



UNIVERSITÀ DI PISA

Learned data structures and their application to DBs



Giorgio VINCIGUERRA
pages.di.unipi.it/vinciguerra/
acube.di.unipi.it



Advanced Algorithms and Applications Lab (A³ Lab)



Paolo Ferragina
Lab leader



Marco Ponza
Postdoc



Giorgio Vinciguerra
PhD student



Antonio Boffa
Master's student



Lorenzo De Santis
Master's student



Francesco Tosoni
Master's student



Alessio Russo
Bachelor's student

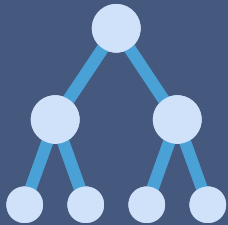
Main sponsors



Our focus areas



Compressed and Learned Data Structures



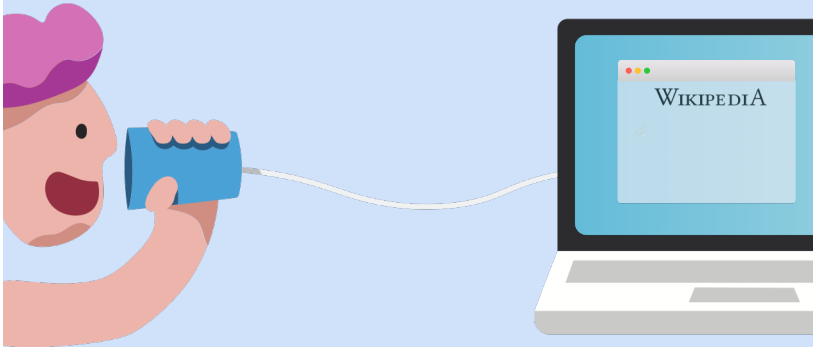
CGCGGC TGAC
AC\$GCGGCTG
C\$GCGGCTGA
CGGCTGAC\$G
CTGAC\$GCGG
GAC\$GCGGCT
GCGGCTGAC\$
GCTGAC\$GCG
GGCTGAC\$GC
TGAC\$GCGGC

Data Compression

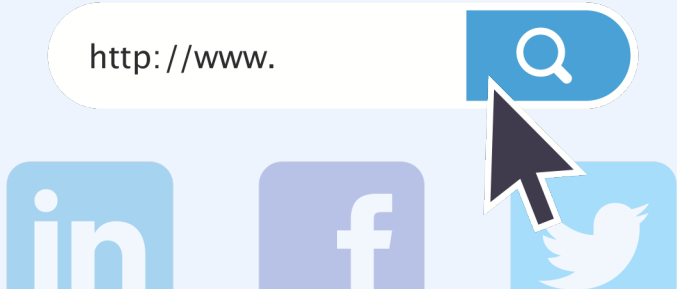


0010100101110101
010100101110111010
111011010010011001

Natural Language Understanding



Web & Social Media

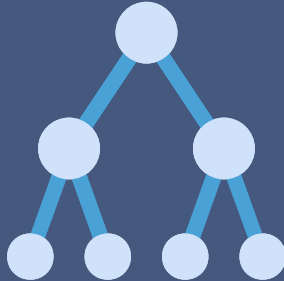


Our focus areas



\$GCGGCTGAC
AC\$GCGGCTG
C\$GCGGCTGA
CGGCTGAC\$G
CTGAC\$GCGG
GAC\$GCGGCT
SCGGCTGAC\$
SCTGAC\$GCG
SGCTGAC\$GC
TGAC\$GCGGC

Compressed
and Learned
Data Structures





Multicriteria
Data Structures

Funded by



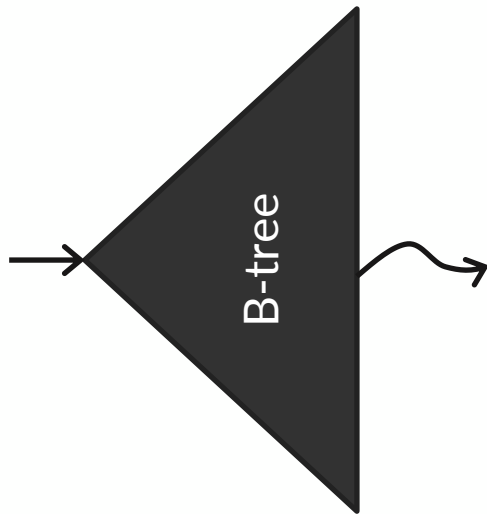
A common database task



Id	Name	Surname	Address	Birthday	City	ZipCode	Email Address	Telephone
598241	Olimpia	Padovesi	Via Valpantena 40	3/22/1997	Pincara	45020	OlimpiaPadovesi@gmail.com	364 5586931
598242	Guglielmo	Siciliano	Via Vipacco 131	12/27/1990	Ceriale	17023	GuglielmoSiciliano@outlook.com	381 9348312
598243	Libero	Padovesi	Via Vico Ferrovia 83	3/28/2000	Dodici Morelli	44040	LiberoPadovesi@gmail.com	346 4151821
598244	Rodolfo	Piccio	Via dei Fiorentini 40	1/30/1991	Napoli	80142	RodolfoPiccio@virgilio.it	316 9920360
598245	Lorenzo	Lucchesi	Via Carlo Alberto 21	10/29/1994	Cernobbio	22012	LorenzoLucchesi@virgilio.it	332 4148918
598246	Guarino	Lucciano	Via A.G. Alaimo 99	12/17/1992	Villa Botteri	15070	GuarinoLucciano@gmail.com	321 9524852
598247	Orestilla	Lorenzo	Via Matteo Schilizzi 38	1/6/1995	San Colombano Certenoli	16040	OrestillaLorenzo@tiscali.it	361 5285109
598248	Agenore	Milani	Via Sergente Maggiore 112	11/5/2000	Valconasso	29010	AgenoreMilani@outlook.com	363 6679100
598249	Sergio	Ricci	Corso Vittorio Emanuele 138	1/5/2000	Cellena	58050	SergioRicci@outlook.it	358 3207432
598250	Erica	Marchesi	Via Campi Flegrei 15	11/11/1998	Zungoli	83030	EricaMarchesi@egmail.com	393 2001401
598251	Durante	Toscano	Piazza Giuseppe Garibaldi 98	7/25/1990	Fornovo Di Taro	43045	DuranteToscano@outlook.it	358 2955558
598252	Belinda	Giordano	Corso Vittorio Emanuele 4	4/7/1997	Polverosa	58010	BelindaGiordano@live.it	333 0933884
598253	Paride	Iadanza	Piazza San Carlo 96	6/29/2001	Stroppiana	13010	Parideladanza@gmail.com	366 7763300
598254	Mafalda	Genovesi	Corso Garibaldi 17	2/13/1999	Selva Di Sora	3039	MafaldaGenovesi@libero.it	327 0211406
598255	Vera	DeRose	Via Giacinto Gigante 77	1/22/1999	Torre Maina	41050	VeraDeRose@outlook.it	339 9092672
598256	Ottavio	Fanucci	Via Antonio da Legnago 126	2/22/1993	Marina Di Gioiosa Jonica	89046	OttavioFanucci@virgilio.it	389 8794516



A common database task



Id	Name	Surname	Address	Birthday	City	ZipCode	Email Address
598241	Olimpia	Padovesi	Via Valpantena 40	3/22/1997	Pincara	45020	OlimpiaF...
598242	Guglielmo	Siciliano	Via Vipacco 131	12/27/1990	Ceriale	17023	Guglielm...
598243	Libero	Padovesi	Via Vico Ferrovia 83	3/28/2000	Dodici Morelli	44040	LiberoPa...
598244	Rodolfo	Piccio	Via dei Fiorentini 40	1/30/1991	Napoli	80142	RodolfoF...
598245	Lorenzo	Lucchesi	Via Carlo Alberto 21	10/29/1994	Cernobbio	22012	Lorenzob...
598246	Guarino	Lucciano	Via A.G. Alaimo 99	12/17/1992	Villa Botteri	15070	GuarinoL...
598247	Orestilla	Lorenzo	Via Matteo Schilizzi 38	1/6/1995	San Colombano Certenoli	16040	Orestilla...
598248	Agenore	Milani	Via Sergente Maggiore 112	11/5/2000	Valconasso	29010	AgenoreM...
598249	Sergio	Ricci	Corso Vittorio Emanuele 138	1/5/2000	Cellena	58050	SergioRic...
598250	Erica	Marchesi	Via Campi Flegrei 15	11/11/1998	Zungoli	83030	EricaMar...
598251	Durante	Toscano	Piazza Giuseppe Garibaldi 98	7/25/1990	Fornovo Di Taro	43045	DuranteG...
598252	Belinda	Giordano	Corso Vittorio Emanuele 4	4/7/1997	Polverosa	58010	BelindaC...
598253	Paride	Iadanza	Piazza San Carlo 96	6/29/2001	Stroppiana	13010	Paridelac...
598254	Mafalda	Genovesi	Corso Garibaldi 17	2/13/1999	Selva Di Sora	3039	MafaldaG...
598255	Vera	DeRose	Via Giacinto Gigante 77	1/22/1999	Torre Maina	41050	VeraDeRo...
598256	Ottavio	Fanucci	Via Antonio da Legnago 126	2/22/1993	Marina Di Gioiosa Jonica	89046	OttavioF...



A common database task



$\text{position}(id) = id + \text{offset}$

Id	Name	Surname	Address	Birthday	City	ZipCode	Email Address
598241	Olimpia	Padovesi	Via Valpantena 40	3/22/1997	Pincara	45020	OlimpiaF...
598242	Guglielmo	Siciliano	Via Vipacco 131	12/27/1990	Ceriale	17023	Guglielm...
598243	Libero	Padovesi	Via Vico Ferrovia 83	3/28/2000	Dodici Morelli	44040	LiberoPa...
598244	Rodolfo	Piccio	Via dei Fiorentini 40	1/30/1991	Napoli	80142	RodolfoF...
598245	Lorenzo	Lucchesi	Via Carlo Alberto 21	10/29/1994	Cernobbio	22012	Lorenzol...
598246	Guarino	Lucciano	Via A.G. Alaimo 99	12/17/1992	Villa Botteri	15070	GuarinoL...
598247	Orestilla	Lorenzo	Via Matteo Schilizzi 38	1/6/1995	San Colombano Certenoli	16040	Orestilla...
598248	Agenore	Milani	Via Sergente Maggiore 112	11/5/2000	Valconasso	29010	AgenoreL...
598249	Sergio	Ricci	Corso Vittorio Emanuele 138	1/5/2000	Cellena	58050	SergioRic...
598250	Erica	Marchesi	Via Campi Flegrei 15	11/11/1998	Zungoli	83030	EricaMar...
598251	Durante	Toscano	Piazza Giuseppe Garibaldi 98	7/25/1990	Fornovo Di Taro	43045	DuranteL...
598252	Belinda	Giordano	Corso Vittorio Emanuele 4	4/7/1997	Polverosa	58010	BelindaC...
598253	Paride	Iadanza	Piazza San Carlo 96	6/29/2001	Stroppiana	13010	Paridelac...
598254	Mafalda	Genovesi	Corso Garibaldi 17	2/13/1999	Selva Di Sora	3039	MafaldaD...
598255	Vera	DeRose	Via Giacinto Gigante 77	1/22/1999	Torre Maina	41050	VeraDeRo...
598256	Ottavio	Fanucci	Via Antonio da Legnago 126	2/22/1993	Marina Di Gioiosa Jonica	89046	OttavioF...

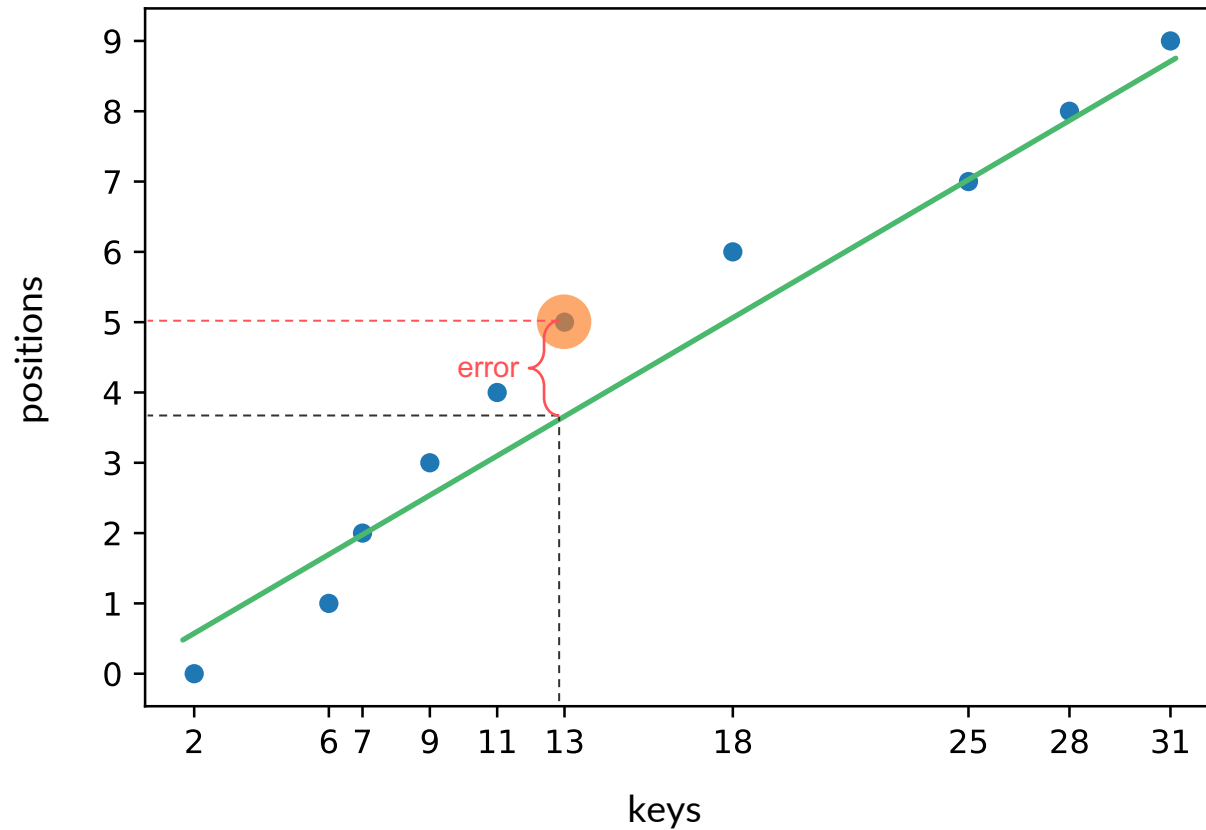


A closer look at realistic data



keys	2	6	7	9	11	13	18	25	28	31
positions	0	1	2	3	4	5	6	7	8	9

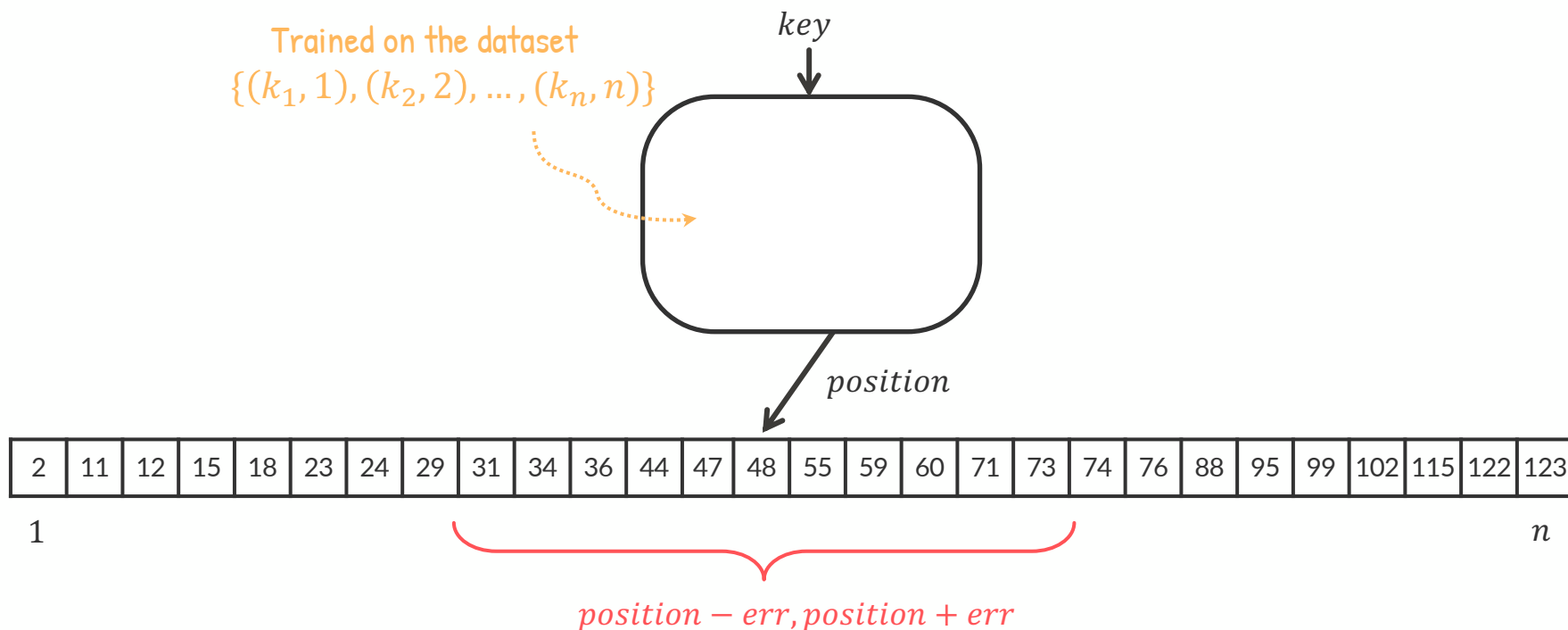
A red curly bracket is positioned under the positions 4 and 5.



B-trees are machine learning models



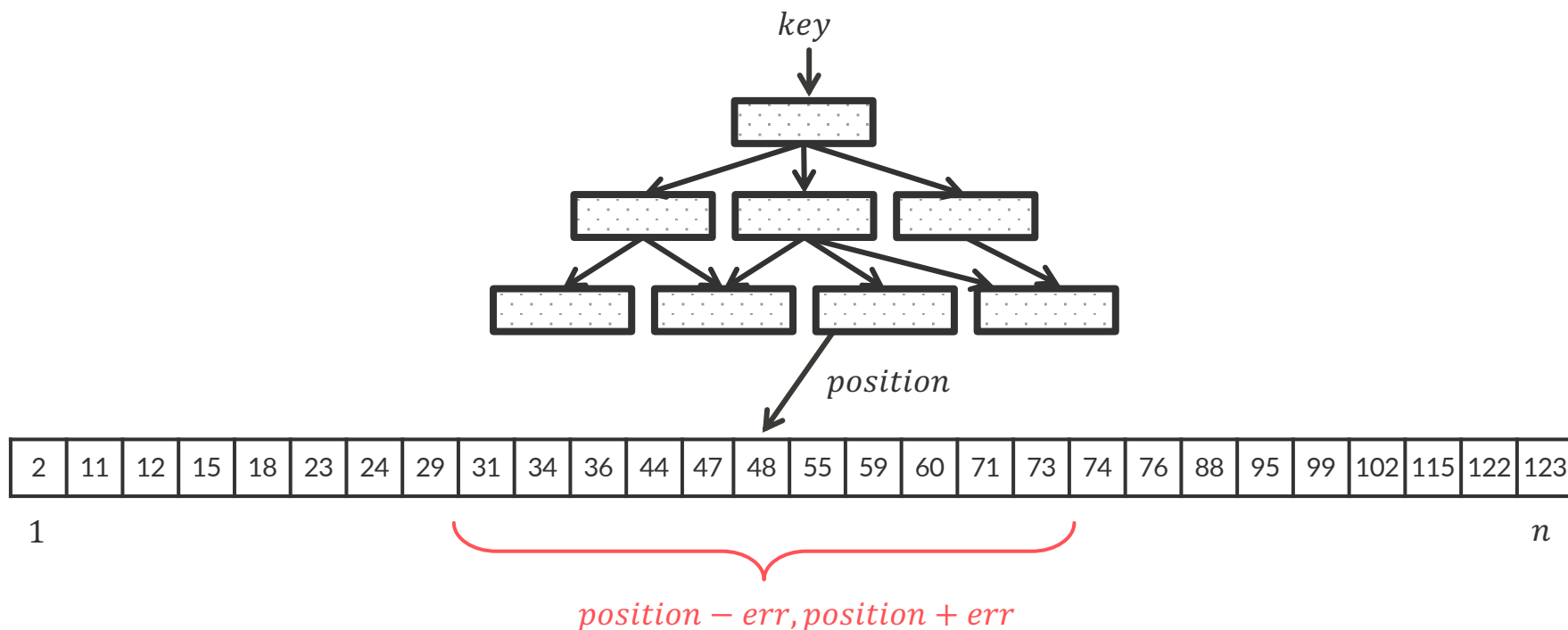
“All existing index structures can be replaced with other types of models, including deep-learning models.” [SIGMOD '18]



B-trees are machine learning models



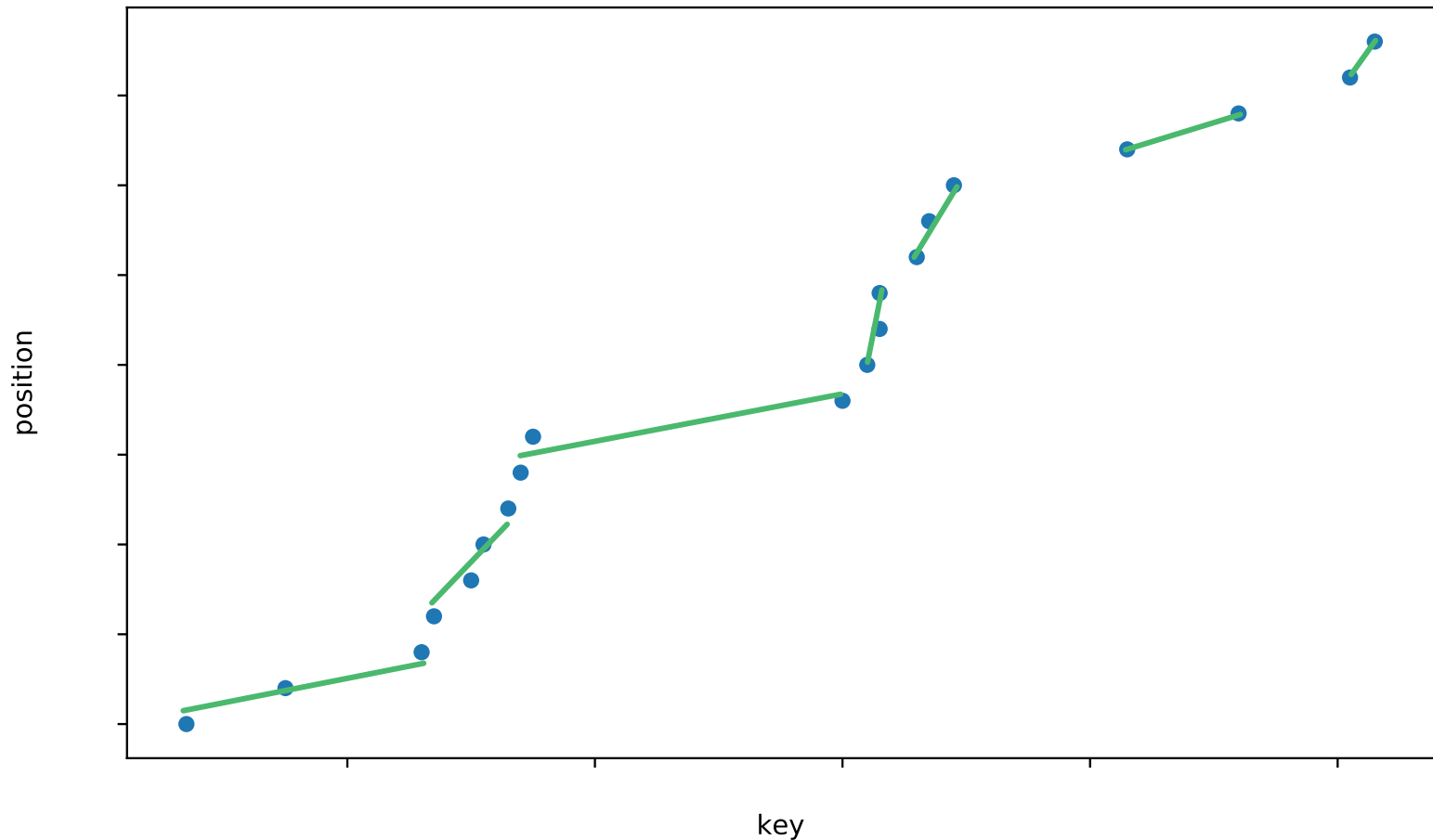
“All existing index structures can be replaced with other types of models, including deep-learning models.” [SIGMOD '18]



The Piecewise Geometric Model index



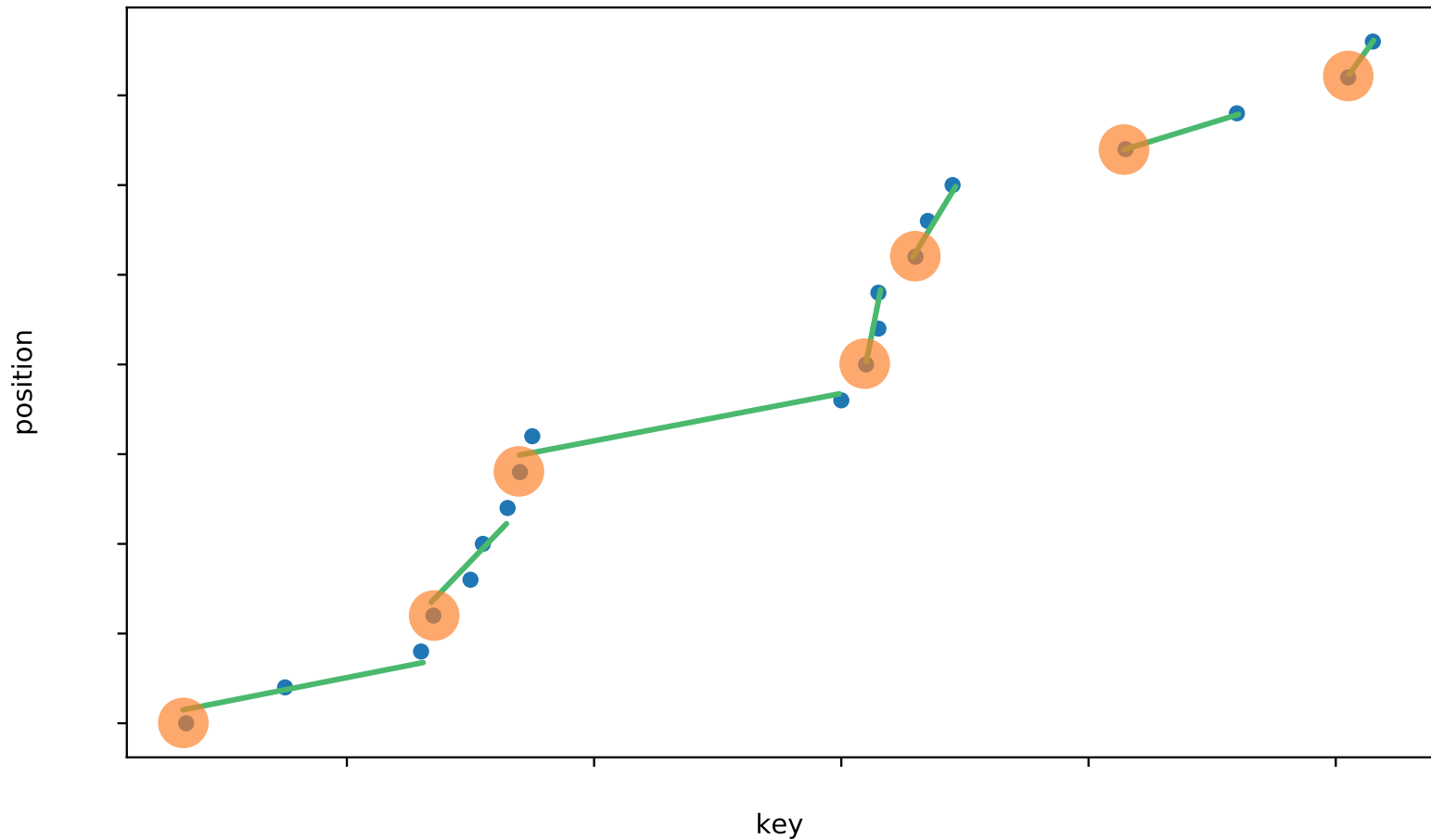
Compute the **opt** piecewise linear approx with max error ε in $O(n)$ time



PGM-index construction



Save the segments as triples (*key*, *slope*, *intercept*)



Memory layout of the PGM-index

~

Segments

(2, sl, ic)	(23, sl, ic)	(31, sl, ic)	(48, sl, ic)	(71, sl, ic)	(76, sl, ic)	(102, sl, ic)
-------------	--------------	--------------	--------------	--------------	--------------	---------------

Input keys

2	11	12	15	18	23	24	29	31	34	36	44	47	48	55	59	60	71	73	74	76	88	95	99	102	115	122	123
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----	-----	-----	-----

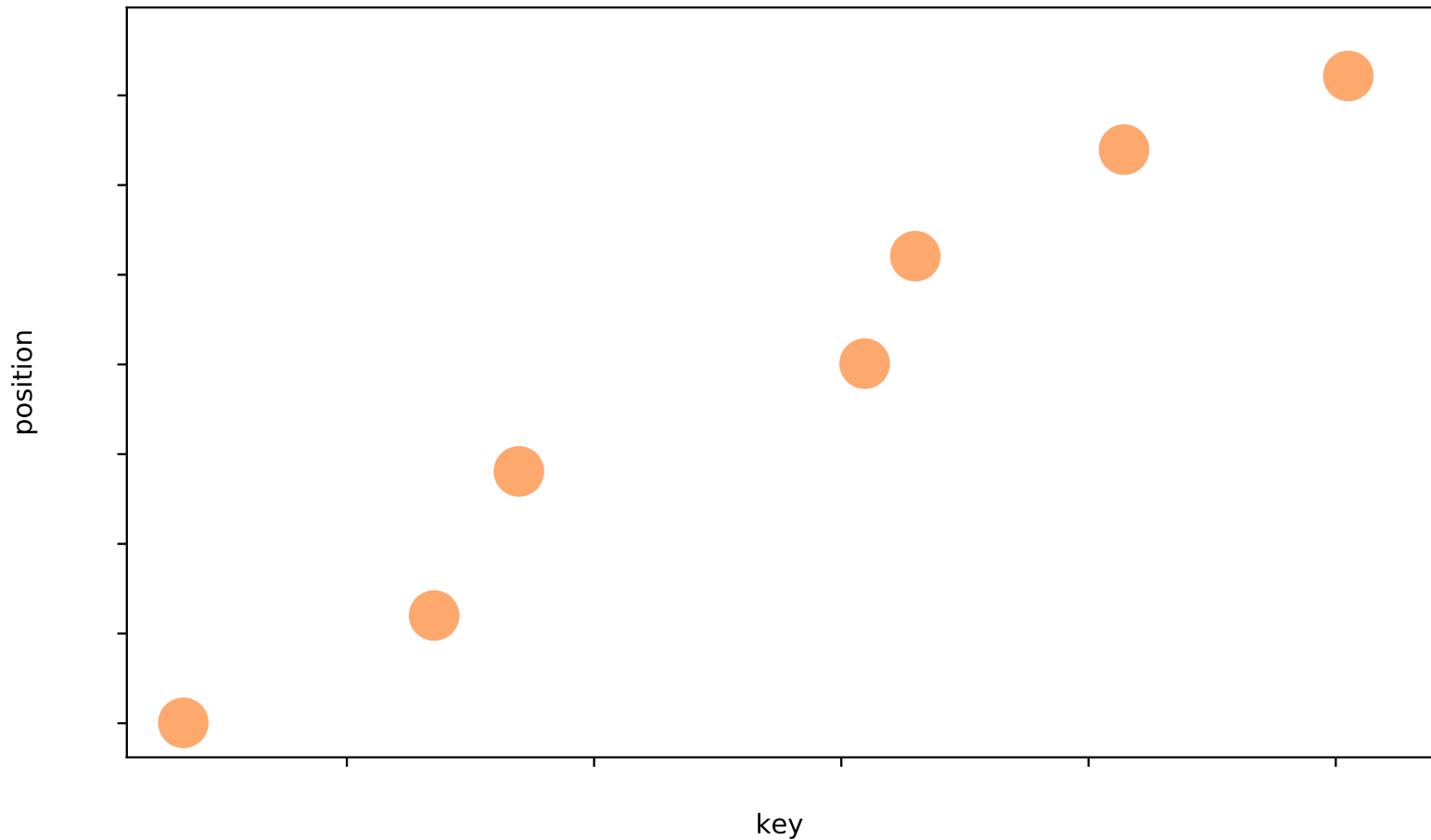
1

n

PGM-index construction



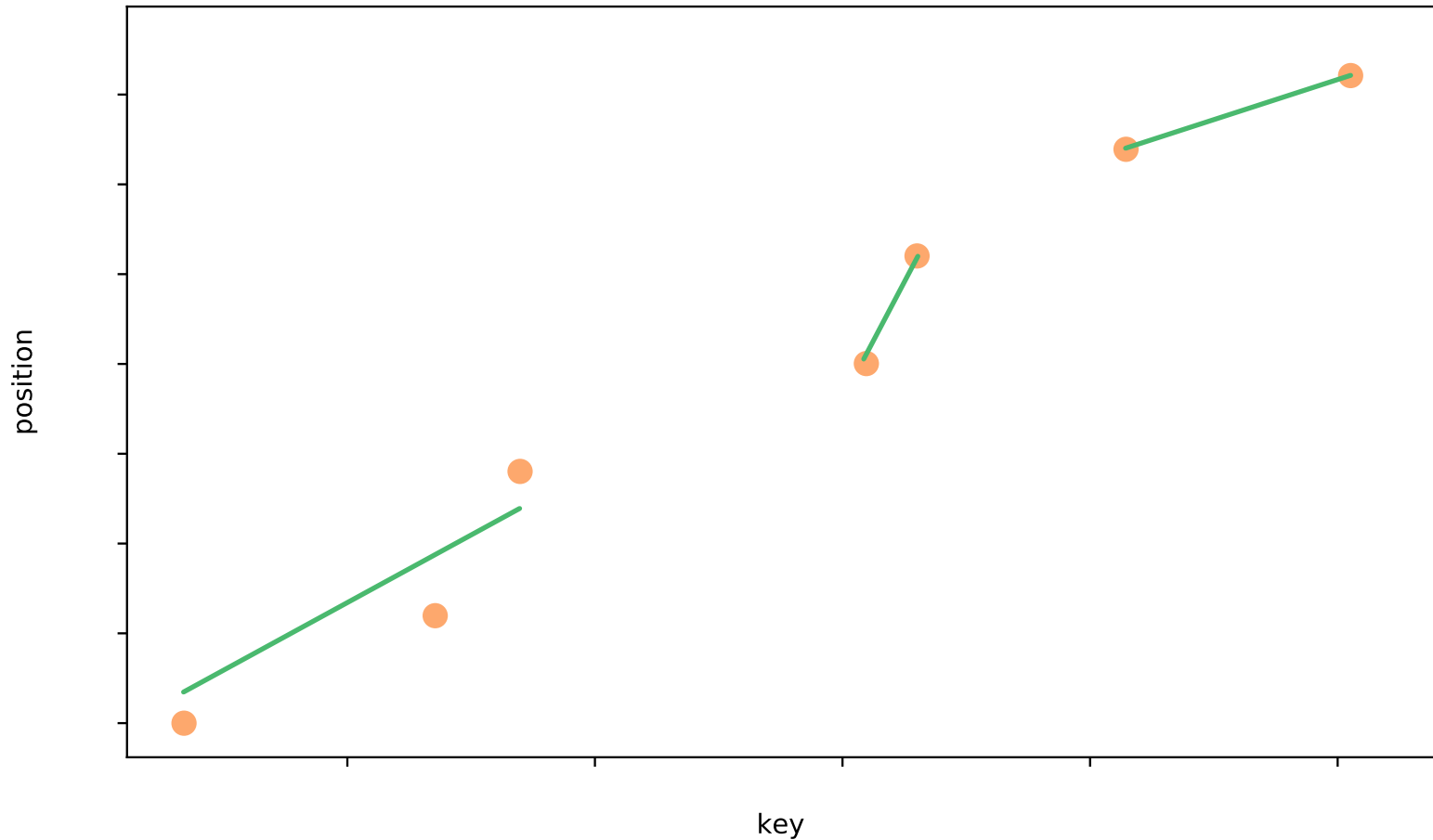
Drop all the points except *segment*.**key**



PGM-index construction



... and repeat!

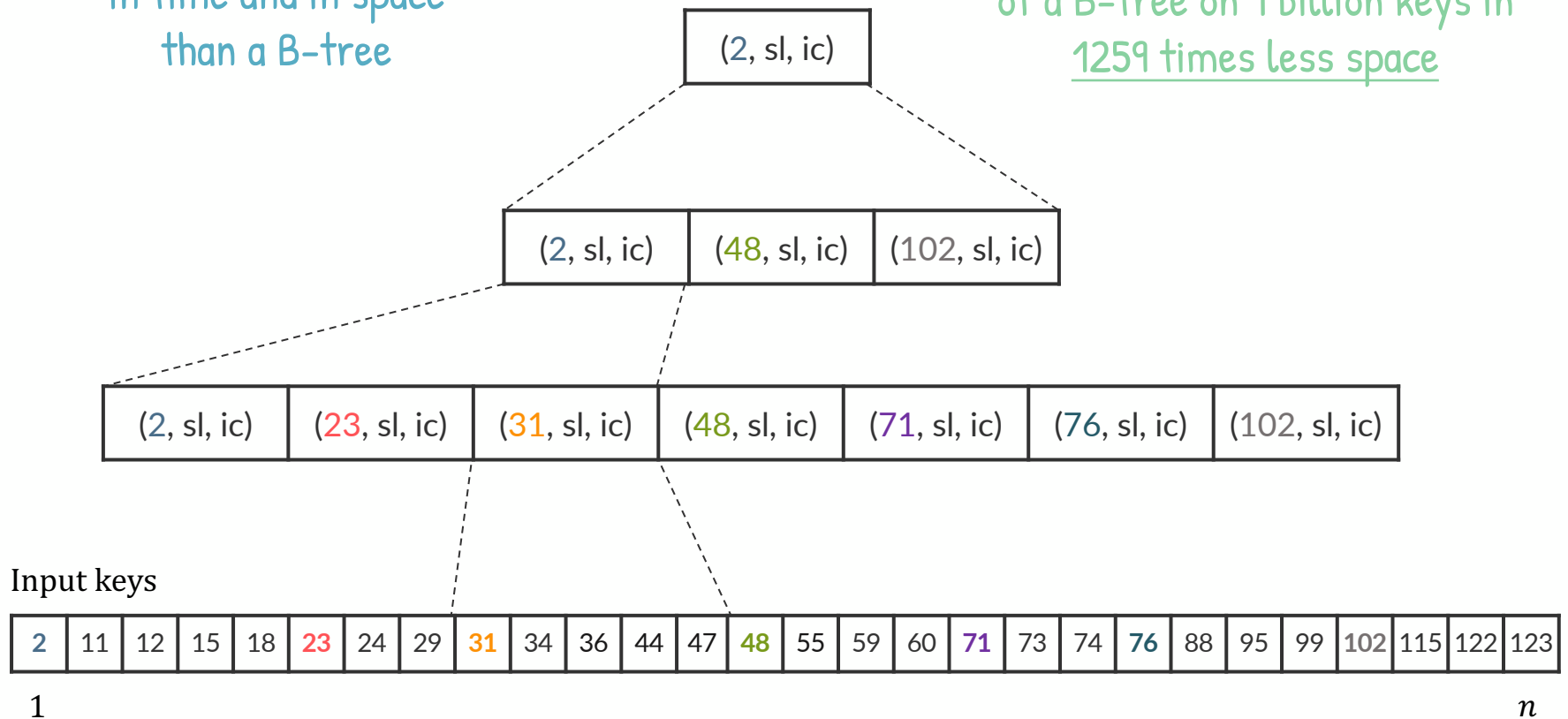


Memory layout of the PGM-index



In theory it can never be worse
in time and in space
than a B-tree

In practice it matched the performance
of a B-tree on 1 billion keys in
1259 times less space





Welcome to the *Advanced Algorithms and Applications Laboratory*, shortly the A³ Lab. We are a research group founded in 2003 by [Prof. Paolo Ferragina](#) and located at the Department of Computer Science at the University of Pisa. We design, analyze and develop effective and efficient algorithms for [Compressed Data Structures](#), [Data Compression](#), [Natural Language Understanding](#), and [Web & Social Media](#).

The A³ Lab received a vast number of governmental and industrial grants to support its researches, spanning from a Yahoo! Research Grant in 2006 to the recent Bloomberg and Google Grants in 2017 and 2018, respectively. Our lab is also part of the [SoBigData H2020 Research Infrastructure](#) dedicated to Big Data and Social Mining (INFRAIA grant agreement #654024). We invite you to have a look at the topics we research on as well as at the full list of the [grants](#) received in the past years.

GCGGCTGAC
C\$GCGGCTG
\$GCGGCTGA
GGCTGAC\$G
TGAC\$GCGG
AC\$GCGGCT
CGGCTGAC\$
CTGAC\$GCG
GCTGAC\$GC
GAC\$GCGGC

Compressed and Learned Data Structures



Natural Language Understanding

Data Compression

001010010111010101
010100101110111010
111011010010011001



Web & Social Media

References



1. Ferragina, P., Vinciguerra, G. (2019). *The PGM-index: a multicriteria, compressed and learned approach to data indexing*. arXiv:1910.06169. <https://pgm.di.unipi.it>
2. Kraska, T., Beutel, A., Chi, E. H., Dean, J., Polyzotis, N. (2018). *The case for learned index structures*. In ACM SIGMOD
3. Idreos, S., Zoumpatianos, K., Hentschel, B., Kester, M. S., Guo, D. (2018). *The data calculator: Data structure design and cost synthesis from first principles and learned cost models*. In ACM SIGMOD
4. Tutorials on self-designing (<https://lnkd.in/eQ4sezi>) and learned (<https://lnkd.in/eKZeNKA>) data structures