Content-aware Data Discovery on VO Catalogs Using Succinct Representations

Dr. Mauricio Araya

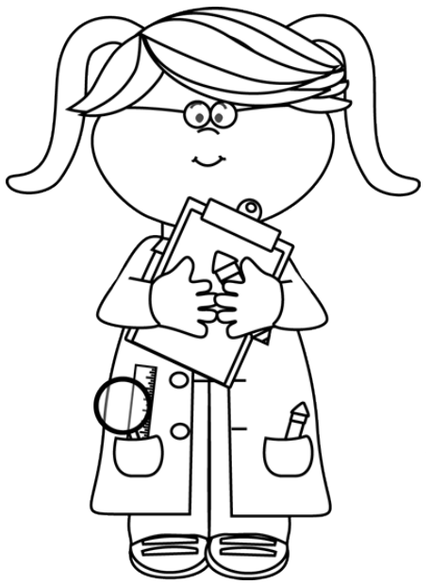
Camilo Saldías, Diego Arroyuelo and Mauricio Solar

1 Content-aware Data Discovery

2 on VO Catalogs Using

3 Succinct Representations

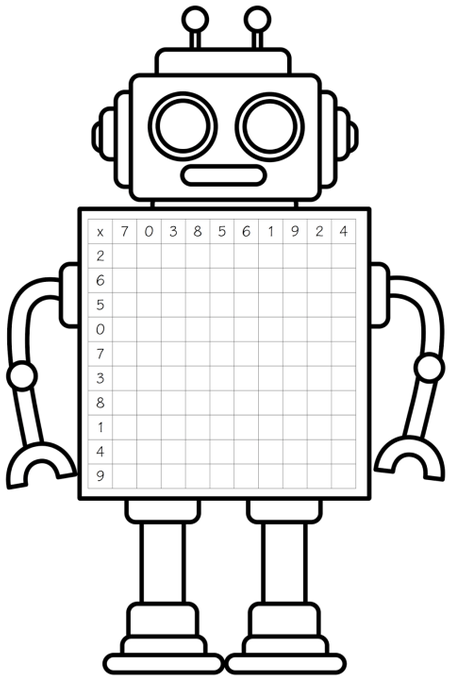
1 Content-aware Data Discovery



Astronomer

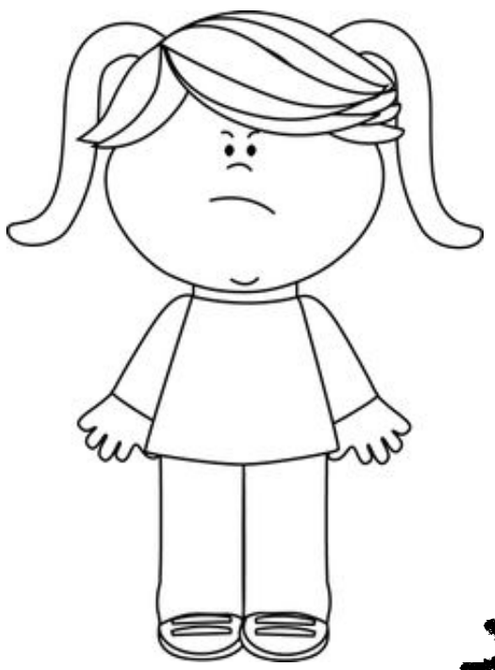
"Give me the metallicities of faint RR Lyrae stars"

"Yes master, here they are all of them!. I've sorted them by inverse luminosity for you."



Archives

1 Content-aware Data Discovery

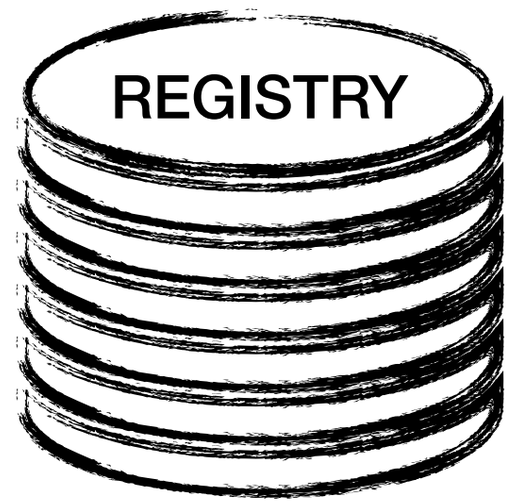


Astronomer

Search for resources
with variable stars tagged

Search for resources
with metallicity information

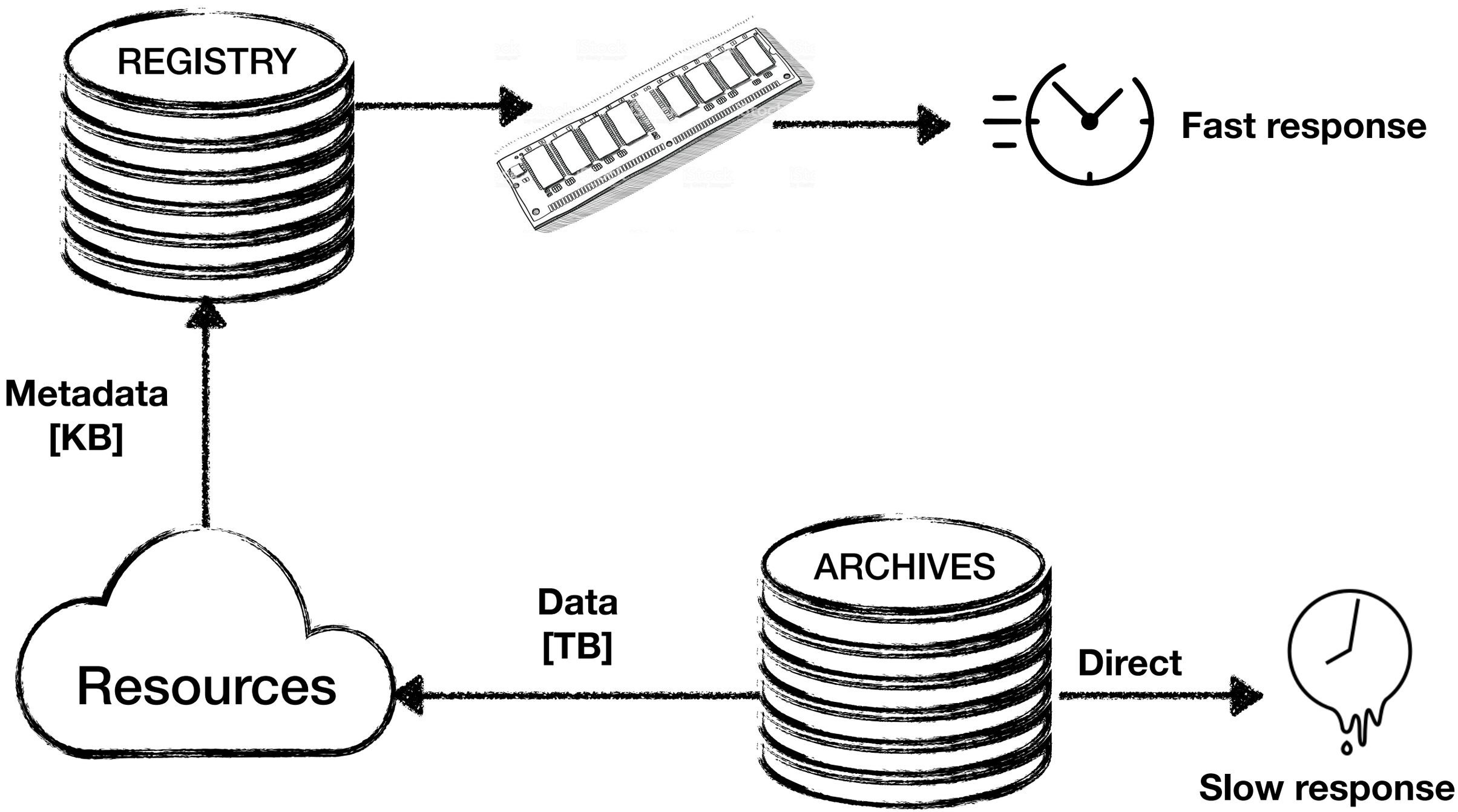
Resources URI



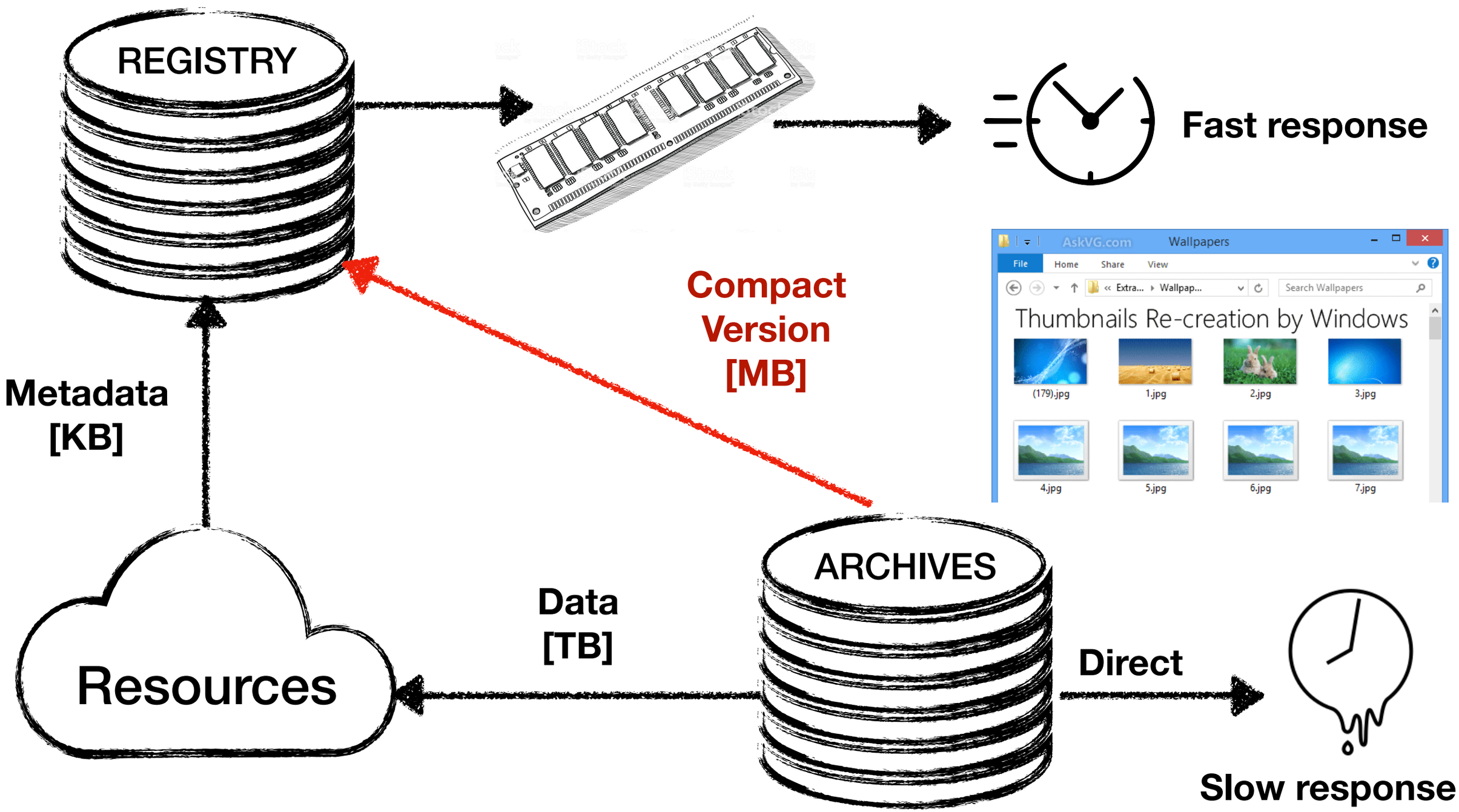
Download data for each resource

Resources

1 Content-aware Data Discovery



1 Content-aware Data Discovery

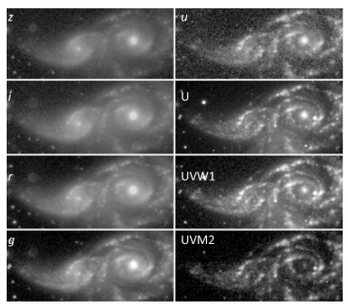


1 Content-aware Data Discovery

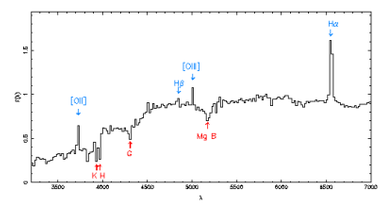
2 on VO Catalogs Using

3 Succinct Representations

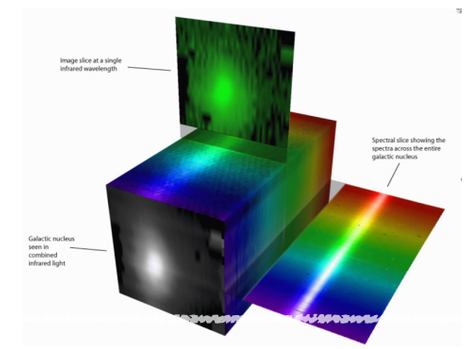
2 on VO Catalogs Using



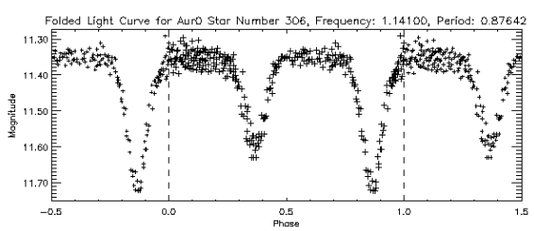
Images



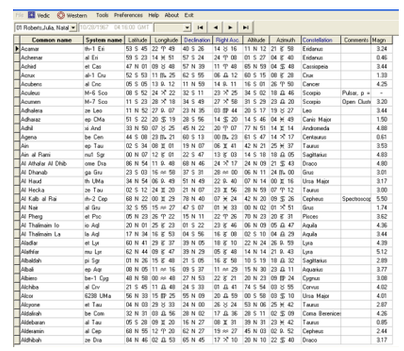
Spectrum



Spectral Cubes



Light Curves



Catalogs

Raw Data

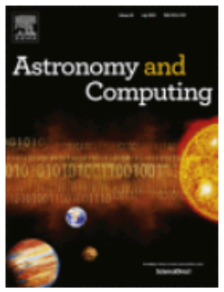
How can be compacted?

2

ELSEVIER

Astronomy and Computing

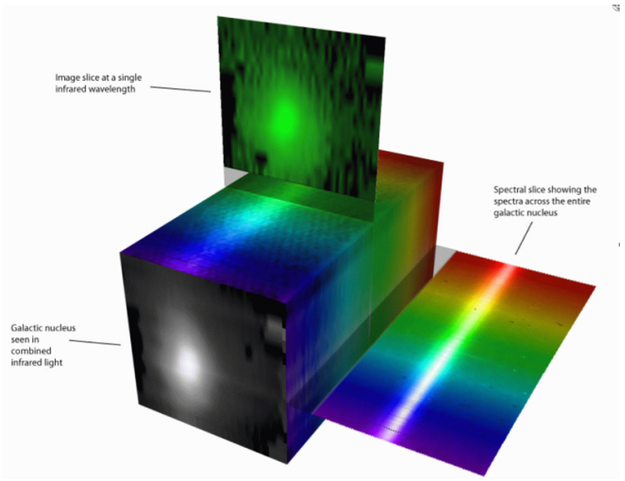
Volume 24, July 2018, Pages 25-35

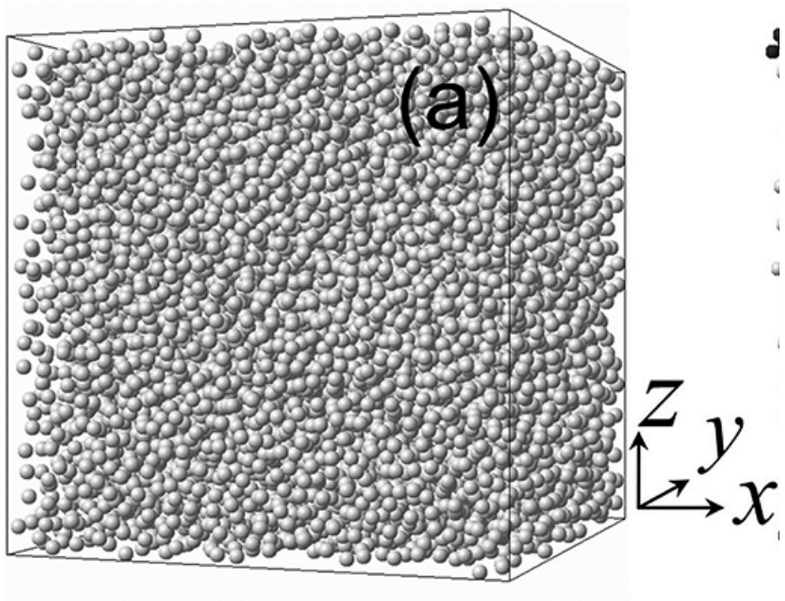


Full length article

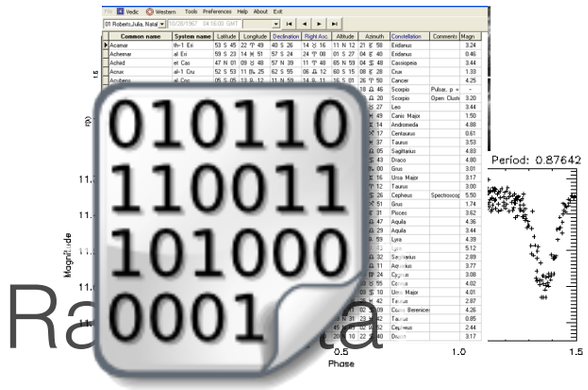
Unsupervised learning of structure in spectroscopic cubes

M. Araya [a], M. Mendoza [a], M. Solar [a], D. Mardones [b], A. Bayo [c]





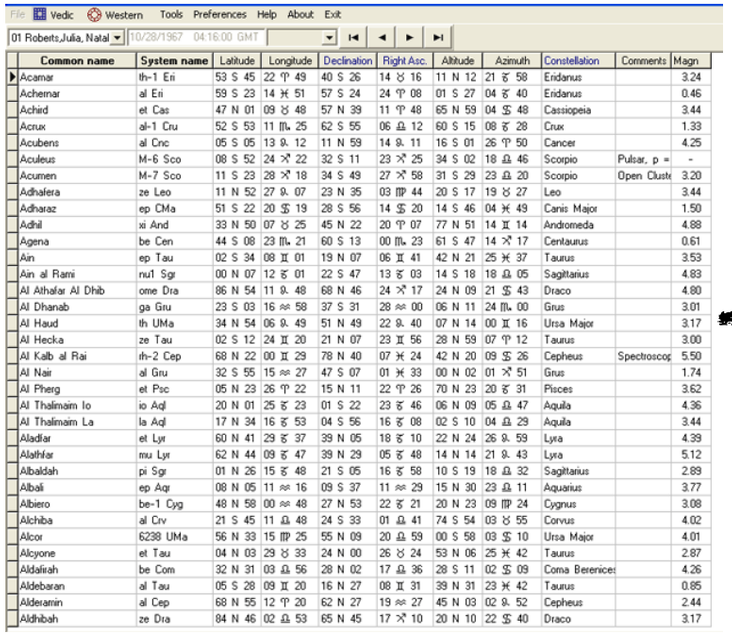
Spectral Cubes



Ra

$$C(\mathbf{x}) \simeq \hat{C}(\mathbf{x};\theta) + \epsilon = \sum_{i}^{k} \alpha_i K(\mathbf{x}, \mathbf{y}_i) + \epsilon,$$

2 on VO Catalogs Using


File Vedic Western Tools Preferences Help About Exit

01 Roberts,Julia, Natal 10/28/1967 04:16:00 GMT

Common name	System name	Latitude	Longitude	Declination	Right Asc.	Altitude	Azimuth	Constellation	Comments	Magn
Acamar	th-1 Eri	53 S 45	22 ♈ 49	40 S 26	14 ♉ 16	11 N 12	21 ♐ 58	Eridanus		3.24
Achernar	al Eri	59 S 23	14 ♓ 51	57 S 24	24 ♈ 08	01 S 27	04 ♐ 40	Eridanus		0.46
Achird	et Cas	47 N 01	09 ♉ 48	57 N 39	11 ♈ 48	65 N 59	04 ♋ 48	Cassiopeia		3.44
Acrux	al-1 Cru	52 S 53	11 ♏ 25	62 S 55	06 ♎ 12	60 S 15	08 ♐ 28	Crux		1.33
Acubens	al Cnc	05 S 05	13 ♌ 12	11 N 59	14 ♌ 11	16 S 01	26 ♈ 50	Cancer		4.25
Aculeus	M-6 Sco	08 S 52	24 ♐ 22	32 S 11	23 ♐ 25	34 S 02	18 ♎ 46	Scorpio	Pulsar, p =	-
Acumen	M-7 Sco	11 S 23	28 ♐ 18	34 S 49	27 ♐ 58	31 S 29	23 ♎ 20	Scorpio	Open Clust	3.20
Adhafera	ze Leo	11 N 52	27 ♌ 07	23 N 35	03 ♍ 44	20 S 17	19 ♉ 27	Leo		3.44
Adharaz	ep CMa	51 S 22	20 ♋ 19	28 S 56	14 ♋ 20	14 S 46	04 ♓ 49	Canis Major		1.50
Adhil	xi And	33 N 50	07 ♉ 25	45 N 22	20 ♈ 07	77 N 51	14 ♊ 14	Andromeda		4.88
Agena	be Cen	44 S 08	23 ♏ 21	60 S 13	00 ♏ 23	61 S 47	14 ♐ 17	Centaurus		0.61
Ain	ep Tau	02 S 34	08 ♊ 01	19 N 07	06 ♊ 41	42 N 21	25 ♓ 37	Taurus		3.53
Ain al Rami	nu1 Sgr	00 N 07	12 ♑ 01	22 S 47	13 ♑ 03	14 S 18	18 ♎ 05	Sagittarius		4.83
Al Athafar Al Dhib	ome Dra	86 N 54	11 ♌ 48	68 N 46	24 ♐ 17	24 N 09	21 ♋ 43	Draco		4.80
Al Dhanab	ga Gru	23 S 03	16 ♒ 58	37 S 31	28 ♒ 00	06 N 11	24 ♏ 00	Grus		3.01
Al Haud	th UMa	34 N 54	06 ♌ 49	51 N 49	22 ♌ 40	07 N 14	00 ♊ 16	Ursa Major		3.17
Al Hecka	ze Tau	02 S 12	24 ♊ 20	21 N 07	23 ♊ 56	28 N 59	07 ♈ 12	Taurus		3.00
Al Kalb al Rai	rh-2 Cep	68 N 22	00 ♊ 29	78 N 40	07 ♓ 24	42 N 20	09 ♋ 26	Cepheus	Spectroscop	5.50
Al Nair	al Gru	32 S 55	15 ♒ 27	47 S 07	01 ♓ 33	00 N 02	01 ♐ 51	Grus		1.74
Al Pherg	et Psc	05 N 23	26 ♈ 22	15 N 11	22 ♈ 26	70 N 23	20 ♑ 31	Pisces		3.62
Al Thalimain Io	io Aql	20 N 01	25 ♑ 23	01 S 22	23 ♑ 46	06 N 09	05 ♎ 47	Aquila		4.36
Al Thalimain La	la Aql	17 N 34	16 ♑ 53	04 S 56	16 ♑ 08	02 S 10	04 ♎ 29	Aquila		3.44
Aladfar	et Lyr	60 N 41	29 ♑ 37	39 N 05	18 ♑ 10	22 N 24	26 ♌ 59	Lyra		4.39
Alathfar	mu Lyr	62 N 44	09 ♑ 47	39 N 29	05 ♑ 48	14 N 14	21 ♌ 43	Lyra		5.12
Albaldah	pi Sgr	01 N 26	15 ♑ 48	21 S 05	16 ♑ 58	10 S 19	18 ♎ 32	Sagittarius		2.89
Albali	ep Aqr	08 N 05	11 ♒ 16	09 S 37	11 ♒ 29	15 N 30	23 ♎ 11	Aquarius		3.77
Albiero	be-1 Cyg	48 N 58	00 ♒ 48	27 N 53	22 ♑ 21	20 N 23	09 ♍ 24	Cygnus		3.08
Alchiba	al Crv	21 S 45	11 ♎ 48	24 S 33	01 ♎ 41	74 S 54	03 ♉ 55	Corvus		4.02
Alcor	6238 UMa	56 N 33	15 ♍ 25	55 N 09	20 ♎ 59	00 S 58	03 ♋ 10	Ursa Major		4.01
Alcyone	et Tau	04 N 03	29 ♉ 33	24 N 00	26 ♉ 24	53 N 06	25 ♓ 42	Taurus		2.87
Aldafirah	be Com	32 N 31	03 ♎ 56	28 N 02	17 ♎ 36	28 S 11	02 ♋ 09	Coma Berenices		4.26
Aldebaran	al Tau	05 S 28	09 ♊ 20	16 N 27	08 ♊ 31	39 N 31	23 ♓ 42	Taurus		0.85
Alderamin	al Cep	68 N 55	12 ♈ 20	62 N 27	19 ♒ 27	45 N 03	02 ♌ 52	Cepheus		2.44
Aldhibah	ze Dra	84 N 46	02 ♎ 53	65 N 45	17 ♐ 10	20 N 10	22 ♋ 40	Draco		3.17

Catalogs

Published

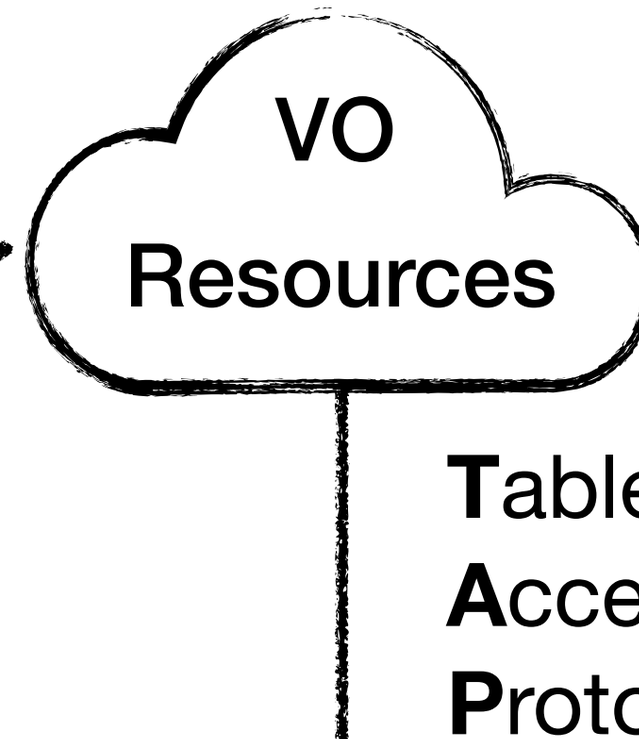
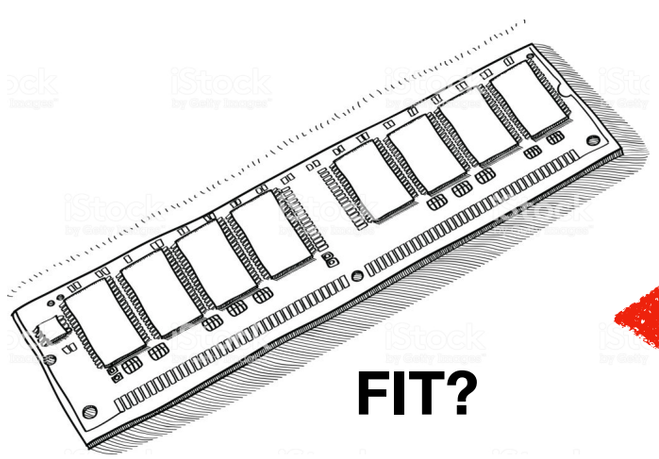


Table
Access
Protocol

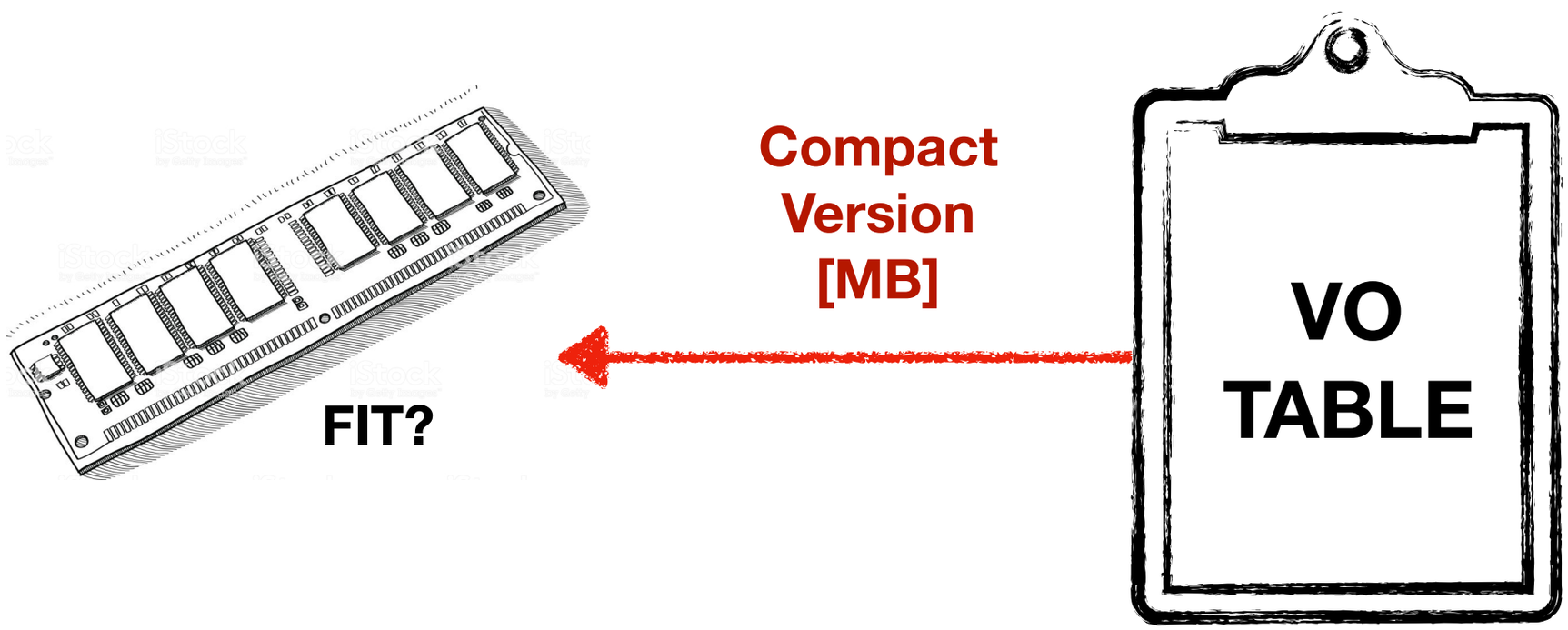
Compact Version [MB]



FIT?

1 Content-aware Data Discovery

2 on VO Catalogs Using



3 Succinct Representations

3 Succinct Representations

Compact Car

Compressed Car

Compact is faster for searching a parking!

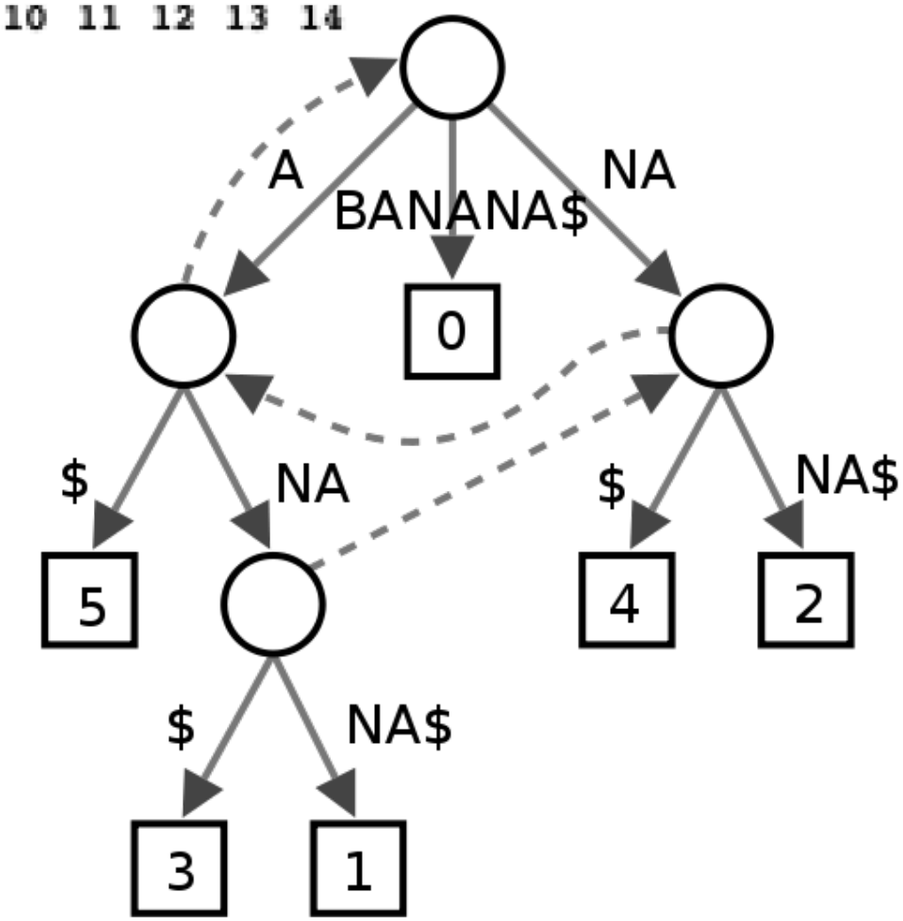
- A compact version of something **can be used!**
- They cannot do it all, but it might be **faster for some tasks**.
- In text mining these are called **succinct representations**

3 Succinct Representations

Suffix Arrays & Trees

S	E	N	S	E	L	E	S	S	N	E	S	S	#
1	2	3	4	5	6	7	8	9	10	11	12	13	14

LCP[i]	i	SA[i]	T[SA[i] ... n]
0	1	14	#
1	2	5	E L E S S N E S S #
1	3	2	E N S E L E S S N E S S #
3	4	11	E S S #
0	5	7	E S S N E S S #
0	6	6	L E S S N E S S #
1	7	10	N E S S #
0	8	3	N S E L E S S N E S S #
1	9	13	S #
2	10	4	S E L E S S N E S S #
1	11	1	S E N S E L E S S N E S S #
1	12	9	S N E S S #
2	13	12	S S #
	14	8	S S N E S S #



- Compressed Suffix Trees
- Wavelet Trees
- Burrows-Wheeler Transform

1 Content-aware Data Discovery
2 on VO Catalogs Using
3 Succinct Representations

TASK
Search for a pattern by relevance

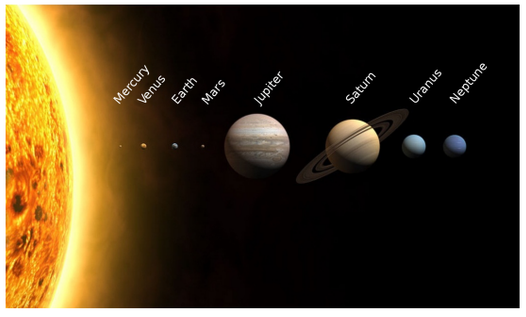
$$T = \langle \langle v_{1,1}, \ldots, v_{1,k} \rangle, \cdots, \langle v_{r,1}, \ldots, v_{r,k} \rangle \rangle$$

Mauricio Solar

Author

Solar

Science Description



Solar System

Star Description

Solar Masses

TASK
Search for a pattern by relevance

$$T = \langle\langle v_{1,1}, \ldots, v_{1,k}\rangle, \cdots, \langle v_{r,1}, \ldots, v_{r,k}\rangle\rangle$$

PostgreSQL

```
SELECT attName, COUNT (attName) AS CountOf
FROM public."{TABLE}"
WHERE value LIKE '%{QUERY}%'
GROUP BY attName;
```

TASK
Search for a pattern by relevance

- High complexity: patterns for which $1000 \leq \rho$;
- Medium complexity: patterns for which $10 \leq \rho < 1000$;
- Low complexity: patterns for which $\rho < 10$.

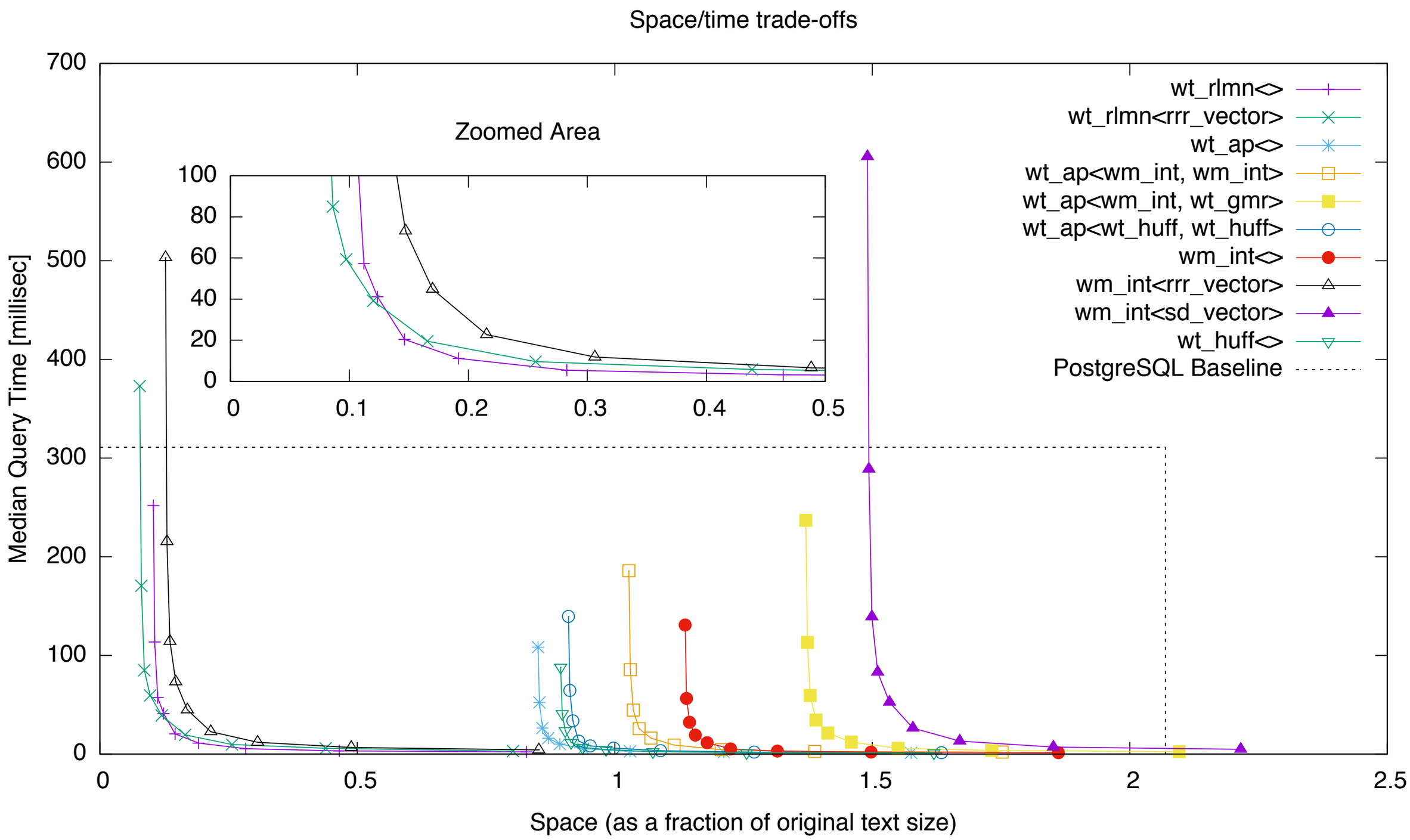
$$\rho = \frac{n_e}{n_c},$$

Word	Occurrences	Access Ratio	Complexity
'eso'	100,000	100009.00	High
'FORS2'	371,196	33809.50	High
'FORS'	143,286	29189.70	High
'GALAXIES'	10,799	53.00	Medium
'CLUSTER'	4,296	39.00	Medium
'SPECTROSCOPY'	2,728	28.00	Medium
'HISTORY'	1,684	9.00	Low
'GROUP'	1,236	8.00	Low
'Imaging'	14,432	7.50	Low

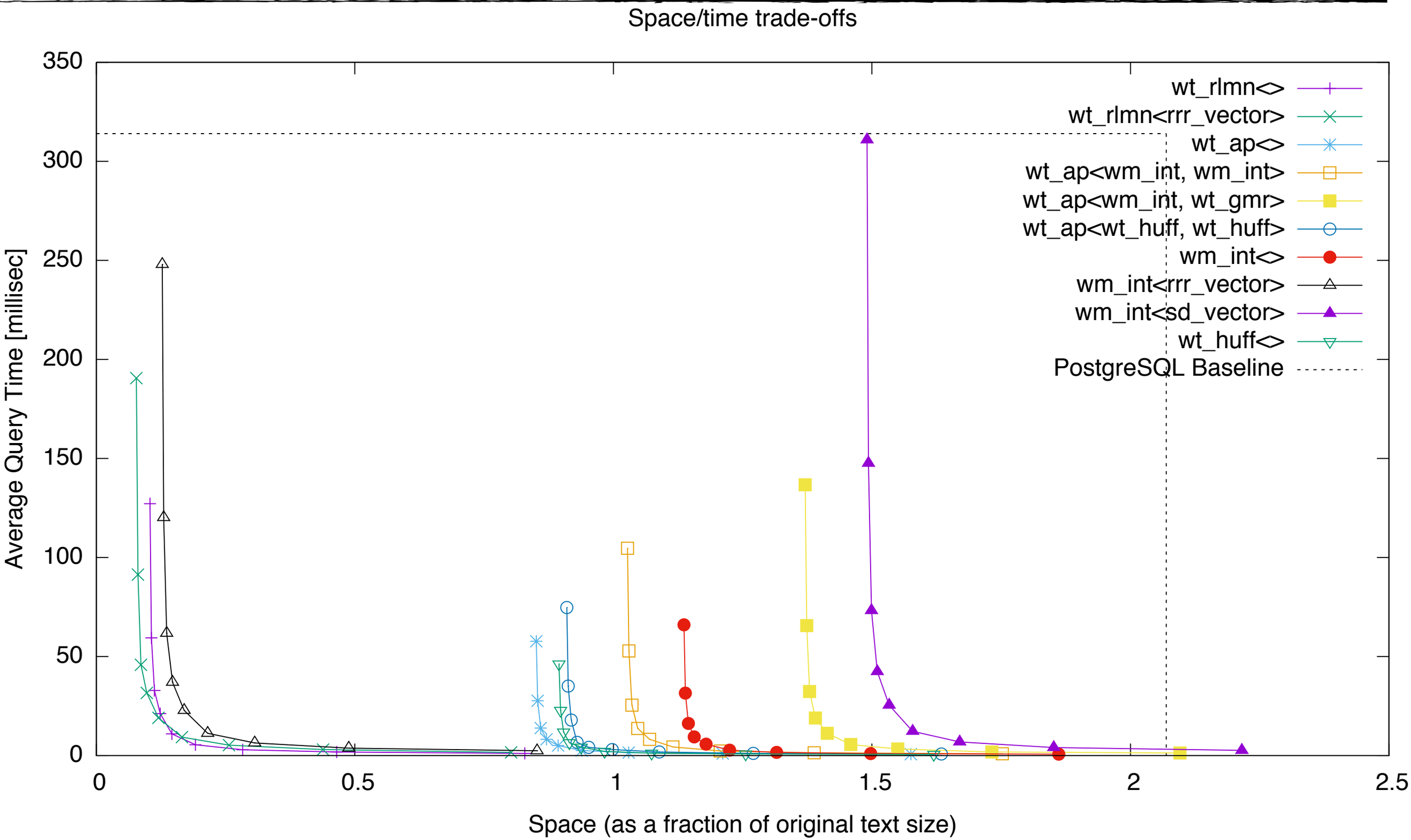
RESULTS
Compression

Data Structure	Space as a fraction of...		
	Original TAP file	Only data (no XML)	String s
wt_rlmn<>	0.26	0.56	0.30
wt_rlmn<rrr_vector>	0.24	0.51	0.27
wt_ap<>	1.01	2.17	1.17
wt_ap<wm_int, wm_int>	1.19	2.55	1.37
wt_ap<wm_int, wt_gmr>	1.53	3.28	1.77
wt_ap<wt_huff, wt_huff>	1.07	2.29	1.24
wm_int<>	1.30	2.78	1.50
wm_int<rrr_vector>	0.29	0.61	0.33
wm_int<sd_vector>	1.65	3.54	1.91
wt_huff<>	1.05	2.26	1.22
PostgreSQL baseline	2.07	4.45	2.40

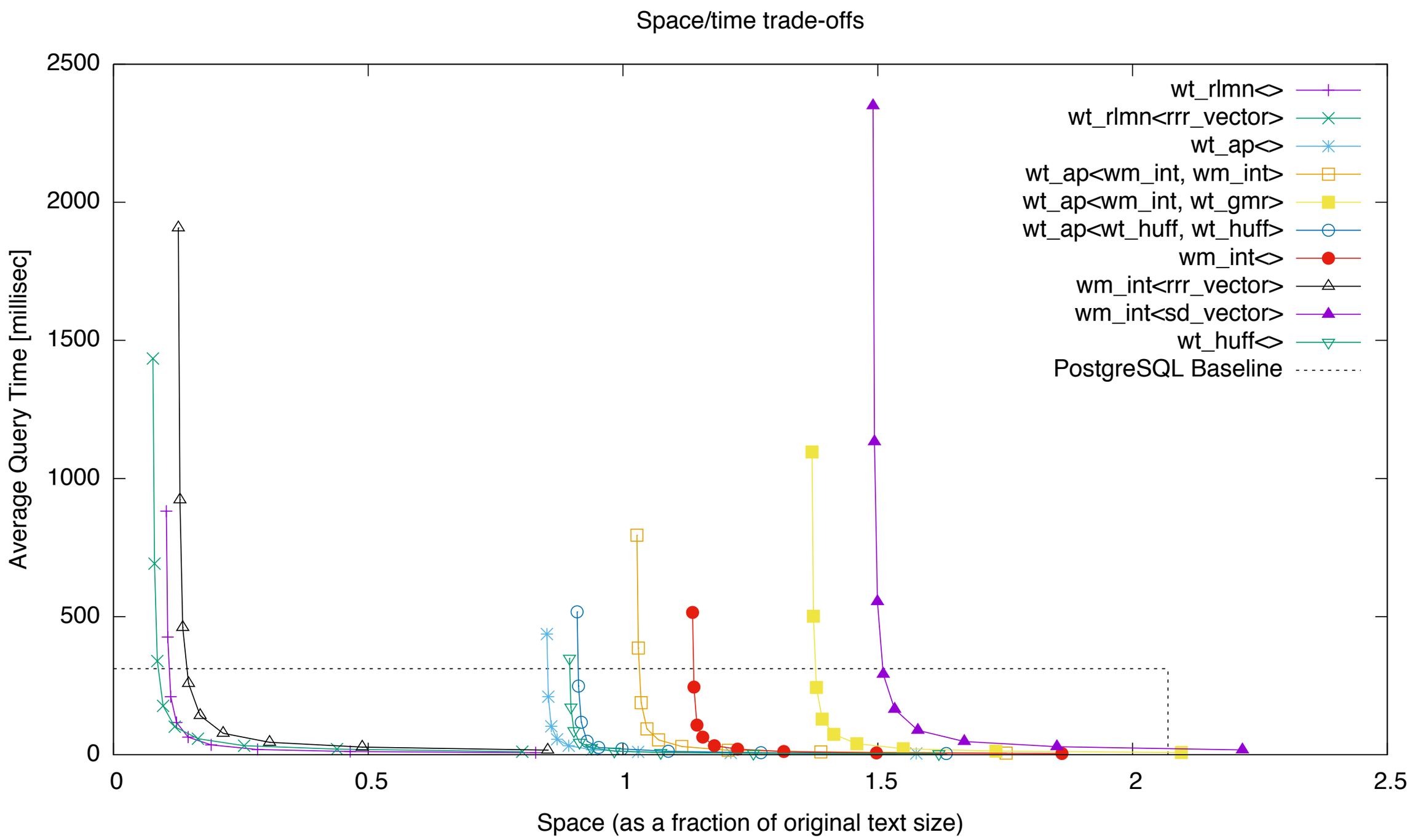
RESULTS
General Performance



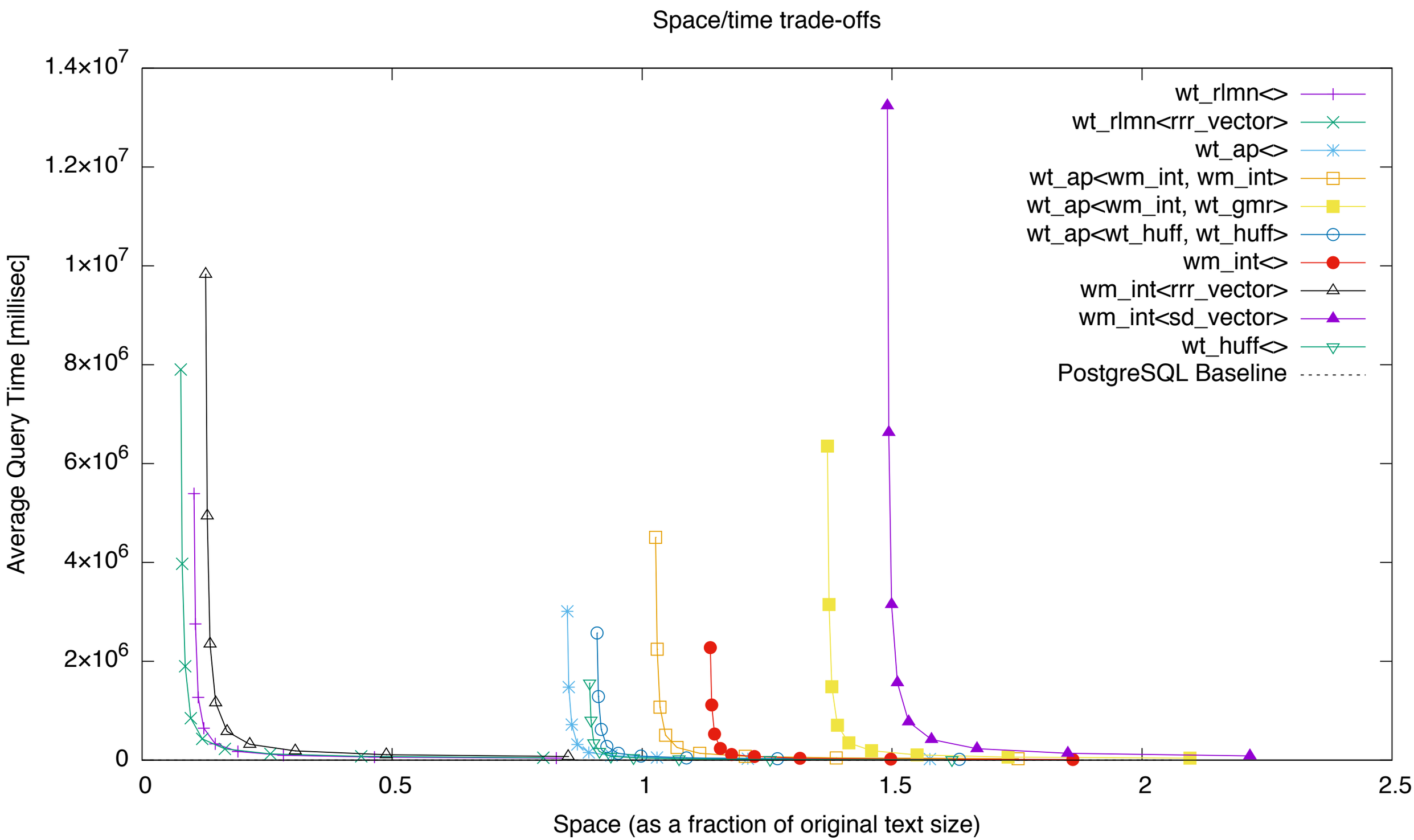
RESULTS
Low Complexity Performance



RESULTS
Medium Complexity Performance



RESULTS
High Complexity Performance



Conclusions

- Retrieves the relevance of a resource w.r.t to a query
- Outputs the columns and the occurrences in a fast way
- Intermediate files are light (15-25%), but not enough
- The worst case is expensive, need to improve.
- **Future:** need to avoid traversing the suffix array intervals
- **Future:** implement this as a service