PARIS DESCARTES

Data Mining/Machine Learning/ Big Data

What's this ?

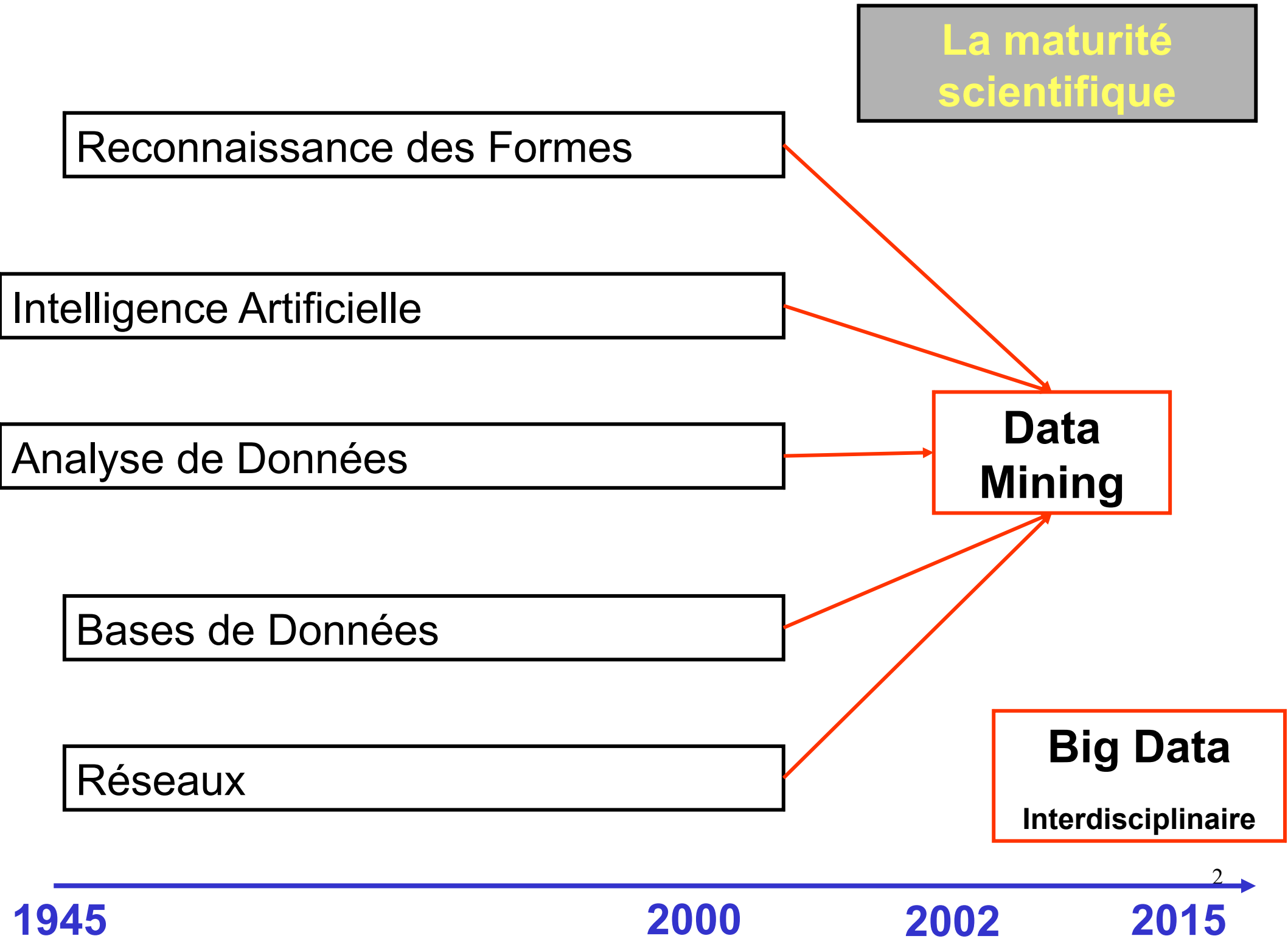
Biblio :

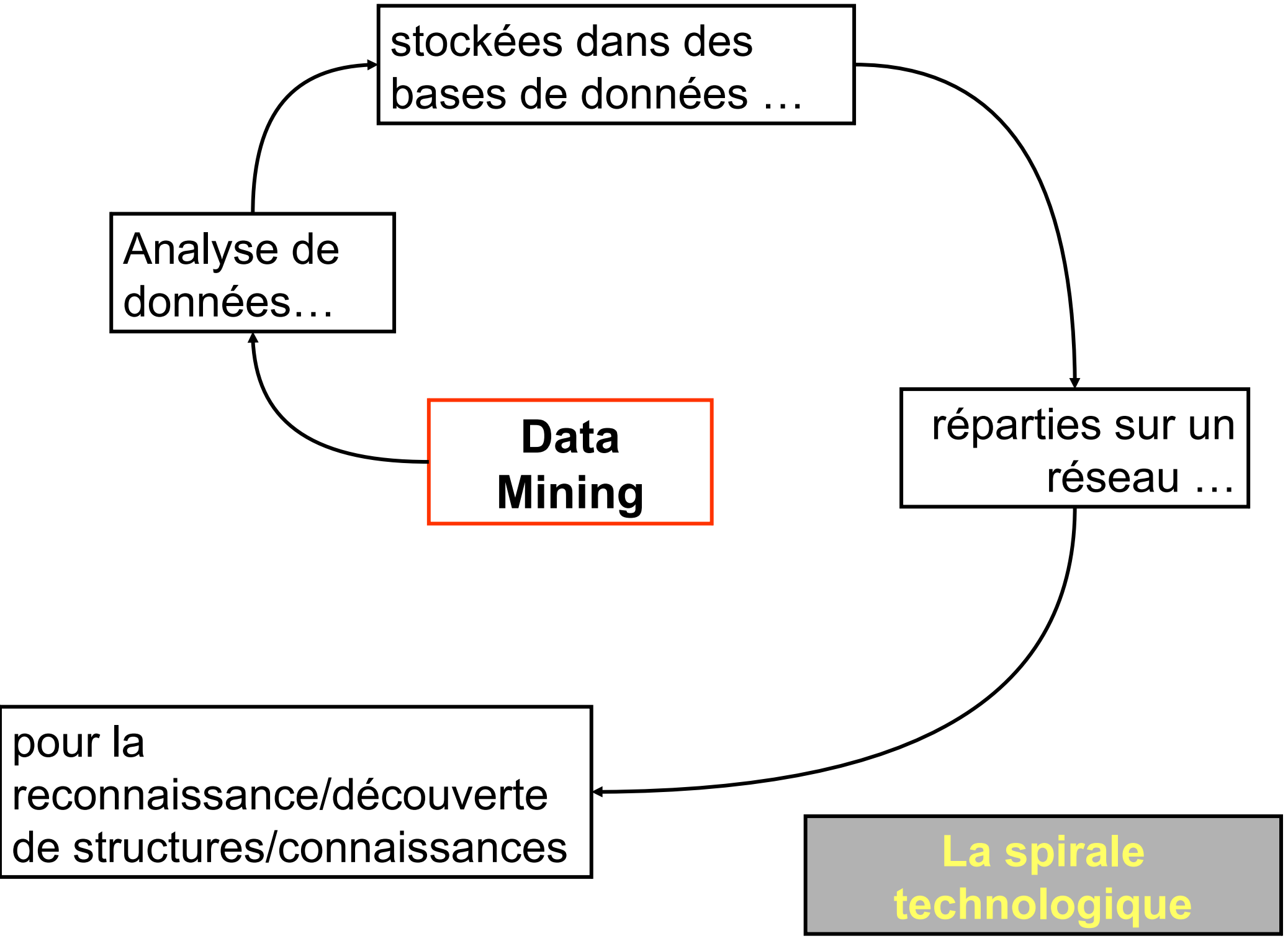
Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning) – 2001 par Pierre Baldi et Søren Brunak

Introduction à la bioinformatique - 2001, par Cynthia Gibas et Per Jambeck. (traduit de l'anglais)

-
- Webs:

http://chem-eng.utoronto.ca/~datamining/dmc/data_mining_map.htm
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0115892
http://svmcompbio.tuebingen.mpg.de/
http://infolab.stanford.edu/~ullman/pub/book.pdf
http://www.math.cmu.edu/~ctsourak/resources.html


La maturité scientifique
Reconnaissance des Formes
Intelligence Artificielle
Analyse de Données
Bases de Données
Réseaux
Data Mining
Big Data
Interdisciplinaire
2
1945
2000
2002
2015


stockées dans des bases de données …
réparties sur un réseau …
pour la reconnaissance/découverte de structures/connaissances
Analyse de données…
Data Mining
La spirale technologique

Exploration de données

Data Mining

Fouille de Données

?

KDD

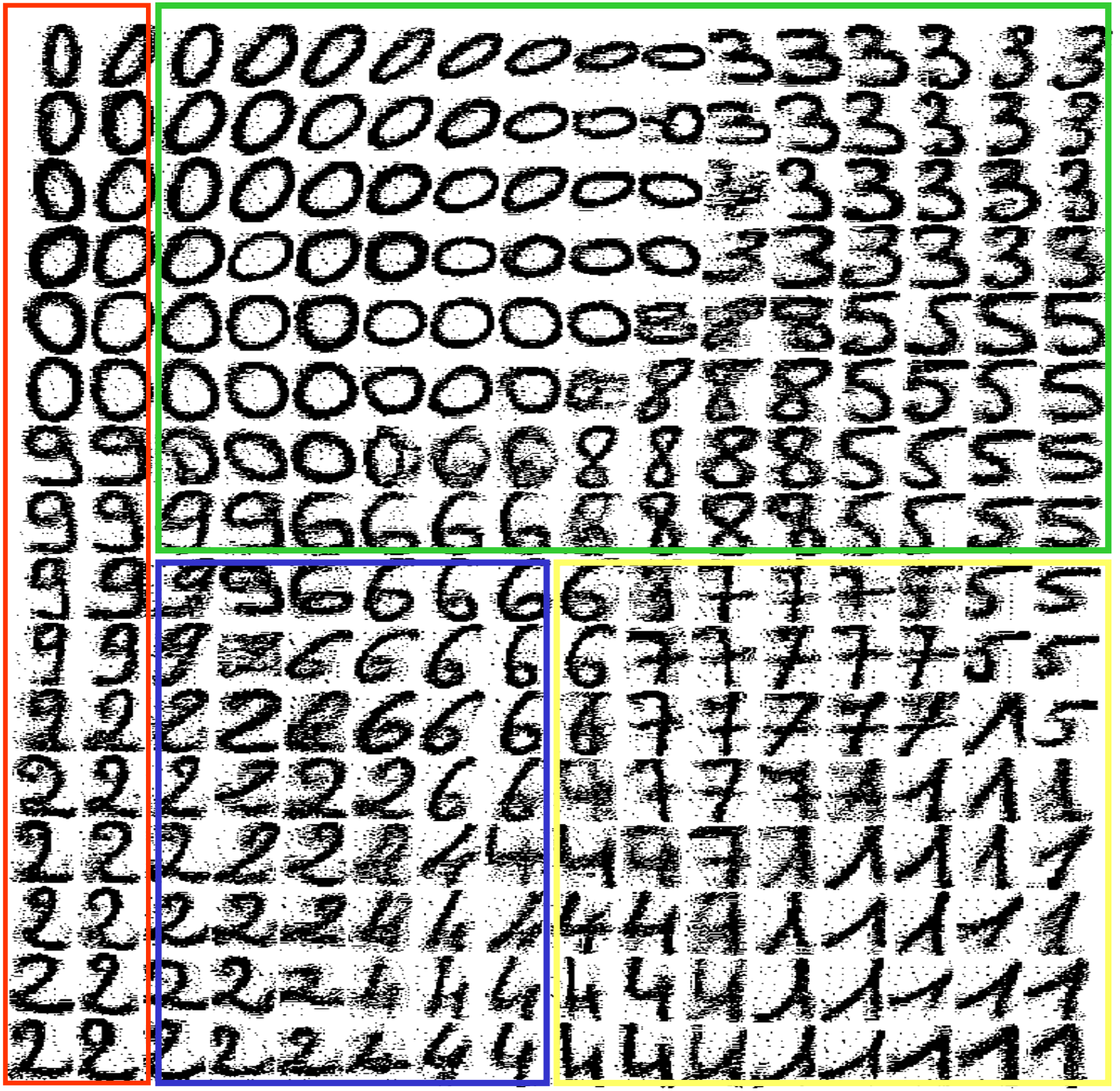
Comment détecter des ressemblances, des structures, des motifs *a priori* ?

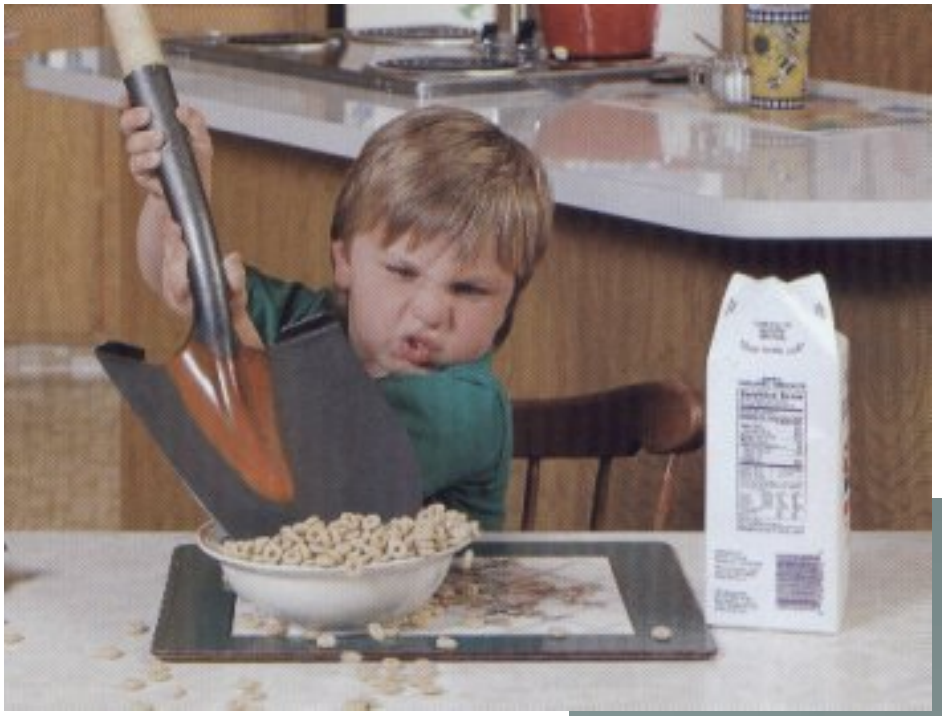
Jean-Paul

Samia

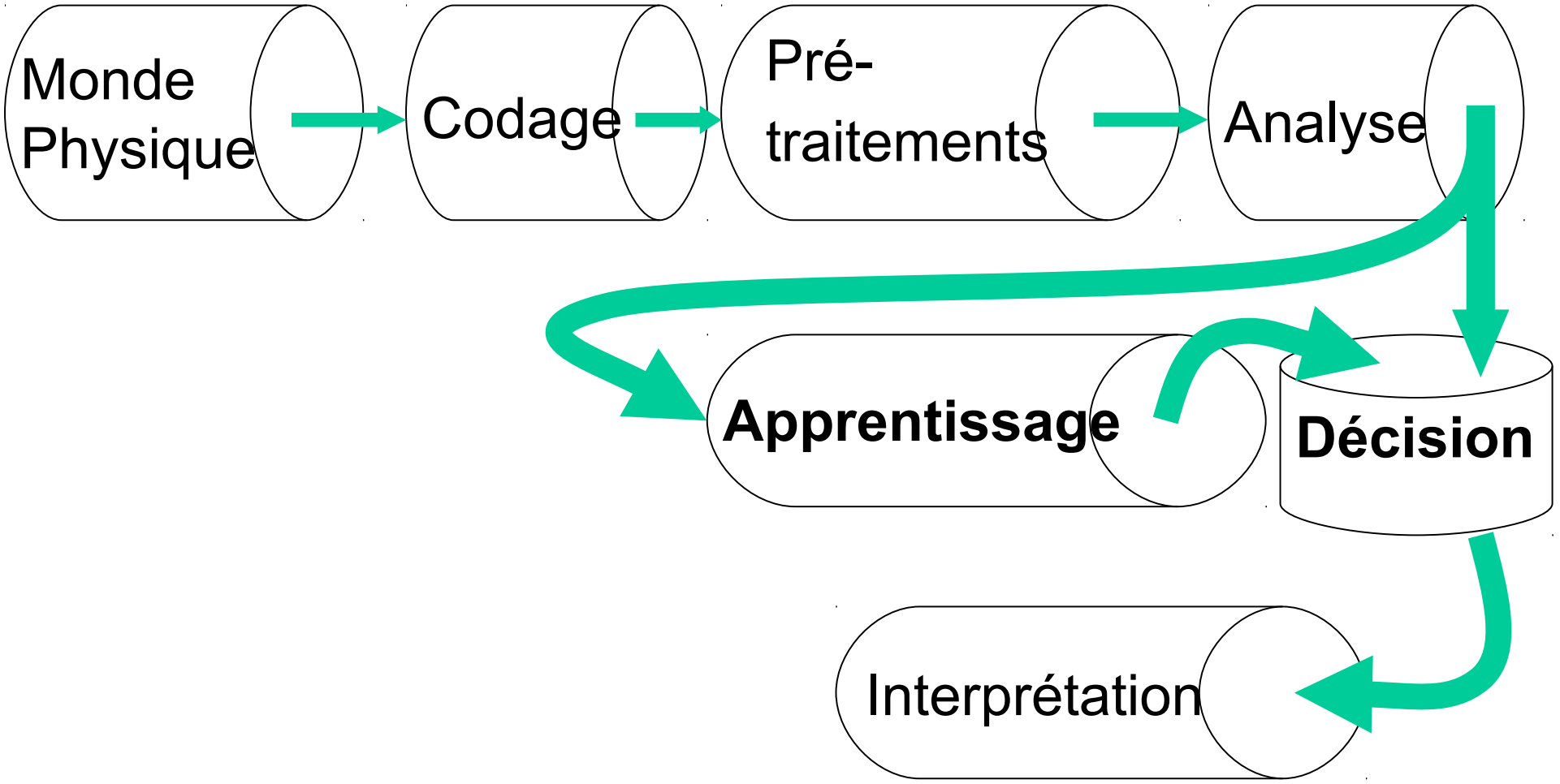
Lin

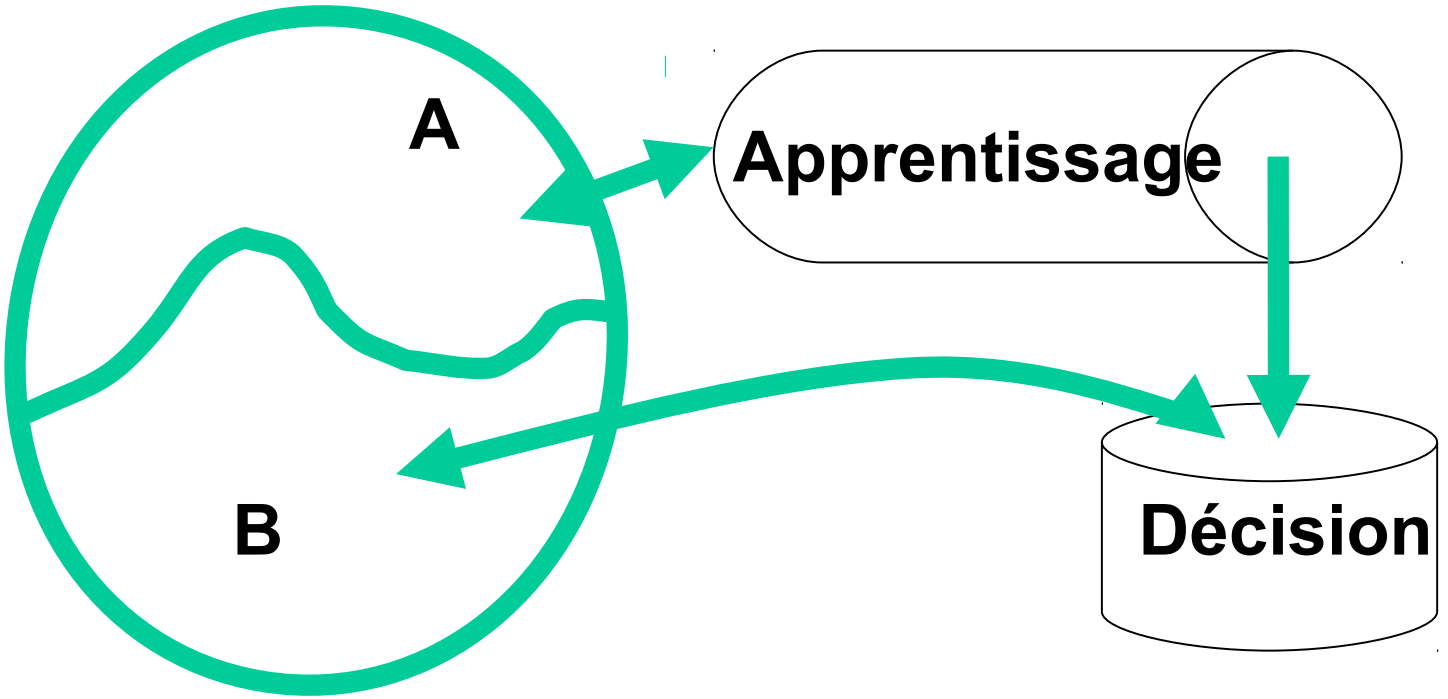
Samy





Système de Reconnaissance de Formes Classique

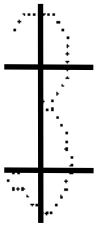




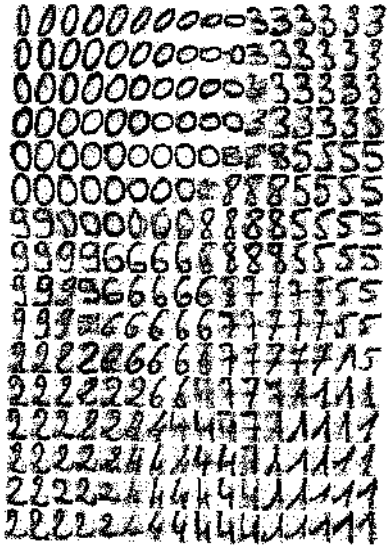
A = Ensemble d ’échantillons pour chaque classe

$A = \varnothing \Rightarrow$ Apprentissage Non Supervisé

$A \neq \varnothing \Rightarrow$ Apprentissage Supervisé

Codage →  Prétraitement →  Analyse → 

- Filtrage du bruit
- Normalisation
- Squelettisation
- Segmentation



Apprentissage

Décision

« 3 »

avec un score de 0.6

On travaille en général sur deux ensembles : un ensemble d'apprentissage A et un ensemble de test T.

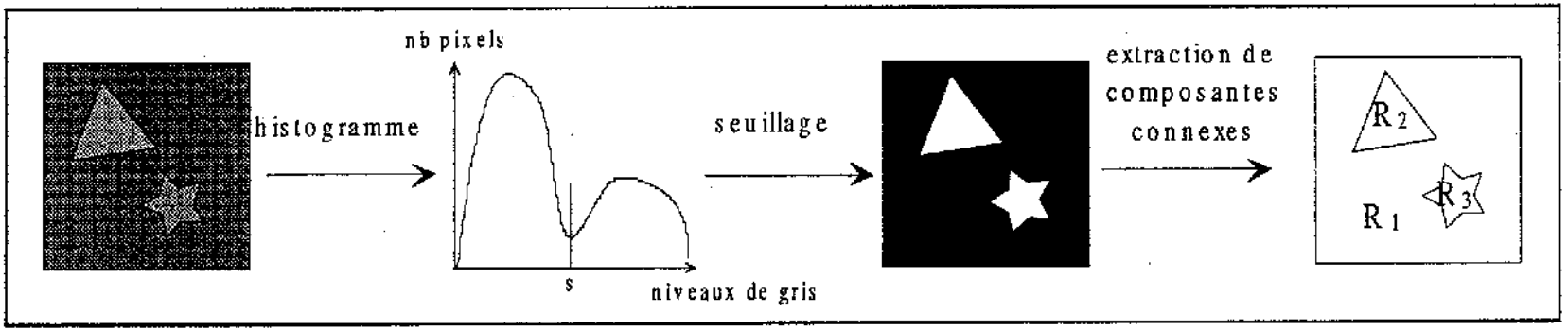
Une estimation du risque réel de l'hypothèse h proposée sur l'ensemble de test T peut être obtenue à partir de la matrice dite de confusion.

Dans le cas binaire par exemple, c'est-à-dire dans le cas du test d'une hypothèse (une classe) indépendamment des autres classes (hypothèses), on a

	'+' prédit	**'-' prédit**
'+' réel	**Vrais positifs**	**Faux négatifs**
'-' réel	**Faux positifs**	**Vrais négatifs**

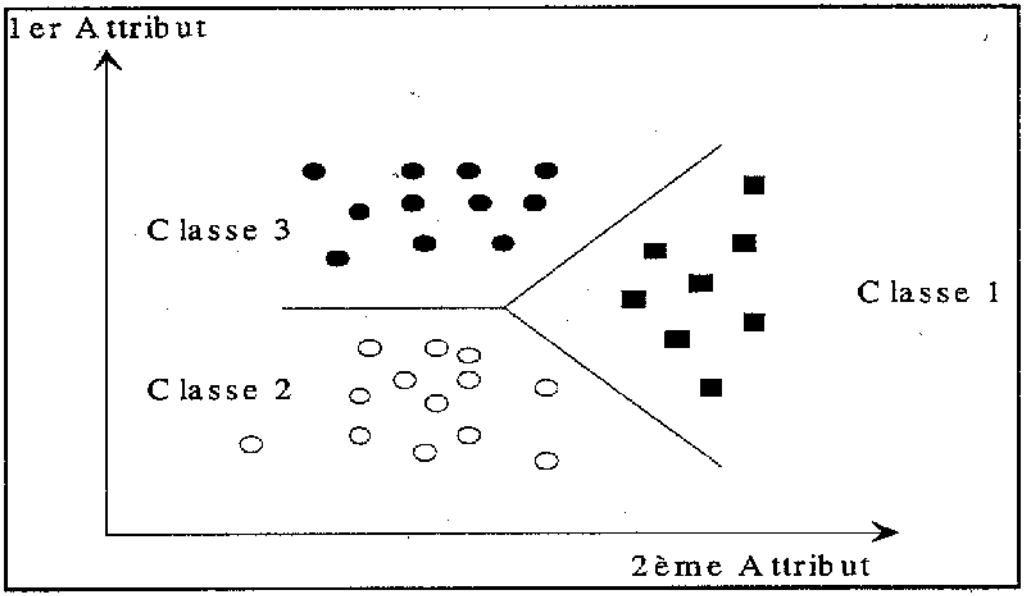
Risque Réel (h) = Somme des termes non diagonaux / Nombre d'exemples
= Somme des exemples mal classés / Nombre d'exemples

Dans le cas NON supervisé, les techniques spécifiques utilisées sont typiques des applications dites de Fouille de Données



Segmentation en régions par seuillage

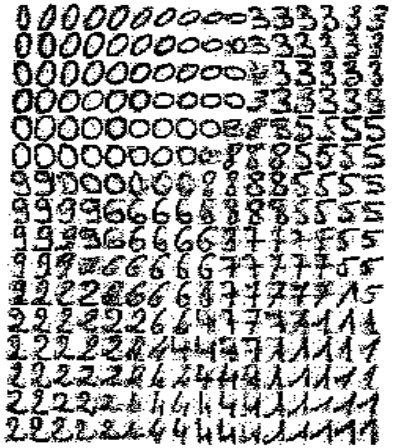
- **Classification**
- **Segmentation**



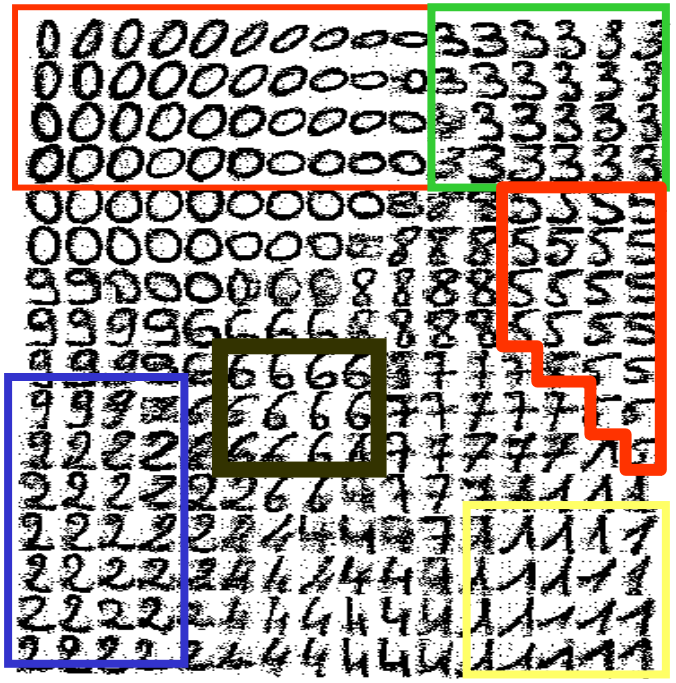
Classification des points dans l'espace des attributs
(3 classes, dimension 2 de l'espace)

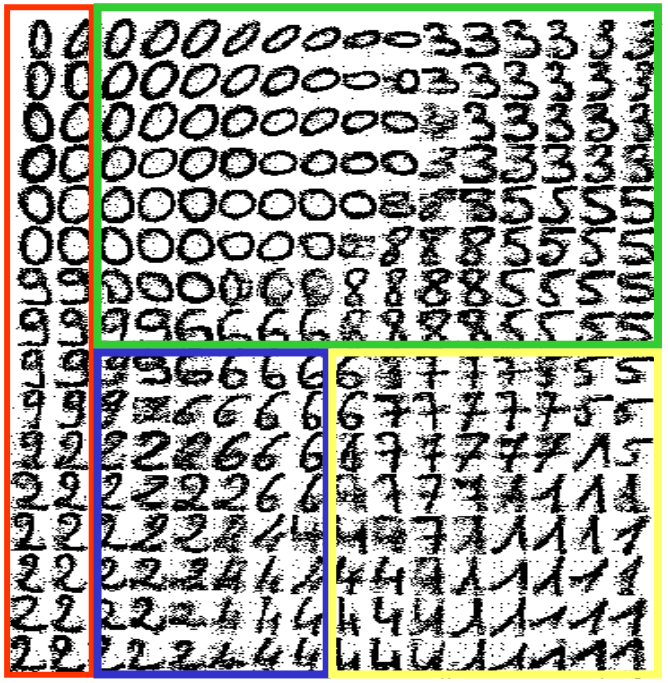
Un problème typique visuel qui pourrait relever de la problématique de la Fouille de Données plus que Reconnaissance des Formes

On donne ces données stockées sur des supports électroniques hétérogènes et non centralisés :



Alors sans intervention de type supervisé (cad sans apprentissage avec exemples), le système parvient à détecter (structurer, extraire) la présence de 10 formes différentes sans forcément les reconnaître, ou bien de 4 scripteurs différents sans forcément les identifier dans un premier temps :





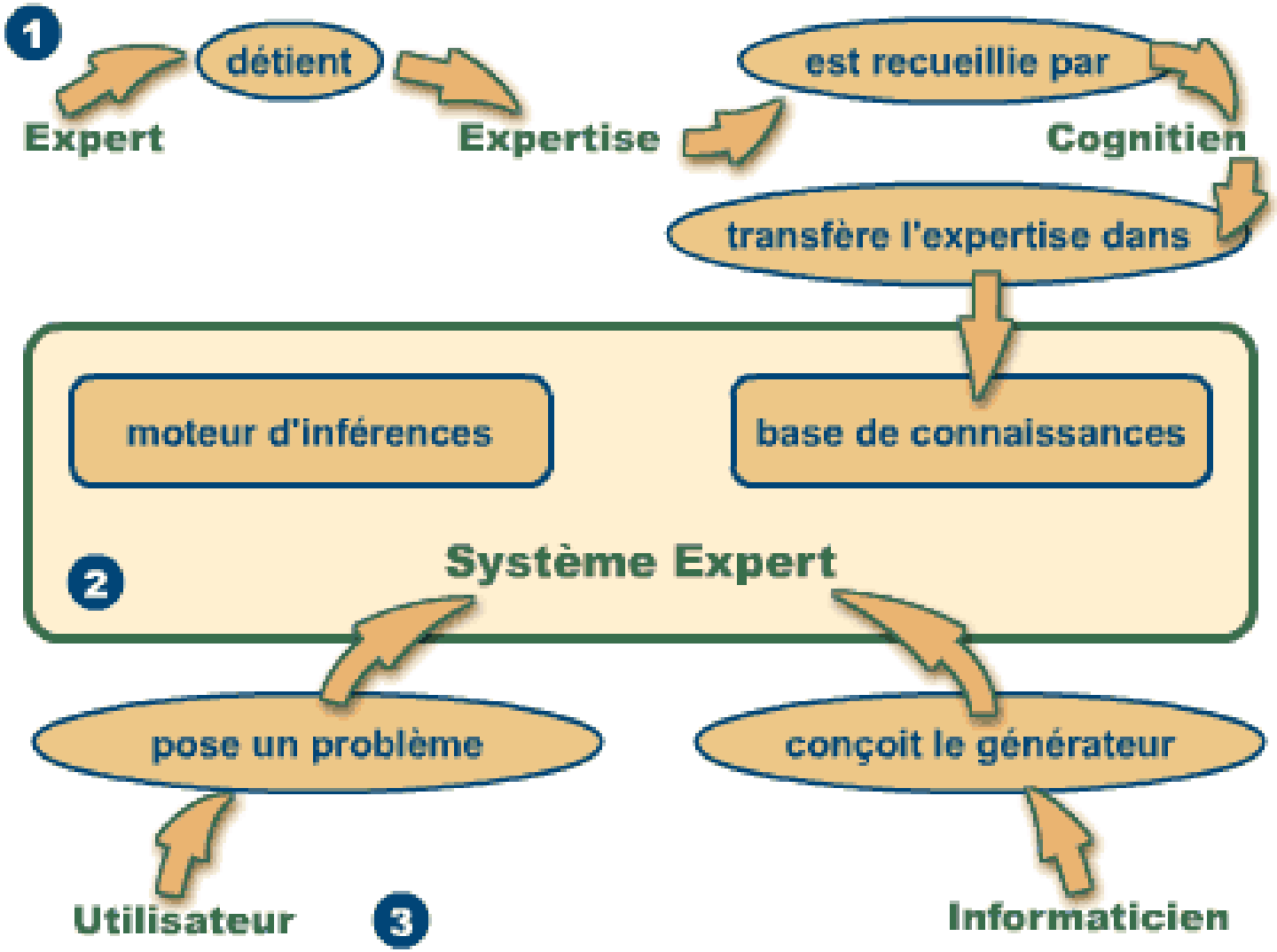
Outre le paradigme de Reconnaissance des Formes, cette intégration nouvelle ou ce paradigme nouveau est la résultante de problématiques arrivées à maturité ou à leur limite comme :

- Les systèmes experts issus de l'IA
- Les bases et les entrepôts de données
- Les protocoles réseaux normalisés

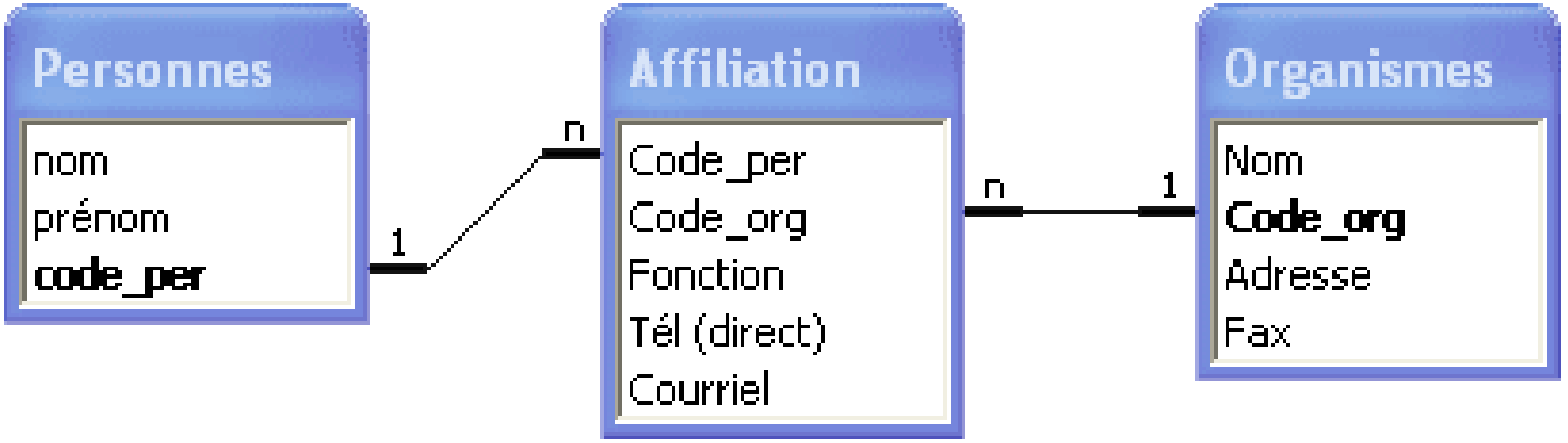
Créer une intelligence des systèmes, avec les potentialités de chacun des outils technologiques intégrés -> le rêve de système pensant plus que pensé

Différence de points de vue entre : SELECTIONNE moi les NOMS des CLIENTS ayant acheté du NUTELLA et du SAVON (requête de type SQL) et je (le logiciel) te (l'utilisateur du logiciel) fais remarquer que les clients qui achète du Nutella achètent aussi du Savon

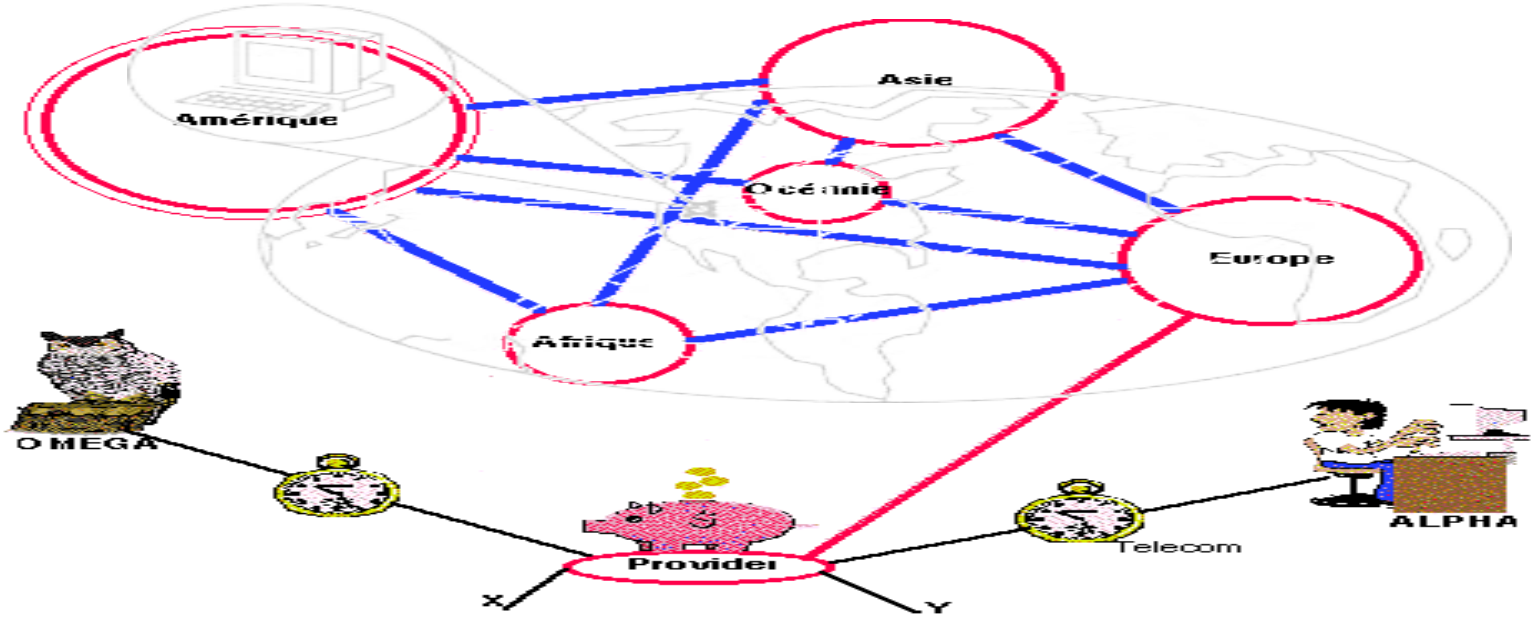
Système Expert Classique



Base de Données Classique

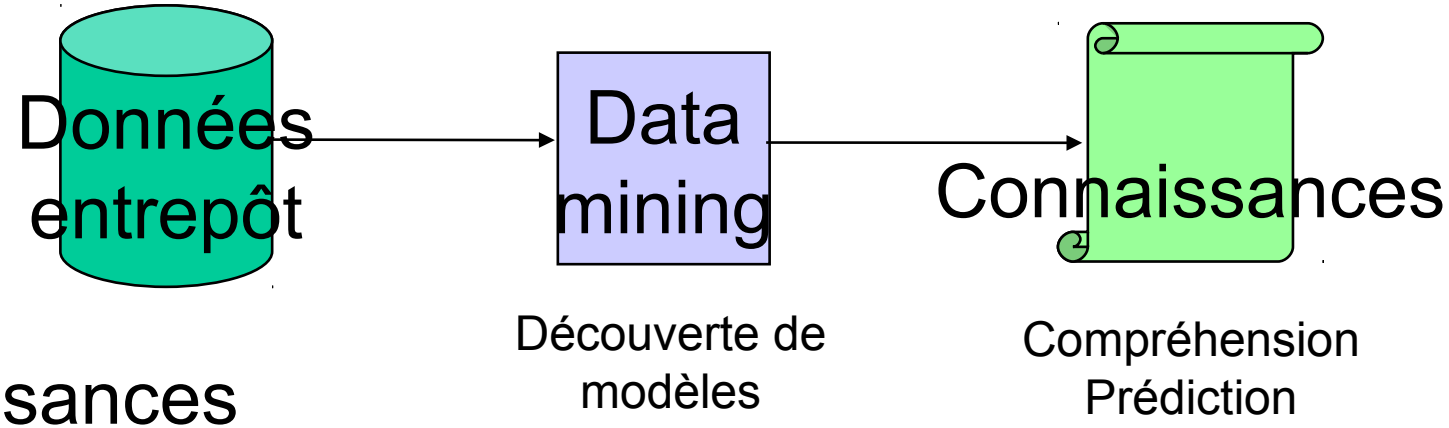


Protocole de Communication Réseau Classique



Qu ’est-ce-que le data mining ?

- Data mining
  - –ensembles de techniques d'exploration de données afin d'en tirer des connaissances (la substantifique moelle) sous forme de modèles présentées à l ’utilisateur averti pour examen



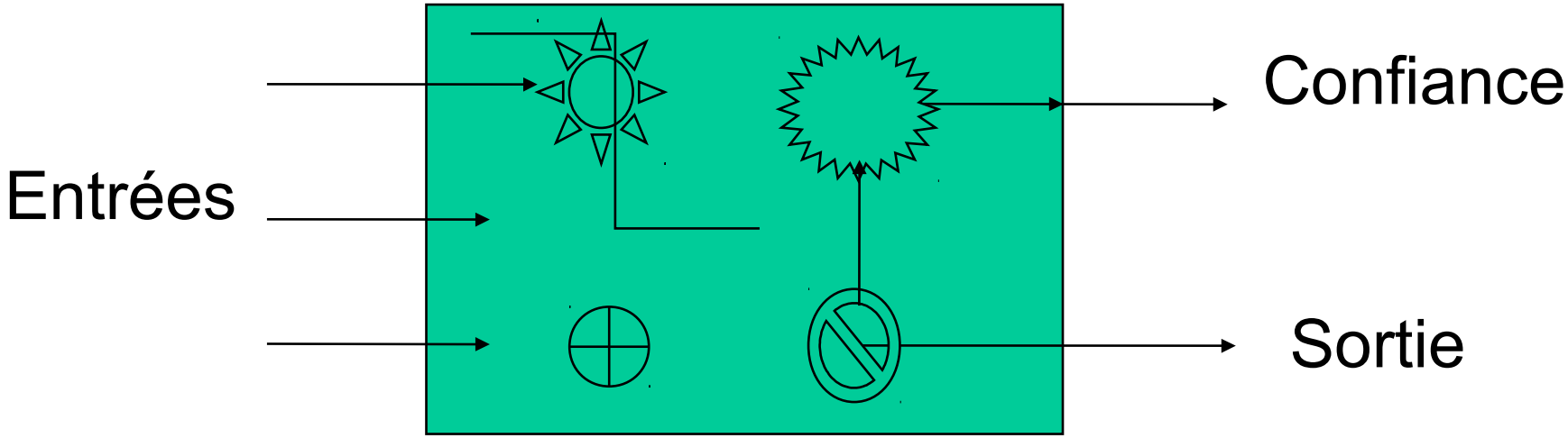
- Connaissances
  - –analyses (distribution du trafic en fonction de l ’heure)
  - –scores (fidélité d ’un client), classes (mauvais payeurs)
  - –règles **(si** facture > 10000 **alors** départ à 70%)

Mécanismes de base

- Déduction : base des systèmes experts
  - schéma logique permettant de déduire un théorème à partir d'axiomes
  - le résultat est sûr, mais la méthode nécessite la connaissance de règles
- Induction : base du data mining
  - méthode permettant de tirer des conclusions à partir d'une série de faits
  - généralisation un peu abusive
  - indicateurs de confiance permettant la pondération

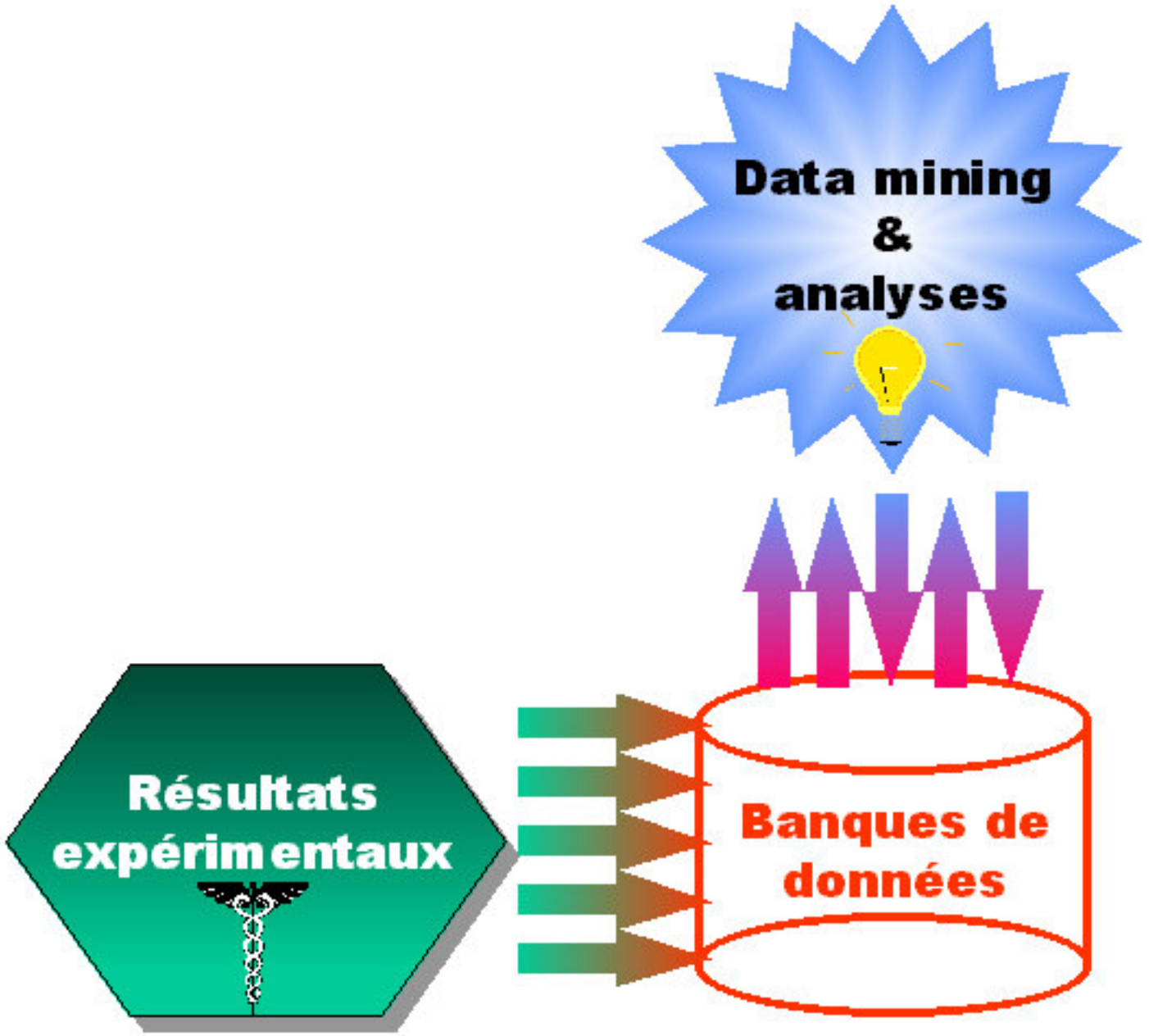
Découverte de modèles

- Description ou prédiction

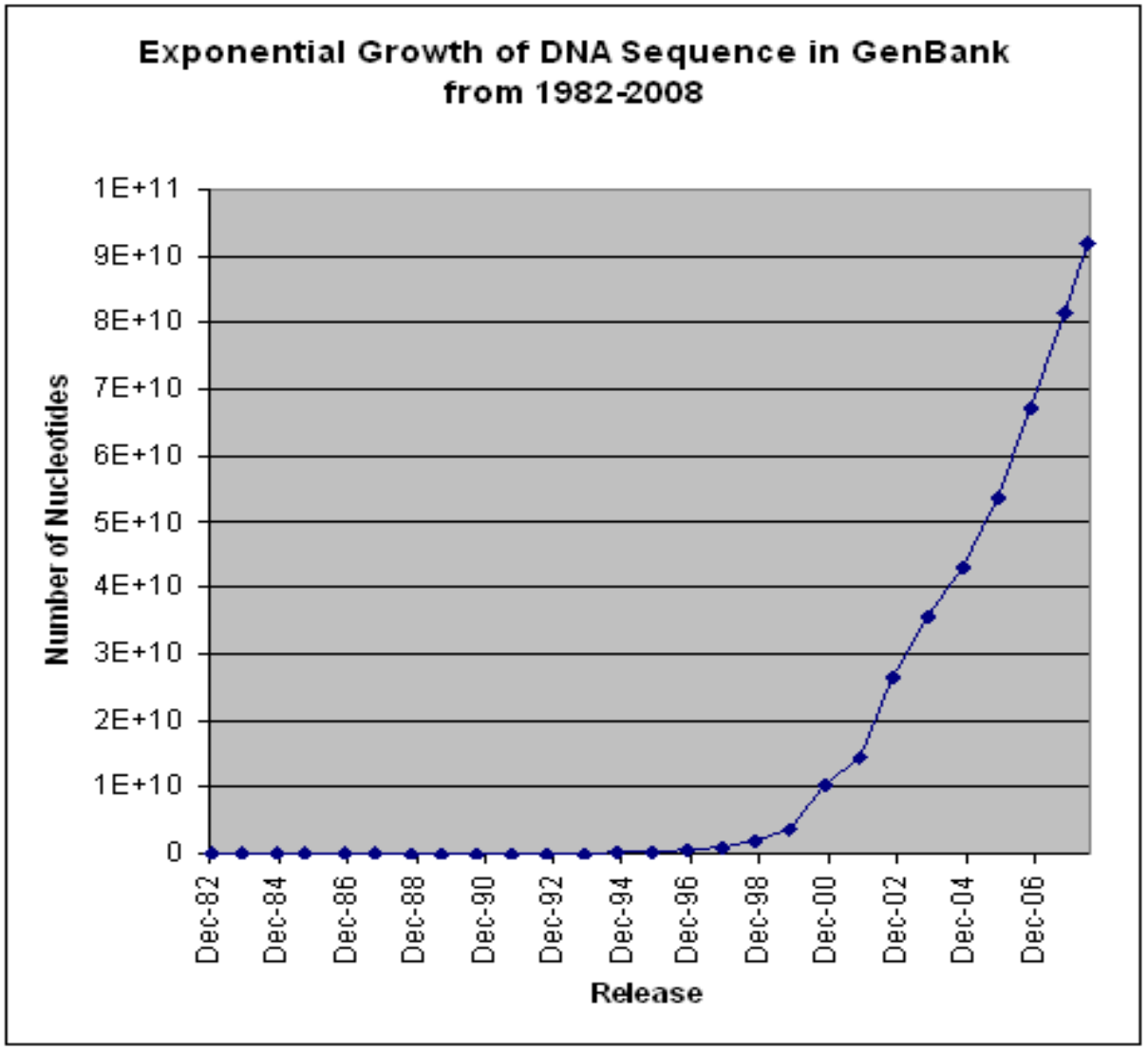


- Apprentissage sur la base
- Utilisation pour prédire le futur
- Exemple : régression linéaire Y = a X + B

Le matériel biologique



A ce compte là, il ne s'agit plus d'apprendre donc de reconnaître mais déjà de comprendre donc de structurer

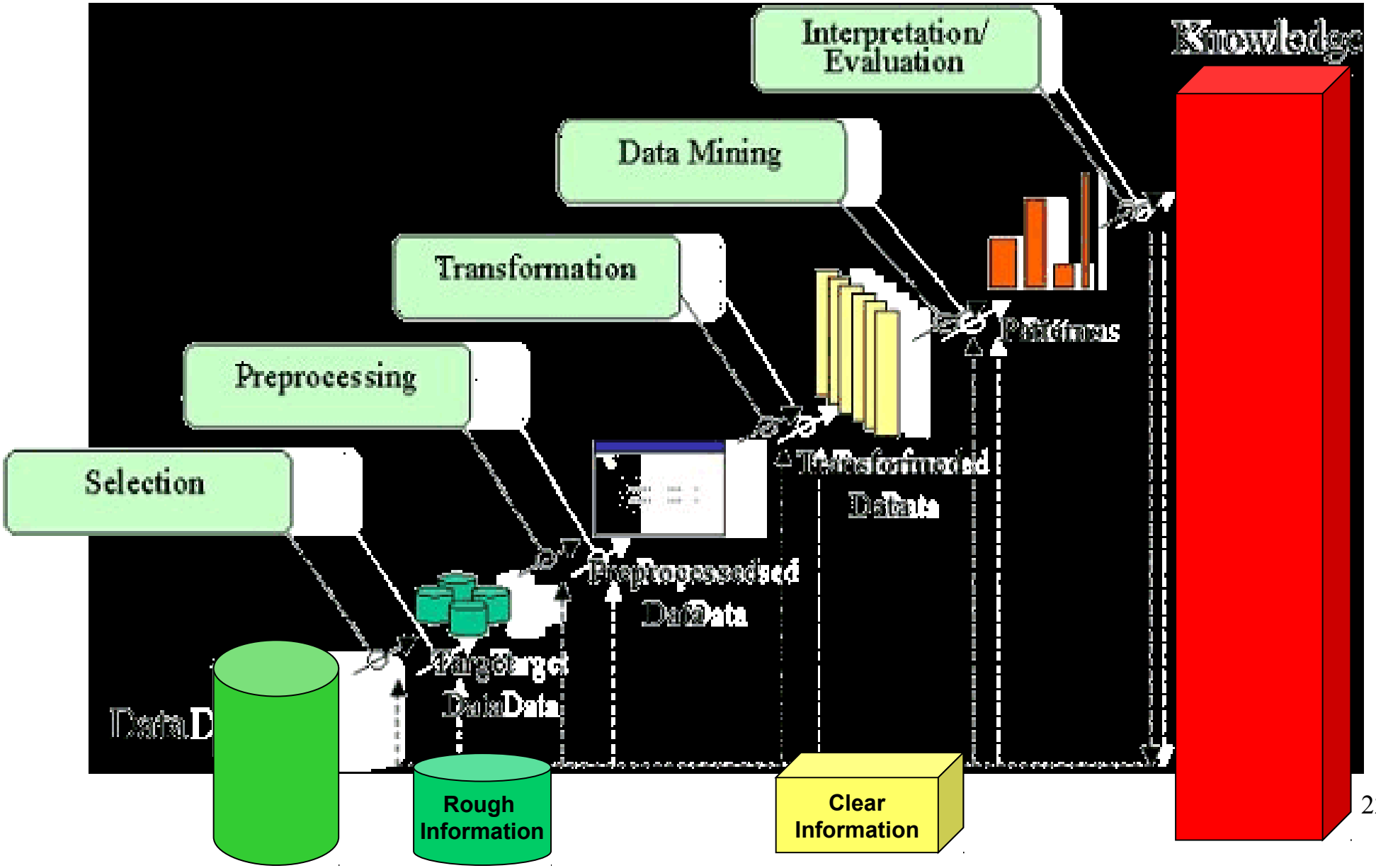


Comment analyser, visualiser, structurer des grandes masses de données réparties, hétérogènes

>cDNA inconnu

AATGCAAGTGCATGCATGCATGCATCGGATCGTACGGATTGCAGTTCGGATTCATAATAA
ATGCGTAAAAACAGTAGTTTCACTAGTTTCAAAAGTTGCATAATACTTGCTGTTCTTCTT
GTTTACCCTAACAGTATGGCTGTTTTCGCTGTTGCTGCTGACGGTATACCTTTCCCTTAC
CACGCTAAATACAGTAACGGTGCTATAAGTCCTCTTCACGTTACTCAAAGTAGTGGTAAC
AGTAGTGTTAAAGCTGAATGGGAACAATGGAAAAGTGCTCACATAACTAGTGACCTTAAC
GGTGCTGGTGGTTACAAATACGTTCAACGTGACATAAACGGTAACACTGACGGTGTTAGT
GAAGGTCTTGGTTACGGTCTTATAGCTACTGTTTGCTTCAACGGTGCTGACAGTAACGCT
CAAACTCTTTACGACGGTCTTTACAAATACGTTAAAAGTTTCGCTAGTGCTAACAACCCT
AACCTTATGGGTTGGCACATAAACAGTAGTAACAACATAACTGAAAAAGACGACGGTATA
GGTGCTGCTACTGACGCTGACGAAGACATAGCTGTTAGTCTTATACTTGCTCACAAAAAA
TGGGGTACTAGTGGTAAAATAAACTACCTTAAAGCTGCTCGTGACTACATAAACAAAAAC
ATATACGCTAAAATGGTTGAACCTAACAACTACACTCTTAAACTTGGTGACATGTGGGGT
GGTAACGACTTCAAAAACGCTACTCGTCCTAGTTACTTCGCTCCTGCTCACCTTCGTATA
TTCTACGCTTACACTGGTGACAAAGGTTGGATAAACGTTGCTAACAAACTTTACACTACT
GTTAACGAAGTTCGTAACAAATACGCTCCTAAAACTGGTCTTCTTCCTGACTGGTGCGCT
GCTAACGGTACTCCTGAAAGTGGTCAAAGTTTCGACTACGACTACGACGCTTGCCGTGTT
CAACTTCGTACTGCTATAGACTACAGTTGGTACGGTGACGCTCGTGCTGCTGCTCAAAGT
GACAAAATGAACAGTTTCATAGCTGCTGACACTGCTAAAAACCCTAGTAACATAAAAGAC
GGTTACACTCTTAACGGTAGTAAAATAAGTAGTAACCACAGTGCTAGTTTCTACAGTCCT
GCTGCTGCTGCTGCTATGACTGGTACTAACACTGCTTTCGCTAAATGGATAAACAGTGGT
TGGGACAAAGTTAAAGACAGTAAAAAATACGGTTACTACGGTGACAGTCTTAAAATGCTT
ATAATGCTTTACATAACTGGTAACTTCCCTAACCCTCTTAGTGACCTTAGTAGTCAACCT
AGTCCTGGTGACCTTAACGGTGACGGTGAAATAGACGAACTTGACATAGCTGCTCTTAAA
AAAGCTATACTTAAACAAAGTACTAGTAACATAAACCTTACTAACGCTGACATGAACCGT
GACGGTGCTATAGACGCTAGTGACTTCGCTATACTTAAAGTTTACCTTTAAT

Un système d'Extraction de Connaissances



Des techniques issus de l'IA et de la RF

Machine learning techniques such as :

- **Arbre de décision**
- **Règles d'association**
- **Réseaux de neurones**
- **Clustering**

Des systèmes combinant les technologies Réseaux et BD

- **SQL**
- **TCP/IP**
- **FTP**
- **Php / mySQL**

Des champs d'applications très diversifiés

- **Commerce – Economie**
- **Web Mining et Marketing**
- **Bio-informatique**
- **Médecine**

Principe global

L'importance pratique et industrielle des procédés d'analyse automatique et intelligente de données, textes, images, sons ou enregistrements électroniques est telle que beaucoup de recherches spécialisées se sont développées.

Nous cherchons ici à en donner une idée et à en dégager les points communs qui sont le propre de la méthodologie de la Fouille de Données.

C'est essentiellement dans la conception des processus de discrimination (ou d'affectation à diverses catégories) que l'on retrouve une méthodologie commune, à quelques variantes près.

En gros une telle fonctionnalité est constituée de plusieurs composantes, correspondant à plusieurs phases de traitement. On en distinguera essentiellement deux, les autres pouvant s'échelonner entre les deux extrêmes :

1. Le prétraitement
2. La découverte de catégories proprement dite

Un fait remarquable en FD est que chaque application fait appel à plusieurs techniques parmi celles présentées ici, avec une interrelation parfois surprenante où l'invention et le flair de l'ingénieurs sont rois.

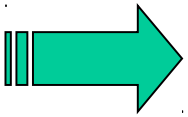
Des algorithmes

Une évolution plus qu'une révolution

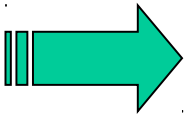
Un cocktail de techniques

Des algorithmes

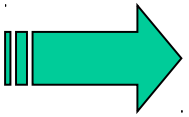
D’inspirations …



Mathématiques : stat. et AD

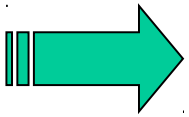


Calculatoires



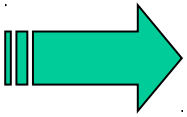
Biologiques

Des algorithmes



Calculatoires

« Clustering »
Arbres de décision
Règles d'association
Programmation dynamique



Biologiques

Réseaux de neurones
Algorithmes génétiques

Des algorithmes

Non Supervisés

Apprentissage *a priori* en mode Découverte

« *Clustering* »

Algorithmes génétiques

Règles d'association

Supervisés

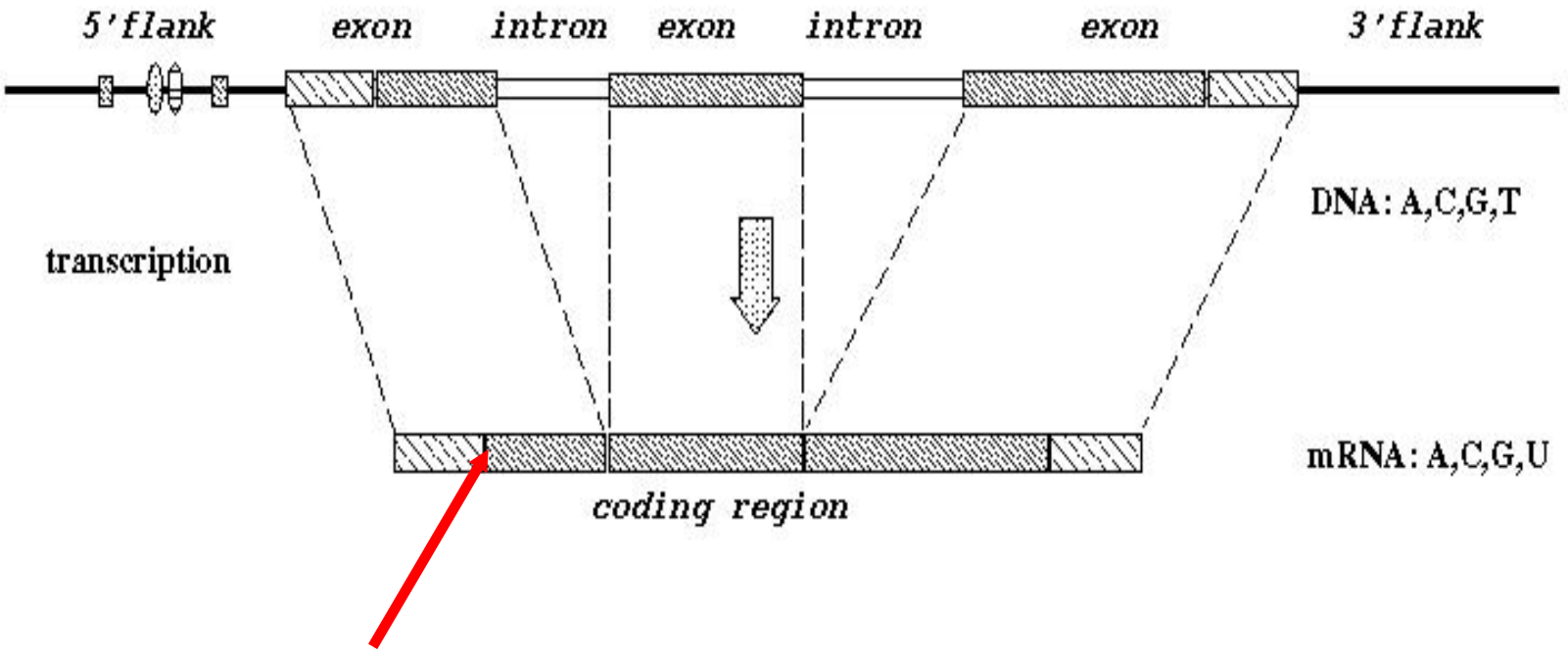
Apprentissage *a posteriori* en mode Reconnaissance - Prédiction

Réseaux de neurones

Arbres de décision

Programmation dynamique

TIS : Translation Initiation Site Recognition/Prediction



Un échantillon de cDNA

```
  299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
.....................................................              80
.............................iEEEEEEEEEEEEEEEEEEEEEEEEE              160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Pourquoi le second ATG est-il un TIS?

A partir de ARNm, ADNc et séquence ADN.

TIS : règle simple du premier AUG (ou ATG si ADN traité) chez les eucaryotes. Mais pas toujours. En plus, erreurs en particulier dans EST + processus biologique de la traduction pas complètement compris.

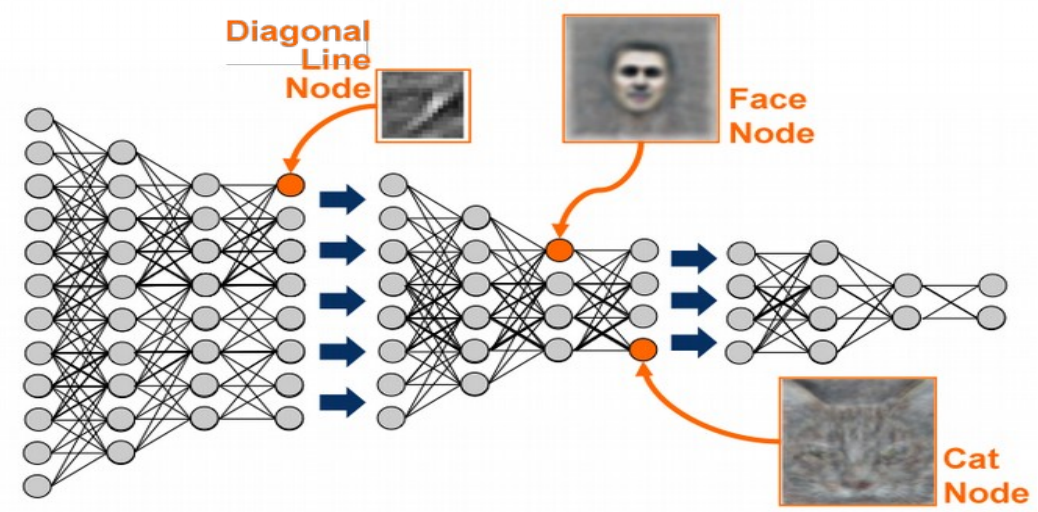
Fig. 1. An example annotated sequence from the dataset of Pedersen and Nielsen. The 4 occurrences of ATG are underlined. The second ATG is the TIS. The other 3 ATGs are non-TIS. The 100 nucleotides upstream of the TIS are marked by an overline. The 100 nucleotides downstream of the TIS are marked by a double overline. The ".", "i", and "E" are annotations indicating whether the corresponding nucleotide is upstream (.), TIS (i), or downstream (E).

Dans cette base, 13375 sites ATG, 25 % sont des vrais *TIS*.

On garde 100 nucléotides *upstream* et 100 nucléotides *dowstream* autour de chaque site potentiel pour essayer de prédire la structure contextuelle expliquant la classification Vrai *TIS* / Faux *TIS*.

Du perceptron, au réseau de neurones artificiels jusqu'aux SVM voir un cours de Machine Learning.

Idée de base :



1. Coder les nucléotides avec un codage binaire par exemple

A : 00
C : 01
G : 10
T : 11

2. Constituer un vecteur de codage numérique (binaire dans ce cas)

00110001010011000010011_TIS_00010100010001010001

3. Nourrir un réseau de neurones plus ou moins sophistiqué pour qu'il apprenne à partir d'exemples : Grand Renouveau actuel de cette modélisation : *Deep Learning*, Yann Le Cun, un Frenchie, à New York, Facebook lab.

→ Un peu boîte noire, mais efficace, et nécessite une interprétation des résultats *a posteriori.*

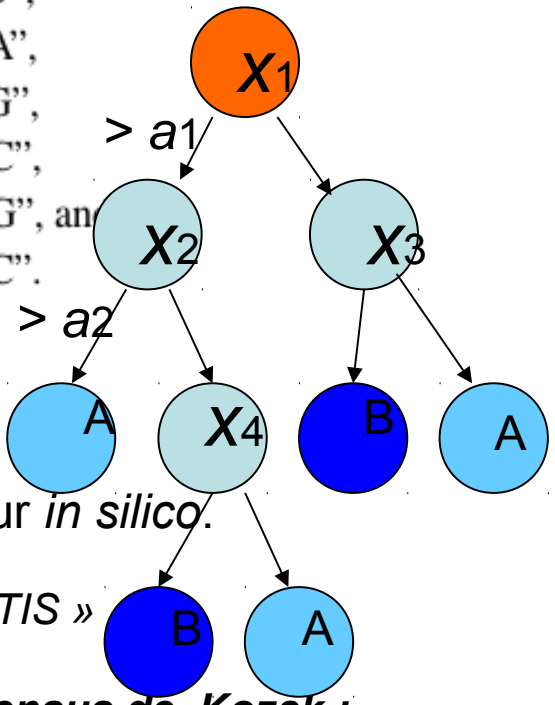
Des techniques plus **explicatives** par *k-gram* etc. et approches génération de caractéristiques, sélection de caractéristiques et intégration de caractéristiques.

Feature Generation : le codage finalement
k-gram : ici suite de k lettres
Si k=3, 4^K combinaisons de trois lettres (cas du codon)
Un « *feature* » peut être un k-gram et sa fréquence d'apparition dans le fragment *upstream* et/ou *downstream, in-frame* ou non etc.
On construit un vecteur de caractéristiques de k-gram et de leurs occurrences pour k variant entre 1 et 5 par exemple : à la fin on a un vecteur à 4 436 composantes par exemple.

Feature Selection : le filtrage intelligent (significatif du point de vue du signal, éventuellement biologique)
On ne garde que des caractéristiques sembl
caractéristiques de la classe *TIS*/ Non *TIS* pa
techniques de corrélations croisées par exer
Ici on ne retient que 9 « *features* » par exem

(1) "position–3",
(2) "in-frame upstream ATG",
(3) "in-frame downstream TAA",
(4) "in-frame downstream TAG",
(5) "in-frame downstream TGA",
(6) "in-frame downstream CTG",
(7) "in-frame downstream GAC",
(8) "in-frame downstream GAG", an
(9) "in-frame downstream GCC".

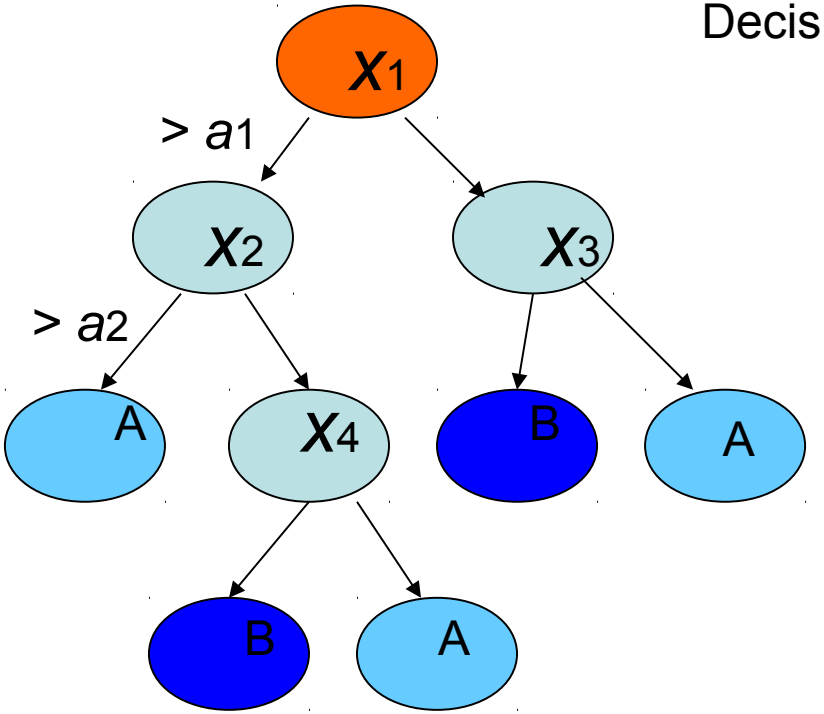
The Curse of Dimensionality In Classification



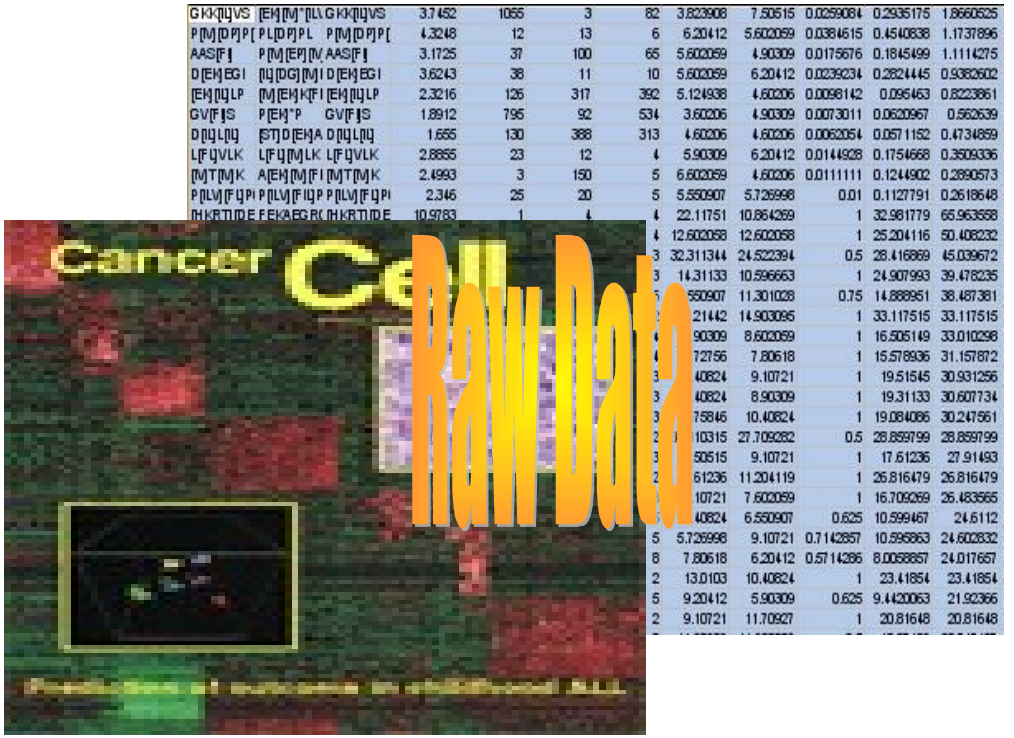
Feature integration : la décision / l'algorithme de classification/prédication
Sur les 9 caractéristiques précédents on entraîne un SVM (séparateur linéaire sophistiqué), un C4.5 (arbre de décision), un classificateur Naïf Bayésien etc. (voir Weka) et on obtient un prédicteur *in silico*.
« if up-ATG = Y and down-STOP >0 then prediction is false TIS »
« if up3-AorG = N and down-STOP <=0 and up3-AorG = Y, then prediction is true TIS »

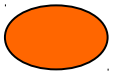
Reste comme toujours à interpréter biologiquement. (Séquence consensus de Kozak : GCC[AG]CCAUGG, îlots CpG (C.G) etc.)

Prognosis based on Gene Expression Profiling

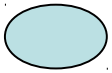


Decision Tree Based In-Silico Cancer Diagnosis

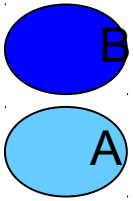




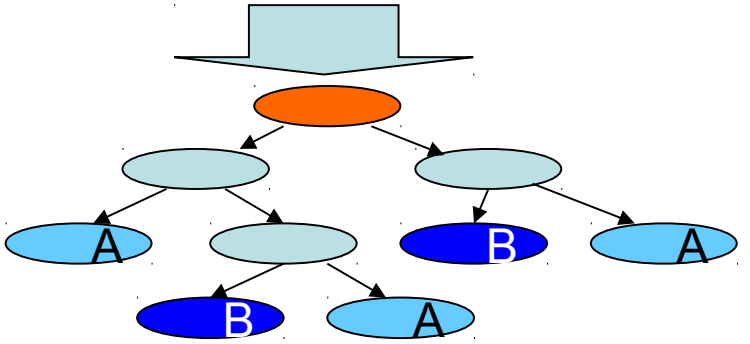
Root node

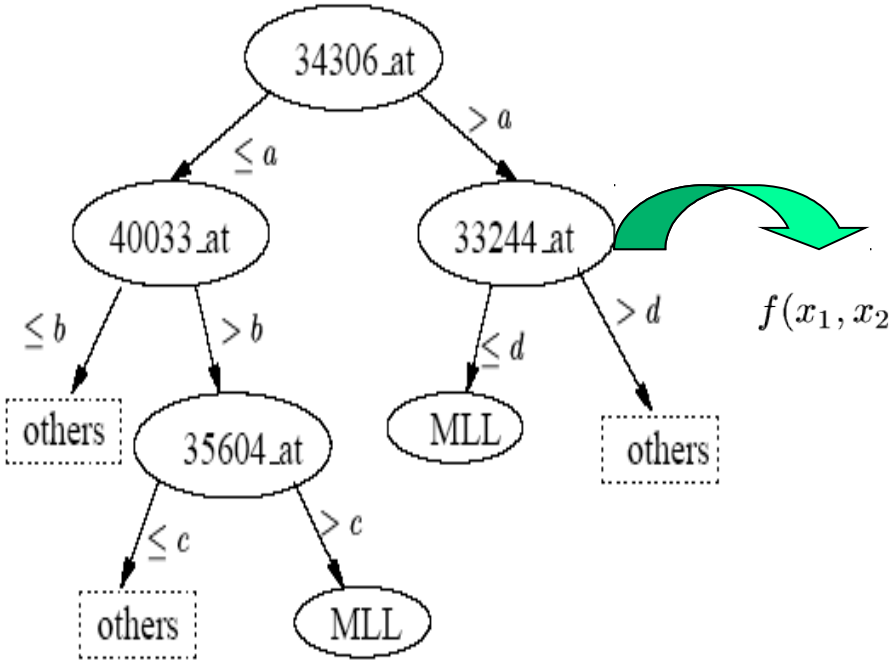


Internal nodes



Leaf nodes





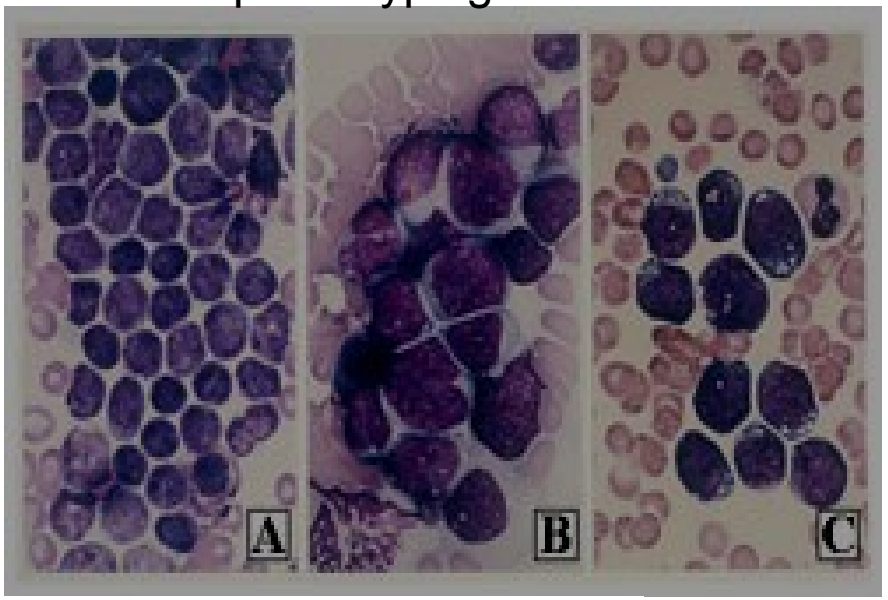
$$f(x_1, x_2, x_3, x_4) = \begin{cases} -1 & \text{if } x_1 \le a,\ x_2 \le b \\ -1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 \le c \\ 1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 > c \\ 1 & \text{if } x_1 > a,\ x_4 \le d \\ -1 & \text{if } x_1 > a,\ x_4 > d \end{cases}$$

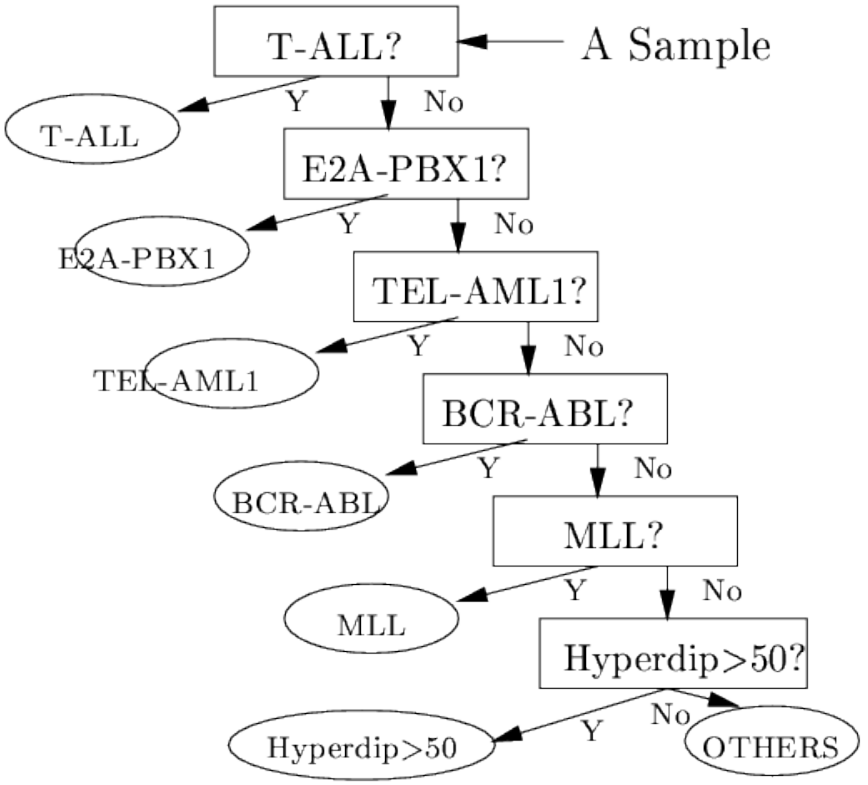
Given a test sample, at most 3 of the 4 genes' expression values are needed to make a decision!

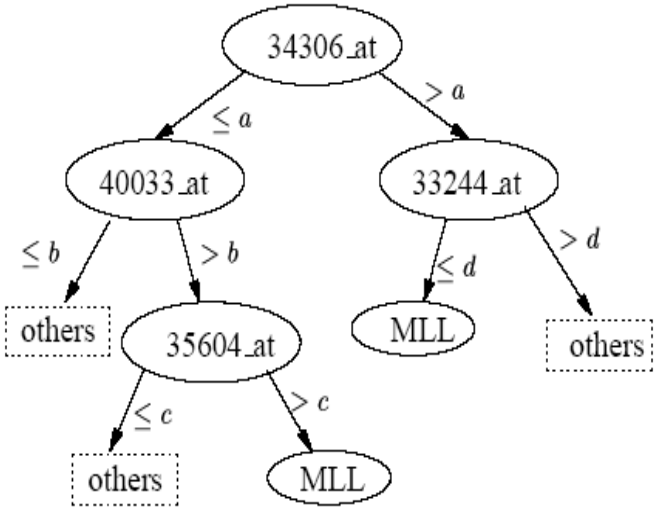
- Yeoh et al., *Cancer Cell* 1:133-143, 2002; Differentiating MLL subtype from other subtypes of childhood leukemia
  - Training data (14 MLL vs 201 others), Test data (6 MLL vs 106 others), Number of features: 12558

Diagnosis of Childhood Acute Lymphoblastic Leukemia (ALL) and Optimization of Risk-Benefit Ratio of Therapy

Immunophenotyping







$$f(x_1, x_2, x_3, x_4) = \begin{cases} -1 & \text{if } x_1 \le a,\ x_2 \le b \\ -1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 \le c \\ 1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 > c \\ 1 & \text{if } x_1 > a,\ x_4 \le d \\ -1 & \text{if } x_1 > a,\ x_4 > d \end{cases}$$

Yeoh et al., *Cancer Cell* 1:133-143, 2002;
Differentiating MLL subtype from other subtypes of childhood leukemia. Training data (14 MLL vs 201 others), Test data (6 MLL vs 106 others), Number of features: 12558

Given a test sample, at most 3 of the 4 genes' expression values are needed to make a decision!

A. G. Hatzigeorgiou. Translation initiation start prediction in human cDNAs with high accuracy. Bioinformatics, 18(2):343–350, 2002.

Précision globale de 94 %

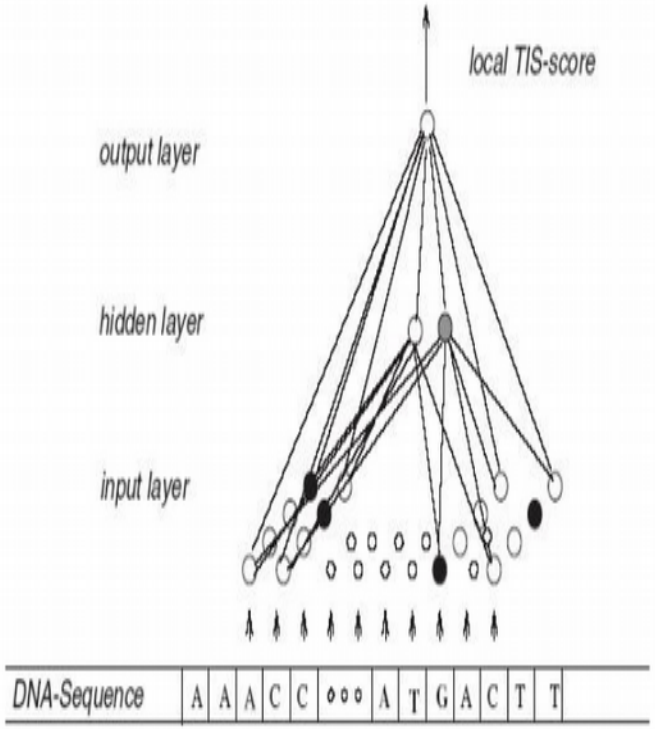


Fig. 4. The consensus ANN of DIANA-TIS. A window of 12 nucleotides is presented to the trained ANN. A high score at the output indicates a possible TIS. (*Image credit: Artemis Hatzigeorgiou.*)

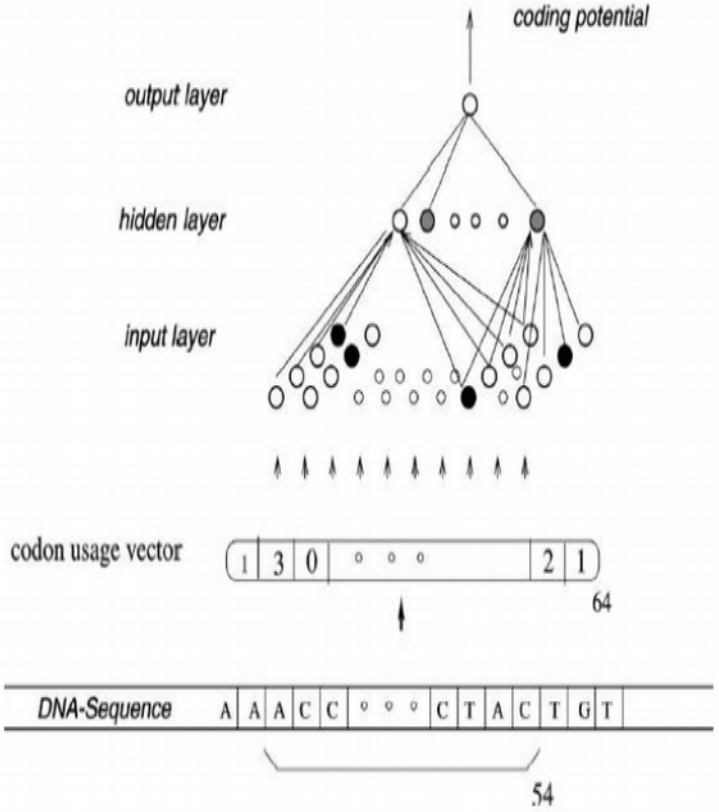
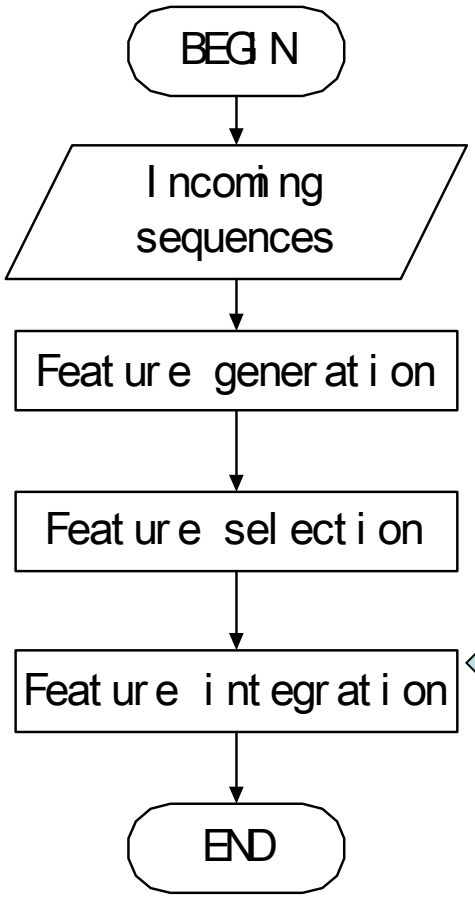
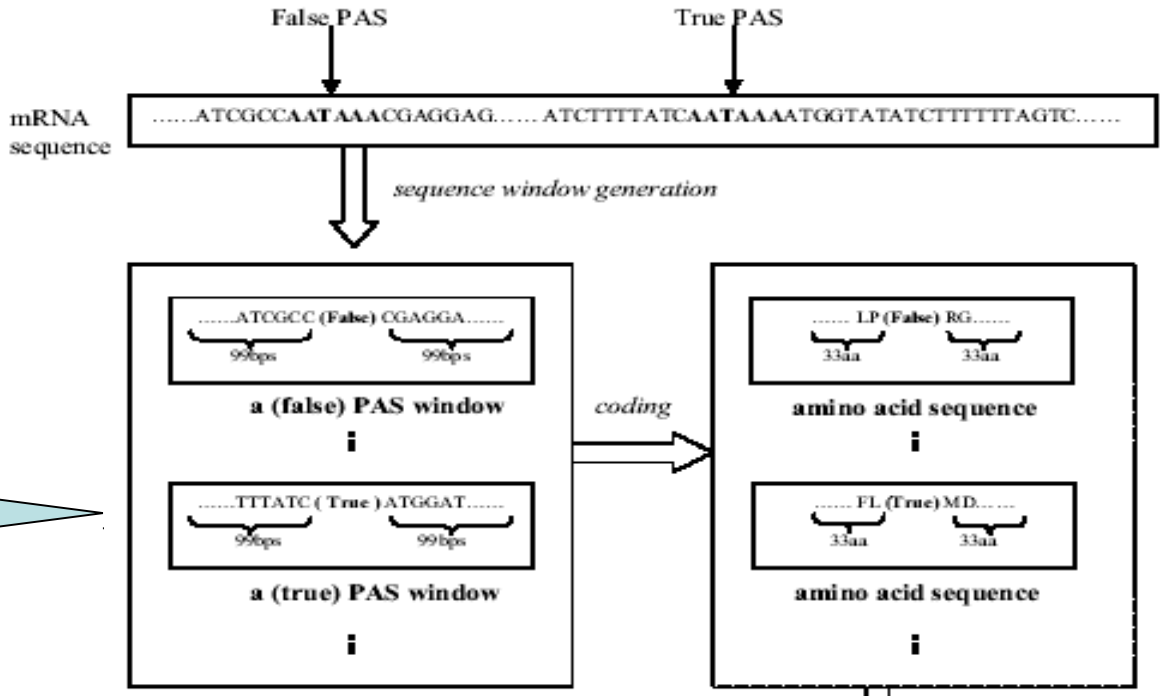


Fig. 5. The coding ANN of DIANA-TIS. A window of 54 nucleotides is presented to the trained ANN. A high score at the output indicates a coding nucleotide. (*Image credit: Artemis Hatzigeorgiou.*)

PAS Prediction





New feature space (total of 925 features + class label)			
42 1-gram amino acid patterns	882 2-gram amino acid patterns	1 bio-know-ledge pattern	class label
UP-A, UP-R, ….,UP-N, DOWN-A, DOWN-R, …., DOWN-N (numeric type)	UP-AA, UP-AR, …., UP-NN, DOWN-AA, DOWN-AR , …, DOWN-NN (numeric type)	UP-T-number (numeric type)	True, False
Frequency as values			
1, 3, 5, 0, 4, …	6, 2, 7, 0, 5, …	10,	False
6, 5, 7, 9, 0, …	2, 0, 3, 10, 0, …	50,	True

SVM in *Weka*