Cache Provisioning for Interactive NLP Services

Jaimie Kelley and Christopher Stewart
The Ohio State University

Sameh Elnikety and Yuxiong He
Microsoft Research

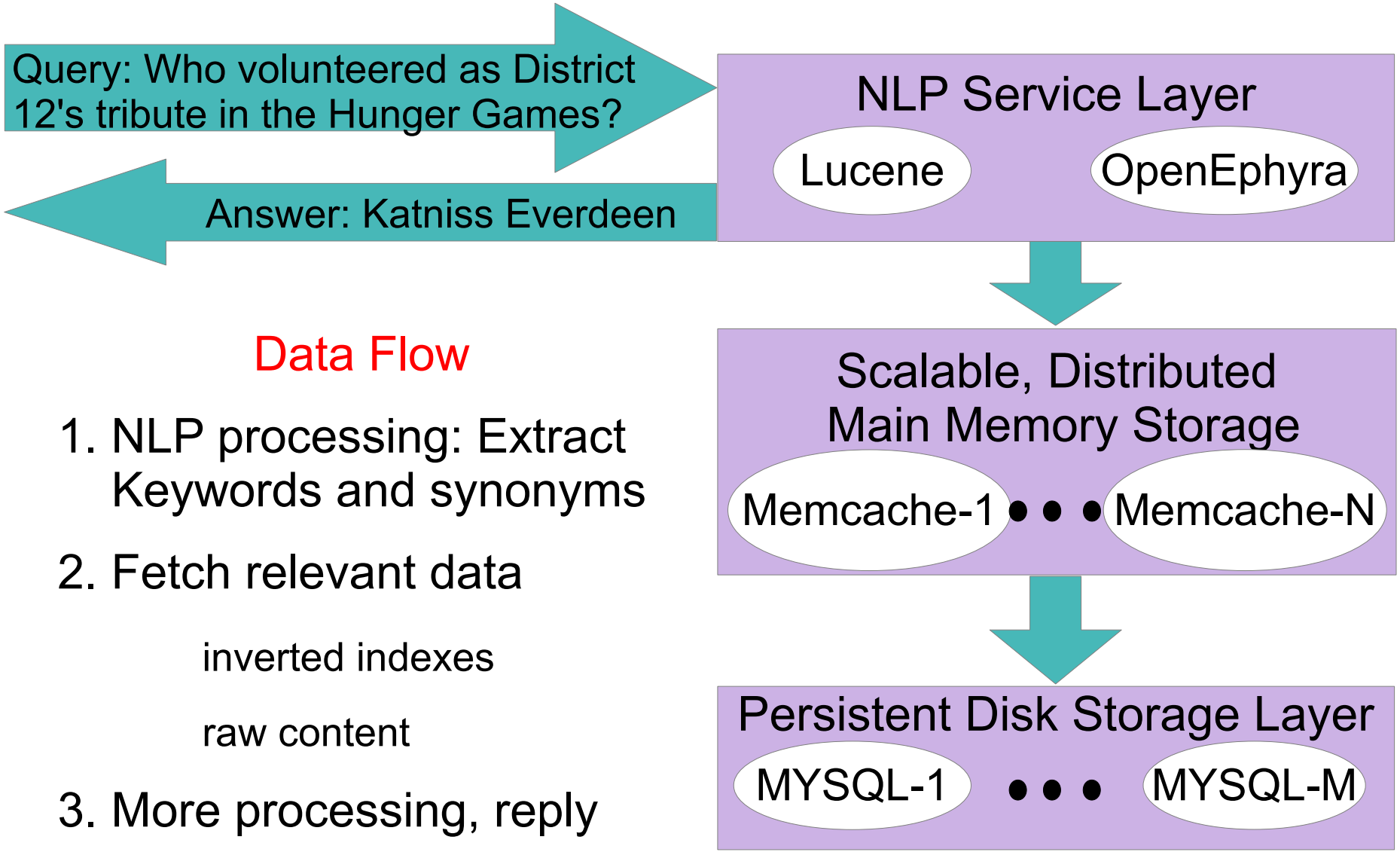
Interactive NLP

"Watson looks at the language. It tries to understand what is being said and find the most appropriate, relevant answer..."

Rob High, IBM Fellow

- Natural Language Processing (NLP) Service
  - Users issue queries via network messages
  - Analyzes and understands human text in context
  - Response times should be fast (bounded latency)
  - Examples: Bing, Google, IBM Watson, OpenEphyra

Interactive NLP



Data Flow

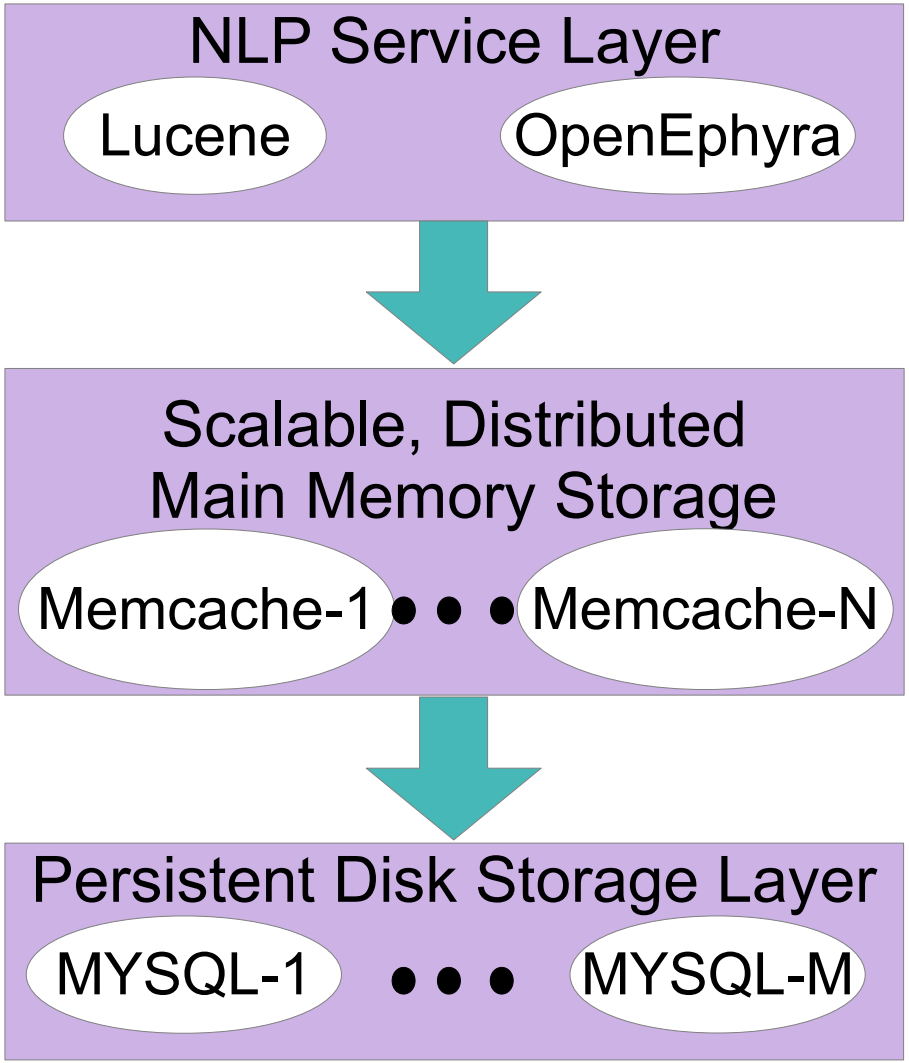
1. NLP processing: Extract Keywords and synonyms
2. Fetch relevant data
   - inverted indexes
   - raw content
3. More processing, reply

Interactive NLP Services

- To compete with Jeopardy champions, IBM Watson had 3 sec. latency bound
- Our experience: 8th graders -> 4 sec. Bound
- Internet services demand sub-second response times

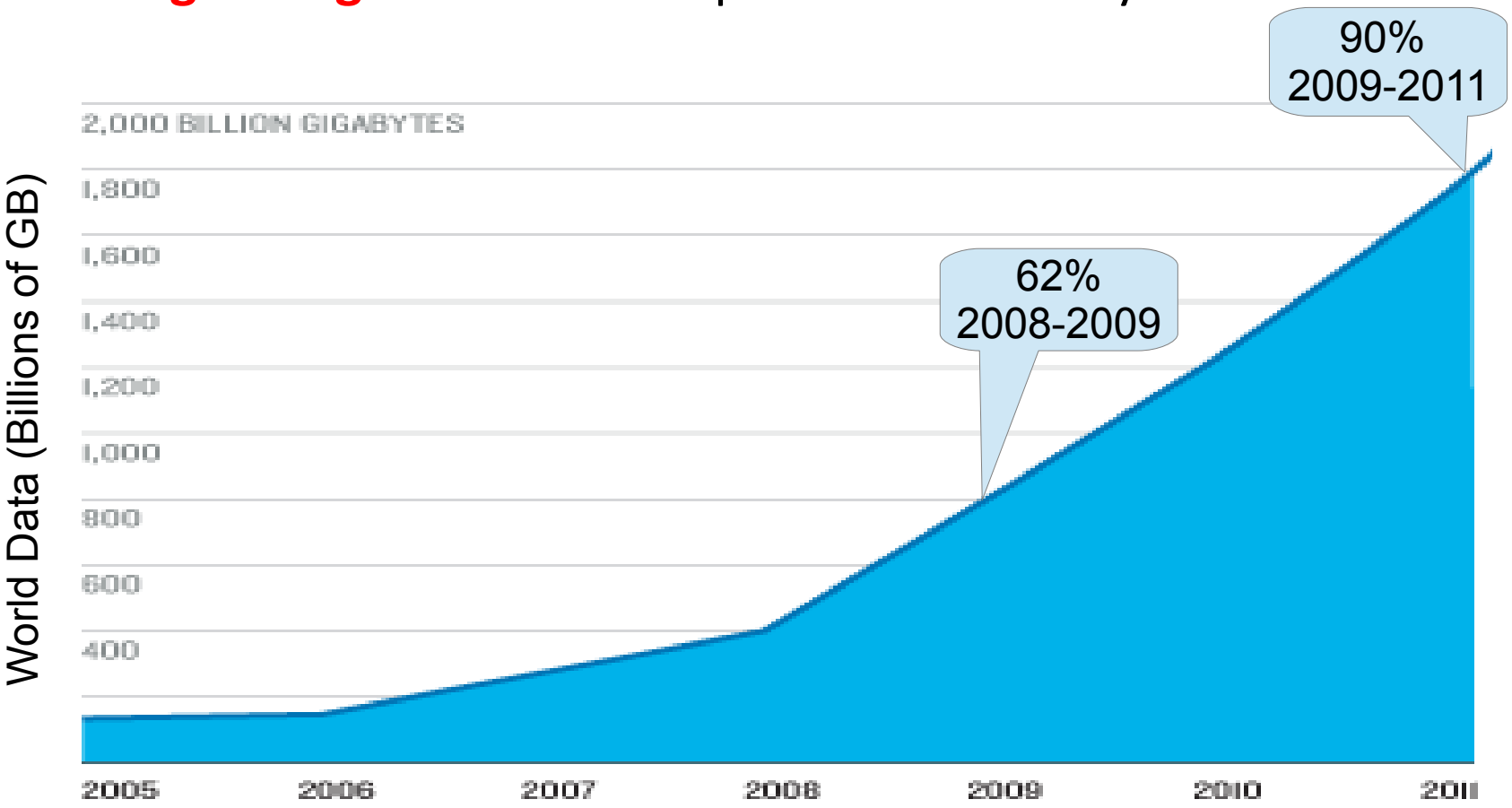
Tight Latency Bounds

- **Access DRAM in parallel**
- **Disk accesses timeout!**



Problem: Data Growth

Data is **growing too fast** to keep in main memory.



Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet

Cache Management

- When the data is too large to fit in cache, what should we evict?

- Traditional Compute Workloads:

  Every data access is needed to answer a query

  Evict least recently used (LRU)

- **NLP services:**

  Only some data are needed (redundant content)

  Remove redundant data with little quality loss

Quality-Aware Cache Management

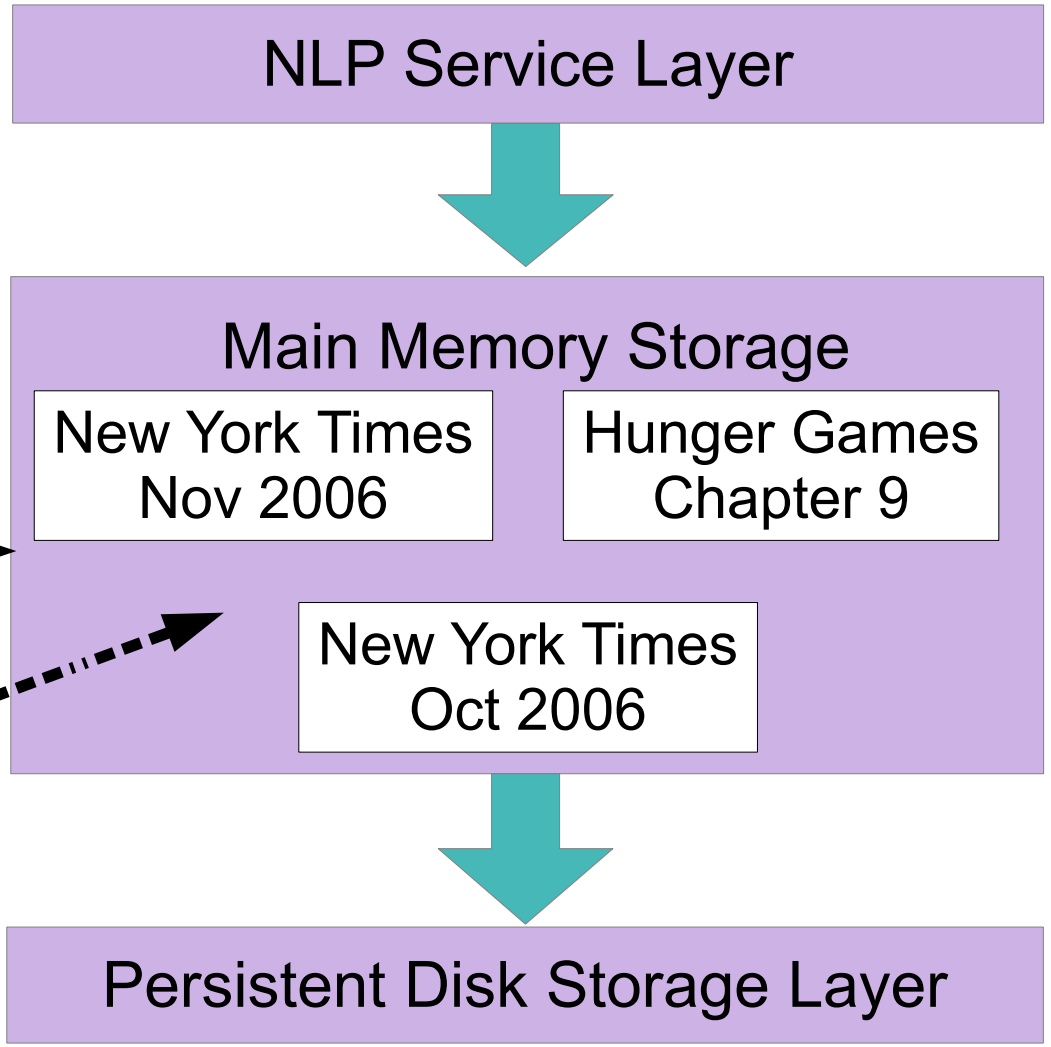
For NLP services, evict data that will cause the least quality loss.

New Data

Harry Potter Chapter 1

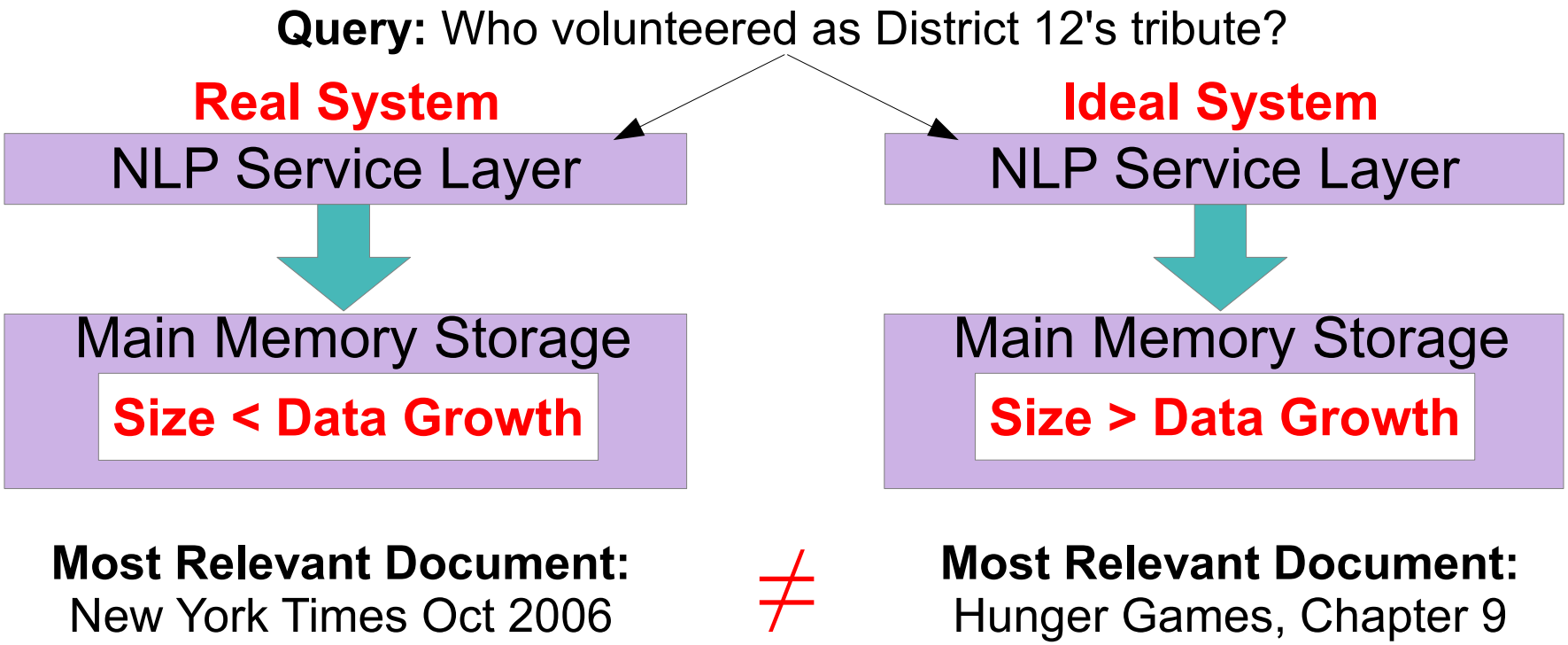
What Data Will LRU Evict?

New York Times or Hunger Games



Our Approach

- Does existing cache management work well for NLP?

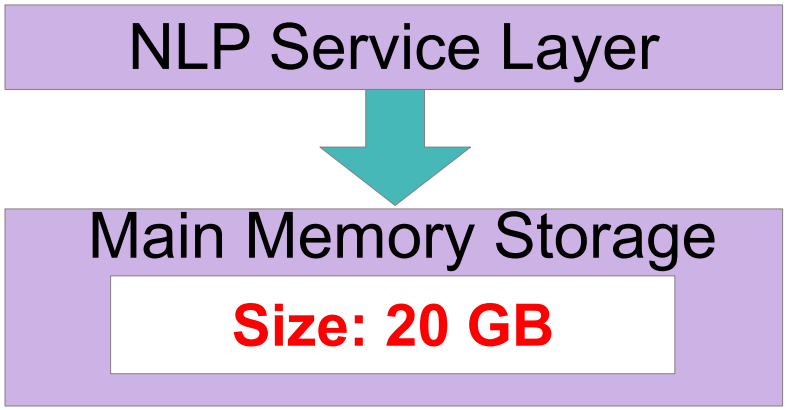


We measure ***quality loss***

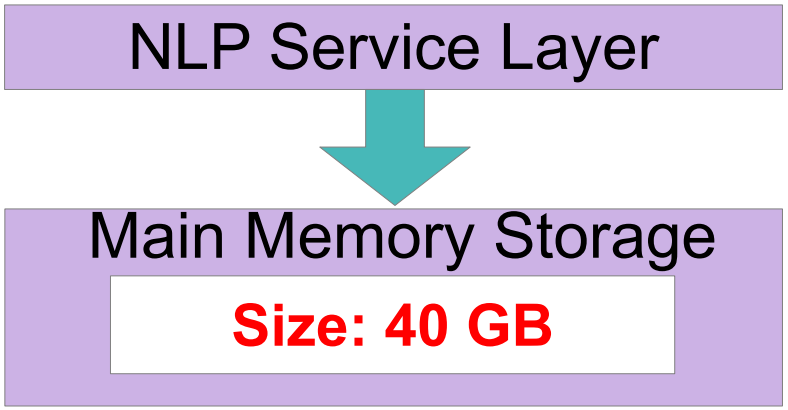
i.e., dissimilarity between real and ideal

Our Approach

- Can quality-aware cache management reduce provisioning costs over time?



Cache Miss Rate: 40%
Avg. Quality Loss: 15%



Cache Miss Rate: 20%
Avg. Quality Loss: 15%

Outline

- Introduction
- Defining Quality Loss
  - Intuition, base model, full model
- Quality Loss in NLP Services
  - Representative queries, data sets, infrastructure, results
- Quality-Aware Cache Provisioning
- Conclusion

Intuition: What is quality loss?

Real System:

Query:
Who volunteered as District 12's tribute?

Answers:

Katniss Foxface

Harry Potter

Peeta Mellark

Jeanine F. Pirro

Ideal System:

Query:
Who volunteered as District 12's tribute?

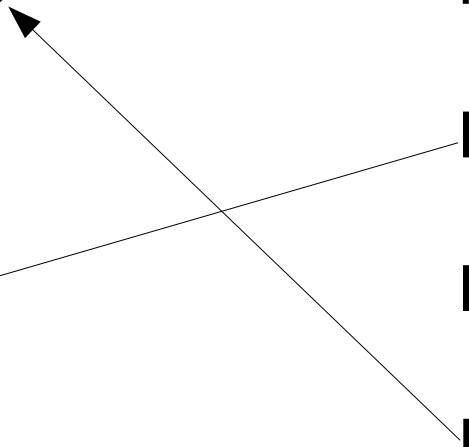
Answers:

Katniss Everdeen

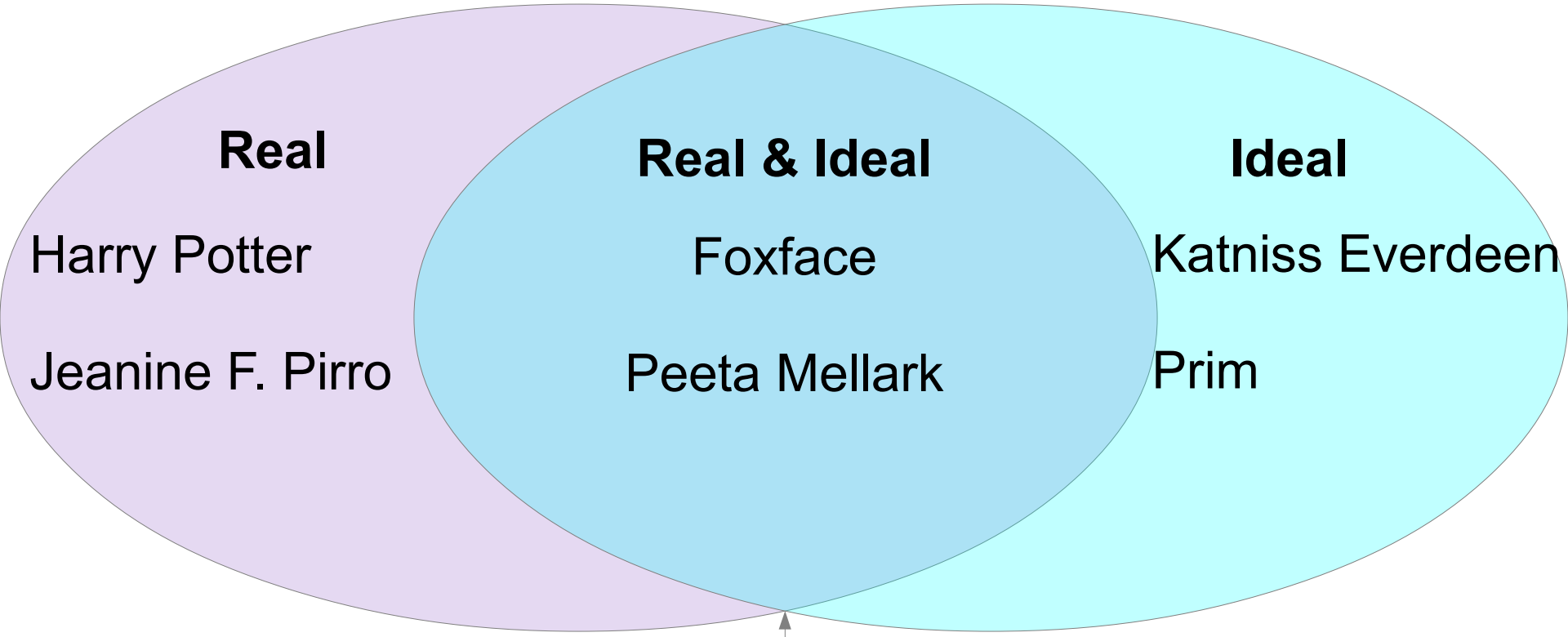
Peeta Mellark

Prim

Foxface



Base Model: Quality Loss



$$S(w, \hat{w}, D, Q) = 1 - \frac{\sum_Q \sum_K \Phi(\sum_{k2} |R_{q,k}(\hat{w}, D) \cap R_{q,k2}(w, D)|)}{|Q|K}$$

Full Model: Quality Loss

- NLP responses present challenges:
  - **Synonyms**
    - Answers from *ideal setup* fall within categories
    - *Real setup* should match categories

      Query: Flowers in Washington State

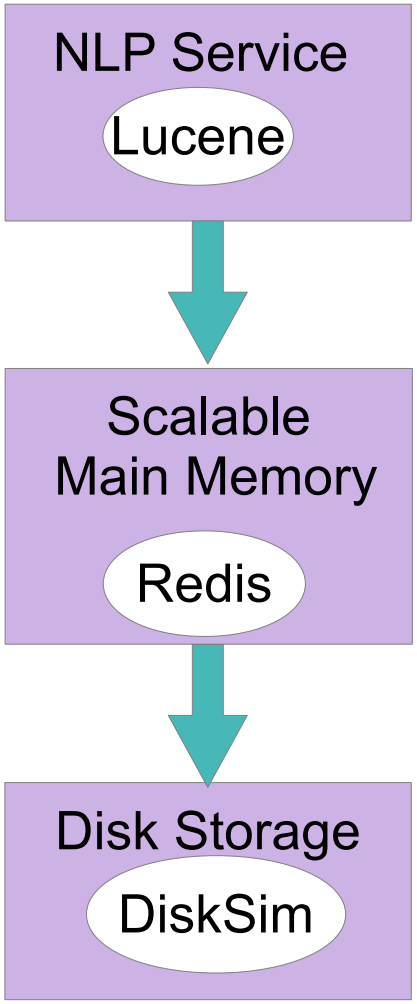
      Answers: florists, gardening, Coast Rhododendron

  - **Noise Tolerance**
    - Answers from the *real setup* can be a superset of answers from *ideal setup*

Outline

- Introduction
- Defining Quality Loss
  - Intuition, base model, full model
- Quality Loss in NLP Services
  - Query trace, data sets, infrastructure, results
- Quality-Aware Cache Provisioning
- Conclusion

Infrastructure



Real NLP Service:

10 sec latency bound

Analyzes keywords

Ranks all indexes requested from storage.

Distributed Cache:

Set max size < | data |

Implemented interface between cache and disk.

Disk Storage:

Two 3-TB hard disks

We used Lucene libraries

Ideal NLP Service:

Exact same processing

Distributed Cache:

No set maximum size

9 GB / cache node

Provision more as needed

Disk Storage:

No timeouts

Key Insight: Ideal setup returns the result created by processing all relevant data without timeouts.

Obtaining a Query Trace

- Google Trends
  - Trace of most popular queries per category 2004-2013
  - Most (over 70%) are multiple word queries

Jan. 2004 Books:	June 2009 Books:	Sept. 2013 Books:
The Bible	The Bible	The Bible
The Lord of the Rings	Alice's Adventures in ...	Fifty Shades of Grey
The Da Vinci Code	The Lord of the Rings	The Great Gatsby
1984	Midnight Sun	Under the Dome
Kama Sutra	1984	The Hunger Games
Romeo and Juliet	Kama Sutra	Psalms
Hamlet	Romeo and Juliet	The Lord of the Rings
Macbeth	Quran	Sword Art Online
To Kill a Mockingbird	Diagnostic and ...	1984
The World Factbook	Diccionario de la lengua...	The 85 Ways to Tie a Tie

Data Sets

New York Times

October 2004 – March 2006

Total Index size: 3 GB

Max Data/Month: 88 MB

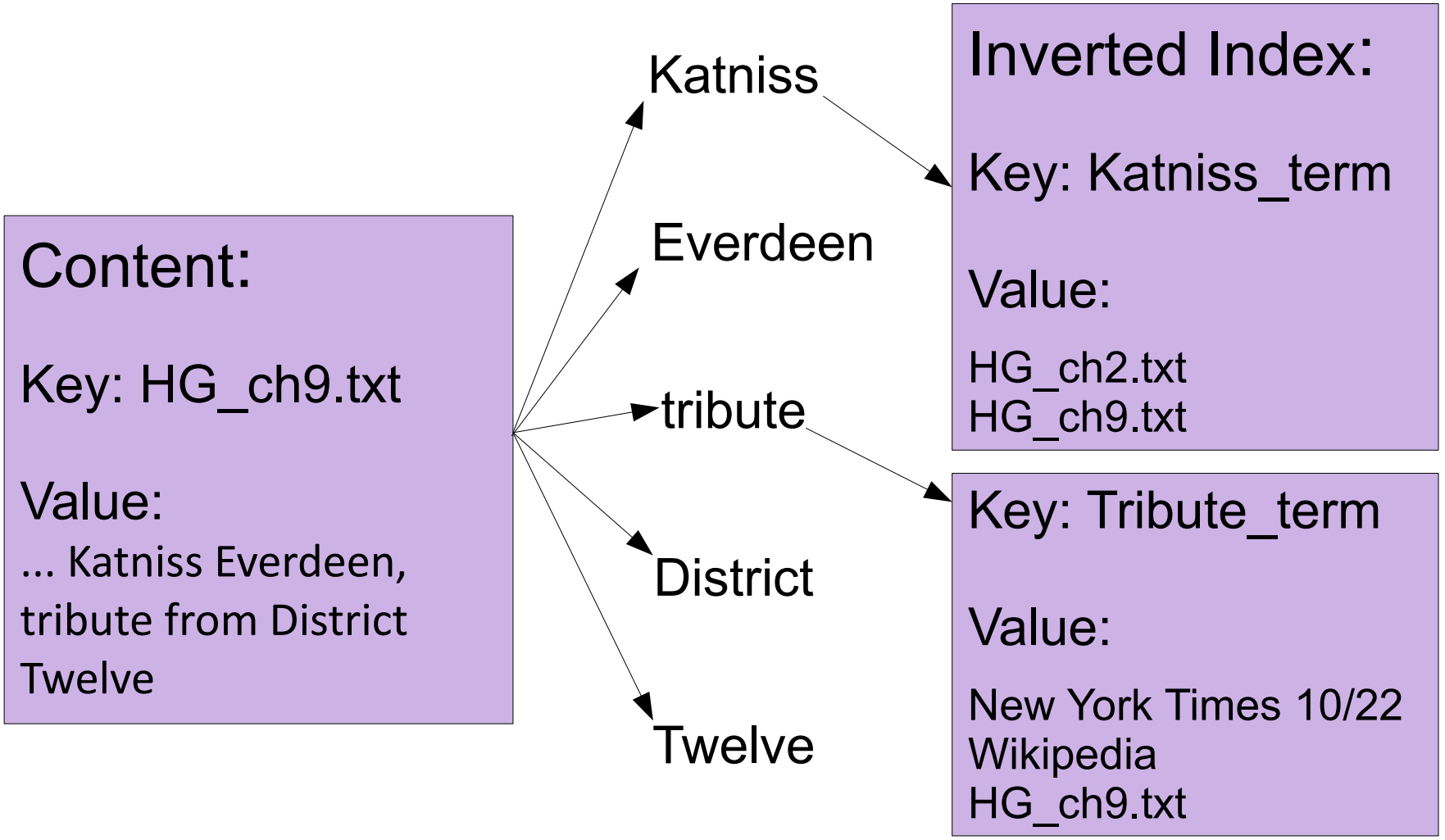
Wikipedia

January 2001 – March 2013

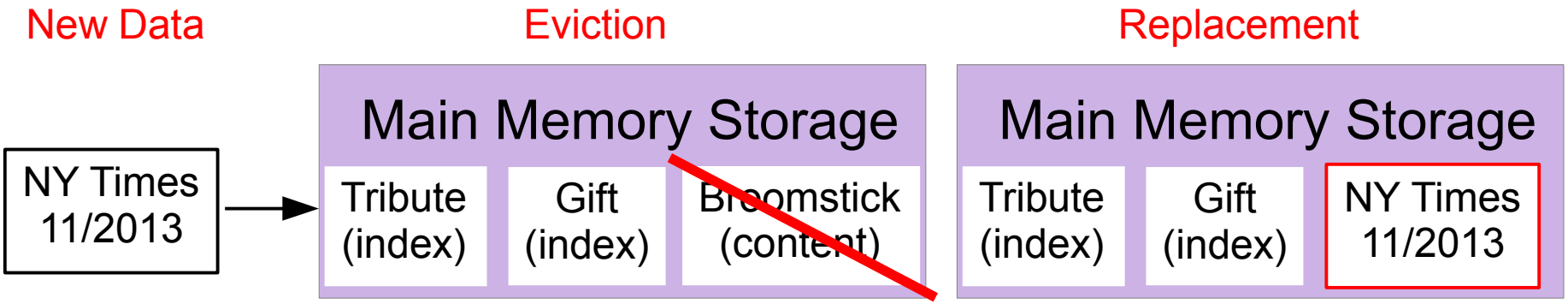
Total Index size: 4.7 TB

Max Data/Month: 30 GB

Types of Data

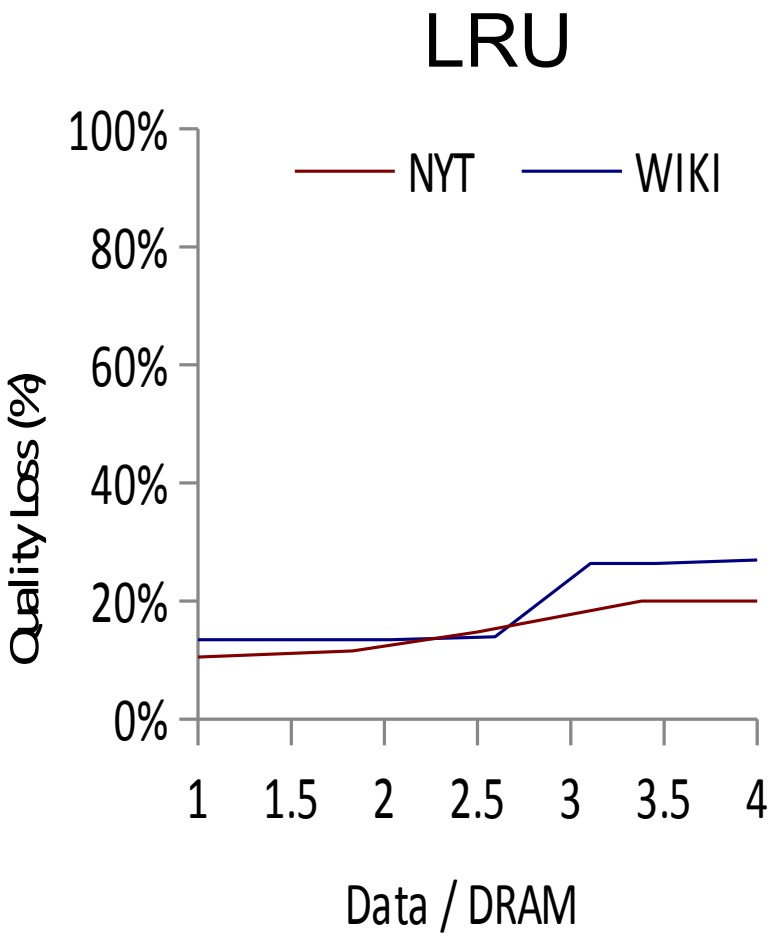


Caching Policies: LRU



- Least Recently Used Cache Management
  - Common approach in distributed stores
  - Implemented in Redis
- Infrequent search terms are sent to disk, unable to be accessed within latency bounds

Caching Policies: LRU



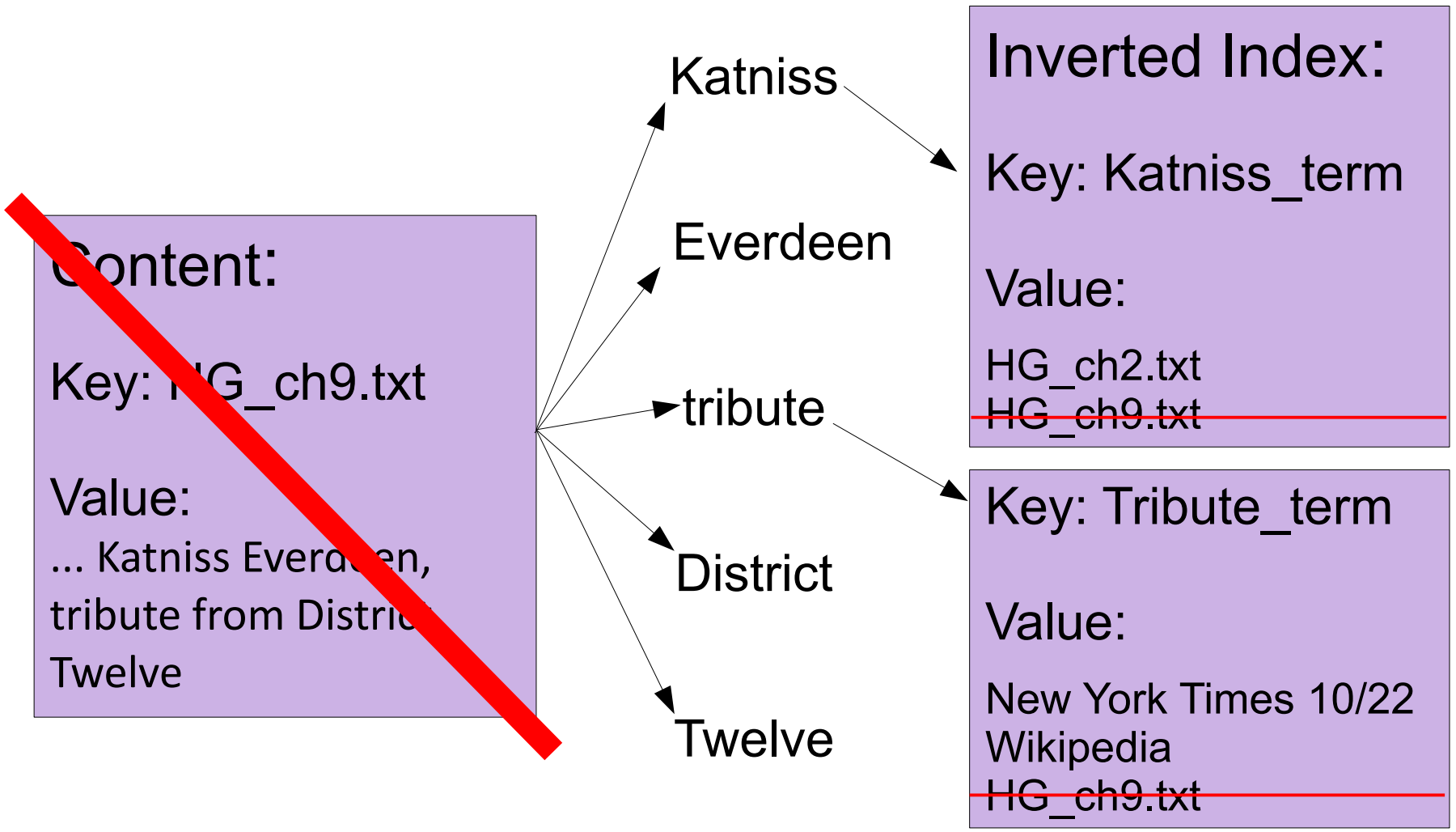
Quality loss rises as terms are evicted

Multiple word queries and single-word synonyms benefit from redundancy

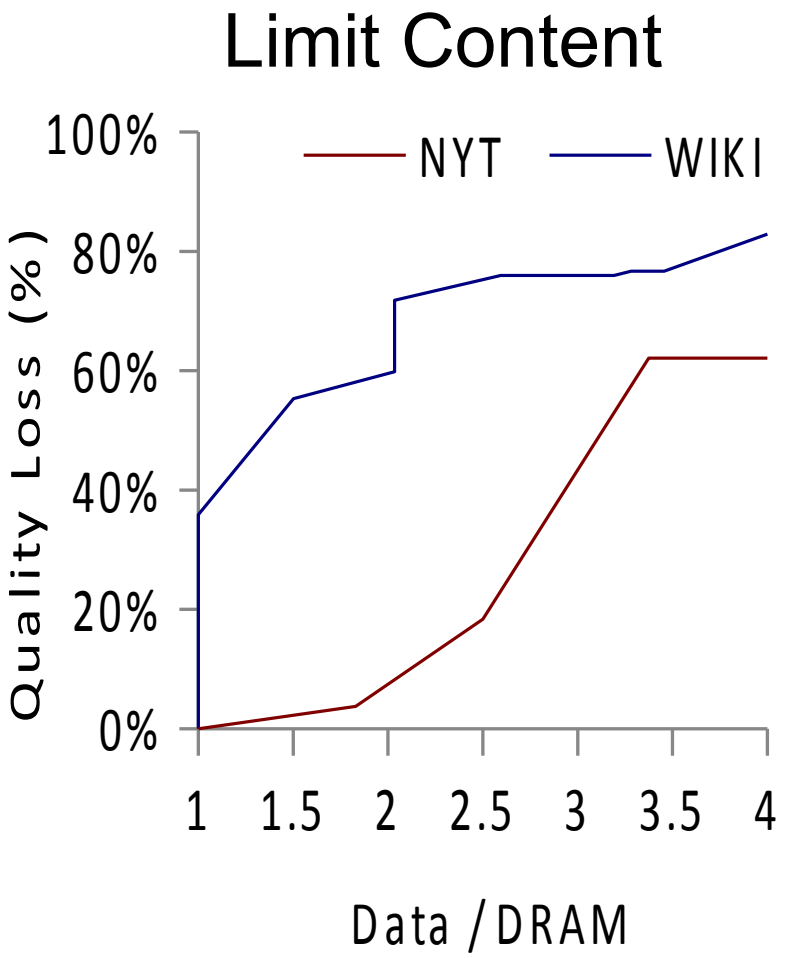
- Non-evicted data may overlap evicted data
- Answers from ideal system can come from non-evicted terms

In both Wikipedia pages and NYT articles, content related to queries were found in non-evicted terms.

Caching Policies: Limit Content



Caching Policies: Limit Content



Quality loss rises as key documents are excluded from index

Documents may contain content that effectively has the same meaning

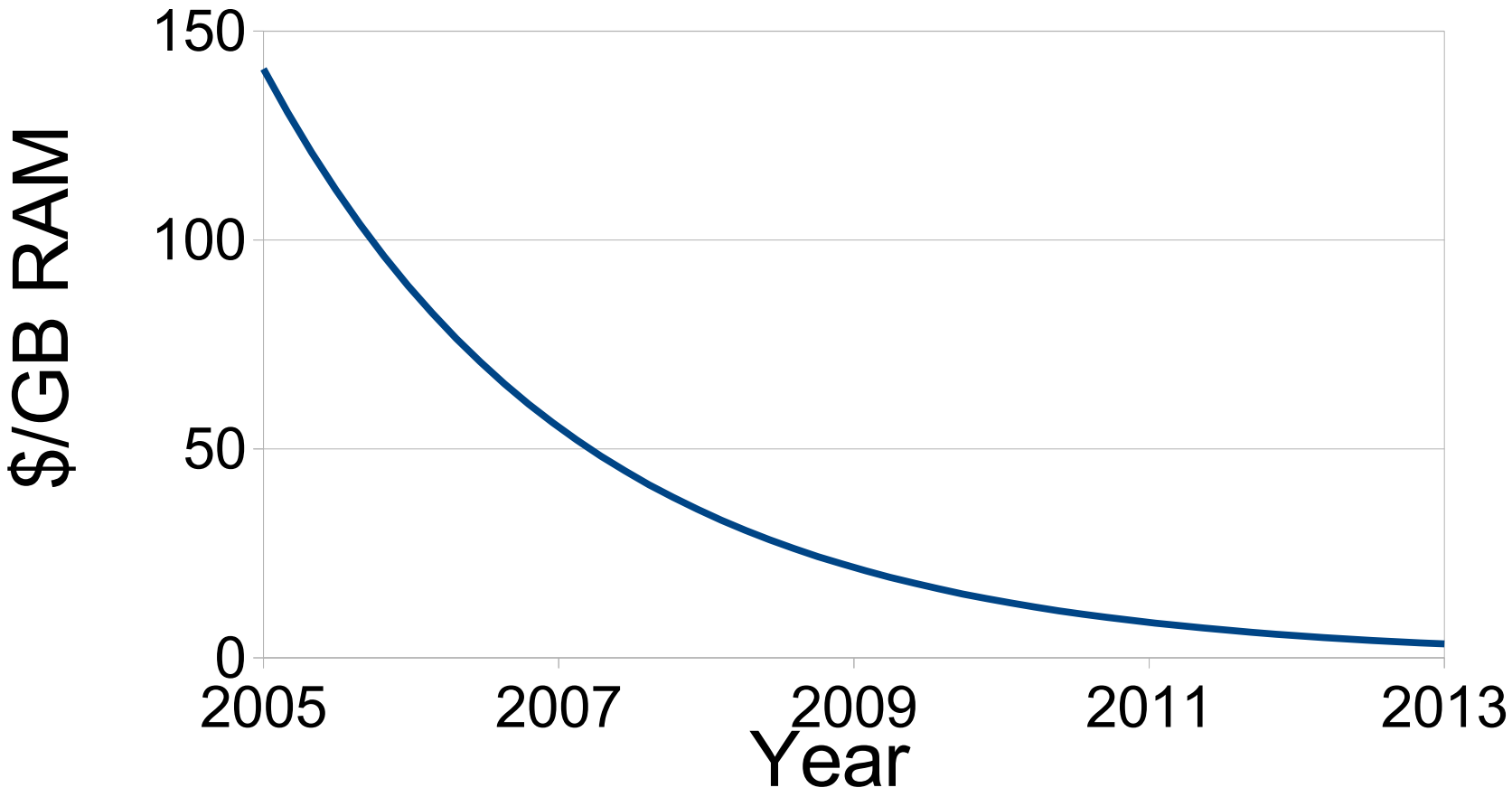
Wikipedia pages are less redundant, as a result of their review process

On average 1 of 2 NYT articles can be removed with low quality loss

Outline

- Introduction
- Defining Quality Loss
  - Intuition, base model, full model
- Quality Loss in NLP Services
  - Representative queries, data sets, infrastructure, results
- Quality-Aware Cache Provisioning
- Conclusion

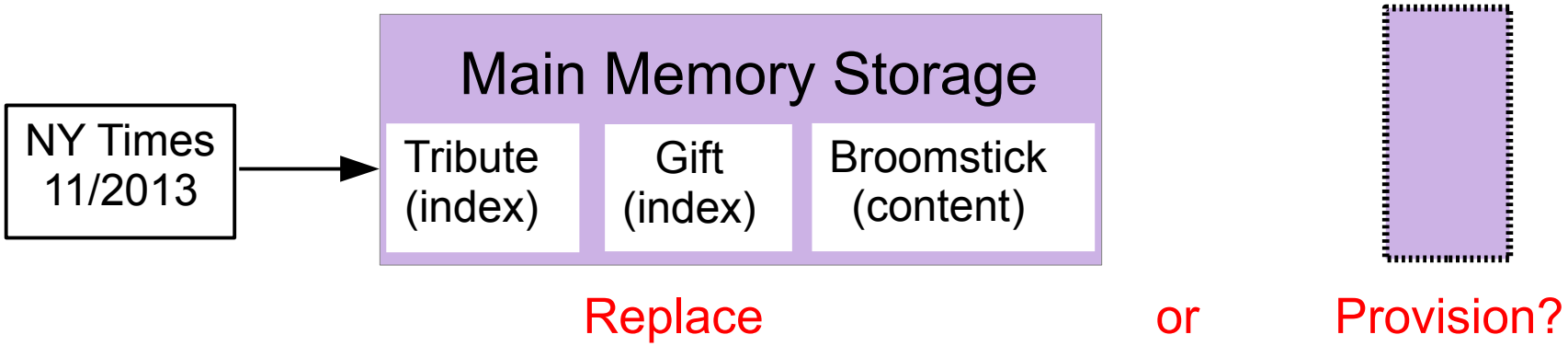
DRAM Trends



Provisioning Cost = #GB RAM * $/GB RAM

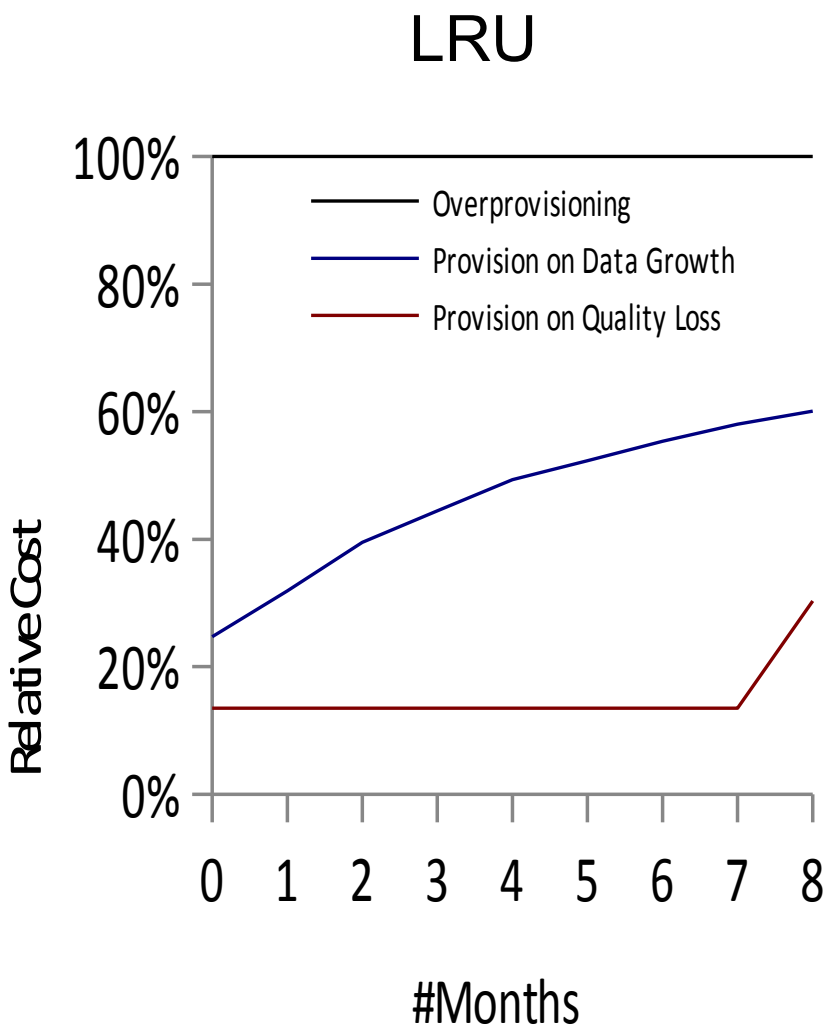
http://www.jcmit.com/memoryprice.htm

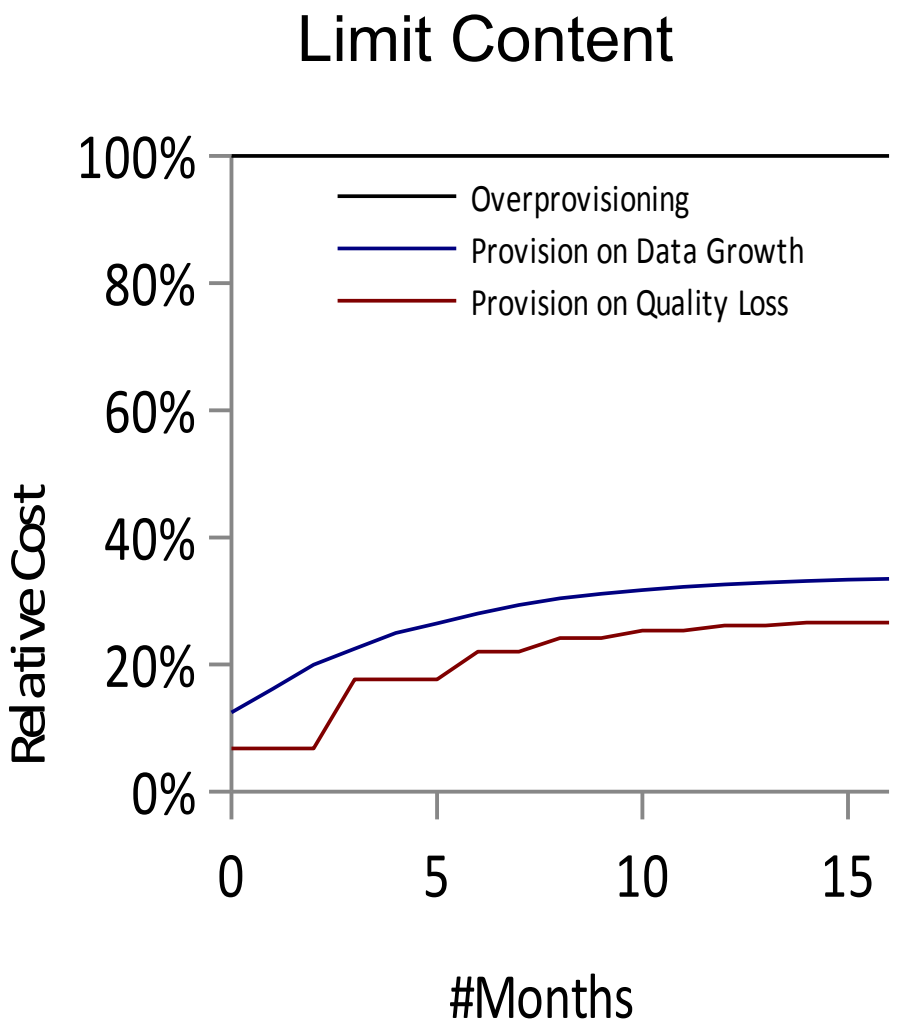
Cache Provisioning Policies



- Over provisioning
  - Always have more RAM available than data
- Provision on Data Growth
  - Wait for more data to be added
- Provision on Quality Loss
  - Wait for quality loss to pass a threshold

Cost Savings





Conclusion

- Data is growing fast, forcing NLP services to respond to queries after accessing only a portion of the data
- NLP services can remove redundant content and/or terms from distributed caches with little quality loss
- New cache management approach: Wait until quality loss occurs before provisioning DRAM to reduce costs