Apprendre des mathématiques and... What else ?

Sébastien Déjean

Université Paul Sabatier, Toulouse III
www.math.univ-toulouse.fr/˜sdejean

22 mars 2016

Mathematics. What else ?"
Mathematics

1. Formation
   - Curriculum Vitae
   - Expérience professionnelle
2. Métier
   - Dans les papiers
   - Dans les faits
3. Activités
   - Analyse de données transcriptomiques
   - Analyse textuelle de documentations techniques
   - Simulation de processus (multi-)fractionnaire
   - Un projet de paléo-anthropologie
   - Modélisation de l'arcade dentaire

INSTITUT de MATHEMATIQUES de TOULOUSE

- **1993** - Baccalauréat série C, mathématiques et sciences physiques, mention AB, (merci l'option latin !)
- **1994** - Diplôme d'Études Universitaires Généralisées (1ère année) Mathématiques et Informatique
- **1994-1997** - Institut Universitaire Professionnalisé, *Mathématiques Industrielles, Calcul Scientifique et Statistique* (DEUG, licence, maîtrise)
- **1998** - Diplôme d'Études Approfondies en mathématiques appliquées
- **2002** - Doctorat en mathématiques appliquées

Stages ingénieur

- Caisse Régionale d'Assurance Maladie (4 mois) - Analyse du calcul du taux de cotisation des accidents du travail
- Siemens Automotive (10 mois) - Modélisation du comportement d'un moteur à essence en vue de réduire la durée d'essais sur banc-moteur

Recherche

Institut National de la Recherche Agronomique - Comparaison de procédures d'estimation dans le cadre des modèles non linéaires à paramètres aléatoires : application à la modélisation de l'évolution temporelle de l'indice de surface foliaire de cultures observées par télédétection spatiale.

Enseignement

Statistique, informatique : IUT, école d'ingénieurs, formation professionnelle.

INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

THÈSE

de Doctorat de l'Université Paul Sabatier - Toulouse III
Spécialité : statistique

COMPARAISON DE PROCÉDURES D'ESTIMATION DANS LE CADRE DES MODÈLES NON LINÉAIRES À PARAMÈTRES ALÉATOIRES

Application à la modélisation de l'évolution temporelle de l'indice de surface foliaire de cultures observées par télédétection spatiale

Sébastien DEJEAN

ANNÉE : 2002

DÉPARTEMENT DE BIOMÉTRIE ET INTELLIGENCE ARTIFICIELLE

Soutenue le 04 juillet 2002 devant la commission d'examen :

Jean-Marc AZAÏS	Directeur et Président
Didier CONCORDET	Examinateur
Gérard DEDIEU	Examinateur
Robert FAIVRE	Examinateur
Christian LAVERGNE	Rapporteur
Michel MESTE	Directeur
Hervé MONOD	Rapporteur

INRA

Ingénieur de recherche en calcul scientifique

Définition de l'emploi type d'après le Référentiel des emplois-types de la recherche et de l'enseignement supérieur (REFERENS) :

L'ingénieur de recherche en calcul scientifique analyse, dans le cadre de projets de recherche, un problème théorique ou une situation d'expérience et d'observation. Il recherche les méthodes d'analyse, conçoit et optimise les outils permettant le traitement du problème. Il s'assure de la pertinence des résultats obtenus.

INSTITUT de MATHÉMATIQUES de TOULOUSE

Statisticien

Équipe de Statistique et Probabilités
Institut de Mathématiques de Toulouse
Université de Toulouse
Centre National de la Recherche Scientifique (CNRS)

Activités au sein de l'IMT

- Recherche appliquée en statistique et probabilités
- Interface entre chercheurs de disciplines différentes
- Domaines privilégiés : biologie, informatique (recherche d'information)

Recherche appliquée en statistique et probabilités

✧ ✧ ✧

Cinq exemples :

- Analyse de données transcriptomiques
- Analyse textuelle de documentations techniques
- Simulation de processus (multi-)fractionnaire
- Un projet de paléo-anthropologie
- Modélisation de l'arcade dentaire

Analyse de données transcriptomiques

En collaboration avec des biologistes de l'Institut National de la Recherche Agronomique (INRA) et de l'Institut National de la Santé Et de la Recherche Médicale (INSERM).

www.math.univ-toulouse.fr/biostat

Thématiques biologiques abordées :

- Fermentation de la levure dans différentes conditions de température et au cours du temps
- Étude de nutrition chez la souris
- Analyse transcriptomique du cancer pancréatique humain
- Recherche sur les obésités
- Folliculogénèse chez la truie
- *S. meliloti* en culture et en symbiose avec la luzerne
- ...

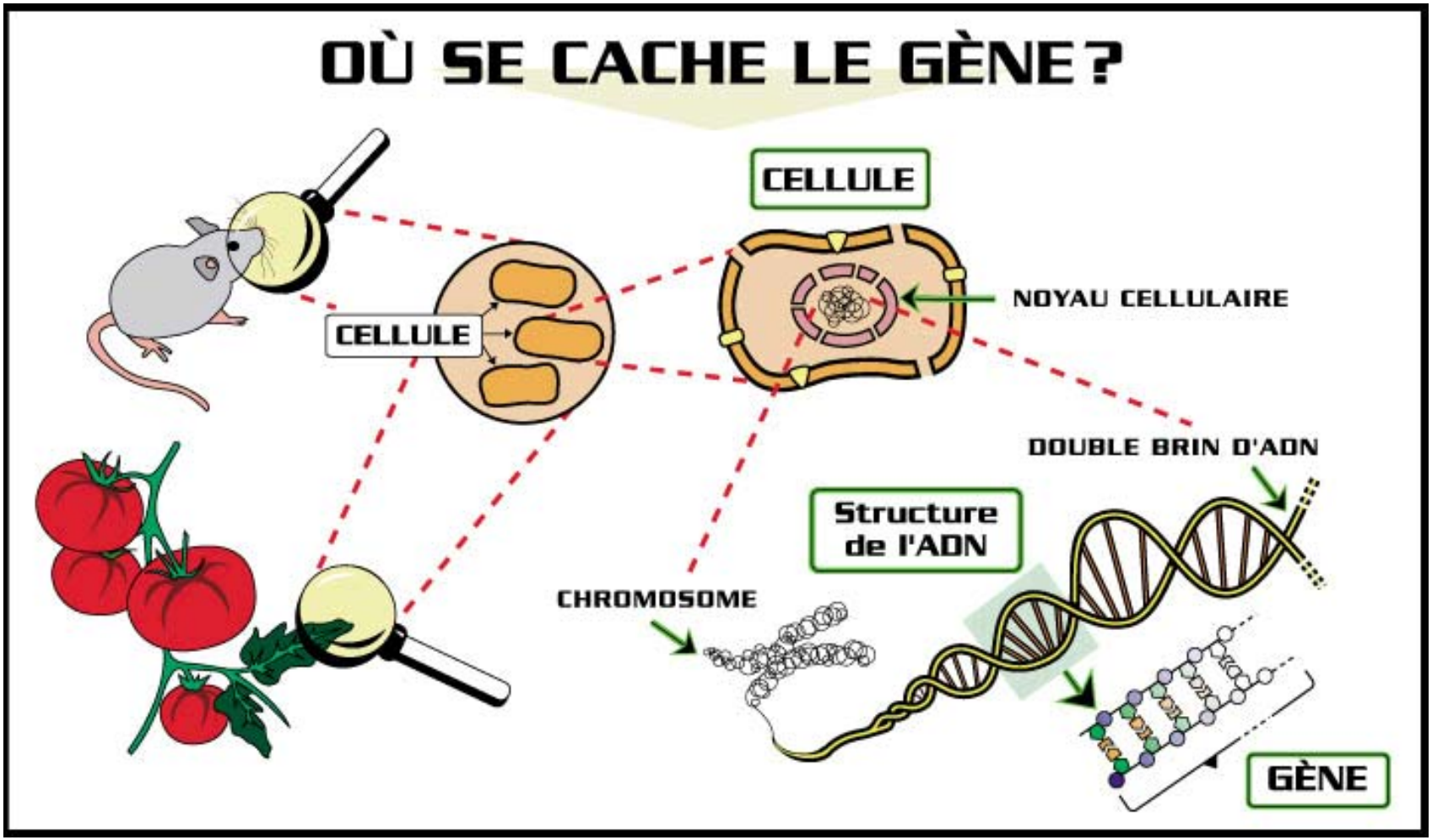
☞ Point commun : technologie des **biopuces** (*microarrays*, un peu dépassée aujourd'hui par les technologies de séquençage), génération d'une quantité importante de données.

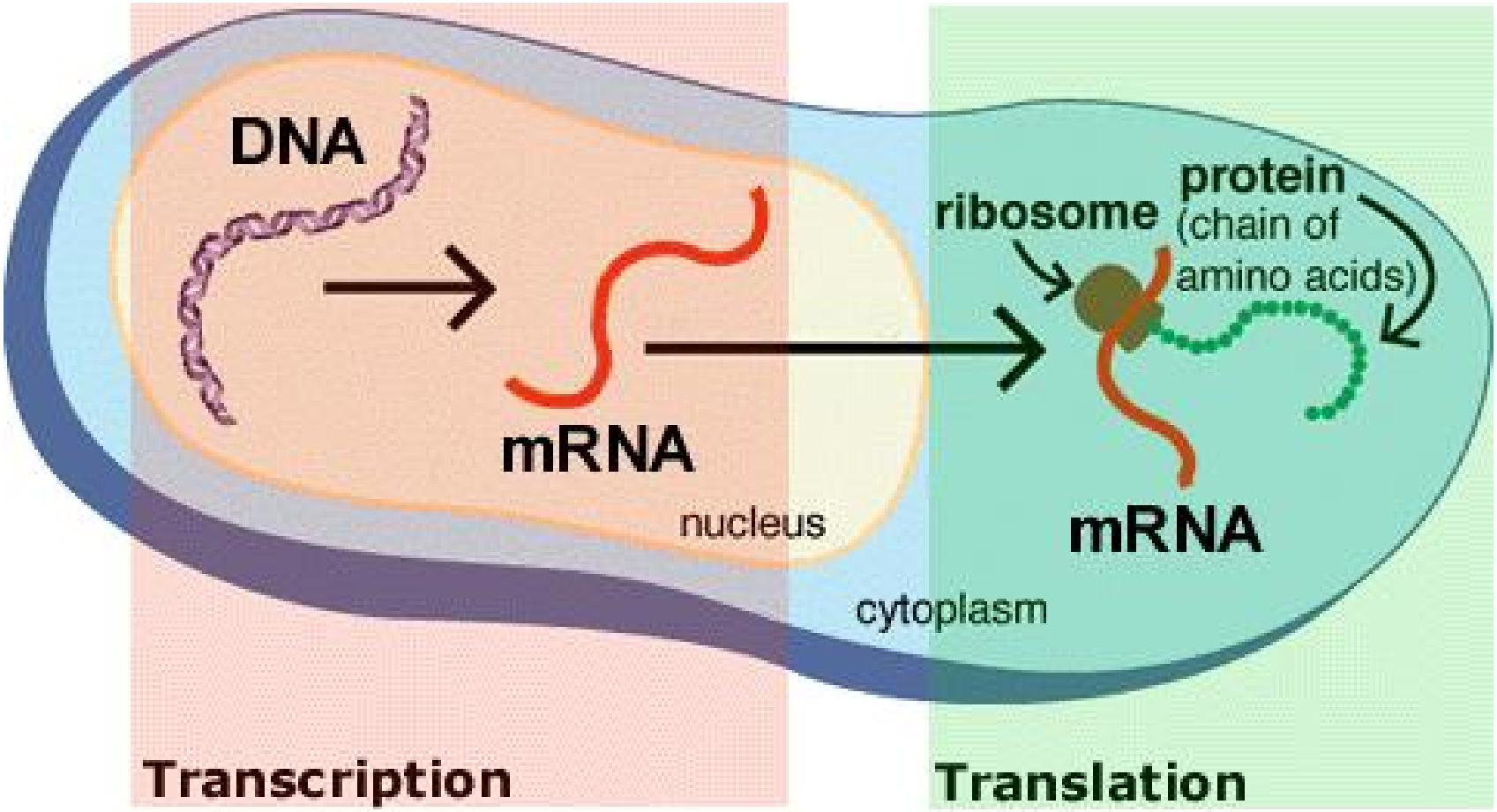
Biopuce ?

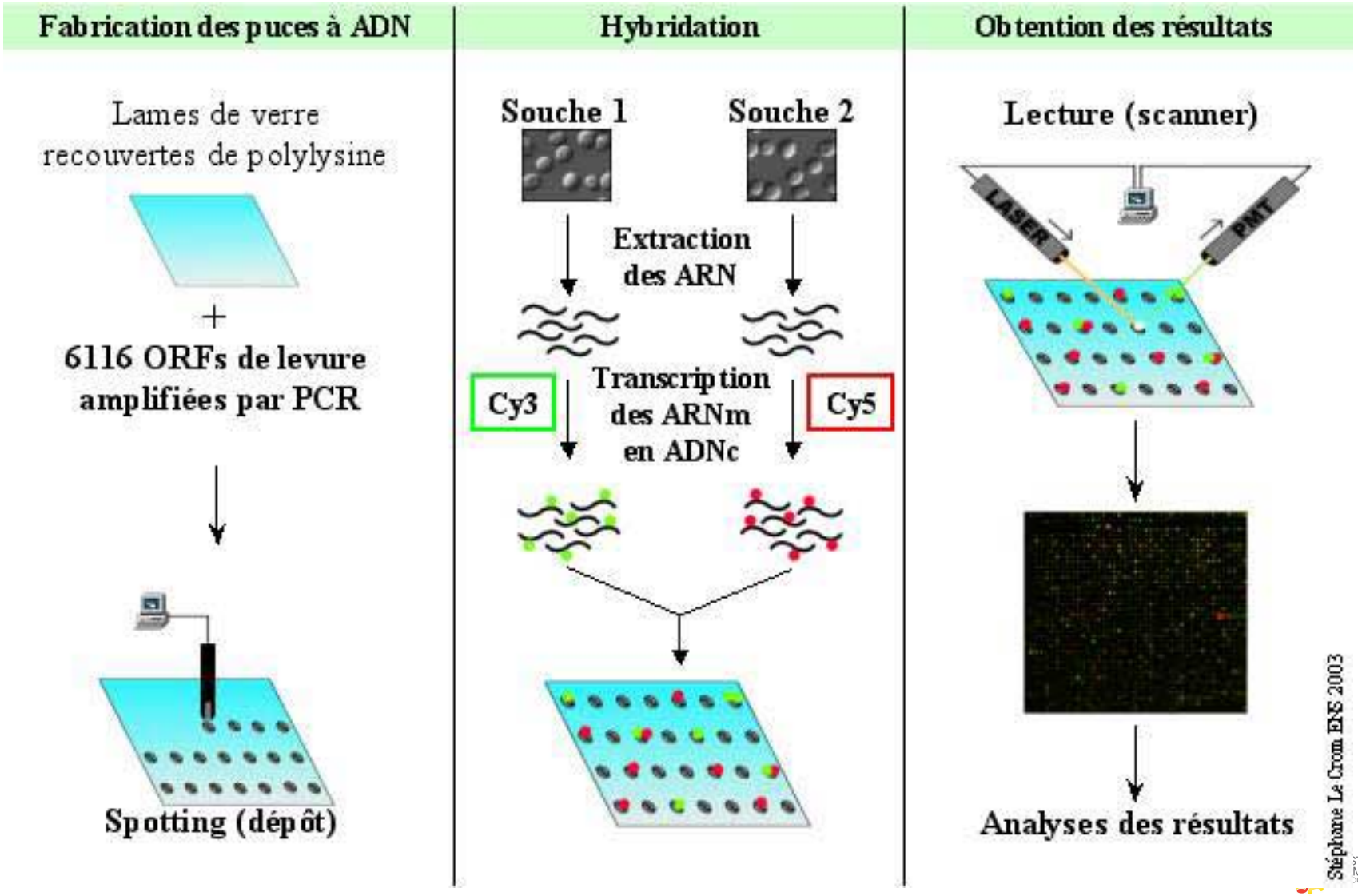
Biopuce ?

Biopuce ?

Rien à voir avec les biopuces


OÙ SE CACHE LE GÈNE ?
CELLULE
CELLULE
NOYAU CELLULAIRE
DOUBLE BRIN D'ADN
Structure de l'ADN
CHROMOSOME
GÈNE


DNA
mRNA
nucleus
ribosome
protein
(chain of
amino acids)
mRNA
cytoplasm
Transcription
Translation


Fabrication des puces à ADN
Hybridation
Obtention des résultats
Lames de verre recouvertes de polylysine
+
6116 ORFs de levure amplifiées par PCR
Spotting (dépôt)
Souche 1
Souche 2
Extraction des ARN
Cy3
Transcription des ARNm en ADNc
Cy5
Lecture (scanner)
LASER
PMT
Analyses des résultats
Stéphane Le Crom ENS 2003

Extrait de données

Étude sur le cancer pancréatique humain

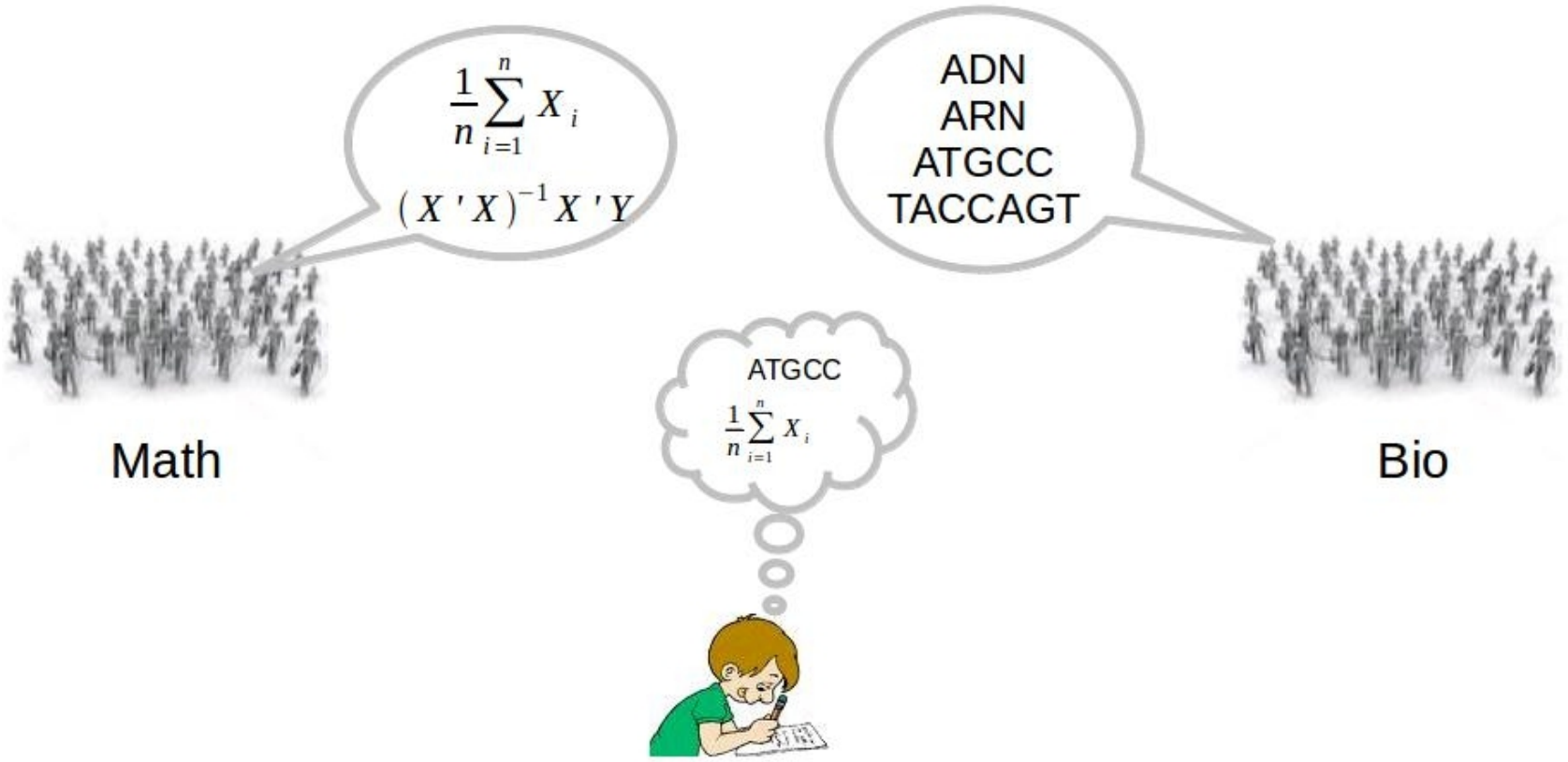
	ASPC1	Bx-PC3	CAPAN1	CAPAN2	NP29	PANC1	MIA-PaCa2	PT1	PT2	PT3	PT4	PT5	PT6	CACO2	....
MAPRE1	1,838	1,736	1,523	2,062	1,353	2,488	2,319	-0,133	0,086	0,555	-0,036	0,238	1,279	2,551	
VIL2	1,458	1,687	1,429	0,788	0,605	0,736	2,243	0,02	0,745	0,25	-0,267	0,19	1,606	0,999	
NME2	3,82	4,452	4,966	4,719	4,031	4,912	5,252	2,958	3,167	3,11	2,743	2,327	3,641	4,141	
NME1	1,819	2,069	3,088	2,648	2,346	3,609	2,85	0,489	1,423	0,53	0,616	0,877	1,353	2,485	
MARK3	0,962	0,363	0,933	1,082	0,446	1,108	0,786	0,004	-0,045	-0,289	0,134	0,193	0,585	1,101	
JUN	2,157	1,417	0,887	-0,204	1,402	1,898	3,404	2,877	2,151	3,219	0,591	2,398	3,606	-0,054	
MYC	2,852	2,965	3,32	2,69	2,997	2,009	3,856	0,376	0,941	1,981	1,225	1,582	1,274	3,028	
FOSL1	2,342	1,996	2,233	1,345	1,963	3,229	3,36	-0,065	0,171	0,812	0,596	-0,774	-0,216	-1,167	
JUNB	-0,486	-0,046	-0,179	-0,649	-0,035	-0,757	-0,642	0,399	0,499	0,56	0,368	-0,3	1,208	-1,231	
AXL	0,741	1,194	-0,433	-0,513	0,326	1,353	1,358	0,018	1,122	0,358	0,501	0,362	1,281	-1,012	
ERBB3	2,733	2,499	2,727	2,35	2,503	1,29	3,142	1,555	0,928	2,503	0,619	0,443	1,877	2,449	
FLT1	2,023	2,674	3,294	3,043	2,686	3,287	3,762	2,178	1,402	2,282	0,679	0,94	2,677	2,436	
...															

Le tableau complet a 868 lignes et 22 colonnes (ce qui n'est pas très volumineux, sans commune mesure avec l'ère du **Big Data**).

Démarche

- Discussions avec les biologistes : quel est le problème ?
- Étude préliminaire du jeu de données : analyse exploratoire élémentaire
- Discussions avec les biologistes : ce que nous trouvons est-il interprétable ?
- Nouvelles questions plus précises
- Mise en œuvre de techniques statistiques spécifiques (discrimination, étude de données cinétiques...)

Résumé en une image



Analyse textuelle de documentations techniques

Dans le cadre d'un contrat avec le Centre National d'Études Spatiales (CNES) "Méthode et Outils de Data Mining pour les Projets Spatiaux" en collaboration avec des informaticiens de l'Institut de Recherche en Informatique de Toulouse (IRIT) et des linguistes de l'Université de Toulouse 2 Le Mirail.

Objectif : proposer une méthode pour la supervision de l'ensemble de la documentation technique générée pour un projet.

Démarche

- Données initiales : corpus de documents (quelques dizaines)
- *Lemmatisation* des documents
- Construire des données numériques (table de contingence)
- Analyse statistique

☞ Lemmatisation : opération consistant à regrouper les formes occurrentes d'un texte ou d'une liste sous des adresses lexicales.

☞ Lemmatisation : opération consistant à regrouper les formes occurrentes d'un texte ou d'une liste sous des adresses lexicales.

? ? ?

Maître Corbeau, sur un arbre perché,
Tenait en son bec un fromage.
Maître Renard, par l'odeur alléché,
Lui tint à peu près ce langage :
"Hé ! bonjour, Monsieur du Corbeau.
Que vous êtes joli ! que vous me semblez beau !
Sans mentir, si votre ramage
Se rapporte votre plumage,
Vous êtes le Phénix des hôtes de ces bois."
À ces mots le Corbeau ne se sent pas de joie ;
Et pour montrer sa belle voix,
Il ouvre un large bec, laisse tomber sa proie.
Le Renard s'en saisit, et dit : "Mon bon Monsieur,
Apprenez que tout flatteur
Vit aux dépens de celui qui l'écoute :
Cette leçon vaut bien un fromage, sans doute. "
Le Corbeau, honteux et confus,

Jura, mais un peu tard, qu'on ne l'y prendrait plus.

Maître Corbeau, sur un arbre perché,
Tenait en son bec un fromage.
Maître Renard, par l'odeur alléché,
Lui tint à peu près ce langage :
"Hé ! bonjour, Monsieur du Corbeau.
Que vous êtes joli ! que vous me semblez beau !
Sans mentir, si votre ramage
Se rapporte votre plumage,
Vous êtes le Phénix des hôtes de ces bois."
À ces mots le Corbeau ne se sent pas de joie ;
Et pour montrer sa belle voix,
Il ouvre un large bec, laisse tomber sa proie.
Le Renard s'en saisit, et dit : "Mon bon Monsieur,
Apprenez que tout flatteur
Vit aux dépens de celui qui l'écoute :
Cette leçon vaut bien un fromage, sans doute. "
Le Corbeau, honteux et confus,

Jura, mais un peu tard, qu'on ne l'y prendrait plus.

maître corbeau arbre percher
tenir bec fromage
maître renard odeur allécher
tenir peu près langage
bonjour monsieur corbeau
être joli sembler beau
sans mentir ramage
rapporter plumage
être phénix hôte bois
mot corbeau sentir pas joie
montrer beau voix
ouvrir large bec laisser tomber proie
renard saisir dire bon monsieur
apprendre flatteur
vivre dépens écouter
leçon valoir bien fromage sans doute
corbeau honteux confus
jurer peu tard prendre plus

☞ Comptage des occurrences

Mots	Occurrences
maître	2
corbeau	4
arbre	1
percher	1
tenir	2
bec	2
fromage	2
renard	2
odeur	1
allécher	1
...	...

☞ Construction d'un tableau de contingence

Mots	**Doc-1**	**Doc-2**	...	**Doc-n**
balise	25	32	...	8
emission	42	21	...	24
transmettre	12	18	...	62
boitier	5	7	...	21
satellite	64	52	...	15
seconde	13	28	...	14
commande	10	21	...	7
transfert	2	41	...	37
amplification	8	7	...	14
batterie	34	12	...	2
...	...	...	...	...

En pratique : quelques dizaines de colonnes et quelques centaines de lignes.

☞ Analyse statistique

- Analyse Factorielle des Correspondances (AFC)
- Classification (*clustering*)

Résultats

- Mise en évidence de groupes de documents homogènes
- Caractérisation de ces groupes en fonction des mots les plus fréquents
- "Apprentissage" du corpus de documents sans le lire

Simulation de processus (multi-)fractionnaire

En collaboration avec des enseignants-chercheurs probabilistes de l'équipe de Statistique et Probabilités.

S. Déjean, S. Cohen (2005)
FracSim: An R package to Simulate Multifractional Lévy Motions
Journal of Statistical Software, vol. 14, issue 18.

Démarche

- *Travaux théoriques*
- Compléments pratiques : simulation
- Programmation (langage C, logiciel statistique R)
- Utilisation d'un super-calculateur (programmation parallèle)
- Visualisation (représentation graphique)

Avertissement

L'image qui va suivre est susceptible de choquer les âmes (in)sensibles... aux mathématiques.

$$Y_{h,n}(t)$$

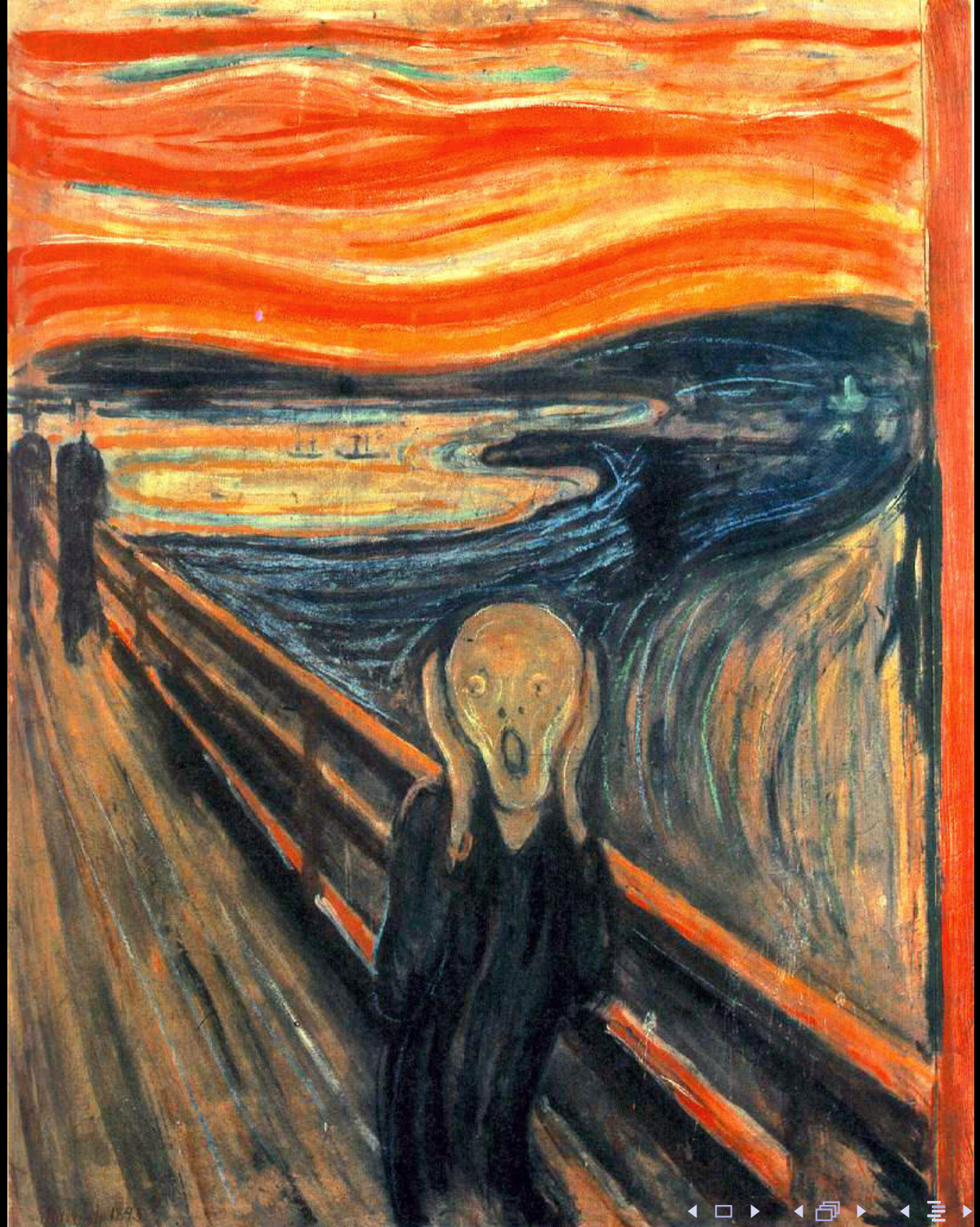
$$Y_{h,n}(t)$$

$$=$$

$$Y_{h,n}(t)$$

$$=$$

$$\sum_{l=1}^{n} \left(\frac{\cos(2\pi\theta_l) \left[\cos\left(\frac{R_l}{m} t.U_l\right) - 1\right] + \sin(2\pi\theta_l) \left[\sin\left(\frac{R_l}{m} t.U_l\right)\right]}{R_l^{h(t)+1/2}} \right)$$



$$Y_{h,n}(t) =$$

$$\sum_{l=1}^{n} \left(\frac{\cos(2\pi\theta_l)\left[\cos\left(\frac{R_l}{m}t.U_l\right) - 1\right] + \sin(2\pi\theta_l)\left[\sin\left(\frac{R_l}{m}t.U_l\right)\right]}{R_l^{h(t)+1/2}} \right)$$

La fonction de **régularité** h(t) est le paramètre-clé de la simulation.

- $\forall t \in E, 0 < h(t) < 1$
- Si $\forall t \in E, h(t) = h$, alors le processus est **fractionnaire**
- Si $h(t)$ varie quand t varie dans E alors le processus est **multi-fractionnaire**

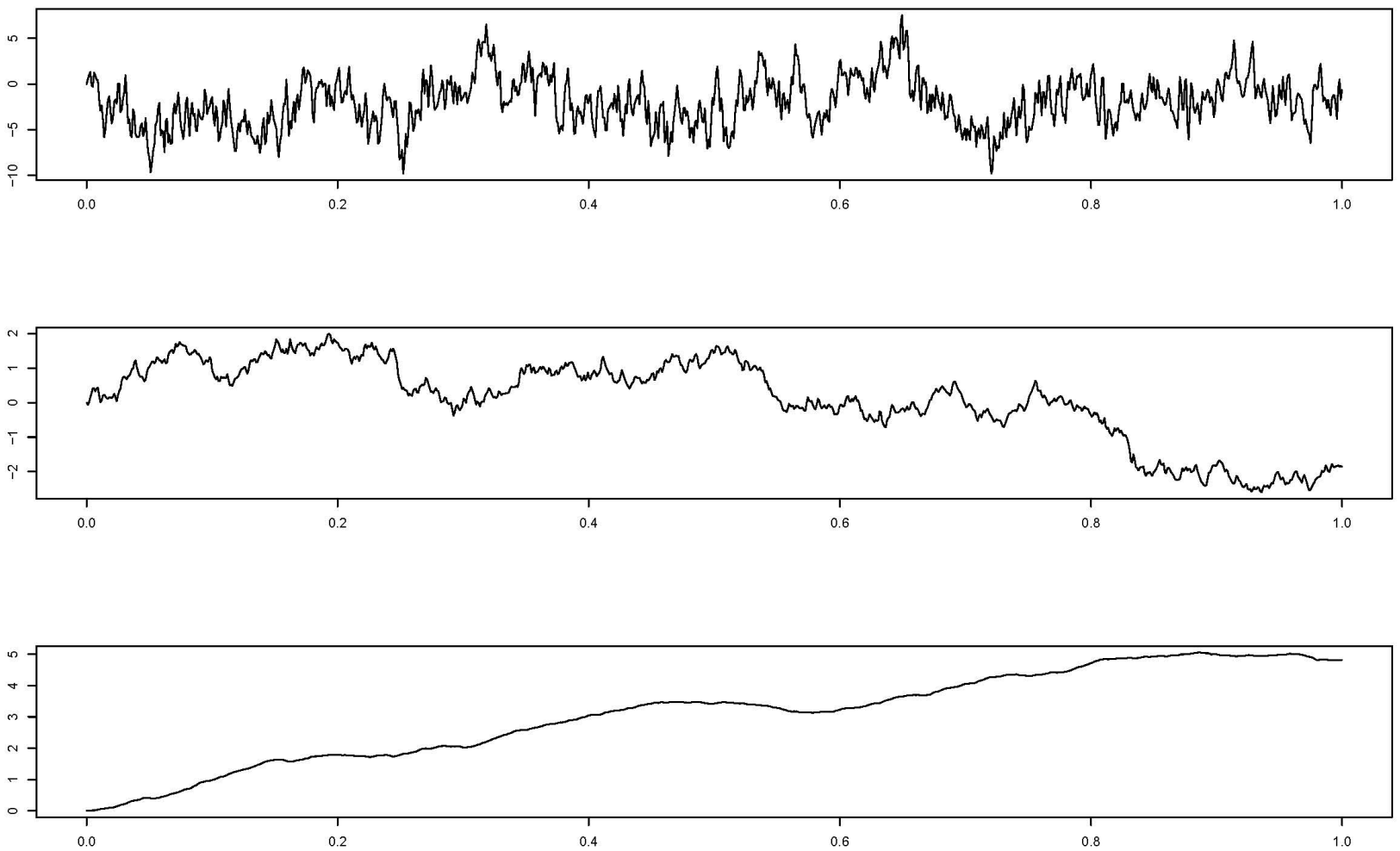
Un programme informatique

Extrait de programme en langage C

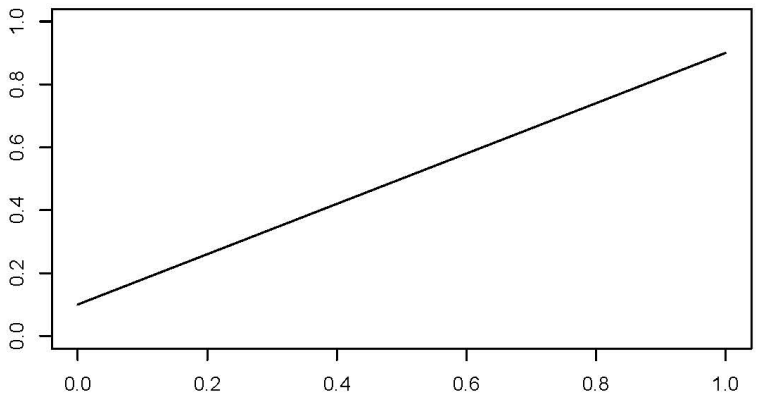
```
...
for (t=0;t<T;t++) nombre de valeurs de t, T termes
for (n=0;n<N;n++) calcul de la somme, N termes
{
      tmp1 = cos(2*M_PI*theta)*(cos(t*R*U)-1);
      tmp2 = sin(2*M_PI*theta)*sin(t*R*U);
      tmp3 = pow(R,0.5+H[t]);
      tmp = (tmp1-tmp2)/tmp3;
      res[t] = res[t] + tmp;
}
...
```

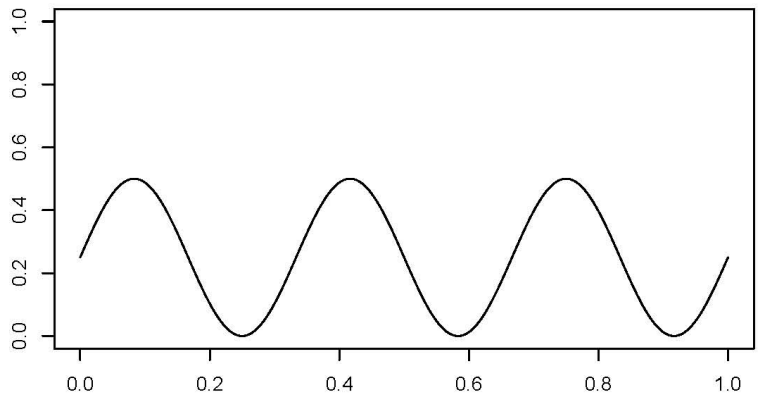
$$\sum_{n=1}^{N}\left(\frac{\cos(2\pi\theta_n)\left[\cos\left(R_n t.U_n\right)-1\right]+\sin(2\pi\theta_n)\sin\left(R_n t.U_n\right)}{R_n^{h(t)+1/2}}\right)$$

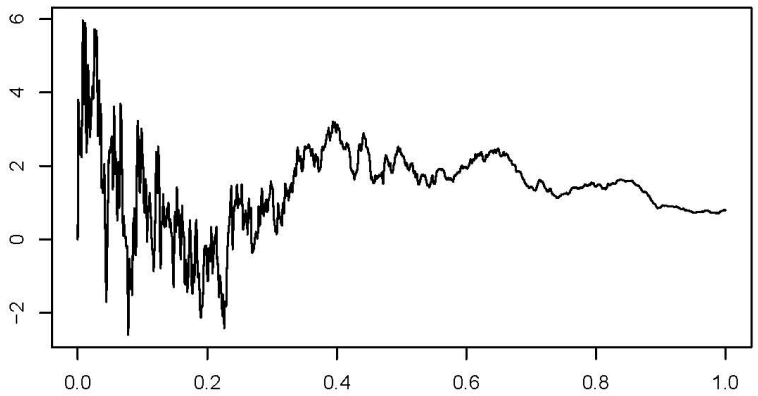
Processus fractionnaire : $h(t) = 0.1 - 0.5 - 0.9$

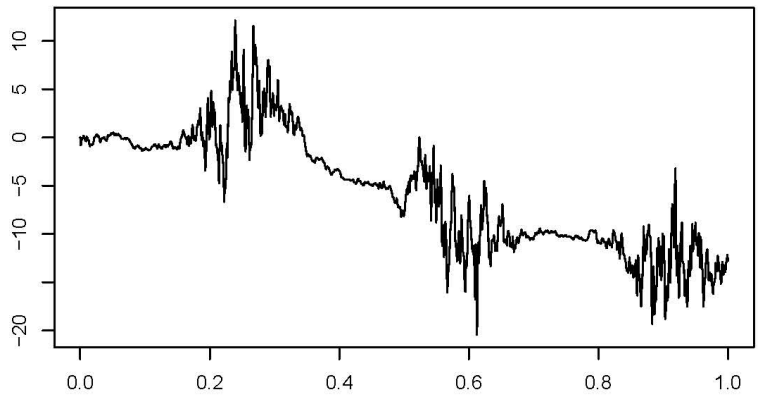


Processus multi-fractionnaire : $h(t)$ croissante - sinusoidale









Temps de calcul

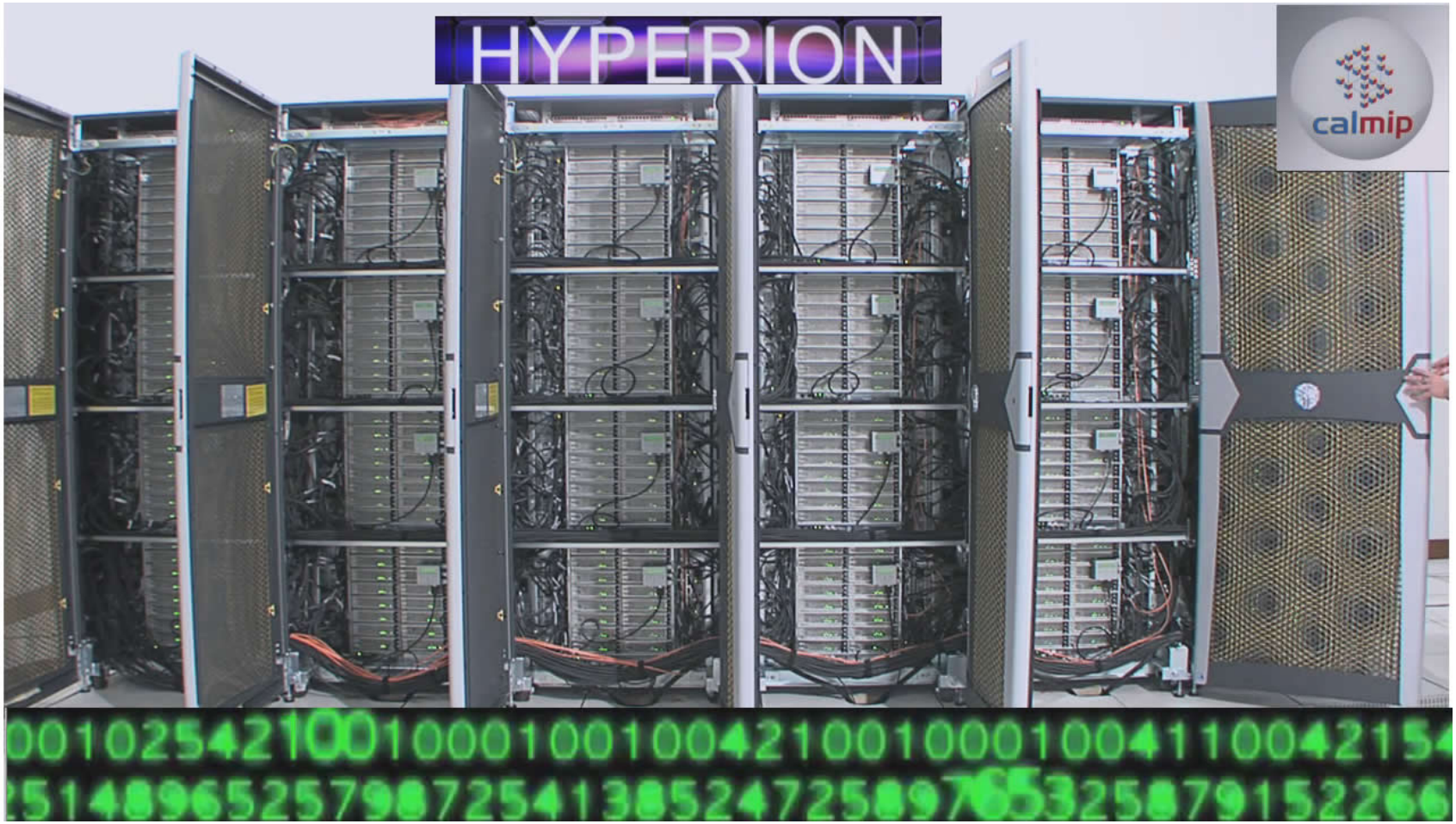
1D	t	N	Durée (s)
Config. 1	10 000	100 000	900 (15 min)
Config. 2	10 000	1 000 000	9000 (2h30)

2D	tx	ty	N	Durée (s)
Config. 1	100	100	10 000	120 (2 min)
Config. 2	100	100	100 000	1200 (20 min)
Config. 3	500	500	10 000	2700 (45 min)

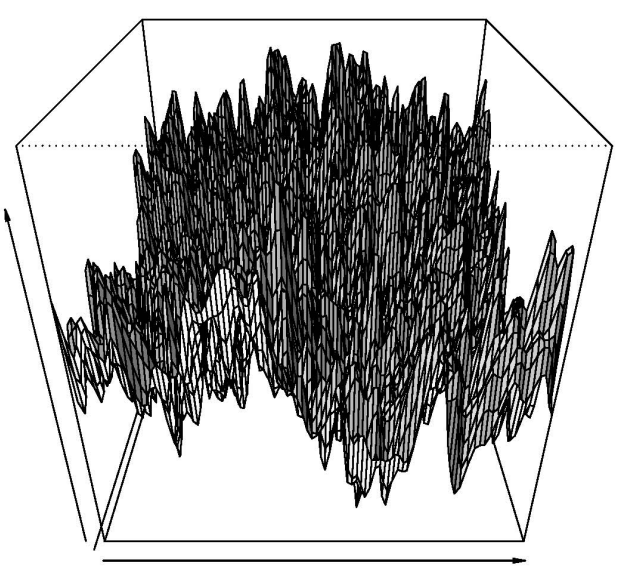
Temps estimé pour tx=ty=500 et N=10 000 000 $\rightarrow$ 27 000 000 secondes soit 7 500 heures (environ 300 jours).

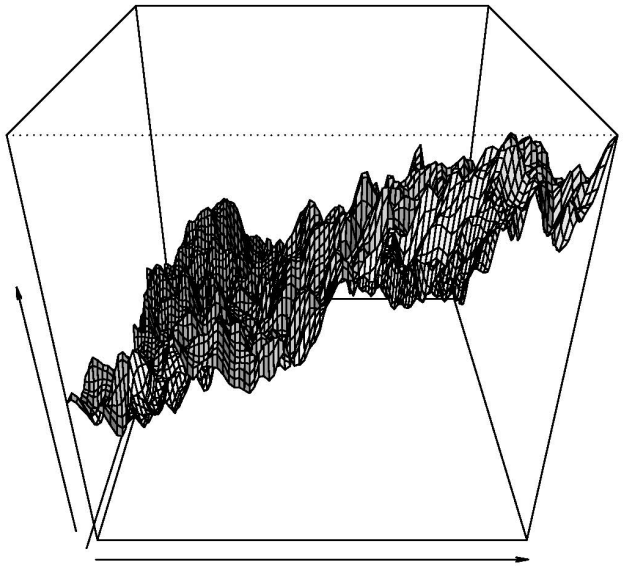
www.rasiel.com
The Oarsmen, 1877 - Caillebotte

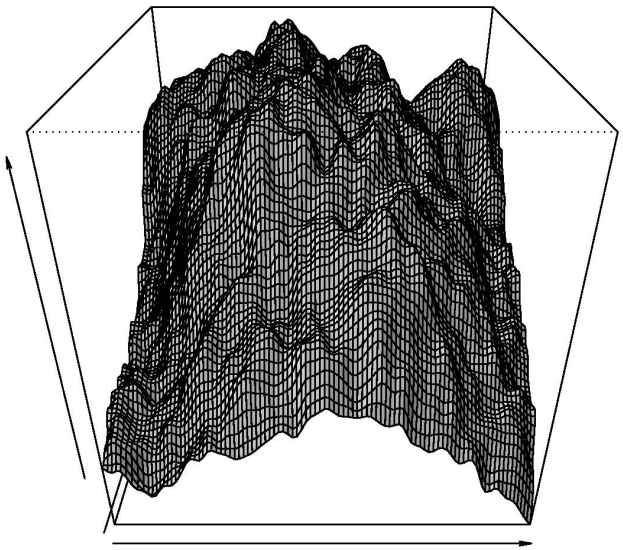
sgi
Altix ICE 8200


HYPERION
calmip

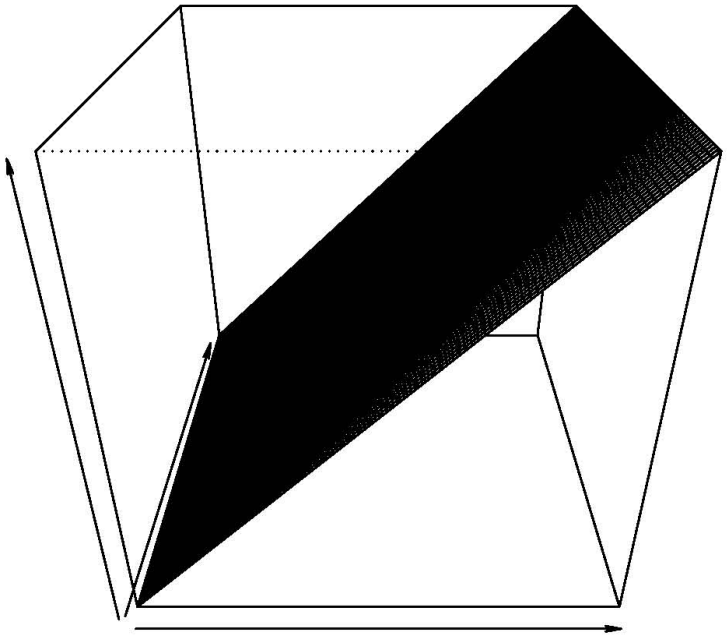
Processus fractionnaire 2D : $h(t) = 0.1 - 0.5 - 0.9$

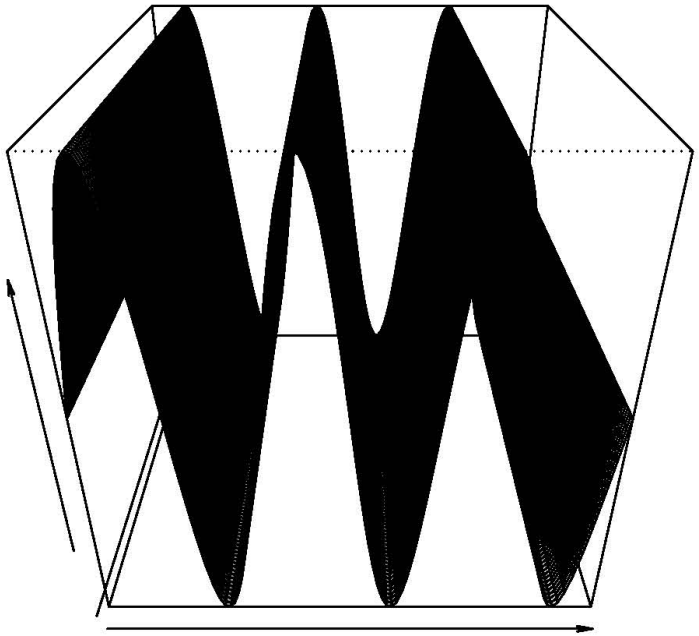




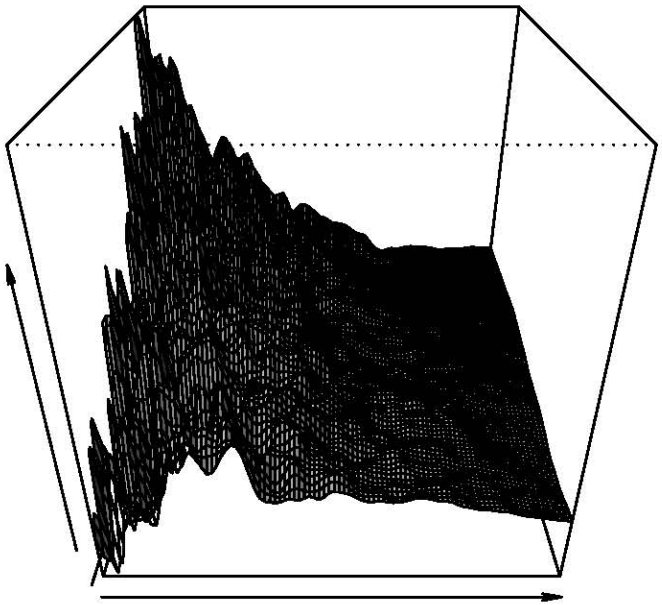


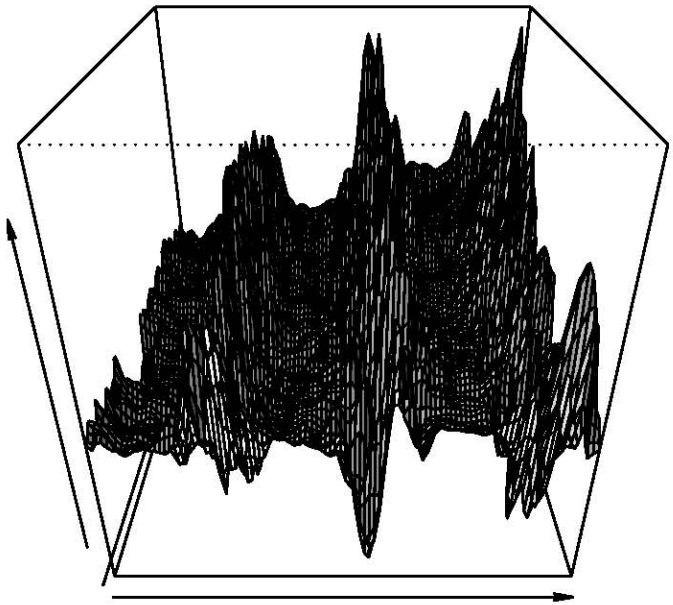
Fonctions de régularité 2D : $h(t)$ croissante - sinusoidale



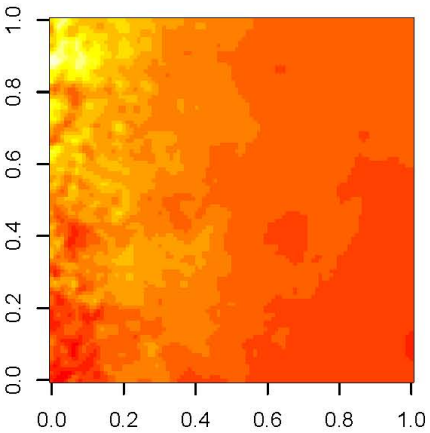


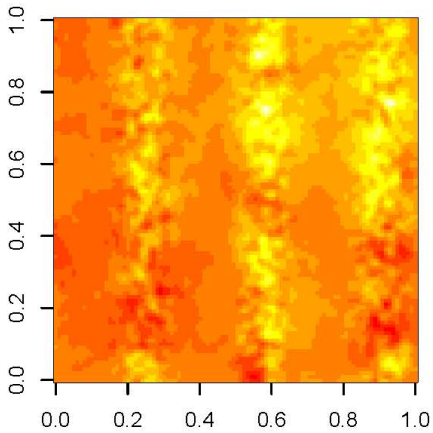
Processus multi-fractionnaire 2D : $h(t)$ croissante - sinusoidale

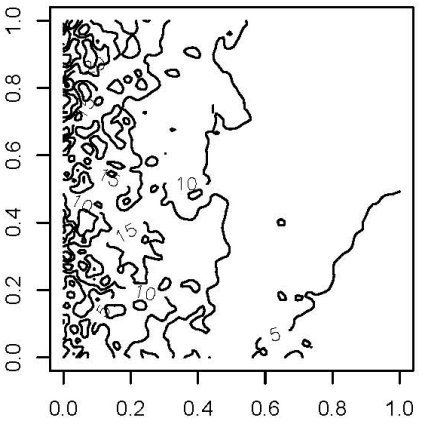


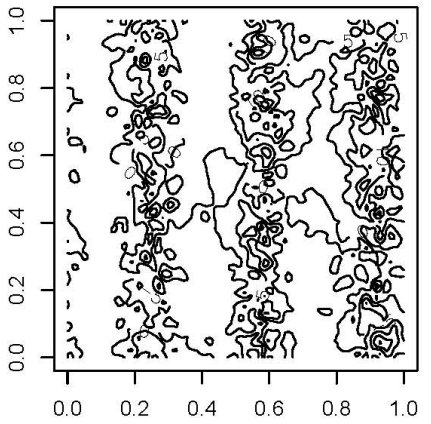


Autres représentations graphiques

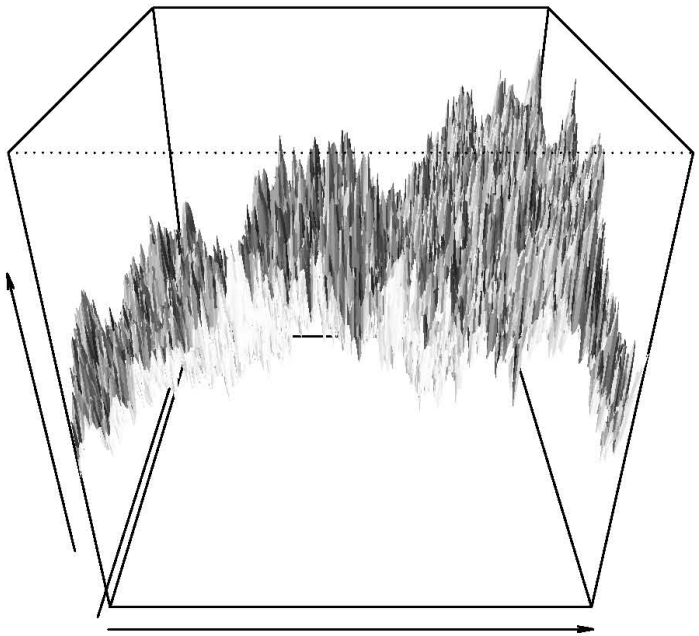


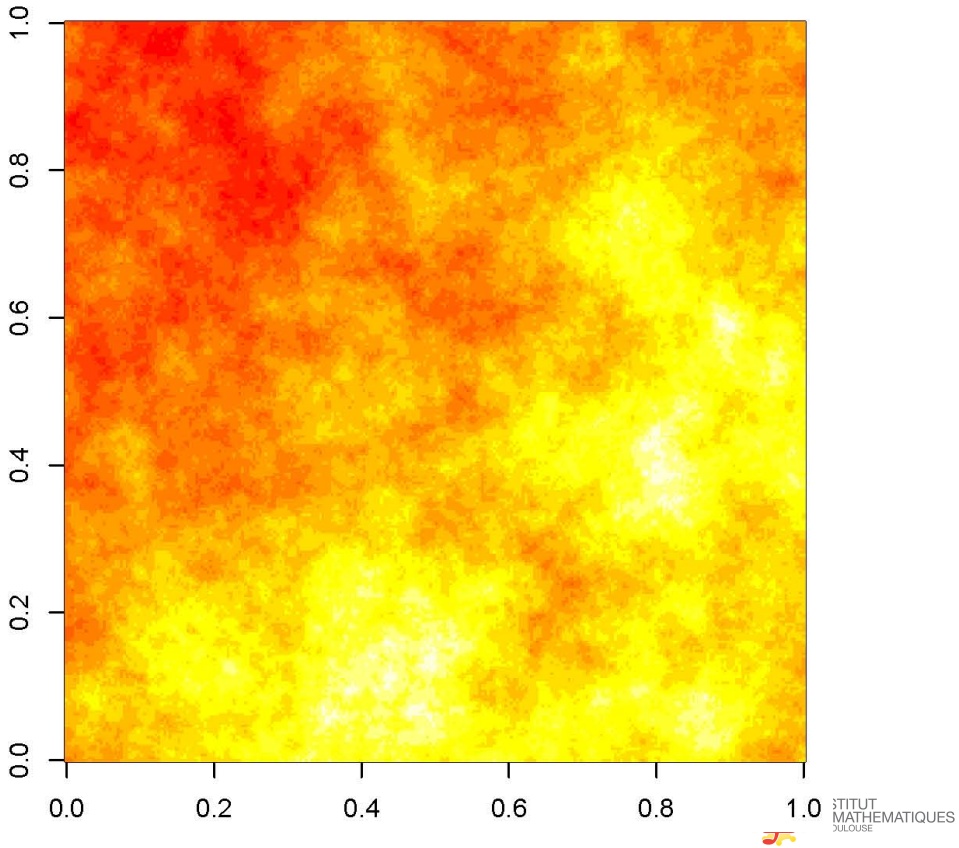




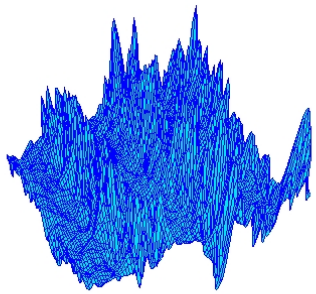


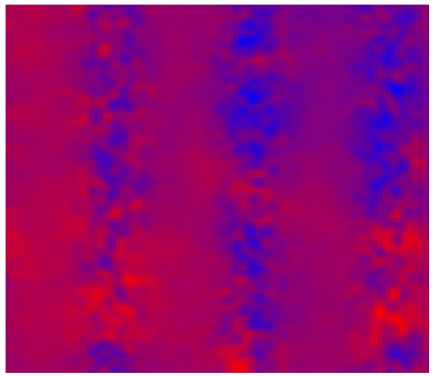
Comparaison d'une surface et d'une image

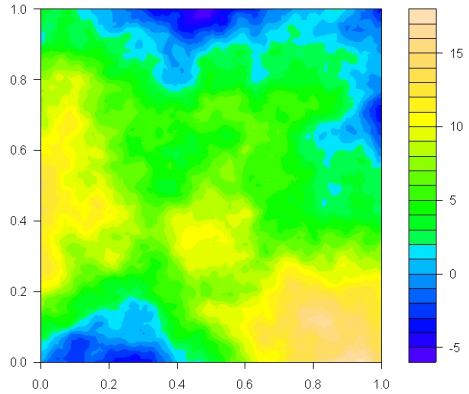


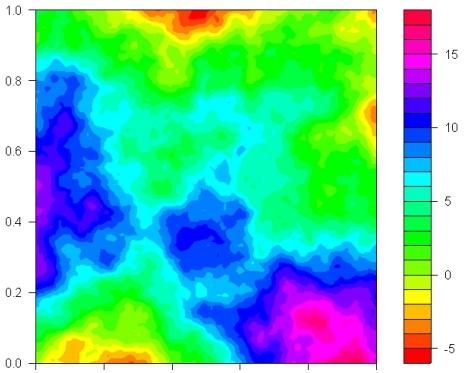


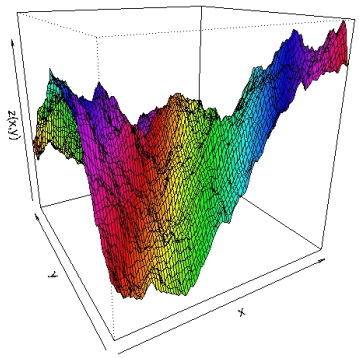
Galerie

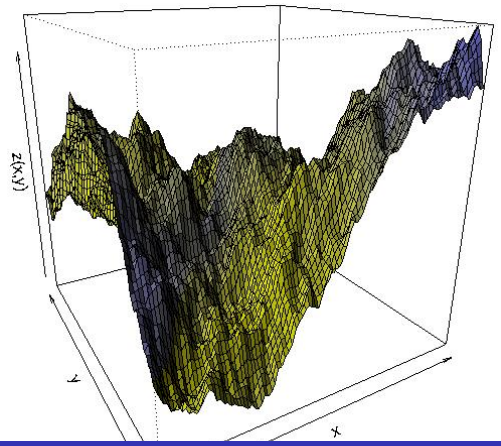





1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
15
10
5
0
-5


1.0
0.8
0.6
0.4
0.2
0.0
15
10
5
0
-5


z(x,y)
y
x


z(x,y)
y
x

Une question ?

Une question ?

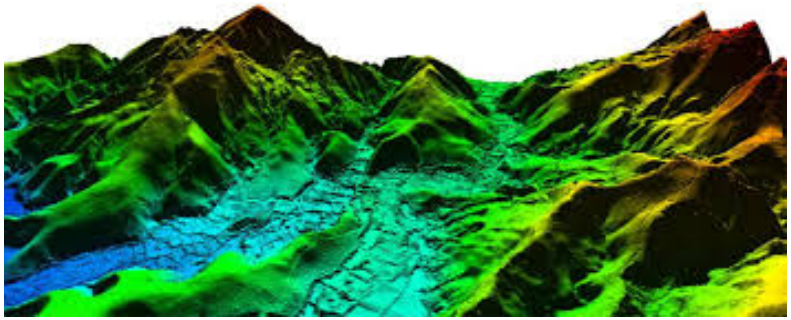
À quoi ça sert ?

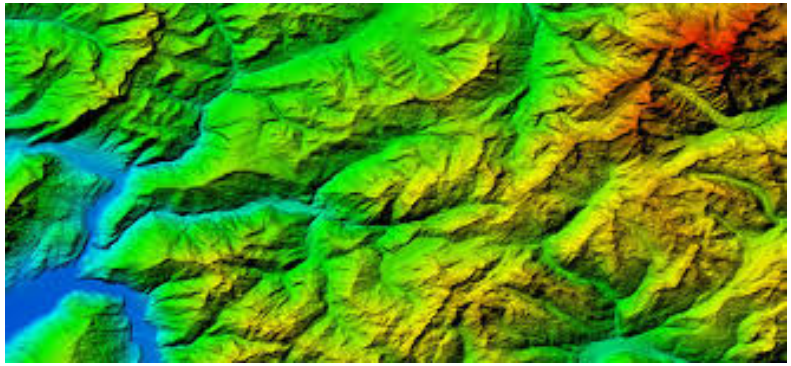
À la com'

Publications | Sébastien Déj...
Publications |
Institut de mathématiques ...
www.math.univ-toulouse.fr
Google
INSTITUT de MATHEMATIQUES de TOULOUSE
Rechercher
ok
Accueil
Coordonnées
Présentation de l'IMT
Membres
Recherche
Séminaires, groupes de travail, colloquium
Congrès et Colloques
Enseignement
Ressources Informatiques de l'IMT
Bibliothèque de mathématiques
Archives
Prix Fermat
Ecole Doctorale
Serveur d'enseignement
Annales de la Faculté des Sciences de Toulouse
FREMIT
MIBS
Institut de Mathématiques de Toulouse
Présentation
L'Institut de Mathématiques de Toulouse (IMT) est une Unité Mixte de Recherche du CNRS (UMR5219), regroupant environ 200 enseignants-chercheurs ou chercheurs permanents, 120 doctorants et 30 post-doctorants, répartis en 3 équipes : Mathématiques Fondamentales (Emile Picard), Mathématiques pour l'Industrie et la Physique (MIP), Statistique et Probabilités (ESP).
Coordonnées
Adresse postale
Université Paul Sabatier
Institut de Mathématiques de Toulouse
118 route de Narbonne
F-31062 Toulouse Cedex 9
Contact
Tél : 05.61.55.67.90
Fax : 05.61.55.75.99
Plan d'accès
ACTUALITÉS
Promotions
- Alexeï Lozinski, Maître de conférence à l'UPS a été promu comme Professeur à l'Université de Franche-Comté (Besançon)
- Marjolaine Puel, Maître de conférence à l'UPS (IUT) a été promue comme Professeur à l'Université de Nice
- Pascale Roesch, Maître de conférence à l'UPS a été promue comme Professeur à l'Université de Marseille.
Nouveaux arrivants pour 2012-2013
4 Professeurs :
- Jean-Yves Dauxois (26e), professeur à l'Université de Franche Comté a été recruté comme Professeur à l'INSA
- Frédéric Ferraty (26e), Maître de conférence à l'UT2, a été recruté comme Professeur à l'UT2
- Aurélien Garivier (26e), Chargé de Recherche CNRS au Laboratoire Traitement et
Annuaire
Nom
Equipe
Choisir une équipe
Rechercher
Espace réservé
Login
Mot de passe
ok
Création ou oubli du mot de passe

Utilisation potentielle (non vérifiée mais plus fun !)

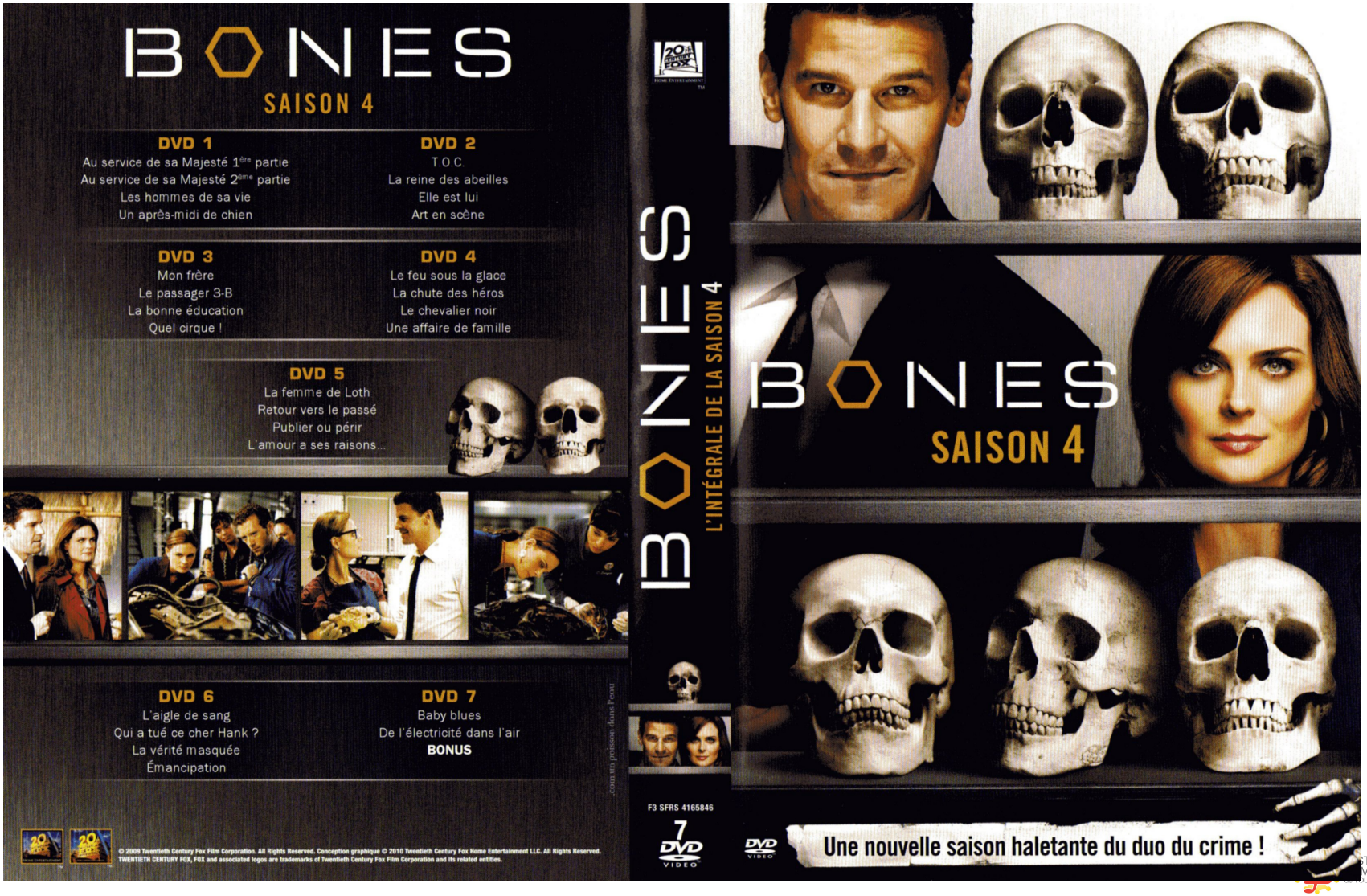
Jeux vidéos : modélisation numérique de terrains pour des simulateurs de vol !





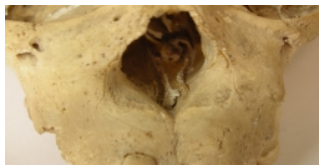
Un projet de paléo-anthropologie

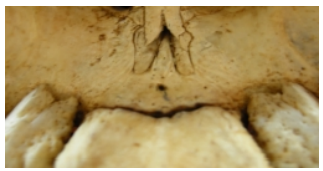
En collaboration avec le Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse (Toulouse)

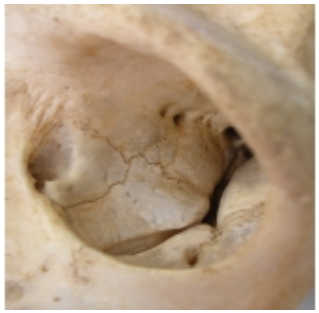

BONES
SAISON 4
DVD 1
Au service de sa Majesté 1ère partie
Au service de sa Majesté 2ème partie
Les hommes de sa vie
Un après-midi de chien
DVD 2
T.O.C.
La reine des abeilles
Elle est lui
Art en scène
DVD 3
Mon frère
Le passager 3-B
La bonne éducation
Quel cirque !
DVD 4
Le feu sous la glace
La chute des héros
Le chevalier noir
Une affaire de famille
DVD 5
La femme de Loth
Retour vers le passé
Publier ou périr
L'amour a ses raisons...
DVD 6
L'aigle de sang
Qui a tué ce cher Hank ?
La vérité masquée
Émancipation
DVD 7
Baby blues
De l'électricité dans l'air
BONUS
© 2009 Twentieth Century Fox Film Corporation. All Rights Reserved. Conception graphique © 2010 Twentieth Century Fox Home Entertainment LLC. All Rights Reserved.
TWENTIETH CENTURY FOX, FOX and associated logos are trademarks of Twentieth Century Fox Film Corporation and its related entities.
BONES
L'INTÉGRALE DE LA SAISON 4
F3 SFRS 4165846
7 DVD VIDEO
BONES
SAISON 4
Une nouvelle saison haletante du duo du crime !

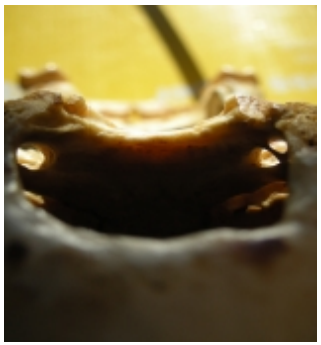
Les données

- **1453** crânes de grands singes : 712 chimpanzés, 396 gorilles, 345 orangs-outan
- espèce et sous-espèce, sexe, âge (J1,J2,J3,A1,A2)
- 40 caractères discrets : présence / absence









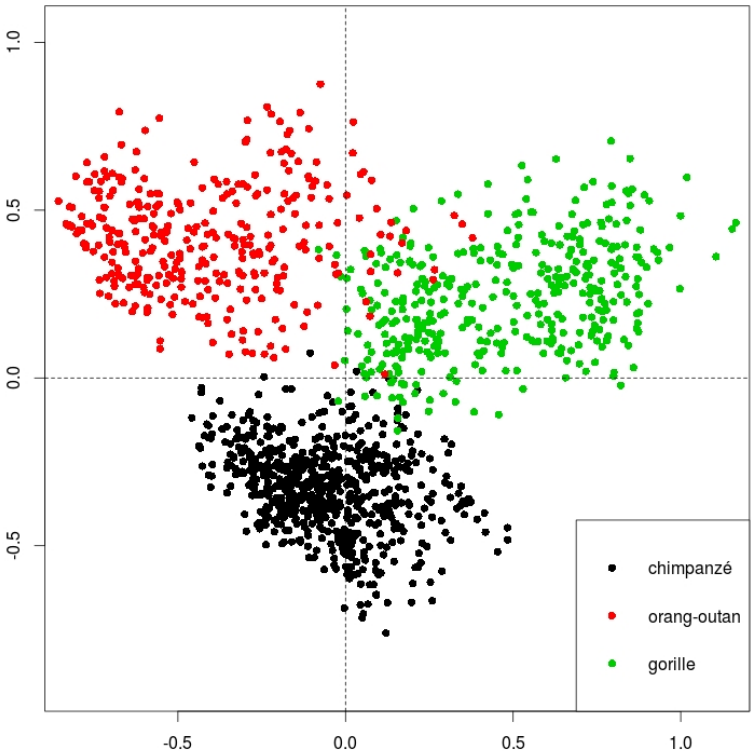
Question

Peut-on identifier les espèces et sous-espèces à l'aide des caractères discrets ?

Méthodes statistiques

- Analyse des Correspondances Multiples
- Forêts aléatoires

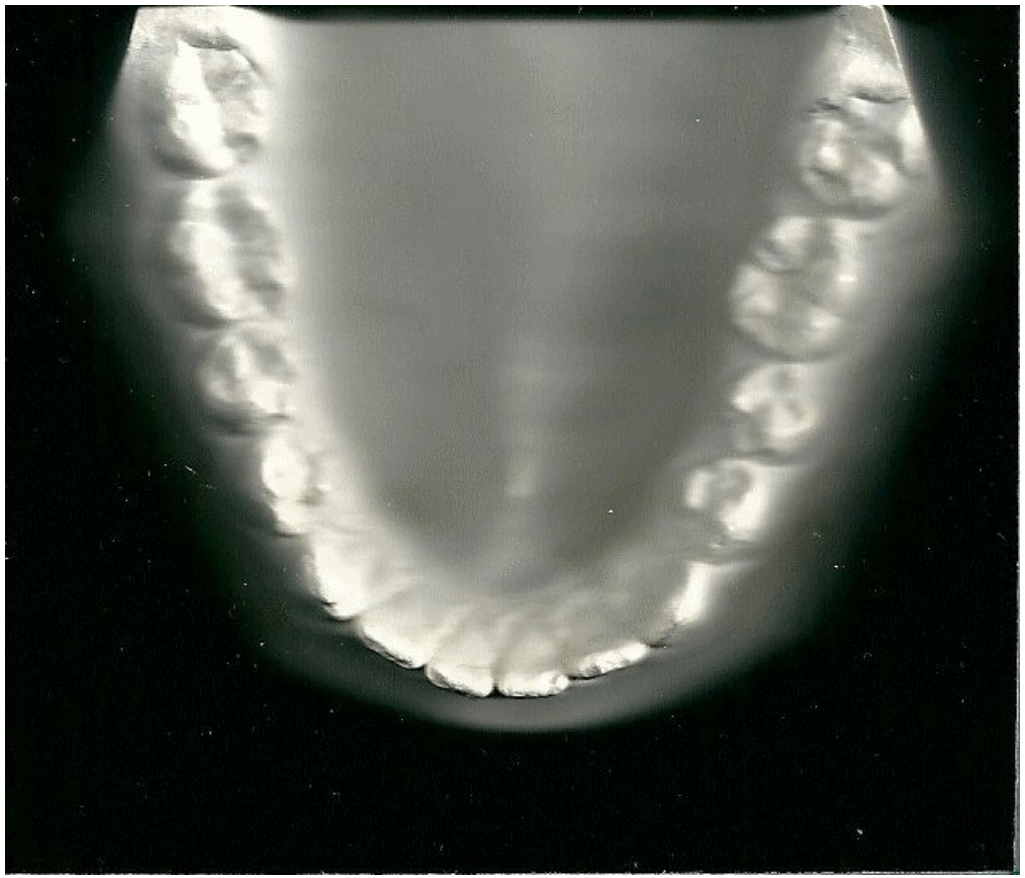
Résultats



Modélisation de l'arcade dentaire

En collaboration avec la faculté de chirurgie dentaire de l'Université Paul Sabatier.

Donnée : radiographie dentaire



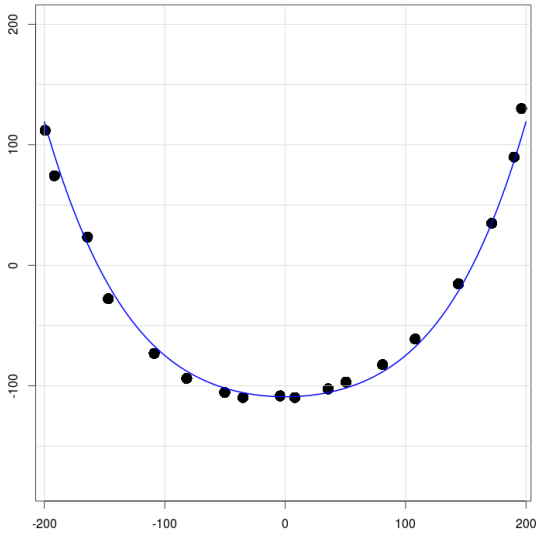
Modèle polynomial

Polynôme de degré 4 :

$$Y = b_0 + b_1.x + b_2.x^2 + b_3.x^3 + b_4.x^4$$

Equation de la courbe bleue (symétrique) :

$$Y = b_0 + b_2.x^2 + b_4.x^4$$



Exemples de questions

- comparer les courbes avant et après traitement
- comparer une courbe à des références (charte)