Narrow Big Data, streamy, kolmogorovská složitost. . .
. . . a několik vzpomínek na ISI WSC 2017

Michal Černý

Katedra ekonometrie
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze

Robust 2018

Odkud přišla motivace?

- 61th ISI World Statistics Congress
- Marakéš (Maroko), 16. – 21. července 2017
- ISI = International Statistical Institute
  - „Statistical Science for a Better World“

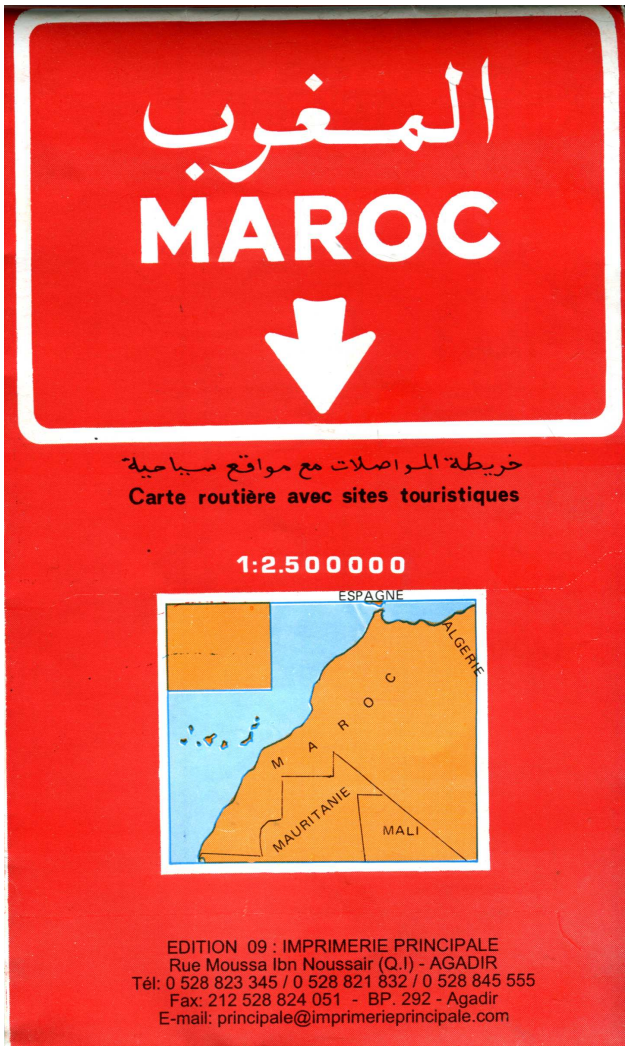
Maroko — co to je?

Maroko — co to je?

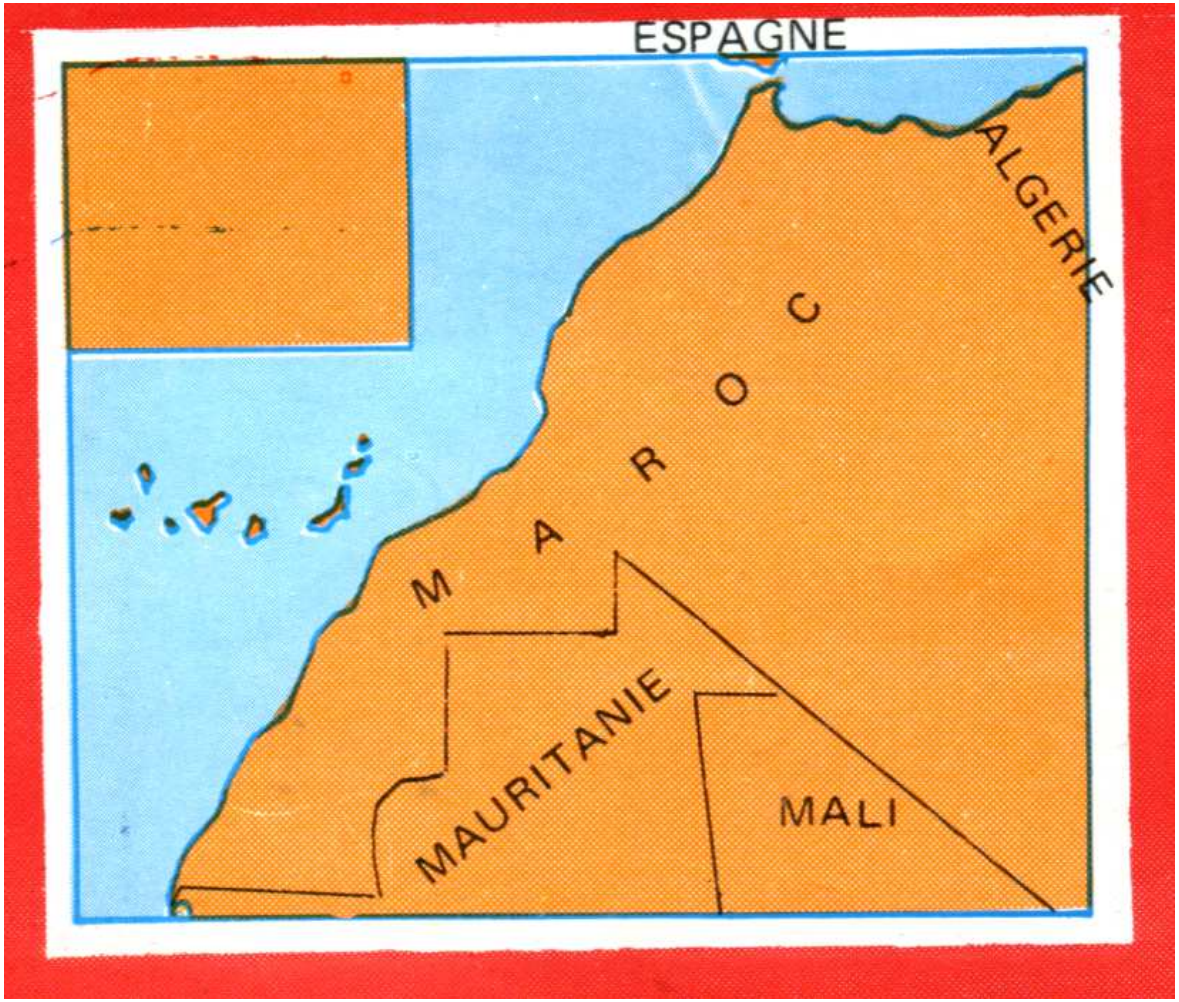
Maroko — co to je?

Maroko — co to je?

Maroko — co to je?

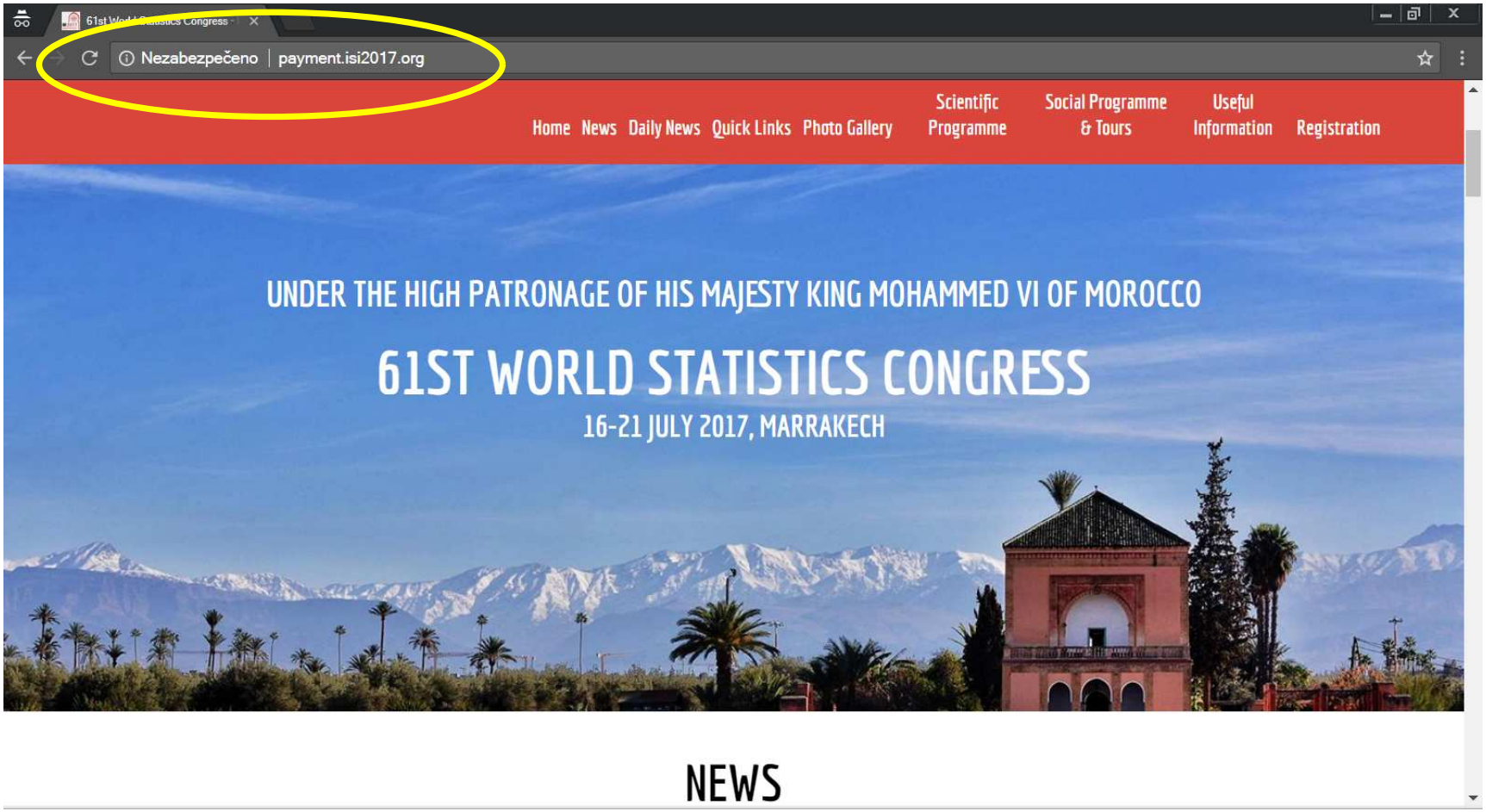


Maroko — co to je?

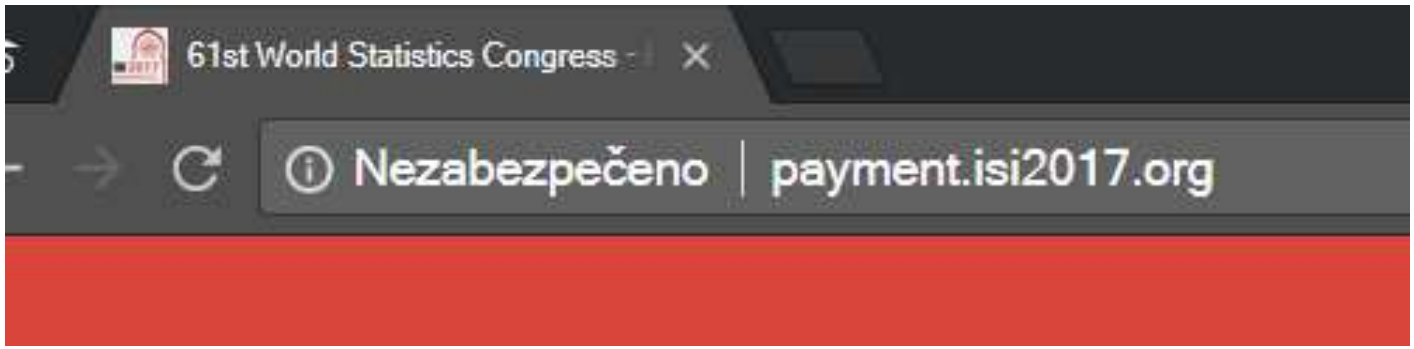


Web konference

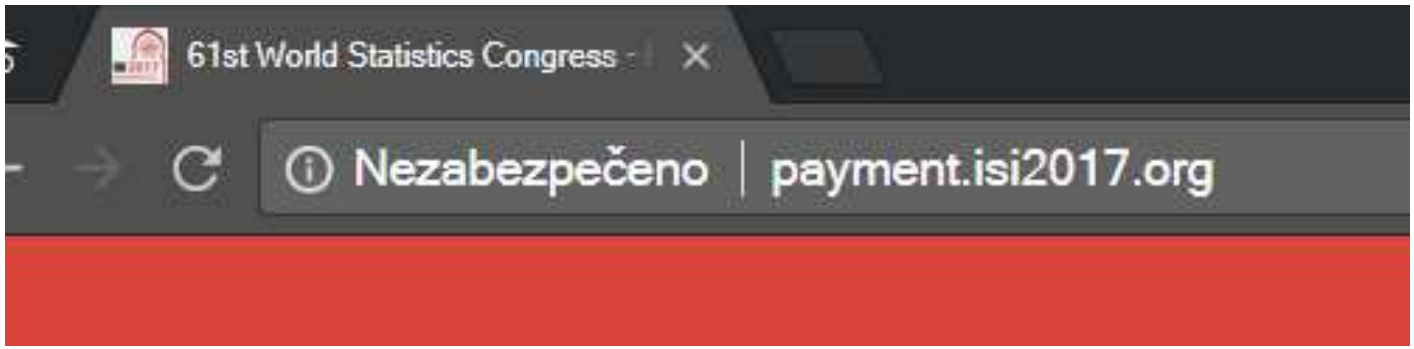
Web konference



Web konference



Web konference



Motto (organizátorů) konference: **Payment!**

Nejčastější téma

Nejčastější téma:

BIG Data.

Nejčastější téma

Nejčastější téma:

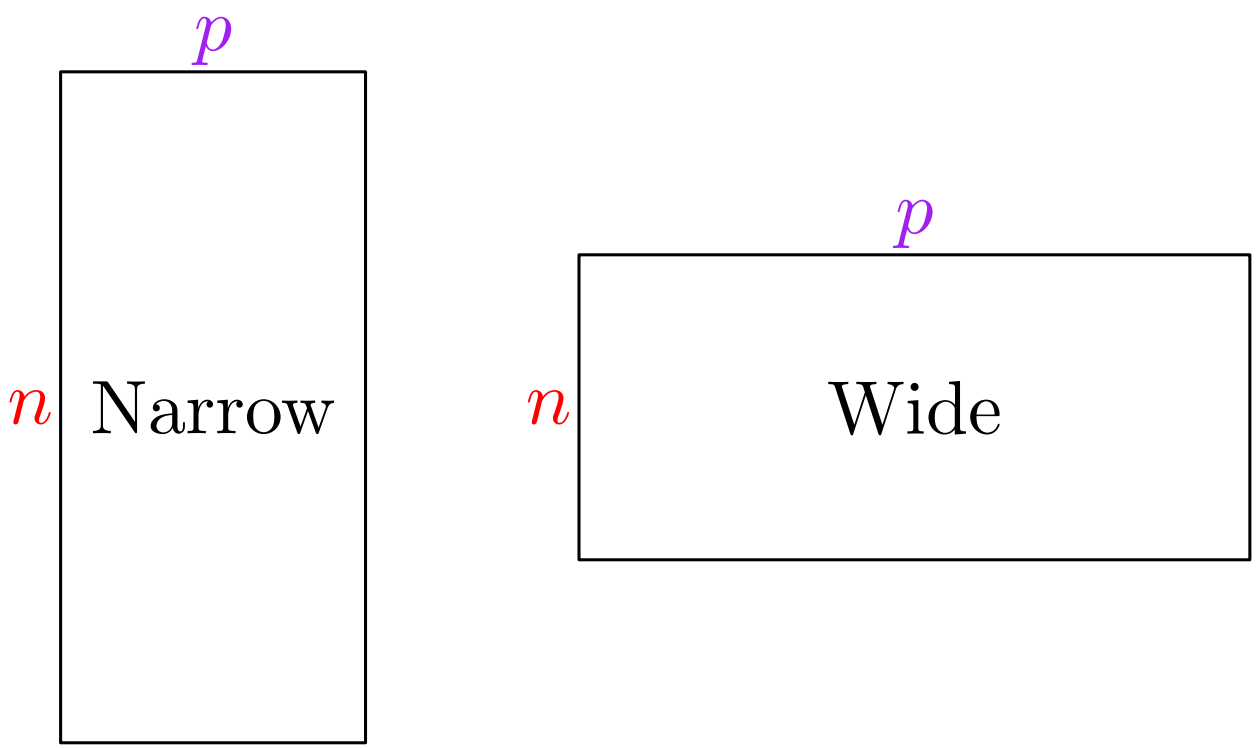
BIG Data.

Nejčastější věta:

Big Data — Yes! That's extremely important!

Big Data: $A \in \mathbb{R}^{n \times p}$

Big Data: $A \in \mathbb{R}^{n \times p}$



Narrow Big Data: Formalizujme to přesněji!

- **Data:** $A \in \mathbb{R}^{n \times p}$

Narrow Big Data: Formalizujme to přesněji!

- **Data:** $A \in \mathbb{R}^{n \times p}$
- **Narrow:** $n \gg p$ [např. $n \approx p^{\log p}$, $n \approx 2^p$, $n \approx 2^{2^p}$ apod.]

Narrow Big Data: Formalizujme to přesněji!

- **Data:** $A \in \mathbb{R}^{n \times p}$
- **Narrow:** $n \gg p$ [např. $n \approx p^{\log p}$, $n \approx 2^p$, $n \approx 2^{2^p}$ apod.]
- **Big:** v paměti lze uložit řádově $p^{O(1)}$ čísel, ne více

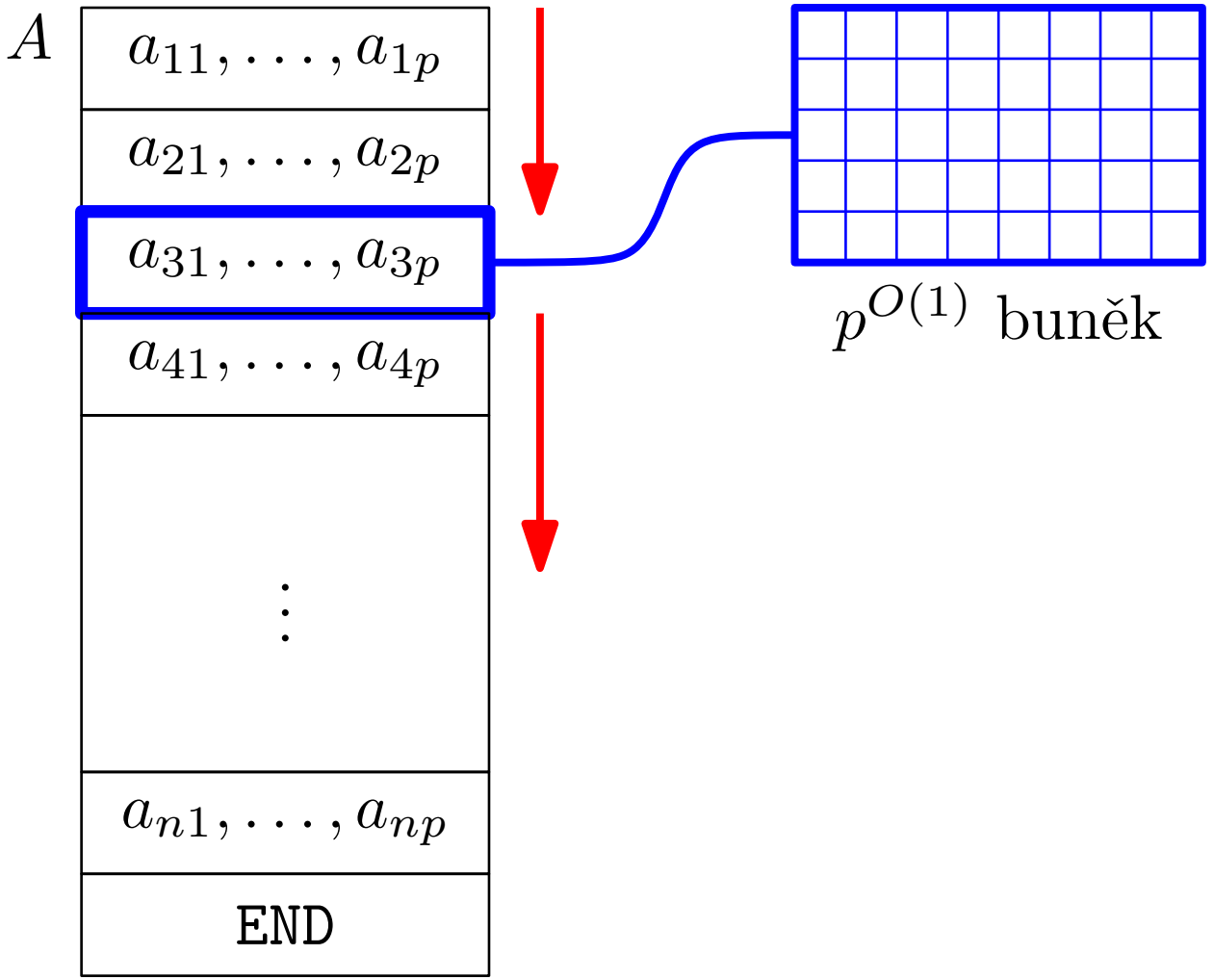
Narrow Big Data: Formalizujme to přesněji!

- **Data:** $A \in \mathbb{R}^{n \times p}$
- **Narrow:** $n \gg p$ [např. $n \approx p^{\log p}$, $n \approx 2^p$, $n \approx 2^{2^p}$ apod.]
- **Big:** v paměti lze uložit řádově $p^{O(1)}$ čísel, ne více
- Data se nevejdou do paměti. Jak se k nim přistupuje?

Narrow Big Data: Formalizujme to přesněji!

- **Data:** $A \in \mathbb{R}^{n \times p}$
- **Narrow:** $n \gg p$ [např. $n \approx p^{\log p}$, $n \approx 2^p$, $n \approx 2^{2^p}$ apod.]
- **Big:** v paměti lze uložit řádově $p^{O(1)}$ čísel, ne více
- Data se nevejdou do paměti. Jak se k nim přistupuje?
- Zde předpokládáme, že jsou organizována jako **stream**.

Streamový model jako Turingův stroj



Dva příklady

- Příklad 1: (Karel Ha) Large Hadron Collider v CERNu
- Příklad 2: (ISI WSC) Let přes Atlantik $\approx$ 650T dat

Co je paměťová buňka?

- **Nepříjemná technikálie:** je třeba omezit velikost paměťové buňky (k uložení jednoho čísla)

Co je paměťová buňka?

- **Nepříjemná technikálie:** je třeba omezit velikost paměťové buňky (k uložení jednoho čísla)
- Paměťová buňka $= (Lp \log n)^{O(1)}$ bitů

Co je paměťová buňka?

- **Nepříjemná technikálie:** je třeba omezit velikost paměťové buňky (k uložení jednoho čísla)
- Paměťová buňka $= (Lp \log n)^{O(1)}$ bitů
- Počítáme s **racionálními čísly**: $A \equiv (a_{ij}) \in \mathbb{Q}^{n \times p}$
- $L = \max_{ij} \text{bitsize}(a_{ij})$

Co je paměťová buňka?

- **Nepříjemná technikálie:** je třeba omezit velikost paměťové buňky (k uložení jednoho čísla)
- Paměťová buňka $= (Lp \log n)^{O(1)}$ bitů
- Počítáme s **racionálními čísly**: $A \equiv (a_{ij}) \in \mathbb{Q}^{n \times p}$
- $L = \max_{ij} \text{bitsize}(a_{ij})$
- Proč to?!
- Paměťová buňka má být dost velká, abychom mohli uložit např. $\sum_{i,j} a_{ij}$ [$\longrightarrow$ výpočetní model nemá být nepřirozeně omezující]

Co je paměťová buňka?

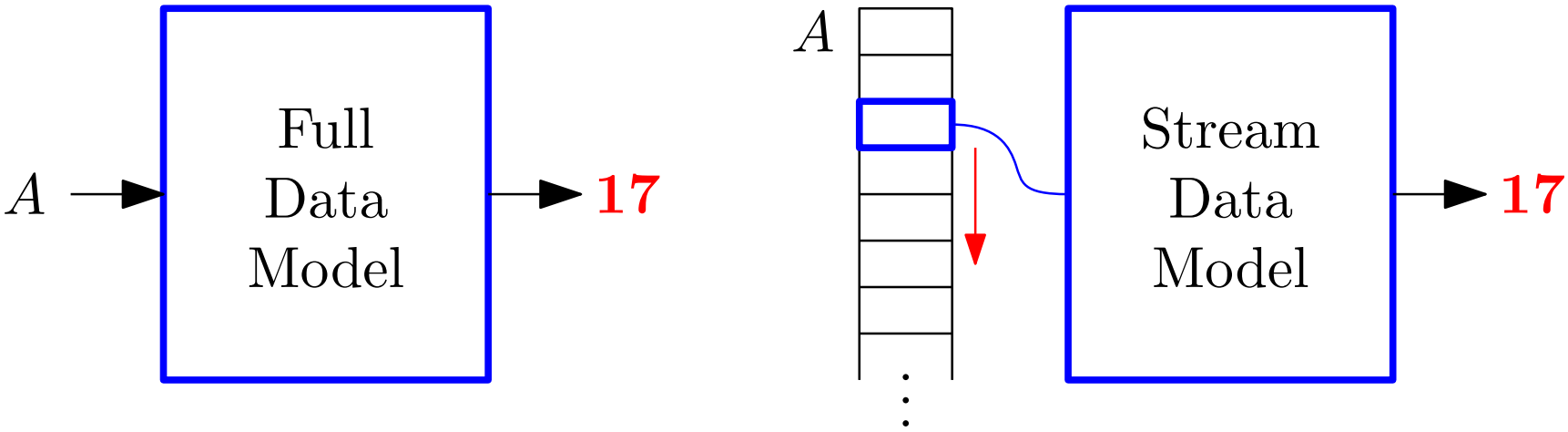
- **Nepříjemná technikálie:** je třeba omezit velikost paměťové buňky (k uložení jednoho čísla)
- Paměťová buňka $= (Lp \log n)^{O(1)}$ bitů
- Počítáme s **racionálními čísly**: $A \equiv (a_{ij}) \in \mathbb{Q}^{n \times p}$
- $L = \max_{ij} \text{bitsize}(a_{ij})$
- Proč to?!
- Paměťová buňka má být dost velká, abychom mohli uložit např. $\sum_{i,j} a_{ij}$ [$\longrightarrow$ výpočetní model nemá být nepřirozeně omezující]
- Ale: do buňky velikosti $\approx Lpn$ už bychom mohli uložit celý dataset [$\longrightarrow$ výpočetní model nemá připouštět „neférˮ podvody]

Otázka

- **Otázka:** Co lze ve streamovém modelu spočítat přesně?

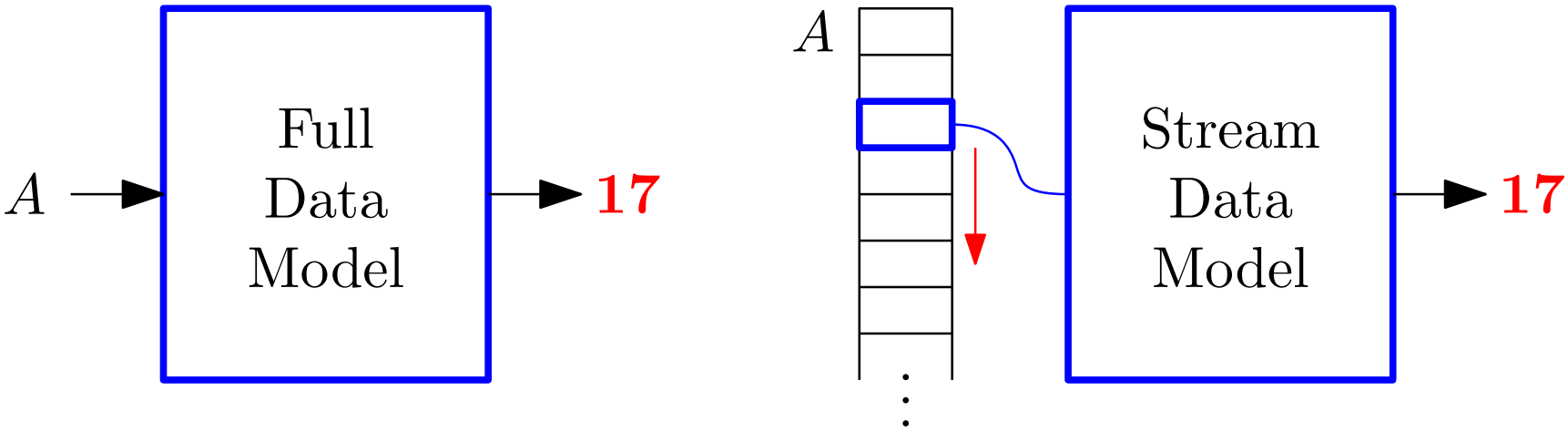
Otázka

- **Otázka:** Co lze ve streamovém modelu spočítat přesně?



Otázka

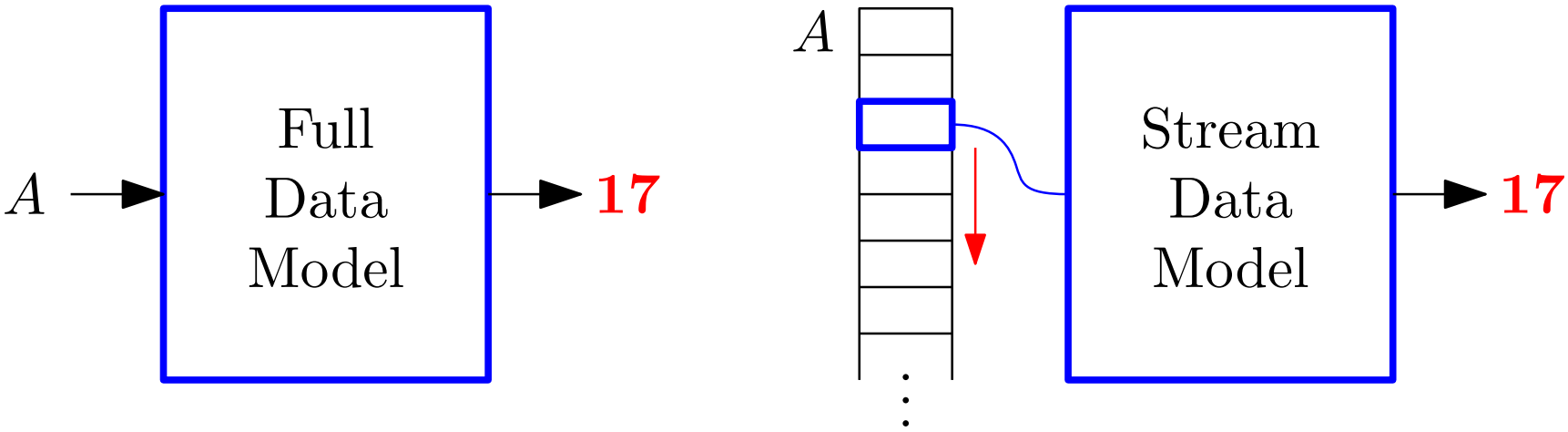
- **Otázka:** Co lze ve streamovém modelu spočítat přesně?



- Příklad: $\frac{1}{n}\sum_i a_i$ . . . **LZE**

Otázka

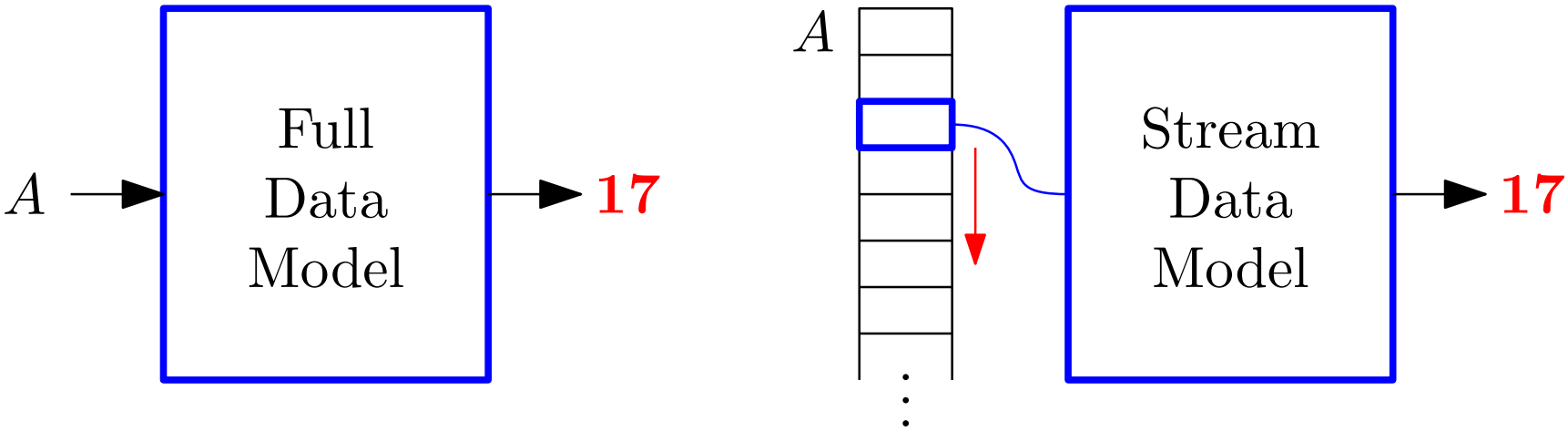
- **Otázka:** Co lze ve streamovém modelu spočítat přesně?



- Příklad: $\frac{1}{n}\sum_i a_i$ . . . **LZE**
- Příklad: $\frac{1}{n}\sum_i a_i^\ell$ . . . **LZE, je-li** $\ell = O(1)$

Otázka

- **Otázka:** Co lze ve streamovém modelu spočítat přesně?



- Příklad: $\frac{1}{n}\sum_i a_i$ . . . **LZE**
- Příklad: $\frac{1}{n}\sum_i a_i^\ell$ . . . **LZE, je-li** $\ell = O(1)$
- Příklad: Empirické kvantily . . . **NELZE**

Negativní důkazy

- Konečná 0-1 posloupnost = string: $y = 01100010 \cdots 1$

Negativní důkazy

- Konečná 0-1 posloupnost = string: $y = 01100010 \cdots 1$
- Kolmogorovská složitost:

$$\mathfrak{K}(y) = \text{délka nejkratšího programu, jenž } y \text{ vytiskne}$$

Negativní důkazy

- Konečná 0-1 posloupnost = string: $y = 01100010 \cdots 1$
- Kolmogorovská složitost:

$$\mathfrak{K}(y) = \text{délka nejkratšího programu, jenž } y \text{ vytiskne}$$

- **Příklad:**
  - string $y = 01010101 \cdots 01$ délky $2m$
  - $\mathfrak{K}(y) = O(\log m)$
  - *Důkaz:* **for** $i := 1$ **to** m; print(01); **end**.

Negativní důkazy

- Konečná 0-1 posloupnost = string: $y = 01100010 \cdots 1$
- Kolmogorovská složitost:

$$\mathfrak{K}(y) = \text{délka nejkratšího programu, jenž } y \text{ vytiskne}$$

- **Příklad:**
  - string $y = 01010101 \cdots 01$ délky $2m$
  - $\mathfrak{K}(y) = O(\log m)$
  - *Důkaz:* **for** $i := 1$ **to** m; print(01); **end**.
- **Věta:** Pro každé m existuje string y délky m t.ž. $\mathfrak{K}(y) \geqslant m$.
- Neformálně: y je „nestlačitelný“, „algoritmicky náhodný“ string

Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$

Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$

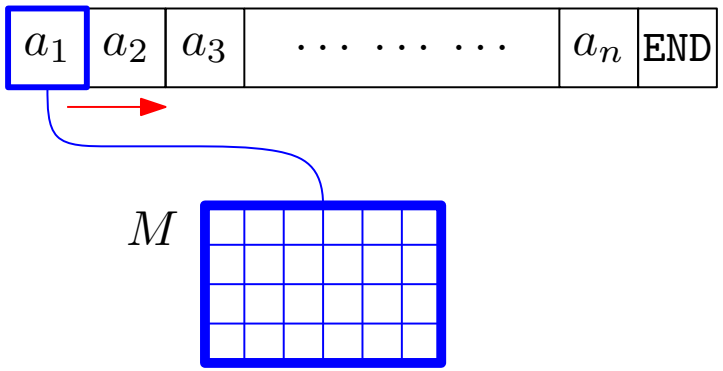
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



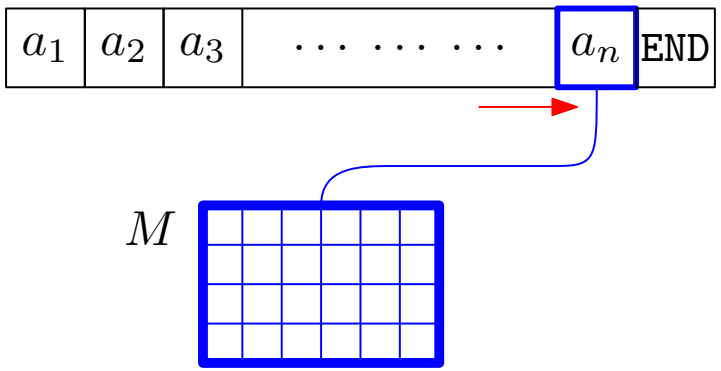
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



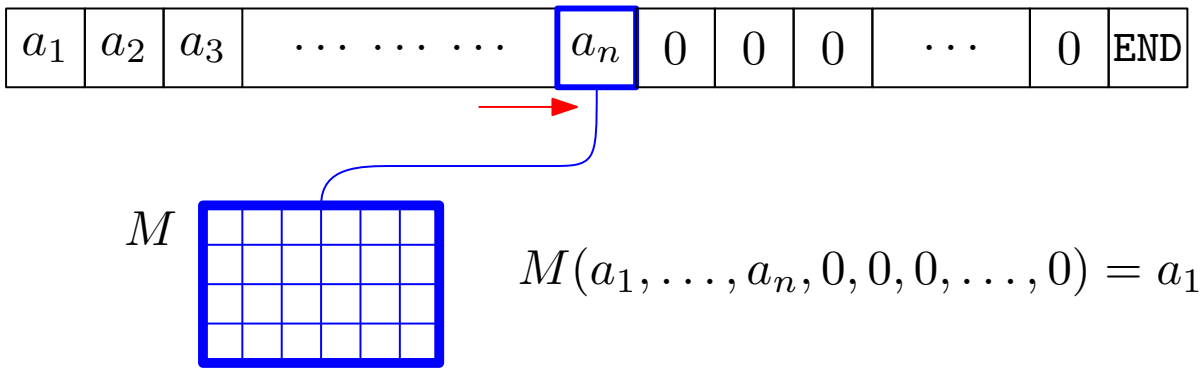
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



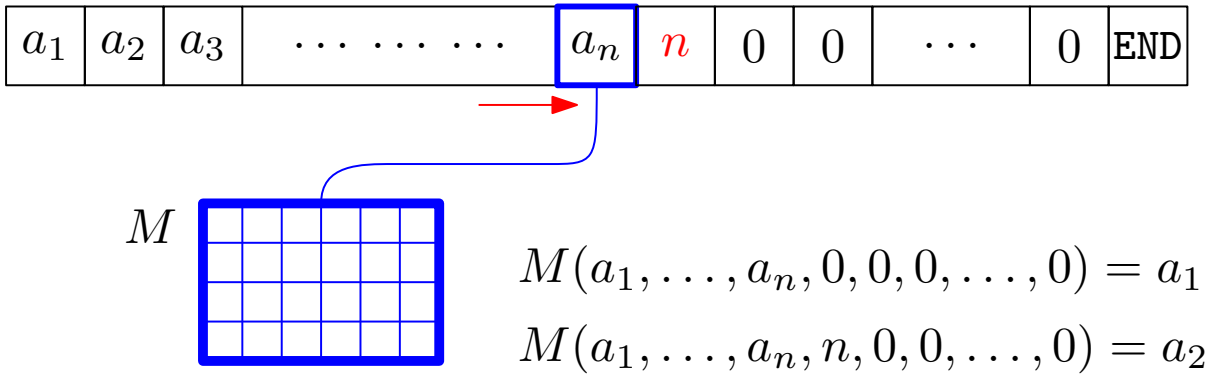
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



$$M(a_1, \ldots, a_n, 0, 0, 0, \ldots, 0) = a_1$$
$$M(a_1, \ldots, a_n, n, 0, 0, \ldots, 0) = a_2$$

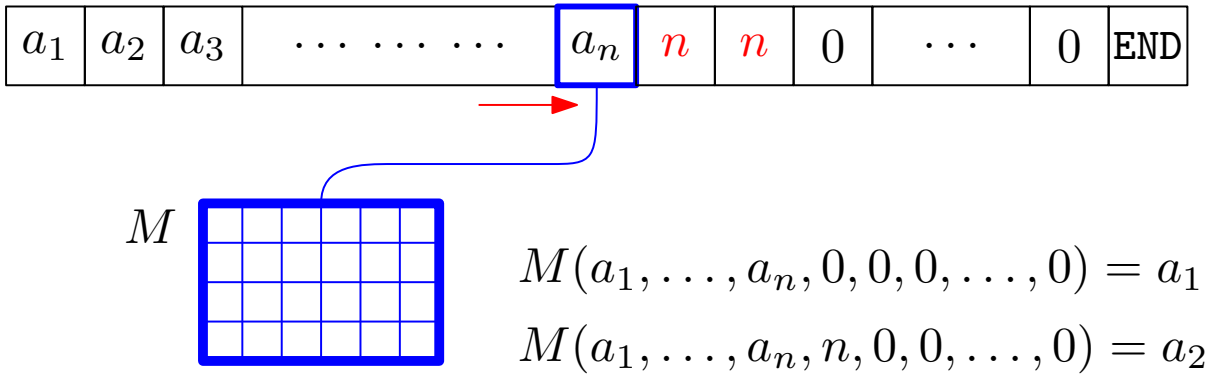
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



$$M(a_1, \ldots, a_n, 0, 0, 0, \ldots, 0) = a_1$$
$$M(a_1, \ldots, a_n, n, 0, 0, \ldots, 0) = a_2$$
$$M(a_1, \ldots, a_n, n, n, 0, \ldots, 0) = a_3$$

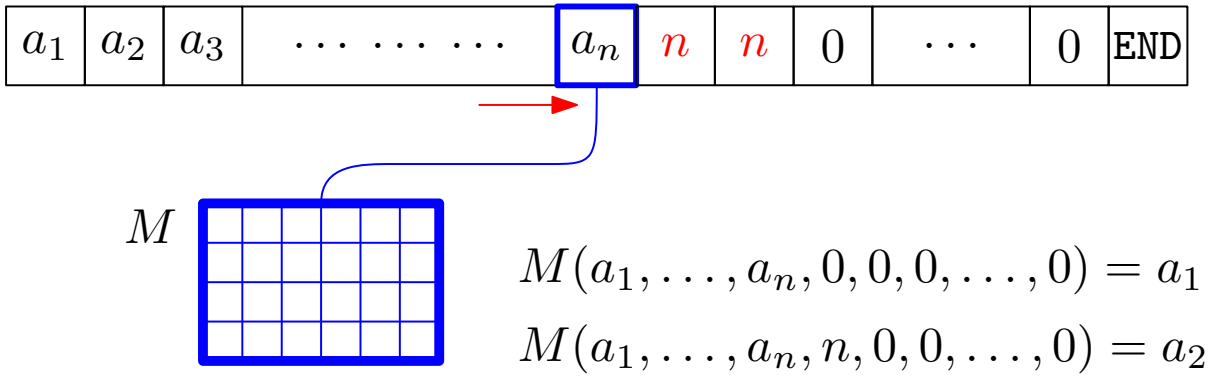
Proč nelze spočítat medián

- **Sporem.** Nechť kdosi sestrojí stream-algo

$$M(a_1, \ldots, a_n) = \text{median}(a_1, \ldots, a_n).$$

- Vezměme $y = y_1\, y_2\, y_3\, \ldots\, y_n$ t.ž. $\mathfrak{K}(y) \geqslant n$
- Vytvořme data

$$a_i = y_1 + y_2 + \cdots + y_i, \quad i = 1, \ldots, n$$



$$M(a_1, \ldots, a_n, 0, 0, 0, \ldots, 0) = a_1$$
$$M(a_1, \ldots, a_n, n, 0, 0, \ldots, 0) = a_2$$
$$M(a_1, \ldots, a_n, n, n, 0, \ldots, 0) = a_3$$

- Právě jsme popsali algo prokazující $\mathfrak{K}(y) \leqslant \log^{O(1)} n$. Spor.

Co negativní výsledky říkají?

Je rozumné klást si otázku:

- Které (běžné) statistické procedury lze stremovaně vyčíslit přesně?
- Důkaz, že to nejde: alibi pro přibližné počítání (např. subsampling)
- Jde-li to: data-reduction techniky nejsou potřeba

Několik pozitivních pozorování

- Regrese $y = X\theta + \varepsilon$

Několik pozitivních pozorování

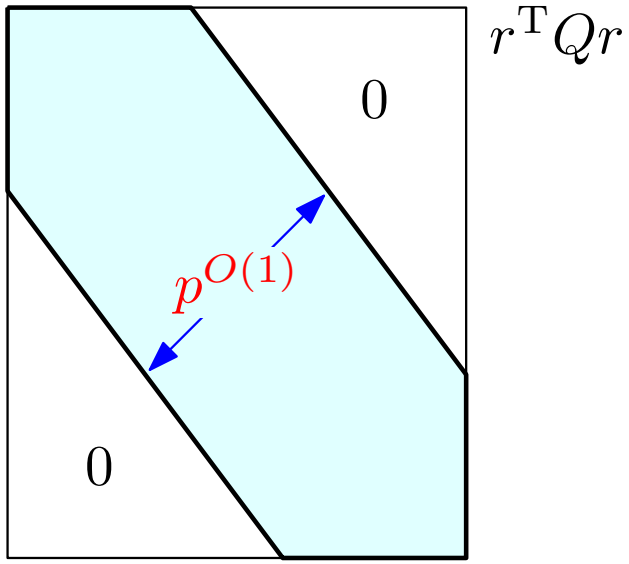
- Regrese $y = X\theta + \varepsilon$
- OLS: $\widehat{\theta} = (X^T X)^{-1} X^T y$ . . . **LZE**

Několik pozitivních pozorování

- Regrese $y = X\theta + \varepsilon$
- OLS: $\widehat{\theta} = (X^T X)^{-1} X^T y$ ... **LZE**
- Ale: vektor residuí $r = y - X\widehat{\theta}$ má délku $n \rightarrow$ nevejde se do paměti

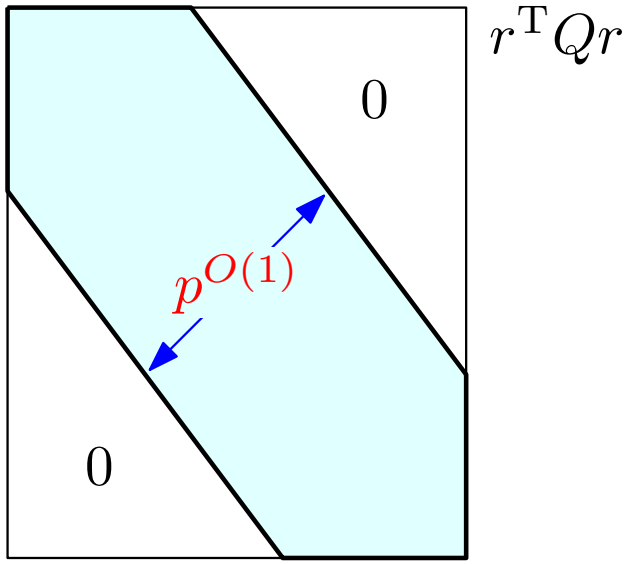
Několik pozitivních pozorování

- Regrese $y = X\theta + \varepsilon$
- OLS: $\widehat{\theta} = (X^T X)^{-1} X^T y$ ... **LZE**
- Ale: vektor residuí $r = y - X\widehat{\theta}$ má délku $n \to$ nevejde se do paměti
- Kvadratická forma v r ... **LZE, je-li dost řídká**



Několik pozitivních pozorování

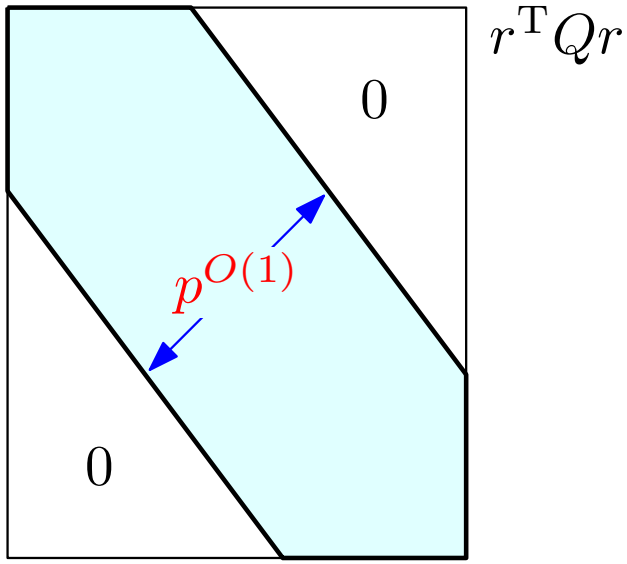
- Regrese $y = X\theta + \varepsilon$
- OLS: $\widehat{\theta} = (X^T X)^{-1} X^T y$ ... **LZE**
- Ale: vektor residuí $r = y - X\widehat{\theta}$ má délku $n \to$ nevejde se do paměti
- Kvadratická forma v r ... **LZE, je-li dost řídká**



- RSS, Durbin-Watson, F-testy, Chow ... **LZE**

Několik pozitivních pozorování

- Regrese $y = X\theta + \varepsilon$
- OLS: $\widehat{\theta} = (X^T X)^{-1} X^T y$ ... **LZE**
- Ale: vektor residuí $r = y - X\widehat{\theta}$ má délku $n \to$ nevejde se do paměti
- Kvadratická forma v r ... **LZE, je-li dost řídká**



- RSS, Durbin-Watson, F-testy, Chow ... **LZE**
- Momenty r řádu $O(1)$ ... **LZE**
  - $\to$ Jarque-Bera ... **LZE**

Občas je to překvapivé...

Dvoustupňové regrese

- **Základní regrese:** $y = X\theta + \varepsilon \rightarrow$ OLS-residua r

Občas je to překvapivé...

Dvoustupňové regrese

- **Základní regrese:** $y = X\theta + \varepsilon \rightarrow$ OLS-residua r
- **Pomocná regrese:** $z(X, y, r) = W(X, y, r)\psi + \delta \rightarrow$ OLS-residua u
- t-test, F-test, LM-test nad u

Občas je to překvapivé...

Dvoustupňové regrese

- **Základní regrese:** $y = X\theta + \varepsilon \rightarrow$ OLS-residua r
- **Pomocná regrese:** $z(X, y, r) = W(X, y, r)\psi + \delta \rightarrow$ OLS-residua u
- t-test, F-test, LM-test nad u

Dva důležité příklady

- Whiteův test na heteroskedasticitu ... **LZE, je-li stupeň Taylorova polynomu v** $W(X, y, r)$ **nanejvýš** $p^{O(1)}$

Občas je to překvapivé...

Dvoustupňové regrese

- **Základní regrese:** $y = X\theta + \varepsilon \rightarrow$ OLS-residua r
- **Pomocná regrese:** $z(X, y, r) = W(X, y, r)\psi + \delta \rightarrow$ OLS-residua u
- t-test, F-test, LM-test nad u

Dva důležité příklady

- Whiteův test na heteroskedasticitu ... **LZE, je-li stupeň Taylorova polynomu v** $W(X, y, r)$ **nanejvýš** $p^{O(1)}$
- Breush-Godfreyův test na AR(s) v disturbancích: **LZE, je-li** $s = p^{O(1)}$

Závěr

Poznámky na závěr

- Lze-li stremovaně počítat přesně, dělejme to (redukce dat vede ke zbytečné ztrátě informace)
- Chceme-li počítat nepřesně, nejprve si dokažme větu, že přesně to nelze
- Zajímavé jsou např. changepoint statistiky typu

$$\max_k \frac{RSS_{1:n} - RSS_{1:k} - RSS_{k+1:n}}{RSS_{1:k} + RSS_{k+1:n}}$$

Děkuji za diskuse

- T. Cipra: rekursivní procedury v autoregresních procesech
- V. Holý: streamované počítání nad vysokofrekvenčními daty
- J. Antoch: za Robust...

Další otázky

- Jaký je přirozený výpočetní model pro Wide Big Data?

Poděkování

Děkuji Vám za pozornost.

Přeji organizátorům ISI WSC 2019, aby se jim podařilo zorganizovat skvělou konferenci.