

Hachage

Distribution de données

D.E ZEGOUR

École Supérieure d'Informatique

ESI

Hachage : Distribution des données

Distribution de Poisson

Supposons

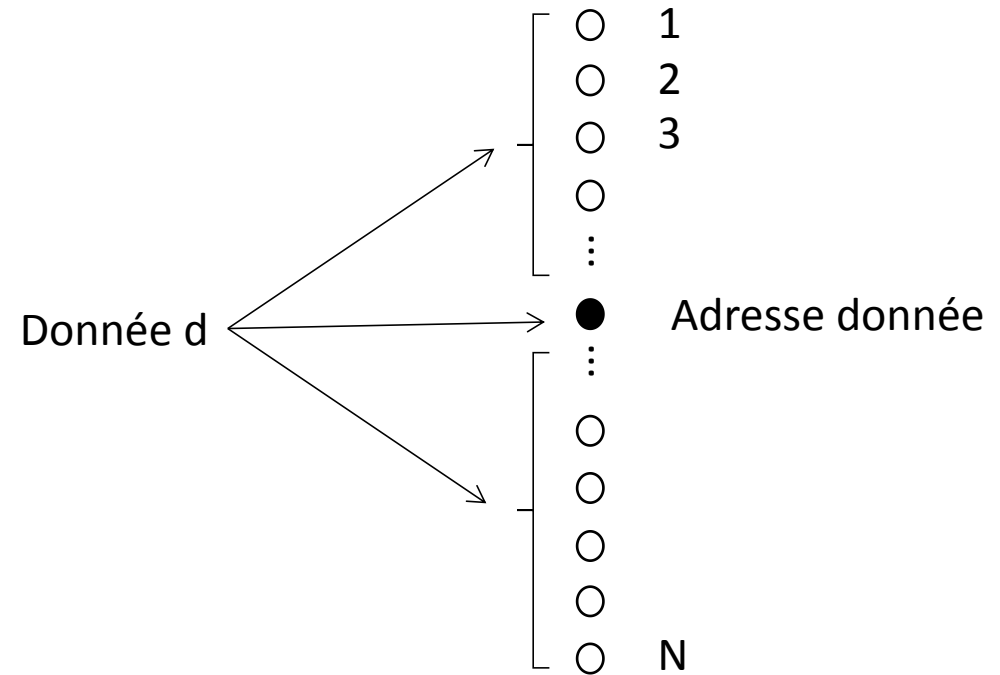
- N adresses possibles
- une adresse donnée
- une donnée d à hacher

Considérons les 2 événements :

A : une adresse donnée n'est pas choisie

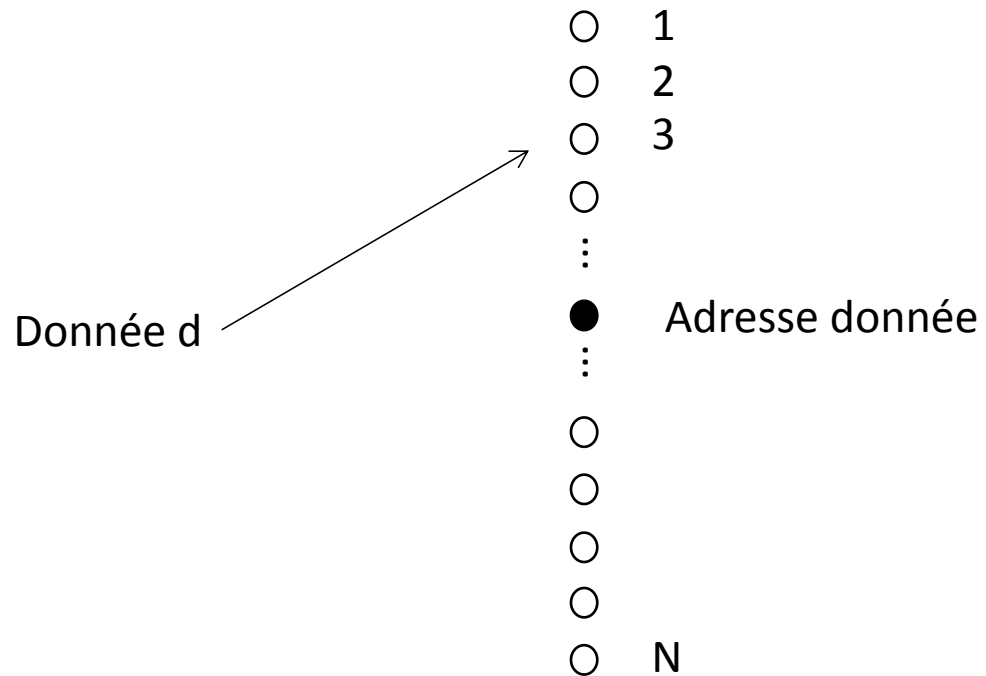
B : une adresse donnée est choisie

Quand une donnée est hachée, l'un des événements (A ou B) se produit pour une adresse donnée.



Hachage : Distribution des données

Distribution de Poisson



Qu'elle est la probabilité qu'une adresse donnée ne soit pas choisie ?

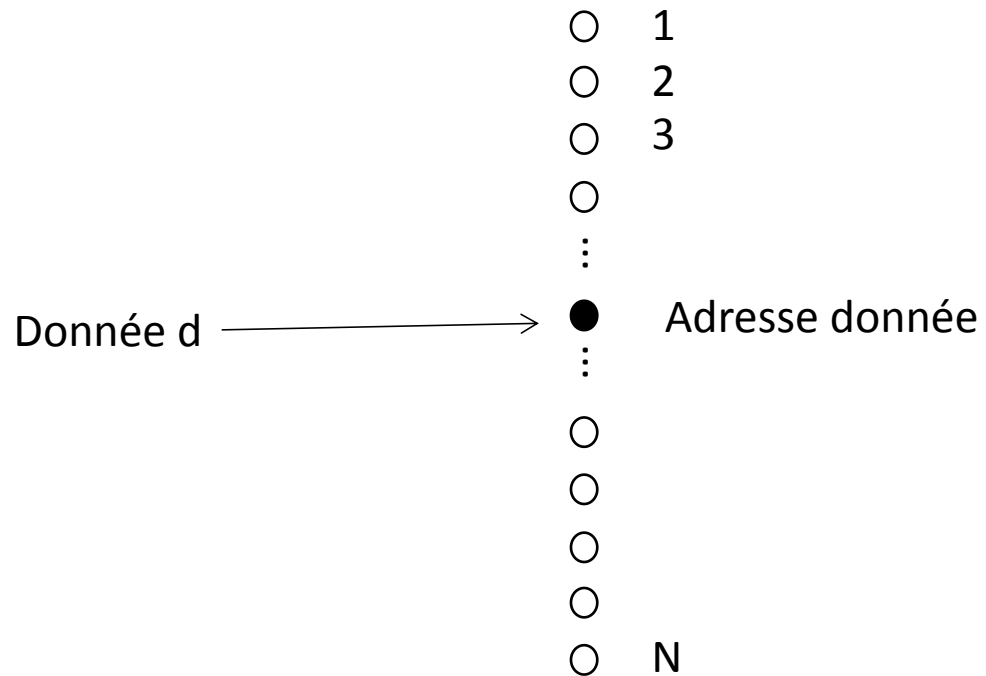
Evènement : A

$$P(A) = a = 1 - 1/N = (N-1)/N = a$$

si $N=10$ alors $a=0.1$

Hachage : Distribution des données

Distribution de Poisson



Qu'elle est la probabilité qu'une adresse donnée soit choisie ?

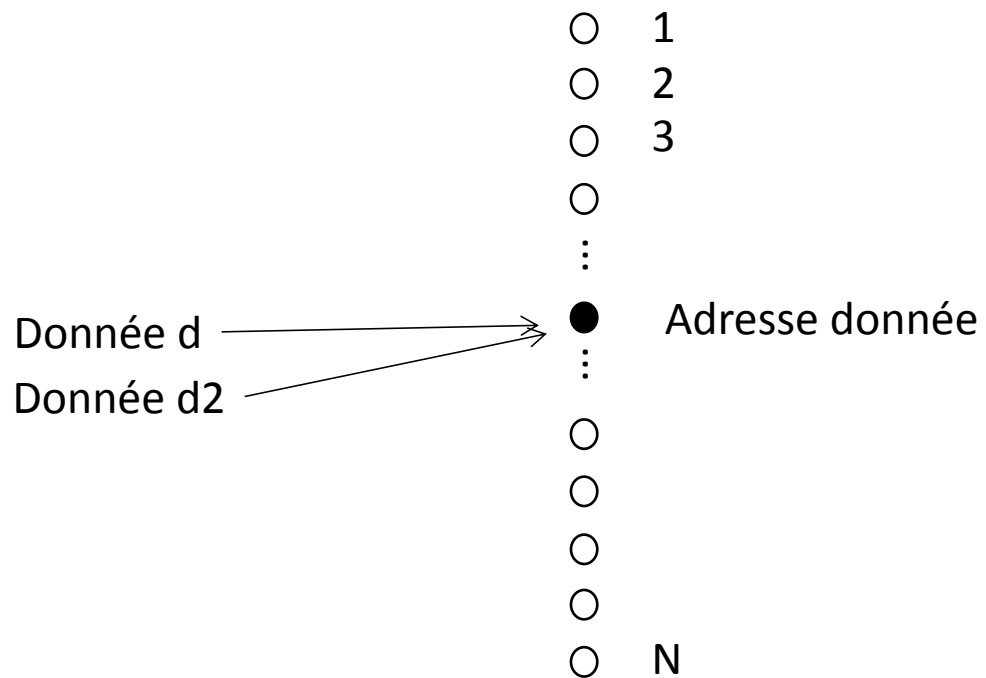
Evènement : B

$$P(B) = b = 1 - 1/N = (N-1)/N$$

si $N=10$ alors $a=0.09$

Hachage : Distribution des données

Distribution de Poisson



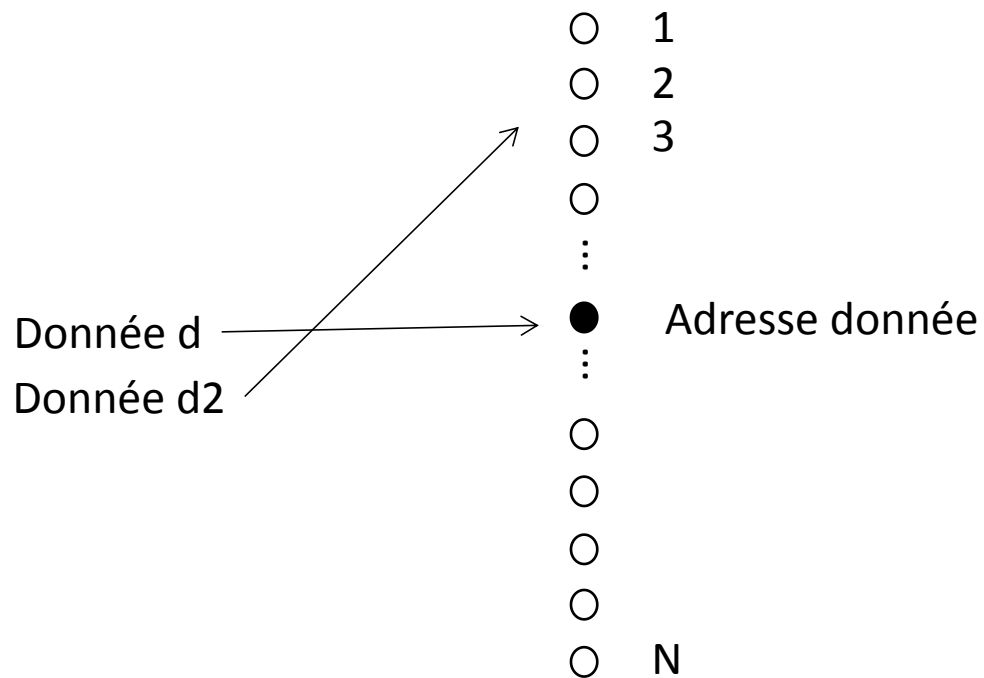
Qu'elle est la probabilité pour que deux données hachent la même adresse ?

Evènement : B B

$P(BB) = b.b = 1/N \cdot 1/N$ (événements indépendants)

Hachage : Distribution des données

Distribution de Poisson



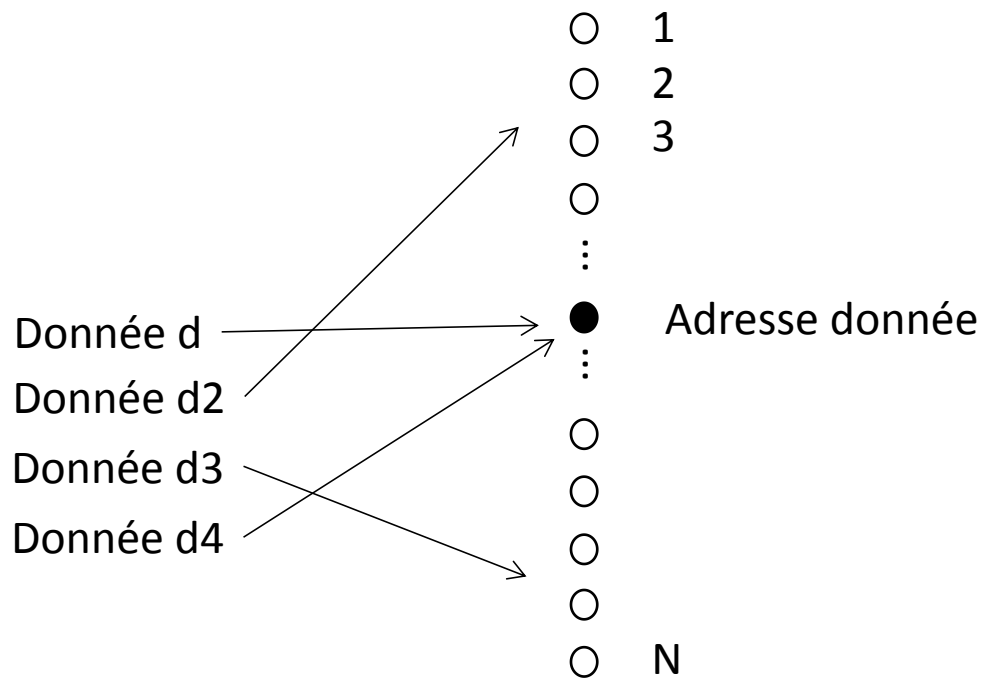
Qu'elle est la probabilité pour que la première donnée soit hachée à l'adresse donnée et que la seconde donnée soit hachée dans une adresse différente de la première ?

Evènement : B A

$$P(BA) = b.a = 1/N \cdot (N-1/N)$$

Hachage : Distribution des données

Distribution de Poisson



Qu'elle est la probabilité associé à l'évènement
BAAB ?

$$P(\text{BAAB}) = b.a.a.b = a^2b^2 = (1/N)^2 (N-1/N)^2$$

Hachage : Distribution des données

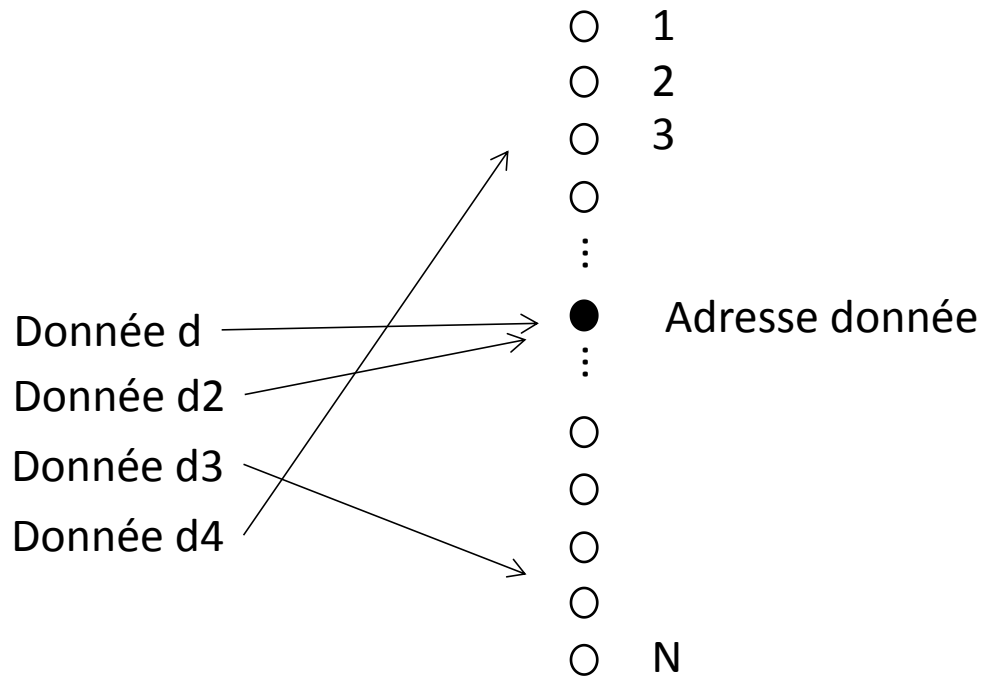
Distribution de Poisson

Probabilité pour que deux données parmi quatre hachent la même adresse ?

Trouver tous les événements possibles:
BBAA, BAAB, BABA, AABB, ABBA, ABAB

Hachage : Distribution des données

Distribution de Poisson



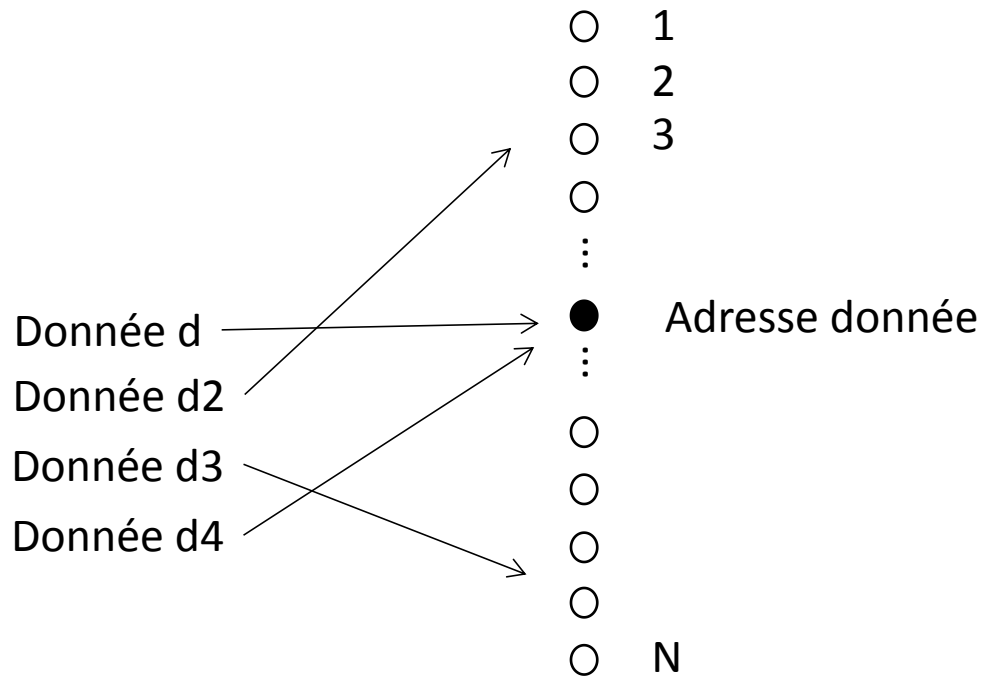
Probabilité pour que deux données parmi quatre hachent la même adresse ?

Trouver tous les événements possibles:

BBAA, BAAB, BABA, AABB, ABBA, ABAB

Hachage : Distribution des données

Distribution de Poisson



Probabilité pour que deux données parmi quatre hachent la même adresse ?

Trouver tous les événements possibles:
BBAA, **BAAB**, BABA, AABB, ABBA, ABAB

Hachage : Distribution des données

Distribution de Poisson

Probabilité pour que deux données parmi quatre hachent la même adresse ?

Trouver tous les événements possibles:
AABB, BAAB, BABA, BBAA, ABBA, ABAB

$$P = P(\text{AABB}) + P(\text{BAAB}) + \dots = 6 a^2 b^2$$
$$P = C_4^2 a^2 b^2.$$

C_4^2 représente le nombre de façons qu'on peut placer 2 A (et 2 B) dans 4 places.

Hachage : Distribution des données

Distribution de Poisson

Généralisation

Probabilité pour que x données parmi r hachent la même adresse ?

x fois l'événement B et $(r-x)$ fois l'événement A.

Trouver tous les événements possibles: C_r^x

$$P(x) = C_r^x \cdot a^{r-x} \cdot b^x$$

avec $C_r^x = x! / (x! (r-x)!)$

Ce qui veut aussi dire : Probabilité qu'une adresse donnée soit choisie x fois et ne soit pas choisie $r-x$ fois.

Hachage : Distribution des données

Distribution de Poisson

Calcul

Si N adresses possibles

$$P(x) = C_r^x a^{r-x} b^x = P(x) = C_r^x (1-1/N)^{r-x} (1/N)^x$$

$P(x=0)$ veut dire probabilité pour qu'une adresse donnée ne soit jamais choisie.

$$P(0) = C_r^0 (1-1/N)^r (1/N)^0$$

$P(x=1)$ veut dire probabilité pour qu'une adresse donnée soit choisie une seule fois

$$P(1) = C_r^1 (1-1/N)^{r-1} (1/N)^1$$

Inconvénient de la formule : difficile à calculer pour N et r grands.

La fonction de POISSON est une bonne approximation.

$$P(x) \approx f(x) = \left(\frac{r}{N} \right)^x \cdot e^{-(r/N)} / x!$$

Hachage : Distribution des données

Synthèse

Si N est le nombre d'adresses possibles,

r est le nombre de données insérées

x est le nombre de données ayant la même adresse (x fois l'évènement B et $r-x$ fois l'évènement A de C_r^x manière possible)

$P(x)$ donne la probabilité que x données hachent la même adresse parmi r données insérées.

$P(x)$ Probabilité qu'une adresse donnée soit choisie x fois et ne soit pas choisie $r-x$ fois

Hachage : Distribution des données

Synthèse

Si N est le nombre d'adresses possibles,

r est le nombre de données insérées

x est le nombre de données ayant la même adresse (x fois l'évènement B et $r-x$ fois l'évènement A de C_r^x manière possible)

$P(x)$ est aussi la proportion d'adresses ayant x données qui lui sont attribuées par hachage.

$N.P(x)$ est le nombre d'adresses qui ont x données attribuées

Ceci nous permet donc de prévoir le nombre de collisions (de données en débordement)

Formule : $N (P(2) + 2P(3) + ...iP(i+1) +)$

Hachage : Distribution des données

Prévention des collisions

Prenons $N = 10\,000$ adresses possibles et $r = 10\,000$ données insérées.
($P(0) = .3679$; $P(1) = .3679$; $P(2) = .1839$; $P(3) = .0613$; etc.)

Quel est le nombre d'adresses qui n'ont aucune donnée attribuée ?
 $10\,000 * P(0) = 3679$

Quel est le nombre d'adresses qui n'ont qu'une seule donnée attribué ?
 $10\,000 * P(1) = 3679$

Quel est le nombre d'adresses qui n'ont que deux données attribuées?
 $10\,000 * P(2) = 1839$
→ 1839 seront en débordement

Quel est le nombre d'adresses qui n'ont que trois données attribuées?
 $10\,000 * P(3) = 613$
→ $613 * 2$ seront en débordement

Mauvaise répartition : des milliers d'adresses avec aucune donnée attribuée.

Plus de 1226 ($= 2 * 613$) données seront en débordement

Hachage : Distribution des données

Réduction des collisions

Montrons à l'aide d'exemples d'une part, comment

l'augmentation du nombre d'adresses possibles

et d'autre part, comment

l'utilisation des cases

peuvent réduire les collisions.

Hachage : Distribution des données

Augmentation de l'espace des adresses

Nous définissons la densité d par : $d = r / N$

(r : nombre de données rangées ; N : nombre d'adresses possibles)

Regardons le comportement des fonctions de hachage (collisions) pour différentes valeurs de d .

$$P(x) = \frac{(r/N)^x \cdot e^{-(r/N)}}{x!} = \frac{d^x \cdot e^{-d}}{x!}$$

$P(x)$ dépend donc du rapport r/N c'est à dire de d .

Aussi, on constate un même comportement pour 500 données distribuées parmi 1000 adresses que 500.000 données distribuées parmi 1 million d'adresses. ($d = 50\%$ pour les deux cas)

Hachage : Distribution des données

Augmentation de l'espace des adresses

Prenons $d = 0.5$ ($N = 1000$ et $r = 500$ données)
avec $P(0) = 0.607$; $P(1) = 0.303$; $P(2) = 0.0758$; $P(3) = 0.0126$; $P(4) = 0.0016$; etc.

Combien d'adresses auront 0 donnée attribuée ?

$$1\,000 * P(0) = 607$$

Combien d'adresses auront 1 donnée attribuée ?

$$1\,000 * P(1) = 303$$

Combien d'adresses auront 2 donnée attribuées ?

$$1\,000 * P(2) = 76$$

Combien d'adresses auront au moins deux données attribuées ?

$$\rightarrow 1\,000 * (P(2) + P(3) + P(4) + \dots) = 90$$

Hachage : Distribution des données

Augmentation de l'espace des adresses

Quel est le nombre de données en débordement ?

$$\rightarrow 1\,000 * (P(2) + 2 * P(3) + 3 * P(4) + 4 * P(5) + \dots) = 107$$

Quel est le pourcentage des données en débordement ?

$$\rightarrow 107 / 500 = 21,4 \%$$

Conclusion:

Si la densité est de 50 %, on peut s'attendre à 78,6% de données rangées dans leur adresse primaire et 21,4 % rangées ailleurs.

Hachage : Distribution des données

Utilisation des cases ($b > 1$)

On accepte b données par adresse possible.

Dans ce cas : $d = r / (b.N)$

b : nombre de données par case
 N : nombre de cases
 r : nombre de données insérées.

	$b=1$	$b=2$
Nombre de données (r)	750	750
Nombre d'adresses (N)	1000	500
Densité (d)	0.75	0.75
Rapport (r/N)	0.75	1.5

Hachage : Distribution des données

Utilisation des cases ($b > 1$)

P(x)	b=1 (r/N = 0.75)	b=2 (r/N = 1.5)
P(0)	0.423	0.223
P(1)	0.354	0.335
P(2)	0.113 (Collisions)	0.251
P(3)	0.033 (Collisions)	0.126 (Collisions)
P(4)	0.006 (Collisions)	0.047 (Collisions)

b=1:
 $1\ 000.[P(2) + 2.P(3) + 3.P(4) + \dots] > 197$

b=2
 $5\ 00.[P(3) + 2.P(4) + 3 P(5) + \dots] > 110$

Nombre de données en débordement
dans chaque cas?

Plus la case est grande, plus les
performances sont meilleures.

Hachage : Distribution des données

Synthèse

Utilisation:

- Insertion avec contrôle(fixer un seuil)
- Bon compromis : chargement de 70 à 80 %

Avantage : Accès rapide à l'information ($O(1)$)

Inconvénients :

- Absence de l'ordre
- Limitation aux données statiques