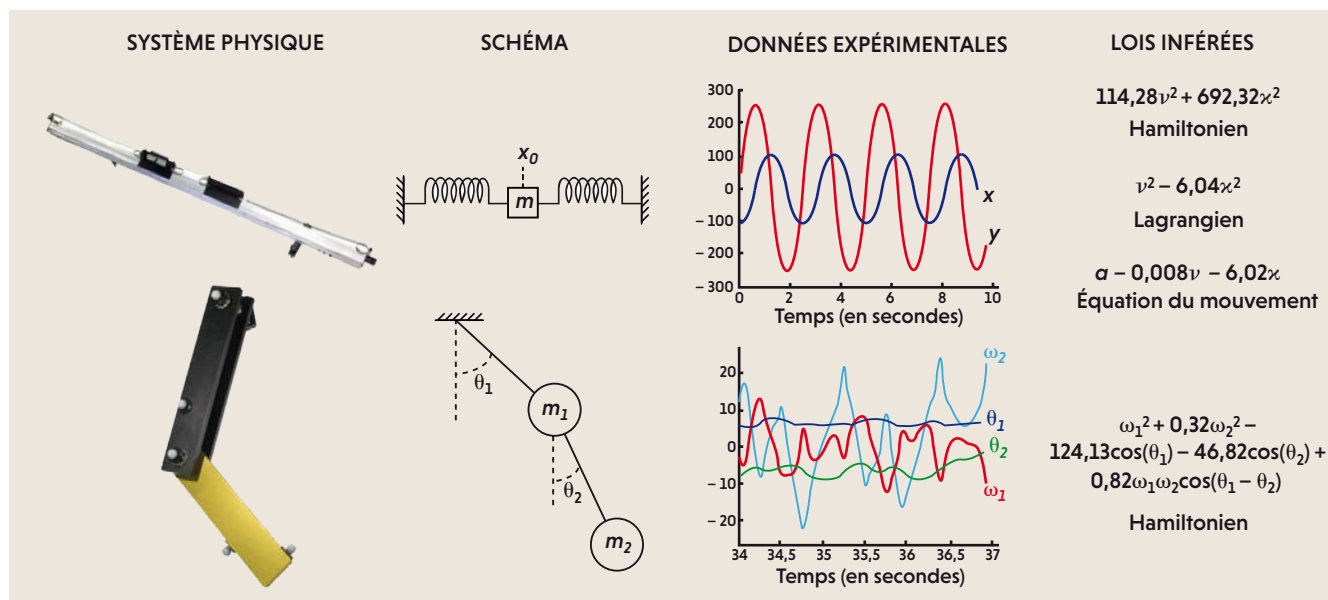




Un algorithme retrouve les lois de la mécanique (à droite) à partir des données (au centre) collectées sur des dispositifs mécaniques (à gauche). En est-il pour autant un scientifique ?

produire, de plus en plus manifeste: l'arrivée du *big data*. Tout le monde en parle, et pour cause: couplé à l'intelligence artificielle (voir *La révolution de l'apprentissage profond*, par Y. Bengio, page 42), le *big data* a déjà, et aura plus encore dans l'avenir, un impact majeur sur les technologies, le travail, les emplois, l'économie (voir *l'entretien avec D. Cohen*, page 8), la médecine et les relations interpersonnelles. Et aussi, à l'évidence, sur les sciences, notamment sur la façon de mener les recherches, et peut-être aussi sur le statut même des lois physiques. De quelle façon ?

## LES PILIERS DE LA SCIENCE

Galilée est à juste titre considéré comme l'inventeur de la science telle que nous la pratiquons depuis quatre siècles, dont les deux grands piliers ont été la théorisation et l'expérimentation: on élabore des hypothèses conduisant à la formulation de lois, qui sont ensuite testées grâce à des expériences spécifiques, souvent très éloignées de la simple observation. Par exemple, lorsque, en 1604, Galilée énonce que tous les corps, quelle que soit leur masse, tombent à la même vitesse, il formule une loi qui n'est ni conforme aux données empiriques de l'époque, ni directement observable (elle ne peut l'être que dans le vide): elle ne sera expérimentalement vérifiée que bien plus tard.

Mais cette façon de faire de la science n'est-elle pas en train de changer sous nos yeux? Avec l'arrivée du *big data*, ce modèle canonique demeurera-t-il toujours vrai? Continuerons-nous d'honorer la pensée spéculative, d'incliner à «la gravité enjouée de l'enquête métaphysique»? L'article de Chris Anderson cité en introduction répondait par la négative à ces questions.

Le *big data* est cet ensemble énorme de données brutes et silencieuses, dont le volume augmente à une vitesse prodigieuse. Ces données choisies mémorisent les «traces» de l'activité d'êtres vivants, de machines, d'objets, et leurs états successifs. Leur collecte est principalement automatique, puis ces données sont analysées par des algorithmes qui y détectent des régularités, par exemple dans le comportement des consommateurs, des machines, des indices économiques, du trafic routier...

À partir de ces régularités, ils infèrent des règles prédictives que nous avons tendance à considérer comme des normes, ou comme des lois générales, voire universelles, alors qu'elles ne sont que la condensation de ce qui a déjà eu lieu: dès lors que le futur qu'elles configurent n'est que du passé extrapolé, elles ne peuvent correctement prédire l'avenir qu'à la condition que celui-ci prolonge le passé, sans surprises ni ruptures ni inventions.

Privilégier de la sorte l'induction, n'est-ce pas faire exagérément confiance à une certaine uniformité du cours de la nature et trop croire que certaines choses n'arriveront pas? L'idée de rupture n'est-elle pas en train de disparaître ?

Certes, le *big data* ouvre des perspectives fascinantes, notamment celle de redécouvrir des lois universelles déjà connues par la simple analyse de données massives. C'est ce qu'ont montré Michael Schmidt et Hod Lipson, de l'université Cornell, à Ithaca, aux États-Unis. Ils ont d'abord accumulé des données de positions, de vitesse, d'angles... enregistrées en certains points d'un dispositif mécanique en mouvement, notamment un double pendule au comportement chaotique (voir la figure ci-dessus). Ensuite, ces informations ont alimenté un algorithme « naïf », c'est-à-dire

➤ dépourvu de toute connaissance en géométrie et en physique. Le programme a réussi à «découvrir» plusieurs lois (les Hamiltoniens, les Lagrangiens, la conservation du moment...). En d'autres termes, l'algorithme s'est mué en un physicien ! Dans ce cas, les lois identifiées étaient connues, mais pourquoi ne pas imaginer qu'un algorithme puisse identifier de nouvelles lois ?

Le *big data* peut également aider à la compréhension d'événements impliquant de très grands nombres de variables quantifiables, tels les phénomènes météorologiques ou climatiques, les comportements électoraux, l'usage des réseaux sociaux...

## FUYEZ L'HÔPITAL !

Toutefois, ce transfert vers les *big data* n'est pas sans risque. De fait, il est aussi possible que nous nous égarions dans l'identification de multiples corrélations, pas forcément bien interprétées. Or une corrélation n'est pas une relation de cause à effet. Elle marque simplement le fait que deux grandeurs semblent dépendre l'une de l'autre, au sens où si l'une augmente, l'autre augmente ou décroît «de la même façon».

Par exemple, il a été récemment démontré dans une université américaine que les étudiants qui sont les plus gros consommateurs d'alcool ont tendance à avoir des résultats scolaires moins bons que les autres. L'alcool est-il la cause directe de cette baisse de niveau ? N'allons pas trop vite en besogne. On peut certes imaginer que l'alcool rende stupide, mais d'autres hypothèses sont aussi envisageables : les étudiants qui boivent de l'alcool seraient à la base de moins bons étudiants ; ou bien ils seraient moins attentifs que les autres les lendemains de soirées arrosées, expliquant leurs déboires aux examens ; ou bien encore, ils boiraient plus que de raison afin de calmer leur peur d'échouer ou pour noyer des chagrins n'ayant rien à voir avec leurs études...

Ainsi, une corrélation manifeste n'est pas nécessairement la manifestation d'une relation de causalité : ce n'est pas parce qu'il y a des grenouilles après la pluie qu'on a le droit de dire qu'il a plu des grenouilles. Mais il arrive très souvent que nous confondions les deux notions, à la manière d'un Coluche conseillant de ne jamais aller à l'hôpital au motif qu'on y meurt en moyenne plus souvent que chez soi... Plusieurs sites recensent ce type de corrélations (voir la figure page ci-contre) relevant de l'effet cigogne : le taux de natalité augmentant comme le nombre de nids de cigognes, on en déduit que les cigognes apportent les bébés ! Il y a matière à sourire, mais le problème n'en demeure pas moins sérieux.

Il y a bien la possibilité qu'avec le *big data*, au lieu de théoriser, nous cédions aux facilités

de l'induction et délaissions le «geste théorique» de type galiléen, celui qui consiste à énoncer des hypothèses portant bien au-delà des données disponibles.

À cet égard, le cas d'Albert Einstein est exemplaire. En 1915, il publiait la théorie de la relativité générale, une conception révolutionnaire de la gravitation, alors qu'il n'avait que très peu de données sur l'Univers à sa disposition : on ignorait, par exemple, pourquoi les étoiles brillent, que d'autres galaxies hormis la

**N'allez jamais à l'hôpital, on y meurt en moyenne plus souvent que chez soi**

nôtre existent, que l'Univers est en expansion, etc. Cela n'a pas empêché les équations d'Einstein de parfaitement s'accommoder de la quantité gigantesque de données recueillies depuis un siècle par les télescopes et les satellites. En outre, elles constituent les fondations d'une véritable cosmologie scientifique, capable d'envisager l'Univers en tant qu'objet physique, doté de propriétés qui le caractérisent «lui-même».

Elles ont même permis de prédire l'existence des ondes gravitationnelles un siècle avant leur première détection, en septembre 2016. Une théorie peut donc non seulement enrichir l'univers des données, mais également agir comme un «treuil ontologique» capable de faire apparaître de nouveaux éléments de réalité. En d'autres termes, la théorie en «dit plus» que les données, notamment par le fait qu'elle explicite des lois que les données n'illustrent jamais que de façon partielle.

Imaginons maintenant que les choses se soient passées dans l'ordre inverse, c'est-à-dire que nous ayons commencé avec toutes les données dont nous disposons aujourd'hui, mais sans avoir à notre disposition la théorie de la relativité générale. Pourrions-nous, par une sorte d'induction théorique permettant de passer des données aux lois, découvrir les