Eléments de statistique

L. Wehenkel

Cours du 9/12/2014
Méthodes multivariées; applications & recherche

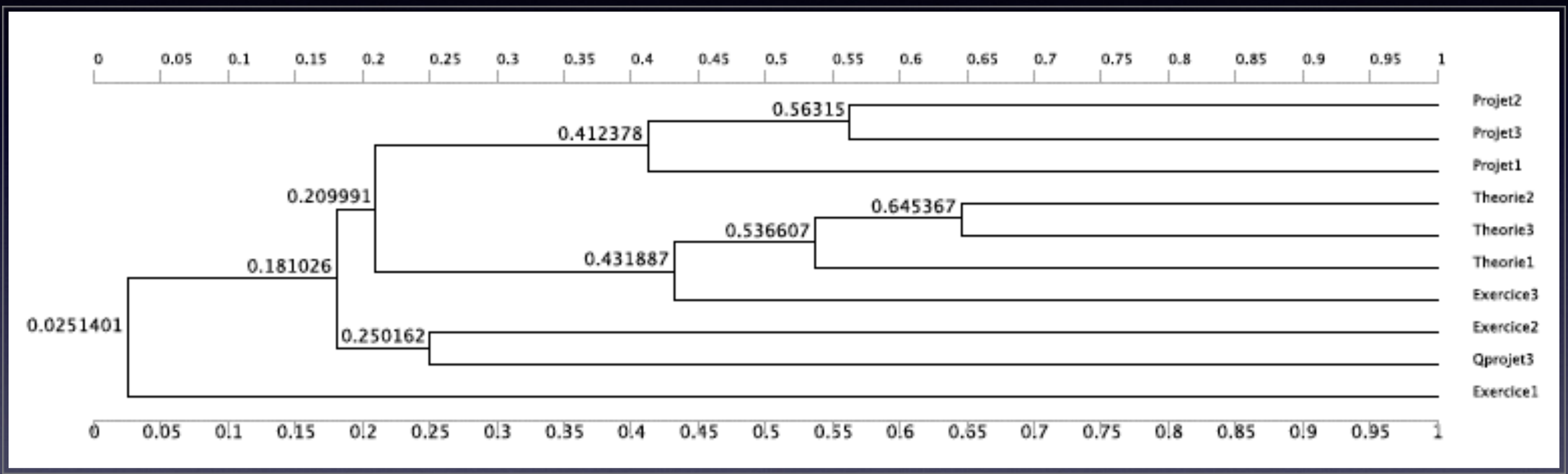
Quelques méthodes d'analyse multivariée

- NB: illustration sur base de la BD 'résultats de probas en juin 2013'
- Méthodes illustrées (explications au tableau)
  - analyse de corrélations (dendrogramme)
  - analyse de similarités (k-means)
  - apprentissage supervisé (arbre de décision)
  - importance de variables (t-test, Extra-Trees)

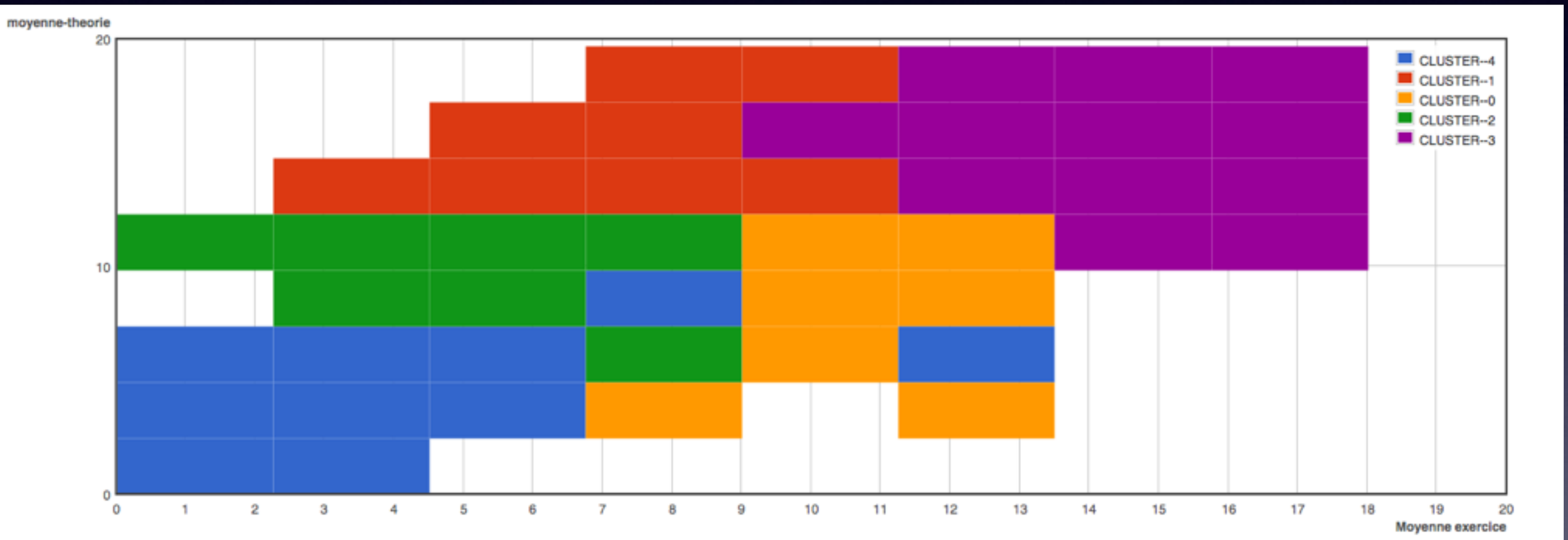
La BD … (pour rappel)

Projet 1	Projet 2	Projet 3	Q projet 3	Théorie 1	Théorie 2	Théorie 3	Exercice 1	Exercice 2	Exercice 3
0	0	0	1	2	3	1	1	0	0
7	8	19	0	6	6	2	3	8	3
14	15	19	16	12	17	14	6	5	17
11	2	0	0	6	11	5	6	14	11
18	19	19	5	16	10	10	2	18	15
18	9	17	7	18	18	5	12	4	6
19	20	17	12	16	11	16	8	16	15
16	12	20	6	18	12	17	16	15	13
19	10	14	11	15	7	6	5	19	1
14	18	19	0	4	2	5	8	3	2
0	0	0	0	5	8	1	11	0	0
0	12	20	3	20	20	16	5	18	0
17	18	14	7	20	15	14	12	12	14
11	14	16	0	18	14	14	4	9	5
20	19	19	14	18	13	17	7	12	13
20	16	20	20	16	13	15	17	12	7
20	14	20	13	13	12	4	0	14	15
13	16	16	1	12	12	10	4	3	1
16	16	19	7	10	9	2	18	0	7
20	14	0	0	20	5	0	6	15	0

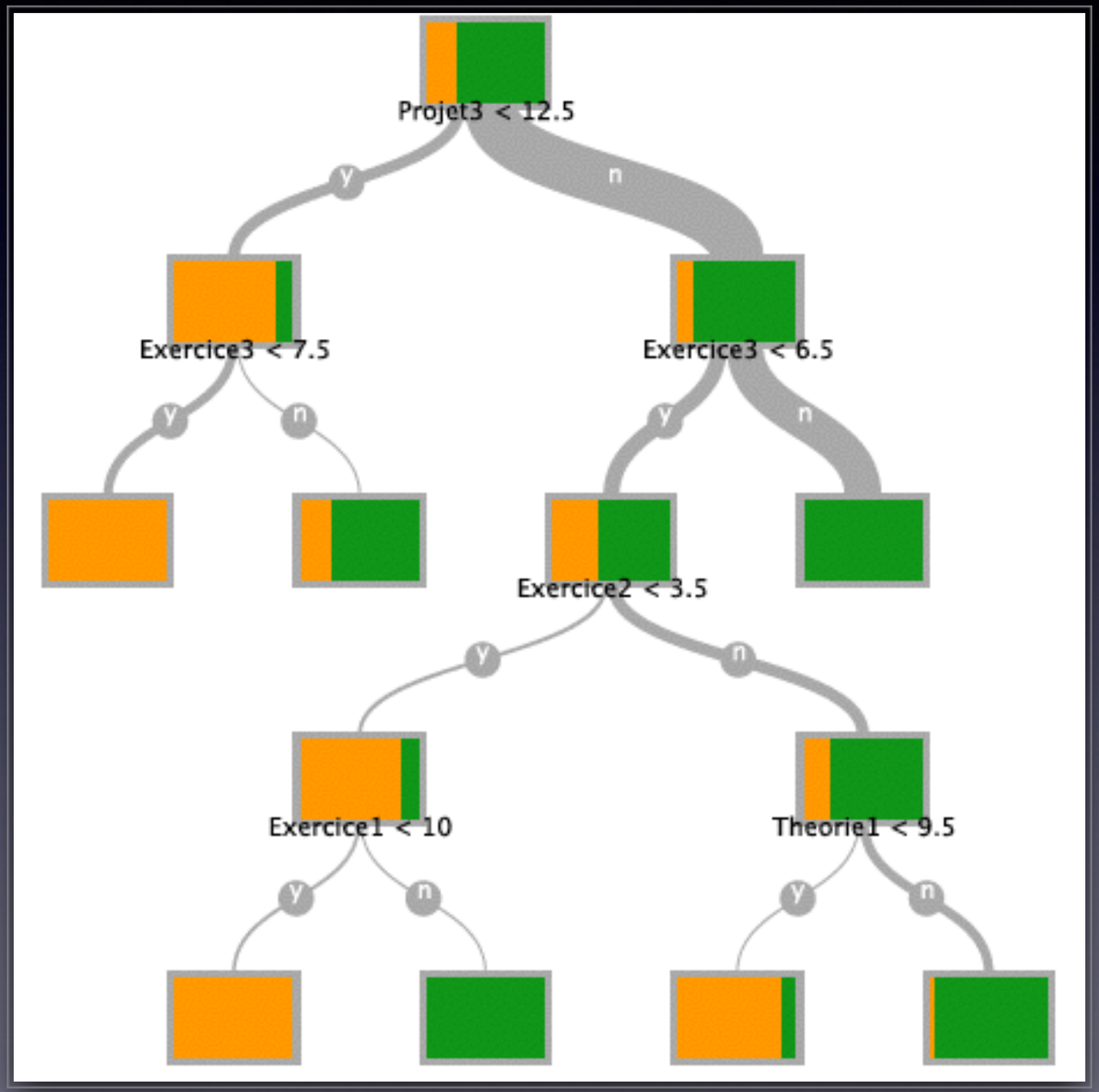
Un dendrogramme



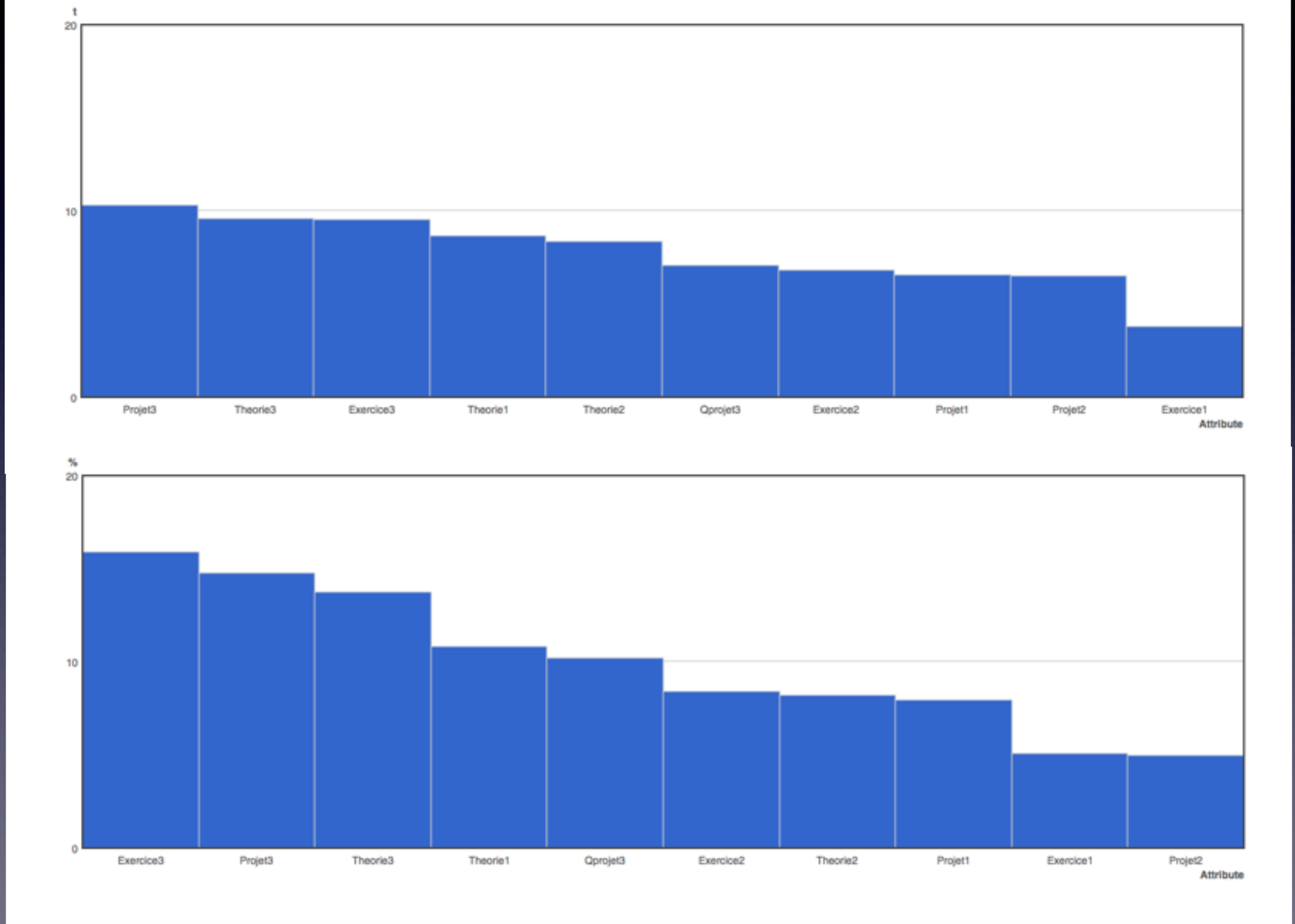
Classification automatique



Un arbre de décision



Importances des variables



Recherches en Systèmes et Modélisation


Ceci n'est pas une pipe.

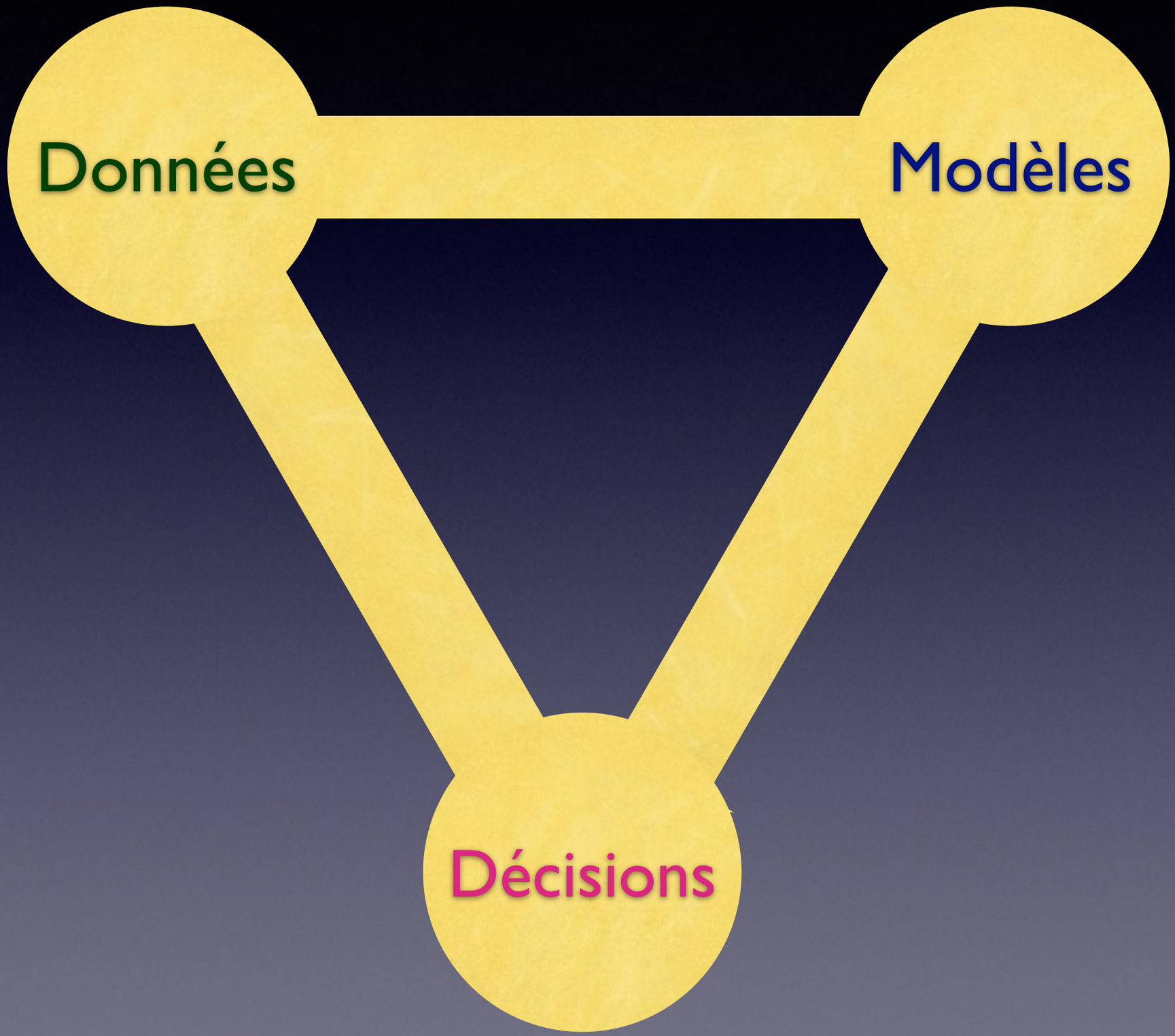
$$\begin{cases} \dot{x} = f(x, y, \lambda) \\ 0 = g(x, y, \lambda) \end{cases}$$

$$z = h_1(x, y, \lambda)$$

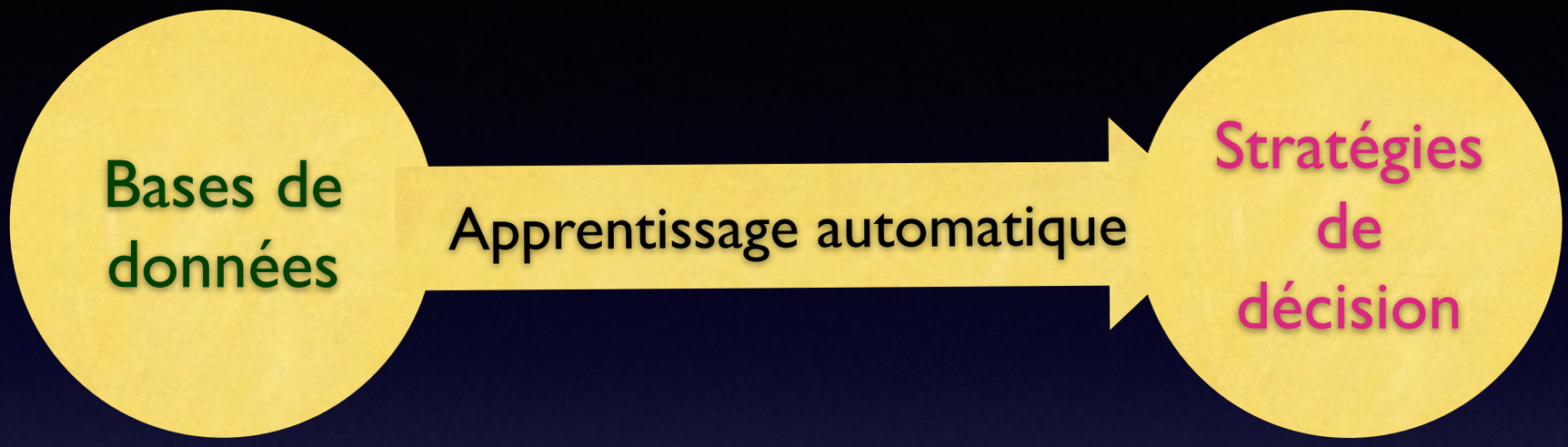
$$s = h_2(x, y, \lambda)$$

Ceci n'est pas un système électrique

Terrain de recherches

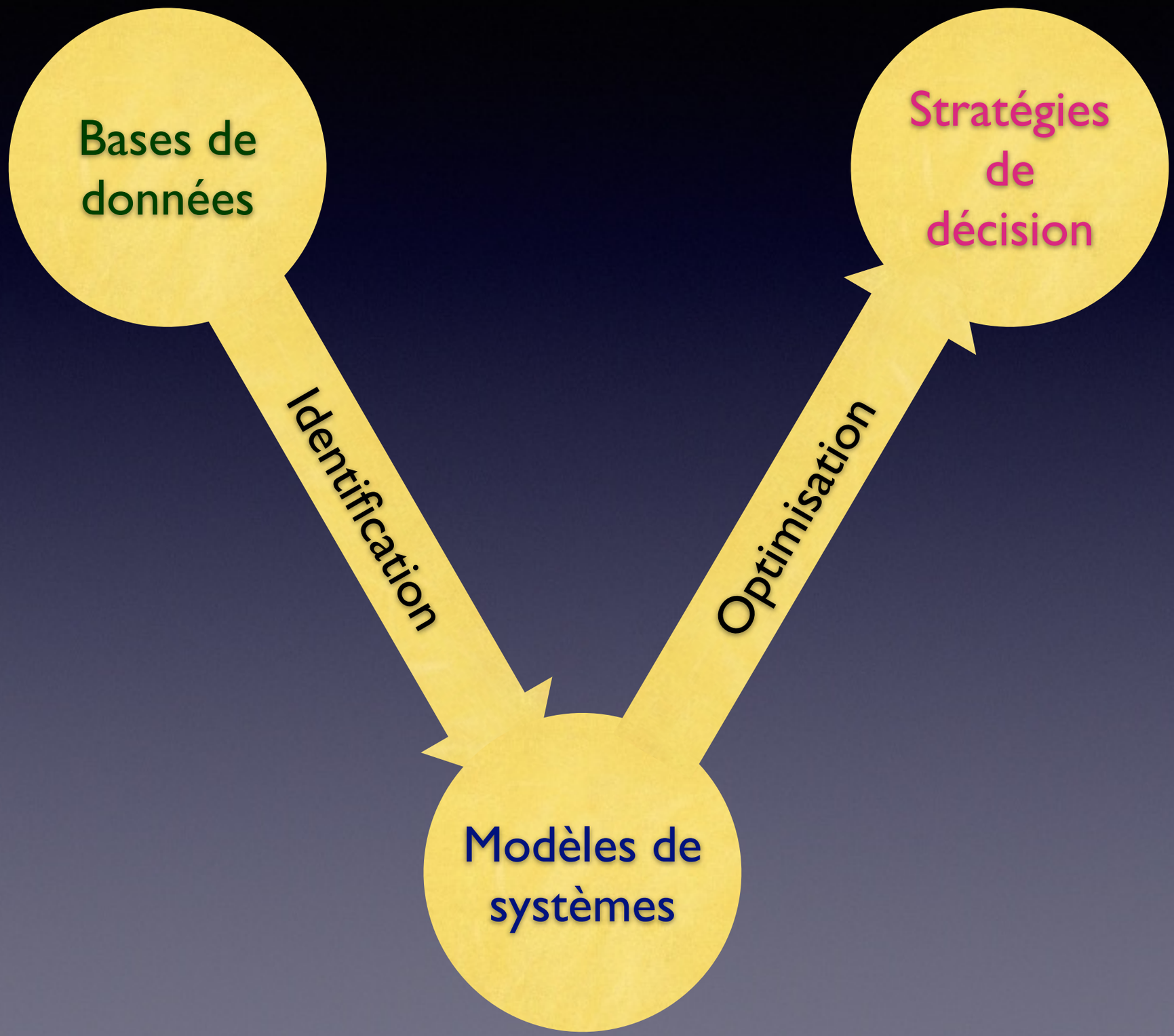


Exploitation de données

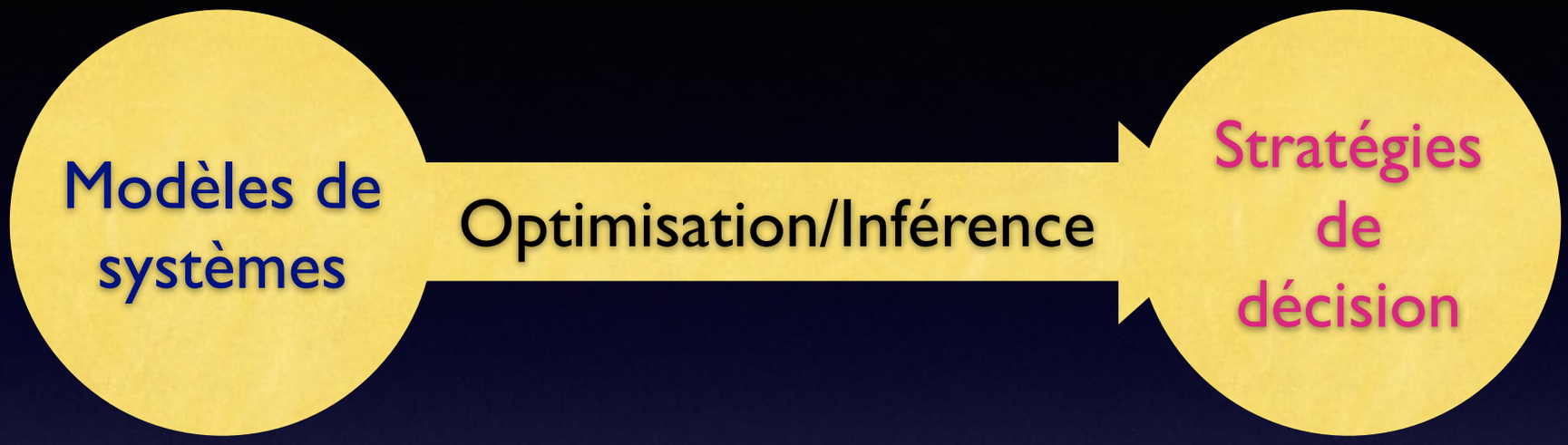


Modèles de systèmes

Exploitation de données

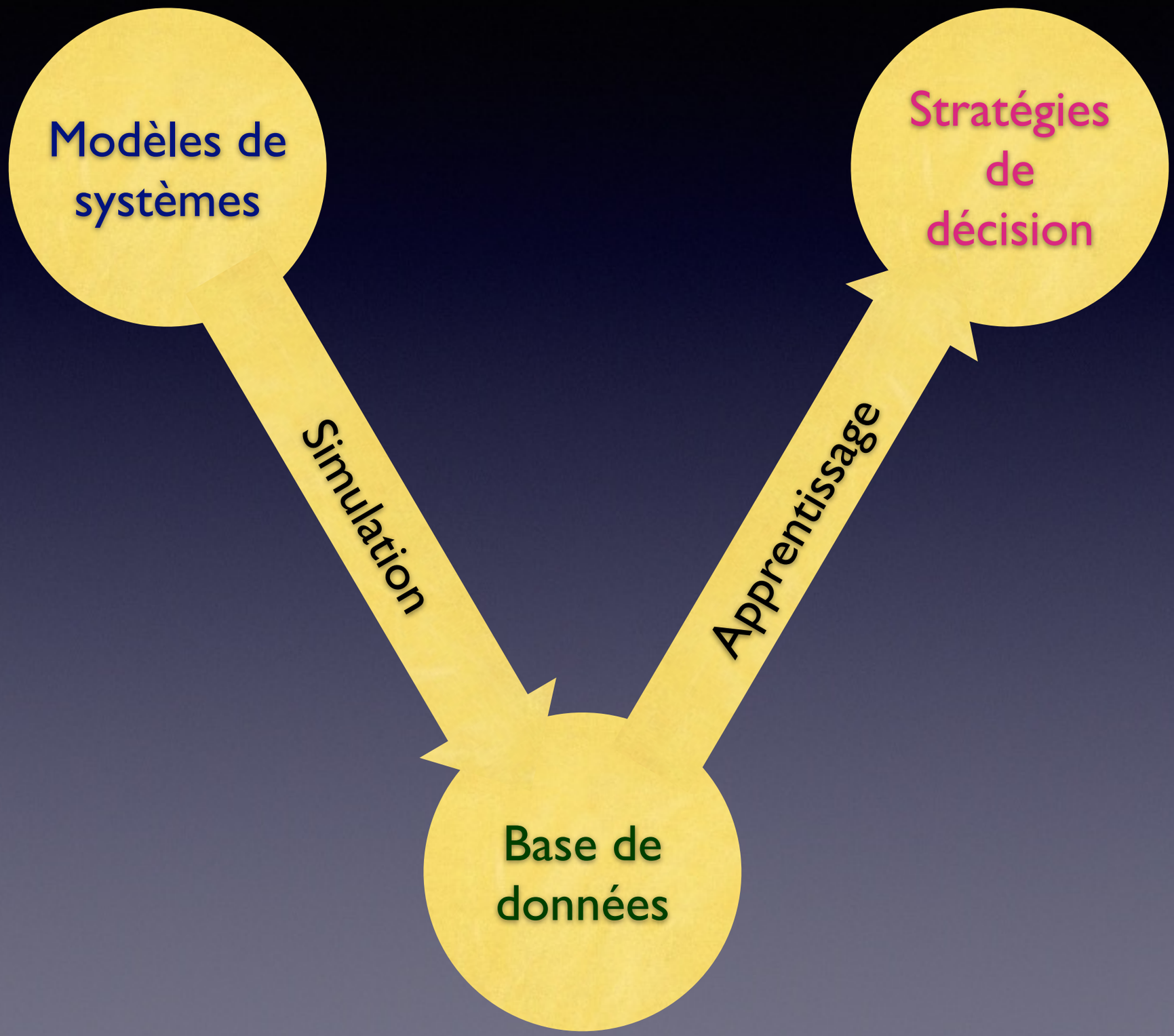


Exploitation de modèles



Base de données

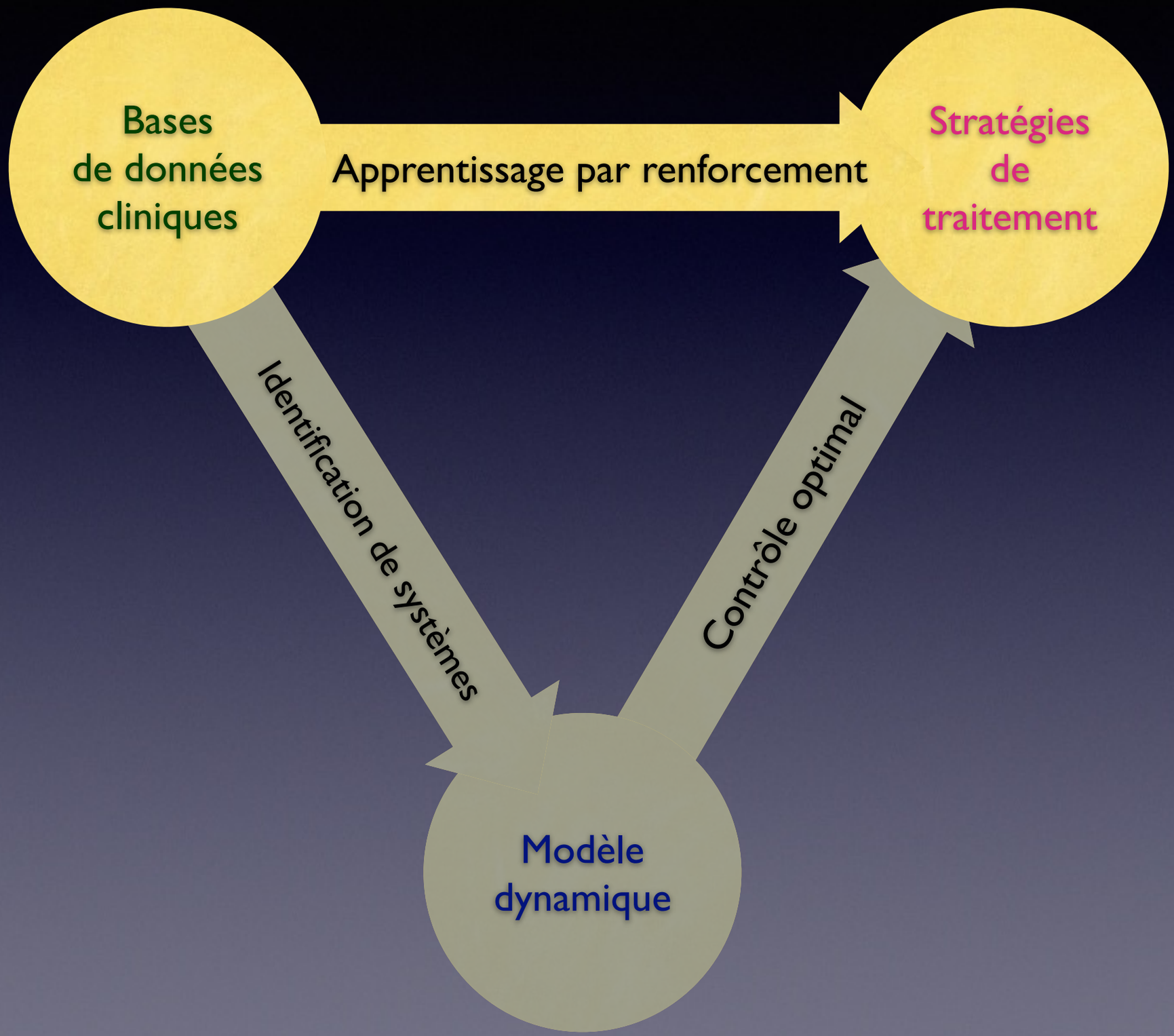
Exploitation de modèles



Exemple de recherche en cours (I)
Amélioration de stratégies de traitement médical des maladies chroniques

- Objectif:
  - des stratégies de traitement médical plus efficaces et mieux adaptées au patient (p.ex. STI pour infection HIV)
- Point de départ:
  - données d'enregistrements cliniques de suivi de patients en traitement

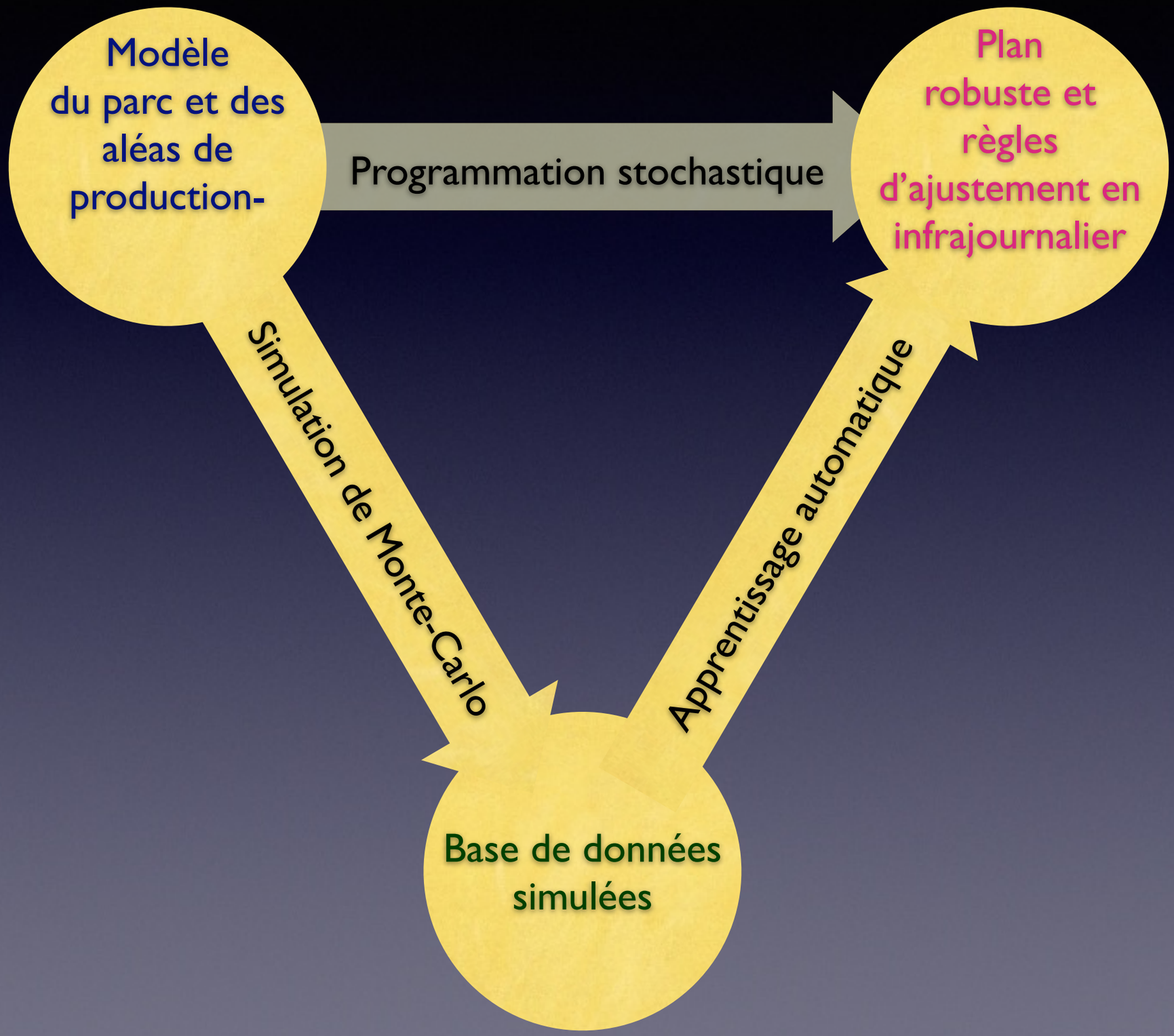
Construction de stratégies de traitement médical



Exemple de recherche en cours (2)
Construction de règles d'ajustement pour la production d'électricité

- Objectif:
  - Construction d'un plan de production robuste et de règles d'ajustement (p.ex. pour l'infra-journalier)
- Point de départ:
  - Modèle détaillé du parc de production et des aléas de consommation et de pertes de groupes

Optimisation de la gestion d'un parc de production



Recherches en méthodes

- Apprentissage automatique
  - supervisé, non-supervisé, par renforcement
  - images, séquences, séries temporelles, graphes
- Optimisation
  - convexe, discrète, stochastique
- Simulation
  - temporelle, échantillonnage de Monte Carlo

Quelques exemples de projets aboutis

Amélioration des systèmes de protection des systèmes électriques

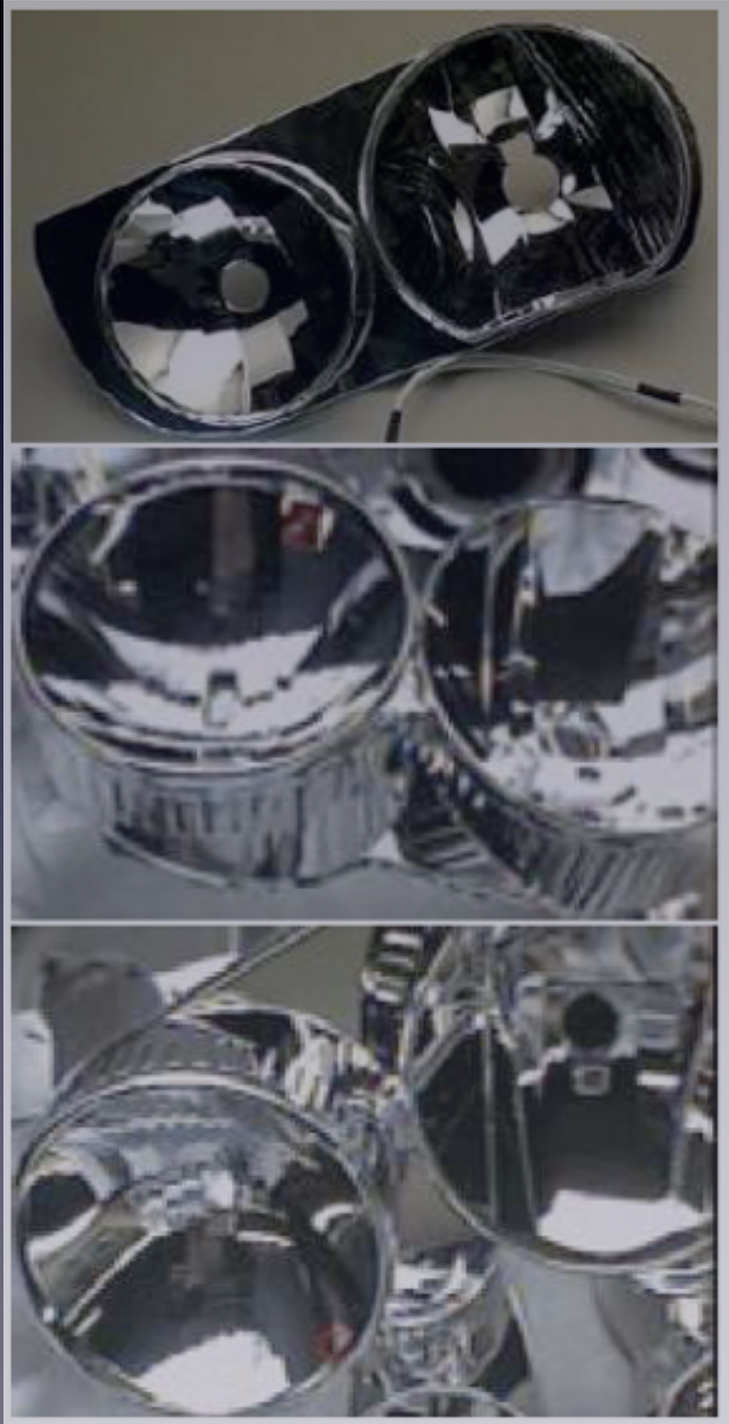
Problem

- Improve emergency control scheme
  - Churchill-Falls power plant
- Reduce probability of blackout

Approach

- 10,000 real-time snapshots sampled (several years)
- Massive time-domain simulations
- Automatically learn decision rules to determine optimal amount of generation and load to trip
- Implement rules in real-time
- New rules enhance security

Contrôle de qualité par système de vision par ordinateur



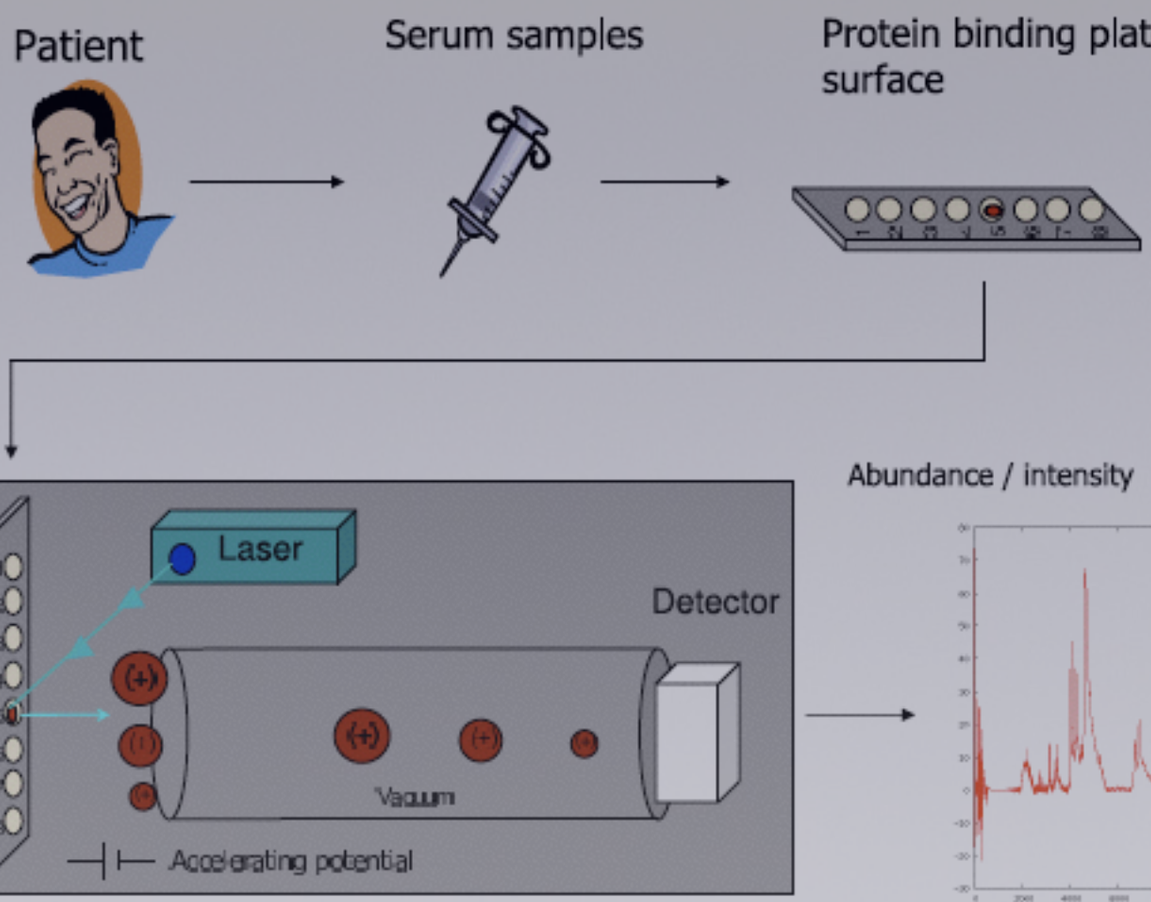
Problem

- Car light reflector manufacturing
- Quality control of aesthetic defects

Approach

- Robotics (handling of reflectors)
- Computer vision (defect detection)
- Extraction of images of defects (10000×300)
- Expert classification into 15 classes
- Build classifiers by automatic learning
- Integration into automatic QC system

Aide au diagnostic médical par analyse de données d'expressions de protéines



SELDI-TOF MS:

Surface Enhanced Laser Desorption/ Ionisation Time of Flight Mass Spectrometry

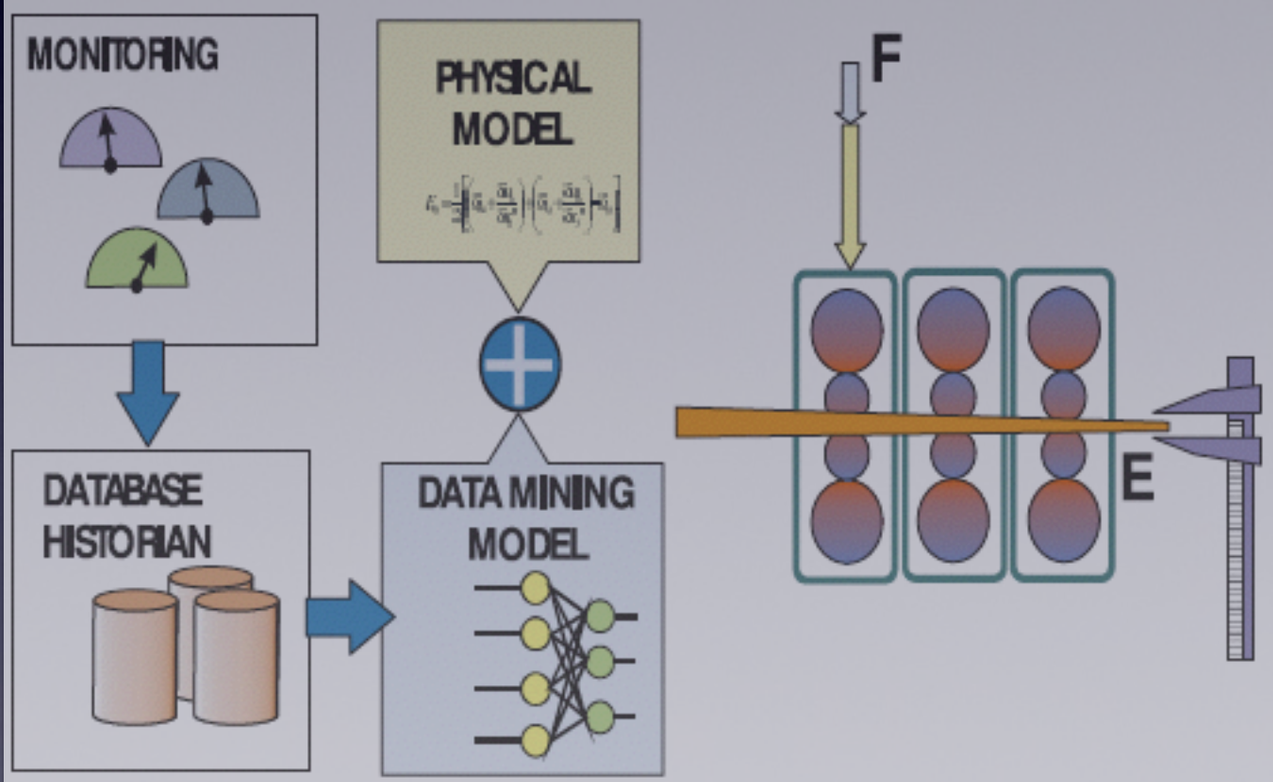
Problem

- Diagnosis of Rhumatoid Arthritis and other inflammatory diseases

Approach [GFd+04]

- Proteomic analysis of serum samples
- Automatic learning to
  - identify biomarkers (protein fragments) specific of disease
  - derive classifier for medical diagnosis

Amélioration du contrôle-commande d'un laminoir de sidérurgie



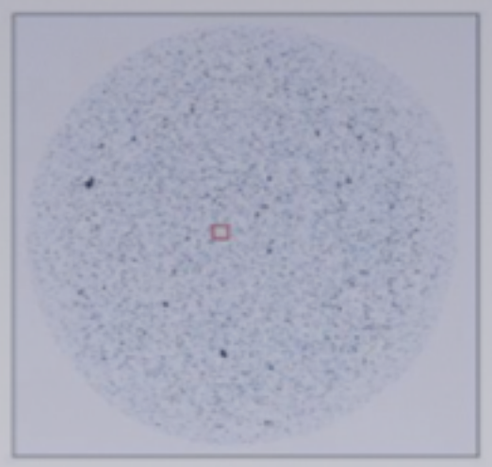
Problem

- Pre-setting of steel-mill controller
- Improve friction force model

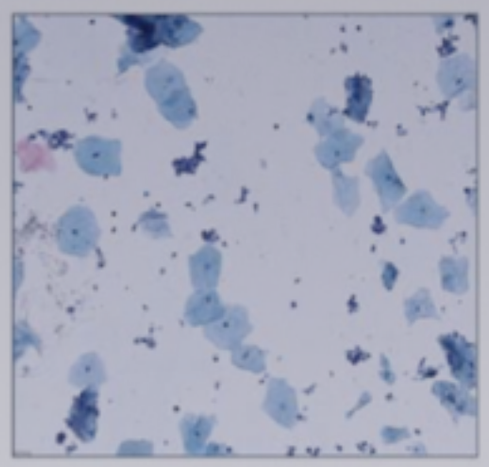
Approach

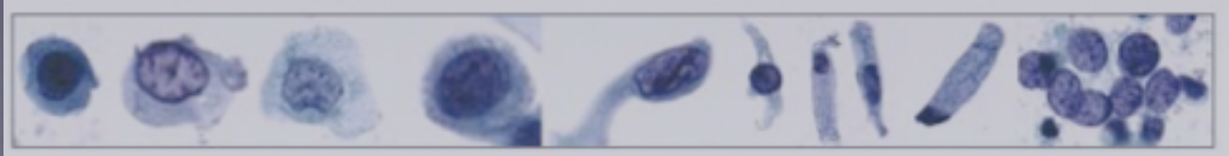
- Collect data from process measurements
- Determine error of physical model
- Automatically learn blackbox model of prediction error
- Combine physical and blackbox model to predict friction forces
- Adaptive pre-setting reduces waste

Aide au diagnostic médical par analyse automatique d'images cytologiques



20 X





Problem

- Early diagnosis of disease based on cytological tests
- Improve detection of rare abnormal cells among tens of thousands of normal cells

Approach

- Pathologists collect a database of normal and abnormal cells
- Automatically learn a cell classification model
- Scan whole-slide digital images (40000×40000 pixels) and rank most suspicious cells for expert review

Le Big Data

- Dans de nombreux domaines, les volumes de données collectées deviennent énormes
- Pour exploiter ces données, de nouvelles disciplines sont développées, à la convergence de l'informatique et des statistiques
- **Data Science**

Some ‘Big Data’ success stories in other fields

NB: I have collected this information from the WEB, and thus have to admit that I can not certify that all this is exactly true.

My purpose here is merely to illustrate some possible uses, sometimes a bit surprising, of Big Data Analytics.

WALMART sells more Strawberry Pop-Tarts when there is a Hurricane !

Just one example of weather related modeling of ‘customer behavior’

Carried out by correlation analysis of a lot of data internal to company and additional data external to it (here the weather data)

Can already be exploited, even if we do not really understand what causes this apparently strange customer behavior

Grocery store optimizes wine selling !

- Lots of data tell us that many consumers tend to choose the ‘average product’, when they have a choice between
  - Cheap
  - Average
  - Expensive

- So, if you want to increase sales of wine in the range of 15-20€ per bottle, it is of interest to place next to them a choice of wines in the range of 40-50€ per bottle
- Even if nobody is going to buy these expensive bottles, putting them in the shop will certainly increase your sales of the type of wine that you actually want to sell !

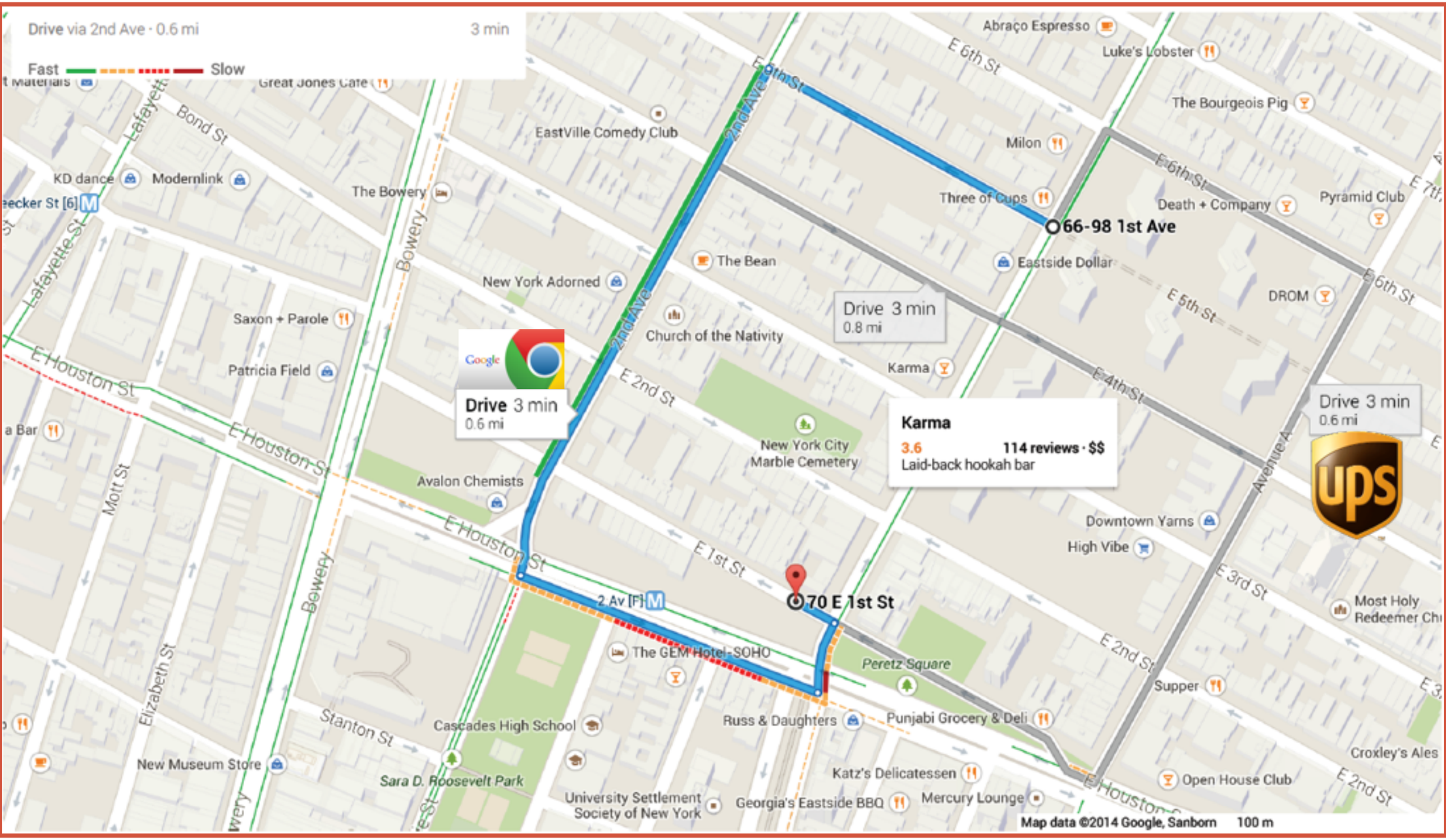
Progress in predicting flu outbreak !

- Centers for Disease Control and Prevention
  - Long established practice for predicting flu outbreak, based on scientific expertise, deep models and specific data collection
  - Google tries out to make such predictions based on search patterns of their users
  - Success story says that Google can predict flu outbreak about 2 weeks before the CDC

Cell-phone payment histories provide basis for bank credit allocation !

- Cell phone provider in India has hundreds of millions of customers
- They found out that the payment histories of their customers can be used for bank credit allocation
- They sold the credit scoring information to Indian banks

UPS ORION beats Google Maps !



Exegesis

- Success is often the result of using the right combination of various heterogeneous (internal + external) data sources
- Big Data means exploitation of very large datasets, so that observed correlations can often be accepted as being facts
- Sometimes, the data analytics yield totally unexpected findings, which turn often out to be of great value.
- Proper algorithms and IT have to be put in place, in order to maintain the information and enable its exploitation in the business of the company

- NB: Correlation analysis must often be combined with human judgment and/or experiments, to yield useful causal models
  - Indeed, “A correlated with B” can mean
    - “A causes B”
    - “B causes A”
    - “Both A and B have a common cause C”