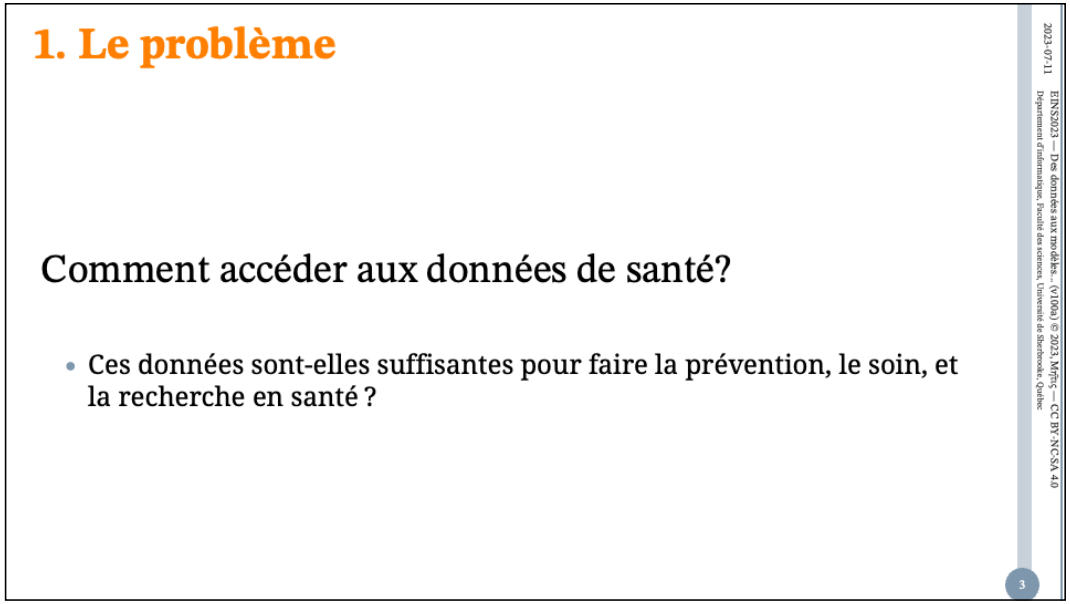

EINS
École d'été interdisciplinaire en numérique de la santé
Des données aux modèles, de la logique aux relations
EINS2023
v100a
2023-07-11
Luc.Lavoie@USherbrooke.ca
© 2018-2021, Μῆτις (http://info.usherbrooke.ca/llavoie)
CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Plan



Ces données sont-elles suffisantes pour faire la prévention, le soin, et la recherche en santé ?

1. Le problème (bis)

Comment accéder aux données *requises pour la prévention, le soin et la recherche* en santé?

- La signification d'une donnée est-elle toujours univoque?
- Est-ce que la donnée est suffisante en elle-même?

1. Le problème (ter)

Comment accéder aux données requises pour la prévention, le soin et la recherche en santé *et les interpréter (correctement)*?

- Ce problème est-il susceptible d'avoir une solution consensuelle, indépendamment des cultures, des sociétés, des états, des ordres professionnels et des intérêts privés?
- Ce problème est-il unique ou ne faudrait-il pas considérer une famille de problèmes, é une famille de solution?

Il y aura plusieurs solutions partielles aux problèmes d'accès aux données requises pour la prévention, le soin et la recherche en santé.

La couverture de ces solutions demeurera partielle et, pour cette raison, il serait souhaitable qu'elle soit interopérable.

L'interprétation de ces données demeurera plurielle et, pour cette raison, il serait souhaitable qu'elle se fasse sur la base de processus et de modèles documentés, ouverts et traçables.

Puisqu'il y aura plusieurs modèles, il importe de se doter d'une démarche pour les élaborer.

Puisqu'il y aura plusieurs modèles, il faut déterminer un méta-modèle facilitant

- *le raisonnement*
- *l'interopérabilité*
- *la documentation*
- *l'ouverture*
- *la traçabilité*

2. Une démarche

- Caractériser le problème
- Caractériser les solutions
- Concevoir, décrire et vérifier les solutions
- Choisir certaines solutions
- Expérimenter et valider celles-ci
- Tirer les conclusions

Explicitons maintenant notre démarche tout en l'appliquant à notre premier problème: trouver LE méta-modèle!

3. Un cas

3.1 Caractérisation du problème
3.2 Caractérisation des solutions
3.3. Description d'une première solution
3.4. Vérification de la solution
3.5. Une première expérimentation
3.6. Les premières conclusions

3. Un cas
3.1 Caractérisation du problème

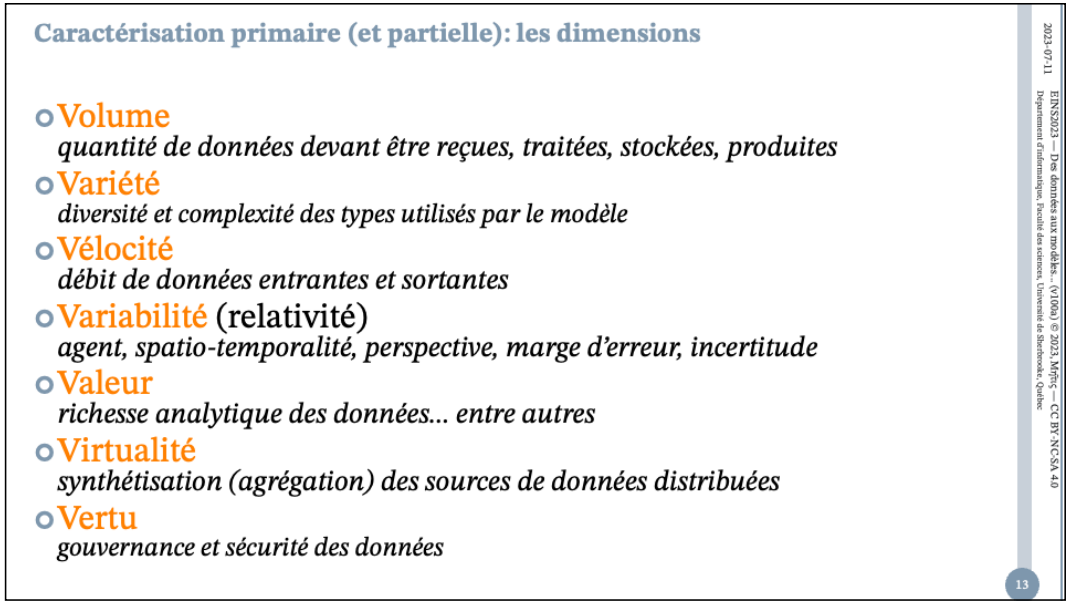
- Cerner un problème afin d'y trouver une solution, passe par la modélisation et l'analyse dudit problème et donc de sa caractérisation.

- Comparer un problème avec d'autres est riche en enseignement, voire en solutions existantes ou adaptables.

- Choisir une solution, parmi d'autres, passe par la comparaison de celles-ci aux autres solutions.

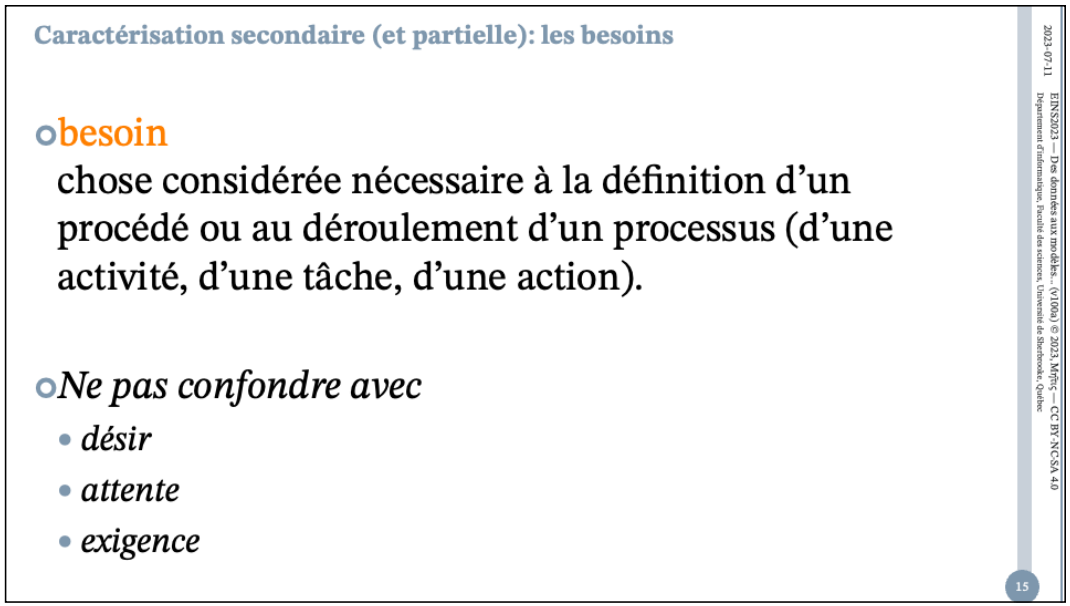
Commençons par deux formes de caractérisation:

- *les dimensions*
- *les problèmes!*



* La mesure de ces dimensions n'est pas aussi triviale qu'on le voudrait.
* Les procédés et les techniques ne font pas (encore) consensus.

- Volume
  pas considérable hors génomique et imagerie
- Variété
  considérable
- Vélocité
  pas considérable
- Variabilité (relativité)
  considérable
- Valeur
  perception très variable
- Virtualité
  considérable
- Vertu
  complexe, sensible, débattue



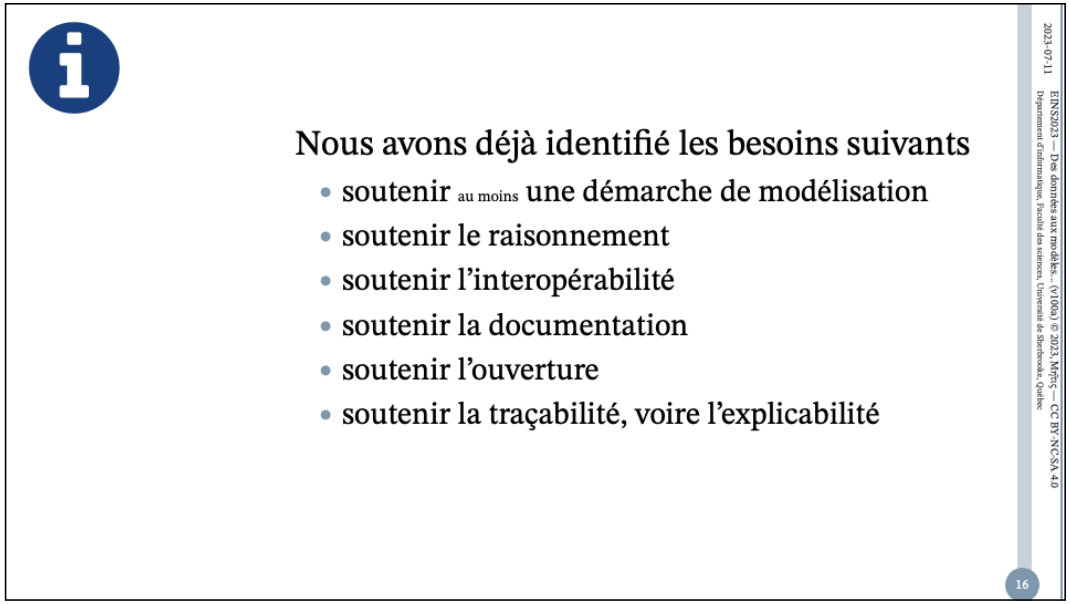
procédé::
...

processus::
...

activité::
ensemble de…

tâche::
action portant sur une chose

chose::
objet matériel ou information



* Voir également article Khnaisser et coll. pour un approfondissement.

3. Un cas
3.2 Caractérisation des solutions

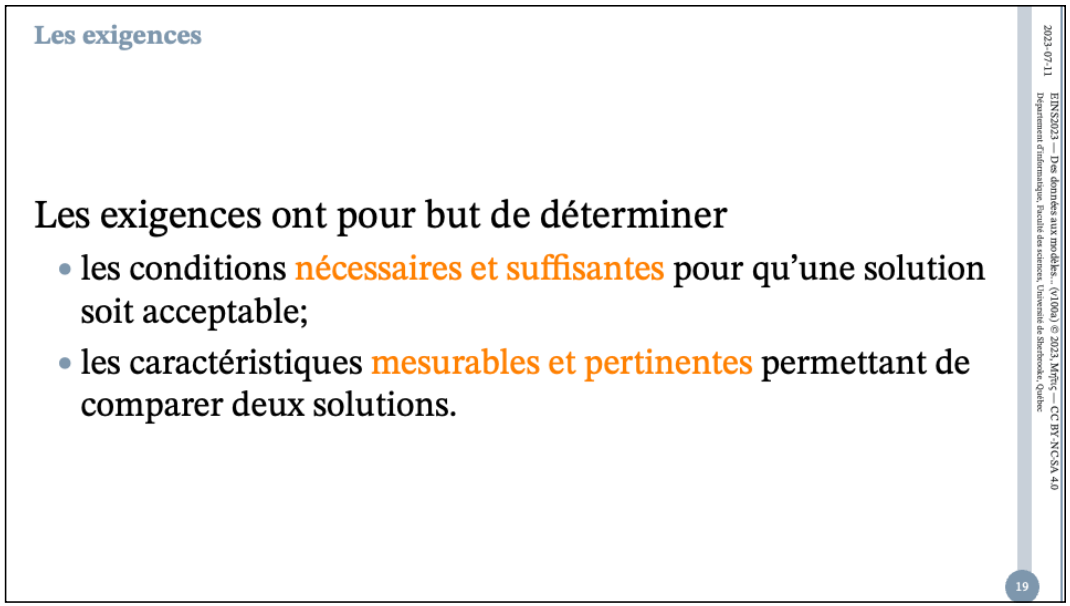
- En présence d'une solution, il faut s'assurer qu'elle en est bien une.
- En présence de plusieurs solutions, il faut pouvoir les comparer.

Commençons par une forme de caractérisation,

- les exigences

et une façon de les utiliser,

- l'adéquation.

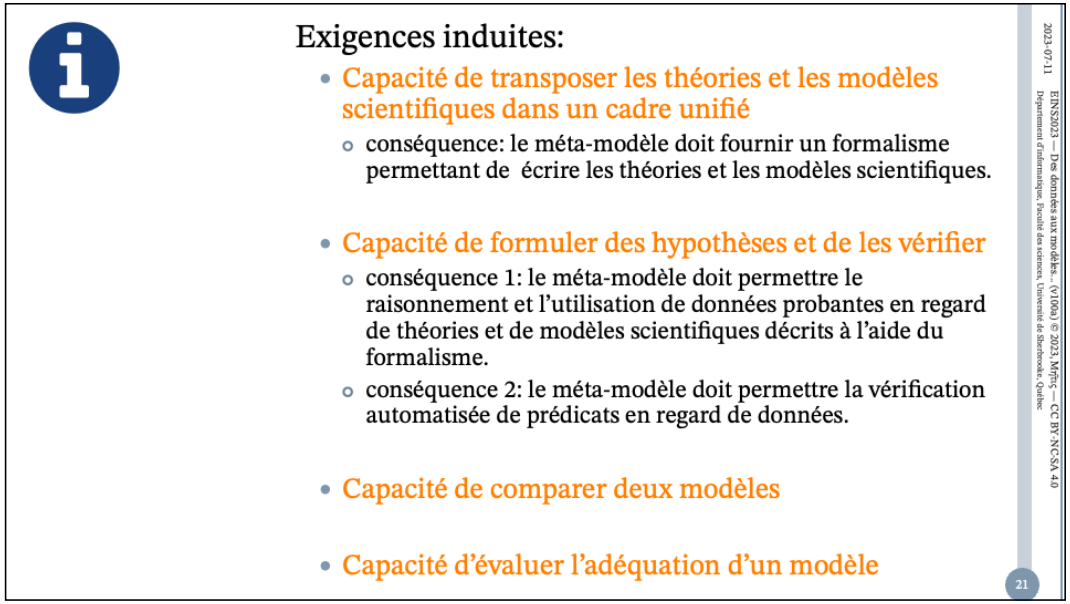


Plusieurs procédés ont été proposés afin de permettre la détermination des exigences.
Ces procédés se distinguent par l'organisation des activités suivantes et les
méthodes préconisées pour chacune.

* exploration
* analyse
* spécification
* vérification
* validation

Exigences pratiques

- Capacité de soutenir un processus de diagnostic puis de choix de traitement
  (fondée sur des données probantes en regard de pratiques médicales reconnues);
  - corolaire 1: capacité de décrire l'état du patient,
  - corolaire 2: capacité de formuler le diagnostic,
  - corolaire 3: capacité de décrire un traitement,
  - ...
- ...



Remarquons que les deux capacités induites sont applicables tant au méta-modèle qu'aux modèles qu'il permet de décrire (modéliser).

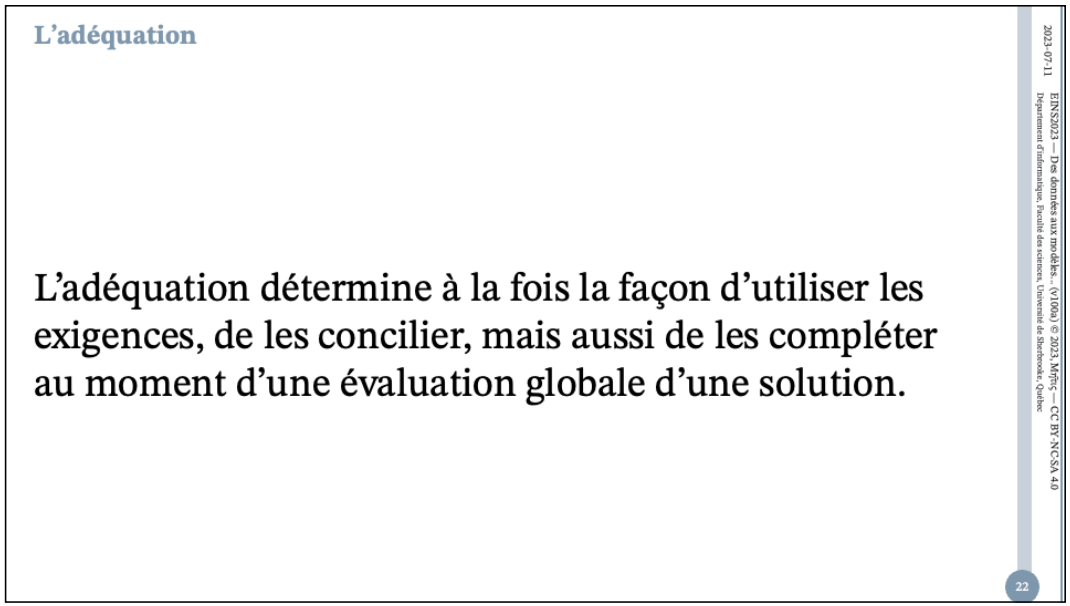
Le processus d'exploration n'est pas terminé :

  * une clarification est certainement possible ;
  * d'autres exigences découlent des besoins exposés ;

Un processus de spécification devrait s'ensuivre afin

  * de s'assurer que chaque exigence soit
    - claire (c'est-à-dire lisible et compréhensible),
    - exacte (c'est-à-dire précise et sans erreurs),
    - complète (c'est-à-dire comprenant tous les éléments requis et tous les éléments nécessaires) et
    - concise (c'est-à-dire sans éléments superflus);
  * d'associer à chaque exigence un critère objectif permettant de vérifier qu'elle est satisfaite.

Finalement, la vérification et la validation devraient s'ensuivre afin de
minimiser les risques de devoir recommencer la modélisation qui en découlera.

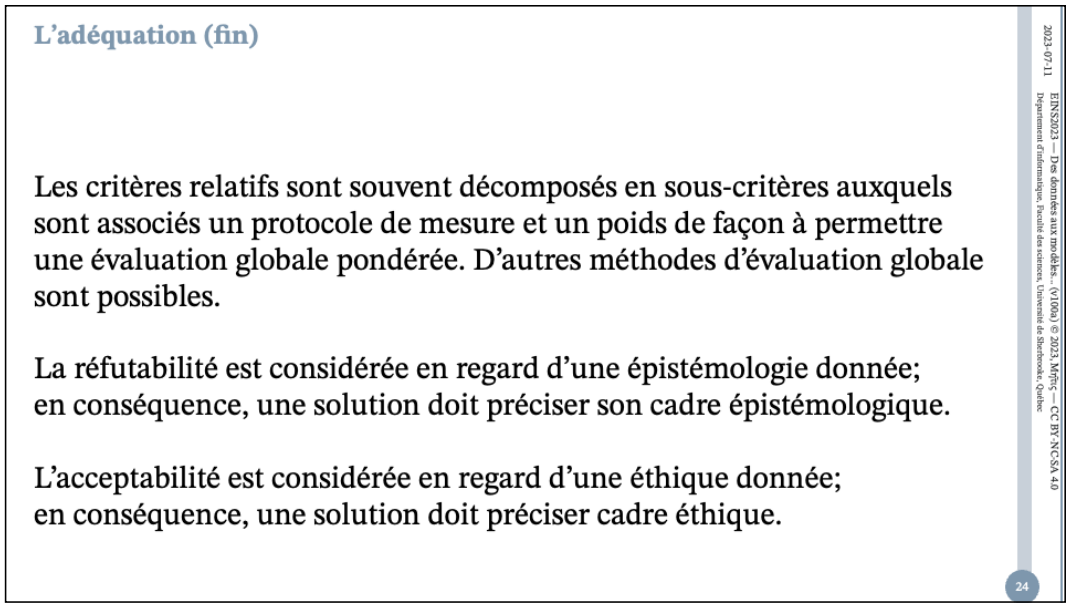


Plusieurs procédés ont été proposés afin de permettre la détermination des exigences.
Ces procédés se distinguent par l'organisation des activités suivantes et les
méthodes préconisées pour chacune.

* exploration
* analyse
* spécification
* vérification
* validation

L'adéquation (suite)

- Critères absolus
  - Validité (conformité au modèle)
  - Efficacité (conformité aux exigences)
  - Cohérence (non-contradiction interne)
- Critères relatifs
  - Complétude (couverture «suffisante» du problème)
  - Efficience (consommation «acceptable» de ressources)
  - Évolutivité (adpatation «aisée» aux changements)
- Méta-critères
  - Réfutabilité (permet d'exprimer et d'évaluer les «falsificateurs potentiels»)
  - Acceptabilité (permet d'exprimer et d'évaluer les «critères éthiques»)

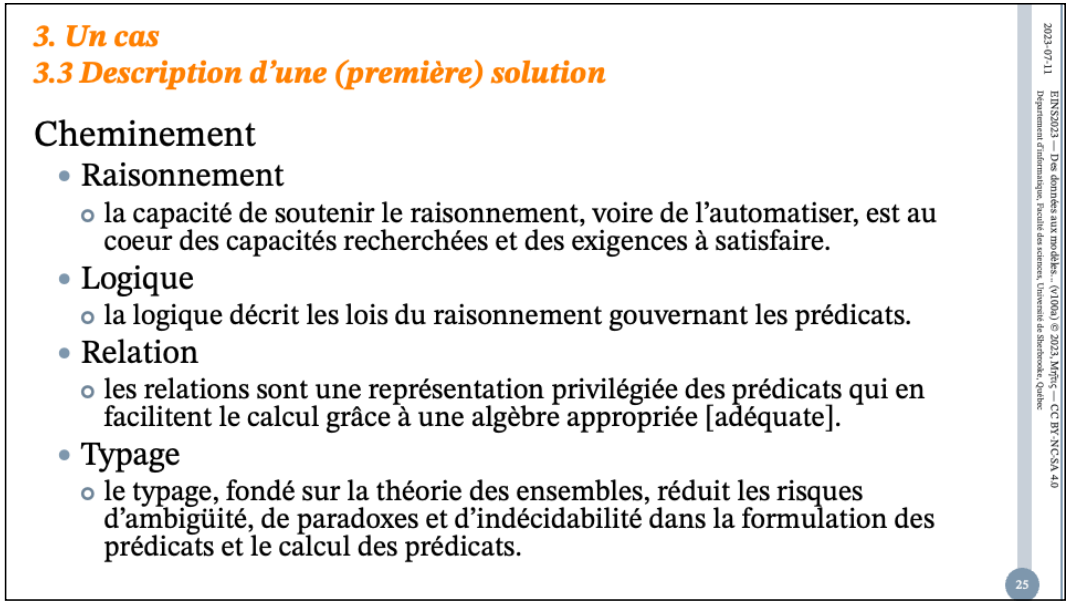


Le cas du méta-modèle.

Au fin de l'exercice, posons

* pour l'épistémologie, celle de Karl Popper avec la réfutabilité discrète
  (de préférence binaire) ;
* pour l'éthique, celle de Karl Popper avec le principe de la nécessaire protection
  de la liberté par l'État… protection qui passe par sa limitation nécessaire
  et suffisante !

Nous y reviendrons après avoir décrit le méta-modèle, au moment de l'évaluer.



Ce cheminement et la théorie relationnelle qui en découle sont le fruit de
travaux que Edgard F. Codd entrepris au cours des années 1960 et mena jusqu'au début
des années 1990. La publication séminale date de 1969.

Typage

- valeur, représentation, type
- type de base et sous-type
- type scalaire et type non scalaire
- type prédéfini et constructeur de type

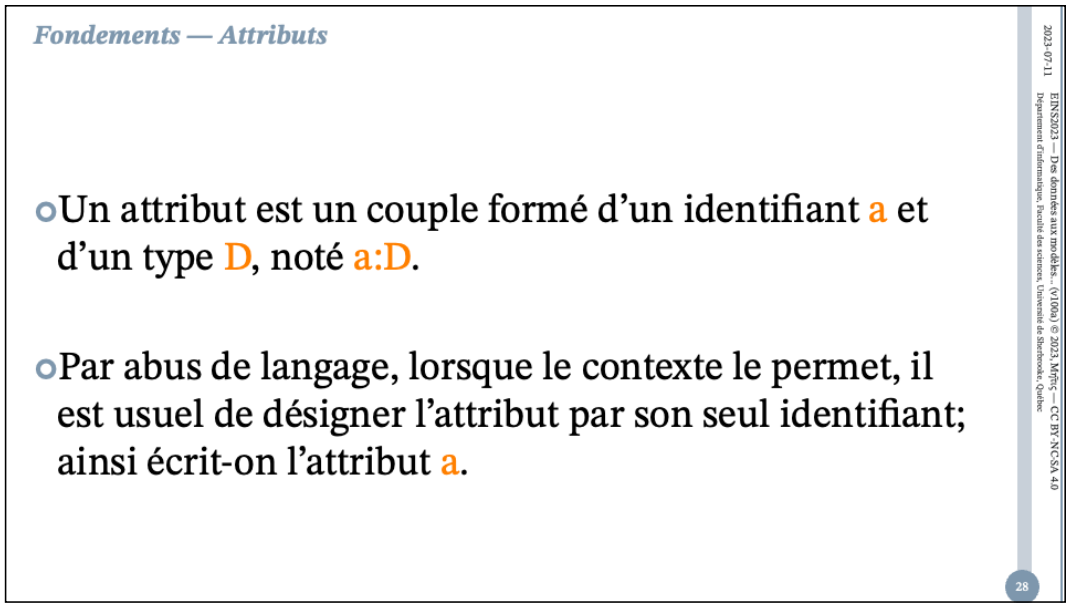
Modèle relationnel

Concepts

- valeur
- variable
- fonction
- état
- procédure
- automatisme

Structure

- attribut
- tuple
- relation
- base



Un tuple correspond à une observation, un «fait».

Un attribut correspond à une dimension de l'observation, à un «aspect» d'un fait.

La relation est un ensemble d'observations de même type, de même «nature».

Deux tuples sont de même type si leurs entêtes sont les mêmes.

Transposé dans le domaine de la logique

- un tuple est une proposition (un énoncé vrai sur le monde);
- une relation est un prédicat (...).

Un prédicat peut être défini

- par énumération (l'ensemble de tous les énoncés vrais et eux seuls);
- par compréhension (la caractérisation nécessaire et suffisante des relations entre les variables).

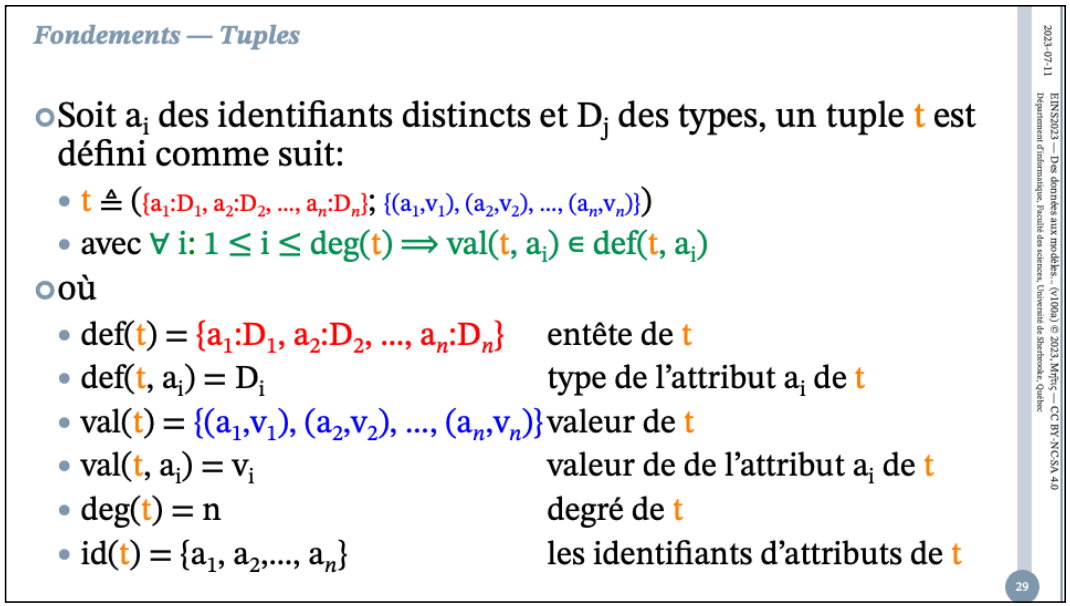
Une base de données est un ensemble de variable de relation définies par leurs valeurs (donc par énumération) et leurs contraintes (donc par compréhension).

En conclusion, une base de données est la représentation d'un système logique.

Dans la pratique, il est souvent difficile d'établir un ensemble de contraintes nécessaires et suffisantes.

Mais on tente de s'en approcher le plus possible.

Pour une théorie des types plus complète, voir (entre autres) IFT 232, IFT 339 et IGE 487.



Un tuple correspond à une observation, un «fait».

Un attribut correspond à une dimension de l'observation, à un «aspect» d'un fait.

La relation est un ensemble d'observations de même type, de même «nature».

Deux tuples sont de même type si leurs entêtes sont les mêmes.

Transposé dans le domaine de la logique

- un tuple est une proposition (un énoncé vrai sur le monde);
- une relation est un prédicat (...).

Un prédicat peut être défini

- par énumération (l'ensemble de tous les énoncés vrais et eux seuls);
- par compréhension (la caractérisation nécessaire et suffisante des relations entre les variables).

Une base de données est un ensemble de relations définies par leurs valeurs (donc par énumération) et leurs contraintes (donc par compréhension). En conclusion, une base de données est la représentation d'un système logique.

Dans la pratique, il est souvent difficile d'établir un ensemble de contraintes nécessaires et suffisantes.

Le cas deg(t) = 0, est important.
Il existe un seul tuple possible (pourquoi?):

- t0 = ({}, {})

Notation simplifiée fondée sur l'ordre d'énumération des attributs
(les identifiants d'attributs et leurs types étant déterminés par ailleurs):

- $t = \langle v_1, v_2, \ldots, v_n \rangle$

Finalement, la notation $t.a_i$ désigne l'attribut $a_i:D_i$ dans def(t).

Fondements — Relations

- Soit a_i des identifiants distincts, D_j des types et t_k des tuples, une relation R est définie comme suit:
  - $R \triangleq (\{a_1{:}D_1, a_2{:}D_2, \ldots, a_n{:}D_n\}; \{t_1, t_2, \ldots, t_m\})$
  - avec $\forall\, i: 1 \leq i \leq card(R) \Longrightarrow def(R){=}def(t_i)$
- Où
  - $def(R) = \{a_1{:}D_1, a_2{:}D_2, \ldots, a_n{:}D_n\}$ entête de R
  - $def(R, a_i) = D_i$ type de a_i de R
  - $val(R) = \{t_1, t_2, \ldots, t_m\}$ valeur de R
  - $deg(R) = n$ degré de R
  - $card(R) = m$ cardinalité de R
  - $id(R) = \{a_1, a_2, \ldots, a_n\}$ identifiants d'attributs de R

Un prédicat peut être défini

- par énumération (l'ensemble de tous les énoncés vrais et eux seuls);
- par compréhension (la caractérisation nécessaire et suffisante des relations entre les variables).

Le cas deg(R) = 0 est important.

Il existe deux relations possibles (pourquoi?):

- R0 = ({}, {})
- R1 = ({}, {t0})

et elles sont très importantes, comme le zéro et le un pour les entiers!

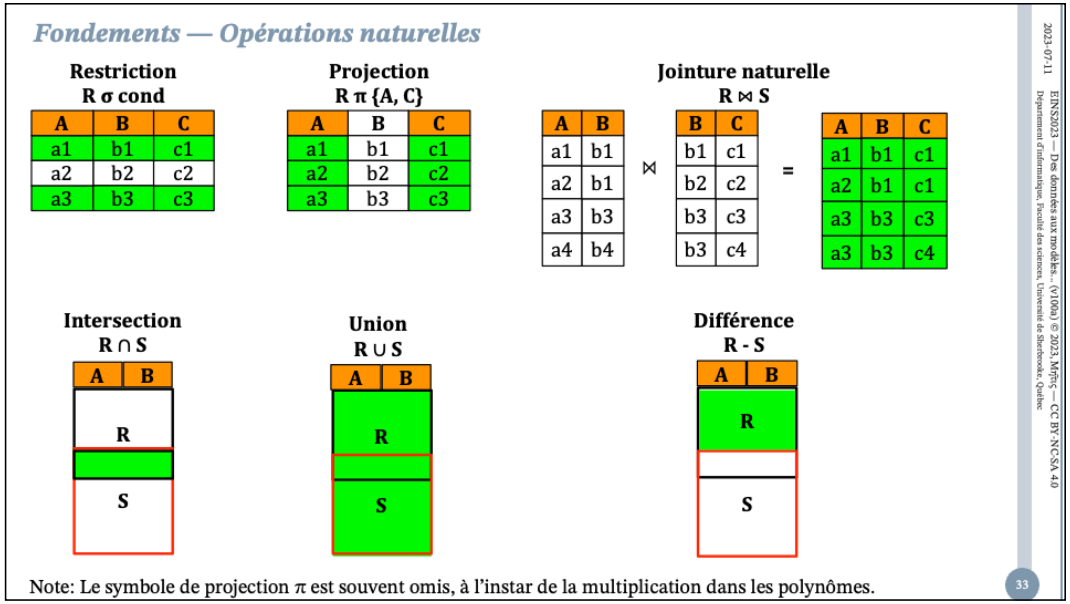
Finalement, la notation $R.a_i$ désigne l'attribut $a_i{:}D_i$ dans def(R).

Fondements — Base «de données»

- Soit v_i des identifiants distincts, D_j des types de relation et r_k des (valeurs de) relations, une base (de données) B est définie comme suit:
  - $B \triangleq (\{v_1{:}D_1, v_2{:}D_2, \ldots, v_n{:}D_n\}; \{r_1, r_2, \ldots, r_m\})$
  - avec $\forall\, i: 1 \leq i \leq card(B) \Longrightarrow def(B, v_i) = def(r_i)$
- Où
  - $def(B) = \{v_1{:}D_1, v_2{:}D_2, \ldots, v_n{:}D_n\}$ entête de B
  - $def(B, a_i) = D_i$ type de a_i de B
  - $val(B) = \{r_1, r_2, \ldots, r_m\}$ valeur de B
  - $deg(B) = n$ degré de B
  - $id(B) = \{v_1, v_2, \ldots, v_n\}$ ensemble des identifiants de variables de relation de B

L'algèbre relationnelle

- union, intersection, différence
- renommage
- projection, restriction, jointure



Opérations relationnels propres: 3

Opérations ensemblistes: 3 (avec le produit cartésien 4)

Opérations structurelles: 1 (renommage)

IGE 487

Le produit est-il vraiment nécessaire?

Et le renommage?

Quel est l'ensemble de base minimal (nécessaire et suffisant)?

Cet ensemble est-il unique?

Voir

DATE, C. J.; DARWEN, H.
Databases, types, and the relational model: The third manifesto.
3rd ed., Addison-Wesley Inc., 2008.
ISBN 0-321-39942-0

Fondements — Opération de renommage

- La présence de l'entête dans chacun des tuples et chacune des relations permet de définir une opération structurelle, le renommage.
- L'entête d'une relation est conservé dans le catalogue du SGBDR.
- Le catalogue est la description des modèles relationnels du SGBDR sous la forme d'une BD dont le modèle relationnel est lui-même dans le catalogue, comme tous les autres modèles relationnels de toutes les autres BD du SGBDR.

Renommage
R ρ A:C

A	B
a1	b1
a2	b2
a3	b3

ρ A:C =

C	B
a1	b1
a2	b2
a3	b3

Pourquoi, et surtout comment, les opérateurs structurels
(aussi appelés meta-opérateurs, tels que le renommage)
peuvent-ils être considérés redondants?

Astuce: prendre le catalogue en considération.

Esquisser un modèle relationnel rudimentaire du catalogue et illustrer votre réponse.

3. Un cas
3.4 Vérification de la solution

La théorie relationnelle est une solution

- acceptable?
- adéquate?

Acceptabilité (satisfaction des exigences)

- La démonstration de la satisfaction des quatre exigences spécifiées plus tôt dépasse la portée et le temps imparti.

- Les écrits de nombreux scientifiques (et parmi les plus grands) le démontrent; écrits qui furent revus, publiés et rendus accessibles.

Adéquation

Critères absolus
- Validité: oui
- Efficacité: oui
- Cohérence: oui

Critères relatifs
- Complétude: suffisante (complétude au sens de Turing démontrée)
- Efficience: la plus efficiente connue parmi les solutions générales
- Évolutivité: pas meilleure que les autres!

Méta-critères
- Réfutabilité...
- Acceptabilité (éthique)...
- À vous de voir, en fonction de l'épistémologie et de l'éthique choisie!

3. Un cas
3.4 Une première expérimentation

- Vive l'APP!

3. Un cas
3.4 Les premières conclusions

- En fait, il s'agit plutôt des conclusions anticipées.
- Autrement dit, les hypothèses à vérifier lors de l'expérimentation.
- Hypothèses découlant de l'analyse du méta-modèle et de celle de la littérature scientifique contemporaine.

Acquis et défis

- Pourquoi présenter acquis et défis conjointement?
- Parce que de nombreux acquis théoriques tardent encore à être mis à disposition, sinon utilisés, en pratique!

Acquis et défis (lot 1)

- Ambigüité:
  - Une certaine amélioration grâce à l'algèbre relationnelle et aux raisonneurs, mais le problème demeure en théorie (lire notre ami Gödel) et en pratique (complexité algorithmique des algorithmes de raisonnement).
- Complétude:
  - La finitude résout théoriquement le problème de complétude... mais persiste toujours en pratique (voir complexité algorithmique et volume de données).
- Données manquantes:
  - Plusieurs solutions ont été proposées dont deux seulement sont solides (Codd:logique_4V et Date:décomosition); malheureusement, aucun langage utilisé couramment ne permet de les utiliser commodément.

Acquis et défis (lot 2)

- Agents:
  - Problème résolu en théorie — en pratique: complexité et expressivité.
- * Axes spatiaux:
  - Problème résolu en théorie — en pratique: outils disponibles, mais la complexité demeure et l'expressivité ainsi que l'efficience demande encore des avancées significatives.
- Axes temporels:
  - Plusieurs approches intéressantes, aucune exhaustive, peu d'outils en pratique... mais qu'est-ce que le temps?
- Cohérence:
  - Que se passe-t-il si on ajoute l'exigence suivante «Capacité de transposer les règles de pratiques médicales» dans le modèle?

Pistes

Utiliser les ontologies appliquées pour décrire les prédicats:

- S'appuyer sur les connaissances.
- Réduire (encore plus) les sources d'ambigüité.

Trois réflexions (en guise de conclusion)

«Du problème à la solution,
en passant par les besoins, les modèles, les exigences et l'adéquation.»

- «Des données aux modèles
en passant par la connaissance et les relations.»

- «Du raisonnement aux relations
en passant par la logique»

Merci à
Pythagore (vers -580 à -495), Socrate (vers -470 à -399),
Platon (vers -428 à -347), Aristote (vers -384 à -322),
Eulcide (vers -300)…
Descartes (1596 à 1650), Pascal (1623 à 1662), Russel (1872 à 1970),
Wittgenstein (1889 à 1951), Gödel (1906 à 1978), Floyd (1936-2001),
Codd (1923 à 2003), Chomsky (1928...), Hoare (1934...) et Date (1941...).