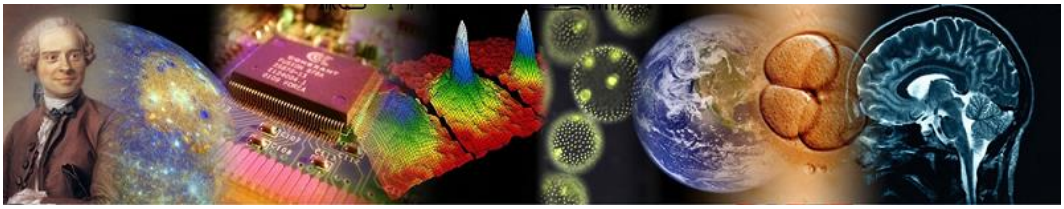

Séminaire du Centre d'Alembert

Centre Interdisciplinaire d'Étude de l'Évolution des Idées, des Sciences et des Techniques

18 avril 2019

Université Paris-Sud/Paris-Saclay, Orsay

La sélection dans tous ses états : fonctions, processus, conséquences

La sélection semble partout, dans la nature comme dans la société. Elle est souvent présentée comme l'outil idéal pour obtenir la meilleure adéquation entre souhaits et possibilités, besoins et ressources. C'est par la sélection qu'émergerait l'excellence. Est-ce le mode de fonctionnement optimisé de toutes les organisations ou un mode de gestion en situation de pénurie ? S'agit-il d'un processus rationnel pour obtenir des résultats interprétables ou d'une contingence ayant modelé l'évolution des espèces ? Comment fonctionne la sélection et existe-t-il des alternatives ? Nous réfléchirons sur les critères et les méthodes, qu'ils soient automatisés ou non, et au-delà nous interrogerons l'impact de la sélection sur le fonctionnement de nos disciplines scientifiques, sur l'établissement des normes, et sur l'organisation de nos sociétés. Pour cela, le Centre d'Alembert fera intervenir des collègues de différents domaines dans le cadre de ce séminaire : science de l'éducation, sciences de la vie, sciences et techniques des activités physiques et sportives, économie, informatique, physique…

Big data, IA, sélection des données : causalités, corrélations, conséquences

Diviyan Kalainathan

Doctorant, Laboratoire de Recherche en (LRI)

Causalité observationnelle : découverte de liens de cause à effet sans expériences randomisées

Résumé

Déterminer les liens de cause à effet afin de comprendre le fonctionnement de mécanismes est primordial et a déjà été utilisé dans tous les domaines : en épidémiologie, physique, sciences sociales, …

Afin de pouvoir déterminer de façon fiable, les expériences randomisées ont toujours été la méthode d'excellence ; mais celles-ci sont parfois chères, contraires à l'éthique, voire même infaisables (par exemple en épidémiologie génétique). Ainsi ont fleuri les méthodes de découverte de liens de cause à effet sans interventions. Basées sur la récolte des données, le but est de déterminer une partie des relations grâce aux propriétés statistiques des données.

Dans cet exposé, nous allons introduire les principaux principes utilisés et les pièges à éviter pour déterminer les relations causales à l'aide de statistiques. Dans un deuxième temps, nous allons présenter des algorithmes employant ces principes efficacement pour produire automatiquement des graphes causaux, et analyser les prédictions de ces algorithmes.

TABLE DES MATIÈRES

Avant-propos

Je m'appelle Christine Eisenbeis. Je suis chercheuse dans ce laboratoire : le LRI [Laboratoire de Recherche en Informatique]. Aujourd'hui, [nous avons] choisi de décliner le problème de la sélection dans tous ses états avec le sujet de l'informatique et, évidemment, on s'est posé la question : Qui est-ce qui va sélectionner derrière ces algorithmes, [derrière] ces intelligences artificielles ?
Que se cache-t-il derrière ces mots ? Ces mots d'IA [Intelligence Artificielle], ne sont-ils pas une boîte noire ? Il y a eu beaucoup de renouveau qui a fait émerger ce mot d'« Intelligence Artificielle » et qui pose beaucoup de questions à cause de deux choses : la première, c'est la puissance des machines qui permet de gérer beaucoup de données, la deuxième, c'est l'élargissement, à tout le monde, de la mise en place de capteurs de données. D'ailleurs on peut se demander si ces capteurs ne sont pas des *voleurs de données*. On parle beaucoup [de ce sujet]. C'est une boîte noire [dont] on va, aujourd'hui, ouvrir deux petites fenêtres, la première pour essayer d'aller voir [ce qu'il] y a comme techniques et [comme] sciences derrière les algorithmes de Big Data (1ère partie présentée par Diviyan) et la deuxième, un regard de sociologue qui va [réfléchir à des] questions [qu'on oublie souvent de se poser en informatique] : « Qu'y a-t-il derrière cette IA et ces Big Data ? Où sont les gens, les humains ? »

La première présentation est faite par Diviyan Kalainathan qui est doctorant, dans ce laboratoire, dans une équipe qui s'appelait TAO (Apprentissage et Optimisation) et qui est devenue TAU (TAckling the Underspecified), sur l'imprécision des finalités du Big Data.

Je vais laisser Diviyan présenter une partie de son doctorat « *Causalité observationnelle. Découverte de liens de cause à effet sans expériences randomisées* » [soutenue le 17 décembre 2019 à l'Université Paris-Saclay]. Je vais également remercier le laboratoire et le Centre d'Alembert qui organise tous ces séminaires qui nous permettent de prendre le temps de nous poser et de réfléchir, et bien sûr, à l'avance, [remercier] les intervenants.

Causalité observationnelle : découverte de liens de cause à effet sans expériences randomisées

[*Temps = 3 minutes et 06 secondes*]

Merci Christine.
Je vais vous faire une présentation sur la causalité observationnelle, donc je vais rester quand même [à] un haut niveau, mais si vous avez la moindre question, n'hésitez pas, parce que je sais que je peux [ne] pas être clair sur certains aspects en causalité.

On va commencer brièvement par introduire ce qu'est la causalité, [en essayant] de poser la définition plus ou moins mathématique, et ensuite, [dans une seconde partie] [on passera] dans le domaine de découvertes de liens de cause à effet dans des *data set* qu'on connaît ou qu'on ne connaît pas vraiment, puis dans une troisième partie, [on verra comment] utiliser ces principes pour en faire des algorithmes. Enfin on conclura sur ce thème.

[*Temps = 3 minutes et 40 secondes*]

1 Introduction

Qu'est-ce que la causalité ? [Sur la] (Diapo 3), on peut voir un monsieur qui va couper une branche et qui va sûrement se faire mal. Donc les douleurs qu'il va percevoir sont l'effet de ce phénomène. Mais quelle [en] est la cause ici ? [S'agit-il] d'un manque de précautions ? S'agit-il d'un manque de connaissances physiques, notamment [au sujet de la force de] gravité ? La causalité : on a une cause et un effet. Les effets peuvent être multiples, les causes aussi.

La première objection qu'on a quand on parle de causalité, c'est que *corrélation* n'est pas *causalité*. (Diapo 4) : on a une petite planche qui dit : « j'ai pris des cours de statistiques et je sais que corrélation n'implique pas causalité, mais je ne sais pas vraiment pourquoi », et on va justement essayer de voir pourquoi.

Dans le cas « corrélation n'est pas causalité », on a ce joli graphique (Diapo 5) ; en ordonnée, on a le nombre de prix Nobel par pays et, en abscisse, on a la consommation de chocolat par pays. On peut voir qu'il y a une très bonne corrélation, assez significative, donc je conseillerais à tout le monde de manger du chocolat plus fréquemment, vous aurez plus de chances d'avoir des prix Nobel. Mais malheureusement, vous vous doutez que ce n'est pas vrai. Justement ici, il y a corrélation et non pas causalité et on va essayer d'expliquer ce cas un peu plus tard dans l'exposé mais vous pouvez taper sur Google « corrélations intéressantes » et vous pourrez voir [qu'on en trouve] d'encore plus absurdes.

[*Temps = 5 minutes et 41 secondes*]

(Diapo 6)
On va définir ce qu'est la causalité, mais tout d'abord par le concept d'opérateurs d'intervention. On va le noter do($X=x_1$), [c'est à dire], qu'on a une variable et je vais la forcer à telle valeur. Par exemple, le monsieur qui était sur sa branche, je vais le forcer à couper sa branche, il n'a pas choisi. C'est l'opérateur d'intervention.

On dit qu'une variable X est une cause d'Y si, en forçant X à une valeur, la variable Y change. Donc là c'est [une définition] en termes de probabilités (diapo 6). Si je force la variable X à prendre la valeur X_1, la probabilité de Y ne changera que si je l'avais forcée à [prendre] une valeur X_2. Modifier la cause a un impact sur l'effet, en [termes de] distribution. C'est tout ce que dit cette formule [de la diapositive 6].

[*Temps = 6 minutes et 45 secondes*]

(Diapo 6)
On a un deuxième exemple où [l'existence de] facteurs génétiques et le fait de fumer impliquent des risques de cancer. Imaginons qu'on force tout le monde à fumer, on remarque que la probabilité d'avoir un cancer va augmenter, bien sûr en supposant qu'on garde des facteurs génétiques constants. On est tous jumeaux, on force une moitié à fumer, l'autre moitié à ne pas fumer, on va remarquer que la moitié qui a fumé a un risque plus élevé d'avoir un cancer. Vous pouvez vous rendre compte qu'il y a un problème. Non seulement, nous ne sommes pas tous jumeaux, désolé, mais on ne peut pas forcer des gens à fumer et [d'autres] à ne pas fumer. Cela [soulève] un vrai problème en causalité parce qu'en physique ou dans d'autres domaines en sciences, on peut faire des expériences randomisées.

Par exemple, je prends des bactéries ou des rats et je fais des expériences : je les clone, etc. et ça marche très bien, mais, dans certains domaines, ça marche beaucoup moins bien, notamment en sciences sociales. On ne peut virer des gens juste pour faire une expérience [et de ce fait], certaines expériences sont chères, coûteuses, d'autres ne sont pas du tout éthiques et certaines sont pratiquement infaisables, notamment le fait qu'on soit tous jumeaux. C'est pour cela qu'intervient la notion de causalité observationnelle. (Diapo 7)

[*Temps = 8 minutes et 22 secondes*]

Causalité observationnelle, c'est à dire qu'on n'a fait aucune expérience dans notre ensemble de données et on va essayer d'[y] trouver des relations de cause à effet en supposant une entrée.

On a des données X, un nombre d'exemples N et des variables, ici différentes variables qu'on va essayer d'ordonner sous la forme d'un graphe, de cause à effet, et bien sûr pas d'intervention possible. A la fin, on [obtiendra] un graphe qui

oriente, qui ordonne toutes ces variables sous forme d'un graphe [qui peut être un arbre] reliant cause et effet, donc pas forcément totalement dirigé, mais ayant quelques relations entre elles.

(Diapo 8)
Un exemple de graphe causal : on a une variable X_1 qui va causer une variable X_3 et une variable X_3 et X_7 qui vont causer une variable X_5. La notion de « functionnal causal model » intervient. C'est un système d'équations : on a une fonction f et des variables qui vont causer les effets.

Par exemple, ici, on a notre exemple avec X_5 : X_3 et X_4 vont entrer dans la fonction, ainsi qu'une variable de bruit qui ajoute la stochasticité dans le système. Tout ce qu'on n'observe pas dans notre système va être introduit dans cette variable et [de ce fait], cette fonction f5 est déterministe. Donc toutes les fonctions d'un *functionnal causal model* sont déterministes. [Il s'agit là] de l'une des modélisations possibles en causalité et c'est celle qu'on sélectionne dans notre cas.

[*Temps = 10 minutes et 14 secondes*]

2 Causal discovery

Ensuite, on va voir les principaux concepts utilisés dans *causal discovery* et les principaux pièges surtout. On peut remarquer des relations statistiques dans nos données qui nous permettent d'orienter nos graphes.

2.1 Dépendance et indépendance causale

Par exemple (Diapo 9), on a A dépendant de C, donc A et C sont dépendants mais conditionnent par rapport à B, A et C deviennent indépendants. En conditionnant par rapport à B, on casse le lien de dépendance entre A et C.

Cette propriété est vérifiée par trois structures sous forme de chaîne ABC, CBA et B qui cause A et C. Vous remarquez tout de suite qu'il y a une relation qui peut être un peu traître.

Par exemple, là [A=>B=>C ou C=>B=>A], on a remarqué que A et C était dépendants, [mais ils sont indépendants dans le cas « A<=B=>C »] Mais si [dans ce dernier cas] on n'observe pas B, qu'il n'est pas dans notre ensemble de données, [alors] on peut avoir des corrélations qui sont « traîtres ». C'est justement l'exemple du chocolat et des prix Nobel : on peut dire qu'il y a une variable B, le PIB du pays, qui cause le fait que des personnes consomment plus de chocolat et, par exemple, [la richesse d'un pays, décrite par le PIB, peut favoriser] l'éducation

[dans ce pays ainsi que le financement de la recherche, ce] qui cause justement une augmentation du nombre de prix Nobel.
C'est ce qu'on [appelle] « avoir un facteur confondant », cette variable qui n'est parfois pas observée qui est assez traître dans nos corrélations.

(Diapo 10)
Une deuxième structure, assez unique : la V-structure. On a deux variables qui sont a priori indépendantes, A et C, qui deviennent dépendantes si on conditionne par rapport à une troisième variable B. Pour essayer de l'expliquer, on va faire un petit exercice en prenant un exemple.

On suppose l'expérience d'un ouvrier dans une entreprise X, [dont le poste de travail nécessite des manipulations plus ou moins compliquées], [avec] des risques d'accident. Bien sûr, si l'ouvrier a peu d'expérience, il a plus de risque d'avoir un accident et [plus] la machine est complexe, [plus] le risque d'accident augmente.

Maintenant on va conditionner par rapport à la variable du milieu. On sait qu'il y a eu un accident et que l'ouvrier était assez expérimenté. Tout de suite, vous pensez [qu'il] doit être dû au fait que la machine doit être assez complexe ; une information vient de transiter de « *worker experience* » vers la complexité de la machine. En supposant que l'ouvrier était assez expérimenté, on peut se dire qu'il a eu un accident parce que la machine était vraiment compliquée : il travaille depuis 20 ans à ce poste, il n'a jamais eu d'accident. C'est une manière de révéler ce type de structure et cette propriété est unique pour une seule structure, ce qui permet d'orienter directement les flèches de A vers B, de C vers B.

[*Temps = 13 minutes et 41 secondes*]

2.2 Le rasoir D'Ockham

(Diapo 11)
Il y a aussi un autre concept dans l'orientation des liens de cause à effet : le rasoir d'Ockham, c'est à dire qu'on suppose que la théorie la plus simple et qui colle à mes données est bonne. C'est un argument de simplicité.

Là, on a deux variables B et A, là, A et B, on a juste retourné la fonction et on remarque que la fonction est assez simple, c'est un sinus, alors qu'ici elle est très complexe. C'est une fonction qui a deux images et c'est assez compliqué de la représenter. [Et partant de là], on se dit que la solution la plus simple, donc celle du sinus, doit sûrement être la solution causale, c'est à dire que A doit causer B parce que c'est plus facile d'aller de A vers B que de B vers A. Par exemple si on a cette valeur de B, on ne sait pas si c'est celle-ci ou celle-là [qu'elle correspond sur le graphe]. C'est donc un deuxième argument de simplicité qui est exploité de plus en plus dans les distributions.

[*Temps = 14 minutes et 47 secondes*]

(Diapo 12)

2.3 Les pièges

Le plus connu est le paradoxe de Simpson. C'est un problème qui intervient si on agrège des données. Ici, [il s'agit] d'un vrai cas médical : on a 2 traitements, A et B, pour des calculs rénaux et là [ce chiffre sur la diapositive correspond aux] succès en fonction de la taille des calculs [rénaux]. Là, pour les petits calculs [rénaux], on a le traitement A [avec] 93 % de réussite et 87,2 % de réussite pour le traitement B. [Pour les gros calculs rénaux, le succès dans le traitement des calculs est respectivement de] 73 % [pour le traitement A] et 69 % [pour le traitement B].
En regardant ce premier tableau, on se dit que le traitement A est sûrement le meilleur car, à chaque fois, il marche mieux que le traitement B dans les différents cas. [Puis], on agrège les résultats [en fusionnant] petits et gros calculs et on observe que le traitement A a seulement 78 % de réussite, contre 83% pour le traitement B. Vous me dites qu'il y a un problème, le traitement A marchait mieux avant et moins bien après. Que s'est-il passé ?
Ici, on a un a priori des médecins sur l'utilisation des traitements. Il y a un facteur confondant qui nous est inconnu et qui est la décision du médecin de choisir le traitement en fonction de la taille du calcul. Ce qui fait que cette relation n'existe pas vraiment. C'est celle-là qu'il faut conserver parce que justement le médecin a choisi le traitement en fonction de la taille du calcul. De ce fait, le traitement B est plus rentable pour les petits calculs et moins rentable pour les gros calculs. C'est un premier piège.

Le deuxième piège est le biais de sélection. (Diapo 13)
Imaginons qu'ici on n'ait pas exactement le même nombre d'opérations qui sont intervenues. Et si on choisit de ne sélectionner que les petits calculs, on peut se tromper, bien sûr, dans notre décision. Ce sont des pièges assez importants à relever et à prendre en compte.
Que s'est-il passé ? C'est le type de calcul rénal qui choisit directement le type de traitement qui va être utilisé en fonction de la taille du [calcul]. Attention aux facteurs confondants et à l'agrégation des résultats parce qu'[ils] peuvent vous mener à [prendre] de mauvaises décisions et à [tirer] de mauvaises conclusions.

[*Temps = 17 minutes et 34 secondes*]

(Diapo 14)
3 suppositions, 3 hypothèses, souvent faites et qui peuvent être critiquables, en découverte observationnelle :

- 1ère hypothèse, assez forte « *Causal Sufficiency* » : pas de facteur confondant en dehors de mes données, ce qui est faux dans la plupart des cas car il y en a toujours, mais on suppose qu'ils n'influent pas trop sur nos conclusions.
- deuxième hypothèse : « *Causal Markov* » : toutes les variables sont indépendantes de leurs non-effets conditionnellement aux parents.
- Dernière hypothèse : « *Causal Faithfulness* » : toute dépendance conditionnelle statistique présente dans les données provient du vrai graphe.

[*Temps = 19 minutes et 10 secondes*]

2.4 Les algorithmes

(Diapo 15)
Le premier type d'algorithme, assez simple, qu'on va voir est sur les propriétés. [Il] consiste à utiliser les différentes propriétés qu'on a vues. Tout d'abord, on trouve un squelette du graphe avec des relations de dépendance et on oriente tous les arcs à l'aide des V-structures, A et C qui causent B, avec les relations qu'on a trouvées et on fait de la propagation de contraintes.

Exemple. (Diapo 18)
[Sur cette structure], il n'y a qu'une seule V-structure : X_3, X_4 qui causent X_5 qu'on retrouve directement ici.
Ensuite, on sait qu'il n'y a pas de V-structure ici, donc X_3 et X_4 ne causent pas X_5, donc on a la propriété statistique qu'on a une équivalence de Markov entre les trois différentes structures et, [en conséquence], on sait qu'on ne peut pas créer de nouvelles V-structures, ce qui fait que, par propagation de contraintes, on peut directement orienter cet arc bleu. Toutefois, [du fait] des propriétés statistiques, on ne peut pas orienter ces directions. Donc ces algorithmes à contraintes ne nous permettent pas d'orienter complètement un graphe, mais seulement [de l'orienter] en partie.

[*Temps = 20 minutes et 41 secondes*]

(Diapo 16)
Les limitations de ces méthodes sont les choix du test statistique. Ici on a juste choisi une corrélation pour le test le plus simple : la corrélation conditionnelle. [Mais], deuxième problème, on doit tester toutes les structures en conditionnant par rapport à tous les sous-ensembles de variables, ce qui fait exploser la complexité de l'algorithme et le temps de calcul qui peut se retrouver exponentiel, notamment un test qui peut prendre moins d'une seconde pour 20 variables peut prendre plus d'un siècle pour 1000 variables.

[*Temps = 21 minutes et 20 secondes*]

(Diapo 17)

Une deuxième approche : les algorithmes à score. On a un graphe candidat par exemple (Diapo 18) et on évalue la solution par rapport aux données [pour voir jusqu'] à quel point le graphe candidat colle aux données. On modélise et on évalue le nouveau graphe et bien sûr si l'erreur est plus faible que le meilleur [résultat] qu'on avait obtenu [auparavant], on conserve le [nouveau] graphe obtenu. Ensuite, on continue à chercher en modifiant des arcs du graphe ; par exemple, on retourne cet arc, on le retire carrément ou on relie ces deux-là, on fait certaines opérations et on regarde si le nouveau graphe colle mieux aux données ou pas. [Ainsi], on fait une recherche heuristique autour des graphes possibles jusqu'à [obtenir] certains critères de convergence ou la stabilité du graphe, c'est-à-dire qu'on ne l'a pas changé depuis un certain nombre d'itérations.

Ces différentes méthodes -que j'ai précédemment évoquées- nous donnent des graphes qui sont partiellement dirigés. Même les graphes à score utilisent ces propriétés [de façon] plus ou moins cachées. Au début, on a un certain ensemble de données, X_1 jusqu'à X_5, et on ne sait pas comment elles sont ordonnées, ni quels sont les causes et les effets.
C'est souvent au niveau de la première étape que ça va casser. Le squelette du graphe va être vide avec des variables qui ne sont pas corrélées entre elles. L'algorithme va s'arrêter ici [n'ayant] aucune relation à donner.

[*Temps = 23 minutes et 37 secondes*]

(Diapo 19)
Avec ces critères de simplicité, on a aussi des méthodes qui sont bivariées. [Il est possible], à partir de seulement deux variables, d'essayer de savoir, entre deux variables X et Y, laquelle cause l'autre. Là c'est le critère de simplicité qu'on a vu précédemment et l'un des modèles les plus connus est le « *Additive Noise Model* » qui suppose que le bruit a une contribution additive dans le mécanisme, contrairement [au] cas le plus général où il serait dans la fonction. Dans ce cas, on a f qui est une fonction continue et E un bruit, indépendant de la cause X, [ce qui] est l'élément le plus important à retenir de cette équation.
Pourquoi ?

On a un parallélogramme, c'est notre distribution, et on va faire la régression dans les deux sens, c'est-à-dire de X vers Y et de Y vers X.
Donc ici, on a supposé que notre mécanisme était linéaire -il l'était effectivement, on a eu un peu de chance- et on regarde les résidus de la régression qui doivent normalement correspondre au bruit. Dans le sens X cause Y, on remarque bien que le bruit est indépendant de la cause, donc indépendant de X. [Dans ce cas], le modèle colle. Si on regarde le résidu dans l'autre sens, donc de la régression de Y par rapport à X, on remarque que le bruit est dépendant de Y, le modèle ne colle pas parce qu'il n'est pas indépendant de la cause.
Directement, on peut dire que X cause Y par cette relation de bruit qui n'est pas indépendante dans l'autre sens.

[*Temps = 25 minutes et 30 secondes*]

(Diapo 20)
Je vais [maintenant] faire un peu de publicité.
Pour nos algorithmes, on essaye d'appliquer ces principes avec des réseaux de neurones, donc je ne vais pas aller dans le détail mais ça revient un peu [au] même concept [que le] graphe (Diapo 8). On va modéliser ces fonctions avec des réseaux de neurones, c'est pratique pour leur flexibilité et leur facilité d'apprentissage, et on arrive à obtenir d'assez bons résultats en utilisant des critères comme les CGNN (Causal Generative Neural Networks) ou des critères SAM dans le 2ème papier (SAM : Strutural Agnostic Model, Causal Discovery and Penalized Adversarial Learning, 2018, ArXiv, Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, Michèle Sebag).
Fin de la pub, merci !

[Temps = 26 minutes et 17 secondes]

(Diapo 21)

3 Conclusion

On va maintenant conclure.
Que peut-on peut dire sur la causalité observationnelle ? Tout d'abord, il faut faire attention parce qu'il n'y a pas de solution, pas d'outil [comme] le couteau suisse qui vous permet de tout faire, et souvent, ce sont des solutions au cas par cas. Il faut [également] faire attention aux conclusions qu'on tire. Il faut toujours essayer de vérifier et de faire des expériences. Donc cela sert aussi à planifier des expériences et en tirer un maximum d'informations. Il faut faire attention à toutes les hypothèses et toujours [voir] si elles sont vérifiées mais cela permet aussi d'avoir une bonne idée de la structure des données et d'avoir des confirmations ou des questions à se poser sur notre ensemble de données.
J'espère que [cet exposé] vous a plu.

Transcription réalisée par Véronique Luec et Julien Gargani.

CENTRE D'ALEMBERT

Centre **I**nterdisciplinaire d'**É**tude de l'**É**volution des **I**dées,des **S**ciences et des **T**echniques
Bâtiment 407 - 91405 ORSAY Cedex - Tél. : 01.69.15.61.90
Courriel : centre.dalembert@universite-paris-saclay.fr
Web : http://www.centre-dalembert.universite-paris-saclay.fr