SocialTrove: A Self-summarizing Storage Service for Social Sensing

Tanvir Amin, Shen Li, Muntasir Rahman, Panindra Seetharamu, Shiguang Wang, Tarek Abdelzaher, Indranil Gupta
University of Illinois at Urbana Champaign

Mudhakar Srivatsa, Raghu Ganti
IBM Research

Reaz Ahmed
University of Waterloo

Hieu Le
TCG Corp

ICAC 2015

Social Sensing

Egypt Unrest

Fukushima Disaster

Sandy Gas Outage

Crimea Annexation

Syria Chemical Attack

Twitter

Facebook

Google+

Instagram

Flickr

Credibility Estimation

Anomaly Detection

Timeline Reconstruction

…

60 Seconds of Social Media*

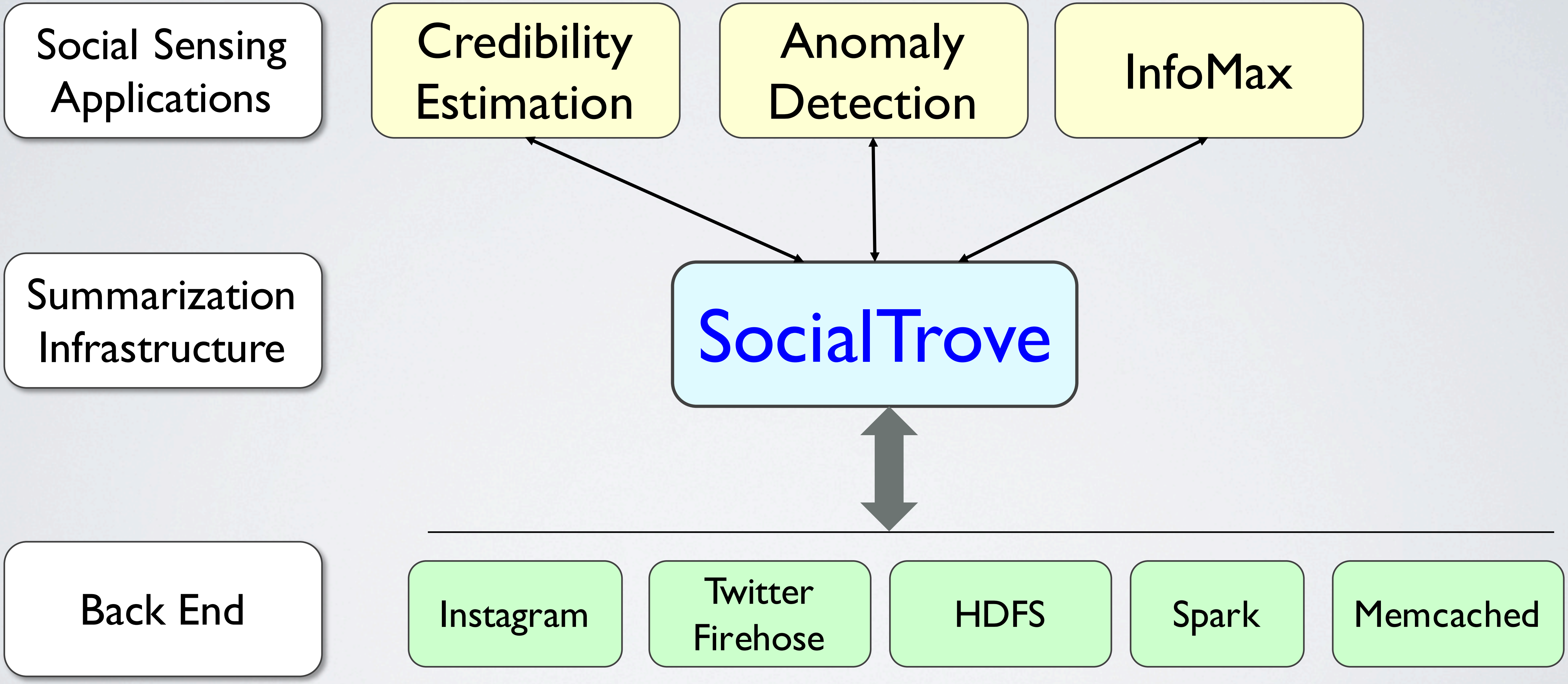
- 2.66 Million Google searches
- 433 Thousand Tweets
- 67 Thousand Instagram photo uploads
- 293 Thousand Facebook Status updates
- 277 Thousand Snaps over SnapChat
- 20 Million Photo views on Flickr
- 11 Thousand LinkedIn searches
- 17 Thousand Walmart transactions
- 20 Thousand Tumblr new photos
- 139 Million emails sent

* http://blog.qmee.com/qmee-online-in-60-seconds/
* http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/

Observations and Goals

- Social Sensing Applications
  - User facing – real-time updates
  - Many users – many queries
- Social Sensing Data Streams
  - Real-time *information overload*
  - Large amounts of *redundant data*
- Obtain a *representative sampling*
  - in a *content agnostic* manner
  - based on *application specified distance metric*
  - at a *configurable granularity*
  - with high query *throughput*
  - and low access *latency*

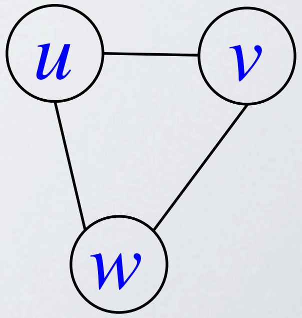
Real-time Social Sensing Stack



Customization Interface

- SocialTrove agnostic to stream content
- Application provides two callbacks
- `Vectorize(u)`
  - Convert stream content u to a feature vector
- `Distance(u.FeatureVector,`
  `v.FeatureVector)`
  - Compute distance between two feature vectors
  - Should be a metric, obey Triangle Inequality

$$\text{distance}(u,v) + \text{distance}(v,w) \geq \text{distance}(u,w)$$



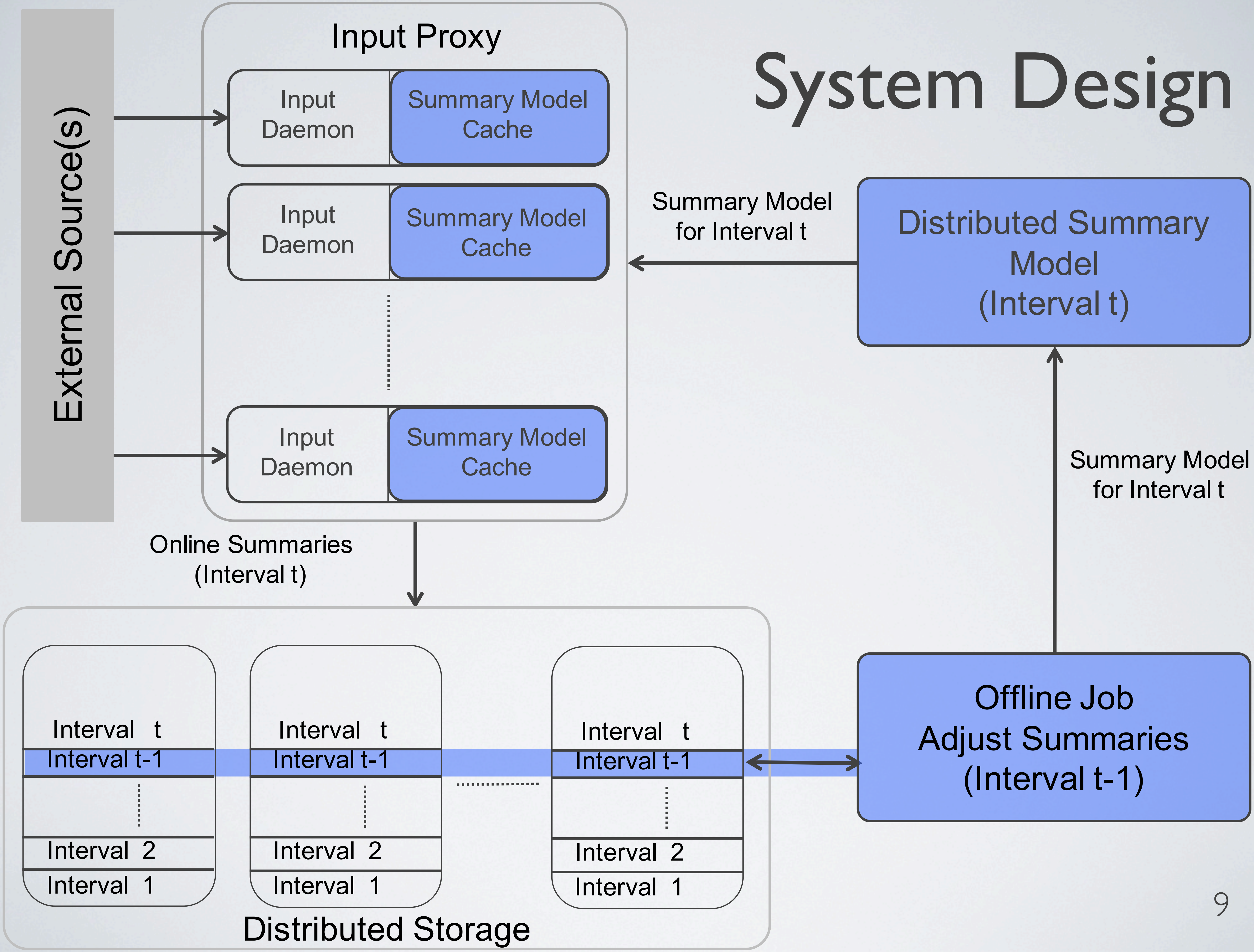
Summary Query Interface

- Transform application specific query.
  - Degree of granularity for summary retrieval.
- Transform SocialTrove query response to application specific results.
- Example:
  - Rank the matching summaries and retrieve top 10.
  - Retrieve matching summaries as a list.
  - Retrieve matching objects and additional metadata.
    - Source ID, Source relations etc.
  - Retrieve summary clusters and member counts.

SocialTrove Runtime

- Maintains a ***summary model*** in ***memory***
  - Nearest neighbor data structure based on application defined distance metric.
  - Encapsulates distance information.
  - Hierarchical organization summarizes data at arbitrary granularity.
  - In-memory model improves throughput and latency.
- Updates limited to once per ***batching interval***
  - Local insertions without synchronizing model.
  - Improves throughput and latency.
  - Temporal locality for social sensing content.
  - Eventual consistency through gossip.

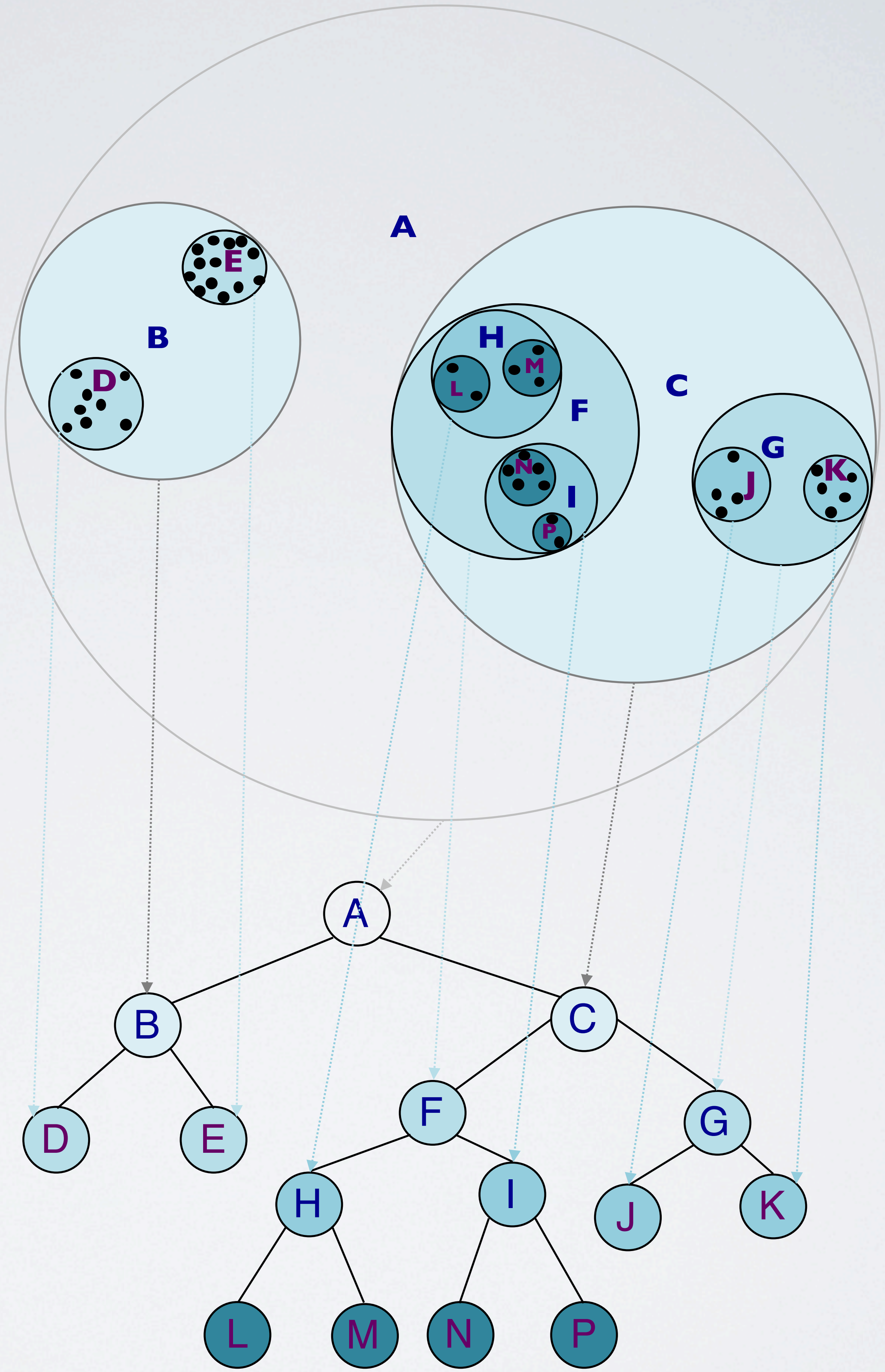
System Design



Summary Model

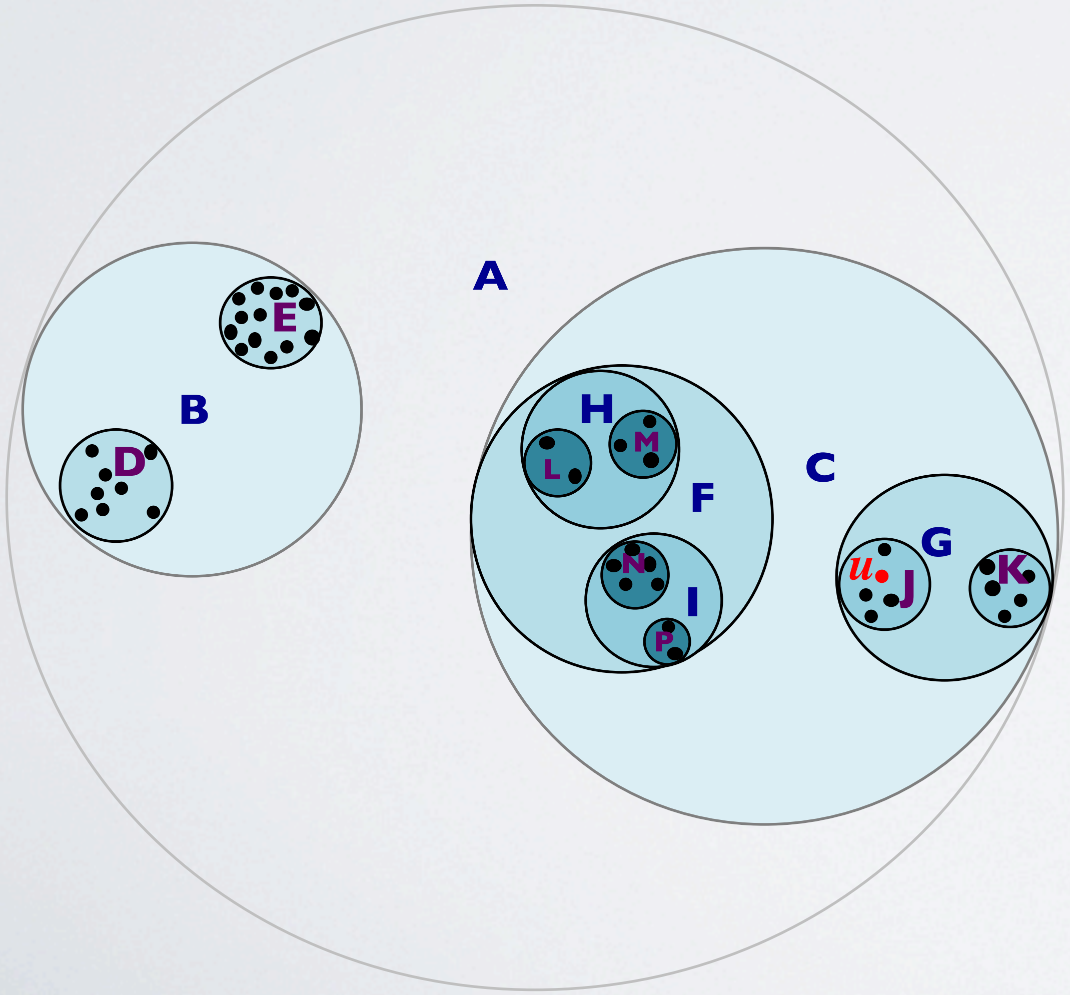
- Represented as a binary tree
    - Constructed using a divide and conquer paradigm.
- Current set of vectors partitioned
    - Into two disjoint sets using *2-means clustering.*
    - Centroids of the two sets become ***two children.***
        - Each set contains points nearest to the centroid of that set, compared to the centroid of the other set.
    - Both sets scheduled for further division in *parallel.*
    - Distance function must satisfy *Triangle Inequality.*
- Generated once per batching interval
    - Updates summary model based on the incoming objects in the last interval.

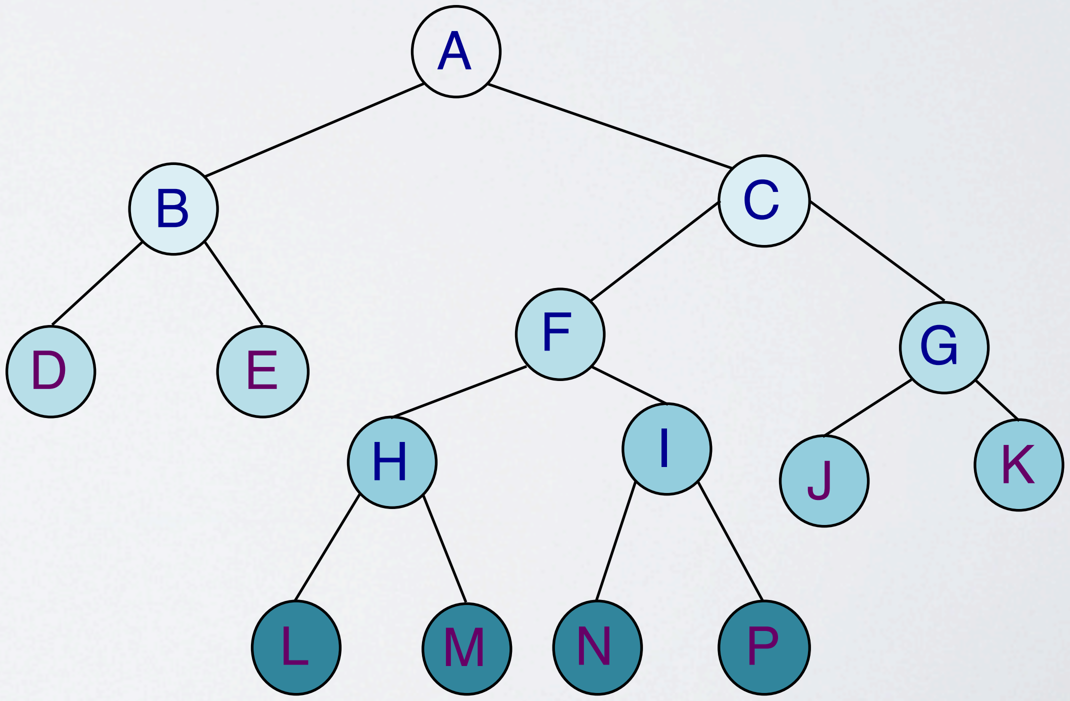
Points to Summary Model



Insertion

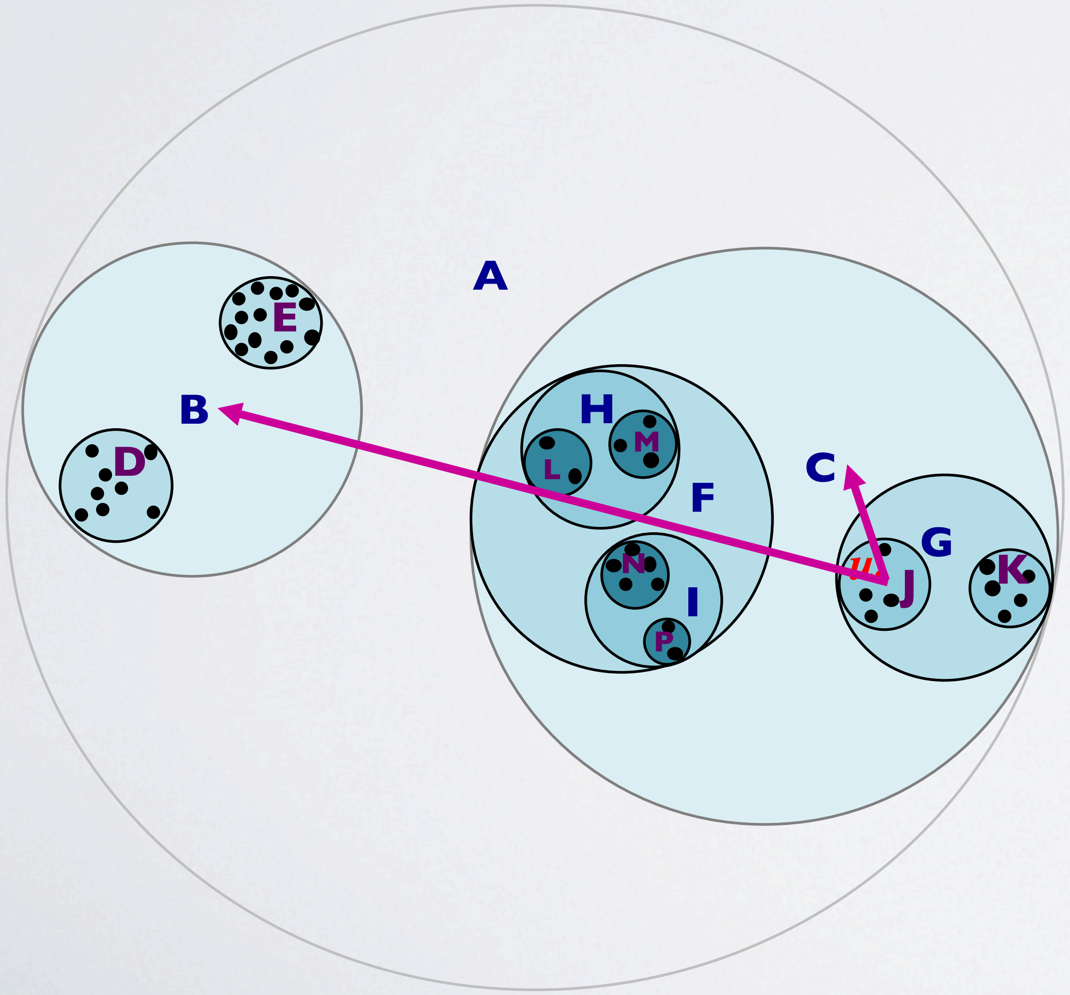
- For example, consider two dimensional space and Euclidean distance
- Insert object u to appropriate cluster

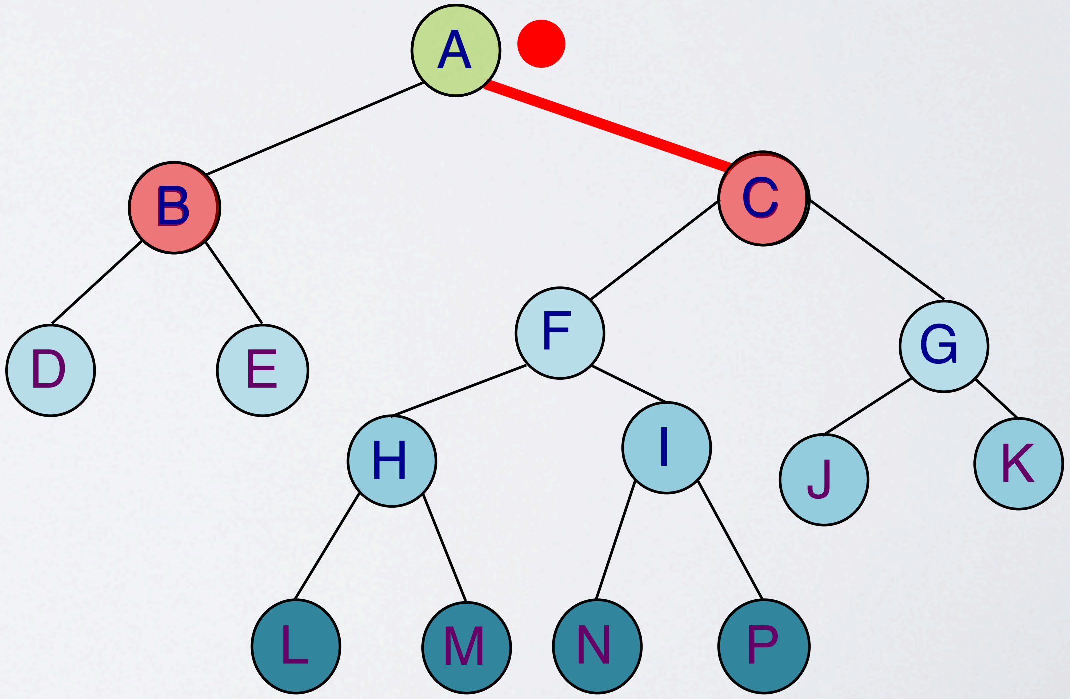




Insertion

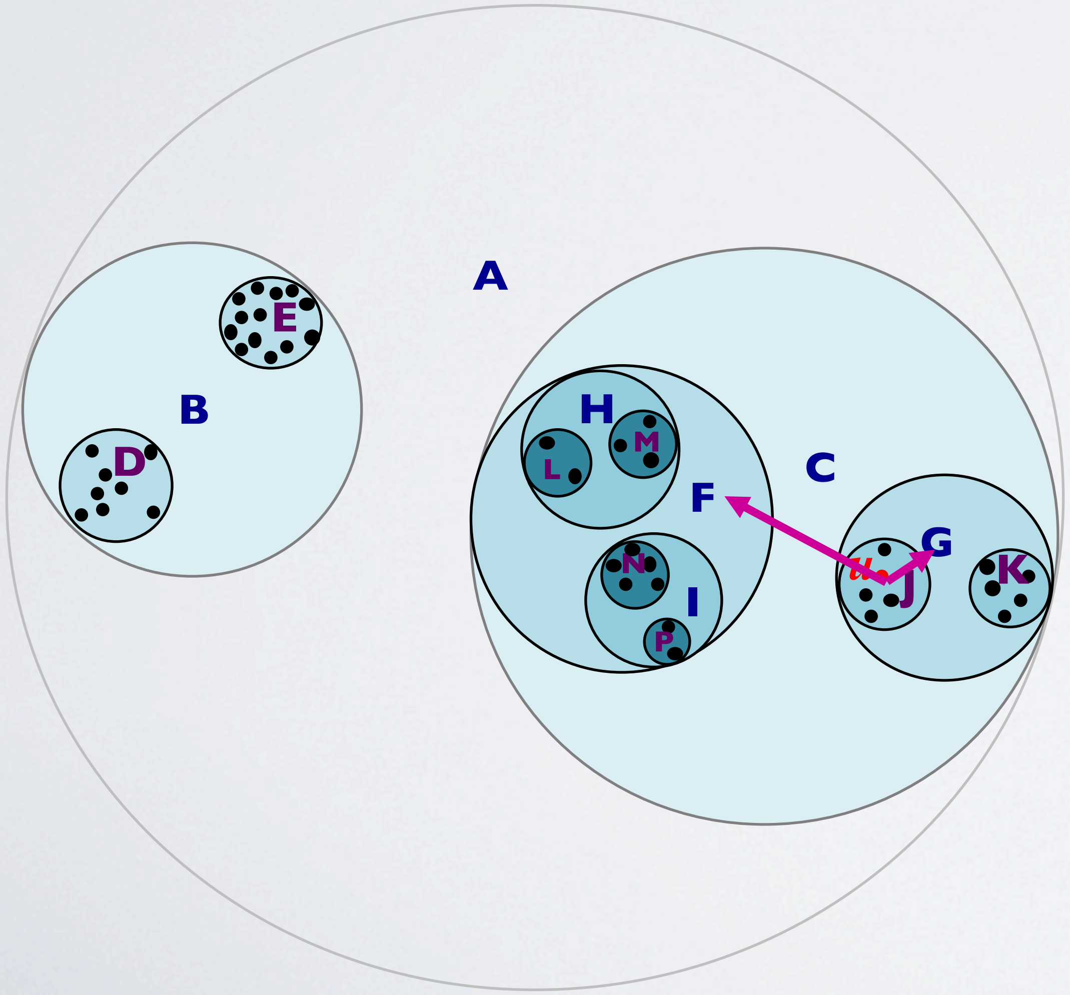
- u arrives at root, A
- u compared to B and C
- u pushed to C

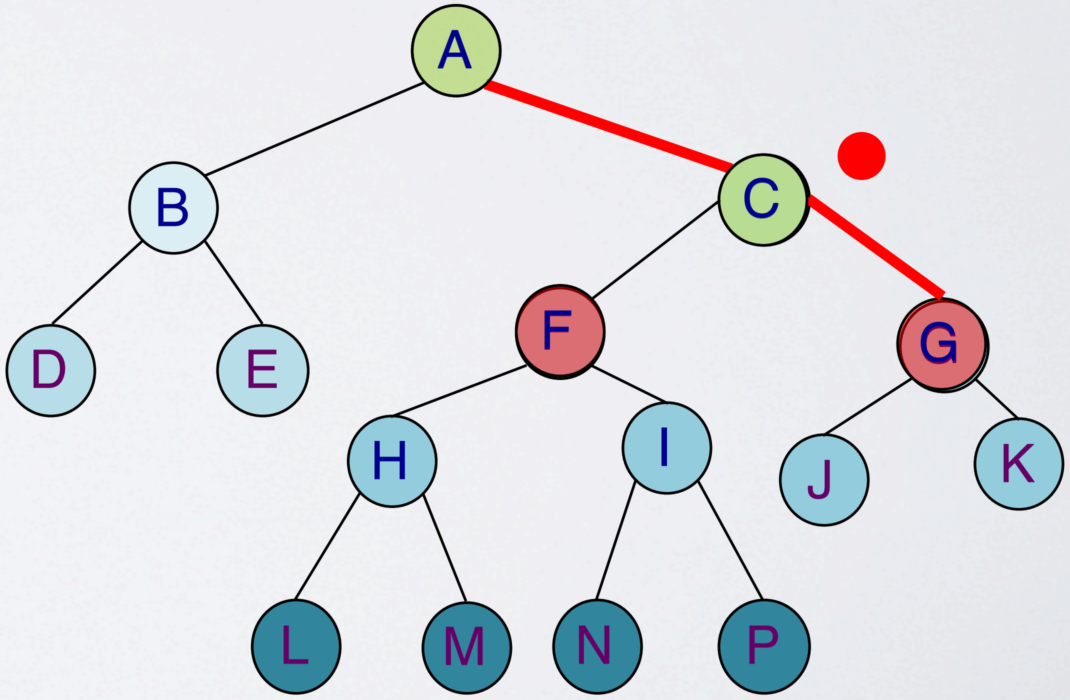




Insertion

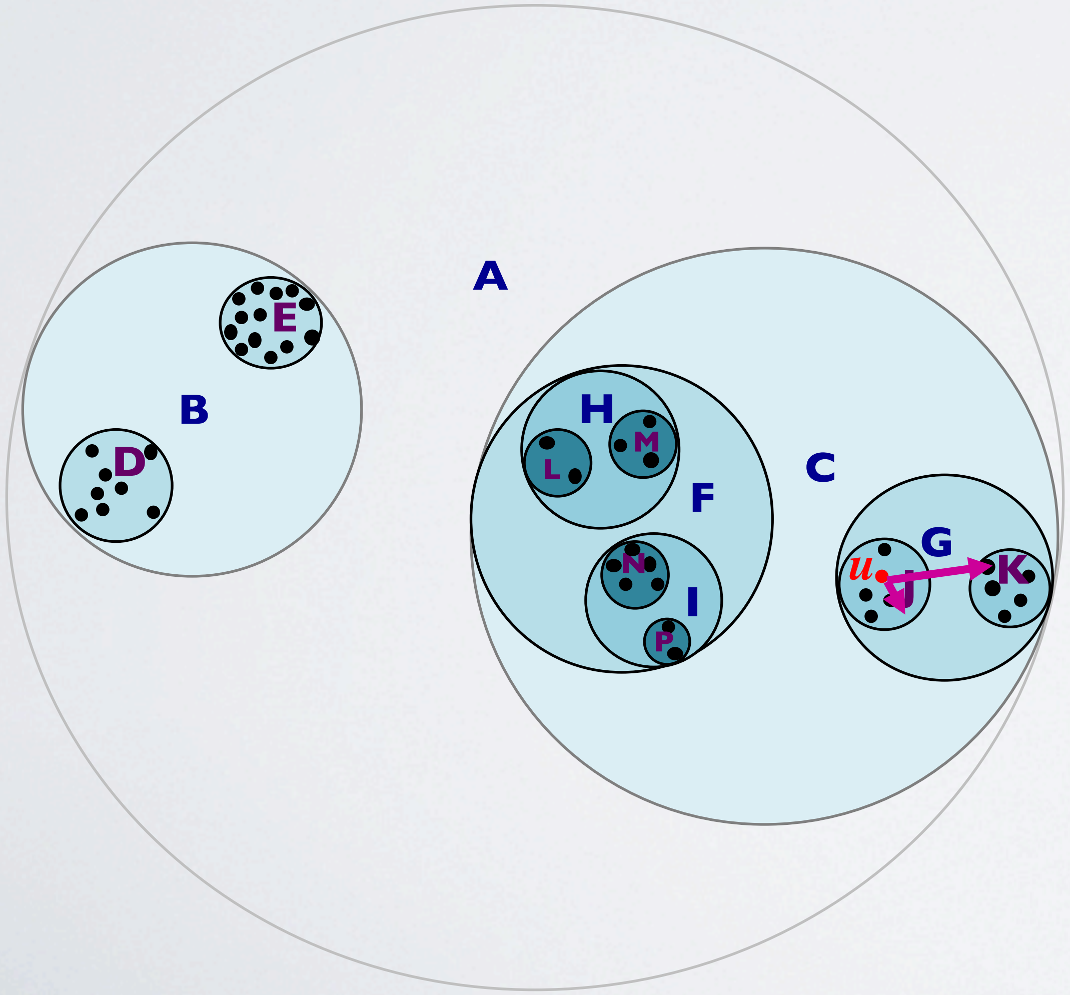
- u compared to F and G
- u pushed to G

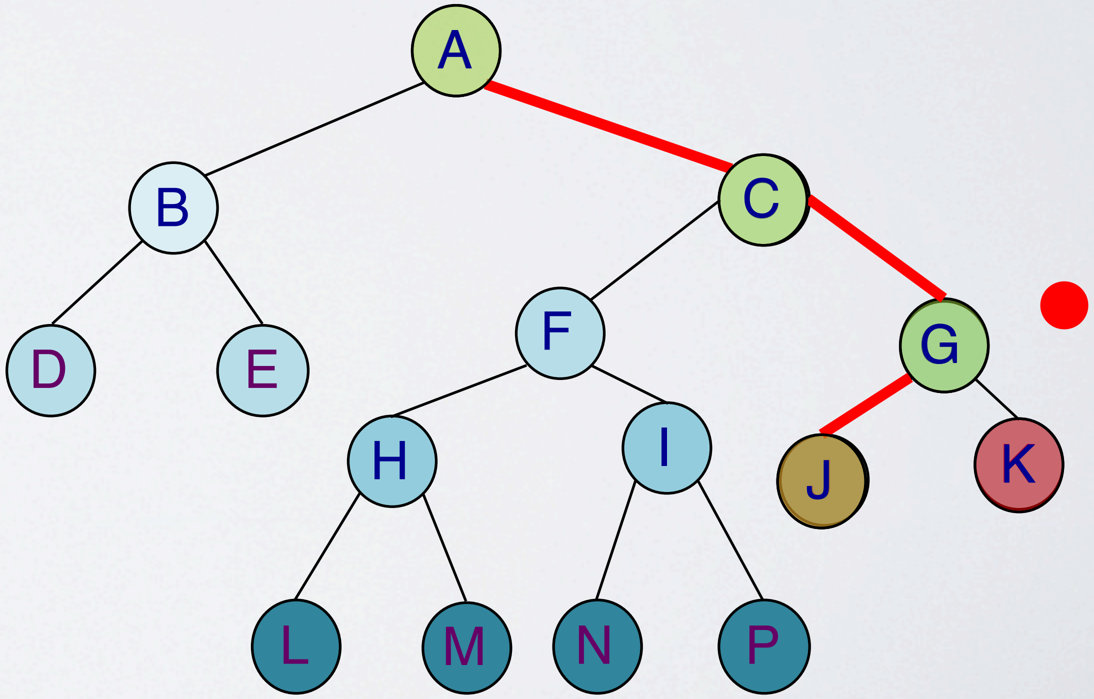




Insertion

- u compared to J and K
- u pushed to J
- u assigned to cluster J





Twitter Application

- Twitter based Fact-finder application
    - Tweets crawled using search api during events.
        - Crimea, Sandy, Fukushima, Egypt Revolution etc.
    - More than 4 million tweets.
- Customization Interface
    - Tweets converted to high dimensional vectors of tokens and frequencies.
    - Used Tanimoto distance function.

$$d(\mathbf{u},\mathbf{v}) = 1 - \frac{\mathbf{u}.\mathbf{v}}{\mathbf{u}^2 + \mathbf{v}^2 - \mathbf{u}.\mathbf{v}}$$

- Summary Query Interface
    - Queries SocialTrove using keywords.
    - Ranks matching summaries according to credibility.

Implementation

- Summary Model
  - Written in Java
  - MapReduce job to generate model runs on Spark.
  - Broadcasts serialized model every interval.
- Client Proxy
  - Java / Python – based on different client applications.
  - Serves matching objects from distributed memcache
- Input Proxy
  - Python – periodically polls twitter via search api
- Distributed Storage
  - HDFS (Hadoop Distributed File System)

SocialTrove Testbed

- Runs on UIUC Green Data Center
  - http://greendatacenters.web.engr.illinois.edu
  - Capacity: 40 large nodes, 40 small nodes
- Testbed Setup
  - 8 nodes
  - Intel Xeon E5-2620 – 12 logical cores
  - 16 GB Ram per node
  - 1 Gbps Ethernet – Star topology
  - 1 TB disk
- One driver machine
  - apollo3.cs.illinois.edu
  - Intel Xeon E5-2690 – 40 logical cores
  - 128 GB Ram
  - 16 TB Raid-5 disk

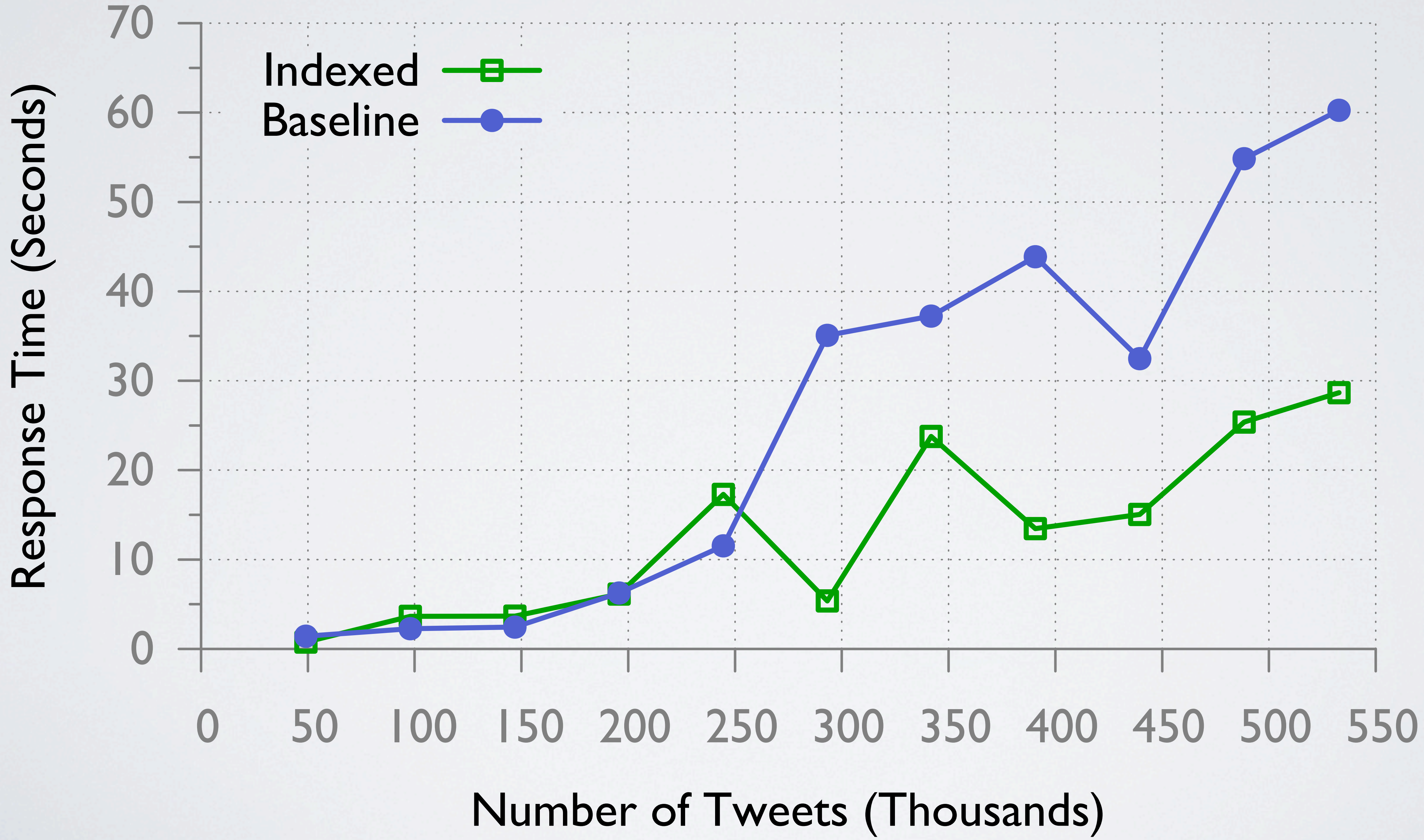
Evaluation

- SocialTrove compared to 4 schemes
  - **Baseline**
    - No summary model computed.
    - Load balanced assignment to storage.
  - **Indexing**
    - No summary model computed.
    - Tweets indexed by keywords.
  - **Summary Baseline**
    - Computes a flat summary model in advance.
  - **Summary Indexing**
    - Computes a keyword indexed summary model.

Response Time

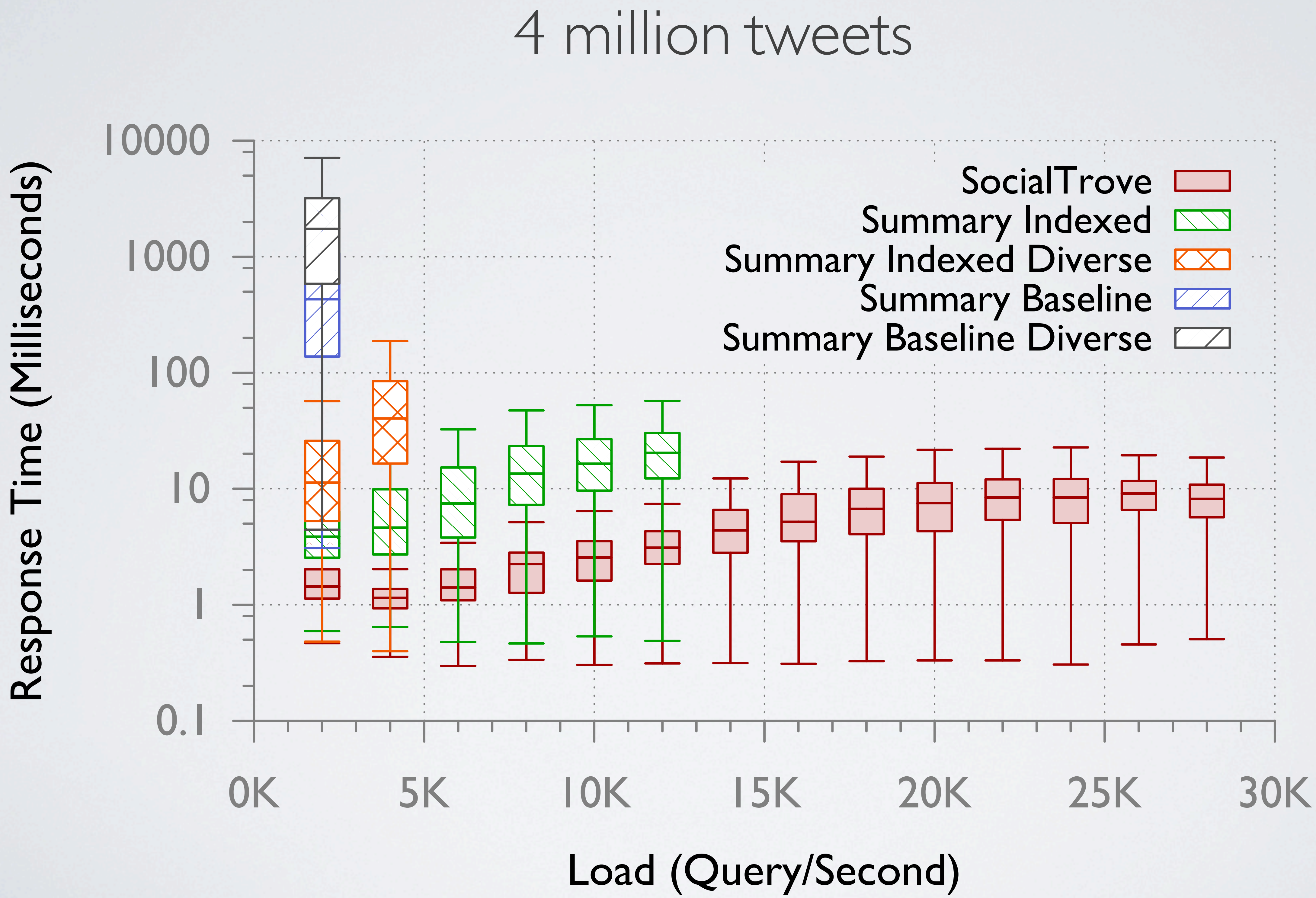
without summary model

Response time is very low with on demand summarization



Response Time

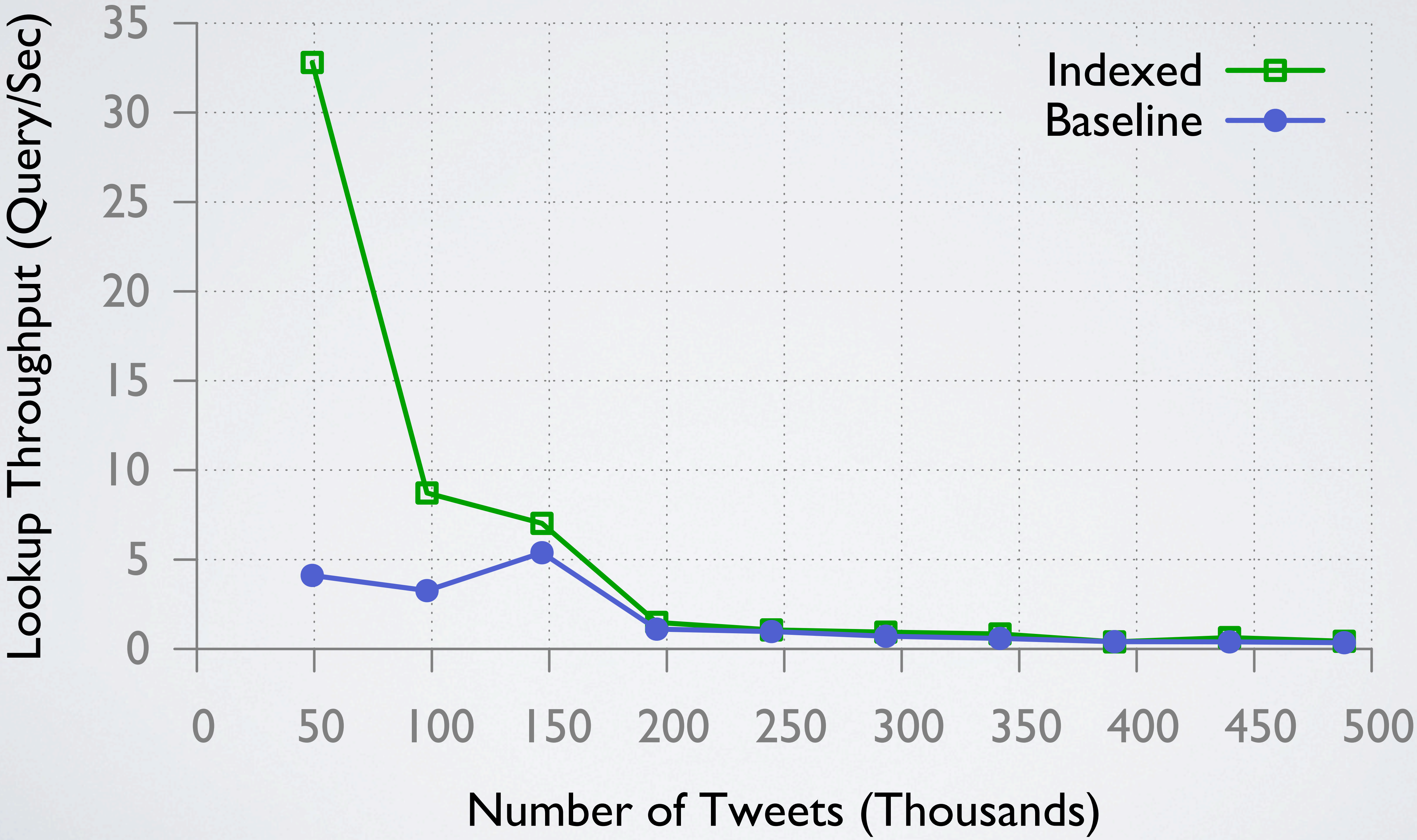
with summary model



Lookup Throughput

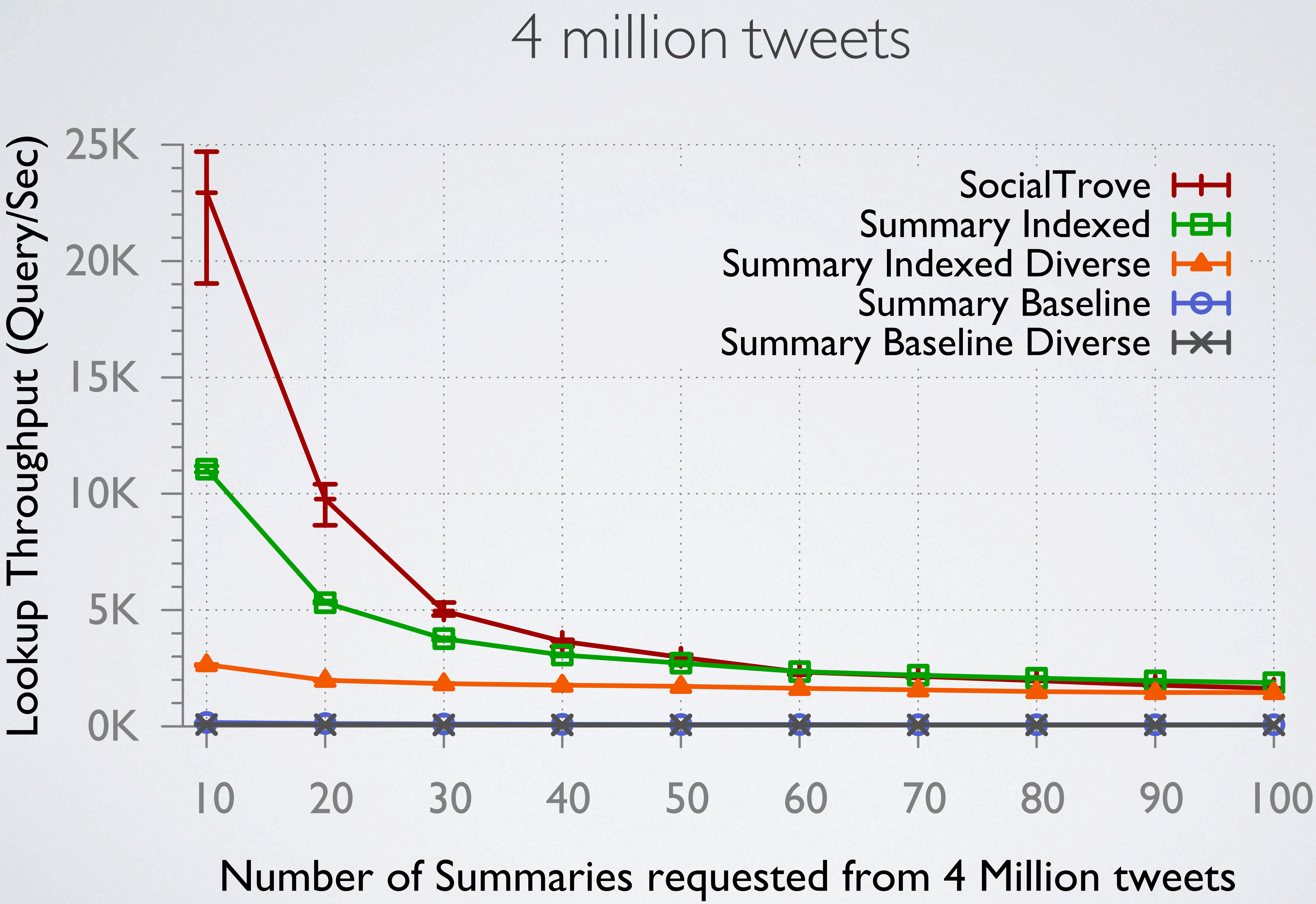
without summary model

On demand summarization is not scalable

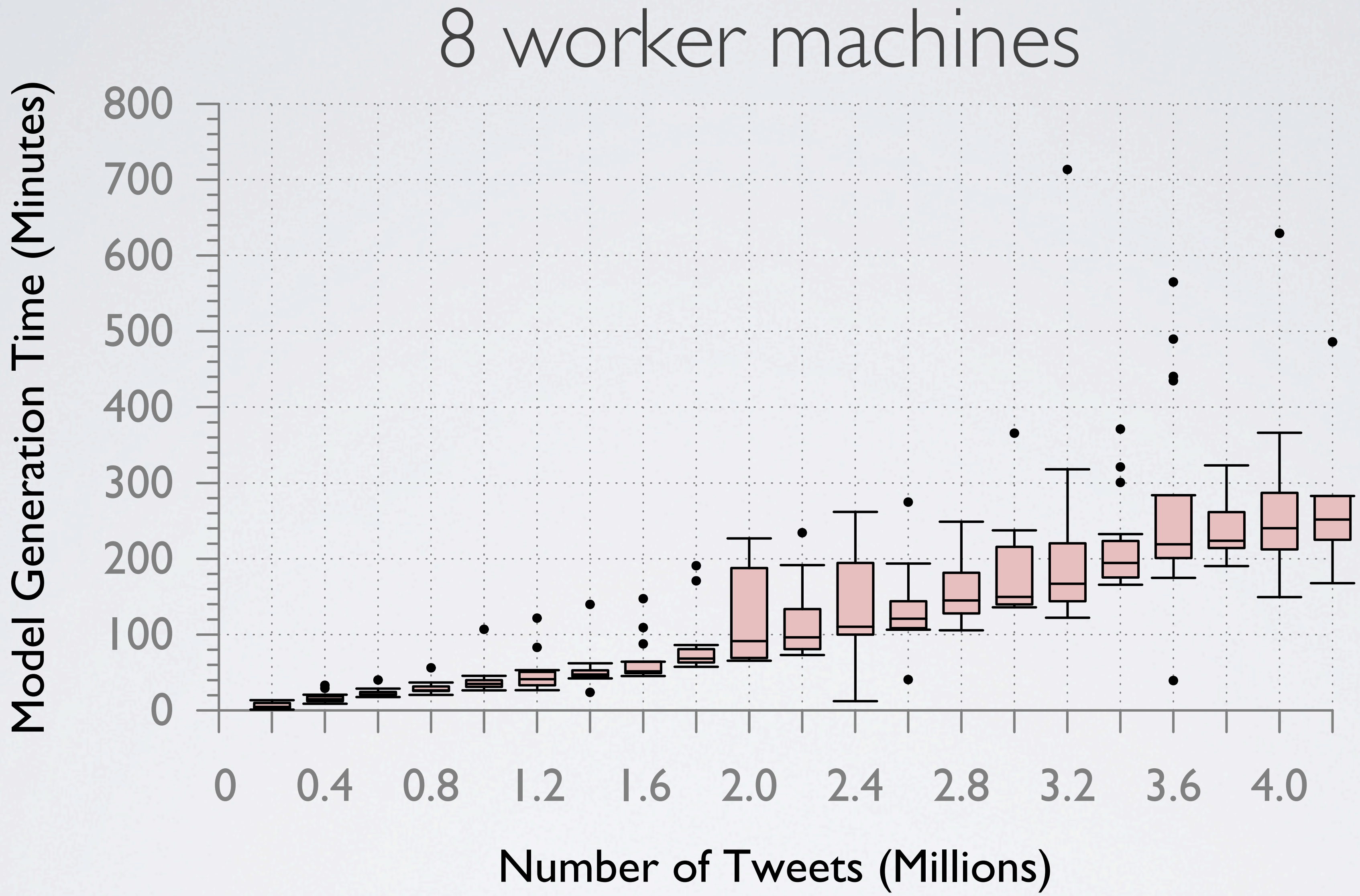


Lookup Throughput

with summary model



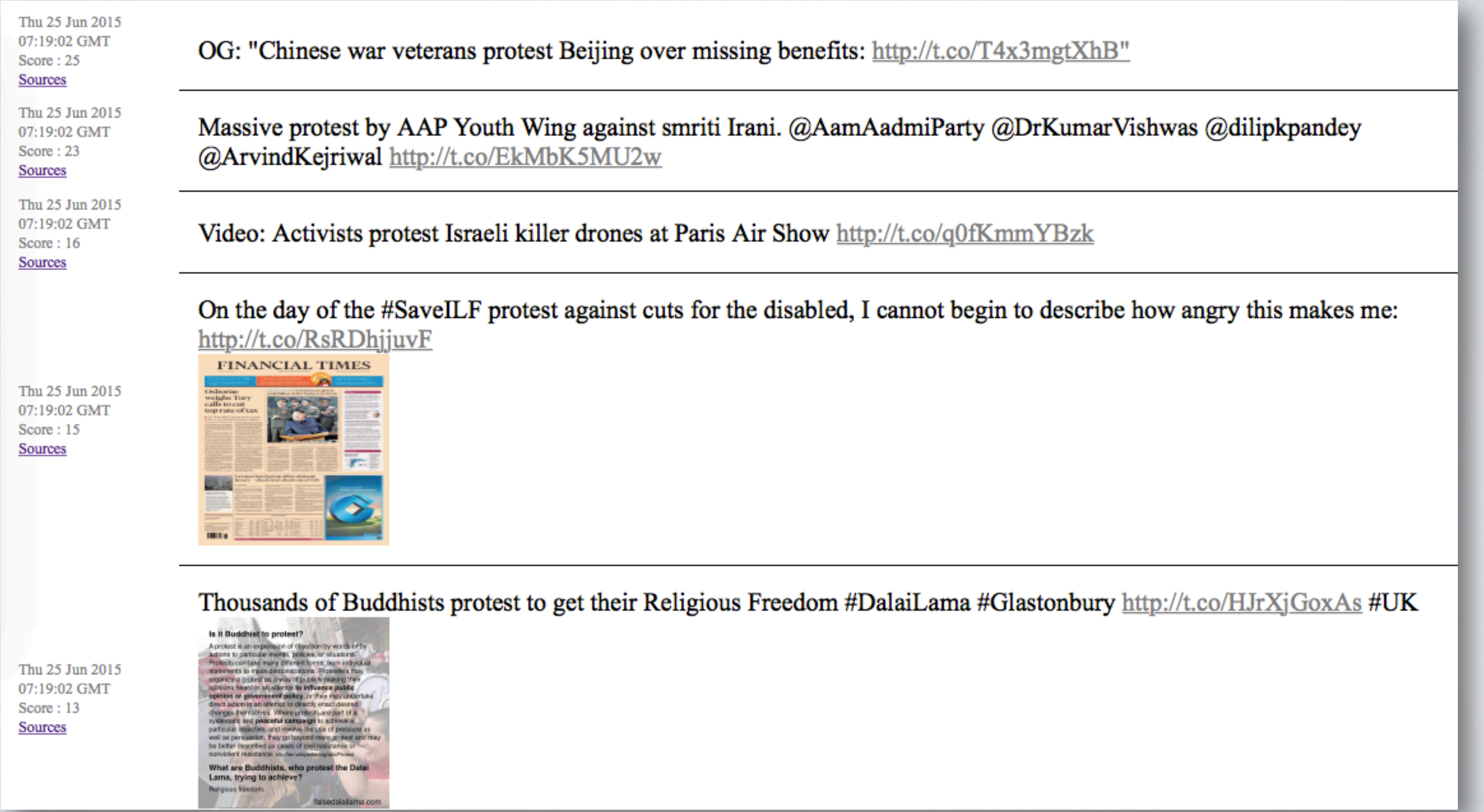
Model Generation Time



Number of Workers	Model Generation Time (4 million tweets)
4	337 minutes
8	239 minutes
11	195 Minutes

Ranked Summary

Protests – Real-time Summary - Thu 25 Jun 2015 07:19:02 GMT



Thu 25 Jun 2015
07:19:02 GMT
Score : 25
Sources

OG: "Chinese war veterans protest Beijing over missing benefits: http://t.co/T4x3mgtXhB"

Thu 25 Jun 2015
07:19:02 GMT
Score : 23
Sources

Massive protest by AAP Youth Wing against smriti Irani. @AamAadmiParty @DrKumarVishwas @dilipkpandey @ArvindKejriwal http://t.co/EkMbK5MU2w

Thu 25 Jun 2015
07:19:02 GMT
Score : 16
Sources

Video: Activists protest Israeli killer drones at Paris Air Show http://t.co/q0fKmmYBzk

Thu 25 Jun 2015
07:19:02 GMT
Score : 15
Sources

On the day of the #SaveILF protest against cuts for the disabled, I cannot begin to describe how angry this makes me: http://t.co/RsRDhjjuvF

FINANCIAL TIMES

Osborne weighs Tory calls to cut top rate of tax

Thu 25 Jun 2015
07:19:02 GMT
Score : 13
Sources

Thousands of Buddhists protest to get their Religious Freedom #DalaiLama #Glastonbury http://t.co/HJrXjGoxAs #UK

Is it Buddhist to protest?

A protest is an expression of objection by words or by actions to particular events, policies, or situations. Protests can take many different forms, from individual statements to mass demonstrations. Protesters may organize a protest as a way of publicly making their opinions heard in an attempt **to influence public opinion or government policy**, or they may undertake direct action in an attempt to directly enact desired changes themselves. Where protests are part of a systematic and **peaceful campaign** to achieve a particular objective, and involve the use of pressure as well as persuasion, they go beyond mere protest and may be better described as cases of civil resistance or nonviolent resistance. [illegible]

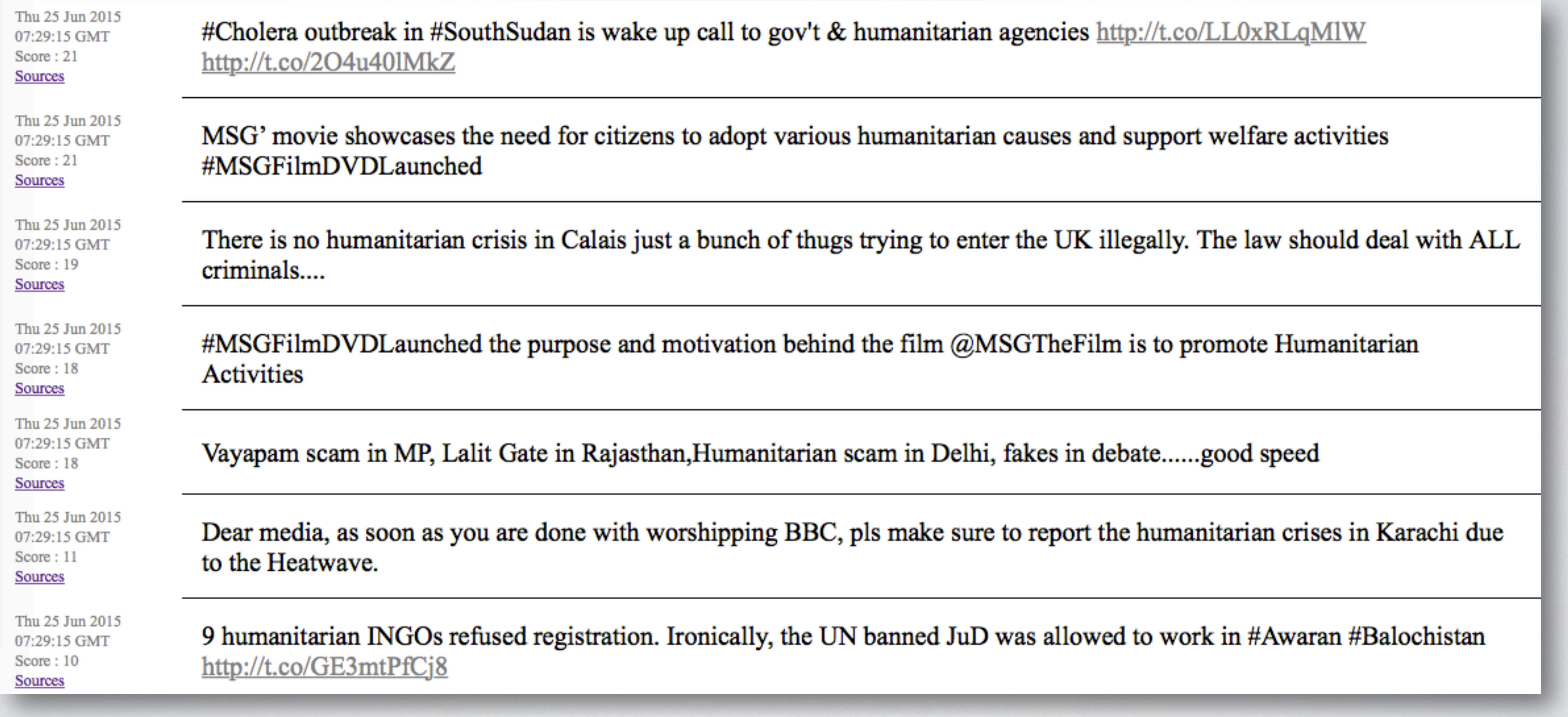
What are Buddhists, who protest the Dalai Lama, trying to achieve?

Religious freedom.

falsedalailama.com

Ranked Summary

Disasters – Real-time Summary - Thu 25 Jun 2015 07:29:15 GMT



Thu 25 Jun 2015 07:29:15 GMT Score : 21 Sources	#Cholera outbreak in #SouthSudan is wake up call to gov't & humanitarian agencies http://t.co/LL0xRLqMlW http://t.co/2O4u40lMkZ
Thu 25 Jun 2015 07:29:15 GMT Score : 21 Sources	MSG' movie showcases the need for citizens to adopt various humanitarian causes and support welfare activities #MSGFilmDVDLaunched
Thu 25 Jun 2015 07:29:15 GMT Score : 19 Sources	There is no humanitarian crisis in Calais just a bunch of thugs trying to enter the UK illegally. The law should deal with ALL criminals....
Thu 25 Jun 2015 07:29:15 GMT Score : 18 Sources	#MSGFilmDVDLaunched the purpose and motivation behind the film @MSGTheFilm is to promote Humanitarian Activities
Thu 25 Jun 2015 07:29:15 GMT Score : 18 Sources	Vayapam scam in MP, Lalit Gate in Rajasthan,Humanitarian scam in Delhi, fakes in debate......good speed
Thu 25 Jun 2015 07:29:15 GMT Score : 11 Sources	Dear media, as soon as you are done with worshipping BBC, pls make sure to report the humanitarian crises in Karachi due to the Heatwave.
Thu 25 Jun 2015 07:29:15 GMT Score : 10 Sources	9 humanitarian INGOs refused registration. Ironically, the UN banned JuD was allowed to work in #Awaran #Balochistan http://t.co/GE3mtPfCj8

Ranked Summary

Crimean Crisis – Offline Summary – February 2014
Thousands at Moscow rally against Russian intervention in #Ukraine
Man in Ukraine plays the piano to help calm down a riot. http://t.co/fdNAc0cfJ2
For Crimea, Google Shows Different Borders Based on Your Location: Russia's Minister of Communications and Mass Media http://t.co/vlHGYlibOC
Militants in eastern #Ukraine were equipped with Russian weapons and the same uniforms as those worn by Russian forces that invaded Crimea.
50,000 #Ukraine supporters march in Moscow protest Russia's intervention in #Crimea.
Some russian tanks on ukrainian border already painted with 'peacekeeping' slogans. How much longer until the 'humanitarian intervention'?
I've been speaking to @BarackObama about the situation in Ukraine. We are united in condemnation of Russia's actions. http://t.co/7Rk2k8iOIK
Ukrainian Defense Ministry says its lone submarine has been taken by Russians.
Ukraine prepares armed response as city seized by pro-Russia forces
Ukraine crisis: Nato warns Russia against further intervention - BBC News

Conclusions

- Summarizes social content stream with arbitrary granularity, in real time.
- Caches a summary model in memory.
  - Scale-out input and client query proxies.
- High throughput and low response time.
  - Thanks to asynchronous updates to summary model, and temporal locality in social stream.
- Outperforms traditional indexing methods.
- Limitation: Evaluation performed only in the context of tweets.