Epistémologie des classifications

Daniel Parrochia

(Université Jean Moulin - Lyon 3)

daniel.parrochia@wanadoo.fr

Avant-propos

- Tout le monde connaît la classification chinoise mentionnée par Borgès et Foucault classant les animaux en :
  - a) appartenant à l'empereur ; b) embaumés ; c) apprivoisés; d) cochons de lait ; e) sirènes ; f) fabuleux ; g) chiens en liberté ; h) inclus dans la présente classification ; i) qui s'agitent comme des fous ; j) innombrables ; k) dessinés avec un pinceau très fin en poil de chameau ; l) et cetera ; m) qui viennent de casser la cruche ; n) qui de loin semblent des mouches.

Le problème est que :

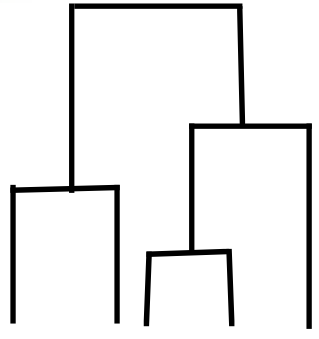
- 1) Les classes s'intersectent ;
- 2) Certaines d'entre elles sont indéfinissables
  - soit parce que la multiplicité qu'elles sont censées contenir est mal définie («et cetera» ou encore «innombrables»),
  - soit parce qu'elle est fonction du temps, et donc en perpétuelle modification («qui viennent de casser la cruche»);
- 3) Une classe au moins (la classe des animaux «inclus dans la présente classification») rend la classification paradoxale
- => Foucault en conclut que cet ordre est pour nous impensable, et il en déduit que, puisqu'il est satisfaisant pour un chinois, tous les ordres sont arbitraires.

Buts de l'exposé

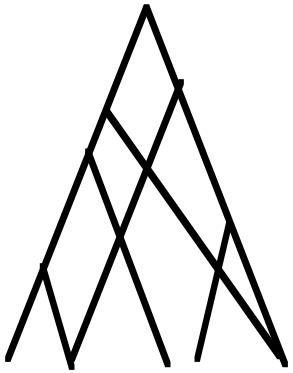
- Le présent exposé voudrait s'opposer point par point à une telle thèse, qui est, à mon avis, totalement erronée.
- Il entend au contraire montrer que :
  - 1) Les ordres ne sont pas arbitraires
  - 2) Certaines classifications sont meilleures que d'autres.
  - 3) On peut penser une théorie universelle des classifications, c'est-à-dire l'existence d'une classification absolue.

Définition

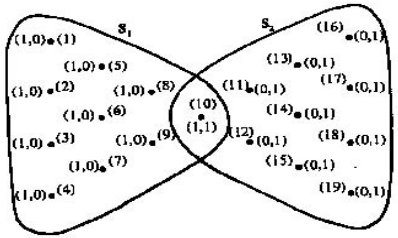
- On appelle « classification » l'opération consistant à distribuer, répartir ou ventiler des individus ou leurs propriétés dans des classes finies ou infinies, hiérarchisées ou non, nettes ou floues, empiétantes ou pas.



Hiérarche de partitions



hiérrchie de recouvrements



partition empiétante

Ellipsoïdes de Neuville

Buts de l'opération

- Simplifier la réalité en réduisant le nombre de ses objets (on substitue aux individus des classes moins nombreuses qu'eux).

- Valider à partir de là certaines hypothèses, faire des prédictions, générer des hypothèses nouvelles

Les deux grands types de classifications

- On peut distinguer, philosophiquement, deux grands types de classifications :
  - Des classifications empiriques ou « phénoménales » (classifications zoologiques, botaniques, médicales, sociologiques, etc.) fondées sur des ressemblances ou dissemblances perceptives ;
  - Des classifications « métempiriques » ou « nouménales » (classifications physico-mathématiques) fondées des invariants plus ou moins abstraits)

1
Les classifications empiriques ou « phénoménales »

Les deux grandes familles de méthodes

- Dans le domaine empirique, deux grandes familles de méthodes taxinomiques se détachent :
  - Les classifications qui procèdent par voie ascendante (qui sont celles qui sont couramment pratiquées par les naturalistes)
  - Les classifications qui procèdent par voie descendante (qui peuvent avoir leur utilité, notamment quand on vise des rangements de type dichotomique ou purement pragmatique - par exemple, dans le domaine bibliothéconomique)

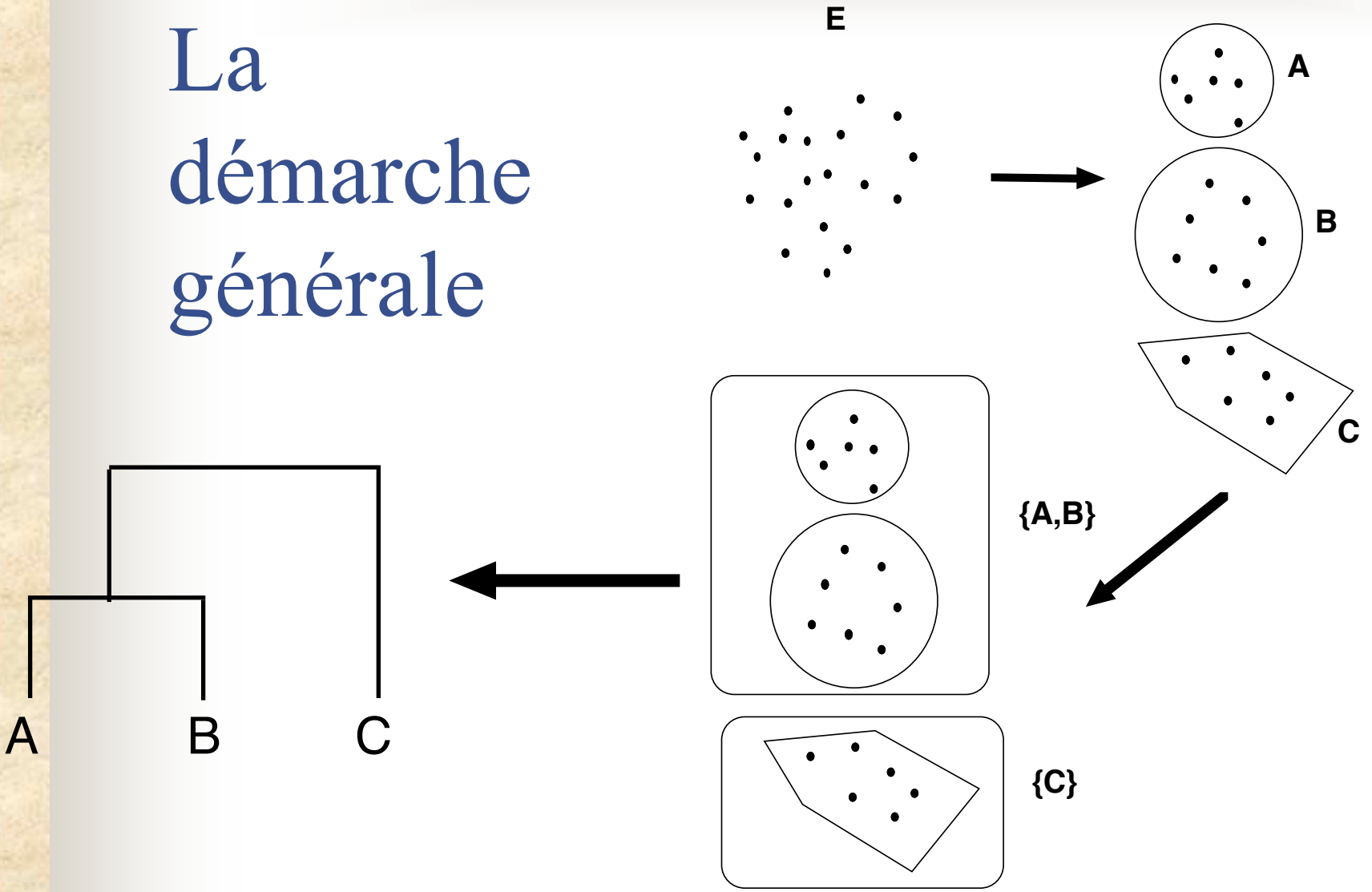
La méthode ascendante.

- Etant donné un ensemble fini d'objets à classer, la démarche générale (cas ascendant) consiste :
- 1° A évaluer les ressemblances entre les individus au moyen d'une mesure de proximité (indice de distance) afin de former des classes.
- 2° A évaluer ensuite la proximité relative des classes elles-mêmes, afin de les regrouper en classes de classes, et de substituer ainsi à l'ensemble de départ un ensemble hiérarchique-ment structuré.

La classification hiérarchique polythétique ascendante

- Les naturalistes du XVIIIe
  - Buffon, *De la manière de traiter et d'étudier l'histoire naturelle*, Paris, 1749
  - M. Adanson, *Histoire naturelle du Sénégal*, Paris, 1757 : « Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances… Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble, afin d'en faire un tout dont les parties soient unies et liées intimement »

La démarche générale



La mesure des proximités

- Pour classer une entité dans un groupe, il faut pouvoir évaluer le degré de proximité des entités à classer de façon précise. Cela se fait aujourd'hui mathématiquement. Deux méthodes sont possibles:
  - Mesurer les ressemblances à l'aide d'un indice de similarité
  - Mesurer les dissemblances à l'aide d'un indice de dissimilarité
- Ces indices sont en très grand nombre et l'on pourrait craindre qu'ils mènent à des préordonnances classificatoires très différentes.
- En fait, il n'en est rien.

Un résultat encourageant

- Lerman (1970) a montré que, lorsque le nombre d'attributs d'un objet variait peu quand on changeait l'indice de similarité, alors, les méthodes fondées sur la donnée d'une pré-ordonnance associée à une similarité prenaient un caractère intrinsèque.
- En d'autres termes, elles ne dépendaient pas du choix de la mesure de similarité.

Les méthodes descendantes

- A l'inverse des constructions ascendantes, les méthodes descendantes procèdent par dichotomies ou scissions successives d'une classe à chaque étape de la démarche.
- Les problèmes rencontrés sont donc :
  - Le choix de la classe à scinder (la plus nombreuse, ou celle de plus grand diamètre, ou encore celle de plus grande dispersion);
  - Le mode d'affectation des objets aux sous-classes (on décidera de mettre par ex dans la même classe les objets possédant la même modalité pour une variable considérée. Mais il y a bien d'autres solutions possibles).

Appréciation des méthodes précédentes Affaiblissements possibles

- Les méthodes ascendantes sont généralement plus fiables que les méthodes descendantes et moins coûteuses en hypothèses (Roux, 1985).
- En tout état de cause, l'informatique permet aujourd'hui des sorties multiples, fondées sur des indices de similarité ou des critères d'agrégation des classes différents..
- Ces méthodes peuvent être multiplement affaiblies :
  - A la place d'une partition (resp. hiérarchie de partitions), on peut se contenter d'un recouvrement (resp. une hiérarchie de recouvrements). (Chandon et Pinson, 1981)
  - A la place de partitions ou classifications nettes, on peut souhaiter des partitions ou classifications floues (Kaufmann, 1977)
  - Arbres hybrides (Carrol et Chang 1973) et groupes empiétants (Shepard et Arabie, 1979) sont aussi possibles

Type de classifications visé par ces modèles

- Classifications zoologiques et botaniques
- Typologie de profils biologiques et pathologiques
- Classifications de données archéologiques
- Taxinomies géographiques (étude régionale de l'agriculture française).
- Classifications de type psycho-sociologique :
  - Typologie des personnages d'enfants dans la littérature enfantine
  - Facteurs de perception de la misère en Europe
  - Taxinomie d'objectifs cognitifs en mathématiques
- Etc.

2
Les classifications physico-mathématiques ou « nouménales »

De nouvelles formes de classifications

- Toutes les classifications ne sont pas de la même nature que celles que nous avons évoquées jusqu'ici.
- Considérons par exemple :
  - La classification de Mendeleiev en chimie
  - La classification des cristaux
  - La classification des particules élémentaires
  - La classification des étoiles
  - Les classifications d'objets mathématiques (groupes, variétés, espaces topologiques, etc.)
- Dans tout ces domaines, il ne s'agit plus de classer les objets selon des ressemblances purement perceptives ou seulement extérieures.

L'opérateur de discrimination

- Dans tous les domaines précédemment cités, l'opérateur qui permet de ventiler les objets dans les classes est lié, de façon intrinsèque, à la nature profonde du domaine à classer et aux théories (physico-mathématiques) qui en rendent compte. Ce peut être,par exemple :
  - La structure du noyau atomique
  - Les symétries fondamentales de la nature
  - Un équilibre entre des lois physiques à l'œuvre dans le phénomène
  - Des invariants mathématiques abstraits.

Spécificité des classifications nouménales

- Dans ce cas-là, il demeure des conventions qui sont celles des dénominations, des unités de mesure et une contingence liée à l'histoire et à l'état de développement de la discipline du domaine à classer;
- Mais, une fois ces conventions posées, il n'y a plus la moindre arbitrarité dans les choix qui sont laissés.
- En d'autres termes, une fois qu'on sait, par exemple, ce qu'est un groupe fini simple, il n'y a qu'une seule classification possible de ces objets, qui est d'ailleurs achevée (17 familles infinies + 26 groupes isolés dits « sporadiques »).

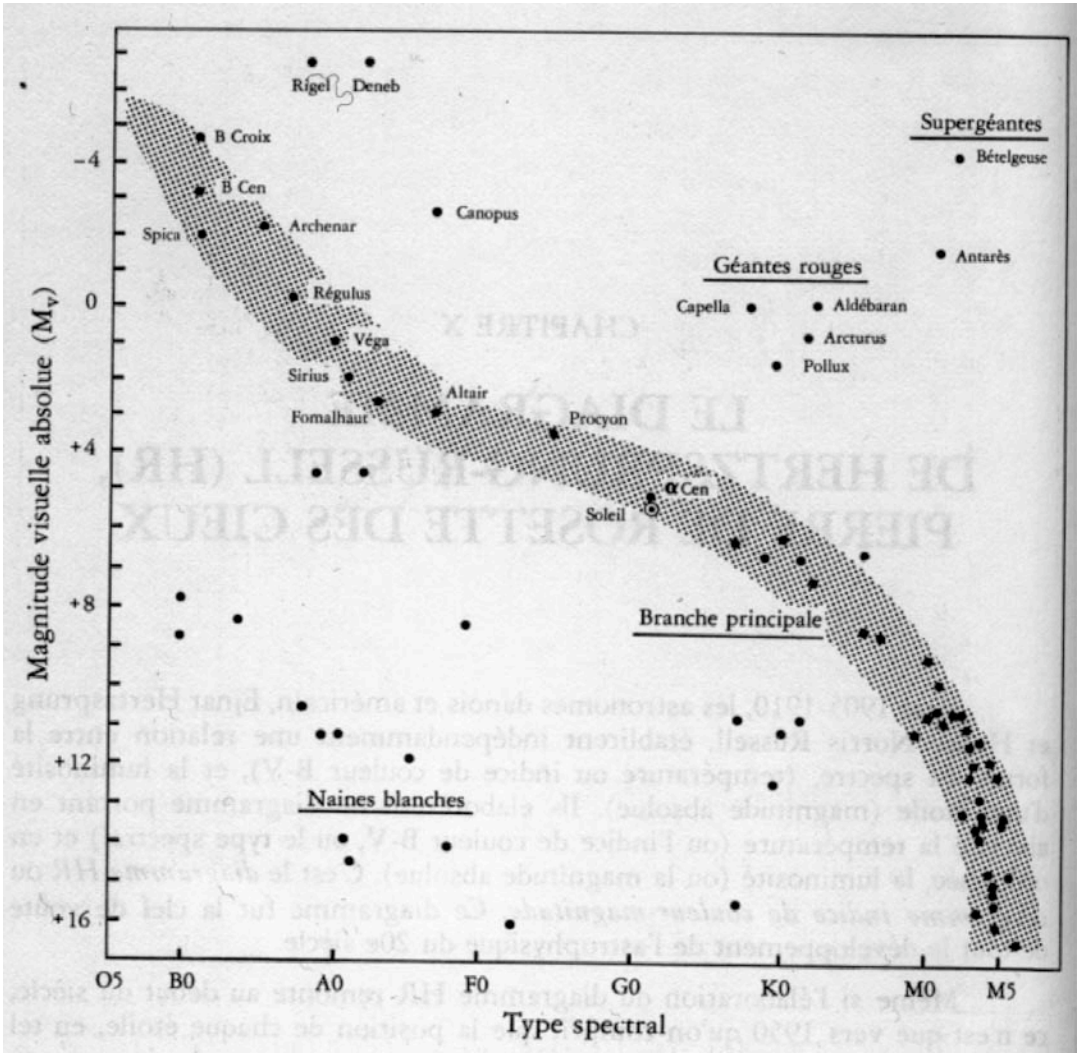
Propriétés fondamentales de ce type de classifications

- Stabilité fondamentale (Mendeleiev = bientôt 2 siècles d'existence)
- Prédictibilité : non seulement l'afflux d'éléments nouveaux ne bouleverse pas la taxinomie existante mais il est prévu par la structure et se loge dans ses trous
- A terme, complétude et exhaustivité de la taxinomie : d'ores et déjà, certaines classifications de structures mathématiques sont considérées comme achevées.
- => Ceci rend évidemment relatif le relativisme.

Exemple : la classification stellaire

- Une étoile (un état stellaire) c'est trois équilibres : hydrostatique, radiatif et convectif, dont on peut prendre connaissance par deux paramètres :
  - La luminosité (magnitude)
  - La composition chimique (celle-ci étant accessible par le type spectral de l'étoile).
- Avec ces deux éléments, les astronomes (resp. danois et américain) Ejnar Hertzsprung et Henry Norris Russell ont construit, vers 1905-1910, le célèbre diagramme HR qui porte leur nom et qui restera valable 10^{50} ans au moins :

Diagramme Hertzsprung-Russell



Sens du diagramme

- La branche principale (le S couché) réunit les étoiles les plus chaudes qui sont aussi les plus lumineuses ainsi que les plus froides qui sont les moins lumineuse. Cette relation régit 90% des étoiles
- Aux extrêmes on a les étoiles qui ont épuisé une partie de leur combustible :
  - Les géantes rouges, très lumineuses mais peu chaudes;
  - Les naines blanches (l'évolution ultérieure) peu lumineuses mais très chaudes..

Autre exemple : classification des cristaux

Systèmes cristallins	*Holoédries*	*Hémiédries*	*Tétartoédries*
Cubique	$\frac{3\,A^4}{3\,P}$ $4\,A^3$ $\frac{6\,A^2}{6\,P}$ C	$3\,A^4\,4\,A^3\,6\,A^2$ $3\,A^2\,4\,A^3\,6\,P$ $3\,A^2\,4\,A^3\,C$	$3\,A^2\,4\,A^3$
Hexagonal	$\frac{A^6}{P}$ $\frac{3\,A'^2}{3\,P'}$ $\frac{3\,A''^2}{3\,P''}$ C	$A^6\,3\,A'^2\,3\,A''^2$ $A^{-6}\,3\,A'^2\,3\,P'$ $A^6\,3\,P'\,3\,P''$ $\frac{A^6}{P}$ C	A^6 A^{-6}
Quadratique	$\frac{A^4}{P}$ $\frac{2\,A'^2}{2\,P'}$ $\frac{2\,A''^2}{2\,P''}$ C	$A^4\,2\,A'^2\,2\,A''^2$ $A^{-4}\,2\,A'^2\,2\,P'$ $A^4\,2\,P'\,2\,P''$ $\frac{A^4}{P}$ C	A^4 A^{-4}
Rhomboédrique	$A^3\,\frac{3\,A^2}{3\,P}$ C	$A^3\,3\,A^2$ $A^3\,3\,P$ $A^3\,C$	A^3
Orthorhombique	$\frac{A^2}{P}$ $\frac{A'^2}{P'}$ $\frac{A''^2}{P''}$ C	$A^2\,A'^2\,A''^2$ $A^2\,P'\,P''$	
Monoclinique	$\frac{A^2}{P}$ C	A^2 P	
Triclinique	C	O	

La cristallographie se développe à travers Romé de l'Isle (1783), Haüy, Berthollet (1811), Beudant et Groth (1830), Dela-fosse.(1840), Pasteur (1848), Bravais (1850). Tant qu'on n'a pas la théorie des groupes, l'inventaire des symétries reste incomplet.

Au XXe siècle, Shoenflies et Fedorow, indépendamment, dénombrèrent sans problème l'existence des 230 groupes ou combinaisons possibles d'opérations de symétrie, dont les 32 classes cristallographiques, les 14 modes réticulaires et les 7 systèmes fondamentaux d'Haüy et Bravais ne sont qu'un cas particulier.

3

Le problème mathématique

Rappel sur l'opération de classification

- L'opération de classification suppose :
  - 1°) La définition d'un certain indice de distance entre les éléments à classer;
  - 2°) La constitution, à partir de cette distance, d'une partition ou d'une préordonnance traduite par un arbre de hiérarchie, qui est fonction de cette distance.
- Essayons de représenter cela de façon un peu plus précise.

Comment définir une distance?

- Etant donné un ensemble de points $S = \{1, 2, ..., n\}$, représentant des objets à classer, on dira qu'une distance $d : S \times S \rightarrow \mathbf{R}$, est une fonction telle que, pour tout couple de points $i,j \in S$, on ait :
  - (1) $d(i,j) \geq 0$
  - (2) $d(i,j) = 0$ si et seulement si $i = j$
  - (3) $d(i,j) = d(j,i)$
  - (4) $d(i,j) \leq d(i,k) + d(k,j)$ (éventuellement)

Qu'est-ce que l'opération de classification?

- Une opération de classification est une fonction classifiante f qui, à tout indice de distance d défini sur un ensemble S de points, associe une partition Γ de S.
- Il est naturel (cf. Jardine et Sibson, 1967) qu'on réclame que cette fonction f ait de bonnes propriétés.
- Trois d'entre elles paraissent absolument nécessaires : l'invariance d'échelle, la richesse, et la consistance.
- Précisons ce qu'on entend par là.

Propriétés de la fonction classifiante f

- *Invariance d'échelle* : Pour toute fonction de distance d et tout $\alpha > 0$, on doit avoir : $f(d) = f(\alpha.d)$ (stabilité par rapport aux unités de mesure)
- *Richesse* : Le domaine de f doit être l'ensemble de toutes les partitions Γ telles que $f(d) = \Gamma$ (Ceci veut dire que,. pour toute partition Γ, on pourra toujours trouver une fonction de distance d sur S, de telle sorte que $f(d) = \Gamma$.)
- *Consistance* : Soit d et d', deux fonctions de distance. Si $f(d) = \Gamma$, et si d' est une transformation de d, alors $f(d') = \Gamma$. (si, par ex. les distances intra-classes sont réduites et les distances interclasses sont augmentées, on aura la même partition résultante.)

Un théorème d'impossibilité

- Jon Kleinberg (2002) a démontré le « théorème d'impossibilité » suivant :
- *Théorème* : Pour tout nombre d'objets $n \geq 2$, il n'y a pas de fonction classifiante f qui satisfasse à la fois la condition d'invariance d'échelle, la condition de richesse et la condition de consistance.
- Faut-il alors renoncer à une théorie générale des classifications?
- Il faut se méfier des soi-disant théorèmes d'impossibilité (cf. le cas d'Arrow en économie). En réalité, pour tourner le problème il suffit d'affaiblir légèrement les exigences de Kleinberg.

L’affaiblissement des exigences

- Si on accepte d’affaiblir les exigences, trois solutions existent :
  - – Ou on garde l’invariance d'échelle et la consistance, et on tolère que la fonction f soit moins riche (cela veut dire que, pour d ≠ d’, f(d) ne mènera pas toujours à une partition très différente.
  - – Ou on maintient la richesse et la consistance et on affaiblit l'exigence d'invariance d'échelle (on réclamera seulement qu’en cas de distances proportionnelles on ait des partitions qui puissent être des raffinements l’une de l’autre)
  - – Ou on garde invariance d’échelle et richesse mais on affaiblit l'exigence de consistance. (Cela veut dire qu’au lieu que, pour une transformation de d, f redonne la même partition, on réclamera seulement qu’elle redonne une partition proche.

4

Conclusion : pour une théorie générale des classifications

Origine du problème : A. Comte

- L'expression « théorie générale des classifications » apparaît dès la 2e Leçon du *Cours de Philosophie positive* et désigne la réflexion théorique (exactement « les travaux philosophiques ») des botanistes et des zoologistes qui sont supposés fournir un « guide certain » en dégageant les véritables principes de l'art de classer.
- Comte développe ensuite ces conceptions dans la 36e leçon (sur la chimie) et surtout dans les leçons 40 à 42 sur la biologie.
- Aujourd'hui, l'idée d'une TGC ne peut prendre qu'une forme mathématique.

Les approches mathématiques du problème

- La constitution d'une théorie générale des classifications peut se faire :
  - Soit sur un mode inductif, en généralisant progressivement les indices, les critères et les structures mis en place par les classificateurs :
  - Soit un mode déductif, à partir de techniques mathématiques générales qui relèvent :
    - Soit de la logique pure (S. Shelah)
    - Soit de la théorie des ensembles (Jech)
    - Soit de la théorie des catégories (R.S. Pierce)

L'idée-limite d'une théorie générale des classifications

- Quelque soit la démarche, une classification générale de toutes les structures finies ou infinies (classification absolue) débouche sur différentes questions fondationnelles en mathématiques :
  - La construction du continu
  - La construction des infinis d'ordre supérieur
- La mathématique, ici, rencontre ses limites car l'existence de structures hiérarchiques infinies (arbres de Kurepa, arbres d'Aronszajn, etc.) repose en fait sur des axiomes indépendants de la théorie ZF.

Questions finales

- Plusieurs questions, finalement, se posent donc au mathématicien qui voudrait développer une théorie générale des classifications :
  - Peut-on exprimer une telle théorie générale dans le cadre de la mathématique ordinaire?
  - Doit-on imaginer plusieurs théories, en fonction des options choisies?
  - L'une de ces théories est-elle plus pertinente qu'une autre ou toutes sont-elles équivalentes?
- Ces questions sont loin des préoccupations ordinaires des taxinomistes, pourtant, elles les commandent. On ne saurait envisager, en fait, de « vraies » classifications sans y répondre.
- Pour l'instant, on n'a pas la réponse à ces questions.Mais un jour viendra où il y en aura une. On aura alors une théorie universelle des classifications.