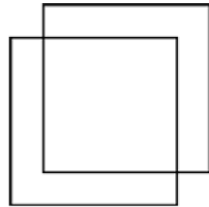


# Z GB na kB: jak naskečovat velká data a neztratit při tom hlavu (ani patu)



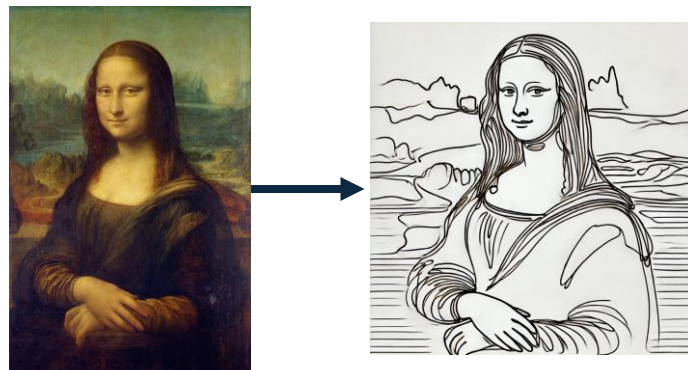
Pavel Veselý

INFORMATICKÝ ÚSTAV  
UNIVERZITY KARLOVY  
Matematicko-fyzikální fakulta  
Univerzita Karlova



# Plán

- Jak velká jsou dnes **velká data**?
- **Skečování dat** – proudové algoritmy
  - Jak hledat **časté prvky**?
  - Jak zjistit **počet různých prvků**?
  - Jak najít  **$k$ -tý nejmenší**, např. **medián**?



# Jak velká jsou **velká data**?

## Sekvence nukleotidů (DNA/RNA)

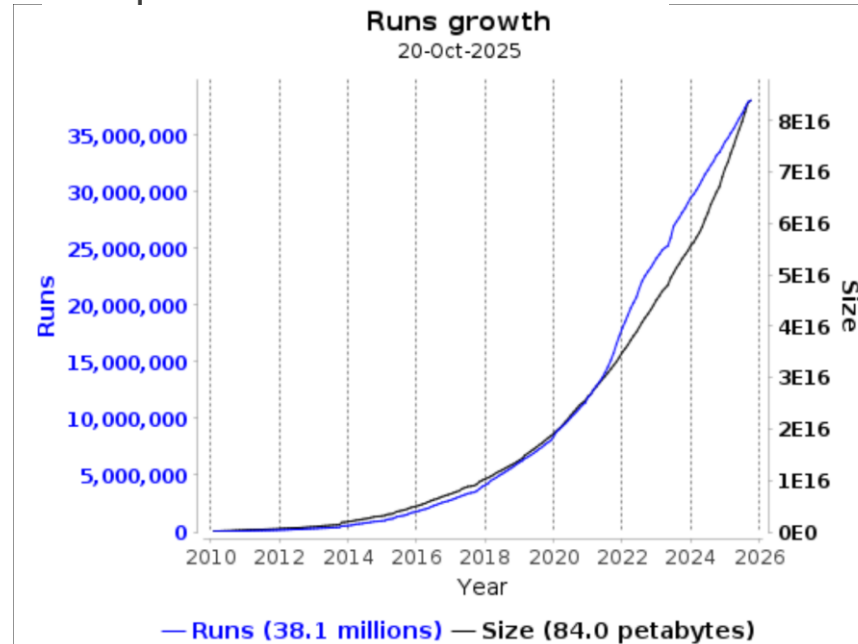
```
ATTAAGGTTTATACCTTCCCAGGTAACAAACCAACC  
AACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAA  
CTTTAAAATCTGTGTGGCTGTCACCTCGGCTGCATGCT  
TAGTGCACTCACGCAGTATAATTAATAACTAATTACT  
GTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTG  
CAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATC  
ATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAA  
GGTAAGATGGAGAGCCTTGTCCCTGGTT...
```

Celkem **67 petabytů**  $\approx 67 \cdot 10^{15}$  bytů!

- Komprimovaná o řád méně
- Potřebujete tisíce běžných HDD



European Nucleotide Archive



Zdroj: European Bioinformatics Institute, ENA

## Jak v tom vyhledávat?

# Co nám řekl **COVID** o velkých datech?

13.1.2020: vědci publikovali genom SARS-CoV-2

GenBank Send to: ▾

## Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC\_045512.1

⚠ This sequence has been updated. [See current version.](#)

[FASTA](#) [Graphics](#)

---

Go to:

LOCUS	NC_045512	30473 bp ss-RNA	linear	VRL 13-JAN-2020
DEFINITION	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome.			
ACCESSION	NC_045512			
VERSION	NC_045512.1			
DBLINK	BioProject: <a href="#">PRJNA485481</a>			
KEYWORDS	RefSeq.			
SOURCE	Wuhan seafood market pneumonia virus			
ORGANISM	<a href="#">Wuhan seafood market pneumonia virus</a>			
	Viruses; Riboviria; Nidovirales; Coronavirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; unclassified Betacoronavirus.			
REFERENCE	1 (bases 1 to 30473)			
AUTHORS	Zhang, Y.-Z., Wu, F., Chen, Y.-M., Pei, Y.-Y., Xu, L., Wang, W., Zhao, S., Yu, B., Hu, Y., Tao, Z.-W., Song, Z.-G., Tian, J.-H., Zhang, Y.-L., Liu, Y., Zheng, J.-J., Dai, F.-H., Wang, Q.-M., She, J.-L. and Zhu, T.-Y.			
TITLE	A novel coronavirus associated with a respiratory disease in Wuhan of Hubei province, China			
JOURNAL	Unpublished			
REFERENCE	2 (bases 1 to 30473)			
CONSTRM	NCBI Genome Project			
TITLE	Direct Submission			
JOURNAL	Submitted (13-JAN-2020) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA			
REFERENCE	3 (bases 1 to 30473)			
AUTHORS	Zhang, Y.-Z., Wu, F., Chen, Y.-M., Pei, Y.-Y., Xu, L., Wang, W., Zhao, S., Yu, B., Hu, Y., Tao, Z.-W., Song, Z.-G., Tian, J.-H., Zhang, Y.-L., Liu, Y., Zheng, J.-J., Dai, F.-H., Wang, Q.-M., She, J.-L. and Zhu, T.-Y.			
TITLE	Direct Submission			
JOURNAL	Submitted (05-JAN-2020) Department of Zoonoses, National Institute			



Máme **genom**  
nebezpečného viru!  
Pojďme udělat  
**mRNA vakcínu!**



Máme nový genom!  
Pojďme **prohledat**  
**databázi**, jestli se  
něčemu nepodobá...

**Kdo byl rychlejší?**

# Co nám řekl **COVID** o velkých datech?

Po **7 měsících**: takřka **remíza**....



Hurá! Vakcína  
prošla 1. testy!




Uff... **dokončil**  
**jsem**  
**prohledávání**...

**nature**

Article | [Published: 12 August 2020](#)

## **Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults**

[Mark J. Mulligan](#), [Kirsten E. Lyke](#), [Nicholas Kitchin](#), [Judith Absalon](#) , [Alejandra Gurtman](#), [Stephen Lockhart](#), [Kathleen Neuzil](#), [Vanessa Raabe](#), [Ruth Bailey](#), [Kena A. Swanson](#), [Ping Li](#), [Kenneth Koury](#), [Warren Kalina](#), [David Cooper](#), [Camila Fontes-Garfias](#), [Pei-Yong Shi](#), [Özlem Türeci](#), [Kristin R. Tompkins](#), [Edward E. Walsh](#), [Robert Frencck](#), [Ann R. Falsey](#), [Philip R. Dormitzer](#), [William C. Gruber](#), [Uğur Şahin](#) & [Kathrin U. Jansen](#)

*Nature* **586**, 589–593 (2020) | [Cite this article](#)

**bioRxiv**

THE PREPRINT SERVER FOR BIOLOGY

Posted August 10, 2020.

## **Petabase-scale sequence alignment catalyses viral discovery**

Robert C. Edgar, Jeff Taylor, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Victor Lin, Dan Lohr, Gherman Novakovsky, Basem Al-Shayeb, Jillian F. Banfield, Anton Korobeynikov, Rayan Chikhi,  Artem Babaian

**nature**

Article | [Published: 26 January 2022](#)

## **Petabase-scale sequence alignment catalyses viral discovery**

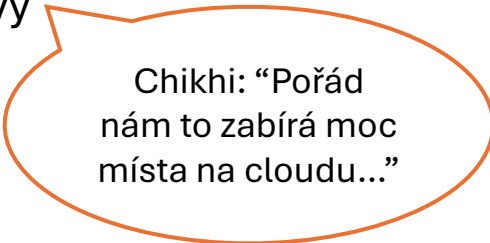
# Nedávný výsledek: Logan [Chikhi et al. 2024]

- databáze DNA a RNA sekvencí
  - **50 petabází**; Dec 2023
  - 2.5 PB komprimovaných dat

→ Nové vědecké objevy

→ **Vyhledávání**

(trvá desítky minut)



Chikhi: “Pořád nám to zabírá moc místa na cloudu...”

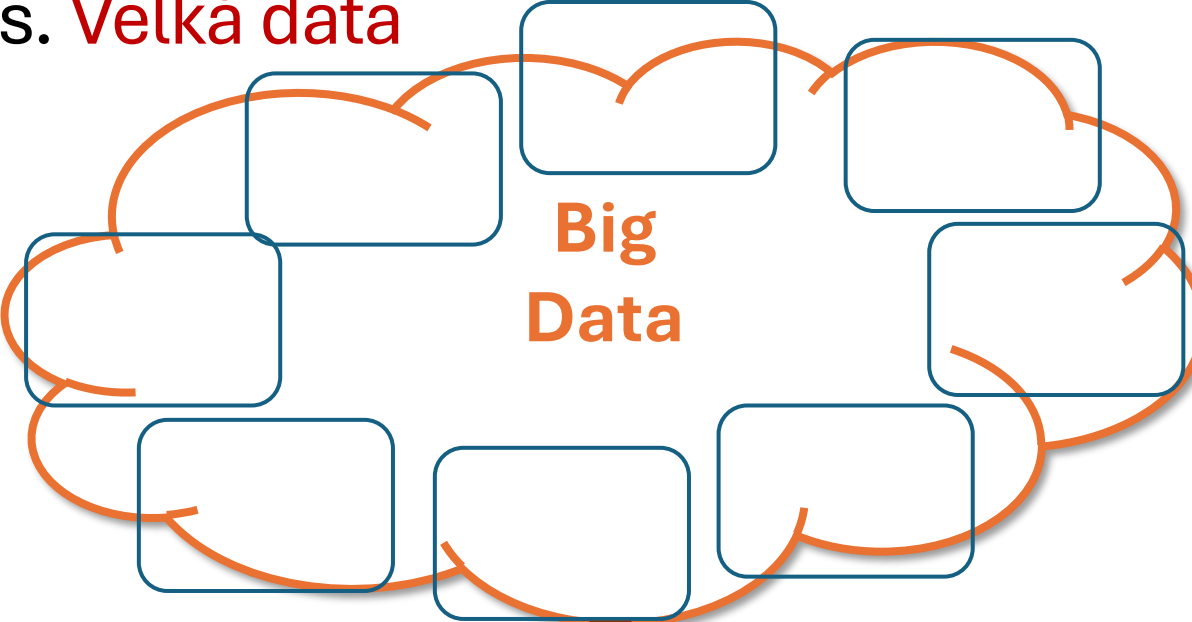
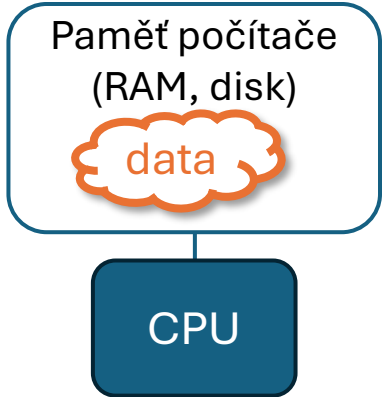
## Založeno na hešování a skečování!

(konkrétně Bloomovy filtery)

# Datové skeče a proudové algoritmy



# Klasické algoritmy vs. **Velká data**



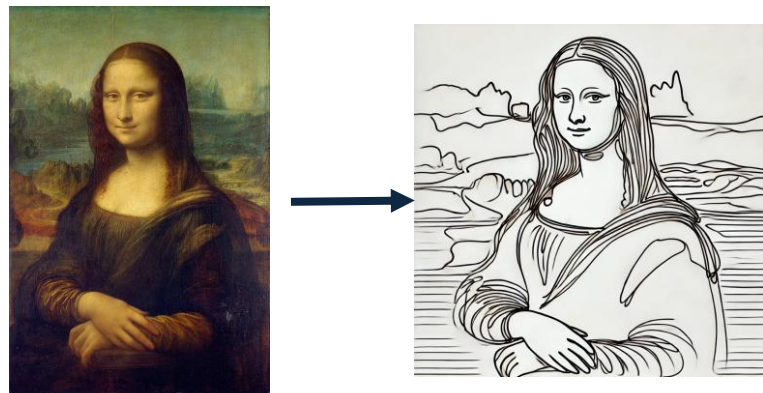
**Co s tím?**

Skeč / shrnutí



# Datové skeče a proudové (streaming) algoritmy

- Jeden průchod přes data
  - S malou pamětí
    - mnohem menší než data
    - např. pouze v řádech kilobytů
- Skeč dat  
= obsah paměti algoritmu
- Cena: ztratíme přesnost
  - Velikost teoreticky  $\leq c_1 \cdot \log^{c_2}(n) + c_3$ 
    - $c_1, c_2, c_3$  jsou konstanty
    - $n$  = velikost dat



Hledání nejčastější prvků

# Velká data v praxi 2: Jaké jsou časté hashtagy na ?

- Cíl: nalézt prvky, které se vyskytují **často**
  - Např. tvoří 1% ze všech hashtagů
- Přesněji: pro dané  $k$  najít prvky, které se vyskytnou  $\geq \frac{n}{k}$  krát
- Paměť ideálně úměrná  $c \cdot k$  pro konstantu  $c$

#Algorithms

#MachineLearning

#Algorithms

#Algorithms

#BigData

#ArtificialIntelligence

#Algorithms

#BigData

#BigData

#Algorithms

#MachineLearning

#BigData

#BigData

#DataScience

#Algorithms

#Algorithms

#DeepLearning

#Programming

#BigData

#Algorithms

#BigData

# Skeč pro nejčastější prvky

- Proudový algoritmus Misra-Gries (z roku 1982)
- Parametr  $k$ 
  - V paměti uloženo  $k$  prvků, každý s počítadlem
    - $P_x$  = počítadlo pro prvek  $x$
- Přejde nový prvek  $y$ :
  1. Pokud  $y$  uložen ve skeči  $\rightarrow$  zvýšíme  $P_y$  o 1
  2. Pokud je ve skeči  $< k$  prvků  $\rightarrow$  přidáme  $y$  s  $P_y = 1$
  3. Jinak (skeč plný a neobsahuje  $y$ ):
    - Snížíme  $P_x$  pro všechny prvky  $x$  ve skeči
    - Ze skeče vymažeme prvky  $x$

„Odčítací  
krok“

**Věta:** Misra-Gries algoritmus najde všechny prvky, které se vyskytují  $\geq \frac{n}{k+1}$  krát.

Navíc vrátí odhad  $P_x$  počtu výskytů  $x$ , který splňuje  $f_x - \frac{n}{k+1} \leq P_x \leq f_x$ , kde  $f_x$  je skutečný počet výskytů.

1. #Algorithms
2. #MachineLearning
3. #Algorithms
4. #Algorithms
5. #BigData
6. #ArtificialIntelligence
7. #Algorithms
8. #BigData
9. #BigData
10. #Algorithms
11. #MachineLearning
12. #BigData
13. #BigData
14. #DataScience
15. #Algorithms
16. #Algorithms
17. #DeepLearning
18. #Programming
19. #BigData
20. #Algorithms
21. #BigData

# Počítání různých prvků

Kolik jsme viděli různých hashtagů?

1. **#Algorithms**
2. #MachineLearning
3. **#Algorithms**
4. **#Algorithms**
5. **#BigData**
6. #ArtificialIntelligence
7. **#Algorithms**
8. **#BigData**
9. **#BigData**
10. **#Algorithms**
11. #MachineLearning
12. **#BigData**
13. **#BigData**
14. #DataScience
15. **#Algorithms**
16. **#Algorithms**
17. #DeepLearning
18. #Programming
19. **#BigData**
20. **#Algorithms**
21. **#BigData**

# Velká data v praxi 3: kolik **různých lidí** vidělo reklamu?



**Studuj  
informatiku  
na Matfyzu**

→

Co vám studium na Matfyzu  
může nabídnout?

Lidí = IP adres

Výzvy:

- **Malá paměť**
  - nelze si uložit, kdo už reklamu viděl
- **Jedna IP adresa započtena jen 1x**
  - bez ohledu na počet výskytů

192.168.45.123

**10.55.78.236**

172.16.254.3

203.0.113.46

170.17.14.55

**10.55.78.236**

185.29.32.176

...

# Počítání různých IP adres

192.168.45.123

0,90073

**10.55.78.236**

**0,02599**

172.16.254.3

0,50782

203.0.113.46

Hešovací funkce

0,97765

170.17.14.55

0,39254

**10.55.78.236**

**0,02599**

185.29.32.176

0,09751

...

...



- IP adresám přiřadíme náhodná čísla
  - Z rozsahu (0, 1)
  - **“Zahešujeme”**
- Pamatujeme si ***k* nejmenších hešů**
  - **Ostatní zahodíme**
  - V praxi např.  $k = 2\ 000$

# Počítání různých prvků: skeč *k-Minimum Values (KMV)*

192.168.45.123

**10.55.78.236**

172.16.254.3

203.0.113.46

170.17.14.55

**10.55.78.236**

185.29.32.176

...

Hešovací funkce  
→



0,90073

**0,02599**

0,50782

0,97765

0,39254

**0,02599**

0,09751

...

Př. pro  $k = 3$ :



**0,02599**

**0,09751**

**0,39254**

0,50782

0,90073

0,97765

Odhad # IP adres:

$$\frac{k - 1}{k\text{-tý nejmenší heš}}$$

V příkladu:  $\frac{2}{0,39254} \approx 5,1$

*k* nejmenších hodnot  
hešovací funkce

# Jak vypadá černá skříňka na hešování?

Kupodivu jednoduše!

- Zvolme prvočíslo  $p$  dost velké
- Náhodně vygenerujeme  $a, b \in \{0, 1, \dots, p-1\}$
- IP adresu převedenou na celé číslo  $x$  zahešujeme jako

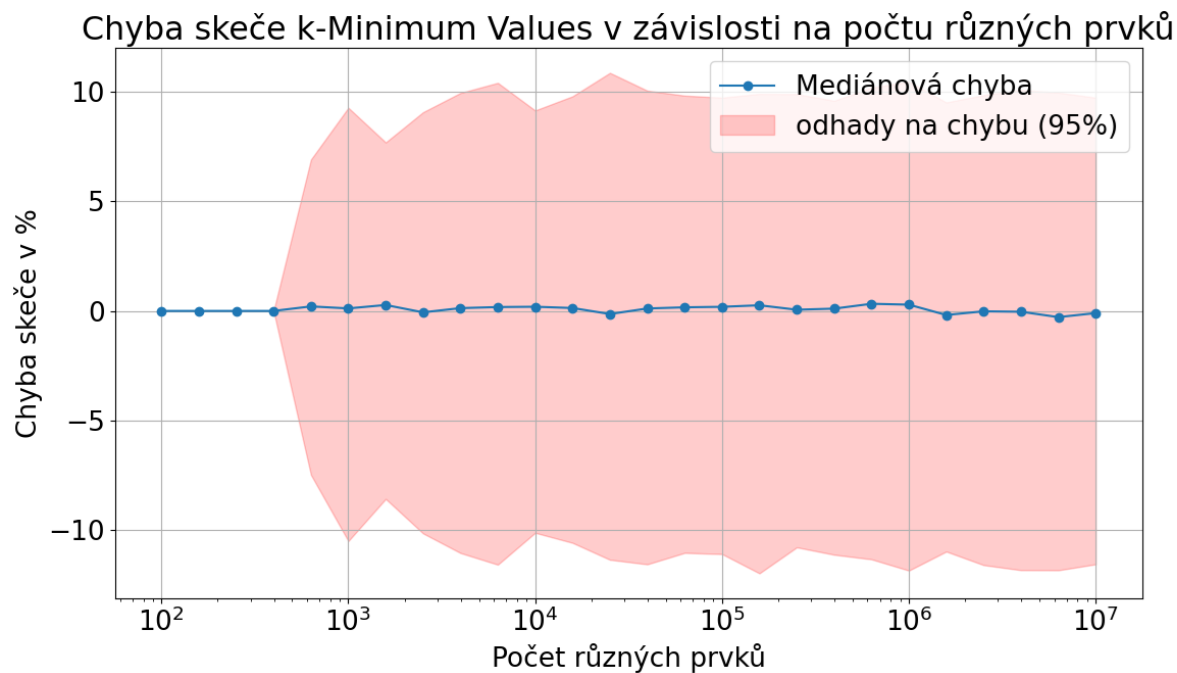
$$\frac{a \cdot x + b \pmod p}{p} \in [0, 1)$$

**Proč tahle funkce funguje dobře?  
Jaké další hešovací funkce lze použít?**



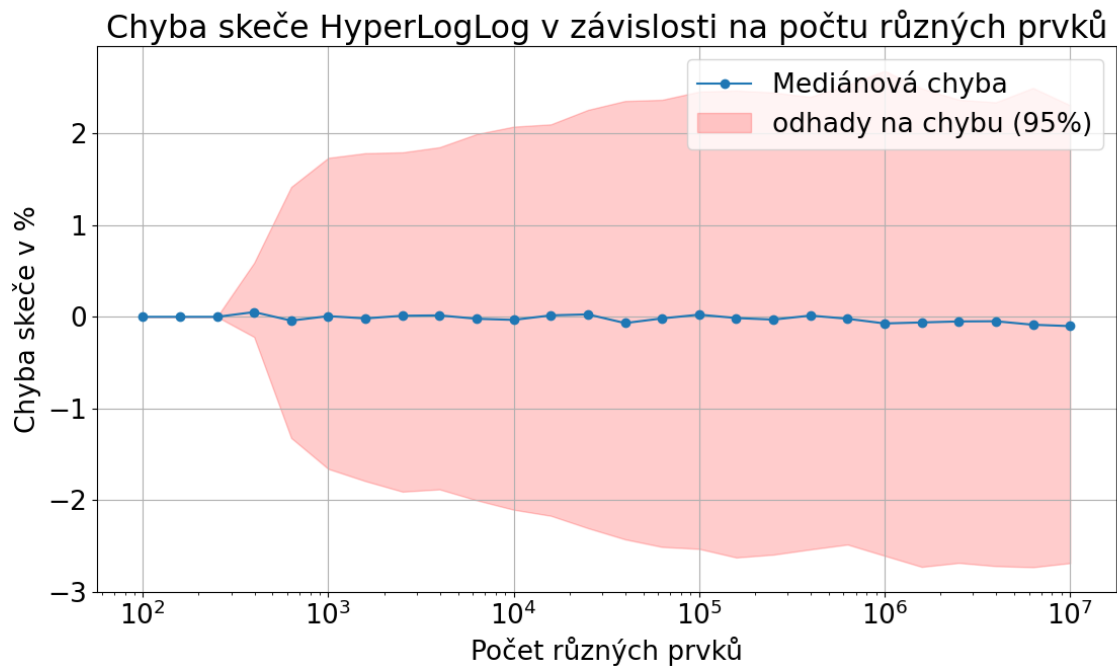
# Chyba skeče *k-Minimum Values* v praxi

- Velikost skeče cca 3 kB
- 3000 opakování na náhodné sekvenci pro daný počet různých prvků



# Chyba lepšího skeče *HyperLogLog*

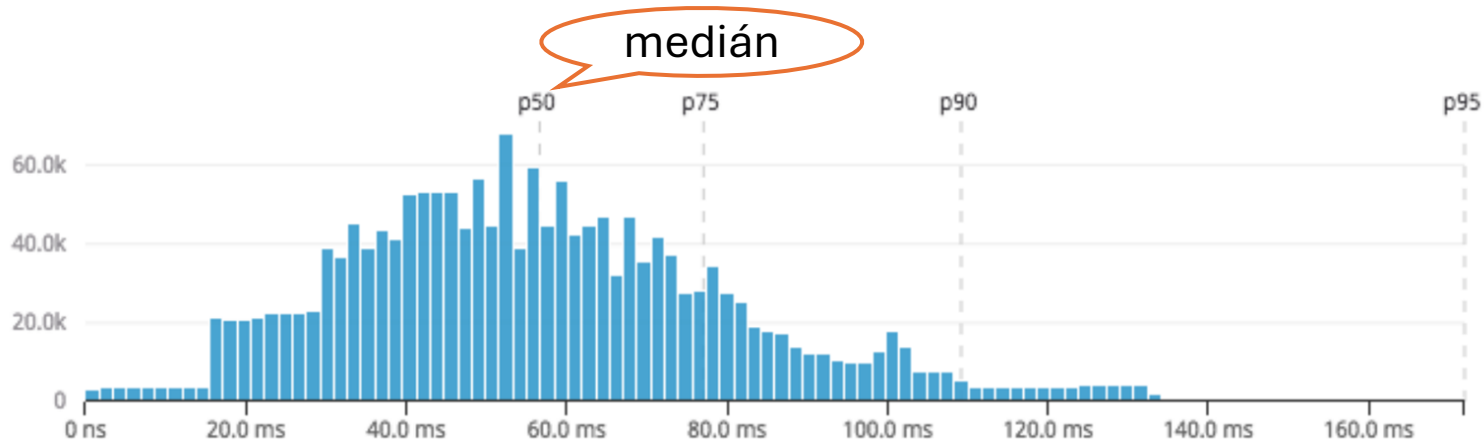
- Velikost skeče cca 2,2 kB
- 3000 opakování na náhodné sekvenci pro každý počet různých prvků



# Hledání mediánu

# Hledání mediánu a $k$ -tého nejmenšího

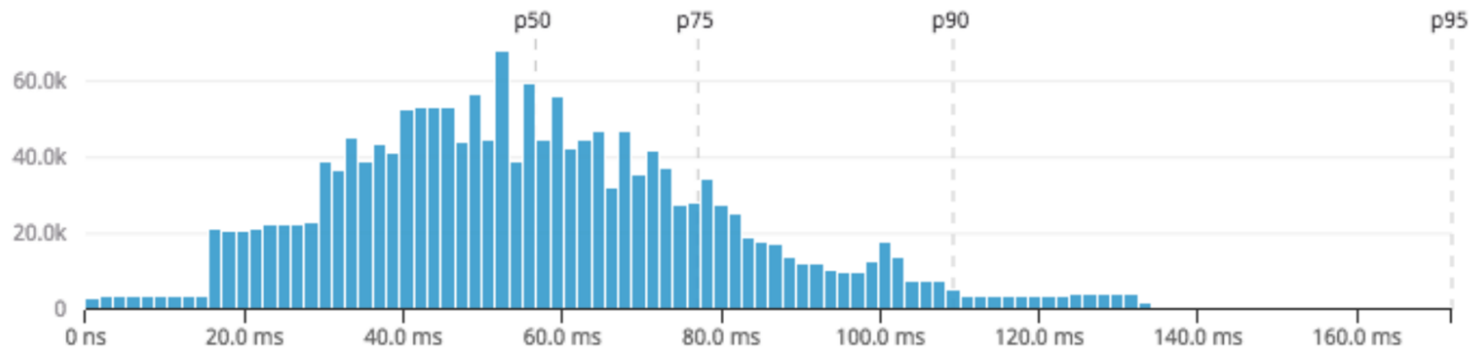
Monitorování kvality služeb na serverech – např. doba zpracování požadavku



- Podobně odhad **ranků**:  $\text{rank}(x) = \text{počet prvků menších nebo rovných } x$
- Proudové algoritmy  $\rightarrow$  chyba  $\pm \varepsilon \cdot n$ 
  - Tedy místo  $k$ -tého nejmenšího vrátíme  $(k \pm \varepsilon \cdot n)$ -tého nejmenšího

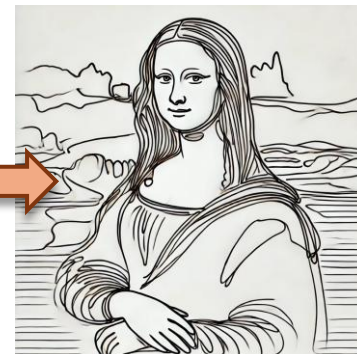
# K čemu je **skečování dat**?

- Hledání **trendů** (časté hashtagy)
- Počítání **různých IP adres** apod.
- Monitorování kvality služeb na serverech – např. doba zpracování požadavku



- **Detekce útoků** v počítačových sítích – např. skenování otevřených portů
- Synchronizace dat po síti
- Přibližná reprezentace velkých množin – např. v Loganu
- **Strojové učení (AI)** – aktualizace modelů po síti
- Publikace statistik se **zachováním soukromí** uživatelů
- ...

# Skečování dat



Děkuji za pozornost!

**#Algorithms**

**#MachineLearning**

**#Algorithms**

**#Algorithms**

**#BigData**

**#ArtificialIntelligence**

**#Algorithms**

**#BigData**

**#BigData**

**#Algorithms**

...

192.168.45.123

**10.55.78.236**

172.16.254.3

203.0.113.46

170.17.14.55

**10.55.78.236**

185.29.32.176

...

>NC\_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

```
ATTTAAAGGTTTATACCTTCCCAGGTAACAACCAACCAACTT
TCGATCTCTTGTAGATCTGTTCTCTAAACGAACCTTTAAAATC
TGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCA
GTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGT
AACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTG
ACCGAAAGGTAAGATGGAGAGCCTTGTG...
```