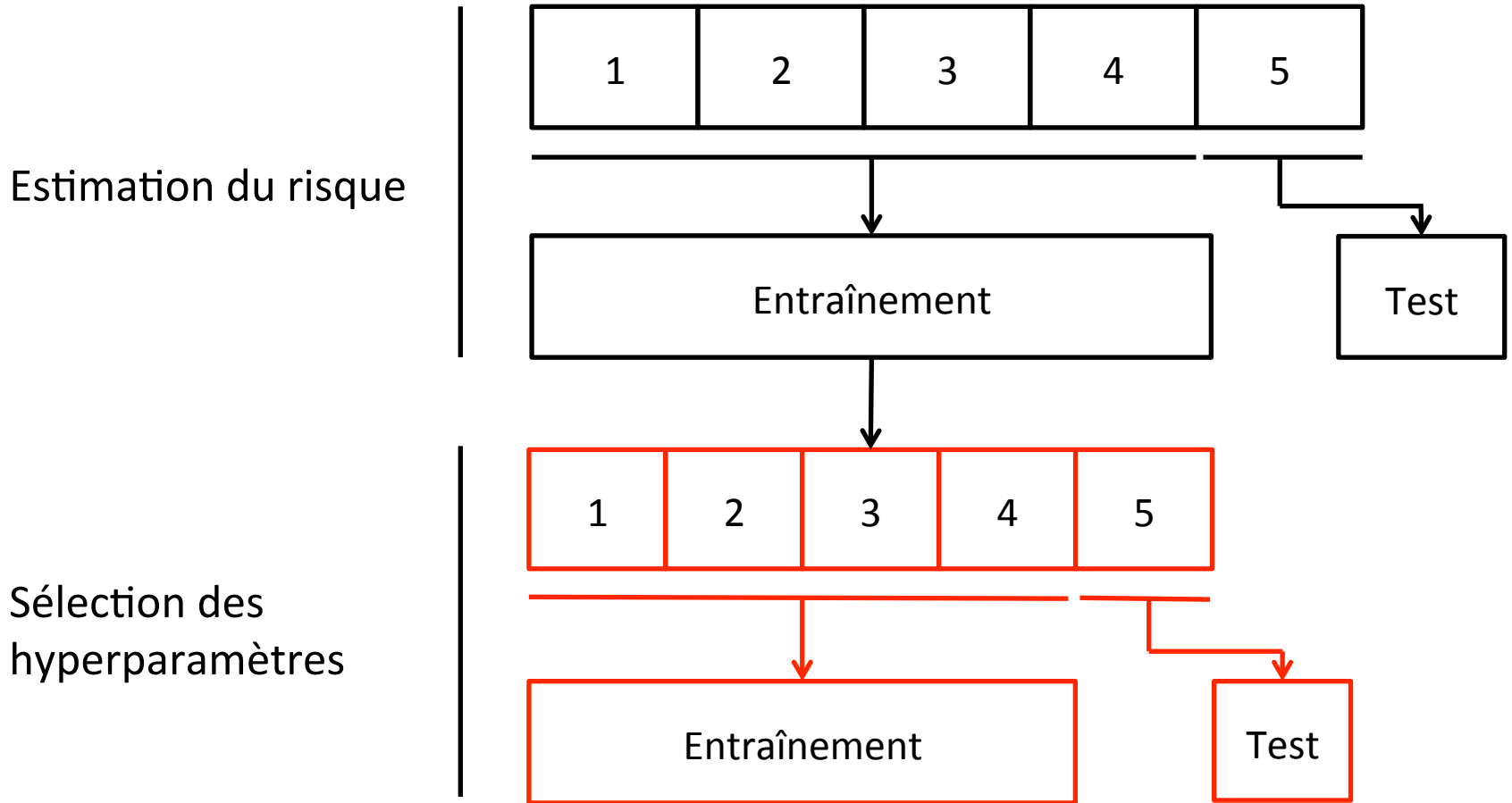
<i>Clostridium Difficile</i>	$ \mathcal{K}  = 32\,823\,803$ , $m = 470$ Source: Jacques Corbeil + Vivian Loo
<i>Pseudomonas Aeruginosa</i>	$ \mathcal{K}  = 132\,487\,288$ , $m = 393$ Source: AstraZeneca
<i>Mycobacterium tuberculosis</i>	$ \mathcal{K}  = 11\,255\,033$ , $m = 154$ Source: PMID25599400
<i>Streptococcus pneumoniae</i>	$ \mathcal{K}  = 10\,542\,251$ , $m = 680$ Source: PMID23644493

# Filtres univariés

- Filtrer certains k-mers avant l'apprentissage selon le résultat d'un test statistique univarié.
- Le test du  $\chi^2$  de Pearson pour l'indépendance mesure la dépendance entre chaque k-mer et les étiquettes.
- p-value: probabilité d'obtenir une certaine valeur du  $\chi^2$  sachant qu'un k-mer est indépendant des étiquettes
- Correction de la FDR par la méthode de Benjamini-Yekutieli



# Validation croisée imbriquée



# Résultats de prédiction

## *Clostridium Difficile*

	SCM	Best	DT	L1SVM	L2SVM
Azithromycine	0.015 (3.2)	0.111 (1.0)	0.056 (8.8)	0.050 (3620.2)	0.041 (2462244.2)
Ceftriaxone	0.070 (2.0)	0.877 (1.0)	0.126 (7.8)	0.067 (623.8)	0.088 (2449563.8)
Clarithromycine	0.015 (3.0)	0.091 (1.0)	0.073 (10.8)	0.071 (1288.6)	0.074 (2463616.8)
Clindamycine	0.025 (2.0)	0.019 (1.0)	0.008 (2.4)	0.008 (971.4)	0.027 (2106950.2)
Moxifloxacin	0.019 (1.0)	0.019 (1.0)	0.019 (1.0)	0.022 (414.0)	0.041 (2487703.6)
Moyenne	0.029 (2.24)	0.223 (1.0)	0.056 (6.16)	0.044 (1383.6)	0.054 (2394015.7)

# Résultats de prédiction

## *Streptococcus Pneumoniae*

	SCM	Best	DT	L1SVM	L2SVM
Benzylopenicillin	0.012 (1.0)	0.009 (1.0)	0.010 (1.6)	0.014 (214.8)	0.014 (675375.6)
Erythromycine	0.031 (2.0)	0.049 (1.0)	0.047 (7.0)	0.045 (378.8)	0.041 (581550.6)
Tetracycline	0.025 (1.2)	0.025 (1.0)	0.025 (1.0)	0.025 (445.5)	0.025 (616914.4)
Moyenne	0.023 (1.4)	0.028 (1.0)	0.027 (3.2)	0.028 (346.3)	0.027 (624613.5)

# Résultats de prédiction

## *Pseudomonas Aeruginosa*

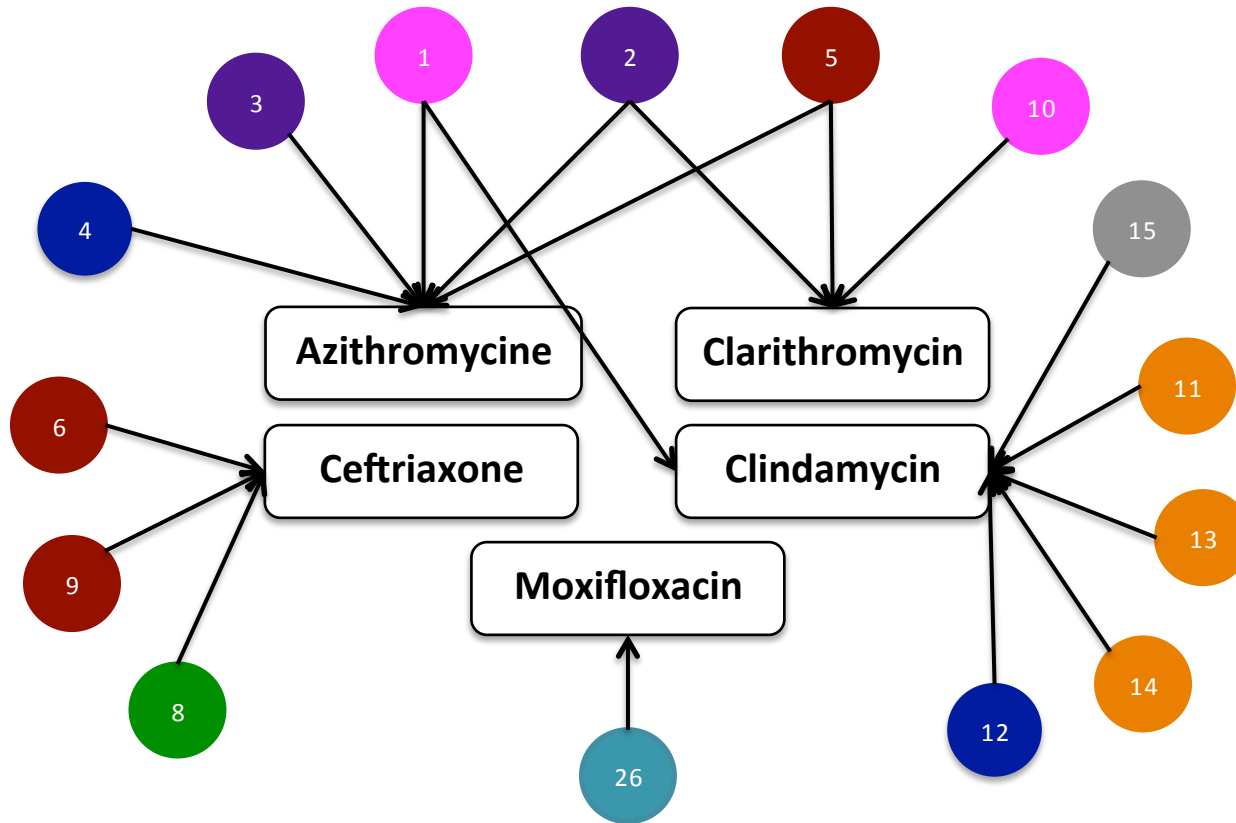
	SCM	Best	DT	L1SVM	L2SVM	Dummy
Amikacin	0.181 (6.0)	0.192 (1.0)	0.195 (21.0)	0.181 (710.0)	0.195 (115537.6)	0.230
Doripenem	0.234 (1.4)	0.264 (1.0)	0.253 (11.4)	0.262 (1532.8)	0.25 (11882.2)	0.378
Meropenem	0.280 (1.8)	0.291 (1.0)	0.253 (4.4)	0.250 (174.4)	0.261 (2961.2)	0.472
Levofloxacin	0.067 (1.4)	0.073 (1.0)	0.067 (1.4)	0.067 (15.6)	0.092 (100928.2)	0.416
Moyenne	0.191 (2.65)	0.205 (1.0)	0.192 (9.55)	0.190 (608.2)	0.2 (57827.3)	0.347

# Résultats de prédiction

## Mycobacterium Tuberculosis

	SCM	Best	DT	L1SVM	L2SVM
Ethambutol	0.209 (1.6)	0.297 (1.0)	0.230 (4.2)	0.202 (208.8)	0.202 (3219.8)
Isoniazid	0.021 (1.0)	0.979 (1.0)	0.014 (1.0)	0.021 (20.2)	0.029 (1550.2)
Rifampicin	0.035 (1.4)	0.035 (1.0)	0.035 (1.4)	0.028 (40.2)	0.029 (180.2)
Streptomycin	0.037 (1.0)	0.963 (1.0)	0.030 (1.4)	0.037 (49.4)	0.037 (1185.6)
Moyenne	0.076 (1.72)	0.569 (1.0)	0.078 (2.0)	0.072 (79.7)	0.074 (1554.0)

# Validation biologique



■ DNA gyrase subunit A

■ Tn6194-like Transposon, other

■ Two-component sensor histidine kinase

■ Transposon Tn6110 and Clostridium Saccharolyticum 23S rRNA m(2)A-2503 methyltransferase

■ Penicillin-binding protein

■ Conjugative transposon FtsK/SpoIIIE-like

■ ErmB rRNA adenine N-6-methyltransferase

■ Other: hypothetical proteins and unmatched k-mers

# Conclusion

- Notre méthode permet d'apprendre des modèles parcimonieux à partir de génomes complets.
- Ces modèles sont composés de règles facilement interprétables.
- Nos modèles ont une forme qui les rend particulièrement faciles à transformer en tests cliniques.
- Travaux futurs: corrélation entre certains k-mers, robustesse aux mutations, modèles plus complexes (DNF-SCM), cancer.

# Remerciements

- Personnes impliquées dans le projet:

Apprentissage Automatique	Bioinformatique
<ul style="list-style-type: none"><li>• Sébastien Giguère</li><li>• Vladana Sagatovich</li><li>• François Laviolette</li><li>• Mario Marchand</li></ul>	<ul style="list-style-type: none"><li>• Maxime Déraspe</li><li>• Frédéric Raymond</li><li>• Jacques Corbeil</li><li>• Paul H. Roy</li></ul>
Lynda Robitaille	



# Références

- Drouin, A., Giguère, S., Sagatovich, V., Déraspe, M., Laviolette, F., Marchand, M., & Corbeil, J. (2014). Learning interpretable models of phenotypes from whole genome sequences with the Set Covering Machine. *arXiv preprint arXiv:1412.1074*.
- Marchand, M., & Shawe-Taylor, J. (2003). The set covering machine. *The Journal of Machine Learning Research*, 3, 723-746.
- World Health Organization. (2014). Antimicrobial resistance. Factsheet 194, 2013.