Comment générer des échantillons uniformes sur des grandes masses de données ?

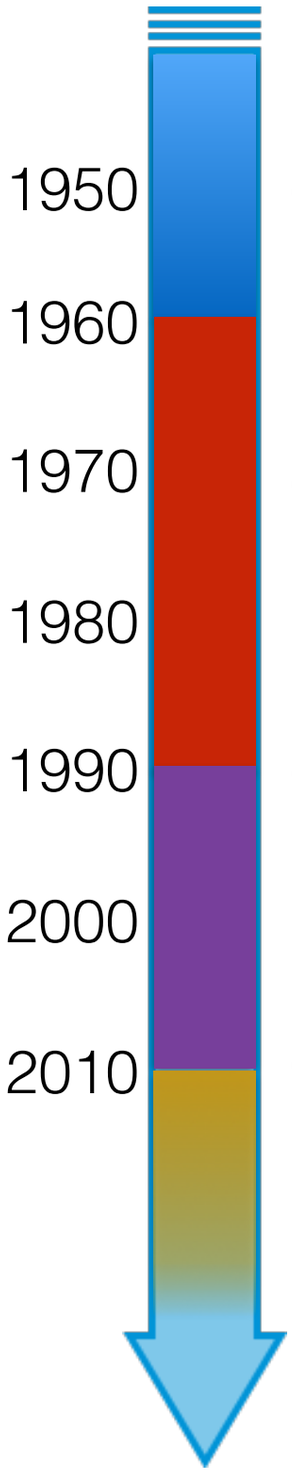
15 Octobre 2016

Yann Busnel

Crest (Ensai) / Inria Rennes – Bretagne Atlantique

9e Colloque Francophone sur les Sondages
Gatineau, Canada

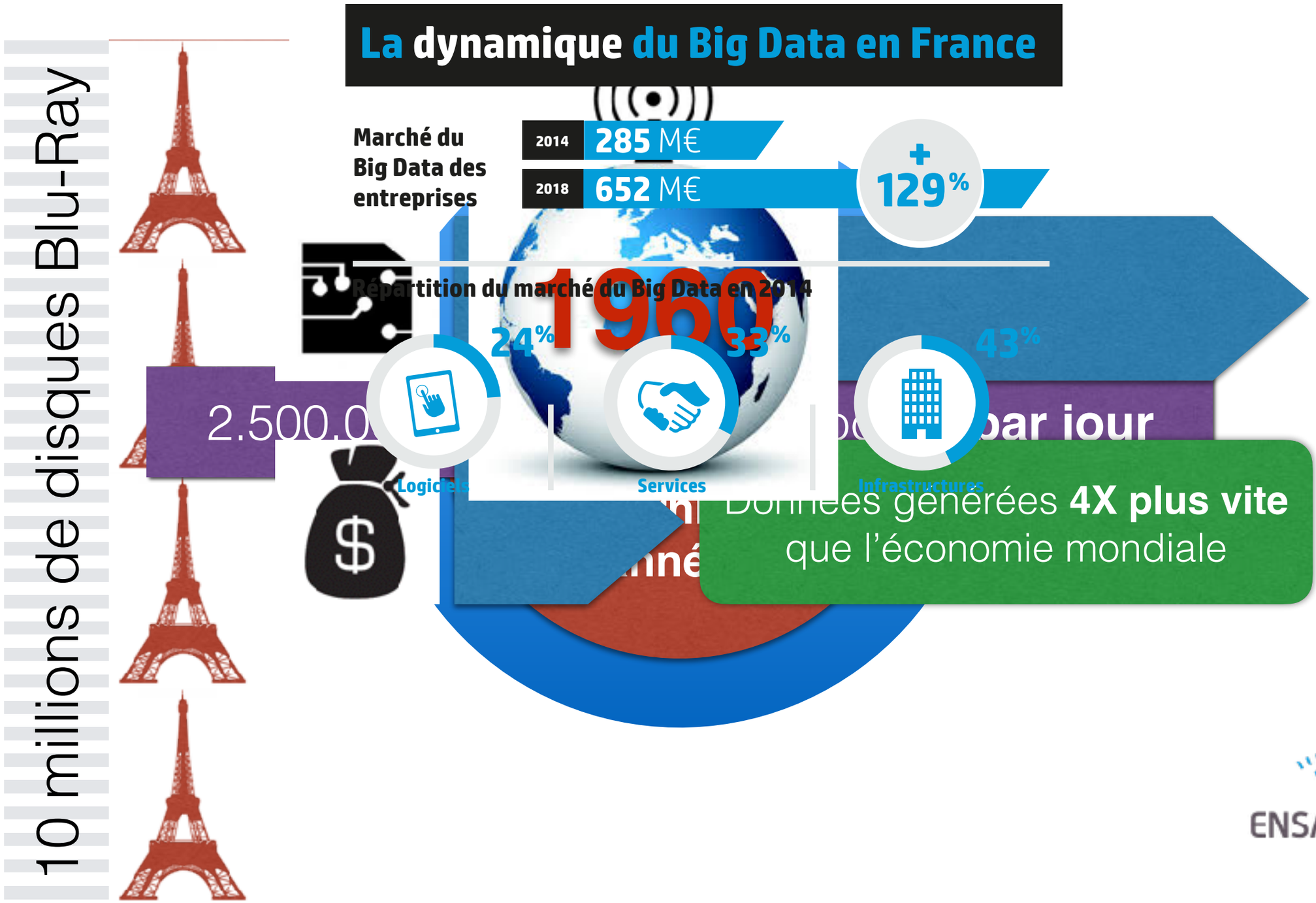
L'escalade de la puissance



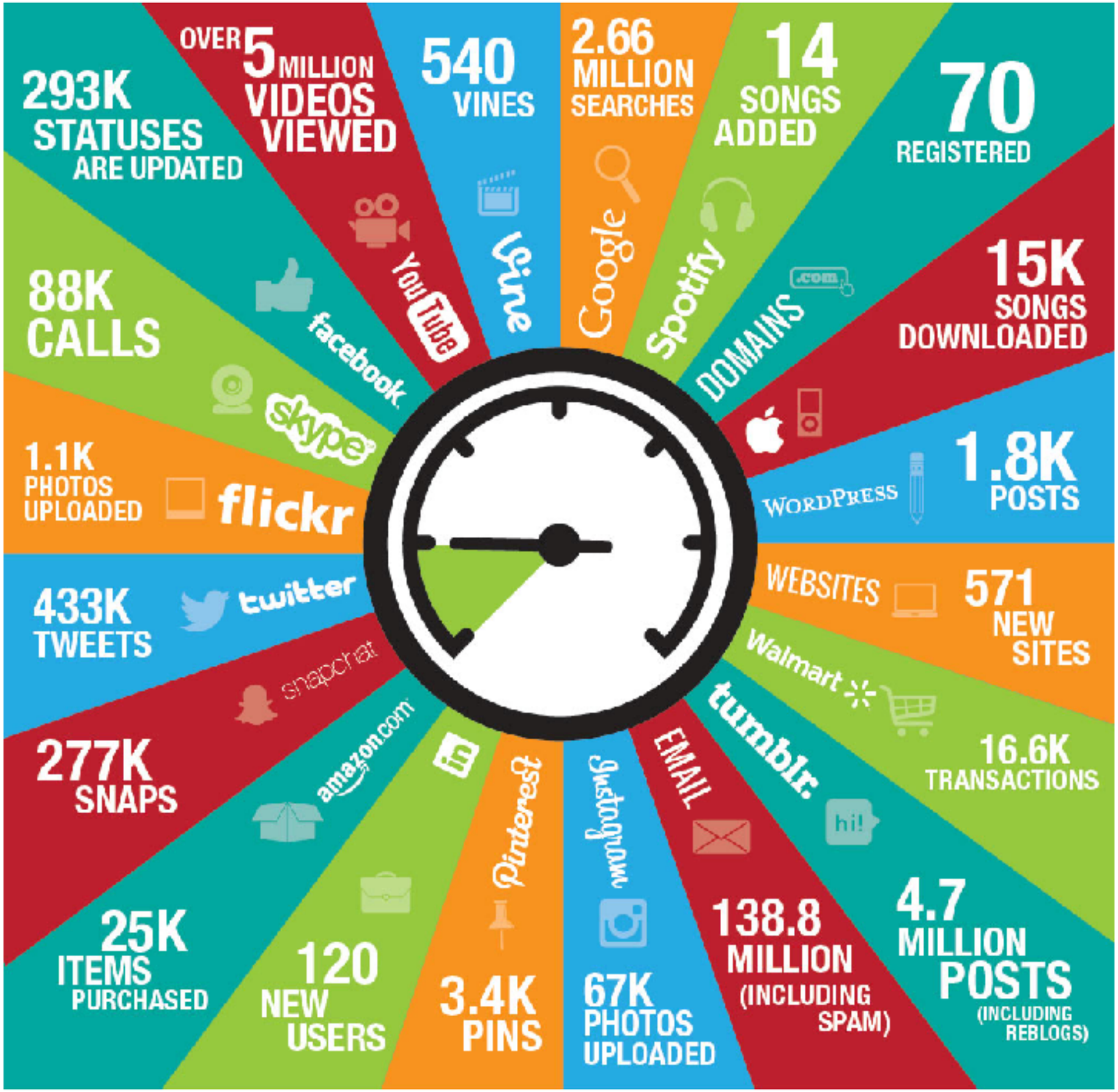
- **Avant 1950 :** la statistique
  - quelques centaines d'individus et quelques variables
  - protocole strict de laboratoire pour une étude scientifique
- **Années 1960-1980** : l'analyse des données
  - quelques dizaines de milliers d'individus et quelques dizaines de variables
  - recueillies de façon rigoureuse pour une enquête précise
- **Années 1990-2000** : le data mining
  - plusieurs millions d'individus et plusieurs centaines de variables
  - recueillies dans le système d'information des entreprises pour de l'aide à la décision
- **À partir des années 2010** : le Big Data
  - plusieurs centaines de millions d'individus et plusieurs milliers de variables
  - de tous types, recueillies dans les entreprises, les systèmes, Internet, pour de l'aide à la décision, de nouveaux services

ENSAI

Quelques chiffres pour se convaincre

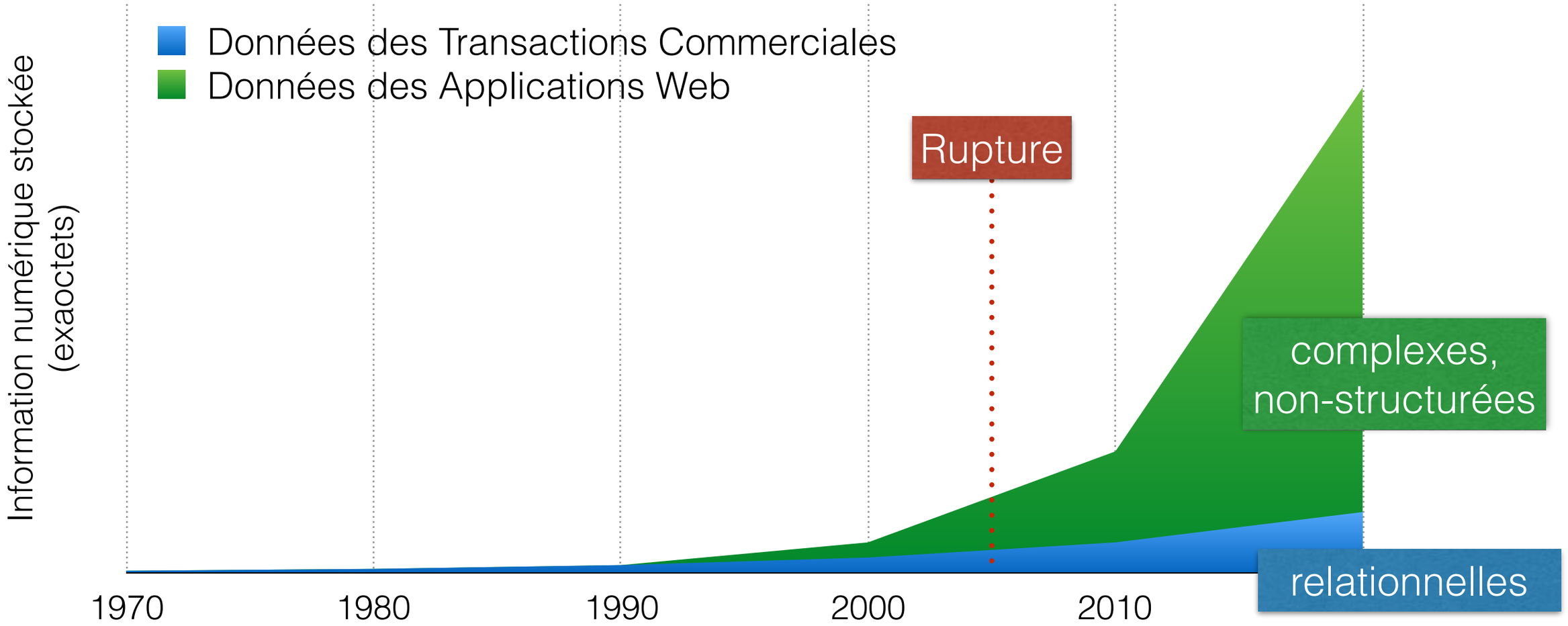


60 secondes en ligne



ENSAI

Variété des données



Il y a 10 ans : explosion du volume des données générées
Les leaders du Web ont du trouver des technologies innovantes

ENSAI

Le Big Data : Buzz ou sujet de fond ?

- Depuis environ 5 ans, le **Big Data** suscite un **intérêt grandissant** auprès des acteurs des systèmes d'information
  - Éditeurs, DSI, intégrateurs, ...
- Concept des **4V**
  - Apporte des **axes de réflexion**
  - Mais : pas de définition claire

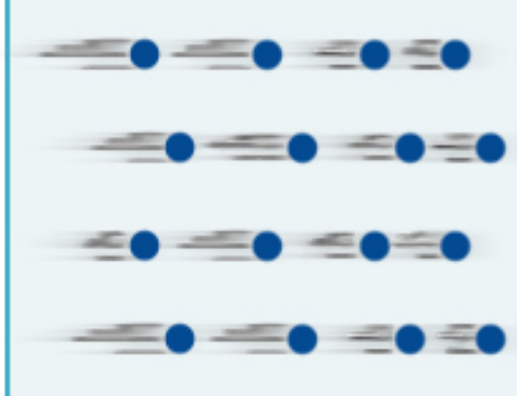
Dimensions du Big Data

Volume

Données brutes d'origine

Terabytes à exabytes de données disponibles

Vitesse



Données dynamiques

Analyse en temps réel, décision en une fraction de seconde

Variété

Données hétérogènes

Format structuré, non structuré, texte, multimédia

Véracité

Données incertaines

Cohérence, fiabilité, qualité et prédictibilité des données

ENSAI

© T. Lombry

« Up to eleven »

Variabilité

Variété

Visibilité

Volume

Validité

Valeur

Venue

Velocité

Véracité

Vocabulaire

Vague

Métrologie des réseaux : Quelques faits

- Plus de 90% des emails sont du spam
- Des milliers de bugs logiciels découvert par an
  - Environ 3 jours pour qu'un virus exploite la faille
  - 30 jours avant la mise à disposition d'un patch
- Environ 8 000 attaques par DoS par jour
- Plus de 150 000 attaques par phishing par an

La surveillance réseaux

- Couramment appelée par son équivalent anglais : Network Monitoring
- Permet de détecter les anomalies de fonctionnement
  - Extraction de données pertinentes
- Utilité : métrologie, dimensionnement dynamique, détection de DoS, ...

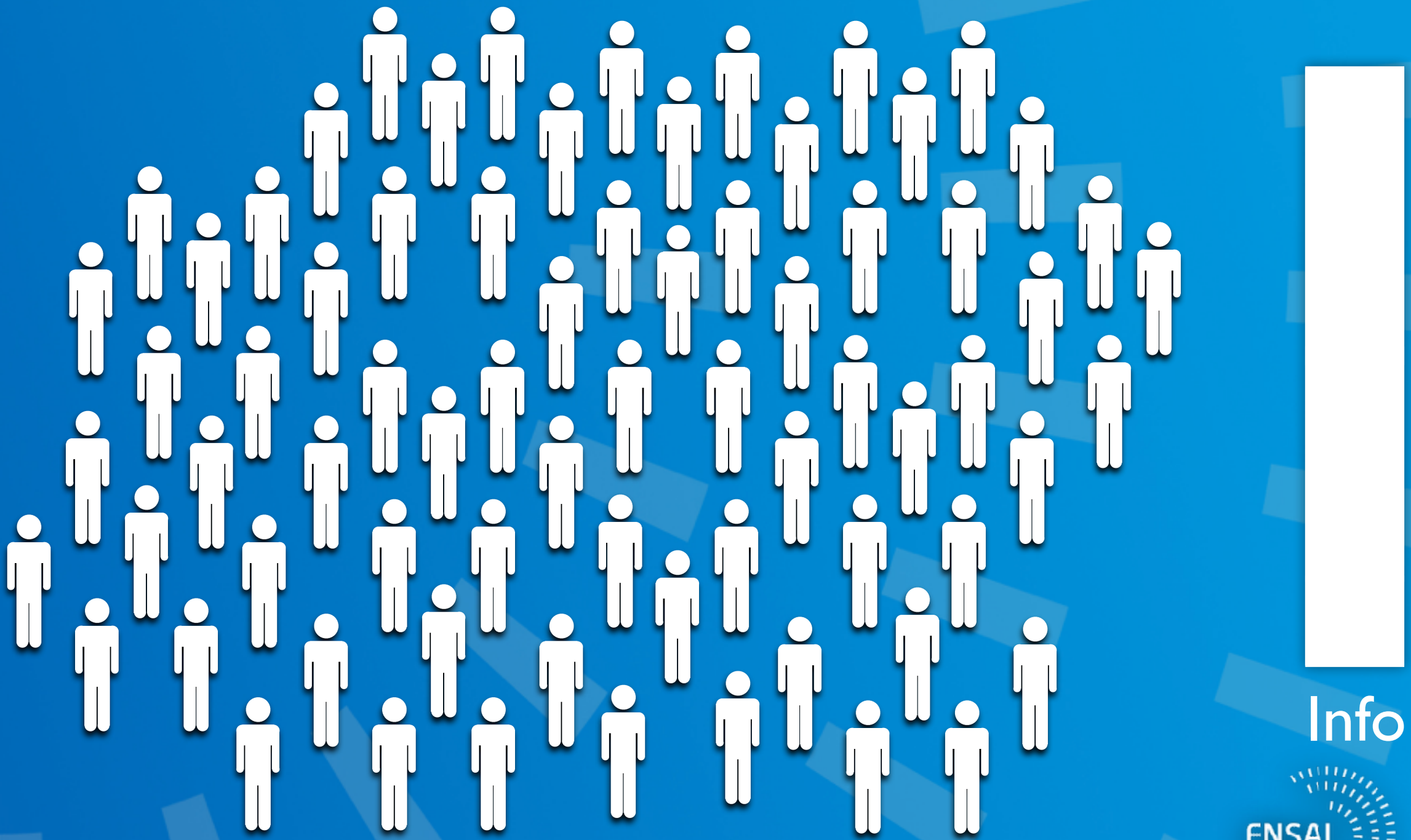
Surveillance par historique

- Fondé sur des résumés d'informations sur une période de temps
  - Essentiel pour comprendre et améliorer la performance
  - Indique le besoin de mise à jour / redimensionnement
  - Justifie les dépenses nécessaires

Surveillance en temps-réel

- Observe continuellement la situation courante (au pire, récente)
  - Utilisé pour comprendre les problèmes/bugs en cours
  - Permet de garantir une réponse rapide à un évènement (bug, attaque, ...)
- Plus consommateur de ressources
- Difficile à mettre en oeuvre

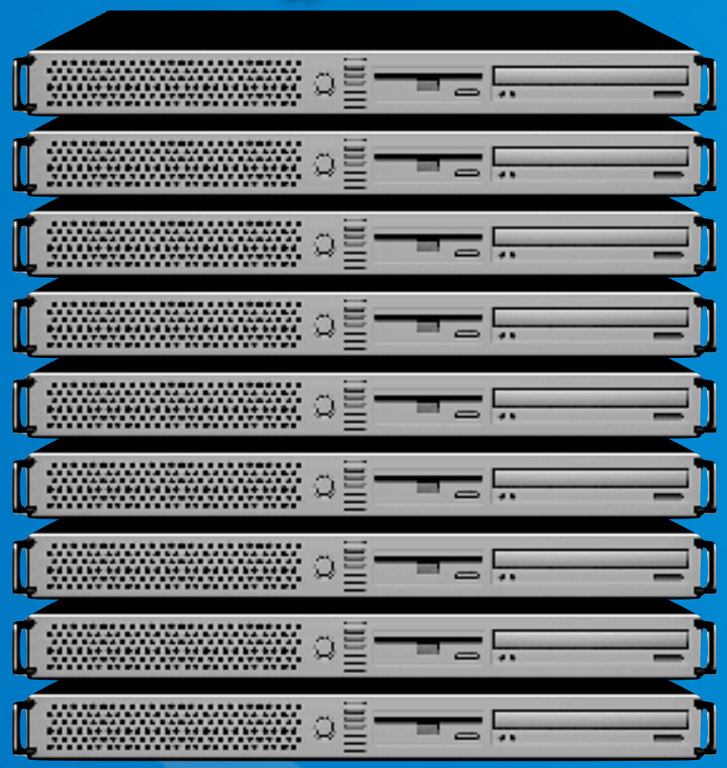
Systèmes d'information répartis



Info

ENSAI

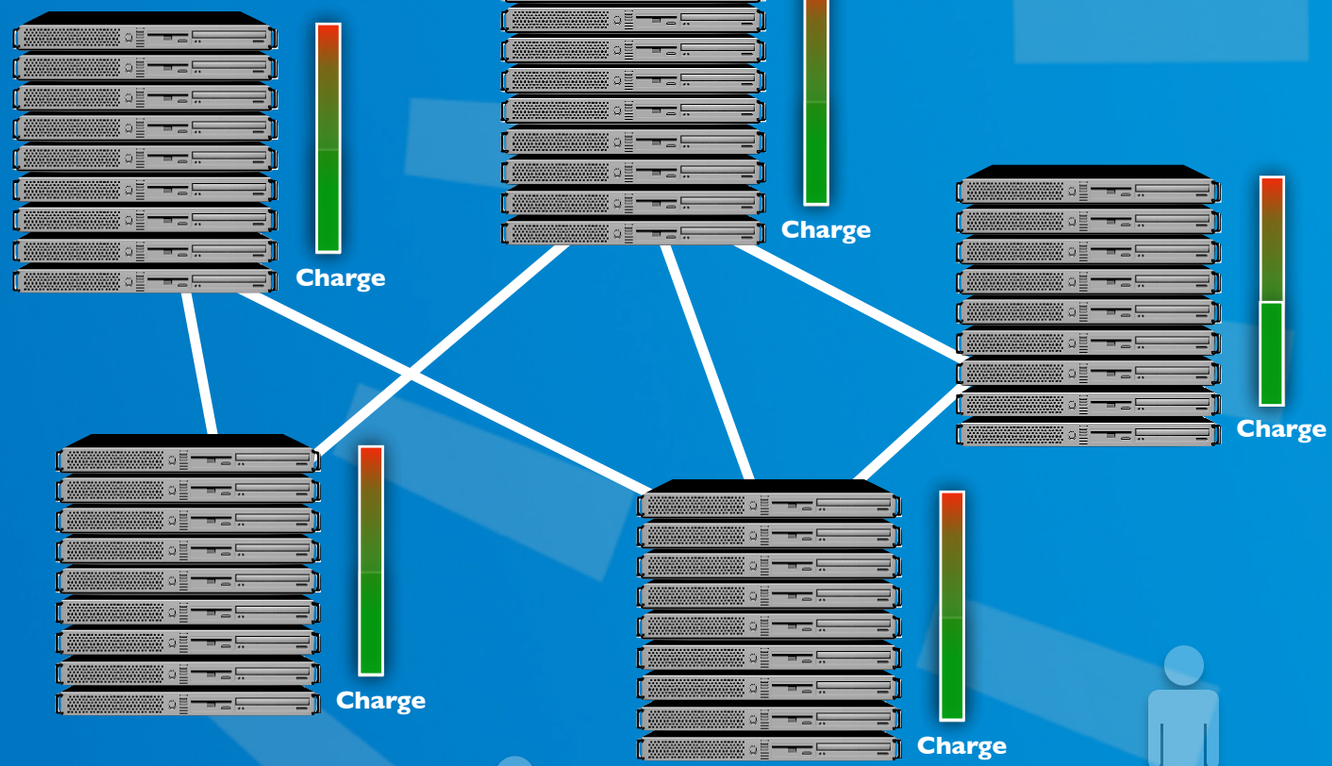
Systèmes d'information répartis



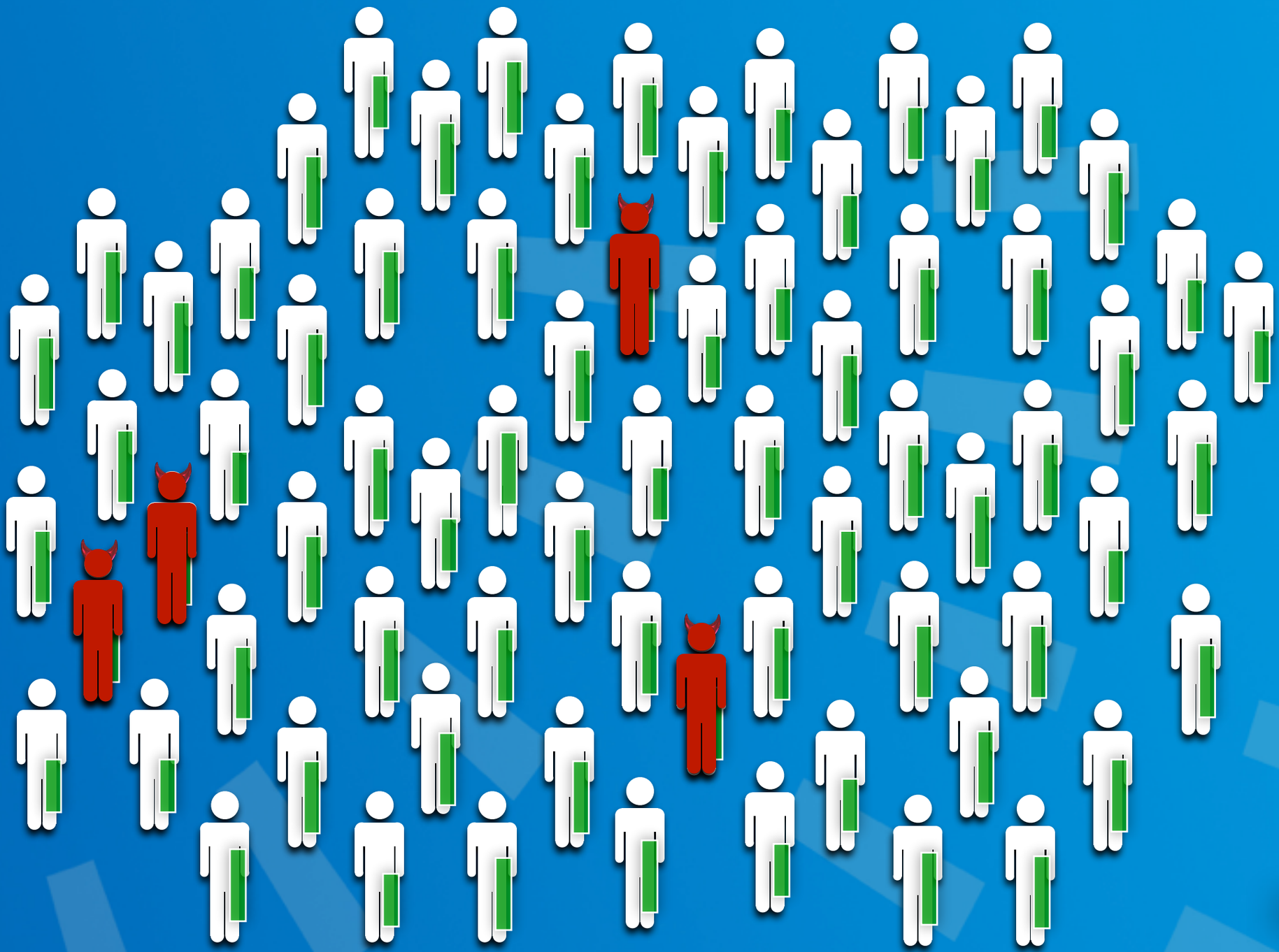
Charge

ENSAI

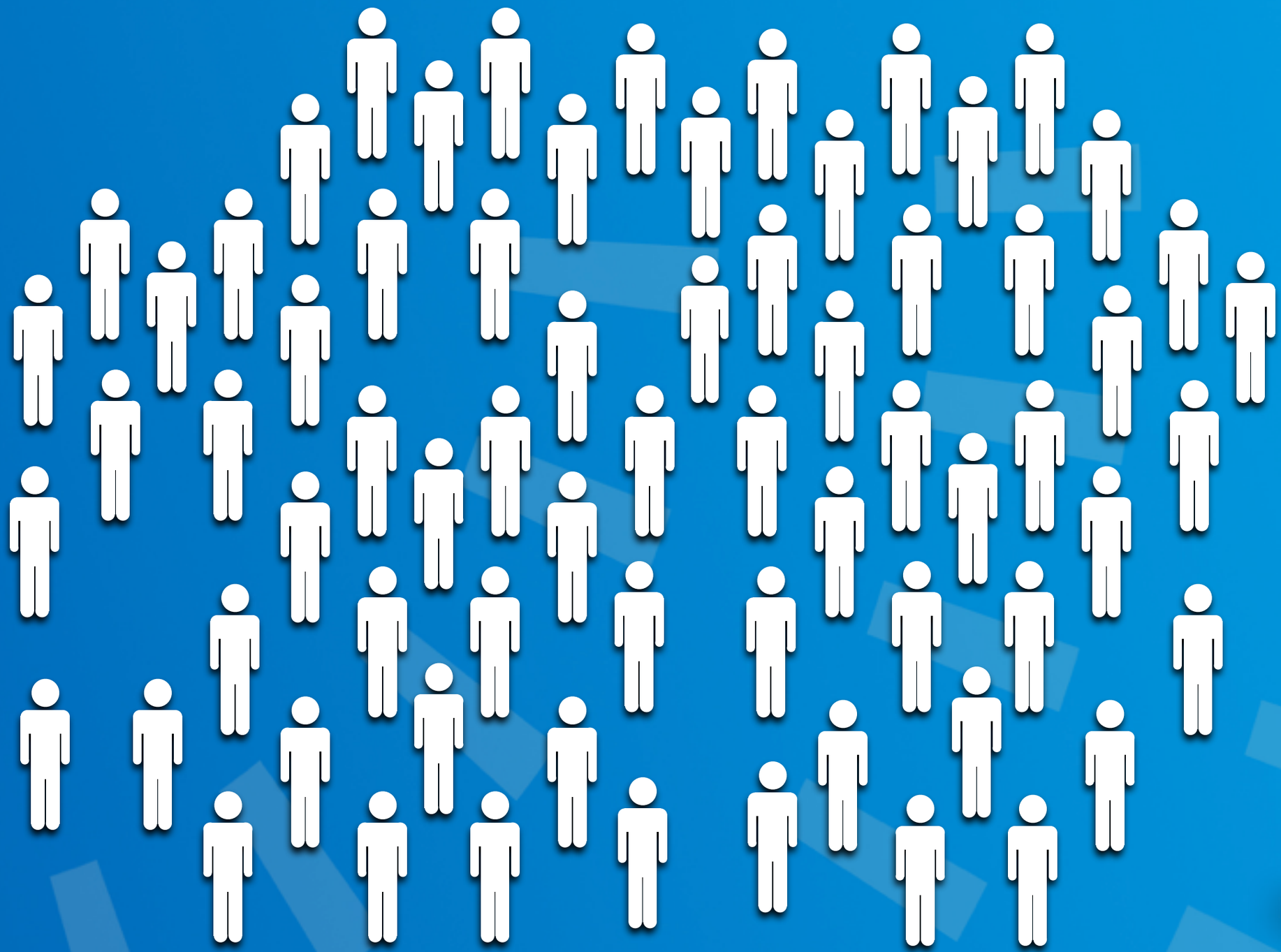
Systèmes d'information répartis



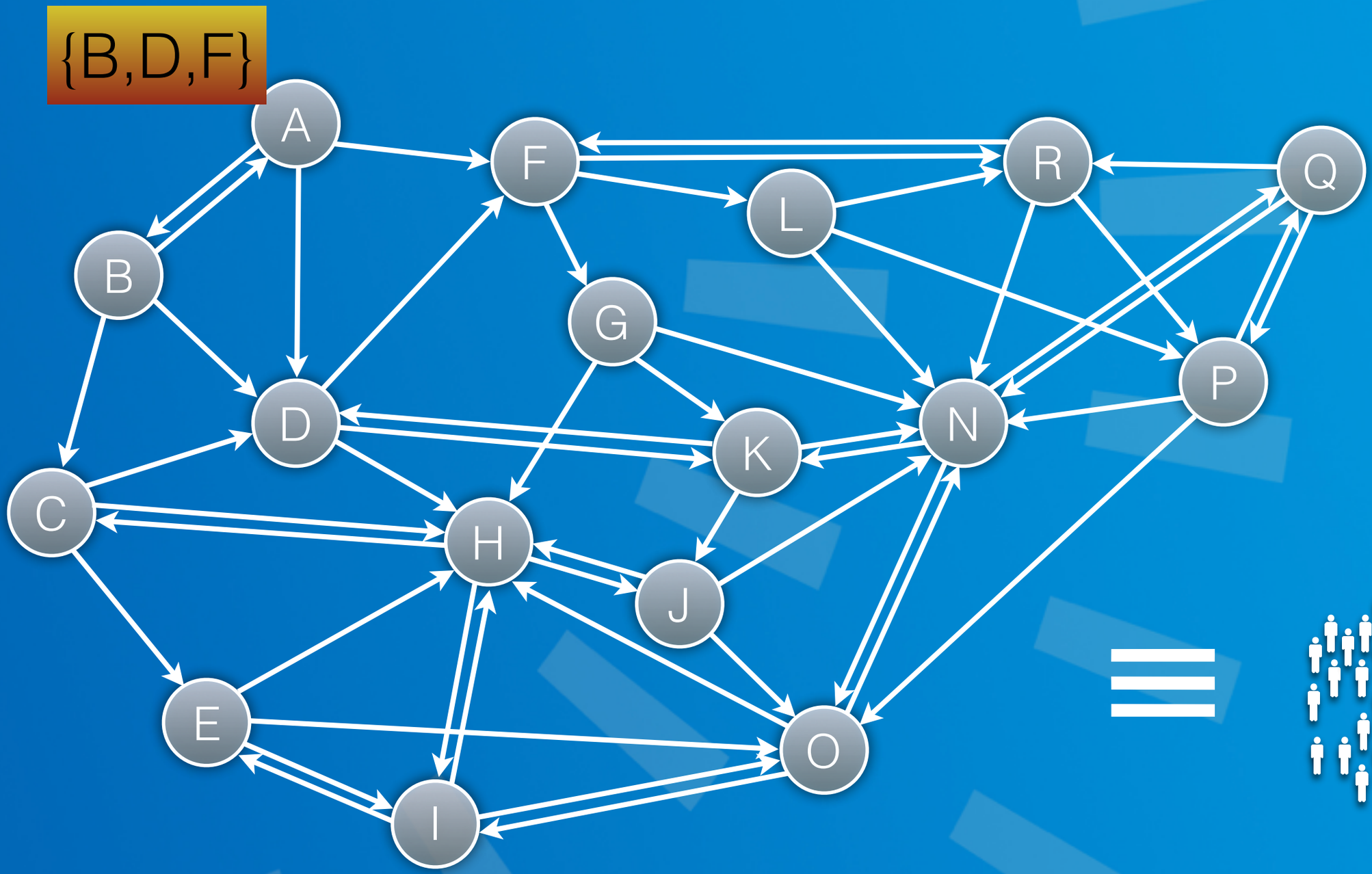
Systèmes d'information répartis



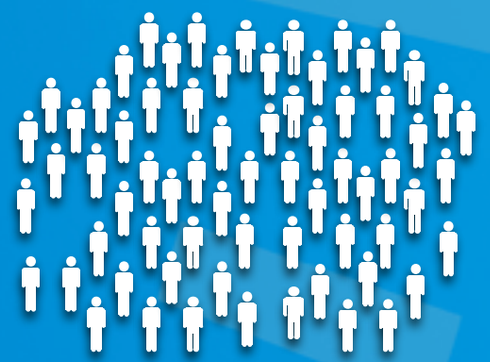
Systèmes d’information répartis



Modélisation



≡



ENSAI

Décentralisation

- Avantage
  - Passage à l'échelle inhérent
  - Capacité d'auto-organisation
  - Tolérance aux défaillances
- Inconvénient
  - Obtention de statistiques globales difficile
  - Surveillance du système extrêmement complexe

Surveillance en temps-réel :

Besoin d’échantillonnage

Temps réel

- Objectif : obtenir des échantillons uniformes récents et en continu
- Service d'échantillonnage uniforme continu :

Obtenir un flux d'échantillon uniforme

- Abstrait par une fonction `getNode()` qui retourne l'adresse d'un noeud présent dans le système

ENSAI

Propriétés

- Uniformité
  - Garantie que chaque nœud possède la même probabilité d'être sélectionné comme échantillon
- Fraicheur
  - Adaptation continue de la population du système
- **Efficacité**
  - Faible quantité mémoire nécessaire
  - Temps de réponse faible

Modèle de flux de données

- Le flux de données est issu d'échantillons obtenus
  - via marche aléatoire, protocole épidémique, observation des routeurs, ...
- Impossible de tout mémoriser tous les identifiants
- Non résistant à des adversaires malveillants

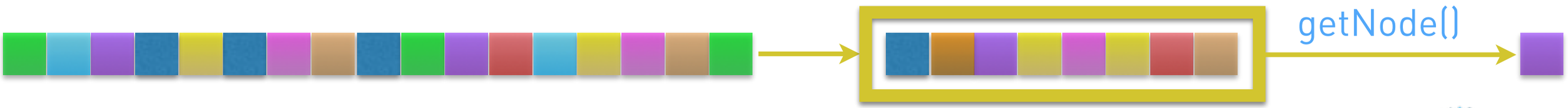
Un peu de formalisme

- Un flux est une sequence d'items $\sigma = < a_1, a_2, \ldots >$
- Chaque a_i est tiré d'un très grand univers $U = \{1, 2, \ldots, N\}$
  - *e.g.* IPv6 addresses: $N = 2^{128}$
- Chaque item peut être reçu à plusieurs reprises dans σ
- Un flux défini implicitement un vecteur de fréquences $(x_1, x_2, \ldots)$ où x_i est le nombre d'occurrences (ou fréquence) de l'item *i* reçu jusqu'à présente dans σ
- Le vecteur de fréquences peut-être vu comme la distribution de probabilité empirique de σ

$$\mathbf{p} = (p_i)_{i \in U} \quad \text{où} \quad p_i = \frac{x_i}{\sum_{j \in U} x_j}$$

Modèles d'adversaire

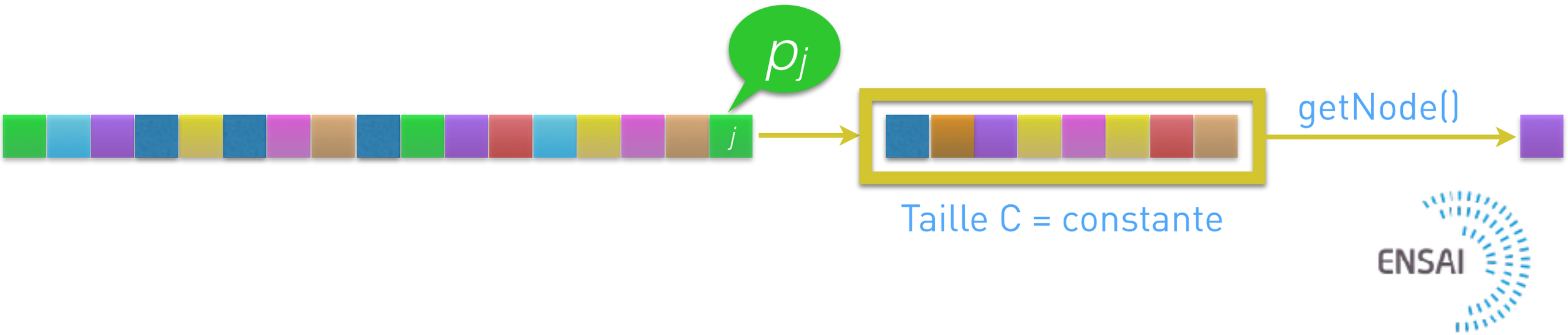
- Capable de manipuler un grand nombre de nœuds
- Adapte ses attaques à la stratégie d'échantillonnage
- Objectif de l'adversaire :
  biaiser les échantillons des nœuds corrects



ENSAI

Echantillonnage uniforme : Version omnisciente

- Hypothèse : Pour chaque item *j* reçu, sa probabilité d'occurrence p_j est connue
- Stratégie omnisciente
  - Si p_j est tout petit : on stocke l'item en mémoire
  - Sinon, on l'ignore la plupart du temps



Echantillonnage uniforme : Version omnisciente

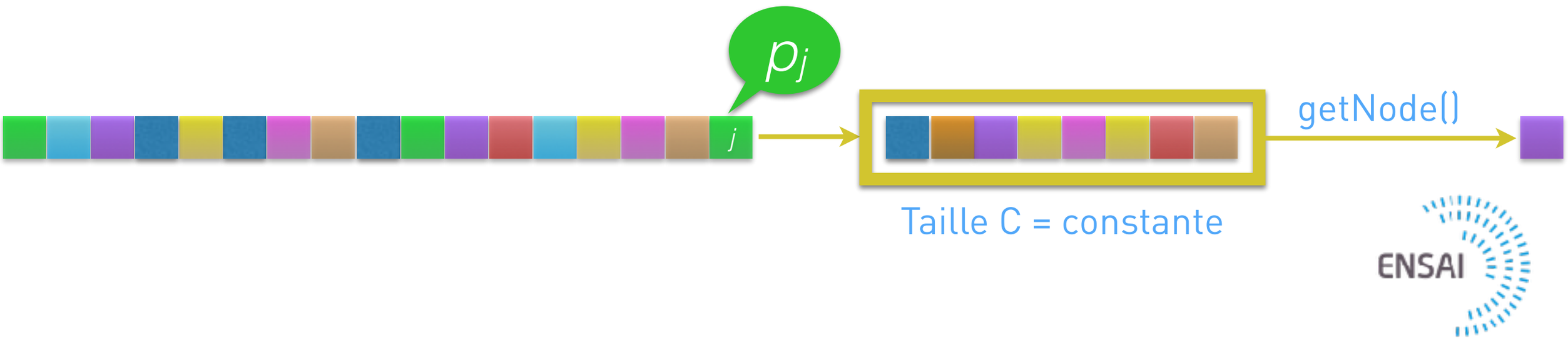
- Algorithme en détail

On insère j dans la mémoire avec une probabilité

$$a_j = \min_{i \in [N]} p_i / p_j$$

Si *j* est inséré, on retire un item *u* de la mémoire selon

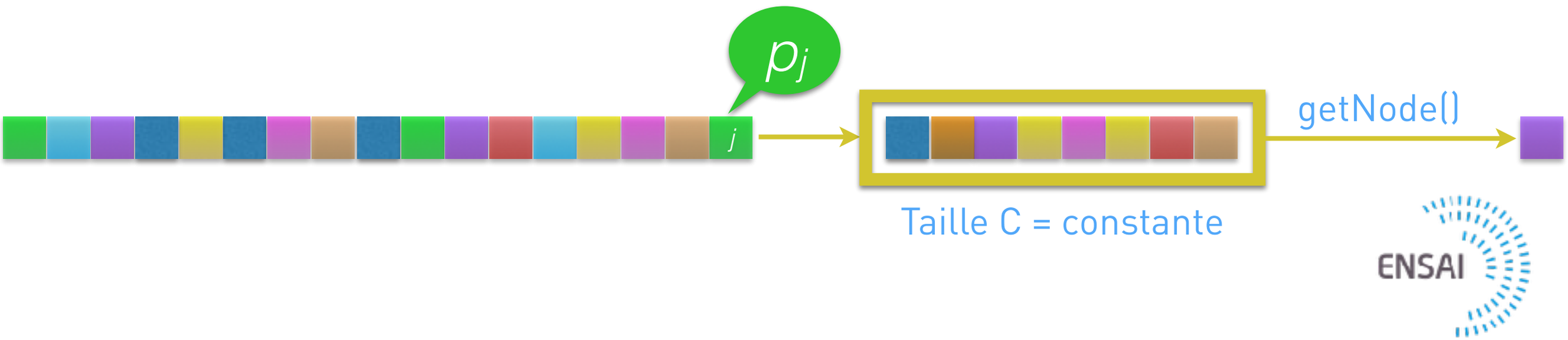
$$r_u = 1 / C$$



ENSAI

Echantillonnage uniforme : Version omnisciente

- Analyse par chaîne de Markov
- Théorème :
  L'algorithme omniscient garantie l'uniformité et la fraicheur en régime stationnaire



Echantillonnage uniforme : Version à l'aveugle

- Hypothèse très (trop) forte pour être mise en pratique
  - Impossible de prédire la distribution des items
- Nécessité d'évaluer p_i en ligne, pour tout i
  - Estimateur de fréquence : Count-Min Sketch

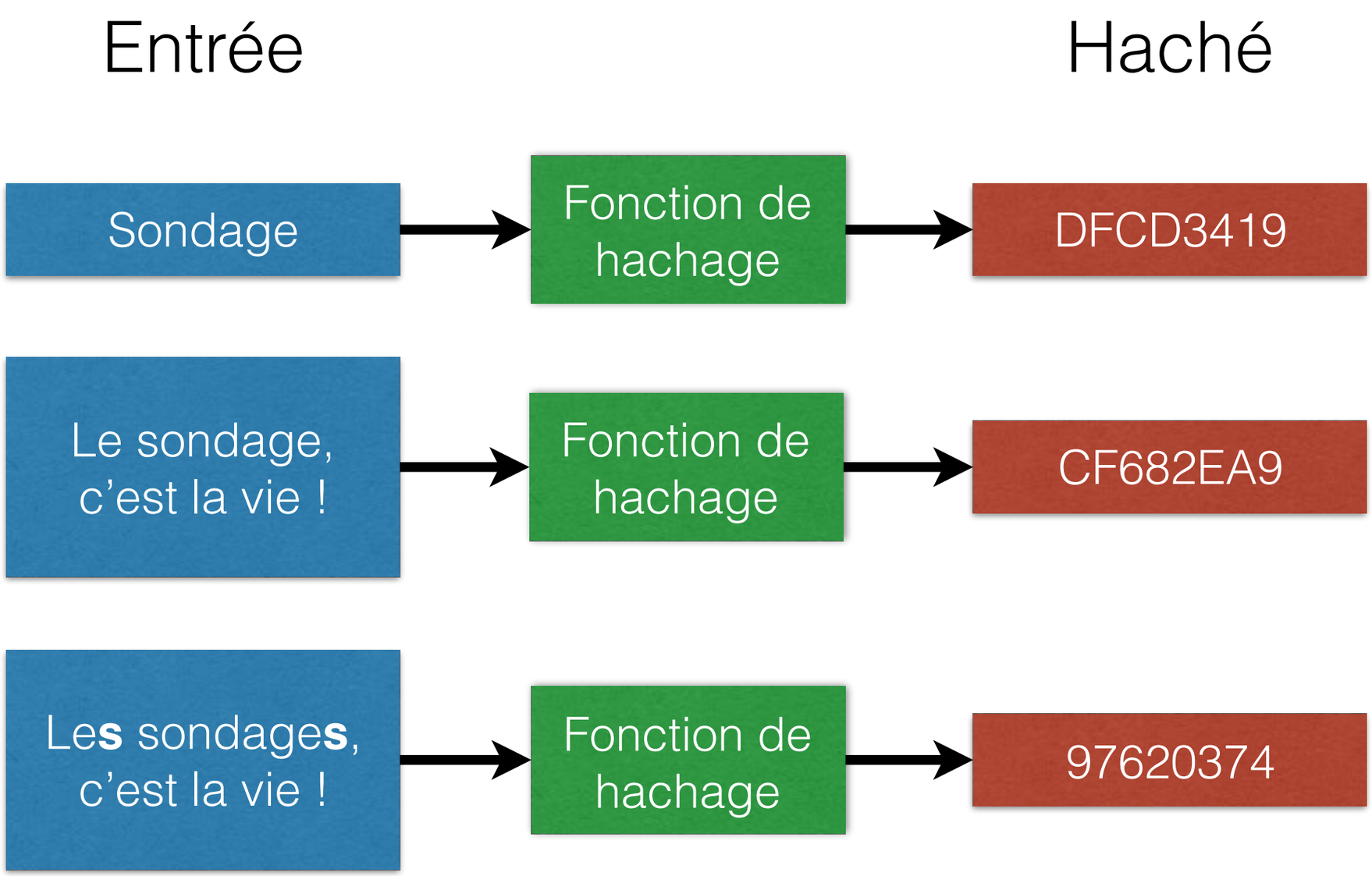
Count-Min Sketch [CM05]

- Estimation des fréquences de tout item
  - (ε,δ)-approximation-additive de x_i
- Maintien d'un vecteur *C* de taille k=2/ε
- Choix d'une fonction de hachage 2-universelle
- Pour chaque item v du flux
  - $C[h(v)]$ ++

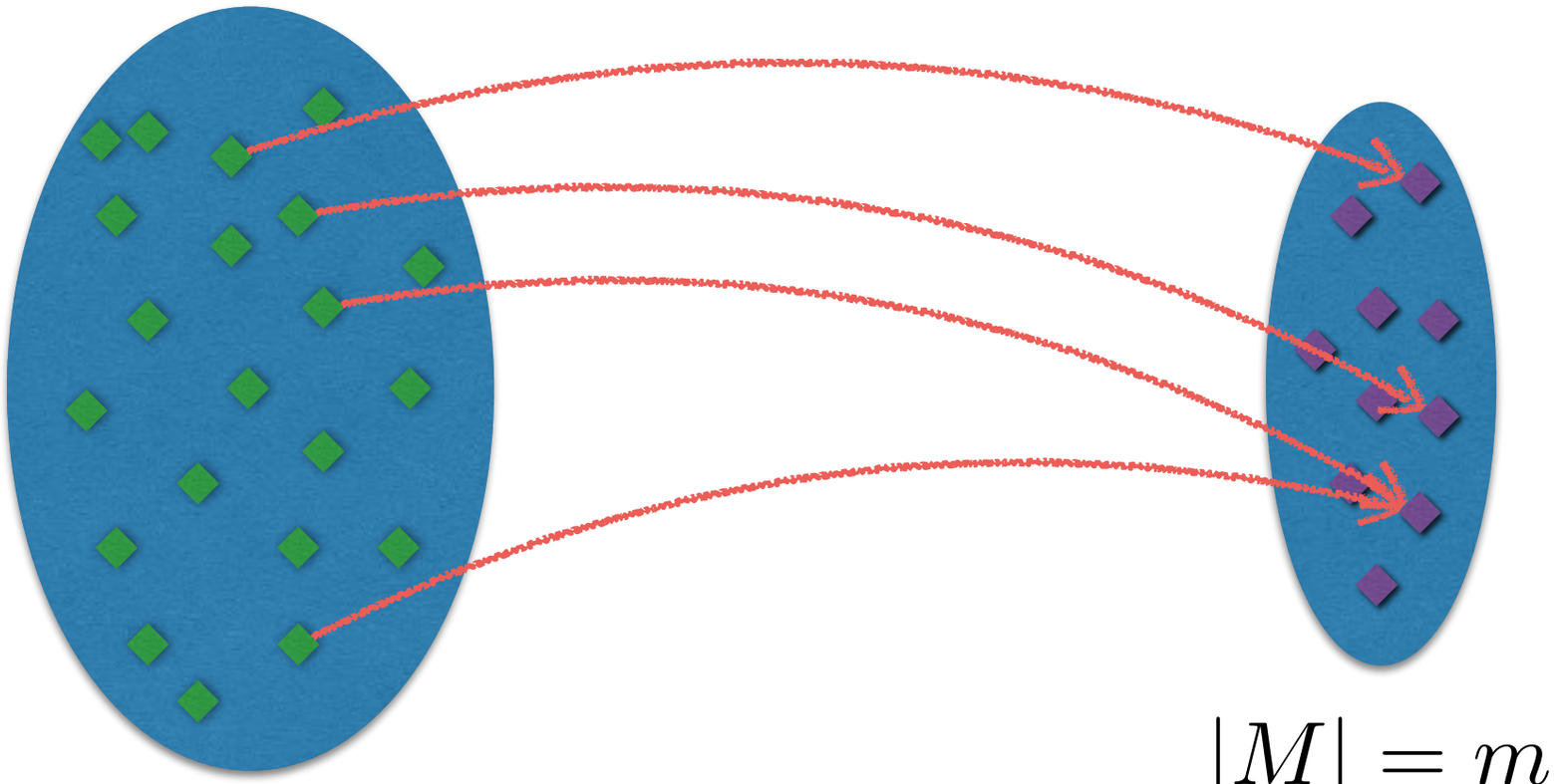
Calcul de 1/**δ** estimateurs en parallèle et retour du minimum

[CM05] G. Cormode and S. Muthukrishnan, An improved data stream summary: the count-min sketch and its applications, in Journal of Algorithms, vol. 55, no. 1, pp. 58–75, 2005.

Qu'est ce qu'une fonction de hachage ?

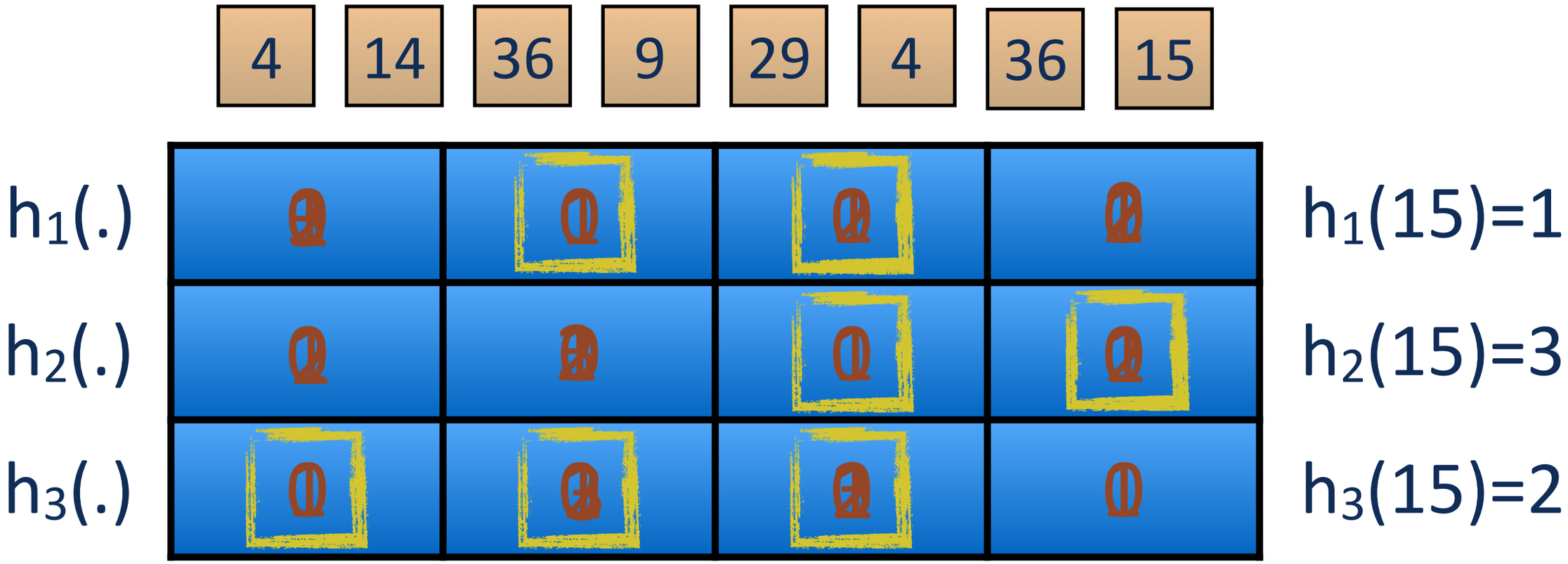


Qu’est ce qu’une fonction de hachage ?



$H = \{h : U \to [m]\}$ est universelle si $\forall x, y \in U, x \neq y : \mathbb{P}_{h \in H}\{h(x) = h(y)\} \leq \frac{1}{m}$

Exemple d'un CM sketch



4

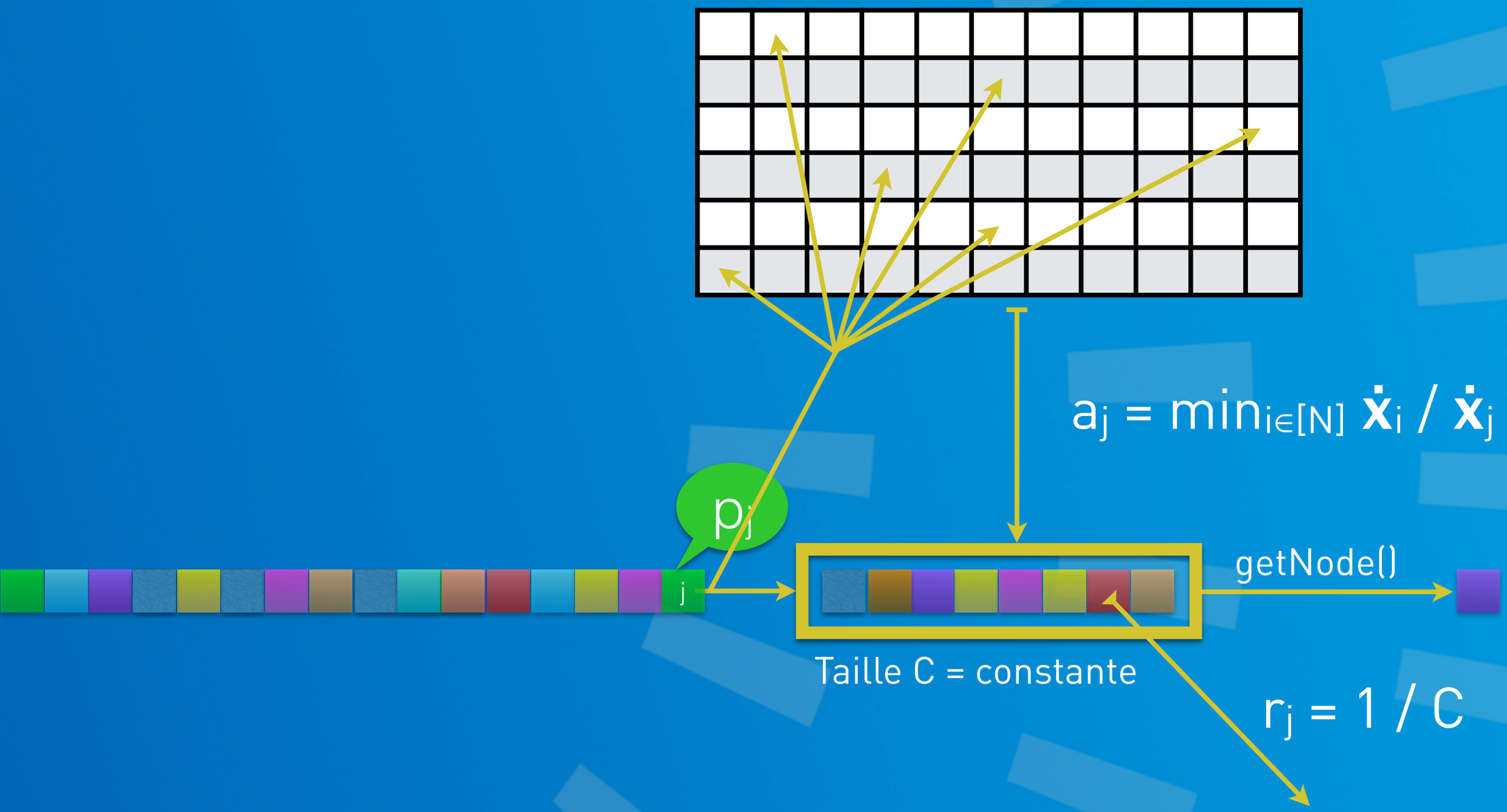
min = 1 min = 1 min = 1 min = 2 min = 1 min = 2

15 29 9 36 14 4

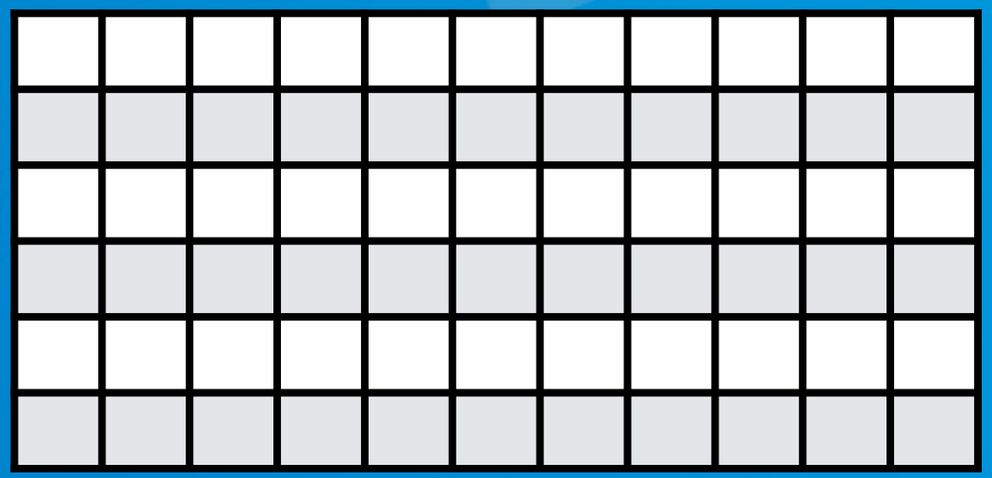
ENSAI

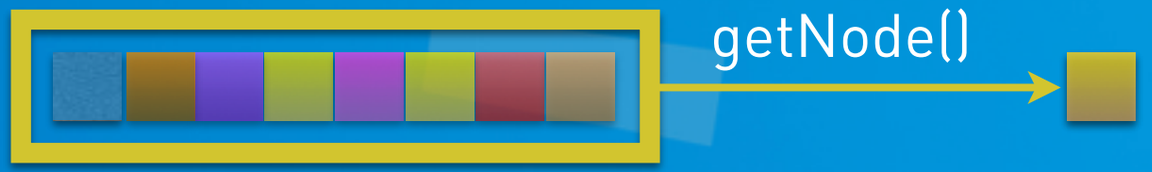
Propriétés de [CM05]

- (ε,δ)-approximation-additive du nombre d'occurrences des items du flux
  - $x_j + \varepsilon(m\text{-}x_j) \geq \dot{x}_j \geq x_j$
    - L'inégalité de droite est toujours vraie
    - L'inégalité de gauche échoue avec une probabilité au plus δ
- Complexité en espace :
  $O(1/\varepsilon \log 1/\delta \, (\log m + \log n))$



Algorithme aveugle





Taille C = constante

Algorithme aveugle

En cas d'attaque du flux

- On a vu que
  - La stratégie omnisciente génère un flux uniforme et « frais » à partir de n'importe quel flux d'entrée
  - La stratégie aveugle émule la version omnisciente par approximation des fréquences des items
- L'unique possibilité d'attaque est d'accroitre artificiellement les fréquences estimées

La stratégie aveugle en présence de collusions

- Attaque ciblée
  - L'adversaire se concentre sur un item j spécifique
  - Doit générer suffisamment d'item o_1, ..., o_l tels que
    - pour chaque ligne s de CM, il existe o_i, $h_s(o_i) = h_s(j)$
- Attaque par inondation
  - L'adversaire souhaite surestimer tous les items
  - Doit générer suffisamment d'item o_1, ..., o_l tels que
    - pour toute case v de CM, il existe o_i, $h_s(o_i) = v$

La stratégie aveugle en présence de collusions

- Attaque ciblée
  - L'adversaire se concentre sur [illegible]ue
  - Doit générer suffi[illegible]que
    - pour [illegible]er tous les items

[illegible]samment d'item o_1, ..., o_l tels que

[illegible] toute case v de CM, il existe o_i, $h_s(o_i) = v$

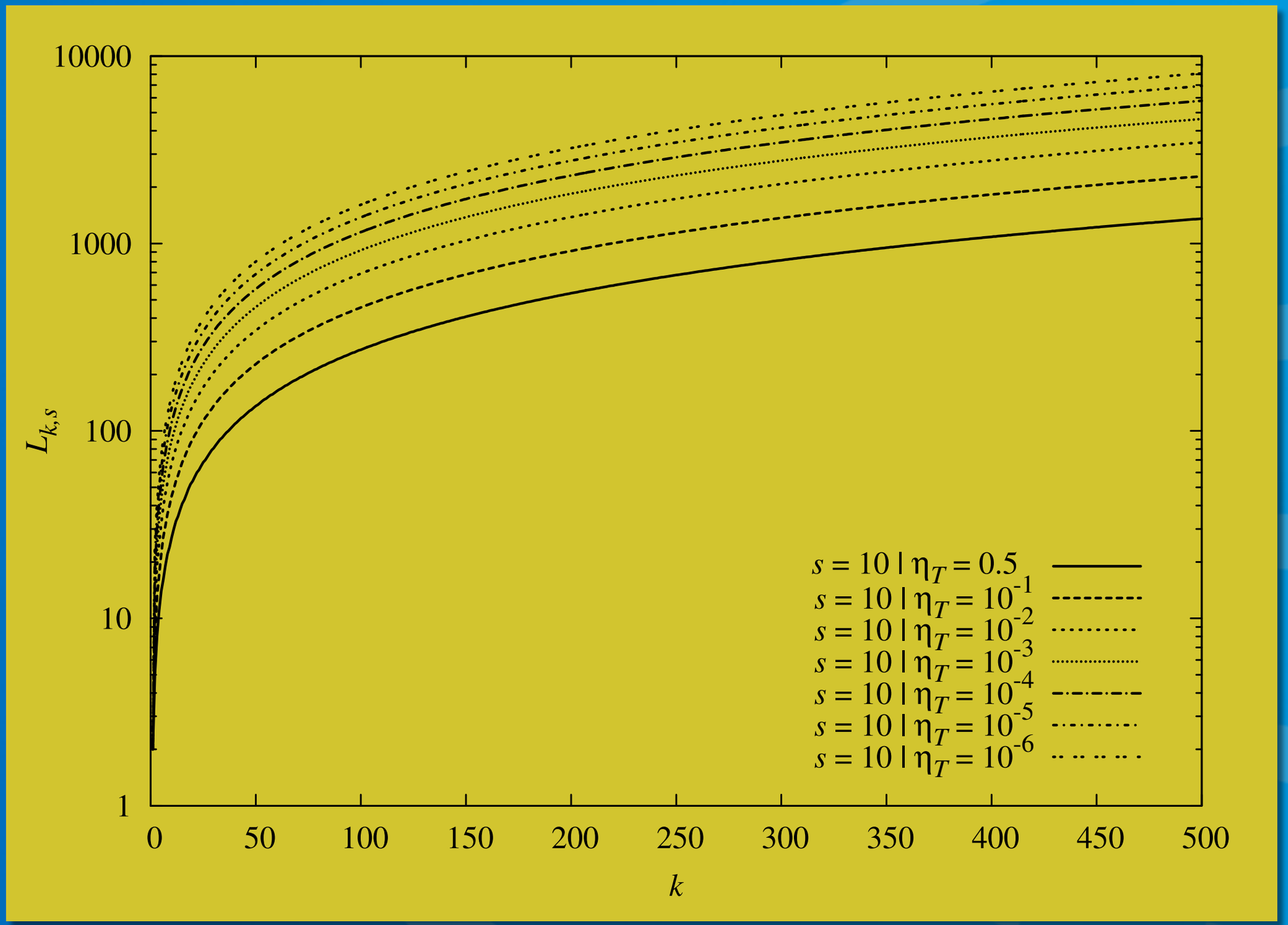
Question

Quel effort doit exercer un adversaire pour réussir ces attaques avec une probabilité 1-η ?

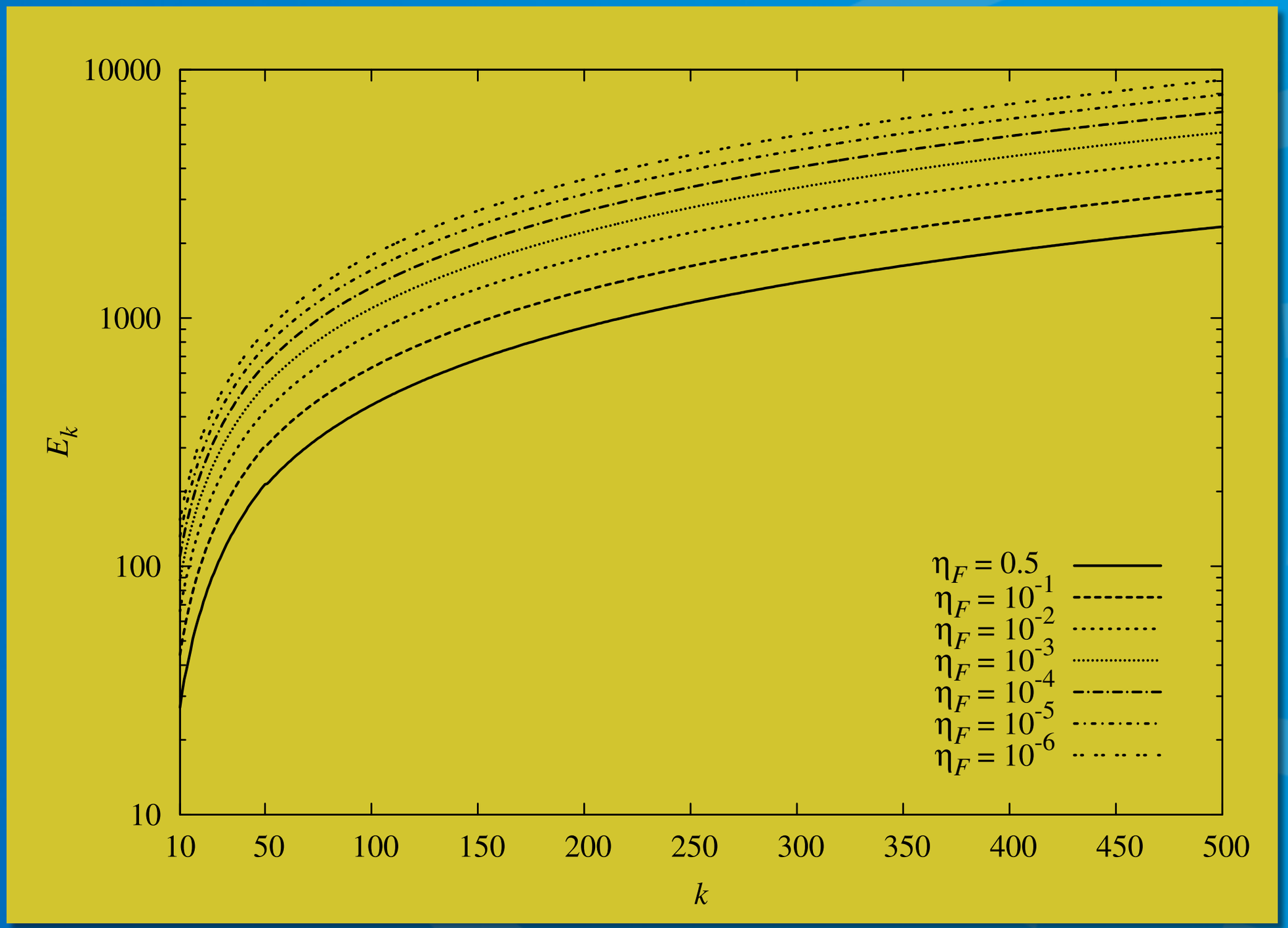
Analyse des attaques

- Modélisation par un problème d'urnes :
  - Chaque case de CM est représenté par une urne
  - Chaque item distinct est représenté par une boule

- On cherche les valeurs de
  - $L_{k,t}$: nombre de balles pour obtenir une collision avec un item j donné
  - E_k : nombre de balles pour obtenir une collision avec tous les items



Analyse des attaques



Analyse des attaques

Settings		η_T or η_F	$L_{k,t}$	E_k
k	t			
10 ($\varepsilon \sim 0.3$)	5 ($\delta \sim 10^{-2}$)	10^{-1}	38	44
10	5	10^{-4}	104	110
50 ($\varepsilon \sim 0.05$)	5	10^{-1}	193	306
50	10 ($\delta \sim 10^{-3}$)	10^{-1}	227	
50	40 ($\delta \sim 10^{-12}$)	10^{-1}	296	
50	5	10^{-4}	537	651
50	10	10^{-4}	571	
50	40	10^{-4}	640	
250 ($\varepsilon \sim 0.01$)	10	10^{-1}	1,138	1,617
250	10	10^{-4}	2,871	3,363

Analyse des attaques

ENSAI

Evaluation de performances

- Evaluation de l'impact d'identifiants malveillants largement surreprésentés
- Traces synthétiques
  - Distributions : Poissons, Pareto, Binomial
- Traces réelles
  - Serveurs http : NASA, ClarkNet, Université de Saskatchewan

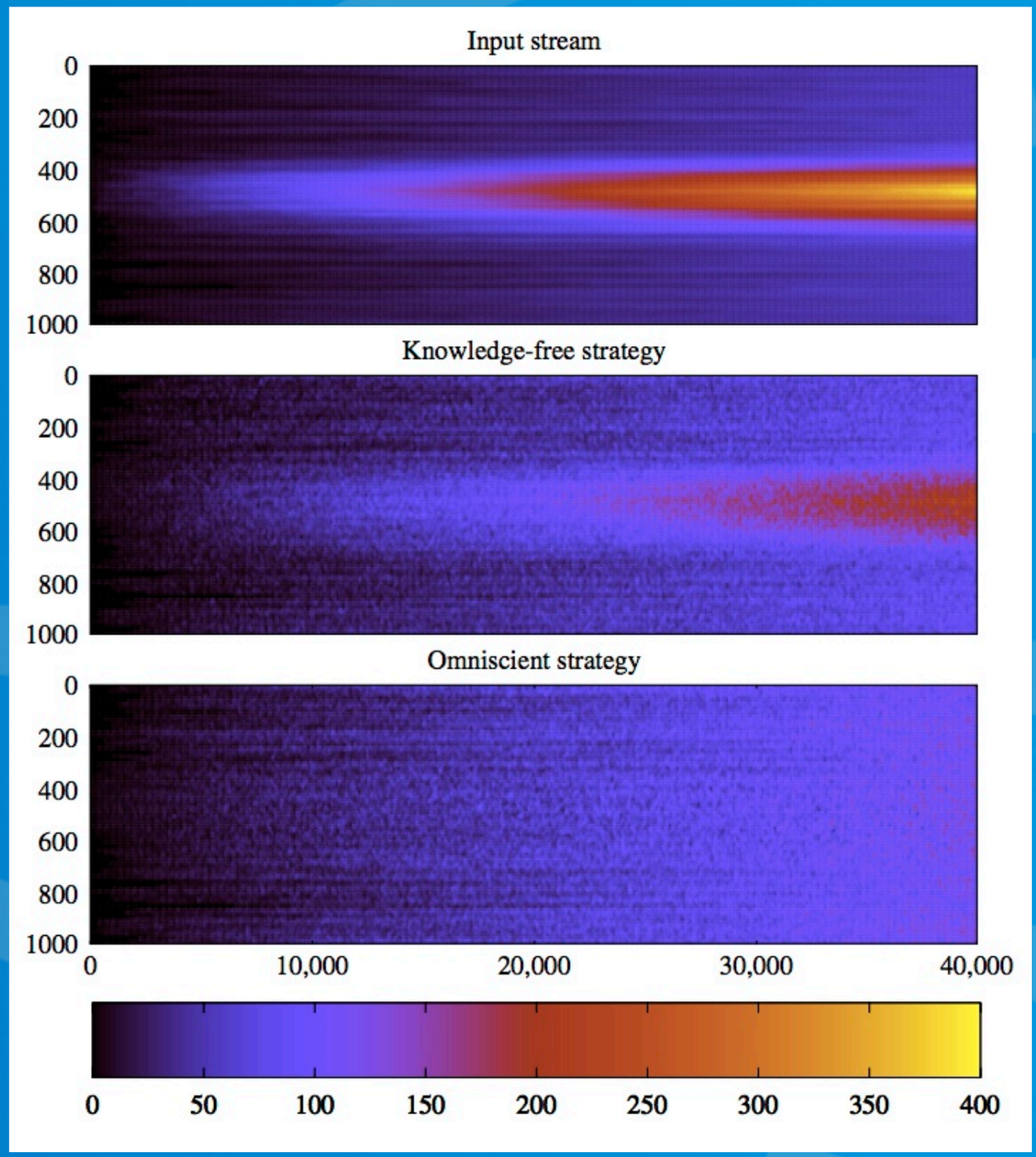
m = 40 000 items

n = 1 000 distincts

c =15

k =15

t =17



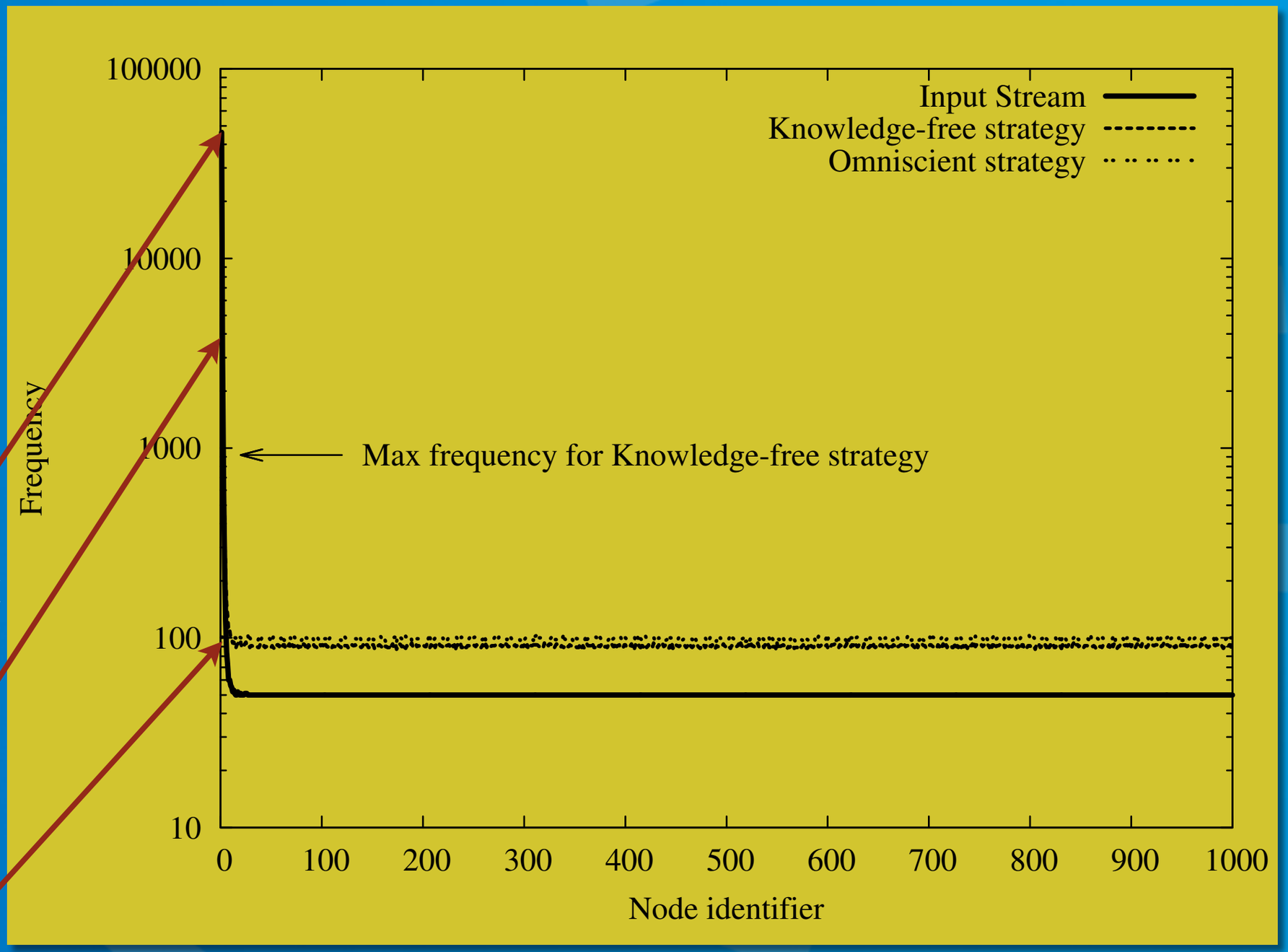
Evaluation de performances

m = 100 000 items
n = 1 000 distincts
c =10
k =10
t =5

L'adversaire injecte 50 000 fois le même item
Tous les autres présent ~50x

La stratégie aveugle réussi à réduire la plus haute fréquence par un facteur 50

La stratégie omnisciente réussi à produire un flux totalement uniforme

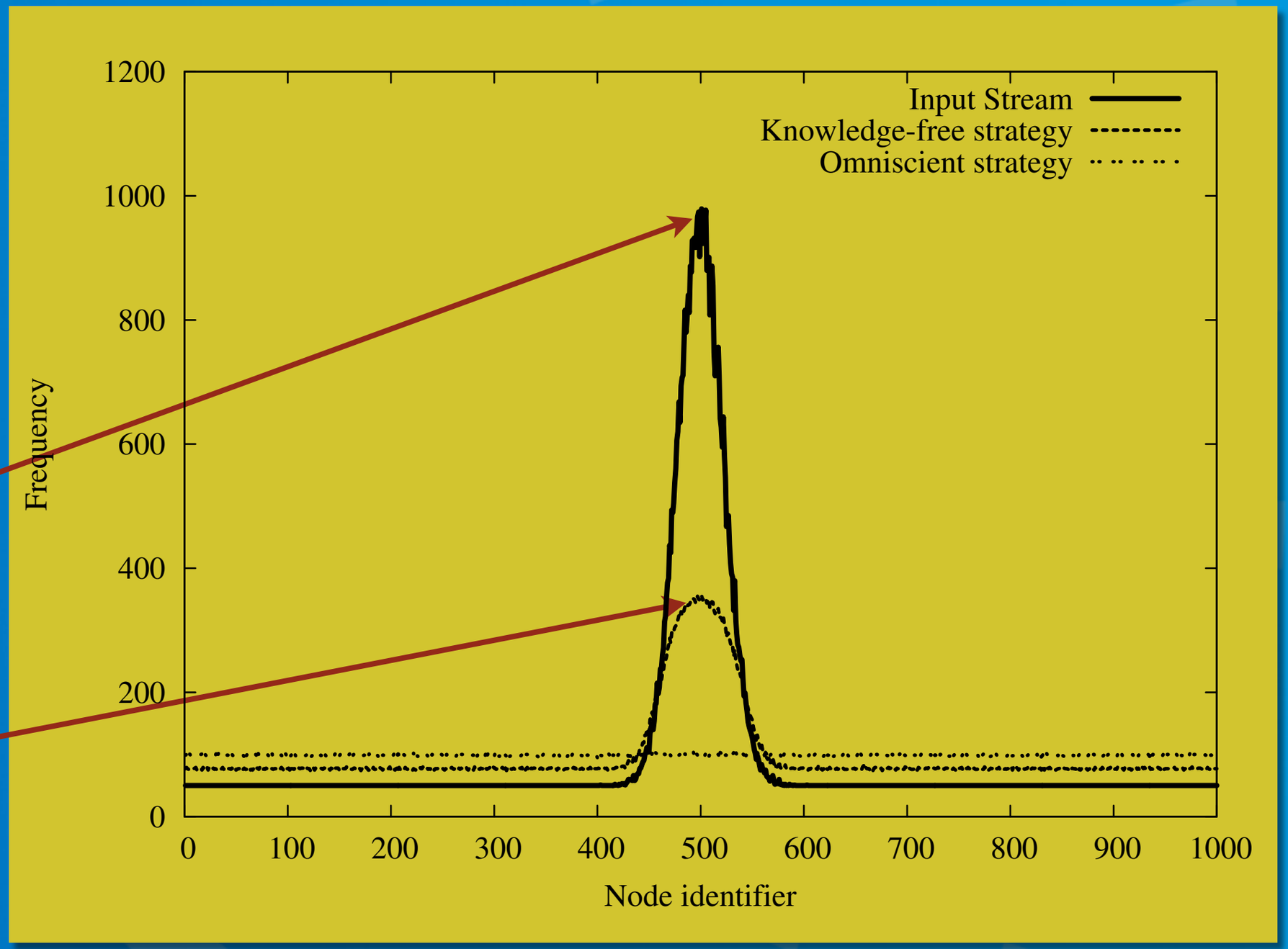


Evaluation de performances

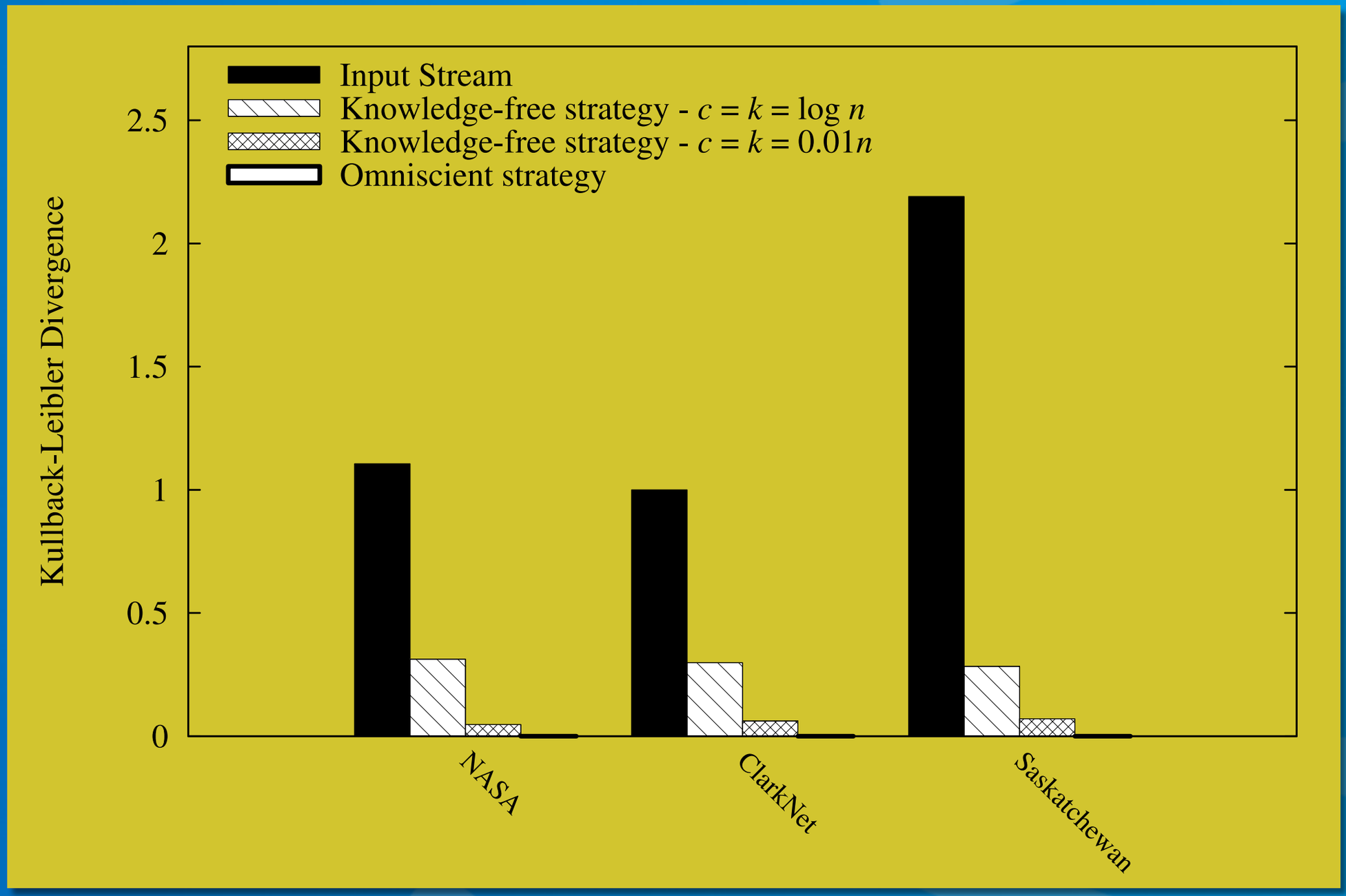
m = 100 000 items
n = 1 000 distincts
c =10
k =10
t =5

Attaque ciblée :
~50 items sur-représentés

D'après le calcul de $L_{k,t}$:
Attaque réussi avec une probabilité de 0.9 si L = 48



Evaluation de performances



m ~ 2 000 000 items / n = 100 000 distincts / t =5

Evaluation de performances

ENSAI

Comment générer des échantillons uniformes sur des grandes masses de données ?

Merci de votre attention

15 Octobre 2016

Yann Busnel

Crest (Ensai) / Inria Rennes – Bretagne Atlantique

9e Colloque Francophone sur les Sondages
Gatineau, Canada