Redescription en analyse de données : exemples variés

François Rioult

CNRS UMR6072 GREYC - Normandy University, Caen, France

19 décembre 2019

Science des données : Analyser pour Décider

La science des données :

1. Analyser :
   - organiser / discrétiser un monde **continu**
   - extraire de la connaissance
2. **pour**
3. Décider :
   - **valoriser** la connaissance
   - prendre des décisions **symboliques**

Décision : susciter l'attention

Susciter l'attention :

- révéler
- recommandation
- générer de l'**engagement**

Ce qu'on sait faire :

- (trans-)action → engagement
- séquence → habitudes
- graphes → relations, recommandation
- texte → polarité, compréhension
- image, son → reconnaissance, description

Variété des données vs. uniformité de la décision

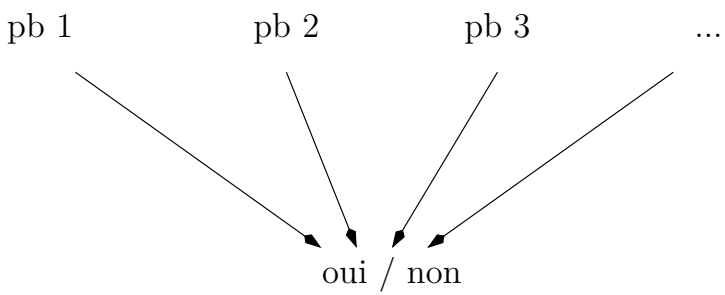
- Variété des données : les traces humaines
  - (trans-)action → engagement
  - séquence, trajectoire → habitudes
  - graphes → relations, recommandation
  - texte → polarité, compréhension
  - image, son → reconnaissance, description
- Uniformité de la décision

Uniformité en décision

Classification supervisée **ou** Classification non supervisée

Déclinaison pour les motifs :

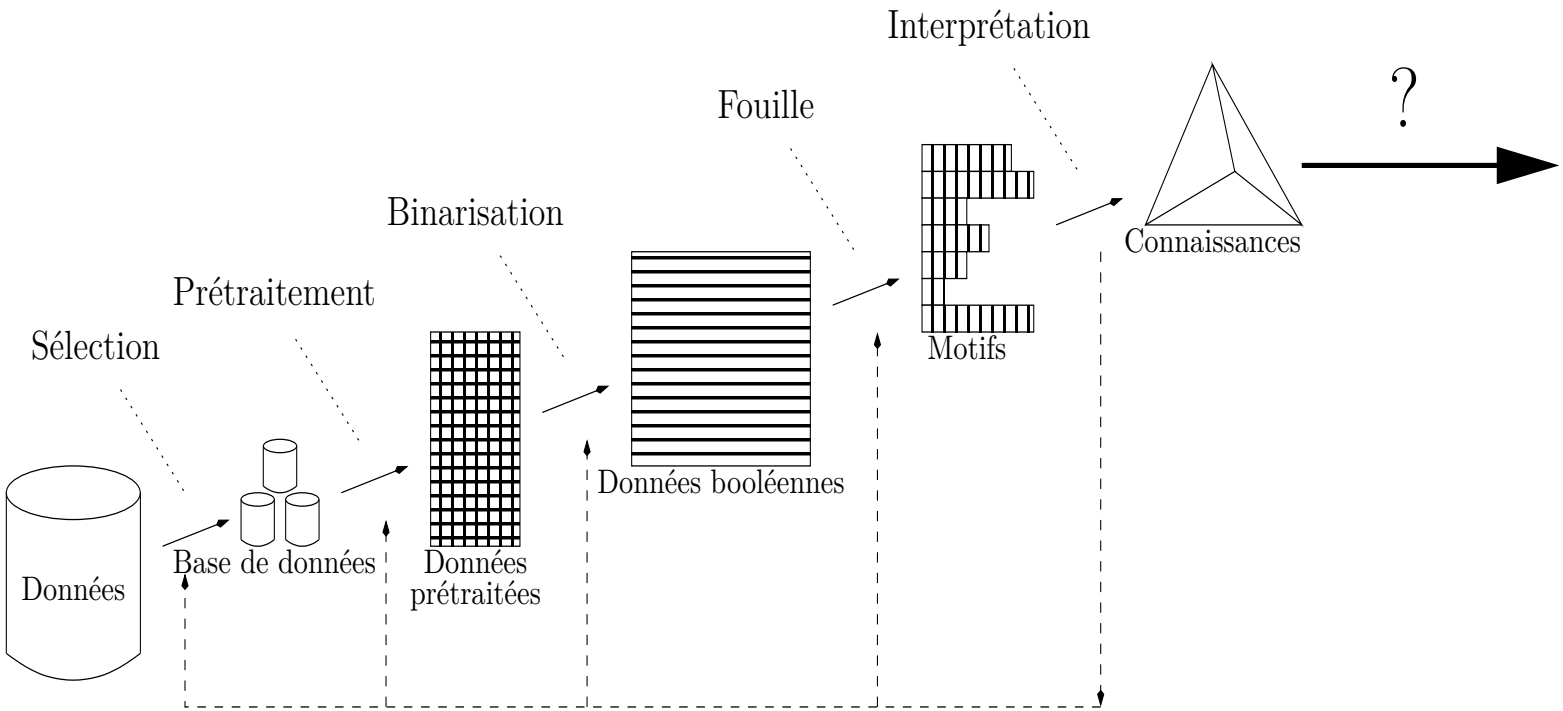
- à base de règles
- à base d'émergents
- à base de fermés
- (early) classification de séquence
- ...



➥ Fouille : méthodes ad-hoc

➥ Comment généraliser ?

Comment prendre la décision ?



En soft computing

$\mathcal{X}$

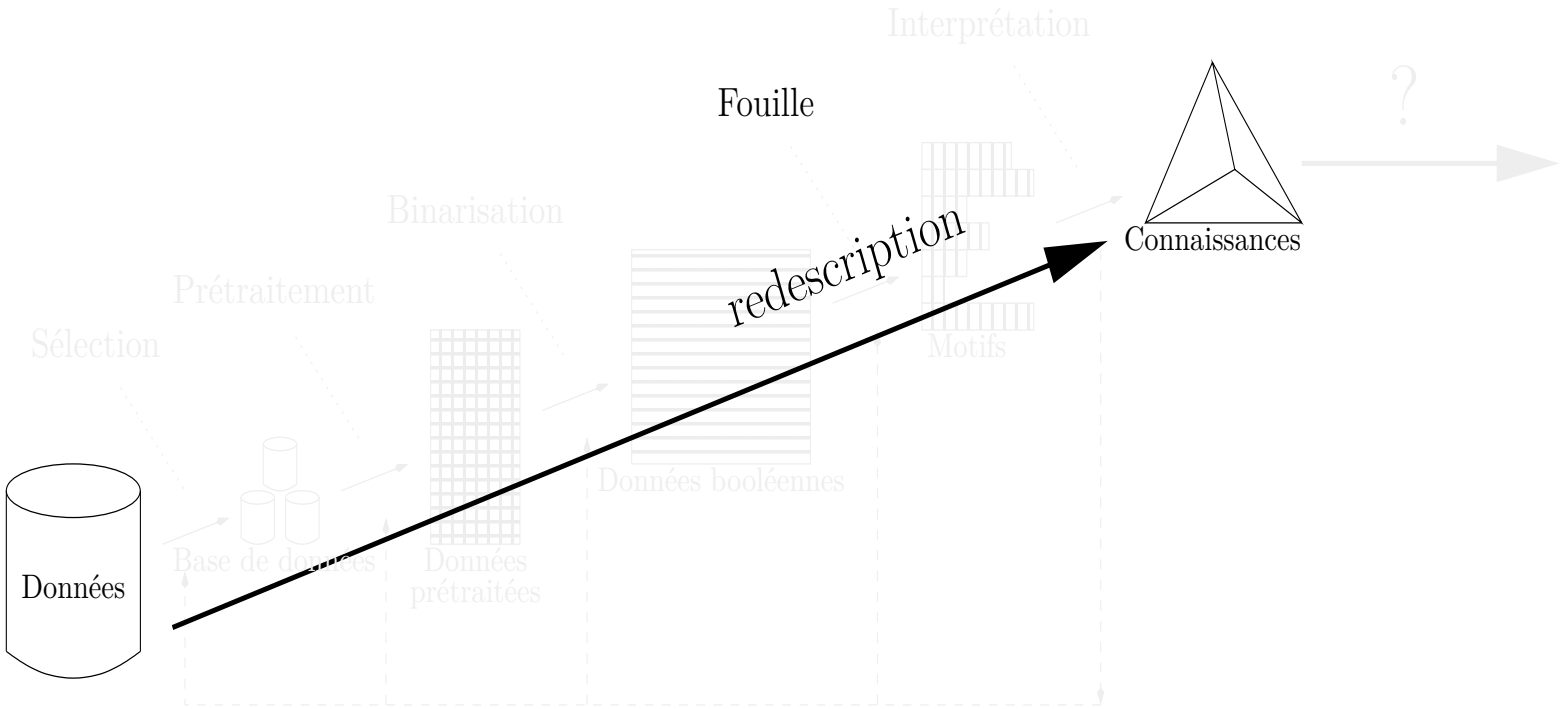
Espace de redescription des données

$\mathcal{X}$ redescription $\mathcal{F}$

Régression

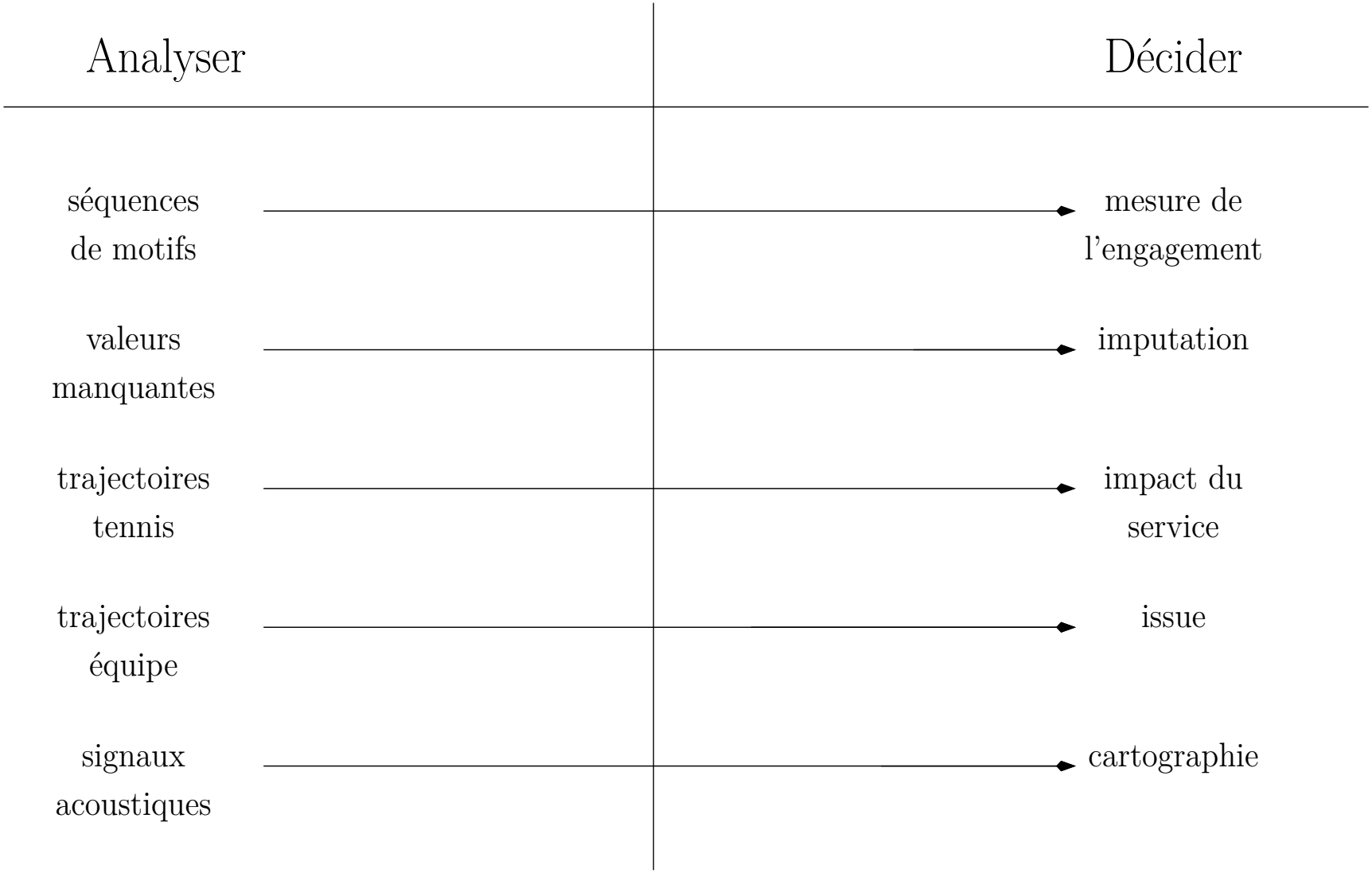
$\mathcal{X}$ redescription $\mathcal{F}$ régression

Redescription de données par la fouille

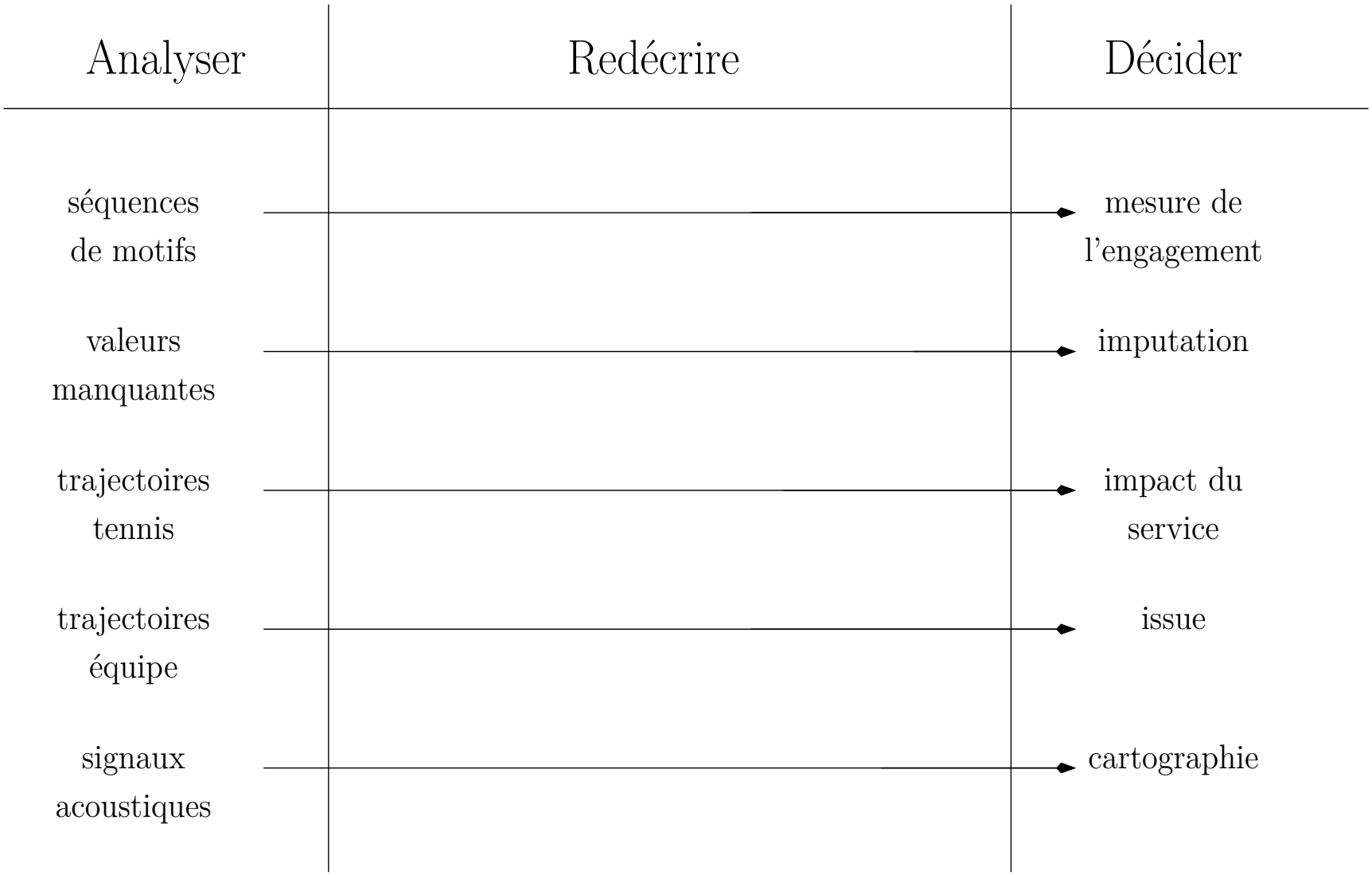


$\mathcal{X}$ redescription $\mathcal{F}$ régression

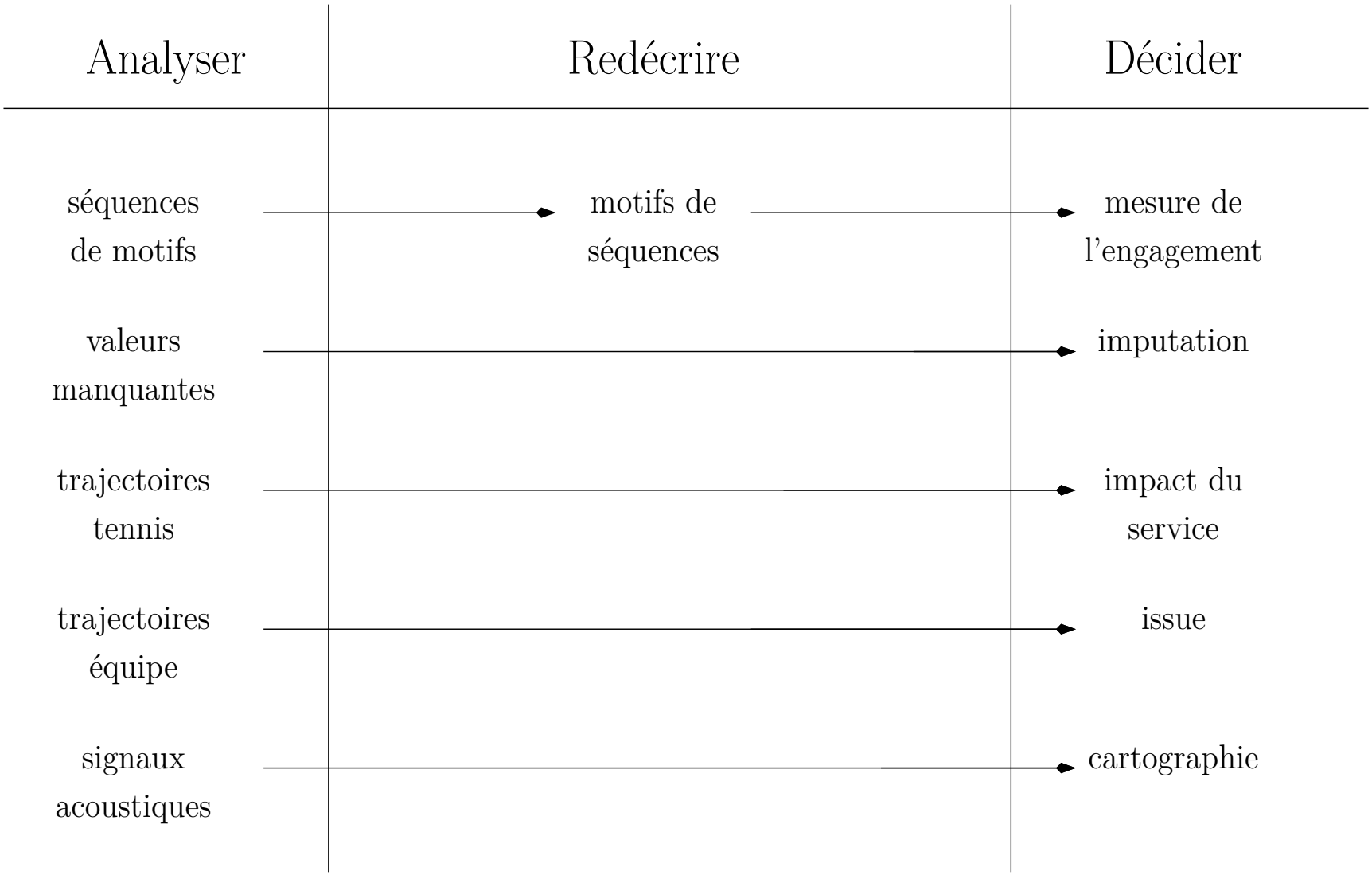
Analyser pour décider



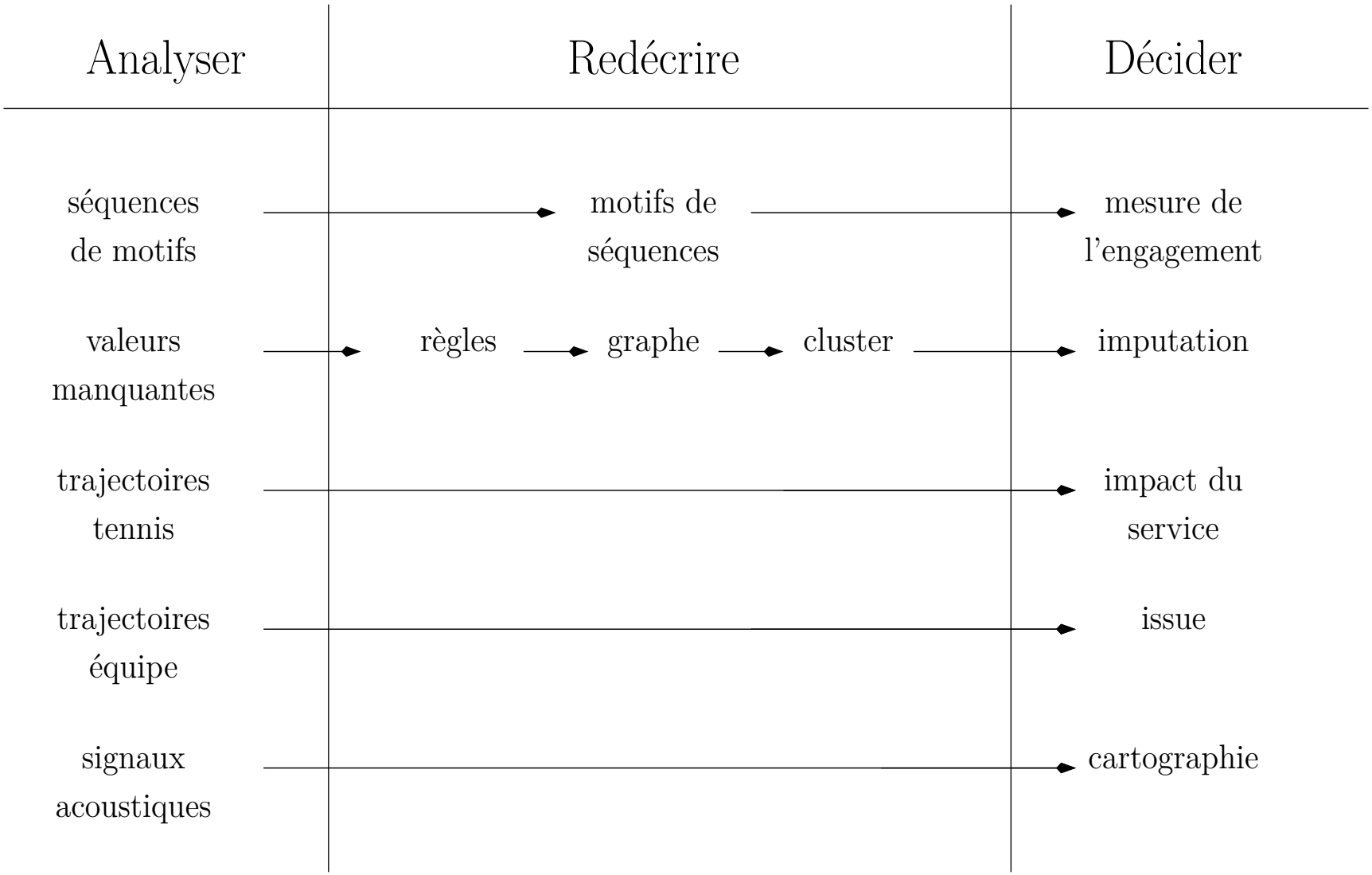
Redescription en fouille



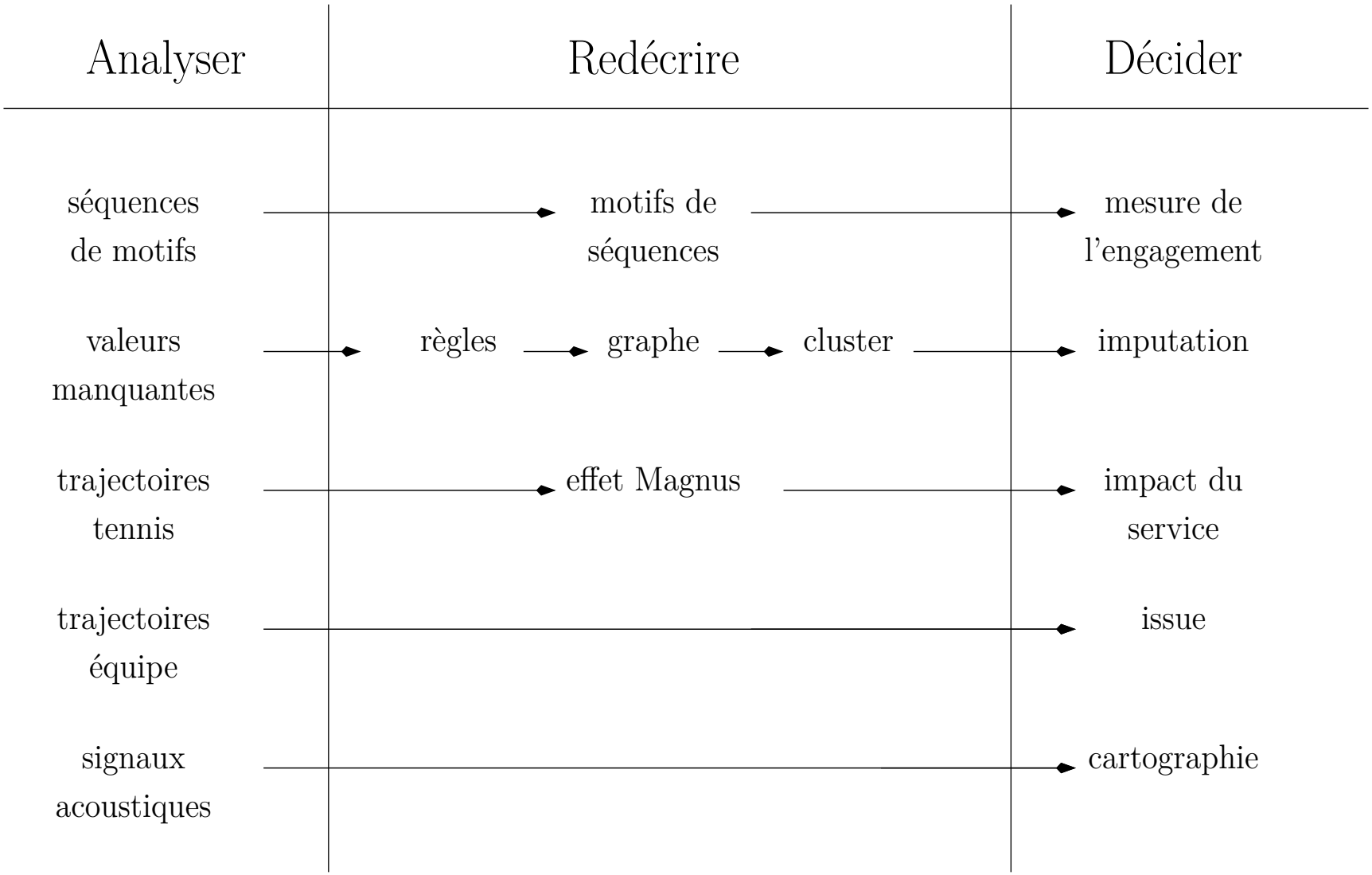
Modélisation du dialogue



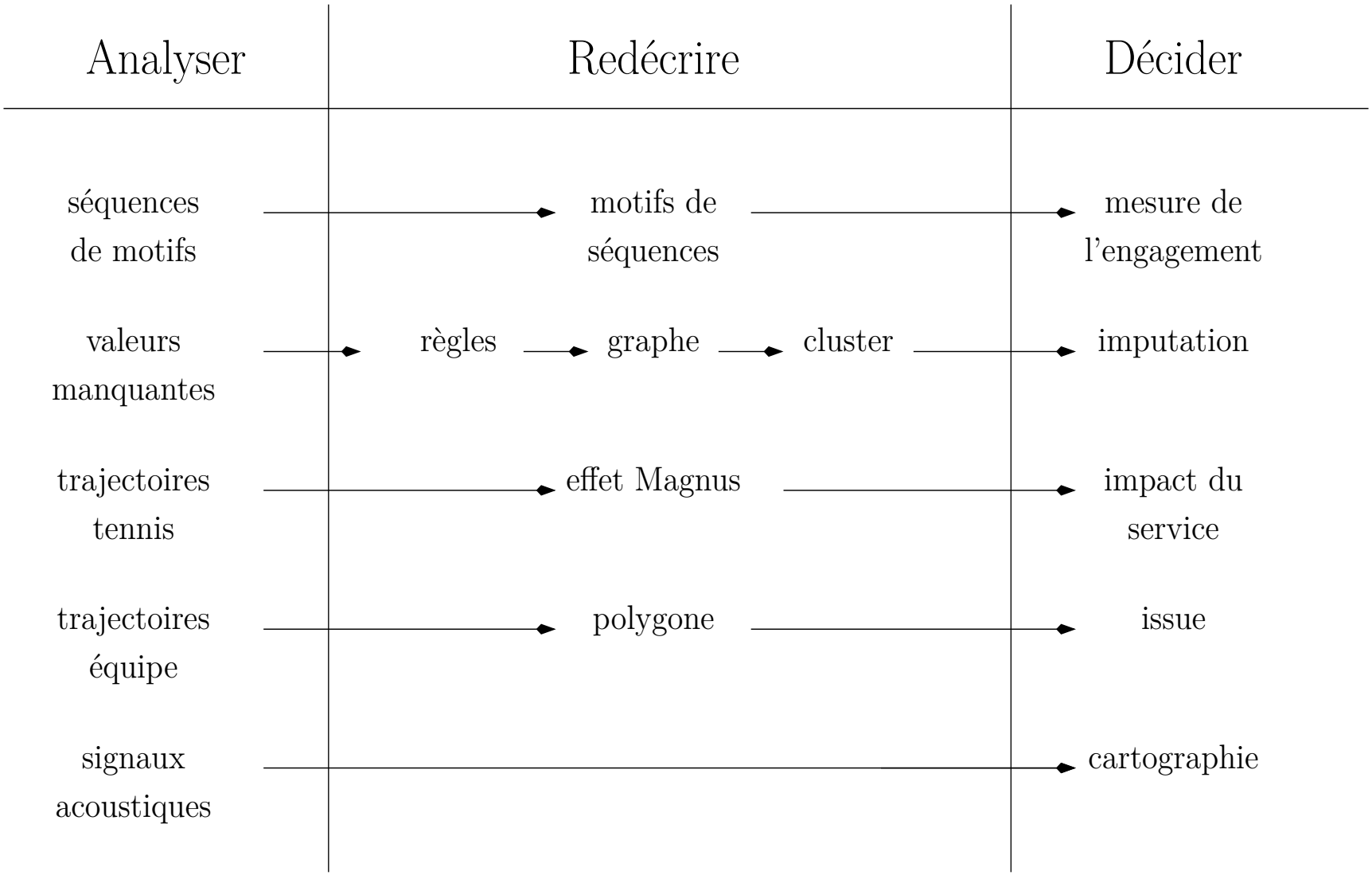
Valeurs manquantes



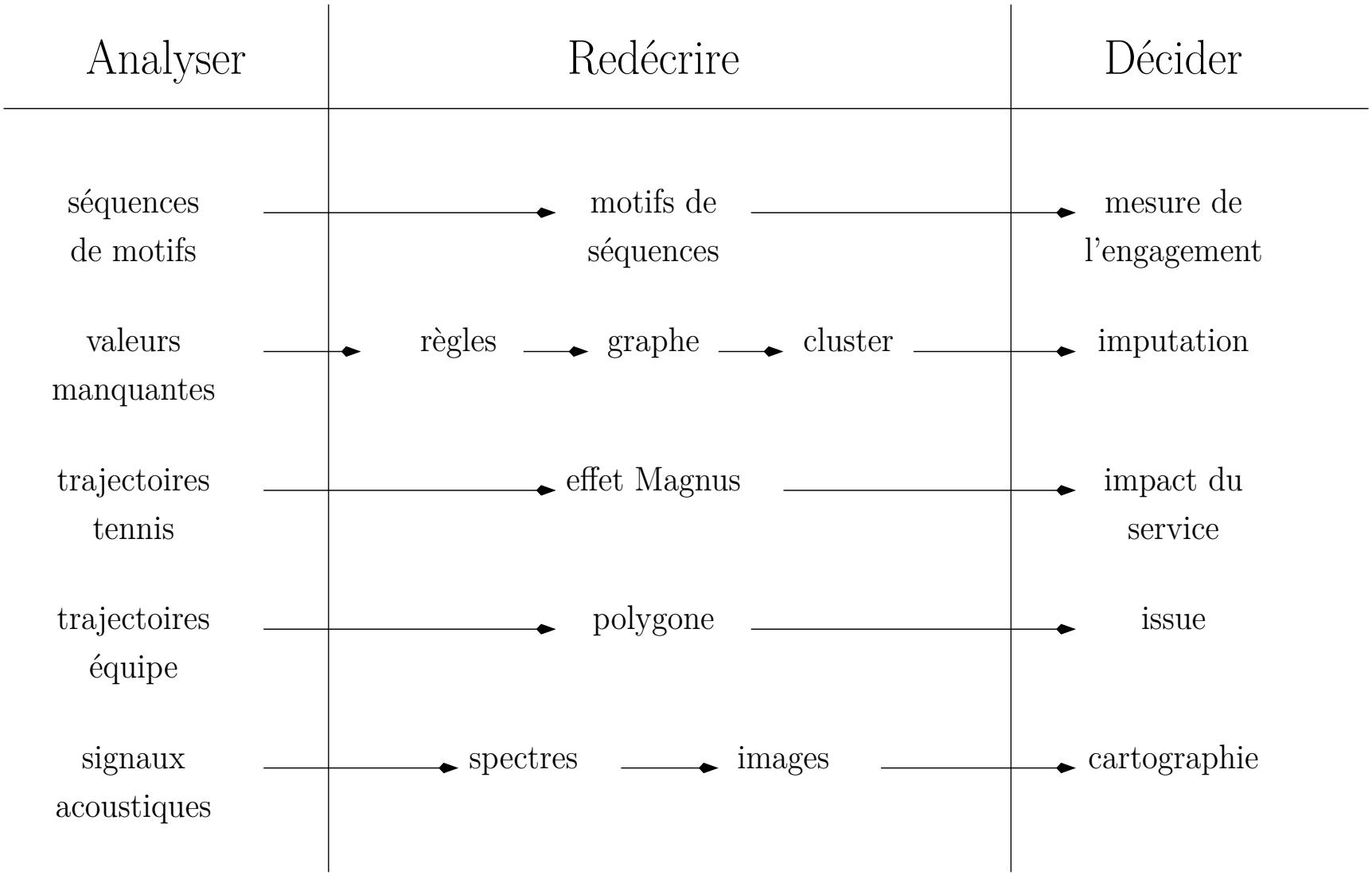
Sport : Tennis



e-Sport


Analyser
Redécrire
Décider
séquences de motifs
motifs de séquences
mesure de l'engagement
valeurs manquantes
règles
graphe
cluster
imputation
trajectoires tennis
effet Magnus
impact du service
trajectoires équipe
polygone
issue
signaux acoustiques
cartographie

Projet AIMS

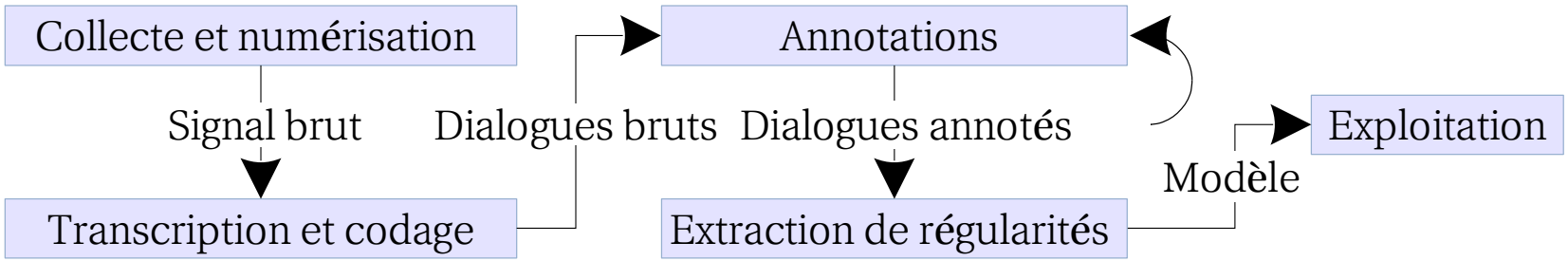

Analyser
Redécrire
Décider
séquences de motifs
motifs de séquences
mesure de l'engagement
valeurs manquantes
règles
graphe
cluster
imputation
trajectoires tennis
effet Magnus
impact du service
trajectoires équipe
polygone
issue
signaux acoustiques
spectres
images
cartographie

Réalisation d'un agent conversationnel

pour narration interactive

Mise en situation dans une classe de primaire

Modèle du dialogue narratif ?



- apport des psychologues
- théorie de l'esprit
- narration chez l'enfant
- les émotions comme vecteur

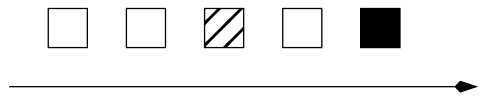
Comment susciter l'engagement de l'interlocuteur ?

À partir d'une suite d'annotations DIT++ :
(Dynamic Interpretation Theory)

Speaker	Breath Group	Dimension	Function
Adult	So, it's morning, chidren are coming to school	Task	Inform
Child	**Yes**	Contact Management	Contact
Adult	Look at this child	Task	Suggestion
Adult	he does not seem happy to be there	Task	Inform
Child	**Yeah, I saw**	Task	Confirm
Adult	And this boy, he has a ball...	Task	Inform
Adult	Look,	Task	Suggestion
Adult	it's Salim, he is calling his friends, [...]	Task	Inform
Child	**Uh Uh**	Time	Stalling

Proposition

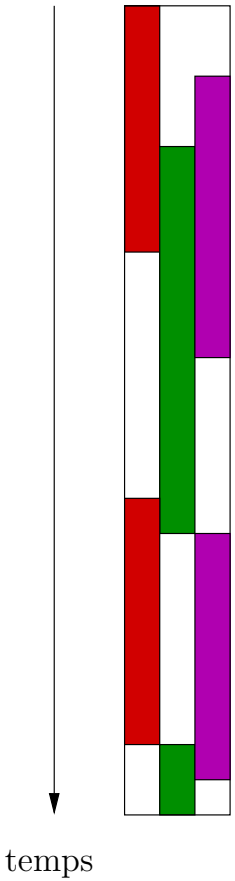
- motifs dédiés
- prédiction d'événement
- sous contrainte d'utilité
- que faire de ces motifs ?



➥ Approche ad-hoc

[PRC12 ; DBLP:journals/ria/SerbanBALCRP14]

Redescription des données



			…	…	…	cible
X						
X						
X		X				
X		X				
X	X	X				
X	X	X				
X	X	X				?
	X	X				
	X	X				
	X	X				
	X					
	X					
	X					
	X					
X	X					
X		X				
X		X				
X		X				
X		X				
X		X				
X		X				
	X	X				
	X					

C4.5
Bayes
random forest
etc.

Redescription des données

Prédire l'engagement :

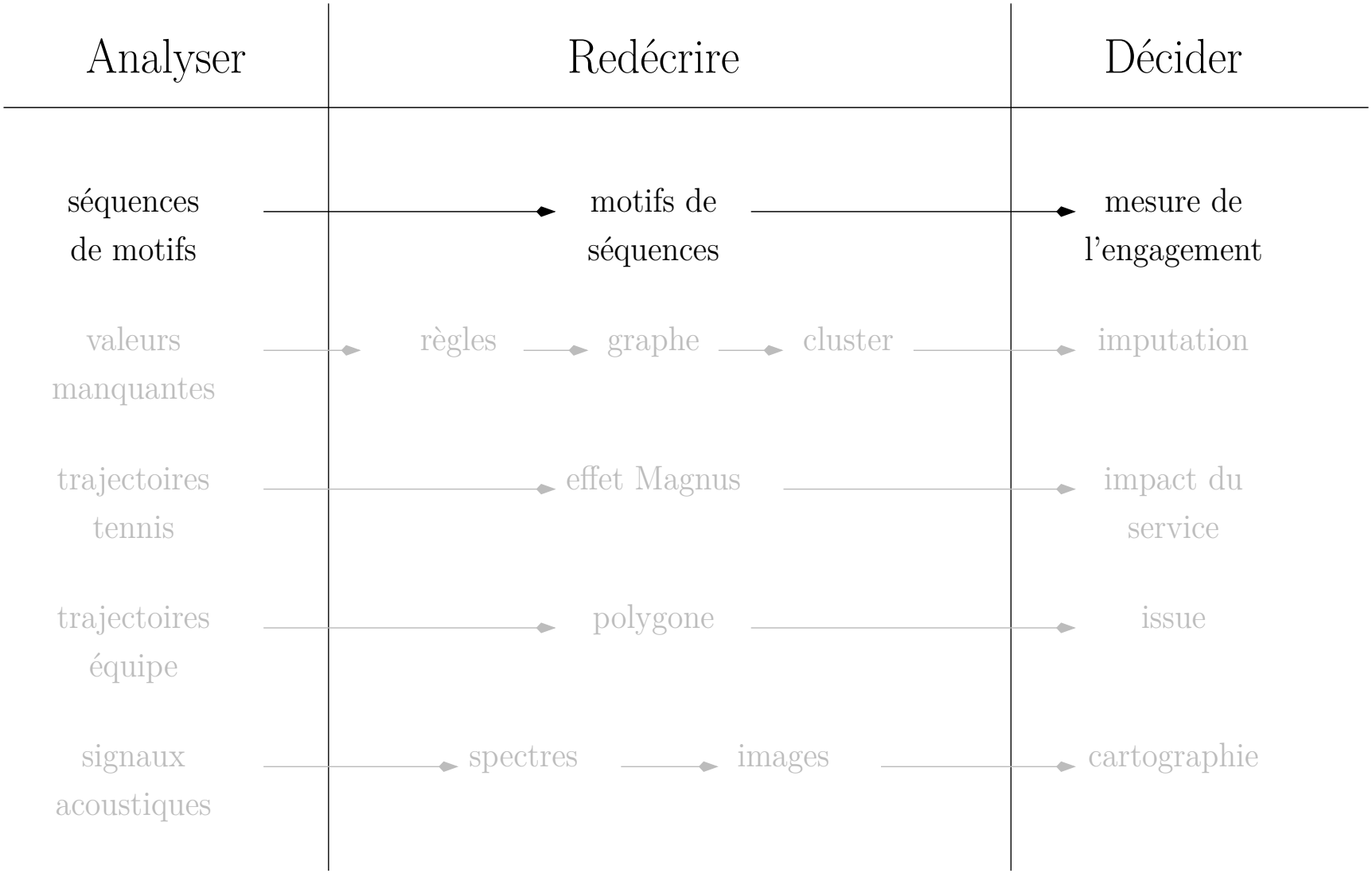
1. extraire des sous-séquences fermées
2. recoder les événements

$$\text{séquences de motifs} \rightarrow \text{motifs de séquences}$$

3. nourrir un classifieur

➥ Prédiction de l'intervention de l'enfant

➥ Mesure de l'engagement [LRC16]


Analyser
Redécrire
Décider
séquences de motifs
motifs de séquences
mesure de l'engagement
valeurs manquantes
règles
graphe
cluster
imputation
trajectoires tennis
effet Magnus
impact du service
trajectoires équipe
polygone
issue
signaux acoustiques
spectres
images
cartographie

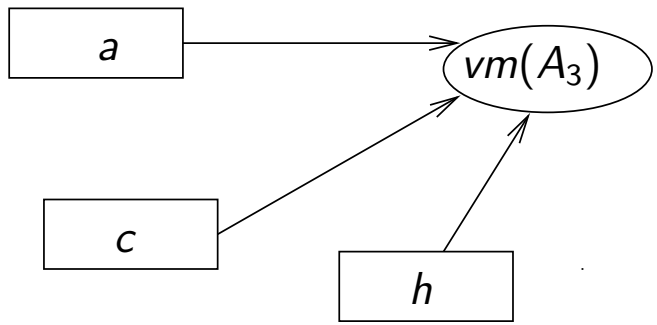
Exemple d'un contexte incomplet

	A_1	A_2	A_3	A_4
o_1	a	c	**?**	h
o_2	**?**	c	e	**?**
o_3	a	c	**?**	h
o_4	a	d	f	**?**
o_5	**?**	c	f	**?**
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	**?**	d	g	**?**

[CI-BENOTHMAN-2009]

Graphes de caractérisation

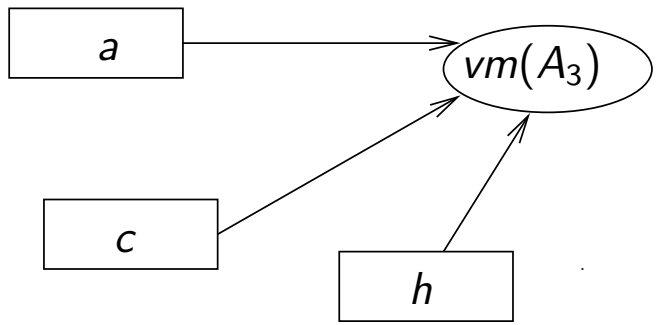
	A_1	A_2	A_3	A_4
o_1	a	c	?	h
o_2	?	c	e	?
o_3	a	c	?	h
o_4	a	d	f	?
o_5	?	c	f	?
o_6	b	?	f	h
o_7	a	?	g	?
o_8	?	d	g	?



[CI-BENOTHMAN-2009]

Graphes de caractérisation

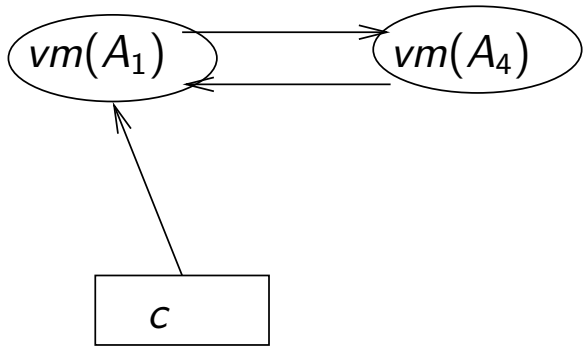
	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	?	c	e	?
o_3	a	c	direct	h
o_4	a	d	f	?
o_5	?	c	f	?
o_6	b	?	f	h
o_7	a	?	g	?
o_8	?	d	g	?

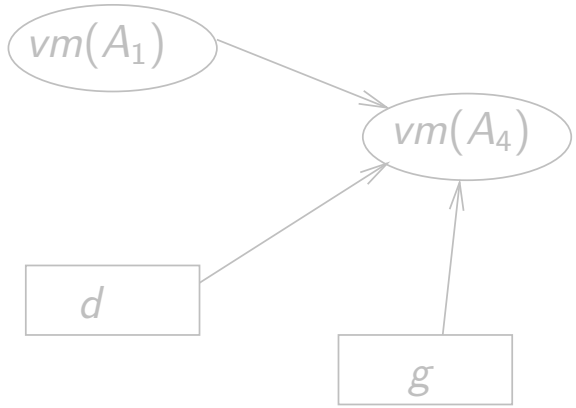


[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	?	c	e	?
o_3	a	c	direct	h
o_4	a	d	f	?
o_5	?	c	f	?
o_6	b	?	f	h
o_7	a	?	g	?
o_8	?	d	g	?

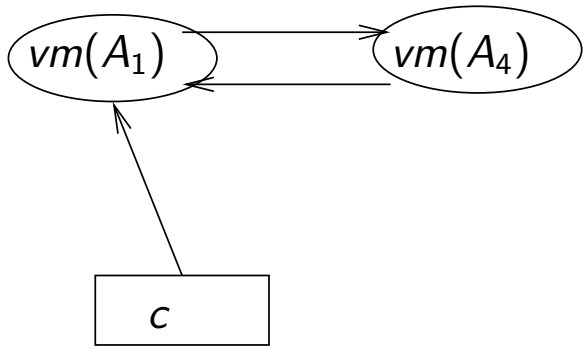


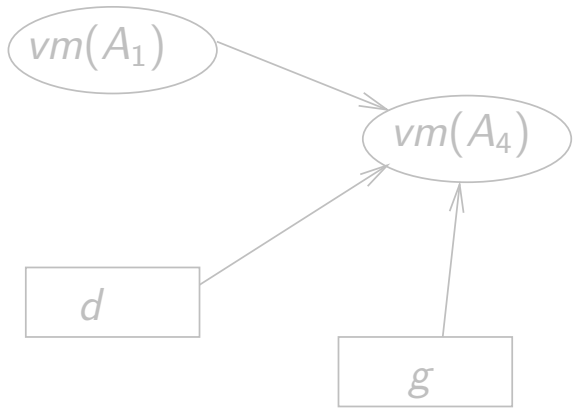


[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	?
o_3	a	c	direct	h
o_4	a	d	f	?
o_5	hybrid	c	f	?
o_6	b	?	f	h
o_7	a	?	g	?
o_8	?	d	g	?

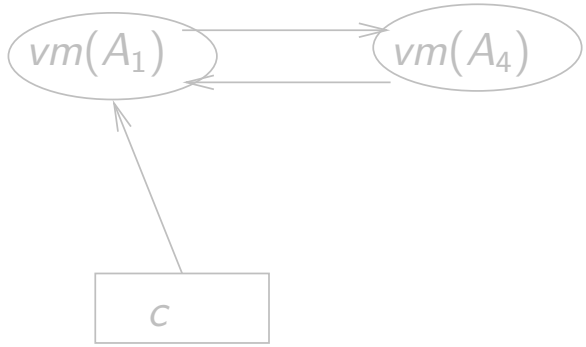


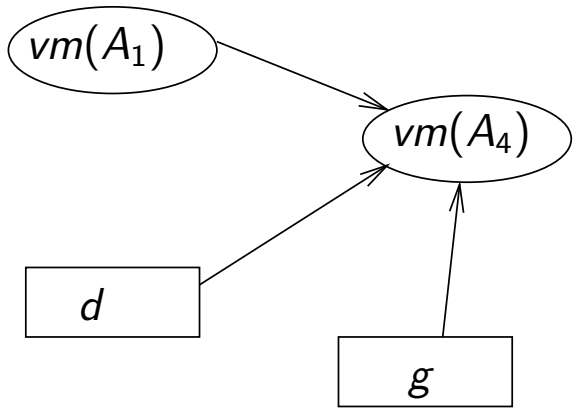


[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	**?**
o_3	a	c	direct	h
o_4	a	d	f	**?**
o_5	hybrid	c	f	**?**
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	**?**	d	g	**?**

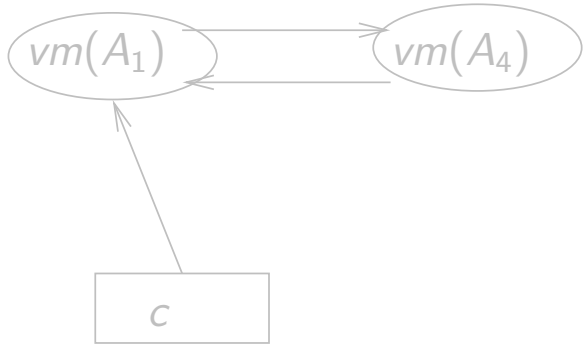


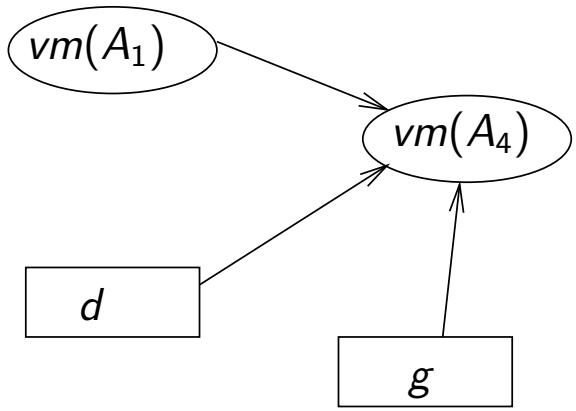


[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	**?**
o_3	a	c	direct	h
o_4	a	d	f	**?**
o_5	hybrid	c	f	**?**
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	random	d	g	**?**

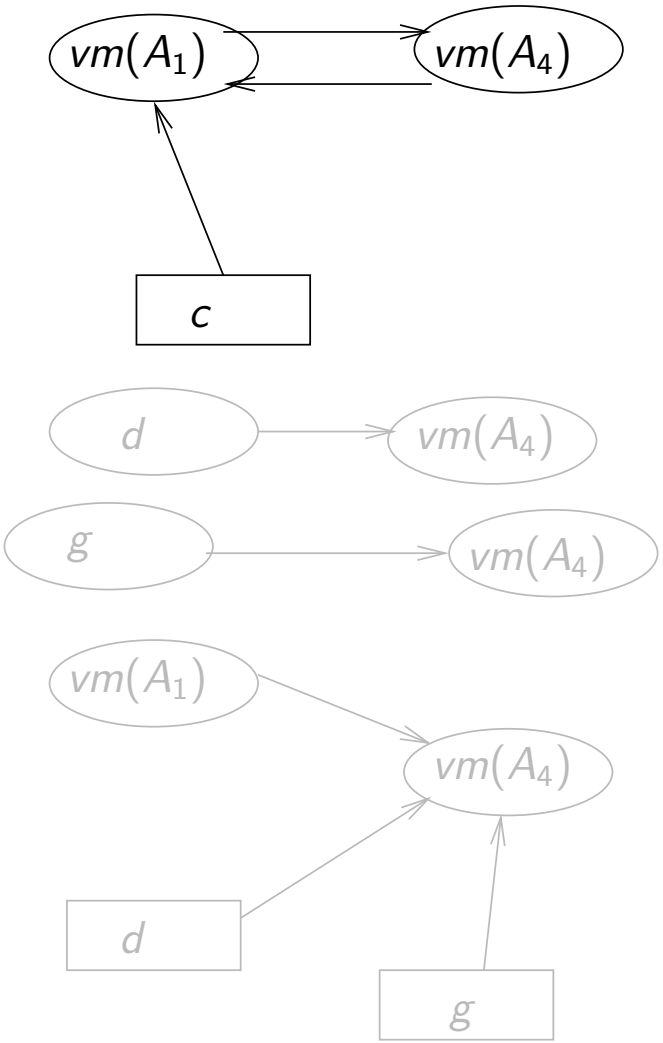




[CI-BENOTHMAN-2009]

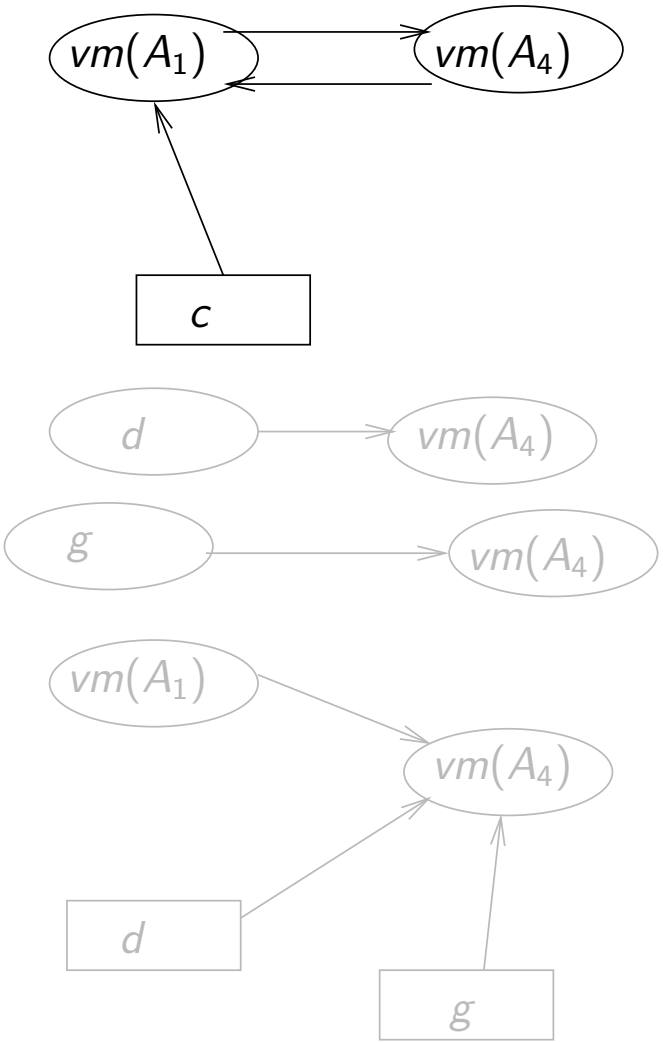
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	?
o_3	a	c	direct	h
o_4	a	d	f	?
o_5	hybrid	c	f	?
o_6	b	?	f	h
o_7	a	?	g	?
o_8	random	d	g	?



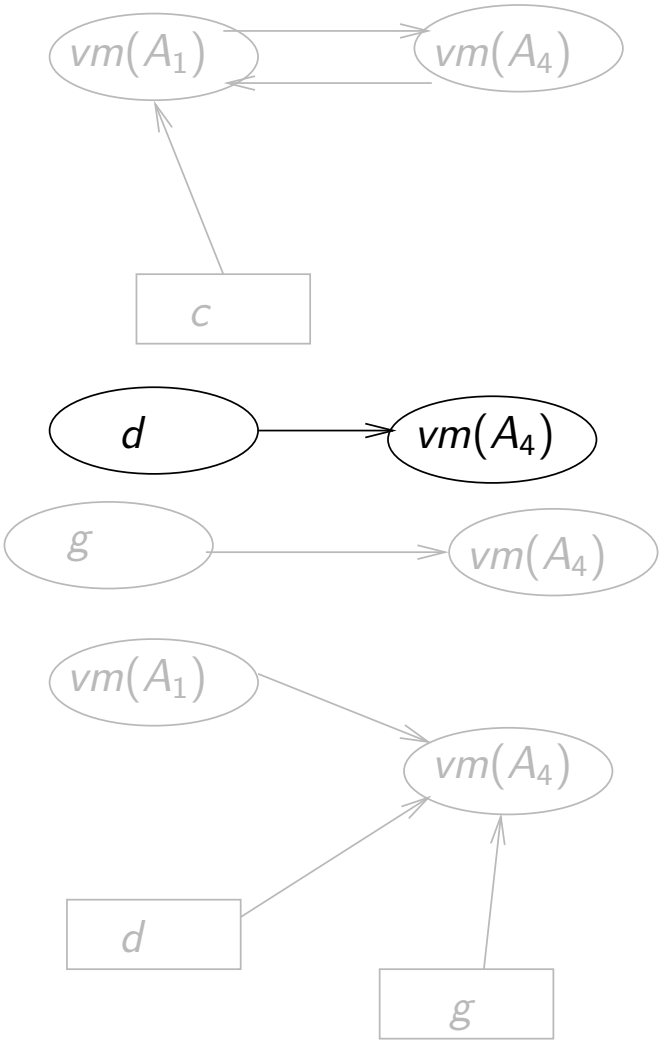
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	**?**
o_5	hybrid	c	f	indirect
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	random	d	g	**?**



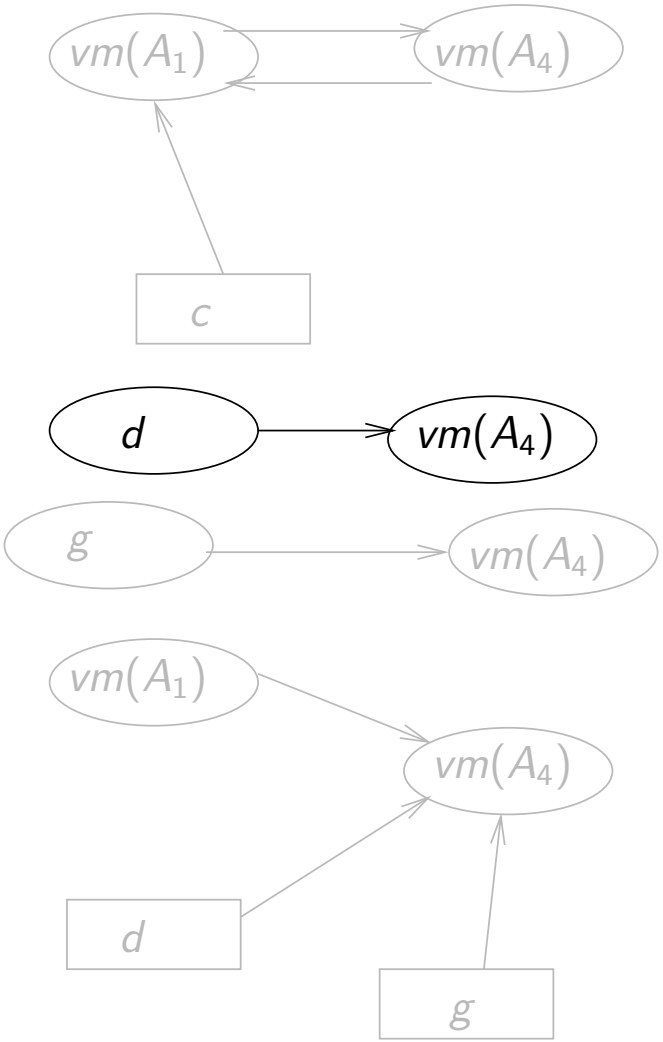
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	?
o_5	hybrid	c	f	indirect
o_6	b	?	f	h
o_7	a	?	g	?
o_8	random	d	g	?



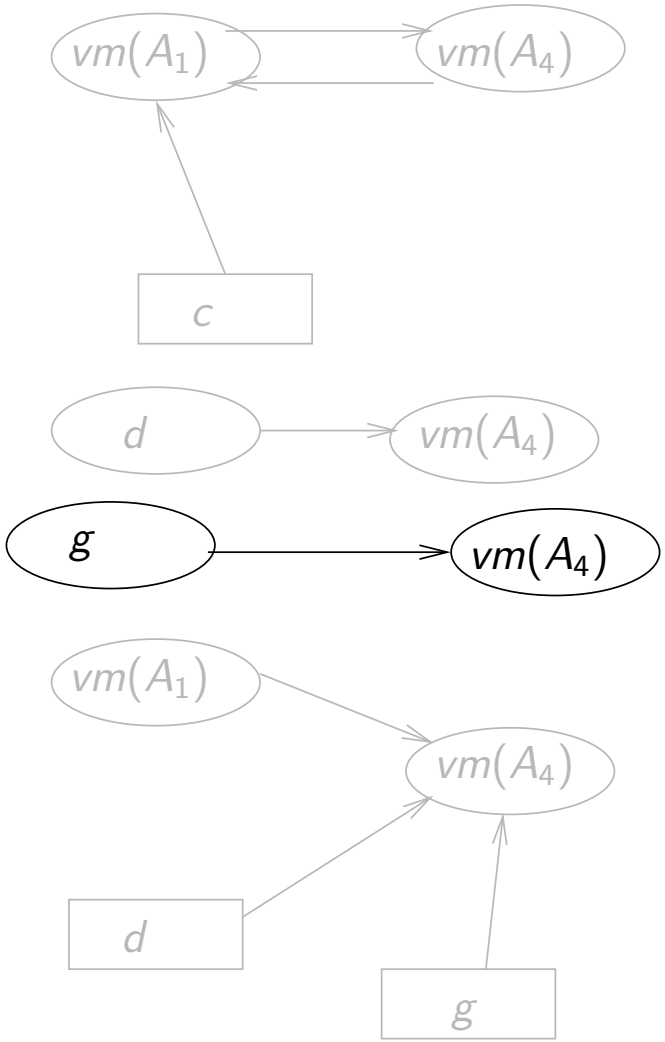
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	?	f	h
o_7	a	?	g	?
o_8	random	d	g	?



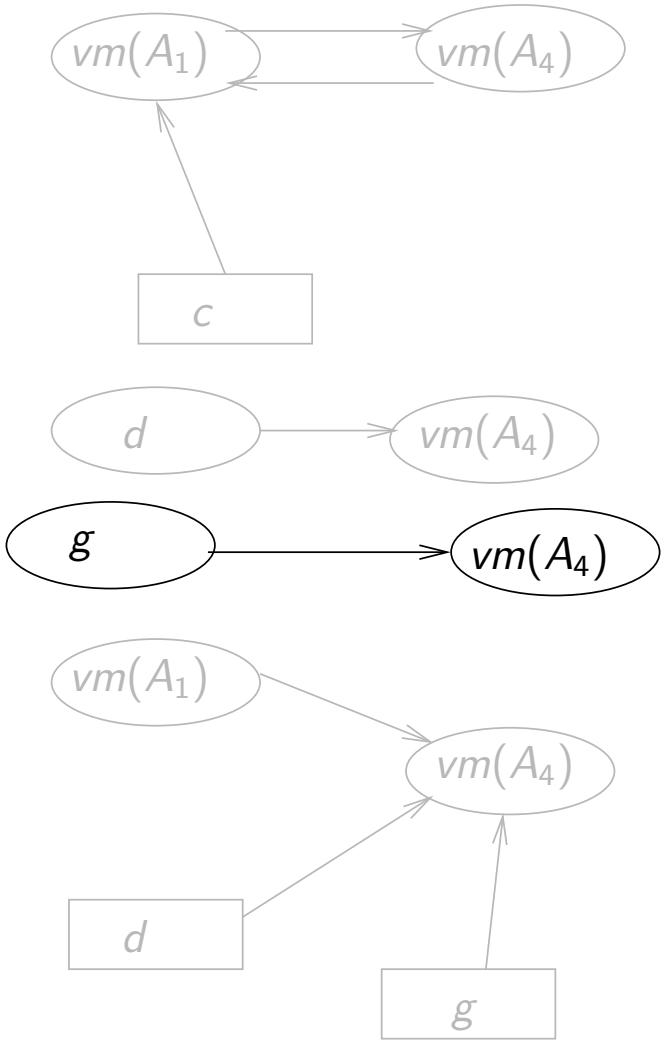
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	?	f	h
o_7	a	?	g	?
o_8	random	d	g	?



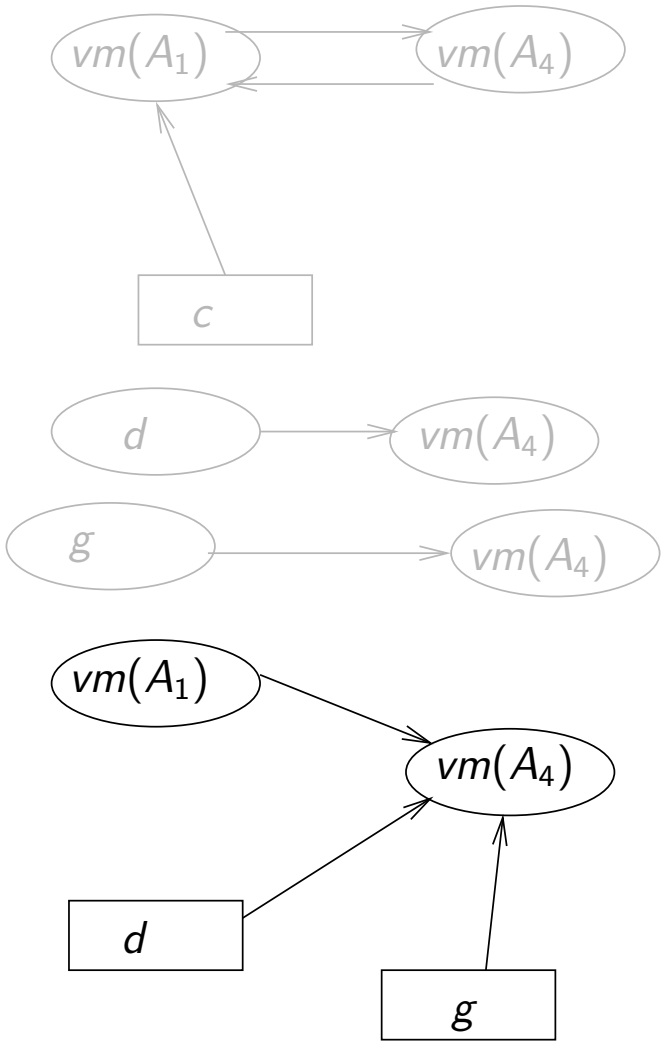
Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	**?**	f	h
o_7	a	**?**	g	direct
o_8	random	d	g	**?**



Graphes de caractérisation

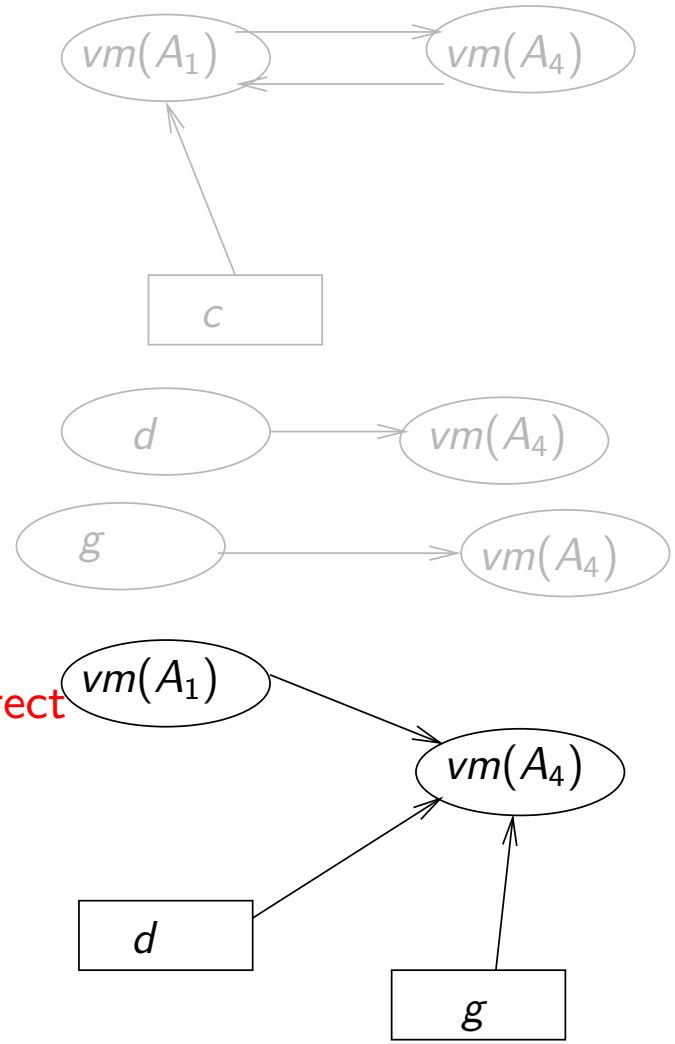
	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	**?**	f	h
o_7	a	**?**	g	direct
o_8	random	d	g	**?**



[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	**?**	f	h
o_7	a	**?**	g	direct
o_8	random	d	g	indirect,direct



[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	?	f	h
o_7	a	?	g	direct
o_8	random	d	g	indirect,direct

[CI-BENOTHMAN-2009]

Graphes de caractérisation

	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	random	f	h
o_7	a	random	g	direct
o_8	random	d	g	indirect,direct

[CI-BENOTHMAN-2009]

Graphes de caractérisation

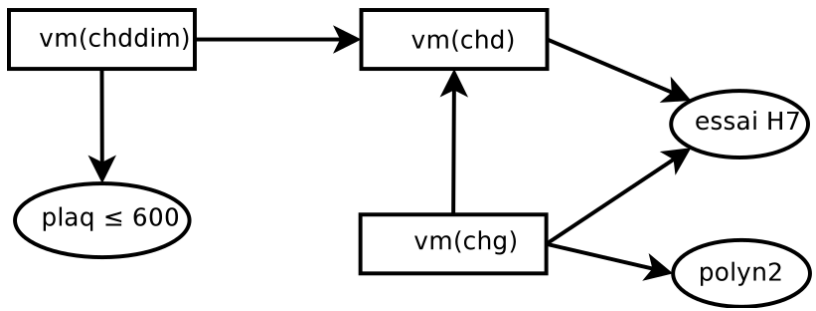
	A_1	A_2	A_3	A_4
o_1	a	c	direct	h
o_2	hybrid	c	e	indirect
o_3	a	c	direct	h
o_4	a	d	f	direct
o_5	hybrid	c	f	indirect
o_6	b	random	f	h
o_7	a	random	g	direct
o_8	random	d	g	indirect,direct

[CI-BENOTHMAN-2009]

Page Rank sur les graphes

Exemple pour la maladie de Hodgkin

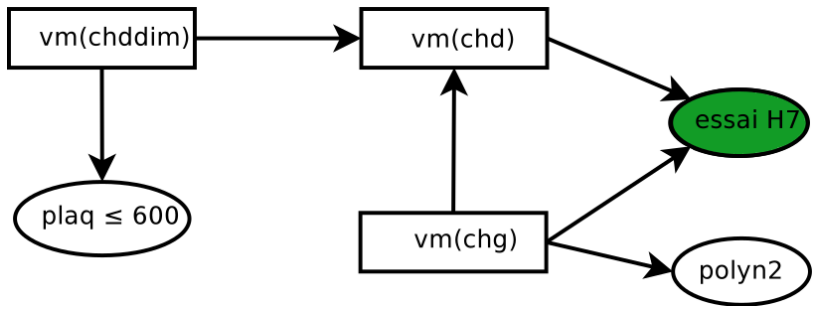
Explication	Probabilité
vm(chd)	
vm(chg)	
vm(chddim)	
plaq $\leqslant$ 600	
essai H7	
polyn2	



Page Rank sur les graphes

Exemple pour la maladie de Hodgkin

Explication	Probabilité
vm(chd)	0.24
vm(chg)	0.08
vm(chddim)	0.15
plaq $\leqslant$ 600	0.08
essai H7	**0.32**
polyn2	0.10



➥ Clustering des valeurs manquantes

➥ Imputation par cluster

Explications des valeurs manquantes

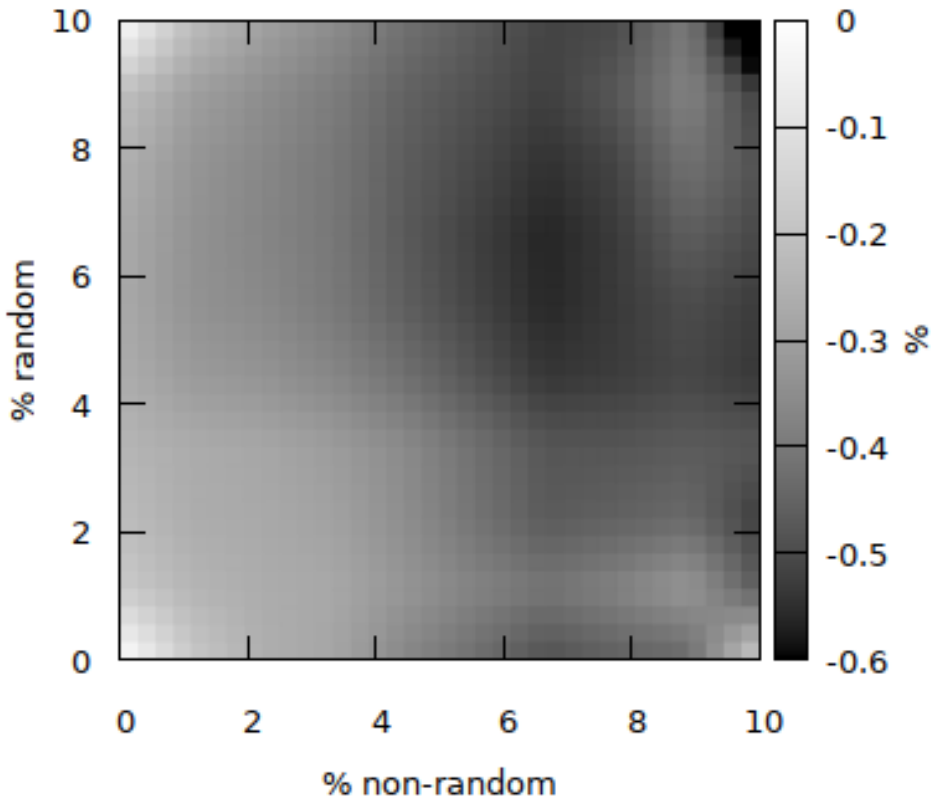
	A_1	A_2	A_3	A_4
o_1	a	c	$?(A_1 = a)$	h
o_2	?	c	e	$?(A_1 =?)$
o_3	a	c	$?(A_1 = a)$	h
o_4	a	d	f	$?(A_2 = d)$
o_5	?	c	f	$?(A_1 =?)$
o_6	b	$?(A_1 = b)$	f	h
o_7	a	$?(A_3 = g)$	g	$?(A_3 = g)$
o_8	$?(A_2 = d)$	d	g	$?(A_2 = d)$

On remplace par la valeur la plus fréquemment présente avec l'explication :

- $o_7 : ?(A_3 = g) \rightarrow A_2 = d$
- $o_8 : ?(A_2 = d) \rightarrow A_1 = a$

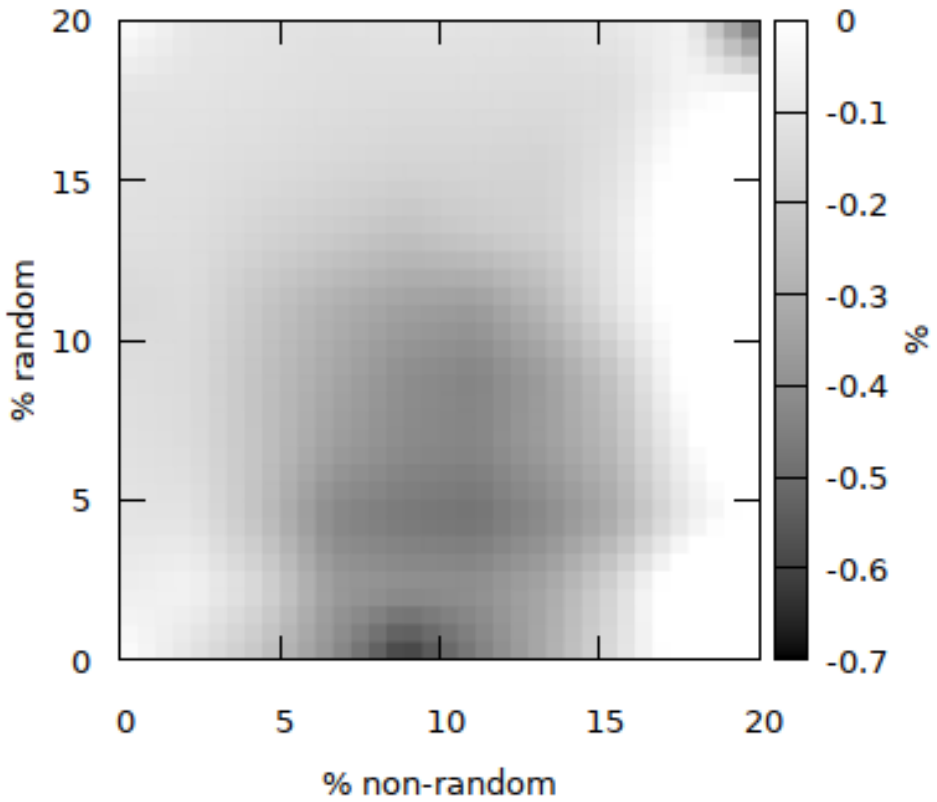
Apport de l'explication

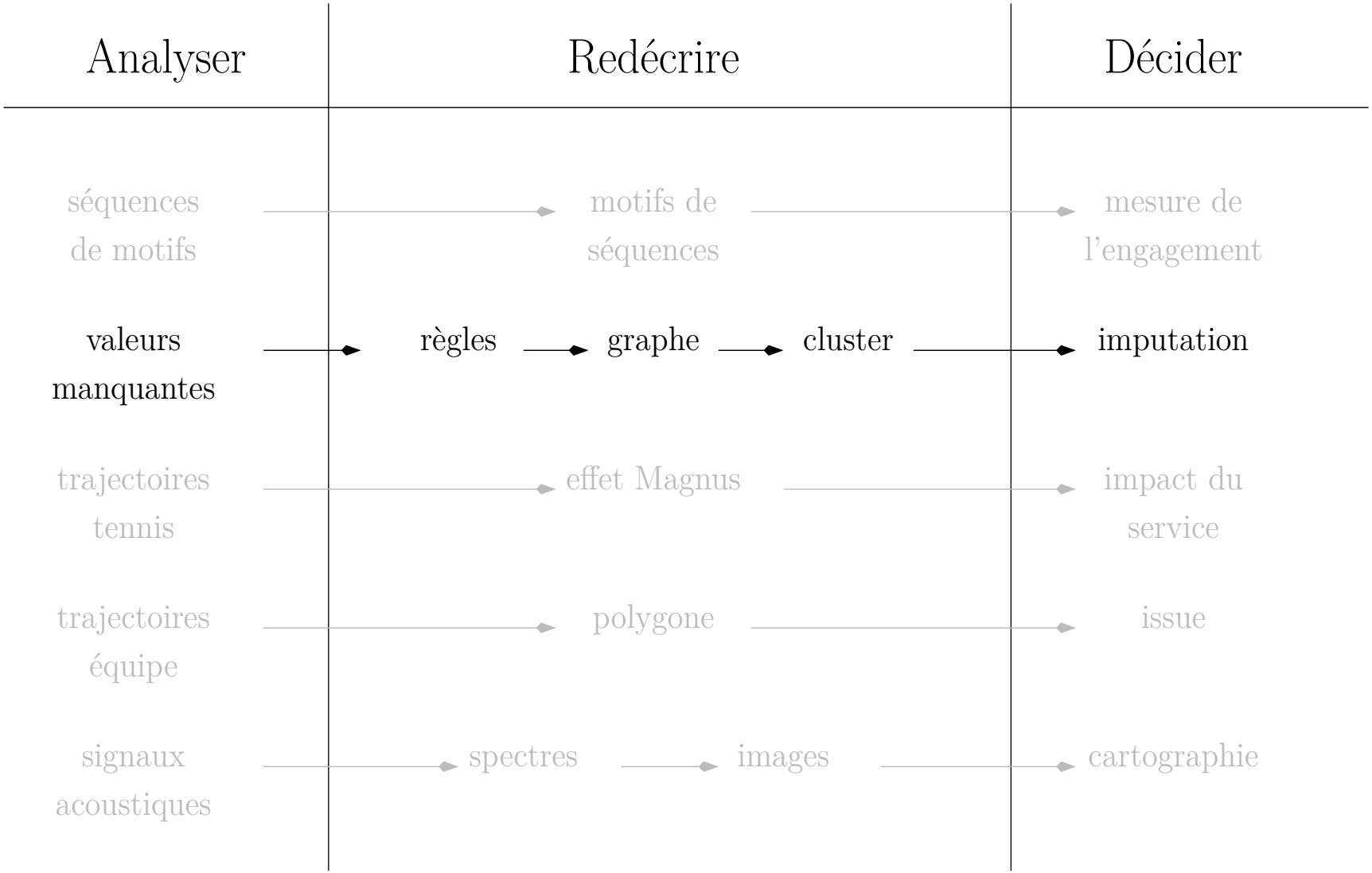
Base Australian



Apport de l'explication

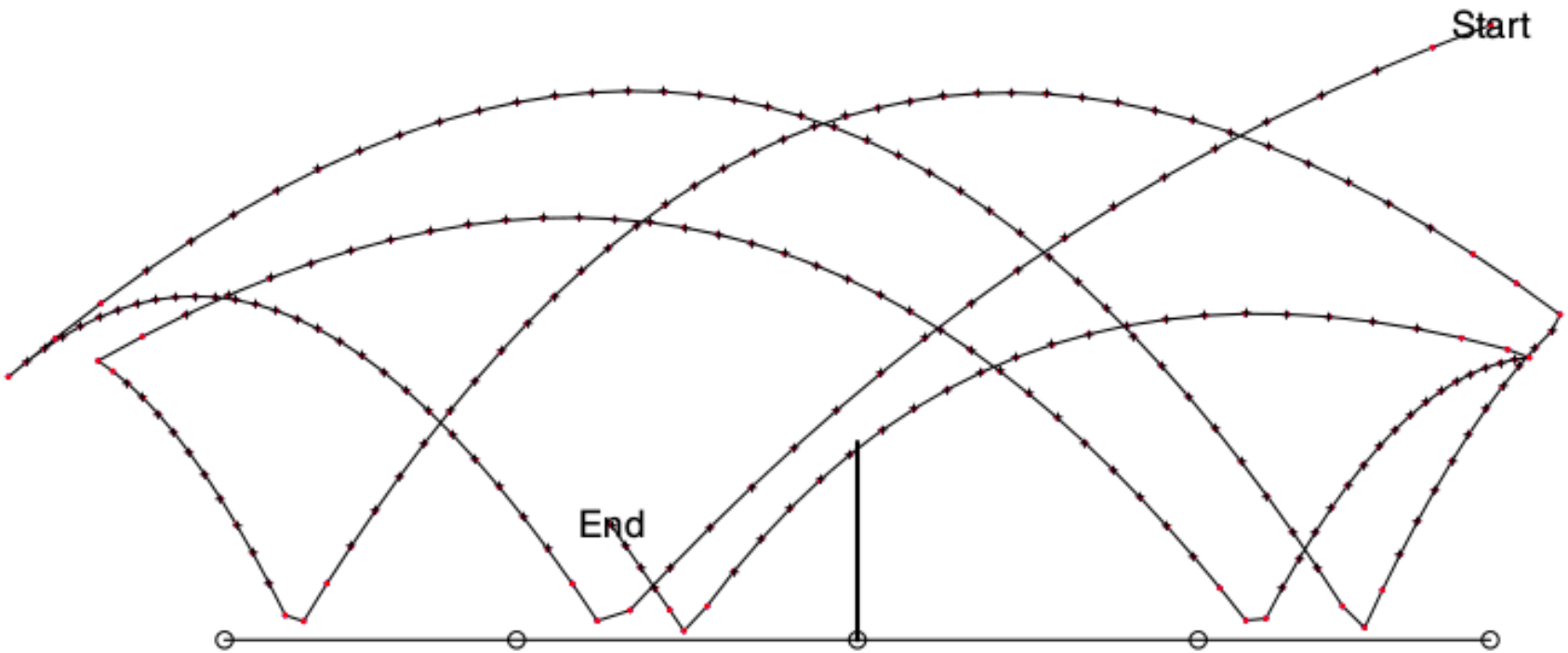
Base Heart



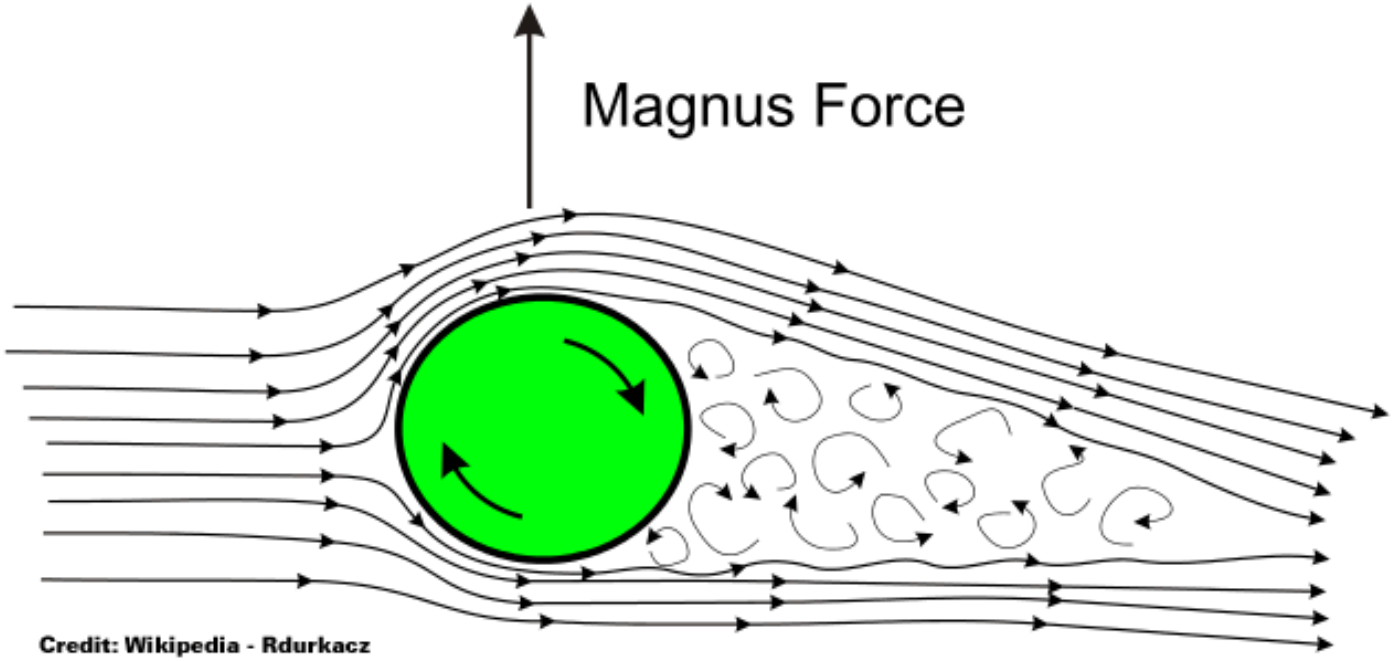

Analyser
Redécrire
Décider
séquences de motifs
motifs de séquences
mesure de l'engagement
valeurs manquantes
règles
graphe
cluster
imputation
trajectoires tennis
effet Magnus
impact du service
trajectoires équipe
polygone
issue
signaux acoustiques
spectres
images
cartographie

Exemple 3 : Tennis

- 5 ans de trajectoires des meilleur-e-s joueu-r-se-s du monde
- système Hawk-Eye
- approximations à 40 Hz de trajectoires physiques



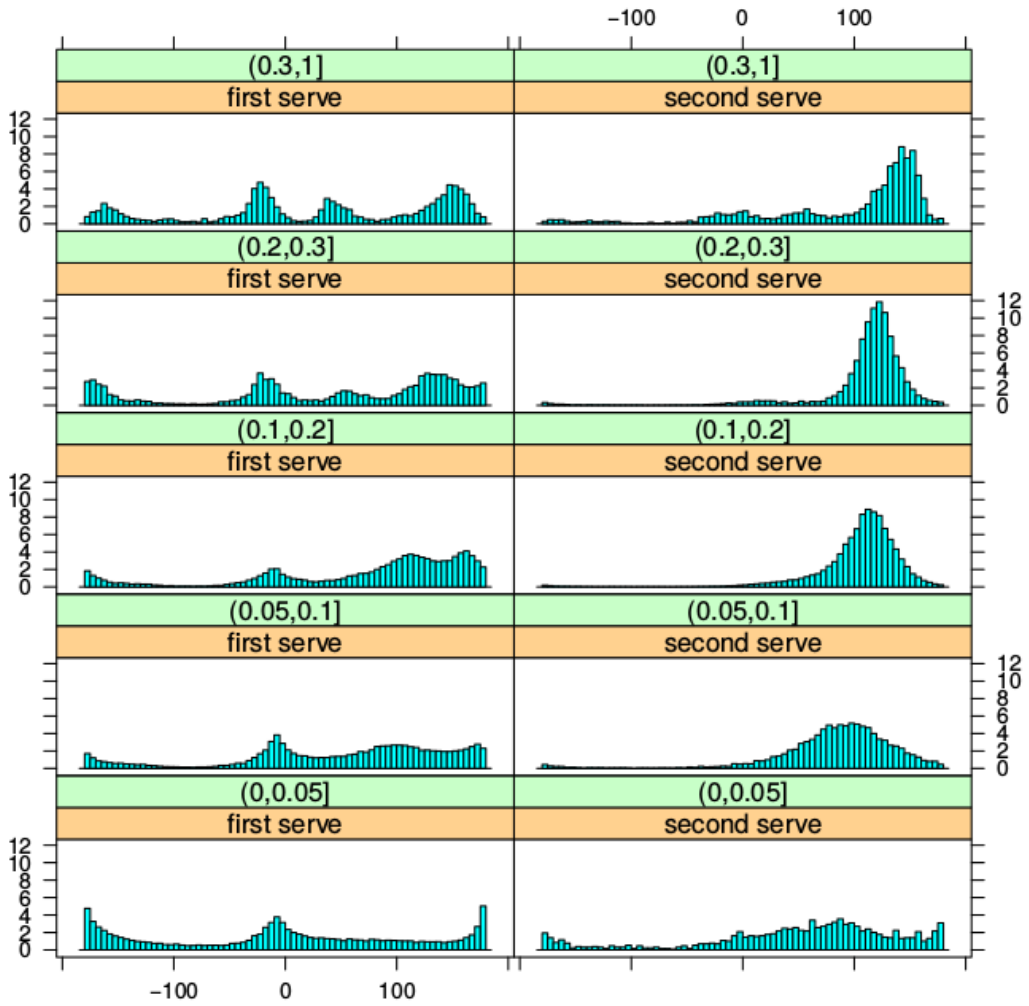
Exemple 3 : Tennis



➥ Approximation linéaire du coefficient de Magnus
à partir des trajectoires 3D

[RMM15]

Exemple 3 : Tennis

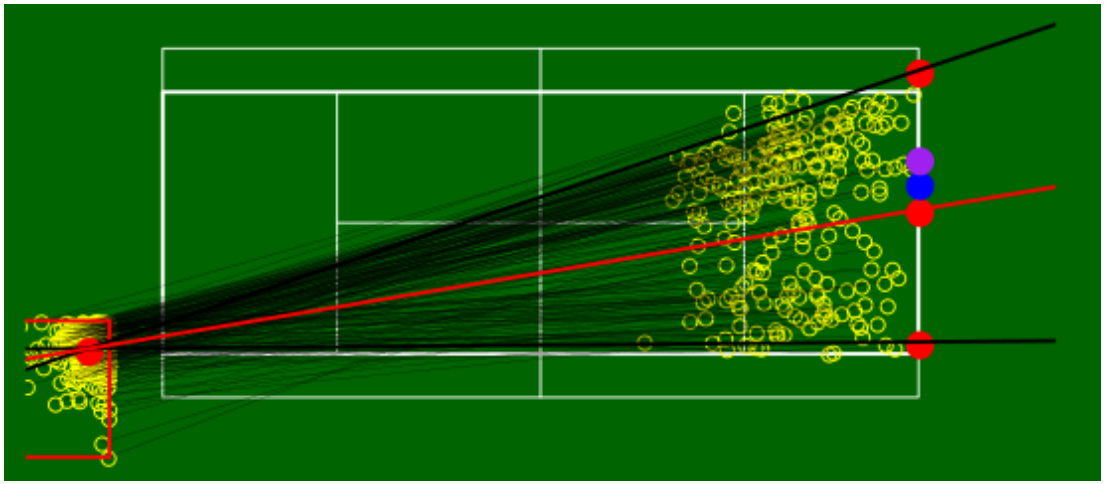


➥ Analyse stratégique du service [SAM11316]

Position du receveur

Principe d'Henri Cochet :

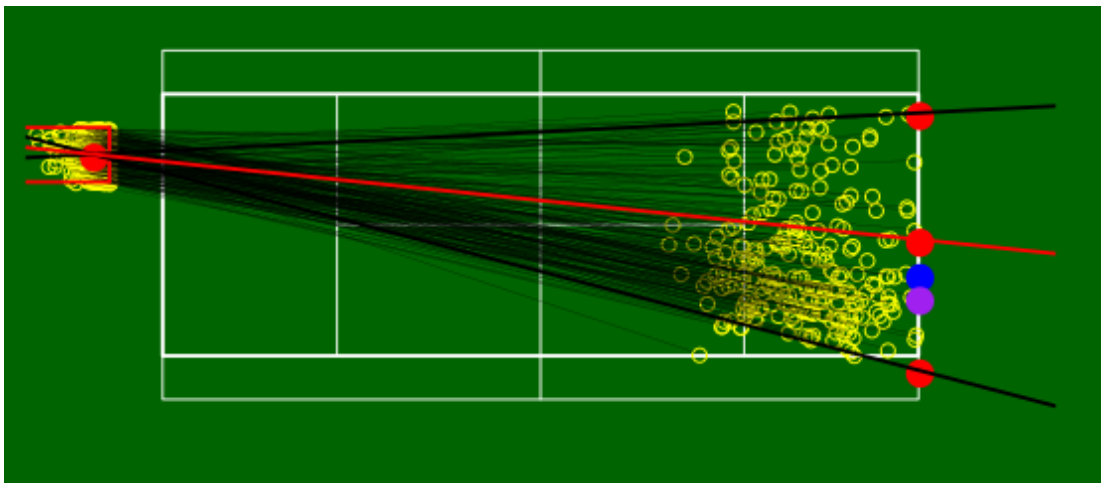
Se positionner sur la bissectrice de l'angle des possibles.



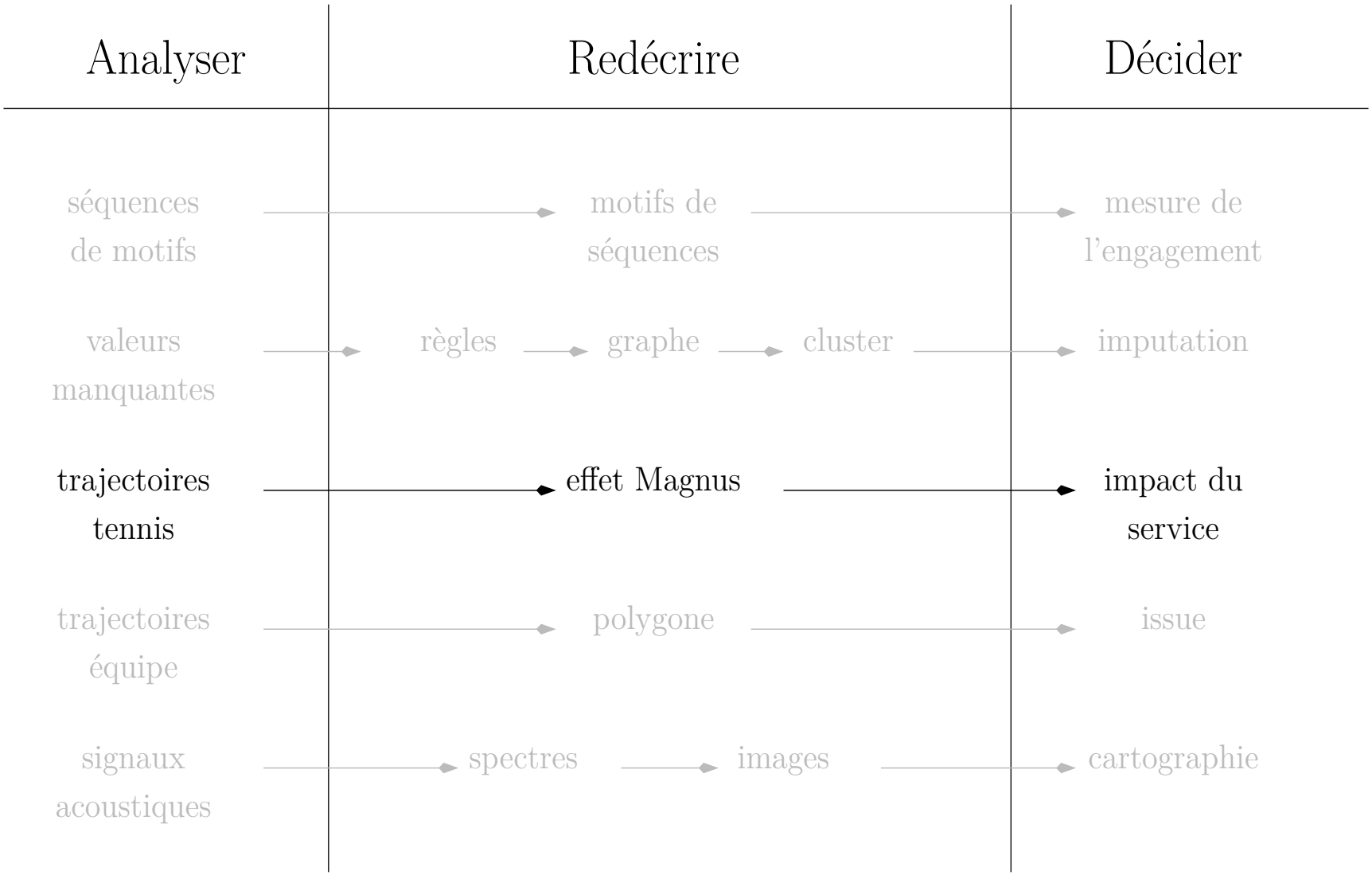
Position du receveur

Principe d'Henri Cochet :

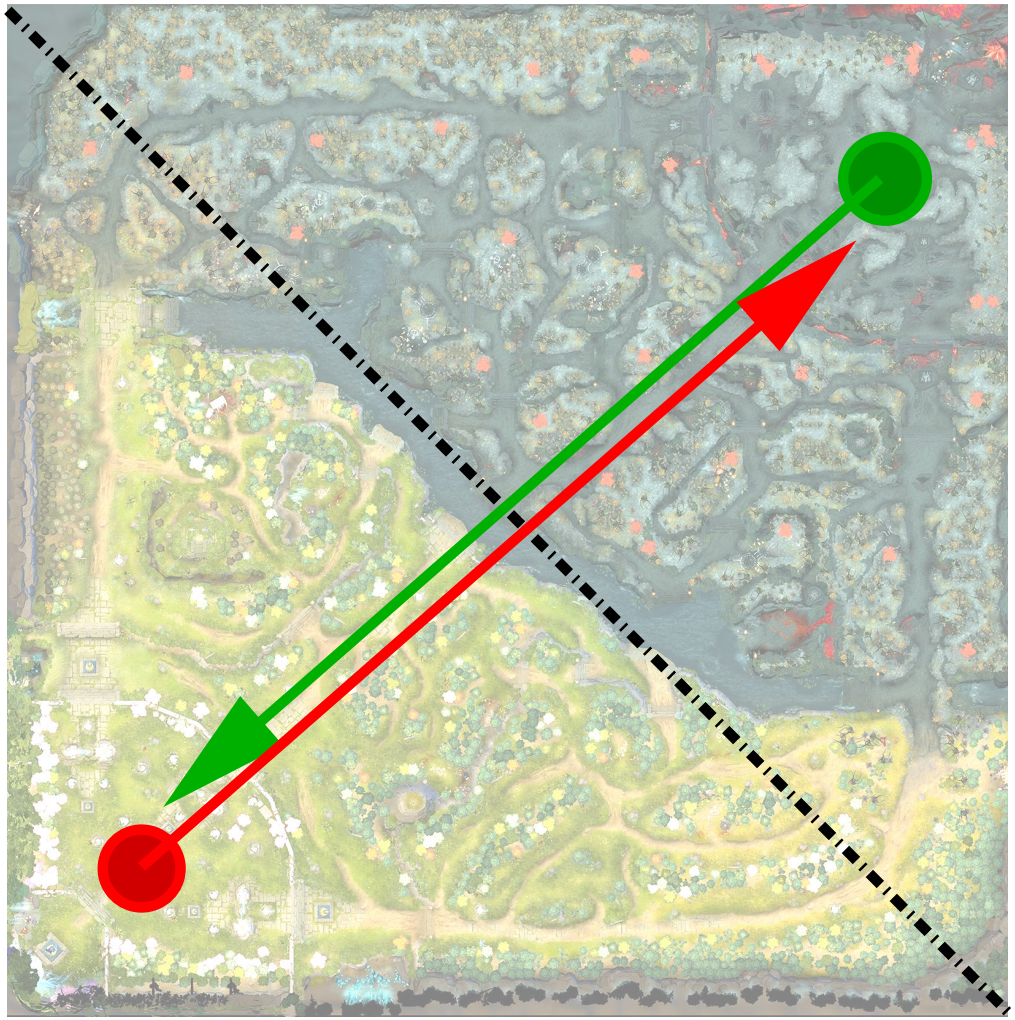
Se positionner sur la bissectrice de l'angle des possibles.



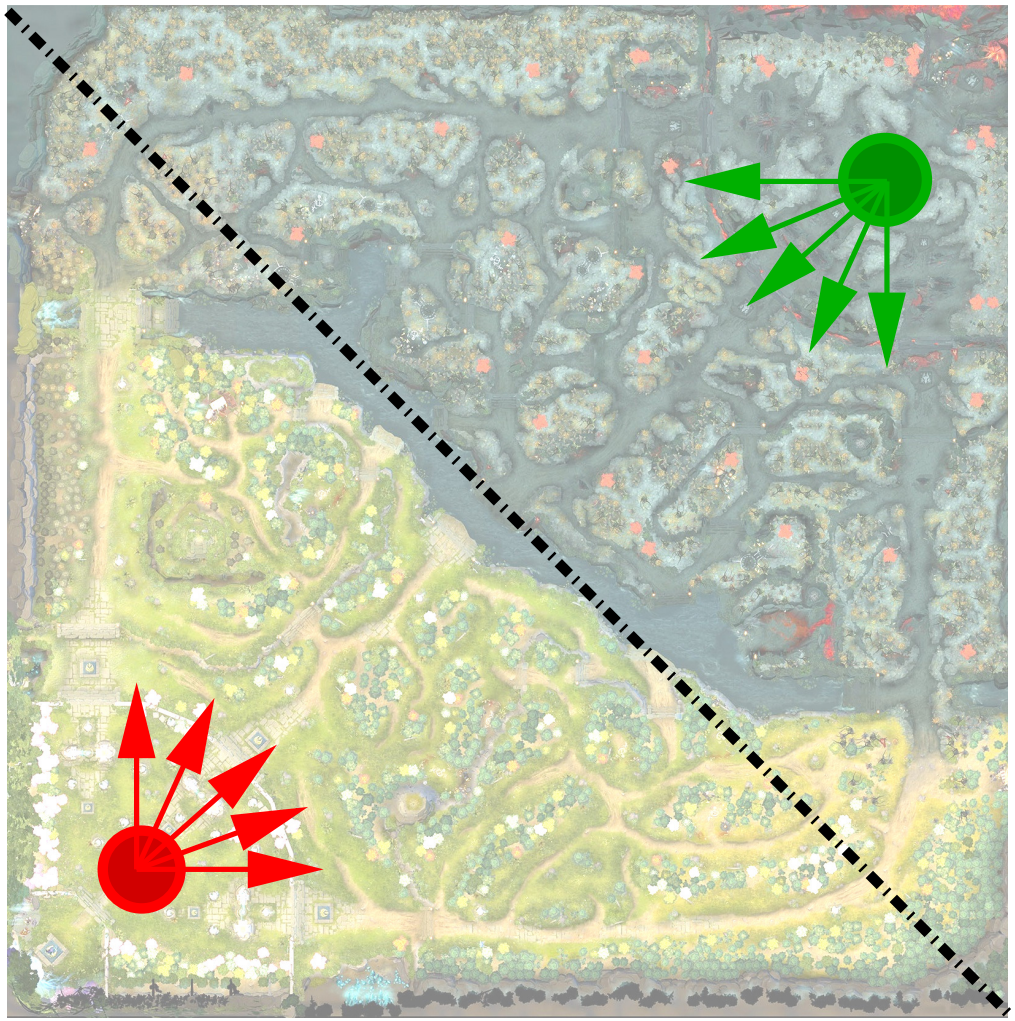
Exemple 3 : Tennis



Exemple 4 : DotA (Defense of the Ancients)



Exemple 4 : DotA (Defense of the Ancients)

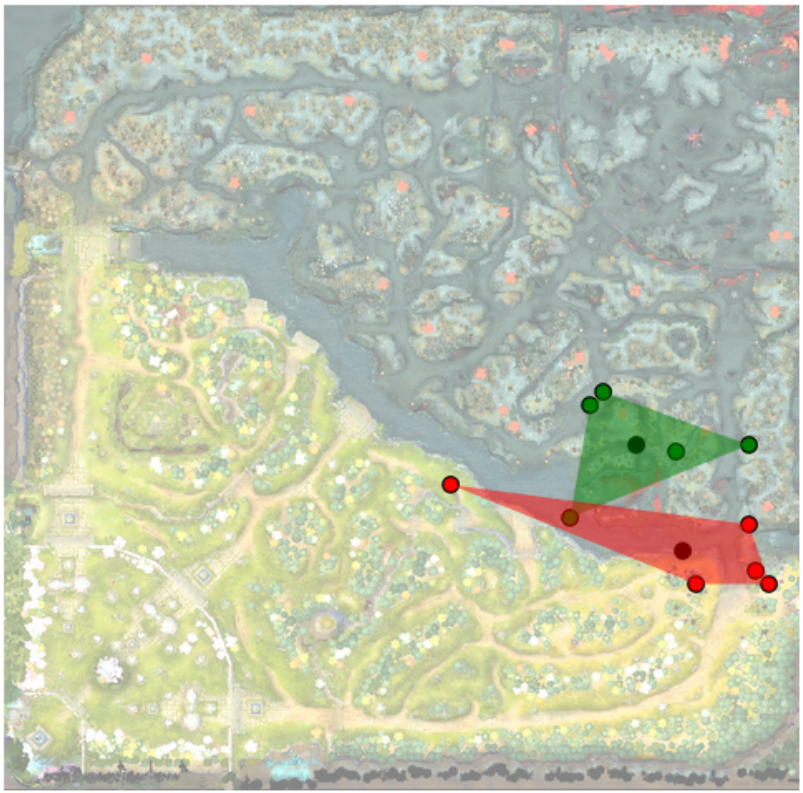


Motivations pour l'étude du e-sport

- un loisir confidentiel
- professionnalisation
- fort potentiel de spectacle
- numérisation d'interactions collectives
- de nombreuses données
- de nombreux problèmes

➥ Pourquoi se priver ?

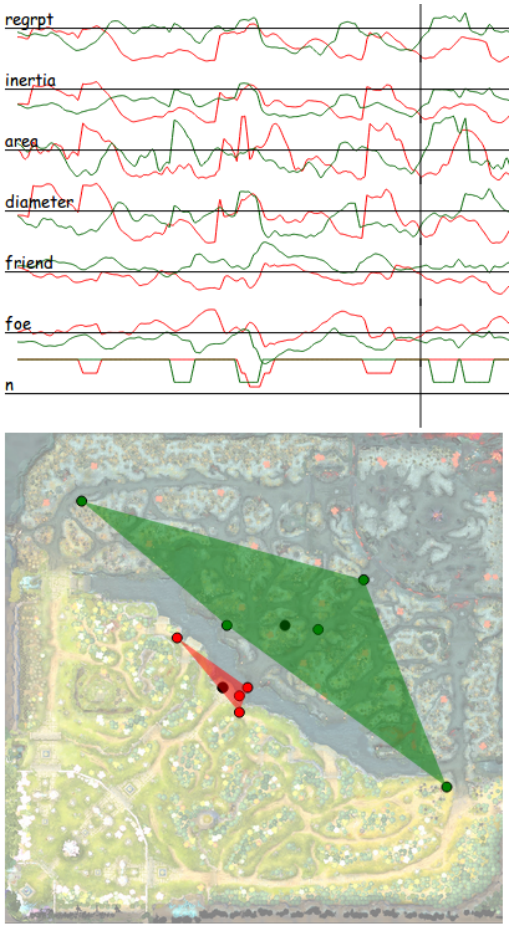
Analyse des propriétés du polygone

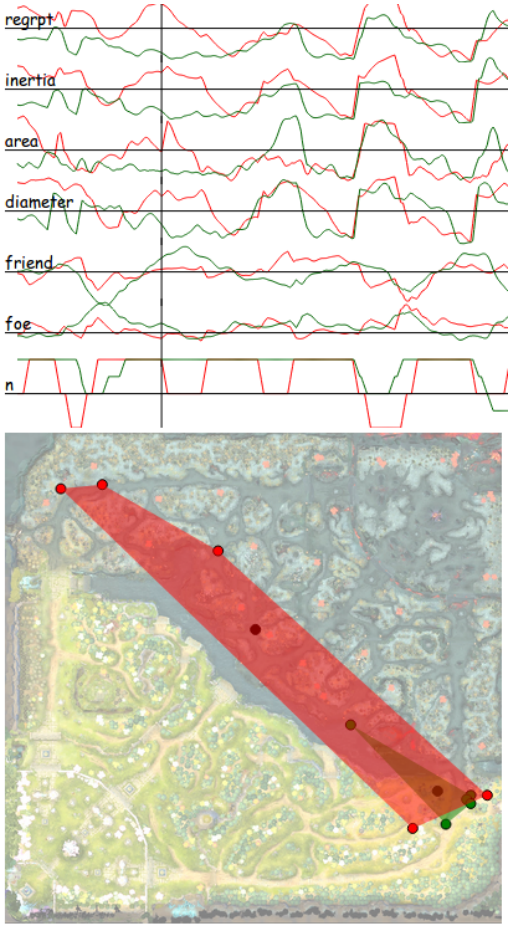


- capacité de regroupement
  vitesse, ordre 1
- inertie
  accélération, ordre 2
- diamètre
- distance à la cible

Prédiction de l'issue du match entre 10' et 20' : 90% AUC
[RNACL-RIOULT-2012 ; Rioult201482]

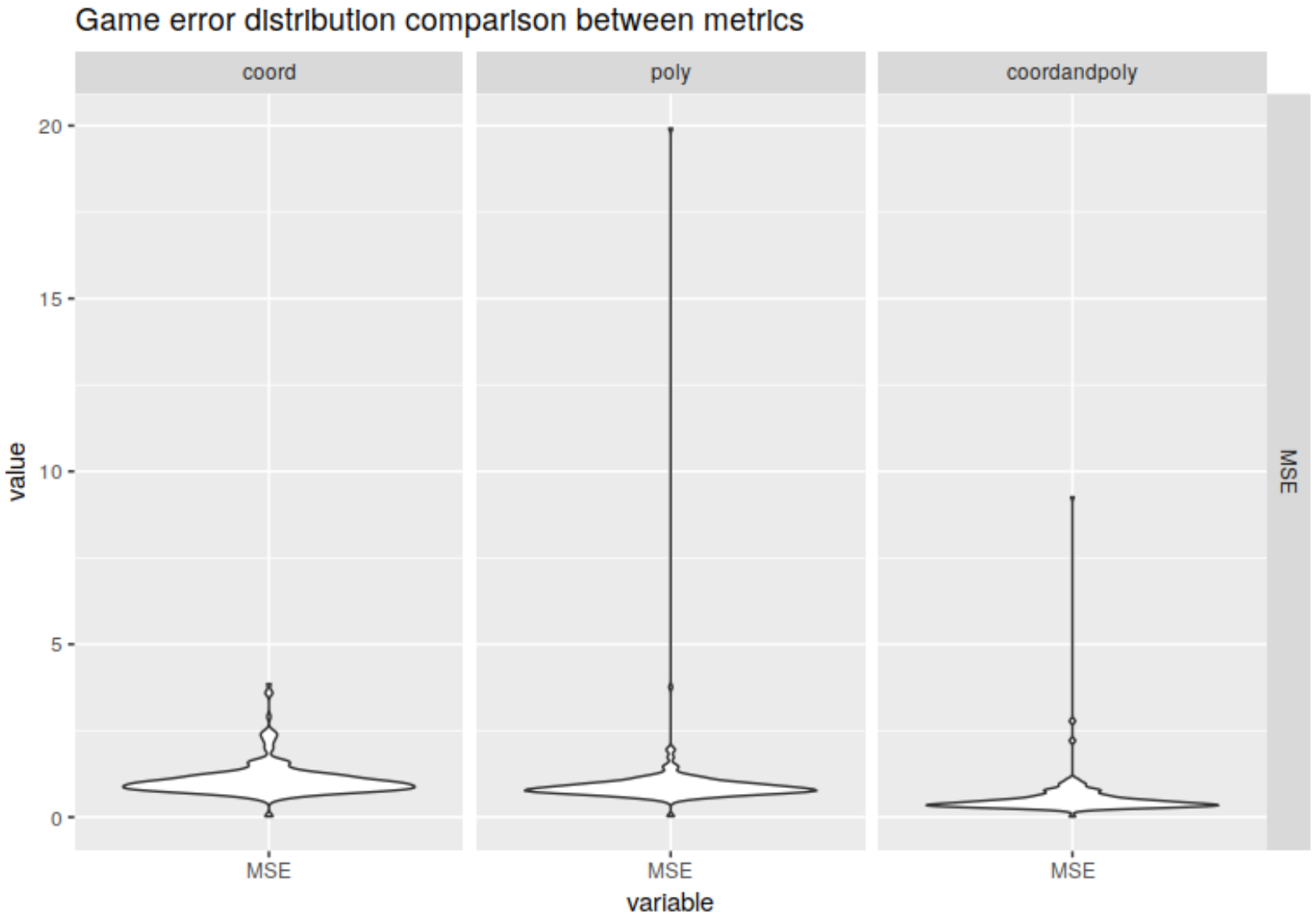
Apport du polygone vs. coordonnées



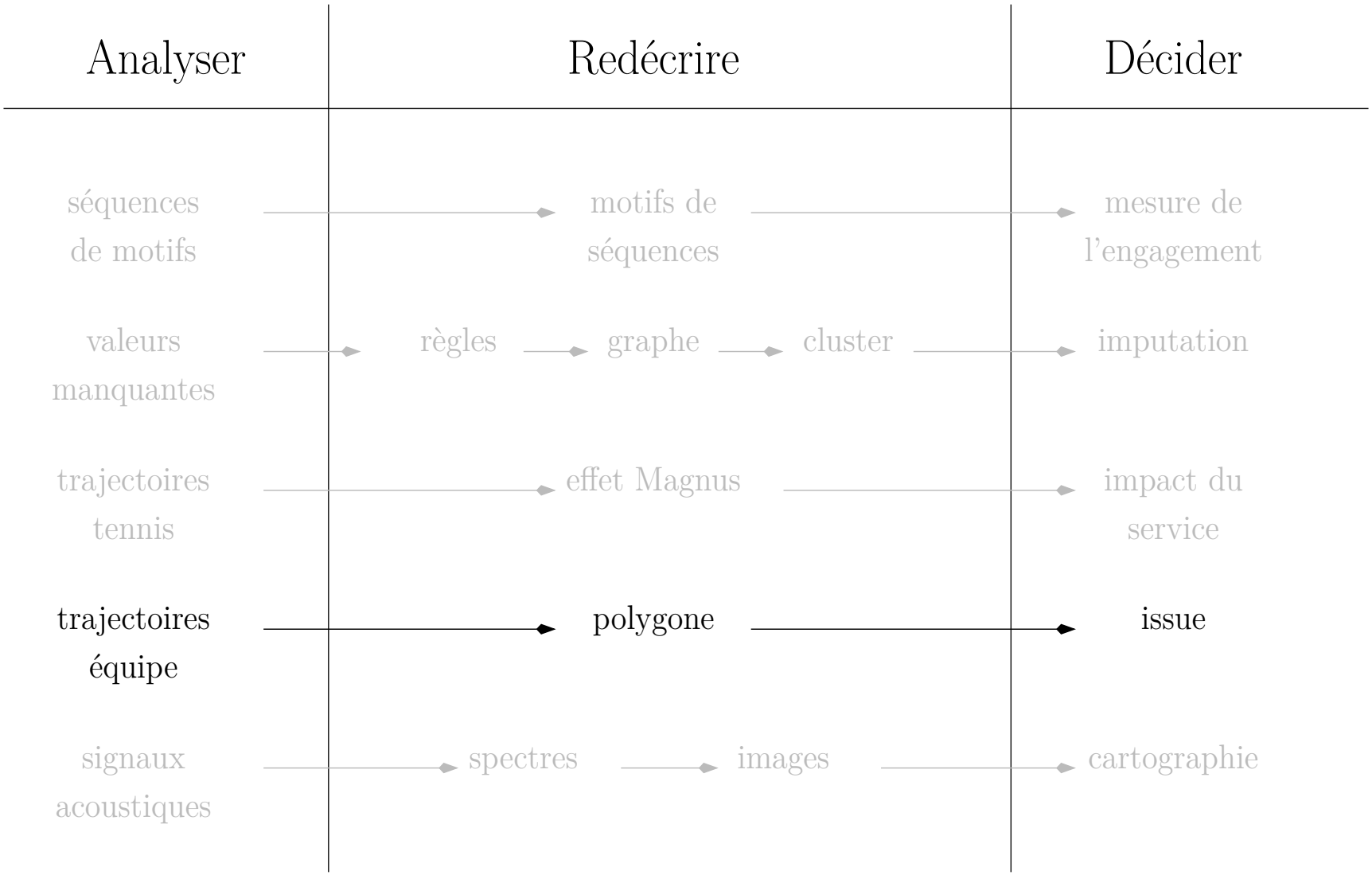


Démo

Apport du polygone vs. coordonnées

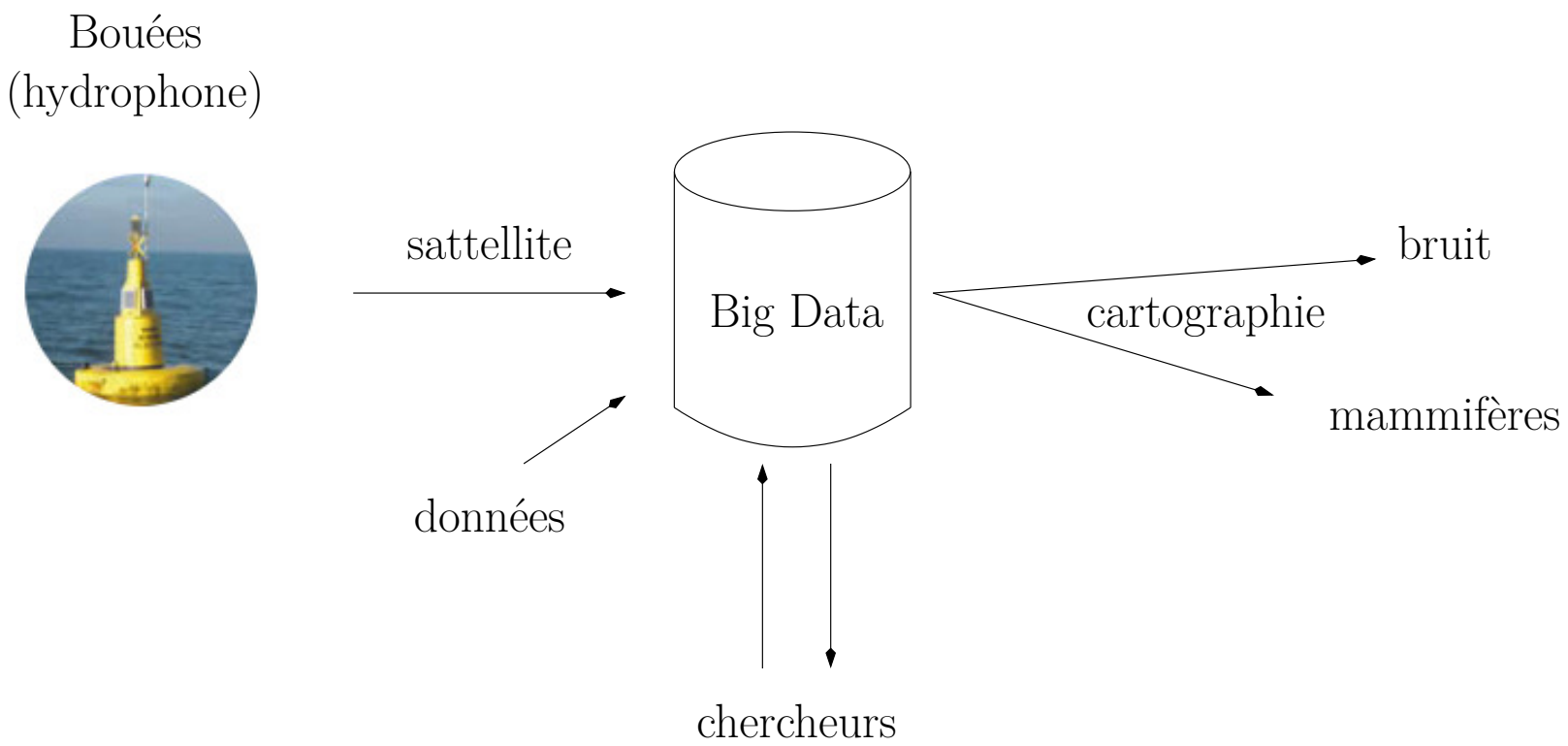


Exemple 4 : e-sport



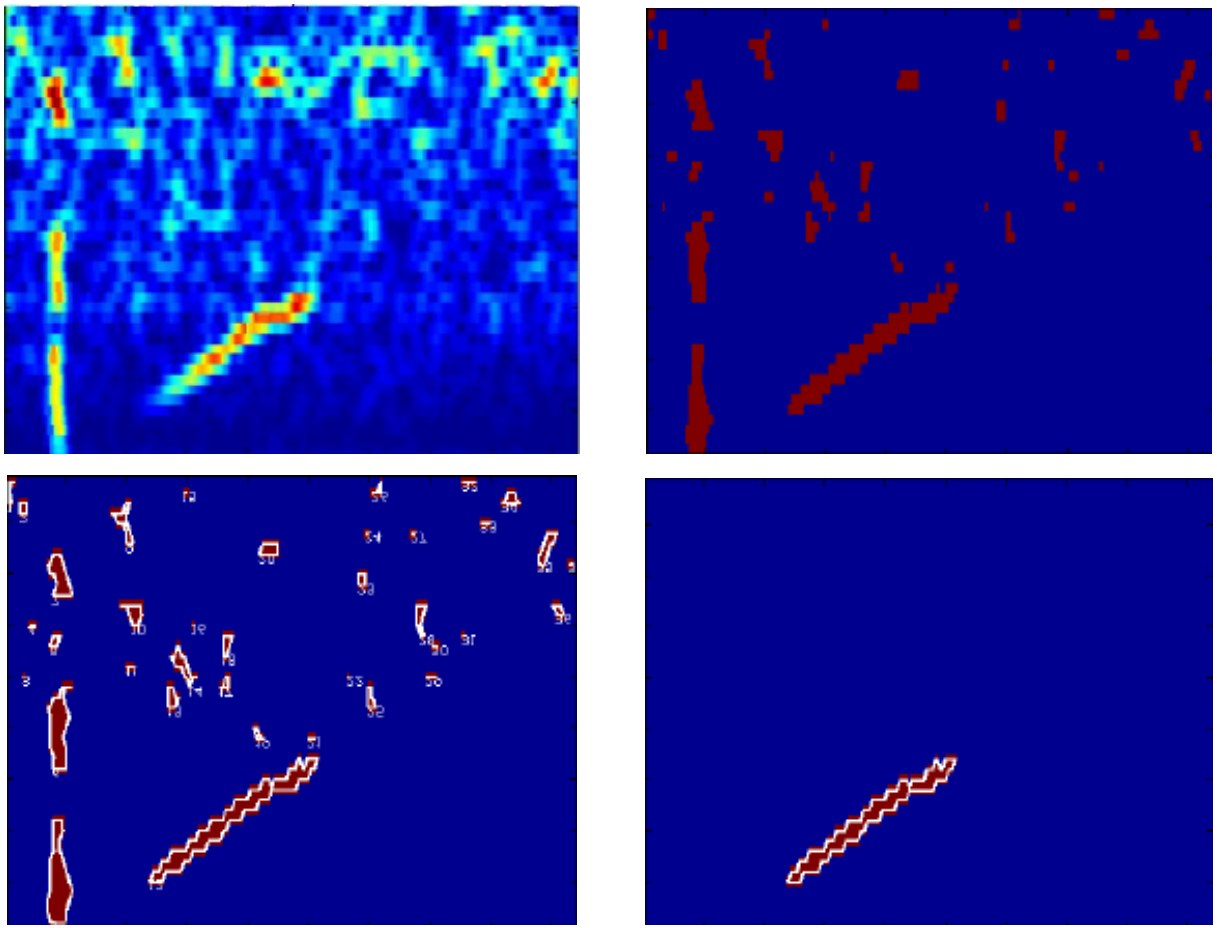
FEDER : Big Data for Sea Sounds

Automated Monitoring System (Automatic Identification System (AIS))



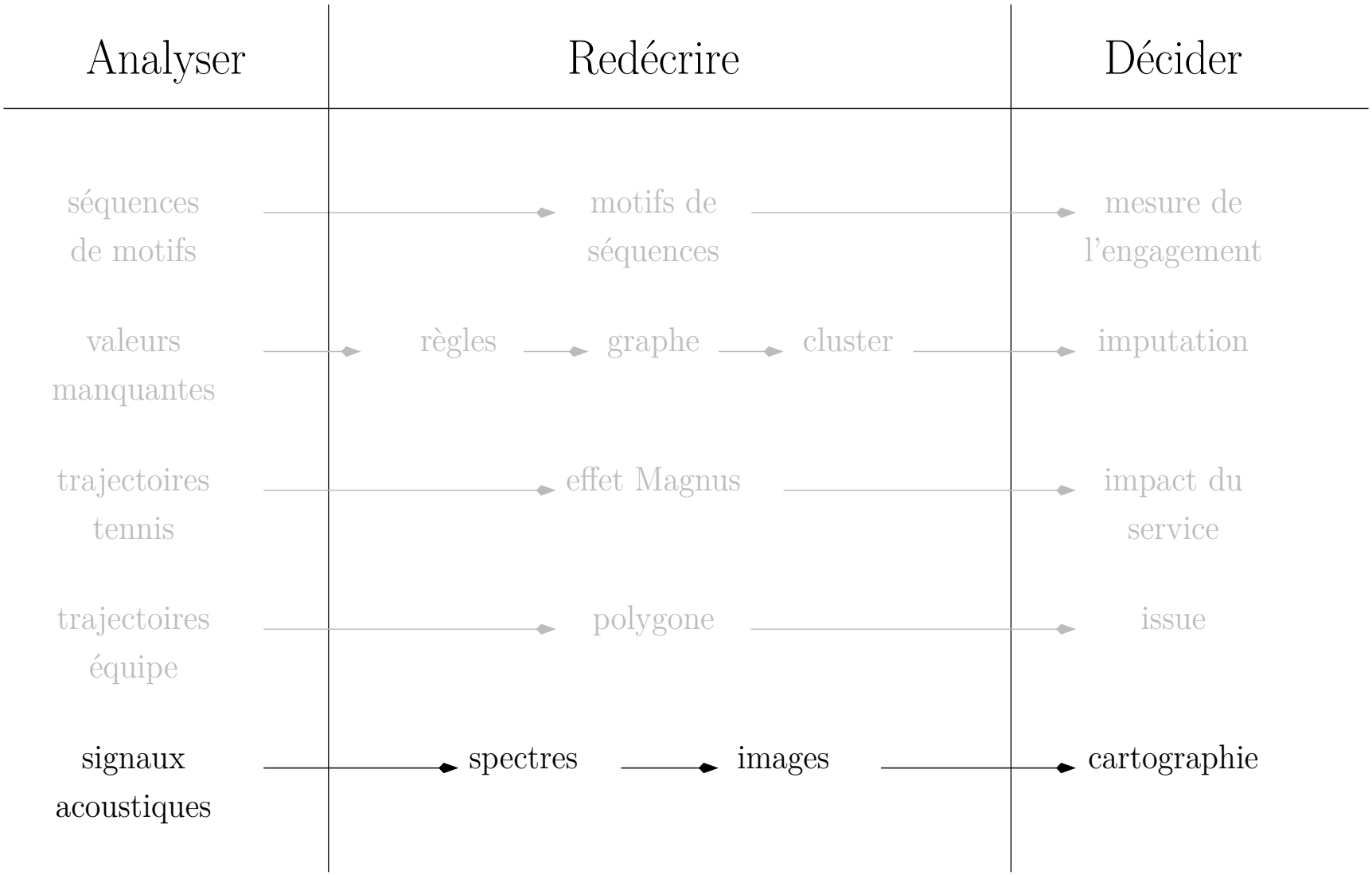
FEDER `http://aims.sinay.fr`, 2 ans de post-doc, Emna Hachicha

FEDER : Big Data for Sea Sounds



(source : [**DBLP:journals/corr/abs-1305-3635**])

FEDER : Big Data for Sea Sounds

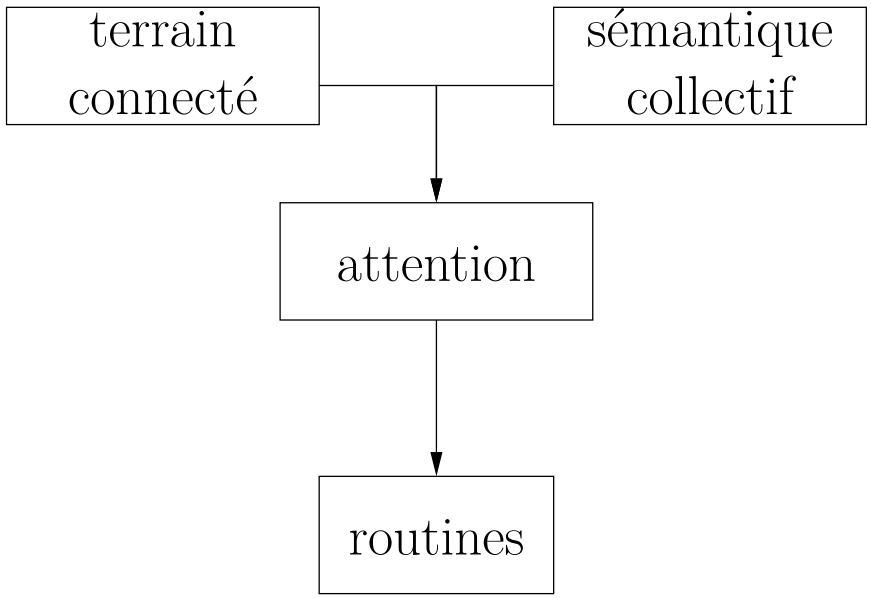


Conclusion

- des recherches en fouille de données
- redescription
  - variété des données
  - transformations
  - uniformité de la décision
  - intégrer la fouille
- une démarche bottom-up

Perspectives

« Sémantique de l'interaction dans le jeu collectif »



Références I