Compter en lançant des pièces

(des tanks allemands à l'algorithme CVM)

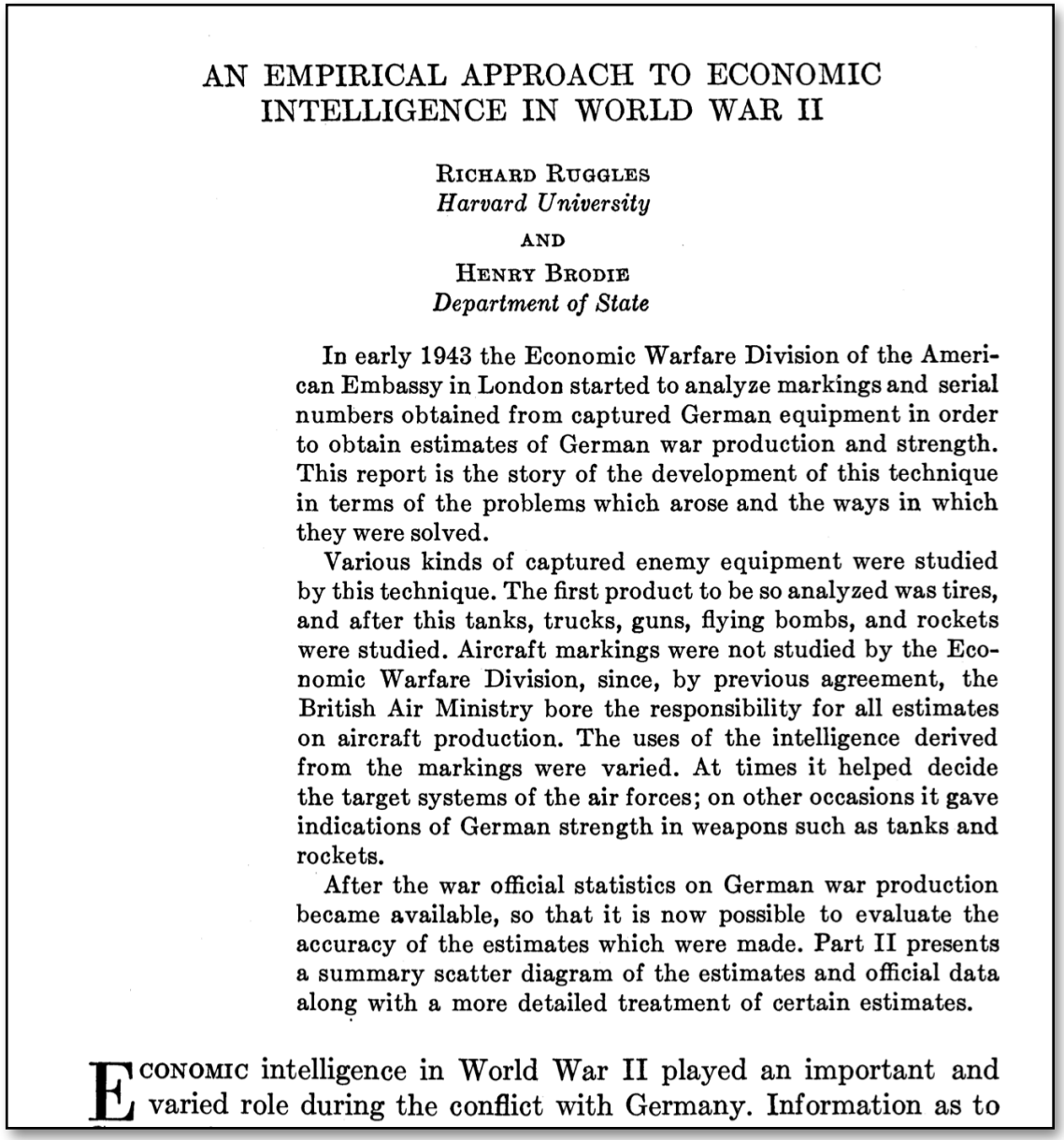
Lucas Gerin (École Polytechnique)

(**1947**) Le début de l'histoire…

The German tank problem

Article scientifique :
R.Ruggles, & H.Brodie. An empirical approach to economic intelligence in World War II. *Journal of the American Statistical Association (1947).*

- Travaux de la *Economic Warfare Division of the American Embassy*
- **But :** Estimer la production d'engins allemands (avions, chars, …)



AN EMPIRICAL APPROACH TO ECONOMIC INTELLIGENCE IN WORLD WAR II

RICHARD RUGGLES
Harvard University
AND
HENRY BRODIE
Department of State

In early 1943 the Economic Warfare Division of the American Embassy in London started to analyze markings and serial numbers obtained from captured German equipment in order to obtain estimates of German war production and strength. This report is the story of the development of this technique in terms of the problems which arose and the ways in which they were solved.

Various kinds of captured enemy equipment were studied by this technique. The first product to be so analyzed was tires, and after this tanks, trucks, guns, flying bombs, and rockets were studied. Aircraft markings were not studied by the Economic Warfare Division, since, by previous agreement, the British Air Ministry bore the responsibility for all estimates on aircraft production. The uses of the intelligence derived from the markings were varied. At times it helped decide the target systems of the air forces; on other occasions it gave indications of German strength in weapons such as tanks and rockets.

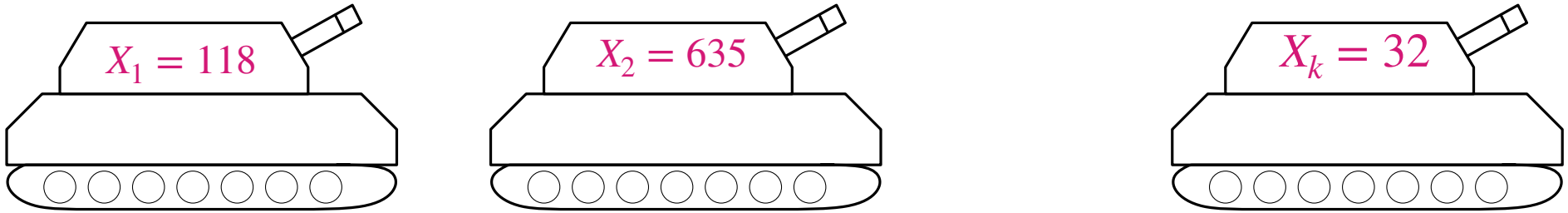
After the war official statistics on German war production became available, so that it is now possible to evaluate the accuracy of the estimates which were made. Part II presents a summary scatter diagram of the estimates and official data along with a more detailed treatment of certain estimates.

ECONOMIC intelligence in World War II played an important and varied role during the conflict with Germany. Information as to

(**1947**) Le début de l'histoire… *The German tank problem*

Entre 1940-1944, **environ 2000 numéros de chassis de tanks** répertoriés

Soient $X_1, \ldots, X_k$ les numéros de série de chassis de k tanks.
Peut-on estimer le nombre n de tanks produits?



Théorème des tanks allemands

Si les numéros sont uniformes entre *1* et *n*,

$$\mathbb{E}\left[\max_{1\leq \ell \leq k} X_\ell\right] = (n+1)\frac{k}{k+1}$$

$$\Rightarrow n \approx \left(\frac{k+1}{k}\max_{1\leq \ell \leq k} X_\ell\right) - 1$$

(**1947**) Le début de l'histoire… *The German tank problem*

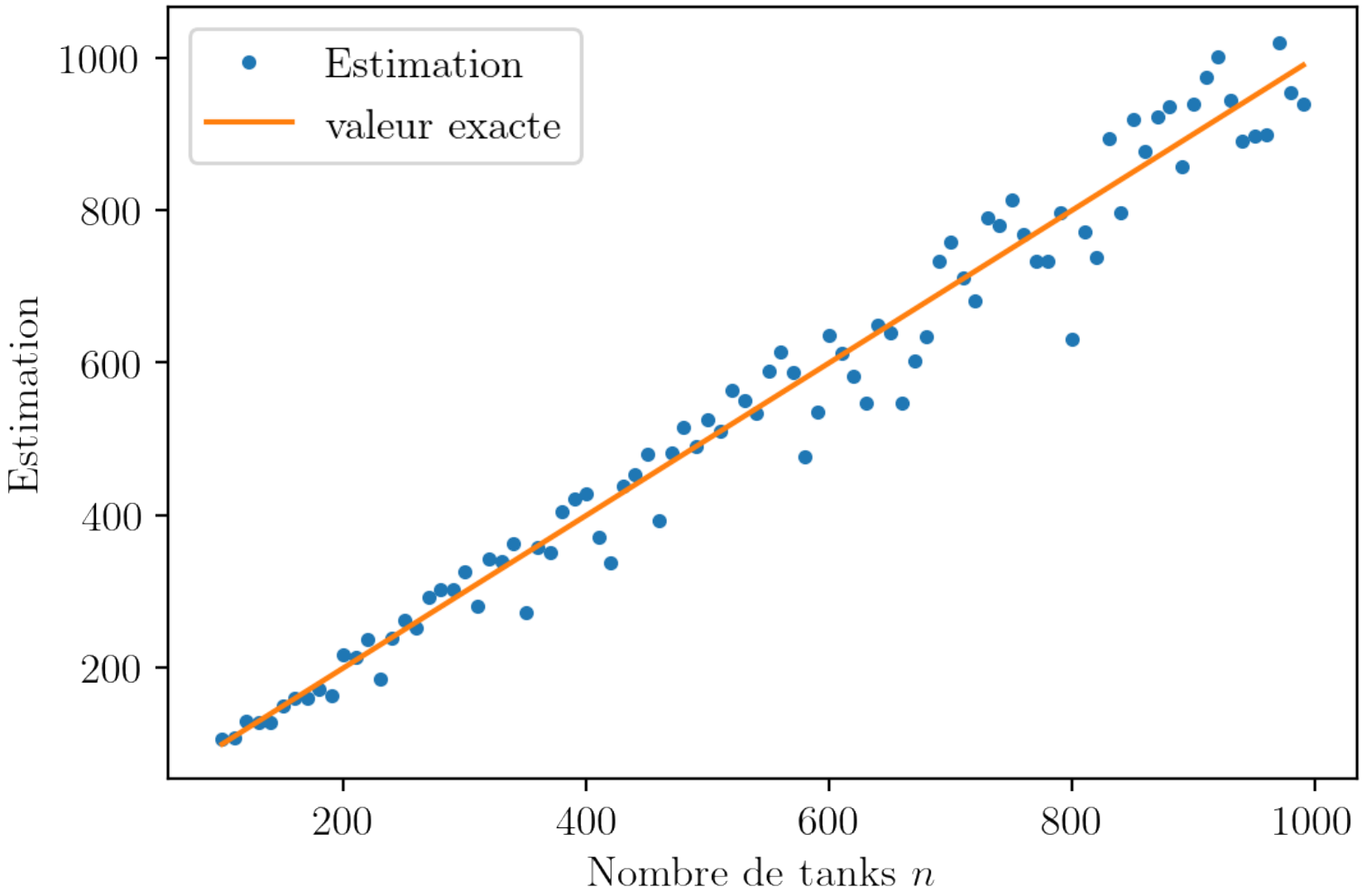
Date	Estimated Monthly Production		Monthly Production Speer Ministry
	Serial Number Estimate	Munitions Record 10 Aug. 42	
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

Source : R.Ruggles, & H.Brodie.

Problème des tanks allemands : Bilan

- Utilisation des probabilités… pour un problème pas aléatoire!
- Il faut un (bon) modèle
- Attention… les maths peuvent servir

Simulations pour $k = 20$ tanks

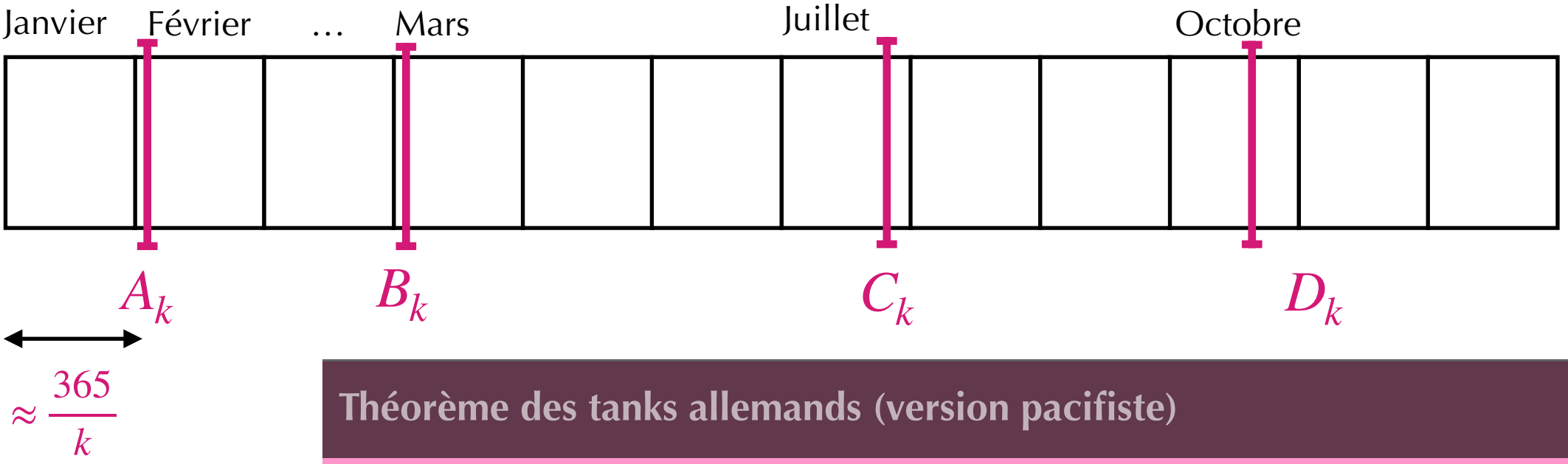


Interlude

Pacifions le théorème

Dans la salle : quel est l'anniversaire

- **le plus tôt dans l'année?** $\rightsquigarrow A_k$
- **le plus tôt après le 1er mars ?** $\rightsquigarrow B_k$
- **plus tôt après le 1er juillet ?** $\rightsquigarrow C_k$
- **plus tôt après le 1er octobre ?** $\rightsquigarrow D_k$



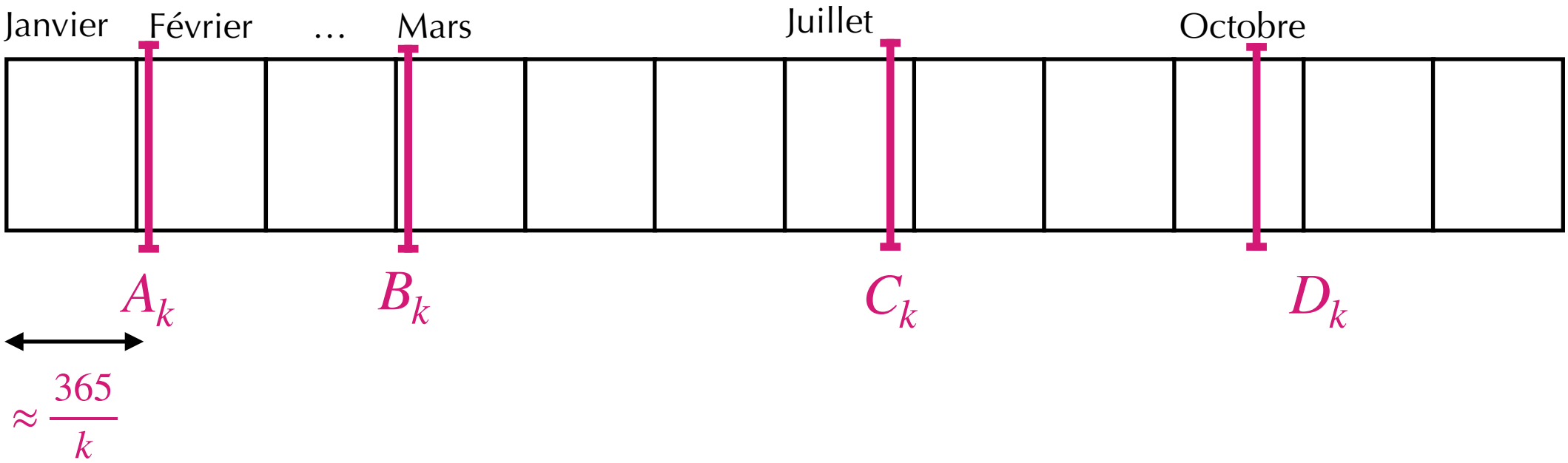
Théorème des tanks allemands (version pacifiste)

Si les anniversaires sont uniformes dans l'année,

$$\mathbb{E}\left[\left(A_k - \textbf{1er Janvier}\right)\right] = \frac{365}{k}, \qquad \mathbb{E}\left[\left(B_k - \textbf{1er Mars}\right)\right] = \frac{365}{k}, \qquad \dots$$

Interlude

Pacifions le théorème



Corollaire

Si les anniversaires sont uniformes dans l'année,

$$k \approx \frac{4 \times 365}{(A_k - \textbf{1er janvier}) + (B_k - \textbf{1er mars}) + (C_k - \textbf{1er juillet}) + (D_k - \textbf{1er octobre})}$$

Plan de l'exposé

1947 Problème des tanks allemands
1978 Algorithme de Morris
2022 Algorithme CVM

Plan de l'exposé

1947 Problème des tanks allemands

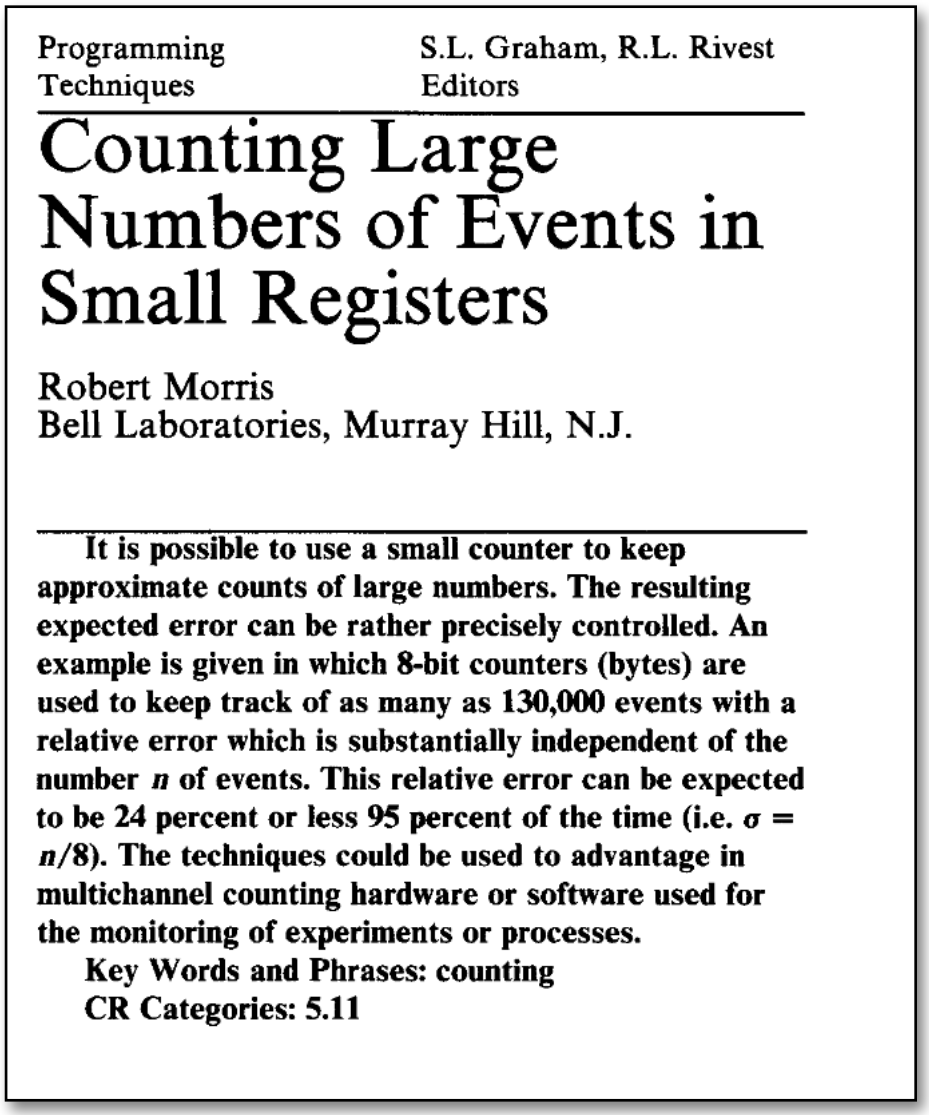
1978 Algorithme de Morris

2022 Algorithme CVM

(**1978**) Suite de l'histoire…

Article scientifique :
Robert Morris (Bell Labs) (1978)

- **But :** Compter le nombre de mots dans un texte donné



Programming Techniques — S.L. Graham, R.L. Rivest Editors

Counting Large Numbers of Events in Small Registers

Robert Morris
Bell Laboratories, Murray Hill, N.J.

It is possible to use a small counter to keep approximate counts of large numbers. The resulting expected error can be rather precisely controlled. An example is given in which 8-bit counters (bytes) are used to keep track of as many as 130,000 events with a relative error which is substantially independent of the number n of events. This relative error can be expected to be 24 percent or less 95 percent of the time (i.e. $\sigma = n/8$). The techniques could be used to advantage in multichannel counting hardware or software used for the monitoring of experiments or processes.

Key Words and Phrases: counting
CR Categories: 5.11

(**1978**) Suite de l'histoire…

Article scientifique :
Robert Morris (Bell Labs) (1978)

Contexte : Traitement du texte

- **But :** Compter le nombre d'éléments dans une liste $x_1, \ldots, x_n$
- On ne peut compter que jusqu'à $2^m - 1$

Contrainte : Mémoire limitée à *m* bits

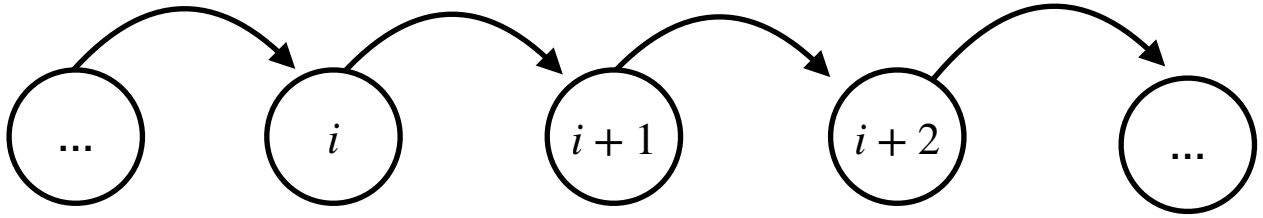
Valeur binaire	compteur
0 0 0 0 0 0	0
0 0 0 0 0 1	1
0 0 0 0 1 0	2
0 0 0 0 1 1	3
0 0 0 1 0 0	4
0 0 0 1 0 1	5
0 0 0 1 1 0	6
0 0 0 1 1 1	7
…	
1 1 1 1 1 0	$2^m - 2$
1 1 1 1 1 1	$2^m - 1$

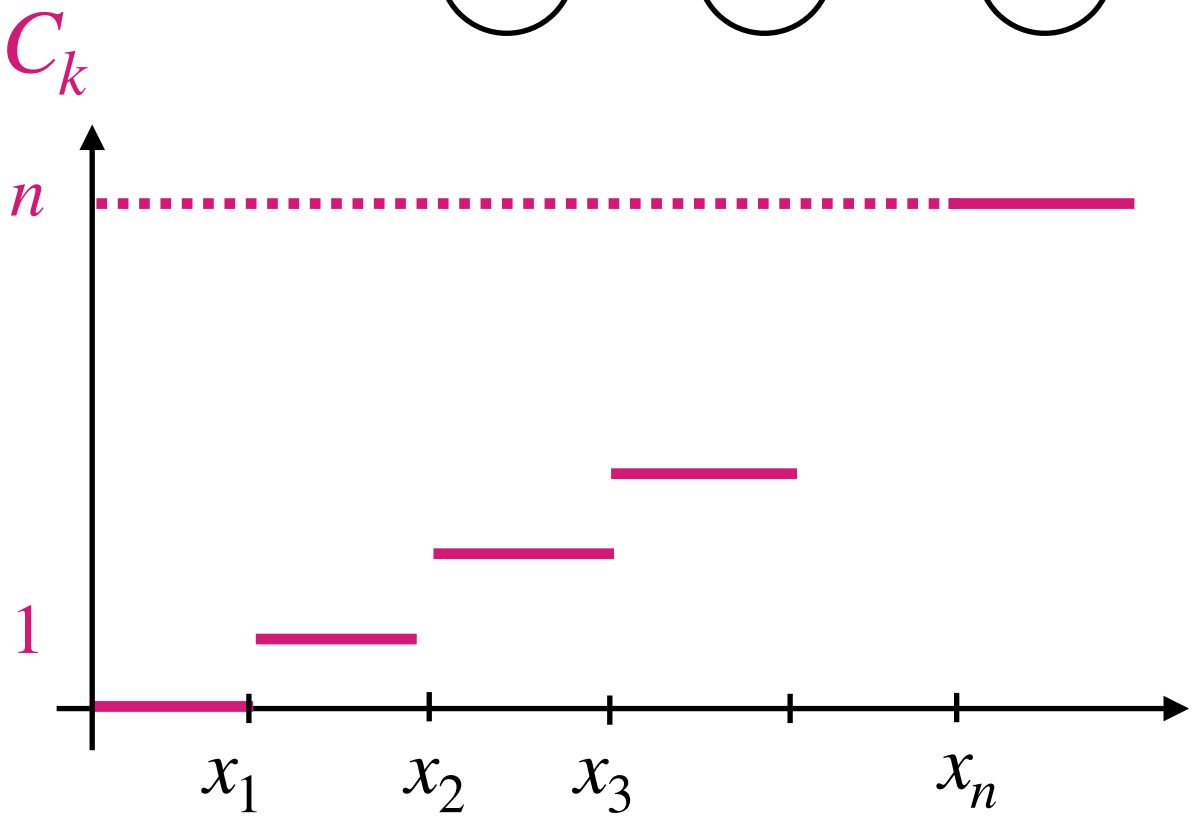
(**1978**) Suite de l'histoire…

• **But :** Compter le nombre d'éléments dans une liste $x_1, \ldots, x_n$

Algorithme classique : compteur $C_1, \ldots, C_k$

Rappel : arrivé à $2^m - 1$ la mémoire est saturée!





Algorithme classique

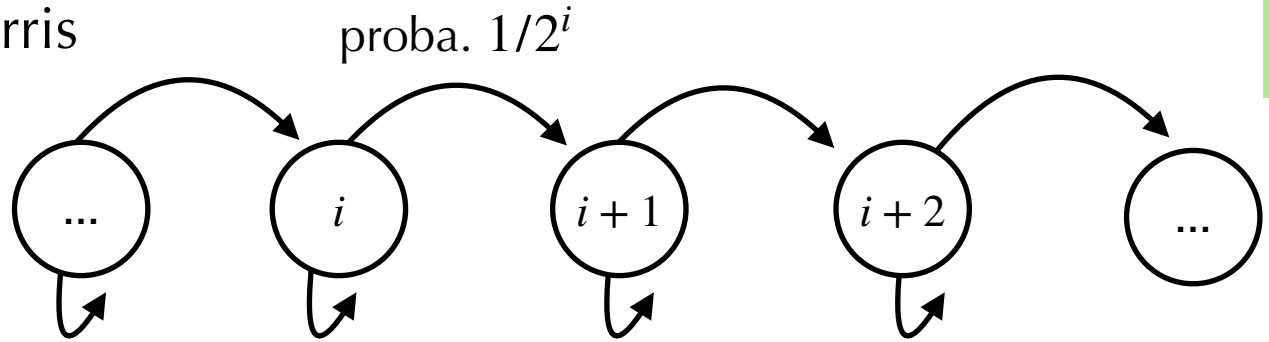
```
C=0
Pour x dans [x_1, ..., x_n]:
    C=C+1

renvoyer C
```

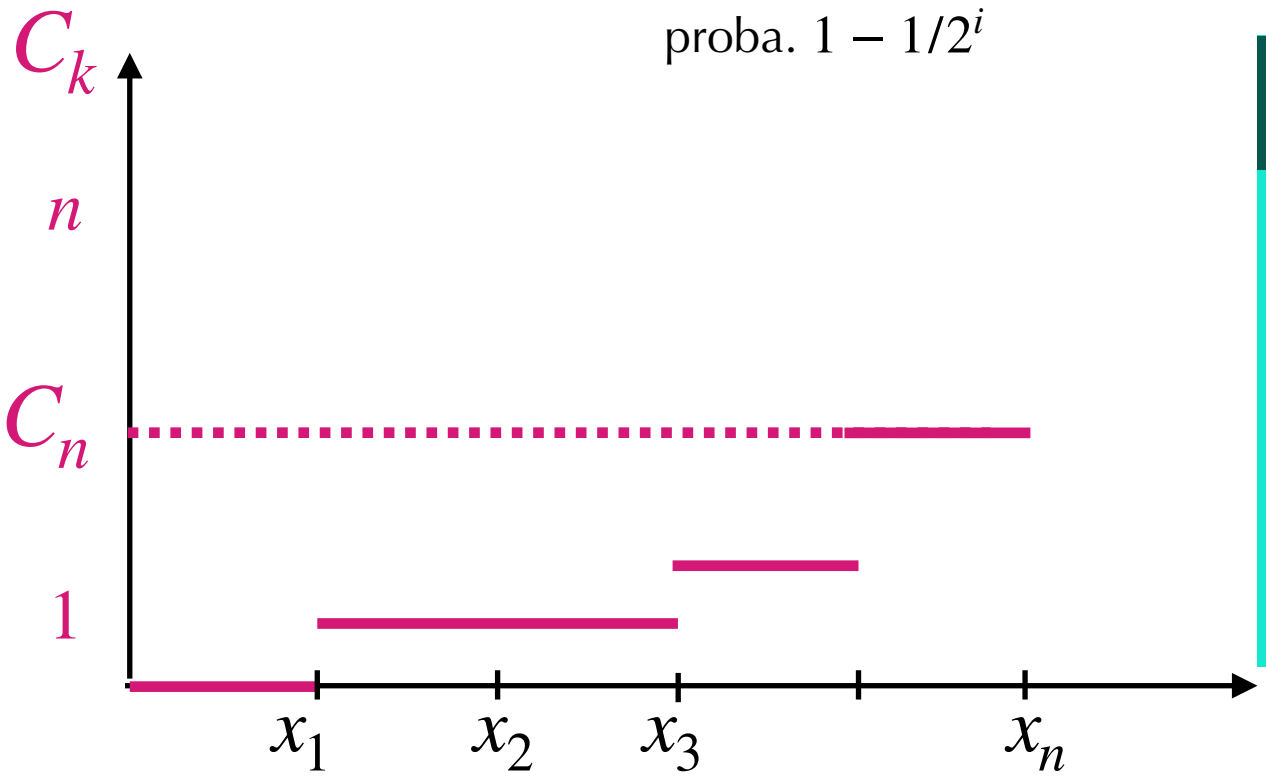
(**1978**) Suite de l'histoire…

- **But :** Compter le nombre d'éléments dans une liste $x_1, \ldots, x_n$

Rappel : arrivé à $2^m - 1$ la mémoire est saturée!

Algorithme de Morris





Algorithme de Morris

```
C=0
Pour x dans [x_1,...,x_n]:
  avec proba. 2^-C faire:
    C=C+1

renvoyer 2^C - 1
```

(**1978**) Suite de l'histoire…

- **But :** Compter le nombre d'élements dans une liste $x_1, \ldots, x_n$

Théorème

Soit C_k la valeur du compteur dans l'algorithme de Morris après x_k. Pour tout k,

$$\mathbb{E}[2^{C_k}] = k + 1.$$

Avec m bits :

- $0 \leq C_n \leq 2^m - 1$
- $1 \leq 2^{C_n} - 1 \leq 2^{2^m - 1} - 1$

A retenir

Avec m bits, on peut compter jusqu'à $2^{2^m-1} - 1$

Algorithme de Morris

```
C=0
Pour x dans [x_1, ..., x_n]:
   avec proba. 2^{-C} faire:
      C=C+1

renvoyer 2^C - 1
```

Plan de l'exposé

1947 Problème des tanks allemands
1978 Algorithme de Morris
2022 Algorithme CVM

(2022) Fin de l'histoire ?

Conférence *European Symposium on Algorithms (2022)*
Sourav Chakraborty, N.V. Vinodchandran, Kuldeep S. Meel.

Contexte : Bases de données

But : Compter le nombre de mots différents dans un texte donné

Distinct Elements in Streams: An Algorithm for the (Text) Book*

Sourav Chakraborty ⓡ
Indian Statistical Institute, India

N. V. Vinodchandran ⓡ
University of Nebraska-Linclon, USA

Kuldeep S. Meel
National University of Singapore, Singapore

Abstract

Given a data stream $\mathcal{A} = \langle a_1, a_2, \ldots, a_m \rangle$ of m elements where each $a_i \in [n]$, the Distinct Elements problem is to estimate the number of distinct elements in $\mathcal{A}$. Distinct Elements has been a subject of theoretical and empirical investigations over the past four decades resulting in space optimal algorithms for it. All the current state-of-the-art algorithms are, however, beyond the reach of an undergraduate textbook owing to their reliance on the usage of notions such as pairwise independence and universal hash functions. We present a simple, intuitive, sampling-based space-efficient algorithm whose description and the proof are accessible to undergraduates with the knowledge of basic probability theory.

2012 ACM Subject Classification Theory of computation → Sketching and sampling

Keywords and phrases Distinct Elements Estimation, Streaming, Sampling

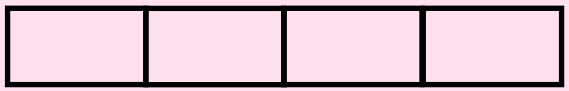
(2022) Fin de l'histoire ?

Conférence *European Symposium on Algorithms (2022)*
Sourav Chakraborty, N.V. Vinodchandran, Kuldeep S. Meel.

Contexte : Bases de données

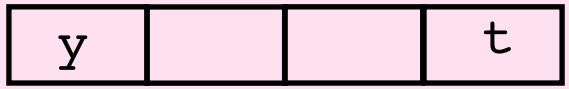
But : Compter le nombre de mots différents dans un texte donné

Algorithme CVM (version un peu simplifiée)

Mémoire = 

pour x dans $[x_1, \ldots, x_n]$:
  si x dans Mémoire, on l'efface

  avec proba 1/2 on ajoute x dans Mémoire



renvoyer $2 \times$ {Nombre de mots dans Mémoire}

Pour La Science (Nov.2024)

R

ENDEZ-VOUS

LOGIQUE & CALCUL

L'ALGORITHME OUBLIÉ

La découverte d'un nouvel algorithme de comptage, reposant sur des principes élémentaires mais pourtant passé inaperçu jusqu'à récemment, étonne les spécialistes.

L'AUTEUR

JEAN-PAUL DELAHAYE
professeur émérite
à l'université de Lille

Il est difficile d'imaginer que les chercheuses et les chercheurs, qui sont nombreux et dont l'intelligence collective ne fait aucun doute, aient pu passer à côté d'un résultat simple au sujet d'un problème important, ou d'une méthode élémentaire de calcul. Poser une nouvelle pierre sur le mur des savoirs qui s'accumulent ne peut, pense-t-on, que demander un gros travail, et ce qu'on ajoutera revêtira forcément une certaine complexité. Eh bien ce concepts informatiques délicats, comme celui de fonction universelle de hachage. Trois chercheurs viennent cependant d'effectuer un progrès inattendu: Sourav Chakraborty, de l'Institut indien de statistiques, à Calcutta, Vinodchandran Variyam, de l'École d'informatique de l'université du Nebraska, aux États-Unis, et Kuldeep Meel, professeur au département d'informatique de l'université de Toronto, au Canada.

L'algorithme proposé a provoqué l'étonne-

Merci!

Messages à emporter…

- Utilisation des probabilités… pour des problèmes pas aléatoires!
- On a besoin de créativité pour trouver de nouveaux algos
- Attention… les maths peuvent servir