Learning-based compressed data structures

Paolo Ferragina, Giorgio Vinciguerra

Università di Pisa

- The deluge of data has made the use of compressed data structures indispensable
- These structures build on two sources of compressibility: statistical properties and repetitiveness of the data
- But there is a new promising kind of regularity to be studied: **approximate linearity**

The predecessor search problem

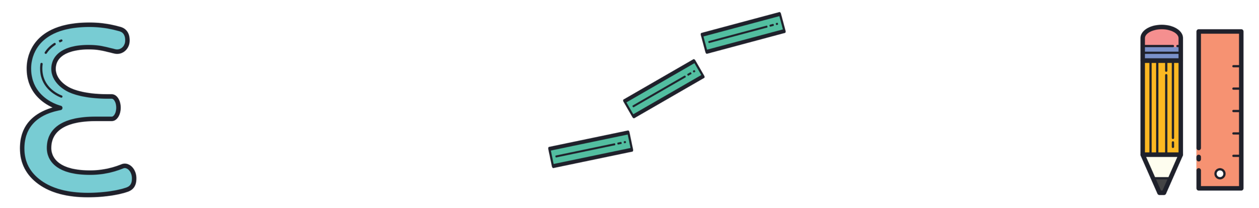
Given n sorted keys implement

$$pred(x) = \text{“largest key} \le x\text{”}$$

Why do we care? Range queries and joins in DBs, conjunctive queries in search engines, IP routing, ...

The PGM-index

pgm.di.unipi.it



Fixed model error ε
Control the size of the search range

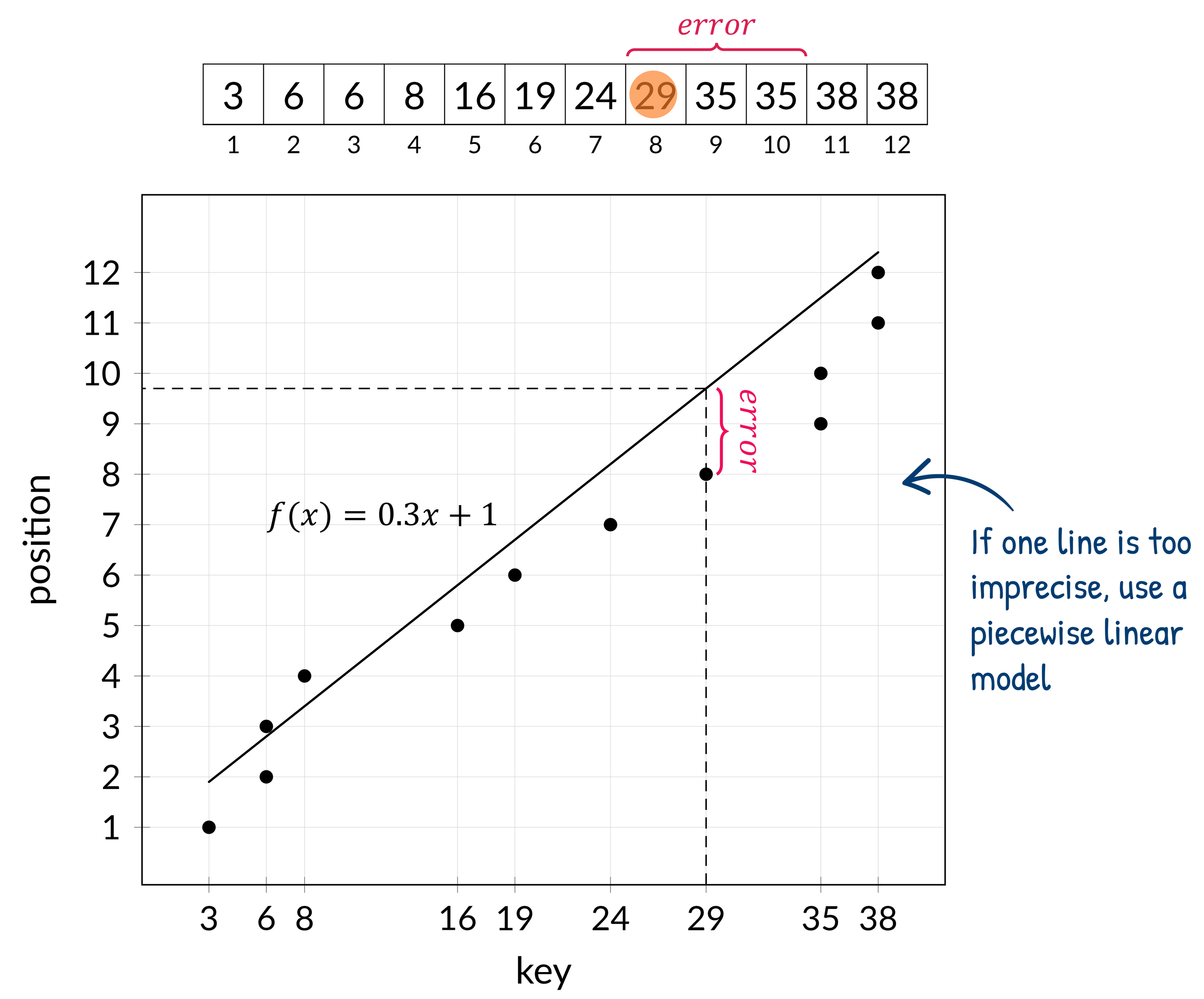
Optimal piecewise linear ε-approx
Fast to construct, captures non-linearities

Recursive design
Adapt to the memory hierarchy

Features

- Optimal time and space complexity guarantees
- Never worse than traditional indexes such as B-trees
- Resistant to adversarial inputs and queries
- Supports multidimensional data and queries
- Auto-tunable to the desired memory usage or query time

Idea: input data as pairs (key, pos)



Experimental results

- As fast as static B-tree, 83× more compressed
- Up to 3× faster than a dynamic B-tree and 1000× more compressed on 10^9 integers
- About 3 seconds to construct on 10^9 integers

Rank/Select dictionaries

Store an integer set S in compressed form, support

$$rank(x) = \text{“\# elements} \le x \text{ in } S\text{”}$$

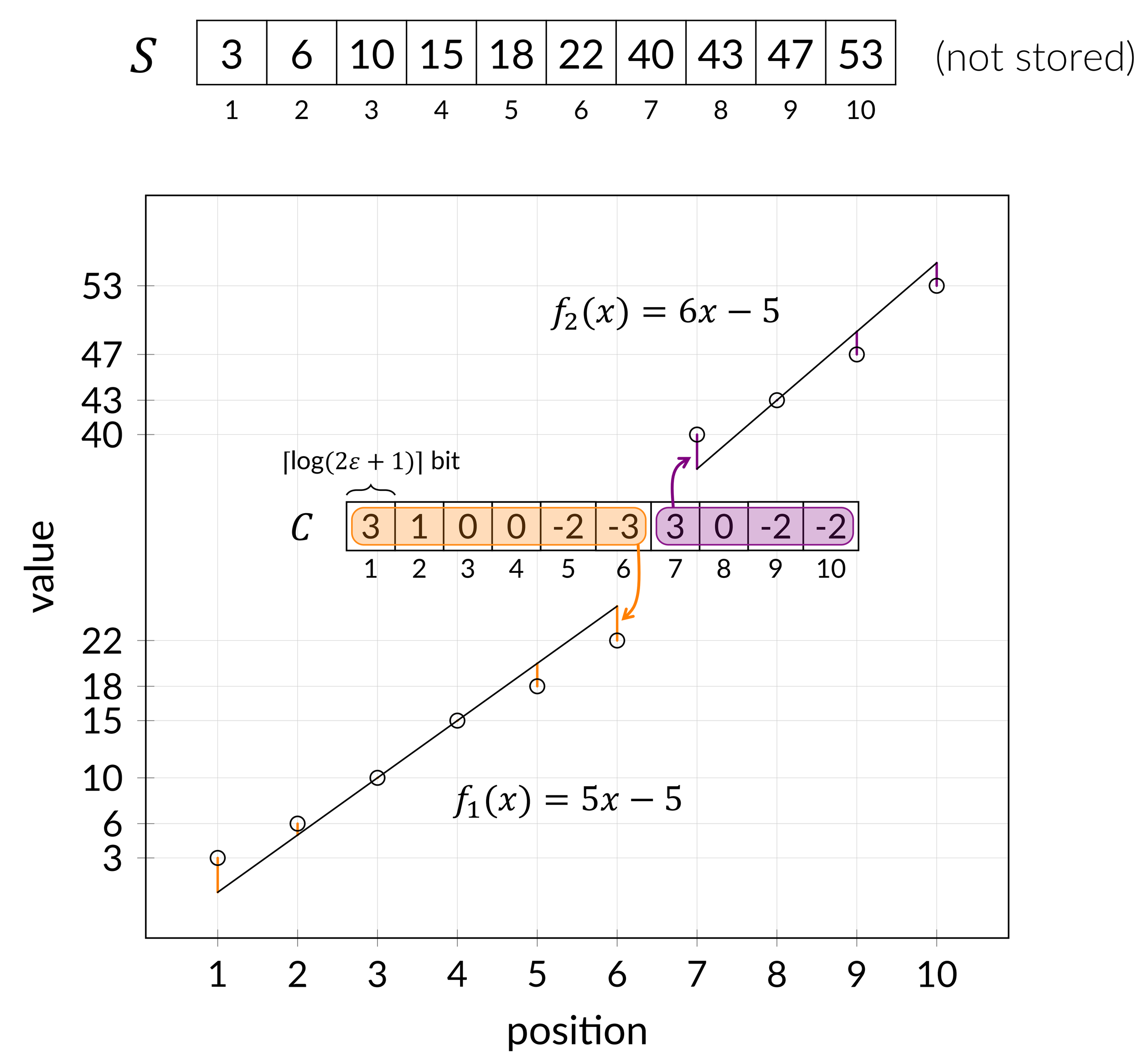
$$select(i) = \text{“}i\text{th smallest element in } S\text{”}$$

Why do we care? Building block of compressed data structures for texts, genomes, graphs, ...

The LA-vector

Orchestrate piecewise linear ε-approximations, corrections and indexing. Faster $select$ and competitive $rank$ wrt well-engineered solutions

Idea: data = segments + corrections



[1] P. Ferragina and G. Vinciguerra. *The PGM-index: a fully-dynamic compressed learned index with provable worst-case bounds*. VLDB, 2020.

[2] P. Ferragina, F. Lillo, and G. Vinciguerra. *Why are learned indexes so effective?* ICML, 2020.

[3] A. Boffa, P. Ferragina, and G. Vinciguerra. *A “learned” approach to quicken and compress rank/select dictionaries*. SIAM ALENEX, 2021.