Deduplication with graph algorithms

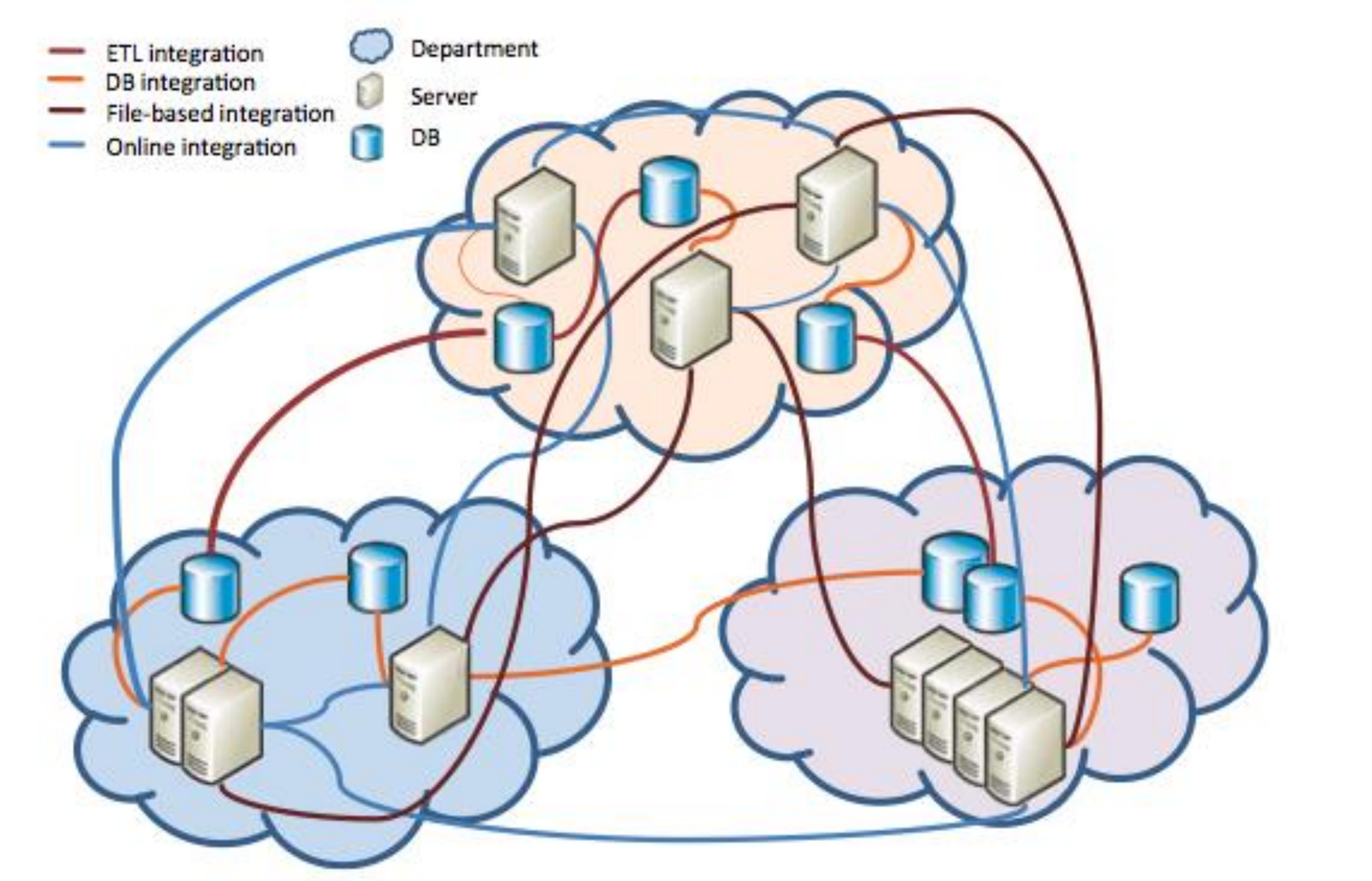
Tom Michiels

Vectr.
Consulting

Hoe kunnen we datascience en machine learning technieken (AI, ChatGPT…) inzetten voor de verbetering van bestaande en de productie van nieuwe openbare statistieken

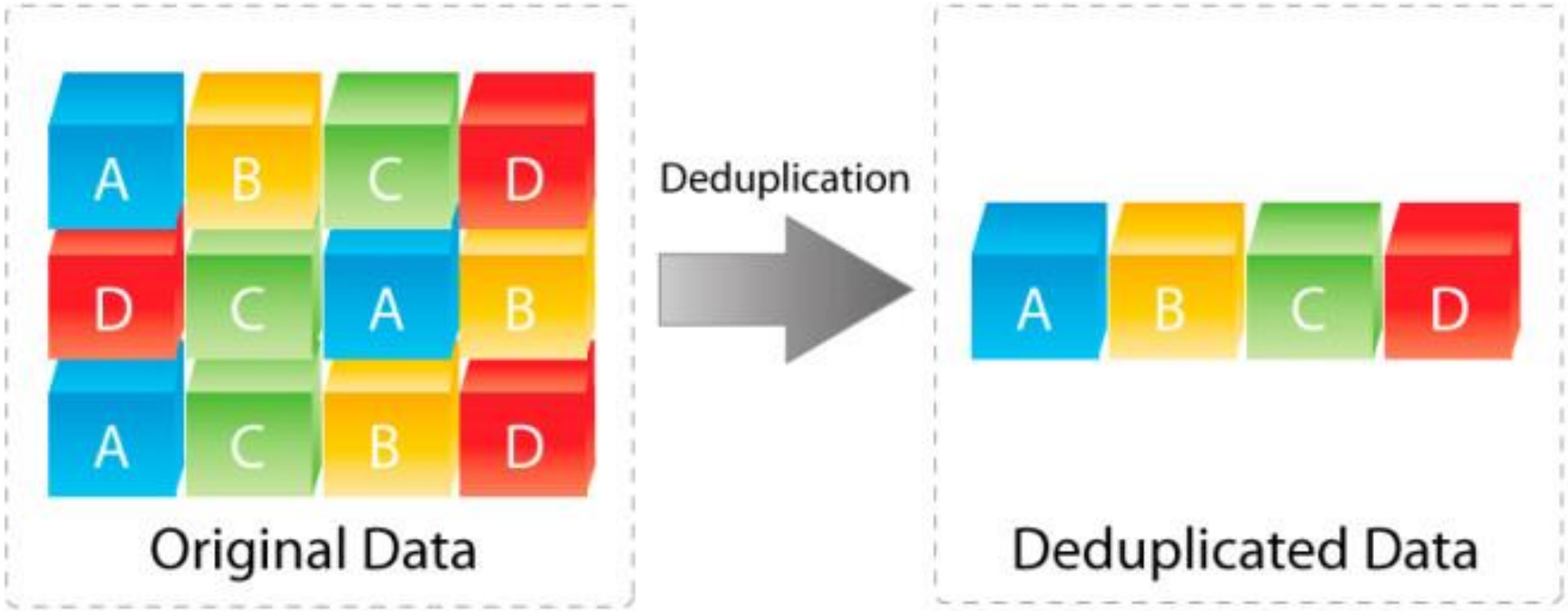
Vectr.
Consulting

Data Quality



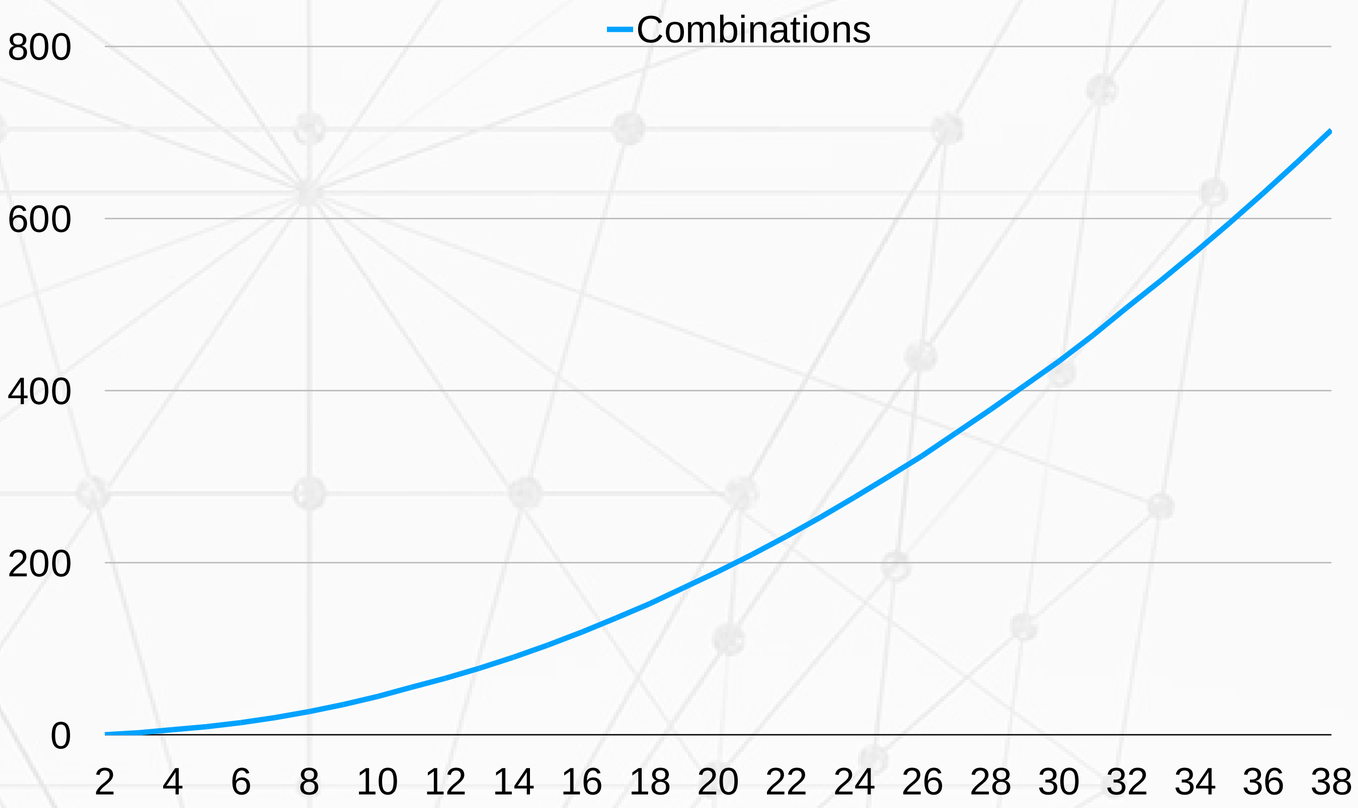
Vectr.
Consulting

Deduplication



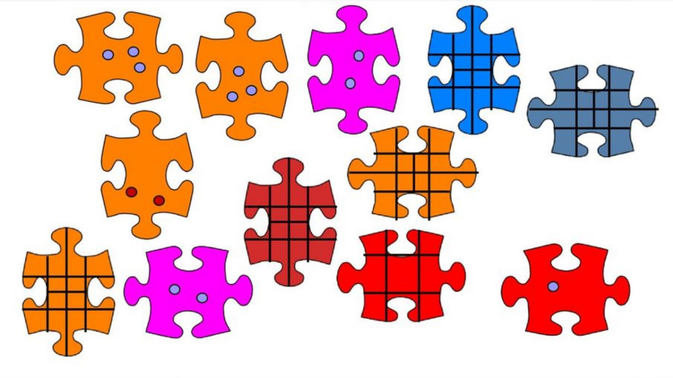
Vectr. Consulting

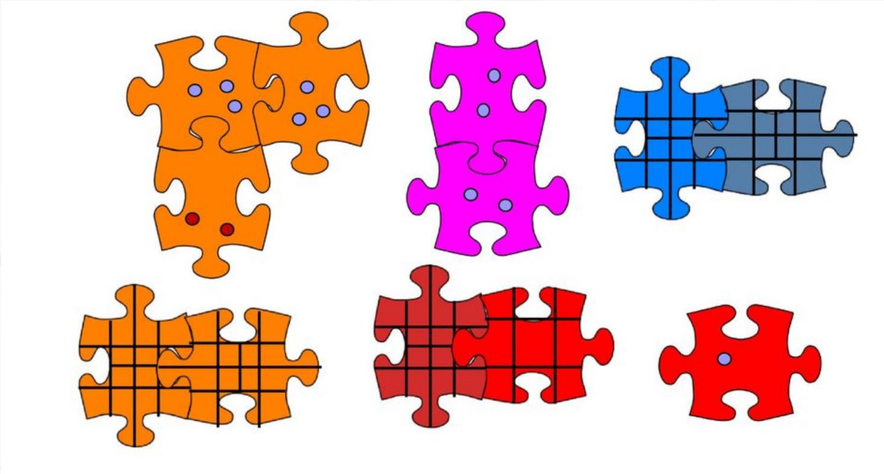
Blocking

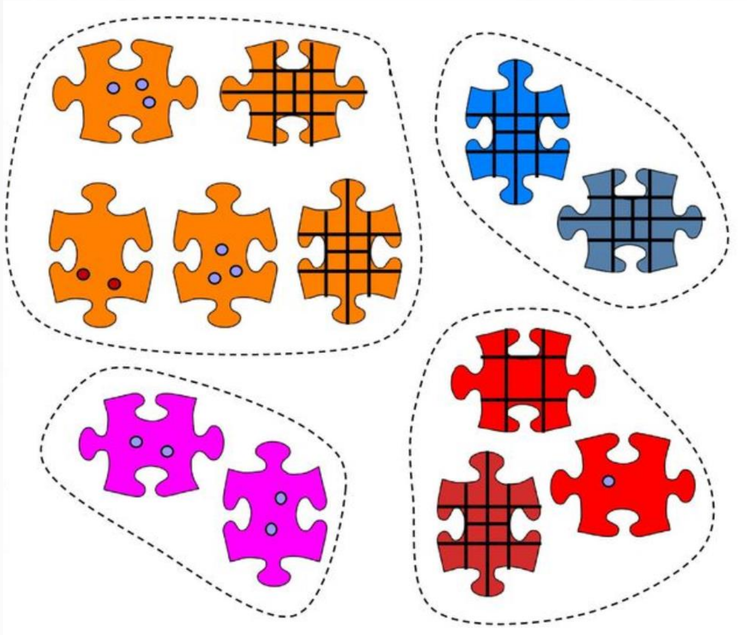


Vectr.
Consulting

Blocking







Vectr.
Consulting

Similarity functions

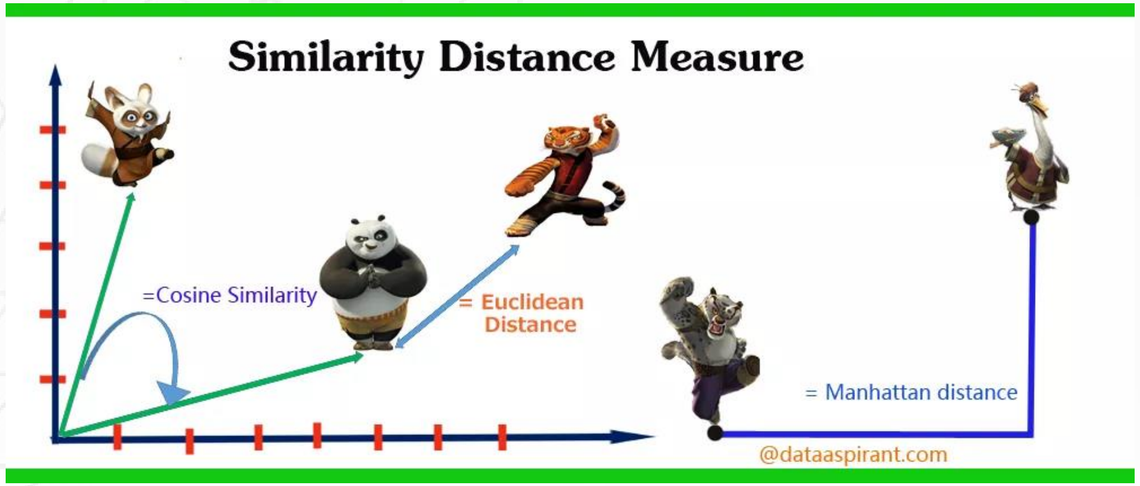
		S	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

← OPERATION #1 - Delete "a"

← OPERATION #2 - Delete "t"

↖ OPERATION #3 - Transform "r" to "n"

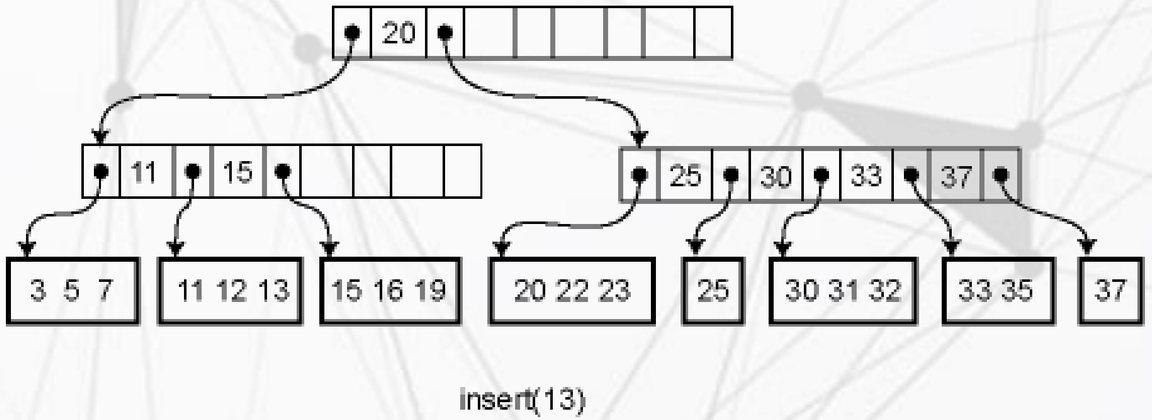
← 3 operations in total



Vectr. Consulting

Traditional Solution

- Index based



Vectr. Consulting

Pairwise Combination problem

Table 1

RecordID	Naam	Voornaam	Tel	Email	Twitter	CookieId
A	Michiels	T		tom@gmail.com	@TmMichiels	3445454
B	Michiels	T	027599608		@TmMichiels	
C	Moreau	M	027599608	t+m@gmail.com		
D	Michiels	T		t+m@gmail.com		
E	Michiels	T		t+m@gmail.com		3556565

Vectr. Consulting

Pairwise Combination problem

Table 1

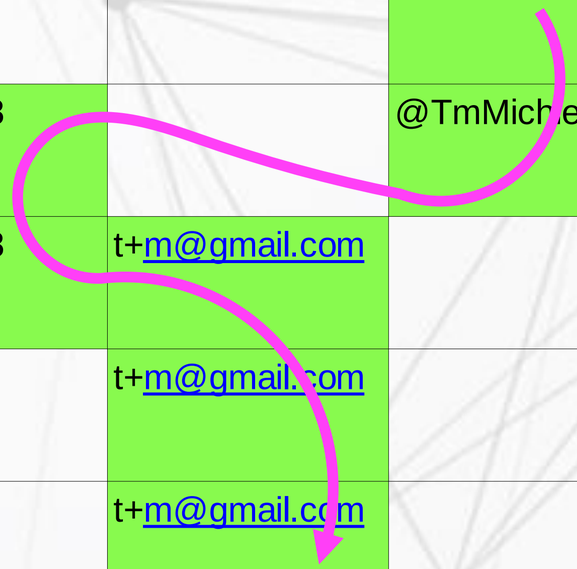
RecordID	Naam	Voornaam	Tel	Email	Twitter	CookieId
A	Michiels	T		tom@gmail.com	@TmMichiels	3445454
B	Michiels	T	027599608		@TmMichiels	
C	Moreau	M	027599608	t+m@gmail.com		
D	Michiels	T		t+m@gmail.com		
E	Michiels	T		t+m@gmail.com		3556565

Vectr. Consulting

Pairwise Combination problem

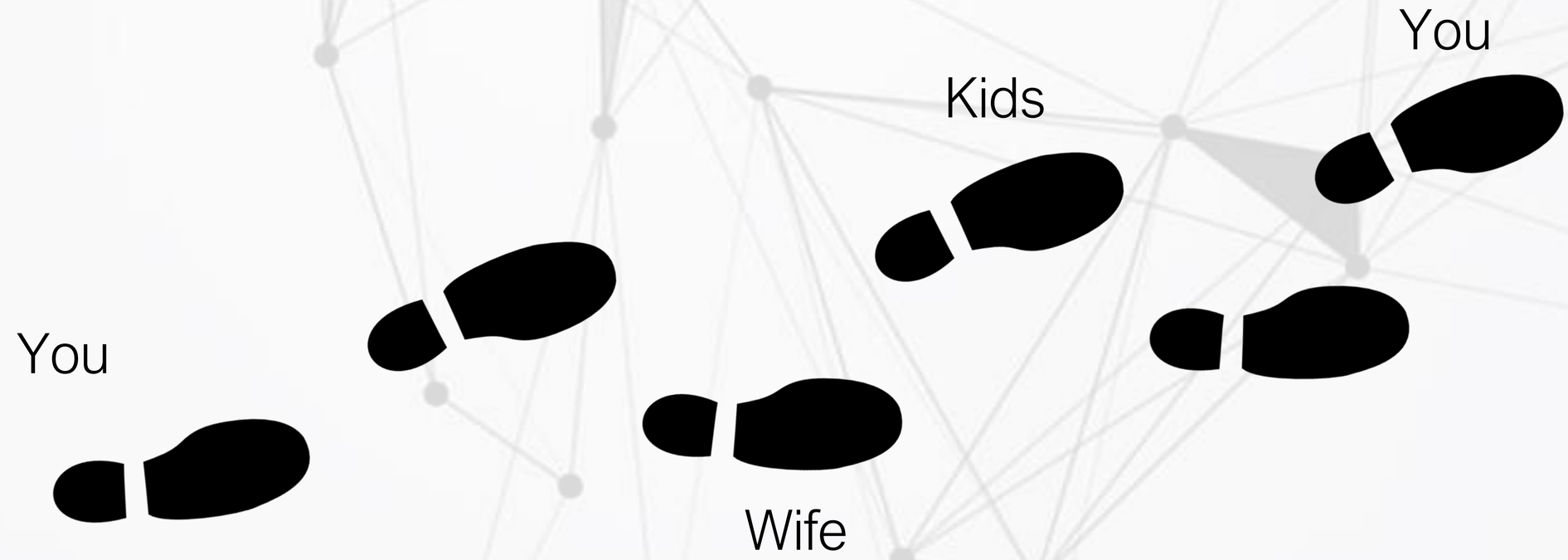
Table 1

RecordID	Naam	Voornaam	Tel	Email	Twitter	CookieId
A	Michiels	T		tom@gmail.com	@TmMichiels	3445454
B	Michiels	T	027599608		@TmMichiels	
C	Moreau	M	027599608	t+m@gmail.com		
D	Michiels	T		t+m@gmail.com		
E	Michiels	T		t+m@gmail.com		3556565



Vectr. Consulting

Graphs to the rescue



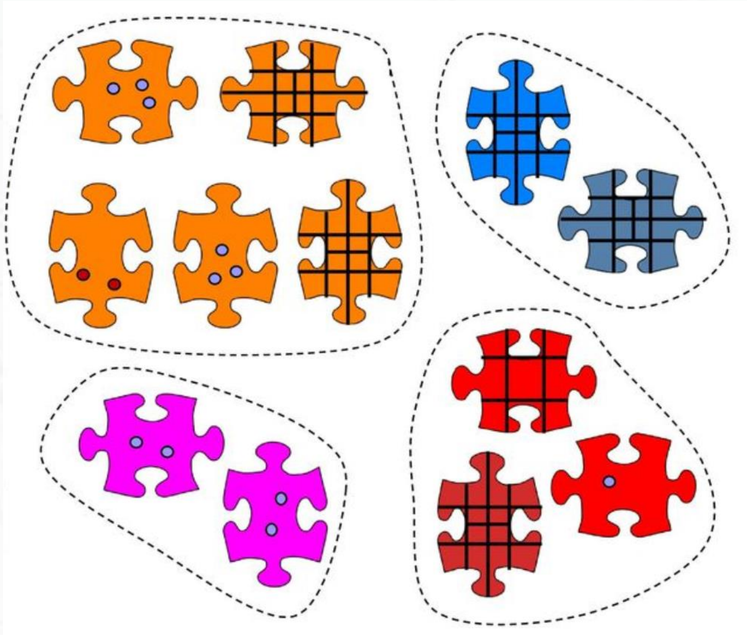
Vectr.
Consulting

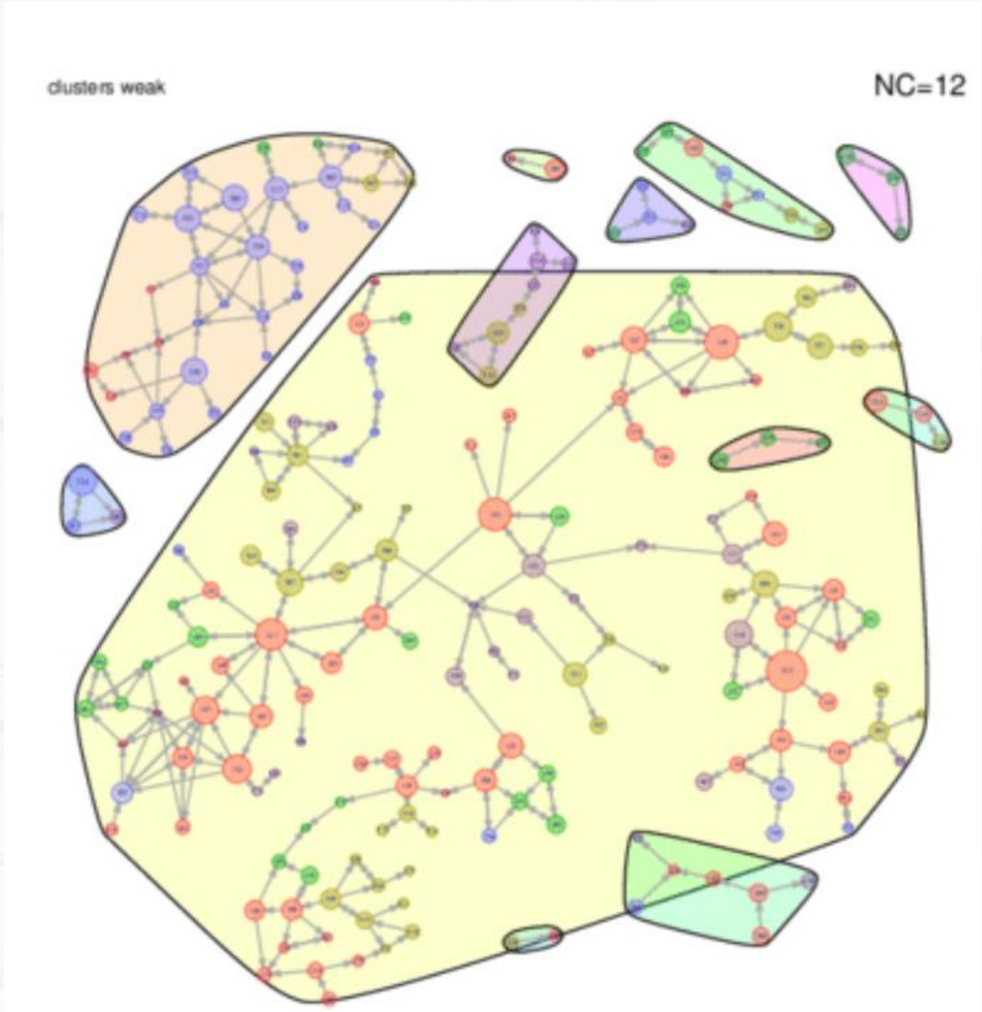
Graphs to the rescue

- Graph locality
- Pointer hopping
- Graph algorithms

Vectr.
Consulting

Graphs to the rescue



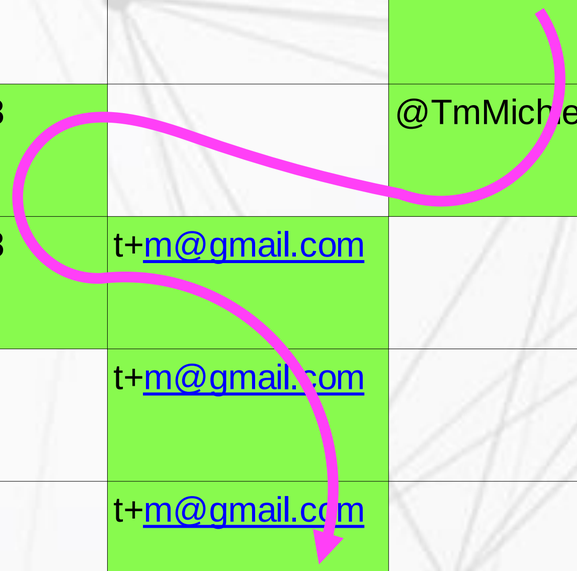


Vectr.
Consulting

Weakly connected component

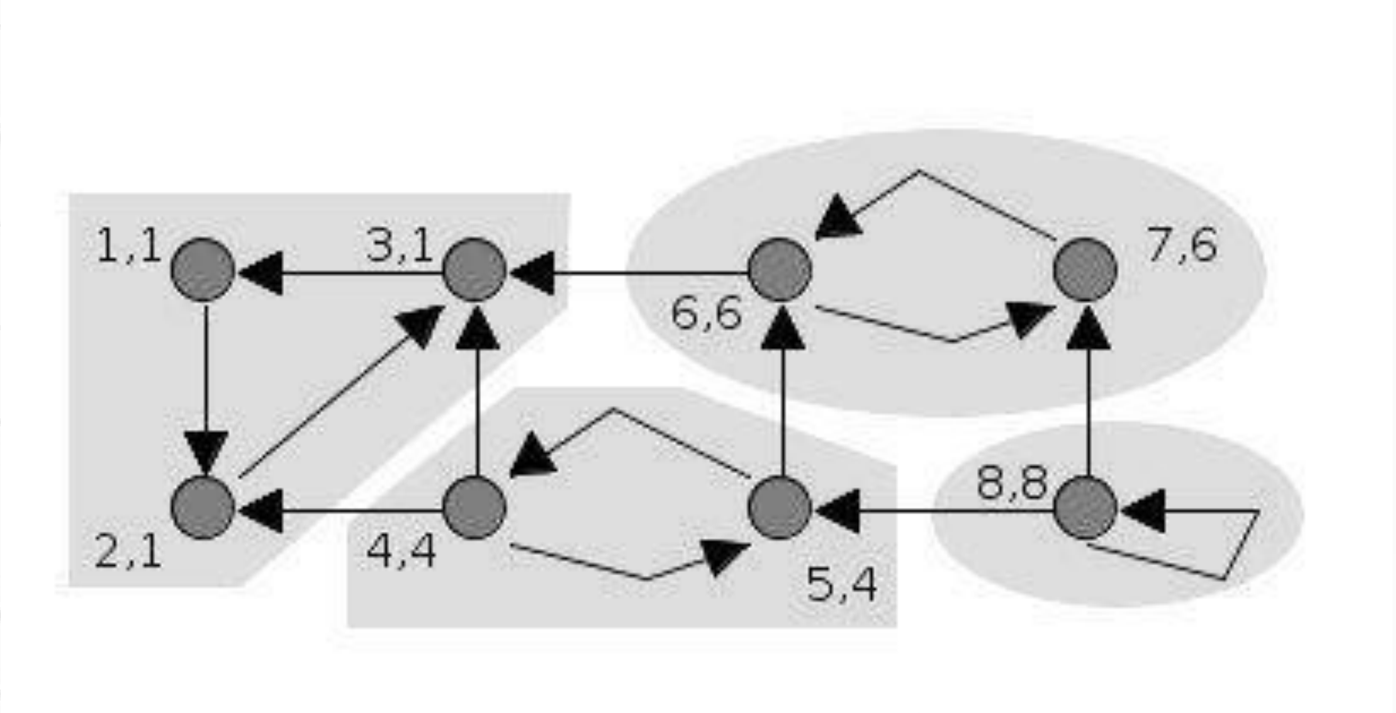
Table 1

RecordID	Naam	Voornaam	Tel	Email	Twitter	CookieId
A	Michiels	T		tom@gmail.com	@TmMichiels	3445454
B	Michiels	T	027599608		@TmMichiels	
C	Moreau	M	027599608	t+m@gmail.com		
D	Michiels	T		t+m@gmail.com		
E	Michiels	T		t+m@gmail.com		3556565



Vectr. Consulting

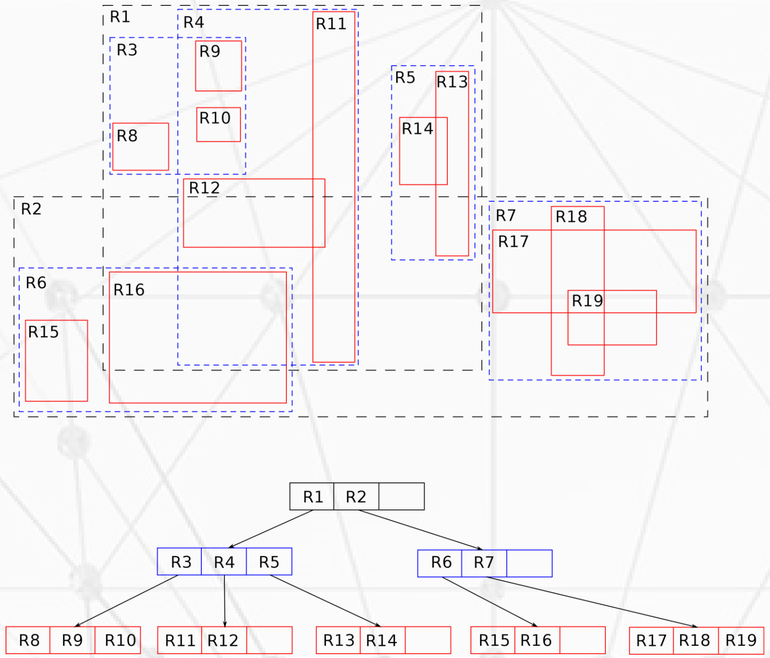
Tarjan’s connected component



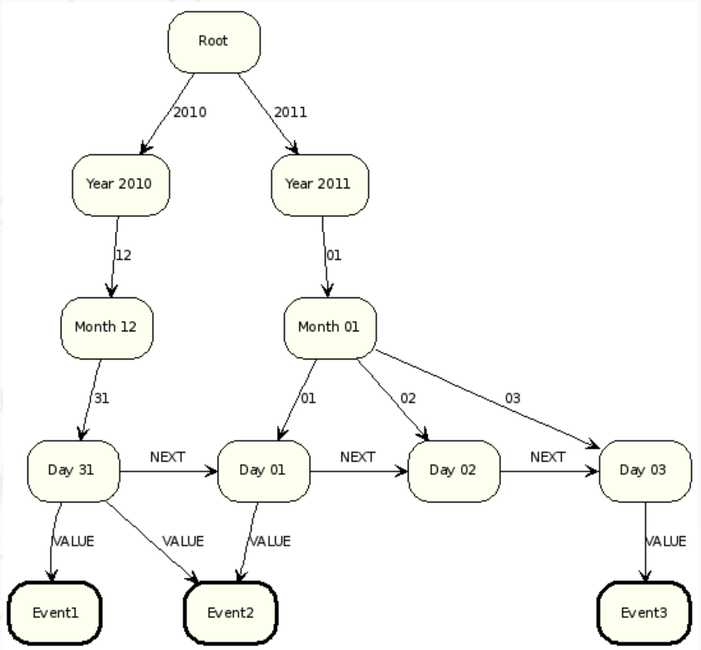
Vectr.
Consulting

Extension - Compare in Space and Time.

R-Tree

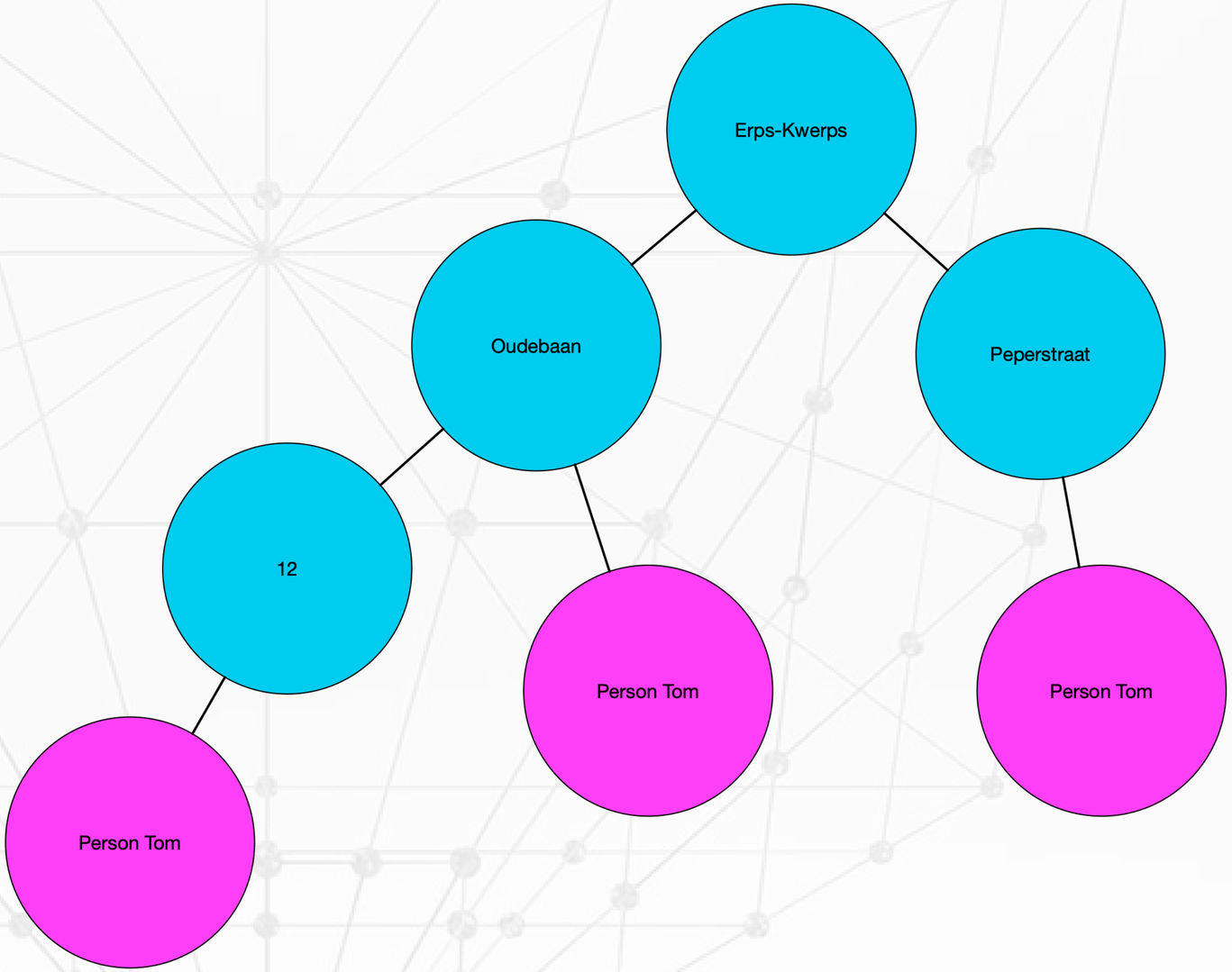


Timetree



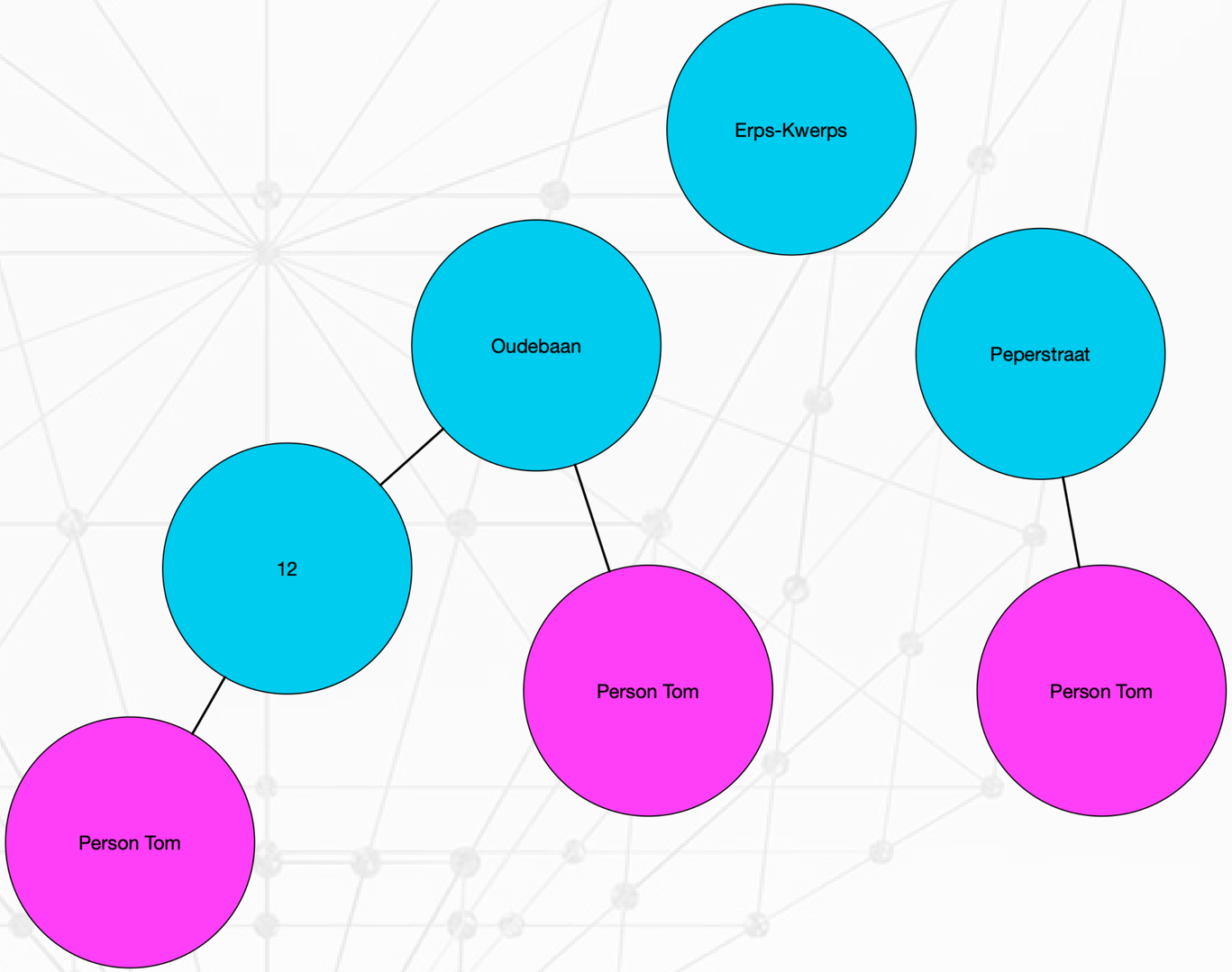
Vectr. Consulting

Extension - Compare in Space and Time.



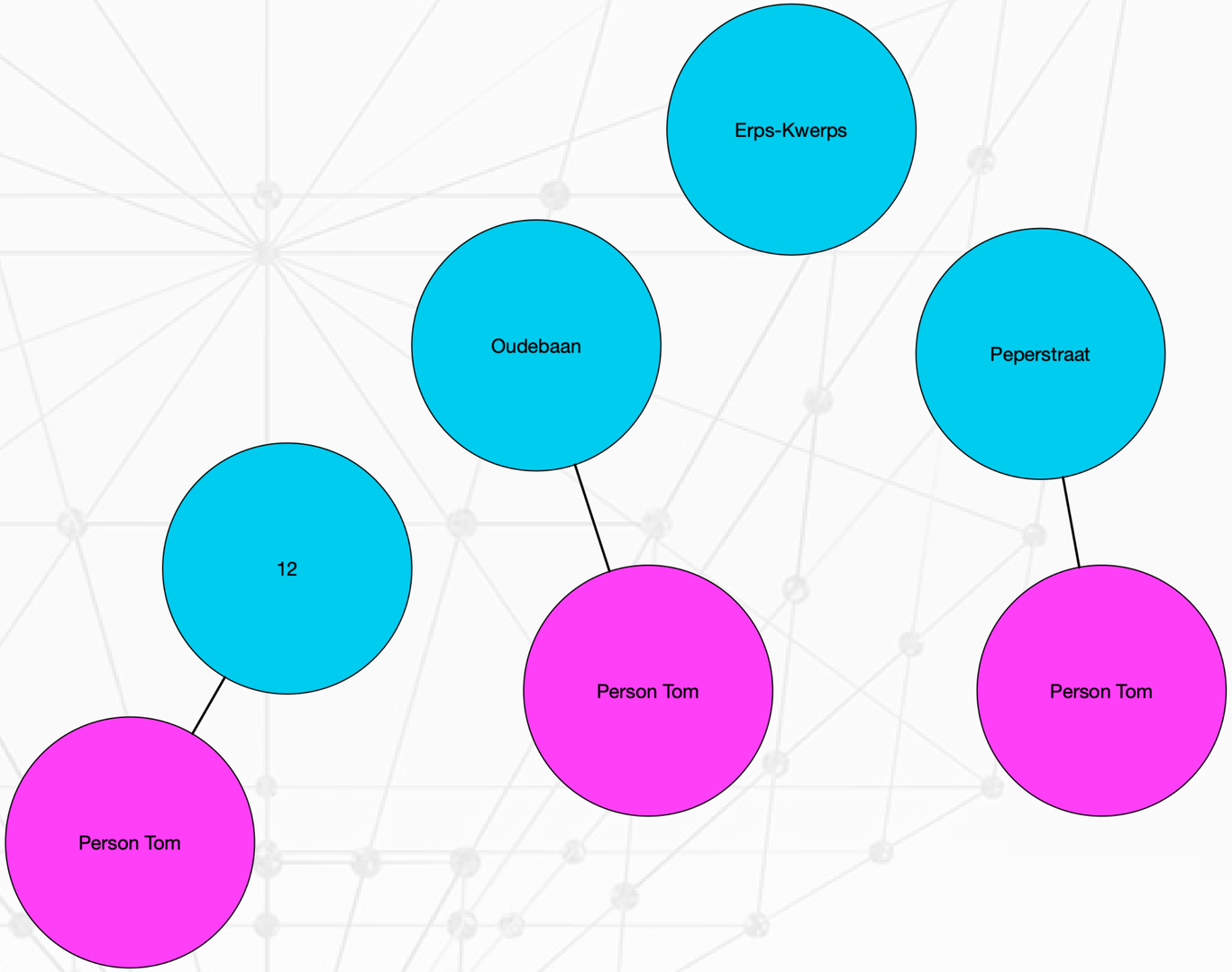
Vectr.
Consulting

Extension - Compare in Space and Time.



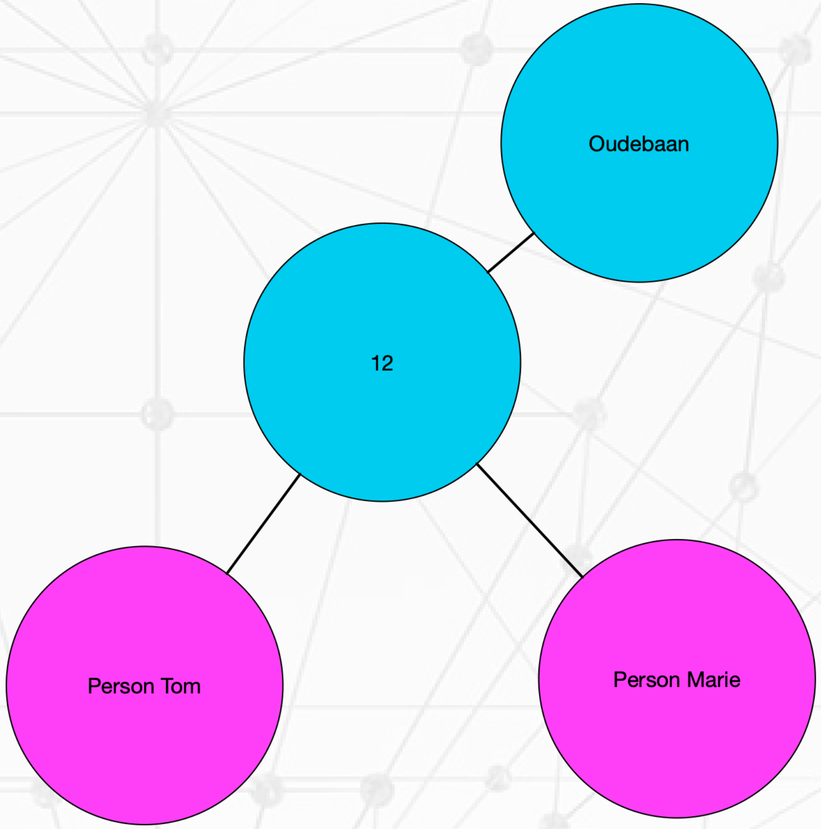
Vectr.
Consulting

Extension - Compare in Space and Time.



Vectr.
Consulting

Added Value Connected Component = Household



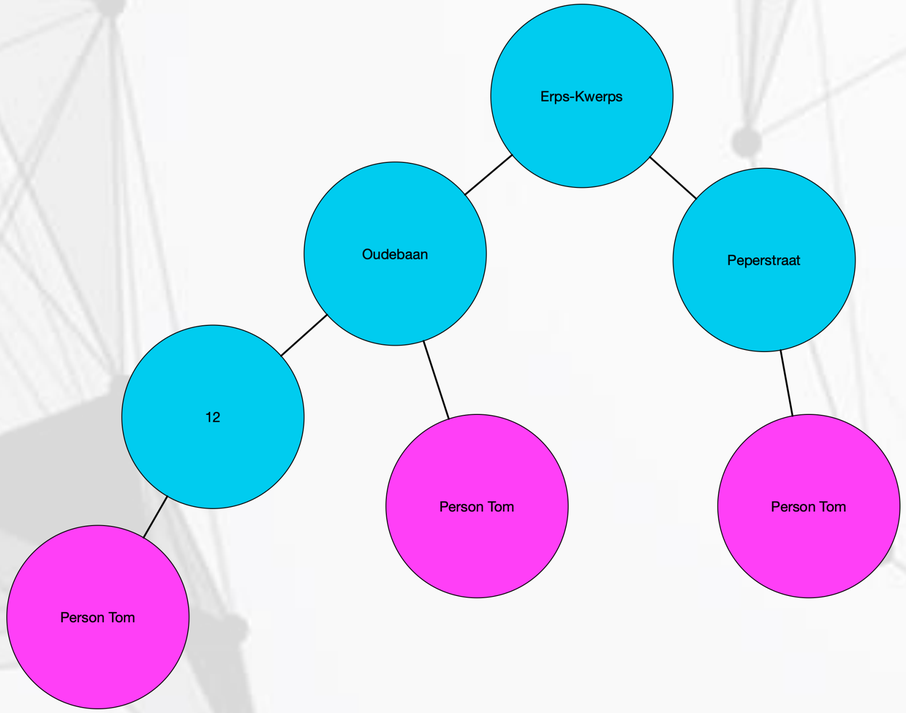
Vectr.
Consulting

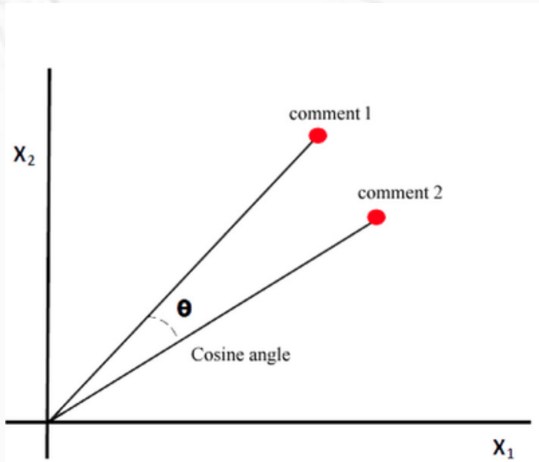
Advantages

- Linking through other records (Paths)
- Finding candidates through trees (Time - Location)
- Deduplication and Cluster detection is 2 step process.

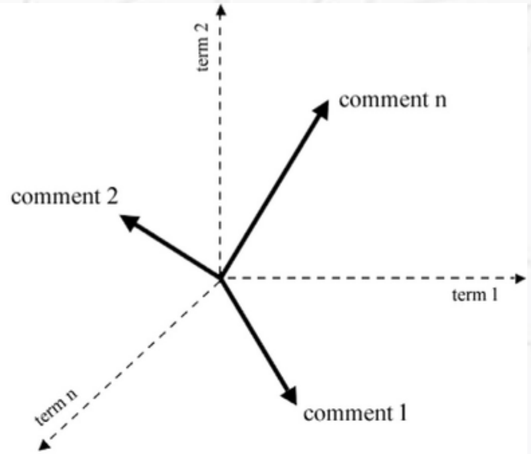
Vectr.
Consulting

Graph2Vec





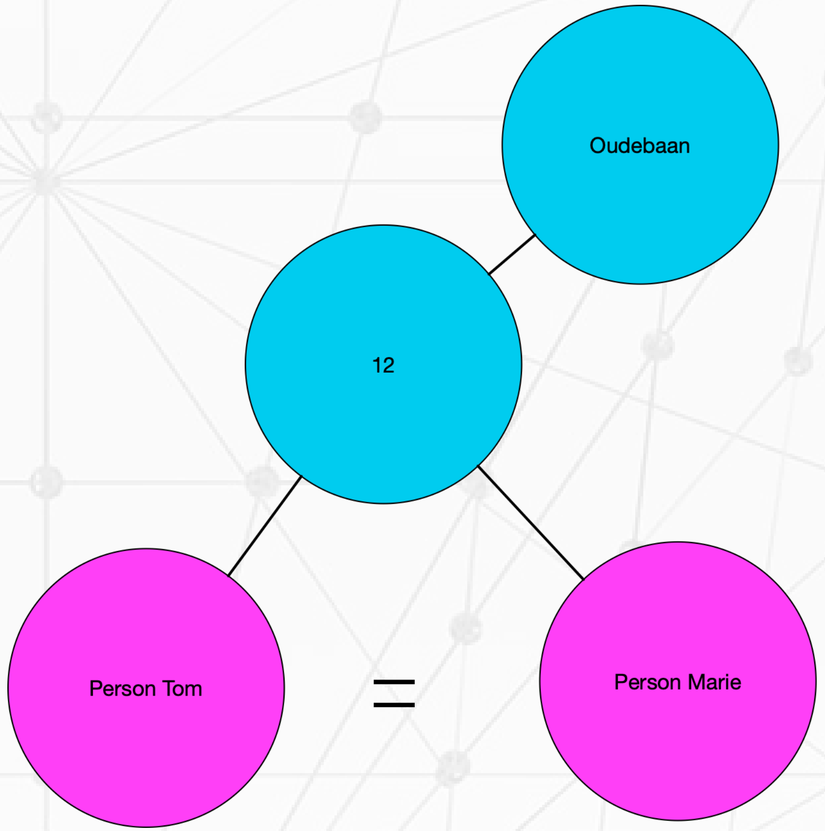
Cosine Similarity



Vector Space Model

Vectr. Consulting

Comparison of pairs



Vectr.
Consulting

Comparison

RecordID	Naam Metric1	Voornaam Metric2	Naam occurence	Voornaam occurence	Match
A - B	0,6	1	0,1	0,1	True
A - C	0,8	0,9	0,1	0,1	False

Vectr. Consulting

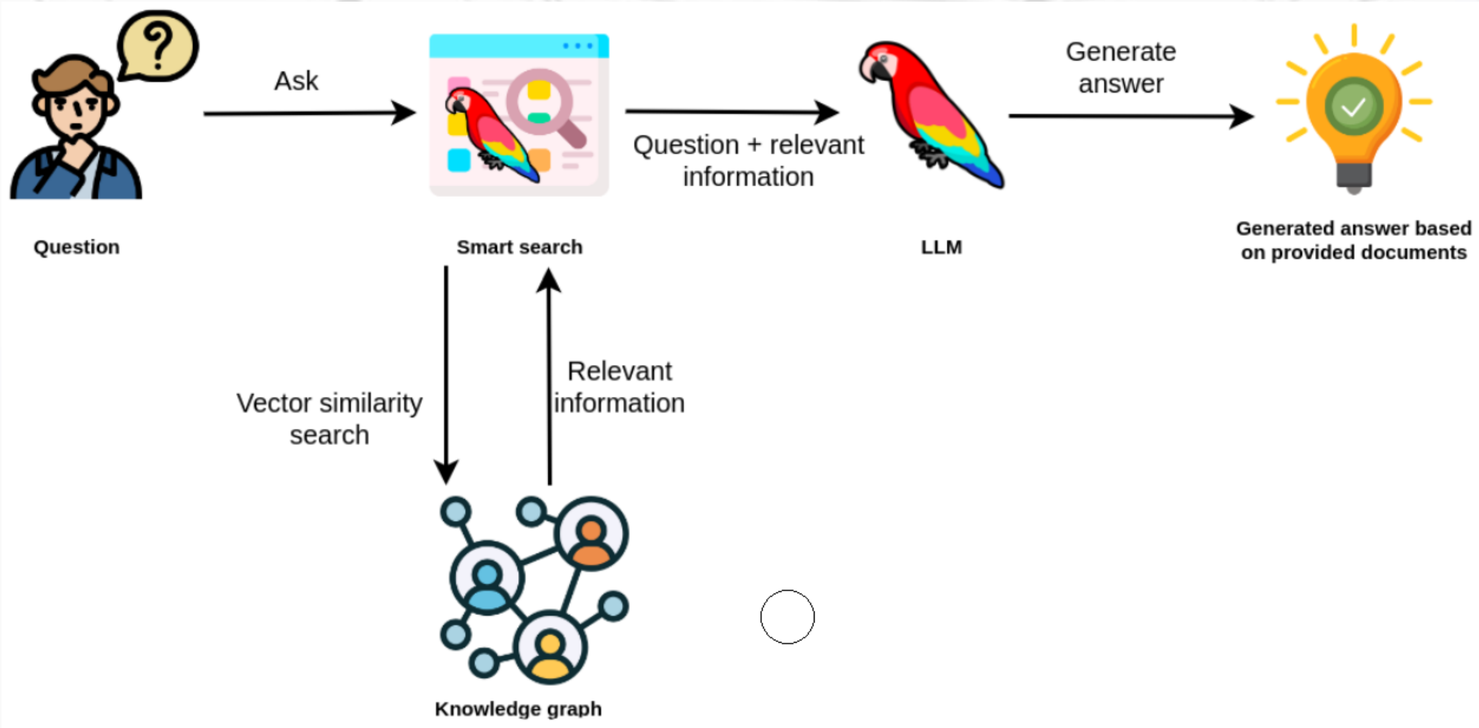
Comparison

- Rule based (Deterministic)
- Machine Learning. Active Learning, Random forests, …

Vectr.
Consulting

Other Possibilities

- Find Documents linked to each other by entities (Same principles as finding Pairs)



Vectr. Consulting

Fraud Detection

- Find things that do not want to be found

Vectr. Consulting

Conclusion

- Data Quality In improves data quality out.
- Graph algorithms can be used in other use cases like

Vectr.
Consulting