POLITECNICO DI TORINO

Higher Methods in Data Science and ML

On the encoding of large, high-dimensional and unorganized datasets

Ulderico Fugacci

30 January 2020

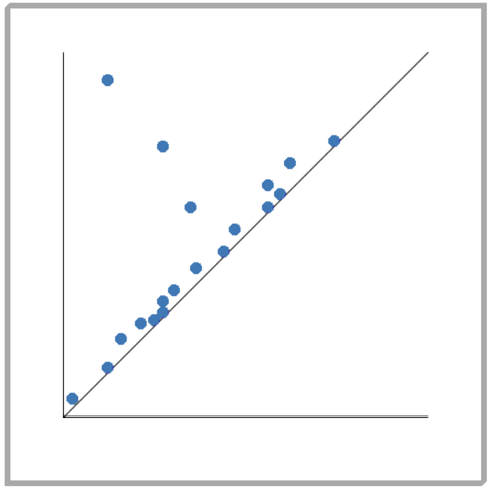
ECCELLENZA 2018-2022

Motivation

Topological Data Analysis (TDA) and Persistent Homology (PH) allow for extracting the core topological information from ***large***, ***high-dimensional*** and ***unorganized datasets***.

E.g. ***point clouds***, ***complex networks***, ***(semi-)metric spaces***

Dataset



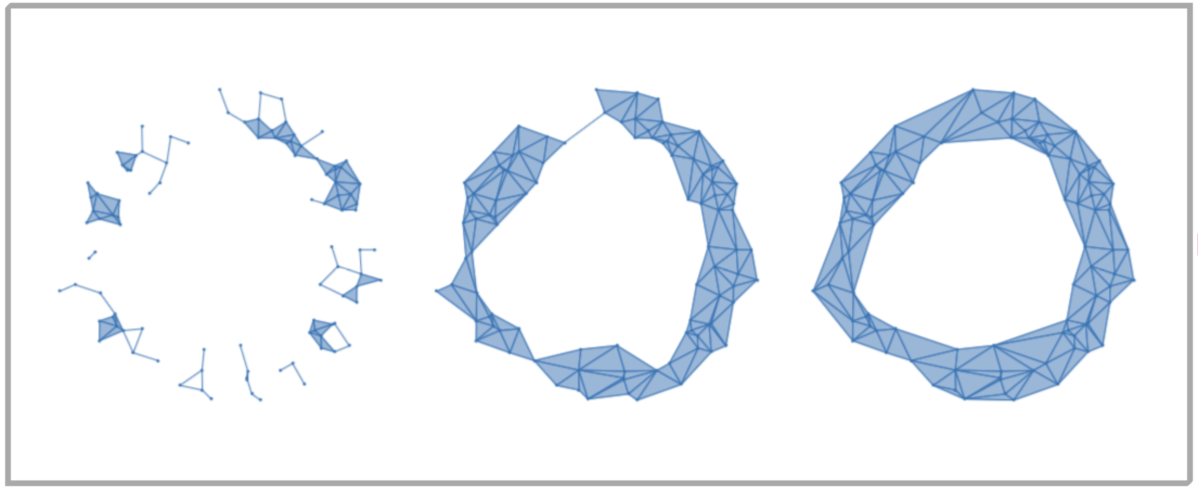
Topological Information

Motivation

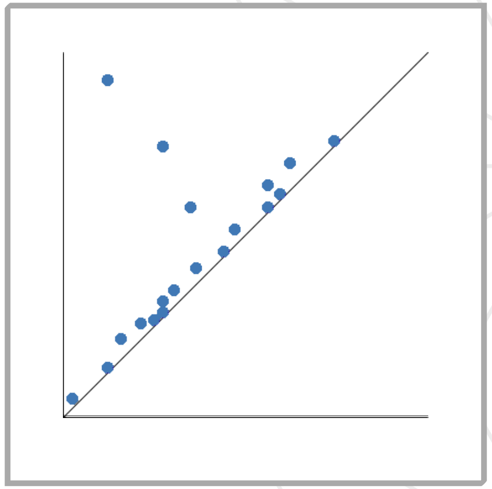
Topological Data Analysis (TDA) and Persistent Homology (PH) allow for extracting the core topological information from ***large***, ***high-dimensional*** and ***unorganized datasets***.

E.g. ***point clouds***, ***complex networks***, ***(semi-)metric spaces***

Dataset



Filtered Simplicial Complex

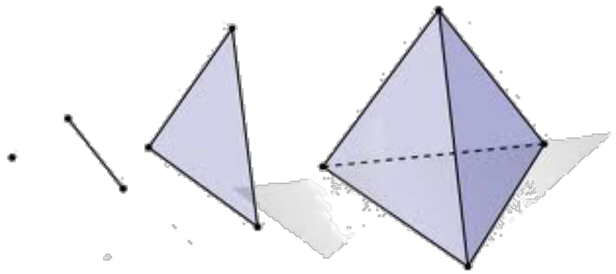


Topological Information

Motivation

Simplicial Complex:

A family K of subsets (called ***simplices***) of a finite set V closed under the operation of taking subsets

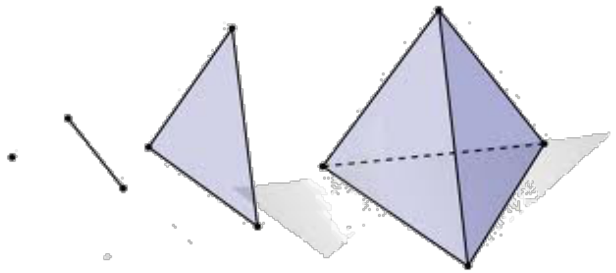


0-, 1-, 2-, 3-simplices

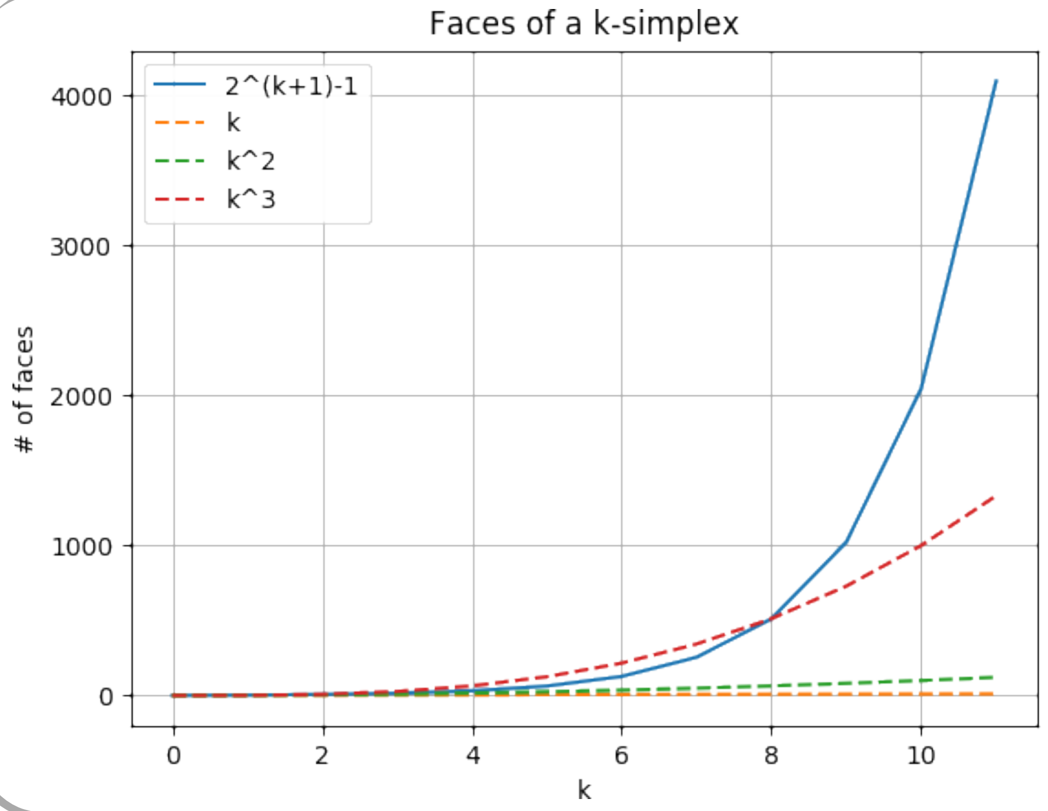
Motivation

Simplicial Complex:

A family K of subsets (called ***simplices***) of a finite set V closed under the operation of taking subsets



0-, 1-, 2-, 3-simplices



Issue:

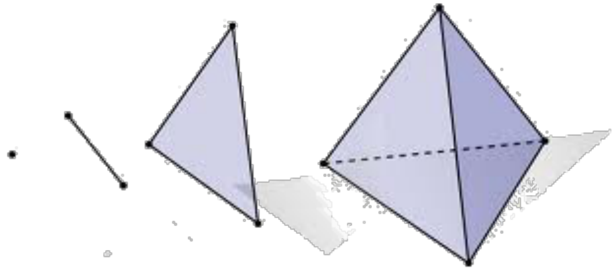
V a point cloud with
$|V| \geq 30$

Filtered complex K associated with V consists of more than a ***billion*** of simplices

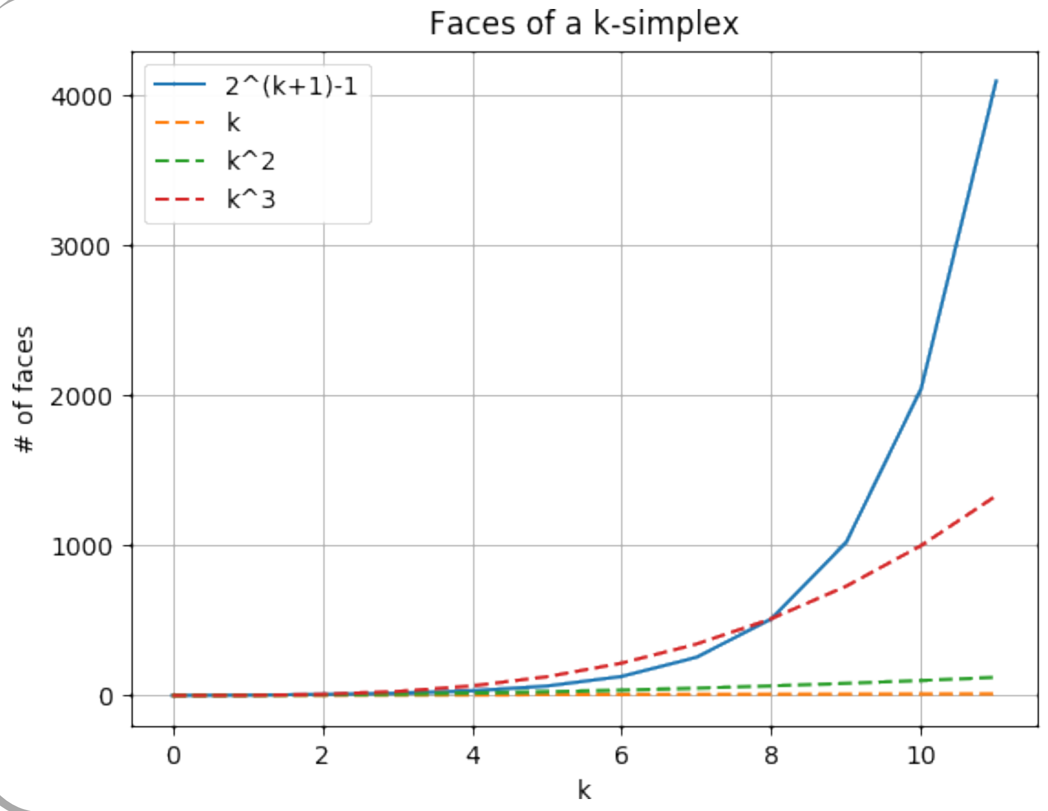
Motivation

Simplicial Complex:

A family K of subsets (called ***simplices***) of a finite set V closed under the operation of taking subsets



0-, 1-, 2-, 3-simplices



Solution:

Development of
compact and efficient data structures
for encoding simplicial complexes

Motivation

POLITECNICO DI TORINO

Outline:

- ***Which info to be stored?***
- ***Data structures:***
  - Simplex-based representations
  - Top-based representations
  - Operator-driven representations
- ***Issues and solutions in adopting top-based representations***
- ***Comparisons***

Motivation

Out of Scope:

- ***Data structures for specific classes of complexes***
  - *E.g., manifolds or complexes of low dimension*
- ***Hierarchical and multi-resolution models***
- ***Construction of a simplicial complex from a dataset***

Data Structures for Simplicial Complexes

POLITECNICO DI TORINO

SmartData

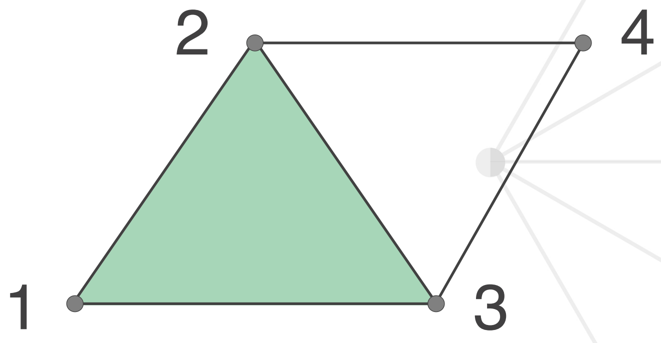
Data structure:

The ***entities*** which a simplicial complex consists of are:

- its ***simplices***

$$K = K_0 \cup K_1 \cup \ldots \cup K_d$$

where K_i is the collection of the i-simplices of K



- the ***topological relations***

$$R_{i,j} \subseteq K_i \times K_j$$

between the simplices of K encoding the (co-)boundary of each simplex

Data Structures for Simplicial Complexes

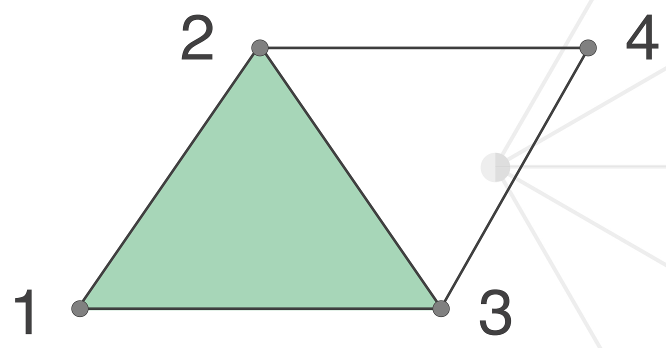
Data structure:

The ***entities*** which a simplicial complex consists of are:

- its ***simplices***

$$K = K_0 \cup K_1 \cup \ldots \cup K_d$$

where K_i is the collection of the i-simplices of K



- the ***topological relations***

$$R_{i,j} \subseteq K_i \times K_j$$

between the simplices of K encoding the (co-)boundary of each simplex

A ***data structure*** for K has to explicitly ***store*** a portion of the above information and to (efficiently) ***retrieve*** the remaining part

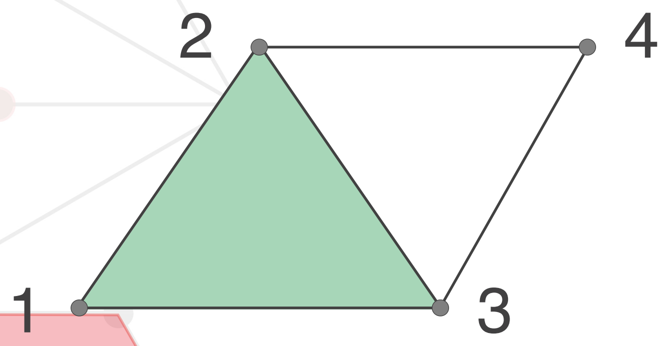
Data Structures for Simplicial Complexes

Topological Relations:

Given a i-simplex σ and a j-simplex τ of K,

$$(\sigma, \tau) \in R_{i,j} \iff \begin{cases} \sigma \subseteq \tau & \text{for } i < j \\ |\sigma \cap \tau| = i \quad (\text{equivalently, } \sigma \cap \tau \in K_{i-1}) & \text{for } i = j \\ \tau \subseteq \sigma & \text{for } i > j \end{cases}$$

Example:



$(12, 123) \in R_{1,2}$ $(12, 24) \in R_{1,1}$ $(12, 1) \in R_{1,0}$

Data Structures for Simplicial Complexes

POLITECNICO DI TORINO

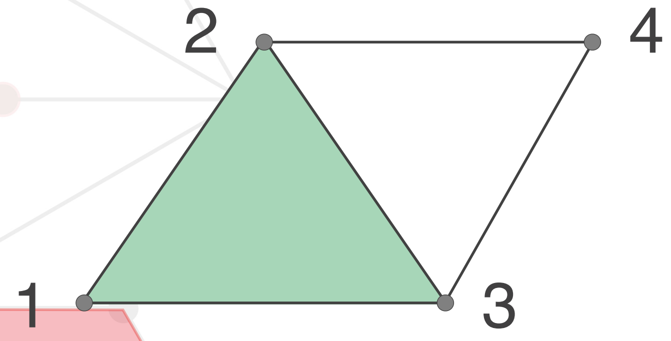
SmartData

Topological Relations:

Given a i-simplex σ and a j-simplex τ of K,

$$(\sigma, \tau) \in R_{i,j} \iff \begin{cases} \sigma \subseteq \tau & \text{for } i < j \\ |\sigma \cap \tau| = i \quad (\text{equivalently, } \sigma \cap \tau \in K_{i-1}) & \text{for } i = j \\ \tau \subseteq \sigma & \text{for } i > j \end{cases}$$

Example:



$(12, 123) \in R_{1,2}$ $\quad$ $(12, 24) \in R_{1,1}$ $\quad$ $(12, 1) \in R_{1,0}$

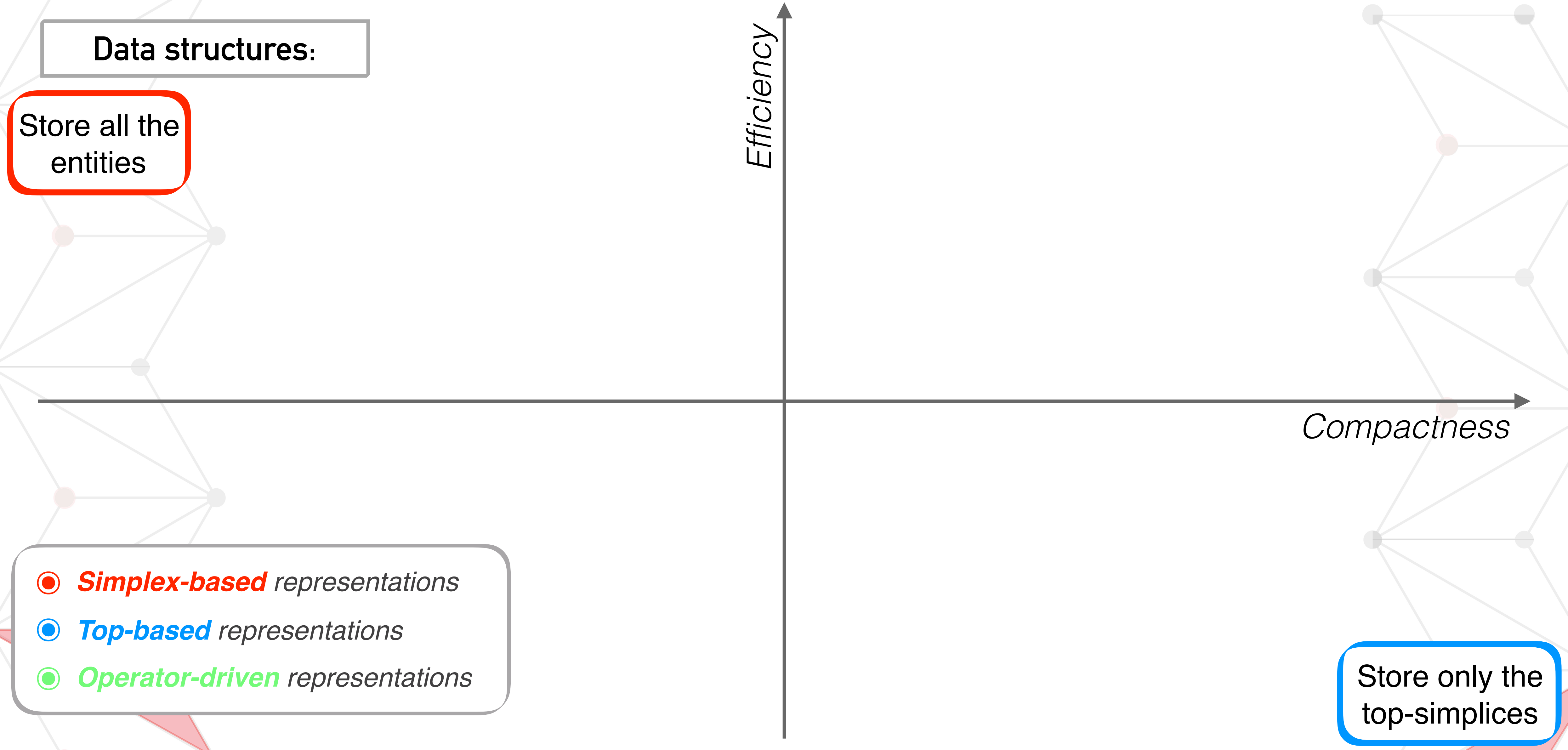
An i-simplex σ is called a ***top simplex*** of K if $R_{i,i+1}$ is empty

Data Structures for Simplicial Complexes

POLITECNICO DI TORINO

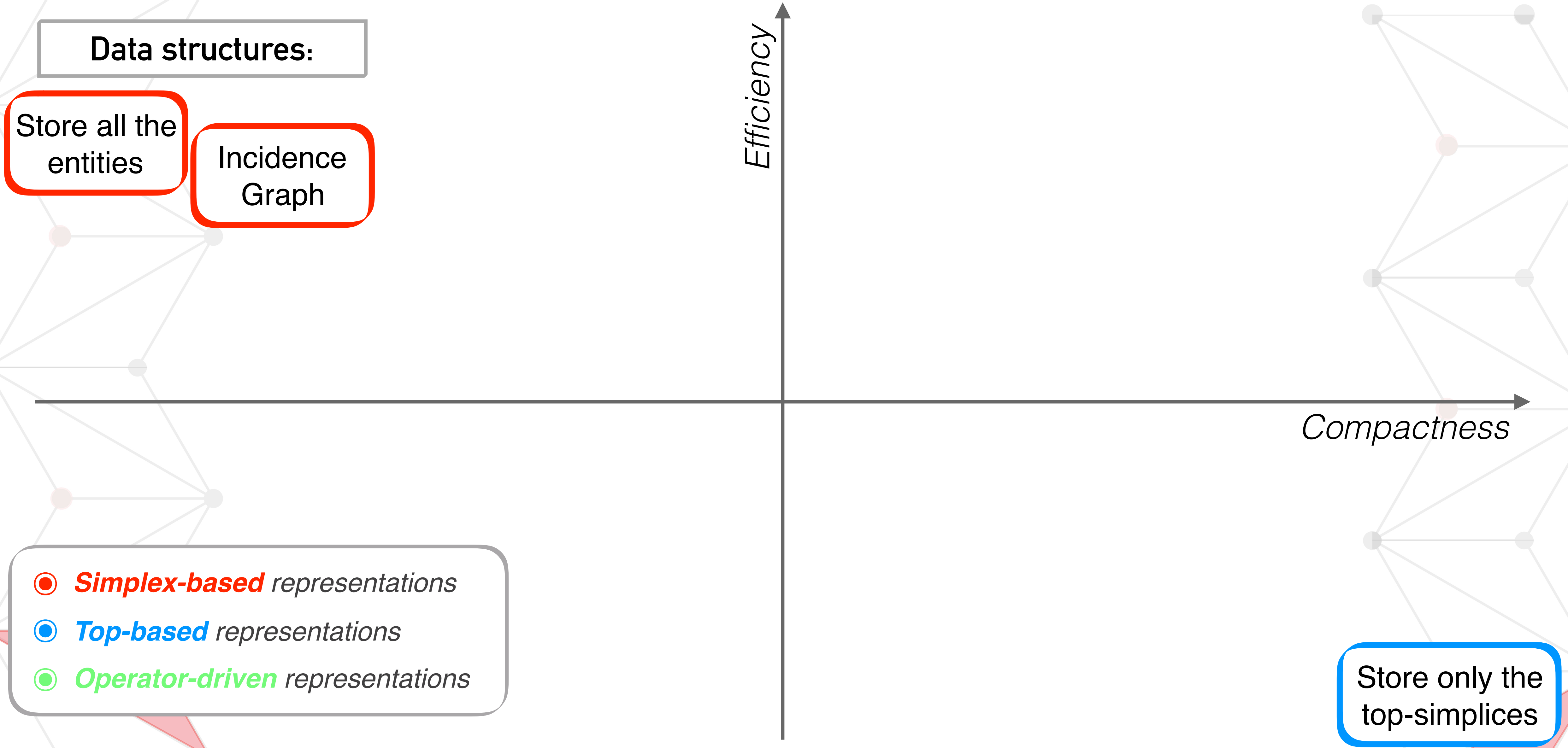
SmartData

Data structures:



- ***Simplex-based*** *representations*
- ***Top-based*** *representations*
- ***Operator-driven*** *representations*

Data Structures for Simplicial Complexes



Data Structures for Simplicial Complexes

POLITECNICO DI TORINO

SmartData

Data structures:

Store all the entities

Incidence Graph

Simplex Tree

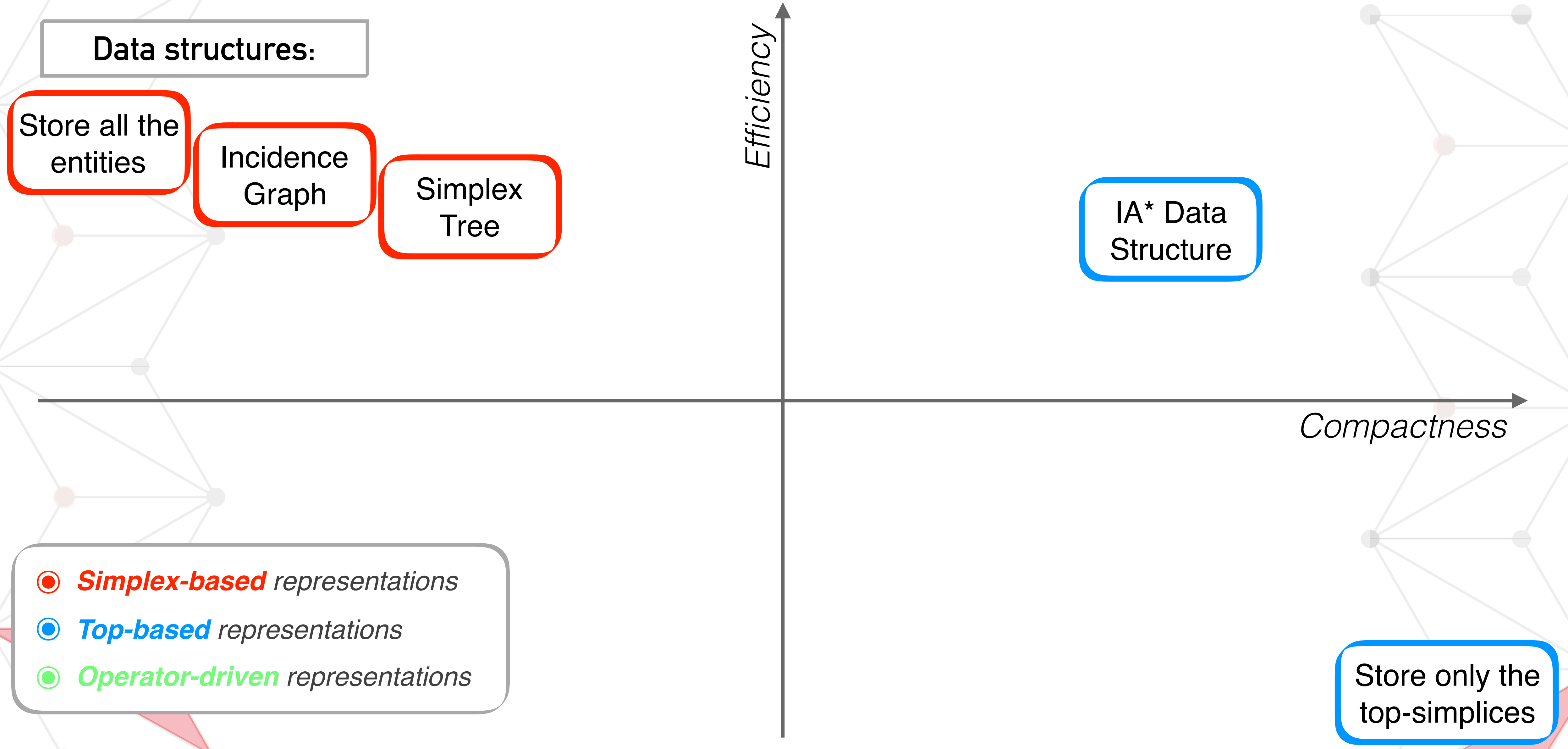
Efficiency

Compactness

- ***Simplex-based*** *representations*
- ***Top-based*** *representations*
- ***Operator-driven*** *representations*

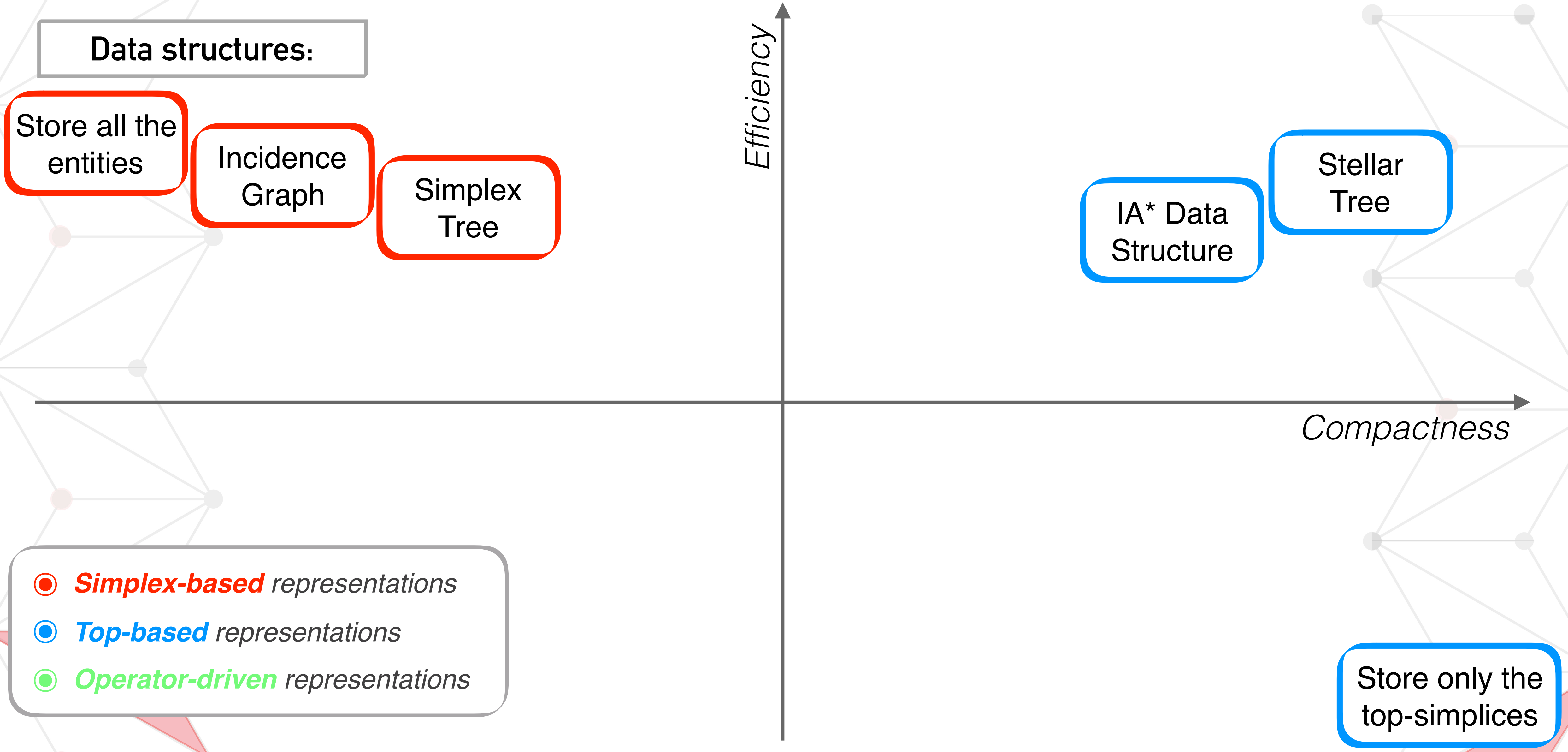
Store only the top-simplices

Data Structures for Simplicial Complexes



Data Structures for Simplicial Complexes

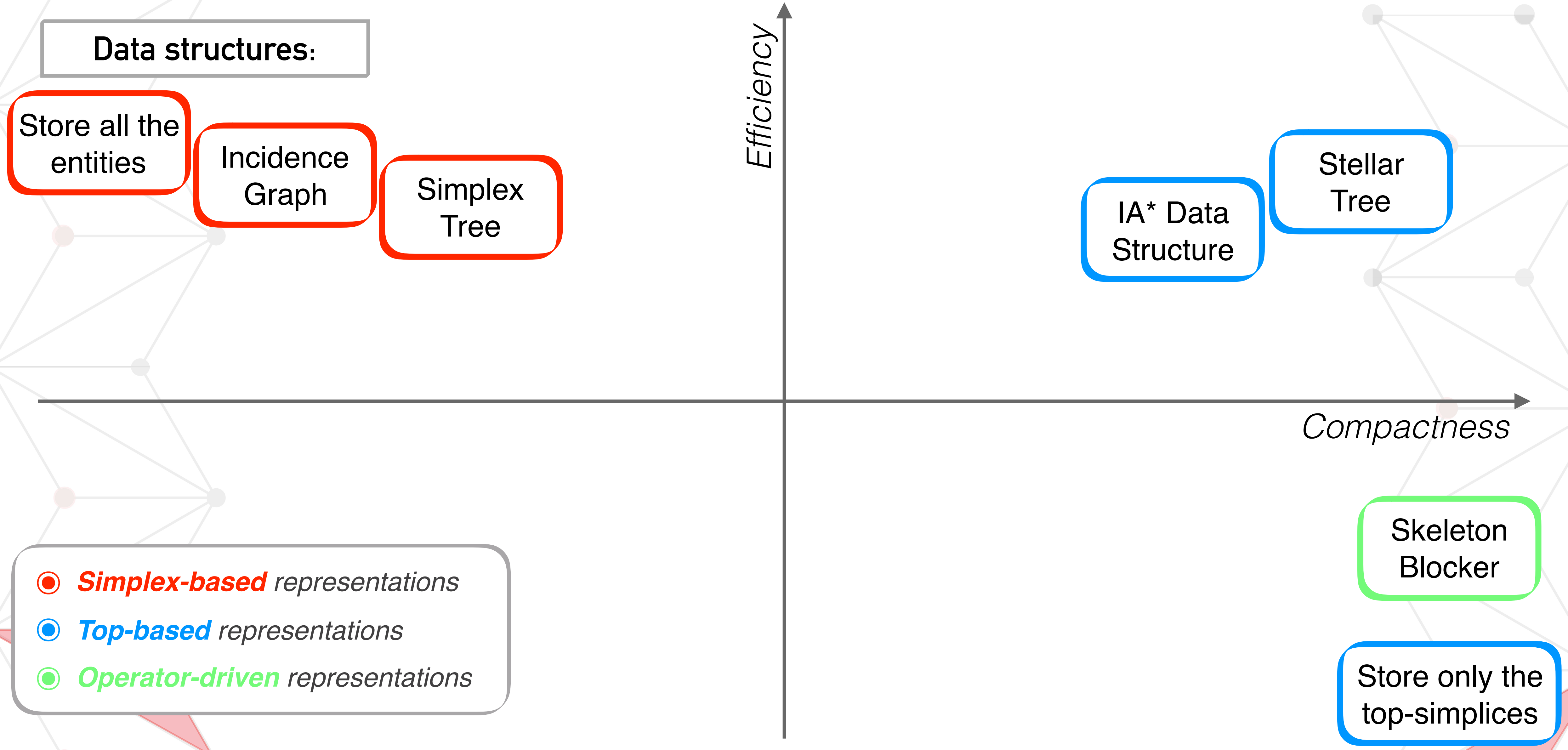
Data structures:



- ***Simplex-based*** *representations*
- ***Top-based*** *representations*
- ***Operator-driven*** *representations*

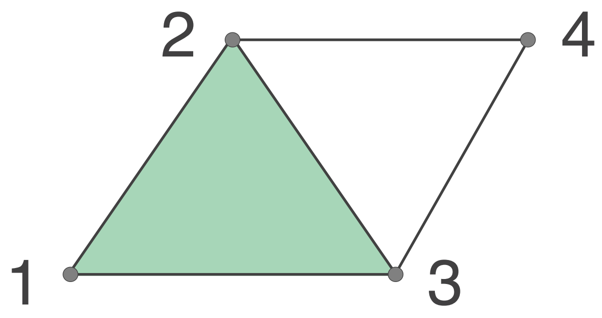
Data Structures for Simplicial Complexes

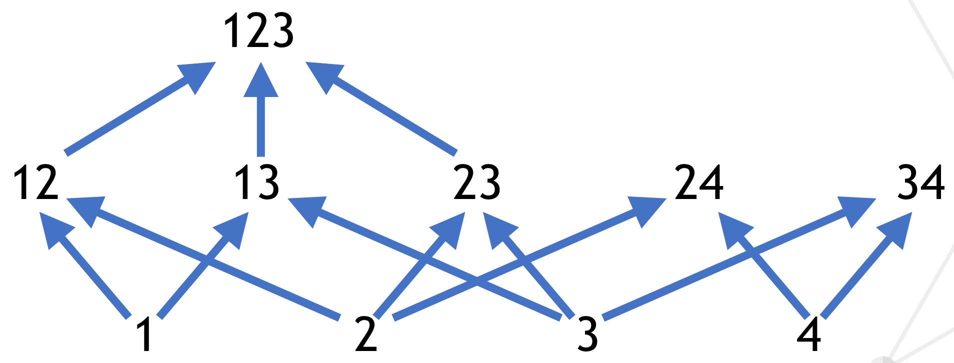
Data structures:



Simplex-based Representations

Incidence Graph:





The simplicial complex K is encoded via a ***directed graph G = (N, A)***:

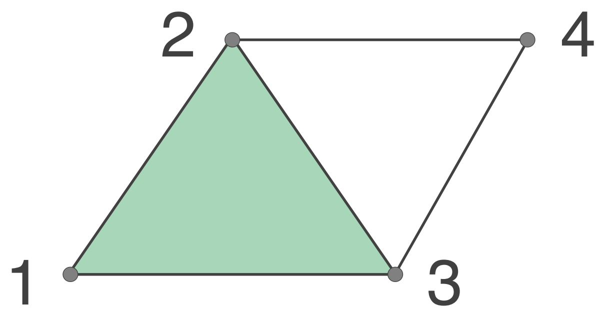
$\boldsymbol{N \longleftrightarrow K}$

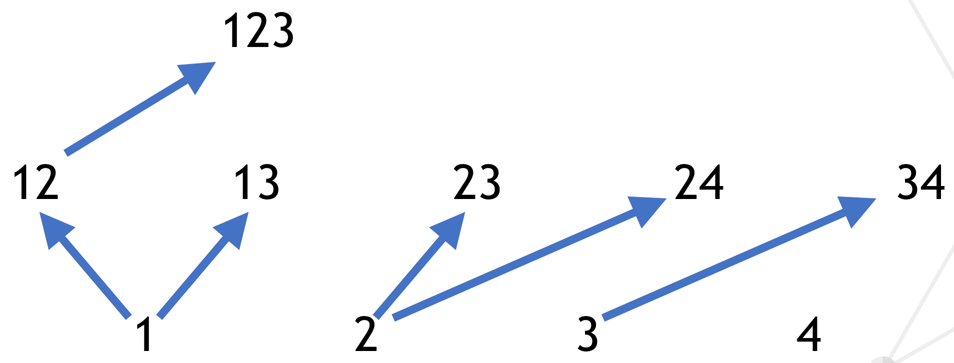
$(\sigma, \tau) \in \boldsymbol{A} \longleftrightarrow (\sigma, \tau) \in \boldsymbol{R_{i,i+1}}$

- All the relations between simplices can be immediately retrieved
- The representation ***size exponentially increases*** with the complex dimension

Simplex-based Representations

Simplex Tree:





The simplicial complex K is encoded via a ***directed graph G = (N, A)***:

$$\boldsymbol{N \longleftrightarrow K}$$

$$(\sigma, \tau) \in \boldsymbol{A} \longleftrightarrow (\sigma, \tau) \in \boldsymbol{R_{i,i+1}} \textit{ and } \boldsymbol{l(\sigma) < l(\tau)}$$

where $l(\sigma)$ denotes the ***maximum value*** taken by the vertices of σ w.r.t. a ***total order*** on K_0

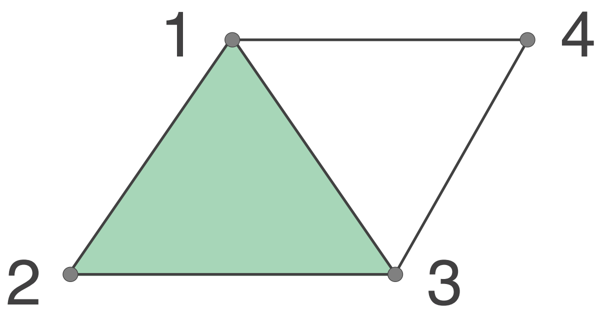
Graph is ***not uniquely determined*** but it depends on the chosen vertex order

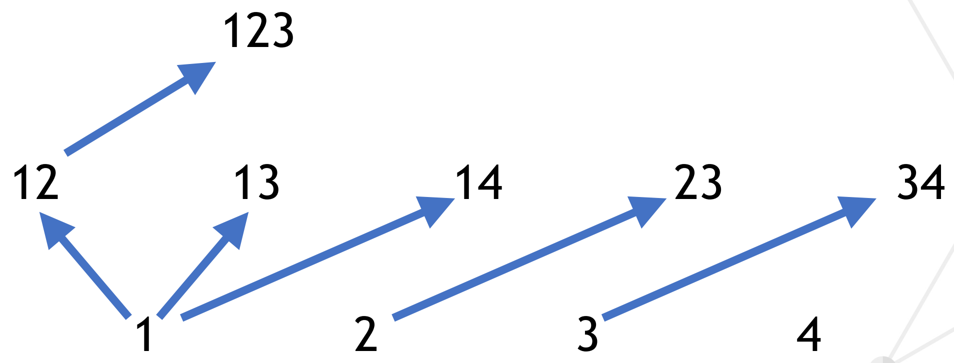
Simplex-based Representations

POLITECNICO DI TORINO

SmartData

Simplex Tree:





The simplicial complex K is encoded via a ***directed graph G = (N, A)***:

$$\boldsymbol{N \longleftrightarrow K}$$

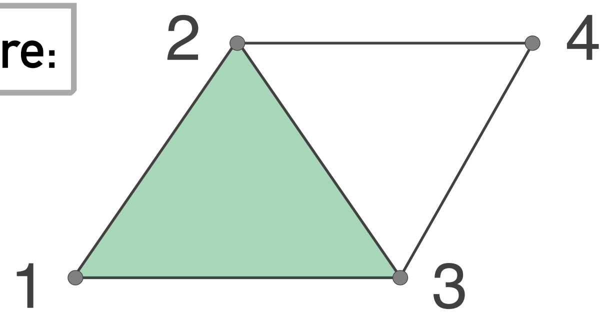
$$(\sigma, \tau) \in A \longleftrightarrow (\sigma, \tau) \in R_{i,i+1} \text{ and } l(\sigma) < l(\tau)$$

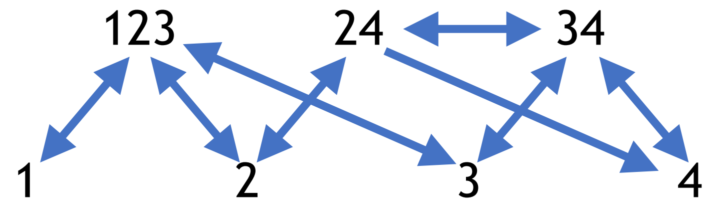
where $l(\sigma)$ denotes the ***maximum value*** taken by the vertices of σ w.r.t. a ***total order*** on K_0

Graph is ***not uniquely determined*** but it depends on the chosen vertex order

Top-based Representations

IA* Data Structure:





The simplicial complex K is encoded via a ***directed graph G = (N, A)***:

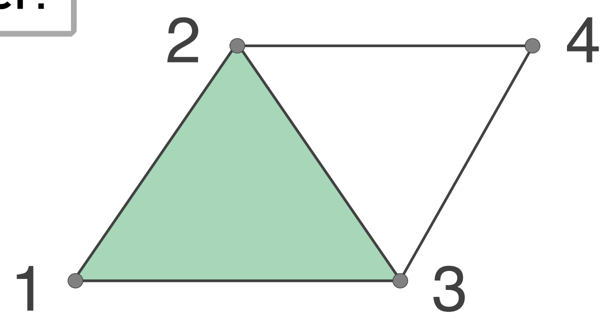
$$N \longleftrightarrow K_0 \cup K_{top}$$

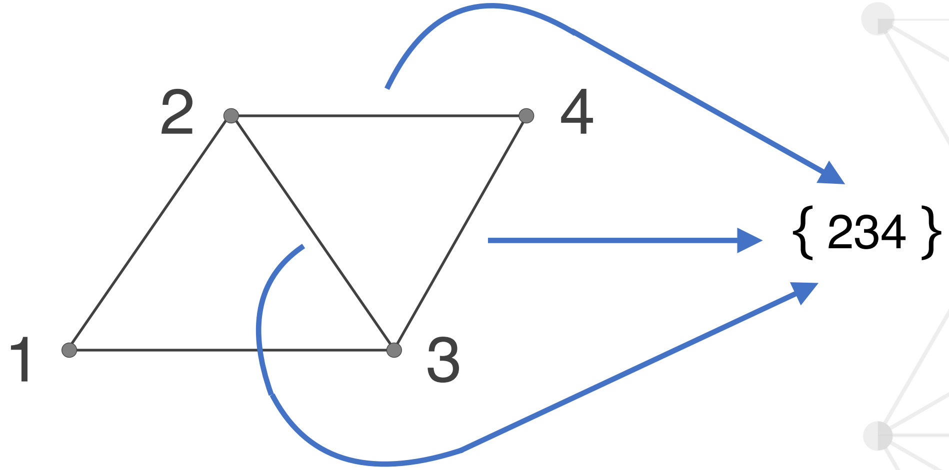
$$(\sigma, \tau) \in A \longleftrightarrow \begin{cases} \sigma \in K_{top} \text{ and } (\sigma, \tau) \in R_{i,0} \\ \sigma, \tau \in K_{top} \text{ and } (\sigma, \tau) \in R_{i,i} \\ \tau \in K_{top} \text{ and } (\sigma, \tau) \in (R_{0,i})^* \end{cases}$$

- ***Compact:*** it explicitly stores just a fraction of the entities of a simplicial complex
- ***Not*** all the relations between simplices are ***immediately available***

Operator-driven Representations

Skeleton Blocker:





The simplicial complex K is encoded by storing its ***1-skeleton*** (i.e. the graph consisting of the 0- and the 1-simplices) and a ***map*** returning, for each 1-simplex σ, the blockers of K containing σ, where:

a simplex τ is a ***blocker*** if τ does not belong to K but all its faces do

- ***Designed for*** flag complexes (e.g. ***VR complexes***) and edge contraction
- Too specific: ***inefficient in any other task***

Possible Issues in Top-based Representations

POLITECNICO
DI TORINO

SmartData

Top-based representations look like promising data structures for encoding a simplicial complex K

But, how to…

1. ***Store information associated to each simplex of K (e.g. labels, gradient, …)?***
2. ***Efficiently perform operators having explicitly stored a fraction of the entities of K?***

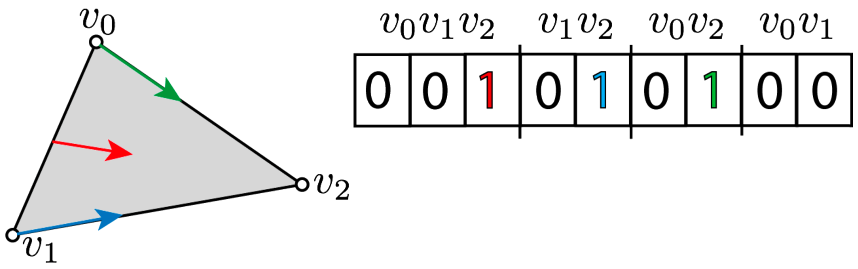
Possible Issues in Top-based Representations

Top-based representations look like promising data structures for encoding a simplicial complex K

But, how to…

1. Store information associated to each simplex of K (e.g. labels, gradient, …)?

Attach information to the top simplices only



2. Efficiently perform operators having explicitly stored a fraction of the entities of K?

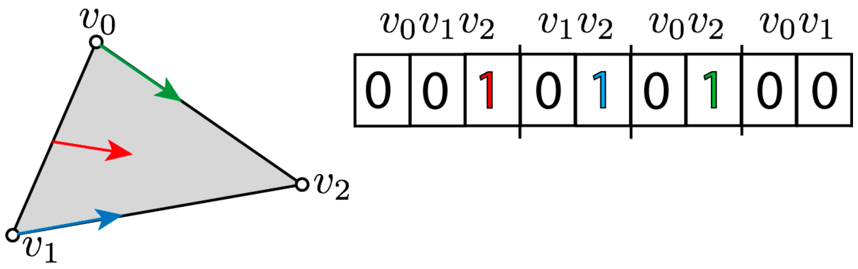
Possible Issues in Top-based Representations

Top-based representations look like promising data structures for encoding a simplicial complex K

But, how to…

1. Store information associated to each simplex of K (e.g. labels, gradient, …)?

Attach information to the top simplices only



2. Efficiently perform operators having explicitly stored a fraction of the entities of K?

Re-define the algorithms performing the operators trying to extract the lowest possible amount of non-explicitly stored entities

Comparisons

POLITECNICO DI TORINO

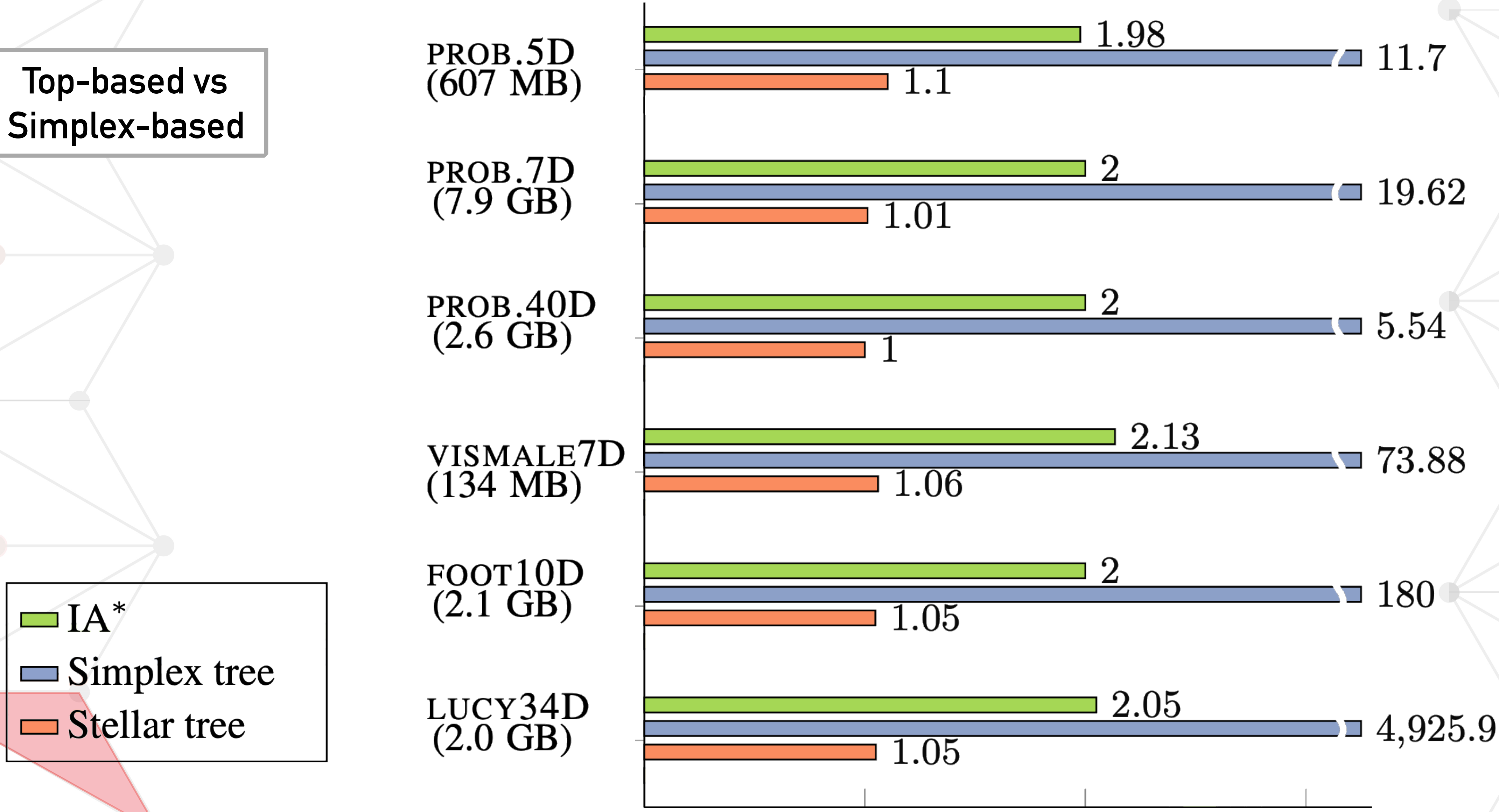
SmartData

Top-based vs Simplex-based

Dataset	d	$\|\Sigma_0\|$	$\|\Sigma_{top}\|$	$\|\Sigma\|$	Storage Cost		
					IA^*	IG	ST
DTI-SCAN	3	0.9M	5.5M	24M	0.97	11.9	2.4
VISMALE	3	4.6M	26M	118M	4.7	-	9.7
ACKLEY4	4	1.5M	32M	204M	6.8	-	12.8
AMAZON01	6	0.2M	0.4M	2.2M	0.12	1.6	0.3
AMAZON02	7	0.4M	1.0M	18.4M	0.28	9.8	1.5
ROADNET	3	1.9M	2.5M	4.8M	0.8	3.3	1.0
SPHERE-1.0	16	100	224	0.6M	0.003	0.9	0.04
SPHERE-1.2	21	100	285	26M	0.0032	-	1.5
SPHERE-1.3	23	100	382	197M	0.0034	-	11.01

Comparisons

Top-based vs Simplex-based



Comparisons

Top-based vs Operator-driven

data	ω		contr.	timings			memory peak	
			edges	**check**	**contr.**	**tot**	**gen.**	**simpl.**
CHICAGO	28	weak		$9.15h$	$2.27m$	$9.19h$	5.6	57.2K
		top	6.38K	$0.01s$	$0.02s$	$0.09s$		7.6
		Skel.		$0.00s$	$0.15s$	$0.15s$	7.8	7.8
	56	weak			out-of-memory		6.2	–
		top	7.99K	$0.04s$	$0.06s$	$0.23s$		10.8
		Skel.		$0.00s$	$0.71s$	$0.71s$	14.1	14.1
ATHENS	63	weak			out-of-memory		11.6	–
		top	27.9K	$0.08s$	$0.11s$	$0.38s$		14.9
		Skel.		$0.00s$	$0.74s$	$0.75s$	26.4	26.8
	126	weak			out-of-memory		10.0	–
		top	31.2K	$0.40s$	$0.49s$	$1.36s$		25.9
		Skel.		$0.01s$	$7.73s$	$7.74s$	66.1	66.7
VISMALE	3.5	weak		$34.3m$	$1.28m$	$40.4m$	1.0K	2.0K
		top	4.23M	$4.34m$	$0.89m$	$7.20m$		2.0K
		Skel.		$0.76m$	$3.34h$	$3.35h$	8.0K	8.0K
FOOT	4.5	weak			killed after 25 hours		7.5K	–
		top	4.69M	$2.89h$	$26.0m$	$3.32h$		10.7K
		Skel.			killed after 25 hours		19.4K	–
LUCY	1.5	weak			killed after 25 hours		7.5K	–
		top	14.0M	$11.9m$	$14.8m$	$32.0m$		15.4K
		Skel.		$23.19s$	$14.6h$	$14.6h$	50.9K	52.1K

In Summary

We have briefly overviewed the ***most common*** ***data structures*** proposed in the literature

Future Directions:

- ***Express*** the most frequently adopted operators ***in terms of top simplices***
- Face the bottleneck concerning the ***construction of a simplicial complex from a dataset*** (maybe proposing an approximated construction?)
- Investigate with your help the connections between simplicial complexes (and the data structures to store them) and ***itemsets in association rule learning***

Thank you!

Main References:

- ***Simplex-based Representations:***
  - Boissonnat, Maria. *The Simplex Tree: an Efficient Data Structure for General Simplicial Complexes.* In Algorithmica, 2014
- ***Top-based Representations:***
  - Canino, De Floriani, Weiss. *IA*: An Adjacency-Based Representation for Non-manifold Simplicial Shapes in Arbitrary Dimensions.* In Computers & Graphics, 2011
  - Fellegara, Weiss, De Floriani. *The Stellar Tree: a Compact Representation for Simplicial Complexes and Beyond.* arXiv preprint, 2017
- ***Operator-driven Representations:***
  - Attali, Lieutier, Salinas. *Efficient Data Structure for Representing and Simplifying Simplicial Complexes in High Dimensions.* In International Journal of Computational Geometry & Applications, 2012
- ***Comparisons:***
  - Fugacci, Iuricich, De Floriani. *Computing Discrete Morse Complexes from Simplicial Complexes.* In Graphical Models, 2019
  - Fellegara, Iuricich, De Floriani, Fugacci. *Efficient Homology-Preserving Simplification of High-Dimensional Simplicial Shapes.* In Computer Graphics Forum, 2019