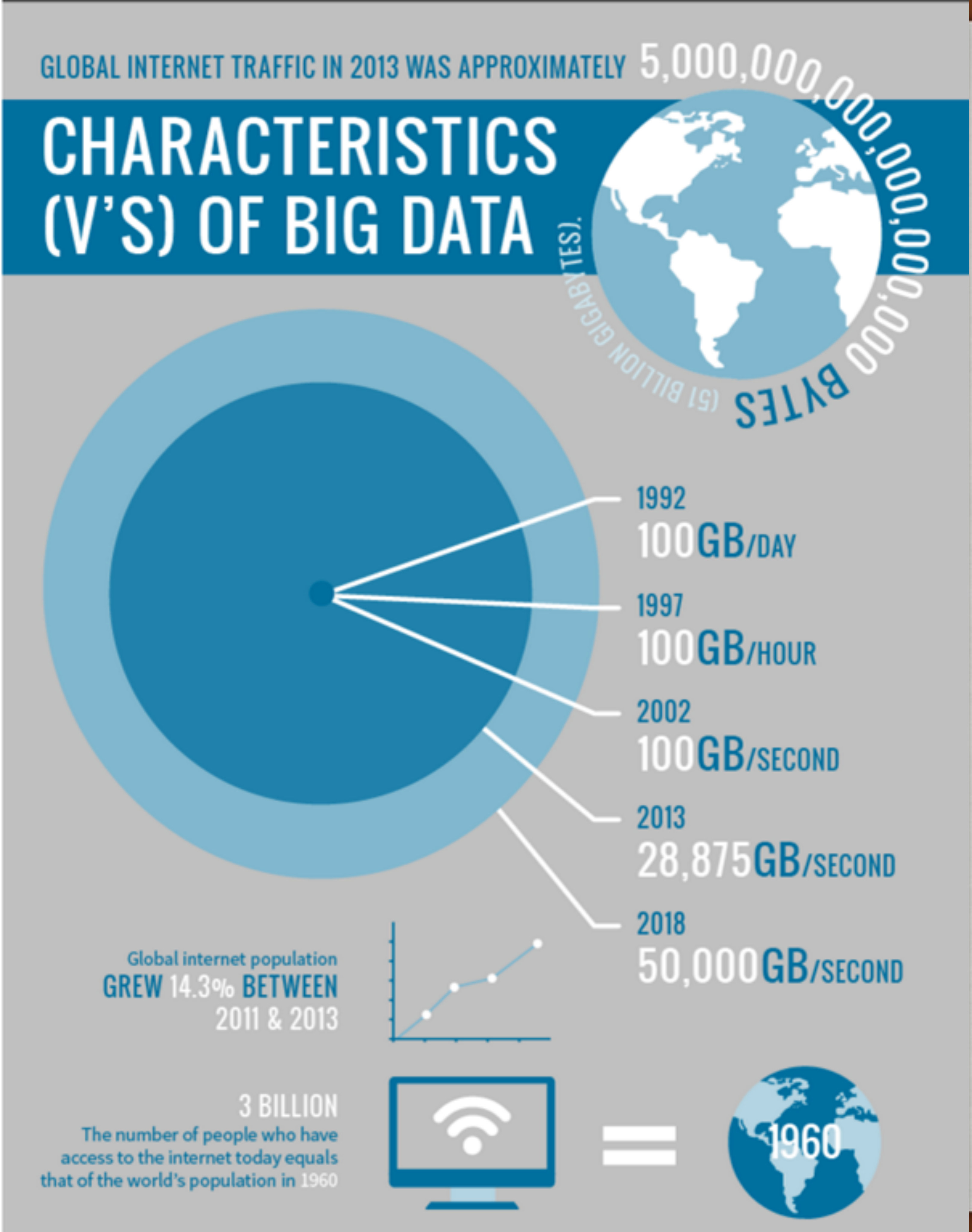
DEEP DIVE INTO SEARCH SYSTEMS

HI!

- Applied Research Engineer - Spam and Relevance team, LinkedIn
- Software Engineer - Search team, LinkedIn
- Undergraduate Thesis - Search team, LinkedIn
- Systems and Network Engineering Intern, LinkedIn
- BITS Pilani - B.E. *(Hons)* Computer Science



GLOBAL INTERNET TRAFFIC IN 2013 WAS APPROXIMATELY 5,000,000,000,000,000,000 BYTES (5 BILLION GIGABYTES).

CHARACTERISTICS (V'S) OF BIG DATA

1992
100GB/DAY

1997
100GB/HOUR

2002
100GB/SECOND

2013
28,875GB/SECOND

2018
50,000GB/SECOND

Global internet population
GREW 14.3% BETWEEN
2011 & 2013

3 BILLION
The number of people who have access to the internet today equals that of the world's population in 1960

Source: vcloudnews

DATA EXPLOSION

- 90% of the world data is created in last two years alone
- The Indexed Web contains at least 4.72 billion pages
- Why I am talking about this?
  - Without retrieval this data has very little use
  - Imagine a world without Search Engines

INFORMATION RETRIEVAL

- Lets define the problem

  Given a query, find material (documents) of an unstructured nature (usually text) from a large collection (corpus)

- Not just web search!
  - E-mail search
  - Search on your laptop
  - Search on Amazon, LinkedIn

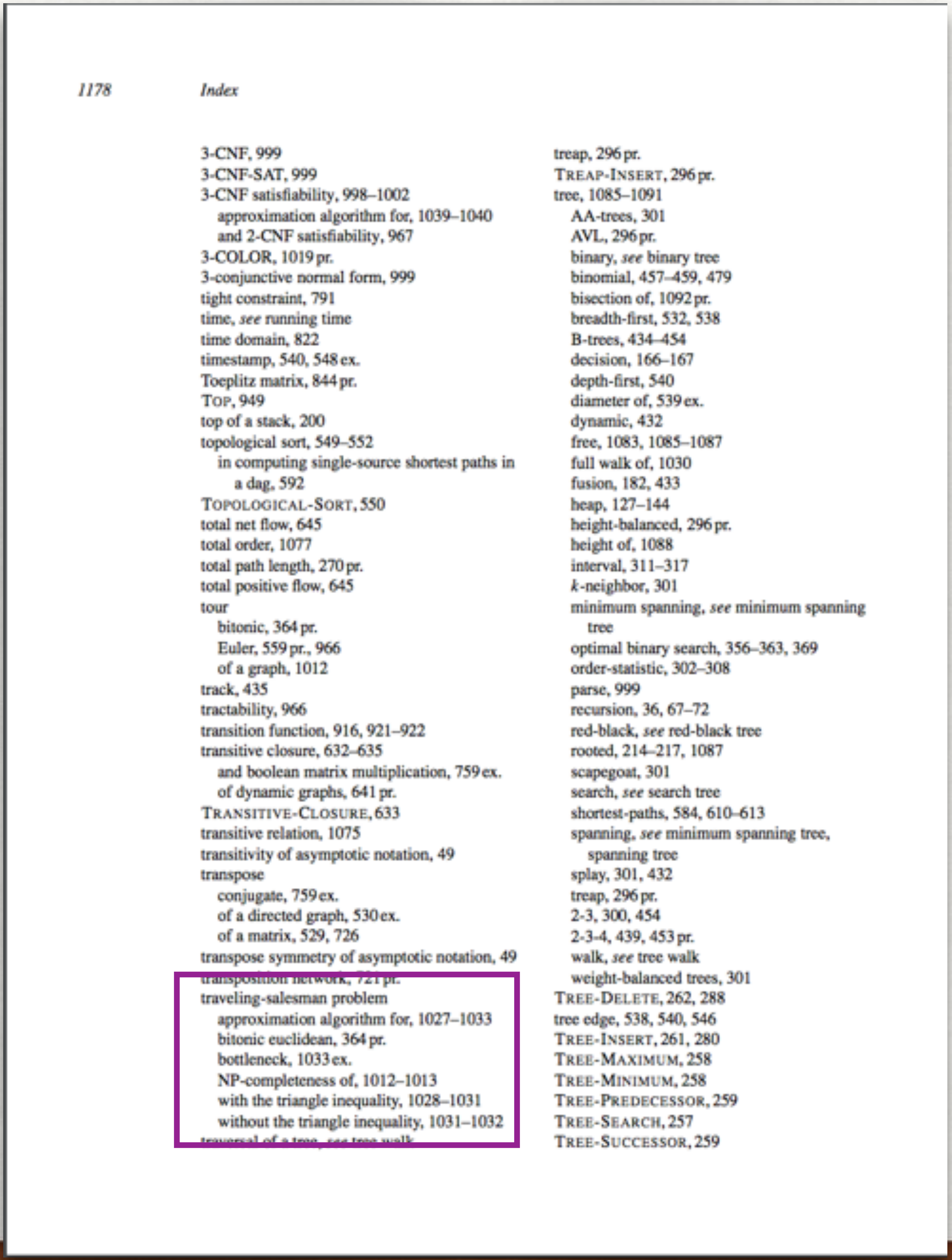
- First step?
  - Collecting Documents!

- What happens after all the documents are collected?
- They are stored in the form of a “Search Index”

SEARCH INDEX

- How should the documents be stored to ensure fast retrieval?
  - Just store all the documents in sequence and go through them
    - Called *grep*
    - More than 4.7 Billion documents on web!
    - 450 M members on LinkedIn
      - 450 M x 1 ms = 125 hours!
  - Any guesses?
    - Hint - How do you search for a term in a book?

SEARCHING FOR TRAVELING SALESMAN?



- Inverted Index!
    - Terms are matched against documents and not vice-versa

INVERTED INDEX

ID	Text
1	Summer is hot. Mangoes are ripe during summer.
2	Summer is warm here
3	Why is summer hot here
4	Later we found why

Term	Document IDS
Mangoes	1
are	1
ripe	1
during	1
summer	1, 2, 3
is	1,2, 3
warm	2
here	2, 3
why	3, 4
hot	1, 3
later	4
we	4
found	4

QUERY

- A query is interpreted as a Boolean Query
  - Examples on LinkedIn Recruiter search -
    - html AND css
    - java OR python
      - implicit - synonyms
    - manager NOT software

- Boolean Retrieval
  - OR / AND
    - term1 - 1,2,3,5,6 | term2 - 1,3,4,5,7,8
    - **MERGING**
      - term1 OR term2 - 1,2,3,4,5,6,7,8
      - term1 AND term2 - 1,3,5
    - Time complexity
      - O[m+n]

Dictionary

Posting List

Term	Freq		Document IDS
Mangoes	1	→	1
are	1	→	1
ripe	1	→	1
during	1	→	1
summer	3	→	1, 2, 3
is	3	→	1,2, 3
warm	2	→	2
here	2	→	2, 3
why	2	→	3, 4
hot	2	→	1, 3
later	1	→	4
we	1	→	4
found	1	→	4

- What Data Structure is used?
  - Posting list -
    - Variable length array vs Linked List
      - Frequency of updates
      - Space and time overhead
  - Accessing terms -
    - Sorted array
    - Hash table
    - B+ trees

THREE PILLARS OF SEARCH

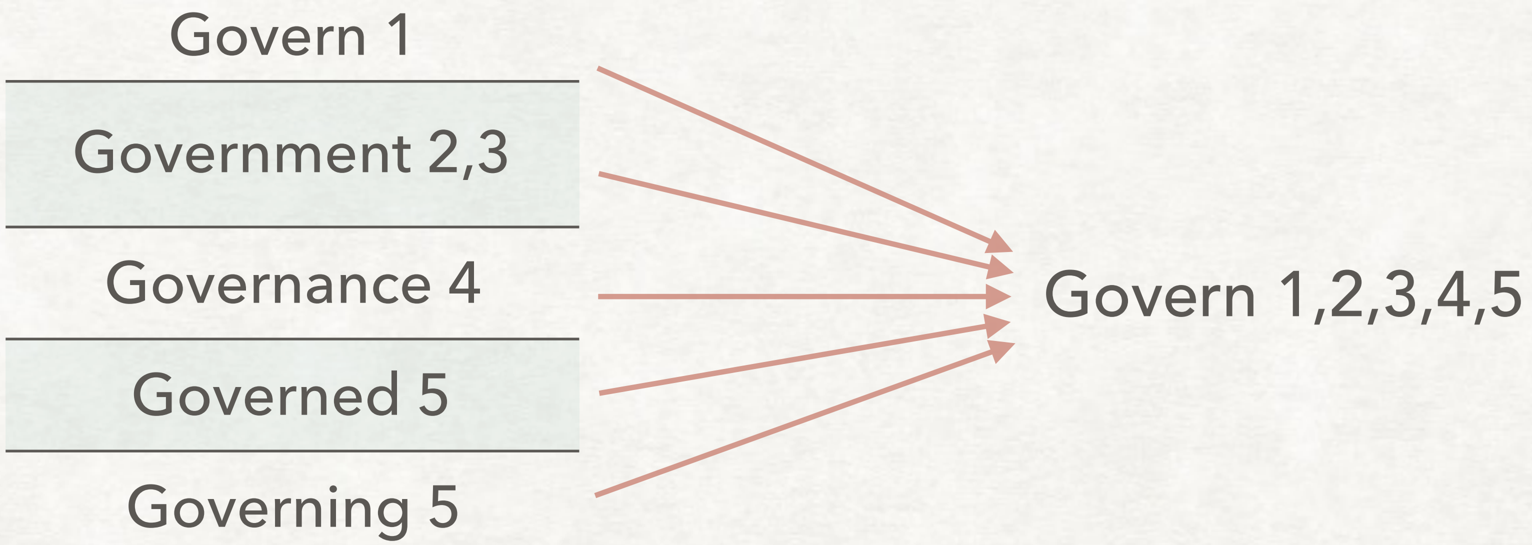
- Any search infrastructure will expose APIs for these three functionalities -
  - Building Index
    - Preprocessing and tokenization
  - Retrieval
  - Ranking

PREPROCESSING

- Document Preprocessing (part of index build)
  - Tokenization for index creation
    - Apply filters like - lowercase, removing punctuations, stopwords
    - Stemming
  - Calculating TF and IDF values for TF-IDF scoring
- Query Preprocessing (part of retrieval)
  - Query tagging
  - Tokenization (should be similar to document preprocessing)
  - Query expansion
    - Synonym expansion

PREPROCESSING

- Stemming



Govern 1

Government 2,3

Governance 4

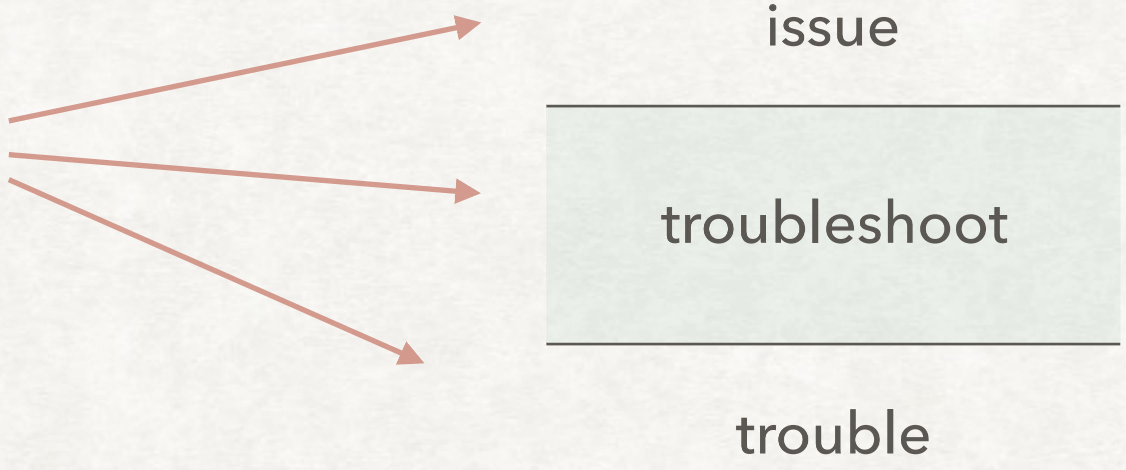
Governed 5

Governing 5

Govern 1,2,3,4,5

- Synonym Expansion

"slow router problem"



issue

troubleshoot

trouble

RETRIEVAL

Term	Document IDS
Mangoes	1
are	1
ripe	1
during	1
summer	1, 2, 3
is	1,2, 3
warm	2
here	2, 3
why	3, 4
hot	1, 3
later	4
we	4
found	4

Query - hot summer

Sample rewritten queries -

- hot AND summer
- (hot OR warm) AND summer
- hot OR summer

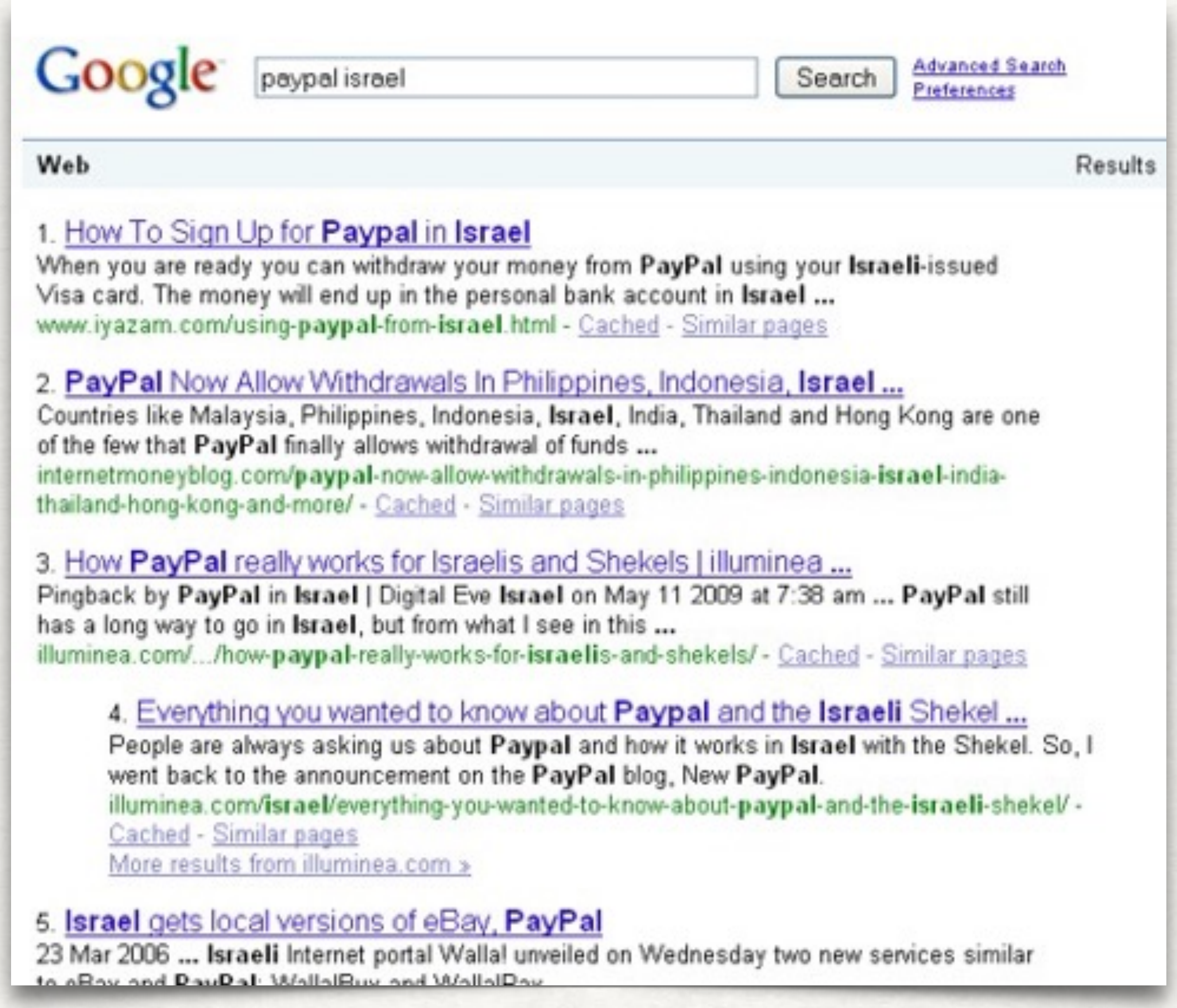
STATIC RANK & EARLY TERMINATION

- A global score of the document
- Each document has one SR but multiple documents can have same SR
- Could be anything from number of connections / followers; length of the documents; Social signals; Page Rank
- Used in early termination
  - Number of documents to score
- Posting Lists are sorted on the basis of Static Rank

RANKING / SCORING

- Retrieved documents are scored based on the combination of -
  - Query dependent features
    - Term Frequency,
    - Min window match,
    - Exact match in title, etc.
  - Query independent features
    - Length of the document,
    - Social Signals,
    - Number of followers of the author, etc.

- Once we have the ranked list of document IDs, we obviously won’t show them to the users!



paypal israel

1. How To Sign Up for **Paypal** in **Israel**
When you are ready you can withdraw your money from **PayPal** using your **Israeli**-issued Visa card. The money will end up in the personal bank account in **Israel** ...
www.iyazam.com/using-**paypal**-from-**israel**.html - Cached - Similar pages

2. **PayPal** Now Allow Withdrawals In Philippines, Indonesia, **Israel** ...
Countries like Malaysia, Philippines, Indonesia, **Israel**, India, Thailand and Hong Kong are one of the few that **PayPal** finally allows withdrawal of funds ...
internetmoneyblog.com/**paypal**-now-allow-withdrawals-in-philippines-indonesia-**israel**-india-thailand-hong-kong-and-more/ - Cached - Similar pages

3. How **PayPal** really works for Israelis and Shekels | illuminea ...
Pingback by **PayPal** in **Israel** | Digital Eve **Israel** on May 11 2009 at 7:38 am ... **PayPal** still has a long way to go in **Israel**, but from what I see in this ...
illuminea.com/.../how-**paypal**-really-works-for-**israeli**s-and-shekels/ - Cached - Similar pages

4. Everything you wanted to know about **Paypal** and the **Israeli** Shekel ...
People are always asking us about **Paypal** and how it works in **Israel** with the Shekel. So, I went back to the announcement on the **PayPal** blog, New **PayPal**.
illuminea.com/**israel**/everything-you-wanted-to-know-about-**paypal**-and-the-**israeli**-shekel/ - Cached - Similar pages
More results from illuminea.com »

5. **Israel** gets local versions of eBay, **PayPal**
23 Mar 2006 ... **Israeli** Internet portal Walla! unveiled on Wednesday two new services similar to eBay and **PayPal**: WallaBuy and WallaPay

- Title, snippet, URL

- How are they stored?

	title	snippet
1	Summer	Hot summer
2	warm	warm in summer
3	hot	summer
4	later	later why

Term	Document IDS
Mangoes	1
are	1
ripe	1
during	1
summer	1, 2, 3
is	1,2, 3
warm	2
here	2, 3
why	3, 4
hot	1, 3
later	4
we	4
found	4

FORWARD INDEX

- Mapping of document ID to forward fields (title / snippet)

	title	snippet
1	Summer	Hot summer
2	warm	warm in summer
3	hot	summer
4	later	later why

Forward Index!

Term	Document IDS
Mangoes	1
are	1
ripe	1
during	1
summer	1, 2, 3
hot	1
is	1,2, 3
warm	2
here	2, 3
why	3, 4
hot	1, 3
later	4
we	4
found	4

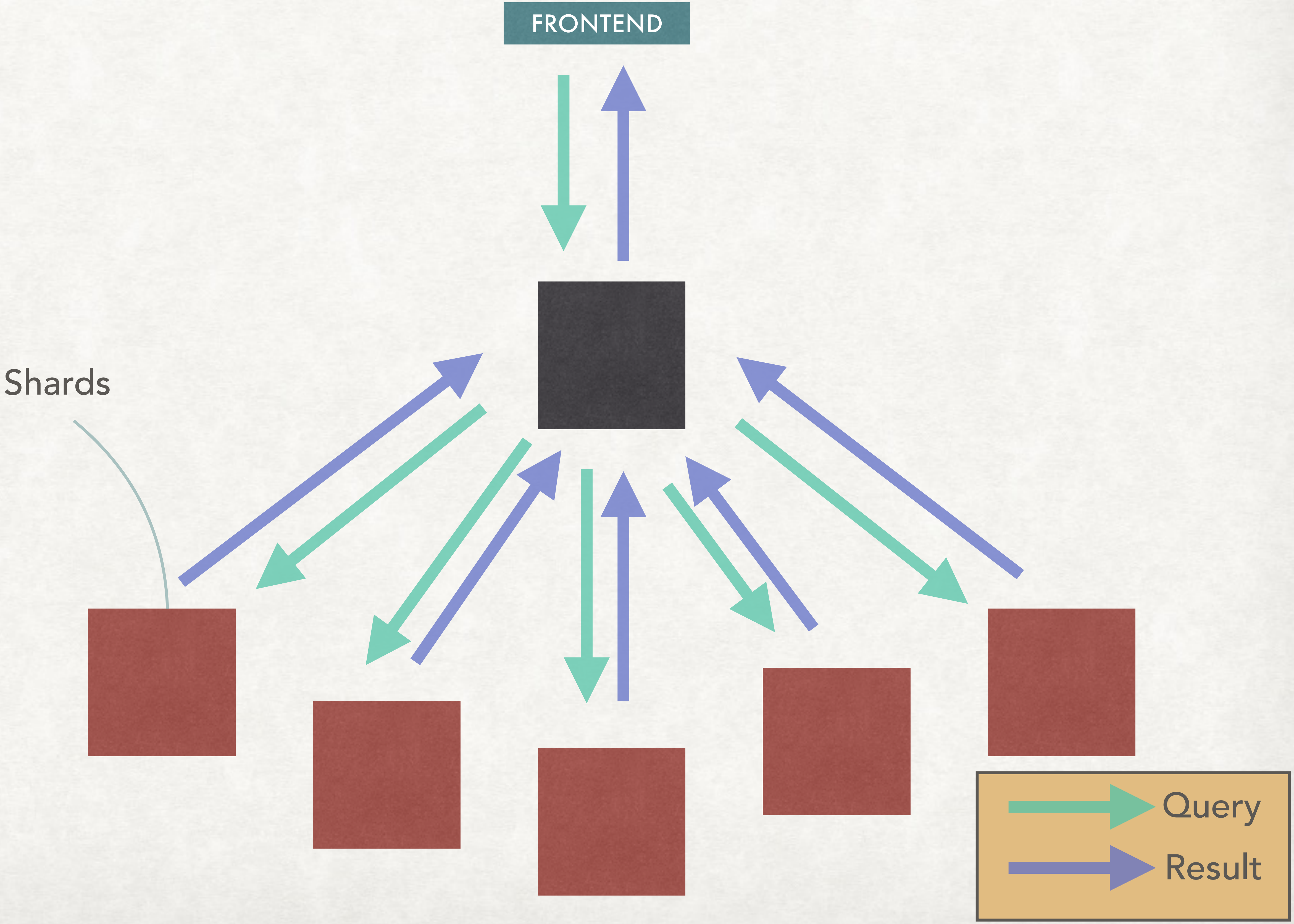
BASIC SEARCH CONCEPTS

- Document
- Query
- Boolean Retrieval model
- Inverted Index
- Forward Index
- Retrieval
- Ranking

SCALING

- All the data won’t fit in one computer
- Even if it does, we should leverage parallelization
- How to break the index into multiple machines?
  - Shards!
  - Let’s consider inverted index first

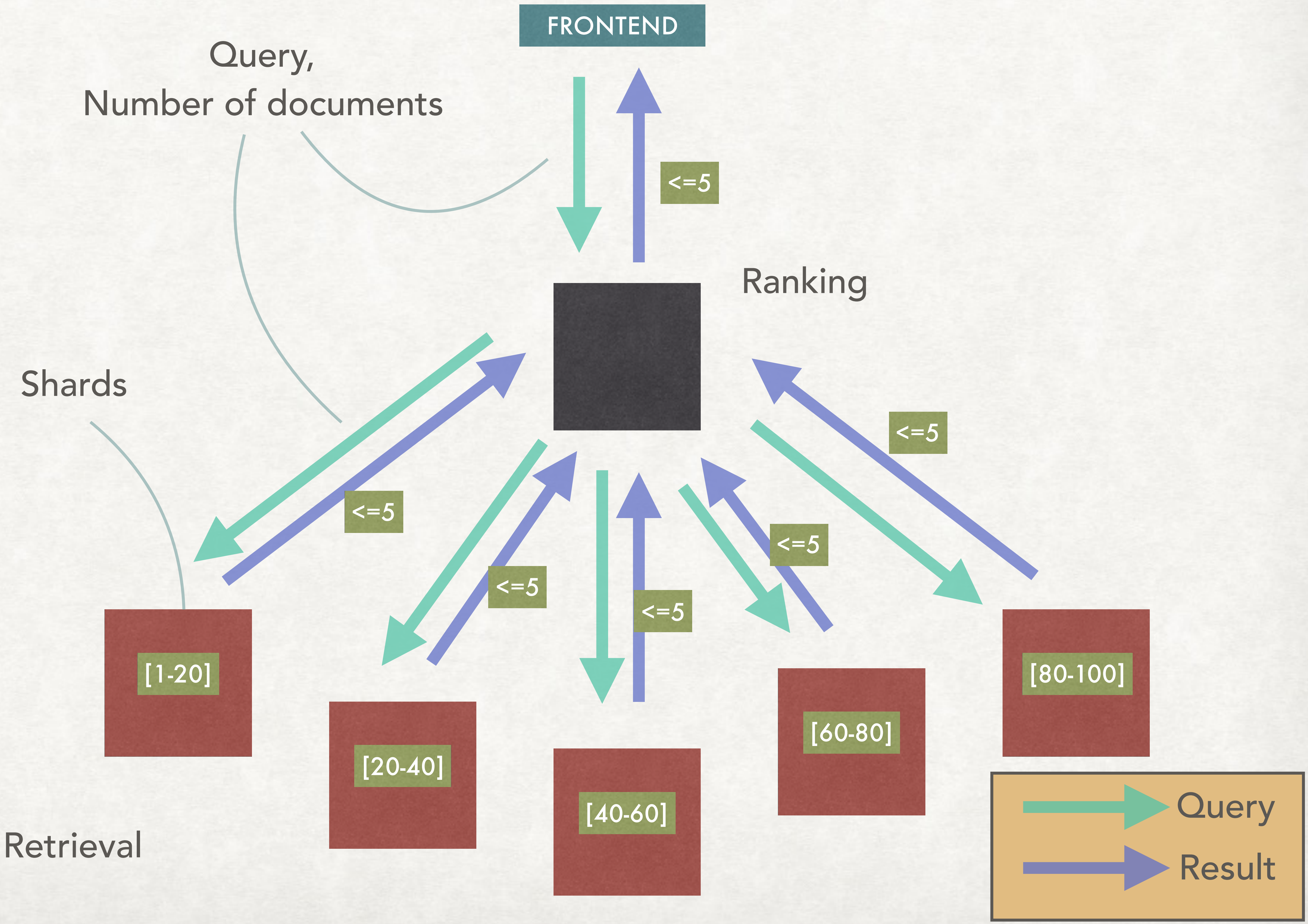
Simple Master Slave Configuration



SCALING

- Document vs Term based partitioning?
  - Failover of a machine
  - Latency
  - Leveraging Parallelization
  - Maintaining forward index
  - All the above favor document based partitioning!

Simple Master Slave Configuration



FRONTEND

Query,
Number of documents

Ranking

Shards

Retrieval

Query - hot OR summer
Number of documents - 2

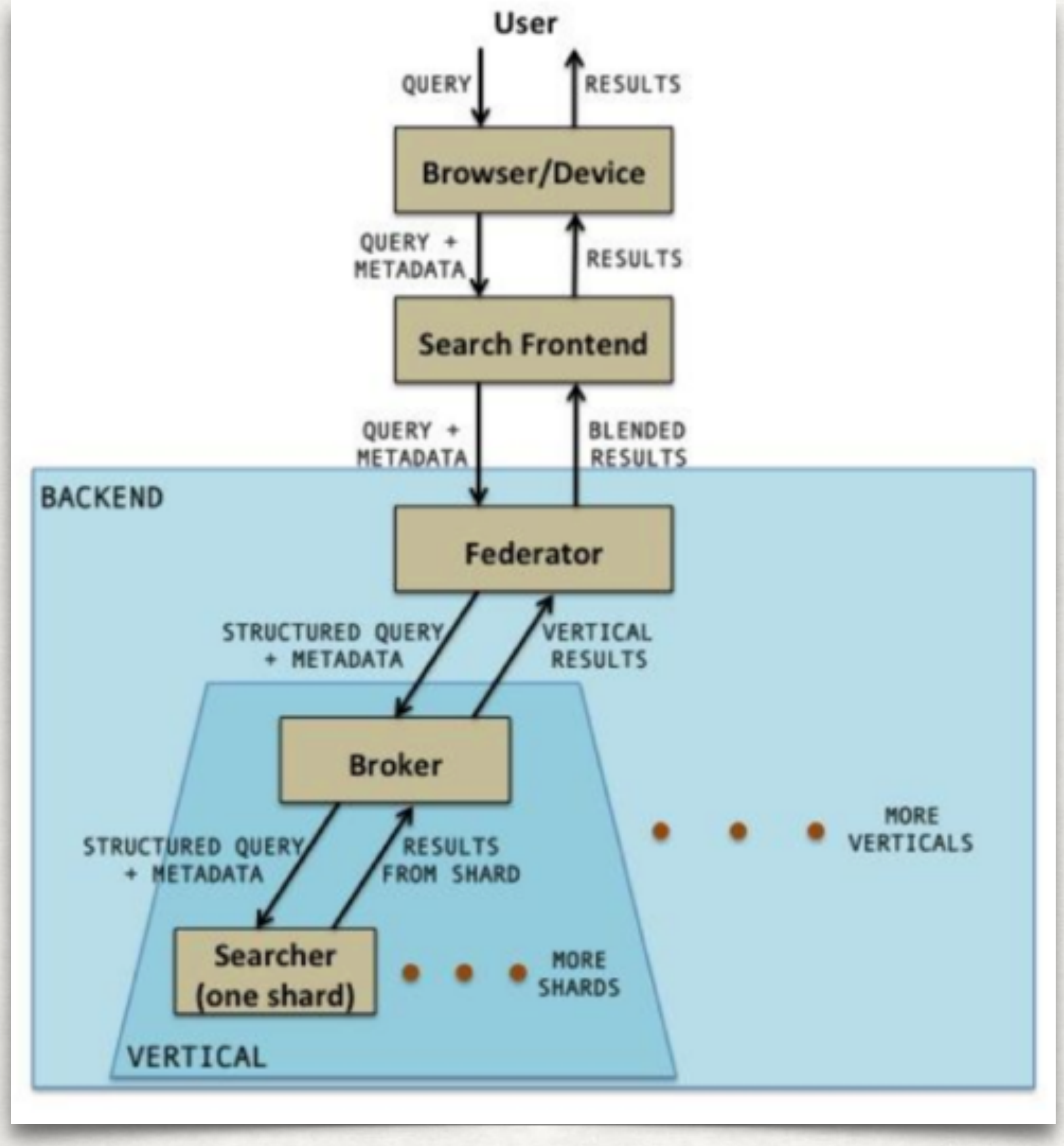
Term	Document IDS
Mangoes	1
are	1
ripe	1
during	1
summer	**1, 2**
hot	1
is	1,2
warm	2
here	2
hot	1

1, 2

Term	Document IDS
summer	3
is	3
here	3
why	3, 4
hot	3
later	4
we	4
found	4

3

LINKEDIN SEARCH STACK (GALENE)



- Federator and Broker
  - Rewrites the Query
  - Fans out
  - Merges the results
  - Requires num of docs to return
- Searcher
  - Operates on single shard
  - Takes rewritten query and retrieves the documents
  - Scores the documents - using query, input metadata, match info
  - Requires num of docs to score

QUESTIONS?

Mansi Gupta

mgupta1410@gmail.com