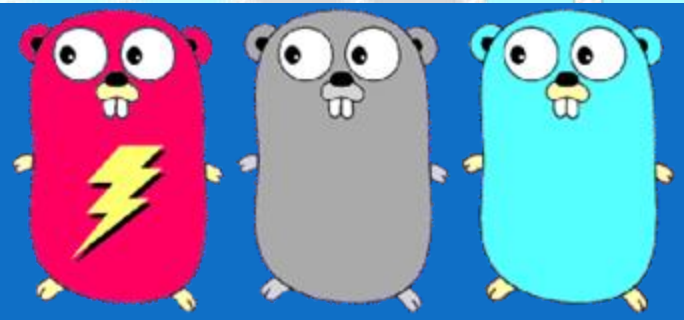
UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

Génie Logiciel pour le Calcul Scientifique



#11

13/02/2025

jean-michel.batto@cea.fr

cea

https://gogs.eldarsoft.com/M2_IHPS

TD3 – intégration

- TD3 noté pour le 21/02/2024
- Dans https://gogs.eldarsoft.com/M2_IHPS/GLCS-CM6-TDXMP
- Nous avons vu le code XMP pour calculer un histogramme à partir d'un fichier txt.
- Le code est « pauvre », il ne donne pas un % par valeur
- Fichier input.txt // formaté avec 10 valeurs par ligne, sauf première ligne

```
100 // le nombre de valeurs ➔ pour allouer la mémoire
1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0,9.0,10.0
2.0,0.0,4.0,5.0,6.0,7.0,8.0,9.0,10.0,11.0
3.0,4.0,5.0,6.0,7.0,8.0,9.0,10.0,11.0,12.0
4.0,5.0,6.0,7.0,8.0,9.0,10.0,11.0,12.0,13.0
```

TD3

- Le problème : pour faire du benchmark, il faut un gros fichier.
- Autre problème : comment permettre un travail individualisé ?

➔ avoir un générateur de fichier sur-mesure.

skewseed : An 8-digit number seed to skew the distribution of integer values.

outputlinecount : Number of lines in the output data set

randomhighint : Highest integer value for random data set

valueperline : Number of values per line

digitcountskew : The precision of the skew is determined by the number of digits (>10 and <300)

➔vous allez utiliser skewseed et y mettre votre matricule ➔ change la distribution

➔outputlinecount est le paramètre à faire varier (20 millions est une limite du code histo.c dans son implémentation actuelle)

Vous ne changez pas les autres valeurs (éventuellement digitcountskew pour avoir un meilleur score d'histogramme)

TD3 : ce qu'on veut évaluer

- Dans un mode conteneurisé, l'espace d'allocation des ressources est finit.
- On peut « dissiper » de la puissance CPU par des optimisations inutiles.
- Par exemple, utiliser 4 nœuds MPI plutôt que 1.
- Par exemple, ne pas paralléliser (impact d'XMP, il y aussi ACC)

- Le code https://gogs.eldarsoft.com/M2_IHPS/GLCS-CM6-TDXMP est une base de travail. Il ne donne pas directement un histogramme de fréquences.

TD3

- Votre travail :
  - Modifier le code histo.c pour y ajouter un chronomètre (attention au link). On veut pouvoir apprécier le temps de traitement de manière précise. Le code histo.c doit vous permettre de calculer un histogramme et de produire une figure pour votre rapport. Vous pouvez utiliser Excel ou un code à vous. Ne passez pas trop de temps sur cette partie.
  - Tester différentes configurations (1 nœud, 2 nœuds, 4 nœuds)
  - Vous allez choisir une configuration (2 nœuds, fichier 10 Millions) qui sera le gold standard. Lorsque vous changez de configuration, vous exprimez le gain **en %** par rapport au gold standard.
  - Tester différentes optimisations du code avec une appréciation en fonction de la taille du fichier d'entrée (de combien il y a un impact).
  - Me transmettre l'histogramme (il s'agit d'une figure avec la valeur des % d'occurrences, % écrit dans un tableau avec 4 décimales). Attention l'histogramme est dépendant du code matricule utilisé.

TD3 – notation ➔ pour la semaine prochaine

Quoi	Intérêt	Impact
Structure de la démarche bien présentée	On comprend le projet. SVP pas de blabla d'IA.	10%
Définition du problème, du concept de partage de la ressource	Les axes d'exploration sont présentés (effet abstraction réseau, coeurs, cache, saturation).	20%
Capacité à produire du code et à modifier le code pour illustrer la démarche	Le code remis doit être commenté. Il permet de vérifier la qualité du travail.	10%
Capacité à mettre en œuvre une IA	On voit l'utilisation d'un ou de plusieurs prompts.	10%
Diagrammes de benchmark	On s'attend à voir si une optimisation est sensible ou non à la taille du fichier d'entrée. Attention aux échelles.	20%
Dépassement du sujet	Capacité à reprendre le problème et à proposer un nouvel angle.	20%
Conclusion synthétique	Les points saillants sont exposés. SVP pas de blabla d'IA.	10%

Dans la forge : votre rapport PDF sans matricule (je l'ai déjà), votre code source avec un README. Pas de fichiers input.txt ni output.txt. Pas de gros fichiers.

Plan du cours #Anonymisation+RGPD

- Introduction aux contraintes réglementaires sur les données personnelles (=survol)
- TD étude d'article : anonymisation / re-identification +K-anonymat
- Présentation d'un algorithme d'anonymisation
- RGPD

Le traitement des données personnelles (avec modification RGPD)

➔principe de l'homme de l'art : responsabilité

- La loi « Informatique et Libertés » 6 janvier 1978
- Modifié par Ordonnance n°2018-1125 du 12 décembre 2018 - art. 1

- Article 1
- **L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.**

Les droits des personnes de décider et de contrôler les usages qui sont faits des données à caractère personnel les concernant et les obligations incombant aux personnes qui traitent ces données s'exercent dans le cadre du règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, de la directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 et de la présente loi.

Le traitement des données personnelles

- La notion de donnée personnelle dans la loi « Informatique et Libertés »
- Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.
- Nom, prénom, adresse postale, e-mail, numéro de téléphone, adresse IP, adresse MAC, plaque d'immatriculation, numéro de sécurité sociale, ….

Le traitement des données personnelles

La notion de donnée sensible

- Les données qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicales des personnes.

Ou

- Les données qui sont relatives à la santé ou à la vie sexuelle de celles-ci

Le traitement des données personnelles

Les champs d'application de la loi « Informatique et Libertés »

- Traitement de données à caractère personnel
- Traitement automatisée des données (nominatives)
- Fichier de données à caractère personnel

Le traitement des données personnelles

- Les données de santé
- Définition par la jurisprudence.
  - Aspects physiques / psychiques de la santé d'une personne (par ex : handicap, fumeur, …)
  - Des informations sur la consommation de drogues, de médicaments
  - Des informations sur la prestation de service de santé à cette personne

Le secret médical

- Les données relatives à la santé ne sont accessibles qu'à des médecins
- Une personne non médecin ne peut accéder qu'à des données préalablement **anonymisées**
- Une distinction et une séparation des données :
  - Les données d'identification
  - Les données de résultats

Introduction aux techniques d'anonymisation

Les organisation possèdent des données personnelles.

-Dans toutes les techniques d'anonymisation, il faut répondre aux questions suivantes :

- i) Est-il toujours possible d'isoler un individu?
- ii) Est-il toujours possible de relier entre eux les enregistrements relatifs à un individu?
- iii) Peut-on déduire des informations concernant un individu?

Introduction aux techniques d'anonymisation

Identifiants

❖Un attribut (ou ensemble d'attributs) qui identifie un enregistrement de façon unique, comme le nom, l'adresse mail.

Quasi-identifiants (QID)

❖ Ce sont les attributs qui ne sont pas eux-mêmes des identifiants uniques, mais peuvent devenir identifiés lorsqu'ils sont combinés avec d'autres ensembles (par l'âge, le sexe ou l'adresse.)

Name	ZipCode	Gender	Age	Disease
Jean	75275	Male	22	Flu
Alex	75278	Male	24	HIV+
Michel	75275	Male	27	Diabetes
Tony	75275	Male	42	HIV+
Sara	75278	Female	25	HIV+
Sandy	75278	Female	29	Diabetes
Lara	75277	Female	23	Cancer

Identifiant **Quasi-identifiants** **Attribut sensible**

Attributs sensibles

❖ Ce sont les attributs que l'on veut conserver exemple le salaire, ou un état de maladie.

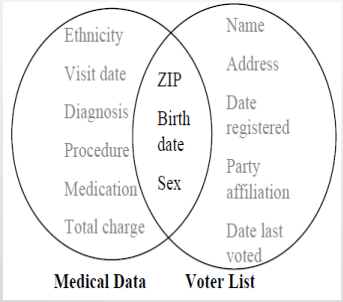
Introduction aux techniques d'anonymisation

1. Pseudonymisation

- Une clé d'identification qui permet d'établir le lien entre les différentes informations des personnes.
- Ces clés d'identification doivent être stockées de manière sécurisée.
- Réalisation très simple avec un algorithme naïf de cryptage ou hachage. (non réversible)
- La combinaison entre l'ensemble des attributs dans deux bases de données peut permettre de retrouver l'individu concerné (exemple de Sweeny).

Sweeny

Etats-Unis en 2001
Le croisement sur une base des données médicale pseudonymisée et une liste électorale avec des données nominatives.



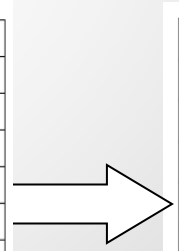
Relier les données

Introduction aux techniques d'anonymisation

2. **Randomisation**
   - **Ajout de bruit** : modifier les valeurs des attributs dans l'ensemble de données pour les rendre moins précis.
   - **Permutation** : mélanger les valeurs des attributs de telle sorte qu'on conserve la distribution exacte de chaque attribut dans l'ensemble de données.
3. **Généralisation // agrégat autour d'une nouvelle classe d'équivalence**
   - **K-anonymat**

Name	ZipCode	Gender	Age	Disease
Jean	75275	Male	22	Flu
Alex	75278	Male	24	HIV+
Michel	75275	Male	27	Diabetes
Tony	75275	Male	42	HIV+
Sara	75278	Female	25	HIV+
Sandy	75278	Female	29	Diabetes
Lara	75277	Female	23	Cancer

Identifiant | Quasi-identifiants | Attribut sensible



Name	ZipCode	Gender	Age	Disease
X	7527*	Male	[20-29]	Flu
X	7527*	Male	[20-29]	HIV+
X	7527*	Male	[20-29]	Diabetes
X	7527*	Male	[40-49]	HIV+
X	7527*	Female	[20-29]	HIV+
X	7527*	Female	[20-29]	Diabetes
X	7527*	Female	[20-29]	Cancer

Un exemple d'un tableau de 3-anonymes

Qualités de l'anonymisation

1. **Le taux de suppression**
   - La quantité de suppression de toutes **les classes d'équivalence** qui sont plus petites que K, doit être inférieure à une limite définie par l'utilisateur.

$$\text{Le taux de Suppression} = \frac{\text{Le nombre des enregistrements supprimés}}{\text{Le nombre total des enregistrements.}}$$

2. **La perte d'information // la disparition des « outliers »**

3. **La protection // la capacité à retrouver un individu ou à relier un groupe ➔ pb des petites valeurs**

TD : Anonymisation

Etude de paper1_Sweeney L. Guaranteeing anonymity when sharing medical data The Datafly

- https://dataprivacylab.org/datafly/index4.html
  - http://dataprivacylab.org/datafly/paper4.pdf
  - L. Sweeney Guaranteeing anonymity when sharing medical data, the Datafly system. *Proceedings, Journal of the American Medical Informatics Association.* (AMIA). Washington, DC: Hanley & Belfus, Inc., 1997.

- Mot anglais : Bin = corbeille
- 1/Quels sont les exemples de de-anonymisation cités dans l'article
- 2/Quelle est la taille des paquets d'information minimale pour des données de santé selon la recommandation citée dans l'article?
- 3/Quelle est l'approche scientifique de l'article ?

Re-identification

Netflix Prize data

Dataset from Netflix's competition to improve their recommendation algorithm

Netflix held the Netflix Prize open competition for the best algorithm to predict user ratings for films. The grand prize was $1 000 000.

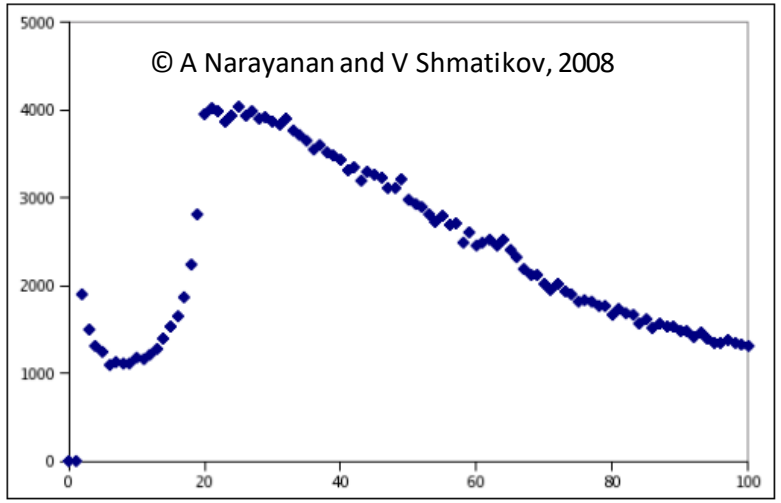


Figure 2. For each $X \leq 100$, the number of subscribers with X ratings in the released dataset.

Film :
Les Dames du Bois de Boulogne (1945)

Id netflix	score	date
2532865	4	2005-07-26
573364	3	2005-06-20
1696725	3	2004-02-27
1253431	3	2004-03-31
1265574	2	2003-09-01
1049643	1	2003-11-15
1601348	4	2005-04-05
1495289	5	2005-07-09
1254903	3	2003-09-02
2604070	3	2005-05-15
1006473	5	2005-05-23
1989892	3	2004-04-06
1517471	4	2003-12-24
1478381	4	2005-05-21
923084	2	2004-11-15
2446292	4	2005-10-06
2554745	3	2003-05-07
1133125	5	2004-08-10
349528	4	2003-08-11
1614895	5	2004-08-29
424958	4	2005-08-02
1390877	3	2003-10-07
2327422	4	2004-08-19
18103	3	2005-07-24
591075	4	2004-03-12

TD : Re-identification

Etude de paper2_shmat_oak08netflix

Narayanan, V. Shmatikov. Robust De-anonymization of Large Sparse Datasets, or How to Break Anonymity of the Netflix Prize Dataset. S&P (Oakland) 2008.

https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

1/Quel est l'enjeu de l'article ?

2/En 2005, que voulait faire Netflix ?

2/Page 3, quelle est l'observation qui permet l'approche de l'attaque ?

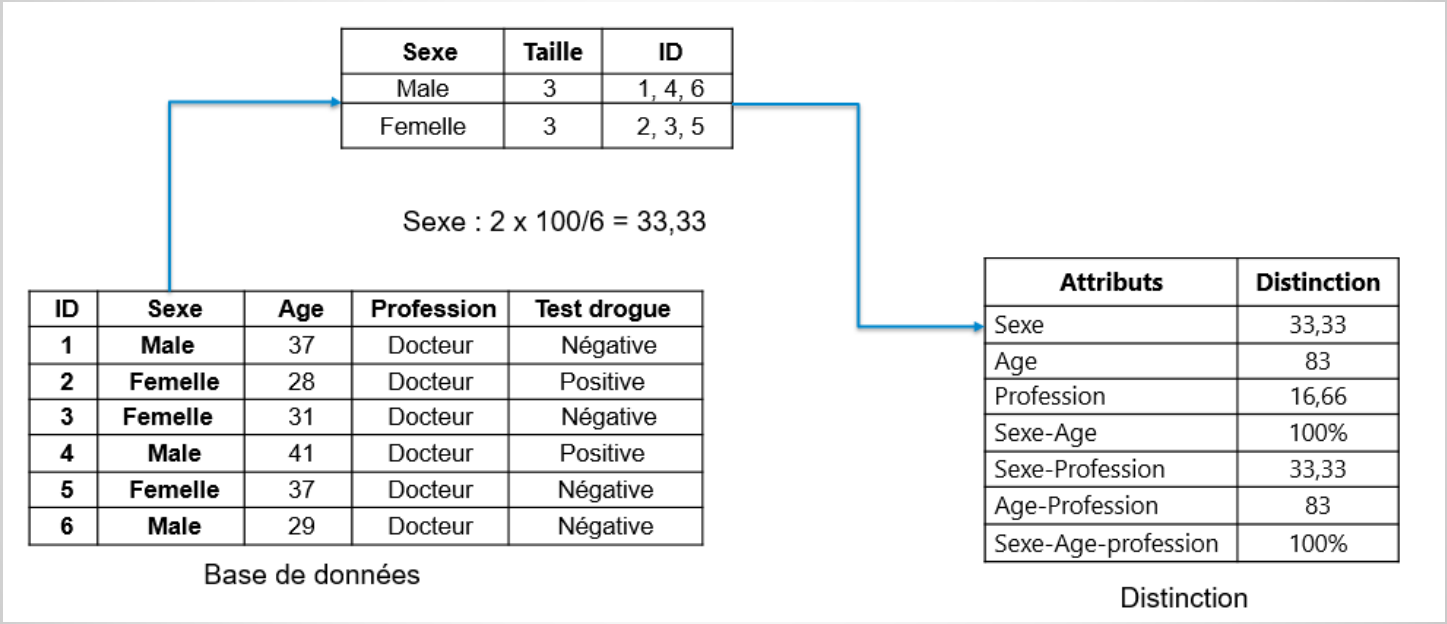
3/Page 8, pourquoi protéger son anonymat ?

4/Page 11, quels sont les enseignements de cet article sur l'anonymat des scores ?

La distinction

- Distinction

$$Distinction = \frac{Nombre\ de\ classes\ d'équivalences * 100}{Nombre\ total\ d'enregistrements}$$



Sexe	Taille	ID
Male	3	1, 4, 6
Femelle	3	2, 3, 5

ID	Sexe	Age	Profession	Test drogue
1	Male	37	Docteur	Négative
2	Femelle	28	Docteur	Positive
3	Femelle	31	Docteur	Négative
4	Male	41	Docteur	Positive
5	Femelle	37	Docteur	Négative
6	Male	29	Docteur	Négative

Base de données

Attributs	Distinction
Sexe	33,33
Age	83
Profession	16,66
Sexe-Age	100%
Sexe-Profession	33,33
Age-Profession	83
Sexe-Age-profession	100%

Distinction

- age-profession : 5 classes / 6 enregistrements = 83,3%
- profession : 1 classe / 6 enregistrements = 16,6%

Les différentes re-identifications

- Approche systémique des attaques
- L'attaque du procureur
  - On connait un individu est on veut extraire une information confidentielle
  - On veut re-identifier l'individu

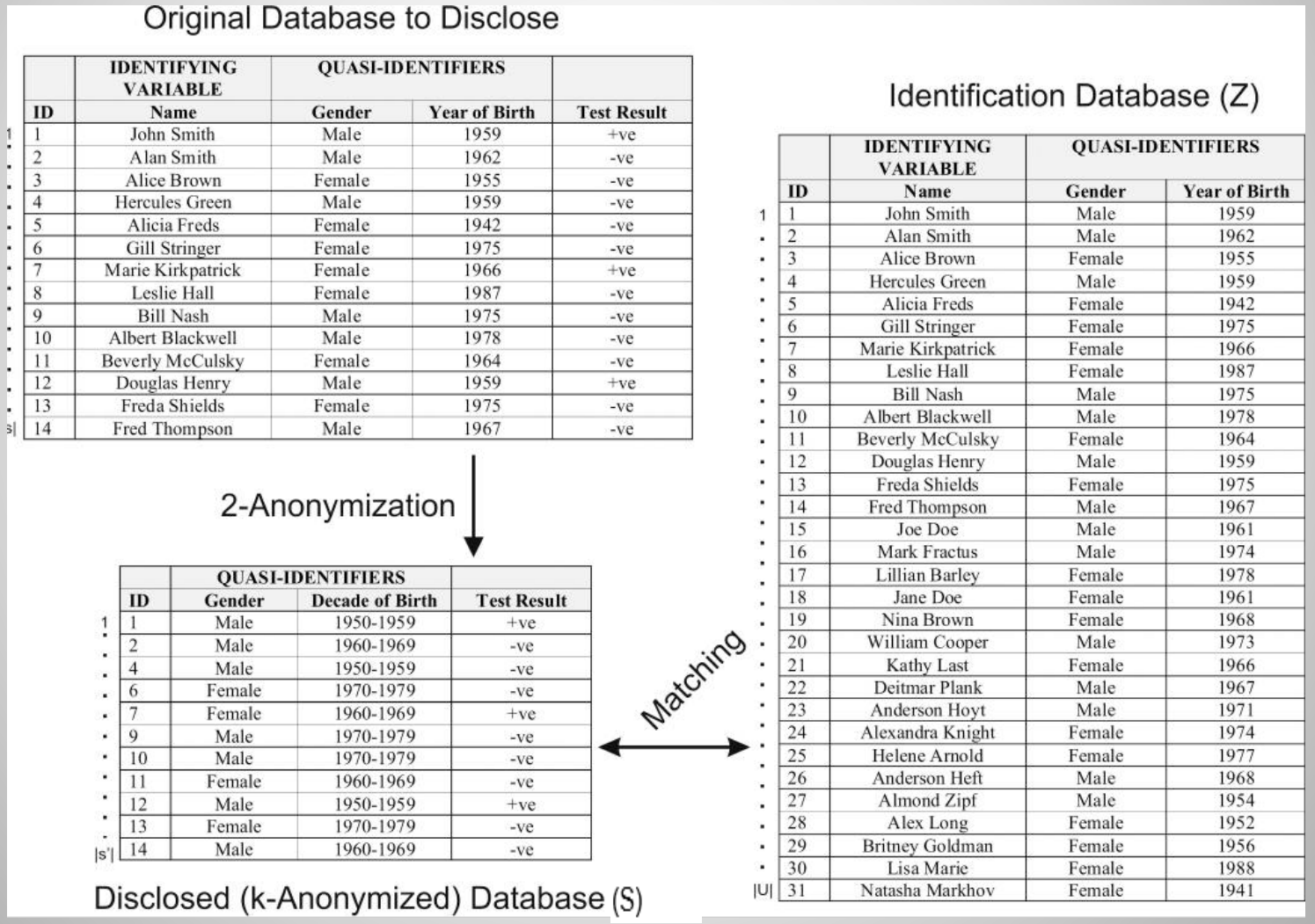
	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Equivalence class		Anonymized table	
Gender	Age	Count	Id
Male	1950-1959	3	1,4,12
Male	1960-1969	2	2,14
Male	1970-1979	2	9,10
Female	1960-1969	2	7,11
Female	1970-1979	2	6,13

- Par exemple la personne est née en 1955, 33% de chance d'être associée à un résultat du test

Les différentes re-identifications

- L'attaque du journaliste :
  - Par rapport à une liste connue, on cherche des individus



Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

2-Anonymization

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Disclosed (k-Anonymized) Database (S)

Matching

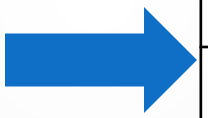
Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Attaque du journaliste On anonymise

Etape 1: Généraliser la base données d'identification

IDENTIFYING VARIABLE		QUASI-IDENTIFIERS	
ID	**Name**	**Gender**	**Year of Birth**
1	John Smith	Male	**1959**
2	Alin Smith	Male	**1962**
3	Alic Brown	Female	**1955**
4	Hercules Green	Male	**1959**
5	Alicia Freds	Female	**1942**
:			
: :			
30	Lisa Marie	Female	**1988**
31	Natasha Markhov	Female	**1941**



IDENTIFYING VARIABLE		QUASI-IDENTIFIERS	
ID	**Name**	**Gender**	**Year of Birth**
1	John Smith	Male	**1950-1959**
2	Alin Smith	Male	**1960-1969**
3	Alic Brown	Female	**1950-1959**
4	Hercules Green	Male	**1950-1959**
5	Alicia Freds	Female	**1940-1949**
:			
: :			
30	Lisa Marie	Female	**1980-1989**
31	Natasha Markhov	Female	**1940-1949**

- **Etape 2:** Calculer toutes les classes d'équivalences pour les données anonymisées.
- [1950-1979]

Equivalence Class		Anonymized table	
Gender	**Age**	**Count**	**ID**
Male	1950-1959	3	1, 4, 12
Male	1960-1969	2	2, 14
Male	1970-1979	2	9, 10
Female	1960-1969	2	7, 11
Female	1979-1979	2	6, 13

- **Etape 3:** Calculer toutes les classes d'équivalences pour les données d'identification généralisées (celles en possession du journaliste).

Equivalence Class		Public Database	
Gender	**Age**	**Count**	**ID**
Female	1940-1949	2	5, 31
Male	1950-1959	4	1, 4, 12, 27
Female	1950-1959	3	3, 28, 29
Male	1960-1969	5	2, 14, 15, 22, 26
Male	1970-1979	5	9, 10, 16, 20, 23
Female	1980-1989	2	8, 30
Female	1960-1969	5	7, 11, 18, 19, 21
Female	1970-1979	5	6, 13, 17, 24, 25

Il s'agit de vérifier le risque

Equivalence Class		Anonymized table	
Gender	**Age**	**Count**	**ID**
Male	1950-1959	3	1, 4, 12
Male	1960-1969	2	2, 14
Male	1970-1979	2	9, 10
Female	1960-1969	2	7, 11
Female	1979-1979	2	6, 13

Matching

Equivalence Class		Public Database	
Gender	**Age**	**Count**	**ID**
Female	1940-1949	2	5, 31
Male	1950-1959	4	1, 4, 12, 27
Female	1950-1959	3	3, 28, 29
Male	1960-1969	5	2, 14, 15, 22, 26
Male	1970-1979	5	9, 10, 16, 20, 23
Female	1980-1989	2	8, 30
Female	1960-1969	5	7, 11, 18, 19, 21
Female	1979-1979	5	6, 13, 17, 24, 25

- **Etape 5:** On s'intéresse au plus grand risque de ré-identification.

Equivalence Class		Anonymized table		Public Database	
Gender	**Age**	**Count**	**ID**	**Count**	**ID**
Male	1950-1959	3	1, 4, 12	4	1, 4, 12, 27
Male	1960-1969	2	2, 14	5	2, 14, 15, 22, 26
Male	1970-1979	2	9, 10	5	9, 10, 16, 20, 23
Female	1960-1969	2	7, 11	5	7, 11, 18, 19, 21
Female	1979-1979	2	6, 13	5	6, 13, 17, 24, 25

- Le plus grand risque correspond à la plus petite classe d'équivalence des données publiques.
- Une probabilité **(3/4)** de ré-identifier une personne dans cette classe.

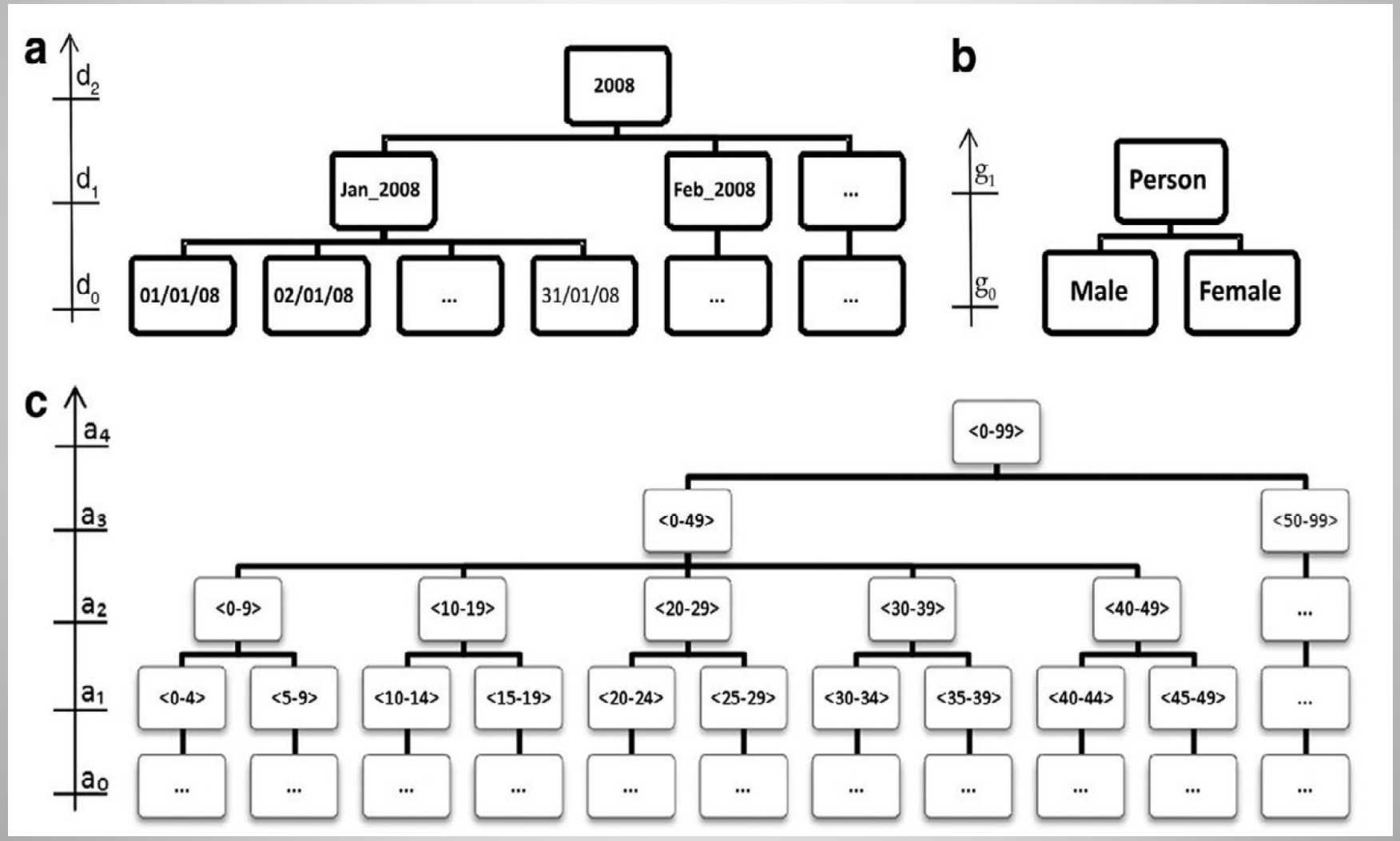
- L'attaque de commercialisation

Equivalence Class		Anonymized table		Public Database		Probability of match
Gender	Age	Count	ID	Count	ID	
Male	1950-1959	3	1, 4, 12	4	1, 4, 12, 27	3/4
Male	1960-1969	2	2, 14	5	2, 14, 15, 22, 26	2/5
Male	1970-1979	2	9, 10	5	9, 10, 16, 20, 23	2/5
Female	1960-1969	2	7, 11	5	7, 11, 18, 19, 21	2/5
Female	1979-1979	2	6, 13	5	6, 13, 17, 24, 25	2/5

- Est une extension de l'attaque du journaliste
- Ici 47% des enregistrements sont « ciblés »

Comment anonymiser?

Construire des classes d'équivalences (date, gender, age)



L-diversité / Utilité

- Supposons qu'on a un attribut (a) avec un domaine Da et Va, le sous-ensemble des valeurs de a. La diversité de d(Va) est calculée par :

$$da(V_a) = \begin{cases} \frac{max(V_a) - min(V_a)}{max(D_a) - min(D_a)} & interval\ values \\ \frac{|distinct(V_a)|}{|D_a|} & discrete\ values \end{cases}$$

- Telle que max(Va), min (Va), max(Da) et min(Da) sont les valeurs maximales et minimale dans Va et Da respectivement, |distinct(Va)| c'est le nombre des valeurs uniques dans Va et IDaI la taille totale du domaine.
- Par conséquent, la diversité d'une classe d'équivalence (gi) qui consiste à plusieurs tuples τ, est calculée par la suite :

$$dt(\tau, A) = \sum_{i=1}^{m} da(\pi_{a_i}(\tau))$$

- Ou |gi| ≥ k et πai(τ) est la diversité d'un attribut a appartient a QID (a1, a2…am)
- Alors la moyenne de la diversité, de toute la classe d'équivalence, représente la perte des données dans un tableau T :
- **Utilité** $= avg(dt(g_1, q), dt(g_2, q), \ldots, dt(g_h, q))$
- Une petite valeur de cette mesure signifie que les tuples des classes sont proches, et donc les données anonymes peuvent être le plus utile.

Protection sur les attributs sensibles

$$protection = avg(\frac{1}{dt(g_1, s)}, \frac{1}{dt(g_2, s)}, \ldots, \frac{1}{dt(g_h, s)})$$

On peut donc avoir 2 contraintes :

- La diversité (utilité des enregistrements)
- La protection

K-anonymat / L-diversité

Application de la k-anonymisation

Nom	Age	Sport pratiqué	Membre
Jules	20	Billard	Ordinaire
Théo	23	Volley	Donateur
Alexis	28	Foot	Donateur
Eleo	30	Relais	Bienfaiteur
Angela	35	Saut en hauteur	Ordinaire
Séraphine	29	Tennis	Donateur
Inès	20	Billard	Emerite
Romain	20	Cricket	Ordinaire
Brigitte	42	Relais	D'honneur

Connaitre l'âge ou le sport, permet d'identifier (=QID)
La variable d'intérêt est le statut (Ordinaire / Donateur / Bienfaiteur / Emerite / D'honneur)
On veut faire de l'anonymisation par agrégat : k = 3

TD en classe

Numéro id	Catégorie	Type de sport	Membre	
1	<22	Non olympique	Ordinaire	€
2	22-29	Balle	Donateur	€€
3	22-29	Balle	Donateur	€€
4	>=30	Athlétisme	Bienfaiteur	€€€
5	>=30	Athlétisme	Ordinaire	€
6	22-29	Balle	Donateur	€€
7	<22	Non olympique	Emerite	€€€€
8	<22	Non olympique	Ordinaire	€
9	>=30	Athlétisme	D'honneur	0

➔**3 classes, avec 3 individus par classe,**

pb il y a une classe qui a perdu la diversité (pas de 3-diversité)

Une solution de K-Anonymat

Numéro id	Catégorie Age	Individuel / Collectif	Membre
1	<25	Individuel	Ordinaire
2	<25	Collectif	Donateur
3	>25	Collectif	Donateur
4	>25	Collectif	Bienfaiteur
5	>25	Individuel	Ordinaire
6	>25	Individuel	Donateur
7	<25	Individuel	Emerite
8	<25	Collectif	Ordinaire
9	>25	Collectif	D'honneur

En découpant autrement, avec 4 classes, on a :

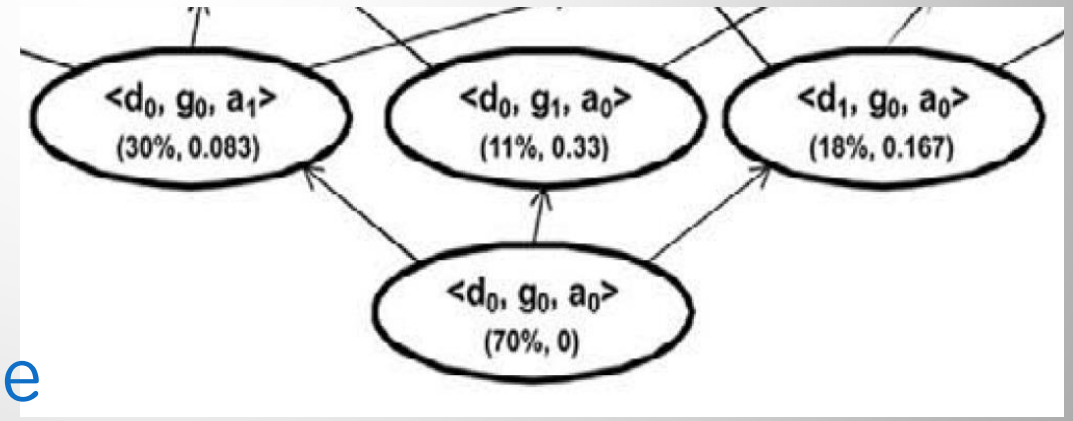
2-anonymité sur l'individu

2-diversité sur la variable d'intérêt

K-anonymisation optimale

- C'est un problème NP-Hard // *on donne une taille K*
- Il s'agit de faire des pavages sur **K** et de trouver un pavage optimal – « généralisation »
- Contraintes d'optimalité : suppression & protection
- **$<d_0,g_0,a_0>$ 70% des enregistrements de la classe sont supprimés**
- **$<d_0,g_0,a_1>$ 30%**
- **$<d_0,g_1,a_0>$ 11%**

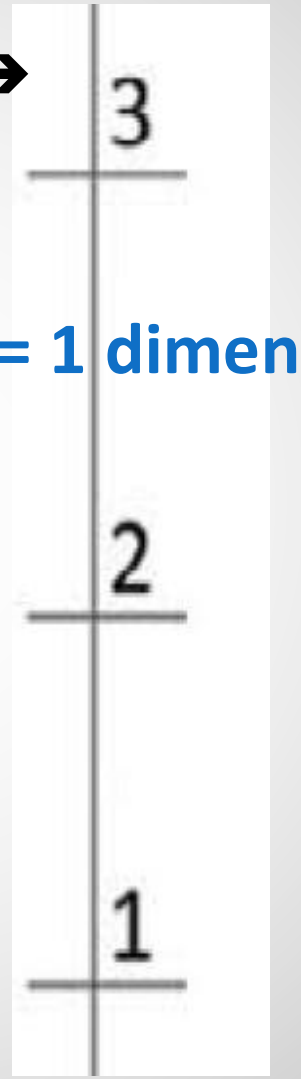
d : date, g : gender, a : age



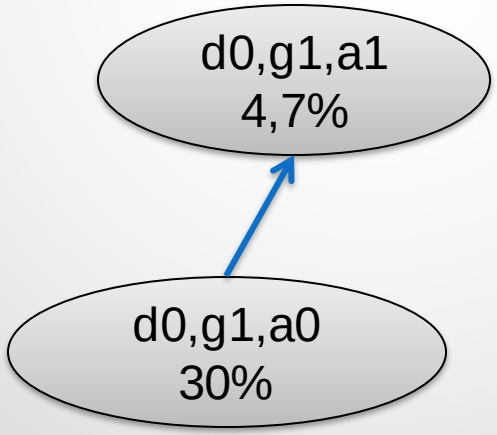
- La protection est calculée pour chaque noeud

K-anonymisation

Nombre de dimensions décrites par le pavage ➔

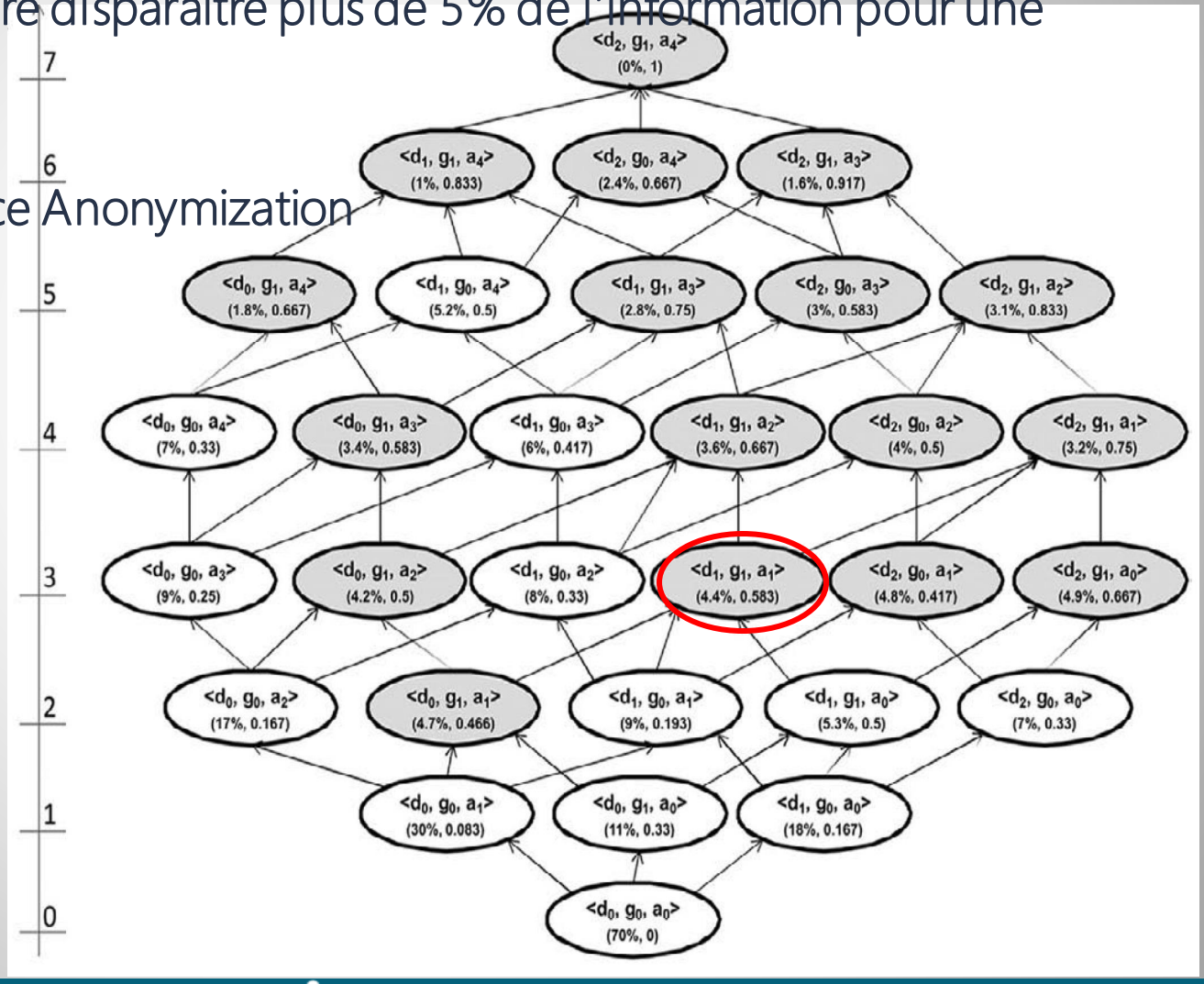


- $<d_0,g_1,a_0>$ 11% ➔ 1 dimension perdue
- Toutes les dates, Personne (perte genre = 1 dimension), âge

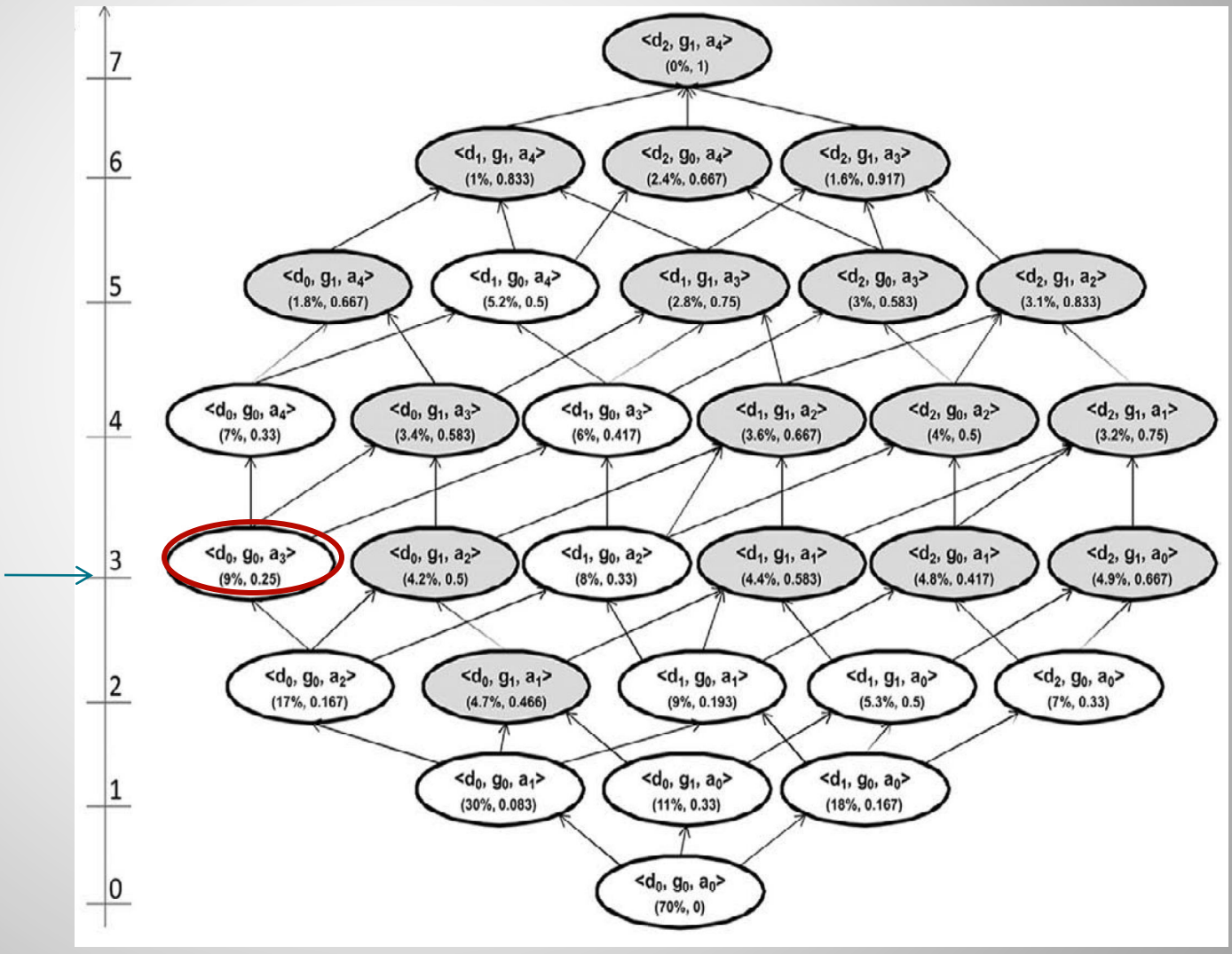


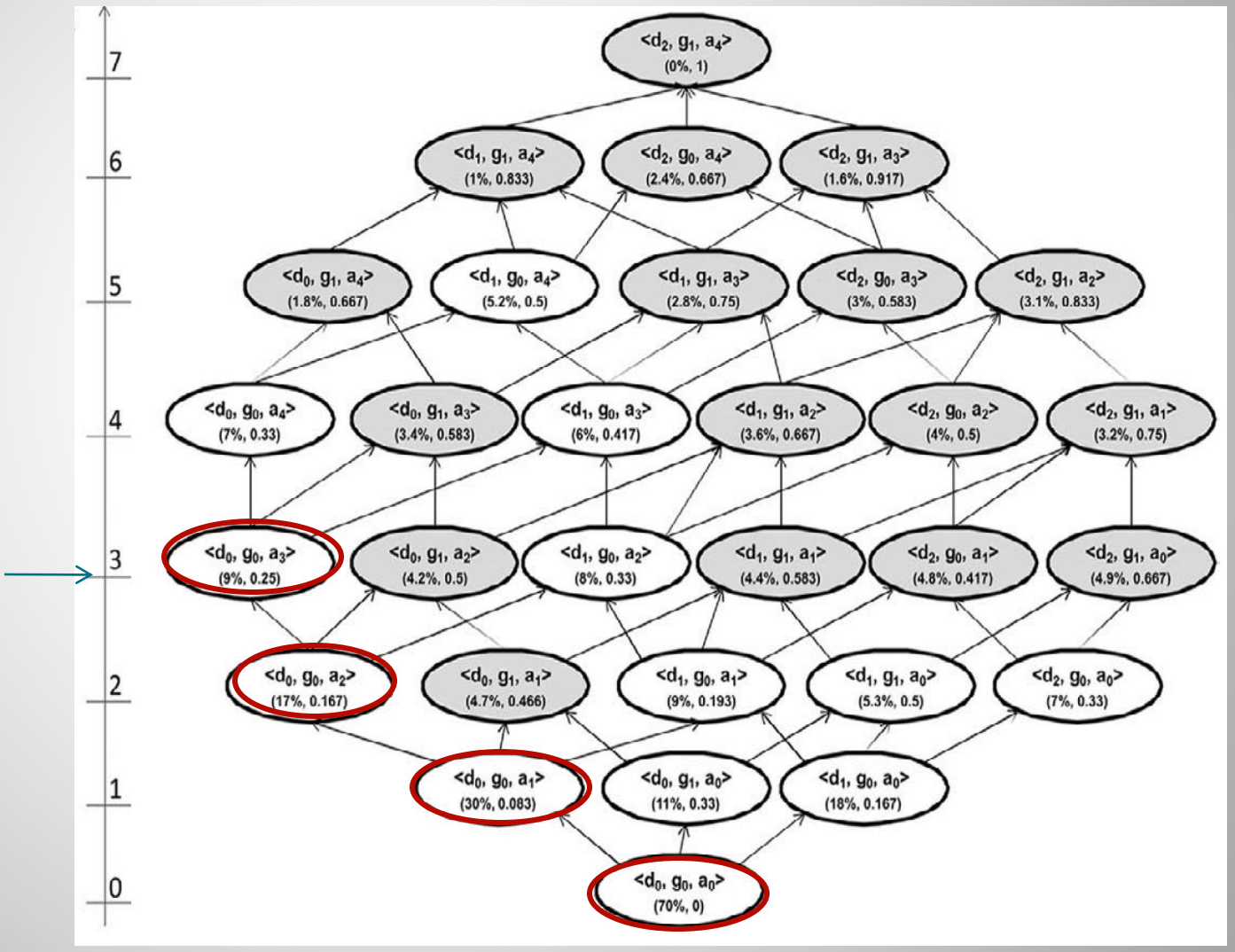
Optimal Lactice Anonymization

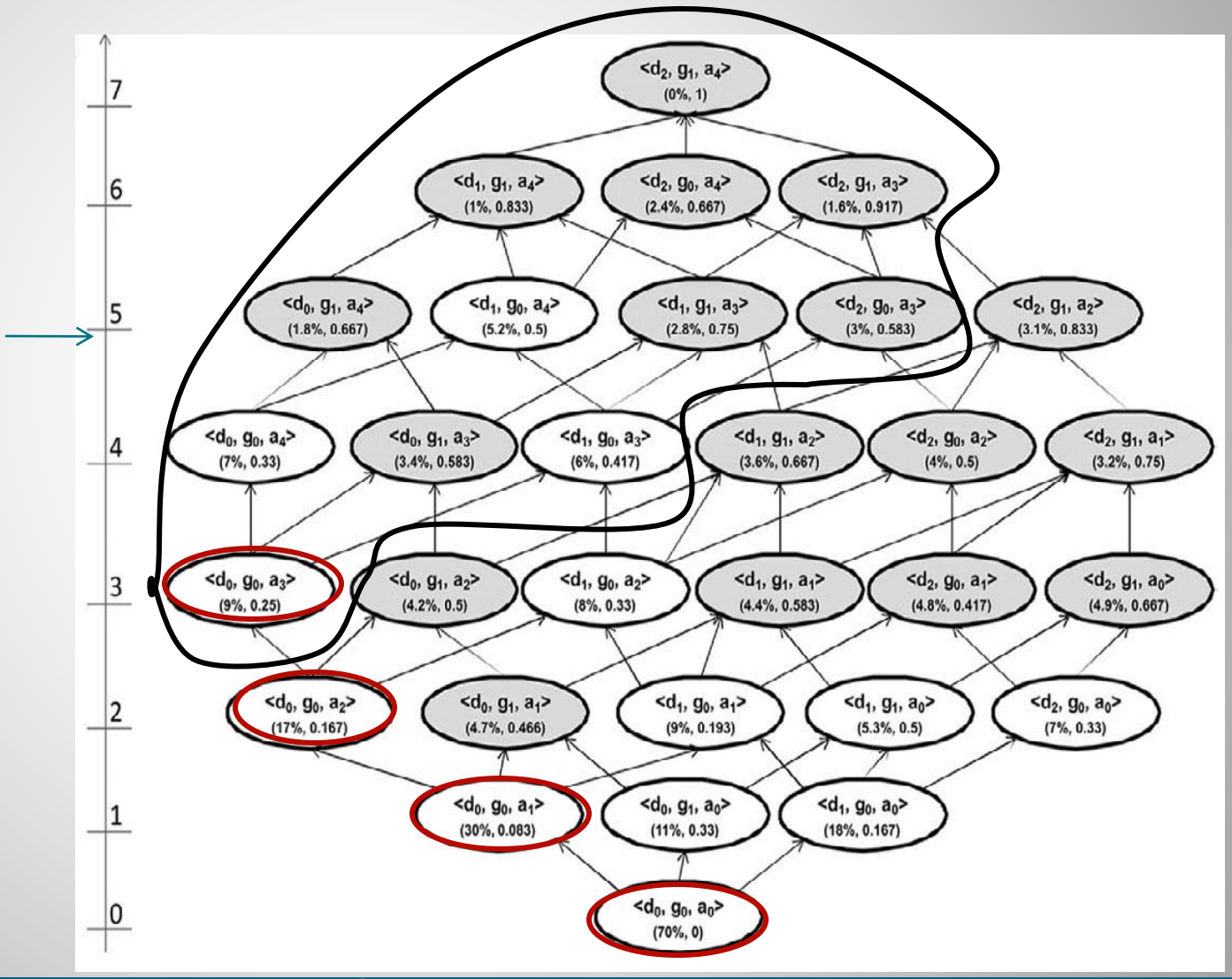
- Limite de Suppression = 5% -
  ratio distinction/protection = 0.6 pour un K donné (K=x).
- ➔on ne veut pas faire disparaitre plus de 5% de l'information pour une classe d'équivalence
- OLA = Optimal Latice Anonymization
  - Recherche binaire
  - Tag prédictif

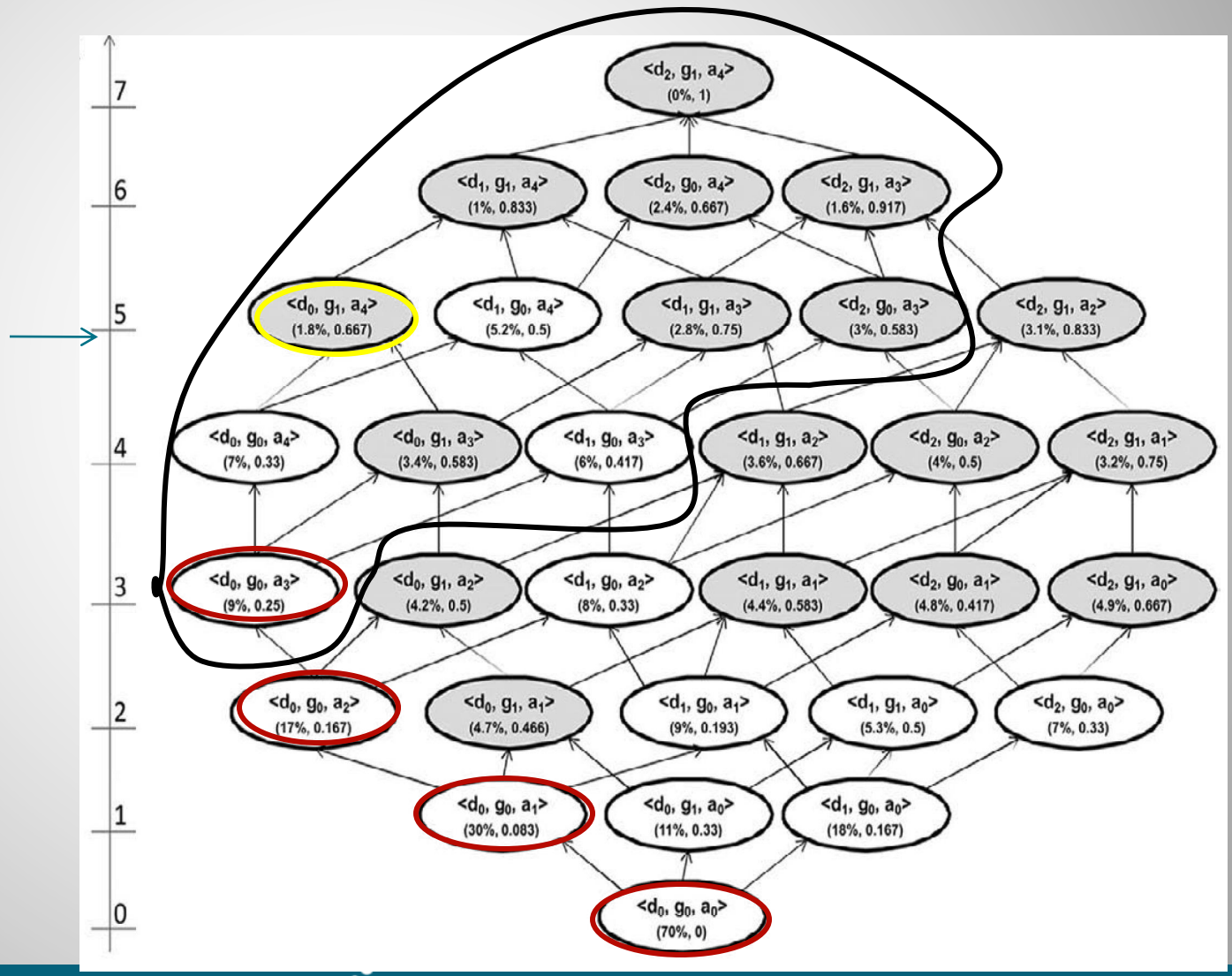


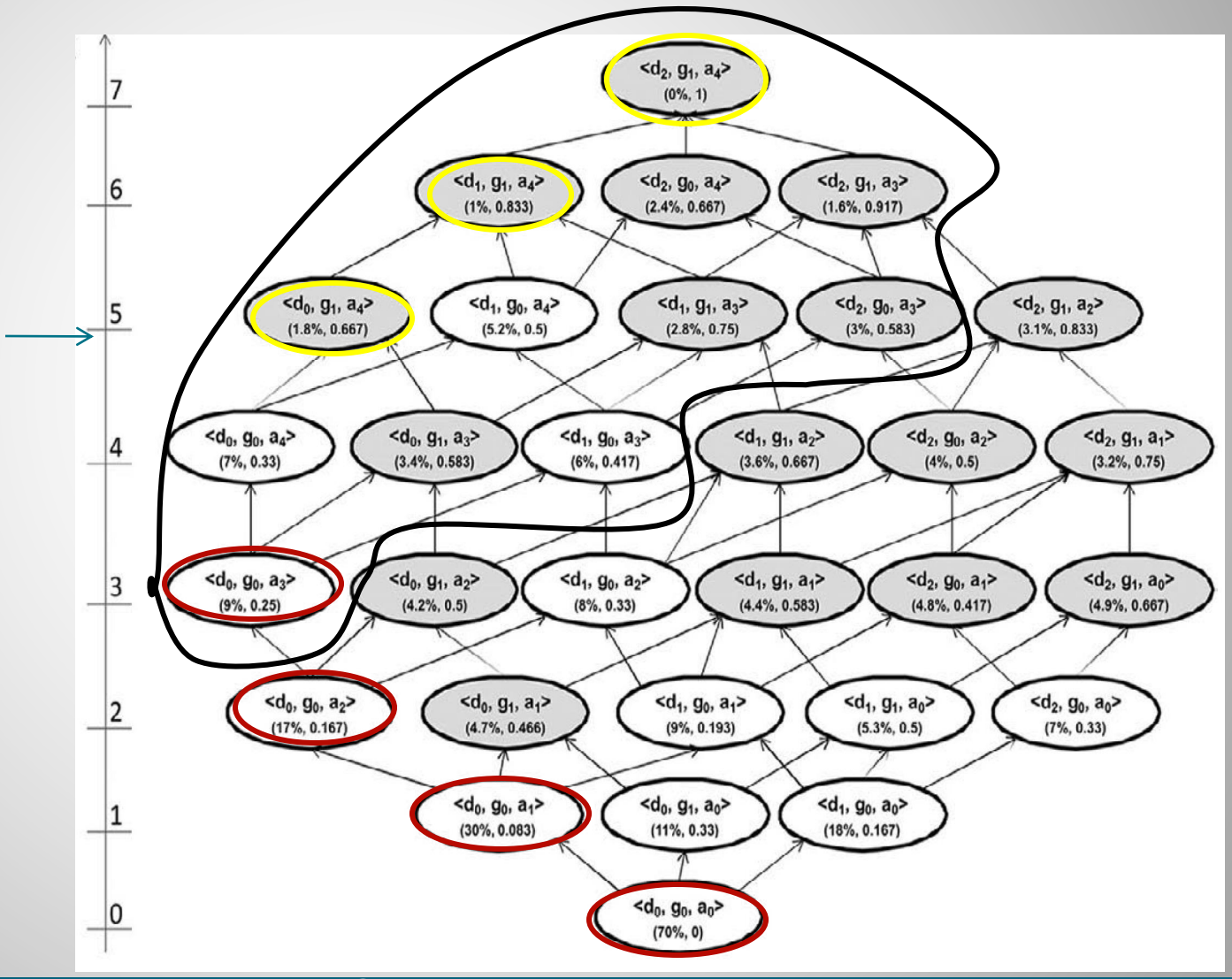
Chaque nœud est en relation

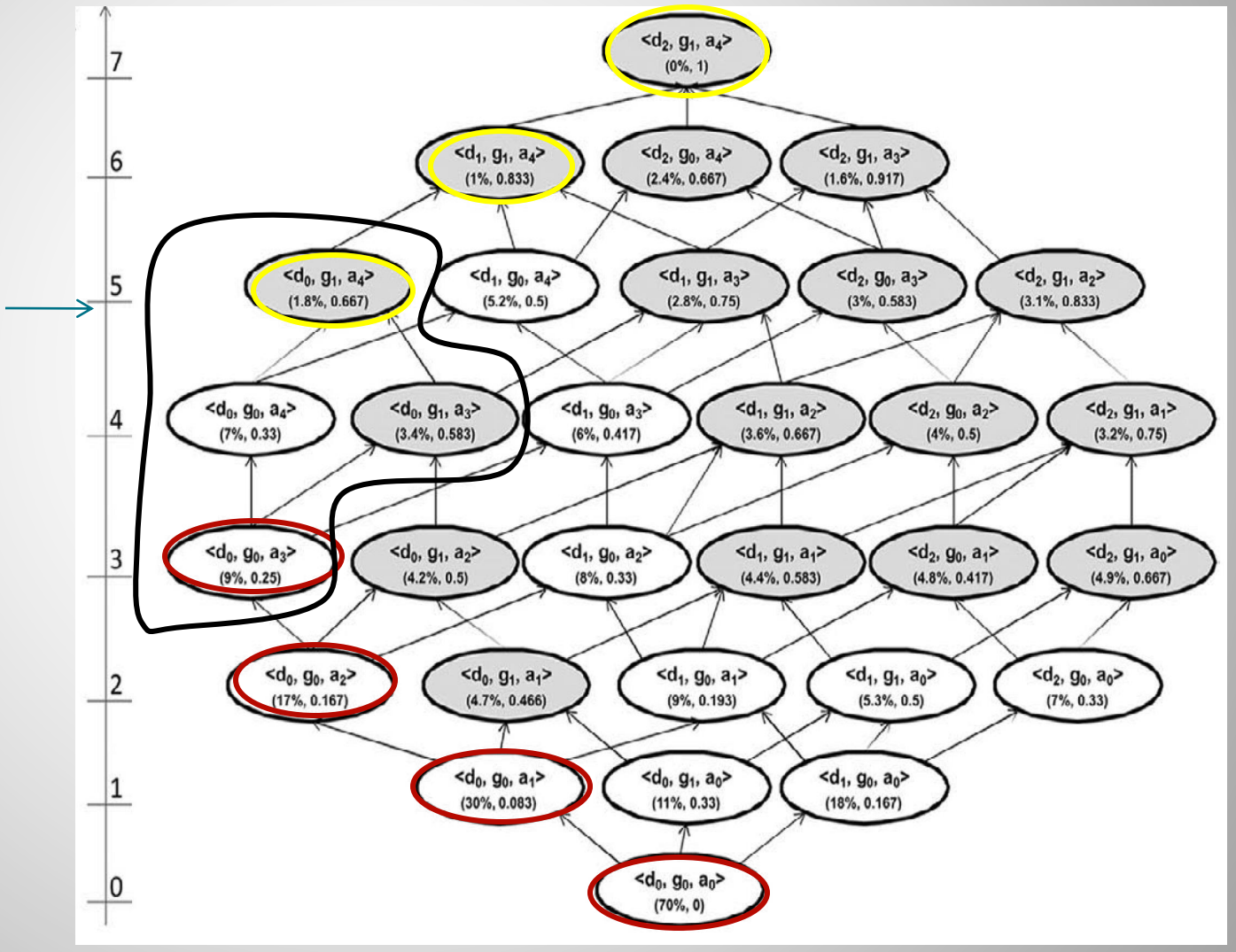


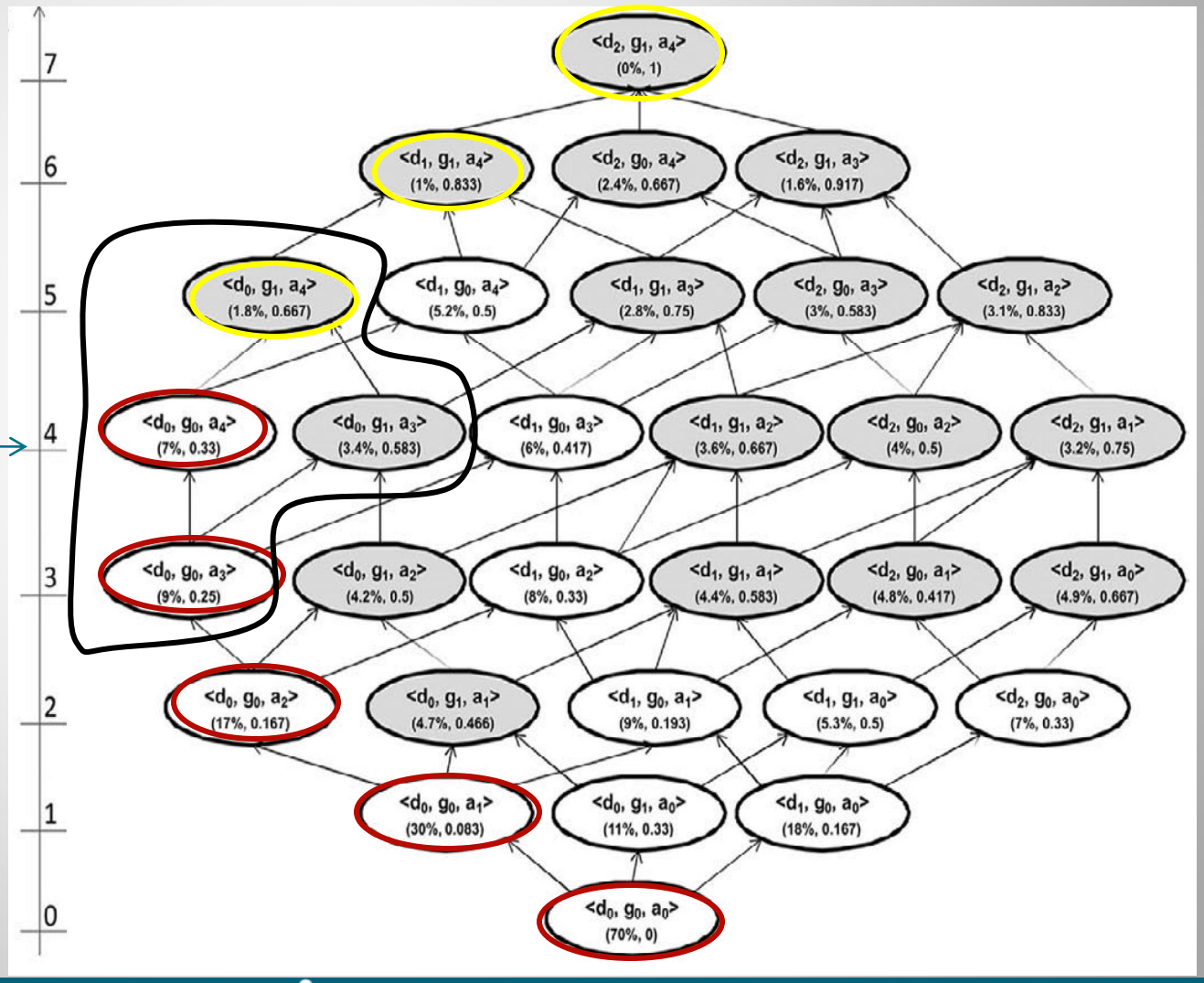

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

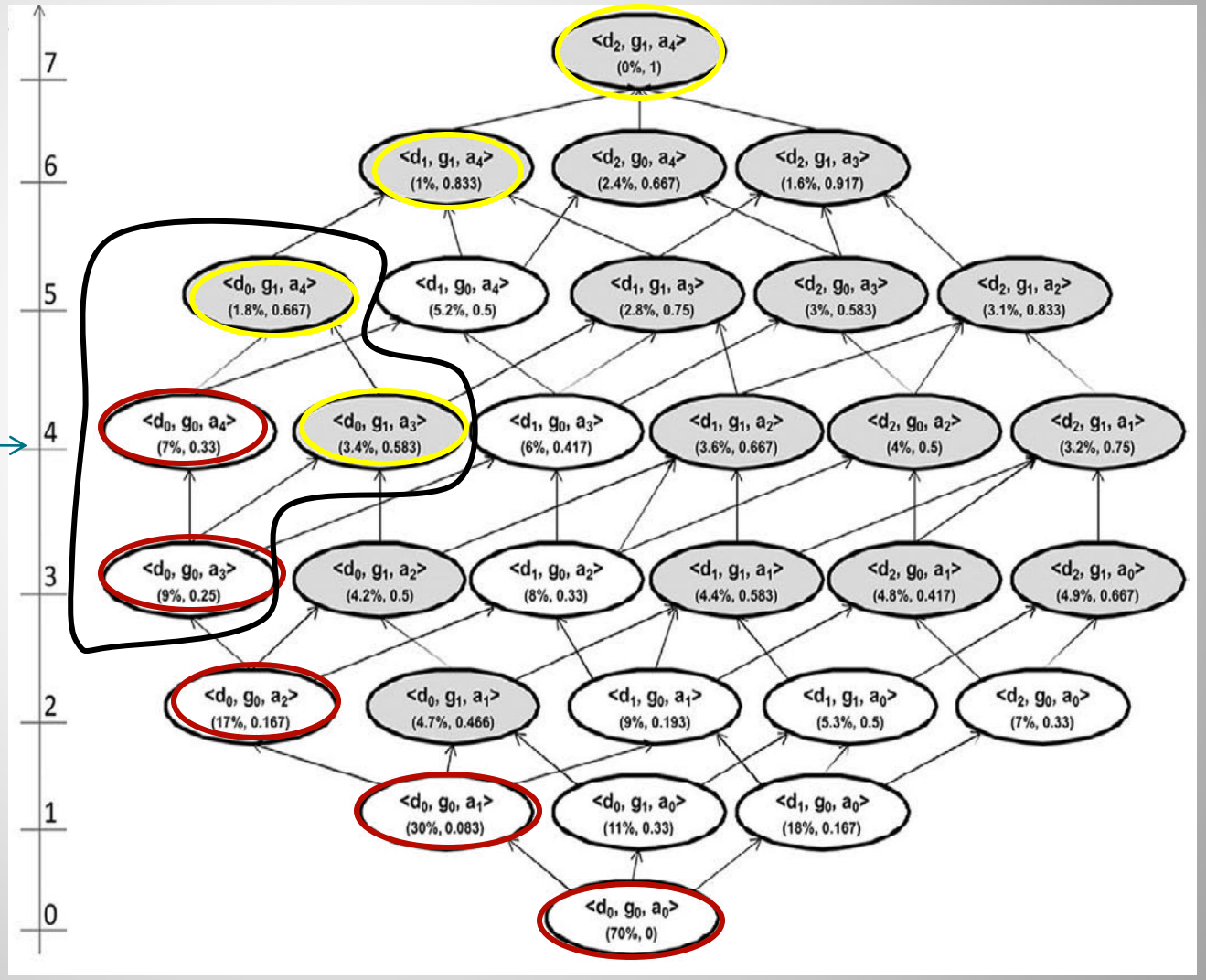

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

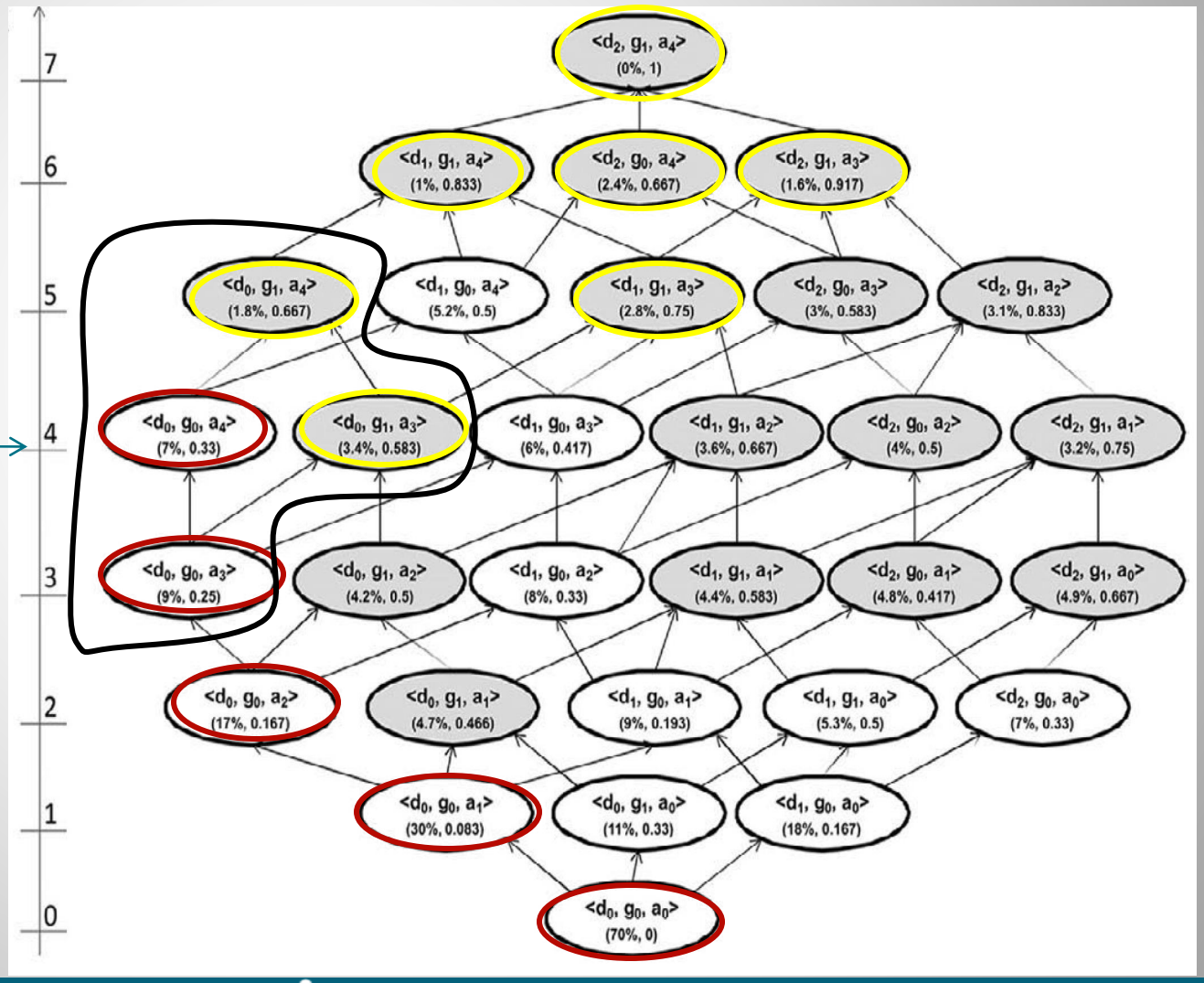

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

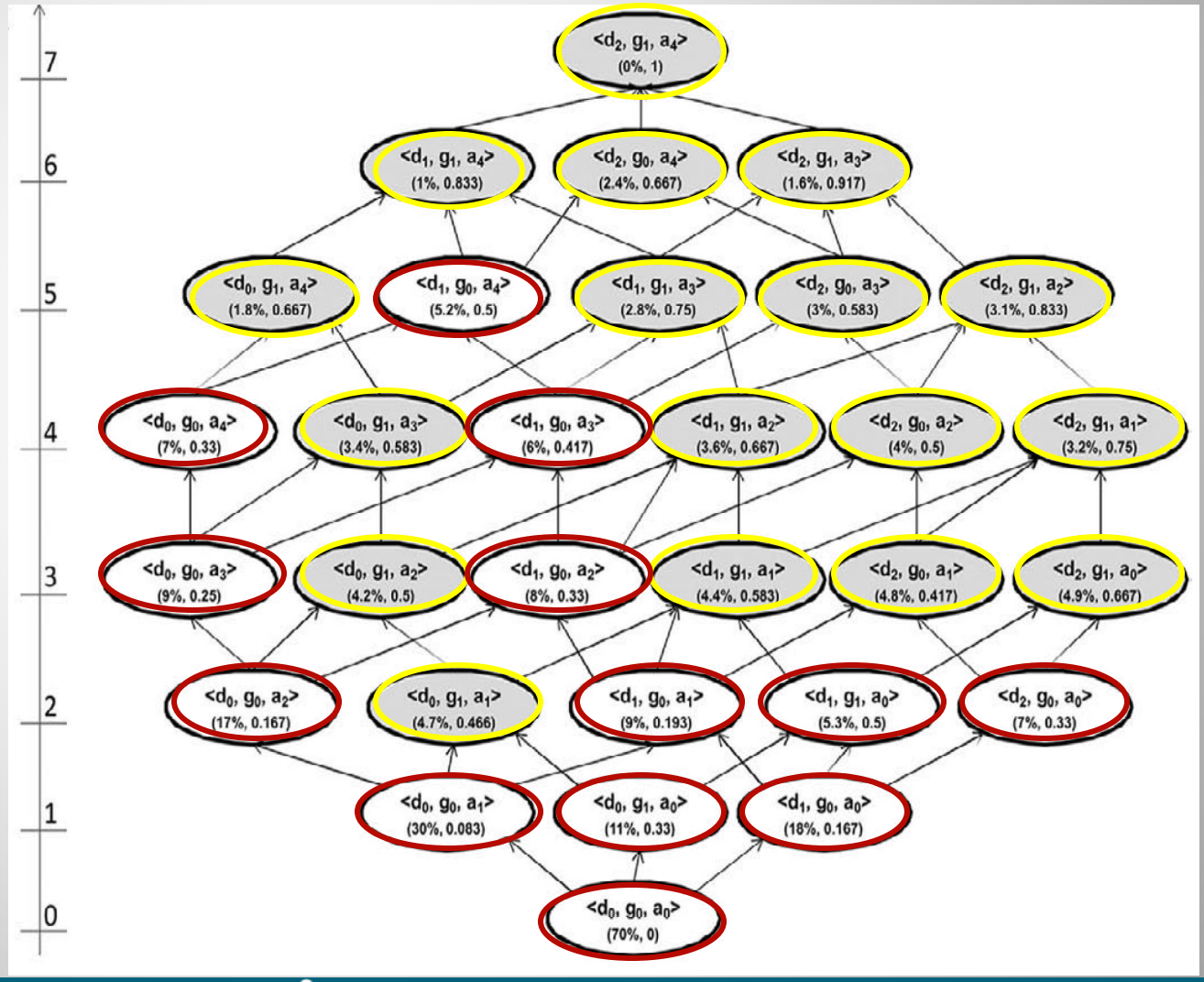

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)


7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

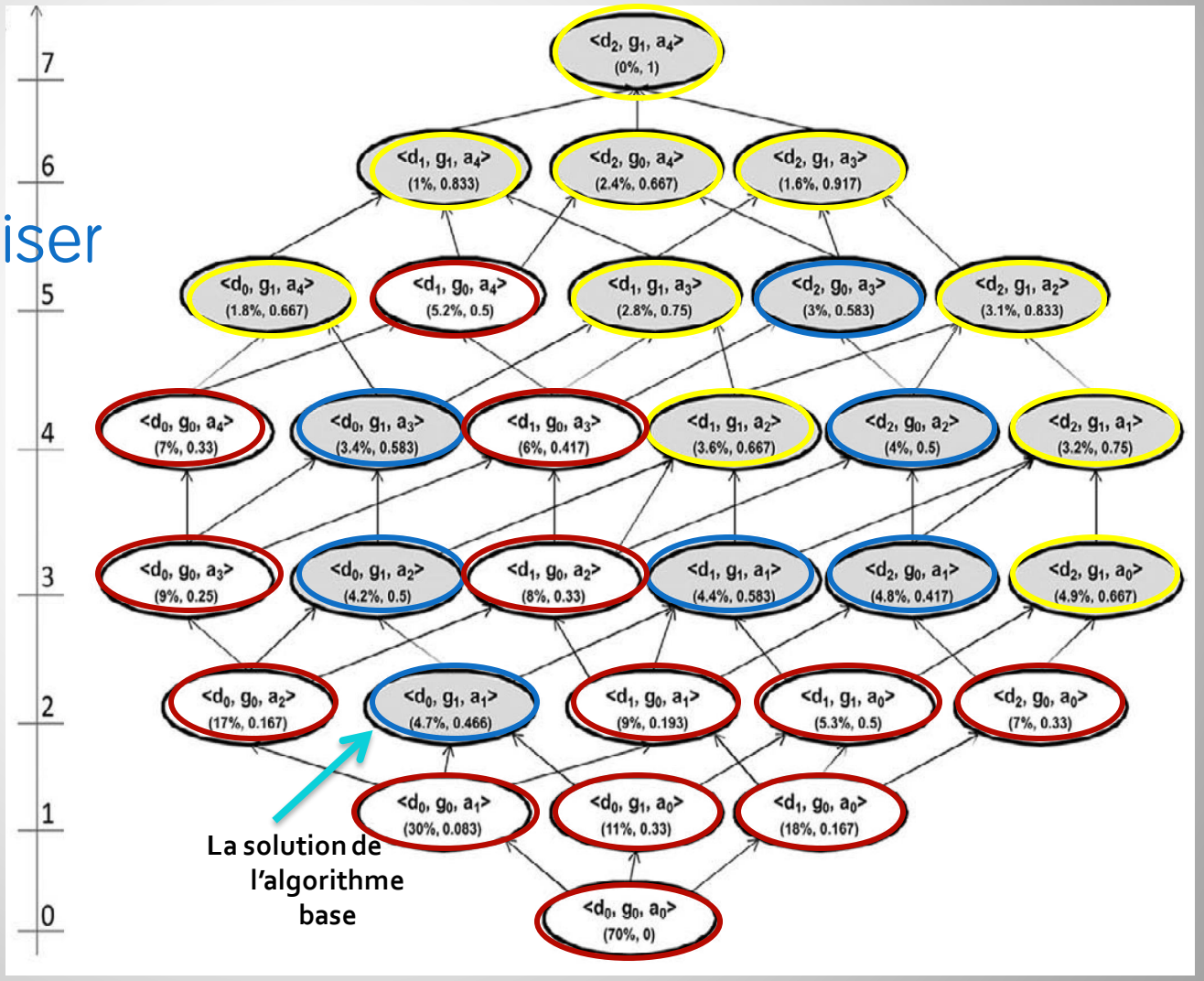

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

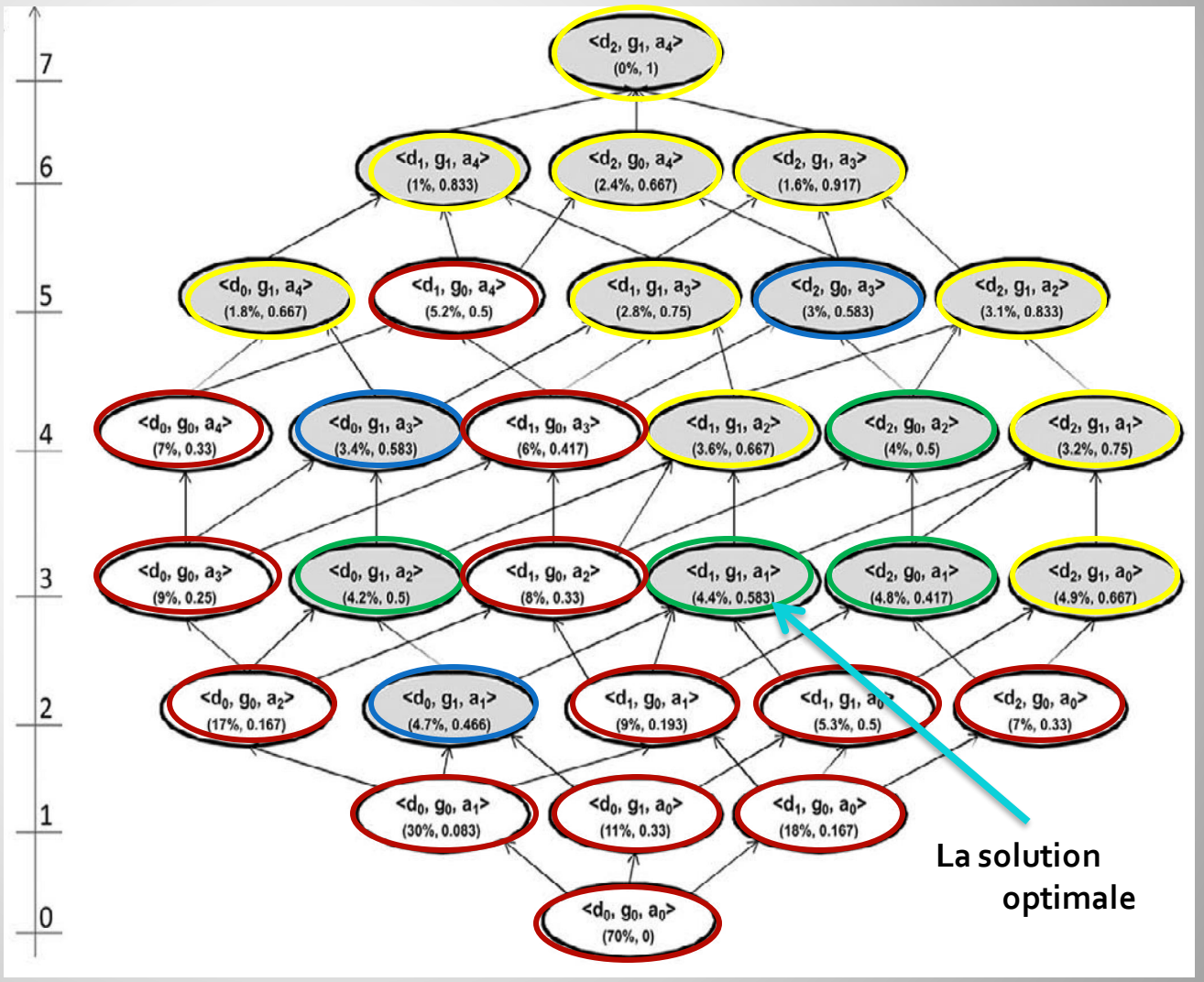

7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)


7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)


7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)

- 2 solutions pour 5%
- On veut maximiser la protection




7
6
5
4
3
2
1
0
<d2, g1, a4> (0%, 1)
<d1, g1, a4> (1%, 0.833)
<d2, g0, a4> (2.4%, 0.667)
<d2, g1, a3> (1.6%, 0.917)
<d0, g1, a4> (1.8%, 0.667)
<d1, g0, a4> (5.2%, 0.5)
<d1, g1, a3> (2.8%, 0.75)
<d2, g0, a3> (3%, 0.583)
<d2, g1, a2> (3.1%, 0.833)
<d0, g0, a4> (7%, 0.33)
<d0, g1, a3> (3.4%, 0.583)
<d1, g0, a3> (6%, 0.417)
<d1, g1, a2> (3.6%, 0.667)
<d2, g0, a2> (4%, 0.5)
<d2, g1, a1> (3.2%, 0.75)
<d0, g0, a3> (9%, 0.25)
<d0, g1, a2> (4.2%, 0.5)
<d1, g0, a2> (8%, 0.33)
<d1, g1, a1> (4.4%, 0.583)
<d2, g0, a1> (4.8%, 0.417)
<d2, g1, a0> (4.9%, 0.667)
<d0, g0, a2> (17%, 0.167)
<d0, g1, a1> (4.7%, 0.466)
<d1, g0, a1> (9%, 0.193)
<d1, g1, a0> (5.3%, 0.5)
<d2, g0, a0> (7%, 0.33)
<d0, g0, a1> (30%, 0.083)
<d0, g1, a0> (11%, 0.33)
<d1, g0, a0> (18%, 0.167)
<d0, g0, a0> (70%, 0)
La solution optimale

En conclusion de l'Anonymisation

- Survol rapide de l'aspect réglementaire
- Les attaques peuvent se faire sur les QID / attributs sensibles
- Il y a différentes méthodes d'attaques
- L'anonymisation optimale est un problème difficile (NP-Hard)
- On peut faire des benchmarks sur la qualité du traitement en utilisant un oracle

RGPD – un exemple

RGPD

- Le RGPD, mise en application le 25 mai 2018
- ➔Harmoniser le panorama juridique européen en matière de protection des données personnelles pour des personnes physiques.

RGPD

- **RÈGLEMENT (UE) 2016/679 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 27 avril 2016**
- Se découpe en 2 éléments :
  - Un règlement général qui couvre l'essentiel du traitement des données au sein de l'U.E.
  - Une directive sur la protection des données – non-étudiée

RGPD

Vision du RGPD par le comptage des mots

Mot	Occurences
pseudonymis*	15
profilage	22
amend*	23
secret	24
sécurité	57
santé	66
respect*	99
sous-trait*	441
contrôl*	519
responsab*	519
caractère	582
personnel*	588
données	841
traitement*	1111

Internet = 10 occurences

L'articulation des contrôles, des responsabilités et des sous-traitants

Portée du RGPD

- S'applique pour les personnes physique dans l'U.E. pour des traitements hors U.E.
  - *Le traitement de données à caractère personnel de personnes concernées qui se trouvent dans l'Union par un responsable du traitement ou* ***un sous-traitant qui n'est pas établi dans l'Union*** *devrait également être soumis au présent règlement lorsque ledit traitement est lié au suivi du comportement de ces personnes dans la mesure où il s'agit de leur comportement au sein de l'Union. Afin de déterminer si une activité de traitement peut être considérée comme un suivi du comportement des personnes concernées, il y a lieu d'établir si les personnes physiques sont suivies sur internet, ce qui comprend l'utilisation ultérieure éventuelle de techniques de traitement des données à caractère personnel* ***qui consistent en un profilage d'une personne physique****, afin notamment de prendre des décisions la concernant ou d'analyser ou de prédire ses préférences, ses comportements et ses dispositions d'esprit.*

RGPD : privacy

- La RGPD introduit « Privacy by design »
  - La protection de la vie privée doit être réalisée par des mesures techniques et organisationnelles adaptées au moment de la conception et lors de l'exécution du traitement.
  - Protection intégrée dans le cycle de vie de la technologie
  - Paramètres de respect de la vie privée dans les services et produits par défaut

RGPD : privacy

- La RGPD introduit « Privacy by design »
  - La protection de la vie privée doit être réalisée par des mesures techniques et organisationnelles adaptées au moment de la conception et lors de l'exécution du traitement.➔ ie. GUID / Salt
  - Protection intégrée dans le cycle de vie de la technologie ➔ ie. chiffrage https
  - Paramètres de respect de la vie privée dans les services et produits par défaut ➔ ie. cookies traceurs doivent être approuvés

RGPD : sous-traitance

- Encadrement étroit des activités des sous-traitants et une responsabilité accrue.
  - ➔Elaborer une **documentation constante** et des missions précises pour les sous-traitants. (article 82)
  - *Tout responsable du traitement ayant participé au traitement est responsable du dommage causé par le traitement qui constitue une violation du présent règlement. Un sous-traitant n'est tenu pour responsable du dommage causé par le traitement que s'il n'a pas respecté les obligations prévues par le présent règlement qui incombent spécifiquement aux sous-traitants ou qu'il a agi en-dehors des instructions licites du responsable du traitement ou contrairement à celles-ci.*

RGPD : analyse du risque

- Analyse du risque
- *(84) le responsable du traitement devrait assumer la responsabilité d'effectuer une analyse d'impact relative à la protection des données pour évaluer, en particulier, l'origine, la nature, la particularité et la gravité de ce risque*

RGPD : analyse du risque

- Exemples de risque
  - Surveillance à grande échelle de zones accessibles au public ➔ ie. floutage
  - Suivi régulier et systématique de personnes ➔ ie. log
  - Mise à disposition de données à des personnes de manière illimitée ➔ ie. qualité de l'anonymisation
  - Traitements soumis à autorisation ou à consultation du DPO

RGPD : mécanisme d'alerte

- Article 40 / Codes de conduite + (85)
- *la notification aux autorités de contrôle des violations de données à caractère personnel et la communication de ces violations aux personnes concernées;*
  - ➔ La notification intervient dans les 72 heures
  - *Si une telle notification ne peut avoir lieu dans ce délai de 72 heures, la notification devrait être assortie des motifs du retard et des informations peuvent être fournies de manière échelonnée sans autre retard indu.*

RGPD : le profilage

- Profilage
- *«profilage», **toute forme de traitement automatisé** de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment **pour analyser ou prédire** des éléments concernant le rendement au travail, la situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique;*
- Droit à s'opposer au profilage
- Pas de profilage ‹13 ans

RGPD : le profilage

- Le profilage ➔ permet de calculer un profil
- (71) *d'obtenir une explication quant à la décision prise à l'issue de ce type d'évaluation et de contester la décision. Cette mesure ne devrait pas concerner un enfant.*

RGPD : la portabilité des données

- Article 20 / Droit à la portabilité des données
- Portabilité des données ➔ permettre à la personne concernée de transférer les données personnelles fournies d'un traitement à un autre (ie. réseau social)

- Incitation à mettre au point des formats interopérables (68)
- *…dans un format structuré, couramment utilisé, lisible par machine et **interopérable**, et de les transmettre à un autre responsable du traitement. Il y a lieu d'encourager les responsables du traitement à mettre au point des formats interopérables permettant la portabilité des données.*

RGPD : DPO

- Délégué à la protection des données = Data Privacy Officer (DPO)

Articles 37 / 38 / 39

➔ personne interne ou prestataire

➔ action gratuite dans l'entreprise où il exerce ses missions. Soumis au secret professionnel.

Tient à jour un registre des traitements de données personnelles.

- Le DPO
  - **d'informer et de conseiller** l'organisme qui vous a désigné, ainsi que ses employés ;
  - **de contrôler le respect du règlement** et du droit national en matière de protection des données ;
  - **de conseiller l'organisme** sur la réalisation d'une analyse d'impact relative à la protection des données et d'en vérifier l'exécution ;
  - **être contacté par les personnes concernées pour toute question ;**
  - **de coopérer avec la CNIL** et d'être son point de contact.
- Le DPO interagit avec le **Responsable de Traitement,** l'équipe Juridique et le Top Management

CNIL Lanceur d'Alerte

- Êtes-vous lanceur d'alerte ? https://www.cnil.fr/fr/lanceurs-dalerte-adresser-une-alerte-la-cnil
  - Ce dispositif de la CNIL est réservé aux personnes physiques qui signalent ou divulguent, sans contrepartie financière directe et de bonne foi, des informations portant sur les données personnelles, et plus particulièrement :
  - une violation du droit de l'Union européenne, de la loi Informatique et Libertés ou du règlement général sur la protection des données (RGPD) ;
    - un crime ;
    - un délit ;
    - une menace ou un préjudice pour l'intérêt général ;
    - une autre violation ou une tentative de dissimulation d'une violation :
    - d'un engagement international régulièrement ratifié ou approuvé par la France ;
    - d'un acte unilatéral d'une organisation internationale pris sur le fondement d'un tel engagement.
- Lorsque les informations n'ont pas été obtenues dans le cadre des activités professionnelles, vous devez en avoir eu personnellement connaissance.

Le règlement relatif à l'intelligence artificielle (RIA)

En date du 13 juin 2024, en vigueur au 1er aout 2024

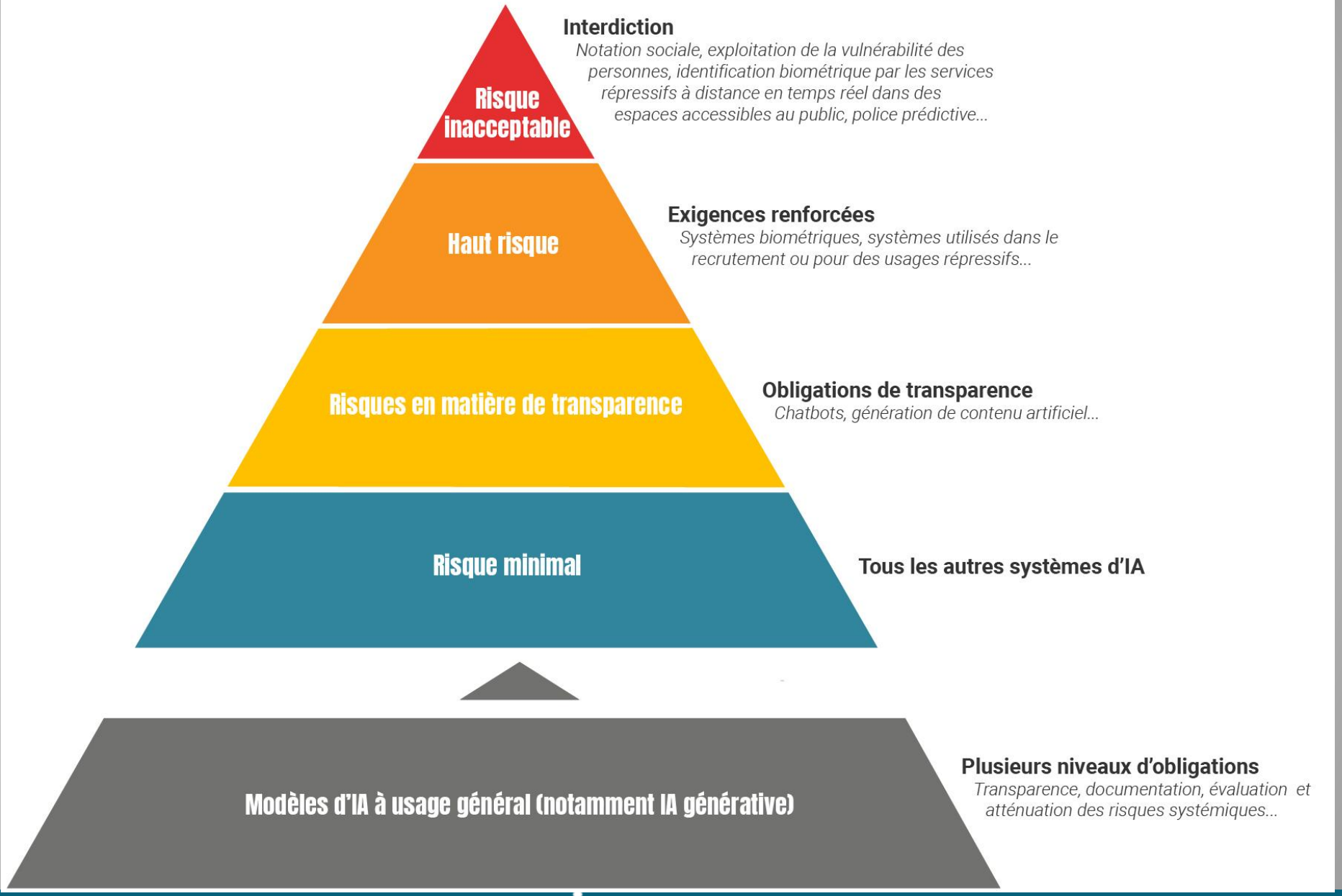
Principes fondateurs ➔ notion de risque élevé si un principe est mis en défaut, avec exemptions si sécurité-défense.

1. action humaine et contrôle humain
2. robustesse technique et sécurité
3. respect de la vie privée et gouvernance des données
4. transparence
5. diversité, non-discrimination et équité
6. bien-être sociétal et environnemental
7. responsabilité

144 pages
Dans l'espace bibliographique
OJ_L_202401689_FR_TXT.pdf

les systèmes d'IA désignés comme étant à haut risque devraient être limités aux systèmes qui ont une incidence préjudiciable substantielle sur la santé, la sécurité et les droits fondamentaux des citoyens dans l'Union et une telle limitation devrait réduire au minimum toute éventuelle restriction au commerce international.

Règlement IA – approche par les risques



L'analyse d'impact sur la protection des données

- **Les risques liés à l'IA à prendre en compte dans une AIPD**
- Les traitements de données personnelles reposant sur des systèmes d'intelligence artificielle présentent des risques spécifiques qu'il convient de prendre en compte :
- les risques pour les personnes concernées liés à des mésusages des données contenues dans la base d'apprentissage, notamment en cas de violation de données ;
- le risque d'une discrimination automatisée causée par un biais du système d'IA introduit lors du développement, par exemple lié à une performance moindre du système pour certaines catégories de personnes ;
- le risque de produire du contenu fictif erroné sur une personne réelle, particulièrement important dans le cas des systèmes d'IA génératives, et pouvant avoir des conséquences sur sa réputation ;
- le risque de prise de décision automatisée causée par un biais d'automatisation ou de confirmation dans le cas où les mesures d'explicabilité nécessaires ne sont pas prises lors du développement de la solution (comme la remontée d'un score de confiance, ou d'informations intermédiaires tel qu'une carte de saillance ou « *saliency map* ») ou si un agent utilisant le système d'IA ne peut pas prendre une décision contraire sans que cela ne lui porte préjudice ;
- les risques liés aux attaques connues spécifiques aux systèmes d'IA tel que les attaques par empoisonnement des données, par insertion d'une porte dérobée, ou encore par inversion du modèle ;
- les risques liés à la confidentialité des données susceptibles d'être extraites depuis le système d'IA ;
- les risques éthiques systémiques et graves liés au déploiement du système, tels que les impacts sur le fonctionnement démocratique de la société, ou encore sur le respect des droits fondamentaux (par exemple en cas de discrimination), et pouvant être pris en compte lors de la phase de développement.
- Enfin, le risque d'une perte de contrôle des utilisateurs sur leurs données accessibles en ligne, une collecte à large échelle étant souvent nécessaire à l'apprentissage d'un système d'IA, notamment lorsque celles-ci sont collectées par moissonnage ou *web scraping*.

Synthèse RIA vs RGPD

	RIA	RGPD
Champ d'application	Le développement, la mise sur le marché ou le déploiement de systèmes et modèles d'IA	Tout traitement de données personnelles indépendamment des dispositifs techniques utilisés (dont les traitements visant à développer un modèle ou système d'IA (données d'entrainement), et les traitements réalisés au moyen d'un système d'IA)
Acteurs visés	Principalement les fournisseurs et déployeurs de systèmes d'IA (dans une moindre mesure les importateurs, distributeurs et mandataires)	Responsables de traitements et sous-traitants
		(dont les fournisseurs et déployeurs soumis au RIA)
Approche	Approche par les risques pour la santé, la sécurité ou les droits fondamentaux, notamment à travers la sécurité des produits et la surveillance du marché en ce qui concerne les systèmes et modèles d'IA	Approche fondée sur l'application de grands principes, l'évaluation des risques et la responsabilisation (accountability)
Modalité principale de l'évaluation de la conformité (non exhaustif)	Évaluation de conformité interne ou par un tiers, notamment au moyen d'un système de gestion des risques et au regard de normes harmonisées	Principe de responsabilité (documentation interne) et outils de la conformité (certification, code de conduite)
Principales sanctions applicables	Retrait du marché ou rappel de produits	Mise en demeure (pouvant enjoindre de mettre le traitement en conformité, de le limiter temporairement ou définitivement, y compris sous astreinte)
	Amendes administratives pouvant aller jusqu'à 35 millions d'euros ou **7% du chiffre d'affaires annuel mondial**	Amendes administratives pouvant aller jusqu'à 20 millions d'euros ou **4% du chiffre d'affaires annuel mondial**

Utilisation d'Audit pour valider un produit à base d'IA

- Cf. referentiel-certification-LNE-processus-IA.pdf

➔ la certification est recommandée

- Exemple page 38

exigence	information visée	client **spécifique**		client **générique**
		doit être **tenu à disposition**	doit être **communiqué**	doit être **communiqué via la fiche produit**
III.3.1	les spécifications relatives à la fonctionnalité d'IA et les critères d'acceptation de chacune des exigences	X		X
III.3.1	éléments de communication avec le client définis lors de la phase de conception	selon ce qui a été défini	selon ce qui a été défini	
III.4.1.3	l'infrastructure (matérielle, système d'exploitation, logicielle), les types de déploiement (cloud public ou privé, on-premise etc.) supportés par la fonctionnalité d'IA et la dépendance à des technologies d'IA sous-jacentes		X	X
III.4.1.4	les interfaces nécessaires à l'usage de la fonctionnalité d'IA		X	X
III.4.1.5	les caractéristiques du domaine d'utilisation visé, notamment les principaux facteurs d'influence sur les performances de la fonctionnalité d'IA		X	X