CISS, Princeton University, 3/19/2014

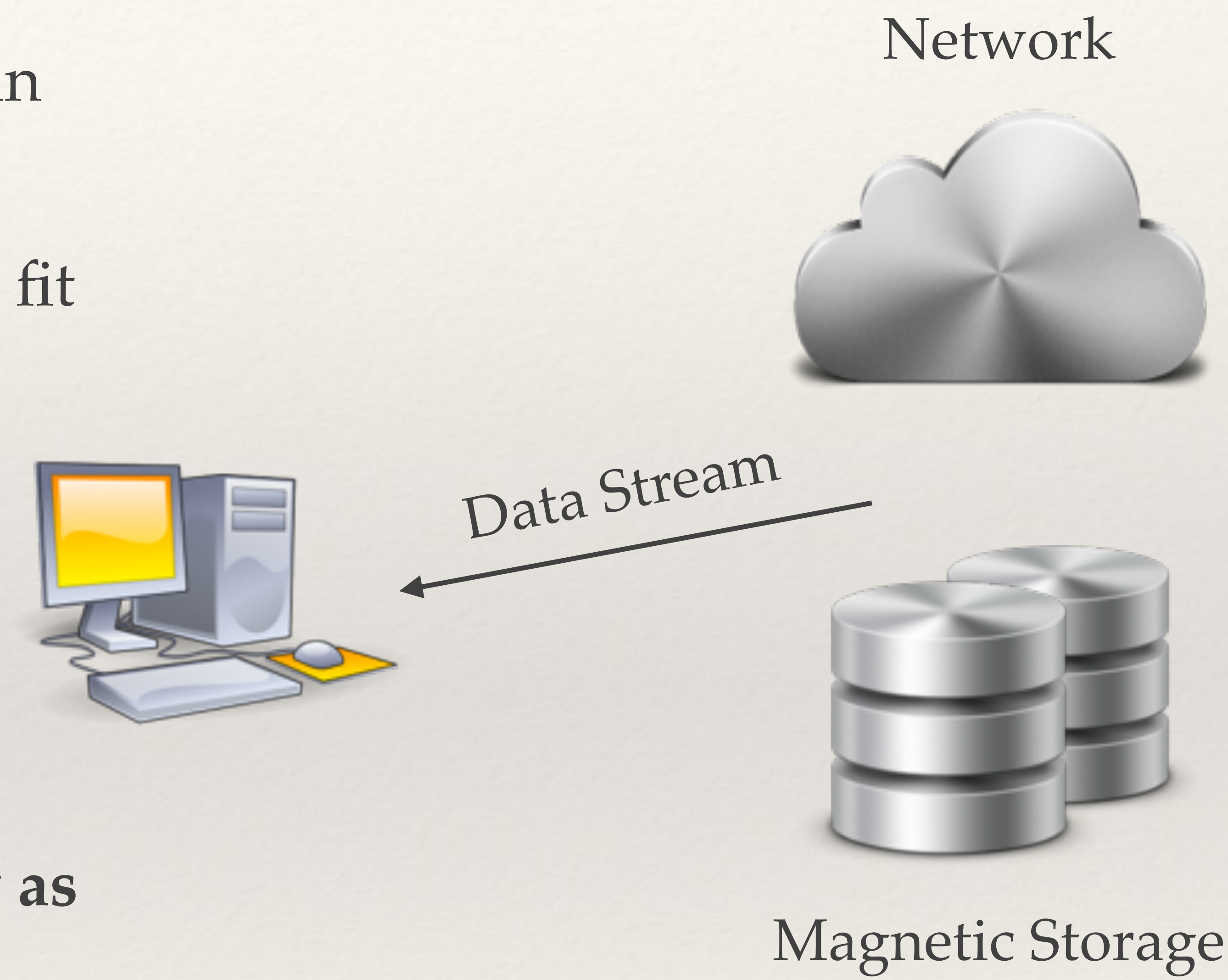
Approximate Sorting of Data Streams with Limited Storage

Farzad Farnoud
Eitan Yakoobi
Jehoshua Bruck

Caltech

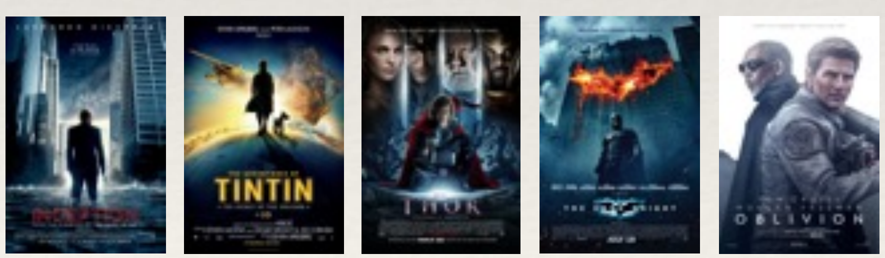
Sorting with Limited Storage

- Sorting is a fundamental operation in data processing
- Data maybe so large that it does not fit in storage and must be sequentially accessed:
  - Streamed data from network
  - Data stored on magnetic storage
- **Not to rearrange data but to approximate its ordering as closely as possible**
- Study of relationship between quality of sorting and available storage



Learning Preference Rankings

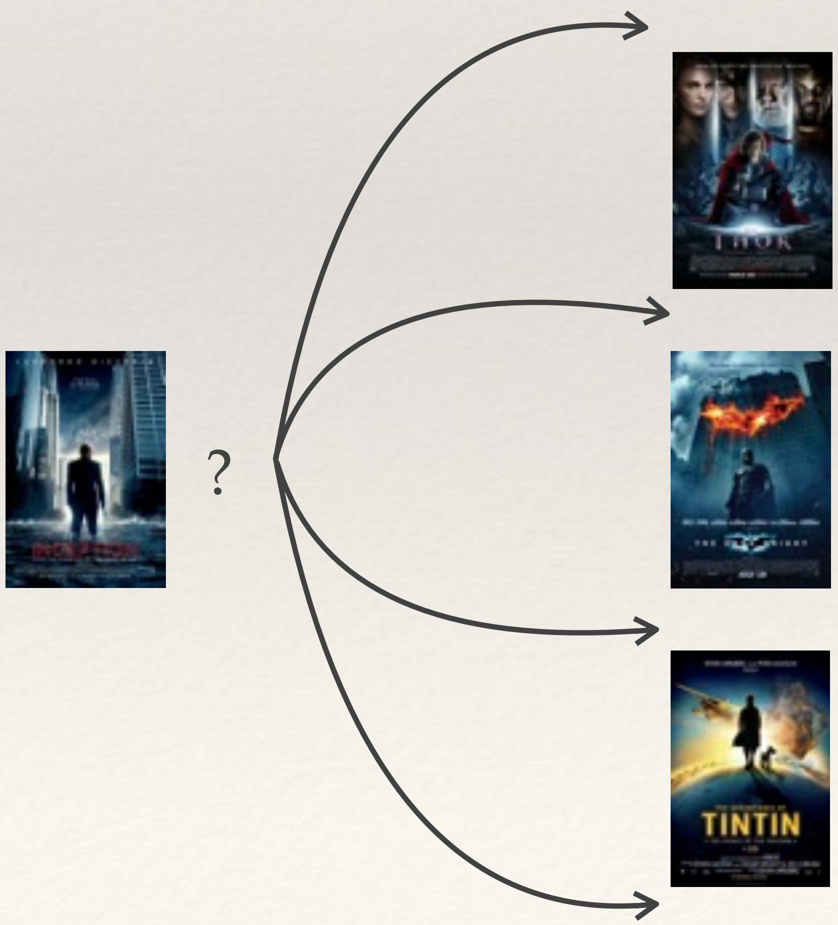
- With minor modification the same setting exists in the context of obtaining a user's ranking of objects that are presented one by one
- User's ranking is useful for recommendation and collaborative filtering
- User can remember only a small number of movies they watched



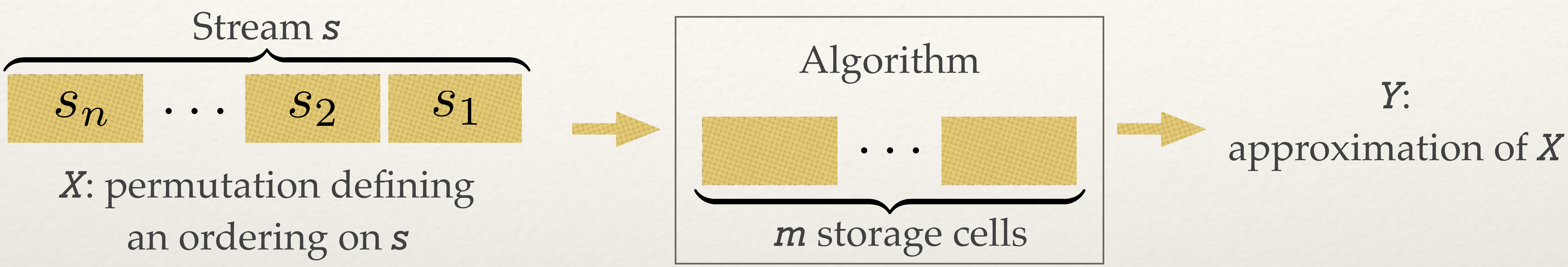
Ranking of movies

Learning Preference Rankings

- With minor modification the same setting exists in the context of obtaining a user's ranking of objects that are presented one by one
- User's ranking is useful for recommendation and collaborative filtering
- User can remember only a small number of movies they watched



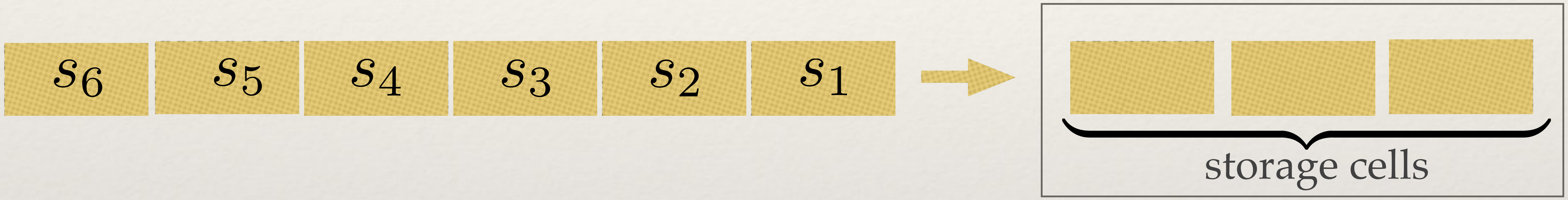
Problem Statement



- If i appears before j in X, then $s_i < s_j$
- To store stream elements, m cells are available; no limitation on other types of storage
- Algorithm can compare any two elements residing in storage
- Deterministic algorithms, X is a random permutation
- Performance measure: *Mutual information* and *distortion* between X and Y

Example

- Suppose X=253461 and m=3



$s_2 < s_1$ $\quad$ $s_2 < s_3 < s_1$ $\quad$ $s_2 < s_3 < s_4$ $\quad$ $s_2 < s_5 < s_4$ $\quad$ $s_2 < s_4 < s_6$

- Output, e.g. Y=235146

Related Work

- J. Munro and M. Paterson. Selection and sorting with limited storage. Theoretical Computer Science, 12(3):315–323, 1980.
- G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited storage. ACM SIGMOD 1998
- Sudipto Guha and Andrew McGregor. Approximate quantiles and the order of the stream. In Proc. 25th ACM Symposium on Principles of Database Systems, pp. 273– 279, 2006.
- A. Chakrabarti, T. S. Jayram, and M. Patrascu. Tight lower bounds for selection in randomly ordered streams. SODA 2008

Performance Measures

- Mutual Information between X and Y
- *Kendall tau* distortion:
  - Counts the number of *pairwise mistakes*
  - \# *transpositions of adjacent elements* taking X to Y
  - Example: $d_\tau(312,123)=2$ since $312 \to 132 \to 123$
- *Weighted Kendall* distortion
- *Chebyshev* distortion

Performance Measures

- Mutual Information between X and Y
- *Kendall tau* distortion
- *Weighted Kendall* distortion:
  - Weight w_i for transposing ith and $(i+1)$st elements
  - Can be used to penalize mistakes in higher positions more
  - Example: $w_1 = 2$, $w_2 = 1$, $d_w(312, 123) = 3$ since $312 \to 132 \to 123$
- *Chebyshev* distortion

Performance Measures

- Mutual Information between *X* and *Y*
- *Kendall tau* distortion
- *Weighted Kendall* distortion
- *Chebyshev* distortion:
  - Also known as l_∞
  - Maximum error in the rank of any element
  - Example: $d_c(35124,12345)=3$

Universal Bounds: Mutual Information

Theorem: For an algorithm that maximizes mutual information, we have

$$\frac{I(X;Y)}{H(X)} \sim \frac{\lg m}{\lg n}$$

In particular, if $m=n^c$, we have $I(X;Y)/H(X) \sim c$

Proof of upper bound:

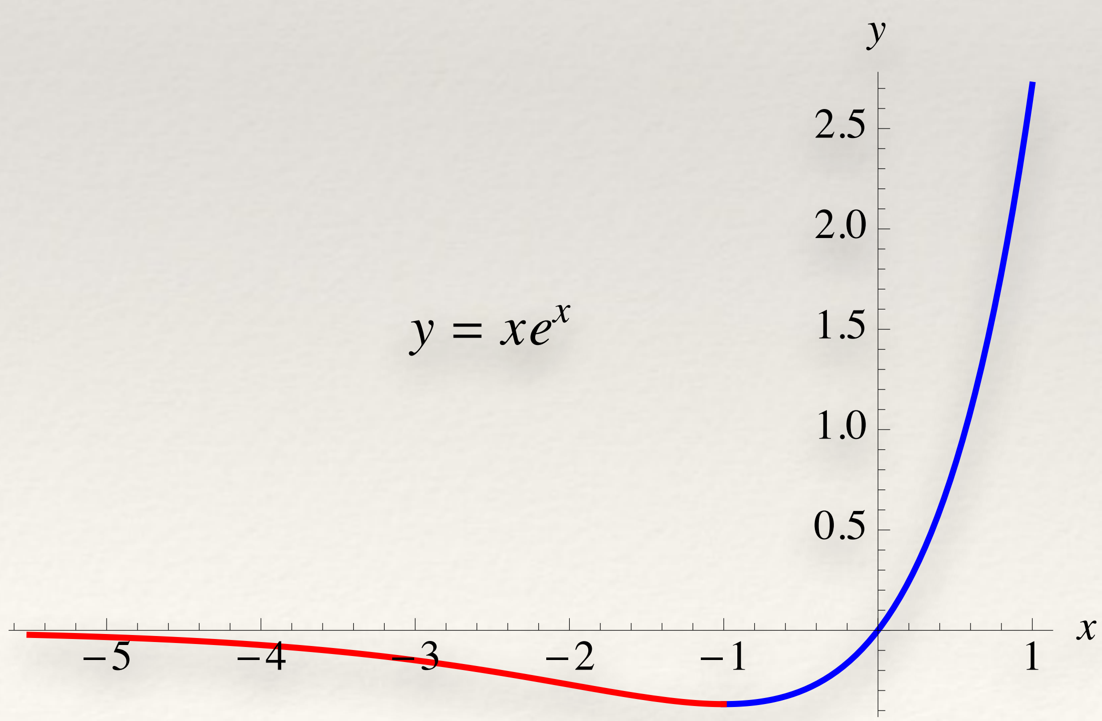
Consider the amount of information obtained by the algorithm:

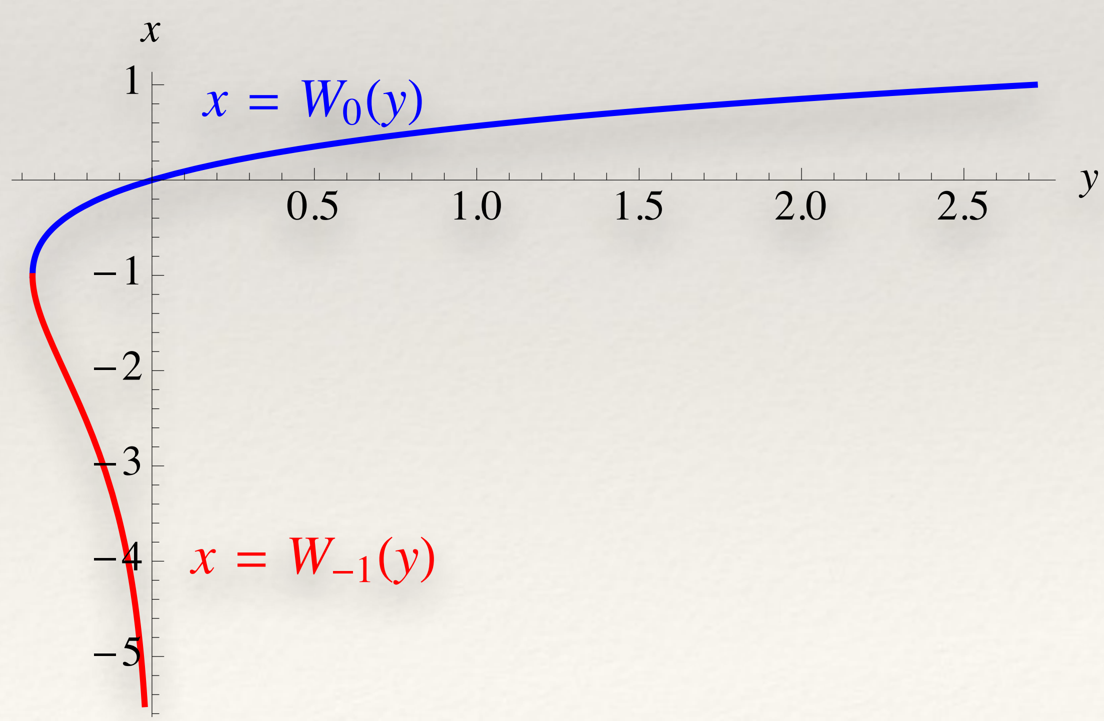
- Each new element is compared with m-1 elements → lg(m) bits
- $I(X;Y) \leq n \lg(m)$, $H(X) \sim n \lg(n)$

Universal Bounds: Kendall Distortion

Theorem: For any algorithm with storage μn and average Kendall distortion δn, if δ is bounded away from zero, then

$$\mu \geq -W_0\left(\frac{-\delta^\delta}{e(1+\delta)^{1+\delta}}\right)(1+o(1))$$



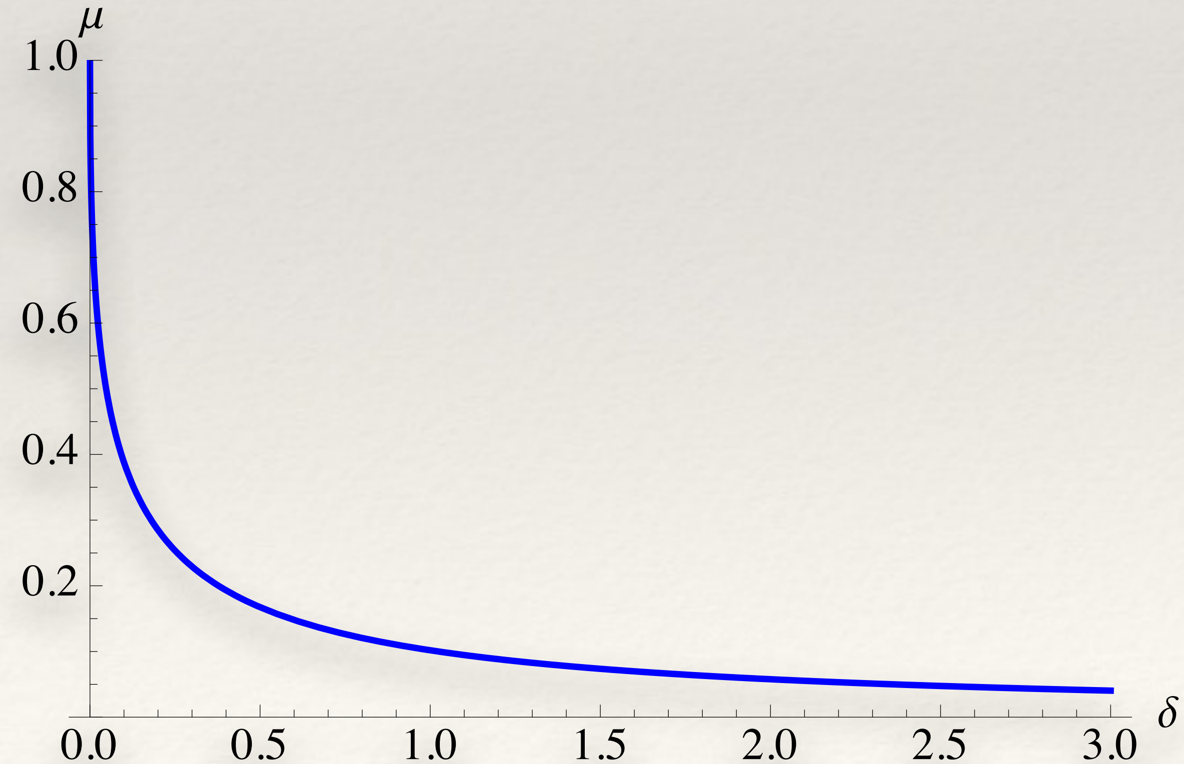


Universal Bounds: Kendall Distortion

Theorem: For any algorithm with storage μn and average Kendall distortion δn, if δ is bounded away from zero, then

$$\mu \geq -W_0\left(\frac{-\delta^\delta}{e(1+\delta)^{1+\delta}}\right)(1+o(1))$$

- As δ increases, we asymptotically have $\mu \geq 1/(e^2\delta)(1+o(1))$



Universal Bounds: Kendall Distortion

Proof outline:

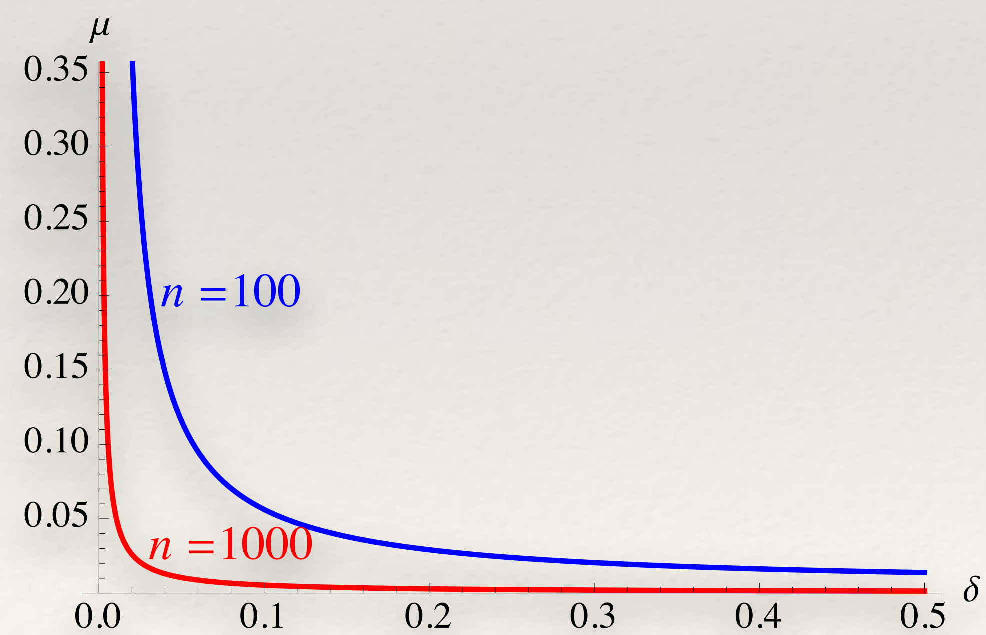
- The number M of outputs of any algorithm is bounded as $M \leq m!(n-m)^m$
- Set of outputs can be viewed as a covering code
- From rate-distortion on permutations [Wang et al. 2013, Farnoud et al. 2014], we find a lower bound on M with respect to δ

Universal Bounds: Chebyshev Distortion

Theorem: For any algorithm with storage μn and average Chebyshev distortion δn, with $2/n \leq \delta \leq 1/2$,

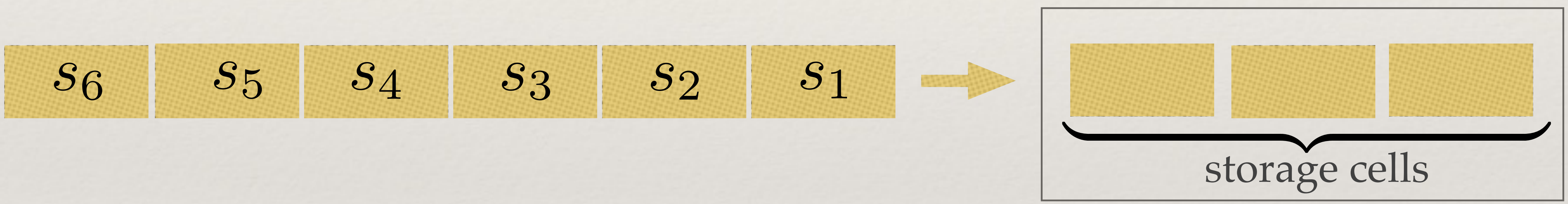
$$\mu \geq -W_0\left(\frac{-(e/2)^{2\delta}}{2\delta n}\right)(1+o(1))$$

- For any fixed δ as n increases, storage requirement becomes a vanishing fraction of n.
- Constant distortion needs at least constant μ



Algorithm

- A simple algorithm:
  - Store the first m-1 elements of the stream, $s_1,\ldots,s_{m-1}$, as *pivots*
  - Compare each new element with the pivots
- Example: Suppose X=253416 and m=3:



$s_2<s_1$ $\quad$ $s_2<s_3<s_1$ $\quad$ $s_2<s_4<s_1$ $\quad$ $s_2<s_5<s_1$ $\quad$ $s_2<s_6<s_1$

- Output Y=234516,
  d_τ(253416,234516)=2, d_c(253416,234516)=2

Algorithm: Performance

Theorem: In terms of mutual information, the algorithm is asymptotically optimal.

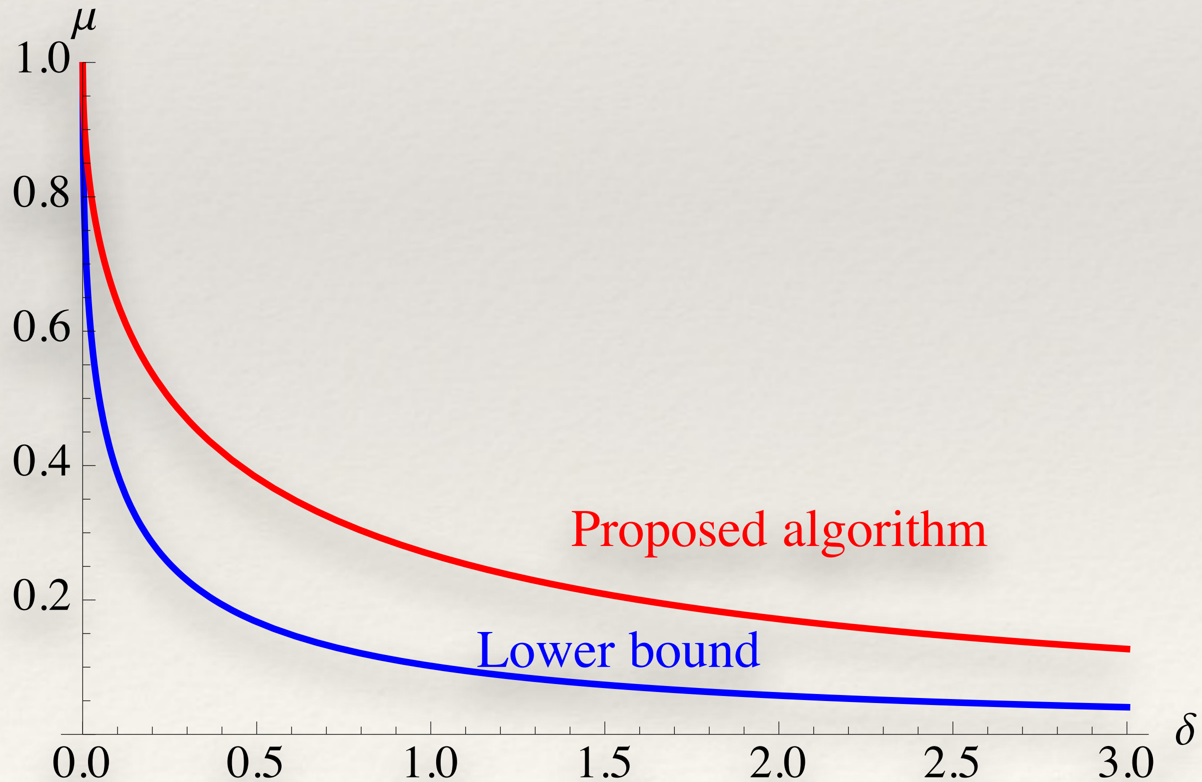
Proof outline:

- Given Y, the permutation X is unknown only in segments bounded by pivots: If Y=234156, then $X \in \{234156, 243156, 234165, 243165\}$
- We write $H(X|Y)$ as a combinatorial sum, bound as $H(X|Y) \leq n \lg(n/m) + O(n)$
- $I(X;Y) = H(X) - H(X|Y) \sim n \lg(m)$, $I(X;Y)/H(X) \sim \lg(m)/\lg(n)$

Algorithm: Performance

Theorem: The algorithm asymptotically requires at most a constant factor as much storage as an optimal algorithm for the same Kendall distortion.

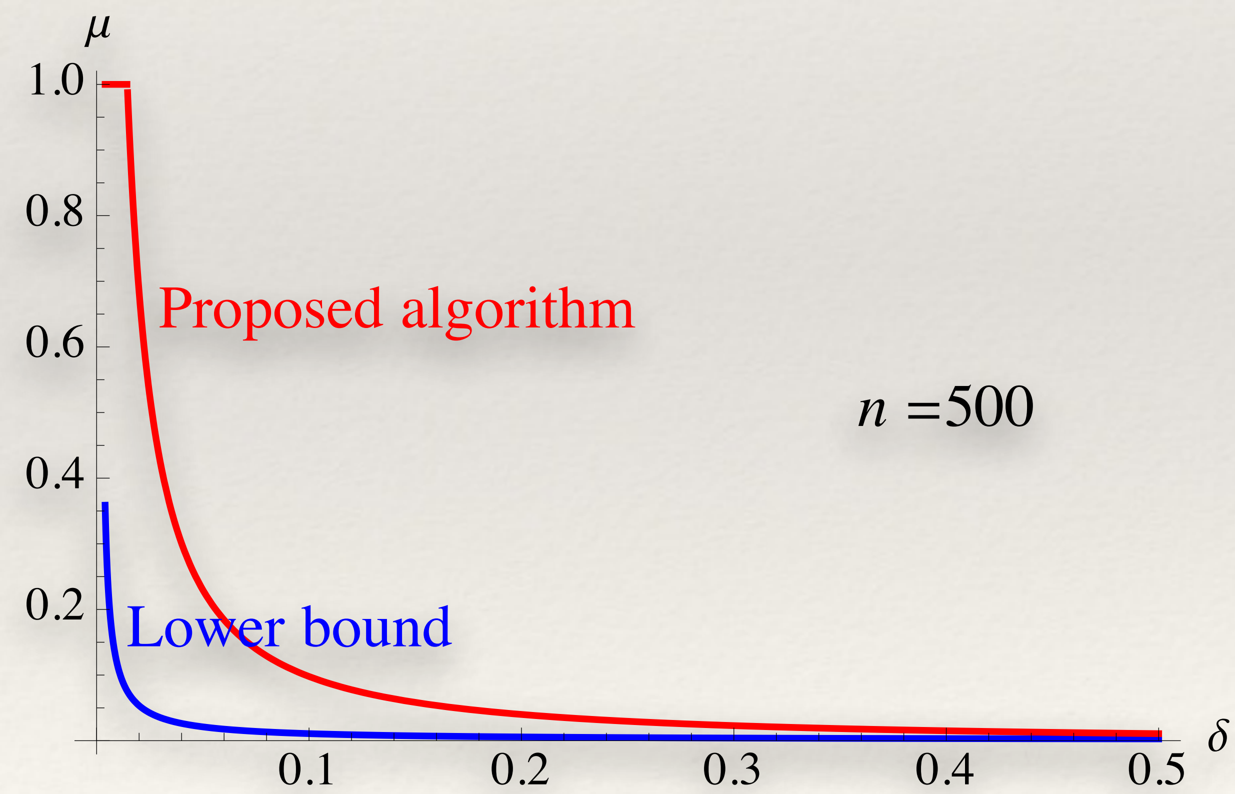
- For large δ, we need $e^2/2 \cong 3.7$ times as much storage.



Algorithm: Performance

Theorem: If the proposed algorithm has storage μn and average Chebyshev distortion δn, with $\delta \leq 1/2$ and δ bounded away from 0, then $\mu \leq W_{-1}(-\delta/e)/(\delta n)$.

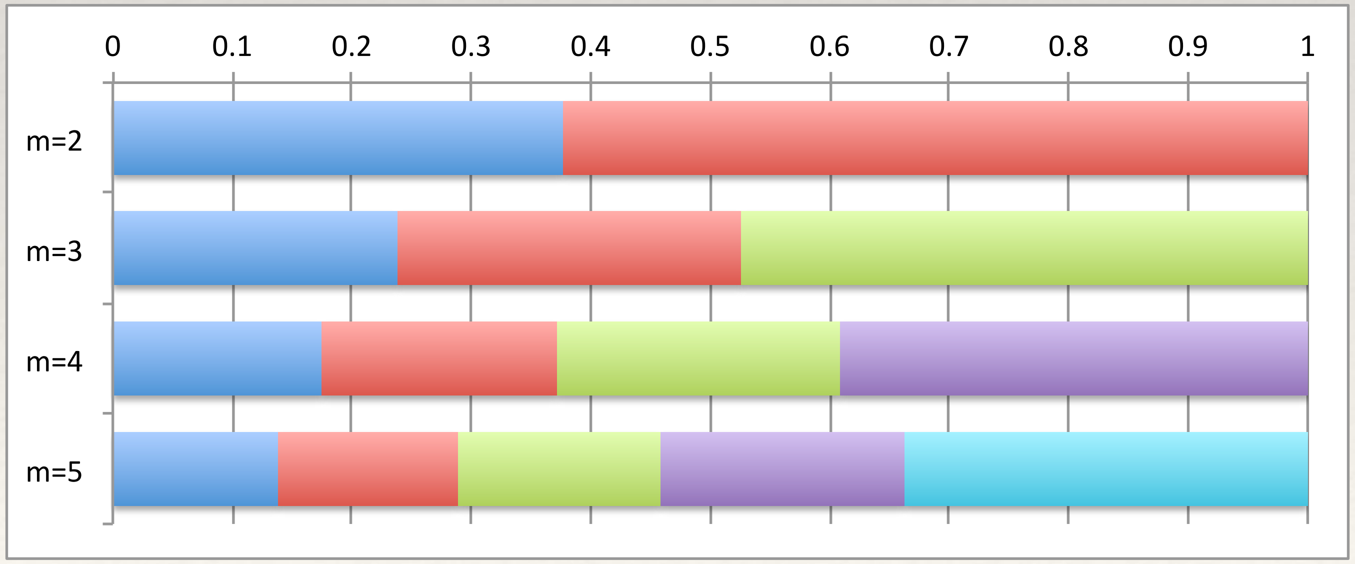
- If δ is bounded away from 0, we need at most a constant times as much storage.
- Since maximum distortion is only n, for vanishing distortion, better algorithm and/or bounds are needed.



Thank You!

Distortion with Weighted Kendall

- What should be the ranks of pivots if errors in higher positions are to be penalized more?
- Use weighted Kendall to model non-uniform importance
- Linearly decreasing weight function: $w_i = 1 + c\,(n-i-1)$:



Remembering last *m* elements

- Finding the best ranking is closely related to the #P-complete problem of *counting the number of linear extensions of a poset*
- Simple algorithm: rank each group of *m* elements and interleave

Theorem: In terms of mutual information, the algorithm is asymptotically optimal. That is, with $m=an^b$, a fraction b of information in X is recovered.

- Better algorithm needed for distortion