GALENE

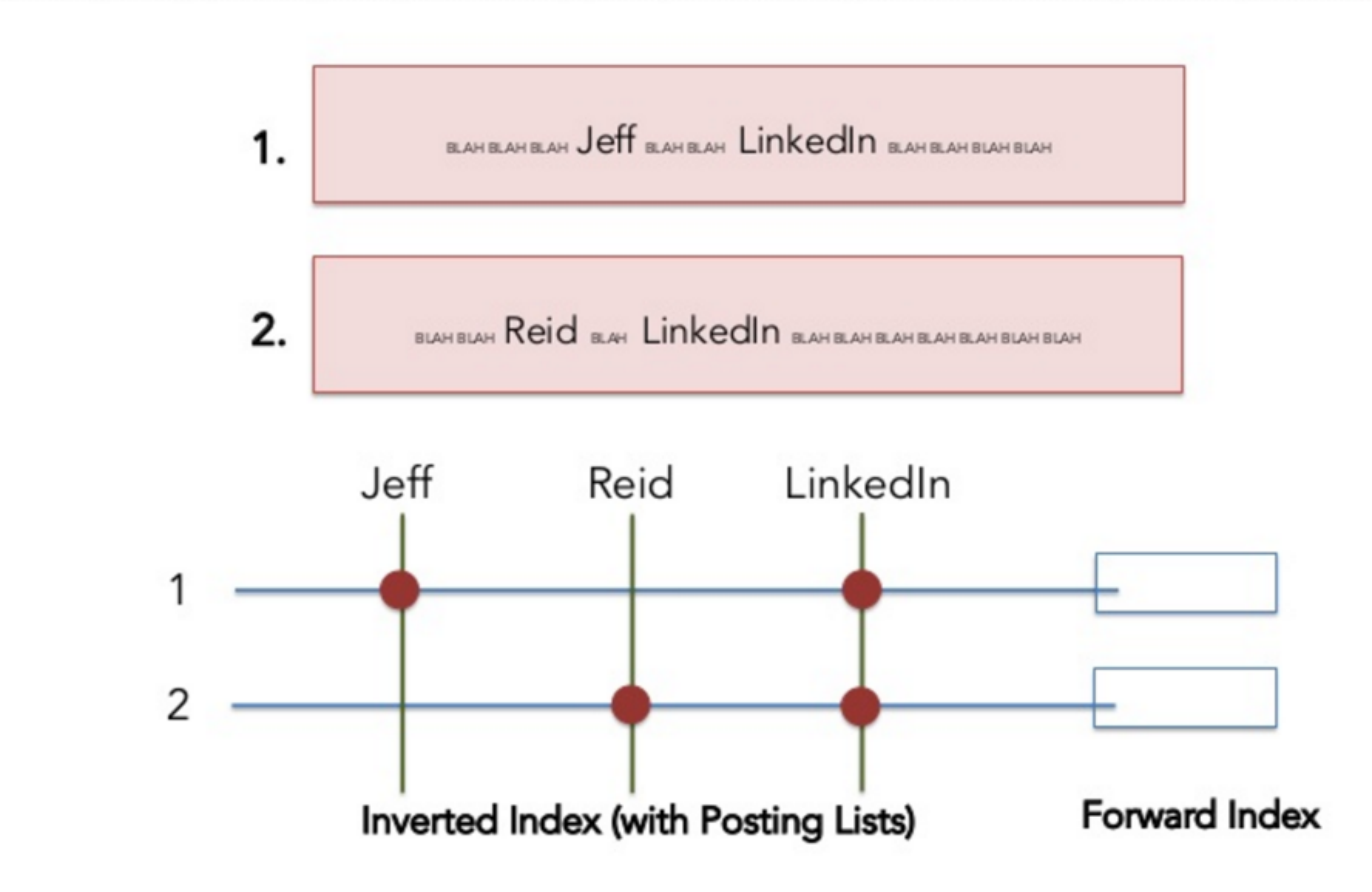
LinkedIn’s Search Architecture

OVERVIEW

- Search at LinkedIn
  - Galene Infrastructure
  - Search Relevance

BASIC SEARCH CONCEPTS

- Document
- Query
- Boolean Retrieval model
- Search Index
    - Inverted Index
      Term -> Doc mapping
    - Forward Index
      Doc -> Metadata mapping
- Posting List
- Relevance



1. BLAH BLAH BLAH Jeff BLAH BLAH LinkedIn BLAH BLAH BLAH BLAH

2. BLAH BLAH Reid BLAH LinkedIn BLAH BLAH BLAH BLAH BLAH BLAH BLAH

Inverted Index (with Posting Lists)

Forward Index

LUCENE

- Open source API that supports -
    - Adding / deleting new documents to index
    - Query construction
    - Retrieving the documents
    - Score the retrieved documents

LEGACY SEARCH

- Lucene based
- Challenges faced -
  - Too many open sourced independent components like *Sensei* for cluster management, *Zoie* for live updates, *Bobo, Cleo, Krati* etc
  - Rebuilding the entire index was difficult given the incremental approach
  - Live updates at entity level granularity
  - Inflexible scoring
  - Many requirements such as offline relevance, query rewriting, reranking, and blending not possible

GALENE

- Lucene is retained as an indexing layer
  We use some elements of Lucene to assist in building indices, build query and retrieving documents

- Major steps - Creating Index; Retrieval; Scoring

- Important Galene features -

  - Offline Index Creation

  - Static Rank and Early Termination

  - Live updates at fine granularity

  - Flexible Relevance Framework

  - Faceting

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 etc
- Used in early termination
  - numToScore
- Posting List sorted on the basis of Static Rank

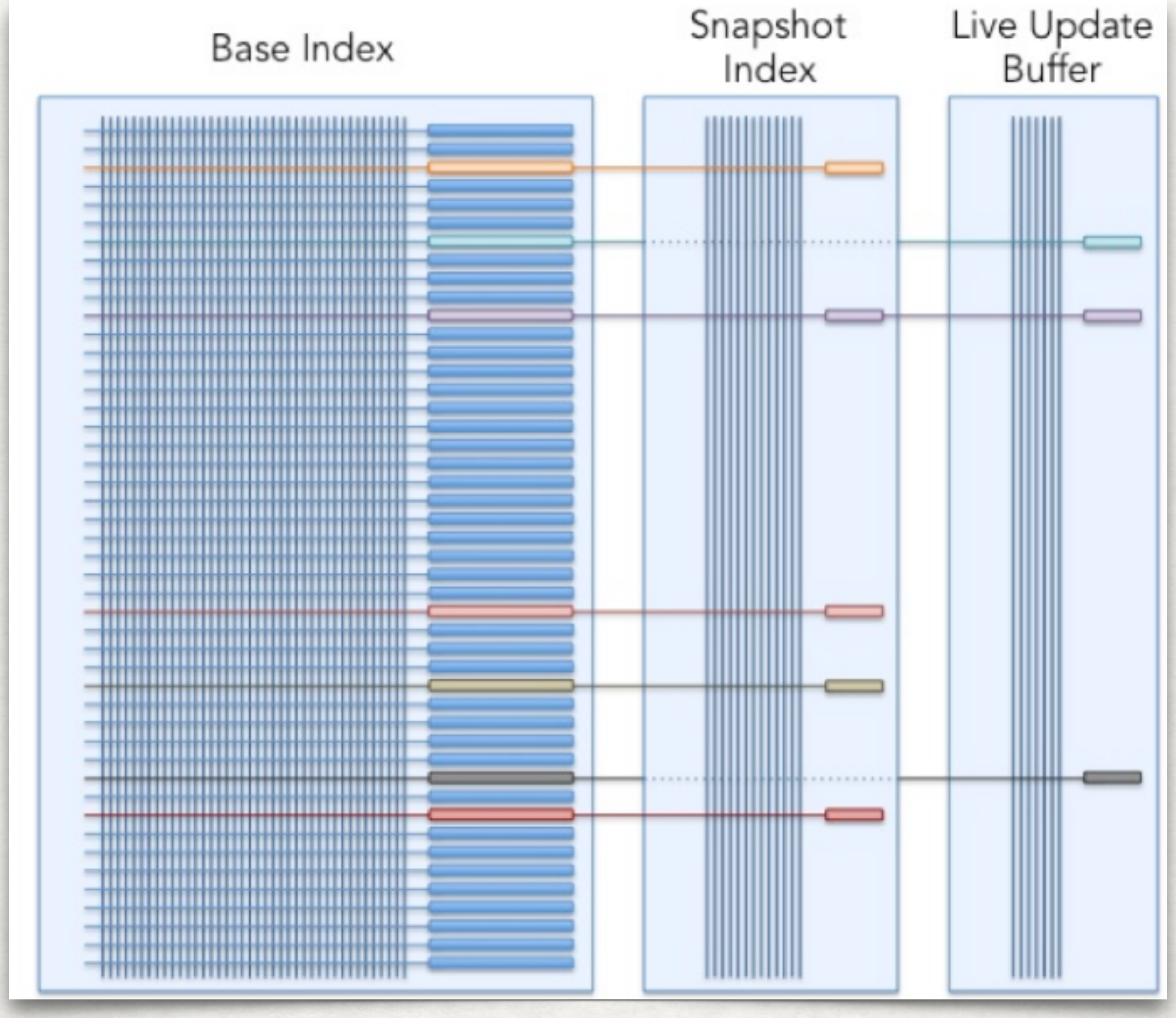
GALENE INDEXING SCHEME

- Base Index
  - Generated periodically offline - every week
  - Lucene Index
  - Contains complex features
- Live update buffer
  - Inverted Index of our own format
  - In-memory
  - Contains incremental updates

GALENE INDEXING SCHEME

- Snapshot Index
  - On disk snapshot of live index
  - Live index is folded regularly - every few hours

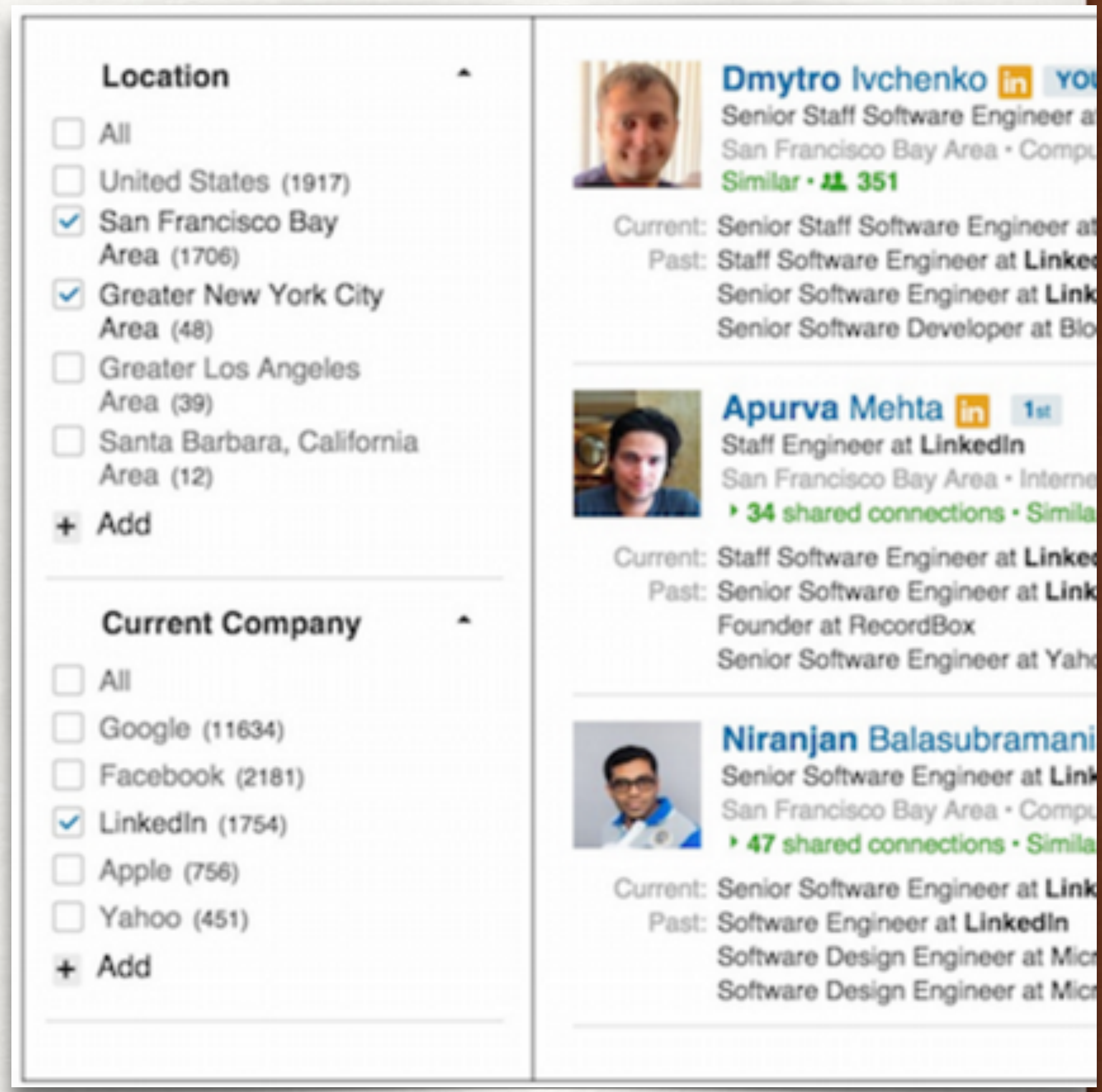
GALENE INDEXING SCHEME



A LOOK AT THE SCHEMA

FACETING

- Discoverable Facets
- Non-discoverable facets
- Early termination and Faceting
  - Discovery and counting
  - Challenge is to do two things at the same time -
    - Approximate facet counts for large values
    - Guarantee exact values for low counts
    - https://engineering.linkedin.com/faceting/many-facets-faceted-search



Location

□ All
□ United States (1917)
☑ San Francisco Bay Area (1706)
☑ Greater New York City Area (48)
□ Greater Los Angeles Area (39)
□ Santa Barbara, California Area (12)
+ Add

Current Company

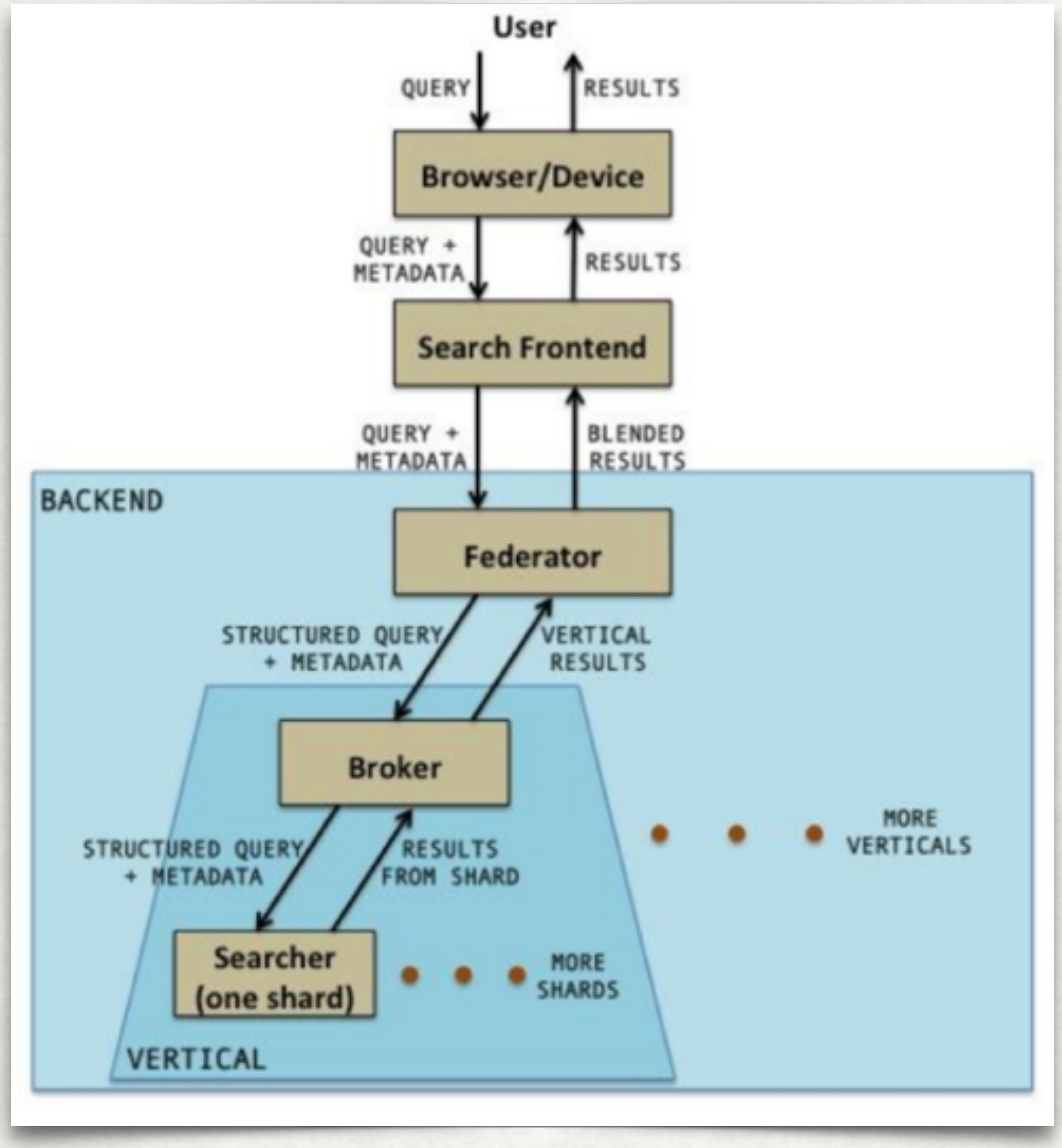
□ All
□ Google (11634)
□ Facebook (2181)
☑ LinkedIn (1754)
□ Apple (756)
□ Yahoo (451)
+ Add

Dmytro Ivchenko in YOU
Senior Staff Software Engineer a
San Francisco Bay Area • Compu
Similar • 351
Current: Senior Staff Software Engineer at
Past: Staff Software Engineer at Linke
Senior Software Engineer at Link
Senior Software Developer at Blo

Apurva Mehta in 1st
Staff Engineer at LinkedIn
San Francisco Bay Area • Interne
34 shared connections • Simila
Current: Staff Software Engineer at Linke
Past: Senior Software Engineer at Link
Founder at RecordBox
Senior Software Engineer at Yah

Niranjan Balasubramani
Senior Software Engineer at Link
San Francisco Bay Area • Compu
47 shared connections • Simila
Current: Senior Software Engineer at Link
Past: Software Engineer at LinkedIn
Software Design Engineer at Mic
Software Design Engineer at Mic

GALENE SEARCH STACK



- Federator and Broker
  - Rewrites the Query
  - Fans out
  - Combines the results
  - numToReturn
  - Plugins - Rewriter and HitMerger

- Searcher
  - Operates on single shard
  - Takes rewritten query and retrieves the documents
  - Scores the documents - using query, input metadata, match info
  - numToScore
  - Plugins - Scorer

SEARCH AS A SERVICE (SEAS)

- Generally one vertical for one type of search / one index
- Searcher, Broker, IDS, Indexer, Live Updater
- IDS deploys the corresponding Base Index shard on all the searchers
- Live updaters receive live updates and generate Kafka event which Indexers and Searchers listen to; Online Transformer
- Indexer generates snapshot index and ships it to the corresponding searcher through IDS

REWRITTEN QUERY LANGUAGE

AIMED TOWARDS MAKING EARLY TERMINATION SUCCESSFUL

- Term Query
  - title:mansi
- Phrase Query
  - "title:mansi title:gupta" [2]
- Boolean Query
  (Required clause, ‘+’; Optional clause, ‘?’; Excluded clause, ‘-')
  - +title:term1 +description:term2
  - +schoolName:stanford -schoolType:primary
  - ?schoolName:stanford ?alternateName:stanford

REWRITTEN QUERY LANGUAGE

AIMED TOWARDS MAKING EARLY TERMINATION SUCCESSFUL

- WOR Query (Weak-OR)
  - WOR title:ibm%5 title:technology%2 title:services%2 [6]
- FLEX Query
  - Diversification
    - ?companyName:ibm [2] ?companyName:oracle [2]
    - ?authorName:jeff [100] ?transcript:jeff [50]
  - Optionality
    - numToScore - 5,
      +companyName:ibm ?companyName:oracle [1]

QUESTIONS?