Comme dit le proverbe, l'union fait la force. Cette vérité maintes fois éprouvée sur les champs de bataille, en politique, en affaires et en amour, s'applique également en statistique. Comme nous allons l'illustrer au moyen du concept de forêt aléatoire, la combinaison d'un ensemble de modèles simples permet souvent d'obtenir de meilleures prévisions que l'emploi d'un seul modèle, aussi complexe et sophistiqué soit-il.

L'union fait la force

Christian Genest
Université McGill
Jean-François Plante
HEC Montréal

« Les diamants sont éternels »

Roméo a récemment rencontré la femme de ses rêves et il veut la demander en mariage. Pour respecter des traditions qui lui sont chères, il désire lui offrir un diamant. En consommateur averti, il cherche d'abord à se renseigner sur les prix et les facteurs qui les influencent. Il déniche pour ce faire un fichier décrivant les principales caractéristiques et le prix de 53 940 diamants.

Dans cet échantillon, le prix moyen d'un diamant est de 3 932,80 $. Toutefois, les prix varient beaucoup. Pour mesurer la dispersion de données $x_1, \ldots, x_n$ autour de leur moyenne

$$m = (x_1 + \cdots + x_n)/n,$$

on calcule généralement la variance expérimentale, soit

$$s^2 = \{(x_1 - m)^2 + \cdots + (x_n - m)^2\}/(n - 1)$$

ou encore sa racine carrée, s, appelée l'écart-type. Dans le cas présent, Roméo trouve s = 3 989,44 $ et $s^2 = 15\,915\,629\,\$^2$. L'écart-type est donc du même ordre de grandeur que la moyenne, ce qui est le signe d'une très grande fluctuation dans les prix.

Classe	Effectif	Prix moyen
0,2 carat ou moins	12	386,17 $
0,2 à 0,4 carat	14 379	739,60 $
0,4 à 0,6 carat	10 057	1 466,49 $
0,6 à 0,8 carat	6 963	2 641,42 $
0,8 à 1 carat	5 027	4 243,70 $
1 à 1,5 carat	12 060	6 513,53 $
1,5 à 2 carats	3 553	11 321,77 $
Plus de 2 carats	1 889	14 951,25 $

Tableau 1.
Prix moyen des diamants en fonction de leur masse en carats. Les intervalles sont ouverts à gauche et fermés à droite.

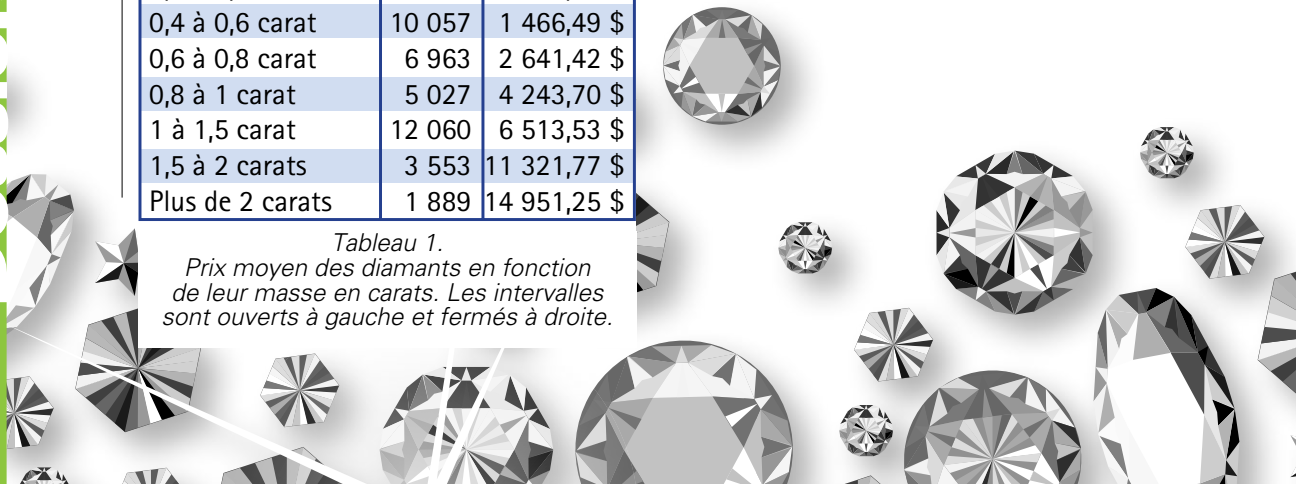
Pourquoi les prix varient-ils autant ?

Si les prix fluctuent tant, c'est que les caractéristiques des diamants sont différentes, à commencer par leur taille, mesurée en carats (un carat équivaut à 200 mg). En regroupant ses données en classes comme au tableau 1, Roméo constate en effet que le prix moyen d'un diamant augmente en fonction de son nombre de carats.

Pour voir dans quelle mesure le nombre de carats d'un diamant permet d'en prédire le prix, supposons un instant que chaque diamant se vende au prix de sa classe, tel qu'indiqué au tableau 1. Autrement dit, supposons que tous les diamants de 0,2 carat ou moins se vendent 386,17 $, que tous ceux qui font entre 0,2 et 0,4 carat se vendent 739,60 $, etc. Ceci n'affecterait pas le prix moyen d'un diamant, qui serait encore de 3 932,80 $, comme on peut s'en rendre compte en calculant la moyenne pondérée des données du tableau 1. Toutefois, la variance serait alors de $13\,416\,227\,\$^2$. Ce modèle de prévision du prix d'un diamant permettrait donc d'expliquer

$$13\,416\,227 / 15\,915\,629 = 84{,}3\,\%$$

de la variation totale. Pas mal, mais comme on va le voir dans la suite, on peut faire mieux.



Construire de meilleures classes

Lorsqu'il a construit le tableau 1, Roméo a créé ses classes arbitrairement. Il s'en est bien tiré, mais son modèle est perfectible. Une bonne manière de l'améliorer, fréquemment employée en pratique, consiste à procéder de façon itérative en maximisant à chaque étape le pourcentage de variance expliquée.

Dans un premier temps, Roméo scindera son échantillon en deux parties C_1 et C_2 selon que le nombre de carats est inférieur ou supérieur à un nombre c. En se servant d'un ordinateur, il trouvera la valeur de c pour laquelle le tableau résultant (qui compte alors deux lignes) explique le plus grand pourcentage possible de la variation totale. Puis, les classes C_1 et C_2 seront elles-mêmes scindées en deux selon le même critère. Ces scissions se feront aux points c_1 et c_2, respectivement. L'approche sera ensuite répétée une troisième fois, ce qui définira au final huit classes, comme c'était le cas au tableau 1. La différence est que cette fois, le choix des classes aura été effectué selon un critère bien précis plutôt qu'au petit bonheur la chance.

Cette procédure itérative peut être résumée graphiquement par une arborescence comme celle de la figure 1, dans laquelle les branches se prolongent de nœud en nœud jusqu'aux feuilles. Chaque feuille représente une classe à laquelle est rattaché un prix. Le tableau 2 correspondant permet d'expliquer 86,7 % de la variance totale, alors que le tableau 1 n'en expliquait que 84,3 %.

En termes mathématiques, le modèle de prévision ainsi créé s'exprime comme une combinaison linéaire de variables indicatrices. En effet, si x représente la masse d'un diamant en carats, alors le prix est déterminé par la formule en encadré :

$$\text{Prix} = 781{,}80 \cdot I_{[0;0{,}465[}(x) + 1\,675{,}05 \cdot I_{[0{,}465;\,0{,}625[}(x) + \cdots + 14\,834{,}69 \cdot I_{[1{,}915;\,+\infty[}(x)$$

où $I_A(x)$ est une variable qui vaut 1 ou 0 selon que la masse x en carats d'un diamant donné se trouve ou non dans l'intervalle A. On dit que c'est un *arbre de régression*. Le terme régression est justifié car la formule servant à prédire le prix est une somme pondérée de variables explicatives. Toutefois, ces variables ne prennent ici que les valeurs 0 et 1 au lieu d'être continues.

Classe	Effectif	Prix moyen
Moins de 0,455 carat	17 289	781,80 $
0,455 à 0,625 carat	7 498	1 675,05 $
0,625 à 0,865 carat	7 255	2 714,34 $
0,865 à 0,995 carat	2 838	3 938,64 $
0,995 à 1,175 carat	9 011	5 672,81 $
1,175 à 1,495 carat	3 814	7 243,42 $
1,495 à 1,915 carats	4 051	10 872,79 $
1,915 carat ou plus	2 184	14 834,69 $

Tableau 2.
Prix moyen des diamants en fonction de leur masse en carats. Les intervalles sont ouverts à droite et fermés à gauche.

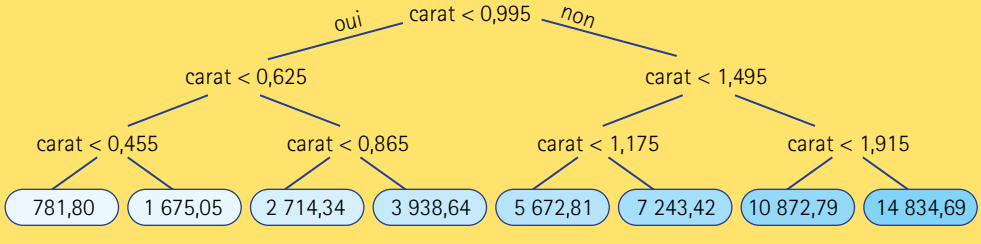


Figure 1. *Arbre obtenu en scindant successivement les groupes de diamants en deux sur la base de leur masse en carats. Chaque scission est faite de façon à maximiser la variance expliquée. Les montants inscrits dans les ovales bleus sont les prix des huit classes.*

Améliorer le modèle en y incorporant d'autres variables

Évidemment, il n'y a pas que le nombre de carats qui sert à fixer le prix d'un diamant. On peut penser par exemple à la clarté, la couleur et la coupe des diamants, ainsi qu'à leurs dimensions (longueur, largeur et profondeur) mesurées en millimètres. Comme toutes ces informations sont disponibles dans le jeu de données de Roméo, il pourrait songer, à chaque fois que vient le moment de scinder une classe en deux, à le faire sur la base de la caractéristique, au choix, qui permet de maximiser le pourcentage de variance expliquée.

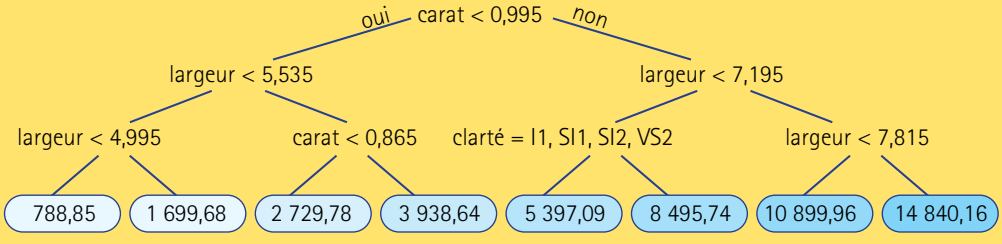


Figure 2. *Arbre obtenu en scindant successivement les groupes de diamants en deux. Chaque scission est opérée sur la base d'une seule des variables disponibles, au choix. La variable et le point de scission sont choisis de façon à maximiser la variance expliquée.*

En procédant ainsi, Roméo obtiendra un modèle statistique plus complexe, comme celui qui est représenté par l'arborescence de la figure 2. Dans ce modèle, la clarté d'un diamant est une caractéristique jugée importante pour en déterminer le prix s'il fait 0,995 carat ou plus et si sa largeur est inférieure à 7,195 mm. En revanche, seule la largeur et le poids jouent un rôle dans la prévision du prix d'un diamant de moins de 0,995 carat. Grâce à ce modèle, on peut dorénavant expliquer 88,9 % de la variance totale. On a donc encore gagné en efficacité, passant de 86,7 % (au tableau 2) à 88,9 % d'explication de la variance totale.

Évaluer un modèle de prévision

Dans le but d'accroître encore la proportion de variance expliquée par son modèle, Roméo pourrait considérer des arbres à 16 feuilles, à 32 feuilles, et ainsi de suite. En poursuivant la démarche à l'ultime, chaque feuille de son arbre représenterait éventuellement un seul diamant et 100 % de la variance observée dans son jeu de données serait alors expliquée. À première vue, ça semble idéal mais c'est aussi dire que le prix de chaque diamant dépend alors de manière unique de ses propriétés. Or au départ, le modèle visait à permettre d'identifier les principales caractéristiques des diamants qui en déterminent le prix. Si on le complexifie à l'extrême, il finira certes par se coller de très près aux données. Ce faisant, l'utilité de ce modèle risque cependant d'être limitée lorsqu'il s'agira de prédire le prix de futurs diamants car il reflétera moins les points communs à tous les diamants que la spécificité de ceux qui ont été utilisés pour sa construction.

Afin d'évaluer la valeur de son modèle comme outil de prévision, Roméo doit donc le valider sur des données qui n'ont pas servi à son élaboration. Pour y arriver, il n'est pas nécessaire de trouver de nouvelles données. Il suffit de n'utiliser qu'une partie du jeu de données initial dans la phase de construction du modèle ; on peut alors procéder à son évaluation avec les données restantes.

Supposons par exemple qu'après avoir tenu compte de son budget, Roméo décide de concentrer son attention sur les 6 042 diamants dont la masse varie entre 1 et 1,05 carat. Il en choisit 1 500 au hasard et les met de côté à des fins de validation. Les 4 542 autres diamants constitueront l'échantillon d'apprentissage qui lui servira à élaborer son modèle.

L'objectif de Roméo étant de prédire au mieux la valeur marchande d'un diamant en fonction de ses propriétés, il mesure l'erreur quadratique moyenne (EQM) de prévision, définie par

$$\text{EQM} = \frac{1}{1\,500} \sum_{i=1}^{1\,500} (\text{prévision}_i - \text{prix}_i)^2,$$

c'est-à-dire la moyenne du carré entre la prévision du prix d'un diamant fournie par le modèle et le prix avéré. Plus l'EQM est petite, plus le modèle est précis. Lorsque le calcul de l'EQM n'est effectué qu'avec les données d'apprentissage, il correspond à la portion non expliquée de la variation totale. Pour référence, l'EQM du meilleur modèle obtenu jusqu'ici par Roméo était de 1 774 211 ^{2}, soit environ 100,0 % – 88,9 % = 11,1 % de 15 915 629 ^{2}.

Après avoir déterminé son arbre de régression au moyen des données d'apprentissage, Roméo s'en sert pour prédire la valeur de chaque diamant de l'échantillon de validation en fonction de ses caractéristiques. Chacune de ces prévisions est ensuite comparée au véritable prix du diamant correspondant. Lorsqu'il est amené à choisir entre deux modèles, Roméo peut ainsi comparer leur performance et opter pour celui qui conduit à la plus petite EQM.

Élagage d'arbres

Les arbres construits jusqu'ici comptaient tous huit feuilles, mais ce nombre était arbitraire. On pourrait complexifier l'arborescence en lui ajoutant des branches. Pour déterminer le nombre optimal de feuilles, on peut se servir de l'EQM calculée sur l'échantillon de validation (1 500 diamants). La figure 3 montre la progression de l'EQM en fonction de la complexité de l'arbre, telle que mesurée par son nombre f de feuilles. La courbe en noir est calculée au moyen de l'échantillon de validation (1 500 diamants), alors que la courbe en rouge est obtenue à partir de l'échantillon d'apprentissage (4 542 diamants).

La figure 3 montre que les deux courbes décroissent rapidement jusqu'aux environs de $f = 50$ feuilles. Noter que sur le graphique, les courbes n'ont été tracées que pour $f \geq 10$. Par la suite, l'EQM calculée à partir de l'échantillon d'apprentissage (en rouge) continue à chuter, donnant faussement l'impression que le modèle ne cesse de s'améliorer. En revanche, l'EQM calculée au moyen de l'échantillon de validation (en noir) atteint un plateau correspondant à la zone dite de « sur-apprentissage », dans laquelle

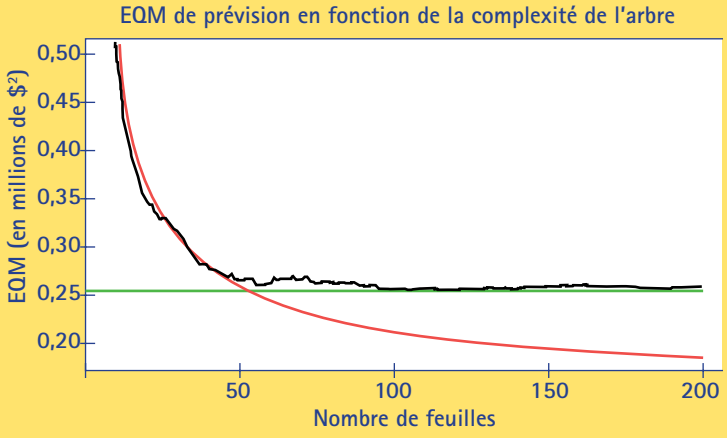


Figure 3. *EQM de prévision en fonction du nombre de feuilles dans un arbre prédisant le prix des diamants de 1 à 1,05 carat en se servant de leurs autres caractéristiques. L'EQM calculée sur l'échantillon de validation est tracée en noir. Son minimum, représenté par une ligne verte, est atteint à 105 feuilles. L'EQM calculée sur les données d'apprentissage est tracée en rouge.*

l'ajout de classes continue de complexifier le modèle sans en améliorer la valeur prédictive. Alors que la courbe rouge est décroissante, la courbe noire atteint un minimum en $f = 105$ feuilles ; sa valeur, qui est alors de 253 696 ^{2}, est indiquée par une ligne verte. En principe, c'est cette solution que Roméo retiendrait, car c'est alors qu'il dispose du modèle de prévision le plus performant en termes de simplicité et de précision. Au-delà de ce seuil, le sur-apprentissage produit des modèles dont la performance se dégrade.

Les quatre C des diamants

La qualité d'un diamant dépend de quatre caractéristiques :

- Carat : Le poids en carat du diamant.
- Coupe : Il existe différentes coupes de diamant, mais la qualité de cette coupe est également un enjeu puisqu'un diamant taillé selon des proportions idéales réfléchira davantage de lumière.
- Clarté : Un diamant peut contenir des impuretés. Selon leur nombre et leur taille, le diamant se voit attribuer une classe telle que VS_1 ou I_3.
- Couleur : Plus un diamant est clair, plus il est recherché. Les lettres de l'alphabet sont utilisées pour coder sa coloration. Jusqu'à F, le diamant est incolore. À partir de N, une teinte jaune est perceptible.

Forêts aléatoires

Sans mentionner à Juliette qu'il est en train de se renseigner sur le prix des diamants, Roméo lui fait part de son forage de données. Sa dulcinée lui demande alors pourquoi il s'astreint à choisir un seul modèle. Elle qui croit énormément au travail d'équipe lui fait valoir qu'en mettant leurs connaissances en commun, ses collègues de bureau parviennent souvent à générer de bien meilleures solutions qu'un seul individu. « Pourquoi alors ne pas mettre à profit les prévisions de plusieurs modèles ? », demande-t-elle.

Pas bête, mais quand bien même Roméo appliquerait son algorithme plusieurs fois aux mêmes données, il obtiendra le même modèle à tous les coups ! Pour générer une diversité salutaire, il doit donc introduire un élément de hasard dans la construction de ses arbres, comme la nature le fait si bien avec le vivant. Après consultation, il envisage alors deux façons de procéder :

i) Chacun des arbres sera construit à partir d'un échantillon d'apprentissage différent, un échantillon artificiel constitué à partir des données de départ en pigeant au hasard des éléments de l'ensemble avec remise (il y a donc un petit risque qu'un même diamant y apparaisse plus d'une fois, mais c'est sans conséquence). De cette manière, ces jeux de données artificiels auront la même distribution que les données originales, mais les modèles qu'on en tirera seront différents.

ii) À chaque étape de construction du modèle, la scission d'une classe ne pourra être basée, disons, que sur trois des huit variables explicatives disponibles. De plus, ces trois variables seront choisies au hasard à chaque étape.

En appliquant les stratégies i) et ii) simultanément à plusieurs reprises, disons 500 fois, Roméo obtient une multitude d'arbres de régression offrant chacun une prévision de prix légèrement différente pour chaque nouveau diamant. C'est ce qu'on appelle une *forêt aléatoire*. Sa prévision finale sera alors obtenue en faisant la moyenne des prévisions.

La validation croisée

Même à l'ère des méga-données, il arrive que les fichiers de données disponibles soient de petite taille et qu'il soit mal avisé d'en consacrer une partie à la validation. Dans pareil cas, on peut recourir à la validation croisée. Pour choisir entre un arbre à huit feuilles et une régression linéaire sans sacrifier de données, par exemple, on peut :

1. Diviser l'échantillon aléatoirement en $P = 10$ parties à peu près égales.
2. Utiliser à tour de rôle chacune des dix parties comme échantillon de validation et les neuf autres pour l'apprentissage.
3. Dans chaque cas, noter l'EQM obtenue, ce qui donne dix valeurs d'EQM pour chacun des deux modèles.
4. Calculer l'EQM moyenne pour chaque modèle. Utiliser cette valeur pour identifier le meilleur modèle.
5. Une fois le meilleur modèle identifié, utiliser toutes les données pour l'ajuster.

Avec la validation croisée, les données d'apprentissage ne servent jamais à mesurer la performance et le modèle final a le bénéfice de s'appuyer sur toute l'information disponible.

Bien que la valeur de P soit en principe arbitraire, on prend très souvent $P = 5$ ou 10 dans la pratique. Quand P est égal à la taille de l'échantillon, la méthode porte le nom de « jackknife ».

Si l'intuition de Juliette est bonne, la performance de cette forêt devrait être supérieure à celle du meilleur arbre. Et de fait, Roméo trouve une EQM d'environ 237 930 ^{2}, soit une valeur qui est 6,2% plus petite que celle obtenue avec le meilleur arbre. Bien sûr, si Roméo répète l'exercice, la nouvelle forêt aléatoire qu'il engendrera sera forcément différente de la première, ce qui conduira vraisemblablement à des résultats quelque peu différents, mais comparables.

Méthodes d'ensemble

Il y a, dit-on, plus d'idées dans deux têtes que dans une. C'est ce qui a motivé les chercheurs en apprentissage automatique à développer les forêts aléatoires et autres méthodes d'ensemble. Pour que l'approche soit viable, il faut bien sûr disposer de moyens informatiques conséquents.

Lorsqu'elle s'appuie exclusivement sur la procédure i), la stratégie décrite plus haut est appelée « bagging », pour « bootstrap aggregating ». Elle peut être appliquée à n'importe quel modèle et bien qu'elle ne soit pas infaillible, elle permet souvent d'obtenir d'excellents résultats.

Le « boosting » est une autre méthode d'ensemble populaire. Plutôt que de construire des modèles en parallèle, on commence par en ajuster un. Puis, on cherche à l'améliorer en faisant appel à d'autres modèles pour réduire l'erreur de prévision.

Même s'il n'y a à ce jour que peu ou pas de résultats mathématiques garantissant l'optimalité des méthodes d'ensemble, on constate qu'en pratique, elles sont souvent très performantes. De nombreux chercheurs tentent actuellement d'en percer les mystères.

À quoi ça sert?

Dans la vie de Roméo et Juliette, les méthodes d'ensemble joueront encore bien des rôles. Quand ils achèteront en ligne des éléments de décoration pour leur nid d'amour, les entreprises de commerce électronique feront appel à des modèles de « gradient boosting » pour associer les produits les plus pertinents aux mots clés de leurs recherches. De même, lorsqu'ils partageront leurs photos de mariage, des modèles de forêts aléatoires imbriqués dans l'architecture des réseaux sociaux détermineront sur quels serveurs en sauvegarder des copies en fonction de la probabilité que des amis de provenances géographiques et de milieux variés veuillent les télécharger. De plus, l'historique de navigation des tourtereaux pourrait avoir une incidence sur les taux que paieront les entreprises spécialisées dans l'industrie du mariage pour afficher leurs services sur la toile et sur les messages publicitaires que leur adresseront les agences de voyage.

Au fil du temps, des outils de prévision de plus en plus perfectionnés contribueront sans doute à faciliter la vie de Roméo et Juliette, et la nôtre. Mais comme les événements récents l'ont montré, la surveillance généralisée de nos moindres faits et gestes comporte aussi des pièges. Espérons que nous aurons collectivement la sagesse de les éviter.

Leo Breiman (1928-2005)

Le regretté statisticien Leo Breiman (1928-2005) est considéré comme l'un des pionniers des forêts aléatoires. Il a grandement contribué à développer les méthodes d'ensemble en général. Il est l'auteur du livre « Classification and Regression Trees » dont une réédition est parue en 2017. Il est reconnu pour avoir contribué à rapprocher la statistique et l'apprentissage machine. Il a décrit les différences qu'il percevait entre ces deux cultures scientifiques dans un article paru en 2001 (voir *Pour en savoir plus*).