Searching over Streams

with Luwak, Kafka and Samza

Alan Woodward - alan@flax.co.uk - @romseygeek

flax
The Open Source Search Specialists

flax

The Open Source Search Specialists

- We build, tune and support fast, accurate and highly scalable search, analytics and Big Data applications
- We use (and create) **open source** software
- We're independent, honest and have 15+ years experience
- We also:
  - Run and attend many events & conferences
  - Write extensively about search & related matters
  - Train and mentor
- We're confluent partners






flax
The Open Source Search Specialists

NLA media access

FT.com FINANCIAL TIMES

REED SPECIALIST RECRUITMENT

Office for National Statistics

News UK

EMBL-EBI

INFOMEDIA

eagle

CabinetOffice

GIG CYMRU NHS WALES Gwasanaeth Gwybodeg Informatics Service

ARACHNYS

CAMBRIDGE UNIVERSITY PRESS

aap

Gorkana Group
Gorkana | durrants | metrica

WESTCOAST

What is a stream?

Log file

flax
The Open Source Search Specialists

Log file

Append

flax
The Open Source Search Specialists

Log file

Append

tail -f

How do you search one?

flax
The Open Source Search Specialists

- Search requires an **inverted index**
- Maps searchable terms to documents in which they appear
- Can also store additional information (positions, frequencies, offsets) for more complex searches, relevancy scoring, highlighting, etc
- Updating is difficult - dealt with in e.g. Lucene by writing multiple small immutable indexes, and merging in the background

FOSDEM

Results for **FOSDEM**

Save

Top / **All**

Lilian @lliillian · 3m

@StOnSoftware @SannyGR @**fosdem** nee staat er volgens mij los van.

View conversation

Henrik Lindberg @hel · 4m

On my way to #**fosdem**

Inderpreet Singh @ip_v1 · 5m

Every once in a while... "@kartben: Where the f@# is my passport when I need it?? :(((#**fosdem** #jfokus :("

FrOSCon @froscon · 9m

Several members of our team will be attending **FOSDEM** this weekend. If you want to meet give us a shout.

miracee @miraceesusanne · 10m

@ScottyTM viel Spass - Grüß Brüssel und alle Freaks der **FOSDEM** - ich trink ein Geb-Bier für Euch mit.

View conversation

Andreas Kupfer @ScottyTM · 11m

On my way to Brussels for #**FOSDEM** with ICE 16. #dbl

FOSDEM

Results for **FOSDEM**

Save

Top / **All**

8 new results

Solarus @Solarus0 · 2m

@lhirlimann deploying a 6to4 proxy is a good practice however. ;) @RatZillaS @**fosdem**

View conversation

Mohamed Hadrouj @hadrouj · 2m

Presentation of Juju by @lazypower at #**FOSDEM** : Full house at the config management room

Searching streams

- Cache all the queries that we're interested in
- Divide the incoming messages into windows, and build indexes on them
- Run cached queries over each index as it becomes available

Log file

Append

1 2 3 4 5 6 7 8 9 10 11 12

yesterday

last hour

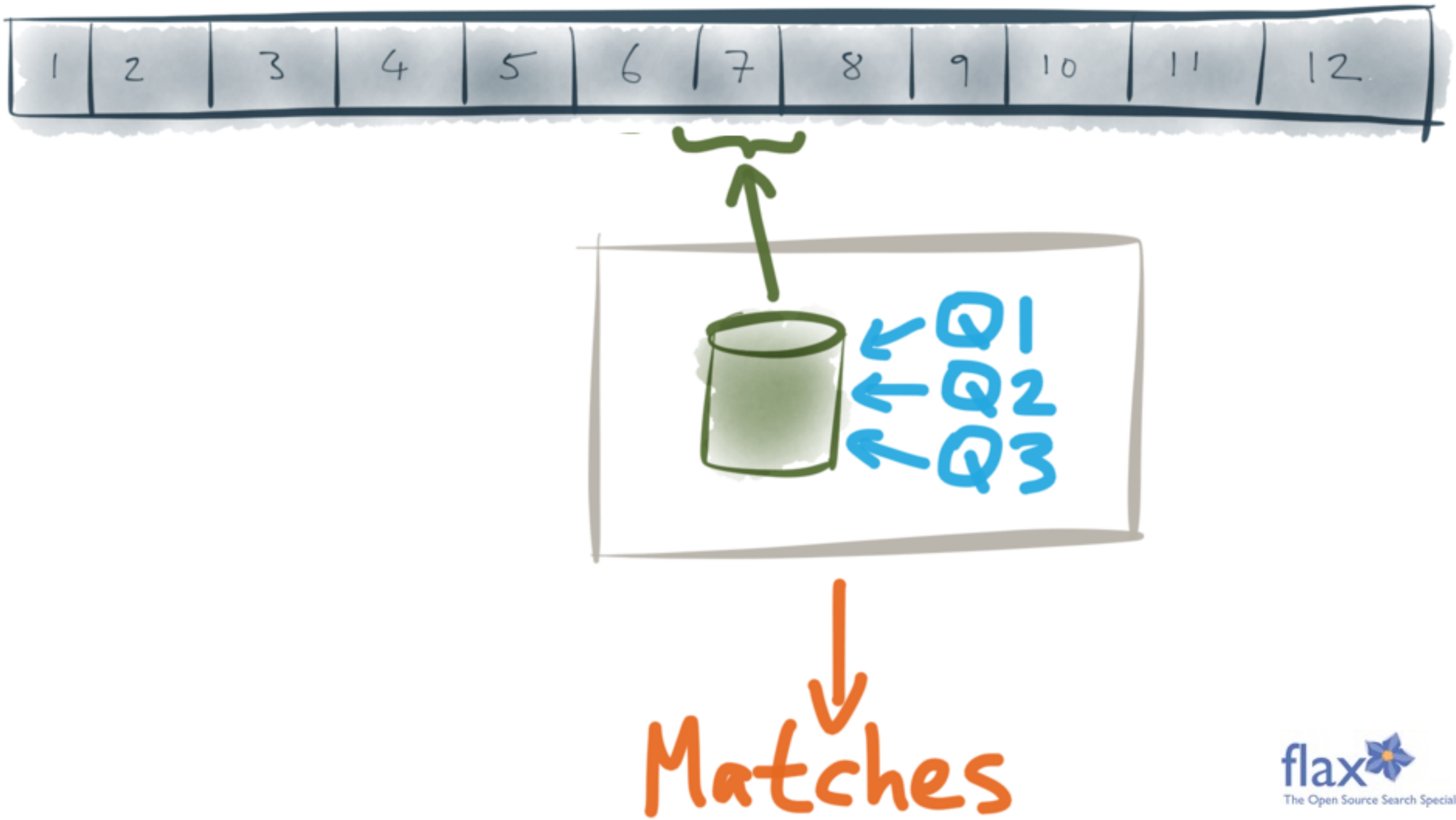
last minute

flax
The Open Source Search Specialists

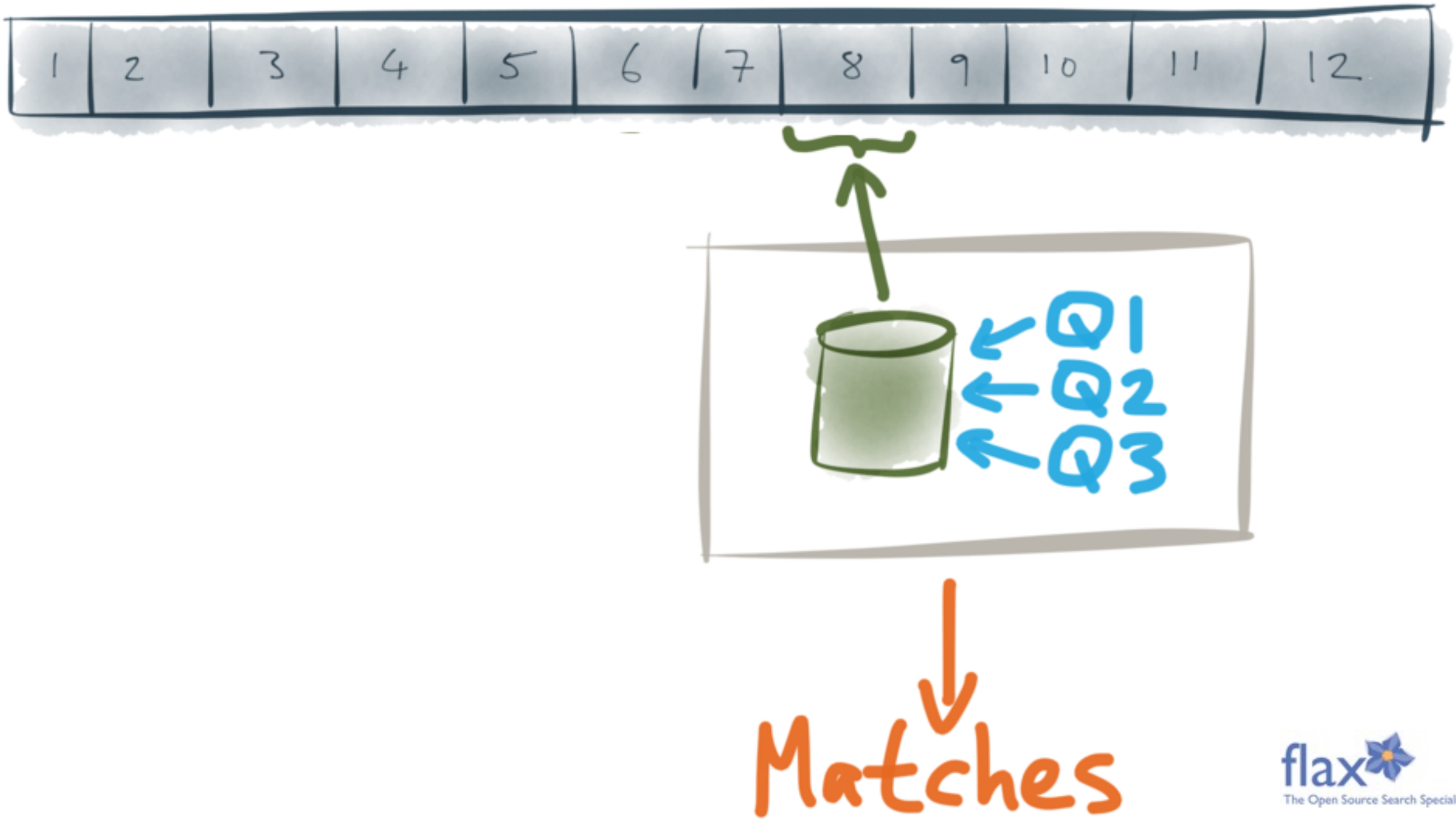
One document, many queries

flax
The Open Source Search Specialists

One document, many queries



One document, many queries



- Example: elasticsearch percolator
- Register queries into the percolator index
- Send documents one-at-a-time to elasticsearch, and it reports back which registered queries match

Does it scale?

flax
The Open Source Search Specialists

Lots of queries = slow!

Complex queries = slow!

flax
The Open Source Search Specialists

Scaling streaming search

- Batch up documents - trade off latency for throughput
- Try and filter out queries that you know won't match

flax
The Open Source Search Specialists

Index of queries

Q1: "WHEELS" NEAR "BUS"

Q2: "WHEELS" NEAR "CAR"

Q3: "WHEELS" OR "BUMPERS"

Index of queries

Q1: "WHEELS" NEAR "BUS" → "BUS"

Q2: "WHEELS" NEAR "CAR" → "CAR"

Q3: "WHEELS" OR "BUMPERS"

→ "BUMPERS", "WHEELS"

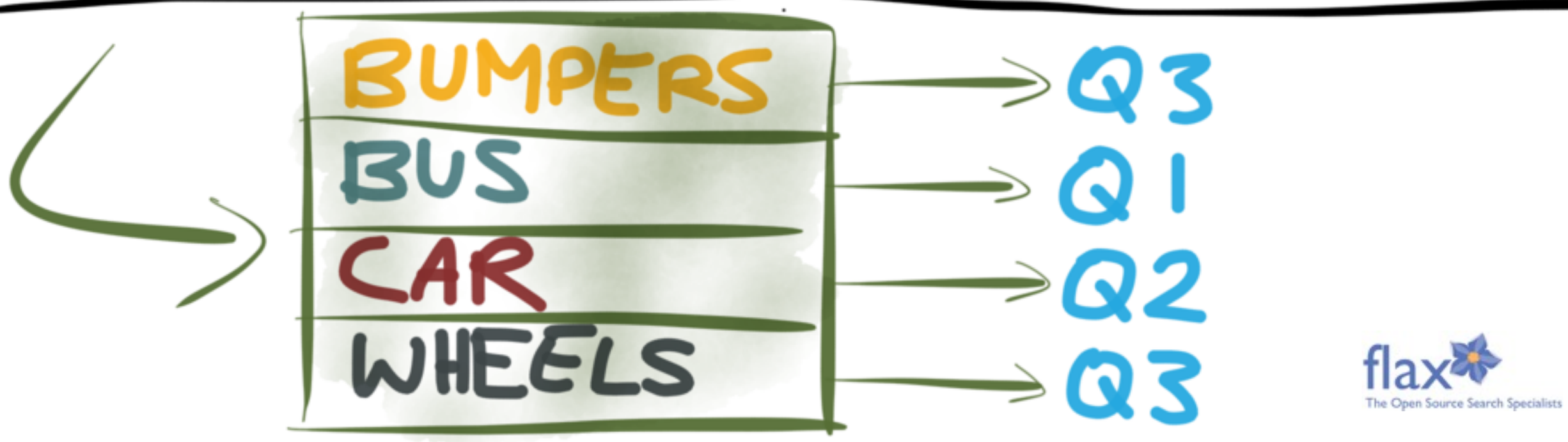
Index of queries

Q1: "WHEELS" NEAR "BUS" → "BUS"

Q2: "WHEELS" NEAR "CAR" → "CAR"

Q3: "WHEELS" OR "BUMPERS"

→ "BUMPERS", "WHEELS"



flax
The Open Source Search Specialists

Document disjunction

"The wheels
on the bus
go round
and round"

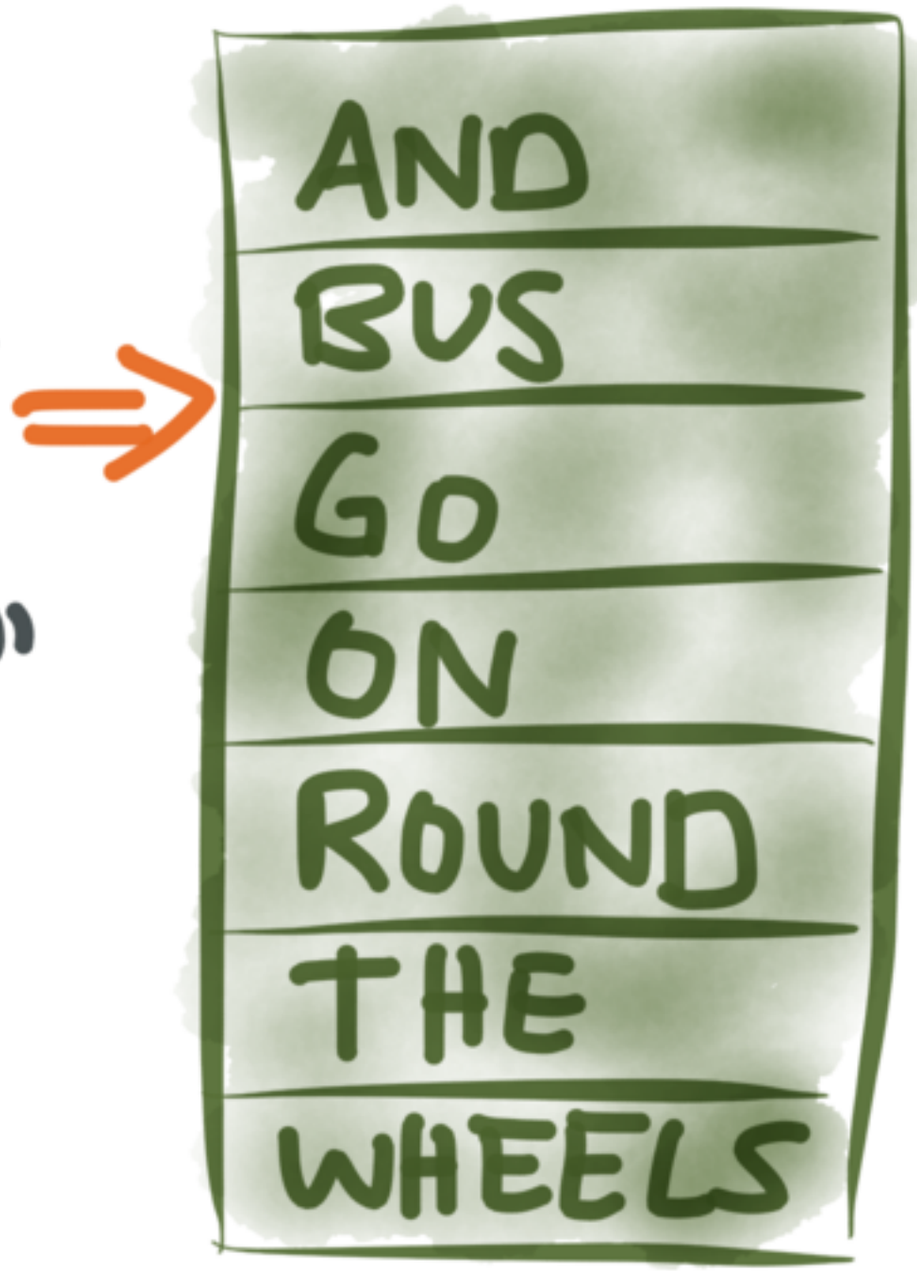
flax
The Open Source Search Specialists

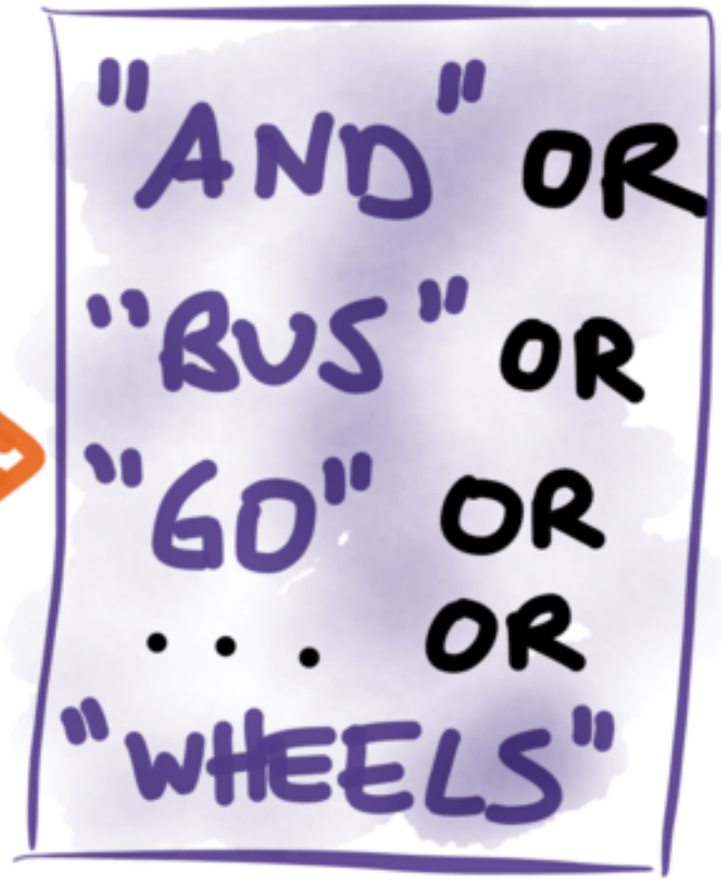
Document disjunction

"The wheels on the bus go round and round"

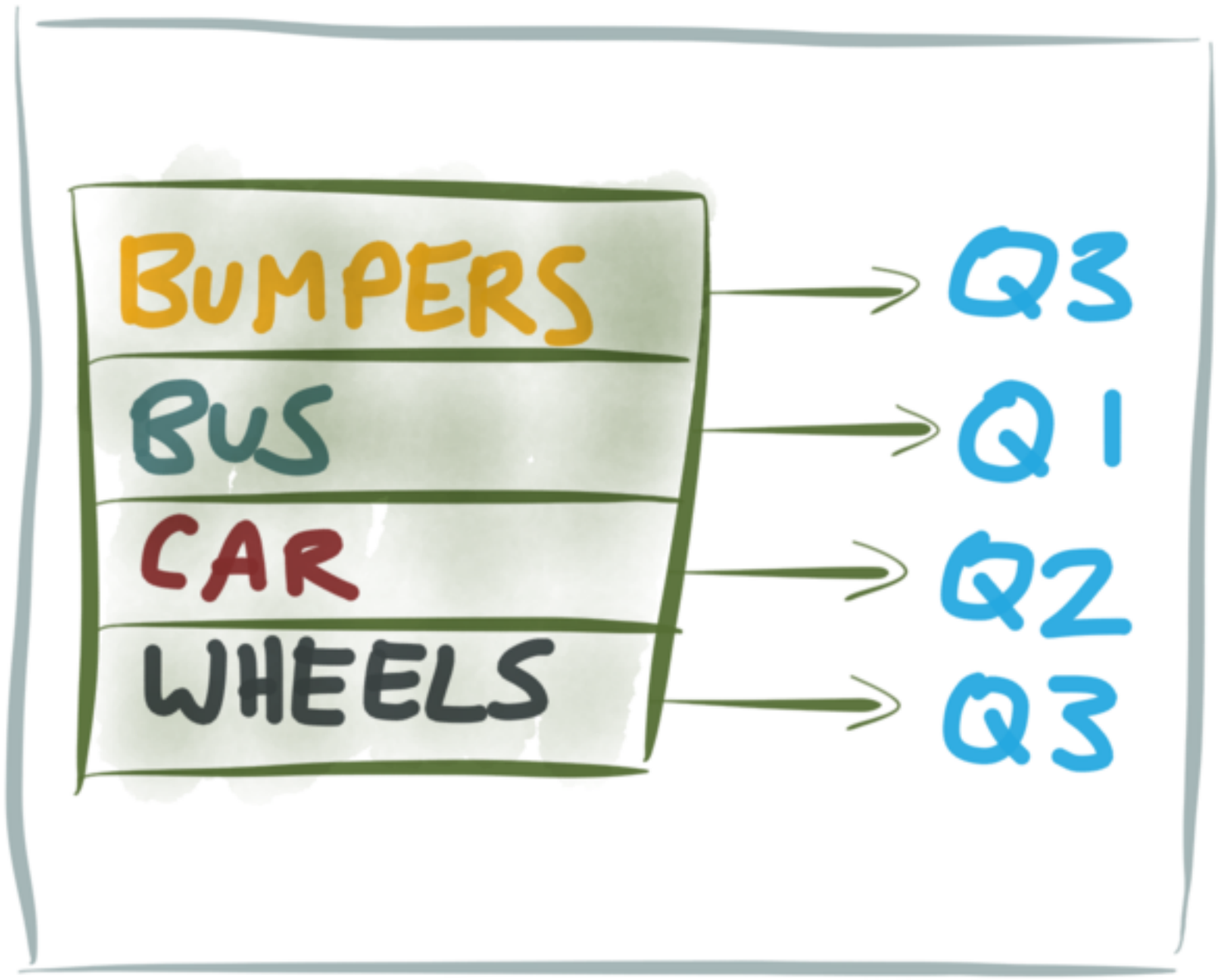
Document disjunction

"The wheels on the bus go round and round"



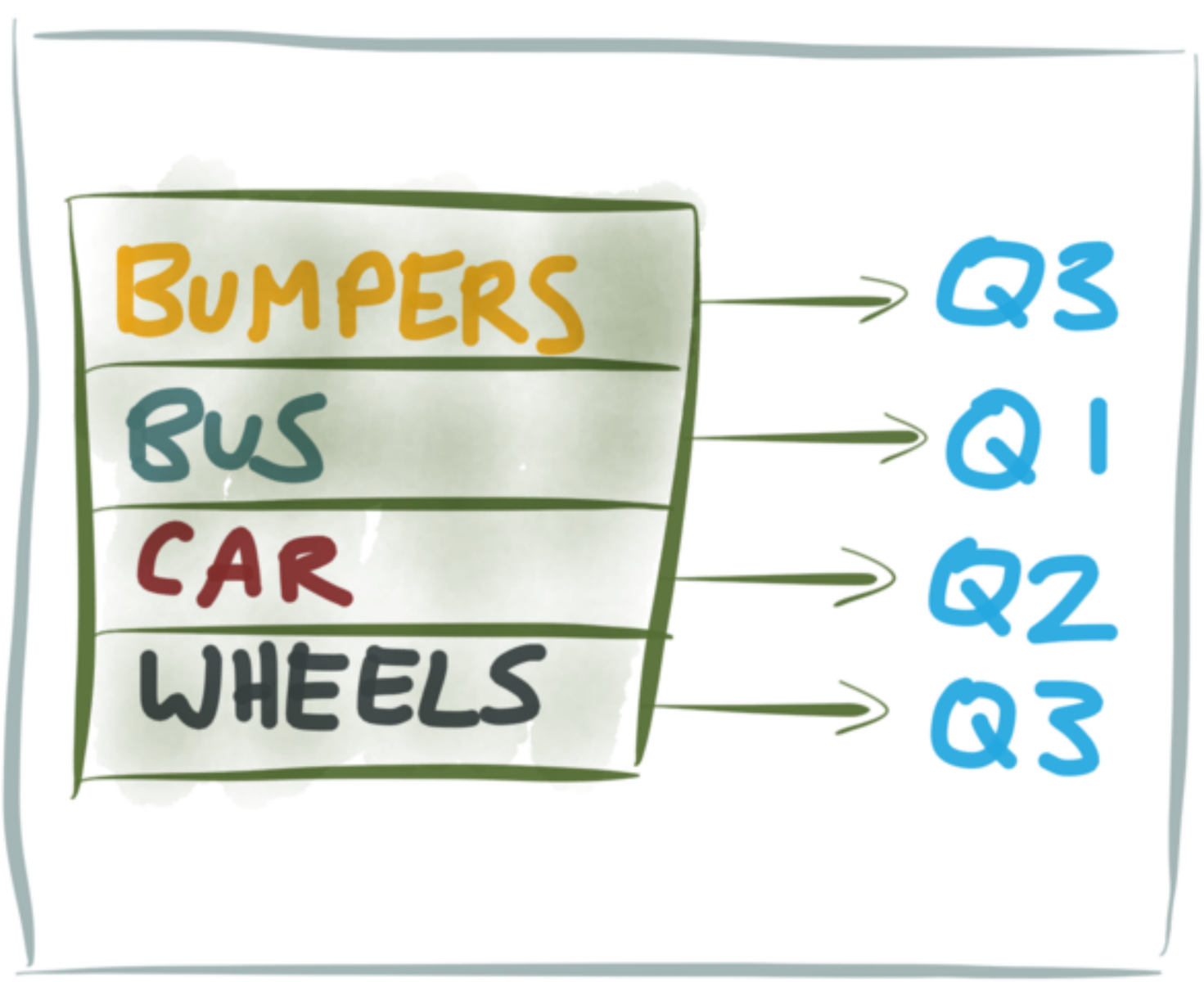


Selecting candidate queries



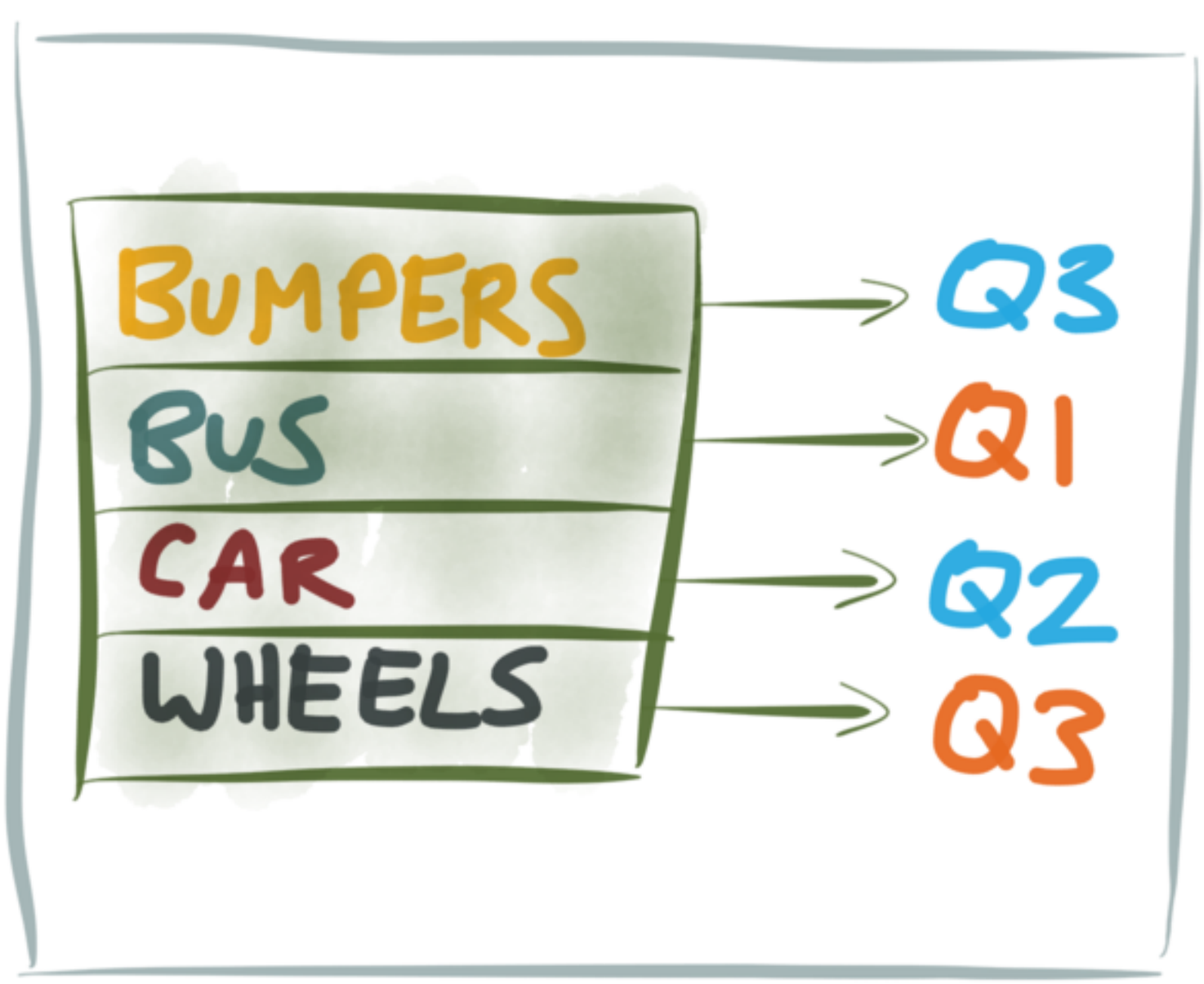
flax
The Open Source Search Specialists

Selecting candidate queries



"AND" OR
"BUS" OR
"GO" OR
...
"WHEELS"

Selecting candidate queries



Q1 and Q3 may match

Q2 doesn't match

Flax Luwak

- Java library for efficiently running queries over document streams
- Builds query indexes and document disjunctions for you
- Can run over single documents for better latency, or document batches for higher throughput
- Much, much faster than just running all queries

flax
The Open Source Search Specialists

Samza - Luwak

- Experimental project to integrate luwak with Kafka topics
- Samza is a stream-processing library that runs within distributed containers (eg Yarn)
- Reads queries from a Kafka topic
- Reads messages from a Kafka topic, and writes results back out to another Kafka topic

flax
The Open Source Search Specialists

Samza - Luwak

- Scales with messages by using Kafka partitions
- Scales with queries by partitioning queries up, and recombining messages using local state

flax
The Open Source Search Specialists

Kafka-Streams

- New stream processing library built directly into Kafka
- In pre-release from Confluent

flax
The Open Source Search Specialists

Questions?

https://github.com/romseygeek/samza-luwak
https://github.com/flaxsearch/luwak

Charlie Hull - charlie@flax.co.uk - @FlaxSearch
Alan Woodward - alan@flax.co.uk - @romseygeek
www.flax.co.uk/blog
+44 (0) 8700 118334

flax
The Open Source Search Specialists